

Exploring Paris to Make an Informed Choice about Relocating



IBM Coursera Data Science Capstone Project

Walter Belluco

May 24th 2020

Summary

This report is concerned with the analysis of real estate, demographic, and location data of the city of Paris to help the reader make an informed decision about where to relocate.

The report is based on publicly available data, from the opendata initiatives of the city of Paris and the French government. Since these data are sometimes expressed with a different geographical segmentation, a preliminary work was necessary to aggregate a single dataset enabling the analysis. Location data concerning the most popular venues was obtained from Foursquare API.

In order to facilitate interpretation, results are presented using a number of visual representations: choropleth maps, dot maps, scatter plots, line plots, heat maps. Content tables are available in the annex.

We looked at different metrics to describe the 80 neighborhoods and 20 districts of Paris:

- rent value of real estate
- selling value of real estate
- location data

The location data from Foursquare was used to group neighborhoods into 4 different clusters. Combining main results into the same map allows to identify at a glance the characteristics of a given neighborhood.

Further work can expand on this analysis to include other relevant criteria. For instance, the enriched dataset can be further developed into an online service providing added-value multi-criteria for real estate scoring.

Table of Contents

Summary	2
Table of Contents.....	3
Introduction and business objectives	4
Data.....	4
Methodology.....	5
Visualizing districts and neighborhoods	6
Grouping rent data by district and creating a reference data frame for real estate data.....	7
Visualizing rent and sales data.....	8
.....	8
Results.....	9
Is it better to buy or to rent?	9
Is population density correlated with real estate value?	11
What are the most common venues for each neighbourhood?	12
Where are similar neighbourhoods, and what do they have in common?	13
Discussion.....	14
Conclusions	16
ANNEX	17

Introduction and business objectives

Every year a number of people relocates to Paris, the French capital and global center of finance, fashion, arts, science, culture, and history.

However, making an optimal choice when relocating can be overwhelming. Which district should be chosen? Is it better to buy or to rent an apartment? A number of factors including renting vs. purchasing price, as well as demographics and the characteristics of the neighborhood must be taken into account.

Several questions may arise when considering where to relocate :

- Is it better to buy or to rent ?
- Does population density affect real estate value ?
- What are the most common venues for each neighbourhood?
- Where are similar neighbourhoods located, and what do they have in common ?

This project aims to create a tool to help make informed decisions based on a comparison of available data : rent prices, sales prices, population density and location data using Foursquare.

Different users will be drawn to different conclusions, according to their preferences and financial capabilities.

Data

A number of sources can be used to gather relevant information about Paris

- <https://parisdata.opendatasoft.com/> : rental data (80 neighbourhoods)
- <https://www.data.gouv.fr/> : administrative districts (20 arrondissements).
- <https://cadastre.data.gouv.fr/> : property sales data in 2019, per district.
- <https://developer.foursquare.com/> : location data.
- <https://fr.wikidia.org/> : population data.

The data was gathered using web-scraping, making calls to the Foursquare API, and downloading geojson files.

Minor data cleansing was done including adding renaming columns, adding district column names, getting rid of duplicates, deleting unnecessary columns.

Methodology

The majority of our exploratory data analysis focused on using longitude and latitude information to make maps and other visuals in order to better understand where neighborhood, districts, and other points of interests were located.

We utilized choropleth maps, dot maps, scatter plots, heat maps and content tables.

We also compared real estate rent and sale values, which are supposed to be correlated with each other, identifying districts having significant deviations.

Relevant data sets were available on a per neighborhood basis or on a per district basis. Reconciliation of the different dataset was necessary prior to analysis.

The geographical datasets consist of 80 neighborhoods arranged in 20 districts.

The rent dataset consisted of 2560 entries and the sales dataset consisted of 30805 relevant entries. In order to avoid outliers in the sales dataset, we restricted our analysis to the apartments having a sales value between 2000 and 40000 Euros per square meter (see also Figure 6 below). The following aggregated table is obtained. Table 1 and table 2 present an example of the rent and sales datasets, respectively.

Cluster analysis of foursquare location data was performed using K-means method. Finally, a summary representation of relevant results was proposed in a single image.

	neighborhood_number	neighborhood	number_of_rooms	furnished	rent_per_sqm	year	city	lat	lon
0	13	Saint-Merri	2	1	35.9	2017	PARIS	48.858521	2.351667
1	22	Odeon	4	0	26.8	2017	PARIS	48.847801	2.336339
2	22	Odeon	4	0	26.4	2017	PARIS	48.847801	2.336339
3	24	Saint-Germain-des-Prés	1	0	32.9	2017	PARIS	48.855289	2.333657
4	24	Saint-Germain-des-Prés	4	0	26.8	2017	PARIS	48.855289	2.333657

Table 1: Real Estate rent (per neighborhood)

	code_postal	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	longitude	latitude	sale_per_sqm
0	1.0	1196000.0	Appartement	3	112.0	2.325288	48.868416	10678.571429
1	1.0	1935600.0	Appartement	2	66.0	2.337119	48.865114	29327.272727
2	1.0	209000.0	Appartement	1	19.0	2.337476	48.866524	11000.000000
3	1.0	752400.0	Appartement	3	75.0	2.345875	48.864218	10032.000000
4	1.0	530000.0	Appartement	2	46.0	2.343732	48.858427	11521.739130

Table 2: Real Estate sales (per district, i.e. per postal code)

Visualizing districts and neighborhoods

First of all, we need to understand where districts and neighborhoods are located.

We used choropleth maps for rent data / sales data and dot maps to identify the district, with a specific color code per district.

We can see from Figure 1 that the Paris district numbers are arranged in a clockwise spiral from the center outwards.

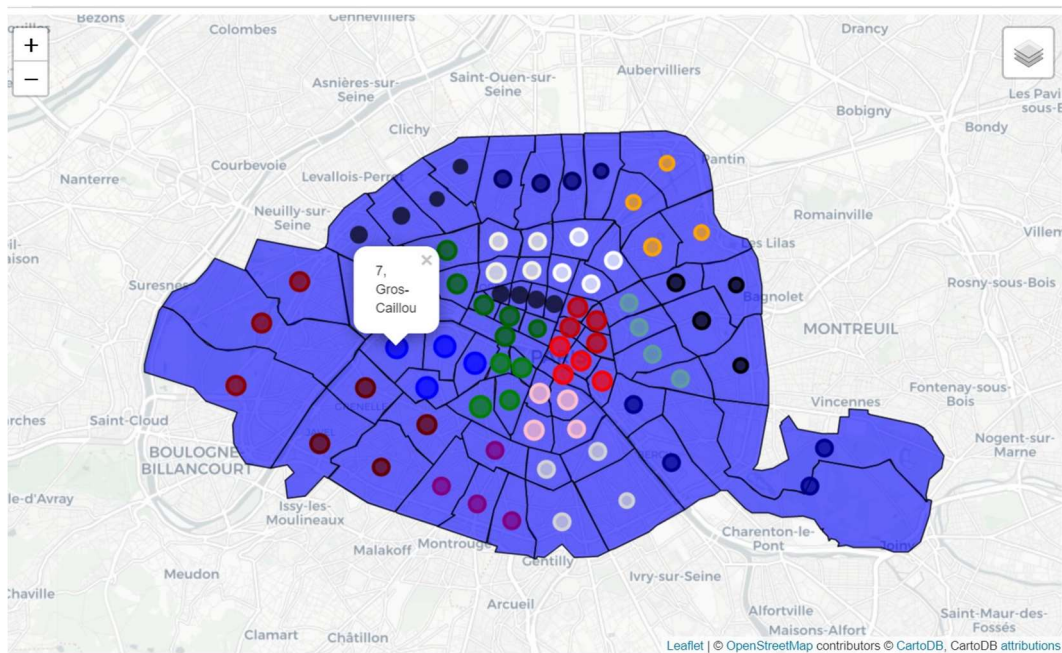


Figure 1: Map of districts and neighborhoods. Color dots represent districts.

Sales data and Rent data were available on a different geographical basis: while rent data were available in a ‘per neighborhood’ basis, sales data were available in a ‘per district’ basis. Paris has 80 neighborhoods in 20 districts, so the “per district” basis is less granular.



	district	rent_per_sqm	lat	lon	district_n	district_name	Surface (hectars)	Population (habitants)	Density (hab/hectar)	dist	sale_per_sqm
0	1	32.586719	48.863654	2.336175	1	Paris 1er arrondissement	183	17 100	93.0	1	13394.495015
1	2	32.372656	48.868391	2.341918	2	Paris 2e arrondissement	99	22 390	227.0	2	11552.363970
2	3	31.996875	48.863027	2.359566	3	Paris 3e arrondissement	117	35 991	307.0	3	12647.411686
3	4	33.124219	48.854680	2.356843	4	Paris 4e arrondissement	160	27 769	173.0	4	13992.603528
4	5	31.432813	48.845083	2.350149	5	Paris 5e arrondissement	254	60 179	236.0	5	13037.819409
5	6	34.000000	48.850975	2.334347	6	Paris 6e arrondissement	215	43 224	201.0	6	15425.011630
6	7	35.500000	48.855609	2.313651	7	Paris 7e arrondissement	409	57 092	139.0	7	15024.348240
7	8	33.293750	48.872656	2.312500	8	Paris 8e arrondissement	388	38 749	99.0	8	12918.311243
8	9	30.760156	48.876807	2.338308	9	Paris 9e arrondissement	218	59 474	272.0	9	11662.342213
9	10	29.134375	48.875401	2.359945	10	Paris 10e arrondissement	289	94 474	326.0	10	10818.260408
10	11	29.081250	48.859727	2.379553	11	Paris 11e arrondissement	367	155 006	422.0	11	10341.288879
11	12	27.430469	48.837620	2.405654	12	Paris 12e arrondissement	637	144 925	227.0	12	10013.739743
12	13	28.360937	48.830449	2.358956	13	Paris 13e arrondissement	715	182 386	255.0	13	9872.322437
13	14	29.286719	48.829511	2.327649	14	Paris 14e arrondissement	564	141 102	250.0	14	11127.463471
14	15	30.209375	48.841559	2.294407	15	Paris 15e arrondissement	848	238 190	280.0	15	10534.075051
15	16	32.045312	48.863524	2.268952	16	Paris 16e arrondissement	791	167 613	211.0	16	11603.811020
16	17	30.257812	48.887411	2.306962	17	Paris 17e arrondissement	567	170 156	300.0	17	11013.844987
17	18	27.013281	48.892599	2.350066	18	Paris 18e arrondissement	601	201 374	335.0	18	10066.068677
18	19	25.269531	48.885874	2.383703	19	Paris 19e arrondissement	679	186 116	274.0	19	8837.313002
19	20	26.195312	48.865261	2.399106	20	Paris 20e arrondissement	598	197 311	329.0	20	9046.189635

Table 3: Real Estate value table.

Visualizing rent and sales data

A consistent trend is observed in rent and sales data, with central and western districts being generally more expensive than eastern districts.

A more detailed analysis is necessary to make an informed decision of renting vs. buying.

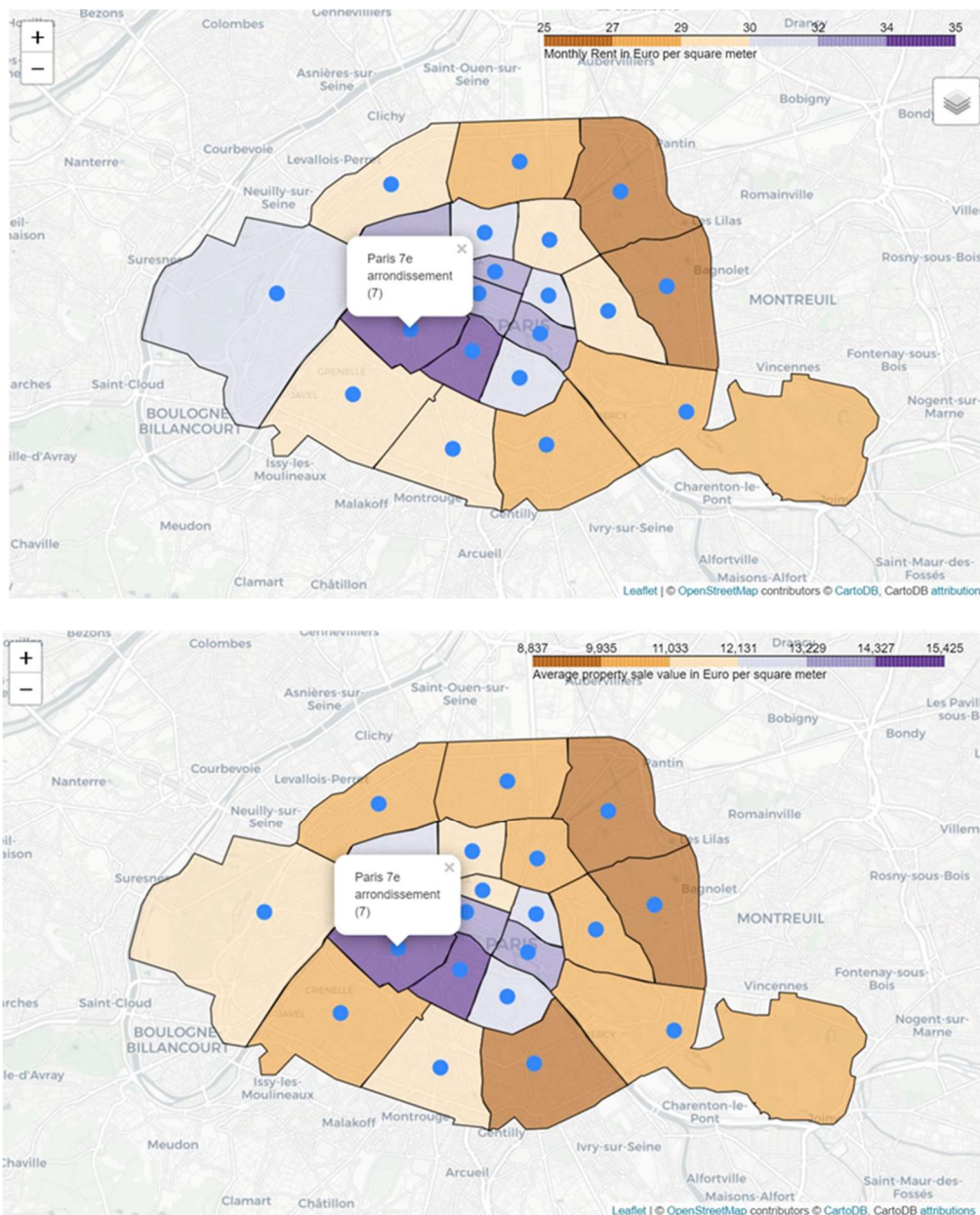


Figure 3: Real Estate Rent data per district (top), sales data per district (bottom).

Results

Is it better to buy or to rent?

Figure 4 represents the correlation between real estate rent and sales value. Rent value per unit surface and Sales value per unit surface are linearly correlated, as one might expect.

Above the trendline, it seems more reasonable to rent than to buy an apartment. For instance, such is the case for District n. 6, which has the highest sales value per square meter.

Below the trendline, it seems more reasonable to buy than to rent. For instance, such is the case for district n. 15, 16 and 2, which have a relatively low sales value compared to the rent value.

This analysis is only preliminary. Before making any buying decision, it should be completed with a more exhaustive investigation concerning financial aspects such as one's capacity to obtain a mortgage, interest rates, trend analysis on sales data in the past year and the forecasted overall economic outlook in Europe, France and Paris.

Also, the peculiar characteristics of the neighborhood should be taken into account.

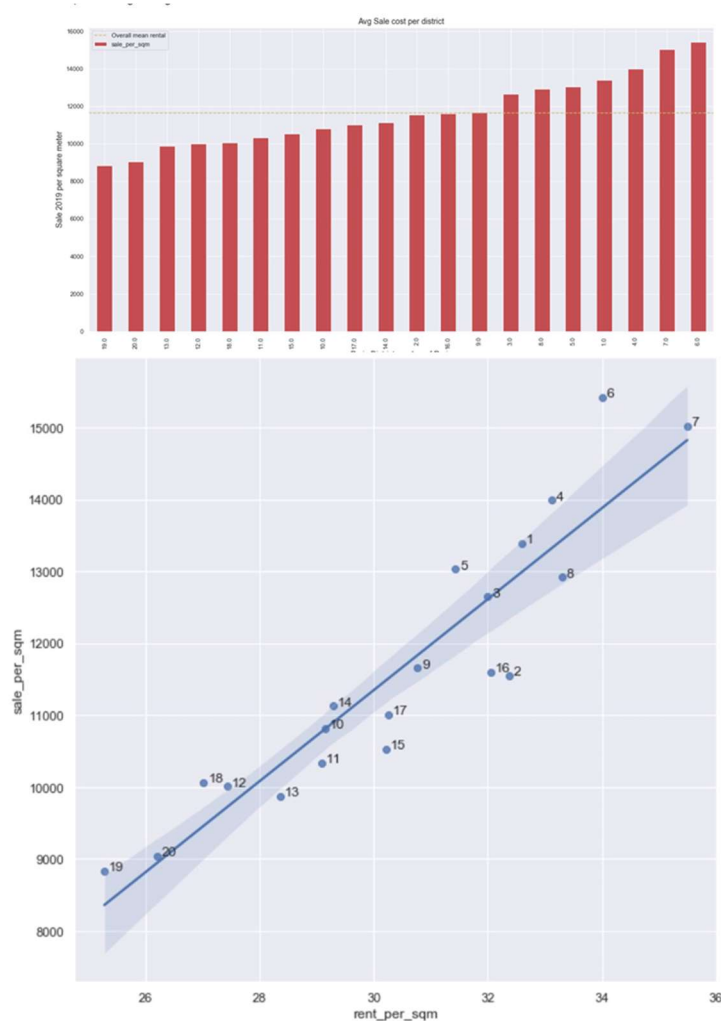


Figure 4: Real Estate Sales value per district (top), Correlation between sales and rent (bottom).

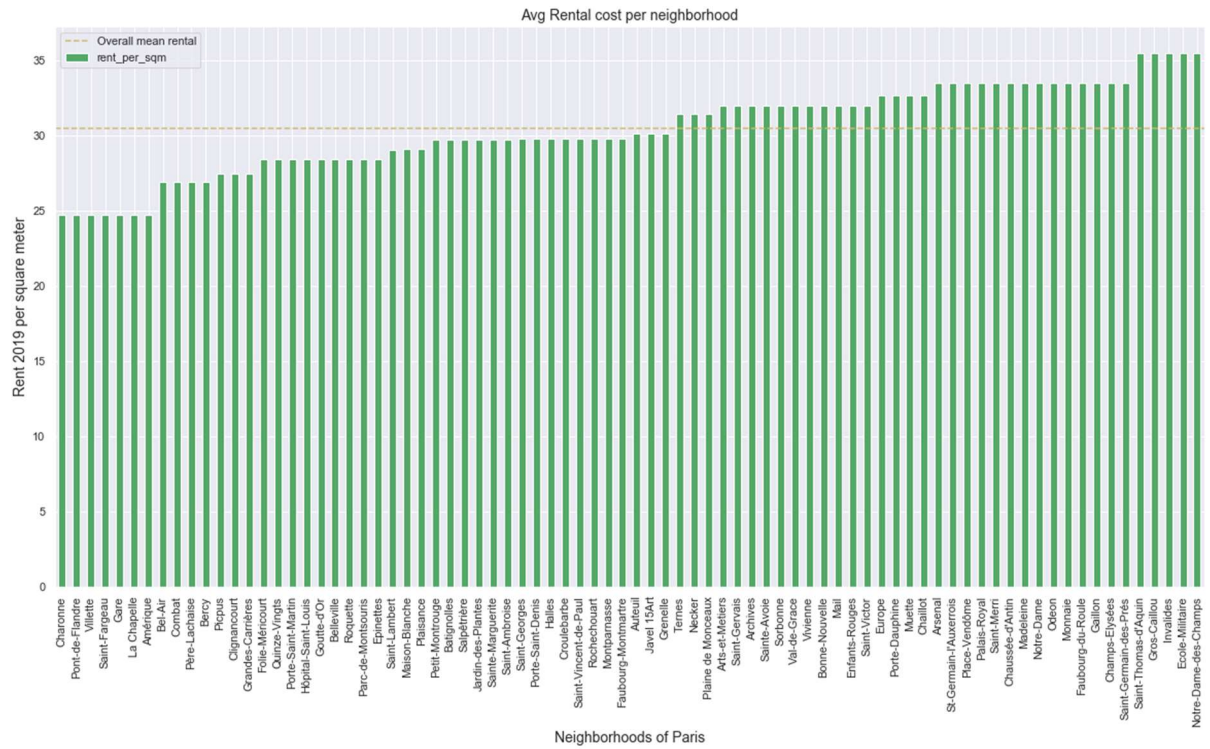


Figure 5: Real Estate Rent value per neighbourhood.

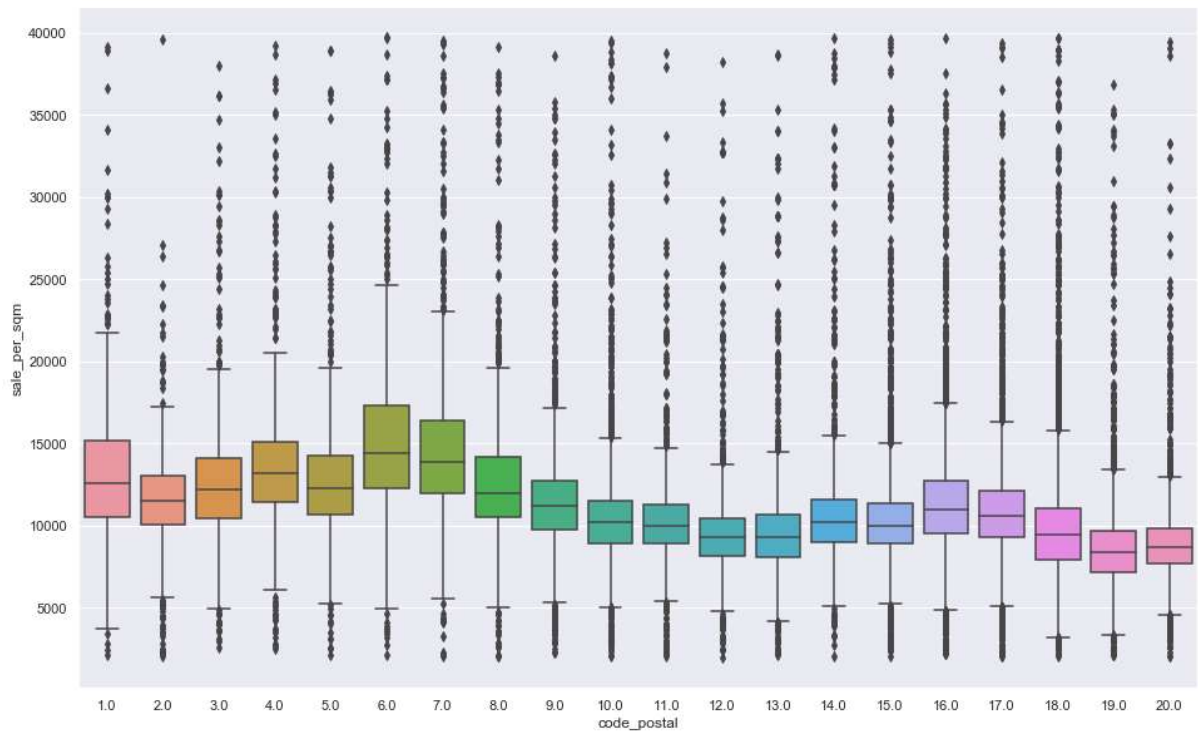


Figure 6: Real Estate Sale value per district.

Is population density correlated with real estate value?

These choropleth maps in Figure 7 represent property value, where the blue dot surface is proportional to population density.

The most expensive districts in terms of property value are also the least populated.

A number of reasons including socio-economic segmentation and accessibility to services and proximity to tourist and business venues could explain this trend.

In order to gain more insight, location analysis is necessary.

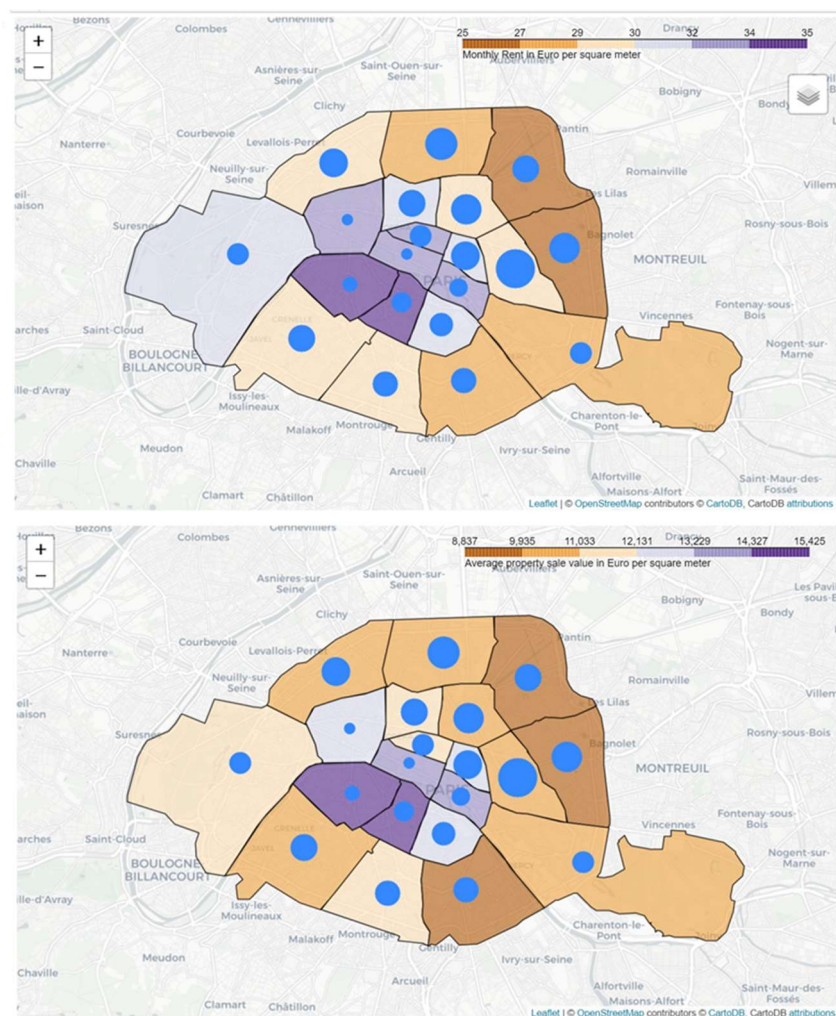


Figure 7 : Top: Monthly rent (colors) vs population density (circles). Bottom : Property value (colors) vs. population density.

What are the most common venues for each neighbourhood?

The 80 neighborhoods in Paris were investigated using Foursquare API. A total of 296 different venue categories were found. For the quantitative analysis we used the first 20 most popular venue categories for each neighborhood.

Neighborhoods and venues were ordered by total number of venues.

Overall, the most common location in Paris is French Restaurant, followed by Hotels and Italian Restaurants.

A heatmap is a compact and efficient way to identify relevant venues. For instance, if you are looking for a Japanese restaurant, the two neighborhoods to go to are Palais Royal and Gaillon.

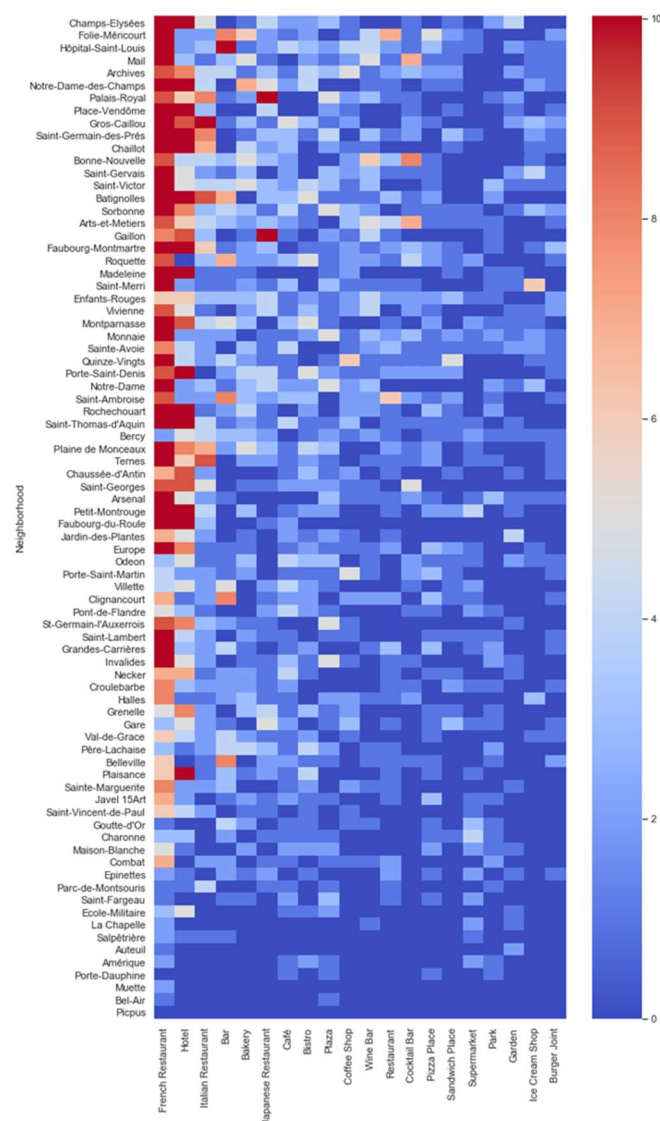


Figure 8 : Number of popular locations per neighborhood.

Where are similar neighbourhoods, and what do they have in common?

The 80 neighborhoods in Paris were clustered into 4 groups based on the most popular venues, using K-means.

As an example, consider Figure 9. the first two most significant venues, French Restaurants and Hotel, are considered here. The scatter plot represents each neighborhood with a circle having the color of the cluster it belongs to.

Based on centroid analysis we could define a name cluster :

- Cluster 0: Exclusive : Glamour and Tourism
- Cluster 1: Going-out
- Cluster 2: Residential
- Cluster 3: Business

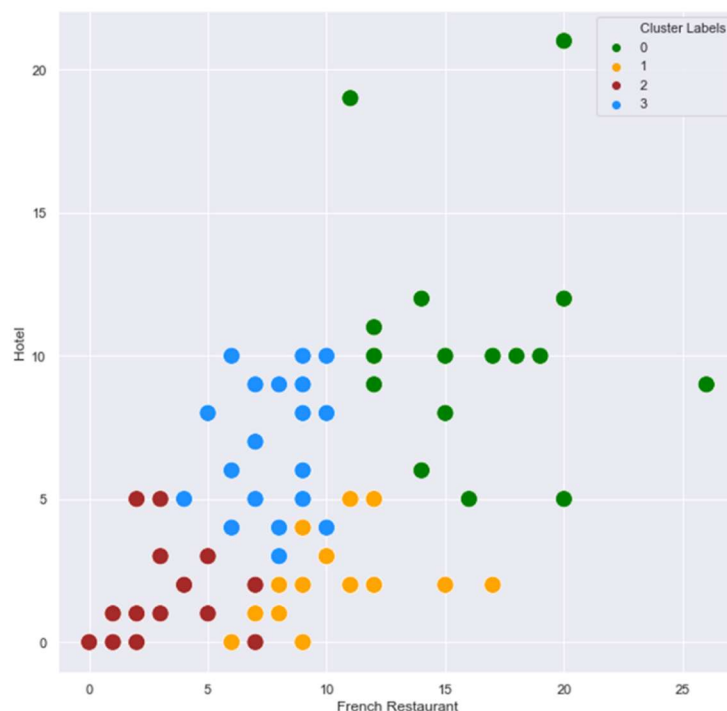


Figure 9 : Segmentation of neighborhood according to the first 2 most common venues (French Restaurant and Hotels).

The number of the first 2 venue categories can be very different in different neighbourhood. We looked at the total number of popular venues for each neighbourhood, and we plotted their relationship with property value (Figure 10). Interestingly, there is no correlation between real estate value and the total number of popular venues in a given neighbourhood.

Linear regression analysis could be used to model rent value as a function of each venue category as it is to be expected that the category of the venues is more important than near overall quantity. This analysis goes beyond the scope of the present investigation.

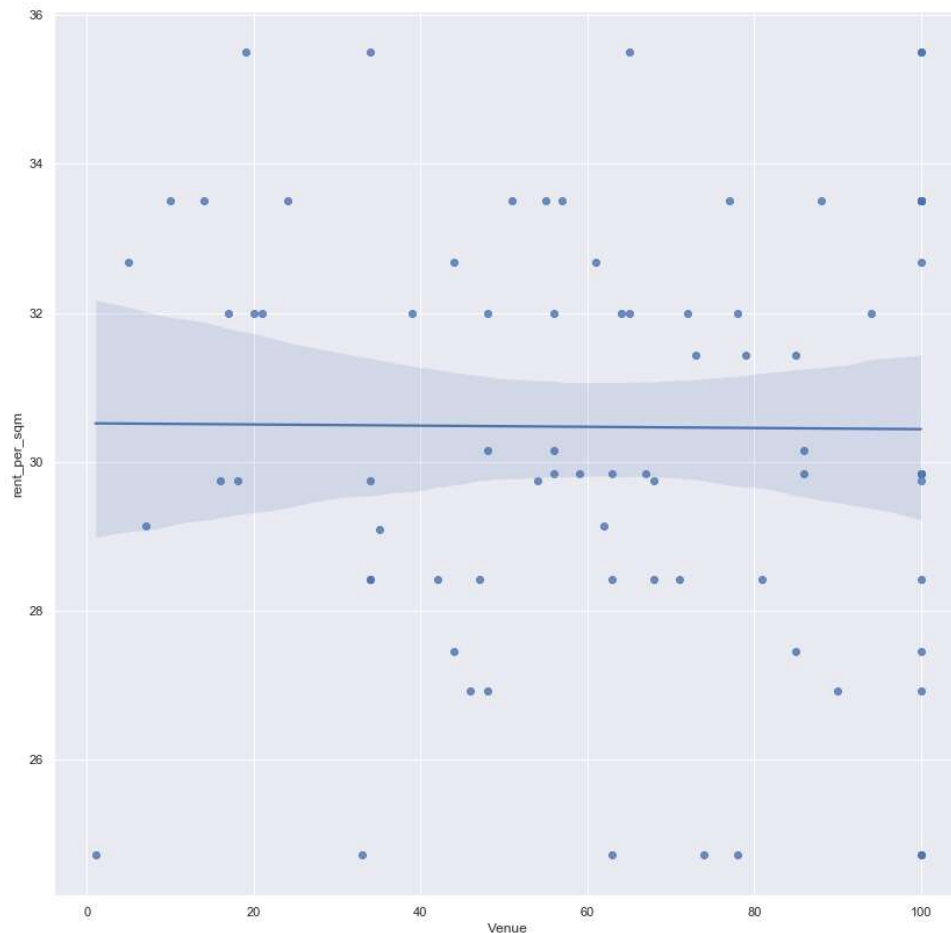


Figure 10: correlation between number of venues and rent

Discussion

- A map with location and characteristics of each cluster, superposed to the real estate value choropleth, is proposed. Based on this representation, the reader can immediately recognize its target neighborhood based on his/her specific criteria. Also, this representation enable further segmentation into subgroups. For instance, cluster 2 “Residential” can be further divided into East (less wealthy) and West (wealthier).
- All data were aggregated in a single dataset, that can be further used to make optimal choices, given a decision vector which captures the weight of each criterion.
- More data could be aggregated to this dataset in order to further refine the choice: crime rate, distance from work, characteristics of the family to be relocated (e.g. availability of special schools for children), ethnic and sociological characteristics of the neighborhood, distance from significant cult-related venues, financial capability, etc. This enriched dataset could pave the way for an online service providing an overall real estate scoring based on its location.
- While the incorporation of more data in our model could provide additional insight, this goes beyond the scope of the present project.

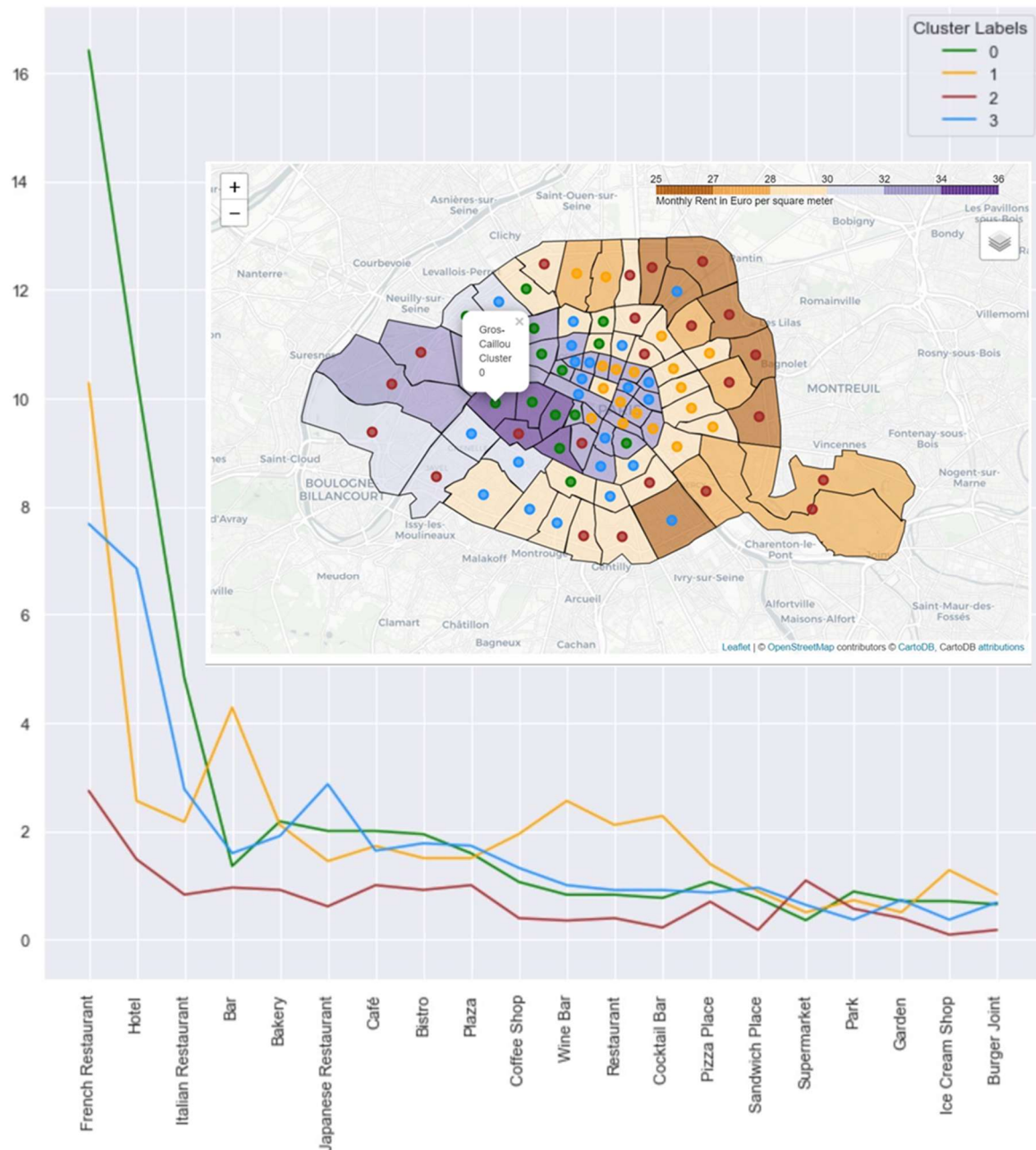


Figure 11: Top 20 Cluster characteristics, cluster location and related monthly rent.

Conclusions

- Several criteria come into play when relocating. This analysis took into account property value, population density and location analysis to provide relevant criteria to make an informed relocation choice.
- All data were aggregated in a single dataframe, that can be used to make optimal choices, for a given a decision vector. This dataset was used to create a visual representation of the location and characteristics of each cluster, superposed to the related real estate value.
- Based on this representation, when relocating to Paris, an informed choice can be made very easily.
- Further work can expand on this analysis to include other relevant criteria. The enriched dataset could be further developed into an online service providing multi-criteria real estate scoring based on location.

ANNEX

[illegible]