Investigation in how long prime ministers of Australia lived, based on the year they were born.*

Daisy Huo

February 5, 2024

1 Introduction

In this paper, we will investigate in the life time of the prime ministers of Australia, based on the year they were born. We will scrape data from Wikipedia (Contributors 2019) using rvest (Wickham 2022), clean it, and then make a graph.

Data was collected, cleaned, and analyzed using the statistical programming software R (R Core Team 2022), with additional support from R packages "tidyverse" (Wickham et al. 2019), "rvest" (Wickham 2022), "randomNames" (Betebenner 2021), "xml2" (Wickham, Hester, and Ooms 2023), "dplyr" (Wickham et al. 2023), "readr" (Wickham, Hester, and Bryan 2024), "janitor" (Firke 2023), "knitr" (Xie 2014), "here" (Müller and Bryan 2020) and "ggplot2" (Wickham 2016).

2 Plan Data

We will start by planning the dataset. We need to plan two aspects. The first is what the simulated dataset will look like, and the second is what the final graph will look like.

The dataset needs to have variables that specify the names of the Prime Ministers, their birth year, death year and the years their lived. Roughly, it should look like Table 1.

We are interested to make a graph with year on the x-axis and the names of the Prime Ministers on the y-axis. Each Prime Minister should be categoried into "Alive" or "Passed Away". A quick sketch of what we are looking for is Figure 1.

*Code and data are available at: https://github.com/dai929/Prime_Minister_of_Canada.git

Table 1: Quick sketch of a dataset that could be useful for analyzing how long each Prime Minister of Australia lived

Prime Minister	Birth Year	Death Year	Years Lived
Eppard, Evelyn	1712	1769	57
Feaman, Nicholas	1717	1816	99
Ossello, Eric	1727	1804	77
Toves, Thien-Kim	1736	1810	74
Gurrola, Vanessa	1762	1851	89

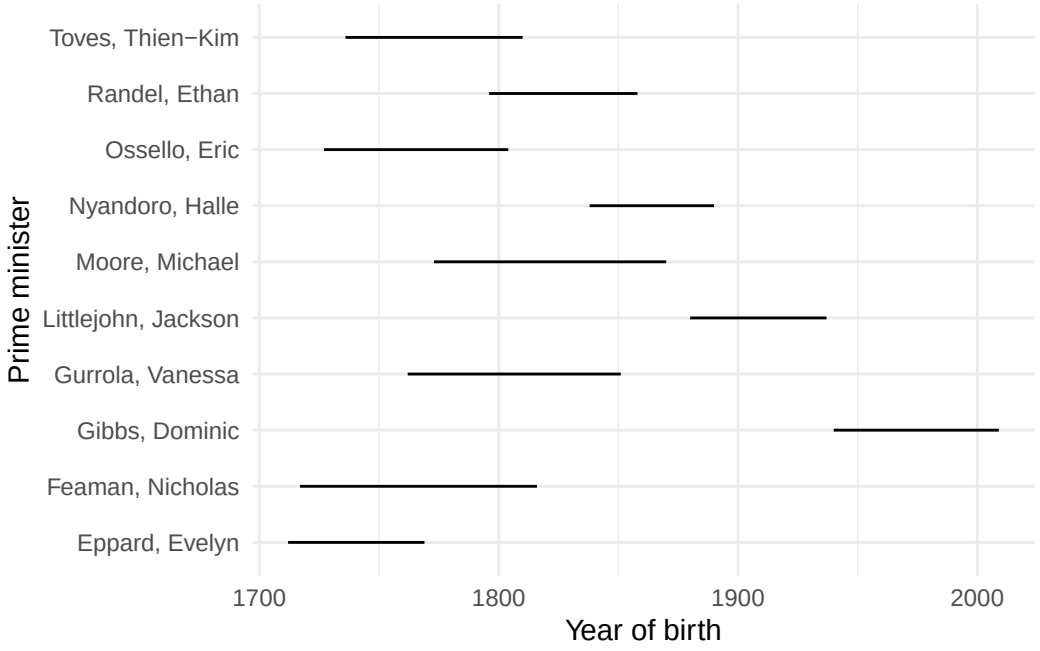

Figure 1: Quick sketch of planned graph showing how long prime ministers of Australia lived

3 Simulate Data

To that end, we will proceed by generating our simulated dataset. We would want a table that has a row for each prime minister, columns for their name, birth and death years. If the prime ministers are still alive, then their death year can be empty. We know that death year should be larger than birth year. Finally, we also know that the years should be integers, and the names should be characters. We want a dataset that looks roughly like Table 2.

Table 2: First ten rows of simulated dataset for analyzing how long each Prime Minister of Australia lived

Prime Minister	Birth Year	Death Year	Years Lived
Eppard, Evelyn	1712	1769	57
Feaman, Nicholas	1717	1816	99
Ossello, Eric	1727	1804	77
Toves, Thien-Kim	1736	1810	74
Gurrola, Vanessa	1762	1851	89
Moore, Michael	1773	1870	97
Randel, Ethan	1796	1858	62
Nyandoro, Halle	1838	1890	52
Littlejohn, Jackson	1880	1937	57
Gibbs, Dominic	1940	2009	69

4 Acquire Data

In interest of how long each prime minister of Autralia lived, we need to acquire a source of data for further web scraping. The Wikipedia page about prime ministers of Australia (Contributors 2019) fits our requirement and can therefore be used as a trusted source of data. As Wikipedia is known as a popular page, which the information contained is highly likely to be correct. The dataset are presented in a table.

First step, we want to save the dataset locally for reproducibility.

```
# A tibble: 6 x 12
  No.   Portrait  Name(Birth-Death)Con~1 `Election(Parliament)` `Term of office`
  <chr> <chr>     <chr>                  <chr>                  <chr>
1 No.   "Portrai~ Name(Birth-Death)Cons~ Election(Parliament)   Took office
2 1     ""        Edmund Barton(1849-19~ 1901 (1st)             1 January1901
3 1     ""        Edmund Barton(1849-19~ 1901 (1st)             1 January1901
4 1     ""        Edmund Barton(1849-19~ 1901 (1st)             1 January1901
5 2     ""        Alfred Deakin(1856-19~ - (1st)                24 September1903
6 2     ""        Alfred Deakin(1856-19~ 1903 (2nd)             24 September1903
# i abbreviated name: 1: `Name(Birth-Death)Constituency`
# i 7 more variables: `Term of office` <chr>, `Term of office` <chr>,
#   Politicalparty <chr>, Ministry <chr>, Monarch <chr>,
#   `Governor-General` <chr>, Ref. <chr>
```

However, in this case, there are too many redundant rows and columns that we do not need in our investigation. Therefore, we will need to clean the dataset.

```
# A tibble: 6 x 1
  raw_text
  <chr>
1 Edmund Barton(1849-1920)MP for Hunter, NSW
2 Alfred Deakin(1856-1919)MP for Ballaarat, Vic[a]
3 Chris Watson(1867-1941)MP for Bland, NSW
4 George Reid(1845-1918)MP for East Sydney, NSW
5 Andrew Fisher(1862-1928)MP for Wide Bay, Qld
6 Joseph Cook(1860-1947)MP for Parramatta, NSW
```

Now that we successfully obtained the parsed data, we need to clean it to match what we wanted. We want a column of the names of Prime Ministers, as well as the columns for birth year and death year. We then need to apply a different expression for the prime ministers who are still alive.

```
# A tibble: 6 x 3
  name          date      born
  <chr>         <chr>     <chr>
1 Edmund Barton 1849-1920 <NA>
2 Alfred Deakin 1856-1919 <NA>
3 Chris Watson  1867-1941 <NA>
4 George Reid   1845-1918 <NA>
5 Andrew Fisher 1862-1928 <NA>
6 Joseph Cook   1860-1947 <NA>
```

Finally, we are left to clean up the columns.

```
# A tibble: 6 x 4
  name           born  died Age_at_Death
  <chr>         <int> <int>        <int>
1 Edmund Barton  1849  1920           71
2 Alfred Deakin  1856  1919           63
3 Chris Watson   1867  1941           74
4 George Reid    1845  1918           73
5 Andrew Fisher  1862  1928           66
6 Joseph Cook    1860  1947           87
```

5 Explore Data

After having the cleaned data, our dataset would look pretty much similar to the sketch we had in Table 1.

Table 3: Cleaned dataset for analyzing how long each Prime Minister of Australia lived

Prime Minister	Birth year	Death year	Age at death
Edmund Barton	1849	1920	71
Alfred Deakin	1856	1919	63
Chris Watson	1867	1941	74
George Reid	1845	1918	73
Andrew Fisher	1862	1928	66
Joseph Cook	1860	1947	87

At this point we would like to make a graph that illustrates how long each prime minister lived (Figure 2). If they are still alive then we would like their lines to run to the end. Moreover, we would like to color this category differently.

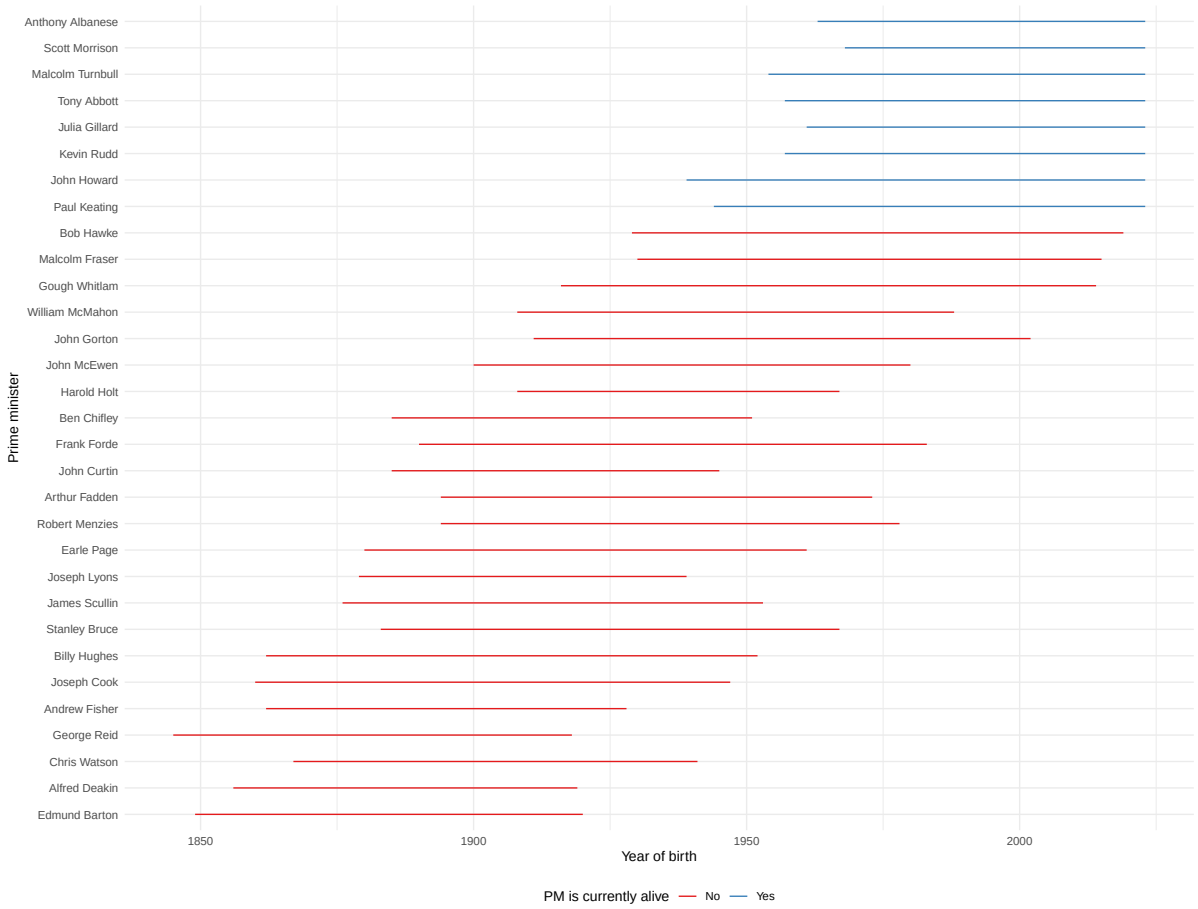

Figure 2: How long each prime minister of Australia lived

6 Discussion

By plotting out the life span of each Prime Minister of Australia, I found that 22 of the 31 Prime Ministers passed away, and 9 of the 31 Prime Ministers were still alive till now. Regarding of the ones that passed away, I am able to find a ascending trend of the number of years that they lived, which implies that recent Prime Ministers tend to live longer than those in the past. This could be caused by the improvement of medical health technology, while the correlation between this topic and the life span of prime ministers requires further study.

This paper took longer than expected in the data cleaning process. However, playing and exploring with the SelectorGadget (Wickham, n.d.) was fun. Next time when I plan to create a similar table and graph, I will try to analyze the raw data more carefully so that I can save some time in the data cleaning process.

References

Betebenner, Damian W. 2021. *randomNames: Function for Generating Random Names and a Dataset*. https://cran.r-project.org/package=randomNames.

Contributors, Wikipedia. 2019. "List of Prime Ministers of Australia." *Wikipedia*. Wikimedia Foundation. https://en.wikipedia.org/wiki/List_of_Prime_Ministers_of_Australia.

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

Müller, Kirill, and Jennifer Bryan. 2020. *Here: A Simpler Way to Find Your Files*. https://cran.r-project.org/web/packages/here/index.html.

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.

———. 2022. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

———. n.d. "SelectorGadget." *Rvest.tidyverse.org*. https://rvest.tidyverse.org/articles/selectorgadget.html.

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://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2023. *Xml2: Parse XML*. https://CRAN.R-project.org/package=xml2.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.