CS591 Data Mechanics – Final Report
Shreya Ramesh
15 December 2016

Introduction

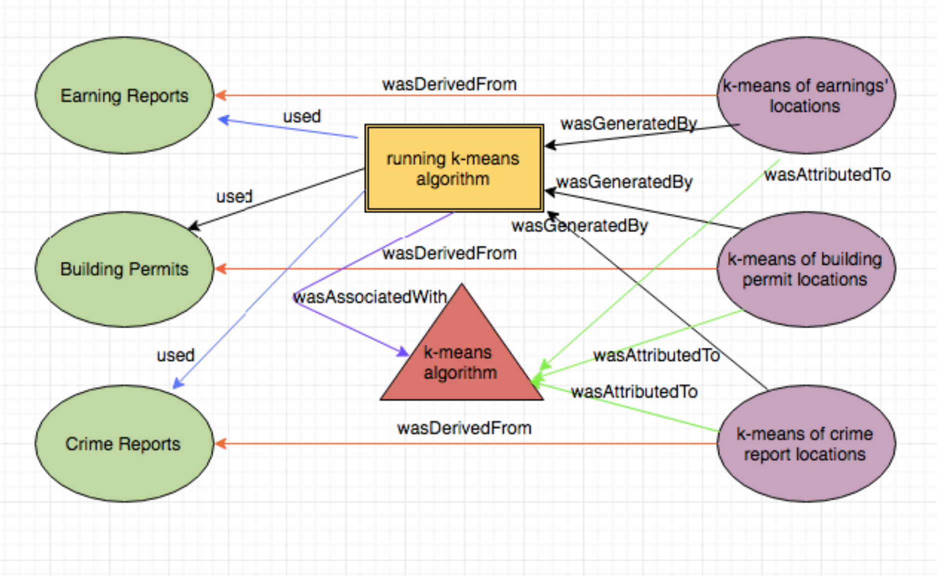
One of the most accurate signs of a neighborhood in the process of being gentrified is the trend in building permits: be it residential buildings having an unusual amount of renovation, or corporate buildings being established.[1] Another marker of gentrification is the crime rates in a neighborhood going down. Therefore, this project examines the building permits, crime reports, and earnings in Boston. The initial goal of this project was to find the particular relationships between building permits, earnings, and crime reports that highlighted whether or not an area was being gentrified.

Datasets

The datasets used from the City of Boston 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. All of the new datasets are of the form: (year, list of 45 means found through the k-means algorithm). 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 calculated because according to the City of Boston website, that is how many zipcodes are approximately in the Greater Boston area. From the analysis of these datasets, the goal was to derive clear information, which showed statistical evidence of gentrification. Below is a diagram of the data provenance, as taught in class:



Algorithms and Techniques

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. Some of the json objects from the API did not have location coordinates, so the Python module geopy (https://pypi.python.org/pypi/geopy) was used to return the latitude and longitude of each location based on zipcode. Then, the k-means of approved building permits was calculated by year. The new dataset maps each year to a list of 45 means.

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

For each Employee Earnings Report, the k-means were calculated by year. Instead of giving each employee a value of 1, the weight was the salary. This way the k-means algorithm will determine where the wealth is most concentrated in Boston. For these datasets as well, geopy module was used to change the zipcode to latitude and longitude coordinates. Since the City of Boston employees aren not an accurate representation of all wealth in Boston the average of each zipcode was calculated by adding up the wealth in each zipcode and dividing it by the number of people in that zipcode to get a more accurate measurement of income in the area.

In retrospect, the project would overall make more sense if the data had been grouped by zipcode in order to see more distinct changes, and to answer the original question of whether gentrification was occurring in an “area.” Another problem with using the k-means algorithm is that it is difficult to visualize the change easily over the third dimension of which year. Grouping the data by year might also be an over-simplification of the problem, and so might only evaluating three years, as that might not be sufficient to see a concrete trend.

The k-means algorithm used is very similar to the one taught in class, but instead of taking in 2 sets of points (M, P), M is defined as 45 points in the 02215 zipcode and the input is a dictionary with the keys being points and the values being the weight. This way in the part of the algorithm, which puts 1 as the weight for each point, based on the point the salary can be used as a weight. For the other two usages of k-means, same code and input dictionaries were used, where every key has a value of 1. A check was implemented where the distance between each of the old means and new means are below a certain threshold, and originally 0.01 was chosen for this project. However, this simply returned the same means each iteration, so for the second component of the project, the algorithm was changed so that the entire means algorithm is run ten times.

Since k-means are not ordered when calculated, for the statistical analysis, a new dataset of points was created with the least distance between them over the 4 years.

Linear regression was chosen as an analysis tool because not only does it provide a basic visualization, but it provides the capability to fit future points around the regression line. The dataset has a key of the data source and the value is a tuple of the coefficients, mean-squared error, and variance score where 1 indicates "perfect prediction".

The p-score and correlation calculations will provide more detailed information on the relationship between the trends and indicate whether or not this shows an overall progression in certain directions. Below are the results obtained:

	Earnings	Building Permits	Crime Reports
2012-2013	0.807	0.928	0.725
2013-2014	0.0	1.0	0.0
2014-2015	0.0345	0.4905	0.5345

Limitations

- 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. This may have concentrated points in certain areas, as values that did not have a latitude and longitude would all receive the same default value for a given zipcode. This likely would have skewed the data.
- 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. However, this may not have sufficed in accurately showing the spread of wealth.

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. Future work will possibly include plotting the results on a map of Boston for better understanding.

Results and Conclusions

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. As mentioned in the feedback for Project 2, a better way of measuring the movement between years might be to measure the angle between each pair of years. This is another possibility for future work as well.

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.