**PROJECT REPORT**

**Introduction and Aim**

The project aims at analyzing the crime trends and establishing a multifaceted inference based on change in crime rates all over the United States for over more than 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.

**Importance of this Analysis**

This type of statistical analysis is important in terms of understanding the trend of crime rates in the 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. The data can also effectively address if the percentage increase in crime rate has been uniform and if not, then by how such increment or decrement varied over the period of five decades.

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.

**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 seeking 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>

The ***population for our study*** is the duration from 1960-2012 for which dataset is provided and the type of study is ‘***Observational***’ as we are observing the data set and drawing inferences based on our statistical analysis. The data is available only for those 50 years and hence, we have taken them. There has been no random sampling of the population. The future scope of this project can be to understand the trend all over United States.

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

* Is there a correlation between Violent Crime rate and Population of each state in the Tri-state area?
* Is there a correlation between the Property Crime rate and Population of each state in the Tri-state area?

The ***sample*** of the dataset is “*years under consideration from 1960-2012”*. The ***level of analysis*** is “*Year”*. In this dataset we come across a detailed description of population of each state of US from (1960-2012) and the number of crimes committed. The crime report includes data of 18 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).

In all our research questions, the independent variables will be either the year or the population of the state while the dependent variable would be the number of crimes taking place. The scales of measurement for the variables are as follows:

|  |  |
| --- | --- |
| **VARIABLE NAME** | **SCALE OF MEASUREMENT** |
| Population | Ratio |
| Violent Crime Rate | Ratio |
| Property Crime Rate | Ratio |

**Expected Outcome**

By the end of the statistical analysis we are expecting to address some of the following things like:

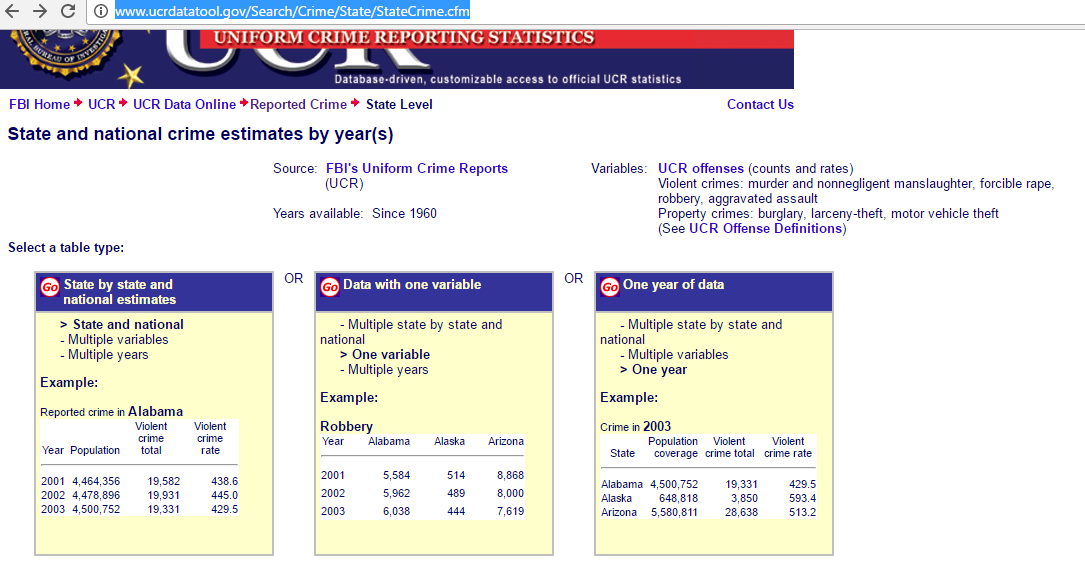
* Relationship between the different crime rates and the population of the states.
* Comparison between the Tri-states on the basis of their crime rates as well as population.

Statistical tests such as correlation and multiple regression would be used to get these results.

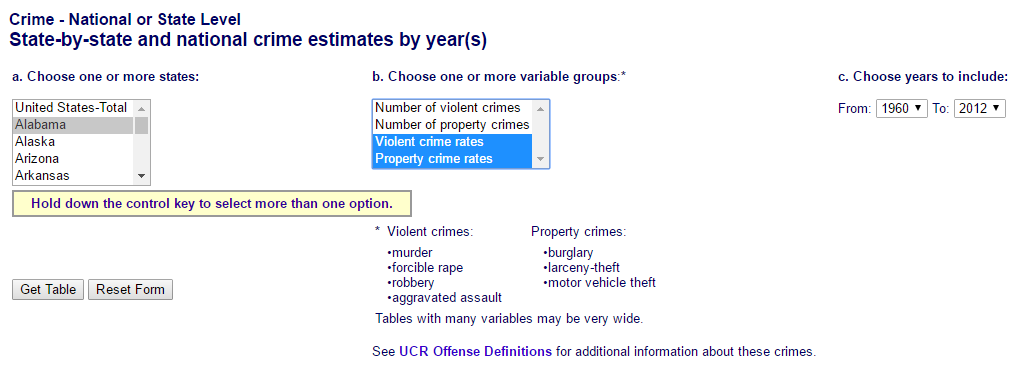
**Steps for Data Cleaning**

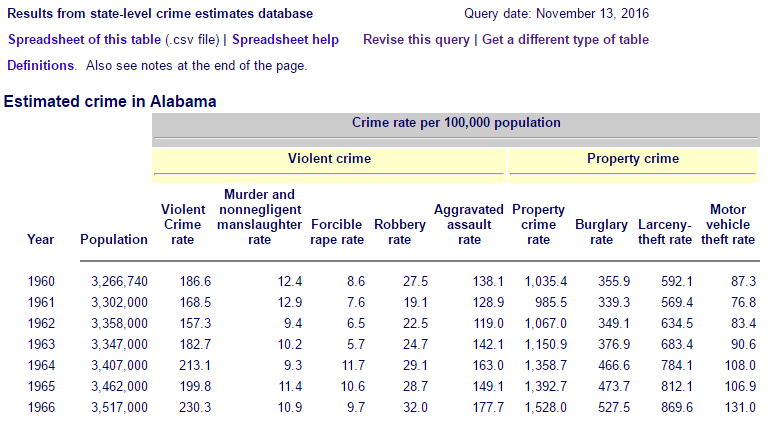
The dataset provided by us is a master dataset with details of crime rates in all the states. For our current research question “Is there a correlation between Violent Crime rate and Population of a particular state in the Tri-state area?” and “Is there a correlation between Property Crime rate and Population of a particular state in the Tri-state area?” we need just a portion of the dataset. We have procured a list of the 3 states (New York, New Jersey and Connecticut) as 3 separate csv files. The steps for obtaining these three datasets are as follows:

1. Go to link: <http://www.ucrdatatool.gov/Search/Crime/State/StateCrime.cfm>. Select the first option ‘State by state and national estimates’ as marked in the image below.



1. You will get a list of options to choose from. Select one particular state in the first option (i.e., New York, New Jersey or Connecticut) , Violent Crime Rates and Property Crime Rate in the second column and the years 1960 to 2012 in the final column. After selecting these options click on get table.



1. Click on ‘Spreadsheet of this table’, as shown by the arrow below, and it will download the csv file.  
   
2. Do this for all the states and you will get 3 different csv files for each state.
3. Each file had certain text written in the header and footer which has to be removed manually so that we can import the dataset in R easily.

The data set obtained does **not** contain any missing values.

**Exploratory Data Analysis**

We have three different datasets, one of each state in the Tri-State area. So the exploratory analysis is done separately for each state.

However, following are the independent and dependent variables in all three datasets.

|  |  |
| --- | --- |
| **Independent Variable** | |
| Population | Ratio |
| **Dependent Variables** | |
| Violent Crime Rate | Ratio |
| Property Crime Rate | Ratio |

We have one independent variable and two dependent variables in the data. The data is the violent crime rate and the property crime rate, in a state, for every year since 1960 to 2012.

**Analysis of data for New York**

**Descriptive statistics of New York**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| Population | 18371025 | 623959.4 | - | - |
| Violent Crime Rate | 731.075 | 263.8111 | 325.4 | 1180.9 |
| Property Crime Rate | 3784.531 | 1277.492 | 1906.7 | 5882 |

**Tests for Normality**

To check if the dependent variables for New York are normal or not, we do **Q-Q plot** of these dependent variables (***Property crime rate and violent crime rate***)

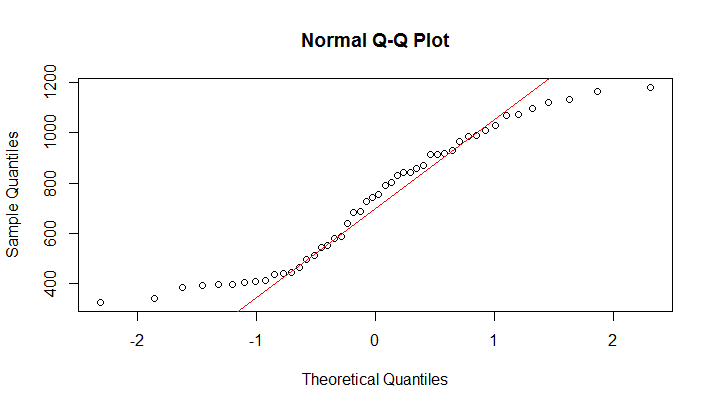


Figure 1: QQ plot for violent crime rate in New York

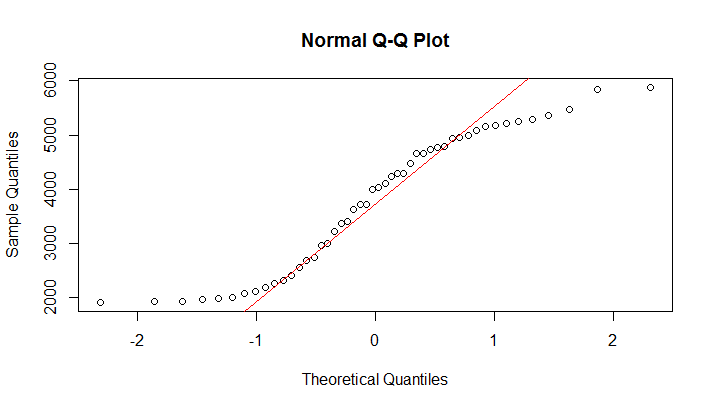


Figure 2: QQ plot for property crime in New York

We notice from the Q-Q Plot that most of the values fall on the line. We can further check it’s normality by the histogram of variables.

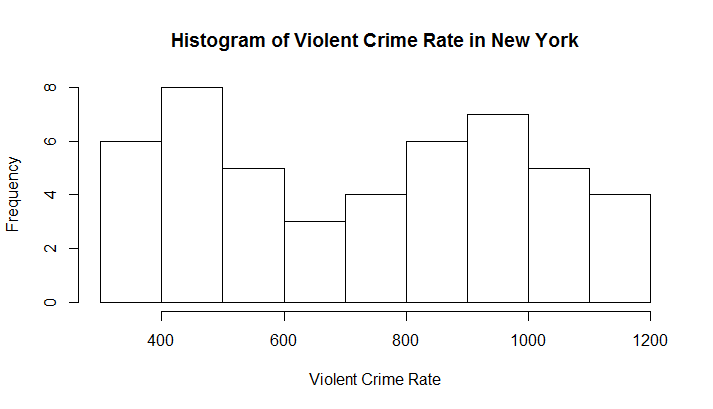


Figure 3: Histogram for Violent Crime Rates in New York

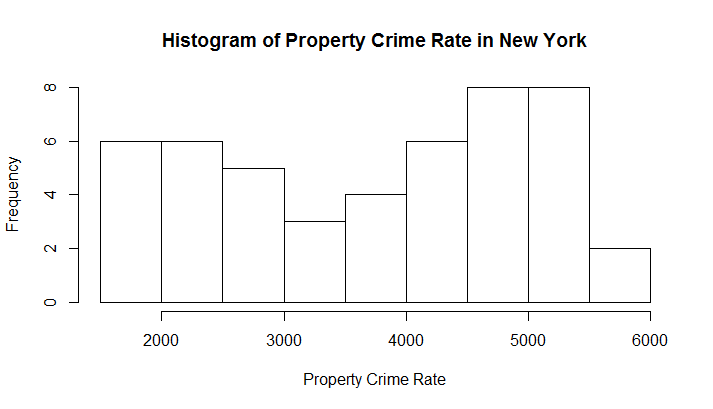


Figure 4: Histogram for property crime rate in New York

The histograms show some deviation from normality but that is because the sample size is very small. So, we cannot conclude that the distributions are not normal just by seeing the histograms.

>cor.test(Population,Violent.Crime.rate)

Pearson’s product-moment correlation

data: Population and Violent.Crime.rate

t = -8.1786, df = 46, p-value = 1.615e-10

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.8647562 -0.6215673

sample estimates:

cor

-0.7697562

>cor.test(Population,Property.crime.rate)

Pearson’s product-moment correlation

data: Population and Property.crime.rate

t = -13.956, df = 46, p-value = 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.9426504 -0.8264948

sample estimates:

cor

-0.8994147

**Graph of Outliers**

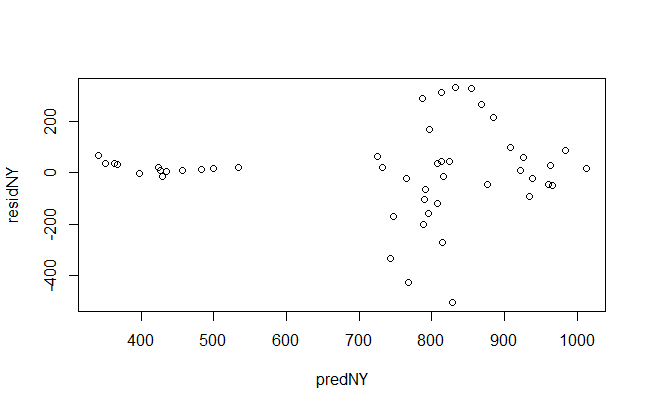
****

Figure 5: Outlier plots for Violent Crimes New York

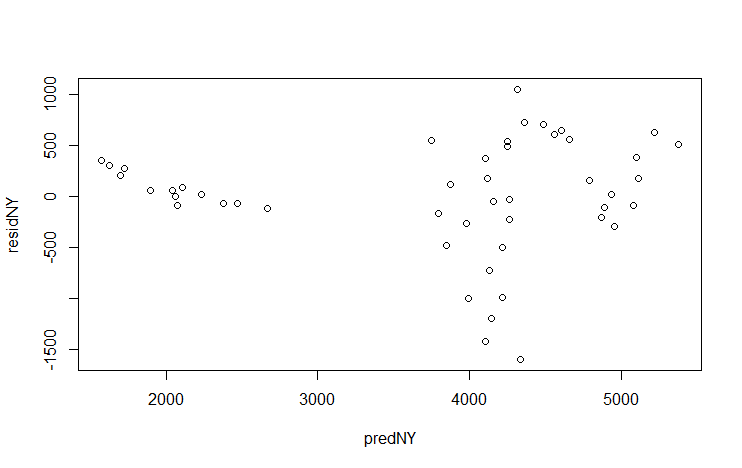


Figure 6: Outliers plot for Property Crimes New York

**Analysis of data for New Jersey**

**Descriptive statistics for New Jersey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| Population | 7686547 | 725889.7 | 6066782 | 8864590 |
| Violent Crime Rate | 404.0075 | 159.5698 | 107 | 647.6 |
| Property Crime Rate | 3452.649 | 1230.404 | 1376.5 | 5797 |

**Test for Normality**

To check if the dependent variables for New Jersey are normal or not, we do **Q-Q plot** of these dependent variables (***Property crime rate and Violent crime rate***)

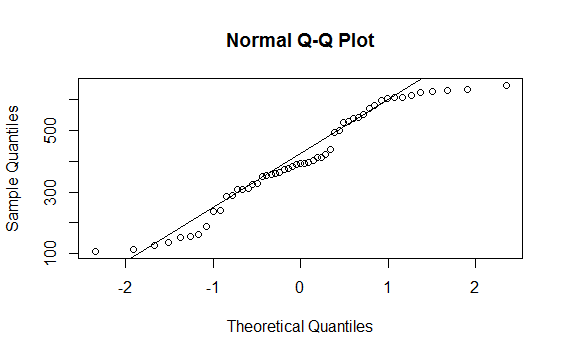


Figure 7: QQ plot for violent crime rate in New Jersey

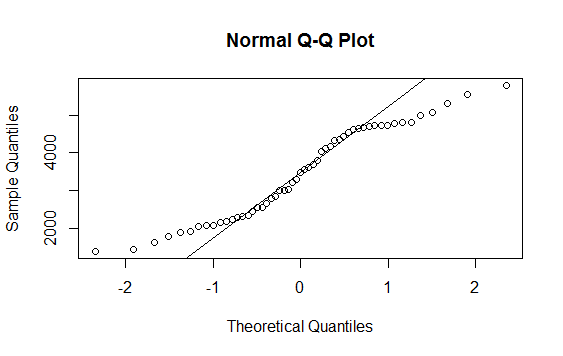


Figure 8: QQ plot for property crime rate in New Jersey

We check the normality of the distribution by plotting a **histogram**. Following are the histograms for the Violent Crime rate and Property Crime rate in New Jersey

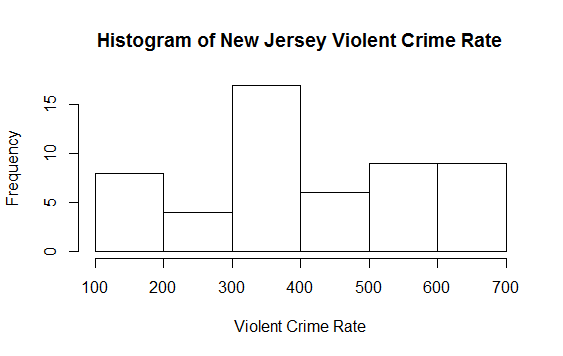


Figure 9: Histogram for violent crime rate in New Jersey

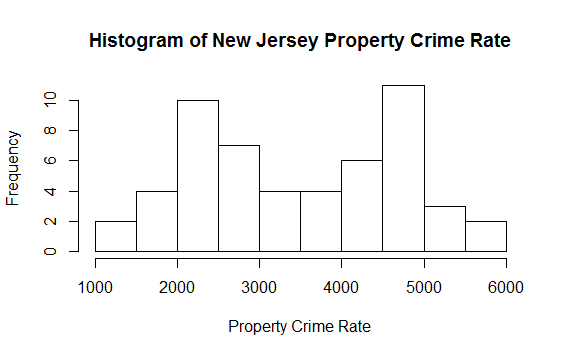


Figure 10: Histogram for property crime rate in New Jersey

These histograms don’t follow the shape of normal distribution; rather they are slightly skewed and bimodal in nature. As seen in the q plots as well although majority of data is compliant to the usual QQ plot there are some outliers which are causing the discrepancies.

In order to deduce whether Crime rates population dependent or not; we perform correlation test in R using following commands in R.

>cor.test(newjersey$Population,newjersey$Violent.Crime.rate)

product-moment correlation

data: newjersey$Population and newjersey$Violent.Crime.rate

t = 2.3262, df = 51, p-value = 0.02402

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.04302151 0.53520374

sample estimates:

cor

0.3097138

>cor.test(newjersey$Population,newjersey$Property.crime.rate)

product-moment correlation

data: newjersey$Population and newjersey$Property.crime.rate

t = -0.48337, df = 51, p-value = 0.6309

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.3317689 0.2065327

sample estimates:

cor

-0.06753096

**Outlier Graphs for Dependent variables of Connecticut**

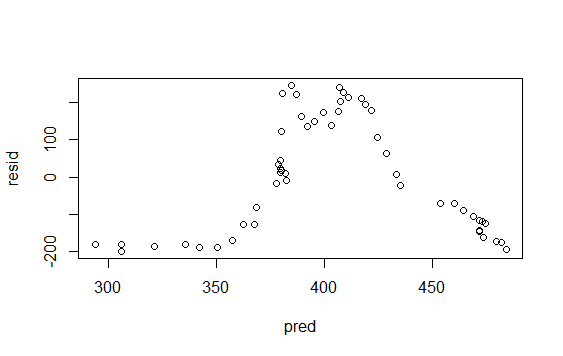


Figure 11: Outlier graph for violent crime in New Jersey

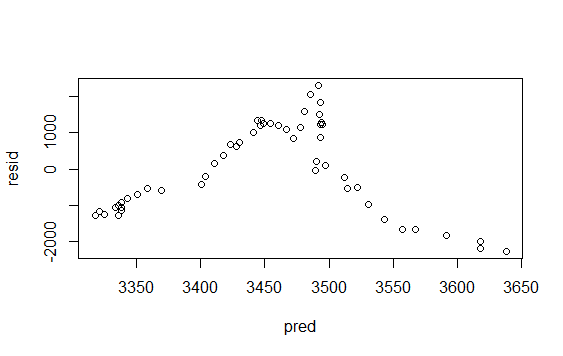


Figure 12: Outlier graph for property crime rate in New Jersey

**Analysis of Data for Connecticut**

**Descriptive statistics for Connecticut**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| Population | 3183763 | 258949.2 | 2535234 | 3590347 |
| Violent Crime Rate | 299.0321 | 139.7463 | 33.6 | 1119.9 |
| Property Crime Rate | 3410.268 | 1210.405 | 553.7 | 5469.2 |

To check if the dependent variables for Connecticut state are normal or not, we do **Q-Q plot** of these dependent variables (***Property crime rate and Violent crime rate***)

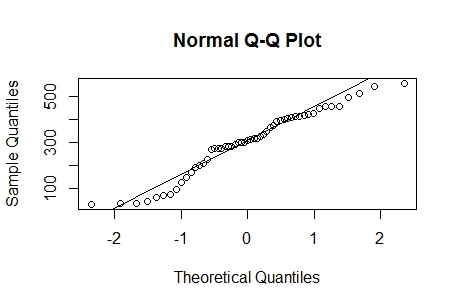


Figure 13: QQ plot for Property Crime Rate

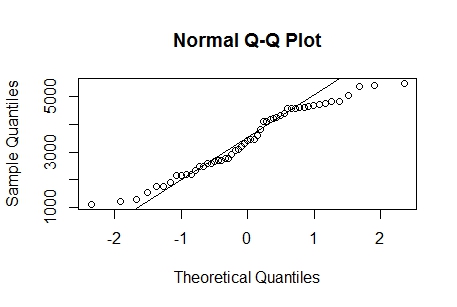


Figure 14: Q-Q plot for violent crime rate

Following are the histograms for the Violent Crime rate and Property Crime rate in Connecticut.

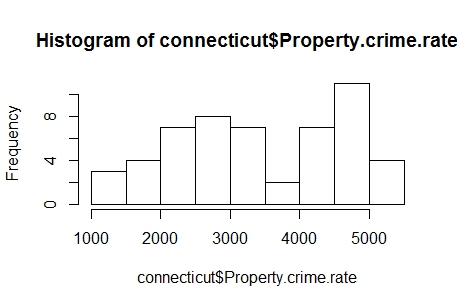


Figure 15: Histogram for property crime in Connecticut

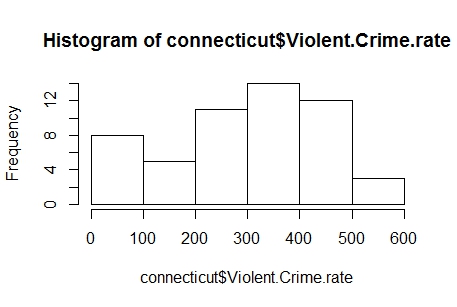


Figure 16: Histogram of violent crime rate in Connecticut

These histograms don’t follow the shape of normal distribution; rather they are slightly skewed and bimodal in nature. As seen in the q plots as well although majority of data is compliant to the usual QQ plot there are some outliers which are causing the discrepancies.

In order to deduce whether Crime rates population dependent or not; we perform correlation test in R using following commands in R

cor.test(connecticut$Population,connecticut$Violent.Crime.rate)

Pearson's product-moment correlation

data: connecticut$Population and connecticut$Violent.Crime.rate

t = 5.5626, df = 51, p-value = 9.851e-07

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.4127644 0.7587661

sample estimates:

cor

0.6144999

cor.test(connecticut$Population,connecticut$Property.crime.rate)

Pearson's product-moment correlation

data: connecticut$Population and connecticut$Property.crime.rate

t = 1.3962, df = 51, p-value = 0.1687

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.08270255 0.43938600

sample estimates:

cor

0.1918806

**Outlier Graphs for Dependent variables of Connecticut**

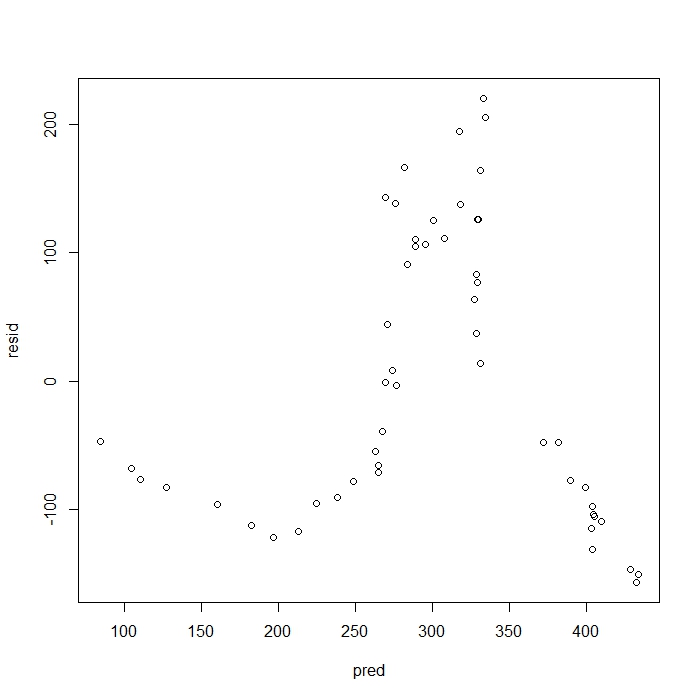


Figure 17Outliers for dependent variable violent crime

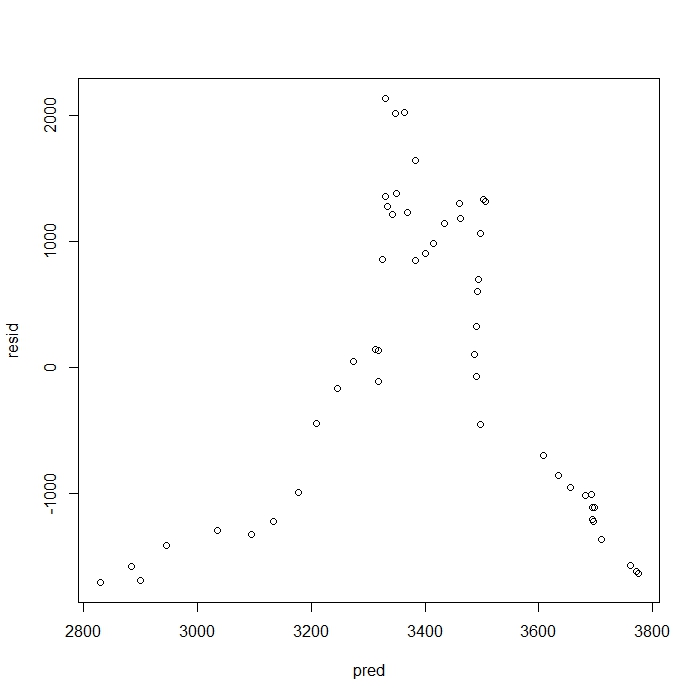


Figure 18: Outliers for dependent variables property crime

**Trend Analysis**

Following is the comparison of the *violent crime rate* in New York since 1960 to 2011.

**SUMMARY**

**New-York**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of Crime** | **df** | **t statistic value** | **p-value** | **correlation** |
| Violent crime | 46 | -8.1786 | 1.615e-10 | -0.769 |
| Property crime | 46 | -13.956 | 2.2e-16 | -0.899 |

Looking at the correlation value we deduce that violent crimes are negatively related to population (Correlation Coefficient **=** -0.769); Property crimes are also negatively related to population(Correlation Coefficient= -0.899)

***Violent Crime rate****:*

***Null Hypothesis: H0*** = Violent Crime rates of New York is dependent on Population of New-York

**Alternate Hypothesis: Ha** = Violent Crime rates of New York is not related to Population of New-York.

***Property Crime rate****:*

***Null Hypothesis: H0*** = Property Crime rates of New York is dependent on population of New-York

**Alternate Hypothesis: Ha** = Property Crime rates of New York is not related to population of New-York

**New Jersey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of Crime** | **df** | **t statistic value** | **p-value** | **corelation** |
| Violent crime | 51 | 2.3262 | 0.024 | 0.3097 |
| Property crime | 51 | -0.48337 | 0.6309 | -0.0675 |

Looking at the correlation value we deduce that violent crime are not related to population **(*correlation coefficient 0.3097)****;* and property crimes are also not related to population ***(correlation coefficient -0.067)***

***Violent Crime rate****:*

***Null Hypothesis: H0*** = Violent Crime rates of New Jersey is dependent on Population of New-Jersey.

**Alternate Hypothesis: Ha** = Violent Crime rates of New Jersey is not related to Population of New-Jersey.

***Property Crime rate****:*

***Null Hypothesis: H0*** = Property Crime rates of New Jersey is dependent on population of New Jersey.

**Alternate Hypothesis: Ha** = Property Crime rates of New Jersey is not related to population of New Jersey.

**Connecticut**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type of Crime** | **df** | **t-statistic value** | **p-value** | **corelation** |
| Violent crime | 51 | 5.5626 | 9.851e-07 | 0.614499 |
| Property crime | 51 | 1.3926 | 0.1687 | 0.1918 |

Looking at the correlation value we deduce that violent crimes are related to population **(*correlation coefficient 0.6)****;* whereas property crimes are not population dependent ***(correlation coefficient 0.19)***

***Null Hypothesis: H0*** = Violent Crime rates of Connecticut is dependent on Population of Connecticut

**Alternate Hypothesis: Ha** = Violent Crime rates of Connecticut is not related to Population of Connecticut

***Property Crime rate****:*

***Null Hypothesis: H0*** = Property Crime rates of Connecticut is dependent on population of Connecticut.

**Alternate Hypothesis: Ha** = Property Crime rates of Connecticut is not related to population of Connecticut.