WebScraping of List of Prime Ministers of Australia*

STA302 Mini Essay 5

Ping-Jen (Emily) Su

February 6, 2024

Table of contents

*Code and data are available at: https://github.com/emisu36/WebScraping-Australia-PM

1 Introduction

WebScraping is performed on the Wikipedia page of "List of Prime Ministers of Australia" with the code from Alexander (2023), and further analyzed.

2 Data

Tools by R Core Team (2022), Wickham et al. (2019) and Firke (2023) is used to clean and analyze the data collected from Wikipedia

```
# 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
```

3 Results

The figure and table is generated with the help of Wickham (2016), Zhu (2021), and Xie (2014).

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

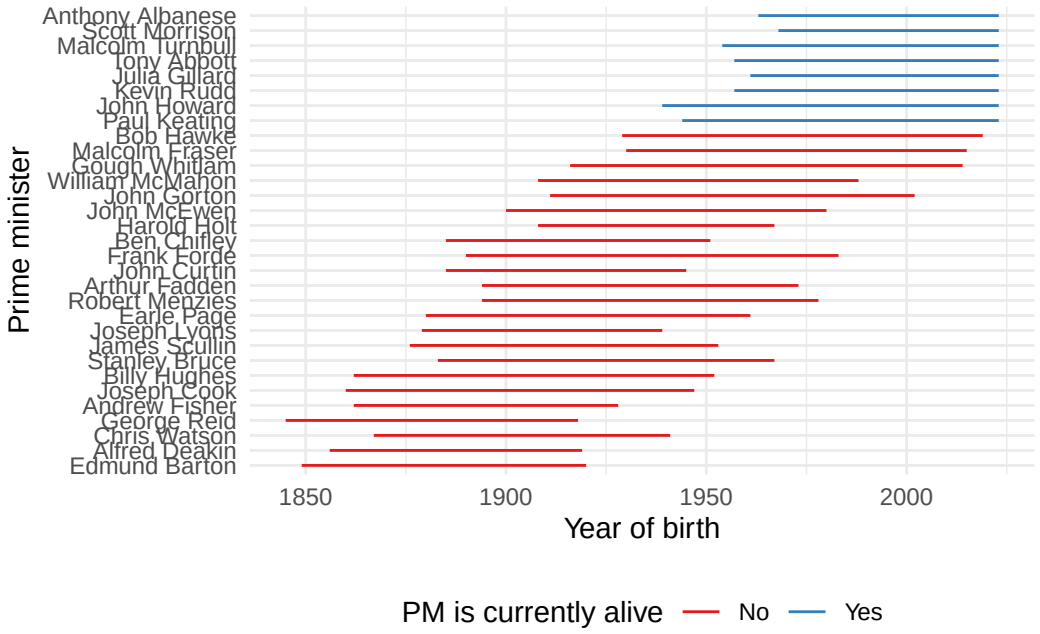

Figure 1: List of Prime Ministers of Australia LifeSpan

4 Discussion

4.1 Findings

The data provides us with information on the years of birth and death of different Australian prime ministers. We can learn and analyze from the data and also from Figure 1.

Life Spans: By examining each Prime Minister's birth and death years (where known), we can determine how long each of them lived.This information allows us to understand the longevity of each leader and compare their life spans.

Historical Timeline: We may construct a timeline of Australian Prime Ministers and comprehend the order of their leadership by looking at the years of birth and death. This aids in placing historical occurrences and political shifts in Australia in context.

Comparison of Ages: In order to evaluate patterns in life expectancy throughout time, we might compare the ages of deceased prime ministers. This approach can shed light on how healthcare and living standards have improved throughout time.

Current Leaders: We can see the present ages of the Prime Ministers who are still living. Understanding the age distribution of political leaders and their possible impact on current actions and policies is made easier with the use of this information.

In general, the information offers insightful information about the characteristics and lifespans of Australian prime ministers, which advances knowledge of the country's political past and leadership.

4.2 Procedure

As for the data source, Wikipedia is a popular resource for learning about a wide range of subjects, including public leaders like prime ministers. Australian prime minister data can be easily scraped and extracted using web scraping techniques, as the Wikipedia page on the subject provides the data in table form.

The data is scraped using the library rvest. After the data has been extracted, I will clean it up by eliminating any extra characters and changing the data types as required. For instance, the age at death for deceased prime ministers is calculated by converting their birth and death years to numerical format. And therefore the figure can be generated successfully.

4.3 Conclusion

Understanding the code and making the code work took longer than expected as the table of the Wikipedia pages do look different from one another. At the same time figuring what should be changed from the original code did take more time than planned. It became fun as everything starts to come together and you start to understand how every piece of code works. From easily searching up what packages every code belongs to implementing all the code and the different parsing techniques are all part of the learning process. Next time I would try graphs that look different or explore other ways to scrap wikipedia pages, such as scraping the second table on the page instead of the first one only.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r and Python*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

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.

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.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.