Partitioning and Comparing Regions of Crime with Regions of Environmental Investment in Boston

Jonathan Silverman, Rohit Matthew

Introduction

A common metric for measuring the overall quality of a city is, not surprisingly, the crime rate. crime is a factor that statisticians and sociologists have linked to everything from overall Police volume and effectiveness, to income and school quality. By analyzing these correlations, one can better understand the contributing factors behind crime and hopefully implement meaningful policy to mitigate its effects.

For our project, we decided to analyze the relationship between Boston's environmental infrastructure and its overall crime rate. However, recognizing that there are historically "bad" areas for crime we realized that creating a one size fits all model for the whole city would not yield entirely meaningful results. Thus we came to the conclusions that k means clustering would be a good way to analyze this information.

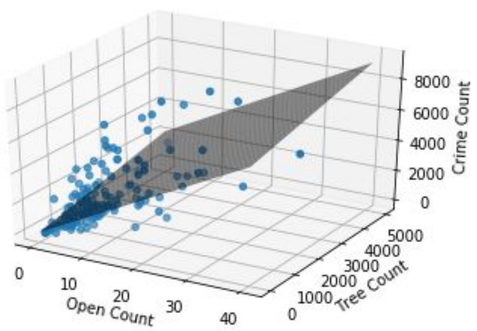
Data Sources

We collected data from five data sources. For Crime we used the Crime Incident Report from 2015 onwards, which contained useful information about the location and type of crime. Then we found sources for the locations of what we will call "Green Assets." These include open spaces, trees, charging stations, and Hubway stations. To convert our data into the feature vectors that we needed for clustering, we partitioned each data point into a cell of a 31x17 matrix, with each cell holding a count of the respective attributes belonging to it.

Methods 1 (Multiple Regression)

We used python to perform the analysis and MongoDB to store the data. After partitioning each datapoint into their respective cells, we decided to try to find which Green Assets best correlated to Crime Rate. Multiple regression is a technique used to estimate the value of a dependent variable from a set of independent variables. In order to estimate the value of the predictors for each value we used the Ordinary Least Squares(OLS) method.

After running this regression with Green Assets as the independent variables and crime as the dependent. We observed each variables confidence intervals to find which ones were significant, through this we were able to recognize that the confidence intervals for hubway stations and charging stations were far too wide. And consequently, we threw them out and ran the regression again with just the tree counts and open counts.



Conclusion 1

This model shows a positive correlation between both open space and trees and Crime. It had an R-squared value of .675, meaning that the model fit the data decently well, in general. But moving forward, we wanted to include more of the geographic aspect

Methods 2 (K-means clustering)

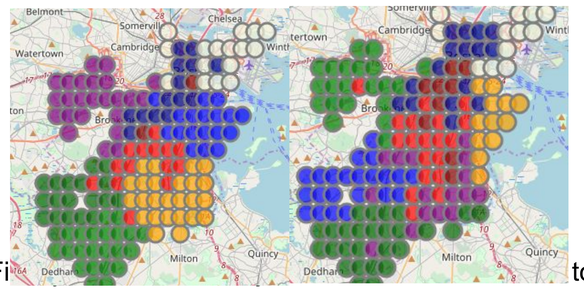
In order to incorporate the geographical aspect of our project, we figured the best way to do this would be to cluster our data points using k means clustering. And compare the clusters for each green asset to the cluster for crime count. To set up the clusters we first had to normalize all the count values and the lat and long parameters. Next we needed find a k value that would work reasonably well with each set of values. So we decided to use Silhouette score as the metric to decide k with. Silhouette score is a method for finding the "strength" of a cluster. The formula for a silhouette coefficient is as follows:

$$s = \frac{b - a}{\max(a, b)}$$

Where a is the the mean distance between a data point and all other points in the same cluster, and b is the mean distance between and the nearest cluster.

From this method, we were able to show that for all the data sets, a local maximum existed around k = 10. So we decided on 10 to be our K-value.

Using count, latitude, and longitude as our feature vector we performed k-means 5 times for each item. Shown below are the results for Open Counts and Crime Counts, respectively.



Fi to compare the similarity of each set of clusters and found that the open count cluster was most similar to the crime count cluster.

Conclusions and Future Research

Although our results point to the idea that open spaces cause crime, we believe that this can be attributed to the simple fact that there are more open spaces in bad areas rather than highly developed areas such as downtown Boston. Going forward, one could increase the granularity of the matrix beyond our 31x17 matrix to increase accuracy. Furthermore, one could perform more Regression testing on each of the clusters to see if the global model is similar to the more local model.