Restaurant diversity in Boston neighborhods

Brian Schaefer

April 17, 2020

1 Introduction

Residents and frequent visitors of Boston are certainly familiar with its famous cultural centers and the food options each one provides. Popular neighborhoods including Chinatown, North End, and the Seaport District are hot spots for those looking to indulge in the city's best Chinese, Italian, and seafood options. But, where should one go for a more diverse selection of restaurants? In this study, we classify neighborhoods in Boston based on restaurant variety and identify the neighborhoods with the most diverse options.

2 Data

To classify Boston's neighborhoods, we need geographical information about the neighborhoods as well as information on the types of restaurants that exist within each neighborhood. For the former, we use geographical coordinates provided by the City of Boston outlining the boundaries of each of Boston's neighborhoods[1]. For the latter, we use Foursquare[2] to gather information on the types of restaurants nearby the geographical center of each neighborhood.

The geographical data are provided in `GeoJSON` format. The dataset consists of multiple sets of `(longitude, latitude)` coordinates that define the boundaries of each neighborhood. We adapt functions from the `geojson_utils` Python package[3] to calculate the coordinates of the centroid and area of each neighborhood given the list of boundary coordinates. Using the Python package `folium`, we can show the neighborhoods on an interactive map (Figure 1).

We use the Foursquare application programming interface (API) to search for restaurants nearby each neighborhood's geographical center using the category "Food"[4]. The Foursquare service returns

[1] https://data.boston.gov/dataset/boston-neighborhoods

[2] https://www.foursquare.com

[3] https://pypi.org/project/geojson_utils/

[4] categoryId 4d4b7105d754a06374d81259, https://developer.foursquare.com/docs/build-with-foursquare/categories/.

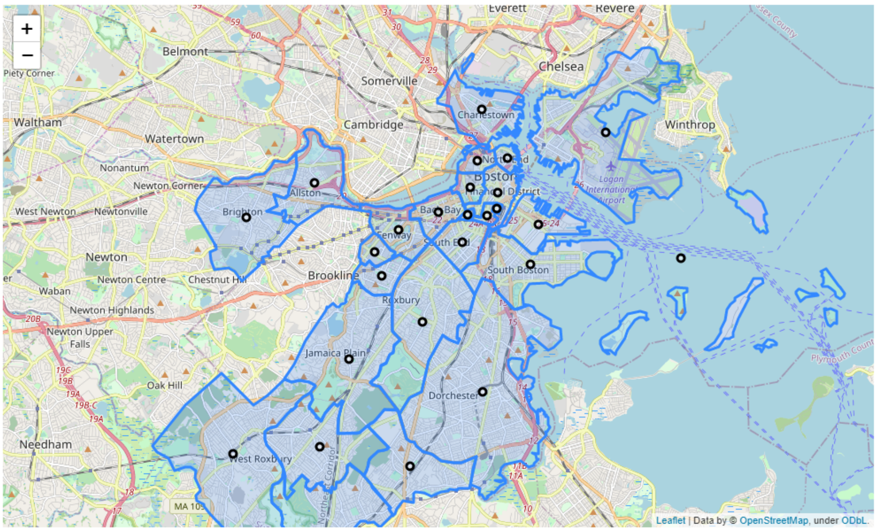


Figure 1: Map of Boston neighborhoods (blue) and geographical centers (black markers).

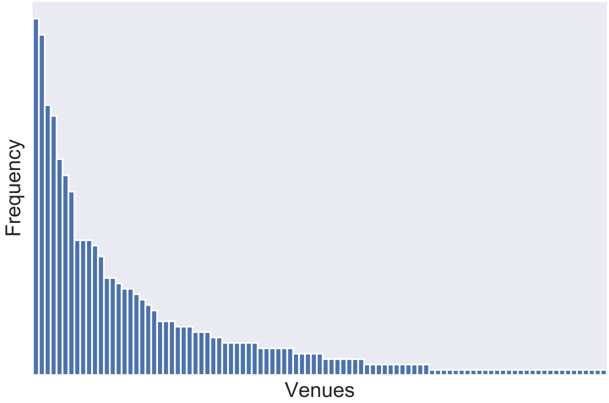


Figure 2: Frequency of venue types across all Boston neighborhoods.

the venue name, geographical coordinates, address, and category name for each of the “Food” venues found in the vicinity of each search point. Since we are interested in unique types of restaurants in each neighborhood, we take the category name for each venue and ignore venues in the same neighborhood with the same name. Figure 2 shows a summary of the relative frequencies of different types of venues across all neighborhoods. There are a few venue types that occur with very high frequency, and a relatively large number of unique venue types.

3 Methodology

We seek to classify neighborhoods based on the number of diverse restaurant types. If we simply count the number of different restaurant types, this will be biased towards neighborhoods that have

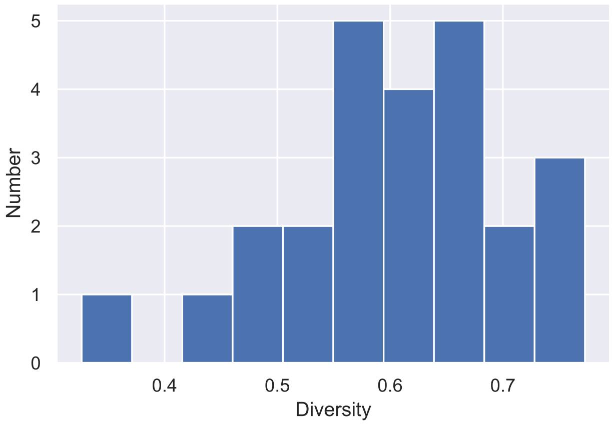


Figure 3: Diversity index for Boston neighborhoods

many restaurants in total. To account for this, we define a “diversity index“ as follows:

$$\text{Diversity} = \frac{\text{Number of unique restaurant types}}{\text{Total number of restaurants}}.$$

We plot a histogram of this metric in Figure 3, and it appears that the neighborhoods in Boston vary significantly in terms of diversity. The data approximately follow a normal distribution, with mean 0.61 and standard deviation 0.1.

To categorize the neighborhoods, we use the k-means clustering algorithm on a few different versions of the dataset. This algorithm allocates data points into each of k clusters by clustering together data points based on the similarity of the features for each data point. The result of this algorithm is a cluster number assignment, ranging from $0...k$. We use the “elbow” method (Figure 4) to determine the optimal number of clusters. By running the algorithm multiple times, with different number of clusters k each time, we can understand how increasing the number of clusters can improve the algorithm’s description of the data. By eye, we choose the point of diminishing returns (k=3 in Figure 4) at which the rate of error improvement slows down (i.e., additional clusters do not provide useful information).

4 Results

We perform three clusterings of the neighborhoods, using (1) the total number of restaurants of each type in each neighborhood, (2) the proportion of restaurants of each type in each neighborhood, and (3) the diversity index for each neighborhood.

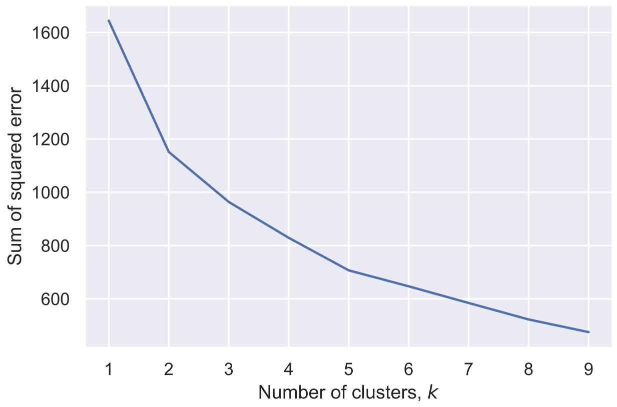
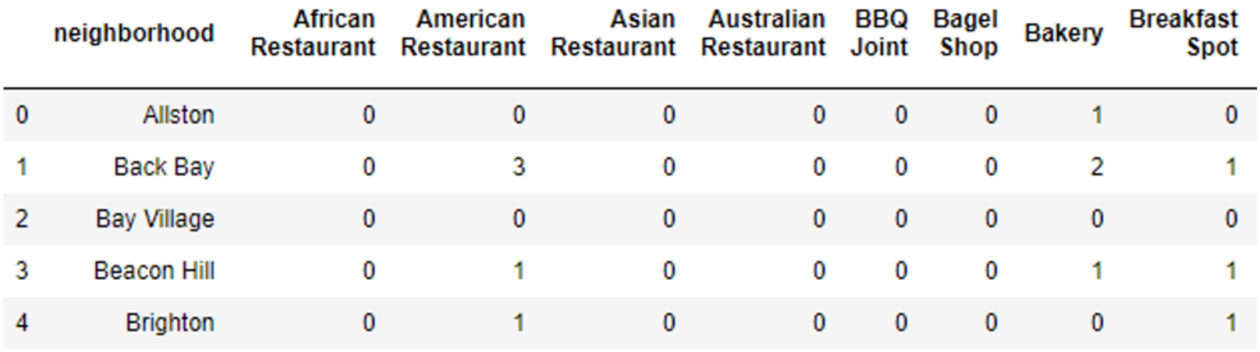


Figure 4: Elbow method for determining number of clusters

4.1 Restaurant counts

Using the Foursquare data, we total the numbers of restaurants labeled within each unique category name to arrive at a dataset of which we show in part below:



	neighborhood	African Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Breakfast Spot
0	Allston	0	0	0	0	0	0	1	0
1	Back Bay	0	3	0	0	0	0	2	1
2	Bay Village	0	0	0	0	0	0	0	0
3	Beacon Hill	0	1	0	0	0	0	1	1
4	Brighton	0	1	0	0	0	0	0	1

For a given neighborhood (leftmost column), each data point is the number of restaurants falling into the category designated by the header above. For example, in Back Bay there are three American Restaurants, two Bakeries, and one Breakfast Spot.

Performing k-means clustering to categorize the neighborhoods into three clusters, we find that the diversity index can vary substantially within each cluster (Figure 5). The first two clusters have a diversity index ranging from $\sim 0.4 - 0.7$, while the last cluster only has one neighborhood with a low diversity index. Upon inspection of the most common restaurant types, we find that cluster 0 consists of neighborhoods with mainly pizza places and cafés, cluster 2 consists of the neighborhood with a preponderance of Italian restaurants (North End), while cluster 1 is more diverse. This is a good start, as it correctly separates North End from the rest of the neighborhoods. However, using

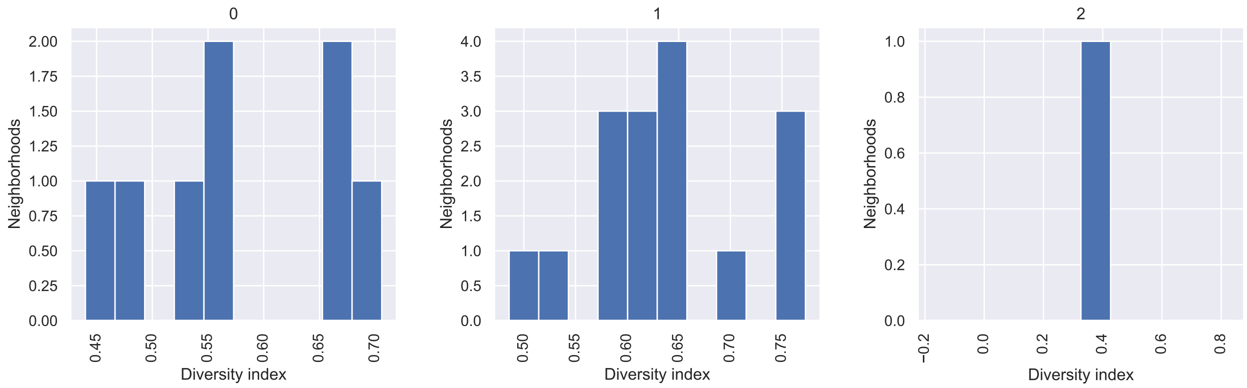
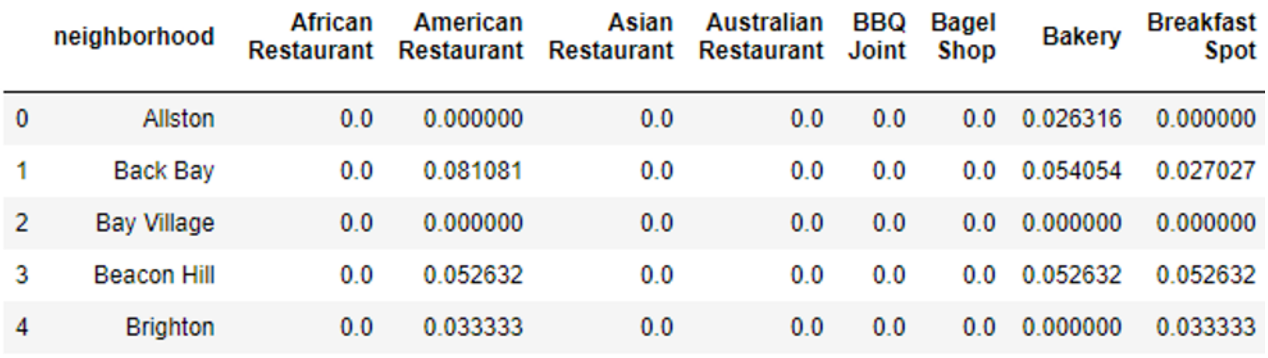


Figure 5: Distribution of diversity index within each cluster, for clustering on data consisting of the counts of each restaurant type.

just the count of restaurants of a given type does not account for the total number of restaurants.

4.2 Restaurant proportions

By dividing each row of the dataset above by the total number of restaurants in each neighborhood, we get a dataset of the proportions of restaurants of each type:



	neighborhood	African Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Breakfast Spot
0	Allston	0.0	0.000000	0.0	0.0	0.0	0.0	0.026316	0.000000
1	Back Bay	0.0	0.081081	0.0	0.0	0.0	0.0	0.054054	0.027027
2	Bay Village	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000
3	Beacon Hill	0.0	0.052632	0.0	0.0	0.0	0.0	0.052632	0.052632
4	Brighton	0.0	0.033333	0.0	0.0	0.0	0.0	0.000000	0.033333

For example, in Back Bay 8.1% of restaurants are American Restaurants.

Performing k-means clustering to categorize the neighborhoods again into three clusters, we still see variance of the diversity index within each cluster (Figure 6). Cluster 0 consists of North End and Bay Village, both of which have Italian restaurants as the most common type. Cluster 1 again consists of neighborhoods with many pizza places and cafés, and cluster 2 is the most varied. We are still not quite able to capture the clustering we want, so we move on to look only at the diversity index.

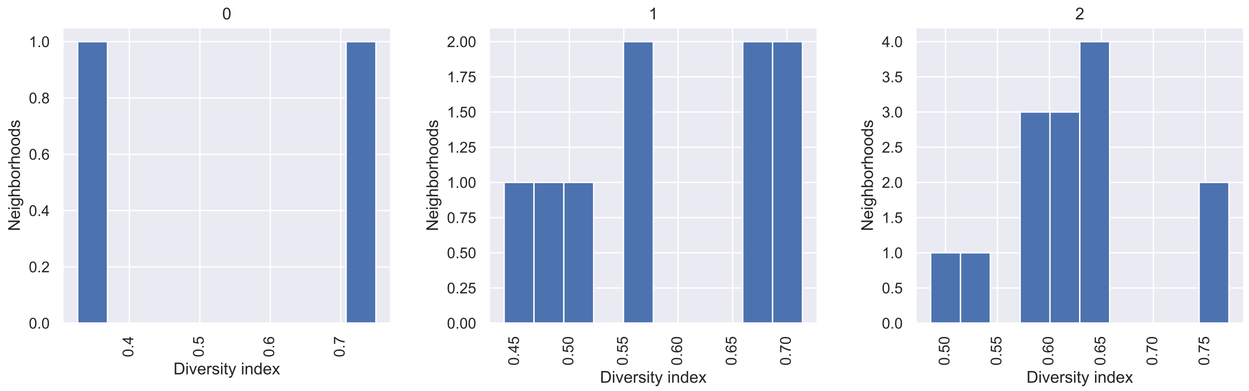


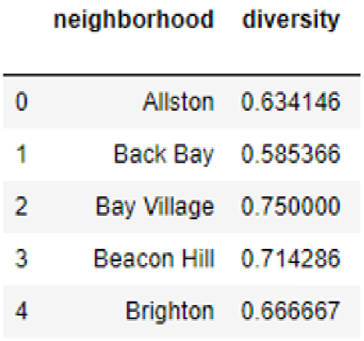
Figure 6: Distribution of diversity index within each cluster, for clustering on data consisting of the proportion of each restaurant type.

Cluster 0	Bay Village, Beacon Hill, Brighton, Mission Hill, ...
Cluster 1	Allston, Back Bay, Charlestonwn, Dorchester, ...
Cluster 2	Chinatown, North End, Longwood, Mattapan

Table 1: Neighborhoods clustered on diversity index

4.3 Diversity index

We use the diversity index to form a much more limited dataset that already captures the gist of the question we set out to answer:



	neighborhood	diversity
0	Allston	0.634146
1	Back Bay	0.585366
2	Bay Village	0.750000
3	Beacon Hill	0.714286
4	Brighton	0.666667

Performing k-means clustering, this time we find that each cluster contains an independent range of diversity index values (Figure 7). Cluster 0 (diversity index > 0.65) is the most diverse, while cluster 2 (diversity index < 0.5) is least diverse. Table 1 summarizes the clustering. The least diverse cluster now contains both Chinatown and North End, as expected for famous cultural centers. The other two neighborhoods are unexpected, but make sense: Longwood is mainly a medical campus, which we would not expect to host diverse food options, and Mattapan has a large Afro-Caribbean influence, so many of the restaurants are expected to be Caribbean or African.

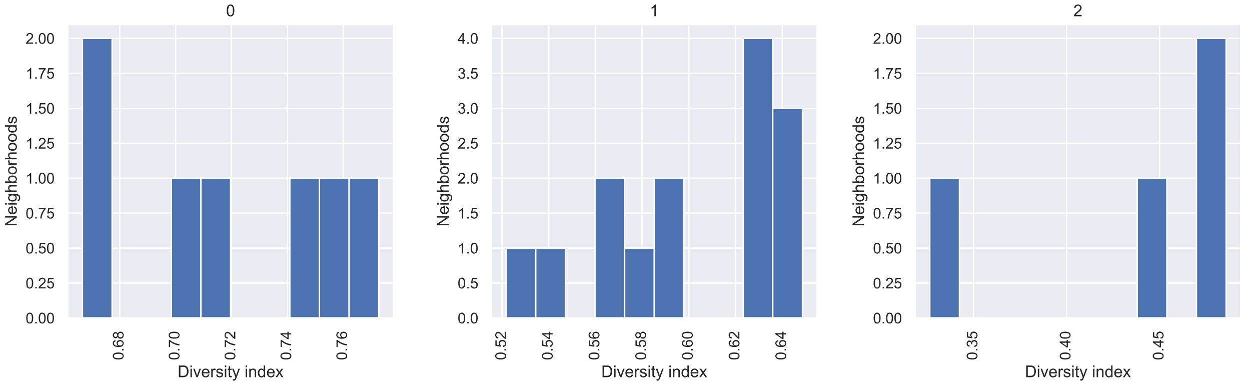


Figure 7: Distribution of diversity index within each cluster, for clustering on only the diversity index.

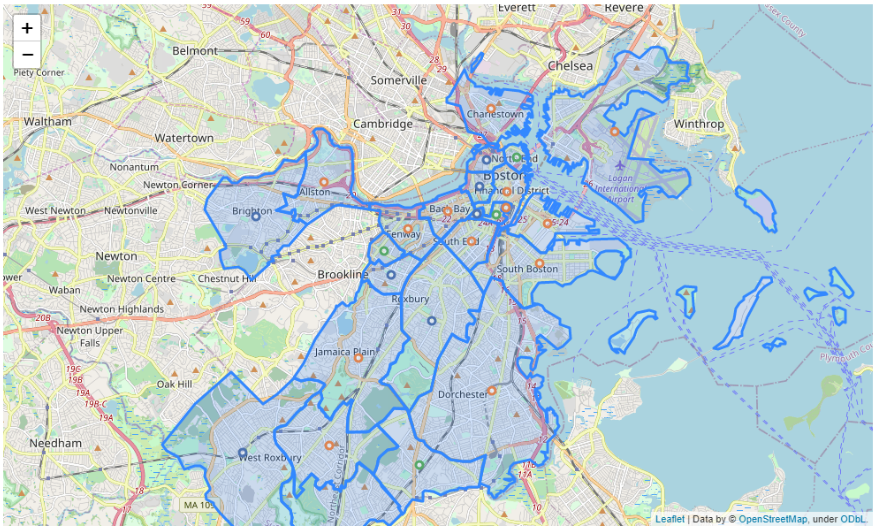


Figure 8: Map of Boston neighborhoods (blue outlines) and geographical centers sorted by cluster (green: least diverse, orange: moderately diverse, dark blue: most diverse).

5 Discussion

The final clustering of neighborhoods makes sense based on common knowledge of Boston's most famous neighborhoods. Are there any patterns in the geographical location of the most diverse food neighborhoods in Boston? In Figure 8, we show a map of the neighborhoods color-coded by cluster. Interestingly, we notice a string of three diverse neighborhoods on the west side of the city center, while the neighborhoods on the east side are less diverse. We also notice a few diverse neighborhoods further from the city center.

6 Conclusion

We clustered neighborhoods in Boston based on the diversity of the restaurants nearby the geographical centers of each neighborhood. As expected, famous cultural centers like Chinatown and North End are deemed some of the least diverse cities, while other neighborhoods like Back Bay are deemed much more diverse. We have compiled a list of neighborhoods offering a variety of dining options that would be great spots for both visitors and new residents to Boston with an adventurous palate.