Looking for the Best Location for an Italian Restaurant

Hongmei Zhu

September 19, 2020

1. Introduction

Assume I own a consulting firm that provides location services for clients. The clients can be anyone who believes that location matters for their business, which can be restaurants, banks, fitness centers, barber shops, and other shops.

During the Pandemic, a lot of restaurants were shut down, and some of them even had to close permanently. As business shutdowns orders are lifted, new restaurants are emerging to meet people's demand for food in big cities. In this capstone project, I take an Italian restaurant as an example of a problem to solve. One of my clients is going to open an Italian restaurant in Boston and he wants us to help him find the best location out of three candidate locations.

1.1 Audience

Because the business we can serve can be all sorts of businesses that have a demand for the best location, our audience are companies, organizations or business persons who want to find the best location to run their business with high performance.

1.2 Problem to solve

The problem to solve is to look for the best location for an Italian restaurant.

1.3 Criteria of the best location

There are different criteria for different businesses that define what is the best location. In this project, the criteria of the best location for the Italian restaurant are as follows:

- There are no more than two Italian Restaurants within 800 meters.
- There is at least one big park within 800 meters.
- There is at least one big mall within 800 meters.

2. Data

2.1 Data sources

(1) Candidate locations

Assume there are three places available for our customer to rent, but they don't know which location will be the best at which to operate an Italian restaurant. So we already know the addresses of three candidate locations in Boston as follows:

- **Location 1**: 48 Deckard St, Boston, MA 02121
- **Location 2**: 52 Mt Vernon St, Boston, MA 02108
- **Location 3**: 270 W Fifth St, Boston, MA 02127

(2) Boston zip code data

Source: https://bostonopendata-boston.opendata.arcgis.com/datasets/53ea466a189b4f43b3dfb7b38fa7f3b6_1

(3) Massachusetts zip code and latitude/longitude information:

Source: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?q=MA

(4) Venue data near candidate locations

Data includes existing restaurants, parks and shopping malls near the three candidate locations, which will be obtained through requests to Foursquare that returns data for venues in JSON format.

(5) Shopping mall data

Shopping mall data will be obtained from this webpage as well as google search results: https://www.bostoncentral.com/shopping.php

2.2 Data processing

The Python libraries pandas, numpy, geopy.geocoders are used to process data.

(1) Boston zip codes and their coordinates

Boston zip code data only has zip code information and doesn't include latitude and longitude information; but I need to know the location of center of each zip code so that I can add zip codes on a map. On the other hand, Massachusetts data includes zip codes and their coordinates expressed as latitude and longitude values. If a Boston zip code is left joined with Massachusetts data on the zip code, then I will have coordinate information along with zip code. Function MERGE in Pandas is used to join two data frames based on the zip code. The result is as shown in Figure 1.

	OBJECTID	Zip	ShapeSTArea	ShapeSTLength	City	State	Latitude	Longitude	Timezone	Daylight savings time flag	geopoint
0	1	2134	3.721936e+07	40794.18240	Allston	MA	42.355147	-71.13164	-5	1	42.355147,-71.13164
1	2	2125	6.476052e+07	62224.52144	Boston	MA	42.316852	-71.05811	-5	1	42.316852,-71.05811
2	3	2110	6.637284e+06	18358.21350	Boston	MA	42.356532	-71.05365	-5	1	42.356532,-71.05365
3	4	2118	3.116158e+07	32353.40762	Boston	MA	42.338724	-71.07276	-5	1	42.338724,-71.07276
4	5	2126	6.078585e+07	45488.39471	Mattapan	MA	42.272098	-71.09426	-5	1	42.272098,-71.09426

Figure 1. Boston zip code and latitude/longitude

(2) Candidate locations and their coordinates

Mapping the candidate locations on a Boston map helps to gain a better understanding of their relative locations. Mapping needs coordinates in latitude and longitude. Since I only know the addresses of these location, they need to be geocoded to produce coordinates in latitude and longitude. The Python library geocoders in geopy provides geocoding functionality that transforms addresses to latitudes and longitudes. The result is as follows:

```
Lat_list = [42.315778819228, 42.35810105, 42.3353948028169]
Long_list = [-71.08507985113481, -71.0669259504036, -71.04899235211268]
```

Lat_list includes three latitude values for three candidate locations, and long_list includes three longitude values for the same.

(3) Shopping malls and their coordinates

Just like the candidate locations, original shopping malls data don't have coordinate information but physical addresses. The Python library geocoders is used again to convert shopping malls addresses to coordinate in latitudes and longitudes. The part of the result is shown in Figure 2.

	mall name	address	city	zipcode	latitude	longitude
0	Prudential Center Boston	800 Boylston St	Boston	2199	42.347172	-71.082506
1	Copley Place	100 Huntington Ave	Boston	2116	42.347231	-71.077584
2	Faneuil Hall Marketplace	4 S Market St	Boston	2109	42.359706	-71.055068
3	South Bay Center	8 Allstate Rd	Boston	2118	42.326896	-71.061791
4	Washington Park Mall	330 Martin Luther King Jr Blvd	Boston	2119	42.318837	-71.084578
5	Longwood Galleria	400 Brookline Ave	Boston	2215	42.338720	-71.107376

Figure 2. The first five shopping malls in Boston and their coordinates

3. Methodology

The entire analysis methods are illustrated in Figure 3, which includes nine steps. The first six steps will be described in this section. The seventh part, Compare Analysis Results, will be described in section 4 Results, and the eighth part, Adjust Method if Needed, will be described in section 5 Discussion. Finally the nineth part Make Decision, is put in section 6 Conclusion.

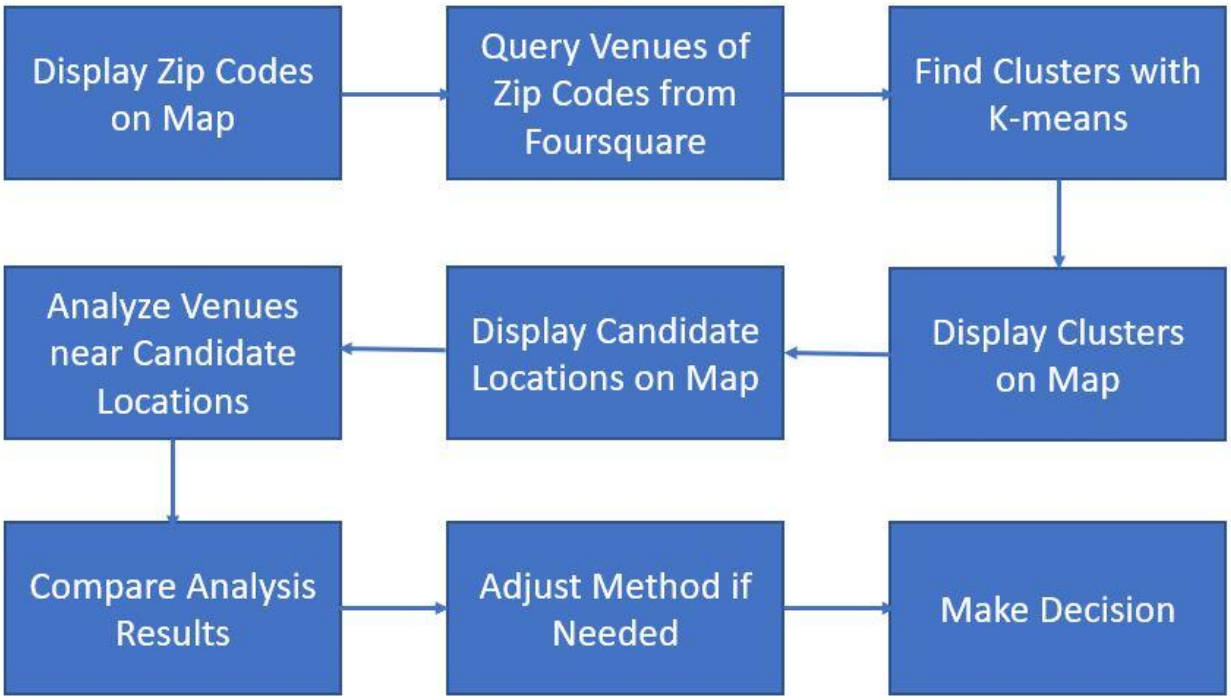

Figure 3. Analysis steps

3.1 Mapping zip codes of Boston, MA

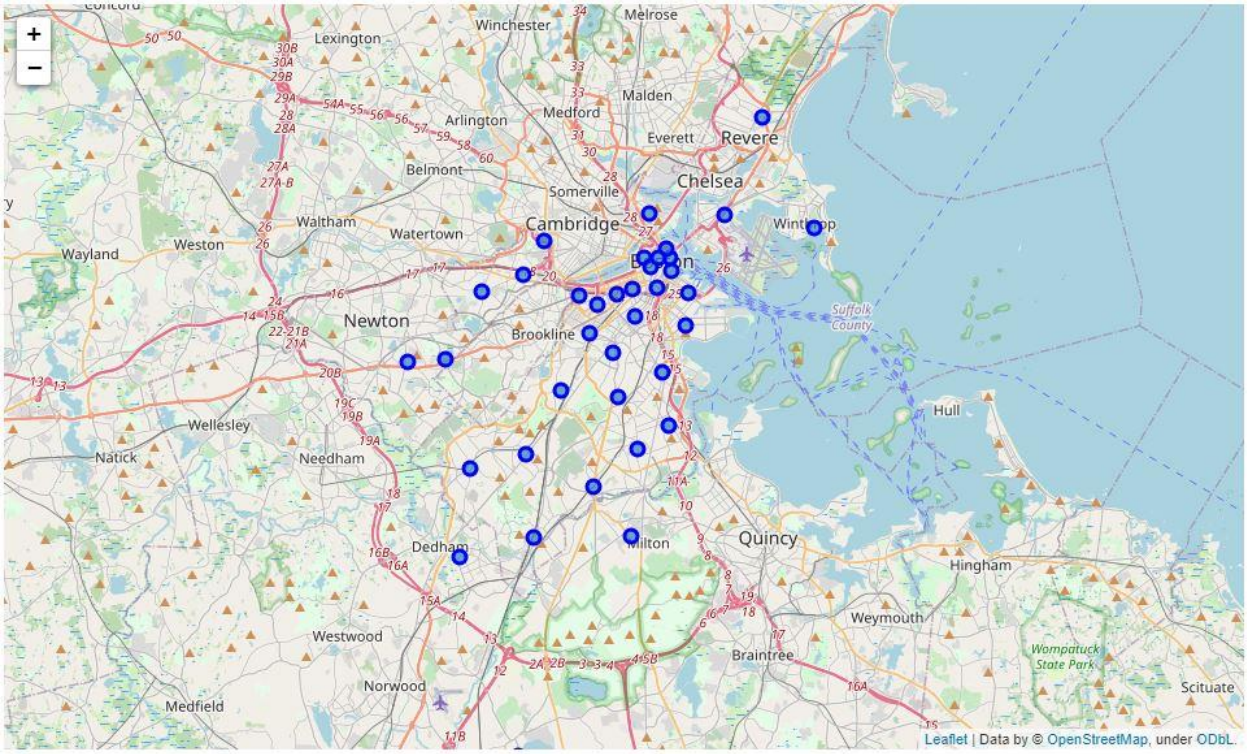

Figure 4. Boston zip code map

Assume that I don't have much knowledge about the city of Boston. Mapping zip codes of the city on a map gives me a general impression about the coverage and geometry of Boston. The zip code map displayed in Figure 4.

3.2 Queries for venues of each zip code

I need to better understand the geographic distribution of all sorts of businesses in Boston before I am able to determine the best location for an Italian restaurant. To gain knowledge of what kinds of businesses are in each zip code in Boston, I need to obtain venue information for each zip code.

The venue request is formed with the coordinate of each zip code and other query parameters, and sent to the Foursquare server. The server returns a response in a JSON file after it receives the request and retrieves data successfully from it's location database. The first five records are shown in Figure 5.

	Zip	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	2134	42.355147	-71.13164	Lulu's Allston	42.355068	-71.134107	Comfort Food Restaurant
1	2134	42.355147	-71.13164	Kaju Tofu House	42.354329	-71.132374	Korean Restaurant
2	2134	42.355147	-71.13164	Fish Market Sushi Bar	42.353039	-71.132975	Sushi Restaurant
3	2134	42.355147	-71.13164	BonChon Chicken	42.353105	-71.130921	Fried Chicken Joint
4	2134	42.355147	-71.13164	Mala Restaurant	42.352960	-71.131033	Chinese Restaurant

Figure 5. The first five venues of zip code 02134 returned from Foursquare

Each row in Figure 5 includes zip code, the coordinates of the zip code, venue name, venue coordinates and venue category. Venue category tells me the type of a venue. I will use venue category information to create a cluster map of businesses.

3.3 K-means algorithm

K-means is an unsupervised algorithm that looks for similarity or dissimilarity within a dataset. It doesn't know any cluster internal structure of the data in advance. It divides the dataset into k non-overlapping subsets with the similarity algorithm. These subsets are also called clusters, and k is the number of resulting clusters. In this project, I use Python library kmens in sklearn to find clusters in the venue data returned by Foursquare.

Venue category information needs to be extracted from data in Figure 5 before I input them into kmeans function to create clusters. The business types of the venue data are on column "Venue Category". I first build a new data frame whose columns are all venue categories. There are 229 columns because there are 229 venue categories. Zip code is added to the data frame as a column. So there are total 230 columns in the data frame. Each record in this data frame has values of zip code, 0 or 1. 1 denotes the category in the corresponding column exists in the zip code, and 0 denotes that category doesn't exist in the zip code. Then the records are grouped by zip code and normalized with mean values for input of the K-means function. Any non-numeric column can't be part of the input, and therefore the zip code column needs to be dropped.

The third step is to retrieve the top 10 most common categories from the normalized dataset by sorting values in each record in a descending manner.

Kmeans is a Python library in the sklearn.cluster module, which takes the normalized data frame as input and output labels for each zip code. Here I choose k = 5 and so label values are integers from 0 to 4, as shown in Figure 6.

	Cluster Labels	Zipcode	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	4	2021	Athletics & Sports	Paintball Field	Farm	Yoga Studio	Falafel Restaurant	Food Truck	Food Service	Food Court	Food & Drink Shop	Food
1	2	2026	Diner	Track	Farmers Market	Gym	High School	Yoga Studio	Falafel Restaurant	Food Service	Food Court	Food & Drink Shop
2	1	2108	Coffee Shop	Pizza Place	Italian Restaurant	Sandwich Place	Restaurant	Plaza	New American Restaurant	Steakhouse	American Restaurant	Historic Site
3	1	2109	Italian Restaurant	Seafood Restaurant	Bakery	Park	Historic Site	Pizza Place	Café	Tourist Information Center	Grocery Store	Hotel
4	1	2110	Boat or Ferry	Seafood Restaurant	Park	Historic Site	Hotel	Harbor / Marina	Coffee Shop	Asian Restaurant	Italian Restaurant	Aquarium

Figure 6. Clusters with top 10 most common venue in each zip code in Boston

From the table in Figure 6 I learned the categories of each label denotes, which is:

- Label 0 - print shop, trail etc.
- Label 1 - all sorts of food and drink places
- Label 2 - farm, gym, school etc.
- Label 3 - park, sport field etc.
- Label 4 - sport field, farm etc.

3.4 Visualization of clusters on the Boston map

After joining zip codes with latitude and longtitide data, the top 10 common venues data have coordinates. I use it as the final cluster data, which is visualized on map in Figure 7 showing 5 clusters in different colors.

The map in Figure 7 tells me that there are all sorts of businesses in downtown areas in Boston, except four zip code areas which are in colors other than blue. A lot of venues are different cuisines of restaurants and drink shops, as well as shopping malls, schools, and so on. Most zip code areas are equally commercial areas and therefore there is no preference for a candidate location at city scale. This means the candidate locations can be in any of these purple zip codes at city scale, or I cannot say which zip code area is better than other zip codes. My interest now is to figure out which location is the best one in terms of meeting the criteria.

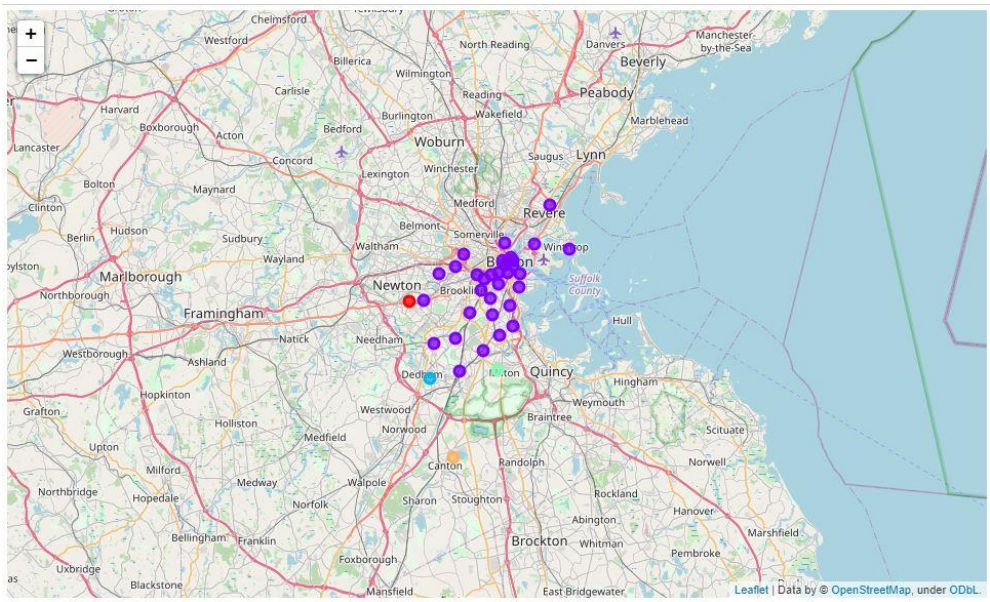

Figure 7. Cluster map of Boston zip code

3.5 Mapping of candidate locations

Viewing the candidate locations on the Boston map provides an intuitive way to help us look for the best location. I visualize them on the map as shown in Figure 8.

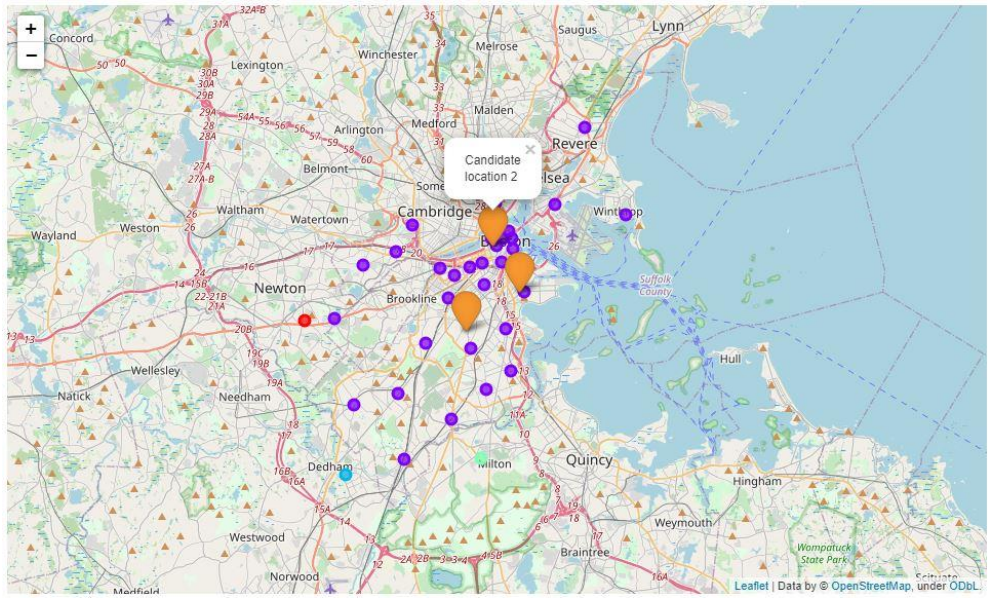

Figure 8. Three candidate locations on Boston map

3.6 Venue analysis of the candidate locations

Knowledge of surrounding businesses near the candidate locations is critical for making a good decision about the best location for the new Italian restaurant. I am very interested in what types of businesses are within 800 meters (half a mile) from each candidate location.

As with the previous method, here I send venue requests to the Foursquare server to obtain venue information near each candidate location. Given that shops in Boston are very dense located, the maximum return number is set to 200 to avoid missing any venues near the candidate locations. The response from the server is saved into a data frame, whose first five results are as shown in the following Figure 9.

	id	venue name	venue latitude	venue longitude	venue category
0	0	Popeyes Louisiana Kitchen	42.318547	-71.082893	Fried Chicken Joint
1	0	The Merengue Restaurant	42.319199	-71.077655	Caribbean Restaurant
2	0	Roxbury YMCA	42.317791	-71.082789	Gym / Fitness Center
3	0	Flames III	42.309020	-71.083061	Caribbean Restaurant
4	0	Walgreens	42.316881	-71.082464	Pharmacy

Figure 9. Venues of the three candidate locations

ID 0 means candidate location 1, 1 means candidate location 2, and 2 is candidate location 3 in the result. Based on the ID of each candidate location, the subset of the result is selected and saved into three data frames. Each data frame is fed into a module called check_best_location that I write with Python. In this module every value of venue category is examined to determine if they are one of the criteria mentioned in section 1.3, specifically the module checks if any parks and shopping malls exist in the venue category. Again, the criteria of the best locations are follows:

- There are no more than two Italian Restaurants within 800 meters.
- There is at least one park within 800 meters.
- There is at least one mall within 800 meters.

4. Results

The results from module check_best_location are listed in Table 1.

Candidate Location	Analysis Output
1	There is no Italian Restaurant within 800 meters There is no park within 800 meters 1 shopping mall(s) found within 800 meters
2	5 Italian Restaurant(s) found within 800 meters 3 park(s) found within 800 meters There is no shopping mall within 800 meters
3	2 Italian Restaurant(s) found within 800 meters 1 park(s) found within 800 meters There is no shopping mall within 800 meters5.

Table 1. Result of best location analysis

5. Discussion

Result of location 1: There is one shopping mall within 800 meters, and there is no Italian restaurant or park. This might be a good place.

Result of location 2: There are 3 parks within 800 meters, which is good, but there are five Italian restaurants within that range already exist, and there is no shopping mall in that range.

Result of location 3: Near this candidate location, there are two Italian restaurants and one park within 800 meters, which is attractive for the new restaurant, but there is no shopping mall nearby.

Because the results convey a mixed message, I'm not able to make a final decision right away according to the information available. Further consideration is needed, such as adding shopping malls on the map could be helpful, as shown in Figure 10.

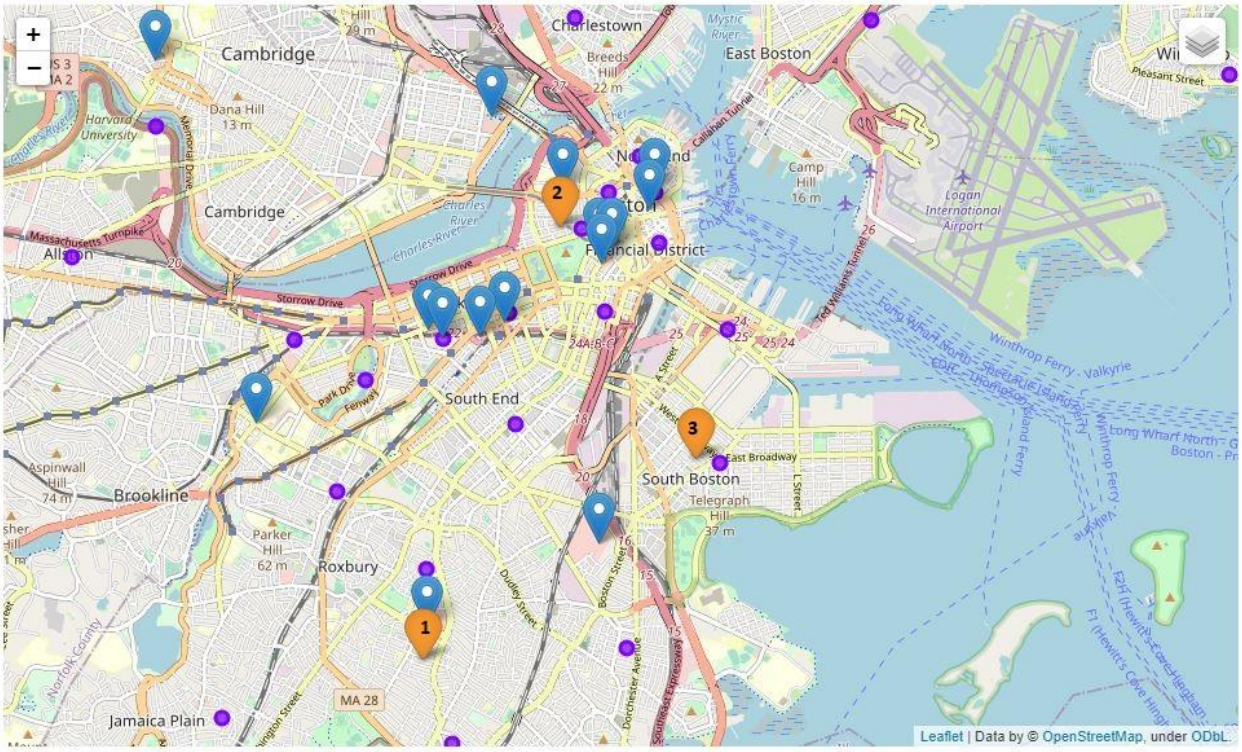

Figure 10. Boston map with shopping malls (blue markers) and candidate locations (orange markers)

By reading the map carefully as shown in Figure 10, I learn candidate location 1 is a little bit far from the center of Boston but there is still a large population in that area. The mall close to it is Washington park mall, which is a big mall attracting heavy customer traffic. Candidate location 2 is located at the center of downtown Boston. Nearby parks attract people but there are five restaurants already there, which means potentially fierce competition among these multiple Italian restaurants, and shopping malls are a little bit far from it. Location 3 is in the southern neighborhoods where there are no big shopping malls; but Thomas park is found near this location. It is a nice park that can attract a lot of people.

I decided to pick one from location 1 and 3, and to determine which I should pick out of the two, I ask our client to provide more information such as rent and seat numbers they prefer. I learn that candidate 1 has the best size with lower rent and therefore I make the final decision that the candidate location 1 is the best for the new Italian restaurant.

6. Conclusion

In this project, my goal is to find the best location from three candidate locations in Boston, Massachusetts for a new Italian restaurant. To reach this goal, I leveraged data science Python libraries such as pandas, geopy, sklearn.cluster, folium, etc. to process, analyze and visualize data. In addition, I utilized Foursquare to obtain venue data near locations of interest. After analyzing data, as well as carefully reading the map displaying locations, I made a final decision that candidate location 1 is the best one to operate the new Italian restaurant. If I had demorgraphic data of each zip code, I could put it on the map, which might help me to make a more accurate decision. Also, I could change K of K-mean to check which one is the most suitable value for creating clusters, which could be another project in the future.