SAN JOSÉ STATE UNIVERSITY

CMPE 274-02 – Project Report

Crime Data Analysis: San Francisco Crime Data

Group 6

Mounica Reddy Kandi
Preethi Billa
Pavan Karthik Gollakaram
Niranjan Reddy Masapeta

Table of Contents

1. Team Details

- Mounica Reddy Kandi
- 016021902
- mounicareddy.kandi@sjsu.edu
- 4083872864

- Niranjan Reddy Masapeta
- 015748122
- niranjanreddy.masapeta@sjsu.edu
- 9165409975

- Preethi Billa
- 015920411
- preethi.billa@sjsu.edu
- 4085494116

- Pavan Karthik Gollakaram
- 015945670
- pavankarthik.gollakaram@sjsu.edu
- 4084621091

2. Project Idea

2.1. Motivation

According to US News, San Francisco has higher crime rate with 439.5 crime rate, higher than similar sized metropolitan areas. San Francisco stands below the mean scale with a 6.9 crime index, faring below the average point. For a city with population close to a million, this means thousands of criminal activities: Arson, Automotive theft, Larceny, Burglary, Aggravated assault, murder, and rape, on a regular basis. For the second wealthiest city in the U.S, their richest among the world, this is a deterring factor be it financially, morally, socially. Thus, our project is focused on making lives of these million people safer through proper usage of the crime data that is available for everyone to make use of.

2.2. Dataset

'Data.world' is the world's largest collaborative open data community that works towards transforming the world's data into knowledge, opportunities. In their process of democratizing data access and championing inclusive, agile processes for data work, dataset containing San Francisco's crime date from January 1st 2003, to 15th March 2015 is made available.

The dataset, with 800,000 records originally derived from the SFPD Crime Incident Reporting System, consists detailed information about each incident. This includes information like the timestamp, address including the latitude and longitude of the crime taken place and the category it belongs to.

Dataset resource link:
https://data.world/data-society/san-francisco-crime-data

A few of the Data Fields:

- **Date** – Incident's date is mentioned here in a standardized format to avoid any further complications.
- **Category –** Clear distinctions according to the laws are made to avoid redundancy.
- **Descript –** The PD verbatim of each case is digitalized as well.
- **DayOfWeek –** This information helps in determining the weekday vs weekend security risks. The similarities and the differences as well.
- **PdDistrict –** The district area in which the crime has taken place.
- **Address –** The precise address in the standardized format is recorded.
- **X (Longitude) –** To further augment the accuracy of the location, precise longitude numbers are included as well.
- **Y (Latitude) -** To further augment the accuracy of the location, precise latitude numbers are included as well.

2.3. Approach

Leveraging multiple open-source tools ranging from Tableau for Business Intelligence for detailed statistics, package of Python and Jupyter Notebook for preprocessing, customizing the analysis, and Tableau for Data Visualization and , we successfully extracted the required visualizations:

1. Day-wise count and respective percentage share of the crime activities.
2. PdDistrict wise percentage share of criminal activities.
3. Count of criminal activities pertaining to each category of the crime.
4. Category wise count of criminal activities pertaining to each district area.

3. System Architecture

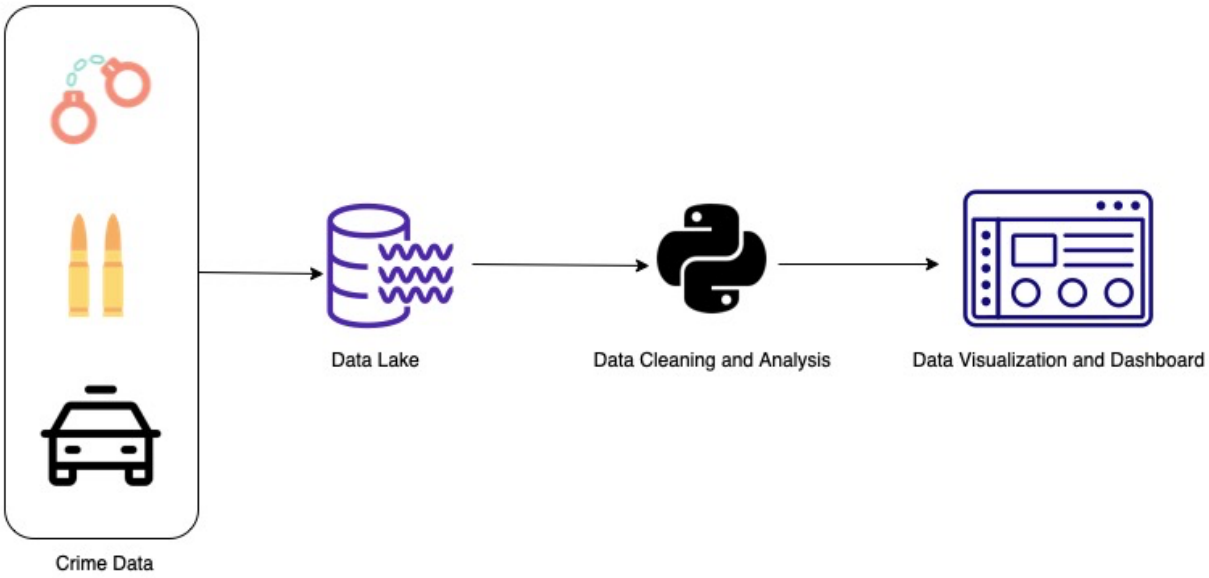


Fig 1: Crime Data Analysis System Architecture

- **Crime Data:** As mentioned above, 'Data.world' provided us with the dataset with all the required data fields.

- **Data Lake:** You can store all your structured and unstructured data in a data lake, which is a centralized repository that works at any scale. You can use data source in Tableau for using it to build dashboard.
- **Data Cleaning and Analysis:**
  - Data Cleaning: The quality with which this first and foremost step in the process is handled, determines a lot about the quality of the result. Thus, we further broke down this process into three parts:
    1. **Missing Values:**
       Deletion or removal of the missing value. Use the mean, median, and mode t o impute missing values. Employ the Sklearn Impute Model (SimpleImputer, I terativeImputer, KNNImputer).
    2. **Outlier Detection:** Among the methods for identifying outliers are boxplots, Z-scores, and Inter Quantile Range (IQR).
    3. **Handling Impurities:** Noisy values and a majority of times the actual data might be missing.
- **Analysis:** Exploratory Data Analysis using Jupyter Notebook for Data preprocessing.
- **Data Visualization and Dashboard:**
  - **Data Visualization:** Leveraging both Jupyter Notebook and Tableau, we visualize the required data trends. Maps, boxplots, graphs are all generated.
  - **Dashboard:** As we have mentioned in the beginning of the project, we created an interactive dashboard in Tableau.

4. Functional and Non-functional requirements

4.1. Functional

- Find patterns to track different crimes depending on areas of interest and time.
- Visualize and categorize crimes based on their features given in the data to study more about criminal behavior.
- User available dashboard for active vigilance.

4.2. Non-functional

- Exploratory Data Analysis using Jupyter Notebook for Data preprocessing.
- Dashboard with visualizations of the given data using Tableau.

4.3. Tech Stack & Tools used:

- Python
- Jupyter Notebook

- Pandas
- Scikit-Learn
- Seaborn
- Folium
- Tableau Public

5. User Scenarios

Official usage:

- Crime department uses our dashboard for future possible crime predictions to prevent them in advance.
- Dashboard also helps in maintaining securities in areas of high crime activity.

Public users:

- Public users might access few sections of the dashboard to stay vigilant in the areas of high criminal activity.

6. Performance Metrics

Software development metrics are quantitative evaluations of a software project or product that aid in management comprehension of the functionality, effectiveness, and productivity of software teams.

- Accuracy of the crime location based on the latitude and longitude.
- Time taken to visually represent the data on the dashboard.
- Observing the data to know more about the criminal behavior.

7. Jupyter Notebook

7.1. Sample Codes

- Importing required modules

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
import seaborn as sns

#from google.colab import drive

#drive.mount('/content/drive')
```

- Import folium library for interactive map design

```
import folium
from folium.plugins import HeatMap, HeatMapWithTime, MarkerCluster
import pytz
```

- Loading Dataset

```
#train = pd.read_csv('/content/drive/MyDrive/train-neeraj.csv', parse_dates=['Dates'])
#test = pd.read_csv('test.csv', parse_dates=['Dates'], index_col='Id')
train = pd.read_csv('train.csv', parse_dates=['Dates'])
```

```
train.tail()
```

	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
878044	2003-01-06 00:15:00	ROBBERY	ROBBERY ON THE STREET WITH A GUN	Monday	TARAVAL	NONE	FARALLONES ST / CAPITOL AV	-122.459033	37.714056
878045	2003-01-06 00:01:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Monday	INGLESIDE	NONE	600 Block of EDNA ST	-122.447364	37.731948
878046	2003-01-06 00:01:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Monday	SOUTHERN	NONE	5TH ST / FOLSOM ST	-122.403390	37.780266
878047	2003-01-06 00:01:00	VANDALISM	MALICIOUS MISCHIEF, VANDALISM OF VEHICLES	Monday	SOUTHERN	NONE	TOWNSEND ST / 2ND ST	-122.390531	37.780607
878048	2003-01-06 00:01:00	FORGERY/COUNTERFEITING	CHECKS, FORGERY (FELONY)	Monday	BAYVIEW	NONE	1800 Block of NEWCOMB AV	-122.394926	37.738212

- Data Cleaning and Processing

```
train['Year'] = train['Dates'].map(lambda x: x.year)
train['Week'] = train['Dates'].map(lambda x: x.week)
train['Hour'] = train['Dates'].map(lambda x: x.hour)
```

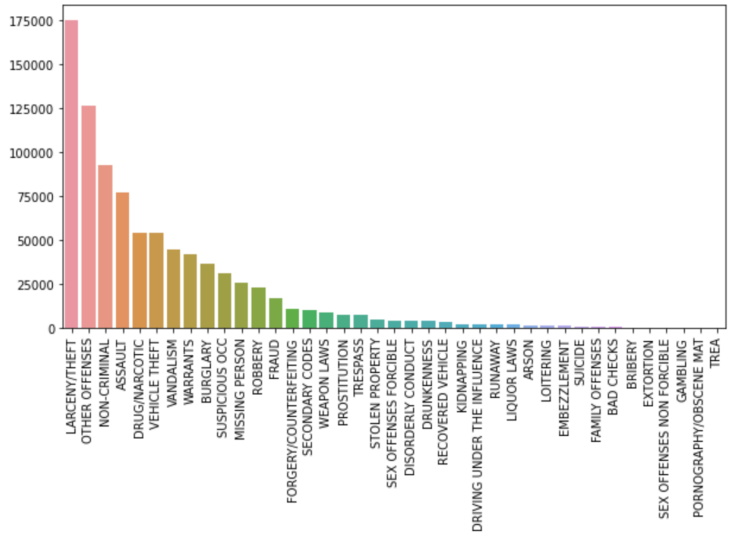
```
train.isnull().sum()
```

```
Dates         0
Category      0
Descript      0
DayOfWeek     0
PdDistrict    0
Resolution    0
Address       0
X             0
Y             0
Year          0
Week          0
Hour          0
dtype: int64
```

7.2. Data Visualizations

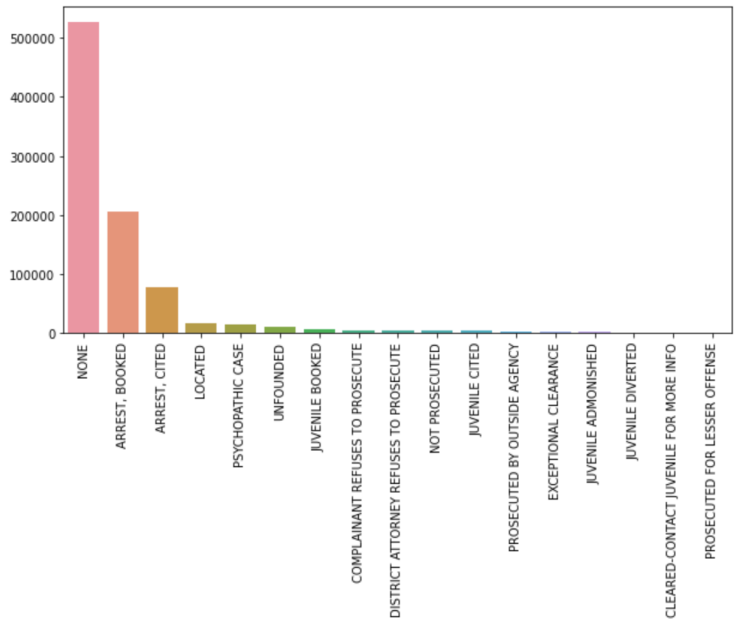
- Barplot showing the count of crimes filtered by category of crimes

```
category_vc = train['Category'].value_counts()
plt.figure(figsize=(10,5))
plt.tick_params(axis='x', rotation=90)
sns.barplot(x=category_vc.index, y=category_vc.values)
plt.show()
```



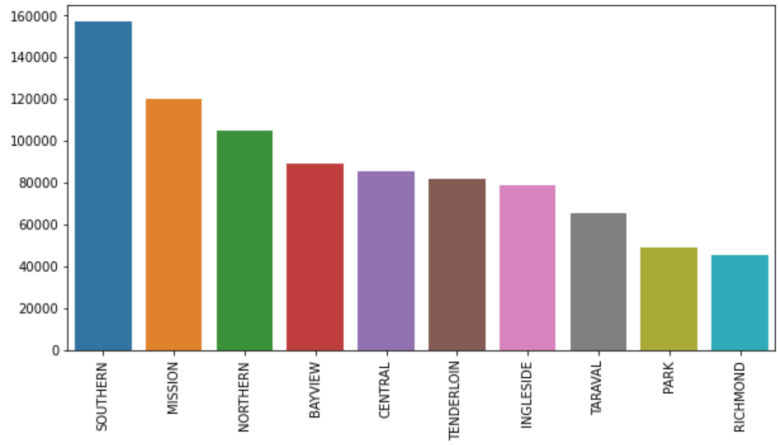
- Barplot showing the count of crimes filtered by resolution for the crimes

```
resolution_vc = train['Resolution'].value_counts()
plt.figure(figsize=(10,5))
plt.tick_params(axis='x', rotation=90)
sns.barplot(x=resolution_vc.index, y=resolution_vc.values)
plt.show()
```



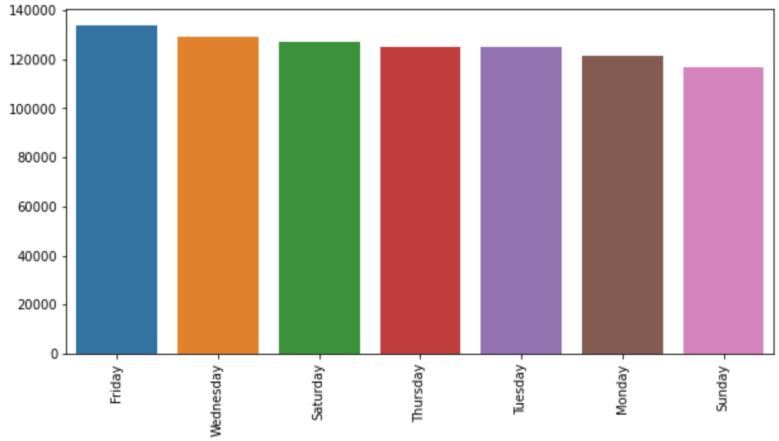
- Barplot showing the count of crimes filtered by district where crimes happened

```
district_vc = train['PdDistrict'].value_counts()
plt.figure(figsize=(10,5))
plt.tick_params(axis='x', rotation=90)
sns.barplot(x=district_vc.index, y=district_vc.values)
plt.show()
```



- Barplot showing the count of crimes filtered by day of the week the crime happened

```
dayofweek_vc = train['DayOfWeek'].value_counts()
plt.figure(figsize=(10,5))
plt.tick_params(axis='x', rotation=90)
sns.barplot(x=dayofweek_vc.index, y=dayofweek_vc.values)
plt.show()
```

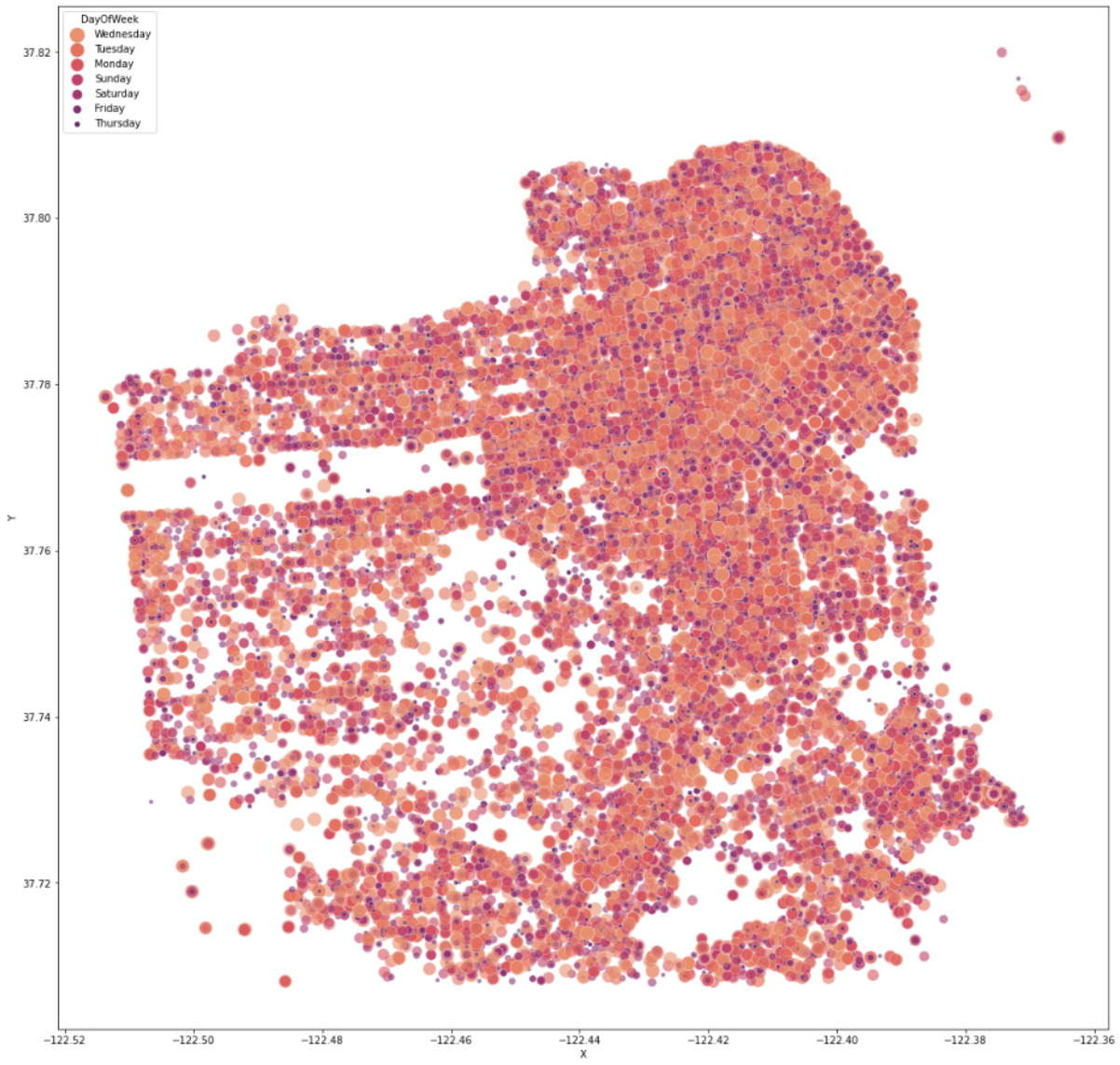


- Scatterplot showing the map of San Francisco filtered by category of crimes

```
fig, ax = plt.subplots(1, 1, figsize=(19, 19))
sns.scatterplot(data=train.iloc[:250000], x='X', y='Y', alpha=0.6, palette='rocket', hue='Category', size='Cat
plt.legend(bbox_to_anchor=(1.0, 1.0), loc='upper left')
plt.show()
```

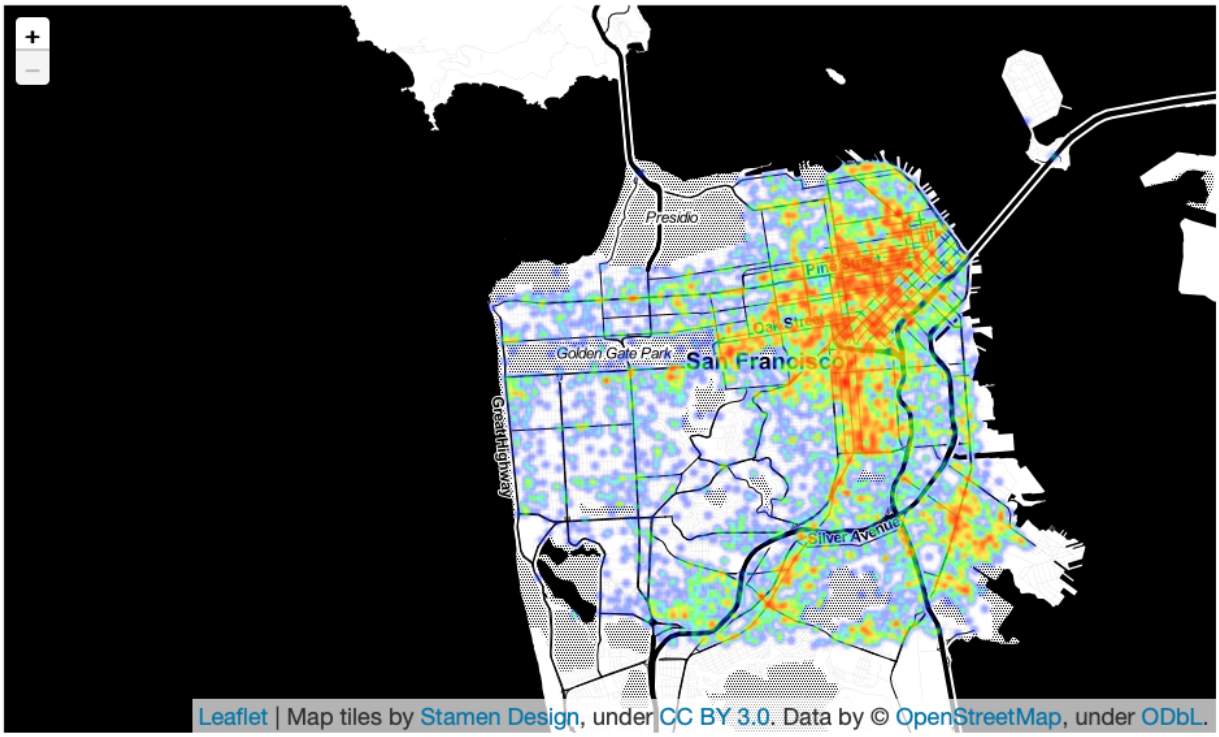
- Scatterplot showing the map of San Francisco filtered by day of the week the crime happened

```
fig, ax = plt.subplots(1, 1, figsize=(19, 19))
sns.scatterplot(data=train.iloc[:50000], x='X', y='Y', alpha=0.6, palette='flare', hue=
plt.show()
```



- Interactive Map using Folium package filtered by Category and Resolution

```
train_query = train[(train['Category']=='WARRANTS') & (train['Resolution']=='ARREST, BO
m = folium.Map(location=[37.774599, -122.425892], zoom_start=13, tiles='stamentoner', m
train_query_geo_list = train_query.values.tolist()
HeatMap(train_query_geo_list, blur=2, radius=3).add_to(m)
m
```

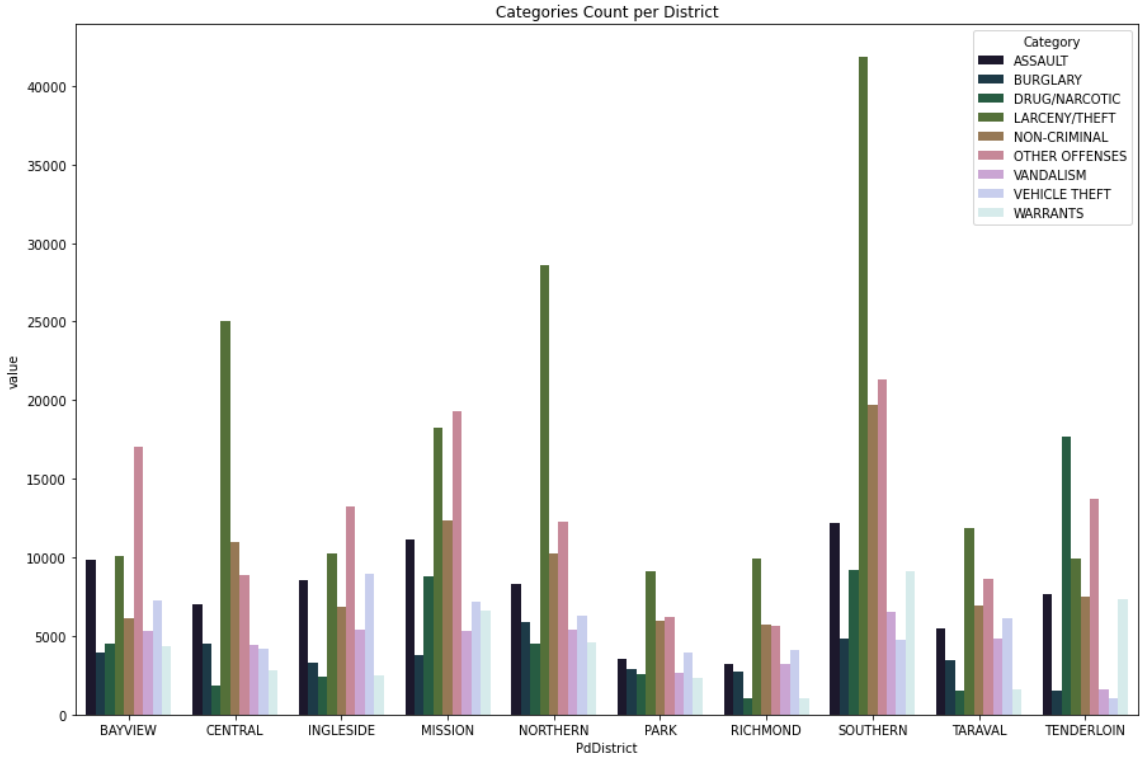


- Barplot showing the count of crimes filtered by Categories Count per District

```
train_50_list = train.loc[train['Category'].isin(frequency_50_list)]
ct_50_list = pd.crosstab(train_50_list['Category'], train_50_list['PdDistrict'])
stack_50_list = ct_50_list.stack().reset_index().rename(columns= {0:'value'})

fig, ax = plt.subplots(1, 1, figsize= (15,10))
sns.barplot(x=stack_50_list['PdDistrict'], y=stack_50_list['value'], hue=stack_50_list[
ax.set_title('Categories Count per District')
```

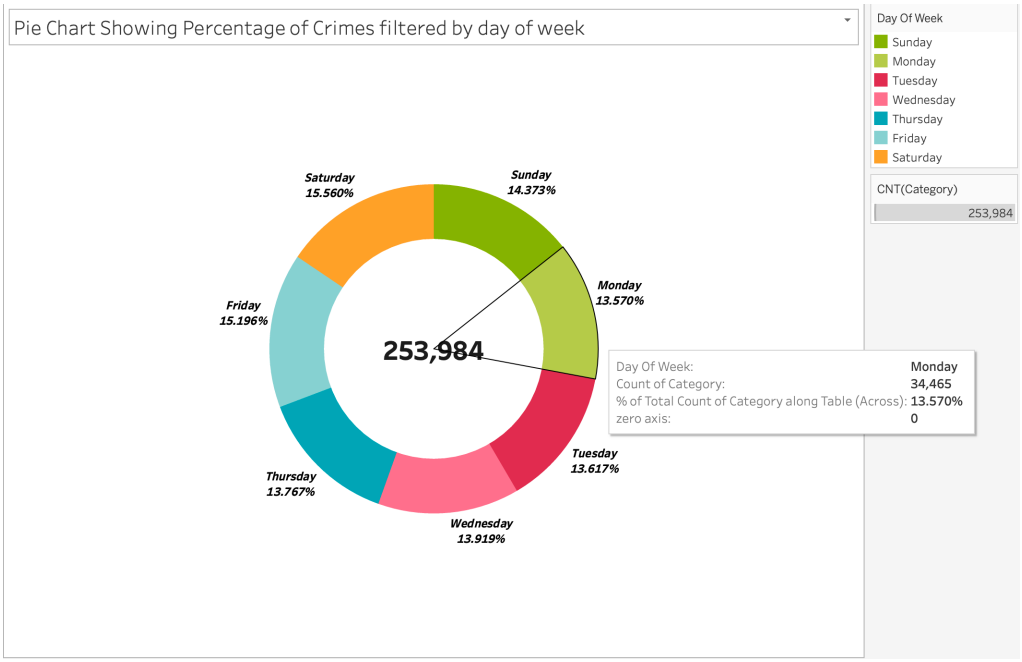
```
Text(0.5, 1.0, 'Categories Count per District')
```



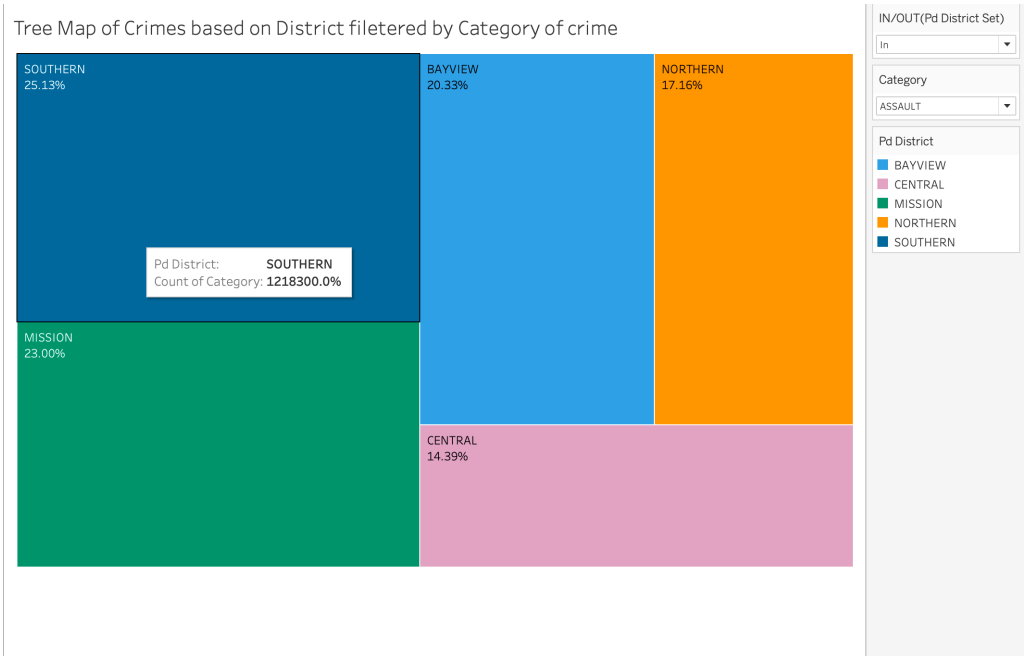
8. Tableau Dashboard

8.1. Data Visualizations

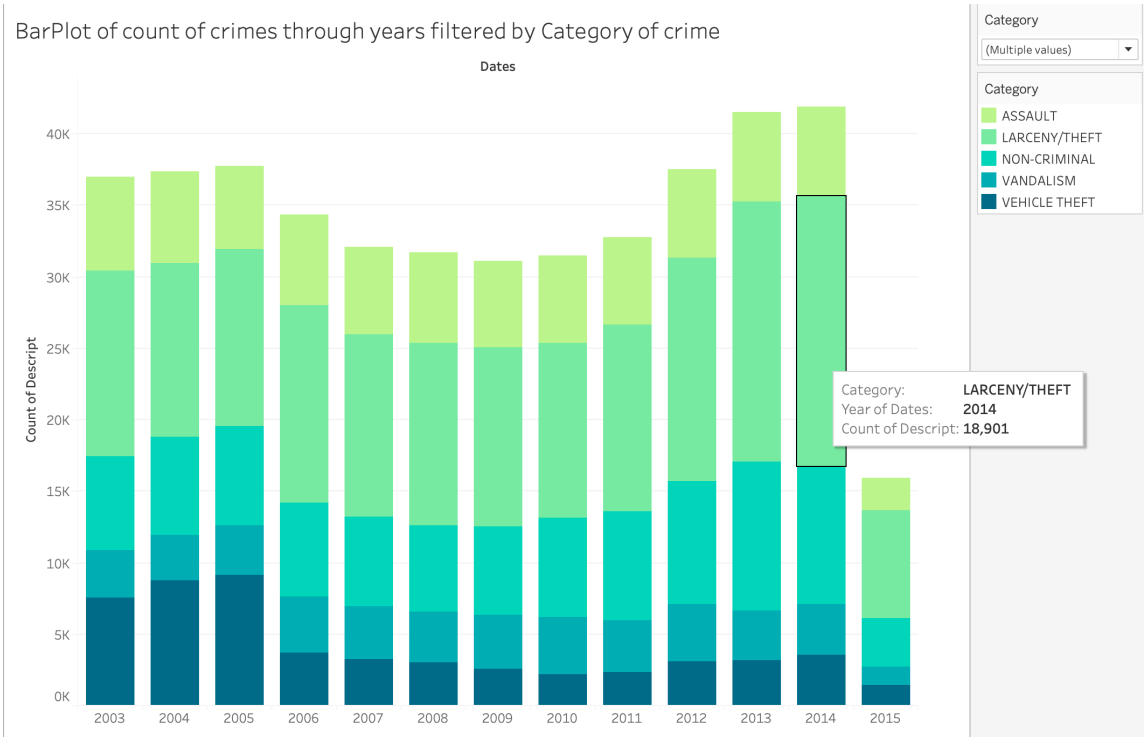
- Pie Chart



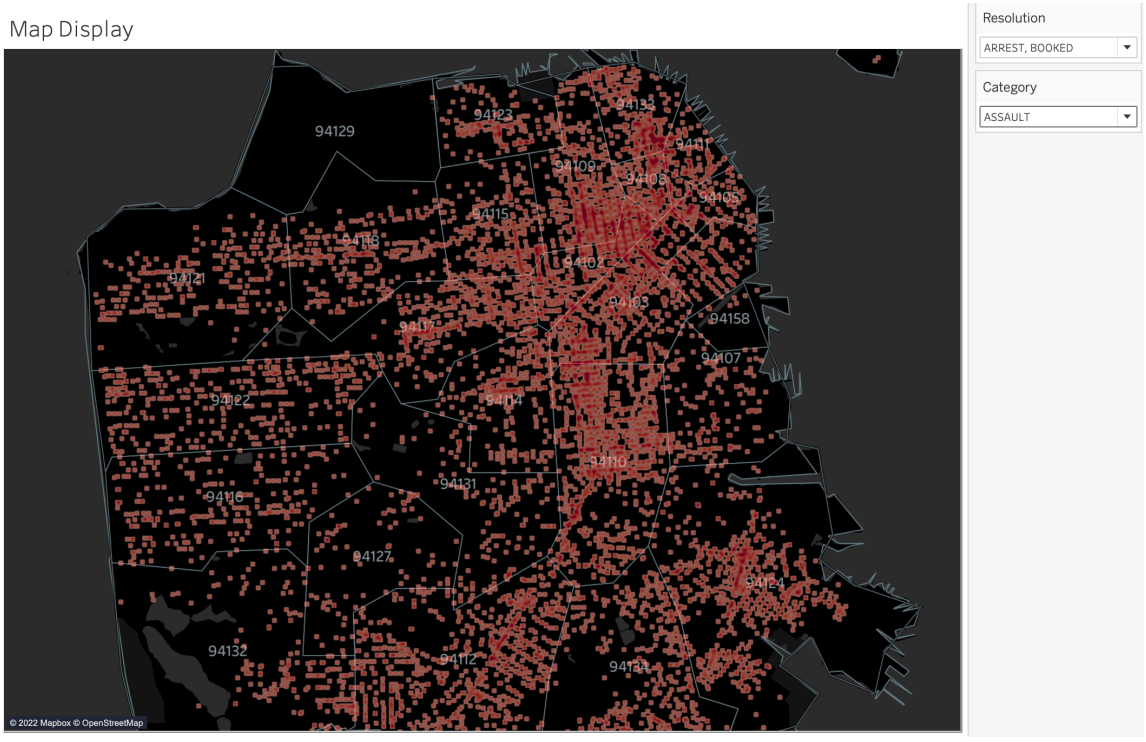
- Tree Map of crimes based on district filtered by category



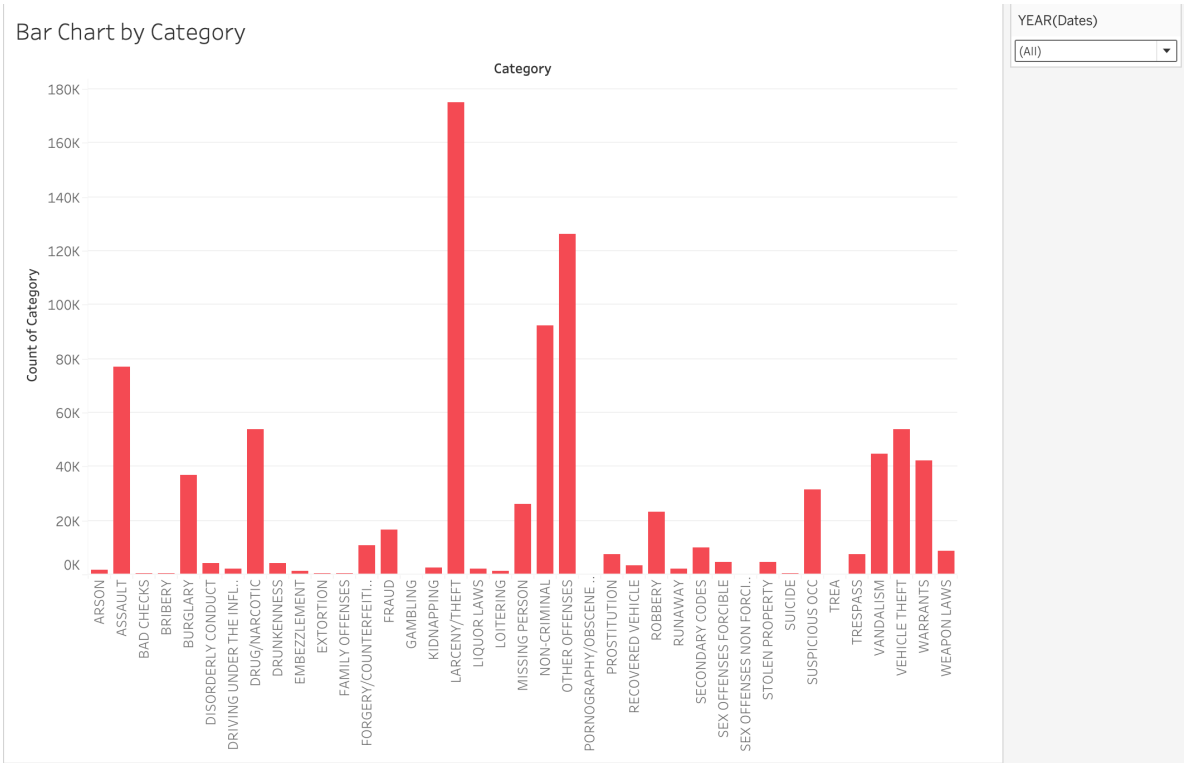
- Bar Plot of crimes through years by category



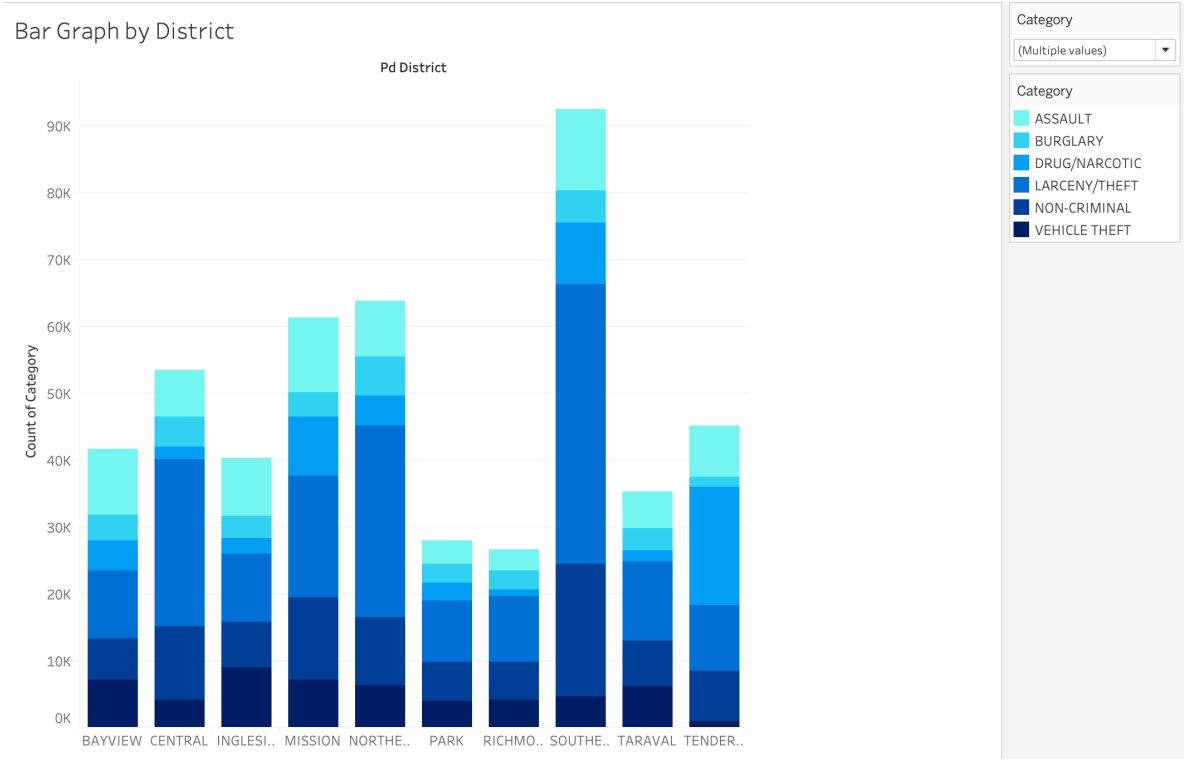
- Interactive Map filtered by Category and Resolution



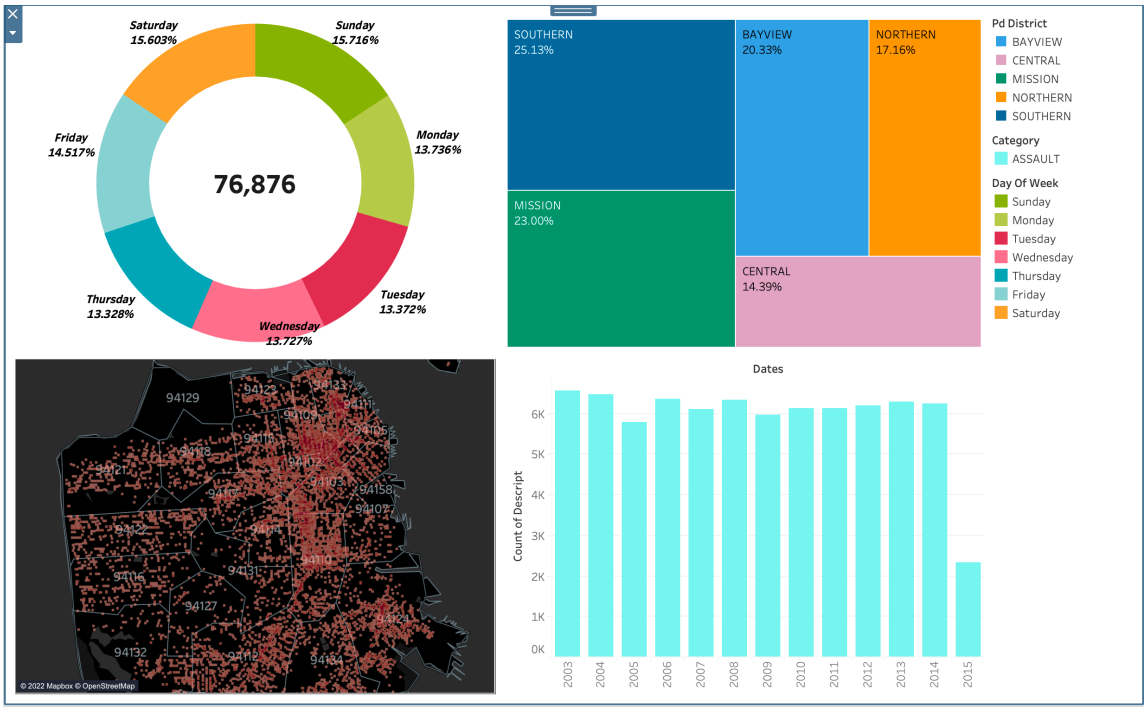
- Bar graph by category of crime

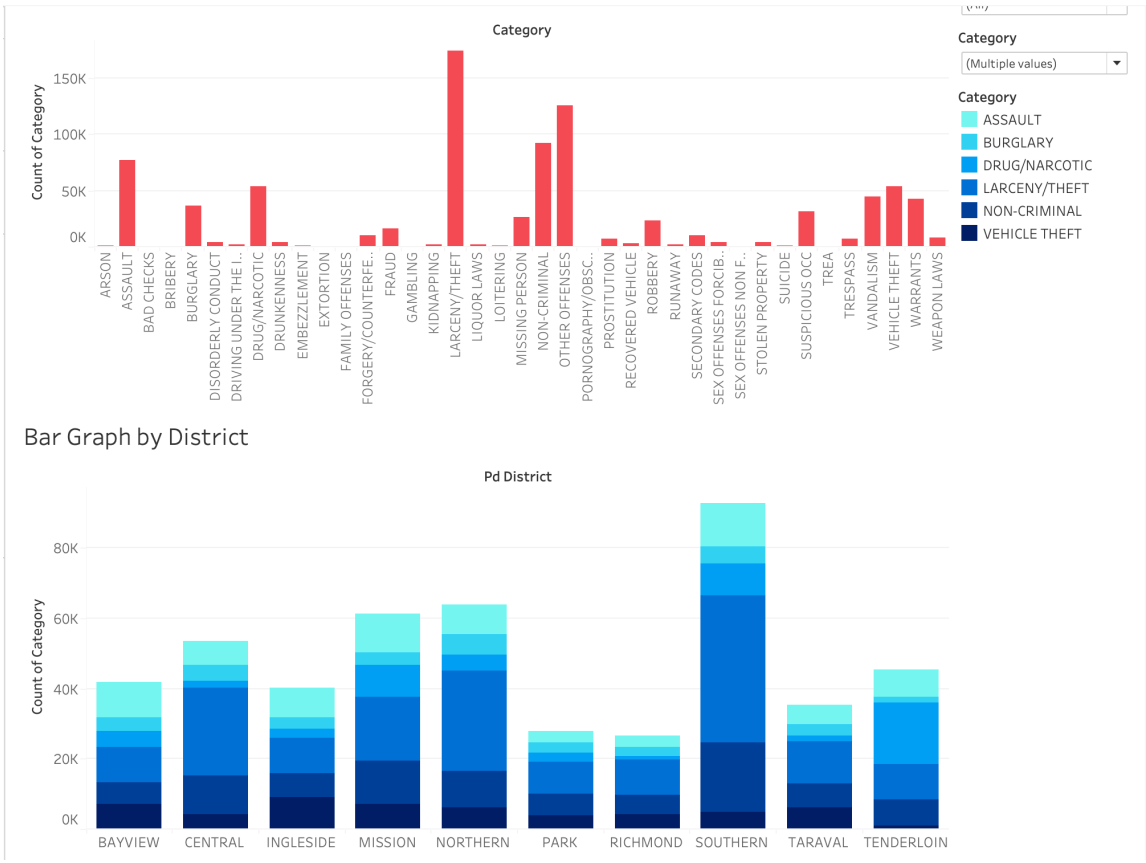


- Bar graph by district



8.2. Dashboard





9. Conclusions

- The age-old assumption that weekends bear witness to more crimes than the weekdays has been asserted but the margin has been very little as to the expected one.

- Comparisons of the area-wise density, district wide count of criminal activities suggest that Southern, Mission areas of the San Francisco were more vulnerable to criminal activities while Central, Bayview had the least susceptibility.

- Compared to 2003, the criminal activities have not been consistent where we saw a huge drop during 2010-2011 but saw augmenting during 2013.

- Over the years, Larceny/theft has consistently been the category with the highest count of activities.

- Most of the criminal activities have not been acted upon i.e., no arrests were made in these cases while some of them are arrested and cited.