BATTLE OF FRENCH NEGHBORHOODS

Final Data Science capstone for IBM data science course on Coursera

1. INTRODUCTION

a. Business Problem

The service we want to provide is simple and can be a real time saving. Assume your customer is about to move out, for any reason you can imagine: looking for a bigger flat or house, looking for a cheaper property, looking for somewhere closer to his new job…

When looking for a new place to live in, you'll be interested in 3 main criterias:

- property description
- property price
- neighbourhood description

Before you start actually visiting some flats or houses, you need to narrow down your search to some specific areas based on the 2 latest criterias.

The idea here is to help doing this first screening by providing a list of neighbourhoods which can fit with what you are looking for, and to provide the corresponding real estate prices. We will characterise neighbourhoods on the basis of the different venues you can find there, as it is a strong marker of the way-of-life you can have in a specific area.

As a practical case we will assume someone is looking for a place to live similar to the Epinettes neighbourhood in Paris, and we will also assume that he is ready to move out to Paris, Toulouse or Bordeaux.

We will provide a list of neighborhoods from those 3 cities with similar venues and their associated real estate prices so he can make his choice, based on his budget.

b. Available data

To solve this problem we will use different datasets, all this data is open-source and it is available from my Github repository.

First I need the list of neighbourhoods for the 3 cities with their geo-coordinates. After some data mining I have been able to find open source data for Paris & Toulouse neighbourhoods. As for Bordeaux I had to go on google maps to get the coordinates & create an excel database:

- *Source « quartier_paris.csv »: https://www.data.gouv.fr/fr/datasets/recensement-population-2013-grands-quartiers-population/*
- *Source « quartier_toulouse.csv »: https://data.toulouse-metropole.fr/explore/dataset/recensement-population-2016-grands-quartiers-diplomes/information/*
- *Source: « quartier_bordeaux.xslx »: from google maps*

Then I need the real estate prices for each of those neighbourhoods. Open source data is available including all the real estates transactions that have been realised in 2019 (last year). Using those and doing some data wrangling I can obtain prices in terms of €/m2, which is the information that I will provide. One can argue that prices have slightly increased since 2019, but this is only marginal and differences from one neighbourhood to the other remain similar.

- *Source « 31.csv », « 75.csv », « 33.csv »: https://cadastre.data.gouv.fr/data/etalab-dvf/latest/csv/2019/*

Eventually I need to obtain the list of venues for each neighbourhood, which is provided by the Foursquare API.

2. METHODOLOGY

a. Database of neighbourhoods with geolocalisation

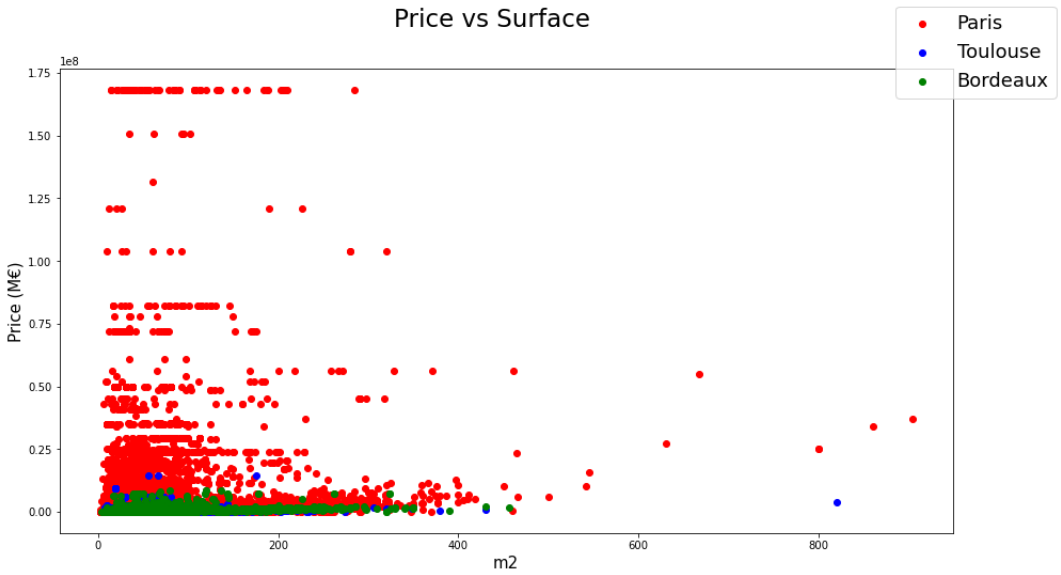
I have extracted neighbourhood name, latitude, longitude and city from the 3 different data sources corresponding to the 3 different cities. After dropping from those files features which are not useful in our analysis, I merged the data to get one single dataframe. To be noticed, I kept the “arrondissement feature”, which is only valid for Paris neighbourhoods, but which can proved useful later on.

	Quartier	Latitude	Longitude	City	arrondissement
0	CARMES	43.596087197	1.44140164967	Toulouse	Toulouse
1	AMIDONNIERS	43.6062858201	1.42339702287	Toulouse	Toulouse
2	MINIMES	43.6191621779	1.43195460628	Toulouse	Toulouse
3	BONNEFOY	43.6195659969	1.45452253663	Toulouse	Toulouse
4	MARENGO - JOLIMONT	43.6117583769	1.46190401897	Toulouse	Toulouse

The dataframe contains 60 neighbourhoods from Toulouse, 80 from Paris and 14 from Bordeaux, which makes a total of 154 different neighbourhoods to be analysed.

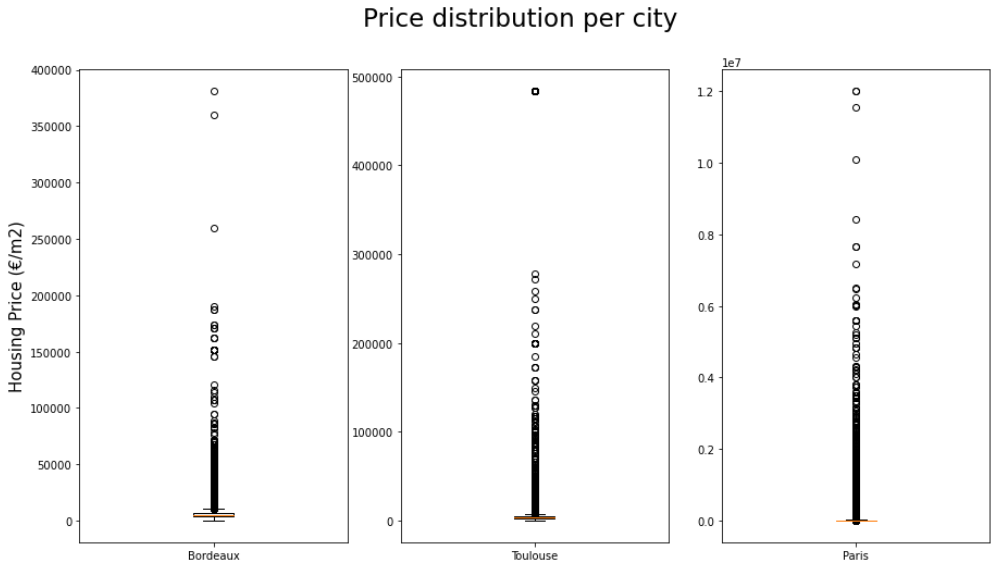
b. Average real estate prices per neighbourhood

After merging our data sources, I get a database of 50081 transactions done in 2019 for the 3 cities, which reduces to 50012 after dropping transaction where the surface of the property or the price isn't provided.



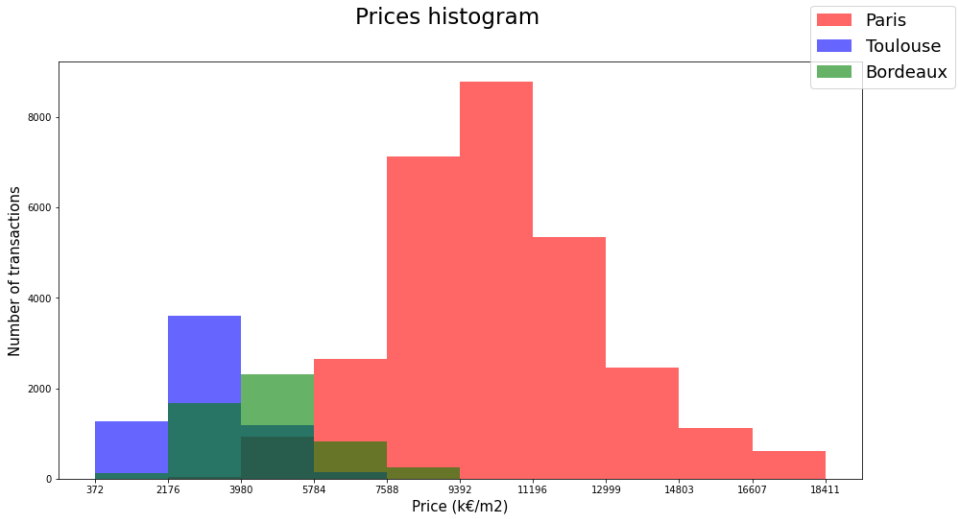
When cross-plotting price (million euro) vs surface (m2), we can't observe any trend, which isn't what was expected, are real estate prices are meant to have some kind of linear relationship with surface, apart from some "exotic" properties.

To better understand the data distribution, we add a feature which we call "prix_m2" which is the price divided by the surface. Then looking at boxplots for the 3 cities, we can observe there are a lot of outliers, especially in Paris (which could already be observed in our previous cross-plot).



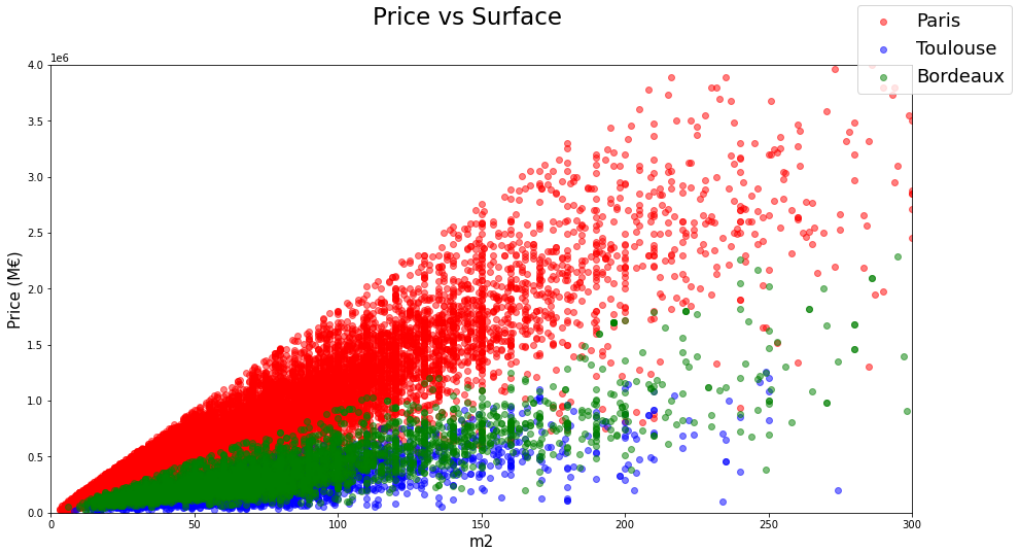
We then perform a series of filtering to remove those properties that are not representative of their neighbourhood real estate prices.

First we remove properties that are above the upper quantile plus one IQR (Inter Quantile Range) and properties that are below the lower quantile less one IQR. Looking at the new boxplot, price distributions per city are then more homogeneous and our statistics are expected to be more representative of the current market average prices. This is also illustrated by the following histogram. To be noticed, there are still 40435 transactions in our dataset.



Now we have cleaned our dataset, we can have a new look at the cross plot price vs surface. The expected trends can be observed:

- Price increase linearly with surface once « exotic » properties have been removed
- Paris is more expensive than Bordeaux & Toulouse
- For a given surface there is still a price spread for each city depending on the neighbourhood



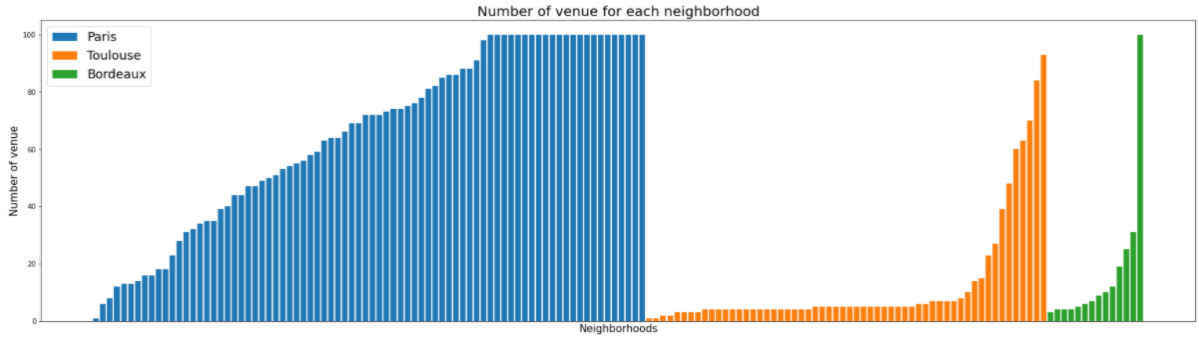
Now the real estate prices dataset is clean and ready for use, we calculate the average housing price per meter square for each neighbourhood from the df_quartier dataframe. To do so we calculate the average over the transactions that have been performed in a 500 m radius around the neighbourhood coordinates. This operation takes several minutes to be performed, we reduce this length by using only transactions from the same "arrondissement" than the neighbourhood.

Eventually this is how our "df_quartier" dataframe looks like, with the neighbourhood name, coordinates, housing price per meter square, and arrondissement (this feature is only meaningful for Paris):

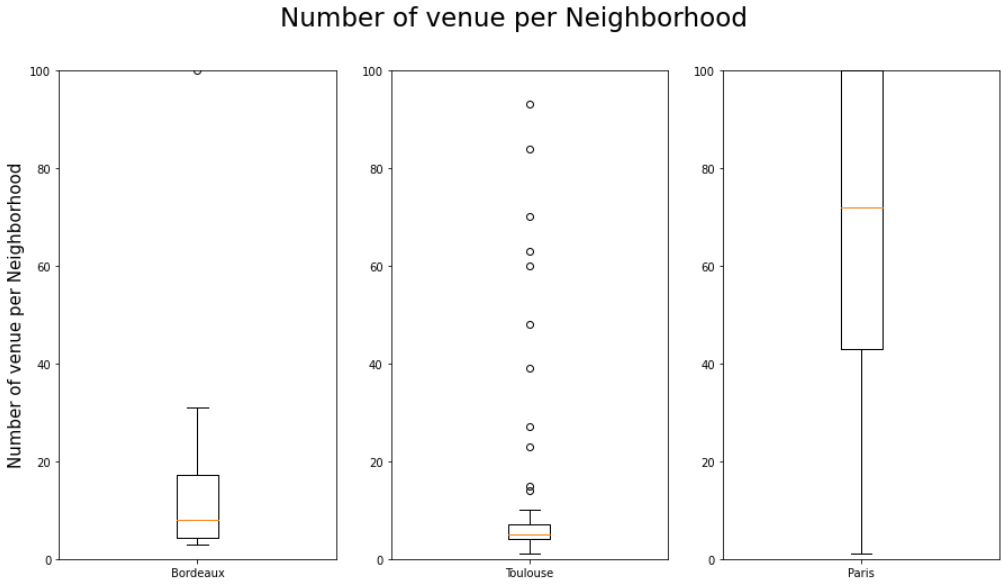
	Quartier	Latitude	Longitude	City	arrondissement	prix_m2
0	CARMES	43.596087197	1.44140164967	Toulouse	nan	4624.42
1	AMIDONNIERS	43.6062858201	1.42339702287	Toulouse	nan	3638.38
2	MINIMES	43.6191621779	1.43195460628	Toulouse	nan	3089.11
3	BONNEFOY	43.6195659969	1.45452253663	Toulouse	nan	3078.21
4	MARENGO - JOLIMONT	43.6117583769	1.46190401897	Toulouse	nan	3003.14

c. Neighbourhood characterisation through venues listing

In addition to the real estate prices, neighbourhoods will be characterised by the venues that can be found there. Venues are from a call to Foursquare API. Requests are limited to a radius of 500 m around the center point of the neighbourhood and to 100 venues per neighbourhood.

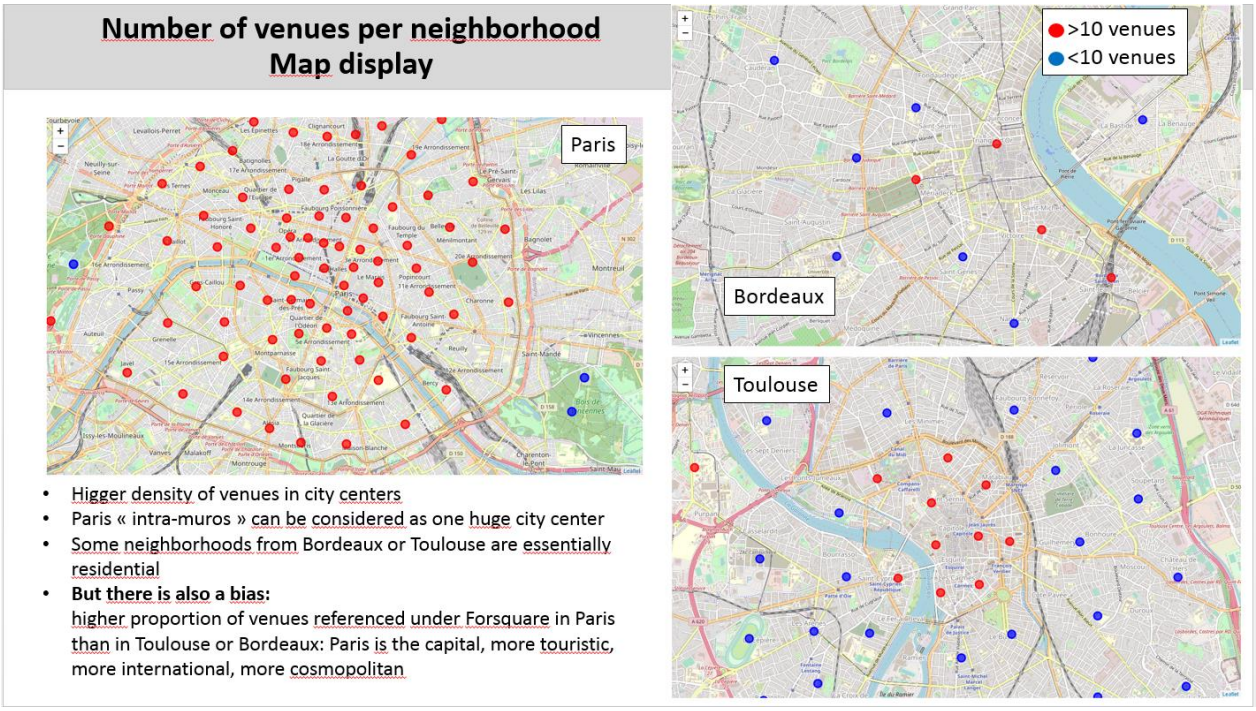


Looking at the distribution of venues per neighbourhood, it appears that while in Paris the distribution is more progressive with several places above the 100 venues cut-off, there are some places in Bordeaux and even more in Toulouse with very few venues registered on Foursquare API. It can be due to 2 reasons: either there are actually very few venues there, or those neighbourhoods are not popular enough or they are not visited by people that contribute to referencing venues on Foursquare and even though there would be several venues in there, they wouldn't appear on Foursquare.



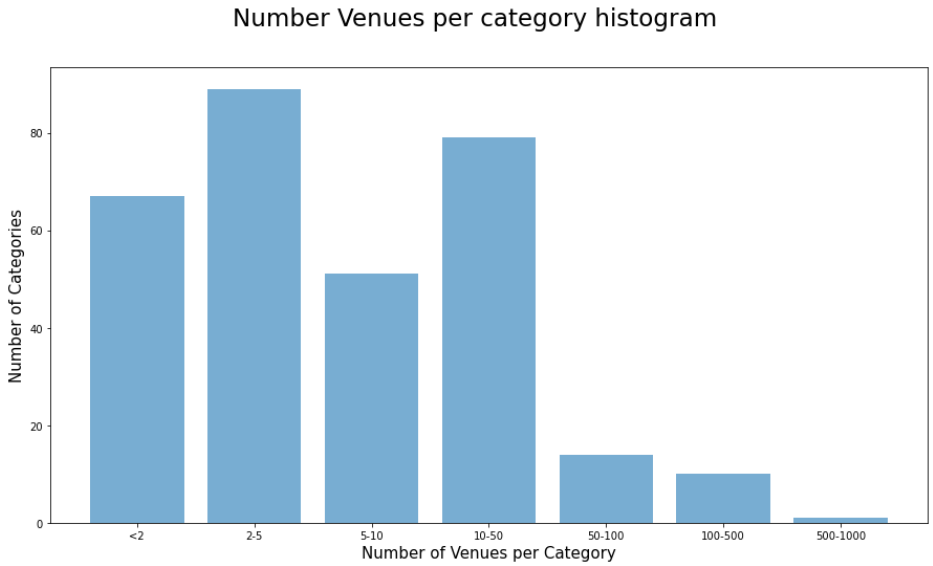
We set a cut-off at 10 venues per neighbourhood, considering that below this cut-off either there are not enough venues for the place to be attractive enough for us, or there are not enough venues on Foursquare to characterise the area and the description wouldn't be relevant enough.

Looking at the spatial distribution, only areas from Toulouse and Bordeaux cities centers are above this cut-off; areas below this cut-off are essentially residential. Paris can be considered as one huge city center, crowded and dynamic, and only areas in the Bois de Boulogne or Bois de Vincennes have less than 10 venues, which makes sense as those areas are woods.



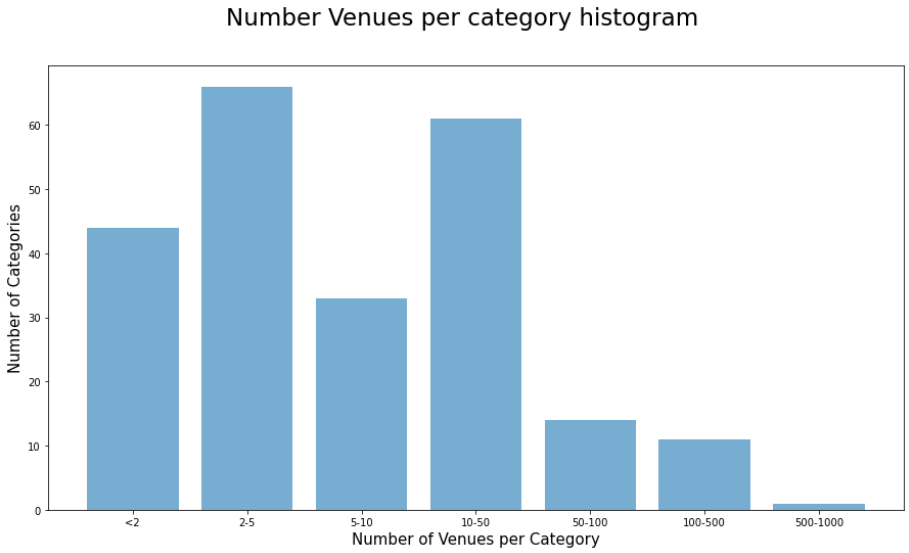
Those neighbourhoods below cut-off are removed from our dataset, which reduced their number from 154 to 93, and the number of venues from 6236 to 5954, meaning that 60% of the neighbourhoods concentrate 95% of the venues from Foursquare.

After those considerations on spatial distribution of venues, we focus on types of venues: there are 319 different categories of venues. Looking at the histogram below, we can see that more than 60 categories gather only one venue each, and more than 80 gather between 2 and 5 venues. Neighbourhoods will be compared on their number of venues belonging to similar venue categories, so categories with only 1 venue won't help clustering the neighbourhoods.



Sometimes this limited number of venue per type is justified, but in many cases, it is only due to the category description being too specific, and two similar venues end-up being classified

under two different names. For example most of the 87 different types of restaurants belong to this group of “too specific” descriptions, by reducing those to 9 we reduce the overall number of categories to 239; and more importantly the number of categories with only 1 venue is reduced from more than 60 to less than 45, and the number of categories with 2 to 5 categories is reduced from 90 to 65.

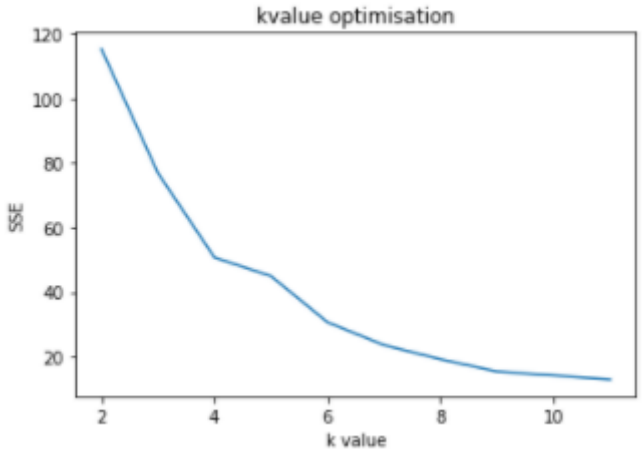


Now we can describe each neighbourhood with its 5 most common type of venues.

	Neighborhood	1st Most common venue	2nd Most common venue	3rd Most common venue	4th Most common venue	5th Most common venue
0	ARNAUD BERNARD	Bar	French Restaurant	Plaza	Sandwich Place	Hotel
1	Amérique	French Restaurant	Supermarket	Plaza	Bed & Breakfast	Pool
2	Archives	French Restaurant	Asian_Restaurant	European_Restaurant	Clothing Store	Art Gallery
3	Arsenal	French Restaurant	Hotel	European_Restaurant	Modern_Food_Restaurant	Asian_Restaurant
4	Arts-et-Métiers	French Restaurant	Asian_Restaurant	Hotel	Wine Bar	European_Restaurant

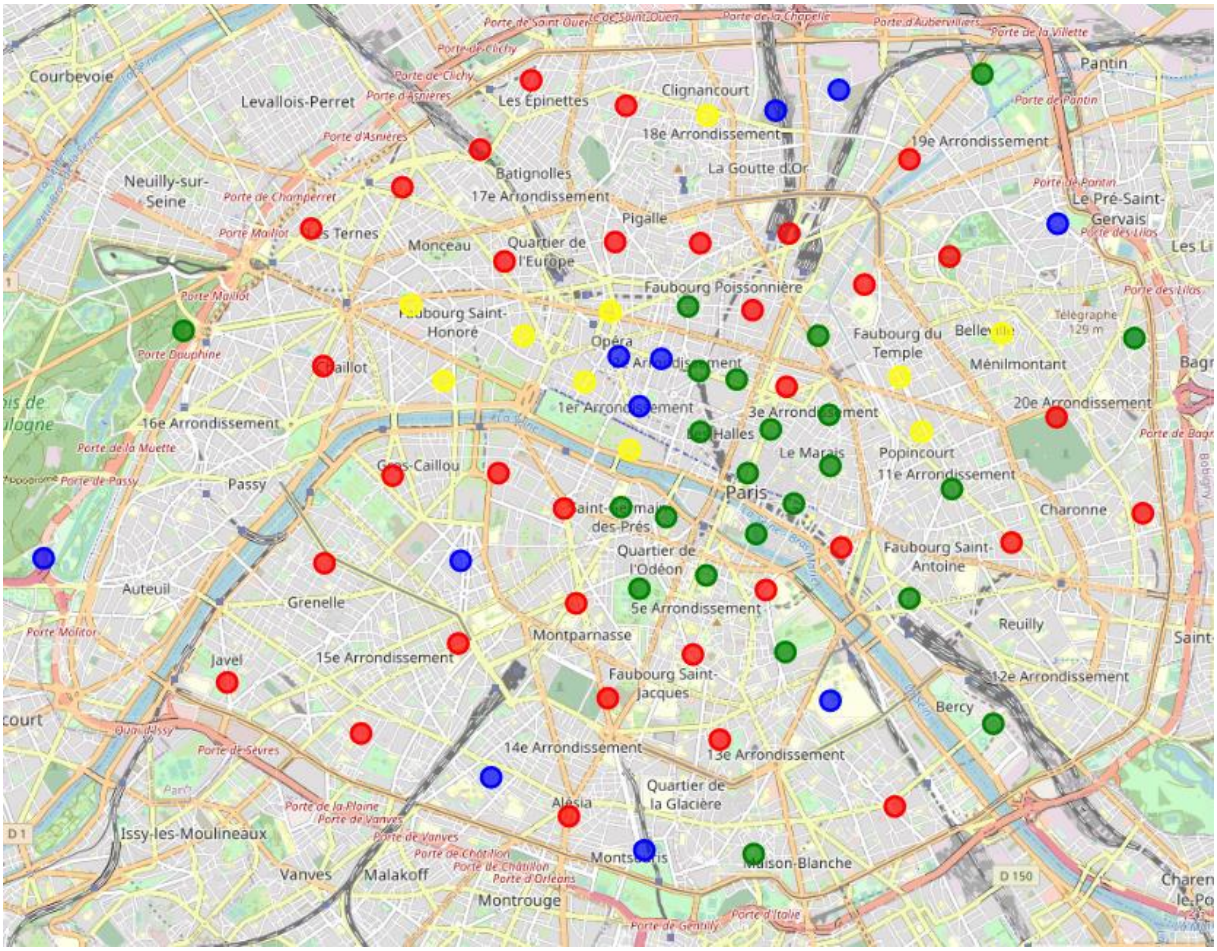
d. Neighbourhood clustering based on venues

Now each neighbourhood from the 3 cities can be described with its average housing price per meter square, and its 5 most common types of venues. The final step is too cluster those neighbourhoods based on their venues, the k-means clustering algorithm will be used here.



First we estimate the optimum k value by calculating the sum of the distance of the neighbourhoods to their cluster centers with scipy. The plot of the results for k varying from 2 to 11 shows that the optimum value is 4.

The algorithm is then run for this k value of 4. The map of Paris neighbourhoods colored per cluster is actually difficult to interpret, although it seems there is some kind of spatial consistency.



3. RESULTS

a. Result

The result is finally there: we can provide the list of neighbourhoods from the same cluster as the Epinettes in Paris, with their 5 most common type of venues and their housing price per square meter. Our customer can now look for more information about places in this reduced list and eventually he can make his choice.

Neighborhood	Latitude	Longitude	City	arrondissement	prix_m2	1st Most common venue	2nd Most common venue	3rd Most common venue	4th Most common venue	5th Most common venue	Cluster
Saint Jean - Belcier	44.83	-0.56	Bordeaux	Bordeaux	4577.33	Cluster	Hotel	Nightclub	Supermarket	Asian_Restaurant	1
SAINT-GEORGES	43.60	1.45	Toulouse	Toulouse	4515.94	French Restaurant	Plaza	Hotel	Bar	Asian_Restaurant	1
MATABIAU	43.61	1.45	Toulouse	Toulouse	4380.18	Cluster	Hotel	French Restaurant	Bar	Asian_Restaurant	1
SAINT-AUBIN - DUPUY	43.60	1.45	Toulouse	Toulouse	4173.31	French Restaurant	Hotel	Bar	European_Restaurant	North_American_Restaurant	1
COMPANS	43.61	1.43	Toulouse	Toulouse	3741.14	French Restaurant	Hotel	Sandwich Place	Bar	Pizza Place	1
Arsenal	48.85	2.36	Paris	4	0.00	French Restaurant	Hotel	European_Restaurant	Modern_Food_Restaurant	Asian_Restaurant	1
Plaine de Monceaux	48.89	2.30	Paris	17	0.00	Cluster	French Restaurant	European_Restaurant	Hotel	Bakery	1
Porte-Saint-Denis	48.87	2.35	Paris	10	0.00	Cluster	Asian_Restaurant	Hotel	French Restaurant	Bakery	1
Père-Lachaise	48.86	2.40	Paris	20	0.00	Cluster	Asian_Restaurant	Bar	Bakery	Bistro	1
Rochechouart	48.88	2.34	Paris	9	0.00	French Restaurant	Hotel	Bakery	Pizza Place	Asian_Restaurant	1
Saint-Georges	48.88	2.33	Paris	9	0.00	Cluster	Hotel	French Restaurant	European_Restaurant	Asian_Restaurant	1
Notre-Dame-des-Champs	48.85	2.33	Paris	6	0.00	French Restaurant	Hotel	North_American_Restaurant	Bakery	Asian_Restaurant	1
Saint-Lambert	48.83	2.30	Paris	15	0.00	French Restaurant	Asian_Restaurant	Hotel	French_Regions_Restaurant	Café	1
Saint-Thomas-d'Aquin	48.86	2.33	Paris	7	0.00	French Restaurant	Hotel	Café	Asian_Restaurant	Coffee Shop	1
Saint-Victor	48.85	2.35	Paris	5	0.00	French Restaurant	Asian_Restaurant	European_Restaurant	Bakery	Hotel	1
Saint-Vincent-de-Paul	48.88	2.36	Paris	10	0.00	Cluster	Asian_Restaurant	French Restaurant	Hotel	African_Restaurant	1
Sainte-Marguerite	48.85	2.39	Paris	11	0.00	French Restaurant	Asian_Restaurant	Bar	European_Restaurant	Hotel	1
Ternes	48.88	2.29	Paris	17	0.00	French Restaurant	European_Restaurant	Asian_Restaurant	Hotel	Modern_Food_Restaurant	1
Val-de-Grâce	48.84	2.34	Paris	5	0.00	Cluster	Asian_Restaurant	French Restaurant	Bar	Hotel	1
Petit-Montrouge	48.83	2.33	Paris	14	0.00	Cluster	Hotel	Asian_Restaurant	French Restaurant	European_Restaurant	1
Necker	48.84	2.31	Paris	15	0.00	Cluster	French Restaurant	Asian_Restaurant	Hotel	Café	1
Arts-et-Métiers	48.87	2.36	Paris	3	0.00	Cluster	French Restaurant	Asian_Restaurant	Hotel	Wine Bar	1
Europe	48.88	2.32	Paris	8	0.00	French Restaurant	Hotel	Asian_Restaurant	Pizza Place	European_Restaurant	1
Batignolles	48.89	2.31	Paris	17	0.00	French Restaurant	Hotel	European_Restaurant	Asian_Restaurant	Bar	1
Chaillot	48.87	2.29	Paris	16	0.00	Hotel	French Restaurant	Asian_Restaurant	European_Restaurant	Art Museum	1
Charonne	48.85	2.41	Paris	20	0.00	Cluster	Supermarket	Asian_Restaurant	Bar	Hotel	1
Combat	48.88	2.38	Paris	19	0.00	French Restaurant	Pool	Restaurant	European_Restaurant	Asian_Restaurant	1
Croulebarbe	48.83	2.35	Paris	13	0.00	Cluster	Asian_Restaurant	French Restaurant	Bar	Sandwich Place	1
Epinettes	48.89	2.32	Paris	17	0.00	Cluster	Asian_Restaurant	French Restaurant	Restaurant	Bakery	1
Gare	48.83	2.37	Paris	13	0.00	Cluster	Asian_Restaurant	Hotel	Sandwich Place	French Restaurant	1

However the resulting list is quite long. Kmeans was run with k equals 4, which is the optimum from our analysis, but a bigger K will provide a reduced list of neighbourhoods that should be even more relevant than the previous. We run again the algorithm with k equals 6, and we get indeed a reduced list. To be noticed, all of them belong to the same cluster too when running the algorithm with k equals 4, which gives some confidence in the results.

Neighborhood	Latitude	Longitude	City	arrondissement	prix_m2	1st Most common venue	2nd Most common venue	3rd Most common venue	4th Most common venue	5th Most common venue	Cluster	Cluster6
Saint Jean - Belcier	44.83	-0.56	Bordeaux	Bordeaux	4577.33	Cluster	Hotel	Nightclub	Supermarket	Asian_Restaurant	1	0
MATABIAU	43.61	1.45	Toulouse	Toulouse	4380.18	Cluster	Hotel	French Restaurant	Bar	Asian_Restaurant	1	0
Arts-et-Métiers	48.87	2.36	Paris	3	0.00	Cluster	French Restaurant	Asian_Restaurant	Hotel	Wine Bar	1	0
Plaine de Monceaux	48.89	2.30	Paris	17	0.00	Cluster	French Restaurant	European_Restaurant	Hotel	Bakery	1	0
Val-de-Grâce	48.84	2.34	Paris	5	0.00	Cluster	Asian_Restaurant	French Restaurant	Bar	Hotel	1	0
Saint-Vincent-de-Paul	48.88	2.36	Paris	10	0.00	Cluster	Asian_Restaurant	French Restaurant	Hotel	African_Restaurant	1	0
Saint-Georges	48.88	2.33	Paris	9	0.00	Cluster	Hotel	French Restaurant	European_Restaurant	Asian_Restaurant	1	0
Père-Lachaise	48.86	2.40	Paris	20	0.00	Cluster	Asian_Restaurant	Bar	Bakery	Bistro	1	0
Porte-Saint-Denis	48.87	2.35	Paris	10	0.00	Cluster	Asian_Restaurant	Hotel	French Restaurant	Bakery	1	0
Petit-Montrouge	48.83	2.33	Paris	14	0.00	Cluster	Hotel	Asian_Restaurant	French Restaurant	European_Restaurant	1	0
Charonne	48.85	2.41	Paris	20	0.00	Cluster	Supermarket	Asian_Restaurant	Bar	Hotel	1	0
Necker	48.84	2.31	Paris	15	0.00	Cluster	French Restaurant	Asian_Restaurant	Hotel	Café	1	0
Montparnasse	48.84	2.33	Paris	14	0.00	Cluster	French Restaurant	Hotel	European_Restaurant	Asian_Restaurant	1	0
Grenelle	48.85	2.29	Paris	15	0.00	Cluster	Hotel	French Restaurant	Asian_Restaurant	Bistro	1	0
Gare	48.83	2.37	Paris	13	0.00	Cluster	Asian_Restaurant	Hotel	Sandwich Place	French Restaurant	1	0
Epinettes	48.89	2.32	Paris	17	0.00	Cluster	Asian_Restaurant	French Restaurant	Restaurant	Bakery	1	0
Croulebarbe	48.83	2.35	Paris	13	0.00	Cluster	Asian_Restaurant	French Restaurant	Bar	Sandwich Place	1	0
Villette	48.89	2.37	Paris	19	0.00	Cluster	Hotel	Bar	Café	French Restaurant	1	0

b. Discussion

As mentioned before, the distribution of number of venues per area let us think there is a bias, with less venues referenced under Foursquare for the cities of Toulouse and Bordeaux than for Paris, and this is probably due to the referencing being less exhaustive for those 2 cities. This creates some bias when comparing neighbourhoods from those 3 cities. It would be interesting to get venues from another source, and compare the distributions.

We have tried also to reduce the number of different type of venues, this work could be continued, as we have focused mainly on restaurants where this typing could obviously be improved, but there are other categories which are too precise and then which don't participate to our comparison of neighbourhoods. It is only a matter of time.

Although the 5 most common venues and the housing price are very useful data which are easy to understand and to use, the result of our clustering is more difficult to interpret. It would be necessary with more time to get additional data to help understanding the results and confirm how relevant our clustering is.

If we want to go further in the neighbourhoods characterisation, data about the sociology of each area would be interesting to consider for example. It would be also very helpful to provide some kind of indicators relative to schools, which are often a key criteria for families: distance to school, reputation of the nearest school, etc.

Eventually, regarding the code itself, calculation of the average housing price per neighbourhood is a bit long, this part of the code at least could be optimised.

4. CONCLUSION

Although our code could be optimised and additional data could prove very helpful in characterising the different neighbourhoods, we have provided a first screening which results in themselves are convenient and which set the basis for a more thorough analysis.