Capstone Project - Airbnb - Paris, France

April 29th 2020

Florence D'AMORE

1- Introduction

1-1. Background

To visit the town Paris, in France, one of types of accommodation used is Airbnb.

A lot of rentals are proposed, and it is difficult to choose the best, when you don't know the city. Airbnb has seen an enormous growth, with the number of rentals listed on its website growing exponentially each year. Making it difficult to choose a location.

1-2. Problem

This project concerns a family who wishes to visit Paris, at walk and by subway.
This family need an apartment, near market, bakery and subway.

The data science can help to prepare their visit.

2- Data acquisition

2-1. Data sources

For this project, the data sourced from :

- the Inside Airbnb website : listings.csv.gz , Paris, on March 15th 2020.
- the Foursquare location data.

The data from *Inside Airbnb* website give the updated listings of rentals by neighborhood; price; type of accommodation; minimum of nights; number of bedrooms.

And the data from *Foursquare location* give the location of market, bakery and subway.

By combining the two datasets, the family can choose the best rental for a pleasant visit in Paris.

2-2. Data cleaning

Concerning the file from Inside Airbnb website, there are 67323 rows and 106 columns.
Only the columns interesting for this project are keeped.

The columns allow to choose the best rental are : 'id', 'name', 'space', 'description','neighbourhood_cleansed', 'latitude', 'longitude', 'property_type', 'room_type', 'accommodates', 'bathrooms', 'bedrooms', 'bed_type', 'amenities', 'price', 'cleaning_fee' and 'security_deposit'.

There aren't duplicate data.
There were not missing values in the *price* columns.
But there were missing values in the *cleaning_fee* and *security_deposit* columns. I supposed that these values are "$0.00".

Prices are in dollars. Values have been made uniform.

As this project is at destination of a family, in the *room_type* column, only the type 'Entire home/apt' has been selected.

After cleaning dataset, there are 58177 rows and 16 columns.

3- Exploratory Data Analysis

The principal point which interested the family is the rental price in function of :

- Paris's neighborhood,
- the number of bedrooms,
- the size of the apartment….

3-1. Mean of rental price

The first information obtained for the price, is the following :

	Price
Count	$58 177.00
Mean	$123.17
Std	$176.40
Min	$0.00
25%	$65.00
50%	$89.00
75%	$130.00
Max	$10 250.00

The mean of rental price is $123.

As the price difference between the maximum and the mean value is significant, we fixed the maximum value at $250.00.
We supposed the other value as outliers.

By limiting the rental price to $250.00 per night maximum, we obtain the following diagram in figure 1 :

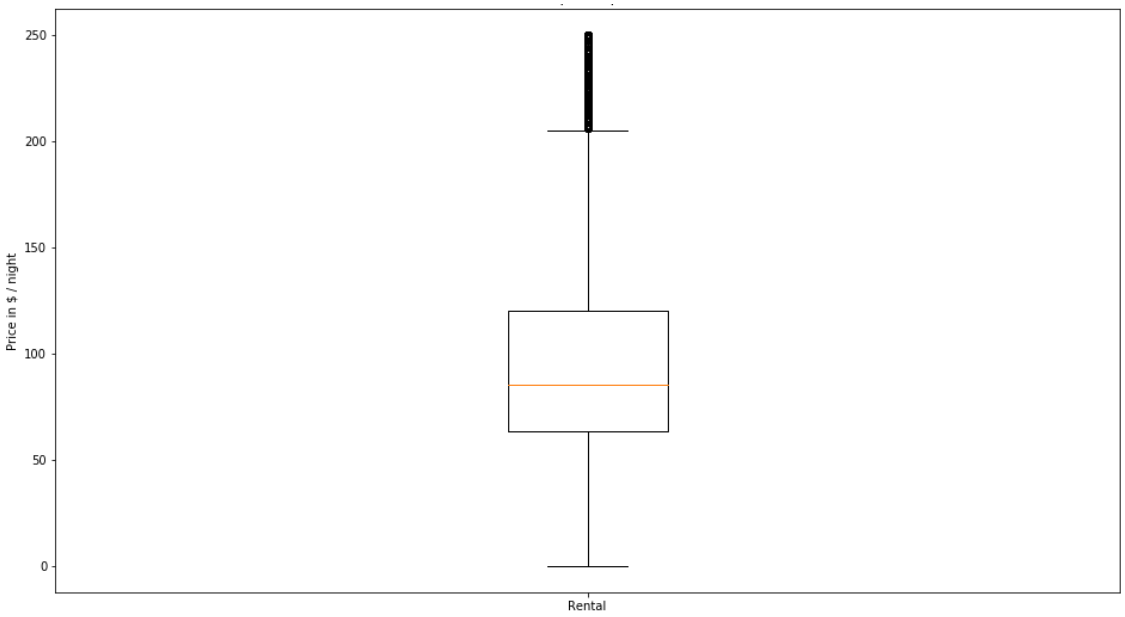

Figure 1 : Box plot of rental price

3-2. Price by coordinates

The figure 2 indicate the repartition of the rental price in according to the coordinates.
The repartition seems to indicate that in the center of the town, the price is high.

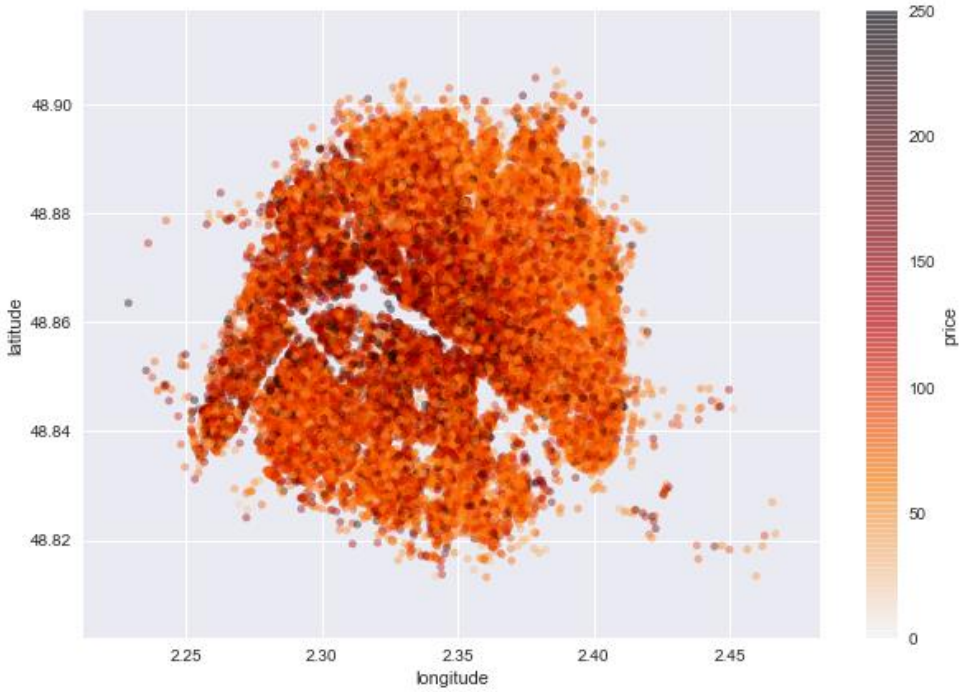

Figure 2 : Distribution of rental price according to the coordinates

3-3. Relationship between the rental price by neighborhood

The following histogram (Figure 3) highlights the larger number of districts, in Paris.

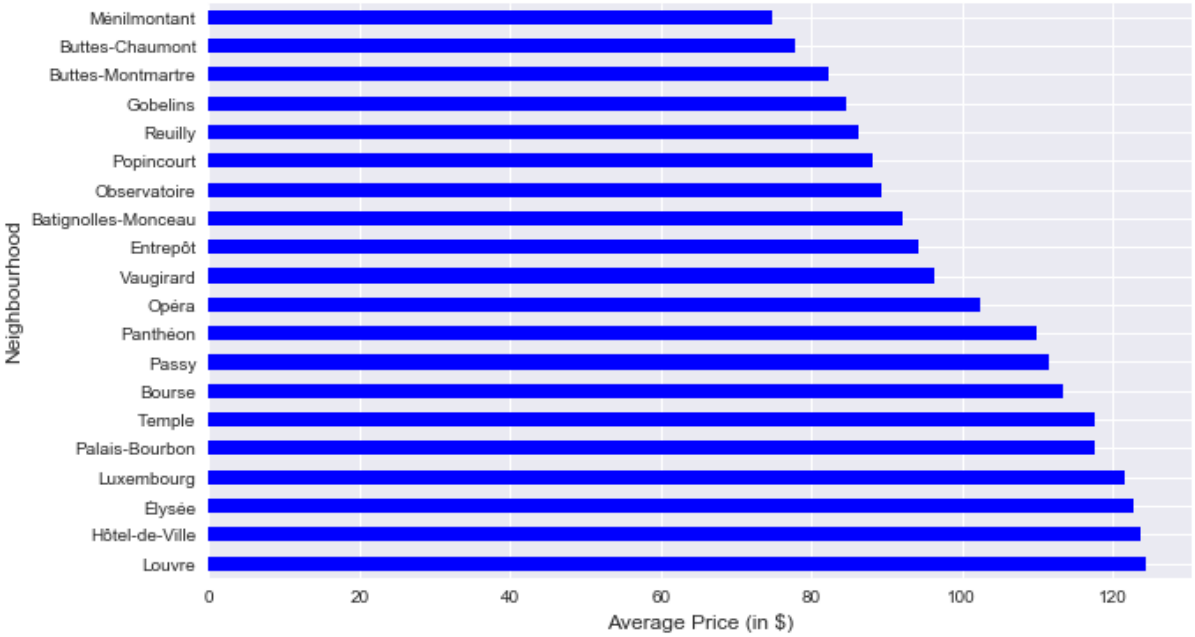

Figure 3 : Average price by neighborhood

The choropleth of rental price by neighborhood, figure 4, confirm that the most central neighborhood of Paris, are the most expensive.

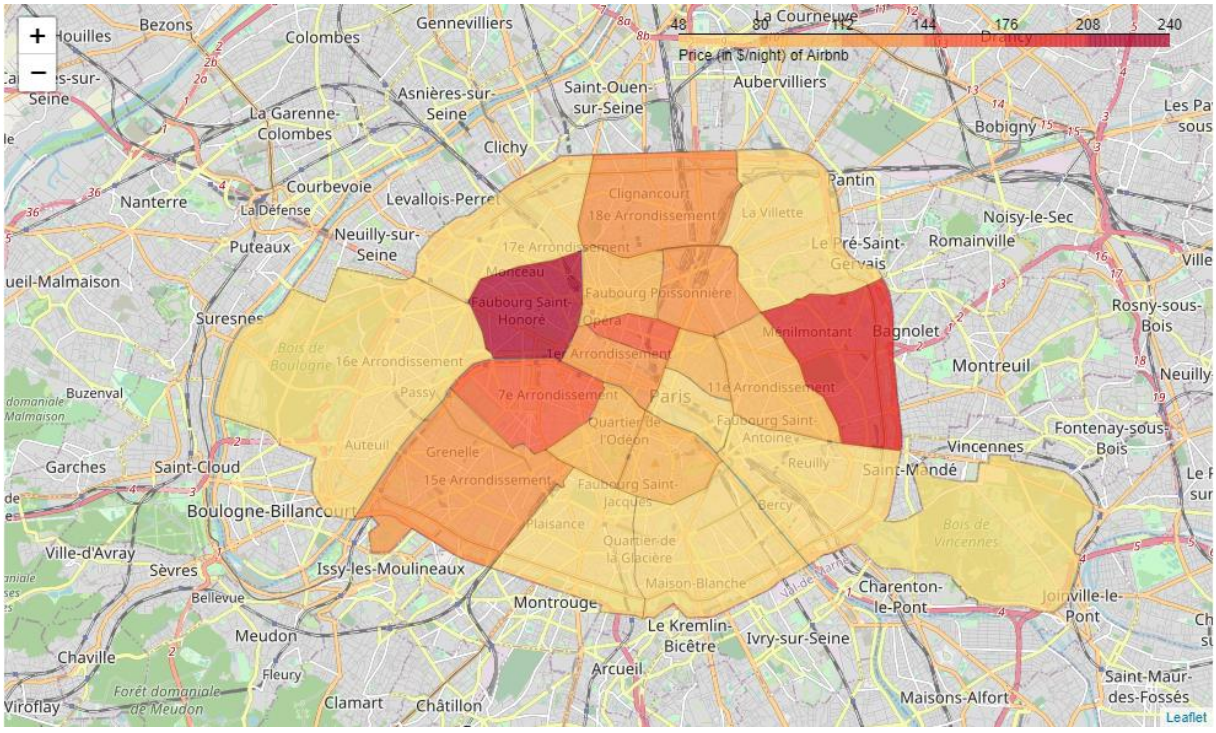

Figure 4 : Choropleth by neighborhood

3-4. Relationship between the rental price by number of bedrooms

We are interested at the distribution of the price, by the number of bedrooms and district, in the following figure 5 :

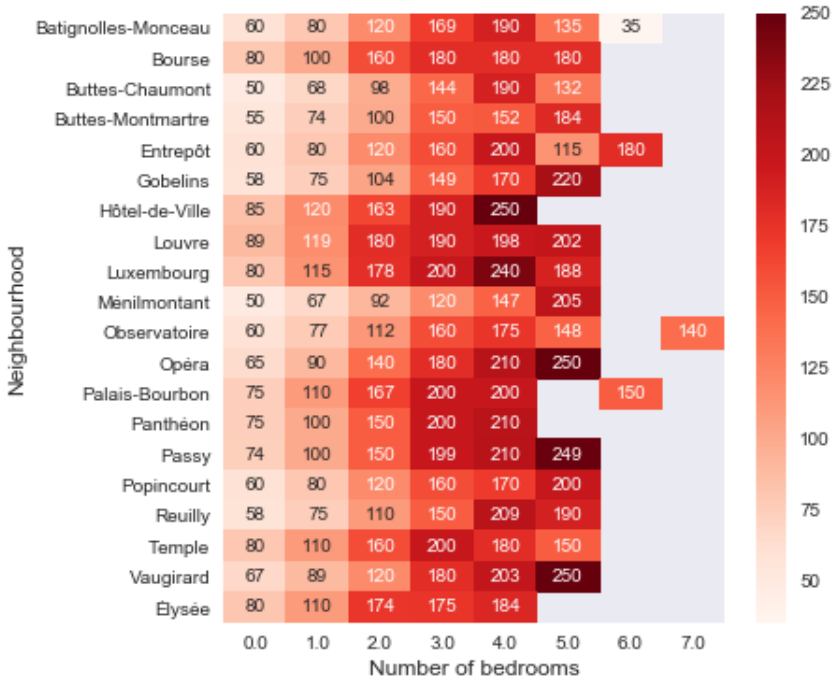

Figure 5 : Relationship between price, number of bedrooms and neighborhood

Unsurprisingly, the fewer rooms there are in rental; the lower the price is.

There is a value is outlier. At the ‘Batignolles-Monceau’ district, for 6 bedrooms, the rental is $35. We have to be careful about the results.

3-5. Amenities

The different amenities hosts offer in the rental are important. Your choice varied in function of the number of the amenities.

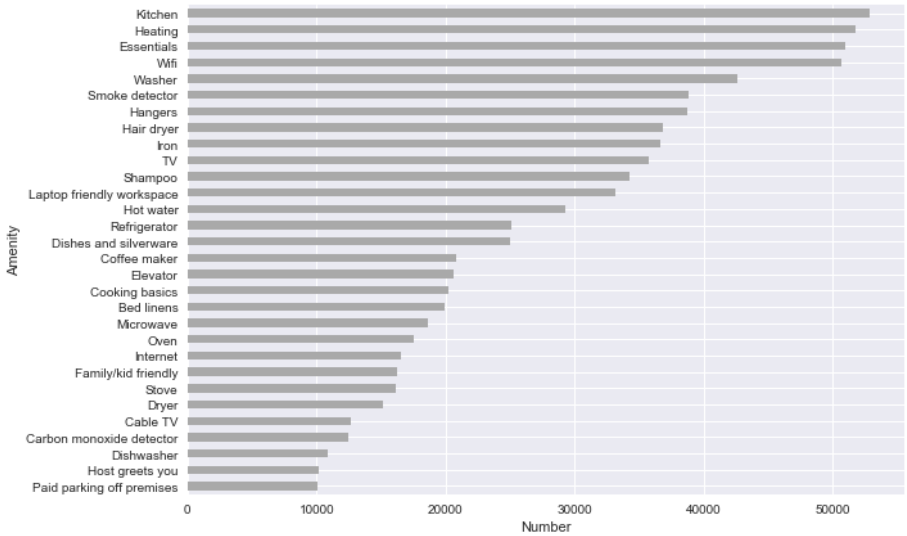

Figure 6 : Number of amenities

As show the histogram at figure 6, the majority of rental has got a kitchen, wifi, washer and TV.
Some hosts mentioned that they have carbon monoxide detector.

3-6. Relationship between the rental price by property type

There are 26 different property type proposed to rental, in Airbnb.
Some type of accommodation seems unrealistic, as an igloo, an island or a plane.

The highest rental price is for the houseboat (Figure 7). And the lower price is for campsite.

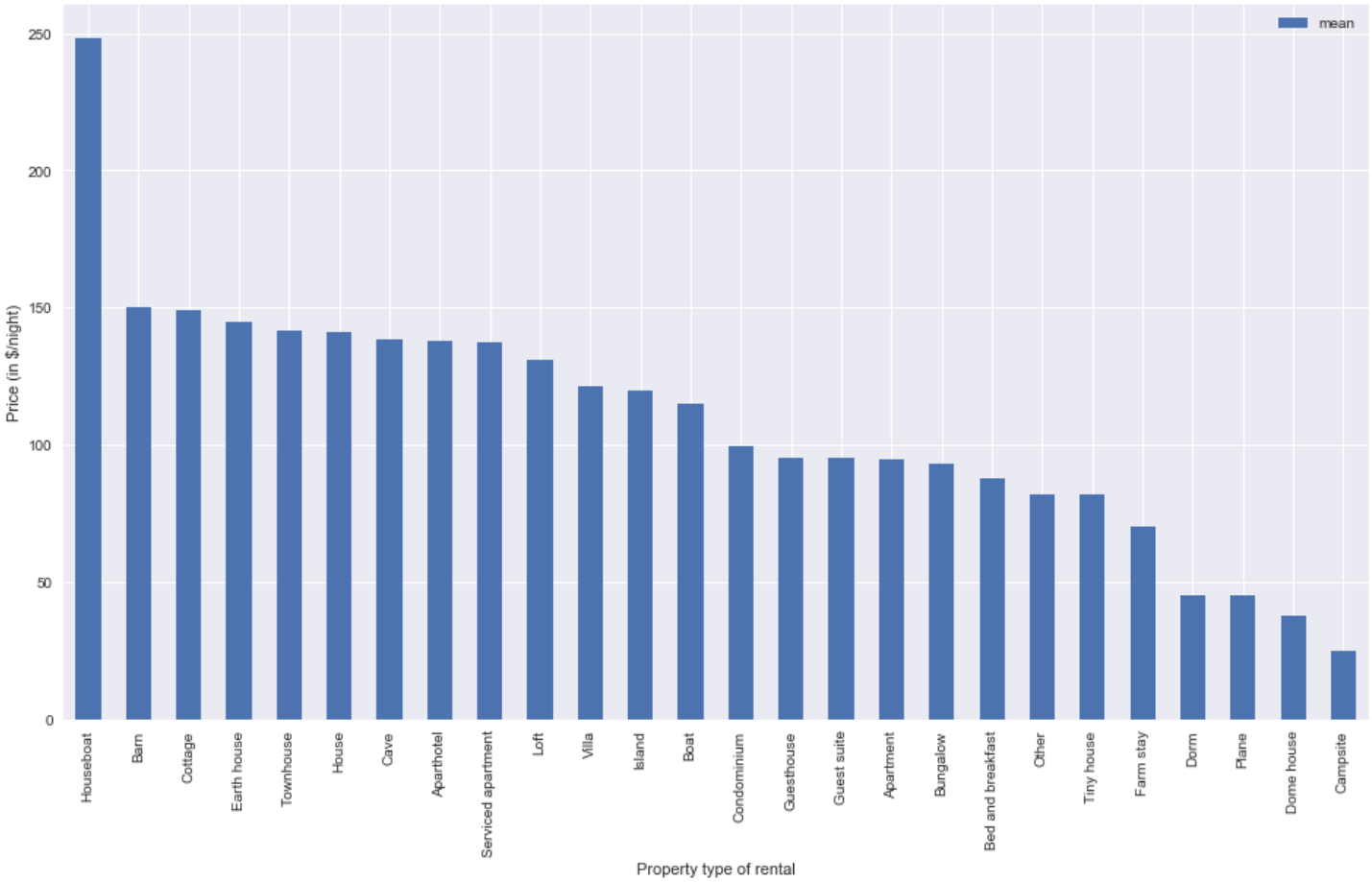

Figure 7 : Rental price by property type

3-7. Relationship between the rental price by security deposit and cleaning fee

They exist a relation between the rental price, and the amount of the security deposit and the amount of the cleaning fee.

It seems that the lower rental price, the lower security deposit and cleaning fee are.
And vice versa. (Figure 8)

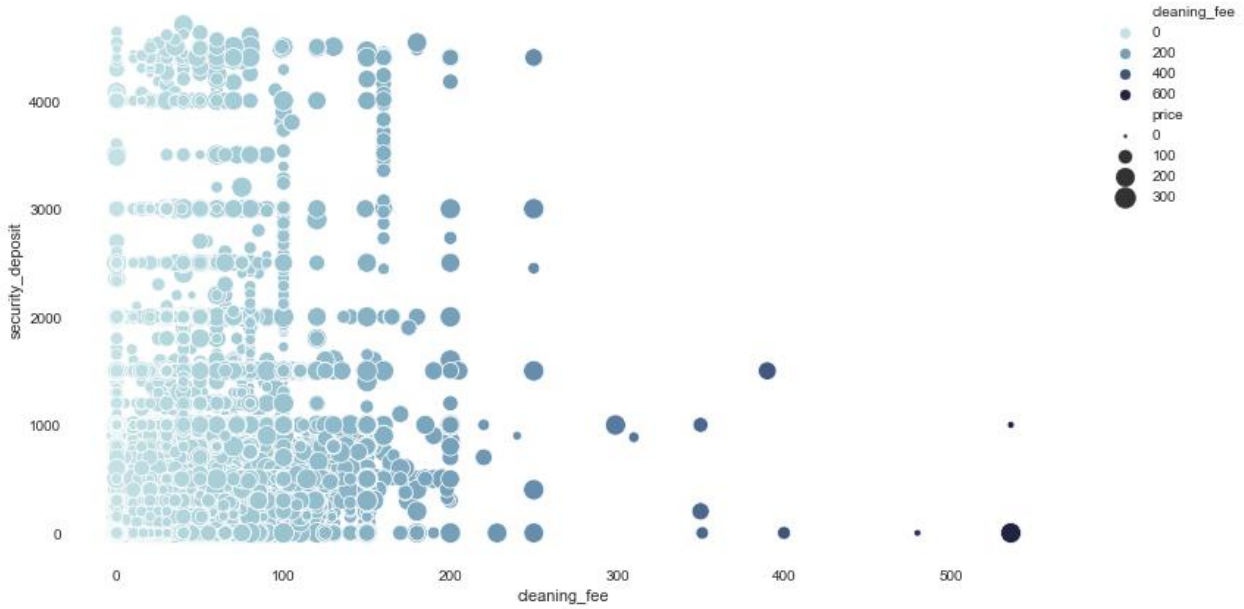

Figure 8 : Relationship between the rental price by security deposit and cleaning fee

4- Predictive modeling

An important factor in the choice of accommodation, is the surface area of the rental.
In this dataset, a lot of values of the square feet column are unknow. We aren't used this column.

But, in the description's column, there are the information that we need.

We extract the information, with the following support :
all double-digit or three-digit numbers that are followed by one of the two characters "s" or "m" (covering "sqm", "square meters", "m^2"…), and may or may not be connected by white space.

We obtain a column "size". In the following, an extract from this column :

id	description	size
3109	I bedroom appartment in Paris 14 Good restaura...	15.0
5396	Cozy, well-appointed and graciously designed s...	30.0
7397	VERY CONVENIENT, WITH THE BEST LOCATION ! PLEA...	40.0
7964	Very large & nice apartment all for you! - Su...	75.0
9359	Location! Location! Location! Just bring your ...	10.0
9952	Je suis une dame retraitée, qui propose un agr...	30.0
10586	BAIL MOBILITE 30 days (1 month) to 300 ...	10.0
10588	LONG TERM RENTAL 12 MONTHS - 9 MONTHS for STU...	12.0
10710	Very close to Place de la Concorde and Madelei...	NaN
11170	The apartment is located in the well known Qua...	31.0

A lot of values are unknow, in the size column.
We try to predicting these values, by linear regression method.

We split data and we obtain the following :

Shape of Training Data: (41529, 2)
Shape of Test Data: (25794, 2)

Shape of X_train: (41529, 1)
Shape of X_test: (25794, 1)

Shape of y_train: (41529,)

Then, we make a linear regression and we make a prediction.

We obtain the following extract :

	price	size
10710	90.0	54.717783
11798	120.0	55.967480
12452	139.0	56.758955
12887	49.0	53.009863
14757	75.0	54.092934

The size values are described in the following :

	Predict size (in m²)
Count	25 794.00
Mean	56.94
Std	13.70
Min	50.97
25%	53.47
50%	54.68
75%	56.38
Max	477.95

The mean of size is 56,94m².

As the difference size between the maximum and the mean value is significant, we fixed the maximum value at 70m².
We supposed the other value as outliers.

By limiting the predict size to 70m², we obtain the following diagram in figure 9 :

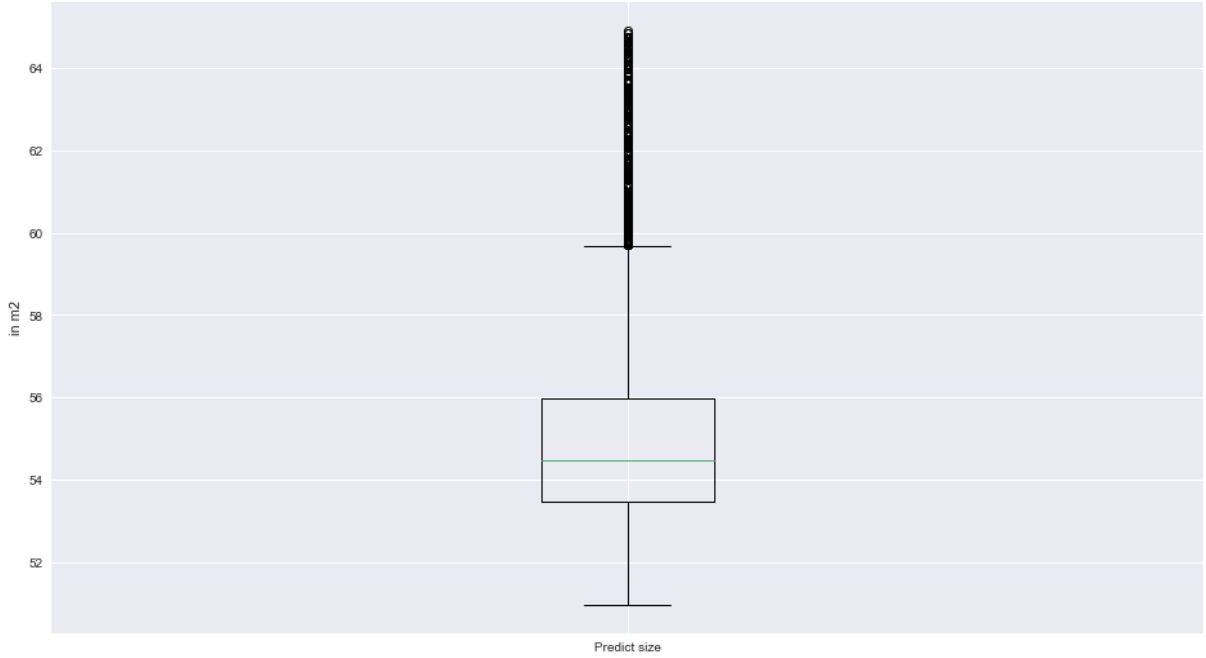

Figure 9 : Mean of the predict size

5- Foursquare location data

The foursquare location data allows to know the localization of bakery, subway and market, near the accommodation.

The figure 10 show the repartition of bakery, market and subway, in the center of Paris, with the different rental price by district.

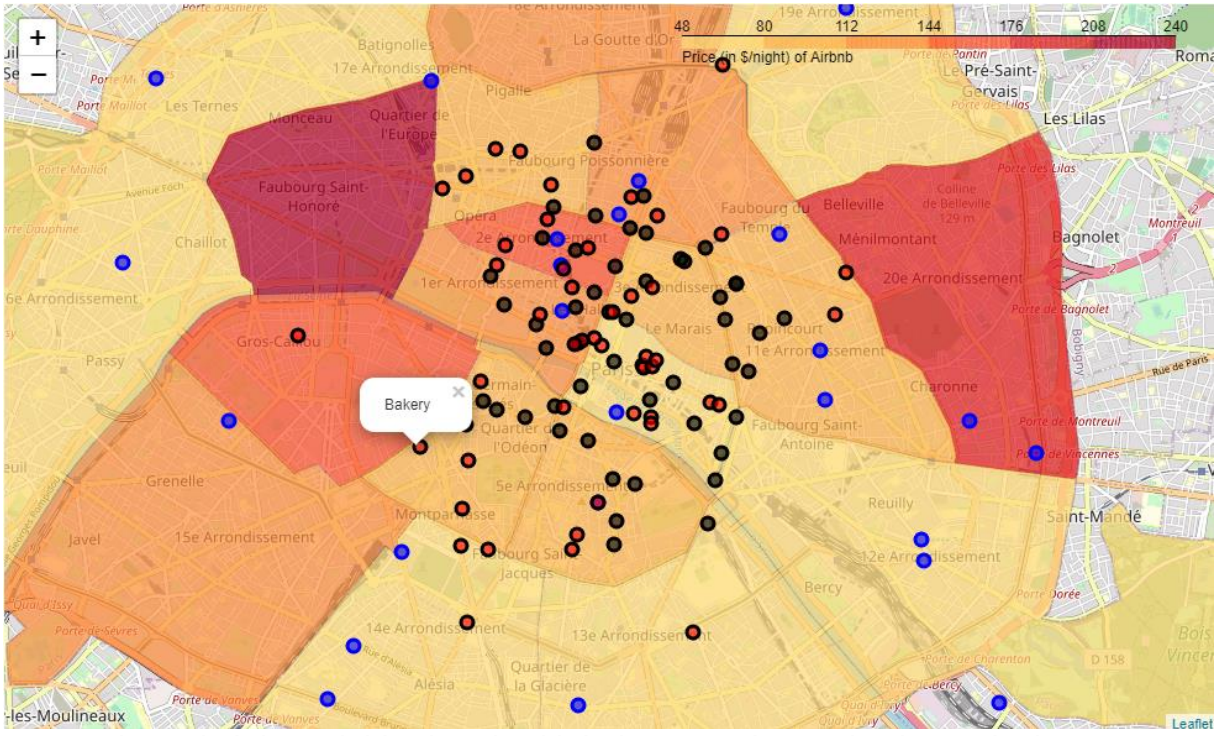

Figure 10 : Choropleth by neighborhood, with bakery, market and subway, in the center of Paris

6- Conclusion

The rental price depends of :

- the geographic position (In the center of Paris, is the most expensive)
- number of bedrooms
- the size of accommodation.

The location in the center of Paris have the most amenities (Market, Subway…), near the location.

There are some values outliers, in the dataset.
You must take all the information before booking, to avoid unpleasant surprises.

Customer satisfaction values should be added to this project.

7- Sources

- "Predicting Prices: XGBoost & Feature Engineering" , by Britta Bettendorf
- "Predicting the Improvement of NBA players", by Zhenfeng Liu : report and presentation
- "Housing Sales Prices & Venues Data Analysis of Istanbul" , by Sercan Yildiz
- GeoJSON data in Paris : arrondissements.geojson