Lifestyle Analysis of Paris

Alexandre Boittin
May 2020

Introduction

Background

Paris is an amazing, multicultural, diverse, dynamic and compact city. Paris is also one of the most expensive cities of the world. After living in the Paris Area for 10 years, my girlfriend and I decided to buy an apartment in Paris to settle in a neighbourhood we like.

Problem

Paris is divided into 20 arrondissements, 80 administrative neighbourhoods, 1181 cadastral sections and 41465 cadastral plots.

Our criteria used to define a good location are:

- Presence of prefered restaurants, activities, métro stations
- Number of parks, opened markets, minimarket, schools
- Distance from the center of Paris
- Housing buying price by meter square

This report will determine if a correlation exists between the previous three independent variables (qualitative and quantitative) and the mean buying price by meter square (target) within a neighbourhood.

In a second stage, we'll try to come out with the most attractive neighbourhoods based on our weighted requirements.

Other correlations could also be assessed:

- What parameter impacts the most the area buying price?
- Does AirBnb houses impact the area buying price?

Data Acquisition and Cleaning

Data Description

To consider the problem the following data sources are used:

Source	Description	Type
Newtable	Last opened restaurants	Website + Personal Notebook
Paris Data	Arrondissements + Administratives Neighborhoods coordinates	API + API
	Houses rental cap 2019	API
	Events and activities	API
	Parks and green spaces	API
	Opened Markets	API
RATP	Global transportation offer (schedules, localisations, lines)	API
Inside Airbnb	Available hosts list for temporary rents	API
Overpass	OpenStreetMap API	API
Etalab	Past real estate buying prices (last 5 years)	Source Code
INSEE	Population	Datasets

More details:

- **Newtable** is a website of restaurants my girlfriend uses to consult whenever she wants a trendy place. It mentions the opening date showing the last tendencies. That's also the opportunity to apply some web scraping applications (e.g. BeautifulSoup).
- **Etalab** is French Government department whose missions are to manage opening and sharing policy of public data (since October, 2019). One of its dataset contains all the last real estate mutation buying prices of France territory. It is updated twice a year.
- **RATP** is the Parisian metro operator. Access to API requires an application.
- **Inside Airbnb** is an open API that shows all available homes and apartments. It will be used to determine the touristic places.
- **Overpass** is a read-only API that serves up custom selected parts of the OpenStreetMap data. We'll use it to geocode addresses and eventually to get other location information.
- **Paris Data** is the Paris city website in partnership with opendatasoft that gathers several datasets under the ODbL (Open Database Licence) as administrative mapping, specific places (park and markets), activities and rental price cap.

Other investigated sources:

- **Foursquare API** is an easy to use database, however it is not the most used in Paris and so doesn't match with our habits.
- **SeLoger** is the most used application for research of housing location or buying in France. However it doesn't provide an official API (only a few customized GitHub Notebooks have been made).
- **Booking.com API** is only open to affiliated partners that should have travel related contents and a high amount of daily visitors.
- **Kaggle** is a "huge repository of community published data & code" but doesn't have good data about Paris.
- **Data Ile de France** is a similar data of Paris Data but for the full Region of Ile de France.
- **Google Maps Geocoding API** should be the best Geocoding service but it is unfortunately not free (0.005 USD per Request).
- **Trip Advisor** does not grant access for purposes of data analysis, research, testing, or similar uses.
- **IGN**, **INSEE** are also rich French government data sources that could be used if needed.
- **SNCF** is the French train transportation company. It also operates some RER lines in Paris (RER C, D, E), but not available yet on the website.

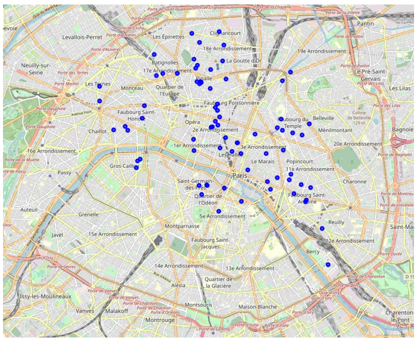
Features Selection

Source	Description	Kept features
Newtable	Last opened restaurants	All
Paris Data	Arrondissements + Administratives Neighborhoods coordinates	All
	Houses rental cap 2019	Not used
	Events and activities	Top 10
	Parks and green spaces	Not used
	Opened Markets	Not used
RATP	Metro Stations	Station positions and line
Inside Airbnb	Available hosts list for temporary rents	Not used
Overpass	OpenStreetMap API	To get locations
INSEE	Population	All

Studied Geographic division

Paris is divided into 20 arrondissements, 80 administrative neighbourhoods, 1181 cadastral sections and 41465 cadastral plots. Studied geographic divisions are the Paris arrondissements. Distances from the center are added to the dataset.

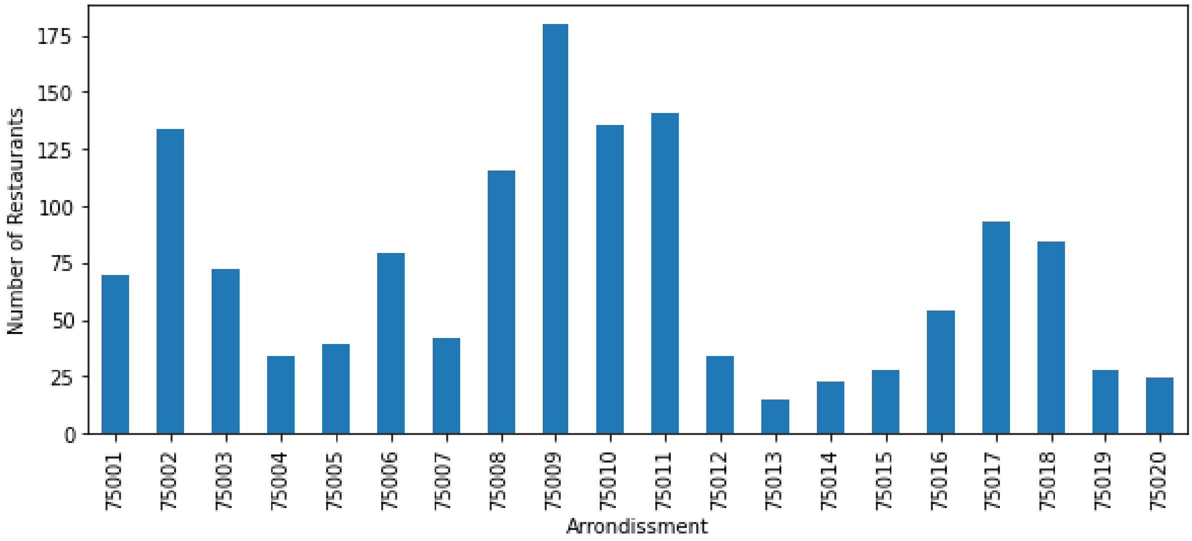
Restaurants



From web scraping, the restaurant dataset has **1485 samples**. Some data is manually cleaned (accents, wrong arrondissements: 750006 rather than 75006, whitespaces).

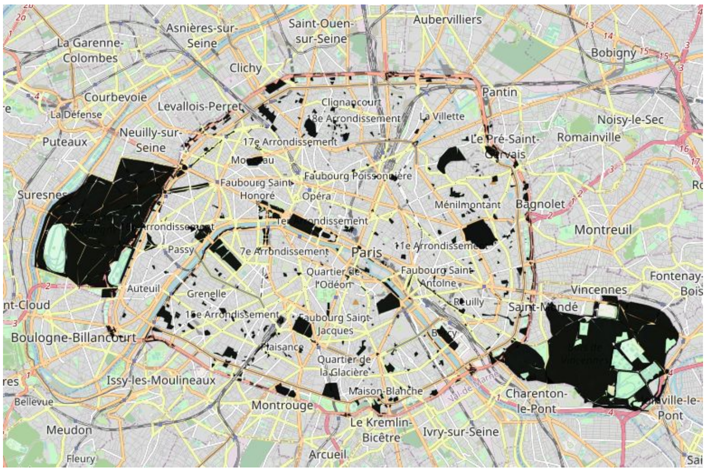
For information, this is where Dataiku has an added value (even excel manages to remove whitespaces).

Using Geocode, here are their position and the repartition per arrondissement below.



Parks

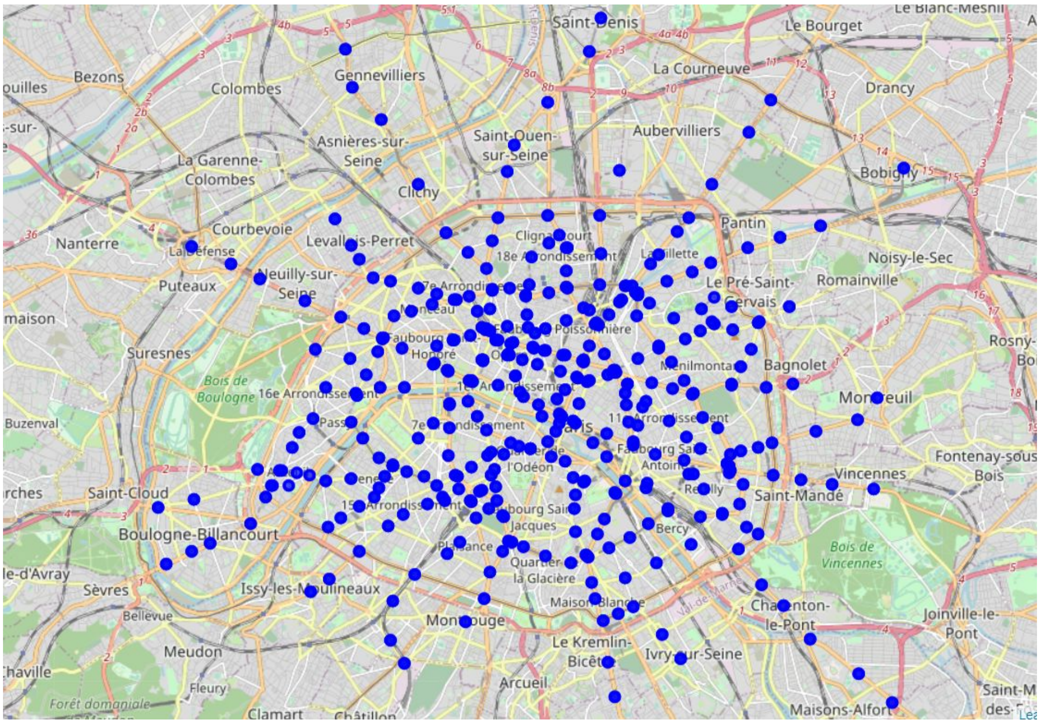
From the GeoJSON file, Paris has 973 parks or included parcels.



RATP

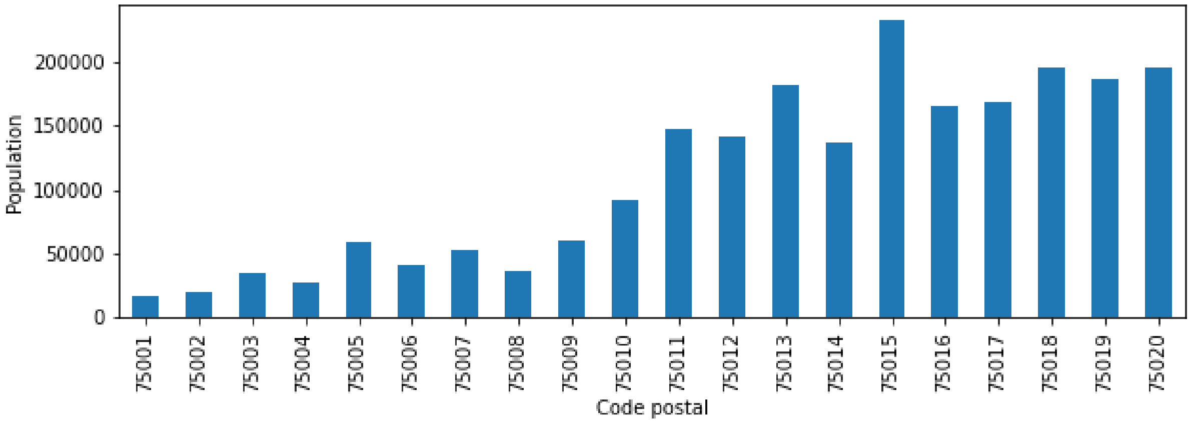
This dataset provides public transport (Metro, Bus, tram) network information, updated every month; provisional schedules, station localisations and possible correspondences. However, the RER C, D and E data are not included as these lines are operated by SNCF, nor the station localisations.
Data is stored in a GTFS format that has a .zip extension containing several .txt files. When API requests are not available (application refused for not compliant IP address), extracting and manipulating all the files on local drive is not easy with an online ipython notebook (e.g. Colab).

In our study, we’ll focus only on the metro line stations. The Paris area has 302 metro stations.



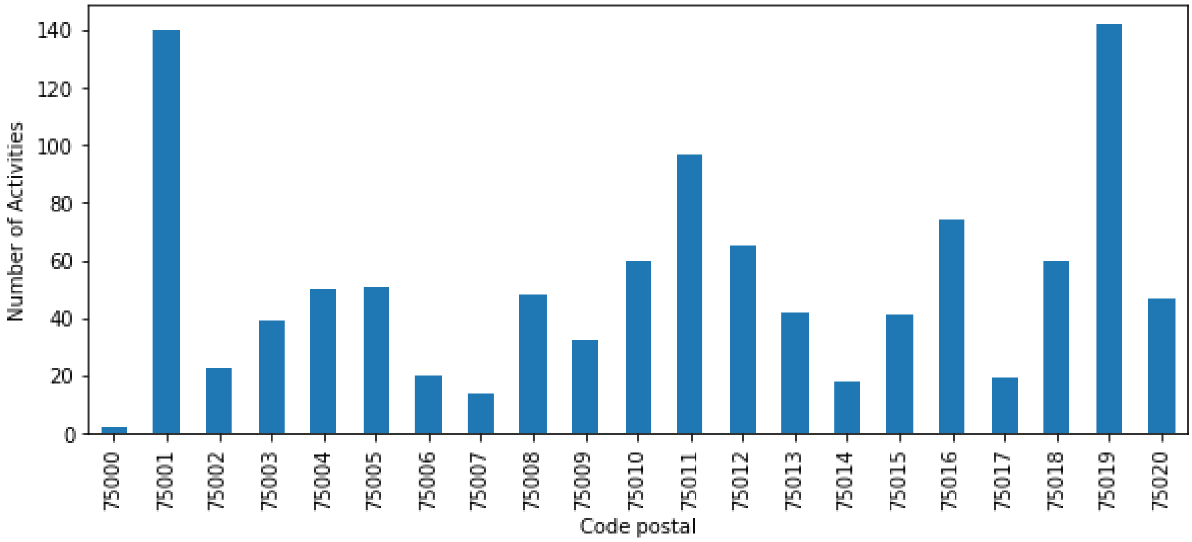
Population

From INSEE, we get the population of each arrondissement in 2016.

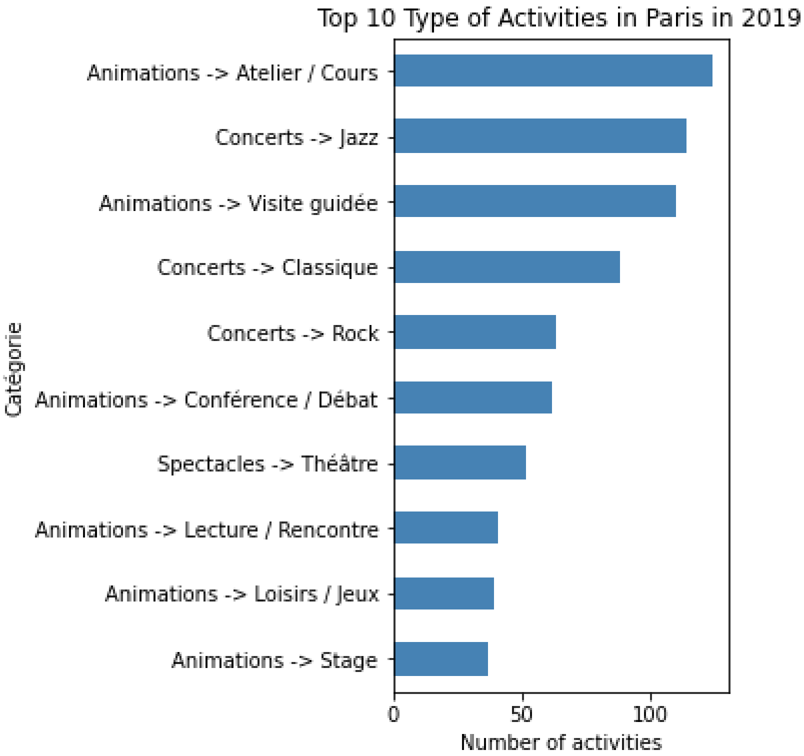


Activities

Dataset from Paris Data provides 1080 events and activities with useful information such positions, prices, schedule and classification category starting from 2019. Here are the number of event per arrondissement;



And the most common activities purposed by the city:



Activities are "one-hot" encoding and only the top 5 activities of each arrondissement are kept.

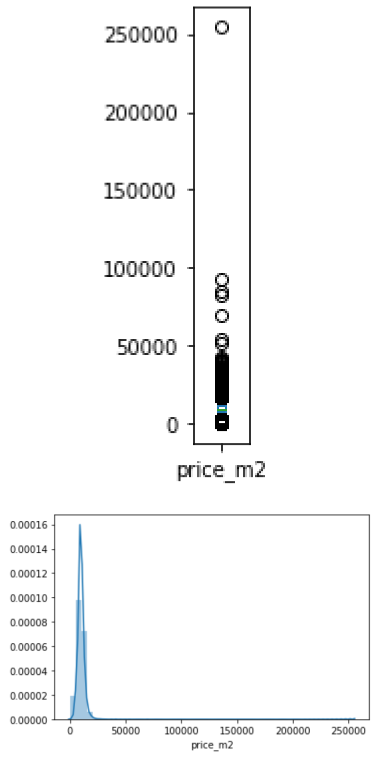
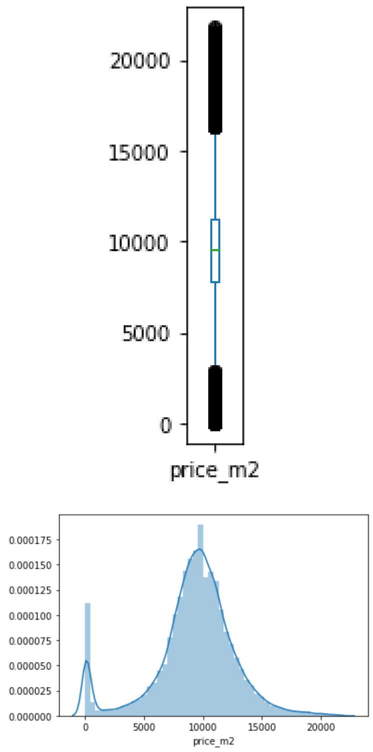
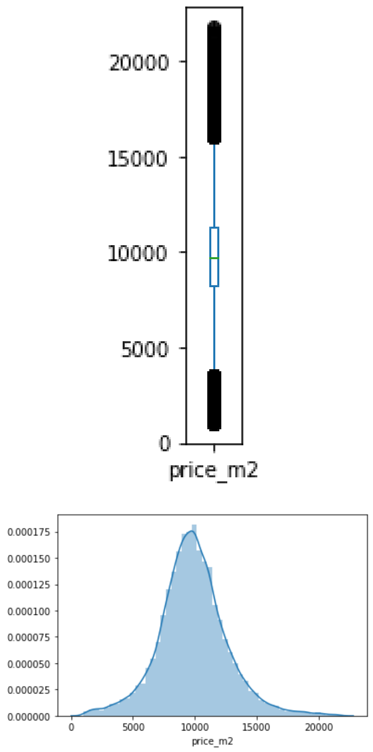
Target: Mean area buying price by meter square

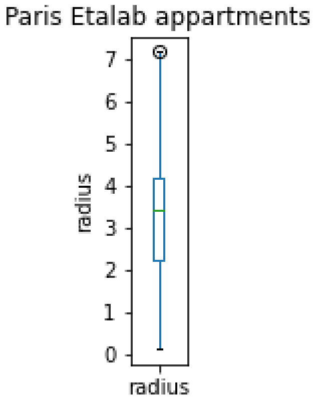
Source	Description	Kept features
Etalab	Past real estate buying prices (last 5 years)	Apartment, sold, last 2 years, without land, one mutation

Etalab uses city INSEE codes as city inputs and mutation buying prices are located in cadastral plots. All unique mutations are defined by the id_mutation. Several information are available; properties (e.g. apartment, outbuilding, commercial premises..), ground floor position and outdoor ground surface, mutation types (Sale, tender, building land), number of involved lots and mutation dates (last 5 years).

As Paris real estate prices change rapidly, **only the last 2 years (2018 - 2019) are extracted**.
However some mutations gathered several appartements on different table rows in addition to the number of lots. Mutations with the same id_mutation are removed.
We thus focus on the **sold mutations** about **an apartment with one row, one lot and without land**.

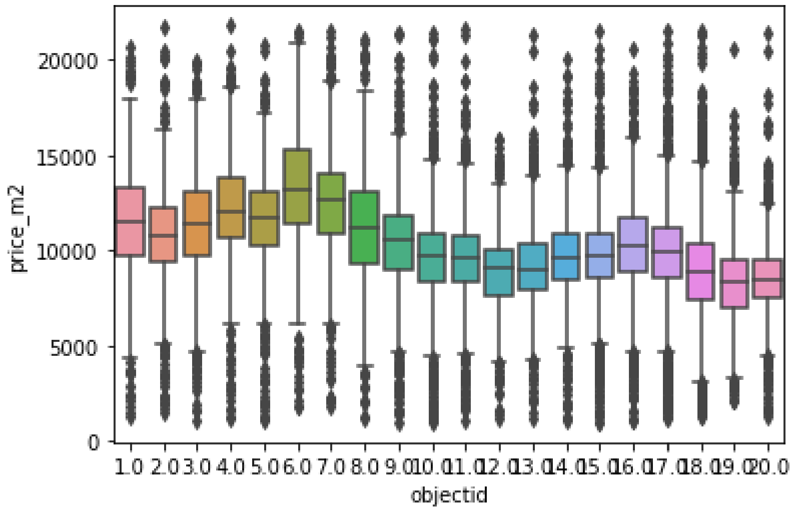
This extraction has some outliers (price/m2 > 50000 euros). Dataset is cleaned for the first time with Z-score on the price/m2 column.
Seeing so many apartments with price/m2 between 0 to 3000 seems wrong. Once values distribution is displayed, we can see a first peak at the beginning of the curve. To correct it, all mutations with price/m2 **below 1,000 euros** are removed from the dataset.

First extraction	Cleaning with Z-Score	Distribution peak Cleaning
		

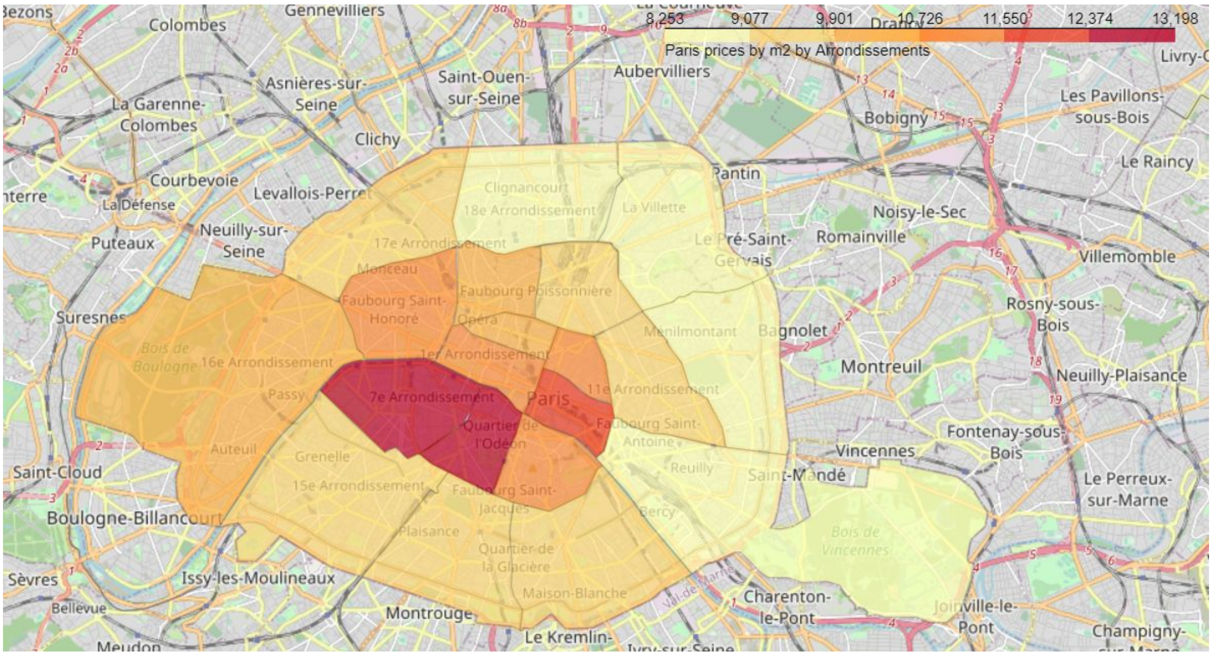


As the opposite, latitude and longitude seemed correct since the first extraction. Mutation distances using the Haversine formula vary from 0 to 7 km from the Paris center point.

After the cleaning, the final dataset has **21855 samples**. Here are the prices by meter square of each arrondissement:



And how they look like displayed on the Paris map:



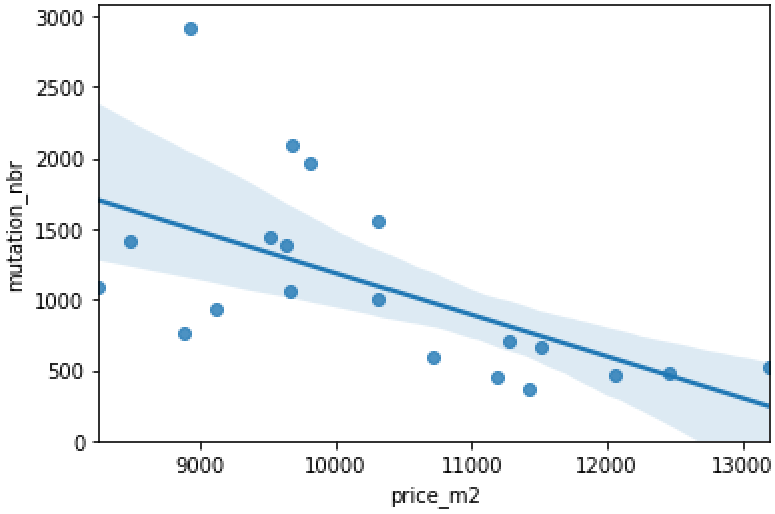
Methodology

Now let’s gather all data by arrondissements.

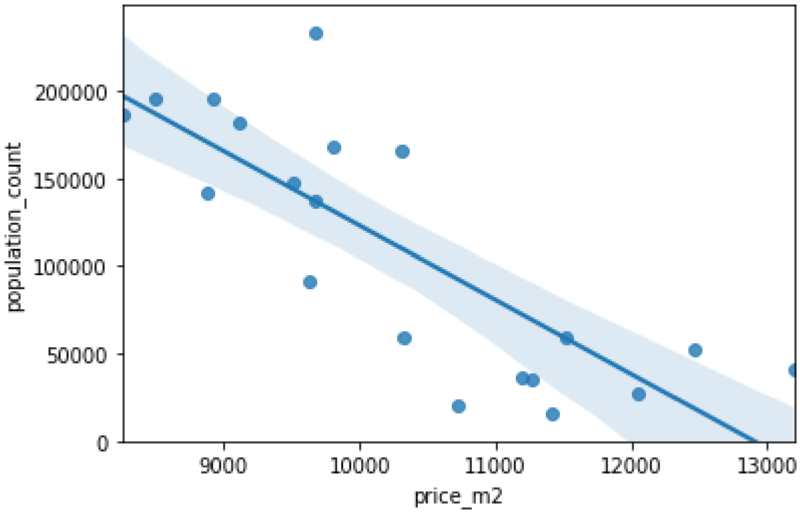
Correlations

Let’s try to find some correlations!

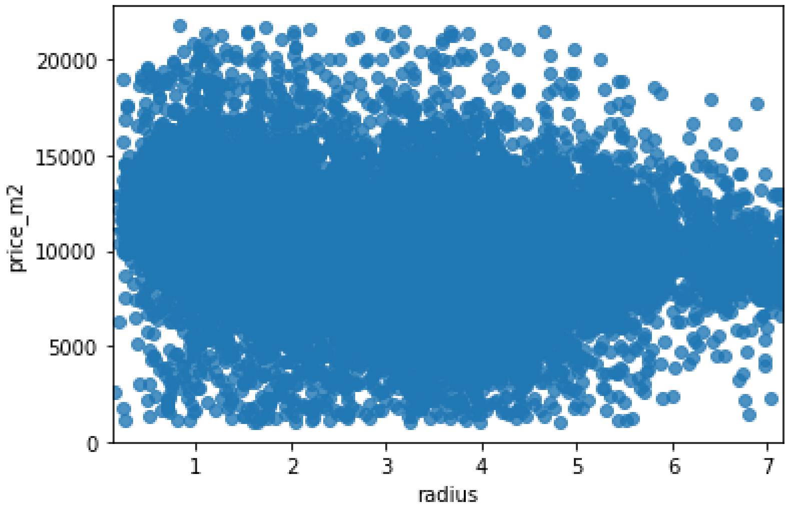
There is a small one (-0.607) between the number of mutations and the price by m2. More expensive the arrondissement is less mutations are.



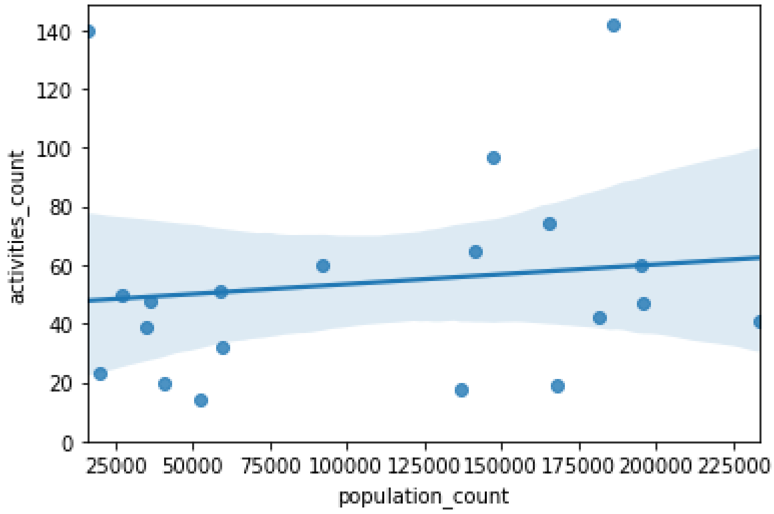
Same for the number of people living in arrondissement (-0.80). Bigger the population is, the cheaper the arrondissement is.



Unfortunately, there is no link between distance from Paris and price.

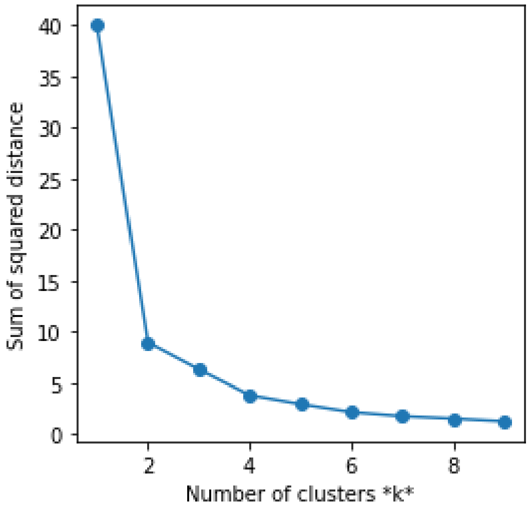


Also, there is no correlation between the number of activities and the population within the arrondissement.

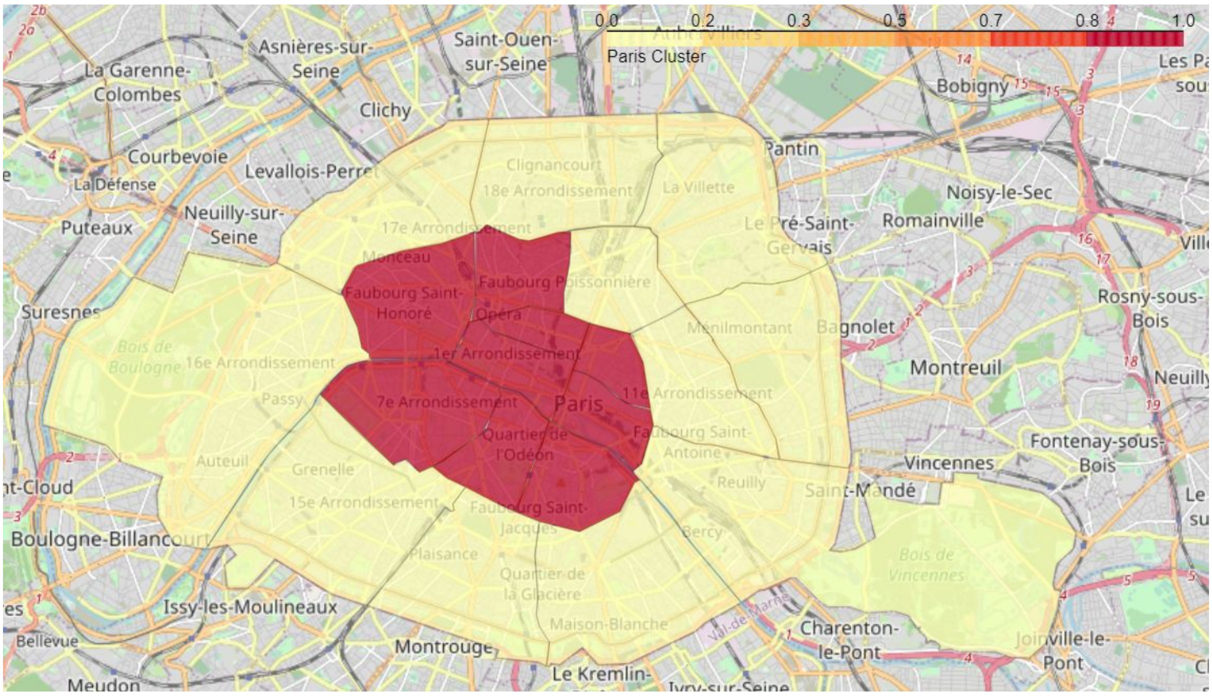


Clusterization

Let’s cluster the arrondissements using K-Means with price and population. Data is first normalized. Using Elbow method, Paris is divided in 2 clusters.



Center cluster has arrondissements with expensive mean price (11572 €/m2) and less populated (38,600 in average). Periphery is cheaper (9296 €/m2) with more population (167,539 in average).



Results

It seems really hard to find explanations of the different Paris' arrondissements prices. However we learnt that there are 2 distinct clusters.

Discussion

More data available and exploitable from the City will help to understand the price differences between Paris arrondissement.

Conclusion

I am glad to see the French administration has opened a lot of their databases since the beginning of 2019!