

The-Battle-of-Neighborhoods-Week-1

Introduction:

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

The objective was to study the different neighbourhoods 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.

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:

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**.

Part 2:

The next step was to plot the map of Paris with its arrondissements using Folium.

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. 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 clusters are determined.

Conclusions:

Cluster 1:

From all 20 districts: The most popular are French restaurants with the exception of 18th (bar), 18th (bar), 20th districts (Pizza) and 6th (Italian). I believe that bars should have shown up in the 3rd, 6th or 10th and 11th districts. The 2nd most popular choice is between hotel and French restaurants.

Cluster 2:

From only 1 district: 12th arrondissement: The most popular choice is the zoo exhibit and the 2nd is the park. It is quite evident as there is Zoo de Vincennes and Park de Vincennes which are quite popular.

Cluster 3:

From only 1 district: It is the 15th district and FourSquare tags it as 'lake' being most popular but realistically, most people go there to see the Eiffel tower and take a river cruise on the river, Seine. The next popular is Pizza, which is obvious after some rigorous sightseeing.

Cluster 4:

From 4 districts: Mostly for French restaurant and Hotel (7, 8, 17 arrondissements: These are quite touristic quartiers of Paris and perhaps the samples taken into account accounts for this bias (if FourSquare users are also rich). The 14th arrondissement is popular among the young and budget travellers.

Summary:

Paris being a very touristic city, the results are bound to be more biased towards hotels and French restaurants as expected. A more detailed study is necessary to find popular venues apart from these. Perhaps a higher order to clustering could improve the results.