Problem Set 05

WRITE YOUR NAME HERE

2021-10-25

Learning goals

- More data wrangling
- Recreate visualizations from the data journalism website FiveThirtyEight.com

Setup

Load necessary packages:

```
library(ggplot2)
library(dplyr)
library(fivethirtyeight)
# For reading CSV spreadsheet files:
library(readr)
```

Question 1: Honor code

For this problem set I worked with:

Question 2: Age of Congress

In this question, you'll be analyzing the age of members of the United States congress over the years. The data of interest is saved in the `congress_age` data frame included in the `fivethirtyeight` package. To understand the data's context, first read:

- The original FiveThirtyEight article the data was used in. In particular the first visualization titled "Average Age of Members of Congress"
- The help file for the `congress_age` data frame by running `?congress_age` in the console.

a) Data wrangling

Take the `congress_age` data frame and perform the data wrangling necessary to create the first visualization in the article. Save the output in a data frame called `avg_congress_age`. Hint: `avg_congress_age` should have 68 rows and 3 variables: `termstart`, `party`, and `mean_age`.

```r
avg_congress_age <- congress_age %>%
  filter(party == "D" | party == "R") %>%
  group_by(termstart, party) %>%
  summarize(mean_age = mean(age))
```

b) Data wrangling

Take the `avg_congress_age` data frame and perform the data wrangling necessary to create a new variable `mean_age_months` that has the mean age of congress members in months. Overwrite the original `avg_congress_age` data frame that had 68 rows and 3 variables with this new data frame that has 68 rows and 4 variables.

```r
avg_congress_age <- avg_congress_age %>%
  mutate(mean_age_months = mean_age * 12)
```

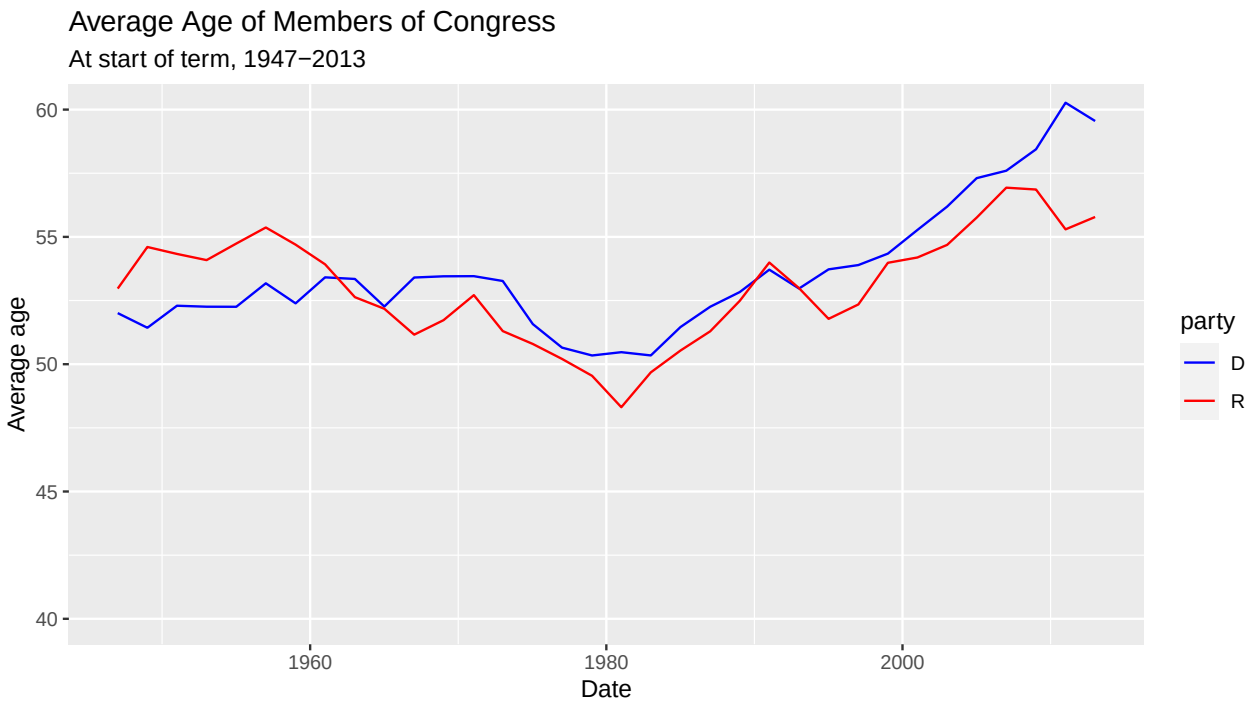
c) Data visualization

Using the `avg_congress_age` data frame, use the `ggplot2` package to recreate the first visualization in the article as follows:

- The values and colors of the lines should match. The legend should be the standard ggplot2 legend.
- The title and subtitle should match. Add the appropriate axes labels.
- Make the y-axis start at 40 and end at 60. Hint: Use google or the ggplot2 cheatsheet seen in ModernDive 2.9.3

In other words, your plot should look like this.

```r
ggplot(avg_congress_age, aes(x=termstart, y = mean_age, col = party)) +
  geom_line() +
  labs(x = "Date", y = "Average age", title = "Average Age of Members of Congress",
       subtitle = "At start of term, 1947-2013") +
  scale_color_manual(values = c("blue", "red")) +
  coord_cartesian(ylim = c(40, 60))
```



Question 3: Titanic

Load the titanic dataset from the internet and take a look at it's contents. It contains demographic information about the 2201 passenengers on the Titanic disaster and information on whether they survived. Note there are 2201 rows in this data, one for each passenger:

```
titanic <- read_csv("https://rudeboybert.github.io/SDS192/static/PS/titanic.csv")
```

a) Overall survival

Using dplyr commands, output a table that displays the counts of survived & died. Your code should print out a table with two columns and two rows of data, along with a "header" row of the variable names.

```
titanic %>%
  group_by(Survived) %>%
  summarize(n = n())
```

```
## # A tibble: 2 x 2
##   Survived     n
##   <chr>    <int>
## 1 No        1490
## 2 Yes        711
```

b) Survival split by sex

Survival split by sex. Using dplyr commands, output a single table that displays the overall survival & death counts of the disaster *split by sex* (as recorded at the time). Your code should print out a table with three columns and four rows of data, along with a "header" row of the variable names.

```
titanic %>%
  group_by(Survived, Sex) %>%
  summarize(n = n())
```

```
## # A tibble: 4 x 3
## # Groups:   Survived [2]
##   Survived Sex        n
##   <chr>    <chr>  <int>
## 1 No       Female   126
## 2 No       Male    1364
## 3 Yes      Female   344
## 4 Yes      Male     367
```

c) BONUS: Passenger ID

Using dplyr commands, output a table that displays only the `passenger_number` of all 17 3rd class female children aboard the ship who died. Your code should print out a table with one column and 17 rows, along with a "header" row of the variable names. Hint: skim through ModernDive Chapter 3 on how to do this.

```
titanic %>%
  filter(Class == "3rd", Age == "Child", Sex == "Female", Survived == "No") %>%
  select(passenger_number)
```

```
## # A tibble: 17 x 1
##    passenger_number
##               <dbl>
##  1               37
##  2              199
##  3              246
##  4              286
##  5              303
##  6              329
##  7              902
##  8             1046
##  9             1240
## 10             1242
## 11             1352
## 12             1374
## 13             1398
## 14             1602
## 15             1661
## 16             1871
## 17             2026
```