Aussie PMs*

Dead or Alive?

Michaela Drouillard

1 March 2023

I've scraped the Wikipedia for Australian Prime Ministers to visualize their birth and death dates. This is for Tutorial 7.

Gathering Data

My data source was the Wikipedia page for "List of prime ministers of Australia". Using the "read_html" function in the rvest [Wickham (2022a)] package in R (R Core Team 2022), I read the HTML content from the Wikipedia page. I wrote the content into a local file named "pms.html", and then read it in again, and named it raw_data.

I used the "html_element" and "html_table" functions from the rvest package to extract data from the HTML content. "html_element" extracts the HTML element with class ".wikitable", and the "html_table" function converts it into a dataframe. I saved both the raw data and the parsed data as RDS files.

Cleaning Data

I used the clean_names function from the janitor package in R to clean the column names in parse_data_selector_gadget, and stored it as a new table (Firke 2021). Then, I selected only the column of interest, which was title name_birth_death_constituency. The values in the column contained the data of interest: the prime ministers names, and their birth and death dates. I removed the first row, which was a duplicate of the column headings.

I used the "separate" function from the tidyr package to split names from birth and death fates at the first "(" character, creating "name" and "not_name" columns (Wickham and Girlich 2022). I used the "str_extract" function from the stringr package (Wickham 2022b) to extract the values for two new columns, "date" and "born", from the "not_name" column. I then used the "mutate" function to remove the "b." text from the "born" column for prime minister who are still alive. Then, I used select to only select the "name", "date", and "born" columns.

*Code and data are available at: https://github.com/michaeladrouillard/AussiePMs_DeadOrAlive.git.

Then, I used the "separate" function again to split the "date" column into "birth" and "died" columns at the "-" character. I used "str_remove"all" from the stringr package to remove the "b." text from the "born" column, and the "if_else" function to replace missing values from the "birth" column with values from the "born" column. (These discrepancies came from the way that dead Prime Ministers and alive Prime Ministers dates were recorded). Then, I turned the born and died columns in integers, and calculated the "Age_at_Death" column by taking the difference between "died" and "born". The "distinct" function from the dplyr package removes any duplicate rows (Wickham et al. 2023).

Results

Table 1: Aussie Prime Ministers, and 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	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
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

Prime Minister	Birth year	Death year	Age at death
Malcolm Turnbull	1954	NA	NA
Scott Morrison	1968	NA	NA
Anthony Albanese	1963	NA	NA

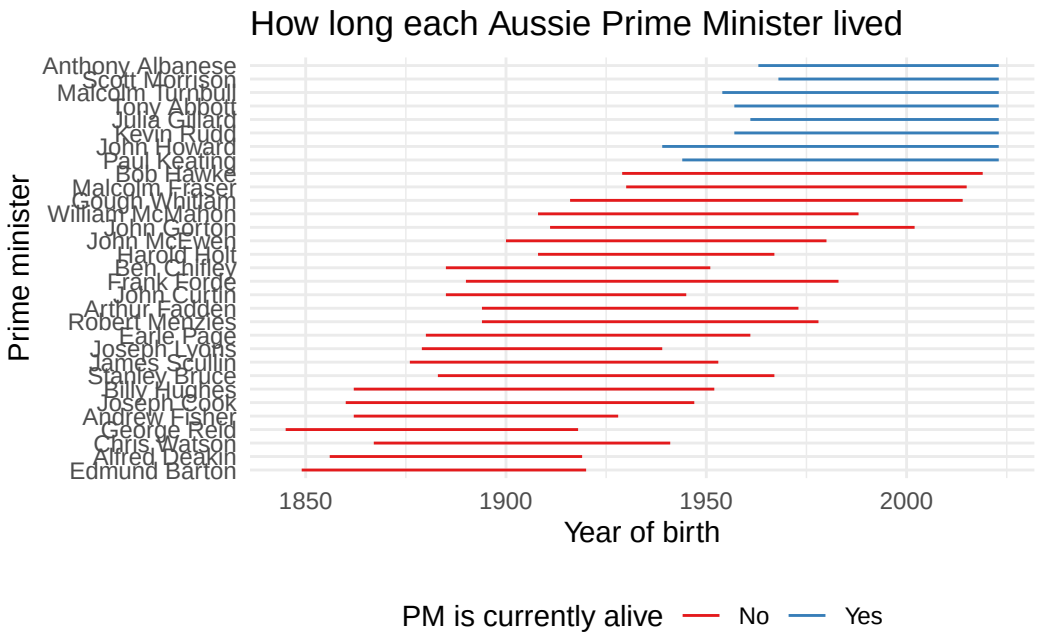

Figure 1: How long each UK prime minister lived

Discussion

Surprisingly, not much in the actual scraping took longer than expected. I was using your example, and the structure was very similar. I only had to change a few lines to have the right values and column titles. The exact moment it became fun was when I gave it a stupid description in the GitHub repo.

What is taking much longer than expected is rendering this damn thing into a PDF. My RStudio is encountering issues with the "as_factor" function in the ggplot code (Wickham 2016).

Oh wait.

I can now shift to past tense because I JUST resolved the issue.

After loading the forcats library (Wickham 2023) in the paper.qmd file, I was able to render the file as a PDF.

I think in the future, I would actually write the code from scratch and not just modify your chunks of code. I always kick myself when I copy paste code because I'm not searing it into my brain the way I should be at this stage of learning.

Back to the Aussies

Gough Whitlam has the longest lifespan of any Australian prime minister. He died at 98 of natural causes. Harold Holt, so far, has the shortest. He disappeared while swimming at Cheviot Beach near Portsea, Victoria, Australia, on December 17, 1867, and his body was never found[1].

The average lifespan of an Australian Prime Minister is 77.86 years, and all but one have been men.

Table 2: Aussie Prime Ministers, BY how old they were when they died

Prime Minister	Birth year	Death year	Age at death
Gough Whitlam	1916	2014	98
Frank Forde	1890	1983	93
John Gorton	1911	2002	91
Billy Hughes	1862	1952	90
Bob Hawke	1929	2019	90
Joseph Cook	1860	1947	87
Malcolm Fraser	1930	2015	85
Stanley Bruce	1883	1967	84
Robert Menzies	1894	1978	84
Earle Page	1880	1961	81
John McEwen	1900	1980	80
William McMahon	1908	1988	80
Arthur Fadden	1894	1973	79
James Scullin	1876	1953	77
Chris Watson	1867	1941	74
George Reid	1845	1918	73
Edmund Barton	1849	1920	71
Andrew Fisher	1862	1928	66
Ben Chifley	1885	1951	66

[1] Omg. I swear to god I was just joking around when I first started this project and put the repo description as "scraping wiki to figure out which Aussie PMs are alive, which ones are dead, and which ones are FAKING IT (kidding) (unless....?)". I had no idea about Holt. Obviously this is tragic. But a part of me hopes that he's faking it, and that I'm accidentally the whistle blower.

Prime Minister	Birth year	Death year	Age at death
Alfred Deakin	1856	1919	63
Joseph Lyons	1879	1939	60
John Curtin	1885	1945	60
Harold Holt	1908	1967	59
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

In Table 3, we can observe that, of the living Prime Ministers, John Howard has already been alive for slightly longer than the average lifespan.

Table 3: Ages of Living Prime Ministers

Prime Minister	Born	Current Age
John Howard	1939	82
Paul Keating	1944	77
Malcolm Turnbull	1954	67
Kevin Rudd	1957	64
Tony Abbott	1957	64
Julia Gillard	1961	60
Anthony Albanese	1963	58
Scott Morrison	1968	53

References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=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.

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

———. 2022b. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2023. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

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, and Maximilian Girlich. 2022. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.