Elizabeth Vetter, James Faller, and Rachael Chandler

Data Science 201: Introduction to Data Science

Professor Carver

Final Paper

12/16/2016

# Predicting Crime Rates

## Starting Question

Our initial goal for this project was to discover if illegal immigration affects other variables such as crime and unemployment rates. Focusing on all of the United States would require very large sums of data, so we decided to work with California, New York, Georgia and Texas. We chose these states because they each have very different policies regarding illegal immigration and have varying amounts of illegal immigrants. We found data from the years 1990-2012 to analyze. Our hope was to find data that represented a clear correlation between the three variables in a way that we could draw an accurate conclusion. If we found a clear correlation, we would hope to be able to predict a change in a state's crime rate given a certain change in the state's illegal immigration rate.

## Sources of Data

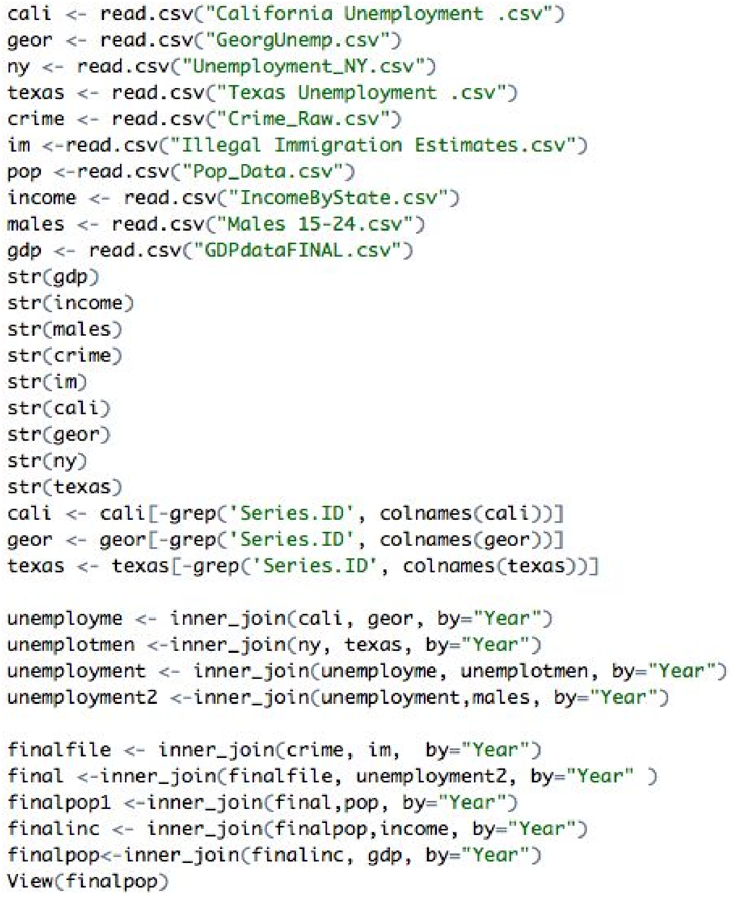
When searching for crime data, we wanted to focus on violent crime only. We found a dataset from the Federal Bureau of Investigation that gave us the number of violent crimes committed per 100,000 people. We knew that we would definitely need to know the yearly population for each of our states. This data from the Federal Bureau of Investigation also included populations, so we utilized that in our analysis. Our unemployment data is from the United States Census Bureau, which gave a very detailed dataset that provided monthly unemployment rates.

When scouring the internet for data on illegal immigration we found most government agencies didn't have much if any accurate data on the topic. We ended up looking for illegal immigration data at a couple of think tanks, eventually we settled on the Pew Research Center. The Pew Research Center had illegal immigration estimates for the years 1990, 1995, 2000, 2005, 2007, 2009, and 2012 based off of the government census. We used an interpolation function on these values to find the missing data between the years. Something we realized when working on our project was that there weren’t very many variables we were working with, and the variables that we did have data for didn’t seem to have any significant trends based on illegal immigration. We decided it would be best to add some other relevant factors to our dataset. Luckily, we found median income per household data for each state in the years 1990-2012 from the Bureau of Economic Research. We were also able to find datasets from the same agency that showed the changes in GDP of industries such as Agriculture, Construction, Retail, Government, ect. Given our knowledge on illegal immigration, we chose to filter our table to provide data on the Farming, Construction and Education Industry. We also got specific data on male population as a percentage of each state from the year 1990-2012. We looked for data on this demographic specifically because the male population age 15-24 causes a significant number of crimes relative to all other demographics. This data was provided by the U.S. census.

**The Analytic Approach**

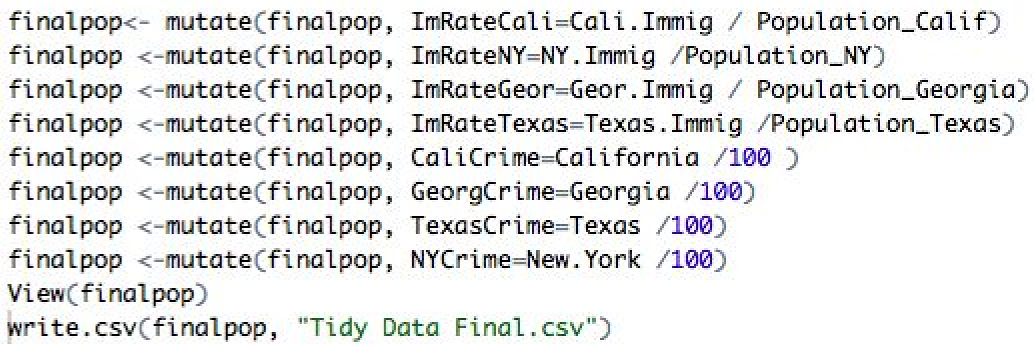
Before we were able to use our data in R, we had to do some tidying in Excel. In our data that we downloaded from our various sources, there were many columns that we didn’t need for our specific analysis. In our original unemployment dataset, monthly unemployment rates were recorded. Therefore, our dataset was 12 times larger than we wanted it to be. In Excel, we used the ‘Average’ function to take an average unemployment rate of each year. We then created a new column of these averages, and deleted other rows in the dataset. Our final tidy unemployment data consisted of 5 columns, “Year”, “Cali.Unemp”, “Geor.Unemp”, “NY.Unemp” and “Texas.Unemp”, with one row for each year. When we found the dataset that specified the GDP per year for specific industries, we wanted to choose industries that we thought would be relevant. Instead of including each industry included in the original dataset, we simplified our table to “Farming”, “Construction” and “Education” GDP’s.

We decided to focus on the years 1990-2012 in all of our separate datasets. Therefore, one of the columns in each of our tables was identical. We read each dataset into R. We then used the “inner_join” command in the dplyr package to join each of the tables together by “Year”. Initially, we used “inner_join” to join together the unemployment rates for each state. We then took that data and joined it with the crime and immigration tables.



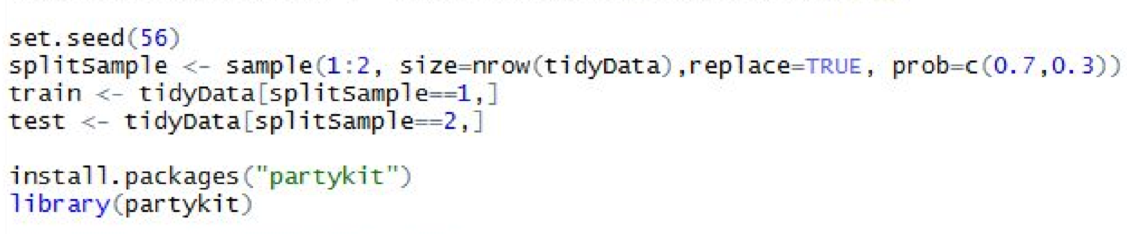
```
cali <- read.csv("California Unemployment .csv")
geor <- read.csv("GeorgUnemp.csv")
ny <- read.csv("Unemployment_NY.csv")
texas <- read.csv("Texas Unemployment .csv")
crime <- read.csv("Crime_Raw.csv")
im <-read.csv("Illegal Immigration Estimates.csv")
pop <-read.csv("Pop_Data.csv")
income <- read.csv("IncomeByState.csv")
males <- read.csv("Males 15-24.csv")
gdp <- read.csv("GDPdataFINAL.csv")
str(gdp)
str(income)
str(males)
str(crime)
str(im)
str(cali)
str(geor)
str(ny)
str(texas)
cali <- cali[-grep('Series.ID', colnames(cali))]
geor <- geor[-grep('Series.ID', colnames(geor))]
texas <- texas[-grep('Series.ID', colnames(texas))]

unemployme <- inner_join(cali, geor, by="Year")
unemplotmen <-inner_join(ny, texas, by="Year")
unemployment <- inner_join(unemployme, unemplotmen, by="Year")
unemployment2 <-inner_join(unemployment,males, by="Year")

finalfile <- inner_join(crime, im,  by="Year")
final <-inner_join(finalfile, unemployment2, by="Year" )
finalpop1 <-inner_join(final,pop, by="Year")
finalinc <- inner_join(finalpop,income, by="Year")
finalpop<-inner_join(finalinc, gdp, by="Year")
View(finalpop)
```

For better data analysis, we needed our variable measurements to be cohesive. Our dataset we were working with measured crime as a number per 100,000 people. We used the “mutate” command to take this data for each state and divide the number by 100. This made the data into a percentage that was easily comparable to the other variables. Similarly, our illegal immigration data was initially measured as quantities representing the number of illegal immigrants. We chose to use “mutate” to divide the number of illegal immigrants by the population. This command changed these numbers to percentages, which was much better for us to work with. We were then able to write a csv “finalpop” which included all of our data.



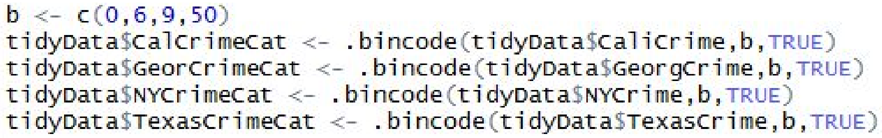
```
finalpop<- mutate(finalpop, ImRateCali=Cali.Immig / Population_Calif)
finalpop <-mutate(finalpop, ImRateNY=NY.Immig /Population_NY)
finalpop <-mutate(finalpop, ImRateGeor=Geor.Immig / Population_Georgia)
finalpop <-mutate(finalpop, ImRateTexas=Texas.Immig /Population_Texas)
finalpop <-mutate(finalpop, CaliCrime=California /100 )
finalpop <-mutate(finalpop, GeorgCrime=Georgia /100)
finalpop <-mutate(finalpop, TexasCrime=Texas /100)
finalpop <-mutate(finalpop, NYCrime=New.York /100)
View(finalpop)
write.csv(finalpop, "Tidy Data Final.csv")
```

Our first approach to analyze this data involved an attempt to make a classification tree, using the partykit package, to predict particular crime rates in every state. We started with California, and used unemployment, immigration, and population data as our predictors. We split the data into test and train datasets using the code below:



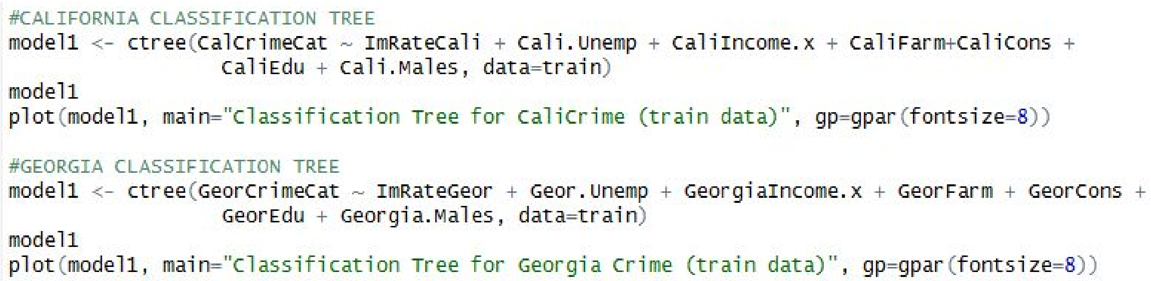
```
set.seed(56)
splitSample <- sample(1:2, size=nrow(tidyData),replace=TRUE, prob=c(0.7,0.3))
train <- tidyData[splitSample==1,]
test <- tidyData[splitSample==2,]

install.packages("partykit")
library(partykit)
```

The classification tree gave no inner nodes, so we decided to add more variables to predict the crime rate. After adding in industry data for construction, education, and farming, as well as the percentage of males between age 15-24, the classification algorithm still gave no inner nodes. Next, we realized that it may be difficult to predict a particular crime rate and a simpler approach may be to predict whether the crime rate will be low, medium, or high. To do this, we added a new variable for each state called "<State>CrimeCategory" that consisted of either a 1, 2, or 3 based on the state's crime rate. 1 corresponds to a crime rate below 3%, 2 correspond to a crime rate between 3 and 6% and 3 corresponds to a crime rate above 6%. The code used to perform this is outlined below:



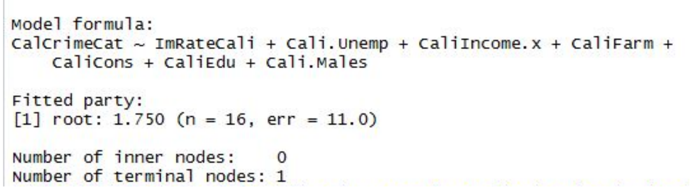
```
b <- c(0,6,9,50)
tidyData$CalCrimeCat <- .bincode(tidyData$CaliCrime,b,TRUE)
tidyData$GeorCrimeCat <- .bincode(tidyData$GeorgCrime,b,TRUE)
tidyData$NYCrimeCat <- .bincode(tidyData$NYCrime,b,TRUE)
tidyData$TexasCrimeCat <- .bincode(tidyData$TexasCrime,b,TRUE)
```

Then, we used commands from the partykit package to create classification trees:

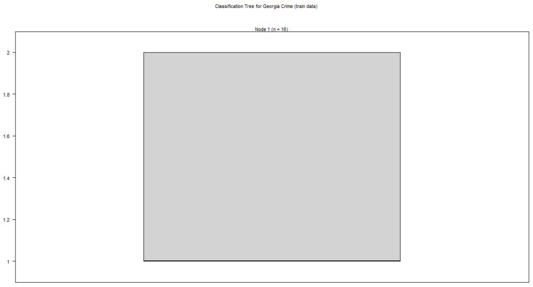


```
#CALIFORNIA CLASSIFICATION TREE
model1 <- ctree(CalCrimeCat ~ ImRateCali + Cali.Unemp + CaliIncome.x + CaliFarm+CaliCons +
                CaliEdu + Cali.Males, data=train)
model1
plot(model1, main="Classification Tree for CaliCrime (train data)", gp=gpar(fontsize=8))

#GEORGIA CLASSIFICATION TREE
model1 <- ctree(GeorCrimeCat ~ ImRateGeor + Geor.Unemp + GeorgiaIncome.x + GeorFarm + GeorCons +
                GeorEdu + Georgia.Males, data=train)
model1
plot(model1, main="Classification Tree for Georgia Crime (train data)", gp=gpar(fontsize=8))
```

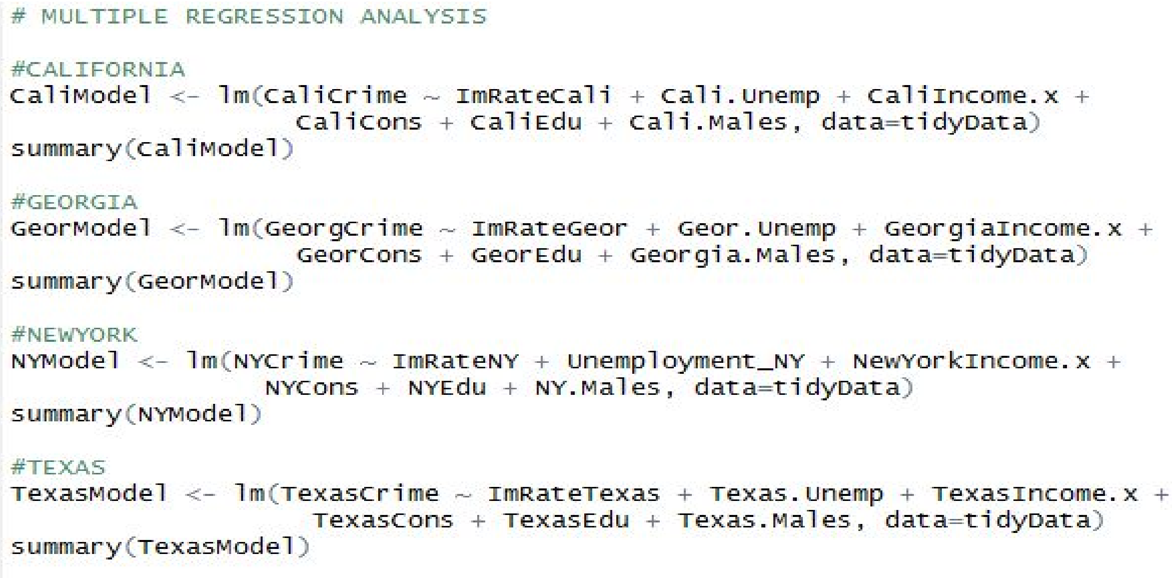
The following output was provided in the console and the plot window:



```
Model formula:
CalCrimeCat ~ ImRateCali + Cali.Unemp + CaliIncome.x + CaliFarm +
    CaliCons + CaliEdu + Cali.Males

Fitted party:
[1] root: 1.750 (n = 16, err = 11.0)

Number of inner nodes:    0
Number of terminal nodes: 1
```

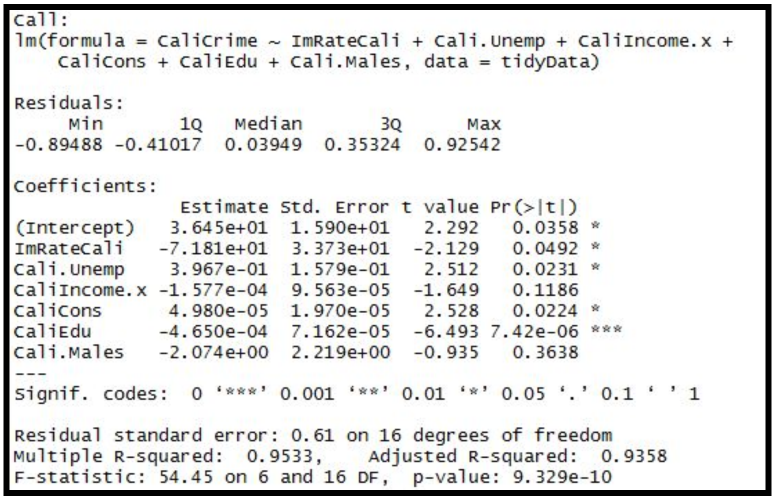


The same result occurred for every state, so we concluded that classification trees were not the optimal approach to analyze our data. Instead, we switched to multiple regression analysis, the examine the impact each variable had on the crime rate for each state. Because our data only consisted of 22 years, we decided to use the California, Georgia, and New York data as our training data and the Texas data as our testing data. The following code ran the regressions:



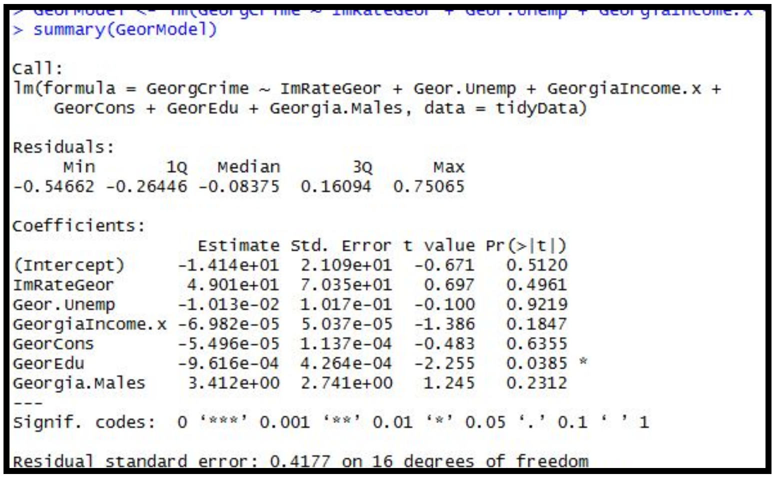
```
# MULTIPLE REGRESSION ANALYSIS

#CALIFORNIA
CaliModel <- lm(CaliCrime ~ ImRateCali + Cali.Unemp + CaliIncome.x +
                    CaliCons + CaliEdu + Cali.Males, data=tidyData)
summary(CaliModel)

#GEORGIA
GeorModel <- lm(GeorgCrime ~ ImRateGeor + Geor.Unemp + GeorgiaIncome.x +
                    GeorCons + GeorEdu + Georgia.Males, data=tidyData)
summary(GeorModel)

#NEWYORK
NYModel <- lm(NYCrime ~ ImRateNY + Unemployment_NY + NewYorkIncome.x +
                  NYCons + NYEdu + NY.Males, data=tidyData)
summary(NYModel)

#TEXAS
TexasModel <- lm(TexasCrime ~ ImRateTexas + Texas.Unemp + TexasIncome.x +
                     TexasCons + TexasEdu + Texas.Males, data=tidyData)
summary(TexasModel)
```

The summary function provided an overview of the model produced as well as the significance of each variable. The summary statistics for California were:



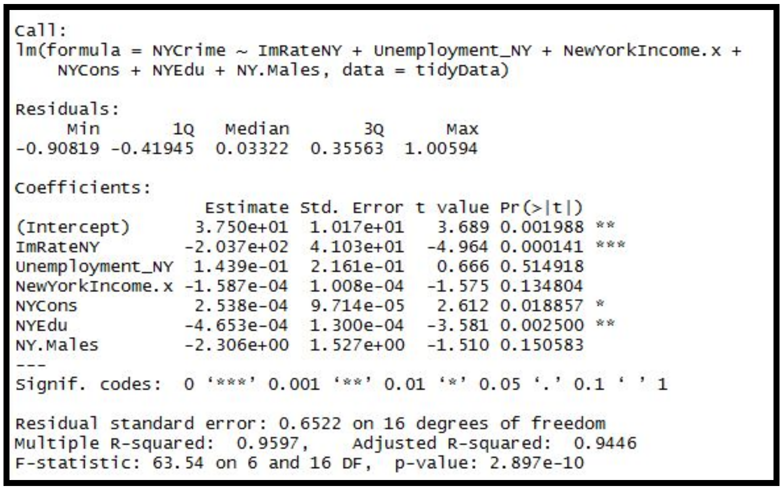
```
Call:
lm(formula = CaliCrime ~ ImRateCali + Cali.Unemp + CaliIncome.x +
    CaliCons + CaliEdu + Cali.Males, data = tidyData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89488 -0.41017  0.03949  0.35324  0.92542

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.645e+01  1.590e+01   2.292   0.0358 *
ImRateCali    -7.181e+01  3.373e+01  -2.129   0.0492 *
Cali.Unemp     3.967e-01  1.579e-01   2.512   0.0231 *
CaliIncome.x -1.577e-04  9.563e-05  -1.649   0.1186
CaliCons      4.980e-05  1.970e-05   2.528   0.0224 *
CaliEdu      -4.650e-04  7.162e-05  -6.493 7.42e-06 ***
Cali.Males   -2.074e+00  2.219e+00  -0.935   0.3638
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.61 on 16 degrees of freedom
Multiple R-squared:  0.9533,    Adjusted R-squared:  0.9358
F-statistic: 54.45 on 6 and 16 DF,  p-value: 9.329e-10
```

The summary statistics for Georgia were:



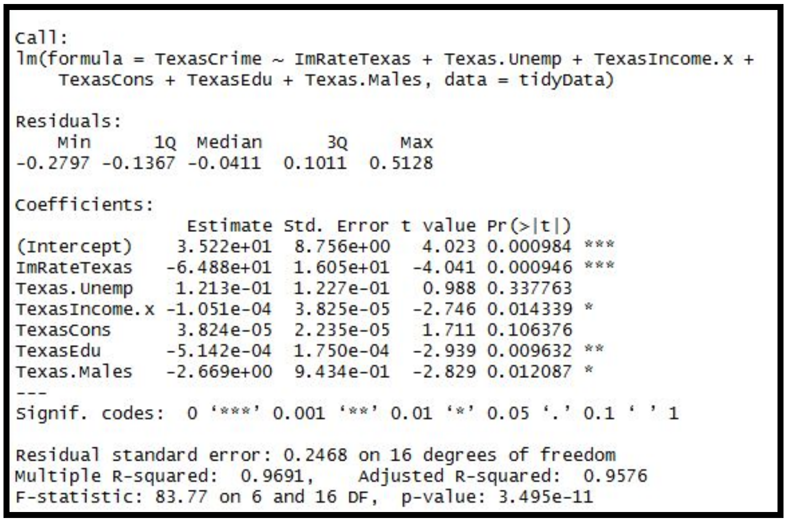
```
> GeorModel <- lm(GeorgCrime ~ ImRateGeor + Geor.Unemp + GeorgiaIncome.x
> summary(GeorModel)

Call:
lm(formula = GeorgCrime ~ ImRateGeor + Geor.Unemp + GeorgiaIncome.x +
    GeorCons + GeorEdu + Georgia.Males, data = tidyData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54662 -0.26446 -0.08375  0.16094  0.75065

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.414e+01  2.109e+01  -0.671   0.5120
ImRateGeor       4.901e+01  7.035e+01   0.697   0.4961
Geor.Unemp      -1.013e-02  1.017e-01  -0.100   0.9219
GeorgiaIncome.x -6.982e-05  5.037e-05  -1.386   0.1847
GeorCons        -5.496e-05  1.137e-04  -0.483   0.6355
GeorEdu         -9.616e-04  4.264e-04  -2.255   0.0385 *
Georgia.Males    3.412e+00  2.741e+00   1.245   0.2312
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4177 on 16 degrees of freedom
```

The summary statistics for New York were:



```
Call:
lm(formula = NYCrime ~ ImRateNY + Unemployment_NY + NewYorkIncome.x +
    NYCons + NYEdu + NY.Males, data = tidyData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90819 -0.41945  0.03322  0.35563  1.00594

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.750e+01  1.017e+01   3.689 0.001988 **
ImRateNY        -2.037e+02  4.103e+01  -4.964 0.000141 ***
Unemployment_NY  1.439e-01  2.161e-01   0.666 0.514918
NewYorkIncome.x -1.587e-04  1.008e-04  -1.575 0.134804
NYCons           2.538e-04  9.714e-05   2.612 0.018857 *
NYEdu           -4.653e-04  1.300e-04  -3.581 0.002500 **
NY.Males        -2.306e+00  1.527e+00  -1.510 0.150583
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6522 on 16 degrees of freedom
Multiple R-squared:  0.9597,    Adjusted R-squared:  0.9446
F-statistic: 63.54 on 6 and 16 DF,  p-value: 2.897e-10
```

The summary statistics for Texas were:



```
Call:
lm(formula = TexasCrime ~ ImRateTexas + Texas.Unemp + TexasIncome.x +
    TexasCons + TexasEdu + Texas.Males, data = tidyData)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2797 -0.1367 -0.0411  0.1011  0.5128

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.522e+01  8.756e+00   4.023 0.000984 ***
ImRateTexas   -6.488e+01  1.605e+01  -4.041 0.000946 ***
Texas.Unemp    1.213e-01  1.227e-01   0.988 0.337763
TexasIncome.x -1.051e-04  3.825e-05  -2.746 0.014339 *
TexasCons      3.824e-05  2.235e-05   1.711 0.106376
TexasEdu      -5.142e-04  1.750e-04  -2.939 0.009632 **
Texas.Males   -2.669e+00  9.434e-01  -2.829 0.012087 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2468 on 16 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9576
F-statistic: 83.77 on 6 and 16 DF,  p-value: 3.495e-11
```

## Description of Final Model

The final regression models for each state are presented below:

California :

$$Crime = 36.45 - 71.81(IM) + .398(UNEM) - .0002(INCOME) + .00005(CONS) - .0005(EDU) - 2.07(MALES)$$

Georgia:

$$Crime = -14.14+49(IM)-.01(UNEM)-.00007(INCOME)-.00005(CONS)-.00096(EDU)+3.41(MALES)$$

New York:

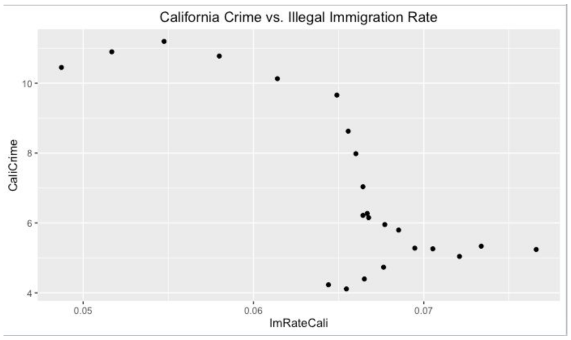
$$Crime = 37.5-203.7(IM)+.14(UNEM)-.00016(INCOME)+.00025(CONS)-.0005(EDU)-2.3(MALES)$$

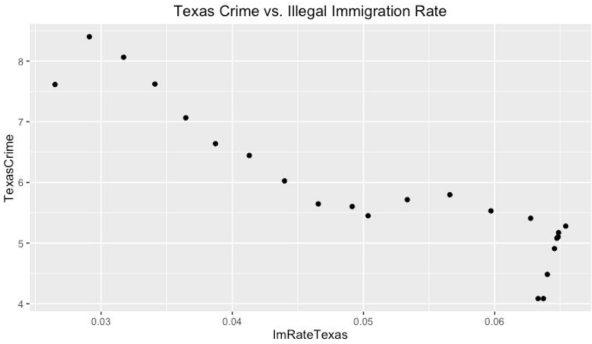
Texas:

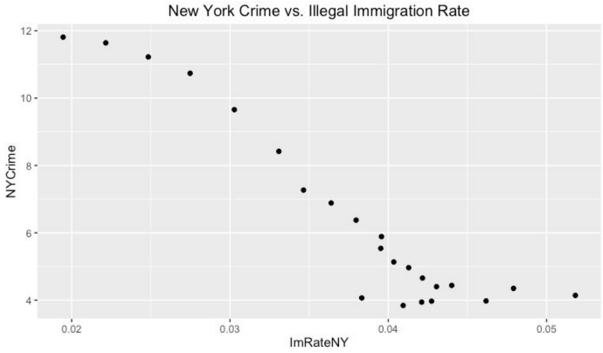
$$Crime = 35.2-64.9(IM)+.12(UNEM)-.0001(INCOME)+.00004(CONS)-.0005(EDU)-2.7(MALES)$$

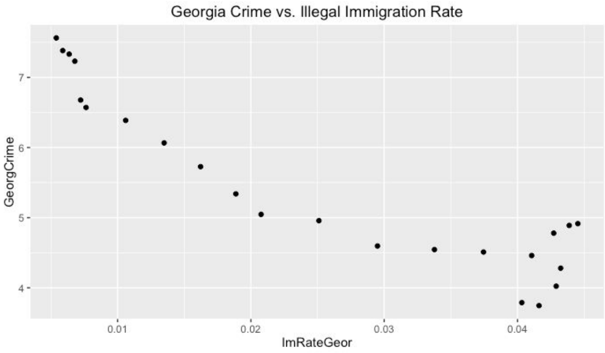
## Relevant Visualizations

These graphs compare Illegal Immigration Rate to Crime Rate:

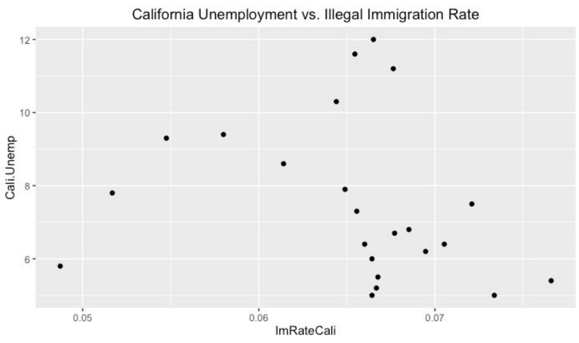


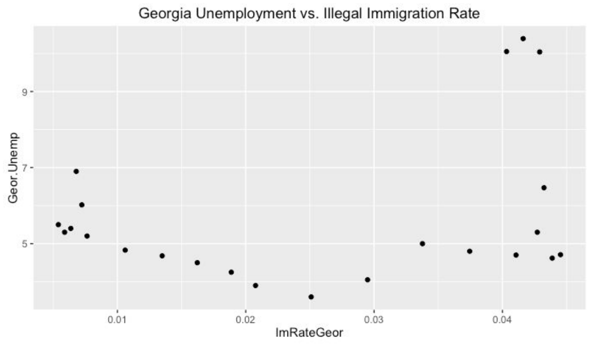


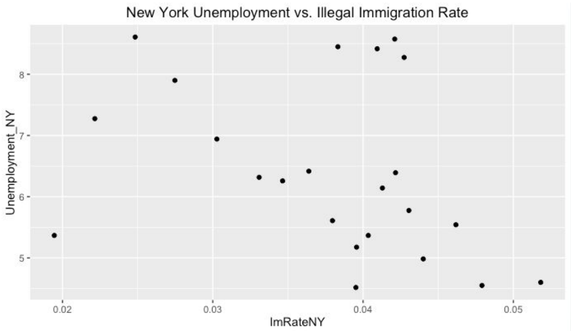


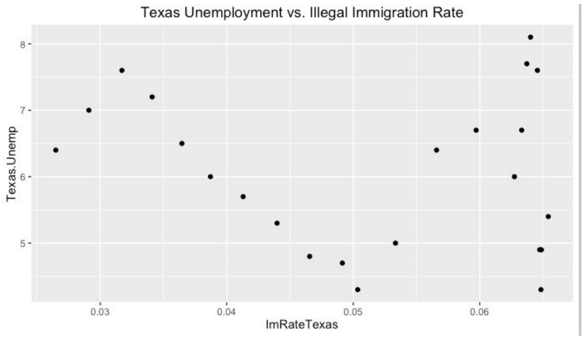


These graphs compare Illegal Immigration Rate to Unemployment Rate:









**Interpretation/Conclusion**

The regression analysis that was run gave very interesting results. Besides Georgia, all of our $R^2$ values were above .95, suggesting that the model fit the data well. The results in every state showed that education was a significant variable in predicting crime. Every coefficient was negative, therefore this suggests that an increase in the value of the education industry caused a decrease in the crime rate. On average, education had the highest t-values, making it the most significant variable in predicting crime rate.

The next most significant variable in predicting crime was immigration. The regression analysis estimated a negative coefficient on immigration for California, New York, and Texas. In Georgia, the coefficient was positive, but the absolute value of the t-value was less than 2, meaning that this variable is insignificant. The t-values for the immigration variable in California, New York and Texas suggested that this variable is significant. Also, the visualizations follow this trend as well by displaying a negative correlation between crime and immigration. As a result, our research supports the claim that an increase in immigration causes a decrease in the violent crime rate in four states. This contradicts popular opinion that immigration rates have a positive impact on crime rate.