Problem Set 03

WRITE YOUR NAME HERE

2021-09-27

Learning goals

- Delve deeper into more data visualizations.
- Get into the habit of labeling elements of visualizations and adding titles. This helps communicate data's *context*.
- Relatedly, making it a point to investigate data's context using whatever information we are provided with. In this problem set's case, by reading *help files*. After all: *numbers are numbers, but data has context*.

Setup

Load necessary packages:

```
library(ggplot2)
library(dplyr)
library(fivethirtyeight)
library(moderndive)
library(lubridate)
```

Getting started

- At the top of this document replace `"WRITE YOUR NAME HERE"` with your name, including the quotation marks. Ex: `"Albert Y. Kim"`
- Knit this file to PDF and read over the questions

Question 1: Honor code

For this problem set I worked with (please indicate even if with no one):

Question 2

Let's analyze the number of campaign stops Clinton and Trump made in the lead up to the 2016 election. This data is in the `pres_2016_trail` data frame included in the `fivethirtyeight` package.

Before we answer any questions, first let's do a brief exploratory data analysis:

1. Read the "help file" associated with this data frame by running `?pres_2016_trail` directly in the Console. Do not include this code in this `.Rmd` file as it might cause an error.
2. Look at the raw values `pres_2016_trail` of by running `View(pres_2016_trail)` directly in the Console. Do not include this code in this `.Rmd` file as it might cause an error.

Now let's take the `pres_2016_trail` data frame and "wrangle" it (i.e. transform it) to count the number of stops both candidates made per week. To do this we'll use code from the `dplyr` package for data wrangling and the `lubridate` package for working with dates and times. Don't worry if you don't understand this code for now, we'll cover it later in this course, in particular ModernDive Chapter 3 on data wrangling.

```
weekly_campaign_stops <- pres_2016_trail %>%
  mutate(week = floor_date(date, unit = "week")) %>%
  group_by(candidate, week) %>%
  summarize(number_of_stops = n())
```

Now look at the raw values of `weekly_campaign_stops` by running `View(weekly_campaign_stops)` directly in the Console. Do not include this code in this `.Rmd` file as it might cause an error.

Part a)

Question: How many rows does `weekly_campaign_stops` have?

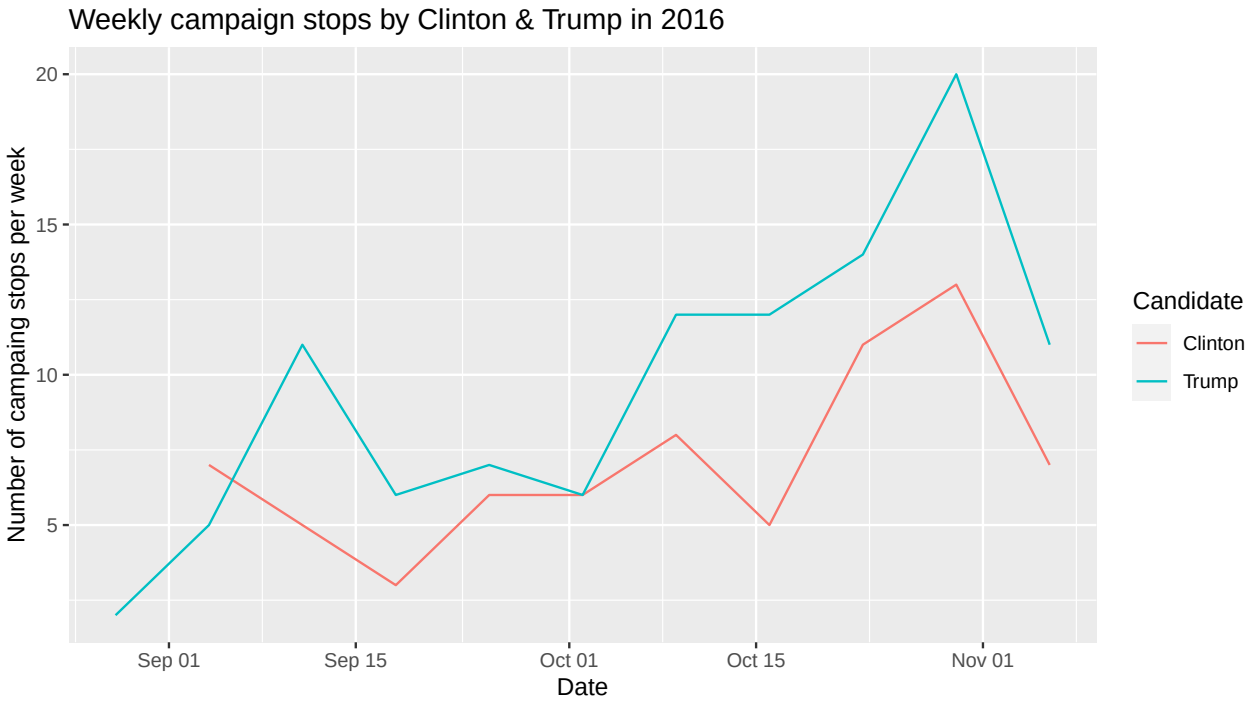
Answer: 21

Part b)

Create a data visualization that allows the reader to compare the number of campaign stops Clinton and Trump made each week leading up to the 2016 election. Note:

1. There is more than one way to do this.
2. Label your axes! Add a title! This helps communicate the data's context to the reader.

```
ggplot(data = weekly_campaign_stops,
       mapping = aes(x = week, y = number_of_stops, col = candidate)) +
  geom_line() +
  labs(
    x = "Date",
    y = "Number of campaing stops per week",
    color = "Candidate",
    title = "Weekly campaign stops by Clinton & Trump in 2016"
  )
```



Part c)

Question: Who made more campaign stops during the election?

Answer: Trump

Question 3

Let's analyze some data on instructor evaluations as given by students at the UT Austin. This data is in the `evals` data frame included in the `moderndive` package.

Before we answer any questions, first let's do a brief exploratory data analysis:

1. Read the "help file" associated with this data frame by running `?evals` directly in the Console. Do not include this code in this `.Rmd` file as it might cause an error.
2. Look at the raw values `evals` of by running `View(evals)` directly in the Console. Do not include this code in this `.Rmd` file as it might cause an error.

Part a)

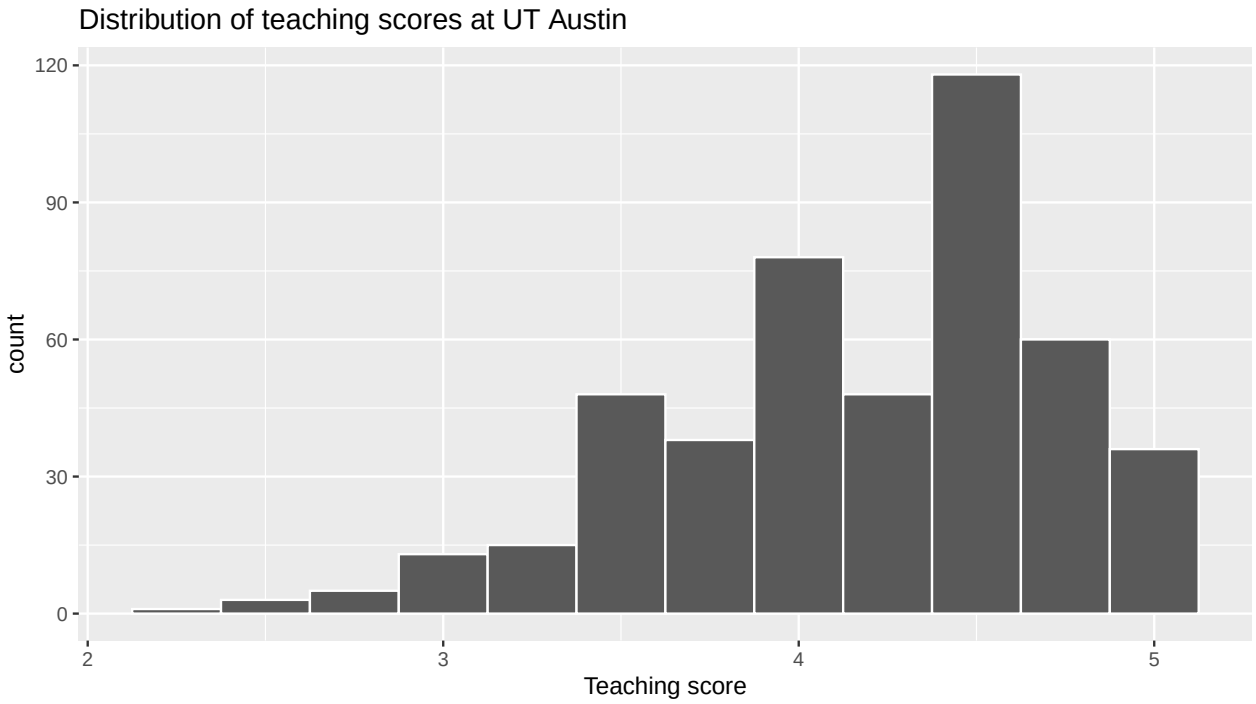
Question: How many unique instructors were considered in this study?

Answer: Looking at the help file for `evals`, even though the data is about 463 courses, there were only 94 unique instructors were considered. This is because instructors often teach more than one class.

Part b)

Visualize the distribution of the numerical variable `score`, which is the teaching score as given by students, with a histogram. Have the bin widths be 0.25 teaching score units.

```
ggplot(data = evals, mapping = aes(x = score)) +
  geom_histogram(binwidth = 0.25, color = "white") +
  labs(x = "Teaching score", title = "Distribution of teaching scores at UT Austin")
```



Part c)

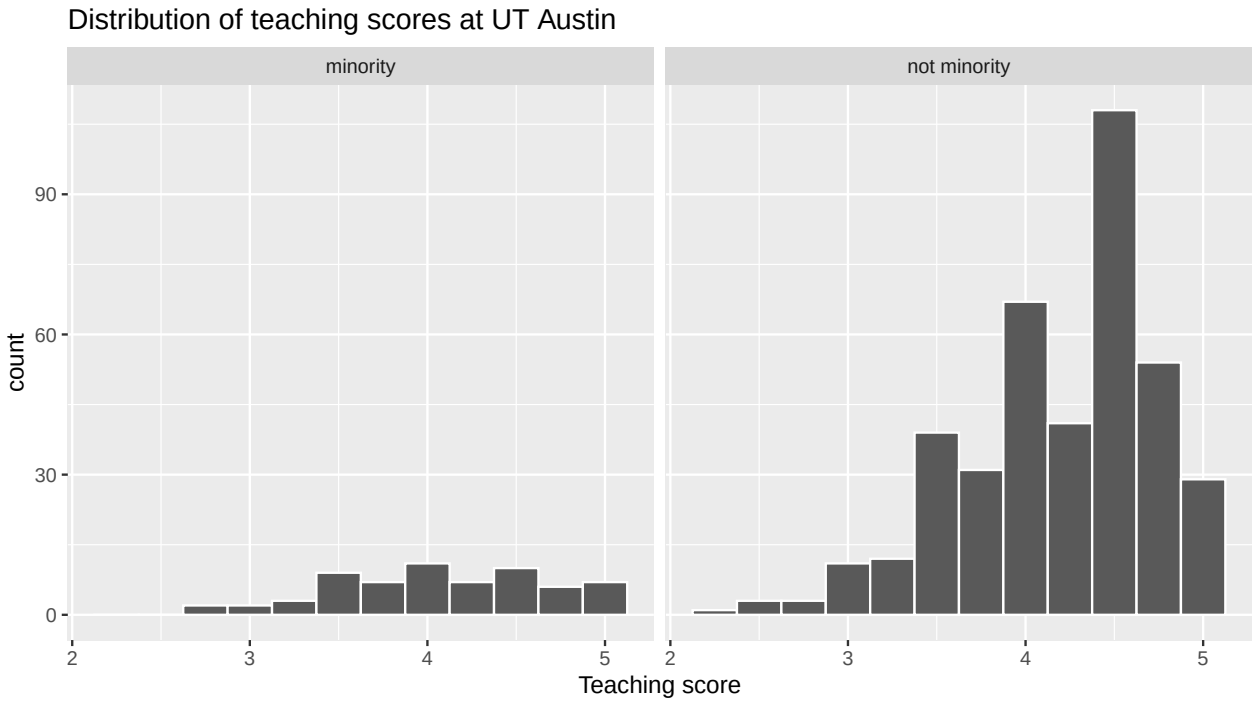
Question: Using only your eyes, what would you say the average teaching score approximately is? You do not compute this value exactly, just visually approximate it. There are no "right" answers to this question, just write down what you think.

Answer: In my opinion (IMO), maybe about 4. Note the data is *left-skewed* because of the tail to the left.

Part d)

Now show the same histogram as above, but minority and non-minority instructors separately. Note you do not need to use any data wrangling to do this, only `ggplot2` code from Chapter 2.

```
ggplot(data = evals, mapping = aes(x = score)) +
  geom_histogram(binwidth = 0.25, color = "white") +
  labs(x = "Teaching score", title = "Distribution of teaching scores at UT Austin") +
  facet_wrap(~ethnicity)
```



Part e)

Question: Using only your eyes, what would you that on average minority or non-minority instructors got higher teaching scores? You do not compute these values exactly, just visually approximate them. There are no "right" answers to this question, just write down what you think.

Answer: Hard to say, especially since there are different numbers of instructors in the two groups. IMO however the non-minorities tended to scores of 4 or higher relatively more often, and thus received on average higher teaching scores.