# (Web)scraping Through the Lives of Australian Prime Ministers*

Maroosh Gillani

January 23, 2024

## 1 Introduction

This paper uses the webscraping techniques taught in Telling Stories with Data (Alexander 2023) on Wikipedia's "List of prime ministers of Australia" article (*Wikipedia: List of Prime Ministers of Australia* 2023)

## 2 Findings

| 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 |
| Billy Hughes | 1862 | 1952 | 90 |
| Stanley Bruce | 1883 | 1967 | 84 |
| James Scullin | 1876 | 1953 | 77 |
| Joseph Lyons | 1879 | 1939 | 60 |
| Earle Page | 1880 | 1961 | 81 |
| Robert Menzies | 1894 | 1978 | 84 |
| Arthur Fadden | 1894 | 1973 | 79 |
| John Curtin | 1885 | 1945 | 60 |

*Code and data are available at: https://github.com/Maroosh-Gillani/STA302_Tutorial_5a

| Prime Minister | Birth year | Death year | Age at death |
|---|---|---|---|
| Frank Forde | 1890 | 1983 | 93 |
| Ben Chifley | 1885 | 1951 | 66 |
| Harold Holt | 1908 | 1967 | 59 |
| John McEwen | 1900 | 1980 | 80 |
| John Gorton | 1911 | 2002 | 91 |
| William McMahon | 1908 | 1988 | 80 |
| Gough Whitlam | 1916 | 2014 | 98 |
| Malcolm Fraser | 1930 | 2015 | 85 |
| Bob Hawke | 1929 | 2019 | 90 |
| Paul Keating | 1944 | NA | NA |
| John Howard | 1939 | NA | NA |
| Kevin Rudd | 1957 | NA | NA |
| Julia Gillard | 1961 | NA | NA |
| Tony Abbott | 1957 | NA | NA |
| Malcolm Turnbull | 1954 | NA | NA |
| Scott Morrison | 1968 | NA | NA |
| Anthony Albanese | 1963 | NA | NA |

From this table, we can see that the longest living prime minister was Gough Whitlan, who died at the age of 98 years old. On the other hand, the prime minister who lived the shortest period of time was Harold Holt, who passed away at the age of 59 years old. The most common range of age for Australian prime ministers to pass away is in their 80s, as 7 prime ministers passed away in that age range. This is closely followed by the age ranges of 70s and 90s, as both have 5 prime ministers each pass away in those age ranges. This is in line with the life expectancy (2020-2022) of Australia, which is 81.2 years for males and 85.3 years for females (Statistics 2022).

As illustrated by Figure 1, it is very interesting to see that 8 Australian prime ministers are still alive. Thus, roughly 26% of prime ministers Australia has had throughout its existence are currently alive. This makes sense, considering Australia is a relatively young country, at only 122 years old constitutionally (Blakeney 2023). This could also be because Australia has had a few long serving prime ministers as well. For example, Robert Menzies served as prime minister for roughly 16 years, and John Howard spent more than 11 years as prime minister (*Wikipedia: List of Prime Ministers of Australia* 2023).
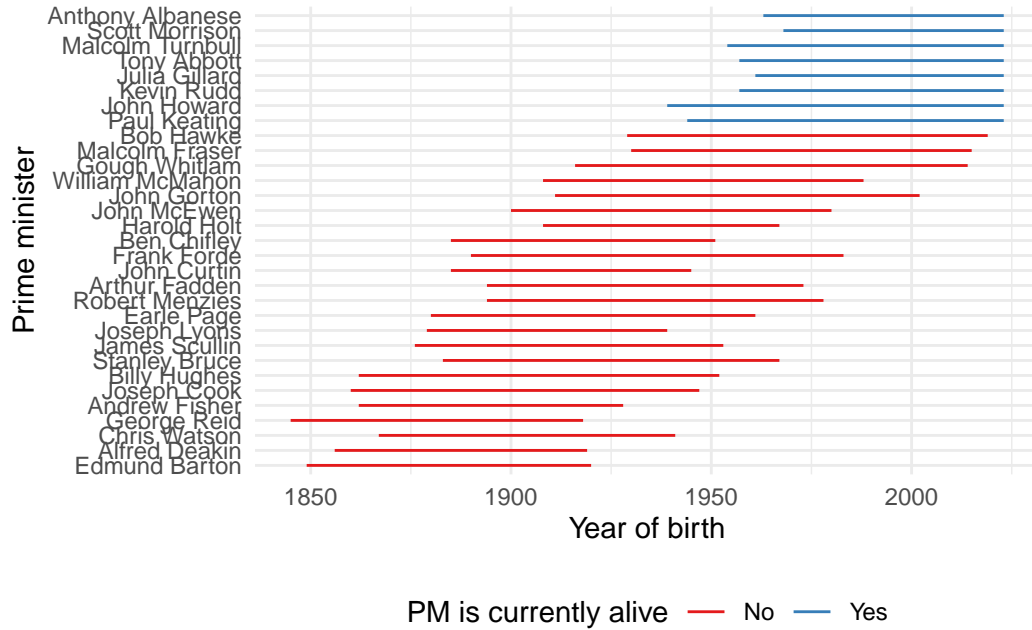
Figure 1: Graph Depicting Australian Prime Ministers and How Long They Lived

## 3 Web Scraping

### 3.1 Data Source

The source of the data used in this paper is Wikipedia (*Wikipedia: List of Prime Ministers of Australia* 2023). From the article, I specifically scraped the table of prime ministers of Australia, which included information such as their portrait, Name, Term of Office, and Political party etc.

### 3.2 Process

For the webscraping process, I used R itself (R Core Team 2023), and the following packages: tidyverse (Wickham et al. 2019), and knitr (Xie 2023). My webscraping code is a modified version of the one displayed in Telling Stories with Data (Alexander 2023). It is essentially the same code, but altered to suit my needs for scraping the Australian prime ministers Wikipedia page. For example, the original code required the column "Prime minister Office (Lifespan)" column from the United Kingdom Prime Ministers Wikipedia Article, whereas I required the "Name (Birth-Death) Constituency" column for this paper. My process was to just follow along the provided code and change it as needed, but this approach ended up causing me a few problems. Firstly, I should have tried to understand the code better before starting to

follow along. This is because there were parts that were left for the reader to resolve themselves, such as the implementation of the clean_names function, which I struggled to realize. Thus, I ended up wasting a lot of time trying to figure out if there was a package that included this function. Another issue I ran into because of this is my lack of understanding with RegEx. The provided code used regular expressions that were suited to parse the United Kingdom prime ministers Wikipedia article specifically. On the other hand, I had to use slightly different regular expressions while parsing (for example, the needed column names were different across the UK and Australian prime minister tables). I did not realize this for a while, which caused a lot of frustration when trying to debug my code. Thus, what I would do differently next time when webscraping is: (a) understand instructions (if provided) much more clearly, and (b) go throught the web page I am scraping more in-depth so I know exactly what to look out for, and thus write my code accordingly. The webscraping process became a lot more fun when I overcame the hurdles mentioned above. It turned out that webscraping is a lot more straightforward than I initially thought (mostly due to my own oversight), which helped me gain more confidence when coding, and thus made the process a lot more fun.

# References

Alexander, Rohan. 2023. *Telling Stories with Data.* Chapman; Hall/CRC. https://tellingsto rieswithdata.com/.

Blakeney, Kyol. 2023. *An Impatient Nation.* https://www.commonground.org.au/article/an-impatient-nation#:~:text=At%20235%20years%20old%20(and,ve%20all%20been%20ther e%20before.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Statistics, Australia Bureau of. 2022. *Life Expectancy.* https://www.abs.gov.au/statistics/pe ople/population/life-expectancy/latest-release.

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.

*Wikipedia: List of Prime Ministers of Australia.* 2023. https://en.wikipedia.org/wiki/List_o f_prime_ministers_of_Australia.

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