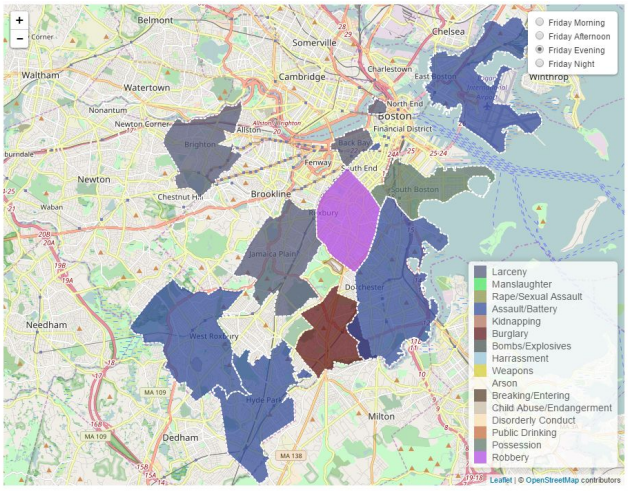
Analysis of Dangerous Crimes in Boston

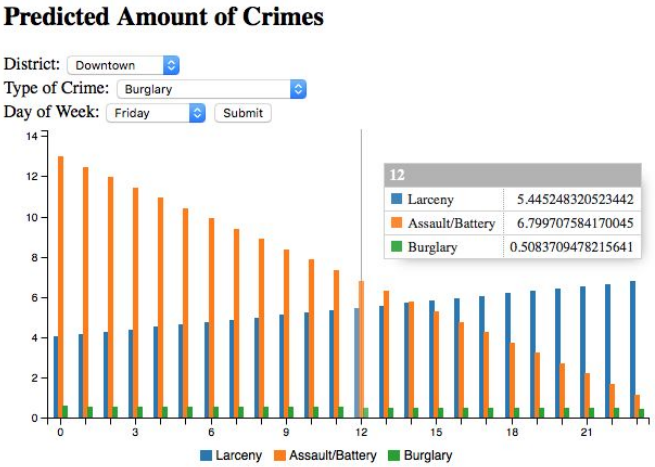
Jacquelyn Andrade, Yao Zhang
CS505 Computational Tools for Data Science
Boston University Computer Science Department

1. Abstract:

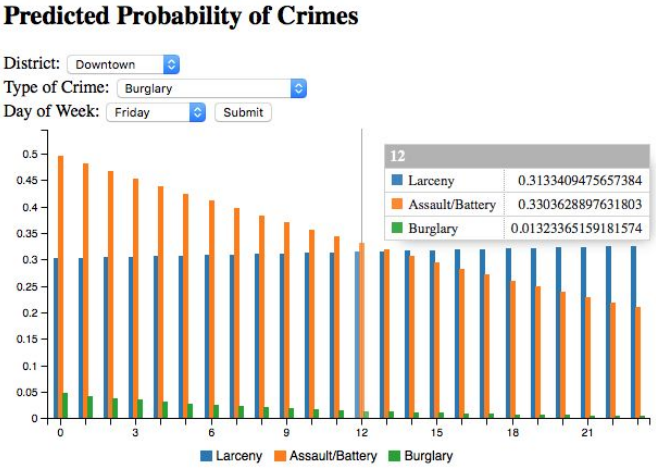
In our project we analyze the types of crime that occur in given locations around Boston on specific times of day. We perform K-means clustering on our data to classify our districts by the most popular crime. After performing this technique, we provide an in-depth review of particular statistics of 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. For example, we do not expect to find a substantial amount of crimes such as DUI in the morning, whereas robberies are more likely to occur at any time of day.



Sample Output of District Map for Friday Evening



Sample Output of Predicted Crimes for Friday in Downtown



Sample Output of Probability of Crimes for Friday in Downtown

2. Introduction:

We have created an interactive user application that will allow our audience to

query search results from our data mining techniques. The user is able query the predicted amount and probability of crimes— in relation to all other crimes—given a district, day of week, and hour. Additionally, the user is able to compare these crimes together on the same visualization. Our categorical and predictive analysis will hope to benefit the Boston Police department in granting more insight into the nature of crimes prevalent in certain areas and improving police response. Given a specific time, day, and district, we hope to be able to provide an appropriate prediction for police to account for the likelihood a certain crime will occur.

For this project we are using the Crime Incident Reports (August 2015 - To Date)[1] dataset downloaded from the City of Boston open data portal. This dataset is the most current and applicable to our analysis. The main motivation of this project is to derive and analyze potentially "dangerous" crimes from this dataset such as:

- 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

[1] Source: https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports-August-2015-To-Date-Source-/fqn4-4qap

We are analyzing the types of crime that occur in given locations around Boston on specific times of day and classifying each district using K-means. We also utilize regression techniques to find correlations and predictors of crime.

The time of day specific for our K-means technique is defined as:

- Morning: 6:00 - 11:00
- Afternoon: 12:00 -16:00
- Evening: 17:00 - 19:00
- Night: 20:00 - 5:00

We use integer values to identify the days of the week:

- Sunday: 1
- Monday: 2
- Tuesday: 3
- Wednesday: 4
- Thursday: 5
- Friday: 6
- Saturday: 7

3. Technique:

3.1 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. Another clustering technique, Gaussian, makes soft assignments: Data points do not belong to a single cluster but have some probability of belonging to each individual cluster. This means some points may have a high probability of belonging to the same category due to their vicinity with each other and similarity of categories. We determined the Gaussian mixture model to be inappropriate for this project because 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.

3.2 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.

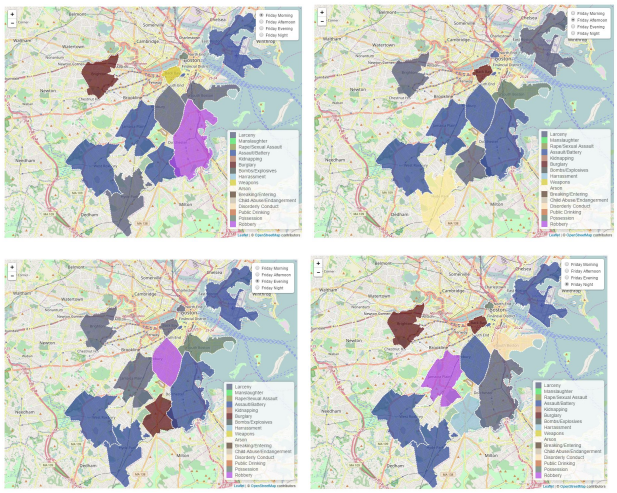
4. Datasets and Experiments:

4.1 Dataset:

This dataset contains field values for the crime incident type, time, day, month, year of incident, weapon used, true false values for whether shooting was involved and if the incident was domestic, along with the coordinates, street name, reported area, and district. Pre-processing is necessary as we need to eliminate row values that are null as well as incident types that are not relevant to our study.

4.2 K-Mean Experiments:

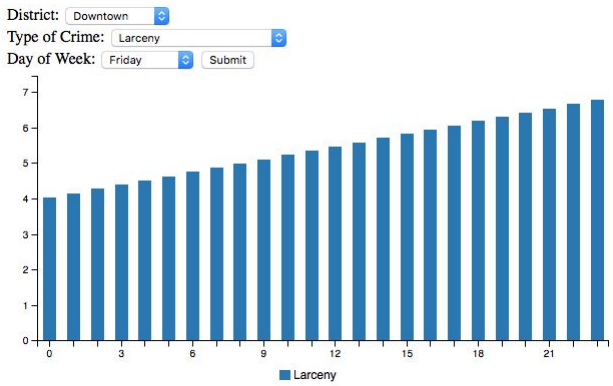
We experimented on the the number of clusters used in K-means and found the value of k with the minimal error to be 12, the total number of unique districts in our dataset.



From left to right on a Friday: [Morning, Afternoon, Evening, Night]

4.3 Linear Regression Experiments:

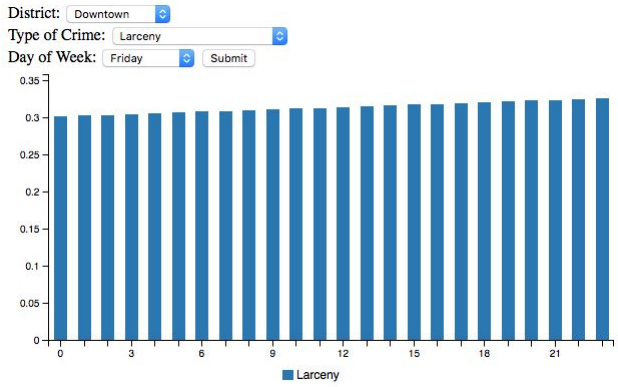
In our linear regression technique, we experimented on the types of crimes that would be used for our regression function. Our linear regression function returns the predicted amount of times a specific crime will occur. We find that in crimes less common in the dataset, the R squared value is low and 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.



Example of Larceny for Friday in Downtown

4.4 Logistic Regression Experiments:

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. Overall, we find Logit Regression to be more informative to the user as opposed to OLS. For crimes that have little to no occurrence, Logit returned a probability close to zero whereas OLS predicts that the crime would occur at least once even if the crime was very unlikely to happen.



Example of Larceny for Friday in Downtown

In both OLS and Logit Regression techniques, we notice that hour is more likely to be a predictor for crime than the day of the week. However, time is not a good determinant of the occurrence of a crime.

5. Results:

In attempts to classify police districts by their popular crimes, our K-means findings coincides with our hypothesis that “normal” crimes (e.g. Assault/Battery, Larceny, Burglary) 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.

Generally, we find that our regression analysis was not an accurate predictor of the occurrence of a crime. In our regression results for each type of crime, the R squared values are moderately low with a value of approximately 0.214 on average and our function value for our logit function is 0.64 on average. The reason for these values are that time and day are not necessarily good predictors of human behavior. There are potentially other factors (e.g. a person’s background) that need to be investigated that motivates an individual to commit a crime.

OLS Regression Results

Dep. Variable:	total_crimes	R-squared:	0.129
Model:	OLS	Adj. R-squared:	0.118
Method:	Least Squares	F-statistic:	11.61
Date:	Sat, 10 Dec 2016	Prob (F-statistic):	1.99e-05
Time:	13:38:20	Log-Likelihood:	-482.43
No. Observations:	3957	AIC:	970.9
Df Residuals:	3954	BIC:	980.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	4.9053	1.135	4.322	0.000	2.663	7.147
day	0.0810	0.198	0.409	0.683	-0.311	0.473
hour	0.2735	0.057	4.811	0.000	0.161	0.386

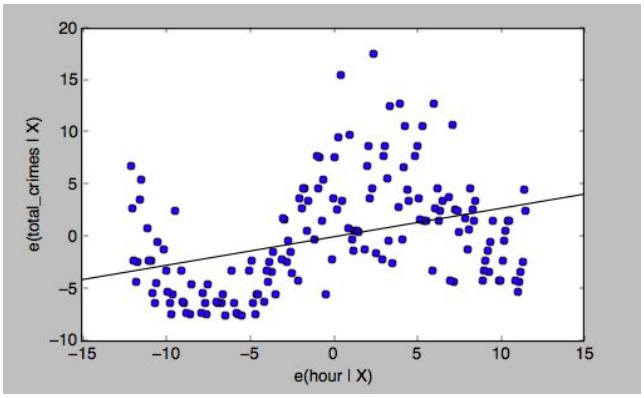
Omnibus:	16.700	Durbin-Watson:	1.101
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18.744
Skew:	0.819	Prob(JB):	8.51e-05
Kurtosis:	3.356	Cond. No.	41.2

Example of Larceny (OLS) on a Tuesday at 1:00 AM

```
Optimization terminated successfully.
         Current function value: 0.652456
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:                is_crime   No. Observations:                 3957
Model:                          Logit   Df Residuals:                     3955
Method:                           MLE   Df Model:                            1
Date:                Sat, 10 Dec 2016   Pseudo R-squ.:                -0.01528
Time:                        13:56:35   Log-Likelihood:                -2581.8
converged:                       True   LL-Null:                       -2542.9
                                        LLR p-value:                    1.000
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
day           -0.1406      0.012    -11.933      0.000        -0.164    -0.118
hour           0.0049      0.004      1.321      0.187        -0.002     0.012
==============================================================================
```

Example of Larceny (Logit) on a Tuesday at 1:00 AM

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.



Example of Larceny on a Tuesday at 1:00 AM

6. 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.

All our source code can be found here:
https://github.com/yazhang28/CS505-Project