Airbnb Pricing Prediction

Amine OMRI

Résumé

Ce projet vise à traiter des données collectées sur le site Aribnb pour la date du 14 décembre 2020 qui comporte toutes les données relatives aux listings a cette date.
Ce projet comporte trois parties principales :

Anayse 01

le premier est une analyse de données qui vise à effectuer une analyse exploratoire des données et à réaliser des visualisations, ce qui inclut le traitement des données nécessaires.

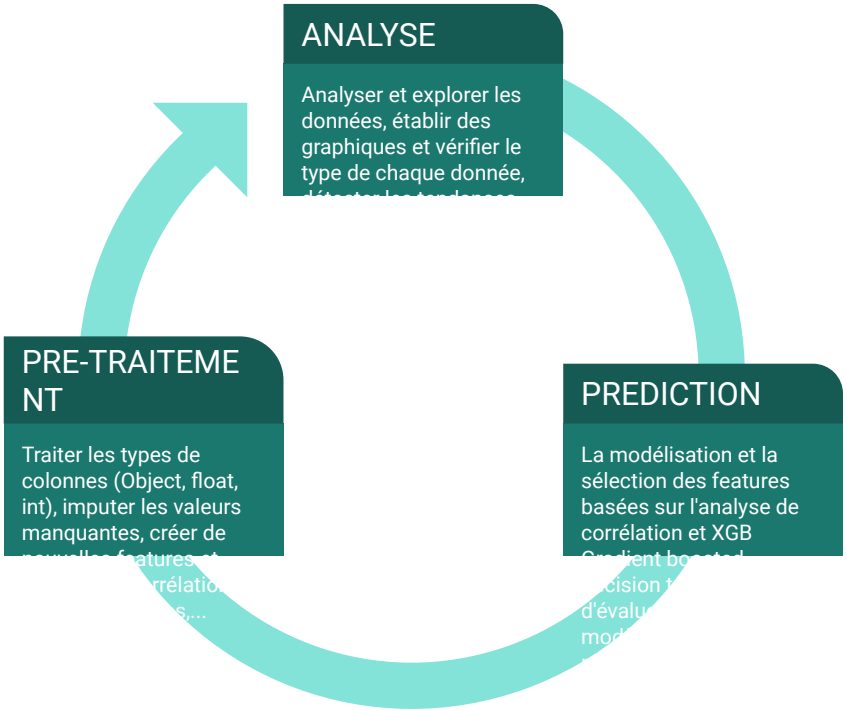
Pré-traitement 02

La seconde consiste à effectuer un prétraitement des données (nettoyage, encodage des données, Features engineering).

Prédiction 03

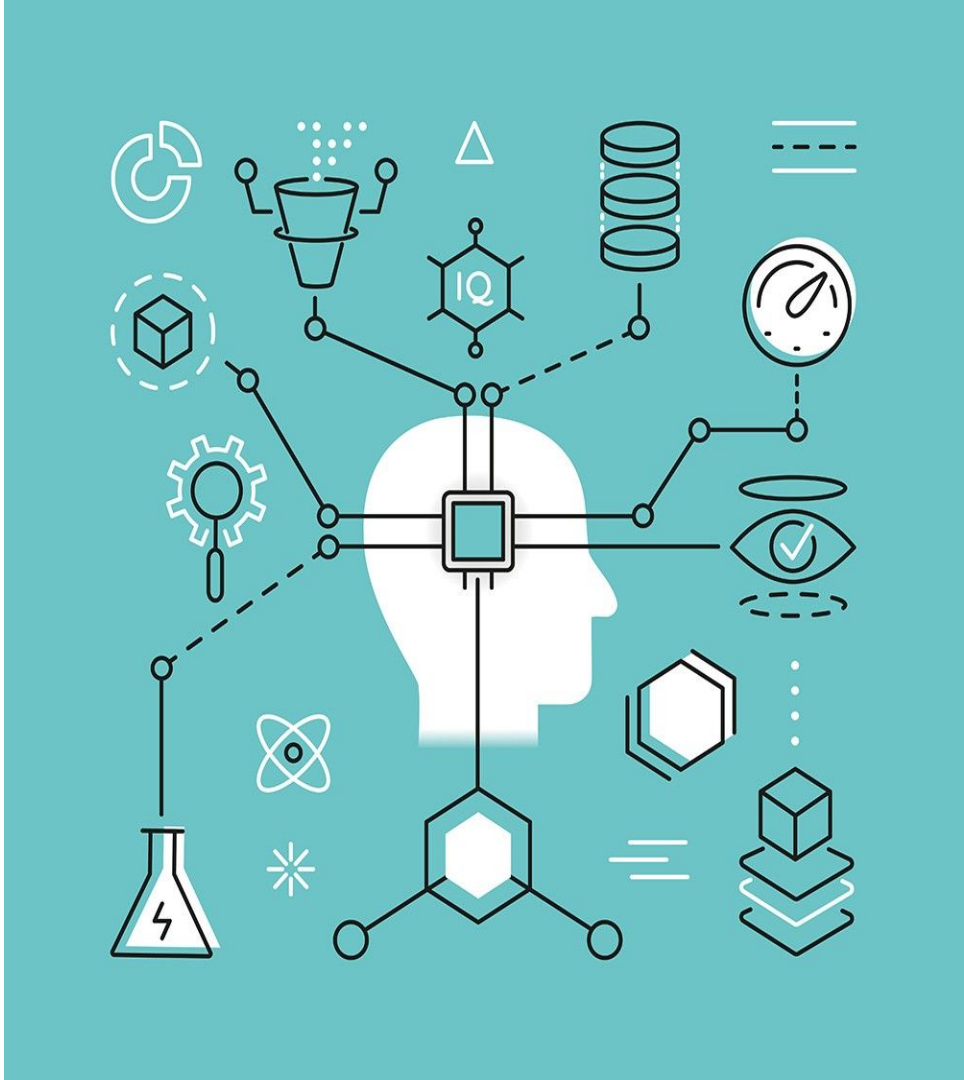
Le dernier est la modélisation et la prédiction qui comprend également certaines étapes de traitement telles que le traitement des valeurs aberrantes et la sélection des features, ainsi que des visualisations.

Cycle du projet



La partie Pré-traitement de données

Nettoyage et encodage et feature engineering



Pré-traitement

1. Nettoyage du dataset listings
2. Traiter les valeurs manquantes
3. One-hot encode les variables catégorielles
4. Feature Engineering:
   - **last_review**: cette colonnne servira à filtrer les listes qui ne sont plus active
   - **host_location**: nous pouvons l'utiliser pour déterminer si l'hôte est local ou non
   - **host_since**: peut être utilisé pour calculer l'expérience des hôtes en fonction du nombres d'années depuis leur première inscription
   - **amenities:** créer des features à partir de cette colonne en Prétraiter la colonne amenities pour extraire tous les valeur amenities possibles et affecter un identifiant à chacun d'entre eux, après avoir défini une fonction d'encodage qui mettra 1 dans l'index correspondant dans une matrice. Et enfin, on va avoir la matrice document - terme en appliquant cette fonction d'encodage à tous les documents du corpus.

Pré-traitement

5 Del meme facon (bag-of-words binaire) créer des features à partir de **host_verifications**

6 Créer des Feature à partir de colonnes de texte **description**

```
def nlp_pipeline(book_texts):
    clean_books = []
    for book in book_texts:
        book = remove_hypens(book)
        book_i = tokenize_text(book)
        book_i = remove_characters_after_tokenization(book_i)
        book_i = convert_to_lowercase(book_i)
        book_i = remove_stopwords(book_i, custom_stopwords)
        book_i = get_lemma(book_i)
        book_i = remove_short_tokens(book_i)
        book_i = keep_only_words_in_wordnet(book_i)
        book_i = apply_lemmatize(book_i)
        clean_books.append(book_i)
    return clean_books
```

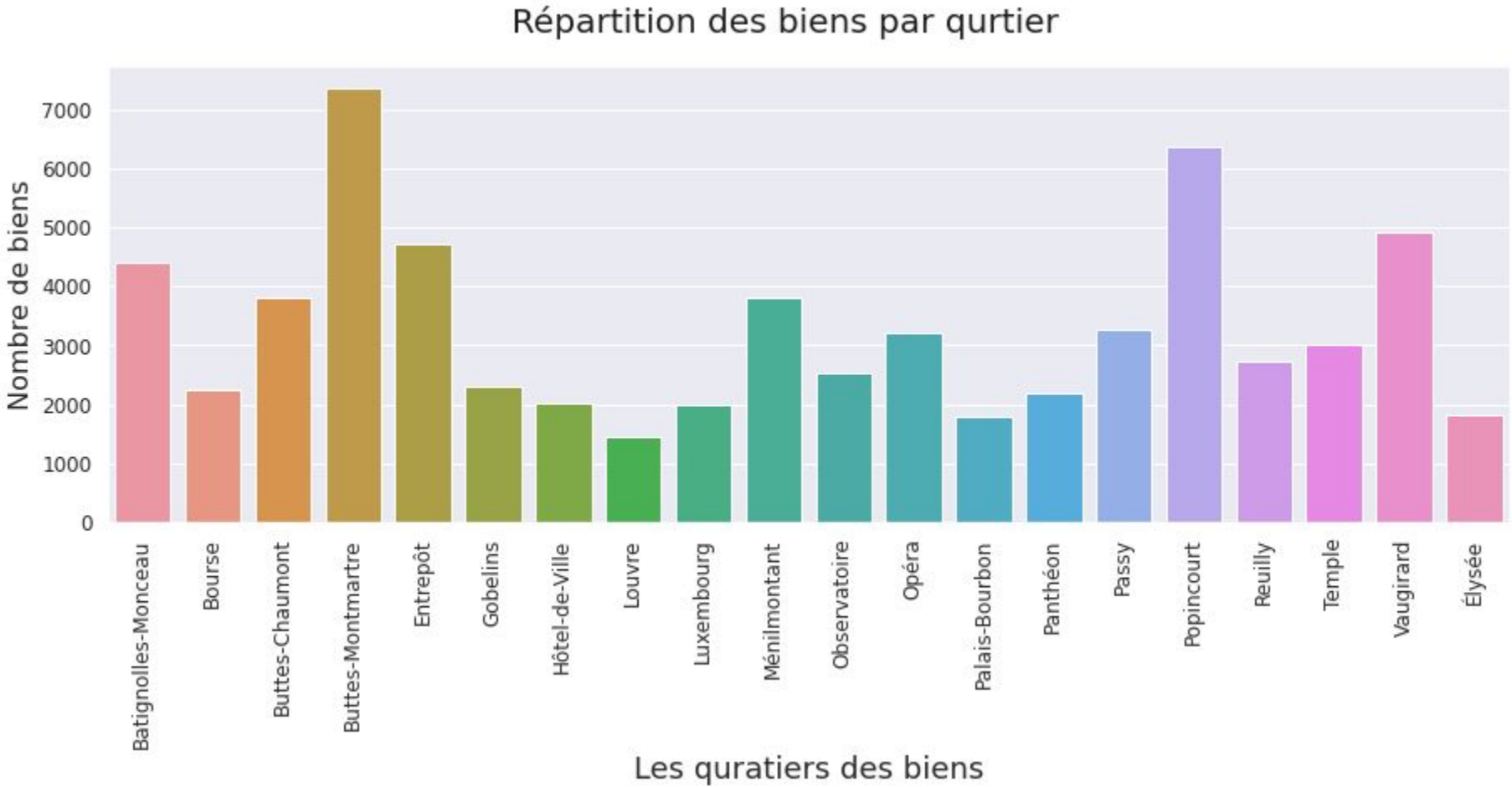
Apré avoir passer les documents par le pipline NLP il faut vectoriser le corpus maintenant que nous avons le corpus nettoyé, nous pouvons utiliser **TfidfVectorizer** pour convertir le texte en format vectoriel.

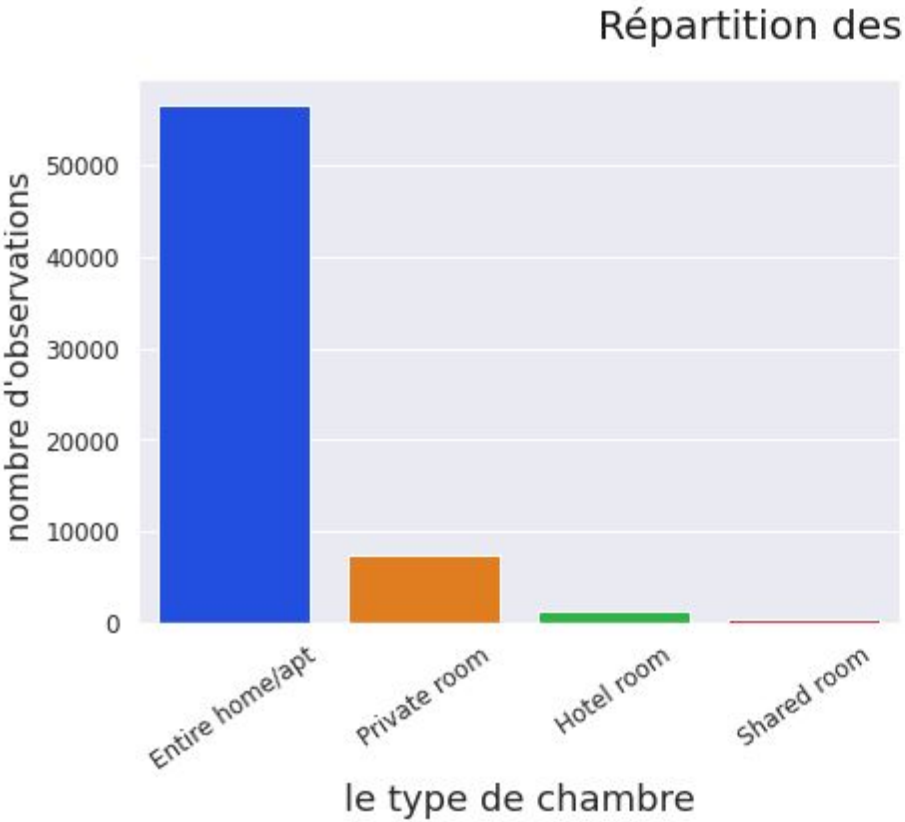
	abbess	able	absence	absolute	absolutely
0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0

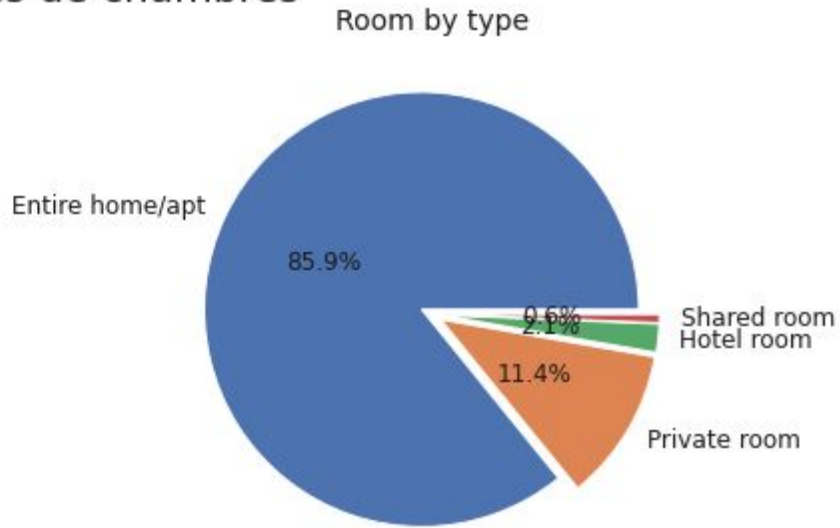
La partie Analyse de données

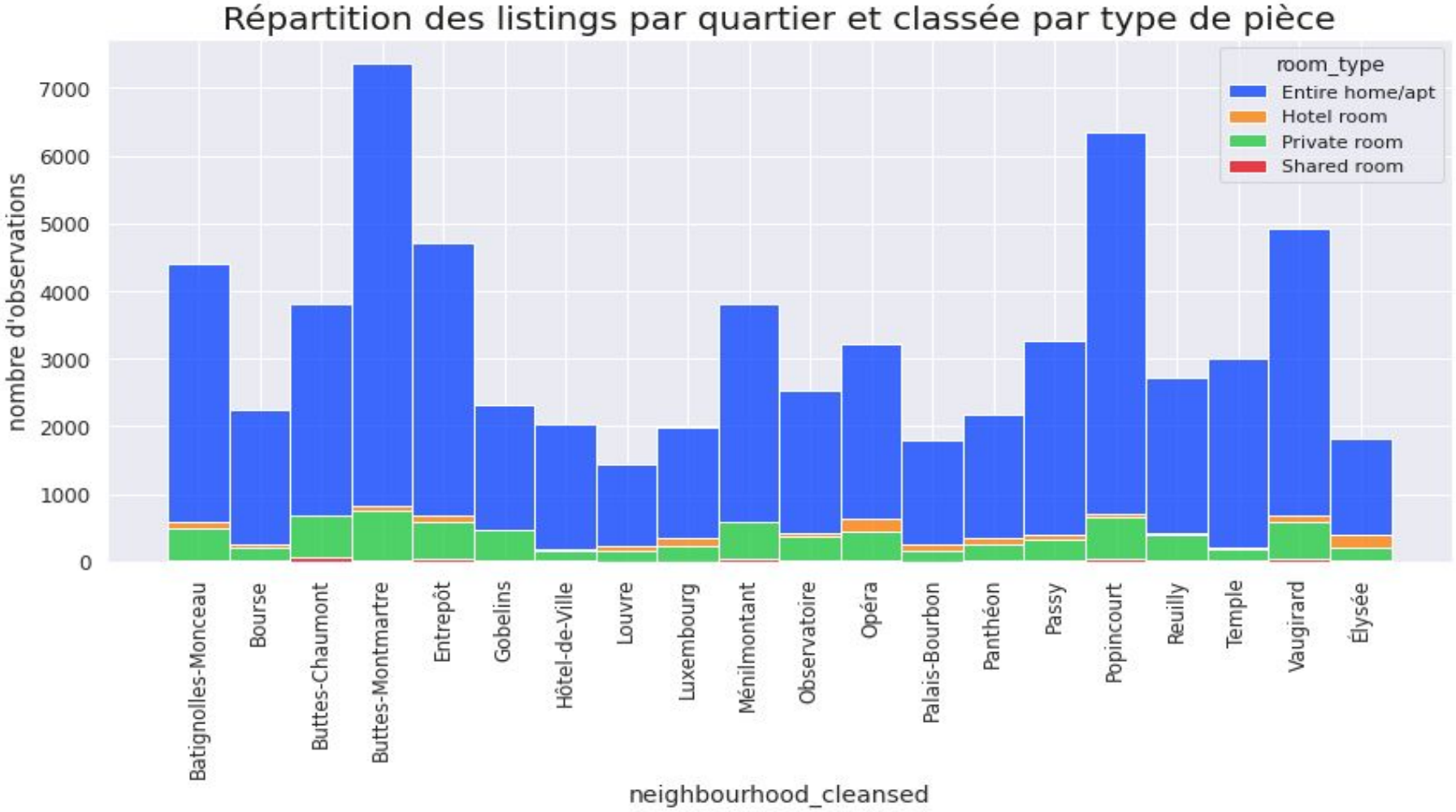
Analyser les données afin de répondre aux questions et d'obtenir des informations sur le prix des listings.

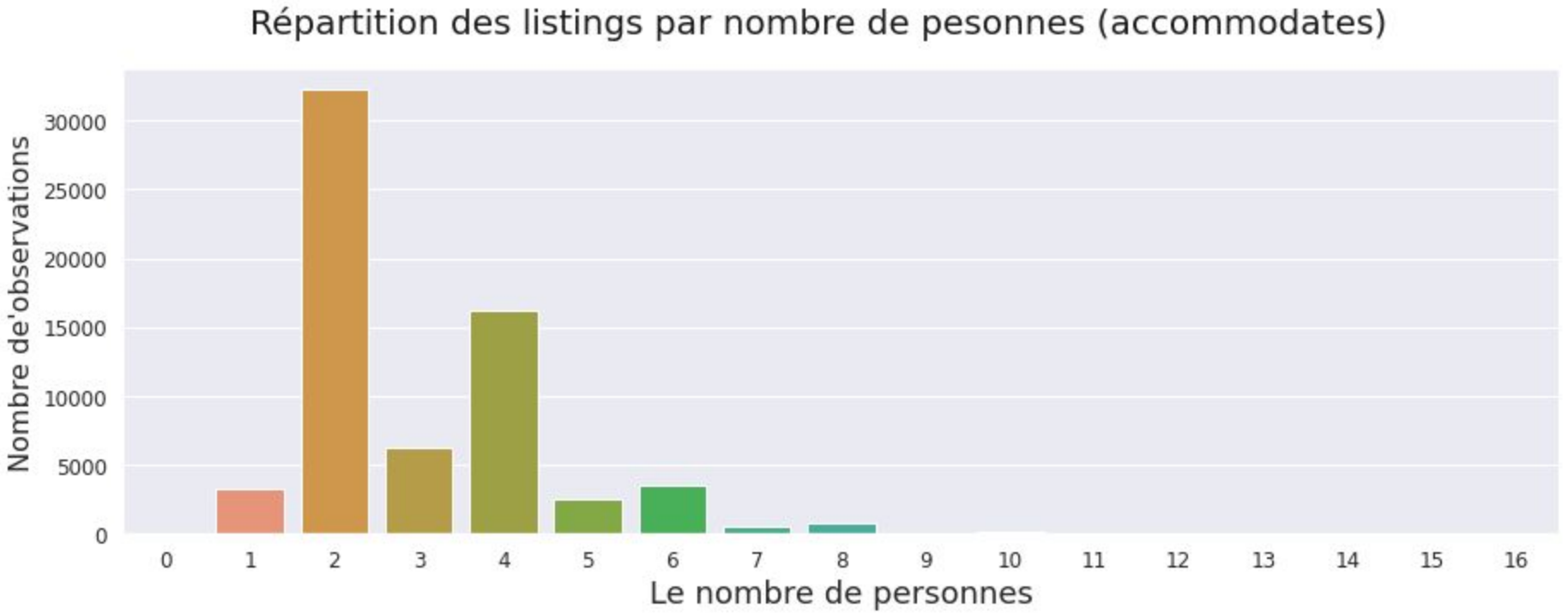

Répartition des listings par property_type
Nombre d'observation
0
10000
20000
30000
40000
50000
Entire apartment
Private room in apartment
Entire condominium
Room in boutique hotel
Entire loft
Room in hotel
Entire house
Private room in condominium
Shared room in apartment
Entire serviced apartment
Entire townhouse
Private room in bed and breakfast
Private room in house
Private room in loft
Room in aparthotel
Private room in townhouse
Room in serviced apartment
Shared room in hostel
Private room in guesthouse
Entire place
Entire guesthouse
Private room in hostel
Room in hostel
Private room
Tiny house
Entire guest suite
Shared room in condominium
Shared room in bed and breakfast
Room in bed and breakfast
Private room in guest suite
Private room in serviced apartment
Entire villa
Boat
Shared room in loft
Entire home/apt
Earth house
Houseboat
Barn
Private room in boat
Dome house
Shared room in tiny house
Entire bungalow
Private room in nature lodge
Entire cottage
Shared room in guesthouse
Private room in earth house
Shared room in townhouse
Shared room in house
Private room in casa particular
Island
Private room in chalet
Private room in houseboat
Private room in villa
Shared room in boutique hotel
Shared room in cabin
Shared room in guest suite
Casa particular
Private room in cabin
Shared room in igloo
Entire floor
Entire bed and breakfast
Shared room in serviced apartment
Cave
Campsite
Camper/RV
Entire chalet
Treehouse
Le type du biens

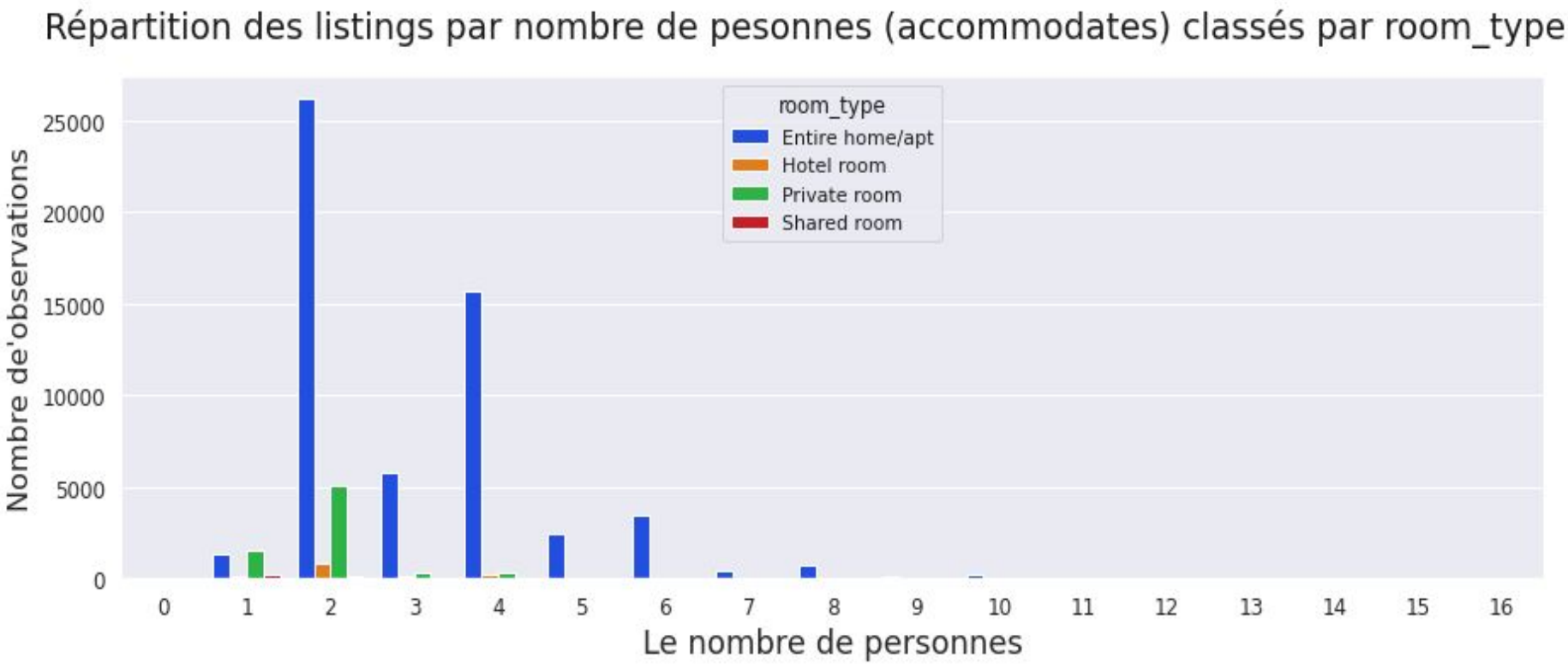

Répartition des biens par qurtier
Nombre de biens
0
1000
2000
3000
4000
5000
6000
7000
Batignolles-Monceau
Bourse
Buttes-Chaumont
Buttes-Montmartre
Entrepôt
Gobelins
Hôtel-de-Ville
Louvre
Luxembourg
Ménilmontant
Observatoire
Opéra
Palais-Bourbon
Panthéon
Passy
Popincourt
Reuilly
Temple
Vaugirard
Élysée
Les quratiers des biens


Répartition des types de chambres
nombre d'observations
50000
40000
30000
20000
10000
0
Entire home/apt
Private room
Hotel room
Shared room
le type de chambre


Room by type
Entire home/apt
85.9%
0.6%
2.1%
11.4%
Shared room
Hotel room
Private room

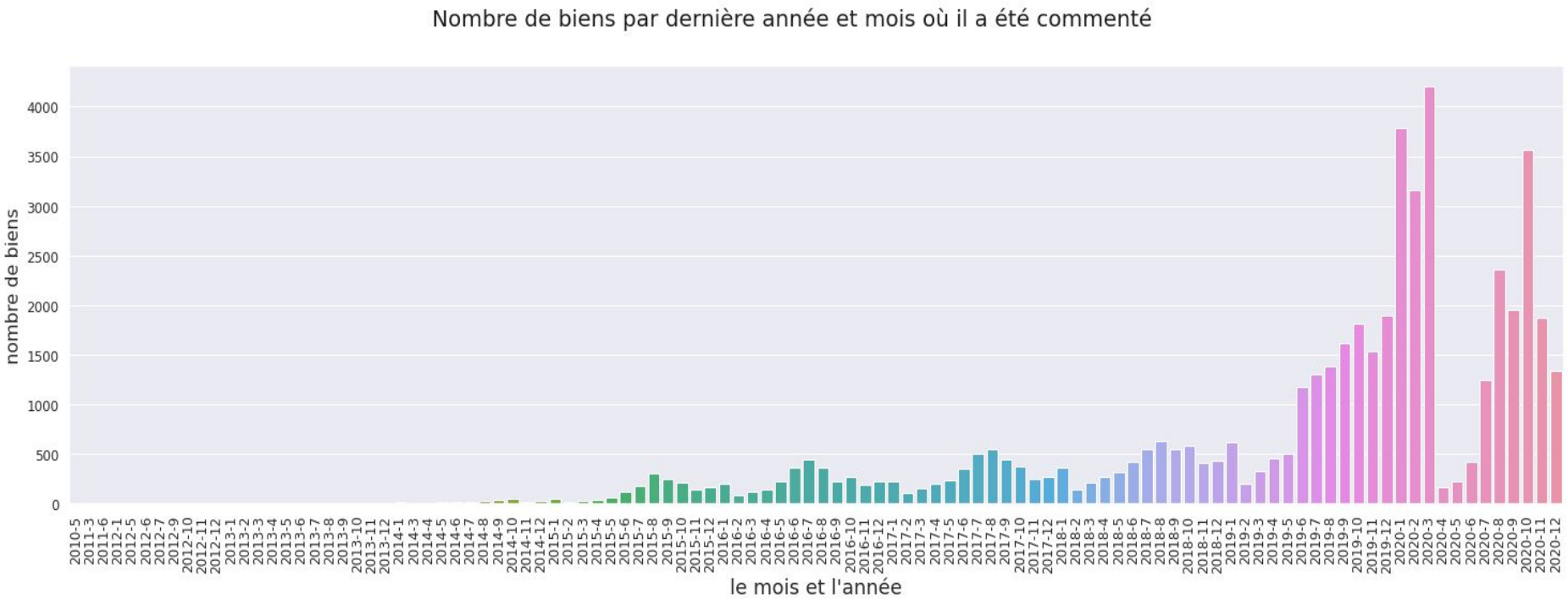

Répartition des listings par quartier et classée par type de pièce
room_type
Entire home/apt
Hotel room
Private room
Shared room
nombre d'observations
0
1000
2000
3000
4000
5000
6000
7000
Batignolles-Monceau
Bourse
Buttes-Chaumont
Buttes-Montmartre
Entrepôt
Gobelins
Hôtel-de-Ville
Louvre
Luxembourg
Ménilmontant
Observatoire
Opéra
Palais-Bourbon
Panthéon
Passy
Popincourt
Reuilly
Temple
Vaugirard
Élysée
neighbourhood_cleansed


Répartition des listings par nombre de pesonnes (accommodates)
Nombre de'observations
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Le nombre de personnes

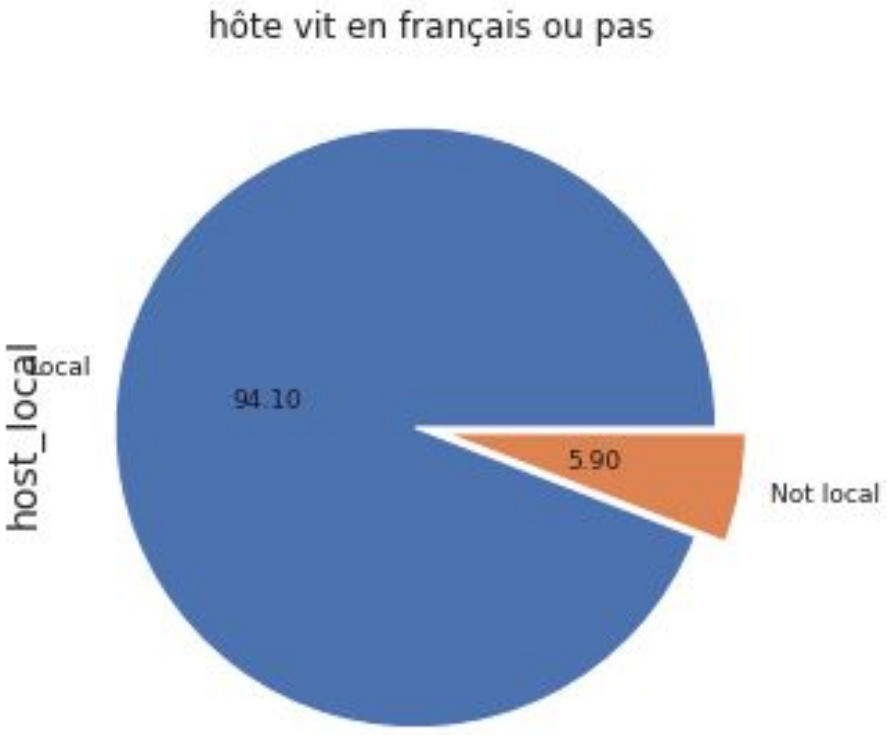

Répartition des listings par nombre de pesonnes (accommodates) classés par room_type
room_type
Entire home/apt
Hotel room
Private room
Shared room
Nombre de'observations
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Le nombre de personnes

last_review: ce champ servira à filtrer les listes qui ne sont plus actives

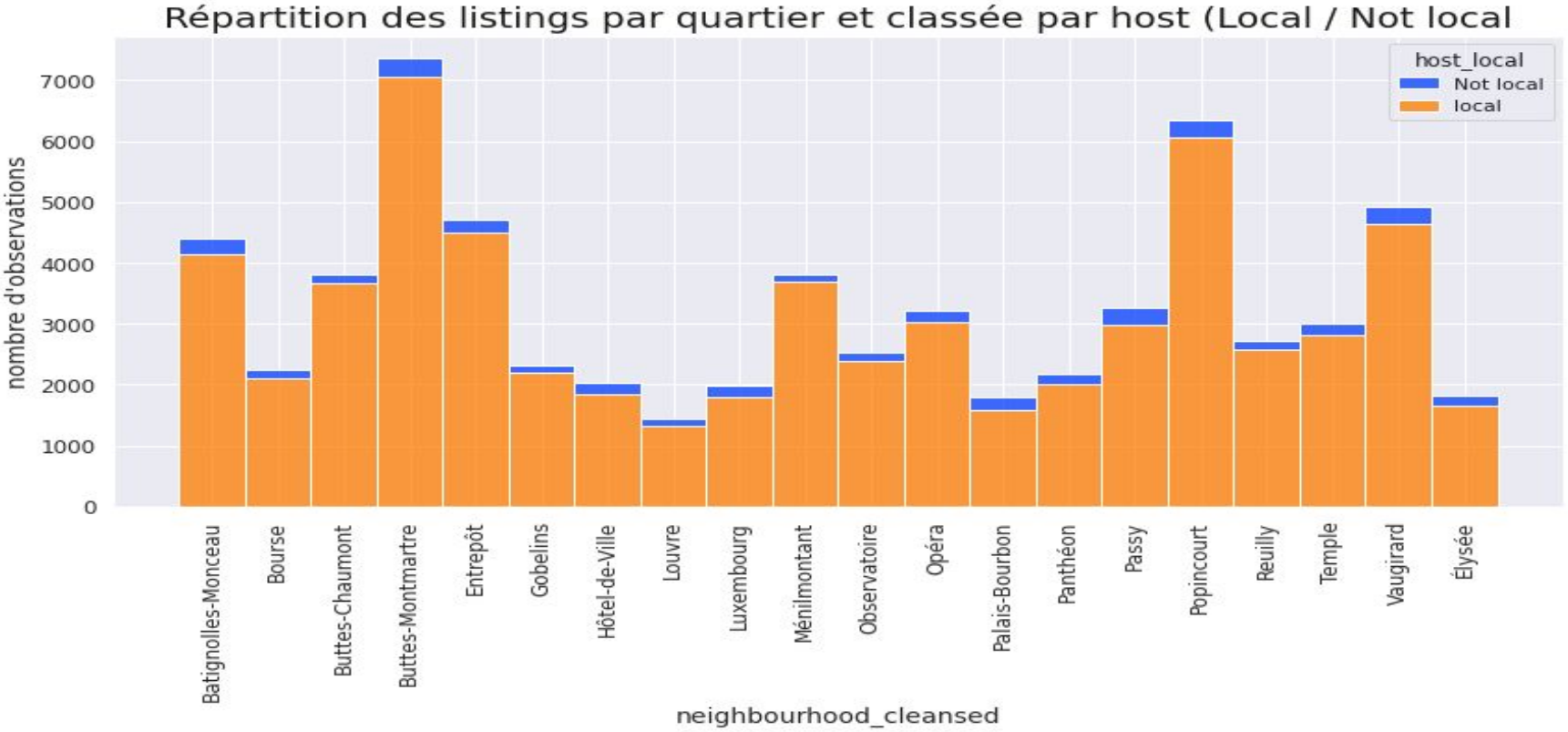
last_review: ce champ servira à filtrer les listes qui ne sont plus actives



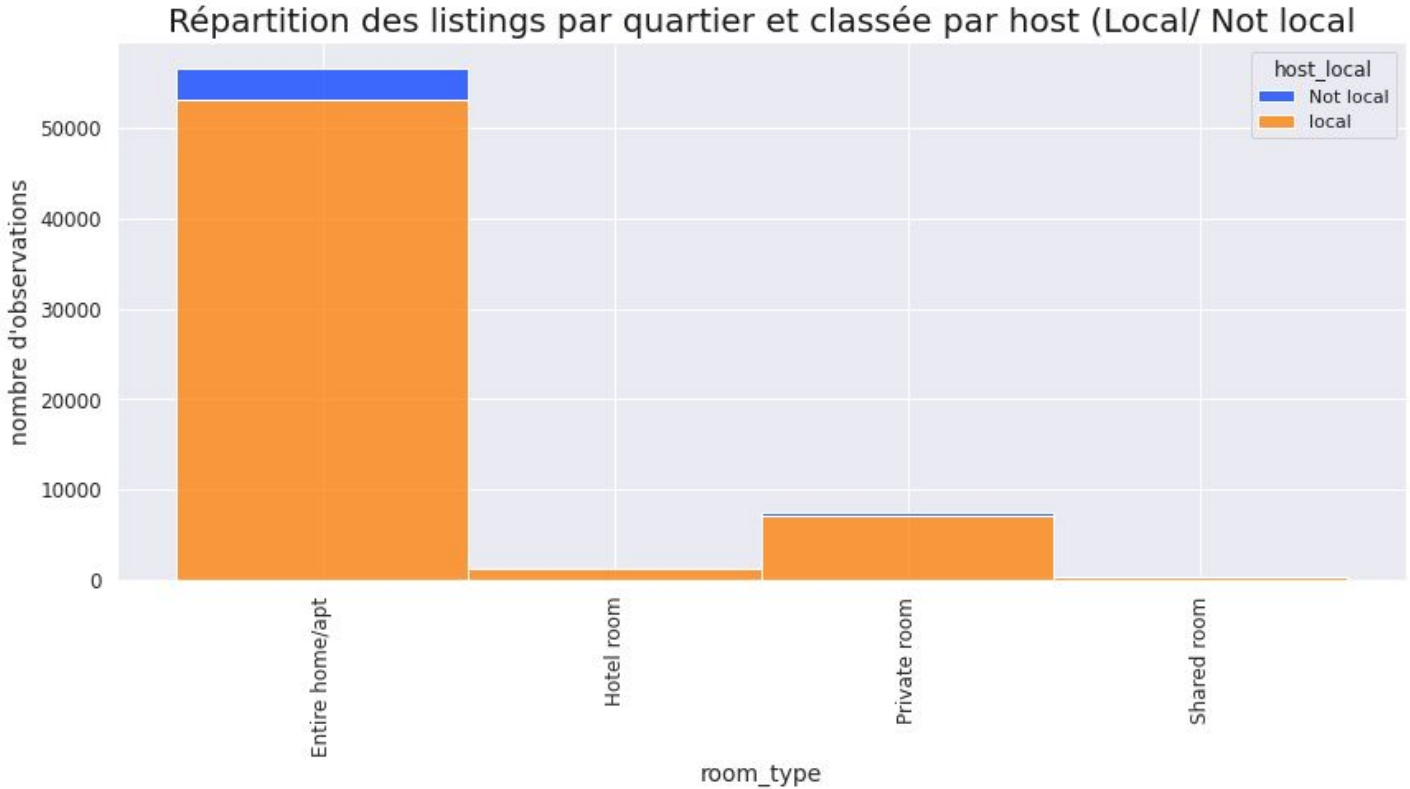
host_location: nous pouvons l'utiliser pour déterminer si l'hôte est local ou non



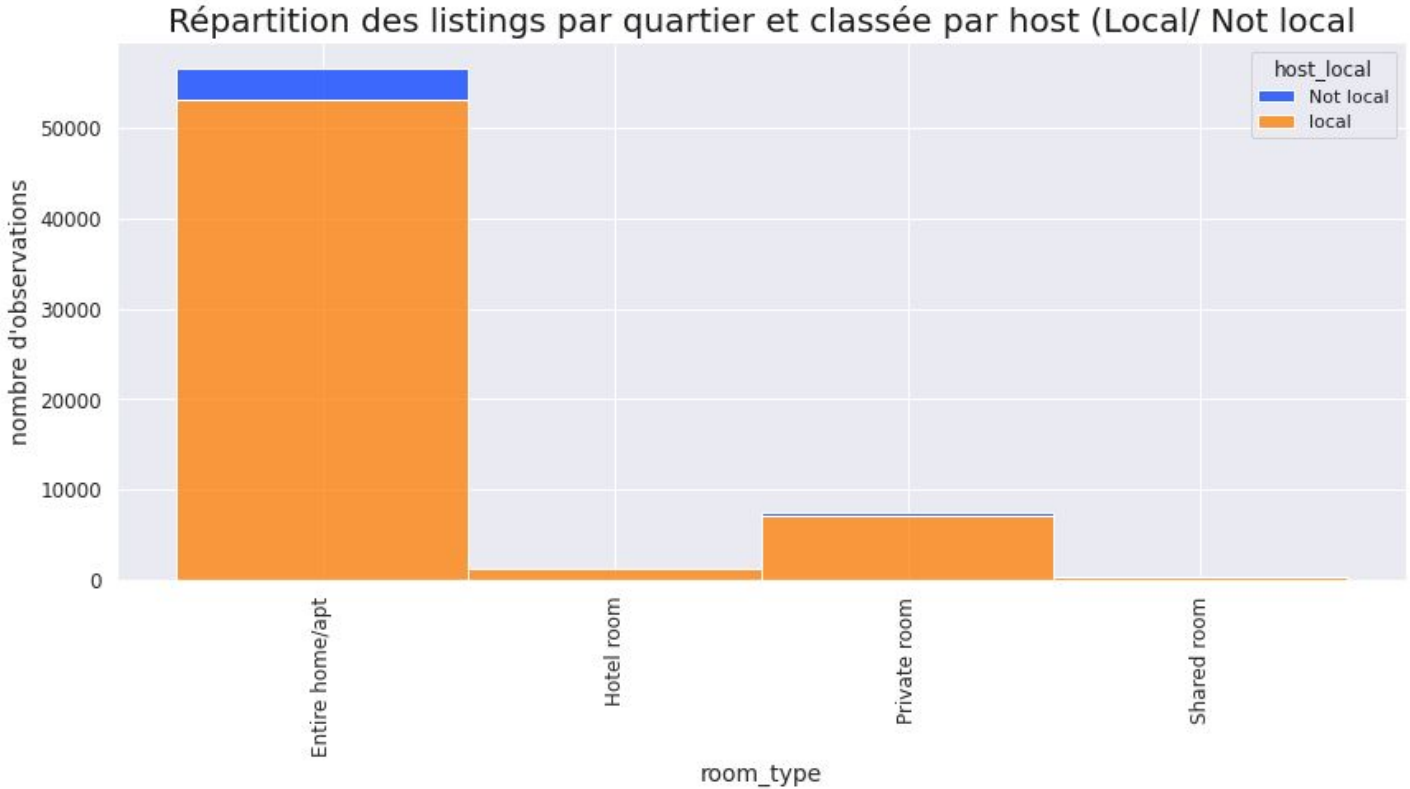
host_location: nous pouvons l'utiliser pour déterminer si l'hôte est local ou non



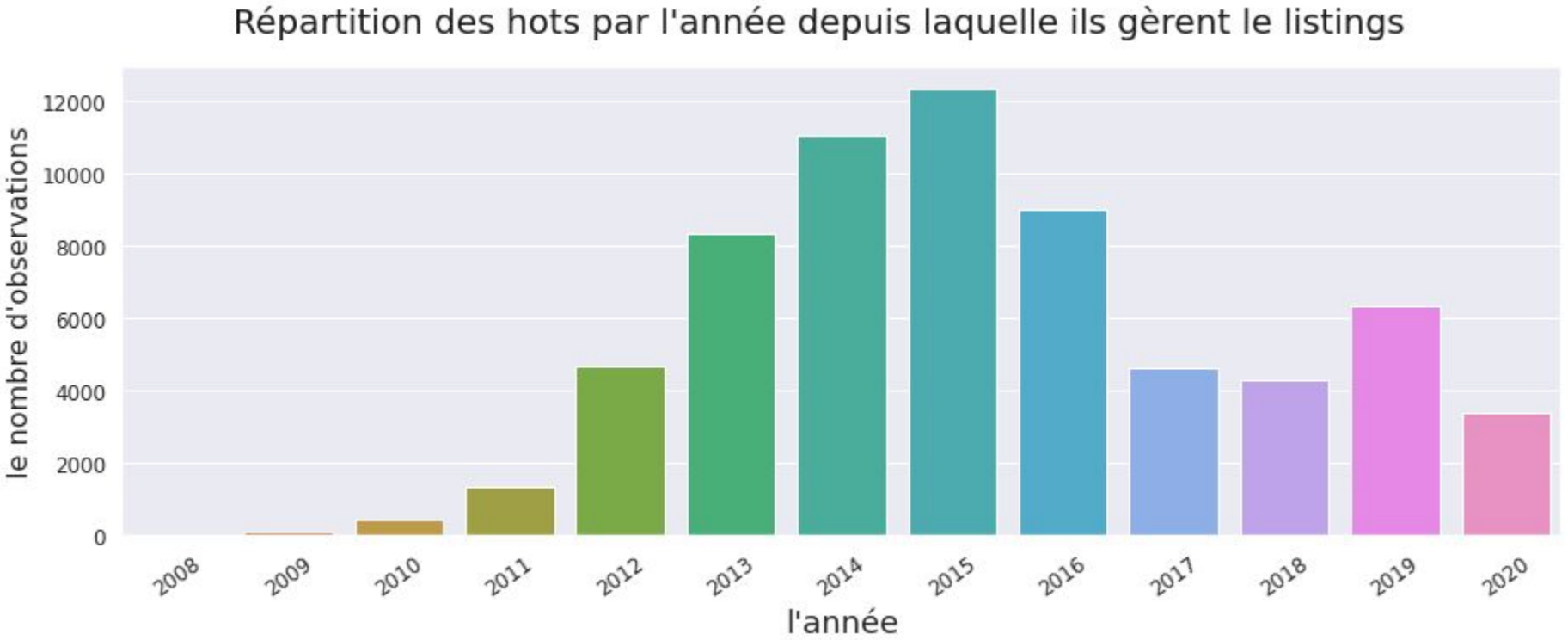
host_location: nous pouvons l'utiliser pour déterminer si l'hôte est local ou non



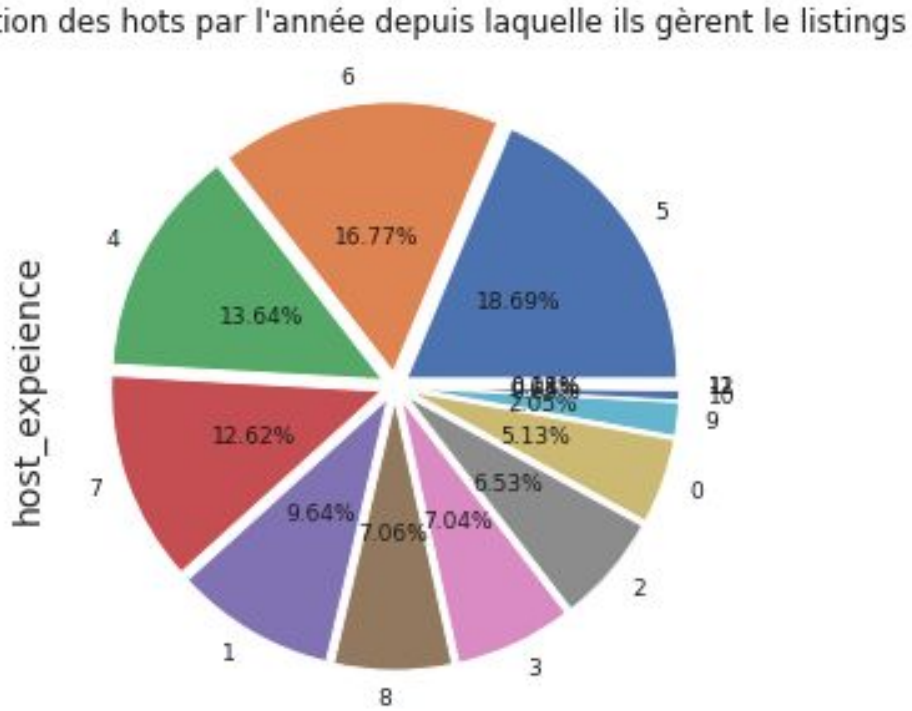
host_location: nous pouvons l'utiliser pour déterminer si l'hôte est local ou non



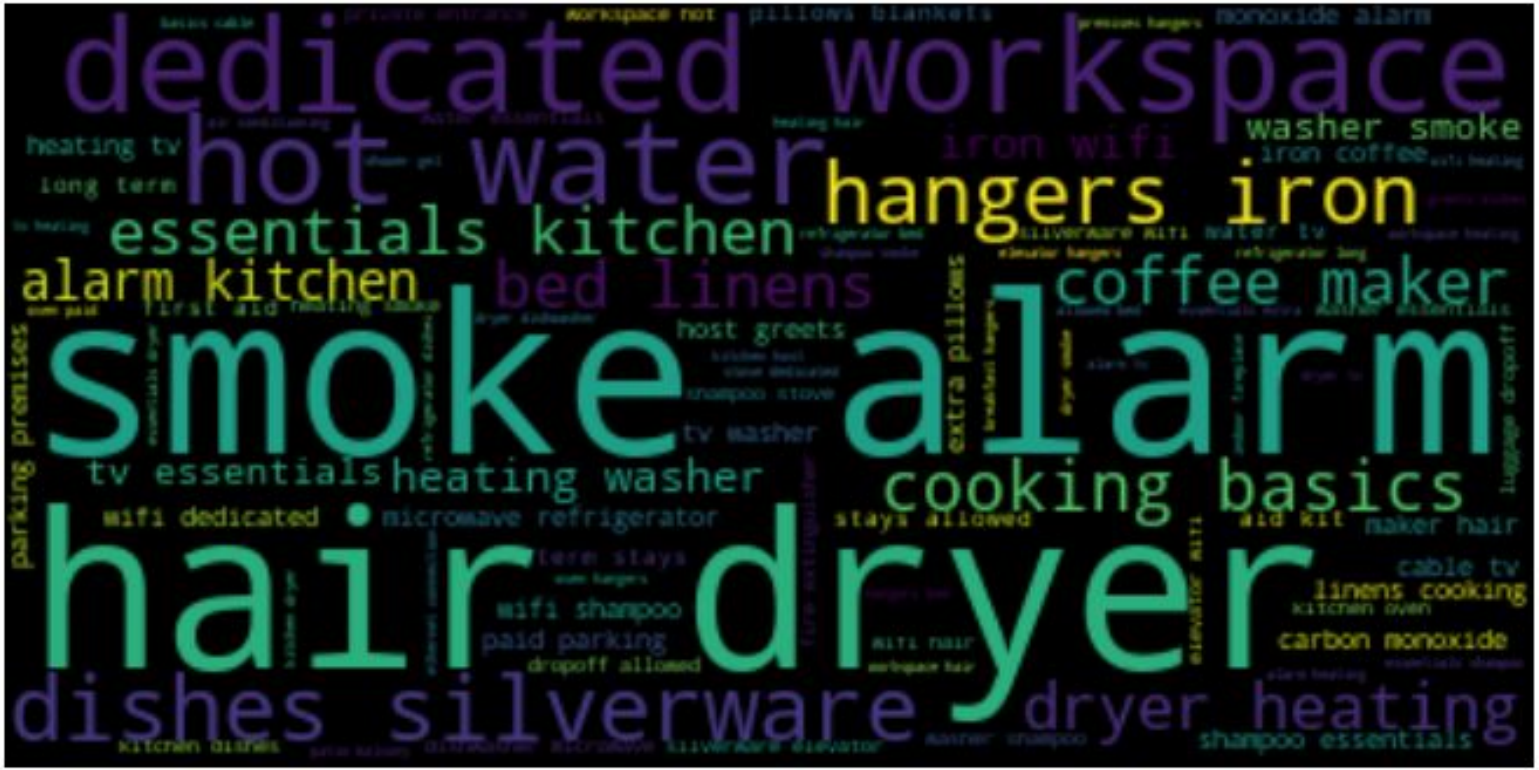
host_since: peut être utilisé pour calculer l'expérience des hôtes en fonction de la durée depuis leur première inscription



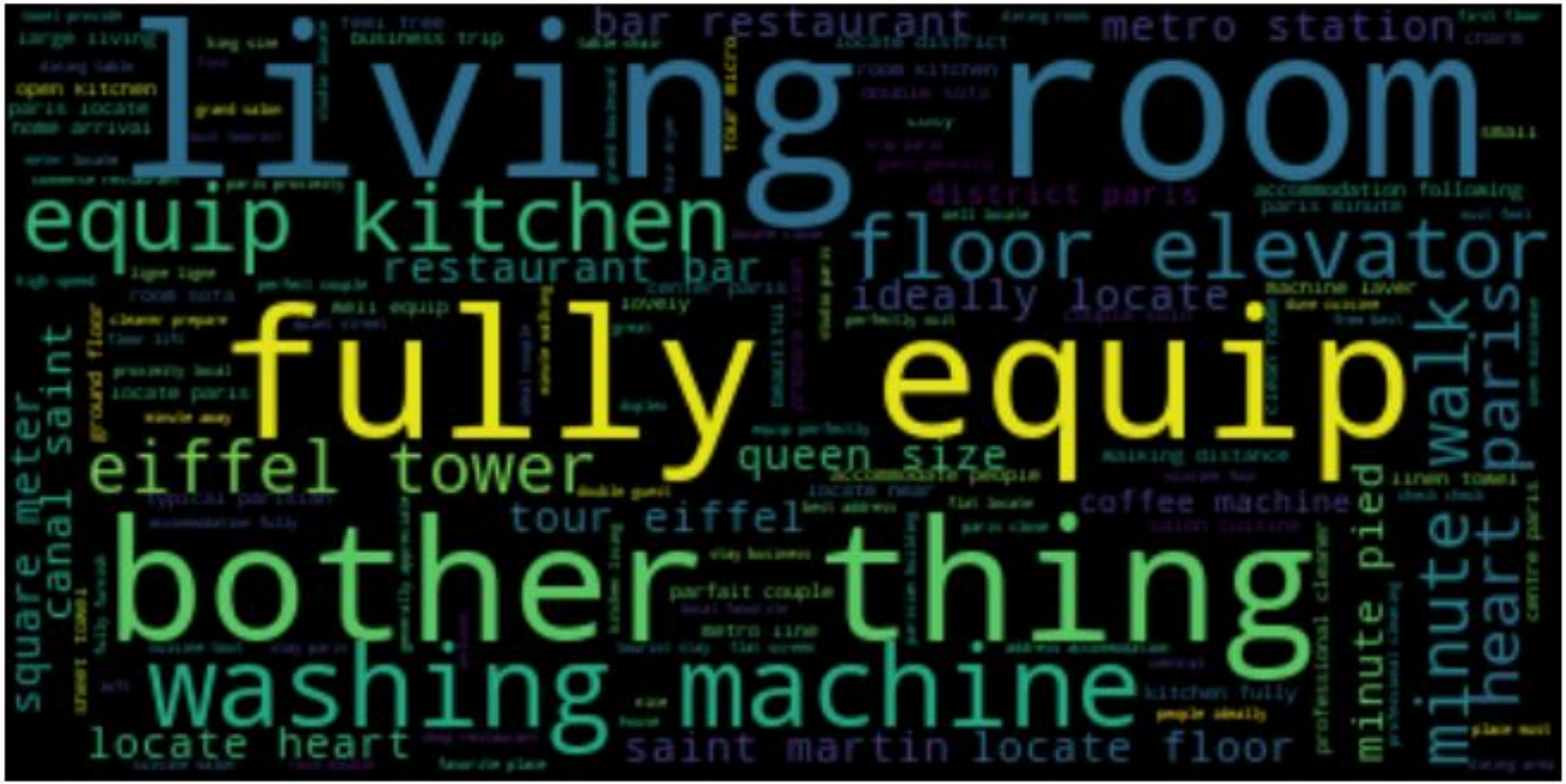
host_since: peut être utilisé pour calculer l'expérience des hôtes en fonction de la durée depuis leur première inscription



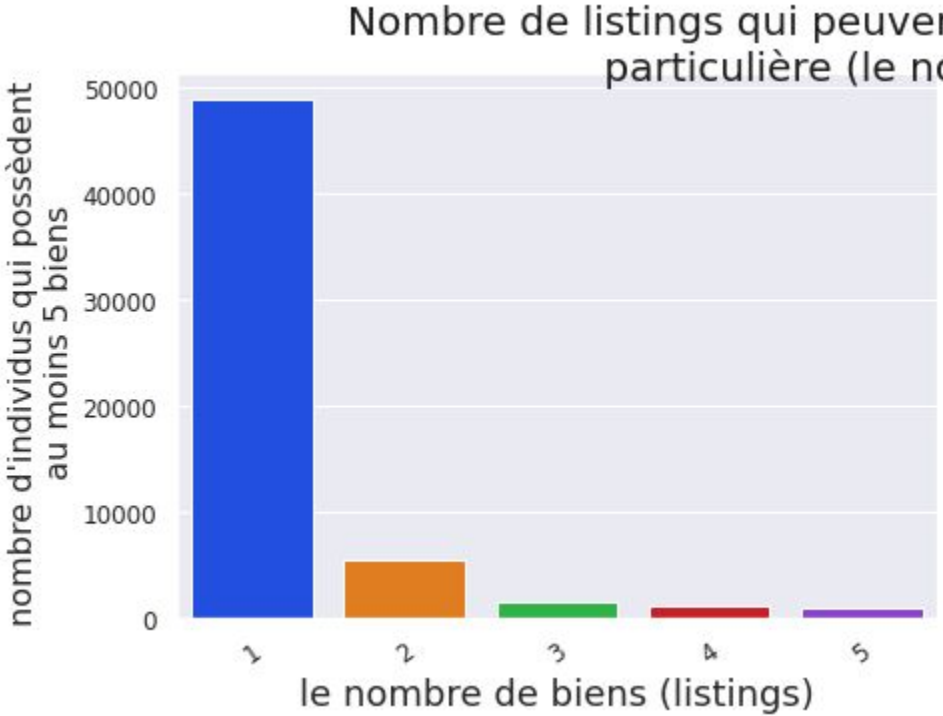
amenities: Créer des features à partir de amenities (équipements)

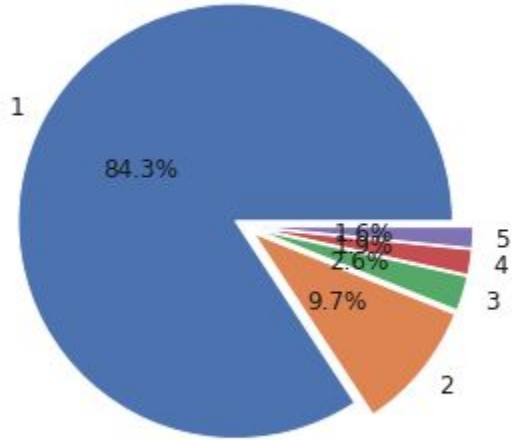


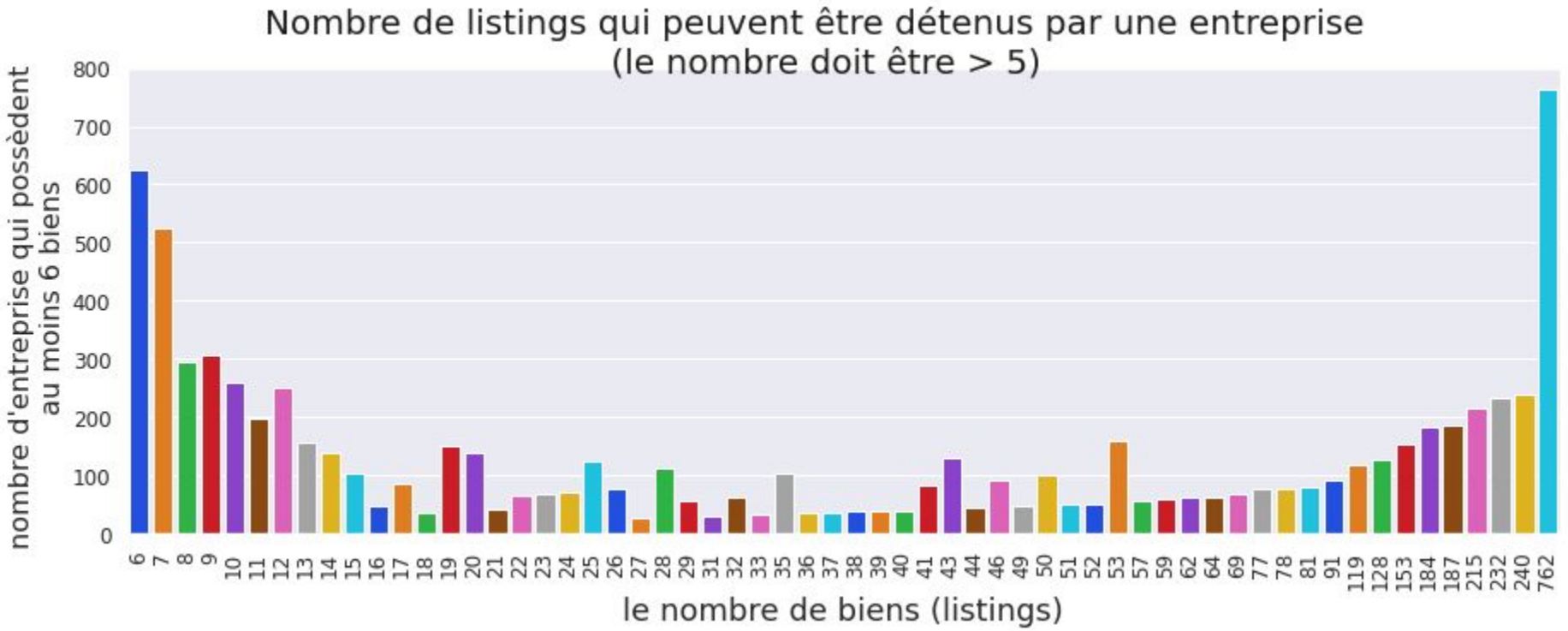
description : extraction de features à partir de la description (colonne textuelle) à l'aide d'un pipline NLP



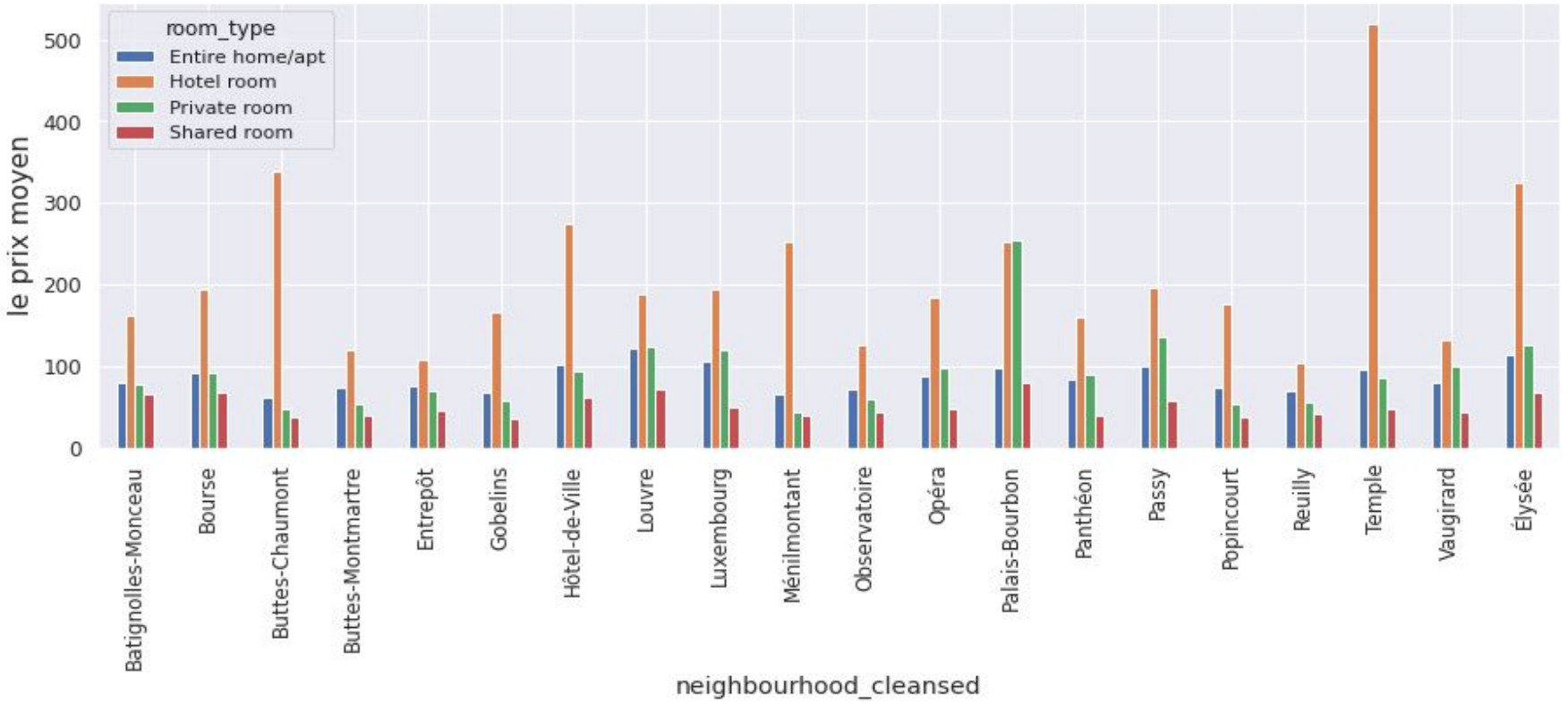
calculated_host_listings_count: valeur continue qui correspond aux nombres effectifs de listings pour les hôtes - mesure permettant de déterminer l'expérience des hôtes ou de distinguer les entreprises et les



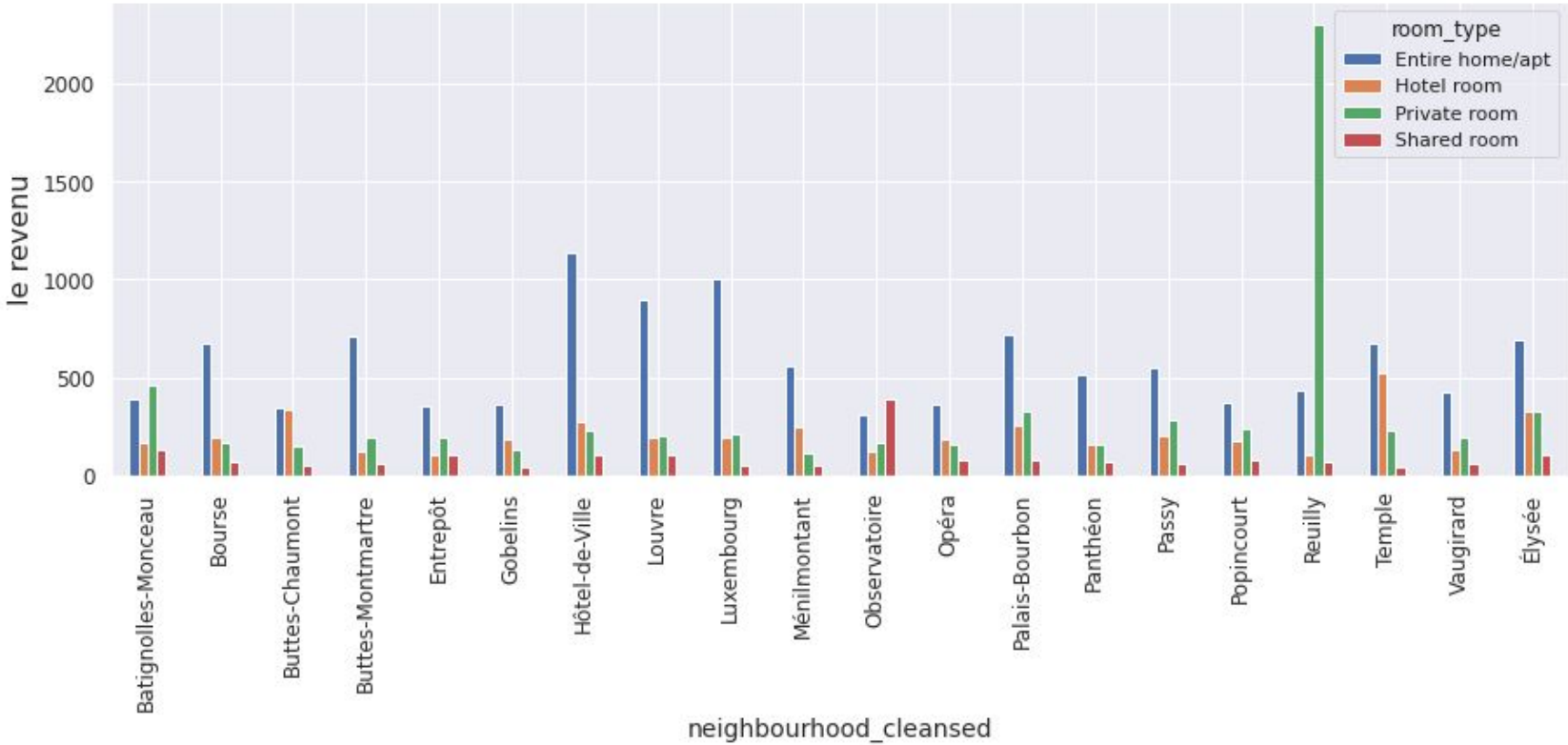



Nombre de listings qui peuvent être détenus par une entreprise
(le nombre doit être > 5)
nombre d'entreprise qui possèdent
au moins 6 biens
0
100
200
300
400
500
600
700
800
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
31
32
33
35
36
37
38
39
40
41
43
44
46
49
50
51
52
53
57
59
62
64
69
77
78
81
91
119
128
153
184
187
215
232
240
762
le nombre de biens (listings)

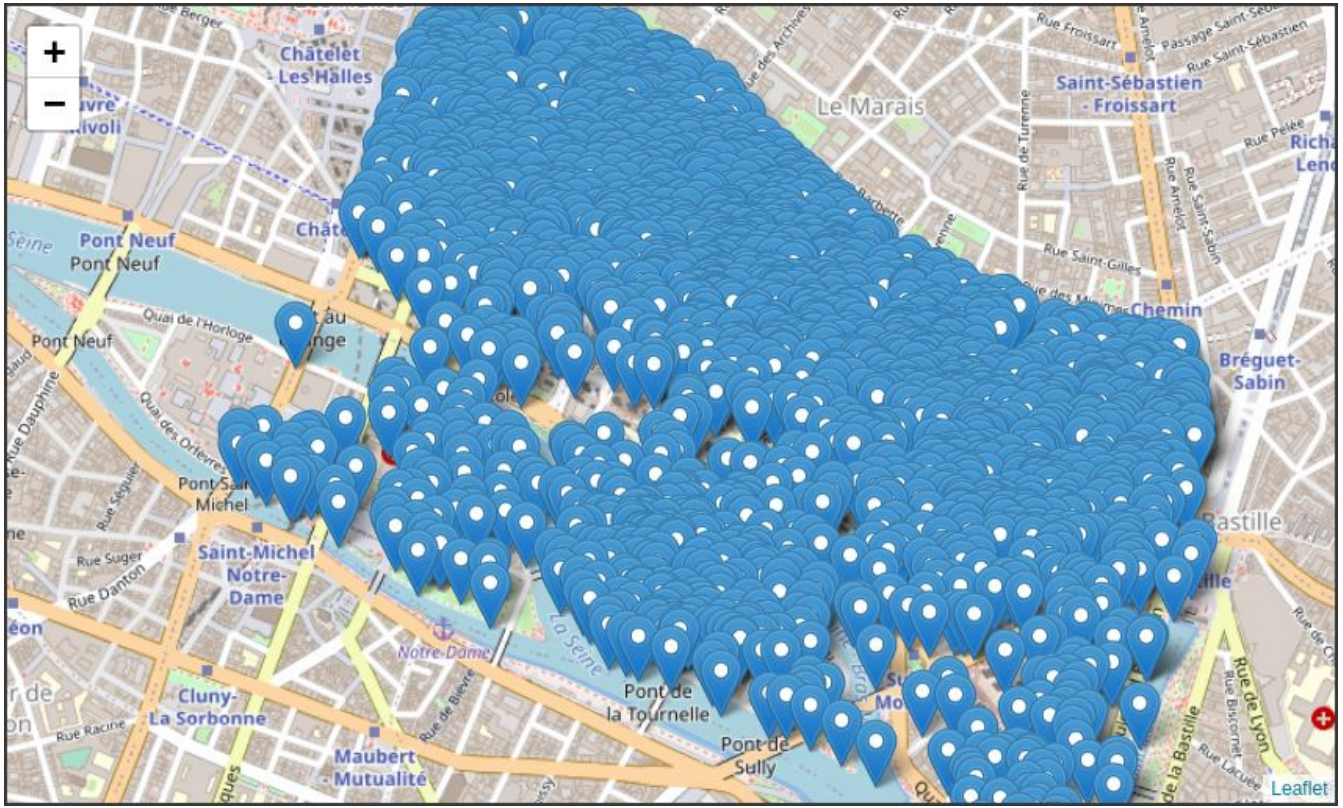
Visualisation du prix moyen du jour pour deux personnes



Calculer les revenus estimés pour chaque listing, les revenus estimés pour chaque listing seront calculés sur la base du prix d'une nuit et du nombre minimum de nuits à partir de la base de données

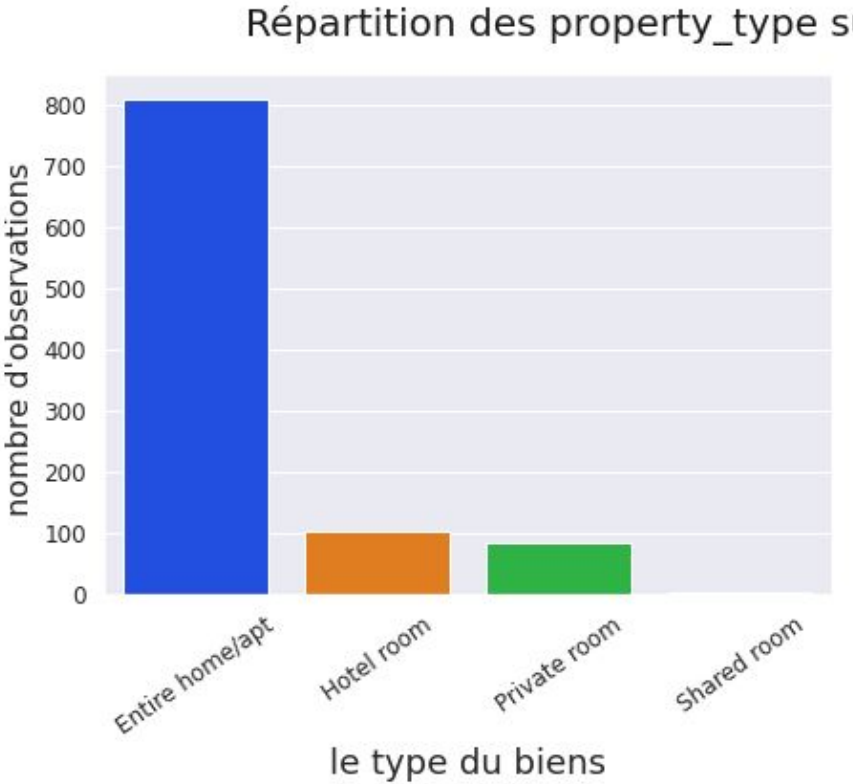


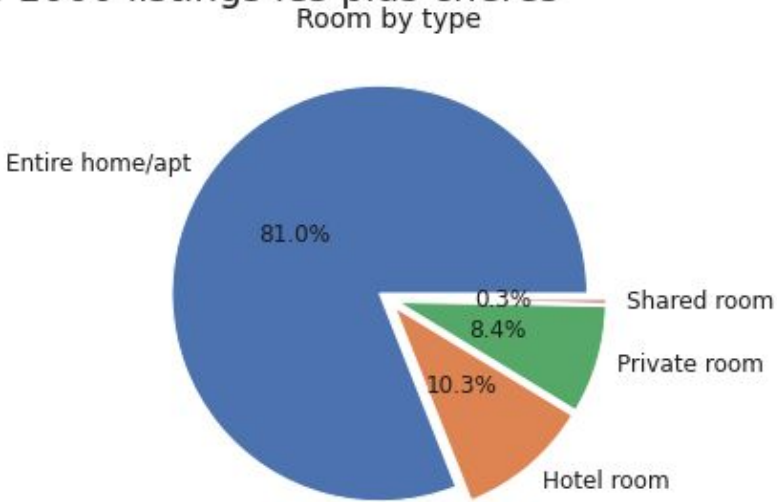
Les listings dans le quartier de l'Hôtel-de-Ville en utilisant folium



Cette fois-ci en se basant sur la colonne prix au lieu de revenu je récupère les 100 listings les plus chers

Type de bien a acherter





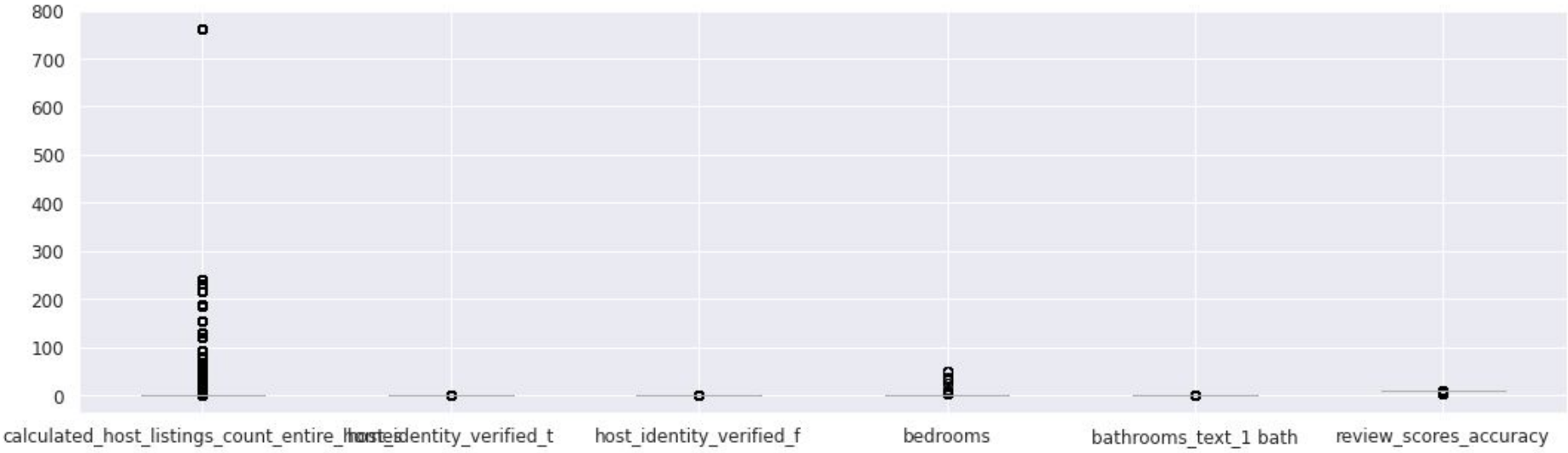
Cette fois-ci en se basant sur la colonne prix au lieu de revenu je récupère les 100 listings les plus chers

La partie prédiction du prix des listings

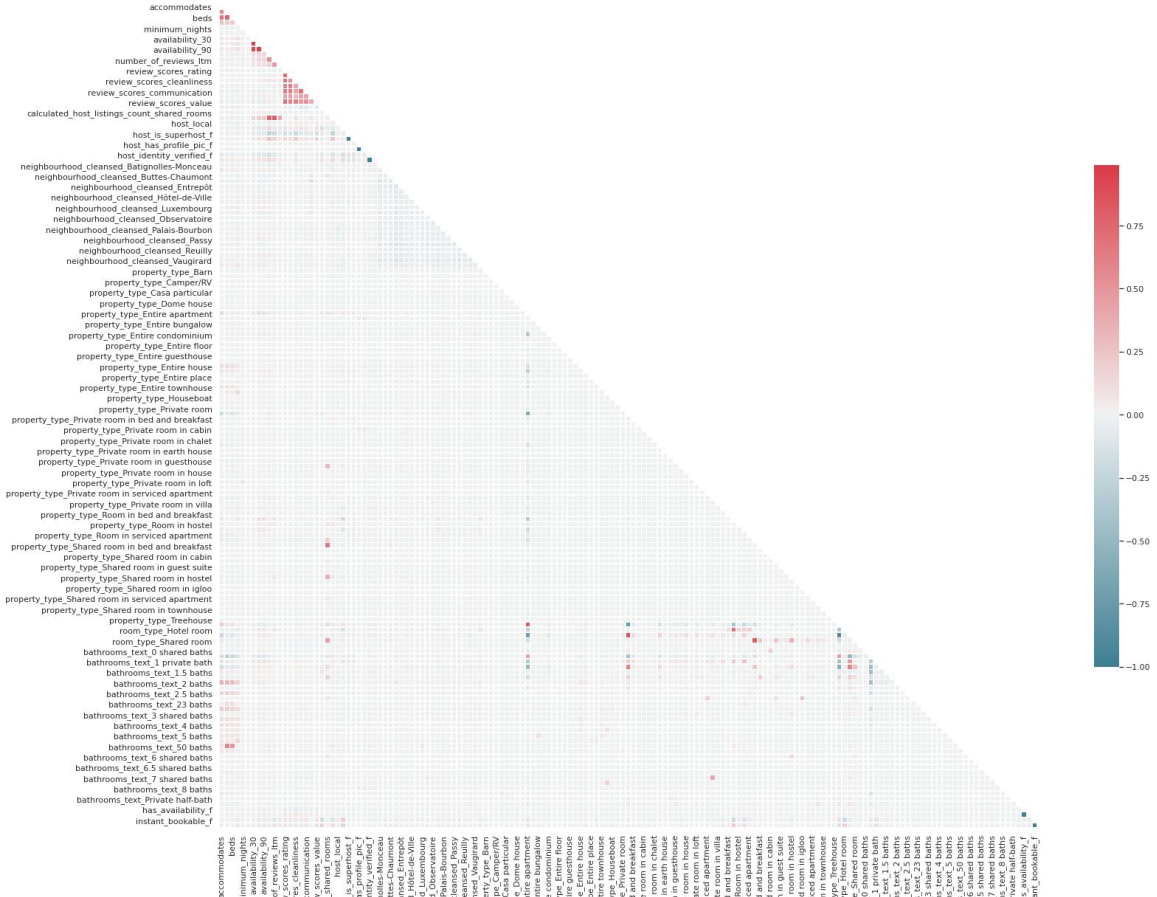
Ici, j'ai fait une sélection de features et une analyse de corrélation afin de trouver le meilleur modèle qui pourrait s'adapter à nos données

Le premier essai : j'ai utilisé un ensemble de données de toutes les colonnes prétraitées sauf celles créées dans la partie ingénierie de données (menities, host_verifications, description)

Suppression des colonnes comportant beaucoup de valeurs aberrantes



établir une graphique collinearity heatmap pour détecter les features corrélées qui pourraient empêcher notre modèle de converger



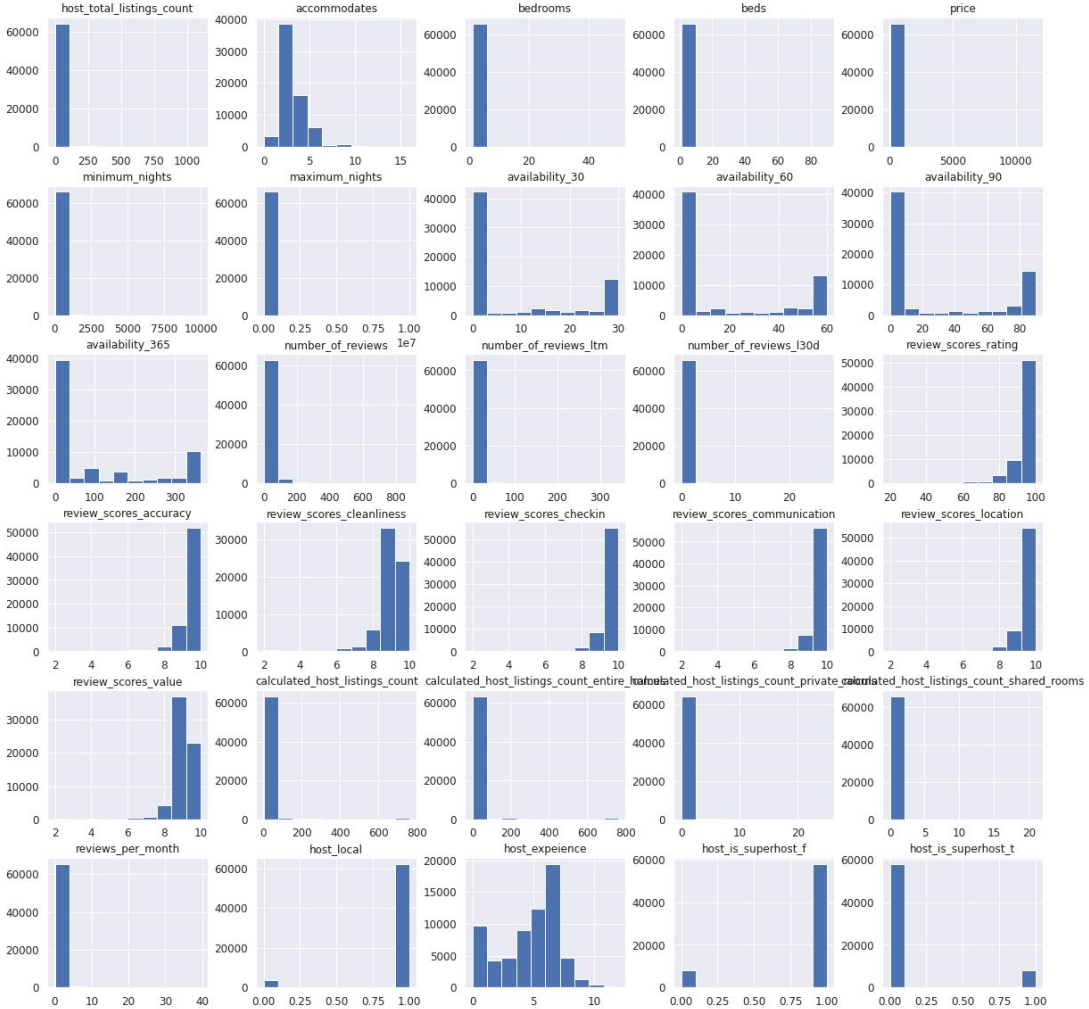
Normalisation et standardisation:

A l'exception de **accommodates** et de **host_experience**, les autres features numériques sont toutes asymétriques et pourraient bénéficier d'une transformation logarithmique

J'ai utilisé différentes méthodes pour standardiser mes données :

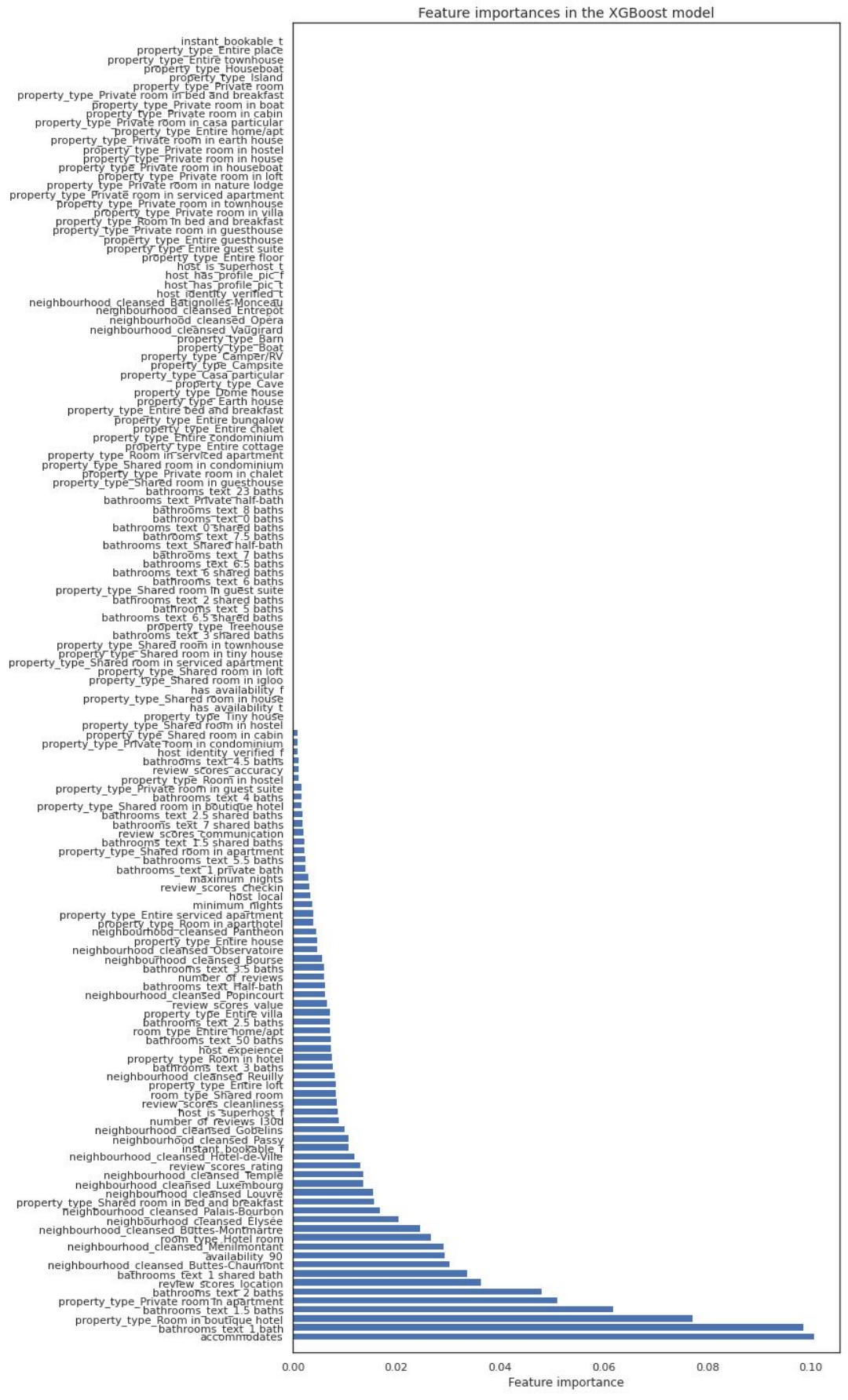
- StandardScaler
- MinMaxScaler
- RobustScaler

Et j'ai gardé MinMaxScaler puisqu'il conserve la structure globale des données



R SQUARED et RMSE

	RL	RIDGE	LASSO	XGB trees
TRAIN	Training RMSE: 0.1914 Training r2: 0.5128	Training RMSE: 0.1915 Training r2: 0.5125	Training RMSE: 0.1914 Training r2: 0.5128	Training MSE: 0.1827 Training r2: 0.5351
TEST	Validation RMSE: 0.191 Validation r2: **0.5193**	Validation RMSE: 0.191 Validation r2: **0.5193**	Validation RMSE: 0.191 Validation r2: **0.5193**	Validation MSE: 0.1843 Validation r2: **0.5361**


Feature importances in the XGBoost model
instant_bookable_t
property_type_Entire place
property_type_Entire townhouse
property_type_Houseboat
property_type_Island
property_type_Private room
property_type_Private room in bed and breakfast
property_type_Private room in boat
property_type_Private room in cabin
property_type_Private room in casa particular
property_type_Entire home/apt
property_type_Private room in earth house
property_type_Private room in hostel
property_type_Private room in house
property_type_Private room in houseboat
property_type_Private room in loft
property_type_Private room in nature lodge
property_type_Private room in serviced apartment
property_type_Private room in townhouse
property_type_Private room in villa
property_type_Room in bed and breakfast
property_type_Private room in guesthouse
property_type_Entire guesthouse
property_type_Entire guest suite
property_type_Entire floor
host_is_superhost_t
host_has_profile_pic_f
host_has_profile_pic_t
host_identity_verified_t
neighbourhood_cleansed_Batignolles-Monceau
neighbourhood_cleansed_Entrepot
neighbourhood_cleansed_Opéra
neighbourhood_cleansed_Vaugirard
property_type_Barn
property_type_Boat
property_type_Camper/RV
property_type_Campsite
property_type_Casa particular
property_type_Cave
property_type_Dome house
property_type_Earth house
property_type_Entire bed and breakfast
property_type_Entire bungalow
property_type_Entire chalet
property_type_Entire condominium
property_type_Entire cottage
property_type_Room in serviced apartment
property_type_Shared room in condominium
property_type_Private room in chalet
property_type_Shared room in guesthouse
bathrooms_text_2.5 baths
bathrooms_text_Private half-bath
bathrooms_text_8 baths
bathrooms_text_0 baths
bathrooms_text_0 shared baths
bathrooms_text_7.5 baths
bathrooms_text_Shared half-bath
bathrooms_text_7 baths
bathrooms_text_6.5 baths
bathrooms_text_6 shared baths
bathrooms_text_6 baths
property_type_Shared room in guest suite
bathrooms_text_2 shared baths
bathrooms_text_5 baths
bathrooms_text_6.5 shared baths
property_type_Treehouse
bathrooms_text_3 shared baths
property_type_Shared room in townhouse
property_type_Shared room in tiny house
property_type_Shared room in serviced apartment
property_type_Shared room in loft
property_type_Shared room in igloo
has_availability_f
property_type_Shared room in house
has_availability_t
property_type_Tiny house
property_type_Shared room in hostel
property_type_Shared room in cabin
property_type_Private room in condominium
host_identity_verified_f
bathrooms_text_4.5 baths
review_scores_accuracy
property_type_Room in hostel
property_type_Private room in guest suite
bathrooms_text_4 baths
property_type_Shared room in boutique hotel
bathrooms_text_2.5 shared baths
bathrooms_text_7 shared baths
review_scores_communication
bathrooms_text_1.5 shared baths
property_type_Shared room in apartment
bathrooms_text_5.5 baths
bathrooms_text_1 private bath
maximum_nights
review_scores_checkin
host_local
minimum_nights
property_type_Entire serviced apartment
property_type_Room in aparthotel
neighbourhood_cleansed_Panthéon
property_type_Entire house
neighbourhood_cleansed_Observatoire
neighbourhood_cleansed_Bourse
bathrooms_text_3.5 baths
number_of_reviews
bathrooms_text_Half-bath
neighbourhood_cleansed_Popincourt
review_scores_value
property_type_Entire villa
bathrooms_text_2.5 baths
room_type_Entire home/apt
bathrooms_text_50 baths
host_expeience
property_type_Room in hotel
bathrooms_text_3 baths
neighbourhood_cleansed_Reuilly
property_type_Entire loft
room_type_Shared room
review_scores_cleanliness
host_is_superhost_f
number_of_reviews_l30d
neighbourhood_cleansed_Gobelins
neighbourhood_cleansed_Passy
instant_bookable_f
neighbourhood_cleansed_Hôtel-de-Ville
review_scores_rating
neighbourhood_cleansed_Temple
neighbourhood_cleansed_Luxembourg
neighbourhood_cleansed_Louvre
property_type_Shared room in bed and breakfast
neighbourhood_cleansed_Palais-Bourbon
neighbourhood_cleansed_Elysée
neighbourhood_cleansed_Buttes-Montmartre
room_type_Hotel room
neighbourhood_cleansed_Ménilmontant
availability_90
neighbourhood_cleansed_Buttes-Chaumont
bathrooms_text_1 shared bath
review_scores_location
bathrooms_text_2 baths
property_type_Private room in apartment
bathrooms_text_1.5 baths
property_type_Room in boutique hotel
bathrooms_text_1 bath
accommodates
0.00
0.02
0.04
0.06
0.08
0.10
Feature importance

Le deuxième essai : J'ai utilisé un ensemble de données de toutes les colonnes prétraitées et aussi celles créées dans la partie ingénierie des données (menities, host_verifications, description)

Mon jeu de données était de taille (65917, 2673)

R SQUARED et RMSE

	RL	RIDGE	LASSO	XGB trees
TRAIN	Training RMSE: 0.1504 Training r2: **0.6172**	Training RMSE: 0.1424 Training r2: **0.6376**	Training RMSE: 0.1421 Training r2: **0.6384**	Training MSE: 0.1714 Training r2: 0.5637
TEST	Validation RMSE: 2.8569801186007695 e+18 Validation r2: **-7.190735871508386 e+18**	Validation RMSE: 0.1567 Validation r2: **0.6056**	Validation RMSE: 0.1572 Validation r2: **0.6043**	Validation MSE: 0.1783 Validation r2: **0.5512**

RANDOM FOREST

AVEC n_estimators = 20	AVEC n_estimators = 100
Training RMSE: 0.0271 Validation RMSE: 0.1774 Training r2: **0.931** Validation r2: **0.5535**	Training RMSE: 0.0234 Validation RMSE: 0.1688 Training r2: **0.9403** Validation r2: **0.5753**

MERCI
POUR VOTRE
ATTENTION

Questions ?