

# Reflection on Analysis of Lifespans of Prime Ministers of Canada\*

Irene Huynh

February 6, 2024

## 1 Results

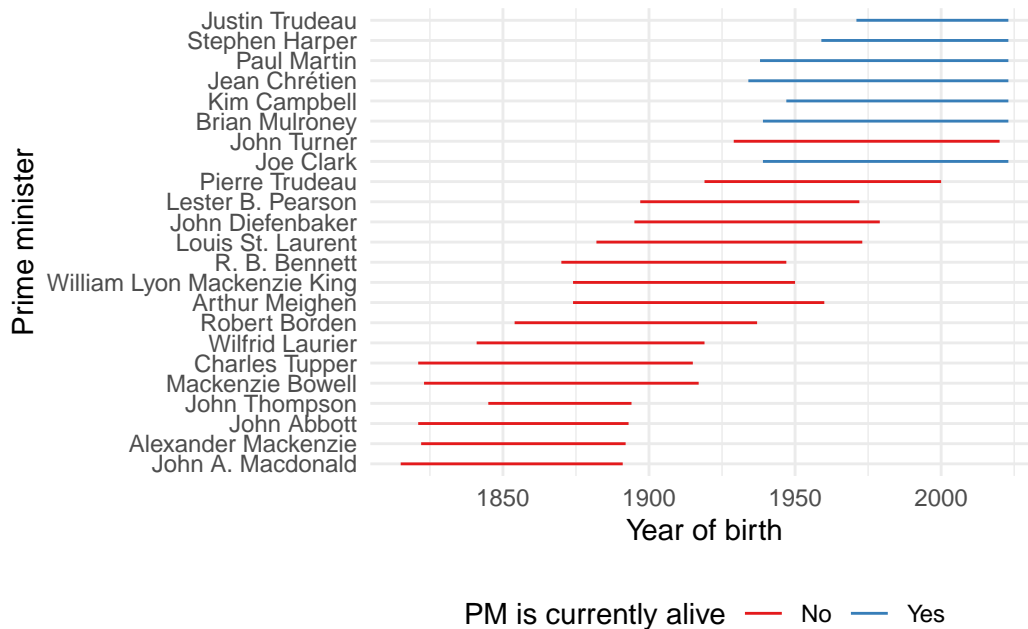


Figure 1: The lifespan of all Prime Ministers of Canada

Figure 1 reveals that John Thompson had the shortest lifespan out of all the prime ministers of Canada and was also much younger than his predecessors. Additionally, almost all prime ministers have died, except for seven that are still alive. This was expected as most of the prime ministers were born more than a hundred years ago.

\*Code is available at: [https://github.com/IreneeeH/STA302\\_Mini-Essay\\_5a.git](https://github.com/IreneeeH/STA302_Mini-Essay_5a.git)

## 2 Planning, Gathering, and Cleaning Data Sources

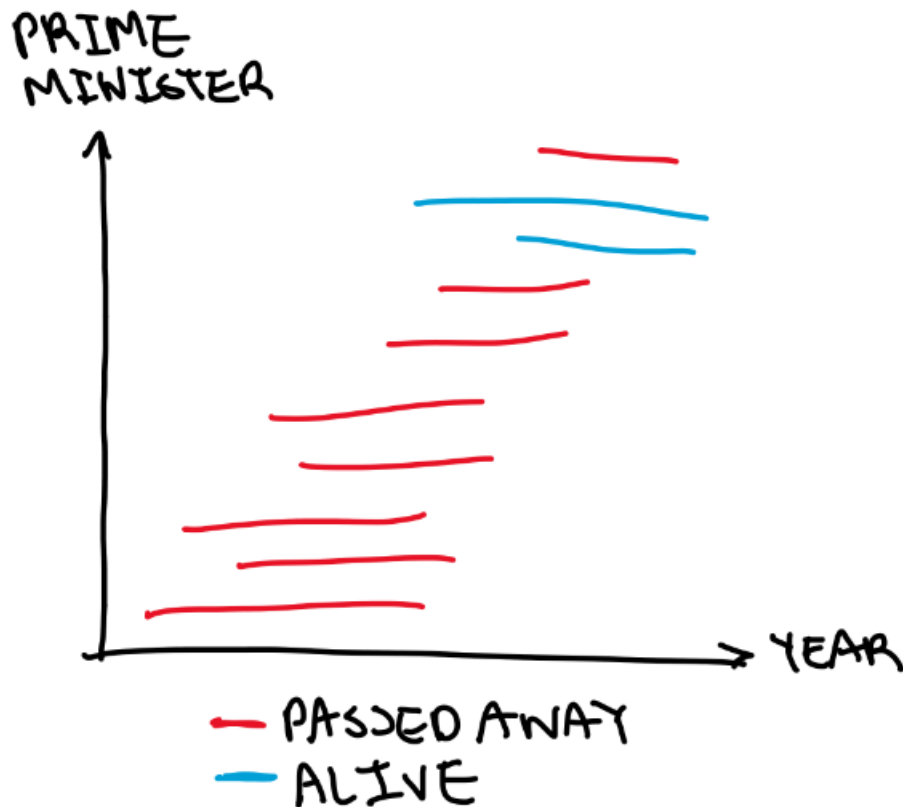
The statistical programming language R was used to retrieve, clean, and process the data (R Core Team 2022). In particular, the following R packages were used: tidyverse (Wickham et al. 2021), rvest (Wickham 2022), and xml2 (Wickham, Hester, and Ooms 2023) for data acquisition; knitr (Xie 2023), janitor (Firke 2023), dplyr (Wickham et al. 2023), and here (Müller 2020) for data cleaning and processing; and ggplot2 (Wickham 2016) for creating figures. Additionally, babynames (Wickham 2021) and pdftools (Ooms 2023) were used for simulating data.

## 3 Planning

A simulation of the expected data set is found below, having the prime minister's name, their birth year, death year, and years lived as columns. The simulation was created using the following R packages: tidyverse (Wickham et al. 2021), babynames (Wickham 2021), and pdftools (Ooms 2023).

```
# A tibble: 10 x 4
  prime_minister birth_year death_year years_lived
  <chr>          <int>    <int>    <int>
1 Kevin          1813     1908      95
2 Karen          1832     1896      64
3 Robert         1839     1899      60
4 Bertha         1846     1915      69
5 Jennifer       1867     1943      76
6 Arthur         1892     1984      92
7 Donna          1907     2006      99
8 Emma           1957     2031      74
9 Ryan           1959     2053      94
10 Tyler         1990     2062      72
```

A sketch of the expected graph can be found below:



## 4 Gathering

The code for the gathering of the data can be found in `scripts/01-download_data.R`

The data utilized in this paper was retrieved from a Wikipedia page of the List of Prime Ministers of Canada (*List of Prime Ministers of Canada* 2023) using the `rvest` R package (Wickham 2022).

### 4.1 Difficulties

This aspect of the paper took longer than expected as there was another element in the original `pms.html` document with a `wikitable` class that precedes the desired table of prime ministers of Canada. This caused `html_element("wikitable")` to use that table instead of the desired one. To resolve this issue, I went into `pms.html` and removed the `wikitable` class from the undesired table. Now, `html_element("wikitable")` is able to identify the correct `wikitable`.

## 5 Cleaning

The code for the cleaning of the data can be found in `scripts/02-data_cleaning.R`

### 5.1 Difficulties

While using the original code provided in *Telling Stories With Data - 7 Gather data* (Alexander, n.d.), cleaning the data was difficult as the birth dates of prime ministers who are still alive did not appear in the final figure. After some trial and error, I realized that the first “born” attribute must be manipulated to include the date: `born = str_extract(not_name, “[[:digit:]]{4}”)`. Additionally, the finalized `cleaned_data` included a final row with extra information from the html file. To remove this, I had to filter it out using `filter(substring(name, 1, 1) != “.”)` as it began with a period.

## 6 Comments

Surprisingly, I was not upset that the code provided to us in *Telling Stories With Data - 7 Gather data* (Alexander, n.d.) did not work flawlessly, rather, it was fun to go through the code and actually understand what was happening and be able to fix the errors. I enjoyed learning how to use and write regular expressions and use unfamiliar functions, such as `html_element()`.

If I were to do this paper again, I would start much earlier to prevent the anxious feeling of not understanding the code and worrying I won’t fix errors in time.

## References

- Alexander, Rohan. n.d. *Telling Stories - 7 Gather Data: Prime Ministers of the United Kingdom*. <https://tellingstorieswithdata.com/07-gather.html#prime-ministers-of-the-united-kingdom>.
- Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. <https://github.com/sfirke/janitor>.
- List of Prime Ministers of Canada*. 2023. [https://en.wikipedia.org/wiki/List\\_of\\_prime\\_ministers\\_of\\_Canada](https://en.wikipedia.org/wiki/List_of_prime_ministers_of_Canada).
- Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. <https://here.r-lib.org/>.
- Ooms, Jeroen. 2023. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. <https://docs.ropensci.org/pdftools/> (website) <https://github.com/ropensci/pdftools#readme> (devel) <https://poppler.freedesktop.org> (upstream).
- R Core Team. 2022. *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*. New York: Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- . 2021. *Babynames: US Baby Names 1880-2017*. <https://github.com/hadley/babynames>.
- . 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 Grolemond, et al. 2021. *Tidyverse: Easily Install and Load the ‘Tidyverse’*. <https://CRAN.R-project.org/package=tidyverse>.
- 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. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.