coursera

IBM

Applied Data Science

The Battle of Neighborhoods:

Opening an Italian Restaurant in Paris

Objectives

We'll attempt to find the best suggestions of locations to open an Italian Restaurant in Paris.

Of course, this is no easy task and the final decision should be made after further on-site investigation, but with some data we can already reduce the area to check to a handful of locations

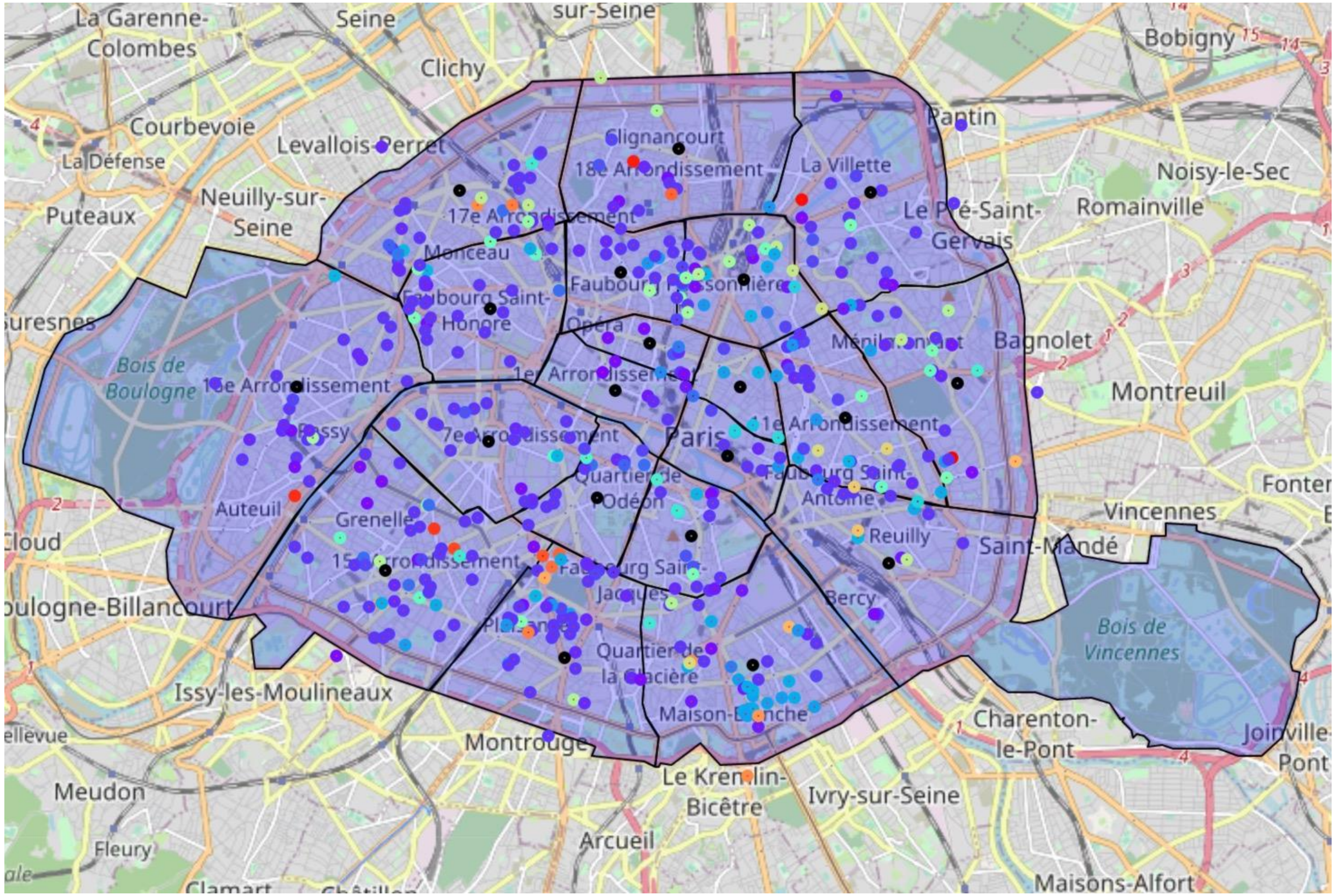
Assumptions:

- Areas in which low restaurant density will be avoided
- Neighborhoods where Italian Restaurants are among the most popular will be favored
- New italian restaurants should be as far as possible from the existing ones as clients might prefer the venues they are used to.
- Areas with many french restaurants will be favored, as they will mainly attract clients that would tend to enjoy italian food as well.

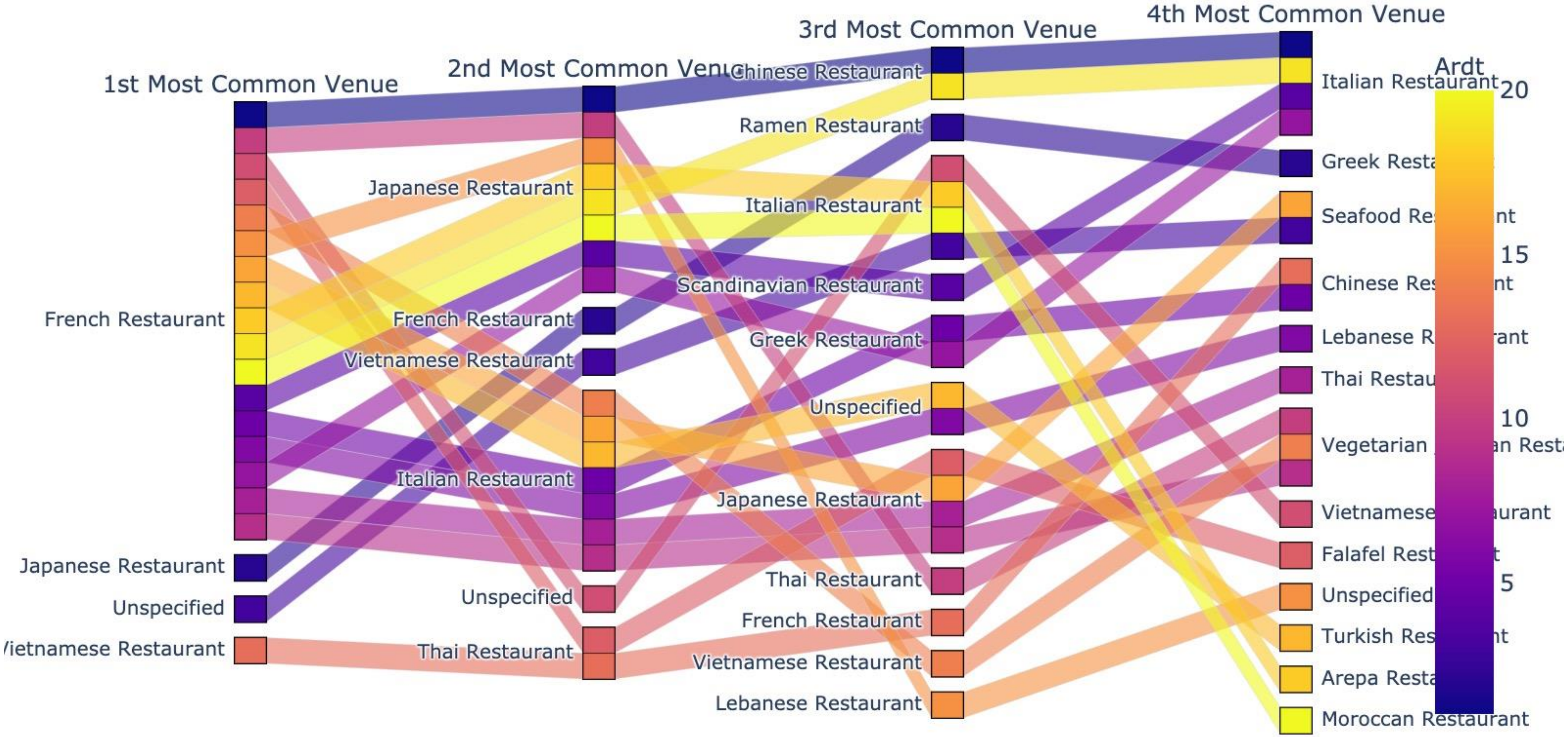
Data

- Shapes and geolocations of each Parisian *arrondissement* from **opendata.paris.fr**
- Venue info from the **FourSquare API**

map plotting all the analysed restaurants, by type >>

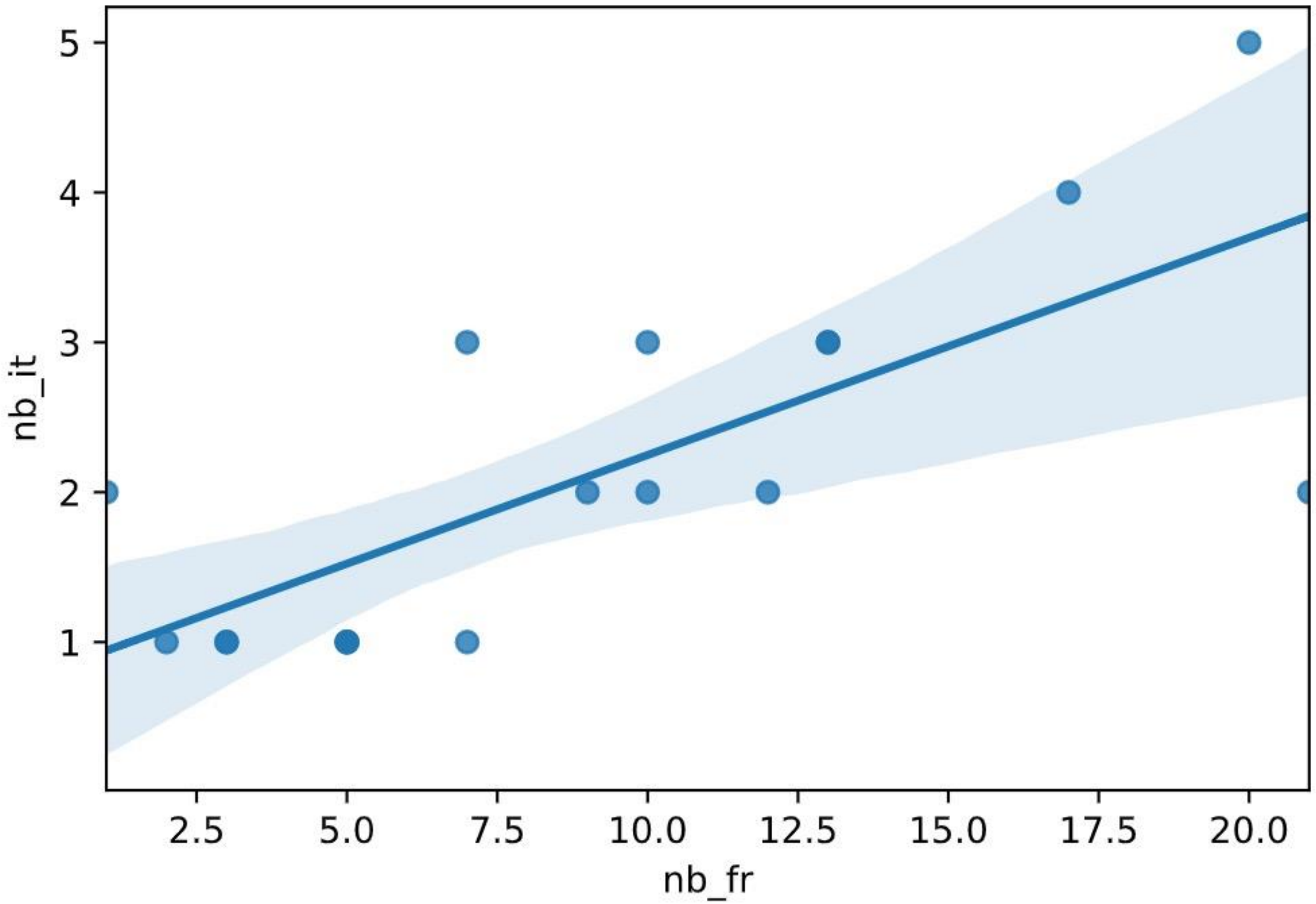


Different districts, different tastes



Verifying our assumptions

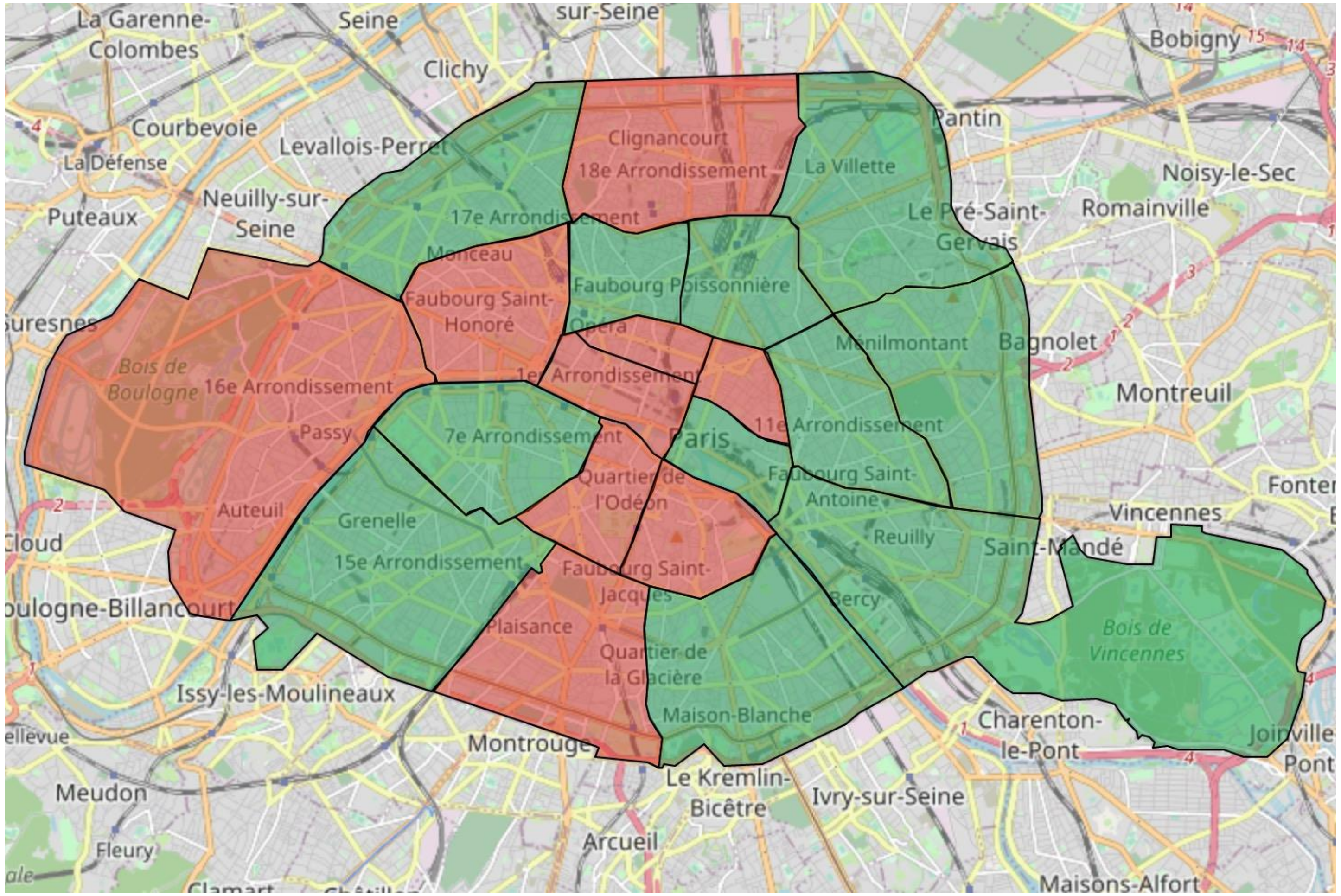
By plotting the number of french restaurants vs italian in each neighborhood, we find that there is a positive correlation between the number of french and italian venues



Identifying areas with less Italian restaurants than average

This allows us to identify which neighborhoods have less italian restaurants relative to the number of french restaurants.

Areas in green would therefore be better to open a new venue as the density of competitors will be lower there

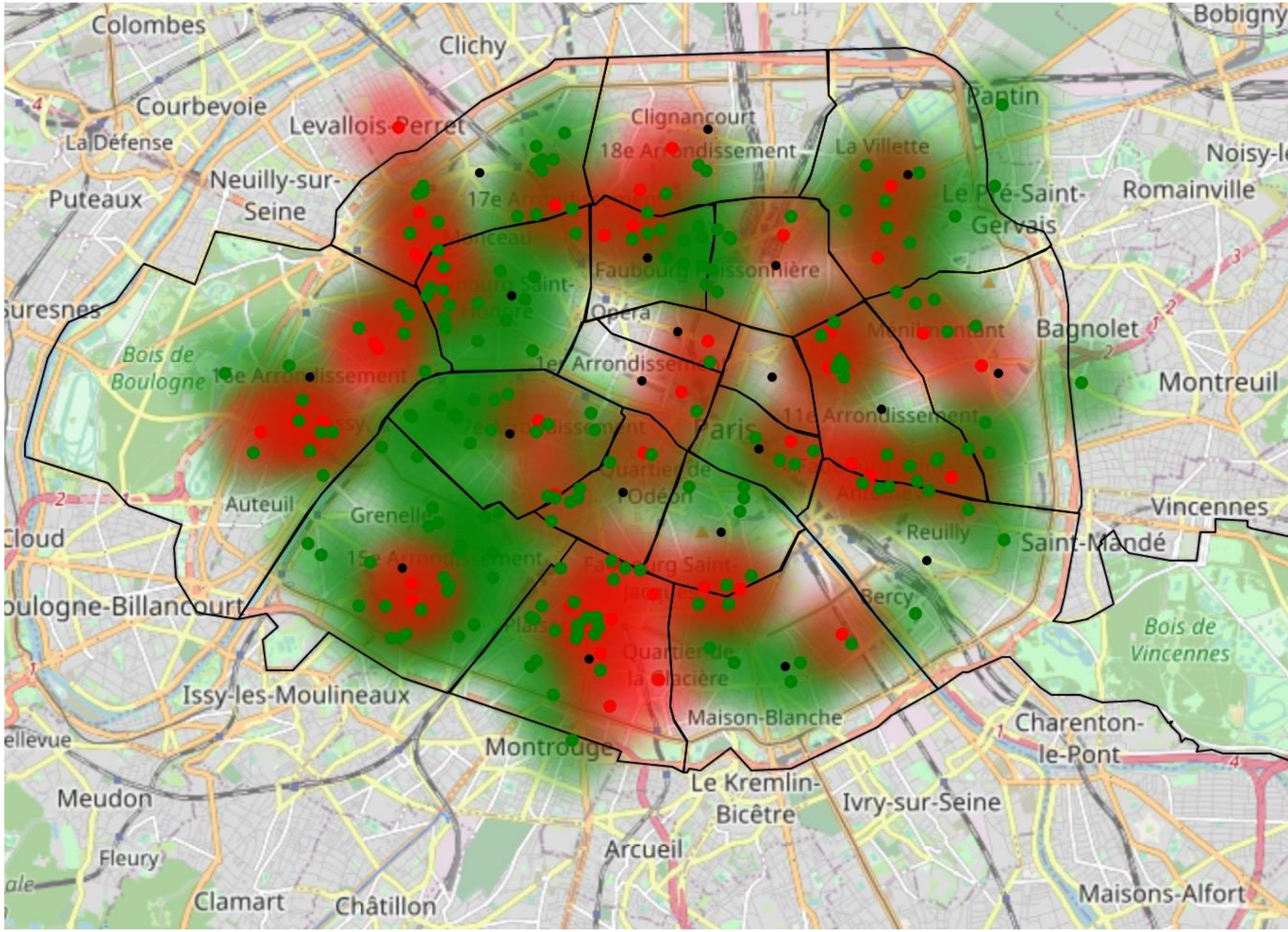


Locating areas to favour and the ones to avoid

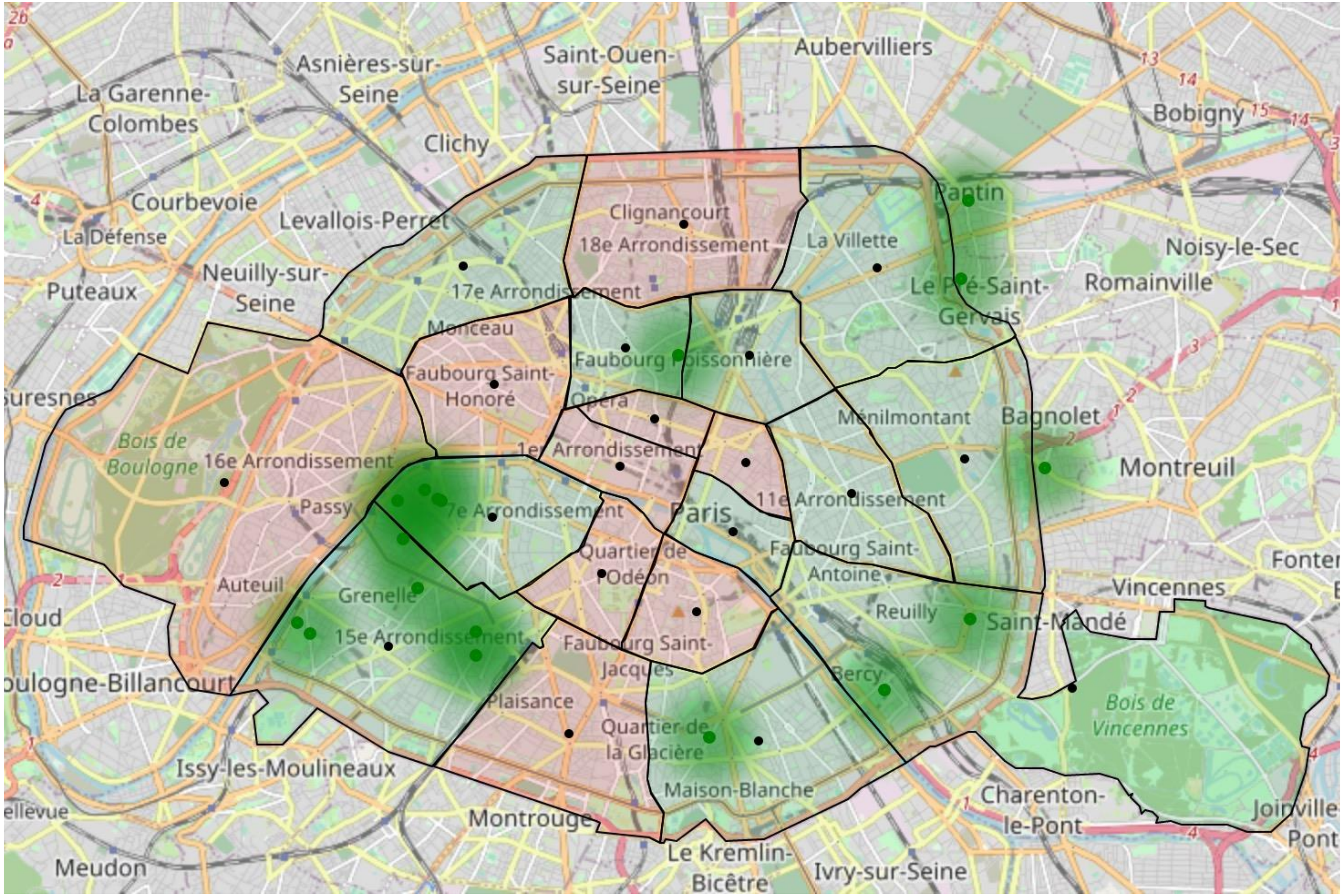
Here we can see the areas where the density of french and italian venues, overlapping.

The red areas (italian) are the ones to avoid.

The green areas (french) are the one we will favour



Isolating favourable areas

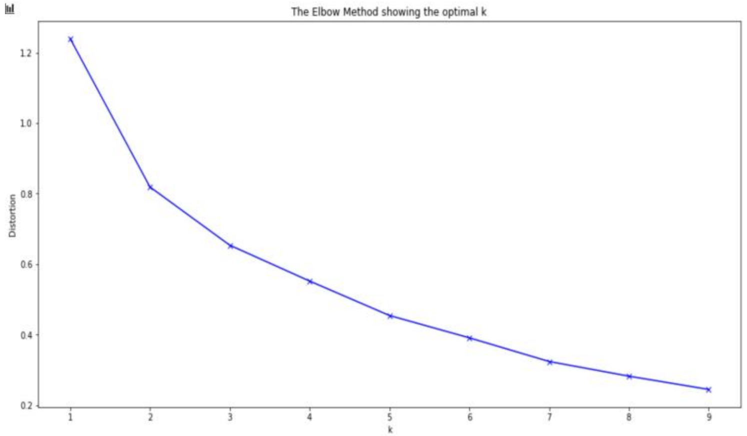


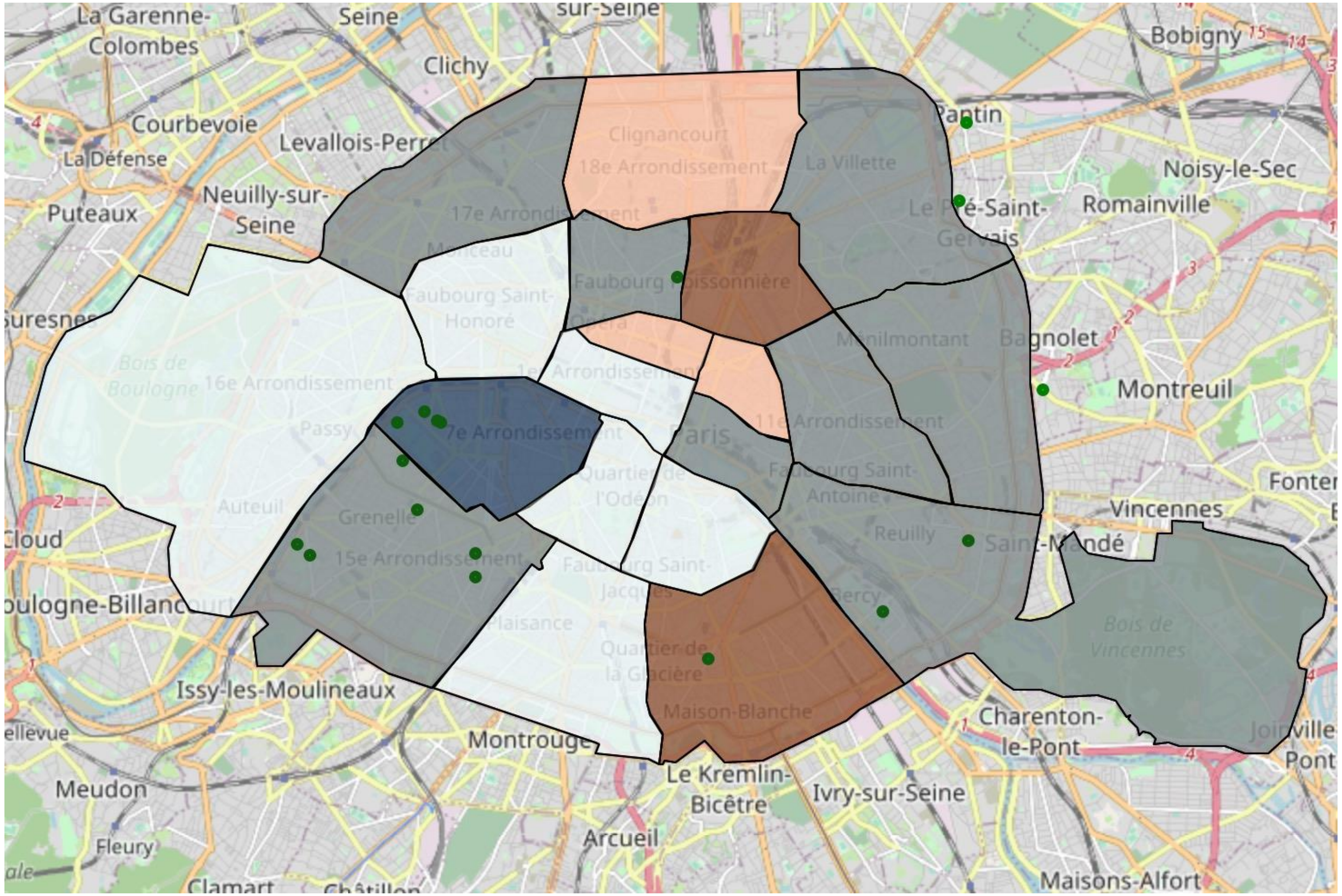
By super-imposing the two previous maps, we obtained this map >>

Cross checking findings using K-means

From the elbow test, below, the chosen value for k was **3**

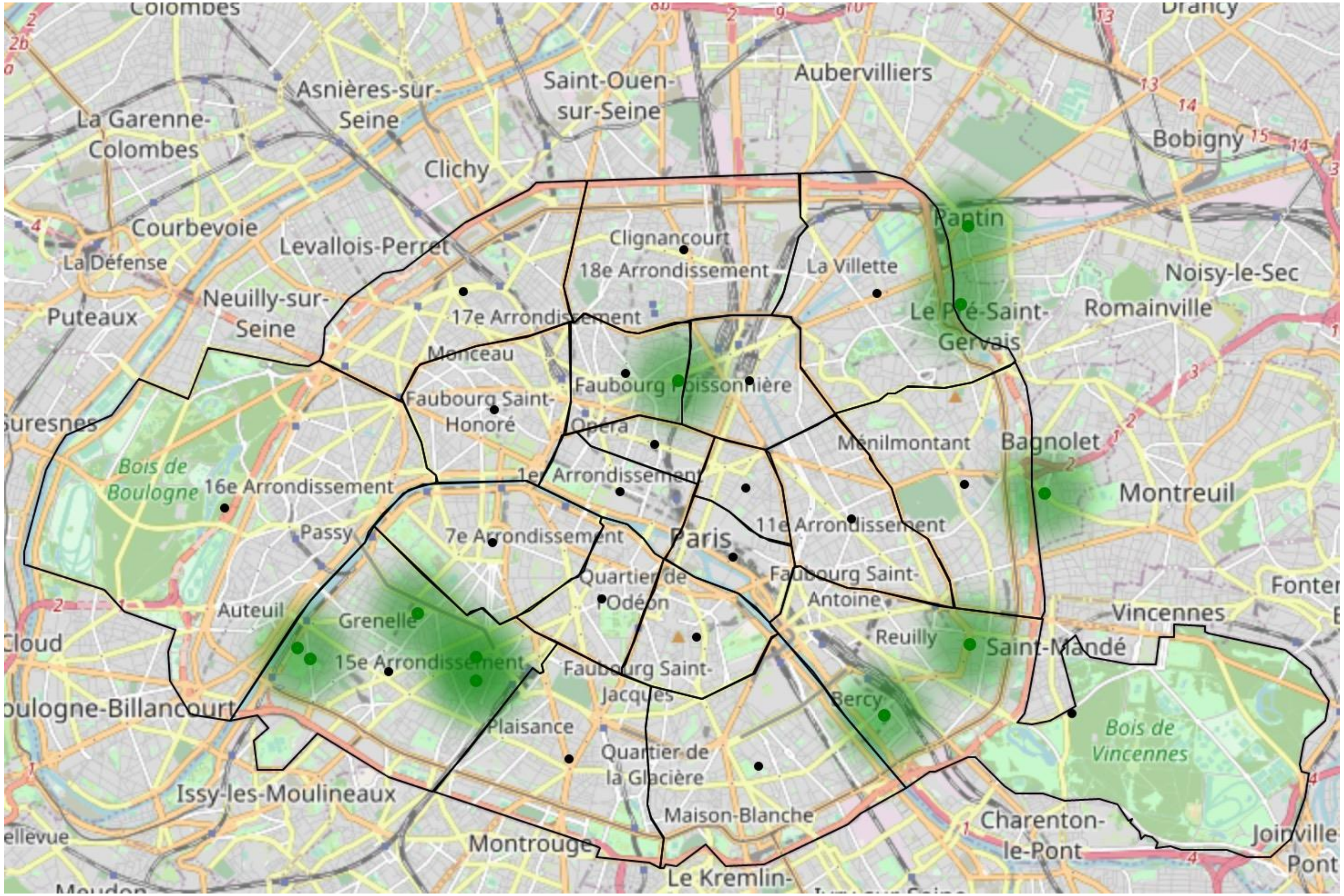
(even though the results were not too clear...)





Colors are for clusters, and darker areas are for areas with deficit.

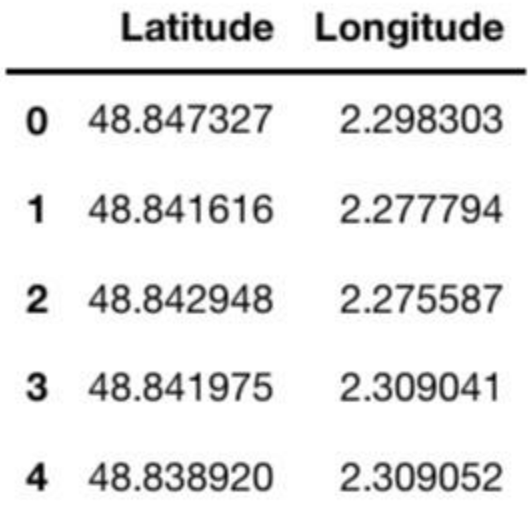
Refining the results



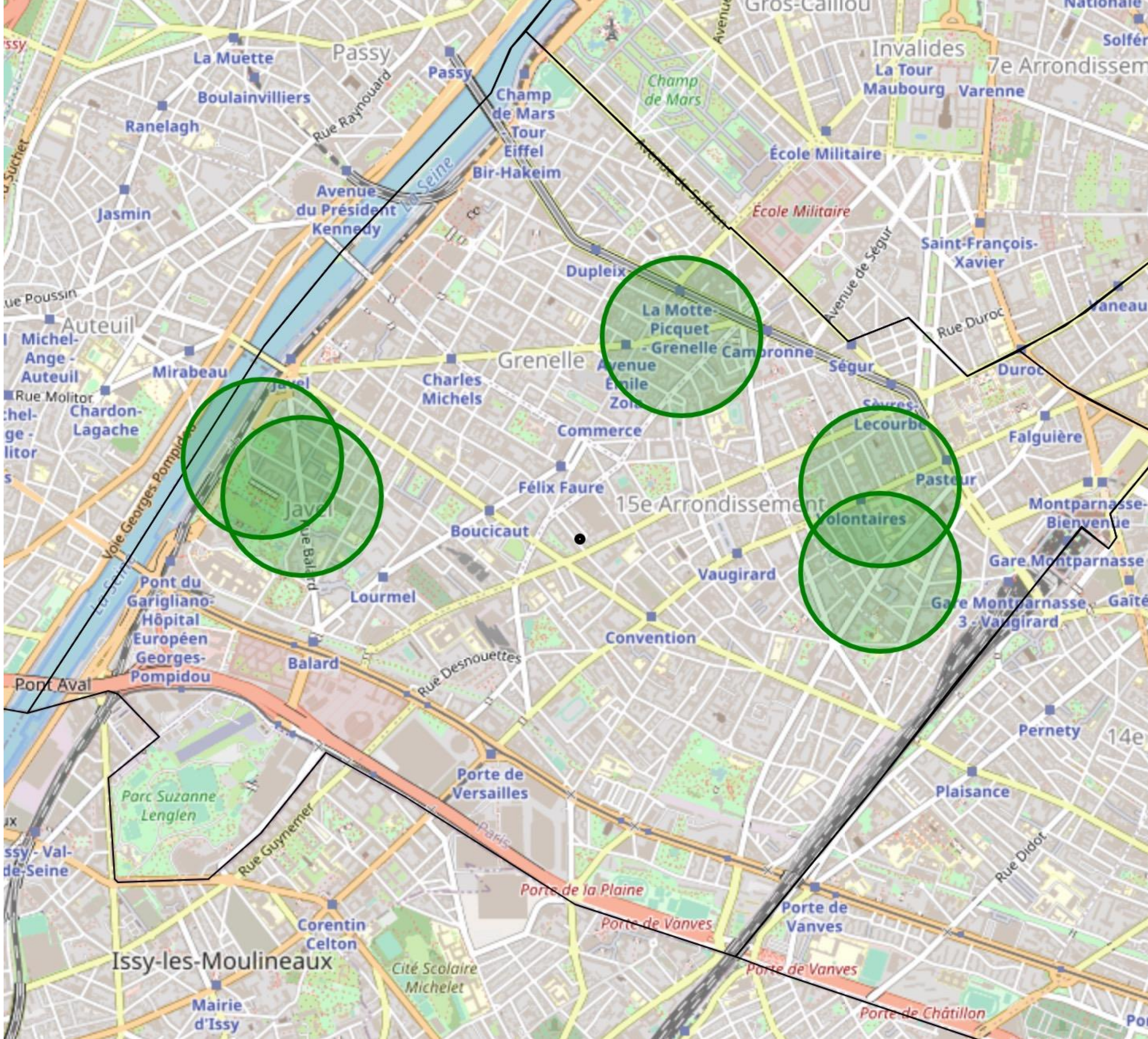
Keeping only the most relevant cluster, we can further reduce our findings.

Results

By combining our analyses, we obtained the following coordinates of places to further investigate:



	Latitude	Longitude
0	48.847327	2.298303
1	48.841616	2.277794
2	48.842948	2.275587
3	48.841975	2.309041
4	48.838920	2.309052



Conclusions

- We've decided to carry several analyses before concluding anything as one dimension of the data can almost never be sufficient on its own, and one should not forget that we have only taken into account a few variables: by cross checking different technical analyses, we increase the likelihood of getting to accurate conclusions.
- For a final decision to be made, it would be interesting to add other datasets to this study, such as the average price per 2 for each neighborhood, demographics, etc. but also, more importantly, to actually go on site to find out what it's like in real life.
- We've shown that with only a small dataset, one could already have a pretty good idea of what's going in a specific area, and target only a handful of locations for further investigations.
- These could lead to conclusions that exactly match the ones above, or shed light on something that had been completely missed, but in any case, a study of the available data can almost always teach us something we didn't know.