Zepeda Week 5 Exercise

PA 466 - Week 5 Exercise

By: Austin Zepeda

Extract a table of articles from the NY Times

Go to the web link https://www.nytimes.com/topic/subject/elections. Extract data from the "Latest" section and create a table comprising three variables (Article Name, Article Description (the short text after the title), and Author Name). Extract information on at least ten articles. Ideally, the table will have three columns and at least ten rows. Make sure to include both the source code and output in your final submission document. [30 pts]

Code

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(rvest)
```

```
Attaching package: 'rvest'

The following object is masked from 'package:readr':

    guess_encoding
```

```
url = "https://www.nytimes.com/section/us/elections"
html = read_html(url)

section = html |>
  html_elements("ol") |>
  head(10)

#use section vector to parse table data

article_name = section |>
  html_elements("h3")
article_name = gsub("<[^>]+>", "", article_name)
length(article_name) # 10 articles
```

```
[1] 10
```

```
article_description = section |>
  html_elements("p.css-1pga48a.e15t083i1") |>
  html_text2()

author_name = section |>
  html_elements("p.css-1y3ykdt.e140qd2t0") |>
  html_text2()
author_name = gsub("^By\\s+", "", author_name)
length(author_name) # only 5, which will mess up the table.
```

```
[1] 5
```

```
# I used ChatGPT to help with this. Prompt: Let's try looking in the broader
# article class "article.css-1l4spti" and if there are authors listed
# "p.css-1y3ykdt.e140qd2t0", then parse them. Otherwise, NA.

# Response:

# Step 1: grab each article
articles <- section %>% html_elements("article.css-1l4spti")

# Step 2: for each article, try to extract the author
author <- sapply(articles, function(article) {
  el <- html_element(article, "p.css-1y3ykdt.e140qd2t0")
  if (is.null(el)) NA_character_ else html_text(el)
})

length(author) # good.
```

```
[1] 10
```

```
author = gsub("^By\\s+", "", author)


ny_times = tibble(
  article_name,
  article_description,
  author
)
ny_times
```

```
# A tibble: 10 x 3
   article_name                                      article_description author
   <chr>                                             <chr>               <chr>
 1 Mamdani's Plan for Government-Run Grocery Stores:~ New York City's De~ Alex ~
 2 Arizona Seventh Congressional District Special El~ Get live results a~ <NA>
 3 Trump Remembers Kirk as a Martyr While Attacking ~ President Trump os~ Tyler~
 4 Minnesota State House District 34B Special Electi~ Get live results f~ <NA>
 5 Boston Mayoral Primary Election Results           Get live results f~ <NA>
 6 Virginia 11th Congressional District Special Elec~ Get live results a~ <NA>
 7 Inside Gavin Newsom's Redistricting Cash Blitz     Total spending cou~ Theod~
 8 City Council President and Pastor Advance in Detr~ Mary Sheffield and~ Mitch~
 9 Detroit Mayoral Primary Election Results          Get live results a~ <NA>
```

```
10 MyPillow Founder Will Not Pay Winnings for Electi~ A court overturned~ Ali W~
```

```r
# I asked ChatGPT to render a more stylized, readable table in the PDF by
# selecting the first 30 characters of each observation.

ny_times_trimmed <- ny_times %>%
  mutate(
    article_name = substr(article_name, 1, 30),
    article_description = substr(article_description, 1, 30),
    author = substr(author, 1, 30)
  )

library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```r
ny_times_trimmed %>%
  kable("latex", booktabs = TRUE, caption = "NY Times Election Articles") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

Table 1: NY Times Election Articles

article_name	article_description	author
Mamdani's Plan for Government-	New York City's Democratic may	Alex Pena, Daniel Vergara, Mar
Arizona Seventh Congressional	Get live results and maps from	NA
Trump Remembers Kirk as a Mart	President Trump oscillated bet	Tyler Pager
Minnesota State House District	Get live results from the spec	NA
Boston Mayoral Primary Electio	Get live results from the 2025	NA
Virginia 11th Congressional Di	Get live results and maps from	NA
Inside Gavin Newsom's Redistri	Total spending could top $200	Theodore Schleifer, Shane Gold
City Council President and Pas	Mary Sheffield and Solomon Kin	Mitch Smith
Detroit Mayoral Primary Electi	Get live results and maps from	NA
MyPillow Founder Will Not Pay	A court overturned a previous	Ali Watkins