How long each prime minister of Australia lived*

Michael Fang

February 27, 2024

1 Analysis

1.1 Source

The project involved gathering data on Australian Prime Ministers from Wikipedia (Wikipedia 2005) using the rvest (Wickham 2022) package in R (R Core Team 2019) for web scraping, along with several other packages (dplyr (Wickham et al. 2023), tidyr (Wickham et al. 2024), stringr (Wickham 2023), babynames (Wickham 2021), janitor (Firke 2023), knitr (Xie 2023)) to manipulate and clean the data. The process began by fetching the raw HTML content of the Wikipedia page listing the Prime Ministers of Australia. This content was then converted to a character string and saved to a file named **`pms.html`** for processing.

1.2 Challenge

The main challenge was parsing the data from the HTML content accurately. This was done by identifying the correct CSS selector (**`.wikitable`**) that contained the data in a table format on the Wikipedia page. The **`html_table()`** function from **`rvest`** was used to extract the table data, which was then processed to clean and format the information appropriately. This involved renaming columns, separating combined data into distinct columns (such as names and birth-death years), and filtering out unwanted rows.

One aspect that took longer than expected was dealing with the inconsistencies in the data format, especially for Prime Ministers who are still alive versus those who have passed away. This required custom handling to accurately extract and calculate ages, as well as ensuring that birth and death years were correctly identified and separated.

[*] Code and data from this analysis are available at: https://github.com/fanger2791/PrimeMinistersAus

1.3 What was enjoyable?

The project became particularly enjoyable during the data cleaning and manipulation phase. Discovering and applying functions from `dplyr` and `tidyr` to transform the raw, messy data into a structured and meaningful dataset was satisfying. It was a practical application of data science techniques that showcased the power of R in handling and cleaning data.

1.4 What would I do differently?

If I were to approach this project again, one thing I would do differently is to spend more time upfront planning the data cleaning steps. Anticipating potential issues with the data format and consistency could streamline the process. Additionally, exploring more advanced text processing techniques or regular expressions to handle the variability in the data might make the cleaning process more efficient and robust.

2 Table

Table 1: Aus Prime Ministers, by how old they were when they died

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	1917	1952	35
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
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

Table 1: Aus Prime Ministers, by how old they were when they died

Prime Minister	Birth year	Death year	Age at death
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

3 Graph

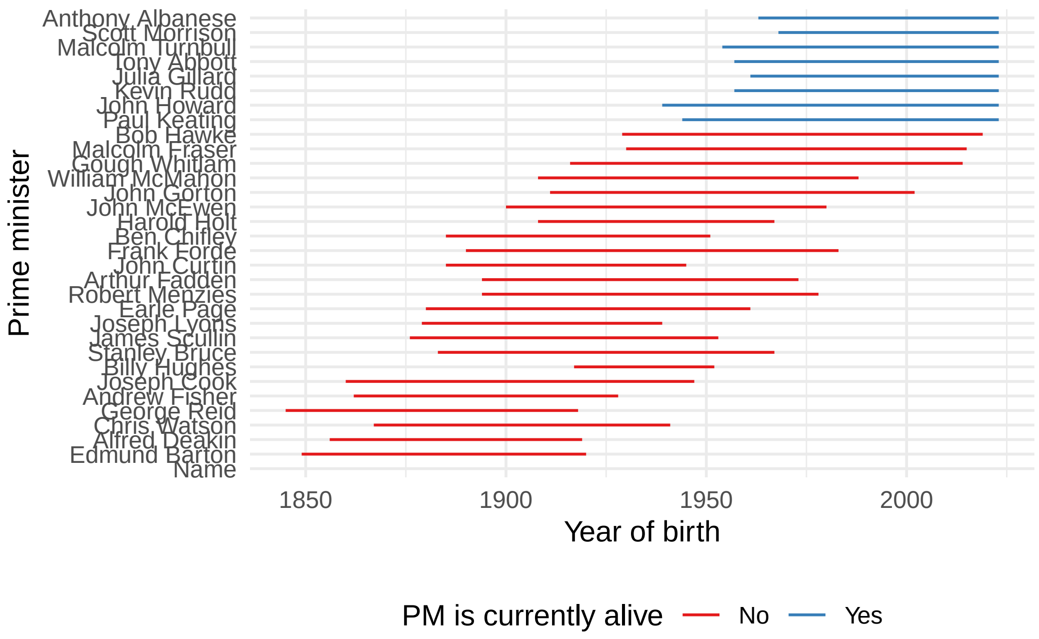

Figure 1: How long each prime minister of Australia lived

4 Conclusion

In summary, this project was a comprehensive exercise in web scraping, data cleaning, and manipulation using R. It highlighted the importance of thoroughly understanding the data source and structure, as well as the challenges and rewards of transforming raw data into a usable format for analysis.

References

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

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Wickham, Hadley. 2021. *babynames: US Baby Names 1880-2017*. https://CRAN.R-project.org/package=babynames.

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

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://cran.r-project.org/web/packages/tidyr/index.html.

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, Hadley Wickham, Davis Vaughan, Maximilian Girlich, and Kevin Ushey. 2024. *Tidyr: Tidy Messy Data*. https://cran.r-project.org/web/packages/tidyr/index.html.

Wikipedia. 2005. "List of Prime Ministers of Australia." /url%7Bhttps://en.wikipedia.org/wiki/List_of_prime_ministers_of_Australia%7D.

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