Capstone Project -The Battle of NeighborhoodsReport

Gabin FODOP

1. INTRODUCTION

1.1 Background

Each year a considerable number of people travel either for vacation, for work or to visit friends and do not have accommodation or sleep. They are then forced to look for accommodation corresponding to their current needs. Once this situation is taken into account, it is important for us to know how these people go about finding their accommodation.

A large part of them turn to hotels or other fast accommodation. Some on the contrary use applications to rent to individuals. The most successful application in this area is AirBnB, but we are not always satisfied and the time to find accommodation is sometimes quite long. Once this problem arises, we ask ourselves how to improve the speed of housing selection? How to guarantee him to stay in a certain comfort?

1.2 Probleme

To obtain a list of all AirBnB accommodations, there are a number of datassets available for free on the internet such as insideairbnb, or even kaggle. But it is also possible to scrape AirBnB data online as in this repositories. For us our data comes from Kaggle.

Our project will therefore be based on certain specifications from our client, to provide him with a rental list that resembles the criteria he has chosen. Finally, to make a classification according to the reconciliation of the different dwellings with k-Mean Clustering and data aggregation from the FourSquare platform To extract the most suitable group for our client

1.3 Who is concern.

This code is for anyone who plans to use the AirBnB platform to rent accommodation today or in the near future. Again, it could be extended to other platforms by ingesting data from these

2 Data acquisition and cleaning

2.1 Data Acquisition

Les donne pour ce projet proviennent d'une composition de source différentes.

La première est le jeu de données insideairbnb du quel on extrait la liste des logements disponible sur AirBnB pour la période allant de 2018 à 2020. Ce jeu de données contient les colonnes suivantes.

- host_response_rate : pourcentage des réponses des utilisateurs
- host_acceptance_rate:
- host_listings_count:
- latitude: position en latitude du bien
- longitude: position en longitude du bien
- city: le dataset etant liee a paris, toutes les villes qui y seront représenté seront identique
- zipcode: adresse postale du bien
- state: Pareil que pour la date, il s'agira ici juste de IDF ou Ile-de-France
- accommodates: Le nombre de personne qui peuvent occuper l'appartement à un moment
- room_type: Le type de logement, il nous faudra faire une exploration pour trouver quels sont les types et leur impact sur le prix
- bedrooms: Nombre de chambre
- bathrooms: Nombre de salle d'eau
- beds: Nombre de lit
- price: Le prix de l'appartement
- cleaning_fee: Frais de nettoyage
- security_deposit: dépôt de garanti
- minimum_nights: nombre de nuit minimum
- maximum_nights: nombre de nuit maximum
- number_of_reviews: nombre de revue

Our second source of data is the Wikipedia page Liste des quartiers administratifs de Paris, which provides us via web scraping with a certain amount of additional information such as population, area, density. We will mainly use population information by neighborhood (municipality)

The third is the list of Parisian arrondissements in GeoJson format. It comes from the france geojson site. It will be used mainly for the visualization of our data and results.

Our last source of data is the FourSquare API which allows us to find the list of stores, restaurants, stores around the place.

2.2 Data Clearing

For the FourSquare and France GeoJson sources, there is no data preparation to do. Most of our processing will be carried out on the dataset containing the list of AirBnBs in Paris.

	host_response_rate	host_listings_count	latitude	longitude	city	zipcode	state	accommodates	room_type	bedrooms	bathrooms	beds	price	cleaning_fee	security_deposit
0	1.00	1.0	48.83349	2.31852	Paris	75014	Île-de-France	2	Entire home/apt	0.0	1.0	0.0	75.0	50.0	0.0
1	1.00	1.0	48.85100	2.35869	Paris	75004	Ile-de-France	2	Entire home/apt	0.0	1.0	1.0	115.0	36.0	0.0
2	1.00	2.0	48.85758	2.35275	Paris	75004	Île-de-France	4	Entire home/apt	2.0	1.0	2.0	115.0	50.0	200.0
3	1.00	1.0	48.86528	2.39326	Paris	75020	Ile-de-France	3	Entire home/apt	1.0	1.0	1.0	90.0	NaN	NaN
4	0.67	3.0	48.85899	2.34735	Paris	75001	Île-de-France	2	Entire home/apt	1.0	1.0	1.0	75.0	200.0	1500.0

Fig 2.1 Visualisation du dataset

The second data source is obtained by web scraping from Wikipedia using the Beautiful Soup library in python.

Les colonnes sont les suivantes :

- Arrondissement : département du quartier
- Quartiers : Nom du quartier
- Population : population du quartier
- Superficie : aire du quartier

Une fois le scaping fait, on obtient le dataset suivant :

	arrondissement	Quartiers	Population	superficie
0	1er	Saint-Germain-l'Auxerrois	1 672	86,9
1	2e	Halles	8 984	41,2
2	3e	Palais-Royal	3 195	27,4
3	4e	Place-Vendôme	3 044	26,9
4	5e	Gaillon	1 345	18,8

Fig 2.2 Wikipedia data

The third source is that of FourSquare, which gives us for a given latitude and longitude the list of shops and restaurants in the vicinity.

We have 7 columns,

- Neighborhood : numero zip de la zonne
- Neighborhood Latitude : latitude of logement
- Neighborhood Longitude : Longitude oflogement
- Venue : name of venue
- Venue Latitude : latitude of venue

- Venue Longitude : Longitude of venue
- Venue Category : Type of venue

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	7992	48.87822	2.35769	Le Delly's	48.878458	2.357852	African Restaurant
1	7992	48.87822	2.35769	Marks & Spencer Food	48.876742	2.358486	Food & Drink Shop
2	7992	48.87822	2.35769	Caves Bardou	48.876635	2.356028	Wine Shop
3	7992	48.87822	2.35769	Marché Saint-Quentin	48.876831	2.355234	Farmers Market
4	7992	48.87822	2.35769	Extérieur Quai - Le Bouillon de l'Est	48.876456	2.357905	Bistro

Fig 2.3 FourSquare values

3 Methodologie:

3.1 Exploration, Data analysis

3.3.1 analyse statistique des Logements.

The describe function is used to quickly obtain the numeric characteristics of the numeric columns of the dataset. We quickly see that the host_acceptance_rate column is completely empty. Finally we also have the number of different values of nan in the other columns.

	host_acceptance_rate	host_listings_count	latitude	longitude	accommodates	bedrooms	bathrooms	beds	minimum_nights	maximum_nights	number_of_reviews
count	0.0	7999.000000	8000.000000	8000.000000	8000.000000	7976.000000	7942.000000	7986.000000	8000.000000	8000.000000	8000.000000
mean	NaN	7.025878	48.864560	2.348739	3.198750	1.248370	1.128494	1.753068	8.759375	546.876000	44.874875
std	NaN	51.031588	0.017641	0.031611	1.569811	0.838492	0.439104	1.172800	36.234546	542.848736	69.075322
min	NaN	0.000000	48.816560	2.230810	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000
25%	NaN	1.000000	48.852230	2.331645	2.000000	1.000000	1.000000	1.000000	2.000000	30.000000	4.000000
50%	NaN	1.000000	48.865400	2.351130	3.000000	1.000000	1.000000	1.000000	3.000000	365.000000	19.000000
75%	NaN	2.000000	48.878740	2.372158	4.000000	2.000000	1.000000	2.000000	5.000000	1125.000000	54.000000
max	NaN	836.000000	48.901010	2.439460	16.000000	7.000000	7.000000	11.000000	1124.000000	10000.000000	783.000000

3.1 Dataset Description

Finally we have fairly large standard deviation values for the columns host_listings_count, minimum_nights, maximum_nights, number_of_reviews.

This is normal, because these are values that vary in ways quite predictable depending on the owner of the home.

3.1.2 Data segmentation

For the rest, the dataset was separated into three for each accommodation possibility.

A first idea for us is to check the outliers, to make sure that a large part of the homes of the same type have close prices.

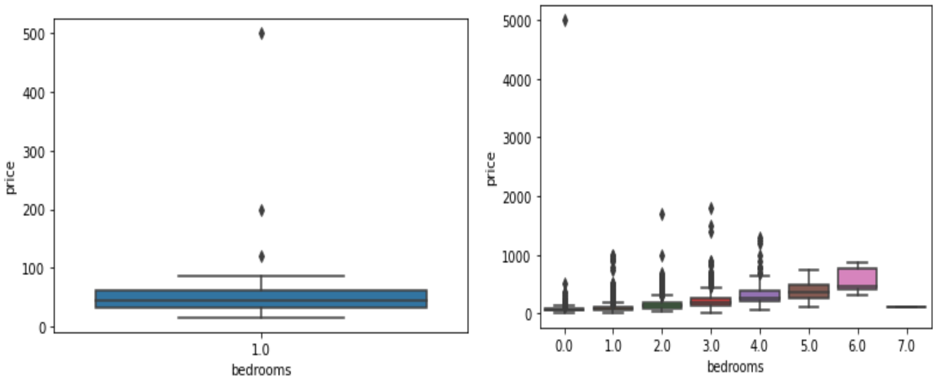

Fig 3.2 Box plot

Display in box plot of the values of two datasets, the first on the left that of the single rooms, on the right that of the entire houses.

3.1.3 Visualization

Visualization on the map of the city of Paris:

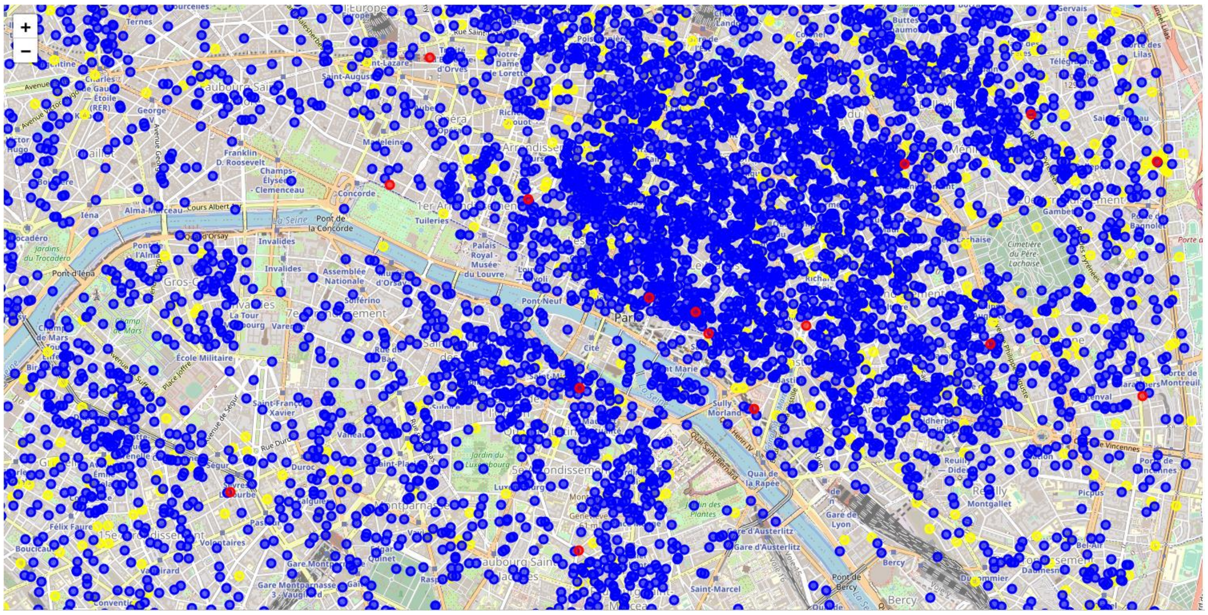

Fig 3.3 En bleu les chambre seules en jaune les Chambres partagées, en rouge les maisons

3.2 Data fusion

Our second mission is to select the values that correspond to the needs of our client. For this we have simulated values. The result of merging our dataset is:

	host_response_rate	host_listings_count	latitude	longitude	city	zipcode	state	accommodates	room_type	bedrooms	security_deposit	minimum_nights	maximum_nights	number_of_reviews	count
790	1.0	1.0	48.86939	2.37471	Paris	75011	Île-de-France	4	Entire home/apt	2.0	200.0	3	14	41	1.0
1046	1.0	1.0	48.87254	2.38944	Paris	75020	Île-de-France	4	Entire home/apt	2.0	0.0	3	365	15	1.0
1758	1.0	1.0	48.83401	2.37115	Paris	75013	Île-de-France	3	Entire home/apt	1.0	0.0	2	1125	18	1.0
4451	1.0	1.0	48.86705	2.34558	Paris	75002	Île-de-France	4	Entire home/apt	2.0	150.0	3	1125	8	1.0
5421	1.0	1.0	48.88185	2.36795	Paris	75010	Île-de-France	4	Entire home/apt	1.0	100.0	1	7	4	1.0

At the end we re-visualize on the map. And we bring them together in clusters with Kmean Clustering.

We therefore obtain 3 housing groups which correspond to the client's needs and which are alike. He is offered 1 from each group and is estimated to like the others in the same group. The final result is in Final.csv.

5 Result

The final values are contained in the result_filtered dataset which gives the following result

	host_response_rate	host_listings_count	latitude	longitude	city	zipcode	state	accommodates	room_type	bedrooms	security_deposit	minimum_nights	maximum_nights	number_of_reviews	count
790	1.0	1.0	48.86939	2.37471	Paris	75011	Île-de-France	4	Entire home/apt	2.0	200.0	3	14	41	1.0
1046	1.0	1.0	48.87254	2.38944	Paris	75020	Île-de-France	4	Entire home/apt	2.0	0.0	3	365	15	1.0
1758	1.0	1.0	48.83401	2.37115	Paris	75013	Île-de-France	3	Entire home/apt	1.0	0.0	2	1125	18	1.0
4451	1.0	1.0	48.86705	2.34558	Paris	75002	Île-de-France	4	Entire home/apt	2.0	150.0	3	1125	8	1.0
5421	1.0	1.0	48.88185	2.36795	Paris	75010	Île-de-France	4	Entire home/apt	1.0	100.0	1	7	4	1.0
6764	1.0	1.0	48.87043	2.35880	Paris	75010	Île-de-France	4	Entire home/apt	1.0	0.0	8	1125	0	1.0
7992	1.0	1.0	48.87822	2.35769	Paris	75010	Île-de-France	2	Entire home/apt	1.0	200.0	4	60	45	2.0

6. Discussion and Conclusion

The idea of this project is to help people choose their ideal AirBnB accommodation among all those available in Paris. Here we are targeting those who do not have the time to make the choice and filter the values for themselves.

Our approach has the drawback of not introducing a certain amount of information such as criminality? The consumption of drugs or the proximity to certain historic buildings in Paris. The future of the project is therefore to add these properties to our predictions to obtain more efficient models