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-base. To put citations in parentheses, use [@key] instead.

```
@Manual{R-base,
  title = {R: A Language and Environment for Statistical Computing},
  author = {{R Core Team}},
  organization = {R Foundation for Statistical Computing},
  address = {Vienna, Austria},
  year = {2024},
  url = {https://www.R-project.org/},
}

@Manual{R-cowplot,
  title = {cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2},
  author = {Claus O. Wilke},
  year = {2024},
  note = {R package version 1.1.3},
  url = {https://wilkelab.org/cowplot/},
}

@Manual{R-devtools,
  title = {devtools: Tools to Make Developing R Packages Easier},
  author = {Hadley Wickham and Jim Hester and Winston Chang and Jennifer Bryan},
  year = {2022},
  note = {R package version 2.4.5},
  url = {https://devtools.r-lib.org/},
}

@Manual{R-dplyr,
  title = {dplyr: A Grammar of Data Manipulation},
  author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis Vaughan},
  year = {2023},
  note = {R package version 1.1.4},
  url = {https://dplyr.tidyverse.org},
}

@Manual{R-forcats,
  title = {forcats: Tools for Working with Categorical Variables (Factors)},
  author = {Hadley Wickham},
  year = {2023},
  note = {R package version 1.0.0},
  url = {https://forcats.tidyverse.org/},
}

@Manual{R-gganimate,
  title = {gganimate: A Grammar of Animated Graphics},
  author = {Thomas Lin Pedersen and David Robinson},
  year = {2025},
  note = {R package version 1.0.9.9000},
  url = {https://gganimate.com},
}

@Manual{R-ggmap,
  title = {ggmap: Spatial Visualization with ggplot2},
  author = {David Kahle and Hadley Wickham and Scott Jackson},
  year = {2023},
  note = {R package version 4.0.0},
  url = {https://github.com/dkahle/ggmap},
}

@Manual{R-ggplot2,
  title = {ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics},
  author = {Hadley Wickham and Winston Chang and Lionel Henry and Thomas Lin Pedersen and Kohske Takahas
  year = {2024},
  note = {R package version 3.5.1},
  url = {https://ggplot2.tidyverse.org},
}

@Manual{R-ggrepel,
  title = {ggrepel: Automatically Position Non-Overlapping Text Labels with
ggplot2},
  author = {Kamil Slowikowski},
  year = {2024},
  note = {R package version 0.9.6},
  url = {https://ggrepel.slowkow.com/},
}

@Manual{R-ggtext,
  title = {ggtext: Improved Text Rendering Support for ggplot2},
  author = {Claus O. Wilke and Brenton M. Wiernik},
  year = {2022},
  note = {R package version 0.1.2},
  url = {https://wilkelab.org/ggtext/},
}

@Manual{R-glue,
  title = {glue: Interpreted String Literals},
  author = {Jim Hester and Jennifer Bryan},
  year = {2024},
  note = {R package version 1.8.0},
  url = {https://glue.tidyverse.org/},
}

@Manual{R-here,
  title = {here: A Simpler Way to Find Your Files},
  author = {Kirill Müller},
  year = {2020},
  note = {R package version 1.0.1},
  url = {https://here.r-lib.org/},
}

@Manual{R-janitor,
  title = {janitor: Simple Tools for Examining and Cleaning Dirty Data},
  author = {Sam Firke},
  year = {2024},
  note = {R package version 2.2.1},
  url = {https://github.com/sfirke/janitor},
}

@Manual{R-kableExtra,
  title = {kableExtra: Construct Complex Table with kable and Pipe Syntax},
  author = {Hao Zhu},
  year = {2024},
  note = {R package version 1.4.0},
  url = {http://haozhu233.github.io/kableExtra/},
}

@Manual{R-knitcitations,
  title = {knitcitations: Citations for Knitr Markdown Files},
  author = {Carl Boettiger},
  year = {2021},
  note = {R package version 1.0.12},
  url = {https://github.com/cboettig/knitcitations},
}

@Manual{R-knitr,
  title = {knitr: A General-Purpose Package for Dynamic Report Generation in R},
  author = {Yihui Xie},
  year = {2025},
  note = {R package version 1.50},
  url = {https://yihui.org/knitr/},
}

@Manual{R-lubridate,
  title = {lubridate: Make Dealing with Dates a Little Easier},
  author = {Vitalie Spinu and Garrett Grolemund and Hadley Wickham},
  year = {2024},
  note = {R package version 1.9.4},
  url = {https://lubridate.tidyverse.org},
}

@Manual{R-pacman,
  title = {pacman: Package Management Tool},
  author = {Tyler Rinker and Dason Kurkiewicz},
  year = {2019},
  note = {R package version 0.5.1},
  url = {https://github.com/trinker/pacman},
}

@Manual{R-patchwork,
  title = {patchwork: The Composer of Plots},
  author = {Thomas Lin Pedersen},
  year = {2024},
  note = {R package version 1.3.0},
  url = {https://patchwork.data-imaginist.com},
}

@Manual{R-plotly,
  title = {plotly: Create Interactive Web Graphics via plotly.js},
  author = {Carson Sievert and Chris Parmer and Toby Hocking and Scott Chamberlain and Karthik Ram and
  year = {2024},
  note = {R package version 4.10.4},
  url = {https://plotly-r.com},
}

@Manual{R-purrr,
  title = {purrr: Functional Programming Tools},
  author = {Hadley Wickham and Lionel Henry},
  year = {2025},
  note = {R package version 1.0.4},
  url = {https://purrr.tidyverse.org/},
}

@Manual{R-raster,
  title = {raster: Geographic Data Analysis and Modeling},
  author = {Robert J. Hijmans},
  year = {2025},
  note = {R package version 3.6-32},
  url = {https://rspatial.org/raster},
}

@Manual{R-readr,
  title = {readr: Read Rectangular Text Data},
  author = {Hadley Wickham and Jim Hester and Jennifer Bryan},
  year = {2024},
  note = {R package version 2.1.5},
  url = {https://readr.tidyverse.org},
}

@Manual{R-readxl,
  title = {readxl: Read Excel Files},
  author = {Hadley Wickham and Jennifer Bryan},
  year = {2025},
  note = {R package version 1.4.5},
  url = {https://readxl.tidyverse.org},
}

@Manual{R-remotes,
  title = {remotes: R Package Installation from Remote Repositories, Including
GitHub},
  author = {Gábor Csárdi and Jim Hester and Hadley Wickham and Winston Chang and Martin Morgan and Dan
  year = {2024},
  note = {R package version 2.5.0},
  url = {https://remotes.r-lib.org},
}

@Manual{R-rmarkdown,
  title = {rmarkdown: Dynamic Documents for R},
  author = {JJ Allaire and Yihui Xie and Christophe Dervieux and Jonathan McPherson and Javier Luraschi
  year = {2024},
  note = {R package version 2.29},
  url = {https://github.com/rstudio/rmarkdown},
}

@Manual{R-scales,
  title = {scales: Scale Functions for Visualization},
  author = {Hadley Wickham and Thomas Lin Pedersen and Dana Seidel},
  year = {2023},
  note = {R package version 1.3.0},
  url = {https://scales.r-lib.org},
}

@Manual{R-sessioninfo,
  title = {sessioninfo: R Session Information},
  author = {Hadley Wickham and Winston Chang and Robert Flight and Kirill Müller and Jim Hester},
  year = {2025},
  note = {R package version 1.2.3},
  url = {https://github.com/r-lib/sessioninfo#readme},
}

@Manual{R-sf,
  title = {sf: Simple Features for R},
  author = {Edzer Pebesma},
  year = {2025},
  note = {R package version 1.0-20},
  url = {https://r-spatial.github.io/sf/},
}

@Manual{R-sp,
  title = {sp: Classes and Methods for Spatial Data},
  author = {Edzer Pebesma and Roger Bivand},
  year = {2025},
  note = {R package version 2.2-0},
  url = {https://github.com/edzer/sp/},
}

@Manual{R-stringr,
  title = {stringr: Simple, Consistent Wrappers for Common String Operations},
  author = {Hadley Wickham},
  year = {2023},
  note = {R package version 1.5.1},
  url = {https://stringr.tidyverse.org},
}

@Manual{R-tibble,
  title = {tibble: Simple Data Frames},
  author = {Kirill Müller and Hadley Wickham},
  year = {2023},
  note = {R package version 3.2.1},
  url = {https://tibble.tidyverse.org/},
}

@Manual{R-tidyr,
  title = {tidyr: Tidy Messy Data},
  author = {Hadley Wickham and Davis Vaughan and Maximilian Girlich},
  year = {2024},
  note = {R package version 1.3.1},
  url = {https://tidyr.tidyverse.org},
}

@Manual{R-tidyverse,
  title = {tidyverse: Easily Install and Load the Tidyverse},
  author = {Hadley Wickham},
  year = {2023},
  note = {R package version 2.0.0},
  url = {https://tidyverse.tidyverse.org},
}

@Manual{R-tigris,
  title = {tigris: Load Census TIGER/Line Shapefiles},
  author = {Kyle Walker},
  year = {2024},
  note = {R package version 2.1},
  url = {https://github.com/walkerke/tigris},
}

@Manual{R-usethis,
  title = {usethis: Automate Package and Project Setup},
  author = {Hadley Wickham and Jennifer Bryan and Malcolm Barrett and Andy Teucher},
  year = {2024},
  note = {R package version 3.1.0},
  url = {https://usethis.r-lib.org},
}

@Article{ggmap2013,
  author = {David Kahle and Hadley Wickham},
  title = {ggmap: Spatial Visualization with ggplot2},
  journal = {The R Journal},
  year = {2013},
  volume = {5},
  number = {1},
  pages = {144--161},
  url = {https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf},
}

@Book{ggplot22016,
  author = {Hadley Wickham},
  title = {ggplot2: Elegant Graphics for Data Analysis},
  publisher = {Springer-Verlag New York},
  year = {2016},
  isbn = {978-3-319-24277-4},
  url = {https://ggplot2.tidyverse.org},
}

@Book{knitr2015,
  title = {Dynamic Documents with {R} and knitr},
  author = {Yihui Xie},
  publisher = {Chapman and Hall/CRC},
  address = {Boca Raton, Florida},
  year = {2015},
  edition = {2nd},
  note = {ISBN 978-1498716963},
  url = {https://yihui.org/knitr/},
}

@InCollection{knitr2014,
  booktitle = {Implementing Reproducible Computational Research},
  editor = {Victoria Stodden and Friedrich Leisch and Roger D. Peng},
  title = {knitr: A Comprehensive Tool for Reproducible Research in {R}},
  author = {Yihui Xie},
  publisher = {Chapman and Hall/CRC},
  year = {2014},
  note = {ISBN 978-1466561595},
}

@Article{lubridate2011,
  title = {Dates and Times Made Easy with {lubridate}},
  author = {Garrett Grolemund and Hadley Wickham},
  journal = {Journal of Statistical Software},
  year = {2011},
  volume = {40},
  number = {3},
  pages = {1--25},
  url = {https://www.jstatsoft.org/v40/i03/},
}

@Manual{pacman2018,
  title = {{pacman}: {P}ackage Management for {R}},
  author = {Tyler W. Rinker and Dason Kurkiewicz},
  address = {Buffalo, New York},
  note = {version 0.5.0},
  year = {2018},
  url = {http://github.com/trinker/pacman},
}

@Book{plotly2020,
  author = {Carson Sievert},
  title = {Interactive Web-Based Data Visualization with R, plotly, and shiny},
  publisher = {Chapman and Hall/CRC},
  year = {2020},
  isbn = {9781138331457},
  url = {https://plotly-r.com},
}

@Book{rmarkdown2018,
  title = {R Markdown: The Definitive Guide},
  author = {Yihui Xie and J.J. Allaire and Garrett Grolemund},
  publisher = {Chapman and Hall/CRC},
  address = {Boca Raton, Florida},
  year = {2018},
  isbn = {9781138359338},
  url = {https://bookdown.org/yihui/rmarkdown},
}

@Book{rmarkdown2020,
  title = {R Markdown Cookbook},
  author = {Yihui Xie and Christophe Dervieux and Emily Riederer},
  publisher = {Chapman and Hall/CRC},
  address = {Boca Raton, Florida},
  year = {2020},
  isbn = {9780367563837},
  url = {https://bookdown.org/yihui/rmarkdown-cookbook},
}

@Book{sf2023,
  author = {Edzer Pebesma and Roger Bivand},
  title = {{Spatial Data Science: With applications in R}},
  year = {2023},
  publisher = {{Chapman and Hall/CRC}},
  url = {https://r-spatial.org/book/},
  doi = {10.1201/9780429459016},
}

@Article{sf2018,
  author = {Edzer Pebesma},
  title = {{Simple Features for R: Standardized Support for Spatial Vector Data}},
  year = {2018},
  journal = {{The R Journal}},
  doi = {10.32614/RJ-2018-009},
  url = {https://doi.org/10.32614/RJ-2018-009},
  pages = {439--446},
  volume = {10},
  number = {1},
}

@Article{sp2005,
  author = {Edzer J. Pebesma and Roger Bivand},
  title = {Classes and methods for spatial data in {R}},
  journal = {R News},
  year = {2005},
  volume = {5},
  number = {2},
  pages = {9--13},
  month = {November},
  url = {https://CRAN.R-project.org/doc/Rnews/},
}

@Book{sp2013,
  author = {Roger S. Bivand and Edzer Pebesma and Virgilio Gomez-Rubio},
  title = {Applied spatial data analysis with {R}, Second edition},
  year = {2013},
  publisher = {Springer, NY},
  url = {https://asdar-book.org/},
}

@Article{tidyverse2019,
  title = {Welcome to the {tidyverse}},
  author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McG
  year = {2019},
  journal = {Journal of Open Source Software},
  volume = {4},
  number = {43},
  pages = {1686},
  doi = {10.21105/joss.01686},
}
```

2.0.1 Data Sources

- Data source 1: [Name of Dataset or Study], [DOI/Link]
- Data source 2: [Name of Dataset or Study], [DOI/Link] # A tibble: 1 x 11 version os system ui language collate ctype tz date pandoc quarto 1 R version~ macO~ aarch~ X11 (EN) en_US.~ en_U~ Amer~ 2025~ 3.2 @~ 1.5.5~