IMPORTING LIBRARIES

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

LOADING DATASET

```
data = pd.read_csv("Airbnb datasets eda.csv")

data.head()
```

```
             id                                               name
host_id  \
0  1.312228e+06         Rental unit in Brooklyn · ★5.0 · 1 bedroom
7130382
1  4.527754e+07  Rental unit in New York · ★4.67 · 2 bedrooms ·...
51501835
2  9.710000e+17  Rental unit in New York · ★4.17 · 1 bedroom · ...
528871354
3  3.857863e+06  Rental unit in New York · ★4.64 · 1 bedroom · ...
19902271
4  4.089661e+07  Condo in New York · ★4.91 · Studio · 1 bed · 1...
61391963

            host_name neighbourhood_group       neighbourhood
latitude  \
0              Walter            Brooklyn        Clinton Hill
40.683710
1            Jeniffer           Manhattan      Hell's Kitchen
40.766610
2              Joshua           Manhattan             Chelsea
40.750764
3  John And Catherine           Manhattan  Washington Heights
40.835600
4      Stay With Vibe           Manhattan         Murray Hill
40.751120

   longitude        room_type  price  ...  last_review
reviews_per_month  \
0 -73.964610     Private room   55.0  ...     20/12/15
0.03
1 -73.988100  Entire home/apt  144.0  ...     1/5/2023
0.24
2 -73.994605  Entire home/apt  187.0  ...     18/12/23
```

```
1.67
3 -73.942500     Private room  120.0  ...     17/09/23
1.38
4 -73.978600  Entire home/apt   85.0  ...    3/12/2023
0.24

  calculated_host_listings_count  availability_365
number_of_reviews_ltm  \
0                            1.0               0.0
0.0
1                          139.0             364.0
2.0
2                            1.0             343.0
6.0
3                            2.0             363.0
12.0
4                          133.0             335.0
3.0

      license  rating bedrooms beds           baths
0  No License       5        1    1  Not specified
1  No License    4.67        2    1              1
2      Exempt    4.17        1    2              1
3  No License    4.64        1    1              1
4  No License    4.91   Studio    1              1

[5 rows x 22 columns]
```

```
data.tail()
```

```
                 id                                               name
\
20765  2.473690e+07  Rental unit in New York · ★4.75 · 1 bedroom · ...

20766  2.835711e+06  Rental unit in New York · ★4.46 · 1 bedroom · ...

20767  5.182527e+07  Rental unit in New York · ★4.93 · 1 bedroom · ...

20768  7.830000e+17  Rental unit in New York · ★5.0 · 1 bedroom · 1...

20769  5.660000e+17  Rental unit in Queens · ★4.89 · 1 bedroom · 1 ...


          host_id host_name neighbourhood_group      neighbourhood
latitude  \
20765  186680487   Henry D           Manhattan    Lower East Side
40.711380
20766    3237504     Aspen           Manhattan  Greenwich Village
40.730580
20767  304317395      Jeff           Manhattan     Hell's Kitchen
40.757350
```

```
20768  163083101   Marissa          Manhattan         Chinatown
40.713750
20769   93827372   Glenroy             Queens          Rosedale
40.658874

       longitude        room_type  price  ...  last_review
reviews_per_month  \
20765 -73.991560     Private room   45.0  ...     29/09/23
1.81
20766 -74.000700  Entire home/apt  105.0  ...     1/7/2023
0.48
20767 -73.993430  Entire home/apt  299.0  ...    8/12/2023
2.09
20768 -73.991470  Entire home/apt  115.0  ...     17/09/23
0.91
20769 -73.728651     Private room  102.0  ...   10/12/2023
4.50

      calculated_host_listings_count  availability_365
number_of_reviews_ltm  \
20765                            1.0             157.0
12.0
20766                            1.0               0.0
1.0
20767                            1.0               0.0
27.0
20768                            1.0             363.0
7.0
20769                            1.0               0.0
62.0

                 license  rating bedrooms beds baths
20765         No License    4.75        1    1     1
20766         No License    4.46        1    2     1
20767         No License    4.93        1    1     1
20768         No License       5        1    1     1
20769  OSE-STRREG-0000513    4.89        1    1     1

[5 rows x 22 columns]
```

EDA

DESCRIPTION OF DATA

```
data.shape
```

```
(20770, 22)

# ROWS  = 20770
# Column = 22

data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20770 entries, 0 to 20769
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              20770 non-null  float64
 1   name                            20770 non-null  object 
 2   host_id                         20770 non-null  int64  
 3   host_name                       20770 non-null  object 
 4   neighbourhood_group             20770 non-null  object 
 5   neighbourhood                   20763 non-null  object 
 6   latitude                        20763 non-null  float64
 7   longitude                       20763 non-null  float64
 8   room_type                       20763 non-null  object 
 9   price                           20736 non-null  float64
 10  minimum_nights                  20763 non-null  float64
 11  number_of_reviews               20763 non-null  float64
 12  last_review                     20763 non-null  object 
 13  reviews_per_month               20763 non-null  float64
 14  calculated_host_listings_count  20763 non-null  float64
 15  availability_365                20763 non-null  float64
 16  number_of_reviews_ltm           20763 non-null  float64
 17  license                         20770 non-null  object 
 18  rating                          20770 non-null  object 
 19  bedrooms                        20770 non-null  object 
 20  beds                            20770 non-null  int64  
 21  baths                           20770 non-null  object 
dtypes: float64(10), int64(2), object(10)
memory usage: 3.5+ MB

#STATISTICAL SUMMARY
data.describe()

                 id       host_id      latitude     longitude  
price  \
count  2.077000e+04  2.077000e+04  20763.000000  20763.000000  
20736.000000   
mean   3.033858e+17  1.749049e+08     40.726821    -73.939179  
187.714940   
std    3.901221e+17  1.725657e+08      0.060293      0.061403  
1023.245124   
min    2.595000e+03  1.678000e+03     40.500314    -74.249840  
10.000000   
```

```
25%    2.707260e+07  2.041184e+07     40.684159    -73.980755
80.000000
50%    4.992852e+07  1.086990e+08     40.722890    -73.949597
125.000000
75%    7.220000e+17  3.143997e+08     40.763106    -73.917475
199.000000
max    1.050000e+18  5.504035e+08     40.911147    -73.713650
100000.000000

       minimum_nights  number_of_reviews  reviews_per_month  \
count    20763.000000       20763.000000       20763.000000
mean        28.558493          42.610605           1.257589
std         33.532697          73.523401           1.904472
min          1.000000           1.000000           0.010000
25%         30.000000           4.000000           0.210000
50%         30.000000          14.000000           0.650000
75%         30.000000          49.000000           1.800000
max       1250.000000        1865.000000          75.490000

       calculated_host_listings_count  availability_365  \
count                    20763.000000      20763.000000
mean                        18.866686        206.067957
std                         70.921443        135.077259
min                          1.000000          0.000000
25%                          1.000000         87.000000
50%                          2.000000        215.000000
75%                          5.000000        353.000000
max                        713.000000        365.000000

       number_of_reviews_ltm          beds
count           20763.000000  20770.000000
mean               10.848962      1.723592
std                21.354876      1.211993
min                 0.000000      1.000000
25%                 1.000000      1.000000
50%                 3.000000      1.000000
75%                15.000000      2.000000
max              1075.000000     42.000000
```

DATA CLEANING

```
# No of missing value in each column

data.isnull().sum()
```

```
id                                  0
name                                0
host_id                             0
```

```
host_name                         0
neighbourhood_group               0
neighbourhood                     7
latitude                          7
longitude                         7
room_type                         7
price                            34
minimum_nights                    7
number_of_reviews                 7
last_review                       7
reviews_per_month                 7
calculated_host_listings_count    7
availability_365                  7
number_of_reviews_ltm             7
license                           0
rating                            0
bedrooms                          0
beds                              0
baths                             0
dtype: int64
```

DROPING MISSING VALUES

```
data.dropna(inplace = True)
```

```
data.isnull().sum()
```

```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
number_of_reviews_ltm             0
license                           0
rating                            0
bedrooms                          0
beds                              0
```

```
baths                             0
dtype: int64
```

```
data.shape
```

```
(20736, 22)
```

```
data.size
```

```
456192
```

DEALING WITH DUPLICATES

```
data.duplicated().sum()
```

```
12
```

```
#DELETING ALL DUPLICATES
data.drop_duplicates(inplace=True)
```

```
data.duplicated().sum()
# ALL DUPLICATED DATAIS DELETED
```

```
0
```

```
data.dtypes
```

```
id                                float64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                             float64
minimum_nights                    float64
number_of_reviews                 float64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count    float64
availability_365                  float64
number_of_reviews_ltm             float64
license                            object
rating                             object
bedrooms                           object
beds                                int64
baths                              object
dtype: object
```

```
data["id"] = data["id"].astype(object)

# DATA TYPE OF ID CHANGED TO OBJECT
data.dtypes
```

```
id                                 object
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                             float64
minimum_nights                    float64
number_of_reviews                 float64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count    float64
availability_365                  float64
number_of_reviews_ltm             float64
license                            object
rating                             object
bedrooms                           object
beds                                int64
baths                              object
dtype: object
```

DATA ANALYSIS

UNIVARIATE ANALYSIS

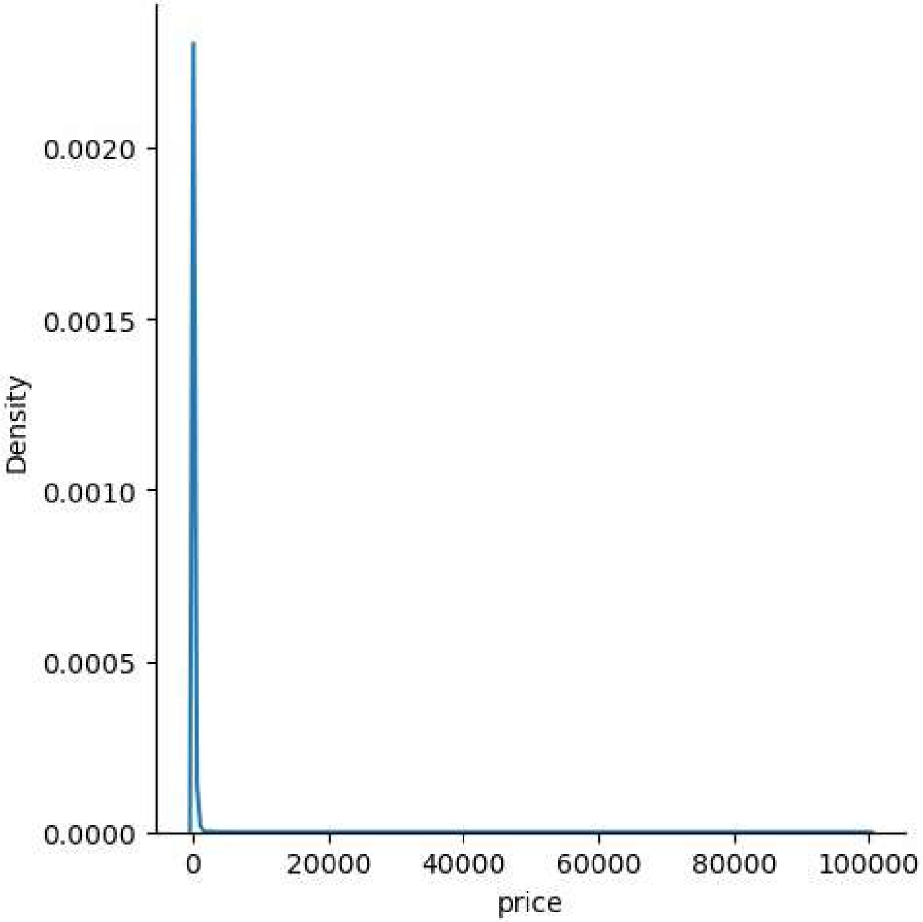
```
# price distribution
data["price"]
```

```
0         55.0
1        144.0
2        187.0
3        120.0
4         85.0
         ...
20765     45.0
20766    105.0
20767    299.0
20768    115.0
20769    102.0
Name: price, Length: 20724, dtype: float64
```

```
sns.displot(data = data,x="price",kind= "kde")
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: 
FutureWarning: use_inf_as_na option is deprecated and will be removed 
in a future version. Convert inf values to NaN before operating 
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

```
<seaborn.axisgrid.FacetGrid at 0x2d56128ad90>
```

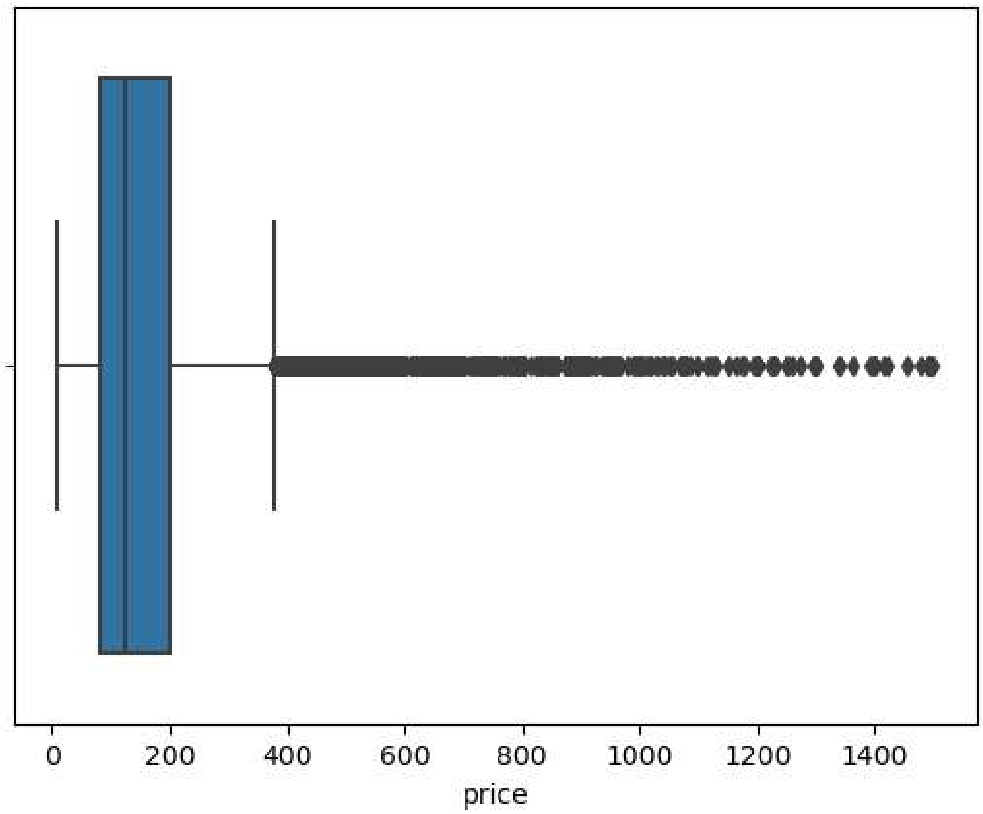


IDENTIFYING OUTLIERS IN PRICE

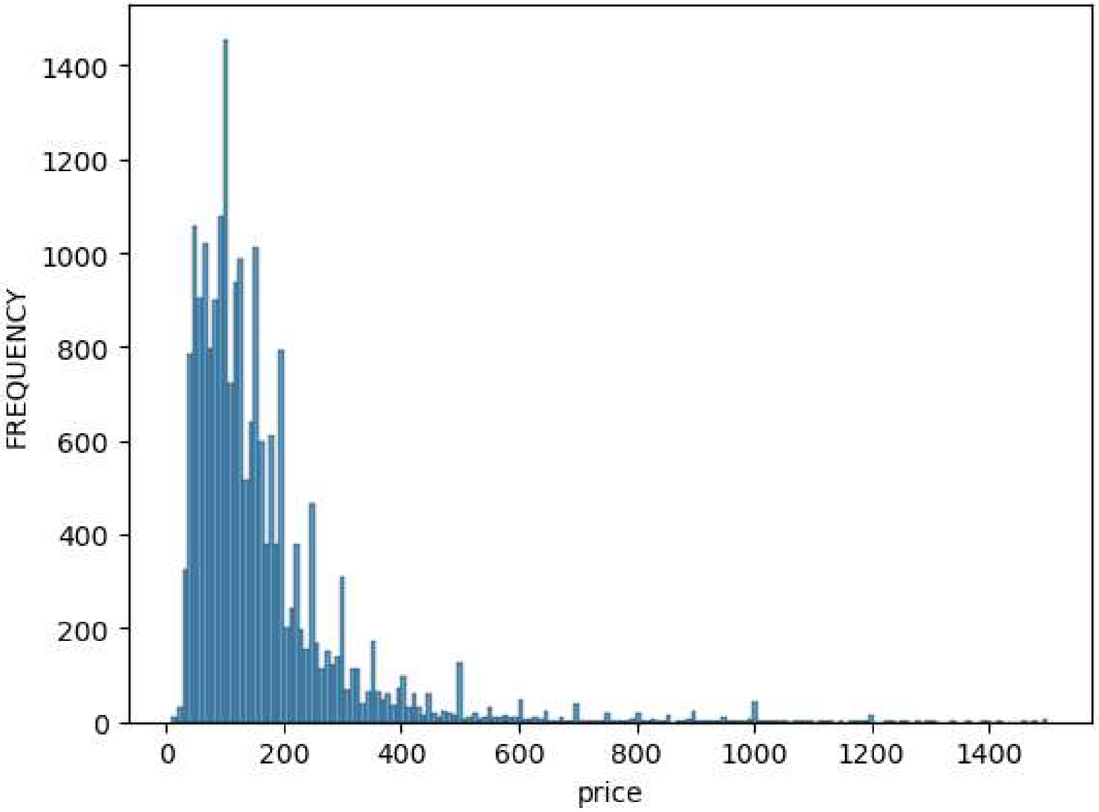
```
df = data[data["price"]<1500]
```

```
sns.boxplot(data = df, x ="price")
```

```
<Axes: xlabel='price'>
```



```
sns.histplot(data = df, x= "price")
plt.ylabel("FREQUENCY")

C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):

Text(0, 0.5, 'FREQUENCY')
```



```
data.columns

Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'number_of_reviews_ltm', 'license',
'rating',
       'bedrooms', 'beds', 'baths'],
      dtype='object')

df.groupby(by = "neighbourhood_group")["price"].mean()

neighbourhood_group
Bronx              58.195946
Brooklyn           72.730111
Manhattan         104.756748
Queens             55.370492
Staten Island      46.255817
Name: price, dtype: float64

df.groupby(by = "neighbourhood_group")["price"].mean()

neighbourhood_group
Bronx              56.132077
```

```
Brooklyn          69.407032
Manhattan        100.395811
Queens            52.915439
Staten Island     43.926998
Name: price, dtype: float64
```

```
df.head()
```

```
                     id
name  \
0             1312228.0         Rental unit in Brooklyn · ★5.0 · 1
bedroom
1            45277537.0  Rental unit in New York · ★4.67 · 2 bedrooms
·...
2  971000000000000000.0  Rental unit in New York · ★4.17 · 1 bedroom ·
...
3             3857863.0  Rental unit in New York · ★4.64 · 1 bedroom ·
...
4            40896611.0  Condo in New York · ★4.91 · Studio · 1 bed ·
1...

     host_id           host_name neighbourhood_group
neighbourhood  \
0    7130382              Walter            Brooklyn        Clinton
Hill
1   51501835            Jeniffer           Manhattan      Hell's
Kitchen
2  528871354              Joshua           Manhattan
Chelsea
3   19902271  John And Catherine           Manhattan  Washington
Heights
4   61391963      Stay With Vibe           Manhattan         Murray
Hill

    latitude  longitude        room_type  price  ...  last_review  \
0  40.683710 -73.964610     Private room   55.0  ...     20/12/15
1  40.766610 -73.988100  Entire home/apt  144.0  ...     1/5/2023
2  40.750764 -73.994605  Entire home/apt  187.0  ...     18/12/23
3  40.835600 -73.942500     Private room  120.0  ...     17/09/23
4  40.751120 -73.978600  Entire home/apt   85.0  ...    3/12/2023

   reviews_per_month calculated_host_listings_count  availability_365
\
0               0.03                            1.0               0.0

1               0.24                          139.0             364.0

2               1.67                            1.0             343.0

3               1.38                            2.0             363.0
```

```
4               0.24                          133.0             335.0

   number_of_reviews_ltm     license  rating bedrooms beds
baths
0                    0.0  No License       5        1    1  Not
specified
1                    2.0  No License    4.67        2    1
1
2                    6.0      Exempt    4.17        1    2
1
3                   12.0  No License    4.64        1    1
1
4                    3.0  No License    4.91   Studio    1
1

[5 rows x 22 columns]
```

```
df.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'number_of_reviews_ltm', 'license',
'rating',
       'bedrooms', 'beds', 'baths'],
      dtype='object')
```

FEATURE ENGINEERING

```
# Price per bed
df["price"] = df["price"]/df["beds"]
df.head()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_13064\622403169.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  df["price"] = df["price"]/df["beds"]
```

```
                     id
name  \
0             1312228.0         Rental unit in Brooklyn · ★5.0 · 1
```

```
bedroom
1           45277537.0  Rental unit in New York · ★4.67 · 2 bedrooms
·...
2  971000000000000000.0  Rental unit in New York · ★4.17 · 1 bedroom ·
...
3             3857863.0  Rental unit in New York · ★4.64 · 1 bedroom ·
...
4           40896611.0  Condo in New York · ★4.91 · Studio · 1 bed ·
1...

     host_id           host_name neighbourhood_group
neighbourhood  \
0    7130382              Walter            Brooklyn        Clinton
Hill
1   51501835            Jeniffer           Manhattan      Hell's
Kitchen
2  528871354              Joshua           Manhattan
Chelsea
3   19902271  John And Catherine           Manhattan  Washington
Heights
4   61391963      Stay With Vibe           Manhattan         Murray
Hill

    latitude  longitude        room_type     price  ...
last_review  \
0  40.683710 -73.964610     Private room   55.0000  ...     20/12/15

1  40.766610 -73.988100  Entire home/apt  144.0000  ...     1/5/2023

2  40.750764 -73.994605  Entire home/apt   11.6875  ...     18/12/23

3  40.835600 -73.942500     Private room  120.0000  ...     17/09/23

4  40.751120 -73.978600  Entire home/apt   85.0000  ...    3/12/2023


   reviews_per_month calculated_host_listings_count  availability_365
\
0               0.03                            1.0               0.0

1               0.24                          139.0             364.0

2               1.67                            1.0             343.0

3               1.38                            2.0             363.0

4               0.24                          133.0             335.0


   number_of_reviews_ltm     license  rating bedrooms beds
baths
```

```
0                    0.0  No License       5        1    1  Not
specified
1                    2.0  No License    4.67        2    1
1
2                    6.0      Exempt    4.17        1    2
1
3                   12.0  No License    4.64        1    1
1
4                    3.0  No License    4.91   Studio    1
1

[5 rows x 22 columns]
```

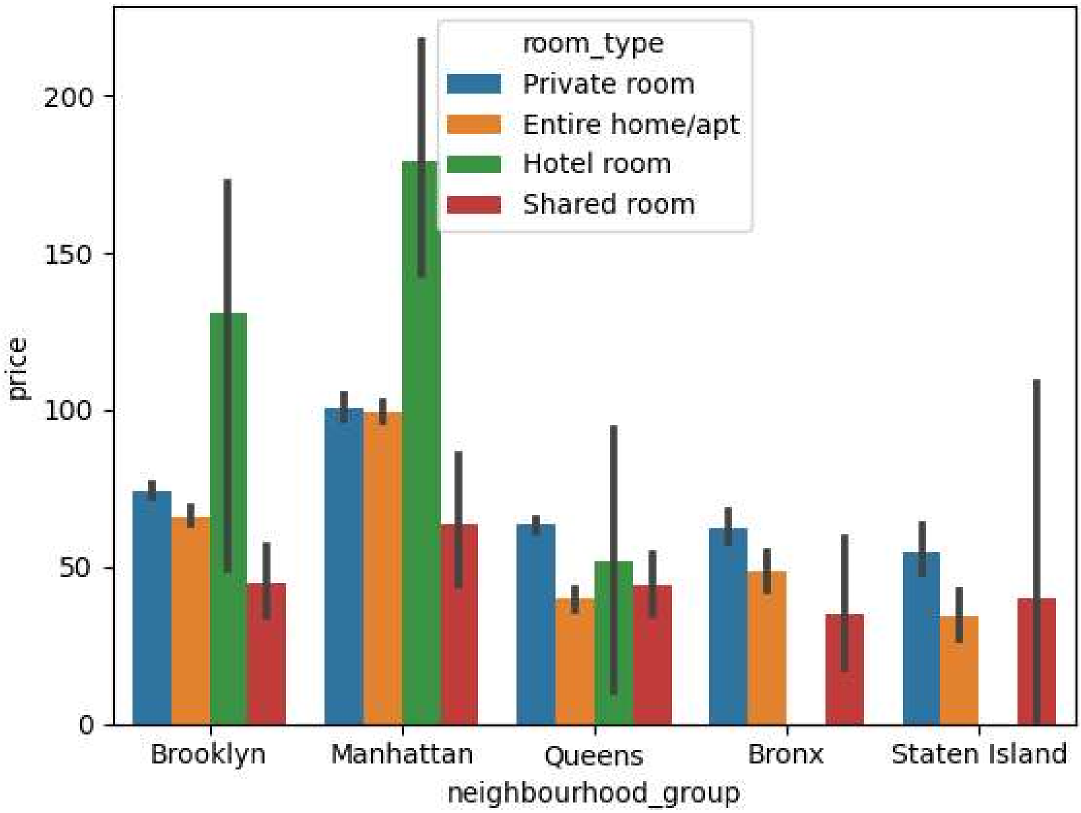
BI-VARIATE ANALYSIS

```
df.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'number_of_reviews_ltm', 'license',
'rating',
       'bedrooms', 'beds', 'baths'],
      dtype='object')
```

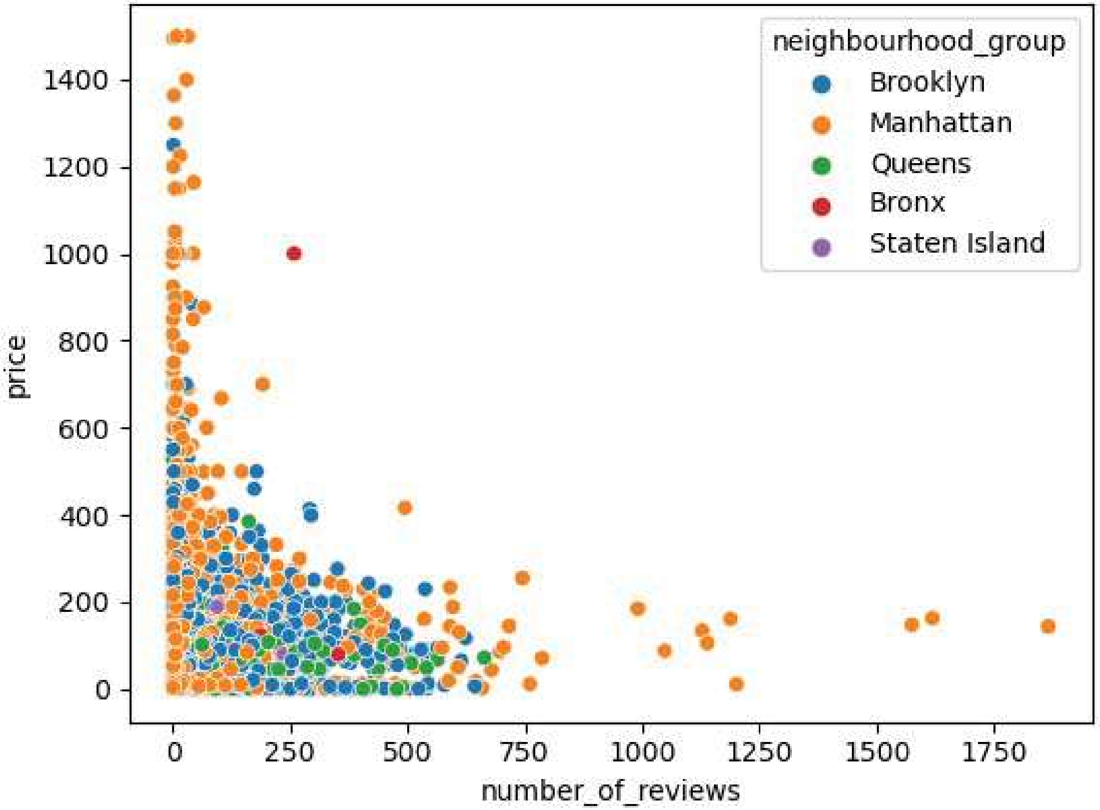
```
# price dependency on neighbourhood
sns.barplot(data= df,x= "neighbourhood_group",y =
"price",hue="room_type")
```

```
<Axes: xlabel='neighbourhood_group', ylabel='price'>
```



```
# Number of reviews
sns.scatterplot(data = df,x = "number_of_reviews",y = "price",hue
="neighbourhood_group")
```

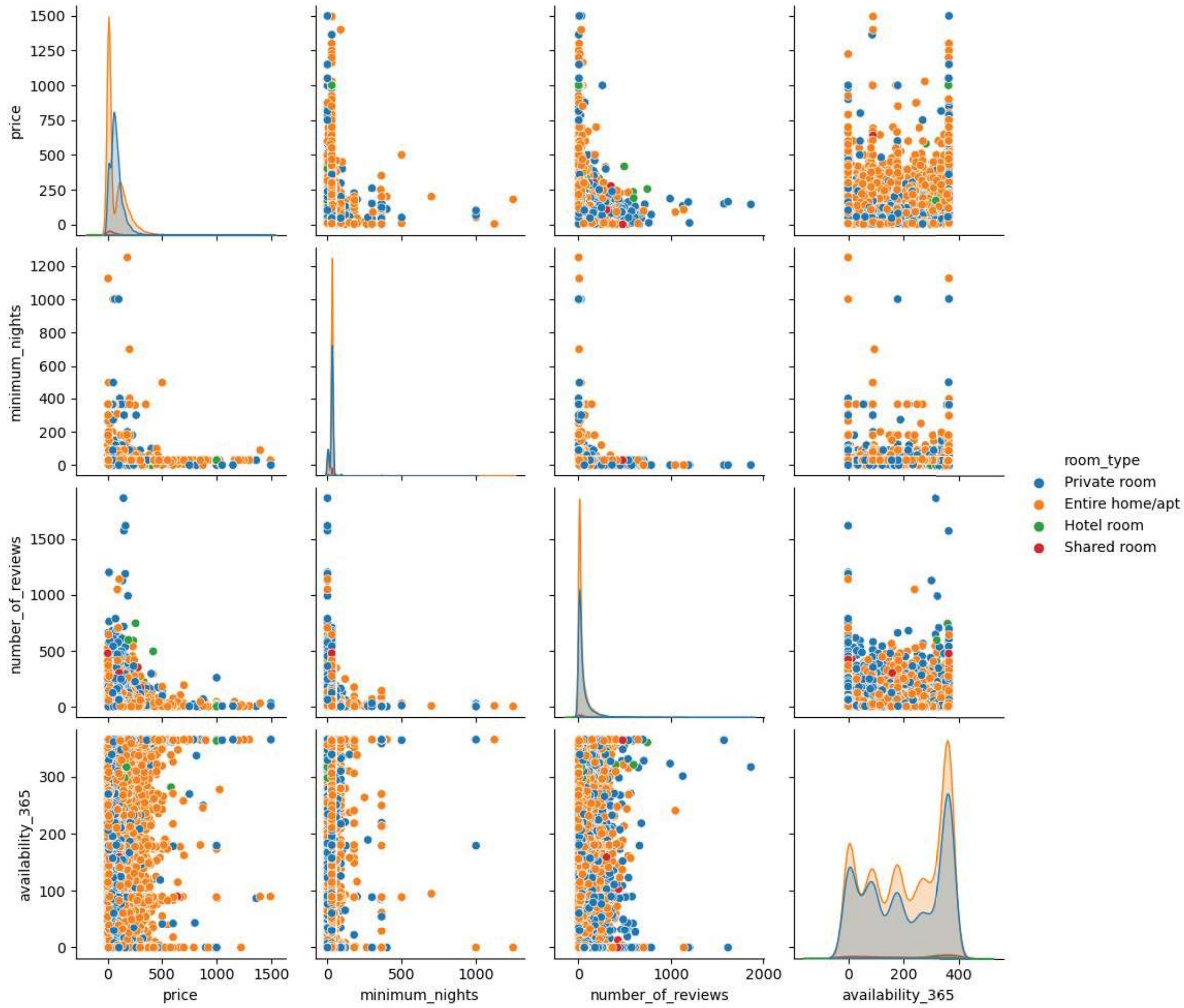
```
<Axes: xlabel='number_of_reviews', ylabel='price'>
```



```
sns.pairplot(data = df, vars =
(["price","minimum_nights","number_of_reviews","availability_365"]),hu
e= "room_type")
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating
instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

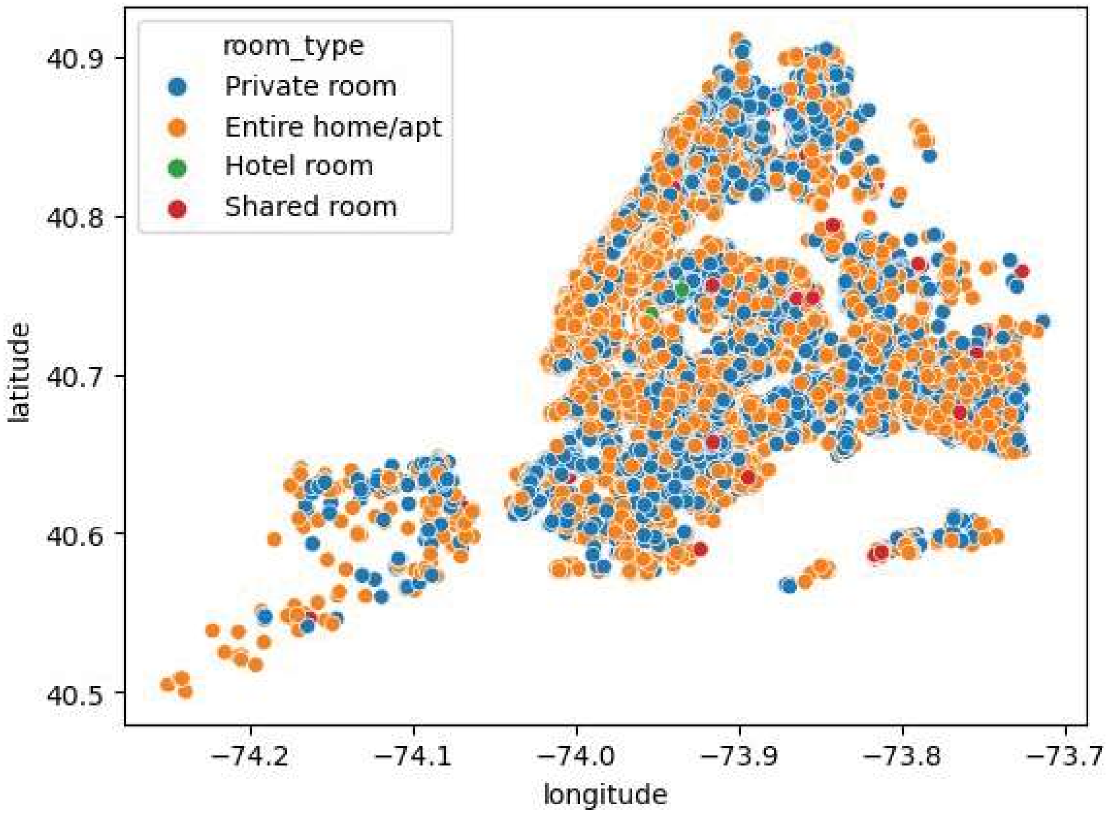
```
<seaborn.axisgrid.PairGrid at 0x2d57f7b2a50>
```



Geographical distribution of Airbnb listings

```
sns.scatterplot(data=df,x="longitude",y= "latitude", hue ="room_type")
```

```
<Axes: xlabel='longitude', ylabel='latitude'>
```



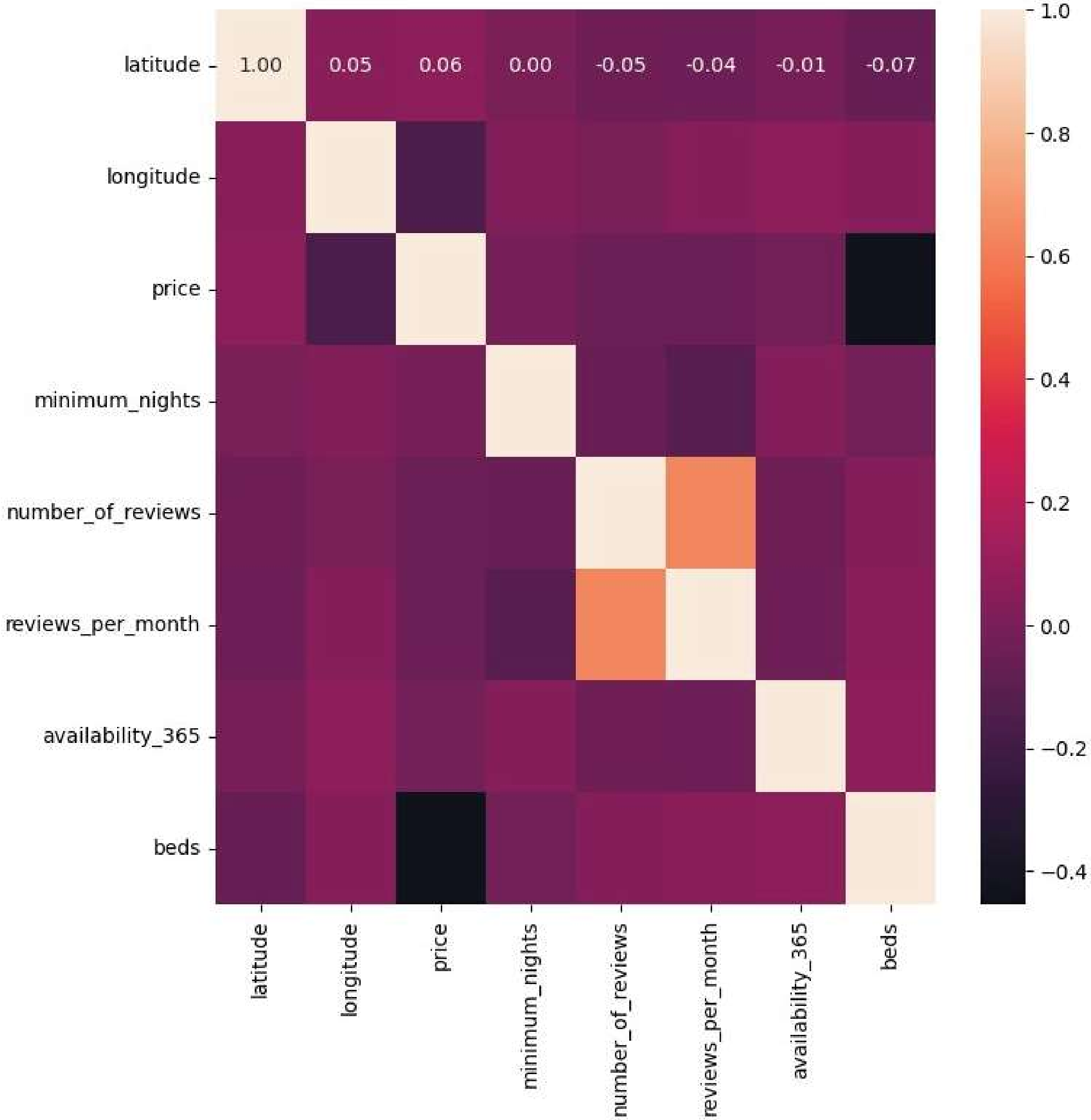
correlation between differnt variable

```
corr =
df[["latitude","longitude","price","minimum_nights","number_of_reviews
","reviews_per_month","availability_365","beds"]].corr()

corr

                   latitude  longitude     price  minimum_nights  \
latitude           1.000000   0.047369  0.063063        0.004590   
longitude          0.047369   1.000000 -0.163829        0.023890   
price              0.063063  -0.163829  1.000000       -0.012361   
minimum_nights     0.004590   0.023890 -0.012361        1.000000   
number_of_reviews -0.047953   0.004820 -0.055789       -0.059049   
reviews_per_month -0.041673   0.041720 -0.054788       -0.122509   
availability_365  -0.005941   0.063523 -0.026115        0.035466   
beds              -0.071753   0.041832 -0.453699       -0.025852   

                   number_of_reviews  reviews_per_month
availability_365  \
latitude                   -0.047953          -0.041673         -
0.005941  
longitude                   0.004820           0.041720
```

```
0.063523
price                      -0.055789          -0.054788         -
0.026115
minimum_nights             -0.059049          -0.122509
0.035466
number_of_reviews           1.000000           0.631005         -
0.049656
reviews_per_month           0.631005           1.000000         -
0.040116
availability_365           -0.049656          -0.040116
1.000000
beds                        0.040071           0.053496
0.065985

                       beds
latitude          -0.071753
longitude          0.041832
price             -0.453699
minimum_nights    -0.025852
number_of_reviews  0.040071
reviews_per_month  0.053496
availability_365   0.065985
beds               1.000000
```

```
plt.figure(figsize = (8,8))
sns.heatmap(data = corr, annot=True,fmt = ".2f")
```

```
<Axes: >
```


latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
availability_365
beds
1.00
0.05
0.06
0.00
-0.05
-0.04
-0.01
-0.07
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
availability_365
beds
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4