# CS591 Data Mechanics – Final Project Shreya Ramesh December 9, 2016

#### Introduction

One of the most accurate signs of a neighborhood in the process of being gentrified is the trend in building permits: be it residential buildings having an unusual amount of renovation, or corporate buildings being established.<sup>1</sup> Another marker of gentrification is the crime rates in a neighborhood going down. Therefore, this project examines the building permits, crime reports, and earnings in Boston.

### **Datasets**

The datasets used from the City of Boston<sup>2</sup> were:

- Approved Building Permits (2006-2016)
- · Crime Incident Reports (July 2012-August 2015)
- Employee Earnings Report 2012
- Employee Earnings Report 2013
- Employee Earnings Report 2014
- Employee Earnings Report 2015

From these, new datasets were assembled of the means of each of the datasets per year. This is to see the progression of the building permits, crime, and wealth over the period of 4 years and attempt to find a correlation between the 3.45 means were used because according to the City of Boston website, that is how many zipcodes are approximately in the Greater Boston area. Below is an image of the Greater Boston areas delineated with zipcodes (on which points may be mapped in the visualization component).



The new dataset for Approved Building Permits only has building permits approved between July 2012 and August 2015 are in the new dataset, to be consistent with the Crime Incident Reports dataset.

The Crime Incident Reports were split up by year and found the k-means of crime incident reports by year. The new dataset maps years to a list of 45 means of locations of the crimes.

#### **Techniques**

As mentioned earlier, the primary calculation was calculating the k-means of each dataset. Below is the algorithm, where  $(x_1-x_n)$  is the set of observations and k is the number of means.

A check was implemented where the distance between each of the old means and new means are below a certain threshold, and originally 0.1 was the value for this project. However, this simply returned the same means each iteration, it was changed so that the entire means algorithm is run 10 times.

$$rg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

## Limitations

#### Datasets

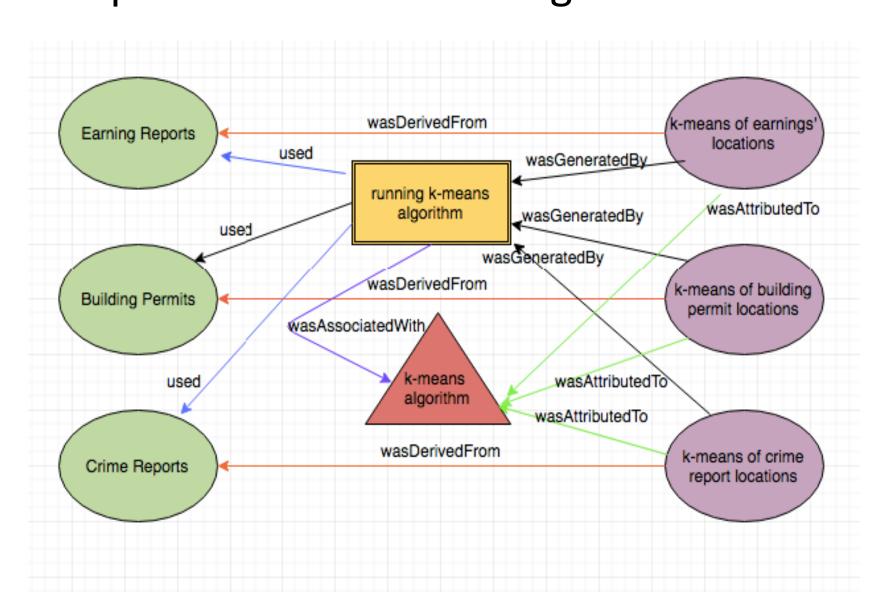
- Some of the json objects from the API did not have location coordinates, so the Python module geopy was used to return the latitude and longitude of each location based on zipcode.
- . Instead of giving each employee a value of 1, the weight acted as the salary. This way the k-means algorithm will determine where the wealth is most concentrated in Boston.
- In retrospect, grouping the data by zipcode may have made analysis and visualization easier.

## Analysis

- Since k-means are not ordered when calculated, so a new dataset was made of points with the least distance between them over the 4 years.
- Grouping the data by year might also be an over-simplification of the problem, and so might only evaluating 3 years.

#### **Data Provenance**

Below is a visualization of the organization of initial datasets, based on the provenance model taught in this class.



## **Statistical Analysis**

The two forms of analysis used were linear regression through the sklearn library and the calculation of correlation and covariance. The formulas for those are below.

$$\operatorname{cov}(X,Y) = rac{1}{n} \sum_{i=1}^{n} (x_i - E(X))(y_i - E(Y)).$$
 $ho_{X,Y} = \operatorname{corr}(X,Y) = rac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$ 

Below are the resulting p-valus. What these results show is little correlation or covariance between the datasets.

	Earnings	Building Permits	Crime Reports
2012	0.807	0.928	0.725
2013	0.0	1.0	0.0
2014	0.0345	0.4905	0.5345

## **Web Service and Visualization**

The results of this project can be viewed through a Flask application. Based on the user's input, the web service will display the results of the statistical analysis on the dataset(s) of building permits, earning reports, crime reports, or all three. Below is a screenshot of the web service component in progress. The visualization will include a graph of the linear regression and maybe a map to better understand the data.

# **Results and Conclusion**

This project could have revealed interesting information about areas of Boston through the statistical analysis, but the results produced did not show any meaningful trends or information. Perhaps the idea of taking the k-means made this problem too simplistic. Future work may include attempting the same analysis but constraining the means by zipcode in order to analyze and visualize results more easily.

#### References

- 1. Atkinson, Rowland. "Measuring Gentrification and Displacement in Greater London." *Urban Studies* 37.1 (2000): 149-65. Web. 29 Sept. 2016.
- 2. "City of Boston | Open Data." *City of Boston*. N.p., n.d. Web. 08 Dec. 2016.