```
# Importing labraries & Loading  Dataset
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv(r"D:\Internship Project\compressed_data.csv")
df
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_4304\3074320747.py:7:
DtypeWarning: Columns (25) have mixed types. Specify dtype option on
import or set low_memory=False.
  df = pd.read_csv(r"D:\Internship Project\compressed_data.csv")
```

	id	NAME	\
0	1001254	Clean & quiet apt home by the park	
1	1002102	Skylit Midtown Castle	
2	1002403	THE VILLAGE OF HARLEM....NEW YORK !	
3	1002755	NaN	
4	1003689	Entire Apt: Spacious Studio/Loft by central park	
...	...	...	
102594	6092437	Spare room in Williamsburg	
102595	6092990	Best Location near Columbia U	
102596	6093542	Comfy, bright room in Brooklyn	
102597	6094094	Big Studio-One Stop from Midtown	
102598	6094647	585 sf Luxury Studio	

	host id	host_identity_verified	host name	neighbourhood group	\
0	80014485718	unconfirmed	Madaline	Brooklyn	
1	52335172823	verified	Jenna	Manhattan	
2	78829239556	NaN	Elise	Manhattan	
3	85098326012	unconfirmed	Garry	Brooklyn	
4	92037596077	verified	Lyndon	Manhattan	
...	...	...	...	...	
102594	12312296767	verified	Krik	Brooklyn	
102595	77864383453	unconfirmed	Mifan	Manhattan	
102596	69050334417	unconfirmed	Megan	Brooklyn	
102597	11160591270	unconfirmed	Christopher	Queens	
102598	68170633372	unconfirmed	Rebecca		

```
Manhattan

              neighbourhood       lat      long        country  ...  \
0                 Kensington  40.64749 -73.97237  United States  ...
1                    Midtown  40.75362 -73.98377  United States  ...
2                     Harlem  40.80902 -73.94190  United States  ...
3               Clinton Hill  40.68514 -73.95976  United States  ...
4                East Harlem  40.79851 -73.94399  United States  ...
...                      ...       ...       ...            ...  ...
102594          Williamsburg  40.70862 -73.94651  United States  ...
102595  Morningside Heights  40.80460 -73.96545  United States  ...
102596            Park Slope  40.67505 -73.98045  United States  ...
102597     Long Island City  40.74989 -73.93777  United States  ...
102598      Upper West Side  40.76807 -73.98342  United States  ...

       service fee minimum nights number of reviews last review  \
0             $193            10.0               9.0  10/19/2021
1              $28            30.0              45.0   5/21/2022
2             $124             3.0               0.0         NaN
3              $74            30.0             270.0    7/5/2019
4              $41            10.0               9.0  11/19/2018
...             ...             ...               ...         ...
102594        $169             1.0               0.0         NaN
102595        $167             1.0               1.0    7/6/2015
102596        $198             3.0               0.0         NaN
102597        $109             2.0               5.0  10/11/2015
102598        $206             1.0               0.0         NaN

        reviews per month review rate number calculated host listings
count  \
0                    0.21                4.0
6.0
1                    0.38                4.0
2.0
2                     NaN                5.0
1.0
3                    4.64                4.0
1.0
4                    0.10                3.0
1.0
...                   ...                ...
...
102594                NaN                3.0
1.0
102595               0.02                2.0
2.0
102596                NaN                5.0
1.0
102597               0.10                3.0
1.0
```

```
102598                NaN                3.0
1.0

        availability 365
house_rules  \
0                 286.0  Clean up and treat the home the way you'd
like...
1                 228.0  Pet friendly but please confirm with me if
the...
2                 352.0  I encourage you to use my kitchen, cooking
and...
3                 322.0
NaN
4                 289.0  Please no smoking in the house, porch or on
th...
...                 ...
...
102594            227.0  No Smoking No Parties or Events of any kind
Pl...
102595            395.0  House rules: Guests agree to the following
ter...
102596            342.0
NaN
102597            386.0
NaN
102598             69.0
NaN

       license
0          NaN
1          NaN
2          NaN
3          NaN
4          NaN
...        ...
102594     NaN
102595     NaN
102596     NaN
102597     NaN
102598     NaN

[102599 rows x 26 columns]

df.columns

Index(['id', 'NAME', 'host id', 'host_identity_verified', 'host name',
       'neighbourhood group', 'neighbourhood', 'lat', 'long',
'country',
       'country code', 'instant_bookable', 'cancellation_policy',
'room type',
```

```
       'Construction year', 'price', 'service fee', 'minimum nights',
       'number of reviews', 'last review', 'reviews per month',
       'review rate number', 'calculated host listings count',
       'availability 365', 'house_rules', 'license'],
      dtype='object')

df.shape

(102599, 26)

df.head()

        id                                                 NAME      host
id  \
0  1001254                Clean & quiet apt home by the park
80014485718
1  1002102                             Skylit Midtown Castle
52335172823
2  1002403               THE VILLAGE OF HARLEM....NEW YORK !
78829239556
3  1002755                                                NaN
85098326012
4  1003689  Entire Apt: Spacious Studio/Loft by central park
92037596077

  host_identity_verified host name neighbourhood group
neighbourhood  \
0            unconfirmed  Madaline            Brooklyn    Kensington

1               verified     Jenna           Manhattan       Midtown

2                    NaN     Elise           Manhattan        Harlem

3            unconfirmed     Garry            Brooklyn  Clinton Hill

4               verified    Lyndon           Manhattan   East Harlem


        lat      long        country  ... service fee minimum
nights  \
0  40.64749 -73.97237  United States  ...       $193            10.0

1  40.75362 -73.98377  United States  ...        $28            30.0

2  40.80902 -73.94190  United States  ...       $124             3.0

3  40.68514 -73.95976  United States  ...        $74            30.0

4  40.79851 -73.94399  United States  ...        $41            10.0
```

```
  number of reviews last review  reviews per month review rate number
\
0                9.0  10/19/2021                0.21                 4.0

1               45.0   5/21/2022                0.38                 4.0

2                0.0         NaN                 NaN                 5.0

3              270.0    7/5/2019                4.64                 4.0

4                9.0  11/19/2018                0.10                 3.0


  calculated host listings count  availability 365  \
0                              6.0             286.0
1                              2.0             228.0
2                              1.0             352.0
3                              1.0             322.0
4                              1.0             289.0

                                          house_rules license
0  Clean up and treat the home the way you'd like...     NaN
1  Pet friendly but please confirm with me if the...     NaN
2  I encourage you to use my kitchen, cooking and...     NaN
3                                                NaN     NaN
4  Please no smoking in the house, porch or on th...     NaN

[5 rows x 26 columns]
```

Checking Missing Values

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102599 entries, 0 to 102598
Data columns (total 26 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              102599 non-null  int64
 1   NAME                            102349 non-null  object
 2   host id                         102599 non-null  int64
 3   host_identity_verified          102310 non-null  object
 4   host name                       102193 non-null  object
 5   neighbourhood group             102570 non-null  object
 6   neighbourhood                   102583 non-null  object
 7   lat                             102591 non-null  float64
 8   long                            102591 non-null  float64
 9   country                         102067 non-null  object
```

```
 10  country code                    102468 non-null  object
 11  instant_bookable                102494 non-null  object
 12  cancellation_policy             102523 non-null  object
 13  room type                       102599 non-null  object
 14  Construction year               102385 non-null  float64
 15  price                           102352 non-null  object
 16  service fee                     102326 non-null  object
 17  minimum nights                  102190 non-null  float64
 18  number of reviews               102416 non-null  float64
 19  last review                     86706 non-null   object
 20  reviews per month               86720 non-null   float64
 21  review rate number              102273 non-null  float64
 22  calculated host listings count  102280 non-null  float64
 23  availability 365                102151 non-null  float64
 24  house_rules                     50468 non-null   object
 25  license                         2 non-null       object
dtypes: float64(9), int64(2), object(15)
memory usage: 20.4+ MB
```

Handle Missing Values

```
'''This code ensures that the 'last review' column is properly
formatted as datetime, missing values in key columns
are appropriately handled, and incomplete records are removed,
preparing the dataset for further analysis or visualization.'''

"This code ensures that the 'last review' column is properly\
nformatted as datetime, missing values in key columns\nare
appropriately handled, and incomplete records are removed,\npreparing
the dataset for further analysis or visualization."

df['last review'] = pd.to_datetime(df['last review'], errors =
'coerce')

df.fillna({'reviews per month':0, 'last review' : df['last
review'].min()}, inplace =True)


df.dropna(subset = ['NAME','host name'], inplace = True)

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

```
id                                0
NAME                              0
host id                           0
host_identity_verified          276
host name                         0
neighbourhood group              26
neighbourhood                    16
```

```
lat                                   8
long                                  8
country                             526
country code                        122
instant_bookable                     96
cancellation_policy                  70
room type                             0
Construction year                   200
price                               239
service fee                         268
minimum nights                      403
number of reviews                   182
last review                           0
reviews per month                     0
review rate number                  314
calculated host listings count      318
availability 365                    420
house_rules                       51867
license                          101947
dtype: int64
```

```
df = df.drop(columns=["license", "house_rules"], errors='ignore')
```

Correct Data Types

```
#Ensure that all columns have the correct data types.

#Removing  the dollar sign and convert into float
df['price'] = df['price'].replace('[\$,]','',regex=True).astype(float)
df['service fee'] = df['service fee'].replace('[\
$,]','',regex=True).astype(float)
```

```
<>:2: SyntaxWarning: invalid escape sequence '\$'
<>:3: SyntaxWarning: invalid escape sequence '\$'
<>:2: SyntaxWarning: invalid escape sequence '\$'
<>:3: SyntaxWarning: invalid escape sequence '\$'
C:\Users\HP\AppData\Local\Temp\ipykernel_4304\3126523377.py:2:
SyntaxWarning: invalid escape sequence '\$'
  df['price'] = df['price'].replace('[\
$,]','',regex=True).astype(float)
C:\Users\HP\AppData\Local\Temp\ipykernel_4304\3126523377.py:3:
SyntaxWarning: invalid escape sequence '\$'
  df['service fee'] = df['service fee'].replace('[\
$,]','',regex=True).astype(float)
```

Remove Duplicates

```
# Now remove duplicates
df.drop_duplicates(inplace=True)
```

Confirm Data Cleaning

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 101410 entries, 0 to 102057
Data columns (total 24 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              101410 non-null  int64
 1   NAME                            101410 non-null  object
 2   host id                         101410 non-null  int64
 3   host_identity_verified          101134 non-null  object
 4   host name                       101410 non-null  object
 5   neighbourhood group             101384 non-null  object
 6   neighbourhood                   101394 non-null  object
 7   lat                             101402 non-null  float64
 8   long                            101402 non-null  float64
 9   country                         100884 non-null  object
 10  country code                    101288 non-null  object
 11  instant_bookable                101314 non-null  object
 12  cancellation_policy             101340 non-null  object
 13  room type                       101410 non-null  object
 14  Construction year               101210 non-null  float64
 15  price                           101171 non-null  float64
 16  service fee                     101142 non-null  float64
 17  minimum nights                  101016 non-null  float64
 18  number of reviews               101228 non-null  float64
 19  last review                     101410 non-null  datetime64[ns]
 20  reviews per month               101410 non-null  float64
 21  review rate number              101103 non-null  float64
 22  calculated host listings count  101092 non-null  float64
 23  availability 365                100990 non-null  float64
dtypes: datetime64[ns](1), float64(11), int64(2), object(10)
memory usage: 19.3+ MB
None
```

```
df = df.drop(columns=["license", "house_rules"], errors='ignore')

df
```

```
               id                                                NAME  \
0         1001254                Clean & quiet apt home by the park
```

```
1        1002102                                Skylit Midtown Castle
2        1002403                THE VILLAGE OF HARLEM....NEW YORK !
4        1003689  Entire Apt: Spacious Studio/Loft by central park
5        1004098         Large Cozy 1 BR Apartment In Midtown East
...          ...                                                ...
102053  57365208               Cozy bright room near Prospect Park
102054  57365760         Private Bedroom with Amazing Rooftop View
102055  57366313     Pretty Brooklyn One-Bedroom for 2 to 4 people
102056  57366865        Room & private bathroom in historic Harlem
102057  57367417                                  Rosalee Stewart

            host id host_identity_verified host name neighbourhood
group  \
0       80014485718            unconfirmed  Madaline
Brooklyn
1       52335172823               verified     Jenna
Manhattan
2       78829239556                    NaN     Elise
Manhattan
4       92037596077               verified    Lyndon
Manhattan
5       45498551794               verified  Michelle
Manhattan
...             ...                    ...       ...                  .
..
102053  77326652202            unconfirmed    Mariam
Brooklyn
102054  45936254757               verified      Trey
Brooklyn
102055  23801060917               verified   Michael
Brooklyn
102056  15593031571            unconfirmed   Shireen
Manhattan
102057  93578954226               verified   Stanley
Manhattan

             neighbourhood       lat      long        country  ...  \
0               Kensington  40.64749 -73.97237  United States  ...
1                  Midtown  40.75362 -73.98377  United States  ...
2                   Harlem  40.80902 -73.94190  United States  ...
4             East Harlem  40.79851 -73.94399  United States  ...
5             Murray Hill  40.74767 -73.97500  United States  ...
...                   ...       ...       ...            ...  ...
102053           Flatbush  40.64945 -73.96108  United States  ...
102054           Bushwick  40.69872 -73.92718  United States  ...
102055  Bedford-Stuyvesant  40.67810 -73.90822  United States  ...
102056              Harlem  40.81248 -73.94317  United States  ...
102057              Harlem  40.81315 -73.94747  United States  ...

       Construction year   price service fee minimum nights  \
```

```
0                  2020.0   966.0       193.0            10.0
1                  2007.0   142.0        28.0            30.0
2                  2005.0   620.0       124.0             3.0
4                  2009.0   204.0        41.0            10.0
5                  2013.0   577.0       115.0             3.0
...                   ...     ...         ...             ...
102053                NaN   696.0         NaN             7.0
102054                NaN   909.0         NaN             1.0
102055                NaN   387.0         NaN             2.0
102056                NaN   848.0         NaN             2.0
102057             2011.0  1128.0         NaN             4.0

        number of reviews  last review  reviews per month  review rate
number  \
0                     9.0   2021-10-19               0.21
4.0
1                    45.0   2022-05-21               0.38
4.0
2                     0.0   2012-07-11               0.00
5.0
4                     9.0   2018-11-19               0.10
3.0
5                    74.0   2019-06-22               0.59
3.0
...                   ...          ...                ...
...
102053               12.0   2019-03-27               0.44
5.0
102054               19.0   2017-08-31               0.72
3.0
102055               50.0   2019-06-26               3.12
4.0
102056                0.0   2012-07-11               0.00
1.0
102057               22.0   2019-06-15               0.85
4.0

        calculated host listings count  availability 365
0                                  6.0             286.0
1                                  2.0             228.0
2                                  1.0             352.0
4                                  1.0             289.0
5                                  1.0             374.0
...                                ...               ...
102053                             1.0               0.0
102054                             2.0               0.0
102055                             2.0             235.0
102056                             1.0               0.0
102057                             1.0             238.0
```

```
[101410 rows x 24 columns]
```

Descriptive Analysis

```
df.describe()
```

```
                  id       host id             lat           long  \
count  1.014100e+05  1.014100e+05  101402.000000  101402.000000
mean   2.920959e+07  4.926155e+10      40.728082     -73.949663
min    1.001254e+06  1.236005e+08      40.499790     -74.249840
25%    1.507574e+07  2.459183e+10      40.688730     -73.982570
50%    2.922911e+07  4.912069e+10      40.722300     -73.954440
75%    4.328308e+07  7.399747e+10      40.762750     -73.932340
max    5.736742e+07  9.876313e+10      40.916970     -73.705220
std    1.626820e+07  2.853703e+10       0.055850       0.049474

       Construction year          price    service fee  minimum nights
\
count      101210.000000  101171.000000  101142.000000   101016.000000

mean         2012.486908     625.381008     125.043998        8.113744

min          2003.000000      50.000000      10.000000    -1223.000000

25%          2007.000000     340.000000      68.000000        2.000000

50%          2012.000000     625.000000     125.000000        3.000000

75%          2017.000000     913.000000     183.000000        5.000000

max          2022.000000    1200.000000     240.000000     5645.000000

std             5.765130     331.609111      66.313374       30.378014


       number of reviews                    last review  reviews per
month  \
count      101228.000000                         101410
101410.000000
mean           27.511854  2018-05-15 21:26:08.721033728
1.163207
min             0.000000            2012-07-11 00:00:00
0.000000
25%             1.000000            2017-07-30 00:00:00
0.090000
50%             7.000000            2019-05-23 00:00:00
0.480000
75%            31.000000            2019-07-01 00:00:00
```

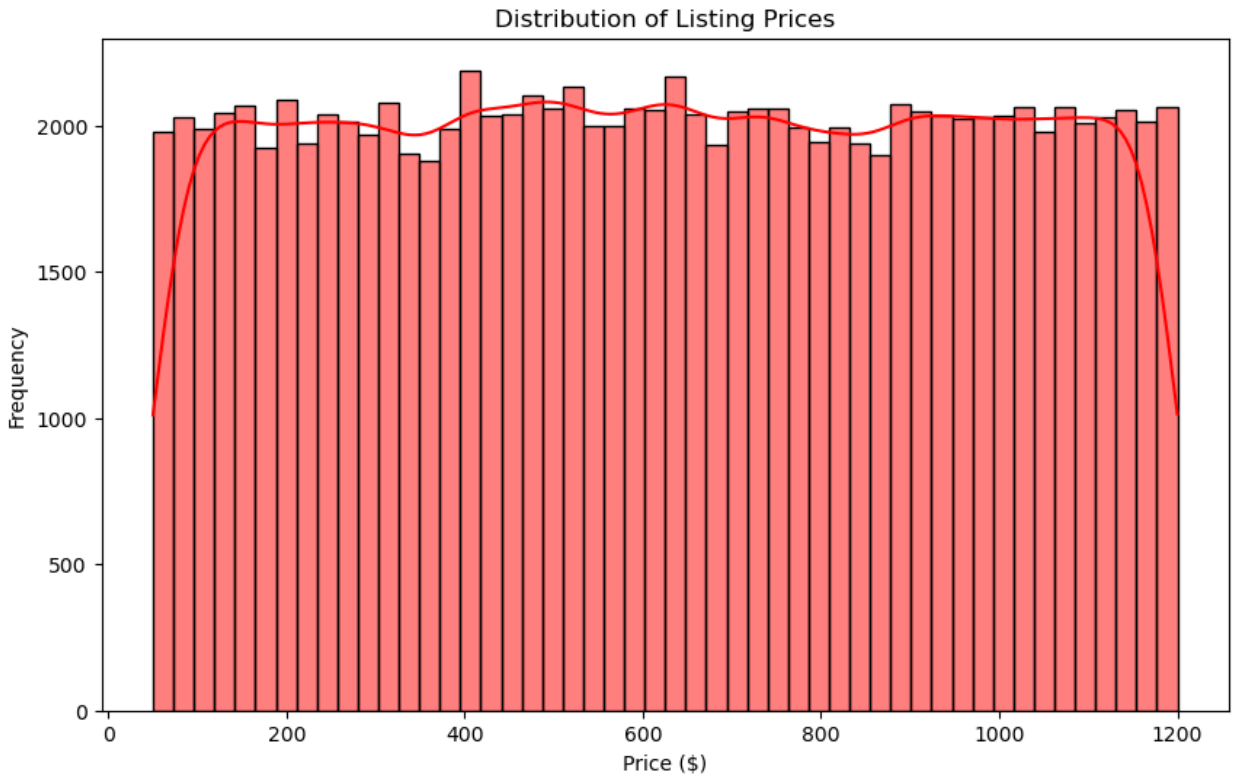
```
1.710000
max          1024.000000             2058-06-16 00:00:00
90.000000
std            49.549258                             NaN
1.683708

       review rate number  calculated host listings count
availability 365
count       101103.000000                   101092.000000
100990.000000
mean             3.278558                        7.948463
141.164660
min              1.000000                        1.000000         -
10.000000
25%              2.000000                        1.000000
3.000000
50%              3.000000                        1.000000
96.000000
75%              4.000000                        2.000000
269.000000
max              5.000000                      332.000000
3677.000000
std              1.285369                       32.328974
135.419199
```

Visualization

Distribution of Prices

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,6))
sns.histplot(df['price'], bins=50, kde=True, color='red')
plt.title('Distribution of Listing Prices')
plt.xlabel('Price ($)')
plt.ylabel('Frequency')
plt.show()
```

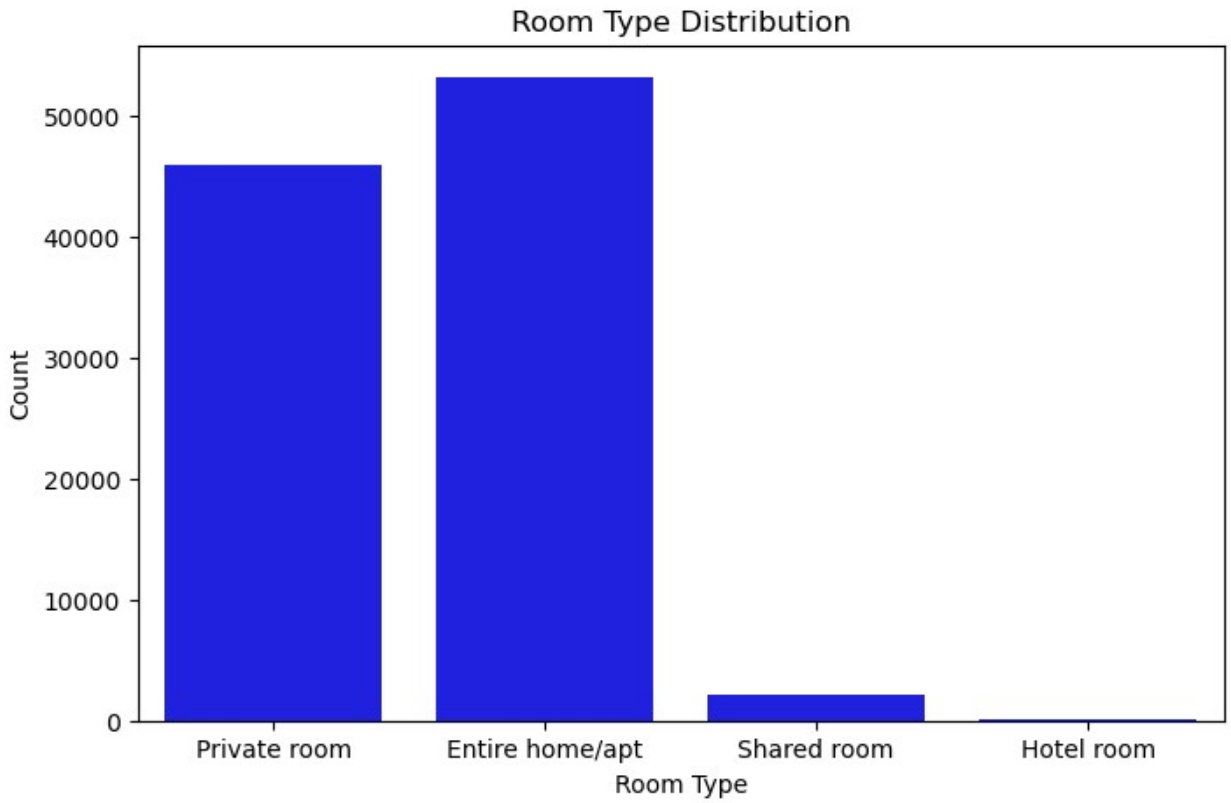


This histogram shows that the vast majority of listings are priced at the lower end, with frequency dropping sharply as prices increase. The steep peak on the left and long right tail indicate a heavily right-skewed distribution, meaning high-priced listings are rare but significantly inflate the price range. This suggests that while most items are affordable, a small number of expensive listings could distort average price metrics.

Room Type Analysis

Analyze the distribution of different room types.

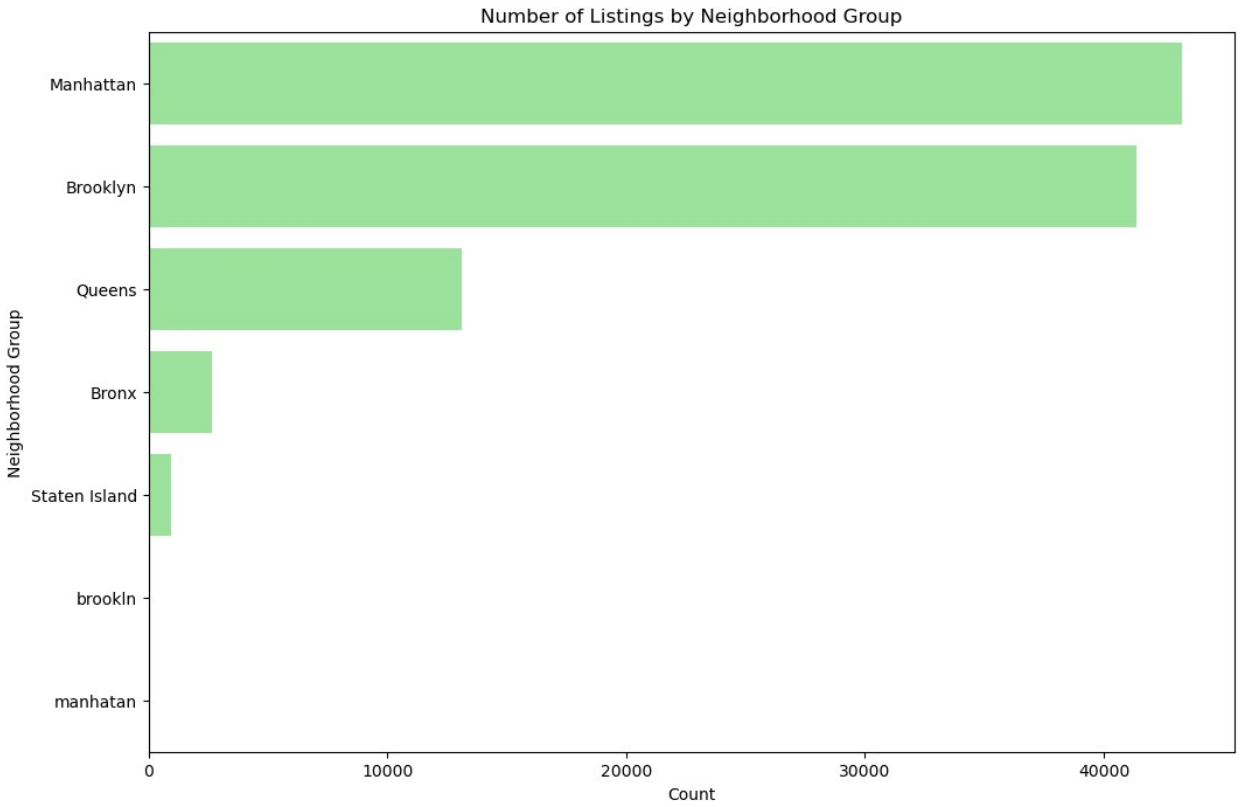
```
plt.figure(figsize=(8, 5))
sns.countplot(x='room type', data=df , color='blue')
plt.title('Room Type Distribution')
plt.xlabel('Room Type')
plt.ylabel('Count')
plt.show()
```



The count plot shows a clear distribution of the different room types available in the Airbnb dataset. The majority of listings are for 'Entire home/apt' and 'Private room', with 'Shared room' and 'Hotel room' being much less common. This insight can be useful for understanding the availability and popularity of different types of accommodations on Airbnb.

Neighborhood Analysis

```
plt.figure(figsize=(12, 8))
sns.countplot(
    y='neighbourhood group',
    data=df,
    color="lightgreen",
    order=df['neighbourhood group'].value_counts().index
)
plt.title('Number of Listings by Neighborhood Group')
plt.xlabel('Count')
plt.ylabel('Neighborhood Group')
plt.show()
```



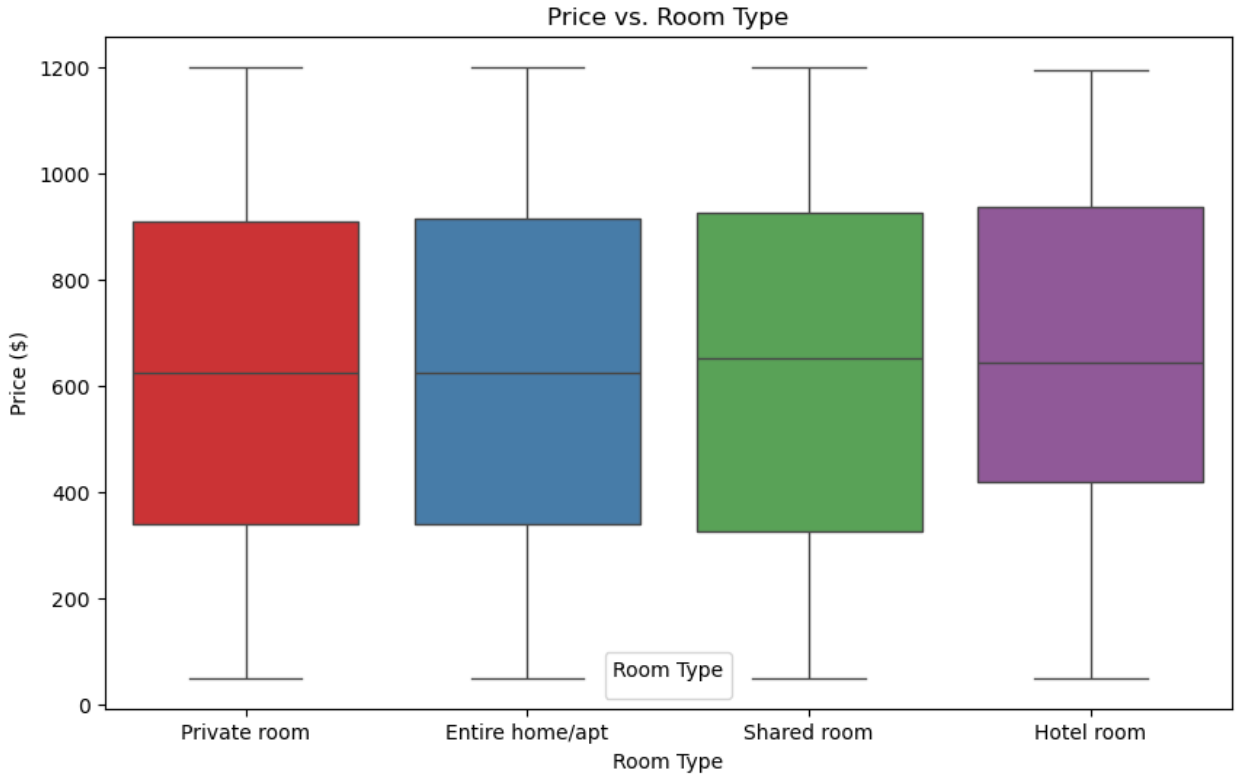
The count plot shows a clear distribution of the number of listings across different neighborhood groups. Manhattan and Brooklyn dominate the listings, suggesting they are prime locations for Airbnb. Queens, Bronx, and Staten Island have fewer listings, indicating less availability or popularity.

Price vs Room Type

Visualize the relationship between price and room type

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='room type', y='price', hue='room type', data=df,
palette='Set1')
plt.title('Price vs. Room Type')
plt.xlabel('Room Type')
plt.ylabel('Price ($)')
plt.legend(title='Room Type')
plt.show()
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_4304\2780626444.py:6:
UserWarning: No artists with labels found to put in legend.  Note that
artists whose label start with an underscore are ignored when legend()
is called with no argument.
  plt.legend(title='Room Type')
```



Price vs. Room Type The box plot provides a detailed view of how prices vary across different room types in the Airbnb dataset. It shows that while 'Shared room' tends to have lower prices, 'Private room', 'Entire home/apt', and 'Hotel room' have higher and more varied price ranges. This visualization helps in understanding the pricing dynamics for different types of accommodations on Airbnb.

```
df.head()
```

```
        id                                                 NAME      host
id  \
0  1001254                Clean & quiet apt home by the park
80014485718
1  1002102                             Skylit Midtown Castle
52335172823
2  1002403               THE VILLAGE OF HARLEM....NEW YORK !
78829239556
4  1003689  Entire Apt: Spacious Studio/Loft by central park
92037596077
5  1004098         Large Cozy 1 BR Apartment In Midtown East
45498551794

  host_identity_verified host name neighbourhood group
neighbourhood  \
0            unconfirmed  Madaline            Brooklyn    Kensington

1              verified     Jenna           Manhattan       Midtown
```

```
2                    NaN     Elise           Manhattan        Harlem

4                verified    Lyndon           Manhattan   East Harlem

5                verified  Michelle           Manhattan   Murray Hill


        lat      long        country  ... Construction year  price  \
0  40.64749 -73.97237  United States  ...             2020.0  966.0
1  40.75362 -73.98377  United States  ...             2007.0  142.0
2  40.80902 -73.94190  United States  ...             2005.0  620.0
4  40.79851 -73.94399  United States  ...             2009.0  204.0
5  40.74767 -73.97500  United States  ...             2013.0  577.0

  service fee minimum nights  number of reviews  last review  \
0       193.0           10.0                9.0   2021-10-19
1        28.0           30.0               45.0   2022-05-21
2       124.0            3.0                0.0   2012-07-11
4        41.0           10.0                9.0   2018-11-19
5       115.0            3.0               74.0   2019-06-22

   reviews per month  review rate number  calculated host listings
count  \
0               0.21                 4.0
6.0
1               0.38                 4.0
2.0
2               0.00                 5.0
1.0
4               0.10                 3.0
1.0
5               0.59                 3.0
1.0

  availability 365
0            286.0
1            228.0
2            352.0
4            289.0
5            374.0

[5 rows x 24 columns]
```

Reviews Over Time

Plot the number of reviews over time.

```
df['last review'] = pd.to_datetime(df['last review'])
reviews_over_time = df.groupby(df['last
review'].dt.to_period('M')).size()
plt.figure(figsize=(12, 6))
reviews_over_time.plot(kind='line',color='red')
plt.title('Number of Reviews Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Reviews')
plt.show()
```