Applied Data Science Capstone Project
The Battle of Neighborhoods

Project report

Author: Irina Chernova

1. Introduction

Problem Background:

London is considered to be one of the world's most important global cities. It is not only one of the largest financial centers but also a city recognized as the global capital for arts and culture, which art venues, museums and galleries are renowned across the world. London exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance etc. It is the most-visited city as well as one of the world's most populous cities, which means that a lot of people every year spend a lot of money for various goods and services.

As it is highly developed city so cost of doing business here is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully. The insights derived from analysis will give good understanding of the business environment which helps provide the most efficient strategy in development of business as well as minimize possible risks and expenses.

Problem Description:

Let’s assume that newly created but very ambitious art dealer company “NNN” is going to open office in London. The field of activity of the company is buying and selling pieces of arts – this is very expensive and highly competitive market. They want to know which place is the best for their new office. So I as a Data Scientist should investigate all boroughs of London in order to make a list of recommendations for them.

An art dealer is a person or company that buys and sells works of art, so I have to consider location of art centers and galleries, artistic workspaces etc. because such facilities provide arts space, visual art gallery space, museum facilities where new artworks exhibit and promote. So the “NNN” company office should be placed closely to such facilities.

Also very important to explore neighborhoods in order to determine how many venues are around including their category, for example: cafes, restaurants, gyms, shopping centers - their description and ratings.

Target audience:

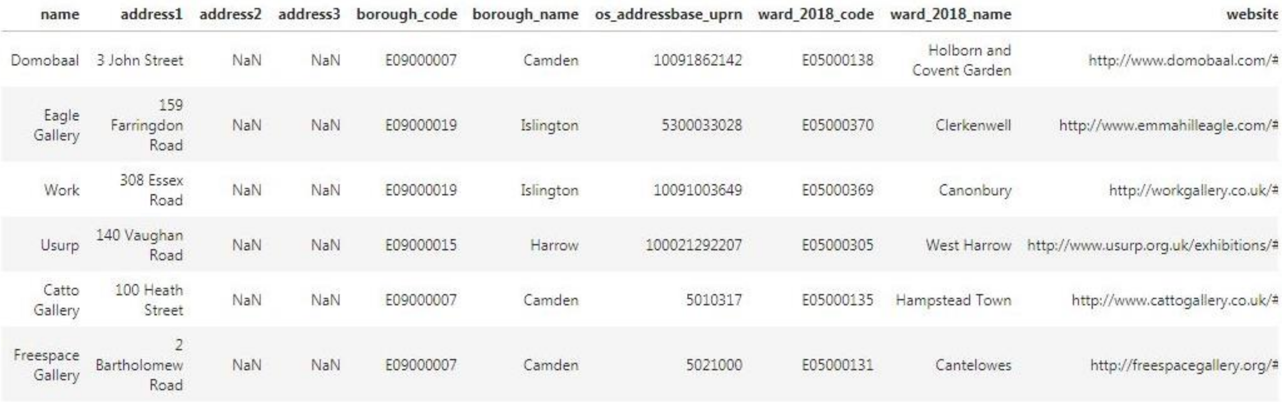
The work is performed for hypothetical Art Dealer Company “NNN”. But not only art dealers could be interested in this exploration, also potential clients or collectors looking for the new opportunities to buy or sell pieces of art could find it useful. As we see our stakeholders are art dealers, artists and customers, who want to buy or sell any work of art.

2. Data

According to business problem described in previous section, I need to study boroughs and neighborhoods of London which are most suitable for the office of ”NNN” company. Thus I’ll be needed a lot of location data concerning art and cultural facilities in London as well as data about social infrastructure and venues around.

Among the variety of on-line data sources I chose the following:

- Dataset on Art Centers and Creative workspaces published by Government of UK under the section Cultural Infrastructure. It is detailed information helpful for understanding location and distribution of the art and cultural facilities in London for further clustering
  Link : https://data.london.gov.uk/dataset/cultural-infrastructure-map
  Format : CSV file



name	address1	address2	address3	borough_code	borough_name	os_addressbase_uprn	ward_2018_code	ward_2018_name	website
Domobaal	3 John Street	NaN	NaN	E09000007	Camden	10091862142	E05000138	Holborn and Covent Garden	http://www.domobaal.com/#
Eagle Gallery	159 Farringdon Road	NaN	NaN	E09000019	Islington	5300033028	E05000370	Clerkenwell	http://www.emmahilleagle.com/#
Work	308 Essex Road	NaN	NaN	E09000019	Islington	10091003649	E05000369	Canonbury	http://workgallery.co.uk/#
Usurp	140 Vaughan Road	NaN	NaN	E09000015	Harrow	100021292207	E05000305	West Harrow	http://www.usurp.org.uk/exhibitions/#
Catto Gallery	100 Heath Street	NaN	NaN	E09000007	Camden	5010317	E05000135	Hampstead Town	http://www.cattogallery.co.uk/#
Freespace Gallery	2 Bartholomew Road	NaN	NaN	E09000007	Camden	5021000	E05000131	Cantelowes	http://freespacegallery.org/#

- Post indexes and boroughs of London as well as the latitude and longitude of each borough of London provided by doogal.co.uk . Common geographical data necessary for accurate localization of boroughs in order to visualize our data for further exploration
  Link : https://www.doogal.co.uk/PostcodeDownloads.php
  Format : CSV file

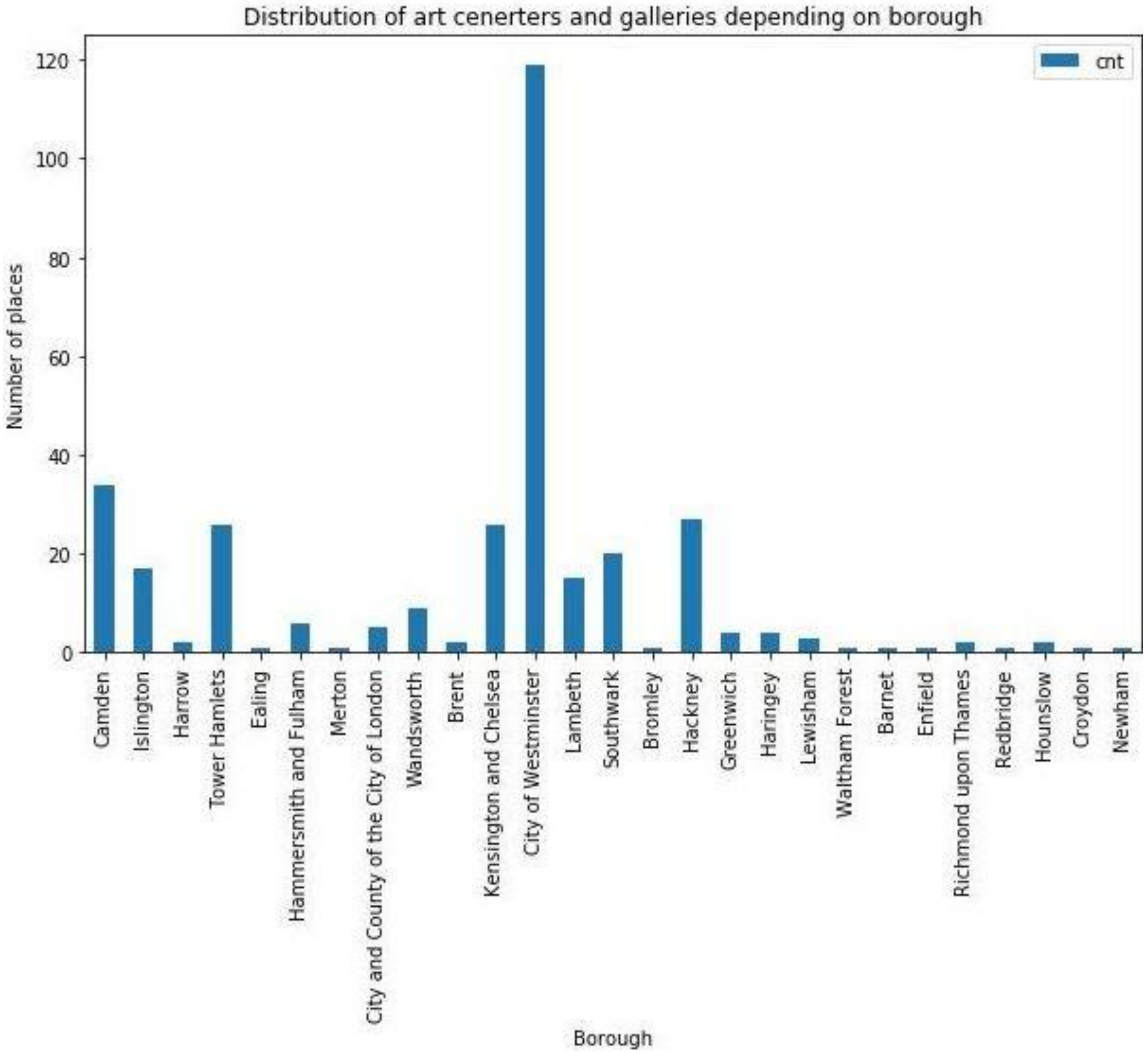
- Foursquare location data and venues information for each borough. This is comprehensive amount of data about various kinds of venues, its location and popularity. I‘m going to use the Foursquare API to explore distribution of different venues for each borough of London
  Link : https://foursquare.com/
  Format : JSON

All *.csv files were uploaded and placed into the project working directory for ease of use

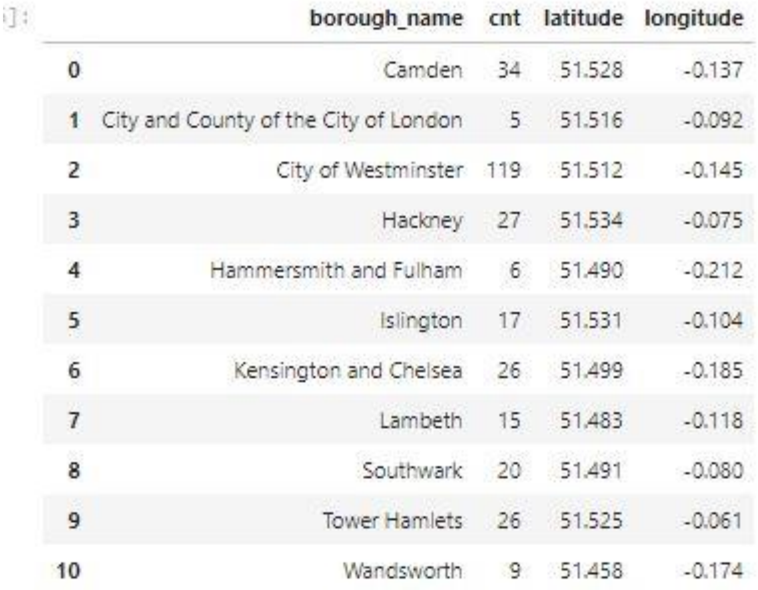
3. Methodology

Data preparation&preprocessing

All necessary data downloaded or scraped from websites has been loaded into the one pandas dataframe, so we would be able to perform some preliminary statistical analysis. For example we could examine the number of art centers&galleries in each borough :



As we can see distribution of places is quite irregular, and we hardly be interested in boroughs where number of art centers less than 4, so we can exclude them from the resulting dataset. After all necessary manipulations our final dataset ready for further exploration:

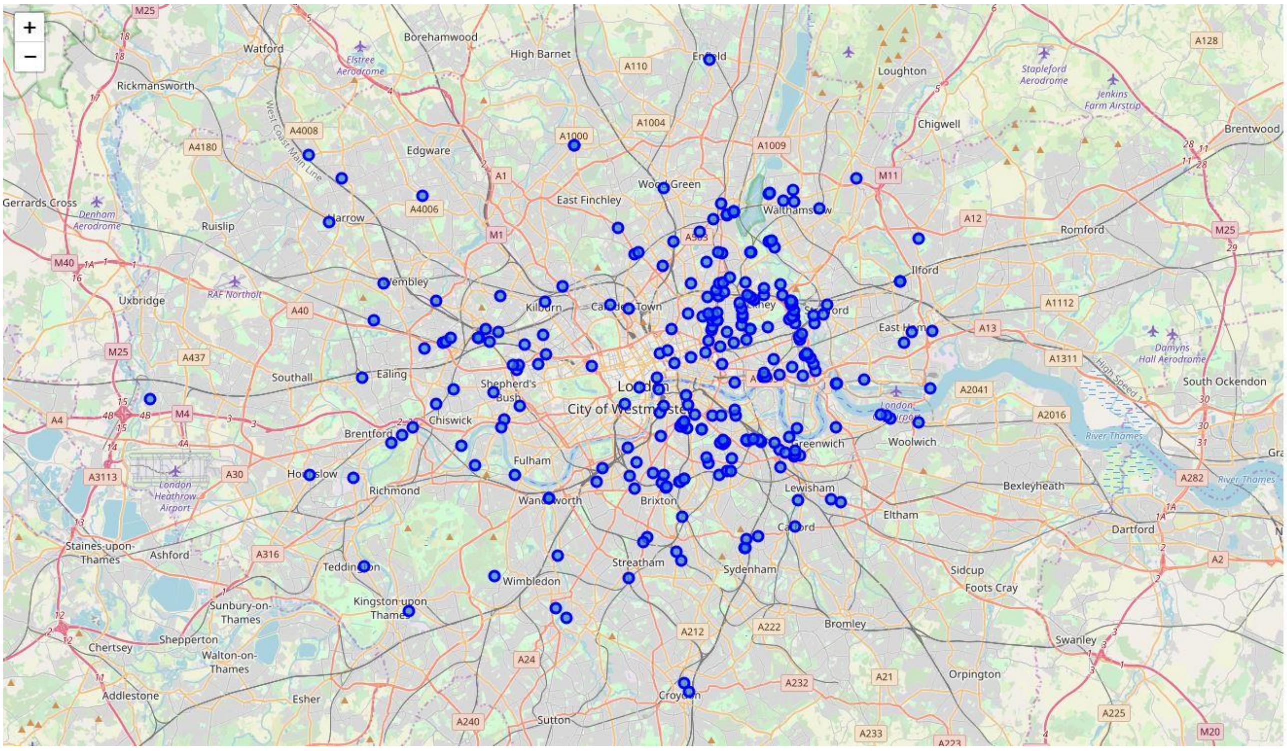


	borough_name	cnt	latitude	longitude
0	Camden	34	51.528	-0.137
1	City and County of the City of London	5	51.516	-0.092
2	City of Westminster	119	51.512	-0.145
3	Hackney	27	51.534	-0.075
4	Hammersmith and Fulham	6	51.490	-0.212
5	Islington	17	51.531	-0.104
6	Kensington and Chelsea	26	51.499	-0.185
7	Lambeth	15	51.483	-0.118
8	Southwark	20	51.491	-0.080
9	Tower Hamlets	26	51.525	-0.061
10	Wandsworth	9	51.458	-0.174

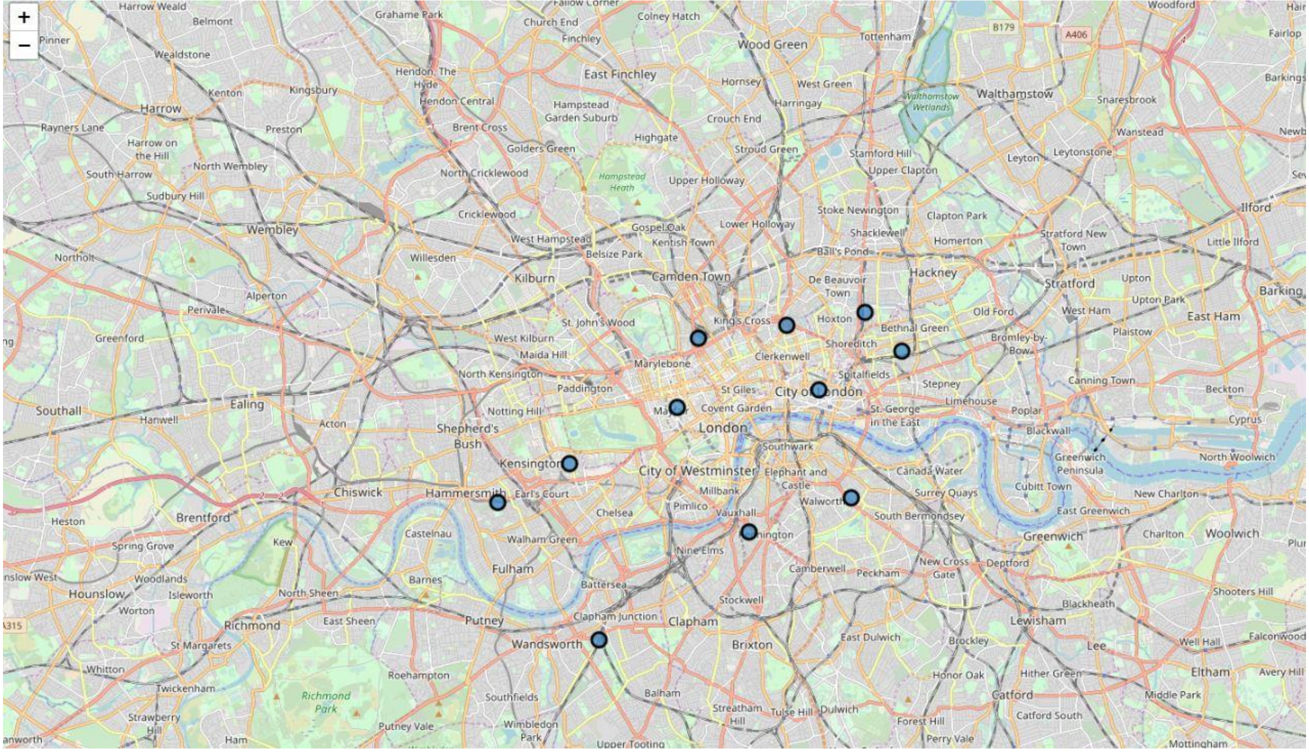
Data visualization

But before we proceed with clustering, let’s take a look at data distribution on the map of London using the

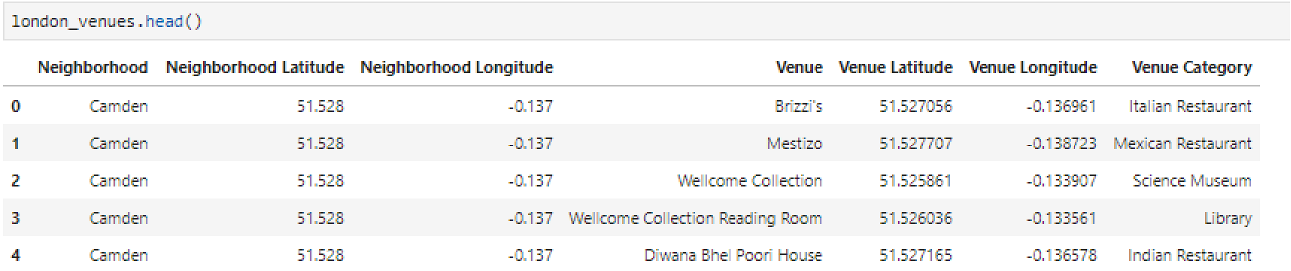
Folium library. Folium is a powerful Python library useful in creation of different type of maps. So, visualization of the initial dataset looks like this :



And location of points of our resulting dataset is shown below:



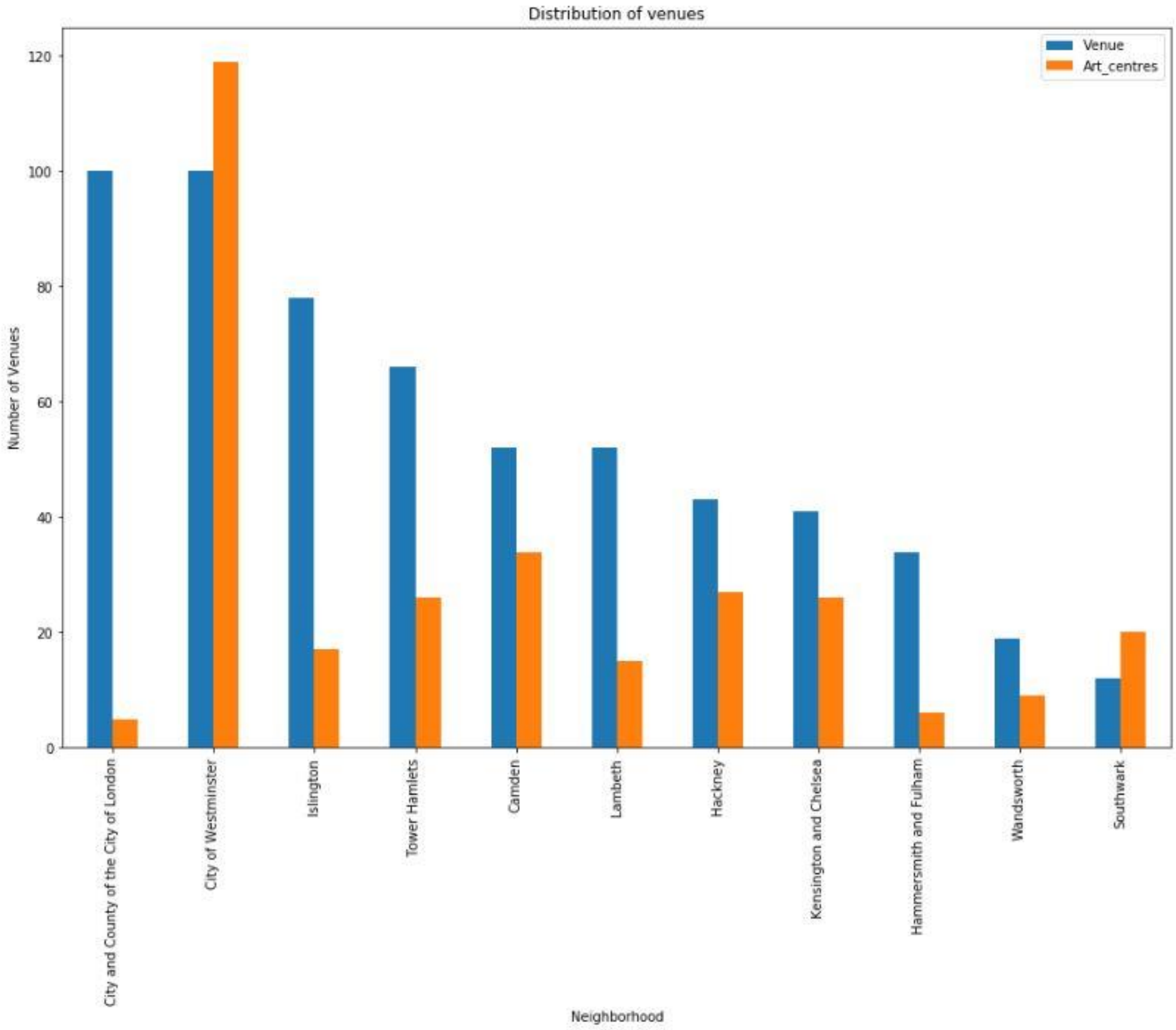
The next step of our analysis is exploration of neighborhoods in selected boroughs. We'll be use the Foursquare API for this purpose. Foursquare, is a local search-and-discovery mobile application which provides search results for its users as well as stores users' check-ins, tips, type and location of venues etc. The Foursquare API allows application developers to interact with the Foursquare database in order to retrieve this information. We use the Foursquare API to acquire number of venues, their type and location for the given boroughs of London. All gathered data is placed into the new data frame as shown below :



```
london_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Camden	51.528	-0.137	Brizzi's	51.527056	-0.136961	Italian Restaurant
1	Camden	51.528	-0.137	Mestizo	51.527707	-0.138723	Mexican Restaurant
2	Camden	51.528	-0.137	Wellcome Collection	51.525861	-0.133907	Science Museum
3	Camden	51.528	-0.137	Wellcome Collection Reading Room	51.526036	-0.133561	Library
4	Camden	51.528	-0.137	Diwana Bhel Poori House	51.527165	-0.136578	Indian Restaurant

Now we would be able to perform some statistic calculation and compare the data. For instance, let’s compare the number of art centers and galleries and total number of different venues depending on borough:

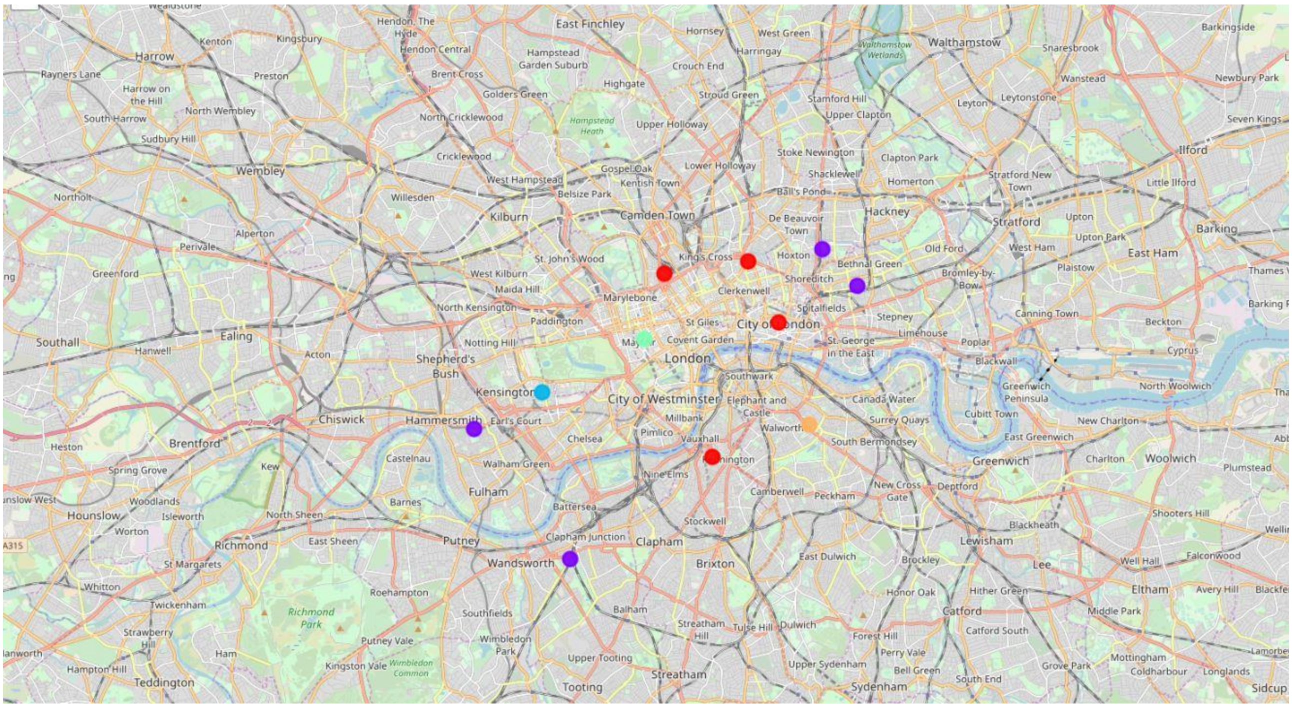


And we going to use the k-means method for clustering neighborhoods using data acquired from the Foursquare API. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. It group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

A cluster refers to a collection of data points aggregated together because of certain similarities. We define a target number k = 5, which refers to the number of centroids we need in the dataset. A centroid is the imaginary location representing the center of the cluster.

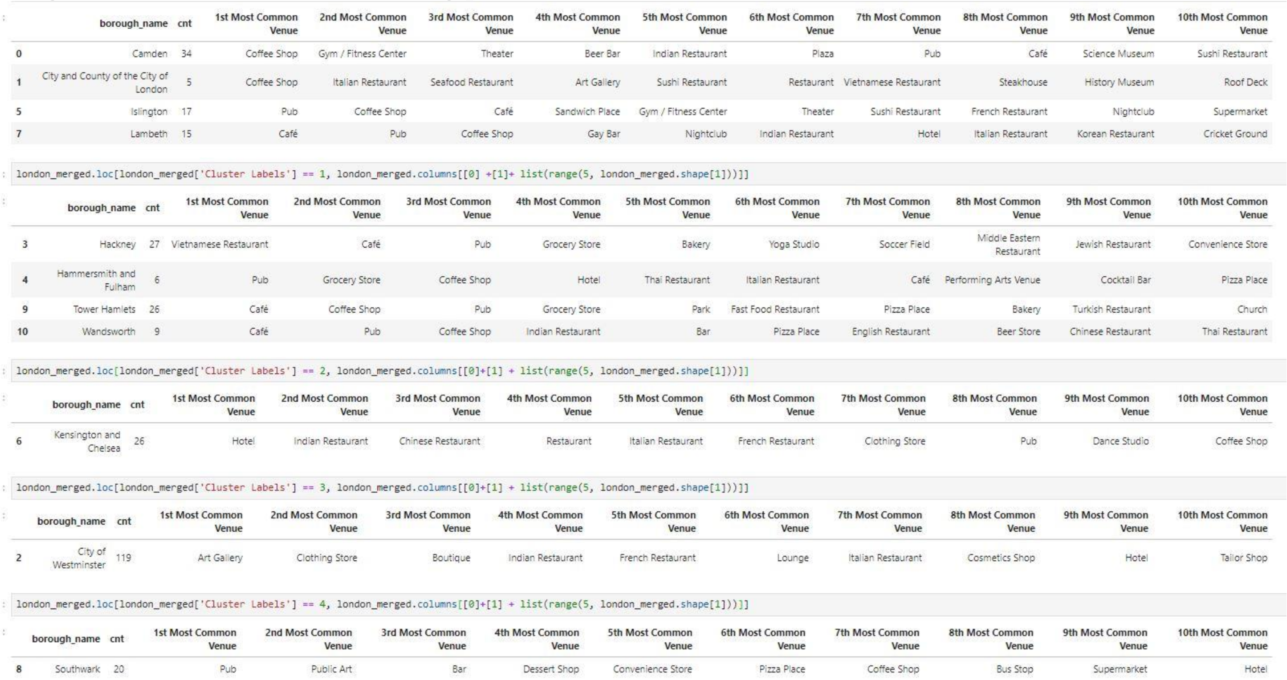
Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

And finally, let’s take a look at clusters obtained by the k-means method :



4. Results

Neighborhoods K-Means clustering provide us the next result:



	borough_name	cnt	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	Camden	34	Coffee Shop	Gym / Fitness Center	Theater	Beer Bar	Indian Restaurant	Plaza	Pub	Café	Science Museum	Sushi Restaurant
1	City and County of the City of London	5	Coffee Shop	Italian Restaurant	Seafood Restaurant	Art Gallery	Sushi Restaurant	Restaurant	Vietnamese Restaurant	Steakhouse	History Museum	Roof Deck
5	Islington	17	Pub	Coffee Shop	Café	Sandwich Place	Gym / Fitness Center	Theater	Sushi Restaurant	French Restaurant	Nightclub	Supermarket
7	Lambeth	15	Café	Pub	Coffee Shop	Gay Bar	Nightclub	Indian Restaurant	Hotel	Italian Restaurant	Korean Restaurant	Cricket Ground

```
london_merged.loc[london_merged['Cluster Labels'] == 1, london_merged.columns[[0] +[1]+ list(range(5, london_merged.shape[1]))]]
```

	borough_name	cnt	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	Hackney	27	Vietnamese Restaurant	Café	Pub	Grocery Store	Bakery	Yoga Studio	Soccer Field	Middle Eastern Restaurant	Jewish Restaurant	Convenience Store
4	Hammersmith and Fulham	6	Pub	Grocery Store	Coffee Shop	Hotel	Thai Restaurant	Italian Restaurant	Café	Performing Arts Venue	Cocktail Bar	Pizza Place
9	Tower Hamlets	26	Café	Coffee Shop	Pub	Grocery Store	Park	Fast Food Restaurant	Pizza Place	Bakery	Turkish Restaurant	Church
10	Wandsworth	9	Café	Pub	Coffee Shop	Indian Restaurant	Bar	Pizza Place	English Restaurant	Beer Store	Chinese Restaurant	Thai Restaurant

```
london_merged.loc[london_merged['Cluster Labels'] == 2, london_merged.columns[[0]+[1] + list(range(5, london_merged.shape[1]))]]
```

	borough_name	cnt	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	Kensington and Chelsea	26	Hotel	Indian Restaurant	Chinese Restaurant	Restaurant	Italian Restaurant	French Restaurant	Clothing Store	Pub	Dance Studio	Coffee Shop

```
london_merged.loc[london_merged['Cluster Labels'] == 3, london_merged.columns[[0]+[1] + list(range(5, london_merged.shape[1]))]]
```

	borough_name	cnt	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
2	City of Westminster	119	Art Gallery	Clothing Store	Boutique	Indian Restaurant	French Restaurant	Lounge	Italian Restaurant	Cosmetics Shop	Hotel	Tailor Shop

```
london_merged.loc[london_merged['Cluster Labels'] == 4, london_merged.columns[[0]+[1] + list(range(5, london_merged.shape[1]))]]
```

	borough_name	cnt	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
8	Southwark	20	Pub	Public Art	Bar	Dessert Shop	Convenience Store	Pizza Place	Coffee Shop	Bus Stop	Supermarket	Hotel

Now we‘re ready to perform analysis of acquired resulting clusters and to make conclusions:

Cluster #3: City of Westminster satisfies our conditions in the best way. But due to the large number of art centers&galleries it’s highly competitive, and doing business here could be difficult for the young company. Also, apparently to rent the office here would be extremely expensive. Thus I wouldn’t recommend this borough for the young unknown company even though they could rent the office there.

Cluster #0 Is more suitable for our “NNN” company. All boroughs of the cluster have highly developed social infrastructure and a lot of venues for the benefit of office employees. As well it provides a lot of possibilities in getting new customers and to be in touch with the artistic life. I would strongly recommend these boroughs as the most suitable place for the office.

Also clusters #2 and #4 appear to be promising and could be appropriate to rent an office.

5. Discussion

Considering the conclusions of the results section, the management of “NNN” company can take a conscious decision about choosing a location of their office based upon their requirements. Result list includes at least 4 boroughs where they can establish their office.

Even though amount of data was limited and some aspects were not taken into account, we got a relevant list of recommendations. For further improvement and refinement of the project model we should use more data and include in our analysis such indicators as number and location of office buildings, rental cost depending on borough, public transport etc. This approach will help us to define more accurate clusters and perform more detailed analysis.

6. Conclusion

Although this study is based on limited data and it doesn’t take into account some important factors, nevertheless it shows the power of the data science analytic approach and tools. It’s amazing how easy and quickly you can get the comprehensive analysis of economic and social infrastructure using only open data. This approach is quite flexible. There are a lot of ways for development and refinement of models and methods. Than more accurate and detailed data you have, than more relevant and precise result you would obtain. The results of such analysis are highly important and helpful in making sensible decision concerning business issues.