# Identifying potential areas for investment, analysis of Paris

# I. Introduction

## A. Business Problem

The city of Paris attracts more than 10 million tourists per year, it is well known for its museums, architecture, shopping and gastronomy. As a consequence, the city have seen its average house price grow by 9.1% in 2018 with an average property price around 625 000 USD. Still, the demand is still greater than the supply and is driven by foreign investors traditionally from US and UK wealthy individual, and recently from Chinese and Indians who are trusting the most pleasant areas of Paris, mainly the 6th, 7th and 8th arrondissement.

In addition, as there is a shortage of disposable land to build, some neighborhood are experiencing positive mutations becoming trendy and much sought after. The best example is the third arrondissement better known as "le Marais" which was literally in ruins and with a high crime rate in the late 80's, though becoming one of the most expensive area in Paris as of today.

In this context and as an investor in real estate (individual or corporation) or more generally for starting a business (restaurant, shops…) it is important to understand what are the characteristics that drive the housing price and the business opportunity of a particular area.

Moreover, based on those findings, is it possible to determine which arrondissement could be the new "Marais" in the near future. Any investor trying to set up a business (restaurants, shops...) in a dynamic area could also use this analysis.

## B. Data description

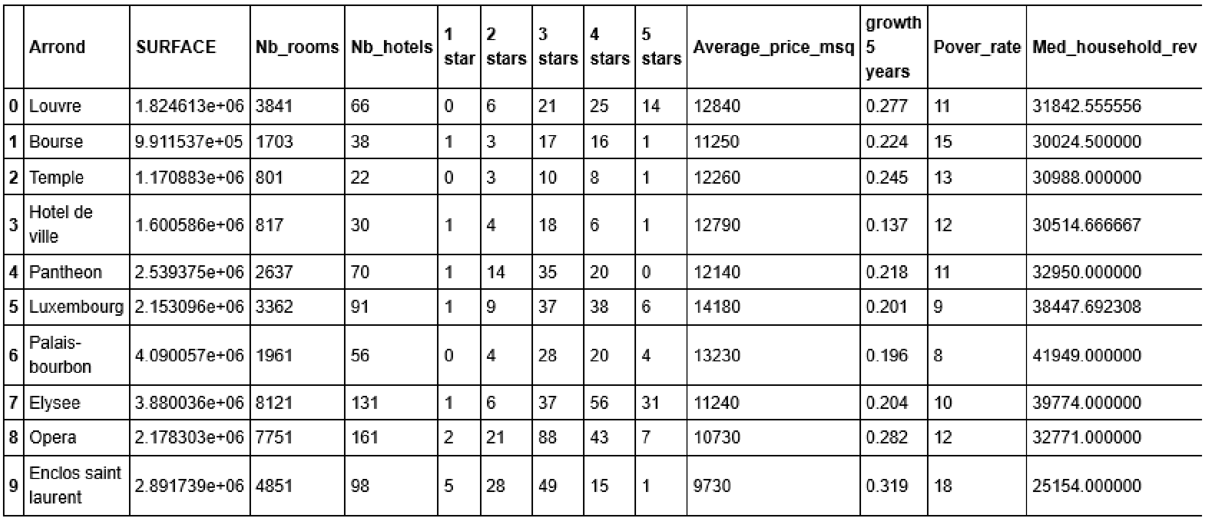
### 1. *Collecting economic and social indicators*

Before using foursquare to obtain venues for each arrondissement[1], I collected from different sources (*Atelier Parisienn de l'urbanisme (APUR), Institut national de la statistique et des etudes économiques (INSEE), data.gouv.fr, Chambre des notaires de Paris*) variables which could represent the attractivity of an arrondissement. Starting from the available economic and social indicators for each arrondissement, I have chosen the following variables to discriminate between the 20 arrondissement of Paris:

[1] Equivalent of neighberhood

1. Surface in sq/meters
2. Number of hotels
3. Number of rooms
4. Number of 1 star hotel
5. Number of 2 star hotel
6. Number of 3 star hotel
7. Number of 4 star hotel
8. Number of 5 star hotel
9. Median revenue per household
10. Poverty rate
11. Average price per square meter
12. Evolution of housing price (compound annual growth over 5 years)
13. Official coordinates of the arrondissement (longitude and latitude)

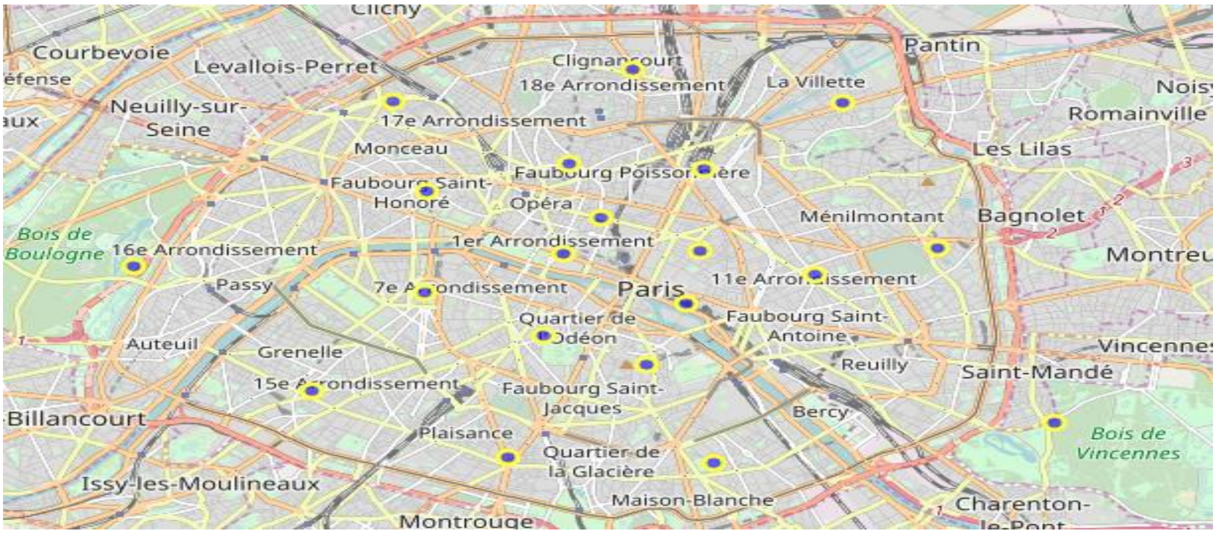
The database has a dimension of 20x14 and looks like the following.



| | Arrond | SURFACE | Nb_rooms | Nb_hotels | 1 star | 2 stars | 3 stars | 4 stars | 5 stars | Average_price_msq | growth 5 years | Pover_rate | Med_household_rev |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Louvre | 1.824613e+06 | 3841 | 66 | 0 | 6 | 21 | 25 | 14 | 12840 | 0.277 | 11 | 31842.555556 |
| 1 | Bourse | 9.911537e+05 | 1703 | 38 | 1 | 3 | 17 | 16 | 1 | 11250 | 0.224 | 15 | 30024.500000 |
| 2 | Temple | 1.170883e+06 | 801 | 22 | 0 | 3 | 10 | 8 | 1 | 12260 | 0.245 | 13 | 30988.000000 |
| 3 | Hotel de ville | 1.600586e+06 | 817 | 30 | 1 | 4 | 18 | 6 | 1 | 12790 | 0.137 | 12 | 30514.666667 |
| 4 | Pantheon | 2.539375e+06 | 2637 | 70 | 1 | 14 | 35 | 20 | 0 | 12140 | 0.218 | 11 | 32950.000000 |
| 5 | Luxembourg | 2.153096e+06 | 3362 | 91 | 1 | 9 | 37 | 38 | 6 | 14180 | 0.201 | 9 | 38447.692308 |
| 6 | Palais-bourbon | 4.090057e+06 | 1961 | 56 | 0 | 4 | 28 | 20 | 4 | 13230 | 0.196 | 8 | 41949.000000 |
| 7 | Elysee | 3.880036e+06 | 8121 | 131 | 1 | 6 | 37 | 56 | 31 | 11240 | 0.204 | 10 | 39774.000000 |
| 8 | Opera | 2.178303e+06 | 7751 | 161 | 2 | 21 | 88 | 43 | 7 | 10730 | 0.282 | 12 | 32771.000000 |
| 9 | Enclos saint laurent | 2.891739e+06 | 4851 | 98 | 5 | 28 | 49 | 15 | 1 | 9730 | 0.319 | 18 | 25154.000000 |

By using the folium method, we can visualize the arrondissement of Paris using the latitude and longitude.

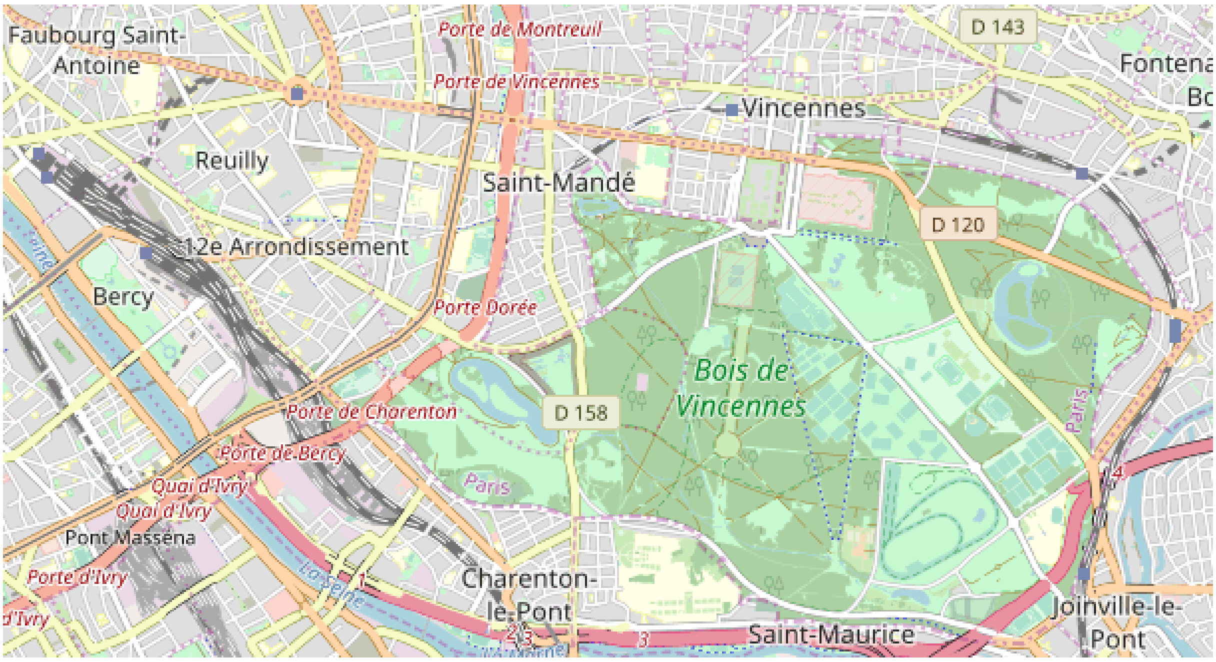
## The city of Paris is composed of 20 arrondissement (neighborhood)



### 2. *Getting venues*

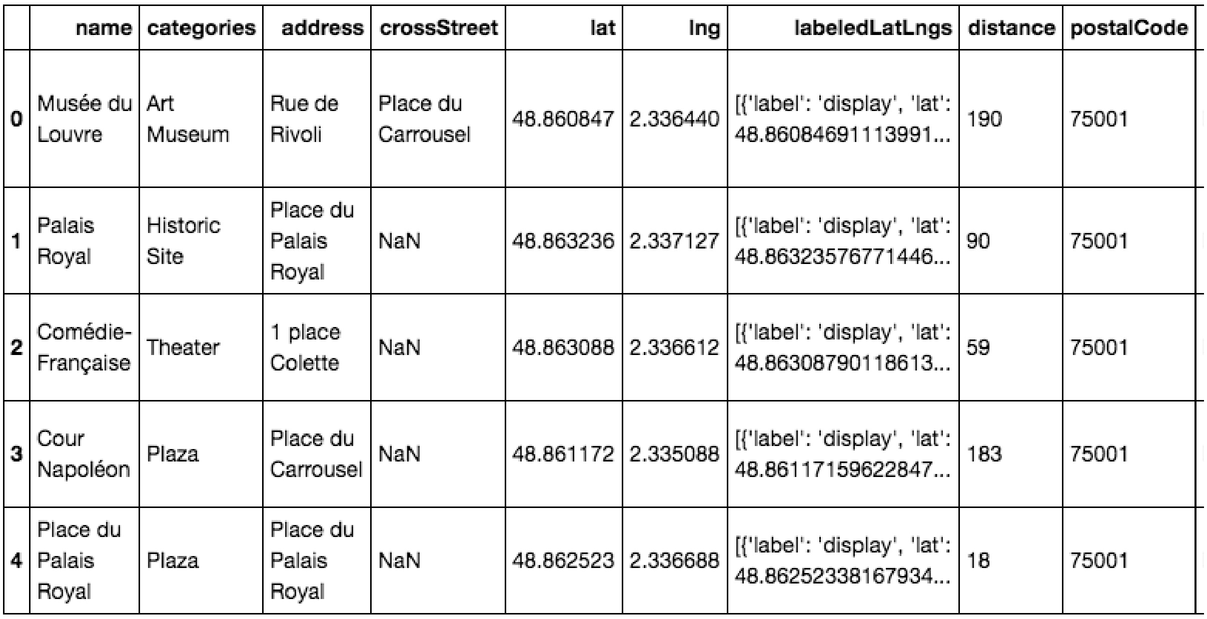
The next step was to add the 50 venues beginning from the center of each arrondissement. There was one adjustment to make relative to the 12th arrondissement (Reuilly), because as you can see below, half of the arrondissement is composed by the "forest of Vincennes". Thus, I subtracted surface of "forest of Vincennes" and moved the center to the modified arrondissement.

**Visualizing "forest of Vincennes" in the 12th arrondissement**



The next step is using foursquare API (2O calls) to get the venues for each arrondissement.

Below is an example of venues call for the first arrondissement.



| | name | categories | address | crossStreet | lat | lng | labeledLatLngs | distance | postalCode |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Musée du Louvre | Art Museum | Rue de Rivoli | Place du Carrousel | 48.860847 | 2.336440 | [{'label': 'display', 'lat': 48.86084691113991... | 190 | 75001 |
| 1 | Palais Royal | Historic Site | Place du Palais Royal | NaN | 48.863236 | 2.337127 | [{'label': 'display', 'lat': 48.86323576771446... | 90 | 75001 |
| 2 | Comédie-Française | Theater | 1 place Colette | NaN | 48.863088 | 2.336612 | [{'label': 'display', 'lat': 48.86308790118613... | 59 | 75001 |
| 3 | Cour Napoléon | Plaza | Place du Carrousel | NaN | 48.861172 | 2.335088 | [{'label': 'display', 'lat': 48.86117159622847... | 183 | 75001 |
| 4 | Place du Palais Royal | Plaza | Place du Palais Royal | NaN | 48.862523 | 2.336688 | [{'label': 'display', 'lat': 48.86252338167934... | 18 | 75001 |

As Paris is very dense, thus having numerous types of venues, I kept the five most appearing categories adding them to the initial database (economic and social indicators of each arrondissement) which is below. The density of the venues implies also that we don't use the number of venues as even by reducing the Limit and the area, the result will be almost the same for all the arrondissement. We have then 20 observations and 19 variables; we will probably have to reduce the dimension before applying the segmentation.

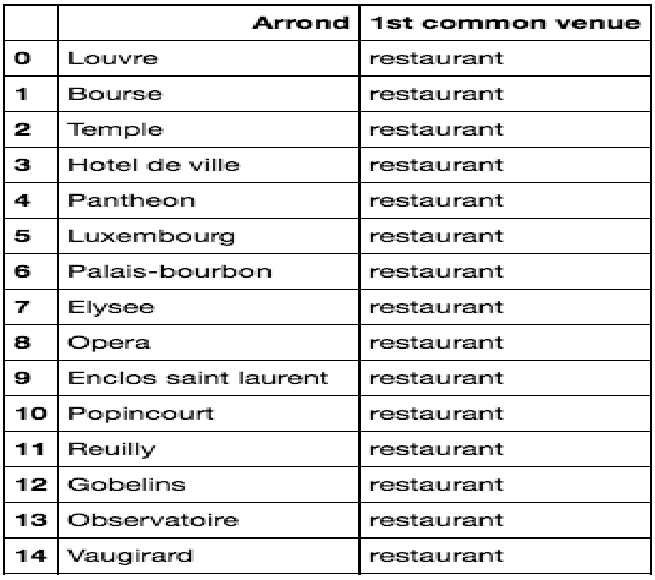
# II. Methodology

## A. Exploratory analysis

Now that we have our database, the first step is to start with an exploratory analysis trough descriptive statistics and data visualization.

First, let us try to reduce the dimension i.e resuming the information contained in our variables if we believe they could be not relevant or redundant. Clearly, the columns containing the coordinates are removed because it does not provide any useful information nor have the power to differentiate between the arrondissement.

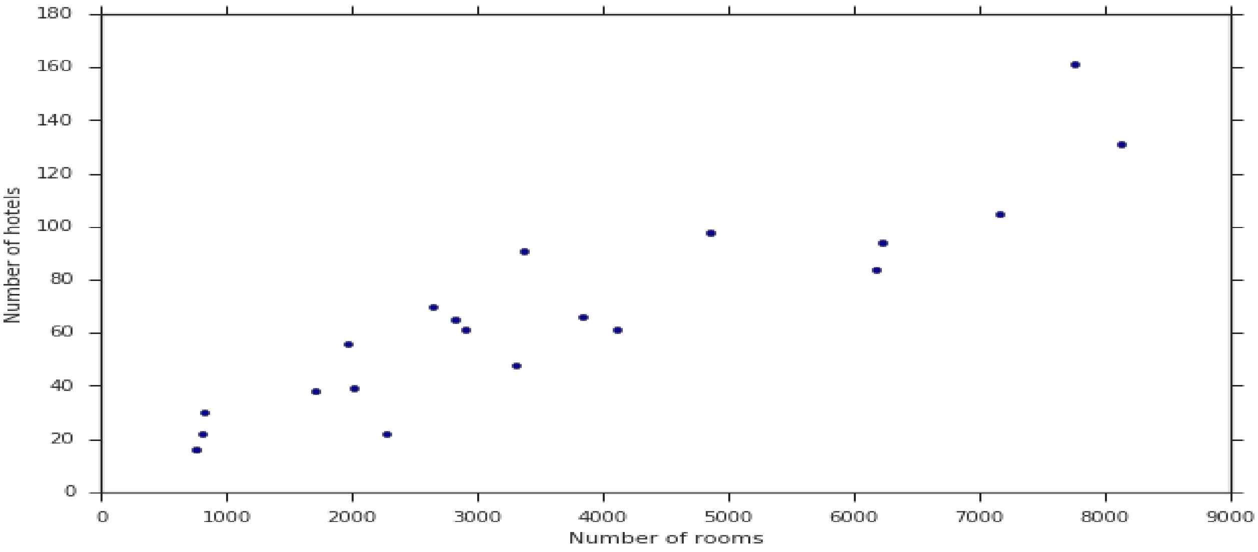
Same thing goes for the column "1st most common venue" as the venue "restaurant" is the same for all the arrondissement.



| | Arrond | 1st common venue |
|---|---|---|
| 0 | Louvre | restaurant |
| 1 | Bourse | restaurant |
| 2 | Temple | restaurant |
| 3 | Hotel de ville | restaurant |
| 4 | Pantheon | restaurant |
| 5 | Luxembourg | restaurant |
| 6 | Palais-bourbon | restaurant |
| 7 | Elysee | restaurant |
| 8 | Opera | restaurant |
| 9 | Enclos saint laurent | restaurant |
| 10 | Popincourt | restaurant |
| 11 | Reuilly | restaurant |
| 12 | Gobelins | restaurant |
| 13 | Observatoire | restaurant |
| 14 | Vaugirard | restaurant |

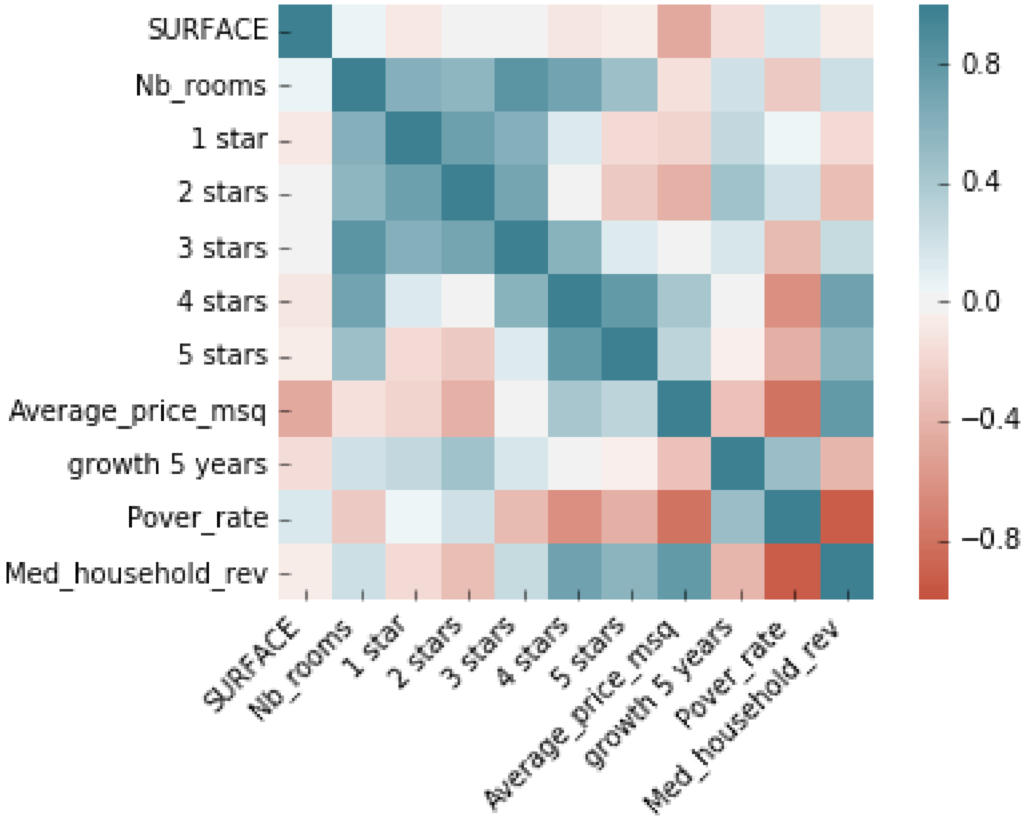
Besides, the number of rooms and the number of hotels are strongly positively correlated (=0.91) and represent the same information, we keep the number of hotels as it is a better indicator of the accommodation capacity.

**Relation between the number of rooms and the number of hotels in an arrondissement**



From the Heatmap correlation, we can see that there are significant positive correlation, between the median household revenue and the number of hotel stars/the average per square meter. There are also negative correlation between the poverty rate and the number of hotels stars/average price per square meter/Median household revenue.

**Display of the correlations between variables**



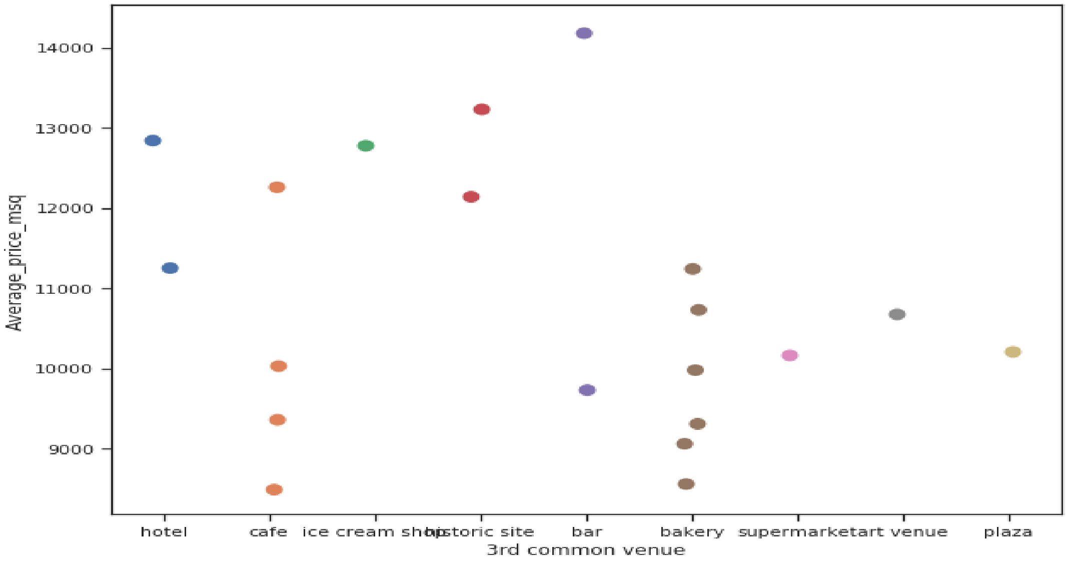
Let us look at their scatterplots to have more insights.

In regards to the Heatmap and the scatter plots (see notebook), we can draw some preliminary and logical conclusions for a hypothetical investor based on the exploratory analysis:

- The higher the median household income in an arrondissement the higher the number of hotels and the price per meter square;
- The lower the poverty rate the higher the business potential of an arrondissement (higher household income, more hotels, higher price per meter square);
- The type of hotels that have the greater capacity of accommodation (number of rooms) are the 3 and 4 stars;
- The higher the surface of an arrondissement the lower the housing price.

For the rest of our study, we will only keep the 2nd most common venue and remove the 3rd, 4rth, 5th most common venue. This choice is justified by plotting the distribution of the type of venue with the other indicators, clearly there is no pattern (see below an example with of the 3rd common venue)

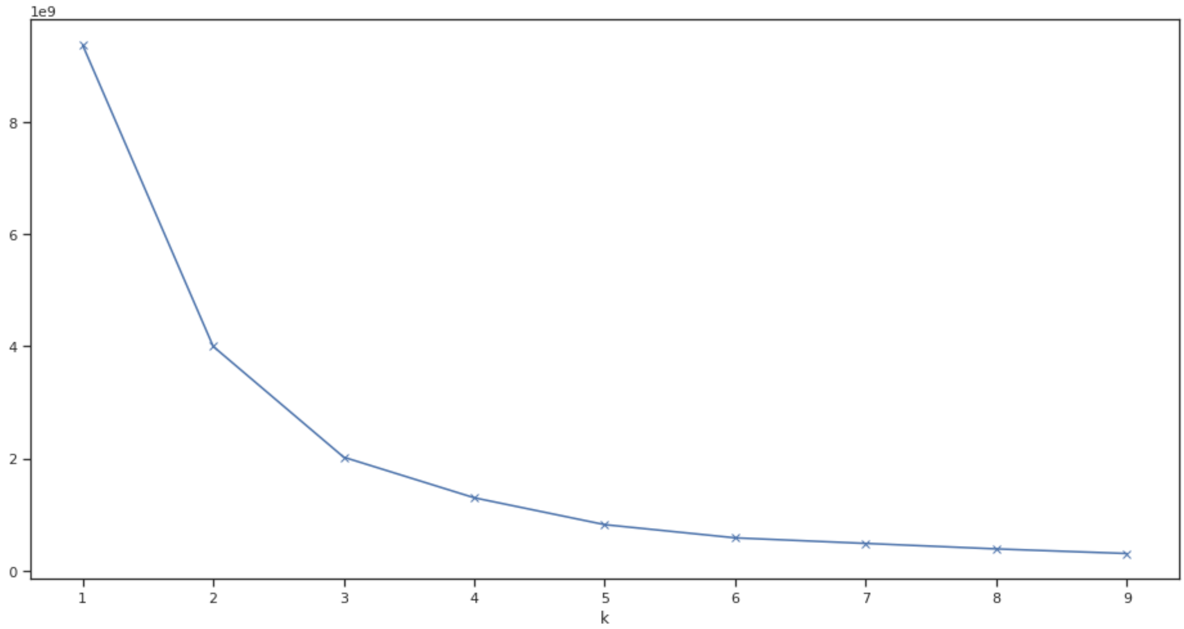
**Distribution of the type of venue and the average price per square meter**



## B. Segmentation

As we don’t have a target variable in our database, we will use unsupervised classification with k-means method. To determine the optimal number of clusters “K”, following is a plot distortion in function of the number of clusters. The optimal number is where there is a bend in the curve at k=3.

**The elbow method showing the optimal K**

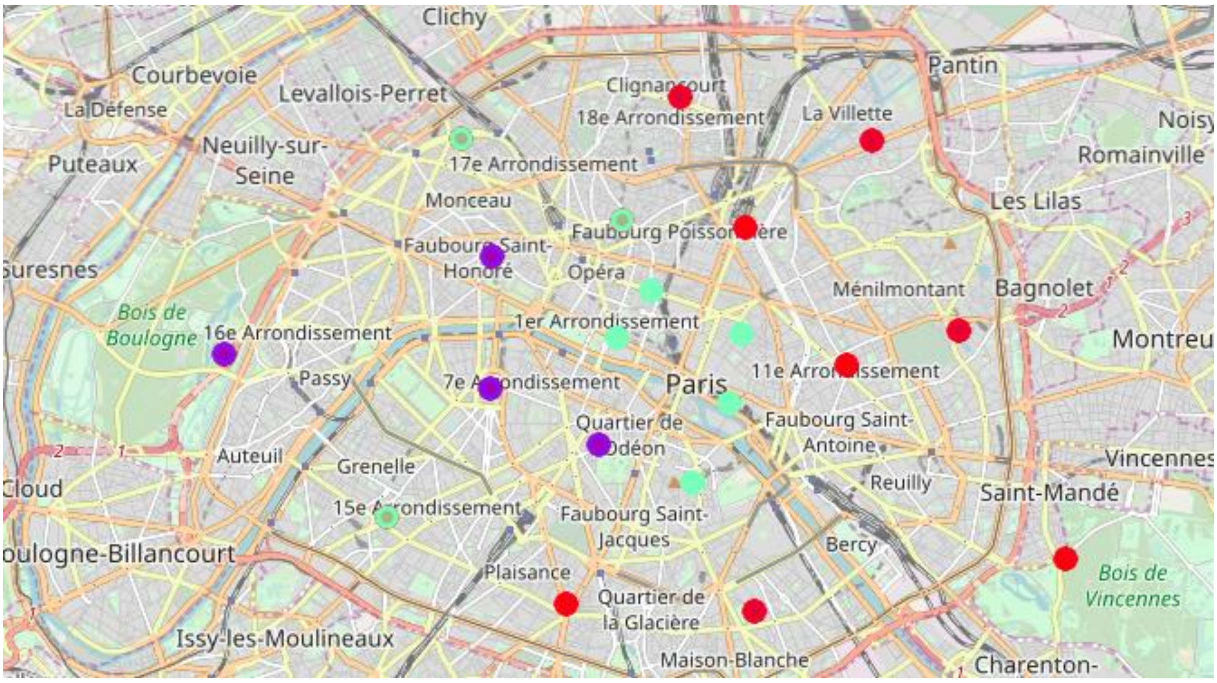


## C. Results

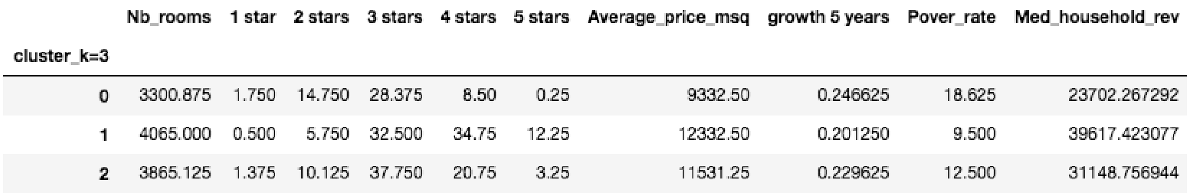
By plotting the three clusters, and before analyzing the characteristics of each arrondissement, we can see that they are geographically separated. There are neighborhoods that are on the east side and outskirts of Paris – cluster 0 –, those that are roughly in the center – cluster 2 – and finally those that on the west side – cluster 1.

Let us look at the specificities of each cluster by calculating the mean for each group.

**Visualization of the three clusters**



Let’s recap the information of each cluster by computing the mean of all variables within each group.



| cluster_k=3 | Nb_rooms | 1 star | 2 stars | 3 stars | 4 stars | 5 stars | Average_price_msq | growth 5 years | Pover_rate | Med_household_rev |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3300.875 | 1.750 | 14.750 | 28.375 | 8.50 | 0.25 | 9332.50 | 0.246625 | 18.625 | 23702.267292 |
| 1 | 4065.000 | 0.500 | 5.750 | 32.500 | 34.75 | 12.25 | 12332.50 | 0.201250 | 9.500 | 39617.423077 |
| 2 | 3865.125 | 1.375 | 10.125 | 37.750 | 20.75 | 3.25 | 11531.25 | 0.229625 | 12.500 | 31148.756944 |

The interpretation of each cluster is straightforward:

- ***Cluster 0: low income neighborhoods***
- ***Cluster 1: high income neighborhoods***
- ***Cluster 2: middle income neighborhoods***

**Cluster’s profile**

| Cluster | Dominant type of hotel (relatively to the other clusters) | Average housing price | Poverty rate | Household income |
|---|---|---|---|---|
| **Low income** | 1 and 2 stars | low | high | low |
| **Middle income** | 3 stars | average | average | average |
| **High income** | 4 and 5 stars | high | low | high |

From an investor point of view, the best neighborhoods for starting a business, building a hotel or acquiring a property, would be the cluster 1 and cluster 2, considered as “sure value”. Choosing between either neighborhood will depend on the risk profile and the financing capacity of the investor.

**Decision matrix for an investor willing to acquire a *property or a building***

| Investor profile | Risk aversion | Capacity of financing | Period of investing | Suitable cluster |
|---|---|---|---|---|
| **Investor 1** | High/average | High | Short/mid/long-term | 1 |
| **Investor 2** | Low | | Mid/long-term | 2 and 0 |
| **Investor 3** | High | Average | Mid/long-term | 2 |
| **Investor 4** | Average/low | | | 2 and 0 |
| **Investor 5** | Low | Low | | 0 |

## D. Discussion

For a city like Paris, getting the venues is not sufficient for segmenting between neighborhoods. There is no clear evidence between the type of venue and the level of social and economic indicators of a neighborhood (household revenue, housing price and poverty rate), this relation could be established by further investigating the type of venues (especially for café, bar and hotel) i.e. trying to classify them for example by the average price of the menu and average price for a hotel room.

The results obtained could be used by an investor as some preliminary insights, though, it is necessary to dig further about the real potential of a neighborhood. For example, the 16th arrondissement is classified as a high-income neighborhood, which is true, but it is also famous for its big and luxurious properties and hotels and much less for its venues, which are much lower relative to other neighborhoods. As a conclusion, the 16th arrondissement is maybe suitable for a wealthy investor in a luxurious property or hotel but not much for a restaurant or a café.

## E. Conclusion

Investing in a property or a building in one the neighborhoods of Paris is quite free of risk due to the growing demand and the touristic hub it constitutes. Thus, choosing between the clusters depend mainly on the profile investor (wealth, risk aversion…). In the other hand, we cannot apply partially the same reasoning for starting a business. Cluster 1 and 2 seems to have the best potential for business but those neighborhoods are heavily saturated by plenty of venues, unless the investment offers something different or innovative, the investor must investigate further.

In addition, some neighborhoods are becoming more attractive or has the potential to emerge, it could be interesting to identify those that are "undervalued" based on their indicators relatively to the cluster it belongs to. By doing a comparative analysis, we can identify some undervalued neighborhoods which could be interesting investing in.

**Identification of potential neighborhoods**

| | | Positioning relative to the average indicators of the cluster | | | |
|---|---|---|---|---|---|
| **Neighborhood** | **Cluster** | **Housing price** | **Poverty rate** | **Household income** | **Evolution of housing price over 5 years** |
| **Reuilly** | 0 | below | below | above | below |
| **Louvre** | 2 | Above | below | above | below |
| **Opéra** | 2 | below | below | above | above |

# Sources

*Atelier Parisien de l'urbanisme (APUR), www.apur.org/fr*

*Institut national de la statistique et des études économiques (INSEE), www.insee.fr*

*data.gouv.fr*

*Chambre des notaires de Paris, paris.notaires.fr*