Starting A New Restaurant in London

BY ANALYZING NEIGHBOURHOODS OF LONDON

Bingqing He | IBM DATA SCIENCE – CAPSTONE PROJECT | 2021.07

1. Introduction

1.1 Background

As one of the world's most important global cities, London is the capital and largest city of England and the United Kingdom. It exerts considerable influence upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation. It is one of the largest financial centers in the world and in 2019, London had the second highest number of ultra high-net-worth individuals in Europe, after Paris.

London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. Its estimated mid-2018 municipal population (corresponding to Greater London) was roughly 9 million, which made it the third-most populous city in Europe. London accounts for 13.4% of the U.K. population.

London is full of restaurants which serves millions of people from the world everyday and the diversity in the population brings the different food habits of people. In this project, the neighbourhoods will be studied, and recommendations will be provided.

1.2 Problem

Data that might contribute to determining the most common venues in the various neighbourhoods include their names and types etc. The main objective of this project is to extract and analyze the right data about various neighbourhoods of London using various data science techniques and suggest our client a suitable location and type for their restaurant.

1.3 Interest

Obviously, those investors who are interested in opening a new restaurant in the city of London would be our target clients. Others who want to find the venues around their residence in the city of London may also be interested.

2. Source of Data

2.1 Neighbourhoods of London

The basic data is extracted from the "List of areas of London" Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London) and the BeautifulSoup library in Python is used for web scraping. After data extracting, there is a detailed list of neighbourhoods present in London.

2.2 Geographical Coordinates

After getting the neighbourhoods data, Python GeoPy library is used for extracting the geographical coordinates of various neighbourhoods. Geographical coordinates are important for the project since they can be used for plotting maps when visualizing the data. With the help of GeoPy, the dataframe has two more columns with latitude and longitude information for each neighbourhood as the figure shown below:

	Neighbourhood	Borough	PostCode	District	Latitude	Longitude
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	51.490860	0.121020
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	51.633296	-0.176466
2	Addington	Croydon[8]	CROYDON	CR0	51.575810	-0.109340
3	Addiscombe	Croydon[8]	CROYDON	CR0	51.472749	-0.203326
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	51.485820	-0.080260

Figure 1. Neighbourhoods with their Latitude and Longitude

2.3 Venue Data

By using FourSquare API, the venue data is extracted and added to the dataset. The data is used to study the venues in various neighbourhoods in London. It provides useful information of various restaurants and other places of interests in the area which helps us understand the competition. This step is the foundation of drawing the conclusion later in the project.

3. Methodology

3.1 Feature Extraction

Feature extraction is carried out by finding the hot spot around the neighbourhoods. In this method, features are the categories represented in the venues and we use binary codes to show whether it exists or not. Specifically, 1 means the category is found in the venue while 0 means not found. After that, all the venues are grouped by the neighbourhoods and the rows show the venues while columns represent the frequency of occurrence of the category.

	Neighbourhood	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	Airport Service	Airport Terminal	Airport Ticket Counter	American Restaurant	...	Windmill	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Xinjiang Restaurant	Yoga Studio	Zoo	Zoo Exhibit
0	Abbey Wood	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Abbey Wood	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Abbey Wood	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Abbey Wood	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Abbey Wood	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 398 columns

Figure 2. Neighbourhoods with venues and category

3.2 Machine Learning

In order to figure out the similarities features that can be useful for categorization, a clustering algorithm called K-Means is implemented. K-Means clustering is one of the simplest and popular unsupervised machine learning algorithms and in this project it helps grouping similar data points together and discovering underlying patterns.

The target number k, which refers to the number of centroids used in the dataset, and the centroids represents the center of the cluster. Every data point is allocated to each cluster by reducing the in-cluster sum of squares. Basically, the K-Means algorithm starts with a group of randomly selected centroids and performs iterative calculations to optimize the positions of the centroids.

In order to find the prior idea about the number of clusters, the Elbow method is implemented in the project. A chart is plotted to compare the errors and numbers of cluster and the optimum target number k is selected. As the figure shown below, 14 is the best choice in the project.

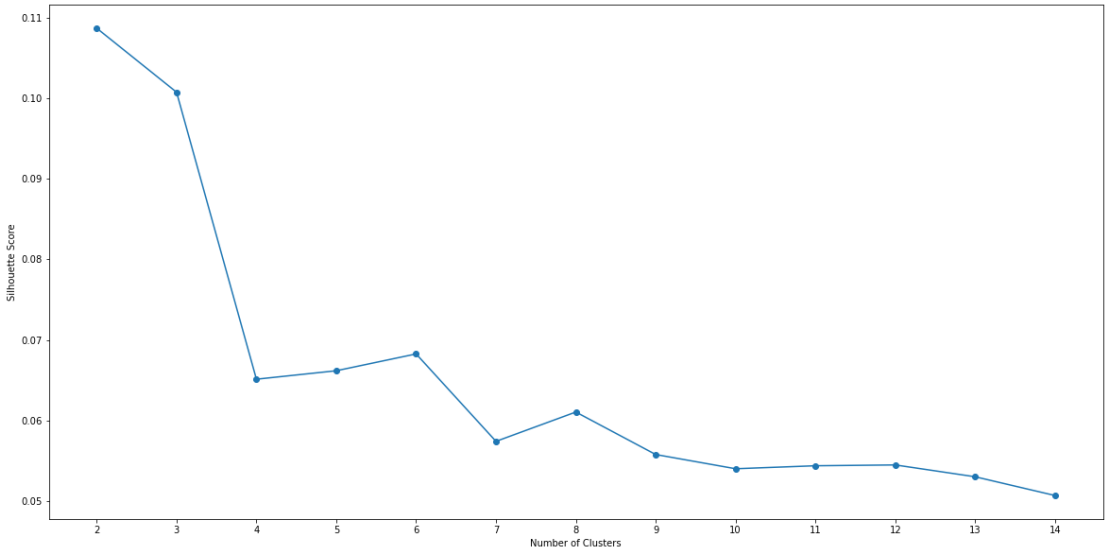

Figure 3. The choice of the centroid in K-Means clustering

3.3 Plotting

Data visualization in a graphical representation of information and data. In this project, several plotting techniques are used to see and understand trends, outliers and patterns in data. The main tool used is the Folium library in Python which plots the map of London and neighbourhoods as well.

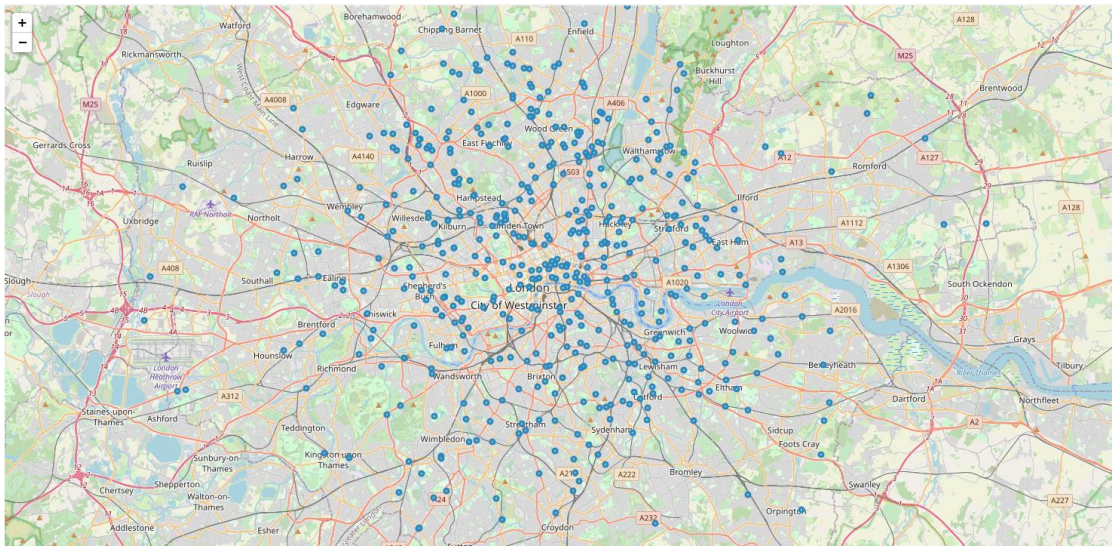

Figure 4. Map of London with neighbourhoods

4. Results

According to the previous sections, when applying the K-Means clustering method and Elbow method to the dataframe of neighbourhoods in London, we use 14 as the centroid number of clusters. After adding more colors in the plot, it is clearly shown that Cluster 2 are grouped in the center of London. Additionally, according to the detailed dataframe shown below, the area has various of different types of food such as Indian food, Japanese food, Korean food etc. as well as a group of coffee shops, drinking bars and pubs.

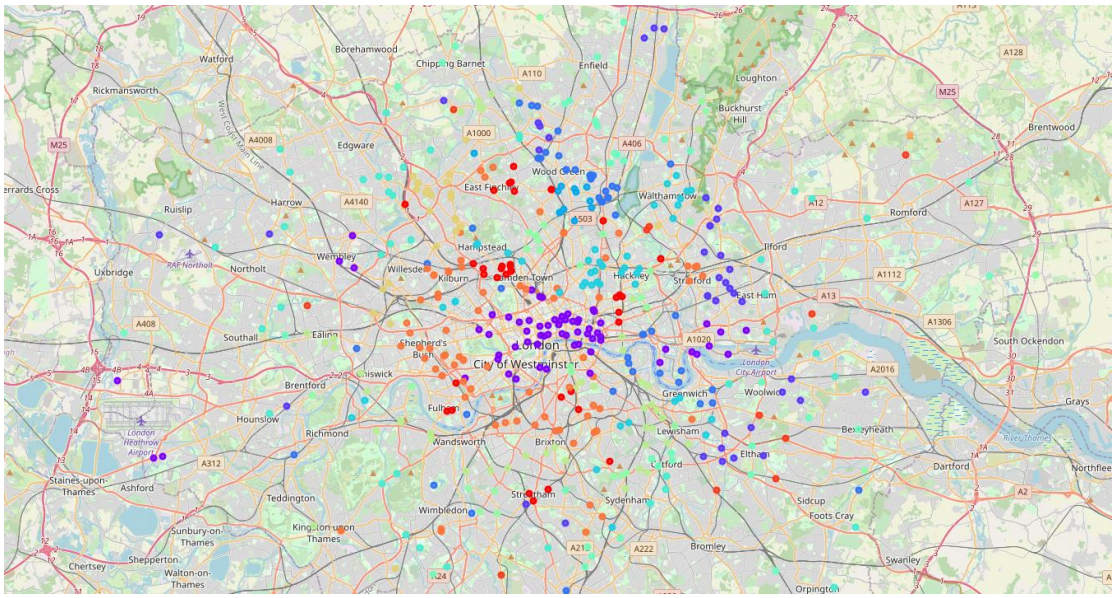

Figure 5. Cluster visualization with Folium

	Neighbourhood	Top 1 Common Venue	Top 2 Common Venue	Top 3 Common Venue	Top 4 Common Venue	Top 5 Common Venue	Top 6 Common Venue	Top 7 Common Venue	Top 8 Common Venue	Top 9 Common Venue	Top 10 Common Venue
6	Aldgate	Hotel	Cocktail Bar	Coffee Shop	Gym / Fitness Center	Pizza Place	Market	Castle	Scenic Lookout	Sushi Restaurant	Event Space
7	Aldwych	Hotel	Steakhouse	Ice Cream Shop	Sushi Restaurant	Bakery	Theater	Restaurant	History Museum	Coffee Shop	Spanish Restaurant
17	Bankside	Coffee Shop	Italian Restaurant	Hotel	Seafood Restaurant	Art Museum	Restaurant	Burger Joint	Grocery Store	Portuguese Restaurant	Wine Bar
18	Barbican	Coffee Shop	Food Truck	Gym / Fitness Center	Hotel	Steakhouse	Bar	Concert Hall	Roof Deck	Vietnamese Restaurant	Cocktail Bar
28	Bayswater	Hotel	Coffee Shop	Gym / Fitness Center	Garden	Gastropub	American Restaurant	Café	Hotel Bar	Pub	Fountain
...	...	...	...	...	...	...	...	...	...	...	...	...
467	Tower Hill	Hotel	Coffee Shop	Cocktail Bar	Gym / Fitness Center	Scenic Lookout	Restaurant	Garden	Castle	Tapas Restaurant	Grocery Store
473	Upminster Bridge	Hotel	Cocktail Bar	Coffee Shop	Steakhouse	Scenic Lookout	Restaurant	Gym / Fitness Center	Castle	Event Space	Trail
494	Wembley	Indian Restaurant	Coffee Shop	Clothing Store	Hotel	Pharmacy	Gym / Fitness Center	Sporting Goods Shop	Sandwich Place	Pizza Place	Ice Cream Shop
495	Wembley Park	Coffee Shop	Hotel	Clothing Store	Indian Restaurant	Gym / Fitness Center	Sporting Goods Shop	Grocery Store	Bar	Pedestrian Plaza	Park
504	West Harrow	Coffee Shop	Hotel	Pub	Bookstore	History Museum	Beer Bar	Restaurant	Wine Bar	Breakfast Spot	Gym / Fitness Center

67 rows × 11 columns

Figure 6. Detailed Top 10 most common venues in cluster 2

5. Discussion

The K-Means model works well and successfully clustered similar nighbourhoods together. After studying all the clusters, it shows that different clusters have different most common venues. People who live in cluster 1 have more interest in spending time in a café, those who live in cluster 6, 8, 9 and 10 prefer pubs while residents in cluster 14 are more likely to walking in a park.

Those who interested in starting a new restaurant in London can go ahead and make a decision on the type and location of the catering business depending on more factors such as parking availabilities and transportation.

6. Conclusion

Business data analysis includes the activities to make strategic decisions, achieve major goals and solve complex problems by collecting analyzing and reporting the most useful information relevant to needs. Information can be about the causes of the current situation, the most likely trends to occur and what should be done as a result.

In this project, Python’s built-in libraries such as Folium and GeoPy, as well as BeautifulSoup API are used for extracting data, analyzing geographical location and plotting graphs and charts. With the help of those data analysis and machine learning techniques, the neighbourhoods of London are studied, and some suggestions are provided for those investors who want to start a new restaurant in the city.