Predict 911 calls for violent and property crimes in San francisco city

By Simantini Patil

Introduction

A 911 call may be a result of an emergency medical need, fire attack, natural disaster or crime. Whatever is the reason, it takes community resources to respond to these requests. Slow response rate to 911 calls can lead to death, destruction of property and high crime rate.

Oftentimes due to lack of data, police department face difficulties to foresee the occurrences of crime/ emergencies and they get caught by surprise. If police department/ related authorities are able to understand the factors which contribute high volume of 911 calls, they can implement proactive strategies to reduce the calls in the future.

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. This project will analyze 911 emergency call data compiled from San Francisco metropolitan area and will conduct a hotspot analysis to determine areas of high call volume.

This project will be developed keeping in mind about the potential users who can take an action to reduce the occurrence of the crimes and those could be San Francisco police department or related authorities. The project will help the authorities to know the crimes that have occurred over the time.

Overview of dataset:

The dataset used in this project is 911 calls data published by City and County of San Francisco available at https://data.sfgov.org/ It consists of the calls for service regarding criminal activity (unverified), from SFPD. Data covers the period 03/31/2016-present.

Dataset has following columns:

	Datatype
address	object
address_type	object
agency_id	int64
call_date	object
call_dttm	object
call_time	datetime64[ns]

```
city                        object
common_location             object
crime_id                     int64
disposition                 object
offense_date                object
original_crimetype_name     object
report_date                 object
state                       object
```

Data Wrangling

There's a couple interesting things about this data-set. The dataset is made available by Socrata, a company based in Seattle that has worked with governments to release open data to the public.

Socrata Open Data API (SODA) allows you to programmatically access a wealth of open data resources from governments.

Notably, this dataset has over two million rows from over two 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 currently 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 want to query from a dataset.

After I fetched 2 million records, I saved them in pandas dataframe. My dataframe was of 2000000 rows X 14 columns dimension.

While analyzing missing data, I observed missing records in 'city ' and 'common_location' columns. I dropped common_location column along with call_date , call_time and agency_id columns as those were not necessary for the analysis.I deleted the rows where city was missing in city column. After dropping all I got around 1900000 plus rows and 10 columns.

While going through the original_crimetype_name column, I observed that there are many radio codes so I downloaded SFPD radio codes xl document.

Radio codes had two columns: Radio codes and Meaning

I then imported Radio_Codes_2016.xlsx file in pandas dataframe and mapped it with the
original dataframe of service calls. By doing this I replaced all the radio codes in original dataframe with their meanings.

Later I observed that many records in original_crimetype_name column are misspelled, written differently or has special characters. I aggregated them under appropriate crime types.

Finally I extracted all the records of Violent crimes such as homicide, rape, robbery and aggravated assault
reference: *https://sfgov.org/scorecards/public-safety/violent-crime-rate-and-property-crime-rate* and property crimes such as Petty Theft, Grand Theft, Burglary, Embezzlement, Strong Arm Robbery, Person breaking in, Stolen vehicle, Broken Window , Vandalism, Extortion, Auto Boost / Strip, Shoplifting which I will leverage later in the analysis.

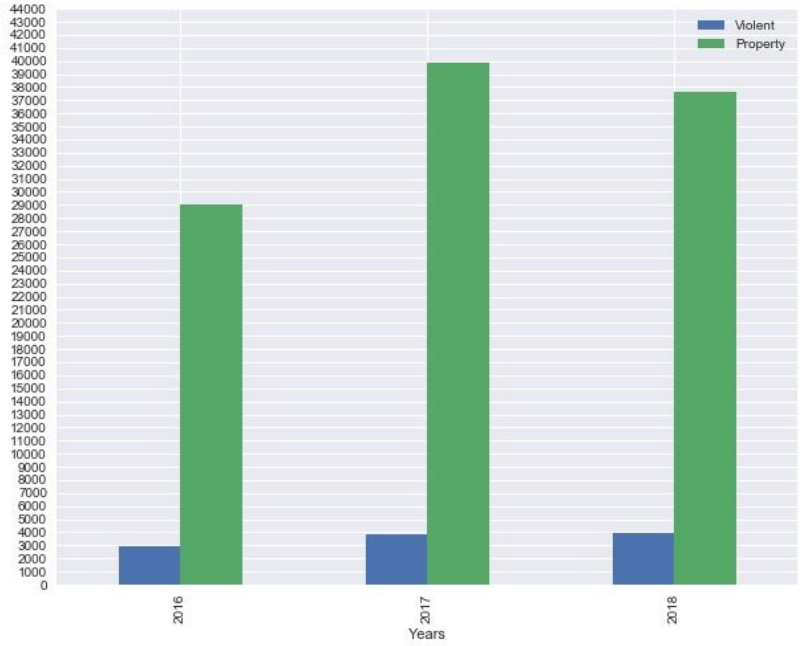
Overall, data cleaning part was not easy . Learning the meaning of radio codes and going through thousands records to group them under appropriate crime type required lot of time.

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.

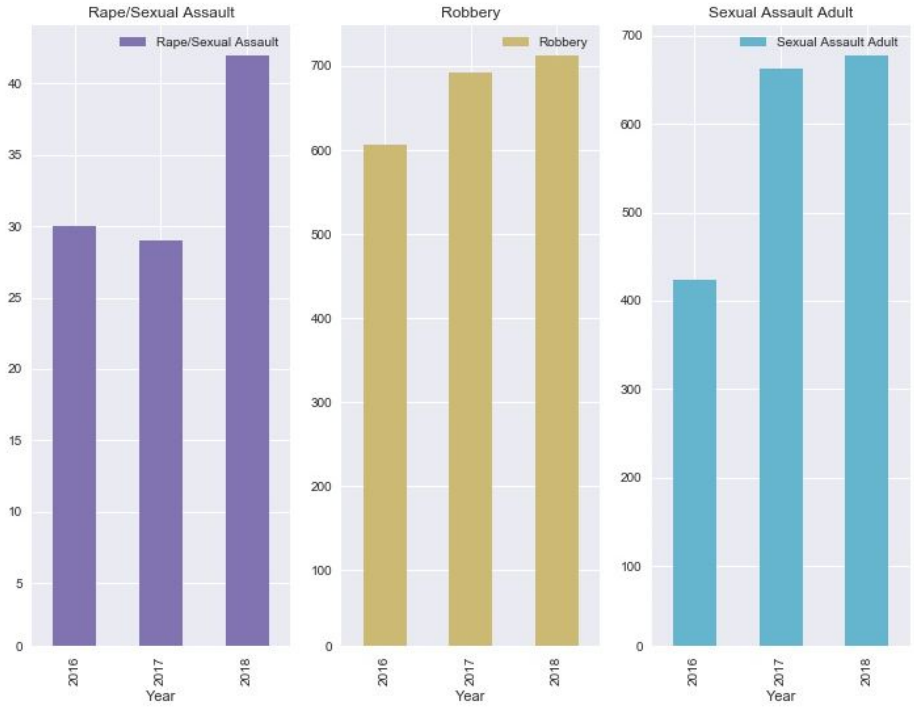


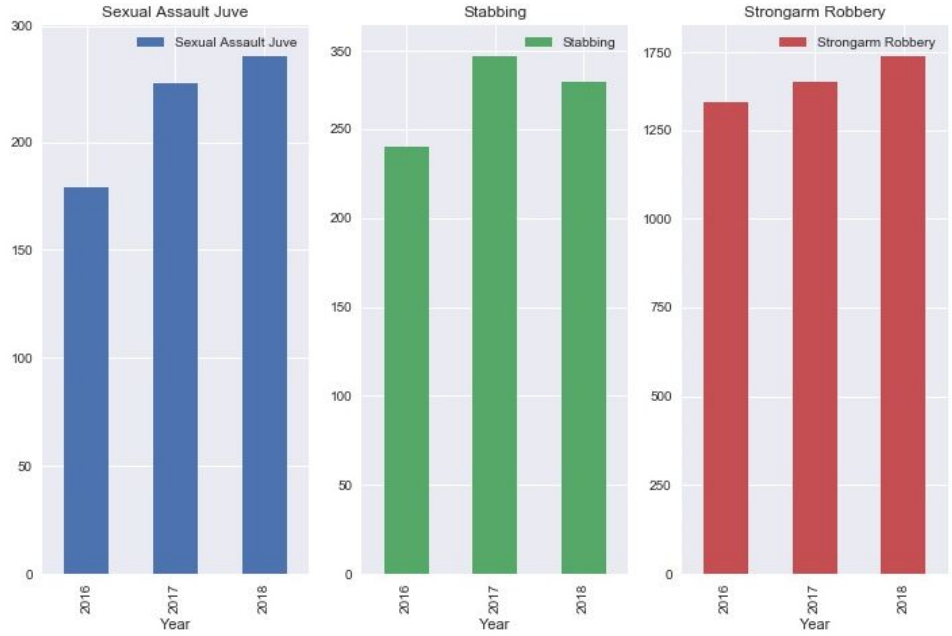
As the data was incomplete for the years 2016 and 2018, it was inappropriate to say that there were more number of calls in 2017. But the fact was alarming to know that

even with incomplete data of year 2018, the calls for some of the crime types were more than Year 2017.

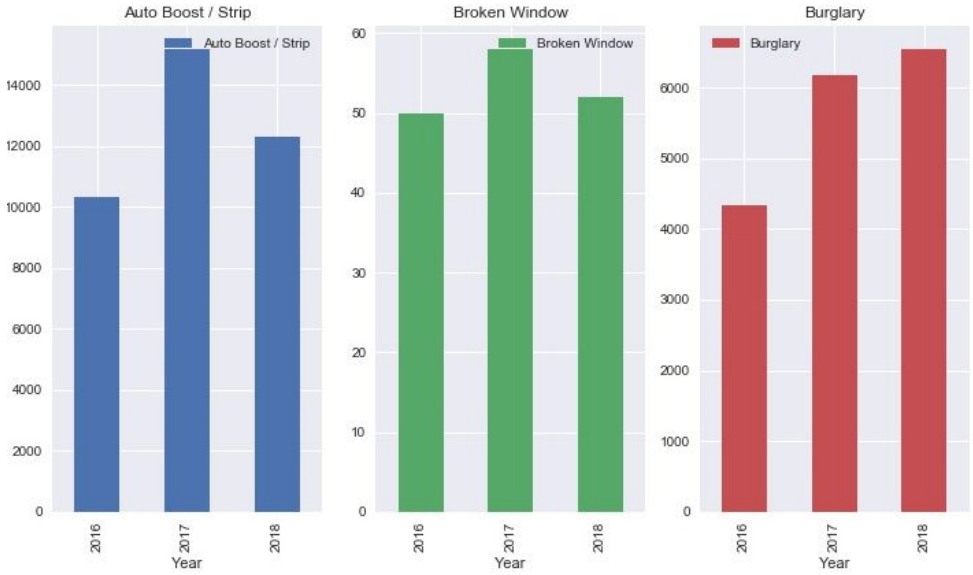
From 2017 to 2018, calls for violent crimes increased by 3.86%, with 9.11% increase in sexual assaults and 3.85% increase in robbery.

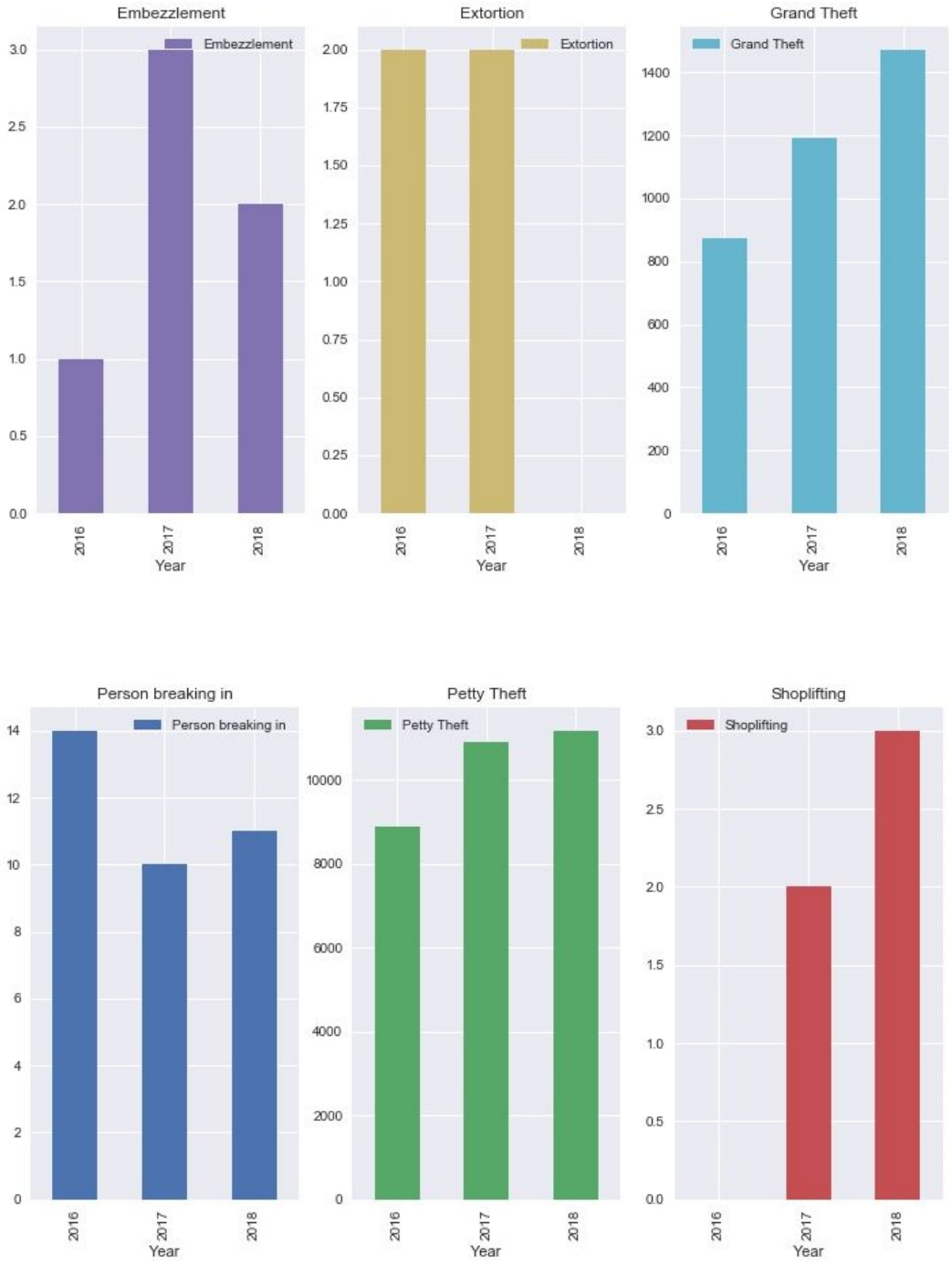
Similarly, an upward trend was observed from 2017-2018 in some of the property crimes such as a 6% increase in Burglary related calls, 23,67% increase in Grand Theft related calls and 2.5% increase in petty thefts related calls.
Plots showing increase/ decrease in calls related to violent crimes:





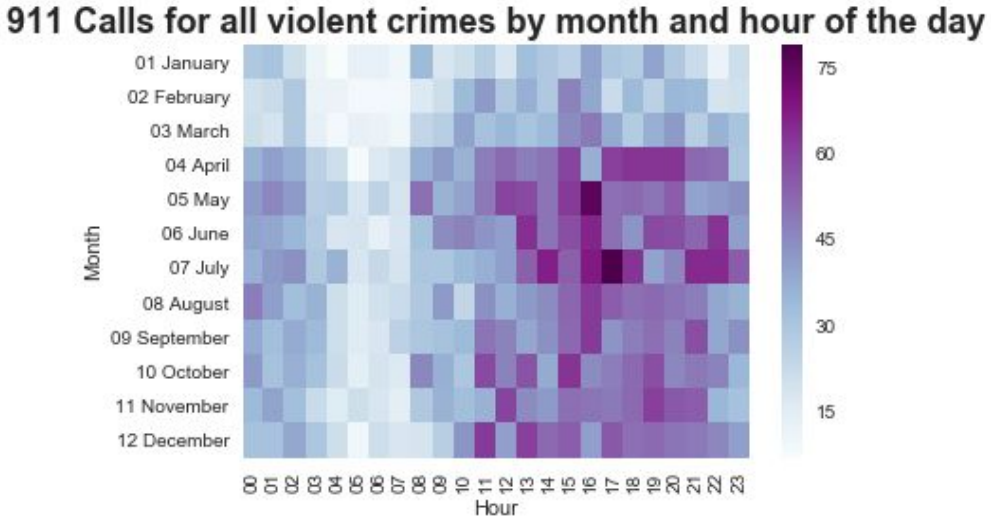
Plots showing increase/ decrease in calls related to violent crimes:

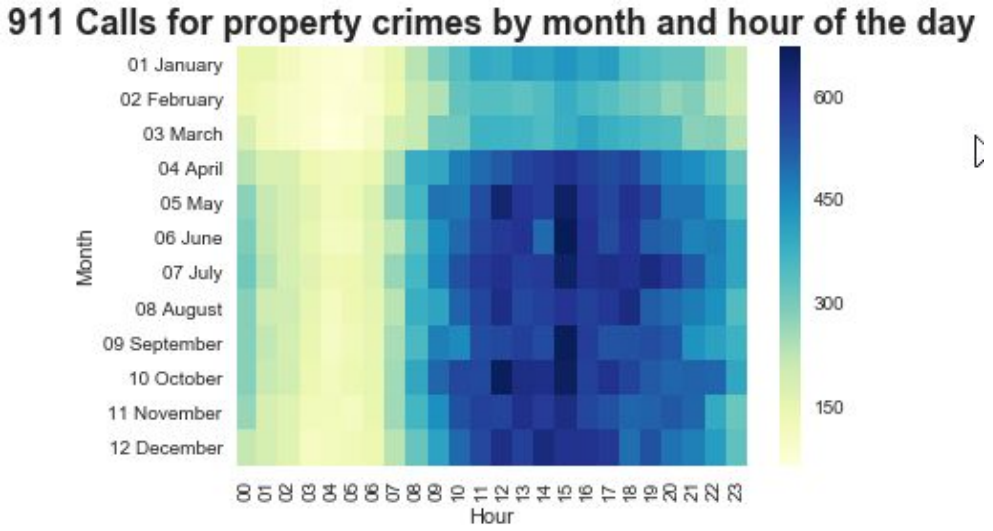




Observed a sudden spike in shoplifting from 2017 to 2018.

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

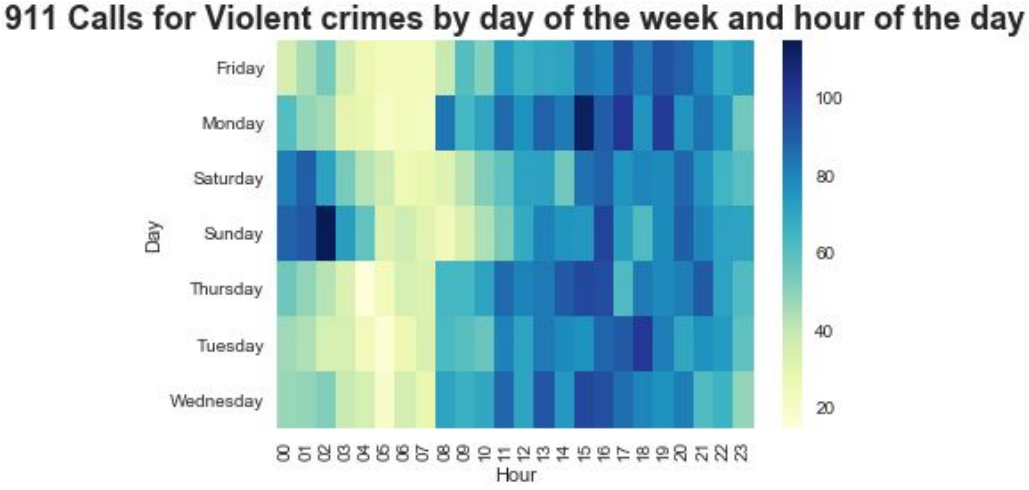


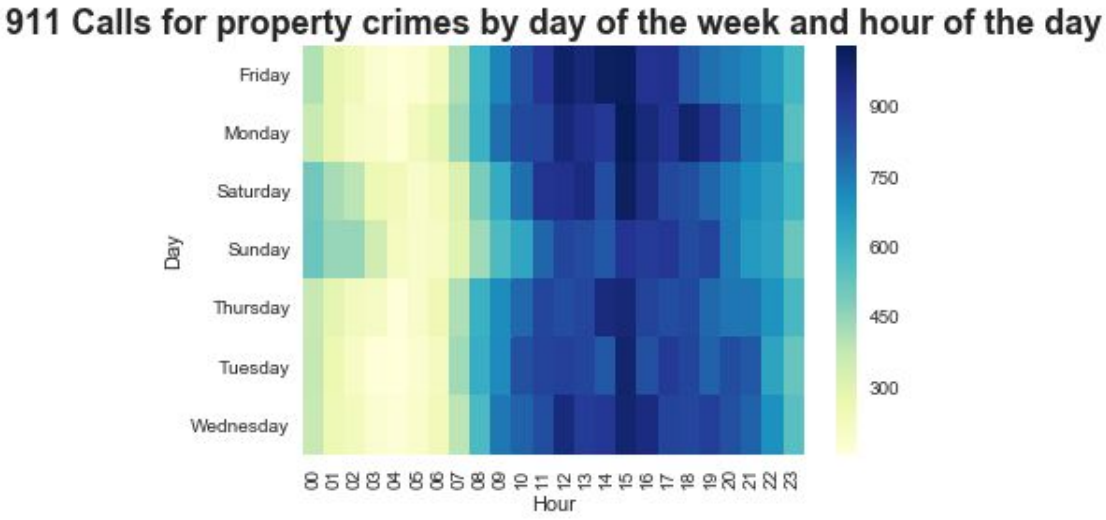


From above heatmaps, it was clear that more calls were being made during summer.

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.

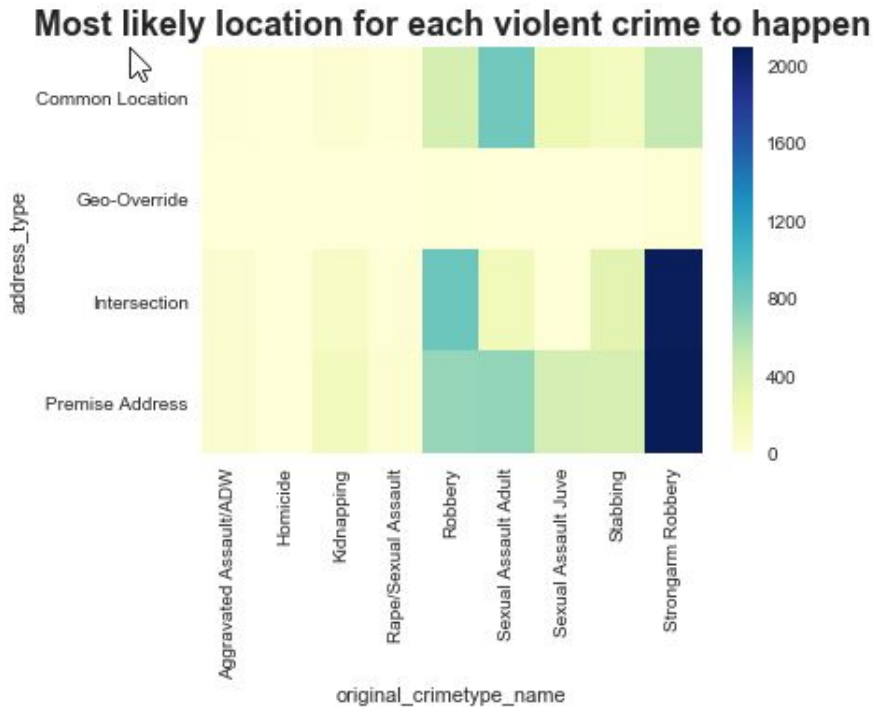
It is also interesting to note that the calls are mostly made from 9:00 am to 2:00 am.One can observe some clear drops in numbers around 3:00 am to 8:00 am in the morning. From the spike in calls right after the drop, it seems like the most frequent crime like robbery might have happened when people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



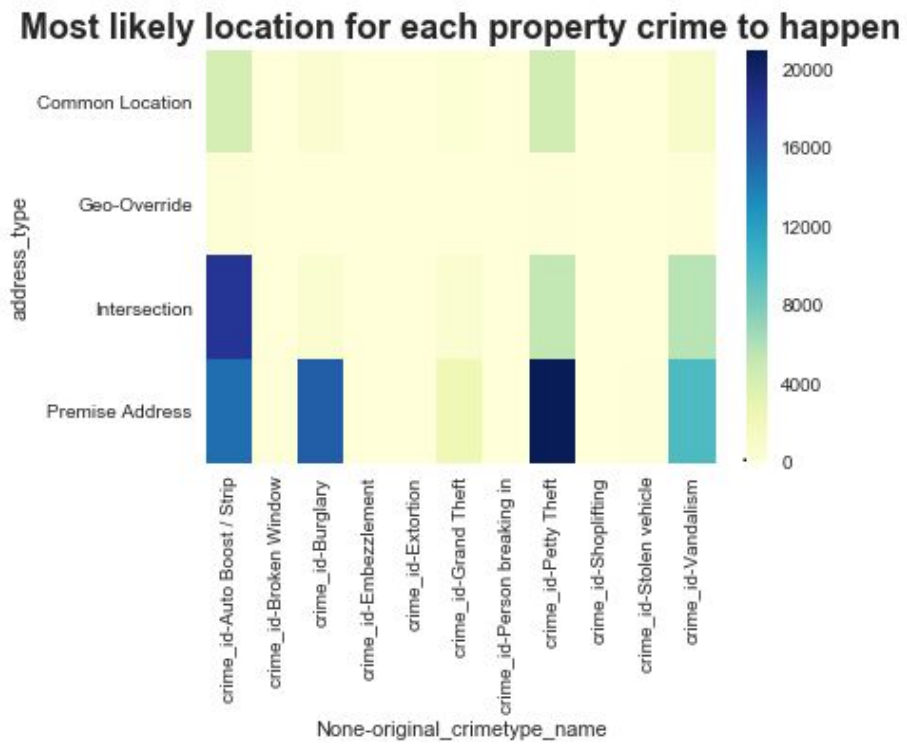


From above heatmap, I found out that the call volume was more at 3.00 pm everyday and there is a spike in number of calls on Monday.

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.



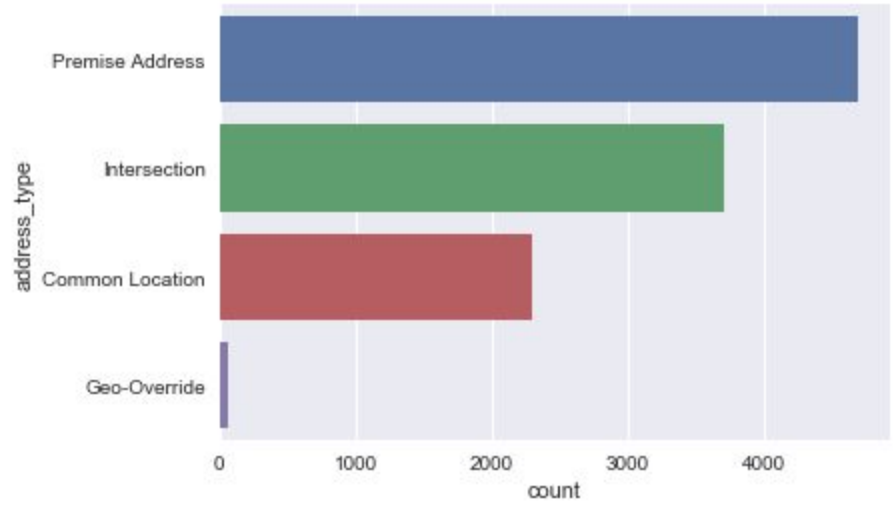
It was alarming to know that most sexual assault related calls reported had common public place address. Strong arm robberies and robberies happened on premise and at intersections.



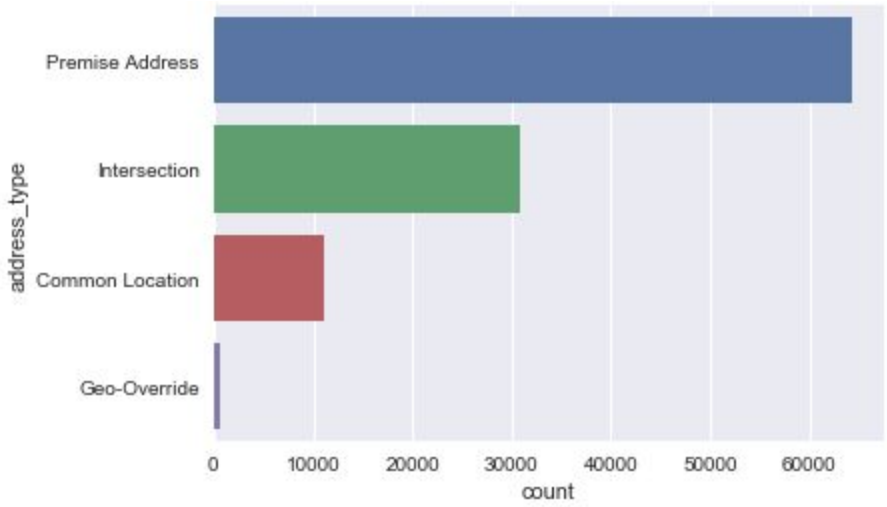
Similarly, most petty thefts and burglaries occurred at premise addresses and auto boosts/ strip happened on intersections.

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:



Exploratory data analysis

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%).

Machine Learning

Predicting future 911 calls for violent and property crimes poses a particularly interesting data challenge because it has both geospatial and temporal dimensions and may be affected by many different types of features like weather, city infrastructure, population demographics, public events, government policy, etc.

Most of the models use extensive features than that were available here.Those include population demographics, employment rates, housing prices, race etc.

I used below mentioned features to predict the number of 911 calls for violent and property crimes from June 2018- December 2018.

Feature Engineering:

From EDA, a significant relevance between time of the call and number of calls was observed. This helped tremendously in manual feature selection process.

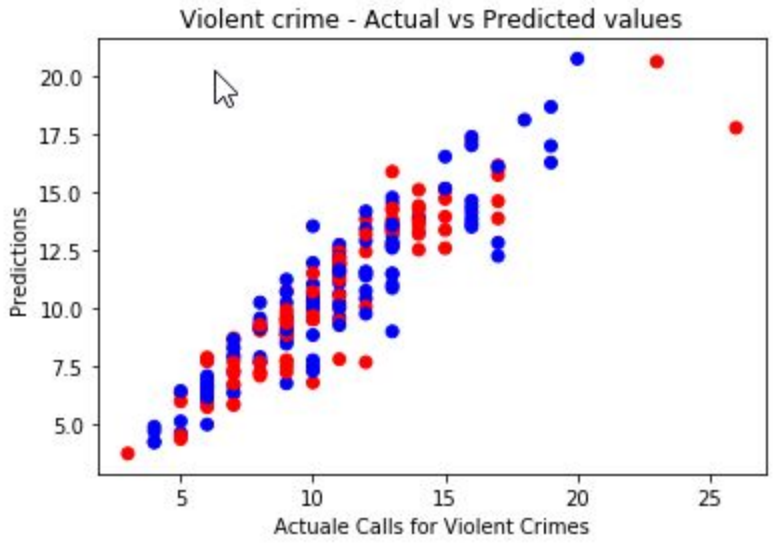
Year : Year of the call received will be extracted from call_dttm column and was converted to int64 data type
Month: Month of the call received will be extracted from call_dttm column and was converted to int64 data type
Date: Date of the call received will be extracted from call_dttm column and was converted to int64 data type
Hour: Hour of the call received will be extracted from call_dttm column and was converted to int64 data type
Day: Day of the week of the call received will be extracted from call_dttm column and Was Encoded to 1-7 for the days from Monday-Sunday
Event: Number of the events per day

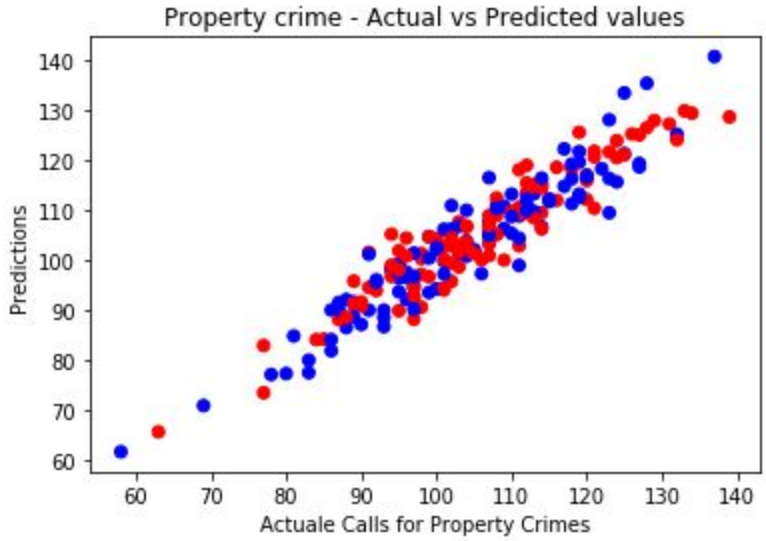
Number of events per day were calculated by grouping the year,month and date of the call and two dataframes (one for violent crimes and property crimes each) were being created with above features.

Predictions:

As the goal was to predict the number of calls for the period June 2018 - December 2018, I splitted the data accordingly. The data from March 2016 - May 2018 was planned to be used for training and data from June 2018- December 2018 was our testing dataset.
Predicting calls for violent crimes and property crimes:

To predict calls for violent crimes and property crimes, the base model was linear regression model. After fitting the model on training set , I tried to predict the events from testing dataset.The accuracy score of linear regression model was 81% for violent crimes and 89% for property crimes.





To know how good the models was, I used different evaluation metrics such as Mean Squared Error and R-Square. (Mean Squared Error is nothing but the difference between the observed value and predicted value, and R-square determines how much of the total variation dependent variable is explained by the variation in independent variable. In our case, event is the dependent variable and Year, Month, Date, Hour and Day are independent variables)

Later I implemented more models on our above problem and checked whether these models perform better than our linear regression model and below are the results.

Model comparisons:

	Linear Regression		Random Forest Regressor		Gradient Boosting Regressor	
	MSE	**r2_score**	**MSE**	**r2_score**	**MSE**	**r2_score**
Calls for Violent crimes	**2.60**	**0.811**	**2.54**	**0.815**	**2.16**	**0.843**
Calls for Property crimes	**20.46**	**0.890**	**23.38**	**0.874**	**21.71**	**0.883**

As we can see for violent crime call prediction, both the MSE and the value of R-square are less with Gradient Boosting Regressor model. Whereas for property crime prediction Linear regression model has small MSE and R-square value.

Final Results:

The following Insights can be given.

1. There is a spike observed in no. of calls for violent and property crimes during summer.
2. Most number calls are expected to happen at 3.00 pm so authorities can implement proactive strategies to reduce the calls in the future at this time.
3. Aggravated Assault is mostly reported from noon till midnight.In short, San Francisco is as safe as any big city neighborhood. But at night, it's wise for everyone to stick to streets with lots of foot traffic, or travel with an equally alert, sensible companion.