Webscraping New Zealand Prime Minister Lifespan Data*

Moohaeng Sohn

February 5, 2024

1 Findings

In New Zealand, prime ministers (also known as PM) are the leader of the political party which has won the largest share of seats forming the government, and stays in the seat as long as the political party keeps the majority (McLean 2023). Though earlier in history of prime ministers, they were given the title "colonial secretary", and later on it was a mix of "premier" and "prime minister"(McLean 2023). After 1906, all leaders were assigned the title of prime minister (McLean 2023). In this paper, we will include colonial secretaries, and premiers as they served essentially the same role as the prime minister.

Table 1 shows a sample of the final dataset that we were able to scrape from Wikipedia page which contained a list of all New Zealand Prime Ministers (Wikipedia contributors 2024). PMs with no death year is still alive. There were 43 prime ministers in New Zealand as of writing this paper, and we have found that 9 past prime ministers of New Zealand are still alive. We also have learned that the shortest living PM was Norman Kirk, dying after living for 51 years. On the other hand, the longest living prime minister is Jim Bolger, who is currently living for 89 years, and is still alive.

Dame Jacinda Ardern is the prime minister with the latest birth year in our dataset, being born in 1980. And Henry Sewell was born in 1807, making Henry Sewell the prime minister with the oldest birth year in our dataset. The latest death of a prime minister occured in 2020, when Mike Moore has died, and Henrey Sewell, the first prime minister (called colonial secretary back then), has the oldest death year of 1879.

There doesn't seem to be a clear trend between the prime minister's birth year and years lived by the prime minister. Figure 1 shows which prime ministers are dead or alive, and how many years they have lived, or has lived so far in case of the PMs who are still alive. It seems like

*Code and data are available at: https://github.com/alexsohn1126/new_zealand_prime_ministers

there is very little correlation between the year of birth and the number of years lived. Even ignoring the currently alive prime ministers does not seem to reveal a clear trend between the birth year and years lived.

Table 1: New Zealand Prime Ministers' Lifespan Data, In the Order of First Day in Office

Prime Minister	Birth Year	Death Year	Years Lived
Henry Sewell	1807	1879	72
William Fox	1812	1893	81
Edward Stafford	1819	1901	82
Alfred Domett	1811	1887	76
Frederick Whitaker	1812	1891	79
Frederick Weld	1823	1891	68
George Waterhouse	1824	1906	82
Julius Vogel	1835	1899	64
Daniel Pollen	1813	1896	83
Harry Atkinson	1831	1892	61
George Grey	1812	1898	86
John Hall	1824	1907	83
Robert Stout	1844	1930	86
John Ballance	1839	1893	54
Richard Seddon	1845	1906	61
William Hall-Jones	1851	1936	85
Joseph Ward	1856	1930	74
Thomas Mackenzie	1853	1930	77
William Massey	1856	1925	69
Francis Bell	1851	1936	85
Gordon Coates	1878	1943	65
George Forbes	1869	1947	78
Michael Joseph Savage	1872	1940	68
Peter Fraser	1884	1950	66
Sidney Holland	1893	1961	68
Keith Holyoake	1904	1983	79
Walter Nash	1882	1968	86
Jack Marshall	1912	1988	76
Norman Kirk	1923	1974	51
Hugh Watt	1912	1980	68
Bill Rowling	1927	1995	68
Robert Muldoon	1921	1992	71
David Lange	1942	2005	63
Geoffrey Palmer	1942		
Mike Moore	1949	2020	71

(continued)

Prime Minister	Birth Year	Death Year	Years Lived
Jim Bolger	1935		
Dame Jenny Shipley	1952		
Helen Clark	1950		
John Key	1961		
Bill English	1961		
Dame Jacinda Ardern	1980		
Chris Hipkins	1978		
Christopher Luxon	1970		

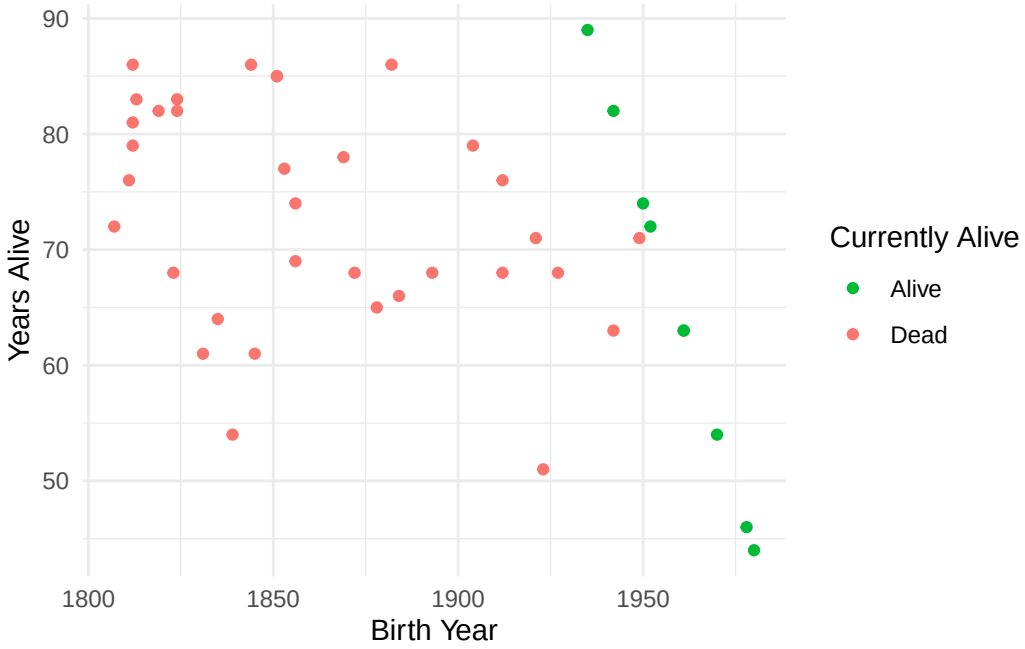

Figure 1: New Zealand Prime Ministers' Birth Year Versus Years Lived

2 Data Source

We have collected data using Open-source Statistical programming language R (R Core Team 2022), along with the following R packages: tidyverse (Wickham et al. 2019), kableExtra (Zhu 2021), here (Müller 2020), xml2 (Wickham, Hester, and Ooms 2023), and rvest (Wickham 2022). We have used rvest (Wickham 2022) R package‘to scrape the data off of a Wikipedia page where the names of New Zealand prime ministers as well as their birth year and death year was recorded (Wikipedia contributors 2024).

Once we got the HTML data from Wikipedia, we were able to use rvest (Wickham 2022) package to select the column that we are concerned about. Then, once we got that column, we had to parse the raw string data and convert it into useful data that we want. Using regex and tidyverse's (Wickham et al. 2019) functions, I was able to extract the information we want from the table that was in the Wikipedia page, and save it as a csv file. There were 2 tables which contained information about New Zealand prime ministers, we had to create 2 tables and then combine them back into one before saving it as a csv file.

The longest part of web scraping was when we were converting raw strings into different data we want. To get the name of each PM, we had to write regex (regular expression) to match patterns within the string which allowed us to extract the data, but there were many edge cases. We couldn't get the first 2 words of each raw string as a lot of them included "The Honourable" or "The Right Honourable", so we removed them with tidyverse (Wickham et al. 2019) functions from every row. But still, getting first 2 words of the new string meant that PMs with 3 word names such as "Dame Jenny Shipley" would not be correctly selected. We noticed that, the last name of a PM was appended by their titles, such as KCMG or GCMG. Therefore we would get whatever comes before the title of those PMs, and that worked decently. It was hard to learn and test these regular expressions as they are not intuitive for people just learning them.

This process became enjoyable when we were able to automate the process of scraping data off of Wikipedia and being able to convert them into a nice csv file in the end with a click of a few buttons. After all the hard work, it was amusing to see whatever was in the Wikipedia page is now in a nicely formatted file, which we can use to analyze and create tables and graphs like we have above.

If we were to do a similar project in the future, we would like to try and make the conversion process going from the raw string into a data table in R easier by making that conversion process a function instead of copy-pasting code and manually changing it like how we did in `02-data_cleaning.R`. This would allow us to not repeat the code, and if there were even more tables, say, 10 tables, we would be able to reuse that function instead of having to copy-paste them 10 times.

References

McLean, Gavin. 2023. "The Role of Prime Minister." *Te Ara Encyclopedia of New Zealand*. Ministry for Culture; Heritage Te Manatu Taonga. https://teara.govt.nz/en/premiers-and-prime-ministers/page-1.

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

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

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.

Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2023. *Xml2: Parse XML*. https://CRAN.R-project.org/package=xml2.

Wikipedia contributors. 2024. "List of Prime Ministers of New Zealand — Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=List_of_prime_ministers_of_New_Zealand&oldid=1195831308.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.