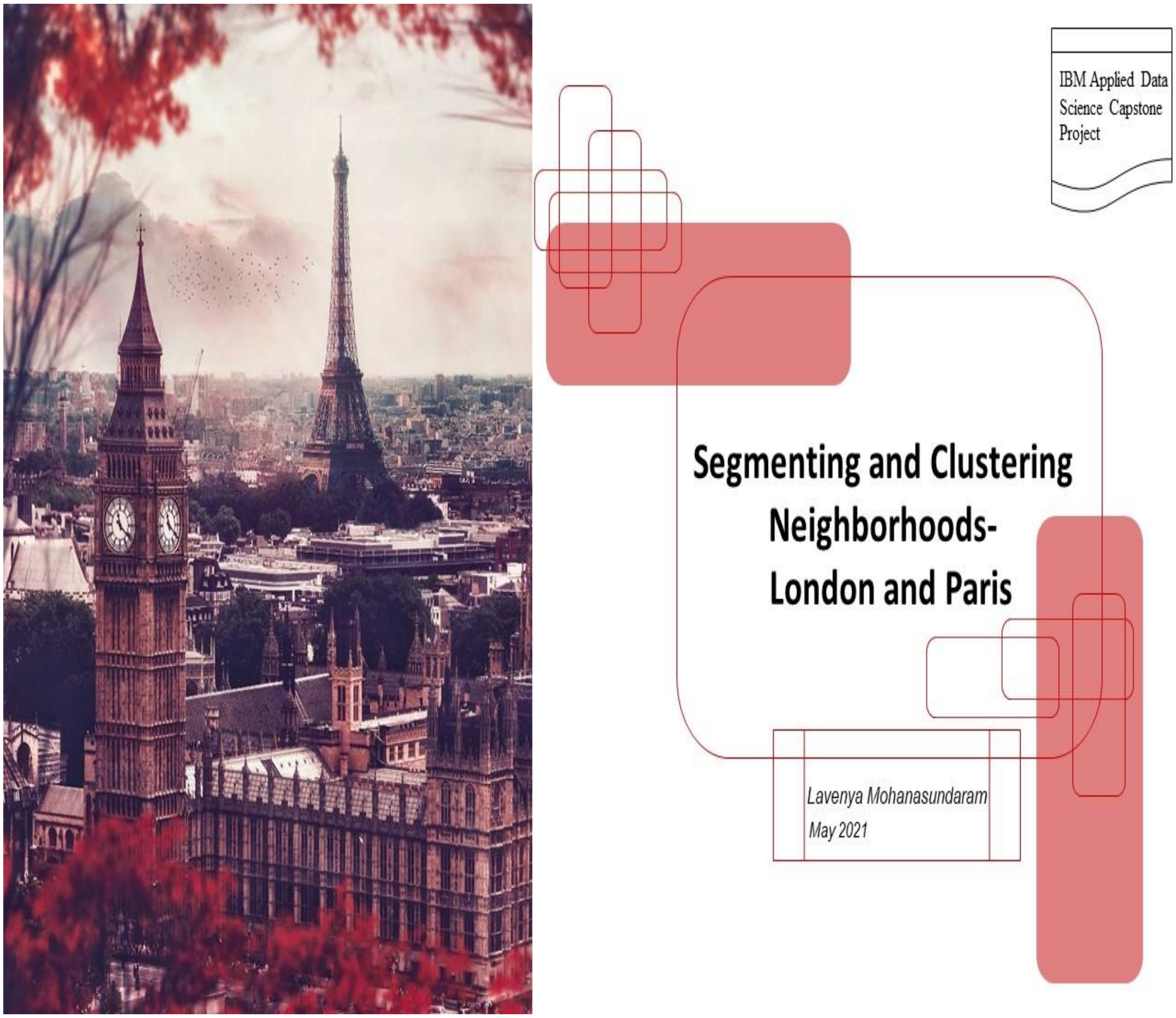


IBM APPLIED DATA SCIENCE CAPSTONE

BATTLE OF NEIGHBORHOODS – LONDON AND PARIS

USING MACHINE LEARNING WITH PYTHON

Table of Contents

1) Introduction

London is a leading global city. London is the capital of England and the United Kingdom; it is also the largest city within the country. It exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation. London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. The London metropolitan area is the third-most populous in Europe, after Istanbul and the Moscow Metropolitan Area, with 14,040,163 inhabitants in 2016.

Paris is the capital and most populous city of France, located in the north-central part of the nation. Since the 17th century, Paris has been one of Europe's major centers of finance, diplomacy, commerce, fashion, gastronomy, science and arts. The City of Paris is part of Île-de-France region, and it is considered as one of economic centers in Europe. It is multicultural city and provides many business opportunities. It was ranked as the second most visited travel destination in the world in 2019, after Bangkok and just ahead of London.

Both London and Paris are found at the heart of two great European nations. London and Paris are quite the popular tourist and vacation destinations for people all around the world. They are diverse and multicultural and offer a wide variety of experiences that is widely sought after.

This project can be useful for those who moves to these cities, to find a good area to build and grow prosperously. In order to get a very good location details that meet this need, the London and Paris are explored through clustering and segmentation based on the London and Paris Post code and proximity to supplies. We try to group the neighborhoods of London and Paris respectively and draw insights to what they look like now.

2) Business Problem

Besides the two being great cities, each of them has their unique winning points as compared to the other. So, if you are planning to embark on a trip or change your residence, and can't quite choose between the two, don't get all stressed up. The aim of this project is to help people to choose their destinations depending on the experiences that the neighborhoods have to offer and what they would want to have. The goal is to help stakeholders and globetrotters to make informed decisions and address any concerns they have including the different kinds of cuisines, provision stores and what the city has to offer.

2.1) Target Audience

The purpose of this project is to help people in exploring better facilities around their neighborhoods. It will help people making smart and efficient decision on selecting great neighborhoods out number of other postal area in both the cites London and Paris. Lots of people are migrating from various cities and needed lots of research for good housing prices, new business and reputed professional places for their children. The tourists can plan accordingly by choosing the neighborhoods in both cities.

This project is for those people who are looking for better neighborhoods and businesses. It will help people to get the awareness of area and neighborhood before visiting these big cities.

3) Data Acquisition

3.1) Data Description

This project will rely on geolocation data for both London and Paris. Postal codes in each city serve as a starting point. Using Postal codes, we use can find out the neighborhoods, boroughs, venues and their most popular venue categories.

For this project we need the following data:

3.1.1) London

To derive our solution, we scrape our data from web source

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

This Wikipedia page has information about all the neighborhoods, we limit it London.

1.borough: Name of Neighborhood
2.town: Name of borough
3.post_code: Postal codes for London.

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

The data for London looks like this:

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728
...	...	...	...	...	...	...
526	Woolwich	Greenwich	LONDON	SE18	020	TQ435795
527	Worcester Park	Sutton, Kingston upon Thames	WORCESTER PARK	KT4	020	TQ225655
528	Wormwood Scrubs	Hammersmith and Fulham	LONDON	W12	020	TQ225815
529	Yeading	Hillingdon	HAYES	UB4	020	TQ115825
530	Yiewsley	Hillingdon	WEST DRAYTON	UB7	020	TQ063804

531 rows × 6 columns

3.1.2) Paris

To derive our solution, we leverage JSON data available from web source

Data Source : https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e

The JSON file has data about all the neighborhoods in France, we limit it to Paris.

1.postal_code: Postal codes for France
2.nom_comm: Name of Neighborhoods in France
3.nom_dept: Name of the boroughs, equivalent to towns in France
4.geo_point_2d: Tuple containing the latitude and longitude of the Neighborhoods.

The data for Paris looks like this:

	datasetid	recordid	fields	geometry	record_timestamp
0	correspondances-code-insee-code-postal	2bf36b38314b6c39dfbcd09225f97fa532b1fc45	{'code_comm': '645', 'nom_dept': 'ESSONNE', 's...	{'type': 'Point', 'coordinates': [2.2517129721...	2016-09-21T00:29:06.175+02:00
1	correspondances-code-insee-code-postal	7ee82e74e059b443df18bb79fc5a19b1f05e5a88	{'code_comm': '133', 'nom_dept': 'SEINE-ET-MAR...	{'type': 'Point', 'coordinates': [3.0529405055...	2016-09-21T00:29:06.175+02:00
2	correspondances-code-insee-code-postal	e2cd3186f07286705ed482a10b6aebd9de633c81	{'code_comm': '378', 'nom_dept': 'ESSONNE', 's...	{'type': 'Point', 'coordinates': [2.1971816504...	2016-09-21T00:29:06.175+02:00
3	correspondances-code-insee-code-postal	868bf03527a1d0a9defe5cf4e6fa0a730d725699	{'code_comm': '243', 'nom_dept': 'SEINE-ET-MAR...	{'type': 'Point', 'coordinates': [2.7097808131...	2016-09-21T00:29:06.175+02:00
4	correspondances-code-insee-code-postal	21e809b1d4480333c8b6fe7addd8f3b06f343e2c	{'code_comm': '003', 'nom_dept': 'VAL-DE-MARNE...	{'type': 'Point', 'coordinates': [2.3335102498...	2016-09-21T00:29:06.175+02:00

3.2) Infrastructures Description

Different kinds of infrastructures in each neighborhood in London and Paris

Data Source:

- ✓ ArcGIS API
- ✓ Foursquare API

3.2.1) 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 neighborhoods of London. The following columns are added to our initial dataset which prepares our data.

1.latitude: Latitude for Neighborhood

2.longitude: Longitude for Neighborhood

3.2.2) Foursquare API

Venue Data:

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in London and Paris; it is used to study the popular venues of different neighborhoods as well as build the unsupervised learning model to cluster neighborhoods.

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. 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 neighborhoods, we then connect to the Foursquare API to gather information about venues inside every neighborhood. For each neighborhood, 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:

1.Neighborhood: Name of the Neighborhood
2.Neighborhood Latitude: Latitude of the Neighborhood
3.Neighborhood Longitude: Longitude of the Neighborhood
4.Venue: Name of the Venue
5.Venue Latitude: Latitude of Venue
6.Venue Longitude: Longitude of Venue
7.Venue Category: Category of Venue

Using these data collected for both London and Paris will allow exploration and examination to build our model. This is a project that will make use of many data science skills, from web scraping, working with API (ArcGIS and Foursquare), data cleaning, data wrangling and map visualization (Folium), Exploratory Data Analysis to perform unsupervised Machine Learning using K-means clustering and Natural Language Processing using word cloud.

4) Methodology

This section provides details for the methodology used in the project. We will be creating our model with the help of Python, so we start off by importing all the required packages.

Package breakdown:

- *Pandas*: To collect and manipulate data in JSON and HTML and then data analysis
- *requests*: Handle http requests
- *matplotlib*: Detailing the generated maps
- *folium*: Generating maps of London and Paris
- *sklearn*: To import K means which is the machine learning model that we are using.
- WordCloud: Data visualization technique used for representing text data in which the size of each word indicates its frequency.

The approach taken here is to explore each of the cities individually, plot the map to show the neighborhoods being considered and then build our model by clustering all of the similar neighborhoods together and finally plot the new map with the clustered neighborhoods. We draw insights and then compare and discuss our findings

4.1) Data Collection

In the data collection stage, we begin with collecting the required data for the cities of London and Paris. We need data that has the postal codes, neighborhoods and boroughs specific to each of the cities.

- To collect the available data for London, we scrape the List of areas of London Wikipedia page to take the 2nd table from https://en.wikipedia.org/wiki/List_of_areas_of_London
- To collect the available data for Paris, we download the JSON file containing all the postal codes of France from https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e Using Pandas, we load the table after reading the JSON file.

4.2) Data Preprocessing

For London, we replace the spaces with underscores in the title. The borough column has numbers within square brackets that we remove using:

we remove the spaces in the column titles and then we add _ between words.

```
wiki_london_data.rename(columns=lambda x: x.strip().replace(" ", "_"), inplace=True)
wiki_london_data
```

For Paris, we break down each of the nested fields and create the dataframe that we need:

We break down each of the nested fields and create the dataframe that we need

```
paris_field_data = pd.DataFrame()
for f in paris_raw.fields:
    dict_new = f
    paris_field_data = paris_field_data.append(dict_new, ignore_index=True)

paris_field_data.head()
```

4.3 Feature Selection

For both of our datasets, we need only the borough, neighborhood, postal codes and geolocations (latitude and longitude). So, we end up selecting the columns that we need by:

```
df1 = wiki_london_data.drop( [ wiki_london_data.columns[0], wiki_london_data.columns[4], wiki_london_data.columns[5] ], axis=1)
```

	London borough	Post_town	Postcode district
0	Bexley, Greenwich [7]	LONDON	SE2
1	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4
2	Croydon[8]	CROYDON	CR0
3	Croydon[8]	CROYDON	CR0
4	Bexley	BEXLEY, SIDCUP	DA5, DA14

```
df_2 = paris_field_data[['postal_code','nom_comm','nom_dept','geo_point_2d']]
df_2
```

	postal_code	nom_comm	nom_dept	geo_point_2d
0	91370	VERRIERES-LE-BUISSON	ESSONNE	[48.750443119964764, 2.251712972144151]
1	77126	COURCELLES-EN-BASSEE	SEINE-ET-MARNE	[48.41256065214989, 3.052940505560729]
2	91730	MAUCHAMPS	ESSONNE	[48.52726809075556, 2.19718165044305]
3	77400	LAGNY-SUR-MARNE	SEINE-ET-MARNE	[48.87307018579678, 2.7097808131278462]
4	94110	ARCUEIL	VAL-DE-MARNE	[48.80588035965699, 2.333510249842654]
...	...	...	...	...
1295	77520	CESSOY-EN-MONTOIS	SEINE-ET-MARNE	[48.50730730461658, 3.138844194183689]
1296	93420	VILLEPINTE	SEINE-SAINT-DENIS	[48.95902025378707, 2.536306342059409]
1297	77130	CANNES-ECLUSE	SEINE-ET-MARNE	[48.36403767307805, 2.990786679832767]
1298	78930	VILLETTE	YVELINES	[48.92627887061508, 1.6937417245662671]
1299	95270	LE PLESSIS-LUZARCHES	VAL-D'OISE	[49.09572967201378, 2.4547564431234923]

1300 rows × 4 columns

4.4 Feature Engineering

Both of our Datasets contains information related to all the cities in the country. We can narrow down and further process the data by selecting only the neighborhoods pertaining to 'London' and 'Paris'

```
df1 = df1[df1['town'].str.contains('LONDON')]
df1
```

	borough	town	post_code
0	Bexley, Greenwich	LONDON	SE2
1	Ealing, Hammersmith and Fulham	LONDON	W3, W4
6	City	LONDON	EC3
7	Westminster	LONDON	WC2
9	Bromley	LONDON	SE20
...	...	...	...
521	Redbridge	LONDON	IG8, E18
522	Redbridge, Waltham Forest	LONDON, WOODFORD GREEN	IG8
525	Barnet	LONDON	N12
526	Greenwich	LONDON	SE18
528	Hammersmith and Fulham	LONDON	W12

308 rows × 3 columns

```
df_paris = df_2[df_2['nom_dept'].str.contains('PARIS')].reset_index(drop=True)
df_paris
```

	postal_code	nom_comm	nom_dept	geo_point_2d
0	75009	PARIS-9E-ARRONDISSEMENT	PARIS	[48.87689616237872, 2.337460241388529]
1	75002	PARIS-2E-ARRONDISSEMENT	PARIS	[48.86790337886785, 2.344107166658533]
2	75011	PARIS-11E-ARRONDISSEMENT	PARIS	[48.85941549762748, 2.378741060237548]
3	75003	PARIS-3E-ARRONDISSEMENT	PARIS	[48.86305413181178, 2.359361058970589]
4	75006	PARIS-6E-ARRONDISSEMENT	PARIS	[48.84896809191946, 2.332670898588416]
5	75004	PARIS-4E-ARRONDISSEMENT	PARIS	[48.854228281954754, 2.357361938142205]
6	75010	PARIS-10E-ARRONDISSEMENT	PARIS	[48.87602855694339, 2.361112904561707]
7	75016	PARIS-16E-ARRONDISSEMENT	PARIS	[48.86039876035177, 2.262099559395783]
8	75008	PARIS-8E-ARRONDISSEMENT	PARIS	[48.8725272662346, 2.312582560420059]
9	75013	PARIS-13E-ARRONDISSEMENT	PARIS	[48.82871768452136, 2.362468228516128]
10	75012	PARIS-12E-ARRONDISSEMENT	PARIS	[48.83515623066034, 2.419807034965275]
11	75005	PARIS-5E-ARRONDISSEMENT	PARIS	[48.844508659617546, 2.349859385560182]
12	75019	PARIS-19E-ARRONDISSEMENT	PARIS	[48.88686862295828, 2.384694327870042]
13	75020	PARIS-20E-ARRONDISSEMENT	PARIS	[48.86318677744551, 2.400819826729021]
14	75007	PARIS-7E-ARRONDISSEMENT	PARIS	[48.85608259819694, 2.312438687733857]
15	75018	PARIS-18E-ARRONDISSEMENT	PARIS	[48.892735074561706, 2.348711933867703]
16	75017	PARIS-17E-ARRONDISSEMENT	PARIS	[48.88733716648682, 2.307485559493426]
17	75015	PARIS-15E-ARRONDISSEMENT	PARIS	[48.84015541860987, 2.293559372435076]
18	75001	PARIS-1ER-ARRONDISSEMENT	PARIS	[48.8626304851685, 2.336293446550539]
19	75014	PARIS-14E-ARRONDISSEMENT	PARIS	[48.82899321160942, 2.327100883257538]

4.5) Data Geocoding

In this project we make use of two infrastructures ArcGIS API and Foursquare API.

Looking over our London dataset, we can see that we don't have the geolocation data. We need to extrapolate the missing data for our neighborhoods. We perform this by leveraging the ArcGIS API. With the Help of ArcGIS API, we can get the latitude and longitude of our London neighborhood data.

As for our Paris dataset, we don't need to get the geo coordinates using an external data source or collect it with the ArcGIS API call since we already have it stored in the geo_point_2d column as a tuple in the df_paris dataframe.

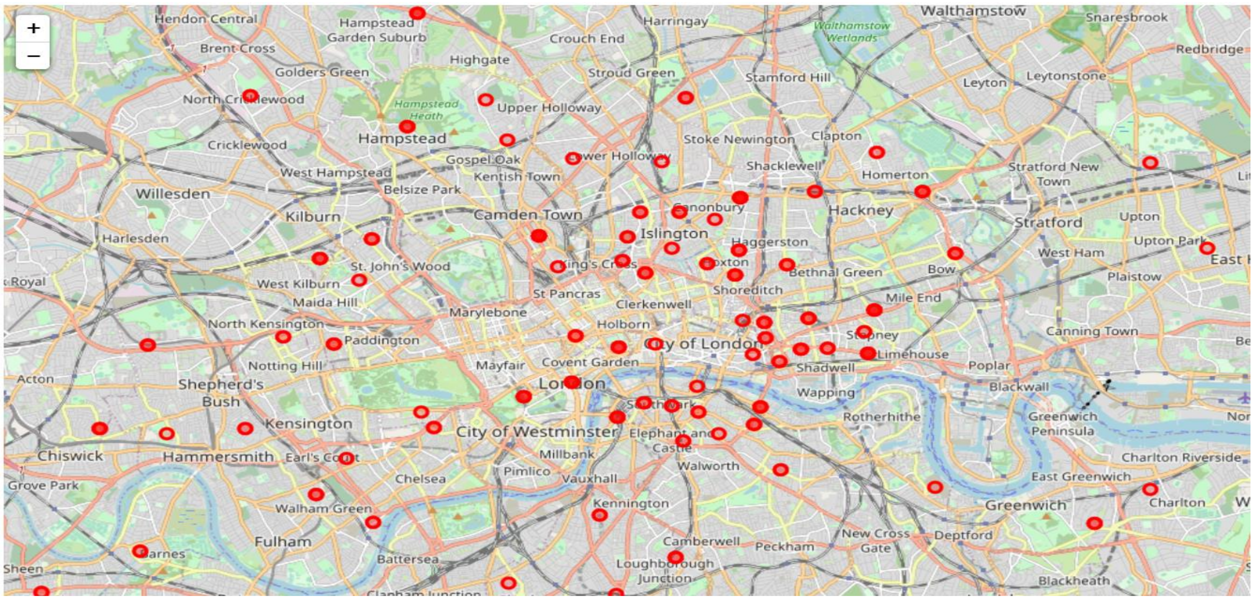
With the help of Foursquare API, a wonderful Geocoder package, we define a function which collects information pertaining to each neighborhood including that of the name of the neighborhood, geo-coordinates, venue and venue categories. After gathering the data, we will populate the data into a pandas dataframe.

4.6) Data Visualization

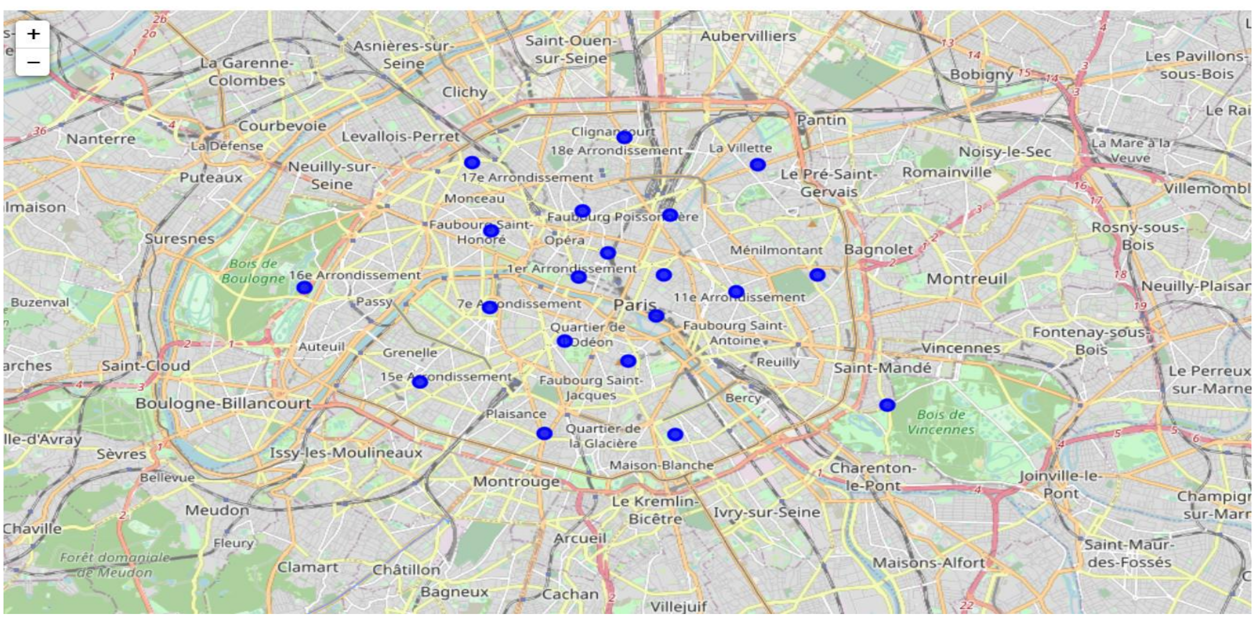
Visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the both cities London and Paris.

we can visualize the maps of London and Paris with the neighborhoods that we collected.

Neighborhood map of London:



Neighborhood map of Paris:



Now that we have visualized the neighborhoods, we need to find out what each neighborhood is like and what are the common venue and venue categories within a 500m radius.

4.7) Data Wrangling - One Hot Encoding

Since we are trying to find out what are the different kinds of venue categories present in each neighborhood and then calculate the top 10 common venues to base our similarity on, we use the One Hot Encoding to work with our categorical datatype of the venue categories. This helps to convert the categorical data into numeric data.

We won't be using label encoding in this situation since label encoding might cause our machine learning model to have a bias or a sort of ranking which we are trying to avoid by using One Hot Encoding.

We perform one hot encoding and then calculate the mean of the grouped venue categories for each of the neighborhoods in London.

Venue categories mean value

We will group the Neighbourhoods and calculate the mean venue categories value in each Neighbourhood

```
London_grouped = London_venue_cat.groupby('Neighbourhood').mean().reset_index()
London_grouped.head()
```

	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Aquarium	Arcade	Arepa Restaurant	...	Vegetarian / Vegan Restaurant	Video Game Store	Vie Re
0	Barnet	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.009747	0.0	0.0
1	Barnet, Brent, Camden	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0
2	Bexley	0.0	0.0	0.0	0.009434	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0
3	Bexley, Greenwich	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0
4	Bexley, Greenwich	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0

5 rows × 323 columns

We perform one hot encoding and then calculate the mean of the grouped venue categories for each of the neighborhoods in Paris like London data.

Venue categories mean value

We will group the Neighbourhoods and calculate the mean venue categories value in each Neighbourhood

```
Paris_grouped = Paris_venue_cat.groupby('Neighbourhood').mean().reset_index()
Paris_grouped.head()
```

	Neighbourhood	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Turkish Restaurant	Udon Restaurant
0	PARIS-10E-ARRONDISSEMENT	0.00000	0.02	0.0	0.0	0.0	0.0	0.00000	0.0	0.030000	...	0.0	0.0
1	PARIS-11E-ARRONDISSEMENT	0.02381	0.00	0.0	0.0	0.0	0.0	0.02381	0.0	0.023810	...	0.0	0.0
2	PARIS-12E-ARRONDISSEMENT	0.00000	0.00	0.0	0.0	0.0	0.0	0.00000	0.0	0.000000	...	0.0	0.0
3	PARIS-13E-ARRONDISSEMENT	0.00000	0.00	0.0	0.0	0.0	0.0	0.00000	0.0	0.186441	...	0.0	0.0
4	PARIS-14E-ARRONDISSEMENT	0.00000	0.00	0.0	0.0	0.0	0.0	0.00000	0.0	0.000000	...	0.0	0.0

5 rows × 200 columns

4.8) Top Venues in the Neighborhoods

In our next step, we need to rank and label the top venue categories in our neighborhoods of London and Paris distinctly.

```
neighborhoods_venues_sorted_london.head()
```

	Neighbourhood	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	Barnet	Pub	Coffee Shop	Park	Bakery	Café	Gastropub	Bus Stop	Cocktail Bar	Indian Restaurant	Train Station
1	Barnet, Brent, Camden	Park	Bus Stop	Pizza Place	Construction & Landscaping	Yoga Studio	Fast Food Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant
2	Bexley	Theater	Pub	Hotel	Monument / Landmark	Bakery	Plaza	English Restaurant	Ice Cream Shop	Cocktail Bar	Art Gallery
3	Bexley, Greenwich	Indian Restaurant	Pizza Place	Home Service	Grocery Store	Fishing Spot	Fish Market	English Restaurant	Escape Room	Ethiopian Restaurant	Event Space
4	Bexley, Greenwich	Lake	Construction & Landscaping	Yoga Studio	Filipino Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farm	Farmers Market

```
neighborhoods_venues_sorted_paris.head()
```

	Neighbourhood	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	PARIS-10E-ARRONDISSEMENT	French Restaurant	Hotel	Bistro	Café	Coffee Shop	Italian Restaurant	Indian Restaurant	Pizza Place	Asian Restaurant	Mediterranean Restaurant
1	PARIS-11E-ARRONDISSEMENT	Restaurant	Café	Italian Restaurant	French Restaurant	Bakery	Vietnamese Restaurant	Pastry Shop	Pizza Place	Cocktail Bar	Plaza
2	PARIS-12E-ARRONDISSEMENT	Zoo Exhibit	Zoo	Bistro	Monument / Landmark	Supermarket	Ethiopian Restaurant	Food & Drink Shop	Flower Shop	Fish Market	Fish & Chips Shop
3	PARIS-13E-ARRONDISSEMENT	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Thai Restaurant	French Restaurant	Juice Bar	Japanese Restaurant	Dessert Shop	Plaza	Coffee Shop
4	PARIS-14E-ARRONDISSEMENT	French Restaurant	Food & Drink Shop	Hotel	Japanese Restaurant	Sushi Restaurant	Bakery	Tea Room	Bistro	Fast Food Restaurant	Italian Restaurant

4.9) Exploratory Analysis– Model Building

4.9.1) K means Clustering – Machine Learning

The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a dataset. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We use K-means clustering technique to cluster the neighborhoods based on the category of venues near the neighborhoods. One important aspect of the k-means model is to determine the number of clusters to use in model development.

We will be using K-Means Clustering Machine learning algorithm to cluster similar neighborhoods together. We will be going with the number of clusters as 5.

K -Mean Clusters for London

```
london_data = london_merged

london_data = london_data.join(neighborhoods_venues_sorted_london.set_index('Neighbourhood'), on='borough')

london_data.head()
```

	borough	town	post_code	latitude	longitude	Cluster Labels	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 Mo Comm Ven
0	Bexley, Greenwich	LONDON	SE2	51.499741	0.124061	3	Lake	Construction & Landscaping	Yoga Studio	Filipino Restaurant	Event Space	Exhibit	Fabric Shop
1	Ealing, Hammersmith and Fulham	LONDON	W3, W4	51.497765	-0.255852	2	Coffee Shop	Park	Playground	Comedy Club	Café	Fish & Chips Shop	French Restaura
6	City	LONDON	EC3	51.513145	-0.078733	2	Coffee Shop	Hotel	Pub	Italian Restaurant	Gym / Fitness Center	Wine Bar	Cocktail Bar
7	Westminster	LONDON	WC2	51.514625	-0.114860	2	Hotel	Pub	Coffee Shop	Café	Sandwich Place	French Restaurant	Restaura
9	Bromley	LONDON	SE20	51.482490	0.119194	2	Bus Station	Campground	Athletics & Sports	Forest	Gym / Fitness Center	Café	Portugue Restaura

K -Mean Clusters for Paris

```
paris_data = paris_combined_data

paris_data = paris_data.join(neighborhoods_venues_sorted_paris.set_index('Neighbourhood'), on='nom_comm')

paris_data.head()
```

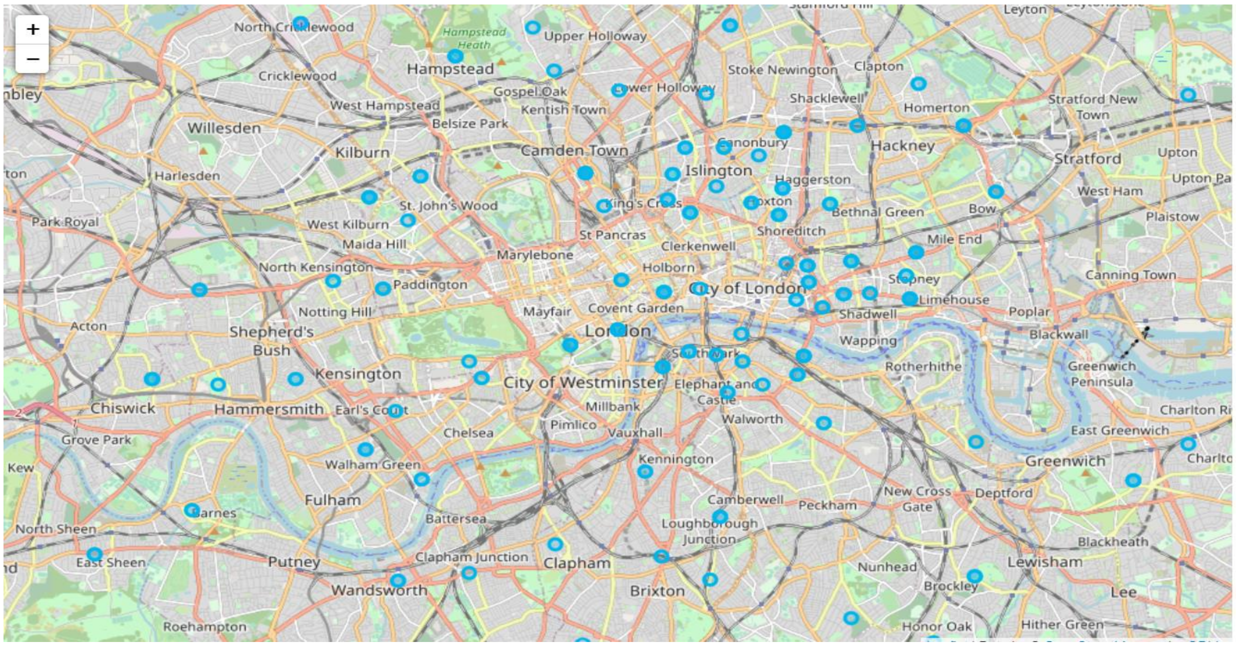
	postal_code	nom_comm	nom_dept	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Mo Comm Venu
0	75009	PARIS-9E-ARRONDISSEMENT	PARIS	48.876896	2.337460	3	French Restaurant	Hotel	Bistro	Japanese Restaurant	Wine Bar	Bakery
1	75002	PARIS-2E-ARRONDISSEMENT	PARIS	48.867903	2.344107	3	French Restaurant	Cocktail Bar	Bakery	Wine Bar	Hotel	Indie Movi Theater
2	75011	PARIS-11E-ARRONDISSEMENT	PARIS	48.859415	2.378741	3	Restaurant	Café	Italian Restaurant	French Restaurant	Bakery	Vietnames Restaurar
3	75003	PARIS-3E-ARRONDISSEMENT	PARIS	48.863054	2.359361	3	French Restaurant	Japanese Restaurant	Coffee Shop	Bakery	Gourmet Shop	Art Galler
4	75006	PARIS-6E-ARRONDISSEMENT	PARIS	48.848968	2.332671	3	French Restaurant	Bakery	Chocolate Shop	Cocktail Bar	Restaurant	Pastry Shop

5) Results

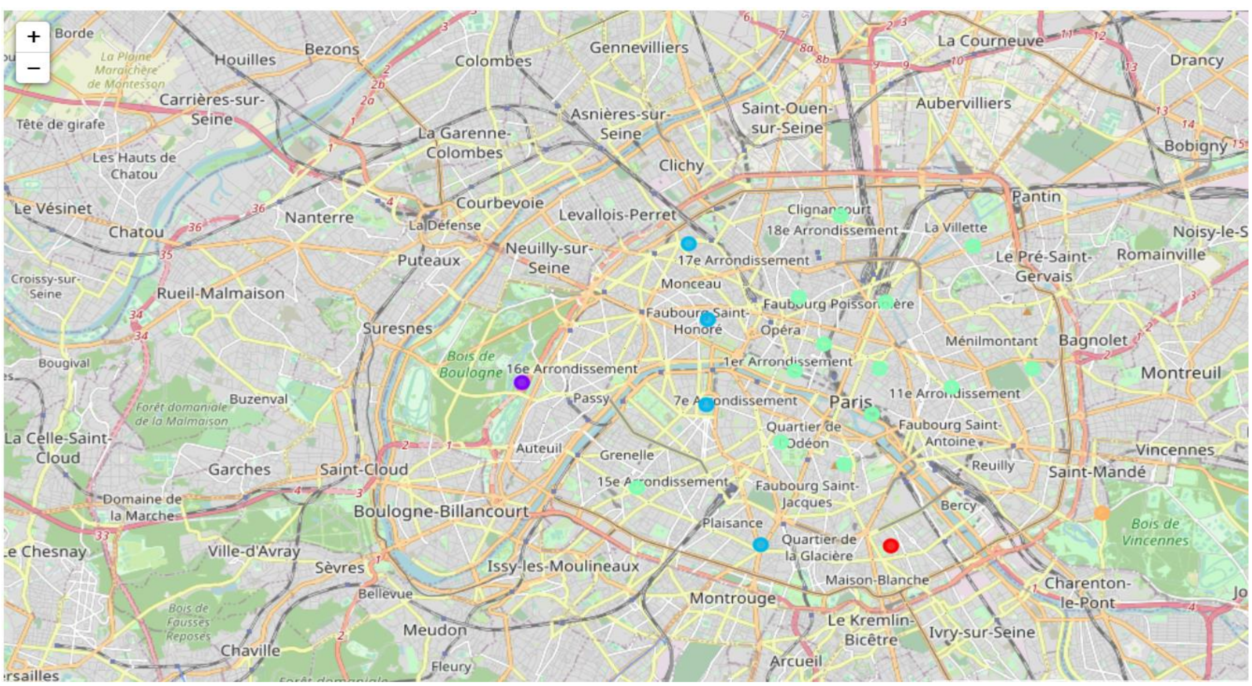
5.1) Visualizing the clustered Neighborhoods

Our data is processed, missing data is collected and compiled. The Model is built. All that's remaining is to see the clustered neighborhoods on the map. Again, we use Folium package to do so.

Map of clustered neighborhoods of London:



Map of clustered neighborhoods of Paris:



5.2) Examining our Clusters

The results from K – Mean Clustering show that we can categorize the neighborhoods into 5 clusters based on the frequency of occurrence.

<u>Clusters in London</u>

Cluster 1:

	town	Cluster Labels	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
121	LONDON	1	Pizza Place	Furniture / Home Store	Park	Yoga Studio	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farm

Cluster 2:

	town	Cluster Labels	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
6	LONDON	2	Coffee Shop	Hotel	Pub	Italian Restaurant	Gym / Fitness Center	Wine Bar	Cocktail Bar	Restaurant	French Restaurant	Sandwich Place
7	LONDON	2	Hotel	Pub	Coffee Shop	Café	Restaurant	Sandwich Place	French Restaurant	Tea Room	Lounge	Art Gallery
9	LONDON	2	Forest	Campground	Bus Stop	Athletics & Sports	Café	Coffee Shop	Gym / Fitness Center	Grocery Store	Park	Japanese Restaurant
10	LONDON	2	Pub	Café	Coffee Shop	Bar	Cocktail Bar	Sandwich Place	Thai Restaurant	Bus Stop	Grocery Store	Indian Restaurant
12	LONDON	2	Pub	Café	Coffee Shop	Bar	Cocktail Bar	Sandwich Place	Thai Restaurant	Bus Stop	Grocery Store	Indian Restaurant
...	...	...	...	...	...	...	...	...	...	...	...	...
521	LONDON	2	Hotel	Coffee Shop	Indian Restaurant	Café	Pub	Pizza Place	Sandwich Place	Gym / Fitness Center	Bar	Korean Restaurant
522	LONDON, WOODFORD GREEN	2	Hotel	Coffee Shop	Pub	Indian Restaurant	Café	Sandwich Place	Monument / Landmark	Theater	Art Gallery	Cocktail Bar
525	LONDON	2	Pub	Coffee Shop	Park	Café	Bakery	Gastropub	Cocktail Bar	Bus Stop	Indian Restaurant	Train Station
526	LONDON	2	Pub	Coffee Shop	Bar	Hotel	Gym / Fitness Center	Café	Italian Restaurant	Bakery	Grocery Store	Thai Restaurant
528	LONDON	2	Coffee Shop	Pub	Café	Grocery Store	Pizza Place	Italian Restaurant	Hotel	Pharmacy	Indian Restaurant	Gastropub

294 rows × 12 columns

Cluster 3:

	town	Cluster Labels	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	LONDON	3	Lake	Yoga Studio	Escape Room	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant

Cluster 4:

	town	Cluster Labels	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
167	LONDON, WELLING	4	Indian Restaurant	Construction & Landscaping	Grocery Store	Fast Food Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farm
457	LONDON, ERITH	4	Indian Restaurant	Construction & Landscaping	Grocery Store	Fast Food Restaurant	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farm

Cluster 5:

	town	Cluster Labels	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
1	LONDON	5	Coffee Shop	Bus Stop	Grocery Store	Park	Argentinian Restaurant	Mediterranean Restaurant	French Restaurant	Comedy Club	Health Food Store	Café
34	LONDON	5	Park	Train Station	Coffee Shop	Tennis Court	Fast Food Restaurant	Bus Stop	Grocery Store	Playground	Comedy Club	Café
99	LONDON	5	Coffee Shop	Bus Stop	Grocery Store	Park	Argentinian Restaurant	Mediterranean Restaurant	French Restaurant	Comedy Club	Health Food Store	Café
141	LONDON	5	Park	Train Station	Coffee Shop	Tennis Court	Fast Food Restaurant	Bus Stop	Grocery Store	Playground	Comedy Club	Café
196	LONDON	5	Coffee Shop	Bus Stop	Grocery Store	Park	Argentinian Restaurant	Mediterranean Restaurant	French Restaurant	Comedy Club	Health Food Store	Café
198	LONDON	5	Coffee Shop	Bus Stop	Grocery Store	Park	Argentinian Restaurant	Mediterranean Restaurant	French Restaurant	Comedy Club	Health Food Store	Café
214	LONDON	5	Park	Train Station	Coffee Shop	Tennis Court	Fast Food Restaurant	Bus Stop	Grocery Store	Playground	Comedy Club	Café
452	LONDON	5	Café	Coffee Shop	Park	Portuguese Restaurant	Grocery Store	Gym / Fitness Center	Japanese Restaurant	Pharmacy	Falafel Restaurant	Escape Room
453	LONDON	5	Café	Coffee Shop	Park	Portuguese Restaurant	Grocery Store	Gym / Fitness Center	Japanese Restaurant	Pharmacy	Falafel Restaurant	Escape Room
499	LONDON	5	Park	Train Station	Coffee Shop	Tennis Court	Fast Food Restaurant	Bus Stop	Grocery Store	Playground	Comedy Club	Café

<u>Clusters in Paris</u>

Cluster 1:

	nom_comm	Cluster Labels	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
10	PARIS-16E-ARRONDISSEMENT	1	Plaza	Lake	Art Museum	Park	Bus Station	Boat or Ferry	French Restaurant	Pool	Gym / Fitness Center	Gym

Cluster 2:

	nom_comm	Cluster Labels	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
3	PARIS-8E-ARRONDISSEMENT	2	French Restaurant	Hotel	Spa	Art Gallery	Corsican Restaurant	Japanese Restaurant	Furniture / Home Store	Resort	Mediterranean Restaurant	Cocktail Bar
14	PARIS-7E-ARRONDISSEMENT	2	French Restaurant	Hotel	Italian Restaurant	Café	Plaza	History Museum	Cocktail Bar	Bistro	Dessert Shop	Cafeteria
16	PARIS-17E-ARRONDISSEMENT	2	French Restaurant	Hotel	Italian Restaurant	Café	Restaurant	Bakery	Japanese Restaurant	Bistro	Plaza	Diner
19	PARIS-14E-ARRONDISSEMENT	2	French Restaurant	Hotel	Japanese Restaurant	Café	Laundromat	Fast Food Restaurant	Tea Room	Bakery	Bistro	Plaza

Cluster 3:

	nom_comm	Cluster Labels	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 Cor
0	PARIS-9E-ARRONDISSEMENT	3	French Restaurant	Hotel	Japanese Restaurant	Bistro	Cocktail Bar	Lounge	Wine Bar	Bakery	Pizza Place	Bar
1	PARIS-2E-ARRONDISSEMENT	3	French Restaurant	Cocktail Bar	Bakery	Wine Bar	Hotel	Salad Place	Bar	Spa	Italian Restaurant	Sandv Place
2	PARIS-11E-ARRONDISSEMENT	3	Restaurant	Café	Pastry Shop	Italian Restaurant	Vietnamese Restaurant	French Restaurant	Asian Restaurant	Bakery	Afghan Restaurant	Sandv Place
6	PARIS-3E-ARRONDISSEMENT	3	French Restaurant	Japanese Restaurant	Coffee Shop	Art Gallery	Italian Restaurant	Gourmet Shop	Bakery	Wine Bar	Cocktail Bar	Burge Joint
7	PARIS-6E-ARRONDISSEMENT	3	Chocolate Shop	Bakery	French Restaurant	Plaza	Cocktail Bar	Fountain	Theater	Italian Restaurant	Pastry Shop	Resta
8	PARIS-4E-ARRONDISSEMENT	3	French Restaurant	Ice Cream Shop	Clothing Store	Pastry Shop	Hotel	Pedestrian Plaza	Plaza	Park	Wine Bar	Italian Resta
9	PARIS-10E-ARRONDISSEMENT	3	French Restaurant	Hotel	Bistro	Café	Coffee Shop	Indian Restaurant	Asian Restaurant	Italian Restaurant	Pizza Place	Burge Joint
11	PARIS-5E-ARRONDISSEMENT	3	French Restaurant	Hotel	Italian Restaurant	Plaza	Bakery	Café	Coffee Shop	Pub	Bar	Lebar Resta
12	PARIS-19E-ARRONDISSEMENT	3	French Restaurant	Bar	Pizza Place	Brewery	Seafood Restaurant	Bistro	Beer Bar	Supermarket	Hotel	Conce Hall
13	PARIS-20E-ARRONDISSEMENT	3	Plaza	Japanese Restaurant	Bakery	Bistro	French Restaurant	Italian Restaurant	Pizza Place	Café	Bar	Hotel
15	PARIS-18E-ARRONDISSEMENT	3	Bar	French Restaurant	Pizza Place	Bistro	Plaza	Restaurant	Café	Italian Restaurant	Supermarket	Vietna Resta
17	PARIS-15E-ARRONDISSEMENT	3	Italian Restaurant	French Restaurant	Hotel	Brasserie	Restaurant	Thai Restaurant	Lebanese Restaurant	Indian Restaurant	Japanese Restaurant	Plaza
18	PARIS-1ER-ARRONDISSEMENT	3	French Restaurant	Japanese Restaurant	Plaza	Hotel	Italian Restaurant	Art Museum	Coffee Shop	Cheese Shop	Thai Restaurant	Theat

Cluster 4:

	nom_comm	Cluster Labels	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
5	PARIS-12E-ARRONDISSEMENT	4	Zoo Exhibit	Zoo	Bistro	Monument / Landmark	Supermarket	Donut Shop	Fish & Chips Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant

Cluster 5:

	nom_comm	Cluster Labels	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
4	PARIS-13E-ARRONDISSEMENT	5	Vietnamese Restaurant	Asian Restaurant	Thai Restaurant	Chinese Restaurant	French Restaurant	Juice Bar	Hotel	Bus Stop	Bookstore	Sandwich Place

5.3) WordCloud -Natural Language Processing

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. For generating word cloud in Python, modules needed are – matplotlib, pandas and wordcloud.

In this project, the wordcloud tool that permit to have a global idea of domination of a category over others. Word cloud is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

We built a word cloud for each 1st Most Common Venue in London to discover dominant Venues and it gives:

We see that Pub, Café, Coffee shop, Hotel, Bus Station, Park and Indian Restaurant are dominant in London (1st Most Common Venue). We see a category named Restaurant. It is a category of restaurant that is not specialized in certain national dishes.

We built a word cloud for each 1st Most Common Venue in Paris to discover dominant Venues and it gives:

We see that French, Restaurant, Plaza, Zoo, Exhibit, Italian, Vietnamese and Japanese are dominant in Paris (1st Most Common Venue). We see a category named Restaurant. It is a category of restaurant that is not specialized in certain national dishes.

6) Discussion

The neighborhoods in London and Paris cluster model provides a label for each neighborhood which is representative of the cluster it belongs to. The cluster labels were then added to the dataframe. The results from the k-means clustering show that we can categorize the neighborhoods into 5 clusters based on the frequency of occurrence for both the cities individually.

Neighborhoods in clusters based on the frequency of occurrence:

Cluster Label	Clusters in London	Clusters in Paris
Cluster 1	Neighborhoods with a low number of frequencies (1 Record)	Neighborhoods with a low number of frequencies (1 Record)
Cluster 2	Neighborhoods with a high number of frequencies (294 Records)	Neighborhoods with a moderate number of frequencies (4 Records)
Cluster 3	Neighborhoods with a low number of frequencies (1 Record)	Neighborhoods with a high number of frequencies (13 Records)
Cluster 4	Neighborhoods with a low number of frequencies (2 Records)	Neighborhoods with a low number of frequencies (1 Record)
Cluster 5	Neighborhoods with a moderate number of frequencies (10 Records)	Neighborhoods with a low number of frequencies (1 Record)

By analyzing these five clusters obtained for cities London and Paris, we can see that some of the clusters are more suited for restaurants, café, plaza, art museums and hotels. These clusters contain a higher degree of restaurants, hotels, multiplex, cafes, bars, other food joints and low degree other of venues like train station, bus station, fish market, gym, performing arts venue and smoke shop, to name a few.

As a cosmopolitan city, the neighborhoods of London offer an eclectic mixture of classic British and multicultural cuisine including Indian, Italian, Turkish and Chinese. London seems to take a step further in this direction by having a lot of Pubs, Restaurants, bars, coffee shops, Fish and Chips shop, Bus Station, Theater and Lake. It has a lot of shopping malls. The main modes of transport seem to be Buses and trains. For leisure, the neighborhoods are set up to have lots of parks, golf courses, zoo, gyms and Historic sites. Thus, the city of London offers a multicultural, diverse and certainly an entertaining experience.

Paris is relatively small geographically. It has a wide variety of cuisines and eateries including French, Thai, Cambodian, Asian, Chinese etc. There are a lot of hangout spots including many Restaurants and Bars. Paris has a lot of Bistro's. Different means of public transport in Paris which includes buses, bikes, boats or ferries. For leisure and sight-seeing, there are a lot of Plazas, Trails, Parks, Zoo, Historic sites, clothing shops, Art galleries and Museums. Thus, Paris seems like the relaxing vacation spot with a mix of lakes, historic spots and a wide variety of cuisines to try out.

7) Conclusion

In this project, I have gone through the process of identifying the business problems, specifying the data required, extracting and preparing the data, visualizing the results, performing machine learning by clustering the data into 5 clusters based on their frequency similarities, tackling and reaching to a definitive solution to business problems for both the cities London and Paris.

The purpose of this project was to explore the cities of London and Paris and see how attractive it is to potential tourists and migrants. We explored both the cities based on their postal codes and then extrapolated the common venues present in each of the neighborhoods finally concluding with clustering similar neighborhoods together.

London and Paris are both vibrant and cultural cities, with a fascinating history and incredible heritage. They are both world class cities and are similar in many ways, but very different in others. We could see that each of the neighborhoods in both the cities have a wide variety of experiences to offer which is unique.

London and Paris seem to offer a vacation stay or ardent getaway with a lot of places to explore, beautiful landscapes, amazing food and a wide variety of culture. London has many venues to explore than Paris. When it comes to integrated transport network, London is best served with 6 international airports (2 in Paris) and almost twice as many bus lines and more overland train lines than Paris. Inclusively, it's up to the stakeholders, immigrants and globetrotters desire to decide which city is preferable more and according to their fondness and considering the factors determined in this project.