Eassy 5A[*]

Sirui Tan

February 24, 2024

1 Introduction

The data employed in this paper was sourced from the wiki. Data cleaning and analysis were conducted using the open-source statistical programming language R (R Core Team (2022)), leveraging functionalities from the tidyverse (Wickham et al. (2019)) suite, including ggplot2 (Wickham (2016)), dplyr (Wickham et al. (2023)), readr (Wickham, Hester, and Bryan (2024)), tibble (Müller and Wickham (2023)),stringr (Wickham (2023)) ,janitor (Firke (2023)) and knitr (Xie (2023)). The detailed procedures for data extraction and cleaning are expounded upon in the subsequent subsections.

2 Data

Some of our data is of penguins (Figure 1), from Horst, Hill, and Gorman (2020).

```
 [1] "no"                              "portrait"
 [3] "name_birth_death_constituency" "election_parliament"
 [5] "term_of_office"                  "term_of_office_2"
 [7] "term_of_office_3"                "politicalparty"
 [9] "ministry"                        "monarch"
[11] "governor_general"                "ref"
```

```
Warning: There were 2 warnings in `mutate()`.
The first warning was:
i In argument: `Age_at_Death = as.numeric(death) - as.numeric(birth)`.
Caused by warning:
```

[*] Code and data are available at:https://github.com/siru1366/mini-eassy-5B.

```
! NAs introduced by coercion
i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

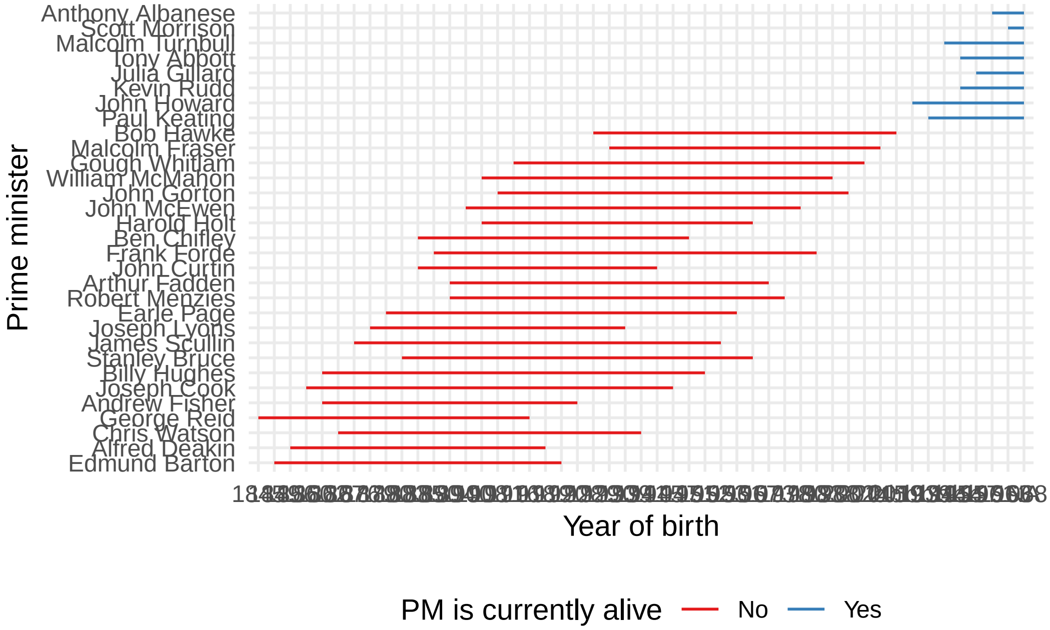

Figure 1: Prime Ministers of the Australia

After gathering and cleaning the data on the prime ministers of Australia, we observed several interesting patterns. The dataset spans from the late 19th century to the late 20th century, covering a significant period of Australian political history. One notable trend is the increasing longevity of prime ministers over time. While early prime ministers such as Edmund Barton and Alfred Deakin had relatively short tenures, later prime ministers like Robert Menzies and John Howard served much longer terms, indicating a potential shift in the political landscape or leadership stability over the years.

Another intriguing observation is the age at which prime ministers assumed office and their life expectancies. While some prime ministers, like John Howard, served well into their older years, others, such as Joseph Lyons, passed away relatively young. Analyzing these age and tenure dynamics can provide insights into the challenges and opportunities faced by political leaders and their impact on governance and policy-making.

Furthermore, the dataset highlights the diversity of backgrounds among Australian prime ministers, including varying birth years and constituencies. Exploring these demographic factors alongside political careers can shed light on the representation and inclusivity within Australian politics throughout history.

The data was sourced from Wikipedia, specifically the page listing the prime ministers of Australia. The process involved web scraping using the rvest package in R to extract relevant information from the webpage. Once extracted, the data was cleaned and transformed using various data wrangling techniques in R, such as splitting columns, extracting numeric values, and handling missing data.

The web scraping process involved identifying the HTML structure of the Wikipedia page and using CSS selectors to target specific elements containing the desired information. This process required careful inspection of the webpage's structure and experimentation with different selectors to accurately capture the relevant data fields.

One challenge encountered during the data collection process was handling inconsistencies in the formatting of the data on the Wikipedia page. For example, variations in the representation of birth and death years required additional preprocessing steps to ensure uniformity and accuracy in the final dataset.

Despite these challenges, the process of gathering and analyzing the data was both educational and rewarding. It provided valuable insights into the history of Australian politics and the individuals who shaped it. Additionally, it offered an opportunity to practice web scraping and data wrangling skills in a real-world context.

In future iterations of this project, I would streamline the data collection process by refining the web scraping code to handle edge cases more effectively and automate repetitive tasks. Additionally, I would explore additional sources of data to enrich the analysis and provide a more comprehensive understanding of the factors influencing political leadership in Australia.

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

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

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.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, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

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

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.