Data Wrangling and Visualization 101 in R - Going through example data sets

T. Florian Jaeger

Summer 2023

Contents

1 This document

This document is part of a series of lectures that introduce students to basic data wrangling, plotting, and general approaches to data analysis. Please see the overview lecture. This and all other lectures are created in R markdown. R markdown combines text with R code, allowing us to see the code and its output, embedded within the text describing the code. If you have the original R markdown file (file extension .Rmd), you can 'knit' the document into an HMTL, PDF, or Word file.

2 Applying these methods to your own data

2.1 Importing your data

2.1.1 DeAngelis & Haefner group

```
##  experiment subject  trialNumber      flowCondition apertureSize
##  1:365      1:243   Min.   :  1.00   Control :188   0:164
##  2:312      2:215   1st Qu.: 18.00   Full    :166   1:148
##             3:219   Median : 38.00   Global  : 86   2:187
##             4:  0   Mean   : 44.64   Local   : 95   3:178
##                     3rd Qu.: 69.00   Opposite: 71
##                     Max.   :131.00   Same    : 71
##
##  probeEccentricity probeAngle   sceneIndex   relativeTilt      absoluteTilt
##  0:338             -15:213    3      :188   Min.   :-26.729   Min.   :-22.587
##  1:339             0  :243    6      : 95   1st Qu.:  0.000   1st Qu.:  0.000
##                    15 :221    4      : 88   Median :  4.901   Median : 10.252
##                               5      : 86   Mean   :  7.177   Mean   :  7.354
##                               7      : 78   3rd Qu.: 15.000   3rd Qu.: 16.600
##                               8      : 71   Max.   : 44.264   Max.   : 42.584
##                               (Other): 71
##   reactionTime       stimulusTime   probeVelX            probeVelY
##  Min.   :  0.0680   Min.   :2     Min.   :-0.0517638   Min.   :-0.2000
##  1st Qu.:  0.9714   1st Qu.:2     1st Qu.:-0.0517638   1st Qu.:-0.2000
##  Median :  1.2694   Median :2     Median : 0.0000000   Median :-0.1932
##  Mean   :  2.5943   Mean   :2     Mean   :-0.0006117   Mean   :-0.1956
##  3rd Qu.:  2.1624   3rd Qu.:2     3rd Qu.: 0.0517638   3rd Qu.:-0.1932
##  Max.   :361.6860   Max.   :2     Max.   : 0.0517638   Max.   :-0.1932
##
##  probeStartLocationX probeStartLocationY probeEndLocationX probeEndLocationY
##  Min.   :1.000       Min.   :0           Min.   :0.9482   Min.   :-0.20000
##  1st Qu.:1.000       1st Qu.:0           1st Qu.:1.0000   1st Qu.:-0.20000
##  Median :1.000       Median :0           Median :1.0000   Median :-0.19319
##  Mean   :1.151       Mean   :0           Mean   :1.1351   Mean   :-0.06728
##  3rd Qu.:1.500       3rd Qu.:0           3rd Qu.:1.4482   3rd Qu.: 0.20000
##  Max.   :1.500       Max.   :0           Max.   :1.5000   Max.   : 0.20000
##
```

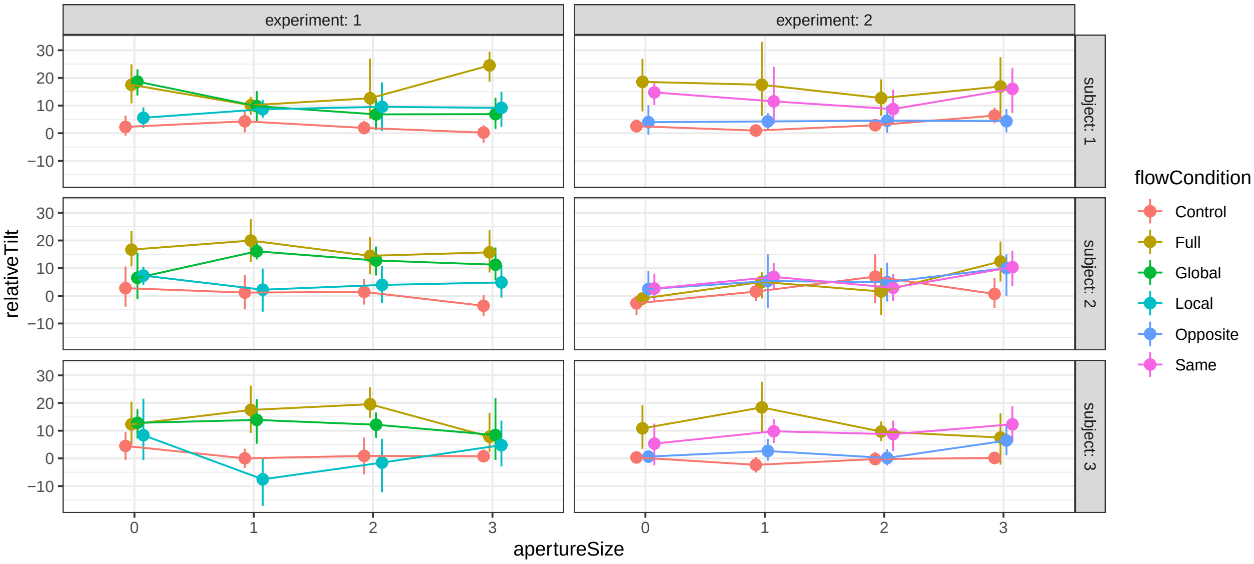
2.1.2 Jaeger group

2.2 Plotting your data

2.2.1 DeAngelis & Haefner group

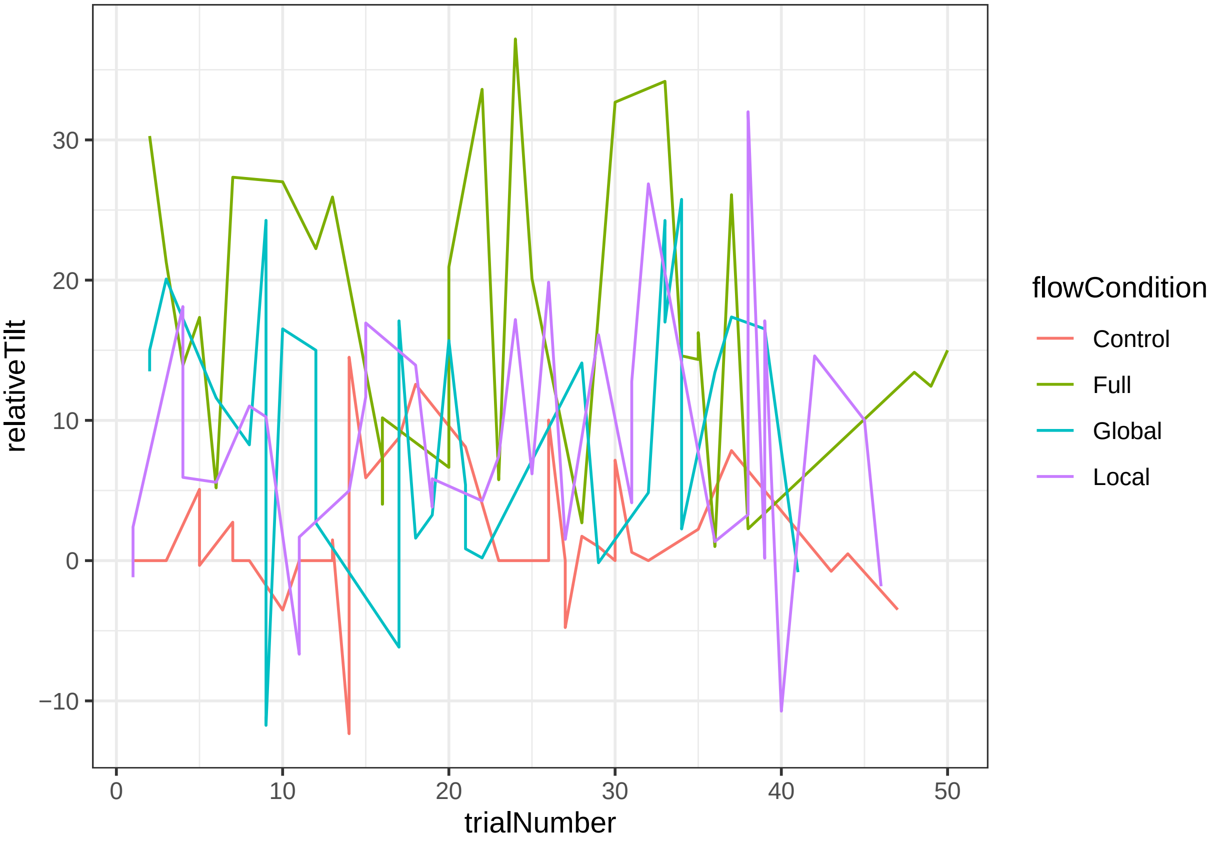
2.2.1.1 Trial exclusions Are there any criteria that would make you think that a trial should be excluded from further analysis?

2.2.1.2 Relative tilt Here is a plot for all subjects that builds on what Ji-ze posted on Slack. **How would you go about changing the titles for the two axes and the condition legend?** Hint: use scales! The R primer I mentioned in the tutorial for the first class talks about scales, too, if you prefer an introduction to reading help files.



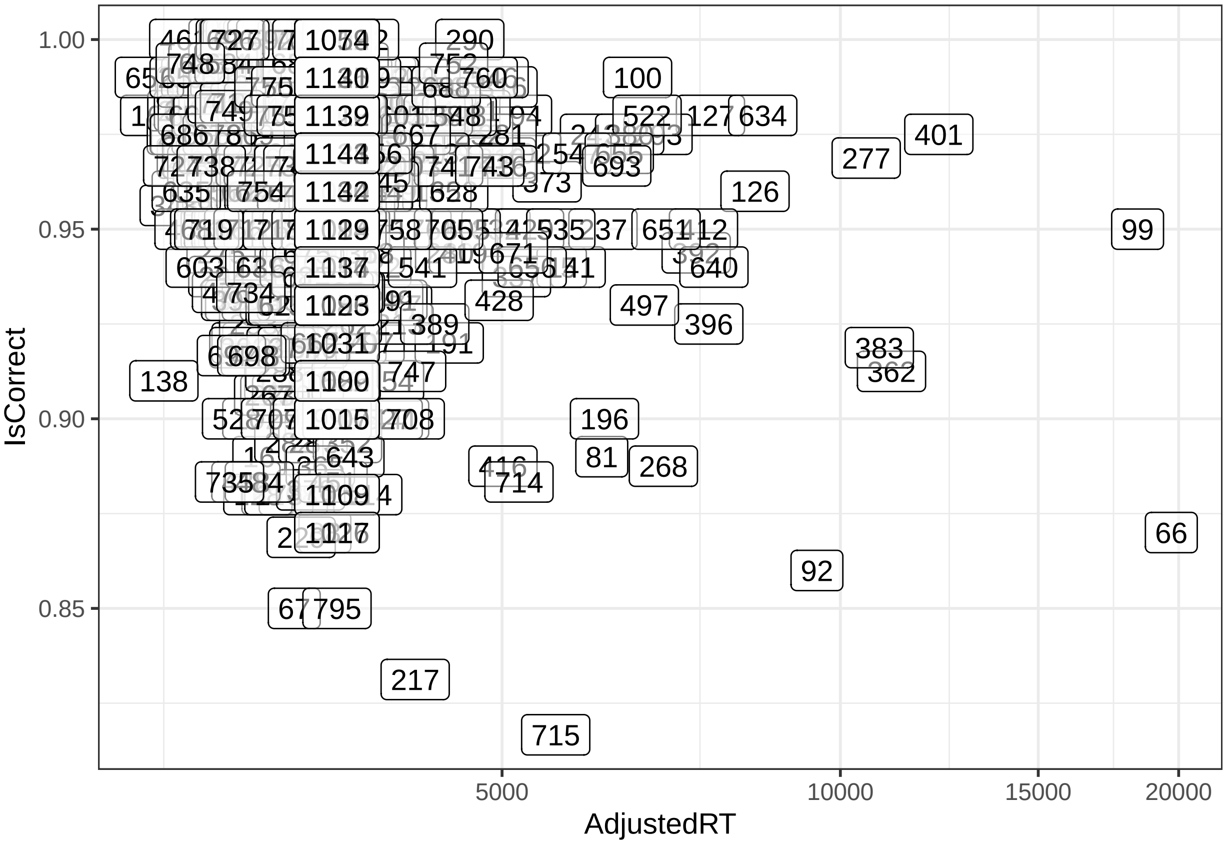
Now let's zoom in on Subject 2 in Experiment 1. Plot the relative tilt as a function of the aperature size and condition, as in the above plot (but only for that subject). **Now think about how you can further show the data separated by probe's angle (probeAngle, 3 values) and eccentricity (probeEccentricity, 2 values).** Hint: you might use shape, transparency (alpha), or faceting to show the data split by additional variables.

2.2.1.3 Plotting changes across trials Some of you also plotted changes across trials. Arya, for example, looked at changes in RTs across trials. Try to plot changes in the relative tilt effect across trials. Here's a plot for Subject 1 in Experiment 1. **How would you combine the data from subjects 1-3 from Experiment 1 and then plot an average for those subjects?** Hint: you don't need to do any manual averaging. Look into geom_smooth, which let's you plot trend lines.



2.2.2 Jaeger group

2.2.2.1 Plotting all subjects' performance during exposure We are getting the average reaction time (AdjustedRT) and accuracy (IsCorrect) of each subject and visualize the distribution of subjects with regard to these two variables. Note that I'm using a log-transformed coordinate system on the x-axis since some RTs can be very high. **What would you conclude from this plot? Should you exclude subjects if they are very slow or fast? Should you look into whether *some* of their trials are very very slow or fast? What would be a good exclusion criterion (if any) based on RTs? Do you think all subjects performed with sufficiently high accuracy to be included in the analysis?**

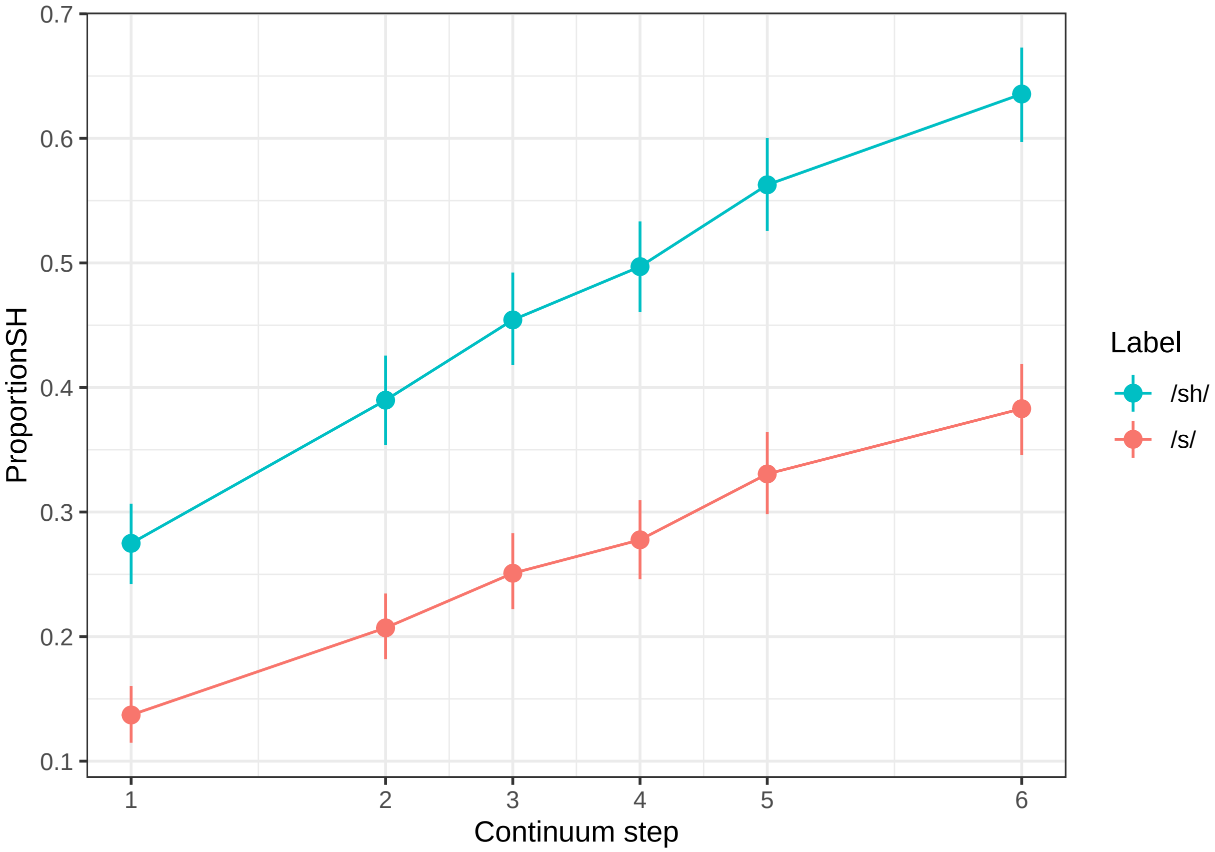


Only after you've settled on an exclusion criterion, try to modify this graph: Specifically, how would you modify this plot to a) only show Experiment 1 and b) color subjects based on the label condition and the pen condition? Hint: labels have both fill (aesthetic fill) and border color (aesthetic color). You can use these two visual means to express the label and pen condition.

How might you additionally indicate in this plot which subjects your exclusion criteria leave for analysis? Hint: you could use a geom_rect to draw a semi-transparent rectangle or you could use geom_segment, geom_hline, and/or geom_vline to draw borders in the RT-by-accuracy space that to indicate your exclusion criteria. Look up those geoms.

2.2.2.2 Plotting all subjects' performance during test Here's the basic plot for the test data. Note that we first summarized the data down to the subject level—i.e., one data point per subject, label condition, and subject. We then get the 95% bootstrapped CIs over those *by-subject means*. This is avoiding overly confident (small) CIs that would result from us failing to acknowledge that repeated measures taken from the same subject are *not* independent of each other. This is different from the motion group, since they are plotting their data separately by subject.

```
## `summarise()` has grouped output by 'Subject', 'Label'. You can override using
## the `.groups` argument.
```



How would you change the y-axis label? How would you make sure that the y-axis actually goes from 0 to 0? Hint: both things can be changed through a scale component. **And, how would you add information about the pen-in-the-hand vs. pen-in-the-mouth condition to this plot?** Hint: shape and linetype provide you with additional visual means, as does faceting.

2.2.2.3 Relating subjects' performance during exposure and test

If you figured out the above, go ahead and try to plot the boundary shift during test against the proportion of shifted words that were rated as words. This will require *combining* the exposure and test data. You will first have to aggregate (summarise) each of the two data down to the by-subject level, and then you can *join* the two data frames (look up ?left_join).

3 Session info

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.3.0 (2023-04-21)
##  os       macOS Ventura 13.4.1
##  system   aarch64, darwin20
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       Europe/Stockholm
##  date     2023-08-19
##  pandoc   3.1.1 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (via rmarkdown)
##
```

```
## - Packages -------------------------------------------------------------------
##  package     * version date (UTC) lib source
##  backports     1.4.1   2021-12-13 [1] CRAN (R 4.3.0)
##  base64enc     0.1-3   2015-07-28 [1] CRAN (R 4.3.0)
##  bit           4.0.5   2022-11-15 [1] CRAN (R 4.3.0)
##  bit64         4.0.5   2020-08-30 [1] CRAN (R 4.3.0)
##  cachem        1.0.8   2023-05-01 [1] CRAN (R 4.3.0)
##  callr         3.7.3   2022-11-02 [1] CRAN (R 4.3.0)
##  cellranger    1.1.0   2016-07-27 [1] CRAN (R 4.3.0)
##  checkmate     2.2.0   2023-04-27 [1] CRAN (R 4.3.0)
##  cli           3.6.1   2023-03-23 [1] CRAN (R 4.3.0)
##  cluster       2.1.4   2022-08-22 [1] CRAN (R 4.3.0)
##  colorspace    2.1-0   2023-01-23 [1] CRAN (R 4.3.0)
##  cowplot     * 1.1.1   2020-12-30 [1] CRAN (R 4.3.0)
##  crayon        1.5.2   2022-09-29 [1] CRAN (R 4.3.0)
##  data.table    1.14.8  2023-02-17 [1] CRAN (R 4.3.0)
##  devtools      2.4.5   2022-10-11 [1] CRAN (R 4.3.0)
##  digest        0.6.33  2023-07-07 [1] CRAN (R 4.3.0)
##  dplyr       * 1.1.2   2023-04-20 [1] CRAN (R 4.3.0)
##  ellipsis      0.3.2   2021-04-29 [1] CRAN (R 4.3.0)
##  evaluate      0.21    2023-05-05 [1] CRAN (R 4.3.0)
##  fansi         1.0.4   2023-01-22 [1] CRAN (R 4.3.0)
##  farver        2.1.1   2022-07-06 [1] CRAN (R 4.3.0)
##  fastmap       1.1.1   2023-02-24 [1] CRAN (R 4.3.0)
##  forcats     * 1.0.0   2023-01-29 [1] CRAN (R 4.3.0)
##  foreign       0.8-84  2022-12-06 [1] CRAN (R 4.3.0)
##  Formula       1.2-5   2023-02-24 [1] CRAN (R 4.3.0)
##  fs            1.6.2   2023-04-25 [1] CRAN (R 4.3.0)
##  generics      0.1.3   2022-07-05 [1] CRAN (R 4.3.0)
##  ggplot2     * 3.4.2   2023-04-03 [1] CRAN (R 4.3.0)
##  glue          1.6.2   2022-02-24 [1] CRAN (R 4.3.0)
##  gridExtra     2.3     2017-09-09 [1] CRAN (R 4.3.0)
##  gtable        0.3.3   2023-03-21 [1] CRAN (R 4.3.0)
##  Hmisc         5.1-0   2023-05-08 [1] CRAN (R 4.3.0)
##  hms           1.1.3   2023-03-21 [1] CRAN (R 4.3.0)
##  htmlTable     2.4.1   2022-07-07 [1] CRAN (R 4.3.0)
##  htmltools     0.5.5   2023-03-23 [1] CRAN (R 4.3.0)
##  htmlwidgets   1.6.2   2023-03-17 [1] CRAN (R 4.3.0)
##  httpuv        1.6.11  2023-05-11 [1] CRAN (R 4.3.0)
##  httr          1.4.6   2023-05-08 [1] CRAN (R 4.3.0)
##  jsonlite      1.8.7   2023-06-29 [1] CRAN (R 4.3.0)
##  knitr         1.43    2023-05-25 [1] CRAN (R 4.3.0)
##  labeling      0.4.2   2020-10-20 [1] CRAN (R 4.3.0)
##  later         1.3.1   2023-05-02 [1] CRAN (R 4.3.0)
##  lazyeval      0.2.2   2019-03-15 [1] CRAN (R 4.3.0)
##  lifecycle     1.0.3   2022-10-07 [1] CRAN (R 4.3.0)
##  lubridate   * 1.9.2   2023-02-10 [1] CRAN (R 4.3.0)
##  magrittr    * 2.0.3   2022-03-30 [1] CRAN (R 4.3.0)
##  memoise       2.0.1   2021-11-26 [1] CRAN (R 4.3.0)
##  mime          0.12    2021-09-28 [1] CRAN (R 4.3.0)
##  miniUI        0.1.1.1 2018-05-18 [1] CRAN (R 4.3.0)
##  munsell       0.5.0   2018-06-12 [1] CRAN (R 4.3.0)
##  nnet          7.3-19  2023-05-03 [1] CRAN (R 4.3.0)
##  pillar        1.9.0   2023-03-22 [1] CRAN (R 4.3.0)
##  pkgbuild      1.4.2   2023-06-26 [1] CRAN (R 4.3.0)
##  pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.3.0)
##  pkgload       1.3.2.1 2023-07-08 [1] CRAN (R 4.3.0)
##  plotly      * 4.10.2  2023-06-03 [1] CRAN (R 4.3.0)
##  prettyunits   1.1.1   2020-01-24 [1] CRAN (R 4.3.0)
##  processx      3.8.2   2023-06-30 [1] CRAN (R 4.3.0)
##  profvis       0.3.8   2023-05-02 [1] CRAN (R 4.3.0)
##  promises      1.2.0.1 2021-02-11 [1] CRAN (R 4.3.0)
##  ps            1.7.5   2023-04-18 [1] CRAN (R 4.3.0)
##  purrr       * 1.0.1   2023-01-10 [1] CRAN (R 4.3.0)
##  R.matlab    * 3.7.0   2022-08-25 [1] CRAN (R 4.3.0)
##  R.methodsS3   1.8.2   2022-06-13 [1] CRAN (R 4.3.0)
##  R.oo          1.25.0  2022-06-12 [1] CRAN (R 4.3.0)
##  R.utils       2.12.2  2022-11-11 [1] CRAN (R 4.3.0)
##  R6            2.5.1   2021-08-19 [1] CRAN (R 4.3.0)
##  Rcpp          1.0.11  2023-07-06 [1] CRAN (R 4.3.0)
##  readr       * 2.1.4   2023-02-10 [1] CRAN (R 4.3.0)
##  readxl      * 1.4.3   2023-07-06 [1] CRAN (R 4.3.0)
##  remotes       2.4.2   2021-11-30 [1] CRAN (R 4.3.0)
##  rlang         1.1.1   2023-04-28 [1] CRAN (R 4.3.0)
##  rmarkdown     2.23    2023-07-01 [1] CRAN (R 4.3.0)
##  rpart         4.1.19  2022-10-21 [1] CRAN (R 4.3.0)
##  rstudioapi    0.15.0  2023-07-07 [1] CRAN (R 4.3.0)
##  scales        1.2.1   2022-08-20 [1] CRAN (R 4.3.0)
##  sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.3.0)
##  shiny         1.7.4.1 2023-07-06 [1] CRAN (R 4.3.0)
##  stringi       1.7.12  2023-01-11 [1] CRAN (R 4.3.0)
##  stringr     * 1.5.0   2022-12-02 [1] CRAN (R 4.3.0)
##  tibble      * 3.2.1   2023-03-20 [1] CRAN (R 4.3.0)
##  tidyr       * 1.3.0   2023-01-24 [1] CRAN (R 4.3.0)
##  tidyselect    1.2.0   2022-10-10 [1] CRAN (R 4.3.0)
##  tidyverse   * 2.0.0   2023-02-22 [1] CRAN (R 4.3.0)
##  timechange    0.2.0   2023-01-11 [1] CRAN (R 4.3.0)
##  tzdb          0.4.0   2023-05-12 [1] CRAN (R 4.3.0)
##  urlchecker    1.0.1   2021-11-30 [1] CRAN (R 4.3.0)
##  usethis       2.2.2   2023-07-06 [1] CRAN (R 4.3.0)
##  utf8          1.2.3   2023-01-31 [1] CRAN (R 4.3.0)
##  vctrs         0.6.3   2023-06-14 [1] CRAN (R 4.3.0)
##  viridisLite   0.4.2   2023-05-02 [1] CRAN (R 4.3.0)
##  vroom         1.6.3   2023-04-28 [1] CRAN (R 4.3.0)
##  withr         2.5.0   2022-03-03 [1] CRAN (R 4.3.0)
##  xfun          0.39    2023-04-20 [1] CRAN (R 4.3.0)
##  xtable        1.8-4   2019-04-21 [1] CRAN (R 4.3.0)
##  yaml          2.3.7   2023-01-23 [1] CRAN (R 4.3.0)
##
##  [1] /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/library
##
## ------------------------------------------------------------------------------
```