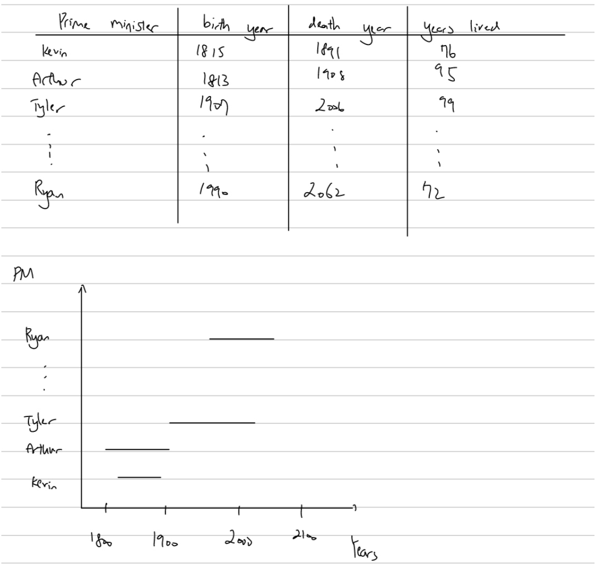
Lifespan of Prime Ministers of Canada

Hyuk Jang

2024-02-06

Lifespan of Prime Minister's of Canada[1]

The question of interest in this paper is how long each prime minister of Canada has lived. The initial plan of the paper is to identify how long each prime minister has lived and the period of time when they lived.his is so that we can use this data to do further analysis in the future. For this reason we need to find and identify a source of data. We need a source of data that is reliable and has some sort of structure. The Wikipedia page about prime ministers of Canada fits both of the criterion's. Below we have a sketch of what we want the data to look like once we have finished cleaning and produced a graph out of it.



Prime minister	birth year	death year	years lived
Kevin	1815	1891	76
Arthur	1813	1908	95
Tyler	1907	2006	99
⋮	⋮	⋮	⋮
Ryan	1990	2062	72

Figure 1: Plan

[1] https://github.com/anggimude/Life-span-of-Canadian-Prime-Ministers

From the website, we scraped the data such as the prime ministers names, birth year, and death year. To do this we have read the website and changed it into html format so that we can use R to clean this data. Within the website scraped, we are only interested in the table that contains names, birth year, and death year of prime ministers of Canada. In the same website there are two tables thus we have made a table named parse_table which imports these data from the html. Then, the data was cleaned so that repeated prime minister names are merged into one so that there is no repetition. Further cleaning was done so that we have prime minister names on one column and their birth year to death year on another column. To find the lifespan of each prime minister we must separate the birth year and death year because it is originally in YYYY-YYYY or b. YYYY format, so that we have two columns, one with the birth year(born) and the other with death year(died). Finally, the lifespan(age at death) was calculated by subtracting died - born. However, there is an issue if we leave it here because there are prime ministers that are still alive and we must also calculate there ages instead of leaving it NA. Despite the fact that this paper is written during 2024, since we did not take into account the exact birth date and death date, it is more accurate to use 2023 for age calculation as it is still early in the year.

Based off of the cleaned data, a graph that illustrates how long each prime minister lived. This cleaned data can be further manipulated to be used for questions like average lifespan of prime ministers of Canada, or produce another graph that includes how many prime ministers are still alive or not.

```
                Prime_Minister born died Age_at_death
1            John A. Macdonald 1815 1891           76
2          Alexander Mackenzie 1822 1892           70
3                  John Abbott 1821 1893           72
4                John Thompson 1845 1894           49
5             Mackenzie Bowell 1823 1917           94
6               Charles Tupper 1821 1915           94
7              Wilfrid Laurier 1841 1919           78
8                Robert Borden 1854 1937           83
9               Arthur Meighen 1874 1960           86
10 William Lyon Mackenzie King 1874 1950           76
```

Cleaning the data took much longer than expected because there were a plethora of issues that arise while the data was being cleaned. For example, the unique_prime_minister was not a data set as I expected, instead it was a value which was the reason the code was not running. Another example is that there were two types of birth_death_year format where one was in YYYY-YYYY and the other was b. YYYY. Due to this the code must be executed for the two different cases to clean up the data. While I was coding I kept running into the issue where one format would work but for the other format values it would keep having an error giving me NA. Figuring how to make the code work for both types of data formats took much longer than I initially expected. Despite the difficulty in certain parts, it was fun whenever I

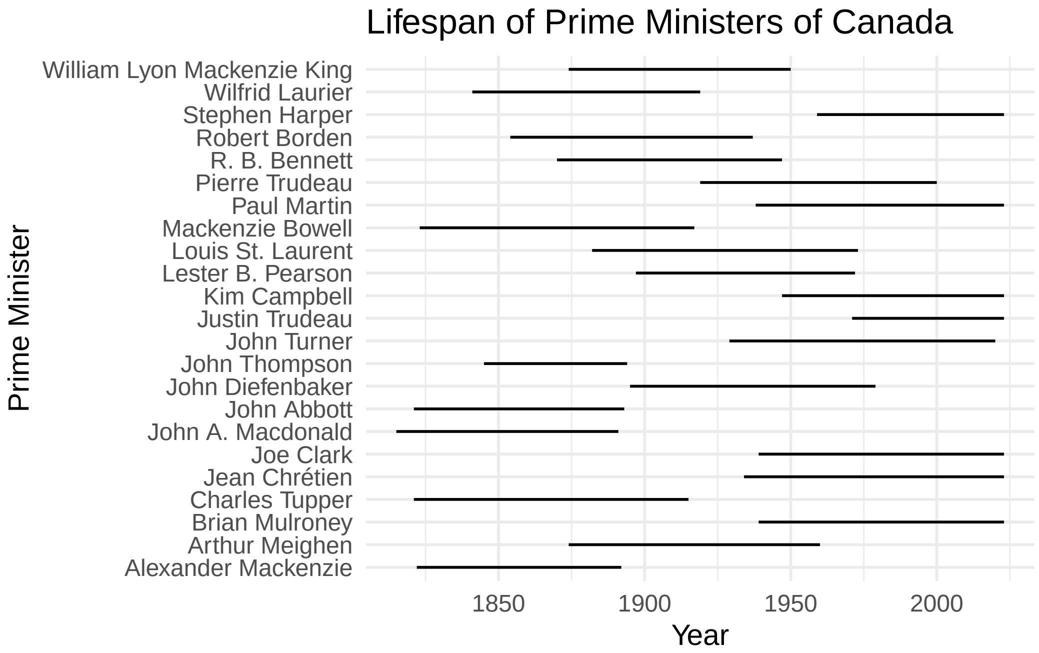


Figure 2: Lifespan of Prime Ministers of Canada

made progress because I was able to manipulate wrong parts of my code to produce data sets that I was thinking of and when it came out to be how I wanted it to be, it gave me joy. If a similar project is done again in the future, I would try to improve the graph and cleaned data so that the graph can include the prime ministers who have passed away in one color and alive in another color and organize the data set so that it is in descending order depending on when each prime minister has passed.

Citation

Firke, Sam. 2021. Janitor: Simple Tools for Examining and Cleaning Dirty Data. https://CRAN.R-project.org/package=janitor.

Müller, Kirill, and Hadley Wickham. 2022. Tibble: Simple Data Frames. https://CRAN.R-project.org/package=tibble.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.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, and Kirill Müller. 2021. Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.

Wickham H (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1, https://github.com/tidyverse/stringr, https://stringr.tidyverse.org.

Wickham H (2023). *rvest: Easily Harvest (Scrape) Web Pages*. https://rvest.tidyverse.org/, https://github.com/tidyverse/rvest.

Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org.