Neighborhoods Analysis: Moving in Boston

Lingyi Cai

May 9th, 2020

I. Introduction

1.1 Background

I moved to Boston, Massachusetts in June 2019 for the relocation of my job. Before moving to Boston, I did some researches online to find an apartment in Boston. Since I do not have a car or bike, I need to either live close to my company or close to public transportation. I made a list of potential apartments I may move in. On the list, there were ten apartments and I listed the rent, pros and cons of each apartment. At the end, I choose the newest apartment(first open for rent in May 2019) on my list which is also close to the place I work. However, after I moved to Boston and lived in that apartment, I realized the location of the apartment is not convenient for me. There are no restaurants, public transportation or entertainment services around the place I lived. Although there's a bus stop(10 minutes walk) close to the apartment, the path to the bus station is currently closed for construction. I have to Uber to work everyday which became one of the biggest monthly expenses for me.

1.2 Problem

I am not satisfied with the location of my apartment and I spent quite a few money on transportation monthly, I need to find a neighborhood that can fulfill my needs as soon as possible. Therefore, I am going to analyze Boston neighborhoods using Foursquare API and find out an area/neighborhood that satisfied my needs.

1.3 Interest

After talking to my friends and coworkers in Boston, I realized I am not the only one who is looking for a new neighborhood to move in. I think this report could help them as well and give them some ideas of which neighborhoods they can move in to.

II. Data acquisition and cleaning

2.1 Data sources

Most details information including longitude, latitude, street name, city, district and zip code can be found in Kaggle datasets here. However, this dataset is very large in size. It contains a lot of duplicate information as well as many other detail information I do not need in this analysis.

2.2 Data cleaning

Before cleaning the data, there were 3,507,168 samples and 11 features in the data. This dataset contains all fourteen Massachusetts counties detail location information. I decided I would only use Middlesex county, Suffolk county and Norfolk county in this analysis because those are the three counties located at the center of the Massachusetts state as well as areas where the majority of people who work in Boston would live in.

The dataset contains many other columns of information I do not need, so I drop five columns that I am not going to use in this project. Moreover, with the same zip code, there are many different locations are recorded in the dataset. Hence, I removed all the rows with the same zip code and only left one row for each zip code. Then, I calculate the missing value rows and found out the missing value rows only occupied a very small portion of the dataset. Thus, I removed all the missing value rows from the dataset.

2.3 Feature selection

After data cleaning, there were 188 samples and 6 features in the data (Table 1).

LON - Longitude

LAT - Latitude

Table 1. First ten rows of the dataset

	LON	LAT	STREET	CITY	DISTRICT	POSTCODE
0	-71.376402	41.986877	FALES ROAD	PLAINVILLE	NORFOLK	2762
1	-71.494588	42.017492	CALIFORNIA AVENUE	BELLINGHAM	NORFOLK	2019
2	-71.457583	42.017947	JENKS STREET	WRENTHAM	NORFOLK	2093
3	-71.288031	42.017653	CEDAR STREET	FOXBOROUGH	NORFOLK	2035
4	-71.442234	42.037219	VAIL DRIVE	FRANKLIN	NORFOLK	2038
5	-71.177637	42.057936	MANSFIELD STREET	SHARON	NORFOLK	2067
6	-71.137240	42.074167	CLEARY DRIVE	STOUGHTON	NORFOLK	2072
7	-71.288320	42.085192	LAFAYETTE LANE	NORFOLK	NORFOLK	2056
8	-71.280177	42.089123	COBBLE KNOLL DRIVE	WALPOLE	NORFOLK	2071
9	-71.279831	42.101279	YONKER PLACE	WALPOLE	NORFOLK	2081

Upon examining the meaning of each feature, it was clear that some of the features were not needed in the report as well as map visualization. The features are dropped including “ID”, “Number”, “Unit”, “Hash” and “Region”. Features I kept including “LON"(longitude), “LAT"(latitude), “Street”, “City”, “District” and “Post code”. The reason I kept “City” and “Street” is that these information is needed when using Foursquare API to explore the neighborhood. Moreover, it would be easier for the audience to read.

Table 2. Simple features selection during data cleaning

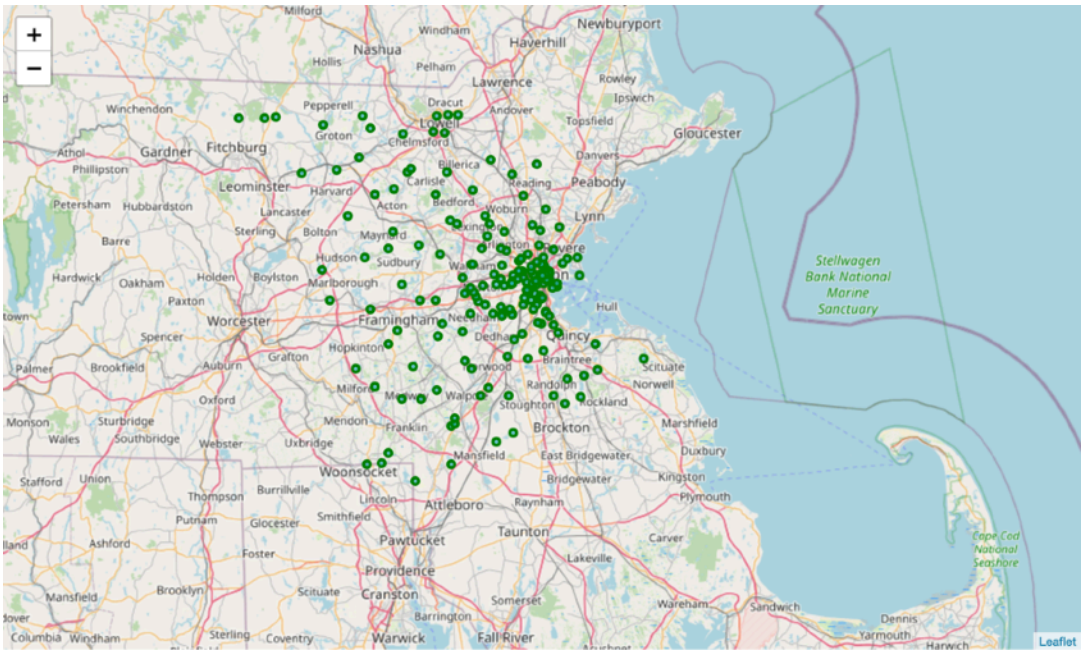
Kept features	Dropped features	Reason for dropping features
LON, LAT, Street, City, District, Postcode	ID, number, unit, hash, region	Dropped features are not needed for this report and will not be used in map visualization

III. Exploratory Data Analysis

3.1 Methodology

As a database, I use GitHub repository and Jupyter nbviewer in my study. My data contains the features Longitude, Latitude, street, city, district and post code of Middlesex county, Suffolk county and Norfolk county.

I used Python Folium library to create a map of Middlesex, Suffolk and Norfolk with neighborhoods superimposed on top.



I utilized the Foursquare API to explore the neighborhoods. I set the limit at 100 venues and the radius 500 meters for each street from their given latitude and longitude information. Here is summary dataset of Boston venues which contains 151 unique street names and 305 unique venue categories. This graph is organized from a dataset of 3368 samples.

Table 3.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
A STREET	82	82	82	82	82	82
ABBOTSFORD STREET	9	9	9	9	9	9
ABERDEEN STREET	39	39	39	39	39	39
ACADIA STREET	10	10	10	10	10	10
ACORN STREET	39	39	39	39	39	39
...	...	...	...	...	...	...
WINSLOW ROAD	2	2	2	2	2	2
WOOD LANE	2	2	2	2	2	2
WOODLAND ROAD	2	2	2	2	2	2
WORCESTER ROAD	6	6	6	6	6	6
YONKER PLACE	1	1	1	1	1	1

We can see street like A street returns 82 venues from Foursquare while streets like Winslow road, Wood lane, Woodland road and Yonker place merely return less than three venues in the given coordinates.

The result does not mean that Foursquare return all the possible venues from the street or neighborhood. The results are the consequences of given latitude and longitude of certain locations in the dataset. Venues and possibilities can always be increased/decreased when more latitude and longitude information are given by information collected. Moreover, we can also change the limit and radius variables in Python which may increase/decrease the results returned by Foursquare.

In summary of this data, 305 venues were returned by Foursquare. Below is summary of a merged table of neighborhoods and venues. The numbers shown are the mean of the frequency of occurrence of each venue category.

Table 4.

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Antique Shop	Aquarium	...	Udon Restaurant	Vegetarian / Vegan Restaurant	Video Game Store	V S
0	A STREET	0.000000	0.0	0.0	0.0	0.0	0.0	0.012195	0.0	0.0	...	0.000000	0.012195	0.0	
1	ABBOTSFORD STREET	0.000000	0.0	0.0	0.0	0.0	0.0	0.111111	0.0	0.0	...	0.000000	0.000000	0.0	
2	ABERDEEN STREET	0.025641	0.0	0.0	0.0	0.0	0.0	0.051282	0.0	0.0	...	0.025641	0.000000	0.0	
3	ACADIA STREET	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	
4	ACORN STREET	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
146	WINSLOW ROAD	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	
147	WOOD LANE	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	
148	WOODLAND ROAD	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	
149	WORCESTER ROAD	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	
150	YONKER PLACE	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.000000	0.000000	0.0	

Then, I created a new dataset which displays top ten venues generated by Foursquare in each neighborhood as table shown below. The table only shows the first five rows of the dataset.

Table 5.

	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	A STREET	Coffee Shop	Hotel	Juice Bar	Italian Restaurant	Bar	Seafood Restaurant	Art Gallery	Clothing Store	New American Restaurant	Asian Restaurant
1	ABBOTSFORD STREET	Monument / Landmark	Plaza	Record Shop	Museum	American Restaurant	Pizza Place	Fast Food Restaurant	Pharmacy	Garden	Exhibit
2	ABERDEEN STREET	Lounge	American Restaurant	Chinese Restaurant	Bakery	Burger Joint	Yoga Studio	Mexican Restaurant	Beer Garden	Big Box Store	Furniture / Home Store
3	ACADIA STREET	Park	Pizza Place	Breakfast Spot	Café	Dive Bar	Diner	Construction & Landscaping	Ice Cream Shop	Convenience Store	Cupcake Shop
4	ACORN STREET	Gourmet Shop	French Restaurant	Pizza Place	Gift Shop	Italian Restaurant	Breakfast Spot	Theater	Kids Store	Bakery	Lake

From the table above, we can see Foursquare return the specific venues from each neighborhood. Venues including hotel, restaurants, bakery, museum and etc. Those information are very helpful for me to analyze and cluster neighborhoods into certain group.

We have some common venue categories in neighborhoods. In order to cluster the neighborhood, I used K-means algorithm. To find the optimal K-mean for my report as well as for my audience, I ran the K-means value from 1 to 10 and compared the outcomes of each K clusters value. At the end, I decided to set my K cluster to value 5 which I believe would fit my report the best.

I merged the cluster labels with my dataset and below is the first five rows of the merged table.

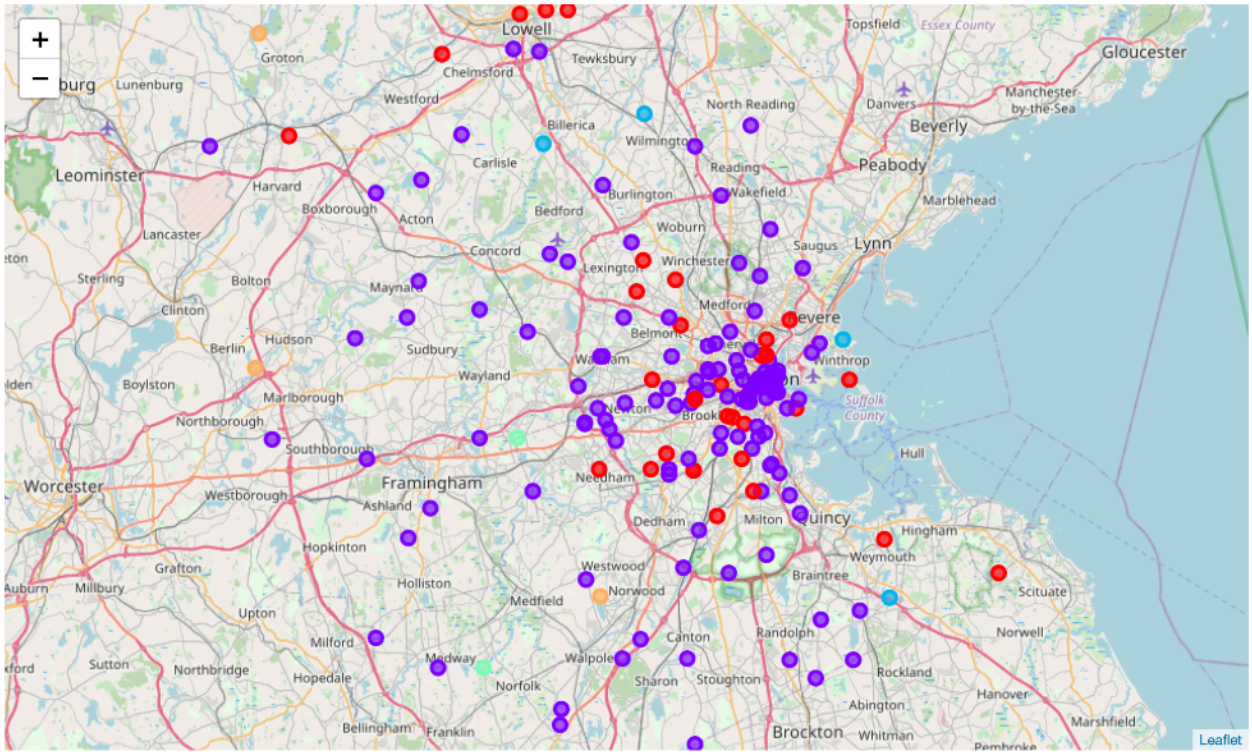
Table 6.

	LON	LAT	Neighborhood	CITY	DISTRICT	POSTCODE	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Con V
1	-71.494588	42.017492	CALIFORNIA AVENUE	BELLINGHAM	NORFOLK	2019	1	Liquor Store	American Restaurant	Pizza Place	Donut Shop	Chinese Restaurant	Conven
3	-71.288031	42.017653	CEDAR STREET	FOXBOROUGH	NORFOLK	2035	0	Spa	Pharmacy	Café	Liquor Store	Sandwich Place	Ch Resta
5	-71.177637	42.057936	MANSFIELD STREET	SHARON	NORFOLK	2067	2	Construction & Landscaping	Exhibit	Women's Store	Falafel Restaurant	Electronics Store	Empa Resta
6	-71.137240	42.074167	CLEARY DRIVE	STOUGHTON	NORFOLK	2072	1	Gym	Women's Store	Exhibit	Dumpling Restaurant	Electronics Store	Empa Resta
8	-71.280177	42.089123	COBBLE KNOLL DRIVE	WALPOLE	NORFOLK	2071	1	Gym	Gym / Fitness Center	Exhibit	Donut Shop	Dumpling Restaurant	Electr

From the table above, a new column “Cluster Labels” is added to the table. Since I set K cluster to five which means the 151 neighborhoods will be divided into five clusters/groups(cluster 0, cluster 1, cluster 2, cluster 3, cluster 4) based on their similarity.

IV. Results

The map below shown the resulting clusters in five different colors dots.



We can then examine each five c lusters by printing out the table.

Table 7. Cluster I

	LAT	POSTCODE	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
3	42.017653	2035	0	Spa	Pharmacy	Café	Liquor Store	Sandwich Place	Chinese Restaurant	Park	Transportation Service	Constructi Landscapi
25	42.208523	2025	0	Café	Baseball Field	Fabric Shop	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Exhi
28	42.234212	2191	0	Park	Exhibit	Donut Shop	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Spa
33	42.289019	2494	0	Park	Sandwich Place	Supermarket	Optical Shop	Gym / Fitness Center	Burrito Place	Donut Shop	Spa	Ice Crea Sh
35	42.406211	2150	0	Pizza Place	Yoga Studio	Fast Food Restaurant	Park	Donut Shop	Grocery Store	Convenience Store	Food	Intersecti

The table above showing the first five rows of cluster one. Cluster one contains more than 30 neighborhoods. From the top ten most common venues we can identify what are around those neighborhoods. For example, the major neighborhoods in cluster one contain Park,

Entertainment service, Spa and Gyms within 500 meters of the given latitude and longitude location.

Table 8. Cluster II

	LAT	POSTCODE	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 Most Common Venue	10th Com V
1	42.017492	2019	1	Liquor Store	American Restaurant	Pizza Place	Donut Shop	Chinese Restaurant	Convenience Store	Bar	Thai Restaurant	Pharmacy	Sanc
6	42.074167	2072	1	Gym	Women's Store	Exhibit	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	F
8	42.089123	2071	1	Gym	Gym / Fitness Center	Exhibit	Donut Shop	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	S
9	42.101279	2081	1	Concert Hall	Women's Store	Falafel Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Exhibit	F
10	42.126208	2343	1	Pharmacy	Liquor Store	Baseball Field	Sandwich Place	Pizza Place	Chinese Restaurant	Gas Station	Bank	Ethiopian Restaurant	Dun Resta
...	...	...	...	...	...	...	...	...	...	...	...	...	
180	42.360117	2201	1	Italian Restaurant	Sandwich Place	Historic Site	Bakery	Coffee Shop	American Restaurant	Park	Pub	Seafood Restaurant	I
181	42.348167	02199	1	Hotel	Clothing Store	Seafood Restaurant	Italian Restaurant	Coffee Shop	Plaza	Ice Cream Shop	Women's Store	American Restaurant	
182	42.367341	02163	1	Park	Gym	College Stadium	Pool	College Hockey Rink	Residential Building (Apartment / Condo)	Soccer Stadium	Squash Court	Tennis Court	Athle S
184	42.383251	2145	1	Mexican Restaurant	Bus Stop	Donut Shop	Liquor Store	Dog Run	Coffee Shop	Asian Restaurant	Arts & Entertainment	Metro Station	S
187	42.358310	2133	1	Coffee Shop	American Restaurant	Historic Site	Italian Restaurant	Steakhouse	Sandwich Place	Restaurant	Cocktail Bar	New American Restaurant	B

Cluster two contains 125 neighborhoods. The most common venues within those neighborhoods are gyms, restaurants, exhibit, stores an entertainment services.

Table 9. Cluster III

	LAT	POSTCODE	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 Most Common Venue	C
5	42.057936	2067	2	Construction & Landscaping	Exhibit	Women's Store	Falafel Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	
22	42.188598	2189	2	Construction & Landscaping	Clothing Store	Brewery	Optical Shop	Women's Store	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	
99	42.544018	1862	2	Construction & Landscaping	Women's Store	Falafel Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Exhibit	
105	42.567261	1876	2	Beach	Construction & Landscaping	Lake	Women's Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Exhibit	
186	42.390782	2151	2	Beach	Construction & Landscaping	Pool	Women's Store	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	

The third cluster includes five neighborhoods. The most common venues for this group including construction & landscaping, beach and restaurants.

Table 10. Cluster IV

	LAT	POSTCODE	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 Most Common Venue	10th Most Common Venue
11	42.134071	2054	3	Skating Rink	Women's Store	Exhibit	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Fabric Shop
51	42.314319	2493	3	Skating Rink	Women's Store	Exhibit	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Fabric Shop

The forth cluster contains two neighborhoods and the most common venues in these two neighborhoods are skating rink, women's store and exhibit.

Table 11. Cluster V.

	LAT	POSTCODE	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 Most Common Venue	10t Mos Commo Venu
21	42.189396	2090	4	Beach	Flower Shop	Falafel Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Exhibit	Fabri Sho
65	42.368337	1749	4	Flower Shop	Women's Store	Fabric Shop	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Event Space	Exhib
111	42.630206	1463	4	Flower Shop	Hot Dog Joint	Women's Store	Exhibit	Dumpling Restaurant	Electronics Store	Empanada Restaurant	Entertainment Service	Ethiopian Restaurant	Ever Spac

The fifth cluster maintains three neighborhoods and the most common venues for those areas are flower shop, restaurants and stores.

V. Discussion

The original dataset contains all fourteen counties' geographic information within Massachusetts state. Before cleaning the data, there were 3,507,168 samples and 11 features in the dataset. However, I narrowed the location down to three counties from fourteen counties. The analysis outcome can vary base on which locations(counties) were chosen as well as when utilizing different approaches in clustering and classification studies.

When deciding the valued of K cluster in my study, I merely run the K cluster variable with different values from 1 to 10 and then printed out and counts the sum of different unique values

within the ten arrays. The reason I chose to divide the dataset into five cluster is that the array was relatively equally divided to five different categories comparing to other arrays. Moreover, I believe when dividing the dataset into too many clusters, it is not very easy for the audience and myself when reading the results.

I ended the study by visualizing the data and clustering information on the Boston map. IN future learning and studies of Data Science, I will consider also adding housing/renting price as well as house/apartment information including size, number of bedroom and number of bathroom into the report.

VI. Conclusion

As a result, the demands of looking for a new neighborhood to move in due to job relocation, budget consideration or other reasons are always high. Some people prefer to go to Real Estate agencies for assistance. As more and more information is available online, many will do research themselves when seeking a new house/apartment. However, there is a large amount of information online and it is not easy to filter the information by merely looking over information posted on different website. By doing this study utilizing Foursquare API and dataset I found online, I have a better understanding of Boston neighborhoods as well as the impacts of Data Science in every aspect of today's society.

VII. Reference

- Kaggle: https://www.kaggle.com/openaddresses/openaddresses-us-northeast#ma.csv
- Foursquare API : https://developer.foursquare.com/