CS591 Project 3 report

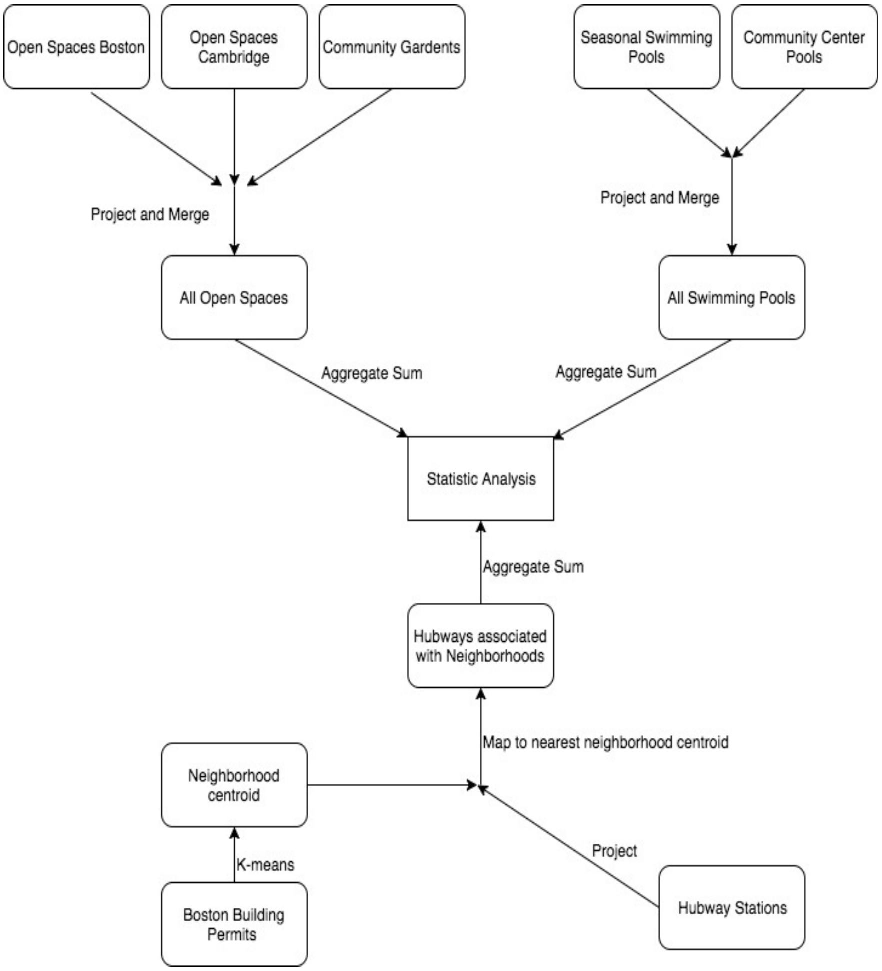
Prof. Lapets

Recreational Places Analysis of Boston's neighborhood

By Wei Ji, Yue Zhou, Yizhi Huang

Introduction:

We are trying to find out how sportive or recreational neighborhoods in Boston are and to see how quantities of different public recreational places or facilities (open spaces, swimming pools and bicycle stations) are correlated. By using these new information, people could have one more perspective to view a neighborhood and decide whether they want to live in such neighborhood. Moreover, government officers could also reflect from this information to see whether a neighborhood need a improvement on the provision of recreational places. Researchers could also incorporate the data we got to study more about Boston's neighborhoods.



Online Data Sets:

1. 'seasonalSwimPools':'https://data.cityofboston.gov/resource/xw3e-c7pz.json'
2. 'communityGardens':'https://data.cityofboston.gov/resource/rdqf-ter7.json'
3. 'openSpaceCambridge':'https://data.cambridgema.gov/api/views/5ctr-ccas/rows.json?accessType=DOWNLOAD'
4. 'waterplayCambridge':'https://data.cambridgema.gov/api/views/hv2t-vv6d/rows.json?accessType=DOWNLOAD'
5. 'openSpaceBoston':'http://bostonopendata-boston.opendata.arcgis.com/datasets/2868d370c55d4d458d4ae2224ef8cddd_7.geojson'
6. 'commCenterPools':’'http://bostonopendata-boston.opendata.arcgis.com/datasets/5575f763dbb64effa36acd67085ef3a8_0.geojson'
7. 'hubwayStations':'http://bostonopendata-boston.opendata.arcgis.com/datasets/ee7474e2a0aa45cbbdfe0b747a5eb032_0.geojson'
8. 'buildingPermits':'https://data.cityofboston.gov/Permitting/Approved-Building-Permits/msk6-43c6/data'

Transformations:

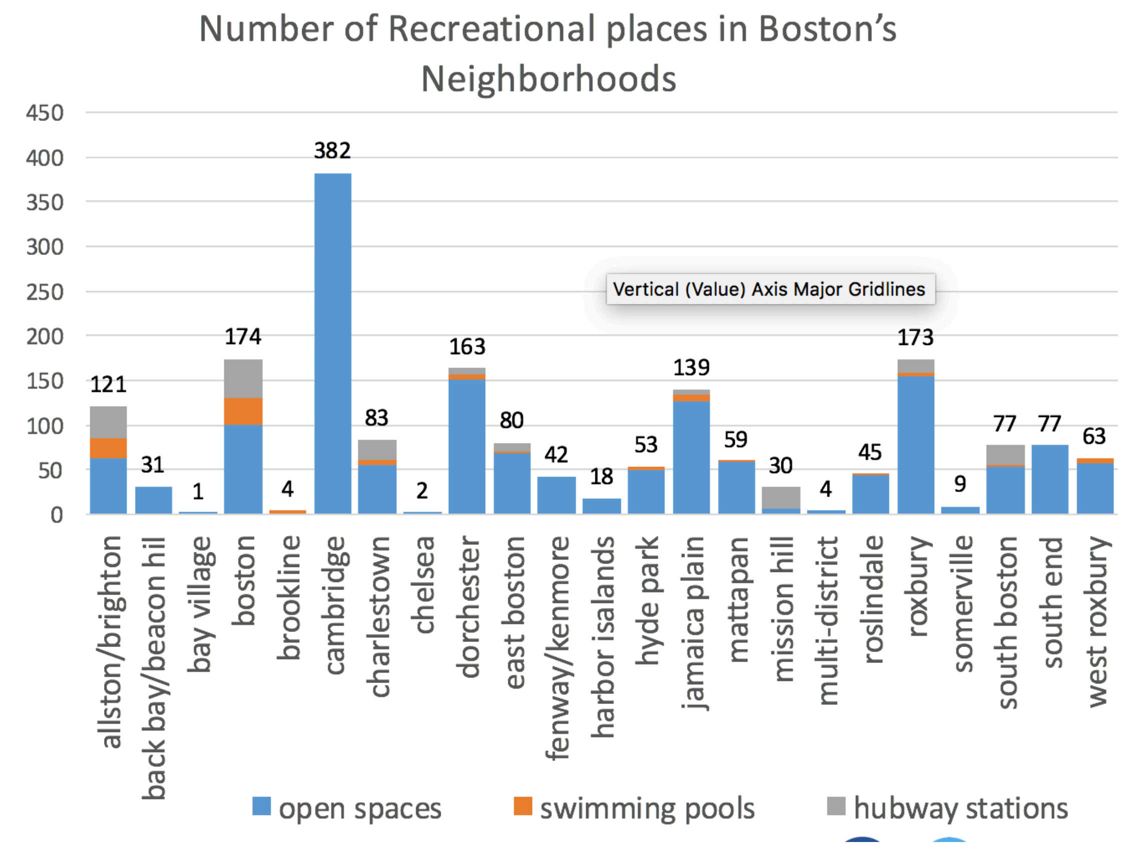
1. Retrieve all data from the 8 public datasets and inserted them into 8 corresponding tables in mongo DB.
2. Do projection to each of the two datasets-'seasonalSwimPools' and 'commCenterPools', so that we get only the pool name and the corresponding neighborhood. And combine the two outcomes into one dataset that contains info of all pools in Boston.
3. Do projection to each of the three datasets-'openSpaceCambridge', 'openSpaceBoston' and 'communityGardens', so that we get only the place name and the corresponding neighborhood. And combine the three outcomes into one dataset that contains info of all open places in Boston.

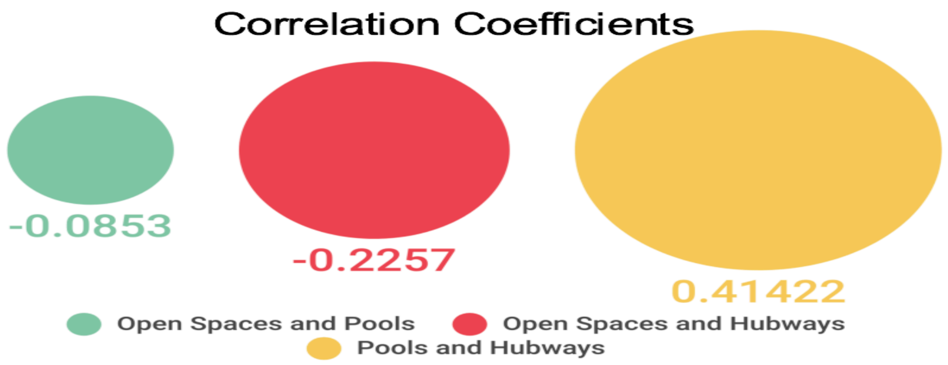
Resulting Data Sets:

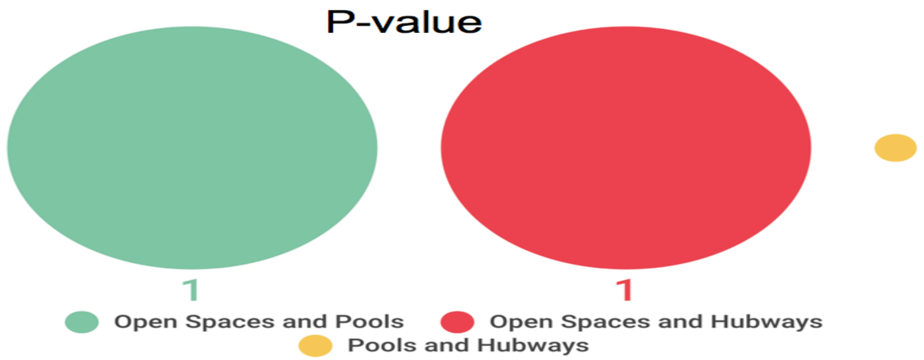
1. 'allPoolsInBoston' - output of combineAllSwimmingPools.py
2. 'allOpenSpacesInBoston' - output of combineAllOpenSpaces.py

Algorithms, Analysis Technique and Tools:

We could not find data about centroid coordinates of neighborhoods in Boston, therefore we calculated an estimation by ourselves. We retrieve new data about all building permits in Boston, and then we use **K-means** algorithm to estimate centroid coordinate for each neighborhood. When we get all neighborhoods' centroids, we incorporate new recreational facility data about hubway (bicycle) stations in Boston. Since the hubway station data include coordinates, we map each hubway station to its closest neighborhood centroid. Then we use map-reduce to count the number of each recreational facility in each neighborhood. At the end, calculate correlation coefficient and p-value b/w number of open spaces and swimming pools, b/w number of open spaces and hubway stations and b/w number of swimming pools and hubway stations.







Conclusion:

We sum up the number of hubway (bike) stations, open spaces and swimming pools for each neighborhood. After that, we use the stats collected above to calculate **correlation coefficient and p-value** b/w number of open spaces and swimming pools, b/w number of open spaces and hubway stations and b/w number of swimming pools and hubway stations. As you could see from the bubble graphs, the results we got are quite different from what we expect. The relationships b/w number of open spaces and swimming pools or number of hubway stations are not very correlated because their p-values are 100%, which means their C.C. are insignificant. On the other hand, the number of pools is relatively more correlated with the number of hubway stations and their p-value is low, which means the correlation is pretty significant.

Future Work:

Since we could not find boundaries data of Boston’s neighborhoods, we calculated the centroid of each neighborhood using a dataset about building permits in Boston, which means that the centroid is not exactly the center of a neighborhood. In the future, we would try to use a dataset about boundaries of Boston’s neighborhoods instead. Moreover, we would associate recreational places with zip codes instead of neighborhood, which is more specific and precise.