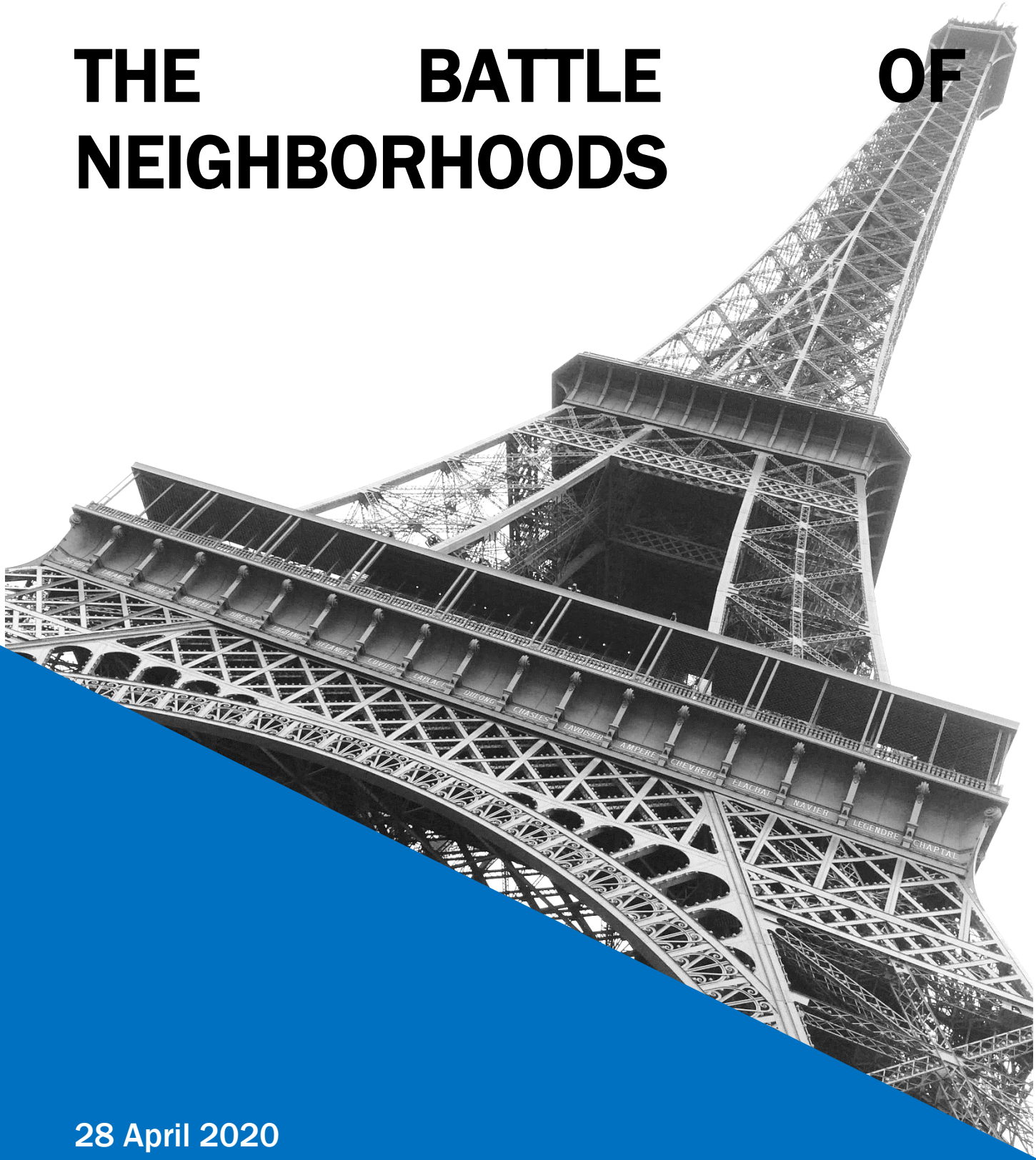
THE BATTLE OF NEIGHBORHOODS



28 April 2020

Julien DELMAS

Best location for a seafood restaurant in Paris, France

Paris is the administrative, economic and demographic capital of France. With over 2.2 million inhabitants, it is also the first touristic destination in the country and it is reknown for its cuisine.

There are hundreds of restaurants in Paris, most of them being French restaurant (mainly "brasserie") or Italian restaurants.

For this capstone project, I decided to suppose that I want to open a seafood restaurant in Paris and I want to find the best place to locate such a restaurant.

This project will be developped in a way that it would be very easy to change the restaurant type (for example change to salad bar, brasserie or fast-food restaurant) by only changing 1 parameter of the code.

Table of Contents

Problem Statement

As stated int the introduction I will try to find the best place to open a new seafood restaurant in Paris.
Paris has a high density of restaurants and finding the best place to open a restaurant is not easy. Consequently, data analysis is probably one of the best ways to approach this problem.

Paris is divided into 20 districts (called "arrondissement" in French). Each of these districts is divided in 4 neighborhoods (called "quartier" in French). Consequently, Paris is divided in 80 neighborhoods.
By analyzing the number of seafood restaurants and their typology, I will try to determine the best neighborhood to open a new seafood restaurant.

The “best place” to open a seafood restaurant can be defined as follows:

- At least one of the following criteria is met regarding competition:
  - ✓ Where there is no or few competition in the seafood restaurant category
  - ✓ Where the competition is on another range of price (cheaper or more premium)
  - ✓ Where the competition is poorly rated
- The population density is high enough to have enough clients (I ignore the impact of tourism for the sake of simplicity)

I will try to answer the following questions:

1. How many seafood restaurants are there in Paris?
2. Where are the best seafood restaurants of Paris located?
3. Where are they located? In which neighborhood / district?
4. In which neighborhood and/or borough should I open a seafood restaurant?

It is important to note that I will develop the project in a way that it would be very easy to change the restaurant type (for example change to salad bar, brasserie or fast-food restaurant) by only changing 1 parameter of the code.

This report can be of interest for any person willing to open a restaurant in Paris.For seafood restaurants, obviously, but it can only be used for other type of restaurants by changing 1 variable in the code.

Data Collection Plan

Data about the City of Paris

I will need to obtain the list of neighborhood with as much data as possible.

Use of Paris Opendata

The city of Paris issued numerous data in Opendata format on the website https://opendata.paris.fr. In particular, the page https://opendata.paris.fr/explore/dataset/quartier_paris/information gives data regarding the neighborhoods of Paris.
The last update of this data is from March 2013 which is sufficient due to the absence of administrive reorganization since this time.

The datafile, which is available in CSV, JSON, Excel, GeoJSON, Shapefile and KML, contains the following columns:

- N_SQ_QU: sequencial id of the neighborhood,
- C_QU: Number of the neighborhood,
- C_QUINSEE: Official number of the neighborhood according to a national format (provided by the Institut national de la statistique et des études économiques, INSEE, i.e. thr National Institute of Statistics and Economic Studies),
- L_QU: Name of the neighborhood,
- C_AR: Number of the district,
- N_SQ_AR: Sequencial link between district and neighborhood,
- PERIMETRE: Perimeter of the neighborhood,
- SURFACE: Area of the neighborhood,
- Geometry X Y: Coordinates of the center of the neighborhood,
- Geometry: Polygon of the boundaries of the neighborhood.

I will use the CSV format because it is easily usable with the *pandas* library and the following fields will be used:

- C_QU: Number of the neighborhood,
- C_QUINSEE: Official number of the neighborhood according to a national format,
- L_QU: Name of the neighborhood,
- C_AR: Number of the district,
- SURFACE: Area of the neighborhood.

In addition, the information regarding the geometry of each neighborhood will be used in GeoJSON format to create maps with Folium.

Use of Wikipedia

The Paris Opendata is missing the population by neighborhood. Consequently, I searched for another source of population per neighborhood and the only I found is Wikipedia (https://en.wikipedia.org/wiki/Quarters_of_Paris).
It is fair mentioning that the population data dates from 1999, which is more than 20 years old. Nevertheless, the population of Paris has not changed much within this period (see

https://en.wikipedia.org/wiki/Demographics_of_Paris#/media/File:Paris_Historical_Population.png) and we can assume that it is also the case with neighborhoods.

The fields available in the Wikipedia table are:

- District (district official name and "also called"),
- Neighborhood (number and name),
- Population in 1999,
- Area,
- Map.

I will not use the area as it is already available in the Paris Opendata dataset and I will not use the map as it is given in an image format.

Consequence

Due to the use of 2 sources, I will need to merge the information for both sources. The neihborhood number is the common key to use for the merger.

Thanks to this merger, I will have one database with the following data:

- Neighborhood code (according to INSEE format),
- Neighborhood number (from 1 to 80 according to Paris format),
- Neighborhood name,
- District number of the neighborhood (from 1 to 20),
- District name of the neighborhood,
- District "also called" name of the neighborhood,
- Population (in 1999) of the neighborhood,
- Area of the neighborhood,
- Perimeter of the neighborhood,
- Latitude and longitude of the neighborhood.

Then, I will be able to add:

- Postal Code which is formed of 75 (numer of the department in France) + 0 + number of the district (from 01 to 20),
- Density of each neighborhood by dividing the population by the area.

In addition, I will have a GeoJSON with the coordinates of the polygon for eahc neighborhood.

Data about the restaurants

I will use Foursquare to obtain the data about the seafood restaurants in Paris.

I will use the "search" method to obtain:

- the list of restaurants for each neighborhood (based on the coordinates of the center of the neighborhood and a radius of 1500 meters which should cover all the neighborhoods (except part of the 2 forests that are at the West and East extremities of the City and contain few restaurants),
- for each restaurant, its id, its name and its location.

Because all the neighborhoods have different sizes and I use only one radius, I will have to remove duplicates and check in which neighborhood belong each restaurant. I can perform this last step by using the GeoJSON obtained previously.

Afterwards, I will obtain more details about each seafood restaurant with the "venue" method of Foursquare:

- Price category,
- Rating of the restaurant,
- Number of likes,
- Number of tips,
- Whether the seafood category is the primary category of the restaurant.

Jupyter notebook

The Jupyter notebook is accessible on GitHub: https://github.com/Julien-D/Coursera_Capstone/blob/master/The%20Battle%20of%20Neighborhoods/The%20Battle%20of%20Neighborhoods.ipynb.

Methodology

Below are the analysis I performed in order to answer to the problem statement. I followed the plan I established earlier (see the plan at https://github.com/Julien-D/Coursera_Capstone/blob/master/The%20Battle%20of%20Neighborhoods/The%20Battle%20of%20Neighborhoods%20-%20Best%20location%20for%20a%20seafood%20restaurant%20in%20Paris%2C%20France%20-%20Draft%20Report%20(Week%201).pdf) and adapted according to results obtained.

Data visualization and analysis regarding neighborhoods of Paris

In order to better understand the neighborhood of Paris, I created:

- A choropleth map of Paris with the population of each neighborhood to understand where the population is mainly situated,
- A choropleth map of Paris with the density of each neighborhood to take into account the size of each neighborhood,
- A scatter plot of the population vs the area per neighborhood,
- A bar chart of the density of population per neighborhood,
- A bar chart of the density of population per district.

Data visualization and analysis regarding restaurants

In order to better understand the location of restaurants, I created:

- A map of the seafood restaurants to see their location,
- A bar chart of the number of seafood restaurants per neighborhood,
- A bar chart of the number of seafood restaurants per neighborhood and per million inhabitant,
- A bar chart of the population for the neighborhoods without seafood restaurant,
- A map representing the rating, the price and whether seafood restaurant is their primary category for each seafood restaurant,
- A scatter plot of price vs. rating of seafood restaurants,
- A bar chart of the average rating of seafood restaurants per neighborhood,
- A scatter plot of the average rating vs. the number of seafood restaurants of each neighborhood,
- A bar chart of the average price of seafood restaurants per neighborhood.

Clustering of the restaurants

I clustered the restaurants based on the Density-based spatial clustering of applications with noise (DBSCAN) algorithm.

The main reasons why I chose this clustering algorithm are:

- It can find clusters of arbitrary shapes,
- It is robust to outliers and we may have many outliers in our data,
- Compared to k-Means, we do not need to specify the number of clusters.

I simulated the DBSCAN algorithm with different sets of input parameters and compare the resuls:

- Coordinates,
- Price,
- Rating,
- Number of tips,
- Number of likes,
- Whether seafood restaurant is the primary category.

The best results were obtained with:

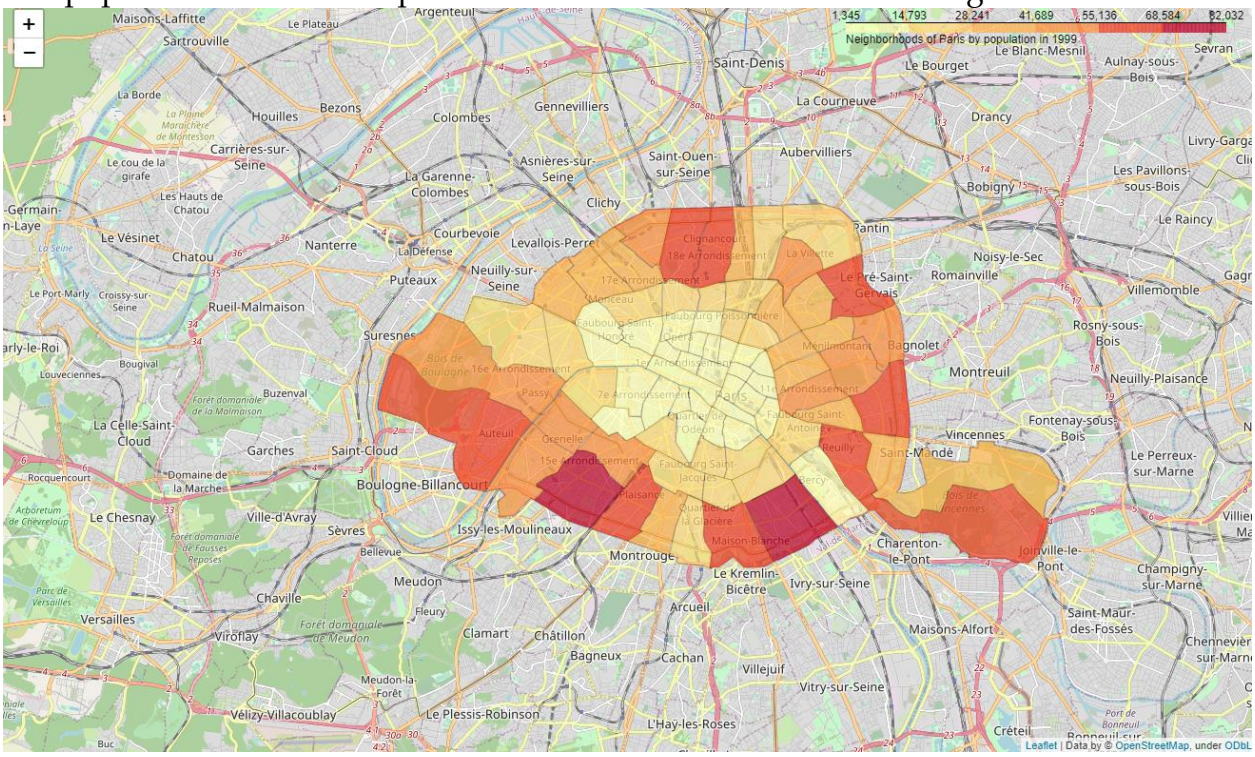
- epsilon = 0.3
- Minimum sample size of 5
- Clustering based on latitude, longitude, price and rating

Nevertheless, it showed lots of outliers so I decided to also cluster the restaurants with the k-Means method and the best results were obtained with 5 clusters.

Results

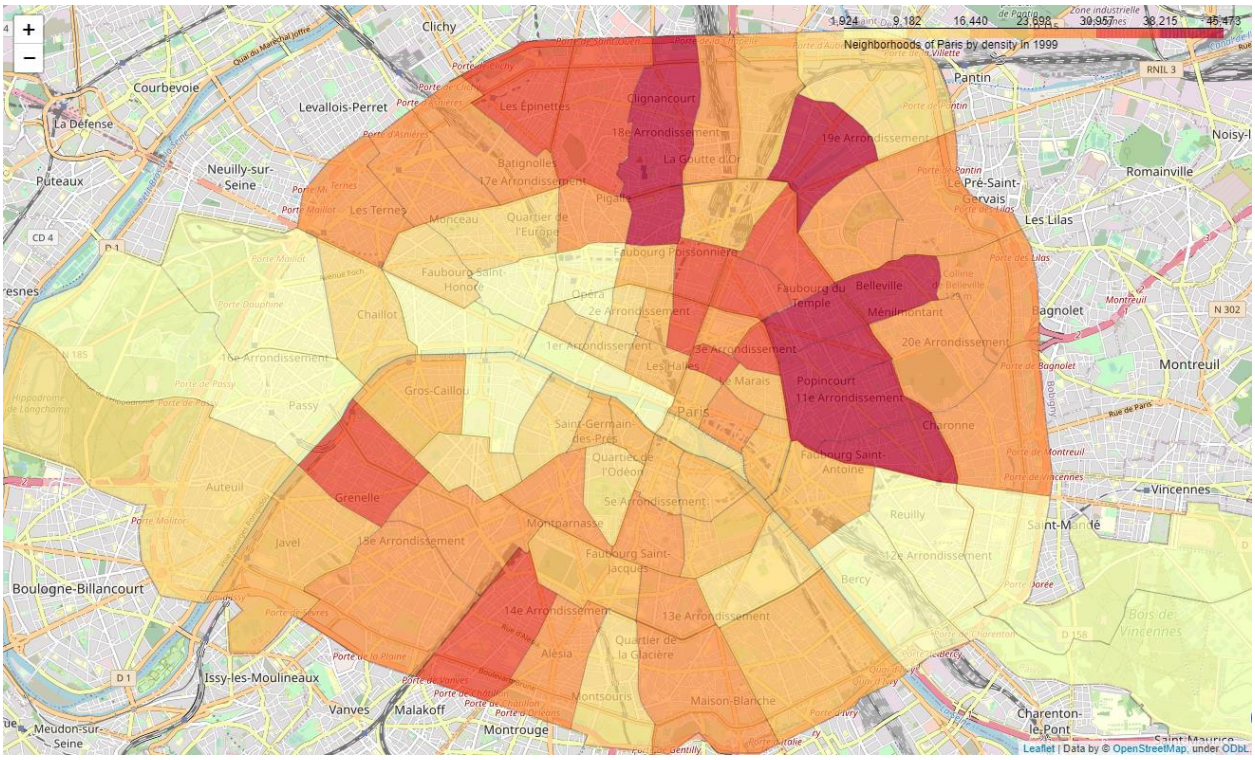
Data visualization and analysis regarding neighborhoods of Paris

The population iof Paris is spread as follows within the different neighborhoods:



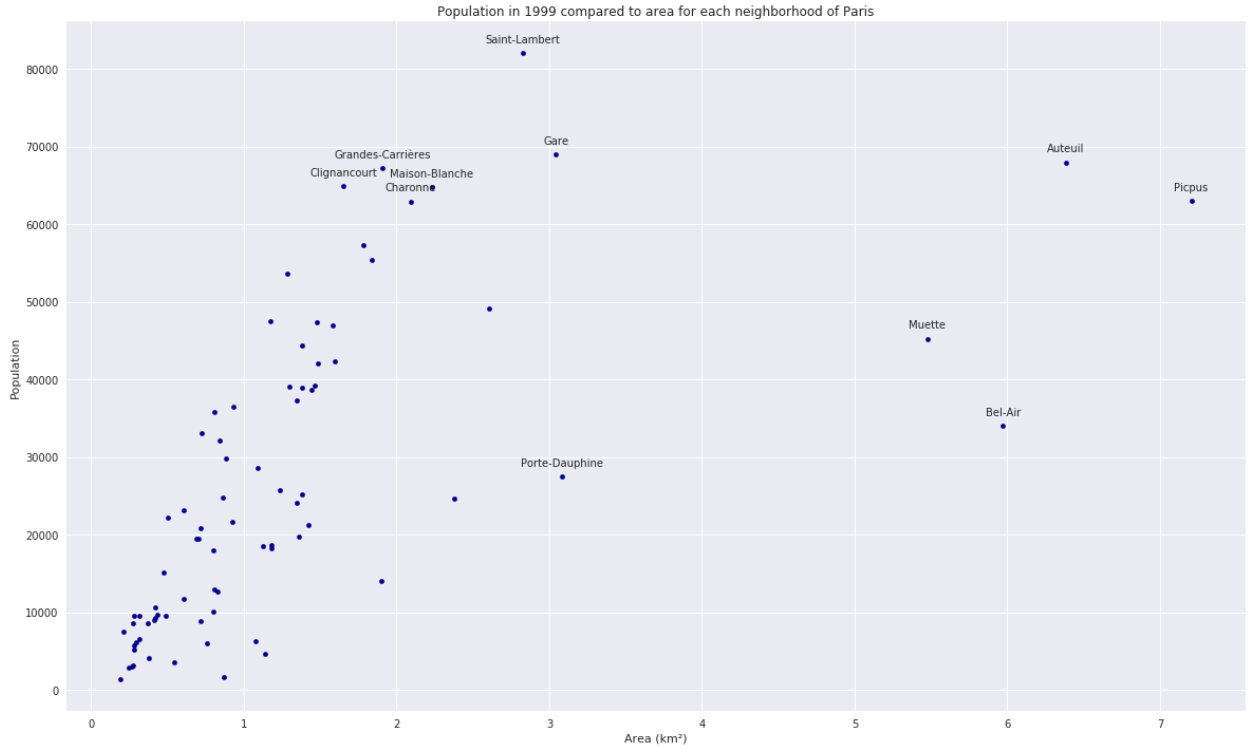
We can immediately see that the nieghborhoods situated in the center of Paris are less populated tat the one situated in the outside part of the city.

Nevertheless, the neighborhoods of the center are smaller than the ones in the outside of the city. A better analysis would be to look at the density (population divided by the area) of each neighborhood:



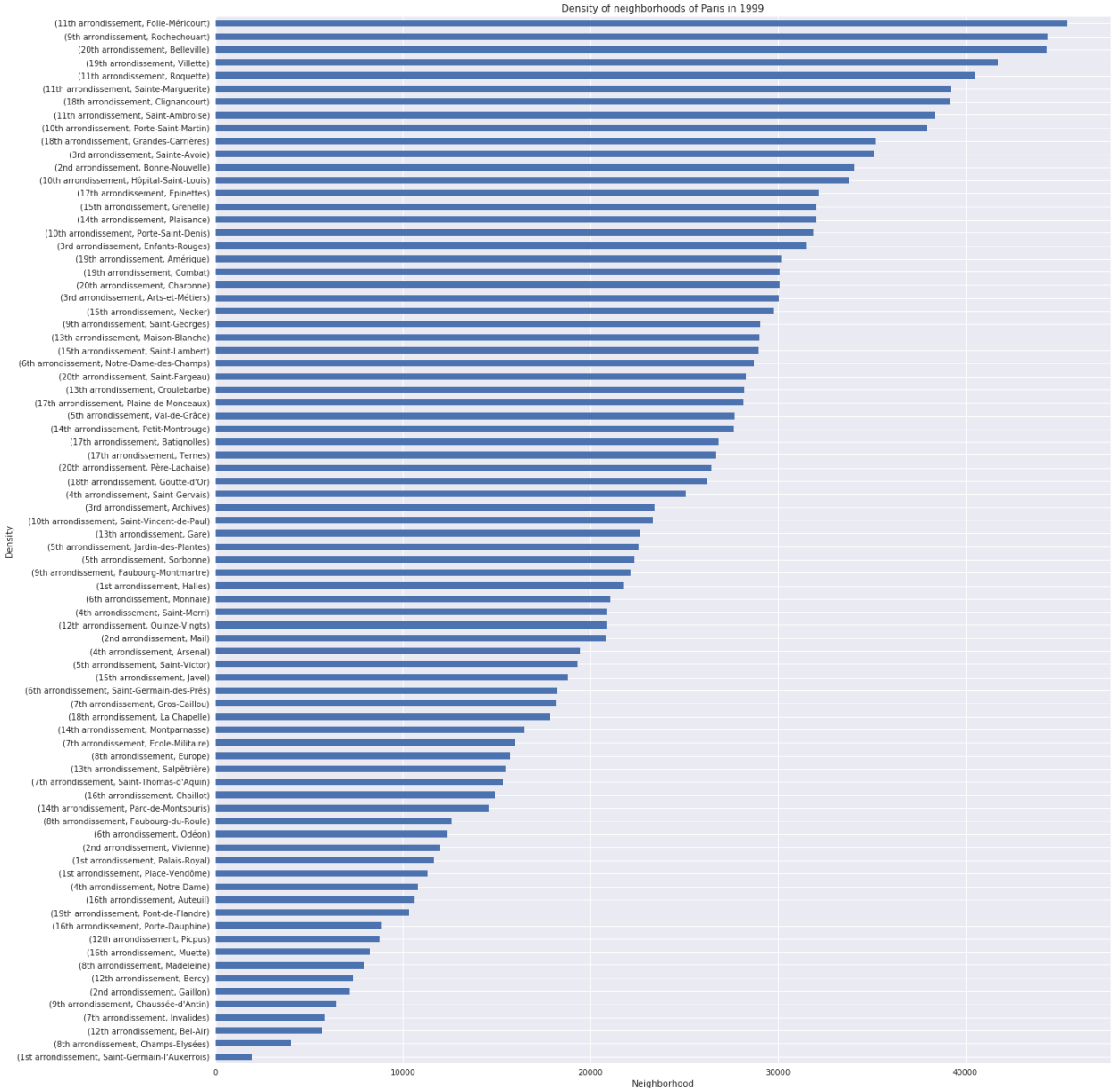
The most densely populated neighborhoods are in the North of Paris and the less densely populated in the South.

A scatter plot of the population vs. the area per neighborhood is as follows:

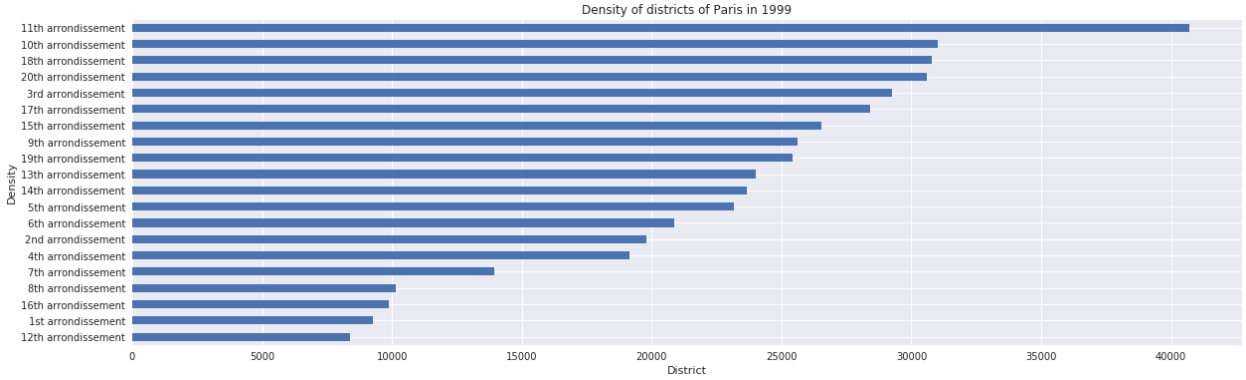


We can see that 4 neighborhoods (Bel-Air – 12th arrondissement, Picpus – 12th arrondissement, Auteuil – 16th arrondissement and Muette – 16th arrondissement) do not follow the average trend. This is due to the presence of 2 huge parks (Boulogne and Vincennes) that are occupying a significant area in these neighborhoods.

A bar chart of the density of population per neighborhood is as follows:



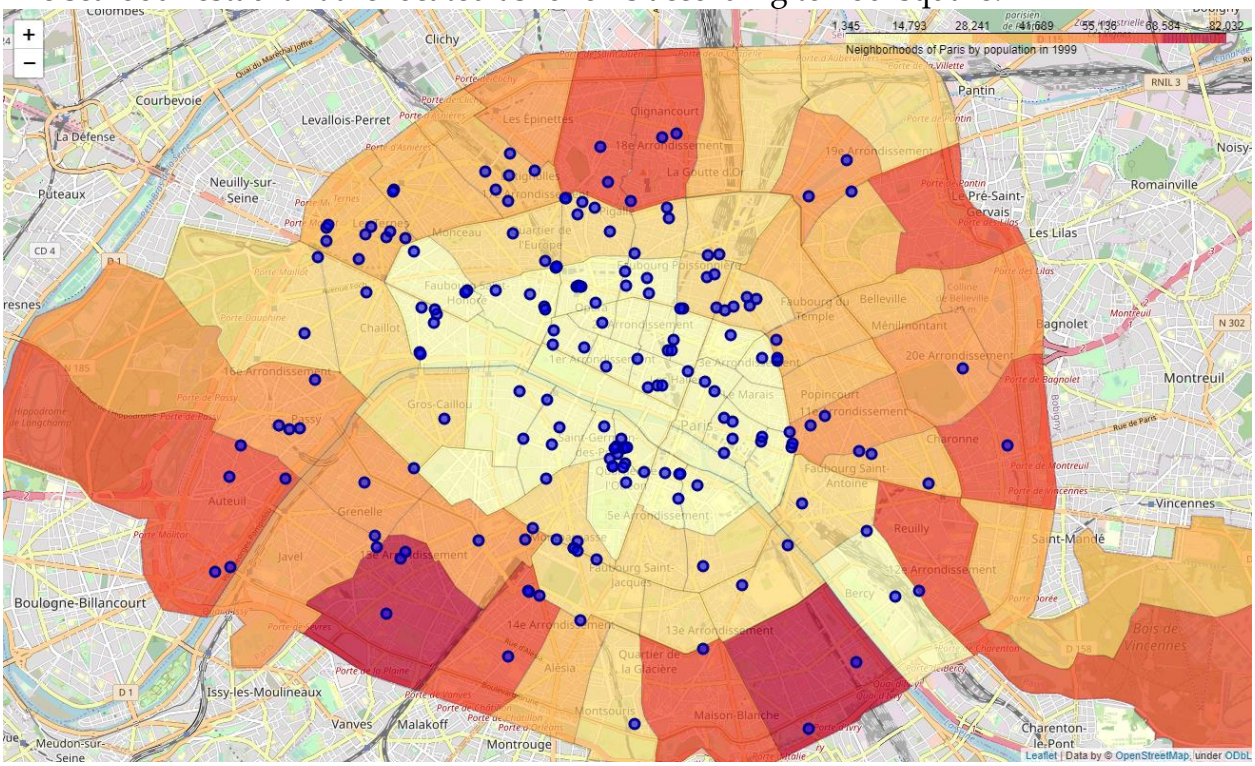
A bar chart of the density of population per district is as follows:



This analysis regarding the density of population will be used for the conclusion.

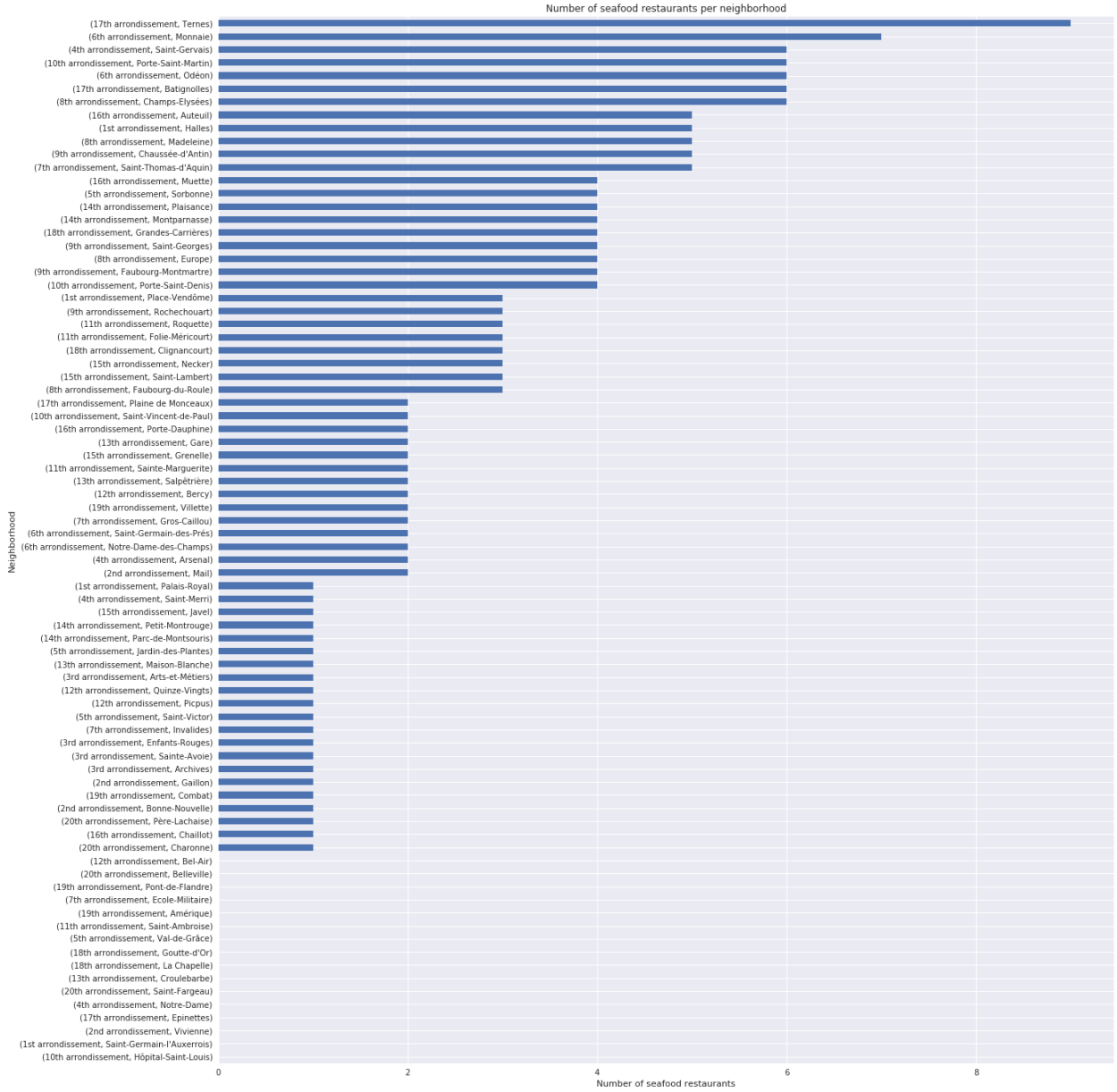
Data visualization and analysis regarding restaurants

The seafood restaurant are located as follows according to Foursquare:

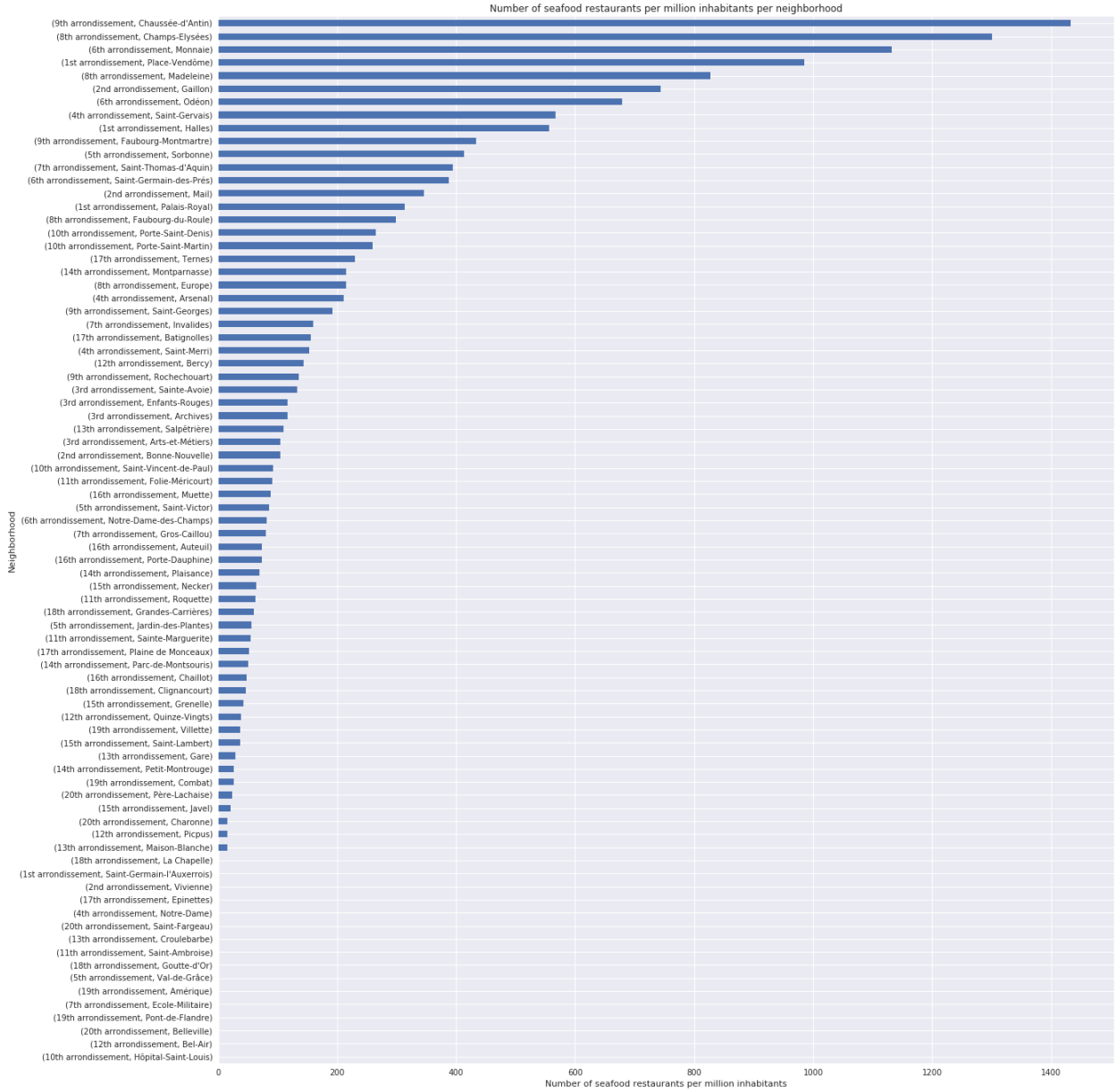


The first thing that we can see is that there is not one seafood restaurant in each neighborhood which could create opportunities. In addition, most of the restaurants are based in neighborhoods with low density.

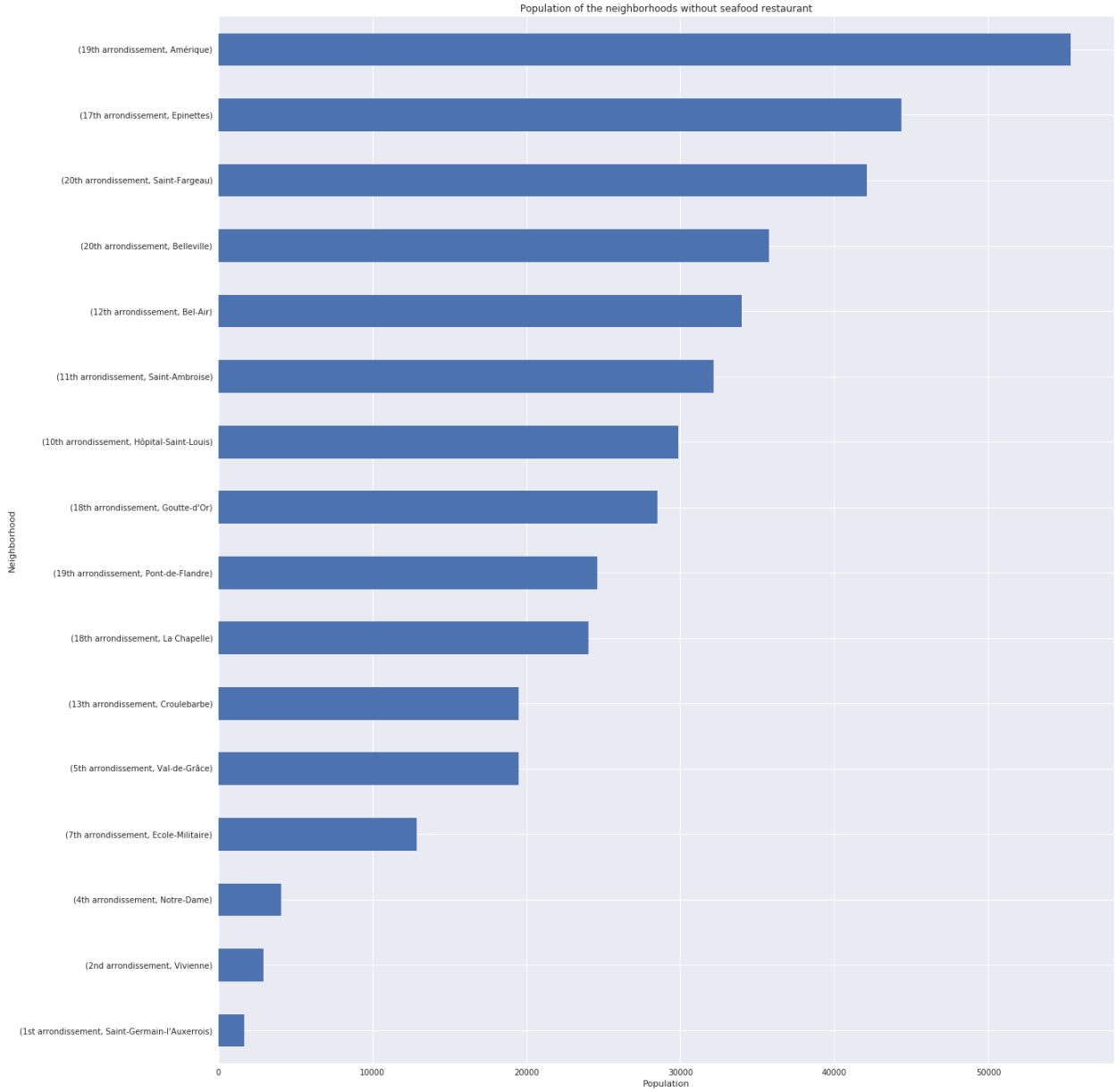
A bar chart of the number of seafood restaurants per neighborhood is as follows:



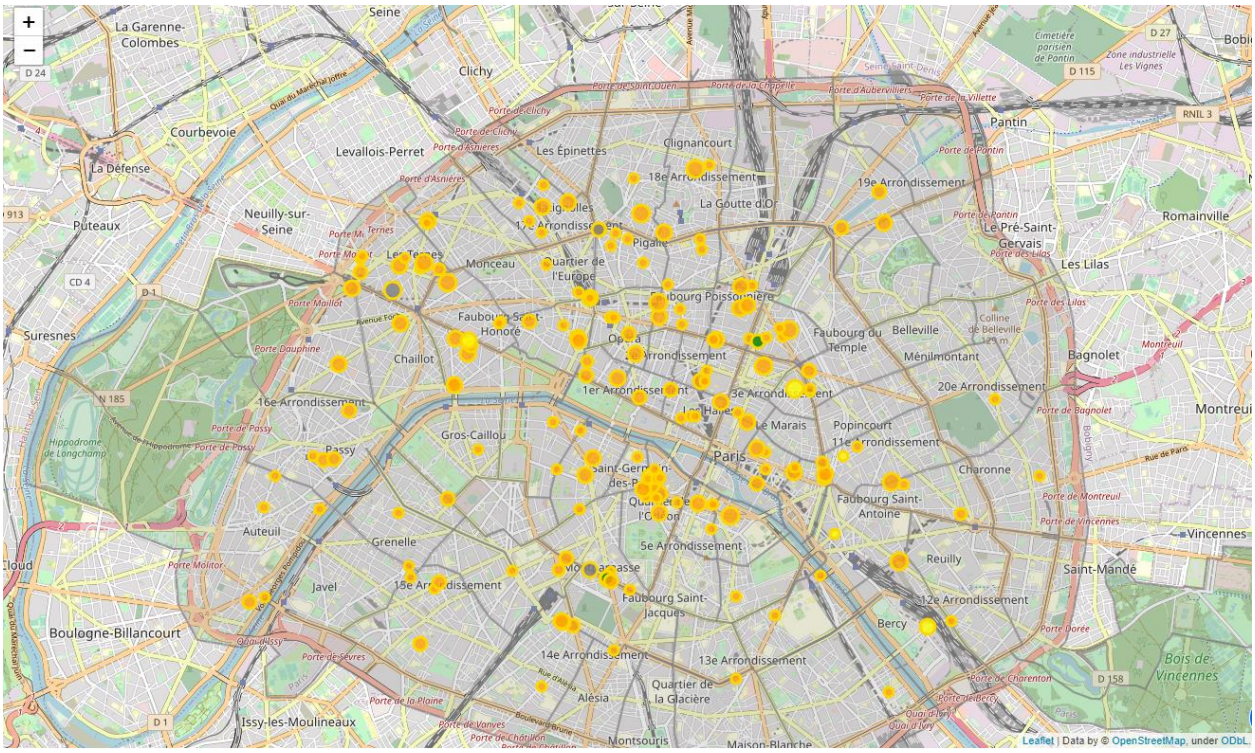
If I "normalize" these numbers by calculating the number of restaurants per millions inhabitants, I obtain the following bar chart:



The 16 neighborhoods without seafood restaurants have the following population:



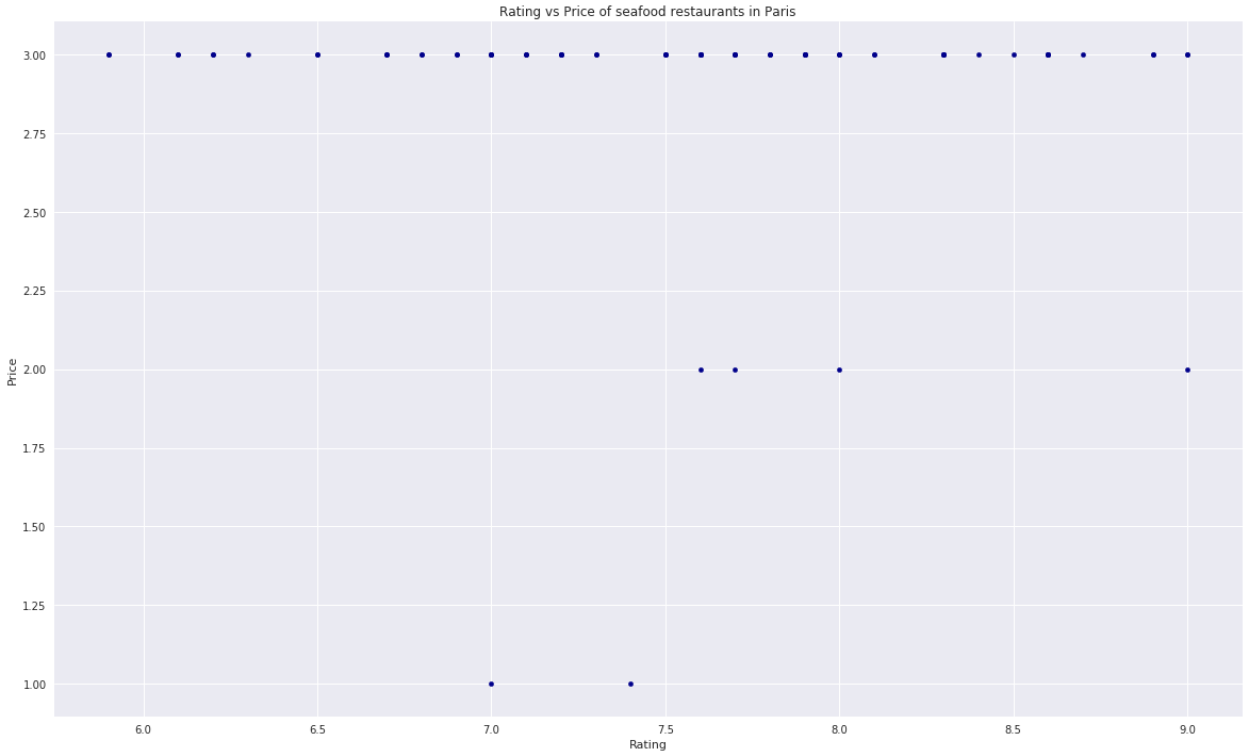
After obtaining detail data about the restaurant, I can perform more sofisticated analysis. Here is a map representing the rating (size of the circle), the price (green, yellow, orange or red depending on the price tier from 1 to 4 – the circle is grey if no information is available about price) and whether seafood restaurant is their primary category for each seafood restaurant (golden border if seafood is the primary category):



I obtain the following information:

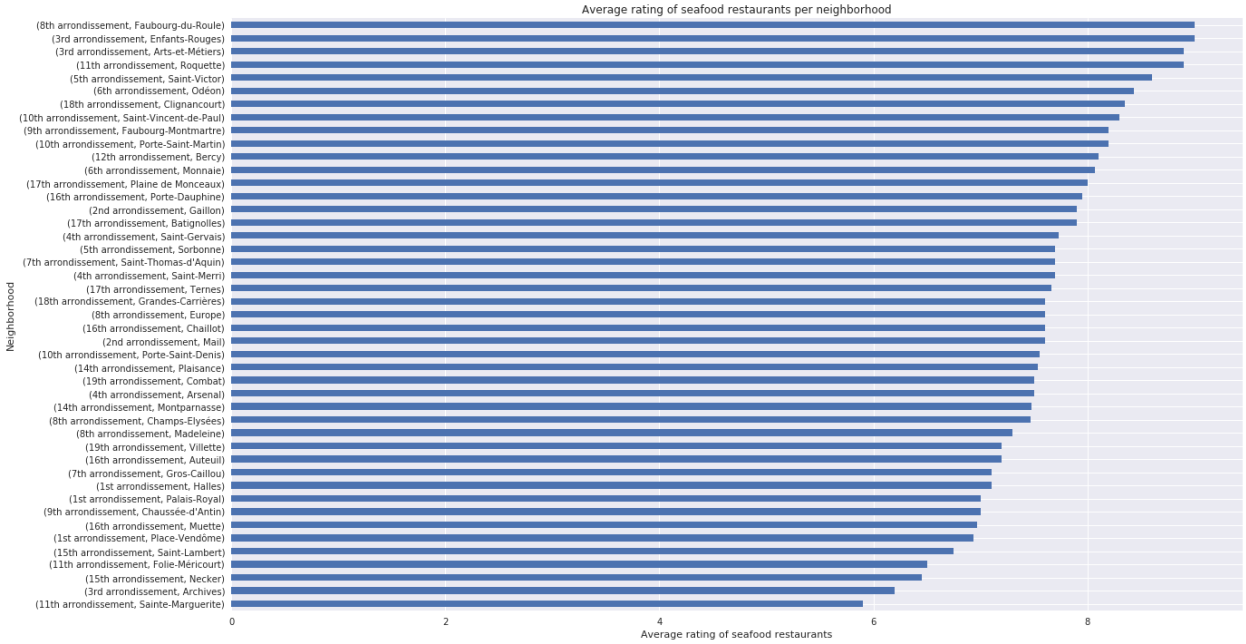
- All the seafood restaurants have the primary category "seafood restaurant". It is not a type of cuisine that is done in addition to another,
- Almost all the circles are orange, corresponding to the "expensive" category. There are also 2 "cheap" restaurants (green) and 4 "average" (yellow). There is no "very expensive" seafood restaurant,
- Ratings are varied (ranging between 5.7 and 9.0) and there is no clear geographic repartition of the ratings

I tried to find a correlation between price and rating and created the following scatter plot:

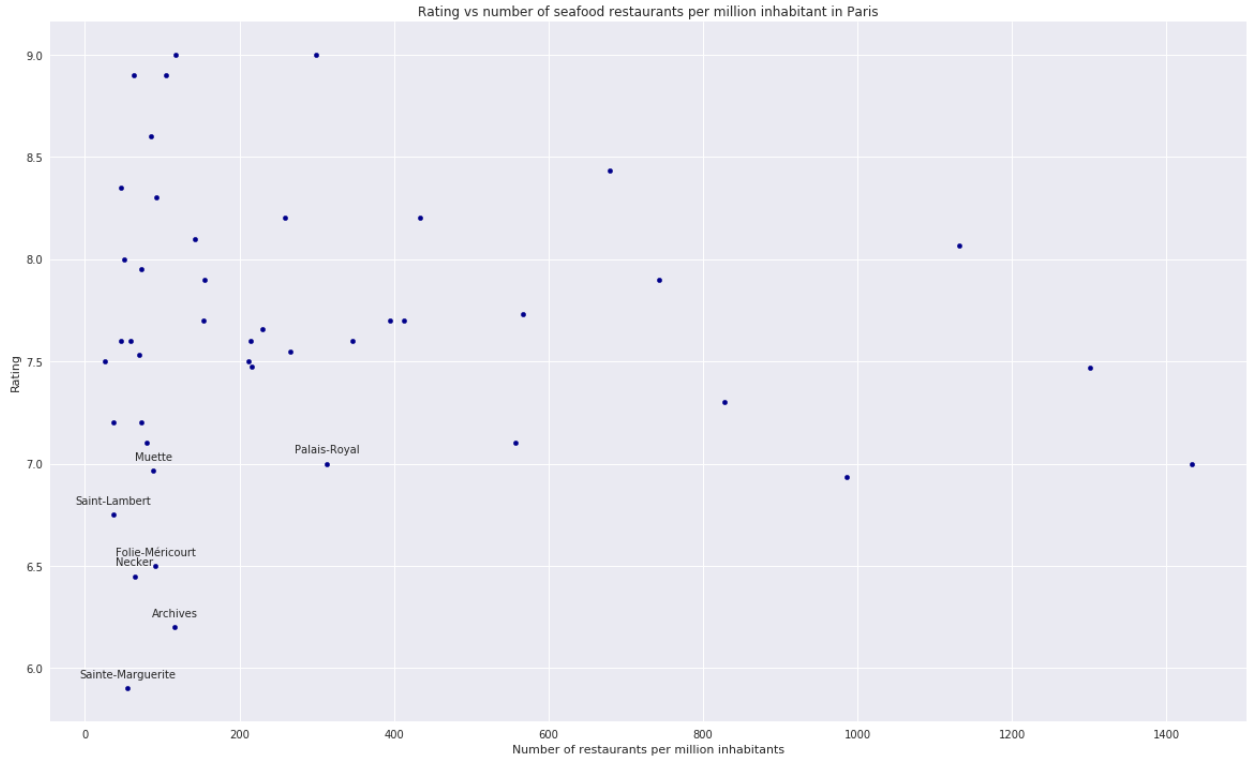


No correlation exist between the 2 criteria.

The analysis of average rating of seafood restaurant per neighborhood is as follows:

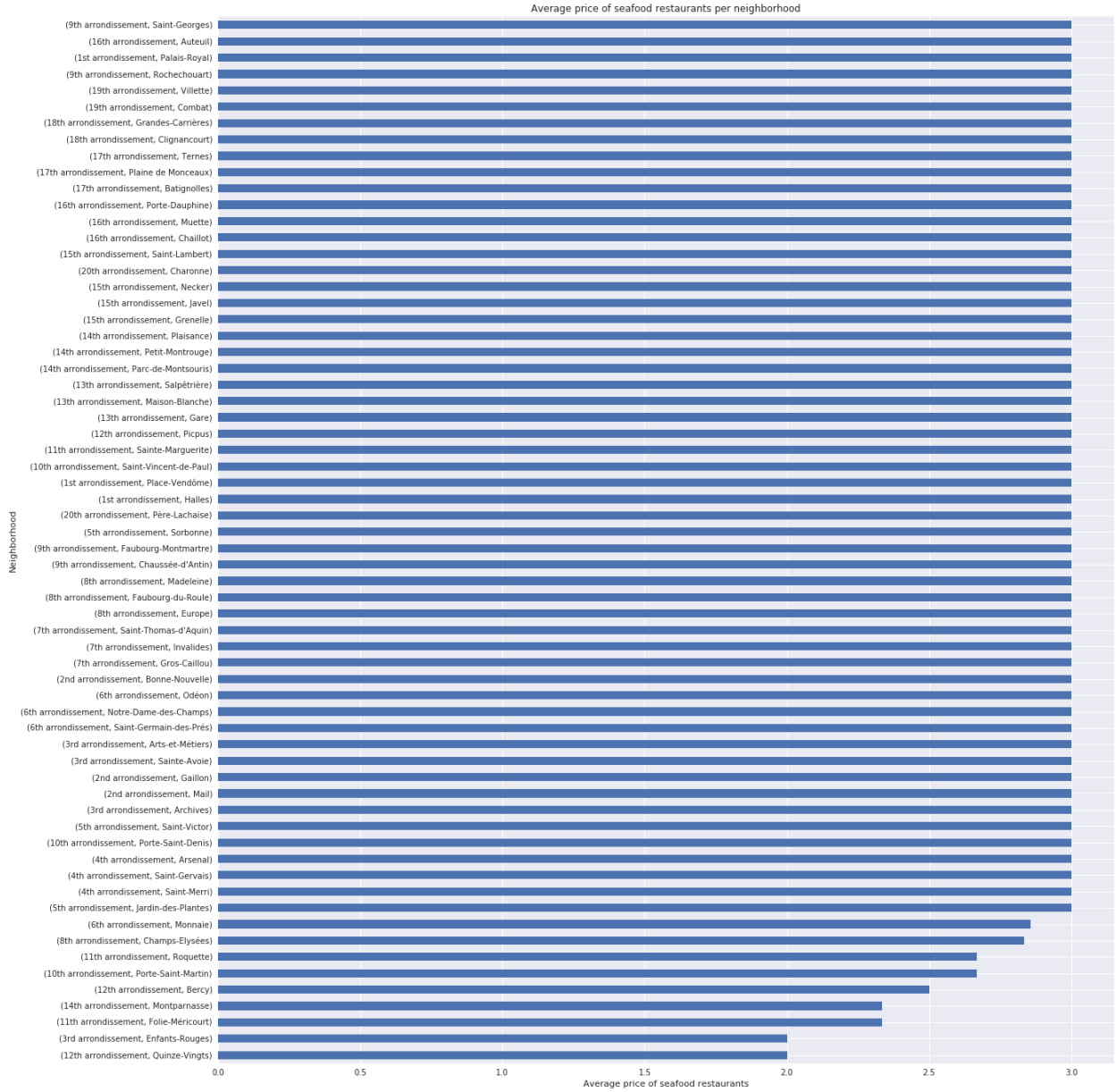


I plotted the average rating vs the number of restaurants per million inhabitants to see if I can find a neighborhood with few poorly rated restaurants which would create opportunities:



What I can see is that we have 4 neighborhood (Sainte-Marguerite – 11th arrondissement, Archives – 3rd arrondissement, Necker – 15th arrondissement and Folie-Méricourt – 11th arrondissement) with seafood restaurants poorly rated and a few number of restaurants.

The analysis of average price of seafood restaurant per neighborhood is as follows:



As seen above, most of the restaurants have a price tier of 3.

Clustering of the restaurants

I clustered the restaurants based on the Density-based spatial clustering of applications with noise (DBSCAN) algorithm.

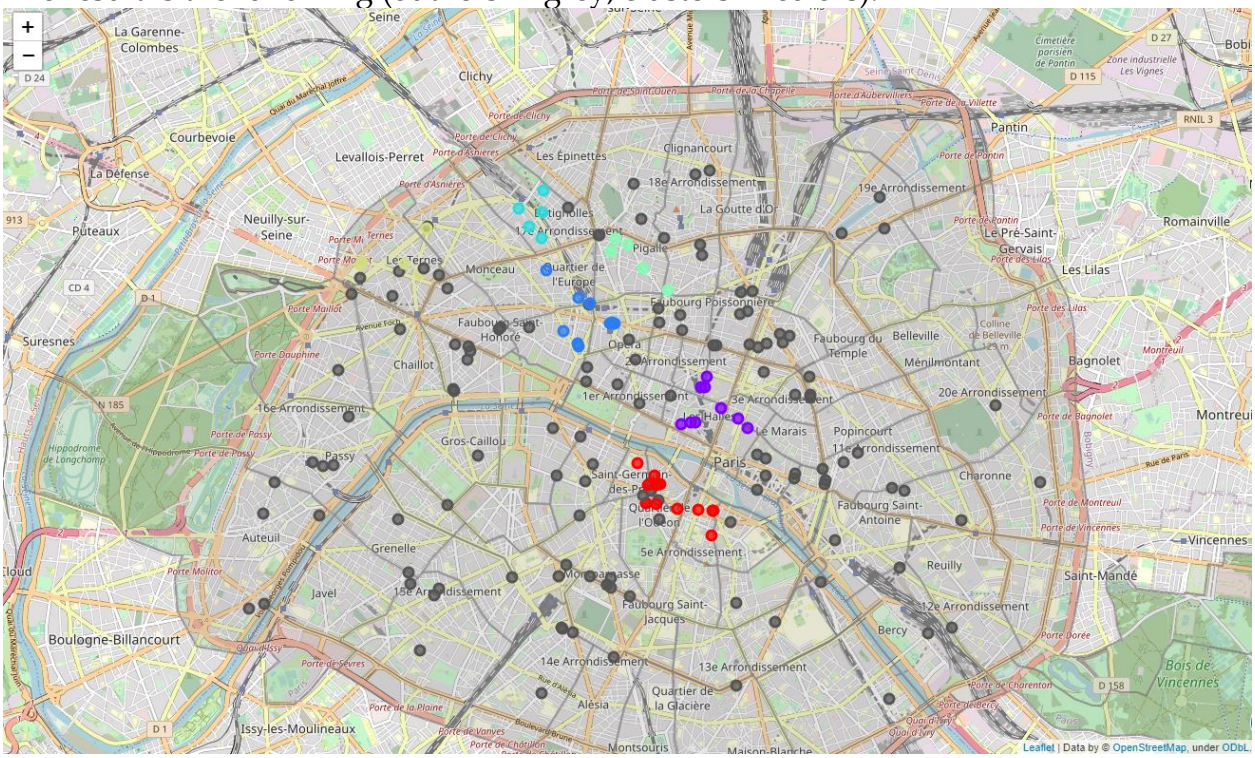
The main reasons why I chose this clustering algorithm are:

- It can find clusters of arbitrary shapes,
- It is robust to outliers and we may have many outliers in our data,
- Compared to k-Means, we do not need to specify the number of clusters.

I ran the alhgorithm with different parameters and the best results were obtained with:

- epsilon = 0.3
- Minimum sample size of 5
- Clustering based on latitude, longitude, price and rating

The result is the following (outliers in grey, clusters in colors):



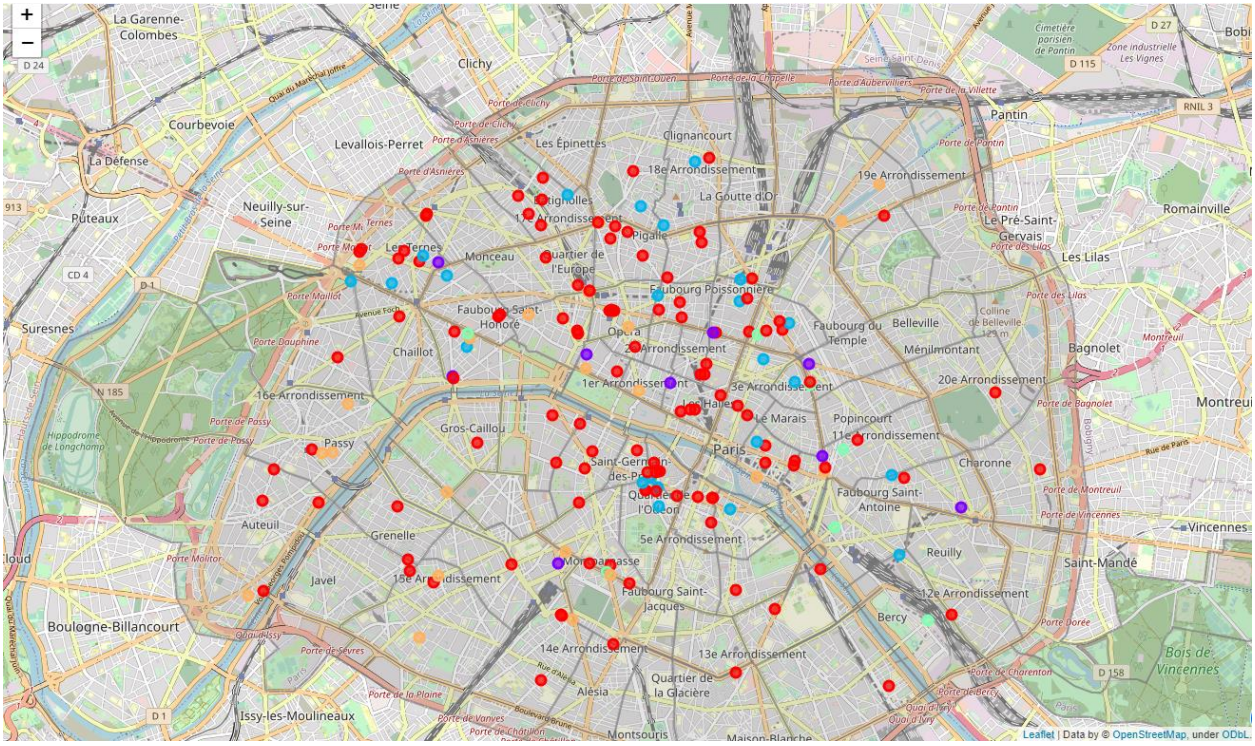
The clusters have the following characteristics:

	Price								Rating							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
ClusterDBSCAN																
-1	128.0	2.90625	0.364519	1.0	3.0	3.0	3.0	3.0	80.0	7.595000	0.804402	5.9	7.000	7.60	8.300	9.0
0	14.0	3.00000	0.000000	3.0	3.0	3.0	3.0	3.0	2.0	7.650000	0.070711	7.6	7.625	7.65	7.675	7.7
1	9.0	3.00000	0.000000	3.0	3.0	3.0	3.0	3.0	3.0	7.666667	0.057735	7.6	7.650	7.70	7.700	7.7
2	11.0	3.00000	0.000000	3.0	3.0	3.0	3.0	3.0	3.0	7.600000	0.100000	7.5	7.550	7.60	7.650	7.7
3	5.0	3.00000	0.000000	3.0	3.0	3.0	3.0	3.0	1.0	7.500000	NaN	7.5	7.500	7.50	7.500	7.5
4	5.0	3.00000	0.000000	3.0	3.0	3.0	3.0	3.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	5.0	3.00000	0.000000	3.0	3.0	3.0	3.0	3.0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We can see that the price is not really taken into account (all the restaurant with a price tier different than 3 are in the "outliers" cluster) and the average rating of each cluster is similar. The location was the only criteria of discrimination.

I was not very satisfied with the DBScan results (due to the high number of outliers and the bad differenciation between the clusters) so I decided to run another clustering algorithm: k-Means.

I ran it a few times and found out that the best results are obtained with 5 clusters:



The clusters have the following characteristics:

	Price								Rating							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
ClusterKMeans																
0	116.0	3.000000	0.000000	3.0	3.0	3.0	3.0	3.0	30.0	7.720000	0.171001	7.4	7.6	7.7	7.90	8.0
1	9.0	3.000000	0.000000	3.0	3.0	3.0	3.0	3.0	9.0	6.188889	0.220479	5.9	6.1	6.2	6.30	6.5
2	22.0	2.954545	0.213201	2.0	3.0	3.0	3.0	3.0	23.0	8.556522	0.282563	8.1	8.3	8.6	8.75	9.0
3	8.0	1.625000	0.517549	1.0	1.0	2.0	2.0	2.0	5.0	7.540000	0.371484	7.0	7.4	7.6	7.70	8.0
4	22.0	3.000000	0.000000	3.0	3.0	3.0	3.0	3.0	22.0	7.018182	0.191824	6.7	6.9	7.0	7.20	7.3

This methodology allowed a better clusterization with clusters with different average pricing and different average rating.

Discussions

Based on the analysis performed above, if I had to open a seafood restaurant in Paris, I would have several options:

1. Neighborhoods with no competition

16 neighborhoods (out of 80, ie. 20%) do not have a single seafood restaurants. Consequently, I could open my new restaurant in one of these.
In particular, the following neighborhoods have a high population but no seafood restaurant:

- Amériques (19th arrondissement)
- Épinettes (17th arrondissement)
- Saint-Fargeau (20th arrondissement)
- Belleville (20th arrondissement)
- Bel-Air (12th arrondissement)

Nevertheless, a further analysis of the typology of the population would be necessary in order to understand the reason why there are no seafood restaurants: seafood restaurants too expensive for this population, this population is not interested in seafood restaurants, business neighborhood, absence of tourists…

2. Neighborhoods with poor competition

The following neighborhoods evidenced the presence of poorly rated seafood restaurants:

- Sainte-Marguerite (11th arrondissement)
- Archives (3rd arrondissement)
- Necker (15th arrondissement)
- Folie-Méricourt (11th arrondissement)
- Saint-Lambert (15th arrondissement)

We can see that the 11th arrondissement and 15th arrondissement seem to concentrate seafood restaurants of poor quality.
In addition, 11th arrondissement is close to 19th and 20th arrondissement which are districts with few restaurants as we saw in the option 1.

3. Differenciation

All the seafood restaurants of Paris have "seafood restaurant" as their primary category. Consequently, one way of differenciating could be to open a restaurant which is not only a seafood restaurant in order to welcome guests that do not like seafood and want to accompany people who want to eat seafood.

With the k-Means method, we can see that 2 out of the 9 restaurants of Cluster 1 (expensive price, low rating) and only 1 of the 22 restaurants of Cluster 2 (expensive price, high rating) are situated in the 11th arrondissement.

Where are the best restaurants

Conclusion

Let's try to answer the 4 questions of the Problem Statement:

1. How many seafood restaurants are there in Paris?

There are 180 seafood restaurants in Paris.

2. Where are they located? In which neighborhood / district?

See the maps above.

3. Where are the best seafood restaurants of Paris located?

The best restaurants are the one of the Cluster 2 defined by the k-Means algorithm (the ones in blue). They are mainly situated in the 6th, 17th and 18th arrondissements.

4. In which neighborhood and/or borough should I open a seafood restaurant?

All the neighborhoods of 11th arrondissements seem to be a good fit as there are few restaurants and they are poorly noted. In addition, the district is close to the 19th and 20th arrondissement in which there are few to no seafood restaurant.


Doing this
project was
very time
consuming but
it wa so
interesting to
do a full
project by
myself!

Acknowledgements

Photo of the Eiffel Tower is from Pixabay: https://pixabay.com/fr/photos/paris-tour-effel-otel-resita-676931
It is under Pixabay license allowing its use in this report:

- All content on Pixabay can be used for free for commercial and noncommercial use across print and digital, except in the cases mentioned in "What is not allowed".
- Attribution is not required. Giving credit to the contributor or Pixabay is not necessary but i appreciated by our community.
- You can make modifications to content from Pixabay.

Photo of the seafood platter is from pxhere: https://pxhere.com/fr/photo/820162.
It is under Creative Commons CC0 license.