Location Analysis for a New Bar Business in Boston

Coursera - Applied Data Science Capstone

James Maxwell

1. Introduction/Business Problem

The goal of this capstone project is to examine the bar scene in Boston with a view to understanding the current location and quantities of the existing bar businesses. A party is interested in opening a new bar in a neighborhood of Boston and is not familiar with the city and would like the location and quantity data as part of a feasibility study to determine whether such a venture is viable. The goal was to cluster similar neighborhoods together from an exiting bar business and generate a count of the existing bars in these neighborhoods.

2. Data collection

The following data is required to enable the study:

- A list of neighborhoods in Boston which defines the geographical scope of the request.
- Latitude and longitude coordinates of the different neighborhoods. This is required to get the maps showing the relative location of each neighborhood within Boston.
- The venue data is extracted which provides information related to existing bar businesses.

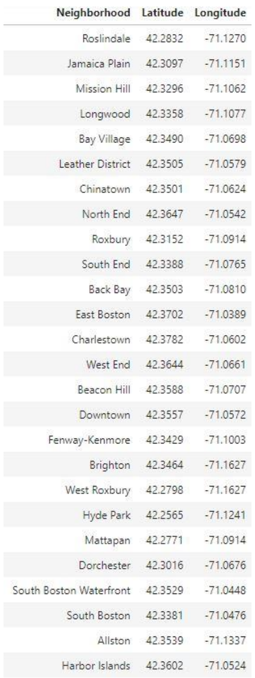
A list of the neighborhoods was obtained from the website of the Boston Planning and Development Agency (http://www.bostonplans.org/getattachment/7987d9b4-193b-4749-8594-e41f1ae27719).

The location coordinates for each of the 26 neighborhoods identified was obtained by a simple search using the google search engine. The neighborhood and location coordinates list were compiled in a single .csv file.

The FourSquare API is used to generate venue data for each of the neighborhoods bar related businesses which in turn is used as the basis for the clustering algorithm.

3. Methodology.

Analysis began with reading in the .csv file with the neighborhood and coordinate data. The resultant dataframe is shown in figure 1.



Neighborhood	Latitude	Longitude
Roslindale	42.2832	-71.1270
Jamaica Plain	42.3097	-71.1151
Mission Hill	42.3296	-71.1062
Longwood	42.3358	-71.1077
Bay Village	42.3490	-71.0698
Leather District	42.3505	-71.0579
Chinatown	42.3501	-71.0624
North End	42.3647	-71.0542
Roxbury	42.3152	-71.0914
South End	42.3388	-71.0765
Back Bay	42.3503	-71.0810
East Boston	42.3702	-71.0389
Charlestown	42.3782	-71.0602
West End	42.3644	-71.0661
Beacon Hill	42.3588	-71.0707
Downtown	42.3557	-71.0572
Fenway-Kenmore	42.3429	-71.1003
Brighton	42.3464	-71.1627
West Roxbury	42.2798	-71.1627
Hyde Park	42.2565	-71.1241
Mattapan	42.2771	-71.0914
Dorchester	42.3016	-71.0676
South Boston Waterfront	42.3529	-71.0448
South Boston	42.3381	-71.0476
Allston	42.3539	-71.1337
Harbor Islands	42.3602	-71.0524

Figure 1

A map of the relative location of the 26 neighborhoods within Boston was produced using a folium map and is shown in figure 2.

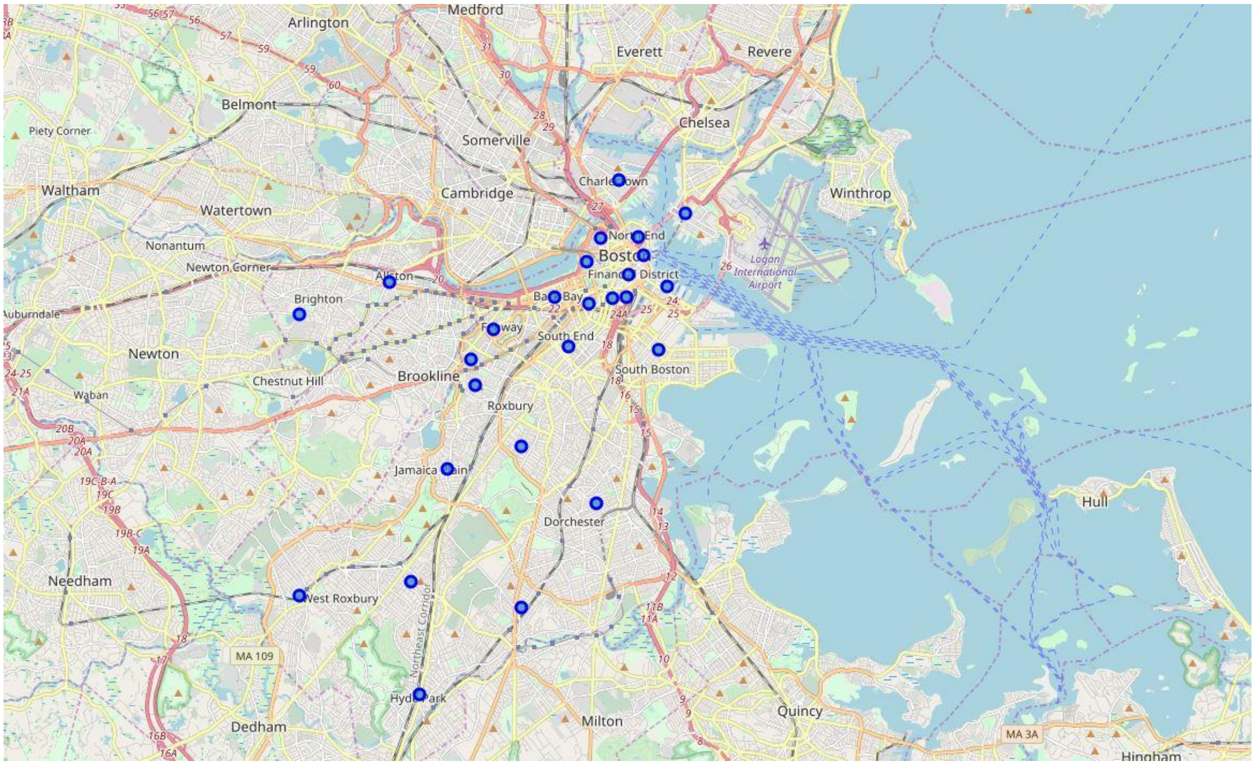
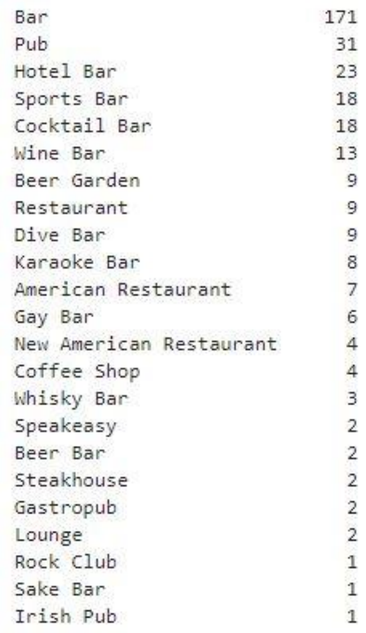


Figure 2

A call to the FourSquare API using the request URL returned venue data for each of the neighborhoods. An additional query was used here restricting the returned venues to be based on search results of the strings ‘bar’ and ‘pub’. The venue search radius was limited to 100 and within a 500m radius of the location information.

Unfortunately, 4 out of 26 neighborhoods did not return any venue data. This is a problem because these 4 neighborhoods do contain bar businesses.

For the returned venue data there were 33 related venue categories and 371 instances of venues for the 22 neighborhoods with data. On closer inspection some venue categories were removed due to only being loosely related to the bar business, leaving the list in figure 3.



```
Bar                        171
Pub                         31
Hotel Bar                   23
Sports Bar                  18
Cocktail Bar                18
Wine Bar                    13
Beer Garden                  9
Restaurant                   9
Dive Bar                     9
Karaoke Bar                  8
American Restaurant          7
Gay Bar                      6
New American Restaurant      4
Coffee Shop                  4
Whisky Bar                   3
Speakeasy                    2
Beer Bar                     2
Steakhouse                   2
Gastropub                    2
Lounge                       2
Rock Club                    1
Sake Bar                     1
Irish Pub                    1
```

Figure 3

The next step in the process is to cluster the neighborhoods into groups of similar neighborhoods based on their venue data. A value of four was chosen for the K-means clustering algorithm and the 4 neighborhoods with no venue data were given their own cluster (i.e a fifth group) after the K-means assigned the neighborhoods with venue data to 4 cluster groups.

One hot encoding is performed on the neighborhood venue data by transforming the data into 0’s and 1’s. The data is then grouped by neighborhood and the mean of the frequency of each venue category is calculated.

This data is input to the k-means algorithm, an unsupervised machine learning algorithm that identified k number of centroids (k=4 in this case) and allocates every point to the nearest cluster. The neighborhood clusters will be formed according to those which have the most venue and venue categories alike.

4. Results

After the 4 neighborhoods (Dorchester, Roxbury, Mattapan and Brighton) that did not return any venue data were assigned to a venue count of 0, the resulting line plot of neighborhood bars is as follows in Figure 4

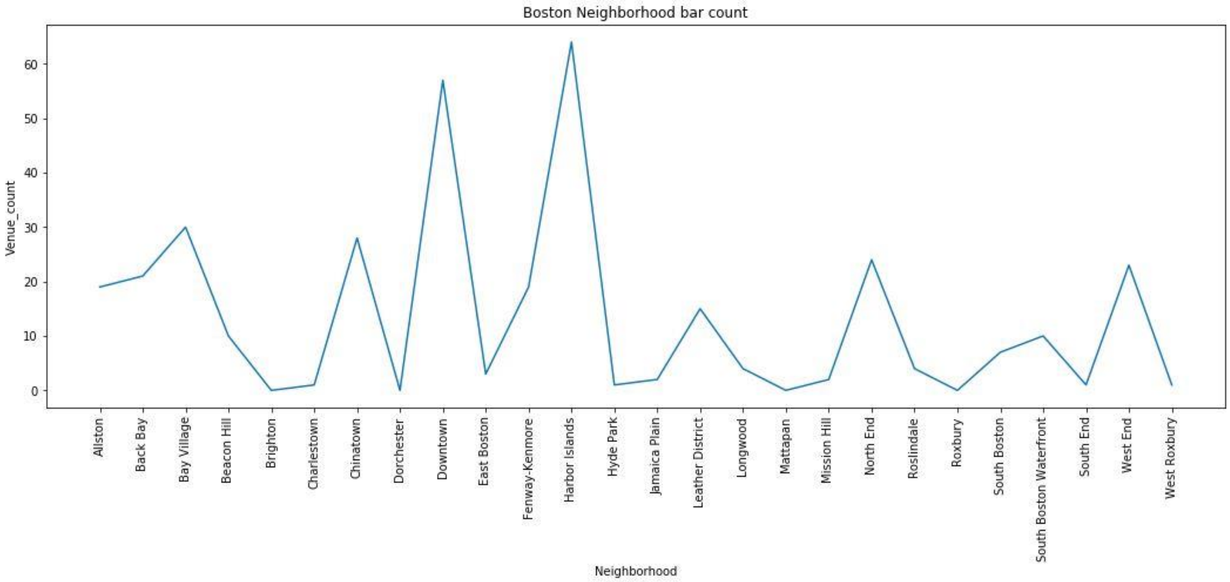


Figure 4

Figure 4 shows that the Harbor Islands neighborhood has the most bars, closely followed by Downtown.

In terms of clustering, the neighborhoods were assigned to 4 clusters through using k-means and one additional cluster to group those that did not have any venue data returned from FourSquare:

Cluster 1 (red): Mission Hill, Longwood, North End, South End, Hyde Park, South Boston Waterfront, South Boston, Allston, Harbor Islands
Cluster 2 (purple): Roslindale, Jamaica Plain, Bay Village, Leather District, Chinatown, Back Bay, East Boston, West End, Beacon Hill, Downtown, Fenway-Kenmore
Cluster 3 (blue): West Roxbury
Cluster 4 (light green): Charlestown
Cluster 5 (orange => no venue data): Brighton, Dorchester, Mattapan and Roxbury

The folium map in Figure 5 displays the neighborhoods designated into 5 clusters.

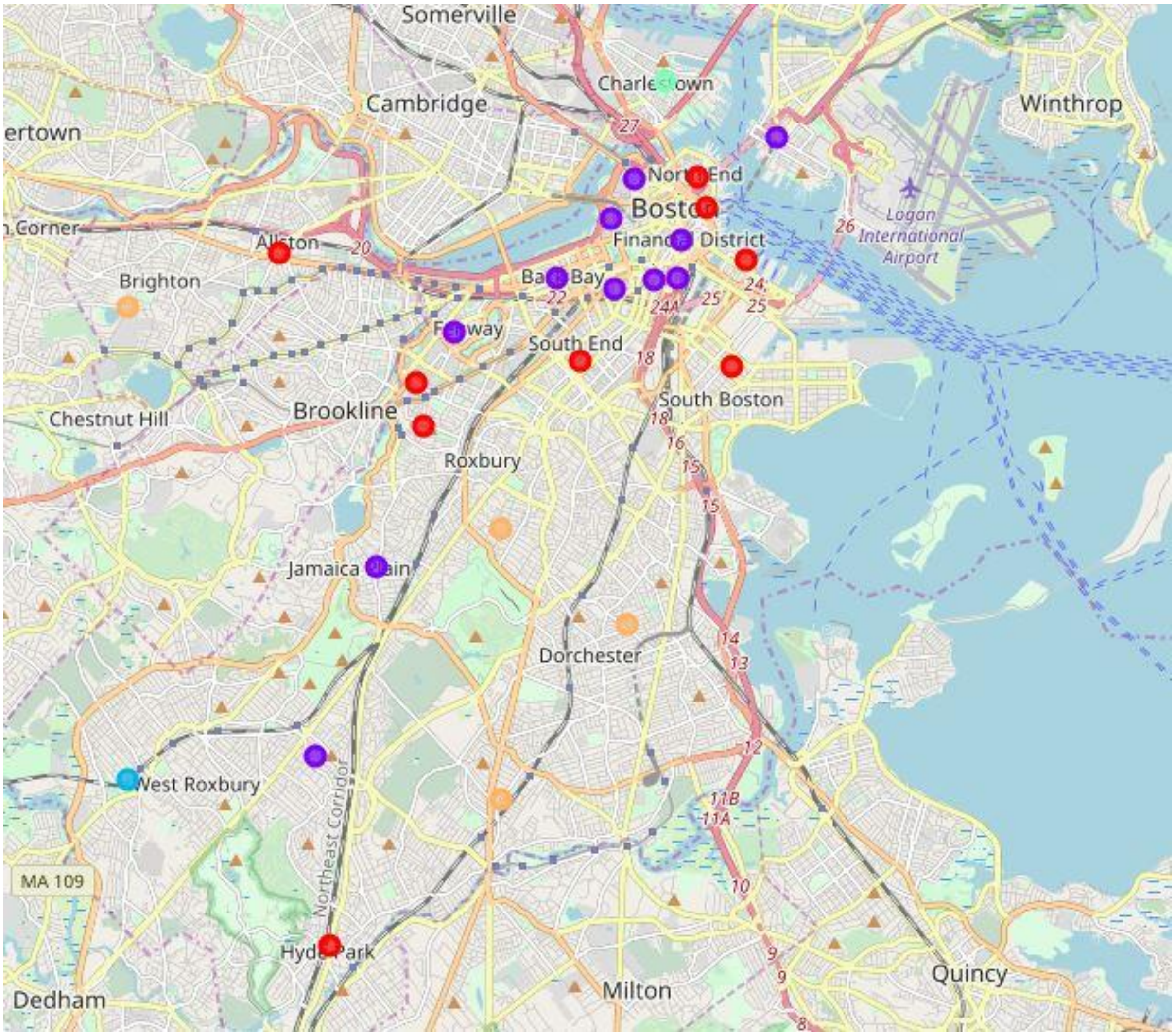


Figure 5

5. Further discussion – observations and recommendations

The available data here is not sufficient to make a decision on locating a new bar business. There are some concerns about the data returned from FourSquare and why venues were missing from 4 neighborhoods. Perhaps adjusting the search radius could help here so debug is required. Some of the original venue categories returned were not relevant and were filtered out of the subsequent analysis.

This data is one input to a final decision regarding locating a new bar. Additionally, other factors need to be considered for each neighborhood such as:

- commercial rental rate data
- demographics
- neighborhood traffic (to include tourist potential)
- median income

6. Conclusion

From relatively little input data relating neighborhoods and their locations, using the FourSquare API returns a rich treasure trove of information regarding exiting venues which make further degrees of analysis possible such as applying the k-means clustering algorithm. Valuable additional information regarding the Boston bar scene has been gathered that will serve as input in the final decision of choosing a neighborhood in which to locate the new bar.