Capstone Project: The Battle of Neighborhoods

# San Francisco Police Department Incident Reports: 2018 to Present

### [*Author: Jithin Prakash Kolamkolly*](https://in.linkedin.com/in/jithinprakashk)

Contents

[San Francisco Police Department Incident Reports: 2018 to Present 1](#_Toc42220871)

[*Author: Jithin Prakash Kolamkolly* 1](#_Toc42220872)

[Summary 2](#_Toc42220873)

[Introduction/Business Problem 2](#_Toc42220874)

[Data Description 2](#_Toc42220875)

[Methodology 3](#_Toc42220876)

[1. Collect Data 3](#_Toc42220877)

[2. Explore and Understand Data 3](#_Toc42220878)

[3. Data Preparation and Pre-processing 5](#_Toc42220879)

[4. Modelling 6](#_Toc42220880)

[**K-Nearest Neighbour** 6](#_Toc42220881)

[**Decision Tree** 7](#_Toc42220882)

[**Logistic Regression** 7](#_Toc42220883)

[5. Evaluation and Testing 7](#_Toc42220884)

[Results 8](#_Toc42220885)

[**Crimes per Category** 8](#_Toc42220886)

[**Time Crime Incident:** 9](#_Toc42220887)

[**Crime per Police District:** 11](#_Toc42220888)

[**Word-Cloud** 12](#_Toc42220889)

[Geo-Location of crime incidents 12](#_Toc42220890)

[Using Foursquare to visualize and infer crime incidents and locality 13](#_Toc42220891)

[Conclusion 14](#_Toc42220892)

[Link to the python code 15](#_Toc42220893)

# Summary

The dataset includes police incident reports filed by officers and by individuals through self-service online reporting for non-emergency cases. Reports included are those for incidents that occurred starting January 1, 2018 onward and have been approved by a supervising officer.

Incident reports filed by officers must be approved by a supervising officer. Once approved and electronically signed by a Sergeant or Lieutenant, no further information can be added to the initial report. A supplemental report for additional information or clarification will be generated if necessary. This means that an individual status will not change on an initial report but may be updated later through a supplemental report. Differentiating among report types can be done using the “Report Type Code” and “Report Type Description” fields.

Incident reports filed online will also be reviewed by a supervising officer. Once approved and electronically signed by a Sergeant or Lieutenant, no further information can be added to the initial report. A supplemental report for additional information or clarification will be generated if necessary. This means that an individual status will not change on an initial report but may be updated later through a supplemental report. You can filter those reports using “Filed Online” as well as the report type fields mentioned above.

Reports can be removed from the dataset in compliance with court orders sealing records as well as for administrative reasons like an active internal affair – administrative and/or criminal investigation.

This data can be used to identify the vulnerable locations in the form of clusters while buying a house or relocating to the area. The same data can be used by the police department to determine the high-risk area and to manage and control the crime scenarios

# Introduction/Business Problem

With our research we hope to find answers to the following questions.

* Does a criminal data base that contains geographical location & basic details of the criminal activity have enough indicators to predict a type of crime?
* Given just a geographic location and time, how accurately can we classify the crime?
* Explore different techniques to improve the results.

# Data Description

In this section I will describe the data that will be used to analyse the police records to find the vulnerable area which can be used to predict the best paces and neighbourhood for reducing the criminal activities within San Francisco. The data I have found is collected from ‘The office of the chief Data Officer – City and County of San Francisco’ (https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783). The Polices Department has developed a report of incidents

In order to develop a sufficient prediction system, we will consider the following fields for the analysis:

* Incident Date
* Incident Category
* Incident Subcategory
* Incident Description
* Resolution
* Police District
* Analysis Neighbourhood
* Latitude and Longitude

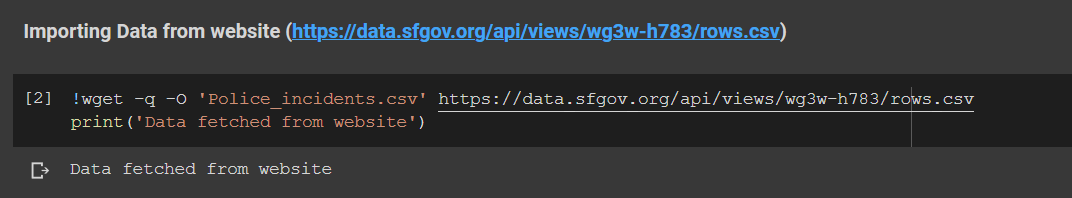
# Methodology

In this part of the report we are going to describe the main components of our analysis and predication system. Our methodology consists of 5 components as shown in figure 1.

Figure 1 – Components of Methodology (CRISP-DM)

## Collect Data

Data is downloaded directly from ‘The office of the chief Data Officer – City and County of San Francisco’ website as follows

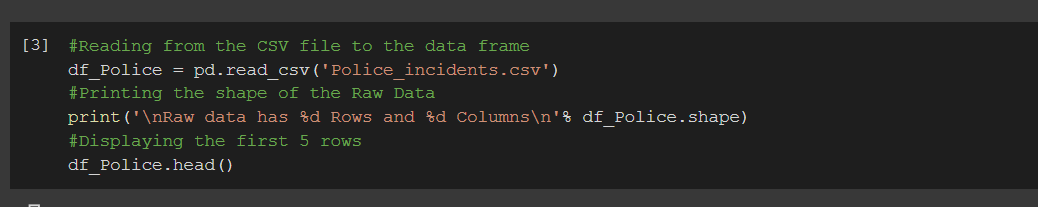


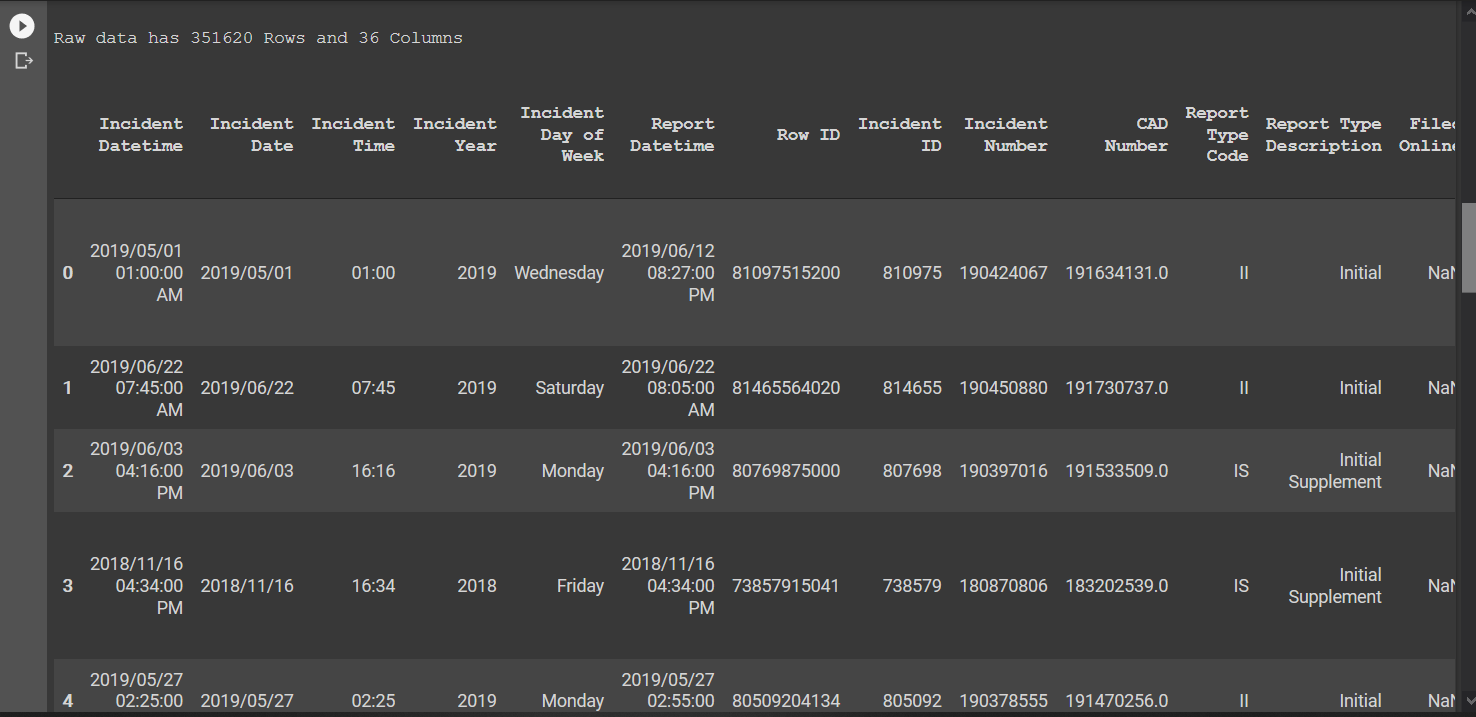
The collected data is raw data which need to be analyzed after proper exploration of data and understanding. After which data need to be organized and standardized.

## Explore and Understand Data

Data is downloaded directly from ‘The office of the chief Data Officer – City and County of San Francisco’ website using the link (<https://data.sfgov.org/api/views/wg3w-h783/rows.csv>)

First five line items are displayed here to get an idea about the data frame and its contents.





Data set consists of more than 350k rows (incidents) and 36 columns (features or attributes). The below table gives an idea about the features/attributes:

|  |  |  |
| --- | --- | --- |
| # | Feature Name | Feature Description |
| 1 | Incident Datetime | The date and time when the incident occurred |
| 2 | Incident Date | The date the incident occurred |
| 3 | Incident Time | The time the incident occurred |
| 4 | Incident Year | The year the incident occurred, provided as a convenience for filtering |
| 5 | Incident Day of Week | The day of week the incident occurred |
| 6 | Report Datetime | Distinct from Incident Datetime, Report Datetime is when the report was filed. |
| 7 | Row ID | An identifier unique to the dataset |
| 8 | Incident ID | This is the system generated identifier for incident reports. |
| 9 | Incident Number | The number issued on the report, sometimes interchangeably referred to as the Case Number |
| 10 | CAD Number | The Computer Aided Dispatch Number |
| 11 | Report Type Code | A system code for report types, these have corresponding descriptions within the dataset. |
| 12 | Report Type Description | The description of the report type |
| 13 | Filed Online | Police reports can be filed online for non-emergency cases. |
| 14 | Incident Code | Incident Codes are the system codes to describe a type of incident. |
| 15 | Incident Category | A category mapped on to the Incident Code |
| 16 | Incident Subcategory | A subcategory mapped on to the Incident Code |
| 17 | Incident Description | The description of the incident |
| 18 | Resolution | The resolution of the incident at the time of the report. |
| 19 | Intersection | The 2 or more street names that intersect closest to the original incident |
| 20 | CNN | The unique identifier of the intersection for reference back to other related base map datasets. |
| 21 | Police District | The Police District reflecting current boundaries |
| 22 | Analysis Neighborhood | Neighborhoods using common real estate and resident definitions |
| 23 | Supervisor District | The districts are numbered 1 through 11 |
| 24 | Latitude | The latitude coordinate |
| 25 | Longitude | The longitude coordinate |
| 26 | point | The point geometry used for mapping features |

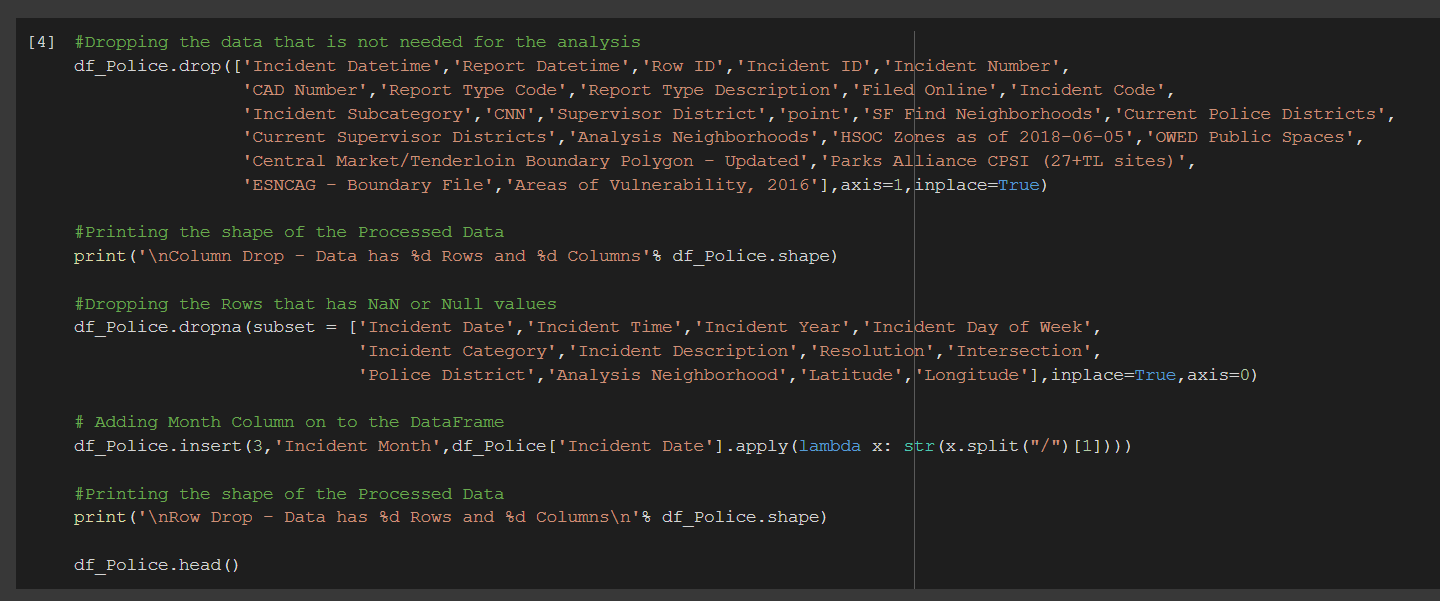
Visualisation of the dataset is important in getting more insights about it and discovering some pattern that might help in the modelling section. For more details on this, please refer the iPython notebook for Week2 – Capstone, File Name: ‘*Week2\_Capstone\_Project.ipynb’*

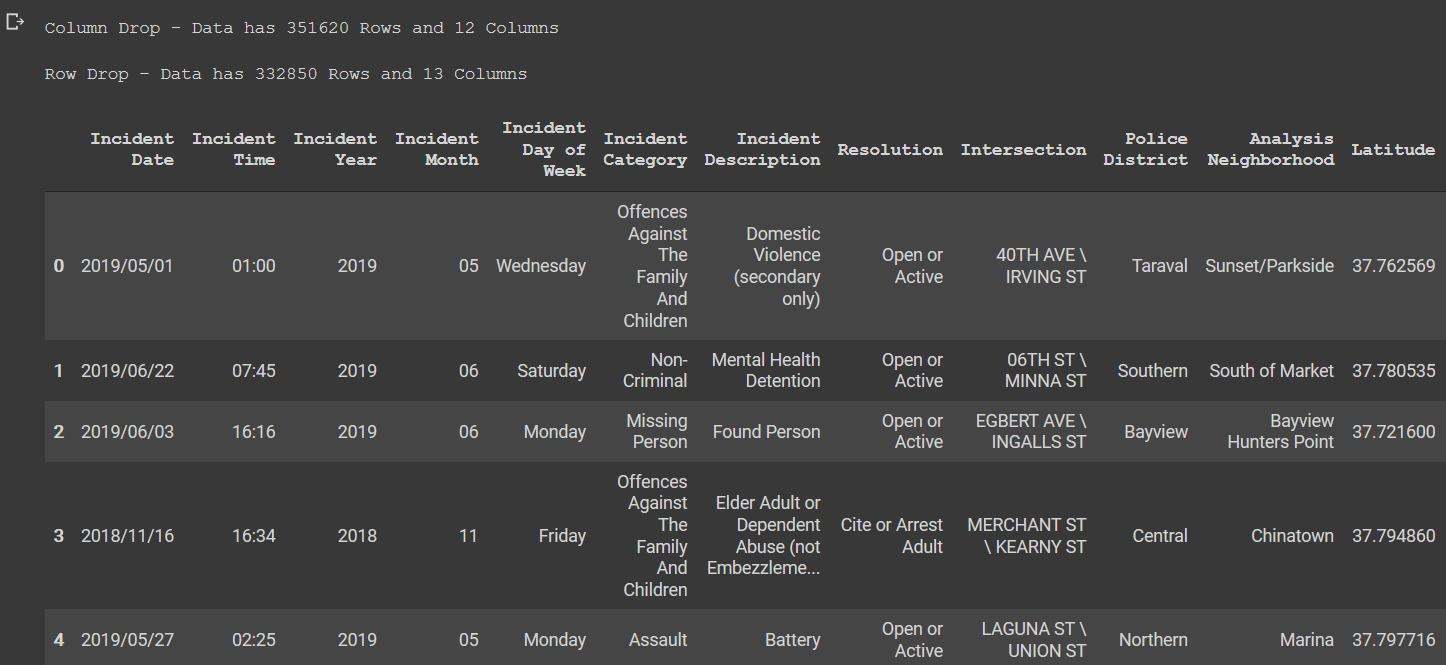
## Data Preparation and Pre-processing

In this component, the dataset is prepared for the modelling process where we choose the machine learning algorithms. To do that, cleaned the data from NaN values and removed the features that are not required for the analysis and modelling. Please find the details as below:

The Final Data frame is prepared with below features:

|  |  |  |
| --- | --- | --- |
| # | Feature Name | Feature Description |
| 1 | Incident Date | The date the incident occurred |
| 2 | Incident Time | The time the incident occurred |
| 3 | Incident Year | The year the incident occurred, provided as a convenience for filtering |
| 4 | Incident Month | The Month the incident occurred |
| 5 | Incident Day of Week | The day of week the incident occurred |
| 6 | Incident Category | A category mapped on to the Incident Code |
| 7 | Incident Description | The description of the incident |
| 8 | Resolution | The resolution of the incident at the time of the report. |
| 9 | Intersection | The 2 or more street names that intersect closest to the original incident |
| 10 | Police District | The Police District reflecting current boundaries |
| 11 | Analysis Neighborhood | Neighborhoods using common real estate and resident definitions |
| 12 | Latitude | The latitude coordinate |
| 13 | Longitude | The longitude coordinate |





## Modelling

There are 13 features/attributes in our prepared dataset. We need to effectively identify the features that has a direct impact on the Crime categories. We need to narrow down our analysis to find the right features which influence the categories of crime.

We can classify the crime based on the time and location of the incident, so we can only include those parameters that is associated with the location and time. Let’s narrow down parameters to

* + - Location (Latitude and Longitude, Police District)
    - Time (Year, Month and Time)

Aim of the project is to effectively classify a category of crime at a given time and location. A model can be built to achieve this.

Since this is under a classification problem, we can use multiple models like

1. K-Nearest Neighbour
2. Decision Tree
3. Logistic Regression
4. Support Vector Machine etc

### **K-Nearest Neighbour**

The Pre-processing of data is done to convert the categorical data into numeric data. Further data is split into train and test data for both X (Dependant) and Y (Target) variable types.

K value is found to be 9 with 0.25 accuracy

Train and test accuracy score are calculated and F1 accuracy score is arrived.

|  |  |
| --- | --- |
| **Train set Accuracy** | **0.3547** |
| **Test set Accuracy** | **0.2472** |
| **F1 Accuracy** | **0.1714** |
| **Jaccard Index Score** | **0.2472** |

### **Decision Tree**

Decision Tree modelling is done with criterion=​gini and depth value of 80 to achieve the proper results.

|  |  |
| --- | --- |
| **Decision Trees’ Accuracy** | **0.2447** |
| **Jaccard Index Score** | **0. 2447** |

### **Logistic Regression**

Logistic Regression modelling is carried out with the data and it was found as below:

|  |  |
| --- | --- |
| **Train set Accuracy** | **0.** **3033** |
| **Test set Accuracy** | **0. 3013** |
| **F1 Accuracy** | **0.** **1395** |
| **Jaccard Index Score** | **0. 3013** |
| **Log Loss** | **2.8177** |

## Evaluation and Testing

In this part we test the modeling algorithms by calculating the accuracy and f1-measure and Jaccard index. We have also search for the best k that can give us the best classification model. In our case k was equal to 9. Refer the table below Please check the file named “*Week2\_Capstone\_Project\_Final.ipynb*” for more details.

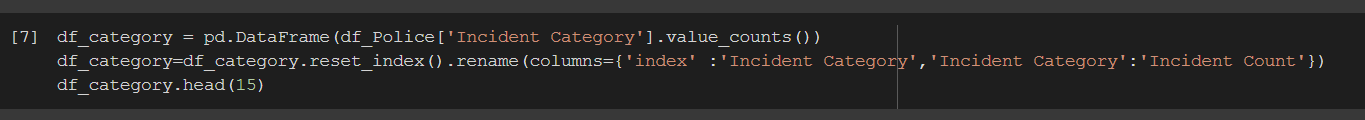
|  |  |  |
| --- | --- | --- |
| **Classifier** | **Parameters** | **Accuracy** |
| K-Nearest Neighbour | K=9 | **0.2472** |
| Decision Tree | Depth = 80 | **0.2447** |
| Logistic Regression | Log loss = 2.82 | **0.3013** |

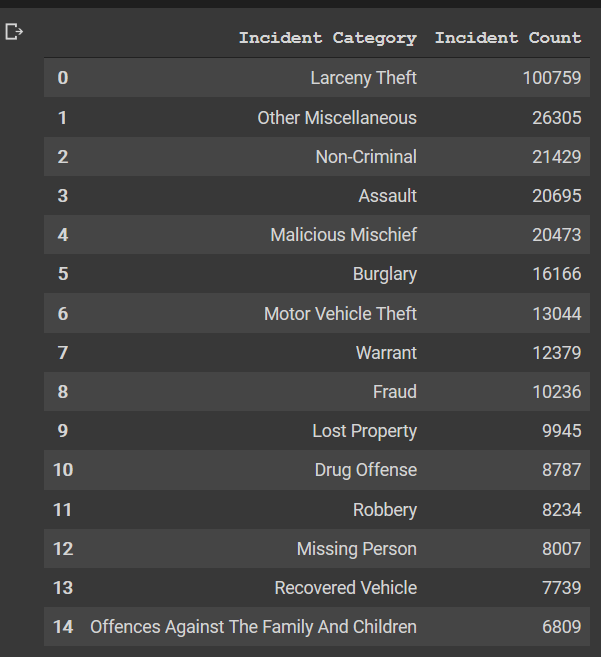
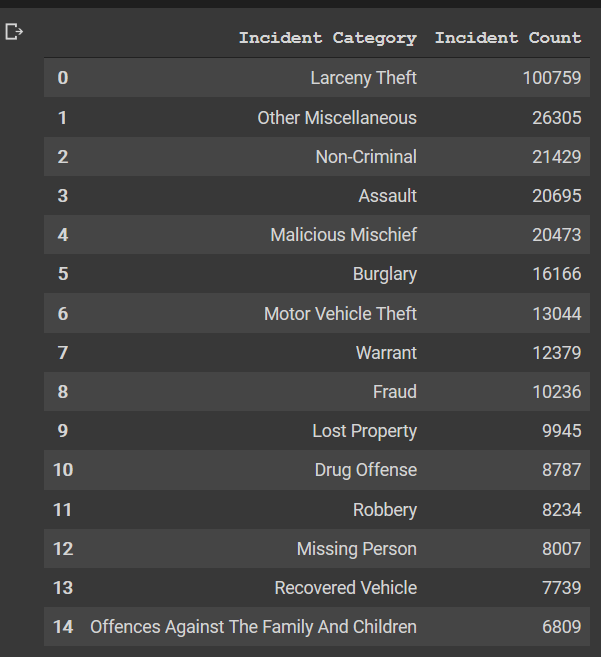
# Results

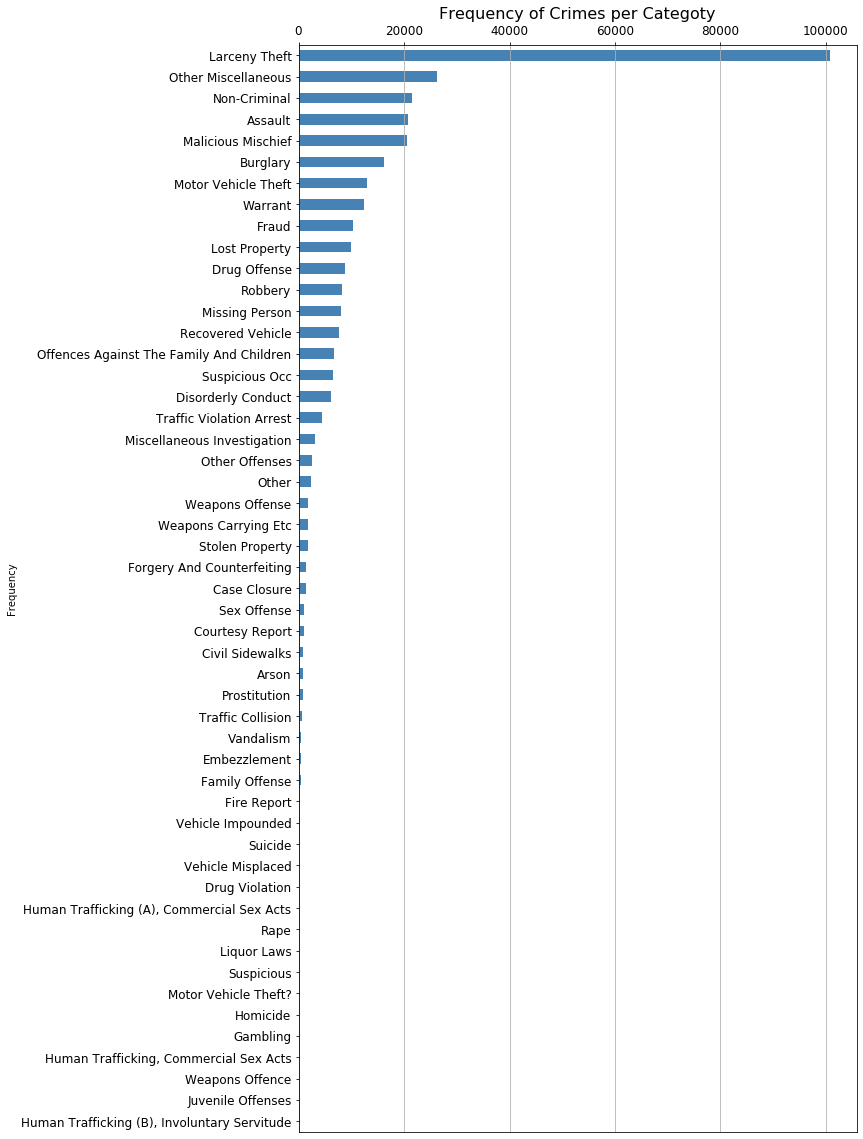
### **Crimes per Category**

A report on classification of crimes based on category gave the following result:

Since there are 51 total categories of crimes, top 15 is taken,



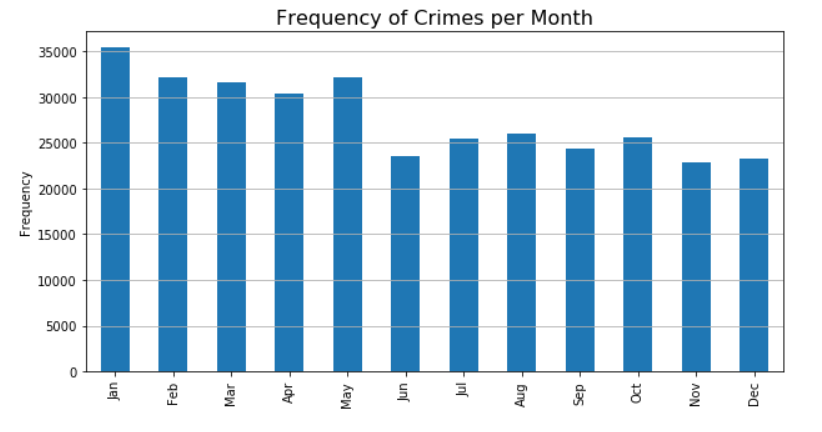


From the visualization it is clear that the Larceny Theft is occurring very frequently and has a high rate compared to all the other incidents. Also, some of the cases are very rare, in visualization/graph it is very evident that the crime cases are not equally distributed.

### **Time Crime Incident:**

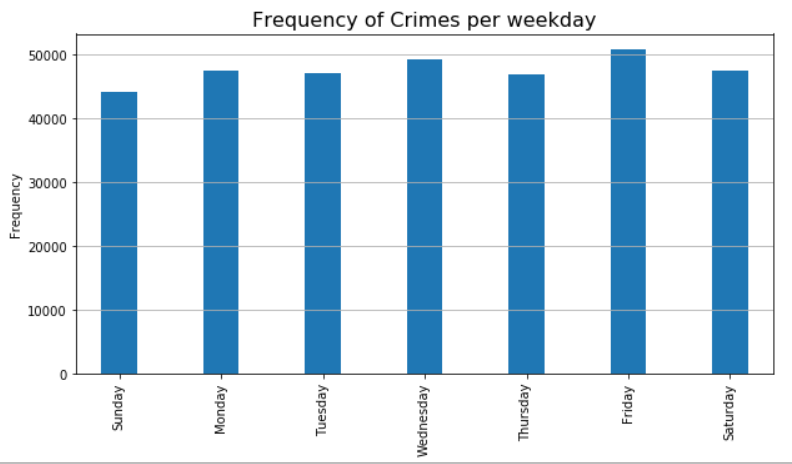
Data is plotted against the time frame to see if there is any connection/relation between the timeframe and the crime frequency.

#### By Month:

****

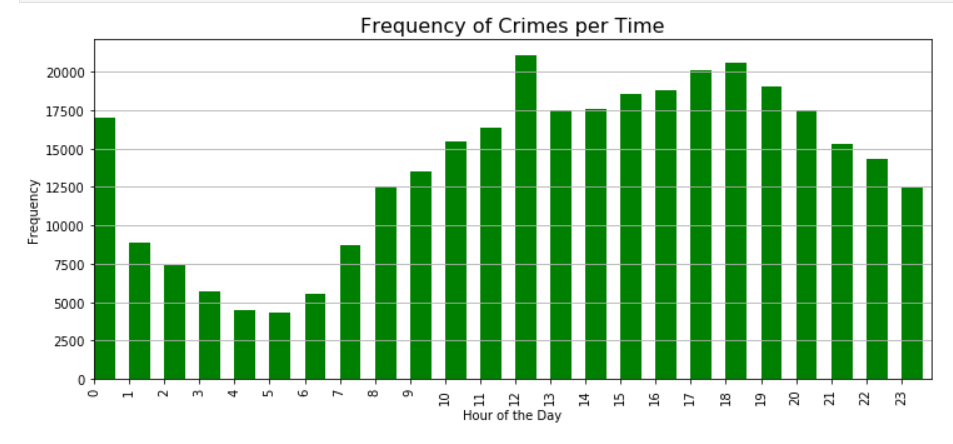
From the bar chart it is clear that the crime incidents happen throughout the year with very small fluctuation in frequency, also it is noticed that the beginning months are reported with more crimes comparatively.

#### By Week Day:



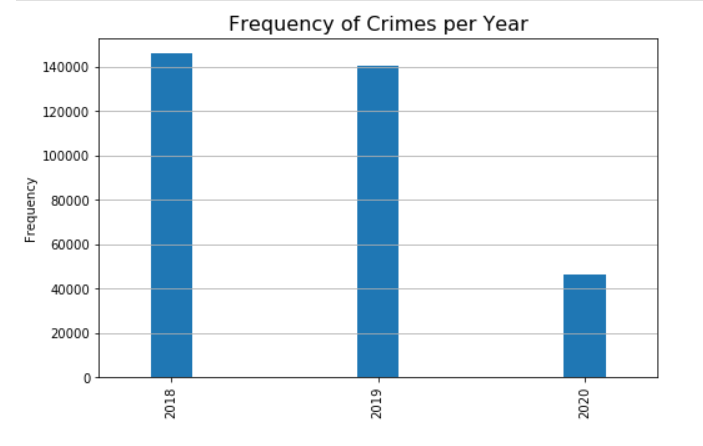
From the graph, crime rates are more on Fridays and less on Sundays. But there is no pattern of any increasing or decreasing trends as such.

#### By Time of the Day:



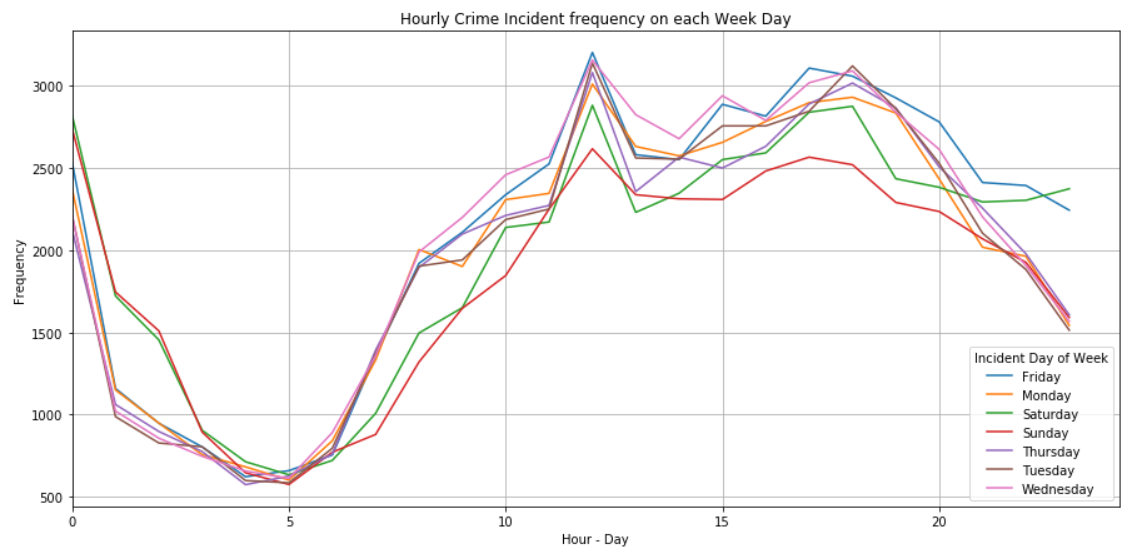
The graph shows that the crime incidents happen more in the evenings and nights than in early morning. From midnight to early morning crime activities are really low. From 3PM to 8PM the crime activities are high.

#### By Year:

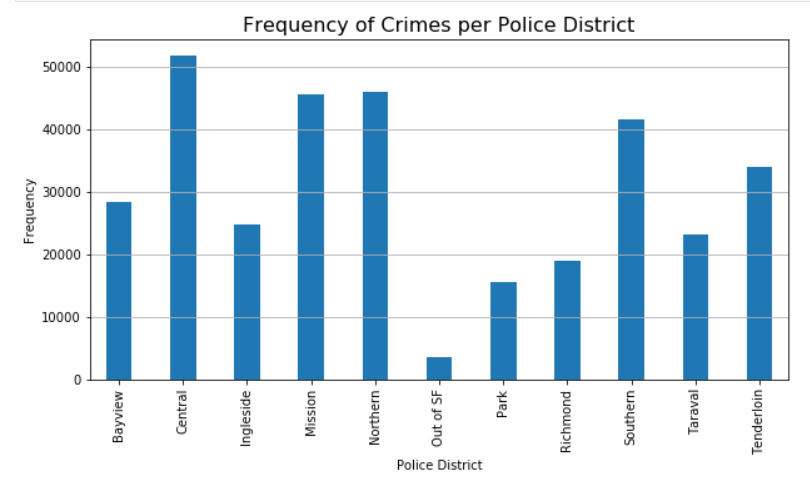


The trend for 2018 and 2019 shows almost same trend with a small decrease in the incident frequency. No inference can be driven for 2020 as it is the current running year.

#### By Hours in each weekday:



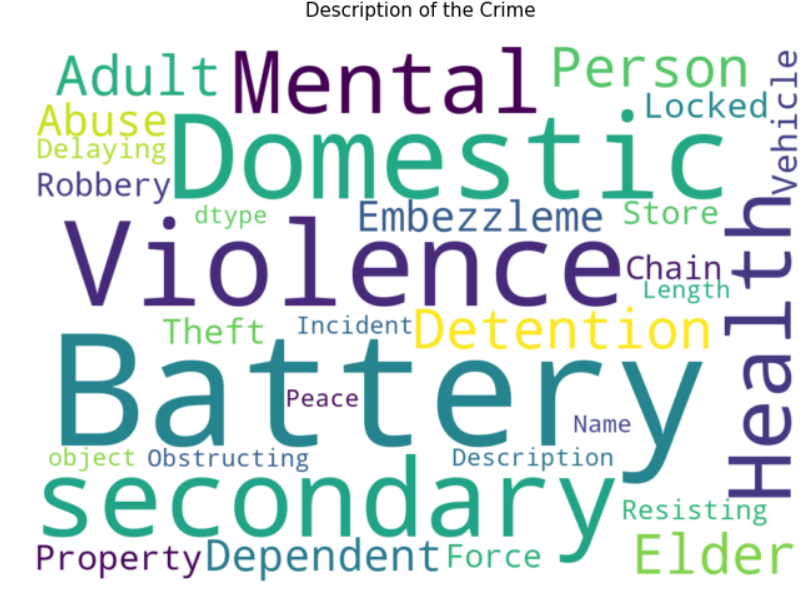
### **Crime per Police District:**

****

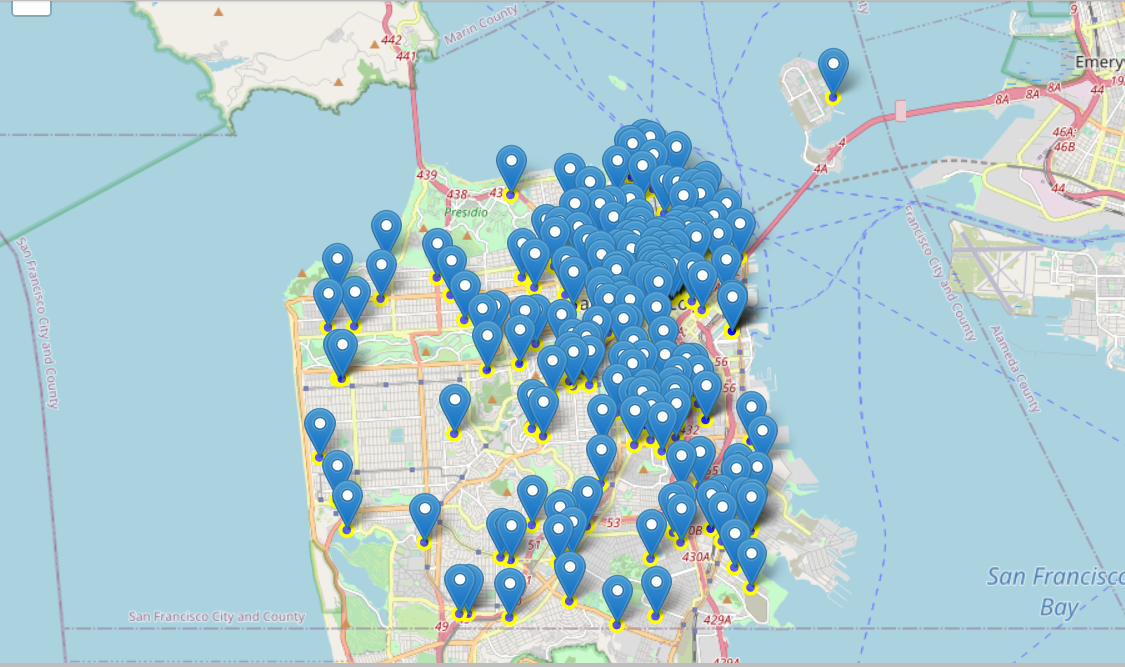
The plot shows that the most crimes has happened in central region where as very less crimes are reported from Out of San Francisco region. Mission, Northern and Southern are around 45000 crimes reported.

### **Word-Cloud**

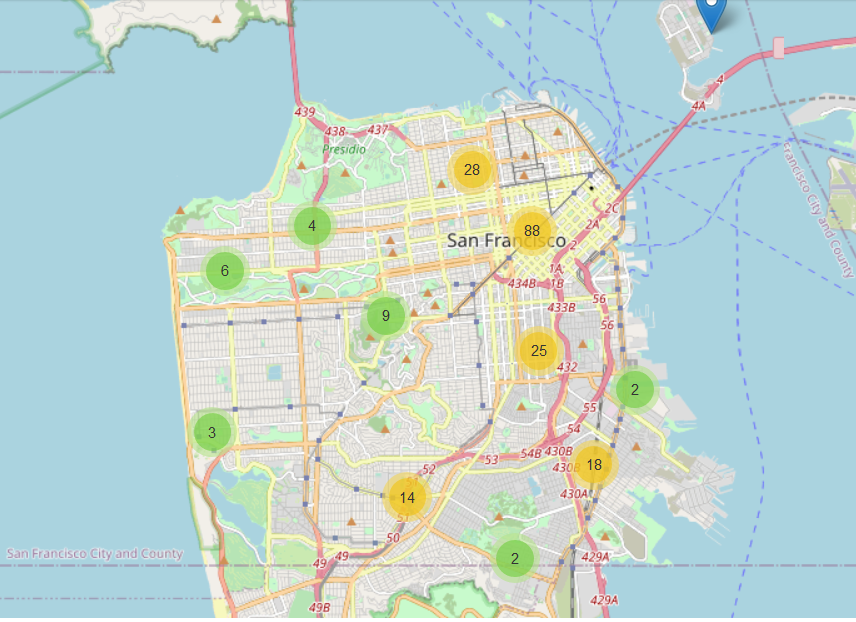
A word-cloud is generated to see the most occurring crime and frequency.



# Geo-Location of crime incidents

The below map shows the first 200 incident in San Francisco. 

Since there are many points, it looks clumsy and untidy, a different approach would be a clustered view as cited below.



# Using Foursquare to visualize and infer crime incidents and locality

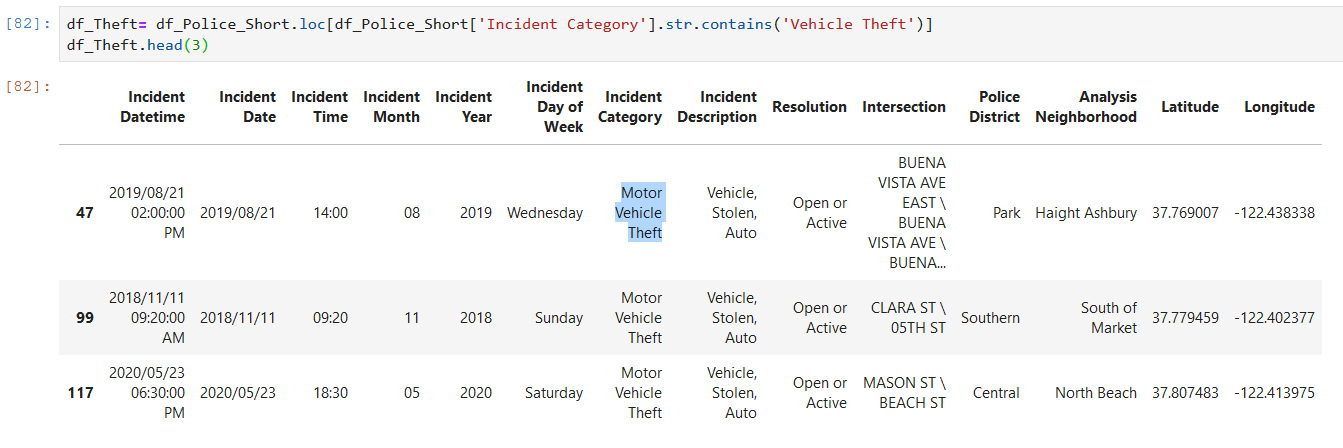
We will make calls to the Foursquare API for different purposes. You will construct a URL to send a request to the API to search for a specific type of venues, to explore a particular business venue, to explore a Foursquare user, to explore a geographical location, and to get trending venues around a location. Also, you will learn how to use the visualization library, Folium, to visualize the results.

The Data Frame is filtered for different crime categories. One of the examples is as shown below:

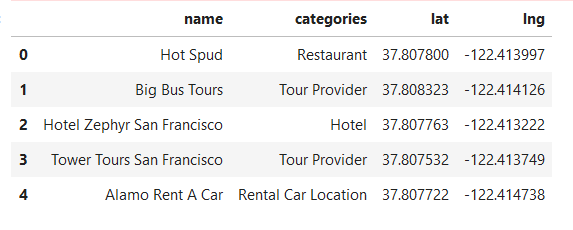
**Category: Vehicle Theft**

Analyze and find the most theft occurred area and obtain its Latitude and Longitude, this can be used to find the venues around the area and can be marked as vulnerable area and need proper monitoring to avoid and eliminate such crimes.

Data frame is obtained as below



Foursquare API for exploring the venues nearby is done to obtain the below data:



The data is plotted to obtain much insights about criticality and vulnerability,



The marked in Red is the crime scene/crime location. The blue marks are the locations that are explored using Foursquare API. From this it can be inferred that the area around is vulnerable and security need to be tightened.

# Conclusion

To eliminate the crime occurrence and to bring public safety and peace in place, it is utmost important to find the category of the crime incident and its nature. Police and security service department need to continuously monitor and tighten the measures to eliminate different categories of criminal activities. In public security services, technology can be used to predict a likely critical violation through the use of data analytics instead of inspecting every joint blindly given the lack of enough manpower for this. The data used to predict critical violation include location and time data. Afterward, places data e.g. Foursquare is used to locate the most affected areas and can be worked towards improving the security and surveillance measures.

# Link to the python code

