Lab 2

Taylor Blair

Math 141, Week 2

Due: Before your Week 3 lab meeting

Goals of this lab

1. Practice decomposing graphs into their `geom`s, the `aes`thetics of those `geom`s, and how the variables map to those `aes`thetics.
2. Reflect on the editorial choices of a graphic and how those choices impact the messages conveyed by the graphic.
3. Practice creating and interpreting `ggplot2` graphs.

Problems

- For each problem, put your solution between the bars of stars.
- For this lab, you don't need to worry about labels and a title for your plots.
- Run the following chunk to load the necessary packages.

```
# Load the necessary packages
library(tidyverse)
library(NHANES)
```

For Problems 1 - 3, we will use the `NHANES` dataset which can be found in the `NHANES` package. Make sure to run the following chunk to load the dataset.

```
# Load NHANES dataset
data(NHANES)
```

Problem 1

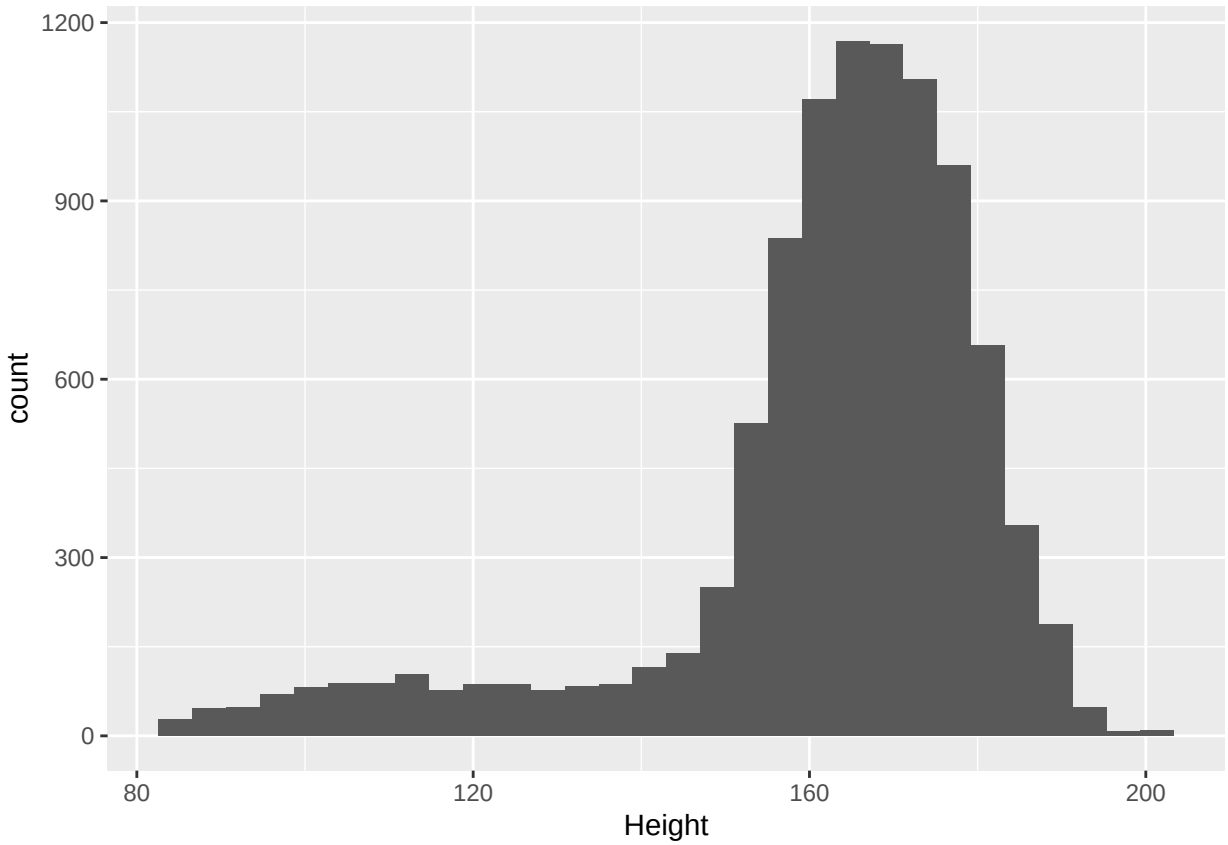
Explore the data a bit. (Leave `eval = FALSE` in the following R chunk. Otherwise, the document may not knit.)

```
# Pull up a data viewer window
View(NHANES)

# Codebook for the variables can be found in the help file
?NHANES
```

a. Fill in the blanks (`___`) to create a histogram of the variable `Height`.

```
# Histogram of height
ggplot(data = NHANES, mapping = aes(x = Height)) +
  geom_histogram()
```



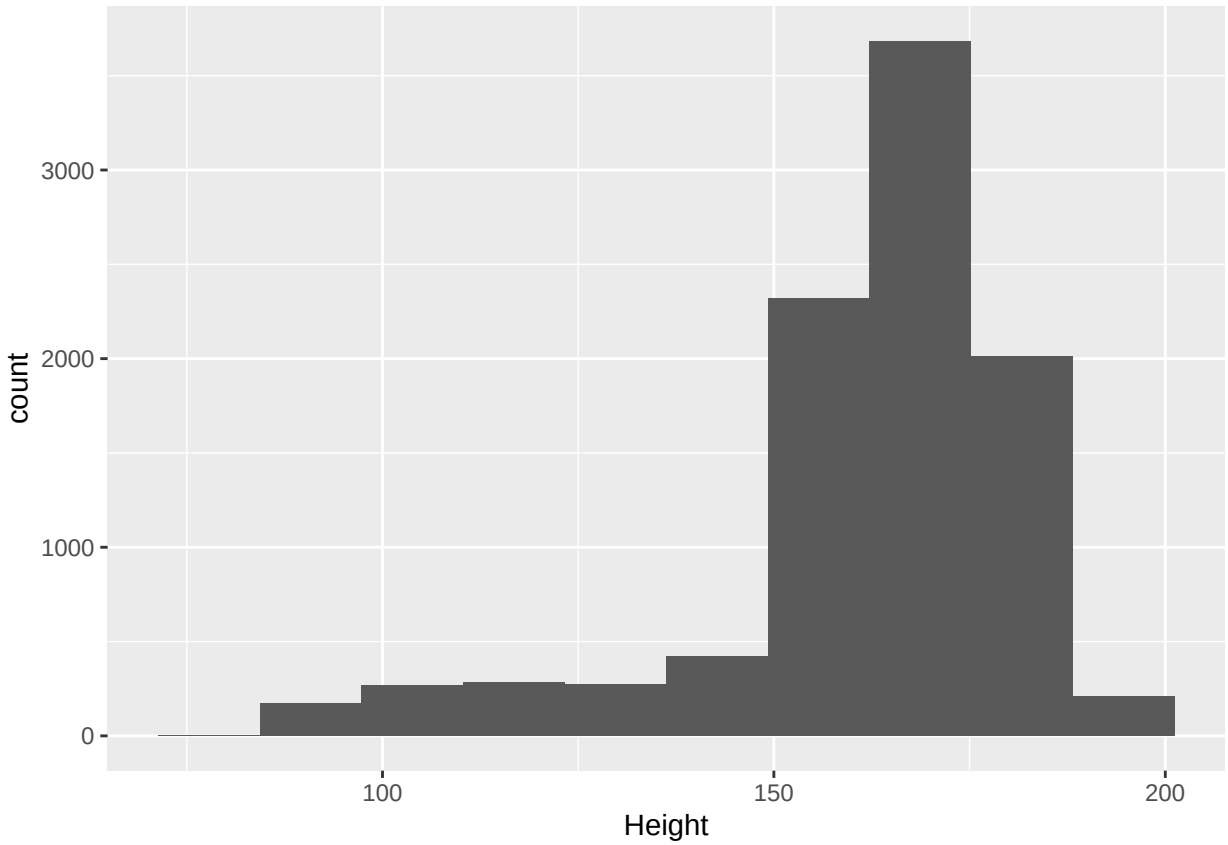
b. Describe the shape of the histogram of `Height`. Explain why this shape makes sense for these data.

- Skewed left
- Centered around ~170
- Max count of 1150
- Bins of approx 4 width or 30 bins?
- Smaller spread
- Makes sense as the general population is going to have a center for height, there will be a max and min height. Afterall people aren't 4 meteres tall

---

c. Look over the messages (printed in red) that appeared when you generated the plot. They should have appeared in the Console or right below the R chunk. The first message reminds us to think about how the data are binned in the histogram. The second message tells us how many values were removed (because of missing values). Let's recreate the histogram but this time set `bins = 10`. Comment on how that changed the histogram. Do you think 10 or 30 bins are better in this case? Justify your answer.

```
ggplot(data = NHANES, mapping = aes(x = Height)) +
  geom_histogram(bins=10)
```



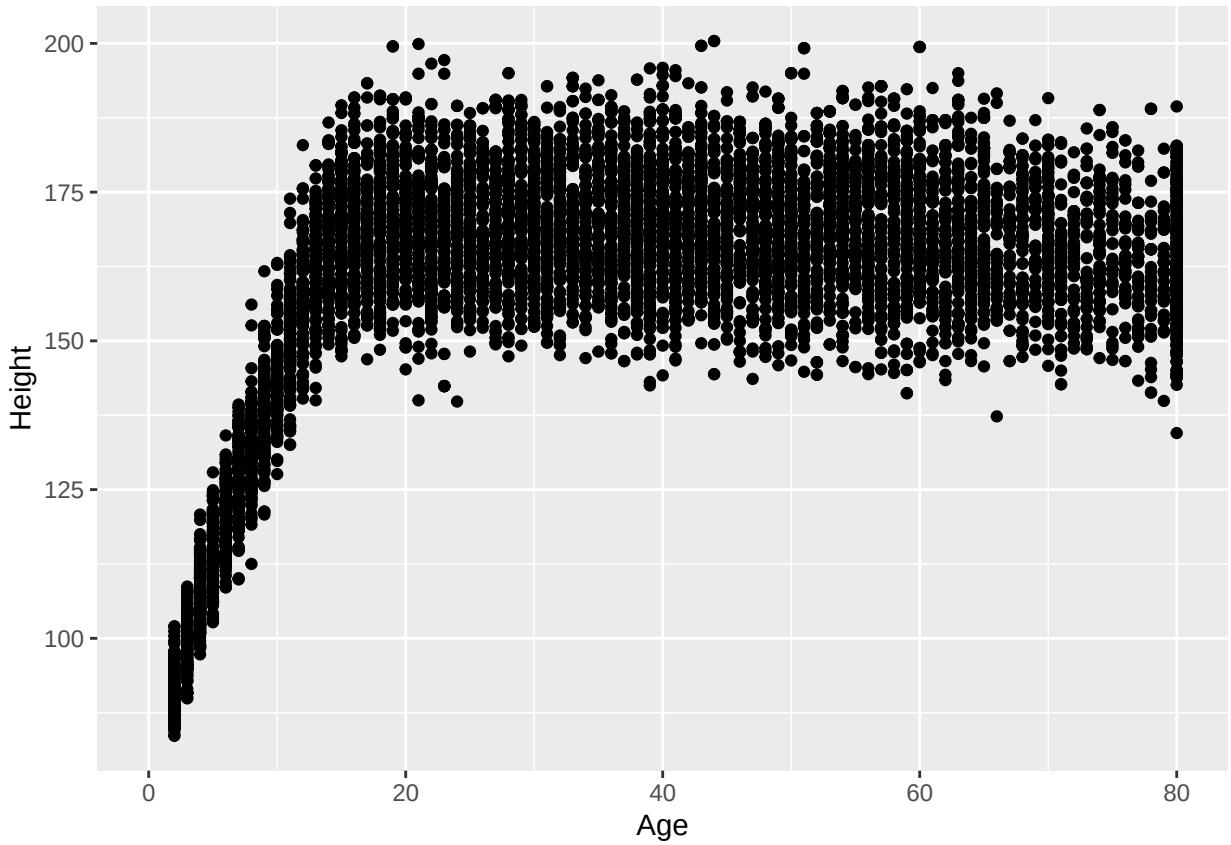
- Although fewer bins does make it easier to visualize the mean and mode, it lacks information on the spread

---

Problem 2

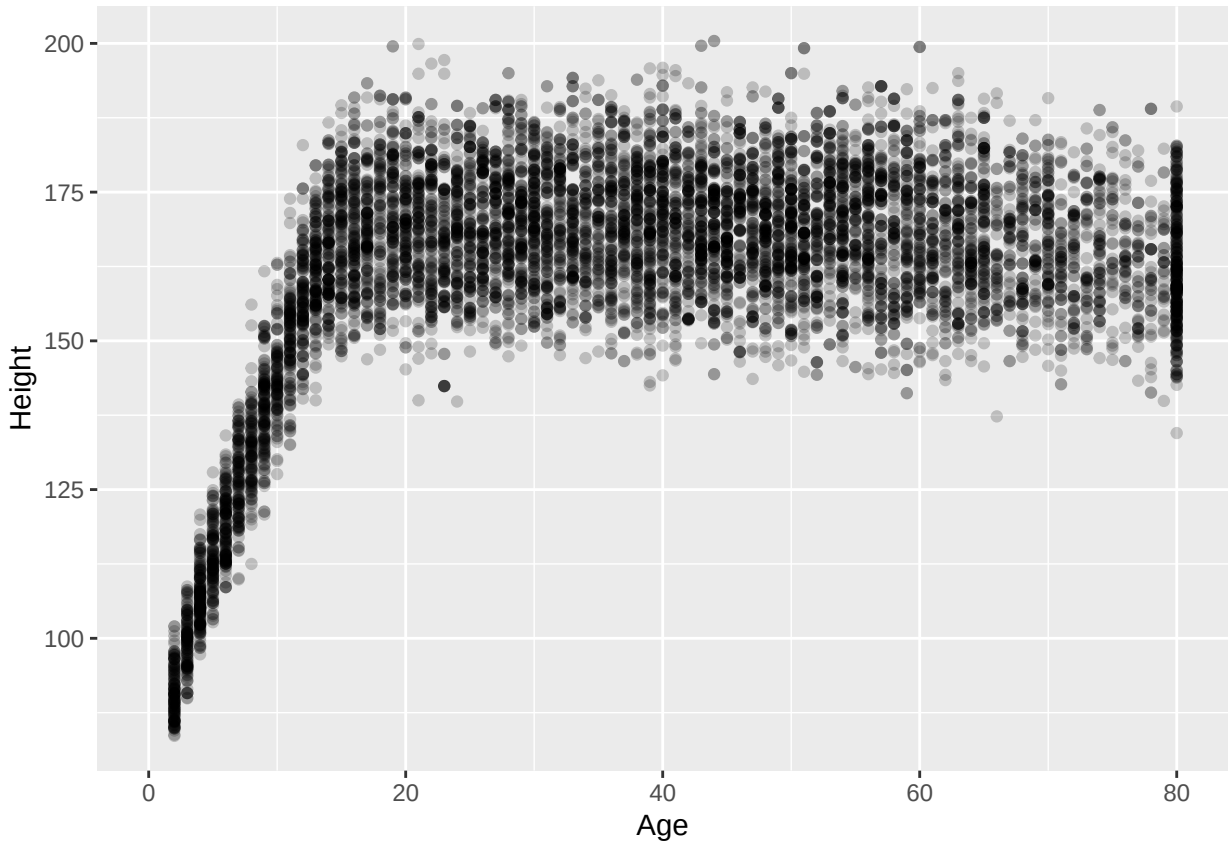
a. Construct a scatterplot with `Age` on the x-axis and `Height` on the y-axis.

```
ggplot(data = NHANES, mapping = aes(x = Age, y = Height)) +
  geom_point()
```



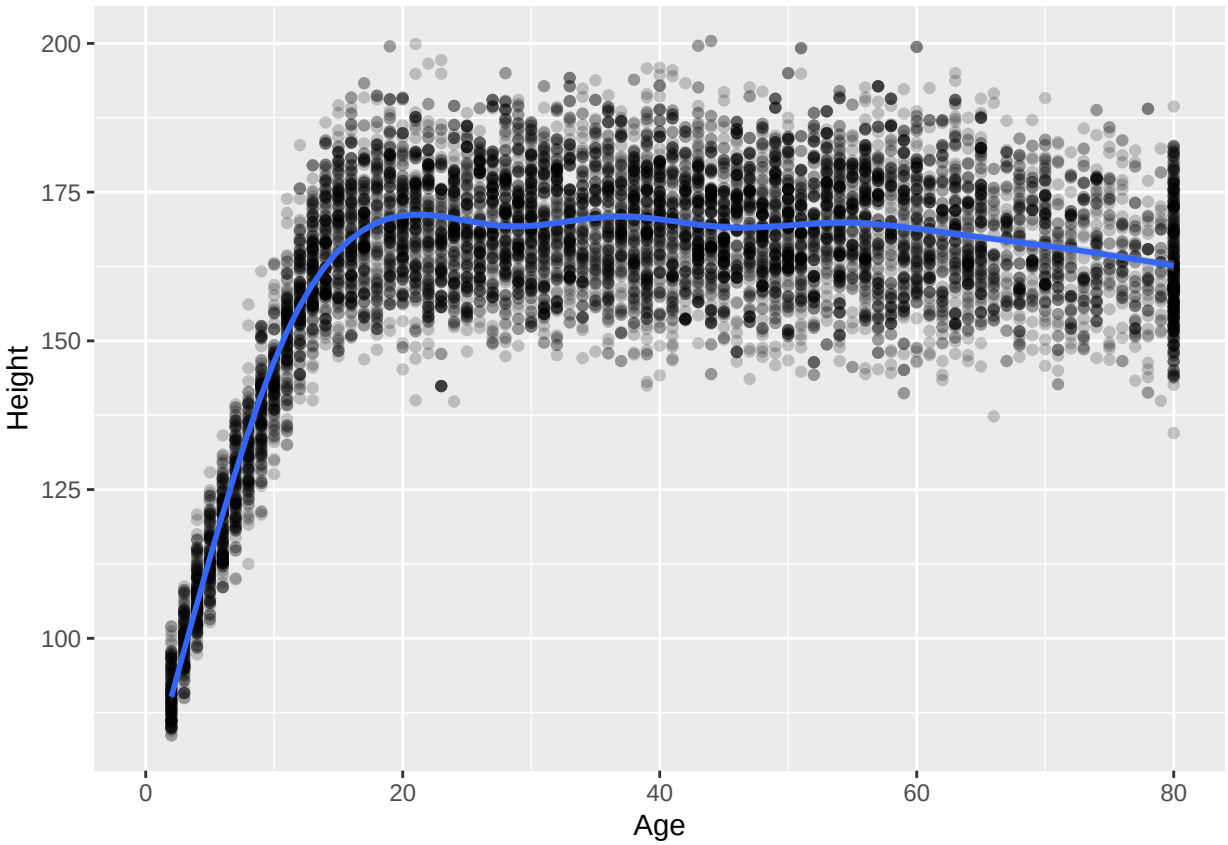
b. Recreate the scatterplot but this time modify the plot to combat the overfitting issue.

```
ggplot(data = NHANES, mapping = aes(x = Age, y = Height)) +
  geom_point(alpha=0.2)
```



c. The geom, `geom_smooth()`, will add a smooth curve that tries to capture the general trend. Add this layer to scatterplot.

```
ggplot(data = NHANES, mapping = aes(x = Age, y = Height)) +
  geom_point(alpha=0.2) + geom_smooth()
```



d. Describe and explain the shape.

- The line created is a polyinomial line. Fitted using newtons method?
- Over time it becomes smoother as the spread becomes wider and trend more linear

---

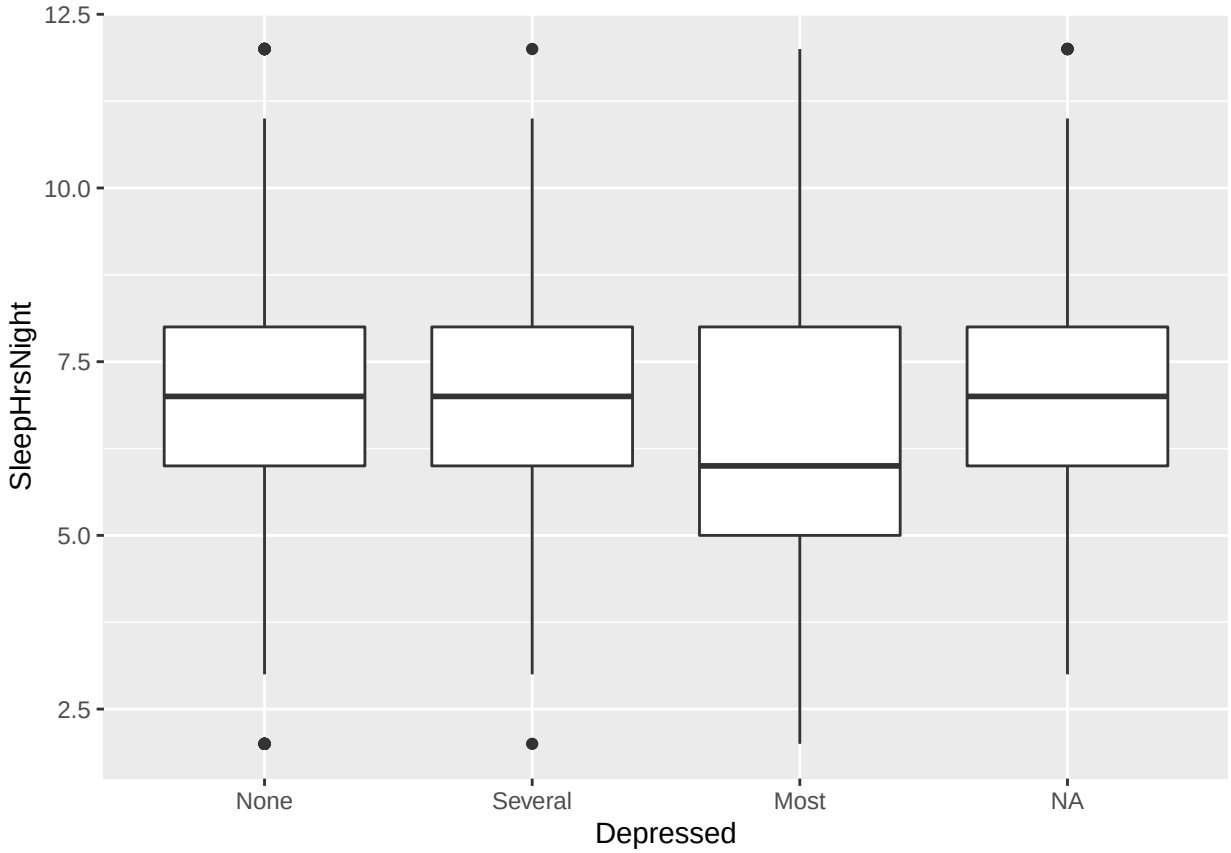
e. Based on the graph, what is the maximum age in the dataset? Why would the researchers round down the ages of the oldest subjects?

- 80 years old
- Anything beyond that was an outlier or the records get fuzzy
- DON'T WANT TO GIVE TO ACCURATE OF AN IDENTITY TO STEAL

Problem 3

a. Construct a side-by-side boxplot of `SleepHrsNight` by the categories of the variable `Depressed`.

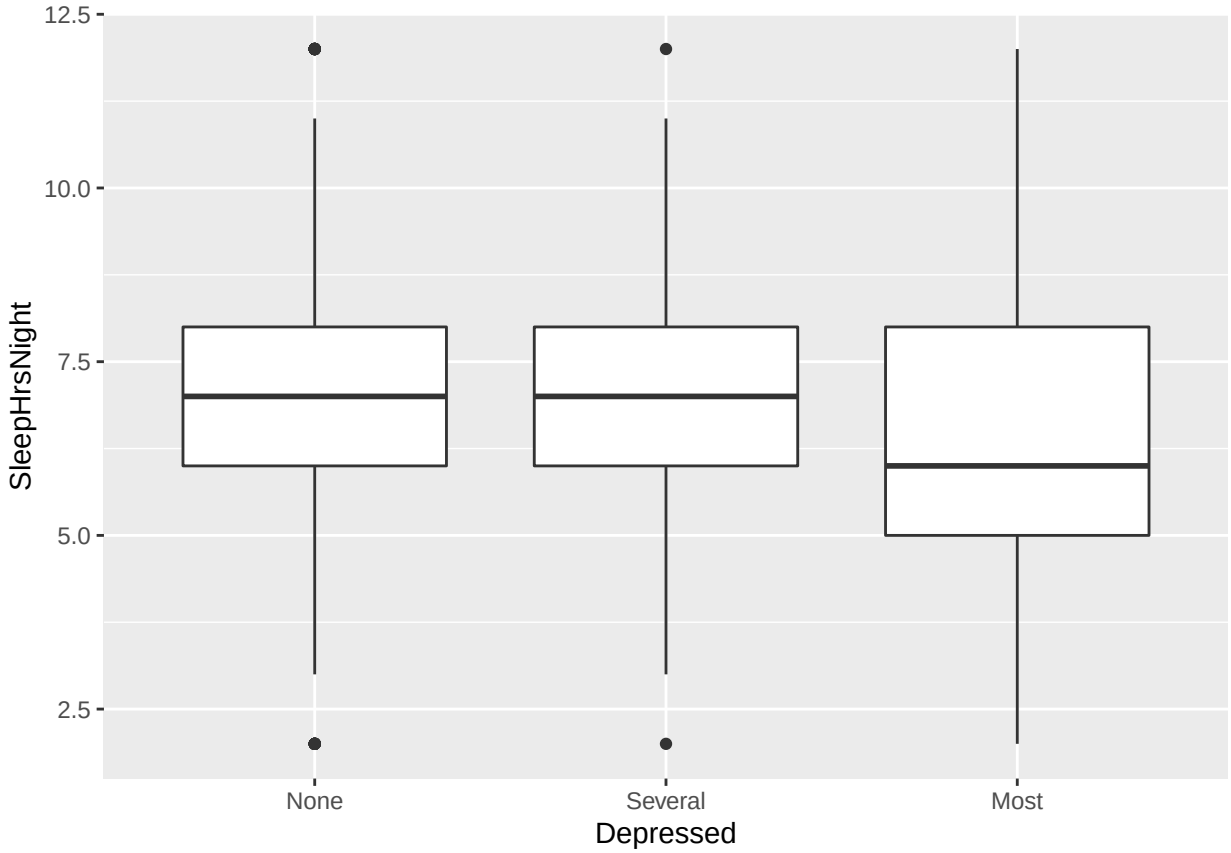
```
ggplot(data = NHANES, mapping = aes(Depressed, SleepHrsNight)) + geom_boxplot()
```



b. The `drop_na()` command removes the rows of a dataset that are NA (missing) for certain variables. Make sure you understand what the code below is doing and then recreate the boxplot from a.

```
# Drops missing values
NHANES2 <- drop_na(NHANES, Depressed, SleepHrsNight)

ggplot(data = NHANES2, mapping = aes(Depressed, SleepHrsNight)) + geom_boxplot()
```



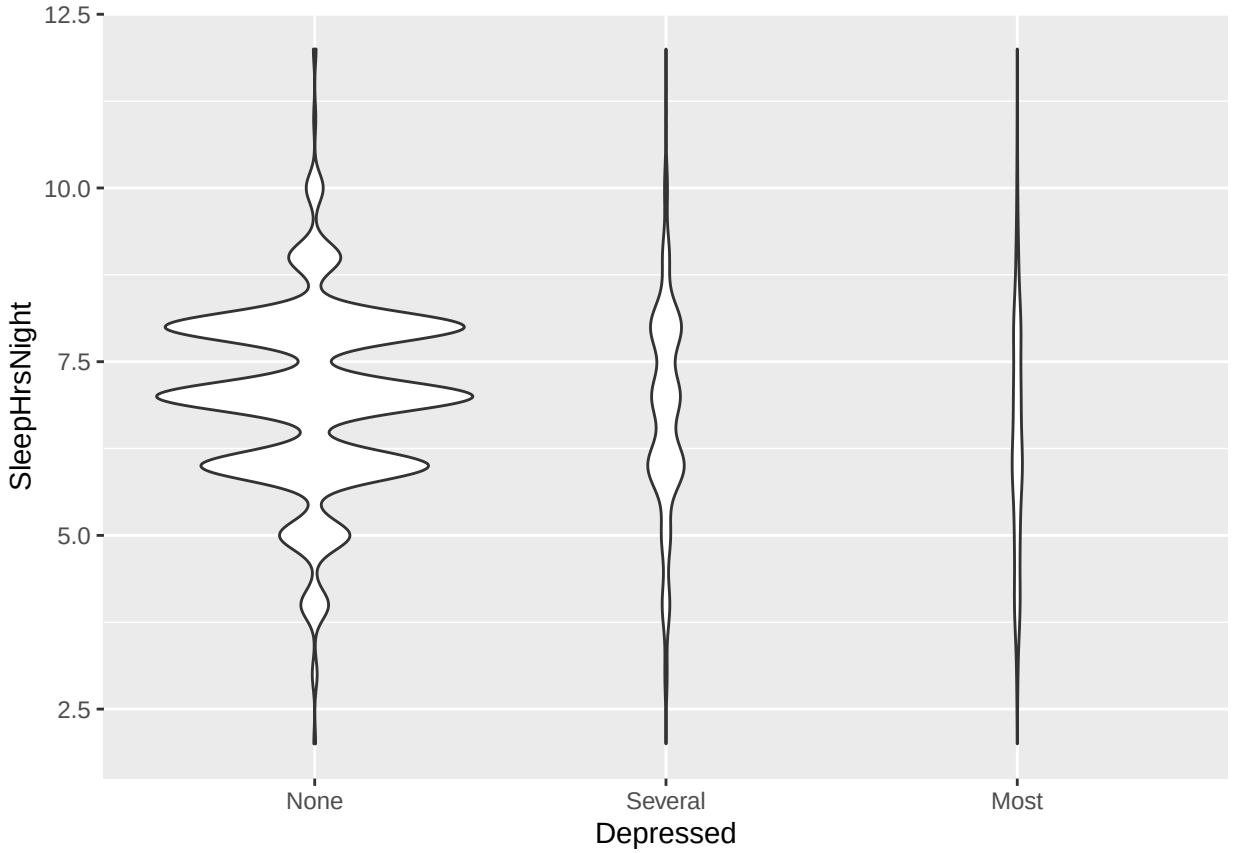
c. Draw some conclusions from the plot in b. Make sure to address the median.

- Median sleep hours a night for an individual that reports no depression is ~7 Hours. Same IQR as none
- Median sleep hours a night for an individual that reports 'several' depression is ~7 Hours. Same IQR as None
- Median sleep hours a night for an individual that reports most depression is ~6 Hours with a greater IQR
- On average, individuals that report greater levels of depression have fewer hours of sleeping

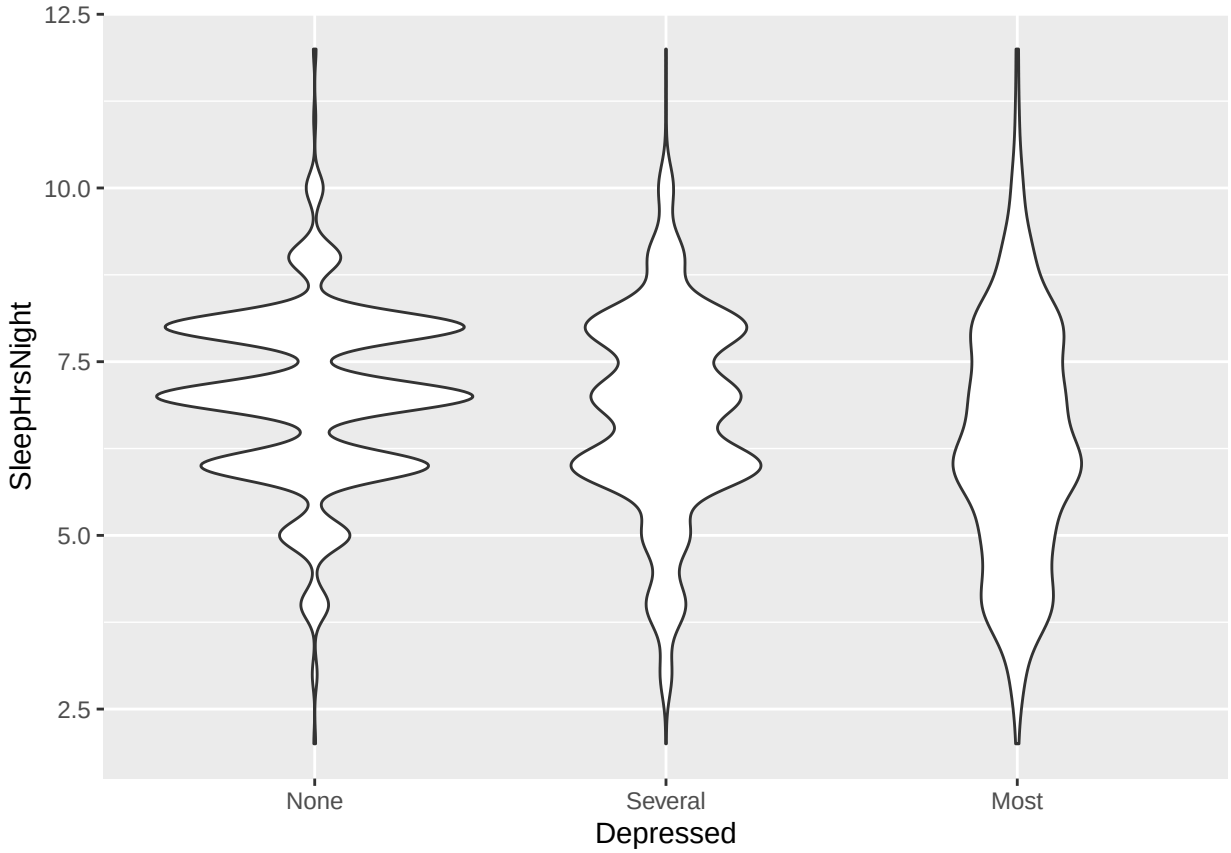
---

d. Construct side-by-side violin plots of `SleepHrsNight` by the categories of the variable `Depressed` (excluding NAs).

```
# Violin plot
ggplot(NHANES2, aes(Depressed, SleepHrsNight)) + geom_violin(scale = "count")
```




```
ggplot(NHANES2, aes(Depressed, SleepHrsNight)) + geom_violin(stat = "ydensity")
```



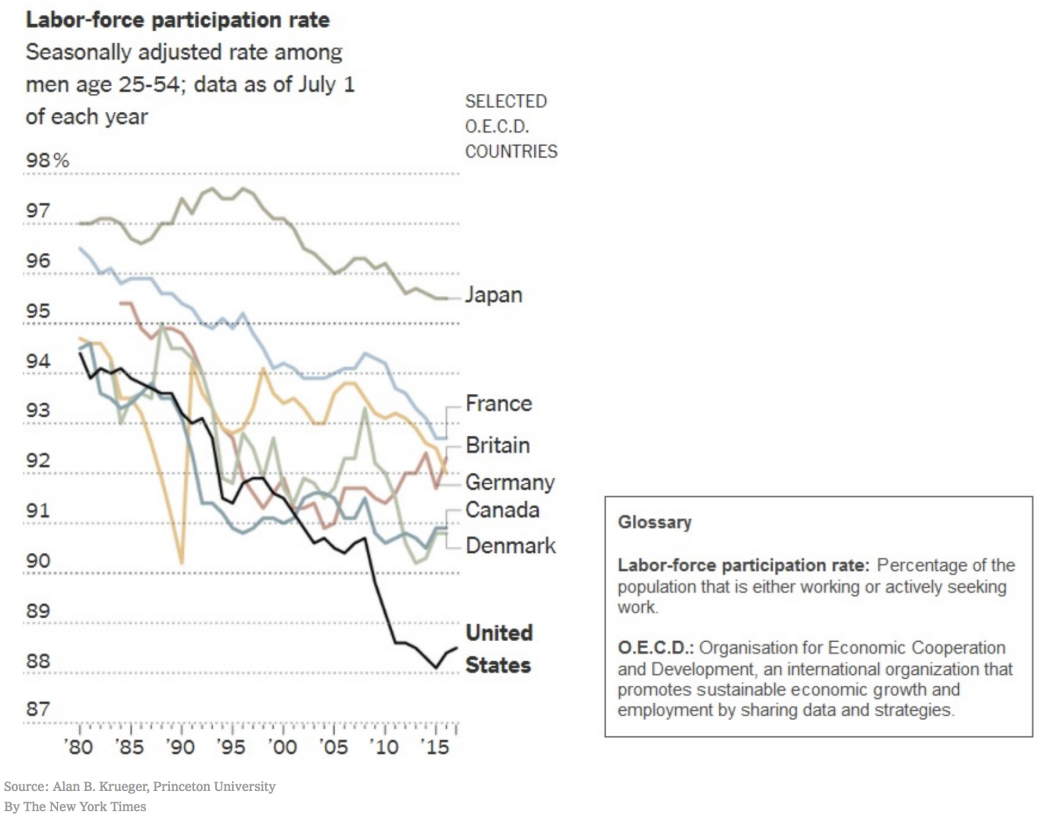
e. Why do you think your violin plots are so *wavy*?

- **Count**
- More individuals report No depression. Making it difficult to tell the distibution of 'Most' depressed
- Likely wavy as respondants give only integer responses leaving gaps in between.
- **Ydensity**
- Even distribution across the categories
- Less wavy in 'Most' Depressed category likely because it contains fewer values and extremes

---

Problem 4

This graphic comes from the NYTimes article entitled Unemployment Is So 2009: Labor Shortage Gives Workers an Edge.



a. What are the variables displayed in this graphic? For each variable, specify if it is categorical or quantitative.

- **Country**
  - Represented by seperate lines and colors +Categorical
- **Year**
  - x-axis
  - 1980-2015
  - Quanitative, can be treated as a category
- **Employmeent participation of men age 25-54 (%)**
  - y-axis
  - 87-98%
  - Quanitative, can be treated as a category

---

b. What **geom** are the variables map to?

- `_line`

---

c. What are the **aesthetic**s of the **geom**? For each **aesthetic**, give the variable that sets the values of that **aesthetic**.

- x location and y location

- Year and Employmeent participation respectively

d. Is other context provided? If so, how does it help the viewer better understand the graphic's story?

- A glossary
- Provides context on the statistics
- Text and color
- Country name next to line

e. What does this graph do well?

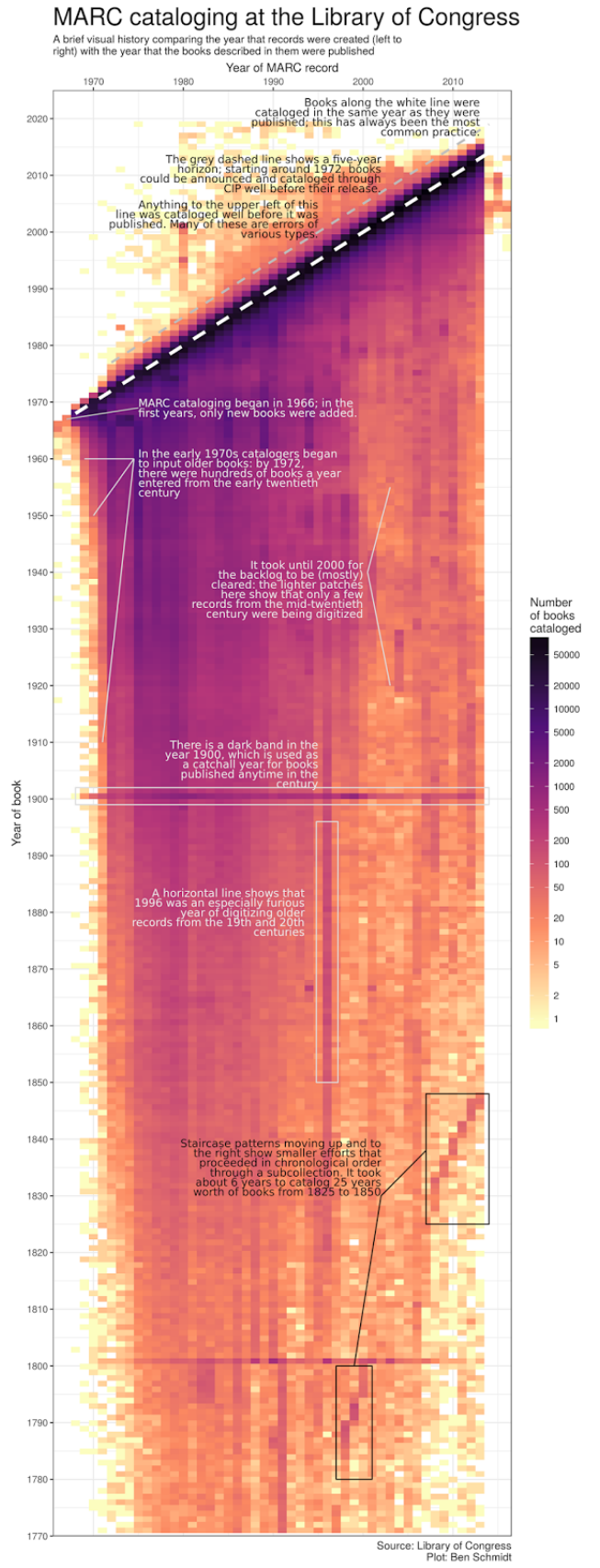
- Conveys the downward trend across selected OCED countries
- Conveys the fluctuation

f. How could this graph be improved?

- More countries
- Further context on fluctuations
- **ADD IN WOMEN**

Problem 5

The following graphic is entitled "A Brief Visual History of MARC Cataloging at the Library of Congress" and was created by Benjamin M. Schmidt so that he could better understand the history of how the Library created their digital card catalogs. The many interesting visual artifacts of the graph also help highlight the significant physical work that was involved in the conversion to digital catalogs.


MARC cataloging at the Library of Congress
A brief visual history comparing the year that records were created (left to right) with the year that the books described in them were published
Year of MARC record
1970
1980
1990
2000
2010
Year of book
2020
2010
2000
1990
1980
1970
1960
1950
1940
1930
1920
1910
1900
1890
1880
1870
1860
1850
1840
1830
1820
1810
1800
1790
1780
1770
Books along the white line were cataloged in the same year as they were published; this has always been the most common practice.
The grey dashed line shows a five-year horizon; starting around 1972, books could be announced and cataloged through CIP well before their release.
Anything to the upper left of this line was cataloged well before it was published. Many of these are errors of various types.
MARC cataloging began in 1966; in the first years, only new books were added.
In the early 1970s catalogers began to input older books: by 1972, there were hundreds of books a year entered from the early twentieth century
It took until 2000 for the backlog to be (mostly) cleared; the lighter patches here show that only a few records from the mid-twentieth century were being digitized
There is a dark band in the year 1900, which is used as a catchall year for books published anytime in the century
A horizontal line shows that 1996 was an especially furious year of digitizing older records from the 19th and 20th centuries
Staircase patterns moving up and to the right show smaller efforts that proceeded in chronological order through a subcollection. It took about 6 years to catalog 25 years worth of books from 1825 to 1850
Number of books cataloged
50000
20000
10000
5000
2000
1000
500
200
100
50
20
10
5
2
1
Source: Library of Congress
Plot: Ben Schmidt

a. What are the variables displayed in this graphic? For each variable, specify if it is categorical or quantitative.

- YAAAAAAAAAAAAAAAAAS, MY FAVORITE GRAPHIC. WE TALKED ABOUT THIS IN DATA FEMINISTS
- **Number of books catologed**
  - Color
  - Z axis
  - Density
  - Log scale
  - 0 to 50,000 books
  - Quanitative
- **Year of MARC record**
- X-axis
- From 1960 to 2017
- Quanitative, can be treated as category
- **Year of Book**
- 1770-2020
- Y-axis
- Quanitative, can be treated as category

---

b. What **geom** are the variables map to?

- Heat map is the general name
- `geom_tile()`
- And add in a z axis for density

---

c. What are the **aesthetic**s of the **geom**? For each **aesthetic**, give the variable that sets the values of that **aesthetic**.

- Each tile corresponds to a density with color being the z value
- **Number of books catologed**
  - Color
  - Z axis
  - Density
  - Log scale
  - 0 to 50,000 books
- **Year of MARC record**
  - X-axis
  - From 1960 to 2017
- **Year of Book**
  - 1770-2020
  - Y-axis

---

d. Is other context provided? If so, how does it help the viewer better understand the graphic's story?

- Text is provided to explain potential trends of high Cataloging

- Subtitle

---

e. What does this graph do well?

- MAKES TAYLOR GUSH ABOUT HOW AMAZING HEATMAPS ARE AND THEIR USAGE FOR CORRELATION.
- Shows individual trends
- Outlier trends
- Explains stair case shapes

---

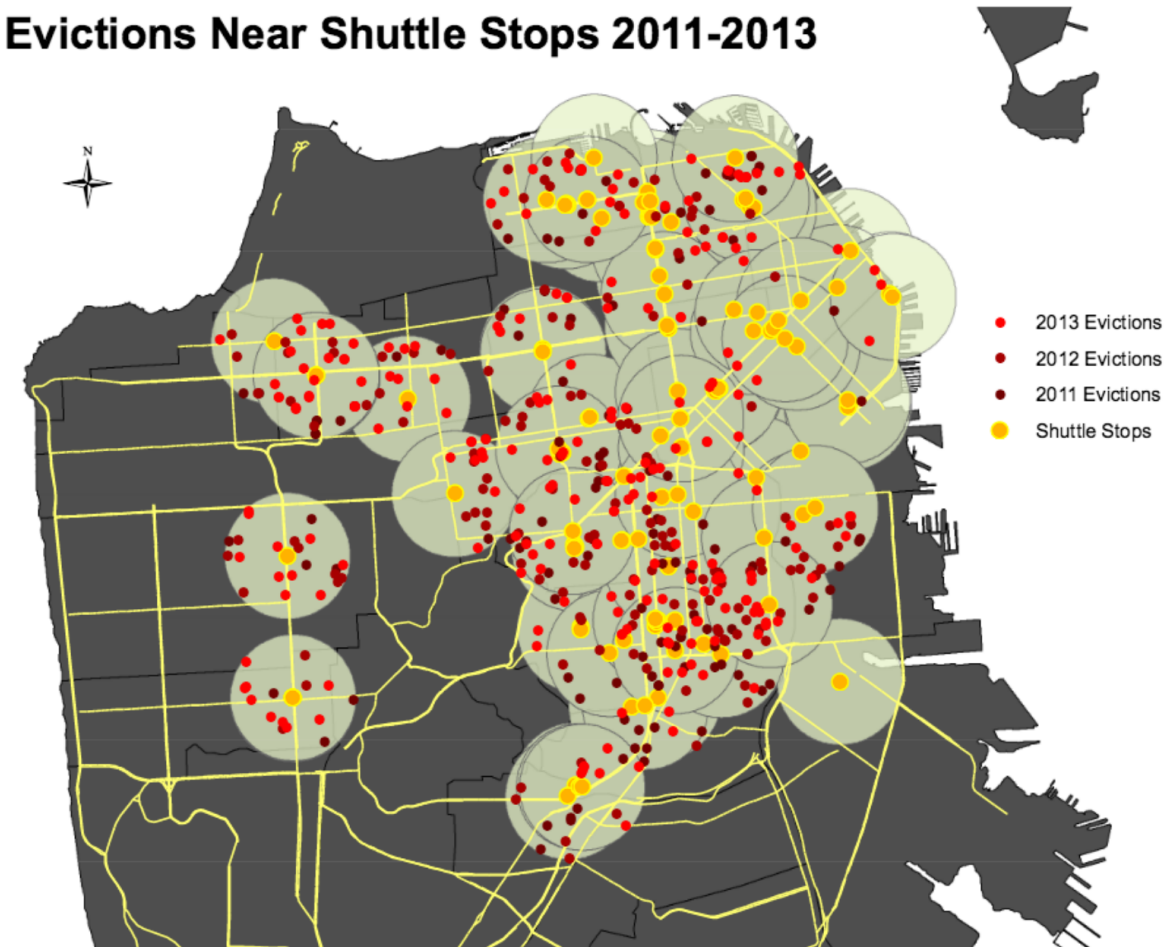
f. How could this graph be improved?

- Explain what happens to books that are republished
- Outliers in the MARC beyond 2018
- Explain in greater detail the 40 year gap and how a book published in 2020 can be in the 1980 MARC Record

---

Problem 6

The "Tech Bus Stop Eviction Map" was created by the Anti-Eviction Mapping Project and plots evictions from 2011-2013 alongside "Google Bus Stops" in San Francisco.



a. What are the variables displayed in this graphic? For each variable, specify if it is categorical or quantitative.

- YAAAAAAAAAAAAAAAAAAAAAAAAAAAAS, MORE DATA FEM GRAPHS
- Year evection
- Color
- Size
- Categorical, can be treated as quanitative
- Appended to a location
- Location eviction
- Overlayed on x and y axis
- within radius of bus stop
- Quanitative
- Bus stop
- Point on a location
- Radius around a location

---

b. What **geom** are the variables map to?

- `geom_point`
  - Alpha
  - Radius
  - Color
- Background is a map

---

c. What are the **aesthetic**s of the **geom**? For each **aesthetic**, give the variable that sets the values of that **aesthetic**.

- Color to represent year eviction
- Radius of bus stop

---

d. Is other context provided? If so, how does it help the viewer better understand the graphic's story?

- North? Rose compass
- A radius around a shuttle stop

---

e. What does this graph do well?

- Evokes emotions
- Tells a story (even if it leaves aside potentialy extenuating facts)

---

f. How could this graph be improved?

- Show me how it compares to the general SF evection rate
- Give stat of rental eviction density% closer to the data point

---

Problem 7

As we discussed in class, when you create a graph you have a lot of editorial choices that can change the story your graph is telling. For example, let's look at two graphs from the 2012 nytimes.com article "One Report, Diverging Perspectives" that both tell a story related to unemployment, but one from "How a Democrat might see things" and another from "How a Republican might see things".

a. Which graph is told from a Democratic perspective and which from a Republican perspective? In justifying your answer, identify the key editorial differences between the graphs.

- **Figure one**
  - *Democrat graph*
  - Shows the recovery after 2008 had a drop of 2.2 percent that was attributed to Obama
- **Figure two**
  - *Republican Graph*
  - Shows a supposed failure, with 43 months of unemployment being greater than 8%

The rate has fallen more than 2 points since its recent peak.

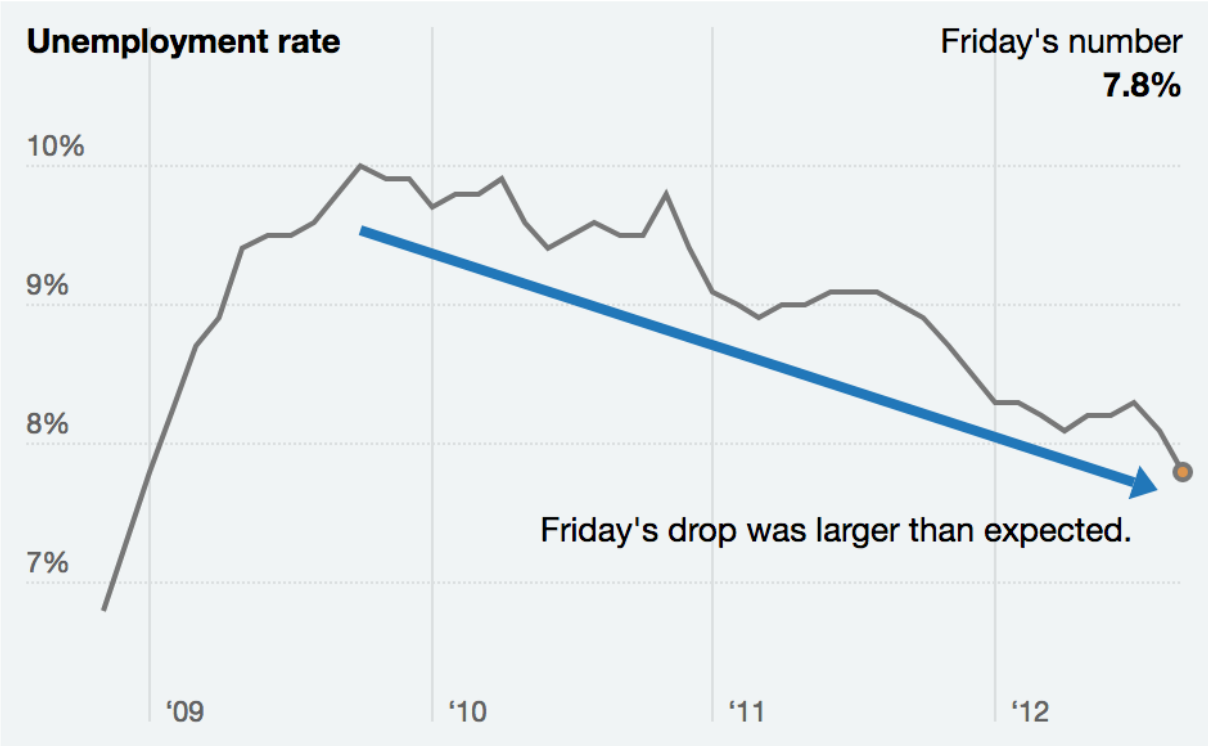


Figure 1: Example 1

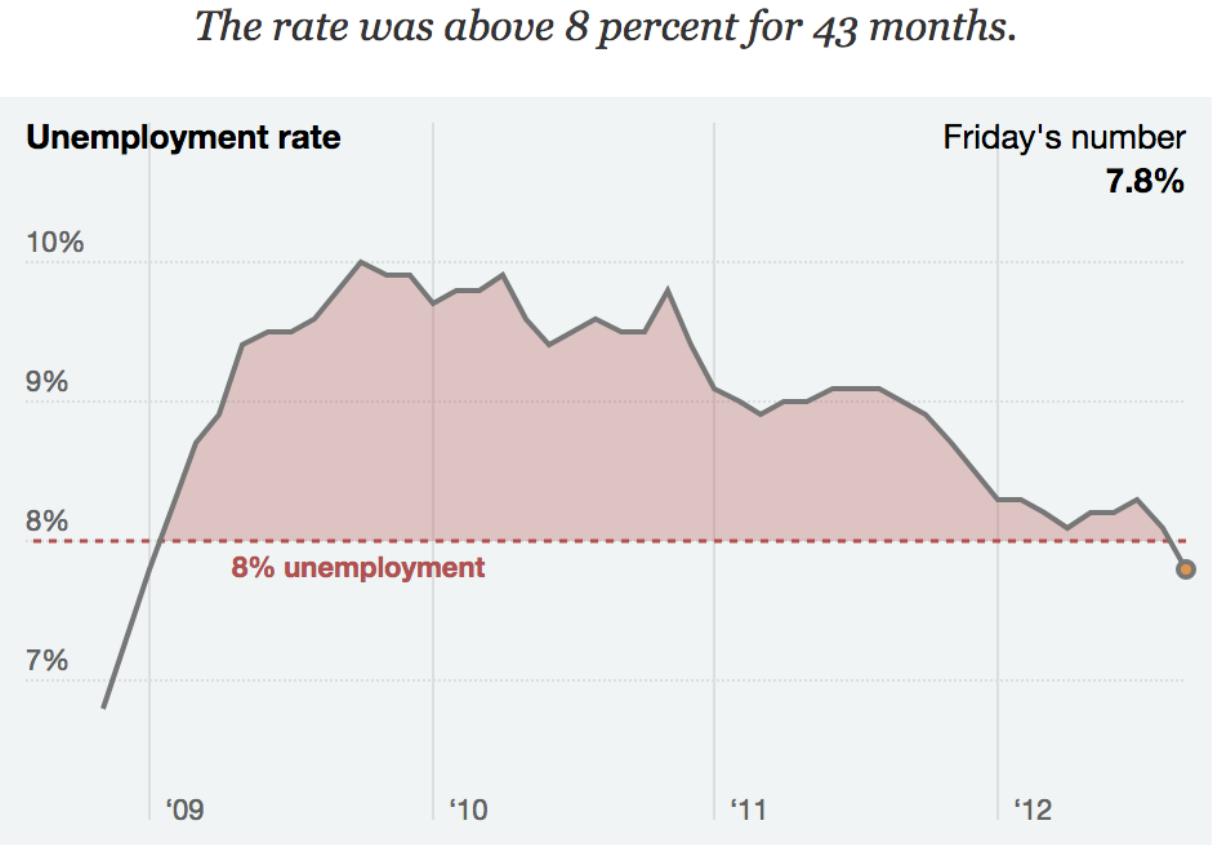


Figure 2: Example 2

b. Are these graphs neutral? Are they factual? Justify your answer.

- They both contain the same data set making them factual, they have the same scales and the same fitting, the only difference is the context provided to each in the form of additional geoms and text.

- They both add context which they use to form an opinionative argument
  - *Republican graph suggests Obama had created an unemployment rate above 8%*
  - *Democrat graph adds context that shows an improvment over time.*
- **Not Nuetral graphs, both use the same data but add context that make them political**
