Assignment 10

Matthew McAvoy

November 17, 2016

1.

The goal of this assignment is to collect data pertaining to crime and education in Florida counties. The data needed to be cleaned and merged, before exploratory analysis could follow.

The first step was moving to the directory where the data was stored. Data was saved from Florida arrest data: http://www.fdle.state.fl.us/cms/FSAC/Data-Statistics/UCR-Arrest-Data.aspx as excel files. The files were then saved as csv. The first file was loaded for exploration.

```
setwd("C:/Users/homur/OneDrive/New College/EDA/Week 11")
```

```
filenames <- list.files("./crime_data", pattern="*.csv", full.names=TRUE)

ldf_a <- read.csv(filenames[1], header=FALSE) %>%
        as.data.frame()
```

2.

To begin cleaning, we wanted to retain rows between the first county and the last.

```
## Slice rows to start at first county and end at last county.
Alachua_label <- which((ldf_a[,1] == "Alachua County") | (ldf_a[,1] =="Alachua") )
Washington_label <- which((ldf_a[,1] == "Washington County") | (ldf_a[,1] == "Washington") )
Alachua_label; Washington_label
```

```
## [1] 4
```

```
## [1] 70
```

Next we adjusted the first year to fit our dimensions and retain only desired columns, being county name, population, and arrests. Additional columns for year and ID were added, ID is formed by pasting the county name with year which were used for joining.

```
start_year <- 2004
ldf_a <- ldf_a %>% select(1:3) %>% slice(Alachua_label: Washington_label) %>%
        mutate(County = paste(V1, "County", sep=" ")) %>%
        mutate(Year = start_year) %>%
        mutate(ID = paste(County, Year, sep="-"))
```

A for-loop was used to load the remaining files corresponding to different years, adjusted as the first year was, and attached to the first data frame. This set of data was finished by adding column names.

```
for (i in 2:length(filenames)) {
    ldf_b <- read.csv(filenames[i], header=FALSE) %>%
        as.data.frame() %>% select(1:3) %>%
        slice(Alachua_label: Washington_label) %>%
        mutate(County = paste(V1, "County", sep=" ")) %>%
        mutate(Year = start_year + i - 1) %>%
        mutate(ID = paste(County, Year, sep="-"))
    ldf_a <- bind_rows(ldf_a, ldf_b)
}
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
names(ldf_a) <- c("V1", "Population", "Arrests", "County", "Year", "ID")
```

We also wanted Education data. I found High School completion by county for years 2009-2014 at http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

Just like the arrest data, the filenames were stored, the first file was adjusted, and the remaining years were added.

```
edu_filenames <- list.files("./education_data", pattern="*.csv", full.names=TRUE)

edu_a <- read.csv(edu_filenames[1], header=FALSE) %>%
        as.data.frame() %>% select(7:8)

edu_year = 2009
edu_a <- edu_a %>% slice(Alachua_label: Washington_label) %>%
    mutate(Year = edu_year) %>%
    mutate(ID = paste(V7, Year, sep="-"))

for (i in 2:length(edu_filenames)) {
    edu_b <- read.csv(edu_filenames[i], header=FALSE) %>%
        as.data.frame() %>% select(7:8) %>%
        slice(Alachua_label: Washington_label) %>%
        mutate(Year = edu_year + i - 1) %>%
        mutate(ID = paste(V7, Year, sep="-"))
    edu_a <- bind_rows(edu_a, edu_b)
}
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
names(edu_a) <- c("County", "HS_Grad_Rate", "Year", "ID")
```

3.

With two data frames of similar size, an inner join was performed that matched on ID, Year, and County.

```
full_data <- inner_join(ldf_a, edu_a)
```

```
## Joining, by = c("County", "Year", "ID")
```

```
head(full_data)
```

```
##         V1 Population Arrests          County Year                    ID
## 1  Alachua    256,232  18,654  Alachua County 2009  Alachua County-2009
## 2    Baker     25,899   1,464    Baker County 2009    Baker County-2009
## 3      Bay    169,562  16,618      Bay County 2009      Bay County-2009
## 4 Bradford     29,085   1,276 Bradford County 2009 Bradford County-2009
## 5  Brevard    555,657  31,353  Brevard County 2009  Brevard County-2009
## 6  Broward  1,744,922  86,327  Broward County 2009  Broward County-2009
##   HS_Grad_Rate
## 1         89.1
## 2         78.6
## 3         85.8
## 4         78.7
## 5         89.9
## 6         87.0
```

The join was successful, just needed to remove a few now unnecessary columns and order how we like.

```
full_dd <- full_data %>% select(-V1,-ID)
full_dd <- full_dd %>% select(County, Year, Population, Arrests, HS_Grad_Rate)
```

Finally year, and rate data was converted to numeric.

```
full_dd[,3] <- gsub(",", "", full_dd[,3])
full_dd[,3] <- as.numeric(full_dd[,3])
full_dd[,4] <- gsub(",", "", full_dd[,4])
full_dd[,4] <- as.numeric(full_dd[,4])
#full_dd[,5] <- gsub(",", "", full_dd[,5])
full_dd[,5] <- as.numeric(full_dd[,5])
str(full_dd)
```

```
## 'data.frame':    390 obs. of  5 variables:
##  $ County      : chr  "Alachua County" "Baker County" "Bay County" "Bradford County" ...
##  $ Year        : num  2009 2009 2009 2009 2009 ...
##  $ Population  : num  256232 25899 169562 29085 555657 ...
##  $ Arrests     : num  18654 1464 16618 1276 31353 ...
##  $ HS_Grad_Rate: num  89.1 78.6 85.8 78.7 89.9 87 72 87.9 83.9 89.8 ...
```

Progress was saved by writing to csv and then re-loaded for initial analysis.

```
write.csv(full_dd, file = "arrest_edu_data.csv")
```

```
data <- read.csv("arrest_edu_data.csv")
data <- data %>% select(-X)
```

4. and 5.

My initial hypothesis is that there is a negative relationship between arrests and high school graduation rates.

High School graduation rate is already in decimal format, however arrests were not. Arrest rate was added to the data by mutating arrests/population.

```
data <- data %>% mutate(Arrest_rate = Arrests/Population)
```

Correlation between High School graduation rate and Arrest rate was performed.

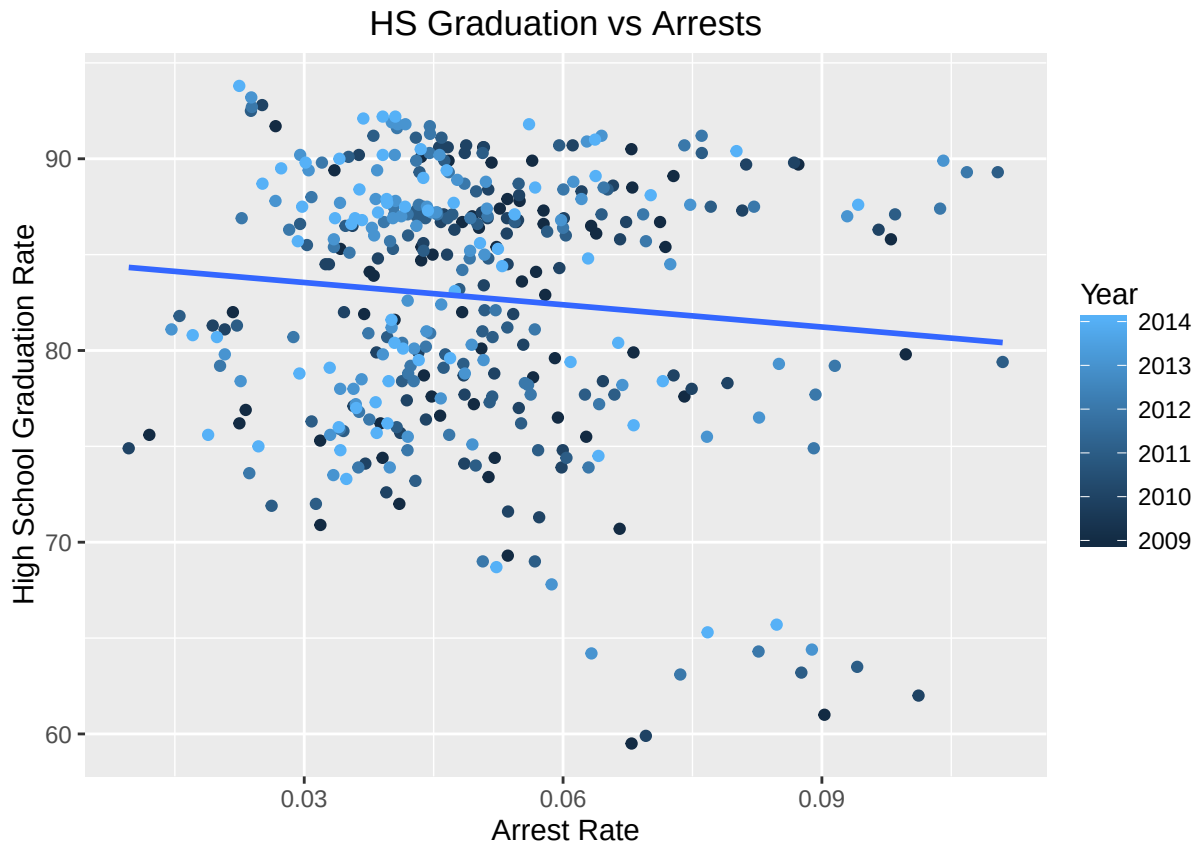
```
data_cor <- data %>% select(HS_Grad_Rate, Arrest_rate, Population)
cor(data_cor)
```

```
##               HS_Grad_Rate Arrest_rate Population
## HS_Grad_Rate     1.0000000 -0.10206670 0.27679525
## Arrest_rate     -0.1020667  1.00000000 0.01766442
## Population       0.2767953  0.01766442 1.00000000
```

While there is a small negative correlation, it isn't very large, actually population has a larger correlation to high school graduation rate.

A plot of Arrests to graduation rate with a best fit line was constructed.

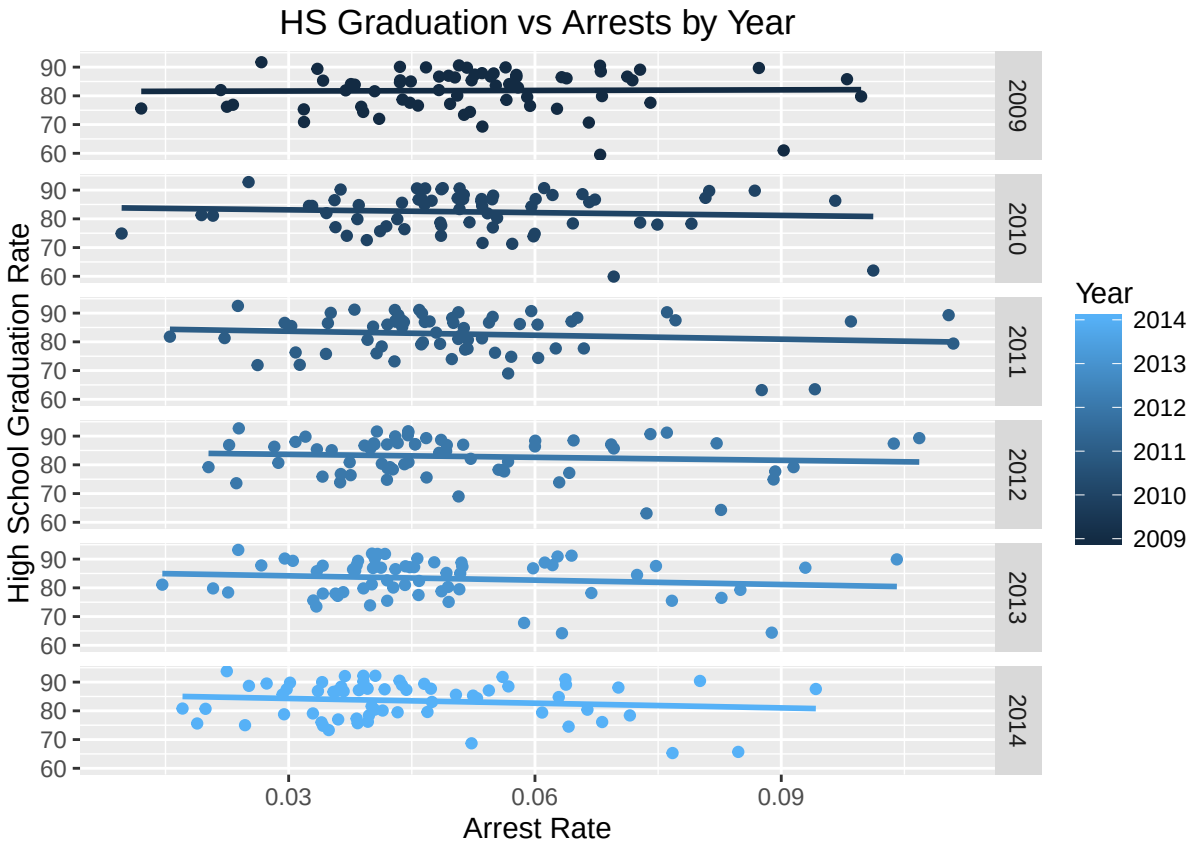
```
ggplot(data = data, aes(x=Arrest_rate, y=HS_Grad_Rate, color=Year)) +
  geom_point() + geom_smooth(method="lm", se=FALSE) +
  labs(x="Arrest Rate", y="High School Graduation Rate", title="HS Graduation vs Arrests")
```



We see evidence to support our hypothesis, that is there appears to be a decrease in high school graduation rate with increased arrests.

We can facet by year as well.

```
ggplot(data = data, aes(x=Arrest_rate, y=HS_Grad_Rate, color=Year)) +
  geom_point() + geom_smooth(method="lm", se=FALSE) +
  facet_grid(Year~.) +
  labs(x="Arrest Rate", y="High School Graduation Rate", title="HS Graduation vs Arrests by Year")
```

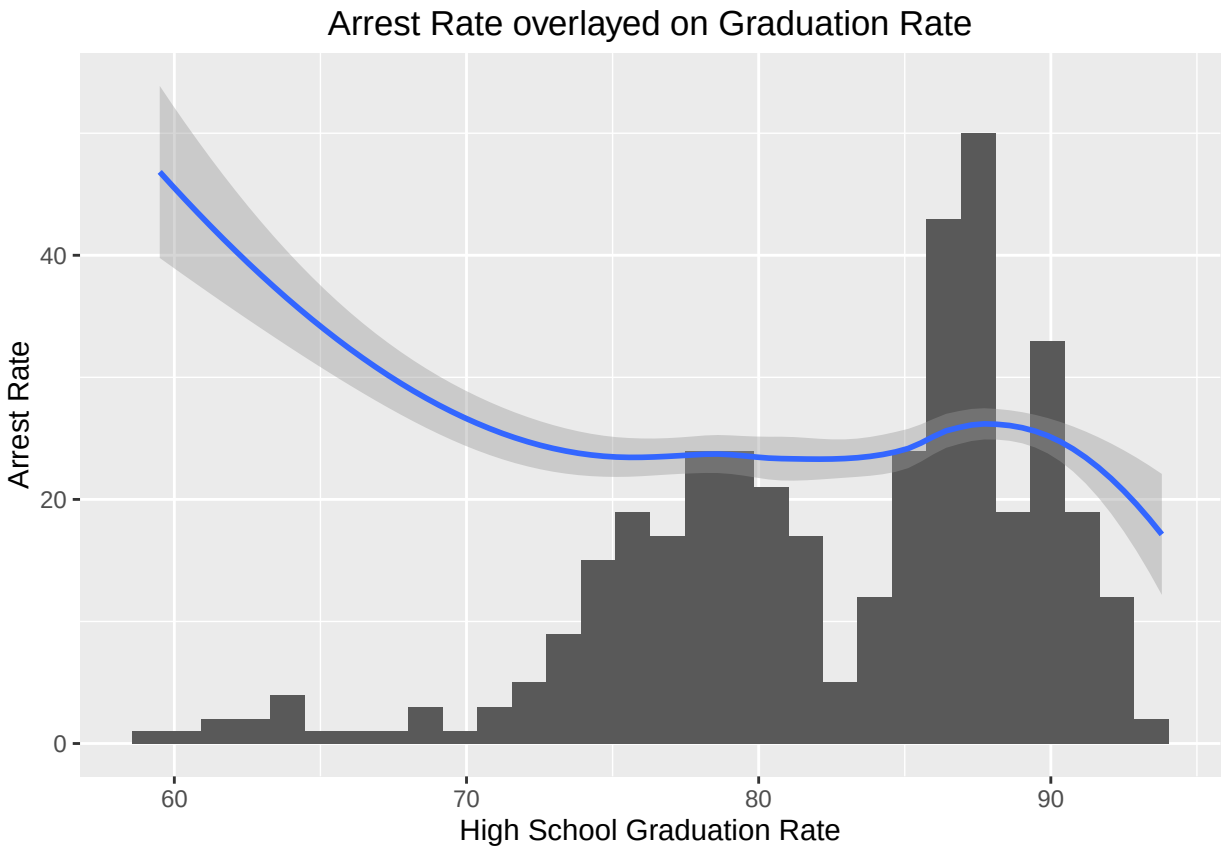


It looks like there was smaller correlation in previous years and stronger in recent years. This might indicate education is playing a more important role in reducing the likelihood of commiting crimes that lead to arrests, or it might mean being arrested has a stronger impact on High School graduation than it did in previous years.

A somewhat interesting finding is plotting a histogram of High School graduation and overlaying a linear model of arrest rate per 500 people (to fit scale). This might say that between about 70-90% graduation rate, there isn't much difference in arrests, but on the ends, improving graduation rates significantly reduced arrests.

```
ggplot(data = data, aes(HS_Grad_Rate)) + geom_histogram() + geom_smooth(aes(y=Arrest_rate*500)) +
  labs(x="High School Graduation Rate", y="Arrest Rate", title="Arrest Rate overlayed on Graduation Rat
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



I have found evidence supporting my initial hypothesis, that there is a negative relationship between High School graduation and arrests. To continue exploration an interesting objective would look at data going back into the 1980's where much different socioeconomic conditions affecting High School graduation and arrests were in play. Additionally gathering demographic data along with subsetting on different crimes would be further avenues of exploration in drawing insight into the relationship between education and arrests in Florida counties.