Segmenting and clustering neighborhoods in Paris

I – Introduction

Problem and discussion of the background

Paris is a densely populated city with more than 2 millions of inhabitant for an area of 105 km^2 so namely a density of about 20,000 inhabitant per km^2 [1, 2]. The city is divided into 20 neighborhoods named the arrondissements of Paris, there are numbered from 1 to 20. In broad outline, the more the arrondissement is far away from the center of the city, the higher is the number associated.
Paris is also one of most touristic places in the world. The city is then combining lots of inhabitants along with many touristic venues like museum, parks, restaurants, and café to satisfy tourists' desires. We want to classify the arrondissements depending on the venues and the density of inhabitants. We could also determine whether a neighborhood is more likely to attract tourist or inhabitants. This study could thus interest people who wnt to settle in Paris and choose their neighborhood according to the type of arrondissement it is and what venues they can find.

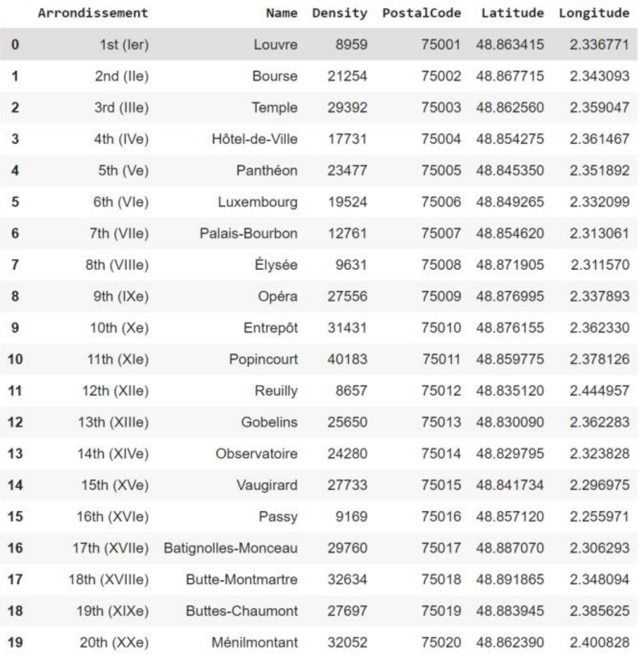
Data description

To analyze the problem, we exploited the following data:

- We've operated web scrapping on the Wikipedia page giving information about the arrondissements of Paris [3], such as the arrondissement number, its name, area, population (2017 estimate) and density [4].
- We've extracted the coordinates of each arrondissement with the geocoder python package [5].
- Foursquare API was used to get the most common venues for each given arrondissement of Paris [6].

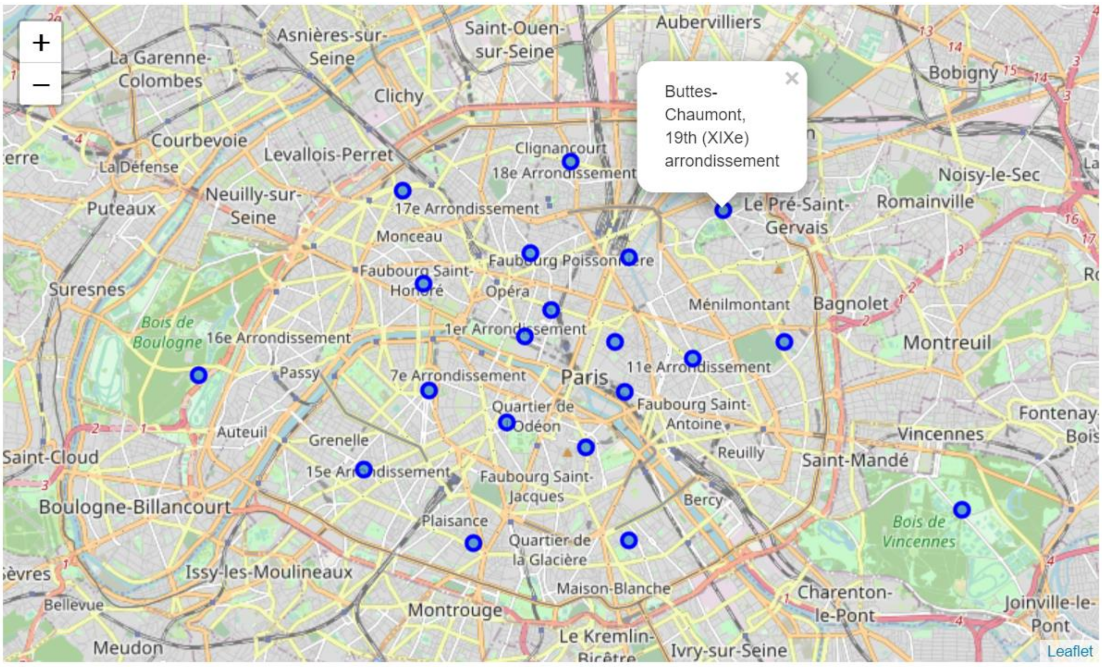
II – Methodology

We have scraped the Wikipedia page to obtain a table with arrondissements of Paris, name, and density. We have also added a column corresponding to the postal code of the arrondissement which is like 750xx where xx is the number of arrondissements. This column was added to be able to use the geocoder package to extract the latitude and longitude of the arrondissements and add corresponding columns in the dataframe.



	Arrondissement	Name	Density	PostalCode	Latitude	Longitude
0	1st (Ier)	Louvre	8959	75001	48.863415	2.336771
1	2nd (IIe)	Bourse	21254	75002	48.867715	2.343093
2	3rd (IIIe)	Temple	29392	75003	48.862560	2.359047
3	4th (IVe)	Hôtel-de-Ville	17731	75004	48.854275	2.361467
4	5th (Ve)	Panthéon	23477	75005	48.845350	2.351892
5	6th (VIe)	Luxembourg	19524	75006	48.849265	2.332099
6	7th (VIIe)	Palais-Bourbon	12761	75007	48.854620	2.313061
7	8th (VIIIe)	Élysée	9631	75008	48.871905	2.311570
8	9th (IXe)	Opéra	27556	75009	48.876995	2.337893
9	10th (Xe)	Entrepôt	31431	75010	48.876155	2.362330
10	11th (XIe)	Popincourt	40183	75011	48.859775	2.378126
11	12th (XIIe)	Reuilly	8657	75012	48.835120	2.444957
12	13th (XIIIe)	Gobelins	25650	75013	48.830090	2.362283
13	14th (XIVe)	Observatoire	24280	75014	48.829795	2.323828
14	15th (XVe)	Vaugirard	27733	75015	48.841734	2.296975
15	16th (XVIe)	Passy	9169	75016	48.857120	2.255971
16	17th (XVIIe)	Batignolles-Monceau	29760	75017	48.887070	2.306293
17	18th (XVIIIe)	Butte-Montmartre	32634	75018	48.891865	2.348094
18	19th (XIXe)	Buttes-Chaumont	27697	75019	48.883945	2.385625
19	20th (XXe)	Ménilmontant	32052	75020	48.862390	2.400828

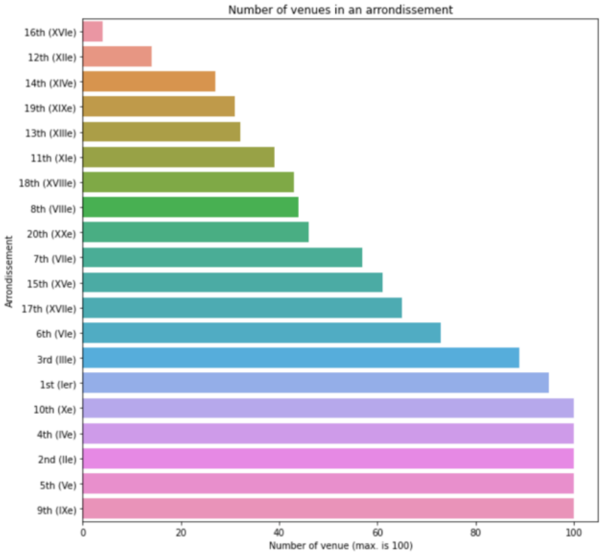
Then, we used python folium library to visualize geographic details of Paris and we put some markers showing the location of arrondissements and name.

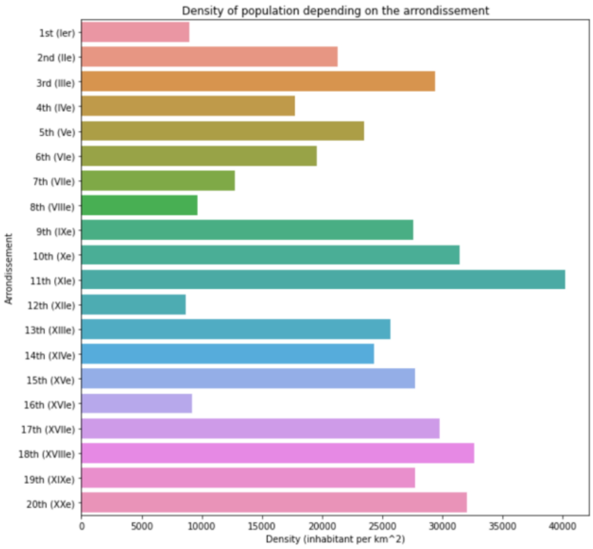


To continue, we used foursquare API to retrieve the existing venues for each arrondissement with a limit of 100 venues in a radius of 500 meter. There are 198 unique venue categories which have been found for all the neighborhoods with this API.

According to the figure below representing the number of venues in an arrondissement, in this small radius of 500 meters, we can observe that the maximum number of venues found by the API is only achieved for the 9th, 5th, 2nd, 4th and 10th (and almost for the 1st) arrondissements. Here we can note that the more the arrondissement is located in the center of Paris, the more venues there will be in a small radius.

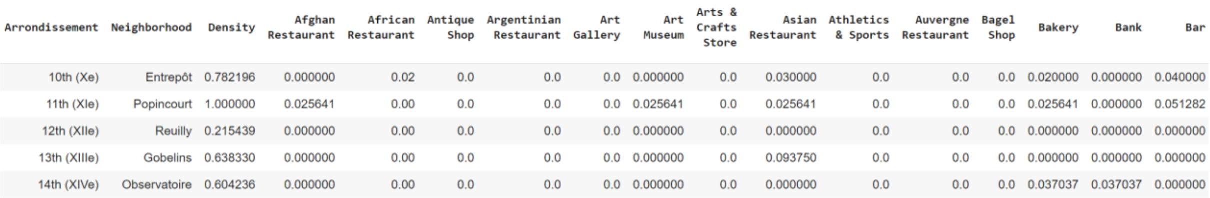
We can thus wonder if theses arrondissements also correspond to the most densely populated arrondissements. The figure below on the right do not show a special trend for these neighborhoods with a high density of venue, they can be densely populated (like 9th and 10th arrondissements) or lightly populated (such as the 1st or 2nd arrondissements).





III – Modeling

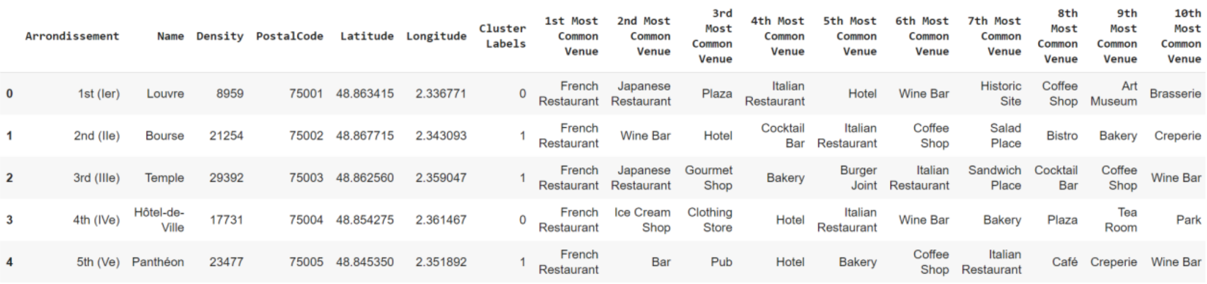
To modelize our data, we choose to use the k-means clustering model. For that, we need to do one-hot encoding (since k-means clustering model can only take numerical data as inputs) on the venue categories appearing in each neighborhood. Next, we must group rows by neighborhood and take the mean of frequency of occurrence of each category. We also need to normalize the density column to obtain a number between 0 and 1 in the same manner as the venue categories. The data looks like this :



Arrondissement	Neighborhood	Density	Afghan Restaurant	African Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auvergne Restaurant	Bagel Shop	Bakery	Bank	Bar
10th (Xe)	Entrepôt	0.782196	0.000000	0.02	0.0	0.0	0.0	0.000000	0.0	0.030000	0.0	0.0	0.0	0.020000	0.000000	0.040000
11th (XIe)	Popincourt	1.000000	0.025641	0.00	0.0	0.0	0.0	0.025641	0.0	0.025641	0.0	0.0	0.0	0.025641	0.000000	0.051282
12th (XIIe)	Reuilly	0.215439	0.000000	0.00	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000
13th (XIIIe)	Gobelins	0.638330	0.000000	0.00	0.0	0.0	0.0	0.000000	0.0	0.093750	0.0	0.0	0.0	0.000000	0.000000	0.000000
14th (XIVe)	Observatoire	0.604236	0.000000	0.00	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.037037	0.037037	0.000000

Next, to train k-means model, we drop the arrondissement and neighborhood columns. We chose k=3 which seems to be a good choice according to the small number of neighborhoods and we fit the model with the dataframe shown above (without arrondissement and neighborhood columns).

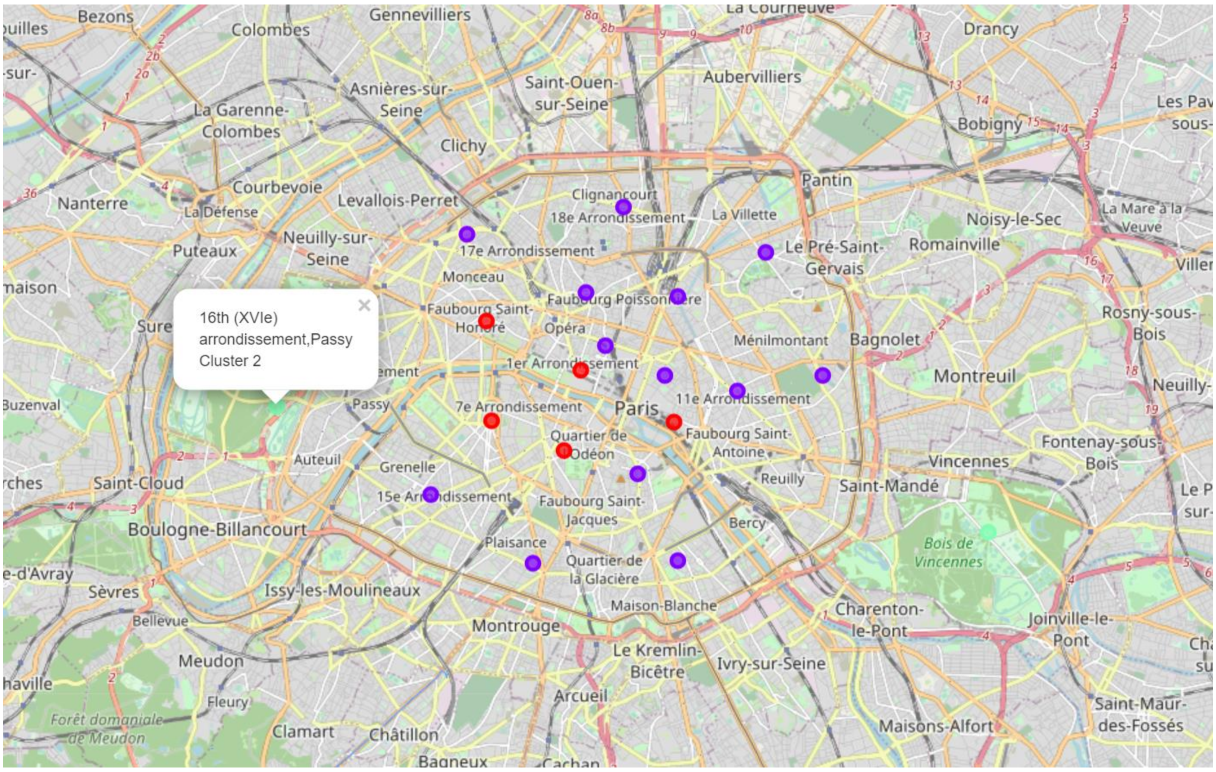
Furthermore, a table showing the 10 most common venues for each neighborhood has been created. A cluster labels column has been added to this table.



	Arrondissement	Name	Density	PostalCode	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	1st (Ier)	Louvre	8959	75001	48.863415	2.336771	0	French Restaurant	Japanese Restaurant	Plaza	Italian Restaurant	Hotel	Wine Bar	Historic Site	Coffee Shop	Art Museum	Brasserie
1	2nd (IIe)	Bourse	21254	75002	48.867715	2.343093	1	French Restaurant	Wine Bar	Hotel	Cocktail Bar	Italian Restaurant	Coffee Shop	Salad Place	Bistro	Bakery	Creperie
2	3rd (IIIe)	Temple	29392	75003	48.862560	2.359047	1	French Restaurant	Japanese Restaurant	Gourmet Shop	Bakery	Burger Joint	Italian Restaurant	Sandwich Place	Cocktail Bar	Coffee Shop	Wine Bar
3	4th (IVe)	Hôtel-de-Ville	17731	75004	48.854275	2.361467	0	French Restaurant	Ice Cream Shop	Clothing Store	Hotel	Italian Restaurant	Wine Bar	Bakery	Plaza	Tea Room	Park
4	5th (Ve)	Panthéon	23477	75005	48.845350	2.351892	1	French Restaurant	Bar	Pub	Hotel	Bakery	Coffee Shop	Italian Restaurant	Café	Creperie	Wine Bar

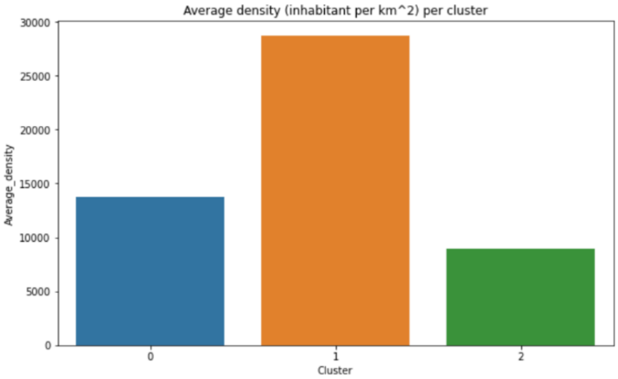
IV – Results analysis and discussion

With follium, we have then created a map to visualize the different clusters in Paris neighborhoods.

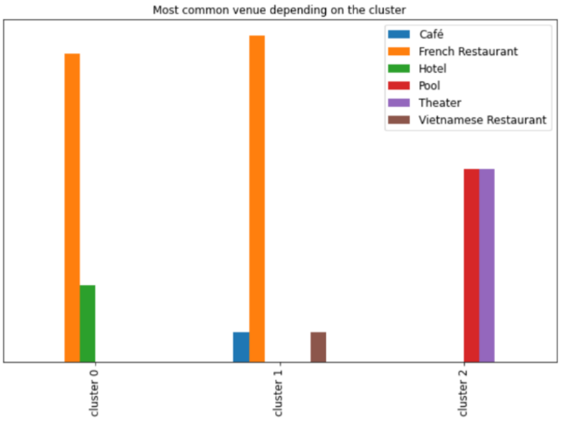


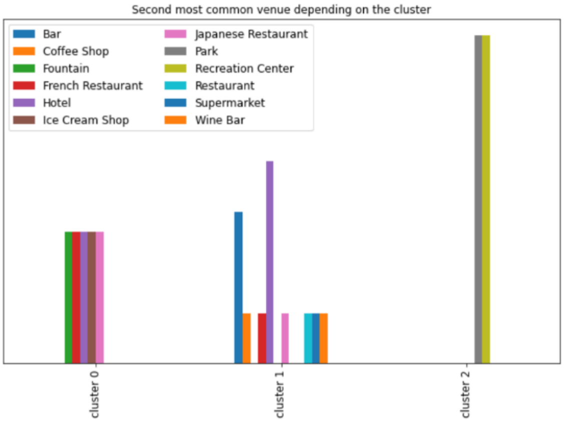
16th (XVIe) arrondissement,Passy
Cluster 2

We can notice that neighborhoods in cluster 0 (in red) seem to be located in the heart of Paris whereas neighborhoods of cluster 2 (in green) are located in the outskirt of Paris and they are characterized by their location near woods (the *Bois de Vincennes* and *Bois de Boulogne*). That is an interesting behavior given that the location of neighborhoods where not considered when modeling with k-means algorithm.



The two bar plots exposed here show the characteristics of the 3 clusters. One major difference between clusters is the density of population. The density of cluster 0, in the heart of Paris, is moderate, the density of cluster 1 is high and the density of cluster 2, in the outskirt of Paris, is quite low. To better understand the way k-means algorithm had clustered the neighborhoods, a bar plot of the 1st most common venue in the different cluster is plotted. The major difference between cluster 0 / 1 and 2 appears to be the presence of French restaurants. Both cluster 0 and 1 have French restaurants as the most common venue whereas for cluster 2 it is pool or theater. When it comes to the second most common venue depending on the cluster, we note that cluster 0 and 1 have many different venues but they still are some places to get together for a drink or a meal: bar, coffee shop, restaurants… We also see that there are some hotels in these neighborhoods. On the other hand, cluster 2 second most common venues are park and recreation centers, so venues to do some activities less focus on sharing drink or food and more into sports or nature. In summary, cluster 0 and 1 have similar venues in their surroundings and their major difference lies in the difference in their density of population. Cluster 2 gathers neighborhoods closer to the nature and of activities focused on sports and nature.





What is very interesting is that when we dig a bit more on the internet to understand the space of touristic places in the different arrondissements, we realize that cluster 0 gather all the arrondissements (1st, 4th, 6th, 7th, 8th) that have the most famous places of Paris. In the 1st arrondissement there are the *Jardin des Tuileries* along with the *Louvre*, in the 4th, *Notre-Dame de Paris*, Pompidou museum, in the 6th, the *Palais du Luxembourg* along with its *Jardins*, the *Pont des Arts* (bridge with love locks), in the 7th, there are the Eiffel Tower and Orsay museum, finally in the 8th, there are the *Arc-de-Triomphe*, the *Grand Palais* and the *Champs-Elysées*… The presence of all these touristic places makes these arrondissements less welcoming to the Parisians and are more focused on tourists, that is why we've observed that these neighborhoods have mid/low density of population whereas it has many similar venues as cluster 1.

V – Conclusion

To conclude, cluster 1 represents the arrondissements that people choose if they prefer being away from tourism and benefit from a real neighborhood life along with some nice place to have a drink or meal with friends. Cluster 0 is made for people who are not bothered to meet too many tourists and like the same king of places than in cluster 0. Finally, cluster 2 is the arrondissements for people who prefer to be in the outskirts of town and access more easily to nature and sport facilities.

References

[1] INSEE, "Comparateur de territoire – Département de Paris (75)," vol.

. [Online]. Available: https://www.insee.fr/fr/statistiques/1405599?geo=DEP-75.

[2] Wikipedia, "List of cities proper by population density." [Online]. Available: https://en.wikipedia.org/wiki/List_of_cities_proper_by_population_density#cite_note-paris-33.

[3] Wikipedia, "Arrondissements of Paris." [Online]. Available: https://en.wikipedia.org/wiki/Arrondissements_of_Paris.

[4] Wikipedia, "Demographics of Paris." [Online]. Available: https://en.wikipedia.org/wiki/Demographics_of_Paris.

[5] "Geocoder package." [Online]. Available: https://pypi.org/project/geocoder/.

[6] Foursquare, "Foursquare API." [Online]. Available: https://developer.foursquare.com/.