**Classification and Crime Predictions in the City of Toronto**

**Introduction:**

In this paper we will strive to identify patterns in crime offences in the City of Toronto over the past few years (2014 to 2019) and try to identify if certain geographical areas are more prone to crime and of what type, thus allowing us to predict the occurrence of it in the future. This analysis intends to help the city effectively organize and appoint police forces and other resources to those areas to aid in controlling crime rates in the coming years.

The theme chosen for this is Predictive Analytics ((Pattern mining, Time-series, Causality, etc.) to predict future trends as well as some Classification algorithms to arrange the data into related groupings.

**Literature Overview:**

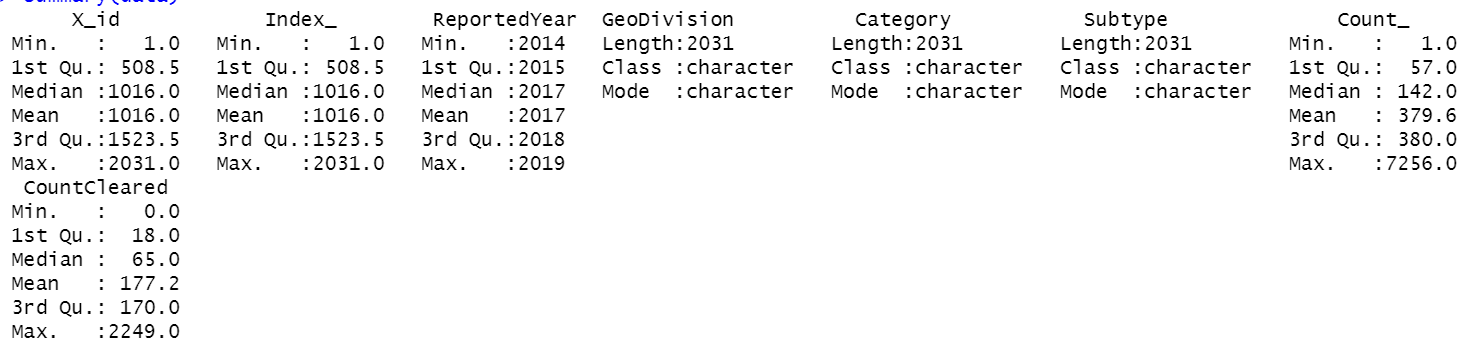
**Dataset:**

The data being used is from About Police Annual Statistical Report - Reported Crimes open dataset on the City of Toronto website.

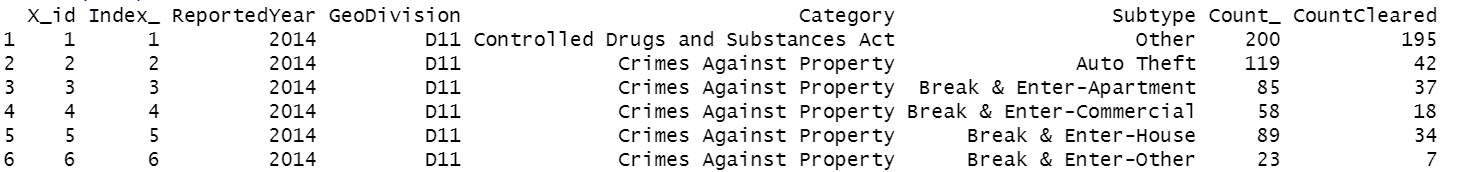
Dataset reference: https://open.toronto.ca/dataset/police-annual-statistical-report-reported-crimes/

Some example data features in this data set are GeoDivision, Crime Category and Count of the number of crimes within it during specific years.

Summary statistics of dataset.



This is how dataset looks like.

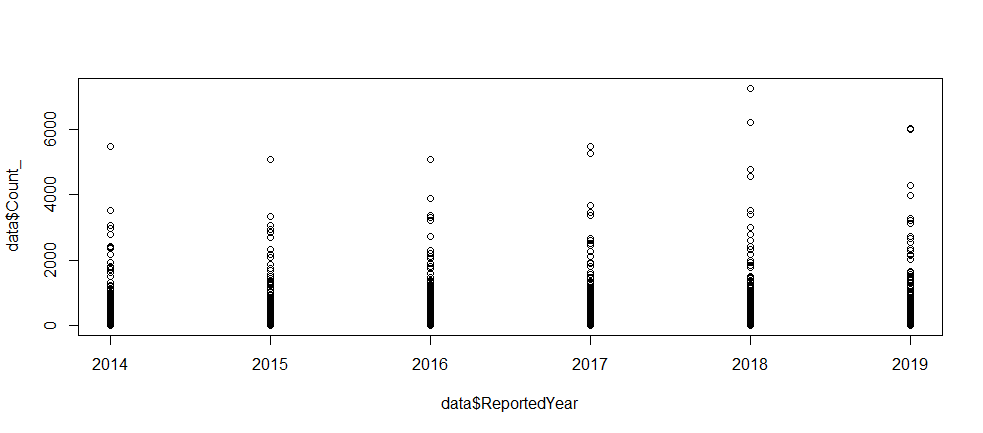


Further looking at the missing data points we found there were no missing data points. Which tells data set has been carefully dealt with.

The dimension of dataset provides an overview of whether if dataset will be able to provide enough insight or not. For this purpose, we analyzed dataset for its dimension, there are 2031 rows and 8 column features.

Now it was time to look into different column features. It is not possible to analyze the whole dataset with all the features it has. Because, there will be a possibility that we are also including useless features in our analysis. Thus, we pinpointed those features which did not contribute anything in understanding the pattern. For the sake of data purity, we did not use id and index features in our analysis.

A basic visualization for an insight of data is as follows. This visualization is between total crimes in a year and number of years. It shows that over all crimes are increasing steadily.



**Approach:**

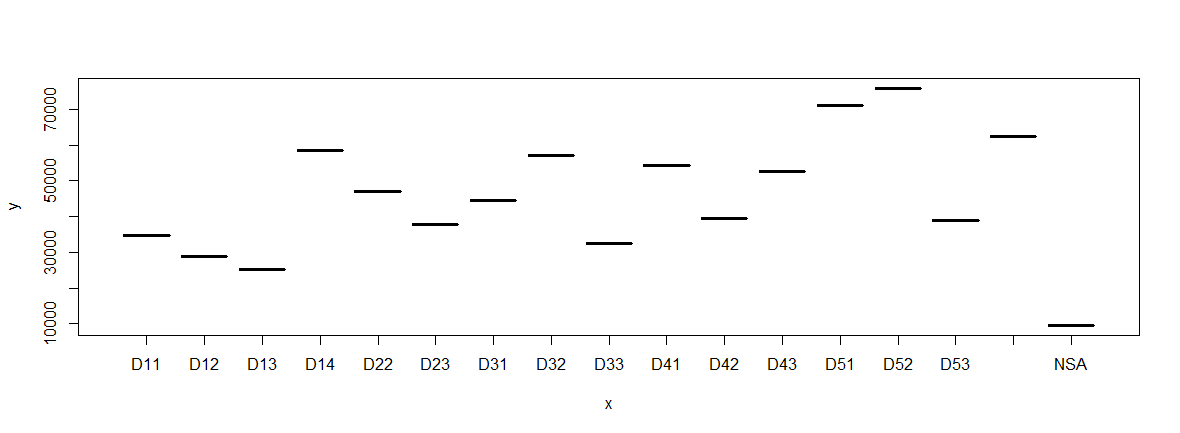
Following is the block diagram of the project flow.



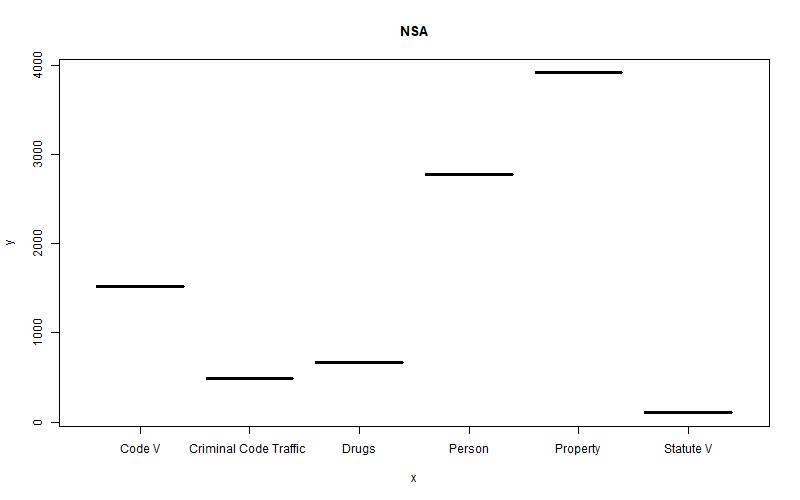
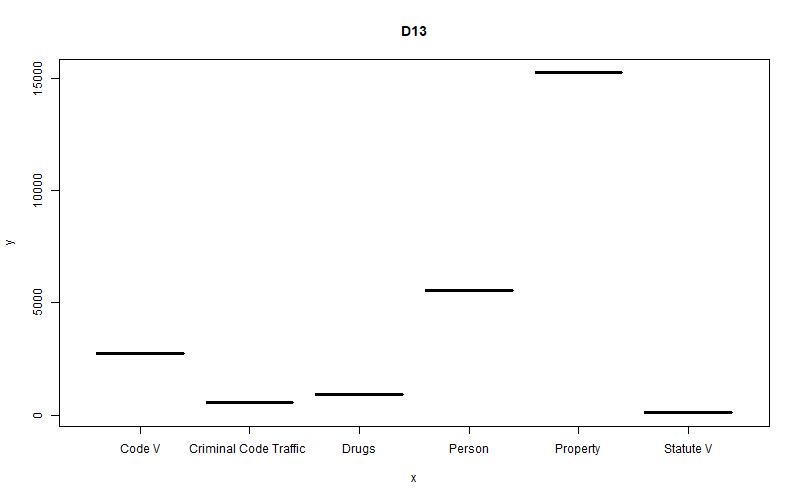
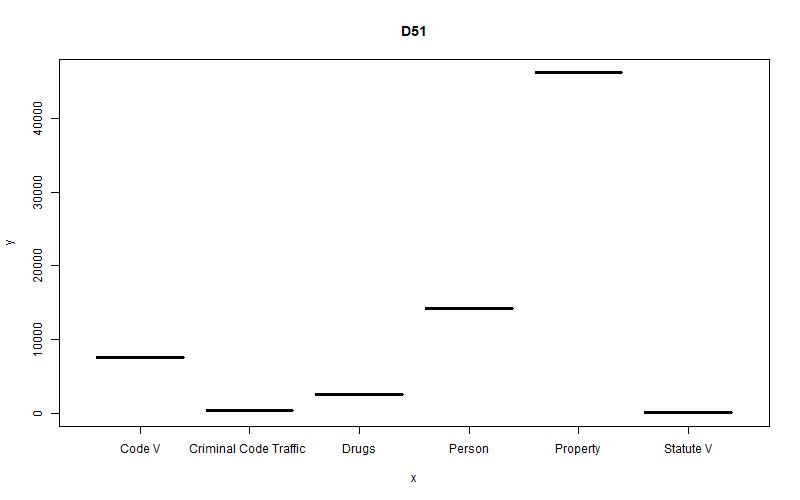
Step 1: Basic data analysis was done in the very first part. Here we became familiar with the overall structure of the data. Statistical summaries of each column were calculated, with analysis of missing data, dimension of dataset and most importantly pinpointing insightful features. All these activities were performed in this very first step.

Step 2: Now we had to do operations aligned with the three research questions. In this step, we researched about the very first question, which was. Is there different GeoDivisions within the city that are more prone to crime and if so, is there any trends in the type of crime categories being seen in those parts?

This research question consists of two parts, first we focused on the first part, which was to identify those geo divisions which were more prone to criminal activities. A visualization was used to conclude that D51 and D52 were the most struck by criminal activities. On the other hand, NSA and D13 are by criminal activities. This could be observed in the following image.



Now when we have concluded the first part. We jumped to second part of the research question. It was to observe the pattern of criminal activity types in the geo divisions.

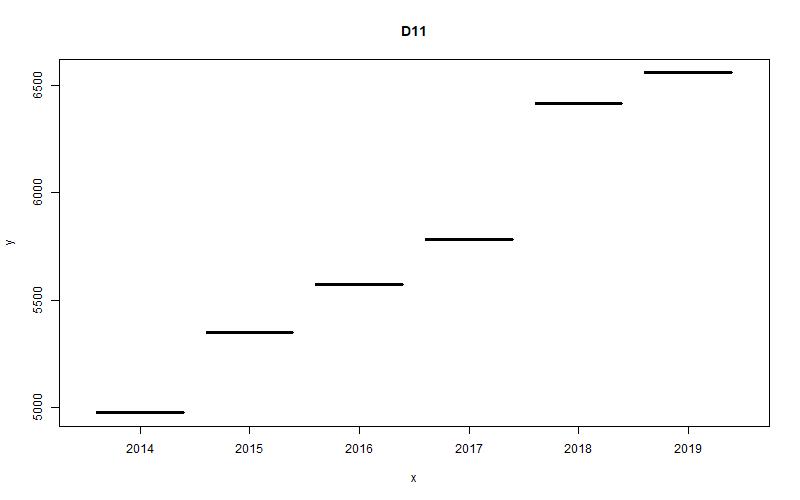


For observing the patterns of the type of criminal activities in highly effected areas and the less affected areas, we observe that patterns remain same almost everywhere. Property crimes are at their peak and the statute V are at lowest.

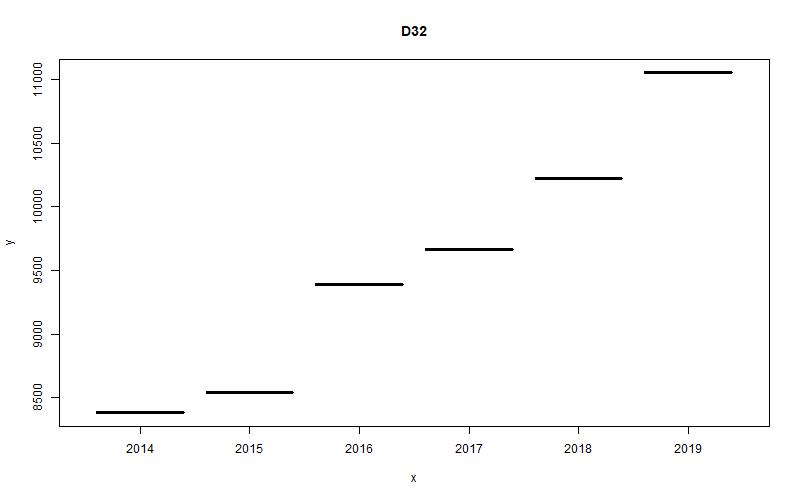
Step 3: Moving towards next research question. Is there geographical locations that have seen a big increase in crime over the years and/or is season a variable to consider?

For this we used visualization and analysis techniques to understand if total crime in a geo division is increasing or decreasing. We found, different divisions had their own patterns of criminal growth but D11, D31, D32 and NSA had been experiencing consistent increase in total crime for the past years.

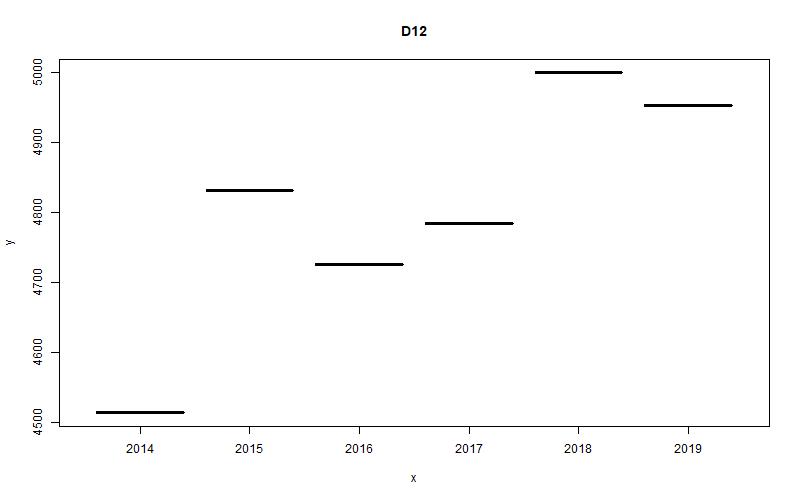
This can be understood by the following images, We can see D11 crimes are at consistent rise.



Same pattern is being observed in the D32.



But D12 is showing a mixed growth and fall of total crimes over years.



Step 3: Now we had to solve the last research question. Based off our findings, what types of crimes in the city can we predict for future years ahead depending on geographical area?

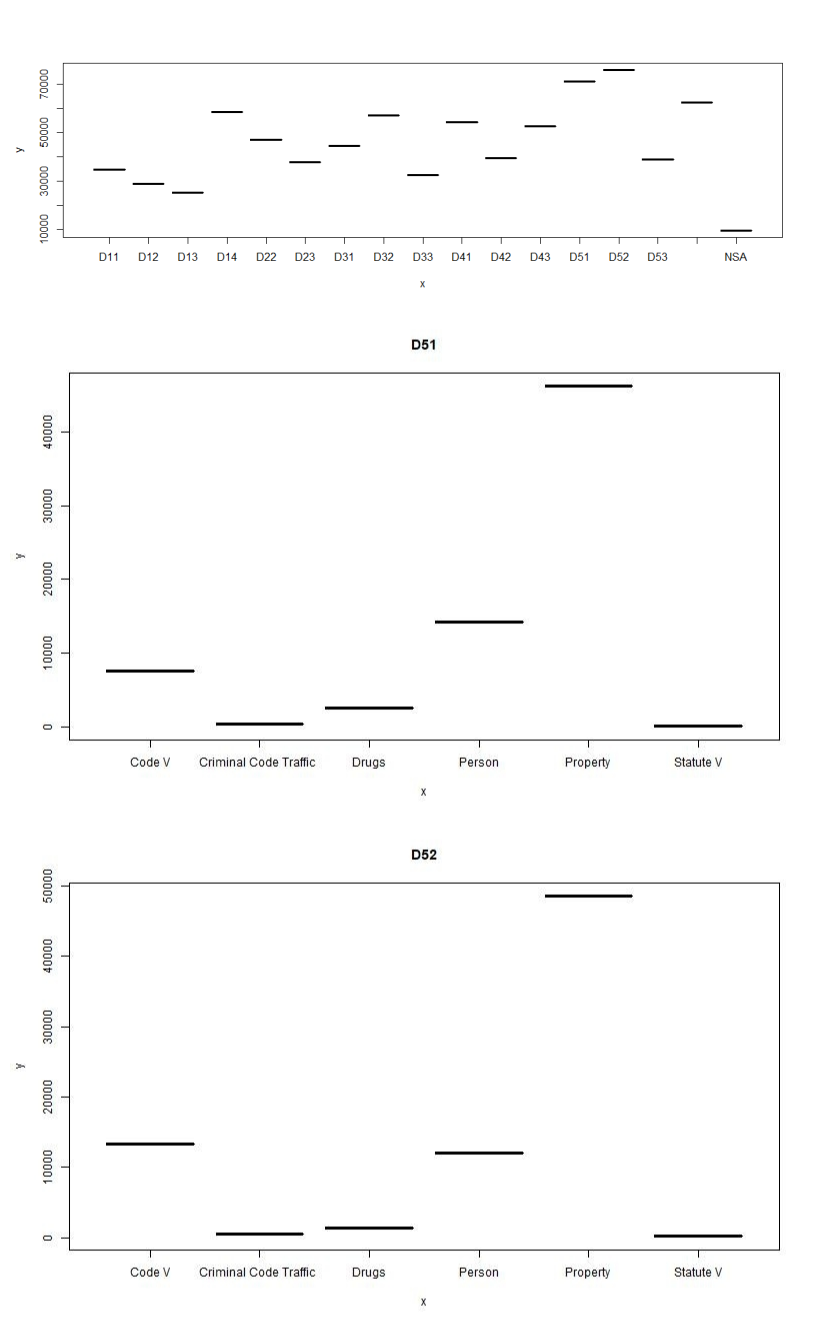
We used machine learning technique known as linear regression to predict the number of crimes. We considered count feature name as the dependent variable and geographical division, year and category of crime as the independent variables.

First, we converted the categorical values to numerical values so we can use them in the linear regression formula. After that, dataset was divided into train and test. The ratio of train and test was set to 70% and 30%. The test dataset values were successfully predicted and compared with the predicted values. A relatively weaker pattern recognition of was observed. This weaker pattern recognition could be seen from the weaker correlation values between all features. We will be discussing more about this in coming sections.

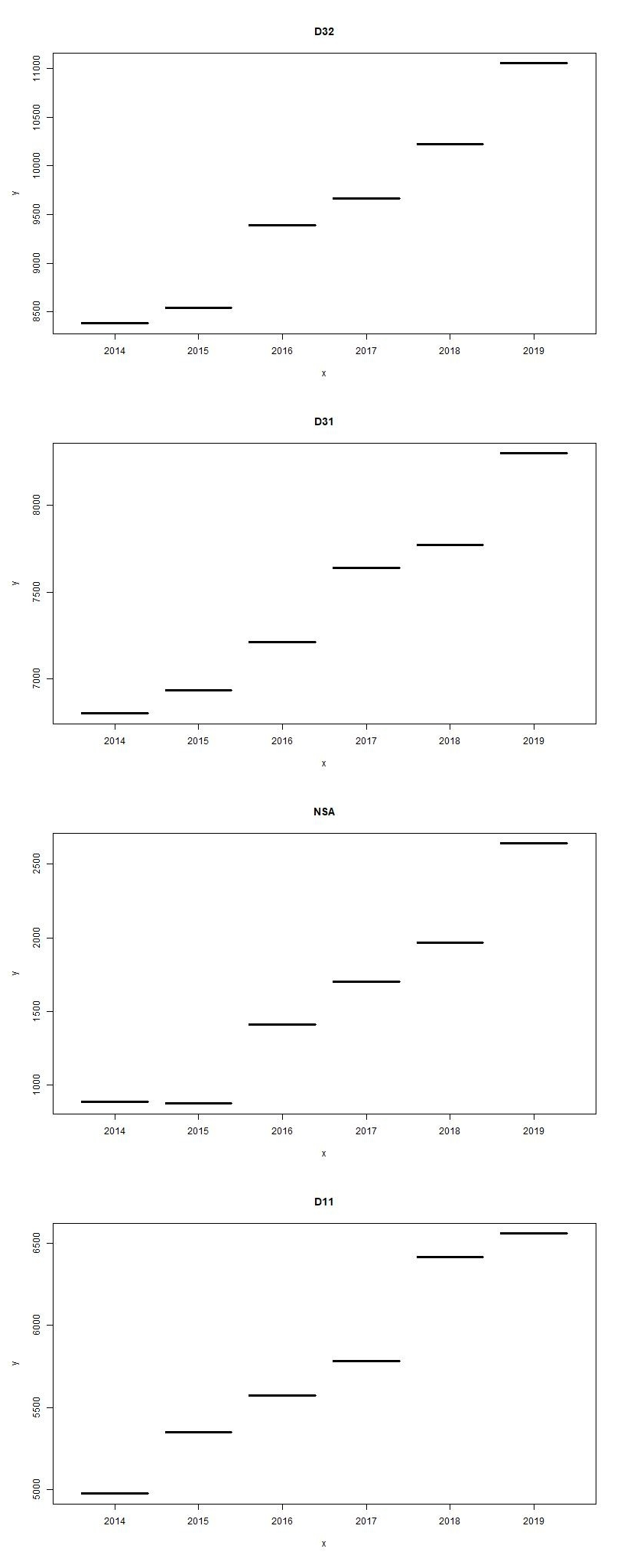
**Results:**

Previously we discussed about implementation and output of every step. But in this section, we will be discussing about the relevant output of each step and what it means.

The Research question 1, for the first part it is very clear that D51 and D52 are top in the criminal activities' magnitude. For the second part It shows, both in D51 and D52 property crimes are the highest. A bird eye view of both visualizations is as following.

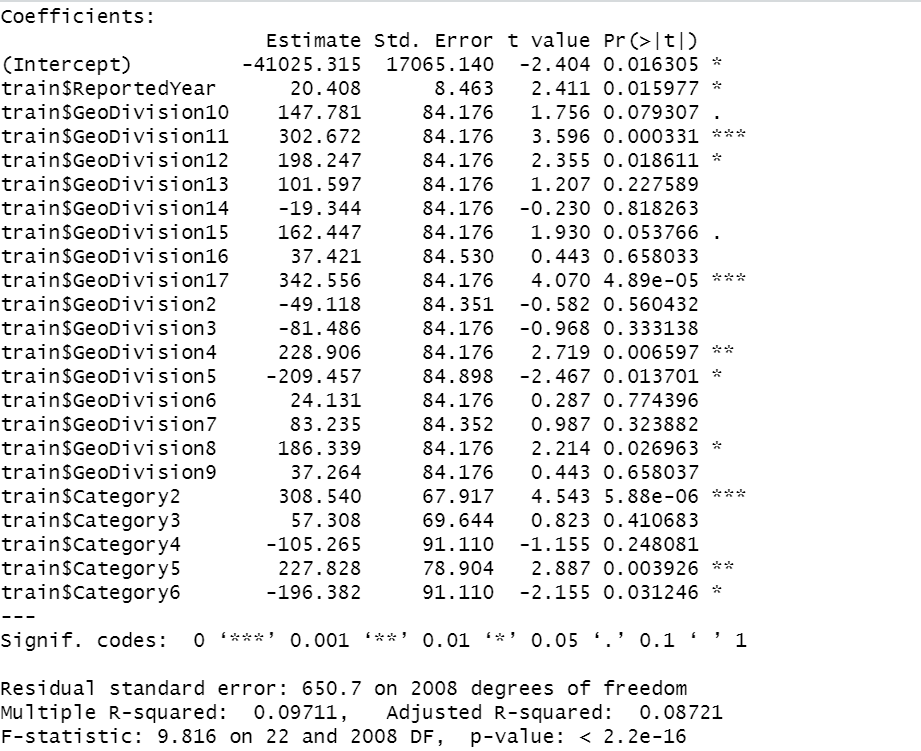


Research question 2 was about recognizing the divisions where trend of crime growth is alarming. Our analysis pinpointed 4 divisions named D11, D31, D32 and NSA where crime growth trend was alarming. Visualizations are as follows.

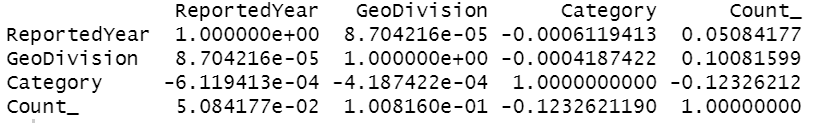


Research question 3 was about regression and predicting the future of crimes in divisions. We applied linear regression model and the summary results of the regression are as follows. The summary of linear regression is as follows.

Note: Lesser the p-value the significant the feature.



Linear regression finds patterns in the dataset and predicts it. By finding correlation among columns of the dataset we can find if there is any relationship/ connection and pattern in the dataset. Correlation ranges from –1 to +1. Zero correlation is bad and means no relationship between columns. The table of correlation between columns of dataset is given below.



Furthermore, we also used k-mean clustering for 5 clusters for the purpose of classifying the data based on their nature. Clustering was based on the crime count feature, which means all data points were clustered on the bases of magnitude of crime. All data points having higher crime count will be in one place and vice versa.

**Conclusion:**

We used the crime dataset with the intent to create a broader impact of the cutting-edge technology we have learned. We analyzed the dataset to extract meaningful information, which could be used by law enforcement agencies.

Initially, We figured out two areas (D51 and D52) out of all were the most effected by criminal activities. Furthermore, same crime type pattern was existing there, as property crimes were the highest in both these divisions. Despite that, we figured out D11, D31, D32 and NSA had a growing trend of crime and later we predicted future of crimes in the areas.

We successfully extracted patterns out of very complicated data, these patterns will be most beneficial for the law enforcement agencies.