Individual Project: What differentiates Crime Resolution?

https://colab.research.google.com/drive/1AS0bPdCshou674mPz_Z26DXTpOfGjjRb

Overview

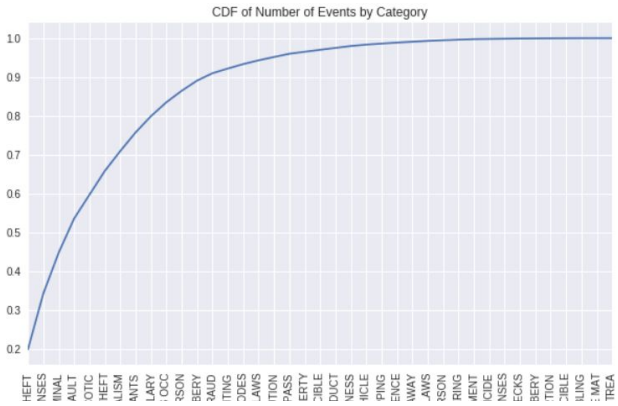
San Francisco was once known for housing some of the world's most notorious criminals on the island of Alcatraz in the 20th century. This project utilizes 12 years of crime data in San Francisco from Kaggle to predict the resolution of crime given location, time and category. In order to better understand the crime data, visualizations are also included.

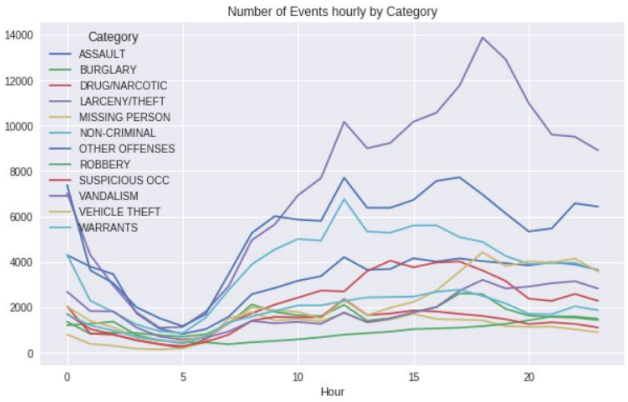
Technique

EDA

In this part, data cleaning is conducted and the attributes of variables are illustrated through visualization tools such as bar chart, line chart, word cloud, histogram and map. Through this process, the following five questions are answered.

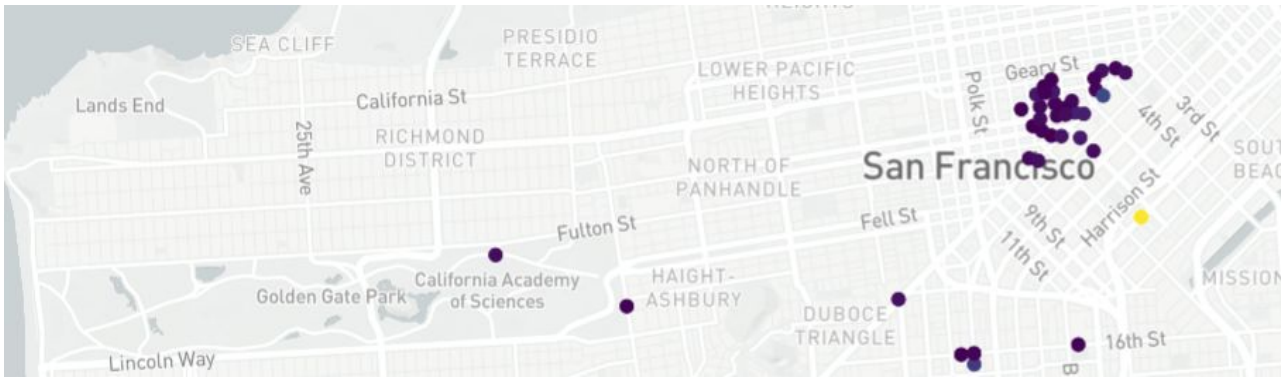
1. What are the top categories of crimes?
2. Which are the top district that has the most crime events?
3. Does Hour make a difference in the number of crime events? What is its correlation with District and Category?





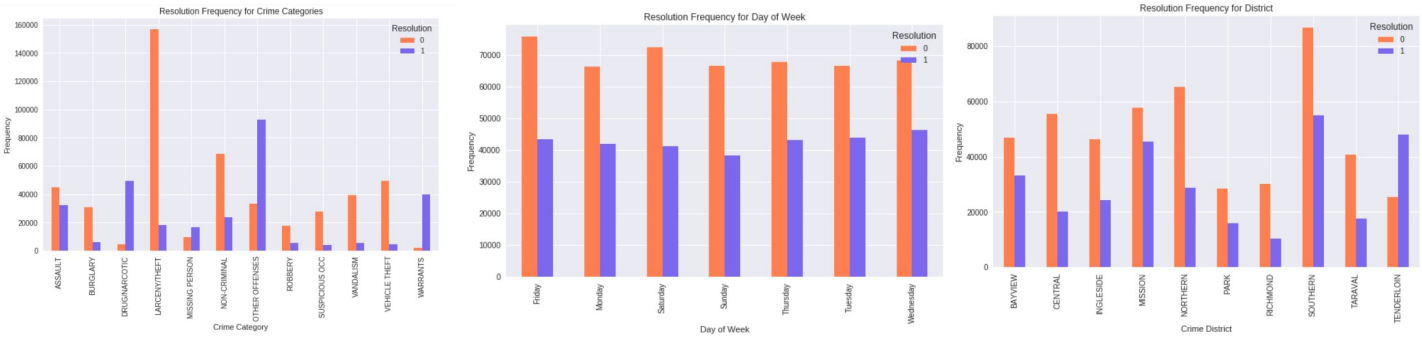
In this step, a cumulative distribution is plotted to show the main crime categories. We can see that 12 categories make up 90% of the crimes and we only retain the top 12 categories in the analysis.

4. What's the density of crime in the different street?



The graph above shows the address with a high rate of crime, most of which are concentrated in downtown. The lighter the color is, the more dangerous the address is. The light yellow point is the one with the most crime records(23k in 12 years).

5. What's the relationship between Resolution and district, crime category and day of week correspondingly?



In order to achieve the goal of identifying whether a crime will be resolved or not, the original nine values of resolution are reassigned to only two values, 0 and 1('None' :0, the others: 1). And we can see from the above three graphs that the

Modeling

With four features("DayOfWeek", "District", "Hour" and "Category) as input, two classification models, logistic regression and KNN, are applied on the data. By comparing the accuracy of each model I finally choose KNN(18) as the best model and the accuracy is 80.50%.

Conclusion

Classification Modeling

- KNN is a relatively good model in this case, with an accuracy of 80%; class '1' is more difficult to predict than class '0'
- Logistic does a bad job using the default parameters, with all predicted to 0

EDA

- Southern Area has the most crimes and Richmond has the least
- 8 categories(out of 39) make up 80% of the crimes; 12 categories make up 90% of the crimes
- Number of Events Hourly by District: nothing append at 5 am, sudden rise at noon and the majority of them occur around 6 pm
- Number of Events Hourly by Category: nothing append at 5 am, most of the events occur at noon; Theft occurs most around 6 pm
- 800 Block of Bryant St has far more crime events than the other Address, as high as 22k(the average is 32)

Other Insights

- Two ways to improve the efficiency of analysis:

  1.focusing on important values of a certain variable 2. regrouping the original categories