Capstone Project
The Battle of Neighborhoods in Paris
Method for determining the best location for a new restaurant

Christophe Coudé

December 14, 2020

1. Introduction

1.1 Background

There are currently around 12,000 restaurants in Paris for 2,150 million inhabitants. That is to say, on average, about 1 restaurant per hectare. That's a lot, the competition is tough. Many restaurants are closing, especially in these times of crisis. A newcomer might then wonder where to set up his (or her) new restaurant. He may then have the choice between :

- finding a place where there are very few restaurants of his specialty in order to attract customers of the local neighborhood who want to change cuisine or,
- setting up his restaurant where there are already many restaurants of his own specialty because even if there is competition, this means that there is an already existing clientele.

The restaurant owner must therefore know what is the distribution of "cuisines" in the different neighborhoods in order to be able to make his choice.

1.2 Business problem and interest

The aim of this project is therefore to offer new restaurant owners a global vision of the distribution of restaurants in Paris thanks to a clustering which makes it possible to know for each district which are the most represented cuisines or regions and those less represented, and a transversale view across the different districts.
According to the strategy chosen by the owner, he (or she) will be able to choose in which district to try settling and the right spot in the district.

Therefore, the problem is :

- to find where the restaurants are located and to know their specialty, for example using the *Foursquare API* [1], and then
- to use machine learning to bring out the general culinary trends and tastes of each neighborhood, for example the unsupervised learning method *Clustering*
- to visualize the position of the various restaurants in the choosen district, for example with *Folium* [2].

2. Data acquisition and cleaning

2.1 Data sources

Worldwide venues description can be accessed through the **Foursquare API** once an account has been created on their site [1].

The different categories of cuisines are referenced on their page of categories [3].

We can see an example of hierarchy describing the kind of cuisine and their ID:

Food
4d4b7105d754a06374d81259

> Afghan Restaurant
> 503288ae91d4c4b30a586d67
> African Restaurant
> 4bf58dd8d48988d1c8941735
>
> > Ethiopian Restaurant
> > 4bf58dd8d48988d10a941735

To request Foursquare, we need the category of restaurant and the GPS coordinates of the neighborhoods of Paris.

It's quite difficult to find accurate GPS coordinates... Each site gives different coordinates... So, for the center of Paris, needed to center the Folium map, I chose the site ***latitude.to*** [4] that I scraped with the library BeautifulSoup.

However, for the coordonates of the 20 districts of Paris, I couldn't find correct coordinates on the Web, so I "manually" used ***Google Map*** [5] to approximately determine them.

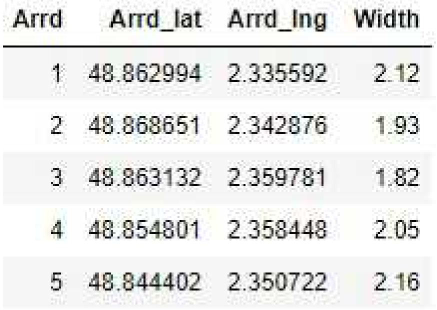
2.2 Data cleaning

There was another difficulty. Foursquare uses circles to demarcate the areas where it will look for the venues. But the Parisian districts are not round.

So I had to manually draw several circles (from 1 to 3) inside the neighborhoods in order to get as many restaurants as possible, even though Foursquare requests only return 30 venues per category each time. So it may be not exactly accurate, but it's the principle.

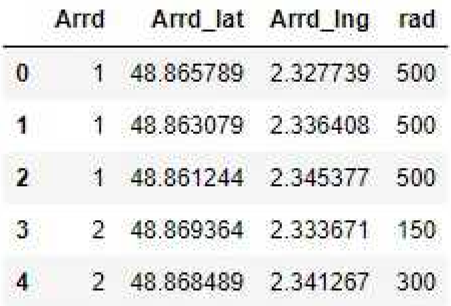
I couldn't draw a unique big circle encompassing each district because Foursquare only brings back 30 venues per search research. I would have lost more venues.

I entered the coordinates of the centers of the districts in my local file
district_centers_coordinates.csv.



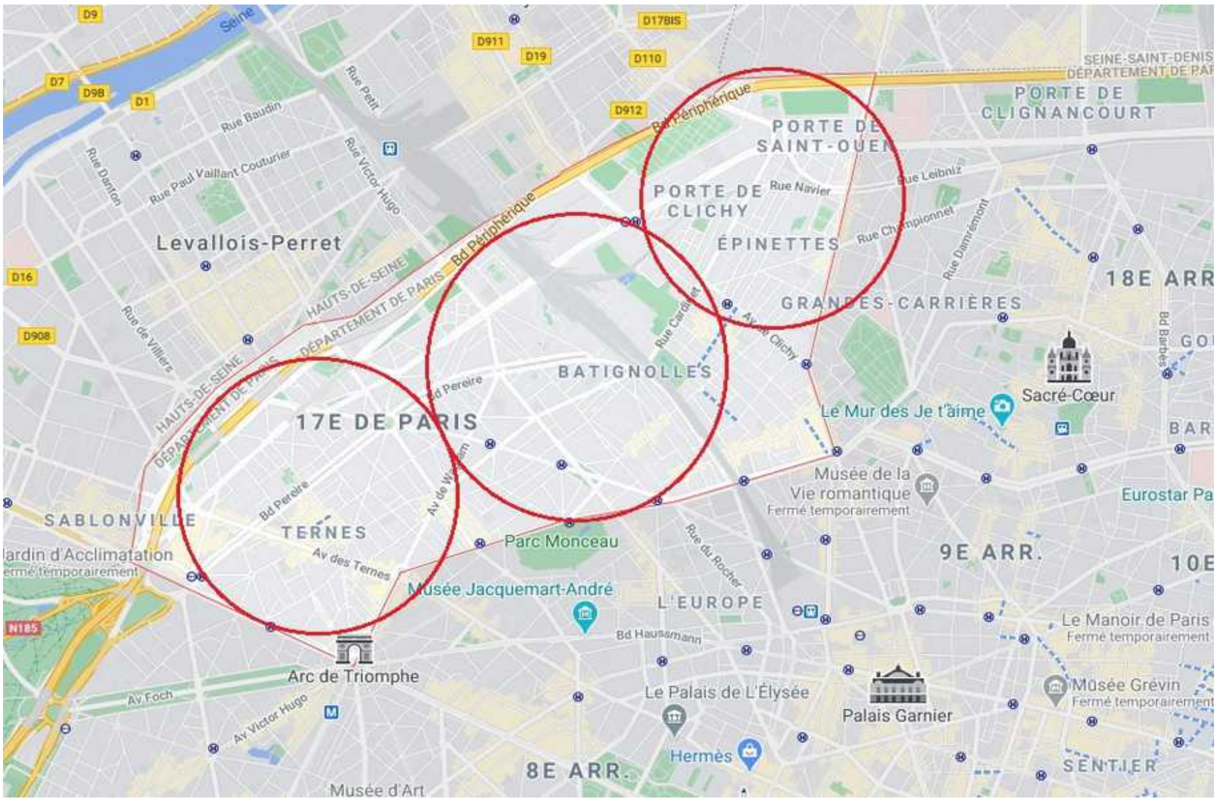
Arrd	Arrd_lat	Arrd_lng	Width
1	48.862994	2.335592	2.12
2	48.868651	2.342876	1.93
3	48.863132	2.359781	1.82
4	48.854801	2.358448	2.05
5	48.844402	2.350722	2.16

And I have entered the centers of the circles and their radius in my local file *district_coordinates.csv.*



	Arrd	Arrd_lat	Arrd_lng	rad
0	1	48.865789	2.327739	500
1	1	48.863079	2.336408	500
2	1	48.861244	2.345377	500
3	2	48.869364	2.333671	150
4	2	48.868489	2.341267	300

Here's an example of how the approximation renders for the 17th district :



2.3 Feature selection

There are about 250 food categories in Foursquare but it was too much because Fousquare limits to 950 requests per day for a free account. So I restricted to 69 categories. Multiplicated by the 46 "circles" in the 20 districts, that makes 3174 requests to Foursquare, i.e. 4 days.

I also added a « region » feature to geographically group most of the cuisine categories. Foursquare had already did it, for example Asia, Africa and so on. But as some venues were only to be find in sub-categories or only in regions, I decided to create a feature to group every country of a continent.

I just made three exceptions to try to highlight some specificities in Paris. French, Indian and Japanese venues. But they can be respectively regrouped in Europa and Asia. For example, we'll see that Japanese restaurants are often more common (5th region place) than America (6th place).

3. Methodology

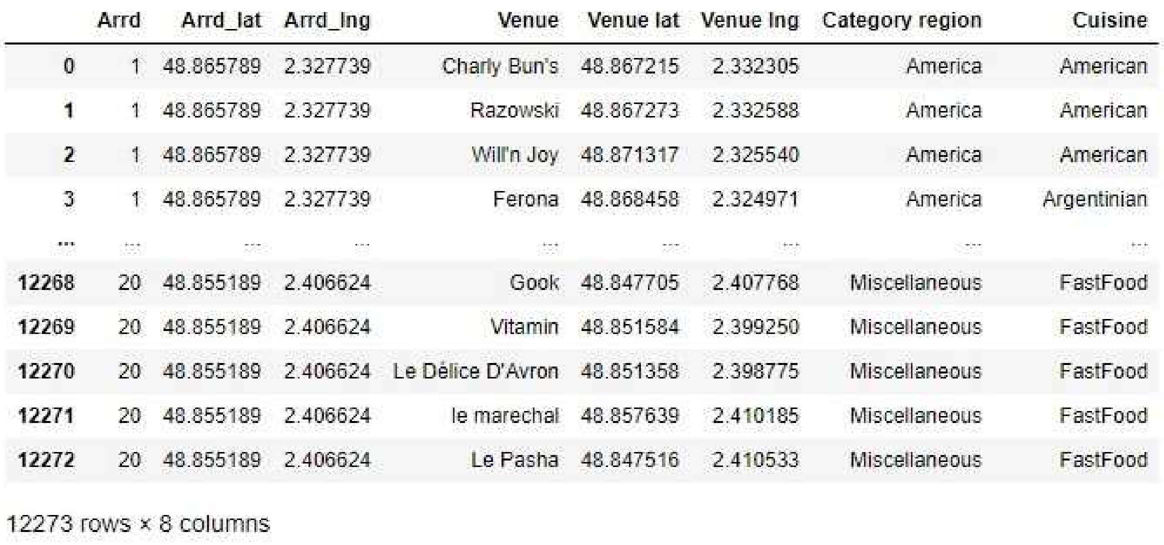
The goal is to find a place that suits the new restaurateur. As a case study, we will take a person who wants to open an Indian restaurant. He chooses the strategy of finding a place in Paris where Indian restaurants are not very numerous and where there are few other restaurants. The method is suitable for any other strategy, it's just to make a choice.

- The first step is to retrieve the data, GPS coordinates of the districts, list of the different restaurants with their category and their GPS coordinates.
- The second step makes it possible to analyze the distribution of restaurant categories in the different districts in a vertical way, district by district, then transversally across the districts thanks to a clustering in order to highlight the major trends in distribution by grouping them together. For example a cluster where Asian restaurants are the majority, a cluster where French restaurants lead, another where fast food is preferred, etc. These two steps allow the new owner to determine in which district he will try to set up his restaurant.
- The third step studies the distribution of restaurants in choosen district by visualizing where the groups of restaurants are located, the areas where there are few or no restaurants, and where the already present Indian restaurants are. The restaurateur will thus have a precise idea of the location where he could set up his restaurant.

If he chose to compete with other Indian restaurants in the districts where they are most represented, the proposed method also makes it possible to do so.

3.1 Exploratory Data Analysis, first step

Google Map having made it possible to retrieve the GPS coordinates, Foursquare provides the coordinates of the venues of Paris. Here is an overview with the beginning and the end of the list:



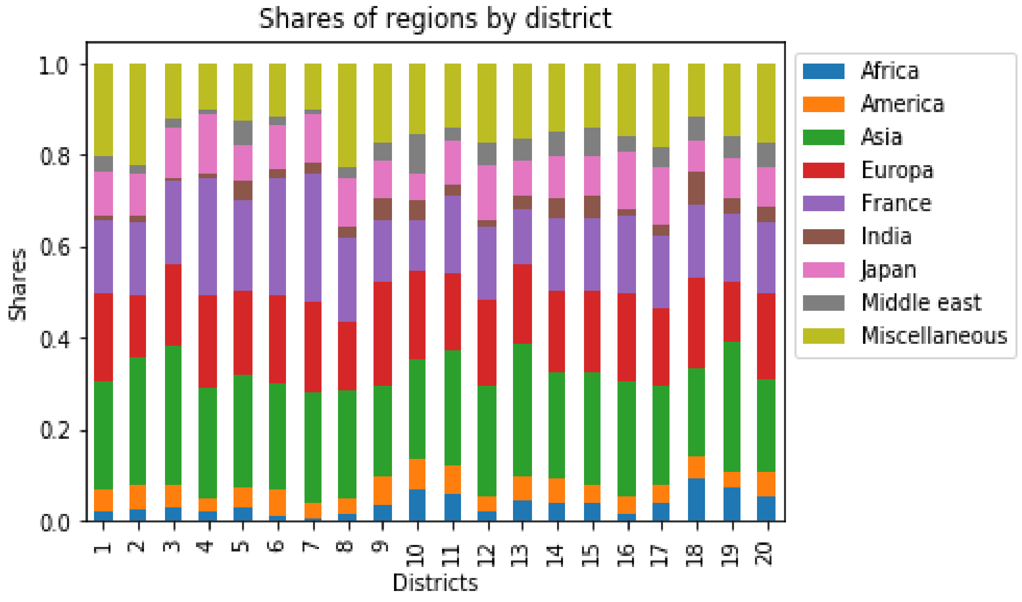
	Arrd	Arrd_lat	Arrd_lng	Venue	Venue lat	Venue lng	Category region	Cuisine
0	1	48.865789	2.327739	Charly Bun's	48.867215	2.332305	America	American
1	1	48.865789	2.327739	Razowski	48.867273	2.332588	America	American
2	1	48.865789	2.327739	Will'n Joy	48.871317	2.325540	America	American
3	1	48.865789	2.327739	Ferona	48.868458	2.324971	America	Argentinian
...	...	...	...	...	...	...	...	...
12268	20	48.855189	2.406624	Gook	48.847705	2.407768	Miscellaneous	FastFood
12269	20	48.855189	2.406624	Vitamin	48.851584	2.399250	Miscellaneous	FastFood
12270	20	48.855189	2.406624	Le Délice D'Avron	48.851358	2.398775	Miscellaneous	FastFood
12271	20	48.855189	2.406624	le marechal	48.857639	2.410185	Miscellaneous	FastFood
12272	20	48.855189	2.406624	Le Pasha	48.847516	2.410533	Miscellaneous	FastFood

12273 rows × 8 columns

3.2 Exploratory Data Analysis, second step

3.2.1 Distribution of the 8 main regions over the 20 districts

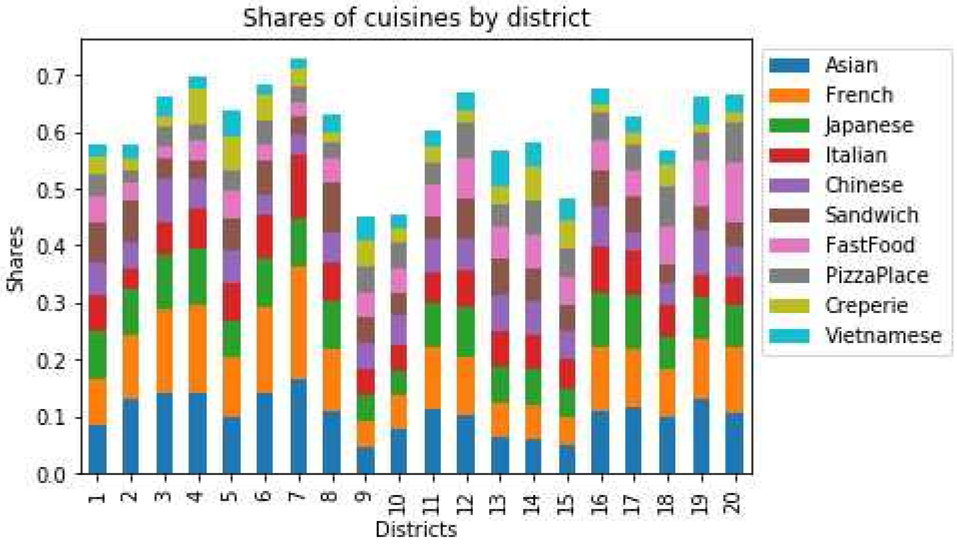
An initial analysis of the large regions shows their distribution over the 20 districts.



We can see that the distributions of the regions are fairly homogeneous « along » the districts.

3.2.2 Distribution of the 10 first cuisines over the 20 districts

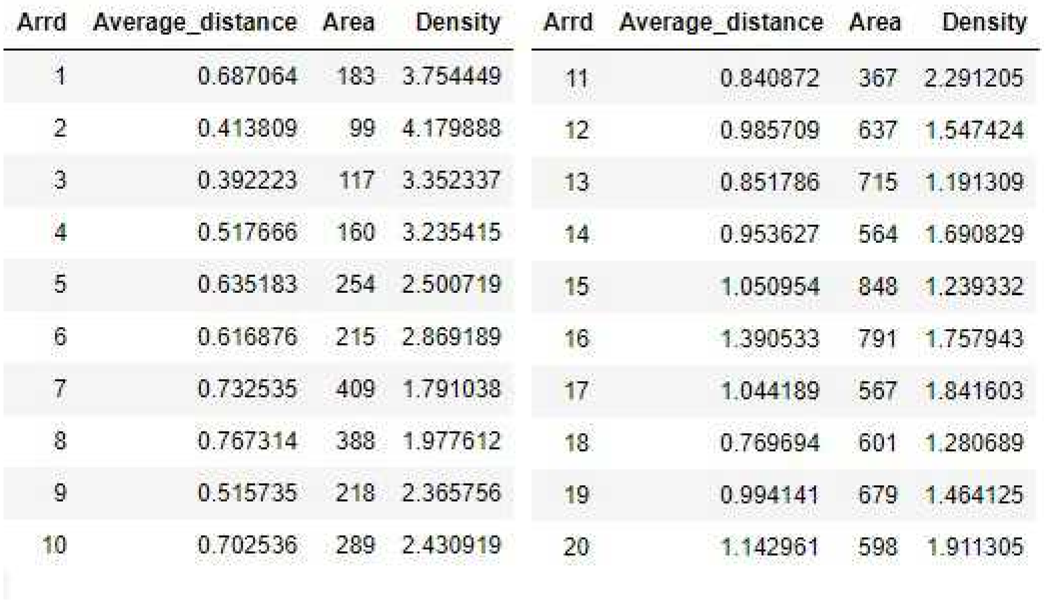
For the cuisines, as there are too many categories to make the chart readable, I only show the ten most represented categories in average.



Here, we can observe that the 10 first categories account for between 40 and 70% of the food venues in the districts. « Fast » restaurants (fast food, pizzas, sandwiches) and takeaways, caterers are omnipresent.

3.2.3 Densities of restaurants in each district

We will see more precisely with the map but we can already get an overall idea of the distance of the restaurants from their district center and their density, reduced to the area of the districts.

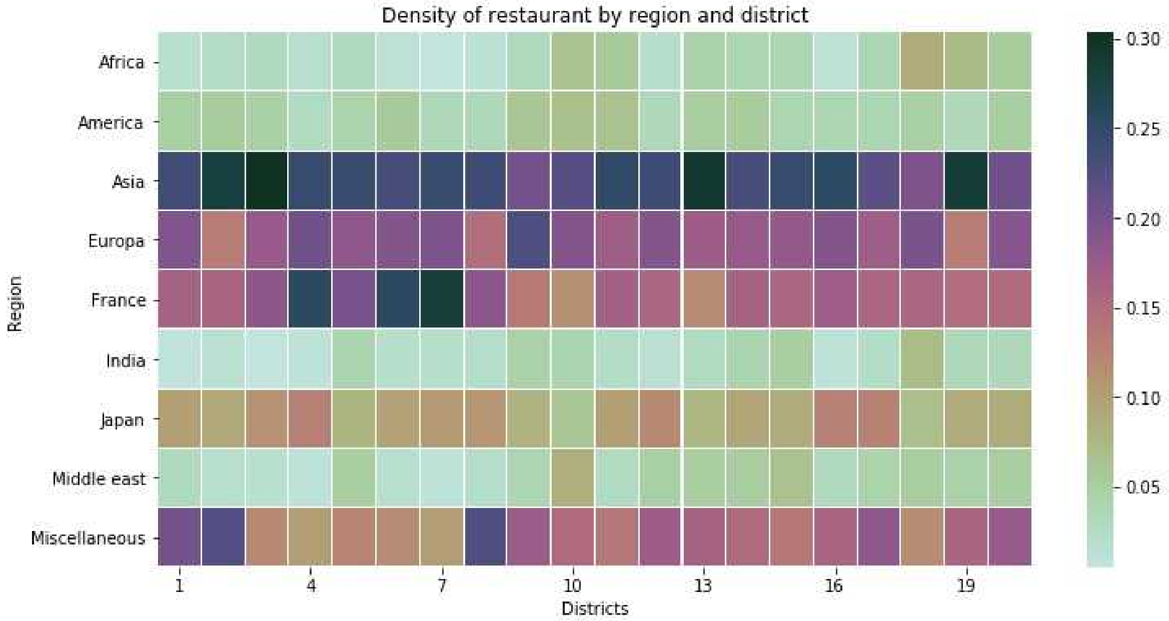


Arrd	Average_distance	Area	Density
1	0.687064	183	3.754449
2	0.413809	99	4.179888
3	0.392223	117	3.352337
4	0.517666	160	3.235415
5	0.635183	254	2.500719
6	0.616876	215	2.869189
7	0.732535	409	1.791038
8	0.767314	388	1.977612
9	0.515735	218	2.365756
10	0.702536	289	2.430919
11	0.840872	367	2.291205
12	0.985709	637	1.547424
13	0.851786	715	1.191309
14	0.953627	564	1.690829
15	1.050954	848	1.239332
16	1.390533	791	1.757943
17	1.044189	567	1.841603
18	0.769694	601	1.280689
19	0.994141	679	1.464125
20	1.142961	598	1.911305

The outer boroughs are the largest (12 to 20, they are the more recent districts incorporated in Paris) whereas the districts 5 to 11 and above all 1 to 4 are the smallest and the oldest. That's were lies a great trading activity (density above 3).

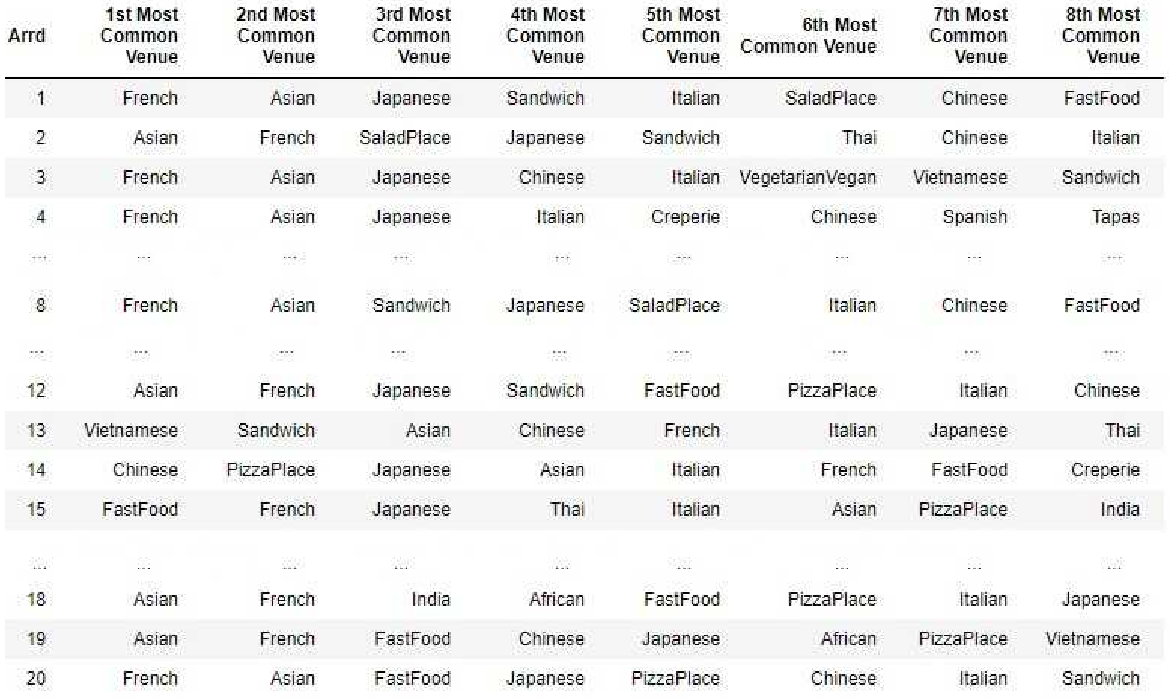
3.2.4 Other representation of « vertical » districts (district by district)

This other representation allows to visualize the distribution of the regions across the districts.



As we can see, the Asia category leads in a large majority of boroughs, followed by France and Europa. Then the "fast" cuisine and Japan.

That is a « regional » view, but can we group districts together according to national cuisines ?
Here appear the specialties classified by district :

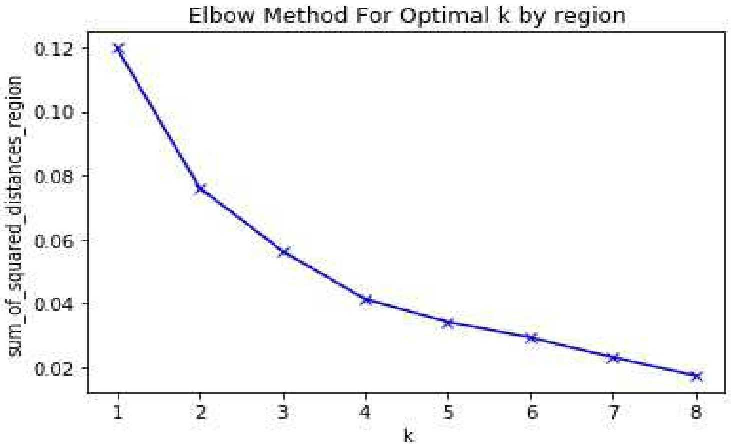


Arrd	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
1	French	Asian	Japanese	Sandwich	Italian	SaladPlace	Chinese	FastFood
2	Asian	French	SaladPlace	Japanese	Sandwich	Thai	Chinese	Italian
3	French	Asian	Japanese	Chinese	Italian	VegetarianVegan	Vietnamese	Sandwich
4	French	Asian	Japanese	Italian	Creperie	Chinese	Spanish	Tapas
...	...	...	...	...	...	...	...	...
8	French	Asian	Sandwich	Japanese	SaladPlace	Italian	Chinese	FastFood
...	...	...	...	...	...	...	...	...
12	Asian	French	Japanese	Sandwich	FastFood	PizzaPlace	Italian	Chinese
13	Vietnamese	Sandwich	Asian	Chinese	French	Italian	Japanese	Thai
14	Chinese	PizzaPlace	Japanese	Asian	Italian	French	FastFood	Creperie
15	FastFood	French	Japanese	Thai	Italian	Asian	PizzaPlace	India
...	...	...	...	...	...	...	...	...
18	Asian	French	India	African	FastFood	PizzaPlace	Italian	Japanese
19	Asian	French	FastFood	Chinese	Japanese	African	PizzaPlace	Vietnamese
20	French	Asian	FastFood	Japanese	PizzaPlace	Chinese	Italian	Sandwich

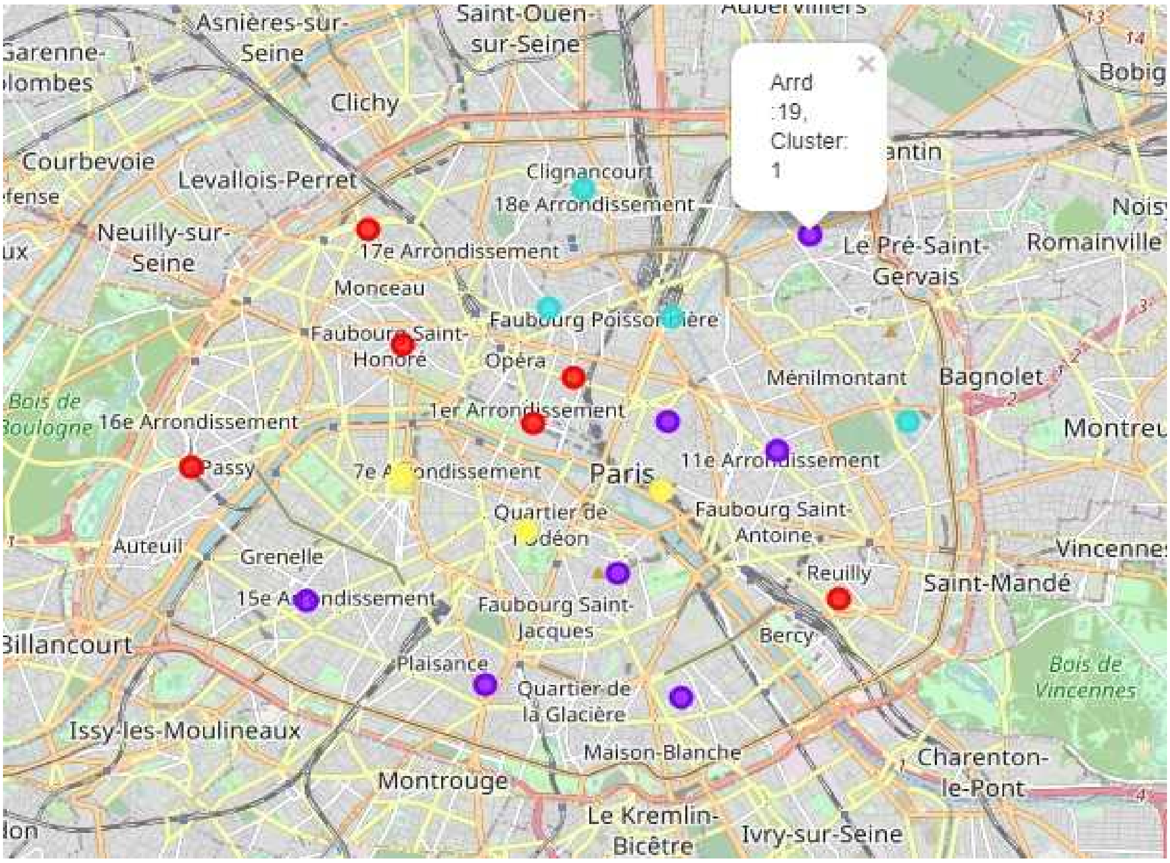
We can see that in the 8th district, « fast » cuisine (sandwiches, salads, fastfood, pizzas are 10th) is very widespread. That's a district where are located a large railway station and numerous business.

3.2.5 Clustering (horizontal representation « district wide »)

Can we retrieve these trends in national and « continental » merging ? Let's take for example the continental case clusterin :
Which number of clusters might we choose ?



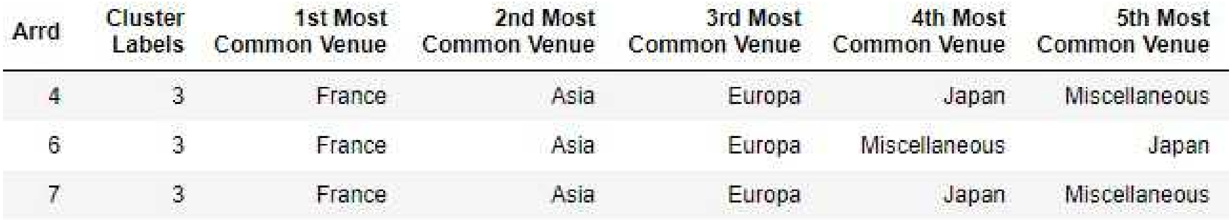
The elbow method suggests that we choose 4 clusters. And indeed, we're going to see that these 4 clusters can be interestingly interpretated. Here's what it renders using *Folium*.



We can see the following clusters :

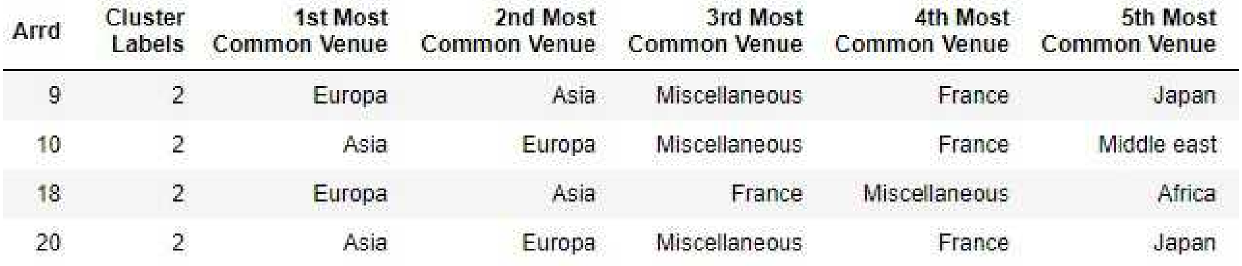
- cluster 0 : districts 1, 2, 8, 12, 16, 17 (red)
- cluster 1 : districts 3, 5, 11, 13, 14, 15, 19 (violet)
- cluster 2 : districts 9, 10, 18, 20 (turquoise)
- cluster 3 : districts 4, 6, 7 (yellow)

Let's take a look at the cluster number 3 :



Arrd	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	3	France	Asia	Europa	Japan	Miscellaneous
6	3	France	Asia	Europa	Miscellaneous	Japan
7	3	France	Asia	Europa	Japan	Miscellaneous

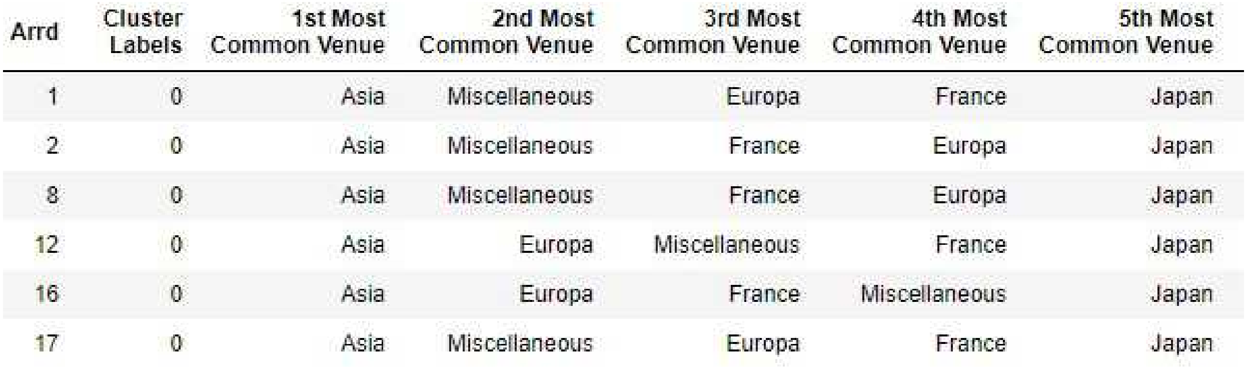
It brings together all the neighbors where France comes fisrt, then Asia and Europa.



Arrd	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	2	Europa	Asia	Miscellaneous	France	Japan
10	2	Asia	Europa	Miscellaneous	France	Middle east
18	2	Europa	Asia	France	Miscellaneous	Africa
20	2	Asia	Europa	Miscellaneous	France	Japan

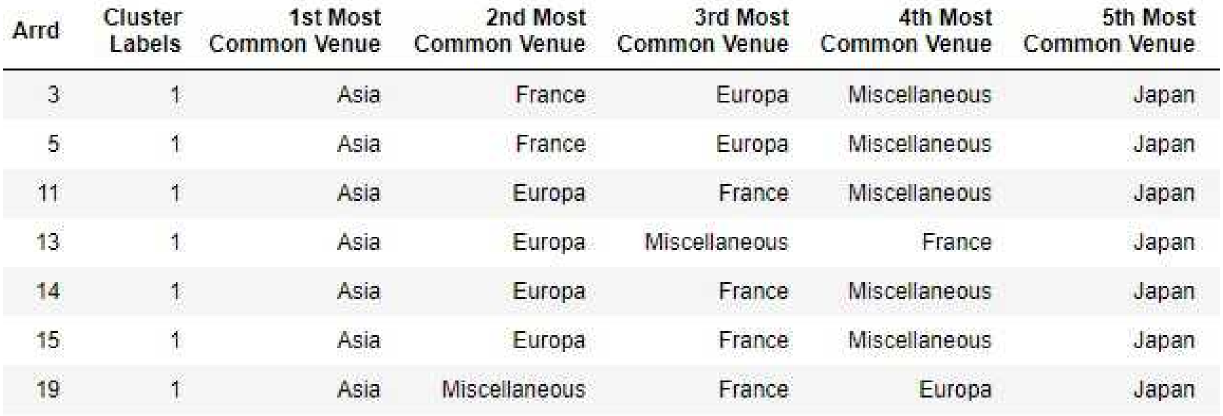
In cluster 2, Asia and Europa are trusting the two first places, the third one being assigned to fast cuisine.

In clusters 0 and 1, Asia comes first, and the distinction lies in the next positions.



Arrd	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	0	Asia	Miscellaneous	Europa	France	Japan
2	0	Asia	Miscellaneous	France	Europa	Japan
8	0	Asia	Miscellaneous	France	Europa	Japan
12	0	Asia	Europa	Miscellaneous	France	Japan
16	0	Asia	Europa	France	Miscellaneous	Japan
17	0	Asia	Miscellaneous	Europa	France	Japan

In cluster 0, the category Miscellaneous mainly before France and Europa, whereas in cluster 1, it's the contrary, France and Europa are before Miscellaneous.



Arrd	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	1	Asia	France	Europa	Miscellaneous	Japan
5	1	Asia	France	Europa	Miscellaneous	Japan
11	1	Asia	Europa	France	Miscellaneous	Japan
13	1	Asia	Europa	Miscellaneous	France	Japan
14	1	Asia	Europa	France	Miscellaneous	Japan
15	1	Asia	Europa	France	Miscellaneous	Japan
19	1	Asia	Miscellaneous	France	Europa	Japan

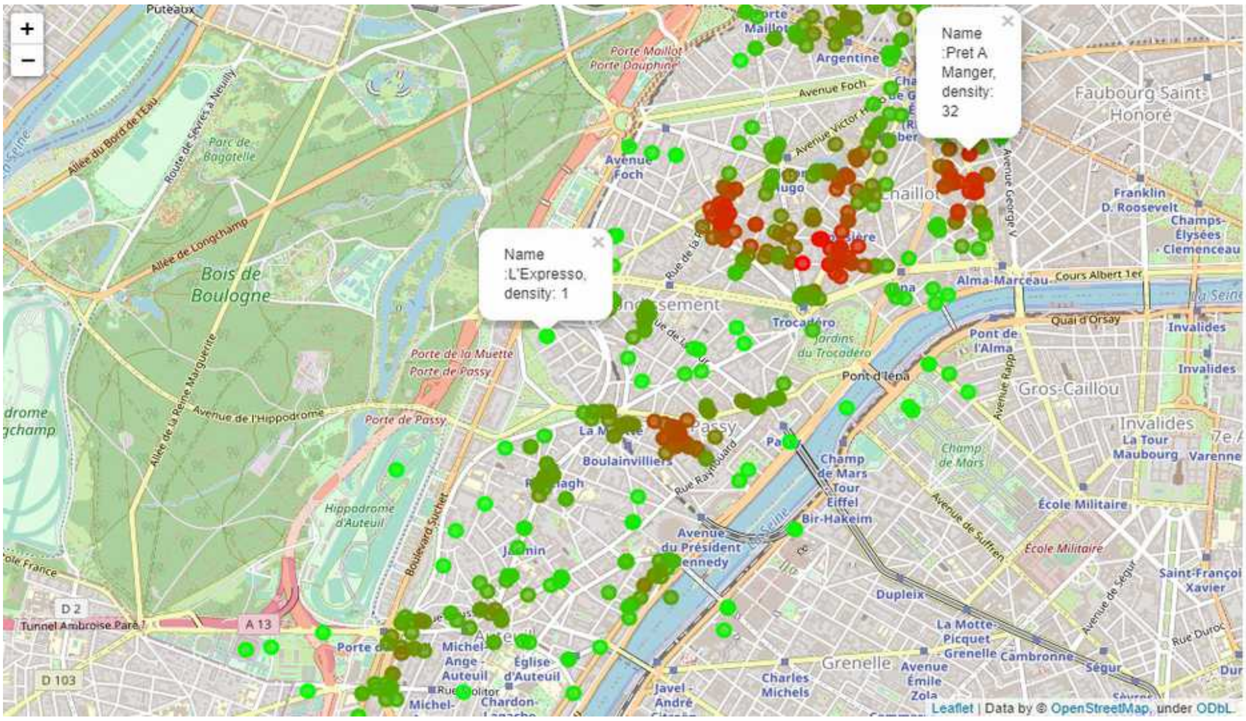
We can note that districts 12 and 13 are sharing the same 5 first categories. The distinction lies in the other categories after.

3.3 Exploratory Data Analysis, third step

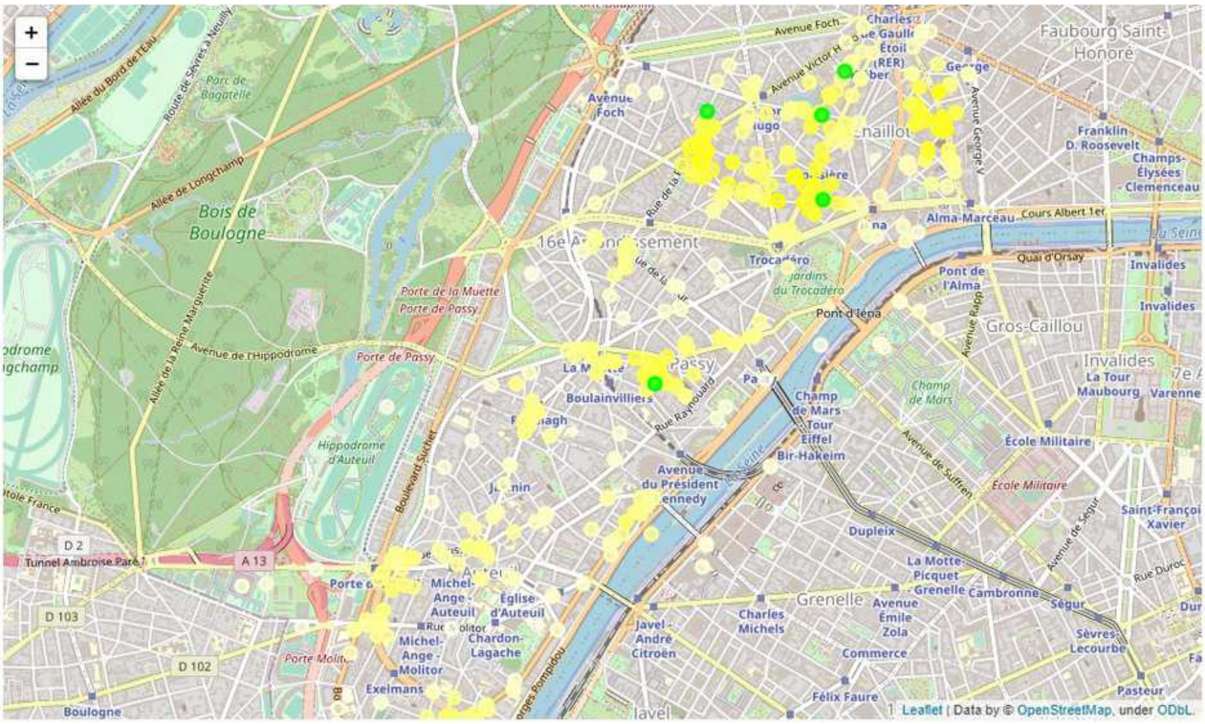
Now that the restaurateur has a good overview of the distribution of restaurants by district and « region », he can choose a district and try to find a good spot.

For our case study, we can choose the opening of an Indian restaurant in the 16th district where the density of restaurant is low as well as the proportion of Indian ones.

Data is good, but maps are better for visualizing. So I calculated for all the restaurants of Paris, a number, called « density » which accounts for the number of other restaurants located in a given circle around the restaurant. I chose a radius of 200 meters. To visualize the density of restaurants throughout the districts, district by district I kept the highest density and created a scale of colors from green to red. The greenest the most isolated, the reddest the less isolated.


+
−
Name :Pret A Manger, density: 32
Name :L'Expresso, density: 1
Bois de Boulogne
Trocadéro
Passy
Tour Eiffel
Leaflet | Data by © OpenStreetMap, under ODbL.

On this map, we clearly see where are the points of higher density. But we've here all the restaurants. What about Indian ones ?
To visualize the positions of the Indian restaurants, as we can see it on the map below, I overlaid two ranges of points.
Pale yellow to bright yellow to show the density of restaurants that are not Indian, and from green to red for the Indian restaurants already present. As there are only 5 Indian restaurants quite distant from each other, they all appear in green.



The owner has now a tool to decide where to set up his new restaurant.

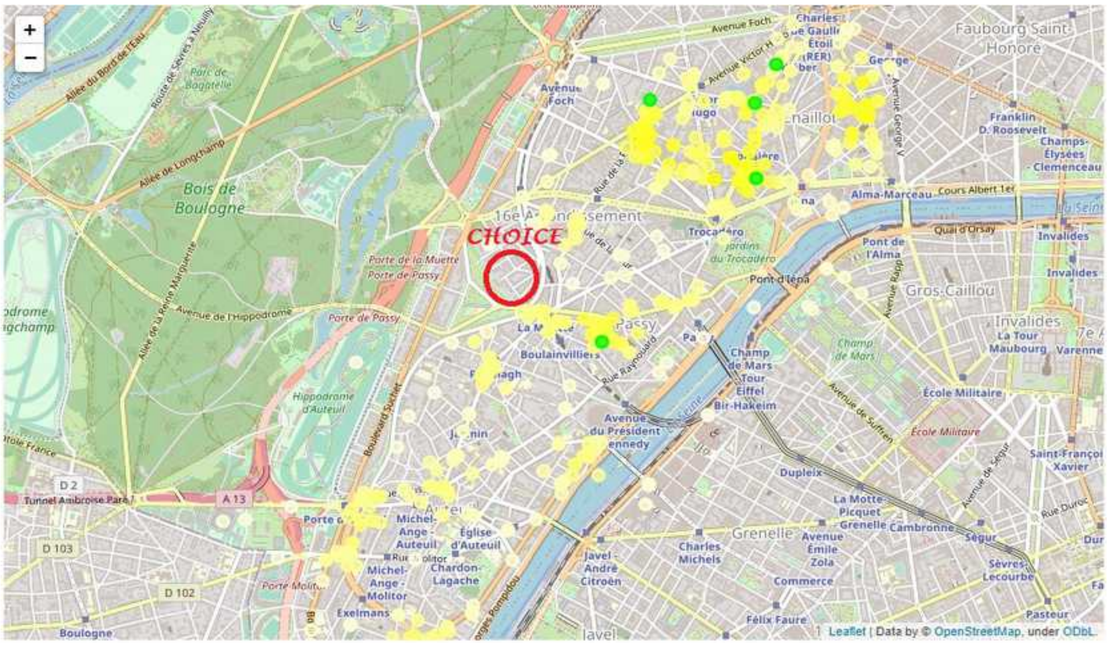
4. Results, difficulties, discussion and possible improvements

- Paris is a big city with a lot of restaurants. The method made it possible to get an idea of the distribution and density of restaurants of all nationalities across all districts.
- Due to the large number of requests, I have limited the number of categories to 69. Normally, all categories should be used. Likewise, one should check what « general » categories like Asia or Japan are accounting for because some Japanese restaurants were in both national category as well as in Asia but others were not in Asia.
- Foursquare using circles to bring back the venues, it is difficult to be exhaustive in the network of districts. This can be awkward when the density of restaurants is high at the crossroads of districts. It would necessary to increase the number of requests.
- Despite this, the method made it possible to correctly define the distribution of restaurants and their density. The clustering highlighted a number of constants in the distribution of cuisines and regions across the districts.

- Thus the « fast cuisine » cluster (number 0) brings together neighborhoods where people do not have much time to eat lunch because they have a train to catch or because they work in places where there are a lot of offices and a lot of queues.
- The cluster 3 brings together neighborhoods that have a fairly old historical and gastronomic tradition (marais, Saint-Germain-des-Près, Invalides, Tour Eiffel) where many French restaurants are located.
- Asia trusts most of the first place ; It would be even more the case in adding inside the categories Japan and India.
- The owner therefore chose the 16th district and we saw that locating the restaurants as well as their density, and separating the India category from the others, allowed him to choose the right location corresponding to the strategy he had chosen. But the method would also allow him to respond to any other strategy.

5. Conclusion

This method allows any restaurateur to analyze the distribution of restaurants and culinary trends by district (density) but also across districts (clustering) in Paris. And it is also valid in any city for an owner coming from another continent for the choice of the city. It is also valid for any other category of venues.
For example, if a person wanted to live in an area where there is a lot of nightlife, or where there are a lot of services, or even on the contrary a peaceful area with few shops. Everything is possible.
Well, of course, this is only the data part of location, density, distribution of the different categories. The future restaurateur must also do a market study depending on the location chosen to see if his business will be viable (local taxes, popularity of restaurants in the neighborhood, level of tourism and so on).



6. References

- [1] Foursquare : https://developer.foursquare.com/developer/
- [2] Folium : https://pypi.org/project/folium/
- [3] Venues categories : https://developer.foursquare.com/docs/build-with-foursquare/categories
- [4] Coordonates of Paris : https://latitude.to/map/fr/france/cities/paris
- [5] Google Map : https://www.google.com/maps/place/Paris