San Francisco Crime Data Analysis

Coursera Data Science Capstone Project

Introduction

Background

Crime has always been a hot topic in different contexts, such politics and tourism. In this project, I'm going to analyze the publicly available crime data in San Francisco from June 2018 to present. The methodology I'm going to use is mainly exploratory data analysis techniques. For exploratory data analysis, I will categorize the data based on different features, such as grouping the data by neighborhoods, crime type, time of the day when the crime happens.

Audience

Many people will be interested in this report. Police will be interested in this report to see the distribution of crimes across different neighborhoods. And they could use this data to distribute the resources they have. Aslo, they could use the data to prevent crimes from happening by dispatching more police forces in hot spots.
Politicians will be interested in this report since they would like to provide a safer society for the public. They could use the report to estimate the resources they're going to put into the police.
Public people will be interested in this report because they would like to know about their neighborhood. And they would avoid going to hot spots where crime would likely to happen.
Also, tourists will be interested in this report since they would like to know where to stay and where to visit because San Francisco is a popular tourist spot.

Data

Data Source

The data I'm going to use is this publicly available data on sfgov website:https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783.
This data set itself has been clearly defined. All attributes are annotated on the website. The geo data is already available for most of the data.

Data Cleaning

After loading the data in Pandas, it is clear that there are missing values in the data set. Since we have an abundance of data(195,142 rows) and what we would like to achieve from this project is a general idea on how crime is distributed, it would be fairly safe for us to drop the rows that have NaN values. This will be reflected in the code in the Jupyter notebook.

Feature Selection

Some of the more important attributes are neighborhoods, date of the incident, day of the incident, time of the incident, and geo location, since they will be used for exploratory data analysis and clustering.

Exploratory Data Analysis

Categorize the data by neighborhood

One interesting aspect of the data is which neighborhood the crime happened. This is easy to calculate by the column “Analysis Neighborhood”. We show the top 10 neighborhood from the data.

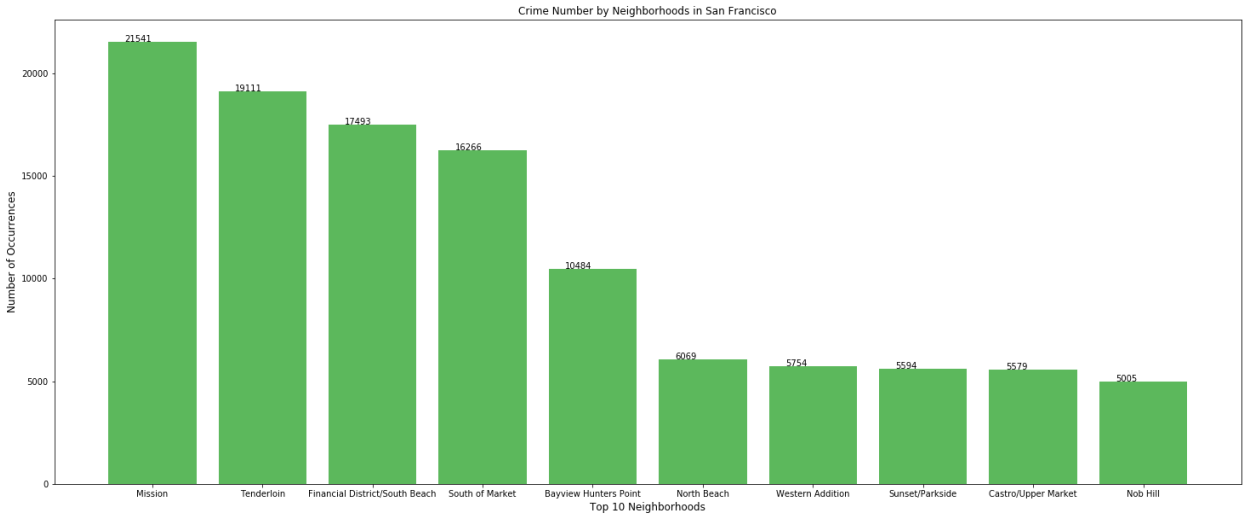

Figure 1 Top 10 Crime Neighborhoods of San Francisco

From the result, it is obvious that Mission has the highest crime, followed by Tenderloin, Financial District, SOMA and Bayview Hunters Point. Also, the top 5 neighborhoods have contributed to a large portion of the crime.

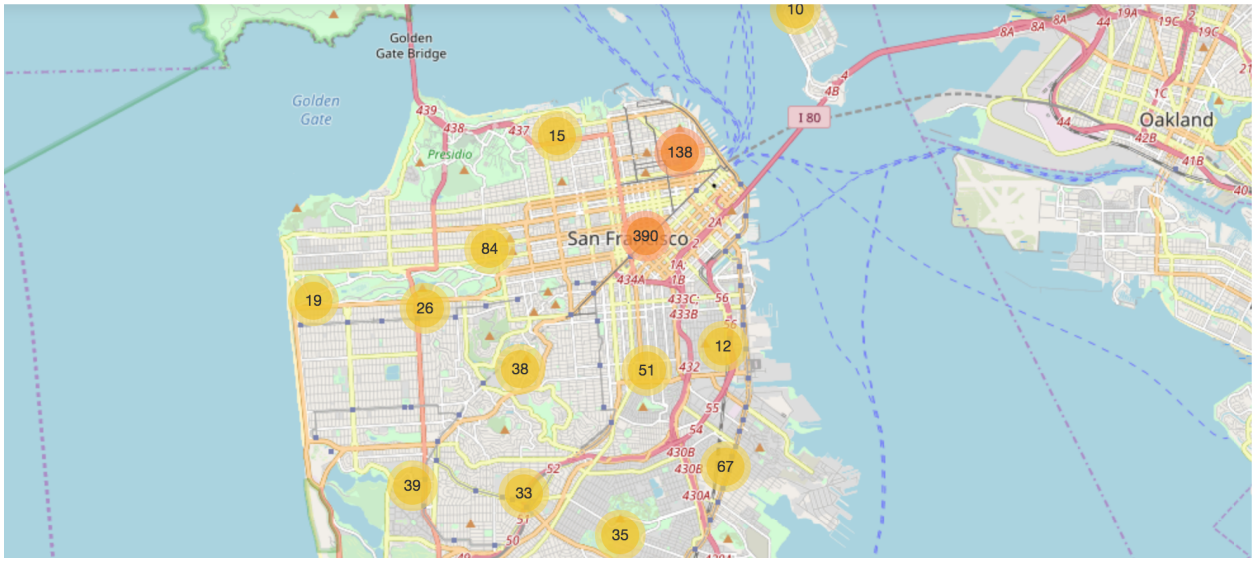

Figure 2 Sample 1000 crime data on map

From the sampled 1000 crime data, it is clear that Mission and Tenderloin have the top number of crimes, followed by Financial District.

Categorize the data by incident category

Another easy analysis we can do is to summarize the data by the category of the crime.

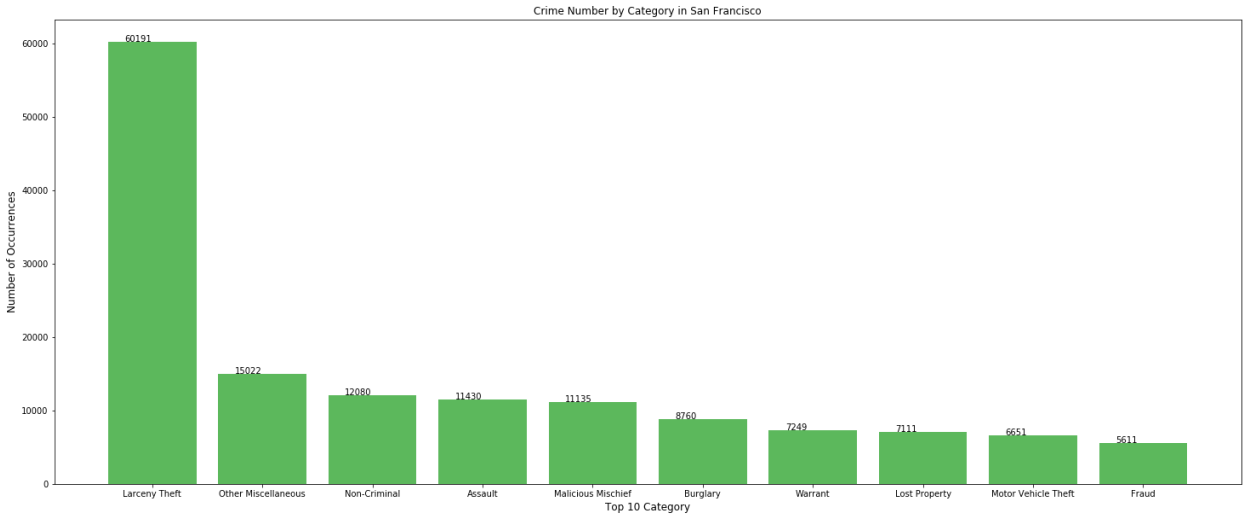

Figure 3 Top 10 crime category

From this plot, it is easy to spot that Larceny Theft has the highest number. Other Miscellaneous and Non-Criminal are general categories and they might include a lot of other sub-categories so we don't really count them as an independent category. And we could say that Assault is the second largest category, followed by Malicious Mischief.

Categorize the data by month

Some other information we could easily get from the data is whether the number of crime will fluctuate throughout the year.

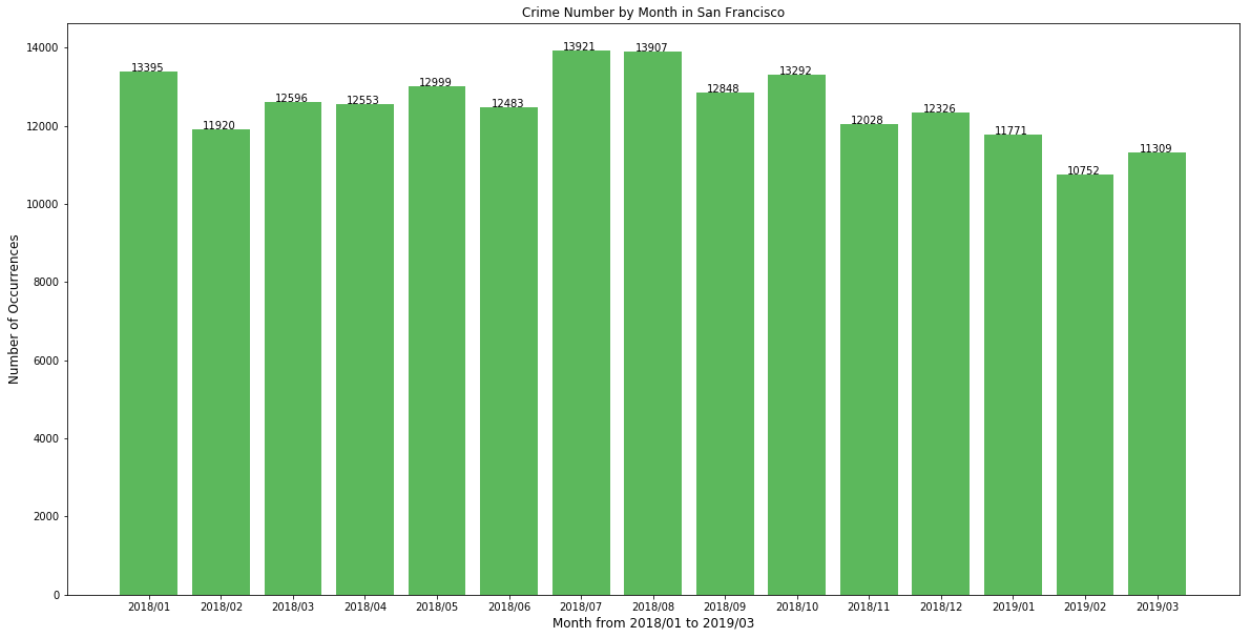

Figure 4 Crime Number by Month in San Francisco

It shows that in 2018, the months that have the most crime are July, August and Jan. It seems that in the summer and at the beginning of the year, there are more crime. Also, we have three months of data from 2019, and the trend of the first three month in 2019 follows the same pattern of that in 2018.

Categorize the data by day of week

Since the data has a column to document which day of the week the crime has happened, it is easy for us to get this information.

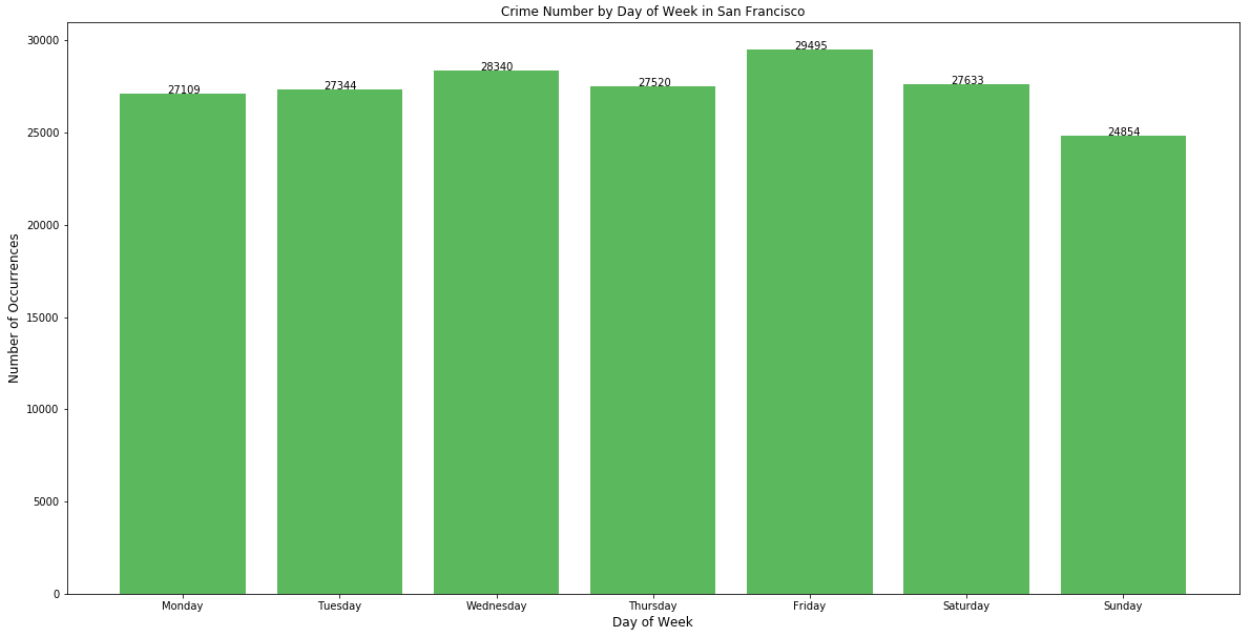

Figure 5 Crime Distribution by Day of Week in San Francisco

It is interesting to see that Friday has the most incidents during the week, which is not surprising either. From Friday to Sunday, the crime is on the downtrend. While from Sunday to Wednesday, it is going up.

Categorize the data by time of day

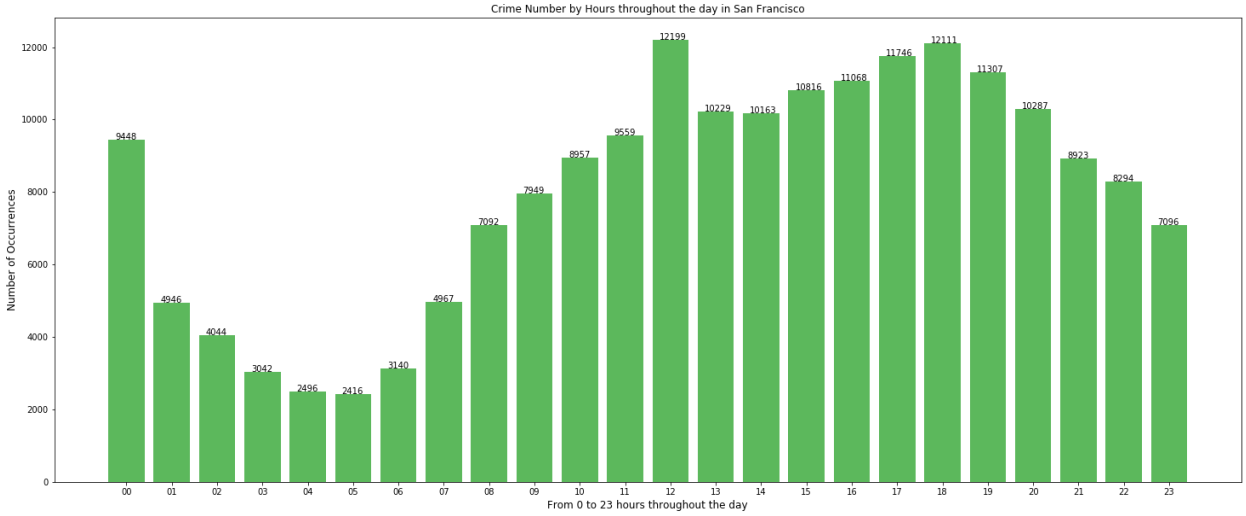

Figure 6 Crime trend throughout the day

This plot is very informative.

- From 5am in the morning to 12pm at noon, the crime number is steadily climbing up, to the peak at noon.
- Then it drops at 1pm, and goes up to peak at 6pm in the afternoon.
- Then it drops from 6pm to 11pm steadily.

- Then at midnight, it jumps again.
- After that, it steadily drops to 5am.
- And a new day starts.

Conclusion

From this report, we could get a clear idea from the data that

- Mission neighborhood has the highest crime number
- Larceny Theft outnumbers far more than other crime categories
- Summer tends to have higher crime throughout the year, followed by Jan
- Friday tends to have the highest number of crime throughout the week
- Crime number throughout the day has a clear trend