CS591 L1 Final Report

Bowen Yang, Jiadong Chen, Xiao Lu,

Professor: Andrei Lapets

Date: 12/13/16

Objective

Massachusetts has the 18th highest crime rate among all the states in US and the crime rate in Boston is 45% higher than the average crime rate in Massachusetts. As residents of Boston, our team decides to find out what is related to crimes geographically.

We would examine and analyze the the locations of crimes in Boston and other geographical information (i.e. the location of liquor stores, high-valued properties and the parks) using mechanisms including k-means and closest distance function we implemented ourselves. We aim to provide some insights for Bostonians to know which places may be unsafe, or even to make Boston safer, via this project.

Resources Used

1. Crime Incident Reports (July 2012 - August 2015): We extracted locations of crimes around Boston from this dataset. (https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports-July-2012-August-2015-Sourc/7cdf-6fgx)

2. Property Assessment 2016: We got the top 10% highest-valued property form this dataset. (https://data.cityofboston.gov/Permitting/Property-Assessment-2016/i7w8-ure5)

3. Liquor Licenses: We acquired the locations of stores that sell alcohol. (https://data.cityofboston.gov/dataset/Liquor-Licenses/hda6-fnsh)

4. School Gardens: We extracted the locations of the locations of school gardens from this dataset. (https://data.cityofboston.gov/Facilities/School-Gardens/cxb7-aa9j)

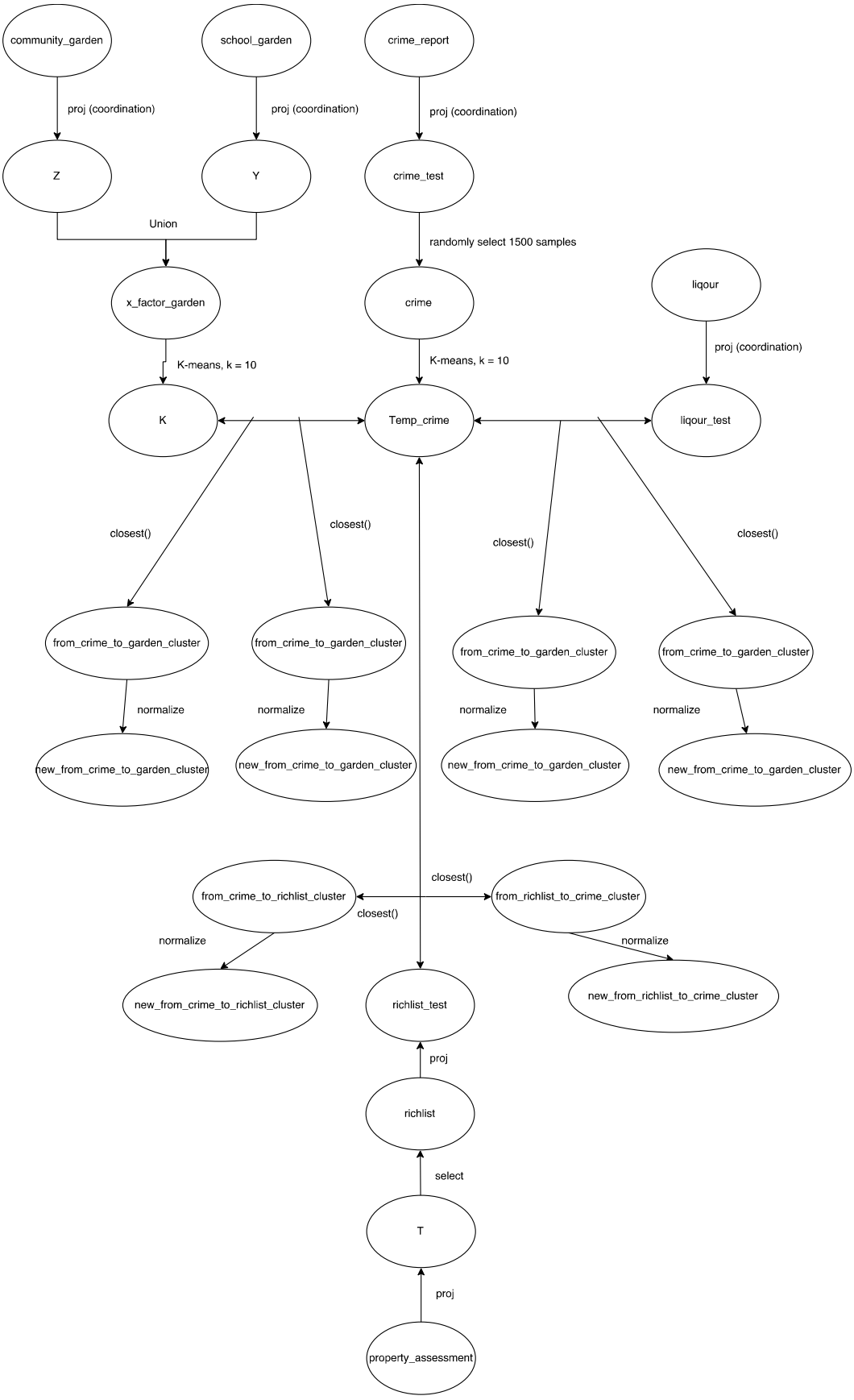
5. Community Gardens: We extracted the locations of the locations of school gardens from this dataset. (https://data.cityofboston.gov/Health/Community-Gardens/cr3i-jj7v)

Basic Algorithm

Firstly, we would utilize *k-means* to make the widely spread locations of crimes, high-valued properties, liquor stores and gardens into 10 clusters for each dataset respectively, then we start analyzing the inner connection between crime rate and the factors listed above. In order to do so, we need to calculate and compare the mean and standard deviation of the closest distance from the crime clusters to the clusters in other three factors. For example, between locations of crime and liquor store, we would find the distance from each crime location cluster to its closest liquor store cluster. Then we do it reversely, finding the distance from each liquor store cluster to its closest crime location cluster. If the mean of the two lists are both low as well as the standard deviation, we can determine that there is a correlation between liquor stores and crime locations.

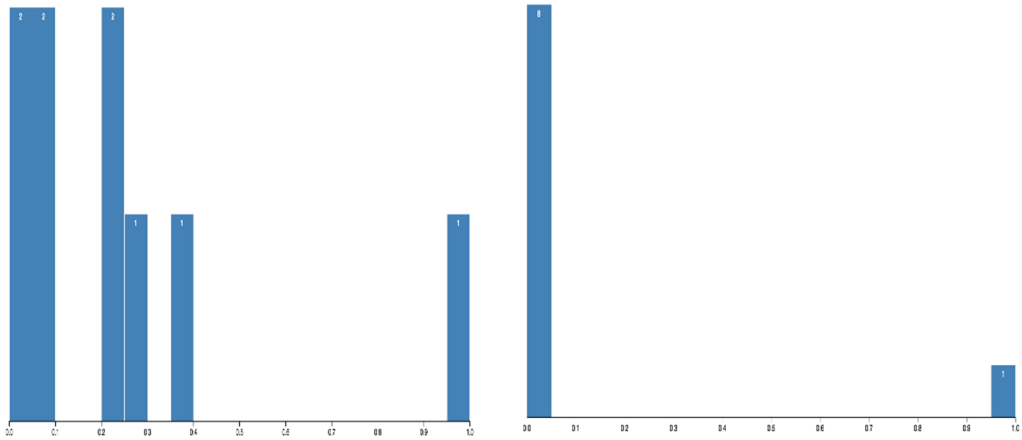
Data Flow

Basically, for all original datasets we used projection to pull out the coordination (and the value of properties in the *richlist* case) that are needed for the next steps. Then we implemented k-means algorithm to make ten clusters for each dataset: K for School Gardens and Community Gardens; *Temp_crime* for Crime Incident Reports, *liquor_test* for Liquor Licenses and *richlist_test* for Property Assessment. After that, we called the function *closest(list1, list2),* a function that computes and finds the shortest distance from every cluster in *list1* to clusters in *list2*, and then returns the mean of the shortest distances.

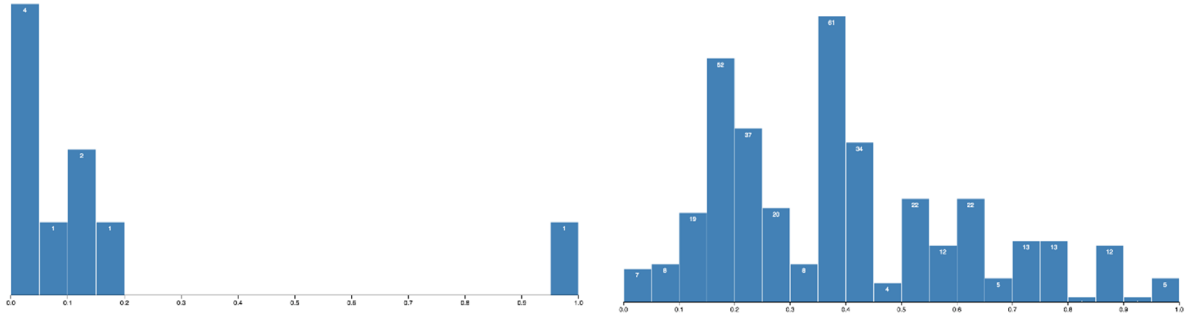

community_garden
school_garden
crime_report
proj (coordination)
proj (coordination)
proj (coordination)
Z
Y
crime_test
Union
randomly select 1500 samples
x_factor_garden
crime
liqour
K-means, k = 10
K-means, k = 10
proj (coordination)
K
Temp_crime
liqour_test
closest()
closest()
closest()
closest()
from_crime_to_garden_cluster
from_crime_to_garden_cluster
from_crime_to_garden_cluster
from_crime_to_garden_cluster
normalize
normalize
normalize
normalize
new_from_crime_to_garden_cluster
new_from_crime_to_garden_cluster
new_from_crime_to_garden_cluster
new_from_crime_to_garden_cluster
closest()
from_crime_to_richlist_cluster
from_richlist_to_crime_cluster
closest()
normalize
normalize
new_from_crime_to_richlist_cluster
richlist_test
new_from_richlist_to_crime_cluster
proj
richlist
select
T
proj
property_assessment

This is how we got *from_crime_to_garden_cluster, from_garden_to_crime_cluster…* Lastly, we normalize each list and generated *new_from_crime_to_garden_cluster, new_from_garden_to_crime_cluster…* facilitating the comparison and visualization.

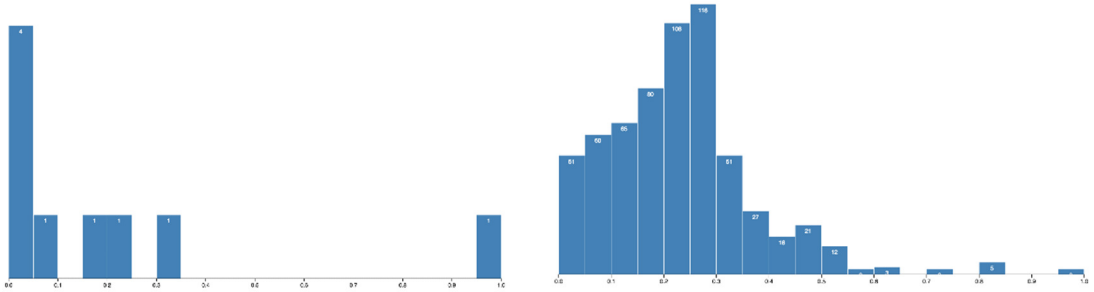
Data Analysis:



Above are the histogram of the normalized shortest distance between garden clusters and crime locations. Based on the graph we believe that there is an outlier in crime location which is so far away from its closest garden. Then we find this outlier and delete it from the points we have, yet the remaining data's STD and average are still high, which means the distance varies a lot and is generally long. Therefore, we believe that there is no apparent correlation between garden location and crime rate.

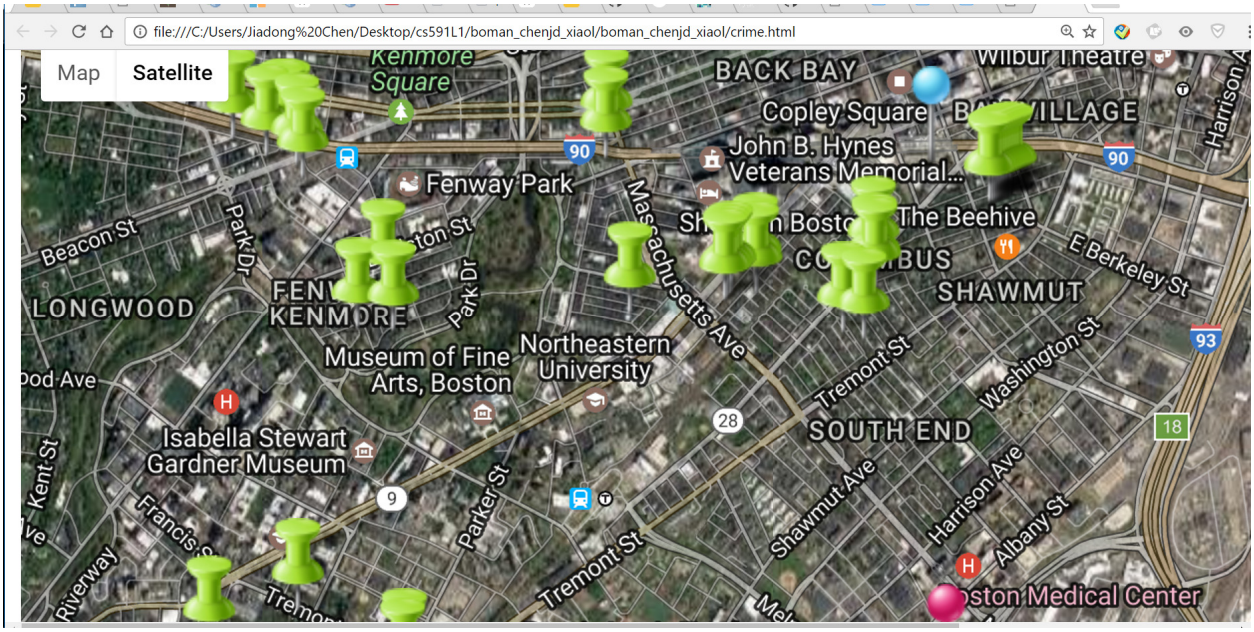


From the above graphs, we can see that the distribution are more normalized and general in all x range, also from the graph not normalized are we able to detect that the average distance is short as well, so we are able to deduce that the distance between each crime incidence location is generally normally distributed around high-valued properties, vice versa.



Following the same steps as above, we found that the average distance is short and the std is small as well, so we believe that there is some connection between liquor store and crime rate.

Visualization:



This is part of the screen shot from crime.html file, which is implemented by google map api. I use blue label to represent main crime locations, pink to garden, green to high-valued properties, then render this interactive graph to the users so that they can be more likely to feel the connection between the factors. Also, it provides users with a concept that which part in Boston is most dangerous, or which part are those high-vlaued properties located.

Conclusion

Since the mean and standard deviation of the shortest distance between the crime spots and liquor stores as well as the distance between crime spots and high-valued properties are relatively low, while the mean and standard deviation of the shortest distance between crime locations and gardens are high, we can conclude that crimes happen more around high-valued properties and liquor stores, and gardens are not necessary related to crimes.

Future Inspiration

1 With this algorithm we can compare more geolocation information with the crime location to find out what is related to crimes in terms of location (geojson data type). Factors such as the income level may influence crime as well.

2 In project3, we tried to implement a website service for finding out the relationship between any geojson datasets and crime. (by uploading a geojson data file into the website and run our algorithm in backend then return result) Yet we failed to do so

because of limited front end Javascript skills, so for now we only have some interactive html file to visualize the result we have so far.