

Finding the best location for an apartment in Paris

by Olivier Chauffert-Yvart - May 2020

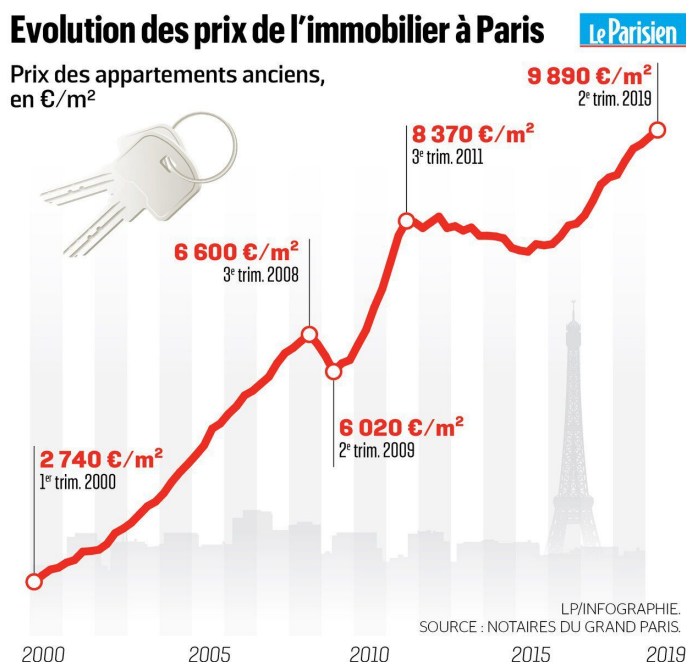
1. Introduction

1.1. Background

Paris is the capital of France and ranks second among the most visited cities in the world¹. As the economic heart of the country, it is also the most populated French city with **2.19 millions** inhabitants as of 2017². Its surface divided in **20 districts** - called *arrondissements* in French - is quite small compared to other capitals: **105.40 km²**. As a Parisian looking for a new place to live, I wondered if data science could help me find the best location for my next apartment, by leveraging open data and information available freely on the Internet.

1.2. Problem

Paris' attractiveness in terms of tourism and economic dynamism, lined with a low increase of new buildings, have caused a **sharp increase in prices** in the last 20 years and make it very difficult to find a place to live. Whether you want to rent a flat or buy an apartment, the competition with other candidates will be fierce.



Real estate price evolution (€/m²) between 2000 and 2019 in Paris, France³

As you can see on this graph, the increase trend from year 2000 is quite steep. The mean price was 2740€ per square meter (m²) in Q1 2000, and it has now exceeded 10.000€/m² in 2020³.

The Parisian market knew two recessions, consequences of the 2008 financial crisis, but they are already part of history. Maybe the Covid-19 pandemic will cause a new decrease of prices.

In this context, how could someone leverage data to choose the best location and find his or her next apartment?

¹ Business Insider: "[The 19 most visited cities around the world in 2019](#)"

² Wikipedia: dzd

³ Le Parisien: "[10 000 euros le m² à Paris : l'évolution du prix de l'immobilier dans la capitale en 5 graphiques](#)"

1.3. Interest

The project I chose to develop may be a source of interest for two categories of people:

- a. **Individuals** planning to buy an apartment in Paris
- b. **Real estate professionals** who have a mandate to find an apartment for their client, with specific research criteria about the location

2. Data acquisition, cleaning and preparation

2.1. Sources

As mentioned above, I tried to find free sources of information on the Internet. My research of data aimed to get a better understanding of **the city** of Paris, its **population**, its **districts**, the repartition of **venues I like to go to**, and its **housing market**.

- A. **Wikipedia**⁴ as a source for administrative details and demographic information;
- B. **Data.gouv.fr**⁵, the open platform for French public data, driven by the Government;
- C. **Foursquare API**⁶, to retrieve the most common venues for a given district of Paris, among a list of selected categories of venues I enjoy;
- D. **Le Bon Coin**⁷, ("The Good Corner"), a website similar to Craigslist or Gumtree, which lets individuals sell nearly anything, including houses and apartments.

2.2. Cleaning & Preparation

The cleaning and preparation operations were the most time-consuming of the project, as data sources were not always well formatted nor reconciled.

- A. For Wikipedia, the districts were listed in a table with population and inhabitants density along history; I kept only the most recent population count, which was from 2017, and calculated the corresponding density (which was missing). I also added some information needed to reconcile these data with the other datasets of the project.
- B. On the open data platform of the French government, I found a GeoJSON file describing the districts of Paris. Two districts had their centers in the middle of the woods, which was not appropriate for a housing research. I managed to modify these centers in more urban areas.
- C. The Foursquare API was leveraged to retrieve data about the most common venues in each district, among a list of categories I selected according to my preferences: bars, nightclubs, comedy clubs, concert halls, music festivals and breweries.
- D. Last but not least, I chose to scrape housing offers on Le Bon Coin website, which aims to match supply and demand, to get a real-life vision of the flats available for acquisition in Paris. I had to define research criteria, like my budget or the minimum number of rooms, identify the profile of a result page to extract a list of offers, and do the same process to extract useful information out of each offer page (140+). The most challenging part was to avoid being blocked by anti-bots cybersecurity measures in place: I made a code running slowly, to mimic the behaviour of a regular human Internet user.

⁴ Wikipedia: [List of Paris districts](#)

⁵ Data.gouv.fr: [Arrondissements](#)

⁶ Foursquare API: [Developers portal](#)

⁷ Le Bon Coin: [French classified ads website](#)

2.3. Aggregation

In order to use all these data simultaneously in *pandas* dataframes, I decided to use the **postal codes** and **district numbers** of each *arrondissement* as matching keys. It was essential to reconcile demographic, geographic, social and commercial information I had to get a holistic view of each district and choose wisely where I would establish my next residence in the future.

3. Methodology

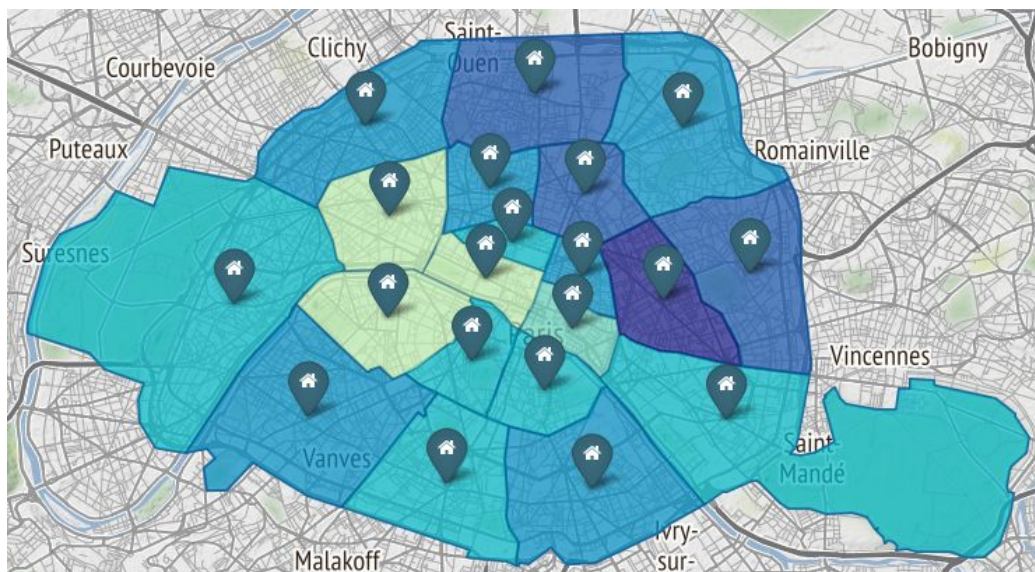
3.1. Exploratory data analysis

3.1.1. First I decided to get administrative information on the city of Paris and its population, in order to build basic knowledge of the subject. **Wikipedia** has a dedicated page about the 20 districts of Paris, and it seemed to be a good starting point. Therefore I scraped this page, cleaned the data, kept only the most recent information about population (2017), and used it as well as surface data to calculate the **population density** (inhabitants/km²) of each *arrondissement*.

```
arrondissements.head()
```

	Num.	Name	Surface (ha)	Pop. 2017	Dens. 2017	PostalCode
0	1er	Louvre	183	16395	8959	75001
1	2e	Bourse	99	21042	21255	75002
2	3e	Temple	117	34389	29392	75003
3	4e	Hôtel-de-Ville	160	28370	17731	75004
4	5e	Panthéon	254	59631	23477	75005

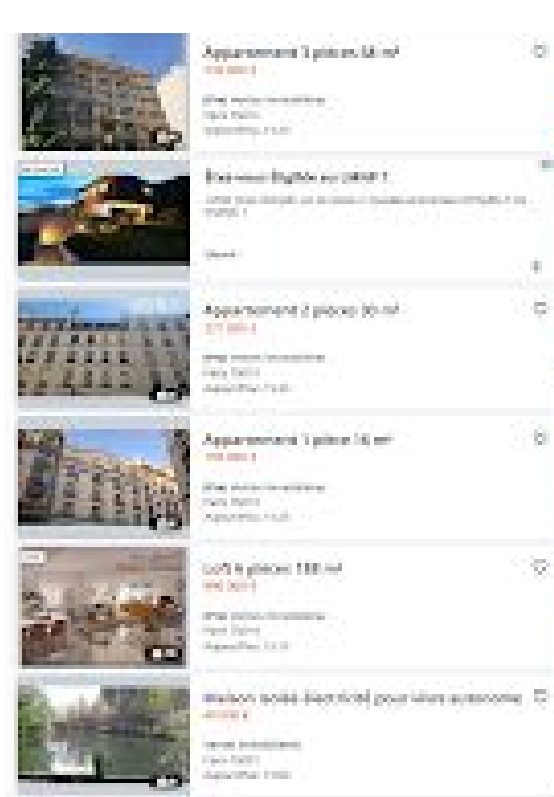
3.1.2. Second, I retrieved **geographic data** from the **French government open data portal**. The GeoJSON file included for each *arrondissement* the location of its center and its legal borders. This would be of great value for data visualization on a map. As two arrondissements - namely 16th and 12th - encompass large woods, I customized the GeoJSON file to relocate their centers in more urban areas. I visualized this on a Choropleth map of Paris and got first insights on how the population is distributed in the city (see below). Markers show the centers of each district.

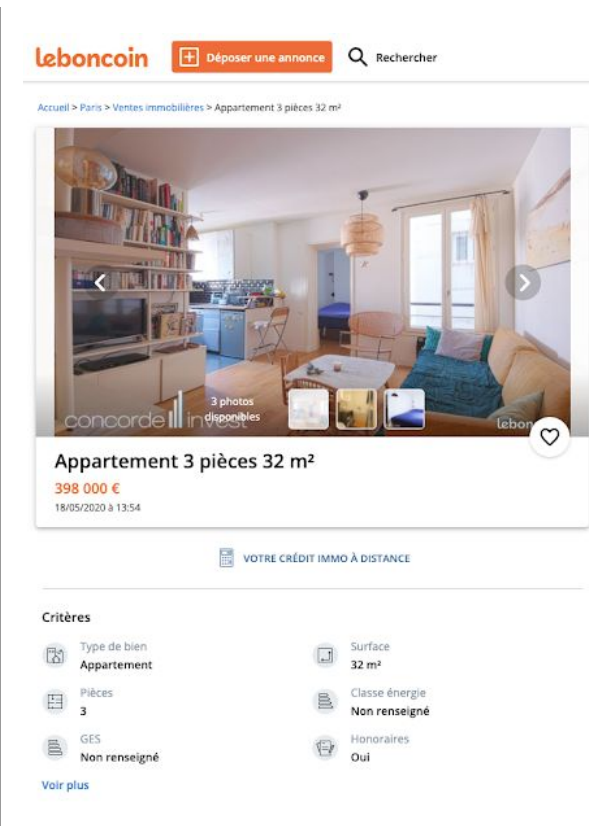


3.1.3. Then I leveraged the **Foursquare API** to get information on the **most common venues per arrondissement**, among selected categories I enjoy: bars, nightclubs, comedy clubs, concert halls, music festivals and breweries. It was interesting to perform the process learnt during the labs of the Applied Data Science module on the city of Paris. The starting point of research was the center of each *arrondissement*, and I limited the radius to 500 meters. I first studied one *arrondissement*, then replicated the process on the 20 districts. It gave me enough data to define a profile per arrondissement, depending on the occurrence of this or that kind of venue.

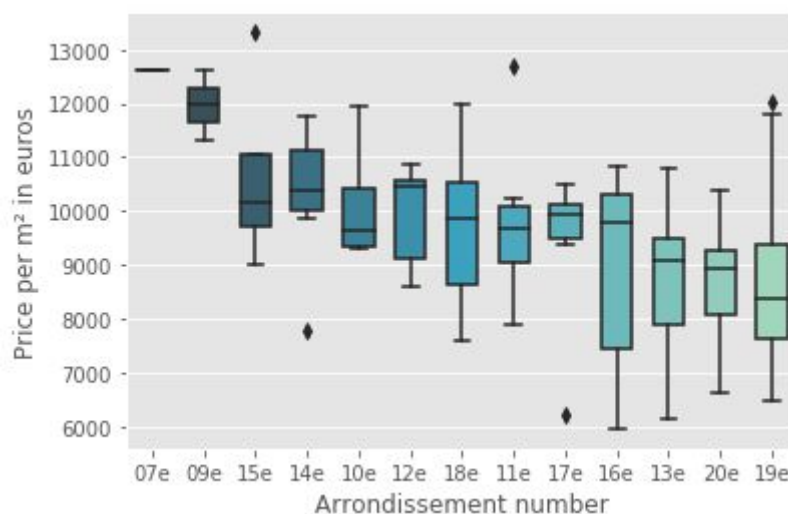
Venue	Venue Latitude	Venue Longitude	Venue Category	Num.	Code
Le Café Blanc	48.862719	2.339578	Bar	75001	1er
Le Musset	48.863811	2.334020	Bar	75001	1er
Jangal	48.864535	2.335138	Nightclub	75001	1er
Bistrot du Jardin	48.861694	2.344087	Pub	75001	1er
Cocorico	48.859124	2.328991	Bar	75001	1er

3.1.4. Finally I challenged myself to scrape real-world data on **real estate offers in Paris**, from the most famous French classified ads website: **Le Bon Coin**. I studied the format of its URLs, result pages and offer pages and had to deal with anti-bots measures. On the left you can see an excerpt from the result pages with 3 offers: I used the result pages (example shown left, below) to retrieve the following: URL, title, price, location (postal code). I had to explore each offer page to get the surface of the apartments, as it was not necessarily precised in the offer's title. Below on the right is an example of a classified ad for an apartment in *Paris 15ème* (15th district).



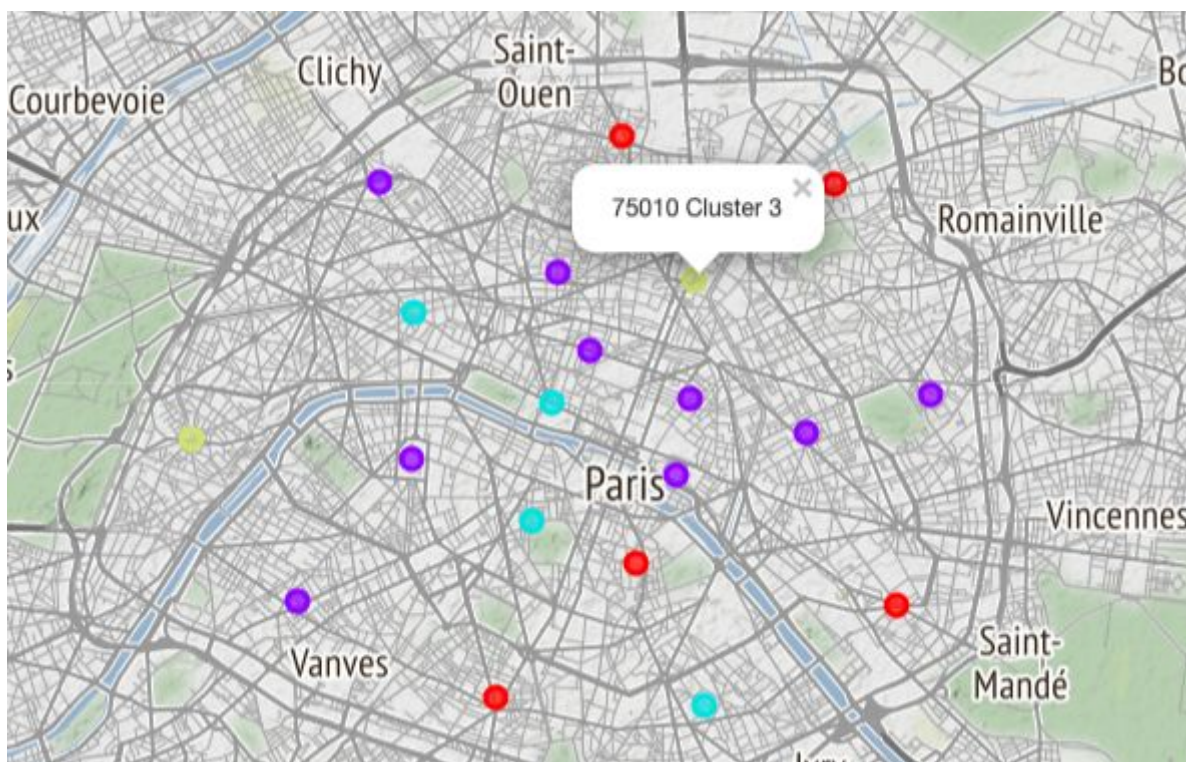


With this load of information, depending on the budget (between 600k€ and 900k€) and minimum number of rooms (4) defined, I was able to aggregate 140 offers from 13 *arrondissements*. After some preparation, which lead me to drop 4 offers whose data was not accurate (missing surfaces or wrong location), I analyzed the distribution of the prices (see below).



3.2. Machine learning

Foursquare data is useful to learn the “anatomy” of a location. I used these data to **cluster** the *arrondissements* of Paris based on the most common venues located in 500m from their centers, according to my criteria. I used the **k-means algorithm** to do so and build **4 clusters** (below).



Concerning bars, nightclubs, comedy clubs, concert halls, music festivals and breweries, the *arrondissements* indicated by the same color belong to the same cluster, and therefore **look alike**.

The example above shows that *arrondissements* 10 and 16 (west), indicated by green points, are similar regarding these kinds of venues and were put in the cluster 3.

Similarly, cluster 0 composed of *arrondissements* marked by red dots shows interesting venues: here is an example of the data behind **cluster 0**, which fits best my interests in sharing glasses with my friends:

Cluster 0: Bars & Wine bars

```
paris_merged.loc[paris_merged['Cluster Labels'] == 0, paris_merged]
```

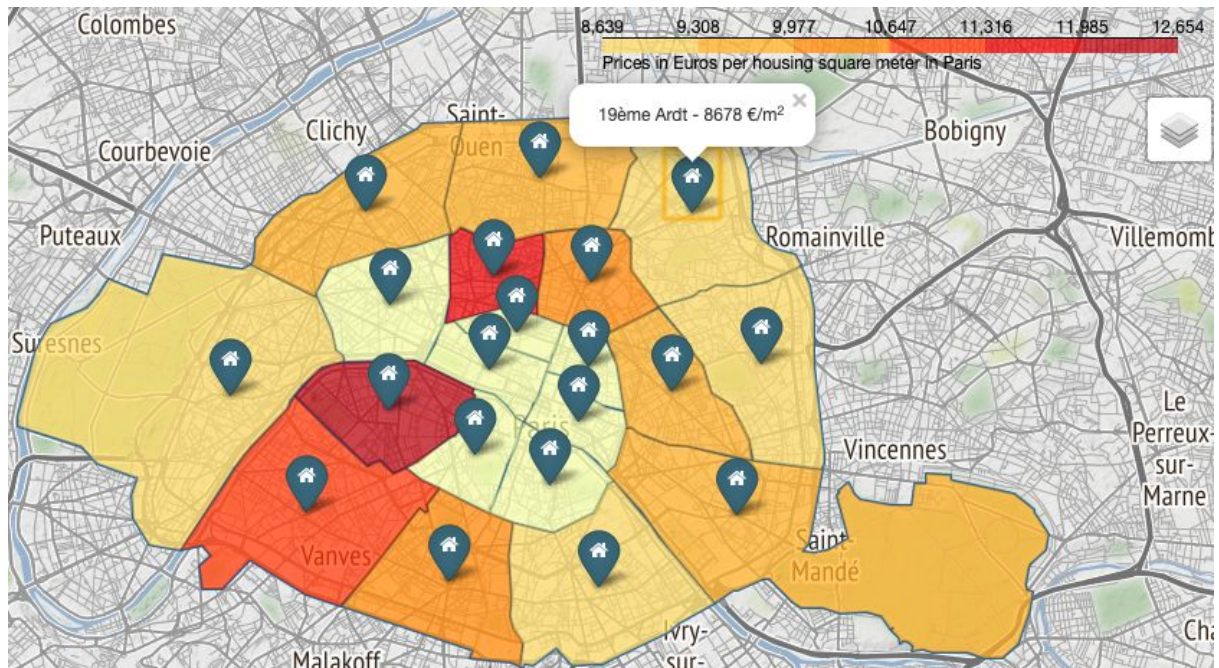
	Name	PostalCode	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
4	Panthéon	75005	0	Bar	Wine Bar	Pub
11	Reuilly (hors bois de Vincennes)	75012	0	Bar	Wine Bar	French Restaurant
13	Observatoire	75014	0	Bar	Wine Bar	French Restaurant
17	Buttes-Montmartre	75018	0	Bar	Dive Bar	Comedy Club
18	Buttes-Chaumont	75019	0	Bar	Brewery	Beer Bar

In the *Buttes-Chaumont* (19th *arrondissement* - Postal code 75019), we can see that the 3 most common venues are **Bars**, **Breweries** and **Beer bars**.



4. Results

The result of these analyses is a **map** where you can visualize the **mean price per square meter** (€/m²) for each *arrondissement*. The Choropleth function of Folium gives us a scale of colors from yellow to red, indicating from the cheapest to the most expensive neighborhood matching our search criteria. As indicated in the project notebook, no offer was found for districts 1 to 6 and 8 (in yellow below) which probably were above my budget limit (900k€). Moreover, prices tend to decrease proportionally with the distance from the center of Paris, and the cheapest districts are located in the east: 13th, 20th and 19th arrondissements. These are also where we found the largest amount of house property offers.



According to the mean prices and the “nightlife” clusters I built, I know now that I can focus my apartment research and efforts in the **19th arrondissement of Paris**, where I will find **venues I like, prices I can afford and an acceptable choice of apartment offers**.

The apartment research is also facilitated thanks to our dataframe with clickable links leading directly to each offer (excerpt below):

```
df_offers[df_offers['PostalCode'] == '75019'].style.format({'PostalCode': coloring, 'Price':
```

	PostalCode	Price	Surface	Price_per_m²	Title	Link	Num.
12	75019	885,000 €	75 m²	11,800 €	Appartement 4 pièces 75 m²	Offer	19e
15	75019	865,000 €	72 m²	12,013 €	Appartement T3-T4 vue Tour Eiffel avec balcon	Offer	19e
41	75019	760,000 €	90 m²	8,444 €	Appartement Canal de l'Ourcq-Parc de la Villette	Offer	19e
44	75019	875,000 €	104 m²	8,413 €	Appartement 5 pièces 104 m²	Offer	19e
59	75019	680,400 €	105 m²	6,480 €	Appartement 5 pièces 105 m²	Offer	19e
61	75019	735,000 €	88 m²	8,352 €	Appartement 5 pièces 88 m²	Offer	19e
79	75019	820,000 €	98 m²	8,367 €	Appartement 4 pièces 98 m²	Offer	19e
80	75019	880,000 €	88 m²	10,000 €	Appartement 4 pièces 88 m²	Offer	19e
82	75019	810,000 €	117 m²	6,923 €	Appartement 6 pièces 117 m²	Offer	19e
89	75019	729,000 €	101 m²	7,217 €	Appartement 5 pièces 101 m²	Offer	19e

5. Discussion

During the course of this project, I noted several improvements which I could have implemented with more time on my hands.

- a. Limiting the research to Le Bon Coin creates a bias, as it gives a limited vision of the market. A better solution could be to scrape and mix information from several websites, like SeLogger, PAP, Logic-Immo... (French real estate websites).
- b. The research criteria on Le Bon Coin also limited the scope of retrieved offer. For a professional use of my program, an exhaustive analysis of the real estate market in Paris should be conducted: we could limit the results only to the location, not on the price or number of rooms.
- c. I selected categories of venues matching my interests, resulting in biased clusters. For an objective vision of the profile of each district, the 'category' parameter should not be used in the Foursquare API request.
- d. Only one offer was found in the 7th *arrondissement* (the most expensive of our analysis): we could have dropped it as it is not enough information to get an idea of the mean price per square meter in this district.
- e. The Foursquare analysis retrieved venues at a distance of 500m from the center of each *arrondissement*. Therefore these venues could be located in another district but we considered they did not, because the postal code of each venue was rarely available. An improvement could be to use a larger radius around the center of each district.

6. Conclusion

During this project I used several data science approaches and techniques learnt during the Applied Data Science module, such as **k-means clustering**, **data scraping** and **geodata visualization**. It was a good opportunity to use real world data and implement a use case for the real estate industry.

An extension of this project could be to aggregate a larger amount of Paris housing offers from several leading real estate websites, use data science to identify the features which have the biggest impact on the price of an apartment or house, and build a model capable of saying if an offer is above or below the market price.