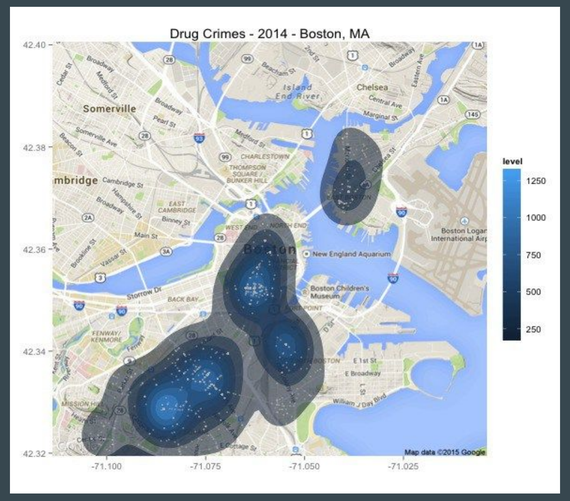
Statistical Correlation of Drug Crime and Youth in the City of Boston

Aditi Dass
Andrew Lee
Benjamin Li
Tony Yao
CS 591 L1
Boston University

Introduction

Drug crime among youth population is a major concern in any major urban populace. People find the possibility of such crimes occurring near children to be a disconcerting reality. However, when said metric is overlayed into city geography, it becomes rather difficult to extrapolate meaningful information.



Project Goals

Our projects goals can be split into two major tracks.

1. Find the distance at which the difference between the number of child establishments surrounding each drug crimes and the number around each standard crime is **maximised**
2. At this distance, find if a **relationship** exists between the number of such establishments and the number of drug crimes

Data

- **Crime Datasets**: Primary dataset was a consolidation of the legacy crimes dataset (July 2012 - Aug 2015) and the current crimes dataset (Aug 2015 - present) found on the data.cityofboston.gov portal.
- **Child Establishment Datasets**: Supplementary datasets contain the locations of establishments that children frequent including public and private schools, public and private daycares and child food services found on the city portal and through web scraping.

Method 1: Objective Function

MAX (| *Average number of child establishments within r distance of a standard crime* - *Average number of child establishments within r distance of a drug crime* |)

The following steps were implemented **for each radius value *r*** of range(0, 5) miles to a precision of 0.1 miles.

Step 1

For each crime and drug crime, the number of children establishments within *r* was found

Step 2

Database from previous step was map reduced to find the number of crimes and drug crimes with *x* child establishments respectively

Step 3

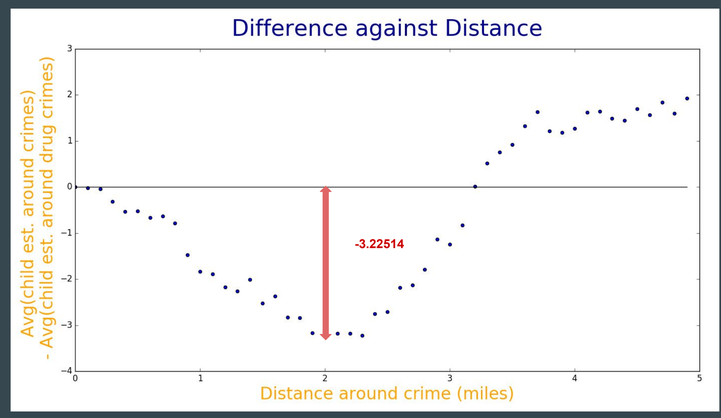
Database from step 2 was further map reduced to find the average number of child establishments around each crime and each drug crime

Step 4

Difference between average establishments around each crime and average establishments around each drug crime found

Distance	Abs Diff
1.8	2.845
1.9	3.169
2.0	**3.225**
2.1	3.178
2.2	3.184

Data from radius > 5 tended towards constant values as distance approaches boston boundaries



Method 2: Regression Analysis

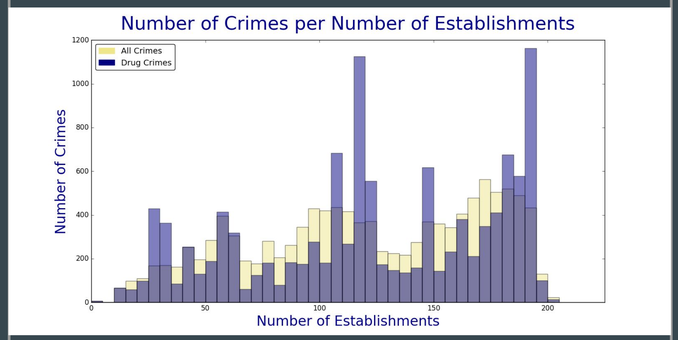
We ran two OLS analyses on our data with the optimized radius of 2.0 miles.

The first compared the total number of establishments with a specific number of all crimes within its radius. For example, a data point in the format: {value : {crimes: 576, total: 27}} means that there are 27 establishments with 576 crimes within its 2 mile radius.

The second analysis compared the total number of establishments with a specific number of drug crimes within its radius. We also created a histogram to visually represent the differences between the normalized number of crimes to number of establishments.

Method 2: Results

	Correlation Coefficient	R-squared Value
All Crimes	0.471	0.775
Drug Crimes	0.186	0.273



Conclusion.

We can see that there are two peaks in our histogram of Number of Crimes per Number of Establishments. These two peaks represent that there are about 1100 crimes with around 125 children establishments in its radius and there are about 1100 crimes with around 190 children establishments in its radius. We suspect that these peaks are a result of a hotspot of crimes within the optimized radius.

We expected to find more drug crimes around child establishments than all crimes. However from the regression analysis results, the higher R-squared value and resulting correlation coefficient of all crimes compared to only drug crimes indicates that our predictions were not in fact the case. We suspect this was due to unreported drug crimes around child establishments.

Future Work

For the Objective Function, the data points represent a definite pattern. Further work involves trying to find the best fit equation of this pattern. Explanations for why avg(establishments per crime) overtakes avg(establishments per drug crime) at radius 3.3 also seems like an intriguing avenue to explore.

Furthermore, we plan to build an interactive graph that responds to the different radius values in order to study any difference in distributions