EDA_Project_AirBnB

October 22, 2024

0.1 About Airbnb

Airbnb is an online marketplace that connects people who want to rent out their homes with those looking for accommodations. Since its inception in 2008, Airbnb has transformed the travel and hospitality industry by offering unique lodging options around the world. Hosts can list their properties—be it a spare room, an entire apartment, or a unique space—on the platform, while guests can find accommodations that suit their preferences and budget.

The platform operates in over 220 countries and regions, with millions of listings available, ranging from budget-friendly shared rooms to luxury villas. With a focus on creating authentic travel experiences, Airbnb has grown to become a popular alternative to traditional hotels, offering travelers more flexibility and local immersion during their stays.

In recent years, Airbnb has also introduced services such as "Experiences," allowing hosts to offer activities and tours, further enhancing guests' travel experiences. This evolution has helped Airbnb become a holistic travel solution, connecting people to places and experiences all over the globe.

1 Python Project EDA & Data Visualization - AirBnB Listings 2024(New York)

1.1 Tasks

1. Importing all dependenices (lib)
2. Loading datasets
3. Initial exploration
4. Data cleaning
5. Data Analysis

1.1.1 Task 1 importing all dependenices (lib)

[59]:
```
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import seaborn as sns
```

```
import matplotlib.pyplot as plt
%matplotlib inline
```

1.1.2 Task 2 loading datasets

[60]:
```
df = pd.read_csv("new_york_listings_2024_.csv", encoding_errors='ignore')
```

1.1.3 Task 3 Initial Exploration

[61]:
```
df.head()
```

[61]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	…	last_review	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	license	rating	bedrooms	beds	baths
0	1.312228e+06	Rental unit in Brooklyn · 5.0 · 1 bedroom	7130382	Walter	Brooklyn	Clinton Hill	40.683710	-73.964610	Private room	55.0	…	20/12/15	0.03	1.0	0.0	0.0	No License	5	1	1	Not specified
1	4.527754e+07	Rental unit in New York · 4.67 · 2 bedrooms ·…	51501835	Jeniffer	Manhattan	Hell's Kitchen	40.766610	-73.988100	Entire home/apt	144.0	…	01/05/23	0.24	139.0	364.0	2.0	No License	4.67	2	1	1
2	9.710000e+17	Rental unit in New York · 4.17 · 1 bedroom · …	528871354	Joshua	Manhattan	Chelsea	40.750764	-73.994605	Entire home/apt	187.0	…	18/12/23	1.67	1.0	343.0	6.0	Exempt	4.17	1	2	1
3	3.857863e+06	Rental unit in New York · 4.64 · 1 bedroom · …	19902271	John And Catherine	Manhattan	Washington Heights	40.835600	-73.942500	Private room	120.0	…	17/09/23	1.38	2.0	363.0	12.0	No License	4.64	1	1	1
4	4.089661e+07	Condo in New York · 4.91 · Studio · 1 bed · 1…	61391963	Stay With Vibe	Manhattan	Murray Hill	40.751120	-73.978600	Entire home/apt	85.0	…	03/12/23	0.24	133.0	335.0	3.0	No License	4.91	Studio	1	1

[5 rows x 22 columns]

[7]:
```
df.tail()
```

[7]:

```
                 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  …     01/07/23               0.48
20767 -73.993430  Entire home/apt  299.0  …     08/12/23               2.09
20768 -73.991470  Entire home/apt  115.0  …     17/09/23               0.91
20769 -73.728651     Private room  102.0  …     10/12/23               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]
```

[31]:
```
df.shape
```

[31]:
```
(20770, 22)
```

[32]:
```
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
```

[33]:
```
df.isna().sum()
```

[33]:
```
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
```

[34]:
```
df.describe().T
```

[34]:

	count	mean	std	min	25%	50%	75%	max
id	20770.0	3.033858e+17	3.901221e+17	2595.000000	2.707260e+07	4.992852e+07	7.220000e+17	1.050000e+18
host_id	20770.0	1.749049e+08	1.725657e+08	1678.000000	2.041184e+07	1.086990e+08	3.143997e+08	5.504035e+08
latitude	20763.0	4.072682e+01	6.029301e-02	40.500314	4.068416e+01	4.072289e+01	4.076311e+01	4.091115e+01
longitude	20763.0	-7.393918e+01	6.140254e-02	-74.249840	-7.398076e+01	-7.394960e+01	-7.391747e+01	-7.371365e+01
price	20736.0	1.877149e+02	1.023245e+03	10.000000	8.000000e+01	1.250000e+02	1.990000e+02	1.000000e+05
minimum_nights	20763.0	2.855849e+01	3.353270e+01	1.000000	3.000000e+01	3.000000e+01	3.000000e+01	1.250000e+03
number_of_reviews	20763.0	4.261061e+01	7.352340e+01	1.000000	4.000000e+00	1.400000e+01	4.900000e+01	1.865000e+03
reviews_per_month	20763.0	1.257589e+00	1.904472e+00	0.010000	2.100000e-01	6.500000e-01	1.800000e+00	7.549000e+01
calculated_host_listings_count	20763.0	1.886669e+01	7.092144e+01	1.000000	1.000000e+00	2.000000e+00	5.000000e+00	7.130000e+02
availability_365	20763.0	2.060680e+02	1.350773e+02	0.000000	8.700000e+01	2.150000e+02	3.530000e+02	3.650000e+02
number_of_reviews_ltm	20763.0	1.084896e+01	2.135488e+01	0.000000	1.000000e+00	3.000000e+00		
beds	20770.0	1.723592e+00	1.211993e+00	1.000000	1.000000e+00	1.000000e+00		

```
number_of_reviews_ltm           1.500000e+01   1.075000e+03
beds                            2.000000e+00   4.200000e+01
```

1.1.4 Task 4 Data cleaning

[35]:
```
df.dropna(inplace=True)

df.isnull().sum()
```

[35]:
```
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
```

[36]:
```
# dealing with duplicates rows
df.duplicated().sum()
```

[36]:
```
12
```

[37]:
```
df.drop_duplicates(inplace=True)
df.duplicated().sum()
```

[37]:
```
0
```

[38]:
```
print(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
```

[39]:
```
df['id'] = df['id'].astype(object)
df.dtypes

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

[39]:
```
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
```

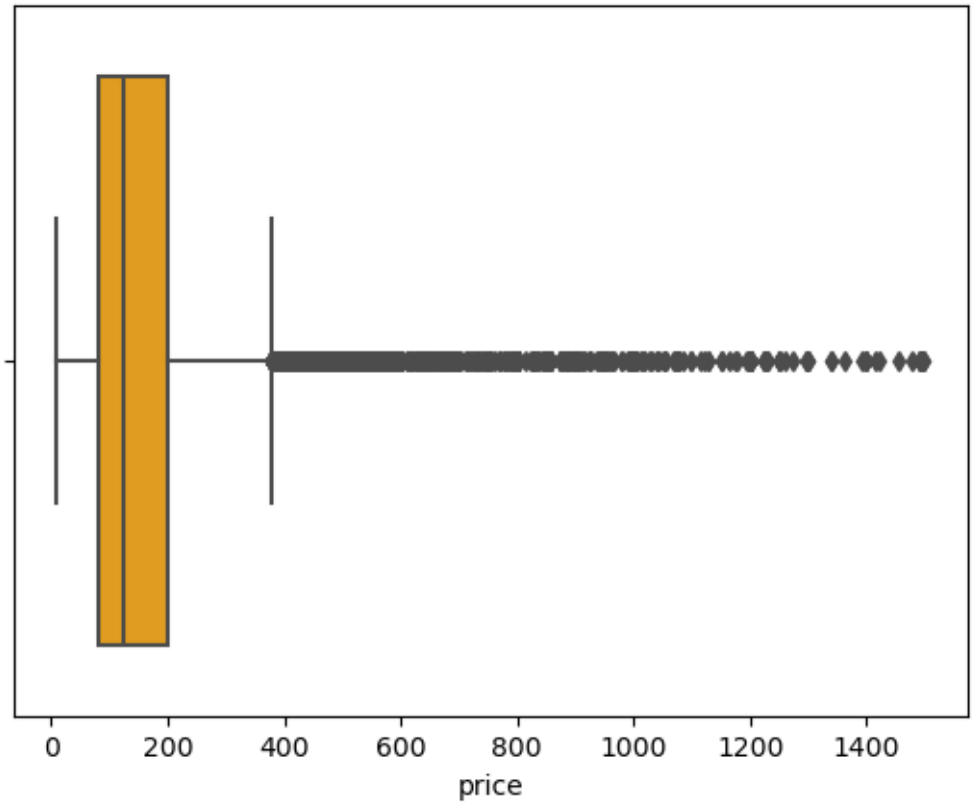
1.1.5 EDA

Task 5: Data Analysis

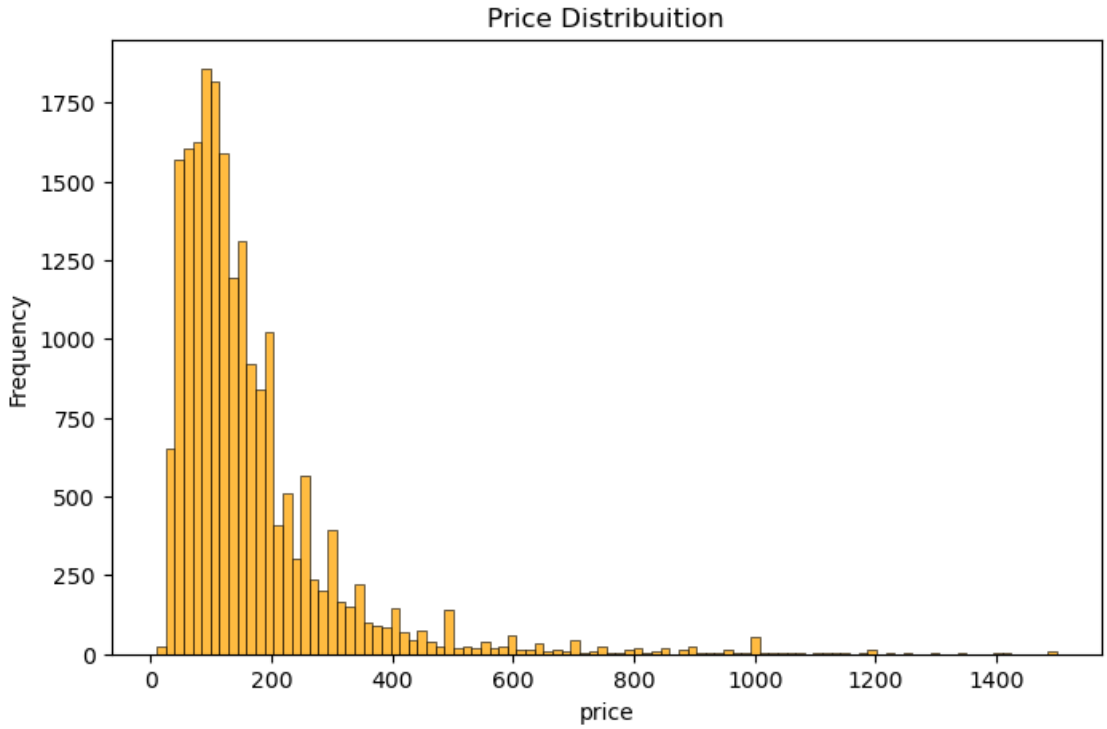
[64]:
```
df = df[df['price'] < 1500]

sns.boxplot(data=df, x='price', color='orange')
```

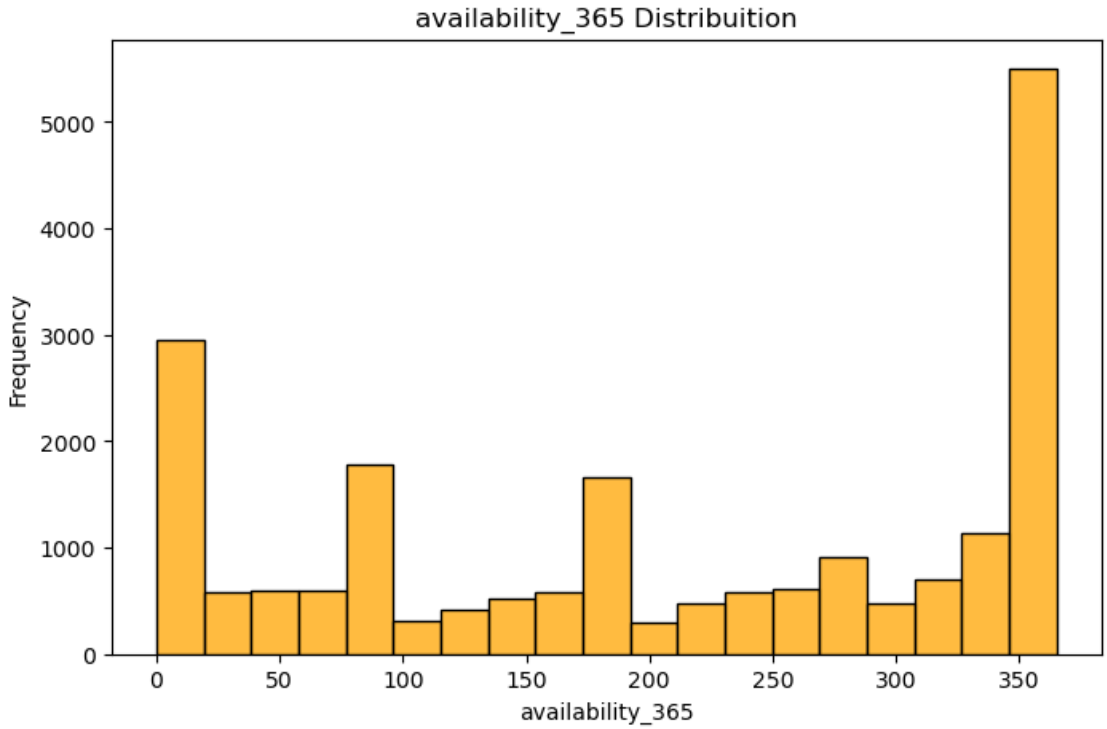
[64]:
```
<Axes: xlabel='price'>
```



[65]:
```
plt.figure(figsize=(8,5))
sns.histplot(df,x='price', bins=100,color='orange')
plt.title('Price Distribuition')
plt.ylabel("Frequency")
plt.show()
```



[66]:
```
plt.figure(figsize=(8,5))
sns.histplot(df,x='availability_365',color='orange')
plt.title('availability_365 Distribuition')
plt.ylabel("Frequency")
plt.show()
```



[67]:
```
df.groupby(['neighbourhood_group','bedrooms']).agg({'price':'mean'})
```

[67]:

neighbourhood_group	bedrooms	price
Bronx	1	87.126420
	2	148.960317
	3	210.971014
	4	266.100000
	5	331.666667
	6	400.000000
	Studio	104.457143
Brooklyn	1	117.361655
	2	206.054822
	3	268.585570
	4	359.721805
	5	349.393443
	6	547.111111
	7	589.200000
	8	916.333333
	9	445.250000
	Studio	143.900285
Manhattan	1	169.209628
	2	277.604722

```
                        3         418.041995
                        4         516.796610
                        5         491.933333
                        6         638.000000
                        7         400.333333
                        9         959.000000
                        Studio    157.763262
Queens                  1          92.020739
                        2         172.379189
                        3         232.530547
                        4         296.467742
                        5         399.153846
                        6         513.400000
                        7         299.000000
                        Studio    113.727273
Staten Island           1          87.873626
                        2         136.526316
                        3         210.347826
                        4         393.125000
                        5         437.000000
                        6         279.000000
                        Studio     80.882353
```

Feature Engineering

[68]:
```
df['price per bed'] = df['price']/df['beds']
df.head()
```

[68]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	…	reviews_per_month
0	1.312228e+06	Rental unit in Brooklyn · 5.0 · 1 bedroom	7130382	Walter	Brooklyn	Clinton Hill	40.683710	-73.964610	Private room	55.0	…	0.03
1	4.527754e+07	Rental unit in New York · 4.67 · 2 bedrooms ·…	51501835	Jeniffer	Manhattan	Hell's Kitchen	40.766610	-73.988100	Entire home/apt	144.0	…	0.24
2	9.710000e+17	Rental unit in New York · 4.17 · 1 bedroom · …	528871354	Joshua	Manhattan	Chelsea	40.750764	-73.994605	Entire home/apt	187.0	…	1.67
3	3.857863e+06	Rental unit in New York · 4.64 · 1 bedroom · …	19902271	John And Catherine	Manhattan	Washington Heights	40.835600	-73.942500	Private room	120.0	…	1.38
4	4.089661e+07	Condo in New York · 4.91 · Studio · 1 bed · 1…	61391963	Stay With Vibe	Manhattan	Murray Hill	40.751120	-73.978600	Entire home/apt	85.0	…	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 price per bed
0  No License       5         1    1  Not specified          55.0
1  No License    4.67         2    1              1         144.0
2      Exempt    4.17         1    2              1          93.5
3  No License    4.64         1    1              1         120.0
4  No License    4.91    Studio    1              1          85.0

[5 rows x 23 columns]
```

[69]:
```
df.groupby(by='neighbourhood_group')['price per bed'].mean()
```

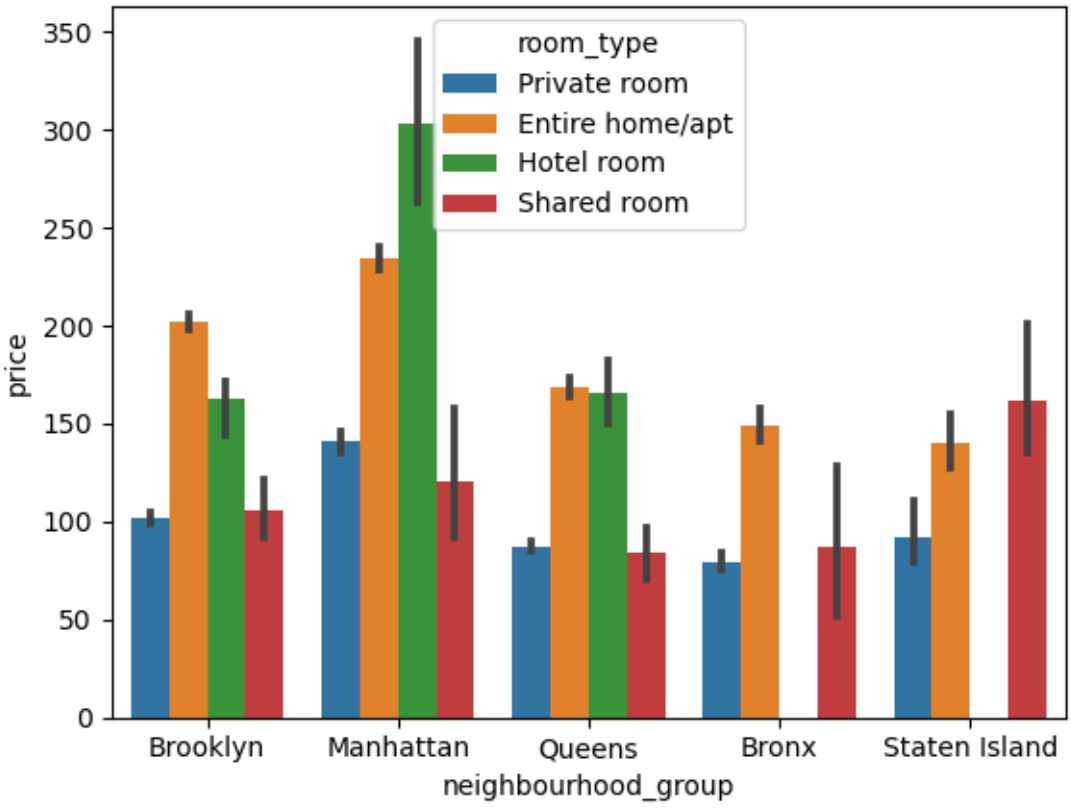
[69]:
```
neighbourhood_group
Bronx             74.713639
Brooklyn          99.788493
Manhattan        138.662489
Queens            76.336210
Staten Island     67.728101
Name: price per bed, dtype: float64
```

[70]:
```
df.columns
```

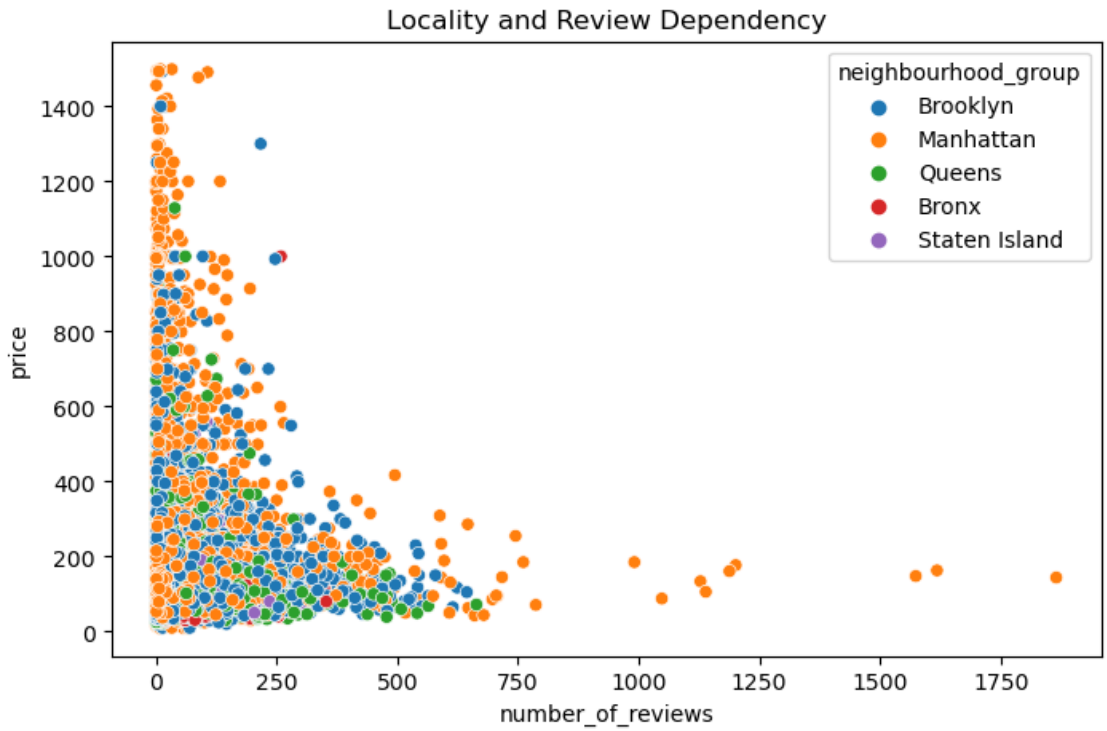
[70]:
```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'number_of_reviews_ltm', 'license', 'rating',
       'bedrooms', 'beds', 'baths', 'price per bed'],
      dtype='object')
```

[71]:
```
sns.barplot(df, x='neighbourhood_group', y = 'price', hue='room_type')
```

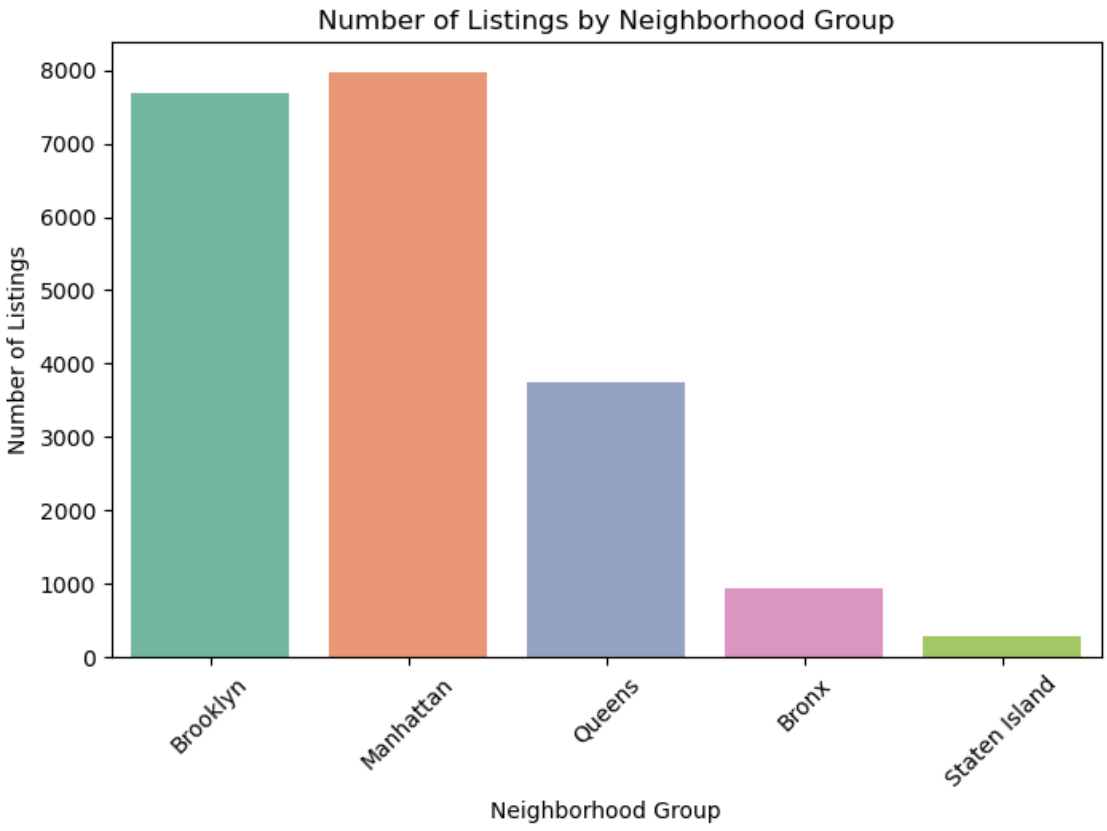
[71]:
```
<Axes: xlabel='neighbourhood_group', ylabel='price'>
```



[72]:
```
plt.figure(figsize=(8,5))
plt.title("Locality and Review Dependency")
sns.scatterplot(df, x='number_of_reviews', y='price', hue='neighbourhood_group')
plt.show()
```

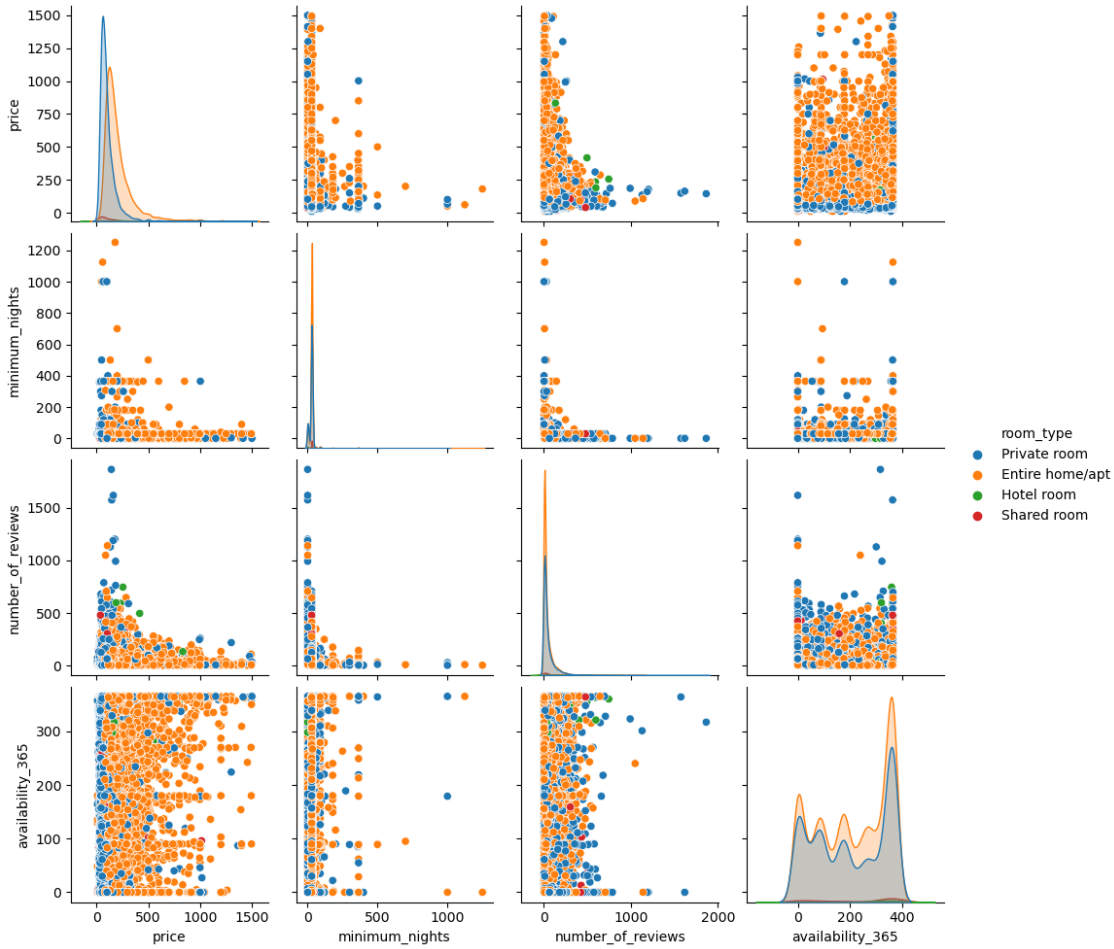


[78]:
```
plt.figure(figsize=(8, 5))
sns.countplot(data=df, x='neighbourhood_group', palette='Set2')
plt.title('Number of Listings by Neighborhood Group')
plt.xlabel('Neighborhood Group')
plt.ylabel('Number of Listings')
plt.xticks(rotation=45)
plt.show()
```



[79]:
```
sns.pairplot(data=df, vars=['price', 'minimum_nights', 'number_of_reviews',␣
↪'availability_365'], hue='room_type')
```

[79]: <seaborn.axisgrid.PairGrid at 0x21298d549d0>


price
minimum_nights
number_of_reviews
availability_365
room_type
Private room
Entire home/apt
Hotel room
Shared room

[ ]:

[ ]:

[ ]: