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.

(**ggplot2?**)

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