lnlzlj1iv

January 22, 2025

1 Projet Airbnb - WU Kylie

But du projet : prédire le logarithme du prix de location d'un Airbnb à partir d'un ensemble de caractéristiques.

1.0.1 Importation des bibliothèques

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[2]:
```
import plotly.express as px
import folium
import json

#from pandas.plotting import register_matplotlib_converters
#register_matplotlib_converters()
```

2 I. Partie exploration qualitative des données

2.0.1 1. Présentation de la base

[3]:
```
# Chargement des données dans un dataframe Pandas
dataset = pd.read_csv("airbnb_train.csv", sep=",")

# Aperçu des données :
dataset.head()
```

[3]:
```
         id  log_price property_type        room_type  \
0   5708593   4.317488         House     Private room
1  14483613   4.007333         House     Private room
2  10412649   7.090077     Apartment  Entire home/apt
3  17954362   3.555348         House     Private room
4   9969781   5.480639         House  Entire home/apt
```

```
                                           amenities  accommodates  bathrooms  \
0  {TV,"Wireless Internet",Kitchen,"Free parking …             3        1.0
1  {"Wireless Internet","Air conditioning",Kitche…             4        2.0
2  {TV,"Wireless Internet","Air conditioning",Kit…             6        2.0
3  {TV,"Cable TV",Internet,"Wireless Internet","A…             1        1.0
4  {TV,"Cable TV",Internet,"Wireless Internet",Ki…             4        1.0

   bed_type cancellation_policy  cleaning_fee  … last_review   latitude  \
0  Real Bed            flexible         False  …         NaN  33.782712
1  Real Bed              strict         False  …  2017-09-17  40.705468
2  Real Bed            flexible         False  …         NaN  38.917537
3  Real Bed            flexible          True  …  2017-09-29  40.736001
4  Real Bed            moderate          True  …  2017-08-28  37.744896

    longitude                                              name  \
0 -118.134410                            Island style Spa Studio
1  -73.909439  Beautiful and Simple Room W/2 Beds, 25 Mins to…
2  -77.031651  2br/2ba luxury condo perfect for infant / toddler
3  -73.924248   Manhattan view from Queens. Lovely single room .
4 -122.430665                      Zen Captured Noe Valley House

      neighbourhood number_of_reviews review_scores_rating zipcode bedrooms  \
0        Long Beach                 0                  NaN   90804      0.0
1         Ridgewood                38                 86.0   11385      1.0
2  U Street Corridor                0                  NaN   20009      2.0
3         Sunnyside                19                 96.0   11104      1.0
4        Noe Valley                15                 96.0   94131      2.0

   beds
0   2.0
1   2.0
2   2.0
3   1.0
4   2.0

[5 rows x 28 columns]
```

[4]:
```
# Informations supplémentaires :
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22234 entries, 0 to 22233
Data columns (total 28 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      22234 non-null  int64
 1   log_price               22234 non-null  float64
 2   property_type           22234 non-null  object
 3   room_type               22234 non-null  object
 4   amenities               22234 non-null  object
 5   accommodates            22234 non-null  int64
 6   bathrooms               22183 non-null  float64
 7   bed_type                22234 non-null  object
 8   cancellation_policy     22234 non-null  object
 9   cleaning_fee            22234 non-null  bool
 10  city                    22234 non-null  object
 11  description             22234 non-null  object
 12  first_review            17509 non-null  object
 13  host_has_profile_pic    22178 non-null  object
 14  host_identity_verified  22178 non-null  object
 15  host_response_rate      16759 non-null  object
 16  host_since              22178 non-null  object
 17  instant_bookable        22234 non-null  object
 18  last_review             17518 non-null  object
 19  latitude                22234 non-null  float64
 20  longitude               22234 non-null  float64
 21  name                    22234 non-null  object
 22  neighbourhood           20148 non-null  object
 23  number_of_reviews       22234 non-null  int64
 24  review_scores_rating    17256 non-null  float64
 25  zipcode                 21931 non-null  object
 26  bedrooms                22208 non-null  float64
 27  beds                    22199 non-null  float64
dtypes: bool(1), float64(7), int64(3), object(17)
memory usage: 4.6+ MB
```

[5]:
```
dataset.shape
```

[5]:
```
(22234, 28)
```

Donc il y a 22233 offres étudiées et 28 caractéristiques initiales.

Les caractéristiques étudiées (intitulés des colonnes) :

[6]:
```
print(dataset.columns)
```

```
Index(['id', 'log_price', 'property_type', 'room_type', 'amenities',
       'accommodates', 'bathrooms', 'bed_type', 'cancellation_policy',
       'cleaning_fee', 'city', 'description', 'first_review',
       'host_has_profile_pic', 'host_identity_verified', 'host_response_rate',
       'host_since', 'instant_bookable', 'last_review', 'latitude',
       'longitude', 'name', 'neighbourhood', 'number_of_reviews',
       'review_scores_rating', 'zipcode', 'bedrooms', 'beds'],
      dtype='object')
```

Je choisis de supprimer les colonnes suivantes : description, first_review, first_review, host_has_profile_pic, host_identity_verified, host_response_rate, host_since, instant_bookable, last_review, name, number_of_reviews, review_scores_rating

car je trouve qu'elles n'influent pas principalement le prix, donc sont moins intéressantes pour travailler sur un dataset ayant des informations plus essentielles.

[7]:
```
# Nouveau dataset :
df = dataset.drop(['description', 'first_review', 'host_has_profile_pic',
                   'host_identity_verified', 'host_response_rate', 'host_since',
                   'instant_bookable', 'last_review', 'name', 'number_of_reviews',
                   'review_scores_rating'], axis=1)
# Et j'ai remarqué qu'il y avait quelques 'NaN' (Not a Number) dans la colonne␣
 ↪des zipcode
# donc je supprime ces lignes
df = df.dropna(subset=['zipcode'])
```

[8]:
```
# Aperçu :
df.head()
```

[8]:
```
         id  log_price property_type        room_type  \
0   5708593   4.317488         House     Private room
1  14483613   4.007333         House     Private room
2  10412649   7.090077     Apartment  Entire home/apt
3  17954362   3.555348         House     Private room
4   9969781   5.480639         House  Entire home/apt

                                           amenities  accommodates  bathrooms  \
0  {TV,"Wireless Internet",Kitchen,"Free parking …             3        1.0
1  {"Wireless Internet","Air conditioning",Kitche…             4        2.0
2  {TV,"Wireless Internet","Air conditioning",Kit…             6        2.0
3  {TV,"Cable TV",Internet,"Wireless Internet","A…             1        1.0
4  {TV,"Cable TV",Internet,"Wireless Internet",Ki…             4        1.0

   bed_type cancellation_policy  cleaning_fee city   latitude   longitude  \
0  Real Bed            flexible         False   LA  33.782712 -118.134410
1  Real Bed              strict         False  NYC  40.705468  -73.909439
2  Real Bed            flexible         False   DC  38.917537  -77.031651
3  Real Bed            flexible          True  NYC  40.736001  -73.924248
4  Real Bed            moderate          True   SF  37.744896 -122.430665

       neighbourhood zipcode  bedrooms  beds
0         Long Beach   90804       0.0   2.0
1          Ridgewood   11385       1.0   2.0
2  U Street Corridor   20009       2.0   2.0
3          Sunnyside   11104       1.0   1.0
4         Noe Valley   94131       2.0   2.0
```

Les intitulés des colonnes sont désormais :

```
[9]: print(df.columns)
```

```
Index(['id', 'log_price', 'property_type', 'room_type', 'amenities',
       'accommodates', 'bathrooms', 'bed_type', 'cancellation_policy',
       'cleaning_fee', 'city', 'latitude', 'longitude', 'neighbourhood',
       'zipcode', 'bedrooms', 'beds'],
      dtype='object')
```

2.0.2 2. Exploration qualitative de la base (distribution des données, répartition)

Pour chacune des colonnes intéressantes, voyons les possibilités d'éléments présents :

```
[10]: print(df['property_type'].unique())
```

```
['House' 'Apartment' 'Townhouse' 'Guest suite' 'Condominium' 'Timeshare'
 'Chalet' 'Guesthouse' 'Bungalow' 'Loft' 'In-law' 'Boat' 'Dorm' 'Other'
 'Bed & Breakfast' 'Camper/RV' 'Villa' 'Boutique hotel' 'Cabin' 'Hostel'
 'Hut' 'Yurt' 'Serviced apartment' 'Castle' 'Vacation home' 'Tent' 'Cave'
 'Tipi' 'Earth House' 'Island' 'Treehouse']
```

```
[11]: print(df['room_type'].unique())
```

```
['Private room' 'Entire home/apt' 'Shared room']
```

```
[12]: print(df['amenities'].unique())
```

```
['{TV,"Wireless Internet",Kitchen,"Free parking on premises","Pets
allowed","Suitable for events",Washer,Dryer,"Smoke
detector",Essentials,Shampoo,"Lock on bedroom door",Hangers,"Hair dryer",Iron}'
 '{"Wireless Internet","Air conditioning",Kitchen,Heating,"Family/kid
friendly","Smoke detector","Carbon monoxide detector","Fire
extinguisher",Essentials,"Lock on bedroom door","24-hour check-in","Hair
dryer",Iron,"translation missing: en.hosting_amenity_50","Self Check-
In",Keypad,"Bed linens",Microwave,"Coffee maker",Refrigerator,"Dishes and
silverware","Cooking basics",Oven,Stove,"Luggage dropoff allowed"}'
 '{TV,"Wireless Internet","Air conditioning",Kitchen,"Free parking on
premises","Pets allowed","Elevator in building",Heating,"Family/kid
friendly",Washer,Dryer,"Smoke detector","Carbon monoxide detector","First aid
kit","Safety card","Fire extinguisher",Essentials,Shampoo,"Lock on bedroom
door","Laptop friendly workspace"}'
 …
 '{TV,Internet,"Wireless Internet","Air
conditioning",Kitchen,Gym,Elevator,"Buzzer/wireless
intercom",Heating,"Family/kid friendly",Washer,Dryer,"Smoke detector","Carbon
monoxide detector","First aid kit",Essentials,Shampoo,Hangers,"Hair
dryer",Iron,"Laptop friendly workspace","Hot water","Bed
linens",Microwave,"Coffee maker",Refrigerator,Dishwasher,"Dishes and
silverware","Cooking basics",Oven,Stove}'
```

```
 '{TV,"Wireless Internet","Air conditioning",Kitchen,Heating,Washer,Dryer,"Smoke
detector","Carbon monoxide detector",Essentials,Shampoo,Hangers,"Hair
dryer",Iron,"Laptop friendly workspace"}'
 '{TV,Internet,"Wireless Internet",Kitchen,"Free parking on
premises",Heating,"Family/kid friendly","Smoke detector","Carbon monoxide
detector","First aid kit","Safety card","Fire
extinguisher",Essentials,Shampoo,"24-hour check-in",Hangers,"Hair dryer","Laptop
friendly workspace"}']
```

Il serait intéressant d'extraire des termes d'équipements, pour voir s'ils réaparaissent dans d'autres descriptions (a priori plus il y en a, plus le prix sera élevé) :

[13]:
```python
import re

# Fonction pour enlever les guillemets, accolades et crochets
def remove_caracteres(s):
    return re.sub(r'["\'\{\}\[\]]', '', s)

# On applique remove_caracteres
amenities_separes = df['amenities'].apply(lambda x: [remove_caracteres(item.
 ↪strip()) for item in str(x).split(',')])

# Garder les éléments uniques
amenities_bis = [item for sublist in amenities_separes for item in sublist]
uniques_amenities = list(set(amenities_bis))

print(uniques_amenities)
```

```
['Waterfront', '', 'Smoke detector', 'Luggage dropoff allowed', 'Cooking
basics', 'Paid parking off premises', 'Iron', 'Path to entrance lit at night',
'Doorman', 'Ground floor access', 'Crib', 'Single level home', 'Private living
room', 'Host greets you', 'Wheelchair accessible', 'Indoor fireplace', 'Air
conditioning', 'Fixed grab bars for shower & toilet', 'Cleaning before
checkout', 'Hot water kettle', 'Lake access', 'TV', 'Smart lock', 'Wide
entryway', 'Extra pillows and blankets', 'Dryer', 'Other pet(s)', 'Family/kid
friendly', 'Lock on bedroom door', 'Beachfront', 'Patio or balcony', 'Pets
allowed', 'Kitchen', 'Shampoo', 'Pocket wifi', 'Buzzer/wireless intercom',
'Fireplace guards', 'Handheld shower head', 'Smartlock', 'Ski in/Ski out',
'Roll-in shower with chair', 'Wide doorway', 'Cable TV', 'Smoking allowed',
'Coffee maker', 'Pack 'n Play/travel crib', 'Carbon monoxide detector', 'Wide
hallway clearance', 'Window guards', 'Washer / Dryer', 'Garden or backyard',
'Free parking on street', 'Hangers', 'Ethernet connection', 'Dog(s)', 'Pool',
'Laptop friendly workspace', 'EV charger', 'Cat(s)', 'Grab-rails for shower and
toilet', 'Flat smooth pathway to front door', 'Children's dinnerware', 'smooth
pathway to front door', 'Bathtub with shower chair', 'Hair dryer', 'Private
bathroom', 'Suitable for events', 'Well-lit path to entrance', 'Baby bath',
'Microwave', 'Bathtub', 'Gym', 'Stove', 'Table corner guards', 'Washer', 'Wide
clearance to bed', 'Free parking on premises', 'Pets live on this property',
'Hot tub', 'Stair gates', 'Flat', 'Essentials', 'BBQ grill', 'Private entrance',
'First aid kit', 'Safety card', 'Elevator in building', 'Self Check-In', 'Step-
free access', 'Doorman Entry', 'Changing table', 'Firm mattress', 'Dishwasher',
'translation missing: en.hosting_amenity_49', 'Lockbox', 'Accessible-height
bed', 'Keypad', '24-hour check-in', 'Heating', 'Refrigerator', 'Outlet covers',
'Wide clearance to shower & toilet', 'Hot water', 'Beach essentials', 'High
chair', 'Wireless Internet', 'Baby monitor', 'Fire extinguisher', 'Dishes and
silverware', 'Breakfast', 'translation missing: en.hosting_amenity_50', 'Room-
darkening shades', 'Disabled parking spot', 'Long term stays allowed', 'Game
console', 'Children's books and toys', 'Bed linens', 'Elevator', 'Oven', 'Air
purifier', 'Babysitter recommendations', 'Internet', 'Accessible-height toilet',
'Firm matress', 'Other']
```

[14]:
```python
print(df['accommodates'].unique())
```

```
[ 3  4  6  1  2 10  7  5 16  8  9 14 12 15 13 11]
```

[15]:
```python
print(df['bathrooms'].unique())
```

```
[1.  2.  3.5 1.5 3.  2.5 0.  nan 5.  4.  6.5 0.5 5.5 6.  7.  4.5 8.  7.5]
```

A priori plus il y a de salles de bains, plus le prix sera élevé.

[16]:
```python
# Afficher les lignes où il y a 5 salles de bain.
lignes_avec_5_salles_de_bain = df.loc[df['bathrooms'] == 5., :]
print(lignes_avec_5_salles_de_bain)
```

```
             id  log_price property_type        room_type  \
503    17257473   6.907755         House  Entire home/apt
1696   18053971   6.678342         House  Entire home/apt
1972   20978478   6.476972         House  Entire home/apt
5237   20761244   4.488636     Apartment      Shared room
8293   15939618   5.796058     Apartment  Entire home/apt
8902   16618193   6.579251         House  Entire home/apt
11146  13216537   4.553877        Hostel     Private room
12698    386966   6.204558         House  Entire home/apt
13286   4092283   6.684612         House  Entire home/apt
15449   3293695   4.094345     Apartment     Private room
16000  19820146   6.212606         House  Entire home/apt
18246  14426987   6.652863         House  Entire home/apt
18254   2500560   6.745236         House  Entire home/apt
20839  20827821   6.476972         House  Entire home/apt
20967  13698365   6.684612         House  Entire home/apt
21171  19213967   6.023448     Apartment  Entire home/apt
21419  15484888   7.166266         House  Entire home/apt
22154  19296720   6.907755         Villa  Entire home/apt

                                               amenities  accommodates  \
503    {TV,"Cable TV",Internet,"Wireless Internet","A…             10
```

```
1696   {TV,"Cable TV",Internet,"Wireless Internet","A…           16
1972   {TV,"Cable TV",Internet,"Wireless Internet",Po…           10
5237   {TV,Internet,"Wireless Internet","Air conditio…            2
8293   {TV,"Cable TV","Wireless Internet","Air condit…           14
8902   {TV,Internet,"Wireless Internet",Kitchen,"Free…            9
11146  {TV,"Cable TV",Internet,"Wireless Internet","A…            2
12698  {TV,"Cable TV",Internet,"Wireless Internet","A…           16
13286  {TV,"Cable TV",Internet,"Wireless Internet","A…           16
15449  {TV,"Wireless Internet","Air conditioning",Kit…            2
16000  {TV,"Cable TV",Internet,"Wireless Internet","A…           16
18246  {TV,"Wireless Internet","Air conditioning",Poo…            8
18254  {TV,"Cable TV",Internet,"Wireless Internet","A…            8
20839  {TV,Internet,"Wireless Internet","Air conditio…           15
20967  {TV,"Cable TV",Internet,"Wireless Internet","A…            6
21171  {TV,"Cable TV",Internet,"Wireless Internet","A…            9
21419  {TV,"Cable TV",Internet,"Wireless Internet","A…           16
22154  {TV,"Cable TV",Internet,"Wireless Internet","A…            8

       bathrooms  bed_type cancellation_policy  cleaning_fee    city  \
503          5.0  Real Bed              strict          True      LA
1696         5.0  Real Bed              strict          True      LA
1972         5.0  Real Bed            flexible         False      LA
5237         5.0  Real Bed            flexible          True     NYC
8293         5.0  Real Bed            flexible          True      LA
8902         5.0  Real Bed              strict          True      LA
11146        5.0  Real Bed              strict         False  Boston
12698        5.0  Real Bed              strict          True      DC
13286        5.0  Real Bed              strict          True     NYC
15449        5.0  Real Bed              strict         False     NYC
16000        5.0  Real Bed              strict          True     NYC
18246        5.0  Real Bed              strict          True      LA
18254        5.0  Real Bed              strict          True      LA
20839        5.0  Real Bed              strict          True      LA
20967        5.0  Real Bed              strict          True     NYC
21171        5.0  Real Bed            flexible         False     NYC
21419        5.0  Real Bed              strict          True      LA
22154        5.0  Real Bed              strict          True      LA

        latitude   longitude          neighbourhood  zipcode  bedrooms  beds
503    34.114527 -118.581026                Topanga    90290       5.0  10.0
1696   34.130161 -118.363875                    NaN    90068       5.0  16.0
1972   33.879094 -118.391052        Manhattan Beach    90266       5.0   5.0
5237   40.853222  -73.935631     Washington Heights    10033       1.0   1.0
8293   34.067207 -118.449054               Westwood    90024       5.0  11.0
8902   33.992222 -118.456121                 Venice    90291       5.0   5.0
11146  42.336059  -71.045924           South Boston    02127       1.0   1.0
12698  38.939301  -77.018718               Petworth    20011       4.0   4.0
13286  40.684922  -73.954891     Bedford-Stuyvesant    11216       8.0  18.0
```

```
15449  40.693242  -73.921260              Bushwick    11221        1.0   1.0
16000  40.715211  -73.946142          Williamsburg  11211.0        8.0  11.0
18246  34.013029 -118.446103             Mar Vista    90066        4.0   4.0
18254  34.129597 -118.341948        Hollywood Hills    90068        4.0   4.0
20839  34.078944 -118.374199          Mid-Wilshire    90048        4.0   4.0
20967  40.782690  -73.983902        Upper West Side    10023        4.0   6.0
21171  40.683663  -73.941006     Bedford-Stuyvesant    11216        5.0   5.0
21419  34.097976 -118.426247  Bel Air/Beverly Crest    90210        5.0   5.0
22154  34.115600 -118.296714        Hollywood Hills    90027        5.0   5.0
```

[17]:
```
print(df['bed_type'].unique())
```

```
['Real Bed' 'Pull-out Sofa' 'Futon' 'Airbed' 'Couch']
```

[18]:
```
print(df['cancellation_policy'].unique())
```

```
['flexible' 'strict' 'moderate' 'super_strict_30' 'super_strict_60']
```

[19]:
```
print(df['cleaning_fee'].unique())
```

```
[False  True]
```

S'il y a des frais de ménage, le prix sera plus élevé.

[20]:
```
print(df['city'].unique())
```

```
['LA' 'NYC' 'DC' 'SF' 'Chicago' 'Boston']
```

[21]:
```
print(df['latitude'].unique())
```

```
[33.78271155 40.70546839 38.91753651 … 40.70674885 40.73853473
 33.76109645]
```

[22]:
```
print(df['neighbourhood'].unique())
```

```
['Long Beach' 'Ridgewood' 'U Street Corridor' 'Sunnyside' 'Noe Valley'
 'West Village' 'Harlem' 'Flushing' 'Westside' 'Upper West Side'
 'Shepherd Park' 'Santa Monica' 'Mission District' 'Murray Hill' nan
 'Chinatown' 'Echo Park' 'Hamilton Heights' 'Mar Vista' 'Encino'
 'Kips Bay' 'Williamsburg' 'West Hollywood' 'Carroll Gardens' 'Downtown'
 'Bedford-Stuyvesant' 'Wicker Park' "Hell's Kitchen" 'Upper East Side'
 'Pasadena' 'Shaw' 'Greenpoint' 'Jackson Heights' 'Clinton Hill'
 'Tompkinsville' 'Torrance' 'Beverly Hills' 'Midtown' 'Financial District'
 'Fort Greene' 'Pacific Heights' 'Mid-City' 'Chelsea' 'Venice'
 'Crown Heights' 'South LA' 'Bushwick' 'Parkchester' 'Glendale'
 'Columbia Street Waterfront' 'East Flatbush' 'Western Addition/NOPA'
 'Hollywood' 'East Village' 'Lower East Side' 'Nolita' 'East Elmhurst'
 'Soho' 'The Rockaways' 'Beacon Hill' 'Forest Hills' 'Chevy Chase'
 'Flatbush' 'Lefferts Garden' 'Park Slope' 'East Harlem' 'Compton'
```

```
'Glover Park' 'Cahuenga Pass' 'East Hollywood' 'Trinidad' 'Elmhurst'
'Westwood' 'Astoria' 'Southwest Waterfront'
'Near Northeast/H Street Corridor' 'Mount Vernon Square'
'Downtown/Penn Quarter' 'Back Bay' 'Silver Lake' 'Nob Hill' 'Whittier'
'Mid-Wilshire' 'Richmond District' 'Inglewood' 'Granada Hills North'
'Coney Island' 'Russian Hill' 'Westlake' 'Jamaica Plain'
'Ditmars / Steinway' 'Midtown East' 'Capitol Hill' 'Monterey Park'
'East Boston' 'Westchester/Playa Del Rey' 'Sunset Park' 'Alphabet City'
'Mattapan' 'Cole Valley' 'Charlestown' 'Fenway/Kenmore' 'Gramercy Park'
'South Beach' 'West End' 'The Castro' 'Telegraph Hill' 'Sheepshead Bay'
'Greenwich Village' 'Washington Heights' 'Toluca Lake' 'Bucktown'
'Jamaica' 'Outer Sunset' 'Eastchester' 'Georgetown' '16th Street Heights'
'Little Italy/UIC' 'Morningside Heights' 'South Boston' 'Adams Morgan'
'Cleveland Park' 'Near North Side' 'Roscoe Village' 'Cathedral Heights'
'South Loop/Printers Row' 'Humboldt Park' 'Studio City' 'North End'
'Los Feliz' 'Malibu' 'Ukrainian Village' 'Hyde Park' 'Avondale'
'North Hollywood' 'South Shore' 'Inner Sunset' 'Hermosa Beach' 'Cerritos'
'Eckington' 'Roslindale' 'Lower Haight' 'Prospect Heights' 'South End'
'Burbank' 'Hollywood Hills' 'Arcadia' 'Sherman Oaks' 'Bernal Heights'
'St. Elizabeths' 'Palisades' 'Borough Park' 'Flatiron District'
'Brighton Beach' 'Gowanus' 'Pilsen' 'Manhattan Beach' 'Duboce Triangle'
'West Los Angeles' 'West Hills' 'El Segundo' 'Roxbury' 'Van Nuys'
'Columbia Heights' 'Bloomingdale' 'Park View' 'Midwood' 'Lakeview'
'Logan Circle' 'Pleasant Plains' 'West Adams' 'Brentwood' 'Kensington'
'Arts District' 'Palms' 'Gardena' 'Bayview' 'South Chicago'
'Magnificent Mile' 'Woodridge' 'Haight-Ashbury' 'Redondo Beach' 'Topanga'
'Lawndale' 'Loop' 'Edgewood' 'Dorchester' 'Dogpatch' 'Carson'
'Windsor Terrace' 'Kingsbridge Heights' 'Corona' 'Downtown Brooklyn'
'Irving Park' 'Rogers Park' 'Bronzeville' 'Glen Park' 'South Pasadena'
"Fisherman's Wharf" 'SoMa' 'Hermosa' 'Valley Village' 'Woodside'
'Altadena' 'River North' 'Logan Square' 'Judiciary Square'
'Greenwood Heights' 'Allston-Brighton' 'South Street Seaport'
'Manor Park' 'Marina Del Rey' 'Alhambra' 'Anacostia' 'Alamo Square'
'Woodland Hills/Warner Center' 'Foggy Bottom' 'Twin Peaks'
'Andersonville' 'Del Rey' 'Richmond Hill' 'Tarzana' 'Tottenville'
'Truxton Circle' 'Bel Air/Beverly Crest' 'Congress Heights'
'Mount Pleasant' 'Bayside' 'Lynwood' 'Barney Circle' 'Oakland'
'Bridgeport' 'Reseda' 'Fresh Meadows' 'Morris Heights' 'Twining'
'Boerum Hill' 'Monterey Hills' 'Fairlawn' 'Pacific Palisades'
'San Gabriel' 'West Town/Noble Square' 'Harbor Gateway' 'Canarsie'
'Dupont Circle' 'Crotona' 'West Farms' 'Old Town' 'West Loop/Greektown'
'Brookland' 'Uptown' 'Brooklyn' 'Brooklyn Heights' 'Brooklyn Navy Yard'
'Gravesend' 'Concourse Village' 'Woodley Park' 'Ingleside'
'Times Square/Theatre District' 'Parkside' 'Atwater Village'
'Boyle Heights' 'Laurel Canyon' 'Concourse' 'Fordham' 'Red Hook'
'Kalorama' 'Arboretum' 'Roosevelt Island' 'Lomita' 'Temple City'
'Montecito Heights' 'Hayes Valley' 'Crestwood' 'Mission Hill'
'Cow Hollow' 'Sun Valley' 'Lake Balboa' 'Highland Park'
```

```
'Mount Washington' 'Petworth' 'Lindenwood' 'Takoma' 'Bensonhurst'
'South Robertson' 'West Ridge' 'Kenwood' 'Tribeca' 'Mission Terrace'
'Hawthorne' 'Potrero Hill' 'Castle Hill ' 'Winnetka' 'Valley Glen'
'East New York' 'San Pedro' 'Inwood' 'Kent' 'Cobble Hill'
'Long Island City' 'Brownsville' 'Little Village' 'Wakefield'
'Buena Vista' 'North Cleveland Park' 'Lincoln Park' 'Kingsbridge'
'El Sereno' 'Park Versailles' 'Crocker Amazon' 'Union Square'
'Michigan Park' 'Morrisania' 'La Crescenta-Montrose' 'Noho' 'Culver City'
'Baldwin Hills' 'Lincoln Heights' 'Bay Ridge' 'Duarte' 'Panorama City'
'Forest Hill' 'Port Morris' 'Sunland/Tujunga'
'Central Northeast/Mahaning Heights' 'Middle Village' 'Van Nest'
'Meatpacking District' 'Burleith' 'Hillbrook' 'Benning Ridge' 'Woodland'
'LeDroit Park' 'Baychester' 'Oceanview' 'West Covina' 'Lakeshore'
'University Heights' 'Maspeth' 'Mott Haven' 'Belmont' 'Deanwood' 'Marina'
'North Beach' 'Tenderloin' 'Civic Center' 'Balboa Terrace'
'New Dorp Beach' 'Norwood Park' 'Skid Row' 'Flatlands' 'Northridge'
'West Portal' 'Glassell Park' 'Albany Park' 'North Hills West'
'West Brighton' 'Meiers Corners' 'Allerton' 'Azusa' 'Downtown Crossing'
'Rosemead' 'Canoga Park' 'Gold Coast' 'Edgewater' 'Takoma Park, MD'
'Mt Rainier/Brentwood, MD' 'North Park' 'Palos Verdes'
'Washington Highlands' 'Glendora' 'Fort Davis' 'Bedford Park'
'Colonial Village' 'Lakewood' 'American University Park'
'South Ozone Park' 'Diamond Heights' 'Battery Park City' 'Lamond Riggs'
'Portage Park' 'Hudson Square' 'Silver Spring, MD' 'Streeterville'
'Navy Yard' 'Co-op City' 'West Puente Valley' 'Rego Park' 'West Roxbury'
'City Island' 'Eagle Rock' 'Bellflower' 'Theater District' 'Downey'
'North Center' 'Riverdale' 'Pleasant Hill' 'Dyker Heights' 'Arleta'
'Cypress Park' 'Woodlawn' 'Excelsior' 'Visitacion Valley' 'Kingman Park'
'Monrovia' 'Great Kills' 'Williamsbridge' 'Norwood' 'Morgan Park'
'Portola' 'River West' 'Near Northeast' 'Mission Bay' 'Armour Square'
'South Whittier' 'Elysian Valley' 'Marine Park' 'Woodhaven'
'Harvard Square' 'The Bronx' 'Tremont' 'Sierra Madre' 'Jefferson Park'
'Dunning' 'East San Gabriel' 'Boystown' 'Presidio Heights' 'Howard Beach'
'La Mirada' 'Soundview' 'Beverly' 'Melrose' 'Carver Langston' 'El Monte'
'Bronxdale' 'Rancho Palos Verdes' 'Ozone Park' 'Garfield Park'
'Little Italy' 'Chatsworth' 'McKinley Park' 'Claremont' 'Naylor Gardens'
'Pico Rivera' 'Douglass' 'Montebello' 'Manhattan' 'Lincoln Square'
'Signal Hill' 'Watts' 'Florence-Graham' 'Brightwood' 'Midland Beach'
'Pelham Bay' 'Sylmar' 'Near West Side' 'Eltingville' 'College Point'
'Wrigleyville' 'Hermon' 'St. George' 'Randall Manor' 'Highbridge'
'Belmont Cragin' 'DUMBO' 'East Los Angeles' 'Kew Garden Hills'
'Santa Fe Springs' 'Huguenot' 'Dupont Park' 'Mount Eden' 'Englewood'
'Covina' 'Bergen Beach' 'New Brighton' 'Hunts Point' 'River Terrace'
'Foxhall' 'Benning' 'La Canada Flintridge' 'Hillcrest' 'Garfield Ridge'
'Stronghold' 'Norwalk' 'Westchester Village' 'San Marino' 'Daly City'
'East Corner' 'Presidio' 'Stapleton' 'Back of the Yards' 'Roseland'
'North Michigan Park' 'West Lawn' 'Ivy City' 'Chestnut Hill' 'Japantown'
'Sea Gate' 'Commerce' 'Throgs Neck' 'Fort Lincoln' 'North Lawndale'
```

```
 'Austin' 'Wesley Heights' 'Randle Highlands' 'La Puente' 'Good Hope'
 'Somerville' 'Winthrop' 'Porter Ranch' 'Longwood' 'Berkley'
 'Fort Wadsworth' 'Elm Park' 'Marshall Heights' 'South El Monte'
 'Eastland Gardens' 'Vinegar Hill' 'Whitestone' 'Rosebank' 'Langdon'
 'Sea Cliff' 'Grymes Hill' 'Montclare' 'Marble Hill' 'Bath Beach'
 'Harbor City' 'Bell' 'Spuyten Duyvil' 'Brookline' 'Bellevue' 'Newton'
 'Watertown' 'West Athens' 'Huntington Park' "O'Hare" 'Grasmere' 'Pacoima'
 'Edenwald' 'South San Gabriel' 'Gerritsen Beach' 'Rossville' 'Greenway'
 'Westmont' 'Port Richmond' 'La Habra' 'Leather District' 'Irwindale'
 'Rolling Hills' 'Grant City' 'Baldwin Park' 'South Gate'
 'Hawaiian Gardens' 'North Hills East' 'Friendship Heights' 'Paramount'
 'Bradbury' 'Spring Valley']
```

```
[23]: print(df['zipcode'].unique())
```

```
['90804' '11385' '20009' '11104' '94131' '10014' '10027' '11355' '90064'
 '10024' '20002' '90404' '94110' '10016' '60657' '02111' '90026' '10031'
 '90066' '90405' '10002' '90094' '91316' '11211.0' '90046' '11231.0'
 '94109' '11221' '91601' '60622' '11249.0' '10019' '10065' '60642' '91105'
 '20001' '11222' '11370.0' '11378' '11205.0' '11233' '10128' '10301'
 '90278' '11206.0' '90048' '10026' '10038' '11205' '90019' '90403'
 '10001.0' '90291' '11225.0' '90015' '11237' '90212' '94108' '10462'
 '91207' '11216' '11212.0' '94117' '90028' '60608' '90815' '11213.0'
 '10003.0' '11207' '10012' '11369.0' '10021' '11216.0' '11693' '10018'
 '02114' '90006' '20008' '20015' '11226' '11225' '11215' '10029.0' '90222'
 '20007' '90068' '90029' '11238.0' '11377' '91101' '90024' '11106' '20024'
 '90016' '11103' '90005' '20005' '02116' '10028' '10025' '91302' '91364'
 '20032' '90039' '11206' '94133' '11237.0' '90601' '60601' '90036' '94118'
 '91766' '10011.0' '90301' '94103' '91344' '94115' '11224' '10075' '20011'
 '90021' '02130' '11105' '90065' '90049' '10022' '20003' '90069' '91754'
 '02128' '90293' '11220' '10001' '10009.0' '02126' '11372' '10010'
 '10002.0' '10023' '02129' '02215' '60626' '10003' '90604' '10035.0'
 '94105' '90249' '94114' '11235' '91790' '10011' '10040' '10036' '91602'
 '11238' '90035' '11422' '94122' '10032' '10469' '60612' '11232' '90014'
 '02210' '90018' '60610' '10013.0' '60618' '60616' '60647' '91423' '02113'
 '02108' '90027' '90265' '60637' '11692' '94121' '91605' '60649' '90254'
 '90703' '11102' '02131' '11101' '02118' '91205' '90023' '10017' '91506'
 '91007' '20010' '20016' '11204' '90025' '91307' '90245' '90803' '02119'
 '91405' '10013' '11230' '02127' '91748' '91107' '91745' '11233.0' '60613'
 '90007' '11211' '11218' '90012' '90034' '10039' '94124' '60617' '90038'
 '60611' '90277' '20018' '90290' '60625' '90260' '60605' '90502' '60641'
 '02125' '90247' '10004' '94107' '90746' '11217' '10463' '91010' '11368'
 '60654' '90808' '90004' '60653' '91301' '90802' '91030' '91502' '90020'
 '90302' '94123' '90650' '91776' '91006' '02124' '91001' '90814' '02134'
 '94115.0' '60607' '90813' '90292' '91803' '20020' '11375' '20037' '60640'
 '11222.0' '11418' '91356' '10307.0' '91403' '90810' '90077' '91767'
 '11364' '94102' '10037.0' '90262' '60609' '91335' '11365' '10453' '90266'
 '10033' '91202' '20019' '90013' '11210.0' '11412' '11373.0' '94117.0'
 '90505' '90032' '11207.0' '91765' '91204' '10027.0' '90272' '90402'
 '90501' '11236.0' '91604' '11229' '91606' '10459' '91367' '10460' '91436'
 '10037' '11203.0' '91106' '60661' '20017' '91789' '91201' '02109' '10304'
 '11201' '10030' '11223' '20012' '10451.0' '02135' '94112' '91321' '94116'
 '90043' '11379' '90033' '90713' '94127' '10468' '11231' '90062' '90232'
 '90057' '11432' '11434' '10044' '91504' '90717' '91780' '90031' '10009'
 '20036' '11370' '94109.0' '90045' '02115' '11691' '91755' '91104' '91352'
 '91325' '11249' '90042' '91411' '90805' '90303' '11414' '90250' '11236'
 '60633' '91505' '90210' '60645' '60615' '90503' '11219' '11226.0' '11203'
 '10473' '91203' '91306' '91401' '90732' '10034' '91801' '90017' '11214'
 '90220' '11201.0' '11212' '60631' '60623' '10466' '91746' '11435' '60614'
 '10472' '90008' '10006' '91387' '10456.0' '94130' '10005' '20268' '91214'
 '90806' '91711' '11210' '91208' '02121' '11209' '94102.0' '91331' '91792'
 '90061' '10451' '91040' '90230' '91311' '91355' '02120' '11213' '94132'
 '91042' '02199' '91791' '60602' '10456' '10458' '90504' '10038.0' '91501'
 '10306' '10035' '11109' '11234' '90056' '91343' '91304' '20006' '91206'
 '7302.0' '10314.0' '10467' '91702' '91770' '90731' '91607' '60660'
 '20912' '20712' '90274' '90401' '91354' '91741' '60805' '02110' '91361'
 '11420' '11368.0' '10280' '20052' '60634' '11208.0' '20910' '10007.0'
 '10007' '11224.0' '10475.0' '11374' '02132' '10464' '90041' '94111'
 '90706' '91103' '90242' '94104.0' '94104' '10459.0' '91390' '10014.0'
 '02136' '91324' '10471.0' '90211' '11228.0' '90280' '93550' '11433'
 '91768' '90241' '94134' '10475' '91786' '11358' '91016' '90010' '10308'
 '11416' '10307' '94114.0' '60643' '02122' '94158' '91108' '90605' '90715'
 '10279' '11234.0' '90807' '11421' '02138' '10704' '10457' '10305' '91024'
 '10452.0' '60630' '90067' '60651' '91775' '90631' '91802' '10310' '90712'
 '90638' '10473.0' '91406' '90044' '90403-2638' '93563' '91377' '91731'
 '90275' '90606' '60624' '91732' '90660' '91303' '90640' '11417' '90037'
 '1m' '90602' '90755' '90059' '90001' '10461' '91342' '10312' '11356.0'
 '91708' '11354' '90304' '10282' '91326' '91351' '91773' '10452' '90305'
 '60659' '90063' '10069' '10018.0' '11361' '11367' '90670' '91733'
 '10457.0' '60621' '10454' '91722' '90022' '94118.0' '90704' '91723'
 '11209.0' '10474' '10453.0' '91011' '11411.0' '94014' '11360' '90248'
 '60638' '90221' '90716' '90034-2203' '60639' '94014.0' '90035-4475'
 '94129' '60636' '10304.0' '11694' '60619' '11429' '91724' '11220.0'
 '11423' '20004' '60629' '91381' '11509.0' '93536' '02467' '90040' '10465'
 '11415' '60606' '60644' '93551' '91744' '11412.0' '11413' '02145' '11001'
 '10010.0' '60302' '11419' '02152' '91020' '93534' '10455' '91340'
 '11362.0' '11363.0' '11372.0' '10303' '11411' '60660-1448' '91750'
 '11221.0' '10471' '90745' '11357' '20816' '11373' '90036-2514' '10026.0'
 '11429.0' '60603' '10308.0' '94401' '60707' '11228' '91210' '90710'
 '90201' '91402' '02186' '02445' '10162' '02458' '02472' '90047' '90255'
 '11426' '11215.0' '60656' '90240' '10004.0' '10305.0' '91384' '90744'
 '93535' '90003' '10128.0' '93543' '11428' '10309' '10463.0' '10270'
 '9004' '10282.0' '11239.0' '93552' '91706' '90011' '90723' '91362'
 '91008' '11436' '10012.0']
```

```
[24]: print(df['bedrooms'].unique())
```

```
[ 0.  1.  2.  5.  3.  8.  4. nan  7. 10.  6.  9.]
```

```
[25]: print(df['beds'].unique())
```

```
[ 2.  1.  3.  8. 10.  4.  5.  6. nan  9.  7. 13. 16.  0. 12. 11. 18.]
```

2.0.3 3. Les corrélations entre le prix et des caractéristiques

Quelques statistiques (moyenne, médiane, écart-type, etc.) :

```
[26]: df.describe()
```

```
[26]:                  id     log_price   accommodates     bathrooms      latitude  \
      count  2.193100e+04  21931.000000  21931.000000  21880.000000  21931.000000
      mean   1.122766e+07      4.783193      3.154348      1.235923     38.463441
      std    6.077360e+06      0.718268      2.141682      0.586293      3.072646
      min    3.362000e+03      2.302585      1.000000      0.000000     33.339002
      25%    6.224753e+06      4.317488      2.000000      1.000000     34.135374
      50%    1.218496e+07      4.700480      2.000000      1.000000     40.662703
      75%    1.639510e+07      5.220356      4.000000      1.000000     40.746395
      max    2.120450e+07      7.600402     16.000000      8.000000     42.390248

                longitude      bedrooms          beds
      count  21931.000000  21907.000000  21900.000000
      mean     -92.267719      1.263477      1.710502
      std       21.669362      0.850969      1.251887
      min     -122.510940      0.000000      0.000000
      25%     -118.340398      1.000000      1.000000
      50%      -76.994992      1.000000      1.000000
      75%      -73.954573      1.000000      2.000000
      max      -70.989359     10.000000     18.000000
```

Conversion des variables catégorielles du dataset en variables indicatrices binaires :

```
[27]: categories = ['property_type', 'room_type', 'accommodates', 'bathrooms',␣
       ↪'bed_type', 'cancellation_policy','cleaning_fee', 'city', 'neighbourhood',␣
       ↪'zipcode', 'bedrooms', 'beds']
      df_bin = pd.get_dummies(df, columns=categories)
```

Voici les variables converties en variables indicatrices binaires, qui à mon avis ont des impacts significatifs sur le prix :

- property_type : différents types de logements (appartement, maison…)
- room_type : indiquant si l'espace loué est un logement entier, une chambre privée ou partagée…
- accommodates et bathrooms pour voir

- bed_type : le type de lit peut aussi influencer les préférences des locataires et donc le prix
- cancellation_policy : peut affecter la flexibilité et la demande
- cleaning_fee pour voir
- city : la localisation géographique (ville) est un facteur pour les prix de location
- neighbourhood : certains quartiers peuvent être plus recherchés que d'autres
- zipcode, bedrooms et beds pour voir
- amenities : nécessite un traitement pour extraire et transformer chaque équipement en une variable binaire, puisqu'il s'agit d'une liste d'éléments

Les variables 'id', 'log_price', 'latitude', 'longitude' n'étaient à mon avis pas nécessaires à convertir en variables dummy.

[28]:
```
df_bin.head()
```

[28]:
```
         id  log_price                                          amenities  \
0   5708593   4.317488  {TV,"Wireless Internet",Kitchen,"Free parking …
1  14483613   4.007333  {"Wireless Internet","Air conditioning",Kitche…
2  10412649   7.090077  {TV,"Wireless Internet","Air conditioning",Kit…
3  17954362   3.555348  {TV,"Cable TV",Internet,"Wireless Internet","A…
4   9969781   5.480639  {TV,"Cable TV",Internet,"Wireless Internet",Ki…

    latitude   longitude  property_type_Apartment  \
0  33.782712 -118.134410                    False
1  40.705468  -73.909439                    False
2  38.917537  -77.031651                     True
3  40.736001  -73.924248                    False
4  37.744896 -122.430665                    False

   property_type_Bed & Breakfast  property_type_Boat  \
0                          False               False
1                          False               False
2                          False               False
3                          False               False
4                          False               False

   property_type_Boutique hotel  property_type_Bungalow  …  beds_6.0  \
0                         False                   False  …     False
1                         False                   False  …     False
2                         False                   False  …     False
3                         False                   False  …     False
4                         False                   False  …     False

   beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  beds_12.0  beds_13.0  \
0     False     False     False      False      False      False      False
1     False     False     False      False      False      False      False
```

```
2     False     False     False      False      False      False      False
3     False     False     False      False      False      False      False
4     False     False     False      False      False      False      False

   beds_16.0  beds_18.0
0       False      False
1       False      False
2       False      False
3       False      False
4       False      False

[5 rows x 1348 columns]
```

Analyse de la répartition des données par catégorie pour les variables catégorielles

- Pour property type :

[29]:
```python
# Pour voir le log_price en fonction du property_type :
log_price_en_fct_property_type = df_bin.groupby([col for col in df_bin.columns␣
 ↪if 'property_type' in col])['log_price'].mean()

print(log_price_en_fct_property_type)
```

```
property_type_Apartment  property_type_Bed & Breakfast  property_type_Boat
property_type_Boutique hotel  property_type_Bungalow  property_type_Cabin
property_type_Camper/RV  property_type_Castle  property_type_Cave
property_type_Chalet  property_type_Condominium  property_type_Dorm
property_type_Earth House  property_type_Guest suite  property_type_Guesthouse
property_type_Hostel  property_type_House  property_type_Hut  property_type_In-
law  property_type_Island  property_type_Loft  property_type_Other
property_type_Serviced apartment  property_type_Tent  property_type_Timeshare
property_type_Tipi  property_type_Townhouse  property_type_Treehouse
property_type_Vacation home  property_type_Villa  property_type_Yurt
False                    False                          False
False                         False                   False                False
False                 False               False                 False
False               False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False           False
False                    False                        False                True
5.289808
                                                    True                 False
4.977980
                       True                         False                False
5.531722
True                     False                        False                False
5.978886
```

```
                                                     True
False                    False                        False                False
4.822138
                                 True                False
False                    False                        False                False
5.220356
        True                     False               False
False                    False                        False                False
5.585600
                                                                    True
False                    False               False                    False
False                        False                False                4.008937
                                  True                              False
False                    False               False                    False
False                        False                False                5.255823
             True                 False                             False
False                    False               False                    False
False                        False                False                4.946918
                                                                         True
False                False                                  False               False
False               False                    False                    False
False                False                 4.994079
                                                   True                  False
False                False                                  False               False
False               False                    False                    False
False                False                 5.010635
                             True                  False                 False
False                False                                  False               False
False               False                    False                    False
False                False                 4.763759
          True               False                 False                 False
False                False                                  False               False
False               False                    False                    False
False                False                 4.291634
                                                                     True
False              False                 False                 False
False                False                                  False               False
False               False                    False                    False
False                False                 4.789230
                                               True                  False
False              False                 False                 False
False                False                                  False               False
False               False                    False                    False
False                False                 3.771639
                     True                      False                 False
False              False                 False                 False
False                False                                  False               False
False               False                    False                    False
False                False                4.678250
                                                                          True
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                4.655889
                                               True                       False
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                4.174387
                           True                False                      False
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                3.588597
True                       False               False                      False
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                5.041489
                                                          True
False                      False               False                      False
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                4.618083
                                      True                False
False                      False               False                      False
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                4.990433
                True                  False               False
False                      False               False                      False
False                     False                 False                False
False                 False                 False               False
False                             False               False
False               False                    False                    False
False                False                5.356599
                                                                       True
False                 False               False                 False
False               False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False               False
False                    False                        False                False
4.547690
                                                  True                 False
False                 False               False                 False
False               False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False               False
False                    False                        False                False
4.755987
                          True                    False                False
False                 False               False                 False
False               False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False               False
False                    False                        False                False
4.823562
                                                                            True
False                   False                False                    False
False               False                 False                      False
False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False               False
False                    False                        False                False
5.113796
                                                        True
False                         False                   False                False
False                 False               False                 False
False               False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False               False
False                    False                        False                False
5.123347
                         True                              False
False                         False                   False                False
False                 False               False                 False
False               False                      False                      False
False                 False                False              False
False                 False               False                False
False               False                    False               False
False                    False                        False                False
```

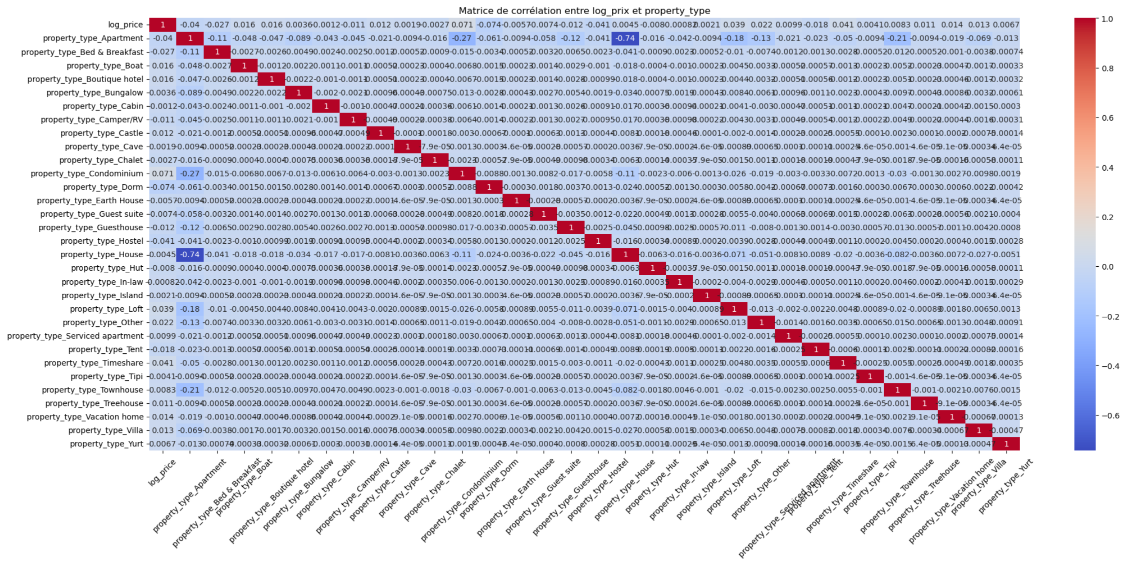
```
4.532817
True                         False                                  False
False                             False                        False                     False
False                     False                   False                     False
False                   False                          False                           False
False                     False                    False                   False
False                     False                   False                    False
False                   False                        False                   False
False                        False                             False                     False
4.762265
Name: log_price, dtype: float64
```

[30]:
```python
# Corrélation forte entre le prix de l'appartement et le property_type ?
# Calcul de la matrice de corrélation
matrice_corr_property_type= df_bin[["log_price"] + [col for col in df_bin.
    ↪columns if "property_type" in col]].corr()
```

[31]:
```python
# heatmap
plt.figure(figsize=(25, 10))
sns.heatmap(matrice_corr_property_type, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et property_type")

plt.show()
```



Je ne remarque pas de grande influence sur le log_prix

- Pour room_type :

[32]:
```
# Pour voir le log_price en fonction du room_type :
log_price_en_fct_room_type = df_bin.groupby([col for col in df_bin.columns if␣
 ↪'room_type' in col])['log_price'].mean()

print(log_price_en_fct_room_type)
```

```
room_type_Entire home/apt  room_type_Private room  room_type_Shared room
False                      False                   True
3.881240
                           True                    False
4.334213
True                       False                   False
5.167591
Name: log_price, dtype: float64
```

[33]:
```
room_type_cols = [col for col in df_bin.columns if 'room_type' in col]
print(room_type_cols)
```

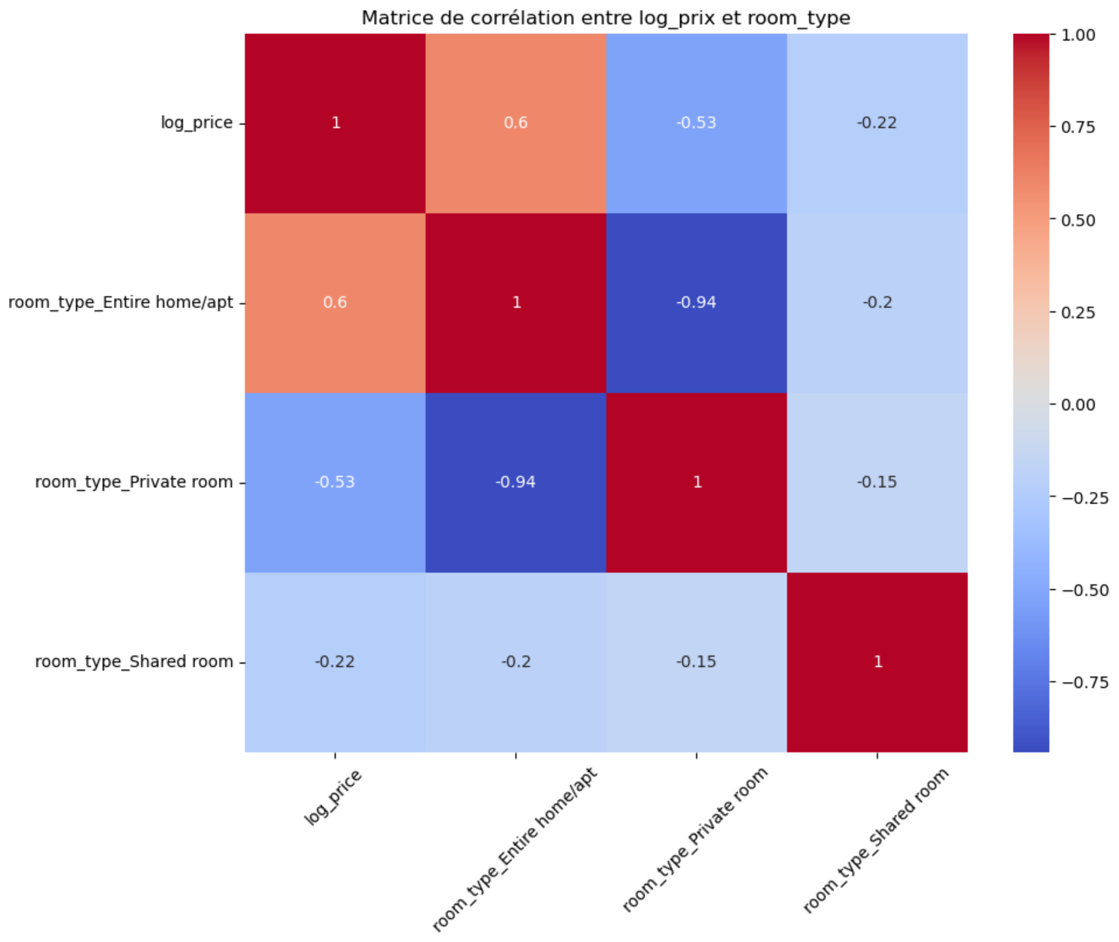
```
['room_type_Entire home/apt', 'room_type_Private room', 'room_type_Shared room']
```

[34]:
```
# Corrélation forte entre le prix de l'appartement et le room_type ?
# Calcul de la matrice de corrélation
matrice_corr_room_type= df_bin[["log_price"] + [col for col in df_bin.columns␣
 ↪if "room_type" in col]].corr()
```

[35]:
```
# heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(matrice_corr_room_type, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et room_type")

plt.show()
```



Je remarque une influence sur le prix

- Pour accommodates :

[36]:
```python
# Pour voir le log_price en fonction du nombre d'accommodates :
log_price_en_fct_accommodates = df_bin.groupby([col for col in df_bin.columns␣
↪if 'accommodates' in col])['log_price'].mean()

print(log_price_en_fct_accommodates)
```

```
accommodates_1  accommodates_2  accommodates_3  accommodates_4  accommodates_5
accommodates_6  accommodates_7  accommodates_8  accommodates_9  accommodates_10
accommodates_11  accommodates_12  accommodates_13  accommodates_14
accommodates_15  accommodates_16
False           False           False           False           False
False           False           False           False           False
False            False            False            False            False
True               6.106292
                                                                   True
```

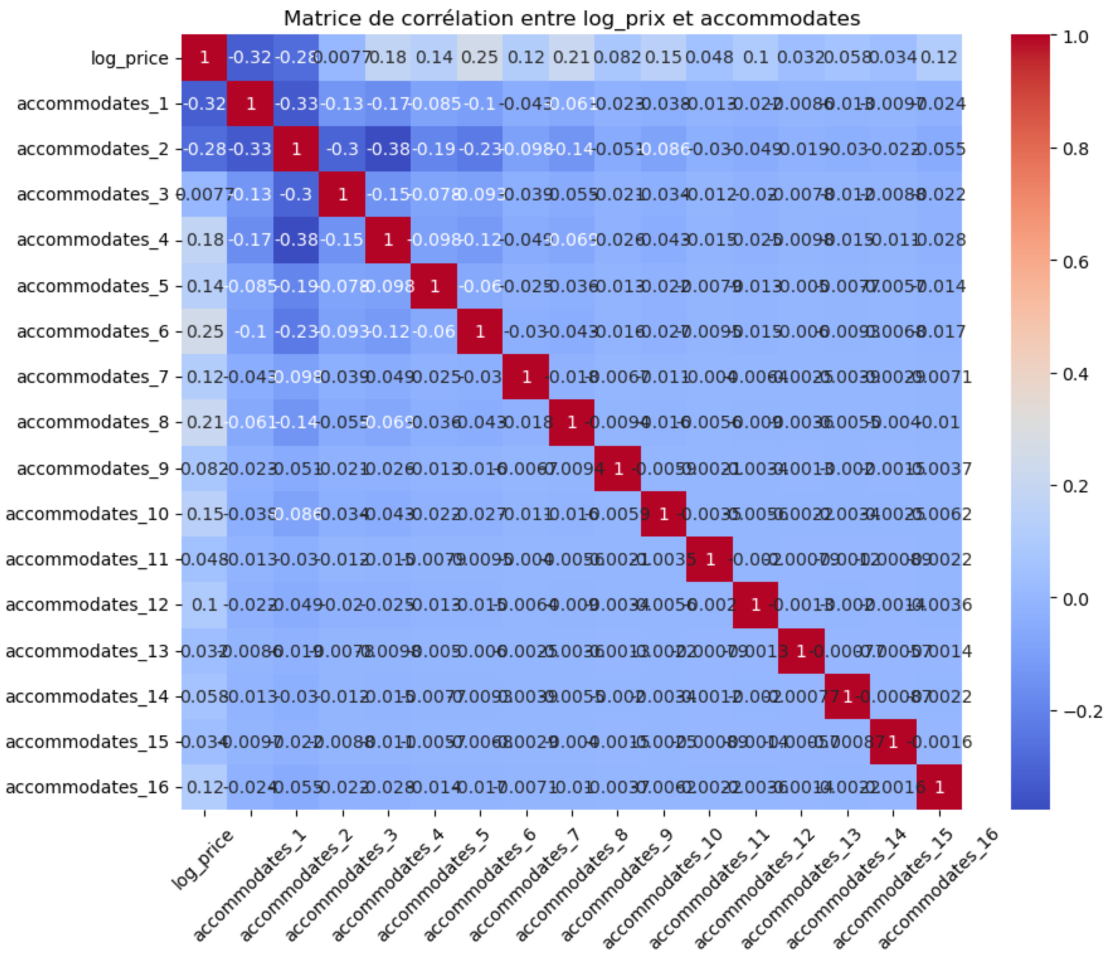
```
False              5.747441
                                                    True             False
False              5.983674
                                   True             False            False
False              5.820231
                  True             False            False            False
False              6.090033
 True             False            False            False            False
False              5.765435
                                                                True
False            False            False            False            False
False              5.848890
                                                True            False
False            False            False            False            False
False              5.773542
                                True            False           False
False            False            False            False            False
False              5.729799
                True            False           False           False
False            False            False            False            False
False              5.532838
True            False           False           False           False
False            False            False            False            False
False              5.451785
                                                                True
False           False           False           False           False
False            False            False            False            False
False              5.223210
                                                True            False
False           False           False           False           False
False            False            False            False            False
False              5.079303
                                True            False           False
False           False           False           False           False
False            False            False            False            False
False              4.799170
                True            False           False           False
False           False           False           False           False
False            False            False            False            False
False              4.554798
True            False           False           False           False
False           False           False           False           False
False            False            False            False            False
False              4.175041
Name: log_price, dtype: float64
```

```
[37]: # Corrélation forte entre le prix de l'appartement et accommodates ?
      # Calcul de la matrice de corrélation
      matrice_corr_accommodates= df_bin[["log_price"] + [col for col in df_bin.
       ↪columns if "accommodates" in col]].corr()
```

```
[38]: # heatmap
      plt.figure(figsize=(10, 8))
      sns.heatmap(matrice_corr_accommodates, annot=True, cmap="coolwarm")

      plt.xticks(rotation=45)
      plt.yticks(rotation=0)
      plt.title("Matrice de corrélation entre log_prix et accommodates")

      plt.show()
```



Je ne remarque pas de grande influence

- Pour bathrooms :

```
[39]: # Pour voir le log_price en fonction des bathrooms :
log_price_en_fct_bathrooms = df_bin.groupby([col for col in df_bin.columns if␣
 ↪'bathrooms' in col])['log_price'].mean()

print(log_price_en_fct_bathrooms)
```

```
bathrooms_0.0  bathrooms_0.5  bathrooms_1.0  bathrooms_1.5  bathrooms_2.0
bathrooms_2.5  bathrooms_3.0  bathrooms_3.5  bathrooms_4.0  bathrooms_4.5
bathrooms_5.0  bathrooms_5.5  bathrooms_6.0  bathrooms_6.5  bathrooms_7.0
bathrooms_7.5  bathrooms_8.0
False          False          False          False          False          False
False          False          False          False          False          False
False          False          False          False          False
4.661688
True             5.176763
                                                                 True
False            7.455810
                                                  True           False
False            6.795084
                                   True           False          False
False            6.498254
                    True           False          False          False
False            6.303097
     True           False          False          False          False
False            6.730003
                                                                      True
False          False          False          False          False          False
6.185231
                                                       True           False
False          False          False          False          False          False
6.443918
                                        True           False          False
False          False          False          False          False          False
5.704211
                         True           False          False          False
False          False          False          False          False          False
6.075479
          True           False          False          False          False
False          False          False          False          False          False
5.576488
                                                                           True
False          False          False          False          False          False
False          False          False          False          False
5.508918
                                                            True           False
False          False          False          False          False          False
False          False          False          False          False
```

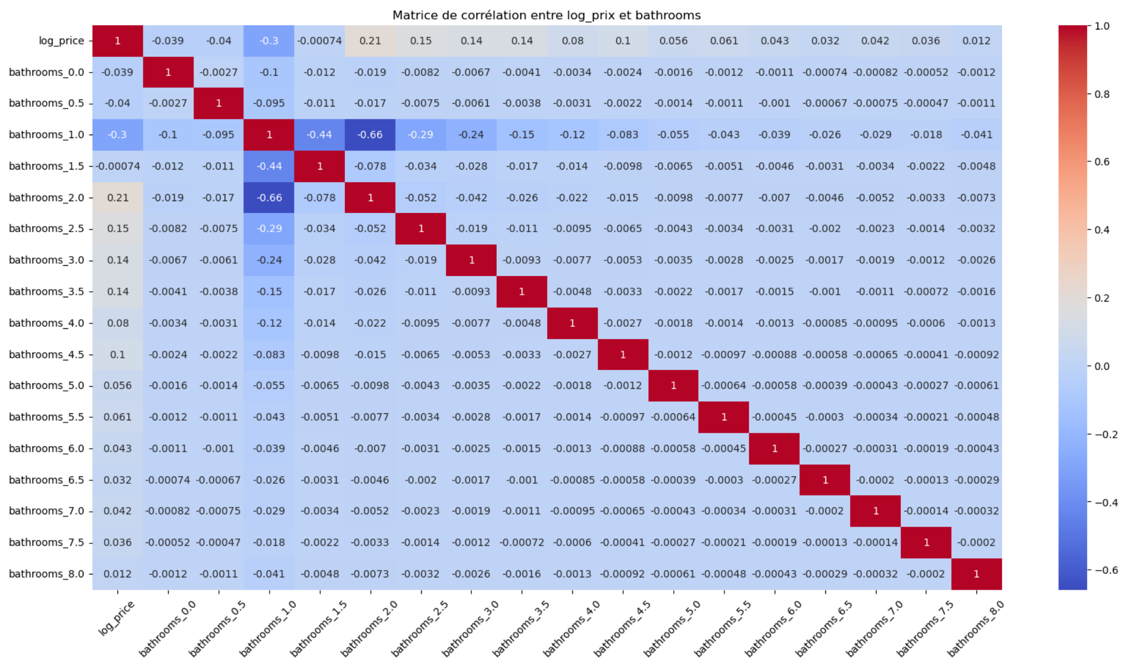
```
5.212290
                                               True           False          False
False          False          False          False          False          False
False          False          False          False          False
4.780835
                              True           False          False          False
False          False          False          False          False          False
False          False          False          False          False
4.671342
               True           False          False          False          False
False          False          False          False          False          False
False          False          False          False          False
4.198718
True           False          False          False          False          False
False          False          False          False          False          False
False          False          False          False          False
4.266018
Name: log_price, dtype: float64
```

[40]:
```python
# Corrélation forte entre le prix et bathrooms ?
# Calcul de la matrice de corrélation
matrice_corr_bathrooms= df_bin[["log_price"] + [col for col in df_bin.columns␣
 ↪if "bathrooms" in col]].corr()
```

[41]:
```python
# heatmap
plt.figure(figsize=(20, 10))
sns.heatmap(matrice_corr_bathrooms, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et bathrooms")

plt.show()
```



Je ne remarque pas non plus de grande influence

- Pour le bed_type :

[42]:
```
# Pour voir le log_price en fonction du bed_type :
log_price_en_fct_bed_type = df_bin.groupby([col for col in df_bin.columns if␣
 ↪'bed_type' in col])['log_price'].mean()

print(log_price_en_fct_bed_type)
```

```
bed_type_Airbed  bed_type_Couch  bed_type_Futon  bed_type_Pull-out Sofa
bed_type_Real Bed
False            False           False           False                  True
4.795401
                                                 True                   False
4.464975
                                 True            False                  False
4.287998
                 True            False           False                  False
4.334119
True             False           False           False                  False
4.299140
Name: log_price, dtype: float64
```

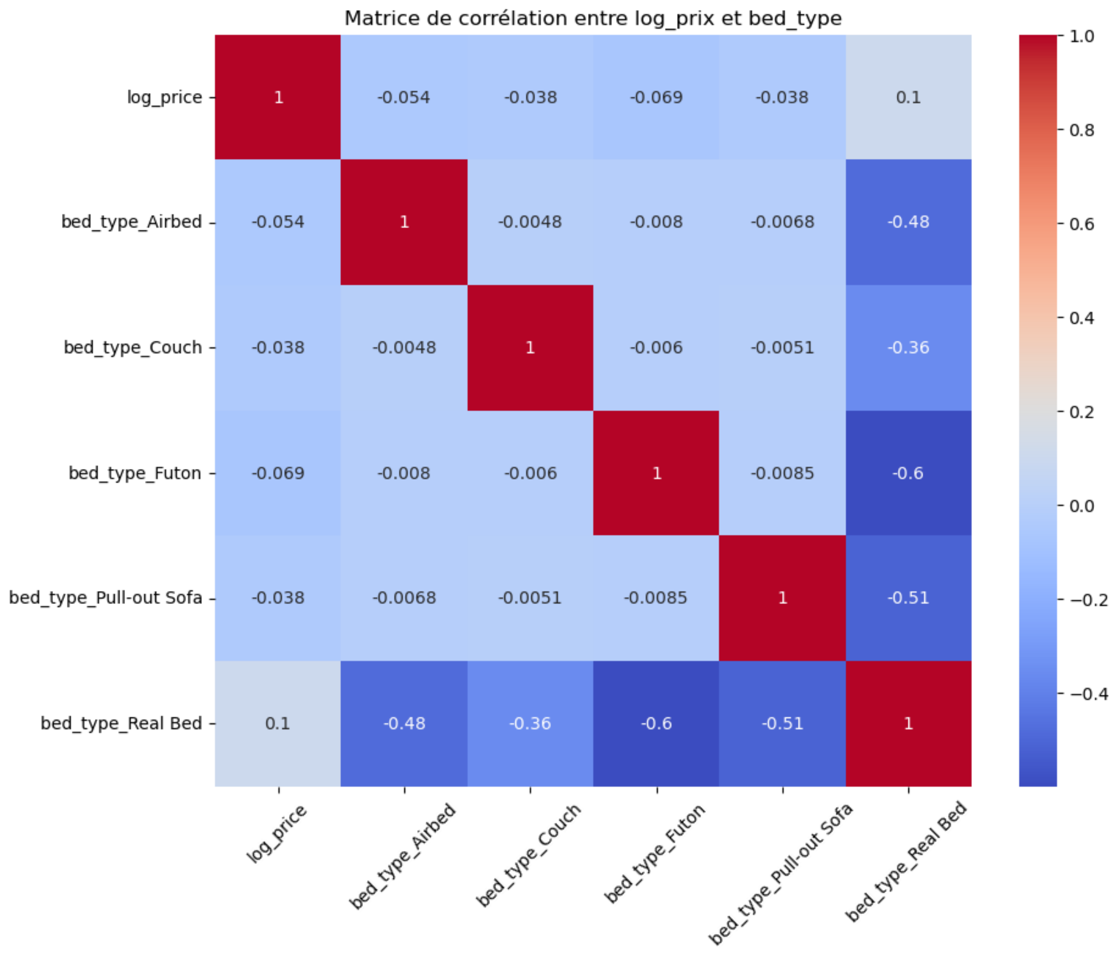
[43]:
```
# Corrélation forte entre le prix de l'appartement et le bed_type ?
# Calcul de la matrice de corrélation
```

```
matrice_corr_bed_type= df_bin[["log_price"] + [col for col in df_bin.columns if␣
↪"bed_type" in col]].corr()
```

```
[44]: # heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(matrice_corr_bed_type, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et bed_type")

plt.show()
```



Je ne remarque pas d'influence majeure non plus

- Pour le cancellation_policy :

[45]:
```
# Pour voir le log_price en fonction du cancellation_policy :
log_price_en_fct_cancellation_policy = df_bin.groupby([col for col in df_bin.
 ↪columns if 'cancellation_policy' in col])['log_price'].mean()

print(log_price_en_fct_cancellation_policy)
```

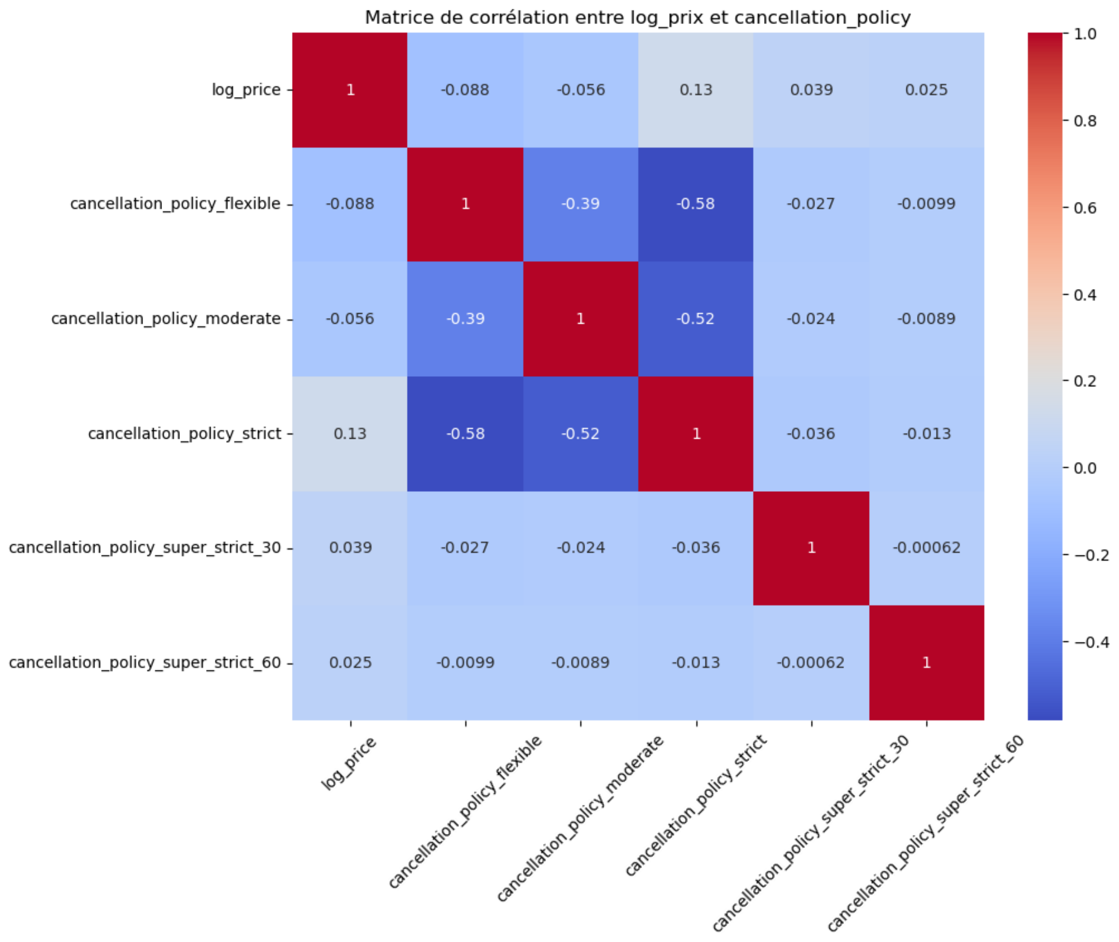
```
cancellation_policy_flexible  cancellation_policy_moderate
cancellation_policy_strict  cancellation_policy_super_strict_30
cancellation_policy_super_strict_60
False                         False                         False
False                                True
5.992353
        True                                 False
5.471440
                                                            True
False                                False
4.886028
                              True                          False
False                                False
4.715328
True                          False                         False
False                                False
4.687225
Name: log_price, dtype: float64
```

[46]:
```
# Corrélation forte entre le prix de l'appartement et le cancellation_policy ?
# Calcul de la matrice de corrélation
matrice_corr_cancellation_policy= df_bin[["log_price"] + [col for col in df_bin.
 ↪columns if "cancellation_policy" in col]].corr()
```

[47]:
```
# heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(matrice_corr_cancellation_policy, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et cancellation_policy")

plt.show()
```



	log_price	cancellation_policy_flexible	cancellation_policy_moderate	cancellation_policy_strict	cancellation_policy_super_strict_30	cancellation_policy_super_strict_60
log_price	1	-0.088	-0.056	0.13	0.039	0.025
cancellation_policy_flexible	-0.088	1	-0.39	-0.58	-0.027	-0.0099
cancellation_policy_moderate	-0.056	-0.39	1	-0.52	-0.024	-0.0089
cancellation_policy_strict	0.13	-0.58	-0.52	1	-0.036	-0.013
cancellation_policy_super_strict_30	0.039	-0.027	-0.024	-0.036	1	-0.00062
cancellation_policy_super_strict_60	0.025	-0.0099	-0.0089	-0.013	-0.00062	1

Ce n’est pas très utile non plus

- Pour le cleaning_fee :

[48]:
```
# Pour voir le log_price en fonction du cleaning_fee :
log_price_en_fct_cleaning_fee = df_bin.groupby([col for col in df_bin.columns␣
  ↪if 'cleaning_fee' in col])['log_price'].mean()

print(log_price_en_fct_cleaning_fee)
```

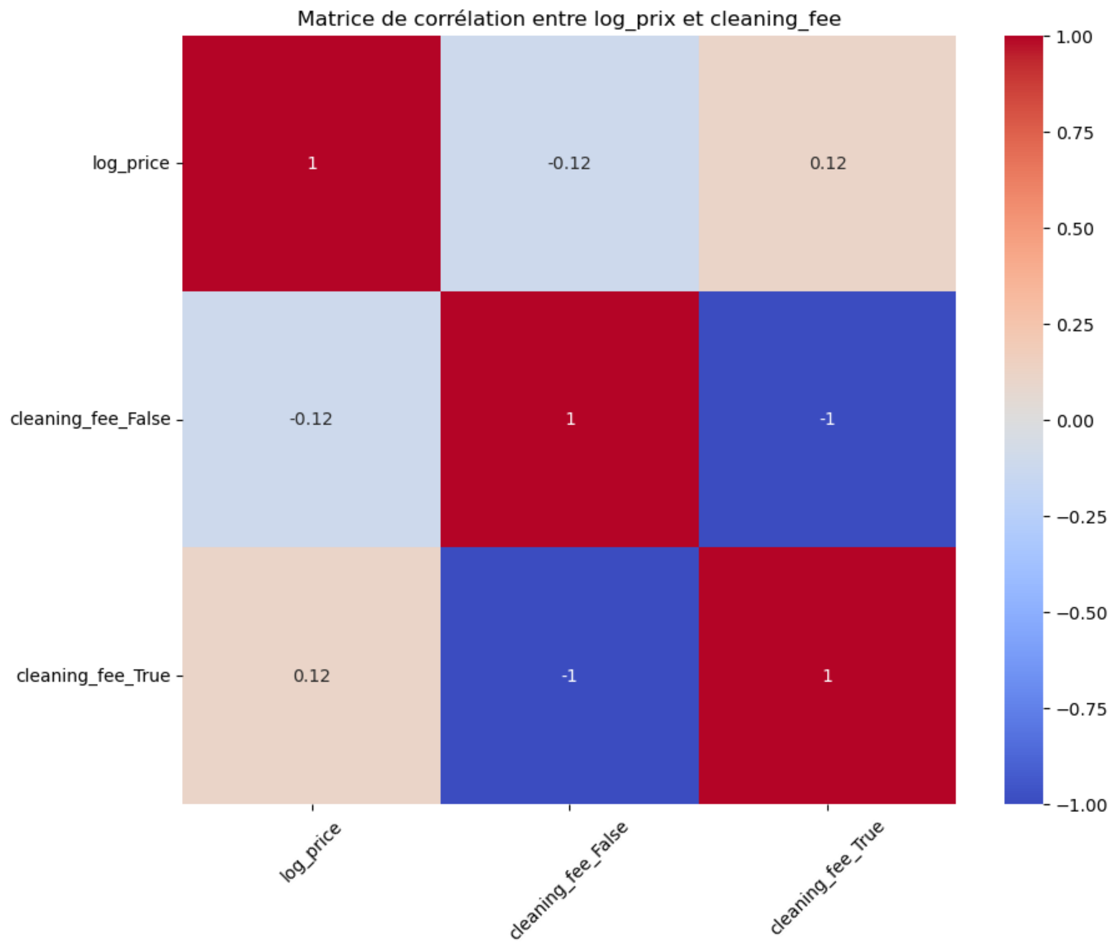
```
cleaning_fee_False  cleaning_fee_True
False               True                 4.833382
True                False                4.642555
Name: log_price, dtype: float64
```

[49]:
```
# Corrélation forte entre le prix de l’appartement et le cleaning_fee ?
# Calcul de la matrice de corrélation
matrice_corr_cleaning_fee= df_bin[["log_price"] + [col for col in df_bin.
  ↪columns if "cleaning_fee" in col]].corr()
```

```
[50]: # heatmap
      plt.figure(figsize=(10, 8))
      sns.heatmap(matrice_corr_cleaning_fee, annot=True, cmap="coolwarm")

      plt.xticks(rotation=45)
      plt.yticks(rotation=0)
      plt.title("Matrice de corrélation entre log_prix et cleaning_fee")

      plt.show()
```



C'est assez intéressant à garder

- Pour la ville :

```
[51]: # Pour voir le log_price en fonction de la ville :
      log_price_en_fct_ville = df_bin.groupby([col for col in df_bin.columns if␣
       ↪'city' in col])['log_price'].mean()
```

```
print(log_price_en_fct_ville)
```

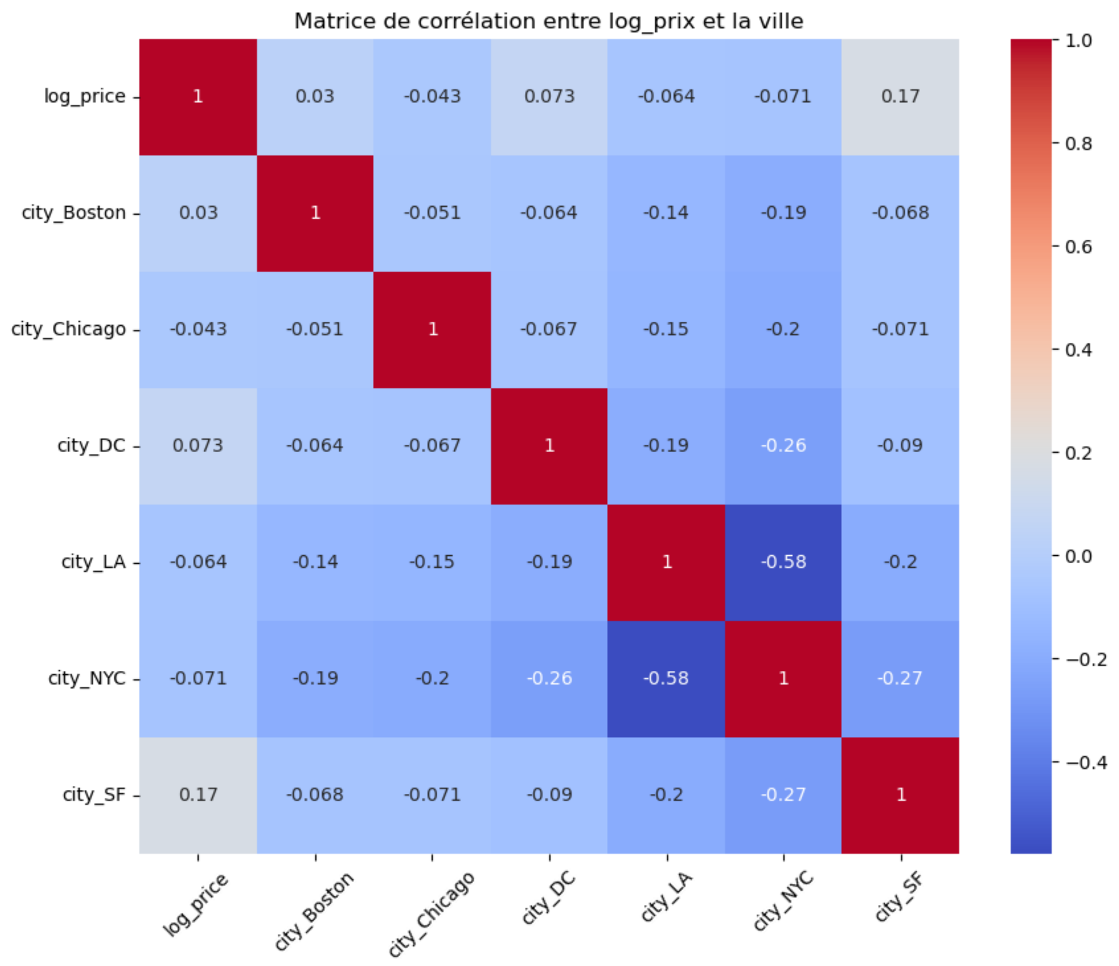
```
city_Boston  city_Chicago  city_DC  city_LA  city_NYC  city_SF
False        False         False    False    False     True       5.183200
                                             True      False      4.724982
                                    True     False     False      4.713099
                           True     False    False     False      4.964046
             True          False    False    False     False      4.650405
True         False         False    False    False     False      4.879681
Name: log_price, dtype: float64
```

[52]:
```
# Corrélation forte entre le prix de l'appartement et la ville ?
# Calcul de la matrice de corrélation
matrice_corr_ville= df_bin[["log_price"] + [col for col in df_bin.columns if␣
 ↪"city" in col]].corr()
```

[53]:
```
#heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(matrice_corr_ville, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et la ville")

plt.show()
```



	log_price	city_Boston	city_Chicago	city_DC	city_LA	city_NYC	city_SF
log_price	1	0.03	-0.043	0.073	-0.064	-0.071	0.17
city_Boston	0.03	1	-0.051	-0.064	-0.14	-0.19	-0.068
city_Chicago	-0.043	-0.051	1	-0.067	-0.15	-0.2	-0.071
city_DC	0.073	-0.064	-0.067	1	-0.19	-0.26	-0.09
city_LA	-0.064	-0.14	-0.15	-0.19	1	-0.58	-0.2
city_NYC	-0.071	-0.19	-0.2	-0.26	-0.58	1	-0.27
city_SF	0.17	-0.068	-0.071	-0.09	-0.2	-0.27	1

On remarque une légère influence

- Pour le voisinage :

[54]:
```
# Pour voir le log_price en fonction du voisinage :
log_price_en_fct_neighbourhood = df_bin.groupby([col for col in df_bin.columns␣
↪if 'neighbourhood' in col])['log_price'].mean()

print(log_price_en_fct_neighbourhood)
```

```
neighbourhood_16th Street Heights  neighbourhood_Adams Morgan
neighbourhood_Alamo Square  neighbourhood_Albany Park  neighbourhood_Alhambra
neighbourhood_Allerton  neighbourhood_Allston-Brighton  neighbourhood_Alphabet
City  neighbourhood_Altadena  neighbourhood_American University Park
neighbourhood_Anacostia  neighbourhood_Andersonville  neighbourhood_Arboretum
neighbourhood_Arcadia  neighbourhood_Arleta  neighbourhood_Armour Square
neighbourhood_Arts District  neighbourhood_Astoria  neighbourhood_Atwater
Village  neighbourhood_Austin  neighbourhood_Avondale  neighbourhood_Azusa
```

neighbourhood_Back Bay neighbourhood_Back of the Yards neighbourhood_Balboa Terrace neighbourhood_Baldwin Hills neighbourhood_Baldwin Park neighbourhood_Barney Circle neighbourhood_Bath Beach neighbourhood_Battery Park City neighbourhood_Bay Ridge neighbourhood_Baychester neighbourhood_Bayside neighbourhood_Bayview neighbourhood_Beacon Hill neighbourhood_Bedford Park neighbourhood_Bedford-Stuyvesant neighbourhood_Bel Air/Beverly Crest neighbourhood_Bell neighbourhood_Bellevue neighbourhood_Bellflower neighbourhood_Belmont neighbourhood_Belmont Cragin neighbourhood_Benning neighbourhood_Benning Ridge neighbourhood_Bensonhurst neighbourhood_Bergen Beach neighbourhood_Berkley neighbourhood_Bernal Heights neighbourhood_Beverly neighbourhood_Beverly Hills neighbourhood_Bloomingdale neighbourhood_Boerum Hill neighbourhood_Borough Park neighbourhood_Boyle Heights neighbourhood_Boystown neighbourhood_Bradbury neighbourhood_Brentwood neighbourhood_Bridgeport neighbourhood_Brighton Beach neighbourhood_Brightwood neighbourhood_Bronxdale neighbourhood_Bronzeville neighbourhood_Brookland neighbourhood_Brookline neighbourhood_Brooklyn neighbourhood_Brooklyn Heights neighbourhood_Brooklyn Navy Yard neighbourhood_Brownsville neighbourhood_Bucktown neighbourhood_Buena Vista neighbourhood_Burbank neighbourhood_Burleith neighbourhood_Bushwick neighbourhood_Cahuenga Pass neighbourhood_Canarsie neighbourhood_Canoga Park neighbourhood_Capitol Hill neighbourhood_Carroll Gardens neighbourhood_Carson neighbourhood_Carver Langston neighbourhood_Castle Hill neighbourhood_Cathedral Heights neighbourhood_Central Northeast/Mahaning Heights neighbourhood_Cerritos neighbourhood_Charlestown neighbourhood_Chatsworth neighbourhood_Chelsea neighbourhood_Chestnut Hill neighbourhood_Chevy Chase neighbourhood_Chinatown neighbourhood_City Island neighbourhood_Civic Center neighbourhood_Claremont neighbourhood_Cleveland Park neighbourhood_Clinton Hill neighbourhood_Co-op City neighbourhood_Cobble Hill neighbourhood_Cole Valley neighbourhood_College Point neighbourhood_Colonial Village neighbourhood_Columbia Heights neighbourhood_Columbia Street Waterfront neighbourhood_Commerce neighbourhood_Compton neighbourhood_Concourse neighbourhood_Concourse Village neighbourhood_Coney Island neighbourhood_Congress Heights neighbourhood_Corona neighbourhood_Covina neighbourhood_Cow Hollow neighbourhood_Crestwood neighbourhood_Crocker Amazon neighbourhood_Crotona neighbourhood_Crown Heights neighbourhood_Culver City neighbourhood_Cypress Park neighbourhood_DUMBO neighbourhood_Daly City neighbourhood_Deanwood neighbourhood_Del Rey neighbourhood_Diamond Heights neighbourhood_Ditmars / Steinway neighbourhood_Dogpatch neighbourhood_Dorchester neighbourhood_Douglass neighbourhood_Downey neighbourhood_Downtown neighbourhood_Downtown Brooklyn neighbourhood_Downtown Crossing neighbourhood_Downtown/Penn Quarter neighbourhood_Duarte neighbourhood_Duboce Triangle neighbourhood_Dunning neighbourhood_Dupont Circle neighbourhood_Dupont Park neighbourhood_Dyker Heights neighbourhood_Eagle Rock neighbourhood_East Boston neighbourhood_East Corner neighbourhood_East Elmhurst neighbourhood_East Flatbush neighbourhood_East Harlem neighbourhood_East Hollywood neighbourhood_East Los Angeles neighbourhood_East New York neighbourhood_East San Gabriel neighbourhood_East Village neighbourhood_Eastchester neighbourhood_Eastland Gardens

neighbourhood_Echo Park neighbourhood_Eckington neighbourhood_Edenwald
neighbourhood_Edgewater neighbourhood_Edgewood neighbourhood_El Monte
neighbourhood_El Segundo neighbourhood_El Sereno neighbourhood_Elm Park
neighbourhood_Elmhurst neighbourhood_Eltingville neighbourhood_Elysian Valley
neighbourhood_Encino neighbourhood_Englewood neighbourhood_Excelsior
neighbourhood_Fairlawn neighbourhood_Fenway/Kenmore neighbourhood_Financial
District neighbourhood_Fisherman's Wharf neighbourhood_Flatbush
neighbourhood_Flatiron District neighbourhood_Flatlands
neighbourhood_Florence-Graham neighbourhood_Flushing neighbourhood_Foggy
Bottom neighbourhood_Fordham neighbourhood_Forest Hill neighbourhood_Forest
Hills neighbourhood_Fort Davis neighbourhood_Fort Greene neighbourhood_Fort
Lincoln neighbourhood_Fort Wadsworth neighbourhood_Foxhall
neighbourhood_Fresh Meadows neighbourhood_Friendship Heights
neighbourhood_Gardena neighbourhood_Garfield Park neighbourhood_Garfield Ridge
neighbourhood_Georgetown neighbourhood_Gerritsen Beach neighbourhood_Glassell
Park neighbourhood_Glen Park neighbourhood_Glendale neighbourhood_Glendora
neighbourhood_Glover Park neighbourhood_Gold Coast neighbourhood_Good Hope
neighbourhood_Gowanus neighbourhood_Gramercy Park neighbourhood_Granada Hills
North neighbourhood_Grant City neighbourhood_Grasmere neighbourhood_Gravesend
neighbourhood_Great Kills neighbourhood_Greenpoint neighbourhood_Greenway
neighbourhood_Greenwich Village neighbourhood_Greenwood Heights
neighbourhood_Grymes Hill neighbourhood_Haight-Ashbury neighbourhood_Hamilton
Heights neighbourhood_Harbor City neighbourhood_Harbor Gateway
neighbourhood_Harlem neighbourhood_Harvard Square neighbourhood_Hawaiian
Gardens neighbourhood_Hawthorne neighbourhood_Hayes Valley
neighbourhood_Hell's Kitchen neighbourhood_Hermon neighbourhood_Hermosa
neighbourhood_Hermosa Beach neighbourhood_Highbridge neighbourhood_Highland
Park neighbourhood_Hillbrook neighbourhood_Hillcrest neighbourhood_Hollywood
neighbourhood_Hollywood Hills neighbourhood_Howard Beach neighbourhood_Hudson
Square neighbourhood_Huguenot neighbourhood_Humboldt Park
neighbourhood_Huntington Park neighbourhood_Hunts Point neighbourhood_Hyde
Park neighbourhood_Ingleside neighbourhood_Inglewood neighbourhood_Inner
Sunset neighbourhood_Inwood neighbourhood_Irving Park neighbourhood_Irwindale
neighbourhood_Ivy City neighbourhood_Jackson Heights neighbourhood_Jamaica
neighbourhood_Jamaica Plain neighbourhood_Japantown neighbourhood_Jefferson
Park neighbourhood_Judiciary Square neighbourhood_Kalorama
neighbourhood_Kensington neighbourhood_Kent neighbourhood_Kenwood
neighbourhood_Kew Garden Hills neighbourhood_Kingman Park
neighbourhood_Kingsbridge neighbourhood_Kingsbridge Heights neighbourhood_Kips
Bay neighbourhood_La Canada Flintridge neighbourhood_La Crescenta-Montrose
neighbourhood_La Habra neighbourhood_La Mirada neighbourhood_La Puente
neighbourhood_Lake Balboa neighbourhood_Lakeshore neighbourhood_Lakeview
neighbourhood_Lakewood neighbourhood_Lamond Riggs neighbourhood_Langdon
neighbourhood_Laurel Canyon neighbourhood_Lawndale neighbourhood_LeDroit Park
neighbourhood_Leather District neighbourhood_Lefferts Garden
neighbourhood_Lincoln Heights neighbourhood_Lincoln Park neighbourhood_Lincoln
Square neighbourhood_Lindenwood neighbourhood_Little Italy
neighbourhood_Little Italy/UIC neighbourhood_Little Village

neighbourhood_Logan Circle neighbourhood_Logan Square neighbourhood_Lomita
neighbourhood_Long Beach neighbourhood_Long Island City neighbourhood_Longwood
neighbourhood_Loop neighbourhood_Los Feliz neighbourhood_Lower East Side
neighbourhood_Lower Haight neighbourhood_Lynwood neighbourhood_Magnificent
Mile neighbourhood_Malibu neighbourhood_Manhattan neighbourhood_Manhattan
Beach neighbourhood_Manor Park neighbourhood_Mar Vista neighbourhood_Marble
Hill neighbourhood_Marina neighbourhood_Marina Del Rey neighbourhood_Marine
Park neighbourhood_Marshall Heights neighbourhood_Maspeth
neighbourhood_Mattapan neighbourhood_McKinley Park neighbourhood_Meatpacking
District neighbourhood_Meiers Corners neighbourhood_Melrose
neighbourhood_Michigan Park neighbourhood_Mid-City neighbourhood_Mid-Wilshire
neighbourhood_Middle Village neighbourhood_Midland Beach neighbourhood_Midtown
neighbourhood_Midtown East neighbourhood_Midwood neighbourhood_Mission Bay
neighbourhood_Mission District neighbourhood_Mission Hill
neighbourhood_Mission Terrace neighbourhood_Monrovia neighbourhood_Montclare
neighbourhood_Montebello neighbourhood_Montecito Heights
neighbourhood_Monterey Hills neighbourhood_Monterey Park neighbourhood_Morgan
Park neighbourhood_Morningside Heights neighbourhood_Morris Heights
neighbourhood_Morrisania neighbourhood_Mott Haven neighbourhood_Mount Eden
neighbourhood_Mount Pleasant neighbourhood_Mount Vernon Square
neighbourhood_Mount Washington neighbourhood_Mt Rainier/Brentwood, MD
neighbourhood_Murray Hill neighbourhood_Navy Yard neighbourhood_Naylor Gardens
neighbourhood_Near North Side neighbourhood_Near Northeast neighbourhood_Near
Northeast/H Street Corridor neighbourhood_Near West Side neighbourhood_New
Brighton neighbourhood_New Dorp Beach neighbourhood_Newton neighbourhood_Nob
Hill neighbourhood_Noe Valley neighbourhood_Noho neighbourhood_Nolita
neighbourhood_North Beach neighbourhood_North Center neighbourhood_North
Cleveland Park neighbourhood_North End neighbourhood_North Hills East
neighbourhood_North Hills West neighbourhood_North Hollywood
neighbourhood_North Lawndale neighbourhood_North Michigan Park
neighbourhood_North Park neighbourhood_Northridge neighbourhood_Norwalk
neighbourhood_Norwood neighbourhood_Norwood Park neighbourhood_O'Hare
neighbourhood_Oakland neighbourhood_Oceanview neighbourhood_Old Town
neighbourhood_Outer Sunset neighbourhood_Ozone Park neighbourhood_Pacific
Heights neighbourhood_Pacific Palisades neighbourhood_Pacoima
neighbourhood_Palisades neighbourhood_Palms neighbourhood_Palos Verdes
neighbourhood_Panorama City neighbourhood_Paramount neighbourhood_Park Slope
neighbourhood_Park Versailles neighbourhood_Park View
neighbourhood_Parkchester neighbourhood_Parkside neighbourhood_Pasadena
neighbourhood_Pelham Bay neighbourhood_Petworth neighbourhood_Pico Rivera
neighbourhood_Pilsen neighbourhood_Pleasant Hill neighbourhood_Pleasant Plains
neighbourhood_Port Morris neighbourhood_Port Richmond neighbourhood_Portage
Park neighbourhood_Porter Ranch neighbourhood_Portola neighbourhood_Potrero
Hill neighbourhood_Presidio neighbourhood_Presidio Heights
neighbourhood_Prospect Heights neighbourhood_Rancho Palos Verdes
neighbourhood_Randall Manor neighbourhood_Randle Highlands neighbourhood_Red
Hook neighbourhood_Redondo Beach neighbourhood_Rego Park neighbourhood_Reseda
neighbourhood_Richmond District neighbourhood_Richmond Hill

neighbourhood_Ridgewood neighbourhood_River North neighbourhood_River Terrace neighbourhood_River West neighbourhood_Riverdale neighbourhood_Rogers Park neighbourhood_Rolling Hills neighbourhood_Roosevelt Island neighbourhood_Roscoe Village neighbourhood_Rosebank neighbourhood_Roseland neighbourhood_Rosemead neighbourhood_Roslindale neighbourhood_Rossville neighbourhood_Roxbury neighbourhood_Russian Hill neighbourhood_San Gabriel neighbourhood_San Marino neighbourhood_San Pedro neighbourhood_Santa Fe Springs neighbourhood_Santa Monica neighbourhood_Sea Cliff neighbourhood_Sea Gate neighbourhood_Shaw neighbourhood_Sheepshead Bay neighbourhood_Shepherd Park neighbourhood_Sherman Oaks neighbourhood_Sierra Madre neighbourhood_Signal Hill neighbourhood_Silver Lake neighbourhood_Silver Spring, MD neighbourhood_Skid Row neighbourhood_SoMa neighbourhood_Soho neighbourhood_Somerville neighbourhood_Soundview neighbourhood_South Beach neighbourhood_South Boston neighbourhood_South Chicago neighbourhood_South El Monte neighbourhood_South End neighbourhood_South Gate neighbourhood_South LA neighbourhood_South Loop/Printers Row neighbourhood_South Ozone Park neighbourhood_South Pasadena neighbourhood_South Robertson neighbourhood_South San Gabriel neighbourhood_South Shore neighbourhood_South Street Seaport neighbourhood_South Whittier neighbourhood_Southwest Waterfront neighbourhood_Spring Valley neighbourhood_Spuyten Duyvil neighbourhood_St. Elizabeths neighbourhood_St. George neighbourhood_Stapleton neighbourhood_Streeterville neighbourhood_Stronghold neighbourhood_Studio City neighbourhood_Sun Valley neighbourhood_Sunland/Tujunga neighbourhood_Sunnyside neighbourhood_Sunset Park neighbourhood_Sylmar neighbourhood_Takoma neighbourhood_Takoma Park, MD neighbourhood_Tarzana neighbourhood_Telegraph Hill neighbourhood_Temple City neighbourhood_Tenderloin neighbourhood_The Bronx neighbourhood_The Castro neighbourhood_The Rockaways neighbourhood_Theater District neighbourhood_Throgs Neck neighbourhood_Times Square/Theatre District neighbourhood_Toluca Lake neighbourhood_Tompkinsville neighbourhood_Topanga neighbourhood_Torrance neighbourhood_Tottenville neighbourhood_Tremont neighbourhood_Tribeca neighbourhood_Trinidad neighbourhood_Truxton Circle neighbourhood_Twin Peaks neighbourhood_Twining neighbourhood_U Street Corridor neighbourhood_Ukrainian Village neighbourhood_Union Square neighbourhood_University Heights neighbourhood_Upper East Side neighbourhood_Upper West Side neighbourhood_Uptown neighbourhood_Valley Glen neighbourhood_Valley Village neighbourhood_Van Nest neighbourhood_Van Nuys neighbourhood_Venice neighbourhood_Vinegar Hill neighbourhood_Visitacion Valley neighbourhood_Wakefield neighbourhood_Washington Heights neighbourhood_Washington Highlands neighbourhood_Watertown neighbourhood_Watts neighbourhood_Wesley Heights neighbourhood_West Adams neighbourhood_West Athens neighbourhood_West Brighton neighbourhood_West Covina neighbourhood_West End neighbourhood_West Farms neighbourhood_West Hills neighbourhood_West Hollywood neighbourhood_West Lawn neighbourhood_West Loop/Greektown neighbourhood_West Los Angeles neighbourhood_West Portal neighbourhood_West Puente Valley neighbourhood_West Ridge neighbourhood_West Roxbury neighbourhood_West Town/Noble Square neighbourhood_West Village neighbourhood_Westchester Village neighbourhood_Westchester/Playa Del Rey

```
neighbourhood_Western Addition/NOPA  neighbourhood_Westlake
neighbourhood_Westmont  neighbourhood_Westside  neighbourhood_Westwood
neighbourhood_Whitestone  neighbourhood_Whittier  neighbourhood_Wicker Park
neighbourhood_Williamsbridge  neighbourhood_Williamsburg  neighbourhood_Windsor
Terrace  neighbourhood_Winnetka  neighbourhood_Winthrop  neighbourhood_Woodhaven
neighbourhood_Woodland  neighbourhood_Woodland Hills/Warner Center
neighbourhood_Woodlawn  neighbourhood_Woodley Park  neighbourhood_Woodridge
neighbourhood_Woodside  neighbourhood_Wrigleyville
False                              False                       False
False                      False                   False                   False
False                        False                   False
False                    False                        False
False                  False                 False                        False
False                  False                          False
False                   False                False                   False
False                         False                        False
False                        False                     False
False                    False                     False                  False
False                      False                       False
False                                False               False
False                     False                  False
False                  False                        False
False                       False                  False
False                  False                        False
False                      False                       False
False                   False                   False                    False
False                         False                     False
False                      False                    False
False                   False                           False
False                      False                   False
False                  False                   False                   False
False                   False                      False
False                          False                 False
False                       False                            False
False                   False                      False
False                  False                        False
False                    False                      False
False                    False                         False
False                     False                      False
False                        False                           False
False                                     False                   False
False                    False                            False
False                           False                 False
False                     False                    False
False                  False                        False
False                       False                False                    False
False                  False                          False
False                   False                     False                   False
False                   False                            False
False                                False                 False
False                  False                        False
False                        False                     False
False                      False                        False
False                      False                         False
False                        False                           False
False                      False                           False
False                    False                   False                    False
False                   False                     False                    False
False                   False                      False
False                 False                    False                    False
False                         False                             False
False                   False                            False
False                          False                   False
False                  False                      False
False                     False                      False
False                         False                  False
False                             False                  False
False                         False                     False
False                        False                    False
False                   False                      False
False                    False                  False
False                              False                     False
False                    False                      False
False                   False                            False
False                      False                         False
False                      False                         False
False                         False                           False
False                       False                         False
False                  False                        False
False                        False                    False
False                    False                          False
False                        False                   False
False                          False                      False
False                    False                    False
False                 False                      False                    False
False                          False                  False
False                    False                         False
False                   False                     False               False
False                           False                       False
False                              False                   False
False                                False                   False
False                    False                      False
False                   False                   False
False                  False                        False
False                       False                           False
False                          False                       False
False                     False                       False
False                         False                       False
False                 False                     False
False                   False               False                    False
False                       False                  False
False                 False                    False
False                     False                    False
False                 False                         False
False                           False                  False
False                        False                               False
False                  False                        False
False                       False                         False
False                  False                       False                  False
False                           False                       False
False                   False                    False                     False
False                         False                        False
False                              False                         False
False                     False                     False
False                              False                           False
False                      False                    False
False                          False                         False
False                         False                       False
False                 False                   False                     False
False                 False                      False
False                               False                    False
False                           False                          False
False                              False                     False
False                  False                  False                       False
False                  False                    False                   False
False                     False                          False
False                  False                    False                False
False                        False                    False
False                          False                    False
False                   False                   False                     False
False                      False                 False
False                          False                      False
False                       False                       False
False                       False                   False
False                           False                              False
False                           False                   False
False                    False                 False
False                        False                    False
False                        False                     False
False                      False                        False
False                         False                   False
False                   False                     False                    False
False                       False                      False
False                    False                           False
False                    False                   False               False
False                        False                       False
False                      False                      False
False                   False               False               False
False                    False                      False
False                        False                         False
False                     False                    False
False                           False                         False
False                            False                      False
False                         False                               False
False                         False                         False
False                    False                            False
False                      False                      False
False                    False                      False                 False
False                          False                  False
False                      False                      False
False                     False                        False
False                      False                                    False
False                        False                  False
False                      False                 False                 False
False                         False                    False
False                            False                            False
False                             False                        False
False                 False                      False
False                   False                   False                 False
False                            False                    False
False                               False                    False
False                         False                    False
False                        False                     False
False                     False                     False
False                    False                                False
False                      False                              False
False                       False                                 False
False                              False
False                                False                   False
False                   False                   False                     False
False                      False                        False
False                          False                  False
False                    False                   False
False                   False                      False
False                   False                         4.607565
                  True                            5.236394
                                                                          True
False                         4.428393
                                                 True                     False
False                         4.870117
                     True                        False                    False
False                         5.002670
```

```
                                                  …
                                      True                    False
False                           False                        False
False                                   False                    False
False                    False                  False                 False
False                        False                  False
False                 False                   False                False
False                             False                        False
False                        False                        False
False                             False                    False
False                  False                  False                      False
False                                  False                                False
False                   False                     False                  False
False                  False                        False
False                       False                  False
False                  False                        False
False                      False                       False
False                   False                   False                    False
False                         False                     False
False                      False                    False
False                   False                          False
False                      False                   False
False                  False                   False                   False
False                   False                      False
False                          False                 False
False                      False                            False
False                   False                      False
False                  False                        False
False                    False                      False
False                    False                         False
False                     False                      False
False                        False                           False
False                                     False                   False
False                    False                            False
False                           False                 False
False                     False                    False
False                  False                        False
False                       False                False                    False
False                  False                          False
False                   False                     False                   False
False                   False                            False
False                                False                 False
False                  False                        False
False                        False                     False
False                      False                        False
False                      False                         False
False                        False                           False
False                      False                           False
False                    False                   False                    False
False                   False                     False                    False
False                   False                      False
False                 False                    False                    False
False                         False                            False
False                   False                            False
False                          False                  False
False                  False                     False
False                     False                    False
False                         False                False
False                             False                False
False                         False                     False
False                        False                    False
False                   False                      False
False                    False                  False
False                              False                    False
False                    False                      False
False                   False                            False
False                      False                          False
False                      False                          False
False                         False                           False
False                       False                        False
False                  False                        False
False                        False                    False
False                    False                          False
False                        False                   False
False                          False                      False
False                    False                    False
False                 False                    False                    False
False                          False                  False
False                    False                        False
False                   False                    False                False
False                           False                      False
False                              False                   False
False                                False                   False
False                    False                      False
False                   False                  False
False                  False                        False
False                       False                       False
False                          False                       False
False                     False                       False
False                         False                       False
False                 False                     False
False                   False                False                    False
False                       False                  False
False                 False                    False
False                     False                    False
False                 False                         False
False                           False                  False
False                        False                               False
False                  False                        False
False                       False                         False
False                  False                       False                  False
False                           False                       False
False                   False                    False                     False
False                         False                        False
False                              False                         False
False                     False                     False
False                              False                           False
False                      False                    False
False                          False                         False
False                         False                       False
False                 False                   False                     False
False                 False                      False
False                               False                    False
False                           False                          False
False                              False                     False
False                  False                  False                       False
False                  False                    False                   False
False                     False                          False
False                  False                    False                False
False                        False                    False
False                          False                    False
False                   False                   False                     False
False                      False                 False
False                          False                      False
False                       False                       False
False                       False                   False
False                           False                              False
False                           False                   False
False                    False                 False
False                        False                    False
False                        False                     False
False                      False                        False
False                         False                   False
False                   False                     False                    False
False                       False                      False
False                    False                           False
False                    False                   False               False
False                        False                       False
False                      False                      False
False                   False               False               False
False                    False                      False
False                        False                         False
False                     False                   False
False                           False                         False
False                            False                      False
False                         False                               False
False                         False                         False
False                    False                        False
False                      False                     False
False                    False                      False                 False
False                          False                  False
False                      False                     False
False                     False                        False
False                      False                                        False
False                        False                  False
False                      False                  False                  False
False                         False                     False
False                            False                            False
False                             False                          False
False                 False                      False
False                   False                   False                 False
False                            False                    False
False                               False                    False
False                         False                     False
False                        False                      False
False                     False                     False
False                    False                              False
False                      False                             False
False                       False                                 False
False                              False
False                                False                   False
False                   False                   False                     False
False                      False                         False
False                          False                   False
False                    False                   False
False                   False                       False
False                   False                         4.206644
           True                       False                   False
False                           False                        False
False                                   False                    False
False                    False                  False                 False
False                        False                  False
False                 False                   False                False
False                            False                         False
False                       False                        False
False                            False                    False
False                  False                  False                      False
False                             False                                False
False                   False                     False                  False
False                  False                        False
False                       False                  False
False                  False                        False
False                      False                      False
False                   False                   False                    False
False                         False                     False
False                      False                    False
False                   False                           False
False                      False                   False
False                  False                   False                   False
False                   False                      False
False                          False                 False
False                       False                            False
False                   False                      False
False                  False                       False
False                    False                     False
False                    False                        False
False                     False                    False
False                        False                           False
False                                     False                    False
False                    False                            False
False                           False                 False
False                     False                    False
False                  False                       False
False                       False                other  False
False                    False                      False
False                   False                            False
False                      False                         False
False                      False                         False
False                         False                           False
False                       False                         False
False                  False                        False
False                        False                    False
False                    False                          False
False                        False                   False
False                          False                      False
False                    False                    False
False                 False                      False                    False
False                          False                  False
False                    False                         False
False                   False                     False               False
False                           False                       False
False                              False                   False
False                                False                   False
False                    False                      False
False                   False                   False
False                  False                        False
False                       False                           False
False                          False                       False
False                     False                       False
False                         False                       False
False                 False                     False
False                   False               False                    False
False                       False                  False
False                 False                    False
False                     False                    False
False                 False                         False
False                           False                  False
False                        False                               False
False                  False                        False
False                       False                         False
False                  False                       False                  False
False                           False                       False
False                   False                    False                     False
False                         False                        False
False                              False                         False
False                     False                     False
False                              False                           False
False                      False                    False
False                          False                         False
False                         False                       False
False                 False                   False                     False
False                 False                      False
False                               False                    False
False                           False                          False
False                              False                     False
False                  False                  False                       False
False                  False                    False                   False
False                     False                          False
False                  False                    False                False
False                        False                    False
False                          False                    False
False                   False                   False                     False
False                      False                 False
False                          False                      False
False                       False                       False
False                       False                   False
False                           False                              False
False                           False                   False
False                    False                 False
False                        False                    False
False                        False                     False
False                      False                        False
False                         False                   False
False                   False                     False                    False
False                       False                      False
False                    False                           False
False                    False                   False               False
False                        False                       False
False                      False                      False
False                   False               False               False
False                    False                      False
False                        False                         False
False                     False                   False
False                           False                         False
False                            False                      False
False                         False                               False
False                         False                         False
False                    False                        False
False                      False                     False
False                    False                      False                 False
False                          False                  False
False                      False                     False
False                     False                        False
False                      False                                        False
False                        False                  False
False                      False                  False                  False
False                         False                     False
False                            False                            False
False                             False                          False
False                 False                      False
False                   False                   False                False
False                            False                    False
False                               False                    False
False                         False                     False
False                        False                      False
False                     False                      False
False                    False                              False
False                      False                              False
False                       False                                 False
False                              False
False                                False                   False
False                   False                   False                     False
False                      False                         False
False                          False                    False
False                    False                   False
False                   False                      False
False                   False                        4.211454
                                                               True
False                      False                   False                   False
False                        False                   False
False                    False                        False
False                  False                 False                        False
False                  False                          False
False                   False                 False                  False
False                         False                        False
False                        False                     False
False                    False                     False                  False
False                      False                       False
False                                False                 False
False                     False                  False
False                  False                       False
False                       False                  False
False                  False                       False
False                      False                       False
False                   False                   False                    False
False                         False                     False
False                      False                    False
False                   False                       False
False                      False                  False
False                  False                   False                  False
False                   False                     False
False                          False                 False
False                       False                            False
False                   False                     False
False                  False                       False
False                    False                      False
False                    False                         False
False                     False                    False
False                        False                           False
False                                     False                   False
False                    False                            False
False                           False                 False
False                     False                    False
False                  False                        False
False                       False                False                    False
False                  False                          False
False                   False                     False                   False
False                   False                            False
False                                the False                 False
False                  False                        False
False                        False                     False
False                      False                        False
False                      False                         False
False                        False                           False
False                      False                           False
False                    False                   False                    False
False                   False                     False                    False
False                   False                      False
False                 False                    False                    False
False                         False                             False
False                   False                            False
False                          False                   False
False                  False                      False
False                     False                      False
False                         False                  False
False                             False                  False
False                         False                     False
False                        False                    False
False                   False                      False
False                    False                  False
False                              False                     False
False                    False                      False
False                   False                            False
False                      False                         False
False                      False                         False
False                         False                           False
False                       False                         False
False                  False                        False
False                        False                    False
False                    False                          False
False                        False                   False
False                          False                      False
False                    False                    False
False                 False                      False                    False
False                          False                  False
False                    False                         False
False                   False                     False               False
False                           False                       False
False                              False                   False
False                                False                   False
False                    False                      False
False                   False                   False
False                  False                        False
False                       False                           False
False                          False                       False
False                     False                       False
False                         False                       False
False                 False                     False
False                   False               False                    False
False                       False                  False
False                 False                    False
False                     False                    False
False                 False                         False
False                           False                  False
False                        False                               False
False                  False                        False
False                       False                         False
False                  False                       False                  False
False                           False                       False
False                   False                    False                     False
False                         False                        False
False                              False                         False
False                     False                     False
False                              False                           False
False                      False                    False
False                          False                         False
False                         False                       False
False                 False                   False                     False
False                 False                      False
False                               False                    False
False                           False                          False
False                              False                     False
False                  False                  False                       False
False                  False                    False                   False
False                     False                          False
False                  False                    False                False
False                        False                    False
False                          False                    False
False                   False                   False                     False
False                      False                 False
False                          False                      False
False                       False                       False
False                       False                   False
False                           False                              False
False                           False                   False
False                    False                 False
False                        False                    False
False                        False                     False
False                      False                        False
False                         False                   False
False                   False                     False                    False
False                       False                      False
False                    False                          False
False                    False                   False               False
False                        False                      False
False                      False                    False
False                   False               False               False
False                    False                      False
False                        False                         False
False                     False                    False
False                           False                         False
False                            False                     False
False                         False                            False
False                         False                        False
False                    False                     False
False                      False                  False
False                    False                      False                False
False                          False                 False
False                      False                  False
False                     False                    False
False                      False                                  False
False                        False                  False
False                      False                    False                  False
False                         False                    False
False                            False                            False
False                             False                          False
False                 False                    False
False                   False                   False                False
False                            False                    False
False                               False                    False
False                         False                     False
False                        False                      False
False                     False                      False
False                    False                          False
False                      False                         False
False                       False                              False
False                             False
False                               False                   False
False                   False                   False                    False
False                      False                         False
False                          False                   False
False                    False                   False
False                   False                       False
False                   False                         5.188996
                                   True                        False
False                      False                   False                   False
False                        False                   False
False                    False                        False
False                  False                 False                        False
False                  False                          False
False                   False                False                   False
False                         False                        False
False                        False                     False
False                    False                     False                  False
False                      False                       False
False                                False               False
False                     False                  False
False                  False                        False
False                       False                  False
False                  False                        False
False                      False                       False
False                   False                   False                    False
False                         False                     False
False                      False                    False
False                   False                           False
False                      False                   False
False                  False                   False                   False
False                   False                      False
False                          False                 False
False                       False                            False
False                   False                      False
False                  False                        False
False                    False                      False
False                    False                         False
False                     False                      False
False                        False                           False
False                                     False                   False
False                    False                            False
False                           False                 False
False                     False                    False
False                  False                        False
False                       False                False                    False
False                  False                          False
False                   False                     False                   False
False                   False                            False
False                                False                 False
False                  False                        False
False                        False                     False
False                      False                        False
False                      False                         False
False                        False                           False
False                      False                           False
False                    False                   False                    False
False                   False                     False                    False
False                   False                      False
False                 False                    False                    False
False                         False                             False
False                   False                            False
False                          False                   False
False                  False                      False
False                     False                      False
False                         False                  False
False                             False                  False
False                         False                     False
False                        False                    False
False                   False                      False
False                    False                  False
False                              False                     False
False                    False                      False
False                   False                            False
False                      False                         False
False                      False                         False
False                         False                           False
False                       False                         False
False                  False                        False
False                        False                    False
False                    False                          False
False                        False                   False
False                          False                      False
False                    False                    False
False                 False                      False                    False
False                          False                  False
False                    False                         False
False                   False                     False               False
False                           False                       False
False                              False                   False
False                                False                   False
False                    False                      False
False                   False                   False
False                  False                        False
False                       False                           False
False                          False                       False
False                     False                       False
False                         False                       False
False                 False                     False
False                   False               False                    False
False                       False                  False
False                 False                    False
False                     False                    False
False                 False                         False
False                           False                  False
False                        False                               False
False                  False                        False
False                       False                         False
False                  False                       False                  False
False                           False                       False
False                   False                    False                     False
False                         False                        False
False                              False                         False
False                     False                     False
False                              False                           False
False                      False                    False
False                          False                         False
False                         False                       False
False                 False                   False                     False
False                 False                      False
False                               False                    False
False                           False                          False
False                              False                     False
False                  False                  False                       False
False                  False                    False                   False
False                     False                          False
False                  False                    False                False
False                        False                    False
False                          False                    False
False                   False                   False                     False
False                      False                 False
False                          False                      False
False                       False                       False
False                       False                   False
False                           False                              False
False                           False                   False
False                    False                 False
False                        False                    False
False                        False                     False
False                      False                        False
False                         False                   False
False                   False                     False                    False
False                       False                      False
False                    False                           False
False                    False                   False               False
False                        False                       False
False                      False                      False
False                   False               False               False
False                    False                      False
False                        False                         False
False                     False                   False
False                           False                         False
False                            False                      False
False                         False                               False
False                         False                         False
False                    False                        False
False                      False                     False
False                    False                      False                 False
False                          False                  False
False                      False                     False
False                     False                        False
False                      False                                        False
False                        False                  False
False                      False                  False                  False
False                         False                     False
False                            False                            False
False                             False                          False
False                 False                      False
False                   False                   False                 False
False                            False                    False
False                               False                    False
False                         False                     False
False                        False                      False
False                     False                     False
False                    False                              False
False                      False                             False
False                       False                                 False
False                              False
False                                False                   False
False                   False                   False                     False
False                      False                         False
False                          False                   False
False                    False                   False
False                   False                       False
False                   False                         4.962115
True                               False                       False
False                      False                   False                   False
False                        False                   False
False                    False                        False
False                  False                 False                        False
False                  False                          False
False                   False                False                   False
False                         False                        False
False                        False                     False
False                    False                     False                  False
False                      False                       False
False                                False               False
False                     False                  False
False                  False                        False
False                       False                  False
False                  False                        False
False                      False                       False
False                   False                   False                    False
False                         False                     False
False                      False                    False
False                   False                           False
False                      False                   False
False                  False                   False                   False
False                   False                      False
False                          False                 False
False                       False                            False
False                   False                      False
False                  False                        False
False                    False                      False
False                    False                         False
False                     False                      False
False                        False                           False
False                                     False                   False
False                    False                            False
False                           False                 False
False                     False                    False
False                  False                        False
False                       False                False                    False
False                  False                          False
False                   False                     False                   False
False                   False                            False
False                                False                 False
False                  False                        False
False                        False                     False
False                      False                        False
False                      False                         False
False                        False                           False
False                      False                           False
False                    False                   False                    False
False                   False                     False                    False
False                   False                      False
False                 False                    False                    False
False                         False                             False
False                   False                            False
False                          False                   False
False                  False                      False
False                     False                      False
False                         False                  False
False                             False                  False
False                         False                     False
False                        False                    False
False                   False                      False
False                    False                  False
False                              False                     False
False                    False                      False
False                   False                            False
False                      False                         False
False                      False                         False
False                         False                           False
False                       False                         False
False                  False                        False
False                        False                    False
False                    False                          False
False                        False                   False
False                          False                      False
False                    False                    False
False                 False                      False                    False
False                          False                  False
False                    False                         False
False                   False                     False               False
False                           False                       False
False                              False                   False
False                                False                   False
False                    False                      False
False                   False                   False
False                  False                        False
False                       False                           False
False                          False                       False
False                     False                       False
False                         False                       False
False                 False                     False
False                   False               False                    False
False                       False                  False
False                 False                    False
False                     False                    False
False                 False                         False
False                           False                  False
False                        False                               False
False                  False                        False
False                       False                         False
False                  False                       False                  False
False                           False                       False
False                   False                    False                     False
False                         False                        False
False                              False                         False
False                     False                     False
False                              False                           False
False                      False                    False
False                          False                         False
False                         False                       False
False                 False                   False                     False
False                 False                      False
False                               False                    False
False                           False                          False
False                              False                     False
False                  False                  False                       False
False                  False                    False                   False
False                     False                          False
False                  False                    False                False
False                        False                    False
False                          False                    False
False                   False                   False                     False
False                      False                 False
False                          False                      False
False                       False                       False
False                       False                   False
False                           False                              False
False                           False                   False
False                    False                 False
False                        False                    False
False                        False                     False
False                      False                        False
False                         False                   False
False                   False                     False                    False
False                       False                      False
False                    False                           False
False                    False                   False               False
False                        False                       False
False                      False                      False
False                   False               False               False
False                    False                      False
False                        False                         False
False                     False                   False
False                           False                         False
False                            False                      False
False                         False                               False
False                         False                         False
False                    False                        False
False                      False                     False
False                    False                      False                 False
False                          False                  False
False                      False                     False
False                     False                        False
False                      False                                        False
False                        False                  False
False                      False                  False                  False
False                         False                     False
False                                 False                                 False
False                                  False                              False
False                     False                       False
False                       False                    False                     False
False                                 False                    False
False                                    False                    False
False                              False                      False
False                             False                       False
False                          False                     False
False                         False                               False
False                           False                              False
False                            False                                  False
False                                   False
False                                     False                    False
False                       False                    False                       False
False                           False                         False
False                               False                   False
False                         False                   False
False                       False                        False
False                       False                          4.798202
Name: log_price, Length: 558, dtype: float64
```

[55]:
```
# Corrélation forte entre le prix de l'appartement et le neighbourhood ?
# Calcul de la matrice de corrélation
matrice_corr_neighbourhood= df_bin[["log_price"] + [col for col in df_bin.
  ↪columns if "neighbourhood" in col]].corr()
```

[56]:
```
# heatmap
plt.figure(figsize=(10, 8))
#sns.heatmap(matrice_corr_neighbourhood, annot=True, cmap="coolwarm")

#plt.xticks(rotation=45)
#plt.yticks(rotation=0)
#plt.title("Matrice de corrélation entre log_prix et le voisinage")

#plt.show()
```

[56]:
```
<Figure size 1000x800 with 0 Axes>
```

```
<Figure size 1000x800 with 0 Axes>
```

Comme il y en a beaucoup, le chargement est très long (et mon pc a presque crashé), mais à mon avis cette donnée joue sur le prix

- Pour les zipcode :

[57]:
```
# Pour voir le log_price en fonction du zipcode :
#log_price_en_fct_zipcode = df_bin.groupby([col for col in df_bin.columns if␣
  ↪'zipcode' in col])['log_price'].mean()
```

```python

#print(log_price_en_fct_zipcode)
```

De même, il y en a beaucoup, mais cette donnée joue aussi sur le prix

- Pour le nombre de chambres :

[58]:
```python
# Pour voir le log_price en fonction du nombre de chambres :
log_price_en_fct_bedrooms = df_bin.groupby([col for col in df_bin.columns if␣
 ↪'bedrooms' in col])['log_price'].mean()

print(log_price_en_fct_bedrooms)
```

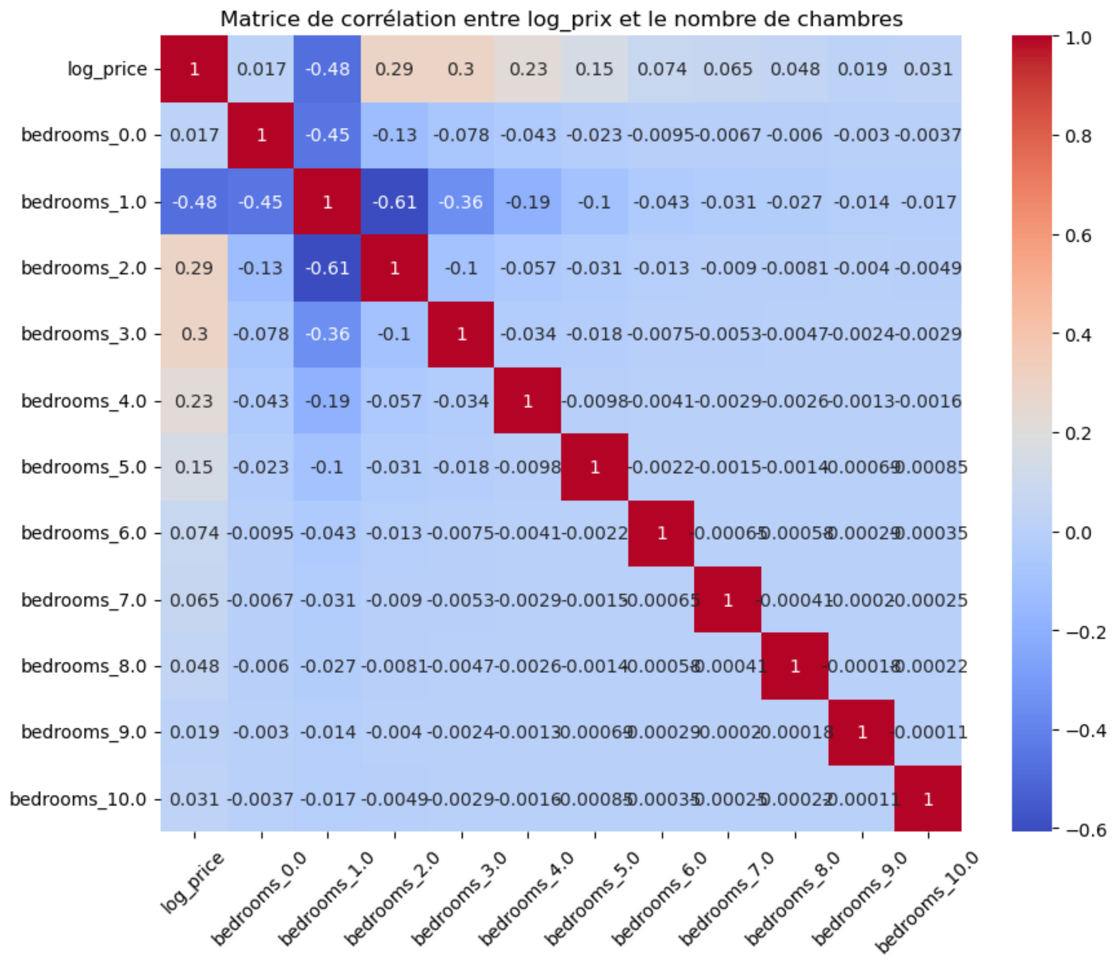
```
bedrooms_0.0  bedrooms_1.0  bedrooms_2.0  bedrooms_3.0  bedrooms_4.0
bedrooms_5.0  bedrooms_6.0  bedrooms_7.0  bedrooms_8.0  bedrooms_9.0
bedrooms_10.0
False         False         False         False         False         False
False         False         False         False         False
4.790867
                                                            True
6.713924
                                              True          False
6.224509
                                True          False         False
6.605355
                  True          False         False         False
6.963010
    True          False         False         False         False
6.541223
                                                                      True
False         False         False         False         False
6.256977
                                                        True          False
False         False         False         False         False
5.978486
                                          True          False         False
False         False         False         False         False
5.644117
                            True          False         False         False
False         False         False         False         False
5.279051
              True          False         False         False         False
False         False         False         False         False
4.543968
True          False         False         False         False         False
False         False         False         False         False
4.820819
Name: log_price, dtype: float64
```

[59]:
```python
# Corrélation forte entre le prix de l'appartement et le nombre de chambres ?
# Calcul de la matrice de corrélation
matrice_corr_bedrooms= df_bin[["log_price"] + [col for col in df_bin.columns if␣
↪"bedrooms" in col]].corr()
```

[60]:
```python
# heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(matrice_corr_bedrooms, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et le nombre de chambres")

plt.show()
```



Cette donnée a une influence sur le prix

- Pour le nombre de lits :

```
[61]: # Pour voir le log_price en fonction du nombre de lits :
      log_price_en_fct_beds = df_bin.groupby([col for col in df_bin.columns if 'beds'␣
       ↪in col])['log_price'].mean()

      print(log_price_en_fct_beds)
```

```
beds_0.0  beds_1.0  beds_2.0  beds_3.0  beds_4.0  beds_5.0  beds_6.0  beds_7.0
beds_8.0  beds_9.0  beds_10.0  beds_11.0  beds_12.0  beds_13.0  beds_16.0
beds_18.0
False     False     False     False     False     False     False     False
False     False     False      False      False      False      False      False
4.476913
                                                                           True
6.684612
                                                                True       False
5.982366
                                                     True       False      False
5.922091
                                          True       False      False      False
5.770915
                               True       False      False      False      False
5.960961
                    True       False      False      False      False      False
4.657515
          True      False      False      False      False      False      False
5.961424
True      False     False      False      False      False      False      False
5.330195
                                                                      True
False     False     False      False      False      False      False      False
5.984289
                                                            True      False
False     False     False      False      False      False      False      False
5.682948
                                                  True      False     False
False     False     False      False      False      False      False      False
5.741630
                                        True      False     False     False
False     False     False      False      False      False      False      False
5.466450
                              True      False     False     False     False
False     False     False      False      False      False      False      False
5.289058
                    True      False     False     False     False     False
False     False     False      False      False      False      False      False
5.003002
          True      False     False     False     False     False     False
```

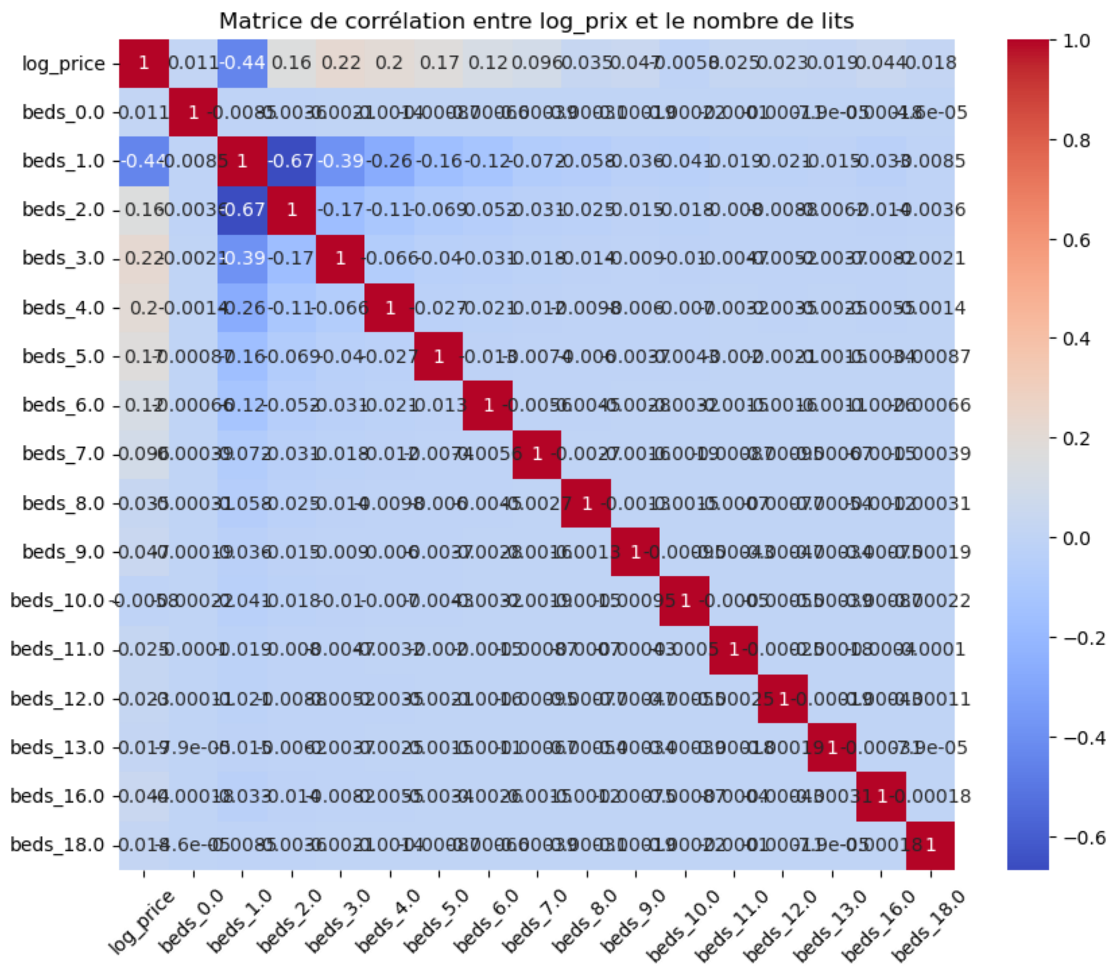
```
False     False     False      False      False      False      False      False
4.530720
True      False     False     False     False     False     False     False
False     False     False      False      False      False      False      False
5.991465
Name: log_price, dtype: float64
```

[62]:
```python
# Corrélation forte entre le prix de l'appartement et le nombre de lits ?
# Calcul de la matrice de corrélation
matrice_corr_beds= df_bin[["log_price"] + [col for col in df_bin.columns if␣
 ↪"beds" in col]].corr()
```

[63]:
```python
# heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(matrice_corr_beds, annot=True, cmap="coolwarm")

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.title("Matrice de corrélation entre log_prix et le nombre de lits")

plt.show()
```



Je remarque une influence sur le prix

Donc les corrélations que je garde :

[64]:
```
colonnes = ['property_type', 'room_type', 'cleaning_fee', 'city',␣
 ↪'neighbourhood', 'zipcode', 'bedrooms', 'beds']
```

Diagrammes (pour voir la distribution des prix en fonction des variables fortement corrélées) :

[65]:
```
fig = px.histogram(df, x = 'log_price', title = 'Distribution du log_price')
fig.show()
```

[66]:
```
# Box plot pour visualiser la distribution du log_price pour chaque type de␣
 ↪propriété
fig = px.box(df, x = 'property_type', y = 'log_price', title = 'log_price par␣
 ↪property_type')
fig.show()
```

[67]:
```
# log_price par type de chambre
fig = px.bar(df, x='room_type', y='log_price', title='log_price par room_type',
             labels={'log_price': 'log_price'})
fig.show()
```

[68]:
```
# Box plot des types de chambres et du log_price
fig = px.box(df, x='room_type', y='log_price', title='log_price par room_type')
fig.show()
```

[69]:
```
# Histogramme des accommodates et du log_price
fig = px.histogram(df, x='accommodates', y='log_price', title='log_price par␣
 ↪Accommodates', histfunc='avg')
fig.show()
```

[70]:
```
# Diagramme en barres des politiques d'annulation et du log_price
fig = px.bar(df, x='cancellation_policy', y='log_price', title='log_price par␣
 ↪cancellation_policy',
             labels={'log_price': 'log_price'})
fig.show()
```

[71]:
```
# Box plot des quartiers et du log_price
fig = px.box(df, x='neighbourhood', y='log_price', title='log_price par␣
 ↪neighbourhood')
fig.show()
```

[72]:
```
# Box plot des codes postaux et du log_price
fig = px.box(df, x='zipcode', y='log_price', title='log_price par zipcode')
fig.show()
```

[73]:
```
# Nuage de points pour visualiser la relation entre le nombre de salles de␣
 ↪bains et le log_price

fig = px.scatter(df, x='bathrooms', y='log_price', title='log_price et␣
 ↪bathrooms',
                 labels={'bathrooms': 'nombre de salles de bain', 'log_price':␣
 ↪'Log Price'})
fig.show()
```

[74]:
```
# Nuage de points pour visualiser la relation entre le nombre de chambres et le␣
 ↪log_price
fig = px.scatter(df, x='bedrooms', y='log_price', title='log_price et bedrooms',
                 labels={'bedrooms': 'nombre de chambres', 'log_price':␣
 ↪'log_price'})
fig.show()
```

- Pour les amenities :

J'essaie de créer des variables binaires à partir de la colonne 'amenities' pour indiquer la présence ou l'absence de certains équipements (en considérant qu'ils ont la même valeur)

[75]:
```python
# Fonction pour extraire les équipements
def extraire_amenities(amenities_str):
    if isinstance(amenities_str, list):
        return amenities_str
    amenities_liste = amenities_str.strip('{}').split(',')
    amenities_extr = [amenity.strip('" ') for amenity in amenities_liste]
    return amenities_extr

df['amenities'] = df['amenities'].apply(extraire_amenities)
```

[76]:
```python
df['amenities']
```

[76]:
```
0        [TV, Wireless Internet, Kitchen, Free parking …
1        [Wireless Internet, Air conditioning, Kitchen,…
2        [TV, Wireless Internet, Air conditioning, Kitc…
3        [TV, Cable TV, Internet, Wireless Internet, Ai…
4        [TV, Cable TV, Internet, Wireless Internet, Ki…
                                 …
22229                                                  []
22230    [TV, Cable TV, Internet, Wireless Internet, Ki…
22231    [TV, Internet, Wireless Internet, Air conditio…
22232    [TV, Wireless Internet, Air conditioning, Kitc…
22233    [TV, Internet, Wireless Internet, Kitchen, Fre…
Name: amenities, Length: 21931, dtype: object
```

Si j'essaie de compter le nombre d'équipements par logement/offre :

[77]:
```python
df['nb_amenities'] = df['amenities'].apply(lambda x: len(x))
```

[78]:
```python
df['nb_amenities']
```

[78]:
```
0        15
1        25
2        20
3        30
4        24
         ..
22229     1
22230    16
22231    31
22232    15
22233    18
Name: nb_amenities, Length: 21931, dtype: int64
```

[79]:
```
# Visualisation de la relation entre le nombre d'équipements et le log_price␣
↪avec un nuage de points :
fig = px.scatter(df, x='nb_amenities', y='log_price', hover_data=['id'])
fig.show()
```

Je veux ajouter la colonne "nb_amenities" à df_bin tout en supprimant la colonne "amenities" du dataframe d'origine df :

[80]:
```
# Suppression de la colonne "amenities" du DataFrame d'origine
df_temp = df.drop('amenities', axis=1)

categories = ['property_type', 'room_type', 'accommodates', 'bathrooms',␣
↪'bed_type', 'cancellation_policy', 'cleaning_fee', 'city', 'neighbourhood',␣
↪'zipcode', 'bedrooms', 'beds']

# Création des variables dummy à partir des colonnes de categories
df_bin = pd.get_dummies(df_temp, columns=categories)

df_bin['nb_amenities'] = df['nb_amenities']
```

[81]:
```
df_bin.head()
```

[81]:
```
         id  log_price   latitude   longitude  nb_amenities  \
0   5708593   4.317488  33.782712 -118.134410            15
1  14483613   4.007333  40.705468  -73.909439            25
2  10412649   7.090077  38.917537  -77.031651            20
3  17954362   3.555348  40.736001  -73.924248            30
4   9969781   5.480639  37.744896 -122.430665            24

   property_type_Apartment  property_type_Bed & Breakfast  property_type_Boat  \
0                    False                          False               False
1                    False                          False               False
2                     True                          False               False
3                    False                          False               False
4                    False                          False               False

   property_type_Boutique hotel  property_type_Bungalow  …  beds_6.0  \
0                         False                   False  …     False
1                         False                   False  …     False
2                         False                   False  …     False
3                         False                   False  …     False
4                         False                   False  …     False

   beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  beds_12.0  beds_13.0  \
0     False     False     False      False      False      False      False
1     False     False     False      False      False      False      False
2     False     False     False      False      False      False      False
```

```
3     False     False     False      False      False      False      False
4     False     False     False      False      False      False      False

   beds_16.0  beds_18.0
0      False      False
1      False      False
2      False      False
3      False      False
4      False      False

[5 rows x 1348 columns]
```

- Pour les cartes :

J'ai aussi utilisé le fichier GeoJSON contenant le jeu de données entier disponible sur le site suivant : https://public.opendatasoft.com/explore/dataset/georef-united-states-of-america-zc-point/export/?location=2,40.5661,39.98938&basemap=jawg.light

[82]:
```
# Scatter plot des latitude, longitude et du log_price
fig = px.scatter(df, x='longitude', y='latitude', color='log_price',␣
 ↪title='Distribution géographique du log_price', color_continuous_scale=px.
 ↪colors.sequential.Viridis)
fig.show()
```

[83]:
```
# Carte centrée sur les données
map = folium.Map(location=[df['latitude'].mean(), df['longitude'].mean()],␣
 ↪zoom_start=10)

# Marqueur pour chaque point avec une couleur basée sur le prix
for idx, row in df.iterrows():
    folium.CircleMarker(
        location=[row['latitude'], row['longitude']],
        radius=5,
        color='crimson',
        fill=True,
        fill_color='crimson',
        fill_opacity=0.5,
        popup=f"Prix: {row['log_price']}"
    ).add_to(map)

map
```

[83]:
```
<folium.folium.Map at 0x1292ae39490>
```

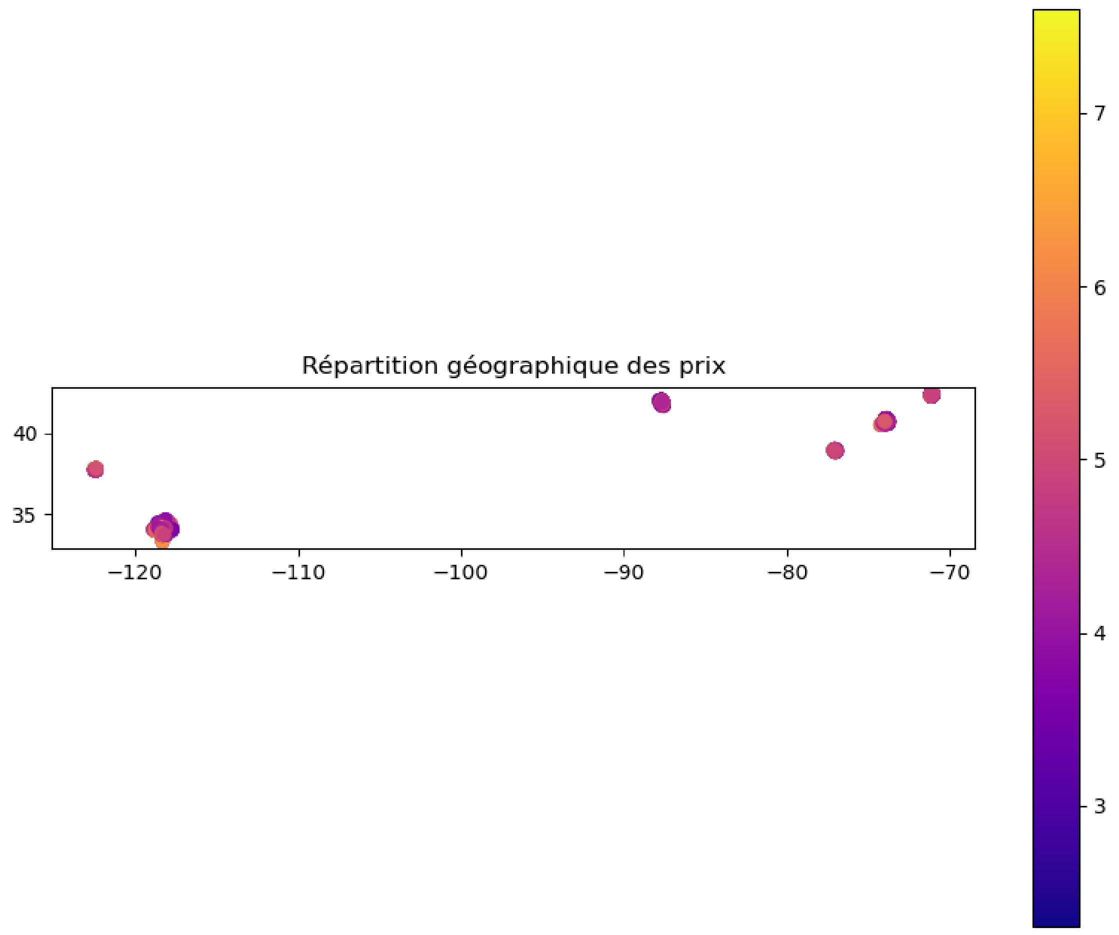
(dézoomer pour voir les Etats-Unis dans son ensemble)

[84]:
```
import geopandas as gpd
```

```
# Conversion du DataFrame en GeoDataFrame
gdf = gpd.GeoDataFrame(
    df, geometry=gpd.points_from_xy(df.longitude, df.latitude)
)

# Création d'une carte de chaleur basée sur les prix
fig, ax = plt.subplots(figsize=(10, 8))
gdf.plot(column='log_price', cmap='plasma', ax=ax, legend=True)
ax.set_title("Répartition géographique des prix")

plt.show()
```



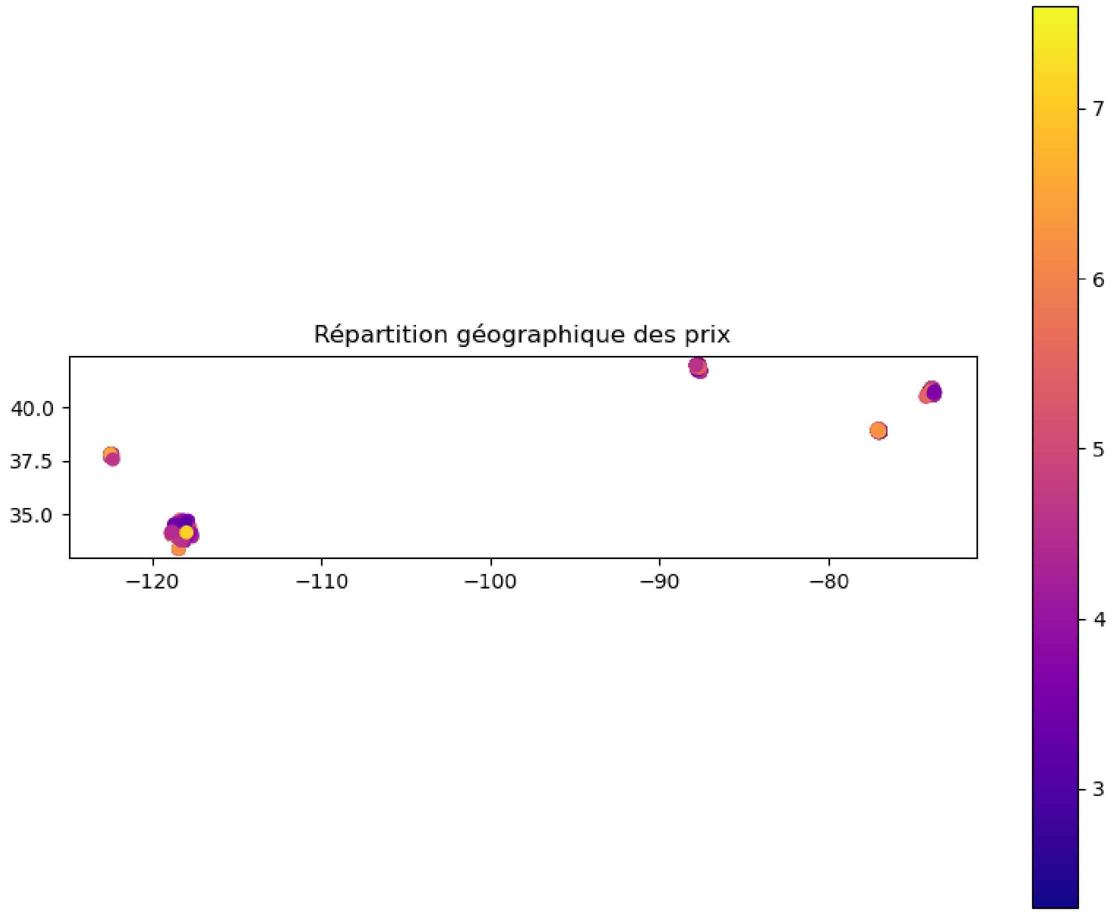
[85]:
```
geo_data = gpd.read_file('usa_zipcodes.geojson')

# Jointure au DataFrame
df = df.merge(geo_data[['zip_code', 'geometry']], left_on='zipcode',␣
 ↪right_on='zip_code')
```

```python
# Conversion en GeoDataFrame
gdf = gpd.GeoDataFrame(df, geometry='geometry')

# Carte
fig, ax = plt.subplots(figsize=(10, 8))
gdf.plot(column='log_price', cmap='plasma', ax=ax, legend=True)
ax.set_title("Répartition géographique des prix")

plt.show()
```



3 II. Partie prédiction

A l'aide du cours et des éléments de https://scikit-learn.org/stable/

3.1 1. Préparation des données

3.1.1 Sélection des données

[86]:
```
X = df_bin.loc[:, df_bin.columns !="log_price"]
y = df_bin["log_price"]  # Prédiction du log du prix
```

[87]:
```
X.head()
```

[87]:
```
         id   latitude   longitude  nb_amenities  property_type_Apartment  \
0   5708593  33.782712 -118.134410            15                    False
1  14483613  40.705468  -73.909439            25                    False
2  10412649  38.917537  -77.031651            20                     True
3  17954362  40.736001  -73.924248            30                    False
4   9969781  37.744896 -122.430665            24                    False

   property_type_Bed & Breakfast  property_type_Boat  \
0                          False               False
1                          False               False
2                          False               False
3                          False               False
4                          False               False

   property_type_Boutique hotel  property_type_Bungalow  property_type_Cabin  \
0                         False                   False                False
1                         False                   False                False
2                         False                   False                False
3                         False                   False                False
4                         False                   False                False

   …  beds_6.0  beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  \
0  …     False     False     False     False      False      False
1  …     False     False     False     False      False      False
2  …     False     False     False     False      False      False
3  …     False     False     False     False      False      False
4  …     False     False     False     False      False      False

   beds_12.0  beds_13.0  beds_16.0  beds_18.0
0      False      False      False      False
1      False      False      False      False
2      False      False      False      False
3      False      False      False      False
4      False      False      False      False

[5 rows x 1347 columns]
```

[88]:
```
y.head()
```

```
[88]: 0    4.317488
      1    4.007333
      2    7.090077
      3    3.555348
      4    5.480639
      Name: log_price, dtype: float64
```

3.1.2 Analyse en Composantes Principales (PCA)

```
[89]: from sklearn.decomposition import PCA

      pca = PCA(n_components=0.95)  # On conserve 95% de la variance
      X_pca = pca.fit_transform(X)
```

3.1.3 Séparation train/test

```
[90]: from sklearn.model_selection import train_test_split

      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
       ↪random_state=42)
```

```
[91]: print ("Dimension pour X_train:", X_train.shape)
      print ("Dimension pour X_test:", X_test.shape)
      print ("Dimension pour y_train:", y_train.shape)
      print ("Dimension pour y_test:", y_test.shape)
```

```
Dimension pour X_train: (17544, 1347)
Dimension pour X_test: (4387, 1347)
Dimension pour y_train: (17544,)
Dimension pour y_test: (4387,)
```

3.1.4 Standardisation

```
[92]: from sklearn.preprocessing import StandardScaler

      scaler_X = StandardScaler()
      X_train_scaled = scaler_X.fit_transform(X_train)
      X_test_scaled = scaler_X.transform(X_test)
```

```
[93]: X_train
```

```
[93]:              id   latitude   longitude  nb_amenities  property_type_Apartment  \
      7240   14609935  40.689492  -73.917084             7                     True
      11169  16002569  42.351714  -71.061075            11                     True
      3226   18449729  34.126614 -118.174707            20                    False
      9379      53579  40.688980  -73.947558            22                     True
      11654   6324190  40.795911  -73.933164            22                     True
```

```
…            …          …           …           …                      …
12144  14100670  40.843931  -73.939488           14                   True
21874  13461289  37.787820 -122.456031            1                   True
5480   10167436  34.104489 -118.373349           10                  False
872    16722017  40.800658  -73.969408            7                   True
16025  13916918  40.809920  -73.958592            5                   True

       property_type_Bed & Breakfast  property_type_Boat  \
7240                           False               False
11169                          False               False
3226                           False               False
9379                           False               False
11654                          False               False
…                                …                 …
12144                          False               False
21874                          False               False
5480                           False               False
872                            False               False
16025                          False               False

       property_type_Boutique hotel  property_type_Bungalow  \
7240                          False                   False
11169                         False                   False
3226                          False                   False
9379                          False                   False
11654                         False                   False
…                               …                     …
12144                         False                   False
21874                         False                   False
5480                          False                   False
872                           False                   False
16025                         False                   False

       property_type_Cabin  …  beds_6.0  beds_7.0  beds_8.0  beds_9.0  \
7240                 False  …     False     False     False     False
11169                False  …     False     False     False     False
3226                 False  …     False     False     False     False
9379                 False  …     False     False     False     False
11654                False  …     False     False     False     False
…                      …  …       …       …       …       …
12144                False  …     False     False     False     False
21874                False  …     False     False     False     False
5480                 False  …     False     False     False     False
872                  False  …     False     False     False     False
16025                False  …     False     False     False     False

       beds_10.0  beds_11.0  beds_12.0  beds_13.0  beds_16.0  beds_18.0
```

```
7240        False      False      False      False      False      False
11169       False      False      False      False      False      False
3226        False      False      False      False      False      False
9379        False      False      False      False      False      False
11654       False      False      False      False      False      False
…           …        …        …        …        …        …
12144       False      False      False      False      False      False
21874       False      False      False      False      False      False
5480        False      False      False      False      False      False
872         False      False      False      False      False      False
16025       False      False      False      False      False      False

[17544 rows x 1347 columns]
```

[94]:
```
X_train_scaled
```

[94]:
```
array([[ 0.55775262,  0.72116905,  0.84459889, …, -0.01307776,
        -0.02826006, -0.00755002],
       [ 0.78653847,  1.26199317,  0.9764602 , …, -0.01307776,
        -0.02826006, -0.00755002],
       [ 1.18856483, -1.414143  , -1.19876596, …, -0.01307776,
        -0.02826006, -0.00755002],
       …,
       [-0.17207368, -1.42134138, -1.20793723, …, -0.01307776,
        -0.02826006, -0.00755002],
       [ 0.90473142,  0.75733826,  0.84218312, …, -0.01307776,
        -0.02826006, -0.00755002],
       [ 0.44390183,  0.76035203,  0.84268249, …, -0.01307776,
        -0.02826006, -0.00755002]])
```

[95]:
```
X_test
```

[95]:
```
             id   latitude   longitude  nb_amenities  property_type_Apartment  \
11319   7919542  38.843454  -76.976974            29                    False
3894    5683405  34.179263 -118.387062            17                    False
12213  19711234  34.093560 -118.273078             7                     True
8802   17112616  34.104154 -118.302914            11                     True
17439  11626795  34.099800 -118.314924            26                     True
…           …        …         …           …                      …
12624  20879290  40.584594  -73.935001            12                    False
10791  17295824  40.689487  -73.927841             5                     True
19388   8909891  40.708304  -74.004147            17                     True
10121  18991641  40.645826  -74.013066            12                     True
12505   7072404  34.023719 -118.487040            12                     True

       property_type_Bed & Breakfast  property_type_Boat  \
11319                          False               False
```

```
3894                          True              False
12213                        False              False
8802                         False              False
17439                        False              False
…                            …                …
12624                        False              False
10791                        False              False
19388                        False              False
10121                        False              False
12505                        False              False

       property_type_Boutique hotel  property_type_Bungalow  \
11319                         False                   False
3894                          False                   False
12213                         False                   False
8802                          False                   False
17439                         False                   False
…                             …                     …
12624                         False                   False
10791                         False                   False
19388                         False                   False
10121                         False                   False
12505                         False                   False

       property_type_Cabin  …  beds_6.0  beds_7.0  beds_8.0  beds_9.0  \
11319                False  …     False     False     False     False
3894                 False  …     False     False     False     False
12213                False  …     False     False     False     False
8802                 False  …     False     False     False     False
17439                False  …     False     False     False     False
…                    …   …       …       …        …       …
12624                False  …     False     False     False     False
10791                False  …     False     False     False     False
19388                False  …     False     False     False     False
10121                False  …     False     False     False     False
12505                False  …     False     False     False     False

       beds_10.0  beds_11.0  beds_12.0  beds_13.0  beds_16.0  beds_18.0
11319      False      False      False      False      False      False
3894       False      False      False      False      False      False
12213      False      False      False      False      False      False
8802       False      False      False      False      False      False
17439      False      False      False      False      False      False
…          …        …        …        …        …        …
12624      False      False      False      False      False      False
10791      False      False      False      False      False      False
19388      False      False      False      False      False      False
```

```
10121      False      False      False      False      False      False
12505      False      False      False      False      False      False

[4387 rows x 1347 columns]
```

[96]:
```
X_test_scaled
```

[96]:
```
array([[-0.54136406,  0.12053842,  0.70332441, …, -0.01307776,
        -0.02826006, -0.00755002],
       [-0.90872297, -1.39701276, -1.20857033, …, -0.01307776,
        -0.02826006, -0.00755002],
       [ 1.39580844, -1.4248973 , -1.20330773, …, -0.01307776,
        -0.02826006, -0.00755002],
       …,
       [-0.37866673,  0.72729001,  0.8405792 , …, -0.01307776,
        -0.02826006, -0.00755002],
       [ 1.27759167,  0.7069618 ,  0.84016743, …, -0.01307776,
        -0.02826006, -0.00755002],
       [-0.68053429, -1.44762106, -1.21318631, …, -0.01307776,
        -0.02826006, -0.00755002]])
```

[97]:
```
scaler_y = StandardScaler()
y_train_scaled = scaler_y.fit_transform(y_train.values.reshape(-1, 1))
y_test_scaled = scaler_y.transform(y_test.values.reshape(-1, 1))
```

[98]:
```
y_train_scaled
```

[98]:
```
array([[-0.8527163 ],
       [-0.26742265],
       [-0.98744698],
       …,
       [-0.25344133],
       [ 0.05698033],
       [-0.56386298]])
```

[99]:
```
y_test_scaled
```

[99]:
```
array([[-1.39553187],
       [-0.32479696],
       [-0.47952626],
       …,
       [ 0.40039475],
       [-0.47952626],
       [ 0.45160777]])
```

3.2 2. Régression linéaire (LinearRegression)

Premier modèle de test : une modélisation de la relation entre la variable à prédire (log_prix) et les variables indépendantes (caractéristiques), en supposant une relation linéaire entre elles

[100]:
```python
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
```

[101]:
```python
# Modèle
regressor_lin = LinearRegression()
```

[102]:
```python
# Entraînement du modèle sur les données d'entraînement
regressor_lin.fit(X_train_scaled, y_train_scaled.ravel())
```

[102]:
```
LinearRegression()
```

[103]:
```python
# Prédictions sur l'ensemble d'entraînement
y_train_pred_lin = regressor_lin.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_lin = regressor_lin.predict(X_test_scaled)
```

[104]:
```python
# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_lin = mean_squared_error(y_train_scaled, y_train_pred_lin)
test_mse_lin = mean_squared_error(y_test_scaled, y_test_pred_lin)

train_rmse_lin = np.sqrt(train_mse_lin)
test_rmse_lin = np.sqrt(test_mse_lin)

train_r2_lin = r2_score(y_train_scaled, y_train_pred_lin)
test_r2_lin = r2_score(y_test_scaled, y_test_pred_lin)

print(f"Entraînement : MSE={train_mse_lin:.3f}, RMSE={train_rmse_lin:.3f},␣
  ↪R²={train_r2_lin:.3f}")
print(f"Test : MSE={test_mse_lin:.3f}, RMSE={test_rmse_lin:.3f},␣
  ↪R²={test_r2_lin:.3f}")
```

```
Entraînement : MSE=0.314, RMSE=0.561, R²=0.686
Test : MSE=67891222065180629262139392.000, RMSE=8239612980303.179,
R²=-68497857738577027335716864.000
```

Plus le score R^2 est proche de 1, plus la prédiction est bonne.

[105]:
```python
# Graphique de résidus : différence entre les valeurs réelles et prédites
residus_lin = y_train - y_train_pred_lin
fig = px.scatter(x=y_train_pred_lin, y=residus_lin, labels={"x": "Valeurs␣
  ↪prédites par la régression linéaire", "y": "Résidus"})
fig.show()
```

[106]:
```
# Distribution des erreurs
fig = px.histogram(residus_lin, nbins=30, labels={"value": "Résidus"},␣
 ↪title="Distribution des erreurs pour la régression linéaire")
fig.show()
```

Conclusion : j'obtiens

- en entraînement : MSE=0.314, RMSE=0.561, R^2=0.686
- en test : MSE=678912220651806292621393920.000, RMSE=8239612980303.179, R^2=-684978577385770273357168640.000

Ici, le modèle s'ajuste très bien aux données d'entraînement, mais en voyant les résultats du test, le modèle semble souffrir d'un sur-apprentissage donc il échoue complètement à généraliser sur les données de test.

3.3 3. Régression par support vector machine (SVR)

L'objectif est de trouver une limite de décision (hyperplan) qui sépare au mieux les jeux de données, en maximisant la marge et en minimisant l'erreur de classification. Cela permet de capturer des relations non linéaires entre les variables et la variable cible (log_prix).

[107]:
```
# Modèle
from sklearn.svm import SVR
regressor = SVR(kernel = 'rbf')
regressor.fit(X_train_scaled, y_train_scaled.ravel())
```

[107]:
```
SVR()
```

(4 min pour charger^)

[108]:
```
# Évaluation des performances sur les ensembles d'entraînement et de test
y_train_pred = regressor.predict(X_train_scaled)
y_test_pred = regressor.predict(X_test_scaled)

train_mse = mean_squared_error(y_train_scaled, y_train_pred)
test_mse = mean_squared_error(y_test_scaled, y_test_pred)
train_rmse = np.sqrt(train_mse)
test_rmse = np.sqrt(test_mse)

train_r2 = r2_score(y_train_scaled, y_train_pred)
test_r2 = r2_score(y_test_scaled, y_test_pred)

print(f"Entraînement : MSE={train_mse:.3f}, RMSE={train_rmse:.3f}, R²={train_r2:
 ↪.3f}")
print(f"Test : MSE={test_mse:.3f}, RMSE={test_rmse:.3f}, R²={test_r2:.3f}")
```

```
Entraînement : MSE=0.314, RMSE=0.561, R²=0.686
Test : MSE=0.432, RMSE=0.658, R²=0.564
```

(+6 min pour charger^)

[109]:
```
# Prédiction d'une nouvelle valeur
new_x = np.zeros(shape=(1, X_train.shape[1]))  # Créer un vecteur de zéros de␣
 ↪la même longueur que X_train
new_x[0, 0] = 6.5  # Remplacer la première caractéristique par 6.5

# Normalisation de la nouvelle donnée
new_x_scaled = scaler_X.transform(new_x)

# La prédiction
y_pred_scaled = regressor.predict(new_x_scaled)
y_pred = scaler_y.inverse_transform(y_pred_scaled.reshape(-1, 1))
print("Log_prix prédit : ", y_pred[0, 0])
```

```
Log_prix prédit :  4.599430606613116
```

```
C:\Users\wwuky\Anaconda\Lib\site-packages\sklearn\base.py:464: UserWarning:

X does not have valid feature names, but StandardScaler was fitted with feature
names
```

On rq si la prédiction du modèle est 0 = 6.5, on obtient : Log_prix prédit : 4.599430606613116

[110]:
```
# Graphique de résidus : différence entre les valeurs réelles et prédites
residus_svr = y_train - y_train_pred
fig = px.scatter(x=y_train_pred, y=residus_svr, labels={"x": "Valeurs prédites␣
 ↪par la SVR", "y": "Résidus"})
fig.show()
```

[111]:
```
# Distribution des erreurs
fig = px.histogram(residus_svr, nbins=30, labels={"value": "Résidus"},␣
 ↪title="Distribution des erreurs pour la SVR")
fig.show()
```

Conclusion : j'obtiens

- en entraînement : MSE=0.314, RMSE=0.561, R^2=0.686
- en test : MSE=0.432, RMSE=0.658, R^2=0.564

Ce modèle semble s'ajuster raisonnablement bien aux données d'entraînement et généralise de manière satisfaisante sur les données de test. Mais il y a une certaine différence de performances entre l'entraînement et le test, ce qui pourrait indiquer un léger sur-apprentissage.

3.3.1 4. Régression par forêt aléatoire (RandomForestRegressor)

C'est un autre algorithme d'apprentissage supervisé utilisé pour résoudre des problèmes de régression, grâce à un ensemble d'arbres ("forêt") de décision entraînés sur des sous-ensembles différents des données d'entraînement, et en utilisant la méthode du bagging (bootstrap aggregating). Il prend en compte les relations non linéaires entre les variables prédictives et la variable cible (log_prix).

[112]:
```
from sklearn.ensemble import RandomForestRegressor
```

[113]:
```
# Modèle
regression_rfr = RandomForestRegressor(n_estimators=100, max_depth=10,␣
 ↪random_state=42)
```

[114]:
```
# Entraînement du modèle sur les données d'entraînement
regression_rfr.fit(X_train_scaled, y_train_scaled.ravel())
```

[114]:
```
RandomForestRegressor(max_depth=10, random_state=42)
```

[115]:
```
# Prédictions sur l'ensemble d'entraînement
y_train_pred_rfr = regression_rfr.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_rfr = regression_rfr.predict(X_test_scaled)
```

[116]:
```
# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_rfr = mean_squared_error(y_train_scaled, y_train_pred_rfr)
test_mse_rfr = mean_squared_error(y_test_scaled, y_test_pred_rfr)
train_rmse_rfr = np.sqrt(train_mse_rfr)
test_rmse_rfr = np.sqrt(test_mse_rfr)

train_r2_rfr = r2_score(y_train_scaled, y_train_pred_rfr)
test_r2_rfr = r2_score(y_test_scaled, y_test_pred_rfr)

print(f"Entraînement : MSE={train_mse_rfr:.3f}, RMSE={train_rmse_rfr:.3f},␣
 ↪R²={train_r2_rfr:.3f}")
print(f"Test : MSE={test_mse_rfr:.3f}, RMSE={test_rmse_rfr:.3f},␣
 ↪R²={test_r2_rfr:.3f}")
```

```
Entraînement : MSE=0.270, RMSE=0.520, R²=0.730
Test : MSE=0.375, RMSE=0.613, R²=0.621
```

[117]:
```
# Graphique de résidus : différence entre les valeurs réelles et prédites
residus_rfr = y_train - y_train_pred_rfr
fig = px.scatter(x=y_train_pred_rfr, y=residus_rfr, labels={"x": "Valeurs␣
 ↪prédites par la régression par forêt aléatoire", "y": "Résidus"})
fig.show()
```

[118]:
```
# Distribution des erreurs
fig = px.histogram(residus_rfr, nbins=30, labels={"value": "Résidus"},␣
 ↪title="Distribution des erreurs pour la régression par forêt aléatoire")
fig.show()
```

Conclusion : j'obtiens

- en entraînement : MSE=0.270, RMSE=0.520, R²=0.730
- en test : MSE=0.375, RMSE=0.613, R²=0.621

Globalement, il semble que ce modèle s'ajuste bien aux données d'entraînement tout en généralisant de manière satisfaisante sur les nouvelles données du test. La différence entre les performances sur l'entraînement et le test n'est pas trop grande, ce qui me paraît être un bon compromis entre biais et variance.

3.3.2 5. Régression ridge (Ridge) (1)

La régression ridge est une méthode de régularisation qui pénalise les coefficients élevés, ce qui réduit la variance du modèle et améliore sa capacité de généralisation, mais avec un petit biais supplémentaire par rapport à la régression linéaire classique.

[119]:
```
from sklearn.linear_model import Ridge
```

[120]:
```
# Modèle
regression_ridge1 = Ridge(alpha=1.0)
```

Test avec alpha = 1

[121]:
```
# Entraînement du modèle sur les données d'entraînement
regression_ridge1.fit(X_train_scaled, y_train_scaled.ravel())
```

[121]:
```
Ridge()
```

[122]:
```
# Prédictions sur l'ensemble d'entraînement
y_train_pred_ridge1 = regression_ridge1.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_ridge1 = regression_ridge1.predict(X_test_scaled)

# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_ridge1 = mean_squared_error(y_train_scaled, y_train_pred_ridge1)
test_mse_ridge1 = mean_squared_error(y_test_scaled, y_test_pred_ridge1)
train_rmse_ridge1 = np.sqrt(train_mse_ridge1)
test_rmse_ridge1 = np.sqrt(test_mse_ridge1)

train_r2_ridge1 = r2_score(y_train_scaled, y_train_pred_ridge1)
test_r2_ridge1 = r2_score(y_test_scaled, y_test_pred_ridge1)

print(f"Entraînement : MSE={train_mse_ridge1:.3f}, RMSE={train_rmse_ridge1:.
 ↪3f}, R²={train_r2_ridge1:.3f}")
print(f"Test : MSE={test_mse_ridge1:.3f}, RMSE={test_rmse_ridge1:.3f},␣
 ↪R²={test_r2_ridge1:.3f}")
```

```
Entraînement : MSE=0.304, RMSE=0.551, R²=0.696
Test : MSE=0.365, RMSE=0.604, R²=0.632
```

[123]:
```
# Graphique de résidus : différence entre les valeurs réelles et prédites
residus_ridge1 = y_train - y_train_pred_ridge1
fig = px.scatter(x=y_train_pred_ridge1, y=residus_ridge1, labels={"x": "Valeurs␣
 ↪prédites par la régression ridge (alpha = 1)", "y": "Résidus"})
fig.show()
```

[124]:
```
# Distribution des erreurs
fig = px.histogram(residus_ridge1, nbins=30, labels={"value": "Résidus"},␣
 ↪title="Distribution des erreurs pour la régression ridge (alpha = 1)")
fig.show()
```

Conclusion : j'obtiens (avec alpha = 1)

- en entraînement : MSE=0.304, RMSE=0.551, R^2=0.696
- en test : MSE=0.365, RMSE=0.604, R^2=0.632

Globalement, il semble que le modèle Ridge s'ajuste de manière satisfaisante aux données d'entraînement et généralise raisonnablement bien sur les nouvelles données du test. La différence entre les performances sur l'entraînement et le test n'est pas trop grande non plus, ce qui semble être un bon compromis entre biais et variance.

3.3.3 Régression ridge (Ridge) (2)

[125]:
```
# Modèle
regression_ridge2 = Ridge(alpha=5) #test avec une autre valeur de alpha
```

[126]:
```
# Entraînement du modèle sur les données d'entraînement
regression_ridge2.fit(X_train_scaled, y_train_scaled.ravel())
```

[126]:
```
Ridge(alpha=5)
```

[127]:
```
# Prédictions sur l'ensemble d'entraînement
y_train_pred_ridge2 = regression_ridge2.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_ridge2 = regression_ridge2.predict(X_test_scaled)

# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_ridge2 = mean_squared_error(y_train_scaled, y_train_pred_ridge2)
test_mse_ridge2 = mean_squared_error(y_test_scaled, y_test_pred_ridge2)
train_rmse_ridge2 = np.sqrt(train_mse_ridge2)
test_rmse_ridge2 = np.sqrt(test_mse_ridge2)

train_r2_ridge2 = r2_score(y_train_scaled, y_train_pred_ridge2)
test_r2_ridge2 = r2_score(y_test_scaled, y_test_pred_ridge2)
```

```
print(f"Entraînement : MSE={train_mse_ridge2:.3f}, RMSE={train_rmse_ridge2:.
↪3f}, R²={train_r2_ridge2:.3f}")
print(f"Test : MSE={test_mse_ridge2:.3f}, RMSE={test_rmse_ridge2:.3f},␣
↪R²={test_r2_ridge2:.3f}")
```

```
Entraînement : MSE=0.304, RMSE=0.551, R²=0.696
Test : MSE=0.364, RMSE=0.604, R²=0.632
```

On obtient sensiblement les mêmes résultats

3.3.4 6. Régression par arbre de décision (DecisionTreeRegressor) (1)

La régression par arbre de décision peut capturer des relations complexes entre les caractéristiques et la variable cible (log_prix), en partitionnant récursivement l'espace des prédicteurs en régions distinctes. Elle prédit la valeur moyenne de la cible dans chaque région. Cependant, le modèle est plus susceptible de sur-apprendre les données d'entraînement si la profondeur de l'arbre n'est pas contrôlée.

[128]:
```
from sklearn.tree import DecisionTreeRegressor
```

[129]:
```
# Modèle
regression_tree = DecisionTreeRegressor(max_depth=5, random_state=42)
```

Test avec une profondeur de 5

[130]:
```
regression_tree.fit(X_train_scaled, y_train_scaled.ravel())
```

[130]:
```
DecisionTreeRegressor(max_depth=5, random_state=42)
```

[131]:
```
# Prédictions sur l'ensemble d'entraînement
y_train_pred_tree = regression_tree.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_tree = regression_tree.predict(X_test_scaled)

# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_tree = mean_squared_error(y_train_scaled, y_train_pred_tree)
test_mse_tree = mean_squared_error(y_test_scaled, y_test_pred_tree)
train_rmse_tree = np.sqrt(train_mse_tree)
test_rmse_tree = np.sqrt(test_mse_tree)

train_r2_tree = r2_score(y_train_scaled, y_train_pred_tree)
test_r2_tree = r2_score(y_test_scaled, y_test_pred_tree)

print(f"Entraînement : MSE={train_mse_tree:.3f}, RMSE={train_rmse_tree:.3f},␣
↪R²={train_r2_tree:.3f}")
```

```
print(f"Test : MSE={test_mse_tree:.3f}, RMSE={test_rmse_tree:.3f},␣
  ↪R²={test_r2_tree:.3f}")
```

```
Entraînement : MSE=0.452, RMSE=0.672, R²=0.548
Test : MSE=0.481, RMSE=0.694, R²=0.514
```

Conclusion : j'obtiens

- en entraînement : MSE=0.452, RMSE=0.672, R^2=0.548
- en test : MSE=0.481, RMSE=0.694, R^2=0.514

Globalement, il semble que ce modèle d'arbre de décision ne s'ajuste pas très bien aux données d'entraînement et ne généralise pas de manière satisfaisante sur les nouvelles données du test.

3.3.5 Régression par arbre de décision (DecisionTreeRegressor) (2)

Avec une profondeur de 10

[132]:
```
regression_tree_2 = DecisionTreeRegressor(max_depth=10, random_state=42)
regression_tree_2.fit(X_train_scaled, y_train_scaled.ravel())

# Prédictions sur l'ensemble d'entraînement
y_train_pred_tree_2 = regression_tree_2.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_tree_2 = regression_tree_2.predict(X_test_scaled)

# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_tree_2 = mean_squared_error(y_train_scaled, y_train_pred_tree_2)
test_mse_tree_2 = mean_squared_error(y_test_scaled, y_test_pred_tree_2)
train_rmse_tree_2 = np.sqrt(train_mse_tree_2)
test_rmse_tree_2 = np.sqrt(test_mse_tree_2)

train_r2_tree_2 = r2_score(y_train_scaled, y_train_pred_tree_2)
test_r2_tree_2 = r2_score(y_test_scaled, y_test_pred_tree_2)

print(f"Entraînement : MSE={train_mse_tree_2:.3f}, RMSE={train_rmse_tree_2:.
  ↪3f}, R²={train_r2_tree_2:.3f}")
print(f"Test : MSE={test_mse_tree_2:.3f}, RMSE={test_rmse_tree_2:.3f},␣
  ↪R²={test_r2_tree_2:.3f}")
```

```
Entraînement : MSE=0.298, RMSE=0.546, R²=0.702
Test : MSE=0.457, RMSE=0.676, R²=0.539
```

3.3.6 Régression par arbre de décision (DecisionTreeRegressor) (3)

Avec une profondeur de 3

[133]:
```
regression_tree_3 = DecisionTreeRegressor(max_depth=3, random_state=42)
regression_tree_3.fit(X_train_scaled, y_train_scaled.ravel())
```

```
# Prédictions sur l'ensemble d'entraînement
y_train_pred_tree_3 = regression_tree_3.predict(X_train_scaled)

# Prédictions sur l'ensemble de test
y_test_pred_tree_3 = regression_tree_3.predict(X_test_scaled)

# Évaluation des performances sur les ensembles d'entraînement et de test
train_mse_tree_3 = mean_squared_error(y_train_scaled, y_train_pred_tree_3)
test_mse_tree_3 = mean_squared_error(y_test_scaled, y_test_pred_tree_3)
train_rmse_tree_3 = np.sqrt(train_mse_tree_3)
test_rmse_tree_3 = np.sqrt(test_mse_tree_3)

train_r2_tree_3 = r2_score(y_train_scaled, y_train_pred_tree_3)
test_r2_tree_3 = r2_score(y_test_scaled, y_test_pred_tree_3)

print(f"Entraînement : MSE={train_mse_tree_3:.3f}, RMSE={train_rmse_tree_3:.
 ↪3f}, R²={train_r2_tree_3:.3f}")
print(f"Test : MSE={test_mse_tree_3:.3f}, RMSE={test_rmse_tree_3:.3f},␣
 ↪R²={test_r2_tree_3:.3f}")
```

```
Entraînement : MSE=0.508, RMSE=0.713, R²=0.492
Test : MSE=0.518, RMSE=0.719, R²=0.478
```

Avec des depths = 10 ou 3, on n'obtient pas de résultat assez satisfaisant non plus : + En augmentant la profondeur maximale de l'arbre à 10, le modèle peut capturer des relations plus complexes entre les caractéristiques et le log du prix. Mais une profondeur trop élevée peut également entraîner un sur-apprentissage. + En définissant la profondeur maximale de l'arbre à 3, l'arbre de décision est assez très simple, avec seulement trois nœuds de décision. Mais une profondeur aussi faible peut entraîner un sous-apprentissage et une mauvaise performance du modèle sur nos données.

[134]:
```
# Graphique de résidus : différence entre les valeurs réelles et prédites
residus_tree = y_train - y_train_pred_tree
fig = px.scatter(x=y_train_pred_tree, y=residus_tree, labels={"x": "Valeurs␣
 ↪prédites par la régression par arbre de décision", "y": "Résidus"})
fig.show()
```

[135]:
```
# Distribution des erreurs
fig = px.histogram(residus_tree, nbins=30, labels={"value": "Résidus"},␣
 ↪title="Distribution des erreurs pour la régression par arbre de décision")
fig.show()
```

3.3.7 Conclusion de la partie

[(E) pour entraînement et (T) pour test]

Régression linéaire : R^2=0.686 (E), R^2 négatif (T), le modèle échoue à généraliser sur les données de test.

SVR : R^2=0.686 (E), R^2=0.564 (T), ce modèle semble moyennement satisfaisant.

Régression par forêt aléatoire : R^2=0.730 (E), R^2=0.621 (T), ce modèle semble assez satisfaisant.

Ridge : R^2=0.696 (E), R^2=0.632 (T), ce modèle semble assez satisfaisant également, mais moins qu'avec la précédente.

Arbre de décision : R^2=0.548 (E) , R^2=0.514 (T), ce modèle semble moyennement satisfaisant.

Par manque de temps, je n'ai pas pu tester d'autres modèles cités sur le site de la documentation de scikitlearn (Régression ElasticNet, régression par gradient boosting…) meme si j'aurais bien voulu…

=> Le modèle de régression par forêt aléatoire me paraît ici être le plus efficace.

3.3.8 6. Prédiction sur le fichier de test

[136]:
```
airbnb_test = pd.read_csv("airbnb_test.csv")
```

[137]:
```
airbnb_test.head()
```

[137]:
```
   Unnamed: 0 property_type        room_type  \
0    14282777     Apartment  Entire home/apt
1    17029381     Apartment  Entire home/apt
2     7824740     Apartment  Entire home/apt
3    19811650         House  Entire home/apt
4    12410741     Apartment  Entire home/apt

                                           amenities  accommodates  bathrooms  \
0  {"Wireless Internet","Air conditioning",Kitche…             3        1.0
1  {"Wireless Internet","Air conditioning",Kitche…             7        1.0
2  {TV,"Cable TV","Wireless Internet","Air condit…             5        1.0
3  {TV,"Cable TV",Internet,"Wireless Internet",Ki…             4        1.0
4  {TV,Internet,"Wireless Internet","Air conditio…             2        1.0

   bed_type cancellation_policy  cleaning_fee city  … last_review  \
0  Real Bed              strict          True  NYC  …  2016-07-18
1  Real Bed              strict          True  NYC  …  2017-09-23
2  Real Bed            moderate          True  NYC  …  2017-09-14
3  Real Bed            flexible          True   SF  …         NaN
4  Real Bed            moderate          True   DC  …  2017-01-22

    latitude   longitude                                    name  \
0  40.696524  -73.991617            Beautiful brownstone 1-bedroom
1  40.766115  -73.989040  Superb 3BR Apt Located Near Times Square
2  40.808110  -73.943756                           The Garden Oasis
3  37.772004 -122.431619         Beautiful Flat in the Heart of SF!
4  38.925627  -77.034596                  Great studio in midtown DC

     neighbourhood number_of_reviews review_scores_rating  zipcode  bedrooms  \
0  Brooklyn Heights                 2                100.0    11201       1.0
```

```
1    Hell's Kitchen                  6                   93.0     10019        3.0
2            Harlem                 10                   92.0     10027        1.0
3      Lower Haight                  0                    NaN   94117.0        2.0
4  Columbia Heights                  4                   40.0     20009        0.0

   beds
0   1.0
1   3.0
2   3.0
3   2.0
4   1.0

[5 rows x 27 columns]
```

[138]:
```
airbnb_test.shape
```

[138]:
```
(51877, 27)
```

Comme j'avais un problème de longueur de données, je vérifie au fur et à mesure :

[139]:
```
print(f"Nombre initial de lignes dans airbnb_test: {len(airbnb_test)}")
```

```
Nombre initial de lignes dans airbnb_test: 51877
```

Même traitement que le fichier entraînement :

[140]:
```
airbnb_test = airbnb_test.drop(['description', 'first_review',␣
 ↪'host_has_profile_pic',
                                'host_identity_verified', 'host_response_rate',␣
 ↪'host_since',
                                'instant_bookable', 'last_review', 'name',␣
 ↪'number_of_reviews',
                                'review_scores_rating'], axis=1)

#airbnb_test = airbnb_test.dropna(subset=['zipcode']) c'est ici que j'avais le␣
 ↪changement de len...!!
```

[141]:
```
print(f"Nombre de lignes après suppression des NA dans 'zipcode':␣
 ↪{len(airbnb_test)}")
```

```
Nombre de lignes après suppression des NA dans 'zipcode': 51877
```

[142]:
```
categories = ['property_type', 'room_type', 'accommodates', 'bathrooms',␣
 ↪'bed_type', 'cancellation_policy', 'cleaning_fee', 'city', 'neighbourhood',␣
 ↪'zipcode', 'bedrooms', 'beds']
airbnb_test_bin = pd.get_dummies(airbnb_test, columns=categories)
```

[143]:
```
colonnes = ['property_type', 'room_type', 'cleaning_fee', 'city',␣
↪'neighbourhood', 'zipcode', 'bedrooms', 'beds']
```

[144]:
```
# Appliquer la fonction extraire_amenities à la colonne 'amenities'
airbnb_test['amenities'] = airbnb_test['amenities'].apply(extraire_amenities)

# pour avoir le nombre d'équipements par offre
airbnb_test['nb_amenities'] = airbnb_test['amenities'].apply(lambda x: len(x))
```

De même qu'avec df_bin, je veux ajouter la colonne "nb_amenities" à airbnb_test_bin, tout en supprimant la colonne "amenities" du dataframe d'origine airbnb_test :

[145]:
```
# Supprimer la colonne "amenities" du fichier d'origine
airbnb_temp = airbnb_test.drop('amenities', axis=1)
```

[146]:
```
# Créer les variables dummy à partir des colonnes catégorielles
airbnb_test_bin = pd.get_dummies(airbnb_temp, columns=categories)

# Ajouter la colonne "nb_amenities" à df_bin
airbnb_test_bin['nb_amenities'] = airbnb_test['nb_amenities']
```

[147]:
```
X.head()
```

[147]:
```
         id   latitude   longitude  nb_amenities  property_type_Apartment  \
0   5708593  33.782712 -118.134410            15                    False
1  14483613  40.705468  -73.909439            25                    False
2  10412649  38.917537  -77.031651            20                     True
3  17954362  40.736001  -73.924248            30                    False
4   9969781  37.744896 -122.430665            24                    False

   property_type_Bed & Breakfast  property_type_Boat  \
0                          False               False
1                          False               False
2                          False               False
3                          False               False
4                          False               False

   property_type_Boutique hotel  property_type_Bungalow  property_type_Cabin  \
0                         False                   False                False
1                         False                   False                False
2                         False                   False                False
3                         False                   False                False
4                         False                   False                False

   …  beds_6.0  beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  \
0  …     False     False     False     False      False      False
1  …     False     False     False     False      False      False
```

```
2  …     False     False     False     False      False      False
3  …     False     False     False     False      False      False
4  …     False     False     False     False      False      False

   beds_12.0  beds_13.0  beds_16.0  beds_18.0
0      False      False      False      False
1      False      False      False      False
2      False      False      False      False
3      False      False      False      False
4      False      False      False      False

[5 rows x 1347 columns]
```

[148]:
```
airbnb_test_bin.head()
```

[148]:
```
   Unnamed: 0   latitude   longitude  nb_amenities  property_type_Apartment  \
0    14282777  40.696524  -73.991617             9                     True
1    17029381  40.766115  -73.989040            15                     True
2     7824740  40.808110  -73.943756            19                     True
3    19811650  37.772004 -122.431619            15                    False
4    12410741  38.925627  -77.034596            12                     True

   property_type_Bed & Breakfast  property_type_Boat  \
0                          False               False
1                          False               False
2                          False               False
3                          False               False
4                          False               False

   property_type_Boutique hotel  property_type_Bungalow  property_type_Cabin  \
0                         False                   False                False
1                         False                   False                False
2                         False                   False                False
3                         False                   False                False
4                         False                   False                False

   …  beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  beds_12.0  \
0  …     False     False     False      False      False      False
1  …     False     False     False      False      False      False
2  …     False     False     False      False      False      False
3  …     False     False     False      False      False      False
4  …     False     False     False      False      False      False

   beds_13.0  beds_14.0  beds_15.0  beds_16.0
0      False      False      False      False
1      False      False      False      False
2      False      False      False      False
```

```
3      False      False      False      False
4      False      False      False      False

[5 rows x 1450 columns]
```

On remarque que X a 1347 columns, et airbnb_test_bin a 1449 columns : mais RandomForestRegressor attend 1347 caractéristiques en input.

J'ai besoin de supprimer les colonnes de airbnb_test_bin qui n'ont pas le même nom que dans X, et d'ajouter celles qui manquent en utilisant une méthode d'imputation pour remplacer les valeurs manquantes par des valeurs estimées (la moyenne des colonnes correspondantes par exemple).

[149]:
```
from sklearn.impute import SimpleImputer

# Obtenir les noms de colonnes communes
colonnes_communes = list(set(X.columns) & set(airbnb_test_bin.columns))

# Supprimer les colonnes de airbnb_test_bin qui ne sont pas dans la liste des␣
 ↪colonnes communes
airbnb_test_bin = airbnb_test_bin[colonnes_communes]

# Ajout des colonnes manquantes à airbnb_test_bin et remplacer les valeurs␣
 ↪manquantes par la moyenne
imputer = SimpleImputer(strategy='mean')
```

[150]:
```
missing_cols = X.drop(columns=colonnes_communes)
print(f"Ajout des colonnes manquantes: {missing_cols.columns}")

airbnb_test_bin = pd.concat([airbnb_test_bin, X.
 ↪drop(columns=colonnes_communes)], axis=1)
airbnb_test_bin_imputed = imputer.fit_transform(airbnb_test_bin)

# Remplacer airbnb_test_bin par les données imputées
airbnb_test_bin = pd.DataFrame(airbnb_test_bin_imputed, columns=airbnb_test_bin.
 ↪columns)
```

```
Ajout des colonnes manquantes: Index(['id', 'property_type_Island',
'neighbourhood_Arboretum',
       'neighbourhood_Commerce', 'neighbourhood_Gerritsen Beach',
       'neighbourhood_Grant City', 'neighbourhood_Harvard Square',
       'neighbourhood_Irwindale', 'neighbourhood_La Habra',
       'neighbourhood_Magnificent Mile', 'neighbourhood_Montclare',
       'neighbourhood_Near Northeast', 'neighbourhood_New Dorp Beach',
       'neighbourhood_Newton', 'neighbourhood_Rolling Hills',
       'neighbourhood_Rossville', 'neighbourhood_South El Monte',
       'neighbourhood_Takoma Park, MD', 'neighbourhood_Watertown',
       'neighbourhood_Woodland', 'zipcode_02138', 'zipcode_02186',
       'zipcode_02458', 'zipcode_02472', 'zipcode_10004.0', 'zipcode_10007.0',
       'zipcode_10012.0', 'zipcode_10162', 'zipcode_10279', 'zipcode_10282.0',
       'zipcode_10307.0', 'zipcode_10308.0', 'zipcode_10309', 'zipcode_10704',
       'zipcode_11001', 'zipcode_11209.0', 'zipcode_11215.0',
       'zipcode_11239.0', 'zipcode_11411', 'zipcode_11412.0',
       'zipcode_11429.0', 'zipcode_11509.0', 'zipcode_1m', 'zipcode_20912',
       'zipcode_60603', 'zipcode_60633', 'zipcode_60660-1448', 'zipcode_60805',
       'zipcode_7302.0', 'zipcode_90034-2203', 'zipcode_90035-4475',
       'zipcode_90036-2514', 'zipcode_9004', 'zipcode_90222',
       'zipcode_90403-2638', 'zipcode_91377', 'zipcode_91708', 'zipcode_91786',
       'zipcode_91802', 'zipcode_94401', 'beds_18.0'],
      dtype='object')
```

[151]:
```
airbnb_test_bin.head()
```

[151]:
```
   property_type_Cave  zipcode_11228  property_type_Serviced apartment  \
0                 0.0            0.0                               0.0
1                 0.0            0.0                               0.0
2                 0.0            0.0                               0.0
3                 0.0            0.0                               0.0
4                 0.0            0.0                               0.0

   zipcode_90022  zipcode_10312  neighbourhood_Rego Park  zipcode_10459.0  \
0            0.0            0.0                      0.0              0.0
1            0.0            0.0                      0.0              0.0
2            0.0            0.0                      0.0              0.0
3            0.0            0.0                      0.0              0.0
4            0.0            0.0                      0.0              0.0

   zipcode_90245  zipcode_90755  bathrooms_2.0  …  zipcode_90036-2514  \
0            0.0            0.0            0.0  …                 0.0
1            0.0            0.0            0.0  …                 0.0
2            0.0            0.0            0.0  …                 0.0
3            0.0            0.0            0.0  …                 0.0
4            0.0            0.0            0.0  …                 0.0

   zipcode_9004  zipcode_90222  zipcode_90403-2638  zipcode_91377  \
0           0.0            0.0                 0.0            0.0
1           0.0            0.0                 0.0            0.0
2           0.0            0.0                 0.0            0.0
3           0.0            0.0                 0.0            0.0
4           0.0            0.0                 0.0            0.0

   zipcode_91708  zipcode_91786  zipcode_91802  zipcode_94401  beds_18.0
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
3            0.0            0.0            0.0            0.0        0.0
```

```
4            0.0            0.0            0.0            0.0        0.0

[5 rows x 1347 columns]
```

```
[152]: airbnb_test_bin.shape
```

```
[152]: (51877, 1347)
```

Mais il faut que les noms des colonnes correspondent à ceux du dataframe de comparaison, et dans le même ordre

```
[153]: # Ordre des colonnes de X (en excluant les colonnes 'id')
x_column_order = [col for col in X.columns if col != 'id']

# Réorganisation des colonnes de airbnb_test_bin
airbnb_test_bin = airbnb_test_bin[x_column_order]
```

```
[154]: # Première colonne
first_column = airbnb_test_bin.iloc[:, 0]

# Ajout de la première colonne à gauche du DataFrame
airbnb_test_bin.insert(0, 'id', first_column)

airbnb_test_bin.rename(columns={'id': 'id'}, inplace=True)
```

Au lieu de None, il faut des False

```
[155]: # Remplacer les valeurs None par False dans airbnb_test_bin
airbnb_test_bin = airbnb_test_bin.fillna(False)
```

```
[156]: airbnb_test_bin.head()
```

```
[156]:         id   latitude   longitude  nb_amenities  property_type_Apartment  \
0  40.696524  40.696524  -73.991617           9.0                      1.0
1  40.766115  40.766115  -73.989040          15.0                      1.0
2  40.808110  40.808110  -73.943756          19.0                      1.0
3  37.772004  37.772004 -122.431619          15.0                      0.0
4  38.925627  38.925627  -77.034596          12.0                      1.0

   property_type_Bed & Breakfast  property_type_Boat  \
0                            0.0                 0.0
1                            0.0                 0.0
2                            0.0                 0.0
3                            0.0                 0.0
4                            0.0                 0.0

   property_type_Boutique hotel  property_type_Bungalow  property_type_Cabin  \
0                           0.0                     0.0                  0.0
```

```
1                          0.0                     0.0                   0.0
2                          0.0                     0.0                   0.0
3                          0.0                     0.0                   0.0
4                          0.0                     0.0                   0.0

   …  beds_6.0  beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  \
0  …       0.0       0.0       0.0       0.0        0.0        0.0
1  …       0.0       0.0       0.0       0.0        0.0        0.0
2  …       0.0       0.0       0.0       0.0        0.0        0.0
3  …       0.0       0.0       0.0       0.0        0.0        0.0
4  …       0.0       0.0       0.0       0.0        0.0        0.0

   beds_12.0  beds_13.0  beds_16.0  beds_18.0
0        0.0        0.0        0.0        0.0
1        0.0        0.0        0.0        0.0
2        0.0        0.0        0.0        0.0
3        0.0        0.0        0.0        0.0
4        0.0        0.0        0.0        0.0

[5 rows x 1347 columns]
```

[157]:
```
X.head()
```

[157]:
```
         id   latitude   longitude  nb_amenities  property_type_Apartment  \
0   5708593  33.782712 -118.134410            15                    False
1  14483613  40.705468  -73.909439            25                    False
2  10412649  38.917537  -77.031651            20                     True
3  17954362  40.736001  -73.924248            30                    False
4   9969781  37.744896 -122.430665            24                    False

   property_type_Bed & Breakfast  property_type_Boat  \
0                          False               False
1                          False               False
2                          False               False
3                          False               False
4                          False               False

   property_type_Boutique hotel  property_type_Bungalow  property_type_Cabin  \
0                         False                   False                False
1                         False                   False                False
2                         False                   False                False
3                         False                   False                False
4                         False                   False                False

   …  beds_6.0  beds_7.0  beds_8.0  beds_9.0  beds_10.0  beds_11.0  \
0  …     False     False     False     False      False      False
1  …     False     False     False     False      False      False
```

```
2  …     False     False     False     False      False      False
3  …     False     False     False     False      False      False
4  …     False     False     False     False      False      False

   beds_12.0  beds_13.0  beds_16.0  beds_18.0
0      False      False      False      False
1      False      False      False      False
2      False      False      False      False
3      False      False      False      False
4      False      False      False      False

[5 rows x 1347 columns]
```

[158]:
```
airbnb_test_bin.shape
```

[158]:
```
(51877, 1347)
```

[159]:
```
# Réentrainer rfr
regression_rfr.fit(X_train, y_train)
```

[159]:
```
RandomForestRegressor(max_depth=10, random_state=42)
```

[160]:
```
y_prediction_finale = regression_rfr.predict(airbnb_test_bin)
```

[161]:
```
print(y_prediction_finale)
```

```
[4.97902898 5.69309137 4.96575722 … 5.21462967 4.25130552 5.23377134]
```

[162]:
```
y_prediction_finale.shape
```

[162]:
```
(51877,)
```

Sauvegarde dans le fichier de prédiction :

[163]:
```
# Récupérer la première colonne de airbnb_test_bin
premiere_col_airbnb_test = airbnb_test.iloc[:, 0]

# Convertir les valeurs en entiers et ensuite en chaînes de caractères
premiere_col_airbnb_test = premiere_col_airbnb_test.astype(str)

# Créer un DataFrame à partir de la première colonne
df_premiere_col = pd.DataFrame(premiere_col_airbnb_test)

# Créer un DataFrame à partir des prédictions
df_predictions = pd.DataFrame(y_prediction_finale)
```

```
[164]: # Vérifier les longueurs avant sauvegarde
       assert len(df_premiere_col) == len(df_predictions), f"Les longueurs des␣
        ↪DataFrames ne correspondent pas: {len(df_premiere_col)} //␣
        ↪{len(df_predictions)}"
```

```
[165]: # Concaténer les deux DataFrames côte à côte
       prediction_fichier = pd.concat([df_premiere_col, df_predictions], axis=1)
```

```
[166]: # Renommer les colonnes
       prediction_fichier.columns = ['', 'logpred']

       # Sauvegarder les prédictions dans un fichier CSV
       prediction_fichier.to_csv("mes_predictions.csv", index=False) # index=False␣
        ↪pour éviter d'ajouter l'index interne à pandas
```

```
[167]: prediction_fichier.head()
```

```
[167]:               logpred
       0  14282777  4.979029
       1  17029381  5.693091
       2   7824740  4.965757
       3  19811650  5.410258
       4  12410741  4.823811
```

Test du fichier :

```
[168]: def estConforme(monFichier_csv):
           votre_prediction = pd.read_csv("mes_predictions.csv")

           fichier_exemple = pd.read_csv("prediction_example.csv")

           assert votre_prediction.columns[1] == fichier_exemple.columns[1],␣
        ↪f"Attention, votre colonne de prédiction doit s'appeler {fichier_exemple.
        ↪columns[1]}, elle s'appelle '{votre_prediction.columns[1]}'"
           assert len(votre_prediction) == len(fichier_exemple), f"Attention, vous␣
        ↪devriez avoir {len(fichier_exemple)} prédictions dans votre fichier, il en␣
        ↪contient {len(votre_prediction)}"

           assert np.all(votre_prediction.iloc[:,0] == fichier_exemple.iloc[:, 0])

           print("Fichier conforme!")

       estConforme("mes_predictions.csv")
```

```
Fichier conforme!
```

```
[ ]:
```