Life Span of the Prime Ministers of Canada*

Kaavya Kalani

February 2, 2024

1 Introduction and Data Source

In this analysis, the primary objective is to visually depict the average life spans of Canadian Prime Ministers. The dataset for this investigation is derived from the comprehensive information available on the official Wikipedia page dedicated to the list of Prime Ministers of Canada https://en.wikipedia.org/wiki/List_of_prime_ministers_of_Canada.

2 Process and Findings

This analysis consisted of me extracting data from Wikipedia, cleaning it and visualising it. This was done using the statistical programming software R (R Core Team 2023) along with the help of multiple libraries: rvest (Wickham 2022), xml2 (Wickham, Hester, and Ooms 2023), tidyverse (Wickham et al. 2019), janitor (Firke 2023), knitr (Xie 2014) and dplyr (Wickham et al. 2023).

The analytical process initiated with data extraction from Wikipedia, using the rvest package to scrape information, subsequently creating a local backup of the page. I then looked for patterns in the HTML that we can use to help us get closer to the data that we want.

I then used SelectorGadget to choose the appropriate table on the page which was the second table. Following that I cleaned my data, focusing on retaining only the essential column while addressing duplicate entries, particularly relevant for prime ministers with multiple terms.

Next, I noticed a pattern in the way information was written in a specific column. I then split and extracted data into name, birth year, and death year for a more organized dataset.

Finally, I calculated the age of all the Prime Ministers who have passed away and that was the last column of my final dataset.

*Code and data are available at: https://github.com/kaavyakalani26/prime_minister_canada.git

Table 1: Sample of cleaned Lifespan data

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

Table 1 provides a glimpse into our cleaned dataset.

I then plot the data to visualise the life span of the prime ministers

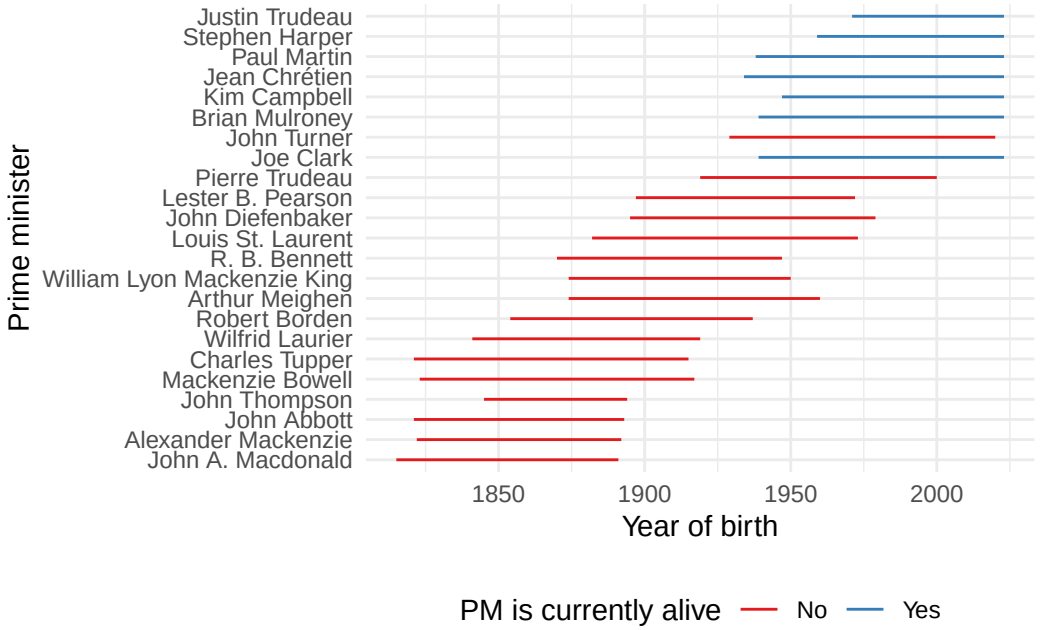

Figure 1: Life Span of the Prime Ministers of Canada

Figure 1 plots the lifespan of the prime ministers of Canada

Looking at the graph, we can see that on an average, most of the lifespans are around 75-80 years. On statistically analysing, the data we find that, among people who have passed away, the mean is 80 years which is very close to what we infer from just looking at the graph. Additonally, smallest life span recorded is 49 years and largest life span recorded is 94 years.

3 Experience with the process

3.1 What took longer than expected?

Figuring out how to split the column information into different columns according to the specific formatting of the table took a long time.

3.2 When did it become fun?

Once I was able to split the column in to appropriate columns, it was fun to calculate the age and plot the graph and visually see the information.

3.3 What would you do differently next time?

Something I would do differently would be that I would understand and plan how to split the column before I code it. This time I essentially started coding and correcting and it wasn't working so I had to take a step back, see what the code is doing and write that part on my own again to make it work for the specific formatting of the table I extracted.

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

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. 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://dplyr.tidyverse.org.

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

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.