Capstone Project: The Battle of Neighborhoods (Week1)

San Francisco Police Department Incident Reports: 2018 to Present

Author: Jithin Prakash Kolamkolly

Contents

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 andLongitude

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.

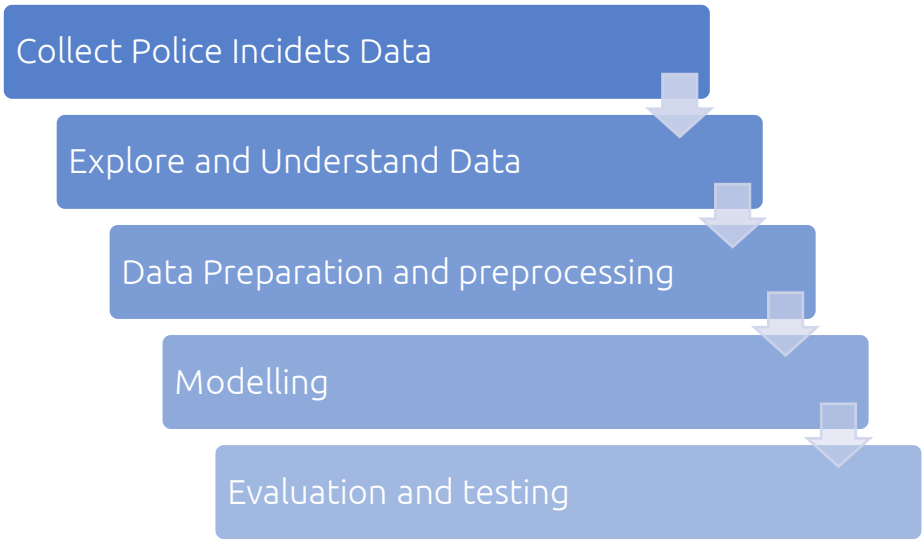


Figure 1 – Components of Methodology (CRISP-DM)

1. Collect Data

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

Importing Data from website (https://data.sfgov.org/api/views/wg3w-h783/rows.csv)

```
[2]  !wget -q -O 'Police_incidents.csv' https://data.sfgov.org/api/views/wg3w-h783/rows.csv
     print('Data fetched from website')

     Data fetched from website
```

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.

2. 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.

```
[3]  #Reading from the CSV file to the data frame
     df_Police = pd.read_csv('Police_incidents.csv')
     #Printing the shape of the Raw Data
     print('\nRaw data has %d Rows and %d Columns\n'% df_Police.shape)
     #Displaying the first 5 rows
     df_Police.head()
```

```
Raw data has 351620 Rows and 36 Columns
```

	Incident Datetime	Incident Date	Incident Time	Incident Year	Incident Day of Week	Report Datetime	Row ID	Incident ID	Incident Number	CAD Number	Report Type Code	Report Type Description	File Online
0	2019/05/01 01:00:00 AM	2019/05/01	01:00	2019	Wednesday	2019/06/12 08:27:00 PM	81097515200	810975	190424067	191634131.0	II	Initial	Nal
1	2019/06/22 07:45:00 AM	2019/06/22	07:45	2019	Saturday	2019/06/22 08:05:00 AM	81465564020	814655	190450880	191730737.0	II	Initial	Nal
2	2019/06/03 04:16:00 PM	2019/06/03	16:16	2019	Monday	2019/06/03 04:16:00 PM	80769875000	807698	190397016	191533509.0	IS	Initial Supplement	Nal
3	2018/11/16 04:34:00 PM	2018/11/16	16:34	2018	Friday	2018/11/16 04:34:00 PM	73857915041	738579	180870806	183202539.0	IS	Initial Supplement	Nal
4	2019/05/27 02:25:00	2019/05/27	02:25	2019	Monday	2019/05/27 02:55:00	80509204134	805092	190378555	191470256.0	II	Initial	Nal

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*'

3. 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

```
[4] #Dropping the data that is not needed for the analysis
    df_Police.drop(['Incident Datetime','Report Datetime','Row ID','Incident ID','Incident Number',
                    'CAD Number','Report Type Code','Report Type Description','Filed Online','Incident Code',
                    'Incident Subcategory','CNN','Supervisor District','point','SF Find Neighborhoods','Current Police Districts',
                    'Current Supervisor Districts','Analysis Neighborhoods','HSOC Zones as of 2018-06-05','OWED Public Spaces',
                    'Central Market/Tenderloin Boundary Polygon - Updated','Parks Alliance CPSI (27+TL sites)',
                    'ESNCAG - Boundary File','Areas of Vulnerability, 2016'],axis=1,inplace=True)

    #Printing the shape of the Processed Data
    print('\nColumn Drop - Data has %d Rows and %d Columns'% df_Police.shape)

    #Dropping the Rows that has NaN or Null values
    df_Police.dropna(subset = ['Incident Date','Incident Time','Incident Year','Incident Day of Week',
                               'Incident Category','Incident Description','Resolution','Intersection',
                               'Police District','Analysis Neighborhood','Latitude','Longitude'],inplace=True,axis=0)

    # Adding Month Column on to the DataFrame
    df_Police.insert(3,'Incident Month',df_Police['Incident Date'].apply(lambda x: str(x.split("/")[1])))

    #Printing the shape of the Processed Data
    print('\nRow Drop - Data has %d Rows and %d Columns\n'% df_Police.shape)

    df_Police.head()
```

```
Column Drop - Data has 351620 Rows and 12 Columns

Row Drop - Data has 332850 Rows and 13 Columns
```

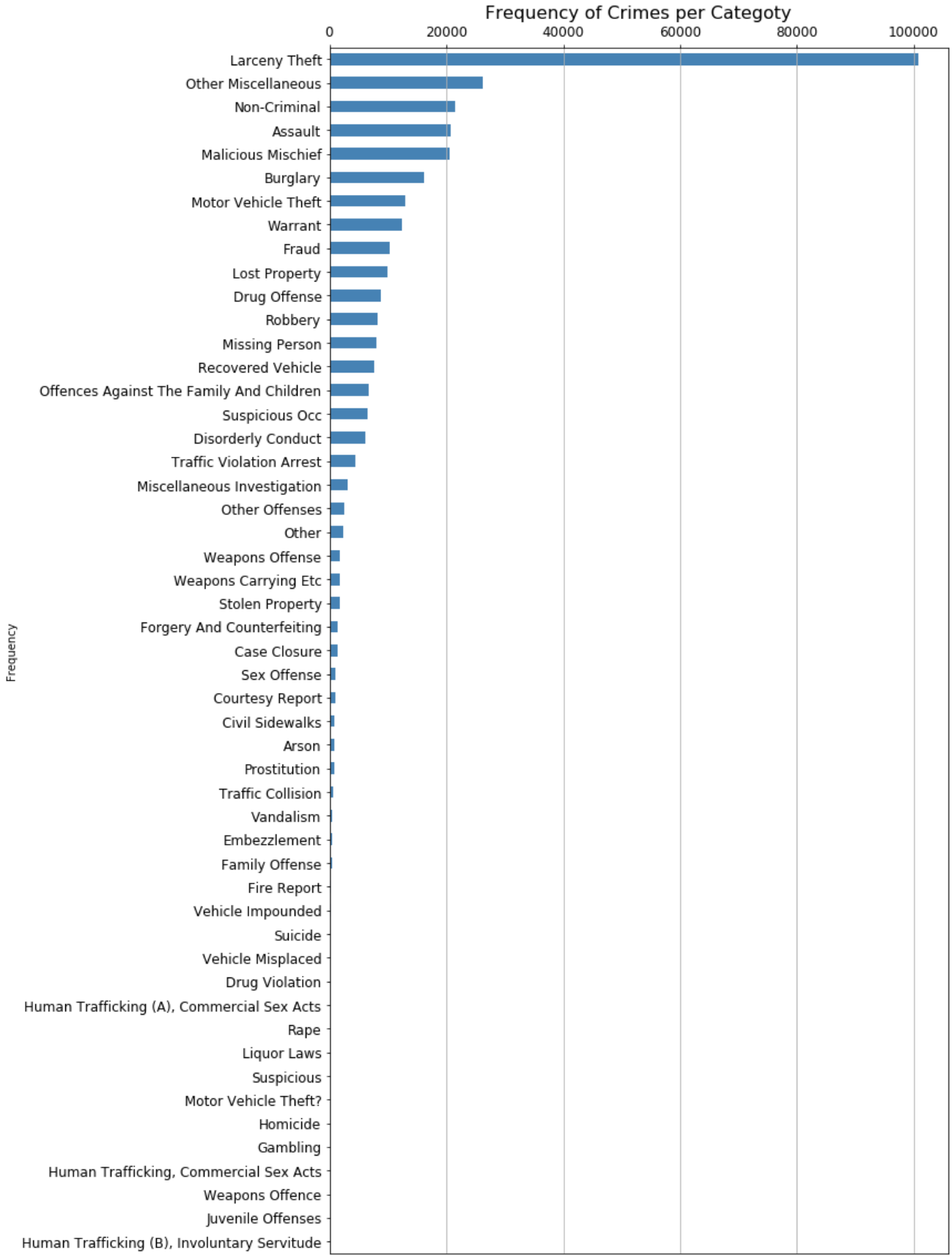
	Incident Date	Incident Time	Incident Year	Incident Month	Incident Day of Week	Incident Category	Incident Description	Resolution	Intersection	Police District	Analysis Neighborhood	Latitude
0	2019/05/01	01:00	2019	05	Wednesday	Offences Against The Family And Children	Domestic Violence (secondary only)	Open or Active	40TH AVE \ IRVING ST	Taraval	Sunset/Parkside	37.762569
1	2019/06/22	07:45	2019	06	Saturday	Non-Criminal	Mental Health Detention	Open or Active	06TH ST \ MINNA ST	Southern	South of Market	37.780535
2	2019/06/03	16:16	2019	06	Monday	Missing Person	Found Person	Open or Active	EGBERT AVE \ INGALLS ST	Bayview	Bayview Hunters Point	37.721600
3	2018/11/16	16:34	2018	11	Friday	Offences Against The Family And Children	Elder Adult or Dependent Abuse (not Embezzleme...	Cite or Arrest Adult	MERCHANT ST \ KEARNY ST	Central	Chinatown	37.794860
4	2019/05/27	02:25	2019	05	Monday	Assault	Battery	Open or Active	LAGUNA ST \ UNION ST	Northern	Marina	37.797716

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,

```
[7] df_category = pd.DataFrame(df_Police['Incident Category'].value_counts())
    df_category=df_category.reset_index().rename(columns={'index' :'Incident Category','Incident Category':'Incident Count'})
    df_category.head(15)
```

	Incident Category	Incident Count
0	Larceny Theft	100759
1	Other Miscellaneous	26305
2	Non-Criminal	21429
3	Assault	20695
4	Malicious Mischief	20473
5	Burglary	16166
6	Motor Vehicle Theft	13044
7	Warrant	12379
8	Fraud	10236
9	Lost Property	9945
10	Drug Offense	8787
11	Robbery	8234
12	Missing Person	8007
13	Recovered Vehicle	7739
14	Offences Against The Family And Children	6809

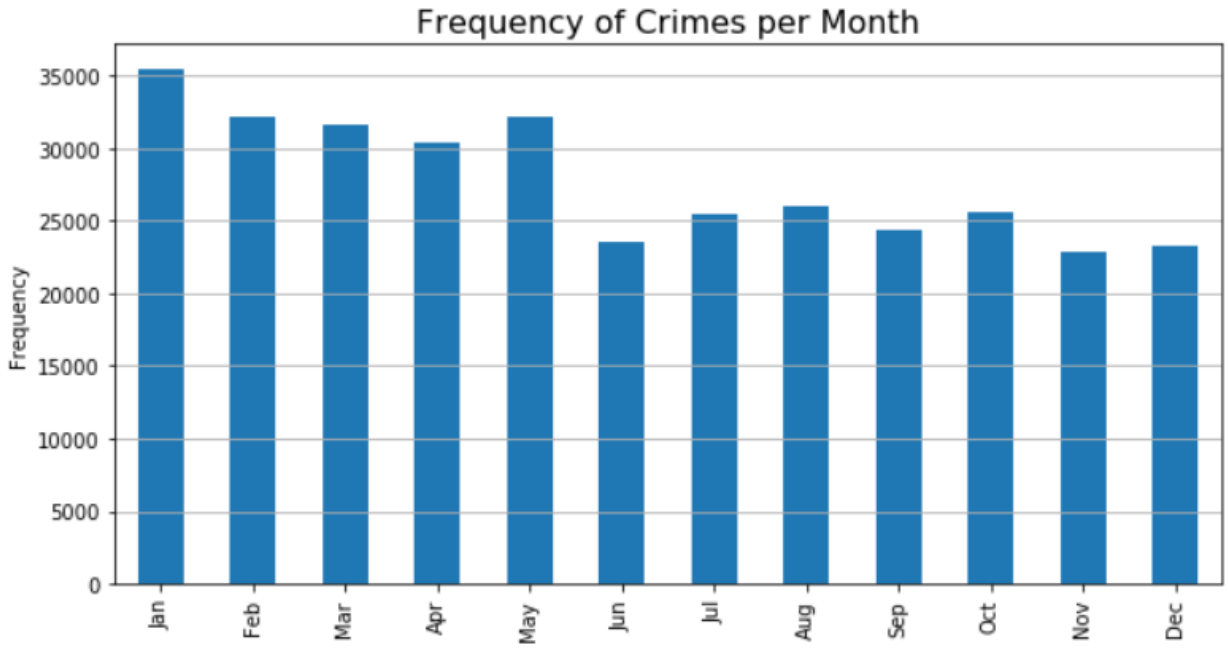


From the visualization it is clear that the Larceny Theft is occurring very frequently and has a high rate compared to all the other incidents. Also, some of the cases are very rare, in visualization/graph it is very evident that the crime cases are not equally distributed.

Time Crime Incident:

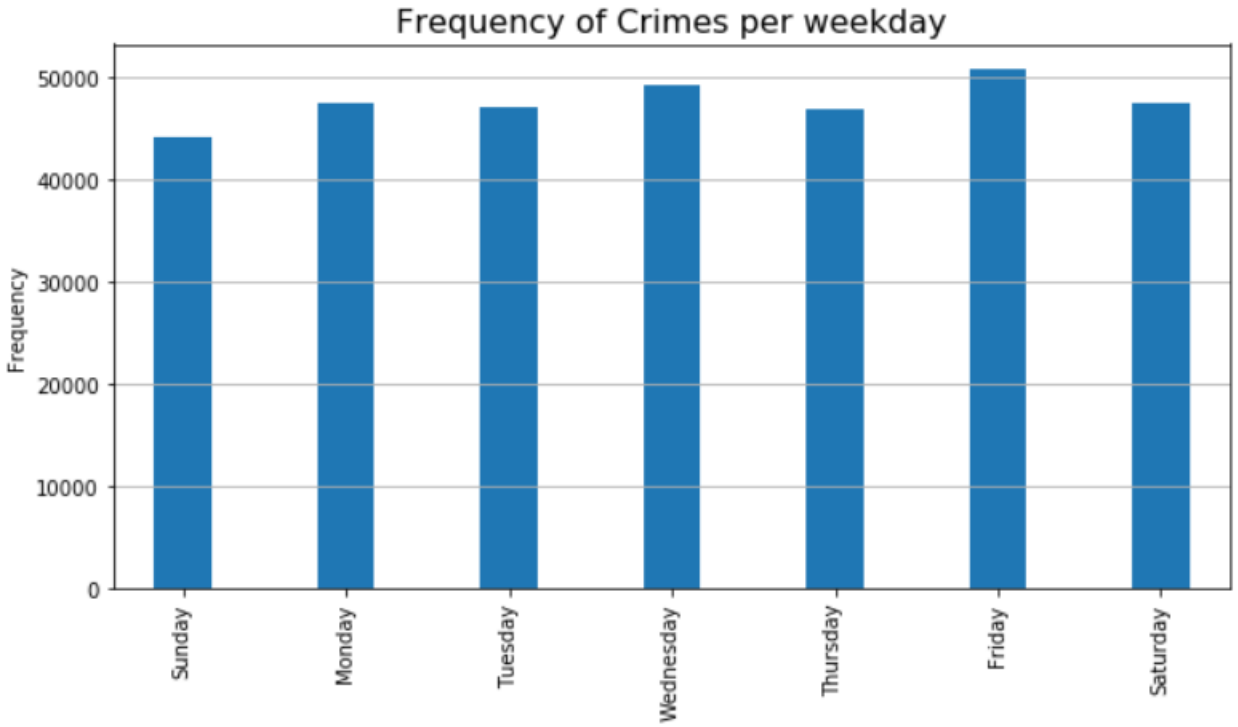
Data is plotted against the time frame to see if there is any connection/relation between the timeframe and the crime frequency.

By Month:



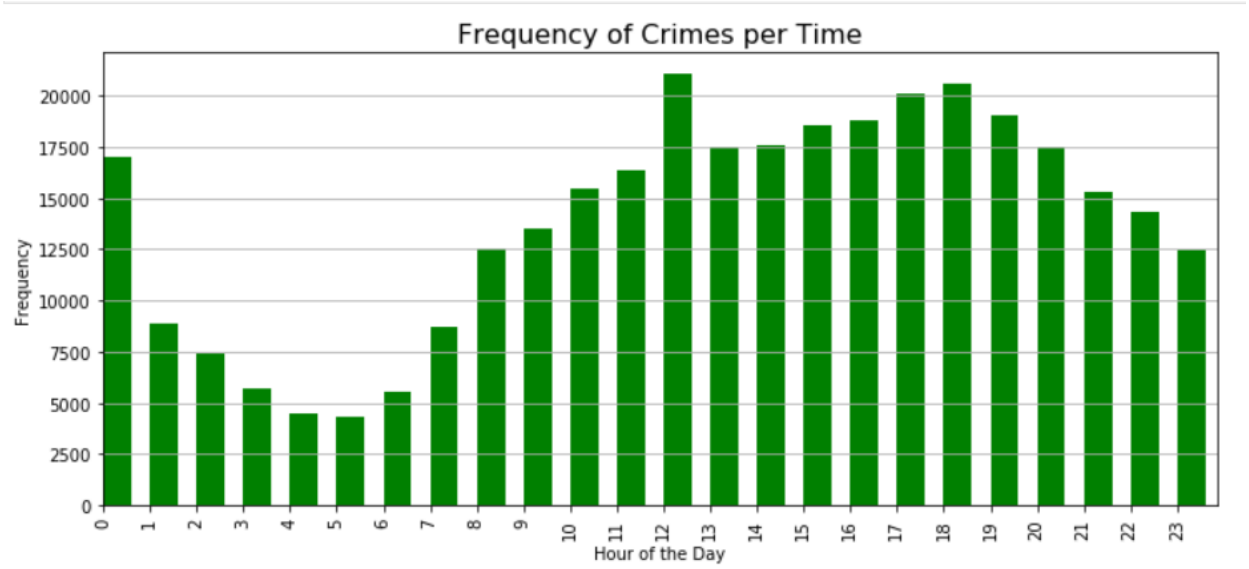
From the bar chart it is clear that the crime incidents happen throughout the year with very small fluctuation in frequency, also it is noticed that the beginning months are reported with more crimes comparatively.

By Week Day:



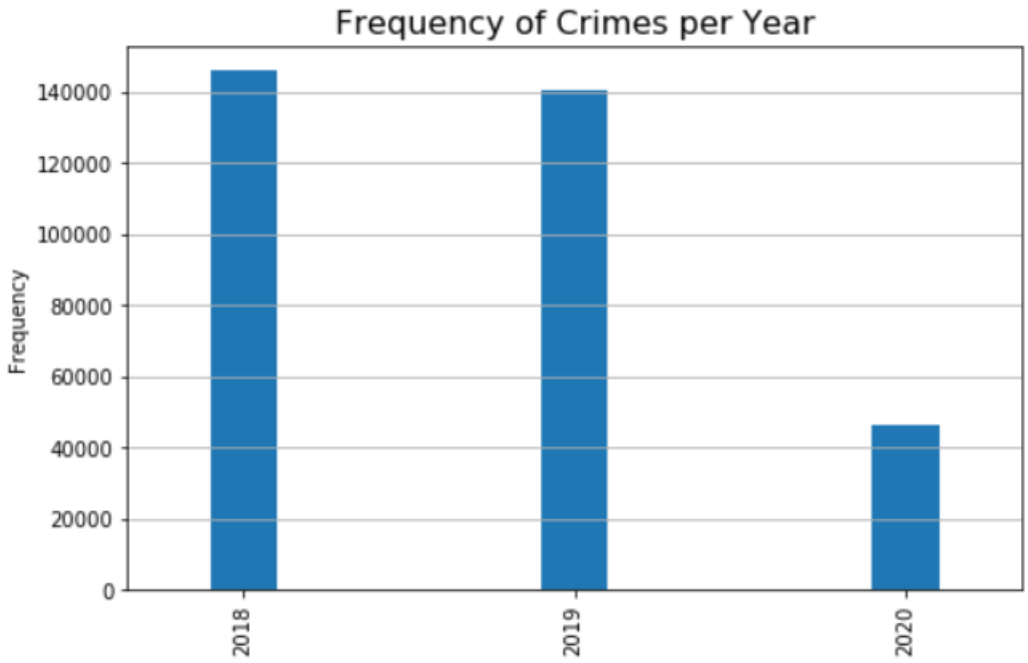
From the graph, crime rates are more on Fridays and less on Sundays. But there is no pattern of any increasing or decreasing trends as such.

By Time of the Day:



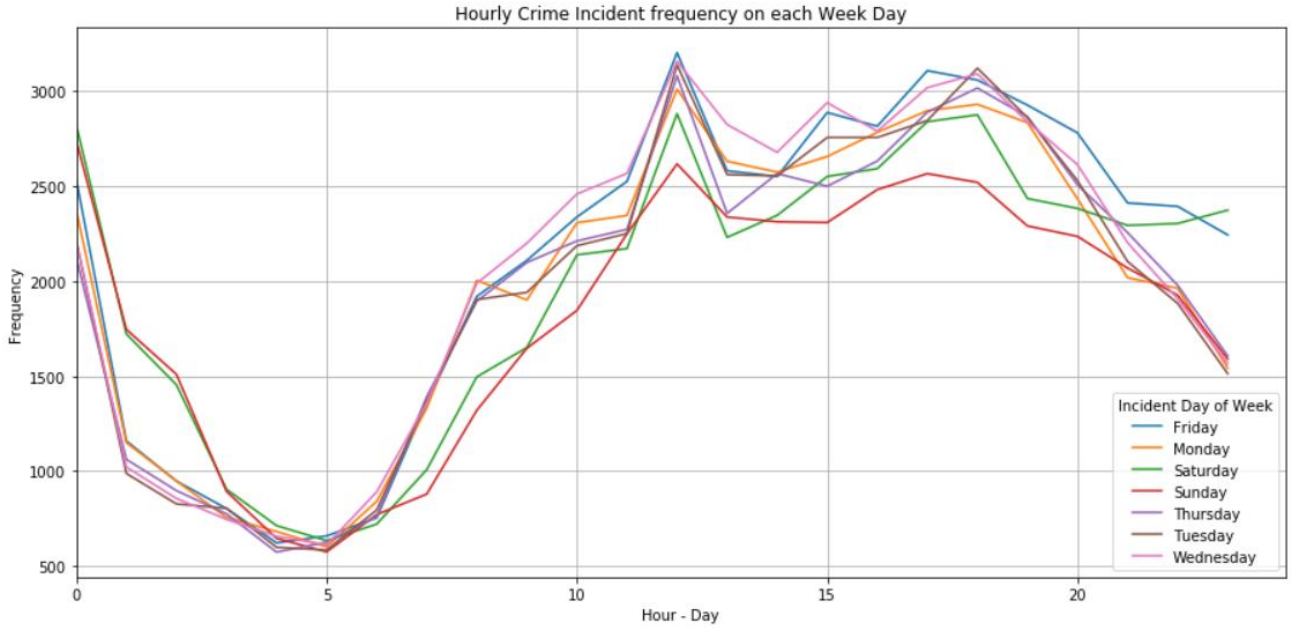
The graph shows that the crime incidents happen more in the evenings and nights than in early morning. From midnight to early morning crime activities are really low. From 3PM to 8PM the crime activities are high.

By Year:

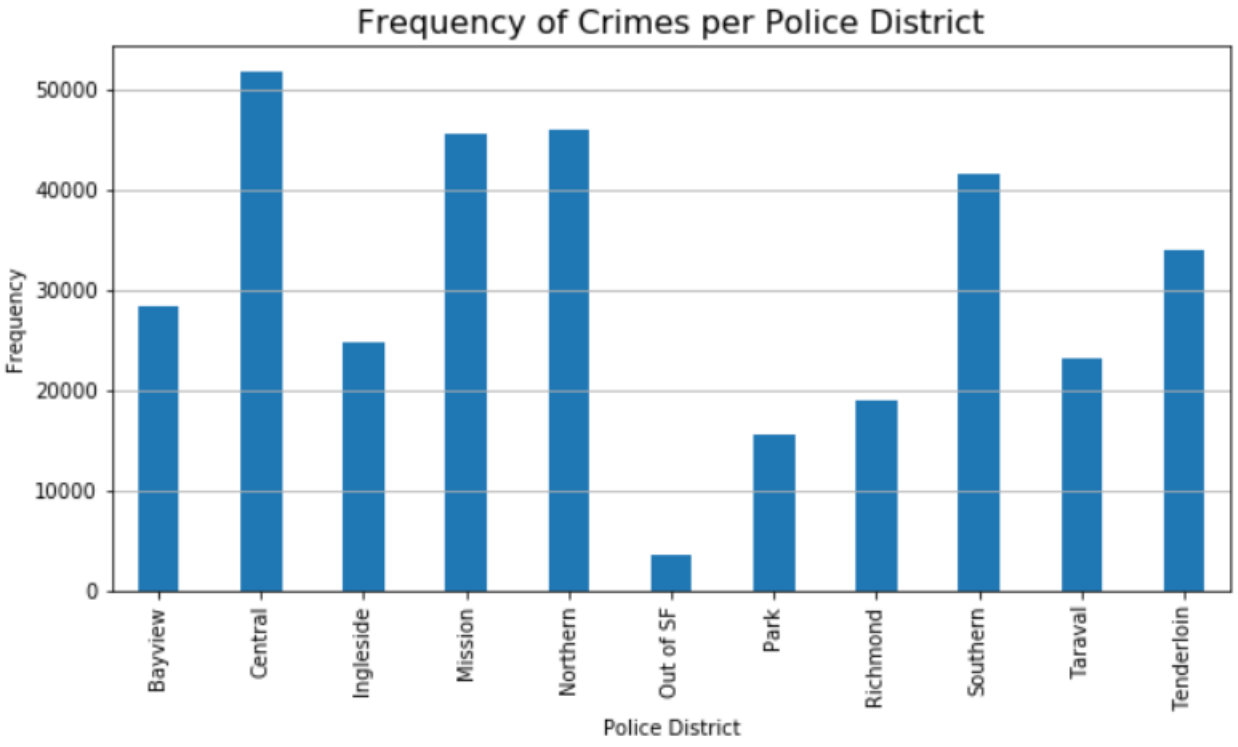


The trend for 2018 and 2019 shows almost same trend with a small decrease in the incident frequency. No inference can be driven for 2020 as it is the current running year.

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.

4. 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

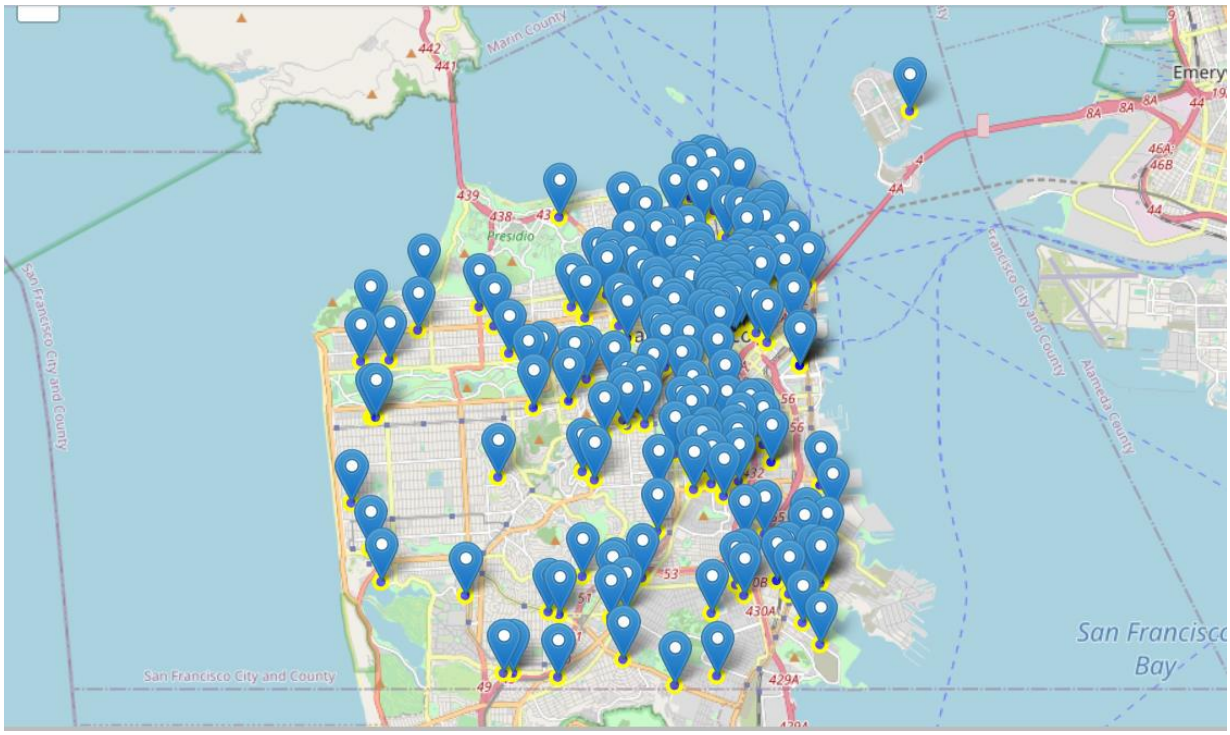
Logistic Regression modelling is carried out with the data and it was found as below:

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

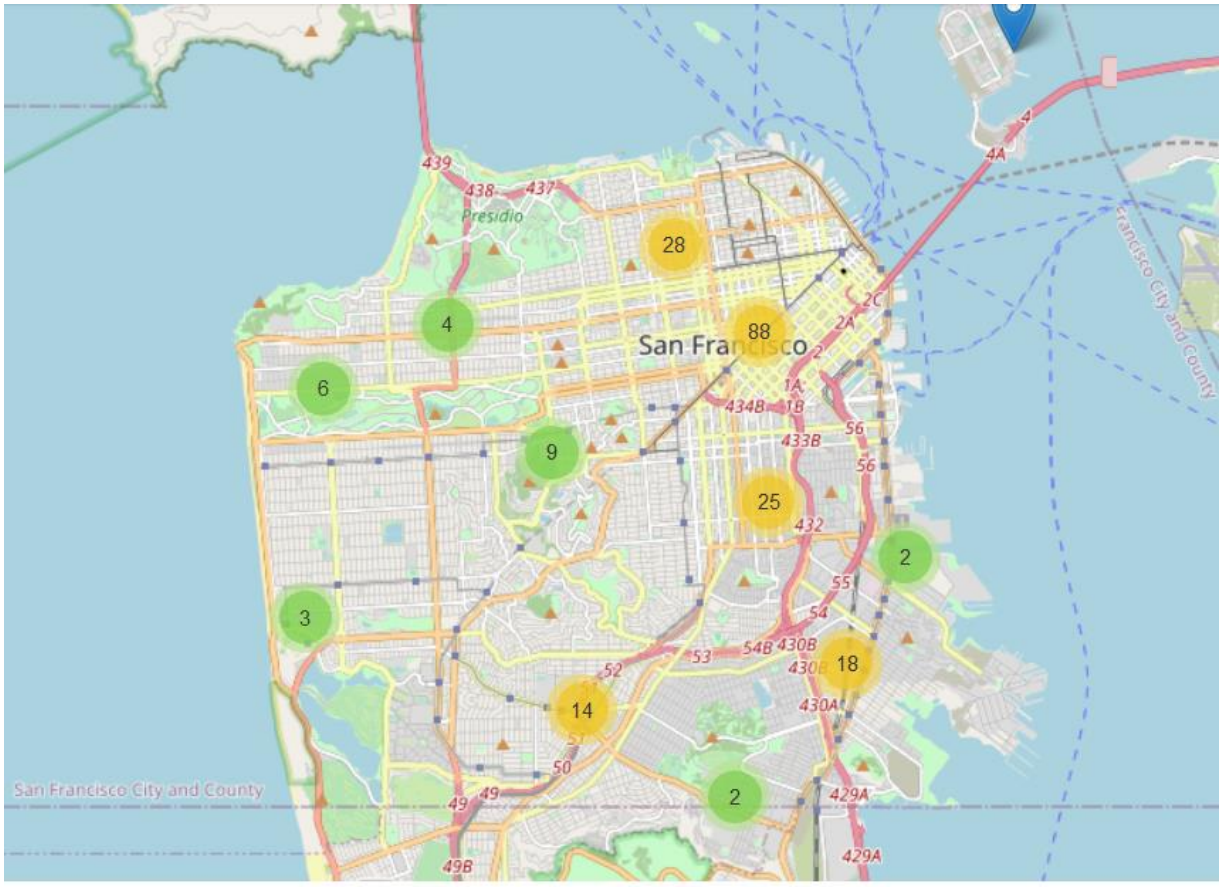
Classifier	Parameters	Accuracy
K-Nearest Neighbour	K=9	0.2472
Decision Tree	Depth = 80	0.2447
Logistic Regression	Log loss = 2.82	0.3013

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 businesses venues

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

```
[82]: df_Theft= df_Police_Short.loc[df_Police_Short['Incident Category'].str.contains('Vehicle Theft')]
      df_Theft.head(3)
```

[82]:

	Incident Datetime	Incident Date	Incident Time	Incident Month	Incident Year	Incident Day of Week	Incident Category	Incident Description	Resolution	Intersection	Police District	Analysis Neighborhood	Latitude	Longitude
47	2019/08/21 02:00:00 PM	2019/08/21	14:00	08	2019	Wednesday	Motor Vehicle Theft	Vehicle, Stolen, Auto	Open or Active	BUENA VISTA AVE EAST \ BUENA VISTA AVE \ BUENA...	Park	Haight Ashbury	37.769007	-122.438338
99	2018/11/11 09:20:00 AM	2018/11/11	09:20	11	2018	Sunday	Motor Vehicle Theft	Vehicle, Stolen, Auto	Open or Active	CLARA ST \ 05TH ST	Southern	South of Market	37.779459	-122.402377
117	2020/05/23 06:30:00 PM	2020/05/23	18:30	05	2020	Saturday	Motor Vehicle Theft	Vehicle, Stolen, Auto	Open or Active	MASON ST \ BEACH ST	Central	North Beach	37.807483	-122.413975

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

	name	categories	lat	lng
0	Hot Spud	Restaurant	37.807800	-122.413997
1	Big Bus Tours	Tour Provider	37.808323	-122.414126
2	Hotel Zephyr San Francisco	Hotel	37.807763	-122.413222
3	Tower Tours San Francisco	Tour Provider	37.807532	-122.413749
4	Alamo Rent A Car	Rental Car Location	37.807722	-122.414738

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.

Link to the python code

Python Code