Opening a Sushi Restaurant in Boston

Choosing a neighborhood to open a Sushi Restaurant in the US's University Capital:

1 – Introduction

The following report will be created with the idea of assisting an entrepreneur seeking to open a Sushi restaurant in the city of Boston, Massachusetts. The business problem, though relatively simple, begs a very fundamental question: *in which neighborhood shall the entrepreneur open his new business?*

Boston is a very diverse city, as its various neighborhoods differ greatly amongst each other when it comes to a wide variety of demographic factors (i.e., race, age, gender, education level). As such, the composition of the businesses that compose each neighborhood varies greatly – some are home to a wide variety of restaurants, while others offer more residential appeals. Some neighborhoods may be over-saturated with Japanese eateries, while others may have a general scarcity of restaurants, indicating a smaller market for restauranteurs. As such, we will attempt to determine which neighborhoods are optimal for the opening of a Sushi Restaurant based on lack of direct competition (other Sushi Restaurants) *and* the presence of a solid dining landscape. To accomplish this goal, we will attempt to identify the neighborhoods that display *both* a low density of Sushi Restaurants, *and* an average density of general dining venues.

2 – Data

The data we will use to solve this problem will come from a variety of different sources. First, we will need to identify **all Boston neighborhoods.** This task we will be able to accomplish by simply grabbing each of their general demographics as outlined by the **Boston Archive**. The **Boston Archive** provides a free database of its neighborhoods' demographic data (including their zip codes), from which we will extract all their identifying labels, as they appear according to the city's officials[1].

Once we have extracted the official denominations of each of the city's neighborhoods, we will require a database of **longitudinal and longitudinal coordinates to match each of the identified neighborhoods**. To achieve this, we will extract a dataset which relates each US Zip Code to its latitudinal and longitudinal coordinates. This database, titled **"US Zip Codes Database"** is compiled by **Simple Maps,** using data from the **IRS, The United States Census, and the American Community Survey.[2]**

Once we have retrieved the coordinates of each of the city's neighborhoods, we will be able to leverage the **Foursquare API** in order to retrieve data regarding the venues that are present within a given radius of the coordinates passed. To achieve this, we will require a Foursquare Developer account capable of performing the **Explore function** (Free accounts will work fine for our

[1] Boston Archive: http://archive.boston.com/news/local/articles/2007/04/15/sixfigurezipcodes_city/
[2] Simple Maps Archive: https://simplemaps.com/data/us-zips

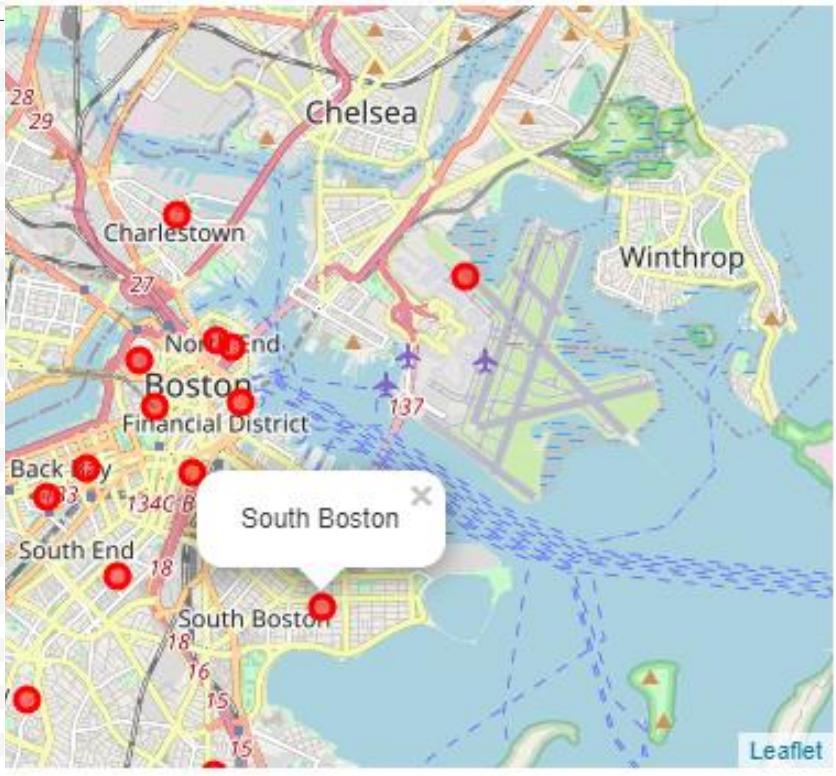
purposes). The data we will extract using the **Foursquare API** will contain the categories, names, and precise coordinates of a significant sample (N = 100) of venues for every neighborhood.

Having these three datasets at our disposal, we should be all set to determine which of the neighborhoods identified will have the most saturated market for Sushi Restaurants, while offering a competitive and appealing dining landscape.

3 – Methodology

Firstly, we will extract the relevant data from the two databases we will choose to pull from. To begin with, we will retrieve every single Boston Neighborhood denomination, their zip code, and their demographics from the **"Boston Archive"** database. At the same time, we will extract every U.S. Zip Code and its coordinates from the "**U.S. Zip Codes"** database. Once both have been extracted, we will associate each Boston Neighborhood / Community with its Latitude and Longitude. Once this is achieved, we can progress to overlay each of the coordinates we have retrieved on a map of Boston, to better understand their geographical disposition. This will be achieved using the "Folium" Python package – and its output will look as reported below (in this case with the South Boston neighborhood highlighted.

	Neighborhood	Zip Code	Latitude	Longitude
0	Beacon Hill	2108	42.35767	-71.06505
1	Markets / Inner Harbor	2109	42.36459	-71.05298
2	Financial District / Wharves	2110	42.35826	-71.05175
3	Chinatown / Tufts-New England Medical Center	2111	42.35007	-71.05911
4	North End	2113	42.36542	-71.05540
5	West End / Back of the Hill	2114	42.36316	-71.06732
6	Fenway / East Fens / Longwood	2115	42.34109	-71.09456
7	Back Bay	2116	42.35062	-71.07564
8	South End	2118	42.33820	-71.07074
9	Roxbury	2119	42.32410	-71.08483
10	Roxbury Crossing	2120	42.33210	-71.09637
11	Roxbury / Grove Hall	2121	42.30589	-71.08583
12	Dorchester / Fields Corner	2122	42.30007	-71.03352
13	Dorchester / Codman Square	2124	42.28572	-71.07105
14	Dorchester / Uphams Corner	2125	42.31545	-71.05594
15	Mattapan	2126	42.27430	-71.09388
16	South Boston	2127	42.33479	-71.03902
17	East Boston	2128	42.37266	-71.01656
18	Charlestown	2129	42.37975	-71.06169
19	Jamaica Plain	2130	42.30986	-71.11486
20	Roslindale	2131	42.28470	-71.12652
21	West Roxbury	2132	42.28025	-71.16145
22	Allston	2134	42.35788	-71.12942
23	Brighton	2135	42.34981	-71.15386
24	Hyde Park	2136	42.25535	-71.12937
25	Cambridge	2138	42.38004	-71.13464
30	Somerville	2143	42.38145	-71.09684
33	North Brighton / Cambridge	2163	42.36636	-71.12076
34	Prudential	2199	42.34747	-71.08202
35	South Boston / Fort Point	2210	42.34781	-71.03905
36	Kenmore / Boston University	2215	42.34778	-71.10287
37	Brookline	2445	42.32599	-71.13396



Having a cohesive relationship between **Neighborhoods** and **Coordinates** is fundamental for the next step, and a correct map like that above is a crucial sanity-check that will carry us through the remainder of the report. Once we have verified the correct geo-localization of each neighborhood of interest, we can leverage the **Foursquare API.**

To do this, we will **require a Foursquare Developer Account**, complete with CLIENT_ID, CLIENT_SECRET and ACCESS_TOKEN. Once this has been properly set-up, we can create an API Request, which will attempt to extract the most popular Foursquare locations around the set of coordinates we have just extracted. Here, we must note some **Foursquare** lingo – we will be performing an **Explore** request, which will return the following set of parameters for each of the coordinates passed through the function: **Venue Name, Venue Category, Venue Latitude, Venue Longitude.** Based on the latitude and longitude data we passed through, we will be able to automatically compile a data-frame that places each returned Venue within its appropriate neighborhood. The sample output will look as described below, in this instance showing the first 5 venues returned in the "**Beacon Hill**" neighborhood:

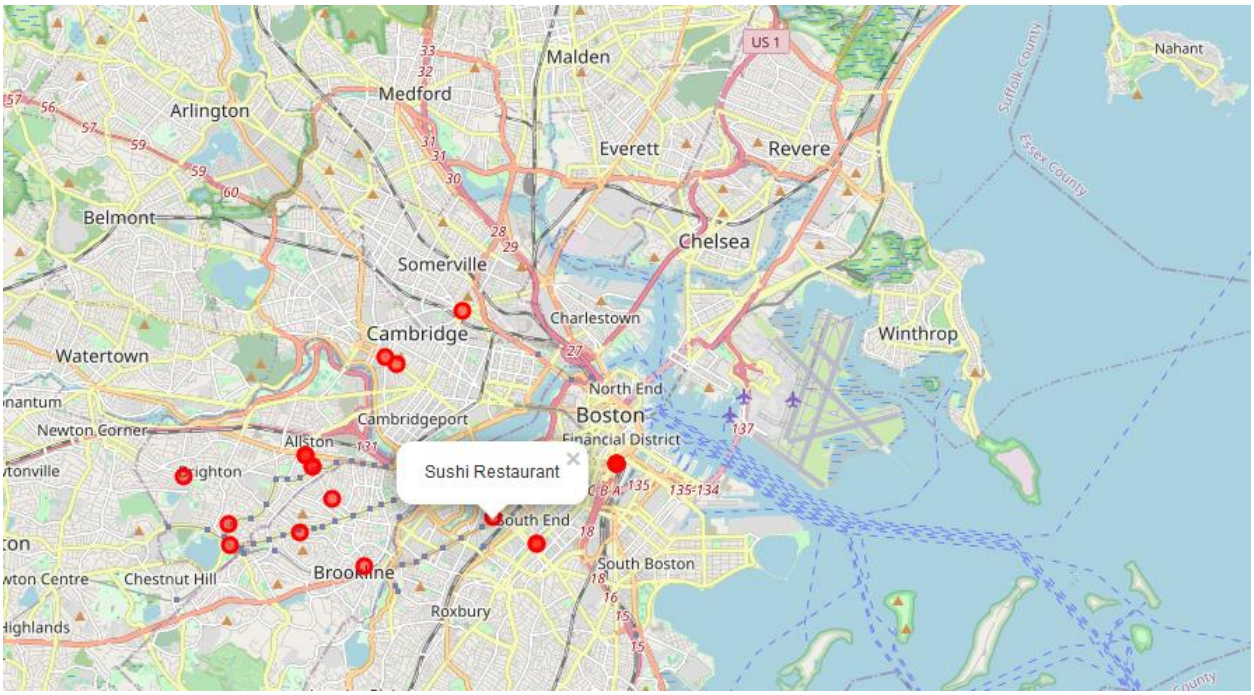
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Beacon Hill	42.35767	-71.06505	Boston Common	42.355487	-71.064882	Park
1	Beacon Hill	42.35767	-71.06505	Boston Athenaeum	42.357481	-71.061838	Library
2	Beacon Hill	42.35767	-71.06505	Yvonne's	42.355664	-71.061466	New American Restaurant
3	Beacon Hill	42.35767	-71.06505	Tatte Bakery & Cafe	42.357904	-71.070439	Bakery
4	Beacon Hill	42.35767	-71.06505	Sam LaGrassa's	42.356870	-71.059960	Sandwich Place

On this pass, the **Foursquare API** returned N = 100 venues per neighborhood, a **sample size** that we can be at least somewhat satisfied with. Within this pass as well, the venues returned spanned a total of N = **187 unique categories**.

One-Hot Encoding: we can now perform a one-hot encoding operation on the dataset at hand, to understand which kinds of venues are most popular within each neighborhood. To do this, we will transpose the dataset above, and represent each venue with a **simple 1**. Once this process is complete, we can retrieve the mean number of venues in each category for every neighborhood and divide it by the total number of venues returned to return the **Frequency of Each Venue Category in each Neighborhood.** Given the frequency of each venue category per neighborhood, we can construct a table that determines the five most popular venue categories in every neighborhood in our dataset. Performing these functions returned the following:

	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	Pizza Place	Bakery	Grocery Store	Rock Club	Fried Chicken Joint	Bar	Seafood Restaurant	Sushi Restaurant	Gym	Mexican Restaurant
1	Back Bay	Park	Hotel	Italian Restaurant	Seafood Restaurant	Clothing Store	American Restaurant	Bakery	French Restaurant	Coffee Shop	Gym
2	Beacon Hill	Park	Coffee Shop	Italian Restaurant	Seafood Restaurant	Bakery	Hotel	Spa	Gourmet Shop	Gym / Fitness Center	Historic Site
3	Brighton	Pizza Place	Bar	Coffee Shop	Gym	Sushi Restaurant	Bakery	Café	Grocery Store	Rock Club	Shoe Store
4	Brookline	Park	Coffee Shop	Pizza Place	Café	Gym	Mexican Restaurant	Hotel	Sushi Restaurant	Italian Restaurant	Liquor Store

Isolating Sushi Restaurants: once we have a better idea of what kind of venues are most popular in every neighborhood, we can zoom into the dataset further by examining Sushi Restaurants alone. An interesting first step in this exploration is to list the **Frequency of Sushi Restaurants in each neighborhood analyzed.** This will provide us an easily read snapshot of how saturated the Sushi Restaurant Market might be in each neighborhood. An easy visual output of this phase is a map of each retrieved Sushi Venue, overlaid on a map of Boston.



Comparing to Total Restaurants: once we have data on the frequency of Sushi Neighborhoods alone, it is interesting to compare the frequency of Sushi Restaurants to the neighborhood's general restaurant frequency. **This will be fundamental in answering our question, as we are attempting to identify a neighborhood where Sushi Restaurants aren't frequent, but where the general populace is prone to visit Dining Venues in general.** We can therefore generate a novel dataset that lists the frequency of Venues in each neighborhood that contain the word "restaurant" in their category, as well as the frequency of "Sushi Restaurant".

	Neighborhood	Restaurant Frequency	Sushi Restaurant Frequency
0	Allston	0.280000	0.03
1	Back Bay	0.220000	0.01
2	Beacon Hill	0.190000	0.01
3	Brighton	0.270000	0.04
4	Brookline	0.230000	0.03

...

K-Nearest-Neighbor Clustering: once we have this dataset, we can perform our final bit of modeling, where we will train a model to group the various neighborhoods, we have explored based on chosen data. We will attempt this in **two ways:**

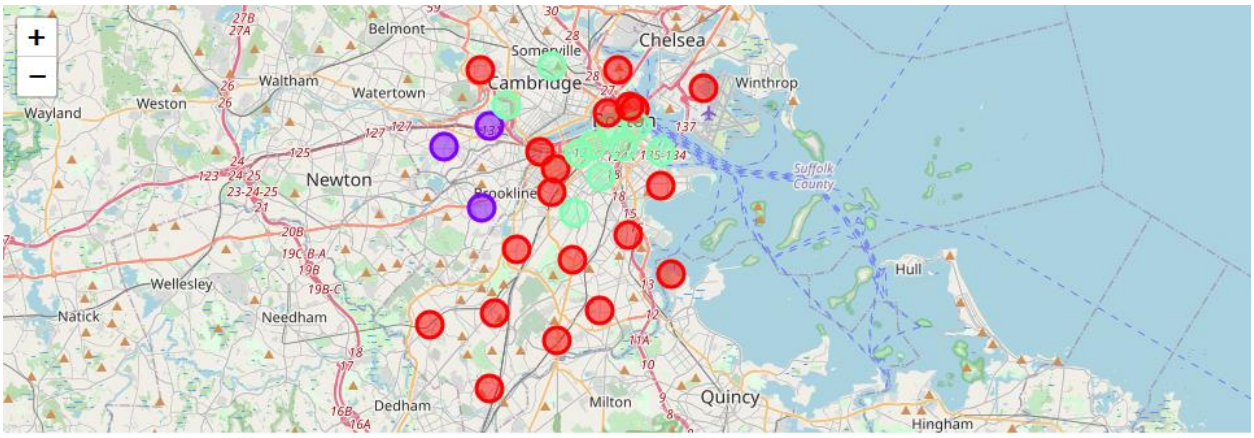
K-Nearest-Neighbor Clustering based on Sushi Restaurant Frequency Alone – in this model, we will only allow for the creation of **N = 3 clusters**, given the relatively low frequency of sushi restaurants. Once we have run and trained the model, we cluster the data based on the frequency of Sushi Restaurants. The following clusters were returned (shown here with the total number of Sushi Restaurants returned displayed next to the Cluster Number):

	Community	Cluster	Total # of Sushi Restaurants
1	Markets / Inner Harbor	0	0.0
4	North End	0	0.0
5	West End / Back of the Hill	0	0.0
6	Fenway / East Fens / Longwood	0	0.0
10	Roxbury Crossing	0	0.0
11	Roxbury / Grove Hall	0	0.0
12	Dorchester / Fields Corner	0	0.0
13	Dorchester / Codman Square	0	0.0
14	Dorchester / Uphams Corner	0	0.0
15	Mattapan	0	0.0
16	South Boston	0	0.0
17	East Boston	0	0.0
18	Charlestown	0	0.0
19	Jamaica Plain	0	0.0
20	Roslindale	0	0.0
21	West Roxbury	0	0.0
24	Hyde Park	0	0.0
25	Cambridge	0	0.0
36	Kenmore / Boston University	0	0.0

	Community	Cluster	Total # of Sushi Restaurants
0	Beacon Hill	2	1.0
2	Financial District / Wharves	2	1.0
3	Chinatown / Tufts-New England Medical Center	2	1.0
7	Back Bay	2	1.0
8	South End	2	2.0
9	Roxbury	2	2.0
30	Somerville	2	1.0
33	North Brighton / Cambridge	2	2.0
34	Prudential	2	1.0
35	South Boston / Fort Point	2	1.0

	Community	Cluster	Total # of Sushi Restaurants
22	Allston	1	3.0
23	Brighton	1	4.0
37	Brookline	1	3.0

Once we have segregated our neighborhoods into Clusters, we can overlay them on a map of Boston to provide an easy visual of how the frequency of Sushi Restaurants change based on the neighborhood. **Red** will be the cluster that returned the lowest number of Sushi Restaurants, **Green** will be the cluster with an average frequency of Sushi Restaurants, while **Purple** will be the cluster with the highest frequency of Sushi Restaurants.



K-Nearest Neighbor Clustering based on both Sushi Restaurant Frequency and General Restaurant Frequency – in this model, we will allow the creation of **N = 8 clusters** to better capture the nuanced differences between Sushi Restaurant Frequency and General Restaurant Frequency. It is likely to prove more fruitful to run this type of K-Nearest Neighbor clustering, as we will be able to create clusters that display clusters with different compositions of Restaurant Frequency to Sushi Restaurant Frequency, such that we may choose a neighborhood with low frequency of Sushi Restaurants, but an average frequency of restaurants in general. The following clusters were returned when we ran a model with N=10 Clusters.

	Community	Cluster	Sushi Frequency	Restaurant Frequency
0	Brookline	0	0.03	0.230000
1	South Boston / Fort Point	1	0.01	0.340000
2	Somerville	1	0.01	0.360000
3	South Boston	1	0.00	0.330000
4	Roxbury Crossing	1	0.00	0.350000
5	Cambridge	2	0.00	0.200000
6	West Roxbury	2	0.00	0.182796
7	Dorchester / Codman Square	2	0.00	0.191919
8	Roxbury / Grove Hall	2	0.00	0.202703
9	Beacon Hill	2	0.01	0.190000
10	Chinatown / Tufts-New England Medical Center	2	0.01	0.190000
11	West End / Back of the Hill	2	0.00	0.190000
12	North Brighton / Cambridge	3	0.02	0.280000
13	Brighton	3	0.04	0.270000
14	Allston	3	0.03	0.280000
15	Dorchester / Fields Corner	4	0.00	0.096774
16	Mattapan	5	0.00	0.135135
17	Hyde Park	5	0.00	0.142857
18	Jamaica Plain	5	0.00	0.162791
19	Markets / Inner Harbor	6	0.00	0.220000
20	Prudential	6	0.01	0.230000
21	Roslindale	6	0.00	0.224719
22	Back Bay	6	0.01	0.220000
23	North End	6	0.00	0.220000
24	East Boston	7	0.00	0.280000
25	Kenmore / Boston University	7	0.00	0.280000
26	Dorchester / Uphams Corner	7	0.00	0.265306
27	Roxbury	8	0.02	0.320000
28	Fenway / East Fens / Longwood	8	0.00	0.310000
29	Charlestown	9	0.00	0.250000
30	Financial District / Wharves	9	0.01	0.250000
31	South End	9	0.02	0.250000

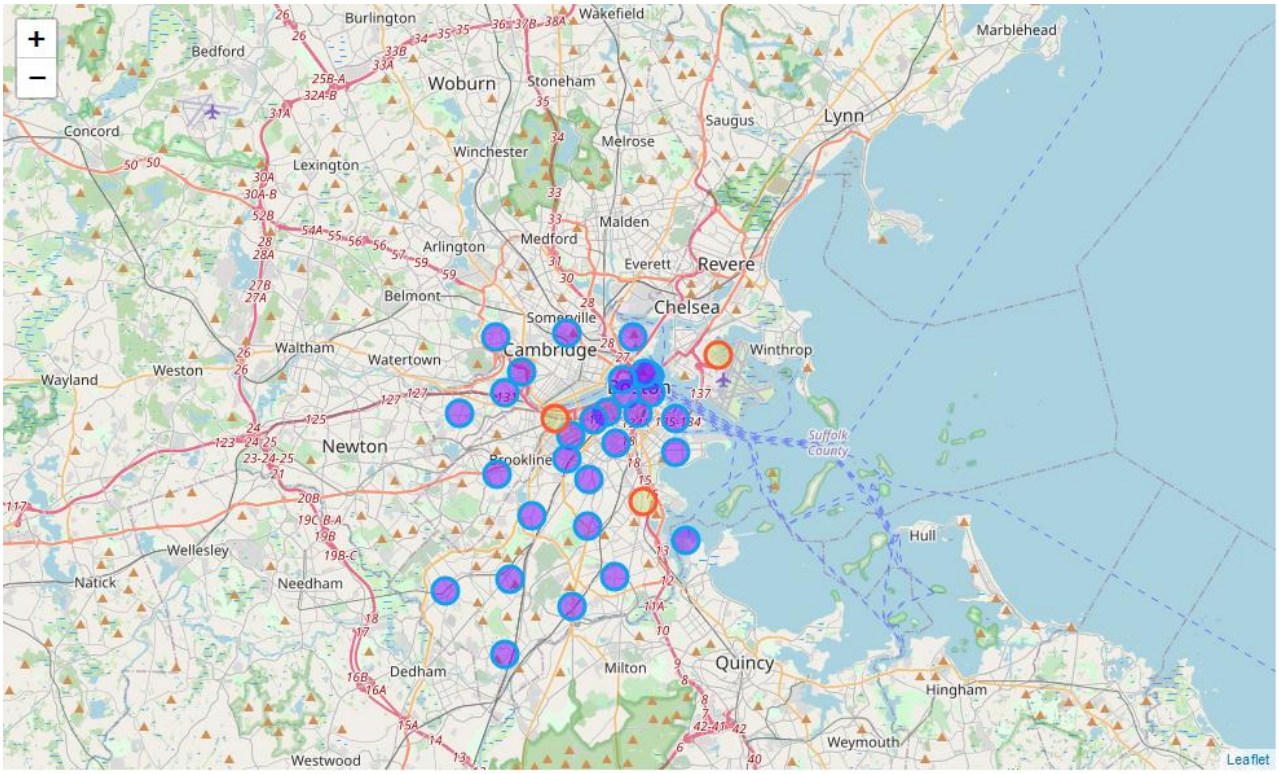
Given the clustering output displayed above, we can immediately identify the clusters that display an average-to-high Restaurant Frequency but a very low frequency of Sushi Restaurants. **Clusters 5&7 display promising results:**

- Cluster 5 displays neighborhoods with abnormally low frequency of both restaurants and Sushi Restaurants.
- Cluster 7 displays neighborhoods with average frequency of restaurants, but with an abnormally low frequency of Sushi Restaurants. This is set is the desired output of our modeling efforts.

Cluster 7 is therefore our desired output, identifying the following neighborhoods as the ideal candidates for neighborhoods in which to open a Sushi Restaurant in Boston. These are:

- East Boston
- Kenmore / Boston University
- Dorchester / Uphams Corner

We can now proceed to plot these results over a map of Boston, displaying Cluster 7 in one color, and every other neighbor in a different color.



Displayed on this map we can see the final output of our efforts: a map of the three most congenial neighborhoods in which to open a Sushi Restaurant in Boston, Massachusetts.

4 – Results

The results of this analytical endeavor were extremely positive. Ultimately, we used Restaurant Frequency and Sushi Restaurant Frequency to cluster neighborhoods and identify which Boston communities were most congenial to opening a new Sushi establishment. Our clustering efforts returned the three following neighborhoods:

- **East Boston**
- **Kenmore / Boston University**
- **Dorchester / Uphams Corner**

When thinking within the frame of our research scope, this cluster perfectly embodies the selection criteria we had set out at the beginning of the research project. We identified neighborhoods where the frequency of Sushi Restaurants was low enough to avoid significant close competition, but neighborhoods that also displayed a bustling restaurant market (as represented by an above-average frequency of restaurants in the given neighborhood).

The results appear to be ideal – the entrepreneur will have an easy choice between three neighborhoods that meet the desired criteria perfectly, all thanks to the ease of K-Nearest Neighbor Clustering.

We could also have chosen to perform our clustering using the frequency of Sushi Restaurants alone. However, the relatively limited dataset, together with the exploratory insight regarding Sushi Restaurants' increased chances of success in neighborhoods with an above-average restaurant density drove us towards the final choice.

All in all, the results displayed surprisingly elegant data, given the relative uncertainty of the dataset extracted from the Foursquare API. It would be irresponsible to say that the metric used in this report is entirely appropriate to answer the question in its entirety, but it certainly provides extremely promising results when trying to determine which neighborhood to choose.

5 – Discussion

The report presented here attempted to identify the best neighborhood in Boston in which to open a new Sushi Restaurant. To do this, we identified all communities in the city, geolocated them, and then requested exploratory data from the Foursquare API, attempting to garner an understanding of each neighborhood's venue make-up.

As an exploratory effort, it was extremely successful, as we were able to identify each community, locate it, and retrieve a sample of venues around the given set of coordinates. We then analyzed each community to determine those which had higher and lower frequencies of Sushi Restaurants, as well as restaurants in general. This returned an extremely interesting dataset, onto which we were able to run different permutations of K-Nearest Neighbor Clustering. We attempted two different methodologies, and the second, feeding the model both the Sushi Restaurant Frequency and the General Restaurant Frequency easily proved the best method to answer our original research question. All in all, we believe the analysis done in this report should not serve as a final answer to the entrepreneur's question, but rather as a jumping-off point for further research.

Further Research: it would be irresponsible to advise an entrepreneur to make his/her decision based on the outcome of this report alone. The dataset retrieved from the Foursquare API was (given our Free Account) relatively limited, and certainly did not provide as full a picture as a researcher might hope. Though the random selection does, to an extent, make up for the relatively low sample size (when it comes to # of Sushi Restaurants), we certainly recommend that the Foursquare API be leveraged to its full potential with a paid account. Another recommendation we would like to promote, particularly when deciding between the three neighborhoods listed in the results, is a correlational study between the neighborhood demographics and the general demographics of the more prolific Sushi eaters. This would allow the entrepreneur to dive deeper into the data and select his/her neighborhood with an even wider arsenal of data.

6 – Conclusions

In conclusion, we have analyzed Boston neighborhoods to identify a cluster that could be deemed most ideal for the opening of a new Sushi Restaurant. K-Nearest Neighbor clustering revealed that **East Boston, Kenmore and Dorchester** were the most suitable neighborhoods given the proposed business question as stated in the introduction. We certainly recommend further research on the topic, but the three neighborhoods' business make-up render them ideal candidates for the entrepreneur's efforts.