Optimal Placement of New Parks in Boston Neighborhoods

Gaspard Etienne, Tanin Tyler Lux, Leo McGann
CS 504 - Lapets

1 Introduction

This project focuses on the optimal placement of new parks throughout select neighborhoods in Boston and the effects open spaces have on a community's overall health. The motivating idea behind this project was if the city was tasked for building a new public park, how would one go about determining the best location to put this new park. Thus, to answer this question, we focused on two factors, one being an area's current access to an open park, which was determined by the distance to the nearest open space, and the other being the area's health. Clearly, favoring areas that are farther away from open spaces would be a desirable location to add a new park. On the other hand, if health and access to open spaces were correlated, then this would provide some justification for favoring unhealthier areas for possible new locations to put a park.

2 Data

2.1 Datasets

In this analysis we retrieved six different datasets from 3 different data sources. The datasets and their descriptions are listed below:

1. Open Spaces
Description: Open spaces of conservation and recreation interest in Boston, Massachusetts, USA, regardless of ownership. GeoJSON of open spaces is provided
Source: Boston Maps Open Data
2. Boston Neighborhoods
Description: The Neighborhood boundaries data layer is a combination of zoning neighborhood boundaries, zip code boundaries and 2010 Census tract boundaries. GeoJSON of each the neighborhoods are provided.
Source: Boston Maps Open Data
3. Parcel 2018
Description: City of Boston 2018 parcels created by the Assessing Department. GeoJson of each individual parcel is provided.
Source: Boston Maps Open Data
4. Property Assessment FY2018

Description: Gives property, or parcel, ownership together with value information, which ensures fair assessment of Boston taxable and non-taxable property of all types and classifications.
Source: Boston Maps Open Data

5. Health Survey Data in Boston
Description: This is the complete dataset for the 500 Cities project 2018 release. This dataset includes 2016, 2015 census tract estimates for 27 measures of chronic disease related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (9).
Source: Center for Disease Control and Prevention

6. Census Tracts GeoJSON
Description: GeoJSON of the census tracts that make up Boston.
Source: Data Mechanics Portal

2.2 Data Collection

The collection of all our data is done in the script getData.py. The accompanying data from each of the six datasets came in JSON form, and so were easily stored locally in MongoDB. Of the six datasets, all but parcel assessments were able be collected with a simple get request to its API. For parcel assessments, to collect the over 170,000 data values, we needed to make repeated queries to the api as the get request returned a maximum of 32,000 data values per request. Thus in our script there is logic implemented to make repeated requests also making sure that we skip over the data values we have already collected.

2.3 Data Transformations

To perform the necessary optimization and statistics analysis, we needed to transform the data that we collected. These transformations are all done in the combineData.py script. Our first transformations is an aggregation of parcel shapes with their assessment values. This is done simply by aggregating by parcel id (PID) which is unique to each parcel. Next, we combined census tract shapes collected from the Data Mechanics Portal with the health statics that are provided by the CDC pertaining to each census tract. Again, this is an aggregation where the key is now census tract id. Following this aggregation we combine parcels with their respective census tract, providing a way to get health statistics for an individual parcel. More specifically, using Shapely and the geoJSON coordinates of each parcel shape and census tract shape, we assign the health statistics of each census tract to each of the parcels that are fully contained

within that census tract. For the health statistics that we used, we project only the percentage of people with obesity, the percentage of people with low physical activity, and the percentage of people with asthma into each parcel using an r-tree index for each census tract to efficiently put parcels into census tracts. We also took an average of these three health statistics to create a new metric which we call a parcel's health score.

After this, using an r-tree index of for each open space, we find the distance of each parcel to its closest space. To do this, take the five nearest open spaces to any parcels, and find the closest one according to the euclidean norm in 2 dimensions with the points being the coordinates. Then we compute the minimum kilometer distance using the Haversine distance from the center of the parcel to any one of the points that make up the open space, and return this value. To get the distance score, we simply take the a parcels ten closest neighboring parcels (including itself), again using r-trees, and return the sum of these ten minimum distances to an open space. The thought process here is that some parcels might be far away from an open space, but if there are no parcels around it then it might not be the best place to put a new park. Instead, we want to favor "areas" that are relatively far away from open spaces, so summing does the trick as if there are a lot of parcels around with relatively larger minimum distances than this would be a good place to add a new park. Now with all the data aggregated we store this data into MongoDB to be used in our optimization and statistics.

3 Methodology

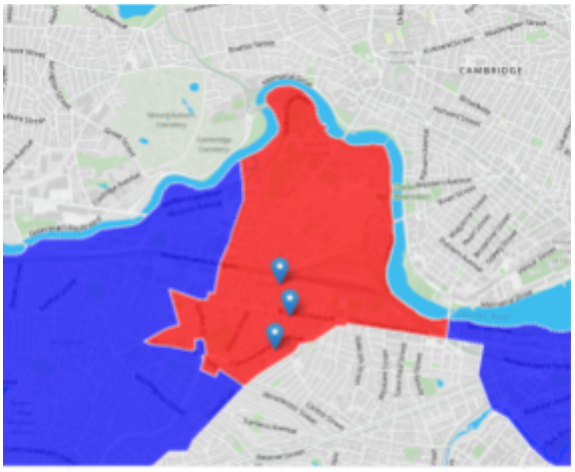
3.1 Optimization

The main purpose of this project is to try to find the optimal place of new parks in each neighborhood in Boston. Before we proceed with the methodology to find these optimal locations, we first want to note that we decided to leave out a few neighborhoods from our analysis. These neighborhoods mostly were neighborhoods defined in or near downtown Boston, where there is little room for new parks. As for determining the optimal placement for the parks that we did choose to analyze, we decided to go with a k-means approach. However, to make sure that k-means favors areas that have less access to open spaces or unhealthier areas, we implement a weighted k-means. To implement a weighted k-means, we computed a weight for each parcel within that neighborhood according to the metric we are interested in, and added that

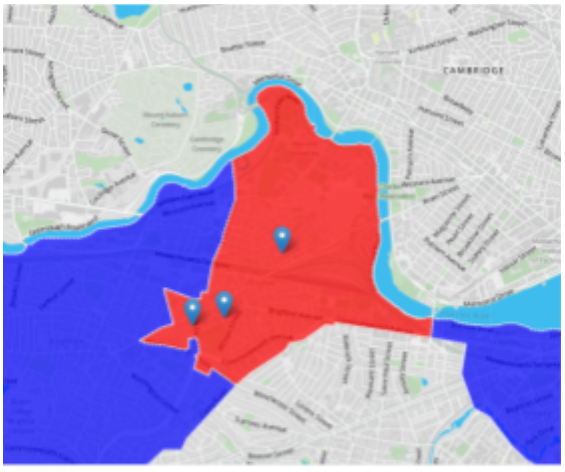
many points to the collection of points that we would run the k-means algorithm on. As mentioned before, we had two metrics, distance scores and metric scores. In the first few test runs of the k-means we found that using these metrics alone did not favor certain areas enough as the integer values between each metrics did not differ by very much. Thus, we standardized each metric for each parcel, or in other words computed

$$distanceStandardized = (distanceScore - \mu_{distanceScore}) \div \sigma_{distanceScore}$$

This was done also for health scores. Here, the means and standard deviations of each metric were neighborhood specific, that is for example a parcel in a neighborhood like Beacon Hill was standardized against the mean and standard deviation of the particular metric in Beacon Hill. Now standardized, we assigned weights accordingly. If the standardized variable lie 1.5 standard deviations away from the mean, we gave it a weight of a 100. If the standardized variable lie 1 standard deviations away from the mean we would give the parcel a weight of 10. Otherwise, every other point would get weighted 1. Favoring higher distance scores and health scores are a direct result of what each metric means. For distance scores, areas farther away from open spaces have larger distance scores, so it makes sense to more heavily weight larger distance scores. Similarly, higher health scores means an unhealthier area, so we would want to weight an unhealthier area more heavily. Thus, we ran two iterations of k-means, one based on distance scores and the other on health scores, for each neighborhood and stored these results. Below is an is an example of the different k-means plots in the neighborhood Allston.



Distance Score K-Means in Allston



Health Score K-Means in Allston

3.2 Statistics

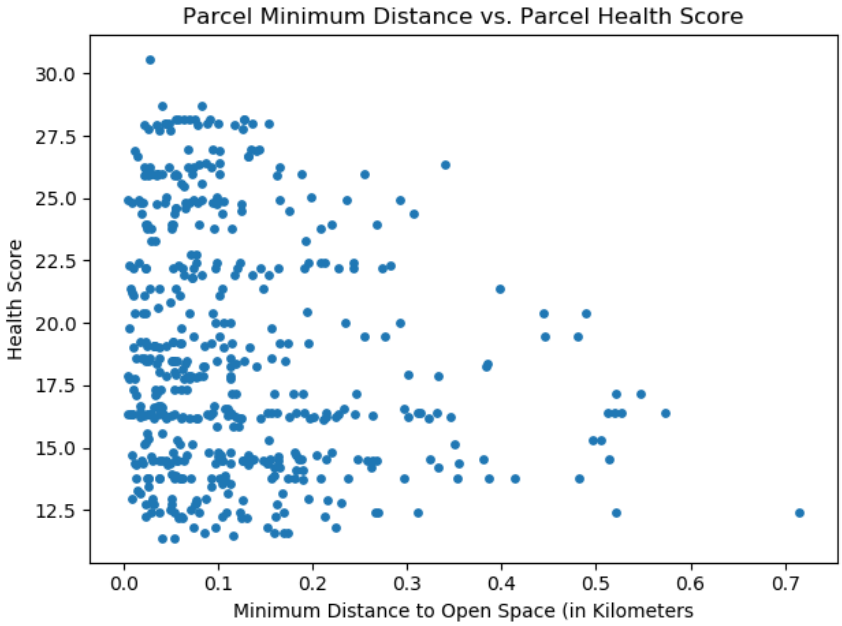
For the statistics portion of this project, we wanted to examine the relationship between an area's access to open space and its health. The reason we chose obesity, low physical activity, and asthma as the three health statistics to focus on was that we believed that these three statistics would most likely be related to the amount of green space an area had. Thus, we computed correlation between a parcels minimum distance with each of the three health statistics as well as the health score which was an average of the three. Below is the table that we got for each neighborhood.

Correlation Between Statistic and Minimum Distance to Open Space

Neighborhood	Obesity	Asthma	Low Physical Activity	Health Score
Roslindale	corr: -.1858 p-value: <.0001	corr: -.1849 p-value: <.0001	corr: -.1864 p-value: <.0001	corr: -.1874 p-value: <.0001
Jamaica Plain	corr: -.0543 p-value: <.0001	corr: -.0461 p-value: .0001	corr: -.0781 p-value: <.0001	corr: -.0665 p-value: <.0001
Mission Hill	corr: .2091 p-value: <.0001	corr: .3492 p-value: <.0001	corr: .4164 p-value: <.0001	corr: .3448 p-value: <.0001
Roxbury	corr: .1124 p-value: <.0001	corr: .0810 p-value: <.0001	Corr: .0720 p-value: <.0001	corr: .0949 p-value: <.0001
South End	corr: -.1549 p-value: <.0001	corr: -.1204 p-value: <.0001	corr: .0251 p-value: .1941	corr: -.0473 p-value: .0142
Back Bay	corr: .0313 p-value: .1996	corr: .1484 p-value: <.0001	corr: .0945 p-value: .0001	corr: .0948 p-value: .0001
Charlestown	corr: -..1437 p-value: <.0001	corr: -.1583 p-value: <.0001	corr: -.1658 p-value: <.0001	corr: -.1582 p-value: <.0001
Beacon Hill	corr: .1447 p-value: <.0001	corr: -.0295 p-value: .2858	corr: ..0355 p-value: .1999	corr: .0359 p-value: .1939
Fenway	corr: -.0393 p-value: .1997	corr: .1535 p-value: <.0001	corr: .0502 p-value: .1008	corr: .0423 p-value: .1661
Brighton	corr: .1336 p-value: <.0001	corr: -.0469 p-value: .0008	corr: ..0863 p-value: <.0001	corr: .1053 p-value: <.0001
West Roxbury	corr: .0144 p-value: .1615	corr: -.0610 p-value: <.0001	corr: -.1307 p-value: <.0001	corr: -.0739 p-value: <.0001
Hyde Park	corr: .0267 p-value: .0171	corr: .0570 p-value: <.0001	corr: .0511 p-value: <.0001	corr: .0416 p-value: .0002
Mattapan	corr: -..1366 p-value: <.0001	corr: .0953 p-value: <.0001	corr: -.1306 p-value: <.0001	corr: -.1293 p-value: <.0001
Dorchester	corr: -.1271 p-value: <.0001	corr: -.1429 p-value: <.0001	corr: -.1305 p-value: <.0001	corr: -.1347 p-value: <.0001

South Boston Waterfront	corr: -.3933 p-value: <.0001	corr: -6306 p-value: <.0001	corr: .3439 p-value: <.0001	corr: -.4540 P-value: <.0001
South Boston	corr: -..0807 p-value: <.0001	corr: -.1210 p-value: <.0001	corr: -.1301 p-value: <.0001	corr: -.1129 p-value: <.0001
Allston	corr: -.3596 p-value: <.0001	corr: -.2228 p-value: <.0001	corr: -.3172 P-value: <.0001	corr: -.3425 p-value: <.0001

Below is also a reservoir sampling of 500 parcels where we plot minimum distance against health score. The compute correlation and p-value for the below sample is (-.1720, .0001) respectively.



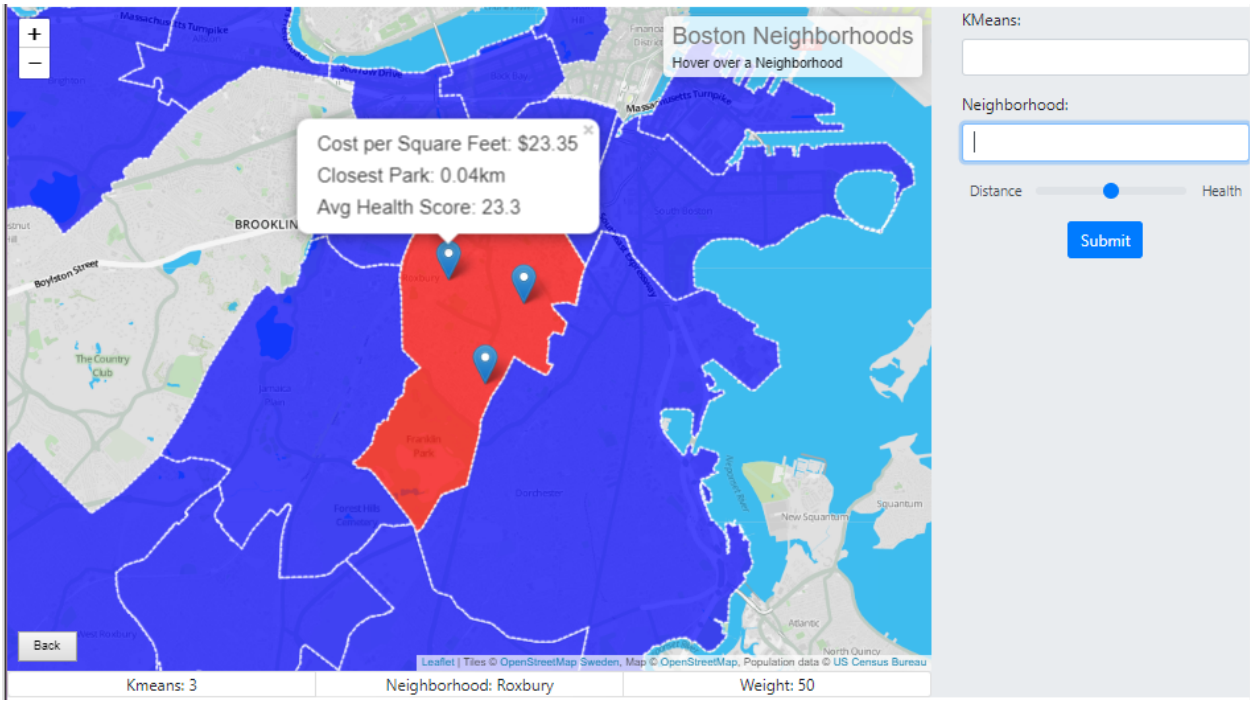
From the correlation table across each neighborhood, we see that some neighborhoods exhibit a positive correlation between minimum distance and health score, while more exhibited a negative correlation. This is actually counter to our hypothesis that the lower the access to open spaces an area has the more unhealthy that area is. A sample from the population, we also see that we have a negative correlation. Thus, it is not clear whether or not open spaces have any effect on improving the healthiness of an area. As a matter of fact, for many neighborhoods it is the contrary. However, these statistics should be taken with a grain of salt. First, this analysis has omitted many other variables like income and demographics that would most likely have an effect on an area's health. Also, if we look at the reservoir sampling, we will see that there seem

to be horizontal streaks corresponding to multiple points having the same health scores. This is a result of the lack of granularity in how we measured the health of a parcel in that we attributed a parcel's health to its census tract. Thus, multiple parcels will lie in the same census tract and thus have the same health score. Thus, we have less variation amongst health scores and thus it is hard to really say being closer to an open space leads to better health.

4 Visualization

4.1 Description

For our visualization, we decide to make an interactive map that lets you choose the neighborhood and number of k-means (from 1 to 5) that you want to display. We also allow the user to toggle the weight of each point between favoring distance scores or favoring health scores. This is implemented through a slider, where a lower slider value attributes to higher preference towards distance scores and a higher slider value corresponds to a higher preference towards health scores. Also, once the k-means are generated, we allow the user to click on the k-mean to get more information about that location. Below is an example of our visualization:



4.2 Implementation

To create this visualization, we employed an interactive client server using Flask. Thus, when the user inputs the number of k-means they want with a weight and a neighborhood, this is

sent as a post request to our server which is then handled by Flask. Flask takes in these inputs, gathers data from the local MongoDB repository, computes these k-means and their associated statistics, and renders these markers onto the map. The average time to handle requests is only a couple of seconds, but on some of the larger neighborhoods it might take longer to generate the k-means as there are significantly more parcels.

5 Conclusion

Based on our statistical analysis on the relationship between distance from an open space and health, we found that the correlation could be either positive or negative across different neighborhoods. It appears however that on the whole population, there is generally a negative correlation between health and minimum distance, which goes against our hypothesis. Most likely, at least just with the data we collected, we aren't able to discern the relationship between minimum distance from a park and health. This could be a result of many different factors, one being there might be other variables, such as economic disparity, that may be acting as an underlying influence on the data. With that being said, we are still able to prioritize less healthy areas in finding the optimal location for new parks, and the importance is not lost as in the future someone might be able to justify such a priority. With that being said, we have created a useful visualization that plots potential locations to place new parks depending on the user's specification. Additionally, we provide statistics about each location, which can serve as yet another factor to analyze whether or not this is a good place to put a new park.

Moving forward, someone can expand this optimization to cater to what someone thought was important in determining new park locations. This would be as simple as collecting the necessary data to apply this metric and changing how you would calculate the weight with this new metric. Also, it would be advisable to dig deeper into the relationship between health and open spaces. This could include gather more specific health data, including more variables, running multivariable regressions, or seeing if there is broader relationship between the density of open spaces in a certain census tract and the overall health of that tract. Hopefully, our analysis will serve as a tool in answer such questions, and a starting point for further analysis.