The Battle of Neighborhoods | Finding a Good Place in Boston, MA

Introduction

Boston Massachusetts is a great place to open up a new restaurant or eatery. It can also be an ideal place to open up museums or shops. Anytime type of venue really. In a busy city it will always be considered a good idea to open one of these establishments.

Boston is a great place and there are plenty of people there to fill our establishment. For a venue to succeed we need to be able to have our venue in the ideal location. In a big city, that can be hard because of the possibility of opening in a dead area. So, the problem is around where we should open this venue. We also need to know what kind of venue we should open.

Any restaurant owner or shop owner looking to open a new establishment in a new city would be interested in analysis like this. We are looking into Boston but in theory this type of analysis could be used for any city. Ideally, after this we should be able to tell where the busiest neighborhoods in the city is. In turn that should be the best place to open a new venue. We will also be able to tell what kind of venue to open in the best area.

Data / Methodology

For our purposes, we first must get the neighborhood locations around the city of Boston. The zip codes of the communities could be used for our purposes. We can use this website, http://archive.boston.com/news/local/articles/2007/04/15/sixfigurezipcodes_city, to get the zip codes that we are going to investigate. From there we can derive the longitude and latitude of these communities. There we can get the most popular venues around these locations.

The Data from that webpage was compiled into a data frame. This scraping was done by reading the web page source and picking out the table information. That gave us the zip code and the neighborhoods that we can use in the rest of our analysis.

From there we used a library to look up the associated longitude and latitude of each of the zip codes. We can add that information to our data frame from earlier. From here we can make sure there are no duplicates or empty entries in our data set.

Now we could use the FourSquare API. The FourSquare API is used to look up location information. They have a useful explore API that can give us a look at the venues around our communities. We can compile all of this information into a new data frame with the type of venue, venue name, longitude, latitude, and the neighborhood they are found in.

Our data set would look something like this:

| Zip Code	Neighborhood	Latitude	Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category

Clustering

Once we have all of our data we can get it prepared for our model. Our original table has a top five rows that looks like this:

Zip Code	Neighborhood	Latitude	Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
02101	Downtown Boston	42.34724	-71.06456	Whole Foods Market	42.345304	-71.063061	Grocery Store
02101	Downtown Boston	42.34724	-71.06456	Turnstyle Cycle	42.345806	-71.063228	Cycle Studio
02101	Downtown Boston	42.34724	-71.06456	Shore Leave	42.345279	-71.063870	Tiki Bar
02101	Downtown Boston	42.34724	-71.06456	Tatte Bakery & Cafe	42.344815	-71.063969	Bakery
02101	Downtown Boston	42.34724	-71.06456	Mike & Patty's	42.348604	-71.067913	Sandwich Place

During the data analysis we can investigate what kind of venues there are by neighborhood and get information like the following:

| Neighborhood	Zip Code	Latitude	Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
Allston	17	17	17	17	17	17	17
Back Bay	100	100	100	100	100	100	100
Beacon Hill	100	100	100	100	100	100	100
Brighton	48	48	48	48	48	48	48
Brookline	107	107	107	107	107	107	107
Brookline Village	13	13	13	13	13	13	13
Cambridge	84	84	84	84	84	84	84
Charlestown	30	30	30	30	30	30	30
Chinatown / Tufts-New England Medical Center	100	100	100	100	100	100	100

We can also create an initial map of Boston and highlight our neighborhoods:
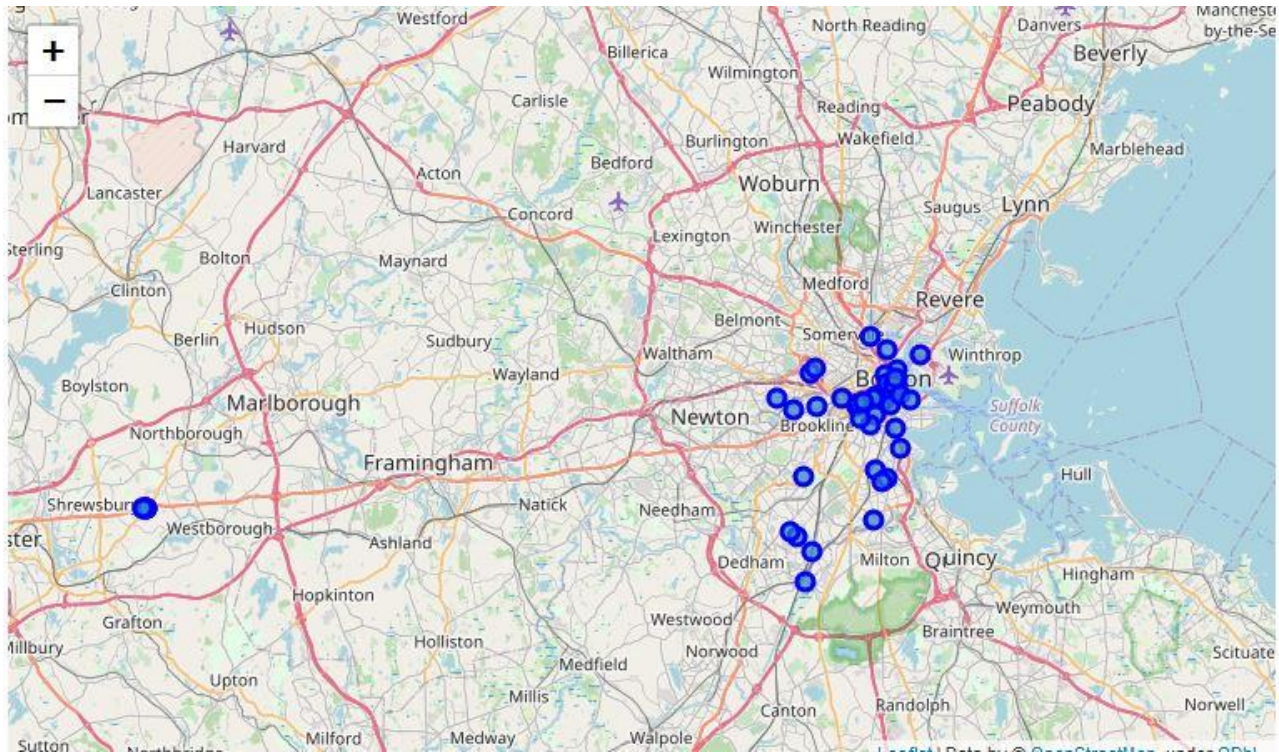
We can identify that we are looking for a solution using clustering. Clustering is an unsupervised machine learning model. It can cluster the data so we have a better idea of where the busiest neighborhoods are. Before running it through our model we need to manipulate our data. Here is what it looks like after the manipulation:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4rth 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
Allston	Gym	Rugby Pitch	Bakery	Gas Station	Squash Court	Grocery Store	Soccer Field	Tennis Court	Beer Garden	Coffee Shop
Back Bay	Spa	Hotel	Gym	Cosmetics Shop	Women's Store	Gym / Fitness Center	Sandwich Place	Coffee Shop	Clothing Store	Seafood Restaurant
Beacon Hill	Coffee Shop	Historic Site	Seafood Restaurant	Sandwich Place	American Restaurant	Pub	Park	Hotel	Plaza	Gastropub
Brighton	Bus Station	Bank	Bakery	Pizza Place	Pub	Coffee Shop	Chinese Restaurant	Café	Smoke Shop	Tanning Salon

We can use the k-means algorithm. This will give us those clusters we were looking for. We run the algorithm for five cluster. We can map those and it will give us a map that looks like this:
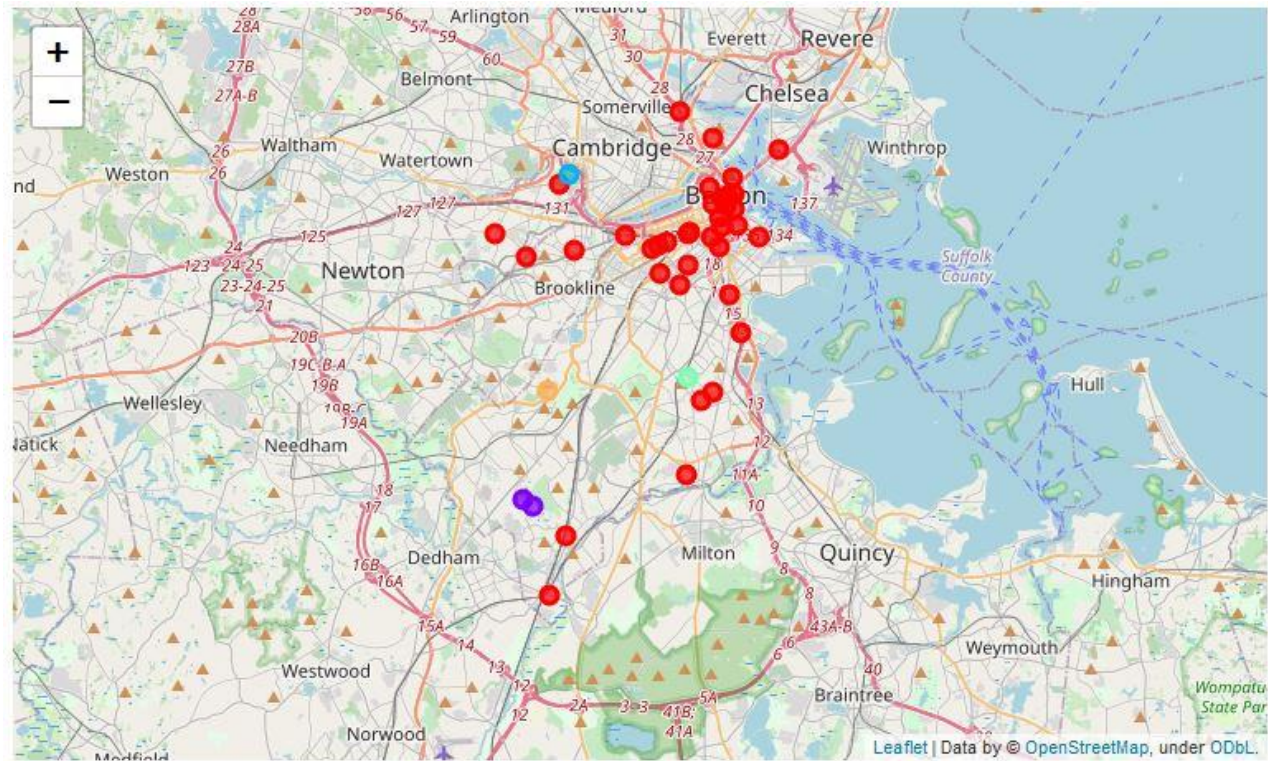
We also have the following information for the clusters and from there we can derive a conclusion.

Cluster 1 (The chart is larger please refer to the notebook associated with the report):

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
Downtown Boston	Coffee Shop	Italian Restaurant	Hotel	Sandwich Place	Seafood Restaurant	Bakery	American Restaurant	Historic Site	Gym / Fitness Center
Beacon Hill	Coffee Shop	Historic Site	Seafood Restaurant	Sandwich Place	American Restaurant	Pub	Park	Hotel	Plaza
Markets / Inner Harbor	Italian Restaurant	Seafood Restaurant	Park	Historic Site	Bakery	Pub	American Restaurant	Sandwich Place	Hotel
Financial District / Wharves	Hotel	Seafood Restaurant	Historic Site	Sandwich Place	Park	Boat or Ferry	Salad Place	Café	Clothing Store

Cluster 2:

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	C...
Roslindale	Lake	Yoga Studio	Performing Arts Venue	Music Venue	Nail Salon	National Park	New American Restaurant	Nightclub	Noodle House	O H
West Roxbury	Lake	Yoga Studio	Performing Arts Venue	Music Venue	Nail Salon	National Park	New American Restaurant	Nightclub	Noodle House	O H

Cluster 3:

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	C...
North Brighton / Cambridge	Park	Residential Building (Apartment / Condo)	Pool	College Hockey Rink	College Stadium	Gym	Noodle House	Opera House	Performing Arts Venue	O S...

Cluster 4:

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
Roxbury / Grove Hall	Food	Garden	Discount Store	Fish & Chips Shop	Pedestrian Plaza	Nail Salon	National Park	New American Restaurant	Nightclub	Noodle House

Cluster 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	C
Jamaica Plain	Home Service	Yoga Studio	Performing Arts Venue	Music Venue	Nail Salon	National Park	New American Restaurant	Nightclub	Noodle House	O H

Conclusion

In Conclusion, we can deduce that we should open an establishment in the first cluster. Cluster one has the most neighborhoods in the area. That means that there is the move foot traffic. We can also see that the type of venue that should open should be a coffee shop or an italian resturant. These seems to be the most popular venues in the area.