Analyzing the NYC Airbnb Market

Overview

Airbnb is an online marketplace connecting people who want to rent out their homes with people looking for accommodations in that locale. It currently covers more than 100,000 cities and 220 countries worldwide.

Data analysis on thousands of listings provided through Airbnb is a crucial factor for the company. Our main objective is to determine the key metrics influencing the listing of properties on the platform. For this, we will explore and visualize the dataset from Airbnb in NYC using basic exploratory data analysis (EDA) techniques. We have found out the distribution of every Airbnb listing based on their location, including their price range, room type, listing name, and other related factors.

Objective

Understanding the factors that influence Airbnb prices in New York City, or identifying patterns of all variables Our analysis provides useful information for travelers and hosts in the city and some of the best insights for the Airbnb business.

Let's Begin

import all the necessary libraries

In [4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

# ignored warnings
import warnings
warnings.filterwarnings("ignore")
```

Import the dataset

In [5]:
```
dataset = pd.read_csv("Airbnb NYC 2019.csv")
dataset.head()
```

Out[5]:

	id	name	host_id	host_name	neighbourhood_group	neighbour
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensi
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Mi
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	H
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinto
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East H

In [6]:
```
dataset.tail()
```

Out[6]:

	id	name	host_id	host_name	neighbourhood_group	
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	

Shape of the dataset

In [7]:
```
dataset.shape
```

Out[7]:
```
(48895, 16)
```

Unique Columns in the dataset

In [8]:
```
dataset.columns
```

Out[8]:
```
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'],
      dtype='object')
```

Information of the dataset

In [9]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

Statistical Description of the dataset

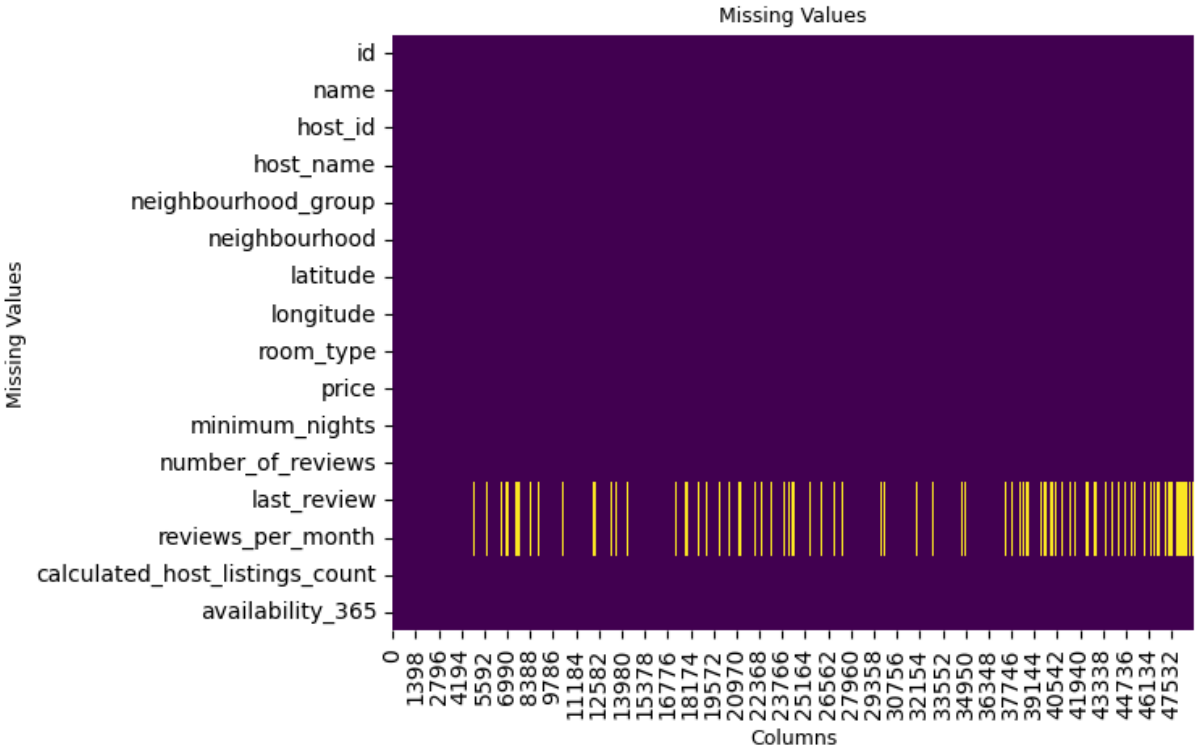
In [10]:
```
dataset.describe()
```

Out[10]:

	id	host_id	latitude	longitude	price
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000

Using Heatmaps to Visualize Missing Values

In [11]:
```
ax = plt.axes()
sns.heatmap(dataset.isna().transpose(), cbar=False, ax=ax, cmap="viridis")
plt.title("Missing Values", fontsize=9)
plt.xlabel("Columns", fontsize = 9)
plt.ylabel("Missing Values", fontsize = 9)
plt.show()
```



Replacing the null values with appropriate values

In [12]:
```
dataset['name'].replace(np.nan, 'Other Hotel', inplace =True)
dataset['host_name'].replace(np.nan, 'other', inplace = True)
```

```
dataset['last_review'].replace(np.nan, 'Not Reviewed', inplace = True)
dataset['reviews_per_month'].replace(np.nan, '0', inplace = True)
```

Checking for Duplicated Records

In [13]:
```
dataset.duplicated().sum()
```

Out[13]: 0

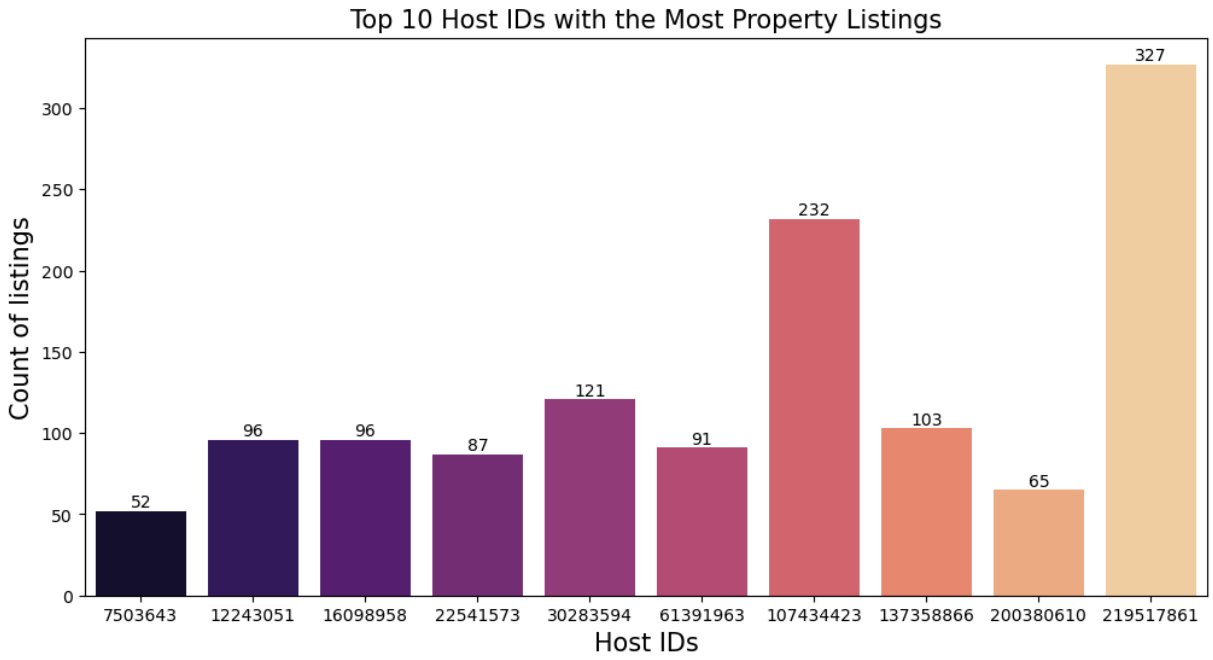
- Now we have performed data cleaning, let's uncover patterns and trends using visualizations —

Which are the top 10 host_id with the highest number of property listings?

In [14]:
```
dataset['host_id'].value_counts().reset_index().head(10)

# Plotting the bar graph to visualize the top 10 host ids with the highest r
plt.figure(figsize=(12, 6))
ax = sns.barplot(x=dataset['host_id'].value_counts().iloc[:10].index, y=data
for bars in ax.containers:
    ax.bar_label(bars)

plt.title("Top 10 Host IDs with the Most Property Listings", fontsize=15)
plt.xlabel("Host IDs", fontsize=15)
plt.ylabel("Count of listings", fontsize=15)
plt.show()
```



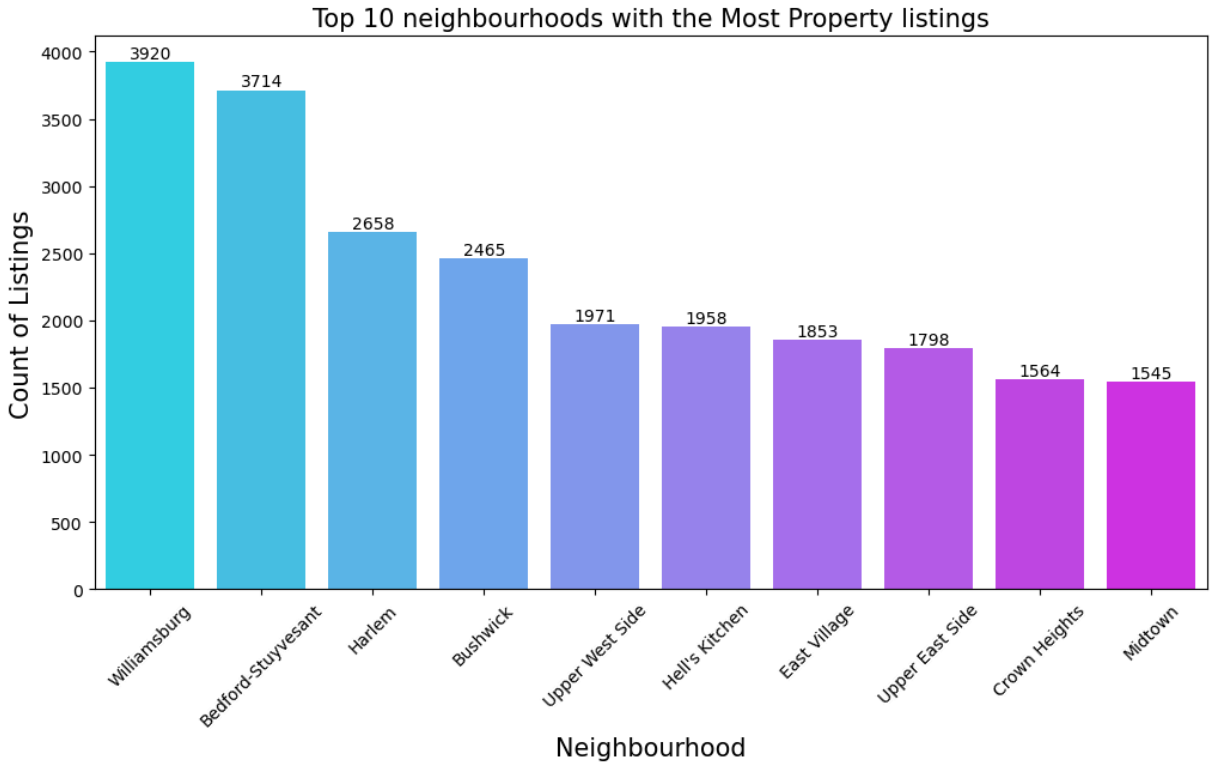
- The Host id 219517861 has the highest number of bookings with the total bookings of 327

What are the top 10 neighborhoods with the most property listings?

In [15]:
```
# Plotting the bar graph to visualize the top 10 neighborhoods with the high
plt.figure(figsize=(12, 6))
ax = sns.barplot(x=dataset['neighbourhood'].value_counts().iloc[:10].index,

for bars in ax.containers:
    ax.bar_label(bars)

plt.title("Top 10 neighbourhoods with the Most Property listings", fontsize=
plt.xlabel("Neighbourhood", fontsize=15)
plt.ylabel("Count of Listings", fontsize=15)
plt.xticks(rotation=45)
plt.show()
```



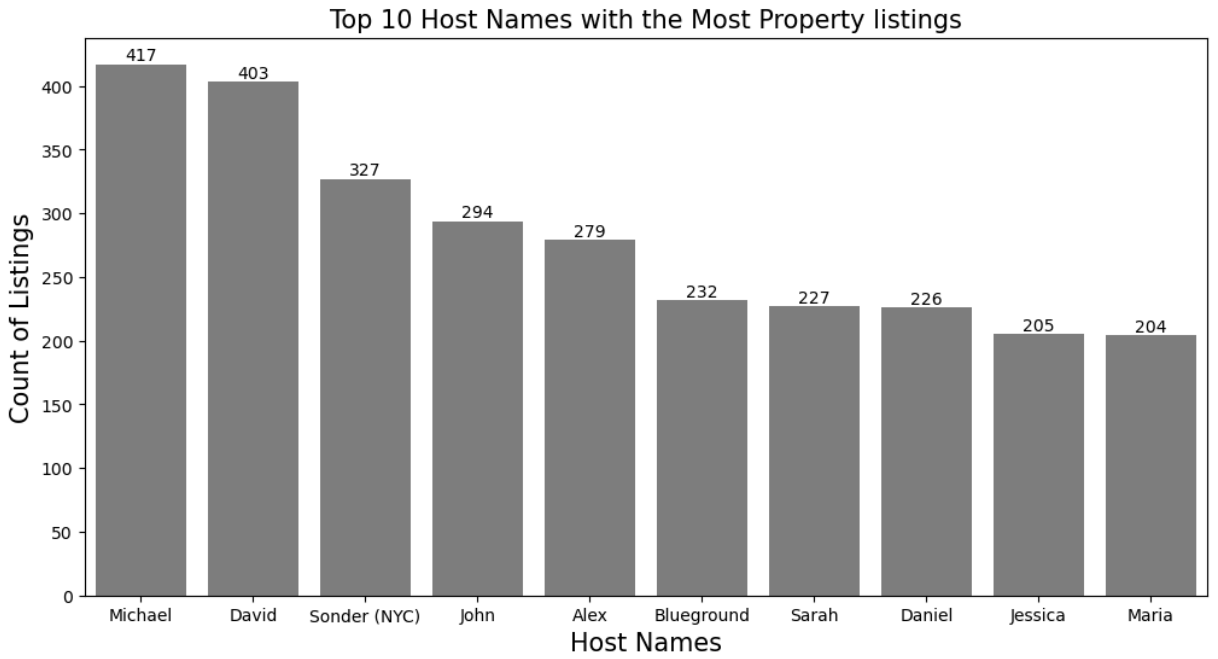
- Williamsburg is the neighborhood with the most listings in NYC with total 3920 listings

What are the top 10 host_name with the highest number of listings?

In [16]:
```
# Plotting the bar graph to visualize the top 10 host names with the highest
plt.figure(figsize=(12, 6))
ax = sns.barplot(x=dataset['host_name'].value_counts().iloc[:10].index, y=da

for bars in ax.containers:
    ax.bar_label(bars)
```

```
plt.title("Top 10 Host Names with the Most Property listings", fontsize=15)
plt.xlabel("Host Names", fontsize=15)
plt.ylabel("Count of Listings", fontsize=15)
plt.show()
```



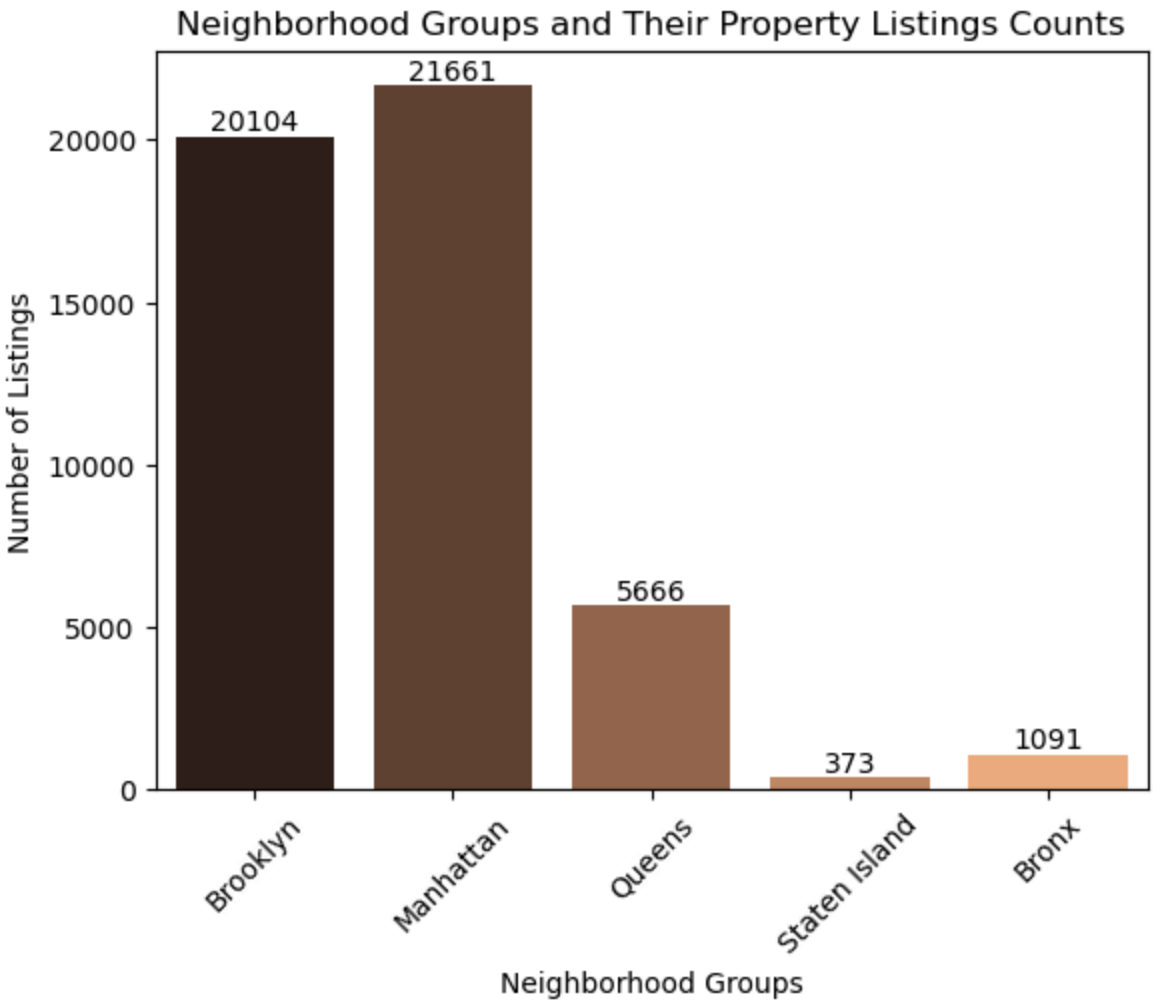
- We can observe that, Michael is the host who is the owner of highest number of property listings with the total of 417 listings

What are the property listing counts for each neighborhood group?

In [17]:
```
listings_per_neighborhood_group = sns.countplot(x='neighbourhood_group', dat

for bars in listings_per_neighborhood_group.containers:
    listings_per_neighborhood_group.bar_label(bars)

plt.title("Neighborhood Groups and Their Property Listings Counts")
plt.xlabel("Neighborhood Groups")
plt.ylabel("Number of Listings")
plt.xticks(rotation=45)
plt.show()
```



- Data shows that, Manhattan is the neighborhood group with the highest number of property listings and Staten Island contributes to the lowest number of listings.

Which neighborhoods belong to each neighborhood group?

In [18]:
```
neighborhoods_by_group = dataset.groupby('neighbourhood_group')['neighbourhc
print(neighborhoods_by_group)
```

```
neighbourhood_group
Bronx            [Highbridge, Clason Point, Eastchester, Kingsb...
Brooklyn         [Kensington, Clinton Hill, Bedford-Stuyvesant,...
Manhattan        [Midtown, Harlem, East Harlem, Murray Hill, He...
Queens           [Long Island City, Woodside, Flushing, Sunnysi...
Staten Island    [St. George, Tompkinsville, Emerson Hill, Shor...
Name: neighbourhood, dtype: object
```

What are the different Types of rooms available in the dataset?

In [19]:
```
dataset['room_type'].unique()
```

Out[19]:
```
array(['Private room', 'Entire home/apt', 'Shared room'], dtype=object)
```

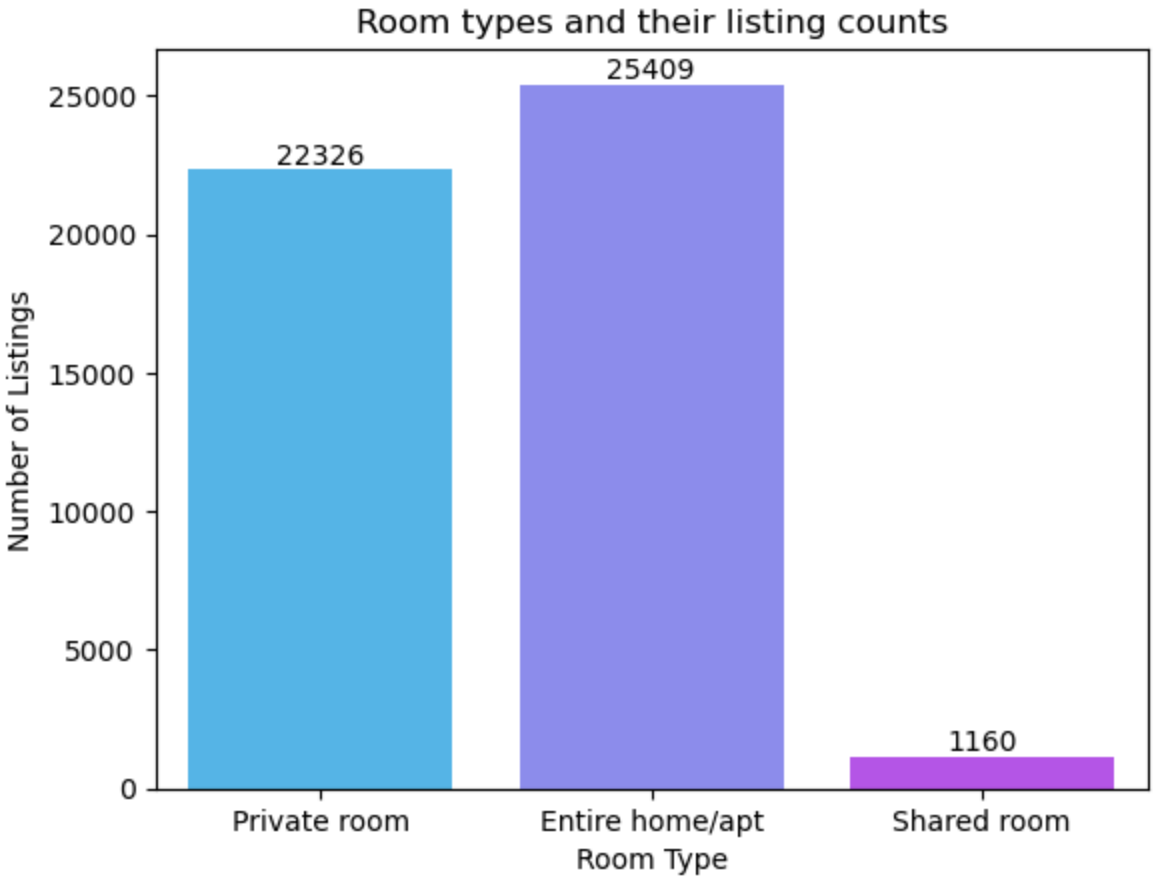
- As we can see, there are mainly 3 categories of rooms available in the Airbnb NYC listings

Which room type contributes to the highest number of listings in Airbnb NYC?

In [20]:
```
room_type_listings = sns.countplot(x='room_type', data=dataset, palette="coo

for bars in room_type_listings.containers:
    room_type_listings.bar_label(bars)


plt.title("Room types and their listing counts")
plt.xlabel("Room Type")
plt.ylabel("Number of Listings")
plt.show()
```



We can observe clearly that “Entire home/apt” room type has the highest number of listings in NYC

The most common room type is “Entire home/apt” room type

And Shared room contributes to the lowest number of listings

What is the Average price for all listings?

In [21]:
```
average_price = dataset['price'].mean()
print("Average price for all listings:", average_price)
```

```
Average price for all listings: 152.7206871868289
```

- The average price for all listings in the dataset is approximately $152.72.

How many hosts have more than one listing?

In [22]:
```
hosts_listings = dataset['host_id'].value_counts()
hosts_multiple_listing = (hosts_listings > 1).sum()
print("Number of hosts with more than one listing:", hosts_multiple_listing)
```

```
Number of hosts with more than one listing: 5154
```

Which neighborhood has the highest average price?

In [23]:
```
highest_avg_price_neighborhood = dataset.groupby('neighbourhood')['price'].m
highest_avg_price = dataset.groupby('neighbourhood')['price'].mean().max()
highest_avg_price_neighborhood, highest_avg_price
```

Out[23]:
```
('Fort Wadsworth', 800.0)
```

How many listings are available throughout the entire year?

In [24]:
```
listings_available_all_year = dataset[dataset['availability_365'] == 365]
listings_365 = listings_available_all_year.shape[0]
print("Number of listings available throughout the entire year:", listings_3
```

```
Number of listings available throughout the entire year: 1295
```

What is the average minimum nights required for a stay?

In [25]:
```
avg_minimum_nights = dataset['minimum_nights'].mean()
print("Average minimum nights required for a stay is:", avg_minimum_nights)
```

```
Average minimum nights required for a stay is: 7.029962163820431
```

How many listings have never been reviewed?

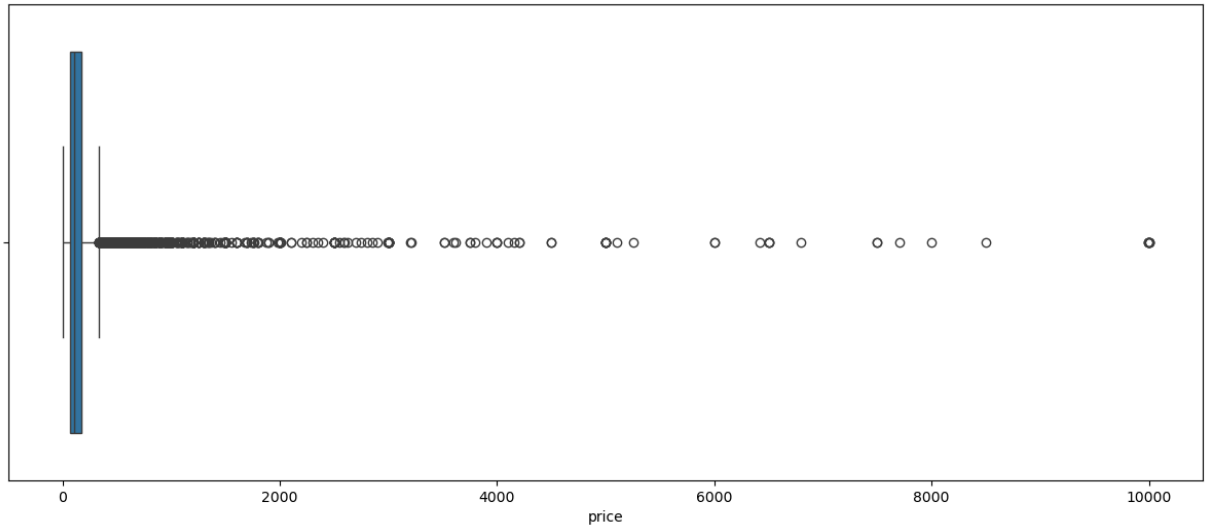
In [26]:
```
listings_never_reviewed = dataset[dataset['number_of_reviews'] == 0]
zero_reviews_count = listings_never_reviewed.shape[0]
zero_reviews_count
```

Out[26]:
```
10052
```

- As per data, 10052 listings have never been reviewed

Analyzing Price Attribute

In [27]:
```
plt.figure(figsize = (15,6))
sns.boxplot(x = dataset['price'])
plt.show()
```



In [28]:
```
# Getting the mathematical answers for the price column
dataset['price'].describe()
```

Out[28]:
```
count    48895.000000
mean       152.720687
std        240.154170
min          0.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

In [29]:
```
# Calculating Interquartile Ranges

Q1 = np.percentile(dataset['price'], 25, interpolation = 'midpoint')

# Third quartile (Q3)
Q3 = np.percentile(dataset['price'], 75, interpolation = 'midpoint')

# Interquaritle range (IQR)
IQR = Q3 - Q1
```

```
lower_fence = Q1 - (1.5 * IQR)
upper_fence = Q3 + (1.5 * IQR)

print('The IQR is:', IQR)
print('The Minimum value (Lower Fence) is:', lower_fence)
print('The Maximum value (Upper Fence) is:', upper_fence)
```

```
The IQR is: 106.0
The Minimum value (Lower Fence) is: -90.0
The Maximum value (Upper Fence) is: 334.0
```

In [30]:
```
# Filter dataset to include only values within the fences
filtered_data = dataset[(dataset['price'] >= lower_fence) & (dataset['price'
```

In [31]:
```
filtered_data.head()
```

Out[31]:

	id	name	host_id	host_name	neighbourhood_group	neighbour
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensi
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Mi
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	H
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinto
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East H

In [32]:
```
filtered_data.shape
```

Out[32]:
```
(45923, 16)
```

What is the average price for each neighborhood group, and which neighborhood group has the highest average price?

In [33]:
```
filtered_data.groupby(['neighbourhood_group'])['price'].describe().T
```

Out[33]:

neighbourhood_group	Bronx	Brooklyn	Manhattan	Queens
count	1070.000000	19415.000000	19506.000000	5567.000000
mean	77.365421	105.699614	145.952835	88.904437
std	47.110940	60.937808	70.473076	53.536041
min	0.000000	0.000000	0.000000	10.000000
25%	45.000000	60.000000	90.000000	50.000000
50%	65.000000	90.000000	135.000000	74.000000
75%	95.000000	140.000000	199.000000	108.000000
max	325.000000	333.000000	334.000000	325.000000

The data shows that Manhattan has the highest average property prices, making it the most expensive area to live in NYC. In contrast, the Bronx has the lowest average property prices, making it the most affordable area in the city.

Queens and Staten Island appear to have similar property prices.

What is the average number of listings per host?

In [34]:
```
total_listings = dataset.shape[0]
unique_hosts = dataset['host_id'].nunique()
```

In [35]:
```
average_listings_per_host = total_listings / unique_hosts
```

In [36]:
```
print("Average Listings per host: ", average_listings_per_host)
```

```
Average Listings per host:  1.3053634834610353
```

How many listings are there per neighborhood group by room type?

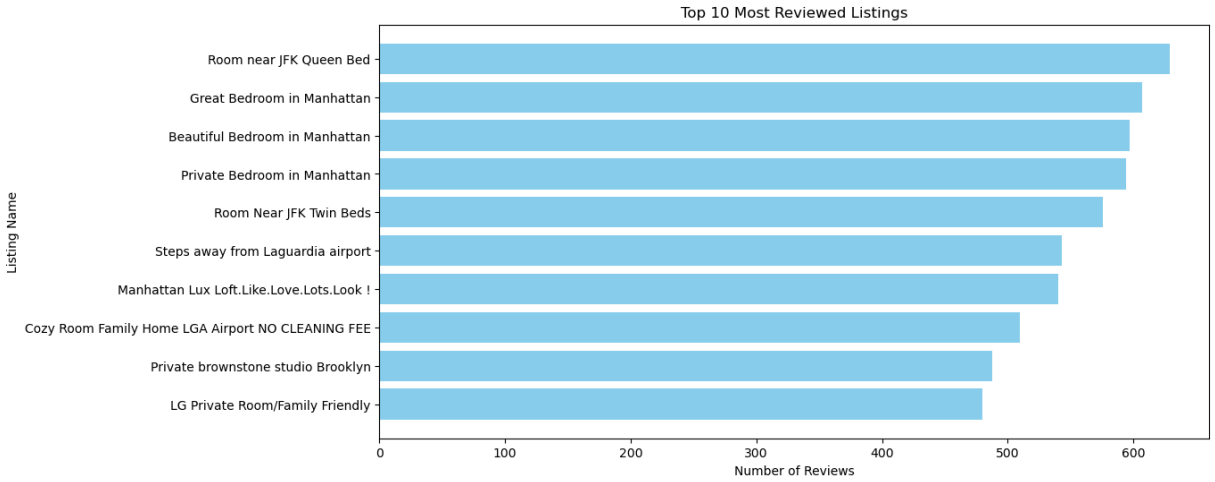
In [37]:
```
listings_per_neighborhood_group_by_room_type = dataset.groupby('neighbourhoo
listings_per_neighborhood_group_by_room_type
```

Out[37]:
```
neighbourhood_group  room_type
Bronx                Private room         652
                     Entire home/apt      379
                     Shared room           60
Brooklyn             Private room       10132
                     Entire home/apt     9559
                     Shared room          413
Manhattan            Entire home/apt    13199
                     Private room        7982
                     Shared room          480
Queens               Private room        3372
                     Entire home/apt     2096
                     Shared room          198
Staten Island        Private room         188
                     Entire home/apt      176
                     Shared room            9
Name: count, dtype: int64
```

What are the top 10 most reviewed listings?

In [38]:
```
top_10_most_reviewed_listings = dataset.sort_values(by='number_of_reviews',

plt.figure(figsize=(12, 6))
ax = plt.barh(top_10_most_reviewed_listings['name'], top_10_most_reviewed_li
plt.xlabel('Number of Reviews')
plt.ylabel('Listing Name')
plt.title('Top 10 Most Reviewed Listings')
plt.gca().invert_yaxis()
plt.show()
```



What is the average price for listings with availability less than 100 days?

In [39]:
```
listings_available_lessthan100_days = dataset[dataset['availability_365']<10
listings_available_lessthan100_days['price'].mean()
```

Out[39]:
```
138.74156660949114
```

In [ ]:
```
#dataframe to csv
filtered_data.to_csv('filtered_data.csv', index='False')
```

Summary of Insights

We analyzed the top 10 hosts to identify the one with the most property listings on Airbnb. Our findings show that Michael holds the top spot with 417 listings, making him the host with the highest number of properties. He is followed by David, who has the second-highest number of listings at 403. Additionally, our data reveals that 5,154 hosts have more than one listing on Airbnb.

The data shows that only 1,295 listings were available year-round, indicating a limited number of consistently accessible properties. Additionally, the average number of listings per host is just one, suggesting that most hosts manage only a single property.

The dataset includes five distinct neighborhood groups: Brooklyn, Manhattan, Queens, Staten Island, and the Bronx. It was found that Manhattan has the highest number of property listings on Airbnb, with a total of 21,661 listings, while Staten Island has the fewest property listings.

Our dataset includes 221 unique neighborhoods, with Williamsburg standing out as the neighborhood with the most property listings, totaling 3,920 listings. It is followed by Bedford-Stuyvesant and Harlem, with 3,714 and 2,658 listings, respectively. Fort Wadsworth has the highest average price, at $800.

The Airbnb NYC listings primarily consist of three room categories: Private room, Entire home/apt, and Shared room. Among these, Entire home/apt has the highest number of listings, making it the most common room type, while the Shared room has the fewest listings. The average price for all listings is $152.72.

According to the dataset, there are a total of 10,052 property listings on Airbnb that have not received a single review. This means that these listings have either not been booked by guests or, if they have been, no feedback has been provided by those who stayed. This lack of reviews could impact the visibility and attractiveness of these listings to potential guests, as reviews often play a significant role in helping users decide on accommodations.

On average, Airbnb listings in this dataset require a minimum stay of 7 nights. This indicates that, generally, hosts prefer longer bookings, possibly to reduce the frequency of turnover and related maintenance tasks.

In [ ]:

Kshitija Chilbule