Density and Specialization of Restaurants in Boston

Alexander Hahn, August 1st, 2020

---

1. Introduction

With life slowly returning to normal after the Covid epidemic, many people will want to start new business and restaurants to cover the gap caused by so many going under.

Boston is a great place to start a business with a thriving biotech and science industry and growing population. But looking into places to start a new business there can be hard without knowing large amounts of information about the city.

1.1 Overview

The idea is to find a place that has has a thriving restaurant scene, showing the area is ripe for future expansions but that doesn't cover the new restaurants niche. For instance, building a sushi place next to another sushi place would more than likely half the profits of both while building a sushi restaurant in a area without may just provide a bump to that restaurant.

1.2 Problem

The idea is to find a location or neighborhood that already has a high density of restaurants but does not have a large amount for a specific niche.

---

2. Data

There are two sets of data required for this project. A list of neighborhoods in Boston with their co-ordinates as well as the location data from four square that we can use to find the location of restaurants in the city.

Using those two pieces, we will look for a variety of restaurants types in the city and build a cluster to find areas of high and low density as well as catalog the types of restaurants in the city.

2.1 Boston neighborhood data

This data consist of names and locations for each of the major regions of the city. It also contains population and density data that we will not be using currently.

2.2 Foursquare Data

We will be cycling through each of the areas and look for the following restaurant types : African, American, Asian, Chinese, Japanese, Korean, Thai, BBQ, Burger, Burger, Fish, French,

German, Greek, Indian, ,Italian, Latin, Mexican, mediterranean, Middle eastern, pizza, Southern, Spanish, Steak, Turkish, Other.

With each of these data sets we will look for them in each of the neighborhoods and then overlay them all into a single pandas frame removing duplicates. This should provide adequate coverage of the city's restaurants.

The data from Foursquare we will use contains information on Venue name, category, latitude and longitude.

---

3. Exploratory Data Analysis

We will need to gather data on the neighborhoods of Boston and their longitude and latitudes in order to get a general overview of the city and begin the process of querying foursquare for restaurant data.

3.1 Problem with Foursquare

While using Foursquare as a free member, you are limited to the number of queries you can ask and the results contained in each of these queries. For instance we can get up to 100 results per query and 2000 per day.

In order to make sure that we have gathered enough data to preform the clustering (which requirers a large amount of data to preform) we will need to run many many queries of varying types and locations and then later merging all the data together to get a larger, more comprehensive map.

We will initially pull in the data for the neighborhoods and then load in a list of probably restaurant categories and pull the entire list for each neighborhood. Doing this will provide

Neighborhood	Restaurant	Queries
26	22	100
	Total Venues :	57200

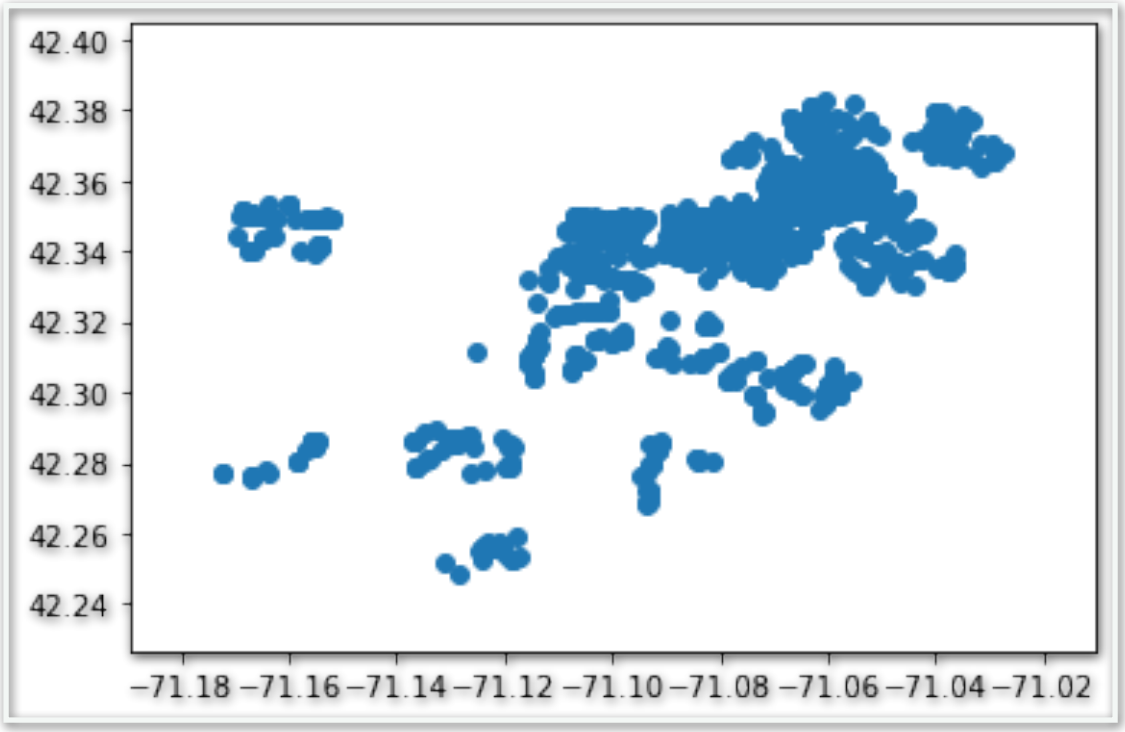
3.2 Data Pulling Results

After pulling all 57200 sets of data and then merging the data removing the un-used data sets, we end up with 3954 unique entries. This is a sharp drop in the number of venues we looked for but it is likely a very full and comprehensive list of venues that will provide strong coverage of the city.

3.3 Initial Mapping

After consolidating the data, we will then pull the data into matplot lib graph to get a general overview of the city and where the distributions are.

As you can see from the graph, there is a large cluster in the top right followed by many other islands of restaurants. This is definitely useful to note as it means that there are in fact islands



of restaurants in the city. This would indicate that we need to search each of these “islands” for places of low density for a specific type of restaurant.

---

4. Predictive Modeling

Now that we have the data in a data frame, have removed the duplicates and see that in fact there are islands for us to look at, it is not time to set up a density based clustering algorithm that will automatically sort the data into various cluster based on the density of the area. It is important to note that we need the algorithm to decide where the cluster are and how big they are. If we set a value then we will not get an accurate picture of the entire city.

4.1 Variables in DB Scan

There are two important variables to look at when starting a density based algorithm. First is the Epsilon value that will be useful in determining how dense the cluster need to be in order to determine if they are separate or unique. The second is the minimum samples value that will show the minimum number of samples needed to form a cluster. For instance, one restaurant outlier should not be considered a cluster as it is too far away from other restaurants and the goal is to find hot spots that need diversity rather than to find empty spaces.

In our test, we will use the following values:

Epsilon	0.001
Minimum Samples	5

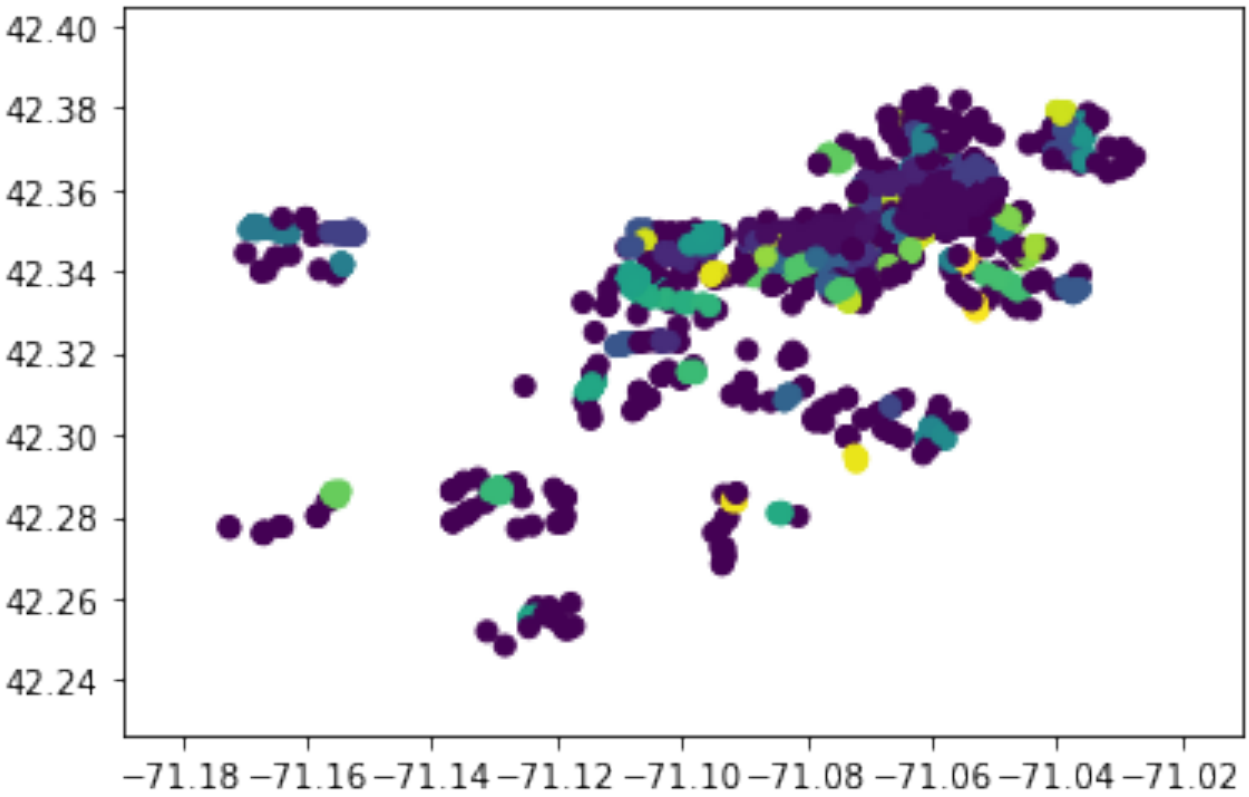
The reason for such a low epsilon is due to the fact that the data is quite small, we are looking at similar levels of difference between the longitude and latitude values.

4.2 Run the DB Scan

We ran the DB Scan and found that the algorithm identified 94 different clusters. The following table shows the density of each of the clusters.

4	12	10	6
6	9	7	83
9	11	21	110
10	21	355	18
6	11	6	10
8	249	22	44
9	69	60	15
6	1544	8	6
14	11	20	18
8	11	5	6
26	21	10	8
8	19	5	8
5	6	13	12
15	5	11	14
7	5	19	5
15	33	17	10
16	6	14	11
9	6	17	6
15	5	6	50
7	15	21	11
11	6	5	6
53	16	8	5
5	7	5	
6	26	7	

As you can see from the data, not all clusters are made equal. We have some that represent the minimum values need to form a cluster, 5 and one that has 1544 unique points. This is likely to be somewhere like the city center with restaurants stacked on top of each other almost or at least covering the city blocks.



The above graph shows the clusters formed all around the city with pockets in the middle and outlier islands showing where the algorithm found lines dividing the two.

Using this information, we now need to find the number of Mexican restaurants per district. This will show us the uniqueness of starting a new Mexican restaurant per each of these clusters show above.

4.3 Finding the balance between density/uniqueness

The ultimate problem we need to resolve is how to distinguish between areas that are of low density and of high uniqueness as apposed to those areas of high density but low uniqueness. We will disregard any areas that contain no Mexican restaurants as they are likely to be too small for our study,

We will use the following formulas to determine the value of an area in terms of density and uniqueness:

> *VALUE = (# OF RESTAURANTS IN CLUSTER/TOTAL RESTAURANTS) * (# OF UNIQUE RESTAURANTS/TOTAL # OF RESTAURANTS IN CLUSTER)*

We will then sort by the largest value and find which area both has many restaurants and has few Mexican ones.

Cluster	Total Value	Cluster	Total Value
1	0.019473950429944400	6	0.0005058168942842690
2	0.0032878098128477500	12	0.0005058168942842690
32	0.0030349013657056100	0	0.0005058168942842690
8	0.002529084471421350	7	0.0005058168942842690
16	0.0015174506828528100	73	0.0005058168942842690
47	0.0010116337885685400	10	0.0005058168942842690
20	0.0010116337885685400	55	0.0002529084471421350
3	0.0007587253414264040	58	0.0002529084471421350
33	0.0007587253414264040	64	0.0002529084471421350
57	0.0005058168942842690	22	0.0002529084471421350
52	0.0005058168942842690	21	0.0002529084471421350
70	0.0005058168942842690	75	0.0002529084471421350

Out of the above graph, we have found that region 1 has the highest value of 0.0195.

---

5. Conclusion

After generating the various clusters, finding the population of venues in each of them then searching for the total uniqueness/density of a region, we have found that region 1 contains the best options for finding a new local. Options 2, 32, 8 and 16 are also prime candidates that have achieved high levels of uniqueness.

The Longitude/Latitude of the prime candidate is -71.0588586089431 and 42.35557291694653 respectfully.

This study does not have its flaw though. Using the weight system to get rid of outliers, we may have inadvertently removed many good options for business by focusing on the “best value”

This also does not show which areas are affordable or in the price range of the future business owner, it nearly provides a tier list of possible locations based on density and uniqueness of the region. So this data in its own right is not the final results of where to place a new business, just a useful indicator of hot spots in the city and low Mexican restaurant areas.

6. Future Directions

There could be many future growth options for the project. We could use unemployment statistics to find areas with potential new employees, we could use affluence values for each district to find a potential price policy for the food. We also do not have information on the viability of many of the business that could come with more historical data. For instance maybe some of the hot spots contain business that open and closed frequently but are still considered prime candidates due to other factors outside of what we are looking.

The project as I see it is only a step or a piece of the larger project and only provides some large scale insights into region that should be looked into.

Tweaking the DB Scan algorithm could also have wide effects such as increasing the size of the clusters, restricting the Epsilon even more to find more precision etc…

Including more data and more varied data on restaurants could also be a highly valuable piece of data. You could use data from other types of restaurants and even other types of business to find varied amounts of information on each district.