SanFrancisco Crime Analysis

This project examines the SanFrancisco police department crime records from 2012 to 2015.

Data Fields

Dates - timestamp of the crime incident

Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.

Descript - detailed description of the crime incident (only in train.csv)

DayOfWeek - the day of the week

PdDistrict - name of the Police Department District

Resolution - how the crime incident was resolved (only in train.csv)

Address - the approximate street address of the crime incident

X - Longitude

Y – Latitude

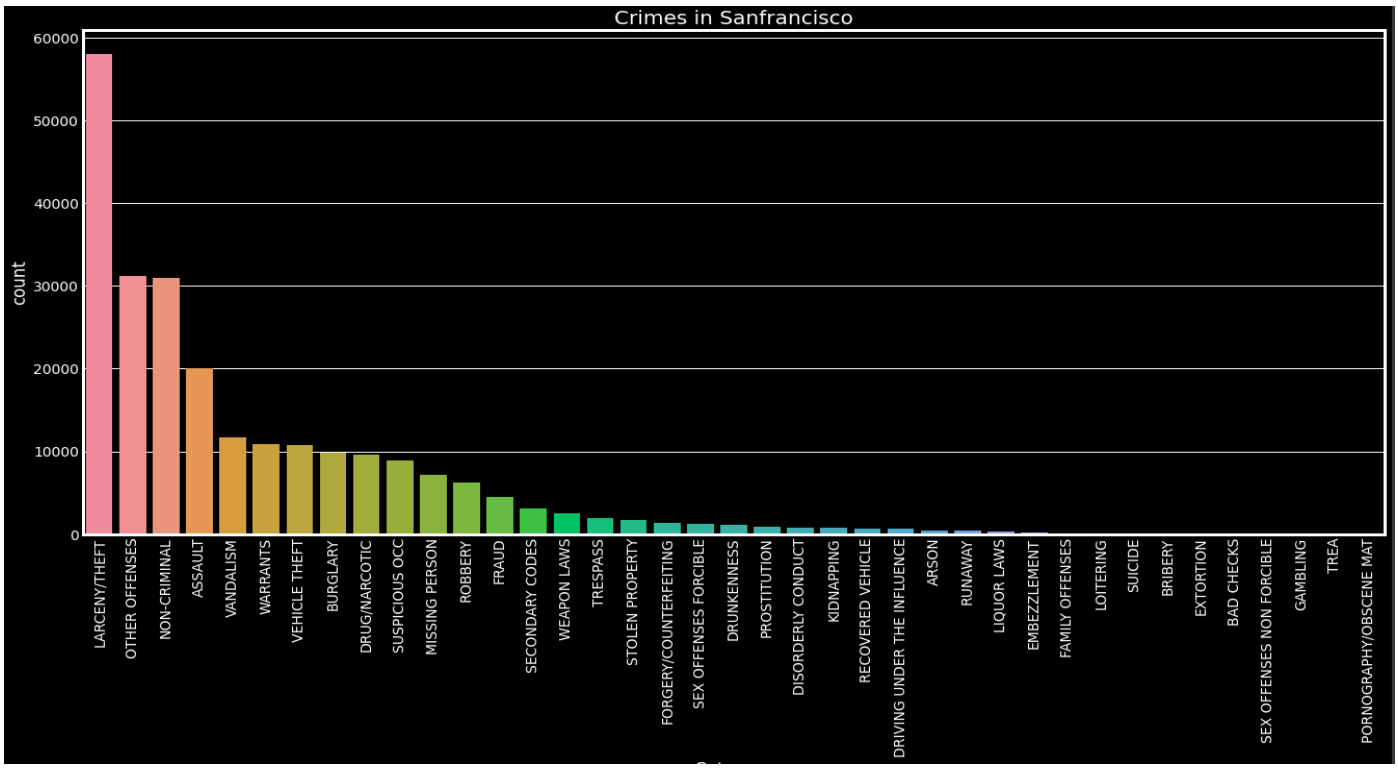
Approach

Python is used for data proceesing and Visualization.

A Deep learning model and a XGBoost model were built for classification and compared their performances.

Data Visualization - Analysis of Crime:

There are 39 category of report incidents available in the dataset in which LARCENY/THEFT was the most common crime with the total of 58021 reported incidents. Similarly, PORNOGRAPHY/OBSCENE MAT is the least occured crime in the city during that time period.



The below fig shows the common resolutions provided for each category of Crimes.

Some intresting facts from the below observations are:

1)More people are getting arrested/booked for the Warrants type of Crimes. Similarly nobody is arrested for Suicidal crimes.

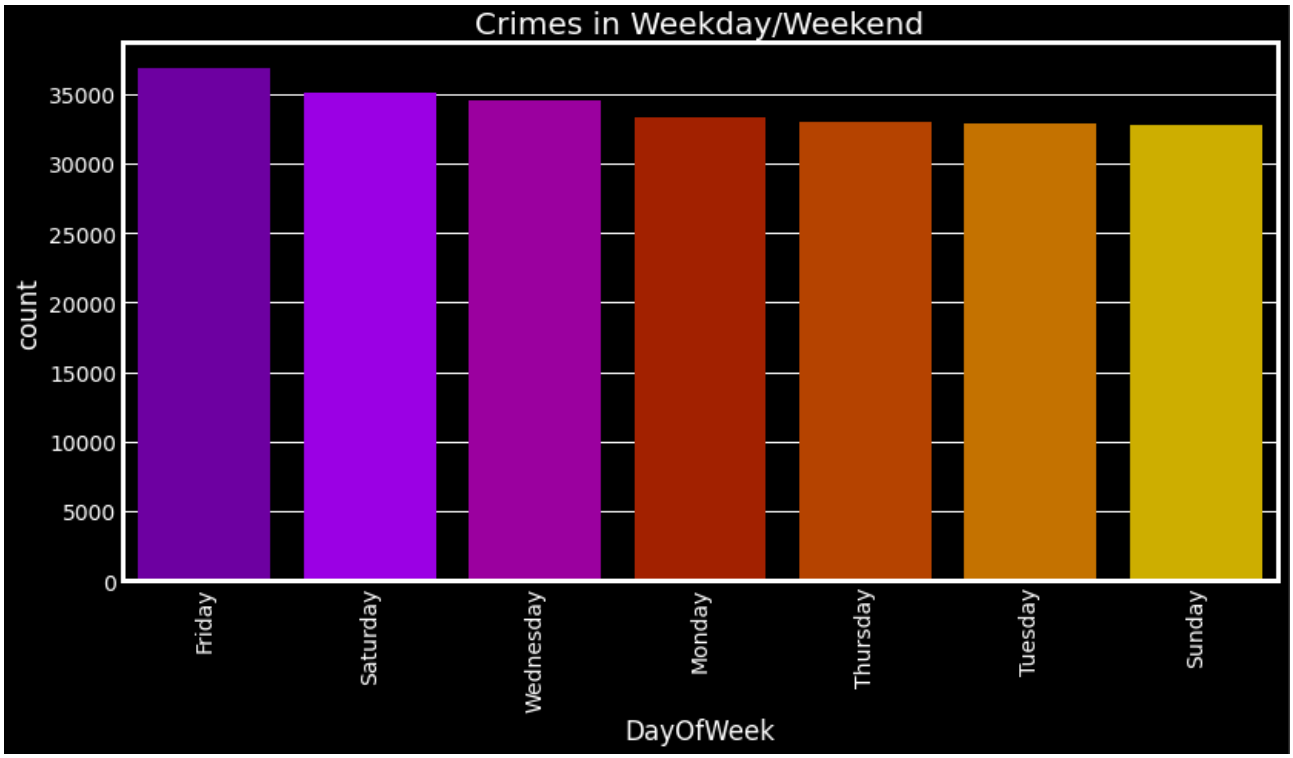
2)For most of the Vehical Theft cases, no action has been taken as a part of resolution.

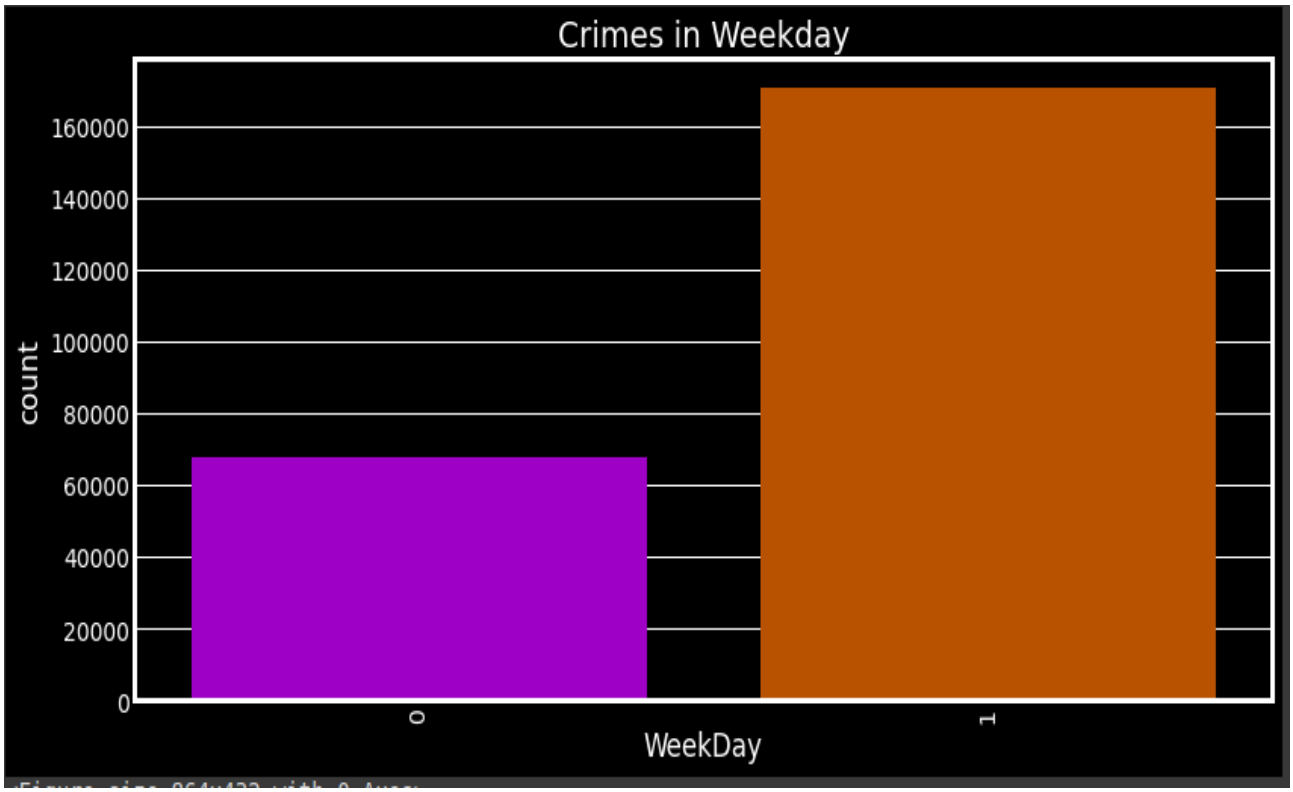
```
WARRANTS
ARREST, BOOKED      90.4
NONE                 7.2
JUVENILE BOOKED      1.2
Name: Resolution, dtype: float64

WEAPON LAWS
ARREST, BOOKED     63.4
NONE               22.2
ARREST, CITED       5.9
Name: Resolution, dtype: float64
```

```
SEX OFFENSES NON FORCIBLE
NONE                                      53.1
ARREST, BOOKED                            15.6
DISTRICT ATTORNEY REFUSES TO PROSECUTE     9.4
Name: Resolution, dtype: float64

STOLEN PROPERTY
ARREST, BOOKED     84.4
NONE                9.2
JUVENILE BOOKED     2.7
Name: Resolution, dtype: float64

SUICIDE
NONE                      56.6
PSYCHOPATHIC CASE         34.3
EXCEPTIONAL CLEARANCE      3.0
Name: Resolution, dtype: float64

SUSPICIOUS OCC
NONE               85.4
ARREST, BOOKED      4.1
UNFOUNDED           3.9
Name: Resolution, dtype: float64

TREA
ARREST, BOOKED    50.0
NONE              50.0
Name: Resolution, dtype: float64

TRESPASS
NONE               35.3
ARREST, BOOKED     31.3
ARREST, CITED      29.8
Name: Resolution, dtype: float64

VANDALISM
NONE               86.7
ARREST, BOOKED      7.8
ARREST, CITED       3.1
Name: Resolution, dtype: float64

VEHICLE THEFT
NONE               92.3
ARREST, BOOKED      4.3
UNFOUNDED           3.1
```

Day wise Distribution of Crimes

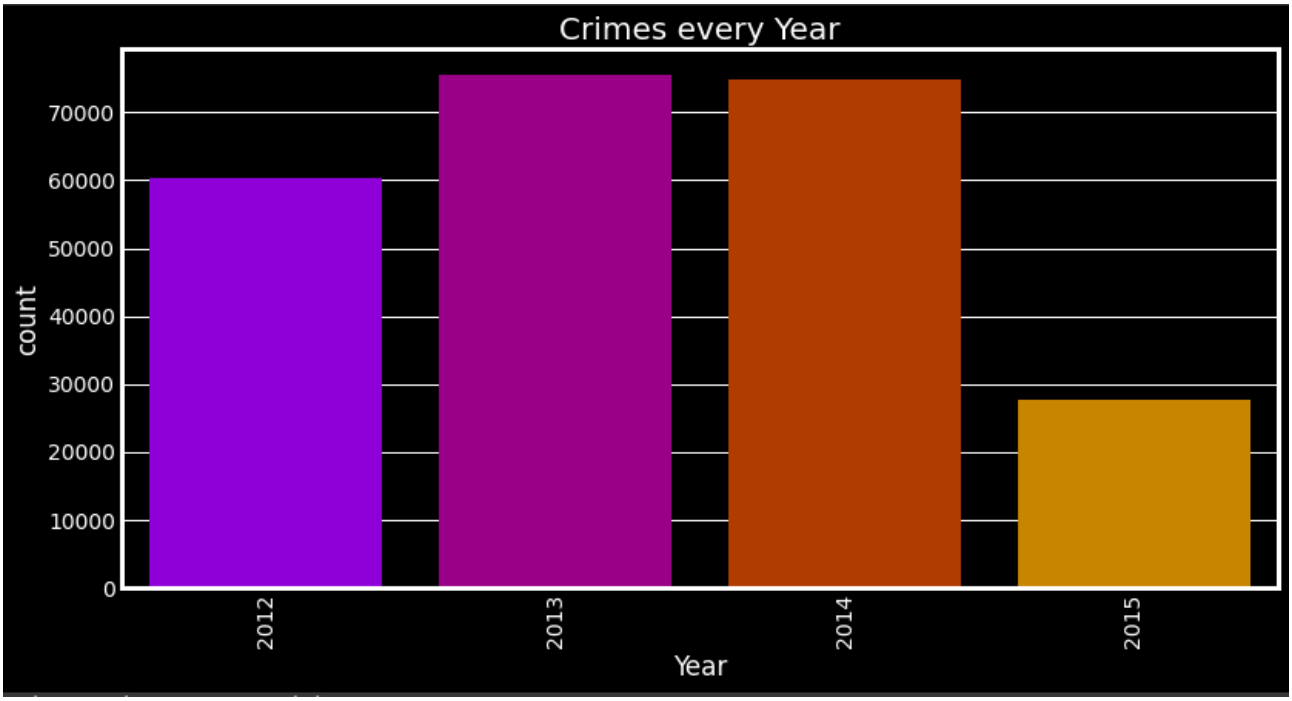
From the below figures, we can conclude that more crimes occured on Friday and compared to weekends, in Weekdays more crimes were reported.





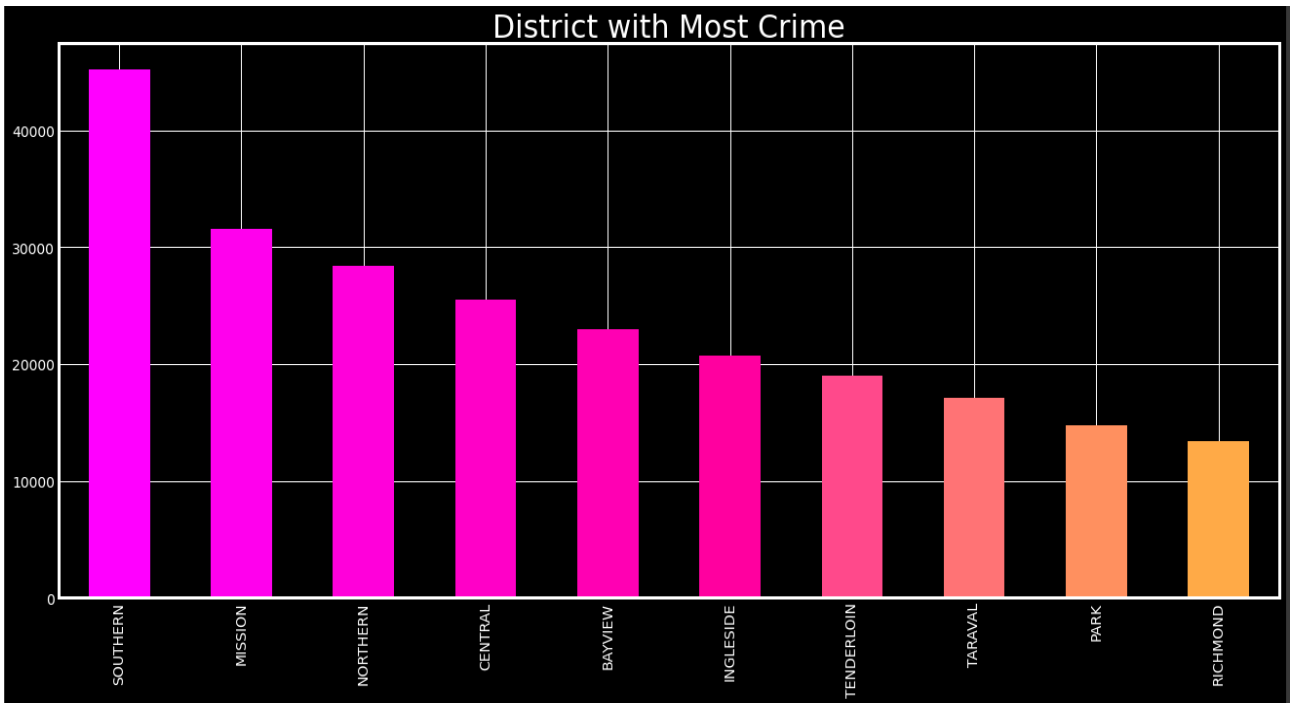
Year wise distribution

There are more crime incidents reported in 2013 and 2104 compared to other years



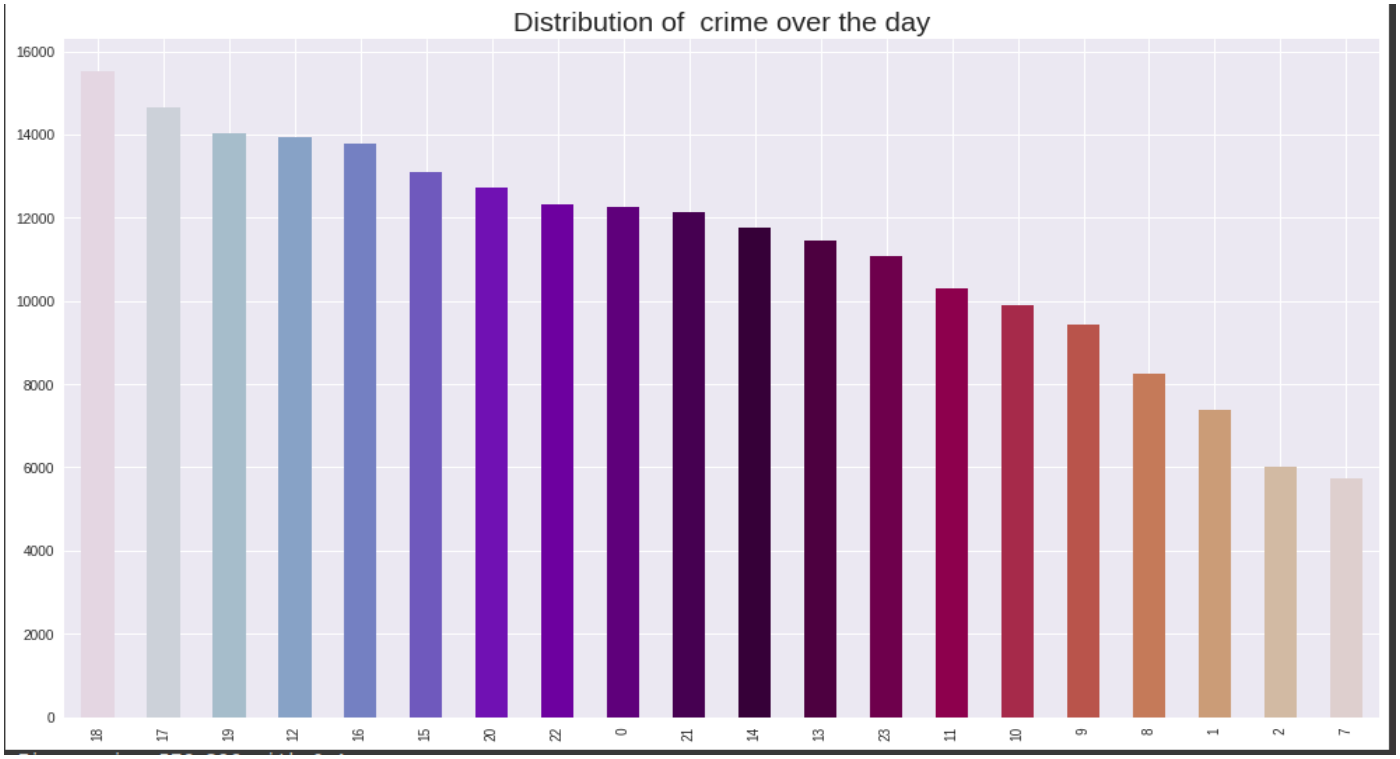
District wise Distribution:

From the below plot, the gaint hotspot is observed in the southern region and relatively less crimes in the Richmond region which can be considered as a safest place in San Francisco.

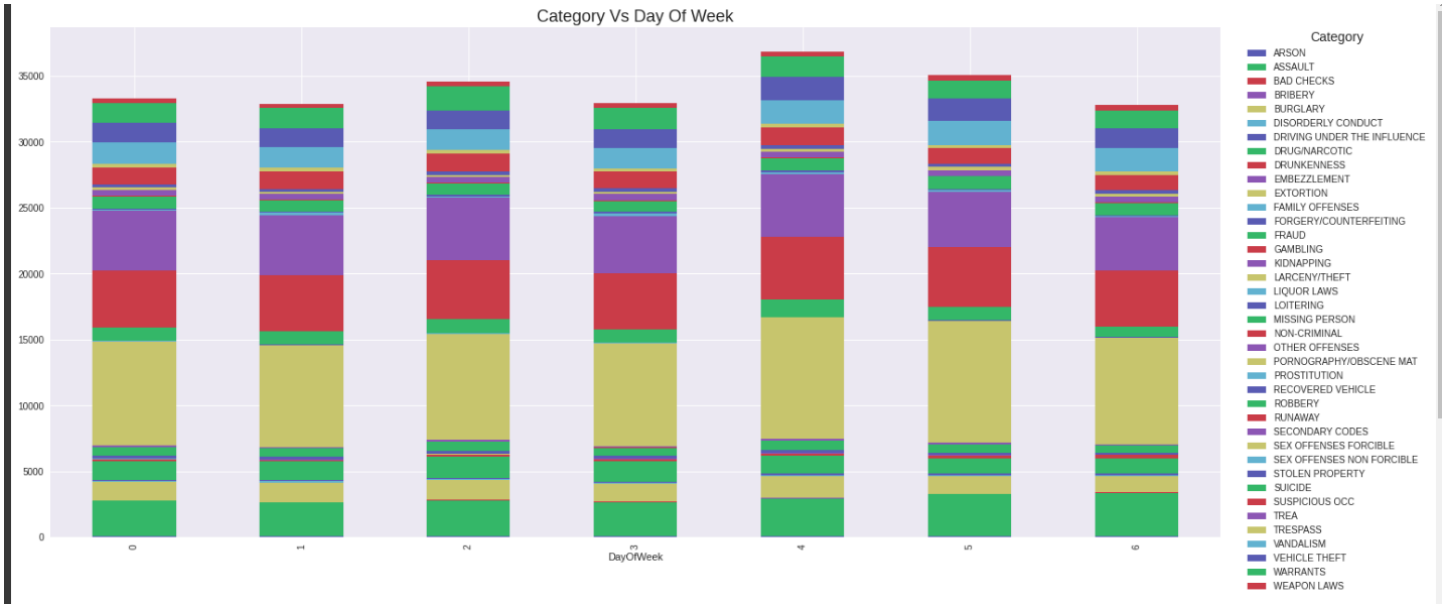


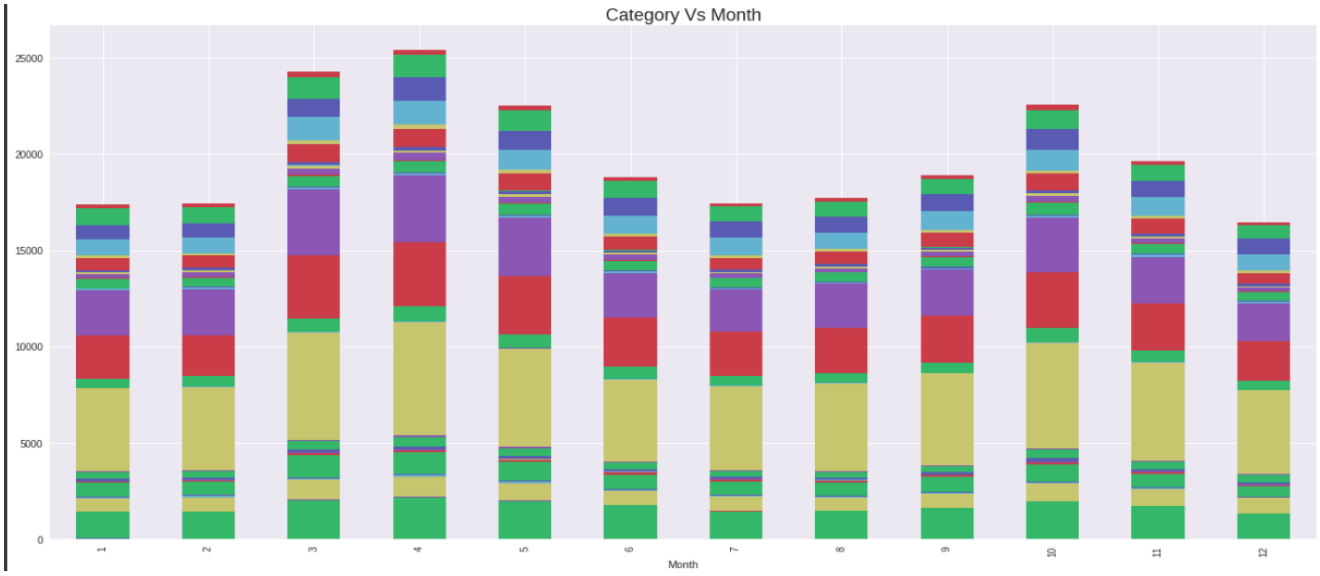
Crime rate over a day:

The below plot concludes that most of the crimes had been reported around 5 to 6 PM and between 3 to 5 AM no crimes were reported.

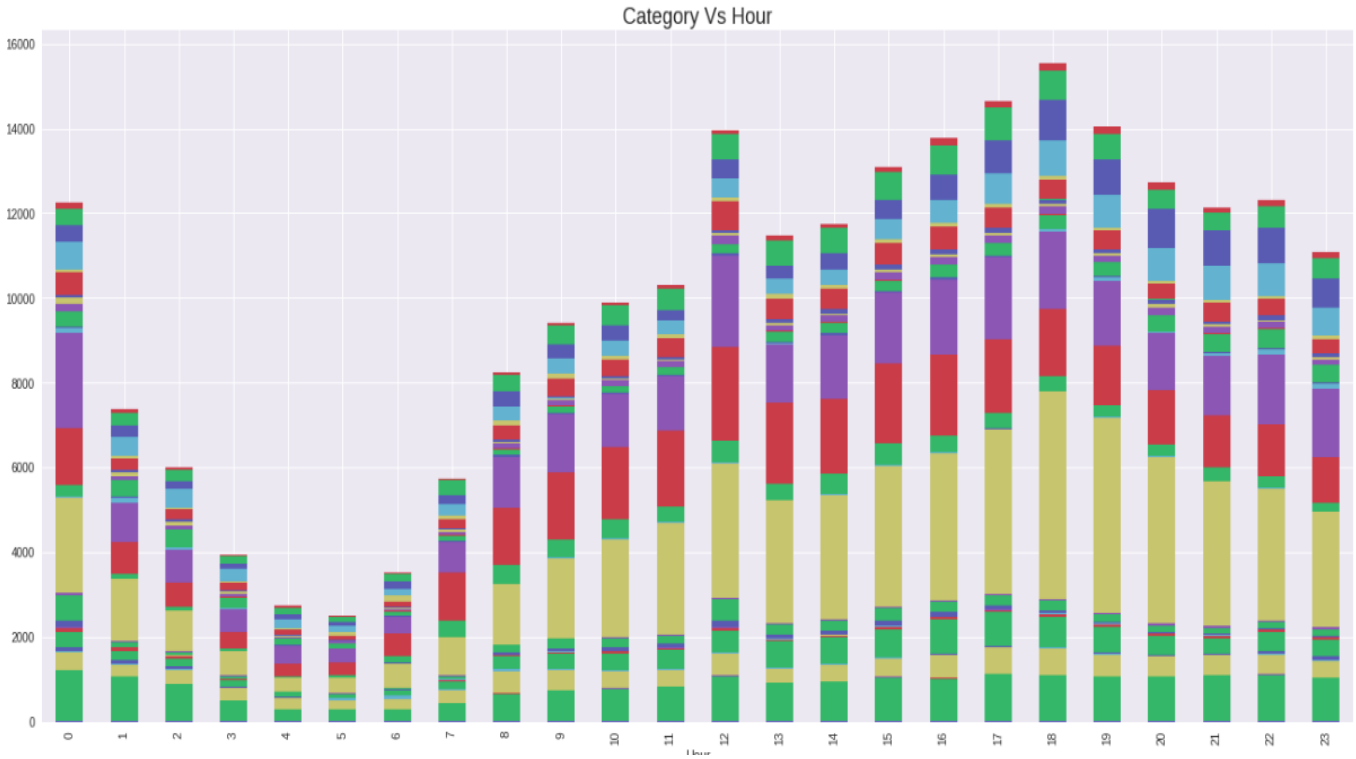


Crime Category Distribution over day of the week, Month and Date as follows



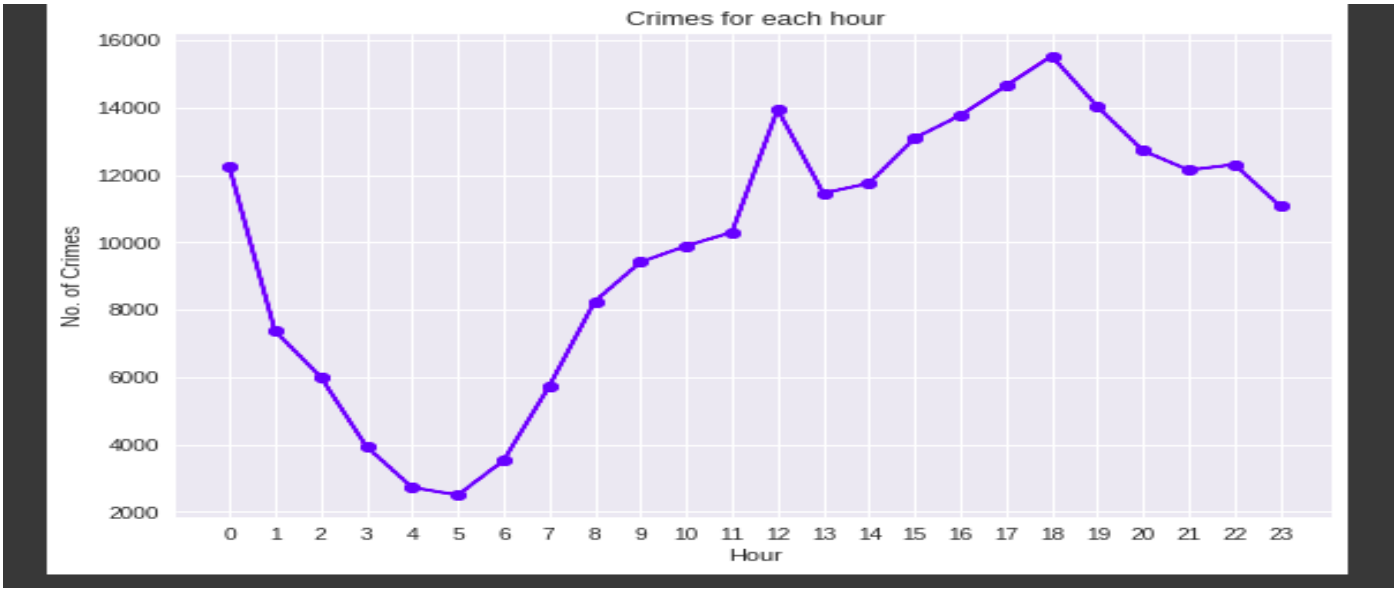


More crimes are reported in the months of March and April as per the above plot.



Hour wise Crime distibution:

From the below fig, 4 AM to 5 AM is the safest time where very less number of crimes were reported and the crimes are at the peack by 12PM , 5 and 6 PM.



Model :

As part of model development, Standardization of features been have done and trained a simple deep neural network with a combination of dense and dropout layers and a XGBoost classifier is trained to predict the crime categories and observations are recorded.

With the deep learning model, accuracy of 24.34% is achieved and with XGBoost 28.39% is achieved.