India Prime Ministers*

Sami El Sabri

February 5, 2024

1 Simulate Data

The goal is obtain a dataset of all Prime Ministers of India to compare their

In this short paper, the objective is to analyze the lifespans of Indian prime ministers, based on their birth years. My approach involves scraping data from Wikipedia using the rvest package in the open-source statistical programming language R (R Core Team 2022), performing data preprocessing, and finally creating a visual representation through graph plotting. Since web scraping is always dependent on how the website is structured, which can change over time, it is important to have a clear idea of the end goal. That is why, to begin, I generated simulated data resembling the structure expected from the Wikipedia scrape. Each entry in our simulated dataset corresponds to a prime minister and it includes fields for their name, birth year, and if they are not currently still alive, their death year. Since India is a fairly new democracy, established in 1948, I expect the birth years to fall within the range of 1900 and 1990, and the death year having to be larger than the birth year. We want the dataset to look as follows:

2 Download Data

As described above, gathering the data means scraping it from a webpage. I chose Wikipedia, since it tends to be rather accurate for factual topics such as Birth and Death years. The article is titled 'List of prime ministers of India' (Wikipedia 2024). The scraping is done using the `read_html` and `write_html` from the `rvest` (Wickham 2021) package, and the code can be reviewed in the appendix (Section 5) for reference.

*Code and data are available at: https://github.com/samielsabri/IndiaPM

3 Clean Data

The data cleaning process involved several step, which were done using the R packages `tidyverse` Wickham et al. (2019) and `kableExtra` Zhu (2021): First, we select the relevant html elements from the scraped raw data, which in this case would be the table of prime ministers. Next, only the relevant rows are selected, which would be `Name(born - died)Constituency` as it includes all the information that we need: Name, Birth year and Death year. However, the information needs to be parsed properly, which took longer than expected due to the difficulty of applying regular expressions. Almost all observations are formatted as the example "First Name(1889–1964) MP for United Province [...]". Therefore, we need to extract everything before the first bracket, the first number after the first bracket and the second number after the first bracket. However, for prime ministers who are still alive, the information is structured like "Narendra Modi(born 1950)". Coming up with regex statements or conditional statements that would correctly parse this text into three distinct pieces of information proved to be surprisingly difficult.

Given these challenges, I adopted a pragmatic approach, splitting the dataset into two categories: one for deceased prime ministers and another for those still alive. Subsequently, I applied distinct regex statements to each dataset, with the intention of reuniting them subsequently. For prime ministers still alive, the death year was artificially set to 2024 for visualization purposes, while ensuring the the `Alive` variable is still set to `TRUE` to avoid any potential conclusion. this should not cause any further confusion. The resulting cleaned dataset (Table 1) closely resembled our earlier simulated table.

4 Interpret and Visualize Data

The data analysis process became most enjoyable once the data was satisfactorily cleaned and could finally be analyzed appropriately. The results revealed intriguing insights into the lifespans of Indian prime ministers. There are definitely variations in lifespan, however they seem to be non-systematic, i.e. there is not a difference between recent and non-recent prime ministers. This is most likely due to the fact that all Prime Ministers assumed office after 1948. Interestingly, 6 out of the 16 prime ministers (38%) aged 90 or older, which is not common in a country with a life expectancy of 70 years. The median and mean ages amongst these prime ministers were 81.5 and 80.375, respectively (Table 2). It would be an interesting point of further inferential analysis how extreme these results are, i.e. how much different the sample of prime ministers is compared to a national sample. The most enjoyable part, then, was the data visualization. The clean dataset allowed for clean visualization. The only complication was the need to make the x-axis continuous and change the scale of years, but after some tinkering, I could achieve a satisfactory result. Figure 1 shows the different lifespans of all Indian Prime Ministers.

Table 1: Cleaned Data of Indian Prime Ministers since 1948

Alive	Name	Birth_Year	Death_Year	Age
TRUE	H. D. Deve Gowda	1933	2024	91
TRUE	Manmohan Singh	1932	2024	92
TRUE	Narendra Modi	1950	2024	74
FALSE	Jawaharlal Nehru	1889	1964	75
FALSE	Gulzarilal Nanda	1898	1998	100
FALSE	Lal Bahadur Shastri	1904	1966	62
FALSE	Indira Gandhi	1917	1984	67
FALSE	Morarji Desai	1896	1995	99
FALSE	Charan Singh	1902	1987	85
FALSE	Indira Gandhi	1917	1984	67
FALSE	Rajiv Gandhi	1944	1991	47
FALSE	Vishwanath Pratap Singh	1931	2008	77
FALSE	Chandra Shekhar	1927	2007	80
FALSE	P. V. Narasimha Rao	1921	2004	83
FALSE	Atal Bihari Vajpayee	1924	2018	94
FALSE	Inder Kumar Gujral	1919	2012	93

Table 2: Summary Statistics of Ages of Indian Prime Ministers since 1948

mean	median	sd
80.38	81.5	14.71

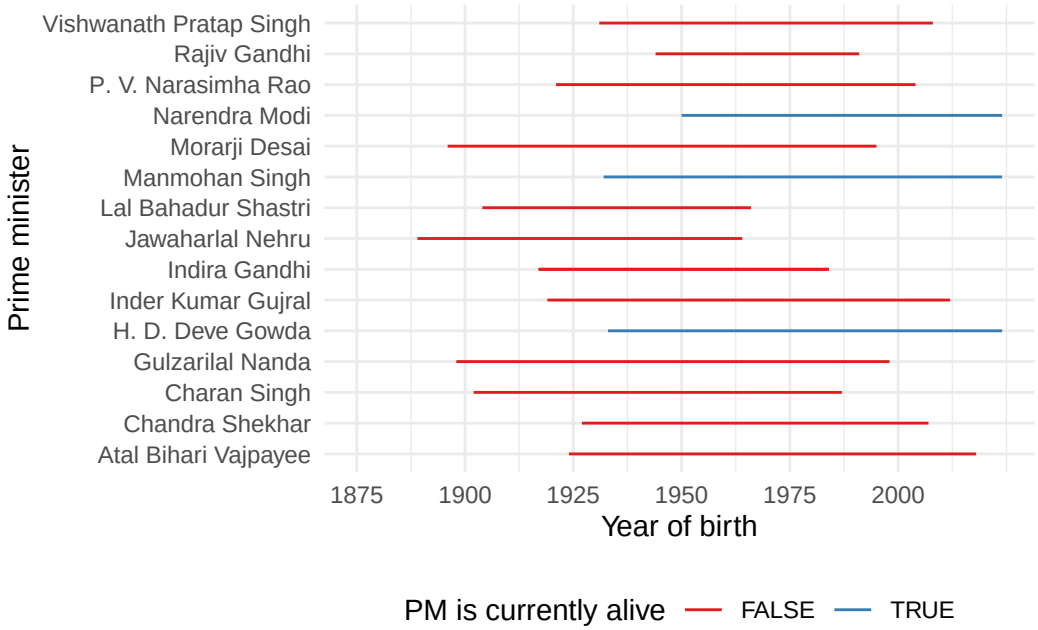

Figure 1: Visualization of the Lifespan of Indian Prime Ministers since 1948

5 Appendix

```
raw_data <-
  read_html(
    "https://en.wikipedia.org/wiki/List_of_prime_ministers_of_India"
  )
write_html(raw_data, "pms.html")
```

```
NULL
```

References

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. 2021. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

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.

Wikipedia. 2024. "List of Prime Ministers of India." https://en.wikipedia.org/wiki/List_of_prime_ministers_of_India.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.