Analysis of Dangerous Crimes in Boston

By Jacquelyn Andrade, Yao Zhang

Introduction:

In our project we analyze the types of crime that occur in given locations around Boston on specific times of day. We perform K-means clustering on our data to classify our districts by the most popular crime. After performing this technique, we provide an in-depth review of particular statistics of crime that occur within a day, and whether time can indeed be used to classify the occurrence of a crime in a district. We use regression techniques to be able to predict the amount of occurrences of a specific crime as well as the probability of the crime occurring out of all crimes within a district. We expect our findings to closely correlate to "typical" crimes associated with certain times of the day.

Datasets:

- Crime Incident Reports (August 2015 - To Date)
  - From City of Boston Data Portal
  - Selected Dangerous Crimes: Larceny, Manslaughter, Driving Under the Influence, Rape/Sexual Assault, Robbery, Assault/Battery, Kidnapping, Burglary, Bombs/Explosives, Harassment, Drugs, Weapons, Arson, Breaking and Entering, Child Abuse/Endangerment, Disorderly Conduct, Public Drinking

Methods:

K-Means:

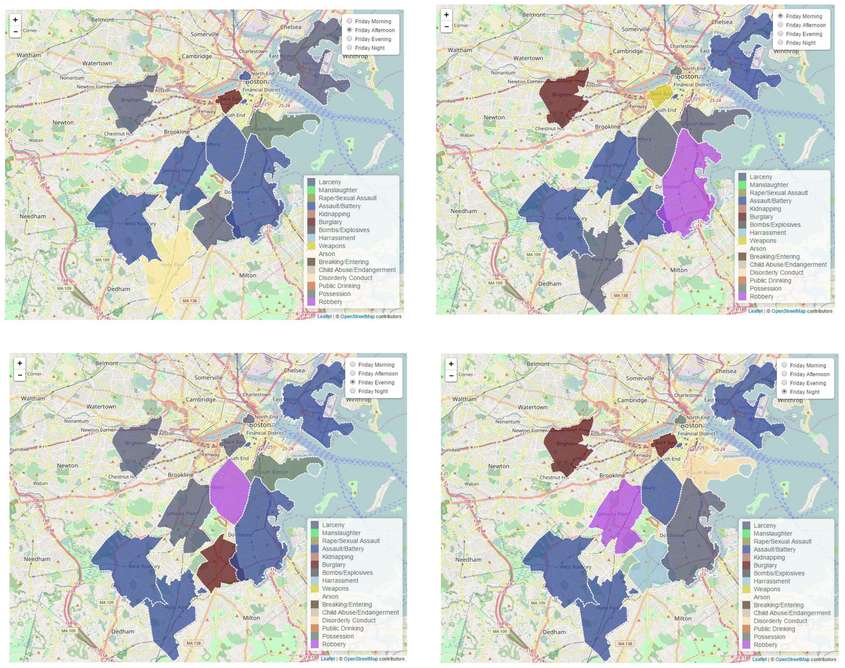
- We decided on using the K-means algorithm to classify our districts by crime popularity. K-means uses hard assignments, such that each point belongs to one cluster or another but never belonging to more than one.
- We wished to distinguish data points specifically for districts without having clusters mix points together.
- Our K-means technique vectorizes on district and offense type. We then found the top unique district and crime type per cluster on a specified time and day of the week.

Regression:

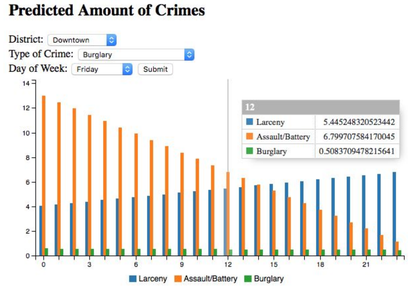
- The purpose of using regression techniques is to test whether time is a good predictor for the types of crime that will occur. This is possible by finding direct predictions on the amount of crimes and the probability of each one occurring.
- For our linear regression technique, we predict the number of occurrences for a crime given the district, day of week, and time.
- Using the logistic regression technique, we analyze the probability of a certain crime occurring given the district, day of week, and time.

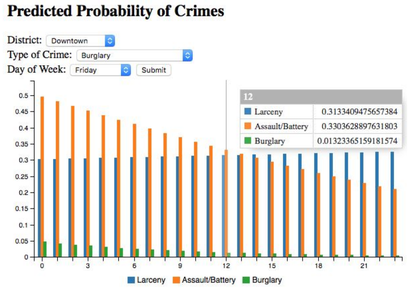
Results:

K-Means:



- In attempts to classify police districts by their popular crimes, our K-means findings coincides with our hypothesis that "normal" crimes would be prevalent in all areas of Boston. These crimes are not specific to location and would occur no matter the time.
- Additionally we find that in certain districts (e.g. South Boston) the criminal activities that occur correlates with what we define as "stereotypical" human behavior. In the South Boston area, "Stereotypical" human behavior on Friday nights can be defined by night activities such as drinking and celebrating. Such activities often involve irrational decisions that more or likely would disturb public peace thus leading to complaints of disorderly conduct.



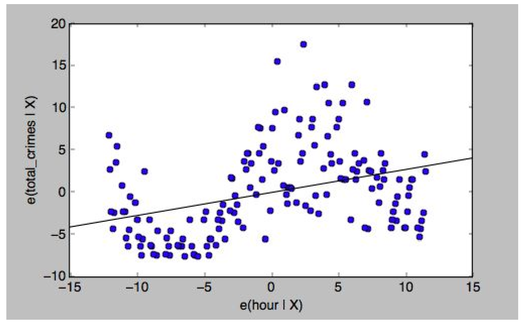


Linear Regression:

- In our linear regression technique, we experimented on the types of crimes that would be used for our regression function. Our linear regression function returns the predicted amount of times a specific crime will occur. We find that in crimes less common in the dataset, the R squared value is low and reduces the accuracy of our predictions. This is expected as our function is not be able to accurately predict the amounts of crimes if there is not sufficient data to spread across all days and hours of the week. For more prevalent and popular crimes, we find our regression statistics to be more informative and accurate.

Logistic Regression:

- In our logistic regression technique, we experiment on the types of crimes to be used in our regression function. Our logit function returns the probability of a specific crime occurring as opposed to all other crimes. We use our dataset on previous crimes to train our regression function in order to make these predictions. We are able to do this with a greater amount of accuracy.



- In each regression graph, our data points deviates significantly from the regression line, which indicates that time can not be the only predictor in attempting to predict human behavior. Our points indicate that our original data does not appear to fit a linear model as we intended it to, and that in future work it would be best to experiment with non-linear models.

Conclusion:

In the future, we hope to increase our number of datapoints in recent crimes and introduce other predictors of crime. We hope that by increasing the number of datapoints, we will increase the accuracy of our regression techniques, as the number of crimes will deviate to a certain number of occurrence at a particular time and day.