IBM Capstone Project:

IBM Capstone Project NY and London Neighbourhoods

Background and Introduction

A company looking to set up premises somewhere should perform an in-depth analysis of where to set up their premises. There may be companies, from the US, UK or from other countries, that are considering setting up public premises in either London or New York, or both. A company based and with premises in NY might want to know which neighbourhood in London is most similar in terms of establishments, or vice versa. This is especially true with companies seeking to set up public premises such as shops and restaurants. An area with a high concentration of restaurants or shops would indicate a potentially high foot fall.

For this assignment, I am going to compare London and New York neighbourhoods using the foursquare API to explore venues in each neighbourhood and then clustering the neighbourhoods on the presence of the venues in each neighbourhood. The aim of the analysis is to help a company with both restaurants and cafes decide where to set up shop in London and New York.

To get the data, it would be necessary to find out the number or proportion of each type of business in each neighbourhood for both cities. Then, clusters can be formed form the data and this data can be used to plot the neighbourhood and clusters and the cluster centres can also be defined, which lets a company know what a particular cluster know. Part of the analysis would be to segment the neighbourhoods and define the segments by the prevalence of different types of establishments.

Data Sources

The data of the names of the neighbourhoods in London can be find using a Wikipedia page which will be scraped. For the NY neighbourhoods, a previous dataset from this course is used. The data for the venues comes from the Foursquare API.

Methodology

Data Sources for Neighbourhoods

We will use requests and Beautiful Soup from website data on neighbourhoods in New York and London - using wikipedia pages on towns in Greater London and the New York Neighbourhood data from a previous session from the course

Find the Geolocation Data for Each Neighbourhood

We will use Geopy, Geopandas or any other Geolocation API Service to find the Latitude and Longitude of the neighbourhoods -Use the foursquare API and the Postcodes to explore the neighbourhoods and return the venues within a radius of 1km of each neighbourhood centre

Clean Data (Changing Venue Categories)

We will change the venue category of certain venues in order to ensure a better analysis and clustering. For example, there may be different types of restaurants in each city (Lebanese restaurant, Italian restaurant, Indian restaurant) and the profile of the cities may differ greatly due to many cultural and historical factors, but for the purposes of the analysis, all these establishments are restaurants. Not making this change might lead to two neighbourhoods to be labelled in different groups but they may actually have a similar number of restaurants or other types of establishment and thus be quite similar. For the purposes of this notebook

Data Preparation

We will clean the data, pass to a dataframe and group the venues by neighbourhood to see the number of venues of each type in a neighbourhood

Clustering

We will use K means to cluster the neighbourhoods in London and New York according to the presence of and quantity of different types of venues.

Cluster Centres

We will pass the Cluster back into the combined data frame and then use the cluster, latitude and longitude of each cluster -Define the cluster centre for each cluster (in terms of the top 10 most common types of venue in each cluster centre)

Grouping the Neighbourhoods

The final stage is grouping the neighbourhoods together so that there are lists of which neighbourhoods

Results

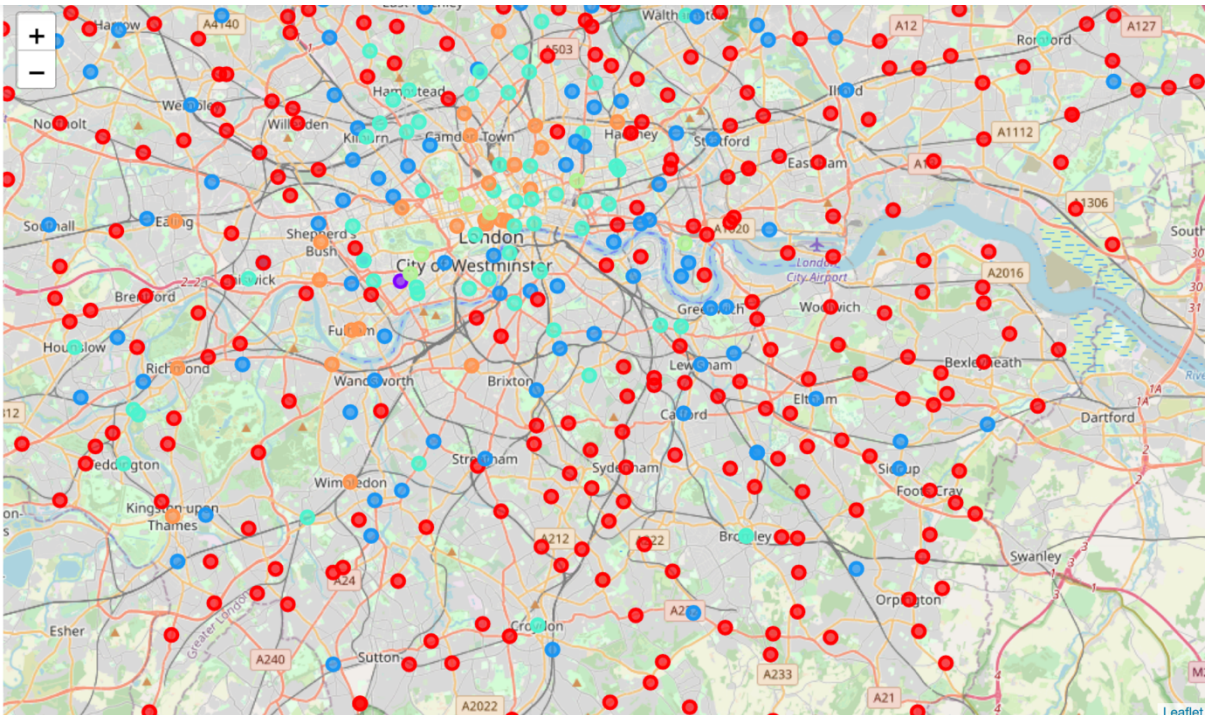

Figure. Map of London with Clusters

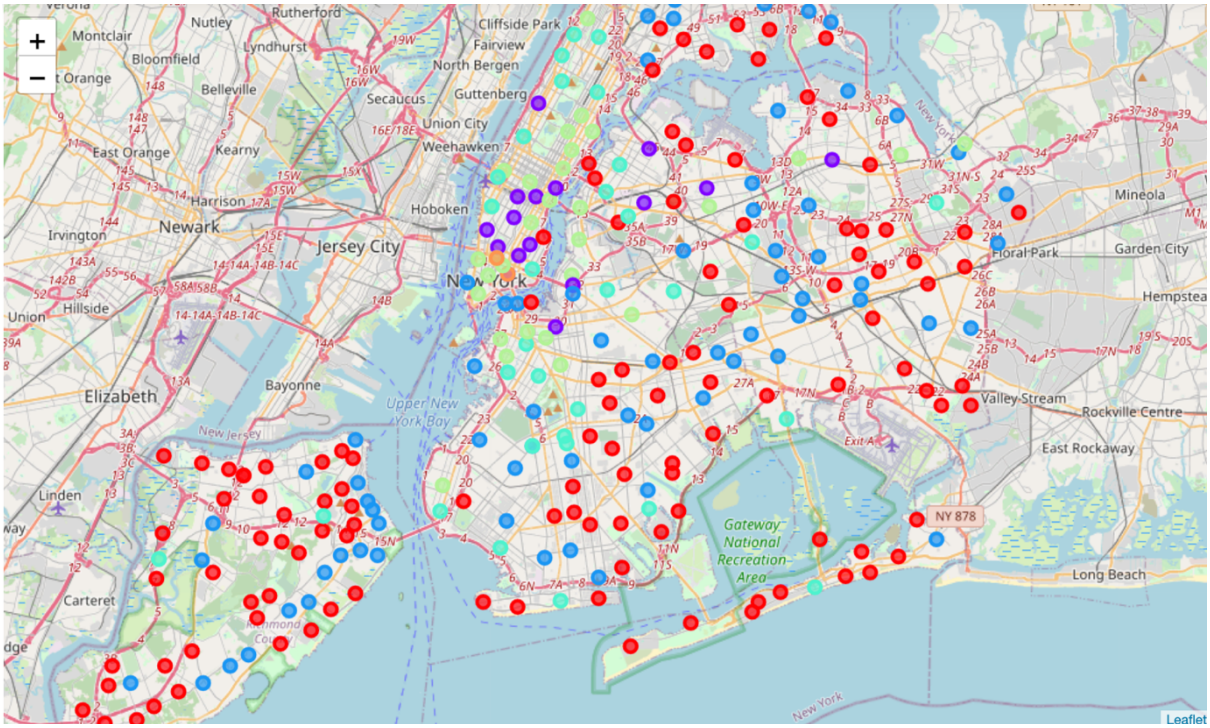

Figure. Map of New York with clusters

Key of clusters

Red - Cluster 0 - lowest number of Restaurants
Purple - Cluster 1 - Highest number of Restaurants
Blue - Cluster 2 – second lowest number of restaurants and other types of venues
Turquoise/Light blue - Cluster 3 – medium to low number of restaurants

Light Green - Cluster 4 - Second highest number of restaurants and high concentration of cafes and bars
Orange - Cluster 5 - Medium number restaurants and a good number of cafes and coffee shops

Top Clusters of Interest – Venue Number Profile

cluster 1

	1
Restaurant	37.444444
Coffee Shop	3.277778
Bakery	3.000000
Bar	2.944444
Pizza Place	2.666667
Hotel	1.888889
Sandwich Place	1.777778
Cocktail Bar	1.777778
Dessert Shop	1.666667
Café	1.555556

cluster 4

	4
Restaurant	24.000000
Coffee Shop	4.142857
Pizza Place	3.142857
Bakery	2.591837
Café	2.346939
Bar	2.061224
Deli / Bodega	1.428571
Hotel	1.346939
Cocktail Bar	1.326531
Sandwich Place	1.285714

cluster 5

	5
Restaurant	19.259259
Coffee Shop	5.592593
Pub	5.518519
Café	4.703704
Bakery	3.037037
Hotel	2.592593
Burger Joint	1.925926
Clothing Store	1.629630
Grocery Store	1.629630
Pizza Place	1.555556

The clusters further away from the centre of London are characterized by lower number of restaurants and any other type of business, as can be expected. Towards the centre of the city, clusters with increasing amounts of restaurants can be seen.

The same trend can be seen with New York city where in Queens, the Bronx and Staten Island, a higher number of areas in lower restaurants can be seen. As you get closer to the centre, i.e., in Manhattan the areas have more restaurants.

An interesting different between the two cities is that London has a greater number of areas in cluster 5 which indicates prevalence of more coffee shops and cafes than New York. New York also has a significantly higher number of neighbourhoods in cluster 4, categorised by having around 20 restaurants 1km from geolocation centre

Discussion

The factor which seems to split the clusters is the number of restaurants followed by other types of landmarks such as pubs, bars and coffee shops.

For a company with restaurants and cafes deciding to set up premises in either city, it seems clear that they should be looking to set up in areas in clusters of higher numbers of restaurants and coffee shops,

New York has a higher number of neighbourhoods with the highest number of restaurants so this might indicate a higher footfall for restaurants and better viability. Given the lack of

neighbourhoods in cluster 1 with the most restaurants in London, companies should look at neighbourhoods in the cluster with the second highest number of restaurants and so on.

However, London has a much higher number of establishments in cluster 5, categories by higher numbers of pubs, bars, restaurants and cafes as well as a good number of restaurants. This indicates that these neighbourhoods would suit very well for coffee shops given their viability and popularity.

The notebook took any venue with a category including the word ‘restaurant’ and changed the category to ‘Restaurant’. This was done in order to standardise the categories. The way New York and London establishments are labelled is likely to differ due to cultural differences, for example bars and pubs where pubs would be more prevalent in the UK. Thus all types of restaurants were changed to restaurant to give a truer indication of establishment density. However, this was only done for restaurants and for a truer analysis of establishment type density, all categories should be revised so that for example coffee house, coffee shop and coffee stand all have the same category, and burger restaurant, pizza restaurant and hot dog restaurant all have the fast-food category. Conversely, if a restaurant chain looking for a deep understanding of the types of restaurants around, the differences between the types of restaurants might be desired but other types of restaurants could be omitted from the analysis.