CRIME CLUSTERING, ITS SHIFTING PATTER AND GENERAL HOUSING EVALUATION

Li Liu, Sibo Zhu | Professor Andrei Lapets
Boston University

OVERVIEW

Security is becoming an increasingly important concern of people nowadays, and among all the datasets available in the CityOfBoston, the crime datasets are ones of the most complete and detailed. In this project, we started on observing the crime clusters using k-means algorithm, then proceeded to investigate two possible causes that lead to the crime clustering, and finally we implemented a rather "naïve" algorithm to evaluate most of the residential buildings in Boston Area.

Due to the computing power and data sets available to us, the objectives are achieved at different level of success, but what we hope to show with this project are some general patterns, and bring more insights to future studies.

PROGRAM AND DATA SETS

Programming language and services used

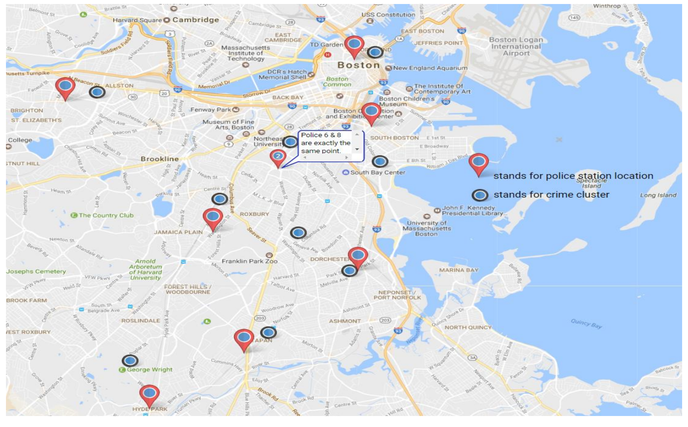
- Python
- MongoDB (Community Edition)

Data sets used

- Crime Incidents Reports (2012 – 2016)
- Entertainment Licenses
- Police Stations
- Liquor Licenses
- Property Assessment (2014 – 2016)

INTRO: CRIME CLUSTER AND POLICE STATION

As an introduction to this project, we will show a view of general crime clusters from 2012 to 2016, calculated with k-means algorithm (k = 9), and for each crime cluster, the closest police station will also be shown. Also, to eliminate some minor or irrelevant crime entries, we filtered the data with a list of interested crime types.



Note from the calculation, within 2 kilometers of each crime cluster there is one police station. Interestedly, such phenomenon can be explained in two ways: The police stations are not very effective since the clusters of crime are near them, or in another way the locations of police stations are correctly chosen because police can get the crime locations as soon as possible.

CLUSTER SHIFTING PATTERN

- In this section, we will discuss the underneath possible pattern and cause that effected the crime clustering. Moreover, the crime clusters now are calculated based on years (2014, 2015 and 2016)

First we consider the shift of clusters of low-value residential properties may effect the shift of the crime clusters.

Low-value residential properties are defined as: from the datasets of property assessments, each building has two attributes: "av_bldg" and "living_area", which mean building value and area of living respectively. Thus we can calculate the cost per area with "av_bldg"/"living_area". The residential property has cost per area that is lower than average cost minus 1.1 standard deviation is classified as low-value residential properties.

The procedure starts with calculating low-value properties clusters and crime clusters of both 2014 and 2015. (K-means with k = 8)

Then for each crime cluster ($crime_n$) in 2014, we group it up with:

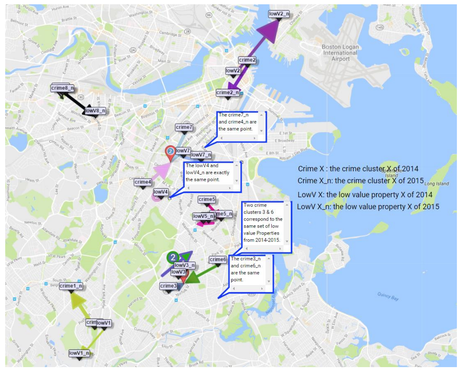
(1) One low-value property cluster in 2014 that has the smallest distance to the crime cluster ($lowV_n$)

  (1) And for this particular low-value property cluster in 2014, we pair it up with one low-value property cluster in 2015 that has least distance to it. ($lowV_{n\_}n$)

(2) One crime cluster in 2015 that has the least distance to it. ($crime_{n\_}n$)

Thus we may consider that from 2014 to 2015, the crime cluster $crime_n$ has moved to $crime_{n\_}n$ with certain vector ***u*** and its corresponding low-value property cluster $lowV_n$ has moved to a $lowV_{n\_}n$ with certain vector ***v***.

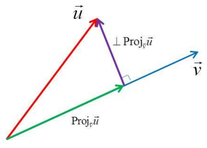
There are total 8 crime clusters, and it will produce 8 pairs of vector ***u*** and ***v***



Then for each pair of vector ***u*** and ***v***, we will project the vector u onto v, produces vector $\boldsymbol{Proj_u}$, this new vector indicates how much crime cluster has shifted in the direction of shifting of the low value property clusters from 2014 to 2015.

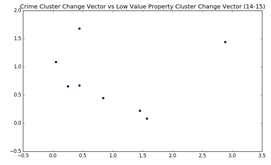
The magnitude of ***v*** and $\boldsymbol{Proj_u}$ are calculated, and stored in a tuple, and again there will be 8 tuples to be produced.

The final steps are to calculate the correlation coefficient of these 8 to tuples, and each tuple is regarded as one single point in a 2d axis. Note that if vector ***u*** resemble vector ***v*** in a great degree, then the point $(|\boldsymbol{v}|, |\boldsymbol{Proj_u}|)$ should be very close to the line y = x. Thus, the correlation coefficient can give us a general information about how related the shifting of crime clusters and the shifting of low-value properties clusters are.



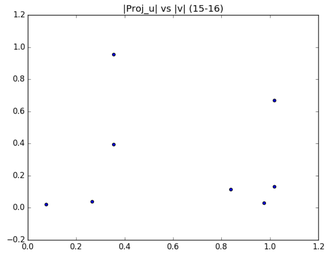
The correlation coefficient we got is 0.000503179425801

Such small coefficient indicates that there is no correlation between the shifting of crime clusters and the shifting of low-value properties clusters.



And the graph on the left shows the results we got from 2015 to 2016 using the same procedure as explained above.

The correlation coefficient is: -0.0119926242743, which is still no where close to 1.



One final note on this section is that we also tried to use information of liquor stores clusters to estimate any shifting of crime clusters. However after we got result and look at the datasets more closely, we found that the liquor stores opening time are really skewed (almost two thirds of the liquor stores issued their licenses within three months of 2013). Thus we decide to not using the result.

GENERAL HOUSING EVALUATION

For this section, we will briefly discuss a general housing evaluation

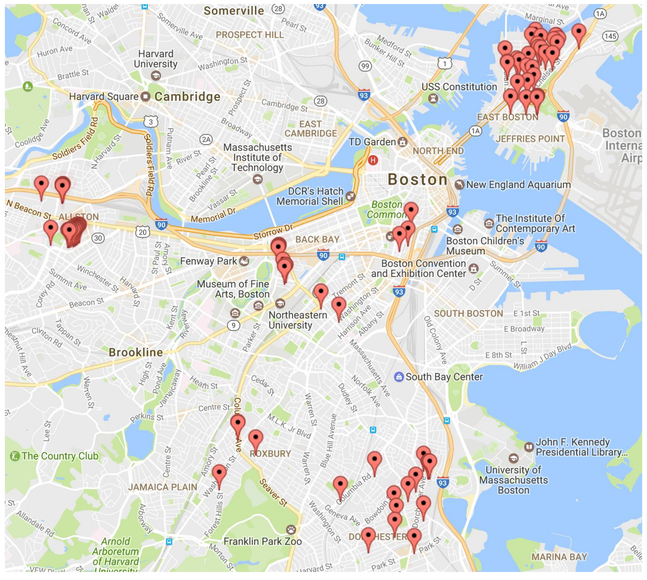
Due to the available and relevant datasets available on the city of Boston datasets, in this scoring system, only three variables will be considered: the distance to the closest crime cluster, the distance to the closest entertainment, and its cost per area.

For people who are buying properties, we provide three very straightforward incentives: The property needs to be as safe as possible, it's a plus if the entertainment center is close to the property, and finally if the price should be as low as possible.

In short, the scoring system can be summarized with following objective function:

Z = (25 - distance to closest entertainment cluster) * 10 - (3 + distance to closest crime cluster) * 10 - (cost per area * 0.5)

Then for each of the residential building retrieved from the 2016 property assessment , we applied the function to it and obtained a list of scores and their corresponding locations. The graph below shows the buildings that have scores higher than average score + 1.85 standard deviations. And it is quiet clear there are 5 clusters of high score buildings spread across the city.



FUTURE WORKS

Regarding the future works of this project, here are the things we think could improve:

- The K-Means algorithm is the core of this project, yet it always takes substantial time to get a result with a size of data around 1500, and with size of 2000 would sometimes cause memory error or just took to long. And this also means sampling of data is required. However, a sample size of 1500 just seem not very sufficient for some large datasets like Crime Incidents. Thus, for future works a better computational devices and power are recommended
- One major component of this project is to analyze how the changing of crime clusters is effected by the changing of clusters of other things such as low-value property. However, the results we get in this project are not very meaningful. Thus some new analyzing techniques could be introduced such as using spatial correlation.
- Crime is a complicated topic, and it may involve many other factors. So it is also useful to consider a new model that can integrate other possible factors.