Crime in Los Angeles

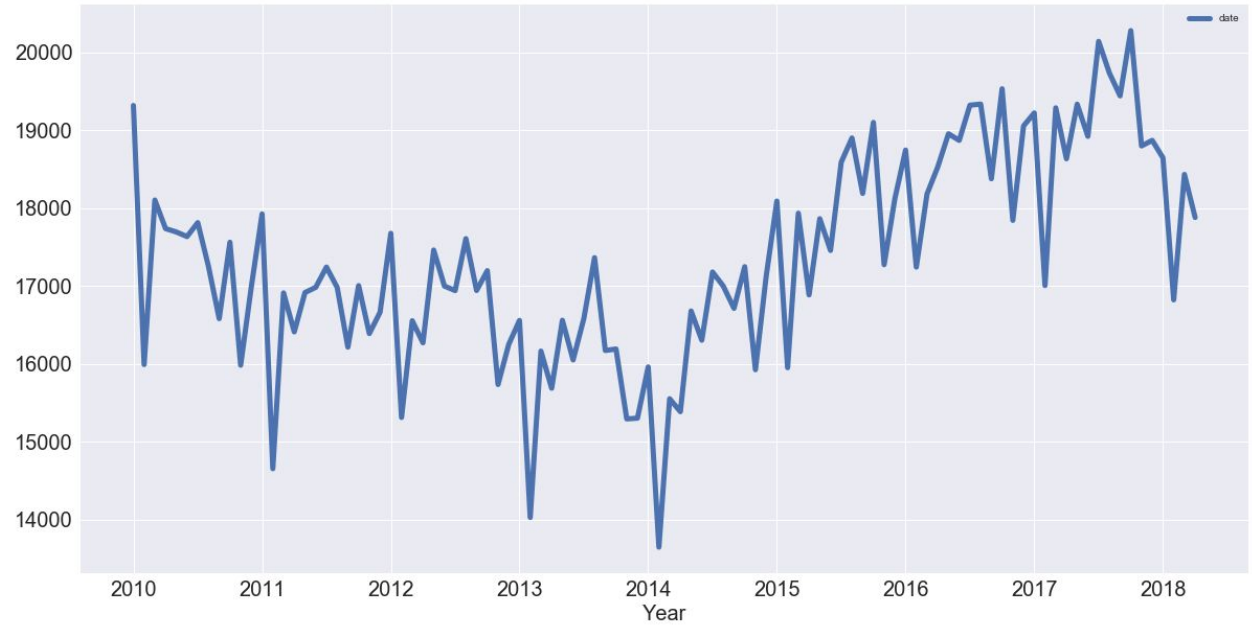
By Jai Bharat Davé

As someone who was born and raised in Southern California, I am deeply interested in helping alleviate the issues we see in our communities. My goal for this project is to develop my data science abilities, conduct a research project from beginning to end, and display my passion for public service.

I obtained the data from the LA city's website. The dataset reflects incidents of crime in the City of Los Angeles dating back from 2010. The data is transcribed from original crime reports that are typed on data, therefore there may be some errors. The provider of the dataset is the Los Angeles Police Department (LAPD) and is refreshed weekly. The dataset contains 1.74 million rows where each row is a crime incident. It contains the date reported, the date occurred, location information, crime description, and victim information. The data is available via API.

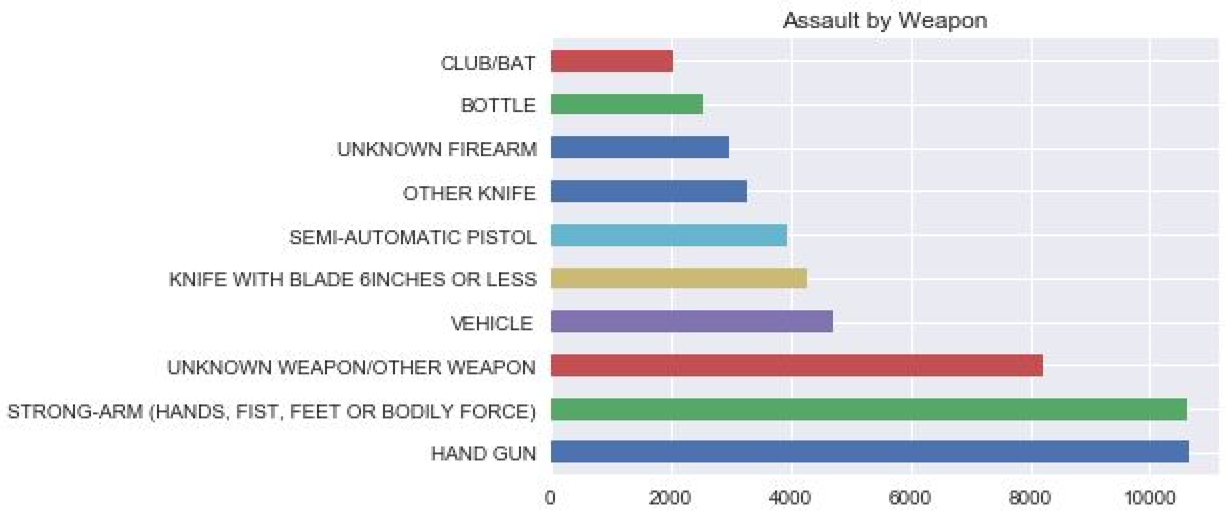
Analyzing the data will be useful to learn more about crime in Los Angeles. Examining the data could lead to finding trends, patterns, and possible ways to lower crime. Police officers could use this data to be more efficient with their patrols, Captains can use it to create procedures to make departments more effective, and policy makers could use it to create legislation that will help alleviate crime.

I investigated where and when crimes are occurring and which populations are more susceptible to being victims of crime. Since the dataset has crime data dating back to 2010 the first thing I looked at was how the crime rate has changed over time.



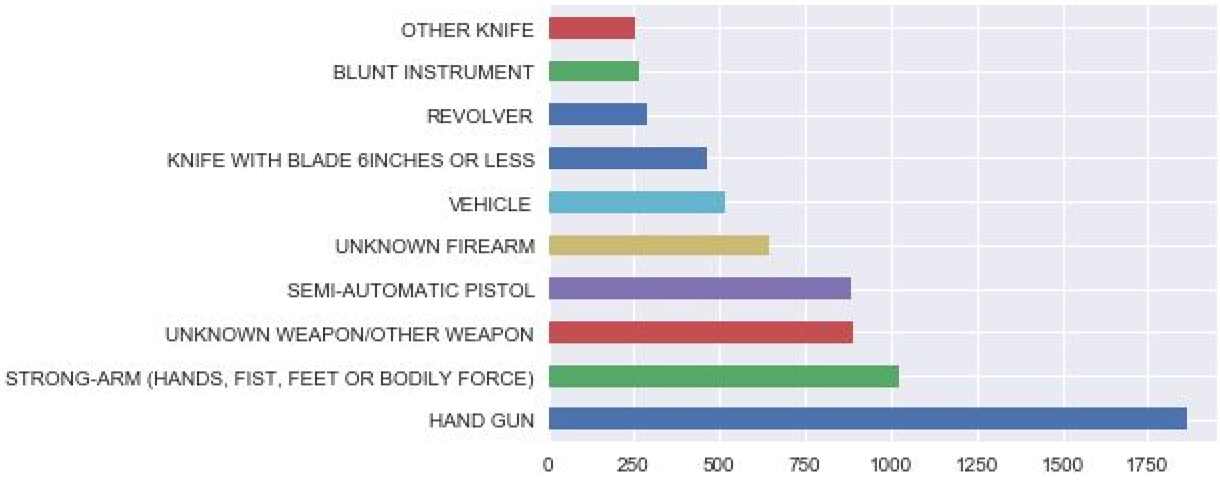
Crime rate appeared to reach it’s low in 2014 and has risen steadily since then.

After getting a sense for the data, I began looking at the basic information, such as what are the most common crimes that occurring in Los Angeles. The ten most frequent crimes in LA are: Battery - Simple Assault, Burglary from Vehicle, Vehicle Stolen, Burglary, Theft Plain, Theft of Identity, Intimate Partner - Simple Assault, Vandalism - Felony, Vandalism - Misdemeanor, and Assault with a Deadly Weapon. I then decided to examine assault with a deadly weapon more closely to see what weapons are most commonly being used.



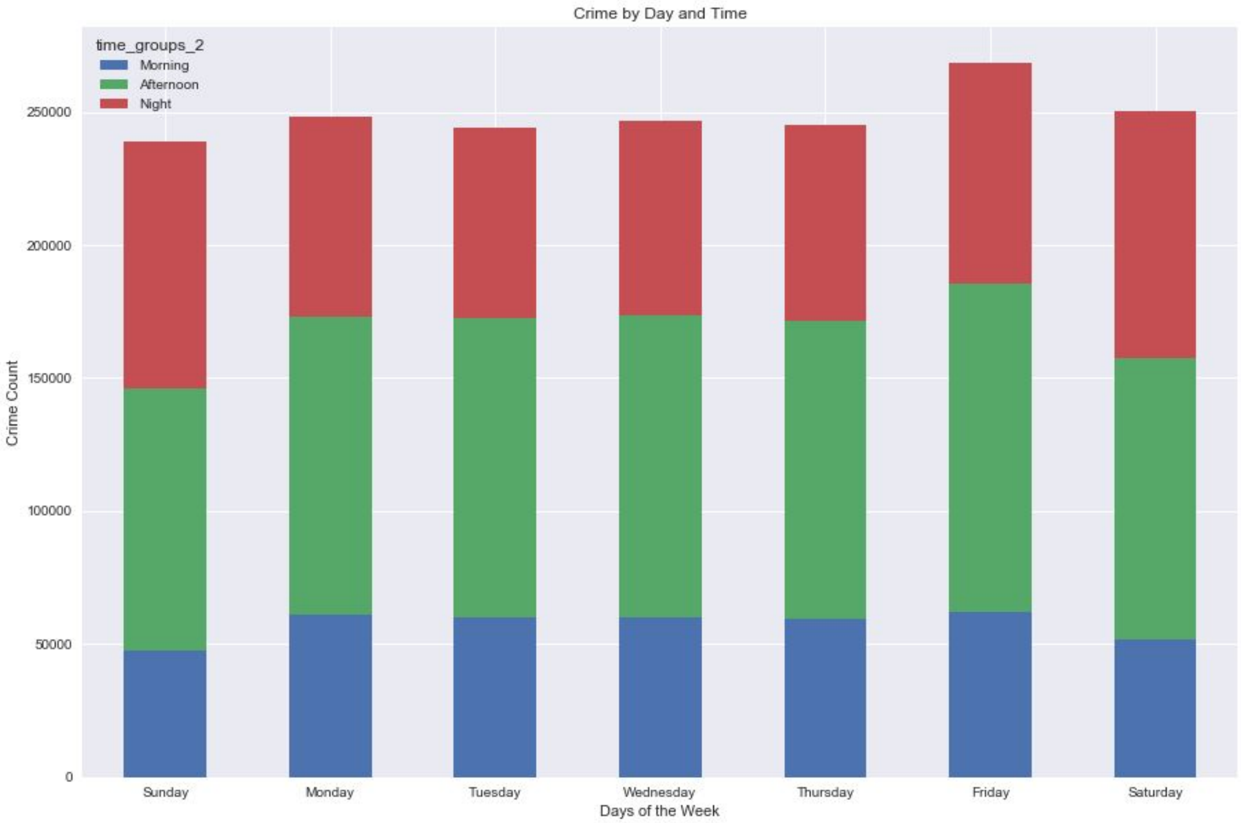
The data shows that handguns are the most commonly used weapon for assaults. Los Angeles has an anonymous gun buyback program. Perhaps by creating similar programs or increasing awareness of the program would reduce the number of assaults with the use of a gun. Making it more difficult to acquire a gun may also help to lower crime.

After examining what kinds of crimes were occurring, I looked at where crimes are occuring. The data shows two areas with significantly more crime than the others. Those two were the 77th street area and the Southwest area. Focusing on the 77th street area, an area in the Southern part of LA, I conducted a proportion test using the stats models. The result was statistically significant. I then determined what crimes were more common in this area and it seemed that assaults were more common than in the general data. I conducted a similar proportion test and determined that it is likely that the proportion of crimes that are assaults in the 77th street area is greater than that of total dataset. Below are the most common weapons used for assaults.



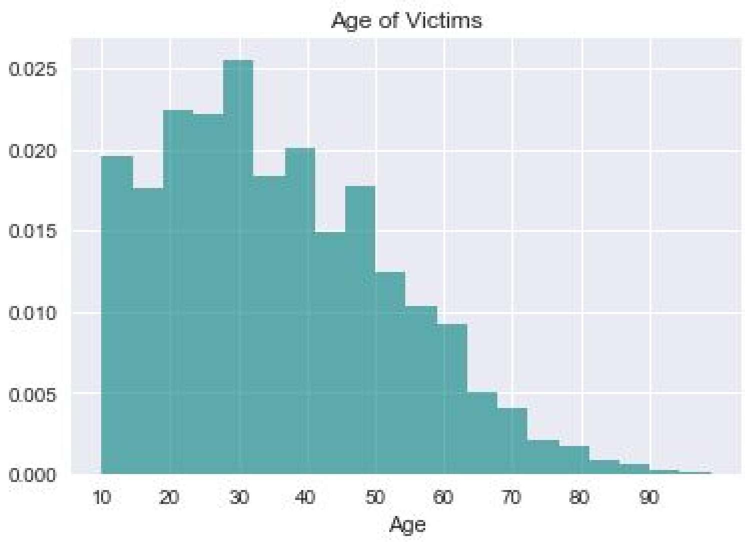
Handguns are by far the most commonly used. From the graph above we can see that the number of assaults with the use of handguns far far surpasses all other types of assaults. Focusing on getting guns off the streets targeted at this area may help lower the number of assaults with the use of a firearm.

After examining what kinds of crimes were occurring, I investigated when crimes were occuring. I put the time column into buckets and created a column indicating the day a crime occurred. I then created a stacked bar chart to show the count of crimes on days and time ranges.



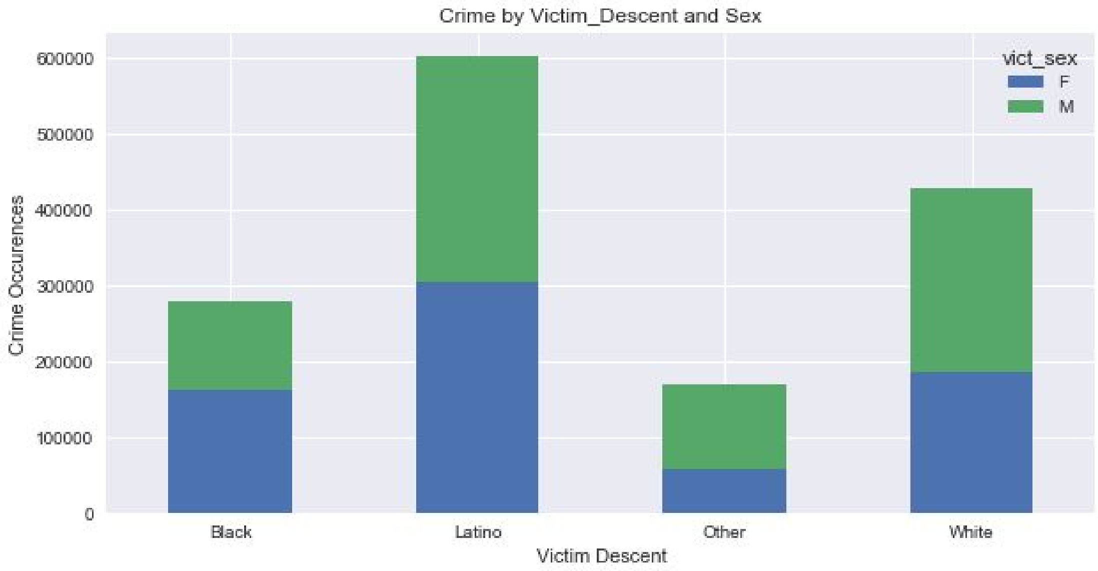
From this graph, it appears that Friday has a higher crime rate than other days. Conducting a proportion test confirms that Friday has a significantly higher crime rate than other days. It appears the difference is in the afternoon. Looking at the most common crimes that occur on Friday afternoons it appears that different kinds of theft are more common. Perhaps this is because individuals have debts that are due at the end of the week and resort to crime out of desperation.

Finally, I studied the data to see what subpopulations where more commonly victimized as a result of crime. To begin I looked at the distribution of victim's age. The average age is 35.98 and the median is 34 which is very close to the median age of Los Angeles as a whole, 34.6.



From this histogram we can see that the portion of the population that are the most victimized are those in their late 20s and early 30s and the risk of victimization appears to drop as individuals get older.

After looking at the distribution of age of victims, I calculated the amount of crimes by victim descent and sex.



Victims are more likely to be hispanic, white, or black. Among black victims females outnumber males. This could possibly be because black victims are less likely to report a crime. This leads me to believe that the number of crimes against the black population is greater than what the data shows. Among white victims it appears that white males are more likely to be victim of crimes than males.

Finally I examined what crimes are most commonly taking place against the homeless population. The mocode 1218 refers to homeless individuals. Three of the five most common crimes experienced by homeless individuals is assault suggesting that a larger proportion of crimes against homeless individuals are assaults.

From my analysis, we can see how the crime rate in Los Angeles has changed over time, when and where crimes are occurring, and who are most commonly victimized by crime. From this we can determine how best to act to alleviate the issues of crime.

After examining the data I wanted to explore whether crimes occur in a systematic way. My hypothesis is that there is a relationship between where a crime is going to occur and when a crime occurs and the type of crime. If a relationship exists we can predict where crimes are going to occur before they happen.

The first thing I needed to determine was my response variable. The dataset contains two geographical data points: the area (district) and latitude and longitude area. The problems with these is that the area is too broad and contains too large of an area so would not be helpful to predict, and latitude and longitude is too specific and would be difficult to predict with accuracy. To get around this I converted latitude and longitude into zip code information. One problem with this is that there are observations that are missing latitude and longitude data. There was no apparent trend for the missing data. I believe these are missing because police reports weren’t being filled completely. The data appeared to be missing at random and so I dropped the rows with missing latitude and longitude data.

Once I determine the response variable I had to decide which features to include in my model. The first variable I wanted to include was time. When modelling I tested to see which groupings are the most effective. The two groupings I tested are three hour buckets and one hour buckets. In all cases using a three hour bucket led to more accurate models.

The second feature I examined was the crime type. The column name corresponding to this feature is “crm_cd”. There are too many distinct crime types to include include in a model after creating dummy variables. To include this in my model I included the top ten types of crime and grouped the rest as other. Having so many crime types grouped together led to it being less effective.

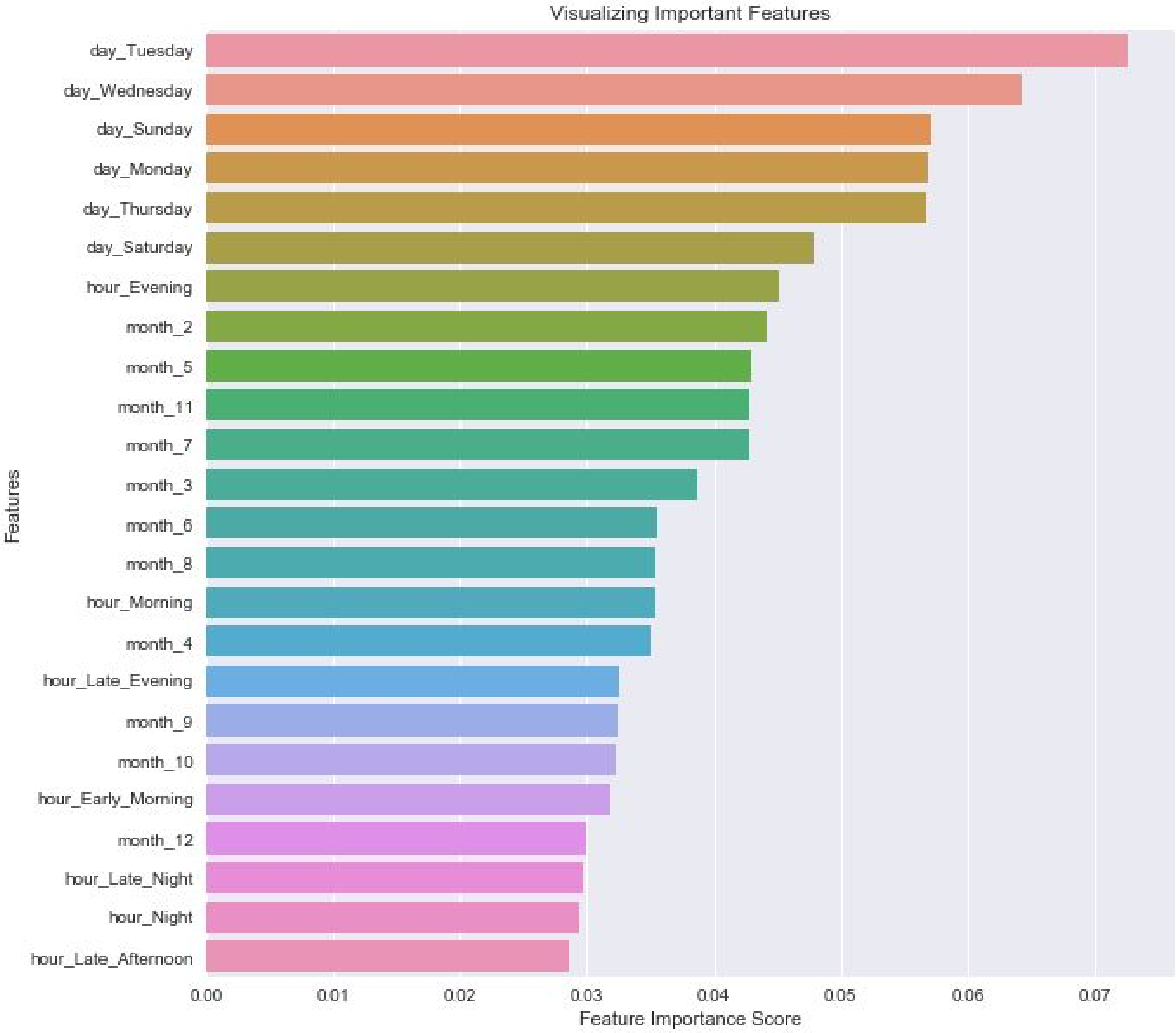
Other variables I tested in my model were area_id which corresponds to the district of interest, the day of the week, and the month. Once I’ve finalized my set of features I need to make the data usable for machine learning. To do this I used one hot encoding for all my features which creates new columns for each level of each feature filled with dummy variables. I also converted my label column, zipcodes, into numerics.

Having numerous categorical variables each with many features makes the dataset extremely wide. To ensure none of the variables are redundant I test for multicollinearity between all the variables using VIF. None of the VIF values were above five suggesting that none of the variables are correlated.

Once I had my features and labels ready I needed to choose which model to use. For a classification with many observations, I decided on using a random forest. First I split the data into training and test sets with 25% of the data held out for testing. I then imported the Random Forest Classifier from sklearn and the created an estimator with 10 estimators. I then trained the model using the training sets and then assigned predictions based on the X test variable. Finally, using the metrics module from sklearn I calculated the accuracy of the model.

After testing different combinations of features and levels I settled on three different models using a Random Forest Classifier. Having limited computing power I had to weigh the tradeoffs between number of features and number of levels per feature. For example, to include all levels of crime I had to collapse time and wasn't able to include months. After testing different combinations, I concluded that grouping time into three hour buckets was more effective than one hour, grouping crime type was ineffective. After testing I settled on three different models.

For the first model I included: area_id, all crime types, time collapsed into three hour buckets, and day of the week. This model had an accuracy score of 29.42%. The second model I swapped out crime for month resulting in an accuracy score of 29.01%. The last model was I created a model on a subset of the data. I isolated all the rows for the 77th Street District, the district with the highest crime rate and created a model using day of the week, month, and time of day. This model had an accuracy score of 32.27% with day of the week having the highest importance, followed by month and time of day. Below is a plot illustrating the importance of different features.



After testing different combinations of features I was not able to create a model that could reliably predict where a crime would occur. For future projects I would like to examine what populations are more susceptible to particular crimes, perhaps, creating a predictive classification which predicts the descent of a victim based on the type of crime and area. In addition to this I would like to combine my current dataset with details about each zip code and see what characteristics of a zip code result in higher crime rates.

My initial objectives was to discover trends in crime data, specifically, where and when crimes occur and who are more prone to being targets of crime. In the end, I was able to identify which areas and times that have statistically significant more crime and what subpopulations are more prone to crime. Another goal I had was to predict where a crime would occur based on other variables using a Random Forest classifier. In the end the most successful model I was able to create only had an accuracy rate of 32.27%. Although I was not able to create a model which reliably predicted where a crime would occur, I believe my overall analysis would be useful to police officers in determining which areas and subpopulations are more prone to crime and policy makers to create legislation that focuses on the more frequent crimes.