SAN FRANCISCO CRIME RATE ANALYSIS
ISYS 812, Summer 2020

Leo Iskandar, Huimin Jiang, and Steven Lu

June 3, 2020

TABLE OF CONTENT

Introduction

The dataset contains the information about the crime rates in different regions of San Francisco in 2016 with other important aspects related to crime, such as the category, the description of crime, the time and date of crime, and the resolution of the criminal case.

Dataset's Name: San Francisco Crime Dataset (On Kaggle: "Sanfrancisco Crime Dataset")

Short Description of Columns:
IncidentNum: Case Number
Category: Type of incident; weapon laws, warrants, assault, etc.
Descript: Description of incident
DayOfWeek: Which day of the week the incident occured
Date: Calendar date of occurrence
Time: Time of occurence
PdDistrict: Police Department's District/Location
Resolution: Kind of punishment given/results of case
Address: Address of incident
X: Latitude of the crime location
Y: Longitude of the Crime
Location: Exact Location
PdId: Police Officer's ID

Number of Rows, Columns, and Missing Values:
Columns: 13
Rows: 150501
Missing Values/NaN/Nulls: 1 Missing value in PdDistrict

Why we chose this dataset:

Cities in the Bay Area are home to many start-up and global technological companies. As there are a lot of international students who pursue their education and career goals in San Francisco, it is considered important to have sufficient understanding of different neighborhoods in this city.

According to the Neighborhood Scout website, San Francisco is deemed as one of the highest crime rate cities in the United States. Based on 2018 data, the crime rate of this city is 64 per one thousand residents. Moreover, more than 99 percent of the communities in California have a lower crime rate than San Francisco. Therefore, given the characteristics of this city, it is beneficial to have more knowledge and awareness about the conditions of its neighborhoods.

Research Statements

Driving Question:

By using the information of crime records from the San Francisco Police Department, how to stay safe in San Francisco?

Sub-questions:

- 1: Which district had the highest incident rates? For this district, which category had the highest incident rates?
- 2: Which category of incident was the largest contributor to crimes in San Francisco?
- 3: What days of week had the highest crime rates? Which month of the year had the highest crime rate? Which time period of the day had the highest time rates?
- 4: What was the percentage of incident cases that did not have any resolution? What are the categories of crime that didn't have any resolution?
- 5: What category of crime had the most unresolved issues? For that category of crime, what was the incident that happened the most?

Data Cleansing Procedures and Documentations

Reading Data and Saving to New Dataframe

The first step before modifying and altering our dataset was to understand our dataset and its variables, its metrics, and what each variable means. After understanding our data set, we needed to import the necessary libraries to begin the cleansing part of the data cycle.

Modifying Missing Values

We then continued to analyze the data frame by looking at the data types of each variable and searched for missing values. We discovered that we had 1 missing value in our 'PdDistrict' column. After close investigation, we were able to analyze that observation and found which district the address of that crime correlated to. We then filled that missing value, using the forward filling method since the observation above it had the same value.

Adding New Columns and Splitting Values

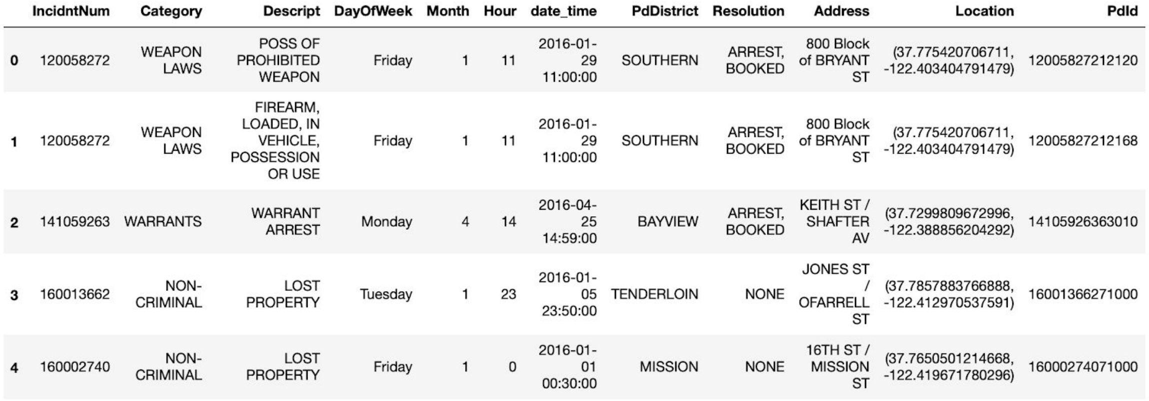
After filling all missing values, we needed to modify the 'Date' Column which is a series with string values. The original data in this column included the date as well as the time. The date showed the exact date when incidents happened, while the time was shown as '00:00:00'. Therefore, the time values in this column were not accurate. However, the data set included another column called 'Time' that included the accurate time of incident. Both columns were string type.

Therefore, We needed to first split string values in the 'Date' column and grab the accurate date value. Then, combine the accurate date values in the 'Date' column with the accurate time values in the 'Time' column, and then convert the combination result to a datetime type for future extractions. A new column called 'date_time' was created to save the datetime type values.

Once the date and time type were converted to datetime, we needed to extract the month and hour of the incidents for future analysis. We created two additional columns, one is called 'Month' to extract month elements from the series 'date_time', and another is called 'Hour' to extract the hour elements. Both columns were relocated to the partnering date variables for easier access and analysis.

Removing Columns

Finally, once the new variables were added, we needed to remove all variables not applicable to our analysis and were duplicates. Which includes the individual 'Date' and 'Time' variables, as well as the 'X' and 'Y' variables which were duplicates of the 'Location' variable. Results of the beginning five observations can be seen below [Figure 1.0].



	IncidntNum	Category	Descript	DayOfWeek	Month	Hour	date_time	PdDistrict	Resolution	Address	Location	PdId
0	120058272	WEAPON LAWS	POSS OF PROHIBITED WEAPON	Friday	1	11	2016-01-29 11:00:00	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	(37.775420706711, -122.403404791479)	12005827212120
1	120058272	WEAPON LAWS	FIREARM, LOADED, IN VEHICLE, POSSESSION OR USE	Friday	1	11	2016-01-29 11:00:00	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	(37.775420706711, -122.403404791479)	12005827212168
2	141059263	WARRANTS	WARRANT ARREST	Monday	4	14	2016-04-25 14:59:00	BAYVIEW	ARREST, BOOKED	KEITH ST / SHAFTER AV	(37.7299809672996, -122.388856204292)	14105926363010
3	160013662	NON-CRIMINAL	LOST PROPERTY	Tuesday	1	23	2016-01-05 23:50:00	TENDERLOIN	NONE	JONES ST / OFARRELL ST	(37.7857883766888, -122.412970537591)	16001366271000
4	160002740	NON-CRIMINAL	LOST PROPERTY	Friday	1	0	2016-01-01 00:30:00	MISSION	NONE	16TH ST / MISSION ST	(37.7650501214668, -122.419671780296)	16000274071000

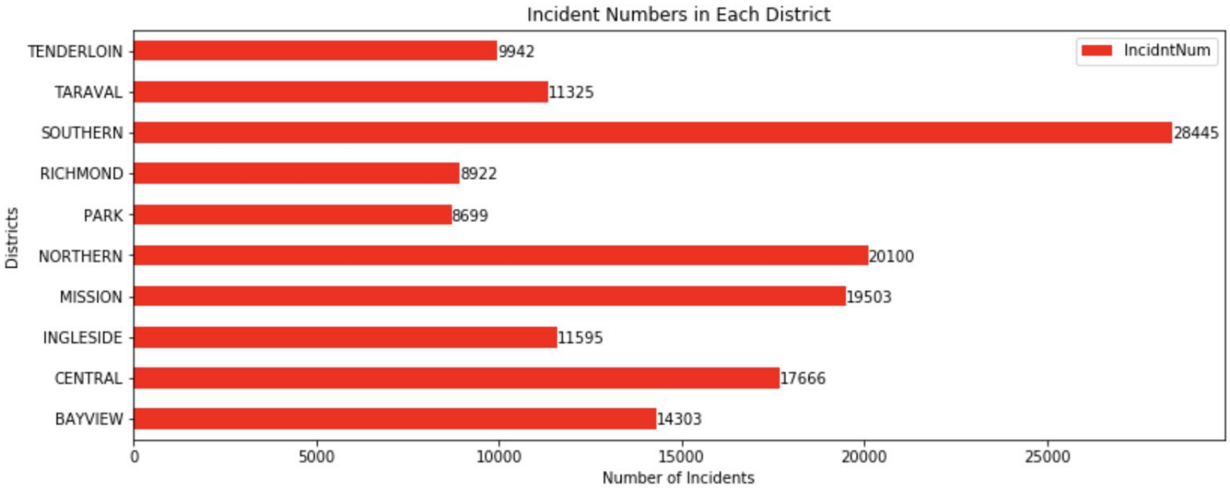
[Figure 1.0] First Few Rows of Data Set

Data Analysis

<u>Sub Question 1:</u>

a. Which district had the highest incident rates?

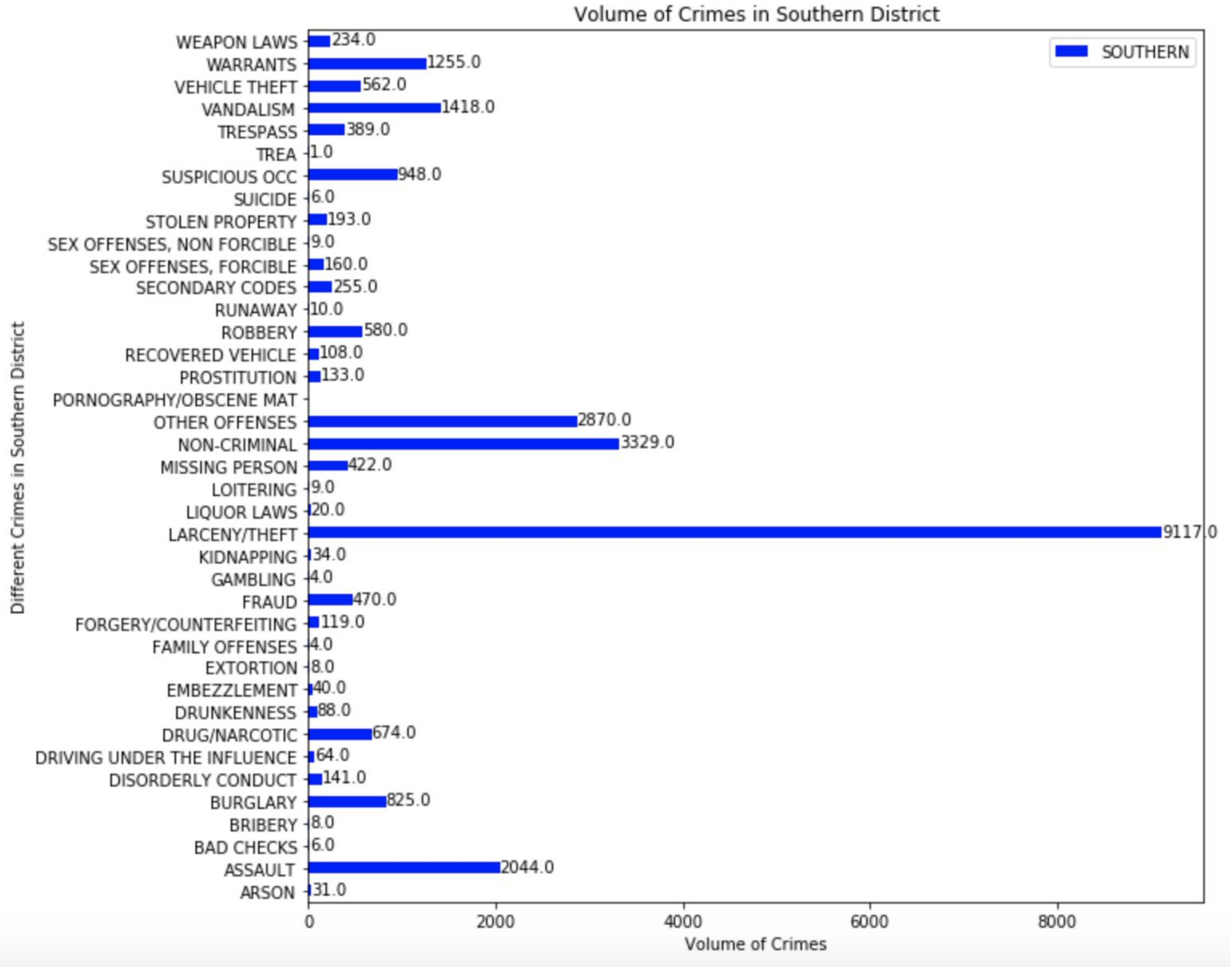
For our first question, we were interested in understanding which district had the highest crime rate and the potential ripple effects it causes in these districts. After close analysis, we were presented with the figure below [Figure 1.1]. The results show that the Southern district had the highest count of incident numbers at an astonishing volume of over 28,000 incidents in just 2016 alone.



[Figure 1.1] Incident Numbers in Each District

b. For this district, which category of crime had the highest incident rates?

After the discovery that the Southern District is the most incident dense area, we decided to further our analysis and focus on the types of crimes that occur in this area. At over 9,000 cases, larceny and theft is the most recurring crime in this district [Figure 1.2].



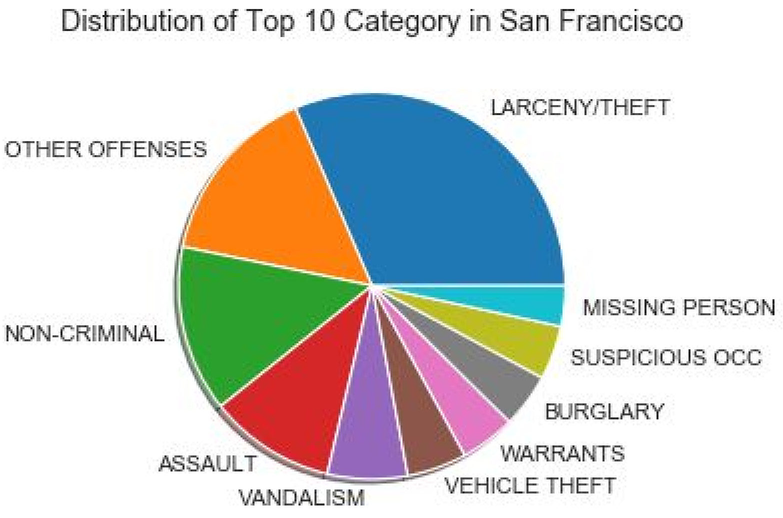
[Figure 1.2] Volume of Different Crimes in Southern District

For our first question, we came to the conclusion that the Southern District had the highest incident rates with larceny and theft being the largest contributing crime.

Sub Question 2:

Which category of incident was the largest contributor to crimes in San Francisco?

In our dataset, there are 39 crime incident categories. To better understand the second question and make our analysis more visualized, we analyzed the top 10 number of crime incident categories. As shown in Figure 2, we found that 'Larceny / Theft' was the largest contributor to crimes in San Francisco, whose percentage was twice the second largest contributor 'Other offenses'.

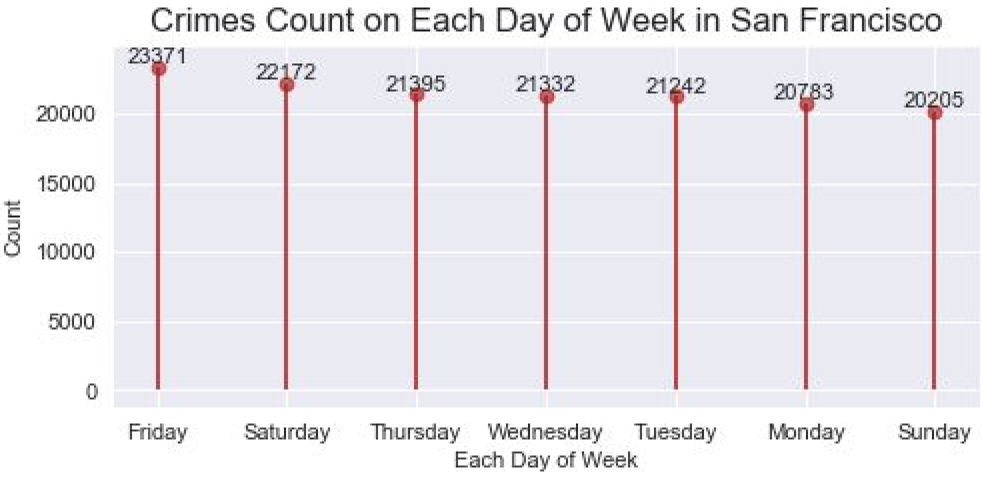


[Figure 2] Distribution of Top 10 Category in San Francisco

Sub Question 3:

a. Which day of week had the highest crime rates?

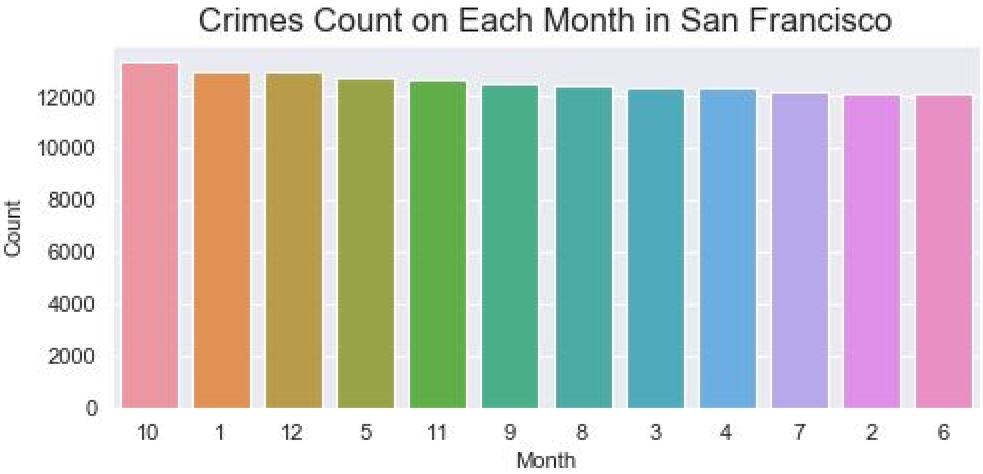
To figure out which day of week had the highest crime rates, we counted the number of crime incidents on each day during 2016. The result showed that there was not much difference in the number of crimes among each day. While crime incidents happened mostly on Friday. One reason we speculated is that Friday was the last day of workday where people prefer shopping or other outdoor activities than other days, which offered a great chance for thefts.



[Figure 3.1] Crimes Count on Each Day of Week in San Francisco

b. Which month of the year had the highest crime rate?

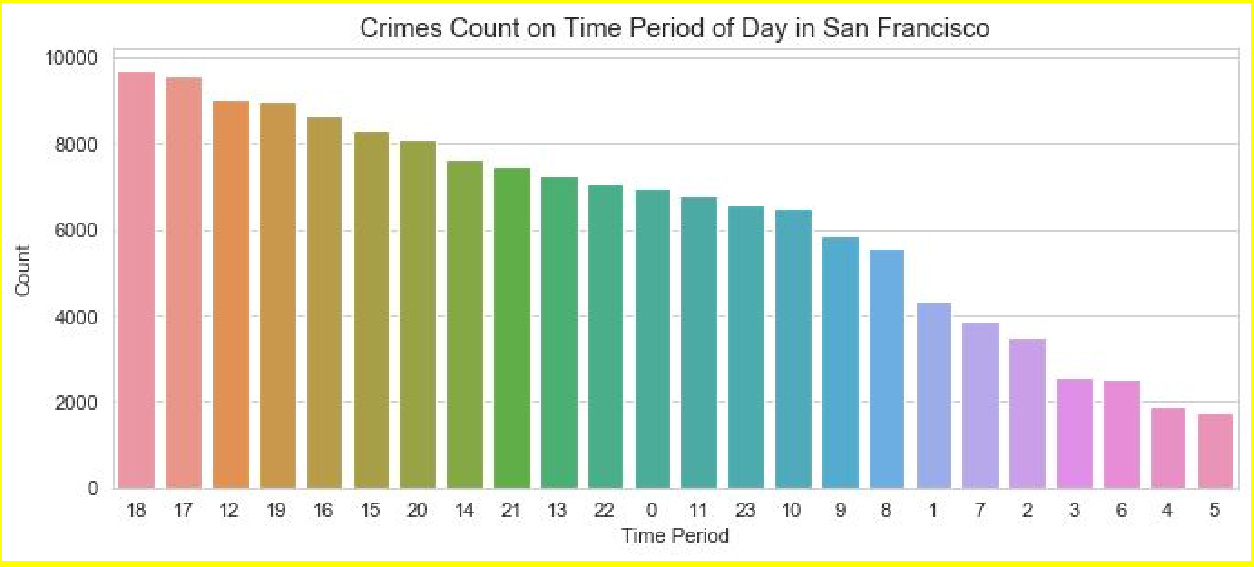
To analyze the number of incidents in each month, we modified data grouped by month vector. As shown in Figure 3.2, the amount of crime incidents that happened during each month in 2016 were close to each other, which were all over 12000 incidents. However, October has the highest count of crimes.



[Figure 3.2] Crimes Count on Each Month in San Francisco

c. Which time period of the day had the highest time rates?

We grouped data by the variable 'Month' to analyze during which period of day the crime incidents happened most frequently. After visual analysis, we found that the periods when crime incidents occurred more often were from 10 am until midnight. Most obvious periods were from 12 pm to 1 pm and 5 pm to 7 pm which were rush hours. However, 6 pm to 7 pm were when the incidents happened most frequently.

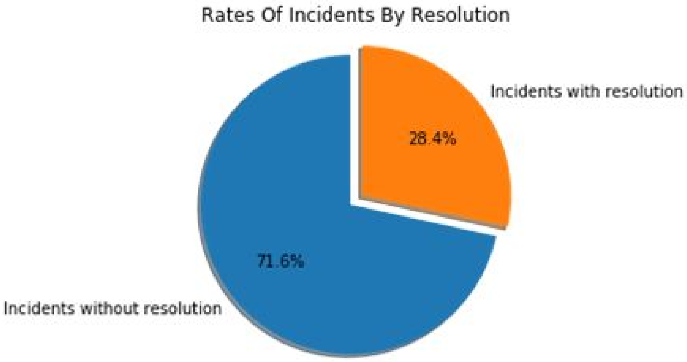


[Figure 3.3] Crimes Count on Time Period of Day in San Francisco

<u>Sub Question 4:</u>

a. What was the percentage of incident cases that did not have any resolution?

People living in San Francisco should be informed of reported incidents without resolution to have better awareness about the inherent risk of certain rates of incidents in this city. Having such information would also provide the benefit of being able to assess the probability of certain cases to be resolved if they happen to undergo a similar situation.



[Figure 4] Rates of Incidents by Resolution

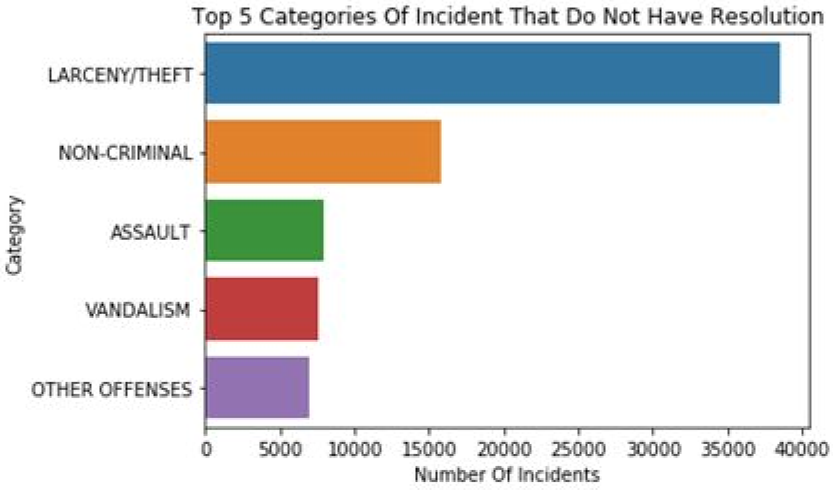
b. What were the categories of crime that did not have any resolution?

To display non-repetitive values from the category of crime without resolution, we used a unique method. Based on the series of 39 categories, all of the categories had cases that remain unresolved. Moreover, those cases constituted 71.6% of all reported incidents in San Francisco in 2016. We used a pie chart to better display the percentage by pointing out the resolution with values “NONE” as incidents without resolution and grouping the other values to another bucket called incidents with resolution.

Sub Question 5:

a. What category of crime that had the most unresolved issues?

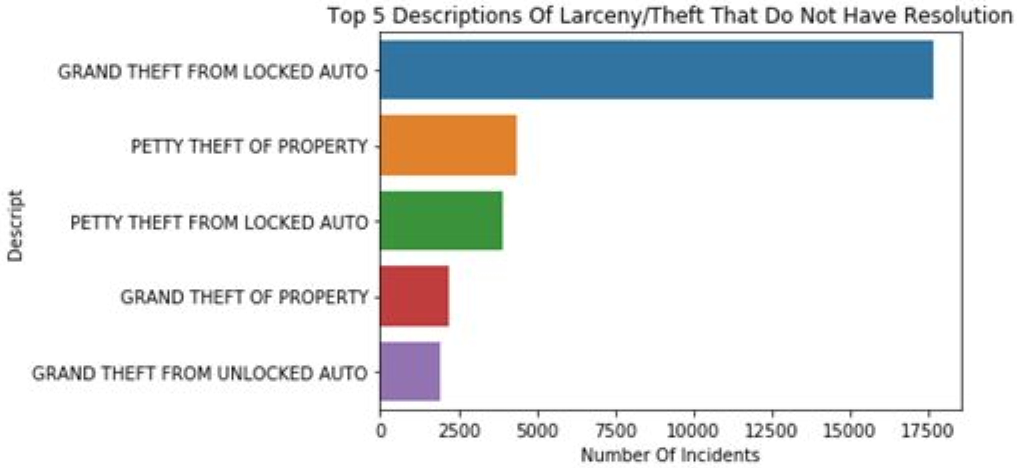
The jupyter file provides the details of each category of incidents and type of description with its occurrences, sorted by the most frequent value. However, to provide better visualization, the bar plots only highlight the top five of the respective groups. In order to do that, we used the concatenation method to create a data frame containing data frames from the top 5 categories of incident.



[Figure 5.1] Top 5 Categories of Incident that Do Not Have Resolution

The chart above tells us that reported cases classified as larceny or theft remain unsettled the most. In this category alone, there were more than 38,000 out of roughly 108,000 unresolved cases. Furthermore, by applying the length method to this category of crime, there were 53 types of description briefly outlining the characteristics of each incident within this category.

b. For that category of crime, what was the incident that happened the most?



[Figure 5.2] Top 5 Descriptions of Larceny/Theft that Do Not Have Resolution

We used another bar plot to display the top 5 types of description within the category larceny/theft. To provide a single data frame for this purpose, we also merged the separated data frames from each type of description by using the concatenation method. Figure 5.2 shows appalling information that larceny or theft from locked automobiles dominates the cases. About half of total theft incidents without resolution is derived from car break-ins. The fact that the cars are locked doesn't stop the criminals from obtaining their target.

Conclusion

Analysis Conclusion:

Our data analysis project consisted of analyzing the various types of crimes in San Francisco from a data set captured in 2016. We were provided information such as the types of crimes, the districts of incidents, time of incident and much more.

Our objective was to develop an analysis that will provide insightful meaning and further information about how individuals can be safe in potentially dangerous districts in San Francisco. We discovered insights such as the Southern District being the most incident dense district with theft being the highest type of crime. We also discovered that in October of 2016, Fridays at roughly 6 pm is when the majority of the crimes occur. We realized that nearly 72% of cases were not resolved with the category of crime being grand theft of locked automobiles. Our team had come to the conclusion that individuals living in San Francisco can be protected by avoiding dangerous districts in San Francisco and avoid bringing their vehicles to these locations, especially in October.

Project Conclusion:

Through this project, we had the opportunity to utilize our vast knowledge of Python programming language on the project analysis. As a result, we became familiar with Python, especially in the sense of manipulating, modifying, analyzing and visualizing our data. Our team has developed a stronger understanding of how the data life cycle progresses and how to extract insightful metrics from our analysis.

Since this course is relatively new to us, remembering all the syntax and steps needed to execute a particular objective was time consuming and challenging. Even though we had fragmented knowledge of how data analysis and data visualization progresses, our team encountered difficulties in developing intricate visualizations due to the various syntax and parameters that each type of graph requires.

Through group discussions and team meetings, we were able to share similar opinions and come together as a team to derive potential solutions to visualizing graphs, merging dataframes, and manipulating our data. Our team has learned to communicate technically and encountered countless technical difficulties applying concepts and manipulations. However, as a collective team, we were able to come to specified results and developed multiple convenient solutions to derive at our desired output.