A Comparison of Bostons Neighborhoods:
Gaining Insights on Major Real Estate Factors

1. Introduction: Business Problem

Background:

The real estate market has always been a bit frightening for first time home buyers. A home, for a large majority of people, will be the largest purchase of their entire life. For that reason, people rightfully want to get it right. There are a large list of factors that go into the consideration of purchasing a home. In larger cities you can expect some added pressure. Housing has becoming increasingly scarce and people are being forced to make quicker and quicker decisions in order to close the deal before someone steals it out from under them.

Problem/ Motivation:

The home buying process is complicated. The information required to make an informed decision on buying a home is vast. That information needs collected and decided upon quickly. Whether you are a person just trying to make an informed and quick decision, you are a person just unfamiliar with the area and need to have a better understanding of all the different neighborhoods, or you are the real estate agent just looking to boost efficiency, I hope to provide some extra information to aid in that process.

We are going to attempt to provide a grouping of neighborhood that have similar characteristics. This will hopefully allow people to look at neighborhoods they might not have considered.

2. Data

Since our goal for this project is to group/cluster similar neighborhoods together, I wanted to collect the information that is most important to choosing a neighborhood to live in. For the purpose of this exercise we are going to be looking at crime data that is provided by data.boston.gov, We are going to be looking at high school rankings based on a USNews.com article, and we are going to be collecting a list of venues of things to do in each neighborhood. The list of venues are going to come from foursquare locations api.

The crime data and the high school rankings are going to be based from 2019 data, but have the potential to be expanded upon later. The venue data will be the most up to date data from December of 2019, since we are actually searching for the data around that time.

Boston Crime Data: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/12cb3883-56f5-47de-afa5-3b1cf61b257b
Boston Neighborhood Geo Spatial Data: https://data.boston.gov/dataset/boston-neighborhoods
Boston Highschool Rankings: https://www.usnews.com/education/best-high-schools/massachusetts/districts/boston-public-schools-111992
Venue data: https://foursquare.com/

explanation of data considerations:

I briefly wanted to address why these fields were chosen and other categories were left out. I feel like the categories listed above are going to be things that you will definitely experience in the neighborhood. Everyone one is concerned with safety. I feel like most people looking to live in a city would like ample things in the area to do, such as coffee shops or museums. I realize high school rankings might not apply to everyone, but I think a majority of people probably consider the education of the area for potential children.
I chose to leave out factors like commute time, because that can vary so much depending on the person applying. I also left out average real-estate prices because I wanted to try to allow for the clustering algorithm later to not be influenced by over inflated prices of an area. I wanted this data comparison to be strictly on what I would call "the merits" of the neighborhood. Just because an area is expensive doesn't inherently make it more fun, more safe, or more educated.
I also chose to leave out things like elementary school and middle school rankings for simplicity. I also didn't want to add to many similar categories at the risk of adding noise. I know early education is important, but for the first pass at this data i'm going to only consider high school rankings.

3. Methodology:

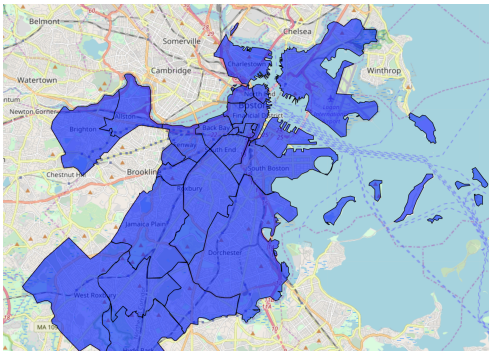
By the nature of our topic it doesn't need preliminary data investigation, but rather our data collection will lead to a greater understanding at the end. We are going to be fetching a variety of data but we aren't going to be looking for particular insights during those stages.

3a. Geo Spatial data: Getting our neighborhoods defined

Since all of this data hinges around the neighborhoods of Boston, it only makes sense than for us to define the neighborhoods and their parameters. So we are going to import the spatial coordinates of the neighborhoods from https://data.boston.gov/dataset/boston-neighborhoods and map these to over top of a folium map.

The data we are pulling from are pulling came in a polygonal form as show below. We were able to take this data and use the python folium library to map out a choropleth map.

geometry
MULTIPOLYGON (((-71.12593 42.27201, -71.12611 ...
POLYGON ((-71.10499 42.32610, -71.10503 42.326...
POLYGON ((-71.09043 42.33577, -71.09050 42.335...
POLYGON ((-71.09811 42.33673, -71.09832 42.337...
POLYGON ((-71.06663 42.34878, -71.06663 42.348...



3b. Boston Crime Data: Organize, Clean, Understand

Our crime data was vast. Fortunately the Boston government website had a very easily downloadable, regularly refreshed, and clearly defined data set. With all of their efforts though it wasn't going to quite match how we needed the data at the end of the project. Since we can see our choropleth map is broken out by neighborhood we are also going to need to organize all of our subsequent data in a neighborhood format. Below you can see some of the variety of columns provided in the original data set.

```
bcd_df = pd.read_csv('Boston_Crime_Data.csv')
bcd_df.head()
```

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE
0	TESTTEST2	423	NaN	ASSAULT - AGGRAVATED	External		0	2019-10-16 00:00:00
1	I92097173	3115	NaN	INVESTIGATE PERSON	C11	355	0	2019-10-23 00:00:00
2	I92094519	3126	NaN	WARRANT ARREST - OUTSIDE OF BOSTON WARRANT	D14	765	0	2019-11-22 07:50:00
3	I92089785	3005	NaN	SICK ASSIST	E13	574	0	2019-11-05 18:00:00
4	I90583827	1402	NaN	VANDALISM	E18	498	0	2019-11-02 05:09:00

```
bcd_df = pd.read_csv('Boston_Crime_Data.csv')
bcd_df.head()
```

	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR	UCR_PART	STREET	Lat	Long	Location
	External		0	2019-10-16 00:00:00	2019	10	Wednesday	0	NaN	RIVERVIEW DR	NaN	NaN	(0.00000000, 0.00000000)
	C11	355	0	2019-10-23 00:00:00	2019	10	Wednesday	0	NaN	GIBSON ST	NaN	NaN	(0.00000000, 0.00000000)
	D14	765	0	2019-11-22 07:50:00	2019	11	Friday	7	NaN	BROOKS ST	NaN	NaN	(0.00000000, 0.00000000)
	E13	574	0	2019-11-05 18:00:00	2019	11	Tuesday	18	NaN	WASHINGTON ST	NaN	NaN	(0.00000000, 0.00000000)
	E18	498	0	2019-11-02 05:09:00	2019	11	Saturday	5	NaN	BRADLEE ST	NaN	NaN	(0.00000000, 0.00000000)

These images don't contain all of our columns, but fortunately we are only interested in a few of them anyway. What we are interested in is the location data ("Lat" and "Long" columns) and the "OFFENSE_DESCRIPTION" column. With an understanding of what type of crimes happened and where they happened, we will be able to take those coordinates and map them to a neighborhood by looping over our geo spatial data.

There was a cleaning process to this data. There were so many rows of data that it really only made sense to look at current year data. This could be improved by looking at the deltas in crime over time, but for the purposes of this project I wanted to stick with current year data. Our data also contained some entries without any location data. This would prevent from mapping it to a neighborhood so we had to remove those rows.

After the cleaning of the data all that was left was formatting. As stated earlier this data needed to be mapped to a neighborhood in order to work with our later analysis. So we ran our coordinates through our “Shape” library that was able to determine if a given point fell within the bounds of the shapes we created over our folium map. The data after that stage just needed to be grouped such that we could a count of crimes per neighborhood. You can see the results of that here, on the right.

Neighborhood	Count of Offenses
Allston	1817
Back Bay	2715
Bay Village	151
Beacon Hill	560
Brighton	2474
Charlestown	1312
Chinatown	566
Dorchester	16642
Downtown	4857
East Boston	2709
Fenway	1958
Hyde Park	3180
Jamaica Plain	3452
Leather District	98
Longwood	278
Mattapan	3367
Mission Hill	1219
North End	582
Roslindale	2143
Roxbury	9304
South Boston	3328
South Boston Waterfront	541
South End	3079
West End	846
West Roxbury	1619

3c. Boston Public School Data: A Web Scraping exercise

For this portion of the project we are going to need to do a bit more work for our data. The Boston crime data was nicely formatted for the most part and was able to just be downloaded from the city governments website. However with the public school data, the metrics I was looking for weren't really available. So we are turning to a US News article that ranks the schools.

It provides a lot of different metrics on each school, but we are going to just focus on college readiness, graduation rates, and enrollment. For a more verbose look at the neighborhoods, you could dig a bit deeper on the different metrics of each school.
So from this article we are going to be scraping the data about these schools.

In this section we did run into some issues with collecting all of the data. The first challenge was getting our web scraping method to collect all of our data. Because the news article was segmented into multiple section with a “load more” button, it wasn’t capturing all of the data. After some sleuthing, I realized that the load more button actually just called to a different page. From that I was able to just run our web scraping technique twice to get all of the data.

Our next challenge was again around the formatting of the data. Some of the entries in the list contained the neighborhood in which the school was located, however some of the schools were listed on the name “Boston” instead. So we needed to take the short list of schools that contained that distinction and map them to the appropriate neighborhoods.

After performing the various cleaning techniques to get our strings formatted, rows aligned, and data aggregated to one row per neighborhood we were left with the resulting data frame. Below is all of the aggregated school data shown.

Neighborhood	Graduation Rate %	College Readiness	Enrollment
Allston	44.000000	0.000000	33.000000
Back Bay	51.500000	0.000000	318.000000
Bay Village	85.000000	34.100000	218.000000
Brighton	70.000000	3.900000	461.000000
Charlestown	55.000000	20.900000	867.000000
Dorchester	69.444444	13.788889	428.222222
East Boston	75.000000	19.100000	1480.000000
Hyde Park	86.333333	31.033333	346.333333
Jamaica Plain	56.250000	17.000000	272.250000
Longwood	97.000000	95.300000	1630.000000
Roslindale	36.000000	0.000000	184.000000
Roxbury	81.333333	18.900000	769.000000
South Boston	70.000000	20.200000	524.000000
South End	32.000000	0.000000	186.000000
West Roxbury	63.000000	10.600000	458.500000

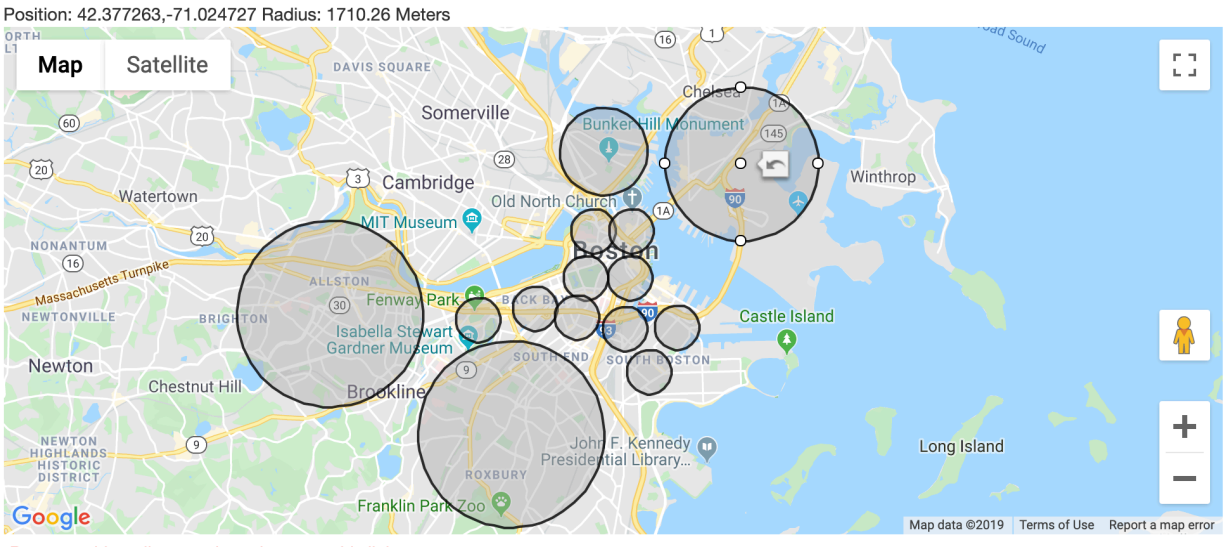
3d. Boston venues: Foursquare API calls to find the fun

Now we are on to our final large chunk of data. We are going to query against the Foursquare location data set in order to grab all of the various venues in Boston. Since there is no way to query all venues for a particular city, we need to be a little creative.

Since each query only allows for a max return of 50 venues, we need to break out our queries. I've utilized a third party tool that allows you to draw a circle over top of a map. This provides its longitude and latitude coordinates and the size of the radius in meters and miles. I've used a series of circles that will cover the whole map of Boston. This will require some later data cleaning to remove duplicates of the overlapping areas, but it gives us a more comprehensive look at the city venue list.

Below you can see roughly how these circles were obtained. I used the tool from the link below, to make this map.
https://www.mapdevelopers.com/draw-circle-tool.php

While this map is not specifically the one I used, this is a quick recreation. The original map contained too many points for easy viewing.



After performing the necessary steps to get all of our venues we then needed to map all of that information to a neighborhood like all the other sections. After the mapping we were left with data the reflects the sample below

	categories	lat	lng	Neighborhood
0	Other Great Outdoors	42.364537	-71.066308	West End
1	Dentist's Office	42.363713	-71.065374	West End
2	Church	42.363131	-71.065304	West End
3	Pool	42.363163	-71.064814	West End
4	Office	42.363739	-71.066325	West End

However upon analysis of this data set I was starting to become worried. My original plan was to have these categories transposed into columns that we will use to perform our Means on later. After sorting through the data I realized that there were literally hundreds of unique categories provided. So in an effort to paint a picture of what a neighborhood had, I needed to reduce these categories into high level categories. After creating the high levels and painfully going through nearly all of the categories and assigning them manual to one of these, I was left with this list

```
food(Restaurant)           306
services                   256
Fun Activities/ Venues     253
pertaining to transit      148
shopping(goods)            140
medical                    117
night life                 107
food(sweets and coffee)     88
shopping(food)              82
religion                    34
```

From this list I felt a lot more comfortable using it to represent what a neighborhood had to offer. This strips some of the unique elements, but it gives a broader picture from which we can draw conclusions later on.

Our data eventually was assigned to its corresponding neighborhood. It need some transformation to allow it to join our other data sets later, but here is a sample of what that data looked like.

		High Level Categories
Neighborhood	**High Level Categories**	
Allston	services	15
	pertaining to transit	11
	shopping(goods)	11
	food(Restaurant)	9
	food(sweets and coffee)	5
	medical	5
	Fun Activities/ Venues	4
	night life	4
	shopping(food)	4
	services	21

3e. Collection: Joining of data

Now we have all three of our data sets. However there is still one last step before we can begin the machine learning portion of the project. We need to join all of our data into one master data frame that is going to be organized based on the neighborhood.

We did this already in the Boston crime data to create our master data frame. Now we just need to format down the other data sets down. Since each data set contains multiple data points per neighborhood we are going to need to group that in a way that makes sense.

After doing some formatting to remove nulls, set the number of decimal points, and other string manipulations to allow for easy joining we are left with the data set below. The image below is just a sample. As you can see we have more rows and columns but this gives an idea of how the final data set was formatted.

	Count of Offenses	Graduation Rate %	College Readiness	Enrollment	Fun Activities/ Venues	food(Restaurant)	food(sweets and coffee)	medical	night life	pertaining to transit	religion	services	shopping(foo
Allston	1817.0	44.0	0.0	33.0	4.0	9.0	5.0	5.0	4.0	11.0	0.0	15.0	4
Back Bay	2715.0	51.5	0.0	318.0	3.0	20.0	4.0	2.0	4.0	4.0	0.0	21.0	2
Bay Village	151.0	85.0	34.1	218.0	1.0	0.0	0.0	0.0	2.0	2.0	1.0	1.0	0
Beacon Hill	560.0	0.0	0.0	0.0	8.0	6.0	4.0	1.0	4.0	2.0	1.0	9.0	0
Brighton	2474.0	70.0	3.9	461.0	2.0	9.0	3.0	2.0	5.0	1.0	2.0	6.0	1
Charlestown	1312.0	55.0	20.9	867.0	23.0	13.0	8.0	3.0	6.0	9.0	2.0	17.0	1
Chinatown	566.0	0.0	0.0	0.0	2.0	21.0	7.0	5.0	0.0	1.0	0.0	4.0	0
Dorchester	16642.0	69.4	13.8	428.2	17.0	37.0	8.0	3.0	4.0	11.0	8.0	34.0	13
Downtown	4857.0	0.0	0.0	0.0	20.0	33.0	12.0	9.0	16.0	13.0	1.0	21.0	2
East Boston	2709.0	75.0	19.1	1480.0	16.0	27.0	3.0	2.0	10.0	22.0	5.0	20.0	9

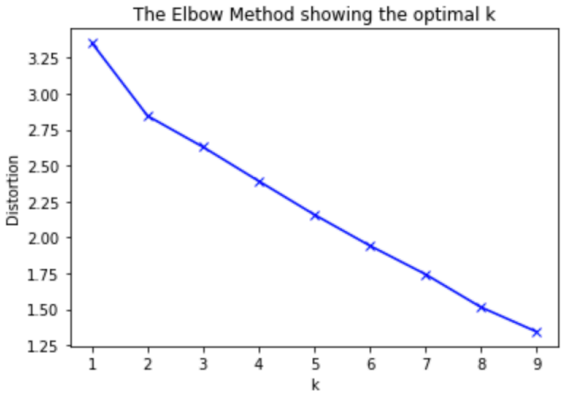
3f. Machine Learning and visualization

We finally have our data formatted, cleaned, and joined in the way we need it to be. Now that we have the data frame in its final form we are going to perform a KMeans clustering on it. This clustering algorithm is going to look at all of the columns and try to assign each neighborhood to a particular group. Based on this grouping we are going to be able to determine which neighborhoods are similar to one another.

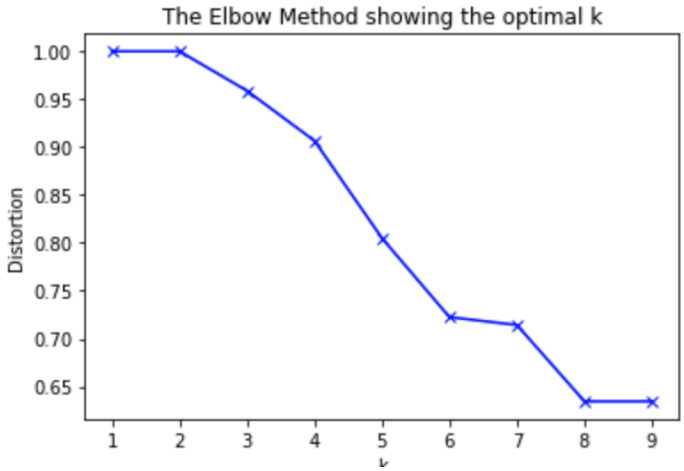
The first step was scaling our data. Feeding in the information directly from the previous section could end up putting more weight on things with higher numbers. This could effect the results of really small or really large neighborhoods. Below is the data frame after performing standard scaling with SKLearn.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	-0.244278	0.189854	-0.551413	-0.647438	-0.641242	-0.228461	0.546257	0.062653	0.089757	1.153510	-0.642999	0.709603	0.293049	0.955123
1	0.020320	0.404721	-0.551413	0.008211	-0.772520	0.878994	0.223469	-0.286412	0.089757	-0.218639	-0.642999	1.447589	-0.232418	1.948452
2	-0.735169	1.364458	1.165170	-0.221841	-1.035076	-1.134560	-1.067685	-0.519121	-0.493666	-0.610682	-0.151294	-1.012366	-0.757885	-0.865979
3	-0.614656	-1.070697	-0.551413	-0.723355	-0.116131	-0.530494	0.223469	-0.402766	0.089757	-0.610682	-0.151294	-0.028384	-0.757885	1.782897
4	-0.050692	0.934725	-0.355088	0.337186	-0.903798	-0.228461	-0.099320	-0.286412	0.381469	-0.806703	0.340411	-0.397377	-0.495152	-0.369314

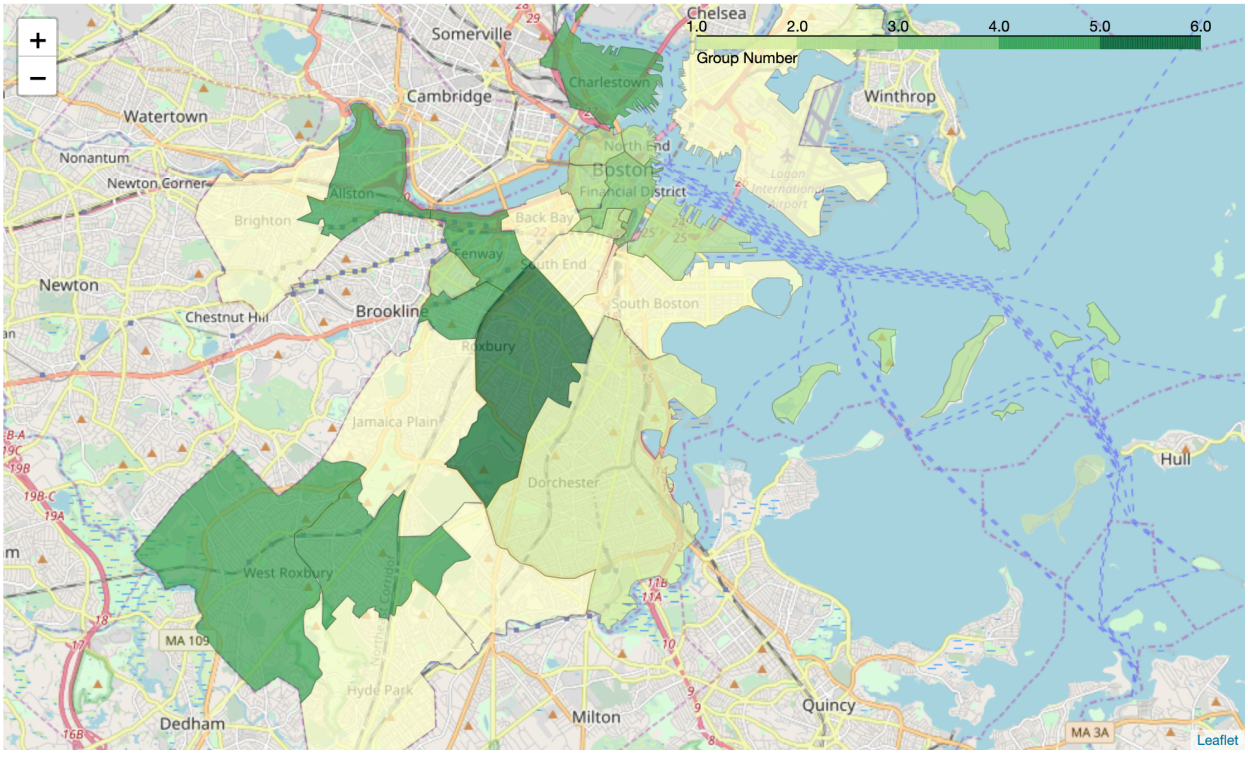
Next we had to determine how many clusters we wanted to group our neighborhoods into. This is a bit challenging because we need to essentially just test a number of groupings and see which one results in the most accurate groupings. A standard method of performing this test is to calculate the distances between the centroid of the cluster and evaluate these groupings. Once we do that for the different values of groups, we plot that and look for an area of sharp change. This change indicates the optimal grouping number.



You can see from the graph above that when performing this method with a euclidean doesn't really provide a strong indicator at any of the pints. So I changed the distance calculation method to that of a Jaccard distance to get the graph below.



Here we can see a sharp change starting at 6 and also one at 8. We are going to choose the first of the changes and set our grouping number to 6. After we apply that number to our Means algorithm we finally get the groups we are looking for. Once we format and at the groups to our data frame we can map it over top of Boston again to get this map.



4. Discussion

We now finally have a complete image of how our neighborhoods relate to each other. Now to be clear, these numbers in no way associate closeness to one another. That is to say if something was labeled a 2, it would not necessarily be closely related to a 1 or a 3. These are just groupings to show that anything within a group is similar to other things within that group. More analysis would need to be done to see to what magnitude these neighborhoods were related they were. I think I've covered a lot of the short comings of the methodology used to generate this map, but I think it is worth repeating. This project is merely a framework from which to build to gain a more accurate picture. For school data, you would probably want more metrics than what was provided. The crime data could probably have been broken down into more granular categories, and the venue data could be more carefully vetted to allow more unique venues to shine through.

With all of that said I think this is a very interesting look at which neighborhoods relate to one another. We can see that a lot of neighborhoods are related in proximity to one another, but it is interesting to see how you can get collections or pockets of similarity. I think that is what a normal person might guess based on an understanding of how most cities behave, but it is interesting to see it visually.It is also interesting to see how in the example of Dorchester and Roxbury that they don't relate to any other neighborhood. They are unique in their characteristics based on this data set.
To circle back to the original purpose of this project, I think it is worth looking at one example. If you have been to Boston, then you have probably visited the Back Bay region. This is where the most wealthy people in the city live. It's right next to all the trendy shopping and dining places. This of course comes with a gigantic real estate price tag. With this information we can see just how many neighborhoods contain similar characteristics to it. One note-able example is East Boston. East Boston real estate is significantly cheaper.This comparison was suppose to be able to draw insights on finding similar neighborhoods that were cheaper but contained major elements that would be of interest to a home buyer. I think from this example we can reasonably assert we achieved our goal.

5. Conclusion

This project has a lot of components to it. There are a wide variety of data sources, each with their own considerations and complications. With all of that data collected, organized, and analyzed I think we were able to provide a valuable foundation for real estate insights.

We were able to capture our most recent crime data; we scraped education data from news websites; we diligently searched for all of the unique things that make of the neighborhoods with the foursquare API. After all of that we performed Kmeans clustering on our data and remapped it over Boston with these groupings.
Through all of our manipulations we gained insight on to the relationship that each neighborhood has to others within the bounds of Boston. This insight could be very valuable for quick decision making for a home buyer or real estate agent alike.