Web Scraping Wikipedia's data on Canadian Prime Ministers*

Timothius Prajogi

February 6, 2024

This data was downloaded, cleaned, parsed, analyzed, and visualized using R (R Core Team 2023), a statistical programming language, with package support from `tidyverse` (Wickham et al. 2019), a collection of libraries which included the following packages that were utilized:

- `ggplot2` (Wickham 2016)
- `dplyr` (Wickham et al. 2023)
- `readr` (Wickham, Hester, and Bryan 2023)
- `tibble` (Müller and Wickham 2023)

Additionally for webscraping libraries xml2 (Wickham, Hester, and Ooms 2023) and rvest (Wickham 2022) were used.

Table 1: A sample of Canadian Prime Minister's Lifespans

Prime Minister	Birth Year	Death Year	Age at death
John A. Macdonald	1815	1891	76
Alexander Mackenzie	1822	1892	70
John Abbott	1821	1893	72
John Thompson	1845	1894	49
Mackenzie Bowell	1823	1917	94
Charles Tupper	1821	1915	94

Findings

Of the 23 Prime Ministers Canada has had, a lot of them die around similar times with other Prime Ministers. For instance the first 4 presidents lived and died around the same time,

*Code and data supporting this analysis is available at: https://github.com/prajogt/web_scraping

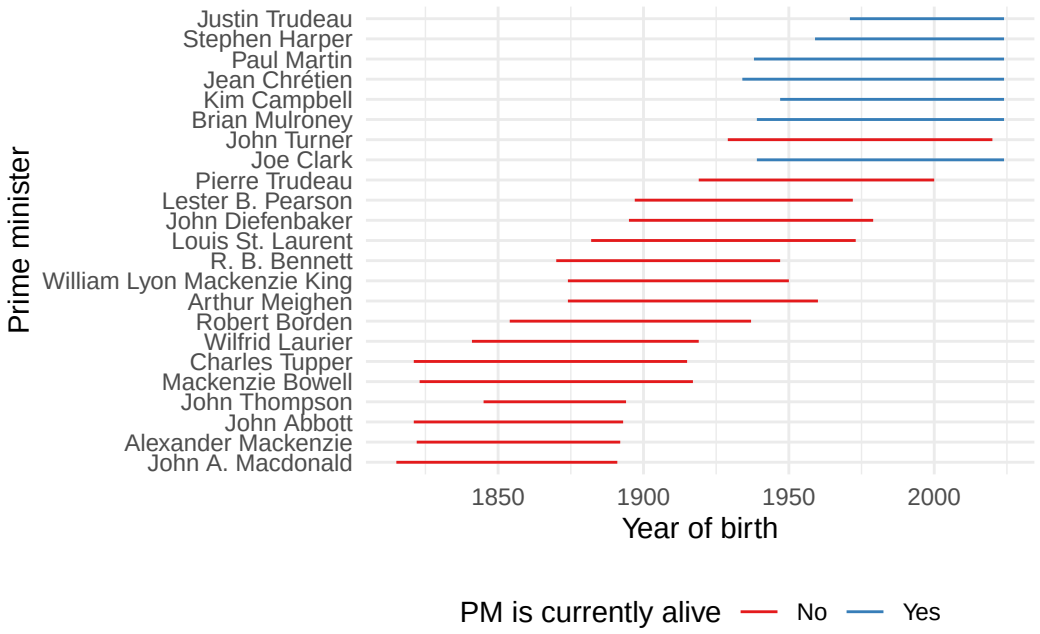

Figure 1: The lifespans of Canadian Prime Ministers

which isn't that interesting in itself since it is the beginning of Canadian independence. Then Mackenzie Bowell to Wilfrid Laurier all died around the time of WW1. Then many of the later Prime Ministers died before the 1980s, but there was a 21 year gap until the next Prime Minister died, and then another 20 year gap until the next.

These gaps were caused by the age differences between the PMs. It was not always the case that older PMs were elected before younger PMs. Some standouts are John Thompson who died relatively young at 48 years of age, and who was also elected young, younger than other Prime Ministers of that time period by around 20 years. Another stand out is Joe Clark, who outlived the Prime Minister who came after him, also due to his relative youth upon election.

Then later, due to Stephen Harpers long term spanning 9 years, a jump in age can be seen in the next Prime Minister, Justin Trudeau, being 12 years younger than his predecessor.

Process

The Data

The data was scraped from Wikipedia's page on Canadian Prime Ministers ("List of Prime Ministers of Canada" 2024). Which contains a table of all of Canada's Prime Ministers, their

year of birth, their year of death, as well as other information not covered in this report such as length of term as Prime Minister, which political party they are in, and the acts and laws they had passed.

Issues in the Process

What took longer than it should have was the SelectorGadget portion of the data scraping. Since there were multiple tables using the `.wikitable` class, the result of `html_table()` was not what was expected. It was as simple as selecting the element from the list that was returned by `html_table()` instead, but the error messages provided by `clean_names()` were not helpful in debugging this.

What was fun?

For me it was sort of fun when we were using regex to parse through the data. From a long line of plain text, we were able to gather the data we wanted for this report.

It also showed the tedium that goes into webscraping, not every website / websource styles their text, html, or css the same, making it difficult to achieve general functionality. Even in this case, if we were to not inspect element Wikipedia's webpage, we wouldn't know which elements and classes held the data we want. `.wikitable` is used to make all Wikipedia tables look the same throughout different articles, but other Wikis and pages may use different tags and methods to present data.

What would I do differently?

In the future I would consider more specific selectors or consider different methods of selection. Especially for larger pages with multiple `.wikitables` it would be difficult to know which one is relevant. Luckily for this page there were only 2, but there are definitely some pages with many more.

Also, better ways of filtering out junk data should be done. When webscraping, the returned value also returned some header / metadata values, which got included in the rows. If I were to have not explicitly printed out all values, it may have gotten through to the end. Possible regex to ensure that only characters, periods and apostrophes would be kept (possible elements in a name), and all "code-like" which would include curly braces, commas and other symbols and therefore be excluded.

References

"List of Prime Ministers of Canada." 2024. https://en.wikipedia.org/wiki/List_of_prime_ministers_of_Canada.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

R Core Team. 2023. *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.

———. 2022. *Rvest: Easily Harvest (Scrape) Web Pages*. https://rvest.tidyverse.org/.

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, 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, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Jim Hester, and Jeroen Ooms. 2023. *Xml2: Parse XML*. https://xml2.r-lib.org/.