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.

Packages cited are Wickham et al. (2024), Wickham (2023), Wickham et al. (2023), Wickham, Vaughan, and Girlich (2024), Firke (2024), Spinu, Grolemund, and Wickham (2024), Hester and Bryan (2024), Wickham and Henry (2025), Müller (2020), Pedersen and Robinson (2025), Slowikowski (2024), Wilke and Wiernik (2022), Wickham, Pedersen, and Seidel (2023), Zhu (2024), Wilke (2024), Pedersen (2024), Sievert et al. (2024), Pebesma (2025), Walker (2024), Hijmans (2025), and Xie (2025).

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

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.

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

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

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/.

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

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

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