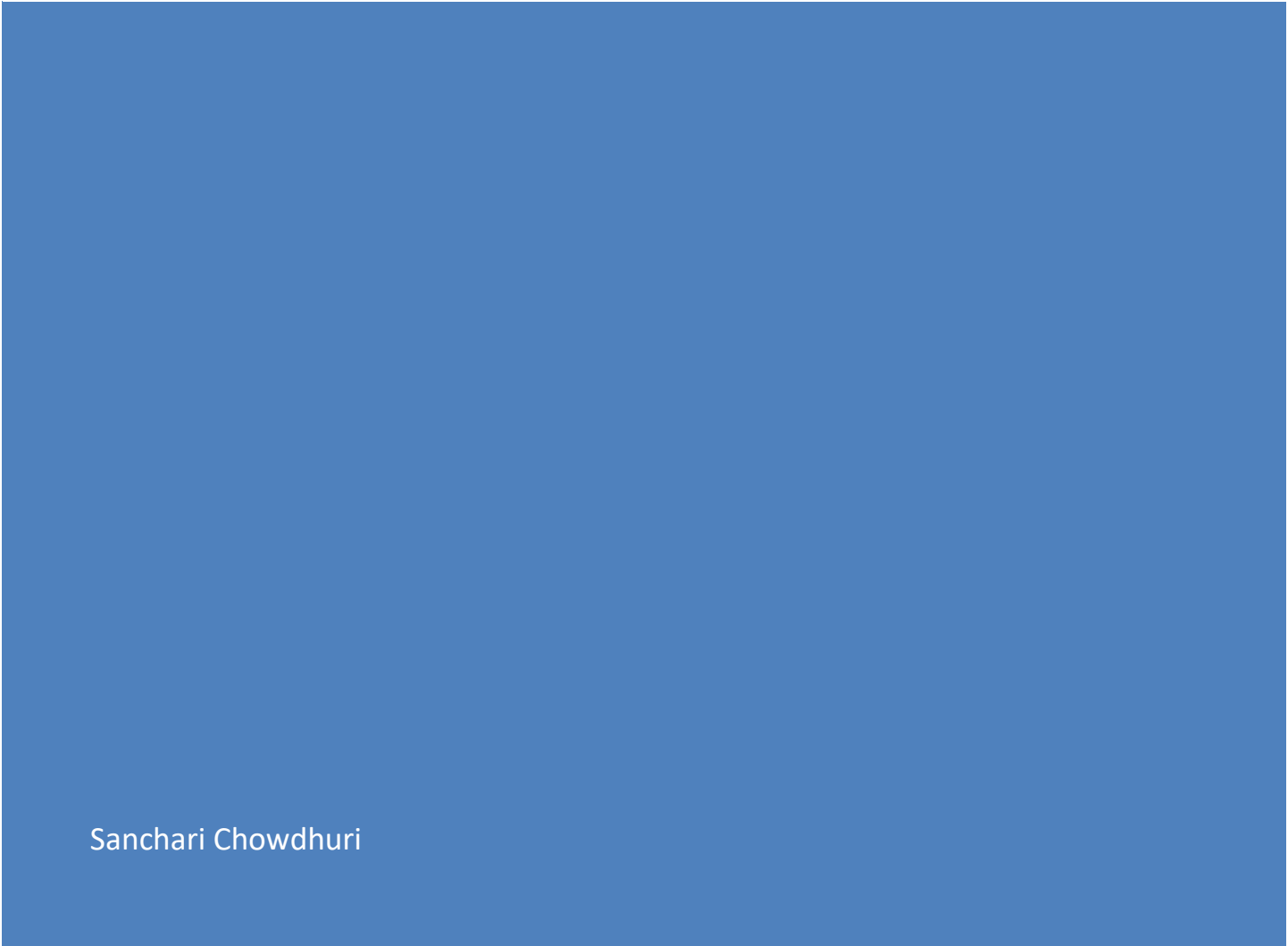




PROJECT REPORT

REGRESSION MODEL FOR THE CRIME RATE IN THE TRI-STATE AREA



Sanchari Chowdhuri

1. Introduction

The project aims at providing statistical analysis of crime trends and establishing a multifaceted inference of population and its impact on crime rate in the tristate area of United States for over a period of five decades (1960-2012). For the given analysis we have chosen a data set of crime statistics which was collected by the FBI from various law enforcement agencies that had participated in the Uniform Crime Reporting Statistics Program starting in 1929. This data is received from over 18,000 city, university/college, county, state, tribal, and federal law enforcement agencies voluntarily participating in the program. The crime data are submitted either through a state UCR Program or directly to the FBI's UCR Program.

The scope of the project elucidates the relationship between population of tristate area, crime rates and how state population influences the violent crime rates and property crime rates. Also with the help of the given dataset we establish a relationship of different violent crimes and how it impacts the total violent crime rate in a given state. A similar relationship for property crime rate is also established in the given paper.

1.1 Importance

This statistical analysis is important for understanding the relationship of population and crime rates in tristate area of United States over a period of five decades. This type of data gives a clear indication of public safety and how crime rates have been influenced with the changing times in society. The intermittent rise and falls of crime numbers with respect to the rising population indicates the way the society has developed.

Potential data user and decision makers for these data analysis can be law enforcements agencies, state planning commission office and lawmakers who can obtain a clear picture as to how to alter the implementation of law and order so that maximum efficiency in maintaining public safety can be obtained.

1.2 Data Collection and License

This data set can be obtained at state-by-state and national crime estimates by year(s). (n.d.). Retrieved September 30, 2016, from <http://www.ucrdatatool.gov/Search/Crime/State/StatebyState.cfm> the data present in the website is public domain and no special permission is required for accessing the data. Details of the same can be found at the US department of Justice Website. <https://www.justice.gov/legalpolicies#copyright>.

1.3 Description of Dataset

The ***population for our study*** are all 50 states of US from 1960-2012 for which dataset is provided while the sample is tristate area (New York, New Jersey and Connecticut). The type of study is '***Observational***' as we are observing the data set and drawing inferences based on our statistical analysis.

Taking the ***level of analysis*** as the "*Tristate area*", we come across a detailed description of population of each state from (1960-2012) and the respective crime rates. The crime report includes data of 7 different

types of crime segregated as **violent crime** (murder and non-negligent manslaughter, forcible rape, robbery, and aggravated assault) and **property crime** (burglary, larceny-theft, and motor vehicle theft).

The **research questions** which can be addressed by the given study are as follows.

1. Does a significant relationship exist between the property crime rate and the IVs - state and population?
2. Does a significant relationship exist between the violent crime rate and the IVs - state and population?

1.4 Expected Outcome

By the end of the statistical analysis we are expecting to find the following outcomes:

- Relationship between the different crime rates and the population of the states.
- Comparison between the Tri-states on the basis of their crime rates.

2. Research Method

Multiple regression is used to learn more about the relationship between several independent variables, also called as the predictor variables, and the dependent variable, also called as the criterion variable. Statisticians use multiple regression to understand the various ways in which a dependent variable can be described by the predictor variables and create a regression equation. This regression equation can be used as a prediction model to predict the outcome in a different setting.

There are several uses of Multiple Linear Regression Analysis. Some of them are:

- Causal Analysis: A correlation analysis identifies the strength of the relationship between the variables, while regression analysis is based on the assumption that there is a dependent or causal relationship between the predictor and criterion variables. Multiple linear regression can be used to identify the strength of that relationship.
- Forecasting an effect: Using the relationship derived from multiple linear regression, we can forecast the effect of changing the value of any of the predictor variable on the independent variable.
- Trend forecasting: From the multiple linear regression analysis model, prediction about the future trends can be made.

2.1 Our study

Our research questions are:

1. Does a significant relationship exist between the property crime rate and the IVs - state and population?
2. Does a significant relationship exist between the violent crime rate and the IVs - state and population?

From the above, we aim to understand if there is a relationship between the crime rate with the two independent variables - State and Population. The crimes are divided into two main types – **Violent crime and Property crime**. Violent crime is further divided into murder and non-negligent manslaughter rate, forcible rape rate, robbery rate, and aggravated assault rate. Property crime rate is divided into Burglary rate, larceny-theft rate, and motor vehicle theft rate.

Since our study involves one categorical predictor i.e. the state, we need to use an indicator variable to dummy code the variable. It consists of 3 levels that are New Jersey, New York and Connecticut. Theoretically, the three states can be dummy coded as shown in the table below:

State	New Jersey	New York
Connecticut (1)	0	0
New Jersey (2)	1	0
New York (3)	0	1

Here, state has three levels – Connecticut (1), New Jersey(2), and New York(3). The three levels can be dummy coded into two variables, New Jersey and New York. If State = 1 then both New Jersey and New York are coded as 0; if state = 2 then New Jersey is 1 and New York is 0; and if state = 3 then New Jersey is 0 and New York is 1.

Our model incorporates the multiple linear regression with dummy coding for the categorical variable.

3. Results

The results for our two research questions are as follows:

3.1 Does a significant relationship exist between the Property Crime Rate and the IVs - state and population?

```
Call:
lm(formula = Property.crime.rate ~ Population + State, data = tristatedata)

Residuals:
    Min       1Q   Median       3Q      Max
-3186.1  -818.8   -28.3   1021.1   2108.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.592e+03  5.592e+02  10.000 < 2e-16 ***
Population   -6.853e-04  1.681e-04  -4.076 7.40e-05 ***
StateNew Jersey  3.128e+03  7.909e+02   3.955 0.000118 ***
StateNew York   1.078e+04  2.564e+03   4.205 4.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1179 on 150 degrees of freedom
Multiple R-squared:  0.1156, Adjusted R-squared:  0.09794
F-statistic: 6.537 on 3 and 150 DF, p-value: 0.0003481
```

The overall test results are $F(3,150) = 6.537$, $p = 0.0003481$. Since p-value is lesser than the significance level ($\alpha = 0.05$) we reject the null hypothesis and conclude that there is a relationship between the Property Crime Rate and the IVs- state and population. The above model takes 'Connecticut' as our baseline or reference variable and 'New Jersey' and 'New York' as the two indicator variables. So, the overall model has $k-1$ levels, i.e. $3-1 = 2$ levels.

Individually, the predictors explain the DV in the following way:

- i. The population of the states is negatively related with the property crime rate. For each 1 unit increase in the population there is a decrease of 6.853×10^{-04} unit in the property crime rate.
- ii. The state of New Jersey has a higher property crime rate as compared to the state of Connecticut by 3.128×10^{03}
- iii. The state of New York has a higher property crime rate as compared to the state of Connecticut by 1.078×10^{04}

The model can be written as

$$\text{PropertyCrimeRate} = 5.592 \times 10^{03} - 6.853 \times 10^{-04}(\text{Population}) + 3.128 \times 10^{03}(\text{StateNewJersey}) + 1.078 \times 10^{04}(\text{StateNewYork})$$

The adjusted R^2 value is 0.097 which means that the predictors collectively represent 9.7% variance in the property crime rate.

3.2 Does a significant relationship exist between the Violent Crime Rate and the IVs - state and population?

```
Call:
lm(formula = violent.Crime.rate ~ Population + State, data = tristatedata)

Residuals:
    Min       1Q   Median       3Q      Max
-423.45 -146.14   18.56  135.31  427.13

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.889e+02   9.029e+01   5.415 2.38e-07 ***
Population   -5.964e-05   2.714e-05  -2.197 0.02953 *
StateNew Jersey  3.735e+02   1.277e+02   2.925 0.00398 **
StateNew York  1.338e+03   4.140e+02   3.232 0.00151 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 190.4 on 150 degrees of freedom
Multiple R-squared:  0.4902, Adjusted R-squared:  0.48
F-statistic: 48.08 on 3 and 150 DF, p-value: < 2.2e-16
```

The overall test results are $F(3,150) = 48.08$, $p < 2.2 \times 10^{-16}$. Since p-value is lesser than the significance level ($\alpha = 0.05$) we reject the null hypothesis and conclude that there is a relationship between the Violent Crime Rate and the IVs - state and population. The above model takes 'Connecticut' as our baseline or reference variable and 'New Jersey' and 'New York' as the two indicator variables. So, the overall model has $k-1$ levels, i.e. $3-1 = 2$ levels.

Individually, the predictors explain the DV in the following way:

- i. The population of the states is negatively related with the violent crime rate. For each 1 unit increase in the population there is a decrease of 5.964×10^{-05} unit in the violent crime rate.
- ii. The state of New Jersey has a higher violent crime rate as compared to the state of Connecticut by 3.735×10^{02}

- iii. The state of New York has a higher property crime rate as compared to the state of Connecticut by 1.338×10^{03}

The model can be written as

$$\text{ViolentCrimeRate} = 4.889 \times 10^{02} - 5.964 \times 10^{-54}(\text{Population}) + 3.735 \times 10^{02}(\text{StateNewJersey}) + 1.338 \times 10^{03}(\text{StateNewYork})$$

The adjusted R^2 value is 0.48 which means that the predictors collectively represent 48% variance in the violent crime rate.

4. Discussion

After studying the results of the regression model conducted on the data we come to conclusion that there is a significant effect of the Independent Variables (State and Population) on the Dependent Variable (Crime Rate). The two linear models obtained for our research questions are:

$$\text{PropertyCrimeRate} = 5.592 \times 10^{03} - 6.853 \times 10^{-04}(\text{Population}) + 3.128 \times 10^{03}(\text{StateNewJersey}) + 1.078 \times 10^{04}(\text{StateNewYork})$$

$$\text{ViolentCrimeRate} = 4.889 \times 10^{02} - 5.964 \times 10^{-54}(\text{Population}) + 3.735 \times 10^{02}(\text{StateNewJersey}) + 1.338 \times 10^{03}(\text{StateNewYork})$$

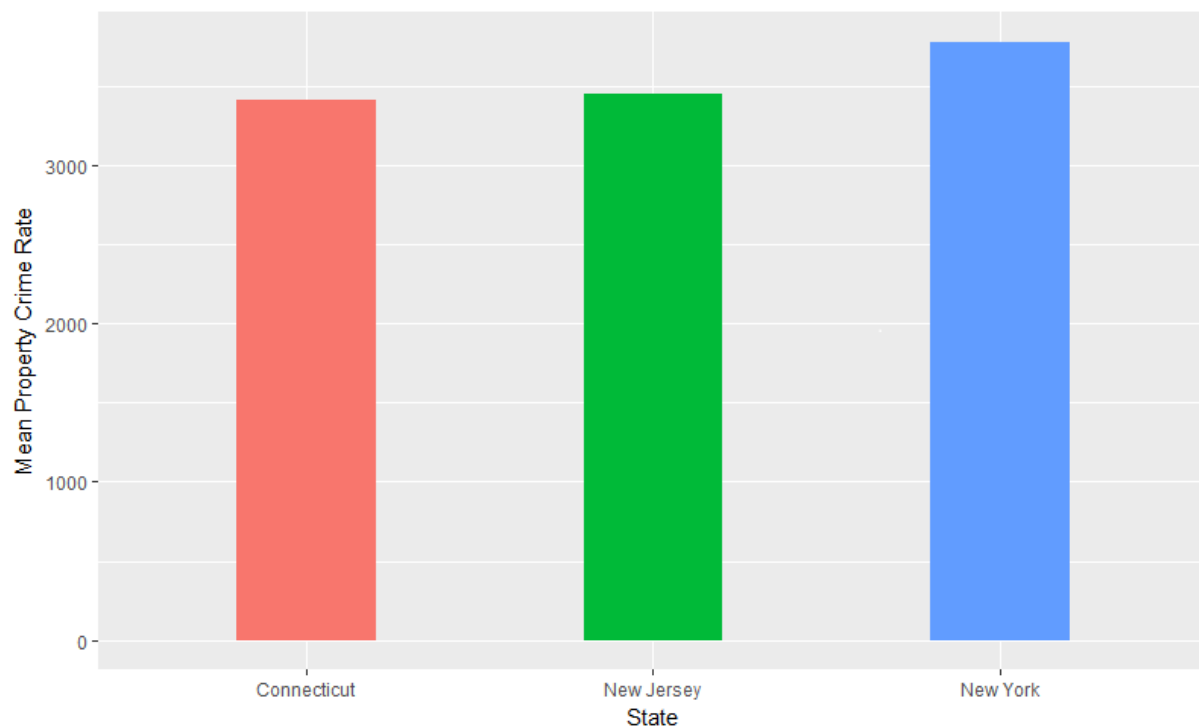


Figure 1: Bar graph: Mean of Property Crime Rate of each state

The bar graphs in Fig.1 and Fig.2 show a comparison of the mean of both the crime rates – Violent crime rate and Property crime rate – for the three states. They show that there is a considerable difference in the violent crime rate of the three states with New York having the highest mean and Connecticut having the lowest. On the other hand, mean property crime rate does not have much difference between the states.

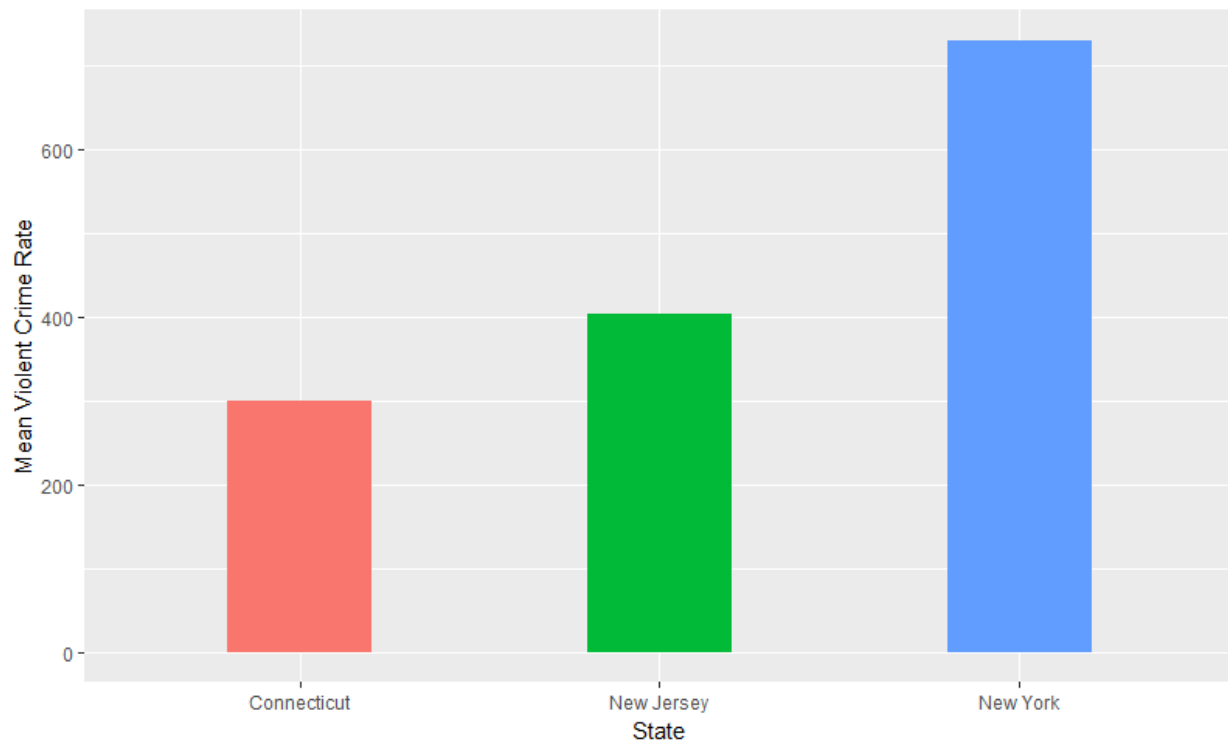


Figure 2: Bar graph for Mean Violent Crime Rate of each state

4.1 Expected vs. Actual Outcome

We expected to see a positive relationship between the population and the crime rates. However, the regression model obtained for both the questions give us a negative relationship between the population and the two crime rates (property crime rate and violent crime rate).

We expected to see a significant difference between the crime rates in the various states. The actual outcome shows that there is a significant difference between the crime rates in the three states of New Jersey, New York and Connecticut.

The order of crime rates (property and violent crime) for the three states are as follows:

$$\text{CrimeRate}_{\text{New York}} > \text{CrimeRate}_{\text{New Jersey}} > \text{CrimeRate}_{\text{Connecticut}}$$

This result is on the lines of what we had expected before running the test.

4.2 Future Scope

Currently, we have just studied the relationship between crime rates in the Tristate area. For future scope we can study the relationship of the crime rates in all the states of United States of America and see if there are any changes in the linear model.

Currently we have not studied the interactions between the two independent variables - state and population. We can study the interaction between the two by performing an interaction regression test.

References

1. Multiple regression with categorical predictor variables - text. Retrieved December 7, 2016, from <http://www.psychstat.missouristate.edu/multibook/mlt08m.html>
2. Uniform crime reporting statistics. (2003). Retrieved December 7, 2016, from <https://www.ucrdatatool.gov/Search/Crime/State/StateCrime.cfm>

REGRESSION MODEL FOR THE CRIME RATE IN THE TRI-STATE AREA

We read the data into the object 'TriStateCrimeData'

```
TriStateCrimeData <- read.csv("F:/MIM/627-Project/TriStateCrimeData.csv")
attach(TriStateCrimeData)
```

Conducting Multiple Linear Regression with the indicator variable as State which has three levels - Connecticut, New Jersey and New York.

Independent variable - population and state Dependent variable - property crime rate

1. Does a significant relationship exist between the property crime rate and the IVs - state and population?

```
m1 = lm(PropertyCrimeRate~Population+State,data = TriStateCrimeData)
summary(m1)
```

```
##
## Call:
## lm(formula = PropertyCrimeRate ~ Population + State, data = TriStateCrimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3186.1  -818.8   -28.3   1021.1   2108.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.592e+03  5.592e+02  10.000 < 2e-16 ***
## Population     -6.853e-04  1.681e-04  -4.076 7.40e-05 ***
## StateNew Jersey  3.128e+03  7.909e+02   3.955 0.000118 ***
## StateNew York   1.078e+04  2.564e+03   4.205 4.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1179 on 150 degrees of freedom
## Multiple R-squared:  0.1156, Adjusted R-squared:  0.09794
## F-statistic: 6.537 on 3 and 150 DF,  p-value: 0.0003481
```

Conducting Multiple Linear Regression with the indicator variable as State which has three levels -

Connecticut, New Jersey and New York

Independent variable - population and state Dependent variable - violent crime rate

2. Does a significant relationship exist between the violent crime rate and the IVs - state and population?

```
m2 = lm(ViolentCrimeRate~Population+State,data = TriStateCrimeData)
summary(m2)
```

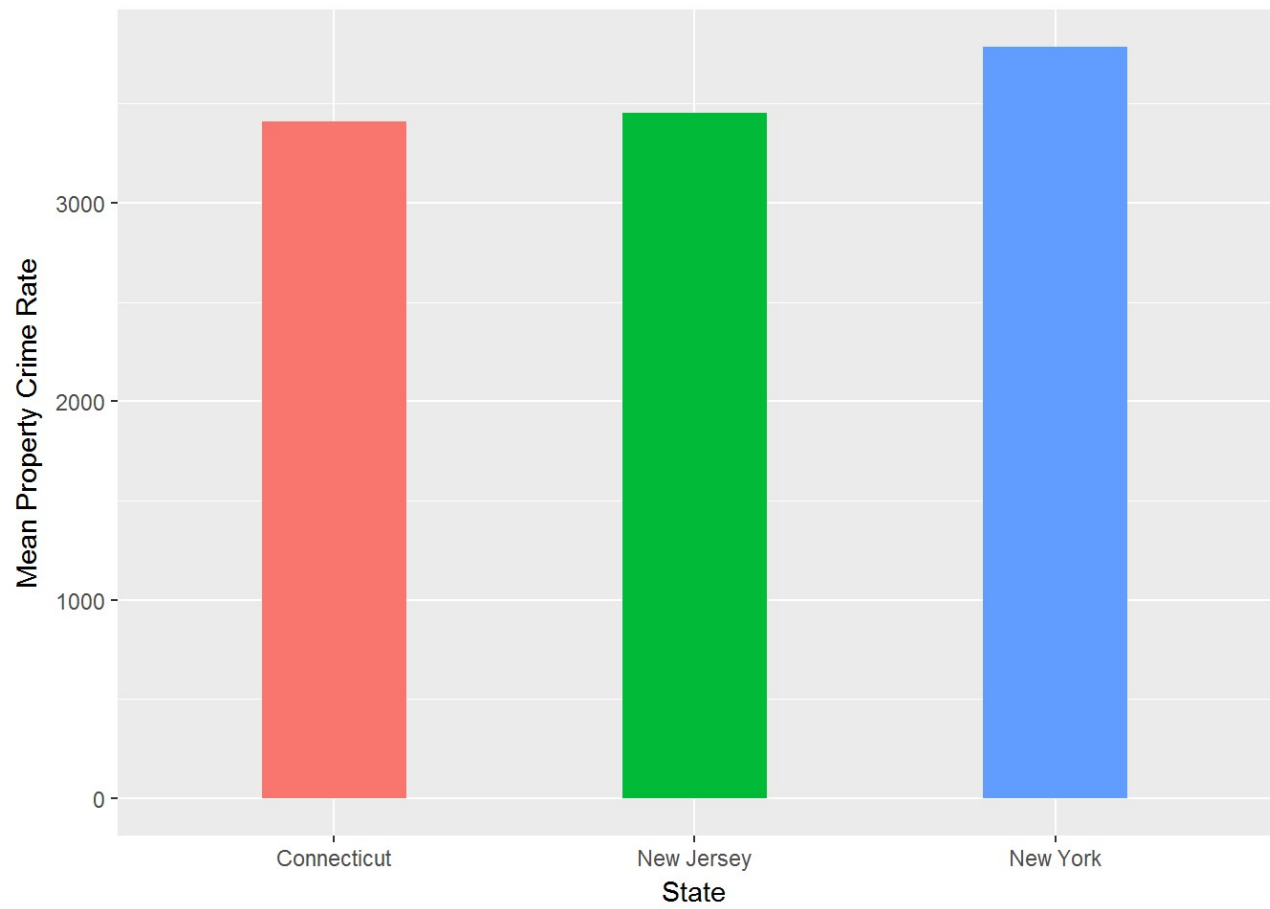
```
##
## Call:
## lm(formula = ViolentCrimeRate ~ Population + State, data = TriStateCrimeData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -423.45 -146.14   18.56  135.31  427.13
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.889e+02  9.029e+01   5.415 2.38e-07 ***
## Population     -5.964e-05  2.714e-05  -2.197  0.02953 *
## StateNew Jersey  3.735e+02  1.277e+02   2.925  0.00398 **
## StateNew York   1.338e+03  4.140e+02   3.232  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 190.4 on 150 degrees of freedom
## Multiple R-squared:  0.4902, Adjusted R-squared:  0.48
## F-statistic: 48.08 on 3 and 150 DF,  p-value: < 2.2e-16
```

Bar graphs showing the mean crime rates for each state of the tristate area

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
res1 <- aggregate(PropertyCrimeRate ~ State, data = TriStateCrimeData, FUN = mean)
ggplot(res1,aes(x=State,y=PropertyCrimeRate))+
  geom_bar(aes(fill = State), stat = "identity", width = 0.40) +
  xlab("State") + ylab("Mean Property Crime Rate")+
  theme(legend.position = "none")
```



```
res2 <- aggregate(ViolentCrimeRate ~ State, data = TriStateCrimeData, FUN = mean)
ggplot(res2, aes(x=State, y=ViolentCrimeRate)) +
  geom_bar(aes(fill = State), stat = "identity", width = 0.40) +
  xlab("State") + ylab("Mean Violent Crime Rate") +
  theme(legend.position = "none")
```

