Life Span of Australian Prime Ministers[*]

Hari Lee

February 25, 2024

1 Plan

In the planning stage, I sketched and simulated a data set mirroring information about Australian Prime Ministers. The sketch outlined the Prime Minister's name, birth year, anticipated lifespan, and the death year.

Then, data was simulated, collected, cleaned, and analyzed using the open-source statistical programming software R (R Core Team 2019). This process involved various packages within R, including `tidyverse` (Wickham et al. 2019), `citervest` (Wickham 2022), `httr` (Wickham 2023), `janitor` (Firke 2023), `here` (Müller 2020), `knitr` (Xie 2023), `tidyr` (Wickham, Vaughan, and Girlich 2023), `lubridate` (Grolemund and Wickham 2011) and `purrr`(Wickham and Henry 2022). Aditionally, code was referenced from `Telling Stories with Data - 7 Gather data` (Alexander 2023).

To simulate the data, I employed the tibble function and included a set seed to ensured reproducibility. For realistic Prime Minister names, I used the (Wickham 2021) data set. From this data set I filtered names with a significant frequency (prop > 0.01). The birth year was then randomly sampled within a range spanning the 18th to the late 20th century. The simulation introduced variability by randomly assigning lifespans between 50 and 100 years. Ultimately, the death year was derived by adding the birth year to the simulated years lived.

2 Data Source

The data was sourced from Wikipedia, a widely used platform for collecting and presenting information. Wikipedia aggregates data from various reputable sources, providing a centralized location for information retrieval. However, it's crucial to acknowledge potential discrepancies or inaccuracies, as Wikipedia content is user-contributed and subject to change.

[*] Code and data from this analysis are available at: https://github.com/hari-lr/Australian-Prime-Ministers-Lifespan/tree/main/outputs

3 Data

This Wikipedia article outlines information about the role and history of the Prime Ministers of Australia (Wikipedia 2005) , offering insights into the constitutional framework, the process of appointment, and the historical context of individuals who have held the position. It details key aspects such as the appointment by the governor-general, the lack of a fixed term for prime ministers, and the factors that might lead to their departure from office, including resignations, electoral losses, and even dismissals.

The table used for this exercise provides a chronological list of the Prime Ministers of Australia, including key details such as their birth and death years, constituency, election and parliamentary terms, political party affiliation, ministry, reigning monarch, and governor-general during their tenure.

3.1 Scrape Data

For the initial part of the code I used the read_html function from the rvest package to scrape the content of the Wikipedia page listing Australian Prime Ministers. Then wrote the HTML content to a file named "pms.html."

3.2 Parse Data

I used html_element and html_table functions from the rvest package to extract tables from the HTML content. Then I parsed the table into a data frame using the clean_names function to standardize column names. The resulting data frame is named parse_data_selector_gadget. I further cleaned the data by renaming the relevant column to "raw_text," selecting only the rows with valid data, and removing duplicate entries.

3.3 Clean Data

To clean the data, I employed the separate function from the tidyr package into the "raw_text" column, splitting it into two columns: "name" and "not_name," based on the opening parenthesis. The str_extract function from the stringr package is then used to extract the "date" (years of service) and "born" (birth year) information from the "not_name" column. The "date" column is further split into "birth" and "died" columns using the separate function. The "born" and "birth" columns are reconciled to handle differences in formatting for Prime Ministers who are still alive. Finally, the table is created with columns "Prime Minister," "Birth year," "Death year," and "Age at death," and duplicate rows are removed.

3.4 Create a Table

I cleaned and formatted data and stored in the aus_cleaned_data data frame. Then, I displayed it as a table using the kable function from the knitr package. The column names are customized, and the first row (which usually contains column descriptions) is excluded with the slice(-1) function.

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
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

4 Results

After analyzing the dataset of Australian Prime Ministers' birth years, death years, and ages at death reveals interesting patterns into the country's political history. The average age at death for the Prime Ministers in the data set is approximately 74 years, reflecting a diverse range of lifespans. Some leaders, like Gough Whitlam, lived notably longer, reaching the age of 98, while others, such as Billy Hughes, had shorter lives.

The data set also highlights the temporal distribution of Prime Ministers, spanning from the late 19th century with Edmund Barton to contemporary figures like Scott Morrison. Examining the age at death provides a historical perspective on the challenges and circumstances faced by these leaders during their terms.

5 Reflection

The most time-consuming part of the data processing was the cleaning phase, particularly the separation of raw data strings into different columns based on the parenthesis. This step required careful consideration of the structure of the data and ensuring that the separation was accurate and consistent. Once I successfully solve this challenge, the data cleaning process flowed more smoothly.

While I wouldn't categorize the experience as "fun," there was a noticeable shift from struggling to a more efficient workflow after cleaning the data.

A key takeaway is the importance of thoroughly understanding the guidelines and conditions for extracting specific data elements. In future projects, I would invest more time analyzing string data structure to make the cleaning process more effective.

References

Alexander, Rohan. 2023. "7 Gather Data." *Telling Stories with Data - 7 Gather Data*. https://tellingstorieswithdata.com/07-gather.html.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://doi.org/10.18637/jss.v040.i03.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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. *httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.

Wickham, Hadley, Mara Averick, Jenny 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, and Lionel Henry. 2022. *purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2023. *tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

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/.