Prime Minister Tutorial - Mini Essay 5a*

Krishiv Jain

February 4, 2024

1 Data Source

The source of the data is Wikipedia, which has a list of Prime Ministers for almost every nation, including India. This list includes a variety of metrics, such as their term in office, their political party, and the head of state during their term. Jawaharlal Nehru was the first and longest-serving prime minister. Indira Gandhi was the only woman to hold the position. Manmohan Singh was the first prime minister from a minority religion. The current prime minister is Narendra Modi. The data source can be used to learn about the history of the prime ministers of India. Moreover, the data source can be used to study the trends in Indian politics. For our purpose, we are concerned with their lifespan.

2 Data Gathering

The data was downloaded, cleaned, and analyzed using the statistical programming language R (R Core Team 2024). The following packages were also used: janitor (Sam Firke 2023), here (Müller and Brya 2020), tidyverse (Wickham et al. 2019), rvest (Wickham 2023), xlm2 (Wickham, Hester, and Ooms 2023), dplyr (Hadley Wickham 2023), tibble (Kirill Müller 2023). The majority of the code was from Telling Stories with Data (Alexander 2023).

Data on their life span was gathered. In the table on Wikipedia, the Prime Minister's lifespan comes under their name. To gather the data, Wikipedia was webscraped. The read_html() function was used to download the data. Since we are only concerned with the table, SelectorGadget was used to help identify the necessary command to access the table, which was ".wikitable", and this was then converted into a table within R. The data was then cleaned to remove unneeded information. The remaining data was of class "character", which contained the Prime Minister's name, and their year of birth and death in brackets. Thus, this was

*Code and data are available at: https://github.com/Krishiv-J/STA302-Mini-essay-5a.git split based on the brackets, and the numbers were classified as the birth year and death year respectively. The difference of the two variables was taken to calculate their age at death.

3 Additional Information

3.1 What took longer than you expected?

Extracting the relevant data after web scrapping the table took longer than expected. I thought that, once I had the data from Wikipedia, getting the specific information I wanted would be straight forward. However, figuring out exactly how to extract the data took much longer than I had expected.

3.2 When did it become fun?

I quite enjoyed seeing the ease with which I was able to web scrape a website such as Wikipedia. Through this method, I believe I can extract relevant data from numerous websites and use the data for my own analysis on a variety of topics. It was also fun once I had the clean data, and was able to use this data, for instance in creating the figure.

3.3 What would you do differently next time you do this?

Next time, I would spend more time on planning the web scraping process and identifying the specific elements of the website that I need. Moreover, I might look into more advanced web scraping methods to improve the data extraction process and reduce the time spent on cleaning the data afterwards.

Table 1: Cleaned Data

name	born	died	Age at Death
Jawaharlal Nehru	1889	1964	75
Gulzarilal Nanda	1898	1998	100
Lal Bahadur Shastri	1904	1966	62
Indira Gandhi	1917	1984	67
Morarji Desai	1896	1995	99
Charan Singh	1902	1987	85
Rajiv Gandhi	1944	1991	47
Vishwanath Pratap Singh	1931	2008	77
Chandra Shekhar	1927	2007	80
P. V. Narasimha Rao	1921	2004	83

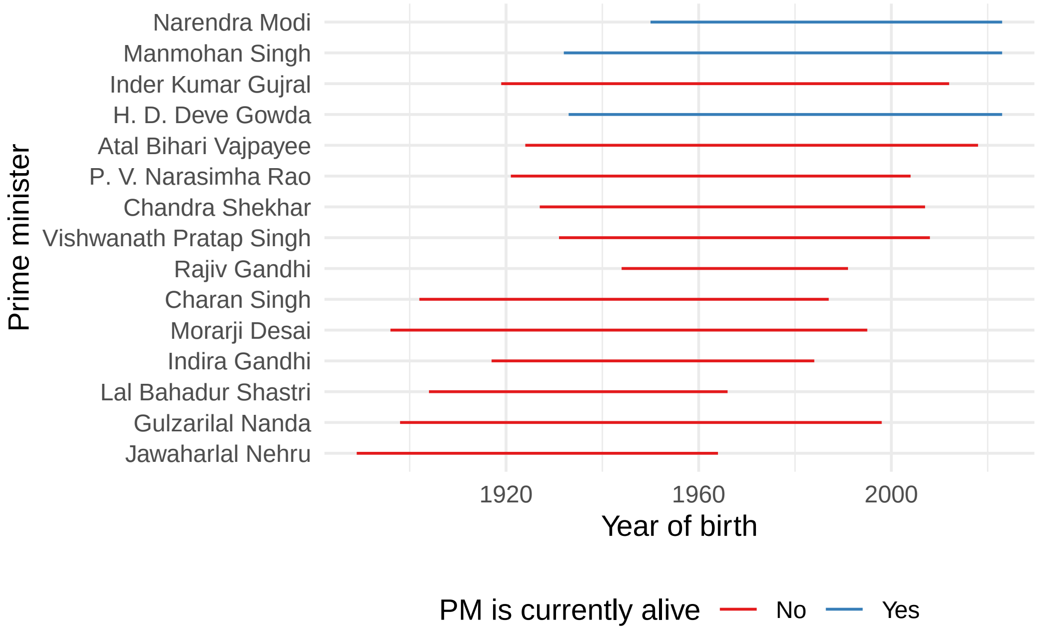

Figure 1: Lifespan of each Indian Prime Minister

References

Alexander, Rohan. 2023. *Telling Stories with Data*. https://tellingstorieswithdata.com/.

Hadley Wickham, Lionel Henry, Romain François. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

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

Müller, Kirill, and Jennifer Brya. 2020. "Here: A Simpler Way to Find Your Files." https://here.r-lib.org/.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sam Firke, Chris Haid, Bill Denney. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Wickham, Hadley. 2023. *Rvest: Easily Harvest (Scrape) Web Pages*.

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