Capstone Project- The Battle of Neighborhoods Final Report

Nigel Chan

1 Introduction

1.1 Background

Boston is the capital and most populous city of the Commonwealth of Massachusetts. Boston is sometimes called a "city of neighborhoods" because of the profusion of diverse subsections; the city government's Office of Neighborhood Services has officially designated 23 neighborhoods.

1.2 Problem

When moving a new city, there are numerous neighborhoods to choose from. One of the considerations to shortlist potential neighborhoods to live in would be the availability of different services within that neighborhood. For example, some people might place a high emphasis on the availability and proximity of restaurants and other services, in addition to other common considerations such as proximity to work, ground transport connectivity, demographics, types of accommodation, density, proximity to parks and other public facilities etc.

In this project, I will specifically be focusing on identifying the different types of amenities (restaurants, cafes, fun, shopping, nightlife - i.e. categories available in Foursquare data) within the various Boston neighborhoods. I will then analyze which neighborhoods have the most diverse range of amenities. I will also include population density as well as median home values as parameters to cluster neighborhoods. Subsequently to analyze the resulting clusters to provide insights into the types of neighborhood clusters produced by the data.

1.3 Interest

This project will target people who are moving to Boston for work and are interested in identifying the best neighborhoods to live in based on the range and number of different services available, as well as its population density and cost of homes (based on the median home value). This is important to those who value convenience and diversity of options near

where they live, as well as have a preference towards how dense the area is, and the relative cost of purchasing a home in those neighborhoods.

2 Data

2.1 Data Sources

The list of Boston's zip codes was obtained form from Boston.gov and use `uszipcode` package to populate the name, latitude, longitude, population density and median home value associated with that zip code.

The second main source of data is Foursquare. Venue information is queried from Foursquare using API calls using the latitude and longitude of each zip code and limiting results to a range of 500m.

2.2 Data Cleaning

In the downloaded file of zip codes, there were duplicate rows with the same zip codes. The duplicates were removed.

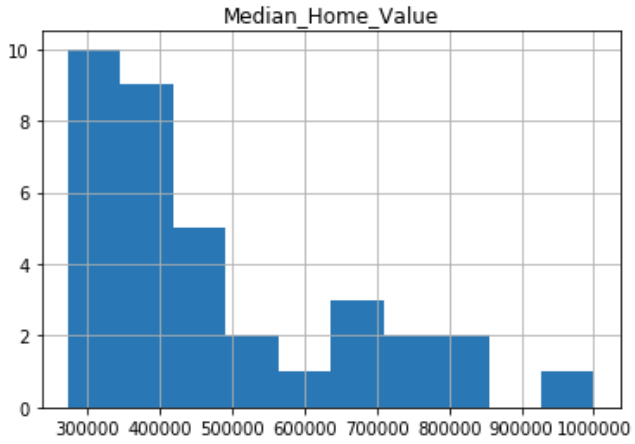
After linking each zip code with its associated data and parameters using the `uszipcode` package, I noted that there were some rows where there was no data available (i.e. data was populated as `NaN`). I removed these rows accordingly.

I also observed that some of the zip codes in Central Boston do not map to a unique neighborhood name, and defaults simply to “Boston”. This is a limitation of the `uszipcode` package. To bypass this issue, I subsequently labeled each zip code on the map by a combination of both its zip code and name to differentiate zip codes with the same name.

3 Exploratory Data Analysis

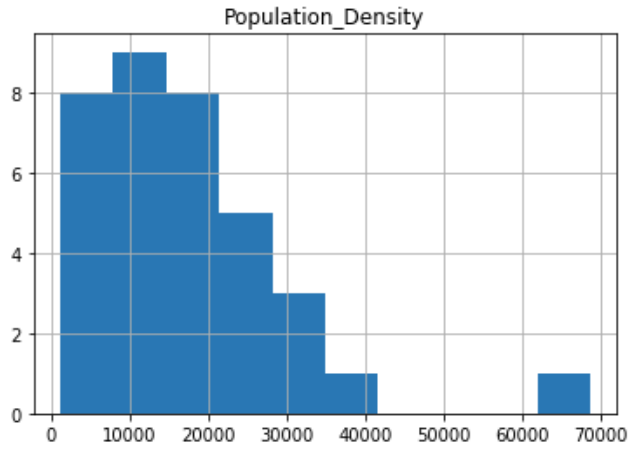
3.1 Distribution of Median Home Value across neighborhoods

The majority of neighborhood median home values are clustered on the lower end of the range of between $300k to around $500k.



3.2 Distribution of Population Density across neighborhoods

The population density of most neighborhoods are clustered around the lower end of range between 0 and 25,000.



3.3 Neighborhoods ranked by diversity of venues

Not surprisingly, the neighborhoods with the most diverse venues were in central Boston.

	ZIP5	Name	Venue Category
0	02199	Boston	63
1	02108	Boston	60
2	02116	Boston	58
3	02110	Boston	55
4	02111	Boston	49
5	02114	Boston	45
6	02118	Boston	44
7	02109	Boston	40
8	02113	Boston	38
9	02115	Boston	37

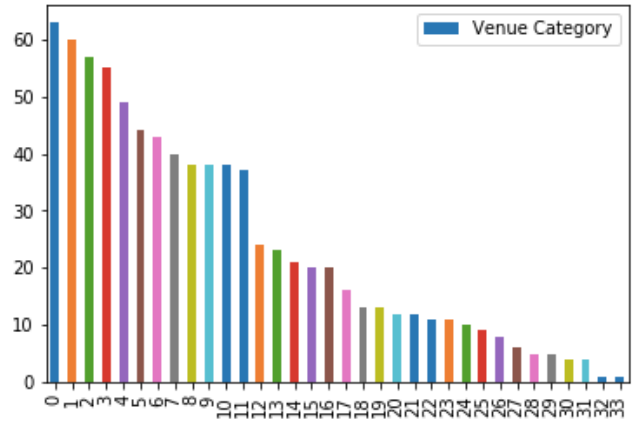
3.4 Neighborhoods ranked by total number of venues

Similarly, the top neighborhoods in terms of total absolute number of venues are also located in central Boston. Note that the number is capped at 100 (limit of each Foursquare query).

	ZIP5	Name	Venue
0	02116	Boston	100
1	02108	Boston	100
2	02109	Boston	100
3	02110	Boston	100
4	02111	Boston	100
5	02113	Boston	100
6	02199	Boston	100
7	02114	Boston	63
8	02118	Boston	59
9	02210	Boston	56

3.5 Distribution of unique venue categories across neighborhoods

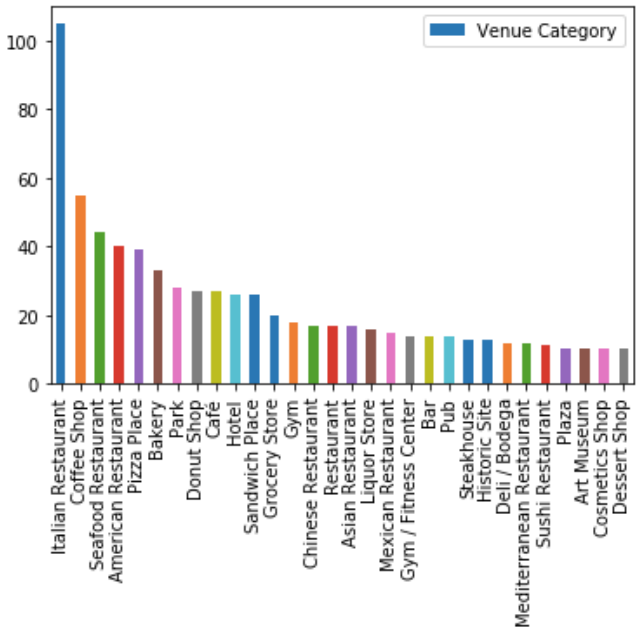
There is a wide spread of the number unique venue categories across the various neighborhoods.



3.6 Top venue categories

There are a total of 1,291 venues across all Boston neighborhoods, with 209 unique categories. A list of the top 10 categories is given below:

	Venue Category
Italian Restaurant	105
Coffee Shop	55
Seafood Restaurant	44
American Restaurant	40
Pizza Place	39
Bakery	33
Park	28
Donut Shop	27
Café	27
Hotel	26

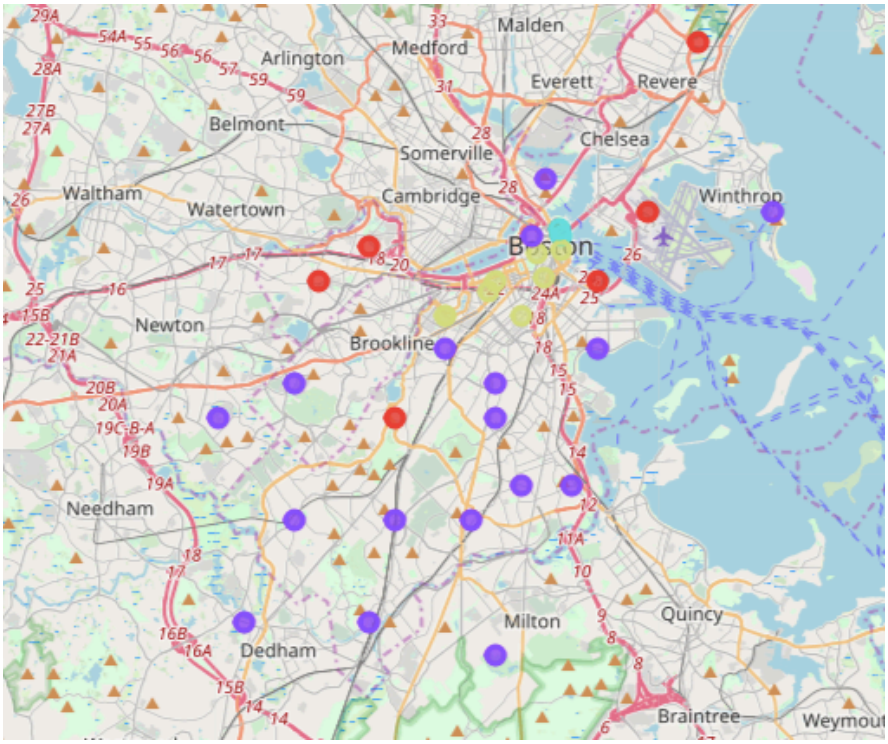


From the above analysis, Italian Restaurants are clearly the most popular category in Boston. In the subsequent clustering, in addition to the total venue categories, it would be important to also specifically define the top 10 categories. This would enable the machine learning algorithm to account also for the most popular and in-demand services.

4 Clustering

4.1 Use of k-means clustering algorithm

I used the k-means algorithm with k=4. The parameters included for clustering included Population Density, Median Home Value, number of Venue Categories and the mean number of the top ten venue categories per neighborhood. Normalization was carried out to ensure the algorithm weighted each parameter equally. This produced the following clusters:



5 Results

5.1 Use of clustering to group neighborhoods together

<u>Cluster 0 (Red) – Affordable with moderate diversity and availability of popular venues types</u>

	ZIP5	Cluster Labels	Italian Restaurant	Coffee Shop	Seafood Restaurant	American Restaurant	Pizza Place	Bakery	Park	Donut Shop	Café	Sandwich Place	Population_Density	Median_Home_Value	Venue Category	Name
18	02128	0	0.000000	0.121212	0.030303	0.090909	0.030303	0.000000	0.000000	0.181818	0.060606	0.030303	8352.0	290200.0	19	Boston
20	02130	0	0.000000	0.080000	0.040000	0.040000	0.040000	0.080000	0.000000	0.040000	0.000000	0.000000	10618.0	392800.0	22	Jamaica Plain
23	02134	0	0.000000	0.050000	0.000000	0.000000	0.050000	0.050000	0.000000	0.100000	0.000000	0.050000	16212.0	361900.0	17	Allston
24	02135	0	0.000000	0.062500	0.000000	0.031250	0.062500	0.093750	0.000000	0.031250	0.062500	0.031250	16236.0	350500.0	24	Brighton
26	02151	0	0.111111	0.055556	0.000000	0.000000	0.000000	0.111111	0.000000	0.055556	0.000000	0.055556	8833.0	306900.0	14	Revere
30	02210	0	0.035714	0.053571	0.107143	0.035714	0.000000	0.017857	0.017857	0.053571	0.017857	0.017857	2256.0	500000.0	36	Boston

Neighborhoods in this cluster seem to have wide representation across the popular venue types as well as have moderate diversity of venue categories. Prices and population density seem to have a wide variance. Geographically the neighborhoods seem to be clustered just outside the central zone.

Cluster 1 (purple) – Generally poor diversity and non-availability of popular venues

	ZIP5	Cluster Labels	Italian Restaurant	Coffee Shop	Seafood Restaurant	American Restaurant	Pizza Place	Bakery	Park	Donut Shop	Café	Sandwich Place	Population_Density	Median_Home_Value	Venue Category	Name
0	02021	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1147.0	425800.0	1	Canton
1	02026	1	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	2412.0	373700.0	9	Dedham
7	02114	1	0.047619	0.015873	0.000000	0.031746	0.079365	0.000000	0.000000	0.031746	0.015873	0.015873	26694.0	479500.0	45	Boston
11	02119	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	15871.0	327900.0	11	Roxbury
12	02120	1	0.041667	0.000000	0.000000	0.000000	0.208333	0.000000	0.041667	0.125000	0.000000	0.041667	24456.0	272900.0	18	Roxbury Crossing
13	02121	1	0.000000	0.000000	0.000000	0.000000	0.071429	0.000000	0.000000	0.000000	0.000000	0.000000	14884.0	311700.0	13	Dorchester
14	02122	1	0.000000	0.000000	0.000000	0.066667	0.066667	0.000000	0.000000	0.066667	0.000000	0.066667	11554.0	320900.0	14	Dorchester
15	02124	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.100000	15913.0	306500.0	10	Dorchester Center
16	02126	1	0.000000	0.000000	0.000000	0.000000	0.066667	0.066667	0.066667	0.066667	0.000000	0.000000	12277.0	282700.0	15	Mattapan
17	02127	1	0.166667	0.000000	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	15744.0	398200.0	8	Boston
19	02129	1	0.000000	0.000000	0.000000	0.000000	0.095238	0.000000	0.047619	0.000000	0.047619	0.000000	12192.0	489900.0	18	Charlestown
21	02131	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	11505.0	355900.0	4	Roslindale
22	02132	1	0.000000	0.000000	0.000000	0.000000	0.125000	0.000000	0.062500	0.000000	0.000000	0.000000	5670.0	393100.0	12	West Roxbury
25	02136	1	0.000000	0.000000	0.000000	0.142857	0.000000	0.000000	0.142857	0.000000	0.000000	0.000000	6207.0	313400.0	7	Hyde Park
27	02152	1	0.181818	0.000000	0.000000	0.000000	0.090909	0.000000	0.000000	0.000000	0.000000	0.000000	8871.0	360200.0	10	Winthrop
28	02186	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2075.0	484000.0	1	Milton
32	02459	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.250000	0.000000	0.000000	3815.0	803000.0	4	Newton Center

Neighborhoods in this cluster appear to have low diversity with a few outliers, and mostly do not have wide representation of the popular venue types.

Cluster 2 (Cyan) – High end with high diversity and available popular venues

	ZIP5	Cluster Labels	Italian Restaurant	Coffee Shop	Seafood Restaurant	American Restaurant	Pizza Place	Bakery	Park	Donut Shop	Café	Sandwich Place	Population_Density	Median_Home_Value	Venue Category	Name
3	02109	2	0.34	0.02	0.07	0.01	0.04	0.07	0.03	0.0	0.03	0.02	21721.0	676700.0	40	Boston
6	02113	2	0.40	0.03	0.04	0.01	0.05	0.06	0.05	0.0	0.03	0.02	68665.0	449100.0	38	Boston

There are only 2 neighborhoods in this cluster and they appear to be very similar in all respects except for population density and median home value. Geographically they appear to be very close together, which should explain the similarities as it is likely that the same venues would overlap across both neighborhoods.

Cluster 3 (Yellow) – High end with high diversity and available popular venues

	ZIP5	Cluster Labels	Italian Restaurant	Coffee Shop	Seafood Restaurant	American Restaurant	Pizza Place	Bakery	Park	Donut Shop	Café	Sandwich Place	Population_Density	Median_Home_Value	Venue Category	Name
2	02108	3	0.040000	0.040000	0.020000	0.050000	0.030000	0.000000	0.010000	0.000000	0.000000	0.040000	27919.0	724500.0	60	Boston
4	02110	3	0.020000	0.080000	0.100000	0.020000	0.010000	0.020000	0.050000	0.010000	0.010000	0.020000	9355.0	724000.0	55	Boston
5	02111	3	0.020000	0.040000	0.020000	0.010000	0.020000	0.060000	0.000000	0.000000	0.020000	0.030000	28542.0	691500.0	49	Boston
8	02115	3	0.019231	0.076923	0.019231	0.038462	0.000000	0.000000	0.019231	0.038462	0.076923	0.019231	40159.0	611800.0	37	Boston
9	02116	3	0.040000	0.050000	0.040000	0.050000	0.000000	0.000000	0.010000	0.000000	0.010000	0.010000	32724.0	807600.0	58	Boston
10	02118	3	0.050847	0.033898	0.016949	0.067797	0.016949	0.016949	0.033898	0.000000	0.016949	0.000000	24075.0	561100.0	44	Boston
29	02199	3	0.020000	0.050000	0.030000	0.050000	0.000000	0.010000	0.010000	0.000000	0.000000	0.010000	19927.0	1000001.0	63	Boston
31	02215	3	0.019231	0.076923	0.019231	0.038462	0.000000	0.000000	0.019231	0.038462	0.076923	0.019231	34190.0	366700.0	37	Boston

The neighborhoods in this cluster is very similar to Cluster 2.

6 Discussion

6.1 Intuition of results

It is interesting to note how the clusters have formed with only a few parameters to reflect a split across neighborhoods that centrally located, slightly outside the city center, and those on the outskirts of the city. This should be what we would intuitively have expected: neighborhoods more centrally located should be denser, more expensive as well as have more diverse and popular venues situated nearby.

6.2 Possible dilutive effect of venue types

However, what was perhaps a little disappointing was the lack of a clear pattern in the population density and median home value within the clusters. There were some outliers within some of the clusters above and this seemed to suggest that having the popular venue columns in the data might have diluted the impact of the other parameters (namely, population density, median home value, and venue categories). It would be interesting to investigate how the results would differ if only the above 3 parameters were clustered again.

6.3 Consider adding more parameters

In revisiting the original premise for this project, this simple project provided a clustering of similar neighborhoods across a limited number of parameters. It might be interesting to consider other parameters that may influence the choice of a neighborhood to live in, such as availability of transportation options, schools, libraries etc.

7 Conclusion

In this project I looked at developing a simple tool to help determine which neighborhoods within Boston were most attractive to live in given just its population density, median home value, diversity of venues and availability of the most popular venues. Through the use of the k-means method, an intuitive clustering of the neighborhoods was obtained. However, there was a lack of a clear pattern among some of the key parameters (population density and median home value) which may suggest that undue weight might have been placed on the data for the popular venue categories. This could be a future area to look into to refine the tool.