SAN FRANCISCO CRIME ANALYSIS

GROUP 9
MOHANA KODIPAKA
SHIVENDRA KUMAR
SNEHALATHA DODDIGARLA
SHREEANSH PRIYADARSHI
MANKARAN SINGH BAHRI
YU CHUANG TSAI

Introduction:

The "San Francisco Crime" dataset from Kaggle has data of crimes committed in 2016, it has details on category of crimes in San Francisco, date and time of the incident, district and exact location where the crime occurred and the resolution reached for the crime. This data set available in Kaggle is borrowed from Coursera and IBM's Data visualization' s site:

Data Set: https://www.kaggle.com/roshansharma/sanfranciso-crime-dataset

The question we are solving here is "Identify the most frequent crime category in San Francisco". Our project includes static graphs and interactive heatmap data visualization to plot exploratory data analysis, decision tree to classify crime categories and monte carlo simulation to calculate the most frequent crime categories.

Main findings:

- With more than 40,000 incidents, "Larceny/Theft" is the most frequent crime, this frequency is more than twice that of the second most frequent crime "Non-Criminal". Interestingly, the majority of these incidences had "None" as the resolution.
- A higher frequency of crimes are committed on Fridays and Saturdays with "Missing Person" having the highest incident rate on Friday
- "Larceny/Theft" has the highest frequency around 6-7PM while "Other Offenses" and "Non-Criminal" crimes have a higher occurrence around noon
- For more than half the crime categories the resolution captured is "None", which is uninformative and misleading
- Juvenile Crime Rate is the highest for "Secondary Codes" with a frequency of 5.3% which is twice the frequency of the second highest juvenile offence "Family Offenses"
- "Larceny/Theft", "Other Offenses" and "Non-Criminal" are the top three crimes in San Francisco based on Monte Carlo Simulation using predicted probabilities from decision tree model

There are a few other Kaggle data set kernels that we referred to get insights around data usage and conclusions derived. Other analyses were focused on exploratory data analysis, data visualization using matplotlib and seaborn. Few of the discussions entailed classification and developing prediction model using KNN or logistic regression.

Computational Setup / Steps of Outline of the steps:

Exploratory Data Analysis:

The focus was to visualize the data set and to understand how different variables impact the dependent variable i.e. crime category. Hence, we plotted various static and interactive maps, to execute this we relied on dictionaries and data frames. For 2 dimensional plots (i.e. 1 variable and its frequency plots) we used dictionaries and for plots with more than 2 dimensions we relied on data frames as it is simpler to pivot the data and cross tab the data. Few of the plots that we used are:

Heatmaps: Mapped crime category across day of the week showing which day of the week has the highest frequency of crime. To get this output, we filtered and retained top 10 crime categories and then grouped the data. The 'crosstab' function was used to get a pivot view of crime category and day of the week. The proportion of crime committed each day of the week was found by using the "div" applied for each crime category.

Line graphs: This chart highlights frequency of crime category per-hour, to get this output we used data frames. We sub set hour from the time column; grouped the data and chose top 10 categories to further pivot the data.

Bar plots: For most of the plots listed below, we used dictionary to aggregate variables and plot.

1. Crimes by District : This plot shows the total number of crimes for each district
2. Crimes vs. Category: This plot shows the total number of crimes per category. In order to better show the plot, we filtered out some categories that had a count less than 1000
3. Crimes vs Week of Day: This plot shows the total number of crimes happened for each day of the week
4. None Action Rate: Ratio of "None" resolutions to the total number of crimes in each Crime Category. We first pivoted the data and group by Category. We used the pivot table to create the plot
5. Arrest Rate: Ratio of number of arrest number divided by the total number of crimes in each Crime Category
6. Juvenile Crime Rate: Number of Juvenile Crimes divided by the total number of crimes in each Crime Category

We applied sort function across a few of the bar graphs to obtain trend and top categories.

Stacked bar graphs: This graph depicts crime category and resolutions for that crime category. To plot this graph, the data preparation was the same as that of Heat map chart. For the final output we included stack=true while plotting a bar graph.

Interactive maps: This heat-map shows the major crime location in San Francisco. Based on the static plot analysis, we got the data for the occurrence of top three crimes. The dynamic heatmap was made using folium package for this data which shows major crime location in San Francisco.

Prediction model :

We build our prediction model on decision tree classifier. In this dataset, the goal is to predict the possibility of each Crime Category by the several predictors. In order to run the model, we created dummy variables for Hour, Day of the Week, and PdDistrict to fit the decision tree and random forest model. We split 70% of the data into training data and 30% of the data into testing data.

Simulation:

After predicting the probability of a particular crime occurring in a particular location, we obtained the average of these probabilities for the 18th hour. Simulation is performed on the selected 30% test data. This average probability was used in a monte carlo simulation of 10,000 times for the

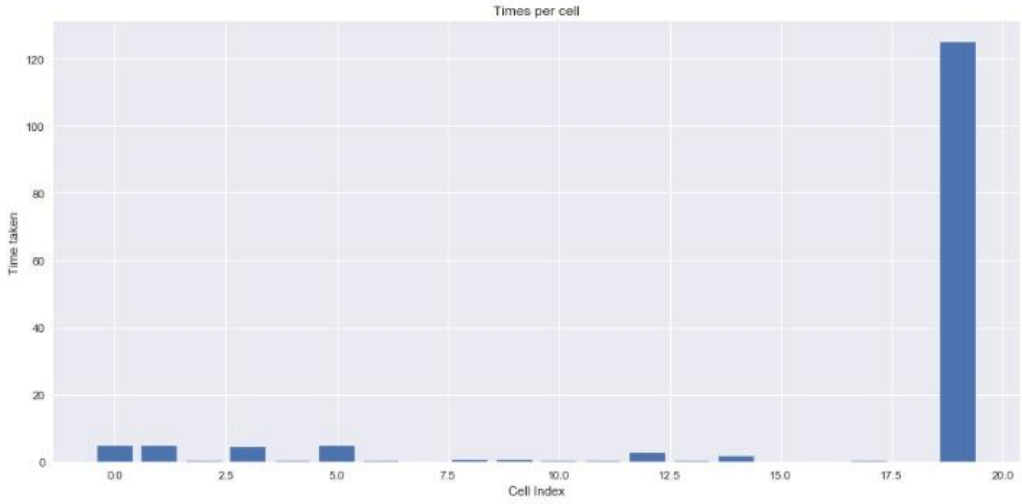
top 10 crime categories for each district. This simulation was performed using dictionaries and it helped us identify the most frequent crime category in San Francisco. This Simulation also helped us verify the crime that occurs most commonly for each district

Computational Challenges

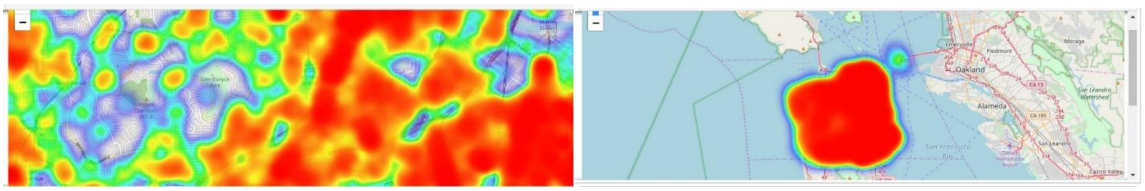
To explore data, we used dictionary to aggregate an attribute(s) and plot each exploration. A number of the EDA graphs were bar plots and we had to revisit this code to aggregate, plot data and make them consistent. In order to overcome this, we leveraged functions. We created a function to aggregate through dictionary, which makes it easier to use and another function to produce bar plots, which makes the graphs consistent and a single step process to plot. While doing Monte Carlo simulation, the main challenge was to use the probabilities of the occurrence of each crime category and then iterate it over each district to get the top three crimes in San Francisco. This was overcomed using multiple dictionaries instead of data frame wherever applicable in order to ensure faster computation, and using iterrows function.

Slowest part of your code:

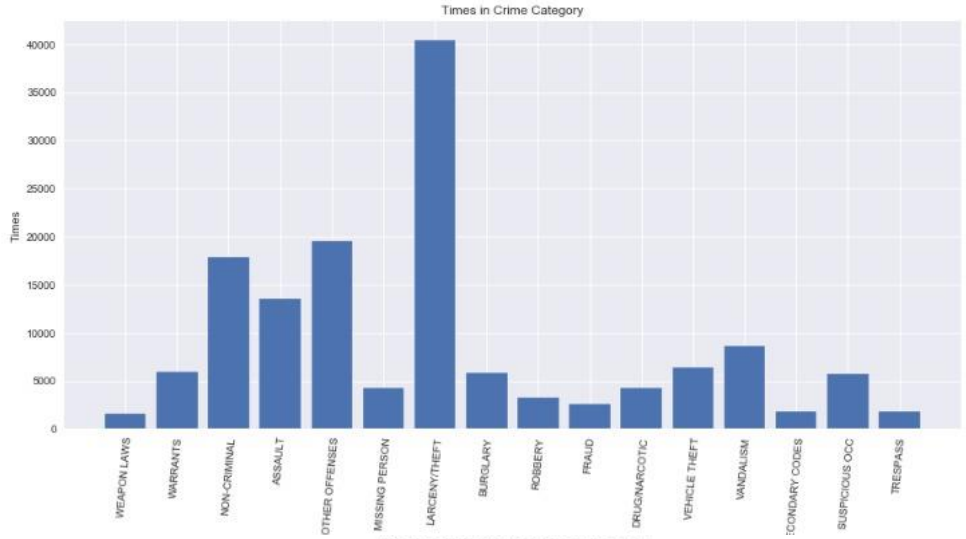
Monte Carlo simulation is the slowest part of our code which can also be seen from the graph. Since we are using data frame, and iterating over each rows for 10000 times, this makes it the slowest part of our code. We could have improved it by using dictionary instead of data frames.

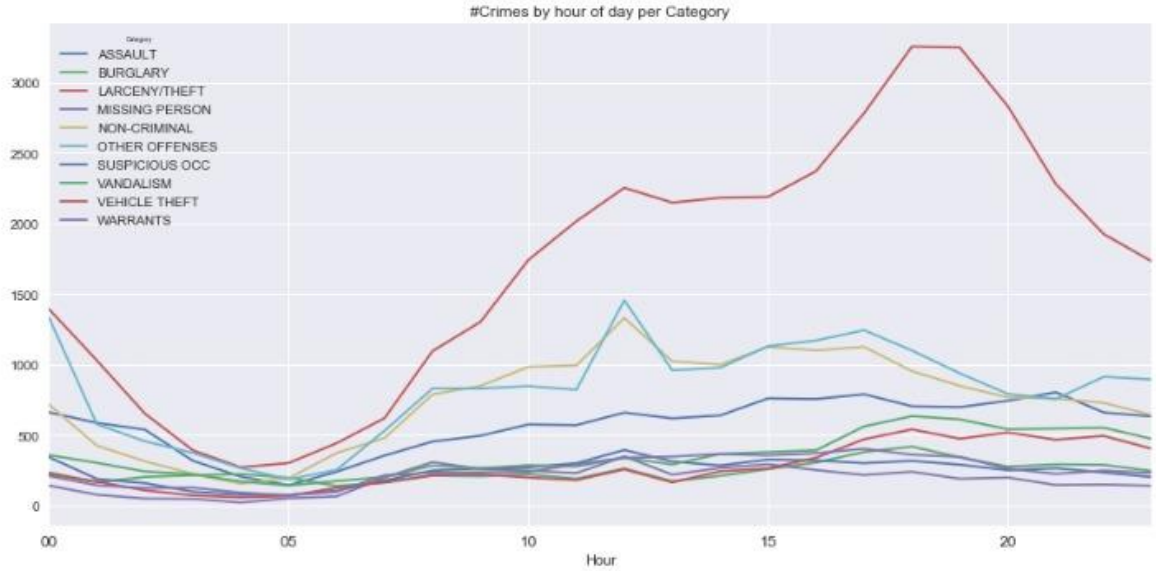


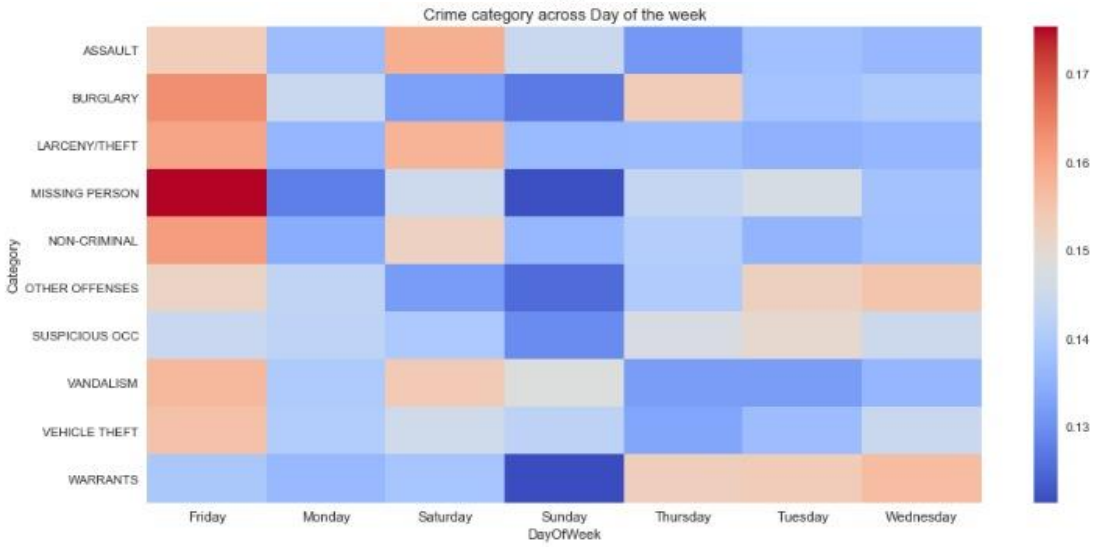
Results

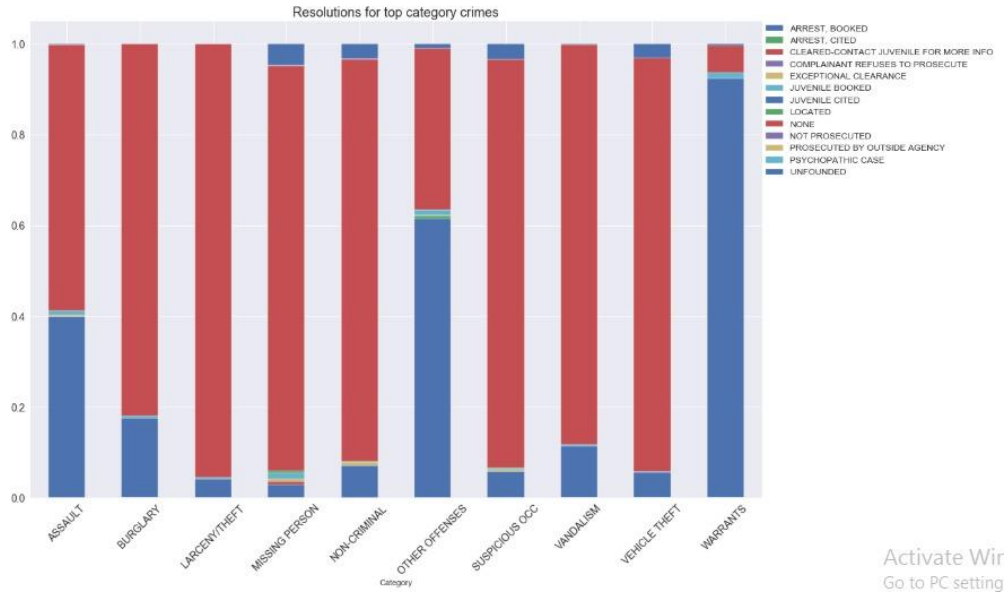


Heatmap of Crime in San Francisco


Times in Crime Category
Times
40000
35000
30000
25000
20000
15000
10000
5000
0
WEAPON LAWS
WARRANTS
NON-CRIMINAL
ASSAULT
OTHER OFFENSES
MISSING PERSON
LARCENY/THEFT
BURGLARY
ROBBERY
FRAUD
DRUG/NARCOTIC
VEHICLE THEFT
VANDALISM
SUSPICIOUS OCC
TRESPASS


#Crimes by hour of day per Category
Category
ASSAULT
BURGLARY
LARCENY/THEFT
MISSING PERSON
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS
3000
2500
2000
1500
1000
500
0
00
05
10
15
20
Hour


Crime category across Day of the week
ASSAULT
BURGLARY
LARCENY/THEFT
MISSING PERSON
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS
Category
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
DayOfWeek
0.17
0.16
0.15
0.14
0.13



EDA graphs highlighted that the top most crime category is "Theft" and the most imposed resolution is "None" for all crime categories. Most of the top crime categories are executed during the weekends specifically Friday and Saturday. Plotting the hour of crime for each category showed that "Larceny/Theft" has the highest frequency around 6-7PM while "Other Offenses" and "Non-Criminal" crimes occur most around noon.

Conclusion:

In conclusion, "Theft, other offences, non-criminal are the most frequent crime categories in San Francisco" is the main finding we derived by using simulations.

Hypotheses to Investigate Further?

Due to limited data, we could only identify "Most frequent crime category". Additional data regarding resolutions, larger time period would be useful to investigate further analysis around 'Crime rate drop/growth' and 'Whether a culprit was charged or not'.

Computational issues we foresee and possible solutions:

The addition of new variables will add complexity to the decision tree classifier. We will need to explore different data structures to ensure faster computation. The addition of more variables will also make it hard to identify the most important variable: we may need to do variable selection or use a boosting/lasso model that gives the most optimized model.

References:

1. https://www.kaggle.com/dbennett/test-map
2. https://www.kaggle.com/mircat/violent-crime-mapping
3. https://www.kaggle.com/dbennett/test-map
4. https://www.kaggle.com/wendykan/don-t-know-what-i-want-to-do-yet