STAT 231: Problem Set 5B

Sean Wei

due by 5 PM on Friday, October 2

This homework assignment is designed to help you futher ingest, practice, and expand upon the material covered in class over the past week(s). You are encouraged to work with other students, but all code and text must be written by you, and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps5B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps5B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER:

1. Justices of the Supreme Court of the United States

a. Confirm (using an R command) that the following Wikipdeia page allows automated scraping: https://en.wikipedia.org/wiki/List_of_justices_of_the_Supreme_Court_of_the_United_States

```
url <- "https://en.wikipedia.org/wiki/List_of_justices_of_the_Supreme_Court_of_the_United_States"
paths_allowed(url)
```

```
##  en.wikipedia.org
```

```
## [1] TRUE
```

b. Go to the List of Justices of the Supreme Court of the United States and scrape the table for the Justices. Write, test, and save your code in an R script called `scrape_justices.R`, and write the data frame out to a csv file called `justices.csv` using the `write_csv` function.

Be sure to push your .R and .csv files to your GitHub repo.

```
## Add your code that is in justices.R to this code chunk.  KEEP the "eval=FALSE" option in this code c
url <- "https://en.wikipedia.org/wiki/List_of_justices_of_the_Supreme_Court_of_the_United_States"
justices <- url %>%
  read_html() %>%
  html_nodes("table")
justices <- html_table(justices[[2]], fill = TRUE)

path <- "/Users/seanwei/Desktop/STAT231-swei1999/homeworks"
write_csv(x = justices, path = paste0(path, "/justices.csv"))
```

c. Load `justices.csv` into this file using the `read_csv` function. Then, run the code given below to create the variable `tenure_length` (a numeric variable containing each justice's tenure on the bench).

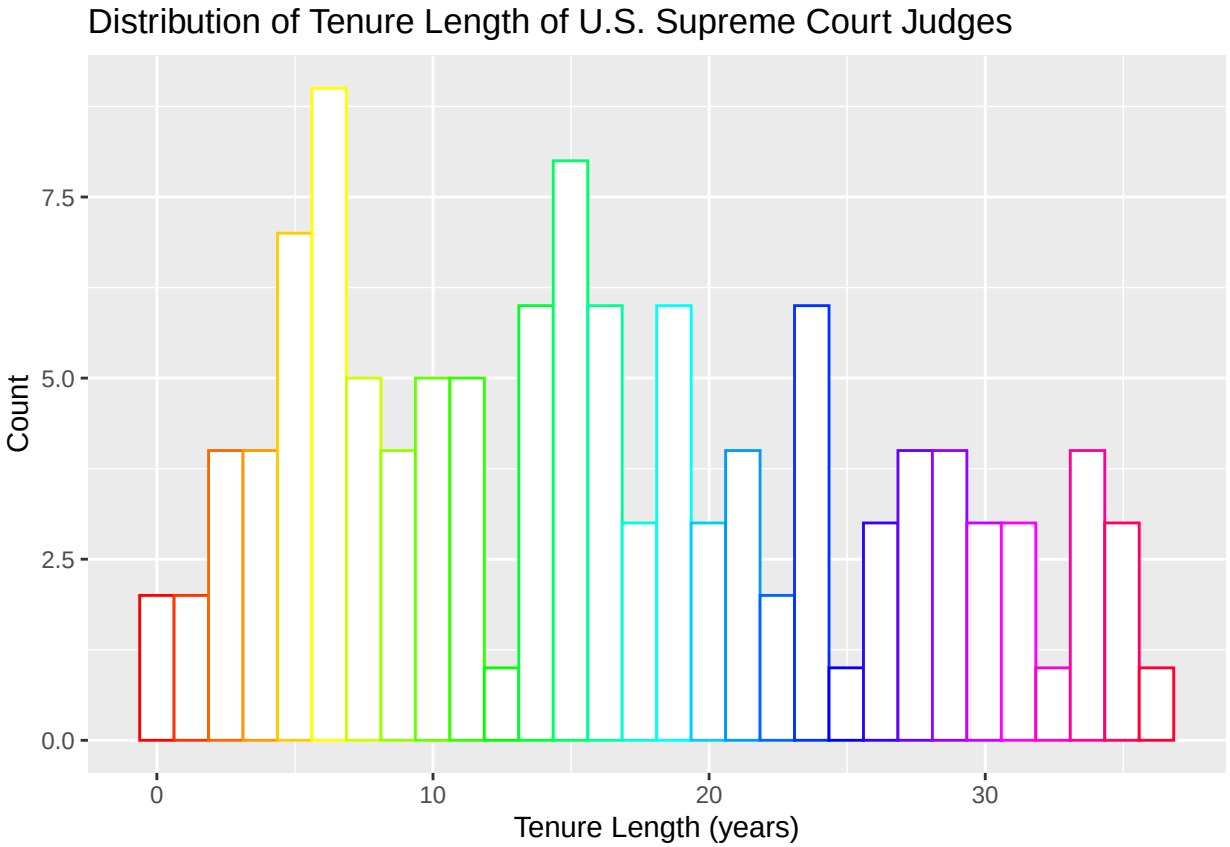
Create a visualization to show the distribution of tenure length of U.S. Supreme Court judges. Interpret the plot.

ANSWER: The histogram below shows the distribution of tenure length of U.S. Supreme Court judges. From the plot, there seems to be around 2 clear peaks, which conveys that the majority of judges are tenured for around 5 and 15 years.

```
justices <- read_csv("/Users/seanwei/Desktop/STAT231-swei1999/homeworks/ps5b/justices.csv")

justices2 <- justices %>%
  clean_names() %>%
  # remove extra line that comes in at end of table
  filter(justice != "Justice") %>%
  # some justices served less than 1 year, adjust their length so can
  # separate correctly
  mutate(tenure_length_temp = case_when(str_detect(tenure_length_d, "year") ~ tenure_length_d
                                        , TRUE ~ paste0("0 years, ", tenure_length_d))) %>%
  separate(tenure_length_temp, into = c("years_char", "days_char"), sep = ","
           , remove = FALSE) %>%
  mutate(tenure_length = parse_number(years_char) + (parse_number(days_char)/365)) %>%
  # create date confirmed as date variable
  separate(date_confirmed_vote, into = c("date_confirmed_vote", "extra")
           , sep = "\\(") %>%
  mutate(date_confirmed = lubridate::mdy(date_confirmed_vote))
```

```
# Distribution of tenure length of U.S. Supreme Court judges
ggplot(data = justices2, aes(x = tenure_length)) +
  geom_histogram(color = rainbow(n = 30), fill = "white") +
  labs(x = "Tenure Length (years)", y = "Count") +
  ggtitle("Distribution of Tenure Length of U.S. Supreme Court Judges")
```


Distribution of Tenure Length of U.S. Supreme Court Judges
Count
7.5
5.0
2.5
0.0
0
10
20
30
Tenure Length (years)

2. Brainy Quotes

a. Confirm (using an R command) that automated scraping of the Brainy Quote webpage (https://www.brainyquote.com/) is allowed.

```
url <- "https://www.brainyquote.com/"
paths_allowed(url)
```

```
##  www.brainyquote.com
```

```
## [1] TRUE
```

b. Life can get frustrating at times. Like when we're trying to Zoom and our internet cuts out. Or when we can't figure out why R's throwing an error when we try to clone a GitHub repo in RStudio. Or, when COVID-19 upends life as we knew it. In these times, it can't hurt to be reminded of the power of persistence, resilience and optimism.

The code in the first R code chunk below scrapes the first 40 quotes returned from a search for "resilience" on BrainyQuote.com. (Do NOT remove the "eval = FALSE" option from that code chunk; you do not want it to evaluate it, i.e. scrape the site, every time you knit this file.)

The code in the second R code chunk below randomly selects a quote and prints it. When you're feeling frustrated, run that code chunk to randomly generate a quote to lift you up (or just make you laugh at the uselessness of the quote; some of them are pretty pathetic . . .).

Note that CSS selector gadget was used to identify the key words to specify in the `html_nodes` function (i.e. ".oncl_q" and ".oncl_a"). These key words will vary depending on what webpage and what particular objects from that webpage you're trying to scrape.

```
quotes_html <- read_html("https://www.brainyquote.com/topics/resilience-quotes")

quotes <- quotes_html %>%
  html_nodes(".oncl_q") %>%
  html_text()

person <- quotes_html %>%
  html_nodes(".oncl_a") %>%
  html_text()

# put in data frame with two variables (person and quote)
quotes_dat <- data.frame(person = person, quote = quotes
                         , stringsAsFactors = FALSE) %>%
  mutate(together = paste('"', as.character(quote), '" --'
                          , as.character(person), sep=""))
```

```
quotes_dat <- read_csv("http://kcorreia.people.amherst.edu/F2021/resilience_quotes.csv")
```

```
## Parsed with column specification:
## cols(
##   person = col_character(),
##   quote = col_character(),
##   together = col_character()
## )
```

```
quote_for_the_day <- quotes_dat[sample(1:nrow(quotes_dat), size = 1),]

quote_for_the_day$together
```

```
## [1] "\"Resilience is woven deeply into the fabric of
Oklahoma. Throw us an obstacle, and we grow stronger.\"
--Brad Henry"
```

Go to BrainyQuote.com and search a different topic (or search an Author) that interests you. Scrape the webpage returned from your search following the same code given above. Save your code in an R script

called `scrape_quotes.R`, and write the data frame out to a csv file called `quotes.csv` using the `write_csv` function.

Be sure to push your .R and .csv files to your GitHub repo.

```
## Add your code that is in quotes.R to this code chunk.  KEEP the "eval=FALSE" option in this code chu
humor_quotes <- read_html("https://www.brainyquote.com/topics/humor-quotes")

quotes <- humor_quotes %>%
  html_nodes(".oncl_q") %>%
  html_text()

person <- humor_quotes %>%
  html_nodes(".oncl_a") %>%
  html_text()

# put in data frame with two variables (person and quote)
humor_quotes_final <- data.frame(person = person, quote = quotes
                         , stringsAsFactors = FALSE) %>%
  mutate(together = paste('"', as.character(quote), '" --'
                          , as.character(person), sep=""))

path <- "/Users/seanwei/Desktop/STAT231-swei1999/homeworks/ps5B"
write_csv(x = humor_quotes_final, path = paste0(path, "/quotes.csv"))
```

c. Load `quotes.csv` into this file using the `read_csv` function. Write code to select *three* of the quotes at random and print them (i.e., set `size = 3` in the `sample` function).

```
set.seed(1823489)
humor_quotes_final <- read_csv("/Users/seanwei/Desktop/STAT231-swei1999/homeworks/ps5b/quotes.csv")
sample(humor_quotes_final$together, 3, replace = FALSE)
```

```
## [1] "\"In conversation, humor is worth more than wit and
easiness more than knowledge.\" --George Herbert"
## [2] "\"The problem with having a sense of humor is often
that people you use it on aren't in a very good mood.\"
--Lou Holtz"
## [3] "\"From there to here, and here to there, funny
things are everywhere.\" --Dr. Seuss"
```