IBM Data Science Capstone

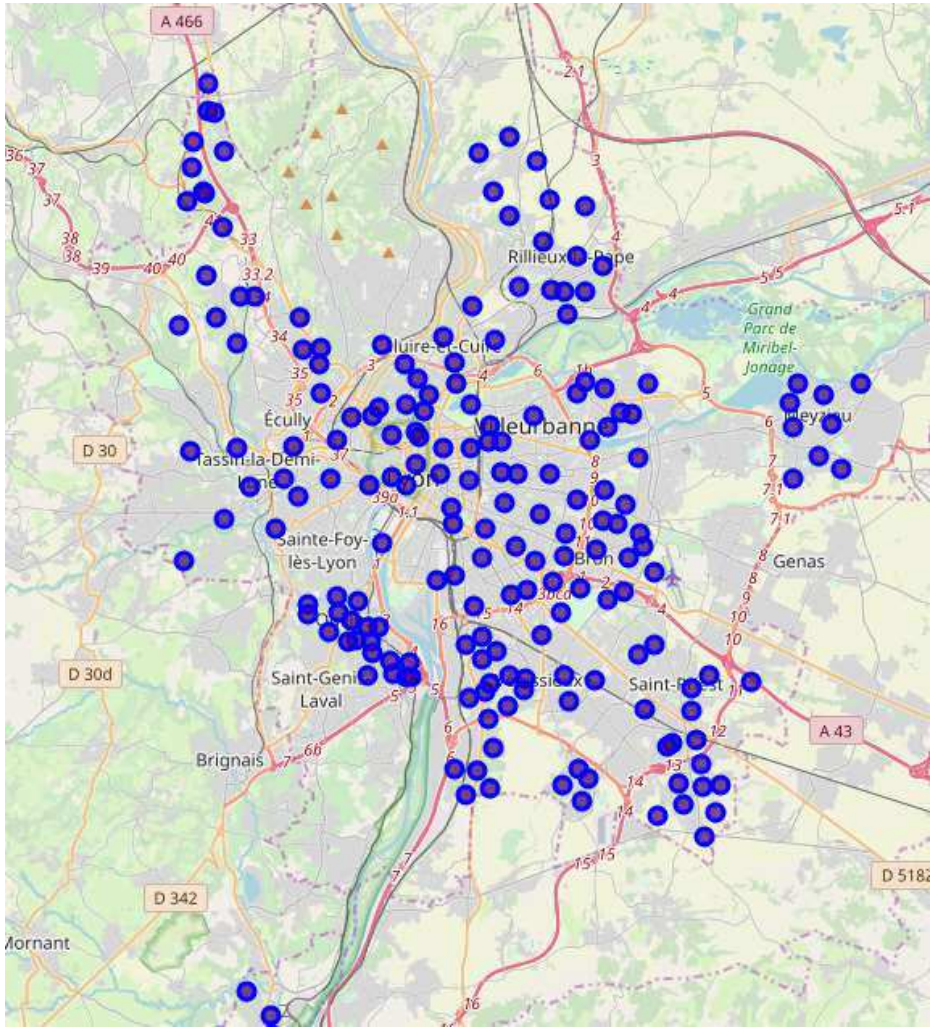
The Battle of Neighborhoods – Final Report

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 and to your preferences you will search for a neighborhood which correspond the best to you.
- 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.

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

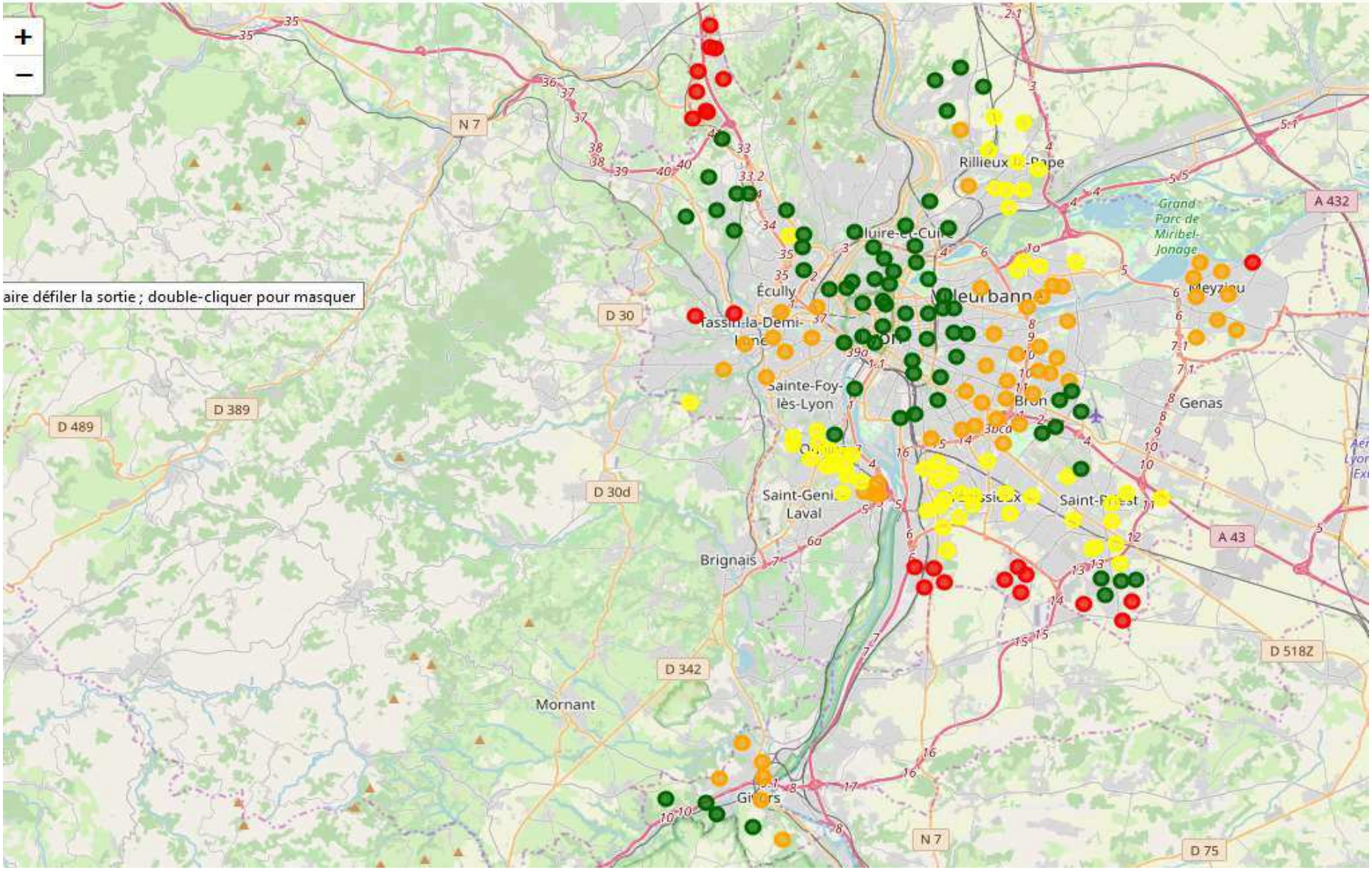
- First we have to identify the neighborhoods of the metropolis of Lyon. On the open platform for French public data: data.gouv.fr we can search for appropriate data including geo-data.
- We also use the data of foursquare to get the different venues of the neighborhoods we identify based on what we 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 can then identify the five or ten most common venues (including schools and hospitals) of each neighborhood as for example below.


A 466
Rillieux-la-Pape
Grand Parc de Miribel-Jonage
Écully
Tassin-la-Demi-Lune
Sainte-Foy-lès-Lyon
Genas
D 30
D 30d
Brignais
A 43
D 342
Mornant

Methodology.

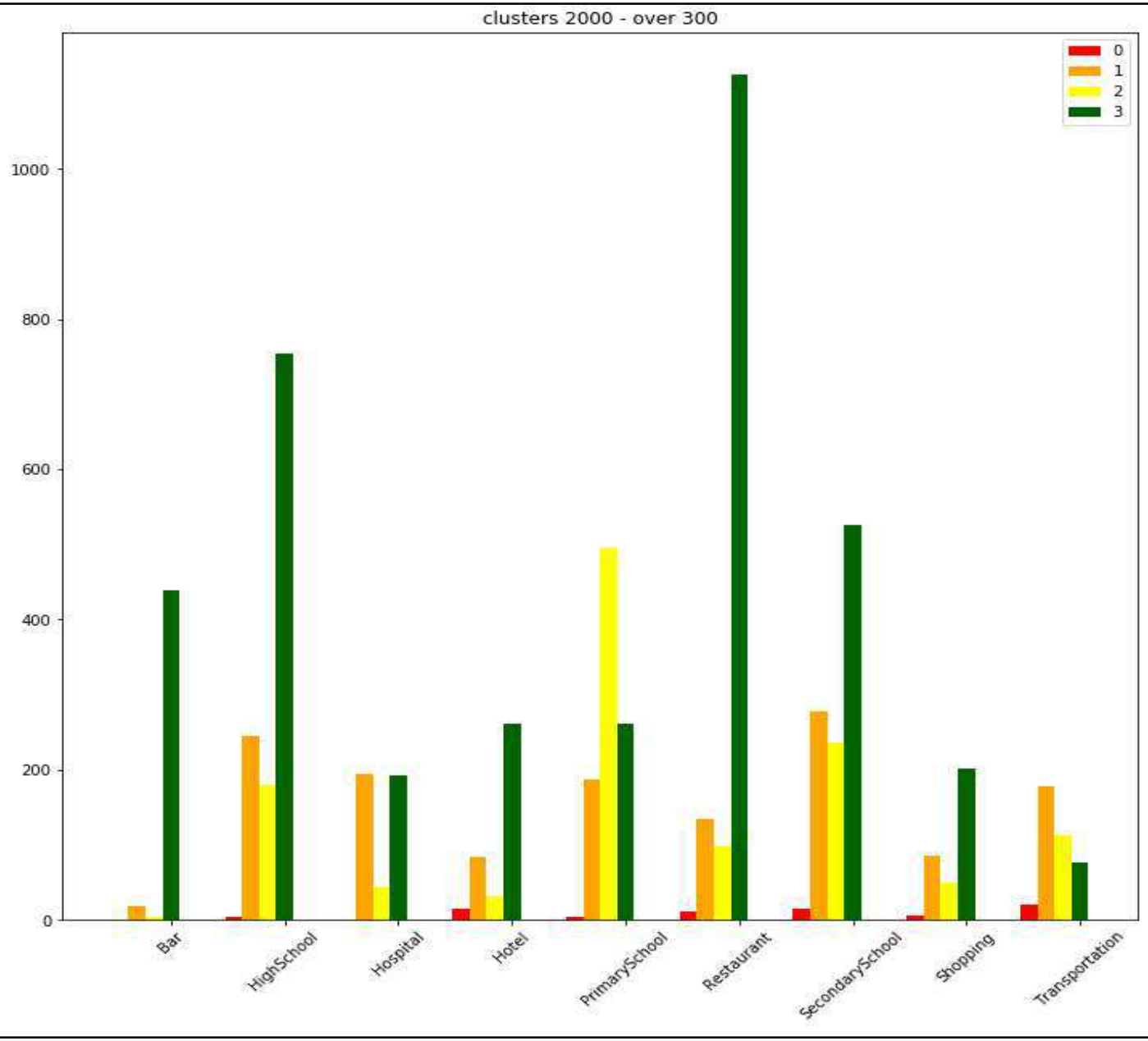
- Because we have data that are not labelled, we use a method for classification which can work in this case: data clustering. And moreover, we chose the K-Means method.
- We will create clusters based on the venues and public establishment which are in a given radius of the neighborhoods. We don’t really 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).

```
map_of_clusters(lyon_merged_2000, 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.


clusters 2000 - over 300
0
1
2
3
1000
800
600
400
200
0
Bar
HighSchool
Hospital
Hotel
PrimarySchool
Restaurant
SecondarySchool
Shopping
Transportation

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.

Conclusion.

- 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.