Data Analysis and Visualizations Results

Clerkenwell vs Back Bay

Kevin Bateni

Introduction and Background

1. Introduction

Which neighborhood is better? Thats is a very subjective question, with a different answer per person. In this post, I am attempting to quantify the answer to this question, by comparing 2 neighborhoods I am familiar with.

The first is Clerkenwell section of London, United Kingdom, and the second is the Back Bay section of Boston, United States.

2. Background

London and Boston are 2 major metropolitan areas with large residential, commuter, and tourist populations. To compare the 2 cities would be difficult, due to the large differences in size (London has a population of 8.9 million as of 2018, while Boston has an estimated population of 692k as of 2019).

To normalize the size, and therefore the count of venues, which will be the primary mechanism for this analysis, we will focus on the two aforementioned neighborhoods.

Background Facts

First, some key facts on the two neighborhoods:
Back Bay has a population of 17,577 (as of 2015). It was officially recognized as a neighborhood of Boston, built on reclaimed land in the Charles River basin, of which construction began in 1859.
The geographical coordinates of Clerkenwell are 51.5237268, -0.1055555.

Clerkenwell as a population of 11,490 (as of 2011). The first known reference to the neighborhood is from 1100, so it has a very long history. The geographical coordinates of Clerkenwell are 51.5237268, -0.1055555

Data

Using the Foursquare API to understand the venues that were located in both of the neighborhoods, I queried for the top 100 venues which were 1200m from the center from their the latitude and longitude locations.

Latitudinal and longitudinal information for Back Bay and Clerkenwell were obtained from the geolocator package in Python. Additional population related data and demographics were obtained from wikipedia pages for Clerkenwell and Back Bay.

Here is a view of the first 5 Venues and category, information obtained using Foursquare API:

Clerkenwell Venues

	name	categories	address	lat	lng
0	The Zetter Townhouse	Hotel	49-50 St John's Sq	51.522849	-0.103658
1	Sushi Tetsu	Sushi Restaurant	12 Jerusalem Passage	51.523348	-0.104015
2	Granger & Co.	Breakfast Spot	50 Sekforde St	51.523504	-0.104629
3	BrewDog Clerkenwell	Beer Bar	45-47 Clerkenwell Rd	51.522401	-0.103835
4	Great Bakery & Deli	Bakery	167-169 Farringdon road	51.524410	-0.110076

Back Bay Venues

	name	categories	address	lat	lng
0	Gre.Co	Greek Restaurant	225 Newbury St	42.349920	-71.081633
1	sweetgreen	Salad Place	659 Boylston St	42.349995	-71.078668
2	The Lenox Hotel	Hotel	61 Exeter Street at Boylston	42.349229	-71.079528
3	Commonwealth Avenue Mall	Park	Commonwealth Ave.	42.351887	-71.080033
4	Boston Marathon Finish Line	Athletics & Sports	560 Boylston St	42.349842	-71.078691

Methodology 1/2

Using the Foursquare API to get the venue data from each neighborhood I queried for the top 100 venues which were 1200m from the center from their the latitude and longitude locations, obtained from the geolocator package in Python.

Now let's look at the data. Firstly we grouped the data by category type and compared the 2 neighborhoods.

Top Venues Categories

Clerkenwell Venues			Back Bay Venues	
Category	Count		Category	Count
Coffee Shop	12		Clothing Store	5
Gym / Fitness Center	6		Coffee Shop	5
Pub	5		Seafood Restaurant	4
Bar	4		Italian Restaurant	4
Wine Bar	3		Ice Cream Shop	4
Falafel Restaurant	3		Hotel	4
Beer Bar	3		American Restaurant	4
Park	3		Spa	4
Hotel	3			
Food Truck	3			

Methodology 2/2

Now let's look at the locations of the venues by plotting them on the map of each city. This visualization allows us to compare the coordinates and placement of each venue to have a different comparison than count.

As you can see by the below figures, the venues in (Fig. 5) Back Bay are located together in a more uniform manner, than those in Clerkenwell (Fig. 4). However Clerkenwell has a major cluster at the top of the map.

Both neighborhoods have a centralized cluster of venues along one street, which indicates a commercial center. This could be good for you if you wanted was access to venues from your home, but not necessarily living right next to them.

Fig 4. Folium Map Plot of Venue Locations — Clerkenwell

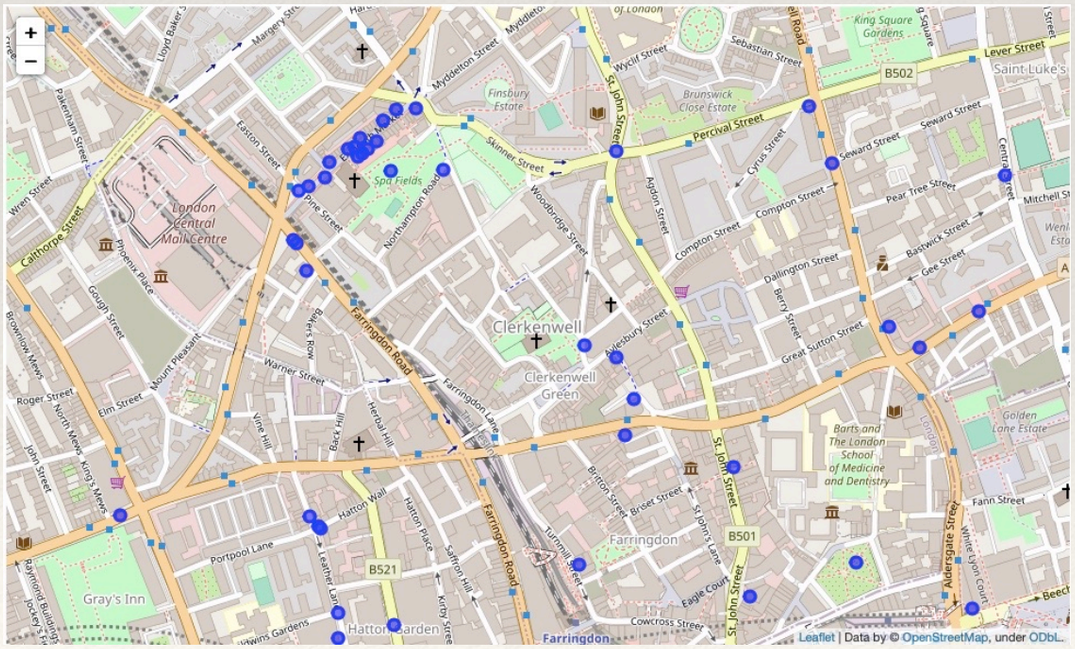
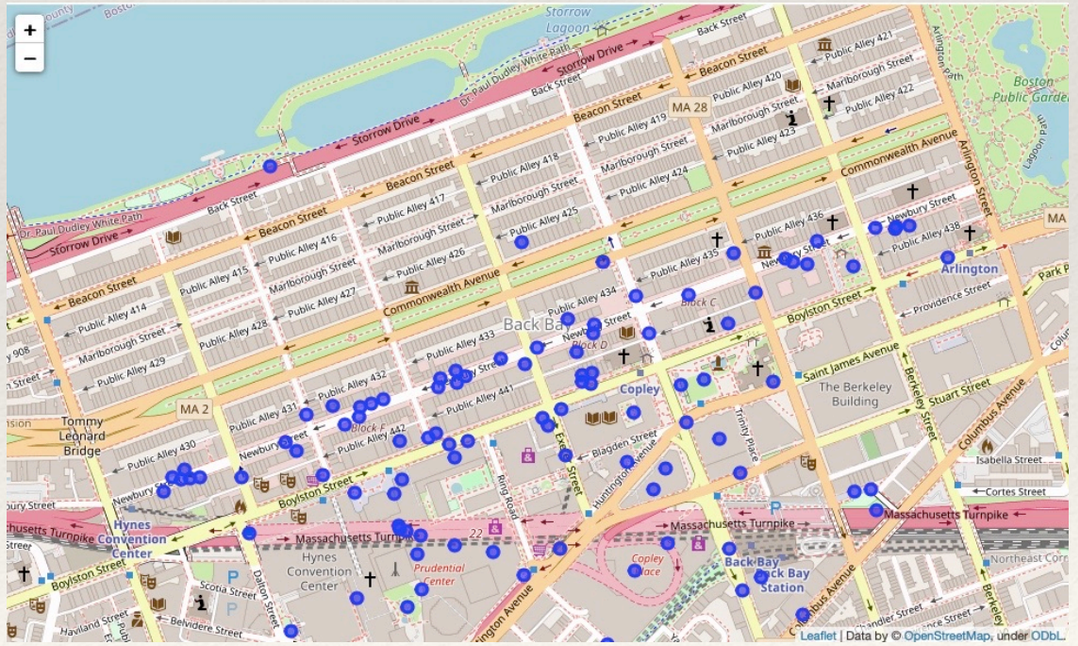


Fig 5. Folium Map Plot of Venue Locations — Back Bay



Results and Discussion 1/3

This brings us to the observation, which neighborhood has a more dispersed population of venues? Let's look at the data to tell us the story. The Foursquare API data has a distance from the center of the neighborhood for the coordinates we entered when doing our search.

Using this data point, we will calculate some descriptive statistics on the distance from the neighborhood centers. We calculate the standard describe statistics using Python's Pandas package. The table in Fig. 6 shows these statistics. As you can tell, Clerkenwell has the highest mean, standard deviation, percentile distribution, and maximum distance between the two neighborhoods.

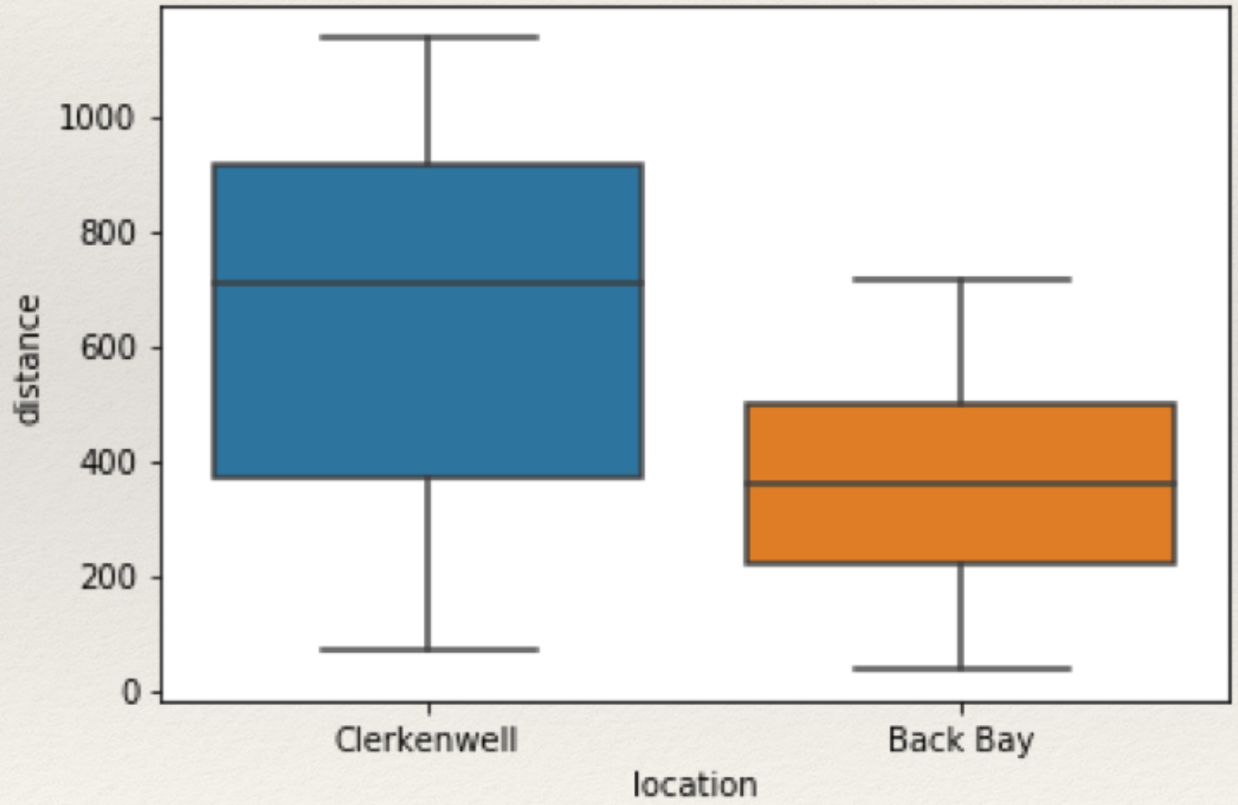
Fig 6. Descriptive Statistics of Venue Distance from Neighborhood Center — Clerkenwell and Back Bay

	Clerkenwell	Back bay
count	100.00	100.00
mean	659.88	359.26
std	279.75	174.94
min	68.00	37.00
25%	369.25	218.25
50%	709.00	359.50
75%	912.25	498.50
max	1135.00	716.00

Results and Discussion 2/3

But is this really the whole story? Let's try another visualization. Using a box plot chart in Python's Seaborn package we are able to see the distribution of the venues. What does the visualization tell us about the venues? The larger blue box for Clerkenwell tells us that the venues in that neighborhood are highly dispersed in their locations.

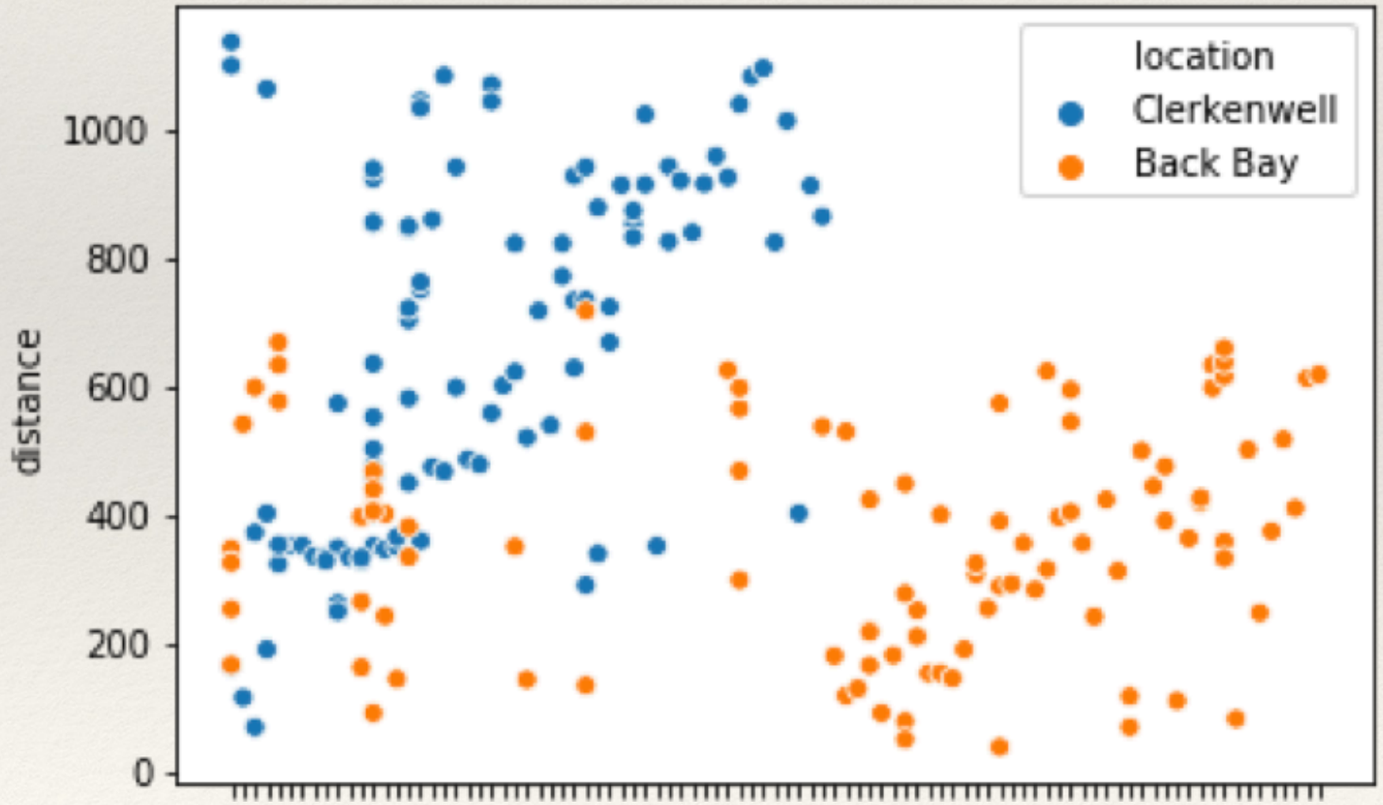
Fig 7. Boxplot Chart of the Venue Distance from Neighborhood Center — Clerkenwell and Back Bay



Results and Discussion 3/3

But this is not all we are able to do. I also performed a more frequently utilized chart in Python's Seaborn package, a scatterplot chart to further investigate the distribution of the venues. Again, the location of the plots for the Back Bay venues are closer to the X axis, indicating a smaller distance from the neighborhood center, additionally they are more uniformly aligned. These three analysis's allow us to say that indeed, Clerkenwell's venues are much more dispersed from the center.

Fig 8. Scatterplot Chart of the Venue Distance from Neighborhood Center — Clerkenwell and Back Bay



Cluster Modeling 1/2

So given these disperse observations of the venue locations, are there other characteristics that we cannot see that would indicate some other correlation between the two neighborhoods? Using machine learning, and specifically the K-Means Clustering model we attempt to ascertain this.

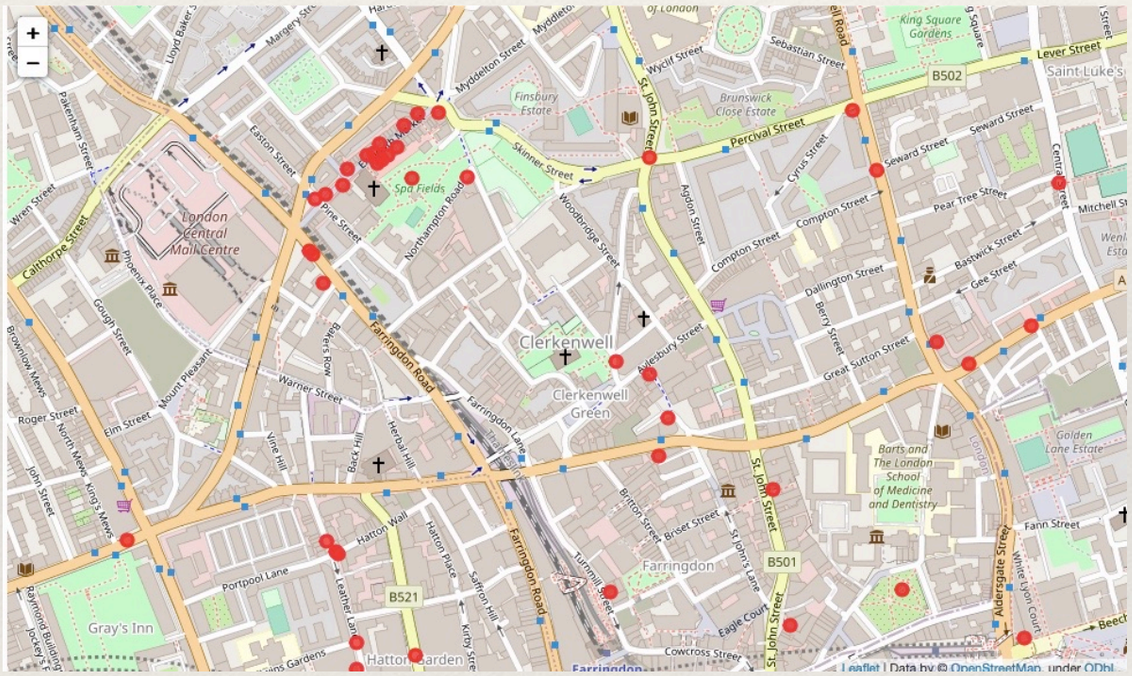
To prepare the model, we first utilize the one hot encoding method to prepare our dat for fitting the model. We combined the venue data and used the location as the label for each neighborhood. However when we fit our model, we did not use the neighborhood as a parameter to avoid influencing the outcome.

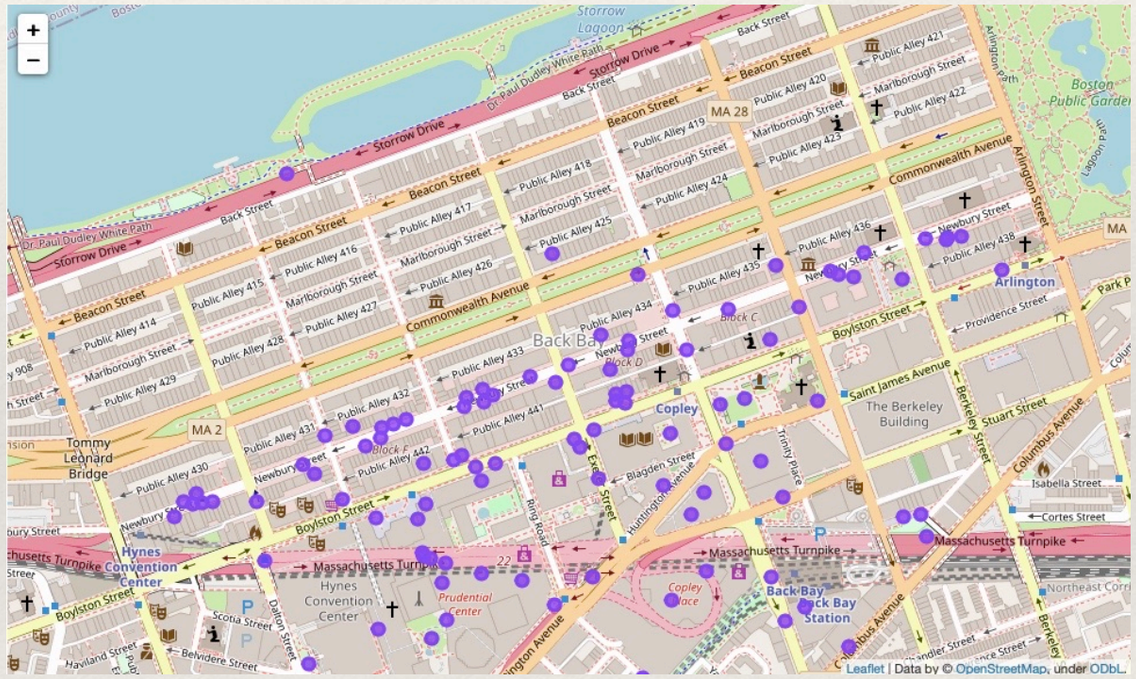
The results ultimately produced two clusters, and aligned all of the Clerkenwell venues into 1 cluster and all of the Back Bay venues into another cluster.

Cluster Modeling 2/2

To visualize this we replotted the venues using a color coded scheme to identify the clusters:

Fig. 9 and 10 Folium Maps of the Venues Clusters using The output of the model I built— Clerkenwell and Back Bay





Conclusion

In conclusion, using a number of different statistical analysis's, visualizations, and building a clustering model, we are able to conclude that the two neighborhoods, Clerkenwell and Back Bay, are very different from each other in terms of location of venues, distribution of venues and ultimately clustering of the venue types. Tis does not inherently mean that one is better than the other, but rather that Clerkenwell venues are located further and more dispersed from the center, and Back Bay venues are not.