12/27/2020

Clustering and Comparing Neighborhoods in London and Paris

IBM Data Science Professional Certificate

Veronique Labeau
COURSERA

TABLE OF CONTENTS

INTRODUCTION

In this project, we will explore, segment and cluster the neighborhoods of two large European cities: **London**, the capital of England, and **Paris**, the capital of France.

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

While London is the capital of England, it is also the largest city within the country. The city stands on River Thames in South East England, with its history stretching back to the Roman times.

On the other hand, Paris, the capital of France, is in the north-central part of the nation. Like London, the city also stands along a river, commonly known as the Seine River. Paris has a rich European history and is regarded to be a global center for culture, fashion, art, and gastronomy.

Business problem

London vs Paris, which is your favorite destination? Picking a favorite city when it comes to London vs Paris can be such a tricky task. It is always a struggle since both cities are famous for their soul-refreshing experiences, as well as iconic attractions. Most tourists find it hard to pick one-holiday destination between the two.

Our goal here is to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. The objective is to help tourists to choose their destinations depending on the experiences that the neighbourhoods have to offer and what they would like to do. This model will also help people to make decisions if they are thinking about migrating to London or Paris or even if they wish to relocate neighbourhoods within the city. Our findings will help stakeholders make informed decisions and address any concerns they have, including the different kinds of cuisines, provision stores and what the city has to offer.

DATA DESCRIPTION

This section describes the processes of acquiring, cleaning, and preparing each dataset used in this project for the next stages. We need geographical location data for both London and Paris to segment and explore their neighborhoods. For both cities, we will essentially need a dataset that contains the boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhoods.

The data below will be used for this analysis. Once downloaded and cleaned up, the data will be combined into one table:

- Borough
- Neighborhood
- Latitude
- Longitude

Data 1: London boroughs

London has a total of 32 boroughs. To explore, analyze and segment them, their longitude and latitude will be added using the link to the following dataset:

- We will scrape London data from the Wikipedia page https://en.wikipedia.org/wiki/List_of_areas_of_London, which has information about all the boroughs. For our study, we will filter the dataset only on the boroughs with London in Post-town column.
- *Locations coordinates*: Since this Wikipedia page lacks information about geographical locations, we will use *ArcGIS API* to complete the dataset with the longitude and latitude data for the boroughs of London.

	Location	Londonborough	Post town	Dialcode	OS grid ref	Postcodedistrict
0	Abbey Wood	Bexley, Greenwich	LONDON	020	TQ465785	SE2
1	Acton	Ealing, Hammersmith and Fulham	LONDON	020	TQ205805	W3
1	Acton	Ealing, Hammersmith and Fulham	LONDON	020	TQ205805	W4
6	Aldgate	City	LONDON	020	TQ334813	EC3
7	Aldwych	Westminster	LONDON	020	TQ307810	WC2

	Borough	Neighborhood	Latitude	Longitude
0	Bexley, Greenwich	Abbey Wood	51.49245	0.12127
1	Ealing, Hammersmith and Fulham	Acton	51.51324	-0.26746
2	Ealing, Hammersmith and Fulham	Acton	51.48944	-0.26194
3	City	Aldgate	51.51200	-0.08058
4	Westminster	Aldwych	51.51651	-0.11968

Data 2: Paris boroughs and neighborhoods

Paris has a total of 20 boroughs (called *arrondissements* in French) and 80 neighborhoods (called *quartiers* in French). To explore, analyze and segment neighborhoods, the longitude and latitude of each of them, as well as the boroughs, will be added using the links to the following dataset:

- *Paris Boroughs*: https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e which is a JSON file data about all the boroughs in France. We will focus our study only on Paris.
- *Paris Neighborhoods*: https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=json&timezone=Europe/Berlin which contains the different neighborhoods of Paris.

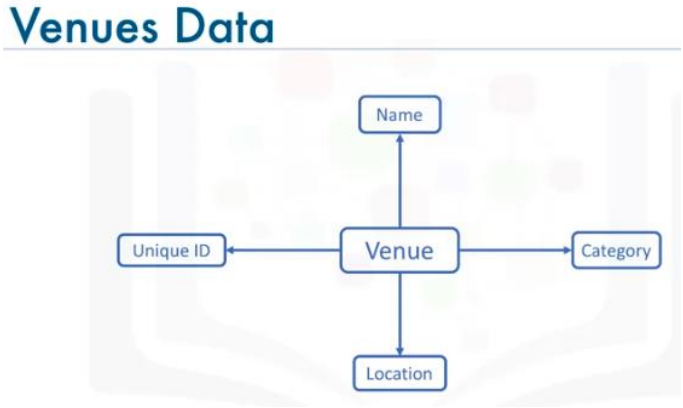
	Borough	IdNeighborhood
0	PARIS-9E-ARRONDISSEMENT	9
1	PARIS-2E-ARRONDISSEMENT	2
2	PARIS-11E-ARRONDISSEMENT	11
3	PARIS-15E-ARRONDISSEMENT	15
4	PARIS-19E-ARRONDISSEMENT	19

	IdNeighborhood	Neighborhood	Latitude	Longitude
0	12	Quinze-Vingts	48.846916	2.374402
1	6	Notre-Dame-des-Champs	48.846428	2.327357
2	14	Petit-Montrouge	48.826653	2.326437
3	19	Pont-de-Flandre	48.895556	2.384777
4	16	Muette	48.863275	2.259936

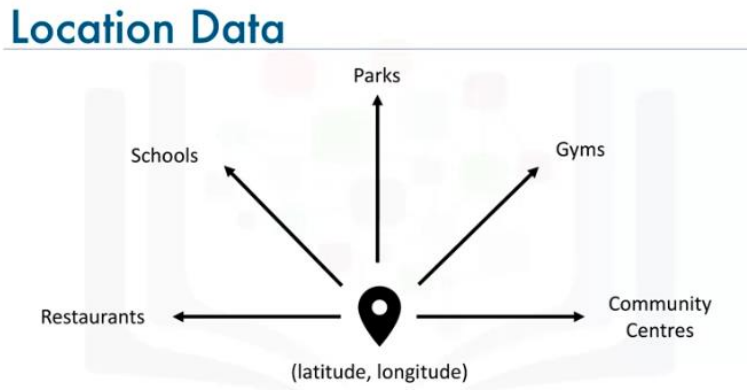
	Borough	Neighborhood	Latitude	Longitude
0	PARIS-9E-ARRONDISSEMENT	Rochechouart	48.879812	2.344861
1	PARIS-9E-ARRONDISSEMENT	Saint-Georges	48.879934	2.332850
2	PARIS-9E-ARRONDISSEMENT	Chaussée-d'Antin	48.873547	2.332269
3	PARIS-9E-ARRONDISSEMENT	Faubourg-Montmartre	48.873935	2.343253
4	PARIS-2E-ARRONDISSEMENT	Gaillon	48.869307	2.333432

Data 3: Venues data

The Venues data describes the venues (restaurants, cafes, parks, museums, etc.) in each neighborhood of the two cities by category. For each neighbourhood, we have chosen the radius to be 500 meters.



To gain that information, we will use "Foursquare" location information (https://Foursquare.com). 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 and returned via a JSON file.



METHODOLOGY

Analytical Approach of the Problem

In this project, the approach taken is to:

- Use the Foursquare API to explore the neighborhoods in London and Paris,
- Get the most common venue categories in each neighborhood,
- Use the *k-Means* clustering algorithm to find similar neighborhoods,
- Use the Folium library to visualize the neighborhoods in both London and Paris and their emerging clusters.

We drew insights and then compare and discuss our findings.

Data Understanding and Preparation

To segment the neighborhoods of both cities and explore them, we essentially need a dataset that contains the related boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of them.

In the data collection stage, we begin by 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. Gaps in data have been identified and plans to either fill or make substitutions have been made accordingly.

London Dataset

A dataset that specifies the neighborhood data for London from Wikipedia page:

- First, we scrape the *List of areas of London Wikipedia* page to get the list of London boroughs and remove any footnote reference numbers with [] in the borough names:

	Location	Londonborough	Post town	Postcodedistrict	Dialcode	OS grid ref
0	Abbey Wood	Bexley, Greenwich	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

Figure 1- London boroughs raw data

- We process only the boroughs that contains the word *London* in the *postal town* column and keep only the columns needed for the next step.
- Finally, we use the Geocoder package with the *arcgis_geocoder* to obtain the latitude and longitude of the needed locations and create the data frame for our study:

	Borough	Neighborhood	Latitude	Longitude
0	Bexley, Greenwich	Abbey Wood	51.49245	0.12127
1	Ealing, Hammersmith and Fulham	Acton	51.51324	-0.26746
2	City	Aldgate	51.51200	-0.08058
3	Westminster	Aldwych	51.51651	-0.11968
4	Bromley	Anerley	51.41009	-0.05683

Figure 2 London data frame

- With the coordinates of London neighborhoods, we use Geopy and Folium libraries to create a map of London with marks for each neighborhood, where name of borough is located and name of neighborhood is superimposed on top. The figure below shows this map, each red circle represents the location of one neighborhood.

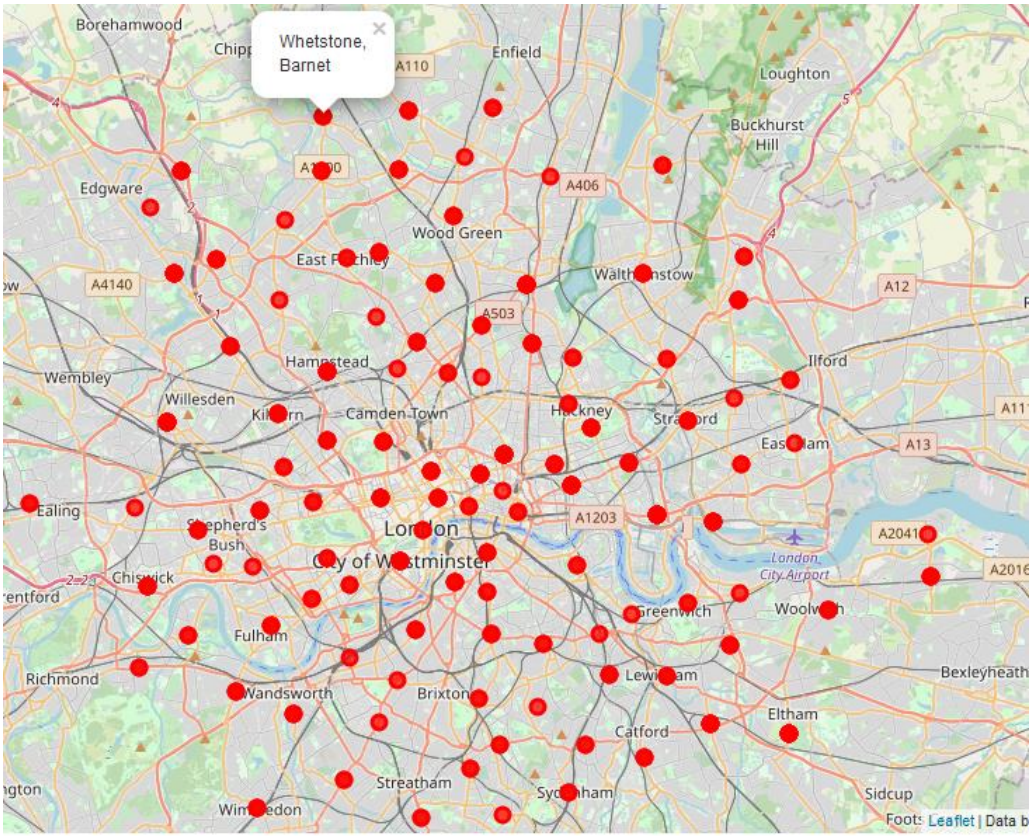


Figure 3 London Map

Paris Dataset

For Paris, the dataset is originally two JSON files that specifies each boroughs and neighborhoods and related coordinates:

- We start by downloading a JSON file containing all the postal codes of France from https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e

```
{'datasetid': 'correspondances-code-insee-code-postal',
 'recordid': '2bf36b38314b6c39dfbcd09225f97fa532b1fc45',
 'fields': {'code_comm': '645',
  'nom_dept': 'ESSONNE',
  'statut': 'Commune simple',
  'z_moyen': 121.0,
  'nom_region': 'ILE-DE-FRANCE',
  'code_reg': '11',
  'insee_com': '91645',
  'code_dept': '91',
  'geo_point_2d': [48.750443119964764, 2.251712972144151],
  'postal_code': '91370',
  'id_geofla': '16275',
  'code_cant': '03',
  'geo_shape': {'type': 'Polygon',
   'coordinates': [[[2.238024349288764, 48.735565859837095],
     [2.226414985434264, 48.75003536744732],
     [2.22450256558849, 48.75882853410981],
     [2.232859032169924, 48.76598806763034],
     [2.250043759055985, 48.761213267519565],
     [2.269288614654887, 48.76063999654954],
     [2.276145972515501, 48.75666127305422],
     [2.283691112862691, 48.748081131389654],
     [2.274517407535147, 48.74072222671912],
     [2.238024349288764, 48.735565859837095]]]},
  'superficie': 999.0,
  'nom_comm': 'VERRIERES-LE-BUISSON',
  'code_arr': '3',
  'population': 15.5},
 'geometry': {'type': 'Point',
  'coordinates': [2.251712972144151, 48.750443119964764]},
 'record_timestamp': '2016-09-21T00:29:06.175+02:00'}
```


Figure 4 JSON Paris boroughs data

- From which we create the following data frame that contains the different boroughs in Paris. Each borough in Paris, called in French *arrondissement,* is determined by a number (from 1 to 20):

	Borough	IdNeighborhood
0	PARIS-9E-ARRONDISSEMENT	9
1	PARIS-2E-ARRONDISSEMENT	2
2	PARIS-11E-ARRONDISSEMENT	11
3	PARIS-15E-ARRONDISSEMENT	15
4	PARIS-19E-ARRONDISSEMENT	19

Figure 5 Paris Arrondissements

- We continue the scraping to complete the data with the neighborhoods and coordinates by downloading a JSON file from this link: https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=json&timezone=Europe/Berlin

```
{'datasetid': 'quartier_paris',
 'recordid': '10c5545818e09beacaff49049e14e571ac404d5c',
 'fields': {'n_sq_qu': 750000015,
  'perimetre': 2878.55965556,
  'geom_x_y': [48.851585175, 2.36476795387],
  'c_qu': 15,
  'surface': 487264.93707154,
  'l_qu': 'Arsenal',
  'geom': {'type': 'Polygon',
   'coordinates': [[[2.368512371393433, 48.85573412813671],
     [2.369003319617131, 48.853741288126166],
     [2.369103414954773, 48.85331137972449],
     [2.369105862149444, 48.8533007254104],
     [2.369106875143625, 48.853296302340944],
     [2.3691141164928142, 48.85326541359504],
     [2.369130104070962, 48.85319674240403],
     [2.369134878036132, 48.85317614812273],
     [2.36913799648954, 48.85316284245766],
     [2.369130098683958, 48.85314585791725],
```


Figure 6 JSON Paris neighborhoods

- We extract the needed information and merge it to the boroughs previously pulled up to obtain the Paris data frame for our study:

	Borough	Neighborhood	Latitude	Longitude
0	PARIS-9E-ARRONDISSEMENT	Rochechouart	48.879812	2.344861
1	PARIS-9E-ARRONDISSEMENT	Saint-Georges	48.879934	2.332850
2	PARIS-9E-ARRONDISSEMENT	Chaussée-d'Antin	48.873547	2.332269
3	PARIS-9E-ARRONDISSEMENT	Faubourg-Montmartre	48.873935	2.343253
4	PARIS-2E-ARRONDISSEMENT	Vivienne	48.869100	2.339461

Figure 7 Paris data frame

- As with London, the following Figure shows a map of Paris and its neighborhood. Each blue circle represents the location of one neighborhood

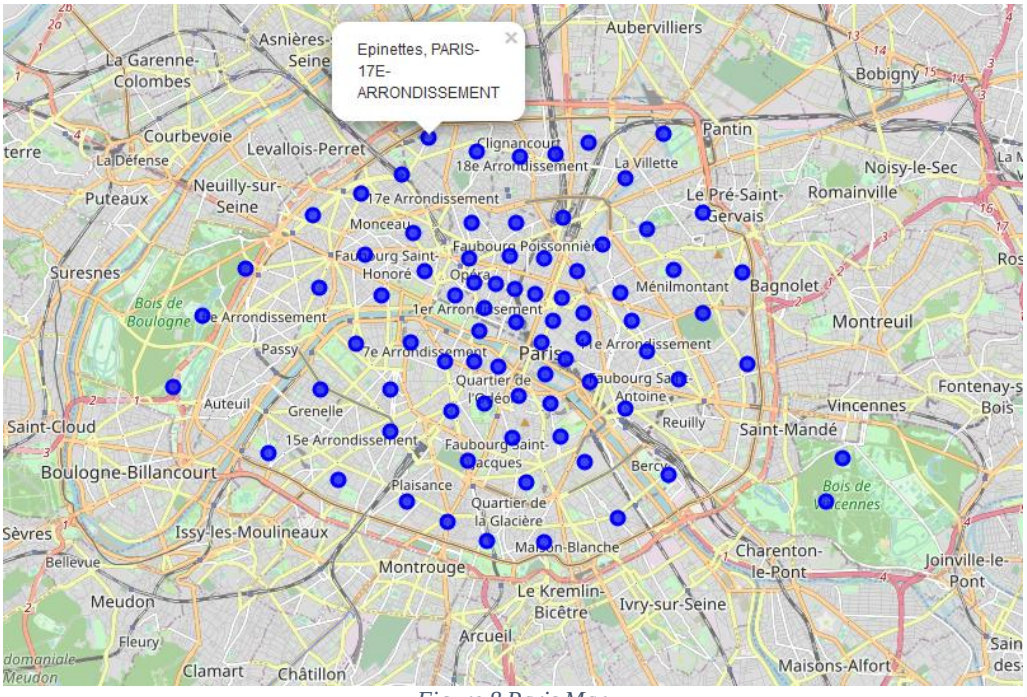


Figure 8 Paris Map

Venues data

For each city, data that describes the venues and the venues categories for each neighborhood is needed. Venues data is retrieved from Foursquare which is a popular source of location and venue data. Foursquare API service is utilized to access and download venues data.

To retrieve data from Foursquare using their API, a URL should be prepared and used to request data related to a specific location. An example URL is the following:

```
https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll
={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)
```

where

- `explore` indicates the API endpoint used,
- `client_id` and `client_secret` are credentials used to access the API service and are obtained when registering a Foursquare developer account,
- `v i`ndicates the API version to use,
- `ll` indicates the latitude and longitude of the desired location,
- `radius` is the maximum distance in meters between the specified location and the retrieved venues,
- `limit` is used to limit the number of returned results if necessary.

Once data frames for both cities are cleaned, we use the Foursquare API to retrieve the information of the common venues in London and Paris neighborhoods.

London venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	51.49245	0.12127	Lesnes Abbey	51.489526	0.125839	Historic Site
1	Abbey Wood	51.49245	0.12127	Sainsbury's	51.492826	0.120524	Supermarket
2	Abbey Wood	51.49245	0.12127	Lidl	51.496152	0.118417	Supermarket
3	Abbey Wood	51.49245	0.12127	Abbey Wood Railway Station (ABW)	51.490825	0.123432	Train Station
4	Abbey Wood	51.49245	0.12127	Bean @ Work	51.491172	0.120649	Coffee Shop

Figure 9 London Venues Data frame

Paris venues

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rochechouart	48.879812	2.344861	Mamiche	48.880112	2.343699	Bakery
1	Rochechouart	48.879812	2.344861	Mikkeller Bar Paris	48.878663	2.345377	Beer Bar
2	Rochechouart	48.879812	2.344861	Les 36 Corneil	48.878997	2.345501	Wine Bar
3	Rochechouart	48.879812	2.344861	Le Potager de Charlotte	48.878924	2.344640	Vegetarian / Vegan Restaurant
4	Rochechouart	48.879812	2.344861	La Ferme Saint Hubert	48.878908	2.345428	Cheese Shop

Figure 10 Paris Venues Data frame

Exploratory Data Analysis

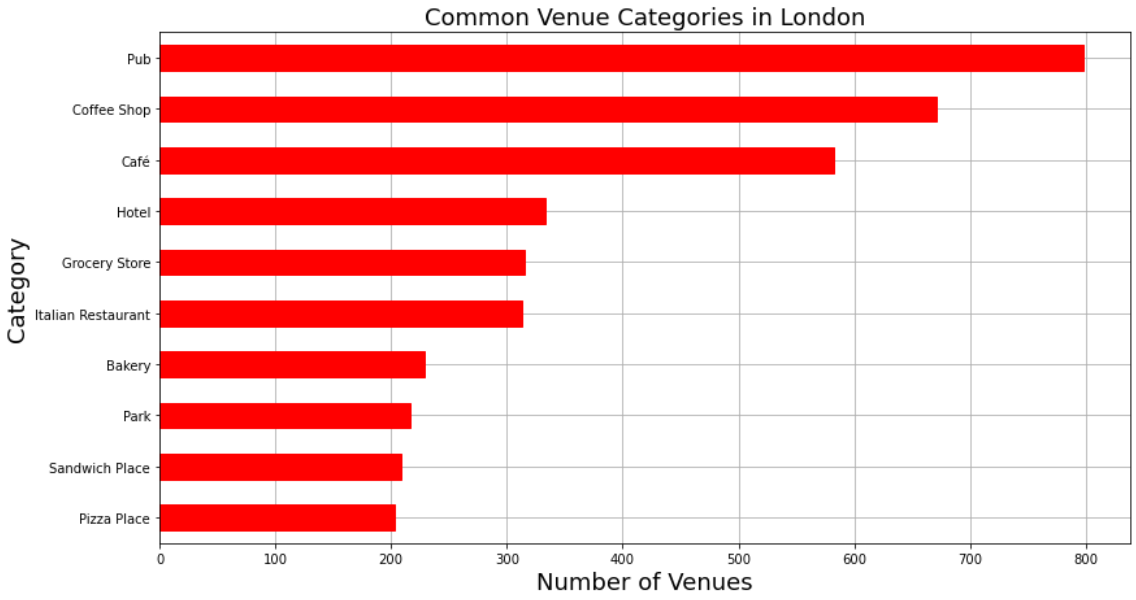
In this section, the datasets produced in the previous section are explored via effective visualizations to better understand the data.

Most Common Venue categories

What are the categories that have more venues than the others in London and in Paris?

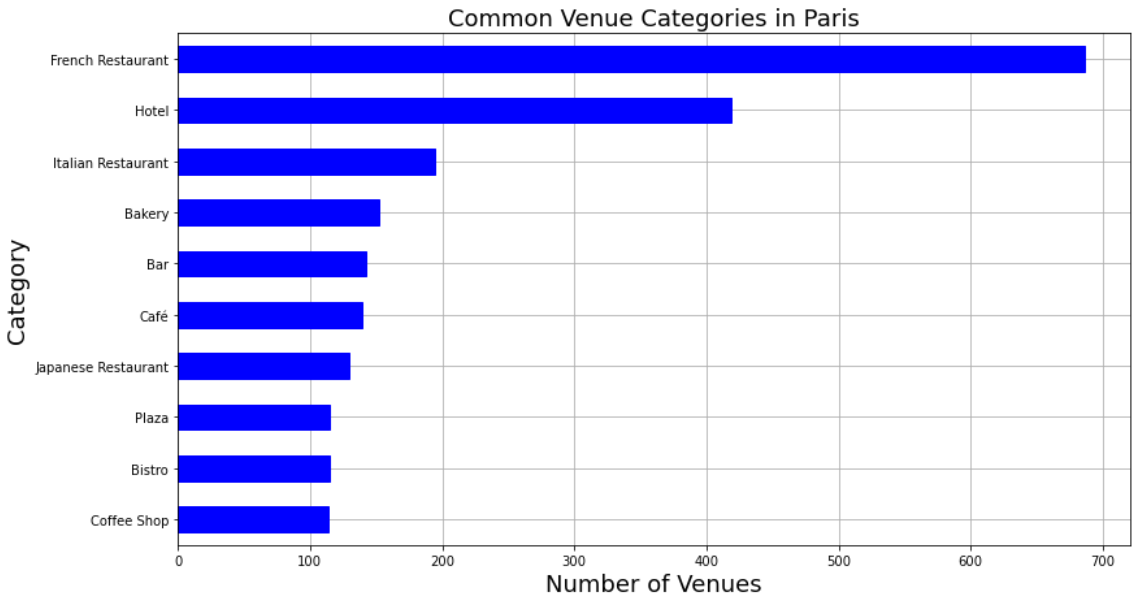
To answer this question, the number of occurrences is counted for each venue category for both cities. The bar plots below show the popularity of the most common venue categories in each city.

London Top 10 venue categories



- The 10 most common venues categories show mostly venues related to food and beverage such as Pub, Coffee shop, Cafe, Grocery store, Sandwich and Pizza place. The 1st type of restaurant listed is Italian.
- The other type of venues are Hotel (ranked 4th) and Park (ranked 8th).

Paris Top 10 Venue Categories



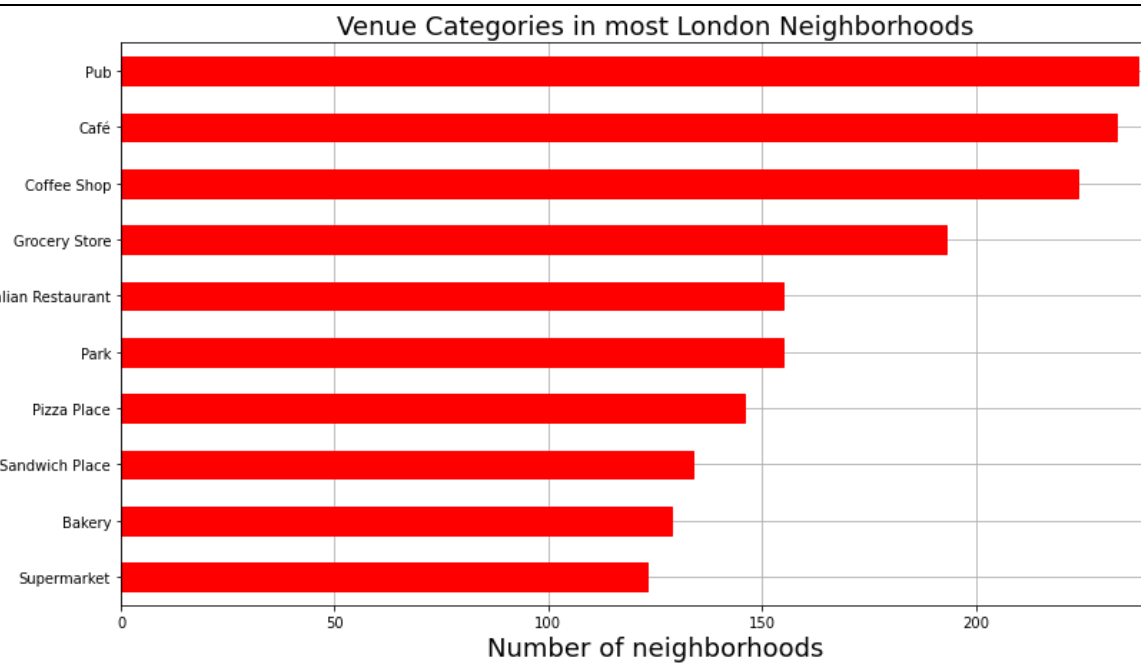
- The 10 most common venue categories show 3 types of restaurants (French, Italian and Japanese) compared to London. While Hotel is ranked in the 2nd position compared to London.

There are similarities between the most common categories in London and in Paris. Indeed, many categories appear in both plots, although some names differ based on cultural terminology: a *Pub* in London is called a *Bar* in Paris.

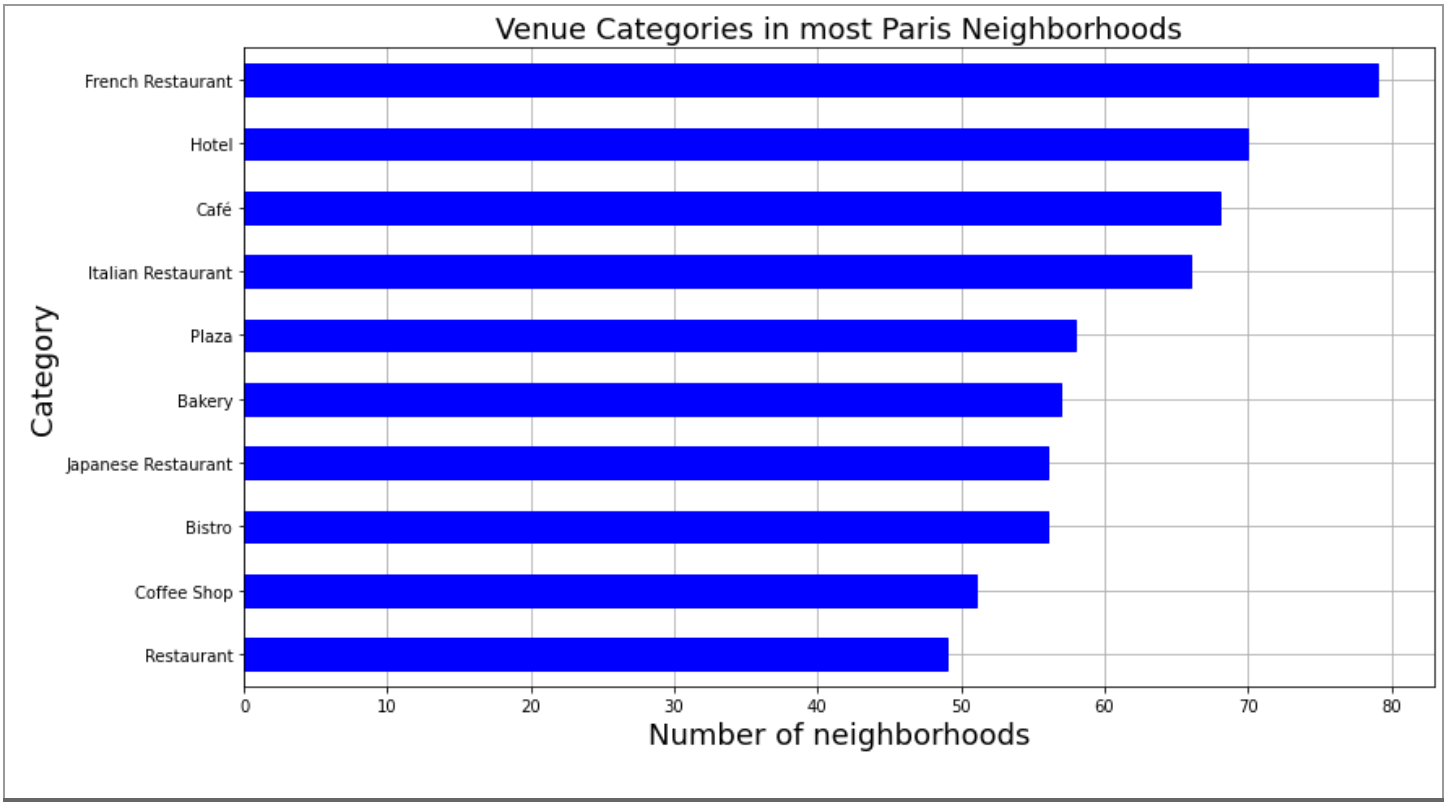
Most Widespread Venue Categories

Now, what are the venue categories that exist in most neighborhood?

London 10 most widespread venue categories



Paris 10 most widespread venue categories



Clustering of Neighborhoods

In this section, clustering is applied on London and Paris neighborhoods to find similar neighborhoods in the two cities. Clustering is the process of finding similar items in a dataset based on the characteristics (features) of items in the dataset.

k-Means Machine learning uses

For segmenting neighborhoods in London and Paris, clustering algorithm is used especially K-mean method. This method is used as unsupervised algorithm. It does not need previous recommendations to build a model. K-mean method is good for segmentation.

K-Means is a type of partitioning clustering, which means it divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar.

Feature Selection

The goal of the clustering is to cluster neighborhoods based on the similarity of venue categories in the neighborhoods. This means that the two things of interest here are the neighborhood and the venue categories in the neighborhood. Thus, the following two features will be selected out of the London and Paris data frames: "Neighborhood" and "Venue Category".

Since the clustering algorithm works only with numerical features, one-hot encoding is then applied on the "Venue Category" feature and the result of the encoding is used for clustering. Once one-hot encoding is applied on the London and Paris data, they are then combined.

The resulting data frame looks like shown below. The venue categories listed in the London data frame are converted to the value 1 as shown in the first row in Figure 11 ("Historic Site" column; and the same applies for all rows).

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Historic Site	History Museum	Hookah Bar	Hostel
0	Abbey Wood	0	0	0	0	1	0	0	0
1	Abbey Wood	0	0	0	0	0	0	0	0
2	Abbey Wood	0	0	0	0	0	0	0	0
3	Abbey Wood	0	0	0	0	0	0	0	0
4	Abbey Wood	0	0	0	0	0	0	0	0

Figure 11 One-Hot encoding London Results

The resulting data frame for Paris after applying the same operations.

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant
0	Saint-Georges	0	0	0	0	0	0	0	0	0
1	Saint-Georges	0	0	0	0	0	0	0	0	0
2	Saint-Georges	0	0	0	0	0	0	0	0	0
3	Saint-Georges	0	0	0	0	0	0	0	0	0
4	Saint-Georges	0	0	0	0	0	0	0	0	0

Figure 12 One-Hot encoding Paris Results

The next step is to aggregate the values for each neighborhood so that each neighborhood is represented by only one row. The aggregation is done by grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category.

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery
0	Abbey Wood	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0
1	Acton	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0
2	Aldgate	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.011364	0.0
3	Aldwych	0.010638	0.0	0.0	0.0	0.010638	0.010638	0.0	0.010638	0.0
4	Anerley	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0

Figure 13 London Aggregated data frame sample

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	Alsatian Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant
0	Amérique	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
1	Archives	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
2	Arsenal	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00
3	Arts-et-Métiers	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01
4	Auteuil	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00

Figure 14 Paris Aggregated data frame sample

Clusters Neighborhoods

Then we apply the clustering algorithm on each data. We create a data frame to show the neighborhoods of London and Paris, the cluster to which each neighborhood belongs, and the most common venue categories in each neighborhood.

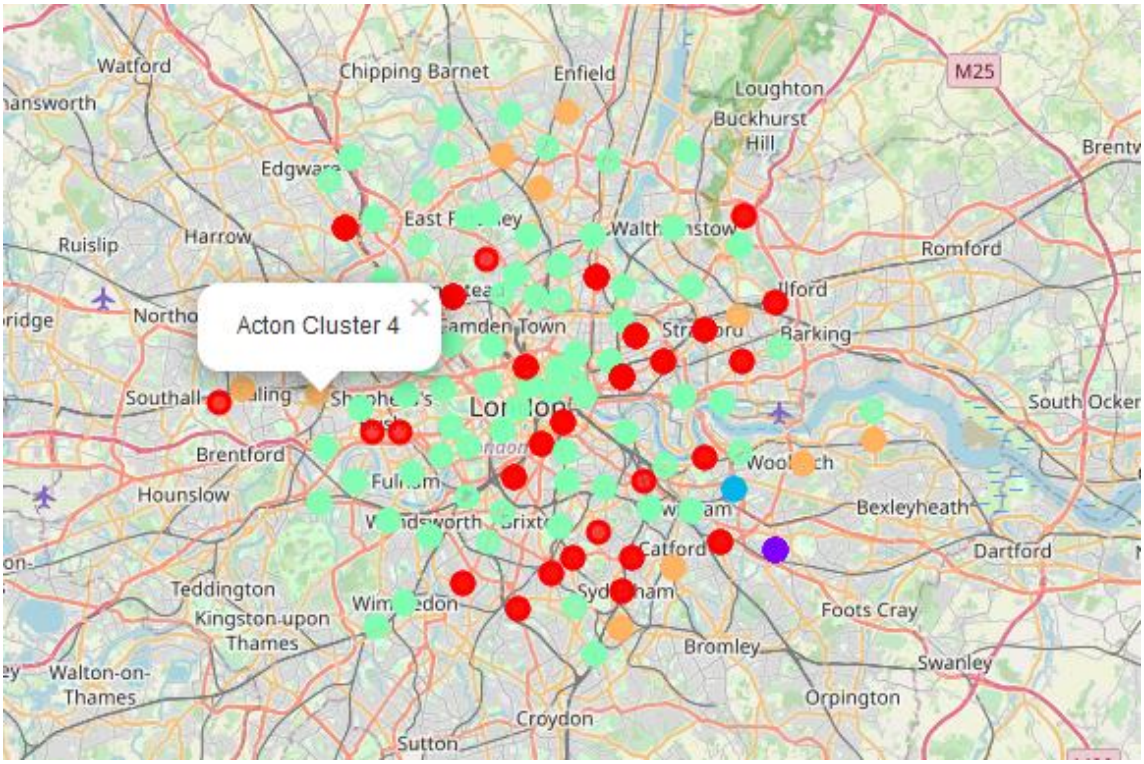
In this project, five (5) clusters are chosen:

- Cluster 0 (Red): 1st Cluster
- Cluster 1 (Violet) : 2nd Cluster
- Cluster 2 (Blue) : 3rd Cluster
- Cluster 3 (Green) : 4th Cluster
- Cluster 4 (Orange) : 5th Cluster

The maps below, constructed using Geopy and Folium libraries, show clustering of the neighborhoods into five clusters.

London Clusters Neighborhoods

For example, we can see *Acton* neighborhood is in the 5th cluster (Cluster 4 on the map below):

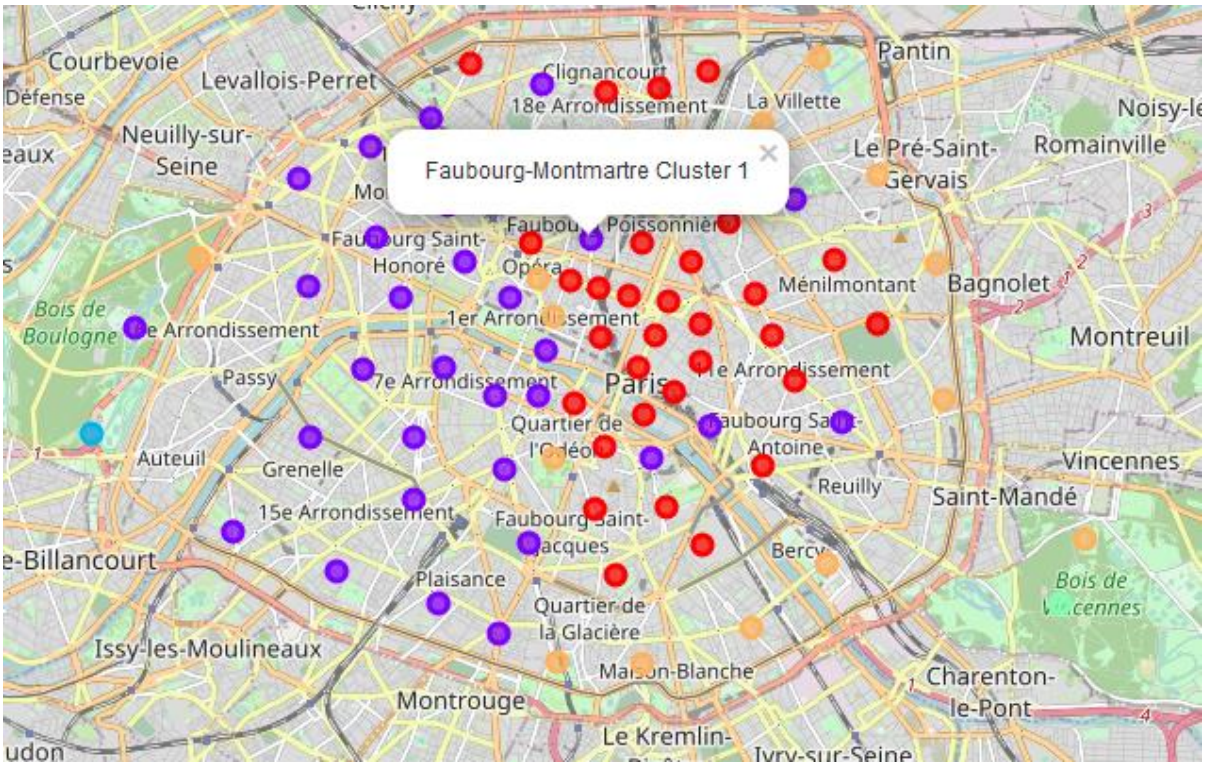


When examining the clusters, we can see the 10 most common venue categories in *Acton* neighborhood:

	Borough	Neighborhood	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 Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bexley, Greenwich	Abbey Wood	51.49245	0.12127	4	Supermarket	Coffee Shop	Train Station	Convenience Store	Platform	Historic Site	Zoo Exhibit	Exhibit	Falafel Restaurant	Farmers Market
1	Ealing, Hammersmith and Fulham	Acton	51.51324	-0.26746	4	Grocery Store	Train Station	Park	Indian Restaurant	Breakfast Spot	Deli / Bodega	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Filipino Restaurant
2	City	Aldgate	51.51200	-0.08058	3	Hotel	Gym / Fitness Center	Restaurant	Salad Place	Coffee Shop	Garden	Italian Restaurant	Pub	Cocktail Bar	Asian Restaurant
3	Westminster	Aldwych	51.51651	-0.11968	3	Pub	Sandwich Place	Theater	Hotel	Japanese Restaurant	Café	Bakery	Clothing Store	Coffee Shop	Chinese Restaurant
4	Bromley	Anerley	51.41009	-0.05683	4	Supermarket	Grocery Store	Convenience Store	Hotel	Fast Food Restaurant	Fish & Chips Shop	Exhibit	Falafel Restaurant	Farmers Market	Filipino Restaurant

Paris Clusters Neighborhoods

For Paris, we can see Faubourg-Montmartre is in the 2nd cluster, noted as Cluster 1 on the map:



When examining the clusters details, the 10 most common venue categories are mostly restaurant venues (French, Italian, Chinese, etc.):

	Borough	Neighborhood	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 Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	PARIS-9E-ARRONDISSEMENT	Saint-Georges	48.879934	2.332850	1	Hotel	French Restaurant	Italian Restaurant	Cocktail Bar	Lounge	Comedy Club	Vietnamese Restaurant	Bistro	Café	Bakery
1	PARIS-9E-ARRONDISSEMENT	Chaussée-d'Antin	48.873547	2.332269	0	Hotel	French Restaurant	Salad Place	Italian Restaurant	Bistro	Department Store	Coffee Shop	Clothing Store	Vietnamese Restaurant	Café
2	PARIS-9E-ARRONDISSEMENT	Rochechouart	48.879812	2.344861	1	French Restaurant	Hotel	Bakery	Vegetarian / Vegan Restaurant	Pizza Place	Park	Record Shop	Restaurant	Coffee Shop	Wine Bar
3	PARIS-9E-ARRONDISSEMENT	Faubourg-Montmartre	48.873935	2.343253	1	French Restaurant	Hotel	Italian Restaurant	Vegetarian / Vegan Restaurant	Coffee Shop	Chinese Restaurant	Cocktail Bar	Japanese Restaurant	Gym / Fitness Center	Bar
4	PARIS-2E-ARRONDISSEMENT	Gaillon	48.869307	2.333432	4	Japanese Restaurant	Hotel	Korean Restaurant	French Restaurant	Chocolate Shop	Bakery	Wine Bar	Plaza	Coffee Shop	Pastry Shop

Combining London and Paris Data

The next step is to combine the two data frames before applying the clustering algorithm to find the similarities between both cities.

The following rules are applied before the combination process:

- To distinguish London neighborhoods from Paris in the new data frame, the name of the respective cities is added to the end of each neighborhood: "_London" and "_Paris".
- When London and Paris do not have the same venue categories, the columns of both data frames are made the same by adding the columns that exist only in London data frame to Paris data frame and vice versa. The newly added columns have a value of 0 for all the rows.

Once combined, we can have the most common categories for each neighborhood in London and Paris by retrieving the five categories with the largest values for each neighborhood.

Clustering London and Paris Data

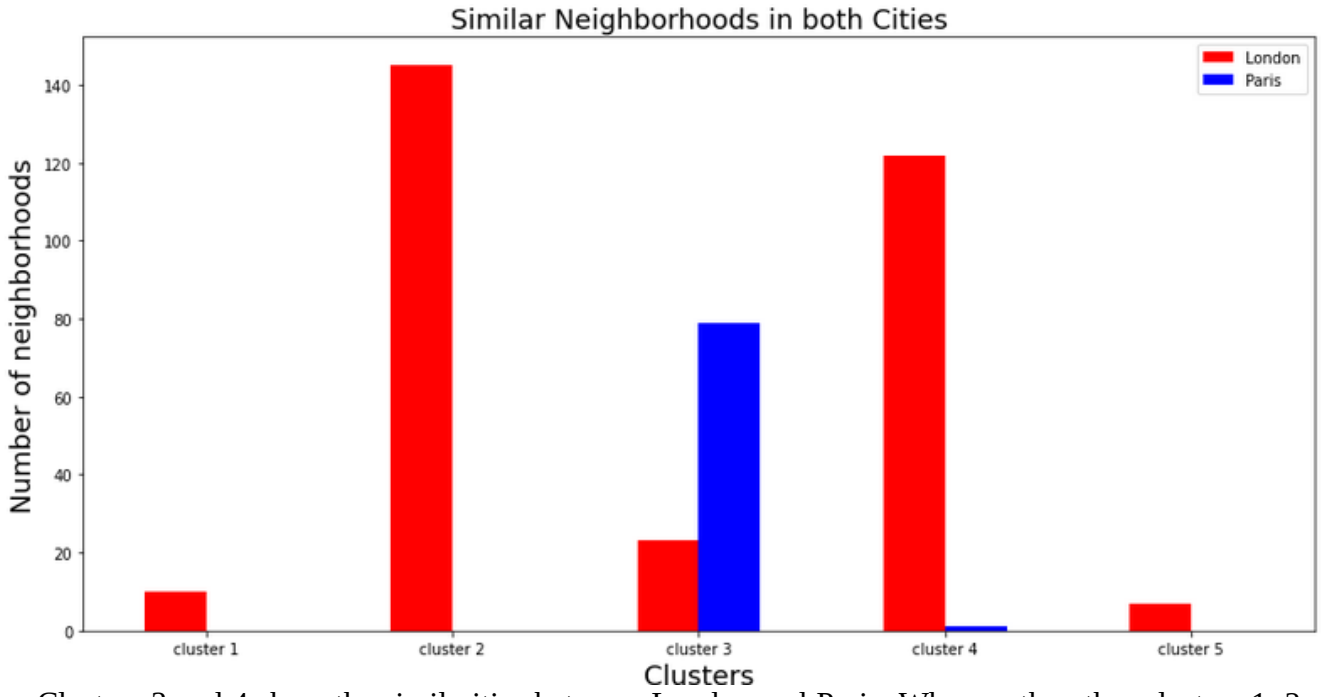
Then we apply the clustering algorithm this time on the combined data. We create a data frame to show the neighborhoods of London and Paris, the cluster to which each neighborhood belongs, and the most common venue categories in each neighborhood.

Neighborhood	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
Winchmore Hill_London	3	Italian Restaurant	Grocery Store	Café	Supermarket	Soccer Field	Ethiopian Restaurant	Event Space	Exhibit	Falafel Restaurant	Farmers Market
Wood Green_London	3	Italian Restaurant	Indian Restaurant	Park	Grocery Store	Bar	Dance Studio	Event Space	Exhibit	Falafel Restaurant	Farmers Market
Woodford_London	1	Bar	BBQ Joint	Grocery Store	Pub	Seafood Restaurant	Film Studio	Ethiopian Restaurant	Event Space	Exhibit	Falafel Restaurant
Woodford Green_London	2	Hotel	Theater	Plaza	Garden	Monument / Landmark	Cocktail Bar	Pub	Restaurant	Tea Room	Boutique
Woodside Park_London	3	Coffee Shop	Bakery	Fast Food Restaurant	Supermarket	Pharmacy	Chinese Restaurant	Thai Restaurant	Theater	Sushi Restaurant	Turkish Restaurant
Woolwich_London	3	Child Care Service	Chinese Restaurant	Convenience Store	Indian Restaurant	Bus Stop	Middle Eastern Restaurant	Grocery Store	Fish & Chips Shop	Zoo Exhibit	Film Studio
Wormwood Scrubs_London	3	Grocery Store	Café	Pub	Gastropub	Pizza Place	Thai Restaurant	Park	Event Space	Gourmet Shop	Greek Restaurant
Amérique_Paris	2	French Restaurant	Supermarket	Pool	Bistro	Park	Bed & Breakfast	Plaza	Café	Zoo Exhibit	Falafel Restaurant
Archives_Paris	2	French Restaurant	Hotel	Coffee Shop	Clothing Store	Bar	Art Gallery	Bookstore	Bistro	Plaza	Cocktail Bar
Arsenal_Paris	2	French Restaurant	Hotel	Park	Gastropub	Plaza	Italian Restaurant	Pedestrian Plaza	Cocktail Bar	Vegetarian / Vegan Restaurant	Seafood Restaurant

RESULTS AND DISCUSSION

Clustering Results

The figure below shows the number of neighborhoods of London and Paris in each of the five resulting clusters.



Clusters 3 and 4 show the similarities between London and Paris. Whereas the other clusters 1, 2 and 5 show the dissimilarities.

Cluster Analysis

The clustering algorithm grouped neighborhoods of London and Paris in 5 clusters based on the similarity between their venues. Now, these clusters will be investigated to see the most common categories in each of them.

The figure below shows the most common venue categories in each cluster. Each common category is displayed as the percentage of venues within the cluster.

Cluster 1

Venue Category	% of venues
Supermarket	27.8689
Convenience Store	16.3934
Train Station	6.55738
Platform	6.55738
Historic Site	6.55738
Coffee Shop	6.55738
Fast Food Restaurant	4.91803

Cluster 2

Venue Category	% of venues
Pub	10.7637
Coffee Shop	5.98383
Café	4.79784
Italian Restaurant	3.07278
Hotel	2.38994
Bakery	2.354
Sandwich Place	2.04852

Cluster 3

Venue Category	% of venues
French Restaurant	10.279
Hotel	8.2574
Italian Restaurant	3.75854
Coffee Shop	3.10364
Café	2.96128
Bakery	2.70501
Bar	2.22096

Cluster 4

Venue Category	% of venues
Café	8.68813
Coffee Shop	8.58429
Grocery Store	7.19972
Pub	5.08827
Pizza Place	3.32295
Park	2.83835
Italian Restaurant	2.49221

Cluster 5

Venue Category	% of venues
Historic Site	20
Bus Stop	20
Golf Course	20
Construction & Landscaping	20
Park	20

Figure 15 Most common venue-categories in each of the 5 clusters

Cluster 3

Drilling down on the cluster 3 which shows interesting data to compare both cities in term of venues:

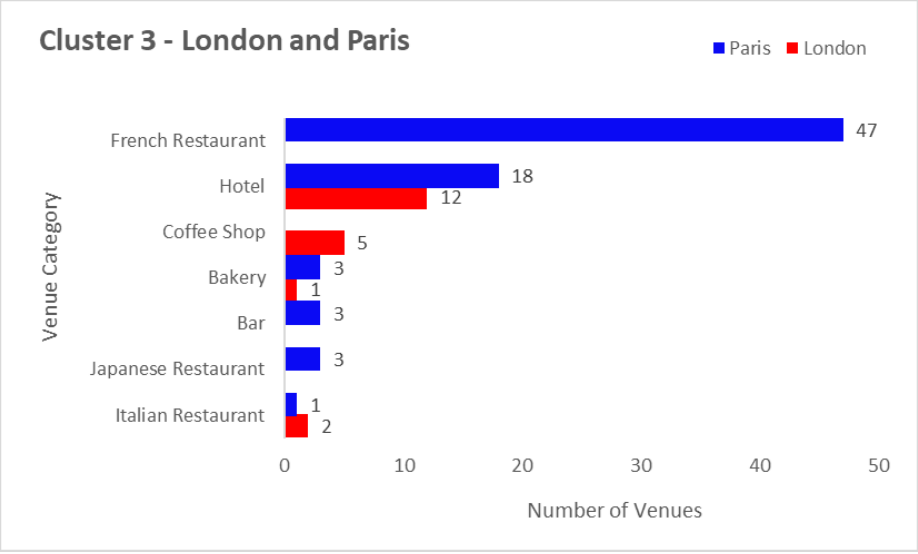


Figure 16 Cluster 3 Venue category detail

Based on the results of the segmentation of the neighborhoods of two cities, the places reported by Foursquare appear to be very oriented towards food and especially restaurants. One option in the future could be to have more data on other categories of venues in order to make a more in-depth comparison in more relevant categories such as tourist sites, transportation, lodging, museums, etc.

CONCLUSION

In this project, the neighborhoods of London and Paris were clustered into multiple groups based on the categories (types) of the venues in these neighborhoods. The results show that there are venue categories that are more common in some cluster than others; the most common venue categories differ from one cluster to the other.

If a deeper analysis, taking more aspects into account, is performed, it might result in discovering different styles in each cluster based on the most common categories in the cluster.