Package 'discovr'

July 22, 2025

Type Package

Title Interactive Tutorials and Data for ``Discovering Statistics Using R and RStudio''

Version 0.2.2

Language en-GB

Maintainer Andy Field <andyf@sussex.ac.uk>

Description Interactive 'R' tutorials and datasets for the textbook Field (2026), ``Discovering Statistics Using R and RStudio'', <https://www.discovr.rocks/>. Interactive tutorials cover general workflow in 'R' and 'RStudio', summarizing data, visualizing data, fitting models and bias, correlation, the general linear model (GLM), moderation, mediation, missing values, comparing means using the GLM (analysis of variance), comparing adjusted means (analysis of covariance), factorial designs, repeated measures designs, exploratory factor analysis (EFA). There are no functions, only datasets and interactive tutorials.

License GPL-3

URL https://www.discovr.rocks, https://github.com/profandyfield/discovr

BugReports https://github.com/profandyfield/discovr/issues

Depends learnr (>= 0.11.4), R (>= 4.2.0)

Imports ggplot2, glue, grDevices, scales

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

NeedsCompilation no

Author Andy Field [aut, cre, cph]

Repository CRAN

Date/Publication 2025-06-06 13:00:10 UTC

Contents

---

acdc *Oxoby (2008) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
acdc
```

Format

A tibble with 36 rows and 2 variables.

Details

AC/DC are one one of the best-selling hard rock bands in history, with around 100 million certified sales, and an estimated 200 million actual sales. In 1980 their original singer Bon Scott died of alcohol poisoning and choking on his own vomit. He was replaced by Brian Johnson who has been their singer ever since. Debate rages with unerring frequency within the rock music press over who is the better frontman. The conventional wisdom is that Bon Scott was better although personally, and I seem to be somewhat in the minority here, I prefer Brian Johnson. Anyway, Robert Oxoby in a playfull paper decided to put this argument to bed once and for all (Oxoby, 2008). Using a task from experimental economics called the ultimatum game, individuals are assigned the role of either proposer or responder and paired randomly. Proposers are allocated $10 from which they have to make a financial offer to the responder (i.e., $2). The responder can accept or reject this offer. If the offer is rejected neither party gets any money, but if the offer is accepted the responder keeps the offered amount (e.g., $2), and the proposer keeps the original amount minus what they offered (e.g., $8). For half of the participants the song 'It's a long way to the top' sung by Bon Scott was playing in the background, for the remainder 'Shoot to thrill' sung by Brian Johnson was playing. Oxoby measured the offers made by proposers, and the minimum offers that responders accepted (called the minimum acceptable offer). He reasoned that people would accept lower offers and propose higher offers when listening to something they like (because of the 'feel-good factor' the music creates). Therefore, by comparing the value of offers made and the minimum acceptable offers in the two groups he could see whether people have more of a feel good factor when listening to Bon or Brian. There were 18 people per group.

These data are approximated from graphs within Oxoby (2008). The object contains the following variables:

- **singer**: the type of teaching method used
- **offer**: offer made (in dollars)
- **mao**: the minimum acceptable offer, MAO, in dollars

Source

www.discovr.rocks/csv/acdc.csv

References

- Oxoby, R. J. (2008). On the efficiency of AC/DC: Bon Scott versus Brian Johnson. *Economic Enquiry*, *47*, 598-602. doi:10.1111/j.14657295.2008.00138.x

`album_sales` *Album sales data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
album_sales
```

Format

A tibble with 200 rows and 5 variables.

Details

Fictitious data that imagines a world where I have a cool job in the music industry. Except, it's not *that* cool because my job is to predict album sales (broadly defined in some way that accounts for physical sales, streams and digital sales). In my little fantasy I collect data from 200 releasures (albums). For each one, I have information about the amount spent advertising the album, the number of sales, the number of plays on radio songs from the album had per week, and a rating of the image of the band. The (fictional) data contains the following variables:

- **album_id**: album identifier.
- **adverts**: advertising budget in thousands of whatever currency is used in your country.
- **sales**: the number of album sales (physical, digital, streams)
- **airplay**: the number of times songs from the album were played on radio thew eek before release
- **image**: a rating of the band's image from scale from 0 (dad dancing at a disco) to 10 (sicker than a dog that's eaten a bag of onions)

Source

www.discovr.rocks/csv/album_sales.csv

`alien_scents` *Alien scents*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
alien_scents
```

Format

A tibble with 50 rows and 4 variables.

Details

The aliens, excited by humans' apparent inability to train sniffer dogs to detect them (see sniffer_dogs), decided to move their invasion plan forward. Aliens are far too wedded to *p*-values in small samples. They decided that they could make themselves even harder to detect by fooling the sniffer dogs by masking their alien smell. After extensive research they agreed that the two most effective masking scents would be human pheromones (which they hoped would make them smell human-like) and fox-pheromones (because they are a powerful, distracting smell for dogs). The aliens started smearing themselves with humans and foxes and prepared to invade. Meanwhile, the top-secret government agency for Training Extra-terrestrial Reptile Detection (TERD) had got wind of their plan and set about testing how effective it would be. They trained 50 sniffer dogs. During training, these dogs were rewarded for making vocalizations while sniffing alien space lizards. On the test trials, the 50 dogs were allowed to sniff 9 different entities for 1-minute each: 3 alien space lizards, 3 shapeshifting alien space lizard who had taken on humanoid form, and 3 humans. Within each type of entity, 1 had no masking scent, 1 was smothered in human pheromones and 1 wore fox pheromones. The number of vocalizations made during each 1-minute sniffing session was recorded.

- **dog_id**: the id of the 50 sniffer dogs
- **entity**: the entity being sniffed by the sniffer dog (alien, alien in humanoid form (shapeshifter), human)
- **scent_mask**: the scent the entity used to mask their natural odour (None, human pheromones, fox pheromones)
- **vocalizations**: the number of vocalizations made by the dog during a 1-minute sniff

Source

www.discovr.rocks/csv/alien_scents.csv

amolad_pal *A Matter of Life and Death palette*

Description

Colour palette based on Iron Maiden's A Matter of Life and Death album sleeve.

Usage

```
amolad_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_amolad(n, type = "discrete", reverse = FALSE, ...)

scale_colour_amolad(n, type = "discrete", reverse = FALSE, ...)

scale_fill_amolad(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale

aesthetics The names of the aesthetics that this scale works with.

scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

palette A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., scales::pal_hue()).

name The name of the scale. Used as the axis or legend title. If waiver(), the default, the name of the scale is taken from the first mapping used for that aesthetic. If NULL, the legend title will be omitted.

breaks One of:

- NULL for no breaks
- waiver() for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

labels One of:

- NULL for no labels
- waiver() for the default labels computed by the transformation object
- A character vector giving labels (must be same length as breaks)
- An expression vector (must be the same length as breaks). See ?plotmath for details.

- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(amolad_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))
```

```
# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_amolad()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_amolad()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_amolad()
```

`angry_pigs` *Video games and aggression example 1*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
angry_pigs
```

Format

A tibble with 336 rows and 4 variables

Details

Angry Birds is a video game in which you fire birds at pigs. A (fabricated) study was set up in which people played Angry Birds and a control game (Tetris) over a 2-year period (1 year per game). They were put in a pen of pigs for a day before the study, and after 1 month, 6 months and 12 months. Their violent acts towards the pigs were counted. The (fictional) data contains

- **id**: participant ID
- **game**: whether the participant had been assigned to play angry pigs or tetris

- **time**: the time at which aggressive acts were measured (Baseline, 1 month, 6 months and 12 months)
- **aggression**: the number of aggressive acts towards pigs

Source

www.discovr.rocks/csv/speed_date.csv

`angry_real` *Video games and aggression example 2*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
angry_real
```

Format

A tibble with 504 rows and 4 variables

Details

Angry Birds is a video game in which you fire birds at pigs. A (fabricated) study was set up in which people played Angry Birds and a control game (Tetris) over a 2-year period (1 year per game). The participant's violent acts in everyday life were monitored before the study, and after 1 month, 6 months and 12 months. The (fictional) data contains

- **id**: participant ID
- **game**: whether the participant had been assigned to play angry pigs or tetris
- **time**: the time at which aggressive acts were measured (Baseline, 1 month, 6 months and 12 months)
- **aggression**: the number of aggressive acts in everyday life

Source

www.discovr.rocks/csv/speed_date.csv

`animal_bride` *Animal bride data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
animal_bride
```

Format

A tibble with 20 rows and 3 variables.

Details

Fictitious data inspired by two news stories that I enjoyed. The first was about a Sudanese man who was forced to marry a goat after being caught having sex with it. I'm not sure he treated the goat to a nice dinner in a posh restaurant before taking advantage of her, but either way you have to feel sorry for the goat. I'd barely had time to recover from that story when another appeared about an Indian man forced to marry a dog to atone for stoning two dogs and stringing them up in a tree 15 years earlier. Why anyone would think it's a good idea to enter a dog into matrimony with a man with a history of violent behaviour towards dogs is beyond me. Still, I wondered whether a goat or dog made a better spouse. I found (but not really) some other people who had been forced to marry goats and dogs and measured their life satisfaction and, also, how much they like animals. The data contains the following variables:

- **wife**: whether the person married a goat or a dog
- **animal**: how much the person likes animals
- **life_satisfaction**: the person's life satisfaction score
- **wife**: Whether the person married a goat or a dog
- **animal**: How much the person likes animals
- **life_satisfaction**: The person's life satisfaction score

Source

www.discovr.rocks/csv/animal_bride.csv

animal_dance *Dancing cats and dogs data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
animal_dance
```

Format

A tibble with 270 rows and 3 variables.

Details

Fictional data about dancing cats and dogs. A researcher was interested in whether animals could be trained to dance. He took 200 cats and 70 dogs and tried to train them to line-dance by giving them either food or affection as a reward for dance-like behaviour. At the end of the week he counted how many animals could line-dance and how many could not. The object contains the following variables:

- **training**: factor describing whether the animal was trained using food of affection as a reward
- **dance**: factor describing whether the cat danced or not
- **animal**: factor describing whether the animal was a cat or a dog

Source

www.discovr.rocks/csv/animal_dance.csv

beckham_1929 *Beckham (1929) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
beckham_1929
```

Format

A tibble with 16 rows and 5 variables.

Details

During my psychology degree I spent a lot of time reading about the civil rights movement in the USA. Instead of reading psychology, I read about Malcolm X and Martin Luther King Jr. For this reason I find Beckham's 1929 study of black Americans a fascinating historical piece of research. Beckham was a black American who founded the psychology laboratory at Howard University, Washington, DC and his wife Ruth was the first black woman ever to be awarded a PhD (also in psychology) at the University of Minnesota. To put some context on the study, it was published 36 years before the Jim Crow laws were finally overthrown by the Civil Rights Act of 1964, and in a time when black Americans were segregated, openly discriminated against and victims of the most abominable violations of civil liberties and human rights (I recommend James Baldwin's superb The fire next time for an insight into the times). The language of the study and the data from it are an uncomfortable reminder of the era in which it was conducted.

Beckham sought to measure the psychological state of 3443 black Americans with three questions. He asked them to answer yes or no to whether they thought black Americans were happy, whether they personally were happy as a black American, and whether black Americans should be happy. Beckham did no formal statistical analysis of his data (Fisher's article containing the popularized version of the chi-square test was published only 7 years earlier in a statistics journal that would not have been read by psychologists). I love this study, because it demonstrates that you do not need elaborate methods to answer important and far-reaching questions; with just three questions, Beckham told the world an enormous amount about very real and important psychological and sociological phenomena. These are the data from that study. The data contains the following variables:

- **profession**: Profession of respondents
- **response**: response to the question as yes or no
- **happy**: frequencies of reponses to a question about whether black Americans were happy
- **you_happy**: frequencies of reponses to a question about whether they personally were happy
- **should_be_happy**: frequencies of reponses to a question about whether black Americans should be happy

Source

www.discovr.rocks/csv/beckham_1929.csv

References

- Beckham, A. S. (1929). Is the Negro happy? A psychological analysis. *Journal of Abnormal and Social Psychology*, 24, 186–190. doi:10.1037/h0072938

biggest_liar *The biggest liar data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
biggest_liar
```

Format

A tibble with 68 rows and 4 variables.

Details

Fictional data based on the World's Biggest Liar competition held annually at the Santon Bridge Inn in Wasdale (in the Lake District, UK). Each year locals are encouraged to attempt to tell the biggest lie in the world. I wanted to test a theory that more creative people will be able to create taller tales. I gathered together 68 past contestants from this competition and noted where they were placed in the competition (first, second, third, etc.); I also gave them a creativity questionnaire (maximum score 60). The data set has four variables

- **id**: Participant id
- **creativity**: Creativity score (maximum score 60)
- **position**: position in competition as a numeric variable from 1 (first place) to 5 (fifth place)
- **novice**: factor coding whether this was the participant's first time in the competition (*first time*) or if they had entered before (*previous entrant*).

Source

www.discovr.rocks/csv/biggest_liar.csv

big_hairy_spider *Big hairy spider data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
big_hairy_spider
```

Format

A tibble with 24 rows and 3 variables.

Details

Is arachnophobia (fear of spiders) specific to real spiders or will pictures of spiders evoke similar levels of anxiety? Twelve arachnophobes were asked to play with a big hairy tarantula with big fangs and an evil look in its eight eyes and at a different point in time were shown only photos of the same spider. The participants' anxiety was measured in each case. The (fictional) data contains the following variables:

- **id**: the participant's first name
- **spider_type**: whether the spider stimulus was a real spider or a photo of a spider
- **anxiety**: the participant's anxiety after exposure to the stimulus

Source

www.discovr.rocks/csv/big_hairy_spider.csv

---

bnw_pal *Brave New World palette*

Description

Colour palette based on Iron Maiden's Brave New World album sleeve.

Usage

```
bnw_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_bnw(n, type = "discrete", reverse = FALSE, ...)

scale_colour_bnw(n, type = "discrete", reverse = FALSE, ...)

scale_fill_bnw(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale aesthetics The names of the aesthetics that this scale works with. scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for `y` axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.
`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(bnw_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_bnw()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_bnw()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_bnw()
```

`bronstein_2019` *Bronstein et al. (2019) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
bronstein_2019
```

Format

A tibble with 947 rows and 5 variables

Details

The rapid increase in 'fake news' and misinformation is a worrying trend in recent years. Perhaps more worrying is how widely some of this news is taken as fact. Researchers have started to look at what characteristics predict susceptibility to fake news. Bronstein et al. (2019) hypothesised that delusion-prone individuals may be more likely to believe fake news because of their tendency to engage in less analytic and open-minded thinking. They conducted two online studies that got merged into a single analysis to test this hypothesis. This object is a subset of variables from their data (I have changed the variable names to match the constructs measured rather than the scales used to measure them). The full dataset is available at doi:10.1016/j.jarmac.2018.09.005.

- **id** (ResponseID in the original dataset): participant ID
- **fake_newz** (ZBelief_Fake in the original dataset): participants viewed 12 fake news headlines, each with a brief description and photo, and rated their accuracy (1 = Not at all accurate, 4 = Very accurate). This variable is the average rating converted to a z-score.
- **delusionz** (ZPDI_Total in the original dataset): Peter's et al Delusion Inventory (PDI), which uses statements such as "Do you ever feel as if there is a conspiracy against you?" to gauge a person's propensity for delusion-like thinking. Again, scores were converted to z-scores.
- **thinkz_open** (ZAOT_Total in the original dataset): open minded thinking was assessed with the Actively Open-minded Thinking (AOT) scale, on which people endorse statements such as "A person should always consider new possibilities" using a six-point scale (1 = strongly disagree, 6 = strongly agree). The total score was again converted to z.
- **thinkz_anal** (ZRF_Total in the original dataset): Analytic thinking was assessed using the Cognitive Reflection Test (CRT), which uses several problems that have intuitive-but-incorrect responses. Participants must override their intuition to get the correct answer. Over 7 items, higher scores (converted to z-scores again) indicate a greater tendency to use an analytic cognitive style.

Source

www.discovr.rocks/csv/bronstein_2019.csv

References

- Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8(1), 108–117. doi:10.1016/j.jarmac.2018.09.005

bronstein_miss_2019 *Bronstein et al. (2019) data with missing values inserted*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
bronstein_miss_2019
```

Format

A tibble with 947 rows and 5 variables

Details

A version of the Bronstein et al. (2019) fake news data (bronstein_2019) but with missing values inserted using MCAR amputation (with the help of the `mice` package and `ampute()` function). For details of variables see bronstein_2019.

Source

www.discovr.rocks/csv/bronstein_miss_2019.csv

References

- Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8(1), 108–117. doi:10.1016/j.jarmac.2018.09.005

catterplot *Catterplot data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
catterplot
```

Format

A tibble with 78 rows and 2 variables.

Details

Fictional data for plotting a catterplot. The object contains the following variables:

- **dinner_time**: the time (hours) since the cat was last fed
- **meow**: How loud the cat's purr is

Source

www.discovr.rocks/csv/catterplot.csv

---

cat_dance *Dancing cats data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
cat_dance
```

Format

A tibble with 200 rows and 2 variables.

Details

Fictional data about dancing cats. A researcher was interested in whether animals could be trained to dance. He took 200 cats and tried to train them to line-dance by giving them either food or affection as a reward for dance-like behaviour. At the end of the week he counted how many animals could line-dance and how many could not. The object contains the following variables:

- **reward**: factor describing whether the cat was trained using food of affection as a reward
- **dance**: factor describing whether the cat danced or not

Source

www.discovr.rocks/csv/cat_dance.csv

`cat_reg` *Cat regression data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
cat_reg
```

Format

A tibble with 200 rows and 7 variables.

Details

Fictional data illustrating how the chi-square test is a linear model. It's about line dancing cats. The object contains the following variables:

- **reward**: whether the cat was trained using food (0) of affection (1) as a reward
- **dance**: Whether the cat danced (1) or not (0)
- **interaction**: the interaction of dance and reward (i.e. dance multiplied by reward)
- **observed**: the observed frequency for the combination of dance and reward
- **expected**: the expected frequency for the combination of dance and reward
- **lmobserved**: the natural logarithm of the observed frequency for the combination of dance and reward
- **lmexpected**: the natural logarithm of the expected frequency for the combination of dance and reward

Source

www.discovr.rocks/csv/cat_regression.csv

`cetinkaya_2006` *Cetinkaya and Domjan (2006) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
cetinkaya_2006
```

Format

A tibble with 59 rows and 6 variables.

Details

Some quail develop fetishes. Really. In studies where a terrycloth object acts as a sign that a mate will shortly become available, some quail start to direct their sexuial behaviour towards the terrycloth object. In evolutionary terms, this fetishistic behaviour seems counterproductive because sexual behaviour becomes directed towards something that cannot provide reproductive success. However, perhaps this behaviour serves to prepare the organism for the 'real' mating behaviour.

Cetinkaya and Domjan (2006) sexually conditioned male quail. All quail experienced the terrycloth stimulus and an opportunity to mate, but for some the terrycloth stimulus immediately preceded the mating opportunity (paired group) whereas others experienced a 2-hour delay (this acted as a control group because the terrycloth stimulus did not predict a mating opportunity). In the paired group, quail were classified as fetishistic or not depending on whether they engaged in sexual behaviour with the terrycloth object.

During a test trial the quail mated with a female and the researchers measured the percentage of eggs fertilized, the time spent near the terrycloth object, the latency to initiate copulation, and copulatory efficiency. If this fetishistic behaviour provides an evolutionary advantage then we would expect the fetishistic quail to fertilize more eggs, initiate copulation faster and be more efficient in their copulations. These are the data from that study. The data contains the following variables:

- **groups**: The group to which each quail belonged (Fetishistics, NonFetishistics, or Control)
- **paired**: whether the terrycloth predicted a mating opportunity (paired) or not (unpaired)
- **egg_percent**: percentage of eggs fetilised by male
- **duration**: Time spent near terrycloth object
- **latency**: Time taken to initiate copulation
- **efficiency**: Copulatory efficiancy

Source

www.discovr.rocks/csv/cetinkaya_2006.csv

References

- Cetinkaya, H., & Domjan, M. (2006). Sexual fetishism in a quail (*Coturnix japonica*) model system: Test of reproductive success. *Journal of Comparative Psychology*, *120*, 427–432. doi:10.1037/07357036.120.4.427

`chamorro_premuzic` *Chamorro-Premuzic, et al. (2008) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
chamorro_premuzic
```

Format

A tibble with 430 rows and 12 variables.

Details

There is some evidence that students tend to pick courses of lecturers they perceive to be enthusastic and good communicators. In a fascinating study, Tomas Chamorro-Premuzic and his colleagues (Chamorro-Premuzic, Furnham, Christopher, Garwood, & Martin, 2008) tested the hypothesis that students tend to like lecturers who are like themselves. The authors measured students' own personalities using a very well-established measure (the NEO-FFI) which measures five fundamental personality traits: neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. Students also completed a questionnaire in which they were given descriptions (e.g., 'warm: friendly, warm, sociable, cheerful, affectionate, outgoing') and asked to rate how much they wanted to see this in a lecturer from -5 (I don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer). The characteristics were the same as those measured by the NEO-FFI. As such, the authors had a measure of how much a student had each of the five core personality characteristics, but also a measure of how much they wanted to see those same characteristics in their lecturer. These are the data from that study. The data contains the following variables:

- **age**: participant age (years)
- **sex**: participant's biological sex
- **stu_neurotic**: Student neuroticism score on the NEO-FFI
- **stu_extro**: Student extroversion score on the NEO-FFI
- **stu_open**: Student openness to experience score on the NEO-FFI
- **stu_agree**: Student agreeableness score on the NEO-FFI
- **stu_consc**: Student conscientiousness score on the NEO-FFI

- **lec_neurotic**: Student rating of how much they wanted the characteristic of neuroticism in their lecturers from -5 (I don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer)
- **lec_extro**: Student rating of how much they wanted the characteristic of extroversion in their lecturers from -5 (I don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer)
- **lec_open**: Student rating of how much they wanted the characteristic of openness to experience in their lecturers from -5 (I don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer)
- **lec_agree**: Student rating of how much they wanted the characteristic of agreeableness in their lecturers from -5 (I don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer)
- **lec_consc**: Student rating of how much they wanted the characteristic of conscientiousness in their lecturers from -5 (I don't want this characteristic at all) through 0 (the characteristic is not important) to +5 (I really want this characteristic in my lecturer)

Source

www.discovr.rocks/csv/chamorro_premuzic.csv

References

- Chamorro-Premuzic, T., Furnham, A., Christopher, A. N., Garwood, J., & Neil Martin, G. (2008). Birds of a feather: Students' preferences for lecturers' personalities as predicted by their own personality and learning approaches. *Personality and Individual Differences*, 44(4), 965–976. doi:10.1016/j.paid.2007.10.032

`child_aggression` *Child aggression data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
child_aggression
```

Format

A tibble with 666 rows and 6 variables.

Details

A study was carried out to explore the relationship between aggression and several potential predicting factors in 666 children who had an older sibling. Variables measured were **parenting_style** (high score = bad parenting practices), **computer_games** (high score = more time spent playing computer games), **television** (high score = more time spent watching television), **diet** (high score = the child has a good diet low in harmful additives), and **sibling_aggression** (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. The data contain the following variables:

- **aggression**: The child's aggression
- **television**: Time spent watching television (high score = more time spent watching television)
- **computer_games**: Time spent playing video games (high score = more time spent playing video games)
- **sibling_aggression**: Agression in older sibling (high score = more aggression seen in their older sibling).
- **diet**: The child's diet (high score = the child has a good diet low in harmful additives).
- **parenting_style**: the parent's parenting style (high score = bad parenting practices).

Source

www.discovr.rocks/csv/child_aggression.csv

---

coldwell_2006 *Coldwell, Pike and Dunn (2006) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
coldwell_2006
```

Format

A tibble with 118 rows and 9 variables.

Details

Coldwell, Pike and Dunn (2006) investigated whether household chaos predicted children's problem behaviour over and above parenting. From 118 families they recorded the age and gender of the youngest child (**child_age** and **child_gender**). They measured dimensions of the child's perceived relationship with their mum: (1) warmth/enjoyment (**child_warmth**), and (2) anger/hostility (**child_anger**). Higher scores indicate more warmth/enjoyment and anger/hostility respectively. They measured the mum's perceived relationship with her child, resulting in dimensions of positivity (**mum_pos**) and negativity (**mum_neg**). Household chaos (**chaos**) was assessed. The outcome variable was the child's adjustment (**sdq**): the higher the score, the more problem behaviour the child was reported to be displaying. These data are from this study. The data contain the following variables:

- **family_id**: The family id
- **child_age**: Age of the youngest child
- **child_gender**: Gender of the youngest child
- **child_warmth**: Perceived warmth of the child to the mother.
- **child_anger**: Perceived anger of the child towards to the mother.
- **mum_pos**: the mother's perceived positivity towards her child.
- **mum_neg**: the mother's perceived negativity towards her child.
- **chaos**: household chaos.
- **sdq**: the child's adjustment on the strengths and difficulties questionnaire (SDQ).

Source

www.discovr.rocks/csv/coldwell_2006.csv

---

cosmetic *Cosmetic surgery data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
cosmetic
```

Format

A tibble with 1376 rows and 7 variables.

Details

Fictitious example based on quality of life predicted from undergoing cosmetic surgery. Cosmetic surgery is on the increase. For example, in the USA, there was a 1600% increase in cosmetic surgical and non-surgical treatments between 1992 and 2002. There are two main reasons to have cosmetic surgery: (1) to help a physical problem; and (2) to change your external appearance when there is no underlying physical pathology. This example uses fictitious data looks at the effects of cosmetic surgery on quality of life. The variables in the data are:

- **id**: The participant id
- **clinic**: Categorical variable that indicates which of 21 clinics the person attended to have their surgery.
- **reason**: Categorical variable that indicates whether the person had or is waiting to have surgery purely to change their appearance or because of a physical reason.
- **base_qol**: Quality of life pre-surgery on a percentage scale (0% = the worst possible quality of life, 100% = the best possible quality of life).
- **post_qol**: Quality of life after cosmetic surgery on a percentage scale (0% = the worst possible quality of life, 100% = the best possible quality of life).
- **days**: The number of days since surgery.
- **bdi**: Levels of depression using the Beck Depression Inventory (BDI).

Source

www.discovr.rocks/csv/cosmetic.csv

`daniels_2012` *Daniels (2012) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
daniels_2012
```

Format

A tibble with 4 rows and 7 variables.

Details

Women (and increasingly men) are bombarded with 'idealized' images in the media and there is a growing concern about how these images affect our perceptions of ourselves. Daniels (2012) showed young women images of successful female athletes (e.g., Anna Kournikova) in which they were either playing sport (performance athlete images) or posing in bathing suits (sexualized images). Participants completed a short writing exercise after viewing these images. Each participant saw only one type of image, but several examples. Daniels then coded these written exercises and identified themes, one of which was whether women self-objectified (i.e., commented on their own appearance/attractiveness). Daniels hypothesized that women who viewed the sexualized images (n = 140) would self-objectify (i.e., this theme would be present in what they wrote) more than those who viewed the performance athlete pictures (n = 117, despite what the participants Section of the paper implies). These are the data from that study. The data contains the following variables:

- **picture**: Whether the picture was of a performance athlete or a sexualized athlete
- **theme_present**: whether a particular theme was present or absent from the participant's writing exercise
- **athletes_body**: frequencies for the theme of the athlete's body
- **admiration**: frequencies for the theme of admiration for the athlete
- **role_model**: frequencies for the theme of the athlete being a role model
- **self_evaluation**: frequencies for the theme of self-evaluation
- **self_physical_activity**: frequencies for the theme of self physical activity

Source

www.discovr.rocks/csv/daniels_2012.csv

References

- Daniels, E. (2012). Sexy versus strong: What girls and women think of female athletes. *Journal of Applied Developmental Psychology*, 33, 79–90. doi:10.1016/j.appdev.2011.12.002

---

dark_lord *Subliminal messages data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
dark_lord
```

Format

A tibble with 64 rows and 3 variables.

Details

Both Ozzy Osbourne and Judas Priest have been accused of putting backward masked messages on their albums that subliminally influence poor unsuspecting teenagers into doing things like blowing their heads off with shotguns. A psychologist was interested in whether backward masked messages could have an effect. He created a version of Taylor Swifts' 'Shake it off' that contained the masked message 'deliver your soul to the dark lord' repeated in the chorus. He took this version, and the original, and played one version (randomly) to a group of 32 people. Six months later he played them whatever version they hadn't heard the time before. So, each person heard both the original and the version with the masked message, but at different points in time. The psychologist measured the number of satanic intrusions the person had in the week after listening to each version. The (fictional) data contains the following variables:

- **id**: the participant's id
- **message**: whether the song had a subliminal satanic message or not
- **intrusions**: number of satanic intrusions in the week after hearing the song

Source

www.discovr.rocks/csv/dark_lord.csv

`davey_2003` *Davey et al. (2003) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
davey_2003
```

Format

A tibble with 60 rows and 4 variables.

Details

Many of us have experienced that feeling after we have left the house of wondering whether we remembered to lock the door, close the window, or remove the bodies from the fridge in case the police turn up. However, some people with obsessive compulsive disorder (OCD) check things so excessively that they might, for example, take hours to leave the house. One theory is that this checking behaviour is caused by the mood you are in (positive or negative) interacting with the rules you use to decide when to stop a task (do you continue until you feel like stopping, or until you have done the task as best as you can?). Davey et al. (2003) tested this hypothesis by asking participants to think of as many things as they could that they should check before going on holiday (**checks**) after putting them into a negative, positive or neutral mood (**mood**). Within each mood

group, half of the participants were instructed to generate as many items as they could, whereas the remainder were asked to generate items for as long as they felt like continuing the task (**stop_rule**). These are the data from that study. The data contains the following variables:

- **id**: Participant id
- **mood**: whether a particular was randomly allocated to a negative, positive or neutral mood induction condition.
- **stop_rule**: whether a particular was randomly allocated to a condition in which they were instructed to undertake a task using an 'as many as can' stop rule or a 'feel like continuing' stop rule.
- **checks**: number of things participants

Source

www.discovr.rocks/csv/davey_2003.csv

References

- Davey, G. C. L., Startup, H. M., Zara, A., MacDonald, C. B., & Field, A. P. (2003). The perseveration of checking thoughts and mood–as–input hypothesis. *Journal of Behavior Therapy and Experimental Psychiatry*, 34(2), 141–160. doi:10.1016/S00057916(03)000351

`df_beta` *DF beta data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
df_beta
```

Format

A tibble with 30 rows and 3 variables.

Details

Fictitious data to illustrate the DF Beta statistic. The tibble contains the following variables:

- **case**: a number from 0-30 indicating the entity (case)
- **x**: Imaginatively named predictor variable
- **y**: The creativity was flowing the day I generated these data - another imaginatively named variable. This time it's the outcome variable.

Source

www.discovr.rocks/csv/df_beta.csv

`discovr` *discovr: Resources for Discovering Statistics Using R and RStudio (Field, 2023)*

Description

The `discovr` package contains interactive `learnr` tutorials and datasets that accompany my textbook Discovering Statistics Using R and RStudio.

Who is the package aimed at?

Anyone teaching from or reading Discovering Statistics Using R and RStudio should find these resources useful.

Interactive tutorials

Getting started:

I recommend working through this tutorial on how to install, set up and work within R and RStudio before starting the interactive tutorials.

Running a tutorial:

To run each tutorial execute

```
learnr::run_tutorial("name_of_tutorial", package = "discovr")
```

Replacing `name_of_tutorial` with the name in bold below. For example, to load the tutorial `discovr_02` execute:

```
learnr::run_tutorial("discovr_02", package = "discovr")
```

- **discovr_01**: Key concepts in R (functions and objects, packages and functions, style, data types, tidyverse, tibbles)
- **discovr_02**: Summarizing data (frequency distributions, grouped frequency distributions, relative frequencies, histograms, mean, median, variance, standard deviation, interquartile range)
- **discovr_03**: Confidence intervals: interactive app demonstrating what a confidence interval is, computing normal and bootstrap confidence intervals using R, adding confidence intervals to data summaries
- **discovr_05**: Visualizing data. The `ggplot2` package, boxplots, plotting means, violin plots, scatterplots, grouping by colour, grouping using facets, adjusting scales, adjusting positions.
- **discovr_06**: The beast of bias. Restructuring data from messy to tidy format (and back). Spotting outliers using histograms and boxplots. Calculating z-scores (standardizing scores). Writing your own function. Using z-scores to detect outliers. Q-Q plots. Calculating skewness, kurtosis and the number of valid cases. Grouping summary statistics by multiple categorical/grouping variables.

- **discovr_07**: Associations. Plotting data with `GGally`. Pearson's *r*, Spearman's Rho, Kendall's tau, robust correlations.
- **discovr_08**: The general linear model (GLM). Visualizing the data, fitting GLMs with one and two predictors. Viewing model parameters with `broom`, model parameters, standard errors, confidence intervals, fit statistics, significance, Bayes factors and Bayesian estimates (using default priors).
- **discovr_09**: Categorical predictors with two categories (comparing two means). Comparing two independent means, comparing two related means, effect sizes, robust comparisons of means (independent and related), Bayes factors and estimation (independent and related means).
- **discovr_10**: Moderation and mediation. Centring variables (grand mean centring), specifying interaction terms, moderation analysis, simple slopes analysis, Johnson-Neyman intervals, mediation with one predictor, direct and indirect effects, mediation using lavaan.
- **discovr_11**: Comparing several means. Essentially 'One-way independent ANOVA' but taught using a general linear model framework. Covers setting contrasts (dummy coding, contrast coding, and linear and quadratic trends), the F-statistic and Welch's robust F, robust parameter estimation, heteroscedasticity-consistent tests of parameters, robust tests of means based on trimmed data, *post hoc* tests, Bayes factors.
- **discovr_12**: Comparing means adjusted for other variables. Essentially 'Analysis of Covariance (ANCOVA)' designs but taught using a general linear model framework. Covers setting contrasts, Type III sums of squares, the F-statistic, robust parameter estimation, heteroscedasticity-consistent tests of parameters, robust tests of adjusted means, *post hoc* tests, Bayes factors.
- **discovr_13**: Factorial designs. Fitting models for two-way factorial designs (independent measures) using both `lm()` and the `afex` package. This tutorial builds on previous ones to show how models can be fit with two categorical predictors to look at the interaction between them. We look at fitting the models, setting contrasts for the two categorical predictors, obtaining estimated marginal means, interaction plots, simple effects analysis, diagnostic plots, partial eta-squared and partial omega-squared, robust models and Bayes factors.
- **discovr_14**: Multilevel models. This tutorial looks at fitting multilevel models using the `lme4` package. It begins with an optional section on checking and coding categorical variables before moving on to show you how to fit and interpret a multilevel model.
- **discovr_14_lme**: Multilevel models. This tutorial looks at fitting multilevel models using the `nlme` package. It begins with an optional section on checking and coding categorical variables before moving on to show you how to fit and interpret a multilevel model.
- **discovr_15**: Repeated measures designs. Fitting models for one- and two-way repeated measures designs using the `afex` package. This tutorial builds on previous ones to show how models can be fit with one or two categorical predictors when these variables have been manipulated within the same entities. We look at fitting the models, setting contrasts for the categorical predictors, obtaining estimated marginal means, interaction plots, simple effects analysis, diagnostic plots, robust models and Bayes factors.
- **discovr_16**: Mixed designs. Fitting models for mixed designs using the `afex` package. This tutorial builds on previous ones to show how models can be fit with one or two categorical predictors when at least one of these variables has been manipulated within the same entities and at least one other has been manipulated using different entities. We look at fitting the

models, setting contrasts for the categorical predictors, obtaining estimated marginal means, interaction plots, simple effects analysis, diagnostic plots, robust models and Bayes factors.

- **discovr_18**: Exploratory Factor Analysis (EFA). Applying factor analysis using the psych package.

Workflow:

The tutorials are self-contained (you practice code in code boxes) so you don't need to use RStudio at the same time. However, to get the most from them I would recommend that you create an RStudio project and within that open (and save) a new R Markdown file each time to work through a tutorial. Within that Markdown file, replicate parts of the code from the tutorial (in code chunks) and use Markdown to write notes about what you have done, and to reflect on things that you have struggled with, or note useful tips to help you remember things. Basically, write a learning journal. This workflow has the advantage of not just teaching you the code that you need to do certain things, but also provides practice in using RStudio itself.

Datasets

See the book or data descriptions for more details. This is a list of available datasets within the package. Raw CSV files are available from the book's website.

- acdc: Data about whether Bon Scott or Brian Johnson is the best singer of AC/DC.
- album_sales: Fictitious data about predicting album sales from advertising, airplay and the band's image.
- alien_scents: Fictitious data about training sniffer dogs to detect alien space lizards when they try to mask their identity with different scents. See also sniffer_dogs.
- angry_pigs: fictitious data about whether playing the video game angry pigs makes people more aggressive towards pigs. See also angry_real.
- angry_real: fictitious data about whether playing the video game angry pigs makes people more aggressive in everyday life. See also angry_pigs.
- animal_bride: Fictitious data about life satisfaction when married to a dog or a goat.
- animal_dance: Fictitious data about training cats and dogs to dance.
- beckham_1929: Data from a study by Beckham (1929).
- big_hairy_spider: Fictitious data about whether anxiety is greater after exposure to real spiders or pictures of spiders.
- biggest_liar: Fictitious data about creativity and telling lies.
- bronstein_2019: Data about whether delusion proneness predicts belief in fake news because of less analytic thinking.
- bronstein_miss_2019: The data in bronstein_2019 but with missing values inserted using MCAR amputation.
- catterplot: Fictitious data for plotting a catterplot.
- cat_dance: Fictitious data about training cats to dance.
- cat_reg: Fictitious data about training cats to dance.
- cetinkaya_2006: data from a study by Cetinkaya and Domjan (2006) about quails with sexual fetishes. Seriously.

- chamorro_premuzic: Data about what students want (personality wise) from their lecturers.
- child_aggression: Fictitious data (based on real research) about predicting aggression in children.
- coldwell_2006: Data predicting childhood adjustment from various parenting variables.
- cosmetic: Fictitious multilevel data predicting quality of life from cosmetic surgery.
- daniels_2012: Data about the effects of sexualised sports images on self-image.
- dark_lord: Fictitious data about the subliminal messages in songs.
- davey_2003: Data about the effects mood and stop rules on checking behaviour.
- dog_training: Data about the training dogs to vocalise when they sniff alien life forms.
- download: Fictitious data about the download music festival and being smelly.
- df_beta: Fictitious data used to illustrate the DF Beta statistic.
- eel: Fictitious data about a randomized control trial to test whether eel therapy is an effective treatment of constipation.
- elephooty: Fictitious data about elephants playing football (soccer).
- escape: Fictitious data about whether I'm a better songwriter than my school bandmate Malcolm.
- essay_marks: Fictitious data about essay marking.
- exam_anxiety: Fictitious data about exam performance, anxiety and revision.
- field_2006: Data that tests a hypothesis that threat information affects children's avoidance of novel animals.
- gallup_2003: Data that tests a hypothesis about why penises have a bell end.
- gelman_2009: Data used to critically evaluate the explanations (and claim) that there are more beautiful women than men in the world.
- glastonbury: More fictitious data about music festivals and being smelly.
- goggles: Fictitious data about whether alcohol affects perception of physical attractiveness.
- goggles_lighting: fictitious data about the moderating effect of lighting on the ratings of attractivenesses of faces after different doses of alcohol.
- grades: Fictitious data about statistics grades.
- ice_bucket: Data about the ice bucket challenge.
- invisibility_base: Fictitious data about how much mischief people would get up to if they had an invisibility cloak using a pre-post study design.
- invisibility_cloak: Fictitious data about how much mischief people would get up to if they had an invisibility cloak using an independent design.
- invisibility_rm: Fictitious data about how much mischief people would get up to if they had an invisibility cloak but using a repeated measures design.
- hangover: fictitious data about the efficacy of different drinks as cures for a hangover.
- hiccups: Fictitious data on digital rectal stimulation and hiccups.
- hill_2007: Data from Hill et al. (2007) testing the effect of different forms of psychoeducation on exercise behaviour.
- honesty_lab: Fictitious data about perceptions of honesty.

- jiminy_cricket: Fictitious data about whether wishing on a star makes you successful.
- johns_2012: Data about whether the colour red is a mating signal to men.
- lambert_2012: Data about whether pornography use is related to relationship commitment and infidelity.
- massar_2012: Data about whether gossiping has an evolutionary function.
- mcnulty_2008: Simulated data to match the results of a study about whether attractiveness is linked to the support given within a relationship.
- men_dogs: Fictitious data about whether men exhibit dog-like behaviours (compared to dogs).
- metal: Fictitious data about whether listening to metal music makes you angry.
- metal_health: Fictitious data about whether listening to heavy metal negatively affects mental health.
- metallica: Data about thrash metal band Metallica.
- miller_2007: Data from Miller et al. (2007) testing the hidden-estrus theory.
- mixed_attitude: Fictitious data about whether different type of imagery in advertising affect ratings of different types of drinks based on the gender identity of the participant.
- murder: Fictitious data about the number of murder each month at three street locations (Ruskin Avenue, Acacia Avenue and Rue Morgue).
- muris_2008: Data about whether you can train children to interpret ambiguous situations in a particular way.
- nichols_2004: Data from the development of the Internet Addiction Scale, IAS (Nichols & Nicki, 2004).
- notebook: Fictitious data about whether watching the film the notebook is emotionally arousing.
- ocd: Fictitious data about interventions for obsessive compulsive disorder.
- ong_2011: Data about social media profile pictures and personality traits.
- ong_tidy: Data about social media profile pictures and personality traits.
- penalty: Fictitious data about predictors of penalty kick success in soccer (or whatever sport you enjoy).
- profile_pic: Fictitious data related to whether the number of friend requests from random people on social media is affected by whether your profile picture depicts you as single or part of a romantic couple.
- pubs: Data illustrating the difference between an outlier and an influential case.
- puppies: Fictitious data related to whether puppy therapy works.
- puppy_love: Fictitious data related to whether puppy therapy works when you adjust for a person's love of puppies.
- raq: Fictitious data relating to a fictional questionnaire about R anxiety that is not an actual questionnaire.
- r_exam: Fictitious data relating to an R exam at two universities.
- reality_tv: Fictitious data relating to whether being on a reality TV show exacerbates personality disorder traits.

- roaming_cats: Fictitious data about how far cats roam from their homes.
- rollercoaster: Fictitious data about how roller-coaster induced fear affects attractiveness ratings.
- santas_log: Fictitious data related to whether the type and quantity of treat consumed on Christmas night affects whether elves successfully deliver presents.
- self_help: Fictitious data about whether self-help books improve relationship satisfaction.
- self_help_dsur: Fictitious data about whether self-help books improve relationship satisfaction compared to statistics books.
- sharman_2015: Data from Sharman & Dingle (2015) about whether listening to metal music increases anger.
- shopping: Fictitious data about shopping.
- sniffer_dogs: Fictitious data about training sniffer dogs to detect alien space lizards when they try to mask their identity with different scents. See also alien_scents.
- social_anxiety: Fictitious (I think) data about whether social anxiety symptoms are specific to social anxiety.
- social_media: Fictitious data about the effects of social media on grammar.
- soya: fictitious data about the effects of eating soya on sperm count.
- speed_date: Fictitious data related to the extent to which interest in dating someone is affected by their looks, personality or the dating strategy they adopt.
- stalker: fictitious data about therapy for stalking.
- students: I can't even remember what this data file contains.
- superhero: fictitious data about whether wearing different superhero costumes leads to more severe physical injuries.
- supermodel: Fictitious data about supermodel salaries.
- switch: Fictitious data relating to whether injuries from playing video console games can be mitigated by a warm up.
- tablets: Fictitious data about predicting the desirability of computing tablets.
- tea_15: Fictitious data based on real data about cognitive functioning and drinking tea.
- tea_716: Fictitious data based on real data about cognitive functioning and drinking tea.
- teaching: Fictitious data about the success of different methods of teaching.
- teach_method: More fictitious data about the success of different methods of teaching.
- text_messages: fictitious data about whether use of messaging apps ruins your grammar.
- tosser: Fictitious data relating to a fictional questionnaire about The Teaching of Statistics for Scientific Experiments, which is fictional.
- tuk_2011: Data about whether needing to urinate helps decision making.
- tumour: fictitious data about mobile phone use and brain tumours.
- tutor_marks: fictitious data comparing 4 tutors marks of the same essays.
- van_bourg_2020: Data from van Bourg et al (2020) relating to whether dogs would release their distressed owners from a box.

- video_games: Fictitious data about the relationship between video game use, callous unemotional traits and aggression.
- williams: Data relating to the development of a questionnaire to measure organizational ability.
- xbox: Fictitious data relating injuries to the type of video console game played and the console it was played on.
- tuk_2011: Data about whether needing to urinate helps decision making.
- zhang_sample: Data about whether performing a maths test under a different name assists performance.
- zibarras_2008: Data from Zibarras, Port, and Woods (2008) relating to the relationship between personality and creativity.
- zombie_growth: fictitious data that mimics a randomised control trial over time testing an intervention to transform zombies back to their pre-zombified state.
- zombie_rehab: fictitious data that mimics a randomised control trial testing an intervention to transform zombies back to their pre-zombified state in different clinics.

Smart Alex solutions

Solutions for end of chapter tasks are available at www.discovr.rocks/solutions/alex/.

Labcoat Leni solutions

Solutions for the Labcoat Leni tasks are available at www.discovr.rocks/solutions/leni/.

Chapter code

Although I recommend working through the interactive solutions, each book Chapter has online code and a downloadable R Markdown file available from www.discovr.rocks/solutions/leni/.

Colour palettes

Colour palettes:

A colour blind-friendly pallette based on Okabe and Ito. Also colour themes based around the studio albums of my favourite band Iron Maiden. If you're wondering why some albums are missing, here's the explanation: X Factor (would basically be 8 shades of grey), Fear of the Dark (terrible album), The Book of Souls (would be 8 shades of black). The following palettes exist.

- amolad_pal: Colour palette (8 colour) based on Iron Maiden's A Matter of Life and Death album sleeve. In `ggplot2` use scale_color_amolad and scale_fill_amolad.
- bnw_pal: Colour palette (8 colour) based on Iron Maiden's Brave New World album sleeve. In `ggplot2` use scale_color_bnw and scale_fill_bnw.
- dod_pal: Colour palette (8 colour) based on Iron Maiden's Dance of Death album sleeve. In `ggplot2` use scale_color_dod and scale_fill_dod.
- frontier_pal: Colour palette (8 colour) based on Iron Maiden's The Final Frontier album sleeve. In `ggplot2` use scale_color_frontier and scale_fill_frontier.
- im_pal: Colour palette (8 colour) based on Iron Maiden's eponymous album sleeve. In `ggplot2` use scale_color_im and scale_fill_im.

- killers_pal: Colour palette (8 colour) based on Iron Maiden's Killers album sleeve. In ggplot2 use scale_color_killers and scale_fill_killers.
- nob_pal: Colour palette (8 colour) based on Iron Maiden's The Number of the Beast album sleeve. In ggplot2 use scale_color_nob and scale_fill_nob.
- okabe_ito_pal: Colourblind-friendly palette (8 colour) from Okabe and Ito. In ggplot2 use scale_color_oi and scale_fill_oi.
- pom_pal: Colour palette (8 colour) based on Iron Maiden's Piece of Mind album sleeve. In ggplot2 use scale_color_pom and scale_fill_pom.
- power_pal: Colour palette (8 colour) based on Iron Maiden's Powerslave album sleeve. In ggplot2 use scale_color_power and scale_fill_power.
- prayer_pal: Colour palette (8 colour) based on Iron Maiden's No Prayer for the Dying album sleeve. Use scale_color_prayer and scale_fill_prayer.
- senjutsu_pal: Colour palette (10 colour) based on the inner gatefold image of Iron Maiden's Senjutsu album album sleeve. In ggplot2 use scale_color_senjutsu and scale_fill_senjutsu.
- sit_pal: Colour palette (8 colour) based on Iron Maiden's Somewhere in Time album sleeve. In ggplot2 use scale_color_sit and scale_fill_sit.
- ssoass_pal: Colour palette (8 colour) based on Iron Maiden's Seventh Son of a Seventh Son album sleeve. In ggplot2 use scale_color_ssoass and scale_fill_ssoass.
- tol_muted_pal: Palette (9 colour) used in the book from Paul Tol. In ggplot2 use scale_color_tol and scale_fill_tol.
- virtual_pal: Colour palette (8 colour) based on Iron Maiden's Virtual IX album sleeve. In ggplot2 use scale_color_virtual and scale_fill_virtual.

References

- Field, A. P. (2023). *Discovering statistics using R and RStudio* (2nd ed.). London: Sage.

Author(s)

Maintainer: Andy Field <andyf@sussex.ac.uk> [copyright holder]

See Also

Useful links:

- https://www.discovr.rocks
- https://github.com/profandyfield/discovr
- Report bugs at https://github.com/profandyfield/discovr/issues

`dod_pal` *Dance of Death palette*

Description

Colour palette based on Iron Maiden's Dance of Death album sleeve.

Usage

```
dod_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_dod(n, type = "discrete", reverse = FALSE, ...)

scale_colour_dod(n, type = "discrete", reverse = FALSE, ...)

scale_fill_dod(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale`

`aesthetics` The names of the aesthetics that this scale works with.

`scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.

- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(dod_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))
```

```
# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_dod()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_dod()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_dod()
```

`dog_training` *Dog training data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
dog_training
```

Format

A tibble with 668 rows and 3 variables.

Details

Fictional data about dogs being trained to vocalize whenever they sniff an alien life form. Essentially dogs were trained using food rewards. One each trial they sniffed an alien and if they made a vocalization they were rewarded with food. This data shows how vocalisations change over blocks of these training trials. The tibble contains the following variables:

- **id**: name of the fictional dog. (Fun fact, the names are real pet names randomly selected from the pet registry in Seattle: https://randommer.io/pet-names)
- **block**: the block of trials (each block represents 100 trials, so block 1 is the result of the first 100 trials and 5 is the result of trials 400-500)
- **vocalizations**: the percentage of trials during which the dog vocalised.

Source

www.discovr.rocks/csv/dog_training.csv

`download` *Download festival data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
download
```

Format

A tibble with 810 rows and 5 variables.

Details

Fictional data about people stinking at music festivals. A biologist was worried about the potential health effects of music festivals. She went to the Download Music Festival and measured the hygiene of 810 concert-goers over the three days of the festival. She tried to measure every person on every day but, because it was difficult to track people down, there were missing data on days 2 and 3. Hygiene was measured using a standardized technique that results in a score ranging between 0 (you smell like a corpse that's been left to rot up a skunk's arse) and 4 (you smell of sweet roses on a fresh spring day). I know from bitter experience that sanitation is not always great at these places and so the biologist predicted that personal hygiene would go down dramatically over the three days of the festival. The object contains the following variables:

- **ticket_no**: the ticket number of the participant as a factor
- **gender**: The gender with which the participant self-identifies as a factor (male, female, non-binary)
- **day1**: the hygiene score from 0 (eau de toilet) to 4 (eau de toilette) on day 1 of the festival
- **day2**: the hygiene score from 0 (eau de toilet) to 4 (eau de toilette) on day 2 of the festival
- **day3**: the hygiene score from 0 (eau de toilet) to 4 (eau de toilette) on day 3 of the festival

Source

www.discovr.rocks/csv/download_festival.csv

`eddiefy` *Iron Maiden Spotify song features data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
eddiefy
```

Format

A tibble with 173 rows and 17 variables.

Details

A dataset containing the song features data from the Spotify API for the studio albums (190-2015) of the greatest band ever, Iron Maiden. Data were obtained using the spotifyr package.

- **artist_name**: Pointless variable that reminds us that the data relate to Iron Maiden
- **album_name**: Name of the album
- **track_name**: Name of the song
- **year**: Year of release of the album
- **danceability**: From the Spotify API: "Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable."
- **energy**: From the Spotify API: "Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy."
- **key**: From the Spotify API: "The key the track is in. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C-sharp/D-flat, 2 = D, and so on."
- **loudness**: From the Spotify API: "The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db."
- **mode**: From the Spotify API: "Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0."

- **speechiness**: From the Spotify API: "Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks."
- **acousticness**: From the Spotify API: "A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic."
- **instrumentalness**: From the Spotify API: "Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0."
- **liveness**: From the Spotify API: "Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live."
- **valence**: From the Spotify API: "A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)."
- **tempo**: From the Spotify API: "The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration."
- **time_signature**: From the Spotify API: "An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure)."
- **duration_ms**: Song length in milliseconds as an integer value.

Source

www.discovr.rocks/csv/eddiefy.csv

`eel` *Eel data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
eel
```

Format

A tibble with 113 rows and 4 variables.

Details

Lo, Wong, Leung, Law, and Yip (2004) describe a case of a 50-year-old man who reported to the emergency department of a hospital with abdominal pain. An X-ray of the man's abdomen revealed the shadow of an eel. The patient claimed that he inserted the eel to 'relieve constipation'. I'm no medic, but this 'remedy' appears counterintuitive. However, it is an empirical question.

To test the hypothesis that an eel might cure constipation, we could do a randomized controlled trial. Our outcome variable would be 'cured' vs. 'not cured'. The main predictor variable would be the intervention condition (eel treatment arm vs. waiting list/no treatment arm). We might also factor in how many days the patient had been constipated before treatment (a proxy of symptom severity). The (fictional) data contains the following variables:

- **id**: Participant id
- **cured**: Whether the participant cured or not after treatment
- **intervention**: Whether the participant was randomized to the no intervention arm of the trial or the intervention arm
- **duration**: the number of days before treatment that the patient had the problem

Source

www.discovr.rocks/csv/eel.csv

References

- Lo, S. F., Wong, S. H., Leung, L. S., Law, I. C., & Yip, A. W. C. (2004). Traumatic rectal perforation by an eel. Surgery, 135, 110–111. doi:10.1016/S00396060(03)00076X

`elephooty` *Elephant football data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
elephooty
```

Format

A tibble with 120 rows and 4 variables.

Details

Fictional data about elephant football. The highlight of the elephant calendar is the annual elephant soccer event in Nepal. A heated argument burns between the African and Asian elephants. In 2010, the president of the Asian Elephant Football Association, an elephant named Boji, claimed that Asian elephants were more talented than their African counterparts. The head of the African Elephant Soccer Association, an elephant called Tunc, issued a press statement that read 'I make it a matter of personal pride never to take seriously any remark made by something that looks like an enormous scrotum'. I was called in to settle things. I collected data from the two types of elephants (Asian or African) over a season and recorded how many goals each elephant scored and how many years of experience the elephant had. The data set has four variables:

- **id**: Elephant id
- **elephant**: Whether the elephant was an Asian elephant or an African elephant
- **experience**: how many years of football experience the elephant had
- **goals**: how many goals the elephant scored during the season

Source

www.discovr.rocks/csv/elephooty.csv

`escape` Escape from inside *data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
escape
```

Format

A tibble with 68 rows and 4 variables.

Details

In my teens I was in a band called Andromeda. I sang, we had a guitarist called Malcolm. We learnt several Queen and Iron Maiden songs and we were truly awful. Suffice it to say, you'd be hard pushed to recognize which Iron Maiden and Queen songs we were trying to play. It's common for bands to tire of cover versions and to get lofty ambitions to write their own tunes. I wrote one called 'Escape From Inside' about the film The Fly that contained the rhyming couplet of 'I am a fly, I want to die' – the great lyricists of the time quaked in their boots at the young new talent on the scene. The only thing we did that resembled the activities of a 'proper' band was to split up due to 'musical differences': Malcolm wanted to write 15-part symphonies about a boy's journey to

worship electricity pylons, whereas I wanted to write songs about flies and dying (preferably both). When we could not agree on a musical direction the split became inevitable. Had I had the power of statistics in my hands back then, rather than split up we could have tested empirically the best musical direction for the band. This study imagines such a world. A study was conducted to see whether I wrote better songs than my old bandmate Malcolm, and whether this depended on the type of song (a symphony or song about flies). The outcome variable was the number of screams elicited by audience members during the songs.

- **id**: Particiant id
- **song_type**: Whether participants listened to a symphony or a song about a fly
- **songwriter**: whether the song was written by Malcolm or Andy
- **screams**: how many screams of anguish participants expelled while listening to the song

Source

www.discovr.rocks/csv/escape.csv

`essay_marks` *Essay mark data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
essay_marks
```

Format

A tibble with 45 rows and 4 variables.

Details

Fictional data about essay marks. A student was interested in whether there was a positive relationship between the time spent doing an essay and the mark received. He got 45 of his friends and timed how long they spent writing an essay (hours) and the percentage they got in the essay (essay). He also translated these grades into their degree classifications (grade): in the UK, a student can get a first-class mark (the best), an upper-second-class mark, a lower second, a third, a pass or a fail (the worst). The data set has four variables

- **id**: Student id
- **essay**: Percentage mark on the essay
- **hours**: hours spend writing the essay
- **grade**: factor that converts the essay percentage to the degree classification of the essay (see general description)

Source

www.discovr.rocks/csv/essay_marks.csv

`exam_anxiety` *Exam anxiety data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
exam_anxiety
```

Format

A tibble with 103 rows and 5 variables.

Details

A psychologist was interested in the effects of exam stress on exam performance. She devised and validated a questionnaire to assess state anxiety relating to exams (called the Exam Anxiety Questionnaire, or EAQ). This scale produced a measure of anxiety scored out of 100. Anxiety was measured before an exam, and the percentage mark of each student on the exam was used to assess the exam performance. These data are fictional. The fictional data contains the following variables:

- **id**: participant id
- **revise**: the time spent revising for the exam (hours)
- **exam_grade**: the percentage score of each student on the exam
- **anxiety**: anxiety score on the EAQ out of 100
- **sex**: whether the participant self-identified as male or female

Source

www.discovr.rocks/csv/exam_anxiety.csv

`field_2006` *Field (2006) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
field_2006
```

Format

A tibble with 381 rows and 3 variables.

Details

Early in my career I looked at the effect of giving children information about entities. In one study (Field, 2006), I used three novel entities (the quoll, quokka and cuscus) and children were told threat information about one of the entities, positive information about another, and given no information about the third (our control). After the information I asked the children to place their hands in three wooden boxes each of which they believed contained one of the aforementioned entities The data from the study has three variables:

- **id**: The participant's id (these do not come from the study data file)
- **info_type**: the type of information given about the animal
- **latency**: the time taken for the child to approach the box (children who had not approached the box within 15s were assumed to be not consenting to that task and were scored as 15s)

Source

www.discovr.rocks/csv/gallup_2003.csv

References

Field, A. P. (2006). The behavioral inhibition system and the verbal information pathway to children's fears. *Journal of Abnormal Psychology*, *115*, 742–752. doi:10.1037/0021843x.115.4.742

`frontier_pal` *The Final Frontier palette*

Description

Colour palette based on Iron Maiden's The Final Frontier album sleeve.

Usage

```
frontier_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_frontier(n, type = "discrete", reverse = FALSE, ...)

scale_colour_frontier(n, type = "discrete", reverse = FALSE, ...)

scale_fill_frontier(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale`

`aesthetics` The names of the aesthetics that this scale works with.

`scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.

- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(frontier_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))
```

```
# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_frontier()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_frontier()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_frontier()
```

`gallup_2003` *Gallup et al. (2003) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
gallup_2003
```

Format

A tibble with 15 rows and 3 variables.

Details

It's something of a wonder how evolution managed to produce such a monstrosity as the human penis. One theory is sperm competition: the human penis has an unusually large glans (the 'bell-end') compared to other primates, and this may have evolved so that the penis can displace seminal fluid from other males by 'scooping it out' during intercourse. Armed with various devices from Hollywood Exotic Novelties, an artificial vagina from California Exotic Novelties, and some water and cornstarch Gallup et al. (2003) put this theory to the test. They loaded the artificial vagina with 2.6 ml of fake sperm and inserted one of three female sex toys into it before withdrawing it: a control phallus that had no coronal ridge (i.e., no bell-end), a phallus with a minimal coronal

ridge (small bell-end) and a phallus with a coronal ridge. They measured sperm displacement as a percentage: 100% means that all the sperm was displaced, and 0% means that none of the sperm was displaced. If the human penis evolved as a sperm displacement device then Gallup et al. predicted: (1) that having a bell-end would displace more sperm than not; and (2) that the phallus with the larger coronal ridge would displace more sperm than the phallus with the minimal coronal ridge. The data from the study has three variables:

- **id**: The participant's id (these do not come from the study data file)
- **phallus**: the type of phallus used (No coronal ridge, minimal coronal ridge and coronal ridge)
- **displace**: percentage of sperm displaced by the phallus

Source

www.discovr.rocks/csv/gallup_2003.csv

References

- Gallup, G. G. J., Burch, R. L., Zappieri, M. L., Parvez, R., Stockwell, M., & Davis, J. A. (2003). The human penis as a semen displacement device. *Evolution and Human Behavior*, *24*, 277–289. doi:10.1016/S10905138(03)000163

`gelman_2009` *Gelman & Weakliem (2009) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
gelman_2009
```

Format

A tibble with 548 rows and 3 variables.

Details

Apparently there are more beautiful women in the world than there are handsome men. Satoshi Kanazawa explains this finding in terms of good-looking parents being more likely to have a baby daughter as their first child than a baby son. Perhaps more controversially, he suggests that, from an evolutionarily perspective, beauty is a more valuable trait for women than for men (Kanazawa, 2007). In a playful and very informative paper, Andrew Gelman and David Weakliem discuss various statistical errors and misunderstandings, some of which have implications for Kanazawa's claims. The 'playful' part of the paper is that to illustrate their point they collected data on the 50 most beautiful celebrities (as listed by People magazine) of 1995-2000. They counted how many male and female children they had as of 2007. If Kanazawa is correct, these beautiful people would

have produced more girls than boys. These are the data from that study. The data contains the following variables:

- **person**: The name of the celebrity
- **child**: whether children are sons or daughters
- **number**: the number of sons/daughters (depending on the value of child) the celebrity has (at the time of the study)

Source

www.discovr.rocks/csv/gelman_2009.csv

References

- Gelman, A., & Weakliem, D. (2009). Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, 97, 310–316.

`glastonbury` *Glastonbury festival data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
glastonbury
```

Format

A tibble with 810 rows and 5 variables.

Details

More fictional data about people stinking at music festivals. The same biologist who was worried about the potential health effects of music festivals and collected data at a heavy metal festival (Download Festival), was worried that her findings might not generalize. To find out whether the type of music a person likes predicts whether hygiene decreases over the festival the biologist measured hygiene over the three days of the Glastonbury Music Festival, which has an eclectic clientele. Her hygiene measure ranged between 0 (you smell like you've bathed in sewage) and 4 (you smell like you've bathed in freshly baked bread). The biologist coded the festival-goer's musical affiliations into the categories 'hipster' (people who mainly like alternative music), 'metalhead' (people who like heavy metal), and 'raver' (people who like dance/ambient stuff). Anyone not falling into these categories was labelled 'no subculture'. The object contains the following variables:

- **ticket_no**: the ticket number of the participant as a factor

- **subculture**: The musical subculture with which the participant self-identifies as a factor (no subculture, hipster, metalhead, raver)
- **day1**: the hygiene score from 0 (eau de toilet) to 4 (eau de toilette) on day 1 of the festival
- **day2**: the hygiene score from 0 (eau de toilet) to 4 (eau de toilette) on day 2 of the festival
- **day3**: the hygiene score from 0 (eau de toilet) to 4 (eau de toilette) on day 3 of the festival
- **change**: the change in hygiene score from day 1 to day 3 of the festival

Source

www.discovr.rocks/csv/glastonbury.csv

`goggles` *Beer goggles effect data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
goggles
```

Format

A tibble with 48 rows and 4 variables.

Details

Fictional data about the beer goggles effect. An anthropologist was interested in the effects of facial attractiveness on the beer-goggles effect. She randomly selected 48 participants. Participants were randomly subdivided into three groups of 16: (1) a placebo group drank 500 ml of alcohol-free beer; (2) a low-dose group drank 500 ml of average strength beer (4% ABV); and (3) a high-dose group drank 500 ml of strong beer (7% ABV). Within each group, half (n = 8) rated the attractiveness of 50 photos of unattractive faces on a scale from 0 (pass me a paper bag) to 10 (pass me their phone number) and the remaining half rated 50 photos of attractive faces. The outcome for each participant was their median rating across the 50 photos. The data set has four variables

- **id**: Participant's id
- **facetype**: Whether the participant rated photos of 'attractive' or 'unattractive' faces
- **alcohol**: The alcohol group to which the participant was assigned. Either a placebo group (who drank 500 ml of alcohol-free beer), a low-dose group (who drank 500 ml of 4% ABV beer), or a high-dose group (who drank 500 ml of 7% ABV beer)
- **attractiveness**: the median rating of the attractiveness of 50 photos from 0 (pass me a paper bag) to 10 (pass me their phone number)

Source

www.discovr.rocks/csv/goggles.csv

`goggles_lighting` *Beer goggles and lighting data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
goggles_lighting
```

Format

A tibble with 208 rows and 4 variables.

Details

Fictional data about the moderating effect of lighting on the beer goggles effect. In previous edx-ample we came across the beer-goggles which suggests that alcohol impairs judgements of facial attractiveness. In this fictional follow-up study a sample of 26 people are given doses of alcohol (0 pints, 2 pints, 4 pints and 6 pints of lager) over four different weeks. They are asked to rate a bunch of photos of faces in either dim or bright lighting. The outcome measure was the mean attractiveness rating (out of 100) of the faces and the predictors were the dose of alcohol and the lighting conditions The data set has four variables

- **id**: Participant's id
- **lighting**: Whether the photos were viewed in dim or bright lighting
- **alcohol**: The dose of alcohol taken before ratings were made
- **rating**: the median rating of the attractiveness of the photos rated from 0 (pass me a paper bag) to 10 (pass me their phone number)

Source

www.discovr.rocks/csv/goggles.csv

grades *Grades data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
grades
```

Format

A tibble with 25 rows and 3 variables.

Details

Fictional data about stats grades. As a statistics lecturer I am interested in the factors that determine whether a student will do well on a statistics course. Imagine I took 25 students and looked at their grades for my statistics module at the end of their first year at university: first class, upper second class, lower second class, third class, pass and fail. I also asked these students what grade they got in their high school maths exams. In the UK GCSEs are school exams taken at age 16 that are graded A, B, C, D, E or F (an A grade is the best). The data set has three variables

- **id**: Student id
- **stats**: Degree classification for a statistics module
- **gcse**: GCSE mathematics classification at age 16

Source

www.discovr.rocks/csv/grades.csv

hangover *Hangover cure data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
hangover
```

Format

A tibble with 15 rows and 4 variables.

Details

A marketing manager tested the benefit of soft drinks for curing hangovers. He took 15 people and got them drunk. The next morning as they awoke, dehydrated and feeling as though they'd licked a camel's sandy feet clean with their tongue, he gave five of them water to drink, five of them Lucozade (a very nice glucose-based UK drink) and the remaining five a leading brand of cola. He measured how well they felt (on a scale from 0 = I feel like death to 10 = I feel really full of beans and healthy) two hours later. He measured how drunk the person got the night before on a scale of 0 = as sober as a nun to 10 = flapping about like a haddock out of water on the floor in a puddle of their own vomit. These data are fictional. The object contains the following variables:

- **id**: participant id
- **drink**: whether the person drank water, Lucozade or Cola as a hangover cure
- **well**: how well the person felt two hours after the hangover cure (0 = I feel like death to 10 = I feel really full of beans and healthy)
- **drunk**: how drunk the person got the night before (0 = as sober as a nun to 10 = flapping about like a haddock out of water on the floor in a puddle of their own vomit)

Source

www.discovr.rocks/csv/hangover.csv

`hiccups` *Hiccups data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
hiccups
```

Format

A tibble with 60 rows and 3 variables.

Details

People have many methods for stopping hiccups (a surprise, holding your breath), and medical science has put its collective mind to the task too. The official treatment methods include tongue-pulling manoeuvres, massage of the carotid artery, and, believe it or not, digital rectal massage (Fesmire, 1988). Let's say we wanted to put digital rectal massage to the test (erm, as a cure for hiccups). We took 15 hiccup sufferers, and during a bout of hiccups administered each of the three procedures (in random order and at intervals of 5 minutes) after taking a baseline of how many hiccups they had per minute. We counted the number of hiccups in the minute after each procedure. These data are fictional. The object contains the following variables:

- **id**: participant id
- **intervention**: the 4 interventions that each participant tried
- **hiccups**: the number of hiccups during the minute after the intervention

Source

www.discovr.rocks/csv/hiccups.csv

---

hill_2007 *Hill et al. (2007) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
hill_2007
```

Format

A tibble with 503 rows and 4 variables.

Details

Hill et al. (2007) examined whether providing children with a leaflet based on the *theory of planned behaviour* increased their exercise. There were four different interventions (intervention): a control group, a leaflet, a leaflet and quiz, and a leaflet and a plan. A total of 503 children from 22 different classrooms were sampled (classroom). The 22 classrooms were randomly assigned to the four different conditions. Children were asked *On average over the last three weeks, I have exercised energetically for at least 30 minutes ___ times per week* after the intervention (post_exercise). The data from the study has three variables:

- **intervention**: The intervention assigned to the classroom (control group, leaflet, leaflet and quiz, leaflet and plan).
- **classroom**: the classroom to which a child belonged

- **pre_exercise**: The exercise score pre-intervention (it's unclear to me from the paper how this was derived from the question asked!)
- **post_exercise**: The exercise score post-intervention (see above)

Source

www.discovr.rocks/csv/hill_2007.csv

References

- Hill, C., Abraham, C., & Wright, D. B. (2007). Can theory-based messages in combination with cognitive prompts promote exercise in classroom settings? *Social Science & Medicine*, 65, 1049–1058. doi:10.1016/j.socscimed.2007.04.024

---

honesty_lab *Honesty lab data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
honesty_lab
```

Format

A tibble with 100 rows and 3 variables.

Details

Fictional data about the honesty lab. Imagine we were interested in how people evaluated dishonest acts. Participants evaluate the dishonesty of acts based on watching videos of people confessing to those acts. Imagine we took 100 people and showed them a random dishonest act described by the perpetrator. They then evaluated the honesty of the act (from 0 = appalling behaviour to 10 = it's OK really) and how much they liked the person (0 = not at all, 10 = a lot). The data set has three variables

- **id**: Participant's id
- **deed**: evaluation of the honesty of the act (from 0 = appalling behaviour to 10 = it's OK really)
- **likeableness**: evaluation of the perpetrator (0 = not at all, 10 = a lot)

Source

www.discovr.rocks/csv/honesty_lab.csv

ice_bucket *Ice bucket challenge data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
ice_bucket
```

Format

A tibble with 23,230 rows and 1 variable.

Details

Google data relating to the ice bucket challenge from 2014. Golfer Chris Kennedy tipped a bucket of iced water on his head to raise awareness of the disease amyotrophic lateral sclerosis (ALS, also known as Lou Gehrig's disease). The idea is that you are challenged and have 24 hours to post a video of you having a bucket of iced water poured over your head in this video you also challenge at least three other people. If you fail to complete the challenge your forfeit is to donate to charity (in this case ALS). The CSV file contains the number of days after Chris Kennedy's initial ice bucket challenge that each of 2,323,452 ice bucket challenge video was uploaded to YouTube. The data here contains a randomly selected 1% of the original data (23,230 cases).

- **upload_days**: the number of days after Chris Kennedy's initial ice bucket challenge that an ice bucket challenge video was uploaded to YouTube

Source

www.discovr.rocks/csv/ice_bucket.csv

im_pal *Iron Maiden palette*

Description

Colour palette based on Iron Maiden's eponymous album sleeve.

Usage

```
im_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_im(n, type = "discrete", reverse = FALSE, ...)

scale_colour_im(n, type = "discrete", reverse = FALSE, ...)

scale_fill_im(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale

aesthetics The names of the aesthetics that this scale works with.

scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

palette A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., scales::pal_hue()).

name The name of the scale. Used as the axis or legend title. If waiver(), the default, the name of the scale is taken from the first mapping used for that aesthetic. If NULL, the legend title will be omitted.

breaks One of:

- NULL for no breaks
- waiver() for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

labels One of:

- NULL for no labels
- waiver() for the default labels computed by the transformation object
- A character vector giving labels (must be same length as breaks)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

limits One of:

- NULL to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(im_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_im()

# Plot some data and apply theme to colour (note UK English)
```

```
ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_im()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_im()
```

`invisibility_base` *Cloak of invisibility data (pre-post design)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
invisibility_base
```

Format

A tibble with 80 rows and 4 variables.

Details

In invisibility_cloak we compared the number of mischievous acts in people who had invisibility cloaks to those without. Imagine we replicated that study, but changed the design so that we recorded the number of mischievous acts in these participants before the study began as well as during the study. The data contains the following variables:

- **id**: participant id
- **cloak**: whether the participant was assigned a cloak of invisibility
- **mischief_pre**: the number of mischievous acts committed during the week before the study
- **mischief**: the number of mischievous acts committed during the week of the study

Source

www.discovr.rocks/csv/invisibility_base.csv

invisibility_cloak *Cloak of invisibility data (independent design)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
invisibility_cloak
```

Format

A tibble with 24 rows and 3 variables.

Details

I got very excited by two news stories implying that scientists had made Harry Potter's cloak of invisibility. Although the newspapers overstated the case, I imagined a future in which we have cloaks of invisibility to test out. Given my slightly mischievous streak, the future me is interested in the effect that wearing a cloak of invisibility has on the tendency for mischief. I take 24 participants and place them in an enclosed community. The community is riddled with hidden cameras so that we can record mischievous acts. Half of the participants are given cloaks of invisibility; they are told not to tell anyone else about their cloak and that they can wear it whenever they liked. I measure how many mischievous acts they performed in one week. The object contains the following variables:

- **id**: participant id
- **cloak**: whether the participant was assigned a cloak of invisibility
- **mischief**: the number of mischievous acts committed during a week

Source

www.discovr.rocks/csv/invisibility.csv

invisibility_rm *Cloak of invisibility data (repeated measures design)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
invisibility_rm
```

Format

A tibble with 24 rows and 3 variables.

Details

I got very excited by two news stories implying that scientists had made Harry Potter's cloak of invisibility. Although the newspapers overstated the case, I imagined a future in which we have cloaks of invisibility to test out. Given my slightly mischievous streak, the future me is interested in the effect that wearing a cloak of invisibility has on the tendency for mischief. I take 12 participants and place them in an enclosed community. The community is riddled with hidden cameras so that we can record mischievous acts. For one week the participants are given cloaks of invisibility, during a different week they are not. I measure how many mischievous acts they performed in each week. These data are the same as in invisibility_cloak but arranged in a repeated measures design. The object contains the following variables:

- **id**: participant id
- **cloak**: whether the participant had access to a cloak of invisibility
- **mischief**: the number of mischievous acts committed during a week

Source

www.discovr.rocks/csv/invisibility_rm.csv

---

jiminy_cricket *Jiminy Cricket data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
jiminy_cricket
```

Format

A tibble with 500 rows and 4 variables.

Details

Fictitious data inspired by my honeymoon at Disney in Orlando. The one blip in my tolerance of Disney, was their obsession with dreams coming true and wishing upon a star. Dreams are good, but a completely blinkered view that they'll come true without any work on your part is not. I think it highly unlikely that merely 'wishing upon a star' will make my dream come true. I wonder if the seismic increase in youth internalizing disorders (Twenge, 2000, 2011) is, in part, caused by millions of Disney children reaching the rather depressing realization that 'wishing upon a star'

didn't work. Anyway, imagine that I collected some data from 250 people on their level of success using a composite measure involving their salary, quality of life and how closely their life matches their aspirations. This gave me a score from 0 (complete failure) to 100 (complete success). I then implemented an intervention: I told people that for the next 5 years they should either wish upon a star for their dreams to come true or work as hard as they could to make their dreams come true. I measured their success again 5 years later. People were randomly allocated to these two instructions. The data contains the following variables:

- **id**: participant id
- **strategy**: whether the person was allocated to the 'hard work' or 'wish upon a star' intervention
- **time**: whether the measure of success was taken before the intervention (pre-intervention) aor after it (post-intervention)
- **success**: the person's success from 0 (complete failure) to 100 (complete success) using my dodgy composite measure.

Source

www.discovr.rocks/csv/jiminy_cricket.csv

`johns_2012` *Johns et al. (2012) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
johns_2012
```

Format

A tibble with 160 rows and 4 variables.

Details

It is believed that males have a biological predispoition towards the colour red because it is sexually salient. The theory suggests that women use the colour red as a proxy signal for genital colour to indicate ovulation and sexual proceptivity. If this hypothesis is true then using the colour red in this way would have to attract men (otherwise it's a pointless strategy). In a novel study, Johns, Hargrave, and Newton-Fisher (2012) tested this idea by manipulating the colour of four pictures of female geneitalia to make them increasing shades of red (pale pink, light pink, dark pink, red). Heterosexual males rated the resulting 16 pictures from 0 (unattractive) to 100 (attractive). These are the data from that study. The data contains the following variables:

- **id**: participant id

- **partners**: sexual experience coded as a factor ('Very little' and 'Some")
- **colour**: colour of the female geneitalia in image
- **attractiveness**: male rating of the attractiveness of the female geneitalia from 0 to 100

Source

www.discovr.rocks/csv/johns_2012.csv

References

- Johns, S. E., Hargrave, L. A., & Newton-Fisher, N. E. (2012). Red is not a proxy signal for female genitalia in humans. *PLoS One*, 7, e34669. doi:10.1371/journal.pone.0034669

`killers_pal` *Killers palette*

Description

Colour palette based on Iron Maiden's killers album sleeve.

Usage

```
killers_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_killers(n, type = "discrete", reverse = FALSE, ...)

scale_colour_killers(n, type = "discrete", reverse = FALSE, ...)

scale_fill_killers(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale`
	`aesthetics` The names of the aesthetics that this scale works with.
	`scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.
	`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).
	`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(killers_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_killers()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_killers()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_killers()
```

---

lambert_2012 *Lambert et al. (2012) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
lambert_2012
```

Format

A tibble with 240 rows and 6 variables

Details

Lambert et al. (2012) found that pornography is related to infidelity. This object contains the data from that study.

- **id**: participant ID (not from the original data)
- **consumption**: pornography consumption on a scale from 0 (low) to 8 (high)
- **ln_porn**: log transformed values of consumption
- **commit**: commitment to the participant's current relationship on a scale from 1 (low) to 5 (high)
- **phys_inf**: whether the person had committed a physical act that they or their partner would consider to be unfaithful (0 = no, 1 = one of them would consider it unfaithful, 2 = both of them would consider it unfaithful)
- **hook_up**: the number of people they had 'hooked up' with in the previous year. (A 'hook-up' was defined to participants as 'when two people get together for a physical encounter and don't necessarily expect anything further')

Source

www.discovr.rocks/csv/lambert_2012.csv

References

- Lambert, N. M., Negash, S., Stillman, T. F., Olmstead, S. B., & Fincham, F. D. (2012). A love that doesn't last: Pornography consumption and weakened commitment to one's romantic partner. *Journal of Social and Clinical Psychology*, 31, 410–438. doi:10.1521/jscp.2012.31.4.410

`massar_2012` *Massar et al. (2012) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
massar_2012
```

Format

A tibble with 83 rows and 4 variables

Details

Everyone likes a good gossip from time to time, but apparently it has an evolutionary function. One school of thought is that gossip is used as a way to derogate sexual competitors – especially by questioning their appearance and sexual behaviour. Apparently men rate gossiped-about women as less attractive, and they are more influenced by the gossip if it came from a woman with a high mate value (i.e. attractive and sexually desirable). Karlijn Massar and her colleagues hypothesized that if this theory is true then (1) younger women will gossip more because there is more mate competation at younger ages; and (2) this relationship will be mediated by the mate value of the person (because for those with high mate value gossiping for the purpose of sexual competition will be more effective). These are the data from that study.

Eighty-three women aged from 20 to 50 (age) completed questionnaire measures of their tendency to gossip (gossip) and their sexual desirability (mate_value). Lambert et al. (2012) found that pornography is related to infidelity. This object contains the data from that study.

- **id**: participant ID (not from the original data)
- **age**: participant age in years
- **gossip**: average response on a tendency to gossip scale. Participants responded to 16 items about their tendency to gossip following the presentation of a scenario. Particpants rated their likelihood to engage in certain behaviours such as 'I would tell negative things about Karen to other people' from 1 (strongly disagree) to 5 (strongly agree). This score is the average response across the 16 items.
- **mate_value**: average response to items from the Self-Perceived Mating Success Scale (each item ranged from 1 = not at all, 5 = very much, so a high score is a high mate value)

Source

www.discovr.rocks/csv/massar_2012.csv

References

- Massar, K., Buunk, A. P., & Rempt, S. (2012). Age differences in women's tendency to gossip are mediated by their mate value. *Personality and Individual Differences*, 52, 106–109. doi:10.1016/j.paid.2011.09.013

`mcnulty_2008` *McNulty et al. (2008) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
mcnulty_2008
```

Format

A tibble with 164 rows and 5 variables

Details

McNulty et al. (2008) found a relationship between a person's attractiveness and how much support they give their partner among newlywed heterosexual couples. These data simulate the results of that study. The object contains the following variables:

- **id**: participant ID
- **attractiveness**: attractiveness of participant
- **support**: support given to partner
- **satisfaction**: relationship satisfaction
- **spouse**: whether the participant is a husband or wife

Source

www.discovr.rocks/csv/mcnulty_2008.csv

References

- McNulty, J. K., Neff, L. A., & Karney, B. R. (2008). Beyond initial attraction: Physical attractiveness in newlywed marriage. *Journal of Family Psychology*, 22, 135–143. doi:10.1037/08933200.22.1.135

---

men_dogs *Are men like dogs data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
men_dogs
```

Format

A tibble with 40 rows and 3 variables.

Details

A psychologist was interested in the cross-species differences between men and dogs. She observed a group of dogs and a group of men in a naturalistic setting (20 of each). She classified several behaviours as being dog-like (urinating against trees and lampposts, attempts to copulate with anything that moved, and attempts to lick their own genitals). For each man and dog she counted the number of dog-like behaviours displayed in a 24-hour period. The (fictional) data contains the following variables:

- **id**: the participant's id
- **species**: whether the participant was a man or a dog
- **behaviour**: number of dog-like behaviours exhibited by the participant in 24 hours

Source

www.discovr.rocks/csv/men_dogs.csv

---

metal *Metal music and anger*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
metal
```

Format

A tibble with 90 rows and 4 variables.

Details

People have claimed that listening to heavy metal, because of its aggressive sonic palette and often violent or emotionally negative lyrics, leads to angry and aggressive behaviour. As a very non-violent metal fan this accusation bugs me (BTW there are some real data on this in sharman_2015). Imagine I designed a study to test this possibility. I took groups of self-classifying metalheads and non-metalheads (fan) and assigned them randomly to listen to 15 minutes of either the sound of an angle grinder scraping a sheet of metal (control noise), metal music, or pop music (soundtrack). Each person rated their anger on a scale ranging from 0 (*All you need is love, da, da, da-da-da*) to 100 (—- *me, I'm all out of enemies*). These data are fictitious.

- **id**: the participant's ID
- **soundtrack**: whether the participant listened to 15 minutes of an angle grinder, metal music or pop music.
- **fan**: whether the participant self-classified as a metal fan (*metalhead*) or not.
- **anger**: self-reported anger after listening to the 15 minutes of sound from 0 (Maria Taylor) to 100 (Corey Taylor)

Source

www.discovr.rocks/csv/metal.csv

`metallica` *Metallica data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
metallica
```

Format

A tibble with 7 rows and 9 variables.

Details

The data show various pieces of information about past and present members of the band Metallica that may or may not be accurate at the time of writing (2019). The data contains the following variables:

- **name**: the band member's name
- **birth_date**: the band member's date of birth
- **death_date**: the band member's date of death (where applicable)
- **instrument**: the instrument played by the band member
- **current_member**: is the member currently in the band? (True or False)
- **songs_written**: the number of songs the band member has contributed to
- **net_worth**: the band member's net worth as of 2019 according to some dodgy website
- **albums**: the number of studio albums each member played on (up to 2020)
- **worth_per_song**: the members net worth per song contributed to

Source

www.discovr.rocks/csv/metallica.csv

`metal_health` *Metal health*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
metal_health
```

Format

A tibble with 2506 rows and 2 variables.

Details

Lacourse et al. (2001) conducted a study to see whether suicide risk was related to listening to heavy metal music. They devised a scale to measure preference for bands falling into the category of heavy metal. This scale included heavy metal bands (Black Sabbath, Iron Maiden), speed metal bands (Slayer, Metallica), death/black metal bands (Obituary, Burzum) and gothic bands (Marilyn Manson, Sisters of Mercy). They then used this (and other variables) as predictors of suicide risk based on a scale measuring suicidal ideation etc. These data are from a fictitious replication. There are two variables representing scores on the scales described above:

- **hm**: the extent to which the person listens to heavy metal music
- **suicide**: the extent to which someone has suicidal ideation

Source

www.discovr.rocks/csv/metal_health.csv

References

- Lacourse, E., Claes, M., & Villeneuve, M. (2001). Heavy metal music and adolescent suicidal risk. *Journal of Youth and Adolescence*, *30*, 321–332. doi:10.1023/A:1010492128537

`miller_2007` *Miller et al. (2007) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
miller_2007
```

Format

A tibble with 296 rows and 4 variables.

Details

Miller and colleagues (2007) tested the *hidden-estrus* theory, which suggests that unlike other female mammals, humans do not experience an *estrus* phase during which they are more sexually receptive, proceptive, selective and attractive. If this theory is wrong then human men should find women most attractive during the fertile phase of their menstrual cycle compared to the pre-fertile (menstrual) and post-fertile (luteal) phase. Miller used the tips obtained by dancers at a lap dancing club as a proxy for their sexual attractiveness and also recorded the phase of the dancer's menstrual cycle during a given shift, and whether they were using hormonal contraceptives. Dancers provided data from between 9 to 29 of their shifts.

- **id**: Dancer's ID.
- **contraceptive**: whether the dancer was currently using oral hormonal contraceptives.
- **cyclephase**: the phase of the dancer's menstrual cycle at the time of a particular shift.
- **tips**: The tips (in US dollars) received during a particular shift

Source

www.discovr.rocks/csv/miller_2007.csv

References

- Miller, G., Tybur, J. M., & Jordan, B. D. (2007). Ovulatory cycle effects on tip earnings by lap dancers: Economic evidence for human estrus? *Evolution and Human Behavior*, 28, 375–381. doi:10.1016/j.evolhumbehav.2007.06.002

mixed_attitude *Imagery and advertising example*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
mixed_attitude
```

Format

A tibble with 180 rows and 5 variables

Details

A marketing researcher was interested in the effects of types of imagery (positive, negative or neutral) on perceptions of different types of drink (beer, wine, water). Participants viewed videos of different drink products in the context of positive, negative or neutral imagery and then rated the products on a scale from –100 (extremely dislike) through 0 (neutral) to 100 (extremely like). Those who identify as men and women might respond differently to the products, so participants self-reported their gender (a between-group variable). The (fictional) data contains the following variables:

- **id**: participant ID
- **gender**: gender identity (self-identify as male or female)
- **drink**: The drink use din the advert (beer, wine or water)
- **imagery**: The valence of the imagery used in the advert (positive, negative, neutral)

Source

www.discovr.rocks/csv/speed_date.csv

murder *Murder in the streets data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
murder
```

Format

A tibble with 36 rows and 3 variables.

Details

Fictitious data about murder. A sociologist wanted to compare murder rates (murder) each month in a year at three high-profile locations in London (street). The data contains the following variables:

- **month**: The month for the reported crime statistics
- **street**: The street location (Ruskin Avenue, Acacia Avenue and Rue Morgue)
- **murder**: the number of reported murders during each month

Source

www.discovr.rocks/csv/murder.csv

---

muris_2008 *Muris et al. (2008) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
muris_2008
```

Format

A tibble with 70 rows and 6 variables.

Details

Anxious people tend to interpret ambiguous information in a negative way. For example, being highly anxious myself, if I overheard a student saying "Andy Field's lectures are really different" I would assume that *different* meant rubbish, but it could also mean 'refreshing' or 'innovative'. Muris, Huijding, Mayer, and Hameetman (2008) addressed how these interpretational biases develop in children. Children imagined that they were astronauts who had discovered a new planet. They were given scenarios about their time on the planet (e.g., *On the street, you encounter a spaceman. He has a toy handgun and he fires at you* ...) and the child had to decide whether a positive (*You laugh: it is a water pistol and the weather is fine anyway*) or negative (*Oops, this hurts! The pistol produces a red beam which burns your skin!*) outcome occurred. After each response the child was told whether their choice was correct. Half of the children were always told that the negative interpretation was correct, and the reminder were told that the positive interpretation was correct.

Over 30 scenarios children were trained to interpret their experiences on the planet as negative or positive. Muris et al. then measured interpretational biases in everyday life to see whether the training had created a bias to interpret things negatively. In doing so, they could ascertain whether children might learn interpretational biases through feedback (e.g., from parents).The data contains the following variables:

- **participant**: a number identifying he participant
- **age**: participant's age in years
- **gender**: self-reported gender of the participant
- **scared**: score on The Screen for Child Anxiety Related Disorders (SCARED)
- **training**: whether the child was assigned to positive interpretation training or negative interpretation training.
- **int_bias**: interpretation bias for everyday events

Source

www.discovr.rocks/csv/muris_2008.csv

References

- Muris, P., Huijding, J., Mayer, B., & Hameetman, M. (2008). A space odyssey: Experimental manipulation of threat perception and anxiety-related interpretation bias in children. *Child Psychiatry and Human Development*, 39, 469–480. doi:10.1007/s105780080103z

---

nichols_2004 *Internet addiction scale (IAS) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
nichols_2004
```

Format

A tibble with 207 rows and 38 variables.

Details

The increasing populatrity (and usefulness) of the Internet has led to the serious problem of internet addiction. To research this construct it's helpful to be able to measure it, so Laura Nichols and Richard Nicki developed the Internet Addiction Scale, IAS (Nichols & Nicki, 2004). This 36-item questionnaire contains items such as *I have stayed on the Internet longer than I intended to* and *My grades/work have suffered because of my Internet use* to which responses are made on a five-point scale (*never*, *rarely*, *sometimes*, *frequently*, *always*). The authors dropped two items because they had low means and variances, and dropped three others because of relatively low correlations with other items. They performed a principal component analysis on the remaining 31 items (N = 207).

- **participant_code**: The participant id
- **gender**: The participant biological sex
- **ias1**: responses (1-5) to the question *I find that I need to use the Internet more to get the same enjoyment as before.*
- **ias2**: responses (1-5) to the question *When I use the Internet now, I do not feel as good as I used to.*
- **ias3**: responses (1-5) to the question *Time spent on the Internet now is not as enjoyable as it was when I first started using the Internet.*
- **ias4**: responses (1-5) to the question *Since I first began using the Internet I would say that the amount of time I spend on line has increased but not the satisfaction.*
- **ias5**: responses (1-5) to the question *I feel depressed, moody or nervous when I am off the internet which goes away when I log on.*
- **ias6**: responses (1-5) to the question *I feel distressed when I am unable to spend as much time on the Internet as I usually do.*
- **ias7**: responses (1-5) to the question *The more time I spend away from the Internet, the more irritable I feel.*
- **ias8**: responses (1-5) to the question *When I attempt to cut back of stop using the Internet I find that the irritability that I experience is relieved by going back on the Internet*
- **ias9**: responses (1-5) to the question *I have stayed on the Internet longer than I intended to.*
- **ias10**: responses (1-5) to the question *I have said to myself 'just a few more minutes on the Internet.'*
- **ias11**: responses (1-5) to the question *I find myself accessing more information on the Internet that I had planned to.*
- **ias12**: responses (1-5) to the question *I find myself doing more things on the Internet than I had intended to*
- **ias13**: responses (1-5) to the question *I have felt a persistent desire to cut down or control my use of the Internet.*
- **ias14**: responses (1-5) to the question *I have attempted to spend less time on the Internet but I have been unable to do so.*
- **ias15**: responses (1-5) to the question *I have tried unsuccessfully to restrict my Internet use because of previous over use.*
- **ias16**: responses (1-5) to the question *I would like to spend less time on the Internet.*

- **ias17**: responses (1-5) to the question *I have walked or driven to campus/work specifically to use the Internet at times when I normally would not go to campus/work*
- **ias18**: responses (1-5) to the question *After being on the Internet late into the night in sleep late the next morning because of my Internet use.*
- **ias19**: responses (1-5) to the question *Once I am on the Internet, I seem to stay on for a long time.*
- **ias20**: responses (1-5) to the question *I am on the Internet so much that I have to make up for the lost time.*
- **ias21**: responses (1-5) to the question *I have missed class/work so that I would have more time to spend on the Internet.*
- **ias22**: responses (1-5) to the question *I have neglected things, which are important and need doing.*
- **ias23**: responses (1-5) to the question *I see my friends less often because of the time that I spend on the Internet.*
- **ias24**: responses (1-5) to the question *I have given up a particular recreational activity in order that I would have more time on the Internet*
- **ias25**: responses (1-5) to the question *At times I have tried to conceal how long I have been on the Internet*
- **ias26**: responses (1-5) to the question *My grades/work have suffered because of my Internet use.*
- **ias27**: responses (1-5) to the question *I have lost sleep because of my Internet use*
- **ias28**: responses (1-5) to the question *The Internet has affected my life in a negative way.*
- **ias29**: responses (1-5) to the question *The people I know through the Internet know me better than my friends at university*
- **ias30**: responses (1-5) to the question *I prefer socializing on the Internet rather than in person with my friends and family*
- **ias31**: responses (1-5) to the question *I feel that life without the Internet would be boring and empty.*
- **ias32**: responses (1-5) to the question *I find myself thinking/longing about when I will go on the Internet again.*
- **ias33**: responses (1-5) to the question *When I feel lonely, I use the Internet to talk to others.*
- **ias34**: responses (1-5) to the question *When I use the Internet, I experience a buzz or a high (i.e., feeling elated).*
- **ias35**: responses (1-5) to the question *I use the Internet as a way of escaping the real world.*
- **ias36**: responses (1-5) to the question *I use the Internet as a way of escaping the "real world."*

Source

www.discovr.rocks/csv/nichols_2004.csv

References

- Nichols, L. A., & Nicki, R. (2004). Development of a psychometrically sound internet addiction scale: A preliminary step. *Psychology of Addictive Behaviors*, 18, 381–384. doi:10.1037/0893164X.18.4.381

nob_pal *The Number of the Beast palette*

Description

Colour palette based on Iron Maiden's The Number of the Beast album sleeve.

Usage

```
nob_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_nob(n, type = "discrete", reverse = FALSE, ...)

scale_colour_nob(n, type = "discrete", reverse = FALSE, ...)

scale_fill_nob(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale aesthetics The names of the aesthetics that this scale works with. scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale. palette A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., scales::pal_hue()). name The name of the scale. Used as the axis or legend title. If waiver(), the default, the name of the scale is taken from the first mapping used for that aesthetic. If NULL, the legend title will be omitted. breaks One of: • NULL for no breaks • waiver() for the default breaks (the scale limits) • A character vector of breaks • A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation. labels One of: • NULL for no labels • waiver() for the default labels computed by the transformation object • A character vector giving labels (must be same length as breaks) • An expression vector (must be the same length as breaks). See ?plotmath for details.

- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(nob_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))
```

```
# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_nob()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_nob()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_nob()
```

`notebook` *The notebook data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
notebook
```

Format

A tibble with 40 rows and 3 variables.

Details

Fictitious data about the film The Notebook. Imagine that a film company director was interested in whether there was really such a thing as a 'chick flick' (a film that has the stereotype of appealing to women more than to men). He took 20 people who mostly self identify as men and 20 who mostly self identify as women and showed half of each sample a film that was supposed to be a 'chick flick' (The Notebook). The other half watched a documentary about notebooks as a control. In all cases the company director measured participants' arousal as an indicator of how much they enjoyed the film. The data contains the following variables:

- **id**: participant ID

- **gender_identity**: gender with which the participant mostly self-identifies
- **film**: whether the person watched The Notebook or a documentary about notebooks
- **arousal**: the person's average physiological arousal (e.g., emotional response) during the film.

Source

www.discovr.rocks/csv/notebook.csv

---

ocd *OCD data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
ocd
```

Format

A tibble with 30 rows and 4 variables.

Details

Fictitious data about interventions for obsessive compulsive disorder. Obsessive compulsive disorder (OCD) is a mental health problem characterized by intrusive images or thoughts that the sufferer finds abhorrent. These thoughts lead the sufferer to engage in activities to neutralize the unpleasantness of these thoughts (these activities can be mental or physical). A group of clinical psychologists were interested in the efficacy of two different interventions for OCD offered at their clinic: cognitive behaviour therapy (CBT) and behaviour therapy (BT). A group who were awaiting treatment acted as a control (a no treatment condition, NT). To gauge the success of therapy, the clinical psychologists measured two outcomes: the occurrence of obsession-related behaviours (actions) and the occurrence of obsession-related cognitions (thoughts) on a single day. Service users were randomly assigned to group 1 (CBT), group 2 (BT) or group 3 (NT). The data contains the following variables:

- **id**: participant ID
- **group**: the group to which service users were assigned (BT, CBT or NT)
- **thoughts**: the number of Number of obsession-related thoughts
- **actions**: the number of Number of obsession-related behaviours

Source

www.discovr.rocks/csv/ocd.csv

okabe_ito_pal *Colourblind-friendly palette*

Description

Colour palette based on Color Universal Design by Okabe and Ito https://jfly.uni-koeln.de/color/.

Usage

```
okabe_ito_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_oi(n, type = "discrete", reverse = FALSE, ...)

scale_colour_oi(n, type = "discrete", reverse = FALSE, ...)

scale_fill_oi(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale aesthetics The names of the aesthetics that this scale works with. scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale. palette A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., scales::pal_hue()). name The name of the scale. Used as the axis or legend title. If waiver(), the default, the name of the scale is taken from the first mapping used for that aesthetic. If NULL, the legend title will be omitted. breaks One of: • NULL for no breaks • waiver() for the default breaks (the scale limits) • A character vector of breaks • A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation. labels One of: • NULL for no labels • waiver() for the default labels computed by the transformation object • A character vector giving labels (must be same length as breaks)

- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(okabe_ito_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.
```

```
classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_oi()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_oi()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_oi()
```

---

ong_2011 *Ong et al. (2011) data: wide/messy format*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
ong_2011
```

Format

A tibble with 275 rows and 12 variables.

Details

A study by Ong et al., (2011) examining the relationship between a person's narcissism and other people's ratings of their profile picture on Facebook. The pictures were rated on each of four dimensions: coolness, glamour, fashionableness, and attractiveness. In addition, each person was measures on introversion/extroversion and narcissism. These data are in messy/wide format. The data contains the following variables:

- **id**: a number identifying he participant

- **grade**: participants grade at school (Sec 1, Sec 2 or Sec 3)
- **age**: participant's age in years
- **sex**: biological sex of the participant
- **status**: frequency of changing ones Facebook status per week
- **attractiveness**: rating of profile picture along the dimension of physical attractiveness (1 = not attractive, 5 = very attractive)
- **fashionable**: rating of profile picture along the dimension of fashionable of profile picture (1 = not fashionable, 5 = very fashionable)
- **glamour**: rating of profile picture along the dimension of glamour (1 = not glamourous, 5 = very glamourous)
- **cool**: rating of profile picture along the dimension of cool (1 = not cool, 5 = very cool)
- **profile**: sum of profile picture ratings
- **extraversion**: score on the NEO Five-Factor Inventory (NEO-FFI) extraversion scale
- **narcissism**: score on the Narcissistic Personality Questionnaire for Children-Revised (NPQC-R)

Source

www.discovr.rocks/csv/ong_2011.csv

References

- Ong, E. Y. L., Ang, R. P., Ho, J. C. M., Lim, J. C. Y., Goh, D. H., Lee, C. S., & Chua, A. Y. K. (2011). Narcissism, extraversion and adolescents' self-presentation on Facebook. *Personality and Individual Differences*, 50, 180–185. doi:10.1016/j.paid.2010.09.022

`ong_tidy` *Ong et al. (2011) data: tidy format*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
ong_tidy
```

Format

A tibble with 1100 rows and 9 variables.

Details

A study by Ong et al., (2011) examining the relationship between a person's narcissism and other people's ratings of their profile picture on Facebook. The pictures were rated on each of four dimensions: coolness, glamour, fashionableness, and attractiveness. In addition, each person was measures on introversion/extroversion and narcissism. These data are in tidy format. The data contains the following variables:

- **id**: a number identifying he participant
- **age**: participant's age in years
- **sex**: biological sex of the participant
- **status**: frequency of changing ones Facebook status per week
- **profile**: sum of profile picture ratings
- **extraversion**: score on the NEO Five-Factor Inventory (NEO-FFI) extraversion scale
- **narcissism**: score on the Narcissistic Personality Questionnaire for Children-Revised (NPQC-R)
- **rating_type**: the dimension along which profile pictures were rated (Attractiveness, Fashionable, Cool, Glamour)
- **rating**: rating of the profile picture from 1 (not attractive/cool/fashionable/glamourous) to 5 (very attractive/cool/fashionable/glamourous)

Source

www.discovr.rocks/csv/ong_2011_tidy.csv

References

- Ong, E. Y. L., Ang, R. P., Ho, J. C. M., Lim, J. C. Y., Goh, D. H., Lee, C. S., & Chua, A. Y. K. (2011). Narcissism, extraversion and adolescents' self-presentation on Facebook. *Personality and Individual Differences*, 50, 180–185. doi:10.1016/j.paid.2010.09.022

`penalty` *Penalty kicks data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
penalty
```

Format

A tibble with 75 rows and 5 variables.

Details

Fictional data set looking at predictors of success of penalty takers in soccer (or whatever sport you enjoy). The outcome variable is whether a penalty is scored or missed. Based on (imaginary) past research there are two factors that reliably predict whether a penalty kick will be missed or scored: (1) the extent to which the penalty taker is prone to worry (measured using the Penn State Worry Questionnaire, PSWQ); and (2) the past success rate of the penalty taker. State anxiety is also likely detrimental effects on performance so it was also measured. The data contain the following variables:

- **id**: Penalty taker's id
- **pswq**: proneness to worry on the Penn State Worry Questionnaire, PSWQ
- **anxious**: state anxiety
- **previous**: The percentage of previous penalties scored (to the nearest percent)

Source

www.discovr.rocks/csv/penalty.csv

---

pom_pal *Piece of Mind palette*

Description

Colour palette based on Iron Maiden's Piece of Mind album sleeve.

Usage

```
pom_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_pom(n, type = "discrete", reverse = FALSE, ...)

scale_colour_pom(n, type = "discrete", reverse = FALSE, ...)

scale_fill_pom(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale aesthetics The names of the aesthetics that this scale works with. scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for `y` axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(pom_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_pom()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_pom()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_pom()
```

---

power_pal *Powerslave palette*

Description

Colour palette based on Iron Maiden's Powerslave album sleeve.

Usage

```
power_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_power(n, type = "discrete", reverse = FALSE, ...)

scale_colour_power(n, type = "discrete", reverse = FALSE, ...)

scale_fill_power(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale

aesthetics The names of the aesthetics that this scale works with.

scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

palette A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., scales::pal_hue()).

name The name of the scale. Used as the axis or legend title. If waiver(), the default, the name of the scale is taken from the first mapping used for that aesthetic. If NULL, the legend title will be omitted.

breaks One of:

- NULL for no breaks
- waiver() for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

labels One of:

- NULL for no labels
- waiver() for the default labels computed by the transformation object
- A character vector giving labels (must be same length as breaks)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

limits One of:

- NULL to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(power_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_power()

# Plot some data and apply theme to colour (note UK English)
```

```
ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_power()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_power()
```

`prayer_pal` *No Prayer for the Dying palette*

Description

Colour palette based on Iron Maiden's No Prayer for the Dying album sleeve.

Usage

```
prayer_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_prayer(n, type = "discrete", reverse = FALSE, ...)

scale_colour_prayer(n, type = "discrete", reverse = FALSE, ...)

scale_fill_prayer(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale` `aesthetics` The names of the aesthetics that this scale works with. `scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale. `palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`). `name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted. `breaks` One of:

- NULL for no breaks
- waiver() for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

labels One of:

- NULL for no labels
- waiver() for the default labels computed by the transformation object
- A character vector giving labels (must be same length as breaks)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

limits One of:

- NULL to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

expand For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function expansion() to generate the values for the expand argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

na.translate Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify na.translate = FALSE.

na.value If na.translate = TRUE, what aesthetic value should the missing values be displayed as? Does not apply to position scales where NA is always placed at the far right.

drop Should unused factor levels be omitted from the scale? The default, TRUE, uses the levels that appear in the data; FALSE includes the levels in the factor. Please note that to display every level in a legend, the layer should use show.legend = TRUE.

guide A function used to create a guide or its name. See guides() for more information.

position For position scales, The position of the axis. left or right for y axes, top or bottom for x axes.

call The call used to construct the scale for reporting messages.

super The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(prayer_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_prayer()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_prayer()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_prayer()
```

---

profile_pic *Profile picture data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
profile_pic
```

Format

A tibble with 80 rows and 4 variables.

Details

A researcher was interested in the effect of profile pictures on social media on unsolicited attention. She took 40 people who had profiles on a social networking website; 17 of them had a relationship status of 'single' and the remaining 23 had their status as 'in a relationship'. We asked these people to set their profile picture to a photo of them on their own (alone) and to count how many friend request they got from random strangers over 3 weeks, then to switch it to a photo of them very obviously as part of a romantic couple and record their friend requests from random strangers over 3 weeks. The (fictional) data contains the following variables:

- **id**: Participant id
- **rel_status**: Whether the participant's relationship status is 'single' or 'in a relationship'
- **profile_pic**: Whether the participant's profile picture depicts them alone or as part of a couple
- **requests**: The number of unsolicited friend requests (in 3 weeks) from random strangers who categorise their sexual orientation such that they are interested in people of the gender of the participant

Source

www.discovr.rocks/csv/profile_pic.csv

---

pubs *Pub data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
pubs
```

Format

A tibble with 8 rows and 2 variables.

Details

Data illustrating the difference between an outlier and an influencial case. The data came to me via David Hitchin, and he in turn got it from Dr Richard Roberts. I have no idea whether it's real or fictitious. The tibble contains the following variables:

- **pubs**: The number of pubs in a particular district of London
- **mortality**: The mortality rate in that district

Source

www.discovr.rocks/csv/pubs.csv

`puppies` *Puppy therapy data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
puppies
```

Format

A tibble with 15 rows and 3 variables.

Details

Despite the increase in puppies on my campus (which can only be a good thing) to reduce stress, the evidence base is pretty mixed. Imagine we wanted to contribute to this literature by running a study in which we randomized people into three groups (dose): (1) a control group, which could be a treatment as usual, a no treatment (no puppies) or ideally some kind of placebo group (we could give people in this group a cat disguised as a dog); (2) 15 minutes of puppy therapy (a low-dose group); and (3) 30 minutes of puppy contact (a high-dose group). The dependent variable was a measure of happiness ranging from 0 (as unhappy as I can possibly imagine) to 10 (as happy as I can possibly imagine). The design of this study mimics a very simple randomized controlled trial (as used in pharmacological, medical and psychological intervention trials) because people are randomized into a control group or groups containing the active intervention (in this case puppies, but in other cases a drug or a surgical procedure). The tibble contains the following variables:

- **id**: Participant id
- **dose**: Treatment group to which the participant was randomly assigned (No puppies (control), 15 minutes of puppy therapy, 30 minutes of puppy therapy)
- **happiness**: Self-reported happiness from 0 (as unhappy as I can possibly imagine being) to 10 (as happy as I can possibly imagine being)

Source

www.discovr.rocks/csv/puppies.csv

`puppy_love` *More puppy therapy data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
puppy_love
```

Format

A tibble with 30 rows and 4 variables.

Details

The researchers who conducted the puppy therapy study in puppies suddenly realized that a participant's love of dogs would affect whether puppy therapy would affect happiness. Therefore, they repeated the study on different participants, but included a self-report measure of love of puppies from 0 (I am a weird person who hates puppies, please be deeply suspicious of me) to 7 (puppies are the best thing ever, one day I might marry one). The tibble contains the following variables:

- **id**: Participant id
- **dose**: Treatment group to which the participant was randomly assigned (No puppies (control), 15 minutes of puppy therapy, 30 minutes of puppy therapy)
- **happiness**: Self-reported happiness from 0 (as unhappy as I can possibly imagine being) to 10 (as happy as I can possibly imagine being)
- **puppy_love**: Self-reported love of puppies from 0 (I am a weird person who hates puppies, please be deeply suspicious of me) to 7 (puppies are the best thing ever, one day I might marry one)

Source

www.discovr.rocks/csv/puppy_love.csv

raq *R Anxiety Questionnaire (RAQ)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
raq
```

Format

A tibble with 2,571 rows and 24 variables.

Details

Fictitious data relating to a fictional questionnaire about R anxiety. I can't stress enough how fictional this example is. Like, don't email me for the questionnaire the whole thing is figment of my mind (and some data simulation). I thought this would be obvious from the questions, but apparently not. Imagine that I wanted to design a questionnaire to measure a trait that I termed 'R anxiety'. I devised a questionnaire to measure various aspects of students' anxiety towards learning R, the RAQ. I generated (in my imagination) questions based on interviews (that never happened in real life) with anxious and non-anxious students and came up with 23 possible questions to include. Each question was a statement followed by a five-point Likert scale: *strongly disagree* = 1, *disagree*, *neither agree nor disagree*, *agree* and *strongly agree* (SD, D, N, A and SA respectively). What's more, I wanted to know whether anxiety about R could be broken down into specific forms of anxiety. In other words, what latent variables contribute to anxiety about R?

With a little help from a few lecturer friends (this never happened in real life) I collected 2571 completed questionnaires. The data are stored in this object with 2,571 rows and 24 columns.

- **id**: The student's id
- **raq_01**: responses (1-5) to the question *Statistics make me cry*
- **raq_02**: responses (1-5) to the question *My friends will think I'm stupid for not being able to cope with R*
- **raq_03**: responses (1-5) to the question *Standard deviations excite me*
- **raq_04**: responses (1-5) to the question *I dream that Pearson is attacking me with correlation coefficients*
- **raq_05**: responses (1-5) to the question *I don't understand statistics*
- **raq_06**: responses (1-5) to the question *I have little experience of computers*
- **raq_07**: responses (1-5) to the question *All computers hate me*
- **raq_08**: responses (1-5) to the question *I have never been good at mathematics*
- **raq_09**: responses (1-5) to the question *My friends are better at statistics than me*

- **raq_10**: responses (1-5) to the question *Computers are useful only for playing games*
- **raq_11**: responses (1-5) to the question *I did badly at mathematics at school*
- **raq_12**: responses (1-5) to the question *People try to tell you that R makes statistics easier to understand but it doesn't*
- **raq_13**: responses (1-5) to the question *I worry that I will cause irreparable damage because of my incompetence with computers*
- **raq_14**: responses (1-5) to the question *Computers have minds of their own and deliberately go wrong whenever I use them*
- **raq_15**: responses (1-5) to the question *Computers are out to get me*
- **raq_16**: responses (1-5) to the question *I weep openly at the mention of central tendency*
- **raq_17**: responses (1-5) to the question *I slip into a coma whenever I see an equation*
- **raq_18**: responses (1-5) to the question *R always crashes when I try to use it*
- **raq_19**: responses (1-5) to the question *Everybody looks at me when I use R*
- **raq_20**: responses (1-5) to the question *I can't sleep for thoughts of eigenvectors*
- **raq_21**: responses (1-5) to the question *I wake up under my duvet thinking that I am trapped under a normal distribution*
- **raq_22**: responses (1-5) to the question *My friends are better at R than I am*
- **raq_23**: responses (1-5) to the question *If I am good at statistics people will think I am a nerd*

Source

www.discovr.rocks/csv/raq.csv

`reality_tv` *Reality TV example*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
reality_tv
```

Format

A tibble with 32 rows and 4 variables

Details

A researcher hypothesized that reality TV show contestants start off with personality disorders that are exacerbated by being forced to spend time with people as attention-seeking as them. To test this hypothesis, she gave eight contestants a questionnaire measuring personality disorders before and after they entered the show. A second group of eight people were given the questionnaires at the same time; these people were short-listed to go on the show, but never did. The (fictional) data contains the following variables:

- **id**: participant ID
- **contestant**: whether the participant was a contestant or was on the short list but never went on the show
- **time**: the time at which personality disorder traits were measured (before or after the show)
- **pd_score**: the score on a personality disorder traits questionnaire

Source

www.discovr.rocks/csv/speed_date.csv

`roaming_cats` *Roaming cats data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
roaming_cats
```

Format

A tibble with 60 rows and 4 variables.

Details

Fictional data about roaming cats. I was interested in the relationship between the sex of a cat and how much time it spent away from home. I had heard that male cats disappeared for substantial amounts of time on long-distance roams around the neighbourhood (something about hormones driving them to find mates) whereas female cats tended to be more homebound. The data set has four variables

- **id**: Cat id
- **time**: Time spent away from home per week
- **sex**: biological sex of the cat as a factor

Source

www.discovr.rocks/csv/roaming_cats.csv

`rollercoaster` *Roaming cats data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
rollercoaster
```

Format

A tibble with 20 rows and 3 variables.

Details

Fictional data based on a study by Meston & Frohlich (2003) that showed that heterosexual people rate a picture of someone of the opposite sex as more attractive after riding a roller-coaster compared to before. Imagine we took 20 people as they came off the Rockit roller-coaster at Universal studios in Orlando and asked them to rate the attractiveness of people in a series of photographs on a scale of 0 (looks like Jabba the Hut) to 10 (looks like Princess Leia or Han Solo). The mean of their attractiveness ratings was the outcome. We also recorded their fear during the ride using a device that collates various indicators of physiological arousal and returns a value from 0, chill, to 10, terrified. This variable is the predictor. The prediction was that fear would be positively associated with ratings of attractiveness.

- **id**: Participant id
- **attractiveness**: Mean attractiveness rating people in a series of photographs from 0 (Jabba the Hut) to 10 (Princess Leia or Han Solo)
- **fear**: fear during the ride measured on a device that collates various indicators of physiological arousal into a value from 0, chill, to 10, terrified)

Source

www.discovr.rocks/csv/rollercoaster.csv

r_exam *R exam data data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
r_exam
```

Format

A tibble with 100 rows and 6 variables.

Details

Fictitious data relating to an R exam at two universities. The tibble contains the following variables:

- **id**: The student's id
- **exam**: first-year R exam scores as a percentage
- **computer**: a measure of computer literacy as a percentage
- **lecture**: percentage of statistics lectures attended
- **numeracy**: a measure of numerical ability out of 15
- **uni**: The university attended (Sussex University or Duncetown University)

Source

www.discovr.rocks/csv/r_exam.csv

santas_log *Self-help book data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
santas_log
```

Format

A tibble with 400 rows and 4 variables.

Details

Let's begin with a Christmas tale. A year ago Santa was resting in his workshop studying his nice and naughty lists. He noticed a name on the naughty list in bold, upper case letters. It said **ANDY FIELD OF UNIVERSITY OF SUSSEX**. He went to look up the file of this Andy Field character. He stared into his snow globe, and as the mists cleared he saw a sad, lonely, friend-less character walking across campus. Under one arm a box of chocolates, under the other a small pink Hippo. As he walked the campus he enticed the young students around him to follow him by offering chocolate. Like the Pied Piper, he led them to a large hall. Once inside, the boys and girls' eyes glistened in anticipation of more chocolate. Instead he unleashed a monologue about the general linear model of such fearsome tedium that Santa began to wonder how anyone could have grown to be so soulless and cruel.

Santa dusted off his sleigh and whizzed through the night sky to the Sussex campus. Once there he confronted the evil fiend that he had seen in his globe. "You've been a naughty boy," he said. "I give you a choice. Give up teaching statistics, or I will be forced to let the Krampus pay you a visit."

Andy looked sad, "But I love statistics," he said to Santa, "It's cool."

Santa pulled out a candy cane, from it emerged a screen. Just as he was about to instruct the screen to call the Krampus, an incoming message appeared: some presents had not been delivered last Christmas!

What was Santa to do? How could he find out what determines whether presents get delivered or not? He panicked.

Just then, Santa heard a sad little voice. It said, "I can help you".

"How? replied Santa.

"My students," he replied, "they can save Christmas. All they need are some data."

With that, Santa looked into his candy screen at the elves who had called him, and turned to Andy. "Tell them what you need."

Andy discovered that to deliver presents Santa uses a large team of elves, and that at each house they usually consume treats. The treats might be Christmas pudding, or sometimes mulled wine. He also discovered that they consume different quantities. Sometimes nothing is left, but other times there might be 1, 2, 3 or even 4 pieces of pudding or glasses of mulled wine. The Elves transmitted a log of 400 of the previous year's deliveries. The (fictional) data contains the following variables:

- **id**: Name of the elf doing the delivery
- **quantity**: How many treats the elf ate before attempting the delivery
- **treat**: which kind of treats were consumed (Christmas pudding or mulled wine)
- **delivered**: were the presents delivered (delivered or not delivered) The (fictional) data contains the following variables:

Source

www.discovr.rocks/csv/santas_log.csv

self_help *Self-help book data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
self_help
```

Format

A tibble with 20 rows and 3 variables.

Details

'Pop psychology' books sometimes spout nonsense that is unsubstantiated by science. I took 20 people in relationships and randomly assigned them to one of two groups. One group read the famous popular psychology book *Women are from Bras and men are from Penis*, and the other read *Marie Claire*. The outcome variable was their relationship happiness after their assigned reading. The (fictional) data contains the following variables:

- **id**: the participant's id
- **book**: whether the participants read *Women are from bras and men are from penis* or *Marie Claire*
- **happy**: the participant's relationship happiness after reading the book assigned to them

Source

www.discovr.rocks/csv/self_help.csv

self_help_dsur *Self-help book vs statistics book data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
self_help_dsur
```

Format

A tibble with 1000 rows and 3 variables.

Details

Twaddle and Sons, the publishers of Women are from Bras and men are from Penis, were upset about my claims that their book was as useful as a paper umbrella. They ran their own experiment (N = 500) in which relationship happiness was measured after participants had read their book and after reading the book you are currently reading. (Participants read the books in counterbalanced order with a six-month delay.) The (fictional) data contains the following variables:

- **id**: the participant's id
- **book**: whether relationship happiness was measured after reading *Women are from bras and men are from penis* or after reading *Discovering statistics using R*
- **happy**: the participant's relationship happiness after reading each book

Source

www.discovr.rocks/csv/self_help_dsur.csv

`senjutsu_pal` *Senjutsu palette*

Description

Colour palette based on Iron Maiden's Senjutsu album inner gatefold sleeve.

Usage

```
senjutsu_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_senjutsu(n, type = "discrete", reverse = FALSE, ...)

scale_colour_senjutsu(n, type = "discrete", reverse = FALSE, ...)

scale_fill_senjutsu(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale` `aesthetics` The names of the aesthetics that this scale works with. `scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for `y` axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.
`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(senjutsu_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_senjutsu()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_senjutsu()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_senjutsu()
```

---

sharman_2015 *Sharman & Dingle (2015) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
sharman_2015
```

Format

A tibble with 117 rows and 4 variables.

Details

There's a perception that listening to extreme music causes anger and associated behavioural problems. As an avid Metal fan and fairly non-angry type of person this sterotype bothers me. Luckily science has come to the rescue. Sharman & Dingle (2015) tested 39 fans of extreme music (metal). Their heart rate was measured at baseline, during a subsequent anger induction and while subsequently listening to music of their choice (which included a lot of bands listed at various point in the acknowledgements of my books). They collected subjective measures too, but this data file contains only the heart rate data from the study.

- **id**: The participant id (the original data had numeric IDs, which I have replaced with randomly generated alpha-numeric codes)
- **music**: Whether the participant was in the music or silence condition
- **phase**: Phase of the experiment (baseline, anger-induction, listening to music)
- **hr**: Heart rate (BPM)

Source

www.discovr.rocks/csv/sharman_2015.csv

References

- Sharman, L., & Dingle, G. A. (2015). Extreme metal music and anger processing. Frontiers in Human Neuroscience, 9. doi:10.3389/fnhum.2015.00272

---

shopping *Shopping and exercise data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
shopping
```

Format

A tibble with 10 rows and 3 variables.

Details

According to some highly unscientific research done by a UK department store chain and reported in Marie Claire magazine, shopping is good for you. They found that the average woman spends 150 minutes and walks 2.6 miles when she shops, burning off around 385 calories. In contrast, men spend only about 50 minutes shopping, covering 1.5 miles. This was based on strapping a pedometer on a mere 10 participants. Although I don't have the actual data, some simulated data based on these means are in this file.

- **sex**: biological sex of the individual
- **distance**: the distance travelled in miles
- **time**: the time spent shopping in minutes

Source

www.discovr.rocks/csv/shopping_exercise.csv

---

sit_pal *Somewhere in Time palette*

Description

Colour palette based on Iron Maiden's Somewhere in Time album sleeve.

Usage

```
sit_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_sit(n, type = "discrete", reverse = FALSE, ...)

scale_colour_sit(n, type = "discrete", reverse = FALSE, ...)

scale_fill_sit(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

n	number of colors
type	discrete or continuous
reverse	reverse order, Default: FALSE
...	Arguments passed on to ggplot2::discrete_scale aesthetics The names of the aesthetics that this scale works with. scale_name **[Deprecated]** The name of the scale that should be used for error messages associated with this scale. palette A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., scales::pal_hue()).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(sit_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_sit()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_sit()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_sit()
```

---

sniffer_dogs *Sniffer dogs*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
sniffer_dogs
```

Format

A tibble with 32 rows and 3 variables.

Details

When the alien invasion comes we'll need spaniels (or possibly other dogs, but lets hope its mainly spaniels because spaniels are cool) to help us to identify the space lizards. The top-secret government agency for Training Extra-terrestrial Reptile Detection (TERD) was put together to test the plausibility of training sniffer dogs to detect aliens. Over many trials 8 of their best dogs (Milton, Woofy, Ramsey, Mr. Snifficus III, Willock, The Venerable Dr. Waggy, Lord Scenticle, and Professor Nose) were recruited for a pilot study. During training, these dogs were rewarded for making vocalizations while sniffing alien space lizards (which they happened to have a few of in Hangar 18). On the test trial, the 8 dogs were allowed to sniff 4 entities for 1-minute each: an alien space lizard, a shapeshifting alien space lizard who had taken on humanoid form and worked undetected as a statistics lecturer, a human, and a human mannequin). The number of vocalizations made during each 1-minute sniffing session was recorded. For more alien lizard and sniffer dog adventures see alien_scents.

- **dog_name**: the name of the sniffer dog
- **entity**: the entity being sniffed by the sniffer dog (alien, alien in humanoid form (shapeshifter), human, human mannequin)
- **vocalizations**: the number of vocalizations made by the dog during a 1-minute sniff

Source

www.discovr.rocks/csv/sniffer_dogs.csv

`social_anxiety` *Social anxiety data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
social_anxiety
```

Format

A tibble with 134 rows and 4 variables.

Details

Anxiety disorders take on different shapes and forms, and each disorder is believed to be distinct and have unique causes. We can summarize the disorders and some popular theories as follows:

- Social Anxiety: Social anxiety disorder is a marked and persistent fear of 1 or more social or performance situations in which the person is exposed to unfamiliar people or possible scrutiny by others. This anxiety leads to avoidance of these situations. People with social phobia are believed to feel elevated feelings of shame.
- Obsessive Compulsive Disorder (OCD): OCD is characterized by the everyday intrusion into conscious thinking of intense, repetitive, personally abhorrent, absurd and alien thoughts (Obsessions), leading to the endless repetition of specific acts or to the rehearsal of bizarre and irrational mental and behavioural rituals (compulsions).

Social anxiety and obsessive compulsive disorder are seen as distinct disorders having different causes. However, there are some similarities. They both involve some kind of attentional bias: attention to bodily sensation in social anxiety and attention to things that could have negative consequences in OCD. They both involve repetitive thinking styles: social phobics ruminate about social encounters after the event (known as post-event processing), and people with OCD have recurring intrusive thoughts and images. They both involve safety behaviours (i.e. trying to avoid the thing that makes you anxious).

This might lead us to think that, rather than being different disorders, they are manifestations of the same core processes (Field & Cartwright-Hatton, 2008). One way to research this possibility would be to see whether social anxiety can be predicted from measures of other anxiety disorders. If social anxiety disorder and OCD are distinct we should expect that measures of OCD will not predict social anxiety. However, if there are core processes underlying all anxiety disorders, then measures of OCD should predict social anxiety. The data contains three variables:

- **spai**: The Social Phobia and Anxiety Inventory (SPAI), which measures levels of social anxiety.
- **iii**: The Interpretation of Intrusions Inventory (III).
- **obq**: Obsessive Beliefs Questionnaire (OBQ), which measures the degree to which people experience obsessive beliefs like those found in OCD.
- **tosca**: The Test of Self-Conscious Affect (TOSCA), which measures shame.

Source

www.discovr.rocks/csv/social_anxiety.csv

References

- Field, A. P., & Cartwright-Hatton, S. (2008). Shared and unique cognitive factors in social anxiety. *International Journal of Cognitive Therapy*, *1*, 206–222. doi:10.1521/ijct.2008.1.3.206

`social_media` *Social media and grammar data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
social_media
```

Format

A tibble with 100 rows and 4 variables.

Details

Imagine we conducted an experiment in which a group of 25 people were encouraged to message their friends and post on social media using their mobiles over a six-month period. A second group of 25 people were banned from messaging and social media for the same period by being given armbands that administered painful shocks in the presence of microwaves (like those emitted from phones). The outcome was a percentage score on a grammatical test that was administered both before and after the intervention. The first independent variable was, therefore, social media use (encouraged or banned) and the second was the time at which grammatical ability was assessed (baseline or after 6 months). These data are fictional. The object contains the following variables:

- **id**: participant id
- **media_use**: Whether the participant was encouraged to use social media or banned from using it
- **time**: the time at which the grammar test was taken: before social media use was manipulated (baseline) and 6 months later
- **grammar**: the score on a grammar test (as a percentage)

Source

www.discovr.rocks/csv/social_media.csv

soya *Soya and sperm counts data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
soya
```

Format

A tibble with 80 rows and 3 variables.

Details

I read a story in a newspaper (yes, back when they existed) claiming that the chemical genistein, which is naturally occurring in soya, was linked to lowered sperm counts in Western males. When you read the actual study, it had been conducted on rats, it found no link to lowered sperm counts, but there was evidence of abnormal sexual development in male rats (probably because genistein acts like oestrogen). As journalists tend to do, a study showing no link between soya and sperm counts was used as the scientific basis for an article about soya being the cause of declining sperm counts in Western males. Imagine the rat study was enough for us to want to test this idea in humans. We recruit 80 males and split them into four groups that vary in the number of soya 'meals' (a dinner containing 75g of soya) they ate per week over a year: no soya meals (i.e., none in the whole year), one per week (52 over the year), four per week (208 over the year), and seven per week (364 over the year). At the end of the year, participants produced some sperm that I could count (when I say 'I', I mean someone else in a laboratory as far away from me as humanly possible). The fictitious data contain the following variables:

- **id**: The participant's id
- **soya**: How many soya meals per week consumed over a year (none, 1, 4 and 7)
- **sperm**: number of sperm cells per milliliter of semen in millions (yes, I did have to Google that)

Source

www.discovr.rocks/csv/soya.csv

`speed_date` *Speed dating data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
speed_date
```

Format

A tibble with 180 rows and 5 variables

Details

Imagine a scientist designed a study to look at the interplay between looks, personality and dating strategies on evaluations of a date. She set up a speed-dating night with 9 tables at which there sat a 'date'. All the dates were stooges selected to vary in their attractiveness (high, average and low), their personality (high charisma, average charisma, writes statistics books), and also the strategy they were told to employ during the conversation (normal or playing hard to get). The dates were trained before the study to act charismatically to varying degrees, and also how to act in a way that made them seem unobtainable (hard to get) or not. As such, across the nine dates/stooges there were three 'high attractive' people one of whom acted charismatically, one who acted normally (average) and another who acted with low charisma, likewise for the three average looking dates and the three low attractiveness dates. Therefore, each participant attending a speed-dating night would be exposed to all combinations of attractiveness and charisma (these are repeated measures).

Upon arrival participants were randomly assigned a blue or red sticker. For the participants with the red sticker the stooges played hard to get (unobtainable) and for those with a blue sticker they acted normally. Over the course a few nights 20 people attended, spent 5-minutes with each of the 9 'dates' and then rated how much they'd like to have a proper date with the person as a percentage (100% = 'I'd pay large sums of money for their phone number', 0% = 'I'd pay a large sum of money for a plane ticket to get me as far away from them as possible'). The (fictional) data contains the following variables:

- **id**: participant ID
- **strategy**: Whether the stooge acted normally or played hard to get
- **looks**: Whether the stooge was rated as high, average or low on looks
- **personality**: Whether the stooge acted with high, average or low charisma
- **date**: rating how much the participant would like to have a proper date with the stooge as a percentage (100% = 'I'd pay large sums of money for their phone number', 0% = 'I'd pay a large sum of money for a plane ticket to get me as far away from them as possible')

Source

www.discovr.rocks/csv/speed_date.csv

`ssoass_pal` *Seventh Son of a Seventh Son palette*

Description

Colour palette based on Iron Maiden's Seventh Son of a Seventh Son album sleeve.

Usage

```
ssoass_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_ssoass(n, type = "discrete", reverse = FALSE, ...)

scale_colour_ssoass(n, type = "discrete", reverse = FALSE, ...)

scale_fill_ssoass(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale`

`aesthetics` The names of the aesthetics that this scale works with.

`scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object

- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(ssoass_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
```

```
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_ssoass()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_ssoass()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_ssoass()
```

`stalker` *Stalking therapy*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
stalker
```

Format

A tibble with 50 rows and 4 variables.

Details

Some fictional data about therapy for stalking. A few years back I was stalked. You'd think they could have found someone a bit more interesting to stalk, but apparently times were hard. It could have been a lot worse, but it wasn't particularly pleasant. I imagined a world in which a psychologist tried two different therapies on different groups of stalkers (25 stalkers in each treatment). To

the first group he gave cruel-to-be-kind therapy (every time the stalkers followed him around, or sent him a letter, the psychologist attacked them with a cattle prod). The second therapy was psychodyshamic therapy, in which stalkers were hypnotized and regressed into their childhood to discuss their penis (or lack of penis), their father's penis, their dog's penis, the seventh penis of a seventh penis, and any other penis that sprang to mind. The psychologist measured the number of hours stalking in one week both before (stalk_pre) and after (stalk_post) treatment.The object contains the following variables:

- **id**: Participant's id code
- **therapy**: Whether the person was assigned to *Cruel to be kind therapy* or *Psychodyshamic* therapy
- **stalk_pre**: number of hours the person spent stalking in one week before therapy
- **stalk_post**: number of hours the person spent stalking in one week after therapy

Source

www.discovr.rocks/csv/stalker.csv

`students` *Students and lecturers data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
students
```

Format

A tibble with 10 rows and 7 variables.

Details

Some fictional data about students and lecturers. The object contains the following variables:

- **name**: Name of person
- **birth_date**: Date of birth (Year-month-day)
- **group**: whether the person is a student or lecturer
- **friends**: how many friends the person has. That's actual friends, not social media friends.
- **alcohol**: Units of alcohol consumed per week
- **income**: income (per anum)
- **neurotic**: Score on a neuroticism scale

Source

www.discovr.rocks/csv/students.csv

`superhero` *Superhero data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
superhero
```

Format

A tibble with 30 rows and 3 variables.

Details

Children wearing superhero costumes are more likely to harm themselves because of the unrealistic impression of invincibility that these costumes could create. For example, children have reported to hospital with severe injuries because of trying 'to initiate flight without having planned for landing strategies' (Davies, Surridge, Hole, & Munro-Davies, 2007). I can relate to the imagined power that a costume bestows upon you; indeed, I have been known to dress up as Fisher by donning a beard and glasses and trailing a goat around on a lead in the hope that it might make me more knowledgeable about statistics. These fictional data contain the severity of injury (on a scale from 0, no injury, to 100, death) for children reporting to the accident and emergency department at hospitals, and information on which superhero costume they were wearing (hero): Spiderman, Superman, the Hulk or a teenage mutant ninja turtle. The fictitious data contain the following variables:

- **id**: The participant's id
- **hero**: The costume being worn at the time of injury (Spiderman, Superman, the Hulk or a teenage mutant ninja turtle)
- **injury**: the severity of injury (on a scale from 0, no injury, to 100, death)

Source

www.discovr.rocks/csv/superhero.csv

supermodel *Supermodel data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
supermodel
```

Format

A tibble with 231 rows and 4 variables.

Details

A fashion student was interested in factors that predicted the salaries of male and female catwalk models. She collected data from 231 models (supermodel.csv). For each model she asked them their salary per day (salary), their age (age), their length of experience as models (years), and their industry status as a model as their percentile position rated by a panel of experts (status). The fictitious data contain the following variables:

- **salary**: The model's salary
- **age**: The model's age (years)
- **years**: The model's experience (years in the industry)
- **status**: Model's status as their percentile position (%) rated by a panel of experts.

Source

www.discovr.rocks/csv/supermodel.csv

switch *Switch: games console injuries*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
switch
```

Format

A tibble with 120 rows and 5 variables.

Details

Fictional data about injuries while playing video games on a console. There are reports of increases in injuries related to playing games consoles. These injuries were attributed mainly to muscle and tendon strains. A researcher hypothesized that a stretching warm-up before playing games would help lower injuries, and that athletes would be less susceptible to injuries because their regular activity makes them more flexible. She took 60 athletes and 60 non-athletes (athlete); half of them played on a Nintendo Switch and half watched others playing as a control (switch), and within these groups half did a 5-minute stretch routine before playing/watching whereas the other half did not (stretch). The outcome was a pain score out of 10 (where 0 is no pain, and 10 is severe pain) after playing for 4 hours (injury).

- **id**: Participant's id
- **athlete**: Whether the participant was an athlete or not
- **stretch**: Whether the participant warmed up with stretching (or not)
- **switch**: Whether the participant played Nintendo Switch games or watched someone else playing
- **injury**: Injury severity (where 0 is no pain, and 10 is severe pain)

Source

www.discovr.rocks/csv/switch.csv

`tablets` *Tablet sales data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tablets
```

Format

A tibble with 240 rows and 4 variables.

Details

A company owner was interested in how to make his brand of (computer) tablets more desirable. He collected data on how cool people perceived a product's advertising to be, how cool they thought the product was, and how desirable they found the product. Am I showing my age by using the word 'cool'? The fictitious data contain the following variables:

- **id**: Participant ID
- **advert_cool**: Perceived 'coolness' of the advertising campaign from 0 (as cool as Andy Field) to 5 (as cool as something that makes you go 'wow, that's sick', or whatever it is that people under the age of 25 say these days)
- **desirability**: The desirability of the product from (0 as desirable as Andy Field) to 10 (I *really* want one of those)
- **product_cool**: Perceived 'coolness' of the product from from 0 (designed by Andy Field) to 5 (Designed by Apple).

Source

www.discovr.rocks/csv/tablets.csv

`teaching` *Method of teaching data (2 groups)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
teaching
```

Format

A tibble with 20 rows and 3 variables.

Details

The data show the score (out of 20) for 20 different students, some of whom are biologically male and others biologically female, and some of whom were taught using positive reinforcement (being nice) and others who were taught using punishment (electric shock)

- **id**: participant ID
- **method**: The type of teaching method used
- **sex**: Biological sex of the individual
- **mark**: The score out of 20 on a test

Source

www.discovr.rocks/csv/teaching.csv

`teach_method` *Method of teaching data (3 groups)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
teach_method
```

Format

A tibble with 30 rows and 3 variables.

Details

To test how different teaching methods affected students' knowledge I took three statistics modules where I taught the same material. For one module I wandered around with a large cane and beat anyone who asked daft questions or got questions wrong (punish). In the second I encouraged students to discuss things that they found difficult and gave anyone working hard a nice sweet (reward). In the final course I neither punished nor rewarded students' efforts (indifferent). I measured the students' exam marks (percentage). This fictional data contains the following variables

- **id**: participant's id
- **group**: The type of teaching method used (Punish, Reward, Indifferent)
- **exam**: The exam mark (%)

Source

www.discovr.rocks/csv/teach_method.csv

`tea_15` *Tea data (small sample)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tea_15
```

Format

A tibble with 15 rows and 3 variables:

Details

One of my favourite activities, especially when trying to do brain-melting things like writing statistics books, is drinking tea. I am English, after all. Fortunately, tea improves your cognitive function – well, it does in old Chinese people at any rate (Feng, Gwee, Kua, & Ng, 2010). I may not be Chinese and I'm not that old, but I nevertheless enjoy the idea that tea might help me think. Here are some (fictional) data based on Feng et al.'s study that measured the number of cups of tea drunk per day and cognitive functioning (out of 80) in 15 people.

- **id**: participant ID
- **tea**: the number of cups of tea a person drinks per day
- **cog_fun**: cognitive functioning (out of 80)

Source

www.discovr.rocks/csv/tea_makes_you_brainy_15.csv

References

- Feng, L., Gwee, X., Kua, E. H., & Ng, T. P. (2010). Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *Journal of Nutrition Health & Aging*, *14*, 433-438.

`tea_716` *Tea data (large sample)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tea_716
```

Format

A tibble with 716 rows and 3 variables:

Details

One of my favourite activities, especially when trying to do brain-melting things like writing statistics books, is drinking tea. I am English, after all. Fortunately, tea improves your cognitive function – well, it does in old Chinese people at any rate (Feng, Gwee, Kua, & Ng, 2010). I may not be Chinese and I'm not that old, but I nevertheless enjoy the idea that tea might help me think. Here are some (fictional) data based on Feng et al.'s study that measured the number of cups of tea drunk per day and cognitive functioning (out of 80) in 716 people.

- **id**: participant ID
- **tea**: the number of cups of tea a person drinks per day
- **cog_fun**: cognitive functioning (out of 80)

Source

www.discovr.rocks/csv/tea_makes_you_brainy_716.csv

References

- Feng, L., Gwee, X., Kua, E. H., & Ng, T. P. (2010). Cognitive function and tea consumption in community dwelling older Chinese in Singapore. *Journal of Nutrition Health & Aging*, *14*, 433-438.

`text_messages` *Messaging apps and grammar example*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
text_messages
```

Format

A tibble with 100 rows and 4 variables

Details

Text messaging and Twitter encourage communication using abbreviated forms of words (if u no wat I mean). A researcher wanted to see the effect this had on children's understanding of grammar. One group of 25 children was encouraged to send text messages on their mobile phones over a 6-month period. A second group of 25 was forbidden from sending text messages for the same period (to ensure adherence, this group were given armbands that administered painful shocks in the presence of a phone signal). The outcome was a score on a grammatical test (as a percentage) that was measured both before and after the experiment. The (fictional) data contains the following variables:

- **id**: participant ID
- **group**: whether the participant was assigned to the text message group or control group
- **time**: the time at which grammar ability was measured (baseline or 6 months later)
- **grammar**: the score on the grammar test as a percentage (%)

Source

www.discovr.rocks/csv/speed_date.csv

`tol_muted_pal` *Tol muted palette*

Description

Colour palette used in the book based on Paul Tol's muted palette https://sronpersonalpages.nl/~pault/data/colourschemes.pdf.

Usage

```
tol_muted_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_tol(n, type = "discrete", reverse = FALSE, ...)

scale_colour_tol(n, type = "discrete", reverse = FALSE, ...)

scale_fill_tol(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colors
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale` `aesthetics` The names of the aesthetics that this scale works with. `scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale. `palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`). `name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted. `breaks` One of: • `NULL` for no breaks • `waiver()` for the default breaks (the scale limits) • A character vector of breaks • A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation. `labels` One of: • `NULL` for no labels • `waiver()` for the default labels computed by the transformation object • A character vector giving labels (must be same length as `breaks`)

- An expression vector (must be the same length as breaks). See ?plotmath for details.
- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(tol_muted_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.
```

```
classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))

# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_tol()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_tol()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_tol()
```

`tosser` *The Teaching of Statistics for Scientific Experiments—Revised (TOSSE-R) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tosser
```

Format

A tibble with 239 rows and 29 variables.

Details

Fictitious data relating to a fictional questionnaire about The Teaching of Statistics for Scientific Experiments. Again, I stress that this example is fictional. I thought the name of the questionnaire would give it away, I mean, no-one is calling a questionnaire TOSSER are they? Don't email me for the questionnaire, it's all made up, you definitley don't want to base your research upon it. Imagine I wanted to revise the 'Teaching of Statistics for Scientific Experiments' (TOSSE) questionnaire,

which is (I mean, it isn't because I made it up) based on Bland's theory that says that good research methods lecturers should have: (1) a profound love of statistics; (2) an enthusiasm for experimental design; (3) a love of teaching; and (4) a complete absence of normal interpersonal skills. These characteristics should be related (i.e., correlated). The revised version of this questionnaire (TOSSE – R) was given to 239 research methods lecturers to see if it supported Bland's theory. Each question was a statement followed by a five-point Likert scale: *strongly disagree* = 1, *disagree*, *neither agree nor disagree*, *agree* and *strongly agree* (SD, D, N, A and SA respectively). The data contains the following variables

- **id**: The student's id
- **q_01**: responses (1-5) to the question *I once woke up in the middle of a vegetable patch hugging a turnip that I'd mistakenly dug up thinking it was Roy's largest root*
- **q_02**: responses (1-5) to the question *Students are like irritating pigeons pecking away at my sanity*
- **q_03**: responses (1-5) to the question *I memorize probability values for the F-distribution*
- **q_04**: responses (1-5) to the question *I worship at the shrine of Pearson*
- **q_05**: responses (1-5) to the question *I still live with my mother and have little personal hygiene*
- **q_06**: responses (1-5) to the question *Teaching others makes me want to swallow a large bottle of bleach because the pain of my burning oesophagus would be light relief in comparison*
- **q_07**: responses (1-5) to the question *Helping others to understand sums of squares is a great feeling*
- **q_08**: responses (1-5) to the question *I like control conditions*
- **q_09**: responses (1-5) to the question *I calculate 3 ANOVAs in my head before getting out of bed every morning*
- **q_10**: responses (1-5) to the question *I could spend all day explaining statistics to people*
- **q_11**: responses (1-5) to the question *I like it when people tell me I've helped them to understand factor rotation*
- **q_12**: responses (1-5) to the question *People fall asleep as soon as I open my mouth to speak*
- **q_13**: responses (1-5) to the question *Designing experiments is fun*
- **q_14**: responses (1-5) to the question *I'd rather think about appropriate dependent variables than meet people*
- **q_15**: responses (1-5) to the question *I soil my pants with excitement at the mention of Factor Analysis*
- **q_16**: responses (1-5) to the question *Thinking about whether to use repeated- or independent-measures thrills me*
- **q_17**: responses (1-5) to the question *I enjoy sitting in the park contemplating whether to use participant observation in my next experiment*
- **q_18**: responses (1-5) to the question *Standing in front of 300 people in no way makes me lose control of my bowels*
- **q_19**: responses (1-5) to the question *I like to help students*
- **q_20**: responses (1-5) to the question *Passing on knowledge is the greatest gift you can bestow an individual*

- **q_21**: responses (1-5) to the question *Thinking about Bonferroni corrections gives me a tingly feeling in my groin*
- **q_22**: responses (1-5) to the question *I quiver with excitement when thinking about designing my next experiment*
- **q_23**: responses (1-5) to the question *I often spend my spare time talking to the pigeons ... and even they die of boredom*
- **q_24**: responses (1-5) to the question *I tried to build myself a time machine so that I could go back to the 1930s and follow Fisher around on my hands and knees licking the floor on which he'd just trodden*
- **q_25**: responses (1-5) to the question *I love teaching*
- **q_26**: responses (1-5) to the question *I spend lots of time helping students*
- **q_27**: responses (1-5) to the question *I love teaching because students have to pretend to like me or they'll get bad marks*
- **q_28**: responses (1-5) to the question *My cat is my only friend*

Source

www.discovr.rocks/csv/tosser.csv

`tuk_2011` *Tuk et al. (2011) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tuk_2011
```

Format

A tibble with 102 rows and 3 variables

Details

Visceral factors that require us to engage in self control (such as a filling bladder) can affect our inhibtory abilities in unrelated domains. In a fascinating study by Tuk, Trampe, and Warlop (2011) participants were given five cups of water: one group was asked to drink them all, whereas another was asked to take a sip from each. This manipulation led one group to have full bladders and the other group relatively empty (urgency). Later on, these participants were given eight trials on which they had to choose between a small financial reward that they would receive soon (SS) or a large financial reward for which they would wait longer (LL). They counted how many trials participants choose the LL reward as an indicator of inhibitory control (ll_sum). The data contains three variables:

- **id**: participant ID
- **urgency**: whether participants were in a high urination urgency condition (they drank everything) or a low urgency condition (they took sips of water)
- **ll_sum**: the total number of LL rewards

Source

www.discovr.rocks/csv/tuk_2011.csv

References

- Tuk, M. A., Trampe, D., & Warlop, L. (2011). Inhibitory spillover: increased urination urgency facilitates impulse control in unrelated domains. *Psychological Science*, 22, 627–633. doi:10.1177/0956797611404901

---

tumour *Mobile phone use and brain tumour data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tumour
```

Format

A tibble with 102 rows and 3 variables

Details

Mobile phones emit microwaves, and so holding one next to your brain for large parts of the day is a bit like sticking your brain in a microwave oven and pushing the 'cook until well done' button. If we wanted to test this experimentally, we could get six groups of people and strap a mobile phone on their heads, then by remote control turn the phones on for a certain amount of time each day. After six months, we measure the size of any tumour (in mm^3) close to the site of the phone antenna (just behind the ear). The six groups experienced 0, 1, 2, 3, 4 or 5 hours per day of phone microwaves for six months. The fictitious data contains three variables:

- **id**: participant ID
- **usage**: how many hours per day were the phones active for (0, 1, 2, 3, 4, or 5 hours)
- **tumour**: Size of any tumour (in mm^3)

Source

www.discovr.rocks/csv/tumour.csv

tutor_marks *Tutor marking data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
tutor_marks
```

Format

A tibble with 32 rows and 3 variables.

Details

It is common that lecturers obtain reputations for being 'hard' or 'light' markers, but there is often little to substantiate these reputations. A group of students investigated the consistency of marking by submitting the same essays to four different lecturers. The outcome was the percentage mark given by each lecturer and the predictor was the lecturer who marked the report. The fictitious data contains three variables:

- **id**: participant's id
- **tutor**: The tutor who marked the work
- **exam**: The mark on the essay (%)

Source

www.discovr.rocks/csv/tutor_marks.csv

van_bourg_2020 *Van Bourg et al. (2020) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
van_bourg_2020
```

Format

A tibble with 201 rows and 6 variables.

Details

Pet dogs often engage in behaviours helpful to their owners (mine likes to cuddle me when I've had a bad day, and in fact when I've had a good day, and now I think of it, pretty much any day regardless of how good or bad its been). It's unclear whether these behaviours are truly prosocial. Can a dog engage in prosocial behaviours that haven't been explicitly trained? Bourg et al (2020) addressed this question by trapping some dog's owners in boxes! In the study 60 dogs were tested in three conditions all of which involved being in a room with large restrainer box (a large acrylic box with holes in the side that could be closed by resting a foam board door across its opening). Each dog had three experiences in the room and each time the experimenters were interested in whether the dog would open the restrainer box within 120 seconds. The order of the 3 experiences was counterbalanced so different dogs completed the experiences in different orders.

- The **food** condition: food was dropped into the restrainer. This condition was to test whether the dog was capable of moving the foam board door to open the box (to get the food).
- The distress condition: the dogs' owner was placed in the restrainer and was instructed to call for help in a distressed tone.
- The reading condition: the dogs' owner was placed in the restrainer and was instructed to read from a magzine at the same pace and in the same tone as in the distress condition.

This data contains a subset of variables from the study, but the full dataset is available in the supplementary materials of the paper doi:10.1371/journal.pone.0231742.s001. The data contains the following variables

- **name**: The dog's name
- **dog_id**: A unique identifier for each dog
- **condition**: Which condition the dog was participating in at the time (distress, food, reading).
- **test_number**: A number form 1 to 3 indicating the order in which the particular condition was administered. For example, 2 would indicate that the data relate to the second of the three tests that the dog experienced.
- **latency**: The time taken to open the box in seconds. If the dog did not open the box a maximum of 120s was recorded.
- **opened_door**: Dis the dog open the restrainer box (1 = yes, 0 = no).

Source

www.discovr.rocks/csv/van_bourg_2020.csv

References

- Van Bourg, J., Patterson, J. E., & Wynne, C. D. L. (2020). Pet dogs (Canis lupus familiaris) release their trapped and distressed owners: Individual variation and evidence of emotional contagion. *PLOS ONE*, 15(4), e0231742. doi:10.1371/journal.pone.0231742

`video_games` *Video game and aggression data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
video_games
```

Format

A tibble with 442 rows and 4 variables

Details

Video games are among the favourite online activities for young people. These games have been linked to increased aggression in youths. Another predictor of aggression and conduct problems is callous-unemotional traits such as lack of guilt, lack of empathy, callous use of others for personal gain. Imagine that a scientist explored the relationship between playing violent video games and aggression. She measured aggressive behaviour, callous-traits, and the number of hours per week they play video games in 442 youths. These fictitious data contains three variables:

- **id**: participant ID
- **agress**: a measure of aggressive behaviour from 0 (no aggression at all) to 100 (extremely aggressive)
- **vid_game**: number of hours per week spent playing video games
- **caunts**: callous unemotional traits measured on the Inventory of Callous-Unemotional Traits (ICU), ranging from 0 (none) to 72 (extreme)

Source

www.discovr.rocks/csv/video_games.csv

`virtual_pal` *Virtual IX palette*

Description

Colour palette based on Iron Maiden's Virtual IX album sleeve.

Usage

```
virtual_pal(n, type = c("discrete", "continuous"), reverse = FALSE)

scale_color_virtual(n, type = "discrete", reverse = FALSE, ...)

scale_colour_virtual(n, type = "discrete", reverse = FALSE, ...)

scale_fill_virtual(n, type = "discrete", reverse = FALSE, ...)
```

Arguments

`n`	number of colours
`type`	discrete or continuous
`reverse`	reverse order, Default: FALSE
`...`	Arguments passed on to `ggplot2::discrete_scale`

`aesthetics` The names of the aesthetics that this scale works with.

`scale_name` **[Deprecated]** The name of the scale that should be used for error messages associated with this scale.

`palette` A palette function that when called with a single integer argument (the number of levels in the scale) returns the values that they should take (e.g., `scales::pal_hue()`).

`name` The name of the scale. Used as the axis or legend title. If `waiver()`, the default, the name of the scale is taken from the first mapping used for that aesthetic. If `NULL`, the legend title will be omitted.

`breaks` One of:

- `NULL` for no breaks
- `waiver()` for the default breaks (the scale limits)
- A character vector of breaks
- A function that takes the limits as input and returns breaks as output. Also accepts rlang lambda function notation.

`labels` One of:

- `NULL` for no labels
- `waiver()` for the default labels computed by the transformation object
- A character vector giving labels (must be same length as `breaks`)
- An expression vector (must be the same length as breaks). See ?plotmath for details.

- A function that takes the breaks as input and returns labels as output. Also accepts rlang lambda function notation.

`limits` One of:

- `NULL` to use the default scale values
- A character vector that defines possible values of the scale and their order
- A function that accepts the existing (automatic) values and returns new ones. Also accepts rlang lambda function notation.

`expand` For position scales, a vector of range expansion constants used to add some padding around the data to ensure that they are placed some distance away from the axes. Use the convenience function `expansion()` to generate the values for the `expand` argument. The defaults are to expand the scale by 5% on each side for continuous variables, and by 0.6 units on each side for discrete variables.

`na.translate` Unlike continuous scales, discrete scales can easily show missing values, and do so by default. If you want to remove missing values from a discrete scale, specify `na.translate = FALSE`.

`na.value` If `na.translate = TRUE`, what aesthetic value should the missing values be displayed as? Does not apply to position scales where `NA` is always placed at the far right.

`drop` Should unused factor levels be omitted from the scale? The default, `TRUE`, uses the levels that appear in the data; `FALSE` includes the levels in the factor. Please note that to display every level in a legend, the layer should use `show.legend = TRUE`.

`guide` A function used to create a guide or its name. See `guides()` for more information.

`position` For position scales, The position of the axis. `left` or `right` for y axes, `top` or `bottom` for x axes.

`call` The `call` used to construct the scale for reporting messages.

`super` The super class to use for the constructed scale

Value

A discrete or continuous scale.

Examples

```
library(scales)
show_col(virtual_pal()(8))

library(discovr)
library(ggplot2)

# Get albums in the classic era from the discovr::eddiefy data.
# I'm not including fear of the dark because it's not in any way classic.
# No prayer for the dying was pushing its luck too if I'm honest.

classic_era <- subset(eddiefy, year < 1992, select = c("energy", "valence", "album_name"))
```

```
# Plot some data and apply theme to color (note US English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_color_virtual()

# Plot some data and apply theme to colour (note UK English)

ggplot(classic_era, aes(x = energy, y = valence, color = album_name)) +
  geom_point(size = 2) +
  theme_minimal() +
  scale_colour_virtual()

# Plot some data and apply theme to fill

ggplot(classic_era, aes(x = album_name, y = valence, fill = album_name)) +
  geom_violin() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_fill_virtual()
```

`williams` *Williams' questionnaire of organizational ability data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
williams
```

Format

A tibble with 239 rows and 29 variables.

Details

Dr Sian Williams (University of Brighton) devised a questionnaire to measure organizational ability. She predicted five factors to do with organizational ability: (1) preference for organization; (2) goal achievement; (3) planning approach; (4) acceptance of delays; and (5) preference for routine. These dimensions are theoretically independent. Williams's questionnaire contains 28 items using a seven-point Likert scale (1 = *strongly disagree*, 4 = *neither*, 7 = *strongly agree*). She gave it to 239 people.

- **participant**: The participant id

- **sex**: The participant biological sex
- **org1**: responses (1-7) to the question *I like to have a plan to work to in everyday life*
- **org2**: responses (1-7) to the question *I feel frustrated when things don't go to plan*
- **org3**: responses (1-7) to the question *I get most things done in a day that I want to*
- **org4**: responses (1-7) to the question *I stick to a plan once I have made it*
- **org6**: responses (1-7) to the question *I enjoy spontaneity and uncertainty*
- **org7**: responses (1-7) to the question *I feel frustrated if I can't find something I need*
- **org9**: responses (1-7) to the question *I find it difficult to follow a plan through*
- **org10**: responses (1-7) to the question *I am an organized person*
- **org11**: responses (1-7) to the question *I like to know what I have to do in a day*
- **org12**: responses (1-7) to the question *Disorganized people annoy me*
- **org13**: responses (1-7) to the question *I leave things to the last minute*
- **org14**: responses (1-7) to the question *I have many different plans relating to the same goal*
- **org16**: responses (1-7) to the question *I like to have my documents filed and in order*
- **org17**: responses (1-7) to the question *I find it easy to work in a disorganized environment*
- **org18**: responses (1-7) to the question *I make to do lists and achieve most of the things on it*
- **org19**: responses (1-7) to the question *My workspace is messy and disorganized*
- **org20**: responses (1-7) to the question *I like to be organized*
- **org21**: responses (1-7) to the question *Interruptions to my daily routine annoy me*
- **org22**: responses (1-7) to the question *I feel that I am wasting my time*
- **org23**: responses (1-7) to the question *I forget the plans I have made*
- **org24**: responses (1-7) to the question *I prioritize the things I have to do*
- **org25**: responses (1-7) to the question *I like to work in an organized environment*
- **org26**: responses (1-7) to the question *I feel relaxed when I don't have a routine*
- **org27**: responses (1-7) to the question *I set deadlines for myself and achieve them*
- **org28**: responses (1-7) to the question *I change rather aimlessly from one activity to another during the day*
- **org29**: responses (1-7) to the question *I have trouble organizing the things I have to do*
- **org30**: responses (1-7) to the question *I put tasks off to another day*
- **org31**: responses (1-7) to the question *I feel restricted by schedules and plans*

Source

www.discovr.rocks/csv/williams.csv

xbox *Xbox: games console injuries*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
xbox
```

Format

A tibble with 40 rows and 4 variables.

Details

Fictional data about injuries while playing video games on a console. A researcher was interested in what factors contributed to injuries resulting from game console use. She tested 40 participants who were randomly assigned to either an active or static game played on either a Nintendo Switch or Xbox One Kinect. At the end of the session their physical condition was evaluated on an injury severity scale.

- **id**: Participant's id
- **game**: Whether the participant played an active or static game
- **console**: The games console used (Nineto Switch or Xbox Kinect)
- **injury**: Injury severity (a score ranging from 0 (no injury) to 20 (severe injury))

Source

www.discovr.rocks/csv/xbox.csv

zhang_sample *Zhang et al. (2013) (subsample)*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
zhang_sample
```

Format

A tibble with 52 rows and 4 variables

Details

Statistics and maths anxiety are common and affect people's performance on maths and stats assignments; women in particular can lack confidence in mathematics (Field, 2010). Zhang, Schmader and Hall (2013) did an intriguing study in which students completed a maths test in which some put their own name on the test booklet, whereas others were given a booklet that already had either a male or female name on. Participants in the latter two conditions were told that they would use this other person's name for the purpose of the test. Women who completed the test using a different name performed significantly better than those who completed the test using their own name. (There were no such significant effects for men.) The data are a random subsample of Zhang et al.'s data with the following variables:

- **id**: participant ID
- **sex**: participant's biological sex
- **name_type**: the booklet condition to which the participant was allocated: Female fake name, Male fake name or Own name
- **accuracy**: the participant's score on the maths test

Source

www.discovr.rocks/csv/zhang_2013_subsample.csv

References

- Field, A. P. (2010). Teaching Statistics. In D. Upton & A. Trapp (Eds.), *Teaching Psychology in Higher Education* (pp. 134-163). Chichester, UK: Wiley-Blackwell.
- Zhang, S., Schmader, T., & Hall, W. M. (2013). L'eggo My Ego: Reducing the Gender Gap in Math by Unlinking the Self from Performance. *Self and Identity*, *12*, 400-412. doi:10.1080/15298868.2012.687012

`zibarras_2008` *Zibarras et al. (2008) data*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
zibarras_2008
```

Format

A tibble with 207 rows and 12 variables.

Details

Zibarras, Port, and Woods (2008) looked at the relationship between personality and creativity. They used the Hogan Development Survey (HDS), which measures 11 dysfunctional dispositions of employed adults: being *volatile*, *mistrustful*, *cautious*, *detached*, *passive_aggressive*, *arrogant*, *manipulative*, *dramatic*, *eccentric*, *perfectionist*, and *dependent*.

- **id**: The participant id
- **volatile**: responses to the question items of the HDS relating to the *volatile* disposition.
- **mistrustful**: responses to the question items of the HDS relating to the *mistrustful* disposition.
- **cautious**: responses to the question items of the HDS relating to the *cautious* disposition.
- **detached**: responses to the question items of the HDS relating to the *detached* disposition.
- **passive_aggressive**: responses to the question items of the HDS relating to the *passive_aggressive* disposition.
- **arrogant**: responses to the question items of the HDS relating to the *arrogant* disposition.
- **manipulative**: responses to the question items of the HDS relating to the *manipulative* disposition.
- **dramatic**: responses to the question items of the HDS relating to the *dramatic* disposition.
- **eccentric**: responses to the question items of the HDS relating to the *eccentric* disposition.
- **perfectist**: responses (1-5) to the question *I have said to myself 'just a few more minutes on the Internet.'*
- **dependent**: responses (1-5) to the question *I find myself accessing more information on the Internet that I had planned to.*

Source

www.discovr.rocks/csv/zibarras_2008.csv

References

- Zibarras, L. D., Port, R. L., & Woods, S. A. (2008). Innovation and the 'dark side' of personality: Dysfunctional traits and their relation to self-reported innovative characteristics. *Journal of Creative Behavior*, 42, 201–215. doi:10.1002/j.21626057.2008.tb01295.x

`zombie_growth` *Zombie growth model*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
zombie_growth
```

Format

A tibble with 564 rows and 5 variables.

Details

In the story within Field (2016) a lot of people get turned into zombies. At the end of the book it is revealed that one of the central characters, Alice, uses a gene therapy that she invented to restore the zombies back to a human state. This dataset relates to her second study in which she tracked efficacy over 12 months after the treatment. The contains measures from 141 zombies measured at four timepoints (baseline and 1, 6, and 12 month follow-up). Zombies were randomly assigned to two arms of the trial (wait list vs. gene therapy) and the outcome was how much they resembled their pre-zombie state (as a percentage).

- **id**: The zombie's id
- **intervention**: a factor that codes which arm of the trial the participant was randomized to (wait list or gene therapy).
- **time**: categorical variable indicating at which phase of the trial resemblance was measured (Baseline, 1 month, 6 months or 12 months).
- **resemblance**: How closely their face resembled their pre-zombified state (100\
- **time_num**: numerical variable indicating how many months since the intervention resemblance was measured.

Source

www.discovr.rocks/csv/zombie_growth.csv

References

- Field, A. P. (2016). *An adventure in statistics: the reality enigma*. London: Sage. https://www.statisticsadventure.com

`zombie_rehab` *Zombie rehab*

Description

A dataset from Field, A. P. (2026). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Usage

```
zombie_rehab
```

Format

A tibble with 190 rows and 6 variables.

Details

In the story within Field (2016) a lot of people get turned into zombies. At the end of the book it is revealed that one of the central characters, Alice, uses a gene therapy that she invented to restore the zombies back to a human state. This dataset relates to her first attempt at an efficacious gene therapy. It contains data from 190 zombies treated at 10 different clinics. Zombies were randomly assigned to two arms of the trial (wait list vs. gene therapy) and the outcome was how much they resembled their pre-zombie state (as a percentage).

- **p_id**: The zombie's id.
- **clinic_id**: id for the clinic attended anonymised as Clinic 1 to Clinic 10.
- **intervention**: a factor that codes which arm of the trial the participant was randomized to (wait list or gene therapy).
- **resemblance**: How closely their face resembled their pre-zombified state (100\
- **zombification**: whether the initial zombification was achieved through low- or high-intensity zombification.
- **months_as_zombie**: the time (in months) that the person had spend in a zombified state before starting the intervention.

Source

www.discovr.rocks/csv/zombie_rehab.csv

References

- Field, A. P. (2016). *An adventure in statistics: the reality enigma*. London: Sage. https://www.statisticsadventure.com

Index