Capstone Project - The Battle of Neighborhoods

David Hincapié

July 23, 2021

1. Introduction

1.1 Background

In today's world the number of air connections increases more every day, the possibility of traveling to any country in the world has increased from the economic point of view and of the availability of flights. Today there are flights to almost any remote place in the world. From an island paradise in the deepest part of the Pacific Ocean to the perpetual snows of Antarctica. This situation has allowed a growth of travelers worldwide who travel for various reasons such as tourism, business, living in a new place, etc. Many times, when a traveler wants to travel, he would long to have many of the comforts that he has at home in the place he is going to travel. When it comes to amenities, it does not necessarily refer to what is inside the house, but to the possibilities for fun, shopping, and other businesses that the neighborhood where you live offers. Being able to get to a new place and have facilities similar to the ones you have at home would be something satisfying for most people. An environment where you can enjoy your stay 100% as if you were in your own home.

1.2 Problem

Alejandro is on a business trip to Paris, France. This trip will last one month in this city. During his stay in paris he will also carry out some tourist activities. He is a 26-year-old Colombian youth. His hobbies are the gym and reading books. His house in Colombia is located in a neighborhood where he has easy access to a gym and a bookstore. As he has to travel to Paris, he wants to find a place to stay that has the places mentioned above nearby and that offers the possibility of finding a Colombian food restaurant or any Latin American food also nearby. Alejandro will use the Airbnb application to find a place that he likes and that fits his budget of minimum 100 euros and maximum 200 per night, but he needs more information to verify that it also adapts to the need of an environment similar to his home in Colombia. It is important that the place of lodging is as close as possible to the central neighborhoods of paris because these are safer and are close to the main tourist destinations of the city. The place must be available every day of the year, it must have at least 20 customer ratings regardless of the average rating it has, a private room would be sufficient and that the last rating would have been made minimally in the year 2020.

1.3 Interest

The main interested party will be Alejandro, who will be able to find the place that best suits his needs in Paris. The project with slight modifications could be used to carry out the same study for other people who have a need similar to Alejandro's.

2. Data acquisition and cleaning

2.1 Data sources

For this project, some Airbnb datasets will be used that are published in the link: http://insideairbnb.com/get-the-data.html . The specific datasets that will be used are related to the city of Paris, which will contain the properties available for rent in the city with all their respective information and the geographical location of the Paris neighborhoods. The data used from the dataset with the help of Foursquare will allow the project to be carried out. The datasets used are available under the Creative Commons CCO 1.0 Universal license.

2.2 Data cleaning

The datasets obtained from the Airbnb page are listed to observe the information they contain and thus select those that really serve for the development of the project. When doing the above, it is observed that of the datasets originally obtained, only two fit the project's requirements. These datasets are stored in the data frame with the names of: dt_listings and map_data. The data sets obtained from Airbnb contain orderly and clean data, therefore the cleaning process is really little or no.

The data frame dt_listings originally has a total of 63090 records and 16 characteristics. Here are some sample records and the names of these features, as well as the type of information they hold.

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights
0	2577	Loft for 4 by Canal Saint Martin	2827	Karine	NaN	Entrepôt	48.86957	2.36127	Entire home/apt	125	3
1	3109	zen and calm	3631	Anne	NaN	Observatoire	48.83191	2.31870	Entire home/apt	60	2
2	5396	Explore the heart of old Paris	7903	Borzou	NaN	Hôtel-de-Ville	48.85247	2.35835	Entire home/apt	47	1
3	7397	MARAIS - 2ROOMS APT - 2/4 PEOPLE	2626	Franck	NaN	Hôtel-de-Ville	48.85909	2.35315	Entire home/apt	90	10
4	7964	Large & sunny flat with balcony !	22155	Anaïs	NaN	Opéra	48.87417	2.34245	Entire home/apt	130	6

Figure 1 Example of the logs

```
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group               float64
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

```
#Number of rows and columns of the dataframe.
dt_listings.shape
```

```
(63090, 16)
```

Figure 2 Column data type and number of records

The records (properties) contained in the original data frame dt_listings were reduced based on the requirements requested by the interested party. Only properties that are between 100 and 200 euros are left, properties that have 20 or more customer ratings, properties that have at least one rating in 2020 or later, properties that are available at least 350 days of the year. The next step was to filter the properties that were in the central area of Paris and finally select only the properties that are a private room. By performing all these filters, the original data frame was reduced to only 12 properties that adapt to the needs of the interested party.

The map_data data frame is left the same since it has the coordinates in geojson format of the Paris neighborhoods.The number of records (neighborhoods) in the data frame is 20. Below is an example of the data frame.

	neighbourhood	neighbourhood_group	geometry
0	Batignolles-Monceau	None	MULTIPOLYGON (((2.29517 48.87396, 2.29504 48.8...
17	Bourse	None	MULTIPOLYGON (((2.35152 48.86443, 2.35095 48.8...
2	Buttes-Chaumont	None	MULTIPOLYGON (((2.38943 48.90122, 2.39014 48.9...
18	Buttes-Montmartre	None	MULTIPOLYGON (((2.36580 48.88554, 2.36469 48.8...
4	Entrepôt	None	MULTIPOLYGON (((2.36469 48.88437, 2.36485 48.8...

```
#Dataframe field types.
map_data.dtypes
```

```
neighbourhood            object
neighbourhood_group      object
geometry                 geometry
dtype: object
```

```
#Number of rows and columns of the dataframe.
map_data.shape
```

```
(20, 3)
```

Figure 3 Column data type and number of records

2.3 Feature selection

After the cleaning process, the dt_listings data frame is left with 12 records and 16 characteristics. From the columns of the original data frame, some are eliminated that are not necessary for the development of the project. The columns listed below are removed from the data frame.

```
dt_listings = dt_listings.drop(['neighbourhood_group'], axis=1)
dt_listings = dt_listings.drop(['minimum_nights'], axis=1)
dt_listings = dt_listings.drop(['reviews_per_month'], axis=1)
dt_listings = dt_listings.drop(['calculated_host_listings_count'], axis=1)
dt_listings = dt_listings.drop(['host_id'], axis=1)
dt_listings
```

Necessary Features	Features not required
Name, neighbourhood, latitude, longitude, room_type, Price, number_of_reviews, last_review, availability_365.	neighbourhood_group, minimum_nights, reviews_per_month,calculated_host_listings_count, host_id, host_name,id.

Based on some of the characteristics that remained, filtering was started to find the properties that were adapted to the needs of the interested party. With these properties already found,the raw material for the use of foursquare was available.

The data frame map_data is left with the original characteristics without any modification

3. Exploratory Data Analysis

For the exploratory analysis, it begins with an overview of the dt_listings data frame that contains all the information on the properties. We proceed to see the statistical details of the data frame, the data types of the characteristics, the correlation between the columns and the number of records it contains.
Once these simple analyzes of the data frame had been carried out, some more complex ones were carried out. It was already known that the total of properties available for rent on Airbnb in the city of paris was 63,090 but it was necessary to know how many of these were in each neighborhood. For this, the properties were grouped by neighborhood, a table and a horizontal bar graph were generated in order to observe these figures.

	total
Buttes-Montmartre	6971
Popincourt	5918
Vaugirard	4782
Entrepôt	4502
Batignolles-Monceau	4168
Ménilmontant	3652
Buttes-Chaumont	3641
Passy	3176
Opéra	3099
Reuilly	2738
Temple	2736
Observatoire	2491
Gobelins	2219
Bourse	2132
Panthéon	2111
Hôtel-de-Ville	1999
Luxembourg	1957
Élysée	1759
Palais-Bourbon	1685
Louvre	1354

Figure 4 As you can see, the neighborhood with the largest number of registered properties is Buttes-Montmartre.

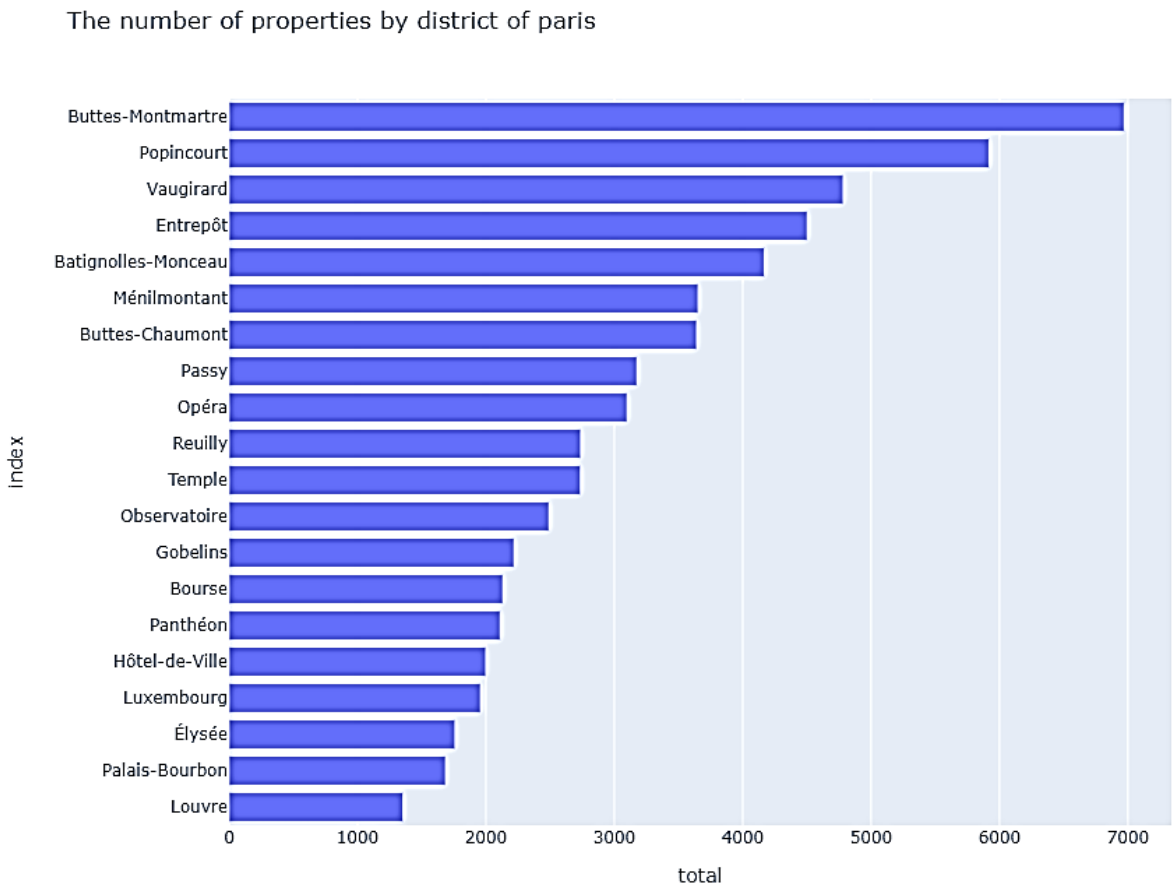


Figure 5 Horizontal bar graph representing previous values.

To take advantage of the data frame map_data that contains the coordinates of the neighborhoods of Paris, a choropleth map was generated with the aforementioned figures. Due to the impossibility of adding the names of the neighborhoods on the choropleth map, a map of the neighborhoods of Paris with their names is added in order to make a parallel between the two.

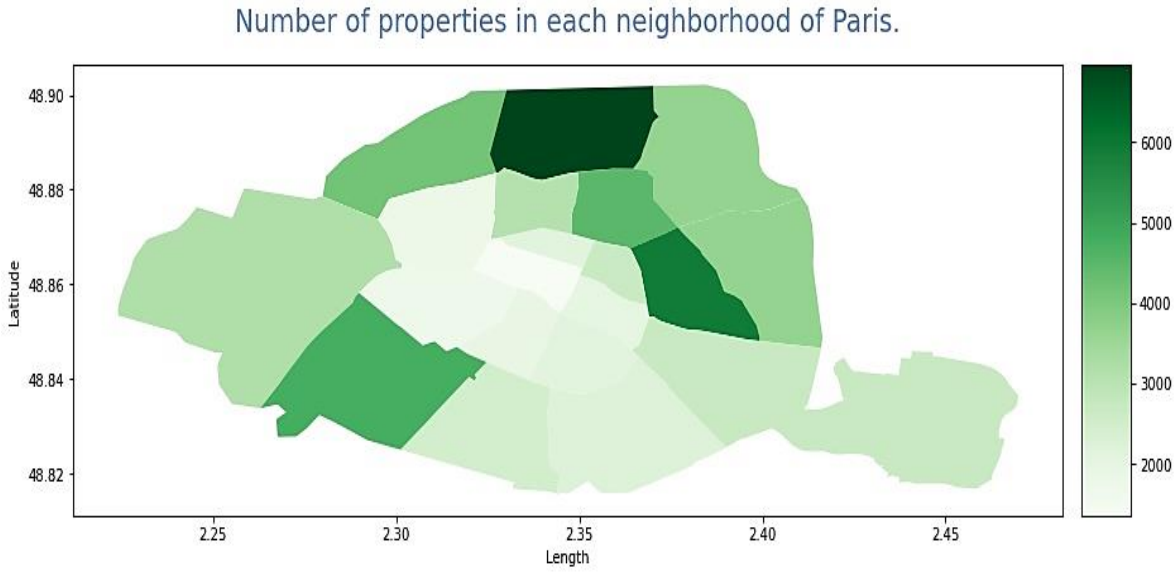


Figure 6 The darkest parts represent the neighborhoods where there are more properties

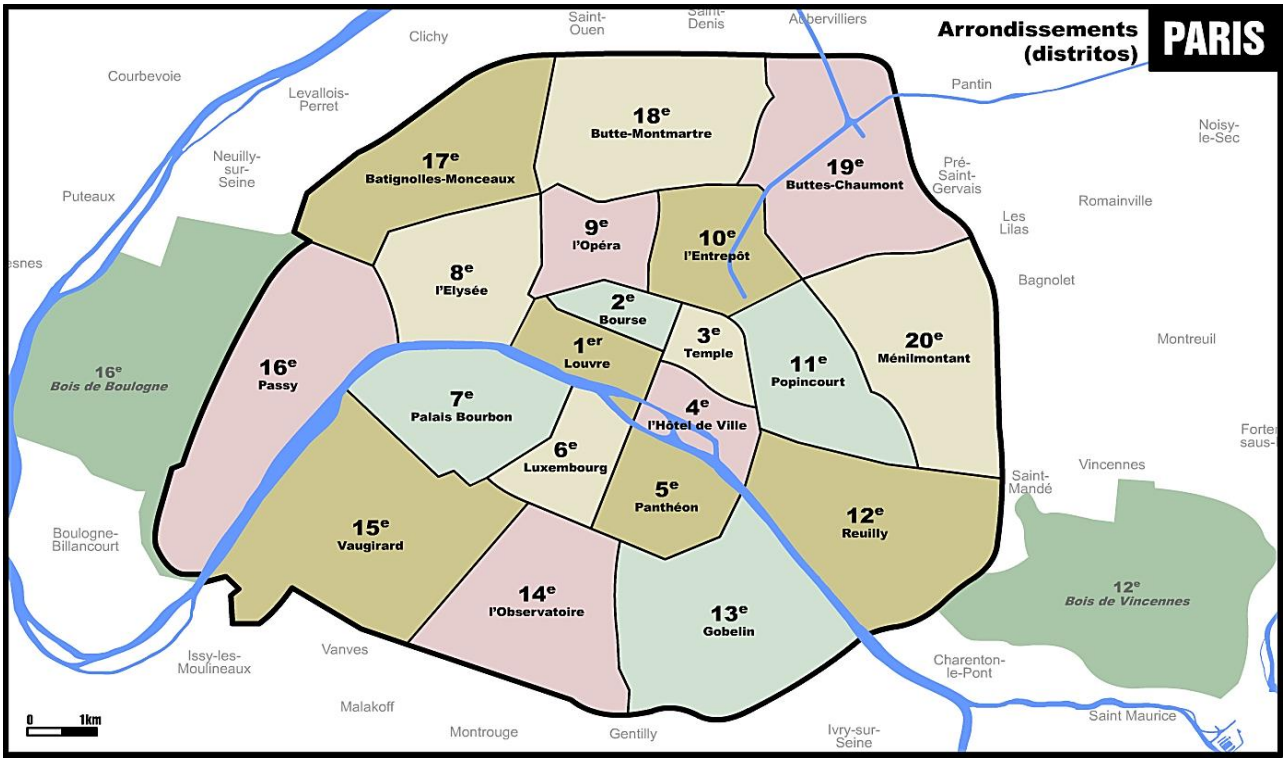
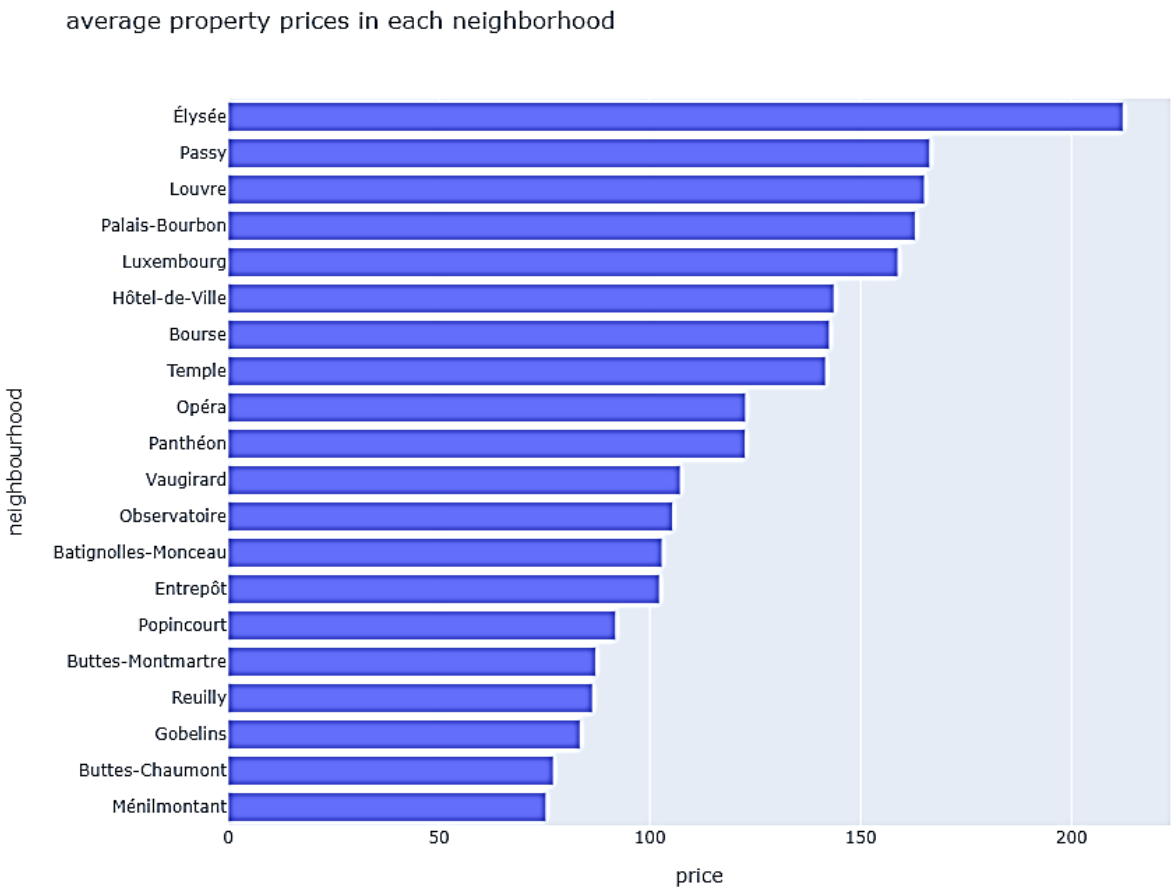


Figure 7 With this map you can make a parallel with the information from the previous one.

Once this analysis was carried out, it was important to know in which neighborhoods of Paris the most expensive and cheapest rents were located. For this, a grouping by neighborhood of the data frame dt_listings and an average of the Price characteristic has been carried out. Once this was done, the table and bar graph were generated.

neighbourhood	price
Élysée	212.259807
Passy	166.402393
Louvre	165.144018
Palais-Bourbon	162.969733
Luxembourg	158.858968
Hôtel-de-Ville	143.691846
Bourse	142.600844
Temple	141.684211
Opéra	122.732172
Panthéon	122.659877
Vaugirard	107.236721
Observatoire	105.427138
Batignolles-Monceau	102.915547
Entrepôt	102.286984
Popincourt	91.924299
Buttes-Montmartre	87.180605
Reuilly	86.422206
Gobelins	83.494817
Buttes-Chaumont	77.126339
Ménilmontant	75.394031

Figure 8 As can be seen, the neighborhood with the highest average price is Élysée and the neighborhood with the lowest price is Ménilmontant.



To take advantage of the data frame map_data that contains the coordinates of the neighborhoods of Paris, a choroplethic map was generated with the aforementioned figures. Due to the impossibility of adding the names of the neighborhoods on the choropleth map, a map of the neighborhoods of Paris with their names is added in order to make a parallel between the two.

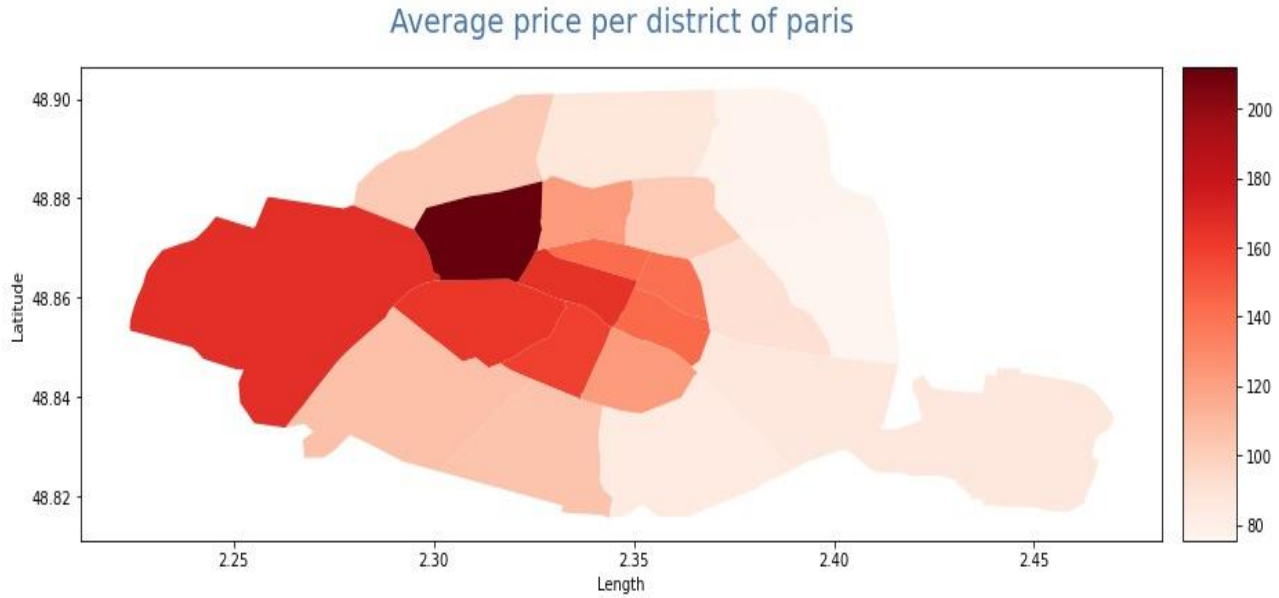


Figure 9 The darkest parts represent the neighborhoods with the highest average price.

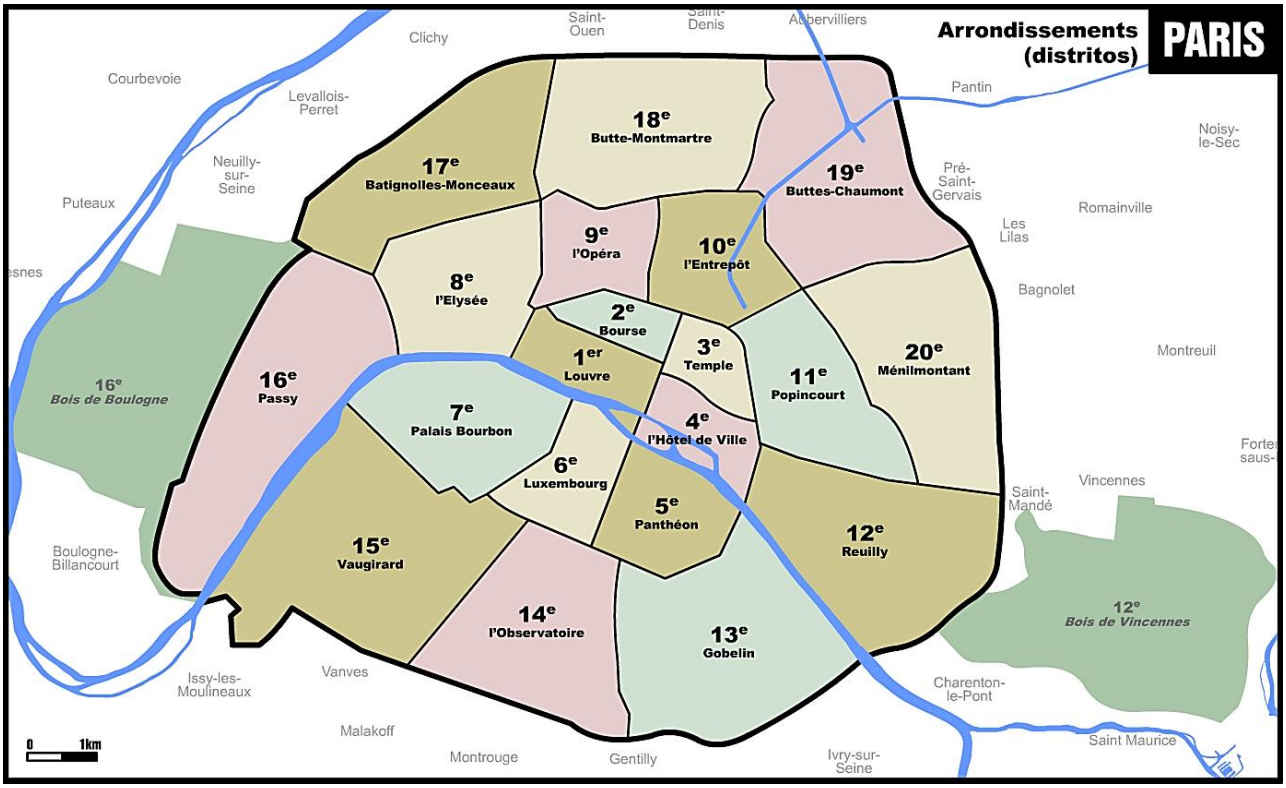


Figure 10 With this map you can make a parallel with the information from the previous one.

The number of properties by type in the city of Paris is also analyzed for this, it is grouped by type of property and the totals are extracted. Table and horizontal bar graphs of the results are made.

	total
Entire home/apt	53825
Private room	7498
Hotel room	1358
Shared room	409

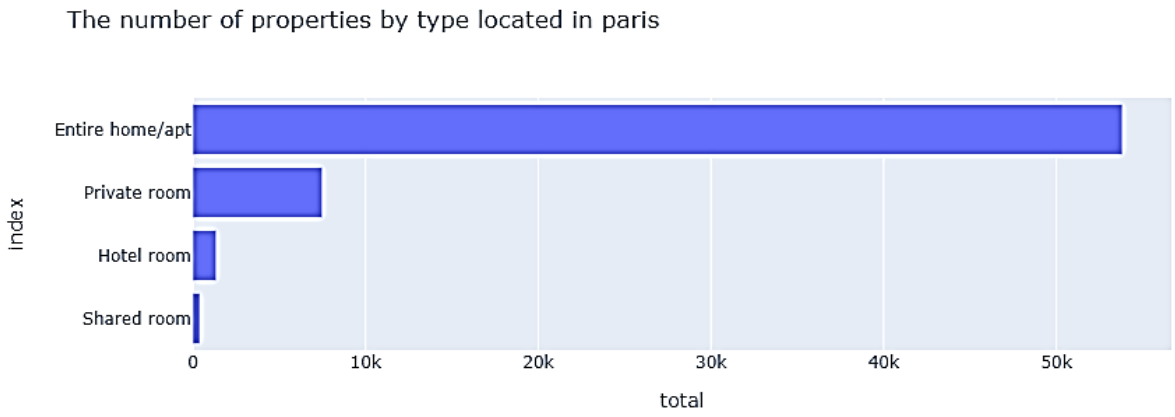


Figure 11 As you can see, the most common type of property is Entire home / apt with 53825 properties.

Once these analyzes had been carried out, the chosen properties were shown on a map, previously a filtering of the dt_listings data frame is carried out based on the client's requirements, leaving as mentioned before, 12 properties that adapt to what the interested party is looking for.

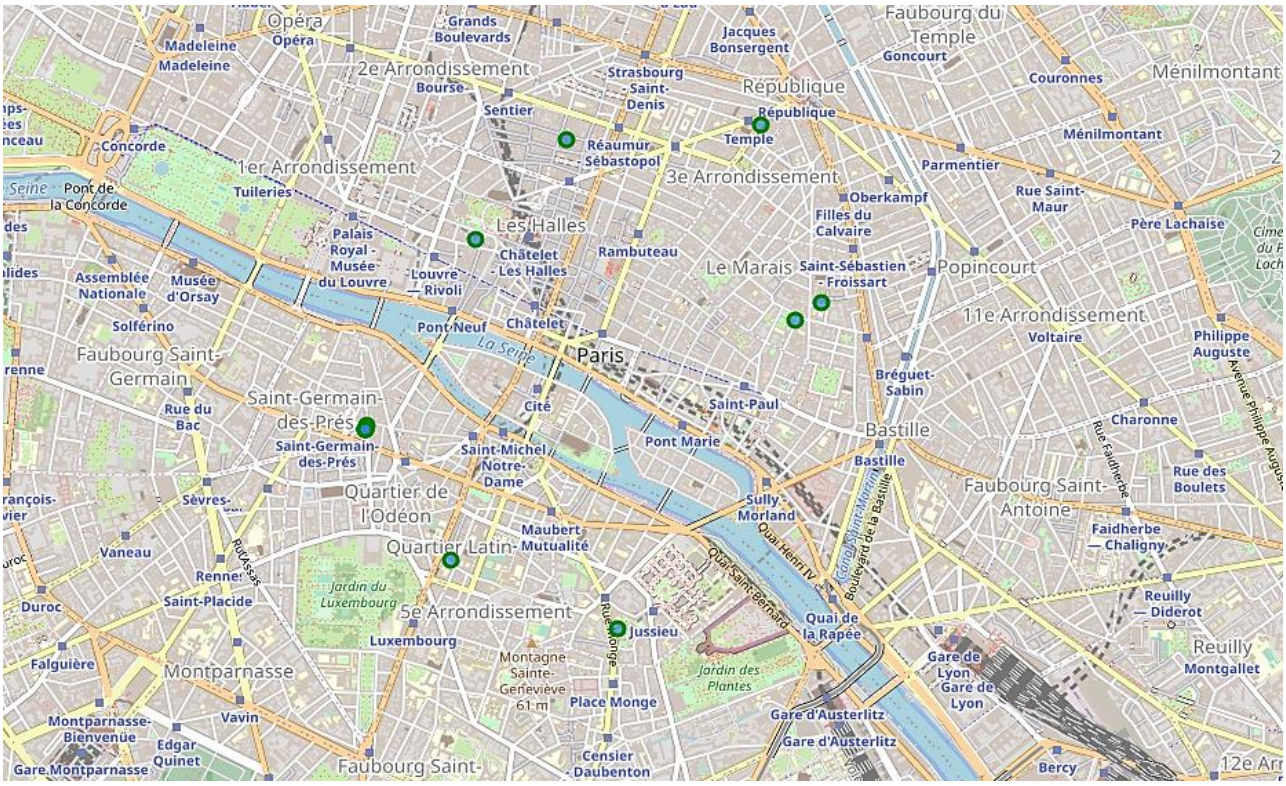


Figure 12 Represented by green circles are the properties.

Once the properties are located, the foursquare platform is used to obtain the places near each property. Once the places have been obtained, a data frame is made with the number of places near each property.

Place airbn	
Double room 2, 3 or 4 beds, Saint-Germain des Prés	44
Hotel Suite with 2 rooms for 3 people in Odeon	44
Premium double room with bed Queensize in Odeon	44
Triple beds hotel room in Saint-Germain des Pres	44
Room in the Center - Louvre, & direct to Airports!	41
Double Room with 4 beds and 2 baths in Odeon	40
Cosy Flat In the heart of paris!	36
Beautiful renovation in The Marais	27
Room and terrace in Le Marais	26
Le Home St Germain	20
French Theory - The Classmate room	19
The Quintessential Parisian Apt	19

Figure 13 Number of places near each property.

The obtained places are included in a total of 87 unique categories. An analysis of the 10 most frequent places is carried out and the 20 places closest to each property within a radius of 200 meters are listed.

```
----Beautiful renovation in The Marais----
                 venue  freq
0               Bistro  0.11
1                Café  0.07
2    French Restaurant  0.07
3         Comedy Club  0.04
4               Diner  0.04
5               Hotel  0.04
6  Japanese Restaurant  0.04
7  Peruvian Restaurant  0.04
8               Plaza  0.04
9   Italian Restaurant  0.04


----Cosy Flat In the heart of paris!----
                 venue  freq
0   Italian Restaurant  0.08
1            Wine Bar  0.08
2     Pedestrian Plaza  0.06
3    French Restaurant  0.06
4            Tea Room  0.06
5         Pizza Place  0.06
6        Cocktail Bar  0.06
7         Pastry Shop  0.03
8  Japanese Restaurant  0.03
9       Ice Cream Shop  0.03
```

Figure 14 10 most frequent places to each property.

	Place airbn	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	...	11th Most Common Venue
0	Beautiful renovation in The Marais	Bistro	Café	French Restaurant	Comedy Club	Diner	Hotel	Japanese Restaurant	Peruvian Restaurant	Plaza	...	Restaurant
1	Cosy Flat In the heart of paris!	Italian Restaurant	Wine Bar	Tea Room	French Restaurant	Pedestrian Plaza	Pizza Place	Cocktail Bar	Ice Cream Shop	Pastry Shop	...	Clothing Store
2	Double Room with 4 beds and 2 baths in Odeon	French Restaurant	Italian Restaurant	Japanese Restaurant	Seafood Restaurant	Sandwich Place	Ramen Restaurant	Café	Hotel	Ice Cream Shop	...	Greek Restaurant
3	Double room 2, 3 or 4 beds, Saint-Germain des ...	Italian Restaurant	French Restaurant	Hotel	Seafood Restaurant	Sandwich Place	Ramen Restaurant	Ice Cream Shop	Café	Japanese Restaurant	...	Greek Restaurant
4	French Theory - The Classmate room	Hotel	Indie Movie Theater	Café	Ice Cream Shop	Tea Room	Sandwich Place	Brasserie	Gourmet Shop	Creperie	...	Bar

Figure 15 The 20 closest places to each property

Finally, a one hot encoding process is carried out based on the Venue Category feature and it is grouped by means of the Place Airbnb feature. The resulting data frame is used for model training.

	Place airbn	African Restaurant	Art Gallery	Art Museum	Asian Restaurant	Bagel Shop	Bakery	Bar	Beer Bar	Belgian Restaurant	...	Sushi Restaurant	Taco Place
0	Beautiful renovation in The Marais	0.000000	0.000000	0.000000	0.000000	0.00000	0.037037	0.000000	0.037037	0.000000	...	0.000000	0.00000
1	Cosy Flat In the heart of paris!	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	...	0.027778	0.00000
2	Double Room with 4 beds and 2 baths in Odeon	0.000000	0.000000	0.000000	0.025000	0.00000	0.000000	0.025000	0.000000	0.025000	...	0.000000	0.00000
3	Double room 2, 3 or 4 beds, Saint-Germain des ...	0.000000	0.000000	0.000000	0.022727	0.00000	0.000000	0.022727	0.000000	0.022727	...	0.000000	0.00000
4	French Theory - The Classmate room	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.052632	0.000000	0.000000	...	0.000000	0.00000

Figure 16 This data frame will be used to train the model.

4. Classification Modeling

To find similarities between properties and classify them, we will use a model based on the k-means algorithm. Which is an unsupervised algorithm. The model is evaluated by the elbow method to find the most consistent value for k.

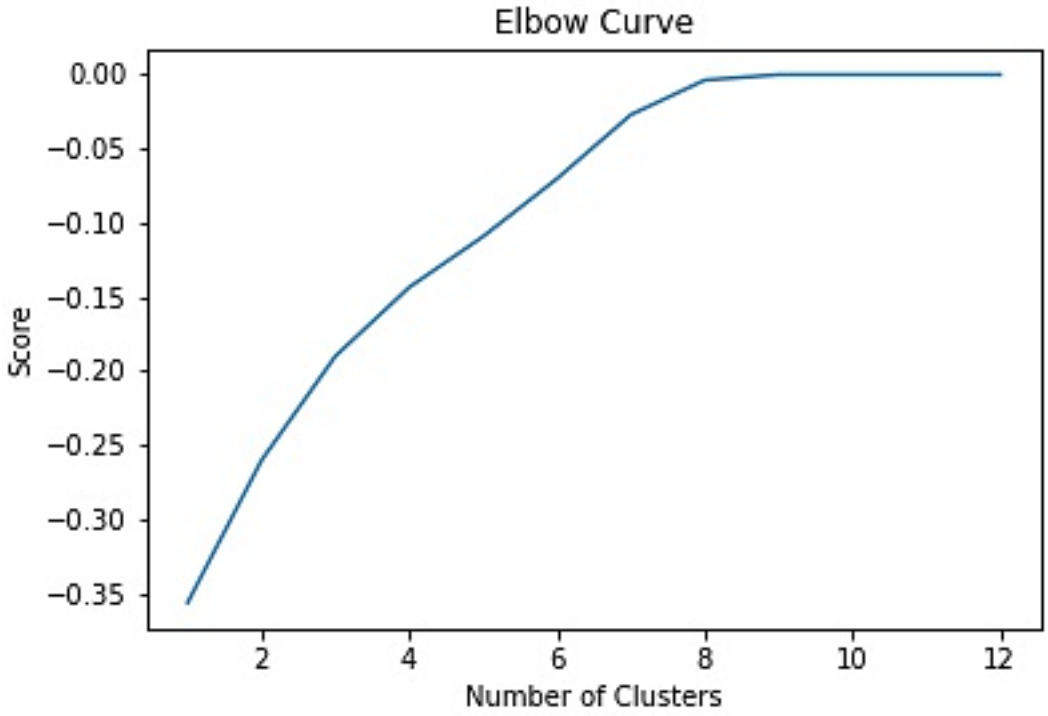


Figure 17 Graph of the elbow method

This model will be trained with a data frame that will contain several characteristics that represent the types of places that exist near the properties. The k centroids of the model will contain a value of 8 that has been selected thanks to the elbow method, therefore it will allow the properties to be classified into 8 clusters.

5. Conclusions

The model offers us the results that will be listed below. The table shows the clusters, the description, the qualification and the percentage of approval.

Nro cluster	Descripcion	Qualification	%approval
0	This cluster contains properties that are too close to each other since it is evident that the nearby places they have are very similar. Of the 20 closest places to this property, only one adapts to the requirements. Although it has the advantage that having so many properties nearby in such a short distance it would make it much easier to change to another property.	1/3	33.3
1	The property that is in this cluster is located near several cultural places which is a great advantage. But it does not meet any of the requirements.	0/3	0.0
2	The property that is in this cluster is located near several food places. One of these adapts to the requirements.	1/3	33.3
3	The property that is in this cluster is located near several cultural places. Only one of these places fits the requirements.	1/3	33.3
4	The property that is in this cluster is located near several food establishments. Only one of these places fits the requirements.	1/3	33.3
5	The property that is in this cluster is located near several food establishments and a bookstore. Two of these places adapt to the requirements.	2/3	66.6
6	The property that is in this cluster is located near several food, clothing and cultural establishments. None of these places suit the requirements.	0/3	0.0
7	The property that is in this cluster is located near several food establishments and gyms. TWO of these locations accommodate requirements.	2/3	66.6

It can be seen that two of the clusters have 2 of the 3 specific places mentioned in the requirements with an approval of 66.6%. This means that the properties located in these two clusters are the ones that have the most option to be Alejandro's Home in Paris.

The two properties are listed below:

	id	name	neighbourhood	latitude	longitude	room_type	price	number_of_reviews	last_review	availability_365
8527	6360244	Cosy Flat In the heart of paris!	Bourse	48.86572	2.34933	Private room	127	222	2021-03-07	359
15235	11229350	Beautiful renovation in The Marais	Temple	48.86637	2.36145	Private room	129	147	2020-08-20	360

The two properties have very similar characteristics, therefore either one would be a good option. In the end, only a few small differences would tip the balance in favor of the property called: "Cozy Flat In the heart of Paris!". The rent is two euros cheaper and the number of reviews is much higher.