Analyzing life-span of individuals on the list of prime ministers in Australia*

Owais Zahid

February 6, 2024

This paper provides a detailed account of working with, cleaning and presenting the data contained in the List of prime ministers of Australia web page.

1 Introduction

Disclaimer: packages used were Tidyverse (Wickham et al. 2019), Janitor (Firke 2023), the R programming language (R Core Team 2022) and the RStudio platform (Posit team 2023).

I decided to scrape the List of prime ministers of Australia wikipedia page using the rvest package (Wickham 2022). As a first step, I visited the page to inspect the layout in case there were multiple tables because I suspected that would cause me trouble during extracting the html content using the aforementioned package and luckily, there was just one table. It is worth noting that this is not a scalable strategy when dealing with hundreds or even dozens of web-pages.

Cleaning the data that I extracted from the page took me the longest time. More specifically, upon extracting web page contents using the read_html function I was met with a table containing duplicate rows and columns with duplicate names as shown below:

```
# A tibble: 108 x 12
   No.   Portrait Name(Birth-Death)Con~1 `Election(Parliament)` `Term of office`
   <chr> <chr>    <chr>                  <chr>                  <chr>
 1 No.   "Portra~ Name(Birth-Death)Cons~ Election(Parliament)   Took office
 2 1     ""       Edmund Barton(1849-19~ 1901 (1st)             1 January1901
 3 1     ""       Edmund Barton(1849-19~ 1901 (1st)             1 January1901
 4 1     ""       Edmund Barton(1849-19~ 1901 (1st)             1 January1901
```

*Code and data are available at: https://github.com/FFFiend/Australian_PM_Lifespan

```
 5 2     ""       Alfred Deakin(1856-19~ - (1st)                24 September1903
 6 2     ""       Alfred Deakin(1856-19~ 1903 (2nd)             24 September1903
 7 2     ""       Alfred Deakin(1856-19~ 1903 (2nd)             24 September1903
 8 3     ""       Chris Watson(1867-194~ - (2nd)                27 April1904
 9 4     ""       George Reid(1845-1918~ - (2nd)                18 August1904
10 (2)   ""       Alfred Deakin(1856-19~ - (2nd)                5 July1905
# i 98 more rows
# i abbreviated name: 1: `Name(Birth-Death)Constituency`
# i 7 more variables: `Term of office` <chr>, `Term of office` <chr>,
#   Politicalparty <chr>, Ministry <chr>, Monarch <chr>,
#   `Governor-General` <chr>, Ref. <chr>
```

2 About the data

We can see that a few president's served a term in office more than once due to the existence of duplicate names within the dataframe. We can also see that the extracted table itself is poorly formatted, with improper indexing that is duplicate and inconsistent. Furthermore, there are three columns with the same name "Term of office" which should really be more like "start of term", "end of term" and "term duration". We also have data on each candidate's political party, and the ministry, monarch and governer general that they served under.

3 Data Cleaning

I tried running distinct() on the entire table in the hopes of extracting all unique entries but when that didn't work, I decided to use dataframe indexing to remove the duplicate row at the top, and then grabbed the 3rd column containing all the president names along with their birth and death years to produce a column containing only distinct entries as follows:

```
# A tibble: 31 x 1
   `Name(Birth-Death)Constituency`
   <chr>
 1 Edmund Barton(1849-1920)MP for Hunter, NSW
 2 Alfred Deakin(1856-1919)MP for Ballaarat, Vic[a]
 3 Chris Watson(1867-1941)MP for Bland, NSW
 4 George Reid(1845-1918)MP for East Sydney, NSW
 5 Andrew Fisher(1862-1928)MP for Wide Bay, Qld
 6 Joseph Cook(1860-1947)MP for Parramatta, NSW
 7 Billy Hughes(1862-1952)MP for West Sydney, NSW (until 1917)MP for Bendigo, V~
 8 Stanley Bruce(1883-1967)MP for Flinders, Vic
 9 James Scullin(1876-1953)MP for Yarra, Vic
```

```
10 Joseph Lyons(1879-1939)MP for Wilmot, Tas
# i 21 more rows
```

At this point all I had to do was parse each entry in the column for the names, birth and death years and produce a table displaying the same data, albeit in distinct columns along with an Age at death column. I proceeded as follows:

```
[1] "Chris Watson"
```

```
[1] "1867-1941"
```

```
[1] "MP for Bland, NSW"
```

I extracted the contents of each row in the 1-column dataframe we obtained previously as shown above, discarded the content after the birth and death years and then created the final table below (using the Knitr (Xie 2023) package):

Table 1: Final table showcasing when each prime minister was born, and died along with their age at death.

Prime Minister	Birth year	Death year	Age at death
Edmund Barton	1849	1920	71
Alfred Deakin	1856	1919	63
Chris Watson	1867	1941	74
George Reid	1845	1918	73
Andrew Fisher	1862	1928	66
Joseph Cook	1860	1947	87

Learning how to create new columns using values from a pre-existing column, as well as separating on strings was a lot of fun as it reminded me of my earlier days of programming in Python, however I would do a couple things differently if I were to repeat this workflow to accomplish a similar task in the future:

- Skim through the documentation for common libraries before starting the paper, as this saves time and helps one focus on the content and data analysis instead of looking up functions and syntax.
- Get some more practice with using R pipes as they're a useful tool and cut down on the amount of code you need to write.

References

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

Posit team. 2023. *RStudio: Integrated Development Environment for r*. Boston, MA: Posit Software, PBC. http://www.posit.co/.

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. 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.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.