The-Battle-of-Neighborhoods-Week-1

Blog: https://wp.me/P6LcRx-3O

Introduction:

The objective was to study the different neighbourhood of Paris to compare the similarities of the popular venues. This is crucial since Paris is a very touristic city and statistical inferences can be extremely biased towards touristic activities and places.

The results would also depend on the type of users using the FourSquare application.

I chose the city of Paris, where I live, so that I could use my first-hand experience to tally the results.

Problem Background:

Paris being a very touristic city, people can get overwhelmed on what to do and with their limited time in the city, it is important to be able to find the most popular venues in each of the districts and decide accordingly. They also they need to know which districts have similar venues and also the popularity ratings per district.

Problem Description:

To discover the different districts of Paris and find districts which are similar and the most popular venues in these districts. **The target audience are the tourists visiting the city of Paris.**

About Paris:

Paris is divided into 20 districts and is situated by the river Seine. Half of the city is on the right bank while the other half is more on the left bank of the river.

Arrondissement	Name	River Bank	Size km
1	Louvre	R	1.826 km2
2	Bourse	R	0.992 km2
3	Temple	R	1.171 km2
4	Hôtel-de-Ville	R	1.601 km2
5	Panthéon	L	2.541 km2
6	Luxembourg	L	2.154 km2
7	Palais-Bourbon	L	4.088 km2
8	Élysée	R	3.881 km2
9	Opéra	R	2.179 km2
10	Entrepôt	R	2.892 km2
11	Popincourt	R	3.666 km2
12	Reuilly	R	16.324 km2
13	Gobelins	L	7.146 km2
14	Observatoire	L	5.621 km2
15	Vaugirard	L	8.502 km2
16	Passy	R	16.305 km2
17	Batignolles-Monceau	R	5.669 km2
18	Butte-Montmartre	R	6.005 km2
19	Buttes-Chaumont	R	6.786 km2
20	Ménilmontant	R	5.984 km2

Methodology:

Part 1: Data Collection

The Wikipedia page for Paris Districts was chosen to extract the basic details of the 20 districts (also called arrondissement) from https://en.wikipedia.org/wiki/Arrondissements_of_Paris

The original idea was to get all "Places of Interest" from each of the individual Wiki pages of each of the arrondissement and use it in conjunction to FourSquare APIs but this was abandoned due to lack of time.

On the Wiki page we see that columns for Arrondissement, Area and Population Density needed some additional data wrangling to make it cleaner.

At the same time, I used data from these two files: mainly for the latitude and longitude. Formats de fichiers plats https://opendata.paris.fr/explore/dataset/arrondissements/download/?format=csv&timezone=Europe/Berlin&use_labels_for_header=true

Formats de fichiers géographiques
https://opendata.paris.fr/explore/dataset/arrondissements/download/?format=geojson&timezone=Europe/Berlin

Again here, I faced issues downloading the files directly in my project, so I had to manually download the csv files and use it in the project.

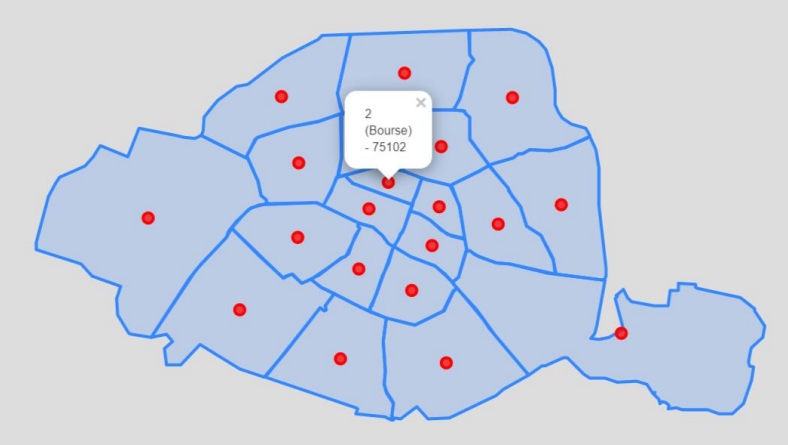
To add here itself, another remark was the dependency of packages that were likely to be used. I faced some issues, but I was able to resolve them. At https://labs.cognitiveclass.ai, I did not find any option to upload the files, so I used Jupyter notebook with Python 3 on my laptop.

So after having scrapped the Wiki page and the csv file containing latitude and longitude, I had the final working dataframe: **paris_et_arrondissements_df**.

Discussion

Part 2:

The next step was to plot the map of Paris with its arrondissements using Folium displaying the district, postal code and the name.



Part 3:

We setup the credentials for Foursquare API

Part 4:

Using FourSquare API calls I used function to retrieve nearby avenues, number of venues per districts, and the unique categories discovered.

The most popular categories of venues were as shown below. French restaurants are, by far, most popular venues.

Venue Category	Count
French Restaurant	176
Hotel	90
Italian Restaurant	50
Café	44
Bistro	42
Bakery	40
Plaza	36
Japanese Restaurant	36
Bar	35

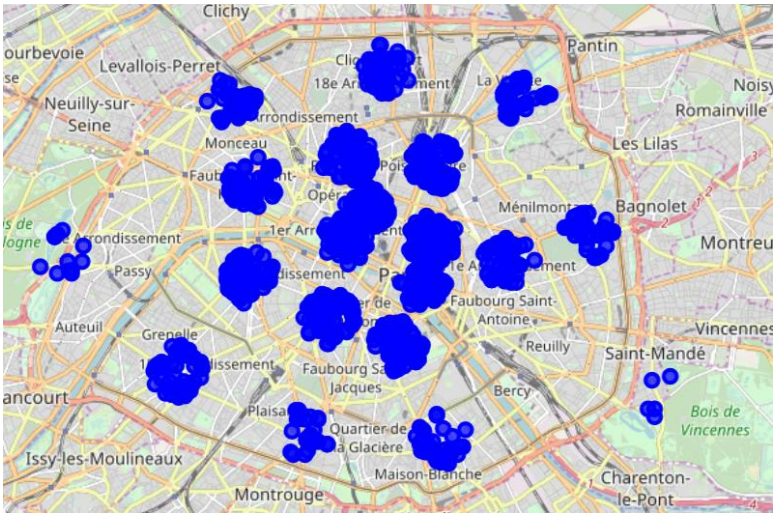
The number of venues returned by the FourSquare API per district is as follows:

Surprisingly, district 15th and 8th returned a smaller number of venues even though they are most frequented districts due to the Eiffel tower and the river cruise being in the 15th and Champs Elysées and the Arc de Triomphe being in the 8th.

District	Count	District	Count
1	100	**11**	64
2	100	**12**	5
3	100	**13**	58
4	100	**14**	28
5	96	**15**	58
6	91	**16**	12
7	100	**17**	43
8	64	**18**	51
9	100	**19**	43
10	100	**20**	51

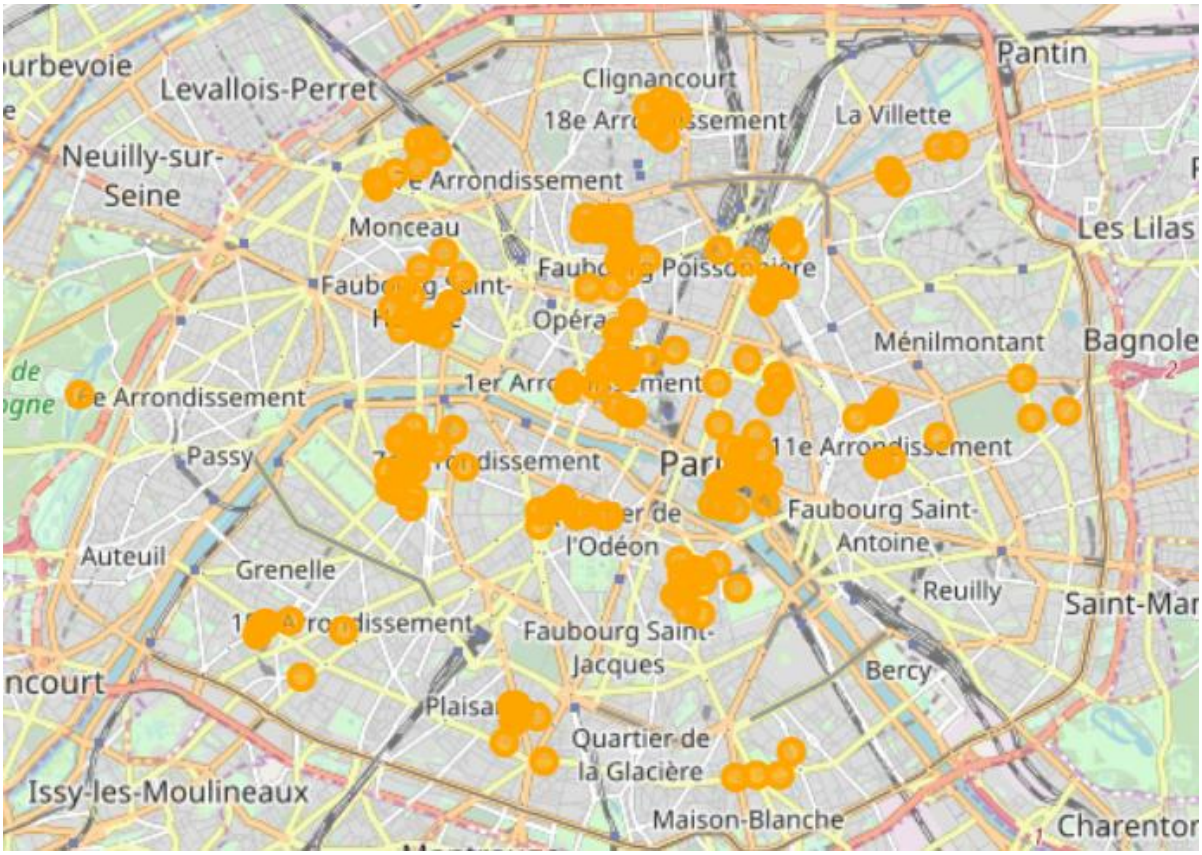
In all 205 categories of venues were returned and 1364 venues were returned. Most number of returned venues were from 1st,2nd ,3rd ,4th ,7th,9th, and 10th districts, each returning 100 venues.

In the map below, we see all the venues returned by FourSqure.



Analysing the top 3 venues: French Restaurant

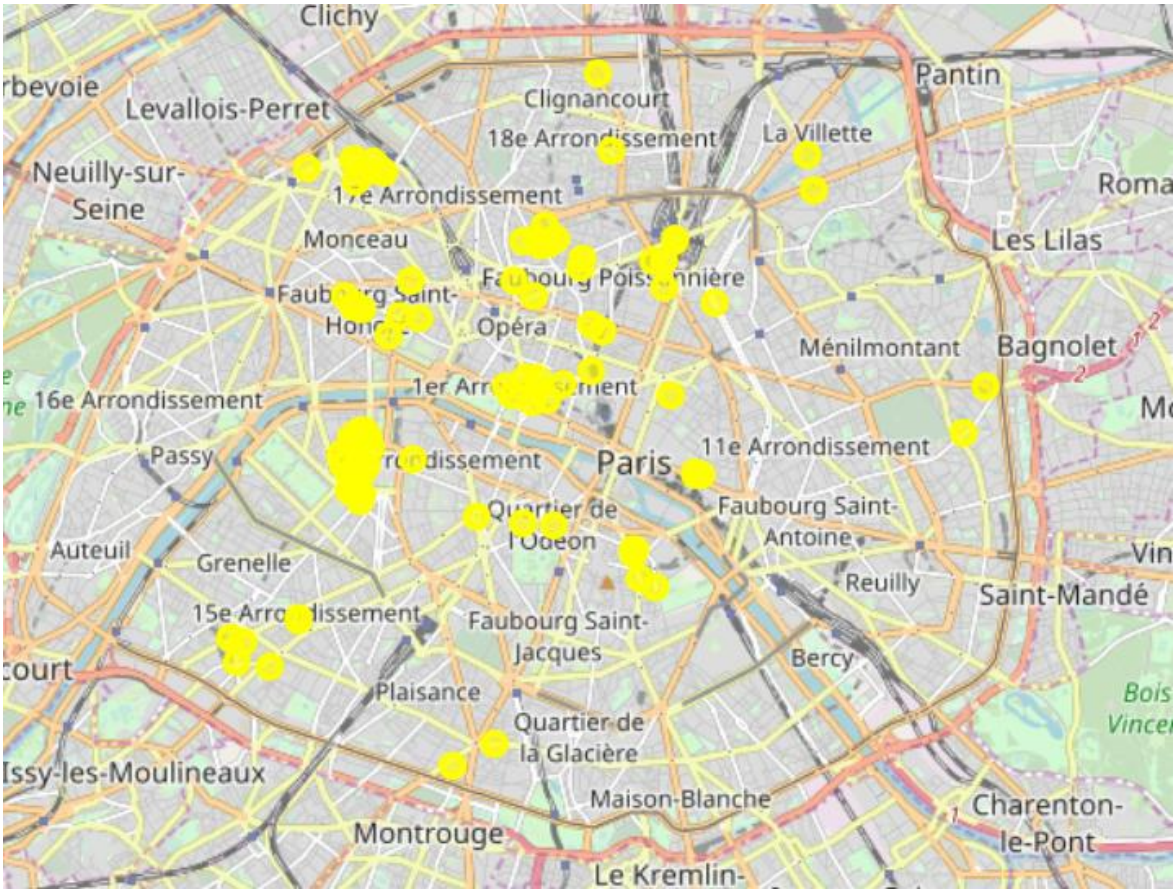
They are concentrated more towards the center and are significantly clustered into smaller pockets.



Analysing the top 3 venues: Hotel

They are concentrated more towards the center and are significantly clustered into smaller pockets.

There are more hotels in the west of Paris compared to the eastern parts of Paris.



Analysing the top 3 venues: Italian Restaurant

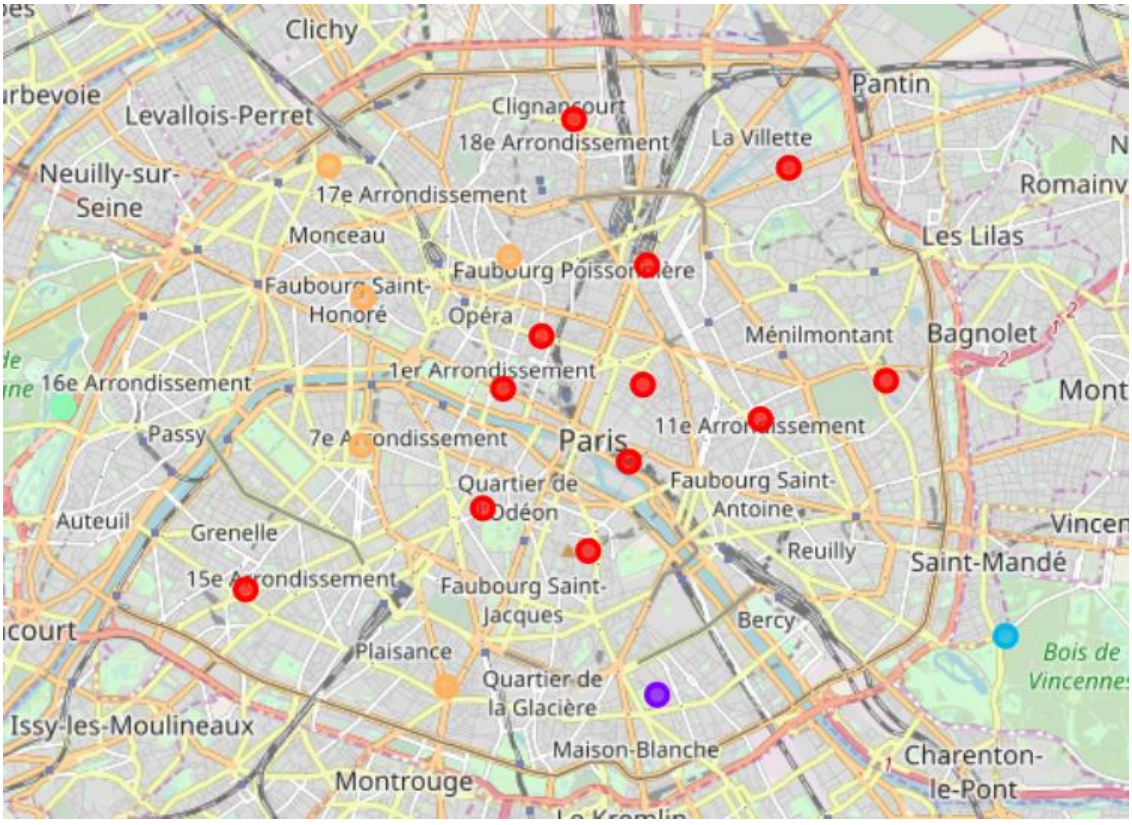
We find more Italian restaurants along the banks of the river Seine compared to elsewhere. The river here in the map is shown by the inverted U curve in blue.

We then use one hot encoding to convert Venue Category into categorical variables for each of the venue's types.

We proceed with grouping the results per each of the districts. We then find the top 10 venues of each neighbourhood

Part 5:

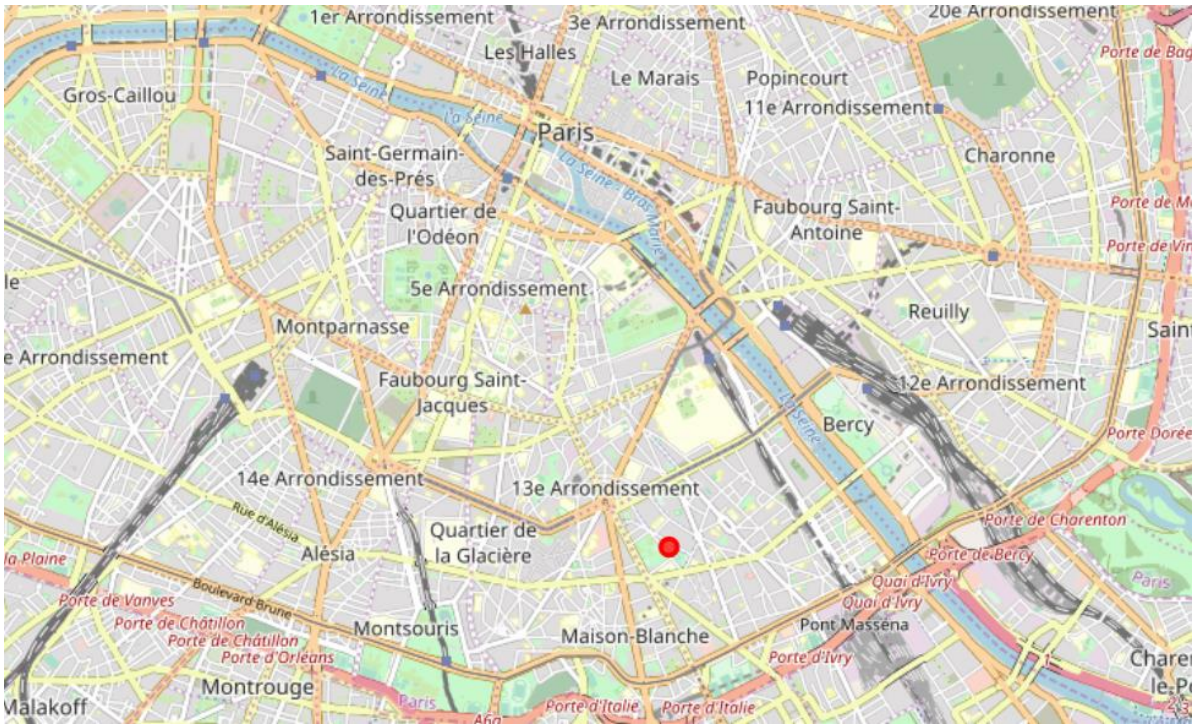
KMeans clustering is applied taking **kclusters = 5** and proceed to visualize these clusters and the most common venues for each of the 4 clusters are determined. The below figure represents the 4 clusters.



Conclusions:

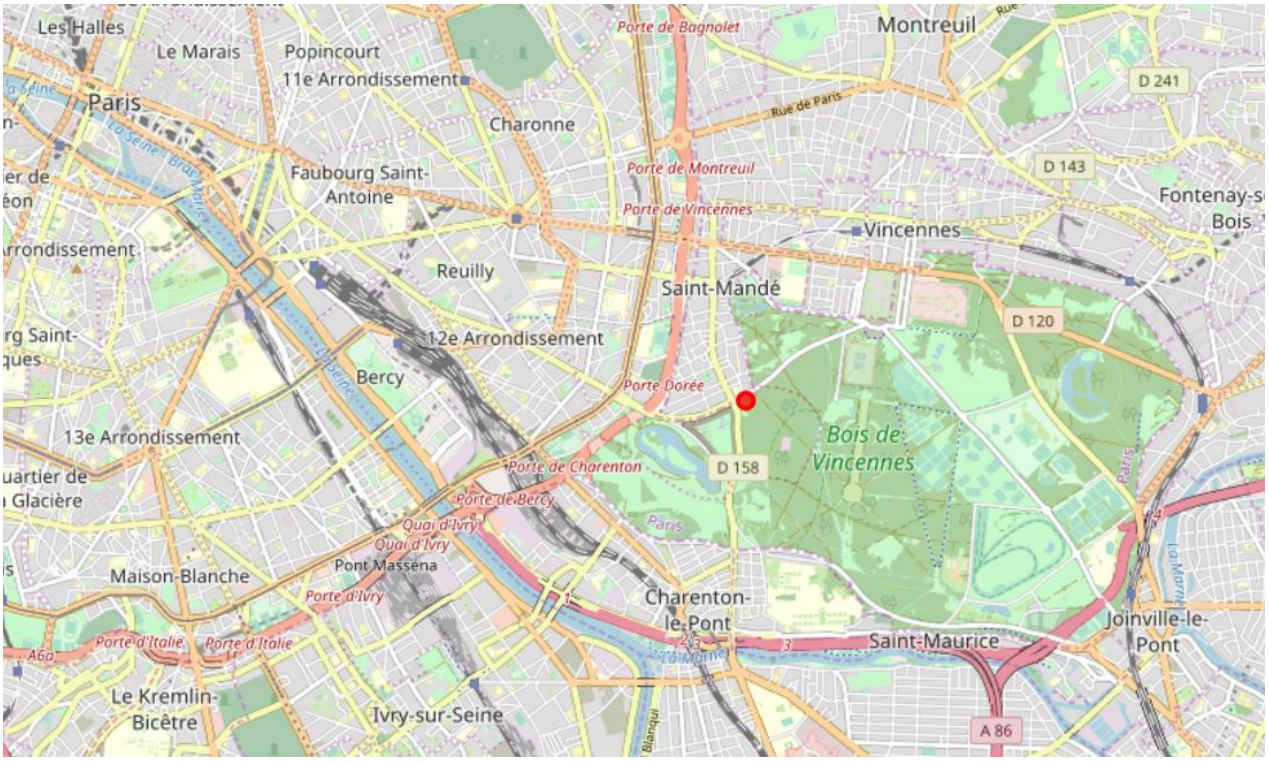
Cluster 1:

It is solely in the 13th district of Paris, also known as ‘Le quartier chinois' and we mainly find Vietnamese / Chinese (or Asian, in general) restaurants there. This district is unique and has rightly been identified as a separate cluster.



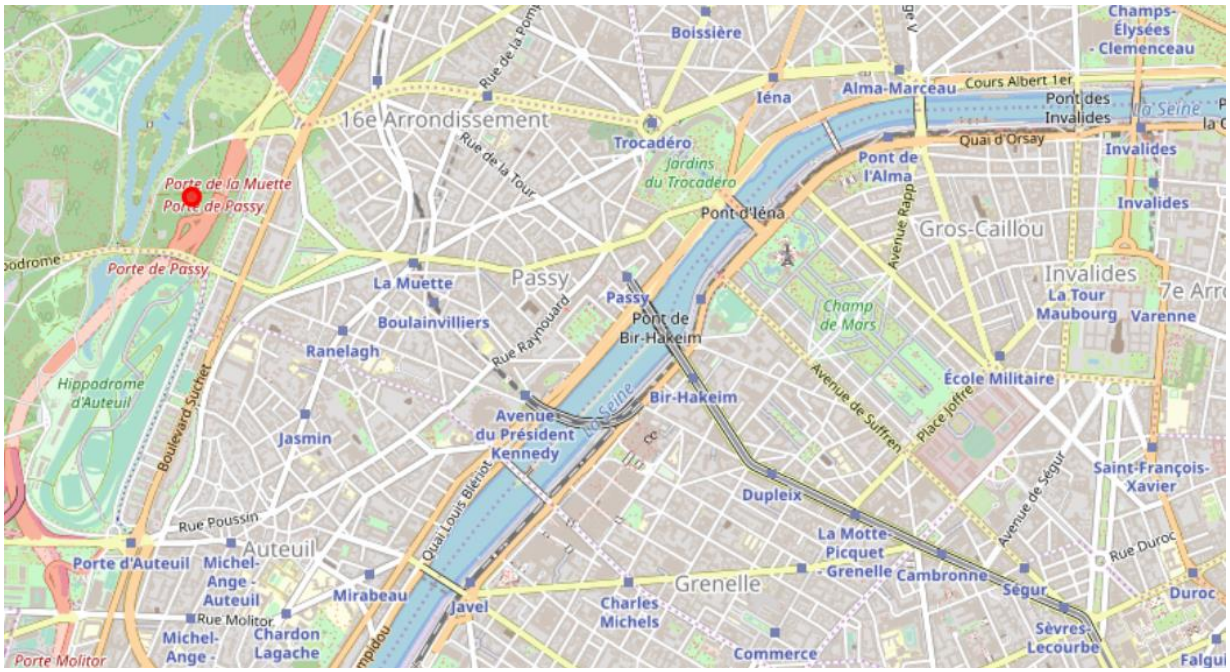
Cluster 2:

It is solely in the 12th district and the most popular venue is a zoo followed by a park. The zoo and the park is actually Zoo/Bois de Vincennes (Park) and is the largest green cover within Paris. Again, this makes this district unique.



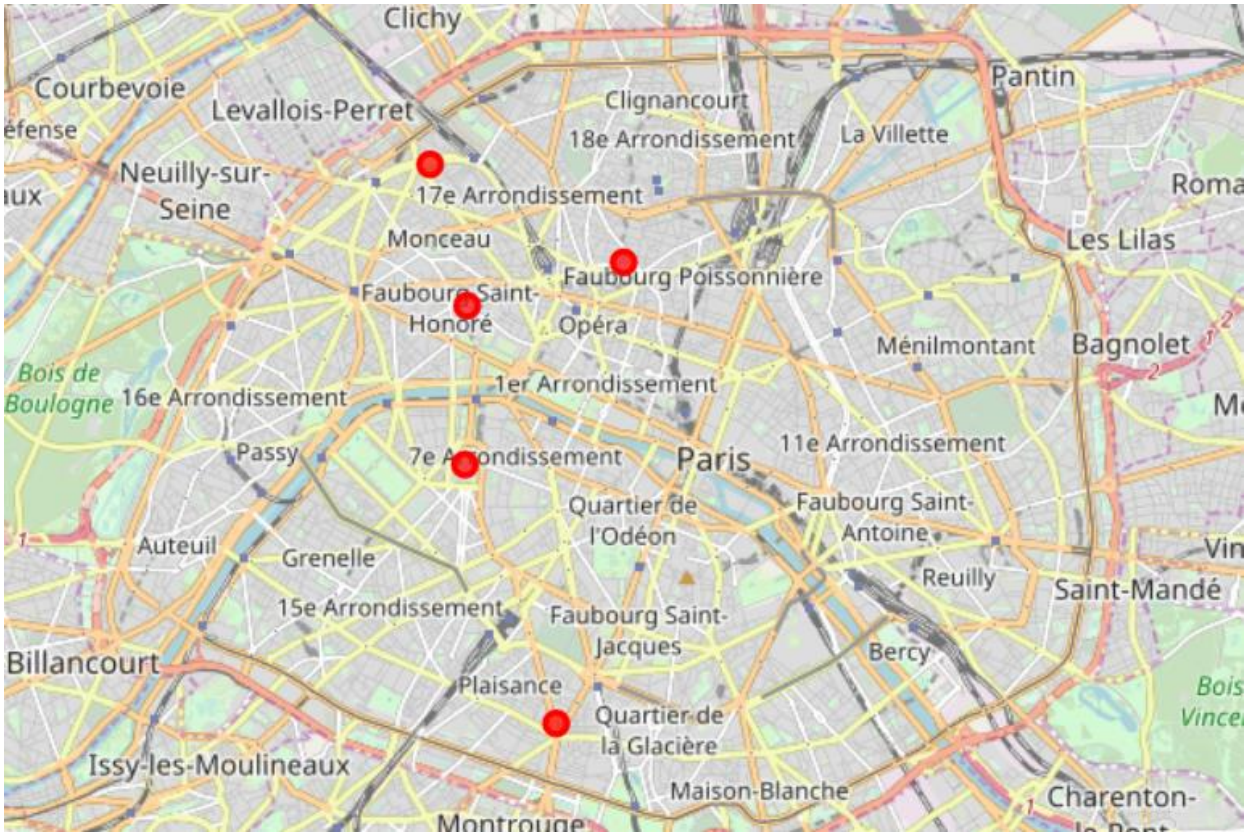
Cluster 3:

It is in the 16th district and the most popular venue is a lake. There is a small lake in Bois de Boulogne, but it is not so popular. But, perhaps the Eiffel Tower and the riven cruise on the Seine in the 15th which are close by are more popular. The 2nd popular venue is a plaza: there exists several boutiques on the periphery of 15th and 16th arrondissements.



Cluster 4:

There are venues from the 7th, 8th, 9th, 14th and 17th districts (French restaurants and hotels are most popular). 7th, 8th and 9th districts are very popular with the tourists while 14th district is preferred by the youngsters.



Summary:

Paris being a very touristic city, the results are bound to be more biased towards hotels and French restaurants as expected.

The KMeans clustering algorithm was quite efficient in identifying unique districts of Paris based on the FourSquare venue inputs.