The Battle of Neighbourhoods

Michal Polski

VS.

Final Project: London vs. New York

1. Introduction / Background

The current world allows us to travel and move between the cities, countries and even continents. The world became more global than ever and the distance does not matter that much anymore. Having in mind the freedom of movement I have decided to look into two cities, which are as similar as different. Global, almost equally large, important and iconic cities. In the same time these two cities represent very different culture and history. The cities which I have chosen are:

- New York
- London

My aim is to help understand the venues landscape in each of the cities.

2. Business Problem

The aim is to help tourists choose their destinations depending on the experiences that the neighbourhoods have to offer and what they would want to have. This also helps people make decisions if they are thinking about migrating to London or New York or even if they want to relocate neighbourhoods within the city. The project findings will help stakeholders make informed decisions and address any concerns they have including the different kinds of cuisines, activity spots, grocery stores and what the city has to offer.

3. Data Description

We require geographical location data for both London and New York. Postal codes in each city serve as a starting point. Using Postal codes we use can find out the neighbourhoods, boroughs, venues and their most popular venue categories.

3.1. London

To derive our solution, We scrape our data from:

https://en.wikipedia.org/wiki/List_of_areas_of_London

This Wikipedia page has information about all the neighbourhoods, we limit it to London only.

a) *borough*: Name of Neighbourhood

b) *town*: Name of borough

c) *post_code*: Postal codes for London

This Wikipedia page lacks information about the geographical locations. To solve this problem we use ArcGIS API

3.2. ArcGIS API

ArcGIS Online enables you to connect people, locations, and data using interactive maps. Work with smart, data-driven styles and intuitive analysis tools that deliver location intelligence. Share your insights with the world or specific groups.

More specifically, we use ArcGIS to get the geo locations of the neighbourhoods of London. The following columns are added to our initial dataset which prepares our data.

a) *latitude*: Latitude for Neighbourhood

b) *longitude*: Longitude for Neighbourhood

3.3. New York Data

Data Source I (json file):

https://cocl.us/new_york_dataset

Geo-spacial data of the New York to get a better understanding of the neighbourhoods in it and their corresponding locations in the Folium map would make certain things clear for the Project. This will be achieved using the acquired data and visualize the same using Choropleth maps.

Data Source II (json file):

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

3.4. Foursquare API Data

The data is going to be collected/acquired from the Foursquare API about the various venues in each neighbourhood of New York city. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, we have chosen the radius to be 500 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

- *Neighbourhood*: Name of the Neighbourhood
- *Neighbourhood Latitude*: Latitude of the Neighbourhood
- *Neighbourhood Longitude*: Longitude of the Neighbourhood
- *Venue*: Name of the Venue
- *Venue Latitude*: Latitude of Venue
- *Venue Longitude*: Longitude of Venue
- *Venue Category*: Category of Venue

Based on all the information collected for both London and New York city, we have sufficient data to build our model. We cluster the neighbourhoods together based on similar venue categories. We then present our observations and findings. Using this data to our stakeholders, so they can take the necessary decision.

4. Methodology

In this chapter we present basic and exploratory analysis, show the application examples and describe machine learning model which will be used for further analysis.

4.1. Exploratory Analysis

The exploratory analysis has been conducted on data related to both cities. The number of venues, neighbourhoods and examples of analysis application.

City	No. of Neighbourhoods	No. of Venues	Unique Venues
New York: Manhattan	40	3172	333
London	50	10276	298

a) New York: Manhattan

Manhattan has 40 unique neighbourhoods, and we can find 333 unique kinds of venues. The list of all venues which are happening / existing in NYC is 3172 long.

Based on exploratory analysis, we can select few neighbourhoods for specific people.

Creating random profiles, lets select 5 examples and assign best neighbourhoods for them:

1. Young, hardworking person, who likes to hit the gym and look for calmer area

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Clinton	Gym / Fitness Centre	Theatre	Italian Restaurant
Marble Hill	Gym	Discount Store	Sandwich Place

2. Student, who likes to party and enjoys exploring restaurants

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
East Village	Bar	Pizza Place	Seafood Restaurant
Gramercy	Bar	Italian Restaurant	Bagel Shop
Inwood	Lounge	Mexican Restaurant	Café

3. Family man, who puts convenience above everything else

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Little Italy	Bakery	Café	Chinese Restaurant
Soho	Clothing Store	Italian Restaurant	Coffee Shop

4. Art student who loves coffee and wants to explore galleries

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Noho	Art Gallery	Italian Restaurant	Mexican Restaurant

5. Nature aficionado who is not fond of crowds

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Morningside Heights	Park	Bookstore	American Restaurant
Roosevelt Island	Park	Coffee Shop	Residential Building (Apartment / Condo)
Stuyvesant Town	Park	Coffee Shop	Bar
Tribeca	Park	Italian Restaurant	Wine Bar

b) London

London has been divided into 50 neighbourhoods, and is characterised by 298 unique venues. The total number of all venues is 10276.

Let's check the best neighbourhoods for profile which we created for Manhattan:

1. Young, hardworking person, who likes to hit the gym and look for calmer area

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Barnet, Brent, Camden	Gym / Fitness Center	Clothing Store	Hardware Store

2. Student, who likes to party and enjoys exploring restaurants

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Camden	**Pub**	Café	Coffee Shop
Camden, Islington	**Pub**	Garden	Bookstore
Greenwich	**Pub**	Grocery Store	Bus Stop
Greenwich, Lewisham	**Pub**	Train Station	Polish Restaurant
Hackney	**Pub**	Café	Coffee Shop
Hammersmith and Fulham	**Pub**	Coffee Shop	Café
Hounslow	**Pub**	Café	Italian Restaurant
Hounslow, Ealing, Hammersmith and Fulham	**Pub**	Café	Italian Restaurant

Lambeth	**Pub**	Café	Coffee Shop
Lambeth, Southwark	**Pub**	Café	Coffee Shop
Lambeth, Wandsworth	**Pub**	Café	Burger Joint
Lewisham	**Pub**	Coffee Shop	Grocery Store
Lewisham, Southwark	**Pub**	Restaurant	Flower Shop
Merton	**Bar**	Pub	Coffee Shop
Southwark	**Pub**	Café	Coffee Shop
Tower Hamlets	**Pub**	Coffee Shop	Chinese Restaurant
Wandsworth	**Pub**	Coffee Shop	Indian Restaurant

3. Family man, who puts convenience above everything else

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Bexley	**Supermarket**	Historic Site	Platform
Bexley, Greenwich	**Supermarket**	Train Station	Platform
Bromley	**Supermarket**	Convenience Store	Hotel
Ealing	**Grocery Store**	Café	Pub
Ealing, Hammersmith and Fulham	**Grocery Store**	Indian Restaurant	Train Station
Islington, Camden	**Grocery Store**	Pizza Place	Coffee Shop

4. Art student who loves coffee and wants to explore galleries

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Bexley	Supermarket	**Historic Site**	Platform
Brent, Harrow	Hotel	**Theater**	Plaza
Redbridge, Waltham Forest	Hotel	**Theater**	Café

5. Nature aficionado who is not fond of crowds

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Islington	Pub	**Garden**	Bookstore

4.2. Clustering (*k*-Means Model)

To better understand the cities, and to detect hidden patterns, we are going to use machine learning tool for clustering. This model will help us find understand and divide different parts of the cities. The model will cluster similar neighbourhoods which will be helpful for stakeholders to limit the number of areas which suit the best their expectations. *k*-Means algorithm is one of the simplest and best approach to data clustering, it helps to create segments of the data or area which are similar to each other. This way,

we will be able to group the neighbourhoods into larger areas, which are similar to each other. Stake holders might be interested in seeing such analysis, to be aware if they can find similar districts, which some small differences (like higher number of gyms).

Method Description:

k-Means algorithm is an iterative algorithm that tries to partition the dataset into *k*-pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way *k*-means algorithm works is as follows:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
   - Compute the sum of the squared distance between data points and all centroids.
   - Assign each data point to the closest cluster (centroid).
   - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster
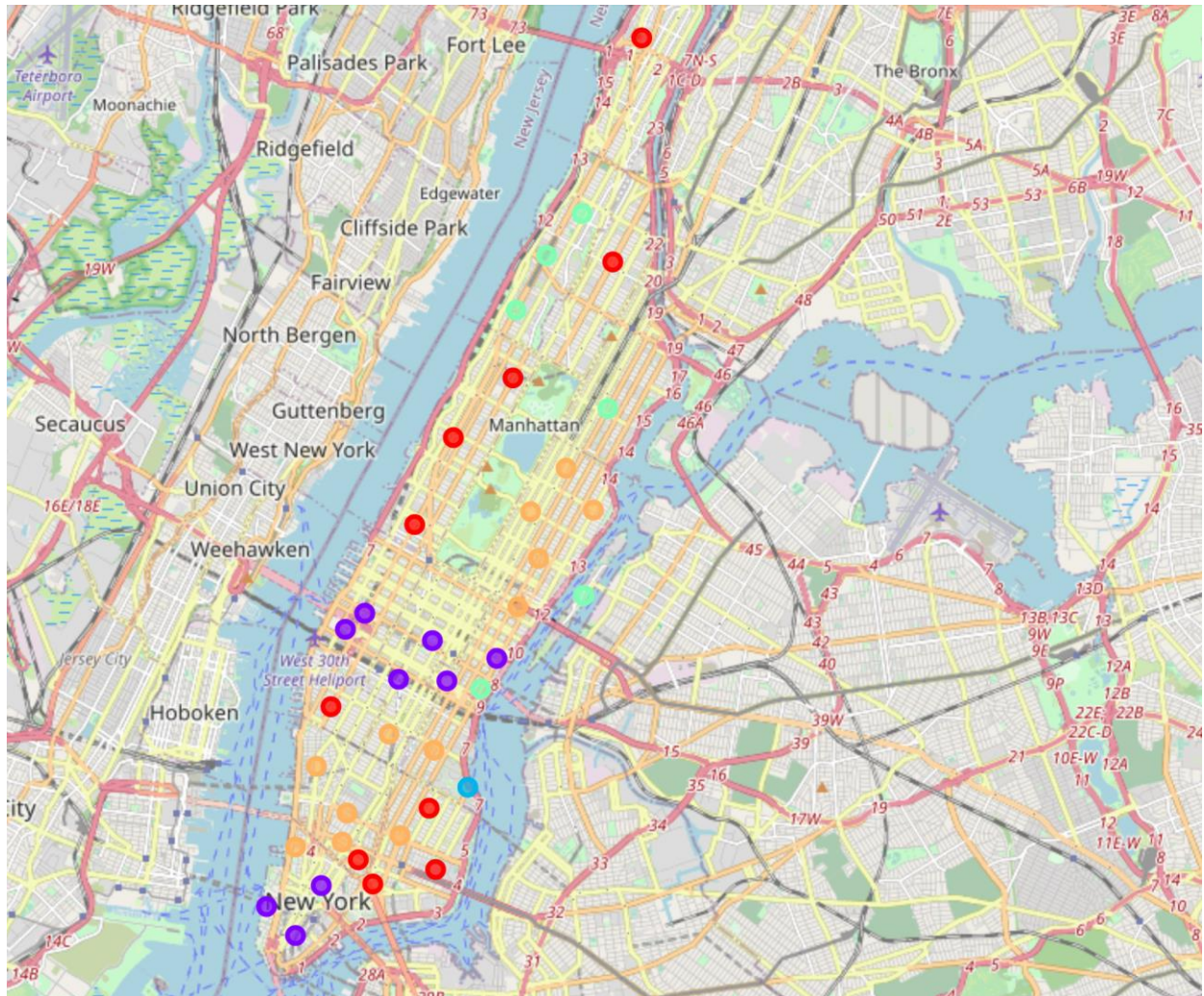
5. Results

The model has been run for two cities and it showed several things around the structure of neighbourhoods in NYC and London. Results will be presented below in form:

a) Graphical Analysis
b) Cluster Analysis

Graphical Analysis

Manhattan Clustering Graphic (k=5):

As can be seen we can notice some larger segments, and surprisingly, similar neighbourhoods are not always next to each other. Model (k=5) has created clusters which are separated by different clusters (examples: red clusters, orange clusters and purple clusters). Based on the picture we can also see that Manhattan is more diverse than London. As we will see in a second, London clusters are more homogeneous and neighbourhoods are similar to each other.

London Clustering Graphic (k=5):
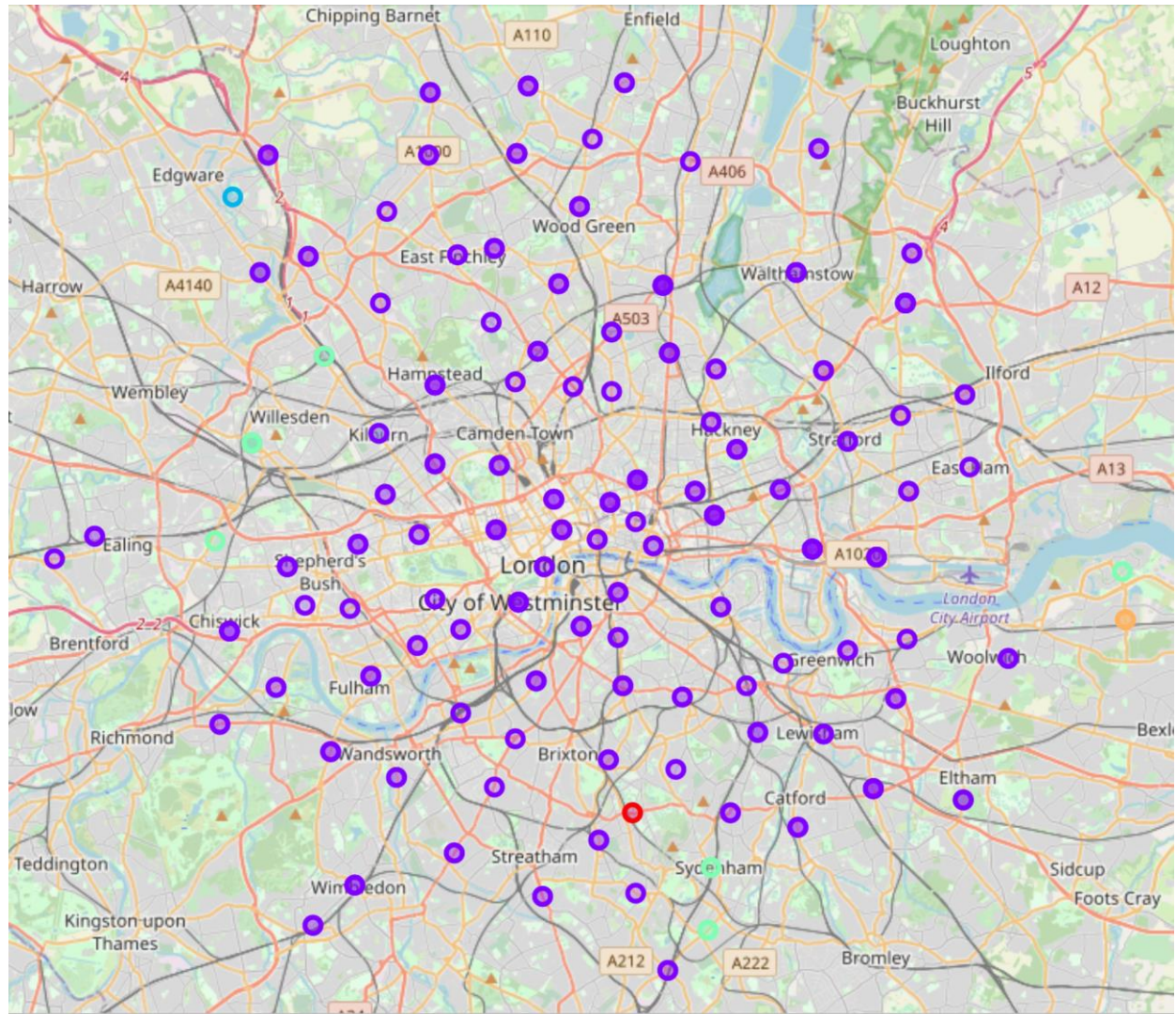
As can be seen at the graphic, for: k equals 5, (purple, green, blue, red and orange). Comparing to Manhattan, neighbourhoods are very similar to each other. Purple is dominant on the map, which means that these neighbourhoods can be characterised by common venues.

Cluster Analysis

Based on model we can check different cluster labels found in the model. For Manhattan, we can present all five labels (0-4) and what are common characteristic around these neighbourhoods.

Now, each cluster can be examined and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, you can then assign a name to each cluster.

Manhattan:

Cluster 1 – Leisure Time (Cafe & Restaurants)

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Chinatown	Chinese Restaurant	Bakery	Dessert Shop
Washington Heights	Café	Bakery	Spanish Restaurant
Central Harlem	African Restaurant	Seafood Restaurant	American Restaurant
Upper West Side	Café	Italian Restaurant	Bakery
Lincoln Square	Café	Plaza	Gym / Fitness Center
Chelsea	Coffee Shop	Bakery	Art Gallery
East Village	Bar	Pizza Place	Seafood Restaurant
Lower East Side	Chinese Restaurant	Latin American Restaurant	Bakery
Little Italy	Bakery	Café	Chinese Restaurant
Manhattan Valley	Mexican Restaurant	Yoga Studio	Bar

Cluster 2 – Tourists and Wellbeing Areas

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Marble Hill	Gym	Discount Store	Sandwich Place
Clinton	Gym / Fitness Centre	theatre	Italian Restaurant
Midtown	Hotel	Clothing Store	theatre
Murray Hill	Coffee Shop	Sandwich Place	Hotel
Battery Park City	Hotel	Coffee Shop	Clothing Store
Financial District	Coffee Shop	Pizza Place	Bar
Civic Centre	Coffee Shop	Spa	Gym / Fitness Centre
Midtown South	Korean Restaurant	Hotel	Japanese Restaurant
Turtle Bay	Coffee Shop	Italian Restaurant	Sushi Restaurant
Hudson Yards	American Restaurant	Gym / Fitness Centre	Hotel

Cluster 3 – Park Cluster

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Stuyvesant Town	Park	Coffee Shop	Bar

Cluster 4 - Diverse Areas

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Inwood	Lounge	Mexican Restaurant	Café
Hamilton Heights	Pizza Place	Café	Coffee Shop
Manhattanville	Coffee Shop	Seafood Restaurant	Sushi Restaurant
East Harlem	Mexican Restaurant	Bakery	Thai Restaurant
Roosevelt Island	Park	Coffee Shop	Residential Building (Apartment / Condo)
Morningside Heights	Park	Bookstore	American Restaurant
Tudor City	Mexican Restaurant	Park	Café

Cluster 5 - Italian Restaurant Cluster

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Upper East Side	Italian Restaurant	Coffee Shop	Exhibit
Yorkville	Italian Restaurant	Coffee Shop	Gym
Lenox Hill	Italian Restaurant	Coffee Shop	Cocktail Bar
Greenwich Village	Italian Restaurant	Clothing Store	Sushi Restaurant
Tribeca	Park	Italian Restaurant	Wine Bar
Soho	Clothing Store	Italian Restaurant	Coffee Shop
West Village	Italian Restaurant	New American Restaurant	Cocktail Bar
Gramercy	Bar	Italian Restaurant	Bagel Shop
Carnegie Hill	Coffee Shop	Café	Yoga Studio
Noho	Art Gallery	Italian Restaurant	Mexican Restaurant
Sutton Place	Coffee Shop	Italian Restaurant	Gym
Flatiron	Italian Restaurant	American Restaurant	Spa

London

Cluster 1 - Coffee Shops and Pubs (part of the data)

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
City	Coffee Shop	Hotel	Italian Restaurant
Westminster	Hotel	Coffee Shop	Sandwich Place
Islington	Coffee Shop	Pub	Café
Barnet	Coffee Shop	Café	Grocery Store
Redbridge	Pub	Coffee Shop	Grocery Store
Barnet	Coffee Shop	Café	Grocery Store
Greenwich	Pub	Grocery Store	Coffee Shop
Hammersmith and Fulham	Pub	Coffee Shop	Café

The most common venue in London is Pub, the next one is coffee shop – these two venues characterise the most of the neighbourhoods in town.

Cluster 2 – Bakery

borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Harrow, Brent	Bakery	Indian Restaurant	Gym

Cluster 3 - Grocery Store & Indian Restaurant

borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Hammersmith and Fulham	Grocery Store	Park	Indian Restaurant
Bromley	Supermarket	Grocery Store	Convenience Store
Brent	Café	Indian Restaurant	Pharmacy
Barnet	Clothing Store	Convenience Store	Music Store
Haringey, Islington	Hotel	Pub	Athletics & Sports
Ealing	Fast Food Restaurant	Bus Stop	Chinese Restaurant
Lewisham	Pub	Gym / Fitness Centre	Indian Restaurant

Cluster 4 – Supermarket & Historic Sites

borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Greenwich	Supermarket	Historic Site	Train Station
Bexley	Supermarket	Historic Site	Platform

Cluster 5 – Flower Shop

borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Southwark	Pub	Flower Shop	Gym / Fitness Center

6. Discussion / Observations / Recommendations

One of the observations regarding the comparison is that cities are very different. Each city offers very different venues. London is more homogeneous, while, Manhattan is more diverse – the neighbourhoods represent very different sets of venues. Also, London might have more to offer, as the number of venues is about 3 times higher. Additionally, the most common venues for London are: Pubs (very British) and Coffee Shops – people love coffee apparently. For Manhattan, the most popular restaurant is Italian, which is expected, having in mind that New York loves Pizza.

Overall, if you are a tourist, and want to decide where to go, based on our model we can recommend New York for someone who likes diversity and Italian food. If you prefer more convenient city (lots of supermarkets) and you like to have a pint – choose London, its full of pubs.

7. Conclusions

In this project, we were trying to find out the characteristics of the two cities, neighbourhoods and check how these neighbourhoods cluster.

k-Means model helped us to better understand the landscape of areas in each city, which can help tourists or new comers to choose best district for their taste.

Both cities have so much to offer, but based on the data, analysis and modelling we can conclude some facts around them.

If you are a fan of Italian cuisine and you like variety you should choose New York. If you prefer to spent leisure time in pub, or you value convenience and shop in supermarket – go with London.