CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

VISITING LONDON CITY

Jorge Álvarez de la Fuente

April 12, 2020

1. INTRODUCTION

London is the capital and largest city of England and the United Kingdom. It is 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. Greater London Built-up Area is the fourth-most populous in Europe with 9,787,426 inhabitants at the 2011 census.

London has been one of the world's top tourism cities for many years, and a key gateway for domestic and international visitors. The total number of visitors in London, including domestic tourists, is estimated at 12 million per year.

As a result, tourism is the second most important sector for the economy of the city after financial services, contributing 12% to its GDP. Tourism, therefore, plays a vital role for London. The sector employs 700,000 people and contributes £36 billion a year to the economy. Tourism also helps to strengthen London's reputation as an open and welcoming global city.

2. BUSINESS PROBLEM

London is a world tourist destination with many visitors each year. The information available concerning tourism in London can be overwhelming and considering trips made to London for holiday purposes lasted 4.8 nights on average in 2019, according to Statista, a good planning is necessary to make the most of the holidays, avoiding downtimes. London has full potential to become the best city to visit in Europe but it still has challenges for visitors such as not using public transport wisely, waiting in long queues as a result of a poor planning and failing to research available attractions.

This project will analyse the venues available for people visiting London in order to look for the best recreational activities and locations in the variety of districts in London city. The venues will be analysed first as a whole and structured afterwards by district for a better comprehension as well.

This project will be interesting to visitors and expats who are considering visiting and relocating to London as they will be able to identify the most attractive locations in London and explore its districts and common venues around each district.

3. DATA

This section describes the data to be used to solve the problem regarding tourism and visits to London city.

3.1 Libraries: The libraries installed and imported to be used in this project are summarized below.

- Numpy: Library to handle data in a vectorized manner. It is used to create cluster labels during the clustering stage.
- Pandas: Library for data analysis. In this project is used to build DataFrames containing Postal Codes, districts, venues, latitudes, longitudes.
- Requests: Library to handle requests. It is used to scrape the list of Postal Codes in London city from Milesfaster website (please, refer to sources section below).
- BeautifulSoup: Library used for pulling data out of html and XML files. It is necessary to scrape information from web pages.
- Geopy: Library used to convert an address into latitude and latitude values. It is used to add the latitudes and longitudes to the districts scraped.
- Json: Library to handle JSON files. In this project is used to make the GET request during the neighborhood exploration stage.
- folium: Library used to create maps. It is used to create maps in London city, containing all districts previously scraped.
- Matplotlib.cm and Matplotlib.colors: Library used to plot a range of colors in London city maps.
- k-means: Library used to apply k-means clustering algorithm used to cluster districts in London city.

3.2 Data Acquisition: The data acquired for this project contains the following columns.

- Postal Code: Contains the postal codes of London city that have the following coding. C: Central, N: North, E: East, W: West, SE: South East, SW: South West, NE: North East, NW: North West.
- District: London city districts within a Postal Code.
- Postal Code Latitude: Latitude of a given Postal Code in London city.
- Postal Code Longitude: Longitude of a given Postal Code in London city.
- Venue: Venue names within a district in London city.
- Venue Latitude: Latitude of a given venue in London city.
- Venue Longitude: Longitude of a given venue in London city.
- Venue Category: Category of a given venue (such as bakery, coffee shop, movie theater, etc.)
- Most common venue: The most common venue within a district ranging from 1st to 10th.

3.3. Data sources: For this project the following data and sources will be used:

- **Milesfaster website** This is a London Hotels specialist website which contains the postal codes and districts of London city. The postal codes and district data of London city will be scraped from Milesfaster using the beautiful soup library of Python, however, it requires an additional data cleaning step to

reorganize and label columns. Milesfaster url is the following: https://www.milesfaster.co.uk/london-postcodes-list.htm

- **ArcGIS** Geocoding is the computational process of transforming a physical address description to a location on the Earth's surface (spatial representation in numerical coordinates). ArcGIS will be used to turn addresses into coordinates, finding addresses interactively and geocoding a table of addresses.

- **Foursquare API** to get the most common venues of given district in London city. Foursquare API has one of the largest database of 105+ million places and is used by over 1250,000 developers. It provides different data information of venues among districts. Foursquare data contains venues, longitude, latitude and postal codes. The information obtained per venue as follows: district, district latitude, district longitude, venue name, venue latitude, venue longitude, venue category. This will be used in the methodology section.

3.4. Scraping the list of postal codes

This section consists in scraping the list of postal codes with their respective districts of London city. The raw data has been extracted from Milesfaster website and then processed parsing the html table using Beautiful Soup library to be converted into a new DataFrame using Pandas library. This table contains the postal codes in two columns simultaneously (Table 1), so these columns should be splitted separately.

Table 1. DataFrame extract containing the table from Milesfaster website with postal codes and districts.

	0	1	2	3
0	E1	Whitechapel, Stepney, Mile End	SE1	Waterloo, Bermondsey, Southwark, Borough
1	E1W	Wapping	SE2	Abbey Wood
2	E2	Bethnal Green, Shoreditch	SE3	Blackheath, Westcombe Park
3	E3	Bow, Bromley-by-Bow	SE4	Brockley, Crofton Park, Honor Oak Park

3.5. Data Cleaning

As mentioned previously, columns 0 and 1 and 2 and 3 should be splitted separately and these columns should be renamed to better descriptive labels. In addition there are NaN values in some rows that must be removed as well. All these steps define the data cleaning process. The results are shown in Table 2.

Table 2. DataFrame extract with the results of the data cleaning process.

	Postal Code	District
0	E1	Whitechapel, Stepney, Mile End
1	E1W	Wapping
2	E2	Bethnal Green, Shoreditch
3	E3	Bow, Bromley-by-Bow

3.6. Adding the latitude and longitude coordinates to the DataFrame

In order to get the coordinates for all districts using ArcGIS, a new function has been created. This function called as get_geocode gets the coordinates (latitude and longitude) for all districts, initializes the location to None and uses a while loop to obtain the coordinates, returning them when called. Two test sampling random chosen postal codes have been performed with the purpose to validate the function created. These tests were successful and the function was applied to the corresponding DataFrame to add the latitudes and longitudes to each postal code and corresponding district(s) (Table 3). This new DataFrame created with the latitudes and longitudes obtained using get_geocode function has been used to generate the venues for each district using Foursquare API.

Table 3. Latitudes and longitudes added to the DataFrame.

	Postal Code	District	Latitude	Longitude
0	E1	Whitechapel, Stepney, Mile End	51.520220	-0.054310
1	E1W	Wapping	51.506282	-0.069426
2	E2	Bethnal Green, Shoreditch	51.526690	-0.062570
3	E3	Bow, Bromley-by-Bow	51.527020	-0.025940

3.7. Using Foursquare API to explore the districts and segment them

The geographical coordinates of London city have been obtained using Nominatim, a tool to search data by name and address (geocoding) and to generate addresses of OpenStreetMap points (reverse geocoding). The coordinates of London found are: 51.5073219, -0.1276474. The districts of London city have been visualized using Folium library. Figure 1 displays the map of London with all districts obtained:

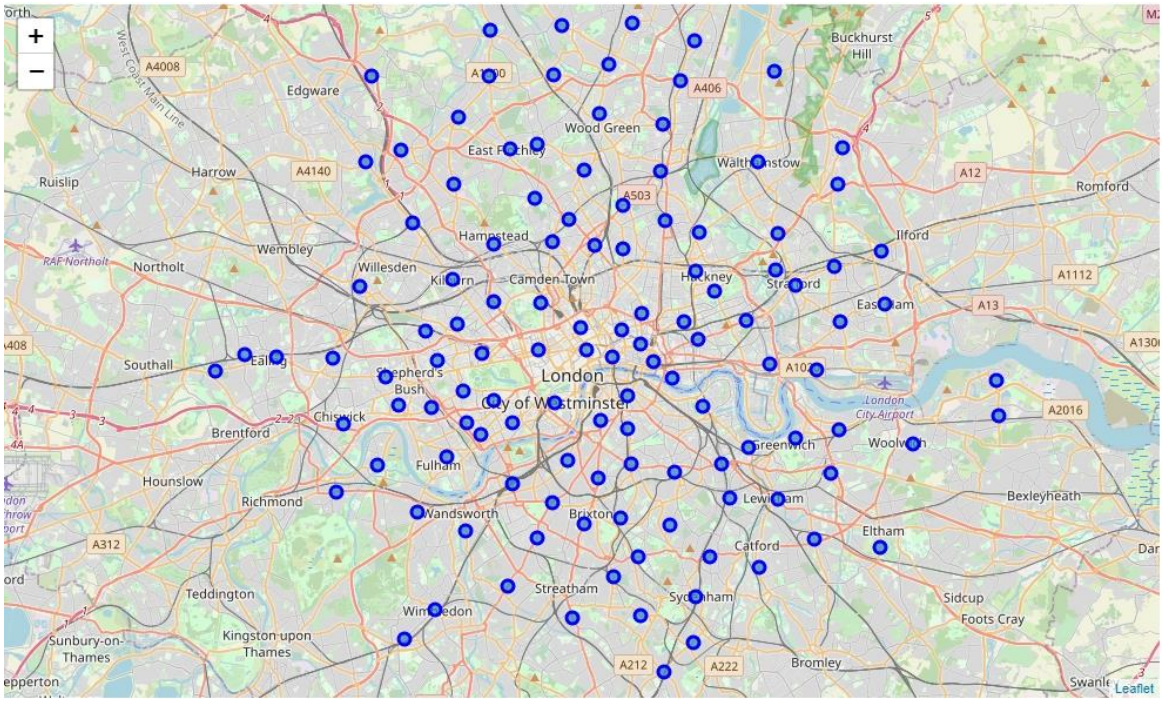

Figure 1. Map of London city with all the districts obtained from Milesfaster website.

The function getNearbyVenues was created to explore the districts in London, including the API request url: https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}
An extract of the full list of districts in London city is shown in Table 4:

Table 4. An extract of the full list of venues in London city.

```
Whitechapel, Stepney, Mile End
Wapping
Bethnal Green, Shoreditch
Bow, Bromley-by-Bow
Chingford, Highams Park
Clapton
East Ham
Forest Gate, Upton Park
Hackney, Dalston
Hackney, Homerton
Leyton
Leytonstone
Manor Park
Plaistow
Poplar, Millwall, Isle of Dogs, Docklands
Stratford, West Ham
Canning Town, North Woolwich, Docklands
Walthamstow
South Woodford
```

The information obtained per venue using the function getNearbyVenues which contains the API request url, the GET request and a new DataFrame created with new columns is the following: district, district latitude, district longitude, venue name, venue latitude, venue longitude and venue category. The function getNearbyVenues retrieved a total of 11,145 venues and an extract of this DataFrame is shown in Table 5. Besides, Table 6 displays an extract of the number of venues per district, grouped by district.

Table 5. Extract of the DataFrame containing venues in London with latitudes, longitudes and categories.

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Rinkoff's Bakery	51.519964	-0.053238	Bakery
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Mouse Tail Coffee Stories	51.519471	-0.058573	Coffee Shop
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Genesis Cinema	51.521036	-0.051073	Movie Theater
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Second Shot	51.527412	-0.056625	Coffee Shop
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Old Street Brewery & Taproom	51.526950	-0.056426	Brewery
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Stepney Green Park	51.517768	-0.047054	Park
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Renegade London Wine	51.527005	-0.056381	Wine Bar
Whitechapel, Stepney, Mile End	51.52022	-0.05431	Mother Kelly's Bottle Shop and Tap Room	51.528413	-0.055843	Beer Bar

Table 6. Extract of the number of venues per district, grouped by district.

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Abbey Wood	19	19	19	19	19	19
Acton	100	100	100	100	100	100
Archway, Tufnell Park	100	100	100	100	100	100
Balham	100	100	100	100	100	100
Barnes, Castelnau	100	100	100	100	100	100
Battersea, Clapham Junction	100	100	100	100	100	100

4. METHODOLOGY

The methodology applied in this project is as follows:

- Foursquare API will be used to explore all venues of all districts and analyse each district in London city.
- The most relevant venues to visit will be retrieved from the entire DataFrame and visualized in bar plots.
- The most common venues will be analysed using word cloud to look for the most common words and translate them into venues.
- Each district will be grouped with the 10 most common venues.
- All districts in London city will be clustered, visualized and examined using k-means algorithm.
- The resulting clusters will be analysed using word cloud to validate the analyses previously performed.

5. EXPLANATORY ANALYSIS

5.1. Analyzing districts

This section consists in performing an explanatory data analysis and deriving some additional information from the raw data. As aforementioned the entire DataFrame comprises 11,145 venues for all districts. A new DataFrame containing random chosen relevant venue categories was created with the purpose of evaluating whether or not London is a good place to visit. Figure 2 displays the results sorted in descending order.

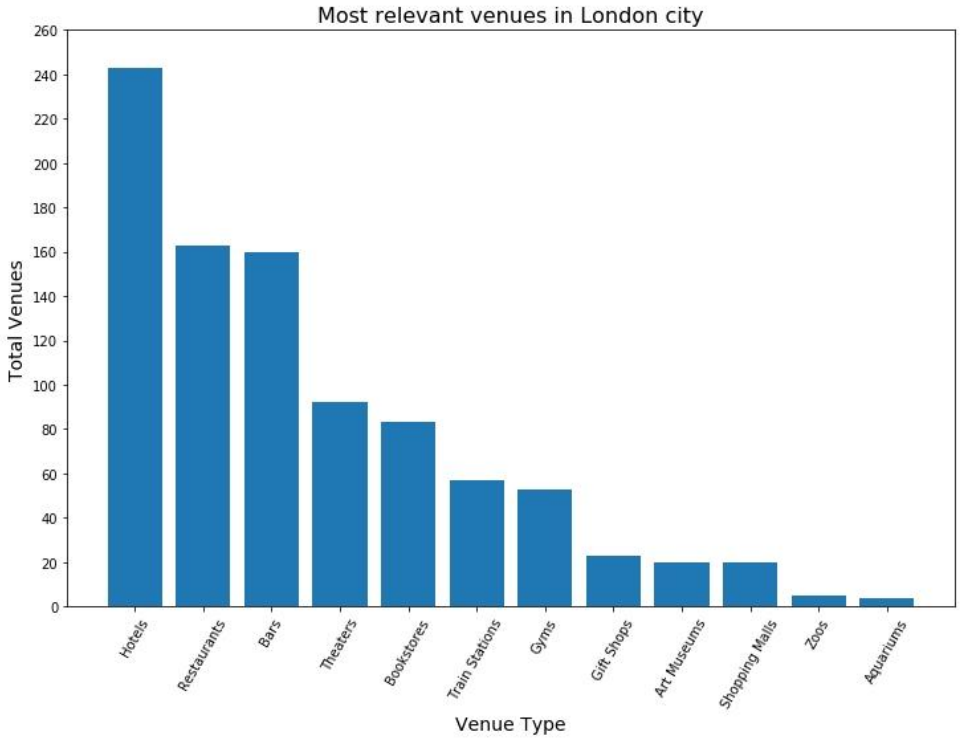

Figure 2. Bar plot with the count of the most popular venues in London city.

As can be noted from the bar plot above, from the venues randomly selected, hotels are the most common venue in London city, followed by restaurants. Hotels were grouped by district to find which districts have most hotels in London city. The results can be visualized below in Figure 3.

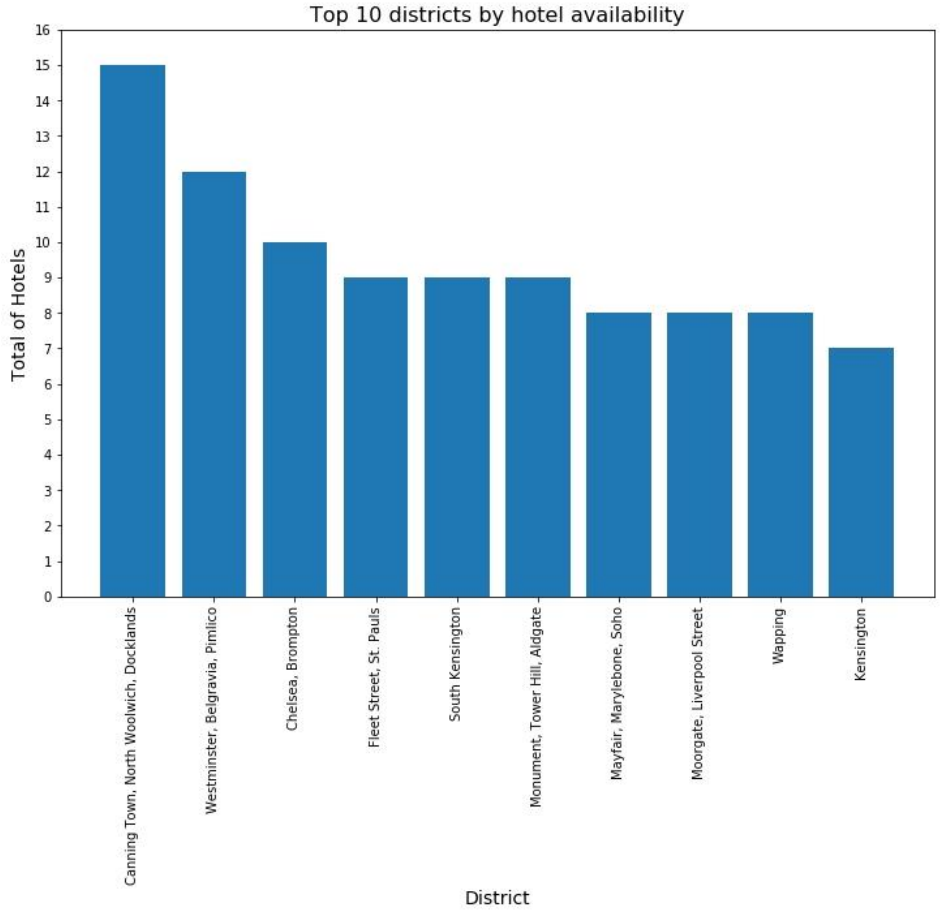

Figure 3. Top 10 districts by hotel availability.

A new DataFrame containing the most popular cuisine types and restaurants in London city was developed. Figure 4 illustrates the number of restaurants by origin or cuisine type in a bar plot.

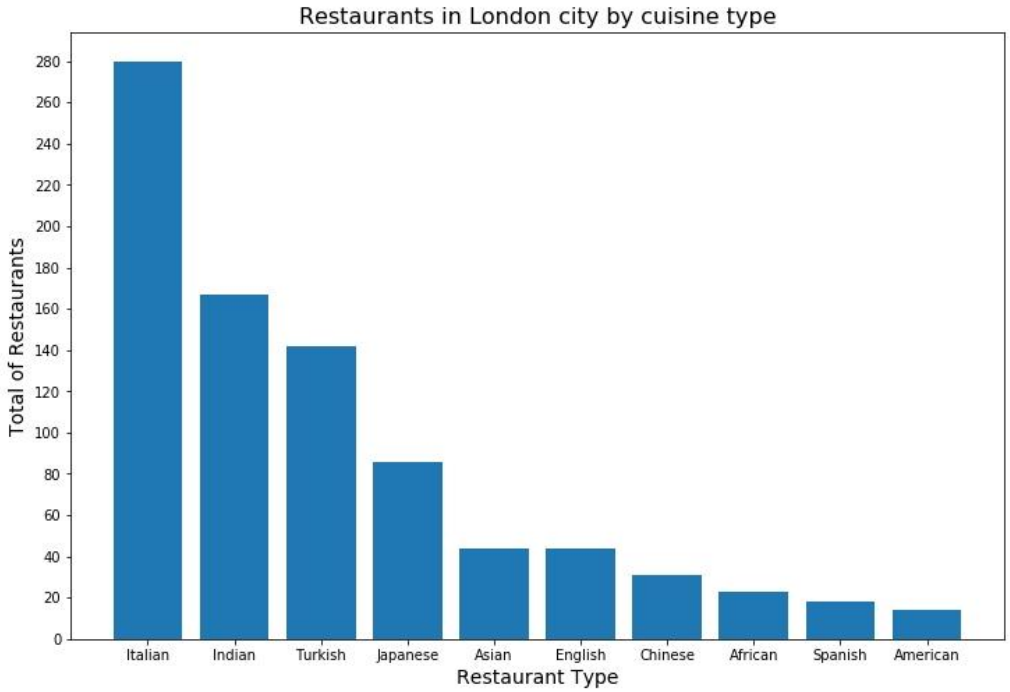

Figure 4. Restaurants in London city by cuisine type or country of origin.

As observed above (Figure 4) Italian and Indian restaurants lead the cuisines in London city. These restaurants were grouped by district. The results are shown below in Figure 5 and Figure 6 respectively.

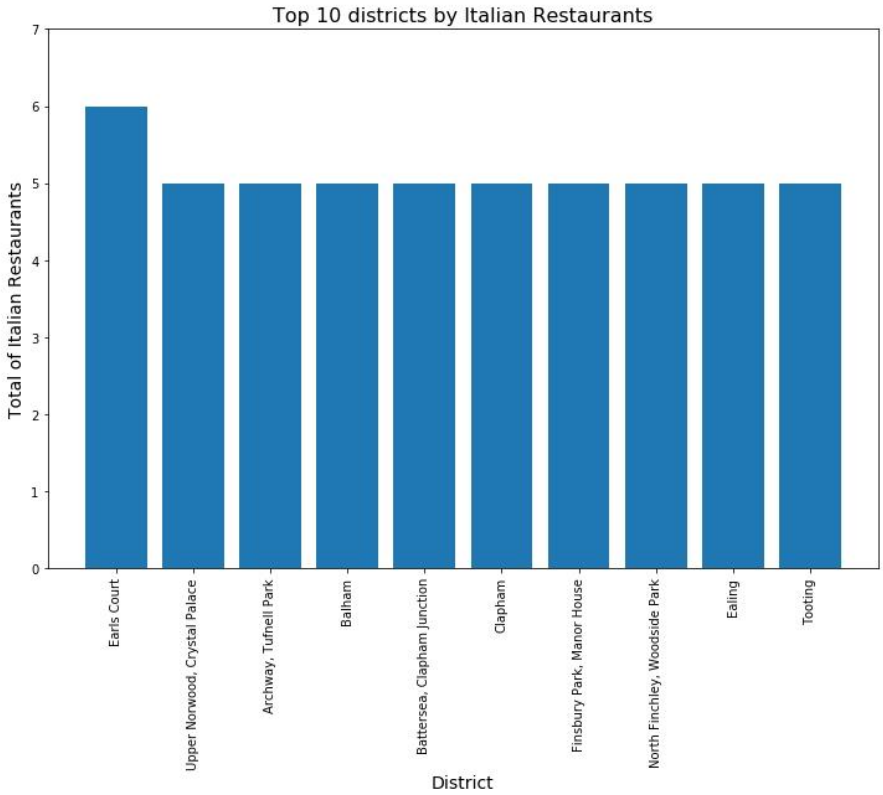

Figure 5. Top 10 districts by Italian Restaurants in London city.

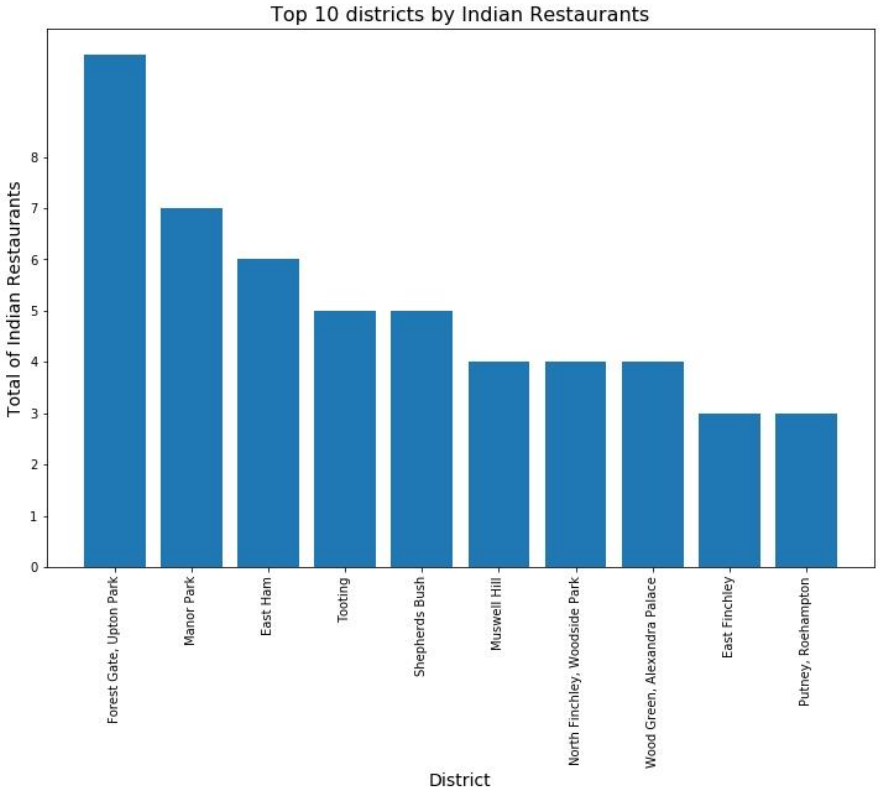

Figure 6. Top 10 districts by Indian Restaurants in London city.

A new function was created with the purpose of revealing the most common venues by district and a new DataFrame was created to display the top 10 venues for each district (Table 7).

Table 7. Extract of the top 10 most common venues in London city sorted by district.

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abbey Wood	Supermarket	Grocery Store	Fast Food Restaurant	Lake	Train Station	Trail	Eastern European Restaurant	Pharmacy	Platform	Warehouse Store
1	Acton	Coffee Shop	Grocery Store	Pub	Gym / Fitness Center	Hotel	Middle Eastern Restaurant	Park	Hookah Bar	Bakery	Gastropub
2	Archway, Tufnell Park	Pub	Coffee Shop	Café	Pizza Place	Bakery	Italian Restaurant	Gastropub	Japanese Restaurant	Trail	Park
3	Balham	Pub	Coffee Shop	Park	French Restaurant	Italian Restaurant	Bakery	Pizza Place	Café	Bar	Burger Joint
4	Barnes, Castelnau	Pub	Park	Café	Coffee Shop	Gastropub	Farmers Market	Restaurant	Grocery Store	Italian Restaurant	Historic Site

Word clouds are commonly used to perform high-level analysis and visualization of text data and was applied to find the most common venues in London city using the previous DataFrame, still sorted by district, in a powerful image using WordCloud library. The results are shown below in Figure 7.

Figure 7. Word cloud displaying the top 10 most common venues in all districts of London city.

5.2. Clustering districts

This section consists in running k-means to cluster the district into 5 clusters. The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. Clustering process includes adding clustering labels, merging clusters to include cluster labels and adding latitudes and longitudes to each district. The resulting clusters are visualized in Figure 8.

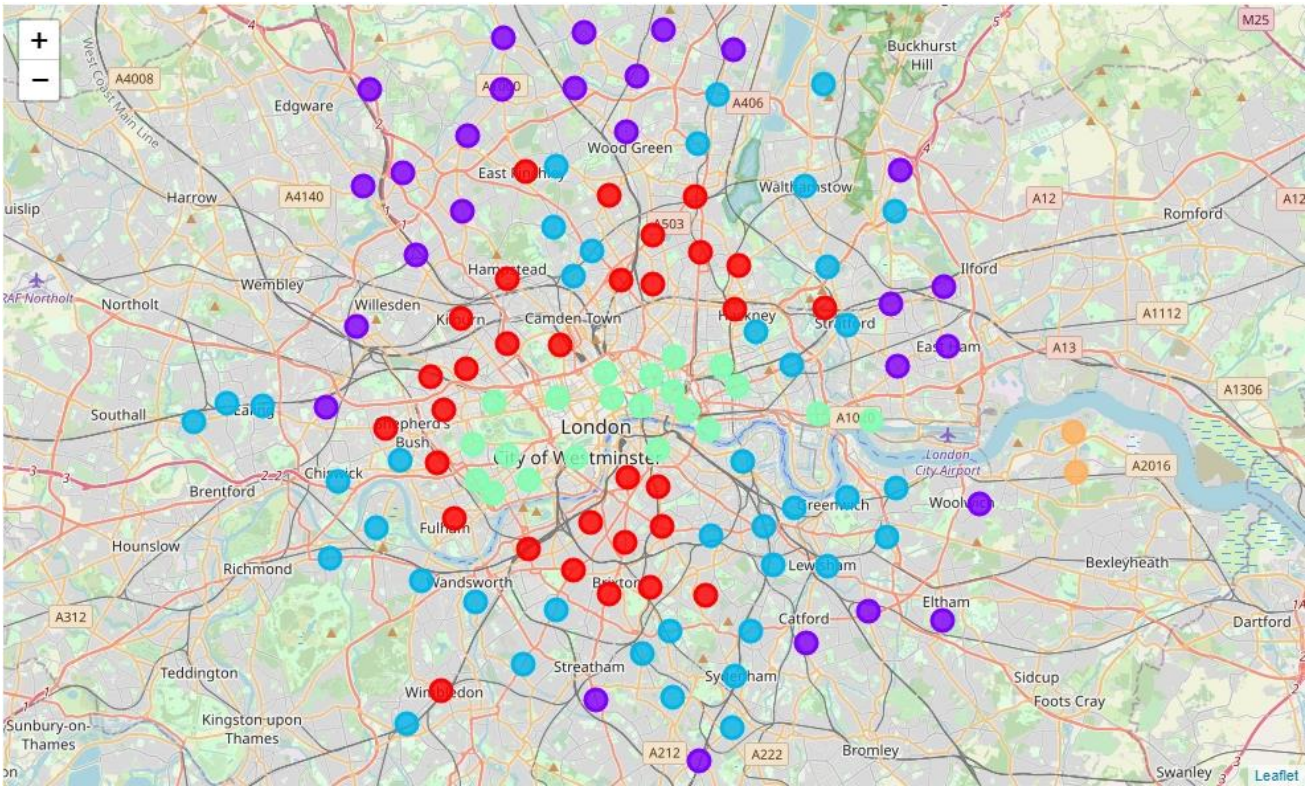

Figure 8. London city map with all clusters represented with different colors.

The legend of the London city map is summarized here below:

- Cluster 0: Red color (London north and south periphery)
- Cluster 1: Purple color (London north outskirts)
- Cluster 2: Light Blue color (London center and south)
- Cluster 3: Light Green color (London center and center-east)
- Cluster 4: Light Orange color (London east outskirts)

5.3. Examining the clusters

Each cluster was examined and set the discriminating venue categories that distinguish each cluster. Each cluster was visualized using word cloud with the purpose of looking for the most common venues in clusters. The resulting word clouds are presented in Figures 9-14.

Figure 9. Cluster 0 word cloud visualization.

Figure 10. Cluster 1 word cloud visualization.

Figure 11. Cluster 2 word cloud visualization.

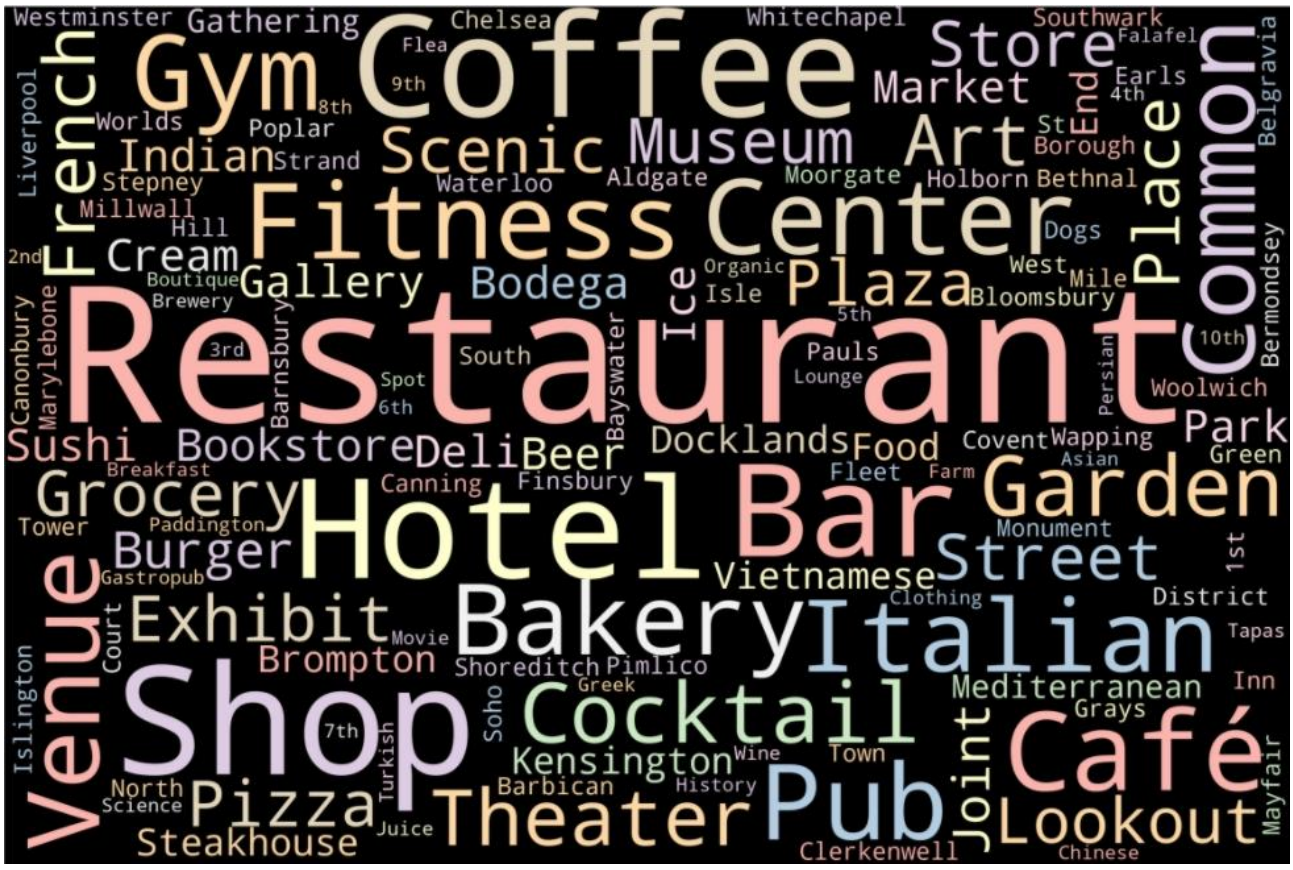

Figure 12. Cluster 3 word cloud visualization.

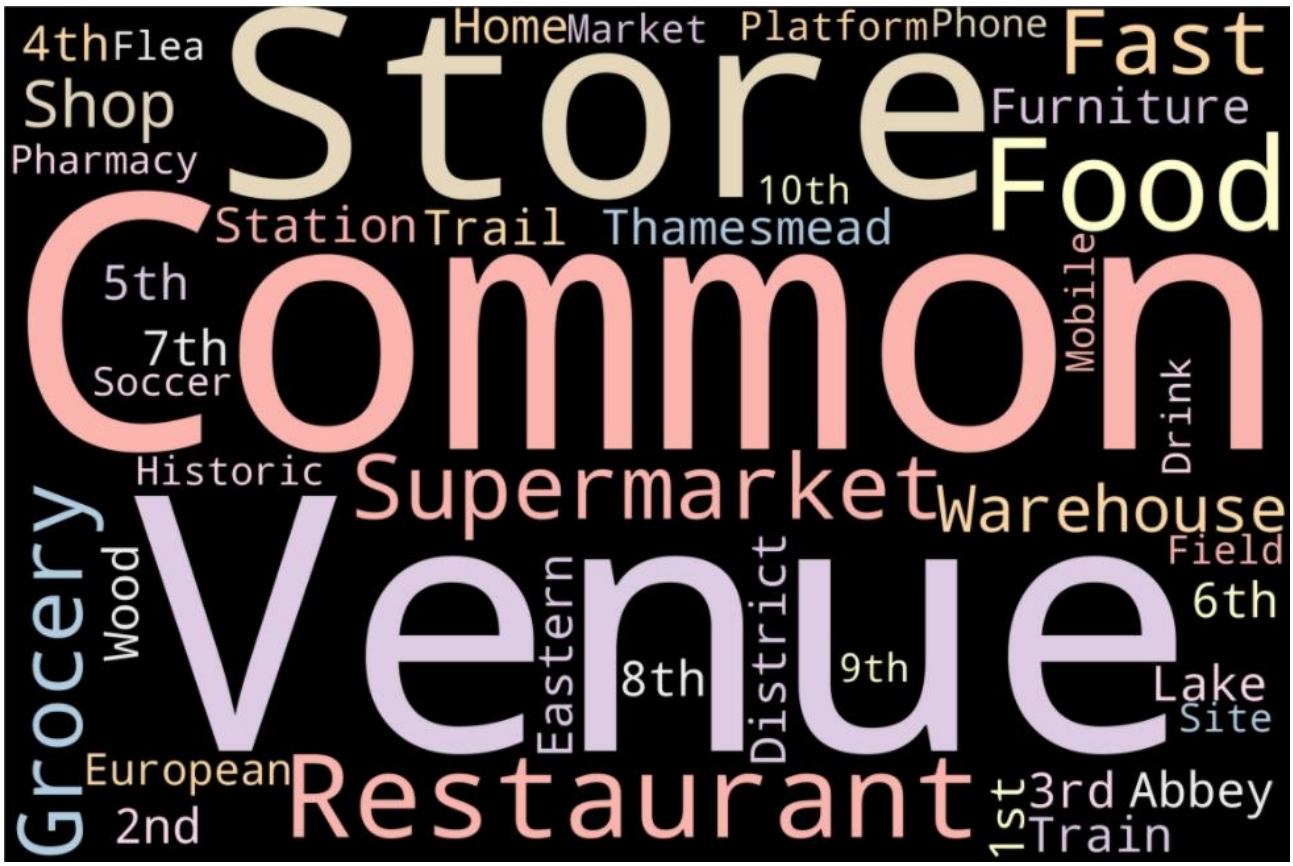

Figure 13. Cluster 4 word cloud visualization.

6. RESULTS AND DISCUSSION

The explanatory analysis revealed that there is a wide variety of venue categories in London city ranging from restaurants, bars, pubs, bakeries to bookstores, art museums, hotels, gyms and so on. It is important to highlight the presence of high number of train stations, which implies that London has a great connectivity to airports and surrounding cities, saving time and increasing mobility for visitors who spent an average of 4.8 nights in 2019 in London, according to Statista. However, as stated in the Business Problem section, a good planning is needed due to the high variety of activities and places to visit.

The scraping and data cleaning processes gave a total of 121 Postal Codes, which implies a total of 32 boroughs, while the London city map showed an uniform distribution of all the districts over London territory. A random analysis containing the most relevant revenues, from the point of view of a visitor, revealed that hotels, restaurants, bars and theaters are the most common venues in London city overall considering all districts. More specifically, the most popular international restaurants are: Italian (with 280 restaurants), Indian (with 167 restaurants), Turkish (with 142 restaurants) and Japanese (with 86 restaurants) in descending order. The difference between the total number of restaurants and the number of restaurants by cuisine type is mainly due to how each label has been categorized in the raw data.

A more detailed analysis of these venues by district has been performed. The districts Canning Town, North Woolwich and Docklands (situated in east and southeast of London city) have the highest hotel availability with a total of 16 hotels, while the districts Earls Court (southwest), Finsbury Park and Manor House (north)

resulted in the districts with the highest italian restaurants availability with a total of 6 each. Moreover, the districts East Ham (east-northeast), Forest Gate (east) and Upton Park (northeast) have the highest availability of indian restaurants with a total of 7 each.

A word cloud analysis of the most common venues of all districts has been developed. The most frequent words found were: Restaurant, Pub, Park, Coffee, Store, Shop, Italian, Hotel, Bakery, Grocery, Gym. This word cloud analysis validates the random analysis previously performed but also revealed other common venues that had not been analyzed yet such as pubs, parks, coffee shops, bakeries, groceries and gyms. A clustering analysis has been performed using k-means algorithm setting 5 clusters (cluster 0, cluster 1, cluster 2, cluster 3, cluster 4) using word cloud to get the most common venues in each cluster. Overall, the most common words found in all clusters were: restaurant, hotel, bar, coffee, pub, grocery, fitness, park, bakery and supermarket validating the previous analyses performed in this respect.

6.1. Limitations and recommendations for further research

There are a number of gaps in the present project that could be benefitted from further research. The explanatory analysis results depend on the accuracy of Foursquare data and the Milesfaster website has limitations as per the number of postal codes included. Furthermore, a more detailed evaluation should have included prices and iconic attractions in London city such as Buckingham Palace and Big Ben.

7. CONCLUSION

London is one of the most popular tourist destinations in Europe with full of things to do and visit. Leaving aside iconic attractions such as Buckingham Palace and Big Ben, this project explored a large number of districts aiming to find the most common venues and where are located geographically. From the visitors perspective, the most relevant venues found in the explanatory analysis were hotels, restaurants, bars, theaters, pubs, bakeries, groceries, coffee shops, gyms. The most popular cuisine types were italian food, followed by indian food, turkish food and japanese food. To sum up, most hotels were found in districts located in southwest and north of London city, while most restaurants were found in districts located in east and northeast of London city. By clustering all the districts with the purpose of creating major zones of interest (including a bigger number of potential locations) it was found that most of the venues revealed during the explanatory analysis are located in London center, center-east, south and north periphery too.

Although the decision of the optimal potential places to visit in London city will be taken by visitors, it can be concluded considering the results of the explanatory analysis that the most popular venues are located in the center, center-east and south periphery of London city. This recommendation is based on additional factors such as proximity of iconic attractions, availability of train stations and closeness to city center.