Data Cleaning

Robert J. Dellinger

April 05, 2025

Contents

1 Introduction

This document outlines the methodology for data cleaning, exploration, and visualization. It is structured to ensure transparency and reproducibility of all analyses.

1.1 Methodology

Briefly describe the methods used in the project, including data sources, cleaning steps, and techniques applied to handle missing or inconsistent data.

1.2 Loading Data

The first step in any data analysis is to load the data. This section outlines the process of importing the data into R, including any necessary transformations or adjustments to ensure compatibility with the analysis.

```
# Example of loading data from a CSV file
# raw_data <- read_csv(here("Data", "Raw", "data_file.csv"))
# raw_data <- read_excel(here("Data", "Raw", "data_file.xlsx"))
```

1.3 Cleaning Data

The data cleaning process involves several steps to ensure the data is in a suitable format for analysis. This includes handling missing values, correcting data types, and removing duplicates.

```
# Example of cleaning data
# - Removing duplicates
# - Handling missing values
# - Converting data types and cleaning white space
# - Renaming columns, etc.

# cleaned_data <- raw_data %>%
#   clean_names() %>%
```

```
#   mutate(column_name = as_factor(column_name)) %>%
#   mutate(date_column = as.Date(date_column, format = "%Y-%m-%d")) %>% # convert to date
#   mutate(numeric_column = as.numeric(numeric_column)) %>% # convert to numeric
#   mutate(accross(everything(), ~str_squish(.))) %>% # clean whitespace
#   drop_na()
```

1.4 Data Exploration

Data exploration is a crucial step in understanding the dataset and identifying patterns or anomalies. This section includes summary statistics, visualizations, and any other relevant analyses to gain insights into the data.

```
# Explore the cleaned data using basic summaries:
# glimpse(cleaned_data)
# summary(cleaned_data)
# str(cleaned_data)
```

1.5 Data Visualization

Data visualization is an essential part of data analysis, allowing for the communication of findings in a clear and effective manner. This section includes various plots and charts to illustrate key insights from the data.

```
# Example of creating a summary table
# summary_table <- cleaned_data %>%
#   group_by(group_var) %>%
#   summarise(mean_value = mean(value_var, na.rm = TRUE)) %>%
#   ungroup() %>%
#   kable() %>%
#   kable_styling(full_width = F, position = "left")

# save_kable(summary_table, file = here(output_path_tables, "summary_table.html"),
#   bootstrap_options = c("striped", "hover", "condensed"),
#   full_width = F, position = "left")
```

```
# Example of visualization plot
# ggplot(cleaned_data, aes(x = var1, y = var2)) +
#   geom_point() +
#   theme_minimal()

# Example of saving a plot
# ggsave(filename = here(output_path_images, "plot_name.png"), plot = last_plot(), width = 6, height =
```

2 Literature Cited

Citing the packages and data used in the analysis is important for reproducibility and transparency. The following code generates a bibliography of all loaded packages. Items can be cited directly within the documentation using the syntax `@key` where key is the citation key in the first line of the entry, e.g., R Core Team (2024). To put citations in parentheses, use `[@key]` instead.

nocite: ‘’

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2024. *Rmarkdown: Dynamic Documents for r.* https://github.com/rstudio/rmarkdown.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition.* Springer, NY. https://asdar-book.org/.

Boettiger, Carl. 2021. *Knitcitations: Citations for Knitr Markdown Files*. https://github.com/cboettig/knitcitations.

Csárdi, Gábor, Jim Hester, Hadley Wickham, Winston Chang, Martin Morgan, and Dan Tenenbaum. 2024. *Remotes: R Package Installation from Remote Repositories, Including GitHub*. https://remotes.r-lib.org.

Firke, Sam. 2024. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hester, Jim, and Jennifer Bryan. 2024. *Glue: Interpreted String Literals*. https://glue.tidyverse.org/.

Hijmans, Robert J. 2025. *Raster: Geographic Data Analysis and Modeling*. https://rspatial.org/raster.

Kahle, David, and Hadley Wickham. 2013. "Ggmap: Spatial Visualization with Ggplot2." *The R Journal* 5 (1): 144–61. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Kahle, David, Hadley Wickham, and Scott Jackson. 2023. *Ggmap: Spatial Visualization with Ggplot2*. https://github.com/dkahle/ggmap.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://tibble.tidyverse.org/.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

———. 2025. *Sf: Simple Features for r*. https://r-spatial.github.io/sf/.

Pebesma, Edzer J., and Roger Bivand. 2005. "Classes and Methods for Spatial Data in R." *R News* 5 (2): 9–13. https://CRAN.R-project.org/doc/Rnews/.

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016.

———. 2025. *Sp: Classes and Methods for Spatial Data*. https://github.com/edzer/sp/.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com.

Pedersen, Thomas Lin, and David Robinson. 2025. *Gganimate: A Grammar of Animated Graphics*. https://gganimate.com.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rinker, Tyler W., and Dason Kurkiewicz. 2018. *pacman: Package Management for R*. Buffalo, New York. http://github.com/trinker/pacman.

Rinker, Tyler, and Dason Kurkiewicz. 2019. *Pacman: Package Management Tool*. https://github.com/trinker/pacman.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Sievert, Carson, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, and Pedro Despouy. 2024. *Plotly: Create Interactive Web Graphics via Plotly.js*. https://plotly-r.com.

Slowikowski, Kamil. 2024. *Ggrepel: Automatically Position Non-Overlapping Text Labels with Ggplot2*. https://ggrepel.slowkow.com/.

Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2024. *Lubridate: Make Dealing with Dates a Little Easier*. https://lubridate.tidyverse.org.

Walker, Kyle. 2024. *Tigris: Load Census TIGER/Line Shapefiles*. https://github.com/walkerke/tigris.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023a. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://forcats.tidyverse.org/.

———. 2023b. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

———. 2023c. *Tidyverse: Easily Install and Load the Tidyverse*. https://tidyverse.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2025. *Readxl: Read Excel Files*. https://readxl.tidyverse.org.

Wickham, Hadley, Jennifer Bryan, Malcolm Barrett, and Andy Teucher. 2024. *Usethis: Automate Package*

and Project Setup. https://usethis.r-lib.org.

Wickham, Hadley, Winston Chang, Robert Flight, Kirill Müller, and Jim Hester. 2025. *Sessioninfo: R Session Information*. https://github.com/r-lib/sessioninfo#readme.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and Teun van den Brand. 2024. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, and Lionel Henry. 2025. *Purrr: Functional Programming Tools*. https://purrr.tidyverse.org/.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. *Devtools: Tools to Make Developing r Packages Easier*. https://devtools.r-lib.org/.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization*. https://scales.r-lib.org.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org.

Wilke, Claus O. 2024. *Cowplot: Streamlined Plot Theme and Plot Annotations for Ggplot2*. https://wilkelab.org/cowplot/.

Wilke, Claus O., and Brenton M. Wiernik. 2022. *Ggtext: Improved Text Rendering Support for Ggplot2*. https://wilkelab.org/ggtext/.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2025. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.