Analysis of 2003-2018 San Francisco Crime Data

I. Introduction

According to a recent report of "Safest Cities in the U.S. released by the WalletHub study on Dec 2, 2019, San Francisco has a safety score of 66.8, ranking 149th place in home & community safety rank, out of 182 major U.S. cities. NeighborhoodScout.com reported that San Francisco has one of the highest crime rates in America compared to all communities of all sizes. There are few cities in the world like San Francisco, that has so much wealth but at the same time such high crime rates. Its leading position in technology and mild weather has attracted talents and tourists from around the world, making it one of the most dynamic cities, both economically and culturally.

According to Open Data Network, both the population and population density have been increasing steadily since 2010 and are projected to continue to grow going forward. This report analyzes the trend of San Francisco crime over the years and aims to help people navigate and explore the city more safely. It can also be used by police for reference for patrolling purposes.

II. Data Acquisition and Cleaning

Our major crime data sets are comprised of 2 data sets from DataSF, an open data website of San Franscisco government. One is "Police Department Incident Reports: Historical 2003 to May 2018"; the other is "Police Department Incident Reports: 2018 to Present". The first data set has over 2 million rows of crime observations and 33 features, covering crime details such as incident number, type, description, date, time, address, location with longtitude and latitude, police district, and resolution etc.

The second data set has over 190,000 rows with more details, including report type, report time, filed online or not, incident subcategory.

We imported the 2 data sets into Jupyter Notebook separately and made a few observations. First, each crime has an incident number without any null values. However, they are not unique identifiers, as one crime incident may include different types of offenses. For example, a person who broke into someone's house may threaten someone in the house with a knife. This would be recorded in the crime database as 2 observations: burglary and assault. To ensure we don't count one crime incident more than once, I removed the duplicated rows by using drop_duplicates('col_name'), which will automatically keep the first row of duplicated ones. I keep that row as it typically indicates the major type of crime.

As the columns and column names of the 2 data sets are also different, we kept the common columns and changed some of the column names of the 2nd data set so we can merge the 2 datasets successfully. As the first data set include part of 2018 crime data, and second data set include part of 2019 data set, we used pd.to_datetime and dt.year to filter the overlapping period and remove the extra 2019 crime data. We used pd.concat to merge the data sets vertically, and saved it as a new csv file. The merged data set has over 1.8 million rows, covering crimes of 16 years from 2003 to 2018.

To identify which columns we need to drop, we used df.isnull( ).sum( ). If the column has an unreasonable number of null values, we drop them from our data set, such as “HSOC Zones as of 2018-06-05” column with 150,258 null values out of 194,356.

To explore factors that may have an impact on the number of San Francisco crimes, I also acquired data sets of San Francisco temperature, unemployment rate, homless population, income level.

I obtained the weather data from website of National Centers for Environmental Information. Due to limitation of the website, the weather data were downloaded in different batches from 2014 to 2018, with all locations in California. We used .loc and a loop function to extract city, date and average temperature from the dataset, and created a new DataFrame.

San Francisco monthly unemployment rate data set was obtained from Fred Economic Research from Jan 1990 to April 2019.

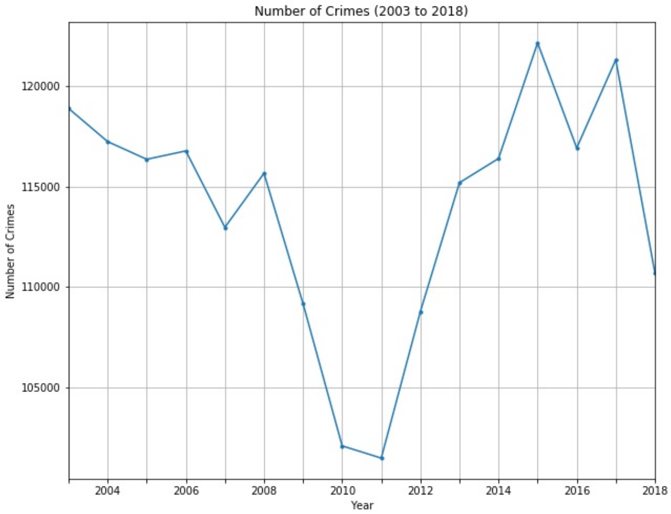
The homeless data set was obtained from “San Francisco Homeless Point in Time Count Reports” on Department of Homeless and Supportive Housing website. The data is collected every other year from 2005 till 2017.

To discover if the crime number is correlated with economic growth, I obtained the San Francisco GDP from 2006 to 2017 from Fred Economic Research and San Francisco median household income for the same time period from Department of Numbers website.

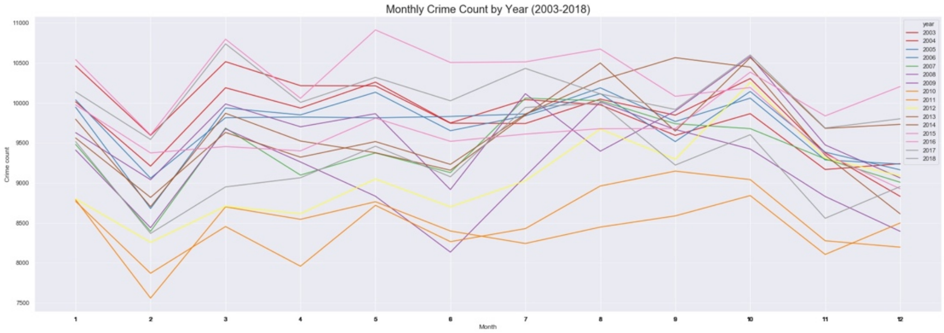
III. Exploratory Data Analysis

3.1 Timing of San Francisco Crime Data

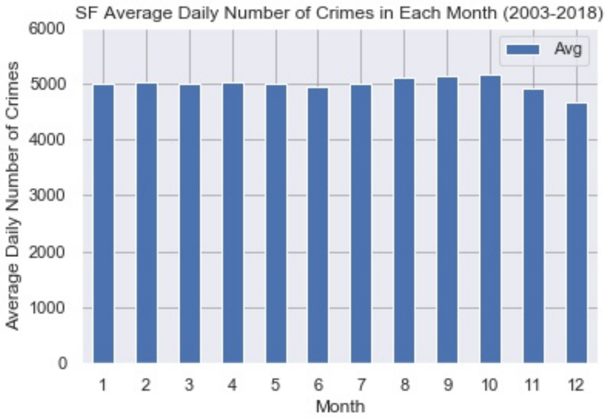
Our cleaned San Francisco crime data set from 2003 to 2018 has 1,822,068 crimes and 17 features. By applying pd.to_datatime( ) to the Date column, we plot the number of crimes by year as shown below. The graph shows a “V” shape of the trend. The annual number of crimes fluctuated before a drastic drop beginning in 2008, then reached bottom in 2011 before a sharp increase. The crime number started to trend down in 2017.



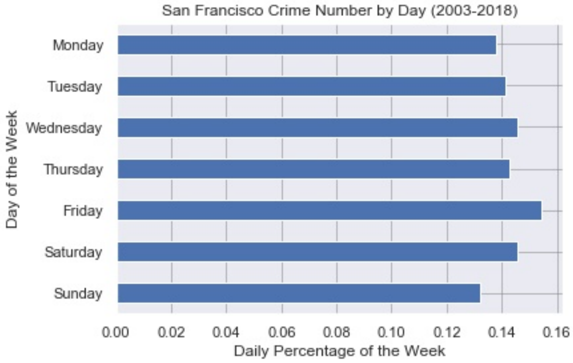
I also looked into the monthly trend by counting the number of crimes in each month, and found October, August and March are the top 3 months. In the graph below, we can see the number of crimes peak in those months in all years.



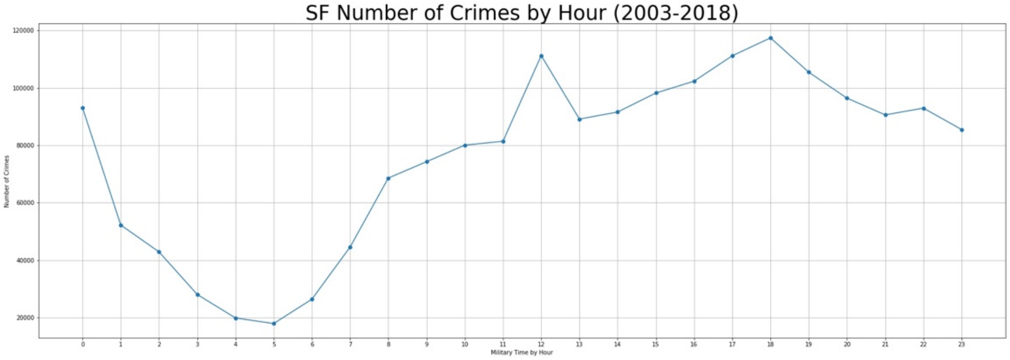
As each month has different number of days, we used numpy array to calculate the average daily number of crimes in each month. Results show that October, September and August have the highest average daily number of crimes, while the holiday months of December and November show the lowest.



DayOfWeek column records the day the crime happened. By using value_counts on this column, we found Friday has highest number of crimes of 281,015, followed by Saturday and Wednesday of 265,328 and 265,323 respectively. The days with least crimes are Sunday and Monday.



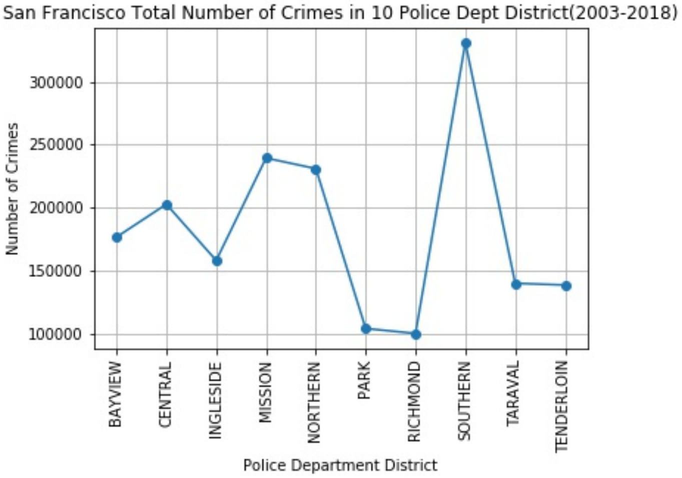
The Time column shows when the time happened by hour and minute in military time. I grouped the data set and plotted the number of crimes by hour. The line plot shows that number of crimes peaked at 12pm and 18pm each day and remained high throughout the day till midnight. The number of crimes is relatively low between 1am to 8am. It dropped to the lowest point at 5am.



3.2 San Francisco Crime Data Analysis by Area

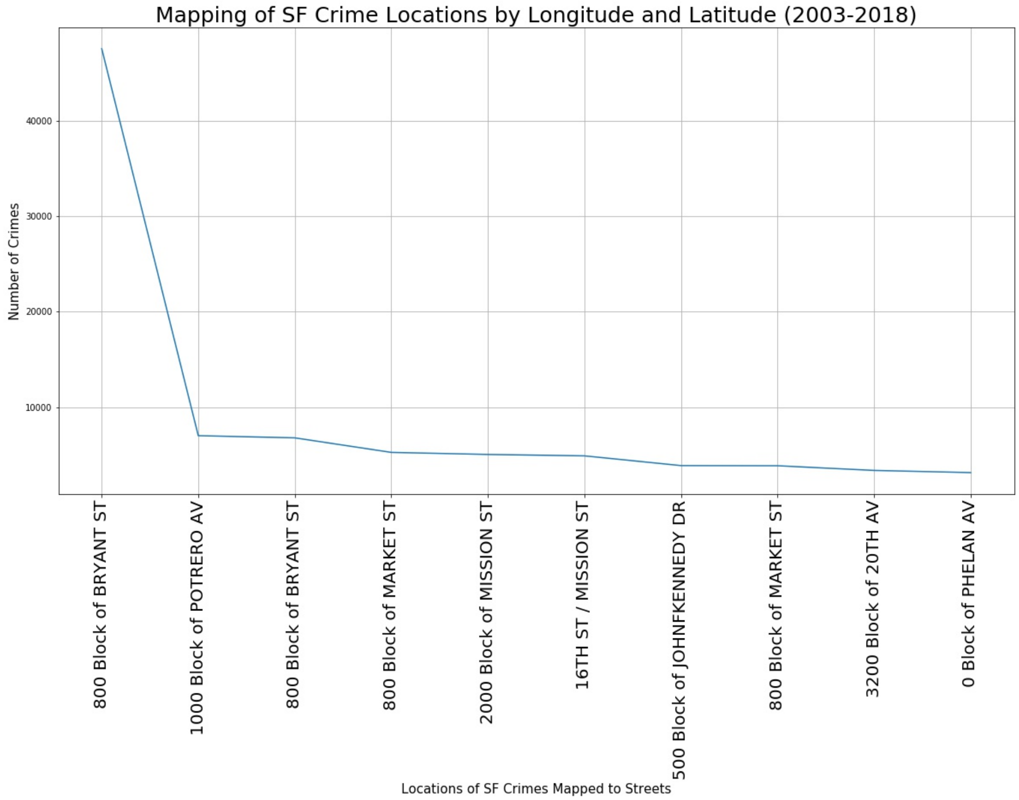
To see the pattern of crime locations, I used the columns of PdDistrict, Address and Location columns. There are 1 null value in PdDistrict and 6415 in Address and Location columns out of 1.8 million observations. Therefore, we consider these missing values immaterial to our analysis. By using df.corr( ), I discovered that these 3 columns is strongly correlated with the number of crimes, with all their correlation coefficiency of above 0.9.

I observed that the formats in PdDistrict and Address in the 2003-2017 and 2018 data sets are different. For example, the PdDistrict of MISSION is written as Mission in 2018 data set. In order to get accurate results for our analysis, I separated them into 2 data sets, and summed the total number of crimes by PdDistrict, then combined them by this column after making the format consistent. Result shows that there are most crimes in Southern, Mission and Northern police district. The districts with least crimes are Richmond and Park. Contrary to our common belief that Tenderloin is considered to be the unsafest neighborhood, the number of crimes is relatively low compared with other neighborhoods.

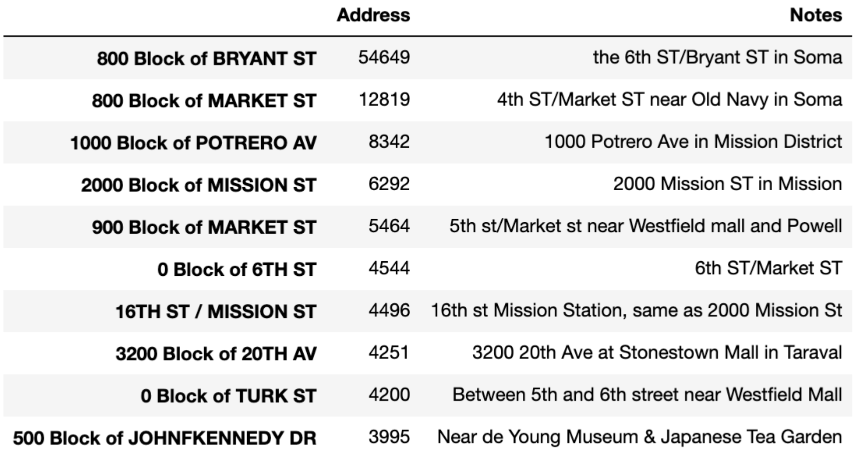


While the formats of streets of crime scenes in the “Address” column are inconsistent in 2 data sets, the “Location” column is recorded in consistent format by using longtitude and latitude in the 2 data sets. We kept both columns for accuracy in our analysis.

By mapping street names to the Location column and plotting this column, we discovered that even though some of the numbers of longitude and latitude are slightly different, they actually refer to the same locations. Therefore, the plot below shows the same location a few times. We can see that the number of crimes at 800 Bryant Street is significantly higher than any other locations. The second highest one is 1000 Potrero Ave in Mission District, followed by 2000 Mission St.

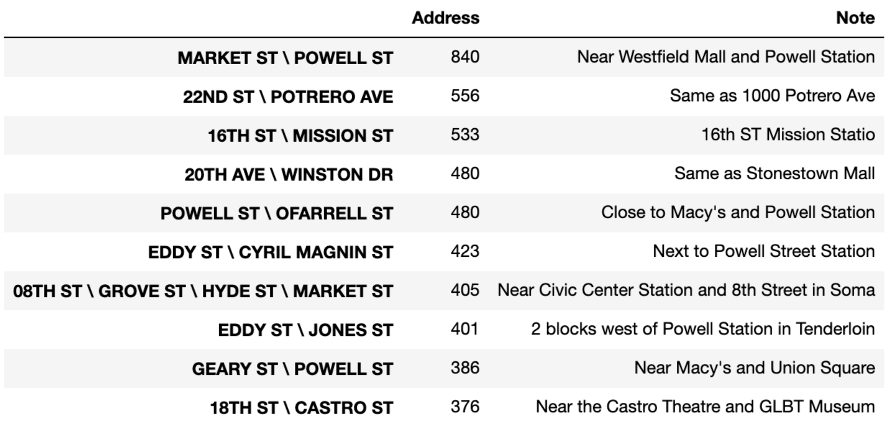


To confirm the above results with more details, we used value_counts( ) on Address column. As the formats of street addresses are different in the original 2 crime data sets, we performed value_counts( ) separately by filtering the merged data set by year. In our 2003-2017 data set, I listed the top 10 streets with most crimes shown, with a Notes column of describing the streets.



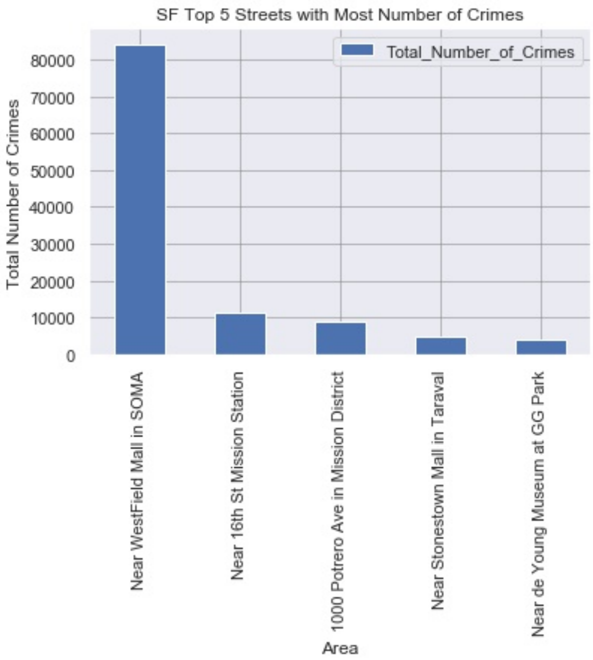
	Address	Notes
800 Block of BRYANT ST	54649	the 6th ST/Bryant ST in Soma
800 Block of MARKET ST	12819	4th ST/Market ST near Old Navy in Soma
1000 Block of POTRERO AV	8342	1000 Potrero Ave in Mission District
2000 Block of MISSION ST	6292	2000 Mission ST in Mission
900 Block of MARKET ST	5464	5th st/Market st near Westfield mall and Powell
0 Block of 6TH ST	4544	6th ST/Market ST
16TH ST / MISSION ST	4496	16th st Mission Station, same as 2000 Mission St
3200 Block of 20TH AV	4251	3200 20th Ave at Stonestown Mall in Taraval
0 Block of TURK ST	4200	Between 5th and 6th street near Westfield Mall
500 Block of JOHNFKENNEDY DR	3995	Near de Young Museum & Japanese Tea Garden

Below is a DataFrame of the top 10 streets with most crimes in 2018.



	Address	Note
MARKET ST \ POWELL ST	840	Near Westfield Mall and Powell Station
22ND ST \ POTRERO AVE	556	Same as 1000 Potrero Ave
16TH ST \ MISSION ST	533	16th ST Mission Statio
20TH AVE \ WINSTON DR	480	Same as Stonestown Mall
POWELL ST \ OFARRELL ST	480	Close to Macy's and Powell Station
EDDY ST \ CYRIL MAGNIN ST	423	Next to Powell Street Station
08TH ST \ GROVE ST \ HYDE ST \ MARKET ST	405	Near Civic Center Station and 8th Street in Soma
EDDY ST \ JONES ST	401	2 blocks west of Powell Station in Tenderloin
GEARY ST \ POWELL ST	386	Near Macy's and Union Square
18TH ST \ CASTRO ST	376	Near the Castro Theatre and GLBT Museum

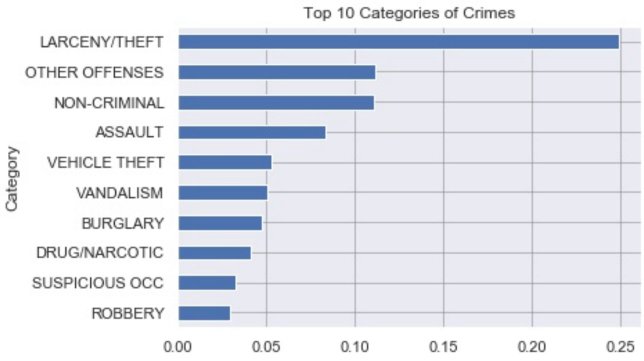
By observing the two lists, we found the results are similar to our location plot, that some of the streets are right next to each other or refer to the same streets. By adding up the number of crimes in the same area from 2003 to 2018, we ranked the top 5 areas as below:



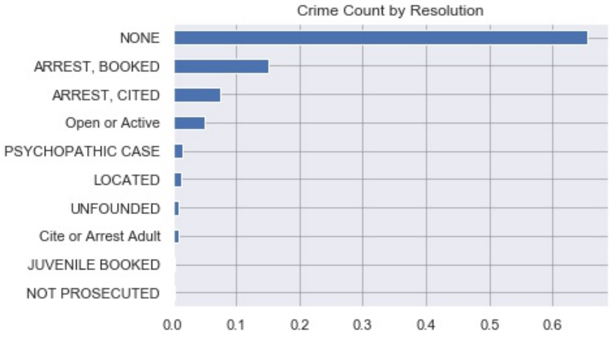
Corresponding to our location plot, the chart above also showed that the streets near Westfield Mall have significantly more crimes than any other streets, followed by the 16th Mission Bart Station. The shopping center Stonestown Mall, despite relatively far from downtown San Francisco, also have a relatively high number of crimes.

3.3 Categories and Description of San Francisco Crime Data

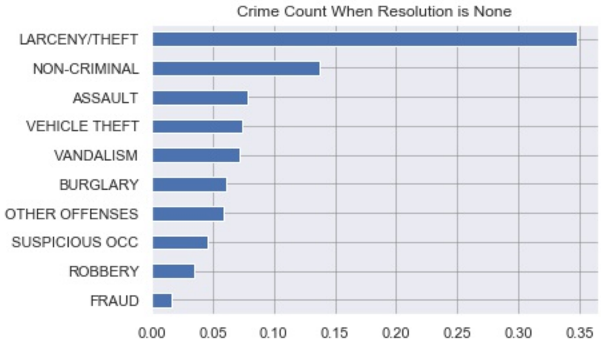
I grouped the data set by Category and found that the top 3 crimes are larceny/theft, other offenses and non-criminal, among which theft is way more common than any other type of crimes. As "other offenses" and "non-criminal" are very vague, I filtered the dataframe and took a closer look at the description. Most offenses are traffic related violations. Most non-criminal cases are lost property and aided cases. Other common crimes include assault, vandalism, and burglary etc.



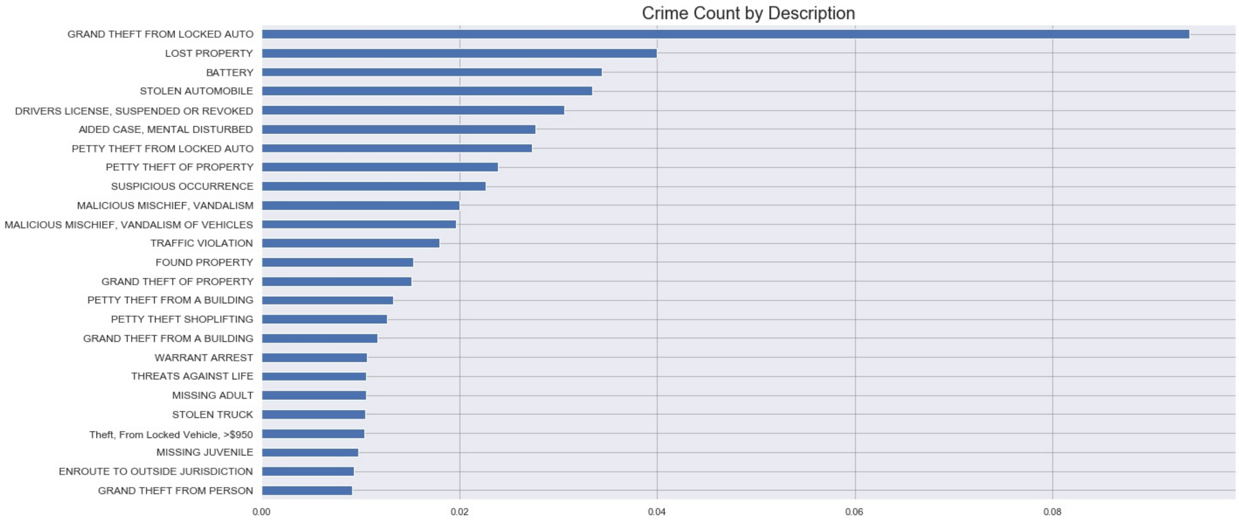
To see how these crimes are solved, I counted the number of crimes by resolution. Our result shows that the resolution for an astounding 65% of crimes is "NONE", which means the police appeared at the crime scene and recorded it, but didn't take action. About 22% of crimes leads to arrests: 15% to "arrests, booked"; 7% to "arrests, cited." Booking means a procedure of police taking the criminal suspect's personal information, recording information about the alleged crime, criminal background checking, fingerprinting, taking mug shots etc. A citation is a written notice issued for a misdemeanor, such as a traffic ticket.



As the number of crimes whose resolution is NONE accounts for a majority of incidents, I looked into what crimes types of crimes are mostly left unsolved. A value_count( ) shows that most of the crimes without any action taken are theft, non-criminal offenses, assault and vandalism.



To further explore nature of the crimes, we performed crime count by description. We plotted the top 25 types by description that accounts for over 50% of the crimes. The bar chart below shows that the top 5 are mostly vehicle related. An overview of the bar chart above, we can see theft from locked vehicle, theft of property, traffic violation, stolen vehicles, battery, vadalism are most common.

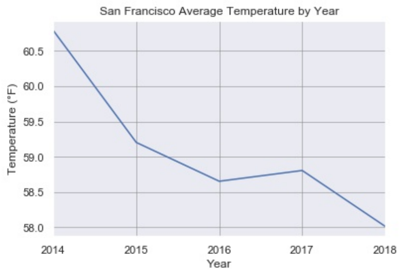


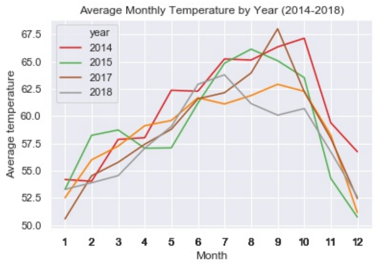
3.4 San Francisco Crime Data and Other Data Sets

To discover what factors would have an impact on crime number, we analyzed our crime data together with temperature, unemployment rate, homeless population and income.

3.4.1 SF Crime Count and Temperature from 2014 to 2018

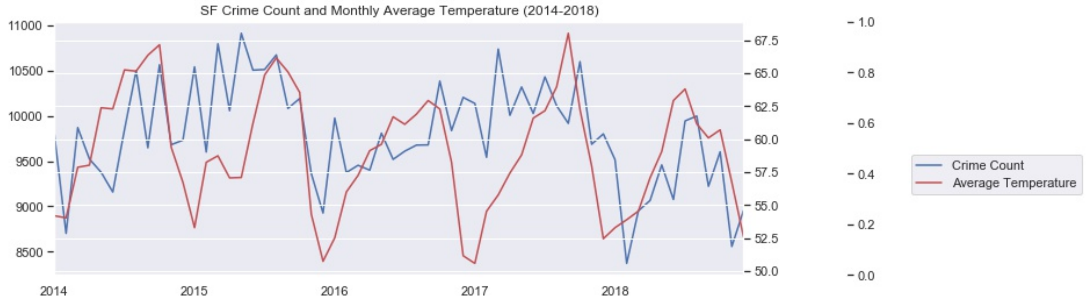
I collected daily average temperature in California from 2014 to 2018 and cleaned up the data set by keeping only the San Francisco average daily temperature over the available years. The annual average temperature shows a downward trend. The warmest time of the year is around September and October in fall. The coldest time of the year is winter around December and January.



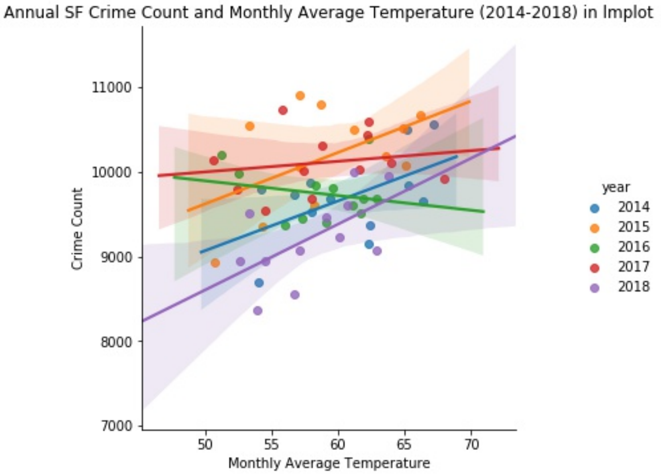


To take a closer look, I calculated the monthly average and combined the data set together with crime data set. Even though the monthly crime count fluctuates a lot, the plot shows that whenever the temperature peaks, the number of crimes would also reach the highest level. When temperature drops to bottom, the number of crimes would hit the lowest level.

The finding corresponds to our analysis about our daily average crime count by month over the years. Typically, November and December in San Francisco winter have lowest crime count; September and October, which are relatively warmer than any other months, have highest crime count.



To confirm the positive relationship between temperature and crime count, we plotted the combined data set by using sns.lmplot, and observed that they are positively related in all 4 years, except for 2016.



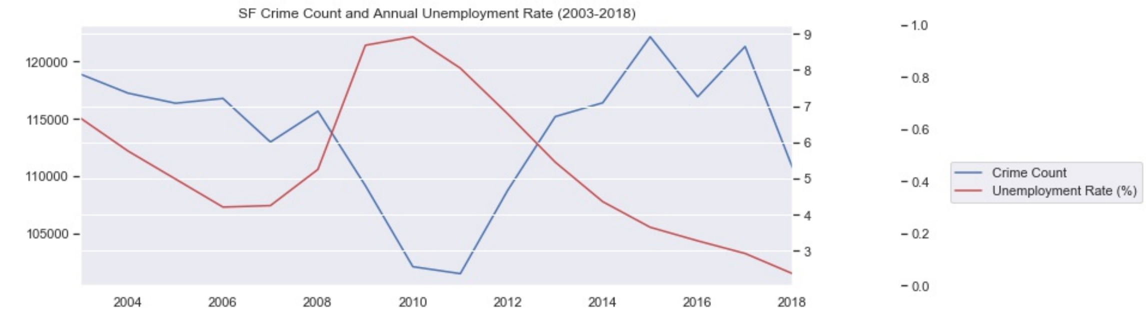
The above analysis shows that temperature is one of the important factors for the number of crimes.

3.4.2 SF Crime Count, Unemployment Rate (2003-2018) and Homeless Population (2005-2017)

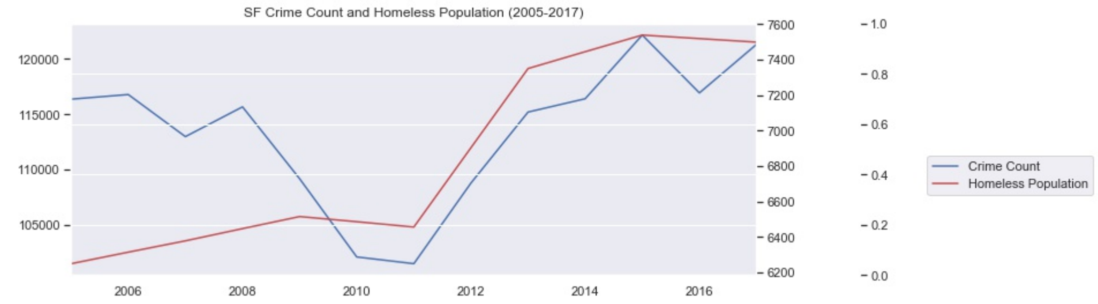
To see if SF crime count is correlated with population in the city, I obtained data sets of SF population, SF unemployment rate and homeless population over the years.

As SF population shows a steady increase over the years, it’s easy to tell that it’s irrelevant to our analysis, as our crime number fluctuates over the years.

I obtained monthly unemployment rate from 1990 to 2018 in San Francisco. By filtering the data set to include the period from 2003 to 2018, and calculating the average annual unemployment rate, I combined it with the crime data set. Our plot shows that in general, the lower the unemployment rate, the lower the number of crimes. However, from 2008 to 2012, which is around the economic downturn, the opposite seems to be true. Therefore, unemployment rate could be one of the factors at play, but we need to explore more other factors.



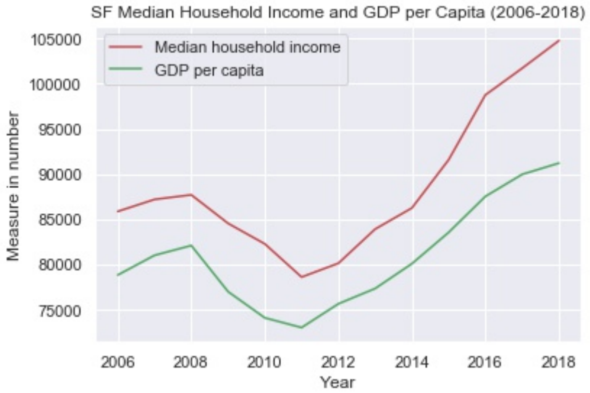
Homelessness is a huge issue in San Francisco. San Francisco counts its homeless population every two years. I collected the homeless data from 2005 to 2017. To combine the 2 data sets for analysis, I used middle point between the available 2 years to fill the missing years such as 2006 and 2008. Our graph shows that starting from 2009, the number of crimes is positively correlated with homeless correlation, and both numbers are trending upward after the economic downturn. The Pearson r value during this time period is as high as 0.94.



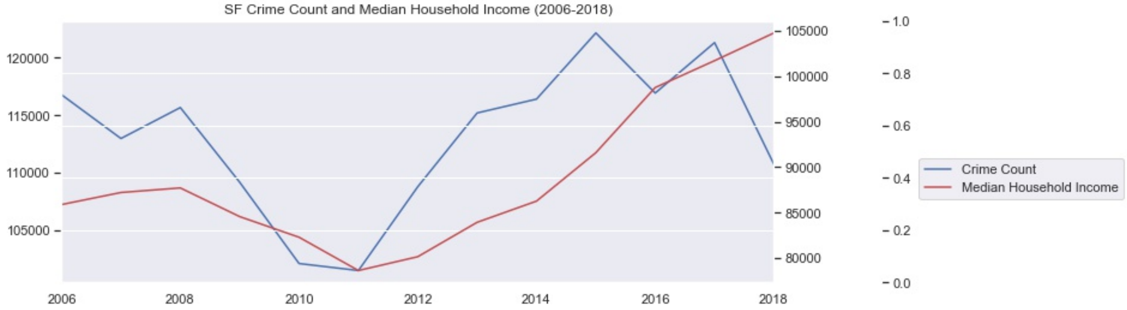
From above, we can see that both unemployment rate and homeless population both play a role in crime count.

3.4.3 SF Crime Count and Economy (2006-2018)

To explore the relationship of crime and economy, I obtained the SF median household income and GDP per capita data from 2006 to 2018. By plotting the 2 features, I found they show a similar trend. Therefore, I kept median household income for the analysis.



Before our analysis, I assume the higher people's income, the lower the number of crimes. However, our plot below suggests otherwise. Crime count hits bottom during recession.



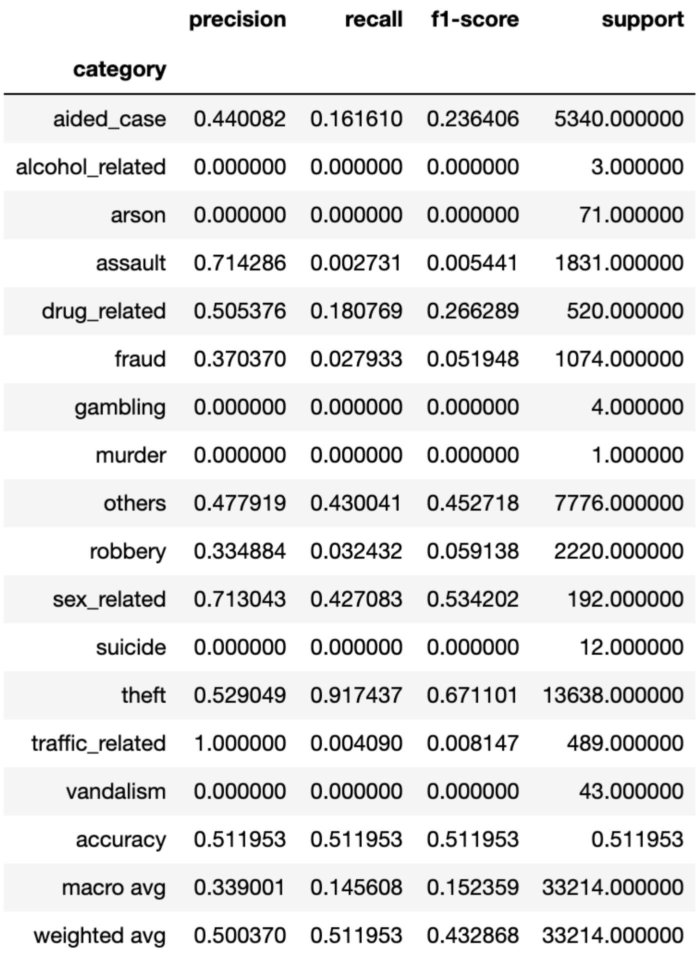
IV. Modeling

With the time, location and resolution of the crime, the dataset can use utilized for predicting the category of crime, which may help people navigate the city more safely. This s is multiclass classification problem. As there are inconsistency in naming the categories over the years, I renamed them and put some specific categories into more generic ones. For example, I would change "LARCENY/THEFT", "STOLEN PROPERTY" and "VEHICLE THEFT" all to a more generic category of "theft". There are 17 types of crimes in total after I cleaned this target column.

As all categorical features must be numerical to be used as input in the model, I used df.replace( ) and LabelEncoder to turn them into numbers, including resolution, PdDistrict, SF Neighborhoods, latitude and altitude. I also used datatime to separate dates into different columns of day, month and year etc. Due to the size of the dataset, I used the most recent year 2018 crime data for modeling.

I split the dataset into 70% of training data and 30% of testing data. XGBoost was utilized because of its high execution speed and better model performance in general. To use this model for multiclass classification, I specified the objective as “multi:softmax” in creating the classifier. I used GridSearchCV and a 3-fold cross-validation for optimal parameters in the model. By using the tuned parameters, the classifier achieved a score of 51.2% in accuracy.

Below is the classification_report for the model performance. In terms of precision, that is the ratio of correctly predicted positive observations to the total predicted positive observations, the model performs best in predicting traffic related, assault and sex related crimes. However, the recall score is very low for traffic related and assault. Recall is the ratio of correctly predicted positive observations within the category. The top 3 categories in recall score are theft, others and sex related. The f1 score, which is the weighted average of precision and recall, is highest for theft, sex-related and other crimes.



	precision	recall	f1-score	support
category				
aided_case	0.440082	0.161610	0.236406	5340.000000
alcohol_related	0.000000	0.000000	0.000000	3.000000
arson	0.000000	0.000000	0.000000	71.000000
assault	0.714286	0.002731	0.005441	1831.000000
drug_related	0.505376	0.180769	0.266289	520.000000
fraud	0.370370	0.027933	0.051948	1074.000000
gambling	0.000000	0.000000	0.000000	4.000000
murder	0.000000	0.000000	0.000000	1.000000
others	0.477919	0.430041	0.452718	7776.000000
robbery	0.334884	0.032432	0.059138	2220.000000
sex_related	0.713043	0.427083	0.534202	192.000000
suicide	0.000000	0.000000	0.000000	12.000000
theft	0.529049	0.917437	0.671101	13638.000000
traffic_related	1.000000	0.004090	0.008147	489.000000
vandalism	0.000000	0.000000	0.000000	43.000000
accuracy	0.511953	0.511953	0.511953	0.511953
macro avg	0.339001	0.145608	0.152359	33214.000000
weighted avg	0.500370	0.511953	0.432868	33214.000000

V. Conclusions

Our analysis shows that the number of crimes peaks in October, September and August, which are the warmest months in San Francisco; December and November, which is around the coldest time of the year, have least crimes.

The number of crimes is highest on Friday, Saturday and Wednesday. The days with least crimes are Sunday and Monday. During each day, the number of crimes is higher during daytime than nighttime. It peaks at 12pm and 18pm and remains relatively low from 1am to 8am. At 5am, the number of crimes drops to the bottom.

Among 10 police districts, South and Mission have highest number of crimes. The districts with least crimes are Richmond and Park. Tenderloin, which is commonly believed the most unsafe district, ranks the 3rd safest neighborhood after Park. Areas around shopping malls or scenic spots that attract huge crowd typically have more crimes. Streets near Westfield Mall have significantly more crimes than any other streets, followed by the 16th Mission Bart Station, 1000 Potrero Ave in Mission and Stonestown Mall in Taraval.

The most common crime types are larceny/theft, other offenses and non-criminal. Most offenses are traffic related violations. Most non-criminal cases are lost property and aided cases. No actions were taken for a staggering 65% of crimes, of which mostly are theft, non-criminal offenses, assault and vandalism.

Based on description of crimes, most are vehicle related, including theft from locked vehicle, traffic violation, stolen vehicles. Other common crimes include theft of property, battery, and vandalism.

By studying the trend of San Francisco crime count over the years with other data sets, I discovered that the homeless population and people's income are highly correlated with the number of crimes. In general, the higher the homeless correlation and people's income, the higher the number of crimes.

By using the above features such as time, location and resolution, I used XGBoost for prediction of crime category, and achieved 52.8% accuracy. The result can be improved if the model runs on gpu so that we can fit more years of historical data into the model, or if we add more categorical features such as temperature and homeless data.

There are other limitations to this analysis. We could explore more about the types of crimes based on locations and time. The categories of crime can be recorded in a more consistent manner, such as "Stolen Property" and "Larceny/Theft" are the same. It would be helpful to our analysis if there's subcategory of crimes. The accuracy of our analysis of correlation between homeless population and crime count can be improved if homeless population count is performed yearly instead of every other year.