Project Summary -

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

This project involved exploring and cleaning a dataset to prepare it for analysis. The data exploration process involved identifying and understanding the characteristics of the data, such as the data types, missing values, and distributions of values. The data cleaning process involved identifying and addressing any issues or inconsistencies in the data, such as errors, missing values, or duplicate records and remove outliers.

Through this process, we were able to identify and fix any issues with the data, and ensure that it was ready for further analysis. This is an important step in any data analysis project, as it allows us to work with high-quality data and avoid any potential biases or errors that could affect the results. The clean and prepared data can now be used to answer specific research.

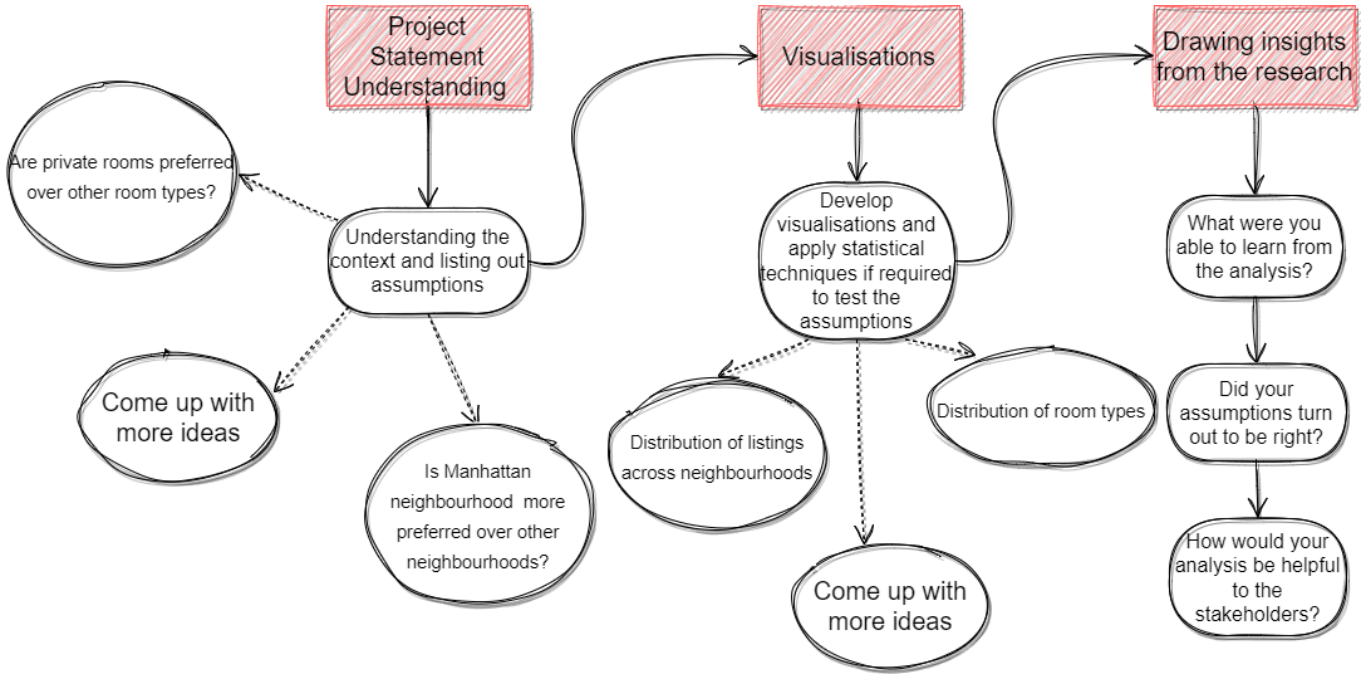
Once the data has been cleaned and prepared, now begin exploring and summarizing it with describe the data and creating visualizations, and identifying patterns and trends in the data. in explore the data, may develop the relationships between different variables or the underlying causes of certain patterns or trends and other methods.

using data visualization to explore and understand patterns in Airbnb data. We created various graphs and charts to visualize the data, and wrote observations and insights below each one to help us better understand the data and identify useful insights and patterns.

Through this process, we were able to uncover trends and relationships in the data that would have been difficult to identify through raw data alone, for example factors affecting prices and availability. We found that minimum nights, number of reviews, and host listing count are important for determining prices, and that availability varies significantly across neighborhoods. Our analysis provides useful information for travelers and hosts in the city.

The observations and insights we identified through this process will be useful for future analysis and decision-making related to Airbnb. and also Our analysis provides useful information for travelers and hosts in the city.

Project Architecture




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

%matplotlib inline


airbnb_data=pd.read_csv("/content/Airbnb Nyc 2019.csv")
airbnb_data
```

	id	name	host_id	host_name	neighbourhood_group	neighbou
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kens
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	M
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	H
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clint
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East H
...	...	...	...	...	...	
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Be Stuy
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bu
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	H
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	Hell's K
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	Hell's K

48895 rows × 16 columns

Next steps: Generate code with `airbnb_data` View recommended plots

UNDERSTANDING THE GIVEN VARIABLES

Listing_id :- This is a unique identifier for each listing in the dataset.

Listing_name :- This is the name or title of the listing, as it appears on the Airbnb website.

Host_id :- This is a unique identifier for each host in the dataset.

Host_name :- This is the name of the host as it appears on the Airbnb website.

Neighbourhood_group :- This is a grouping of neighborhoods in New York City, such as Manhattan or Brooklyn.

Neighbourhood :- This is the specific neighborhood in which the listing is located.

Latitude :- This is the geographic latitude of the listing.

Longitude :- This is the geographic longitude of the listing.

Room_type :- This is the type of room or property being offered, such as an entire home, private room, shared room.

Price :- This is the nightly price for the listing, in US dollars.

Minimum_nights :- This is the minimum number of nights that a guest must stay at the listing.

Total_reviews :- This is the total number of reviews that the listing has received.

Reviews_per_month :- This is the average number of reviews that the listing receives per month.

Host_listings_count :- This is the total number of listings that the host has on Airbnb.

Availability_365 :- This is the number of days in the next 365 days that the listing is available for booking.

Data Exploration and Data Cleaning

```
airbnb_data.shape
```

```
(48895, 16)
```

This dataset has around 48,895 observations with 16 columns and it is a mix between categorical and numeric values.

```
airbnb_data.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
```

```
airbnb_data.isnull().sum()
```

```
id                                0
name                             16
host_id                           0
host_name                        21
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                   10052
reviews_per_month             10052
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

```
# Replacing  null values

airbnb_data["name"]=airbnb_data["name"].fillna("Unknown")
airbnb_data["host_name"]=airbnb_data["host_name"].fillna("xyz")
airbnb_data["last_review"]=airbnb_data["last_review"].fillna("0")
airbnb_data["reviews_per_month"]=airbnb_data["reviews_per_month"].fillna("0")
```

```
airbnb_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
dtype: int64
```

```
# check duplicate rows in dataset
airbnb_data = airbnb_data.drop_duplicates()
airbnb_data.count()
```

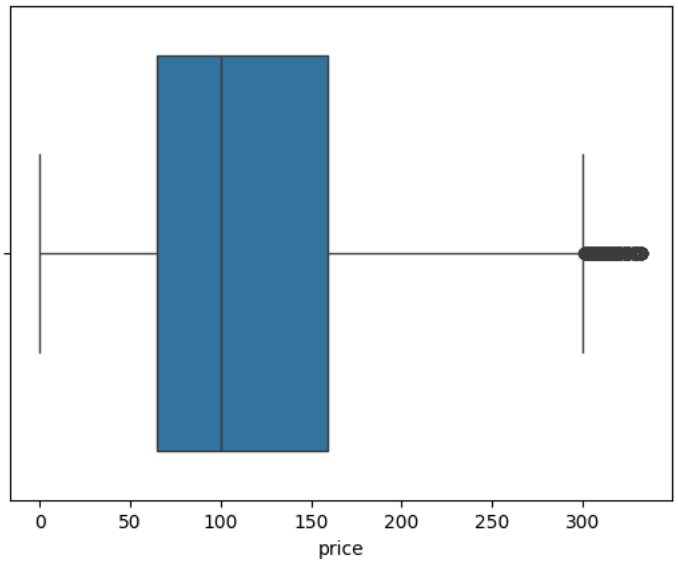
```
id                                48895
name                              48895
host_id                           48895
host_name                         48895
neighbourhood_group               48895
neighbourhood                     48895
latitude                          48895
longitude                         48895
room_type                         48895
price                             48895
minimum_nights                    48895
number_of_reviews                 48895
last_review                       48895
reviews_per_month                 48895
calculated_host_listings_count    48895
availability_365                  48895
dtype: int64
```

Describe the Dataset and removing outliers

```
# describe the DataFrame
airbnb_data.describe()
```

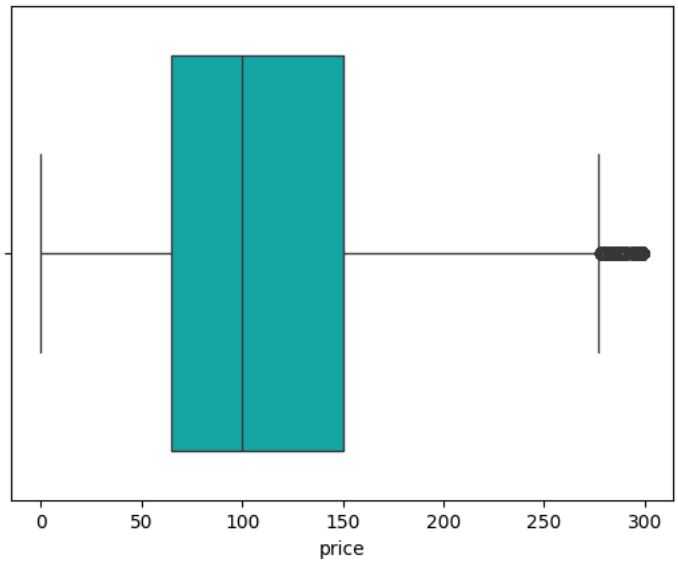
	id	host_id	latitude	longitude	price	minimum_n
count	3.633400e+04	3.633400e+04	36333.000000	36333.000000	36333.000000	36333.0(
mean	1.422079e+07	4.471385e+07	40.728640	-73.953169	147.358022	6.9(
std	8.427656e+06	5.381382e+07	0.054364	0.044079	230.736064	21.0
min	2.539000e+03	2.438000e+03	40.499790	-74.242850	0.000000	1.0(
25%	6.967039e+06	5.905590e+06	40.689100	-73.982400	69.000000	2.0(
50%	1.421927e+07	2.213123e+07	40.722280	-73.955970	105.000000	3.0(
75%	2.152909e+07	6.139196e+07	40.763490	-73.938240	170.000000	5.0(
max	2.891183e+07	2.179093e+08	40.911690	-73.712990	10000.000000	1250.0(

```
sns.boxplot(x = airbnb_data['price'])
plt.show()
```



```
# writing a outlier function for removing outliers in important columns.
def iqr_technique(DFcolumn):
  Q1 = np.percentile(DFcolumn, 25)
  Q3 = np.percentile(DFcolumn, 75)
  IQR = Q3 - Q1
  lower_range = Q1 - (1.5 * IQR)
  upper_range = Q3 + (1.5 * IQR)                        # interquantile range

  return lower_range,upper_range

lower_bound,upper_bound = iqr_technique(airbnb_data['price'])
airbnb_data = airbnb_data[(airbnb_data.price>lower_bound) & (airbnb_data.price<upper_bound)]


# so the outliers are removed from price column now check with boxplot and also check shape of new Dataframe!

sns.boxplot(x = airbnb_data['price'],color="c")
print(airbnb_data.shape)
```

```
(44977, 16)
```




```
# so here outliers are removed, see the new max price
print(airbnb_data['price'].max())
```

```
299
```

Data Visualization

(1) Distribution Of Airbnb Bookings Price Range Using Histogram

```
plt.figure(figsize=(12, 5))
sns.set_theme(style='darkgrid')
sns.distplot(airbnb_data["price"],color="c")
plt.title("AIRBNB BOOKING PRICE DISTRIBUTION")
plt.xlabel("Price")
plt.ylabel("Density")
```

```
<ipython-input-26-968b12c68064>:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(airbnb_data["price"],color="c")
Text(0, 0.5, 'Density')
```

**observations : **

The range of prices being charged on Airbnb appears to be from 20 to 330 dollars , with the majority of listings falling in the price range of 50 to 150 dollars.

The distribution of prices appears to have a peak in the 50 to 150 dollars range, with a relatively lower density of listings in higher and lower price ranges.

There may be fewer listings available at prices above 250 dollars, as the density of listings drops significantly in this range.

---

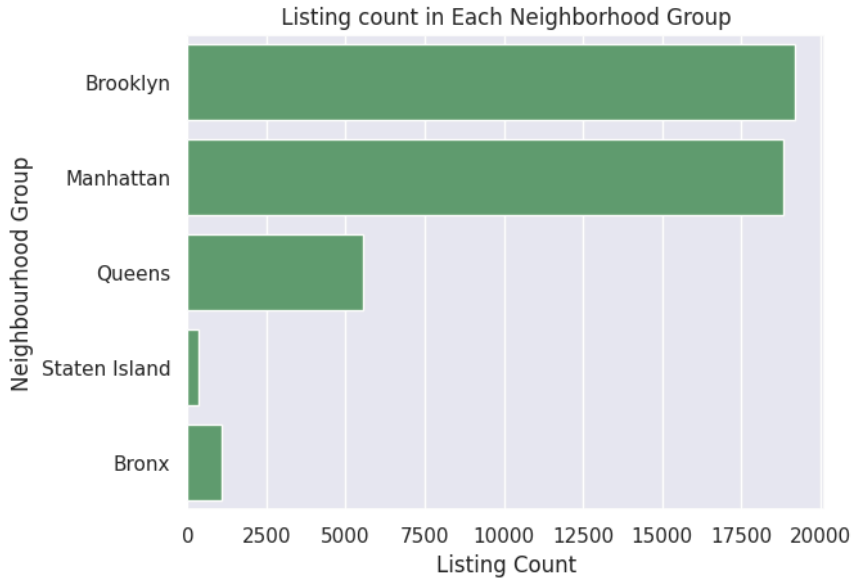
(2) Total Listing/Property count in Each Neighborhood Group using Count plot

```
neighbourgroup_avg_price=airbnb_data.groupby("neighbourhood_group")["calculated_host_listings_count"].count().sort_val
neighbourgroup_avg_price
```

```
neighbourhood_group
Brooklyn         19191
Manhattan        18826
Queens            5535
Bronx             1063
Staten Island      362
Name: calculated_host_listings_count, dtype: int64
```

```
sns.countplot(airbnb_data["neighbourhood_group"],color="g")
plt.title(" Listing count in Each Neighborhood Group ")
plt.xlabel("Listing Count")
plt.ylabel("Neighbourhood Group")
```

```
Text(0, 0.5, 'Neighbourhood Group')
```



Observations -->

As we can see from the chart above Manhattan neighborhood has the highest number of Airbnb's

1. Manhattan
2. Brooklyn
3. Queens
4. Bronx
5. Staten Island And Manhattan and Brooklyn has more than 75% of the AirBnb's.

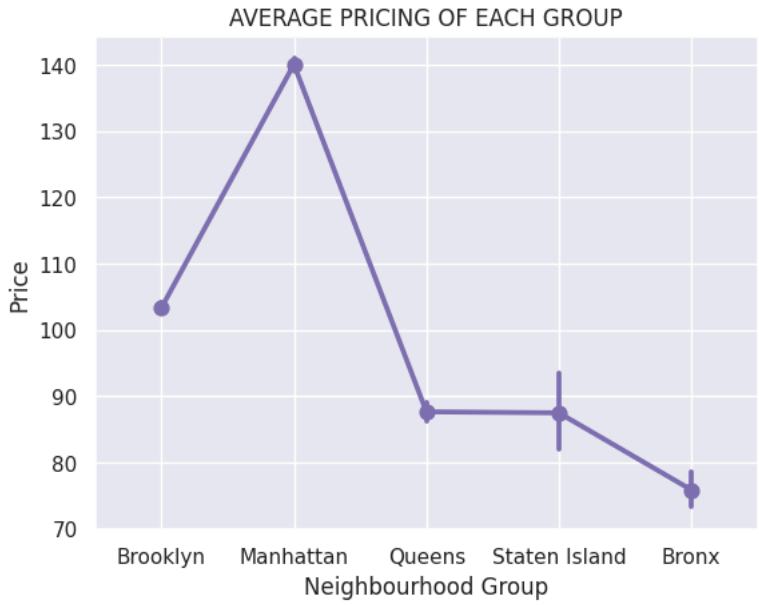
---

(3) Average Price Of Each Neighborhood Group using Point Plot

```
neighbour_avg_price=airbnb_data.groupby("neighbourhood_group")["price"].mean().sort_values(ascending=False)
neighbour_avg_price
```

```
neighbourhood_group
Manhattan        140.092000
Brooklyn         103.345214
Queens            87.649684
Staten Island     87.488950
Bronx             75.858890
Name: price, dtype: float64
```

```
sns.pointplot(x=airbnb_data["neighbourhood_group"],y=airbnb_data["price"],color="m")
plt.title("AVERAGE PRICING OF EACH GROUP")
plt.xlabel("Neighbourhood Group")
plt.ylabel("Price")
plt.grid(True)
```



Observations -->

The average price of a listing in New York City varies significantly across different neighborhoods, with Manhattan having the highest 146 dollars/day average price and the Bronx having the lowest near 77 dollars/day.

In second graph price distribution is very high in Manhattan and Brooklyn. but Manhattan have more varity in price range, you can see in second violinplot.

The average price increases as you move from the outer boroughs (Bronx, Brooklyn, Queens, and Staten Island) towards the center of the city (Manhattan).

The average price in queens and Staten Island is relatively similar, despite being in different parts of the city.

The data suggests that the overall cost of living in New York City is higher in the center of the city (Manhattan) compared to the outer boroughs. This is likely due to the fact that Manhattan is the most densely populated and commercially important borough, and therefore has higher demand for housing in the centrally located neighborhoods

---

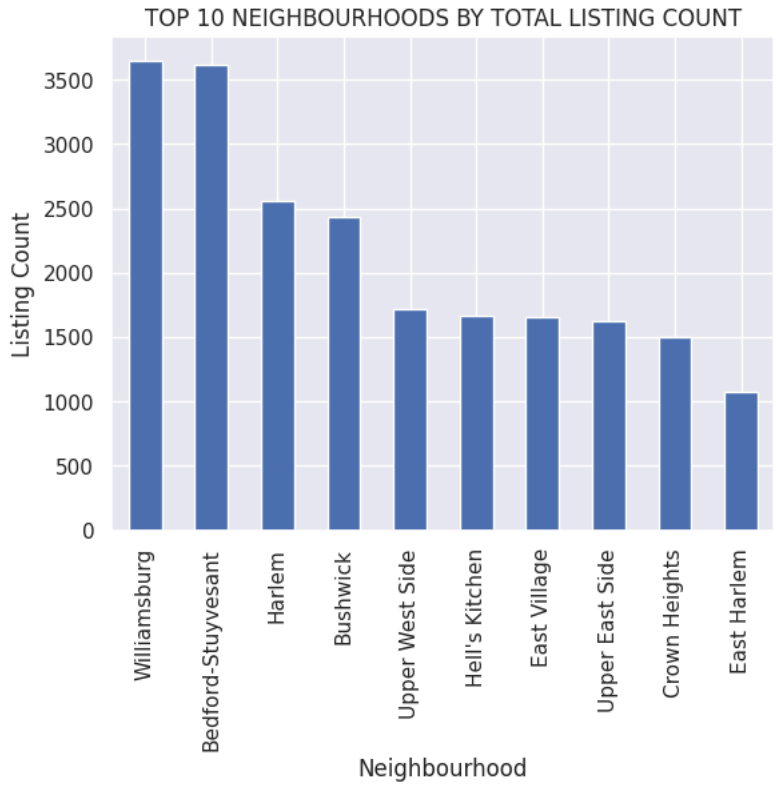
(4) Top Neighborhoods by Listing/property using Bar plot

```
top_neighbour_count=airbnb_data.groupby("neighbourhood")["calculated_host_listings_count"].count().sort_values(ascendi
top_neighbour_count
```

```
neighbourhood
Williamsburg          3653
Bedford-Stuyvesant    3616
Harlem                2556
Bushwick              2430
Upper West Side       1721
Hell's Kitchen        1663
East Village          1660
Upper East Side       1629
Crown Heights         1500
East Harlem           1070
Name: calculated_host_listings_count, dtype: int64
```

```
top_neighbour_count.plot(kind="bar",color="b")
plt.title("TOP 10 NEIGHBOURHOODS BY TOTAL LISTING COUNT")
plt.xlabel("Neighbourhood")
plt.ylabel("Listing Count")
```

```
Text(0, 0.5, 'Listing Count')
```



Observations -->

The top neighborhoods in New York City in terms of listing counts are Williamsburg, Bedford-Stuyvesant, Harlem, Bushwick, and the Upper West Side.

The top neighborhoods are primarily located in Brooklyn and Manhattan. This may be due to the fact that these boroughs have a higher overall population and a higher demand for housing.

The number of listings alone may not be indicative of the overall demand for housing in a particular neighborhood, as other factors such as the cost of living and the availability of housing may also play a role.

---

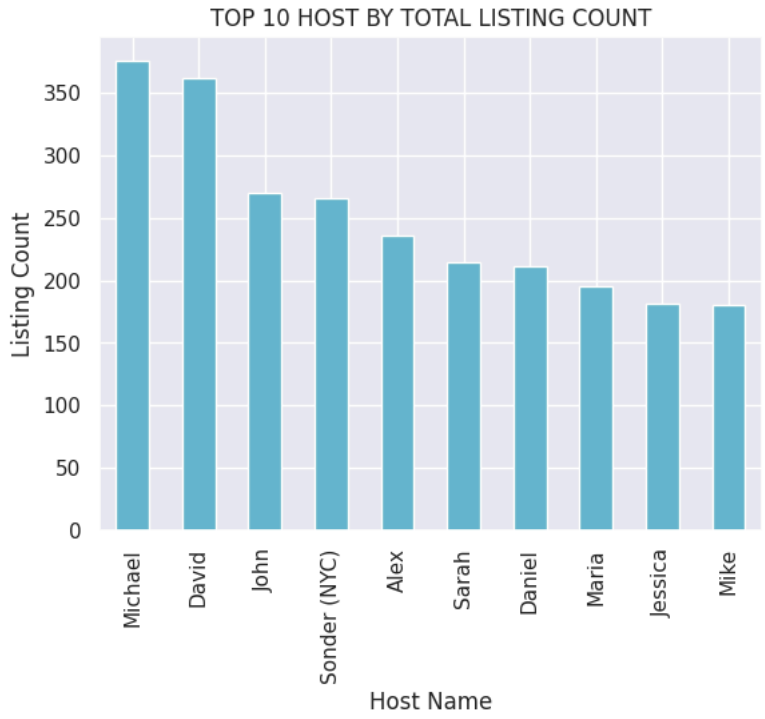
(5) Top Hosts With More Listing/Property using Bar chart

```
top_host=airbnb_data.groupby("host_name")["calculated_host_listings_count"].count().sort_values(ascending=False)[:10]
top_host
```

```
host_name
Michael         376
David           362
John            270
Sonder (NYC)    266
Alex            236
Sarah           215
Daniel          211
Maria           195
Jessica         181
Mike            180
Name: calculated_host_listings_count, dtype: int64
```

```
top_host.plot(kind="bar",color="c")
plt.title("TOP 10 HOST BY TOTAL LISTING COUNT")
plt.xlabel("Host Name")
plt.ylabel("Listing Count")
```

```
Text(0, 0.5, 'Listing Count')
```



Observations -->

The top three hosts in terms of total listings are Michael, David, and John, who have 383, 368, and 276 listings, respectively.

There is a relatively large gap between the top two hosts and the rest of the hosts. For example, john has 276 listings, which is significantly fewer than Michael's 383 listings.

In this top10 list Mike has 184 listings, which is significantly fewer than Michael's 383 listings. This could indicate that there is a lot of variation in the success of different hosts on Airbnb.

There are relatively few hosts with a large number of listings. This could indicate that the Airbnb market is relatively competitive, with a small number of hosts dominating a large portion of the market.

---

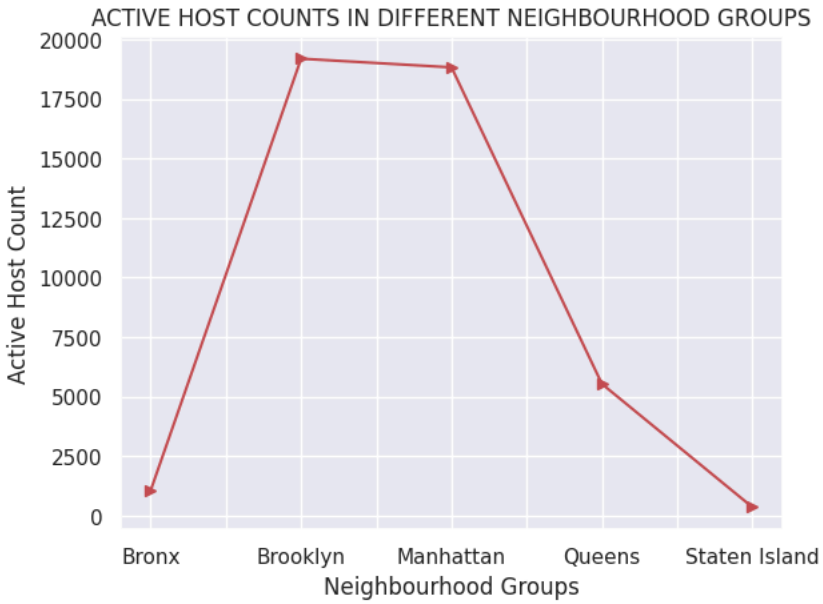
(6) Number Of Active Hosts Per Location Using Line Chart

```
active_host_count=airbnb_data.groupby("neighbourhood_group")["calculated_host_listings_count"].count()
active_host_count
```

```
neighbourhood_group
Bronx             1063
Brooklyn         19191
Manhattan        18826
Queens            5535
Staten Island      362
Name: calculated_host_listings_count, dtype: int64
```

```
active_host_count.plot(kind="line",marker=">",color="r")
plt.grid(True)
plt.title("ACTIVE HOST COUNTS IN DIFFERENT NEIGHBOURHOOD GROUPS")
plt.xlabel("Neighbourhood Groups")
plt.ylabel("Active Host Count")
```

```
Text(0, 0.5, 'Active Host Count')
```



Observations -->

Manhattan has the largest number of hosts with 19501,Brooklyn has the second largest number of hosts with 19415.

After that Queens with 5567 and the Bronx with 1070. while Staten Island has the fewest with 365.

Brooklyn and Manhattan have the largest number of hosts, with more than double the number of hosts in Queens and more than 18 times the number of hosts in the Bronx.

---

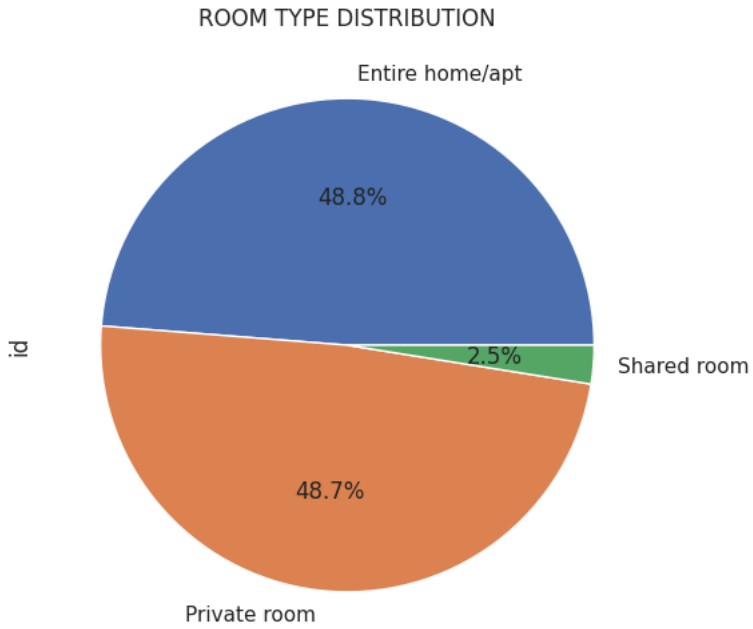
(7) Total Counts Of Each Room Type

```
room_types=airbnb_data.groupby("room_type")["id"].count()
room_types
```

```
room_type
Entire home/apt    21929
Private room       21915
Shared room         1133
Name: id, dtype: int64
```

```
plt.figure(figsize=(6,6))
room_types.plot(kind="pie",autopct="%.1f%%")
plt.title("ROOM TYPE DISTRIBUTION")
```

```
Text(0.5, 1.0, 'ROOM TYPE DISTRIBUTION')
```



Observations -->

The majority of listings on Airbnb are for entire homes or apartments, with 22784 listings, followed by private rooms with 21996 listings, and shared rooms with 1138 listings.

There is a significant difference in the number of listings for each room type. For example, there are almost 20 times as many listings for entire homes or apartments as there are for shared rooms.

The data suggests that travelers using Airbnb have a wide range of accommodation options to choose from, including private rooms and entire homes or apartments

---

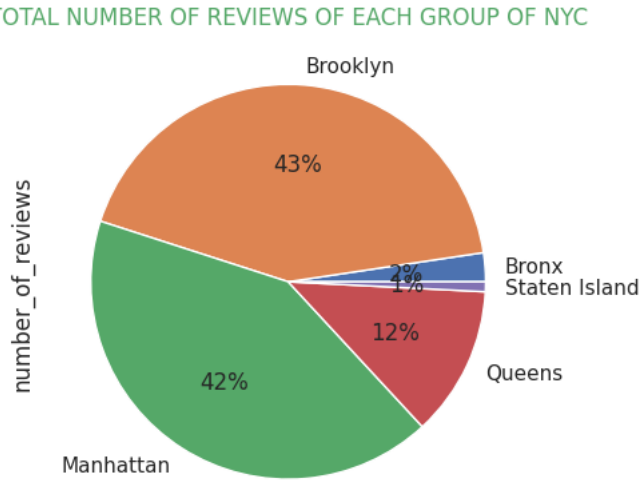
(8) Total Reviews by Each Neighborhood Group using Pie Chart

```
review_total=airbnb_data.groupby("neighbourhood_group")["number_of_reviews"].count()
review_total
```

```
neighbourhood_group
Bronx             1063
Brooklyn         19191
Manhattan        18826
Queens            5535
Staten Island      362
Name: number_of_reviews, dtype: int64
```

```
review_total.plot(kind="pie",autopct="%.f%%")
plt.title("TOTAL NUMBER OF REVIEWS OF EACH GROUP OF NYC",color="g")
```

```
Text(0.5, 1.0, 'TOTAL NUMBER OF REVIEWS OF EACH GROUP OF NYC')
```



Observations -->

Brooklyn has the largest share of total reviews on Airbnb, with 43.3%, followed by Manhattan with 38.9%.

Queens has the third largest share of total reviews, with 14.2%, followed by the Bronx with 2.6% and Staten Island with 1.0%.

The data suggests that Airbnb is more popular in Brooklyn and Manhattan compared to the other neighborhood groups.

Despite having fewer listings, Brooklyn has more reviews on Airbnb compared to Manhattan. This could indicate that Airbnb users in Brooklyn are more likely to leave reviews, or that the listings in Brooklyn are more popular or successful in generating positive reviews. It is worth noting that there could be a number of other factors that could contribute to this difference in reviews, such as the quality of the listings or the characteristics of the travelers who use Airbnb in these areas.

---

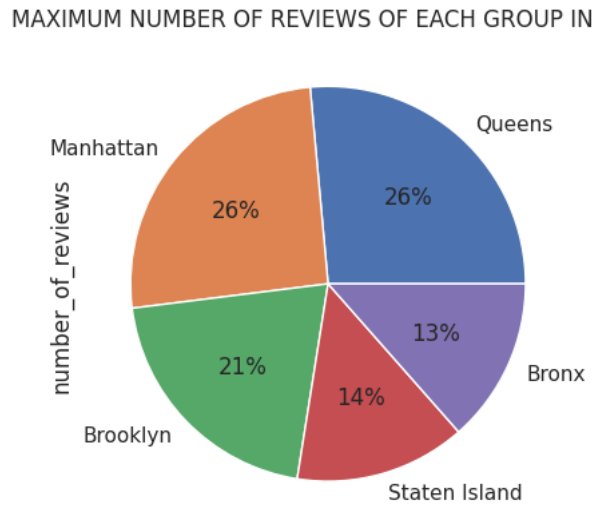
(9) Number of Max. Reviews by Each Neighborhood Group using Pie Chart

```
review_max=airbnb_data.groupby("neighbourhood_group")["number_of_reviews"].max().sort_values(ascending=False)
review_max
```

```
neighbourhood_group
Queens           629
Manhattan        607
Brooklyn         488
Staten Island    333
Bronx            321
Name: number_of_reviews, dtype: int64
```

```
review_max.plot(kind="pie",autopct="%.f%%")
plt.title("MAXIMUM NUMBER OF REVIEWS OF EACH GROUP IN NYC")
```

```
Text(0.5, 1.0, 'MAXIMUM NUMBER OF REVIEWS OF EACH GROUP IN NYC')
```



Observations –>

Queens and Manhattan seem to be the most popular neighborhoods for reviewing, as they have both high number of maximum reviews.

Queens has the highest percentage of reviews at 26.5%, but it has the third highest number of listings, behind Manhattan and Brooklyn. This suggests that Queens may be a particularly popular destination for tourists or visitors, even though it has fewer listings compared to Manhattan and Brooklyn.

Manhattan and Brooklyn also have a high percentage of reviews, at 25.5% & 20.5%. This indicates that it is a popular destination for tourists or visitors as well. (number of listings higher than queens)

Overall, this data suggests that Queens, Manhattan, and Brooklyn are the most popular neighborhoods for tourists or visitors, based on the high number of reviews they receive.

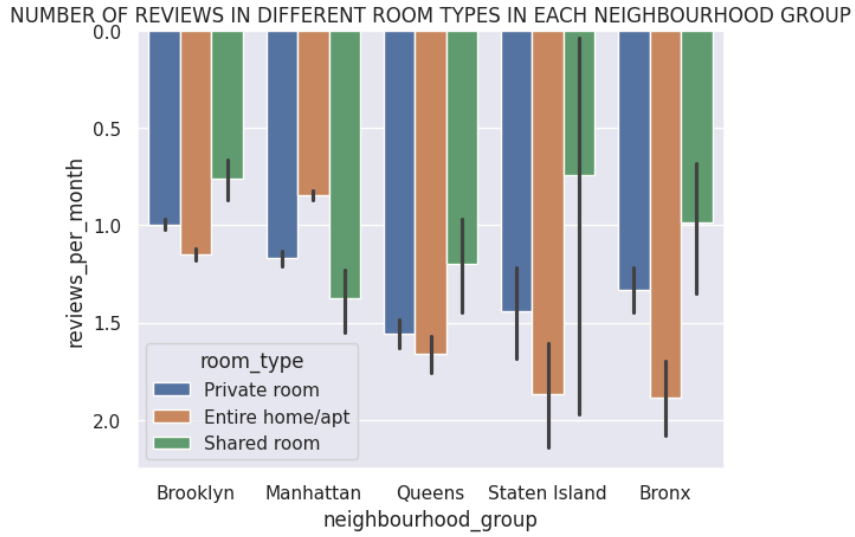
---

(10) most reviewed room type per month in neighbourhood groups

```
sns.barplot(x=airbnb_data["neighbourhood_group"],y=airbnb_data["reviews_per_month"],hue=airbnb_data["room_type"])
```

```
plt.title("NUMBER OF REVIEWS IN DIFFERENT ROOM TYPES IN EACH NEIGHBOURHOOD GROUP")
```

```
Text(0.5, 1.0, 'NUMBER OF REVIEWS IN DIFFERENT ROOM TYPES IN EACH NEIGHBOURHOOD
GROUP')
```



Observations –>

1. Brooklyn has Highest number of Private Rooms.

2. Manhattan has Highest number of Entire Home/Apartment.
3. Manhattan has Highest number of Shared Room.

---

(11) Stay Requirement counts by Minimum Nights using Bar chart

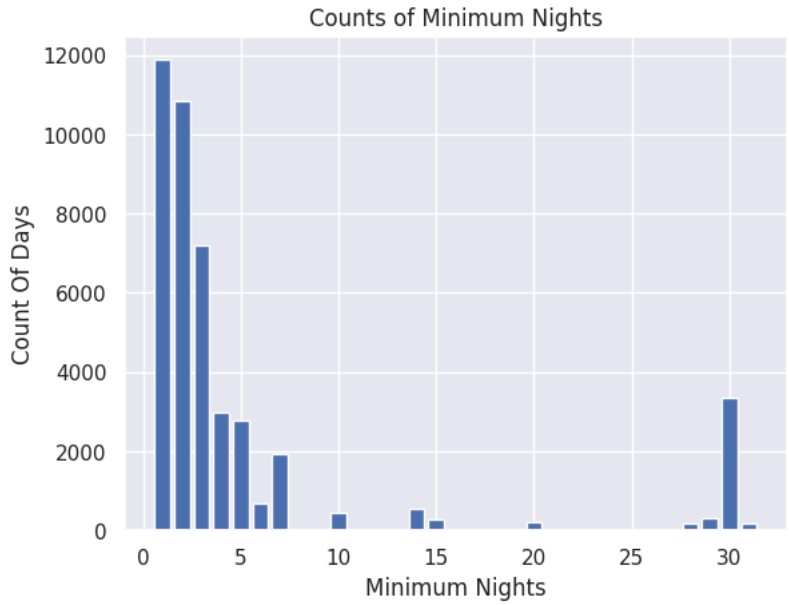
```
min_nightc=airbnb_data.groupby("minimum_nights").size().reset_index(name='count').sort_values('count',ascending=False)
min_nightc
```

	minimum_nights	count
0	1	11881
1	2	10847
2	3	7210
29	30	3350
3	4	2988
4	5	2760
6	7	1918
5	6	669
13	14	536
9	10	456
28	29	322
14	15	268
19	20	215
30	31	182
27	28	173

Next steps: Generate code with min_nightc View recommended plots

```
count=min_nightc["count"]
min_night=min_nightc["minimum_nights"]
plt.bar(min_night,count)
plt.xlabel("Minimum Nights")
plt.ylabel("Count Of Days")
plt.title("Counts of Minimum Nights")
```

```
Text(0.5, 1.0, 'Counts of Minimum Nights')
```



Observations -->

The majority of listings on Airbnb have a minimum stay requirement of 1 or 2 nights, with 12067 and 11080 listings, respectively.

The number of listings with a minimum stay requirement decreases as the length of stay increases, with 7375 listings requiring a minimum stay of 3 nights, and so on.

There are relatively few listings with a minimum stay requirement of 30 nights or more, with 3489 and 189 listings, respectively.

---

(12) Correlation Heatmap Visualization

```
# Calculate pairwise correlations between columns
corr=airbnb_data.corr()
corr
```

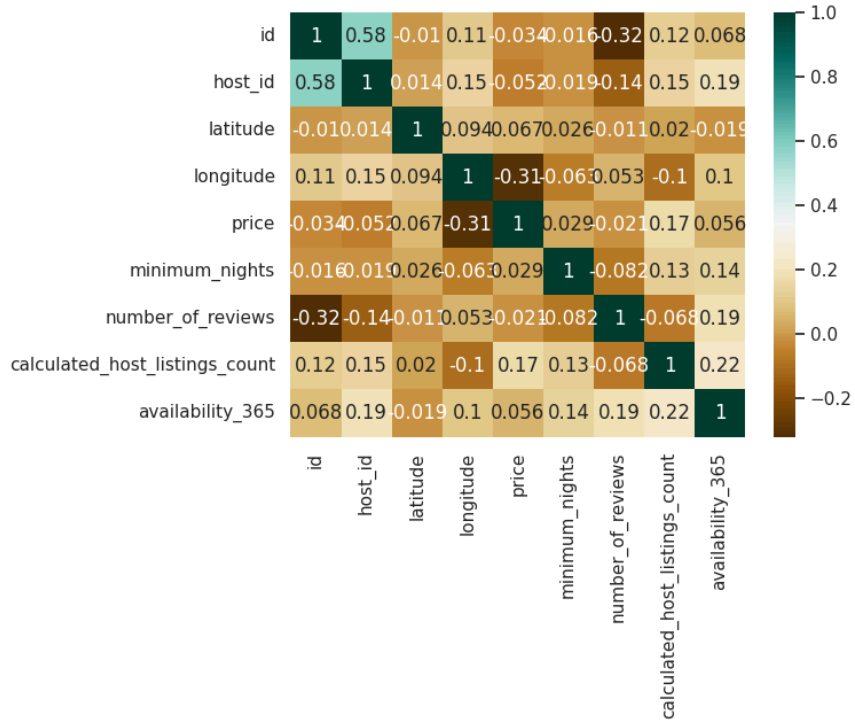
```
<ipython-input-71-3044a3f09410>:1: FutureWarning: The default value of numeric_only :
  corr=airbnb_data.corr()
```

	id	host_id	latitude	longitude	price	mini
id	1.000000	0.579887	-0.010079	0.105418	-0.033853	
host_id	0.579887	1.000000	0.013793	0.148663	-0.051731	
latitude	-0.010079	0.013793	1.000000	0.094165	0.066675	
longitude	0.105418	0.148663	0.094165	1.000000	-0.307471	
price	-0.033853	-0.051731	0.066675	-0.307471	1.000000	
minimum_nights	-0.015866	-0.019224	0.026031	-0.063088	0.029468	
number_of_reviews	-0.319709	-0.135420	-0.011413	0.052625	-0.020572	
calculated_host_listings_count	0.118975	0.147845	0.020304	-0.103692	0.166843	
availability_365	0.068291	0.190773	-0.019410	0.100499	0.055862	

Next steps: Generate code with corr View recommended plots

```
# Visualize correlations as a heatmap
sns.heatmap(corr, cmap='BrBG',annot=True)
```

```
<Axes: >
```



Observations -->

There is a moderate positive correlation (0.58) between the host_id and id columns, which suggests that hosts with more listings are more likely to have unique host IDs.

There is a weak positive correlation (0.17) between the price column and the calculated_host_listings_count column, which suggests that hosts with more listings tend to charge higher prices for their listings.

There is a moderate positive correlation (0.23) between the calculated_host_listings_count column and the availability_365 column, which suggests that hosts with more listings tend to have more days of availability in the next 365 days.

There is a strong positive correlation (0.58) between the number_of_reviews column and the reviews_per_month column, which suggests that listings with more total reviews tend to have more reviews per month.

---

(12) Pair Plot Visualization

```
sns.pairplot(airbnb_data)
plt.title("FEATURE DISTRIBUTION")
plt.show()
```