Property market in Paris

Comparing prices vs neighborhood’s services

Nuno Alagoa, 2020-03-16

1 Introduction

The property market is one of the most speculative markets today.

Although housing is considered a basic necessity, it’s also seen as one of the most profitable investments, making it difficult for most people to buy or even rent property in cities’ center.

Paris is one of the most expensive cities in the world when it comes to buying or renting properties. According to the Global Property Guide’s website [1], Paris is the 7th most expensive city in the world to buy property and the 10th when it comes to renting. For someone who want’s to rent a property in Paris, a question that frequently arises is “Am I paying this much because this neighborhood offers a good range of services or is it just due to speculation?”. The intent of this study is to provide some insights to help a future tenant answering this question.

There are numerous factors that influence property price and it’s not possible to consider it all in this study, so I will try to focus on relevant services for neighborhood’s inhabitants (for instance, the existence of hotels is not relevant, but the existence of banking services may be).

For this study I will map the concept of neighborhood to Paris’ *arrondissements*, which are administrative divisions inside the city of Paris. This concept makes it easier to retrieve data already confined to a particular neighborhood from official sources.

By the end of this report, someone with the intent of buying/renting a property in Paris, will be able to choose between one of the twenty neighborhoods (*arrondissements*) according to the density of services available and the type of services.

2 Data

When acquiring data for solving a given problem, there are two main characteristics that are crucial for the project's success:

- Data accuracy
- Data freshness

Data accuracy is most probably guaranteed by obtaining data from reliable and official sources. Freshness it's not always possible but it's desirable, otherwise there's the risk of trying to address a problem with outdated data.

For this study I tried to guarantee these characteristics when researching for potential sources.

From the beginning of this study, it was my intention to display means of visualizing the neighborhoods on a map and to achieve that I would need geographic information to help me establish the boundaries of each neighborhood. I managed to obtain a GeoJSON file from https://www.data.gouv.fr [2], a governmental source with data provided by Paris City Hall (data from 2019).

Since the GeoJSON file only defines boundaries but not the coordinates of a central point of the neighborhood, I obtained a second data set from https://opendata.paris.fr [3] with data also provided by the Paris City Hall (data from 2013, since there were no changes since then). Besides the coordinates of the central point, this source also provided the name and number of each neighborhood (each *arrondissement* has a number that administratively identifies it).

The above data sources allowed me to have identification, borders and coordinates of Paris' neighborhoods:

1. Louvre
2. Bourse
3. Temple
4. Hôtel-de-Ville
5. Panthéon
6. Luxembourg
7. Palais-Bourbon
8. Élysée
9. Opéra
10. Entrepôt

11. Popincourt
12. Reuilly
13. Gobelins
14. Observatoire
15. Vaugirard
16. Passy
17. Batignolles-Monceau
18. Buttes-Montmartre
19. Buttes-Chaumont
20. Ménilmontant

To retrieve information about the services provided in each neighborhood, the FourSquare API was used. I defined a radius around a central point of each neighborhood and explored the services provided within that radius for some specific categories (the ones that might bring most value to a prospective buyer/renter):

1. Arts & Entertainment
2. College & Univerity
3. Food
4. Nightlife Spot
5. Outdoors & Recreation
6. School
7. Police Station
8. Fire Station
9. Post Office
10. Parking
11. Medical Center
12. Shop & Service
13. Travel & Transport (manipulated after extraction to eliminate the *Travel* component)
14. Spiritual Center

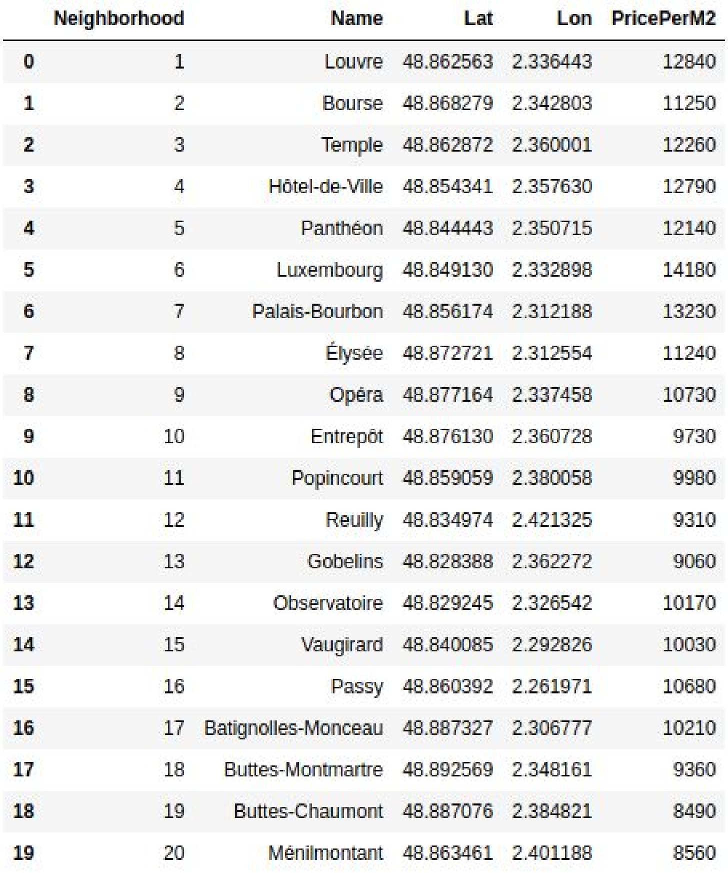
Finally I needed property price data per Paris' neighborhood. Since I was unable to find a reliable and updated source for rental prices, I had to work with sale prices per m^2 of used properties, which is still reliable for this purpose since rental prices tend to be proportional to sale prices.

This data was scrapped (there was no file available for download) from https://droit-finances.commentcamarche.com [4], with data provided by Paris Chamber of Notaries as a result of effective property sales on 2019 third quarter.

3 Methodology

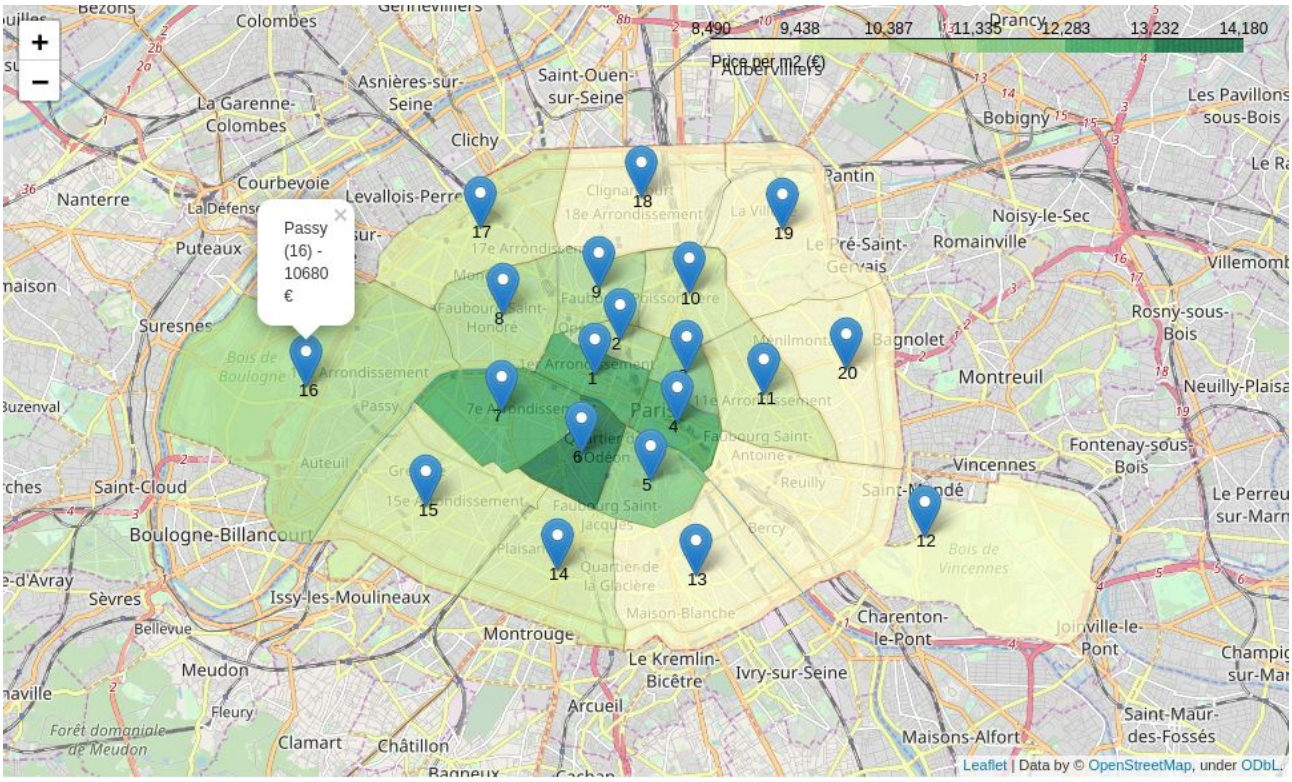
Having all the data sources identified, it was time to start building a Pandas dataframe by selecting relevant data from those sources and by combining features to obtain more meaningful features.

The initial dataframe was assembled with neighborhood information (IDs, names and coordinates) from Paris City Hall and prices per m^2 from Paris Chamber of Notaries. It had the just a fraction of the intended data and it looke like this:



	Neighborhood	Name	Lat	Lon	PricePerM2
0	1	Louvre	48.862563	2.336443	12840
1	2	Bourse	48.868279	2.342803	11250
2	3	Temple	48.862872	2.360001	12260
3	4	Hôtel-de-Ville	48.854341	2.357630	12790
4	5	Panthéon	48.844443	2.350715	12140
5	6	Luxembourg	48.849130	2.332898	14180
6	7	Palais-Bourbon	48.856174	2.312188	13230
7	8	Élysée	48.872721	2.312554	11240
8	9	Opéra	48.877164	2.337458	10730
9	10	Entrepôt	48.876130	2.360728	9730
10	11	Popincourt	48.859059	2.380058	9980
11	12	Reuilly	48.834974	2.421325	9310
12	13	Gobelins	48.828388	2.362272	9060
13	14	Observatoire	48.829245	2.326542	10170
14	15	Vaugirard	48.840085	2.292826	10030
15	16	Passy	48.860392	2.261971	10680
16	17	Batignolles-Monceau	48.887327	2.306777	10210
17	18	Buttes-Montmartre	48.892569	2.348161	9360
18	19	Buttes-Chaumont	48.887076	2.384821	8490
19	20	Ménilmontant	48.863461	2.401188	8560

Although it had only pretty basic information, in conjunction with the GeoJSON (from Paris City Hall) was enough to depict on a Folium map all neighborhoods with a color coded layer that easily allowed to see property prices along neighborhoods. I also added some popups displaying the neighborhood's name and number, with the price per m^2 in Euro.



This visualization helped me reinforce what is already common knowledge: centrality is one of the factors that most contribute to higher properties' prices. But the real question is if this behavior happens just for centrality sake or is it due to the services provided on central neighborhoods (or the density of services)? If we compare *Passy* (16) with *Gobelins* (13), they are both peripheral neighborhoods (*Passy* maybe even more than *Gobelins*), but the price per m^2 in *Passy* is almost 18% higher than in *Gobelins*. Is the availability of services the key factor to justify this price difference?

To help answering these questions it was necessary to identify what services were provided by each neighborhood and to accomplish this I needed to retrieve this data from FourSquare API and merge it into my dataframe. I decided to use the *explore* endpoint of the API, that allows users to search for venues on a given area, but since I was not interested in all kind of venues I would need to complement the search with a *categories* endpoint to limit the kind of venues with interest to this study. To initiate this data retrieval I would have to answer these questions:

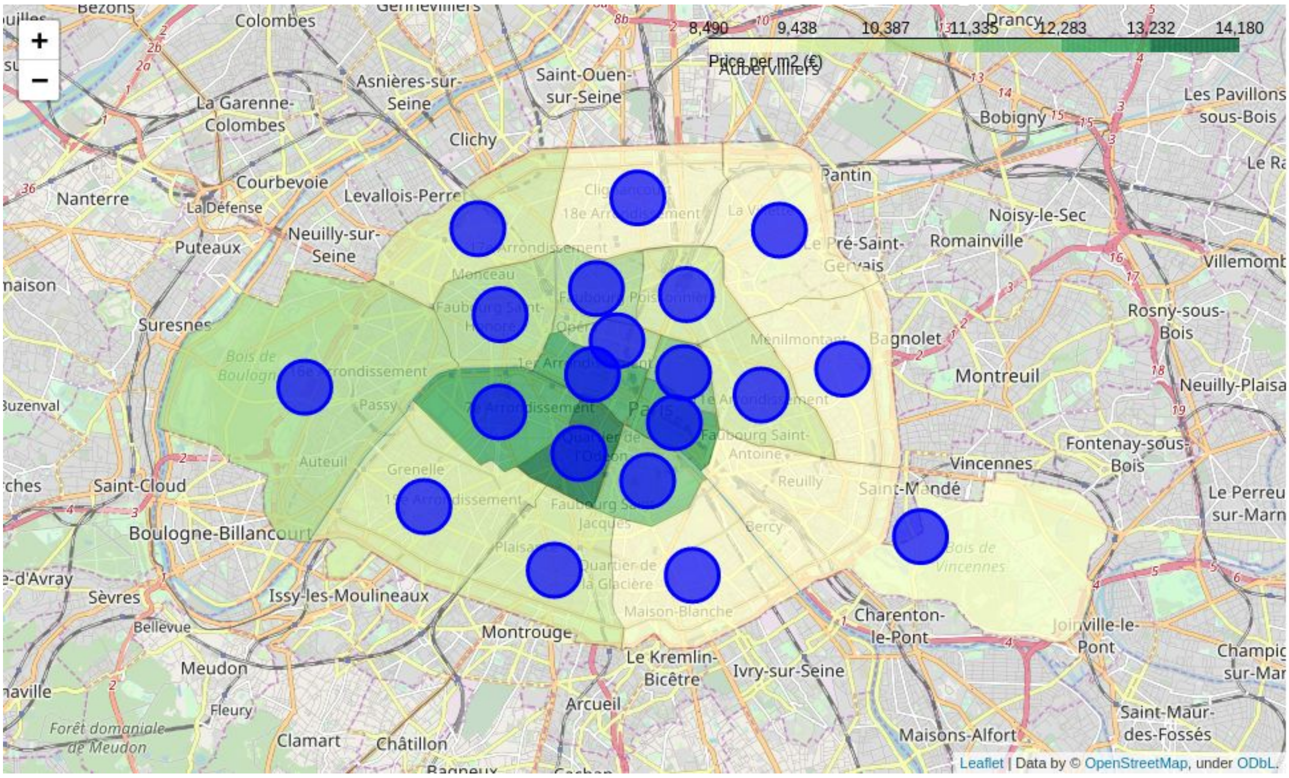
- What kind of venues would be meaningful to this study?
- What radius should I use for the search area?
- Should I limit the number of venues and what should that value be?

The FourSquare documentation has a list of predefined categories and subcategories that are intended to use with the *categories* endpoint. After studying the documentation I opted for a group of categories that fitted what I considered to be basic pillars of what most people look for when choosing a neighborhood: security, mobility, education, entertainment and services. The following table shows how FourSquare categories fall into this group:

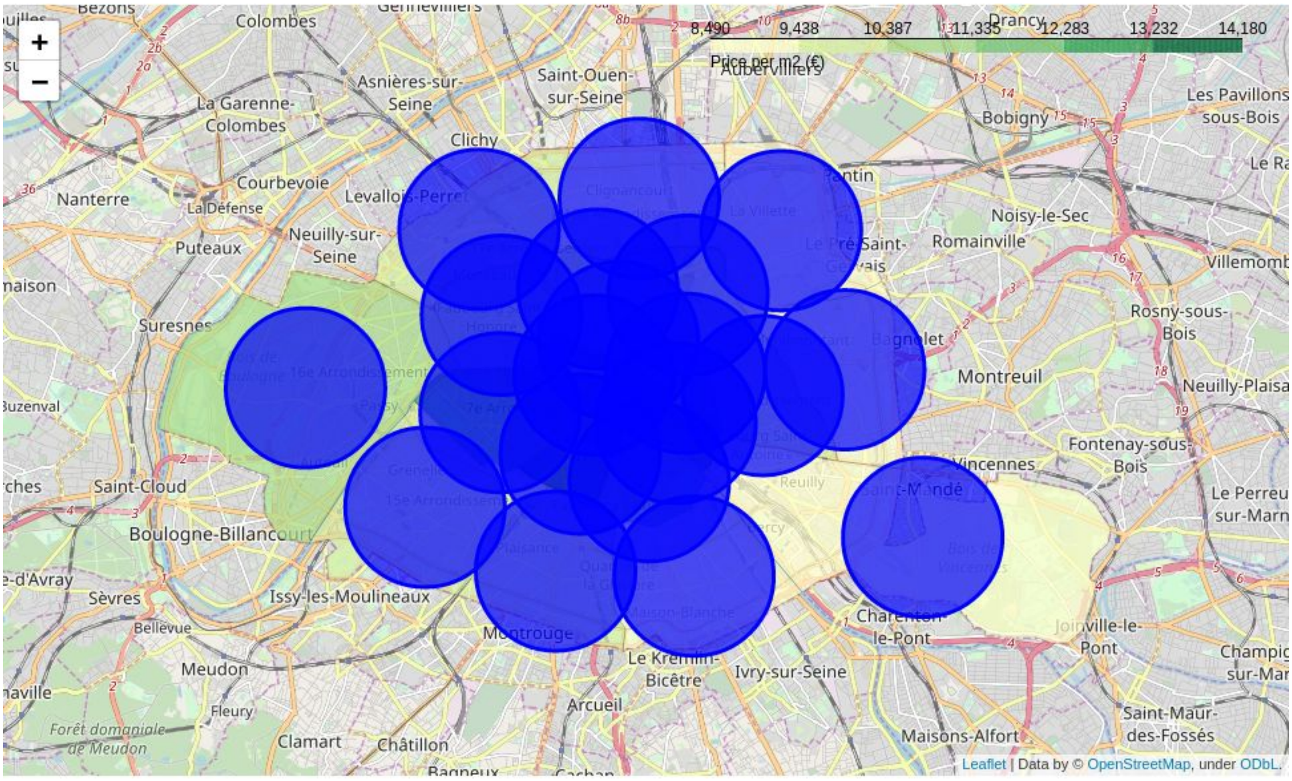
Security	Mobility	Education	Entertainment	Services
Police Station	Parking	College & University	Arts & Entertainment	Post Office
Fire Station	Transport	School	Food	Shop & Service
Medical Center			Nightlife Spot	Spiritual Center
			Outdoors & Recreation	

FourSquare doesn’t have a *Transport* category per se but I used the *Travel & Transport* category and then eliminated all the venues that were not considered as *Transport* (such as hotels, B&Bs, hostels, travel agencies, etc.).

When it came to define a radius for the search area I had several doubts about it. If I picked a small radius I would be unable to capture the essence of peripheral neighborhoods, that are typically bigger than central neighborhoods as we can see in the following map, depicting a 500m radius:



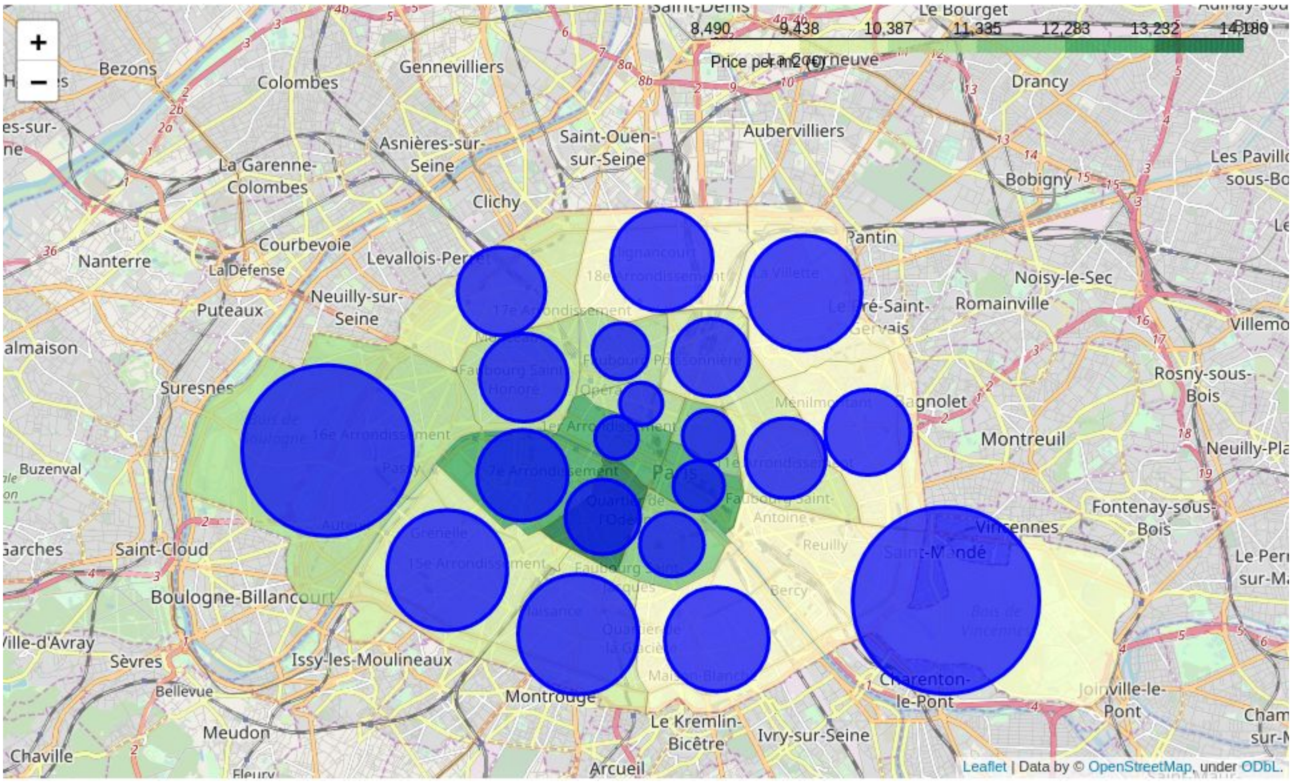
On the other end, choosing a bigger radio that would capture the peripheral neighborhoods features, would result on severe overlapping for central neighborhoods. The following map depicts a radius of 1500m and it’s obvious the advantage of covering peripheral neighborhoods and the disadvantage of the overlapping:



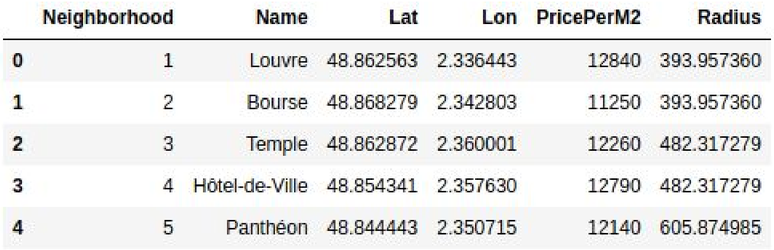
My final decision was to use a dynamic radius that would adjust to the neighborhood's size. To define each neighborhood's radius I used the following algorithm:

1. Build a function that calculates the distance between two points given its coordinates (using the haversine formula [5])
2. Build a distance matrix where indexes and columns are all the central points of each neighborhood and the intersection of each two points is the result of applying the haversine formula to it
3. Choose the minimum distance for each neighborhood (representing the distance to the central point of the closest neighborhood) and divide it by two, since we want the radius to stay at the midpoint between neighborhoods to avoid overlap

This algorithm is not perfect because it doesn't optimize the radius, it just chooses the maximum radius that guarantees the nonexistence of overlapping. However it seems to me a better approach than specifying a fixed radius that will be too big for some neighborhoods and to small for others. The following map shows the dynamic radius applied. You will see that for irregularly shaped neighborhoods there will be a worse coverage but in general it's an improvement over the fixed radius approach.



The radius was also included on the dataframe since it would be used later to calculate the density of services.



	Neighborhood	Name	Lat	Lon	PricePerM2	Radius
0	1	Louvre	48.862563	2.336443	12840	393.957360
1	2	Bourse	48.868279	2.342803	11250	393.957360
2	3	Temple	48.862872	2.360001	12260	482.317279
3	4	Hôtel-de-Ville	48.854341	2.357630	12790	482.317279
4	5	Panthéon	48.844443	2.350715	12140	605.874985

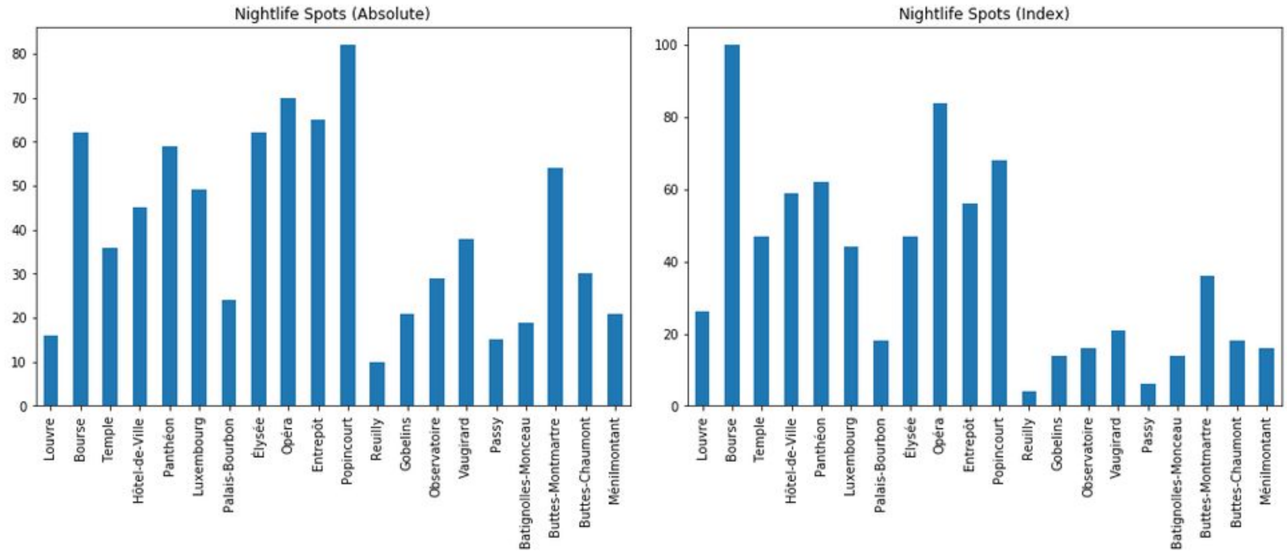
The last decision regarding FourSquare’s API was the limit of venues to retrieve and I choose to define it as one hundred per category, because it seems a reasonable compromise between the amount of data retrieved and the time it takes to retrieve it.

The retrieved data was grouped by neighborhood and category to allow me to add to the dataframe a column per category (named after the category) with the number of venues per neighborhood.

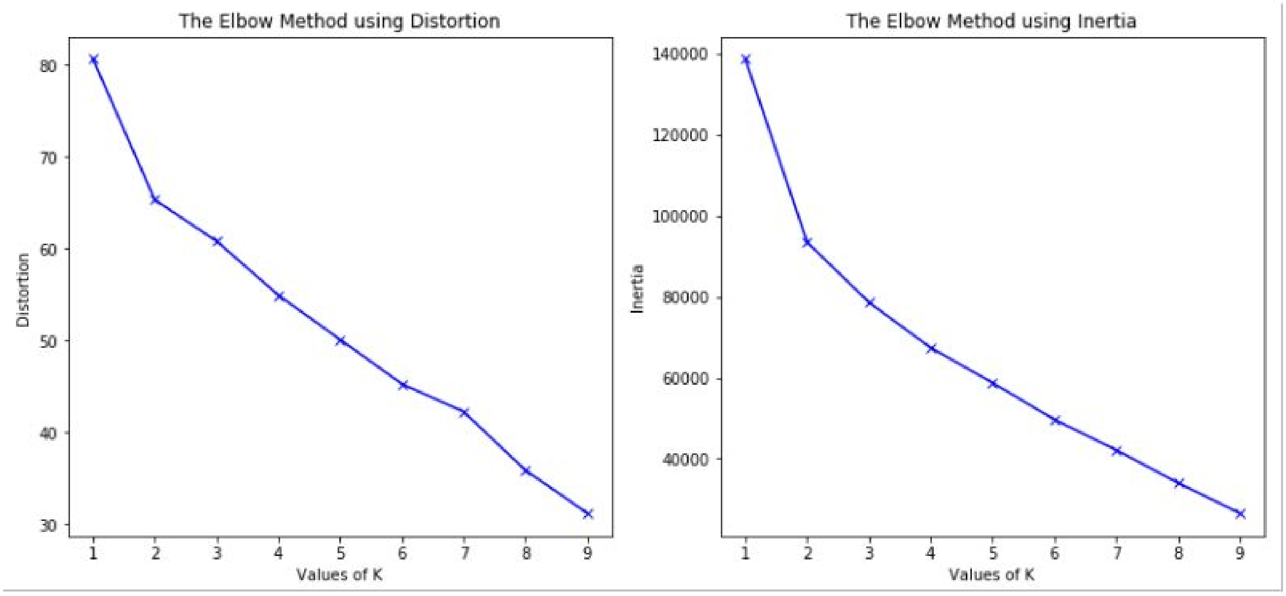
The next step was to turn this data into meaningful data, since it doesn’t make much sense compare absolute values. Having 82 nightlife spots doesn’t make *Popincourt* (11) a more suitable neighborhood to explore nightlife than *Bourse* (2) with its 62 spots, because the explored area was roughly half the size of *Popincourt* (11).

To overcome this issue I created new features that represented the density of services per category in each neighborhood. The value assigned to this new feature took the form of an index ranging from 1 to 100 (number of venues divided per radius and normalized to fit this range).

We can clearly see the difference between comparing absolute values or index values on the following plots regarding the category *Nightlife Spots*:



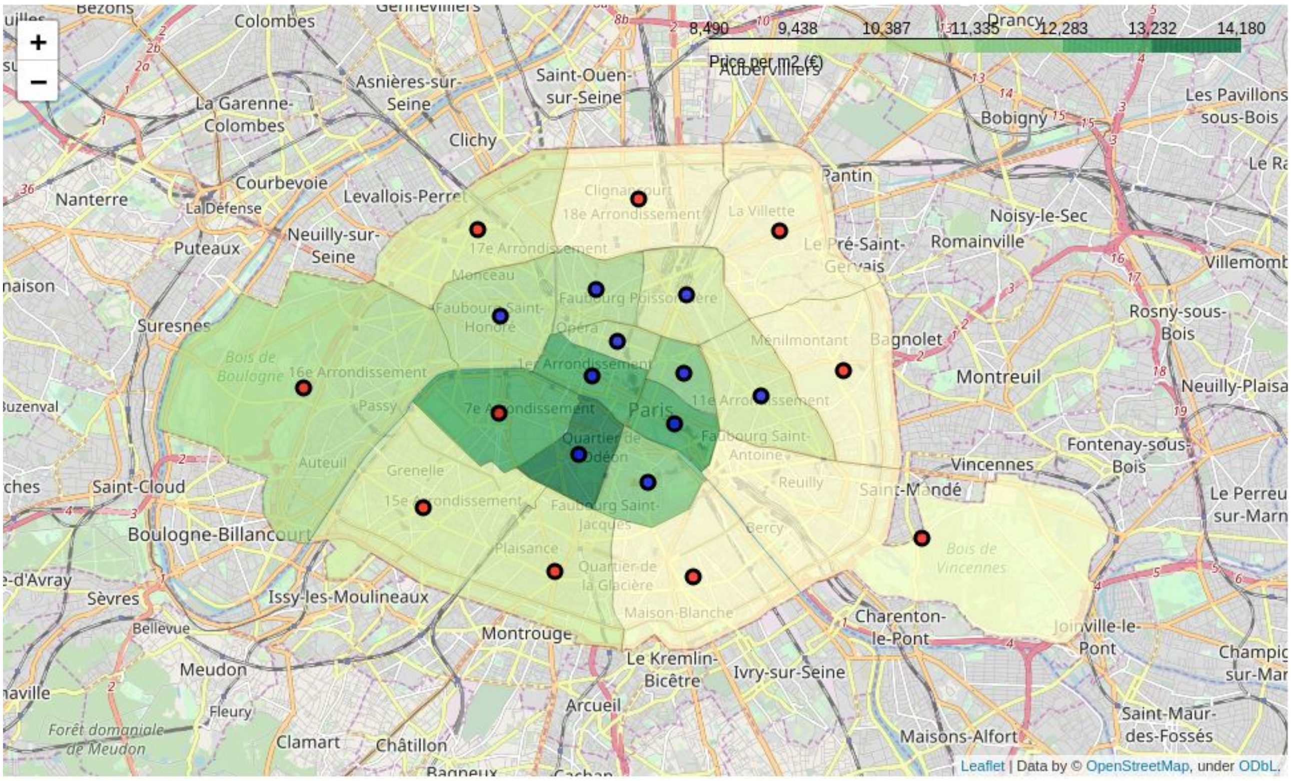
Proceeding with the analysis I decided to apply K-means clustering to the neighborhoods based on the indexes previously calculated. To decide how many clusters should the algorithm detect, I used the elbow method that indicated that two (either using distortion or inertia) clusters should be formed.



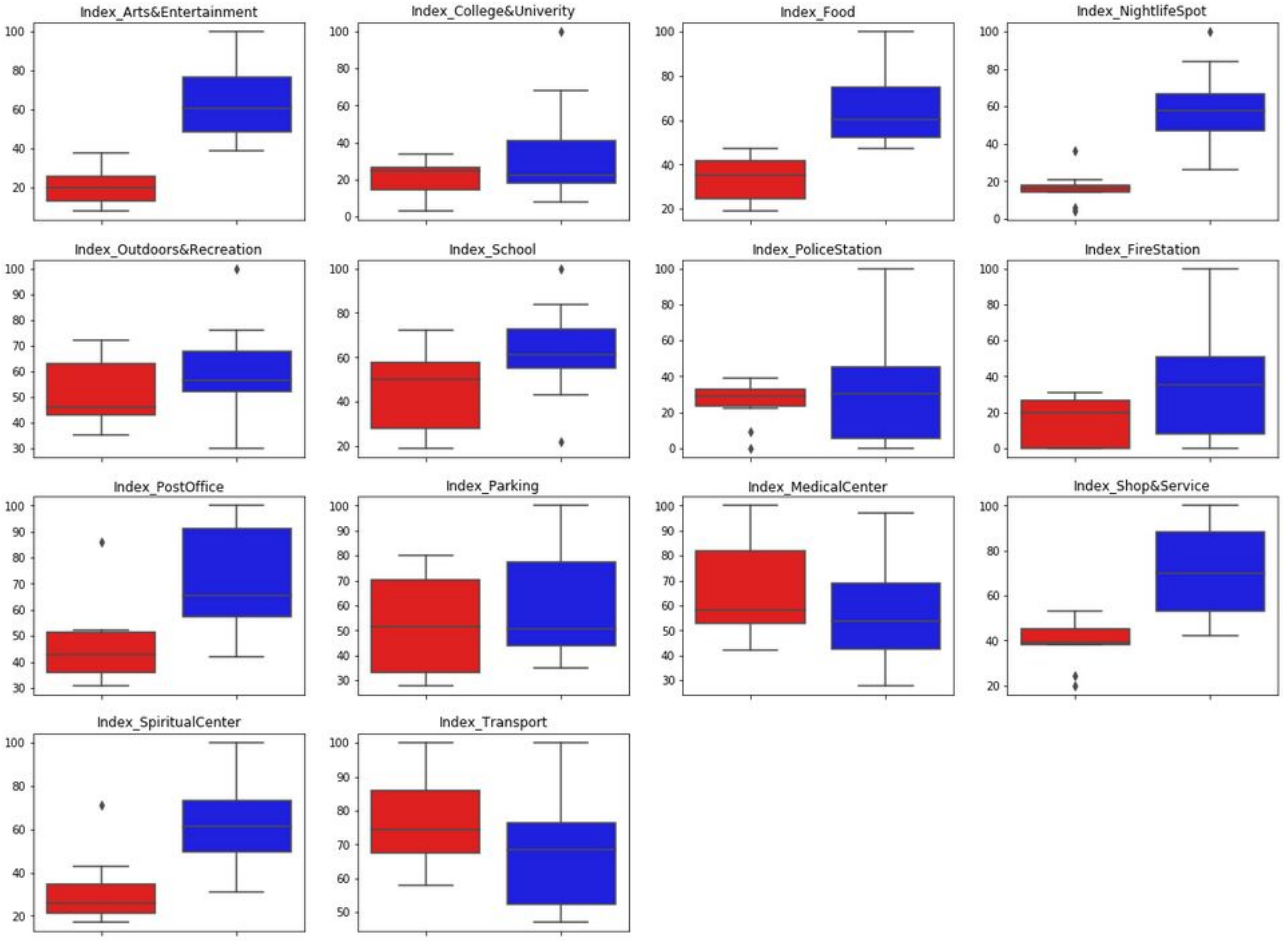
After applying K-means the two clusters formed were surprisingly (or maybe not) an almost perfect map of peripheral and central neighborhoods. The only central neighborhood that was placed on the peripheral neighborhood's cluster was *Palais-Bourbon* (7).

The interesting about this was that no geographical or price information was provided to the K-means algorithm and even though, solely based on the density of services, a separation was made between inner and outer neighborhoods (with the exception above mentioned).

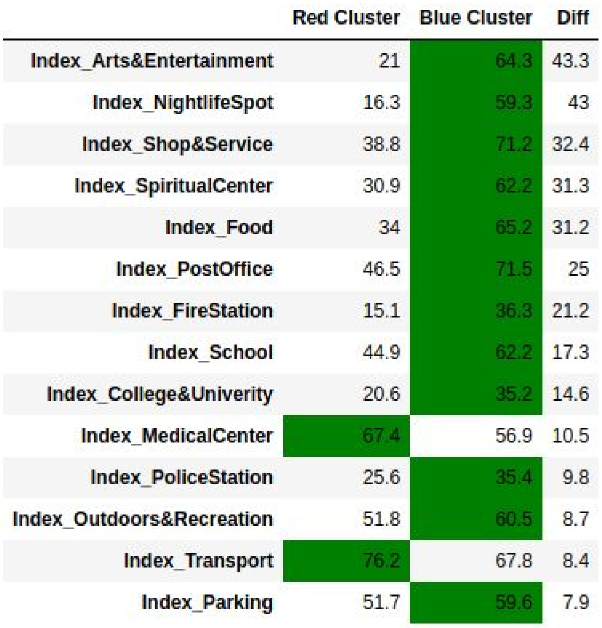
From now on, on this document, I will refer to the clusters as the Red Cluster and the Blue Cluster according to the following map (I will also use the same color scheme on every plot):



The clustering information was also added to the dataframe and several plots where used to analyze and compare category indexes between neighborhoods and clusters.

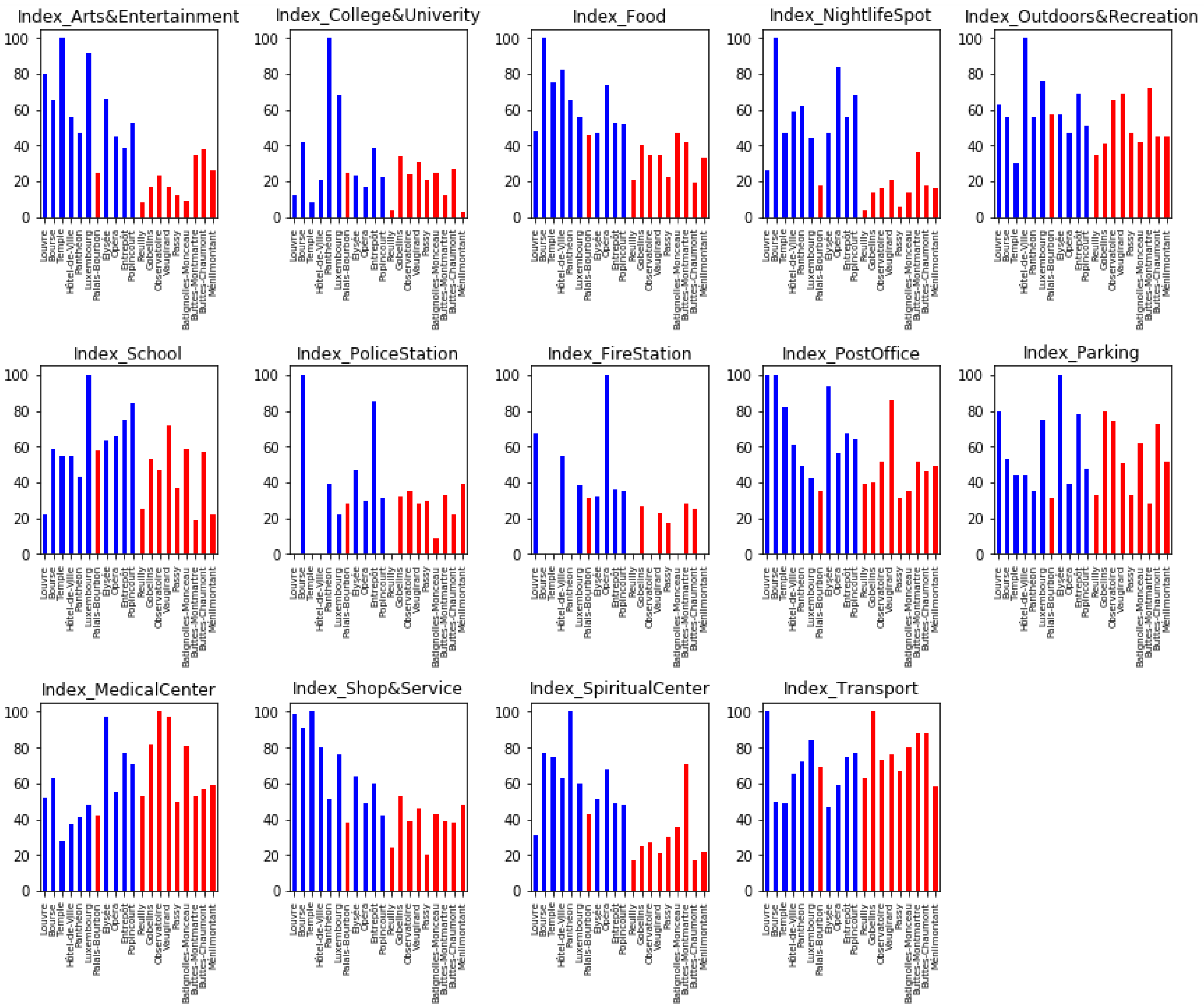


I also decided to summarize the difference between the mean index values per cluster and category.



	Red Cluster	Blue Cluster	Diff
Index_Arts&Entertainment	21	64.3	43.3
Index_NightlifeSpot	16.3	59.3	43
Index_Shop&Service	38.8	71.2	32.4
Index_SpiritualCenter	30.9	62.2	31.3
Index_Food	34	65.2	31.2
Index_PostOffice	46.5	71.5	25
Index_FireStation	15.1	36.3	21.2
Index_School	44.9	62.2	17.3
Index_College&Univerity	20.6	35.2	14.6
Index_MedicalCenter	67.4	56.9	10.5
Index_PoliceStation	25.6	35.4	9.8
Index_Outdoors&Recreation	51.8	60.5	8.7
Index_Transport	76.2	67.8	8.4
Index_Parking	51.7	59.6	7.9

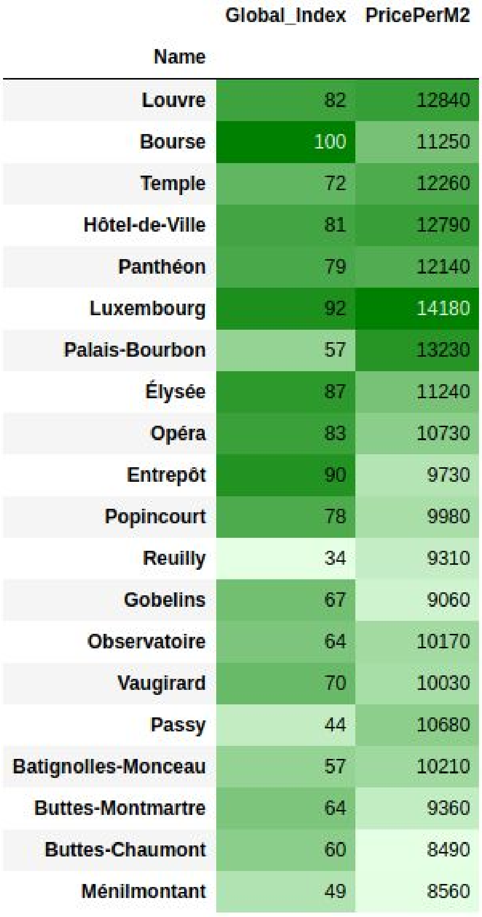
The above boxplots/table were useful to compare clusters but for comparing particular neighborhoods lacked detail, so I opted to use some plots that allowed to do that kind of comparison.



The last step of the analysis was to compare neighborhood's price per m^2 against the offerings in terms of services. To achieve that I created a new feature that was an aggregate index that resumes all the category's indexes in a single number. This index is nothing more than the sum of all indexes, normalized to a range of values between 0 and 100.

On a customer tailored model a function could be built to provide distinct weights to categories, according to what a customer most values in a neighborhood.

To better visualize this index against price per m^2 I introduced the following and table:



Name	Global_Index	PricePerM2
Louvre	82	12840
Bourse	100	11250
Temple	72	12260
Hôtel-de-Ville	81	12790
Panthéon	79	12140
Luxembourg	92	14180
Palais-Bourbon	57	13230
Élysée	87	11240
Opéra	83	10730
Entrepôt	90	9730
Popincourt	78	9980
Reuilly	34	9310
Gobelins	67	9060
Observatoire	64	10170
Vaugirard	70	10030
Passy	44	10680
Batignolles-Monceau	57	10210
Buttes-Montmartre	64	9360
Buttes-Chaumont	60	8490
Ménilmontant	49	8560

This table allows us to pose questions such as why should we pay so much for property in *Palais-Bourbon* or why a neighborhood with so much to offer as *Entrepôt* has such a competitive price.

4 Results

When we look at property prices distribution in Paris it’s quite clear that central neighborhoods tend to have much higher prices than other neighborhoods. This is the “normal” behavior everywhere and not only in Paris.

Understanding why this happens is the key for someone that wants to move into the city and needs to be sure what will be the benefits of paying considerably more to stay in a central neighborhood.

There were some characteristics that were considerably more widespread in almost all central neighborhoods when compared to the remaining:

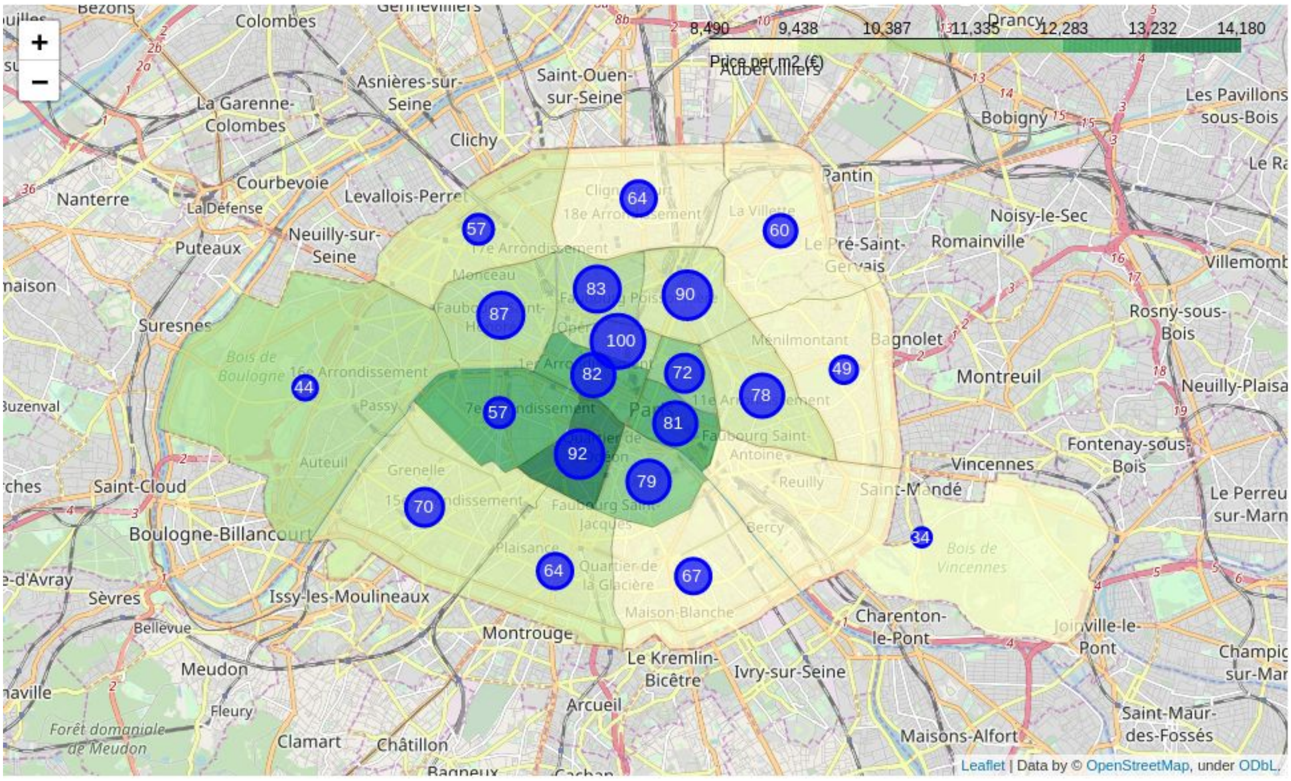
- Arts and entertainment facilities
- Nightlife establishments
- Spiritual centers
- Restaurants and other food related establishments

- Shops and services (ATMs, banks, supermarkets, lawyers, IT services, etc.)

The above represent the most significant differences, but central neighborhoods have better density of services in almost all categories, with the following exceptions (with little relevance due to minimal differences):

- Medical centers
- Transport

The choice of a particular neighborhood over others can be made by analyzing the charts provided in this report in accordance to each one’s preferences, but as more general view I provide a map with a global index superimposed over the price per m^2 of each neighborhood.



This map allows to promptly discard some of the neighborhoods before digging into all the detailed information about each category of services.

5 Discussion

This study supports, in general, the idea that the higher prices of central neighborhoods are somehow justified by the range and density of services offered. However there are three particular cases in this study that deserve a special mention.

Palais-Bourbon is a neighborhood with a global index of 57 (16th in the ranking) but despite that is the 2nd most expensive neighborhood. Why is that? Taking a closer look we see that more than one third of this neighborhood is comprised of the Eiffel Tower, Champ de Mars (a huge park), Les Invalides (huge museum and park), several governmental buildings (military school, UNESCO, ministries, etc.) and several small parks. What this means is that the density of services is highly compromised because it takes into account the whole neighborhood but one third of it is “unusable”.

Another interesting case is *Passy* with a global index of 44 (19th in the ranking) but despite that is the most expensive peripheral neighborhood (10th most expensive when including all neighborhoods). Are we facing a similar problem to *Palais-Bourbon*? In fact we are. This neighborhood has more than half of its area occupied by the biggest green area in Paris, the Bois de Boulogne.

The last case is *Reuilly* with the worst global index across Paris (34). This case suffers from two different problems: more than half of its area is a green area and since this neighborhood as a very irregular shape, the radius excluded a considerable part of it (and excluded area is of the urban type).

These three neighborhoods are somehow outliers in this dataset and deserved this closer look.

If I was asked to make suggestions based solely on this study, these would be my picks (the three cases mentioned above are excluded from the recommendations due to what was explained before):

- *Bourse* is a central neighborhood that ranked 1st in the global index and when it comes to price it occupies the middle of the table as the 7th most expensive. This is the best buy when it comes to service density
- *Entrepôt* is located half way between central and peripheral neighborhoods, but with a global index of 90 (3rd in the ranking) and a price per m^2 of 9730€ (6th less expensive), it’s the right buy with great ratio services/price

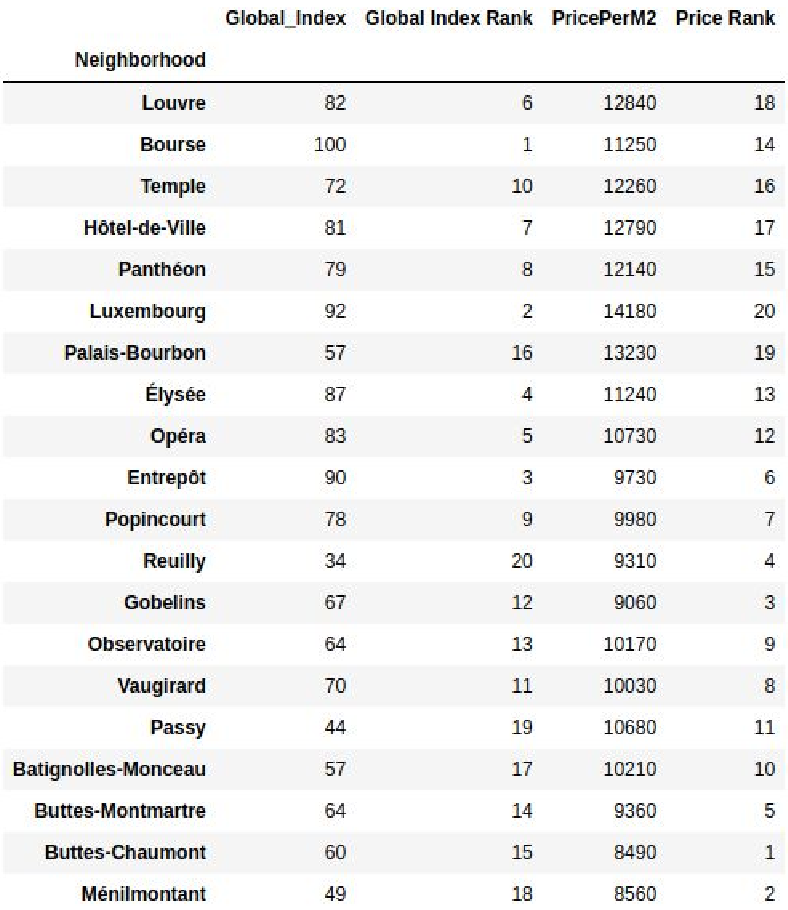
6 Conclusion

With so many variables at stake it’s not easy to build a model to accomplish an apparently easy objective of suggesting a particular neighborhood to a future tenant that could fulfill its needs.

One of the most difficult aspects to achieve it’s to obtain the services constrained to a particular neighborhood and even more difficult to capture the plenitude of the neighborhood, since the venues are searched in circular areas which hardly map to the real boundaries.

Even with these constraints I think that this study can provide some guidelines for future tenants in Paris, particularly for those who do not know the city and don’t have the time to take a deep dive into it. The document provides some detail about the types and density of services available in each neighborhood and at the same time a more general approach for those who don’t want to drill into each category.

For those who just want a quick reference here’s a final table with the global index, price per m^2 and respective rankings per neighborhood.



	Global_Index	Global Index Rank	PricePerM2	Price Rank
Neighborhood				
Louvre	82	6	12840	18
Bourse	100	1	11250	14
Temple	72	10	12260	16
Hôtel-de-Ville	81	7	12790	17
Panthéon	79	8	12140	15
Luxembourg	92	2	14180	20
Palais-Bourbon	57	16	13230	19
Élysée	87	4	11240	13
Opéra	83	5	10730	12
Entrepôt	90	3	9730	6
Popincourt	78	9	9980	7
Reuilly	34	20	9310	4
Gobelins	67	12	9060	3
Observatoire	64	13	10170	9
Vaugirard	70	11	10030	8
Passy	44	19	10680	11
Batignolles-Monceau	57	17	10210	10
Buttes-Montmartre	64	14	9360	5
Buttes-Chaumont	60	15	8490	1
Ménilmontant	49	18	8560	2

Sources:

[1] https://www.globalpropertyguide.com/most-expensive-cities

[2] https://www.data.gouv.fr/en/datasets/r/4765fe48-35fd-4536-b029-4727380ce23c

[3] https://opendata.paris.fr/explore/dataset/arrondissements/download/?format=csv&timezone=Europe/London&lang=fr&use_labels_for_header=true&csv_separator=%3B

[4] https://droit-finances.commentcamarche.com/faq/7409-immobilier-a-paris-prix-au-m2-des-arrondissements

[5] https://en.wikipedia.org/wiki/Haversine_formula