Search for locations for the establishment of a new yoga studio in Paris

1. Introduction

1.1. Background

There are nowadays more than 2.5 million yoga practitioners in France. Over the last 15 years, more and more people have been practicing yoga. Studio managers speak of an annual increase in attendance of 15 to 20%[1]. Faced with this ever-increasing demand, new yoga studios are set up every year. However, these same managers reveal us the difficulty to set up a profitable activity, because of the relatively low price of the courses given and the low number of profitable time slots. The choice of a new location for a yoga studio should therefore not only be based on the existence of competition but perhaps more on the possibility of concentrating the clientele and the cost of the premises.

1.2. Problem

The difficulty of obtaining a profitable yoga studio requires, during looking for a location to take into account other criteria than that of the competition's location. This study therefore aims to identify the client population and its areas of residence. With the additional study of the transport offer to increase the potential clientele and the rental rate zones we seek to reveal the optimal zones for the establishment of a new yoga studio.

1.3. Interest

This study could be of interest mainly to entrepreneurs who wish to open their first yoga studio or to expand their course offerings with additional premises. It will show them the areas that are most free of competition and most likely to concentrate a potential clientele by relying on nearby residents as well as on more distant customers by choosing a properly serviced area. It will also enable them to target the least expensive areas possible in terms of rental costs.

---

[1] [1]Redaction JDD. Le Journal du Dimanche. Véroniqua Leuilliot [online] ,accessed 22 May 2020. Available at : https://www.lejdd.fr/Sport/yoga-lage-du-business-3433355

2. Data acquisition

2.1. Data sources

The first grid to identify the areas of interest being the Parisian districts, we will use the open data from the data.gouv.fr site (postcode, surface area, coordinates of the centre and its perimeter, INSEE code). The Foursquare API enabled us to obtain all the places (name, category, coordinates) of Paris present in their database. As the API limit is 100 elements per query, we will have to divide the betting area into circles with a radius of 500 m, to be sure to extract the available data set. These data will enable us to identify the yoga studios but also the profile of each district. The number of yoga studios obtained being unrealistic because it is too small, it will be completed thanks to the API of the Sirene V3.0 database of the website ODS (opendatasoft)

The data on the distribution of the population (number, age groups and types of jobs by district) has been downloaded beforehand and is available on the INSEE website. Transport data (names of stations, lines, coordinates) are available on the « iledefrance mobilités » website. Affiliation to the boroughs of these stations will be possible thanks to the Address API of geo.api.gouv.fr, which makes it possible to obtain an address based on geographical coordinates. Finally, the average rental prices come from the geolocaux site. As the site is protected against code scrapping, but the effort being reasonable, the data was manually taken from the source code.

2.2. Data cleaning

The data obtained through the Foursquare API required first of all the removal of the duplicates obtained because of the methodology used. In addition, the postal code field was filled in more or less rigorously, which necessitated correcting the format of certain elements by deleting strings of characters. For 2-3 items, the postal code was incorrect and had to be changed. Finally, the places with the postcode 'Paris' or 'France' were simply deleted because they were few and irrelevant. Finally, the data were restricted to the districts of Paris (postal code 75001-75020) in order to remove unnecessary data.

Concerning the data from the Sirene V3.0 database, the query was made on the name 'yoga', which necessitated the deletion of data that did not correspond to the type of activity entered. Similarly, it was found that some establishments were duplicated, probably as a result of a move. Studios not currently active were therefore also deleted from the table. In total, half of the studios obtained via the Sirene V3.0 database were deleted.

Concerning the population distribution data, the document provides a large and rather redundant amount of information: e.g. on an age group, then on the same gender age group. After a quick data exploration, I identified the redundancies. For a question of efficiency it was simpler and quicker not to remove them but I did not use them.

The rest of the imported data required no correction.

3. Exploratory Data Analysis

3.1 Concentration of Yoga Studios

Once we have obtained the list of yoga studios, we can see the heterogeneity of the quantitative and geographical concentration of yoga studios by borough:

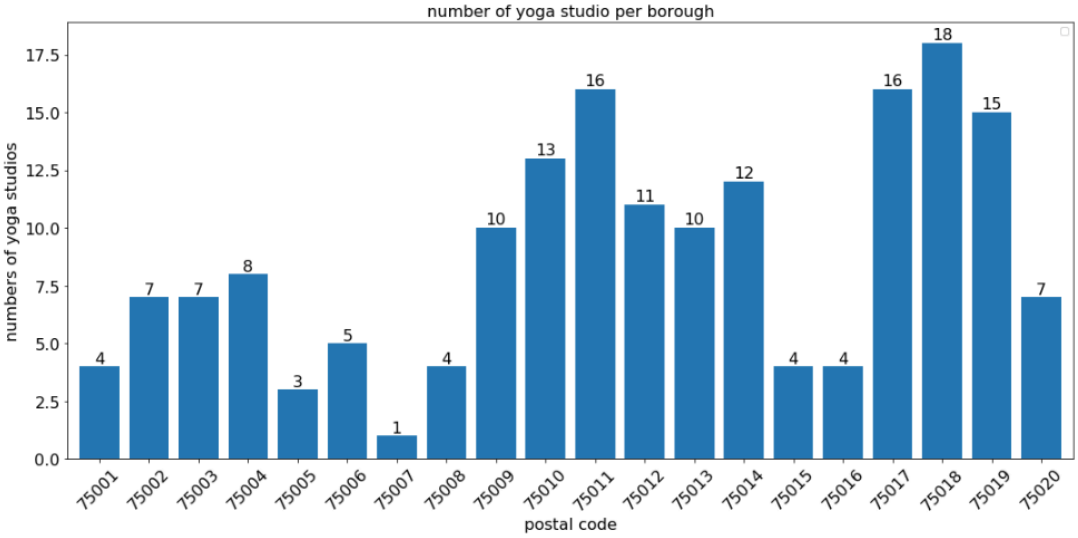

Fig. 1: representation of the number of yoga studios by district

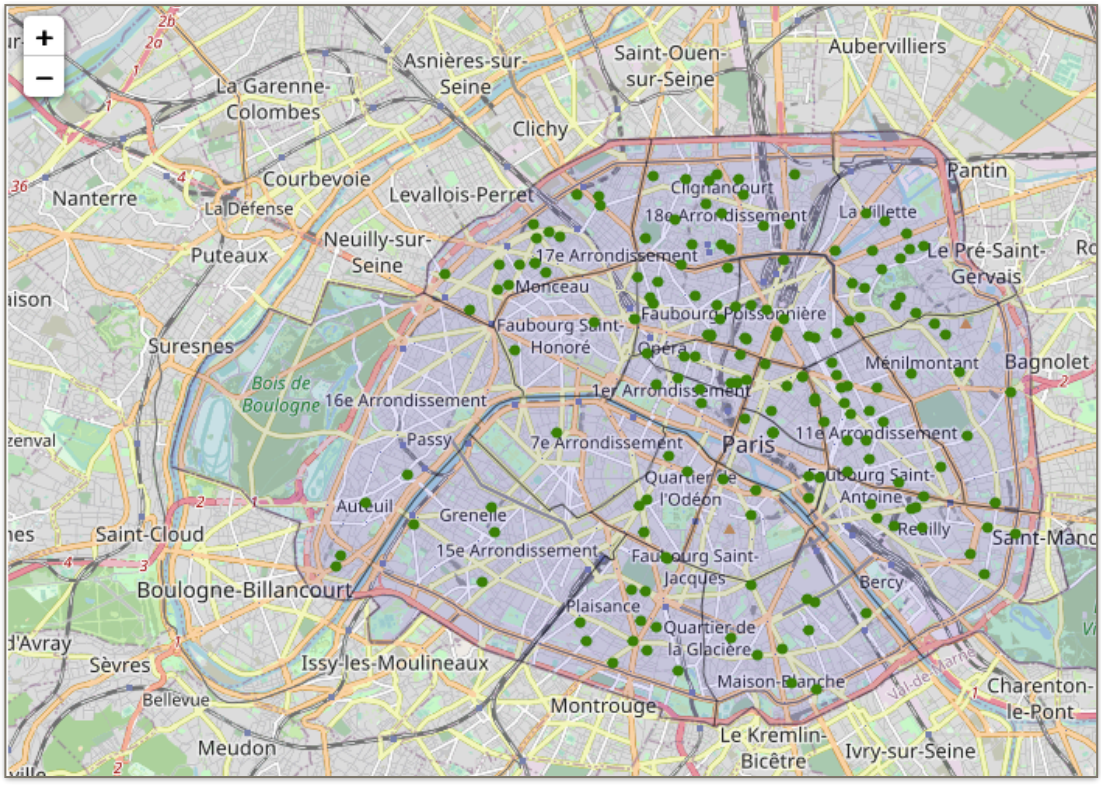

Fig.2: geographical representation of yoga studios per boroughs

We note that yoga studios are more established in the north and east of central Paris.

3.2 Profile of Paris districts according to their commercial and tourist activities

To try to explain the observed heterogeneity, we will establish a profile of the different Parisian districts based on commercial establishments (restaurants, bars, hotels, shops, etc.) and tourist sites (monuments, museums, squares, etc.). To do this, we have calculated for each district, the average of each category of places, then keep the 10 most frequent ones. We were then able to group the districts by similar profiles.

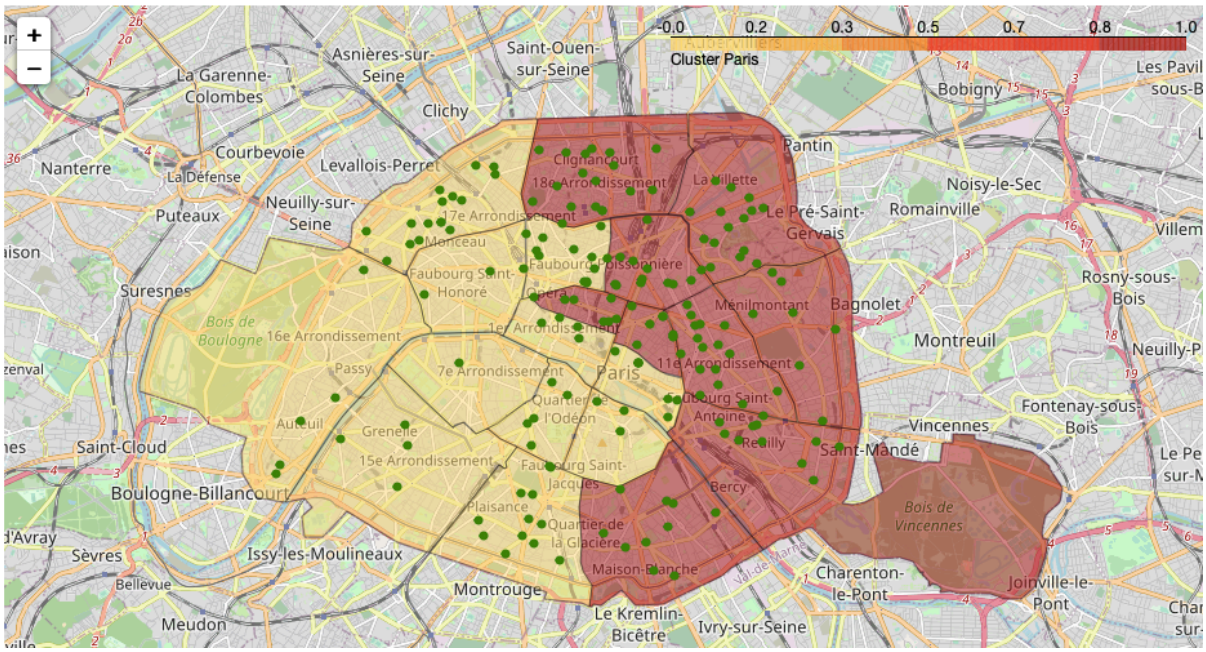

Fig.3 : Clusters of boroughs according to their commercial and tourist activity and location of yoga studios

We find a west-east cutout of the city. The two most frequent establishments per cluster are French restaurants and hotels for the yellow cluster and French restaurants and bars for the red cluster. By studying their presence in the cluster, we can see that the yellow cluster has about twice as many hotels and a third more French restaurants than the red cluster. Conversely, it has only one third of the bars present in the red cluster. We can thus suppose that the west of Paris welcomes more tourists than the east of Paris. The east of Paris is more populated by Parisians, and therefore probably has more potential customers. Moreover, the presence of bars in the red cluster may lead us to suppose that a larger share of the young working population resides there.

	Hotel			French restaurants			Bar		
	Min	**Mean**	**Max**	**Min**	**Mean**	**Max**	**Min**	**Mean**	**Max**
Yellow cluster	0.036	0.095	0.156	0.121	0.151	0.212	0.003	0.021	0.045
Red cluster	0.033	0.055	0.086	0.075	0.103	0.156	0.015	0.056	0.109

Fig. 4: Comparison of the presence of the 2 most frequented locations of the 2 identified Clusters

3.3 Profile of the population of the Paris district clusters

Since the data for the population distribution are in number of inhabitants, it was first necessary to calculate the percentages for each age group and each type of wage activity.

Then we calculated the average percentage for each category and the sum of the population for each cluster :

	Population	0_14	15_29	30_44	45_59	60_74	≥ 75
Yellow cluster	1 004 693	13 %	24 %	21 %	18 %	15 %	9 %
Red cluster	1 201 795	14 %	23 %	25 %	19 %	13 %	6 %

Fig. 5: Age distribution of the population of the 2 identified Clusters

	High-level professional	Mid-level professional	Employee	Labourers	Pensioners	Company director	Others
Yellow cluster	32 %	13 %	10 %	3 %	19 %	5 %	18 %
Red cluster	29 %	16 %	13 %	5 %	17 %	3 %	16 %

Fig6: Population distribution by salary activity in the 2 identified Clusters

The averages carried out by cluster on population segments and activities show us that the yellow cluster is made up of a population over 60 years old (22%) is higher than the red cluster (18%) while the active population aged 30 to 59 years old (28%) is lower than the red cluster (33%).Our hypothesis that the boroughs where many yoga studios are located correspond to those with a larger young and active population seems to be validated by these new data.

However, we note that the 13th, 14th and 17th districts all belong to the yellow cluster and have 10, 12 and 16 yoga studios respectively. This corresponds to an average to high number of studios per borough. We cannot therefore select the choice of boroughs according to these clusters. On the other hand, it will be possible to use it as a criterion for the final choice by the stakeholders.

3.4 Profile of the boroughs according to Parisian population distribution

Our previous analysis showed us that there may be a link between population distribution and the presence of yoga studios. We therefore investigated which type of population is correlated with the number of yoga studios:

	aged 30- 44 years old	Mid-level professional	Employee	Labourers
Pearson Correlation Coefficient	0.5920	0.5576	0.6093	0.6093
P-value	0.00596	0.01063	0.01393	0.00434

Fig7: Relevant correlation coefficient between different population groups and number of yoga studios

Pearson's correlation coefficients tell us that the number of yoga studios is positively correlated with similar strength for all four criteria. The P-values, on the other hand, tell us that the certainty of the correlation for the criteria "aged 30-44 years old" and "labourer" is greater than for the other 2 criteria.

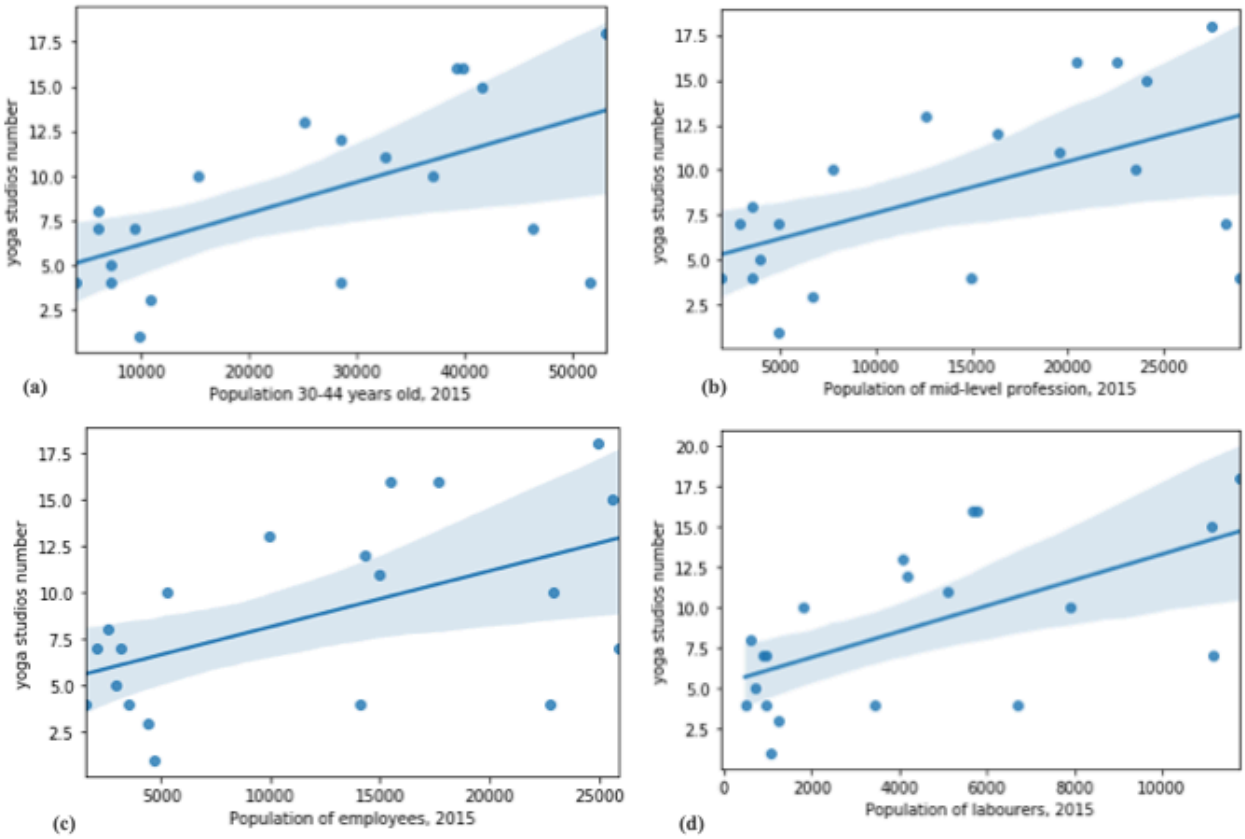

Fig8 : Representation of correlations between different population groups and number of yoga studios

Legend :

(a) correlation between the population aged 30-44 and the number of yoga studios

(b) correlation between the population with an intermediate profession and the number of yoga studios

(c) correlation between the population with an employed occupation and the number of yoga studios

(d) correlation between the population with a blue-collar occupation and the number of yoga studios

We were able to perform new population groupings according to these 4 criteria, the average values of which are as follows:

	aged 30- 44 years old	Mid-level professional	Employee	Labourers	All activity	Studios number	Studio density
Orange cluster	24.3%	16.4%	15.8%	7.1%	39.3%	13.3	2.18
White cluster	22.1%	14.4%	11.9%	3.7%	30.0 %	9.3	2.10
Light blue cluster	27.0%	16.1%	11.3%	4.2%	32.2%	10.6	5.3
Dark blue cluster	18.1%	11.1%	9.4%	2.4%	22.9%	3.4	1.0

Fig. 9: Average figures of the population distribution criteria correlated to the yoga studios number for the four clusters.

We get four clusters of boroughs:

- The orange cluster corresponds to the boroughs with the highest population of the 3 categories of activities (intermediate occupations, employees, blue-collar workers) and more particularly with the highest percentage of blue-collar workers. It is also observed that it has the highest average number of yoga studios per borough.
- The light blue cluster corresponds to the boroughs where the population aged 30 to 44 is the most represented and where intermediate professions are predominant. These districts also have the highest density of yoga studios.
- The white cluster is similar to the sky blue cluster, with a slight decrease in the representation of the population aged 30 to 44 (4%), intermediate occupations (2%) and blue-collar activities (1%). It also has half the density and a lower average number of studios per borough than the first two clusters.
- The dark blue cluster corresponds to the boroughs with the lowest percentage of the population aged 30 to 44 and the lowest proportion of the 3 categories of activities. It has the lowest average density and the lowest average number of yoga studios per borough.

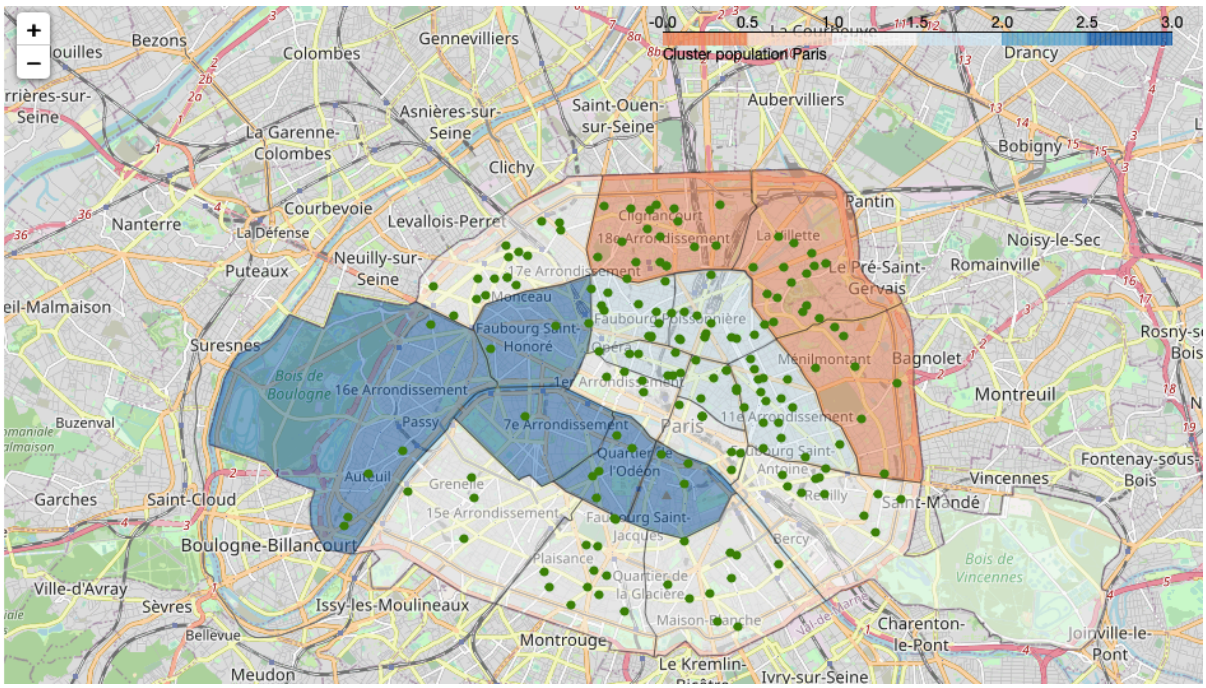

Fig. 10: Grouping of boroughs according to their population aged 30 to 44 years, working in an intermediate occupation, employees or manual workers

The presence of yoga studios is positively correlated to the four criteria studied but with greater certainty for the population aged 30 to 44 years old and those with a labour force activity. The proportion of workers (max average: 7%) being always lower than that of the age group (max average: 26%) we propose to privilege the age criterion of the population. This enables us to classify the following clusters in this order: light blue, orange, white, dark blue. The light blue and orange clusters being already the most equipped in terms of yoga studios, we propose to focus on the white cluster. Indeed, although the dark blue cluster has the fewest yoga studios we judge from the figures that it is not sufficiently populated with potential customers unlike the white cluster. The districts corresponding to the white cluster are as follows: 75001,75004, 75012 to 75015,75017.

3.5 Profile of boroughs and areas around transport stations according to commercial rental prices

From the data of the average rental price of commercial premises per station, we can calculate the average prices by rounding. This enables us to create a cluster of districts according to the average price, as well as a cluster of areas centred on metro stations according to their average price.

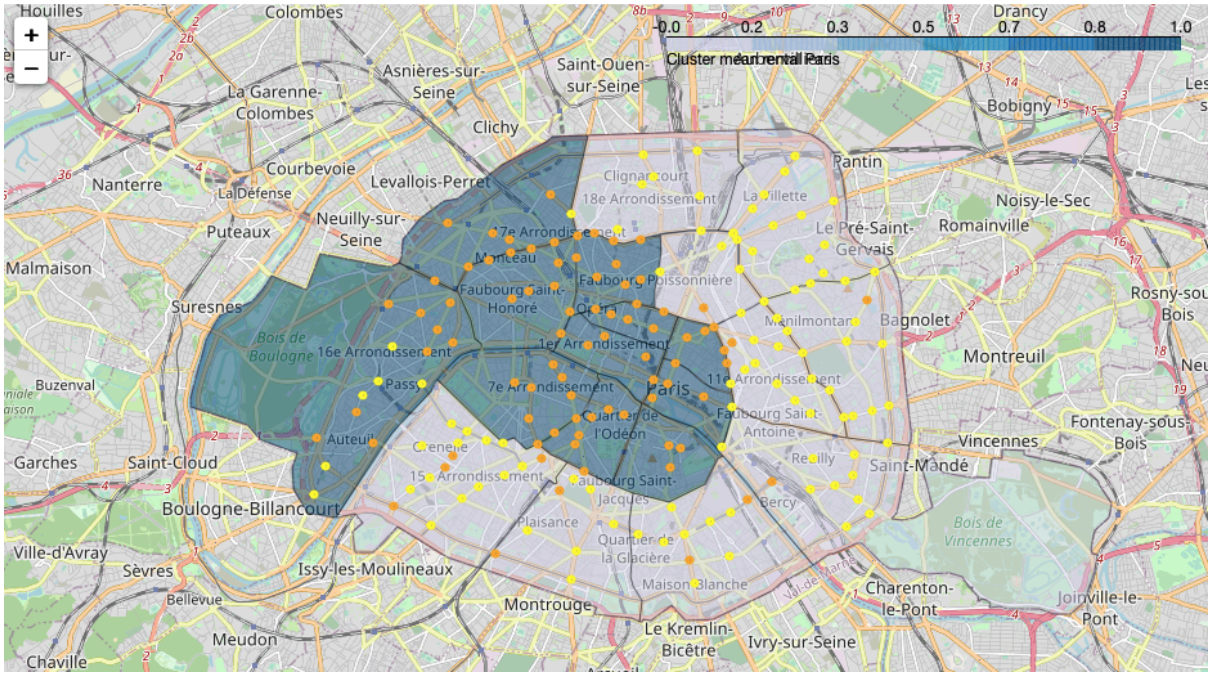

Fig. 11: Grouping of boroughs and metro stations according to their average rental prices for commercial premises

	Boroughs		Stations	
	Cluster dark blue	Cluster light blue	Cluster yellow	Cluster orange
Min	463 €	298 €	202 €	424 €
Max	597 €	385 €	420 €	763 €

Fig. 12: Average extreme values of average rental prices for commercial premises by grouping of boroughs and metro stations

As explained in the introduction, getting a profitable yoga studio is difficult. That is why we propose to focus only on the boroughs with the lowest prices. Cross-checking this selection with the previous one, we are left with only four boroughs: 75012 to 75015.

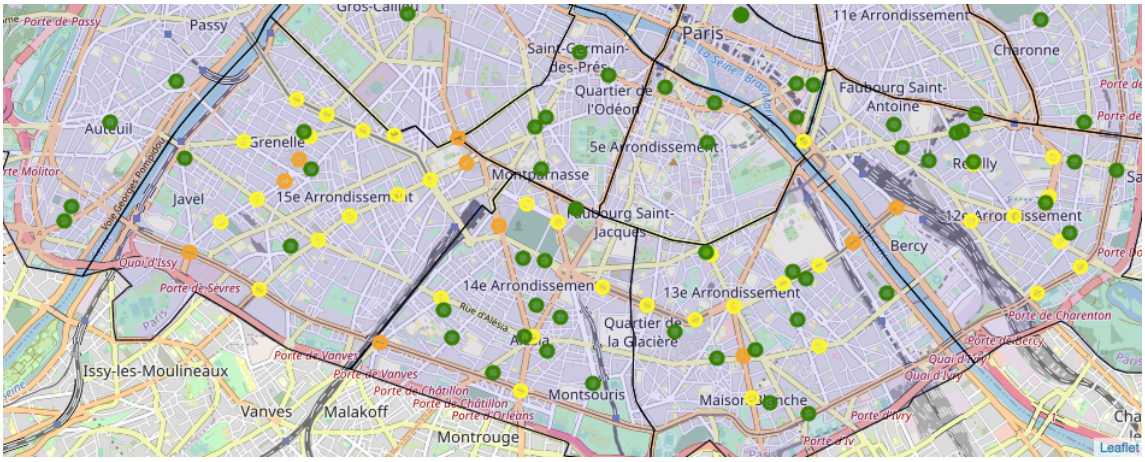

Fig.13:Representation of metro stations according to their price range (orange and yellow) and existing yoga studios for pre-selected boroughs

For the rest of the study we keep only the stations (yellow) belonging to the lowest price range.

3.6 Ease of transport

In order to be able to increase the number of customers, we want to offer an area that is well served by transport. We have decided that the stations must have at least two transport lines. We have calculated this data from the imported data. Otherwise, we want to propose as an additional criterion, the transport weight of the boroughs according to the number of available lines and the number of hubs (stations with at least 3 transport lines) existing in the borough. The number of lines per station and the transport weight were calculated from the imported data. We obtain the following areas :

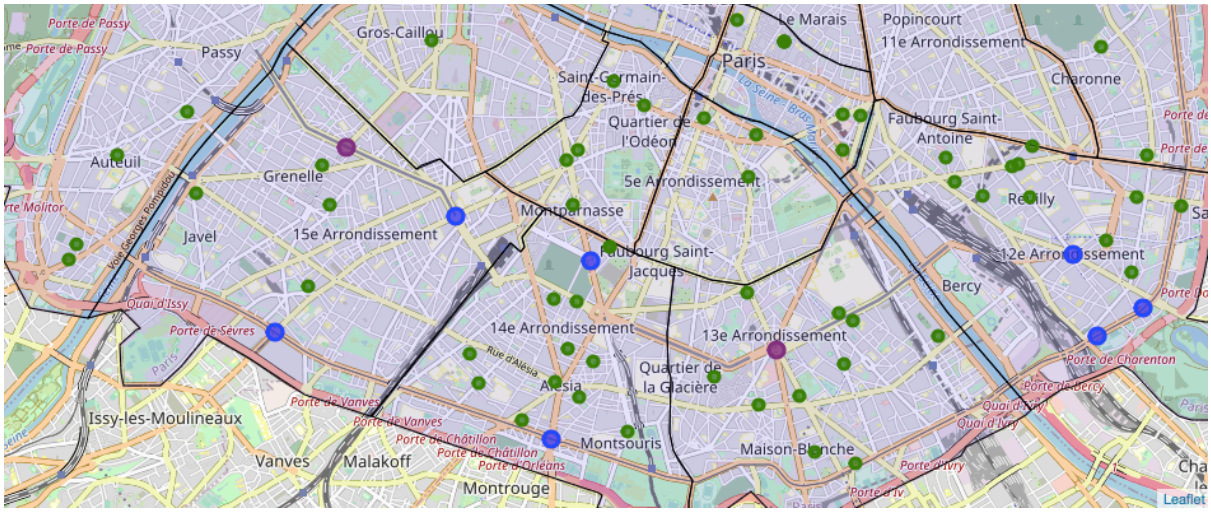

Fig. 14: Representation of subway stations according to their lines numbers (blue and purple) and existing yoga studios for the pre-selected boroughs

Station	Rental price (€/m^2/year)	Number of lines	Postal code	Transport weight	Touristic
Pasteur	342	2	75015	4.95	Yes
La Motte Picquet Grenelle	343	3	75015	4.95	Yes
Porte Doréee	311	2	75012	7.23	No
Porte de Versailles	315	2	75015	4.95	Yes
Porte d'Orléans	302	2	75014	2.06	Yes
Porte d'Italie	389	3	75013	3.56	No
Porte de Charenton	275	2	75012	7.23	No
Raspail	420	2	75014	2.06	Yes
Daumesnil	336	2	75012	7.23	No

Fig. 15: Data from selected stations

3.7 Final Decision Tracks

We got nine stations, nine possible zones, four of which are located at the far ends of the city. Since we are looking to increase our clientele by using transportation from different directions, we find these four stations less interesting. Instead, we propose to focus on the stations closer to the centre of Paris: Pasteur, La Motte Picquet Grenelle, Raspail, Porte d'Italie, Daumesnil.

Stakeholders will have to weight the following criteria in order to refine the final selection: rental price, number of lines, transport weight, number of studios present in the district, the district's tourist profile.

4. Conclusion

The goal of this project was to identify areas of Paris populated by potential clients in order to help stakeholders narrow down the search for an optimal location for a new yoga studio. By studying the profile of the boroughs based on data on population distribution and density of yoga studios, we first identified the general boroughs that warrant further analysis (75001, 75004, 75012-75015, 75017). Next, we grouped the boroughs according to the rental prices of the premises. The superimposition of the different groupings carried out allowed us to reduce the interest boroughs to the 12th, 13th, 14th, and 15th boroughs. Taking into account the offer of transport revealed areas of interest at the mesh of the transport station. These areas could serve as a starting point for final exploration by stakeholders.

The final decision on the optimal location of yoga studios will be made by the stakeholders based on the specific characteristics of the boroughs and the locations in each recommended zone, taking into account additional factors such as competition for the same type of yoga practice, the attractiveness of each location (size of residential population), the proximity of Vélib stations, the availability of real estate, prices, the social and economic dynamics of each zone, and so on.