School of Information Studies
SYRACUSE UNIVERSITY

SAN FRANCISCO CRIME CLASSIFICATION

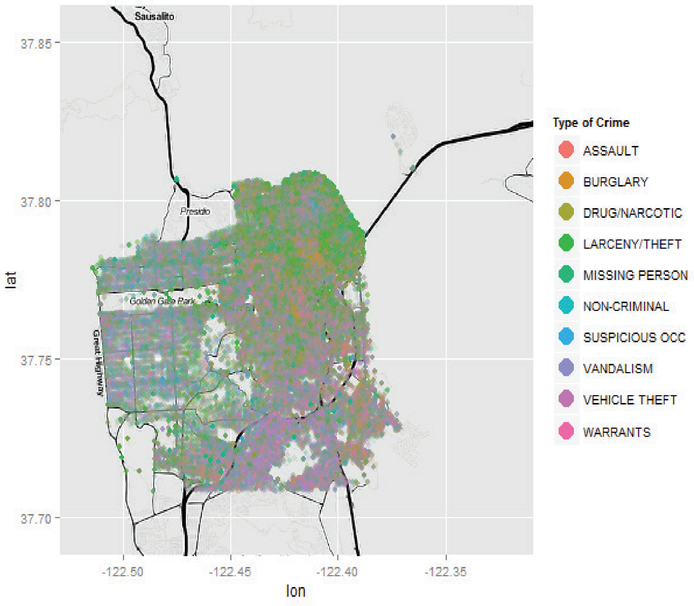
DATA QUESTIONS: Which crimes in SF City require maximum attention? What are our top crimes, district-wise? | Which days of the week experience the most crime occurances? | How are our crimes categories co-related?

Problem Description:

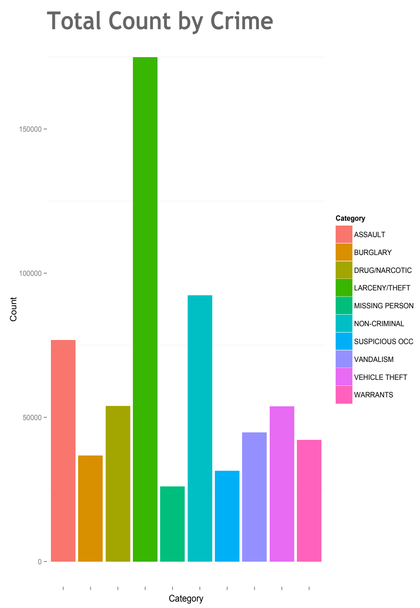
From 1943 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more fir its techology scene than its crimninal past. But, with the rising wealth inequality, housing shortages, and a profileration of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. From Sunset to SOMA, and Marina to Excelsior, this dataset provides nearly 12 years of crime reports from actross all of San Francisco's neighborhoods. Our objective here is to predict the crime category, given a day, and location.

GOAL: The prediction of crime is a step towards a safer society. Our system will aid police in stopping crimes preemptively by allowing them to better assess likely scenarios. Knowing which crimes are more likely to happen, based on date, time, and location, will give law enforcement the necessary tools to better society.

TOP 10 CRIMES - SPATIAL ANALYSIS (Figure 1)

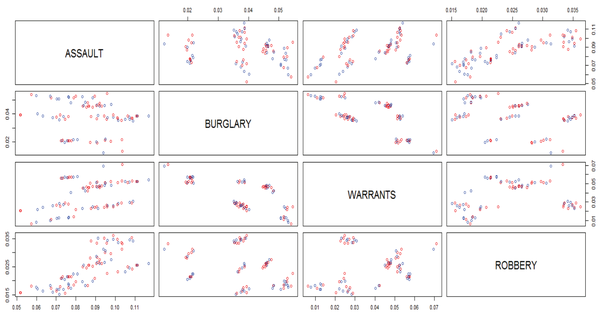


TOP 10 CRIMES - CRIME COUNT(Figure 2)

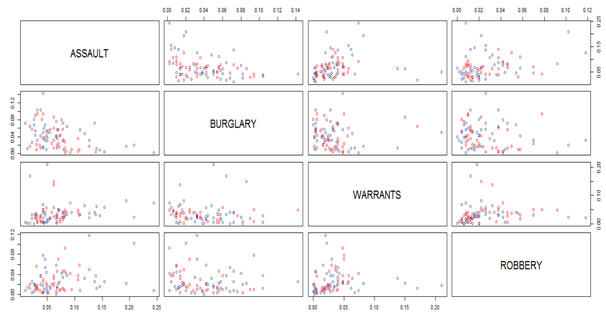


- Figure 1 and 2: The first two charts explain the spread of the top ten ten crimes that have occurred over the twelve years. After spatially mapping the data, the bar chart is used to get an accurate count of each crime category.
  From the two graphs it evident that Larcency/Theft is the most frequently occurring crime and it is also the most spread out in terms of location.
  — The average crime count is 632929 and larceny accounts for 27% of the top ten crimes
- Figure 3 and 4: The below charts are used to categorize the crimes based on weekday and district. The crime category is evaluated against each of the districts and week day.
  All twelve years of data is used for this analysis. We can find some interesting patterns using these graphs.
  For example, Drug and narcotic related incidents have occurred the most in Tendorlin and the Southern district is the most affected by theft.
  We can also find that vehicle theft seems is relatively more on the weekends than on the weekdays.
- Figure 5 and 6: MODEL BUILDING FOR PREDICTION
  We created a regression model and two data mining algorithms for our analysis.
  The multiple predictor regression model was run against each category of crime and the model results were summarized.
  The model was used to predict results using the test file. The random forest algorithm was used to find the probability of each type of crime against each test record.
  The figure 5 and 6 explain correlation between the predicted outputs. For example, in both models assault and warrants are highly correlated suggesting an increase in warrants may lead to an increase assaults. Likewise the plots can be used to find correlations between multiple categories of crime.

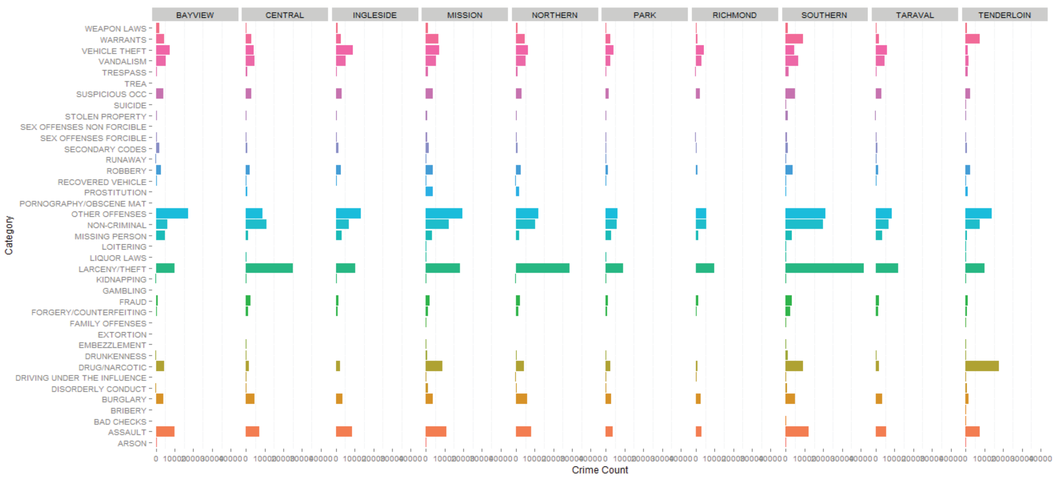
REGRESSION MODEL - PREDICTION PLOT (Figure 5)



RANDOM FOREST MODEL - PREDICTION PLOT (Figure 6)



CATEGORIZED CRIME COUNT BASED ON DISTRICT (Figure 3)



CATEGORIZED CRIME COUNT BY DAY OF WEEK (Figure 4)