Data Mining

CSCI-6674

**Analytics and Insights Team**

**Project Report Phase-4**

Master’s in Computer science

University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING, West Haven, CT

Submitted to:

Prof. Dr. Reza Sadeghi

Table of Contents

# Project Report: Phase - 4

**Team Details**

| | | |
|---|---|---|
| Team Name | **:** | **Analytics and Insights Team** |

**Team Members**

| | | |
|---|---|---|
| 1. Vyshampa Maringanti | : | vmari2@unh.newhaven.edu (Project Lead) |
| 2. Akhila Gade | : | agade1@unh.newhaven.edu (Team Member) |
| 3. Bagyasree Chitikela | : | bchit1@unh.newhaven.edu (Team Member) |

**Research Question:**

Can we predict the number of offenses for a given precinct within the next three months based on historical crime data of the New York Police Department (NYPD) by using Negative Binomial Regression?

**Research Objective:**

The main objective of this project is to create a predictive model using Negative Binomial Regression1 to predict the number of crime occurrences/offences for a given precinct of the NYPD2.This prediction will aid the police department in determining the required patrolling manpower per precinct.

**Dataset Description:**

The dataset covers the data of all crimes reported to the New York Police Department from 2006 to the end of last year i.e.,2019. This is a large dataset consisting of 6.98 million records with 35 attributes. Given, the large size, we could filter data to a more manageable size and per our project requirement.

## Data set Source:

https://www1.nyc.gov/site/nypd/stats/crime-statistics/citywide-crime-stats.page.

# Exploration Techniques -EDA

EDA(Exploratory Data Analysis) is one of the data analysis methodologies used to summarize dataset characteristics with statistical numbers and graphs. In our project we implemented EDA on NYPD Dataset. NYPD maintains monthly crime data based on police incident reports across New York. The data is saved on the NYPD website with each month corresponding to an Excel file. Pre-processing algorithms in Python were used to connect directly to Excel files in the web source to retrieve the required information.

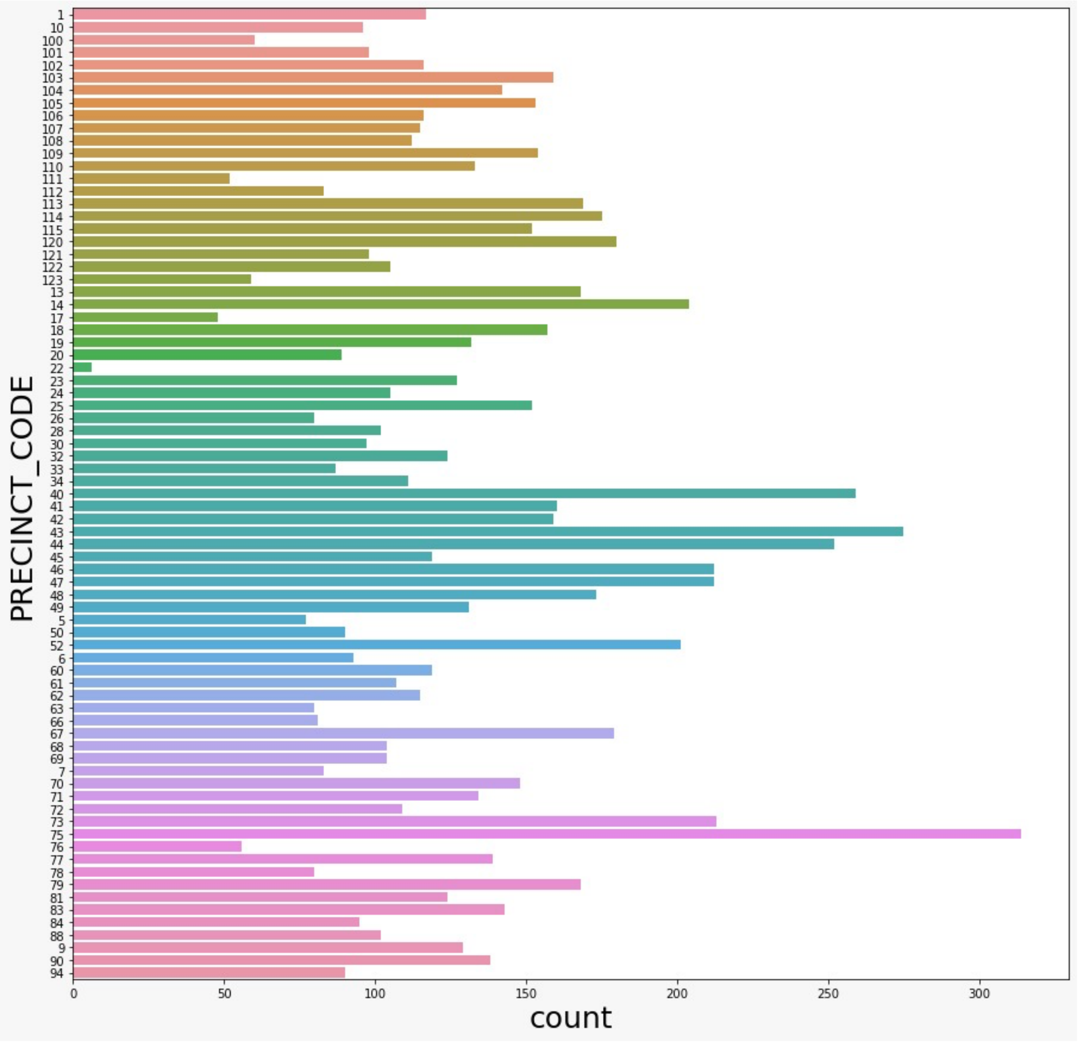
A lot of statistical and machine learning algorithms could be used as an EDA technique. Here, we picked a few primary techniques to illustrate.

| **Exploratory Method** | | **Visualization Technique** |
|---|---|---|
| **Univariate** | Categorical | Count Plot |
| | Numerical | Histogram |
| **Bivariate** | Categorical ‘Vs’ Numerical | Box Plot, Scatter plot, Pie chart, violin plot, bar plot |
| | Numerical ‘Vs‘ Numerical | Line Plot or Replot |

# Data Exploration from Different Perspectives

1. **Unique Value Count:**

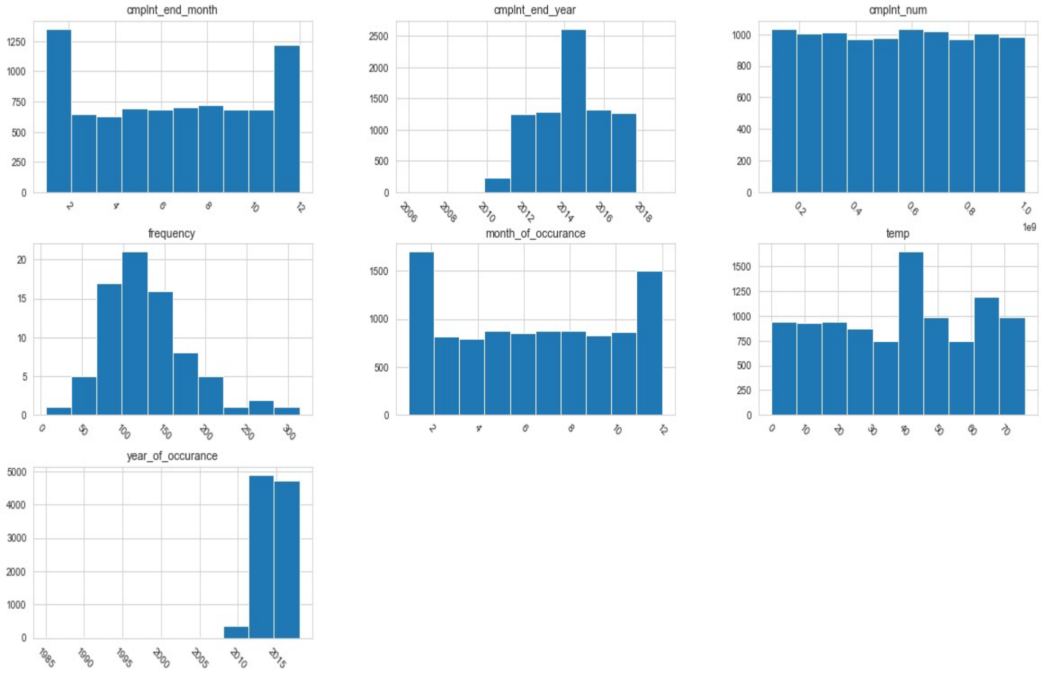
One of the first items that was used during data exploration is to see how many unique values there are in the categorical columns. This gives you an idea of what the data are about. The unique number of categorical columns in the dataset of the precincts is seen here.



The categorical column with the maximum number of unique values is Precinct_code-75 which  has received highest number of complaints or unique values that are approximately 350. This column has 77 unique precinct codes among which the highest number of complaints filed are in precinct-75.This category has minimum value of precinct noted as-precinct-22. If we observe maximum number of unique values is in "precinct-75" column, which means that the most of the valuable information required for research is mainly around different complaints in precinct-75.

2. **Histogram:**

It offers details on the spectrum of values in which most values fall. It also provides details as to whether there is a skew in the results. If we make a histogram in the **month_of_occurance**, it displays the highest and lowest number of complaints per month. Similarly, **cmplnt_end_month** column describes in which month maximum and minimum number of crimes that have been successfully completed by suspects. The temp is the label encoded column of **Addr_pct_cd** which is number of complaints at specific precinct. **Frequency (count(temp))** chart shows the highest and lowest frequency values.

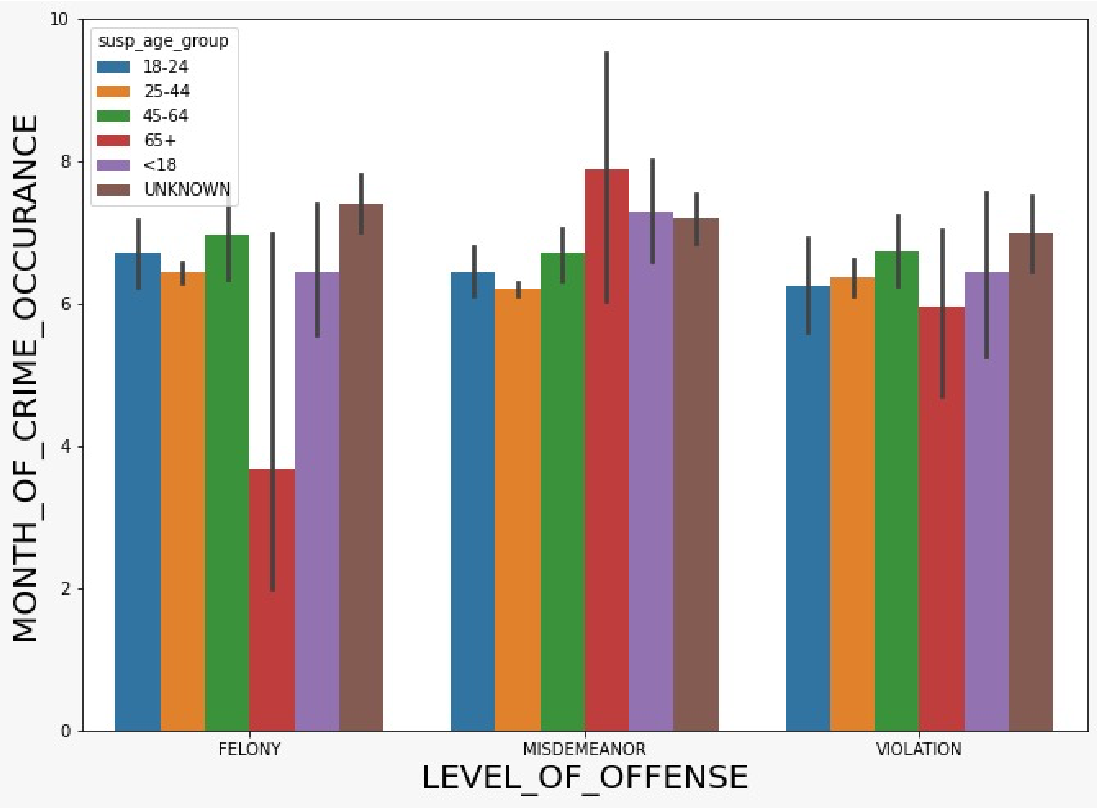


In the above Graphs Frequency graph represents normal distribution curve. Hence, implementation of Linear regression model is possible.

### 3. Cluster size Analysis:

Grouping data together makes it possible for us to get that high-level perspective. Data groups allow us to look at groups rather than individual data points first. This grouping is also called as clustering or segmentation. As a first step in segmentation, cluster size analysis reveals how data can be separated into various categories.

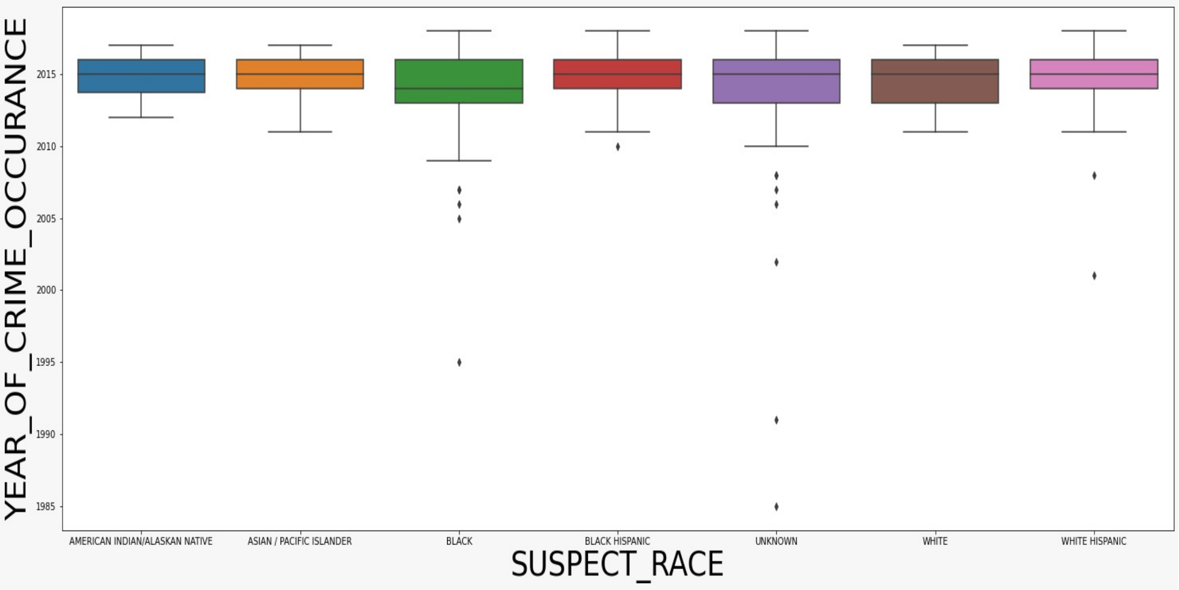
As we may observe, we split all data into three classes. We have clusters that are more or less of the same size. Each group represents different levels of offenses such as FELONY, MISDEMEANOR and VIOLATION. In the below graph we can observe the highest level of crime occurred in the month august and the age group of the suspect was 65+ and the level of offense was MISDEMENOR. Lowest offense was also done by the suspects with age group 65+ (offense type= FELONY).



## 4. Outlier Perspective:

By using outlier detection method we tried to find out the unusual data which is outlier detection/ Anomaly Detection. Outliers represent unusual, rare, anomaly or exceptional. Outliers analysis also helps to enhance the quality of exploratory data analysis.

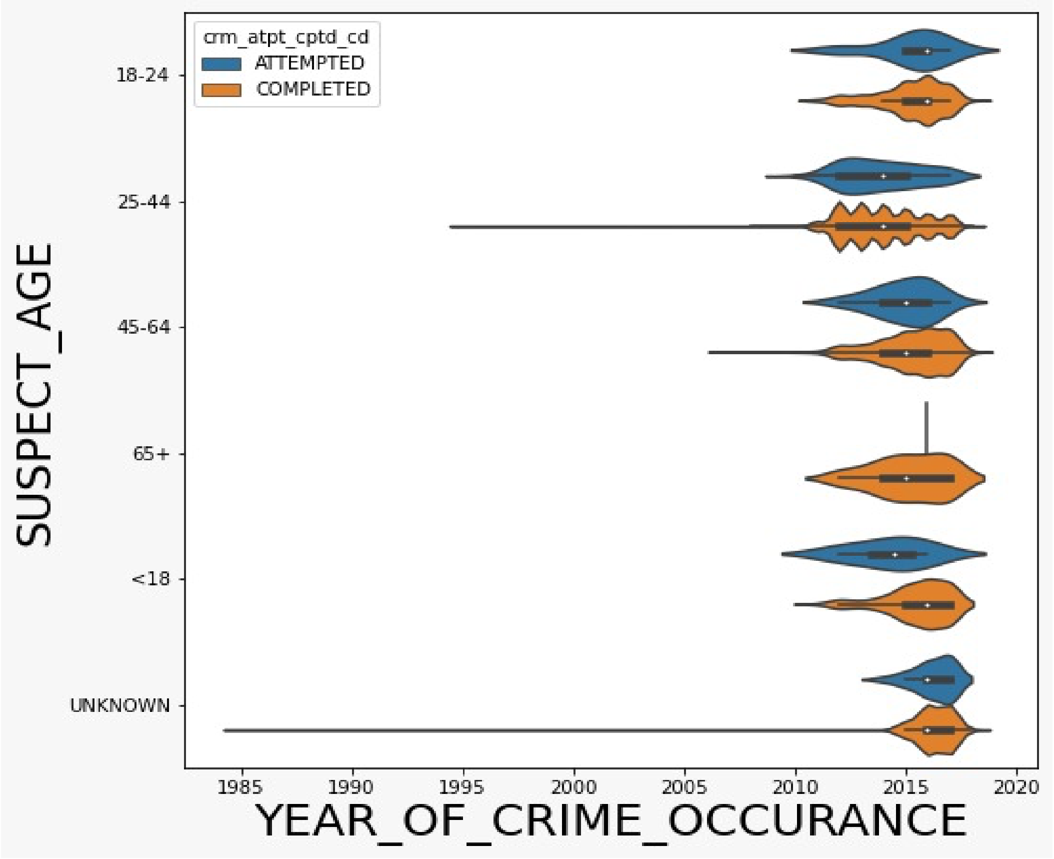
The below graph represents box plot between the columns of **Year_of_crime_occurance** (Numerical )and **Suspect_Race(Categorical).** The below box plot explains the Unknown has maximum outliers as 6. Next highest outliers are in Black race. Least outliers are around white Hispanic race. Also, it can be deduced that most of the crimes were occurred between 2012-2018.



### Violin Plots:

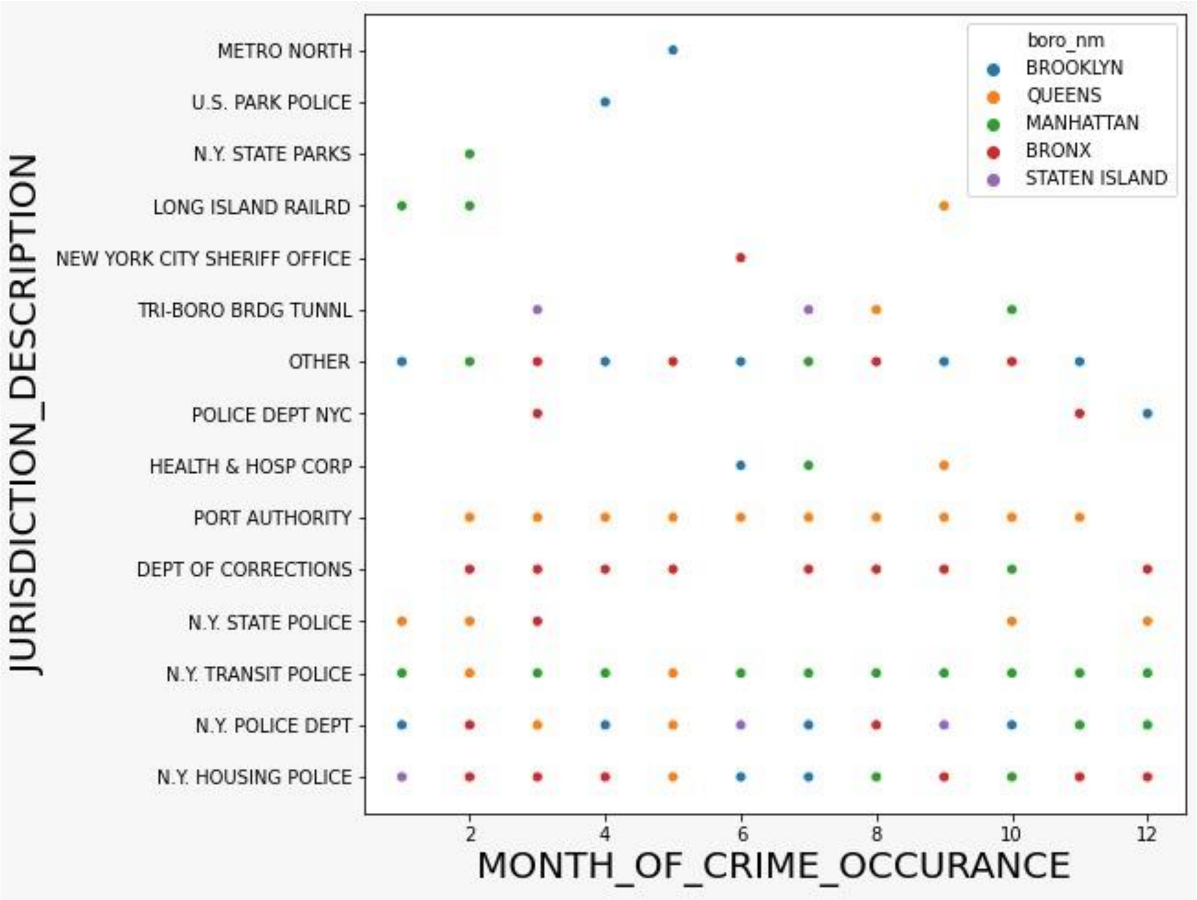
Violin plot is a method for plotting numeric results.Here we used one numerical and one categorical value to plot violin graph. It is very similar to a box plot, with the addition of a rotated kernel density plot on either side. Also, Outliers are defined with a solid black line.In the below violin chart Suspect_Age is in the y-axis and x-axis as Year_Of_Crime_Occurance.

Blue label denotes the Number of crimes attempted which is very high between the period of 2012- 2017. Yellow label denotes the successfully completed crimes by the suspects with the age group of 65+. Also, we can observe that all these suspects were completed the crimes successfully that they have attempted.This insight might help while doing data modelling.When it comes to outliers most of the outliers were detected during the years 1995 and 2010 at the age group of the suspect is around 44+.



5. **Scatter Plots:**

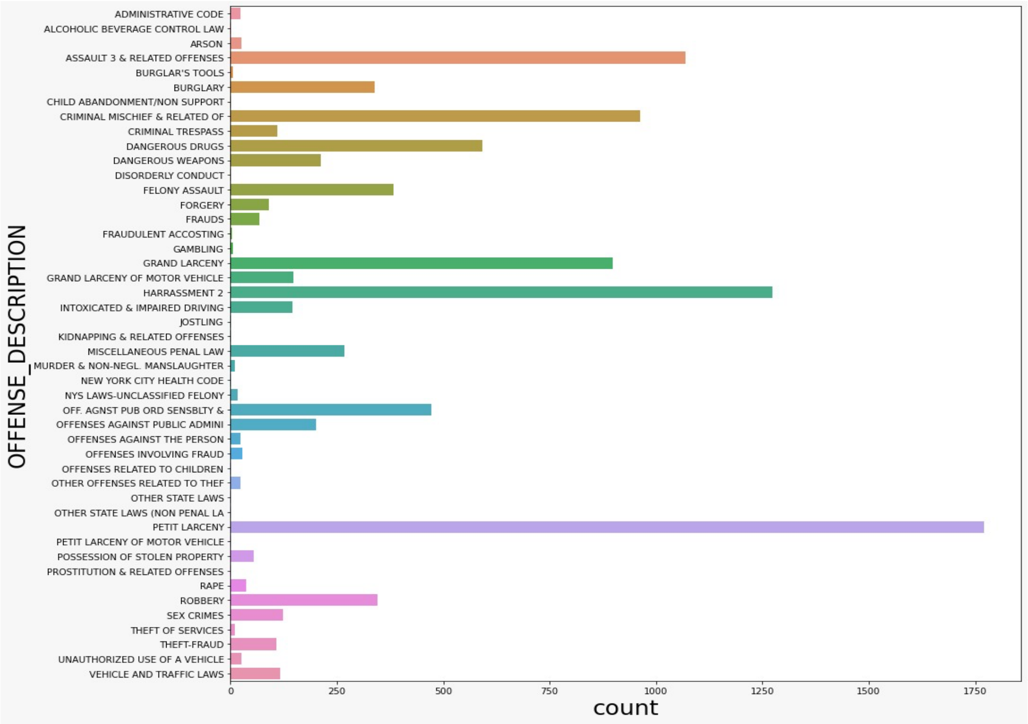
Scatter plot was built between Jurisdiction and Month of crime. If we look at the spread of the crimes in the scatter plot, we can deduce that most of the crimes were occurred in March in BRONX boro jurisdiction. Even though there are the same number of dots in February but there were more outliers in the data. One more observation can be made by looking at the port authority orange dots in the graph which is detection of continuous crime  from the months between February and November.



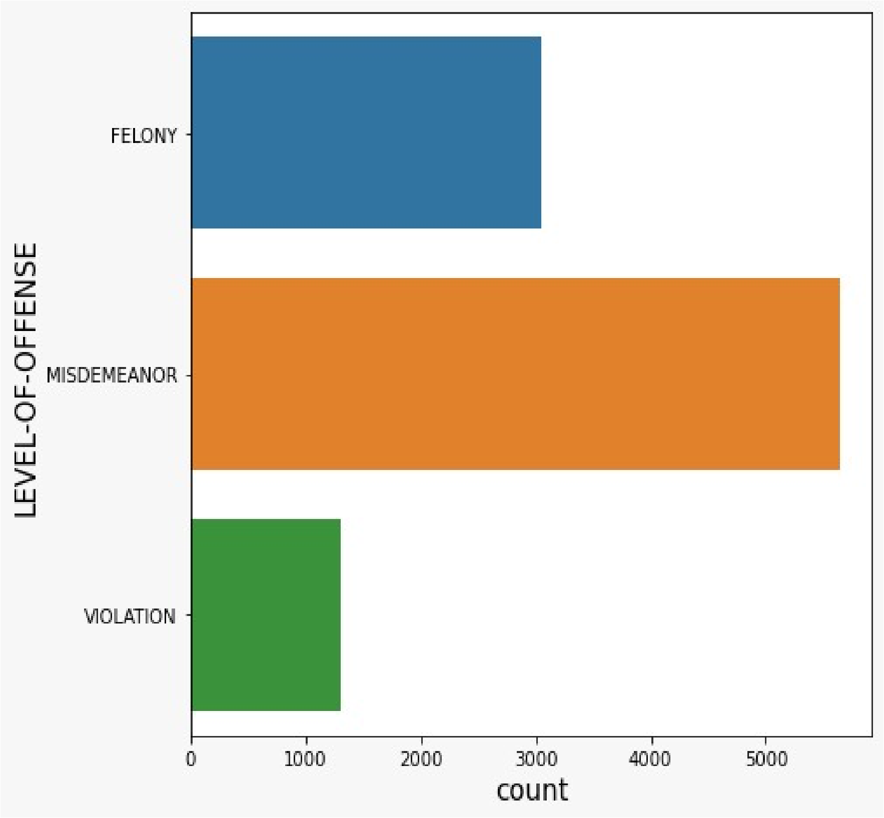
## 6. Categorical Perspective:

One of the very first things that could be useful during data exploration is to see how many unique values there are in the categorical columns. This gave us an idea of what the data are really about.

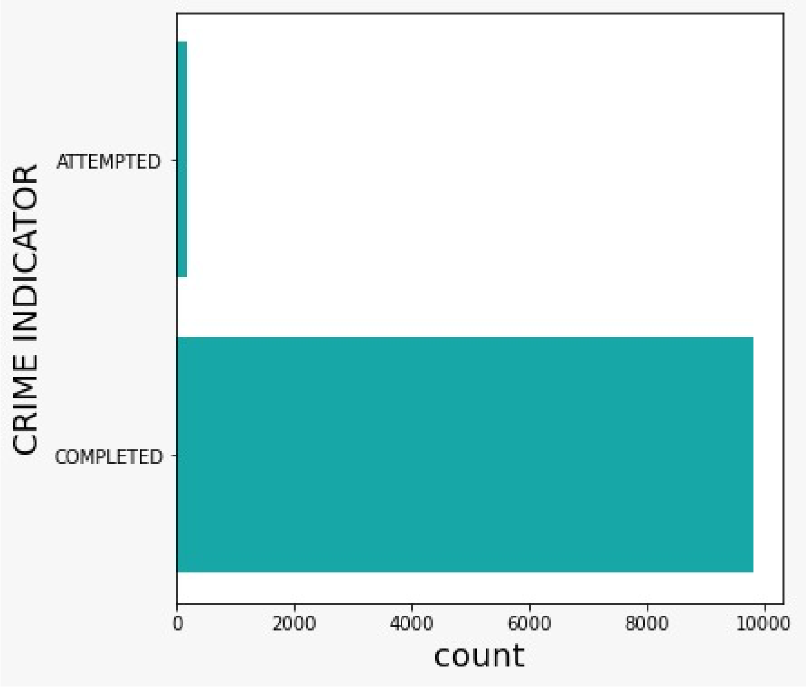
a) **Offense Description:** The unique number of categorical columns in the dataset of the unique Offense Descriptions is seen here. x-axis represents the count of offenses and y-axis denotes the Offense description numbered from 1 to 45 labels. Highest number of crimes were for the offense **Petit Larceny** which is more than 1750.



b) **Level_Of_Offense:** The below graph represents the Level of Offense as a unique categorical value. By looking at the graph we can observe that **Misdemeanor** was the most (more than 5000) committed offense among all these offenses listed. As you can see the green bar in the graph which has the Least level of offense was recorded as Violation.

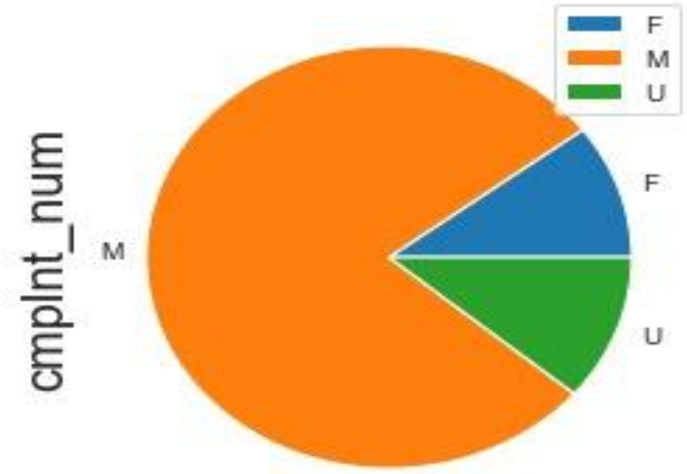


c) **Crime Indicator:** The below graph represents the crimes which have been completed or attempted. Here, the Crime indicator indicates that number of crimes which have been successfully completed or attempted by the suspect. We can observe that a greater number of crimes have been successfully implemented by the suspect with the count of approximately 9800, whereas the  crimes that have been interrupted or failed were less than 100.



7. **Categorical Vs Numerical :**

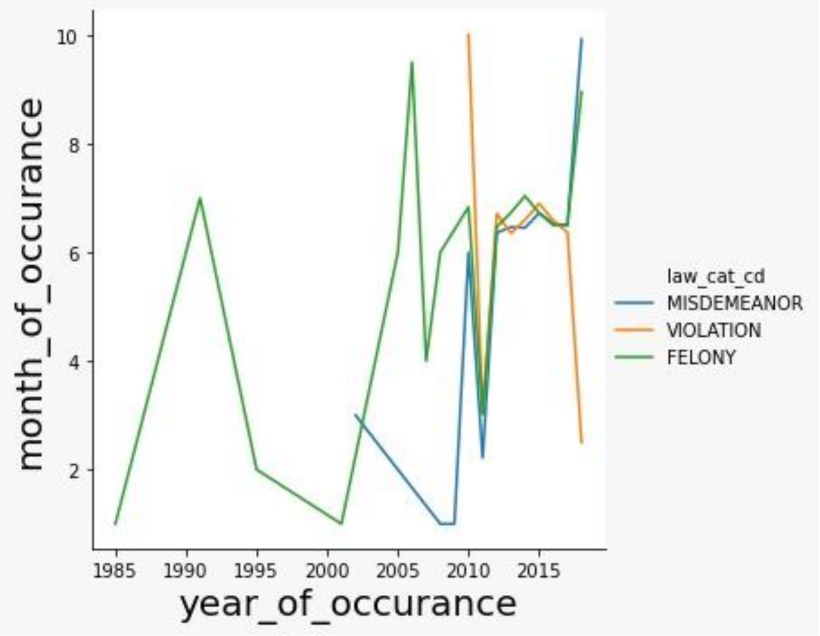
The below pie chart is plotted between suspect sex (Categorical )and complaint number (Numerical). Suspect_sex feature has three categories namely, Male(M), Female(F) and Unknown(U). The reason for having  Unknown category is because we do not have enough information about suspect gender. We observed from the graph that a greater number of complaints have been filed against Male category.



## 8. Numerical Vs Numerical:

In this graph x-axis and y-axis were generated from a single feature called “Cmplnt_fr_dt”. The feature 'cmplnt_fr_dt' has yyyy-mm-dd format, by using DatetimeIndex() function in pandas we divided this into two separate columns as year and month. We also used ‘Law_cat_cd’ (level of offense) attribute as a hue in order to get more insights about the data.

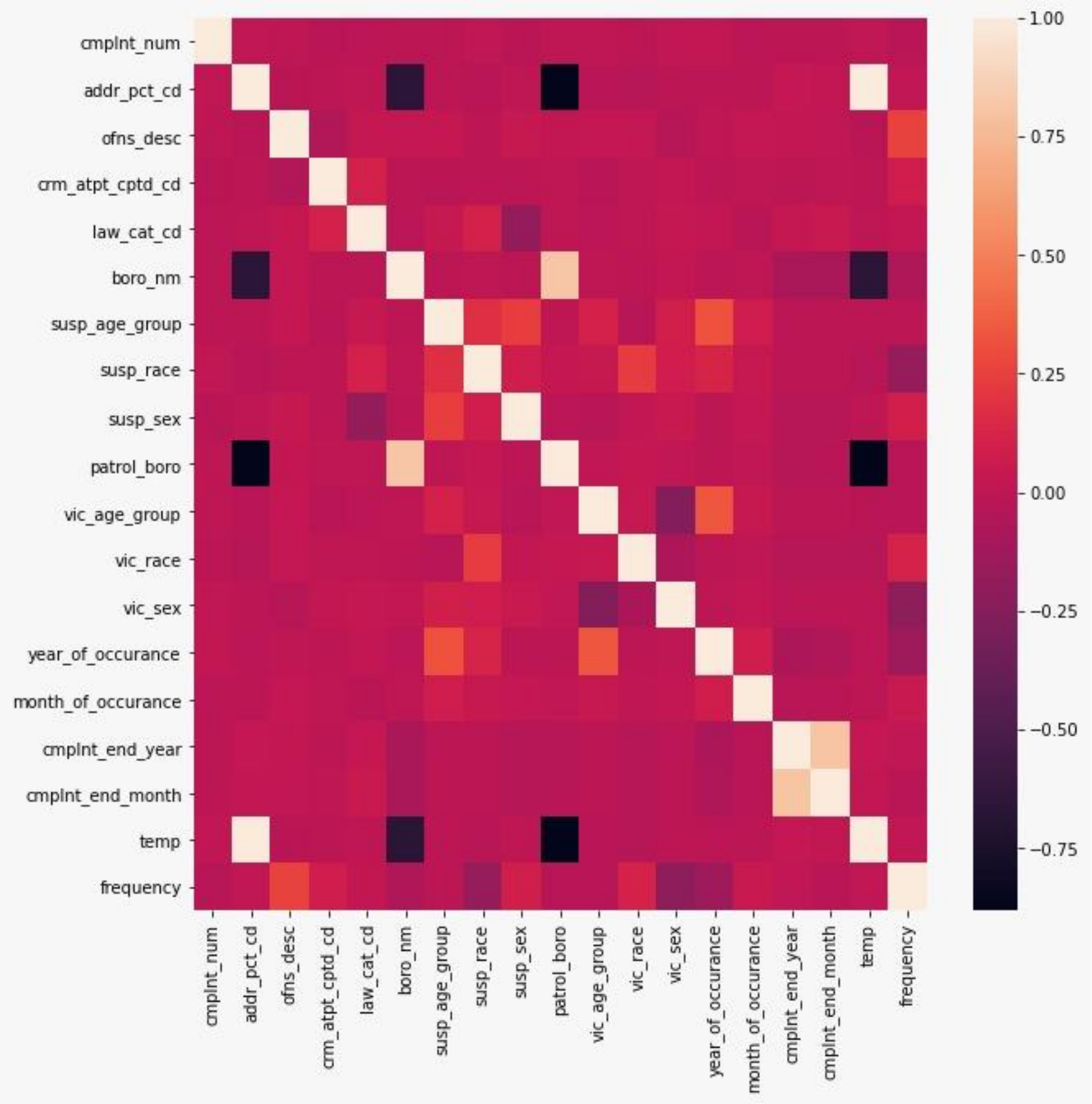
The observations from the graph are that, FELONY was the most occurring level of offense from 1985 to 2015. Highest value of VIOLATION was observed in October 2010 but decrease in VIOLATION can be observed after 2015.It can also be deduced that MISDEMEANOR trend is increasing after 2011.



## 9. Correlation Between Categorical Features:

Heat map is a very useful data analysis technique for discovering a connection between Categorical variables. There are several categorical columns in our NYPD crime dataset.

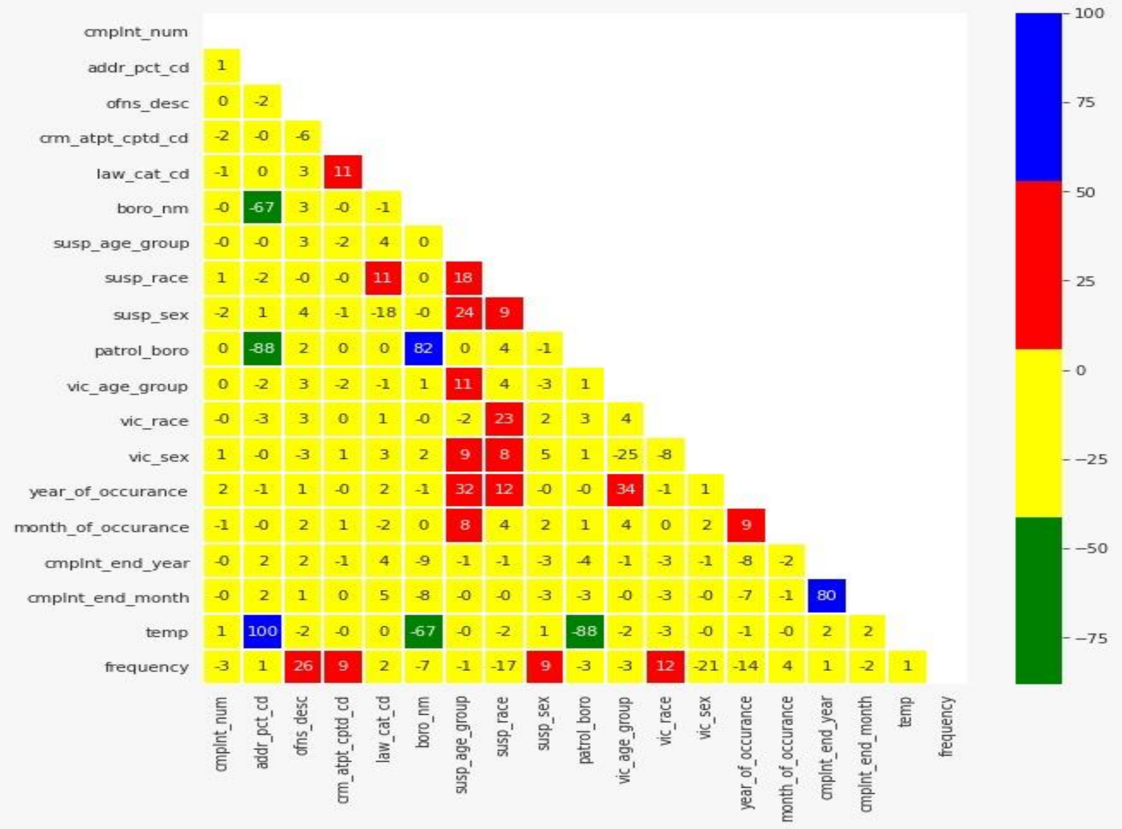
The below is correlation map by using Heat map. High correlations are shown with light color and low correlations are shown in dark colors like black and dark red. In the below map the top correlations are between **Cmplnt_end_year** and **Cmplnt_end_month .** The medium correlation is between **Patrol_boro** and **Boro_nm**, **Ofns_desc** and **frequency**, **Year_of_occurance** and **susp_age_group**. Many other categories come under low correlation for example **vic_age_group** and **vic_sex**.



**Heat Map Mask:** Correlation tests the dependency of two variables. It also tests "how two variables are dependent on each other " and "how strongly they connect" means an increase in one factor and an increase in another. To verify this dependency, we implemented heat map mask by using Seaborn Library.

Seaborn Heatmap feature has a mask statement that lets us to select the elements from the input data frame. In our example, we want to mask the upper triangular elements to make the lower triangle correlation map.

In the below heat map dark blue and dark red represents Positive correlation and Negative correlation respectively. Hence, green color  represents Negative correlation between the features and blue represents positive correlation between the selected features.



# Summary

1. Histogram: As per the histogram we can conclude that frequency(histogram) graph represents normal distribution. Hence implementation of linear regression model is possible.
2. Unique value count: The categorical column with maximum number of unique values is precinct code-75 which has received highest number of complaints.
3. Cluster size analysis: Highest level of crime occurred in the month of august and age group of suspects was 65+ and level of offense was misdemeanor.
4. Outlier Perspective: Most of crimes are occurred in 2012-18. most suspected race was white Hispanic and black.
5. Violin plot: Number of crimes attempted which is very high between the period of 2012-2017. successfully completed crimes by the suspects with the age group of 65+.
6. Scatter plots: Port authority (orange dots) in the graph which is detection of continuous crime from the months between February and November.
7. Categorical Perspective:

   offens desc: Highest number of crimes were for the offense Petit Larceny.

   Level of offense : we can observe that Misdemeanor was the most (more than 5000) committed offense among all these offenses listed.
8. Categorical Vs Numerical: We observed from the graph that a greater number of complaints have been filed against Male category.
9. Numerical Vs Numerical : The observations from the graph are that FELONY was the most occurring level of offense from 1985 to 2015.

10. Correlation between Categorical Features: In the heat map the top correlations are between Cmplnt_end_year and Cmplnt_end_month. The medium correlation is between Patrol_boro and Boro_nm, Ofns_desc and frequency, Year_of_occurance and susp_age_group. It is also observed that few features have strong correlation, and few got very less correlation.

# References

https://towardsdatascience.com/15-data-exploration-techniques-to-go-from-data-to-insights-93f66e6805df