

Web Scraping Wikipedia's data on Canadian Prime Ministers*

Timothius Prajogi

February 5, 2024

The data was scraped from Wikipedia's page on Canadian Prime Ministers ("List of Prime Ministers of Canada" 2024).

These data statistics were 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

*Code and data supporting this analysis is available at: <https://github.com/prajogt/prime.git>

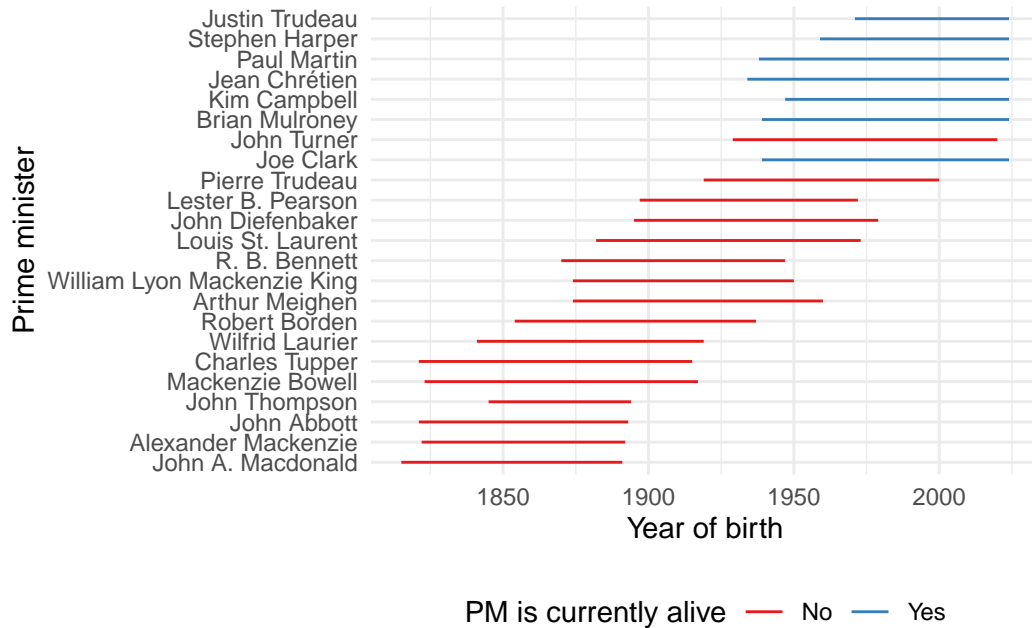


Figure 1: The lifespans of Canadian Prime Ministers

Findings

ASDLKASLKDm

- “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 Golemund, 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/>.