Group 127 : Analyzing the Crime in Chicago

First Name	Last Name	Monday or Tuesday class
Zeshan	Sayed	Tuesday(online)
Vikhyat	Tandon	Tuesday(online)

Table of Contents

1. Introduction

Crime in Chicago is a topic of conversation from many years etc, as crime in Chi-city is dissipated throughout the city with respect to location, time, etc. It becomes very difficult for lawful forces to control the crime rate in Chicago city. Through this aspect of reducing the crime rate, we as a team developed a Public Safety Application in DemonHacks(Depaul University)m which got a great recognition on several platforms which motivated my team to do depth analysis on the problems that people of Chicago city are facing. This idea of detailed analysis on the crime rate in Chicago will help to predict the crime rate each day, which will help the lawful public safety to deploy their forces effectively. This idea of predicting the crime rate will help the main project to determine the IoT servers locations depending upon the density of the crime rate in the city.

2. Data

The data is from **Kaggle**, here is the link:-
https://www.kaggle.com/currie32/crimes-in-chicago

The data-set contain data from 2012 to 2017 and it contains 14 lakhs rows

The dataset is of the crime domain. It contains close to a million records and has 24 features.
The features are as follows:

1. **ID:** identifies unique record.
2. **Case Number:** indicates crime report which is unique to the incident.
3. **Date:** date when the crime occurred.
4. **Block:** block address where the crime occurred.
5. **IUCR:** It indicates Illinois uniform crime reporting code. It is linked to primary type and description features.
6. **Primary Type:** It describes the IUCR codes.
7. **Description:** A subCategory description of primary type feature.
8. **Location Description:** It describes the location where the crime occurred.
9. **Arrest:** indicates whether an arrest was made.
10. **Domestic:** indicates whether the crime occurred was domestic or not.
11. **Beat:** It indicates which beat marshal was present at that incident.
12. **District:** It indicates the police district where the incident occurred.
13. **Ward:** It indicates the ward number of the city where the incident occurred.
14. **Community Area:** It indicates the community area code where the incident occurred.
15. **FBI Code:** It indicates the crime classification per the FBI.
16. **X coordinate:** Describes the location where the actual incident occurred.
17. **Y coordinate:** Describes the location where the actual incident occurred.
18. **Year:** Describes the year when the crime has occurred.

19. **Updated on:** Indicates when the record has been updated.
20. **Latitude:** Indicates the latitude of the location where the incident has occurred.
21. **Longitude:** Indicates the longitude of the location where the incident has occurred.
22. **Location:** Describes the location where the incident has occurred.

3. Problems to be Solved

There is lot of crime in Chicago and near by region due to which citizens are devastated. The crime is spread over different regions at different period of time. The time of the crime event contributes towards a crime scene. Using the predictive policing technique i.e., a multidimensional optimization problem we try to utilize the data available to us in finding the instances of crime over time and across geographies.

4. Data Processing

Before starting to clean the data, it was important to understand how the data was organize. For this we asked ourselves few questions which are as follows:

1. What are the fields?
2. What are the datatypes?
3. Are they sorted or not?
4. How are they stored?
5. Are there any Null values?

For getting answers of above result we used **str()** and **summary()** commands, below are the outputs of the commands on dataset:-

1) str()

```
> str(crime.data1)
'data.frame':   1456714 obs. of  23 variables:
 $ X                   : int  3 89 197 673 911 1108 1130 1801 1868 1
 $ ID                  : int  10508693 10508695 10508697 10508698 10
 $ Case.Number         : Factor w/ 1456599 levels "","161884","2234
 $ Date                : Factor w/ 582146 levels "01/01/2012 01:00:0
 $ Block               : Factor w/ 32774 levels "0000X  I94/EXIT 12
 $ IUCR                : Factor w/ 365 levels "0110","0141",..: 52 5
 $ Primary.Type        : Factor w/ 33 levels "ARSON","ASSAULT",..: 1
 $ Description         : Factor w/ 342 levels "$500 AND UNDER",..: 1
 $ Location.Description: Factor w/ 143 levels "","ABANDONED BUILDIN
 $ Arrest              : Factor w/ 2 levels "False","True": 2 1 1 1
 $ Domestic            : Factor w/ 2 levels "False","True": 2 2 1 1
 $ Beat                : int  1022 313 1524 1532 1523 631 133 215 2
 $ District            : num  10 3 15 15 15 6 1 2 24 7 ...
 $ Ward                : num  24 20 37 28 28 8 3 3 40 17 ...
 $ Community.Area      : num  29 42 25 25 25 44 35 38 1 67 ...
 $ FBI.Code            : Factor w/ 26 levels "01A","01B","02",..: 1
 $ X.Coordinate        : num  1154907 1183066 1140789 1143223 11398
 $ Y.Coordinate        : num  1893681 1864330 1904819 1901475 19016
 $ Year                : int  2016 2016 2016 2016 2016 2016 2016 20
 $ Updated.On          : Factor w/ 959 levels "01/01/2016 03:52:56
 $ Latitude            : num  41.9 41.8 41.9 41.9 41.9 ...
```

Fig 1: Output of str() function

2) **summary()**

```
> summary(crime.data1)
       X                 ID              Case.Number                          Date
 Min.   :      3   Min.   :   20224   HZ140230:      6   01/01/2012 12:01:00 AM:    166   001XX N STATE ST
 1st Qu.:2698636   1st Qu.: 9002709   HY346207:      4   01/01/2013 12:01:00 AM:    122   0000X W TERMINAL S
 Median :3063654   Median : 9605776   HZ403466:      4   01/01/2012 12:00:00 AM:    115   008XX N MICHIGAN A
 Mean   :3308606   Mean   : 9597550   HZ554936:      4   01/01/2015 12:01:00 AM:    110   076XX S CICERO AVE
 3rd Qu.:3428849   3rd Qu.:10225766   HV217424:      3   01/01/2014 12:01:00 AM:    104   0000X N STATE ST
 Max.   :6253474   Max.   :10827880   HW486725:      3   01/01/2016 12:01:00 AM:    104   064XX S DR MARTIN
                                      (Other) :1456690   (Other)                 :1455993   (Other)
                Description                          Location.Description   Arrest          Domestic
 SIMPLE                  :150600   STREET                        :330471    False:1079242   False:1236660   M
 $500 AND UNDER          :136036   RESIDENCE                     :233530    True : 377472   True : 220054   I
 DOMESTIC BATTERY SIMPLE:130700   APARTMENT                     :185023                                    M
 TO VEHICLE              : 75801   SIDEWALK                      :160891                                    M
 OVER $500               : 74906   OTHER                         : 55774                                    3
 TO PROPERTY             : 71694   PARKING LOT/GARAGE(NON.RESID.): 41768                                    M
 (Other)                 :816977   (Other)                       :449257
```

Fig 2: Output of summary(crimedata)

From the screenshots one can infer the answers of the questions above. Also, we can figure out that every incident has a unique Case.Number, which is used to plot unique values.
Once we get the overview of the data, we start with the process of data cleaning.

Data Cleaning:

Chicago Crime data consist of qualitative data. This data contained a number of duplicate values, null values and improper imputed values.
Thus, data cleaning was the most important part for predicting the values with less error.

1. ***Removing Duplicate Values*:**

   We can note that there are some columns where values are duplicated, i.e., there are two or more rows have same value.For example, there are three rows in the data that have a case value equal to HT572234. These duplicated rows are needed to be removed.
   The duplicate values were removed with the help of **duplicated()** function.

2. ***Removing Null Values:***

   Similarly as in duplicate values, we check for null values with **is.na()** function. Removing these rows stands as an important factor because null values make the prediction difficult due to it outlying features. These values are important for model creation but they are needed to be filled logically. For example, we cannot take the mean of longitude and latitude from the columns and filled it for the missing values. These values can be either filled by mining the data or can be removed. Since this project is more of an

analytics we concentrated more on analysis rather than mining and then doing analysis.We can also use **which()** function to find the missing value in a particular row.

3. *Removing Improper Imputed Values:*

Improper imputed values means illogical values for example “CASE” and “CASE#” value which are improper due to imports and export of data in to different platforms.

4. *Extracting Data from date column.*

The “**date**” column gives an approximate date and time stamp about the crime incident. We check the date column as how it is stored with the help of head() function which shows few rows of the head of the data.

Now we know that date is stored as a factor variable. To extract the date from the date column we convert the date to date object with the help of **POSIXlt() function**. Also, we tested the conversion on a variable first and then used it in actual column, this removes the risk of changing the actual date column to NA due to any mistake in the **POSIXit() function**. Then we processed the date to acquire date from the value.

```
> summary(crime.data$Date)
01/01/2012 12:01:00 AM 01/01/2015 12:01:00 AM 01/01/2015 12:00:00 AM 01/01/2012 12:00:00 AM (
                   126                     97                     88                     78
01/01/2014 12:00:00 AM 06/01/2012 09:00:00 AM 03/01/2012 09:00:00 AM 10/01/2012 09:00:00 AM (
                    67                     66                     61                     59
08/01/2012 09:00:00 AM 01/01/2015 09:00:00 AM 05/01/2012 09:00:00 AM 06/01/2014 09:00:00 AM (
                    55                     54                     53                     52
```

Fig 3: Original timestamp values

```
> summary(crime.data$date)
2012-01-01 2012-06-01 2012-07-01 2013-01-01 2013-05-01 2012-09-01
      1332       1185       1166       1159       1135       1132
2012-06-24 2012-06-13 2012-08-25 2012-08-01 2012-03-21 2012-05-23
      1077       1075       1075       1073       1072       1070
```

Fig 4: Extracted date values

5. *Extracting Time from date column.*

After performing the above step, we already have the whole date and time stamp in date object. Here we just need to extract the timestamp convert it to 24hrs format and then store it in another variable. Also we created a timetag to sort the date between 4 time stamps and then visualize it accordingly.

Extracted time and saving it to new variable.

```
> summary(crime.data$time)
09:00:00 10:00:00 12:00:00 08:00:00 07:00:00 11:00:00 03:00:00 02:00:00 06:00:00 01:00:00
   58208    49545    47836    46693    37528    37252    37014    36914    36562    35798
06:30:00 12:01:00 05:30:00 04:30:00 11:45:00 07:45:00 12:15:00 10:45:00 09:45:00 08:45:00
   16911    16471    15695    15640     7240     7103     7088     6880     6813     6765
```

Fig 5: Extracted time values

Overall heads of time, date and Date:

```
> head(crime.data$Arrest)
[1] 1 0 0 0 0 0
> head(crime.data$time)
[1] 11:40:00 09:40:00 11:31:00 10:10:00 10:00:00 10:35:00
720 Levels: 01:00:00 01:01:00 01:02:00 01:03:00 01:04:00 01:05:00 01:06:00 01:07:
> head(crime.data$date)
[1] 2016-05-03 2016-05-03 2016-05-03 2016-05-03 2016-05-03 2016-05-03
1840 Levels: 2012-01-01 2012-01-02 2012-01-03 2012-01-04 2012-01-05 2012-01-06 20
> head(crime.data$Date)
[1] 05/03/2016 11:40:00 PM 05/03/2016 09:40:00 PM 05/03/2016 11:31:00 PM 05/03/20
571328 Levels: 01/01/2012 01:00:00 AM 01/01/2012 01:00:00 PM 01/01/2012 01:04:00
>
```

Fig 6: Head values of time, date and original timestamp

6. ***Categorizing similar but despaired values.***

   There were two fields in the data which had the same description of the crime incident. For example in Primary.Type column we can see that there are two different types of same crime for example, “Non-Criminal” and “NonCriminal”. We can infer from the Primary.Type column that there are 31 crime types but 4 out of them are same, thus to remove the redundancy and to make data unique and accurate we combine this these crime types and save it into one type with the help of if.else() function.

Data Processing:

Data processing contains adding new columns to the data for the analysis purpose because one cannot directly infer **day, month, and year** from the date object. Qualitative data needs a lot of preprocessing and thus we did a detailed analysis on every column and counted the class relative frequency and determined is this column valuable to be converted. Resulting we converted few to binary like **Arrest** which consisted TRUE or FALSE to binary as 1 or 0.

```
> head(crime.data$Arrest)
[1] 1 0 0 0 0 0
> summary(crime.data$Arrest)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2614  1.0000  1.0000
> head(crime.data$Arrest)
[1] 1 0 0 0 0 0
>
```

Fig 7: Head of Arrest value

Furthermore, we created week column as well to analyse the trend accordingly with weekday() and month() functions.

To be removed
It was being difficult to handle data with having 30 variables, so we created a new table with only values that are affecting the prediction. We arranged the values with unique() functions and added the value of the x variables to the new data, which we named as crime.model.

5. Methods and Process

Once we cleaned the data, we get the number of rows in the finalized cleansed data. The number of rows counted is approximately **14 lakhs rows** due to which we decided to use **Hold-out** approach in order to cut the data into **train** and **test**.

As there is a presence of a **times series** in our Chicago crime data, we separated the time part from leaving behind the date part in the time i.e., month, day and year Date function. Obtaining the time series, we were able to plot the count of crime that is happening in Chicago on a particular day for the whole data series we have taken. For plotting the time series we used the **hchart** function in the the **highcharter** library.

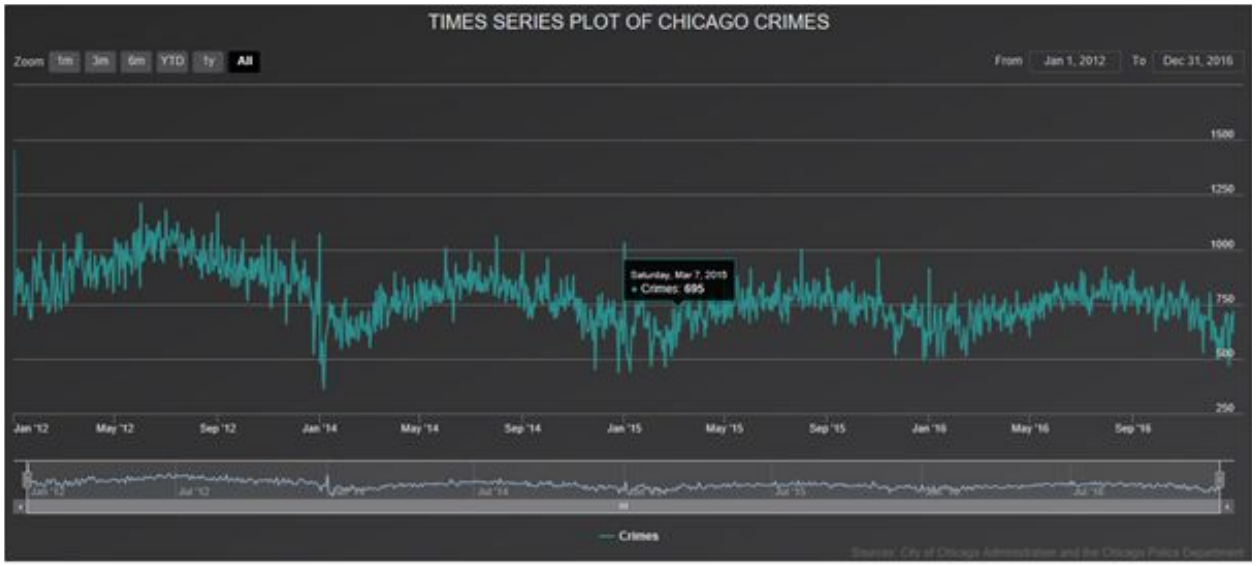

Fig 8: Times Series Plot of Chicago Crimes

From fig. 8 we can clearly observe that the Chicago Crime data is stationary i.e., over a period of time, the values of mean and variance remains constant. To prove that the data is stationary we take the help of **QQplot** and **Histogram.**

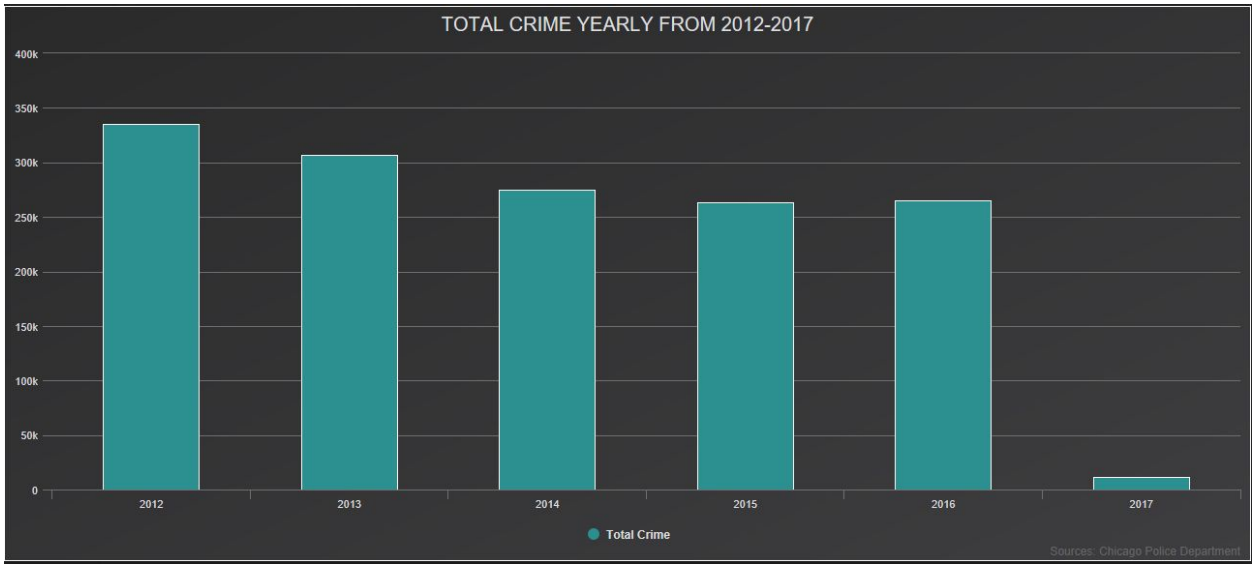

Fig 9: Chicago Crime Rate from 2012 - 2017

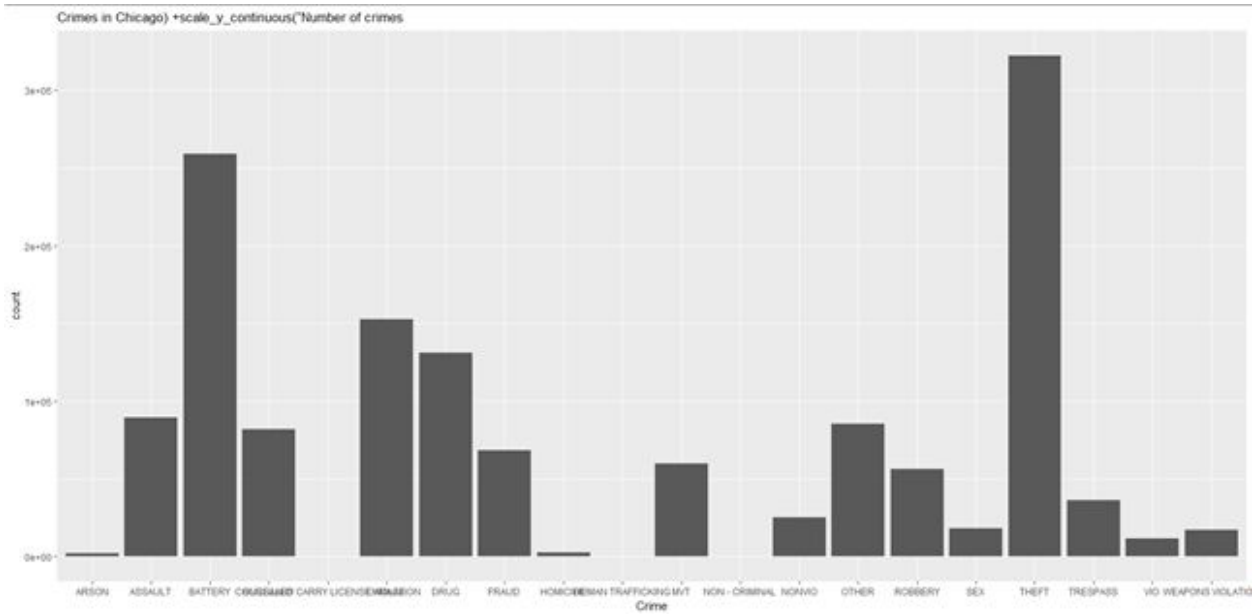

Fig 10: Count of various types of Crime in Chicago

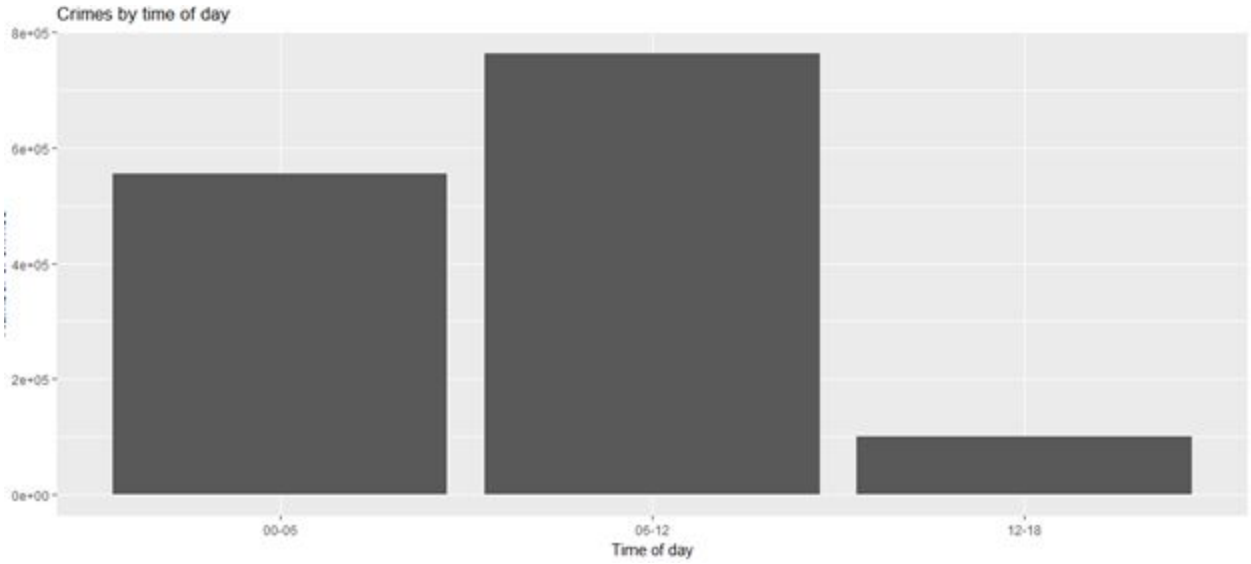

Fig 11: Crimes in Chicago by the time of day

In fig. 11, we can see the plot of the crimes happening in chicago during the day. From the chart, we can observe that most the crime in chicago happens during the morning hours. The city of Chicago is not at all safe during the late night hours as well which can be figured out seeing the representation for crimes happened between 12am to 5am.

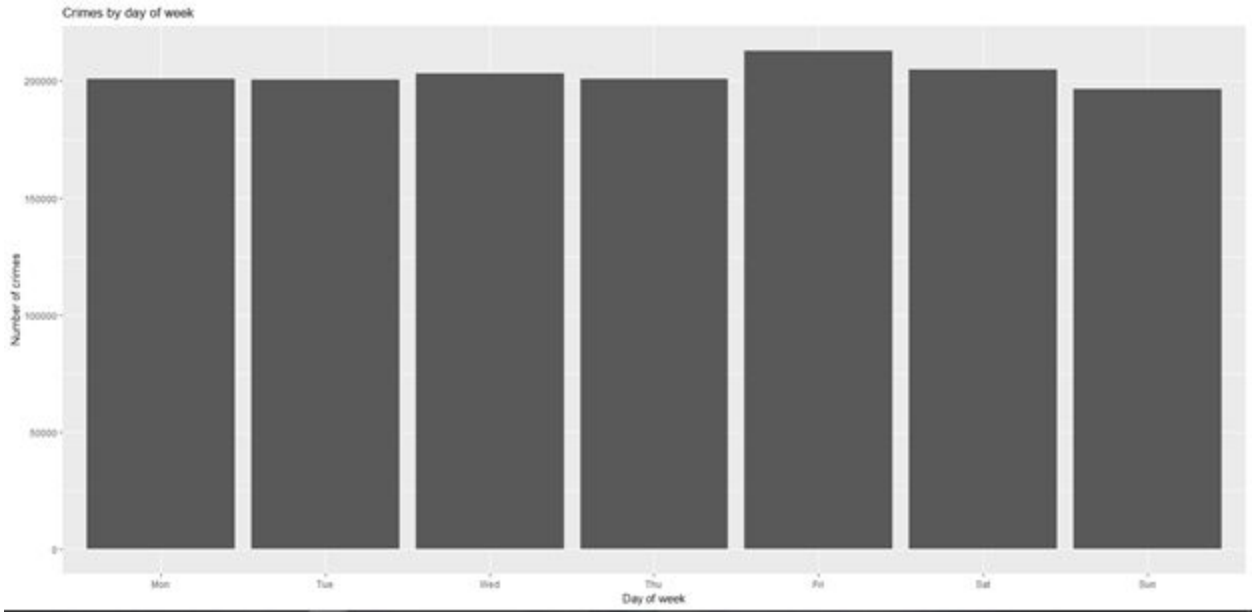

Fig 12: Crimes in Chicago by days of week

From fig. 12, we can clearly figure out that the maximum crime in chicago happens during the weekend days i.e., Friday, Saturday and Sunday with Friday being the day having the highest count. The trend remains similar on weekdays, still the count of crime happening on those days is pretty high.

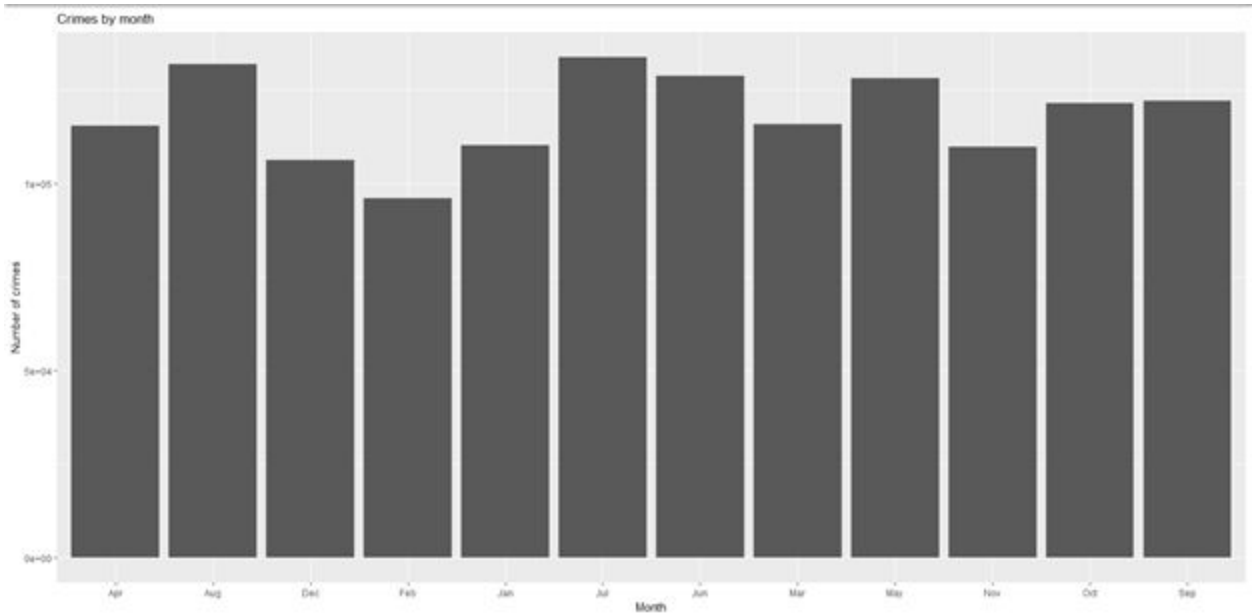

Fig 13: Crimes in Chicago monthwise

Fig. 13 explains the trend of Chicago crime on a monthly perspective. From the trend, we observe that as winters sets in and is in its peak i.e., in the months of November, December, January and February, the crime rate starts to decrease starting November and remains

comparatively low till February. The reason behind this might be the harsh winter conditions which preserve in Chicago. From the representation, we can also observe that once the weather starts improving, i.e., from the month of March, the crime rate in Chicago also starts increasing gradually.

With the value to be predicted we found the correlations and eliminated the columns which are not useful or do not have a satisfactory results on the output. As understanding the nature of the problem is critical for creating a model within a given set of constraints. Thus, we cannot predict exactly when and where the next crime is going to happen but we can derive actionable insights in the form of statistics.

Now, our data set has all the crime incidents that were recorded in the past given date. After combining the new data columns we get beats in which no crime occurred or may be not recorded. For these rows, we see NAs in the count and arrest fields. We can replace the NAs with zero to indicate that there were no crimes, and therefore no arrests, in these beats on these days.

On of the main assumptions we make when working with linear models is that linear regressions and analysis of variance is that the residual errors follow a normal distribution. Often we observe that the response variable of interest is categorical or discrete, and not continuous. There are two problems which apply when we are working with an ordinary linear regression model. The first problem is that, many distributions of count data are positively skewed as many values are 0. The skewed distribution cannot be normal because of the presence of so many 0's in the data set. The second issue is that which is theoretically impossible, but is quite likely that the regression model may end up producing negative predicted values. Now considering our data, we too had many inputs in the Chicago crime data which didn't had values i.e., they were blank or zero. Therefore, any attempts to transform the data to normal distribution failed. Considering these mentioned factors, we compared our Chicago crime data on two models i.e., the Poisson model and the Negative Binomial model.

The definition of the **Poisson model** assumes that the mean and variance of the errors to be equal which in most cases isn't true. There are cases where when the variance of errors is greater than the mean just like in our Chicago crime data example (though it can smaller too). When we are working with the over-dispersed Poisson model, we need to observe the estimates about how larger the variance is than the mean. This estimate helps us to correct the effects of the large variance on the p-values.

The **Negative Binomial Model** is an extension of the the Poisson distribution model where the distribution's parameter is itself considered to be random. The reason for this randomness could be the higher value of variance of the data than the mean of the data.

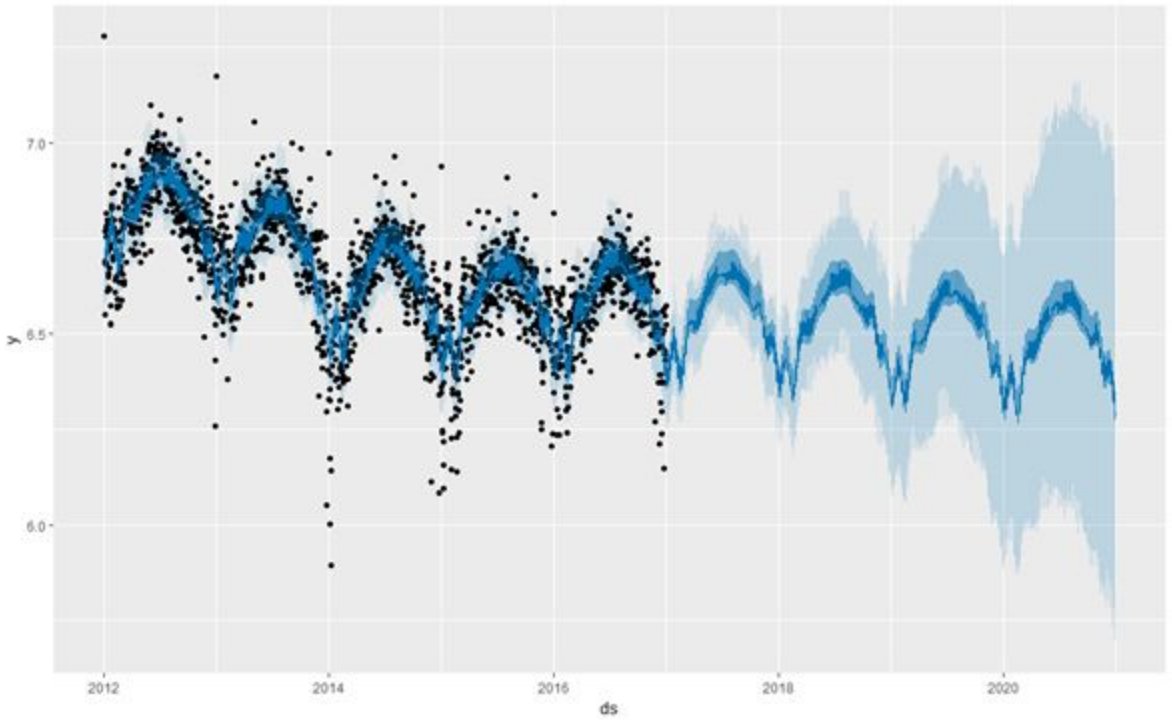

Fig 14: Prediction analysis of Crimes in Chicago

In fig. 14, we can see the Prediction of the number of crimes which is going to happen from 2017 to 2020 keeping the current crime count of crimes in Chicago i.e., from 2012 to 2016 in mind. For forecasting, we are using the facebook library **Prophet.** Prophet is a procedure for forecasting time series data. It is an additive model where non-linear trends are fit with yearly and weekly seasonality, and even considers holidays. The model works best when we are working with daily periodicity data which consists of one historical data as well. In our case,the historical data is the Chicago crime data we have from 2012 to 2016. The input to **prophet** is always a dataframe with two columns namely **ds** and **y.** The column **ds** (datestamp) column represents the date or the datetime values while **y** column is numeric and consists the measurement of the data we want to forecast.

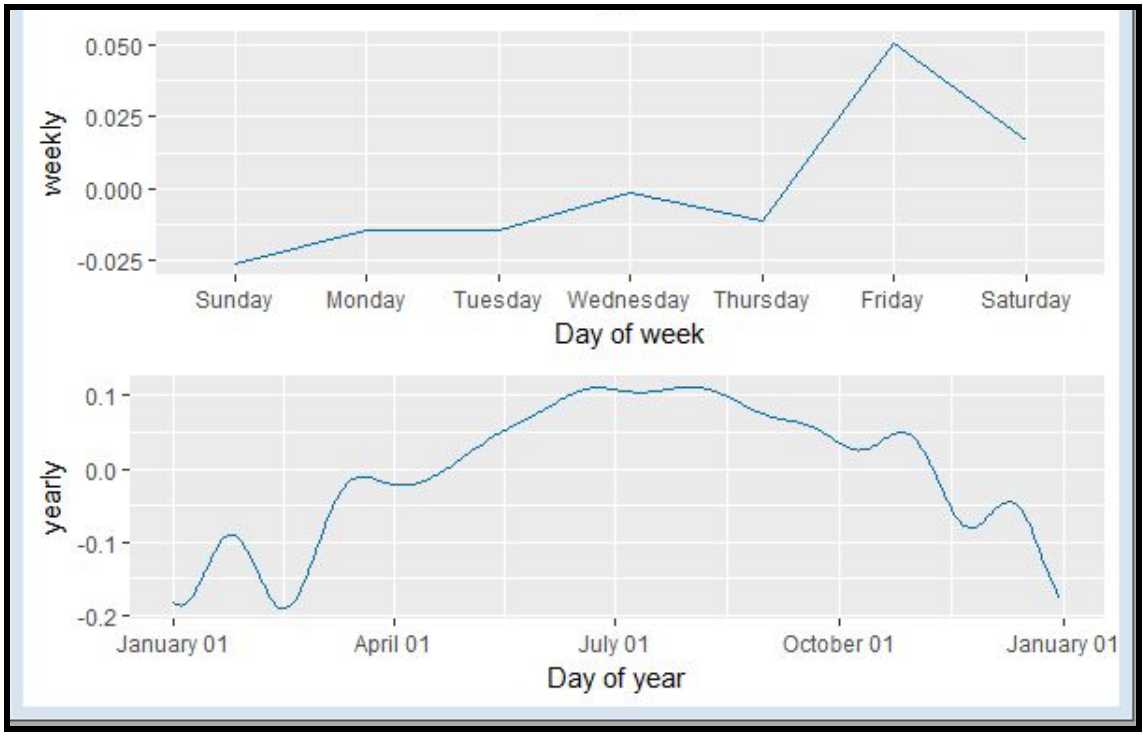

Fig 15: Monthly and Yearly analysis using Prophet

In our case, from fig 14 **ds** is the date value of the chicago crime while **y** is the count of crime committed in Chicago. When forecasting using the prophet library we use the log value of the data we want to forecast for i.e., **log(value of data for forecasting).** As seen in fig 15, we tried to check the results obtained in our early analysis in fig 12 and fig 13, to see if the results matched those obtained from prophet. As seen in fig 15, we can see in the weekly analysis that the amount of crimes occurring on the weekend is way higher than what happens during weekdays even though the weekdays number is pretty high. But on weekends the number substantially increases and in our earlier analysis in fig 12, we found the same points. Similarly, in fig 15 we can see the crime rate in Chicago goes down during the months of extreme winters i.e., from November to February and then starts to increase from the month March and onwards as the weather conditions get better. Our previous prediction from the previous results in fig 13 also states the same facts.

6. Evaluations and Results

6.1. Evaluation Methods

In this project we have used 2 evaluation methods for selecting the best model between negative binomial model and Poisson Model. The two evaluations methods we used are AIC value and RMSE matrix to determine between the two models.

```
Call:
glm.nb(formula = count ~ past.crime.1 + past.crime.7 + past.crime.30 +
    crime.trend + policing + factor(day) + season + I(past.crime.30^2),
    data = model.data, init.theta = 14.0536319, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8559  -0.7453  -0.2365   0.5002   8.7683

Coefficients:
                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)         -1.504e+00  8.013e-03 -187.698  < 2e-16 ***
past.crime.1         1.274e-02  5.382e-04   23.675  < 2e-16 ***
past.crime.7        -3.122e-02  3.412e-04  -91.509  < 2e-16 ***
past.crime.30        3.259e-02  1.206e-04  270.360  < 2e-16 ***
crime.trend          3.551e+00  2.569e-02  138.203  < 2e-16 ***
policing             1.829e-01  9.645e-03   18.966  < 2e-16 ***
factor(day)Mon      -6.052e-02  3.614e-03  -16.746  < 2e-16 ***
factor(day)Sat      -4.593e-02  3.604e-03  -12.744  < 2e-16 ***
factor(day)Sun      -8.469e-02  3.629e-03  -23.338  < 2e-16 ***
factor(day)Thu      -5.964e-02  3.615e-03  -16.499  < 2e-16 ***
factor(day)Tue      -6.026e-02  3.613e-03  -16.679  < 2e-16 ***
factor(day)Wed      -4.715e-02  3.603e-03  -13.089  < 2e-16 ***
seasonspring         5.114e-02  2.796e-03   18.289  < 2e-16 ***
seasonsummer         3.264e-02  2.796e-03   11.674  < 2e-16 ***
seasonwinter         1.635e-02  2.919e-03    5.599 2.16e-08 ***
I(past.crime.30^2) -7.016e-05  4.326e-07 -162.191  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(14.0536) family taken to be 1)

    Null deviance: 881337  on 500111  degrees of freedom
Residual deviance: 528092  on 500096  degrees of freedom
AIC: 1821919

Number of Fisher Scoring iterations: 1
```

Fig 16: Negative Binomial Model summary

From fig 16, we can infer that the variables have **p-value** less than **0.05 and AIC value** is **1821919.**

```
[Previously saved workspace restored]

> sqrt(mean((test.data$count - crime.model.pred)^2))
[1] 1.820897
> summary(save)
Error in object[[i]] : object of type 'closure' is not subsettable
> summary(save)
```

Fig 17: RMSE value of Negative Binomial Model

From the Fig 17, we can infer that the **RMSE** value is **1.820897**

```
> summary(crime.model.poisson)

Call:
glm(formula = count ~ past.crime.1 + past.crime.7 + past.crime.30 +
    policing + crime.trend + factor(day) + season + I(past.crime.30^2),
    family = poisson, data = model.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0248  -0.8020  -0.2540   0.5547  10.7076

Coefficients:
                    Estimate Std. Error  z value Pr(>|z|)
(Intercept)       -1.409e+00  7.406e-03 -190.186  < 2e-16 ***
past.crime.1       1.258e-02  4.775e-04   26.353  < 2e-16 ***
past.crime.7      -2.773e-02  3.036e-04  -91.353  < 2e-16 ***
past.crime.30      3.114e-02  1.095e-04  284.395  < 2e-16 ***
policing           1.719e-01  8.653e-03   19.871  < 2e-16 ***
crime.trend        3.275e+00  2.347e-02  139.525  < 2e-16 ***
factor(day)Mon    -5.973e-02  3.255e-03  -18.347  < 2e-16 ***
factor(day)Sat    -4.536e-02  3.243e-03  -13.986  < 2e-16 ***
factor(day)Sun    -8.396e-02  3.272e-03  -25.659  < 2e-16 ***
factor(day)Thu    -5.865e-02  3.255e-03  -18.017  < 2e-16 ***
factor(day)Tue    -5.887e-02  3.254e-03  -18.092  < 2e-16 ***
factor(day)Wed    -4.592e-02  3.242e-03  -14.161  < 2e-16 ***
seasonspring       5.009e-02  2.522e-03   19.860  < 2e-16 ***
seasonsummer       3.230e-02  2.513e-03   12.855  < 2e-16 ***
seasonwinter       1.564e-02  2.648e-03    5.905 3.52e-09 ***
I(past.crime.30^2) -6.723e-05  3.859e-07 -174.183  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1023793  on 500111  degrees of freedom
Residual deviance:  618269  on 500096  degrees of freedom
AIC: 1777399
```

Fig 18: Poisson Model summary

From the fig 18, we can infer that all the x variables have **p-value less** than **0.05** and **AIC value** came out to be **1777399**

```
Residual deviance:  618269  on 500096  degrees of freedom
AIC: 1786164

Number of Fisher Scoring iterations: 5

> crime.model.poisson.predict <- predict(crime.model.poisson, test.data, type= "response")
> sqrt(mean((test.data$count - crime.model.poisson.predict)^2))
[1] 1.716278
> validate <- data.frame(test.data$count, crime.model.poisson.predict)
```

Fig 19: RMSE value of Poisson Model

From fig 19, we can infer that the **RMSE** value is **1.716278**

6.2. Results and Findings

Depending upon the AIC values first we selected the model with low AIC and then calculated the Root Mean Square Standard Error,which came out to be as follows:-

Models	AIC Value	RMSE value
Negative Binomial Model	1821919	1.820897
Poisson Model	1777399	1.716278

Table 1: AIC and RMSE values

From the above table we can see that **AIC value** of **Poisson Model** is **less** than **AIC value** of **Binomial Model**. Also when we plot the actual vs predicted depending upon the RMSE value we can see that Poisson model predicted line has less error as compared to Negative Binomial Model. Thus strengthening our analysis making the Poisson Model the best fit.

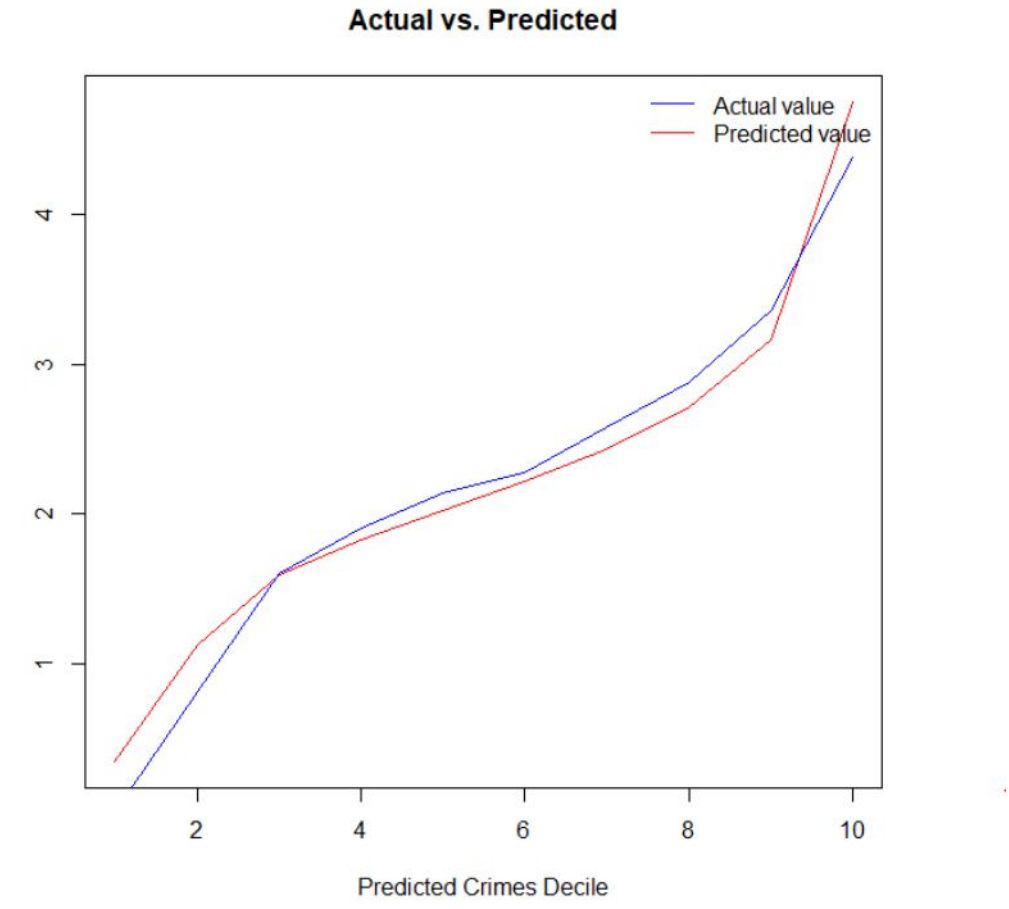

Fig 20: Representation of Actual vs Predicted Value for Negative Binomial Model

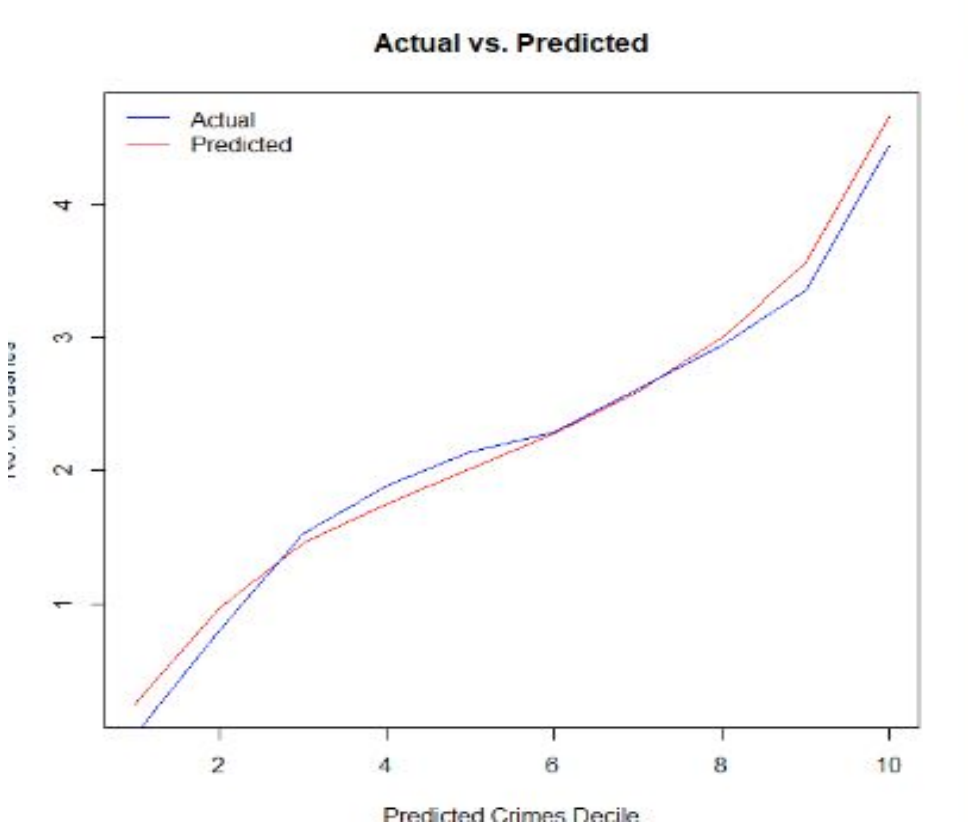

Fig 21: Actual vs Predicted values for Poisson Model

7. Conclusions and Future Work

7.1. Conclusions

We predicted the Time Series data with multivariate statistical approach using Negative Binomial Model and Poisson Model from which Poisson Model came out to be a best fit with respect to AIC and RMSE evaluation. Also, we came to know that for a model to be more accurate it is very important to clean the data meticulously. Data cleaning is the most important step for the model to predict better, which our case is the Poisson Model.

7.2. Limitations

The limitations to this project is that we used multivariate statistical approach because variance came out to be more that mean due to presence of categorical variables. This type of overdispersion is assumed to be true by these models. Also, spatial analysis on such model can increase the efficiency.

7.3. Potential Improvements or Future Work

We have implemented temporal analysis which is time series, but with the help of spatial analysis,for, example taking locations, longitude and latitude under consideration heat maps can be generated in order to analysis the crime location wise.