Machine Learning

# SFO Crime Classification

# Project report


**Team Name** :- William Shakespeare

**Team Members** :-

Raja Rakshith - IMT2016087

K. Viraj Bharadwaj - IMT2016093

D Sriram Siddhartha Reddy-IMT2016103


## Introduction

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz.

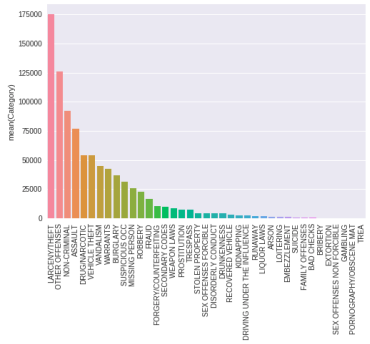
Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay.

From Sunset to SOMA, and Marina to Excelsior, this competition's dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, you must predict the category of crime that occurred.

### Analysing dataset(visualisation)

## About Category Column:

The number of crimes done in each category.

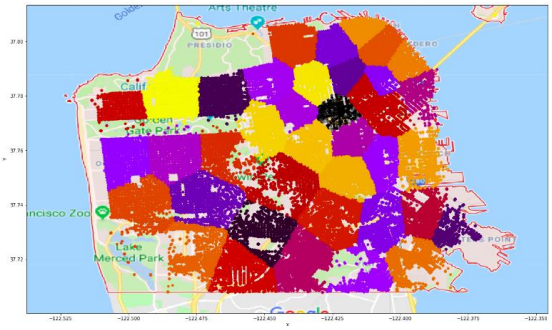



## Preprocessing:

The X and Y columns had outliers.So these were replaced by the median of respective columns

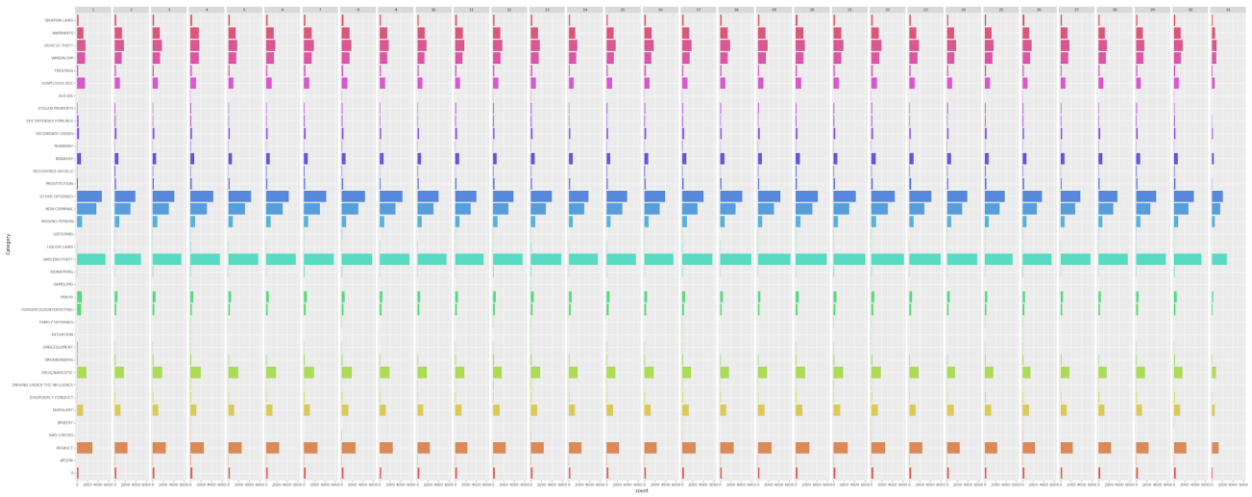
## Visualization of Different features used:

1)



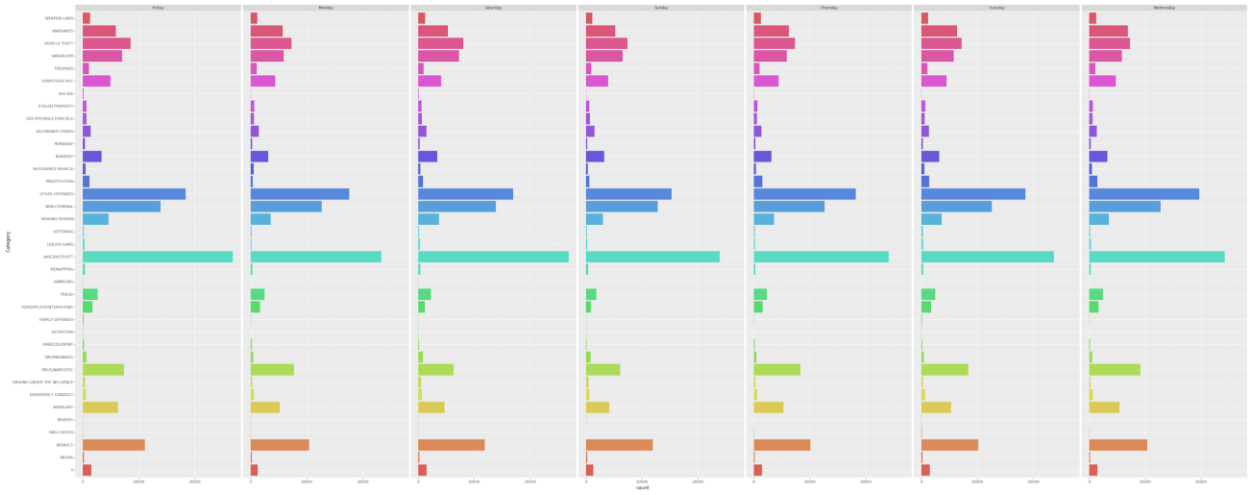

The coordinates of the crime location visualised on the map , they are clustered using k map algorithm .the number of clusters used are 40

2)



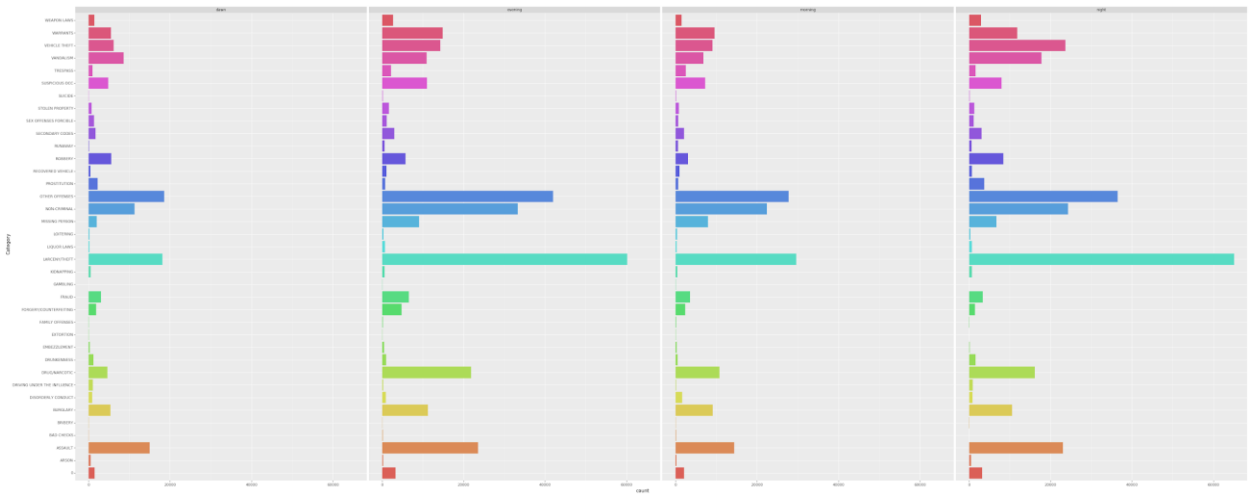
The number of crimes done on a day (1-31)of a month for each crime

3)



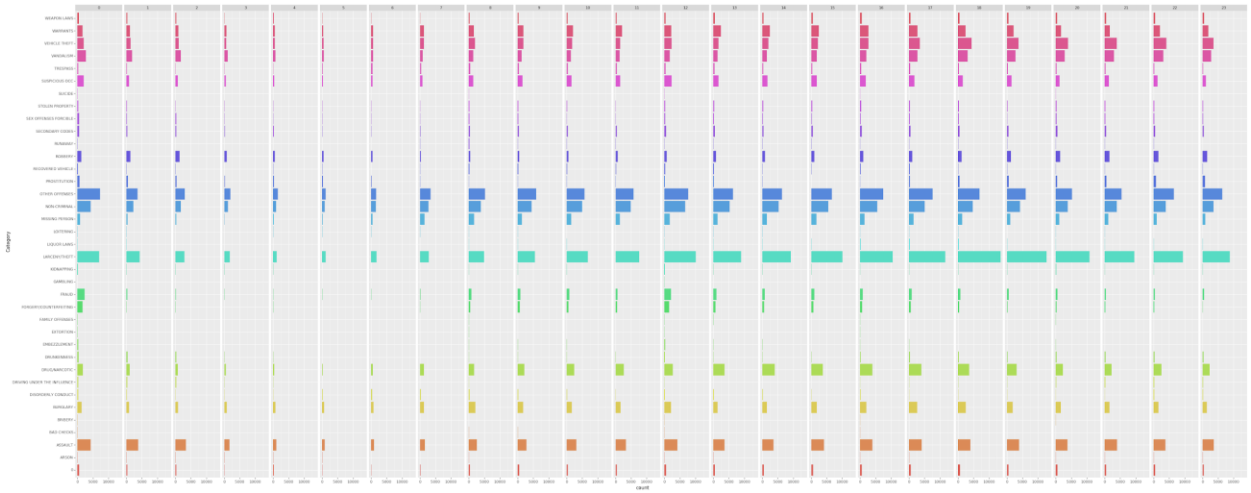
The number of crimes done on a day (1-7)of a month for each crime

4)



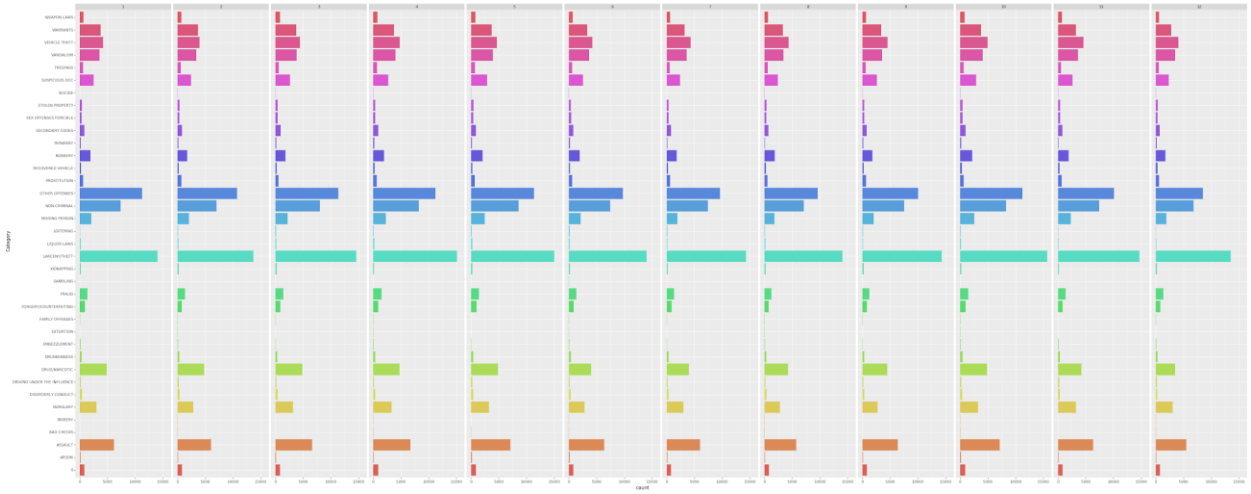
The number of crimes done on different parts of day for each crime

5)



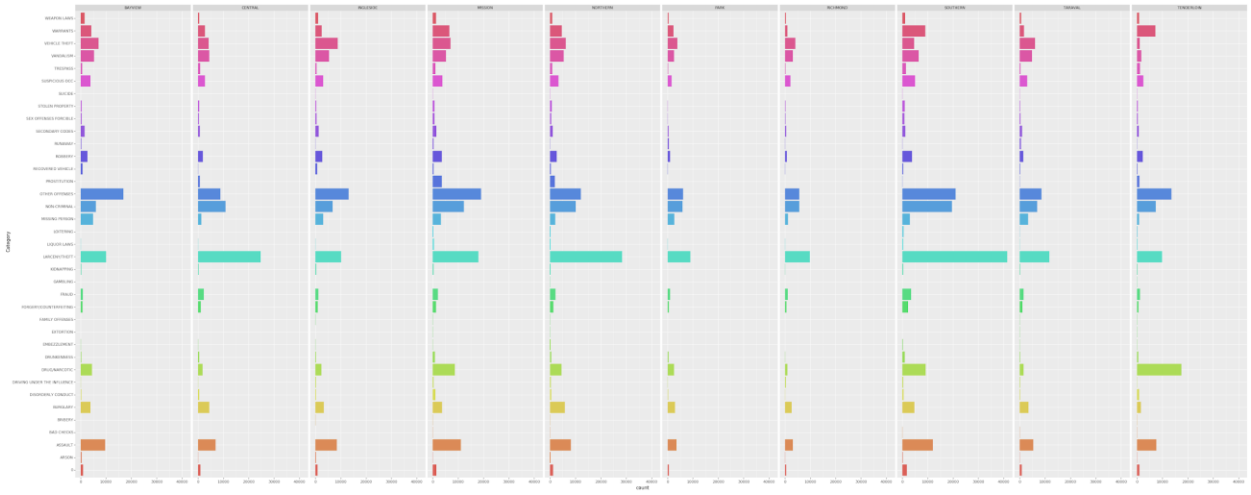
The number of crimes done on different hours (1-24) of a day for each crime.

6)



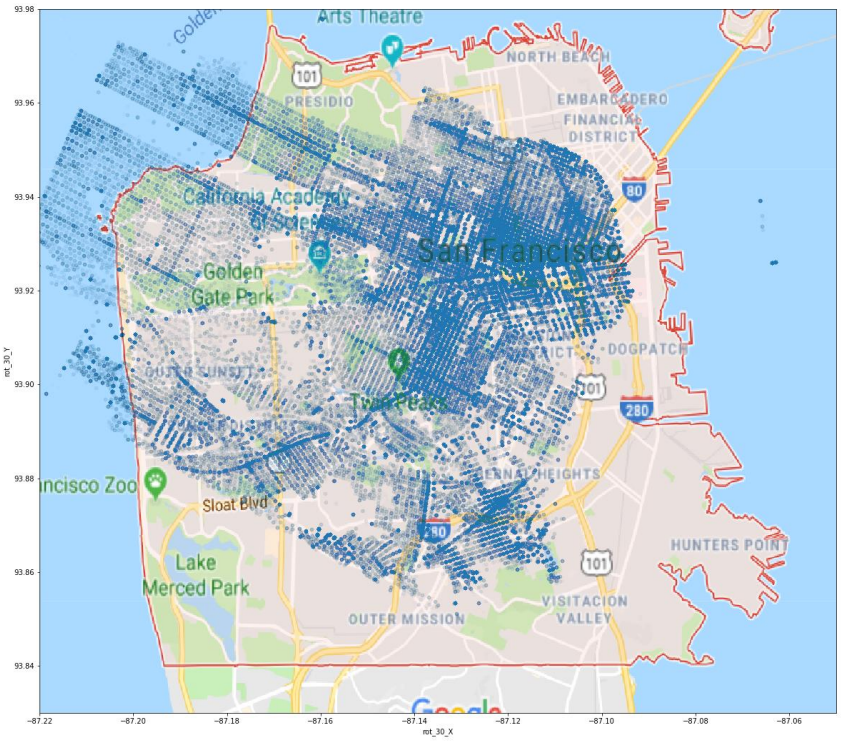
The number of crimes done on different months (jan-dec)of a year .

7)



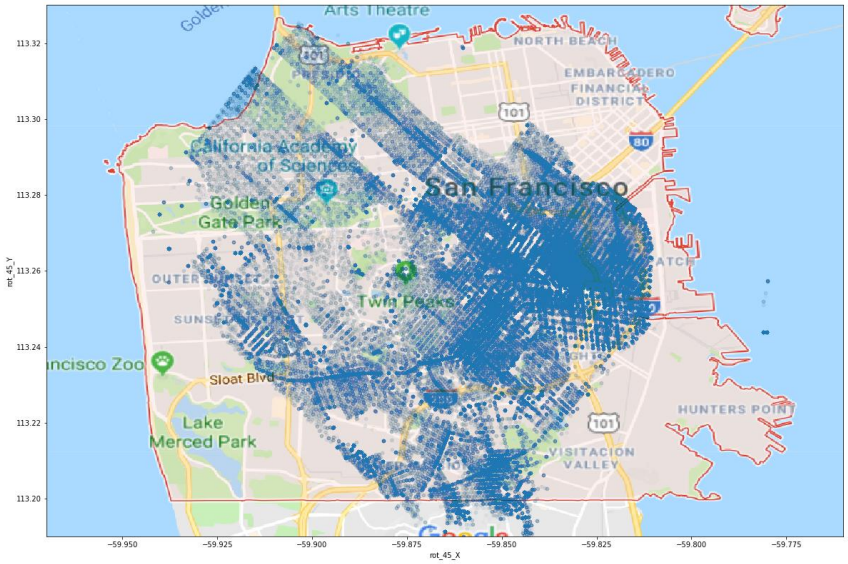
The number of crimes done in different districts per crime

8)



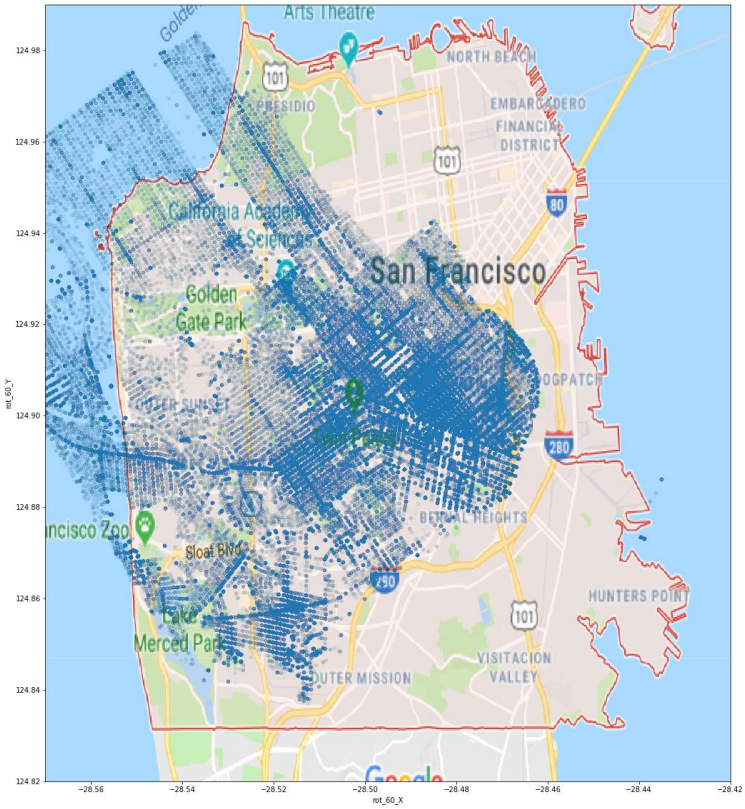

Visualisation of coordinates of all the crimes when we rotate the X and Y axes  by 30 degrees.

9)



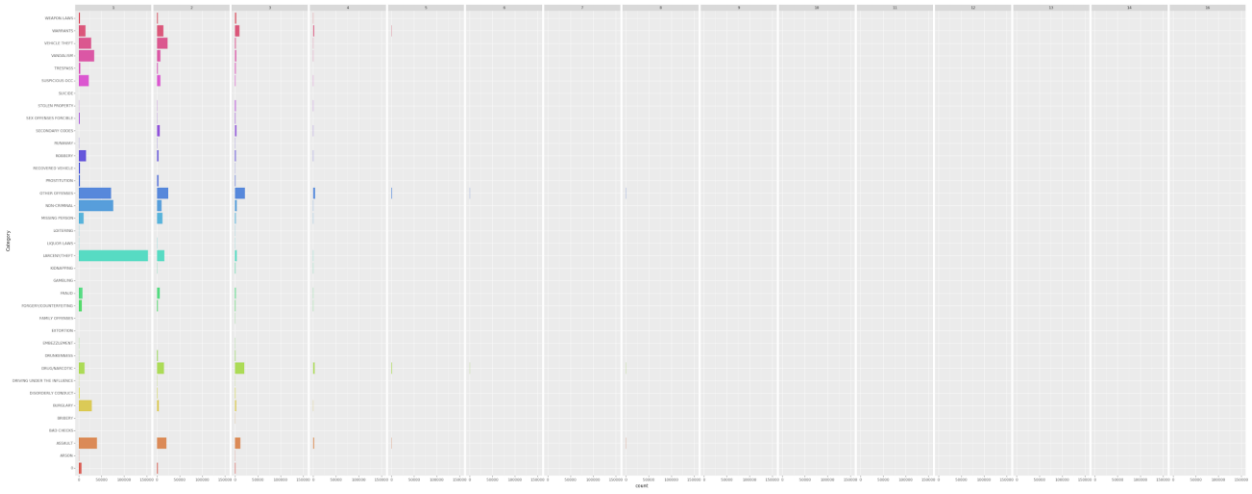

Visualisation of coordinates of all the crimes when we rotate the X and Y axes  by 45 degrees.

10)




Visualisation of coordinates of all the crimes when we rotate the X and Y axes  by 60 degrees.

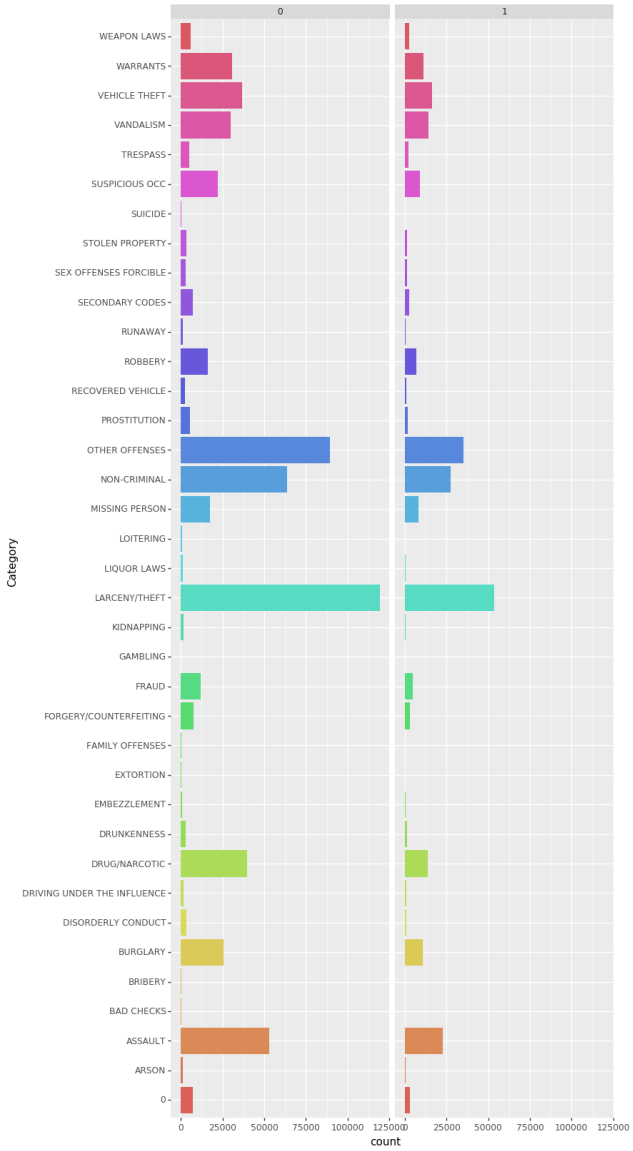
11)



1 in the graph means a crime incident is recorded as a crime involving 2 different categories of crime.Simarly 2 means a crime incident is recorded as a crime involving 3 different categories of crime and so on.

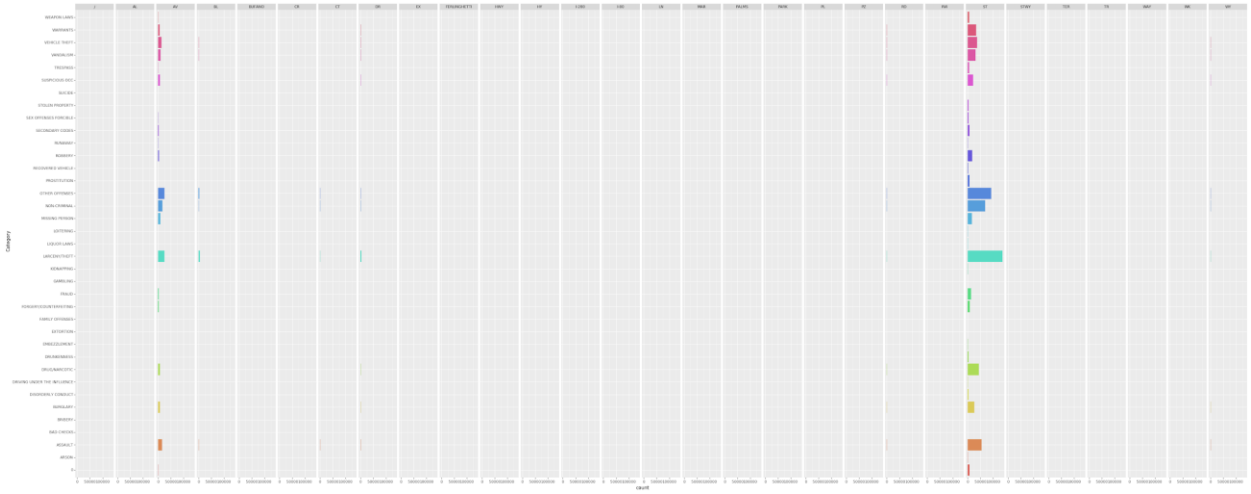
The graph indicates the count of a single category of crime appearing when a crime incident is recorded as a crime involving different categories of crimes.

12)



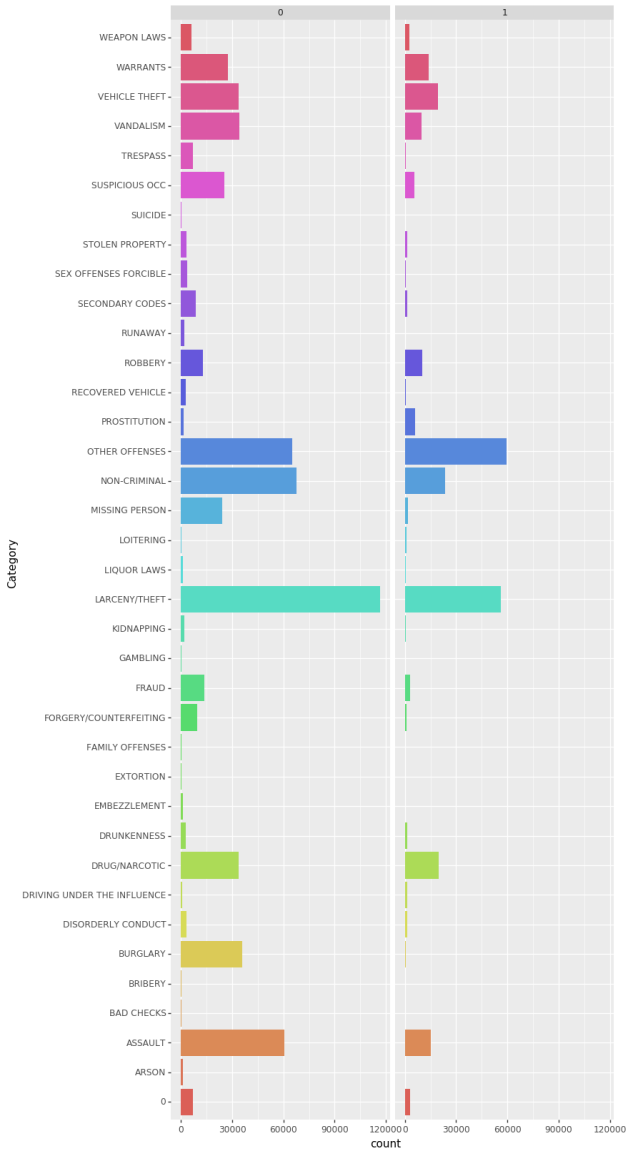

The number of crimes done on a street (0)vs street corner (1)for each crime

13)



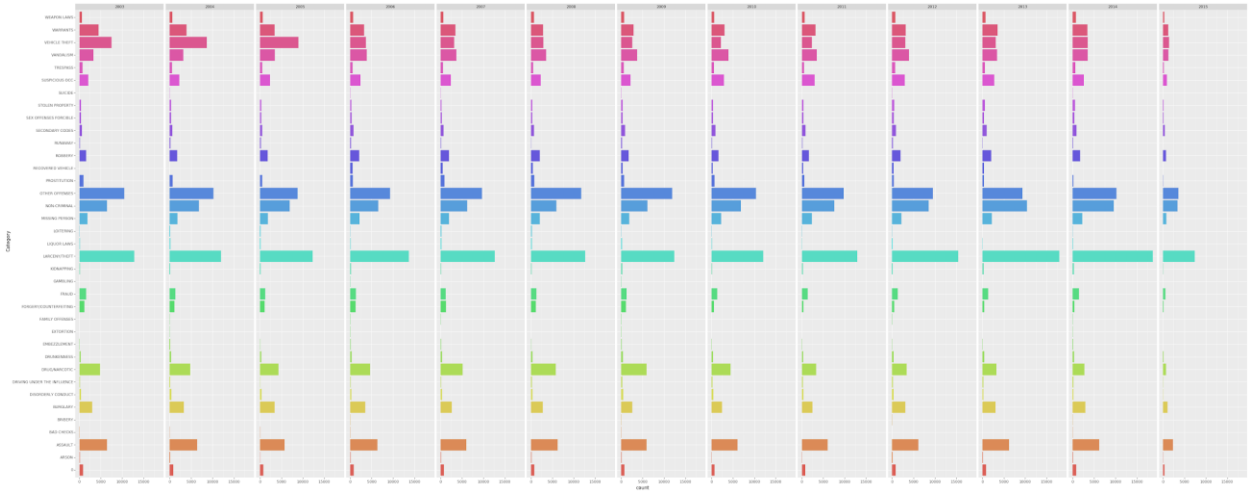
The number of crimes done on each street of a particular address.

14)




The number of crimes done on weekend(1) vs non weekend(0) for each crime

15)



The number of crimes done in every year for each crime.

## Feature engineering:

### Hour:

The crime is classified into 24 parts each for one hour.It used directly without any encoding

### Day of week:

The crimes done are classified based on the day of the weekI.e 1-7.they are used directly without any encoding.

### Day of month:

The crimes done are classified based on the day of the month i.e 1-31.they are used directly without any encoding

### Month:

The crimes done are classified based on the month of a year, i.e 1-12.One hot encoding is used for this column.

### Year:

The crimes done are classified based on year,i.e 2003-2015.One hot encoding is used for this column.

### Weekend:

If the day of the crime falls on the weekend then they are grouped together

### Street name :

Crimes are grouped by on which street the crime has been done , one hot encoding is used.

### Street corner:

Whenever there are two street names in the address column separated by "/" then it is taken as a street corner , crimes are separated on whether its a corner or not.one hot encoding is used.

### Simultaneous crimes:

Crime that occurred at the same time and address are grouped together.

### Seasons:

The year is divided into four seasons

months(1-2) and month (12)-winter

months (3-5)-spring

months (6-8)-summer

months(9-11)-autumn.

Crimes are grouped on based on this.One hot encoding is used for this column.

### Dayparts:

The 24 hr period is divided into four parts dawn(12-6) ,morning(6-12) ,evening(12-6) , night(6-12).the crimes are classified into one of the four categories using "Hour" column. One hot encoding is used for this column.

### Rot_theta_X and Rot_theta_Y:

A Rotation of theta°  has been applied to the latitude(X) and longitude(Y) columns.

Now the new coordinates in these columns are

X1=Y*sin(theta)+X*cos(theta)

Y1=Y*cos(theta)-X*sin(theta)

The different values used for theta are 30°,45°,60°.

Crimes have been classified using these columns.

### Radial_r:

The radial distance of a particular point is calculated by using the formula.

r=√X^2+Y^2 where X and Y are latitude and longitude of a particular point in the map of San Francisco.

Crimes have been classified using this column.

### Cluster:

Using Kmeans 40 different clusters are formed from  X and Y columns.

Crimes have been classified  based on this column.

One hot encoding is used for this column.

### CategoryNum:

The column to be predicted is label encoded.

## Models used:

### Naive Byes:

Eval Metric:

Validation Log Loss-2.672

### Random Forest:

Parameters:

N_estimators=300

Max_depth=25

Eval Metric

Validation Log Loss-2.29

Test Log Loss-2.31

### XgBoost:

Parameters:

Eta=0.4

num_rounds=70

Objective ='multi:softprob'

Eval Metric:

Validation log loss:2.15

Log Loss- 2.17657

XgBoost gave us the best result.

## G-drive link for pkl file

https://drive.google.com/folderview?id=1hnh0tP6ZOH0EOhGIe2iJmcgKmtUJq8Cr