Clustering the Neighborhoods of Boston

By Athokshay Ashok

I. Introduction and Business Problem

Boston is a historically rich and economically booming city that is always bustling with life and energy. With a number of universities and companies at its heart, Boston is a hub for all kinds of businesses. However, due to its vibrancy, the city of Boston has become extremely competitive in terms of housing. Various sources report that the cost of living in Boston is 48% greater than the national average. For families moving into the city, it can be confusing as to where specifically in the city they should move to, especially since the cost of living and type of venues varies greatly in each neighborhood.

Boston has 22 neighborhoods, each of which is vastly different from the others in terms of demographics, venues, and per capita income. While some neighborhoods are lined with designer stores and fancy restaurants, others are quieter suburbs that are located next to universities. From bubble tea stores in Chinatown to streets full of pubs downtown, there is a vast cultural difference between the areas as well. The locations and venues in each neighborhood tell us a lot about the affordability of the region, and how compatible a family would be in that neighborhood.

The purpose of this project is to use data from the Foursquare API to access geographical data regarding venues all over the city of Boston and classify them by neighborhoods using a K-Means Clustering ML algorithm. The average housing price of each neighborhood will also be examined in parallel to provide a holistic view of each region of Boston that families may want to consider before moving. Finally, maps of the area will be generated using the Folium library in Python to show the clusters by location.

II. Data Description

1. List of Boston Neighborhoods: https://en.wikipedia.org/wiki/Neighborhoods_in_Boston. This website gives an overview of how the city of Boston is split into regions and gives a list of the neighborhoods.
2. Latitude.to: https://latitude.to/. This website returns the coordinates of any mentioned location. Since the data for the coordinates of the neighborhoods of Boston does not

already exist in a tabular format to extract, this website will be used to manually get the coordinates.

3. Foursquare API : https://developer.foursquare.com/. This API will be used for accessing venues at or near the desired locations. It allows us to make 500 premium calls a day that return information about each venue such as coordinates, venue category, neighborhood of the venue, etc. Credentials for a developer account were used to obtain a client ID and client secret.
4. Zillow: www.zillow.com. This website is widely used for buying and selling houses, and contains data about average housing prices of the Boston neighborhoods.
5. Analyze Boston: https://data.boston.gov/dataset/boston-neighborhoods. This website contains the geojson data for the neighborhoods of Boston which can be used to generate choropleth maps.

III. Methodology

To begin with, I created a Pandas dataframe of the neighborhoods of Boston and their latitudes and longitudes, which I accessed from the data sources mentioned in the previous section and manually created lists out of them. Below is the head of the dataframe which displays 5 of the 22 neighborhoods.

	Neighborhood	Latitude	Longitude
0	Allston	42.3539	-71.1337
1	Back Bay	42.3503	-71.1337
2	Bay Village	42.3490	-71.0698
3	Beacon Hill	42.3588	-71.0707
4	Brighton	42.3464	-71.1627

Figure 1: Boston Neighborhoods Dataframe

After accessing the exact latitude and longitude of Boston, MA using the geocode library, I used the Folium python library which is a powerful tool for graphing geographical maps. Setting the frame of the map to the city of Boston, I displayed each neighborhood as a marker on the map using the coordinates from the above dataframe.

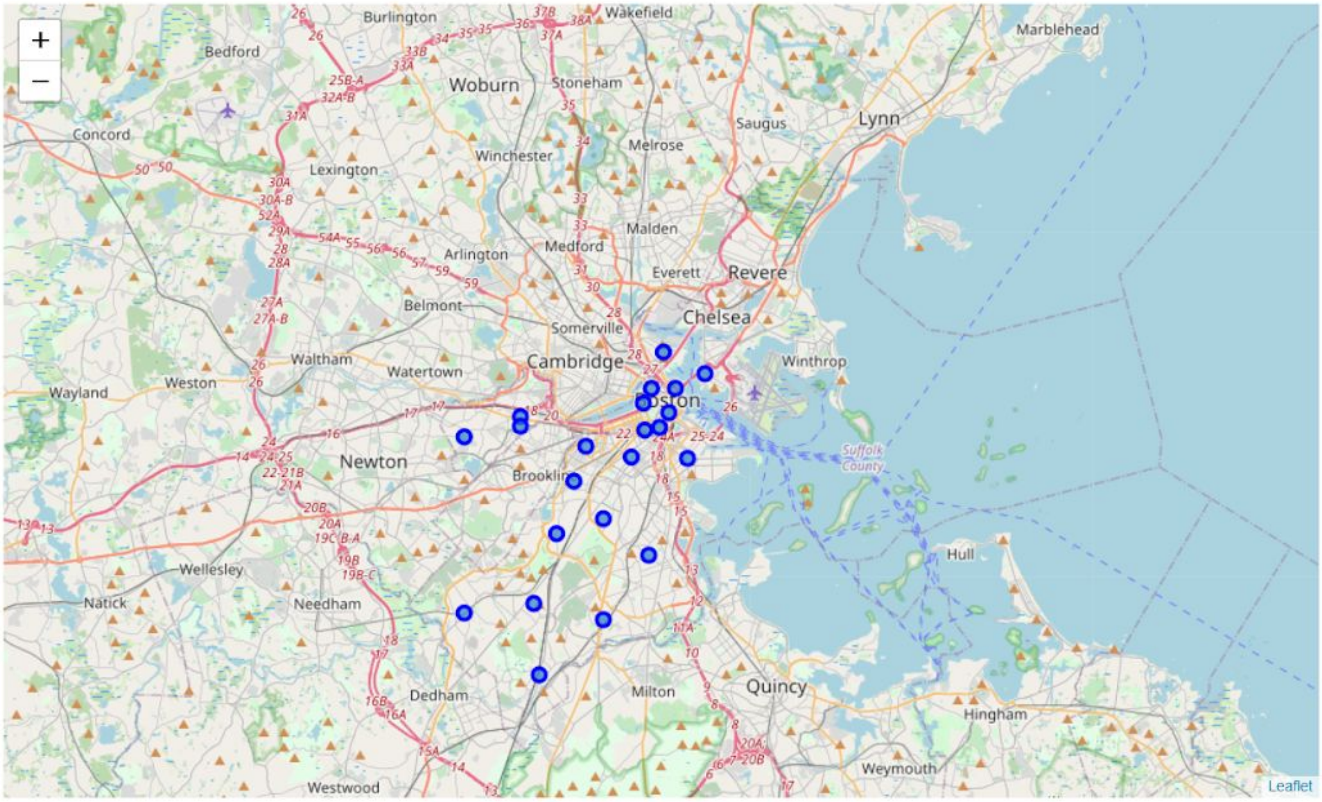


Figure 2: Map of the Neighborhoods of Boston using Folium

Next, I connected to the Foursquare API using my client ID and client secret from the developer account to access information regarding venues in each of the neighborhoods. I set the limit to 100 venues and the search radius to 700 meters for each neighborhood from their given latitude and longitude. Below is a head of the list of venues, the category of the venue, and coordinate locations that Foursquare returned all grouped together in a dataframe.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allston	42.3539	-71.1337	Lulu's Allston	42.355068	-71.134107	Comfort Food Restaurant
1	Allston	42.3539	-71.1337	Fish Market Sushi Bar	42.353039	-71.132975	Sushi Restaurant
2	Allston	42.3539	-71.1337	Azama Grill	42.354422	-71.132358	Falafel Restaurant
3	Allston	42.3539	-71.1337	Brighton Music Hall	42.352844	-71.132565	Rock Club
4	Allston	42.3539	-71.1337	Mala Restaurant	42.352960	-71.131033	Chinese Restaurant
5	Allston	42.3539	-71.1337	Whole Heart Provisions	42.353745	-71.137189	Vegetarian / Vegan Restaurant
6	Allston	42.3539	-71.1337	Kaju Tofu House	42.354329	-71.132374	Korean Restaurant
7	Allston	42.3539	-71.1337	BonChon Chicken	42.353105	-71.130921	Fried Chicken Joint
8	Allston	42.3539	-71.1337	Deep Ellum	42.353844	-71.136923	Gastropub
9	Allston	42.3539	-71.1337	Thai Place	42.353185	-71.133784	Thai Restaurant

Figure 3: List of Nearby Venues for Each Neighborhood from Foursquare

Grouping the venues by neighborhood, I generated a data frame of how many venues were extracted for each of the 22 neighborhoods. Shown below is the head of the dataframe.

Neighborhood	Venue
Allston	100
Back Bay	78
Bay Village	100
Beacon Hill	62
Brighton	40

Figure 4: List of the Number of Venues Returned, Grouped by the Neighborhood

It is clear that while we were able to reach the limit of 100 venues for some neighborhoods, the Foursquare API returned less than 100 for others and this is a result of the latitude and longitude values used. To increase the data set, more precise coordinates can be passed in but for the sake of consistency, we will use the same coordinate information.

I also noted that of the 1357 total venues that were generated for the 22 neighborhoods, there were 231 unique categories of venues. After one-hot encoding each of these categories as features of the data set and standardizing the values, I generated a dataframe of the top 10 categories of venues found in each neighborhood. Below is the head of the dataframe.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allston	Korean Restaurant	Pizza Place	Chinese Restaurant	Thai Restaurant	Coffee Shop	Bakery	Bar	Rock Club	Sushi Restaurant	Donut Shop
1	Back Bay	Korean Restaurant	Pizza Place	Chinese Restaurant	Mexican Restaurant	Rock Club	Bar	Japanese Restaurant	Gastropub	Sushi Restaurant	Donut Shop
2	Bay Village	American Restaurant	Spa	Theater	Hotel	Seafood Restaurant	Italian Restaurant	Gym	Mexican Restaurant	Gourmet Shop	Wine Bar
3	Beacon Hill	Park	Hotel Bar	Italian Restaurant	Pizza Place	American Restaurant	Restaurant	Hotel	Playground	Café	Gourmet Shop
4	Brighton	Pizza Place	Chinese Restaurant	Sushi Restaurant	Bank	Coffee Shop	Supplement Shop	Lake	Donut Shop	Dry Cleaner	Noodle House

Figure 5: Top 10 Venue Categories for Each Neighborhood

Since there are common venue categories for the neighborhoods, I used a K-means clustering algorithm, which is a popular unsupervised learning model, to generate clusters of the neighborhoods based on the common venues. To determine the optimal k value, I used the elbow method with the Calinski Harabasz metric and timings set to false. From the chart below, it was evident that 5 clusters should be made of the neighborhoods.

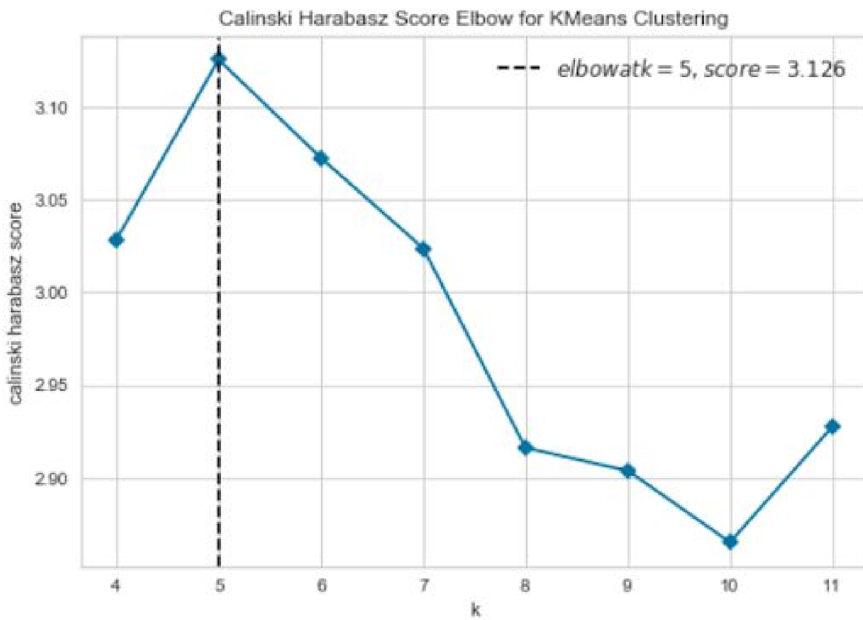


Figure 6: Elbow Method to Determine Optimal k Value

IV. Results and Discussion

From the clustering model, cluster labels were generated for each of the neighborhoods, as shown in the head of the dataframe below.

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mc Comm Ven
0	Allston	42.3539	-71.1337	2	Korean Restaurant	Pizza Place	Chinese Restaurant	Thai Restaurant	Coffee Shop	Bakery	Bar	Rock Club	Su Restaura
1	Back Bay	42.3503	-71.1337	2	Korean Restaurant	Pizza Place	Chinese Restaurant	Mexican Restaurant	Rock Club	Bar	Japanese Restaurant	Gastropub	Su Restaura
2	Bay Village	42.3490	-71.0698	4	American Restaurant	Spa	Theater	Hotel	Seafood Restaurant	Italian Restaurant	Gym	Mexican Restaurant	Gourn Sh
3	Beacon Hill	42.3588	-71.0707	4	Park	Hotel Bar	Italian Restaurant	Pizza Place	American Restaurant	Restaurant	Hotel	Playground	Ca
4	Brighton	42.3464	-71.1627	2	Pizza Place	Chinese Restaurant	Sushi Restaurant	Bank	Coffee Shop	Supplement Shop	Lake	Donut Shop	Dry Clear

Figure 7: Neighborhoods with Cluster Labels

The idea was to identify what category of venue each cluster was most closely associated with. For example, if all the neighborhoods where the 1st most common venue was a pizza place were placed in one cluster, then we could assign that cluster a label of “Pizza”. However, when I examined the clusters, there did not seem to be a clear correlation between the 1st most common venue and the neighborhoods in the cluster, so I decided to join the top three most common venues for each neighborhood and display them as data points on the map.

In the map shown below, the neighborhoods are categorized into clusters by color and clicking on any of the neighborhoods will tell you the cluster that the neighborhood belongs to and the top 3 most common venue categories in the area.

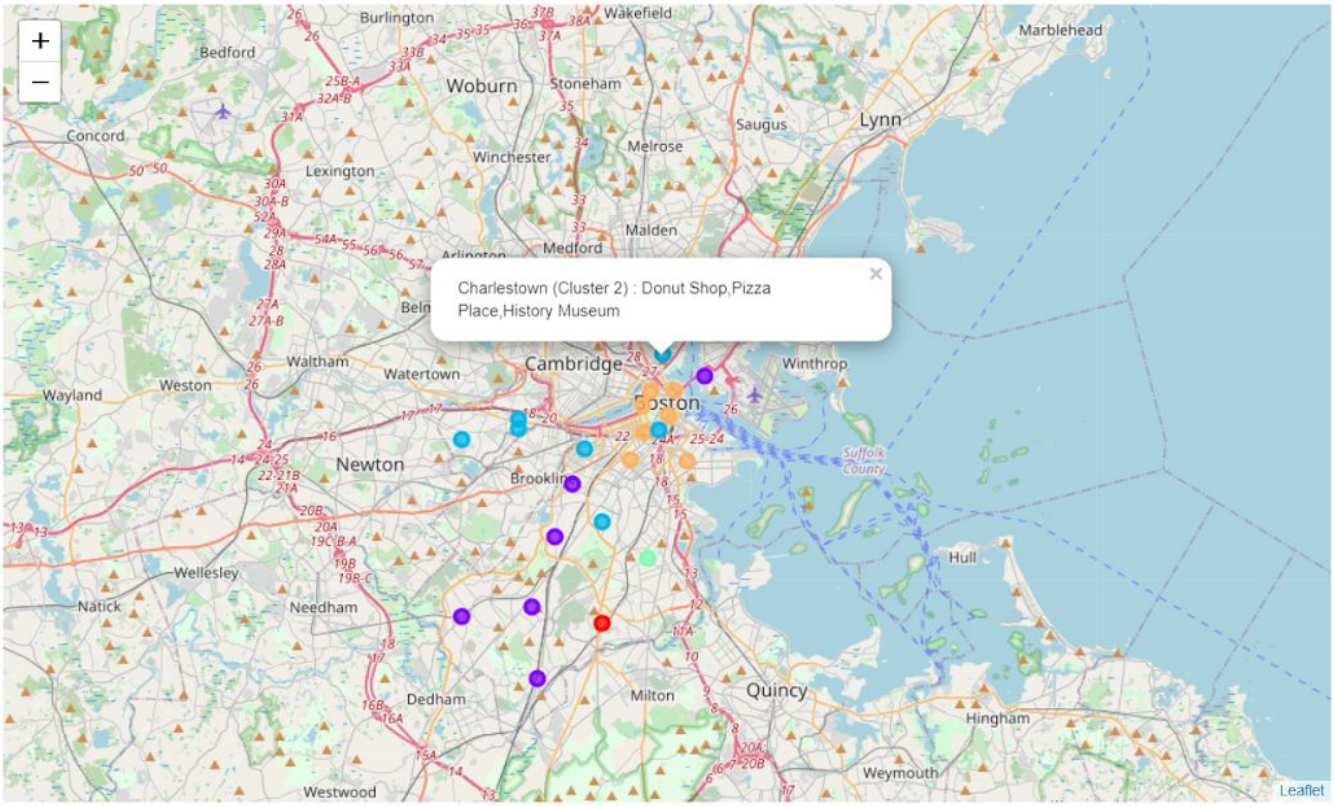
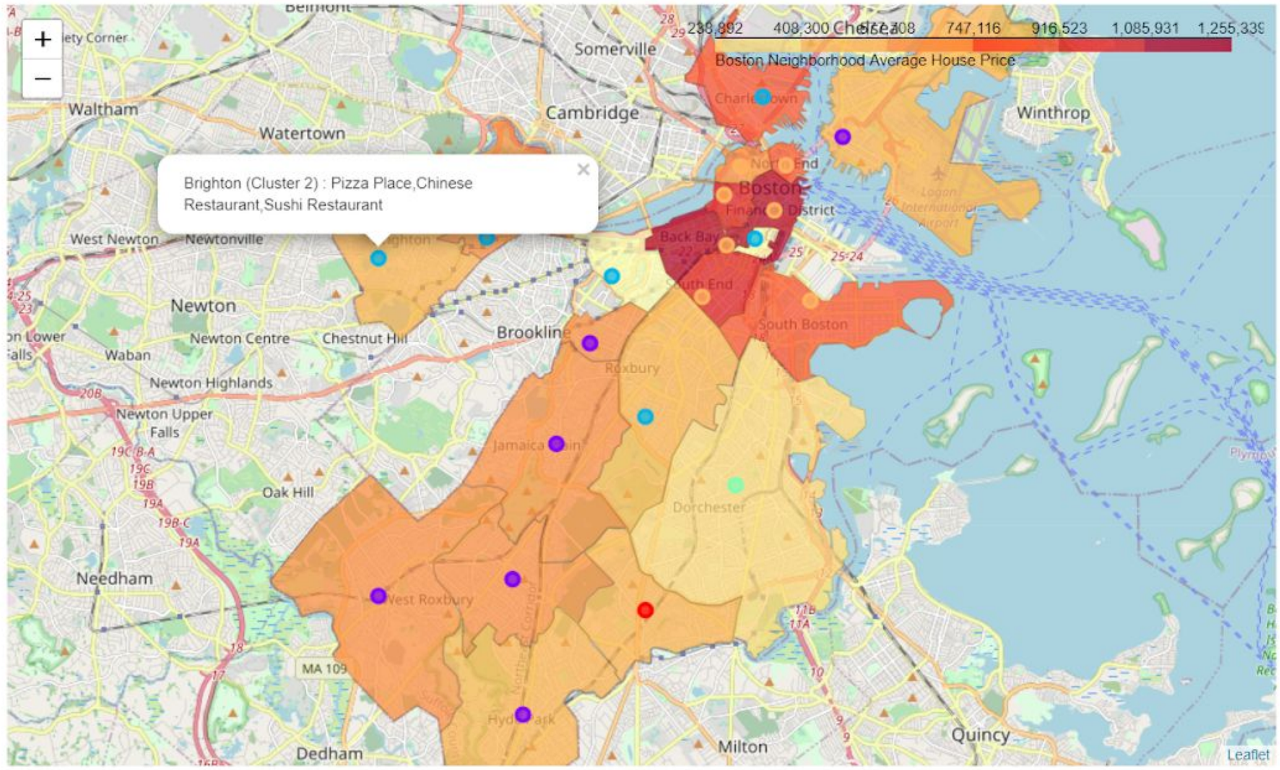


Figure 8: Clustered Neighborhoods with Top 3 Most Common Venue Categories Labels

To understand how the average housing price varies with the neighborhood and the clusters, I generated a choropleth map using the boston geojson data acquired from Analyze Boston. Since this geojson data had more than 22 neighborhoods, I had to clean it first to include only the 22 neighborhoods that we are concerned with. The average housing prices were found on Zillow. Adding the above labeled cluster points to the choropleth map generated the following result.



<u>Figure 9: Choropleth Map of Boston Based on Average Housing Price Along with Neighborhood Clusters By Venue Category</u>

The clusters make more sense now with the average housing prices. We can see that at the heart of the city where the housing prices are high, the most common venues are “fancy” places such as spas, Italian restaurants, hotels, and theaters. These correspond to the cluster points marked in orange. In the neighborhoods with less expensive housing prices, the most common venues are more “common” places such as coffee shops, pizza places, and Chinese restaurants. These primarily correspond to the cluster points marked in blue and purple. It appears that there is a correlation between the housing prices and the category of venues located in the neighborhoods. For someone who is single, looking to explore the city life, and has the means to do so, living in the neighborhoods that lie at the heart of the city will be the best bet. For families with kids looking for a moderate lifestyle, the neighboring areas are the safest choices though it may be a bit of a travel to explore the lavish lifestyle of the inner city.

V. Conclusion

In this project, I explored the categories of most common venues and the average housing prices of the 22 neighborhoods in Boston and discovered that there is a strong correlation between the two. Depending on the lifestyle you wish to have, you may have to compromise on one of the two above mentioned factors. For families moving to the city in search of affordable living, there are quite a few options of neighborhoods to move to depending on the types of venues located in the area. In some instances, people may wish to live with others of their ethnicity or country of origin, and these can be reflected in the most common venues of the areas. Boston is a culturally extremely diverse city, so it can be safe to say that no two places will be the same.

Given more time and resources, I would like to investigate other factors such as the average school ratings of the neighborhoods, which are important for families to consider when moving. No such dataset was available for the Boston schools, but I hope to expand on this project in the time to come.