Lab 11 - Data, Aesthetics, & Geometries

Johnathan Hsu

November 9, 2017

Complete the following exercises below. Knit together the PDF document and commit both the Lab 11 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Which variables in your dataset are you interested in visualizing? Describe the level of measurement of these variables and what type of geography you think is appropriate to represent these variables. Give your reasoning for choosing the `geom_()` you selected.

I am interested in visualizing sentencing length, race, age, offense level, criminal history.

2. Is your data in the proper format to visualize the data in the way you want? Why or why not? *If you need/want to change the structure of your data, do it below.*

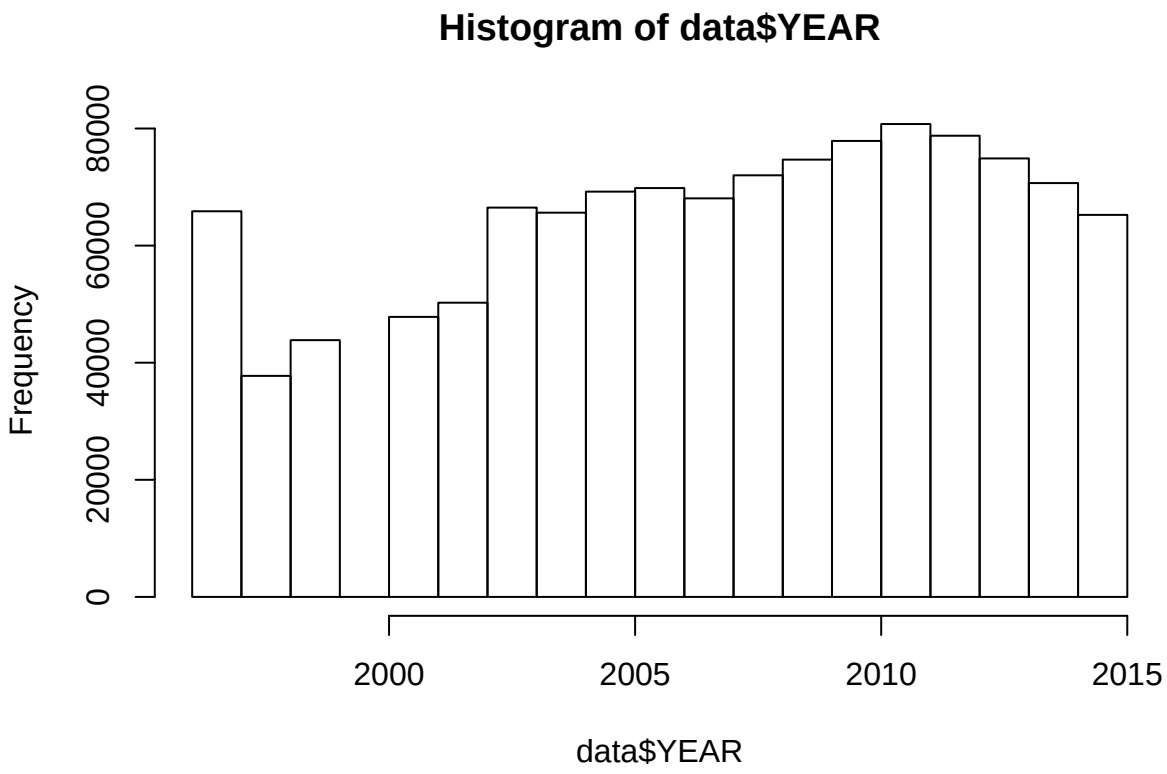
They arre in the format that I want becasue they are all in integer, which means that I can manipulate them to be factors and show scale (e.g. education can be from 0-5, low to high education, respectively). The variables are tidy, as there is no overlap in the information presented.

```
## # A tibble: 6 x 15
##   USSCIDN MONRACE  YEAR MONSEX CITIZEN EDUCATN   AGE STATMIN STATMAX
##     <int>   <int> <int>  <int>   <int>   <int> <int>   <int>   <int>
## 1  246845       2  1996      1       1       1    20     120     540
## 2  248876       2  1996      0       1       1    54       0     252
## 3  248986       2  1996      0       1       2    21       0     240
## 4  252632       1  1996      0       1       1    25      60     480
## 5  252645       1  1996      0       3       1    35       0      36
## 6  252646       1  1996      1       1       1    21       0      60
## # ... with 6 more variables: DISPOSIT <int>, TOTPRISN <int>,
## #   XFOLSOR <int>, XCRHISSR <int>, DISTRICT <int>, MONCIRC <int>
```

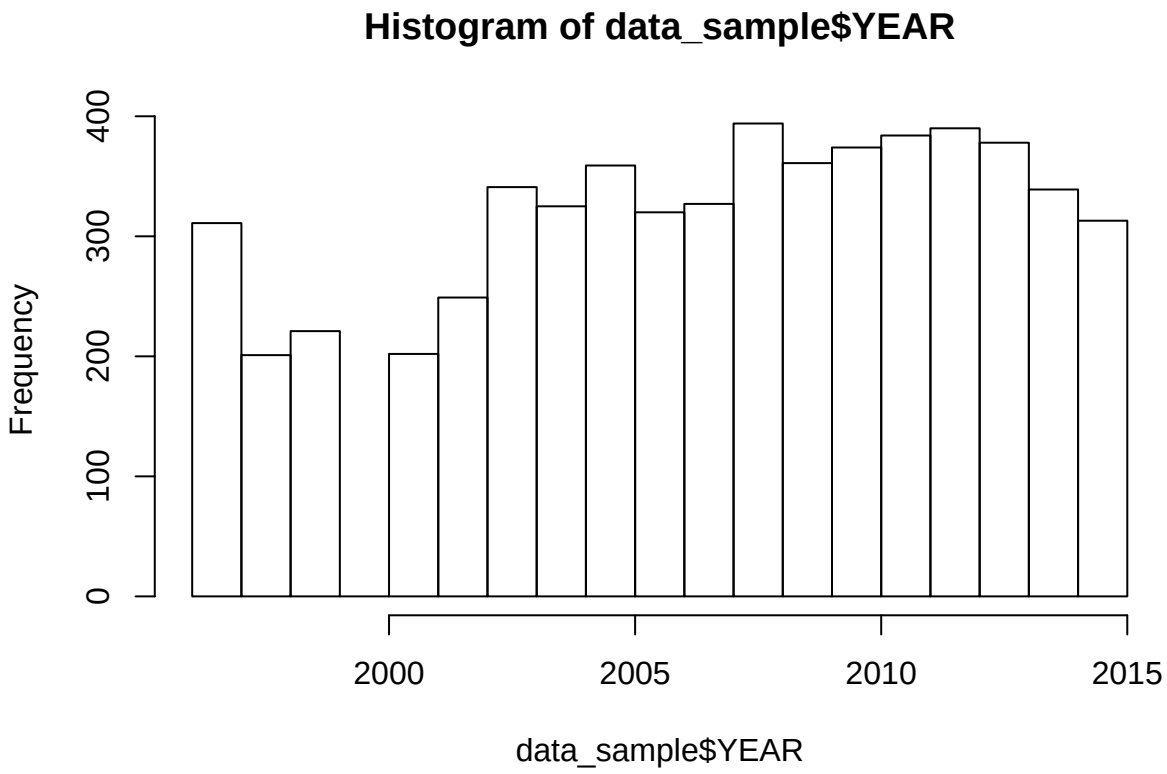
3. Create at least two different exploratory plots of the variables you chose using the skills we covered in class today. What types of mapping aesthetics did you choose and why? What do these plots tell you about your data?

I chose a scatterplot to see the distribution of different sentences over different two dominant race in the criminla justice system.

```
# Sampling data to not overcrowd the scatterplot
data_sample <- sample_n(data, 6000)

# checking that the years are representative
hist(data$YEAR)
```




```
hist(data_sample$YEAR)
```




```
data_prebooker <- data_sample %>%
  filter(MONRACE <= 2) %>%
  filter(YEAR < 2005)
```

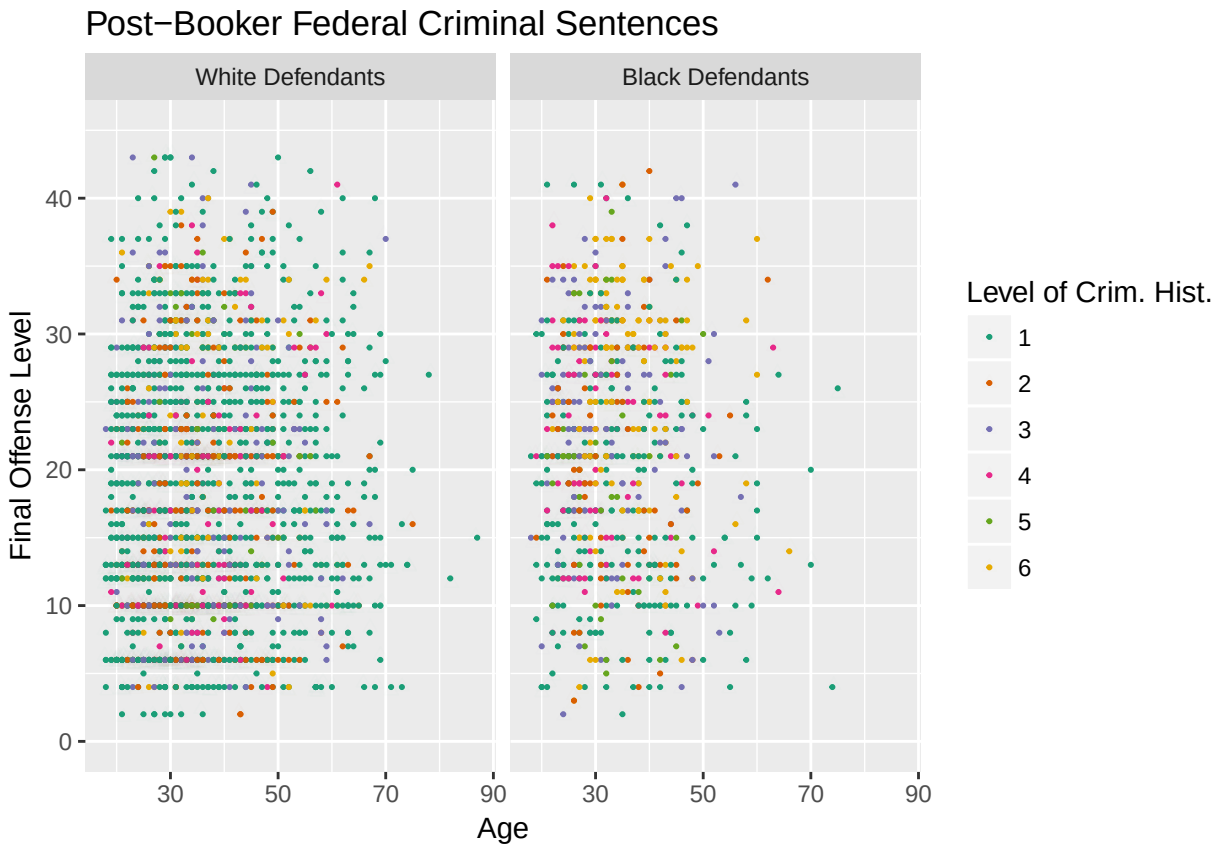
```
data_postbooker <- data_sample %>%
  filter(MONRACE <= 2) %>%
  filter(YEAR > 2005)

race_labels <- c('1' = "White Defendants",
                 '2' = "Black Defendants",
                 '3' = "Native American",
                 '4' = "Asian/Pacific Islander",
                 '5' = "Other")


post_booker_scatter <- ggplot(data_postbooker, aes(x = AGE, y = XFOLSOR, col = as.factor(XCRHISSR))) +
                               geom_point(size = 0.4) + scale_x_continuous(name = "Age") +
                               geom_jitter(alpha = 0.01, shape = 2)

pre_booker_scatter <- ggplot(data_prebooker, aes(x = AGE, y = XFOLSOR, col = as.factor(XCRHISSR))) +
                               geom_point(size = 0.4) + scale_x_continuous(name = "Age") +
                               geom_jitter(alpha = 0.01, shape = 2)
```

4. Create at least three variations of the plots you've already made by modifying some of the arguments we covered in class (i.e. `position`, `scale`, `size`, `linetype` etc.). Do any of these modifications help you understand your data better? Why or why not? Do any of them create a misleading interpretation of the relationships between your variables? If yes, how so?

I added a lot of things to the first two graphs (see below) - they are better because we can now compare the spread between two races, they are labeled, and the dots aren't fat and clumped up. They include jitter, alpha, legends, and many more.

```
# pre and post booker scatterplots
post_booker_scatter <- ggplot(data_postbooker, aes(x = AGE, y = XFOLSOR, col = as.factor(XCRHISSR))) +
                               geom_point(size = 0.4) + scale_x_continuous(name = "Age") +
                               labs(title = "Post-Booker Federal Criminal Sentences", y = "Final Offens
                               geom_jitter(alpha = 0.01, shape = 2) +
                               scale_color_brewer(name = "Level of Crim. Hist.", type = "qual", palette
                               guides(fill = guide_legend(title="Race")) +
                               scale_y_continuous(limits = c(0,45)) +
                               facet_grid(.~ MONRACE, labeller = as_labeller(race_labels))


pre_booker_scatter <- ggplot(data_prebooker, aes(x = AGE, y = XFOLSOR, col = as.factor(XCRHISSR))) +
                               geom_point(size = 0.4) + scale_x_continuous(name = "Age") +
                               labs(title = "Pre-Booker Federal Criminal Sentences", y = "Final Offense
                               geom_jitter(alpha = 0.01, shape = 2) +
                               scale_color_brewer(name = "Level of Crim. Hist.", type = "qual", palette
                               guides(fill = guide_legend(title="Race")) +
                               scale_y_continuous(limits = c(0,45)) +
                               facet_grid(.~  MONRACE, labeller = as_labeller(race_labels))

post_booker_scatter
```

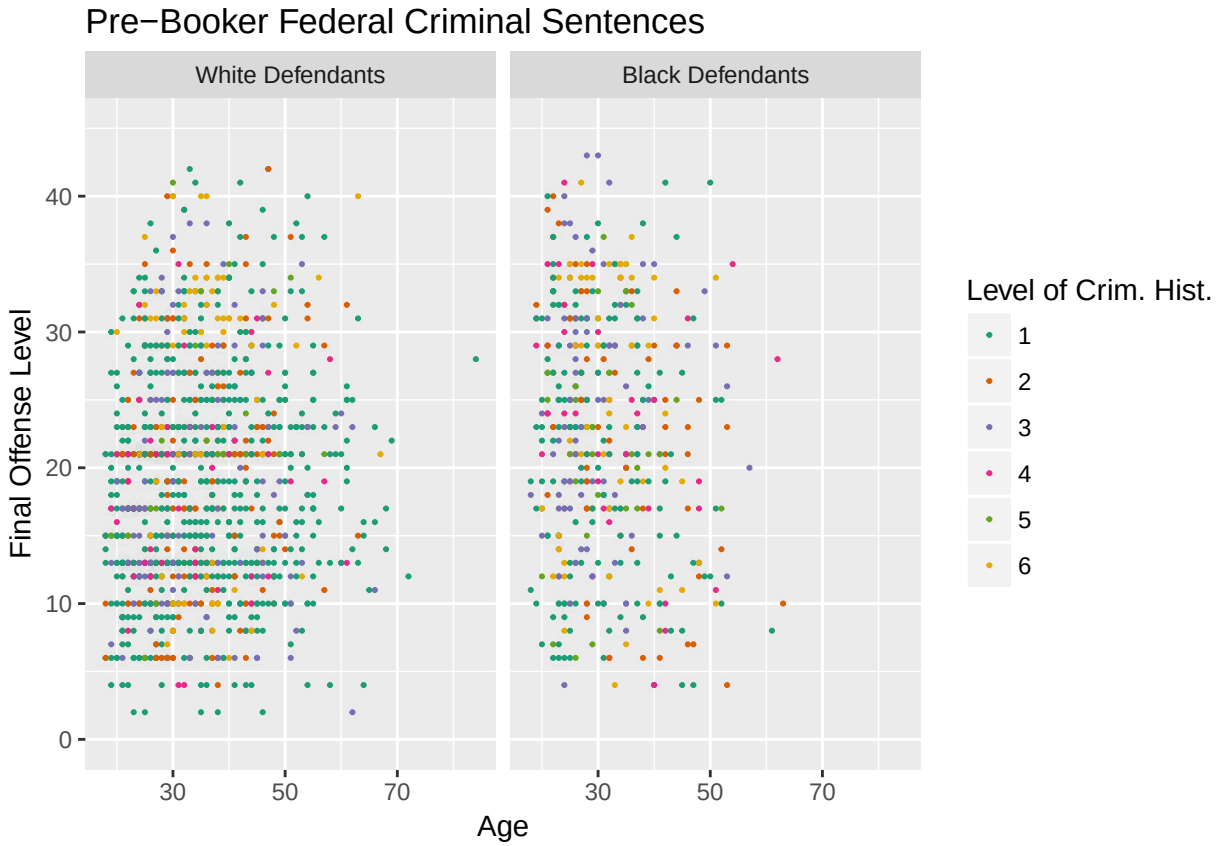
```
## Warning: Removed 19 rows containing missing values (geom_point).

## Warning: Removed 19 rows containing missing values (geom_point).
```




```
pre_booker_scatter
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```



5. From the plots you've created thus far, do any of them seem appropriate for a general audience? Why or why not? If so, what do you think you'd still need to do to make them more suitable as explanatory visualizations?

I think the plot is great for the general audience because it's easy to compare the spread of the two graphs and determine the differences.

I think the graphs from #2 (Pre and Post 2005 Booker) would make the most sense. My goal is to be able to put the right enough amount of sample or subset data to make the two graphs still readable.