Capstone Project 1: Milestone Report

Mapping and Visualizing 911 calls for violent and property crimes in San Francisco and predict future calls

Introduction :

Today, San Francisco is known more for it's tech scene. But, with rising wealth inequality and housing shortages, there is no scarcity of crime in San Francisco. According to SF police, Dispatchers in San Francisco receive between 3,500-4,000 calls per day. Ninety percent of the calls are for Police. The other 10% are for Fire/Medical services. Approximately half of all calls are 911 calls, and the other half are non-emergency police calls.

In the last two decades it has become very easy for the public to access emergency dispatch system via the 9-1-1 function. Ease of service has been a two-edged sword for the police and other emergency agencies. The 9-1-1 function has had a major impact on the number of police calls for service. The increasing number of calls handled by emergency dispatch systems has also increased the necessity for handling these calls efficiently while meeting the needs of the public.

It is important for police departments to establish strategies that enable them to respond effectively and rapidly to emergency situations.

In this project, I will use a machine learning model to predict the number of 911 calls which will help the authorities to distribute resources fairly by needs.

Data :

The data used for this project originates from sfgov.org . It has been published by City and county of San francisco.It consists of the calls for service regarding criminal activity (unverified), from SFPD covering the period 03/31/2016-present and has information about the call time, crime type and the demographic information of the event reported.
The data can be accessed from:
https://data.sfgov.org/Public-Safety/Police-Department-Calls-for-Service/hz9m-tj6z

Below are the features available in the dataset:

1. crime_id : Unique number to identify the call
2. original_crimetype_name : crime name and Radio codes (Each crime type has a unique Radio code. These are *used by law enforcement* to keep communication succinct. Ex. 187 - Homicide)

3. report_date : Report date
4. call_date : date on which the call was made
5. offense_date : date of the offense occurred
6. call_dttm : date and time of the call received
7. disposition : different types of disposition codes (A disposition code is a unique code given to the action taken after the call, it is transmitted at the conclusion of every call. Ex: SFD - SFFD Medical Staff Engaged)
8. address : address of the event reported
9. city : city of the event reported
10. state : state of the event reported
11. agency_id : Agency id
12. address_type : the type of the location the event occured. It can be a premise address, an intersection, a common location/ public place or a geo encoded address.
13. Common_location- address of the public place the event occurred

In addition to above dataset, I used radio codes data. For better understanding of the crime types, I replaced original_crimetype_name records with their equivalent crime names from Radio_codes_2016.xlsx file obtained from the same source https://data.sfgov.org/Public-Safety/Police-Department-Calls-for-Service/hz9m-tj6z.

Data wrangling and Cleaning:

As part of data wrangling my first step was to retrieve the data from https://data.sfgov.org/. I noticed that the data was being updated daily and was available for download and through an API. I decided to make use of their API to get the real time data.

Notably, the dataset had over two million rows from over three years of calls data. I initially tried reading in the data-set using the API but got only 1000 rows.

After some research, I learned that at present SODA API has a limit of returning 1000 rows at a time when querying the dataset. To query more than 1000 rows, I added '$limit=' parameter to json url which set a limit on how much I wanted to query from a dataset.

After reading the data in json format I converted it to pandas dataframe named ‘calls_for_service’. I spent time identifying null values, deleting the rows with null values and dropping the unuseful columns. After taking care of Null values, I cleaned the data by removing white spaces and special characters which was mostly observed in original_crimetype_name column.

Column Original_crimetype_name in the dataframe had some entries of radio codes instead of the actual crime name . Radio codes are brevity codes used in voice communication by law enforcement. I merged radio codes excel file with calls_for_service dataframe to replace the radio codes with it's more understandable meaning.

For further analysis, I created two separate data frames for calls for violent crime and calls for property crime.

Dataframe ‘Calls_for_violent_crimes’ was created with calls data reporting Homicide, Robbery, Strong arm Robbery, Aggravated Assault/ADW , Rape/Sexual Assault, Sexual Assault Adult , Sexual Assault Juve, Kidnapping and Stabbing.

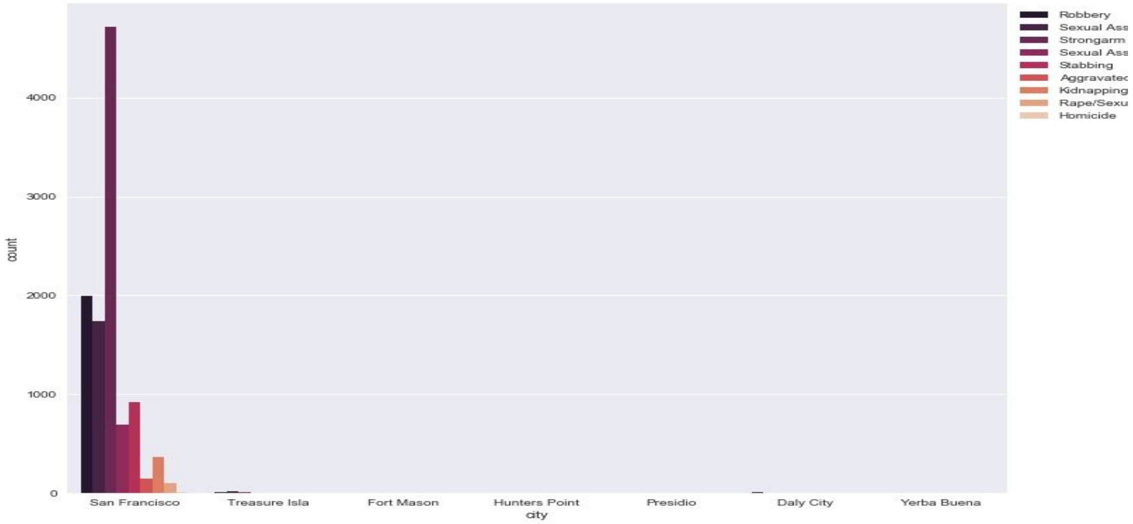
Whereas dataframe ‘Calls_for_property_crimes’ had the calls data reporting property crimes such as Petty Theft','Grand theft, Burglary, Embezzlement, Person breaking in, Stolen vehicle, Broken Window, Vandalism, Extortion, Auto Boost / Strip and Shoplifting.

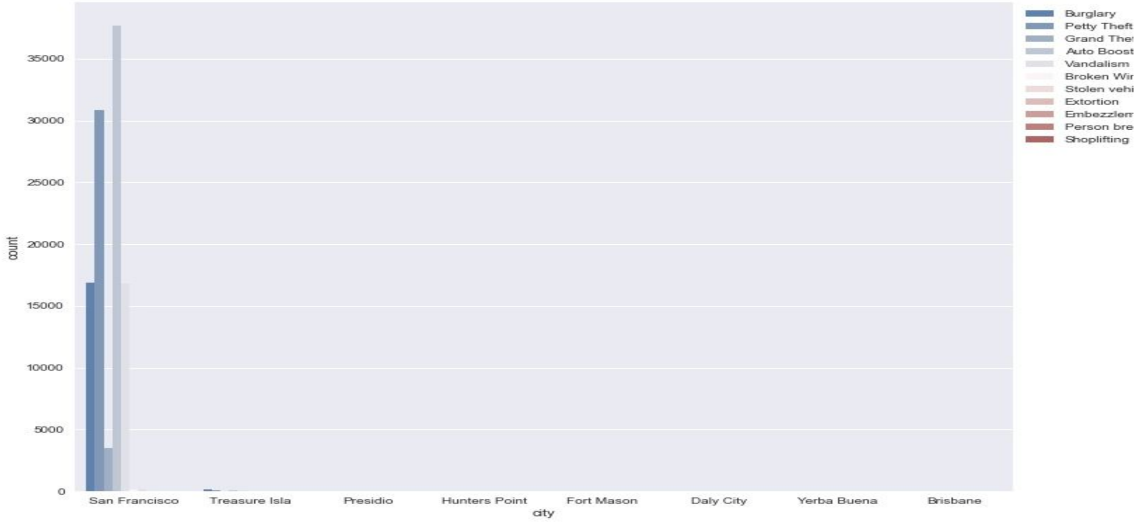
The preprocessing stage can often take the majority of the time spent on a data analysis project. Fortunately for the sake of the analysis this part didn’t consume as much time as it could have due to the data being fairly clean to begin with.

Data Story:

Once data wrangling and cleaning part was done, I started to explore the data to find correlation between different variables (features) from the dataset.

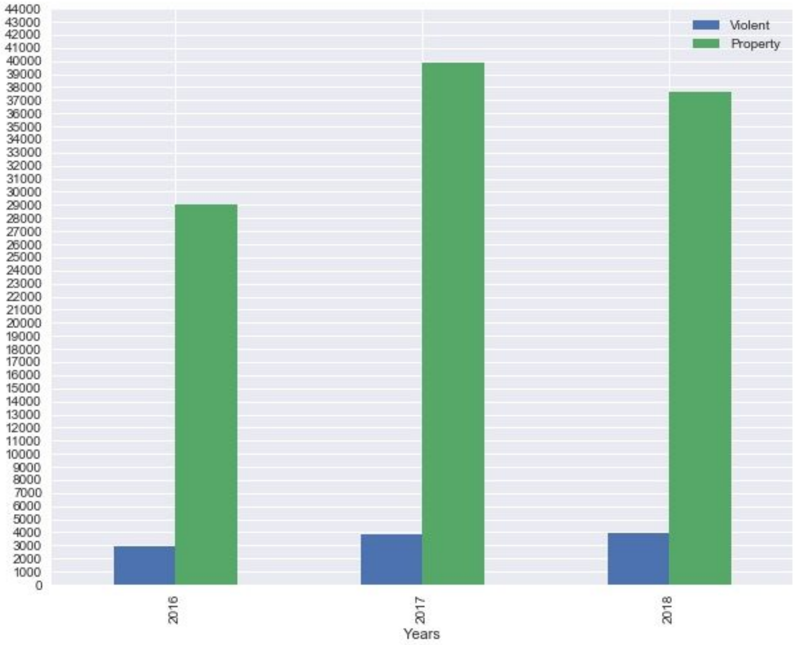
I tried initially plotting counterplots for the different violent and property crimes per city and observed that being a bigger city among all, San Francisco city has fairly large amount of calls received.





I wanted to check how the number of calls has changed over the years. With a bar plot below it was observe that the number of calls for property crime increased in 2017 and started decreasing in 2018 .

Whereas calls reporting Violent crimes increased in 2017 and kept increasing in 2018.

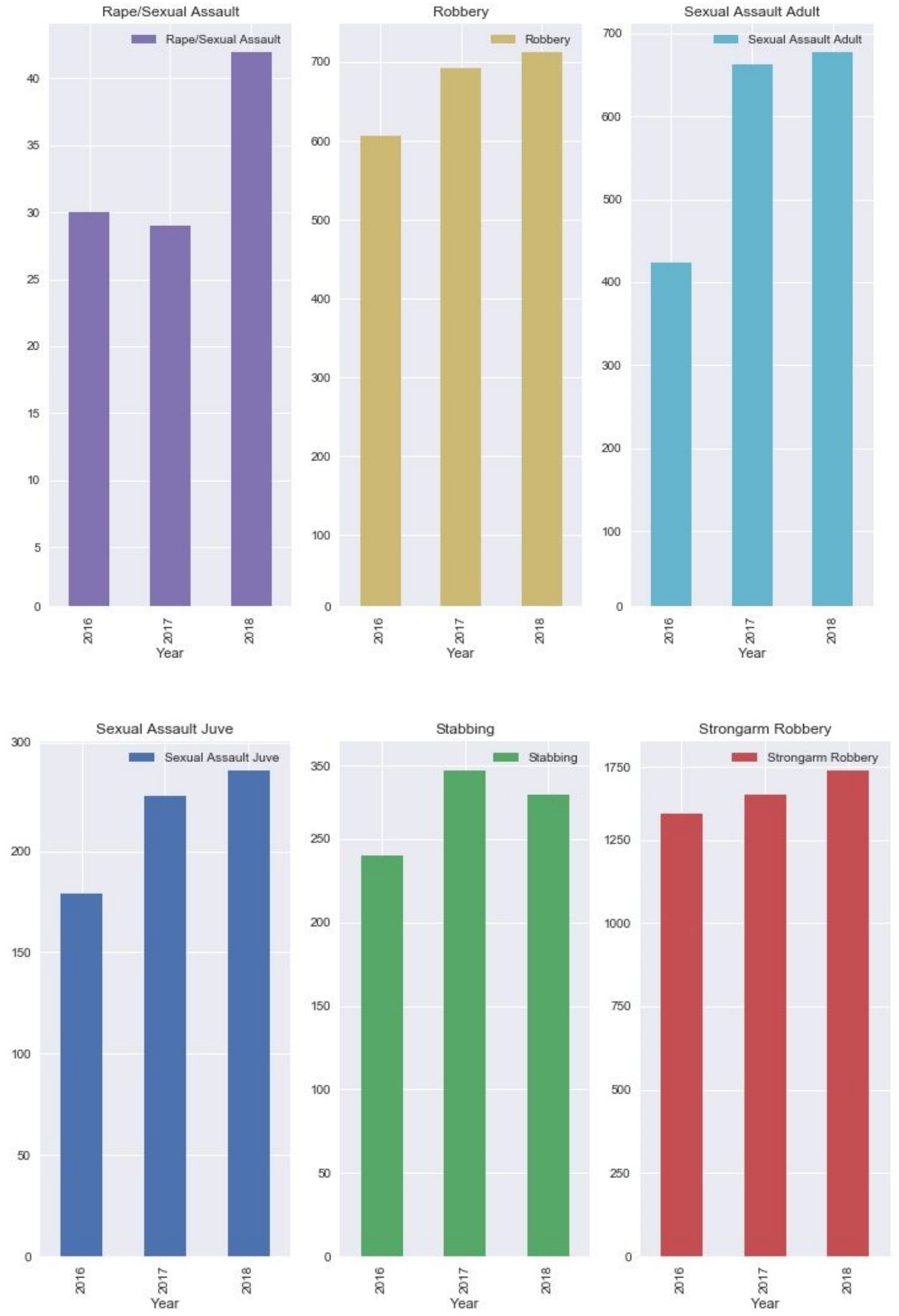


As the data was incomplete for the years 2016 and 2018, it was inappropriate to say that there were more number of calls in 2017. But the fact was alarming to know that even with incomplete data of year 2018, the calls for some of the crime types were more than Year 2017.

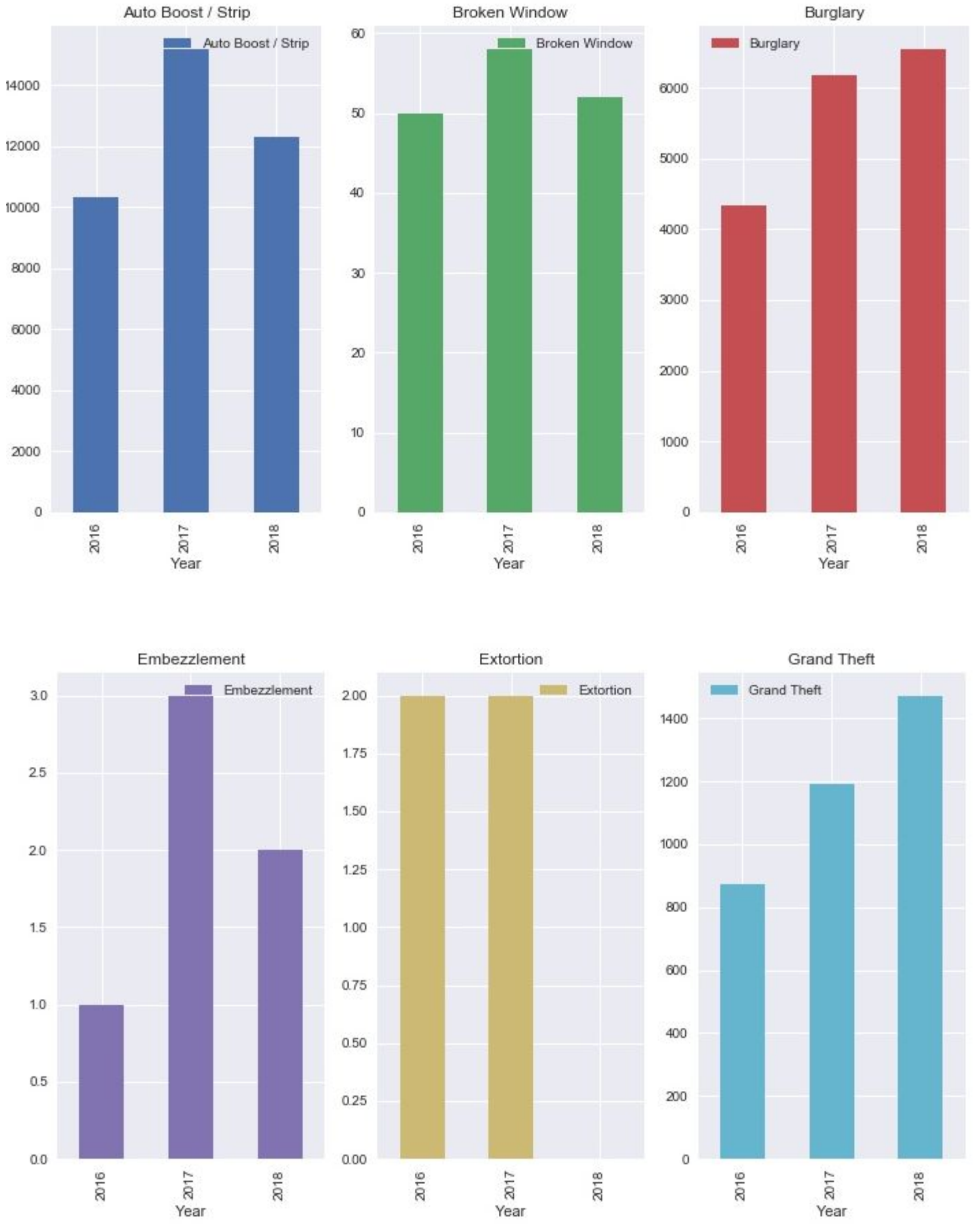
From 2017 to 2018, calls for violent crimes increased by 3.86%, with 9.11% increase in sexual assaults and 3.85% increase in robbery.

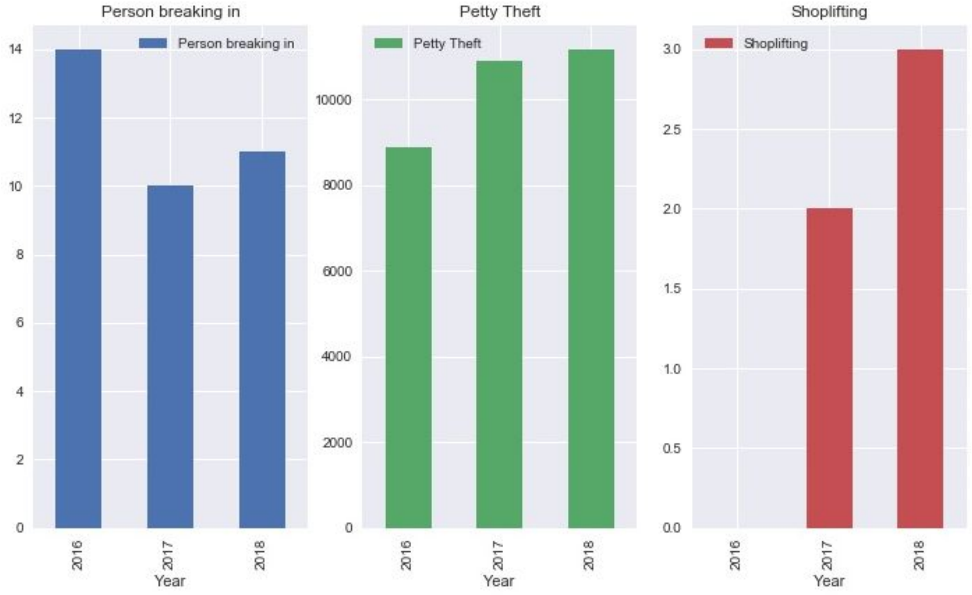
Similarly, an upward trend was observed from 2017-2018 in some of the property crimes such as a 6% increase in Burglary related calls, 23,67% increase in Grand Theft related calls and 2.5% increase in petty thefts related calls.

Plots showing increase/ decrease in calls related to violent crimes:

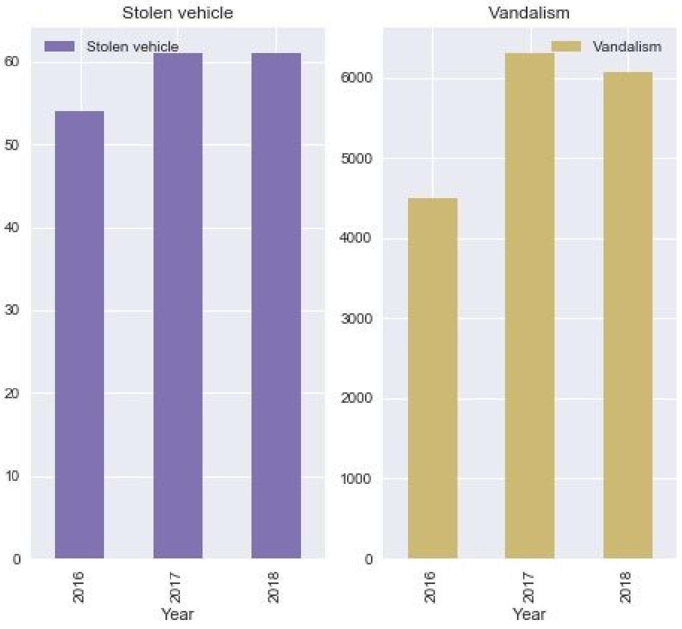


Plots showing increase/ decrease in calls related to violent crimes:

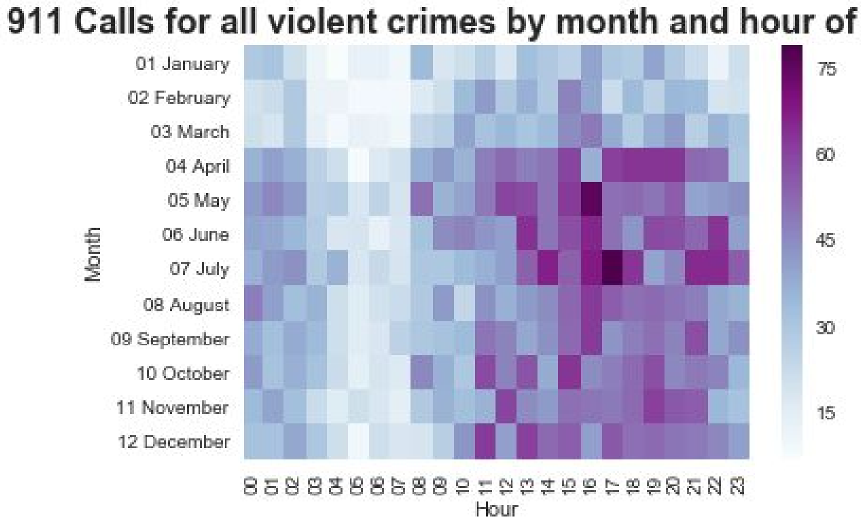




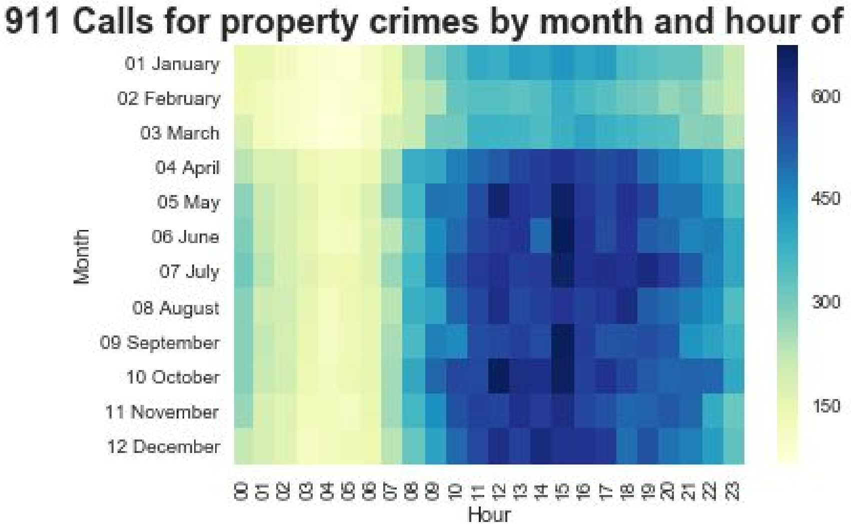
Observed a sudden spike in shoplifting from 2017 to 2018.



Further,911 Calls were plotted by month, day and hour of the day using Seaborn Heatmaps


911 Calls for all violent crimes by month and hour of the day

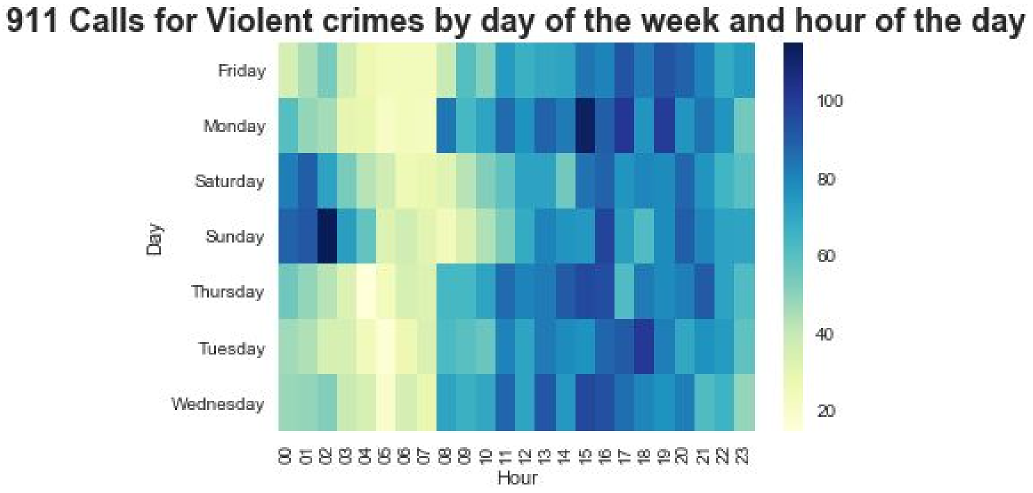
.

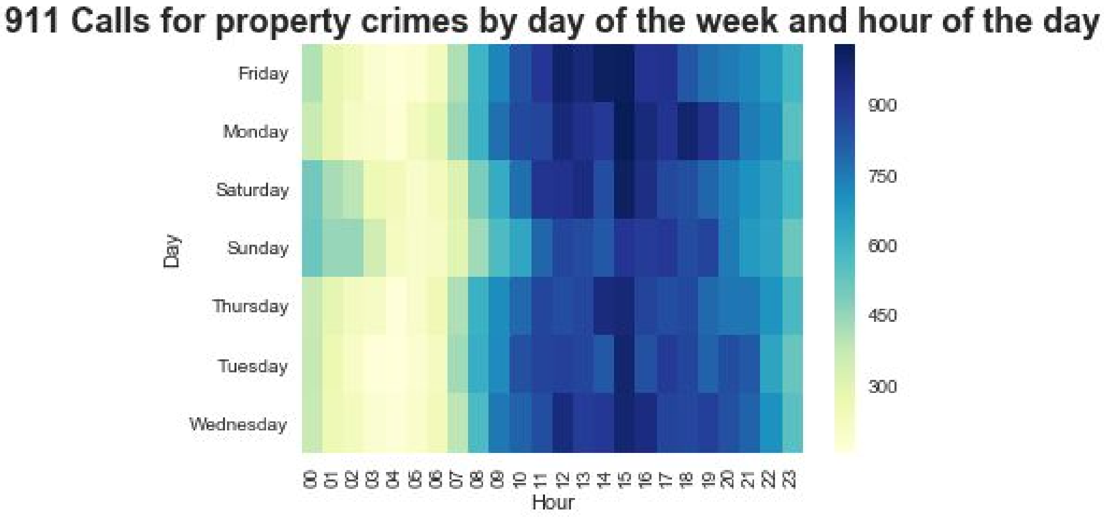

911 Calls for property crimes by month and hour of the day

From above heatmaps, it was clear that more calls were being made during summer.

This proves the old theory that,increased temperatures in summer drive many out of doors and to leave windows open in their homes.Also, increased daylight hours can lengthen the amount of time people spend away from their homes raise the amount of people in public and the amount of time that homes are left empty. Others point to the effect of students on summer vacation, who are otherwise occupied with schooling during other seasons. It is also true that those suffering from heat-induced discomfort become more aggressive and likely to act out.

It is also interesting to note that the calls are mostly made from 9:00 am to 2:00 am.One can observe some clear drops in numbers around 3:00 am to 8:00 am in the morning. From the spike in calls right after the drop, it seems like the most frequent crime like robbery might have happened when people are asleep and businesses are closed.

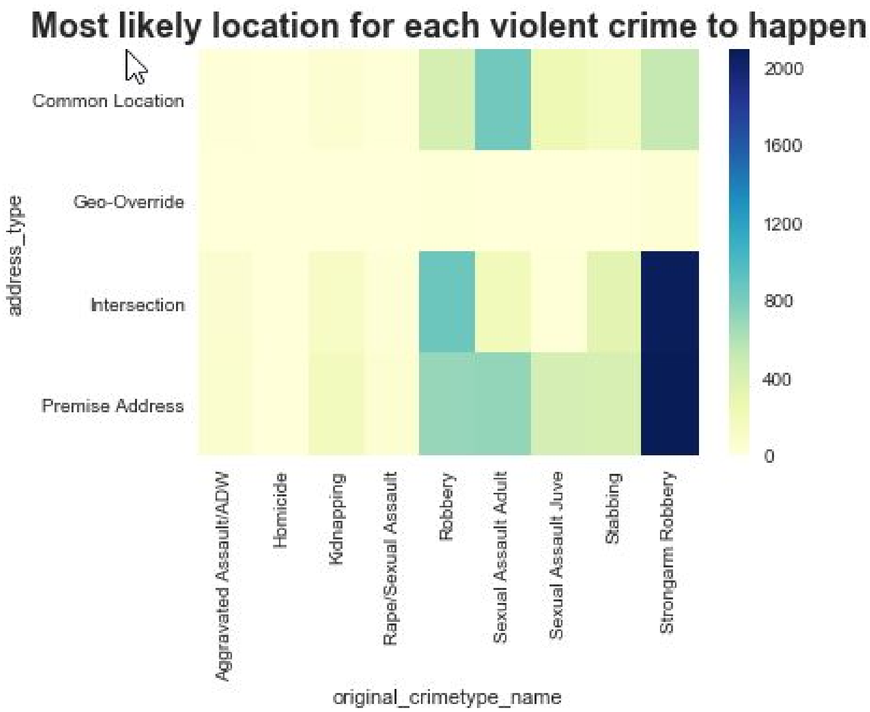

911 Calls for Violent crimes by day of the week and hour of the day


911 Calls for property crimes by day of the week and hour of the day

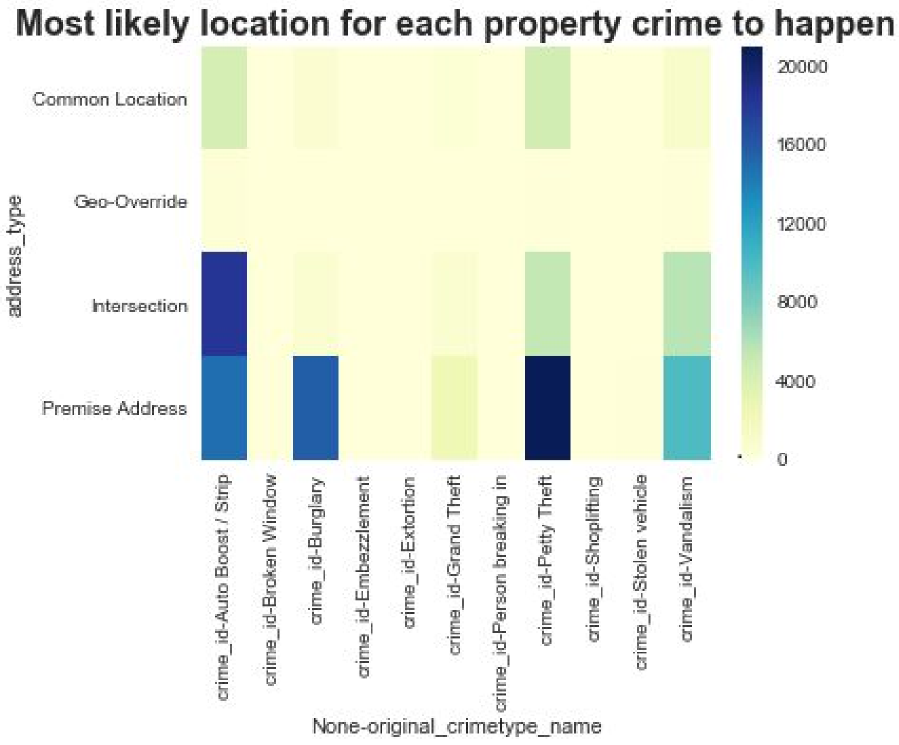
From above heatmap, I found out that the call volume was more at 3.00 pm everyday and there is a spike in number of calls on Monday.

It was also observed that more calls were made on weekend from midnight to 4.00 am.

Next in the exploratory analysis, I found out the most likely address type for each crime type. These was derived from a heatmap plotting using address types and crime types.



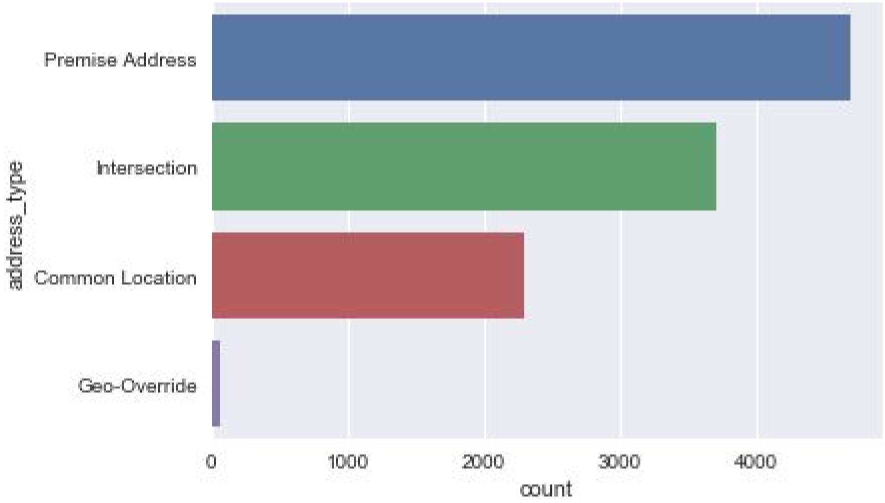
It was alarming to know that most sexual assault related calls reported had common public place address. Strong arm robberies and robberies happened on premise and at intersections.



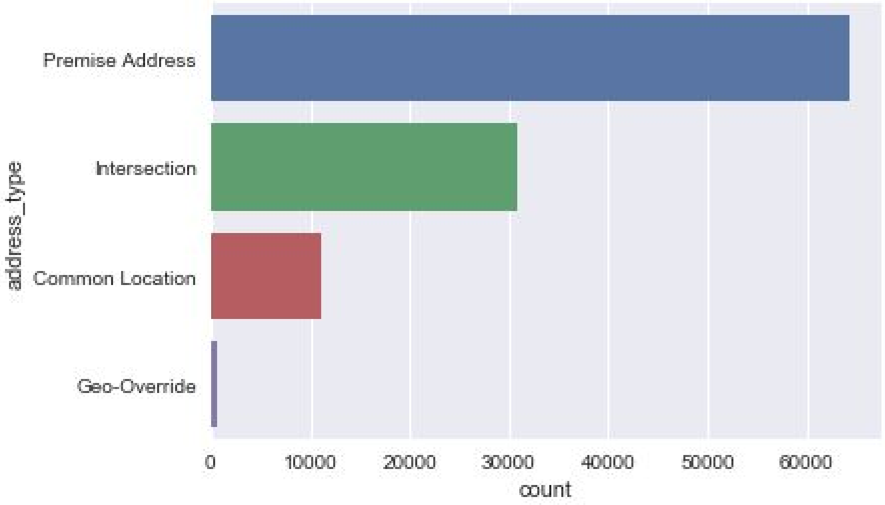
Similarly, most petty thefts and burglaries occurred at premise addresses and auto boosts/ strip happened on intersections.

From plotting the crimes per address types, it was found that premises in San francisco are more vulnerable to both violent and property crimes.

Violent Crimes per address type:



Property Crimes per address type:



<u>Exploratory data analysis :</u>

From visualizing data, I could address the question of whether or not the specific types of calls for service have changed over time, what are the most busy months for the dispatchers and what time and day has more call volume.

From data visualization I found out that most calls were being made in summer but to prove this I performed a hypothesis test.

Violent crime hypothesis testing:

1. The Null hypothesis : There is no difference in number of calls for service in summer and Spring
   Alternative hypothesis : There is a difference in number of calls for service in summer and Spring

Null hypothesis was rejected as p-value= 3.6965315773436525e-06 which was less than the level of significance i.e (1%).

Property crime hypothesis testing:

As per the Crime report, Burglary was the only property crime to increase last year in San Francisco, so let's start with below hypothesis relating calls reporting burglaries.

2. The Null hypothesis : There is no difference in number of calls for service in summer and winter
   Alternative hypothesis : There is a difference in number of calls for service in summer and winter
3. The Null hypothesis : There is no difference in number of calls for service in summer and Spring
   Alternative hypothesis : There is a difference in number of calls for service in summer and Spring

Both null hypothesis were declined as p-value (Hypothese 1: p-value: 3.735964927553748e-50, Hypothesis 2: 3.6912232284393345e-08) which were less than the level of significance i.e (1%).