Prime ministers of New Zealand

STA302 Mini essay 5a: Web Scrapping

Brandon Benson

2024-02-05

Github repo: https://github.com/brndnjbenson/New_Zealand_PMs_lifespan.git

---

Data Source

The data source containts all of the administered Prime Ministered of New Zealand, obtained from Wikipedia where the data are presented in a table form. Using SelectorGadget which is included in rvest, the data is web scrapped into R with package xml2 and rvest, which then read_html() is used followed by html_element() to extract the desired data from Wikipedia.

Along with the names of the Prime Ministers, the data also includes the number order of Prime Minister, details of Constituency, Election and Parliament information, term of office, political party, and government details for each listed Prime Minister.

After importing the data set which includes all of the columns named parse_data_selector_gadget, the data are then organized using html_table(), and the process of data cleaning then followed.

Data

Table 1: New Zealand Prime Ministers, arranged in the order of administration. , and their Birth year and how long have they lived until the year they died.

Prime Minister	Birth Year	Death Year	Life Span
Henry Sewell	1807	1879	72
Sir William Fox	1812	1893	81
Sir Edward Stafford	1819	1901	82
Alfred Domett	1811	1887	76

Prime Minister	Birth Year	Death Year	Life Span
Sir Frederick Whitaker	1812	1891	79
Sir Frederick Weld	1823	1891	68
George Waterhouse	1824	1906	82
Sir Julius Vogel	1835	1899	64
Daniel Pollen	1813	1896	83
Sir Harry Atkinson	1831	1892	61
Sir George Grey	1812	1898	86
Sir John Hall	1824	1907	83
Sir Robert Stout	1844	1930	86
John Ballance	1839	1893	54
Richard Seddon	1845	1906	61
William Hall-Jones	1851	1936	85
Sir 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
Sir Sidney Holland	1893	1961	68
Sir Keith Holyoake	1904	1983	79
Sir Walter Nash	1882	1968	86
Sir Jack Marshall	1912	1988	76
Norman Kirk	1923	1974	51
Hugh Watt	1912	1980	68
Sir Bill Rowling	1927	1995	68
Sir Robert Muldoon	1921	1992	71
David Lange	1942	2005	63
Sir Geoffrey Palmer	1942	NA	NA
Mike Moore	1949	2020	71
Jim Bolger	1935	NA	NA
Dame Jenny Shipley	1952	NA	NA
Helen Clark	1950	NA	NA
Sir John Key	1961	NA	NA
Sir Bill English	1961	NA	NA
Dame Jacinda Ardern	1980	NA	NA
Chris Hipkins	1978	NA	NA
Christopher Luxon	1970	NA	NA

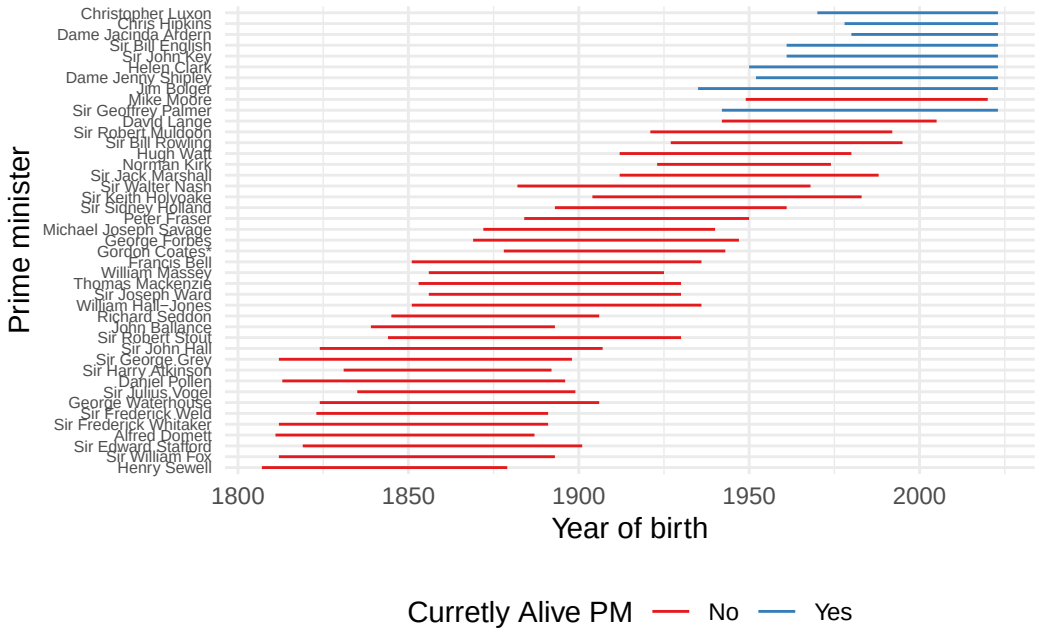

Figure 1: The Lifespan of every New Zealand Prime Ministers, with names of Prime Minister arranged from bottom to top in order of administration.

Table 1 shows lifespan or ages of all Prime Ministers administered for the country New Zealand since the establishment of the nation, including their birth and death year. There has been a total of 43 Prime Ministers in New Zealand's political history, with the first Prime Minister Henry Sewell born in the early 1800s. This is the earliest birth year for a Prime Minister, whilst on the other hand respectively to the current year, Dame Jacinda Ardern is born the latest out of the other Prime Ministers on year 1980, and also the youngest at 43 years old. Looking at the life span of the Prime Ministers, there are three individuals with the longest lifespan at 86 years old. These include Sir George Grey, Sir Robert Stout, and Sir Walter Nash, all born in the 1800s.

With Figure 1, we can observe the distribution pattern of the Prime Ministers lifespan over time, where the red colored data marks those have passed and blue for those who are still alive. The most prominent observation in Figure 1 is that there are significantly more dead Prime Ministers than the alive ones. Using Table 1, only 9 Prime Ministers are currently alive, with the oldest being Jim Bolger in 1935. The remaining Prime Ministers whom are dead has life span in the range of 50 to 80 years. All of the alive individuals are born 1930s and over, but they are not all the latest administered Prime Ministers. In the last 10 administrations, 1 out of 10 Prime Ministers died which is Mike Moore in 2020. Sir Geoffrey Palmer is the earliest Prime Ministers to not have died yet. Just over a half of New Zealand Prime Ministers are born before the 1900s, and are all dead. All of the alive Prime Ministers are born after 1950s in Figure 1, and only 11 Prime Ministers made it after the year 2000.

Data Source

The data source containts all of the administered Prime Ministered of New Zealand, obtained from Wikipedia where the data are presented in a table form. Using SelectorGadget, the data is web scrapped into R with package xml2 and rvest, which then read_html() is used followed by html_element() to extract the desired data from Wikipedia.

Along with the names of the Prime Ministers, the data also includes the number order of Prime Minister, details of Constituency, Election and Parliament information, term of office, political party, and government details for each listed Prime Minister.

After importing the data set which includes all of the columns named parse_data_selector_gadget, the data are then organized using html_table(), and the process of data cleaning then followed.

Reflections

The part which took the longest is figuring out how to use the tool SelectorGadget in helping with web scrapping. In some times, it became a trial and error process but eventually, all the needed data was gathered after some attempts. The process of mutating of the parsed_data were also very time consuming, because the focus is to only collect the name of the Prime Ministers, as that column also contains other information like their Constituency, Year alive, and Title therefore a removal of these unwanted details were applied.

The project became fun during the data clean process, which includes the said mutating process. This challenges my ability to prepare a presentable data, where being proactive in finding and removing unwanted details and keeping the necessary ones helps to develop an intuition in data cleaning. Then putting the prepared data set into a graph also challenges my ability to create a tidy and well presented graph containing data that I've thoroughly reorganized.

To improve, I would use a more efficient method in data cleaning to reduce the time consuming parts of that stage. On some parts, I manually removed unwanted rows and altered the birth years and lifespan values because it resulted in NA when it shouldn't. This issue happened to about 5 rows, which therefore I decided manually mutate these rows however, if there were more rows that needed alterations, I would result with using another method which offers a more efficient effect.

References

Wikipedia contributors. (2024, January 15). List of prime ministers of New Zealand. In *Wikipedia, The Free Encyclopedia*. Retrieved 18:59, February 5, 2024, from https://en.wikipedia.org/w/index.php?title=List_of_prime_ministers_of_New_Zealand&oldid=1195831308

R Core Team (2021). R: A language and environment for statistical computing. R Foundation fo
https://www.R-project.org/.

Wickham H (2021). _babynames: US Baby Names 1880-2017_. R package version 1.0.1,
<https://github.com/hadley/babynames>.

Wickham H, Hester J, Ooms J (2023). _xml2: Parse XML_. R package version 1.3.6,
https://github.com/r-lib/xml2, <https://xml2.r-lib.org/>.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G,
Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K,
Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K,
Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_,
4(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

Wickham H (2022). _rvest: Easily Harvest (Scrape) Web Pages_. R package version
1.0.3, https://github.com/tidyverse/rvest, <https://rvest.tidyverse.org/>.

Firke S (2023). _janitor: Simple Tools for Examining and Cleaning Dirty Data_. R
package version 2.2.0, https://sfirke.github.io/janitor/,
<https://github.com/sfirke/janitor>.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of
Data Manipulation_. R package version 1.1.4, https://github.com/tidyverse/dplyr,
<https://dplyr.tidyverse.org>.

Wickham H, Vaughan D, Girlich M (2023). _tidyr: Tidy Messy Data_. R package
version 1.3.0, https://github.com/tidyverse/tidyr, <https://tidyr.tidyverse.org>.

Xie Y (2023). _knitr: A General-Purpose Package for Dynamic Report Generation in
R_. R package version 1.45, <https://yihui.org/knitr/>.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and
Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In
Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing
Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595