DATA IMPORT EXAMPLE

Example of how I import excel data

This work entitled

Data import example Example of how I import excel data

was compiled by

Márton Kiss MD

This document has been meticulously compiled by the author, who assures the application of the finest methodologies and the most comprehensive professional knowledge available at the time of writing. The author guarantees that every effort has been made to ensure the accuracy and reliability of the information contained within, reflecting a rigorous approach to research, analysis, and attention to detail.

Márton Kiss MD, Applied Biostatistician 4242.10.10

Chapter 1 Summary

When handling the good ol' source data if given in excel I like to import it as detailed here. With a second excel sheet with labels/colnames to my liking. I find this useful in the long run if the data has more than 3-5 columns as this lets me quickly adapt to new iterations by the client (eg. 'I just inserted a column…'). Enjoy.

Chapter 2 Data import

When the document is long, I like to put this section in a separate *.r* file and *source()* it in the main document. It basically:

- Reads the 'labels' as I typed them in the descriprion excel sheet and attaches them to the dataset
- transforms the columns flagged in the description as appropriate

```r
descriptor <-
  file.path(data_directory,"description.xlsx") %>%
    readxl::read_excel(skip = 0)

# make a list to be used as labels for 'labelled::'
labs <-
  # take a 1xN matrix
  matrix( descriptor$description,
          ncol = length(descriptor$description)
        ) %>%
 as.list() %>%
 # add column names as 'names'
 `names<-`(descriptor$name_new)

# Making the actual dataset
data <- file.path(data_directory,"Iris.xls") %>%
  readxl::read_excel( ., skip = 0) %>%
                  mutate( across( .cols = which( descriptor$trf ==
        "factor"),
                                  .fns = as.factor
                                  ),
                           across( .cols = which( descriptor$trf == "numeric"),
                                  .fns = as.numeric # removing potential '?',
        'NA', '.' etc.
                                  ),
                           across( .cols = which( descriptor$trf == "date"),
                                  .fns = as_date # works only if excel
        recognizes the date..(?)
                                  )
                         ) %>%
  `colnames<-`( descriptor$name_new) %>%
  labelled::`var_label<-`(   labs  ) %>%
  mutate(
    # making up a new variable as an example
    newvar = petal_width + sepal_width
    ) %>%
  # This is how to set up the labels for new variables
  `var_label<-`( list(newvar = "This is my new example variable, adding
        up petal and sepal width")) %>%
  # One of these days I'm gonna learn when to use this
  rowwise() %>%
  mutate(
    mock_ID = runif(1, .1, 20) %>% round %>% as.integer
  ) %>%
  set_variable_labels( mock_ID = "Mock ID")
```

The data should be read properly, doing nothing fancy just reporting the first 20 lines after switching the labels back (maybe there is another better way…?)

Numeric representation of species	These are the width of the petals	These are the length of the petals	These are the width of the sepals	These are the length of the sepals	Character representation of the species	This is a date column to illustrate transformations	This is my new example variable, adding up petal and sepal width	Mock ID
1	0.20	1.40	3.50	5.10	Setosa	2022-01-01	3.70	9
1	0.20	1.40	3.00	4.90	Setosa	2022-01-02	3.20	18
1	0.20	1.30	3.20	4.70	Setosa	2022-01-03	3.40	11
1	0.20	1.50	3.10	4.60	Setosa	2022-01-04	3.30	17
1	0.20	1.40	3.60	5.00	Setosa	2022-01-05	3.80	15
1	0.40	1.70	3.90	5.40	Setosa	2022-01-06	4.30	19
1	0.30	1.40	3.40	4.60	Setosa	2022-01-07	3.70	9
1	0.20	1.50	3.40	5.00	Setosa	2022-01-08	3.60	16
1	0.20	1.40	2.90	4.40	Setosa	2022-01-09	3.10	12
1	0.10	1.50	3.10	4.90	Setosa	2022-01-10	3.20	7
1	0.20	1.50	3.70	5.40	Setosa	2022-01-11	3.90	12
1	0.20	1.60	3.40	4.80	Setosa	2022-01-12	3.60	19
1	0.10	1.40	3.00	4.80	Setosa	2022-01-13	3.10	16
1	0.10	1.10	3.00	4.30	Setosa	2022-01-14	3.10	15
1	0.20	1.20	4.00	5.80	Setosa	2022-01-15	4.20	8
1	0.40	1.50	4.40	5.70	Setosa	2022-01-16	4.80	8
1	0.40	1.30	3.90	5.40	Setosa	2022-01-17	4.30	6
1	0.30	1.40	3.50	5.10	Setosa	2022-01-18	3.80	14
1	0.30	1.70	3.80	5.70	Setosa	2022-01-19	4.10	10
1	0.30	1.50	3.80	5.10	Setosa	2022-01-20	4.10	8

3 Notes

Information regarding the compilation of this document:

```
knitr::opts_chunk$set(comment = NA)
sessionInfo() %>%
    report::report() %>%
    cat()
```

Analyses were conducted using the R Statistical language (version 4.3.0; R Core Team, 2023) on Windows 10 x64 (build 19045), using the packages rmarkdown (version 2.23; Allaire J et al., 2023), effects (version 4.2.2; Fox J, Weisberg S, 2019), carData (version 3.0.5; Fox J et al., 2022), flextable (version 0.9.2; Gohel D, Skintzos P, 2023), lubridate (version 1.9.2; Grolemund G, Wickham H, 2011), DHARMa (version 0.4.6; Hartig F, 2022), huxtable (version 5.5.2; Hugh-Jones D, 2022), labelled (version 2.12.0; Larmarange J, 2023), emmeans (version 1.8.7; Lenth R, 2023), nlme (version 3.1.162; Pinheiro J et al., 2023), gtsummary (version 1.7.1; Sjoberg D et al., 2021), ggplot2 (version 3.4.2; Wickham H, 2016), readxl (version 1.4.2; Wickham H, Bryan J, 2023), roxygen2 (version 7.2.3; Wickham H et al., 2022), dplyr (version 1.1.2; Wickham H et al., 2023), knitr (version 1.43; Xie Y, 2023), pagedown (version 0.20; Xie Y et al., 2022) and kableExtra (version 1.3.4.9000; Zhu H, 2023).

3.1 References

- Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2023). *rmarkdown: Dynamic Documents for R*. R package version 2.23, https://github.com/rstudio/rmarkdown. Xie Y, Allaire J, Grolemund G (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338, https://bookdown.org/yihui/rmarkdown. Xie Y, Dervieux C, Riederer E (2020). *R Markdown Cookbook*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9780367563837, https://bookdown.org/yihui/rmarkdown-cookbook.
- Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, 3rd edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html. Fox J, Weisberg S (2018). “Visualizing Fit and Lack

of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals." *Journal of Statistical Software*, *87*(9), 1-27. doi:10.18637/jss.v087.i09https://doi.org/10.18637/jss.v087.i09. Fox J (2003). "Effect Displays in R for Generalised Linear Models." *Journal of Statistical Software*, *8*(15), 1-27. doi:10.18637/jss.v008.i15https://doi.org/10.18637/jss.v008.i15. Fox J, Hong J (2009). "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package." *Journal of Statistical Software*, *32*(1), 1-24. doi:10.18637/jss.v032.i01https://doi.org/10.18637/jss.v032.i01.

- Fox J, Weisberg S, Price B (2022). *carData: Companion to Applied Regression Data Sets*. R package version 3.0-5, https://CRAN.R-project.org/package=carData.
- Gohel D, Skintzos P (2023). *flextable: Functions for Tabular Reporting*. R package version 0.9.2, https://CRAN.R-project.org/package=flextable.
- Grolemund G, Wickham H (2011). "Dates and Times Made Easy with lubridate." *Journal of Statistical Software*, *40*(3), 1-25. https://www.jstatsoft.org/v40/i03/.
- Hartig F (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.6, https://CRAN.R-project.org/package=DHARMa.
- Hugh-Jones D (2022). *huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats*. R package version 5.5.2, https://CRAN.R-project.org/package=huxtable.
- Larmarange J (2023). *labelled: Manipulating Labelled Data*. R package version 2.12.0, https://CRAN.R-project.org/package=labelled.
- Lenth R (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.7, https://CRAN.R-project.org/package=emmeans.
- Pinheiro J, Bates D, R Core Team (2023). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-162, https://CRAN.R-project.org/package=nlme. Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and*

S-PLUS. Springer, New York. doi:10.1007/b98882https://doi.org/10.1007/b98882.

- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
- Sjoberg D, Whiting K, Curry M, Lavery J, Larmarange J (2021). "Reproducible Summary Tables with the gtsummary Package." *The R Journal*, *13*, 570-580. doi:10.32614/RJ-2021-053 https://doi.org/10.32614/RJ-2021-053, https://doi.org/10.32614/RJ-2021-053.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.
- Wickham H, Bryan J (2023). *readxl: Read Excel Files*. R package version 1.4.2, https://CRAN.R-project.org/package=readxl.
- Wickham H, Danenberg P, Csárdi G, Eugster M (2022). *roxygen2: In-Line Documentation for R*. R package version 7.2.3, https://CRAN.R-project.org/package=roxygen2.
- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.2, https://CRAN.R-project.org/package=dplyr.
- Xie Y (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.43, https://yihui.org/knitr/. Xie Y (2015). *Dynamic Documents with R and knitr*, 2nd edition. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1498716963, https://yihui.org/knitr/. Xie Y (2014). "knitr: A Comprehensive Tool for Reproducible Research in R." In Stodden V, Leisch F, Peng RD (eds.), *Implementing Reproducible Computational Research*. Chapman and Hall/CRC. ISBN 978-1466561595.
- Xie Y, Lesur R, Thorne B, Tan X (2022). *pagedown: Paginate the HTML Output of R Markdown with CSS for Print*. R package version 0.20, https://CRAN.R-project.org/package=pagedown.
- Zhu H (2023). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/, https://github.com

/haozhu233/kableExtra.

This document was compiled at:

```
Sys.time()
```

```
[1] "2023-07-06 17:35:31 CEST"
```

```
save.image(file = here::here("inst", "states", "dat_import_out.Rdata"))
```