Rensselaer Polytechnic Institute

Exploring U.S. Crime and Population Change by Time and Location

Using Clustering and Regression Models on the FBI UCR Dataset

Scout Oatman-Gaitan

Data Analytics - ITWS 6600

Thilanka Munasinghe

May 4, 2020

Contents:

1. Abstract and Introduction

1.1 Abstract and Introduction

The goal of my project is to understand crime trends in the United States both at a national level and also at a state level. I will explore these rates and crime totals over time and see how the data is changing to try to gain some insight on which types of crimes are committed most in the United States. At the state level, I want to understand what areas are more prone to crime and what areas aren't. I want to explore where states rank in crime totals versus crime rates. Obviously states with high populations like CA, NY, and TX will have high crime totals because they have more people. Looking at crime rates will give us better insight into what states have 'worse' crime because we can see that scaled data. For these purposes I plan on developing both a clustering model and a regression model.

1.2 Motivation

The motivation for this assignment is to have a better understanding of the crime climate in America. It is to find problematic areas and safer areas and try to gain some insight into where to look for questions we have. Ideally, this process could give some direction to councils trying to advise these specific locations.

2. Data Description

2.1 About the Dataset

I used data gathered by the FBI, specifically the Uniform Crime Reporting Program (UCR). I used two forms of this data[1]. First, I used data at a national level that reports the year (1997-2016), population, crime rates for various categories, and crime totals for various categories. Note that crime rates are per 100,000 inhabitants. Second, I used a bigger dataset that included a breakdown by state. This dataset included information for each state from 1997-2014 and the same crime rates and totals by category. The first dataset was available in full from FBI UCR but the second dataset required joining for 50 tables. Both tables included a “legacy definition rape” category and a “revised definition rape” category. The revised definition was composed of mostly null values so I ignored this category. I had planned to also look at a county level but struggled to find good data for this.

In addition to this main data, I also utilized a data set[2] from the FCC with the corresponding state Federal Information Processing System (FIPS) codes. This would help with geographical visualizations.

2.2 Reference Material

FBI UCR is a federally recognized program and therefore provides reputable data. There are many sources of reference for this data as it is well-known and widespread.

[1] https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/topic-pages/tables/table-1
[2] https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt

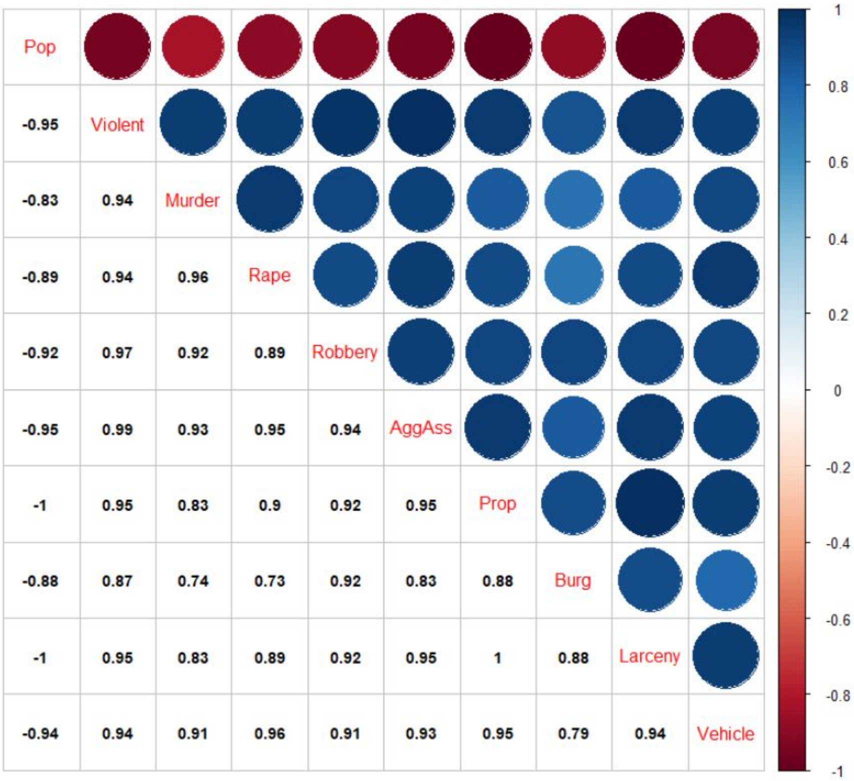
3. Analysis

3.1 Data Cleaning and Preparation

The first thing I had to do with the national level data was make the columns numeric. Next I rescaled the data so that columns are better compared. I used the method:

$$f(x) = \frac{x - min(x)}{max(x) - min(x)} * 100$$

Next I had to split the data into two different data frames – one for the crime rates and one for the crime totals. I made a correlation matrix for the rates data:



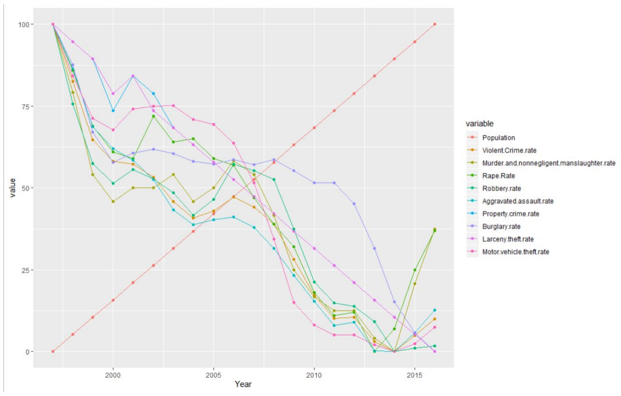
Pop									
-0.95	Violent								
-0.83	0.94	Murder							
-0.89	0.94	0.96	Rape						
-0.92	0.97	0.92	0.89	Robbery					
-0.95	0.99	0.93	0.95	0.94	AggAss				
-1	0.95	0.83	0.9	0.92	0.95	Prop			
-0.88	0.87	0.74	0.73	0.92	0.83	0.88	Burg		
-1	0.95	0.83	0.89	0.92	0.95	1	0.88	Larceny	
-0.94	0.94	0.91	0.96	0.91	0.93	0.95	0.79	0.94	Vehicle

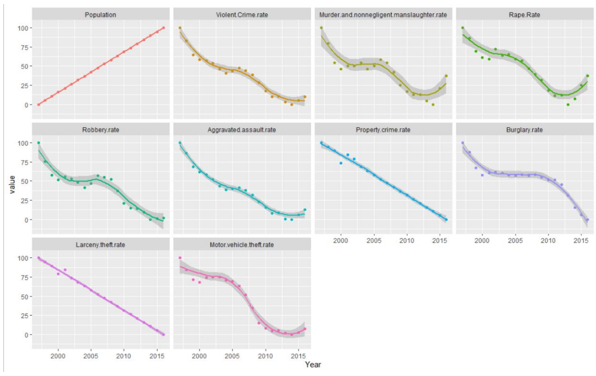
We can see negative correlation across the board for the population. We know that the population is increasing through the years and that all crime categories are decreasing and seem to have strong correlations to each other.

For the state level data, first I used a left join with a data frame that had fips codes so that my data would have the associated fips codes for states. This would make visualizing geographic data easier later on. Next, like I did for the national level, I scaled the data using the same function as above and I split the data frame so that rates and totals were separate. I decided to do my future analysis in rates as it was more telling given that it accounted for population. Rates were per 100,000 people.

3.2 Time Based Analysis

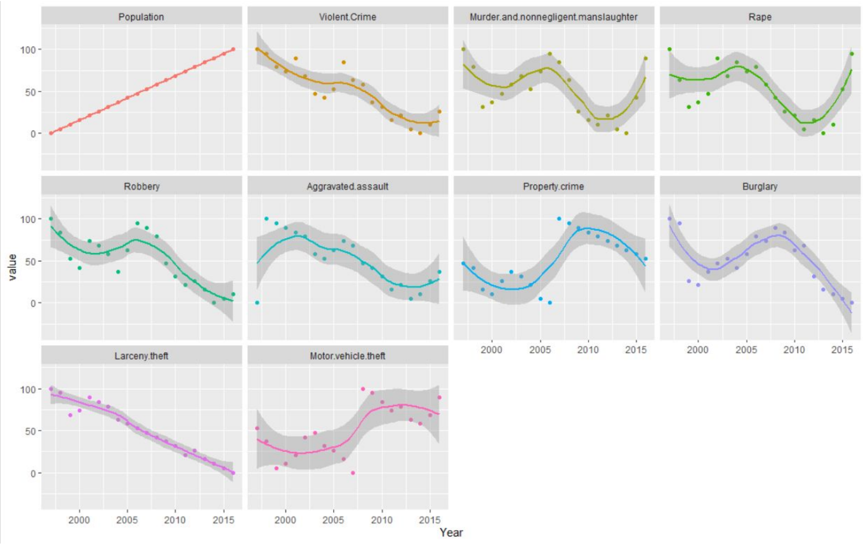
National-level rates:





At a national level we can explore the population, crime rates, and crime totals. The visuals above disregard the national level totals. The red increasing line is the population. All other lines, all decreasing, are the various crime rates. We can see that rates like larceny and property crime decrease linearly whereas other rates are decreasing in other manners. Violent crime, murder, rape, aggravated assault, and motor vehicle theft all have a bit of an uptick at the end.

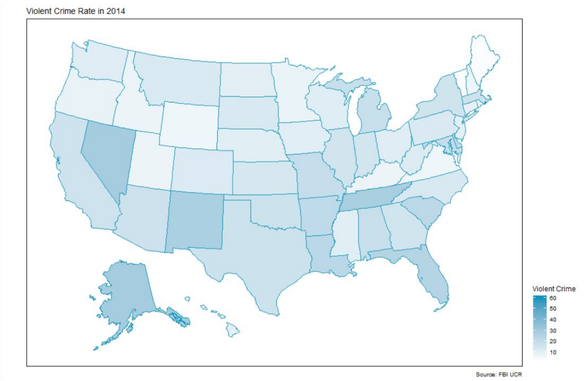
National-level totals:

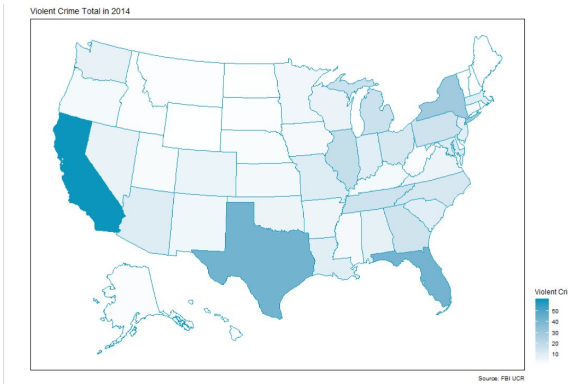


The visuals above disregard the national level rates we explored above and instead focus on totals. The red increasing line is the population. All other lines are the various crime rates. We immediately see that there is much more variation in the data. It looks like property crime and motor vehicle theft are actually increasing but we know from the rates data that the rates are decreasing. This is a case of an increasing population. That uptick that I mentioned in the rates data is much more pronounced here with the exception of larceny which is still very linear. While this data is very interesting and has merit in other analysis, I decided that I was more interested in the rates since it accounts for the change in population.

3.3 Geographical Analysis

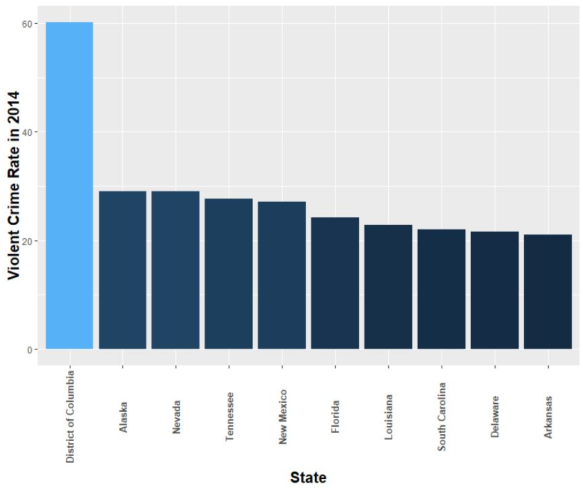
Violent crime rate versus violent crime total in 2014:

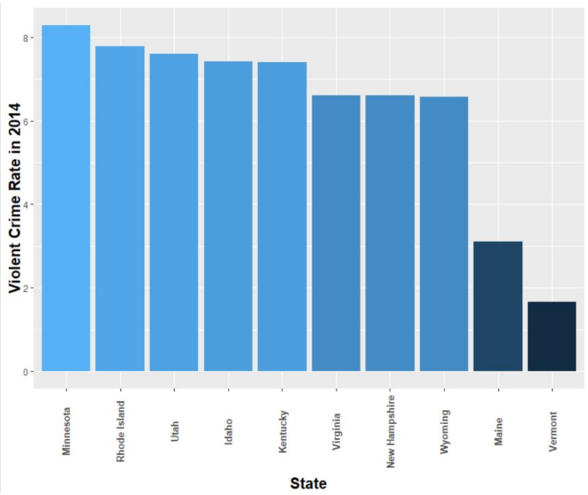




These graphics seek to drive home the point of exploring rates versus totals. On the left we see the violent crime rate in 2014 and on the right we see the violent crime totals in 2014. By looking at the totals, CA, TX, FL, and NY are by far the most offensive states. The top 4 most populated states are those states in that order. Therefore, that graphic doesn't give us much insight on the distribution of crime but more so on the population. On the other hand, the graphic on the right accounts for the population and shows just the rate per 100,000 inhabitants. From that visual, we see that Nevada, New Mexico, and Tennessee tend to be top offenders.

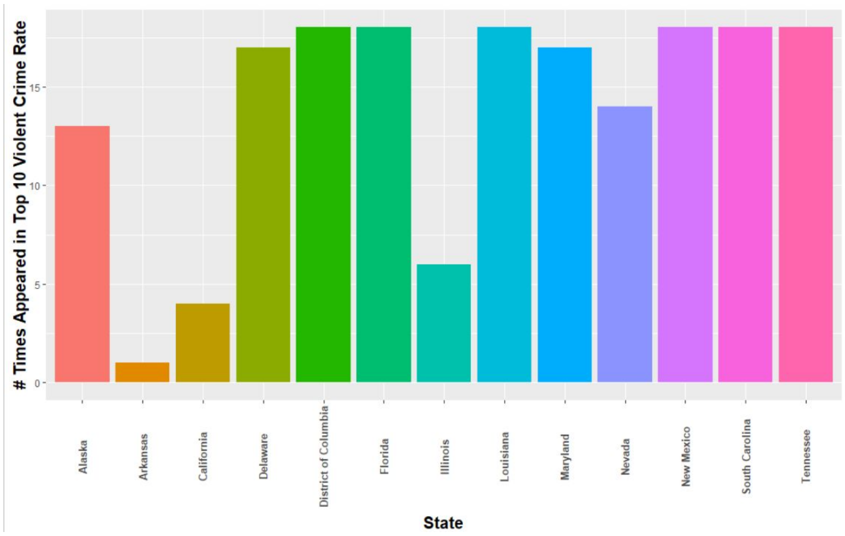
Top 10 violent crime rates by state in 2014 versus bottom 10 states:





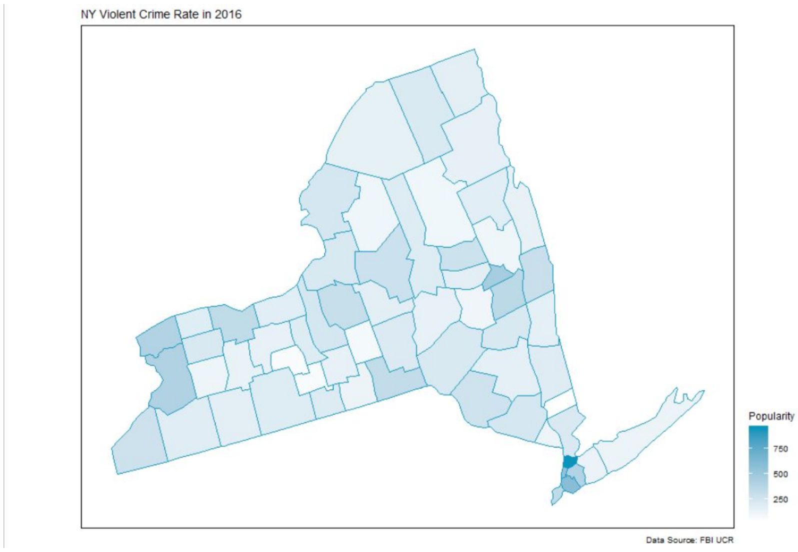
These graphics show the top 10 and bottom 10 states for violent crime rates in 2014. We see that Washington DC has the highest rate by a considerable margin followed by Alaska, Nevada and Tennessee. Vermont had the lowest rate followed by Maine, Wyoming, and New Hampshire.

The following plot looks at the number of times a state appeared in the top 10 for violent crime rates from 1997 to 2014:



This plot follows the top 10 states in violent crime rate by showing how many times a time was in the top 10 from 1997 to 2014. Washington DC, Florida, Louisiana, New Mexico, South Carolina, and Tennessee were all in the top 10 every single year for the past 17 years. Delaware and Maryland were similar, appearing in the top 10 16 out of the past 17 years.

I had planned to incorporate some data level data too but struggled to find the appropriate data, i.e. data that spans crimes, time, and counties. I would have had to do this for 50 files. Here is an example of what I intended to do with NY in 2016:



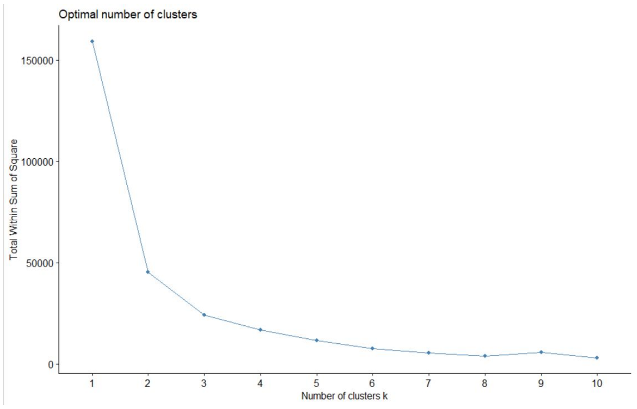
4. Model Development and Application of Models

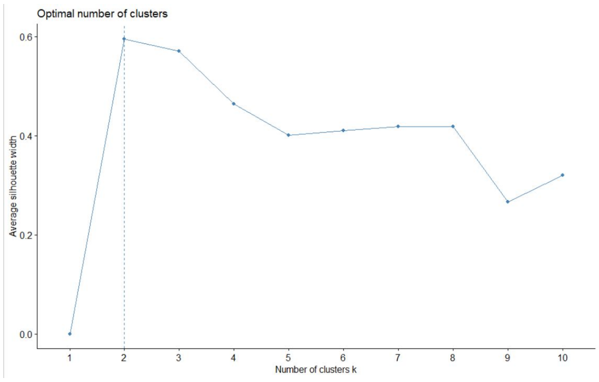
4.1 Model 1 - Clustering

4.1.1 Clustered National Level Data

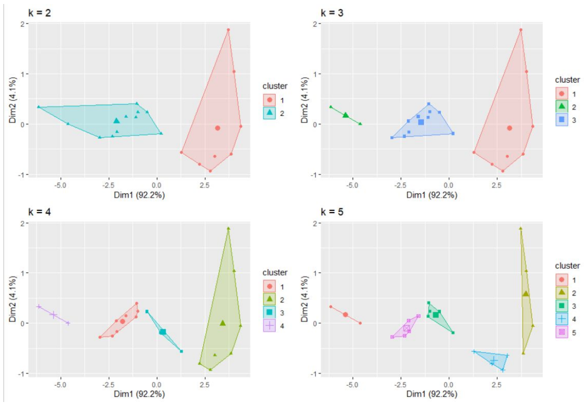
First, I did clustering on the national level data. This data has shown a sharp increase in the size of population and a sharp decrease in crime across all categories. It would then only make sense that the clustering is broken by year.

First to understand the optimal amount of clusters, I made an elbow chart and a silhouette chart. Here they are:

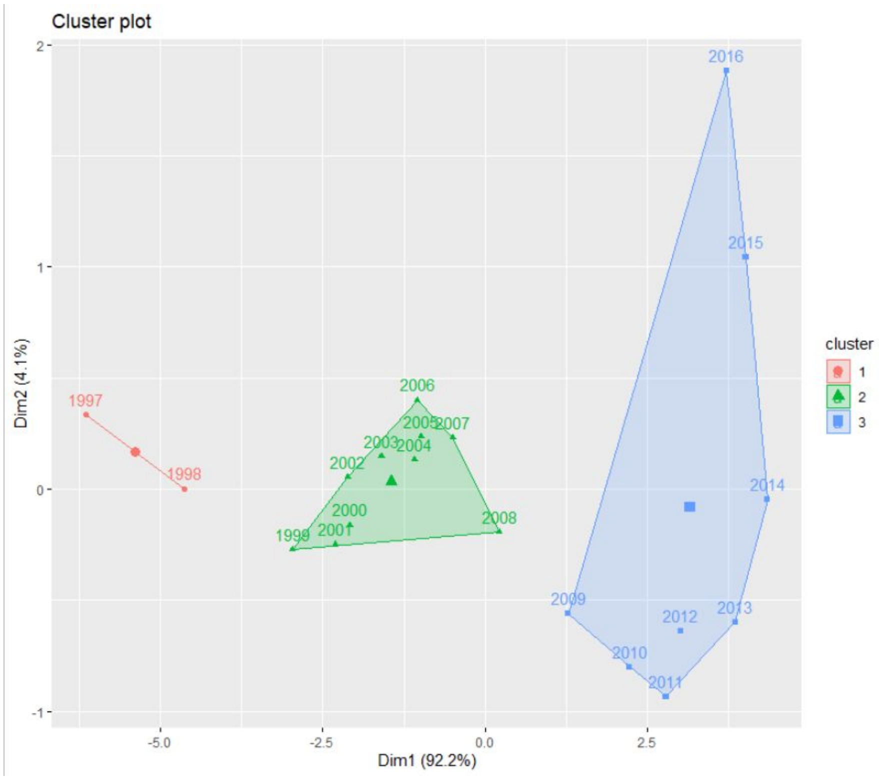




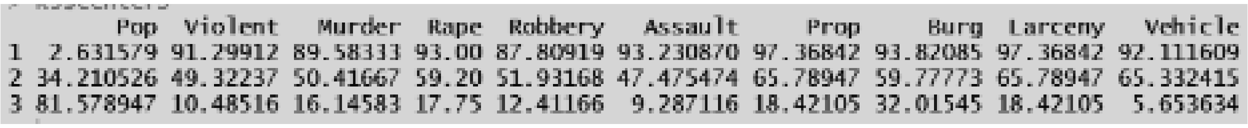
From these two graphics, it would seem that the optimal number of clusters is either 2 or 3. Here is a visual showing clusters 2, 3, 4, and 5:



From this, I make the judgment call that 3 clusters is ideal because I want to understand those two points on the far left side (cluster 2 when k=3). Here is the visual with 3 clusters:



You can see that the years read chronologically from left to right. Cluster 1 has 1997-1998, cluster 2 has 1999-2008, and cluster 3 has 2009-2016. We can assume that these clusters are defined by an increasing population and decreasing rates from left to right. Here are the definitions:

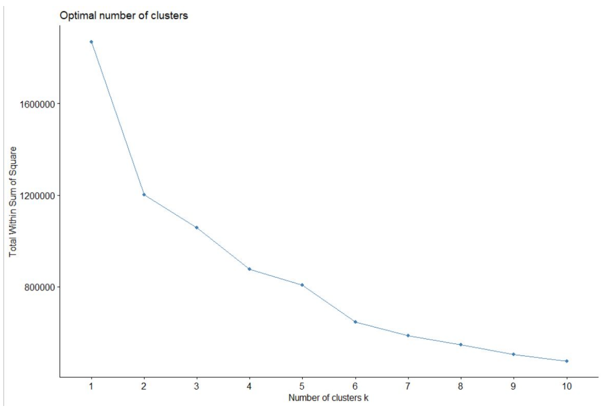


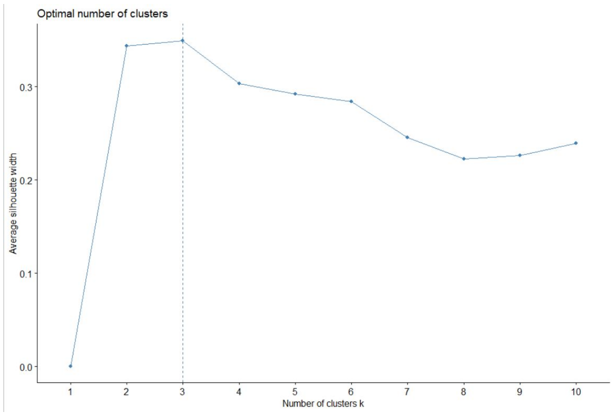
```
         Pop  Violent   Murder  Rape  Robbery   Assault     Prop     Burg  Larceny   Vehicle
1   2.631579 91.29912 89.58333 93.00 87.80919 93.230870 97.36842 93.82085 97.36842 92.111609
2  34.210526 49.32237 50.41667 59.20 51.93168 47.475474 65.78947 59.77773 65.78947 65.332415
3  81.578947 10.48516 16.14583 17.75 12.41166  9.287116 18.42105 32.01545 18.42105  5.653634
```

We can see that our ideas above were correct. As time progresses, the population is steeply increasing and across every category of crime we can see a decrease in the rate. Most notably, vehicle theft and assault have decreased the most.

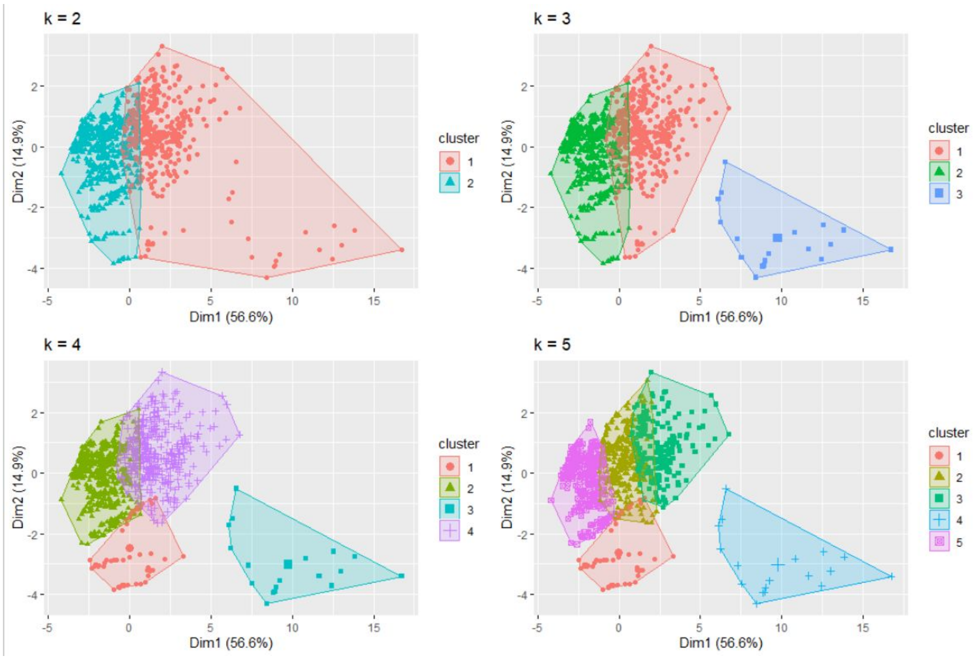
4.1.2 Clustered State Level Data

Next, I performed clustering on the state level data. Like before, first I did an elbow chart and a silhouette chart to try to understand the optimal number of clusters:

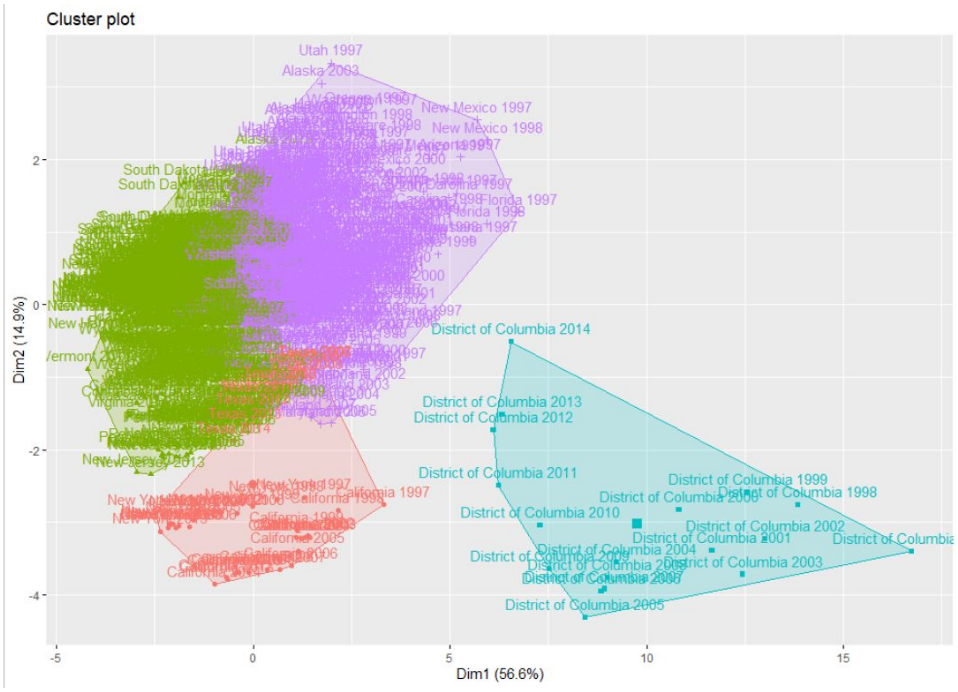




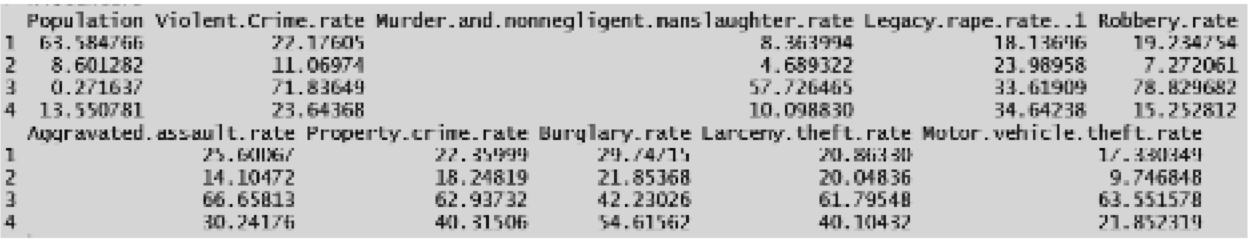
These don’t tell us too much as the lines aren’t very smooth. Here is a sample of the data ranging from two clusters to five:



It looks to me like three or four clusters are ideal. I will work with four given that there is so much data (51 territories over 17 years). Here is the clustering with four:



These results are very interesting. Cluster three, the blue cluster, is composed totally of Washington D.C. for all years (1997-2014). That means that DC was so different from the rest of the data that it remains its own cluster for 17 years. Unfortunately, this is because of high rates, not low rates. Cluster 1, the red cluster, is composed of NY, CA, and some TX. This is likely due to those states having the highest populations over time. Finally, clusters 2 and 4 are similar but ultimately split by something. Here we see the centers:



```
  Population Violent.Crime.rate Murder.and.nonnegligent.manslaughter.rate Legacy.rape.rate..1 Robbery.rate
1  63.584766           22.17605                                  8.363994            18.13696    19.234754
2   8.601282           11.06974                                  4.689322            23.98958     7.272061
3   0.271637           71.83649                                 57.726465            33.61909    78.829682
4  13.550781           23.64368                                 10.098830            34.64238    15.252812
  Aggravated.assault.rate Property.crime.rate Burglary.rate Larceny.theft.rate Motor.vehicle.theft.rate
1                25.60067            22.35999      29.74715           20.86330                17.330349
2                14.10472            18.24819      21.85368           20.04836                 9.746848
3                66.65813            62.93732      42.23026           61.79548                63.551578
4                30.24176            40.31506      54.61562           40.10432                21.852319
```

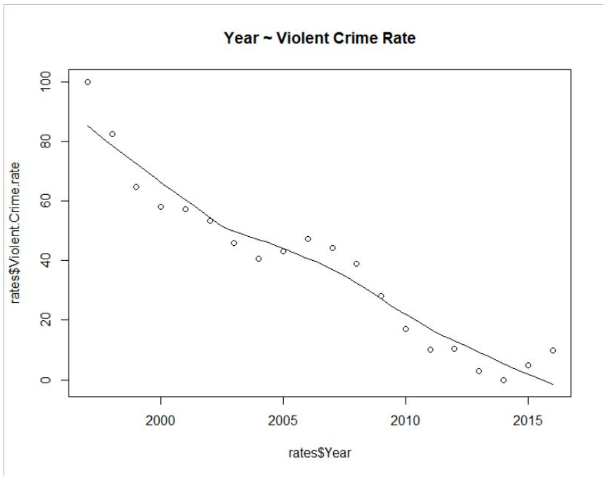
It appears cluster 4 has higher rates over time and a higher population whereas cluster 2 has a lower population and lower rates. Cluster 1 is defined largely by their high population and cluster 3, DC, has the highest rates in most categories (not in rape or burglary).

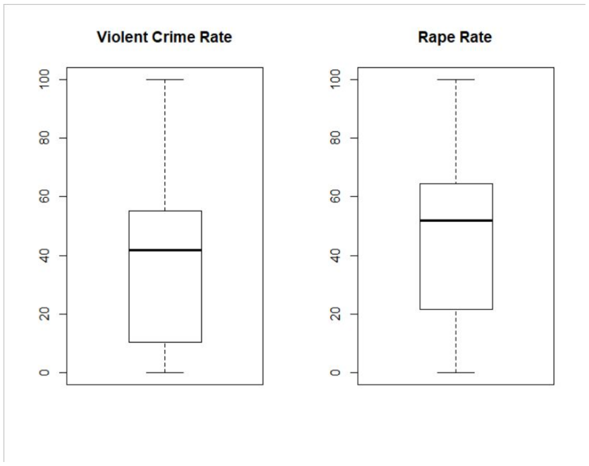
4.2 Model 2 - Regression

I performed some linear regression analysis on the nation-level data to try to understand future trends of certain crime rates and what other variables had the most impact on those rates. My prediction is that most rates will be most affected by time and population since we see that as time goes on and population grows, all rates decrease. I will build two models, one for the violent crime rate and one for the rape rate.

4.2.1 Regression on Violent Crime Rate

The goal here is to create a prediction regression model to understand violent crime rate trends and variable importance. Below are some graphics regarding violent crime rate:





On the left we can see that violent crime rates are decreasing over time. On the right we can see the boxplot for the violent crime rate and the rape rate. They are shaped similarly. They don’t have outliers. Their upper and lower bounds are the same (0 and 100 respectively) as the data has been scaled. The shape is on the lower end with Q1 around 10 and Q3 around 60.

The correlation between the violent crime rate and the population is -0.954.

When we fit a model to the full dataset, we get a strong R-squared that rounds to 1. Our f-statistic is high (2.59e+06) and our p-value is low ($p < 2.2e-16$). The strongest significant variables at *** or less than 0.001 significance are the rape rate, robbery rate, and the aggravated assault rate. Proving my hypothesis wrong, population and year are not significant. The only other significant variable is the murder rate at a significance level of 5%.

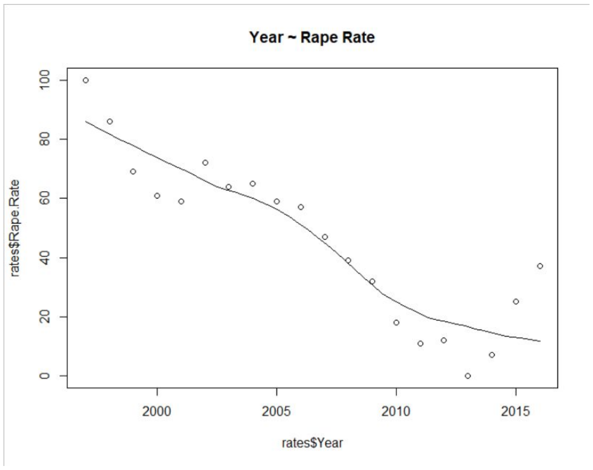
Next, I split the data with 80% train and 20% test to build a prediction model. This model shows the same ranking of variable importance except now the murder rate is no longer significant. Our actual versus predicted table looks like the following:

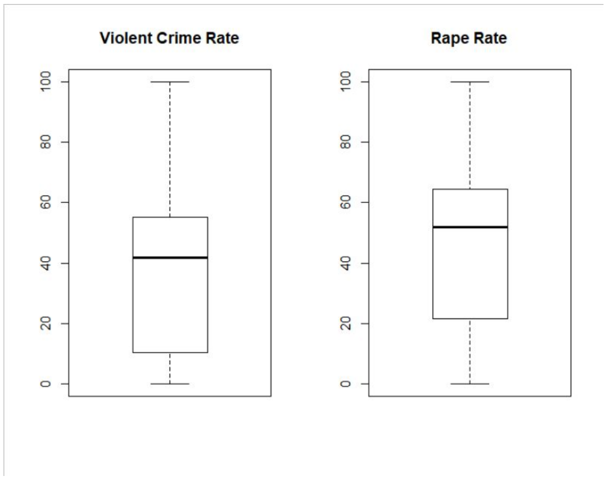
Actual	Predicted
100	99.952
57.298	57.404
0.00	0.067
9.904	9.873

As you can see, the model did a great job of predicting the actual violent crime rate. It has a min-max-accuracy of 74.86%. As for the MAPE, because we have an actual score of 0, we run into a “divide by 0” issue. If we replace that 0 value with 0.1, we get a MAPE of 8.49%. This model performs well.

4.2.2 Regression on Rape Rate

The goal with this model is similar to the violent crime rate. We want to develop a prediction model for the rape rate to understand trends and variable importance. Here are similar graphics as above:





Like with the violent crime rate, the rape rate is decreasing. The box plot is similar too, showing Q1 around 20 and Q3 around 65. There are no outliers.

The correlation between the rape rate and the population is -0.894. This is less strong than the correlation between the violent crime rate and the population but still strong nevertheless.

When we fit a model to the full dataset, we get a strong R-squared that rounds to 0.9997. Our f-statistic is high (4216) and our p-value is low ($p < 2.2e-16$). The strongest significant variables at a less than 0.001 significance level are the violent crime rate, robbery rate, and the aggravated assault rate. Proving my hypothesis wrong, population and year are not significant. The only other significant variable is the murder rate at a significance level of 5%. These results are very similar to the violent crime rate model.

Like with the last model, I split the data with 80% train and 20% test to build a prediction model. This model shows the same ranking of variable importance except now the murder rate is no longer significant. Our actual versus predicted table looks like the following:

Actual	Predicted

100	101.184
59	56.353
7	5.327
37	37.751

As you can see, the model did a great job of predicting the actual rape rate. It has a min-max-accuracy of 92.11% and a MAPE of 7.90% (there was no divide by 0 issue). This model performs really well. According to the min-max-accuracy and the MAPE, it performs better than the violent crime model.

These models ended up being very similar because of their strong correlation. We could predict that a model exploring the robbery rate and the aggravated assault rate would be similar to these two models too because of their correlation. A model exploring other crime rates that aren't related to these four would likely produce different results and would be an interesting area to explore further.

5. Conclusions and Discussion

5.1 Interpreting Results and Conclusion

At a surface level, as time goes on from 1997 to 2016 we see a consistent increase year to year in the population and a decrease in every single crime rate. Something is changing in society that enables the amount of people to go up and the amount of crime to go down.

Unfortunately, over time we do see the same areas as the most troublesome for the violent crime rate. Washington DC, Florida, Louisiana, New Mexico, South Carolina, and Tennessee were all in the top 10 every single year for the past 17 years. Delaware and Maryland were similar, appearing in the top 10 16 out of the past 17 years. Even though in general the rates are decreasing, these areas remain the most troublesome and should investigate ways to reduce their crime rates.

Building on that, Washington DC is a ***very*** troubled area, topping the charts in most crime rate categories over the last 17 years. The crime rates there exceed any other territory and must be worked on.

Populous areas aren’t necessarily the most crime ridden but taking a look at county level data could show that cities have higher rates. That would be an interesting next step.

Finally, we found through regression analysis that violent crime, rape, and aggravated assault are for some reason highly correlated.

5.2 Summarizing the Overall Process

My first goal was getting familiar with the data and cleaning it up. I created some visualizations, understood some important statistics around the data, and understood general trends. Next, I had to decide which ways I wanted to drill down into the data. With so many parameters like year, crime category, rate or totals, and areas, I had to target my analysis somewhere specific. Further, I had to decide modelling techniques. I found clustering to be informative as it would display similar and dissimilar areas and give insights like we got with the DC data. Regression seemed like a natural choice in the nation-level data.

Finally, it’s important to take a step back and understand results which is what I did above in 5.1.

5.3 Future Work

It would be worthwhile to have differently targeted questions and drill down on the data in different ways, i.e. looking at a specific crime category in a specific area or a specific period of time. This would be good data to create a dashboard for so that the user can drill down as they see fit.

Keeping up with current data and new year’s information is a good place to keep this project going. Finally as I mentioned before, getting a new level of data at a county level could be useful to specific states and areas.

6. Appendix

6.1 GitHub URL

Check out my work at:
https://github.com/scoutog/DataAnalyticsSpring2020/tree/master/FBI

6.2 Code

6.3.1 National-Level Code with Regression Models

```
library(BBmisc)
df <- read.csv("fbi_crimes.csv", nrows = 20)
df <- df[,1:20]
colnames(df)[colnames(df) == 'ï..Year'] <- 'Year'
summary(df)

sapply(df, class)
cols.num <- c(2,3,5,7,9,11,13,15,17,19,14,18)
df[,cols.num] <- sapply(df[cols.num],as.numeric)
sapply(df, class)

all <- c(2:20)
rescale <- function(x) (x-min(x))/(max(x) - min(x)) * 100
df[,all] <- sapply(df[all], rescale)
summary(df)
###################################################
library(tidyverse)
ggplot(data=df, aes(Year, Population)) +
  geom_point() + geom_line()

library(ggplot2)
library(reshape2)
d <- melt(df, id.vars="Year")

ggplot(d, aes(Year,value, col=variable)) +
  geom_point() + geom_line()

# Separate plots
ggplot(d, aes(Year,value, col=variable)) +
  geom_point(show.legend = FALSE) +
  stat_smooth(show.legend = FALSE) +
  facet_wrap(~variable)

###################################################
rate <- c(1,2,4,6,8,10,12,14,16,18,20)
rates <- df[,rate]

number <- c(1,2,3,5,7,9,11,13,15,17,19)
numbers <- df[,number]
```

```
d_rates <- melt(rates, id.vars="Year")

ggplot(d_rates, aes(Year,value, col=variable)) +
  geom_point() + geom_line()

# Separate plots
ggplot(d_rates, aes(Year,value, col=variable)) +
  geom_point(show.legend = FALSE) +
  stat_smooth(show.legend = FALSE) +
  facet_wrap(~variable)

d_nums <- melt(numbers, id.vars="Year")

ggplot(d_nums, aes(Year,value, col=variable)) +
  geom_point() + geom_line()

# Separate plots
ggplot(d_nums, aes(Year,value, col=variable)) +
  geom_point(show.legend = FALSE) +
  stat_smooth(show.legend = FALSE) +
  facet_wrap(~variable)

### CORRELATION
rates2 <- rates
colnames(rates2) <-
c("Year","Pop","Violent","Murder","Rape","Robbery","AggAss","Prop","Burg","Larceny","Vehicle")

res <- cor(rates2)
round(res, 2)

library("Hmisc")
res2<-rcorr(as.matrix(rates2[,2:11]))

library(corrplot)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)

corrplot(res2$r, type="upper", order="hclust",
         p.mat = res2$P, sig.level = 0.01, insig = "blank")
# Insignificant correlations are leaved blank
corrplot(res2$r, type="upper", order="hclust",
         p.mat = res2$P, sig.level = 0.01, insig = "blank")

xyz <- cor(rates2[,2:11])
corrplot.mixed(xyz, lower.col = "black", number.cex = .9)
corrplot(xyz, order="hclust", addrect=3, col = cm.colors(100))
corrplot(xyz, type = "lower", order = "hclust", tl.col = "black", tl.srt = 45)

####################################
####### Time Series Analysis #######
####### (Not used in report) #######

library(forecast)
```

```
violent.ts <- ts(rates$Violent.Crime.rate,start = c(1997),end = c(2016),freq=1)

nValid <- 12
violent_train <- window(violent.ts, start = c(1997), end = c(2010))
violent_valid <- window(violent.ts, start = c(2010), end = c(2016))

plot(violent_train,xlab="Time",ylab="Violent Crime Rate",ylim=c(min(rates$Violent.Crime.rate),
max(rates$Violent.Crime.rate)),bty="l")

## B
train.lm.season <- tslm(violent_train ~ trend, lambda=0)
train.lm.season.pred <- forecast(train.lm.season, h=6, level=0)
print(summary(train.lm.season))

plot(train.lm.season.pred,  ylab = "Violent Crime Rate", ,bty='l',xlab = "Time",xaxt="n",
ylim=c(min(rates$Violent.Crime.rate), max(rates$Violent.Crime.rate)),xlim = c(1997,2016), main = "", flty
= 2, col="red")
axis(1, at = seq(1997, 2016, 1))
lines(violent_train)
lines(train.lm.season$fitted, lwd = 2, col="blue")
lines(violent_valid)
grid()
lines(c(2010, 2010), c(min(rates$Violent.Crime.rate), max(rates$Violent.Crime.rate)),lwd=3,col="red")
text(1998, 30000000, "Training",cex=1)
text(2002, 40000000, "Validation",cex=1)
text(1992, 65000000, "Air", cex=1.5)

accuracy(train.lm.season$fitted, violent_valid)

library(TTR)
sma_vio <- SMA(violent.ts,n=1)
plot.ts(sma_vio)
vio_fore <- HoltWinters(violent.ts, gamma=F)
plot(vio_fore)

#####################################
####### Regression Analysis ########
#####################################
scatter.smooth(x=rates$Year, y=rates$Violent.Crime.rate, main = "Year ~ Violent Crime Rate")
scatter.smooth(x=rates$Year, y=rates$Rape.Rate, main = "Year ~ Rape Rate")

#Box plot for outlier
par(mfrow=c(1,2))
boxplot(rates$Violent.Crime.rate, main="Violent Crime Rate")
boxplot(rates$Rape.Rate, main="Rape Rate")

# Check density plot
library(e1071)
par(mfrow=c(1, 2))
plot(density(rates$Violent.Crime.rate), main="Density Plot: Violent Crime Rate", ylab="Frequency",
sub=paste("Skewness:", round(e1071::skewness(rates$Violent.Crime.rate), 2)))
polygon(density(rates$Violent.Crime.rate), col="red")
plot(density(rates$Rape.Rate), main="Density Plot: Rape Rate", ylab="Frequency",
sub=paste("Skewness:", round(e1071::skewness(rates$Rape.Rate), 2)))
```

```
polygon(density(rates$Rape.Rate), col="red")

# Correlation against time and population
cor(rates$Violent.Crime.rate, rates$Year)
cor(rates$Violent.Crime.rate, rates$Population)

cor(rates$Rape.Rate, rates$Year)
cor(rates$Rape.Rate, rates$Population)

#Model
vio.lm.fit <- lm(Violent.Crime.rate ~., data=rates)
summary(vio.lm.fit)

rape.lm.fit <- lm(Rape.Rate ~ ., data=rates)
summary(rape.lm.fit)
# Model assessment
model <- vio.lm.fit
modelSummary <- summary(model)
modelCoeffs <- modelSummary$coefficients
beta.estimate <- modelCoeffs["Rape.Rate", "Estimate"]
std.error <- modelCoeffs["Rape.Rate", "Std. Error"]
t_value <- beta.estimate/std.error
p_value <- 2*pt(-abs(t_value), df=nrow(rates)-ncol(rates))
f_statistic <- model$fstatistic[1]
f <- summary(model)$fstatistic
model_p <- pf(f[1], f[2], f[3], lower=FALSE)

t_value # larger t value is better
p_value # low is better
f_statistic
model_p
AIC(model)
BIC(model)

# Prediction models
set.seed(100)
trainingRowIndex <- sample(1:nrow(rates), 0.8*nrow(rates))
train <- rates[trainingRowIndex, ]
test  <- rates[-trainingRowIndex, ]

vio.mod <- lm(Violent.Crime.rate ~., train)
vio.pred <- predict(vio.mod, test)
summary(vio.mod)
AIC(vio.mod)

actuals_preds <- data.frame(cbind(actuals=test$Violent.Crime.rate, predicteds=vio.pred))
correlation_accuracy <- cor(actuals_preds)
actuals_preds

min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
min_max_accuracy
# => 74.86%, min_max accuracy
actuals_preds[3,1] <- 0.1
mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
mape # 8.498%

# RAPE MODEL

rape.mod <- lm(Rape.Rate ~., train)
rape.pred <- predict(rape.mod, test)
summary(rape.mod)
AIC(rape.mod)

actuals_preds <- data.frame(cbind(actuals=test$Rape.Rate, predicteds=rape.pred))
correlation_accuracy <- cor(actuals_preds)
head(actuals_preds)

min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
min_max_accuracy
# => 92.11%, min_max accuracy
mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
mape # 7.90%
```

6.3.2 State-Level Code

```
library(tidyverse)
df <- read.csv("crimes_state_time.csv")
df<-df[,-c(7,17)]
rescale <- function(x) (x-min(x))/(max(x) - min(x)) * 100
df[,c(3:21)] <- sapply(df[,c(3:21)], rescale)
rate<-c(1,2,3,13:21)
total<-c(1:12)
rates_df <- df[,rate]
total_df <- df[,total]

all <- c(3:12)

library(BBmisc)

fips <- read.csv("us-state-ansi-fips.csv")
names(fips)[names(fips) == "stname"] <- "State"

rates_df <- left_join(rates_df, fips, by = "State")
total_df <- left_join(total_df, fips, by = "State")

rates2010 <- subset(rates_df, Year == 2010)
names(rates2010)[names(rates2010) == "State"] <- "state"
names(rates_df)[names(rates_df) == "State"] <- "state"
names(total_df)[names(total_df) == "State"] <- "state"
##########
library(gtrendsR)
library(usmap)

orange <- "#0C95BC"

plot_usmap(data = rates2010, values = "Population",  color = orange, labels=FALSE) +
  scale_fill_continuous( low = "white", high = orange,
                         name = "Popularity", label = scales::comma
  ) +
  theme(legend.position = "right") +
  theme(panel.background = element_rect(colour = "black")) +
  labs(title = "Population in 2010", caption = "Source: FBI UCR")


plot_usmap(data = rates2010, values = "Population", include =  c(.south_atlantic, .mid_atlantic,
.new_england ), color = orange, labels=TRUE) +
  scale_fill_continuous( low = "white", high = orange,
                      name = "Popularity", label = scales::comma
  ) +
  theme(legend.position = "right") +
  theme(panel.background = element_rect(colour = "black")) +
  labs(title = "US East Coast Population", caption = "Source: FBI UCR")


#################
summary(rates_df$Year)

x <- "2014violent.jpg"
y <- 2014
z <- "Violent Crime Total in 2014"

#jpeg(x, width = 568, height = 376)
plot_usmap(data = subset(total_df, Year == y), values = "Violent.crime.total",  color = orange,
labels=FALSE) +
  scale_fill_continuous( low = "white", high = orange,
                      name = "Violent Crime", label = scales::comma
  ) +
  theme(legend.position = "right") +
  theme(panel.background = element_rect(colour = "black")) +
  labs(title = z, caption = "Source: FBI UCR")
#dev.off()

############


library(plyr)
library(dplyr)
library(ggplot2)
library(lubridate)
library(plotly)
library(RColorBrewer)
library(choroplethrMaps)
library(choroplethr)
library(tm)
library(wordcloud)
library(RColorBrewer)

by_state <- rates_df  %>% filter(Year == 2014) %>% group_by(state) %>% select(state, Year,
Violent.Crime.rate) %>%
  arrange(desc(Violent.Crime.rate))

head(by_state, 10)
```

```
tail(by_state, 10)

ggplot(head(by_state, 10), aes(reorder(state, -Violent.Crime.rate), Violent.Crime.rate, fill =
Violent.Crime.rate)) +
  geom_bar(stat = "identity") + xlab("State") + ylab("Violent Crime Total in 2014") +
  theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4, face = "bold"),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        axis.title.x = element_text(face = "bold", size = 15, vjust = -0.35),
        axis.title.y = element_text(face = "bold", vjust = 0.35, size = 15)) +
  theme(legend.position = "none" )

ggplot(tail(by_state, 10), aes(reorder(state, -Violent.Crime.rate), Violent.Crime.rate, fill =
Violent.Crime.rate)) +
  geom_bar(stat = "identity") + xlab("State") + ylab("Violent Crime Total in 2014") +
  theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4, face = "bold"),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        axis.title.x = element_text(face = "bold", size = 15, vjust = -0.35),
        axis.title.y = element_text(face = "bold", vjust = 0.35, size = 15)) +
  theme(legend.position = "none" )


top.states <- head(rates_df  %>% filter(Year == 1997) %>% group_by(state) %>% select(state, Year,
Violent.Crime.rate) %>%
       arrange(desc(Violent.Crime.rate)),10)




i <- 1998
while (i < 2015) {
  top.states <- rbind(top.states, head(rates_df  %>% filter(Year == i) %>% group_by(state) %>%
select(state, Year, Violent.Crime.rate) %>%
                                         arrange(desc(Violent.Crime.rate)),10))
  i = i+1
}

counts <- table(top.states$state)
counts <- as.data.frame(counts)
counts <- counts[counts$Freq != 0,]

ggplot(counts, aes(Var1, Freq, fill = Var1)) +
  geom_bar(stat = "identity") + xlab("State") + ylab("# Times Appeared in Top 10 Violent Crime Rate") +
  theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4, face = "bold"),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        axis.title.x = element_text(face = "bold", size = 15, vjust = -0.35),
        axis.title.y = element_text(face = "bold", vjust = 0.35, size = 15)) +
  theme(legend.position = "none" )

top.states[top.states$state == "Maryland",]
top.states[top.states$state == "North Carolina",]
top.states[top.states$state == "Ohio",]


###################
```

```
by_state <- total_df  %>% filter(Year == 2014) %>% group_by(state) %>% select(state, Year,
Legacy.rape..1) %>%
  arrange(desc(Legacy.rape..1))

head(by_state, 10)
tail(by_state, 10)

ggplot(head(by_state, 10), aes(reorder(state, -Legacy.rape..1), Legacy.rape..1, fill = Legacy.rape..1)) +
  geom_bar(stat = "identity") + xlab("State") + ylab("Rape Crime Total in 2014") +
  theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4, face = "bold"),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        axis.title.x = element_text(face = "bold", size = 15, vjust = -0.35),
        axis.title.y = element_text(face = "bold", vjust = 0.35, size = 15)) +
  theme(legend.position = "none" )

ggplot(tail(by_state, 10), aes(reorder(state, -Legacy.rape..1), Legacy.rape..1, fill = Legacy.rape..1)) +
  geom_bar(stat = "identity") + xlab("State") + ylab("Rape Crime Total in 2014") +
  theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4, face = "bold"),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        axis.title.x = element_text(face = "bold", size = 15, vjust = -0.35),
        axis.title.y = element_text(face = "bold", vjust = 0.35, size = 15)) +
  theme(legend.position = "none" )


top.states <- head(total_df  %>% filter(Year == 1997) %>% group_by(state) %>% select(state, Year,
Legacy.rape..1) %>%
                    arrange(desc(Legacy.rape..1)),10)




i <- 1998
while (i < 2015) {
  top.states <- rbind(top.states, head(total_df  %>% filter(Year == i) %>% group_by(state) %>%
select(state, Year, Legacy.rape..1) %>%
                                      arrange(desc(Legacy.rape..1)),10))
  i = i+1
}

counts <- table(top.states$state)
counts <- as.data.frame(counts)
counts <- counts[counts$Freq != 0,]

ggplot(counts, aes(Var1, Freq, fill = Var1)) +
  geom_bar(stat = "identity") + xlab("State") + ylab("Rape Crime Total in 2014") +
  theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4, face = "bold"),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        axis.title.x = element_text(face = "bold", size = 15, vjust = -0.35),
        axis.title.y = element_text(face = "bold", vjust = 0.35, size = 15)) +
  theme(legend.position = "none" )


#################
```

```
rates20 <- subset(rates_df, Year == 1997)


plot_usmap(data = rates20, values = "Violent.Crime.rate", include =  c(.north_central_region), color =
orange, labels=TRUE) +
  scale_fill_continuous( low = "white", high = orange,
                        name = "Violent Crimes Rate", label = scales::comma,
                        limits = c(80,870),) +
  theme(legend.position = "right") +
  theme(panel.background = element_rect(colour = "black")) +
  labs(title = "North Central US Violent Crime Rate", caption = "Source: FBI UCR")



ggplot(data=rates_df, aes(Year, Violent.Crime.rate)) +
  geom_point() +
  geom_smooth(color='red')
```

6.3.3 National-Level Clustering

```
library(BBmisc)
df <- read.csv("fbi_crimes.csv", nrows = 20)
df <- df[,1:20]
colnames(df)[colnames(df) == 'ï..Year'] <- 'Year'
summary(df)

sapply(df, class)
cols.num <- c(2,3,5,7,9,11,13,15,17,19,14,18)
df[,cols.num] <- sapply(df[cols.num],as.numeric)
sapply(df, class)

all <- c(2:20)
rescale <- function(x) (x-min(x))/(max(x) - min(x)) * 100

df[,all] <- sapply(df[all], rescale)
summary(df)

rate <- c(1,2,4,6,8,10,12,14,16,18,20)
rates_df <- df[,rate]

number <- c(1,2,3,5,7,9,11,13,15,17,19)
numbers_df <- df[,number]

row.names(rates_df)<-rates_df[,1]
rates_df <- rates_df[,-1]

colnames(rates_df) <-
c("Pop","Violent","Murder","Rape","Robbery","Assault","Prop","Burg","Larceny","Vehicle")

library(tidyverse)
library(cluster)
library(factoextra)
```

```
fviz_nbclust(rates_df, kmeans, method = "wss")
fviz_nbclust(rates_df, kmeans, method = "silhouette")

k2 <- kmeans(rates_df, centers = 2, nstart = 25)
str(k2)
k2
fviz_cluster(k2, data = rates_df)

k3 <- kmeans(rates_df, centers = 3, nstart = 25)
k4 <- kmeans(rates_df, centers = 4, nstart = 25)
k5 <- kmeans(rates_df, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = rates_df) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = rates_df) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = rates_df) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = rates_df) + ggtitle("k = 5")

library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)

fviz_cluster(k3, data=rates_df)
k3$centers


fviz_cluster(k5, data=rates_df)
k5$centers

#2016 highest pop, lowest prop, burg, larceny

#1997 lowest pop, highest every other crime

set.seed(123)
```

6.3.4 State-Level Clustering

```
library(tidyverse)
df <- read.csv("crimes_state_time.csv")
df<-df[,-c(7,17)]
rate<-c(1,2,3,13:21)
total<-c(1:12)
rates_df <- df[,rate]
total_df <- df[,total]
all <- c(3:12)
rates_df$state.year <- paste(rates_df$State,rates_df$Year)

# set row names to the state year
row.names(rates_df)<-rates_df[,13]
# remove the state, year, and state.year columns
rates_df <- rates_df[,-c(1,2,13)]
```

```
# normalize input variables
rescale <- function(x) (x-min(x))/(max(x) - min(x)) * 100
rates.df.norm <- sapply(rates_df,rescale)
rates.df.norm

# add row names: utilities
row.names(rates.df.norm)<-row.names(rates_df)

###########################
library(tidyverse)
library(cluster)
library(factoextra)

fviz_nbclust(rates.df.norm, kmeans, method = "wss")
fviz_nbclust(rates.df.norm, kmeans, method = "silhouette")

k2 <- kmeans(rates.df.norm, centers = 2, nstart = 25)
str(k2)
k2
fviz_cluster(k2, data = rates.df.norm)

k3 <- kmeans(rates.df.norm, centers = 3, nstart = 25)
k4 <- kmeans(rates.df.norm, centers = 4, nstart = 25)
k5 <- kmeans(rates.df.norm, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = rates.df.norm) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = rates.df.norm) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = rates.df.norm) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = rates.df.norm) + ggtitle("k = 5")

library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)

fviz_cluster(k4, data=rates.df.norm)
k4$centers

# Cluster 3 ALL DC (lowest pop, highest violent, murder, robbery, aggass, property, larceny, vehicle)
# Cluster 1 - NY, CA, TX (highest pop, high robbery, moderate else)
# Cluser 2 - splits remaining states (lowest levels of most)
# Cluster 4 - splits remaining states (higher pop, higher violent, murder, most rape,
#                                aggass, prop, most burgarly, larceny, vehicle)
```