Exploratory Data Analysis-AirBnb

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

df = pd.read_csv(r"C:\Users\adars\OneDrive\Pictures\Documents\
AirBnb.csv")

df.head(1)
```

```
          id                                          name  host_id
host_name  \
0  1312228.0  Rental unit in Brooklyn · ★5.0 · 1 bedroom  7130382
Walter

  neighbourhood_group neighbourhood  latitude  longitude     room_type
price  \
0            Brooklyn  Clinton Hill  40.68371  -73.96461  Private room
55.0

   ...  last_review  reviews_per_month calculated_host_listings_count
\
0  ...     20/12/15               0.03                             1.0


   availability_365  number_of_reviews_ltm     license  rating
bedrooms beds  \
0               0.0                    0.0  No License       5
1    1

           baths
0  Not specified

[1 rows x 22 columns]



df.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
```

```
df.shape
```

```
(20770, 22)
```

```
df.isnull()
```

```
          id   name  host_id  host_name  neighbourhood_group
neighbourhood  \
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
...      ...    ...      ...        ...                  ...
...
20765  False  False    False      False                False
False
20766  False  False    False      False                False
False
20767  False  False    False      False                False
False
20768  False  False    False      False                False
```

```
False
20769  False  False    False      False                False
False

       latitude  longitude  room_type  price  ...  last_review  \
0         False      False      False  False  ...        False
1         False      False      False  False  ...        False
2         False      False      False  False  ...        False
3         False      False      False  False  ...        False
4         False      False      False  False  ...        False
...         ...        ...        ...    ...  ...          ...
20765     False      False      False  False  ...        False
20766     False      False      False  False  ...        False
20767     False      False      False  False  ...        False
20768     False      False      False  False  ...        False
20769     False      False      False  False  ...        False

       reviews_per_month  calculated_host_listings_count
availability_365  \
0                  False                           False
False
1                  False                           False
False
2                  False                           False
False
3                  False                           False
False
4                  False                           False
False
...                  ...                             ...
...
20765              False                           False
False
20766              False                           False
False
20767              False                           False
False
20768              False                           False
False
20769              False                           False
False

       number_of_reviews_ltm  license  rating  bedrooms   beds  baths

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

...                     ...      ...     ...       ...    ...    ...

20765                 False    False   False     False  False  False

20766                 False    False   False     False  False  False

20767                 False    False   False     False  False  False

20768                 False    False   False     False  False  False

20769                 False    False   False     False  False  False


[20770 rows x 22 columns]
```

```
df.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

```
df.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
```

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

```
df.duplicated()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
20765    False
20766    False
20767    False
20768    False
20769    False
Length: 20736, dtype: bool
```

```
#Return Duplicate Values

df.duplicated().sum()
```

```
0
```

```
#droping duplicates
df.drop_duplicates(inplace=True)
```

```
df.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
```

```
#Changing Data Types

df['id']=df['id'].astype(object)
df['host_id']=df['host_id'].astype(object)
df.dtypes
```

```
id                                 object
name                               object
host_id                            object
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
```

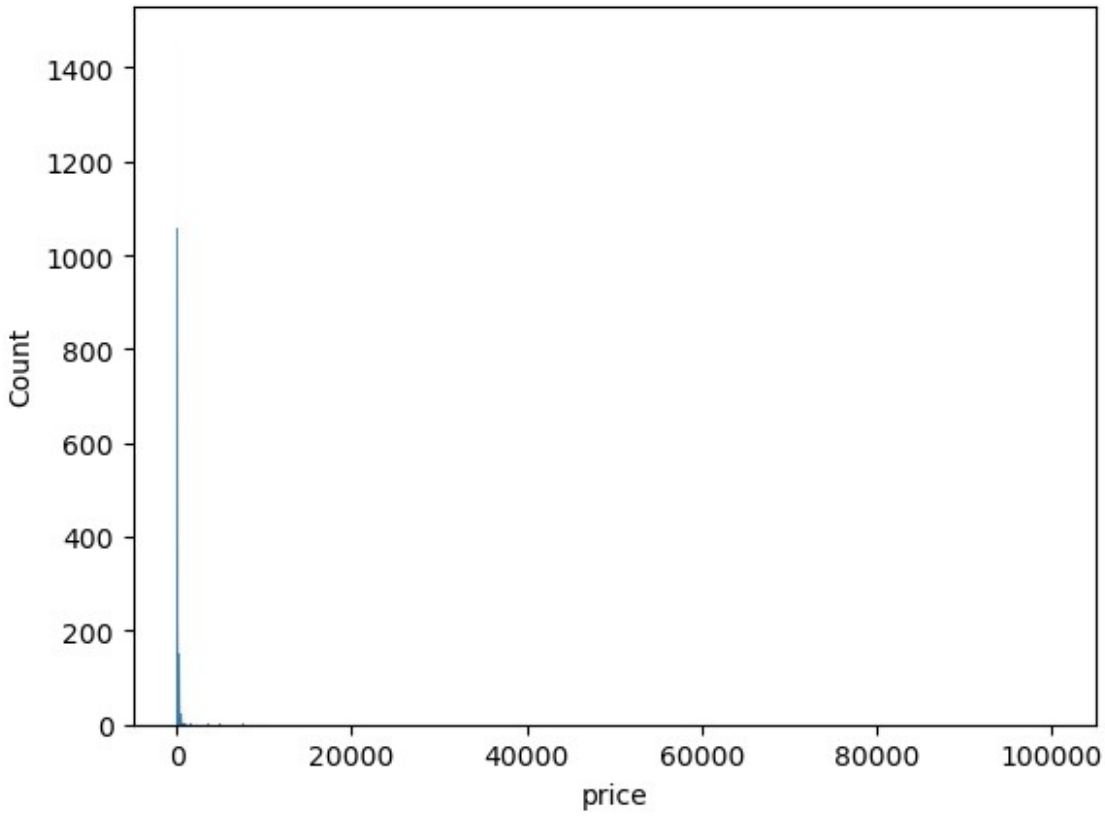
Data Analysis

Univariante Analysis

```
import matplotlib.pyplot as plt
import seaborn as sns
```

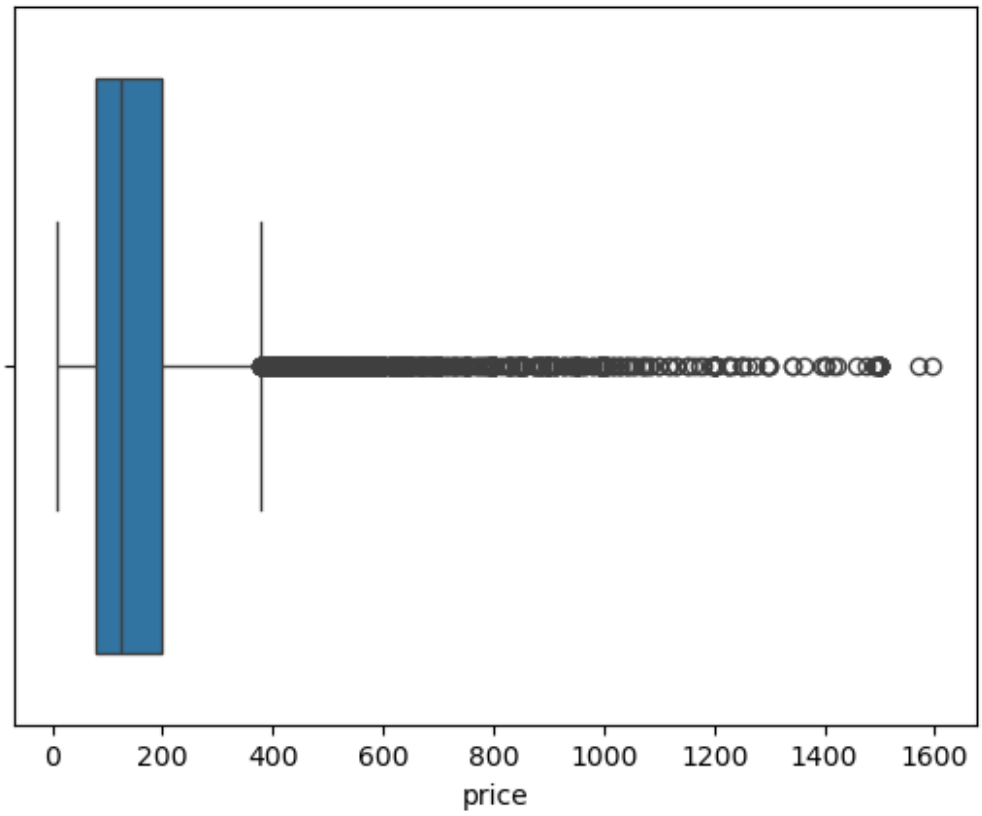
Price Distribution

```
sns.histplot(data=df, x='price')
plt.figure(figsize=(5, 4))
plt.show()
```



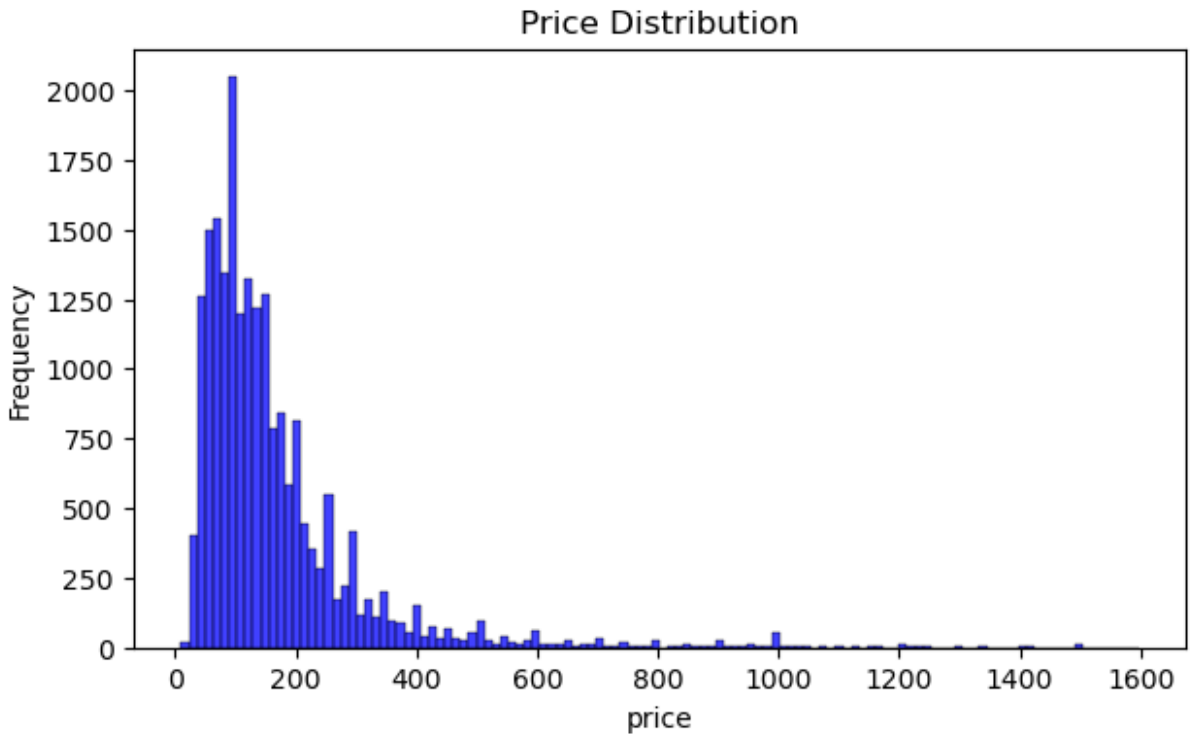

```
<Figure size 500x400 with 0 Axes>
```

```
sns.boxplot(data=df, x='price')
plt.figure(figsize=(5, 4))
plt.show()
# Outliers in price
```



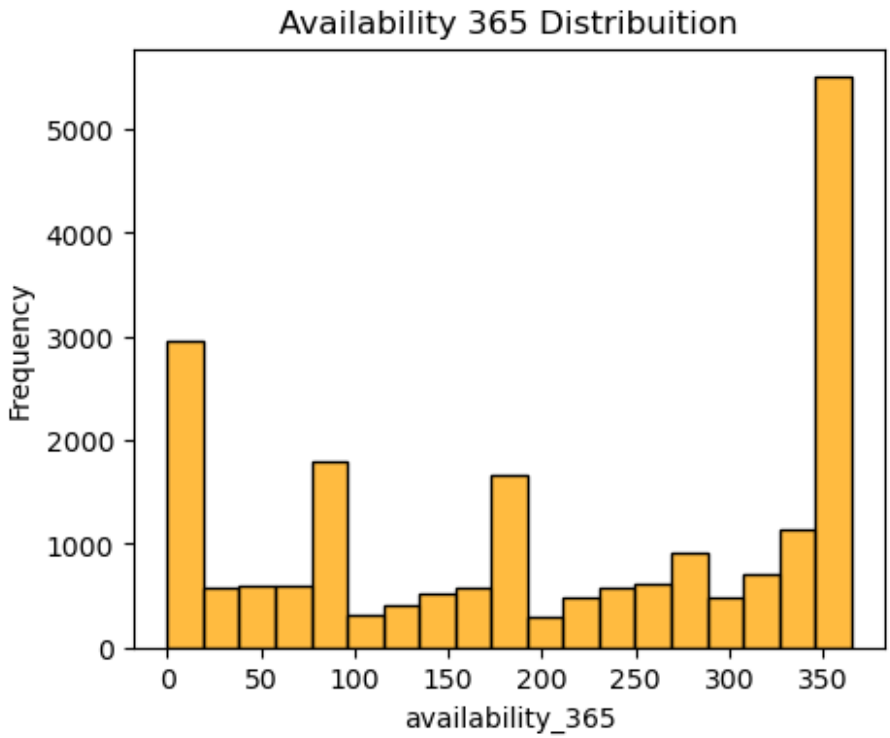

```
<Figure size 500x400 with 0 Axes>
```

```
plt.figure(figsize=(7, 4))
df = df[df['price']<1600]
sns.histplot(data=df, x='price', color='blue', bins=120)
plt.ylabel("Frequency")
plt.title("Price Distribution")
plt.show()
```




```
plt.figure(figsize=(5,4))
sns.histplot(data=df, x='availability_365', color='Orange')
plt.title('Availability 365 Distribuition')
plt.ylabel('Frequency')
```

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



Feature Engineering

```
df.groupby(by='neighbourhood')['price'].mean()

neighbourhood
Allerton            100.828571
Arden Heights       133.750000
Arrochar            124.818182
Arverne             203.059701
Astoria             112.404432
                          ...
Windsor Terrace     160.824561
Woodhaven            91.928571
Woodlawn            133.500000
Woodrow             143.500000
Woodside             79.117647
Name: price, Length: 221, dtype: float64


#Price per Beds
```

```
df['price per bed']= df['price']/df['beds']
```

```
df['price per bed']
```

```
0         55.0
1        144.0
2         93.5
3        120.0
4         85.0
         ...
20765     45.0
20766     52.5
20767    299.0
20768    115.0
20769    102.0
Name: price per bed, Length: 20658, dtype: float64
```

```
df.groupby(by='neighbourhood_group')['price per bed'].mean()
```

```
neighbourhood_group
Bronx             74.713639
Brooklyn         100.006647
Manhattan        139.249856
Queens            76.336210
Staten Island     67.728101
Name: price per bed, dtype: float64
```

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

```
neighbourhood_group
Bronx            107.990506
Brooklyn         155.488542
Manhattan        205.395742
Queens           121.681939
Staten Island    118.780069
Name: price, dtype: float64
```

Bivariante Analysis

```
# Price vs Avilability

plt.figure(figsize=(5,4))
sns.scatterplot(data=df, x='availability_365', y='price',
color='lightgreen')
```

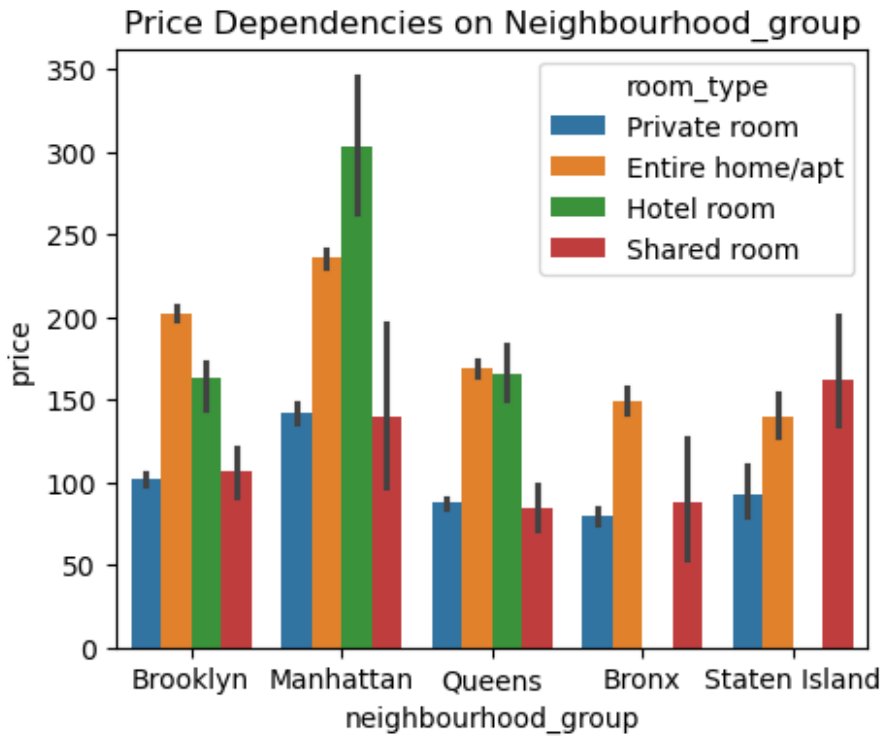
```
plt.title('Price vs Avilability')
plt.ylabel('price')
```

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


```
#df.columns
```

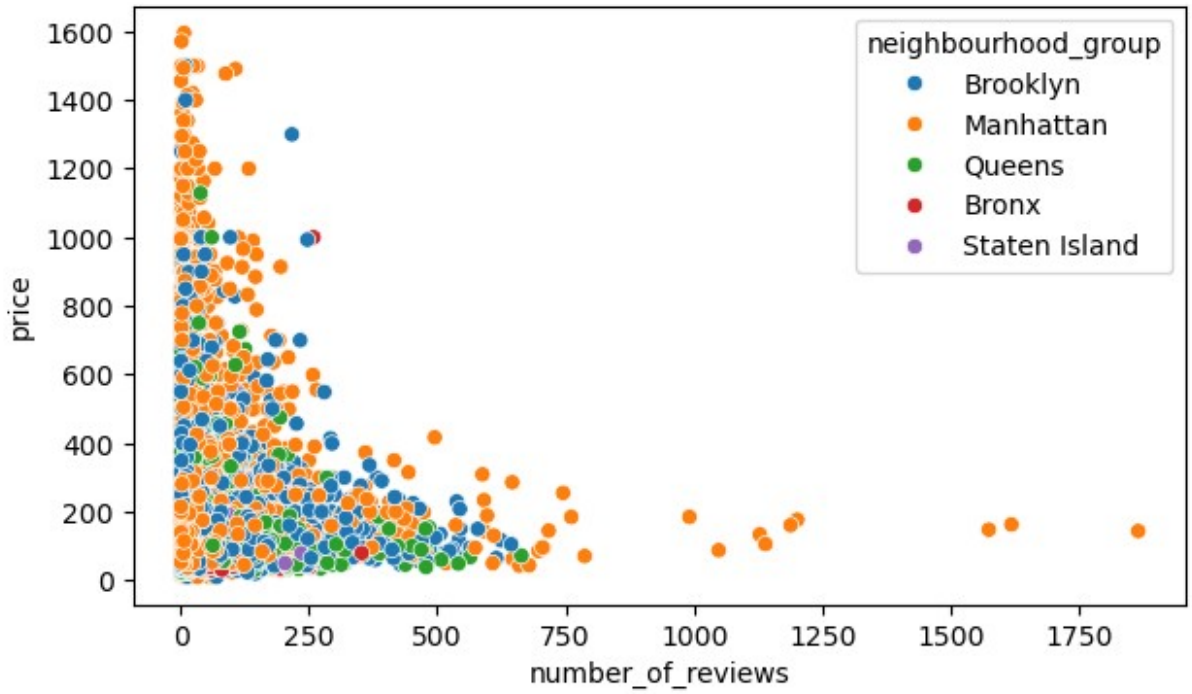
```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'number_of_reviews_ltm',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

```
plt.figure(figsize=(5,4))
sns.barplot(data=df, x='neighbourhood_group', y='price',
hue='room_type')
plt.title('Price Dependencies on Neighbourhood_group')
plt.show()
```




```
# Price Relationship with Neighourhood_group

plt.figure(figsize=(7, 4))
sns.scatterplot(data=df, x='number_of_reviews', y='price',
hue='neighbourhood_group')
```

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

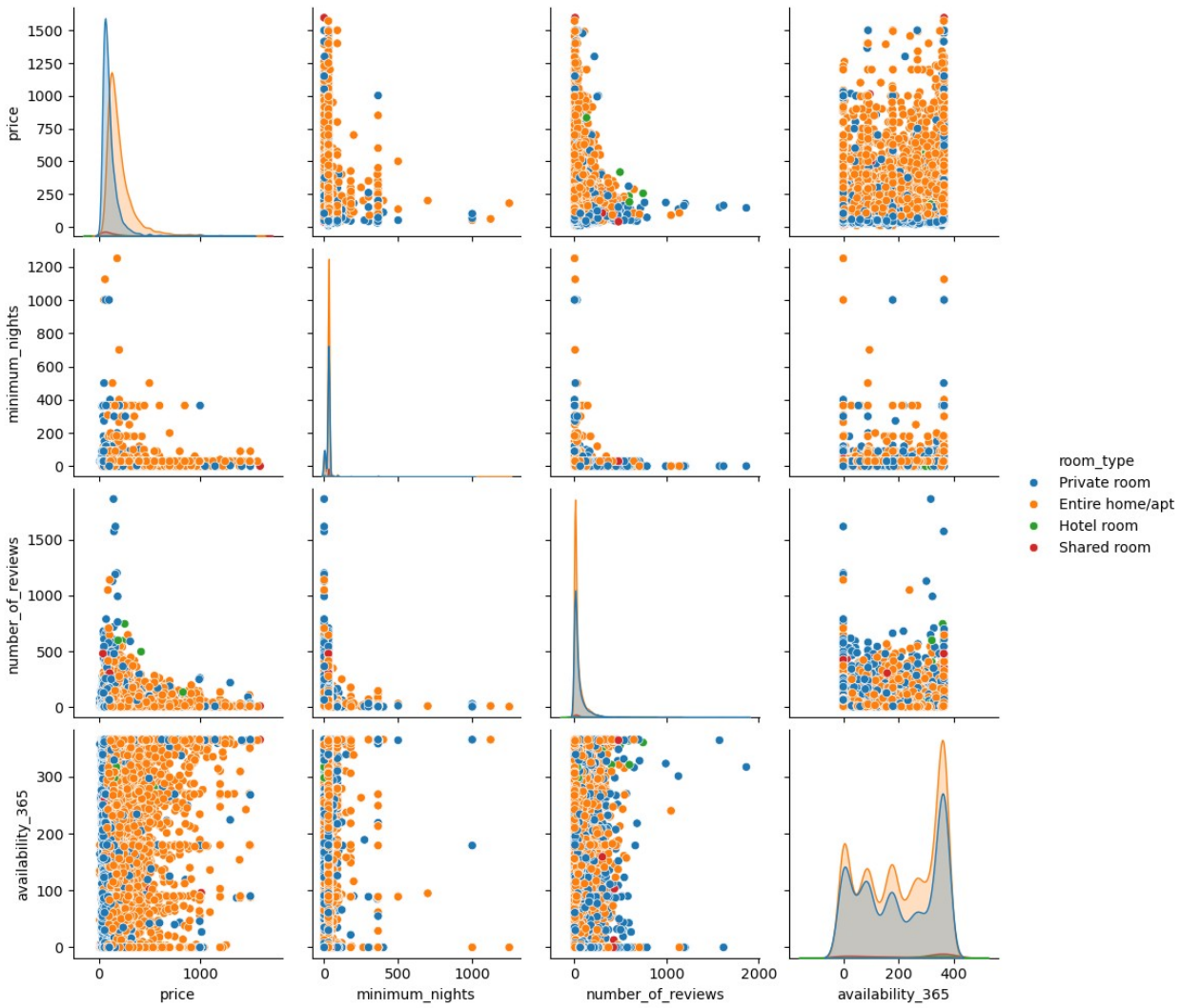



```
df.columns
```

```
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', 'price per bed'],
      dtype='object')
```

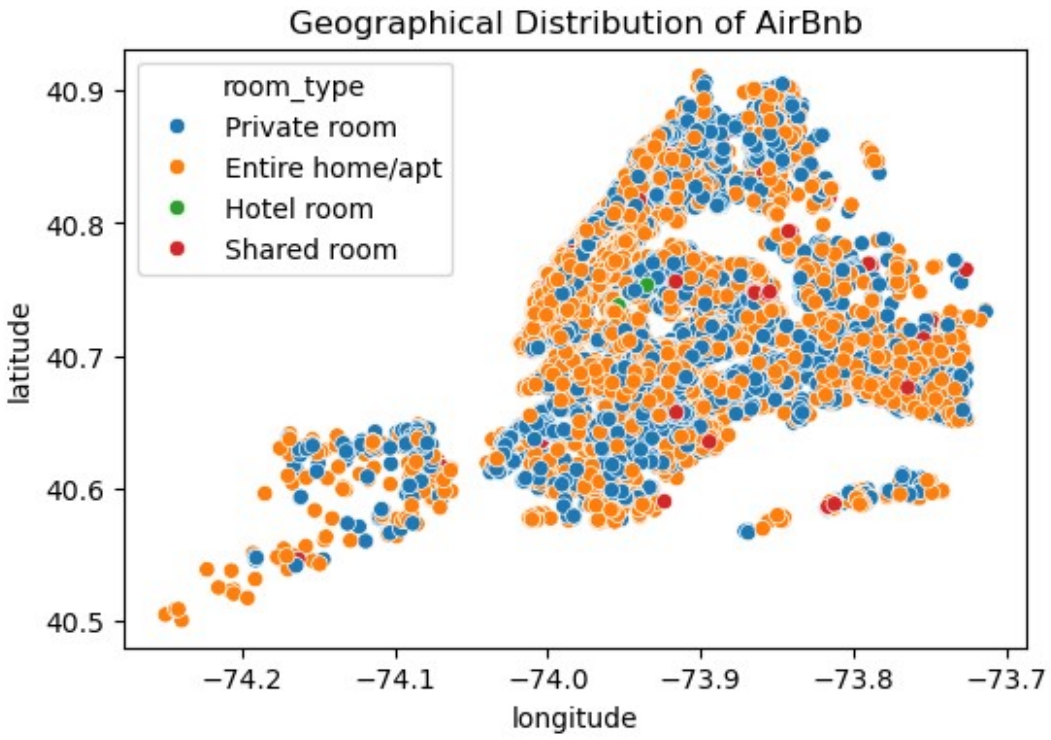
```
sns.pairplot(data=df, vars=['price', 'minimum_nights',
'number_of_reviews', 'availability_365'], hue='room_type')
```

```
<seaborn.axisgrid.PairGrid at 0x2989ee2a570>
```



Geographical Analysis

```
plt.figure(figsize=(6, 4))
sns.scatterplot(data=df, x='longitude', y='latitude', hue='room_type')
plt.title('Geographical Distribution of AirBnb')
plt.show()
```



Correlation

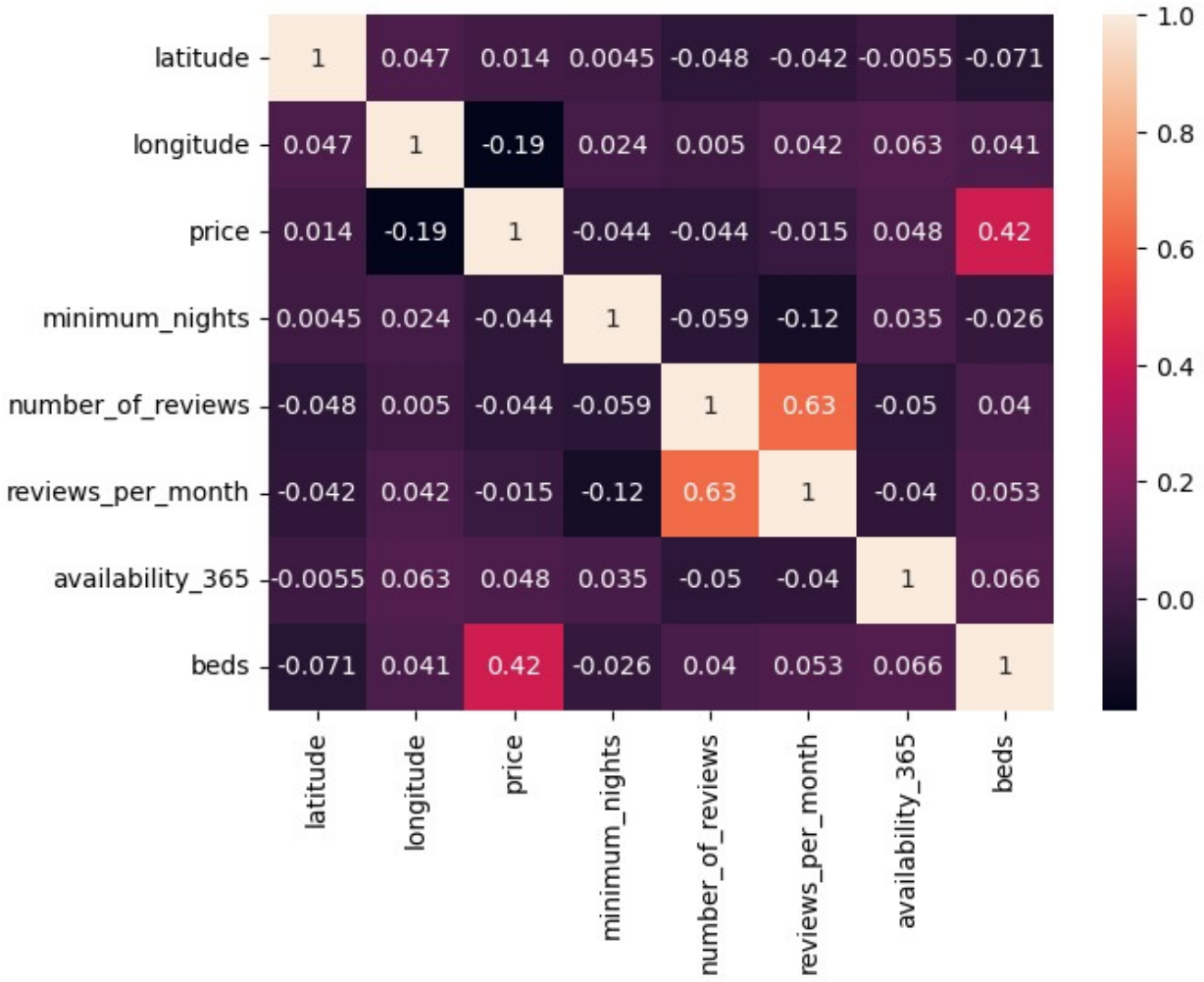
```
# Correlation of one variable with other numerical columns

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   0.047150  0.014285        0.004546
longitude           0.047150   1.000000 -0.191878        0.023922
price               0.014285  -0.191878  1.000000       -0.043549
minimum_nights      0.004546   0.023922 -0.043549        1.000000
number_of_reviews  -0.047900   0.005049 -0.044268       -0.058992
reviews_per_month  -0.041821   0.041874 -0.014609       -0.122526
availability_365   -0.005472   0.063105  0.047975        0.035241
beds               -0.070552   0.041423  0.415024       -0.026132

                    number_of_reviews  reviews_per_month
availability_365  \
```

```
latitude                    -0.047900           -0.041821          -
0.005472
longitude                    0.005049            0.041874
0.063105
price                       -0.044268           -0.014609
0.047975
minimum_nights              -0.058992           -0.122526
0.035241
number_of_reviews            1.000000            0.630981          -
0.049738
reviews_per_month            0.630981            1.000000          -
0.040231
availability_365            -0.049738           -0.040231
1.000000
beds                         0.039519            0.052874
0.066299

                          beds
latitude            -0.070552
longitude            0.041423
price                0.415024
minimum_nights      -0.026132
number_of_reviews    0.039519
reviews_per_month    0.052874
availability_365     0.066299
beds                 1.000000
```

```
plt.figure(figsize=(7, 5))
sns.heatmap(data=corr, annot=True)
```

```
<Axes: >
```


latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
availability_365
beds
1 0.047 0.014 0.0045 -0.048 -0.042 -0.0055 -0.071
0.047 1 -0.19 0.024 0.005 0.042 0.063 0.041
0.014 -0.19 1 -0.044 -0.044 -0.015 0.048 0.42
0.0045 0.024 -0.044 1 -0.059 -0.12 0.035 -0.026
-0.048 0.005 -0.044 -0.059 1 0.63 -0.05 0.04
-0.042 0.042 -0.015 -0.12 0.63 1 -0.04 0.053
-0.0055 0.063 0.048 0.035 -0.05 -0.04 1 0.066
-0.071 0.041 0.42 -0.026 0.04 0.053 0.066 1
1.0
0.8
0.6
0.4
0.2
0.0