Table des matières

IBM Data Science Capstone

The Battle of Neighborhoods – Final Report

1. Description of the problem and a discussion of the background.

Suppose you arrive in a new town for living, and you don't know where to live. According to your needs, for example find a school for your children, not being too far from working, and to your preferences, for example having a park near your home for Sunday walking or restaurants because you're a gourmet, you will search for a neighborhood which correspond the best to you. The difficulty is that you don't know this new place and the different neighborhoods with their specificities and it will take time before you have a good knowledge of the different neighborhoods.

So how can you do? In this Capstone, I propose to answer to this question for the metropolis of Lyon using different data to give an answer to a newcomer depending on his choices.

Why the metropolis of Lyon? First, because it is where I am living for ten years, so I have a good knowledge of its neighborhoods. And a presentation of this metropole is a second kind of answer.

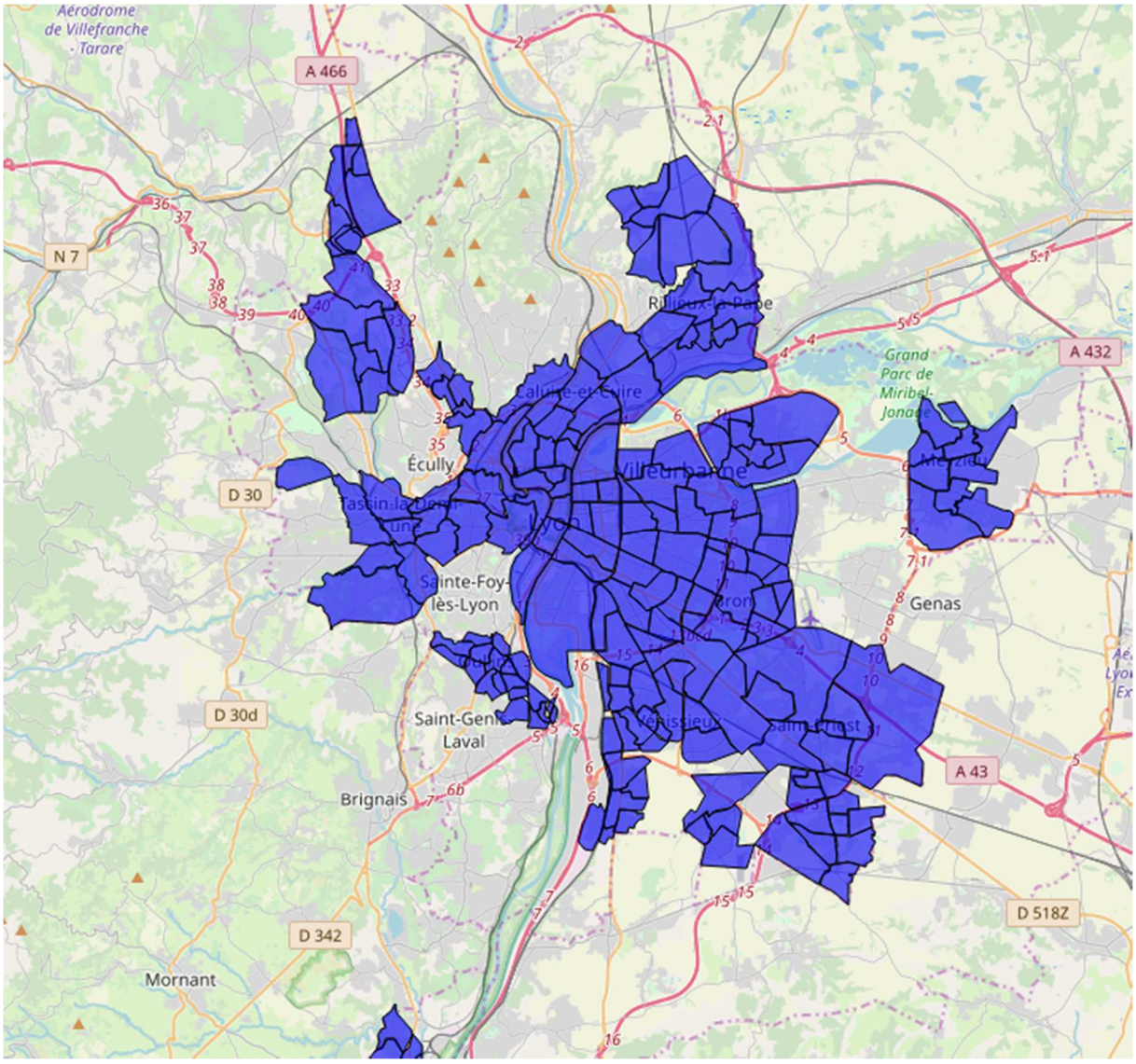
The metropolis of Lyon is one of the most economic and touristic attractive metropolises in Europe with a reasonable cost of life. It's a crossroads in the heart of Europe with performing transportations, economically dynamic with sectors of excellence, universities and research centers and a place where life is good with a great environment: a lot of green spaces, many historical sites (10% of the city of Lyon is listed as a UNESCO World Heritage Site), a rich cultural life and multiple sports and leisure activities. And do not forget that Lyon is the capital of gastronomy with its famous "bouchons" and great cooking chefs.

Because many people arrive each year in Lyon and his metropolis which population is growing faster than the French average, find a neighborhood where to live according to your way of life can be useful. And data may help to answer to this question.

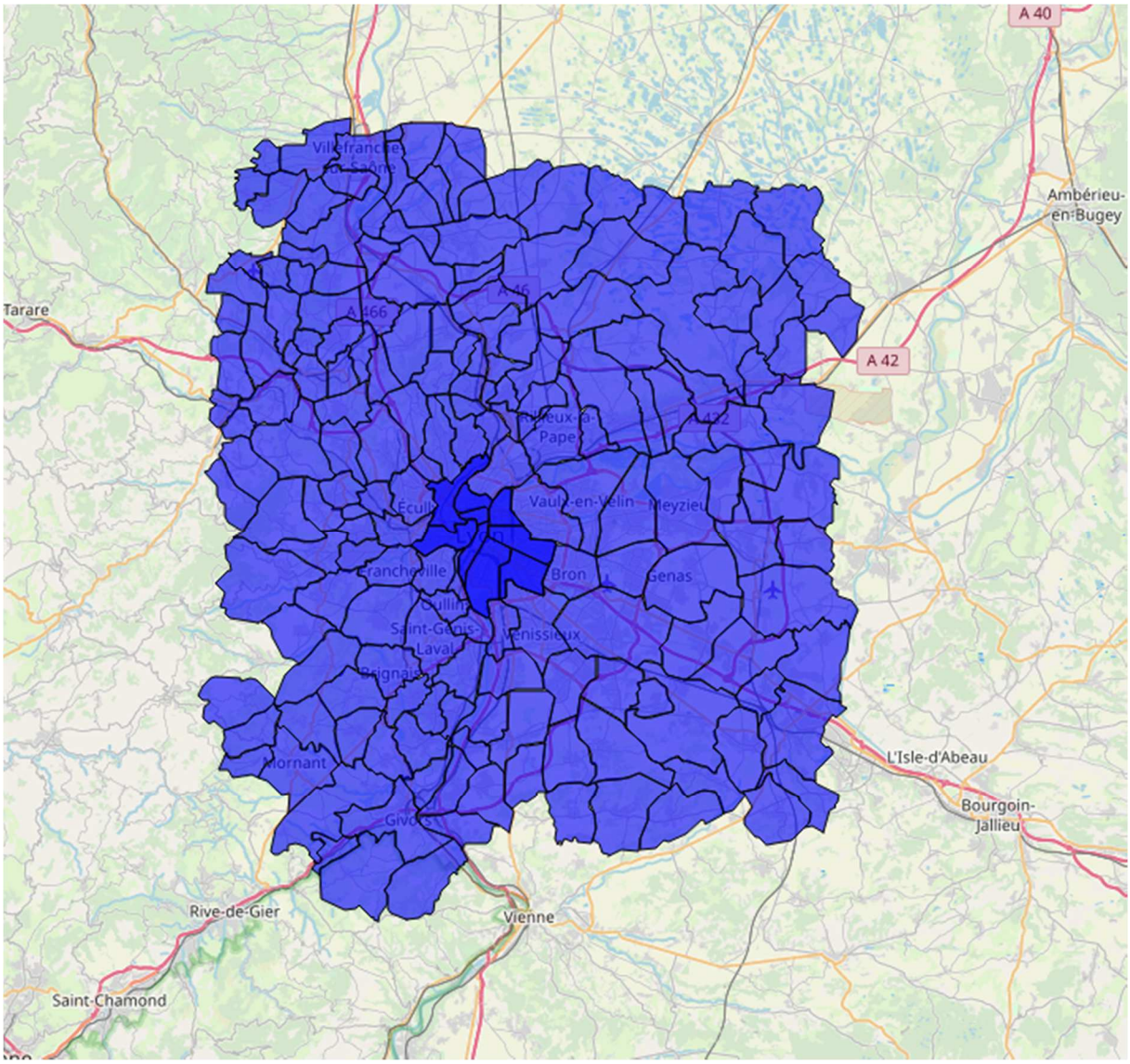
2. Description of the data and how it will be used to solve the problem.

First we have to identify the neighborhoods of the metropolis of Lyon. On the open platform for French public data: data.gouv.fr we can search for appropriate data including geo-data.

We investigate and find three kind of sources which can be our neighborhoods on a subdivision of the platform about Lyon: data.grandlyon.com. The first are the administrative districts of the town of Lyon. It is not good because it doesn't cover the metropolis and also, the administrative districts (called "arrondissement" in French) are much larger than the idea of a neighborhood. The second are the neighborhoods of the metropolis "Quartiers des communes de la Métropole de Lyon". You can see below the map.



We note that not all neighborhoods are represented, and some are not contiuous. So we look for the third source which are the cities of the metropolis, in the sense of the French administration called a “commune”. You can see below the map of the cities of the metropolis of Lyon and in darker blue the districts of Lyon mentioned above.



Conversely, this perimeter is much too wide for our purpose. So we decide to keep the second source “Quartiers des communes de la Métropole de Lyon” which is the closest to our needs.
With more time on it, it will be appropriate to complete those data with the missing neighborhoods, but they are not the main places of the metropolis of Lyon therefore it will not be so limiting for us.

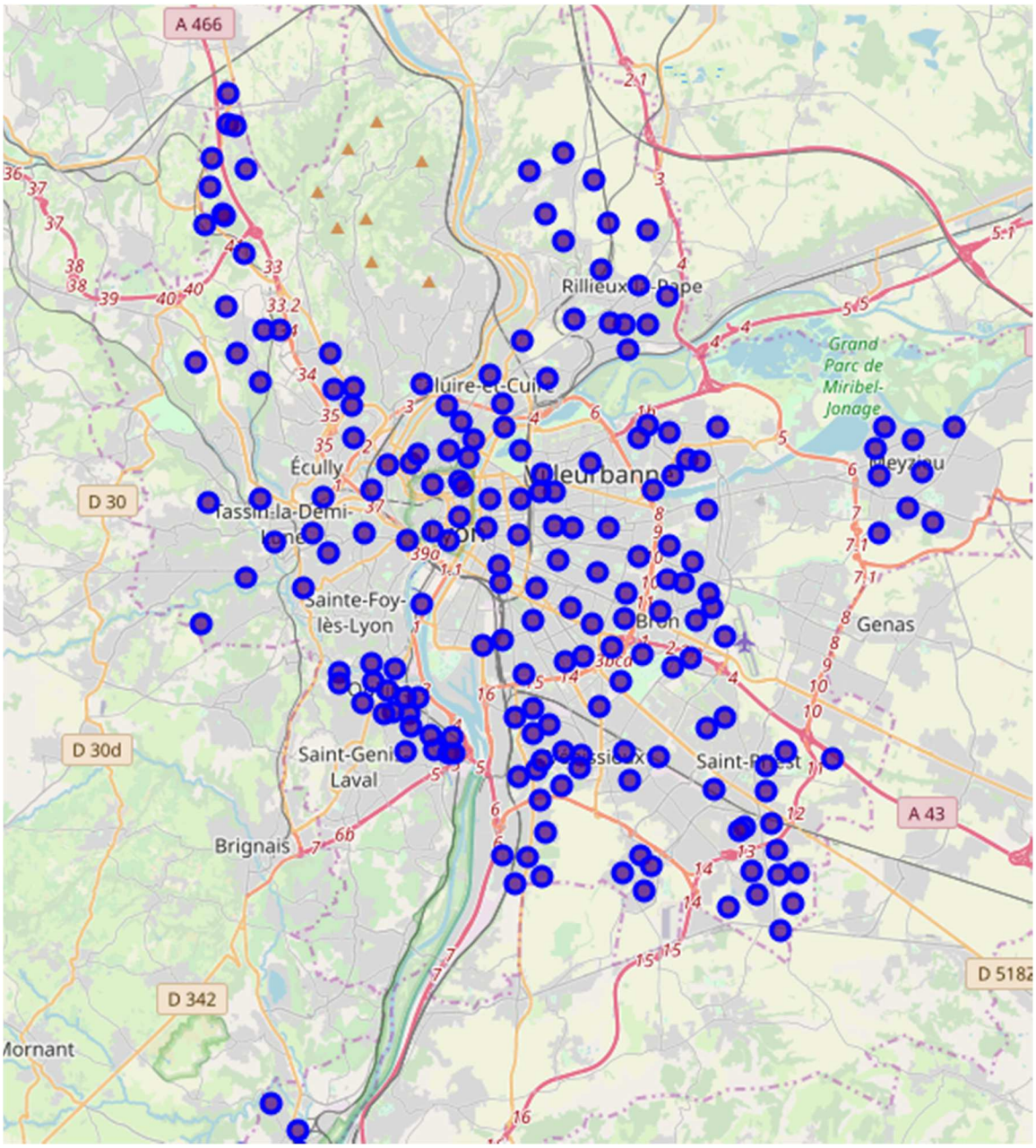
FOURSQUARE

We also use the data of foursquare to get the different venues of the neighborhoods we identify based on what we study previously in this course.

We also want to add some data that are not in the foursquare venues. This are the location of schools (primary, secondary and high schools) and hospitals. We find those

geo-data on the same platform data.gouv.fr, three sources for the schools (one for each kind of school) and one for hospital.

The data for the neighborhoods give us the coordinates of the limit and not a latitude and longitude, so we calculate the centroid to get the lat and long of each neighborhoods as you can see on the map below.



Based on those data, we can run Foursquare API to get the venues of each neighborhoods in a radius of 1000 meters. After that we concatenate to it the schools and hospital in the same radius to complete our data.

We can then identify the five or ten most common venues (including schools and hospitals) of each neighborhood as for example below.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Bussière	Restaurant	PrimarySchool	Pharmacy	FoodStore	Flea Market
1	Cadière	Harbor / Marina	Supermarket	Shopping	Services	Restaurant
2	Centre-ville	PrimarySchool	Theater	TakeAway	SecondarySchool	FoodStore
3	Chassagnes	HighSchool	Park	Sport	SecondarySchool	Pool
4	Clavelière	HighSchool	Pharmacy	Shopping	Services	SecondarySchool

Some neighborhoods will not have so many venues even not at all, so we will drop them.

This is what we will use to answer to the initial question. The next section will explain the methodology.

3. Methodology.

Because we have data that are not labelled, we use a method for classification which can work in this case: data clustering. And moreover, we chose the K-Means method.

We will create clusters based on the venues and public establishment which are in a given radius of the neighborhoods. We don't really know what the radius value must be, so we will try different one from 2000 meters to 250 meters (2000 – 1500 – 1000 – 750 – 500 – 250).

We also don't know what the best number of clusters for clustering is. We will use the elbow method to choose it.

- Step 1: we prepare the data for clustering.

  First, we get the public data from the platform data.gouv.fr and create dataframe for each radius to determine the public establishments in the area of each neighborhoods. We get 6 dataframe.

  After that, we get the foursquare data. We want to have the venues of each neighborhoods in a given radius from 2000 to 250 meters. To avoid too many requests to foursquare (because we are limiting), we will first request for the radius of 2000 meters. We create the dataframe "lyon_venues_2000" with those data. Based on this (which contain the latitude and longitude of each neighborhoods and venues) we can calculate the five dataframe for the other radius chosen.

  For memory, the calculation of the distance of two points on a sphere is given by the formula of haversine :

$$2r\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

  Where r is the radius of the sphere, for the Earth we take 6367445 meters,
  φ is the latitude and λ the longitude of the points expressed in radian.

  Next we create a top category for the foursquare data because the native category we get is too large. We have done this in a simple excel file and have imported it in the notebook for use: typevenue.csv. We add this "venue type" to the data.

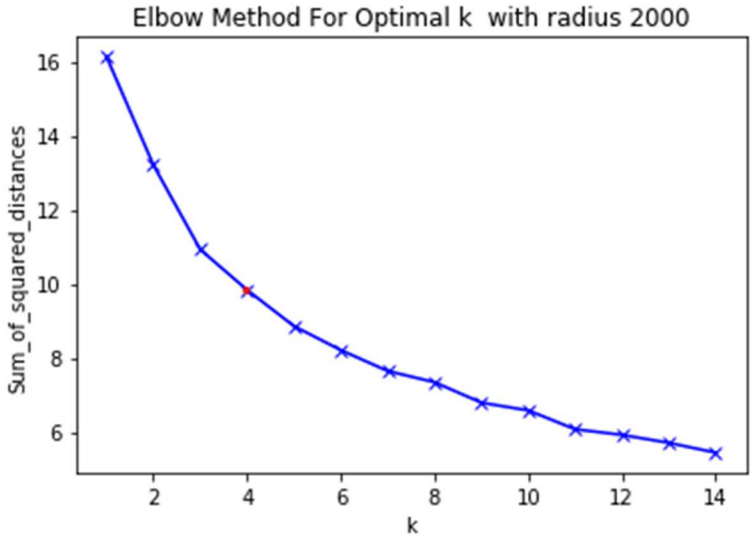
After that we clean the foursquare data because we identify some venue categories that are not real venue for us (for example a venue with the category “neighborhood”). Those venues are put in the venue type “A supprimer” wich is the french for “to delete”.

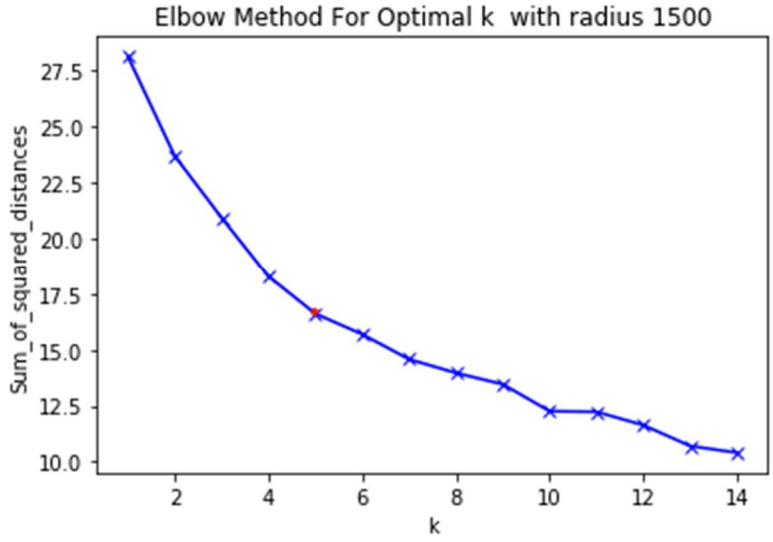
We then create the dataframe for the clustering using the function *get_dummies* previously used on the venues data. Before finishing our preparation by grouping the neighborhoods by mean of the number of venues, we aggregate the education and health data we get on the public platform with the foursquare one.

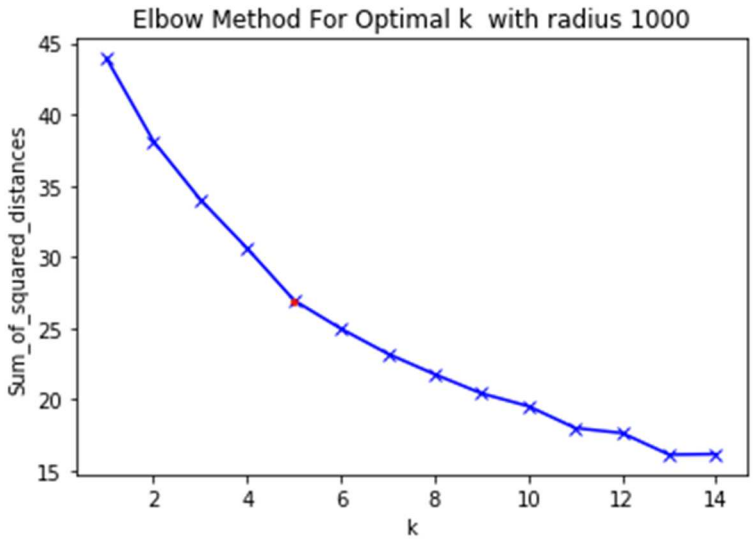
- Step 2: Clustering.

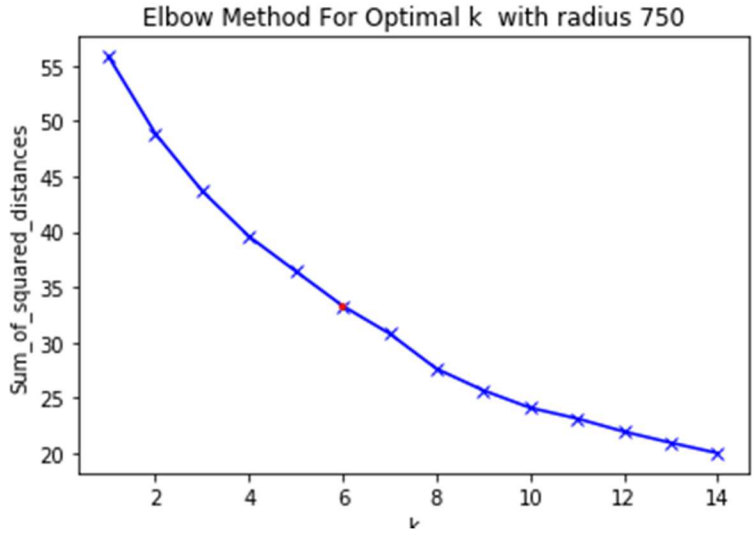
  We use the K-Means clustering method for different number of clusters and for each set of data from different radius. As you can see below, we obtain a visualization of the results for each set by looking the sum of squared distances against the number of clusters.

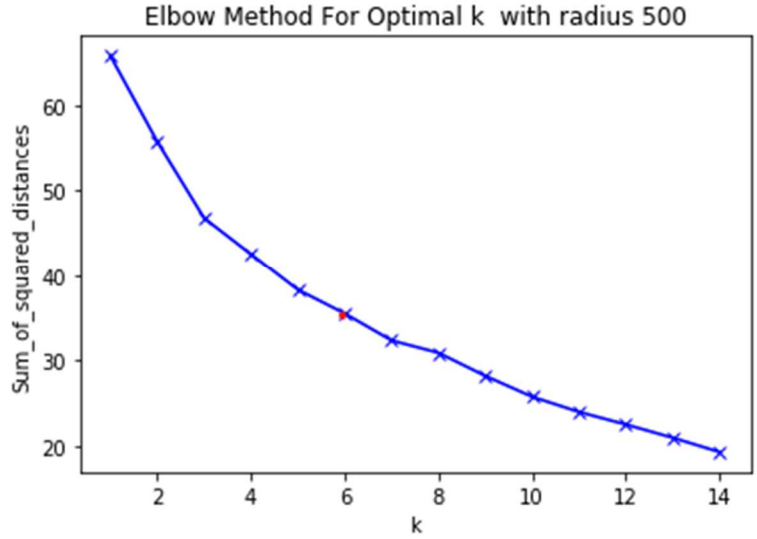
  We put in this report a red point to mark the number of clusters choosed for each radius studied.

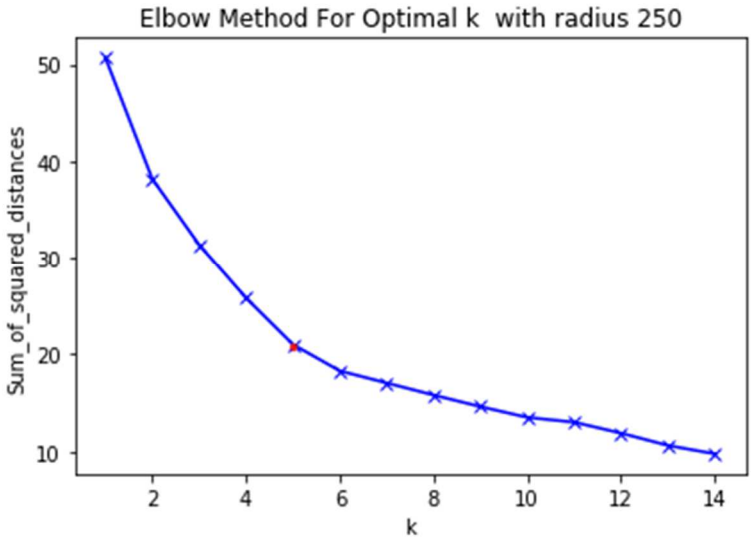





Elbow Method For Optimal k with radius 1000
Sum_of_squared_distances
k
45
40
35
30
25
20
15
2
4
6
8
10
12
14

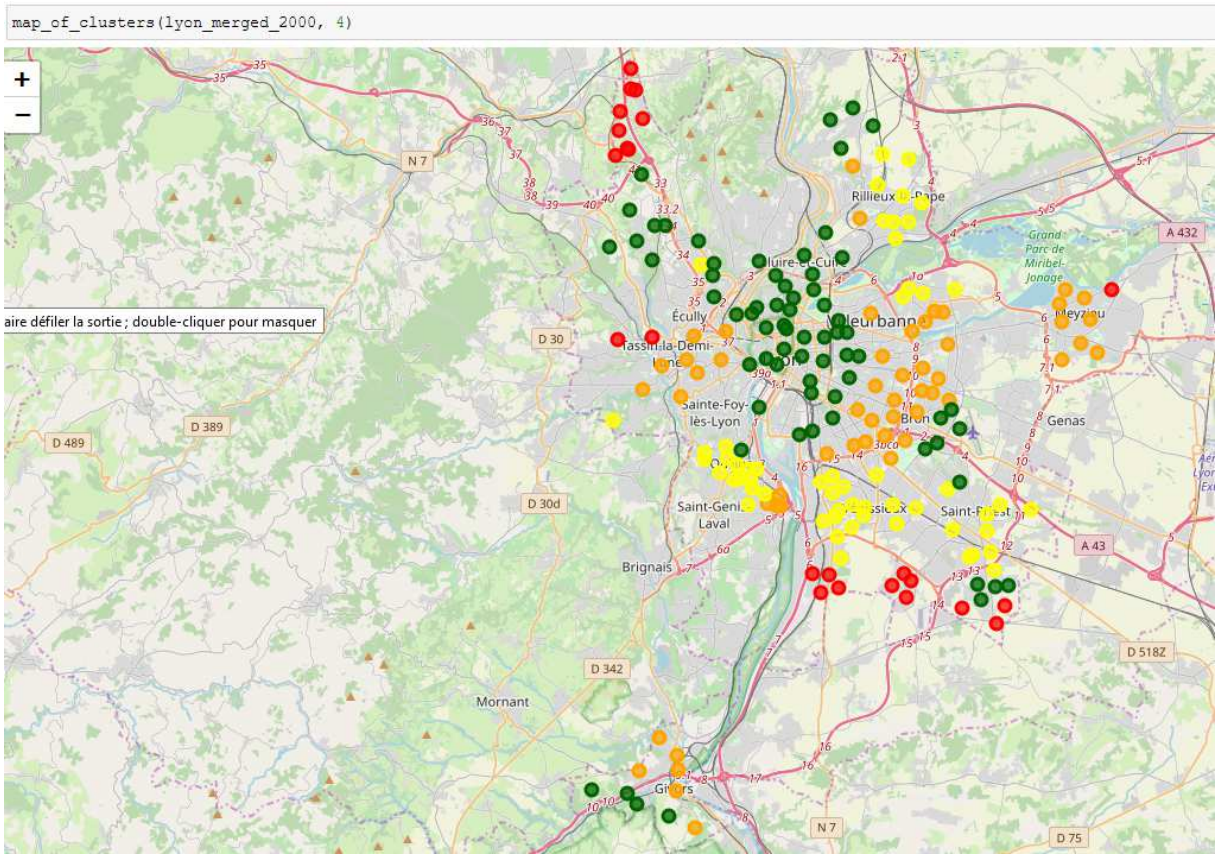

Elbow Method For Optimal k with radius 750
Sum_of_squared_distances
55
50
45
40
35
30
25
20
2
4
6
8
10
12
14


Elbow Method For Optimal k with radius 500
Sum_of_squared_distances
k
60
50
40
30
20
2
4
6
8
10
12
14

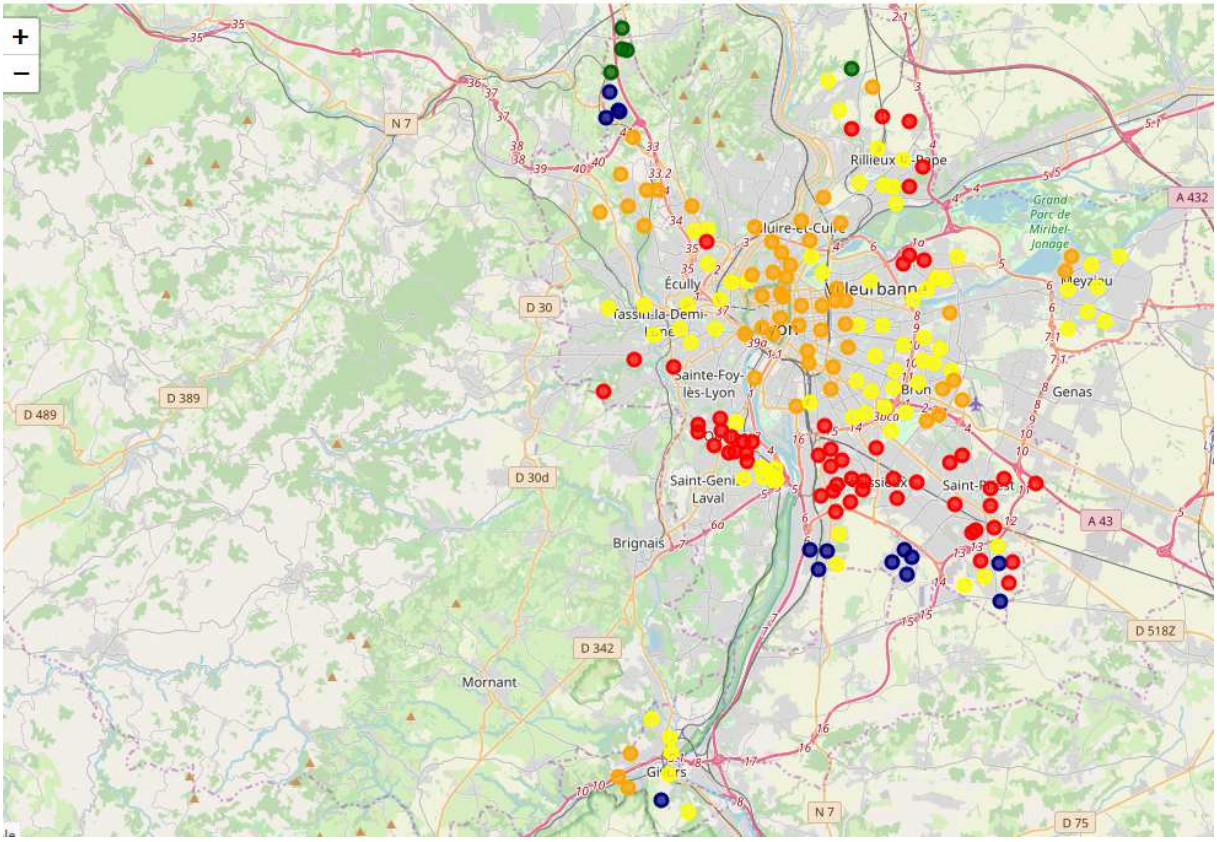


Finally, we do the clustering for each radius with the chosen number and visualize the result on a folium map.

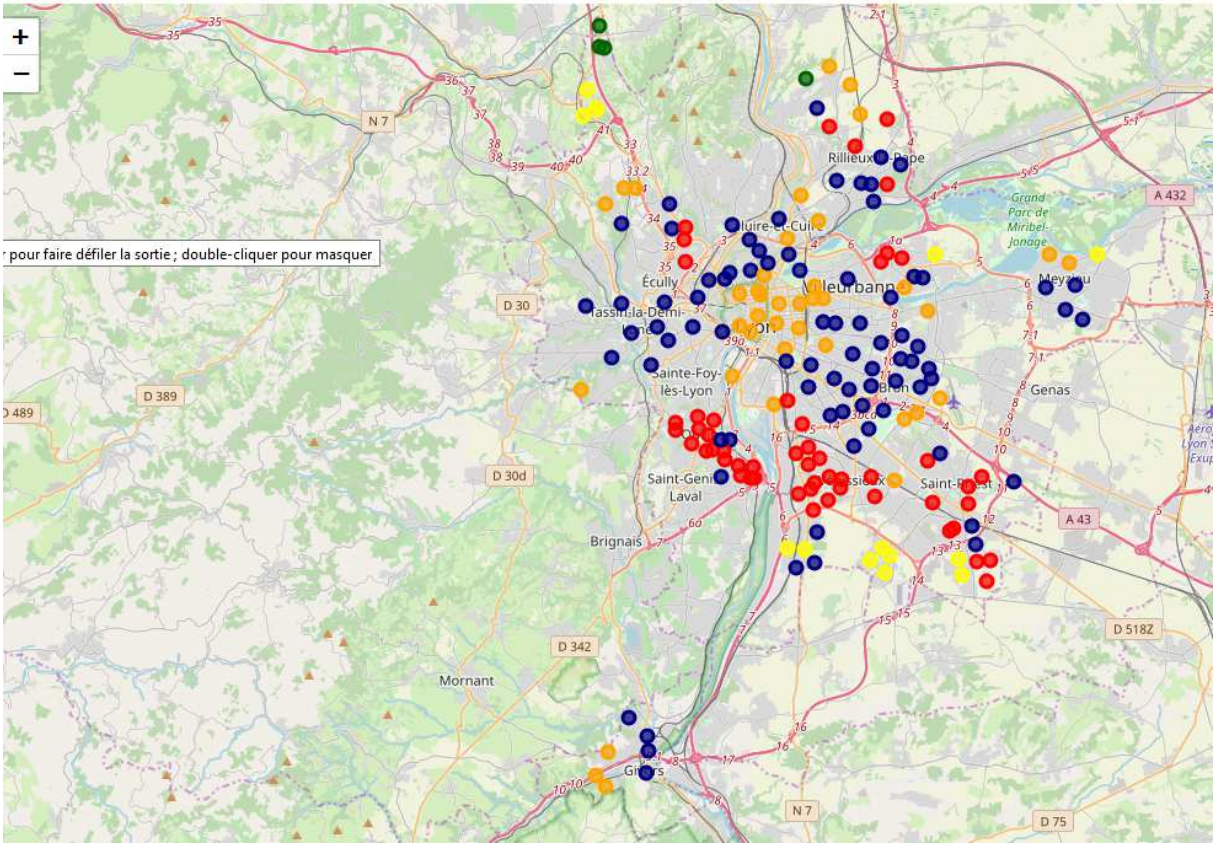
Radius	Number of clusters
2000	4
1000	5
1500	5
750	6
500	6
250	5



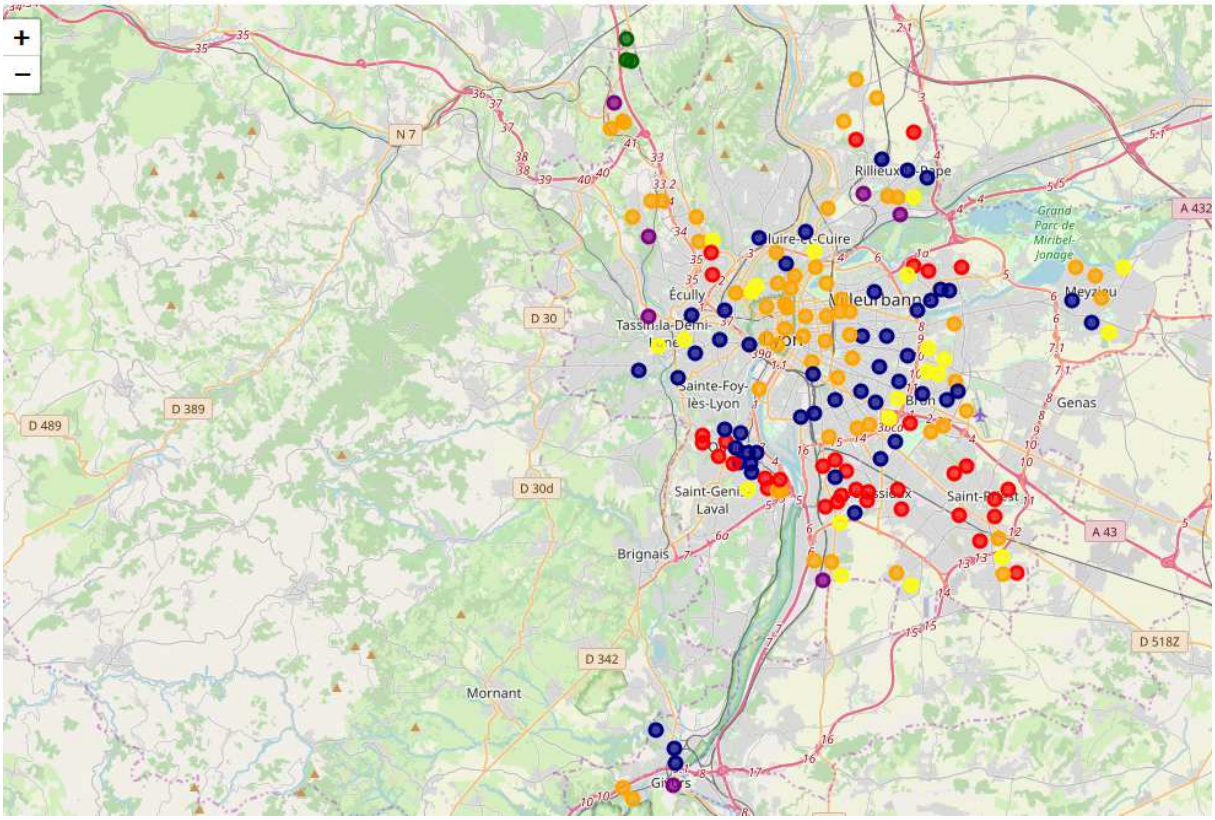
```
map_of_clusters(lyon_merged_1500, 5)
```



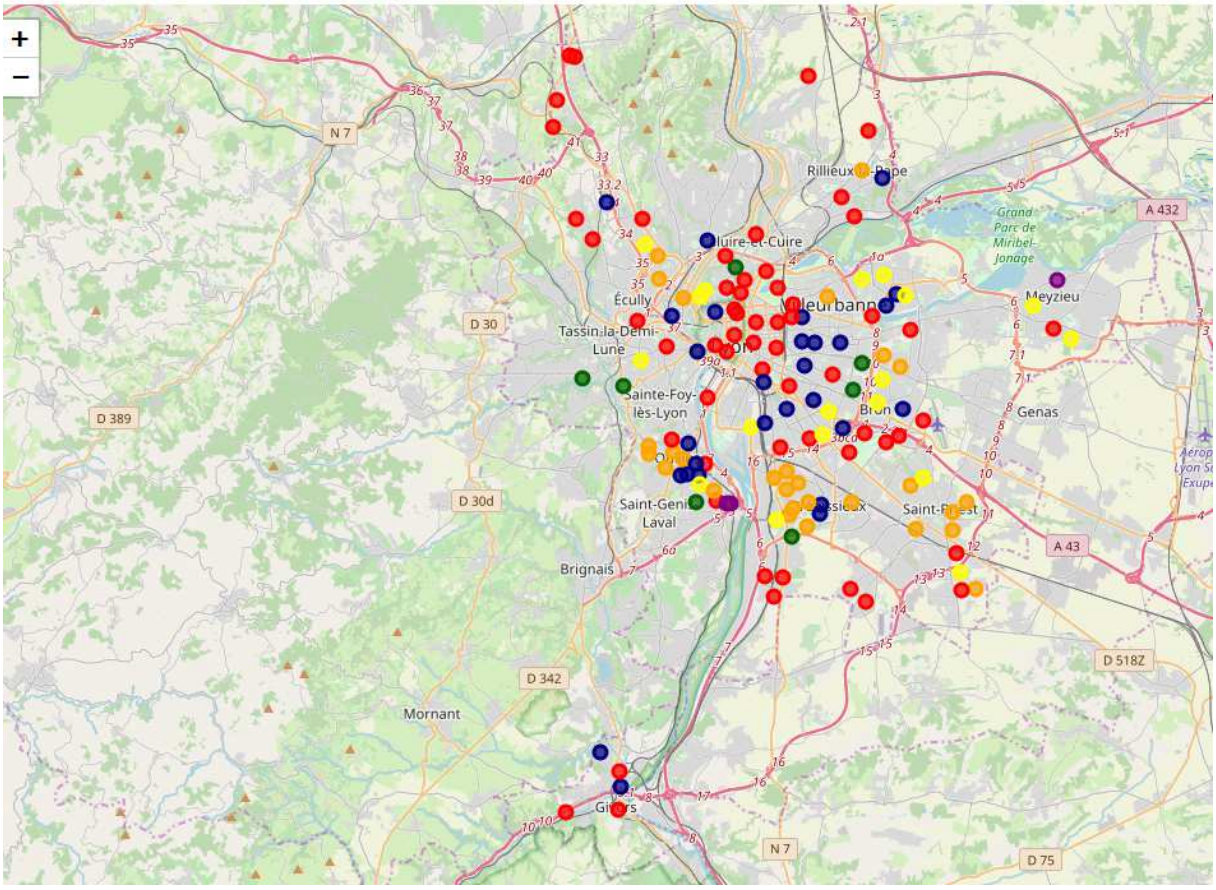

```
map_of_clusters(lyon_merged_1000, 5)
```



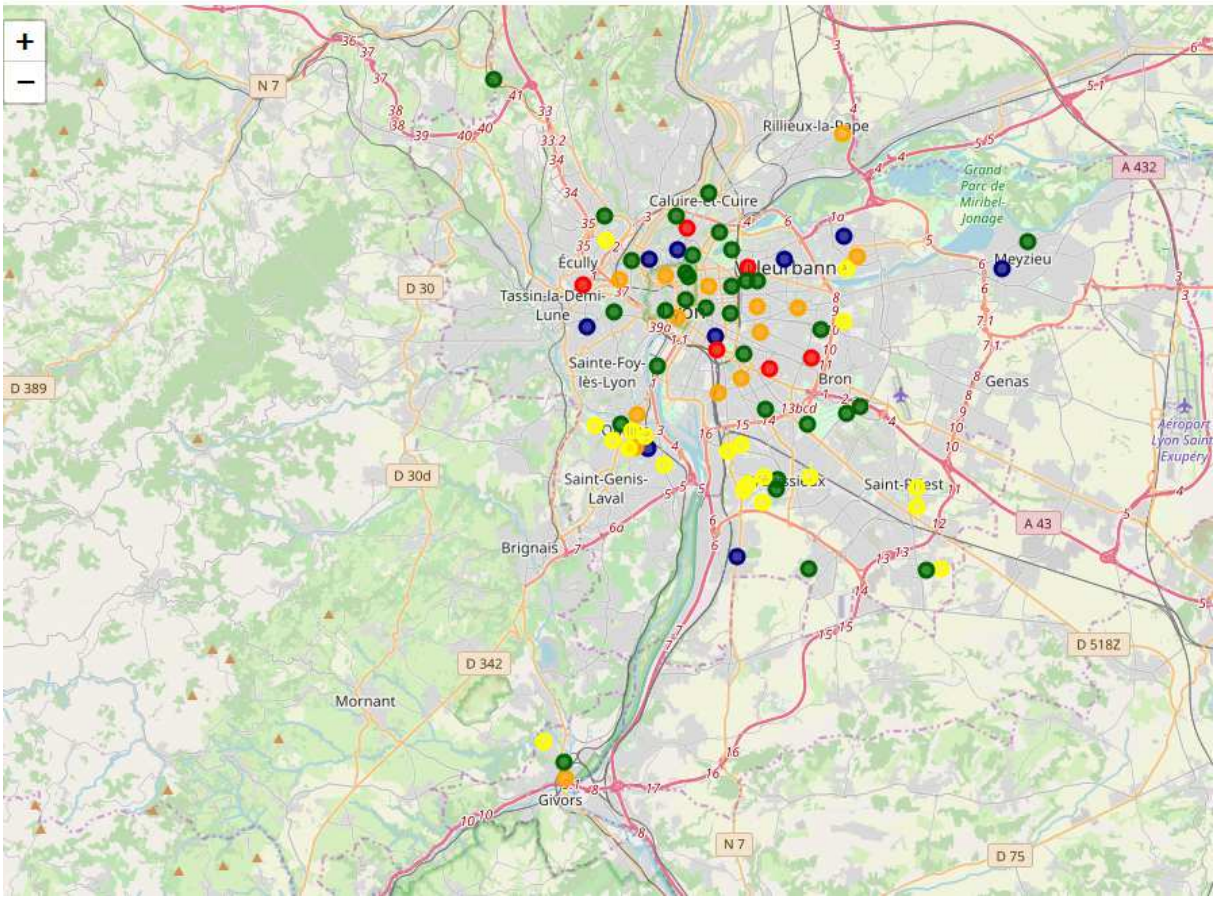
```
map_of_clusters(lyon_merged_750, 6)
```




```
map_of_clusters(lyon_merged_500, 6)
```



```
map_of_clusters(lyon_merged_250, 5)
```



4. Results.

We see there are some similarities in the different clusters we obtain for the different chosen radius, especially for high radius.

Logically, while the radius decrease, the number of venues decrease also. We note also that with low radius we get some clusters with a very small number of venues.

Now, according to your needs and preferences, we could propose to you different neighborhoods where to live.

To go further, we look which venues are most frequent in the different clusters we get for each set. We restitute many barplots for each set (using the corresponding color used in the maps) for the venues whose total number is greater, between and less than given numbers as shown in the tables below. All of them can be seen in the notebook.

Radius 2000

Barplots	less than	more than
1		300
2	300	100
3	100	40
4	40	20
5	20	10
6	10	

Radius 1500

Barplots	less than	more than
1		150
2	150	50
3	50	15
4	15	10
5	10	5
6	5	

Radius 1000

Barplots	less than	more than
1		100
2	100	50
3	50	20
4	20	10
5	10	5
6	5	

Radius 750

Barplots	less than	more than
1		50
2	50	20
3	20	8
4	8	3
5	3	

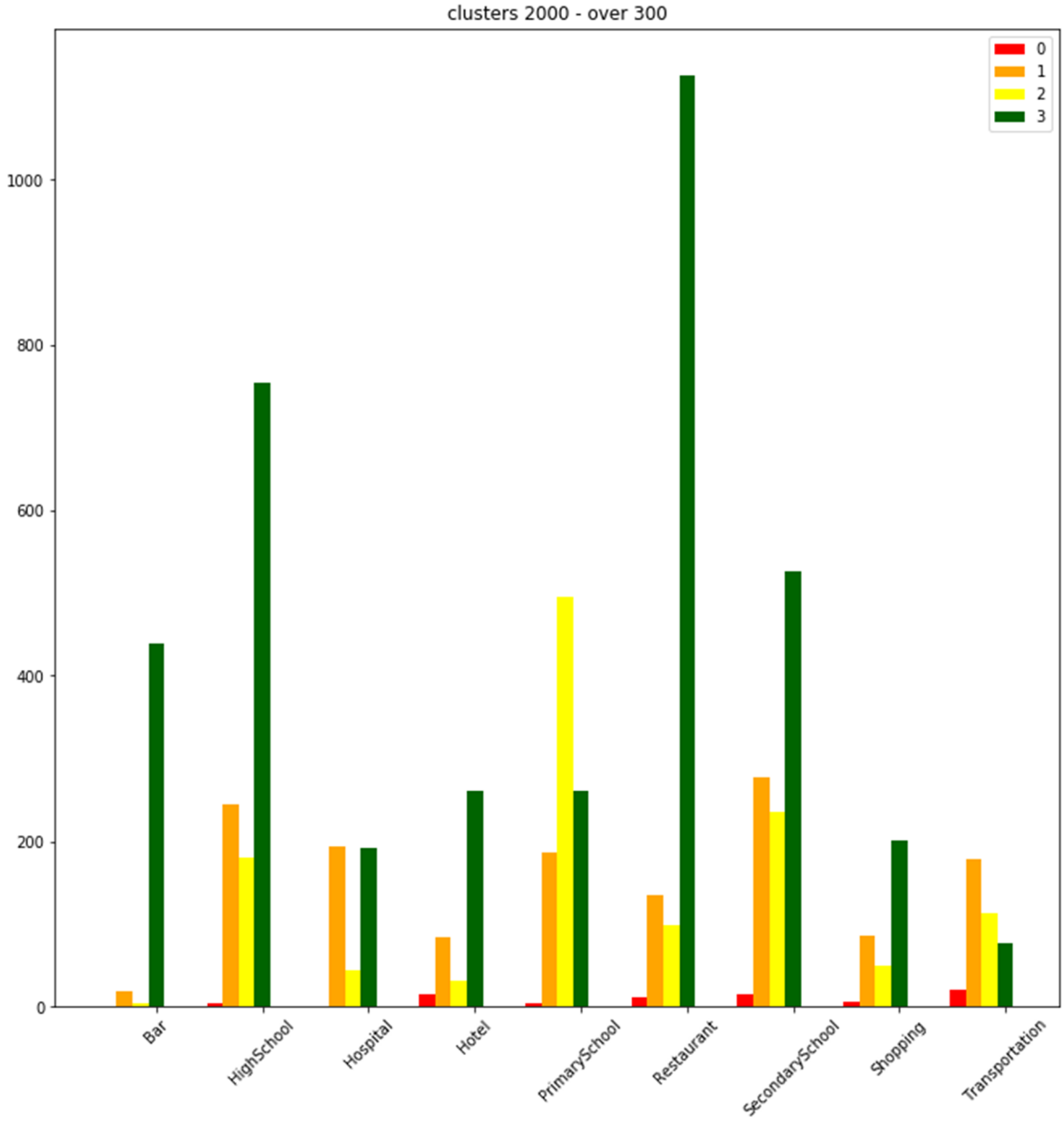
Radius 500

Barplots	less than	more than
1		25
2	25	12
3	12	4
4	4	2
5	2	

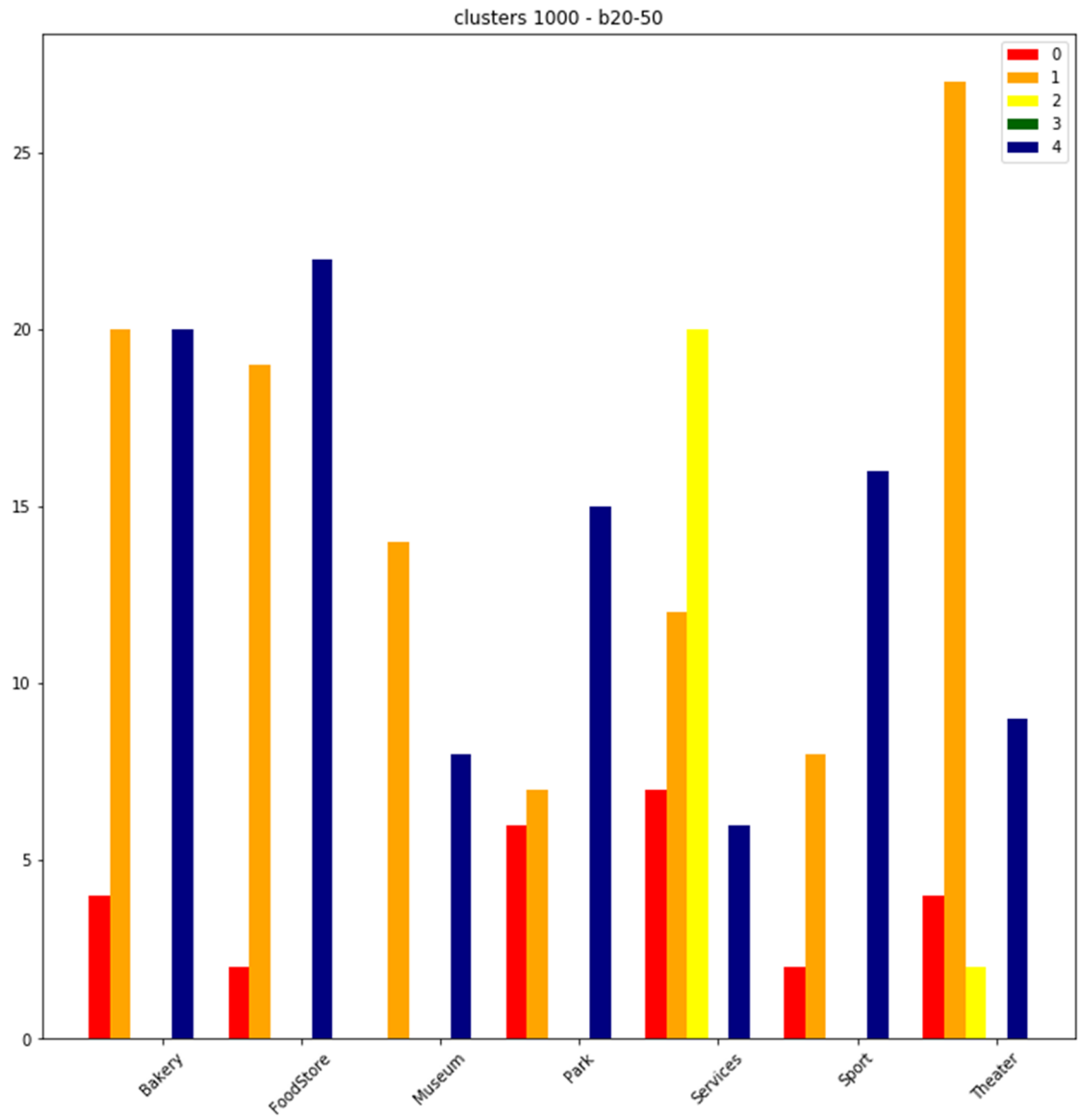
Radius 250

Barplots	less than	more than
1		8
2	8	3
3	3	2
4	2	

For example, this the barplot of the number of venues for each cluster of the set radius 2000 and the total number of venues above 300.



Another example for radius 1000 and total number between 20 and 50.



The choose of the radius restrained should be associate to the preferences of people, for example if you prefer walk around your neighborhood for shopping, the radius 250 or 500 will better discriminate the different neighborhood. But if you like driving for getting around a large radius would be better.

5. Discussion.

The results obtained allow us to give a first answer to our initial question. We note, however, that this answer should be refined. On the one hand, additional information on neighborhoods, for example rental prices or environmental data, would be useful to better respond. On the other hand, it would be better to think about how to collect the needs of the users to better meet them. This could be done using a survey for example and it would identify the missing data to be included in our modeling.

6. Conclusion.

In this notebook we have used some data to answer a simple question, using a clustering method. What we obtained is modest and shows that a more in-depth but much more important work would be necessary to develop a professional solution. On the one hand by widening the scope of the data used. On the other hand, this solution could, for example, include upstream a survey to collect needs and downstream a scoring system to determine the best neighborhoods to offer. In addition, thinking about how to integrate user feedback in order to improve the service could be useful. Finally, before developing this solution, a market study would be necessary.