IBM DATA SCIENCE CAPSTONE

Supercities: London and New York
Melting Pot or Unique Diversity?

Juan Trujillo

Introduction

London and New York are often described as two of the most cosmopolitan cities in the world.[1] They also compete for the title of world's financial hub,[2] and are often high on the polls around the best place to live.[3]

Both have a very similar population (ca. 8.5 million inhabitants) and similar surfaces (London has 1572 square kilometers vs New York 1214 square kilometers).

They are two of the most vibrant and thriving urbs in the world and they share a common official language. Beyond these similarities, what degree of overlapping exists between New York and London if we look at their distribution of amenities, shops, bars and other elements of the urban orography? How similar are their neighbourhoods if we would combine them into a single super-city? Are their city areas and the distribution of services and venues homogeneous to a point where they can be traceable in their counterpart?

This examination proposes an empirical approach to verify a theory of gentrification that tends to see a continuous trend towards the amalgamation and blending of highly developed cities into 'melting pots' that compose diversity into homogeneity.[4]

[1] https://www.worldatlas.com/articles/the-most-cosmopolitan-cities-in-the-world.html
[2] https://uk.reuters.com/article/uk-britain-banks/brexit-helps-new-york-take-top-finance-spot-from-london-survey-idUKKCN1SY08L
[3] https://www.businessinsider.com/new-york-city-vs-london-comparison-2015-5?r=US&IR=T
https://www.bloomberg.com/opinion/articles/2019-06-04/why-london-is-better-than-new-york
[4] https://en.wikipedia.org/wiki/Melting_pot

Fig.1 Comparison of both cities across 6 dimensions (Demographics, Quality of Living, Environment, Leisure, Transportation, General Info) Source: https://versus.com/en/london-vs-new-york

Problem Description

As part of this project, we plan to combine data obtained from the Foursquare Places API surveying the places and offerings in the different neighbourhoods in New York and London as if they were a single super-city.

We will obtain information about all the neighbourhoods in New York and all tube stations in London in order to cluster similar areas between both cities across the Atlantic.

We want to identify the commonalities and singularities in terms of activities that these cities offer in each one of their neighbo(u)rhoods. We want to understand if there are neighbourhood traits that are unique to one of the cities or whether there are commonalities to certain areas in both. Is the Soho in London similar in its places to its eponymous in New York?

Can you obtain the same services in Chelsea and the Upper East Side? Are hip cafés more pervasive in Shoreditch or in the Meatpacking District?

A friend is planning to move to New York and has her eyes in moving to Battery Park. She is currently living in Lewisham in London. Is there much resemblance in the services and offerings of these two areas? What are the types of places she will be finding more often in this new neighbourhood? Will she find he favourite tandoori near her new home?

Data Description

In the case of New York, we have a well-documented list of neighbourhoods grouped in community boards.[5] We will utilise the data provided by Cognitive Class[6] that already includes geo-location information to determine the data points that will be used to cluster the information.

London is composed of 32 wards, however, these are large geographical areas that can span over a hundred square kilometres (e.g. Bromley or Hillingdon). In order to be more granular, we will use the pervasive and entrenched nominal usage of tube stations in London. With over 300 stations, Londoners tend to refer to their closest station when describing their neighbourhood. Transport For London (TFL) provides a data set containing information about all their stations including their geolocation. The information is provided in the form of kml and xds files.[7] It is true that this choice brings some bias to the study since there are parts of London that are not well connected to the tube, particularly the south. On the other hand, this approach will offer additional granularity and will cover the

[5] https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City
[6] https://cocl.us/new_york_dataset
[7] https://api-portal.tfl.gov.uk/docs

most densely populated and representative areas in the nowadays old Roman settlement of Londinium.[8]

We will use the Foursquare Places API[9] to describe the sites and venues in these neighbourhoods and we will use the data provided in order to cluster the neighbourhoods across both cities as if we were handling a contiguous territory. We will limit the number of venues due to limitations in the API to 200 within 1,000 meters of each neighbourhood based on its geolocation.

With these three sets of data (COCL New York neighbourhoods, TFL London tube stations, and Foursquare Places API) we will review the most common venues in a cluster of this super-city to review how much similarity there is across the neighbourhoods of both cities.

Methodology

K-means is a popular algorithm in the category of unsupervised learning that minimizes dissimilarity in a data subset. However, results can vary depending on the random location of the initial centroids.

To prevent this, the experiment has been executed 10 times using the Sci-Kit Learn library and the clustering K-means module.[10] k-means++ has been used to initialize the centroids. In all instances, the results were very similar.

The other main decision to be made when using K-means is the number of centroids initialised. The 'elbow method' is often used to identify the right number of centroids as a function of the reduction of the cost function, calculated as the sum of the squared distances. In this particular case, this method is not fully conclusive since there is no critical inflexion but rather a relatively smooth and continuous descent in the cost function in relation to the number of Ks.

[8] https://en.wikipedia.org/wiki/Londinium
[9] https://developer.foursquare.com/places-api
[10] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html K-means ++ ' selects initial cluster centers for k-mean clustering in a smart way to speed up convergence'.

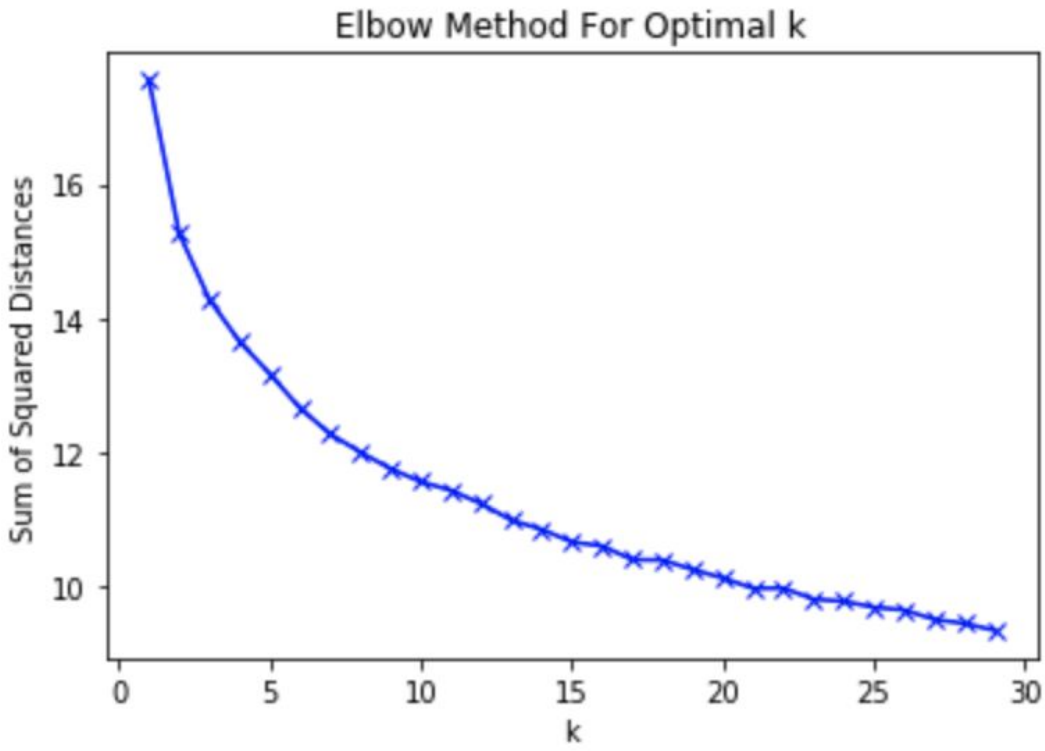

Fig. 2 Result of Applying the Elbow Method

After applying this method to calculate the optimal K, K=10 was the parameter utilised since it provided with a balanced relation between the number of clusters and cost function whilst allowing visualising the information when plotted on a map.

The data utilised to compare the different locations in these two cities was the FourSquare Places API. This API provides with information about the venues and sites within a radius to the selected locations. The radius was limited to 1,000 meters and the number of venues was restricted to 200.

With these criteria, we still obtained over 500+ 'place types', also called venues' categories. This included from Tiki bars, fountains, tailor shops or volleyball courts among a wide array of places. A full list can be found in the Jupyter Notebook available on Github.[11]

[11] https://github.com/pazamorta/Coursera_Capstone/blob/master/Supercity%20Clustering%20-%20London%20vs%20New%20York.ipynb

For each one of the 600 locations, we performed a one-hot encoding for each place type and calculated the frequency of each venue category per location. We ended up with a dataframe of 600 rows with 511 types of venues across the super-city.

Finally, we used the k-means algorithm to cluster the 600 observations based on their features. We plotted these 10 clusters on the maps of New York and London using the Folium library, generated pie charts that check the number of observations per city to establish the overlapping between these areas, and highlighted the most frequent venue type across the locations identified to provide us an intuitive idea of the predominant type of venue in a given cluster.

Results

After aggregating the cities and clustering them into 10 groups based on the 511 venue types identified across both cities according to the data obtained from Foursquare Places API, no significant overlapping across the neighbourhoods of both cities is evidenced.

A full analysis can be found in the Jupyter Notebook in Github

https://github.com/pazamorta/Coursera_Capstone/blob/master/Supercity%20Clustering%20-%20London%20vs%20New%20York.ipynb

The neighbourhoods tend to be clustered within the same city, hinting at a similar configuration of venues and establishments prototypical of certain areas of both cities.

Out of the 10 clusters identified, 8 are exclusive to each one of the cities (London: Cluster #0, Cluster #3, Cluster #5 - New York: Cluster #2, Cluster #4, Cluster #6, Cluster #8, Cluster #9). These clusters are characteristic to either city.

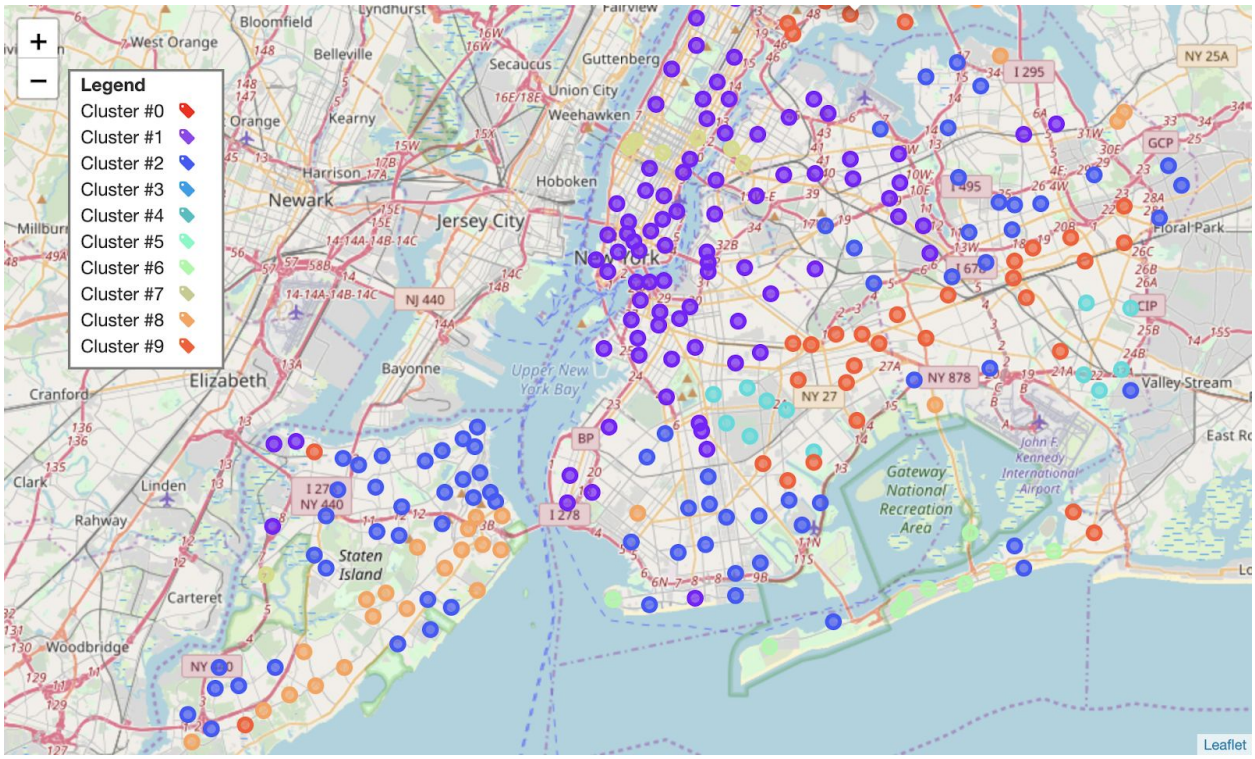

Fig.3 New York Clusters plotted on a map using Folium

Among the 2 other clusters, there is also little overlapping: in Cluster #1 London neighbourhoods have an 8.1% representation. The inverse happens in Cluster #7 where New York neighbourhoods have slightly over 7.6% of the total.

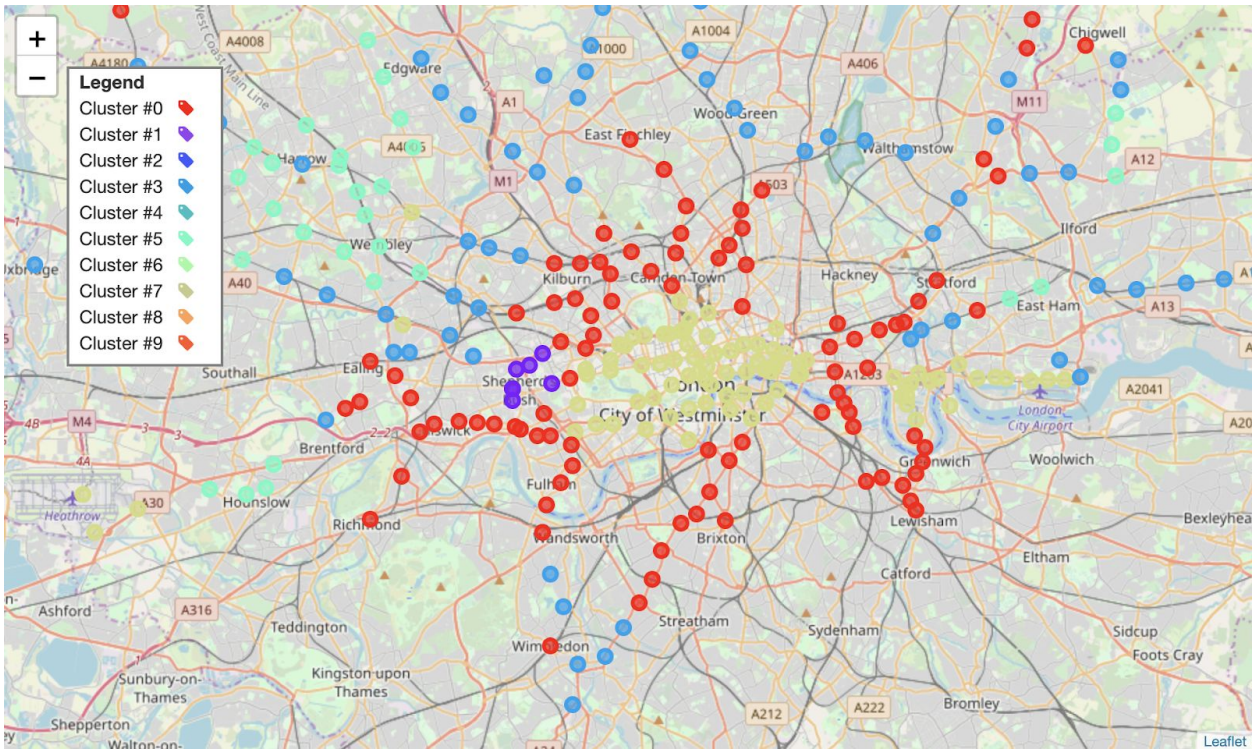

Fig.4 London Clusters plotted on a map using Folium

A prototypical categorization of these clusters based on the most common type of venue and their placement within the city would allow us to differentiate these 10 clusters in the following areas that I have labeled based on the venues just with the intention to provide a memorable subject to the cluster.

Clustering of Neighbourhoods

1. Cluster #0 - The Suburban Posh - Exclusive to London

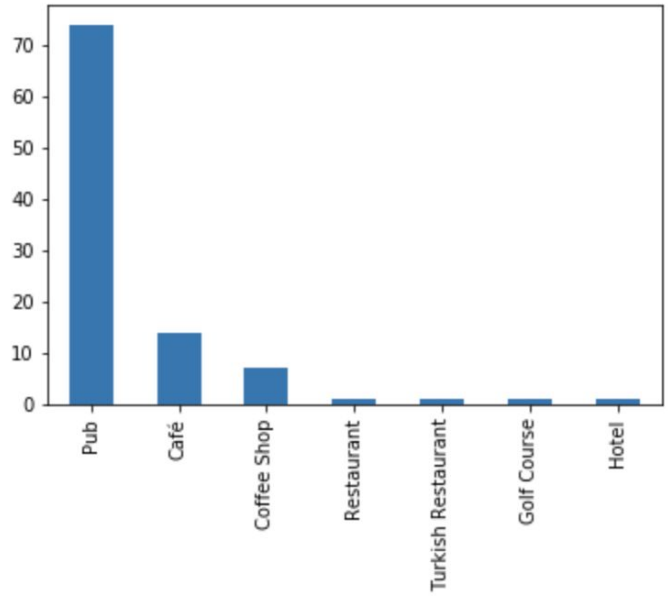

Fig. 5 Aggregated values for the most common venue for all neighbourhoods in Cluster #0

2. Cluster #1 - The Resourceful Center - Prototypical New York with 8.1% London neighbourhoods

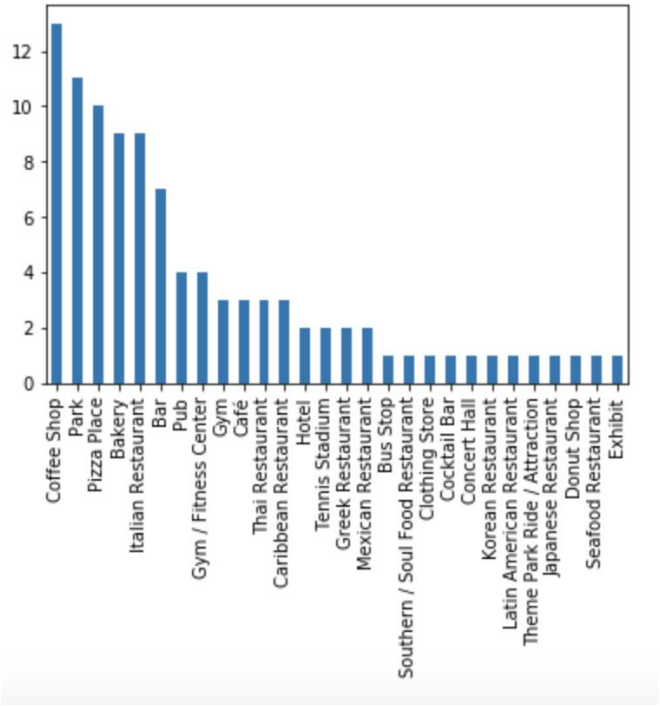

Fig. 6 Aggregated values for the most common venue for all neighbourhoods in Cluster #1

3. Cluster #2 - The Italian Quartiere - Exclusive to New York

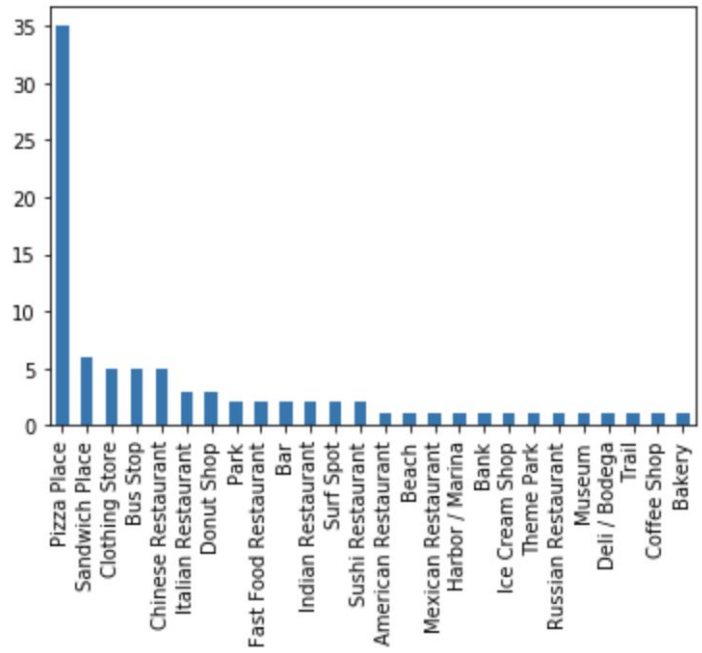

Fig. 7 Aggregated values for the most common venue for all neighbourhoods in Cluster #2

4. Cluster # 3 - The Suburban Comfy - Exclusive to London

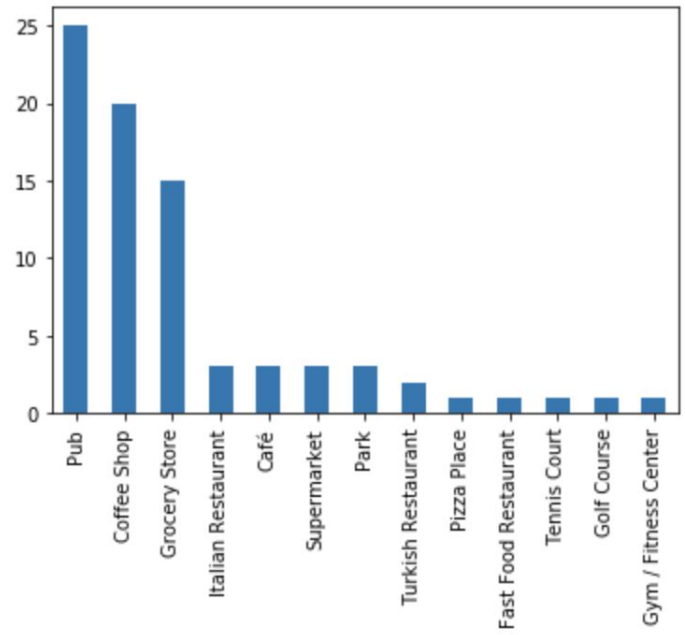

Fig. 8 Aggregated values for the most common venue for all neighbourhoods in Cluster #3

5. Cluster #4 - The Barrio Outskirts - Exclusive to New York

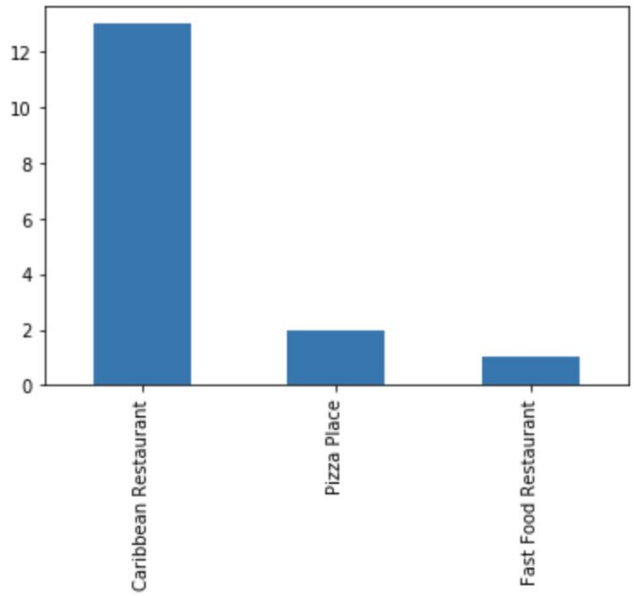

Fig. 9 Aggregated values for the most common venue for all neighbourhoods in Cluster #4

6. Cluster #5 - The Residential Diverse - Exclusive to London

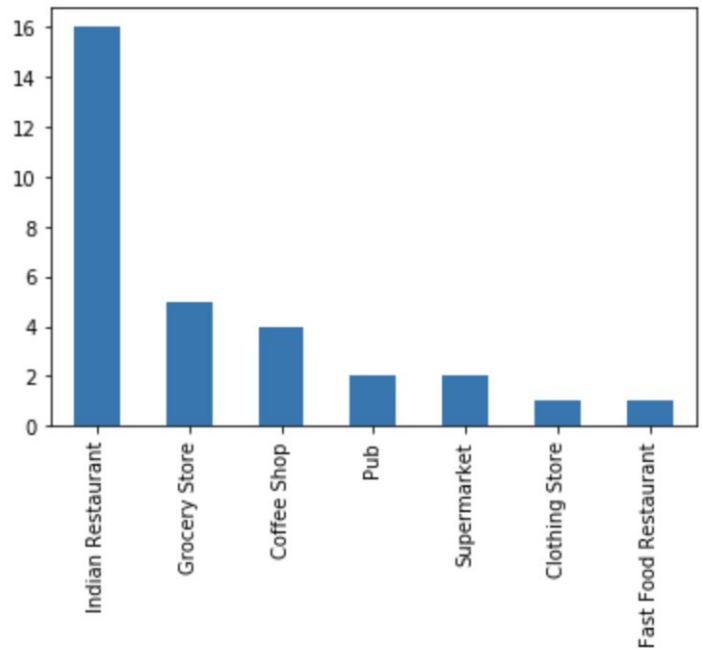

Fig. 10 Aggregated values for the most common venue for all neighbourhoods in Cluster #5

7. Cluster #6 - The Coastal City - Exclusive to New York

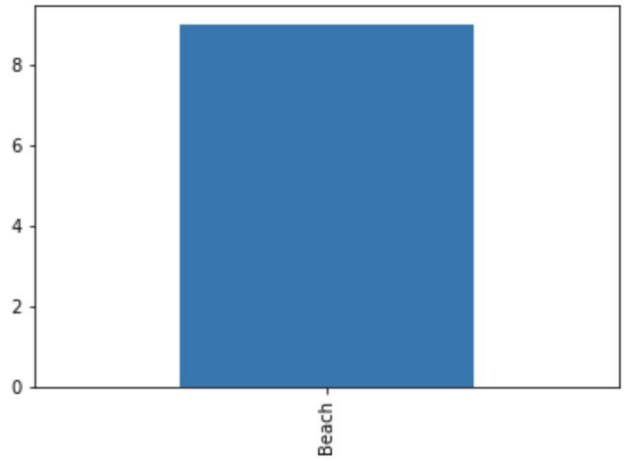

Fig. 11 Aggregated values for the most common venue for all neighbourhoods in Cluster #6

8. Cluster #7 - Theatres and Turistic - Prototypical London with 7.6% of New York neighbourhoods

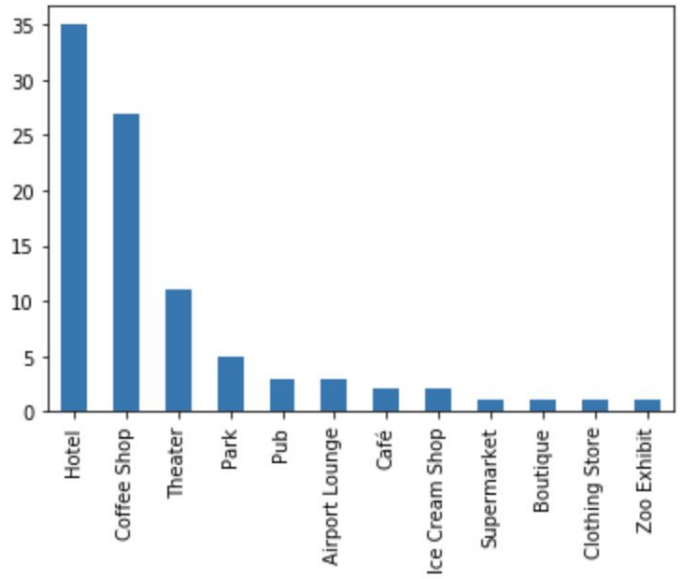

Fig. 12 Aggregated values for the most common venue for all neighbourhoods in Cluster #7

9. Cluster #8 - The Metropolitan Fringes - Exclusive to New York

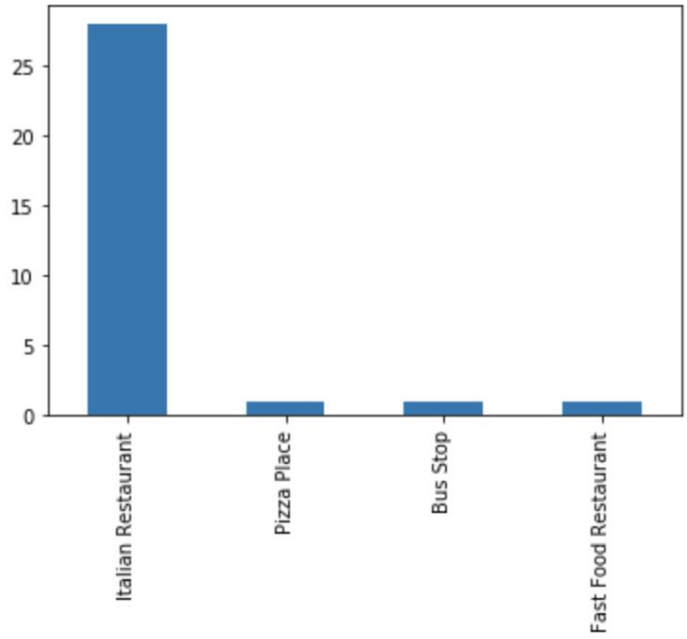

Fig. 13 Aggregated values for the most common venue for all neighbourhoods in Cluster #8

10. Cluster # 9 - The Modest Suburb - Exclusive to New York

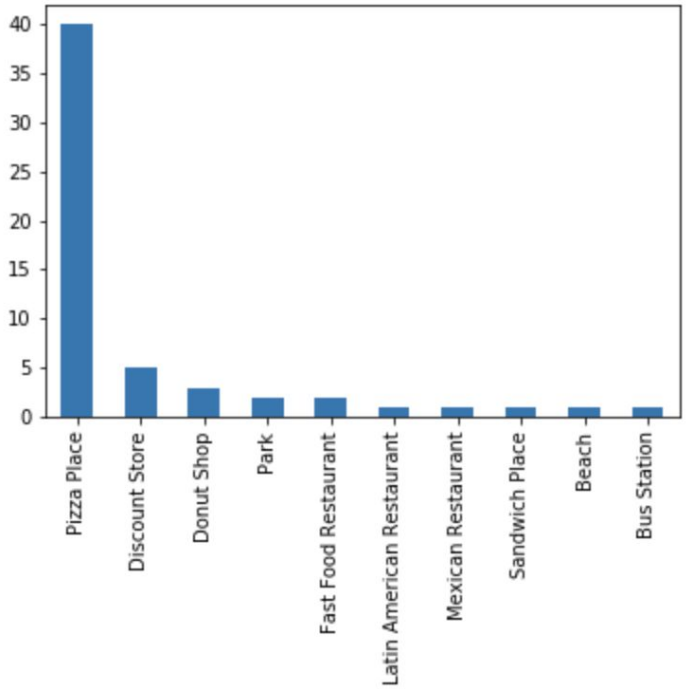

Fig. 14 Aggregated values for the most common venue for all neighbourhoods in Cluster #9

Discussion

The results clearly indicate a lack of overlapping between neighbourhoods in London and New York. According to the k means clustering algorithm applied to the Foursquare Venue data, each city has characteristic constellations of venues that are not present in its counterpart.

The k means clustering algorithm is one of the most popular unsupervised learning algorithms. Clustering attempts at grouping similar data points based on Euclidean distance and variance minimisation.

As part of this analysis, we utilised the k means algorithm in the SciKit Learn library. The main parameters that have an impact on the outcome of this algorithm are the number of clusters and its initialisation, the number of iterations and the distance, and the variant of the algorithm used.

We attempted the exercise with 3 different ks (7, 10, 15) and k-means++ as a way to attain quicker convergence of the centroids. Different random states for centroid initialization were tested. The end result always had a similar degree of separation between the clusters generated for both cities with minimal overlapping between neighbourhoods across the two cities.

A future direction of investigation to validate these results could include the usage of a different data set to determine the neighbourhoods in London. The TFL data is not homogeneous around population or venue density. It also does not comprise similar areas in extension. On the other hand, intuitively, metro stations tend to provide an indirect indication of density and we would tend to assume a certain correlation between distance in tube stations and reduction in the population and venue density.

The TFL data is biased towards certain areas being more transited and there are areas of London without TFL transport and connected via train stations. However, other patterns to define neighbourhoods might bring their own biases. Recent articles have disreputed other approaches to geospatial analysis, specifically those based on zip/postal codes.[12]

[12] https://carto.com/blog/zip-codes-spatial-analysis/ "ZIP Codes, that they do not represent an actual area on a map, but rather a collection of routes that help postal workers effectively deliver mail."

Conclusion

The examination of a virtual super-city composed by the New York neighbourhoods and London areas around the Tube stations has shown that each one of these urbs has a very unique flair in terms of the venues and commercial cum entertainment environments that they exhibit.

A preliminary conclusion indicates hints at how different cities generate habitats that are specific to those cities and that can span across multiple neighbourhoods within the same city but can rarely be found in other cities.

Despite the apparent demographic similarity of these two cities, the configurations of their neighbourhoods in terms of locales and services' offerings are specific to each one of them.

If we were to move to a neighbourhood with a similar range of venues, we will definitely do better in staying in the same city and moving neighbourhoods rather than crossing the Atlantic.

A corollary drawn from this investigation: melting pots are unique and diverse when replicated in different social, historical and geographical contexts. Despite referring to a similar trend of convivence in gentrified cities, melting pots remain diverse and uniquely configured by their social habitat.

"[T]he problem with ZIP Codes is that:

1. They don't represent real boundaries, but rather routes
2. They don't represent how humans behave"