# San-Francisco Venues Crime Analytics

## 1.Introduction

### 1.1 Background

San Francisco is the most densely populated and large city in the state of California. San Francisco is one of the 20 fastest growing cities in the United States. It's estimated that thousands of new residents will call San Francisco home by 2040, and the city Planning Director said that the city would need more than 92,000 more housing units and 191,000 new jobs to accommodate this growth. The Bay Area as a whole will need 1.1 million additional jobs and 660,000 new housing units to provide for an estimated 2.1 million more people who will move to the city by 2040.
Source: worldpopulationreview.com/us-cities/san-francisco-population

### 1.2 Problem

San Francisco is the nation's leader in property crime. Burglary, larceny, shoplifting, and vandalism are included under this ugly umbrella. The rate of car break-ins is particularly striking: in 2017 over 30,000 reports were filed, and the current average is 51 per day. Other low-level offenses, including drug dealing, street harassment, encampments, indecent exposure, public intoxication, simple assault, and disorderly conduct are also rampant.
Source : https://www.city-journal.org/san-francisco-crime

### 1.3 Interest

SFPD would be interested to see analytics around the number of incidents by day and hour, so that it can estimate the resource required for controlling crime and ensure economic growth of its commercial areas

## 2. Data acquisition and cleaning

### 2.1 Data Sources

For data relating for number of crimes, SFPD website has the data that can be downloaded in CSV format. The file has 26 Columns. But for this exercise, we took Incident Day of the Week, Incident Time, Incident Category, Analysis Neighbourhood along with its Latitude and Longitude.

For data relating to Venues, FourSquare developer access was used to pull information on Neighbourhood, Neighbourhood Latitude & Longitude, Venue, Venue Latitude and Longitude along with Venue Category.

### 2.2 Data Cleaning.

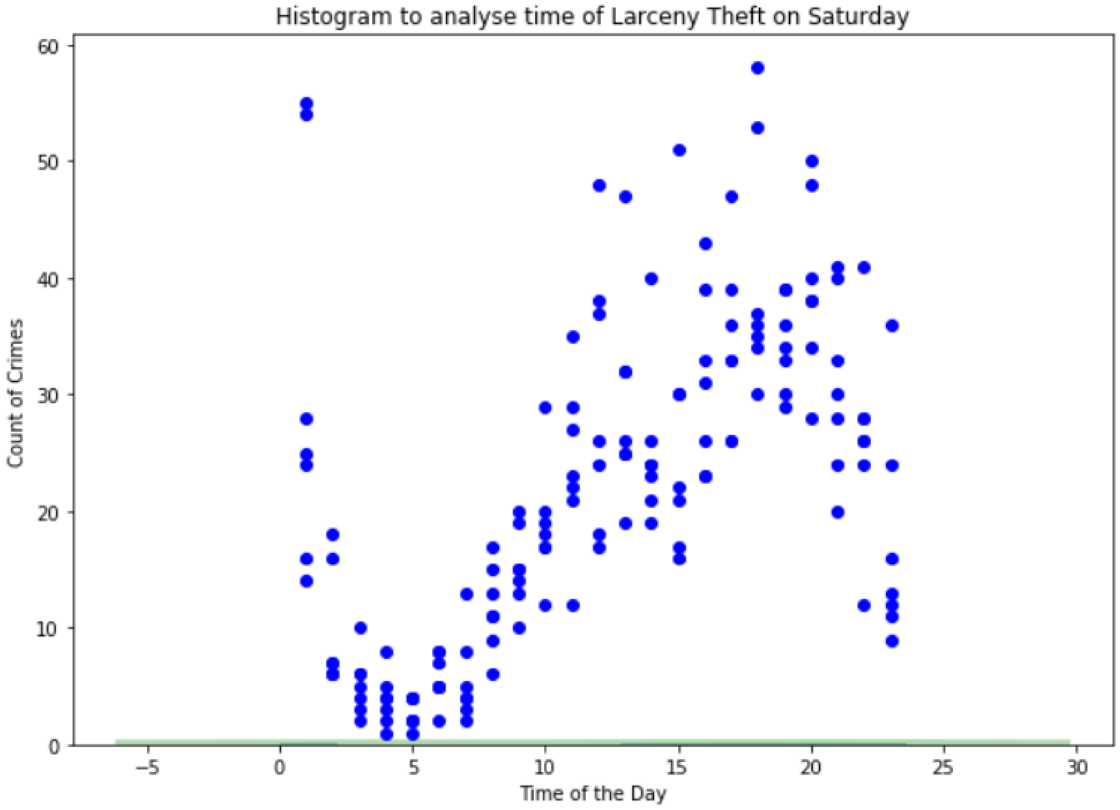
Due to sheer volume of records and for computational speed, I had to limit the incident of Crimes for the month of June 2019. There weren't any issues with data quality with the data downloaded from SFPD website, but it did involve in Data Wrangling to come out with meaningful insights and visualisation. Time had to binned for easy analysis along with converting the format of Time stamp and converting Week day into Categorical numerical group

Similarly, beautiful soup was used to retrieve data from Four Square using web scraping but the data quality was good and needed data wrangling for insights and visualisation

### 2.3 Feature Selection

Primary focus on answering the problem statement of optimising the deployment of SFPD resource, the features selected for analysis was Type of Crime, Neighbourhood, Time and Day of the Crime, Latitude and Longitude of Crime.

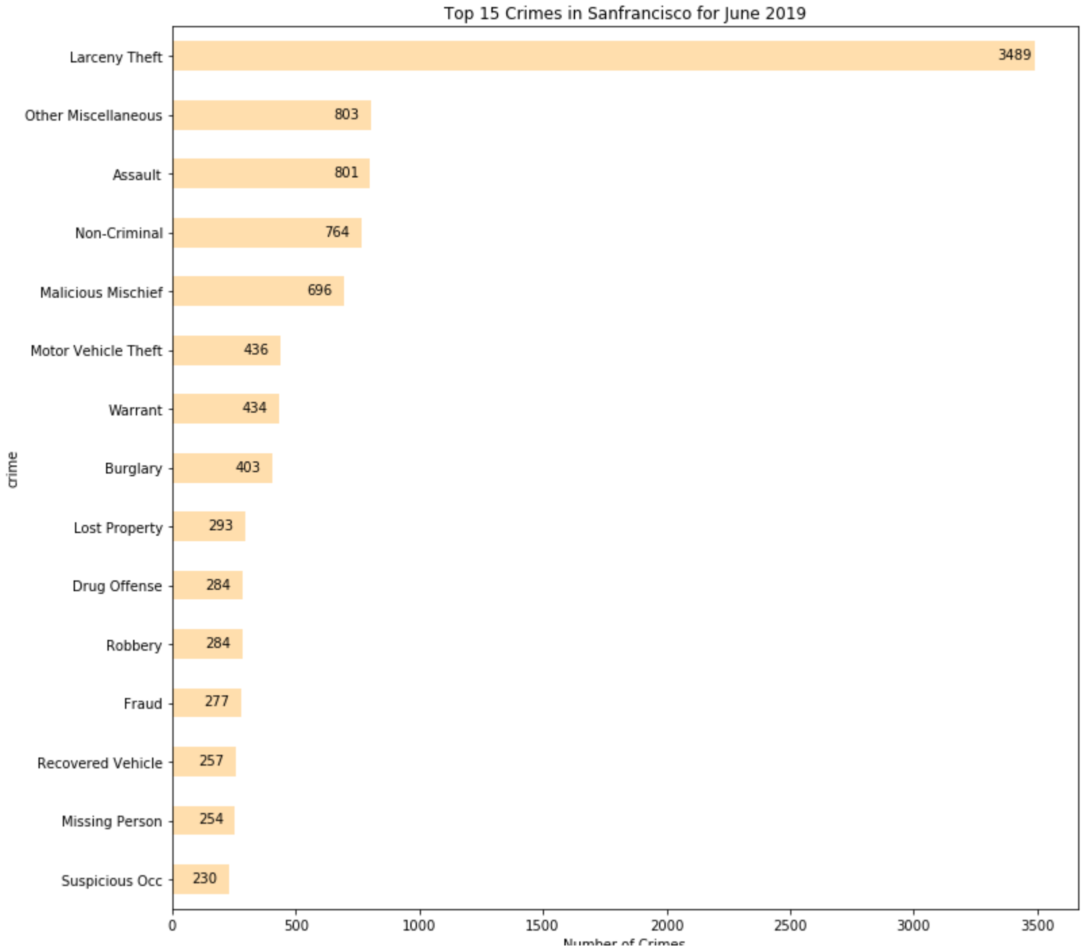
I created a scatter plot to analyse the relationship



## 3. Exploratory Data Analysis

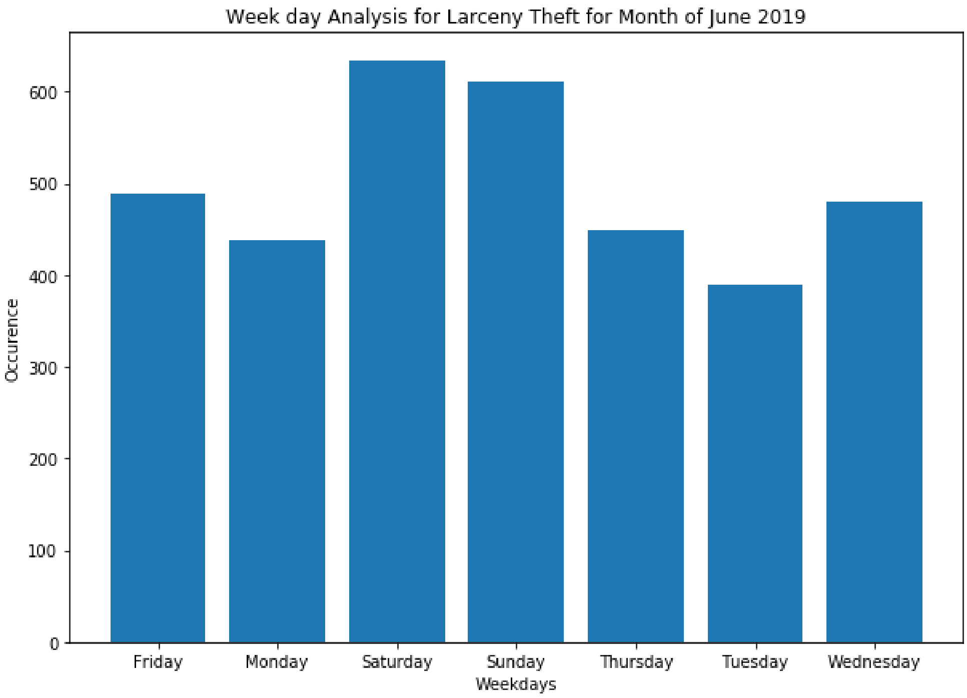
The first of the analysis involved looking at the highest type of the Crime in the San Francisco. Since the Category had more than 30 types of Crimes, I limited the analysis to the top 15 based on the count and in relation to the Venues

### 3.1 Top 15 Crimes in San-Francisco



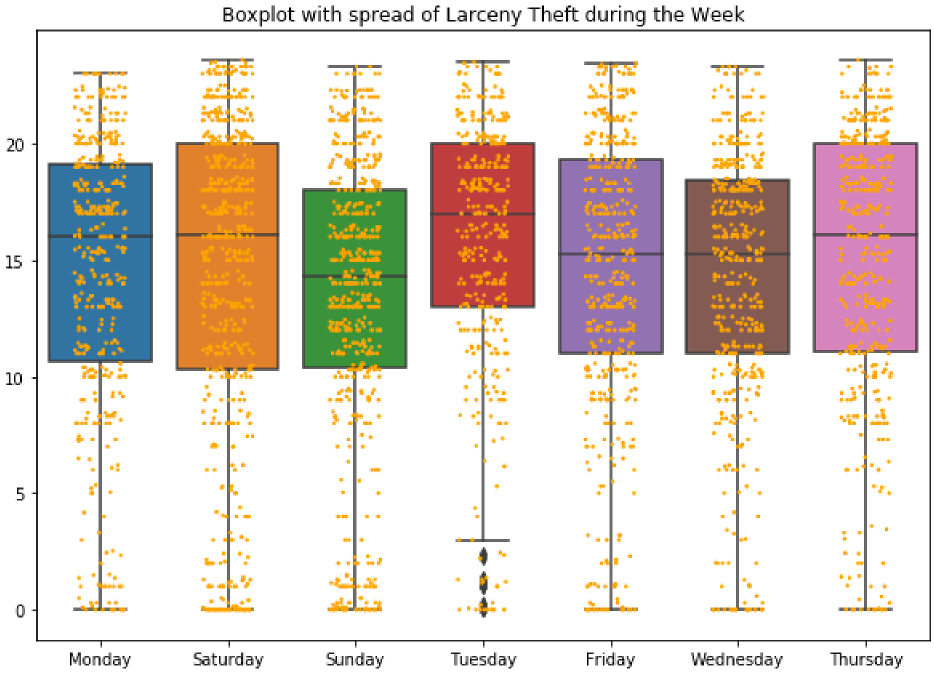
**3.2 Larceny Crime by Weekday**

Analysis shows that Larceny theft is the highest count of crimes for the month of June 2019 at 3489 count. There are other types of crime which are of similar nature like burglary which is breaking into a house and Robbery which uses force for theft but the legal definition is different for each of these crimes Considering the magnitude of crime belonging to Larceny Theft, we will focus our analysis on Larceny for this exercise taking into account Weekday and time



The Barchart shows higher larceny counts for Saturday and Sunday compared to other week days This aligns with our assumption of higher crowds in venues during holidays will lead to higher counts of crimes Lets plot this against the box plot and visualise the data to get a fair idea of the spread for each days of the week

### 3.3 Larceny Crime during the time and day of the Week



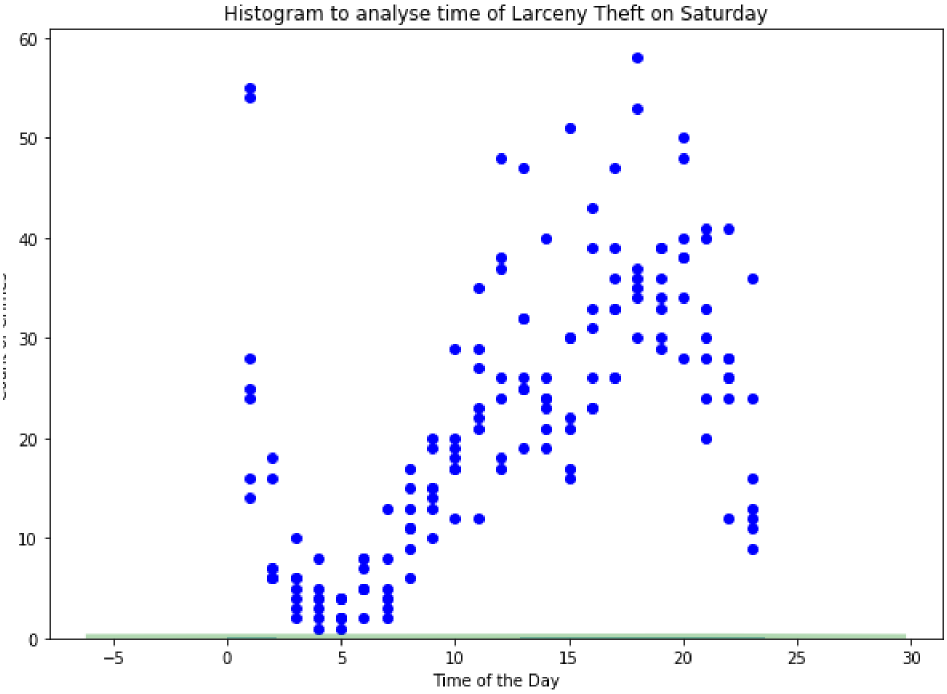
The visualisation shows the count of crimes increases post Afternoon till late midnight and then shows a decreasing trend that lasts from early morning to early afternoon.

The Boxplot also shows the typical time zones of the crimes happening during different weekdays, notably Sunday which shows crime increasing in morning but decreasing by night as compared to other weekdays, indicating of the fact that people retire Sunday evening to get ready for Monday work.

On the other hand Saturday the counts starts increasing in morning but continues till late night, indicating of the fact that crowds tend to hang around in venues during weekends celebrating till late night

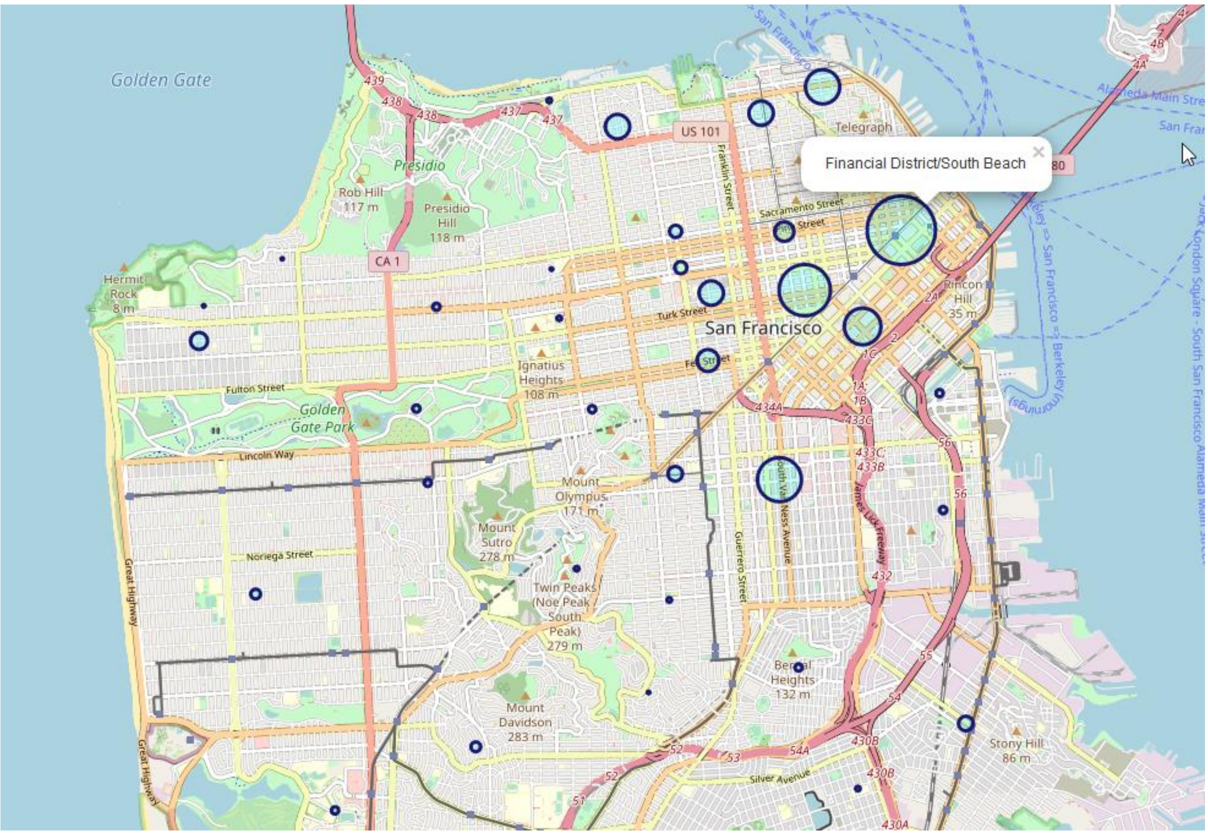
I will proceed with Plotting histogram for to analyse the theft occurrence on Saturday across various times

### 3.3 Larceny Crime during the time for Peak time Saturday



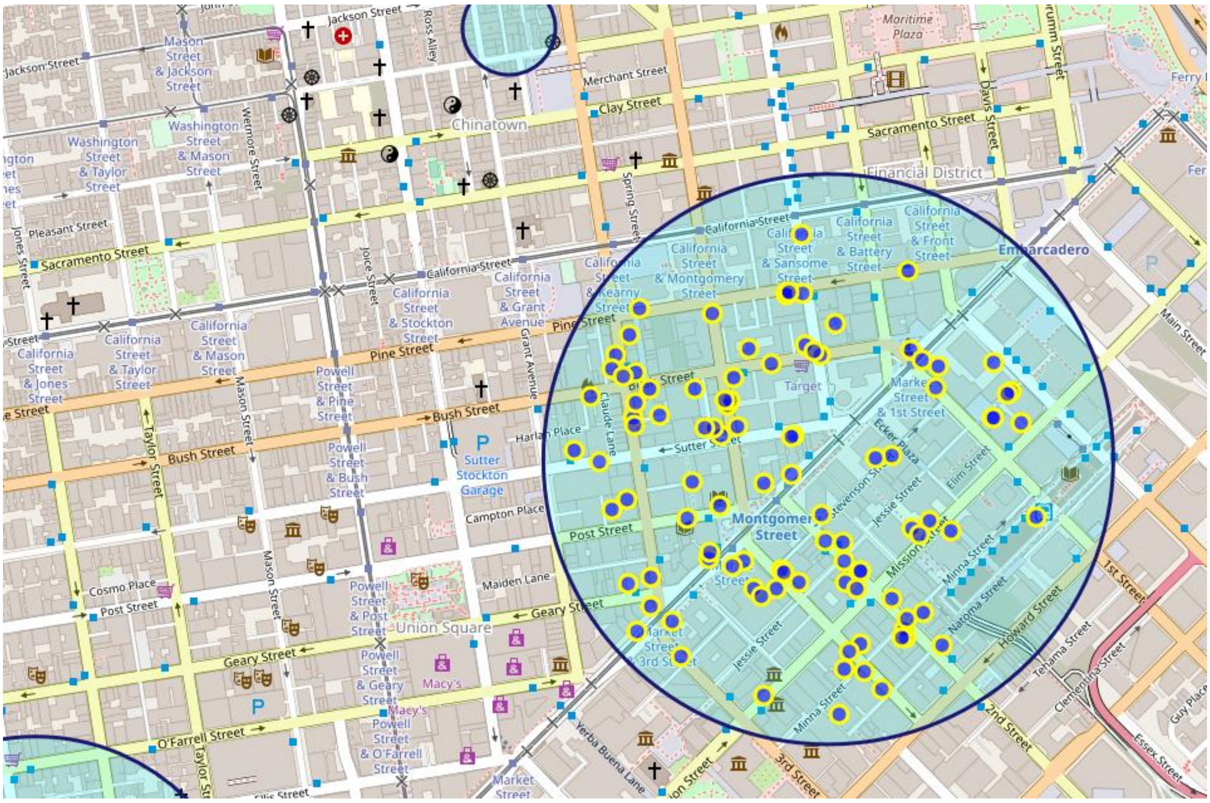
Interestingly the scatter plot shows the trend for Saturday which starts with crowd moving into Venues by Morning and the peak reaches at Late evening to Midnight

### 3.4 Visualisation of the location with higher counts of Larceny Crime



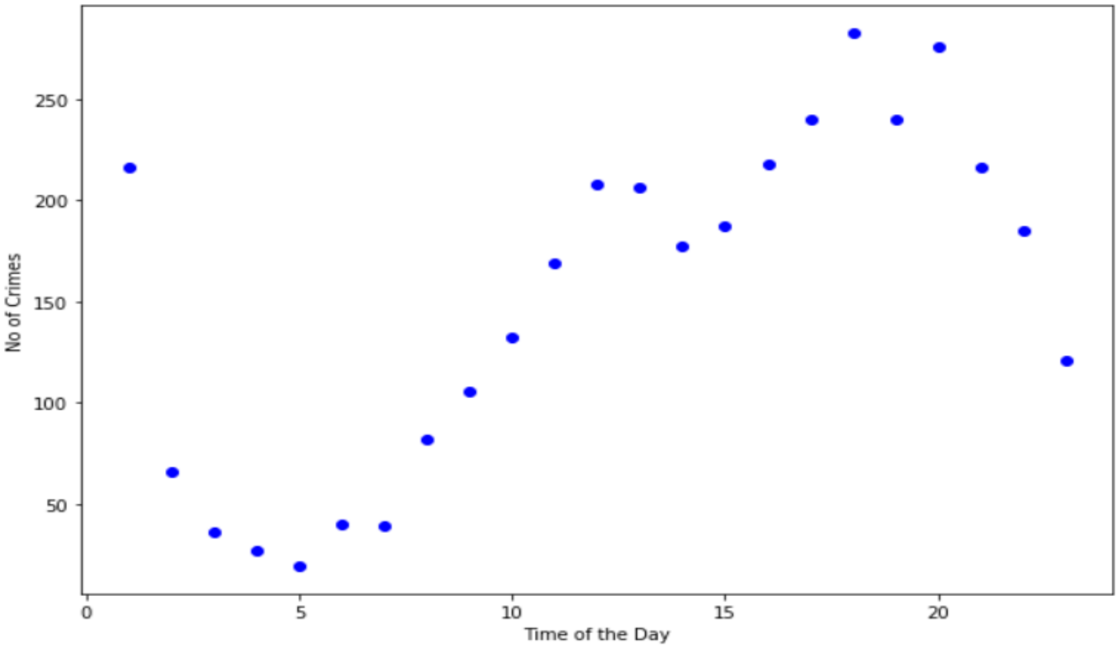
Using Folium map its now easy to check the area with highest count of Larceny Crime, the top being Financial District/South Beach.

### 3.4 Visualisation of spread of Venues in Financial District/South Beach



## 4. Predictive Modelling

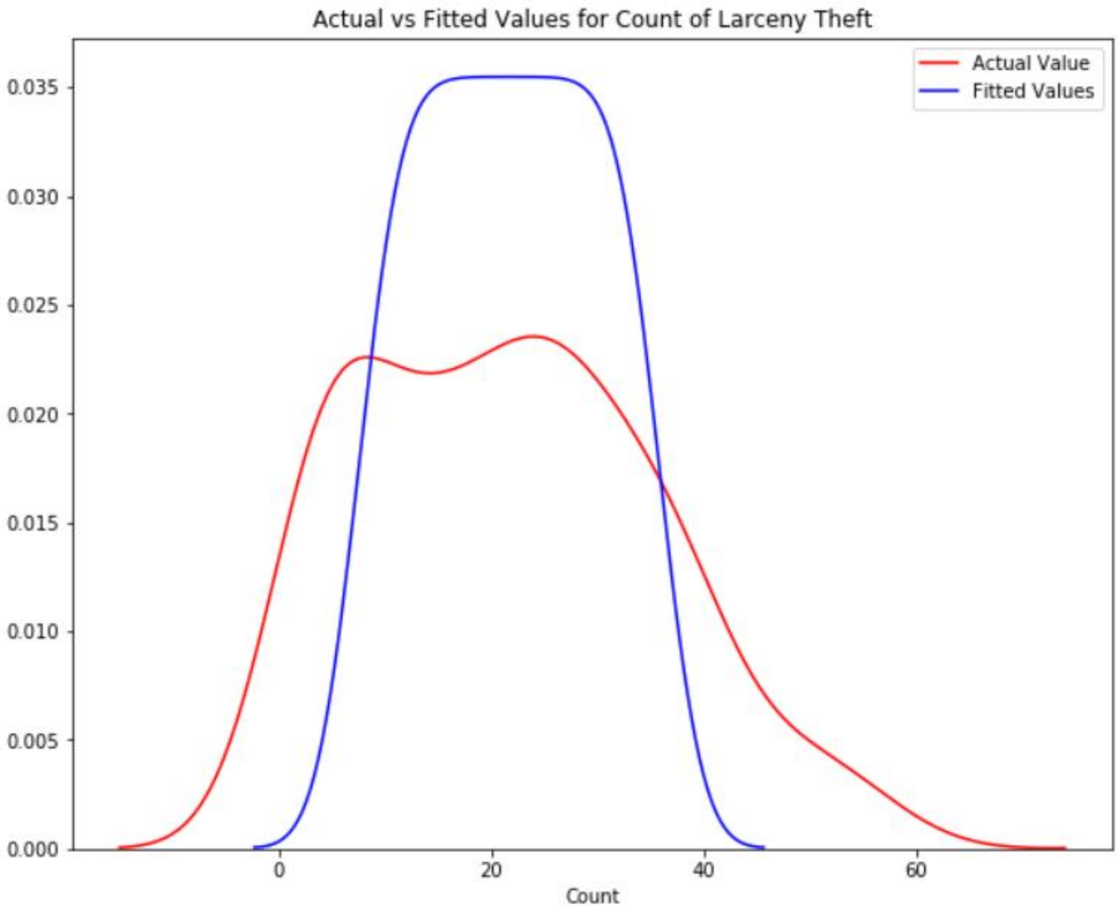
The scatter plot is showing a nonlinear relationship, I will refine this further by summing up all the crimes for each day for different binned time..



### 4.1 Linear Regression

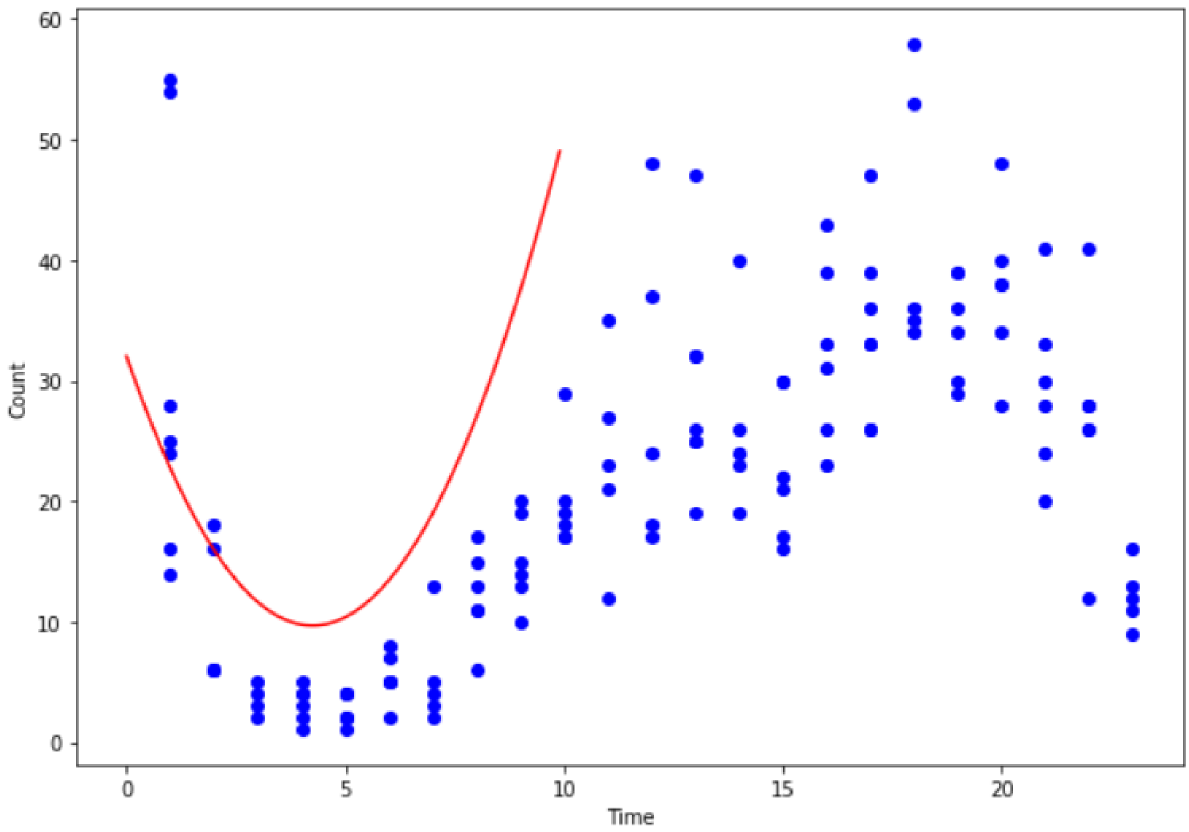
I then proceeded to fit a simple linear regression to see how the model performed with the following results
Intercept is: 5.55
Coefficient is:[0.35403727 1.22543761]
R2-score: 0.35

The below diagram shows the fitment clearly showing that simple liner regression is not suitable for the data



### 4.2 Polynomial Regression with Degree 3

I then proceeded with polynomial regression with degree 3 and below are the results
Intercept: [32.02728127]
Mean absolute error: 6.65
Residual sum of squares (MSE): 76.12
R2-score: 0.31



My analysis is that Non Linear regression model should be used to build a model with better accuracy and fitment considering the volatility of the crimes that occurs during different time and day of the week.

## 5. Conclusion

In this exercise, I studied the relationship of Weekday, Time and mainly one particular Crime "Larceny Theft", as this was the highest recorded crime in San-Francisco. There is a trend that emerges showing higher crime count on Saturday and Sunday and the spread of time for the crime, mainly targeting towards late evening to early morning on Weekends.

The relationship is non linear and I was not able to get a get prediction model using simple linear model or polynomial models, this means that predictive model should be build using Non Linear Modelling techniques using a Cubic Function

### 6. Future Direction

Future direction lies in creating a model with non linear technique that will enable SFPD to predict the number of crimes that can happen based on history for Week day and time of the day that will enable them to plan resources. Furthering the capability will be to use Folium visualisation to analyse the areas most effected by the crime and the venues around the area. Clustering approach can be used to find relationship between type of venue and number of crimes so that SFPD can deploy resources at these points as preventive measures