

✓ **Project Name - Airbnb Booking Analysis**

Project Type - Exploratory Data Analysis

Contribution - Individual

Name - Aastha Gadwal

✓ **Project Summary -**

Purpose of the Analysis The aim of this analysis is to understand the factors influencing Airbnb pricing in New York City and to identify patterns among all variables. This study offers valuable insights for both travelers and hosts, as well as actionable recommendations for Airbnb business strategies.

Project Overview The project began with a comprehensive exploration and cleaning of the dataset to prepare it for analysis. During the data exploration phase, we examined various attributes of the data, such as data types, missing values, and value distributions. The cleaning process addressed inconsistencies, such as errors, missing entries, duplicate records, and outliers.

By resolving these issues, we ensured the dataset was of high quality and ready for detailed analysis. This foundational step is critical in any data-driven project to minimize biases and inaccuracies that could distort findings. With the dataset refined and prepared, we proceeded to answer specific research questions.

Data Exploration and Summarization After cleaning the dataset, we analyzed it further by summarizing the data, creating visualizations, and identifying patterns and trends. This exploration involved uncovering relationships between different variables and understanding the underlying causes of observed patterns.

We utilized data visualization techniques to present and interpret the Airbnb dataset effectively. Graphs and charts were generated to illustrate key insights, with observations and interpretations accompanying each visualization. These visual aids facilitated a deeper understanding of trends and relationships in the data that might not have been evident from raw data alone.

Key Findings Our analysis revealed significant trends and factors influencing Airbnb prices and availability in New York City. For example:

Minimum nights, number of reviews, and host listing counts are critical determinants of pricing. Availability varies considerably across neighborhoods, reflecting different market dynamics.

These insights are not only valuable for Airbnb hosts aiming to optimize pricing strategies but also for travelers seeking cost-effective accommodation options.

Implications and Applications The findings from this analysis provide a foundation for future research and decision-making related to Airbnb operations. Travelers and hosts alike can leverage these insights to make informed choices, while businesses in the industry can use them to refine strategies and improve market positioning.

➤ **GitHub Link -**

↳ 1 cell hidden

✓ **Problem Statement**

- 1) Which neighborhoods in New York City are the most popular for Airbnb rentals?
- 2) How do rental prices and availability differ across neighborhoods?
- 3) What trends have emerged in New York City's Airbnb market over time? Are there noticeable changes in the number of listings, pricing, or occupancy rates?
- 4) Are there specific patterns in the types of properties rented on Airbnb in New York City? Which property types tend to be more popular or command higher prices?
- 5) What factors appear to influence the pricing of Airbnb rentals in New York City? Are there specific variables correlated with rental costs?
- 6) Which areas in New York City are most suitable for hosts to invest in properties that are affordable yet attract high traffic?
- 7) How does the length of stay vary for Airbnb rentals across different neighborhoods in New York City? Are there neighborhoods that cater more to longer stays or shorter visits?
- 8) Is there a relationship between the ratings of Airbnb listings in New York City and their prices? Do higher-rated properties tend to be more expensive?
- 9) What is the total number of reviews and the maximum reviews recorded for each neighborhood group in New York City?
- 10) Which room type is reviewed the most in each neighborhood group, and how does this trend vary monthly?
- 11) What are the best property locations for travelers seeking Airbnb rentals in New York City?
- 12) Which locations in New York City offer the best opportunities for Airbnb hosts to maximize bookings and profitability?

13)How do rental prices vary across neighborhood groups in New York City? What are the patterns and differences?

✓ Define Your Business Objective?

To analyze the Airbnb market in New York City to uncover key trends, pricing strategies, and factors influencing demand, providing actionable insights for hosts to maximize profitability and for travelers to make informed booking decisions.

✓ General Guidelines : -

1. Well-structured, formatted, and commented code is required.
2. Exception Handling, Production Grade Code & Deployment Ready Code will be a plus. Those students will be awarded some additional credits.

The additional credits will have advantages over other students during Star Student selection.

[Note: - Deployment Ready Code is defined as, the whole .ipynb notebook should be without a single error logged.]



3. Each and every logic should have proper comments.
4. You may add as many number of charts you want. Make Sure for each and every chart the following format should be answered.

Chart visualization code

- Why did you pick the specific chart?
 - What is/are the insight(s) found from the chart?
 - Will the gained insights help creating a positive business impact? Are there any insights that lead to negative growth? Justify with specific reason.
5. You have to create at least 20 logical & meaningful charts having important insights.

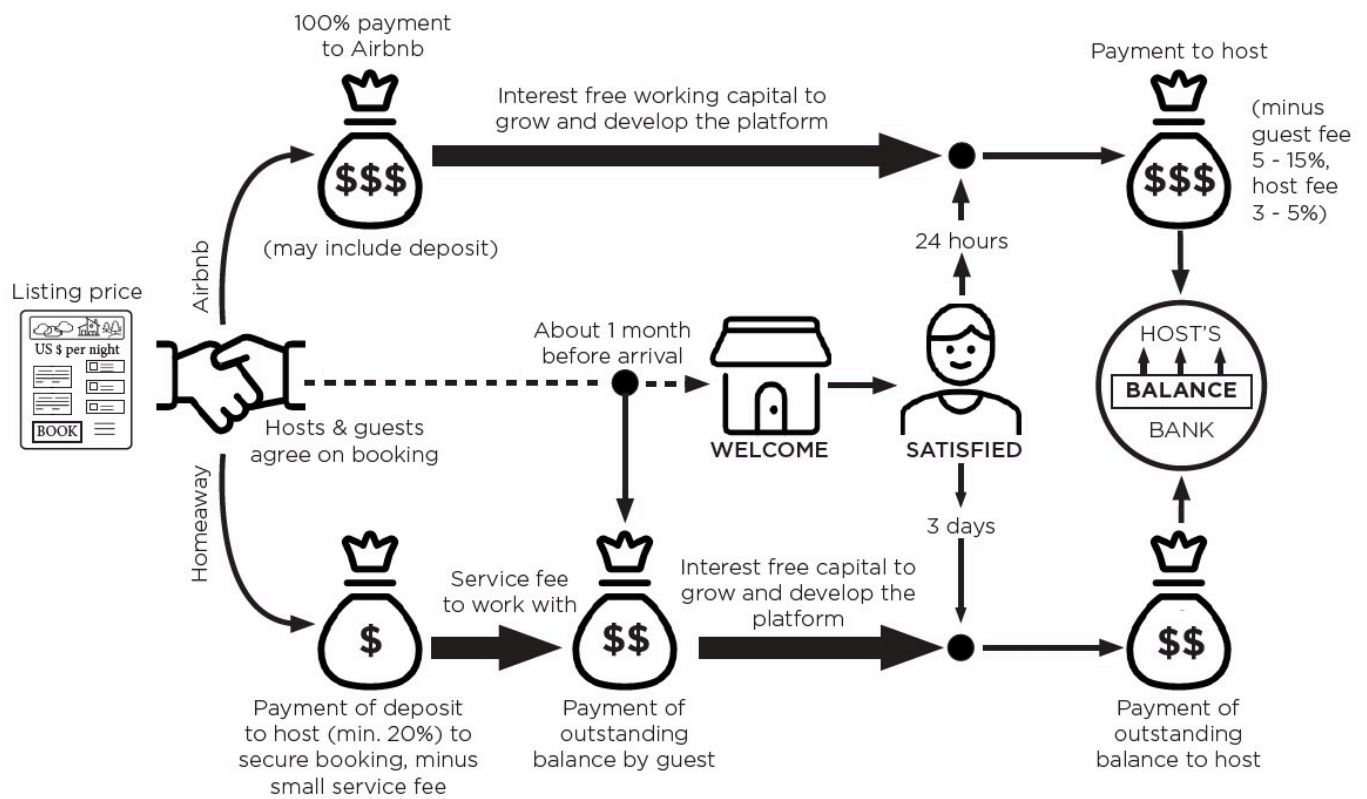
[Hints : - Do the Vizualization in a structured way while following "UBM" Rule.

U - Univariate Analysis,

B - Bivariate Analysis (Numerical - Categorical, Numerical - Numerical, Categorical - Categorical)

M - Multivariate Analysis]

✓ *Let's Begin !*



✓ *1. Know Your Data*

✓ Import Libraries

```
# Importing all the required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import drive
```

✓ Dataset Loading

```
#Load Dataset
file_id="1uUFvm-Murnx3UclbWBjJZLn69E-sncZf"
url = f"https://drive.google.com/uc?id={file_id}"
```

```
data = pd.read_csv(url)
display(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	Mi
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	Bus
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 data](#)

[View recommended plots](#)

[New interactive sheet](#)

Dataset First View

```
# Dataset First Look
data.head()
```



	id	name	host_id	host_name	neighbourhood_group	neighbourhood	lat
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40

Next steps:

[Generate code with data](#)[View recommended plots](#)[New interactive sheet](#)

Dataset Rows & Columns count

```
# Dataset Rows & Columns count
data.shape
```



(48895, 16)

Dataset Information

```
# Dataset Info
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
```

▼ Duplicate Values

```
# Dataset Duplicate Value Count
data.duplicated().sum()
```

 0

▼ Missing Values/Null Values

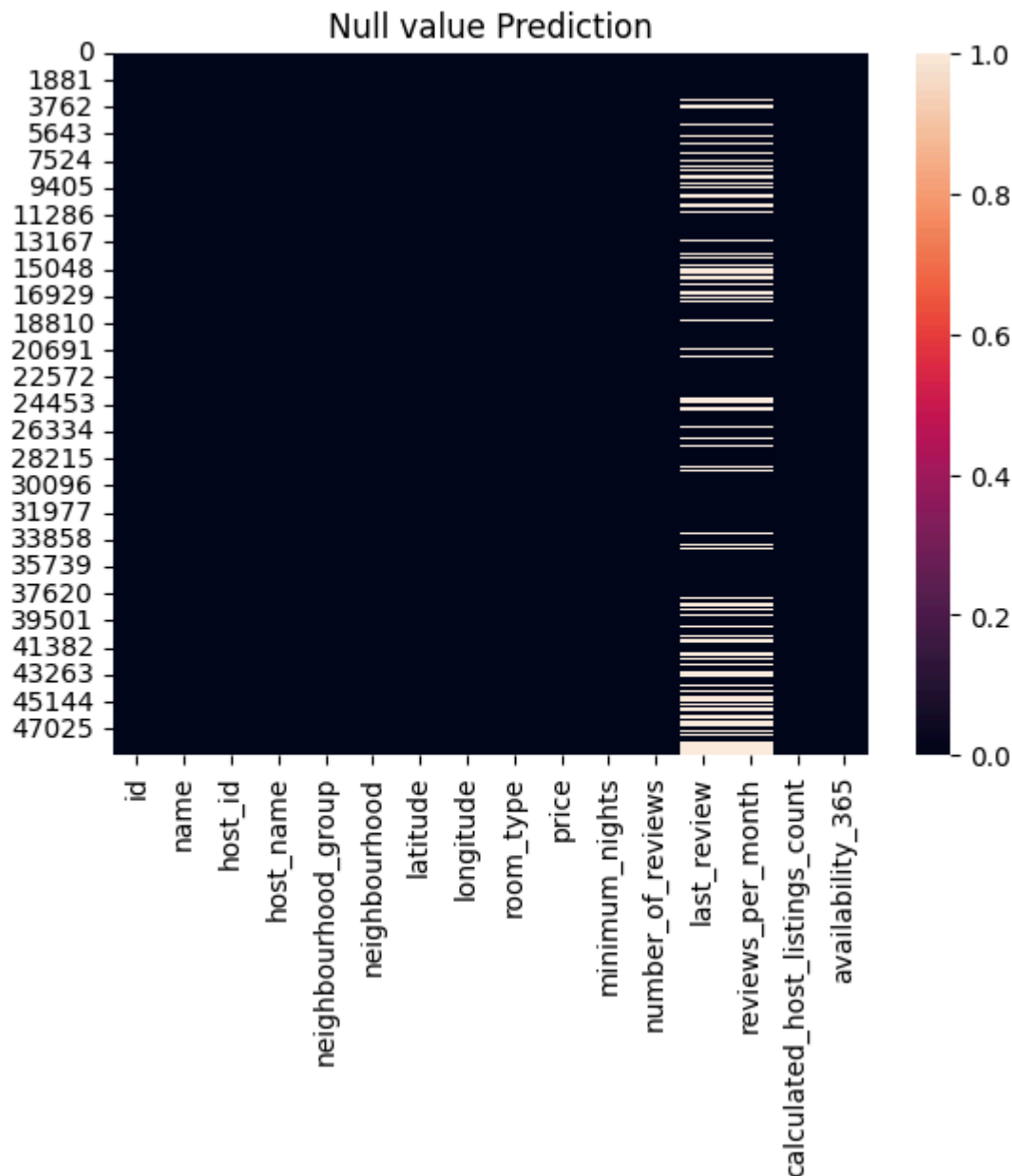
```
# Missing Values/Null Values Count
data.isnull().sum()
```



	0
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

```
# Visualizing the missing values
sns.heatmap(data=data.isnull())
plt.title('Null value Prediction')
plt.show()
```

✓ What did you know about your dataset?

- 1) host_name and name does not have many null values, so first we are good to drop those rows.
- 2) Now, the columns last_review and reviews_per_month have total 10052 null values each.
- 3) last_review column is not required for our analysis as compared to number_of_reviews & reviews_per_month. We're good to drop this column
- 4) So, name, host_name, neighbourhood_group, neighbourhood and room_type fall into categorical variable category.
- 5) While id, host_id, latitude, longitude, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365 are numerical variables

✓ 2. Understanding Your Variables

```
# Dataset Columns
data.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365'],
      dtype='object')
```

```
# Dataset Describe
data.describe()
```

```

count    4.889500e+04  4.889500e+04  48895.000000  48895.000000  48895.000000  48895.00
mean    1.901714e+07  6.762001e+07   40.728949   -73.952170   152.720687    7.02
std     1.098311e+07  7.861097e+07    0.054530    0.046157   240.154170   20.51
min     2.539000e+03  2.438000e+03   40.499790   -74.244420    0.000000    1.00
25%     9.471945e+06  7.822033e+06   40.690100   -73.983070    69.000000    1.00
50%     1.967728e+07  3.079382e+07   40.723070   -73.955680   106.000000    3.00
75%     2.915218e+07  1.074344e+08   40.763115   -73.936275   175.000000    5.00
max     3.648724e+07  2.743213e+08   40.813060   -73.712990  10000.000000  1250.00
```

✓ Variables Description

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.

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

calculated_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.

✓ Check Unique Values for each variable.

```
# Check Unique Values for each variable.
data.nunique(axis=0)
```



	0
id	48895
name	47905
host_id	37457
host_name	11452
neighbourhood_group	5
neighbourhood	221
latitude	19048
longitude	14718
room_type	3
price	674
minimum_nights	109
number_of_reviews	394
last_review	1764
reviews_per_month	937
calculated_host_listings_count	47
availability_365	366

dtype: int64

✓ 3. *Data Wrangling*

✓ Data Wrangling Code

```
# Write your code to make your dataset analysis ready.
```

```
#STEP1 - Identifying missing values
```

```
data.isnull().sum() # Counting missing values in each column
```



	0
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

```
#STEP2 - Dropping rows with missing values(from the columns with very less missing values
```

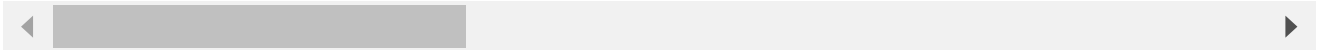
```
data.dropna(subset=['name', 'host_name'])
```

```
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	Mi
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	Bus
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 data

View recommended plots

New interactive sheet

```
data.shape #changes are done in original dataframe
```



(48895, 16)

```
#STEP3 - Removing duplicates
data.duplicated().sum() #Checking for duplicates
```

#As output is zero hence no duplicates are present

 0

#Dropping 'last_review' column

```
data = data.drop(['last_review'], axis=1)
```

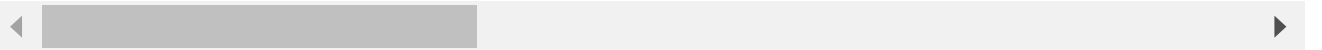
last_review column is not required for our analysis as compared to number_of_reviews & reviews_per_month. We're good to drop this column.

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	Mi
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	Bus
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 × 15 columns



Next steps:

[Generate code with data](#)

[View recommended plots](#)

[New interactive sheet](#)

data.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 15 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 reviews_per_month   38843 non-null float64
13 calculated_host_listings_count 48895 non-null int64
14 availability_365     48895 non-null int64
dtypes: float64(3), int64(7), object(5)
memory usage: 5.6+ MB
```

#STEP4 - Filling the null value with Unknown

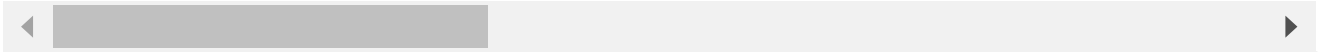
#Filling missing values of 'last_review' and 'review_per_month' column with 0

```
data.fillna({'reviews_per_month':0},inplace=True)
display(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	Mi
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	Bus
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 × 15 columns



Next steps:

Generate code with data

View recommended plots

New interactive sheet

```
data.isnull().sum()      #no null values present
```



0

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
reviews_per_month	0
calculated_host_listings_count	0
availability_365	0

dtype: int64

#STEP5 - changing the data type

#Converting datatype of the mentioned column

```
data['reviews_per_month']=pd.to_datetime(data['reviews_per_month'])  
data.dtypes
```



0

id	int64
name	object
host_id	int64
host_name	object
neighbourhood_group	object
neighbourhood	object
latitude	float64
longitude	float64
room_type	object
price	int64
minimum_nights	int64
number_of_reviews	int64
reviews_per_month	datetime64[ns]
calculated_host_listings_count	int64
availability_365	int64

dtype: object

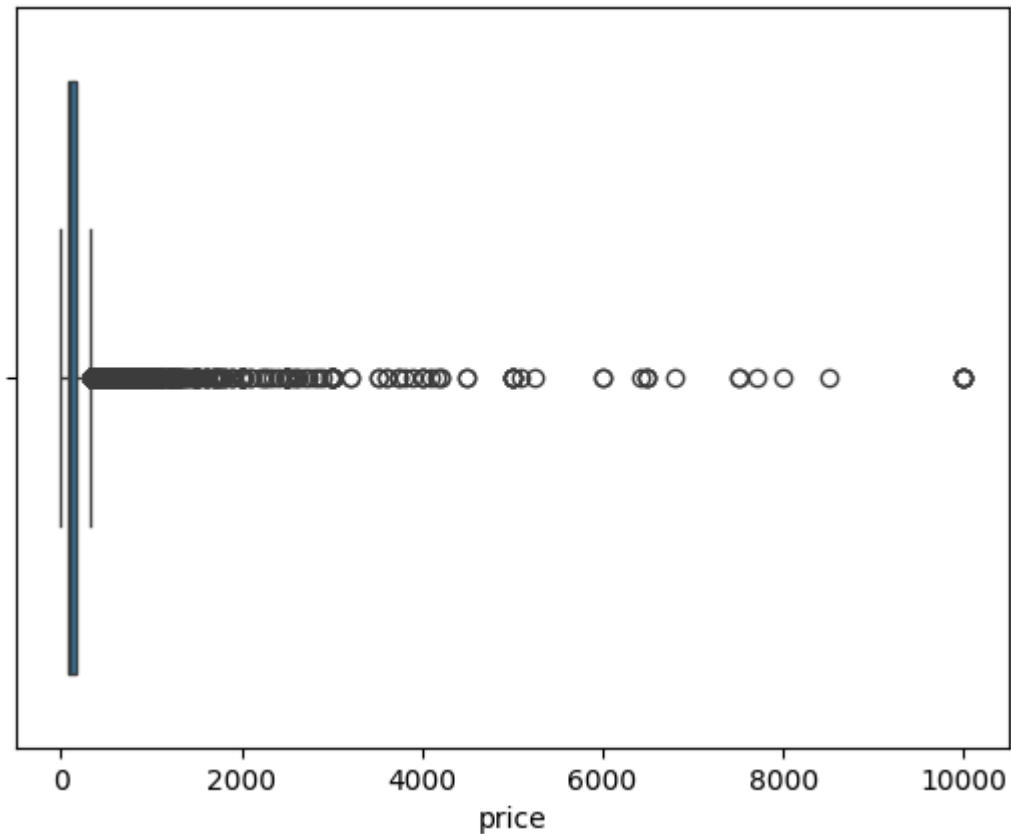
#STEP6 - Handle Outliers

#As price column is very important so we will handle outlier of 'price' column first

```
sns.boxplot(data=data,x='price')  
plt.title('Handling Outlier')  
plt.show()
```



Handling Outlier

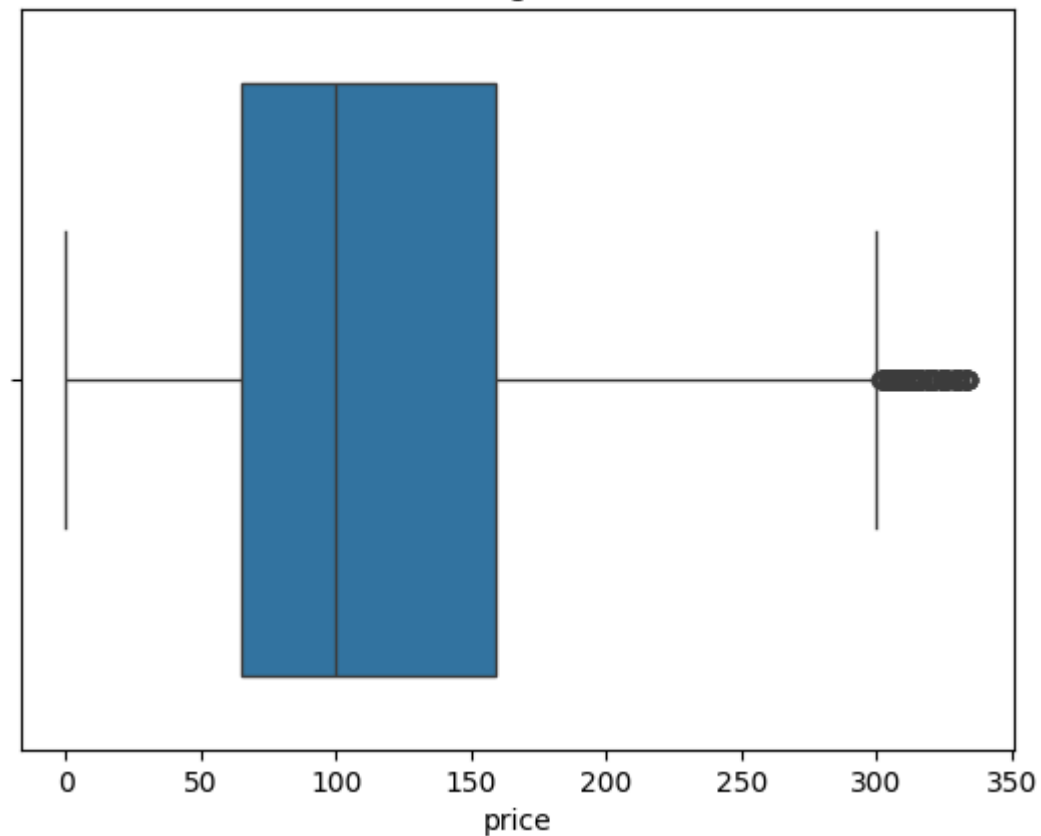


```
#Using IQR technique
Q1 = data['price'].quantile(0.25)
Q3 = data['price'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
data = data[(data['price'] >= lower_bound) & (data['price'] <= upper_bound)] #tc

#Outliers are removed. Let's check again,the box plot and shape of the data
sns.boxplot(data=data,x='price')
plt.title('Handling Outlier')
plt.show()
print(data.shape)
```



Handling Outlier



(45923, 15)

#STEP - Creating new feature

```
#data['last_review_year']=data['last_review'].dt.year
```

```
data['reviews_per_month_year']=data['reviews_per_month'].dt.year
```

```
data['reviews_per_month_year']
```



<ipython-input-144-eaf9cd39f0a7>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/usage/10min.html#working-with-copies>
data['reviews_per_month_year']=data['reviews_per_month'].dt.year

reviews_per_month_year	
0	1970
1	1970
2	1970
3	1970
4	1970
...	...
48890	1970
48891	1970
48892	1970
48893	1970
48894	1970

45923 rows × 1 columns

dtype: int32

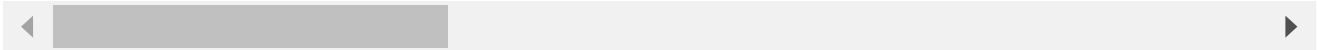


data



	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	Mi
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	Bus
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

45923 rows × 16 columns



Next steps:

[Generate code with data](#)

[View recommended plots](#)

[New interactive sheet](#)

data.columns



```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',  
      'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',  
      'minimum_nights', 'number_of_reviews', 'reviews_per_month',  
      'calculated_host_listings_count', 'availability_365',  
      'reviews_per_month_year'],  
      dtype='object')
```

- ✓ What all manipulations have you done and insights you found?

Answer Here.

4. Data Vizualization, Storytelling & Experimenting with charts : Understand the relationships between variables

✓ Chart - 1

Price Distribution using Histogram

```
# Chart - 1 visualization code

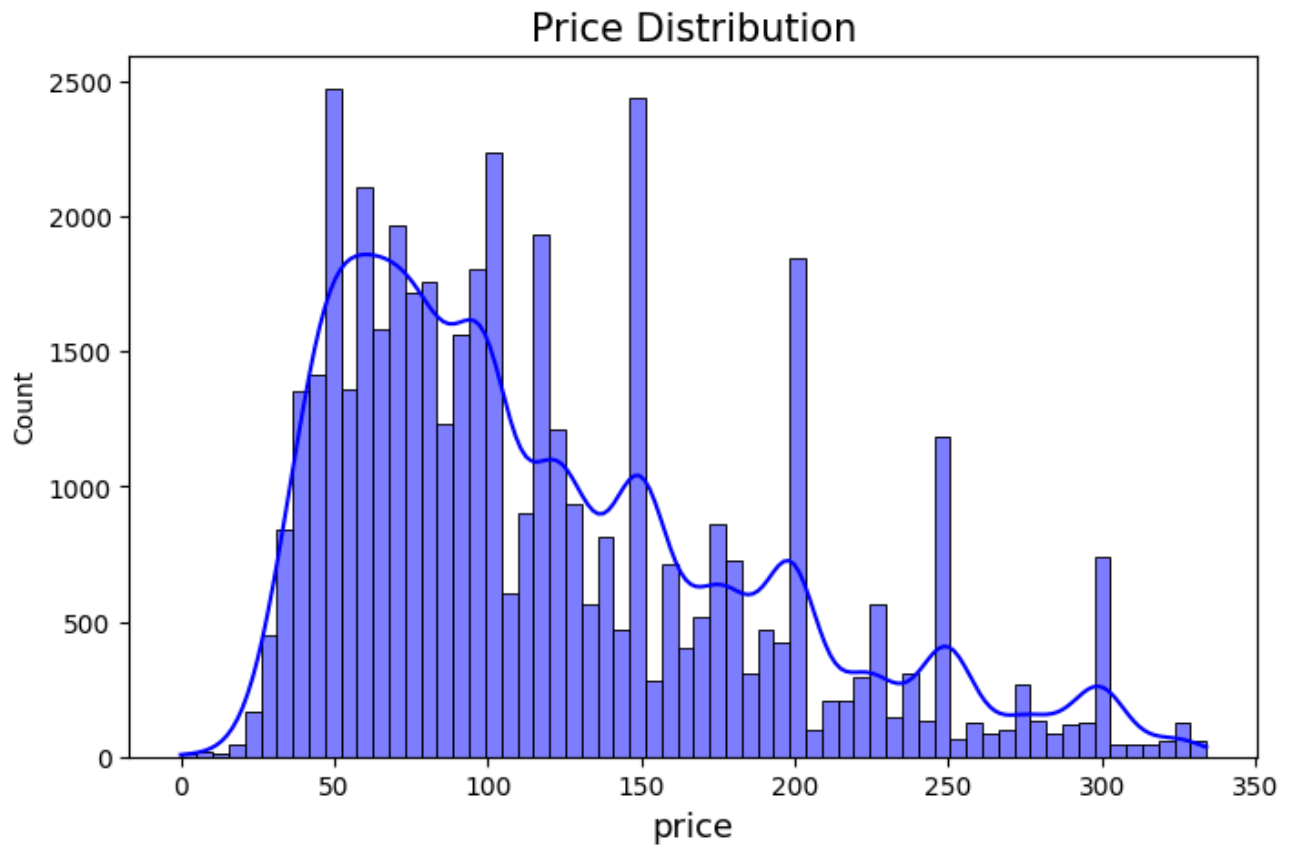
plt.figure(figsize=(8,5))

#Creating the plot using seaborn histplot function
sns.histplot(data=data,x='price',kde=True,color='blue')

#Adding title to the plot
plt.title("Price Distribution",fontsize=15)

#Customization of the plot
plt.xlabel("price",fontsize=13)

plt.show()
```

✓ 1. Why did you pick the specific chart?

The histogram was chosen because it effectively represents the distribution of continuous variables like price. It helps visualize how prices are spread across different ranges, while the KDE (Kernel Density Estimate) curve provides a smooth outline of the distribution, making patterns more apparent.

✓ 2. What is/are the insight(s) found from the chart?

The histogram shows the frequency of listings at different price points. If most values are clustered around lower price ranges, it indicates affordability. The presence of outliers, such as listings with significantly higher prices, can also be identified, hinting at luxury accommodations or anomalies in the dataset.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, gained insights will help creating a positive business impact. Understanding price distribution helps identify market trends and pricing strategies. For example, it can guide hosts in setting competitive prices or identifying opportunities for premium offerings. It also aids Airbnb in understanding market segmentation and targeting specific customer needs.

✓ Chart - 2

Total count of each room type using Pie Chart

Chart - 2 visualization code

```
#Calculating the number of different types of rooms inn Airbnb dataset
room_type_count_df=data['room_type'].value_counts().reset_index()
```

```
#Rename the columns
room_type_count_df.columns=['Room_type','Total_counts']
```

```
#Printing the dataframe
room_type_count_df
```

	Room_type	Total_counts	
0	Entire home/apt	22789	
1	Private room	21996	
2	Shared room	1138	

Next
steps:

[Generate code with room_type_count_df](#)
[View recommended plots](#)
[New interactive sh](#)

```
#Creating Pie Chart
```

```
#Set figure size
```

```
plt.figure(figsize=(10, 6))
```

```
#Set the labels
```

```
x=['Entire home/apt', 'Private room', 'Shared room']
```

```
y=[25409,22326,1160]
```

```
#Create pie chart
```

```
plt.pie(y,labels=x,autopct='%1.1f%%',colors=['red','green','yellow'])
```

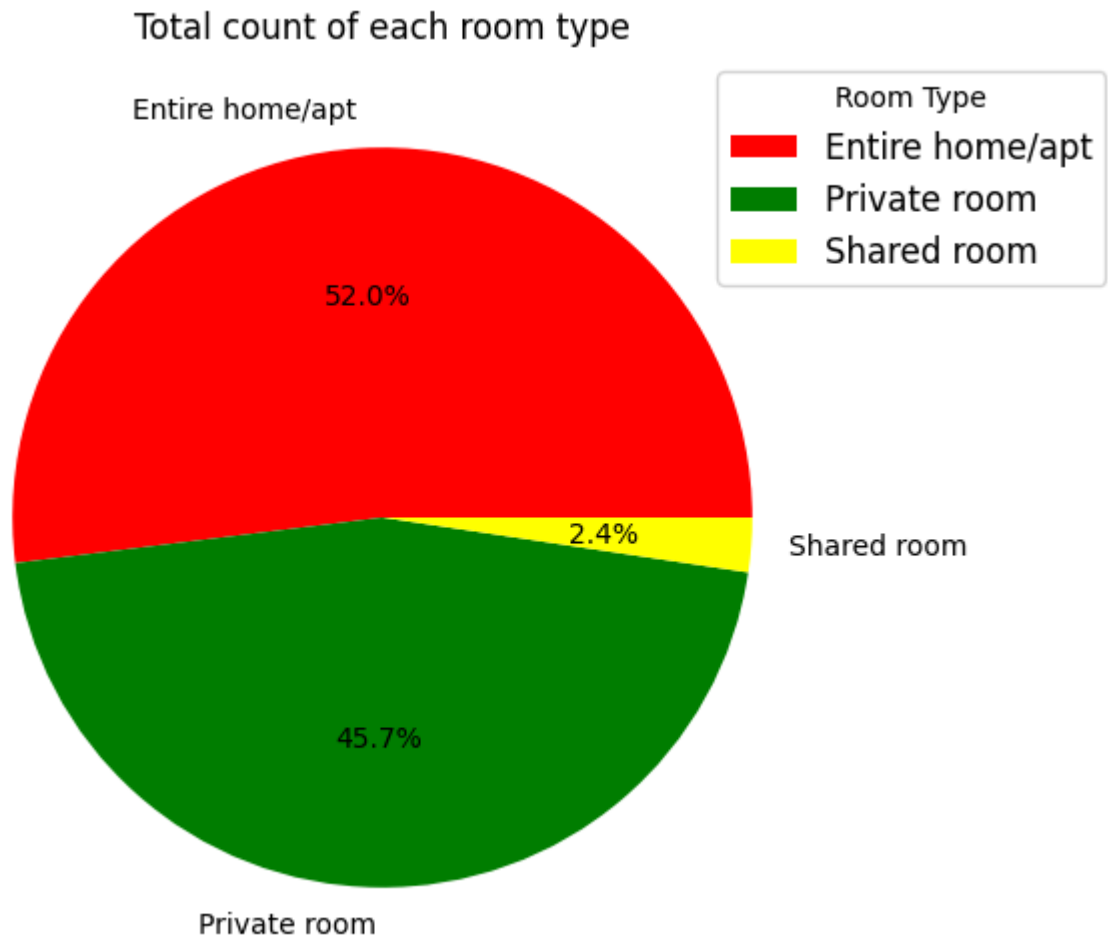
```
#Assign a title
```

```
plt.title('Total count of each room type')
```

```
# Add a legend to the chart
```

```
plt.legend(title='Room Type', bbox_to_anchor=(0.8, 0, 0.5, 1), fontsize='12')
```

```
#Show the chart  
plt.show()
```



✓ 1. Why did you pick the specific chart?

The pie chart was chosen because it effectively represents proportional data, showcasing the distribution of different room types in the Airbnb dataset. It provides a clear visual comparison of percentages, helping stakeholders easily identify the dominant room types.

✓ 2. What is/are the insight(s) found from the chart?

The chart reveals that "Entire home/apt" is the most common room type, constituting 52% of the listings. "Private room" follows at 45.7%, while "Shared room" accounts for a mere 2.4%. This indicates that the majority of users prefer accommodations offering more privacy.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, the insights can positively impact the business by guiding hosts to focus on listing "Entire home/apt" and "Private room" types, which are in higher demand. However, the low percentage of "Shared room" listings suggests limited interest, which could lead to negative growth if investments are heavily focused on this category. To avoid this, marketing and pricing strategies should target the most popular room types.

```
# This is formatted as code
```

✓ Chart - 3

No. of host per location using line chart

```
# Chart - 3 visualization code
```

```
# create a new DataFrame that displays the number of hosts in each neighborhood group
hosts_per_location=data.groupby('neighbourhood_group')['host_name'].count().reset_index()
```

```
#Rename the columns
```

```
hosts_per_location.columns=['Neighbourhood_group','Host_counts']
```

```
#Display the dataframe
```

```
hosts_per_location
```

	Neighbourhood_group	Host_counts	
0	Bronx	1069	
1	Brooklyn	19406	
2	Manhattan	19497	
3	Queens	5565	
4	Staten Island	365	

Next
steps:

[Generate code with hosts_per_location](#)
[View recommended plots](#)
[New interactive sh](#)

```
#Visual representation
```

```
#Set figure size
```

```
plt.figure(figsize=(10,5))
```

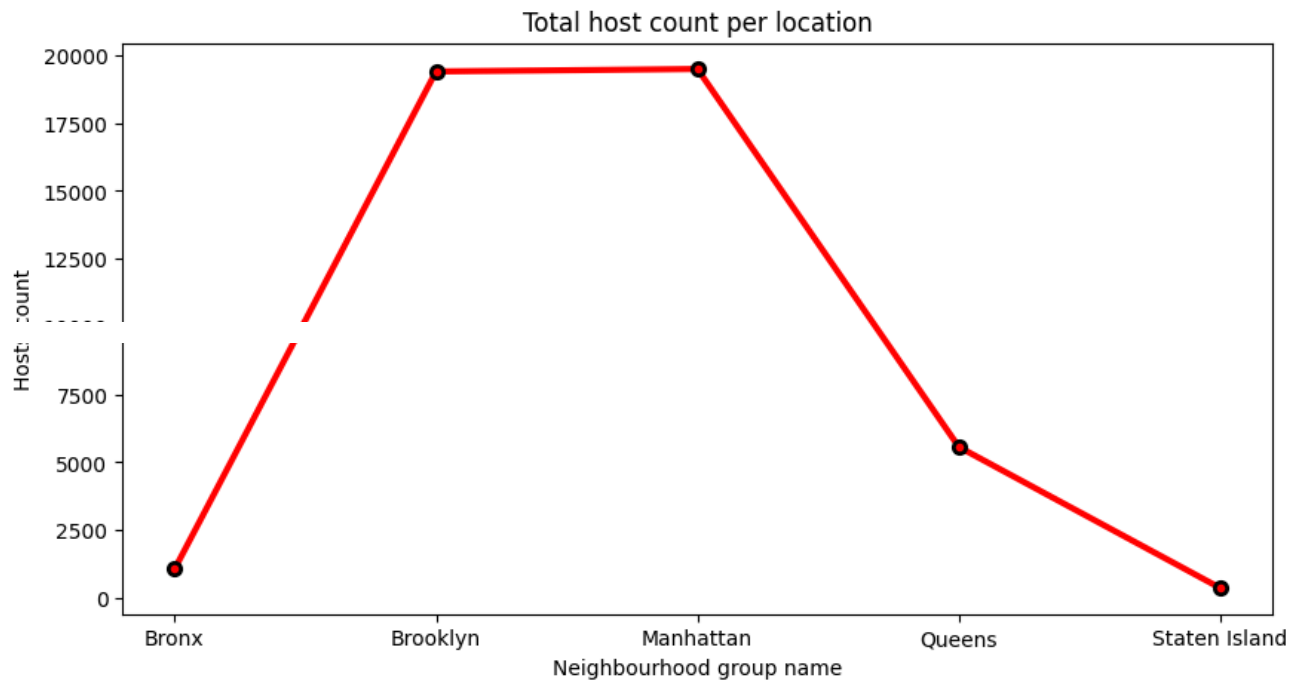
```
#Create line chart
```

```
plt.plot(hosts_per_location['Neighbourhood_group'],hosts_per_location['Host_counts'],color
```

```
#Customization
```

```
plt.title('Total host count per location')  
plt.xlabel('Neighbourhood group name')  
plt.ylabel('Hosts count')
```

```
#Show the plot  
plt.show()
```



✓ 1. Why did you pick the specific chart?

The line chart was selected as it effectively shows trends and comparisons across categories, making it easier to observe how the number of hosts varies across different neighborhood groups in a sequential and visually connected manner.

✓ 2. What is/are the insight(s) found from the chart?

The chart highlights that Manhattan and Brooklyn have the highest number of hosts among all neighborhood groups. This suggests these areas are the most active regions for Airbnb hosting, likely due to their popularity among tourists and travelers.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, the insights can positively impact the business by focusing on improving services in Manhattan and Brooklyn, where host activity is highest, ensuring continued growth. However, the lower host counts in other neighborhood groups might indicate underutilized potential or lower demand, which could limit revenue growth in those areas if not addressed with targeted marketing or incentives.

✓ Chart - 4

Top hosts with more listings using Bar chart

```
# Chart - 4 visualization code
```

```
# create a new DataFrame that displays the top 10 hosts in the Airbnb NYC dataset based o  
top_listers= data['host_name'].value_counts().reset_index()
```

```
#Rename the columns  
top_listers.columns=['Host_name','Total_listing']
```

```
#Display the dataframe  
top_listers.shape
```

```
↔ (11008, 2)
```

```
#Visual representation
```

```
# # Get the top 10 hosts by listing count  
top_hosts = data['host_name'].value_counts()[:10]
```

```
# #Set figure size  
plt.figure(figsize=(15,7))
```

```
# #Create bar chart  
sns.barplot(data=top_hosts,palette='hsv')
```

```
# #Customization  
plt.xlabel('top10_hosts', fontsize=13)  
plt.ylabel('total_NYC_listings', fontsize=13)  
plt.title('top 10 hosts on the basis of no of listings', fontsize=15)
```

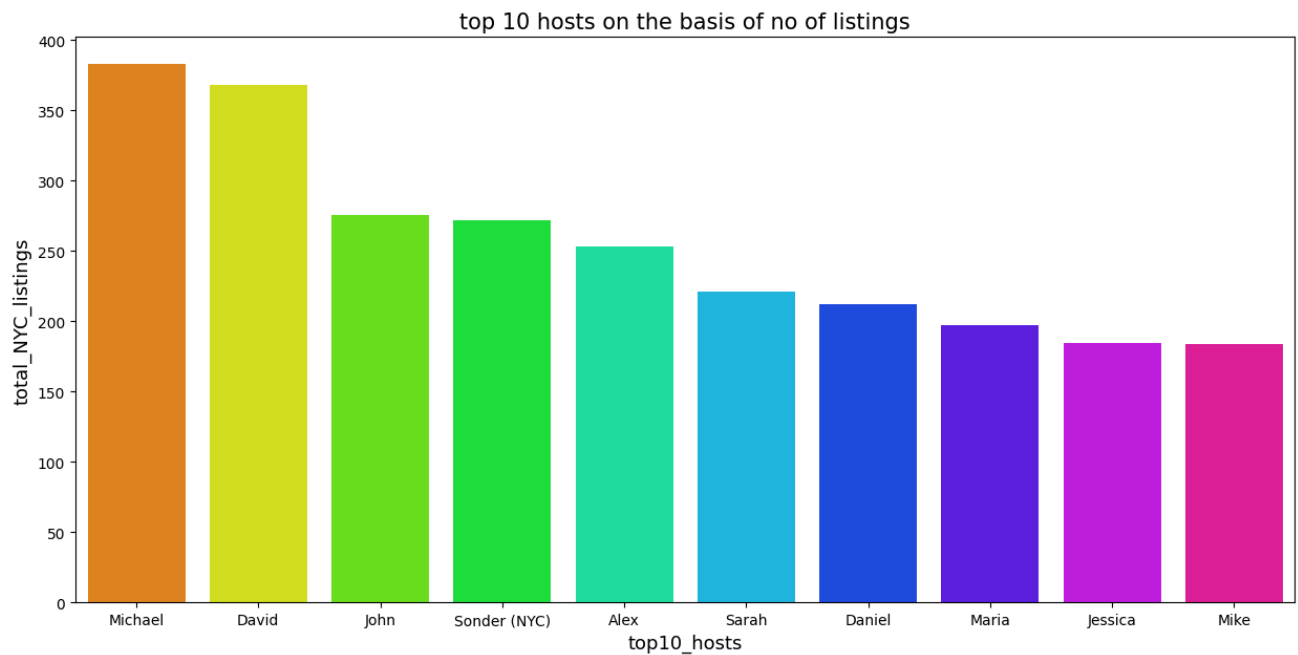
```
#Display  
plt.show()
```



<ipython-input-153-b758fbb63eab>:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.

```
sns.barplot(data=top_hosts,palette='hsv')
```



✓ 1. Why did you pick the specific chart?

The bar chart was chosen as it is well-suited for comparing discrete categories (hosts) and clearly highlights the top 10 hosts with the highest number of listings. It provides an easy-to-read visual representation of the distribution.

✓ 2. What is/are the insight(s) found from the chart?

The chart identifies the top 10 hosts with the most listings in the Airbnb NYC dataset. This indicates that a few hosts, like Michael David, dominate the market, managing multiple properties, which could be professional property managers or large-scale hosts.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights can positively impact the business by identifying key players who could be engaged for partnerships, promotional campaigns, or loyalty programs. However, reliance on a few top hosts might pose risks if their activity decreases or shifts to competitors, potentially leading to a disproportionate loss in listings

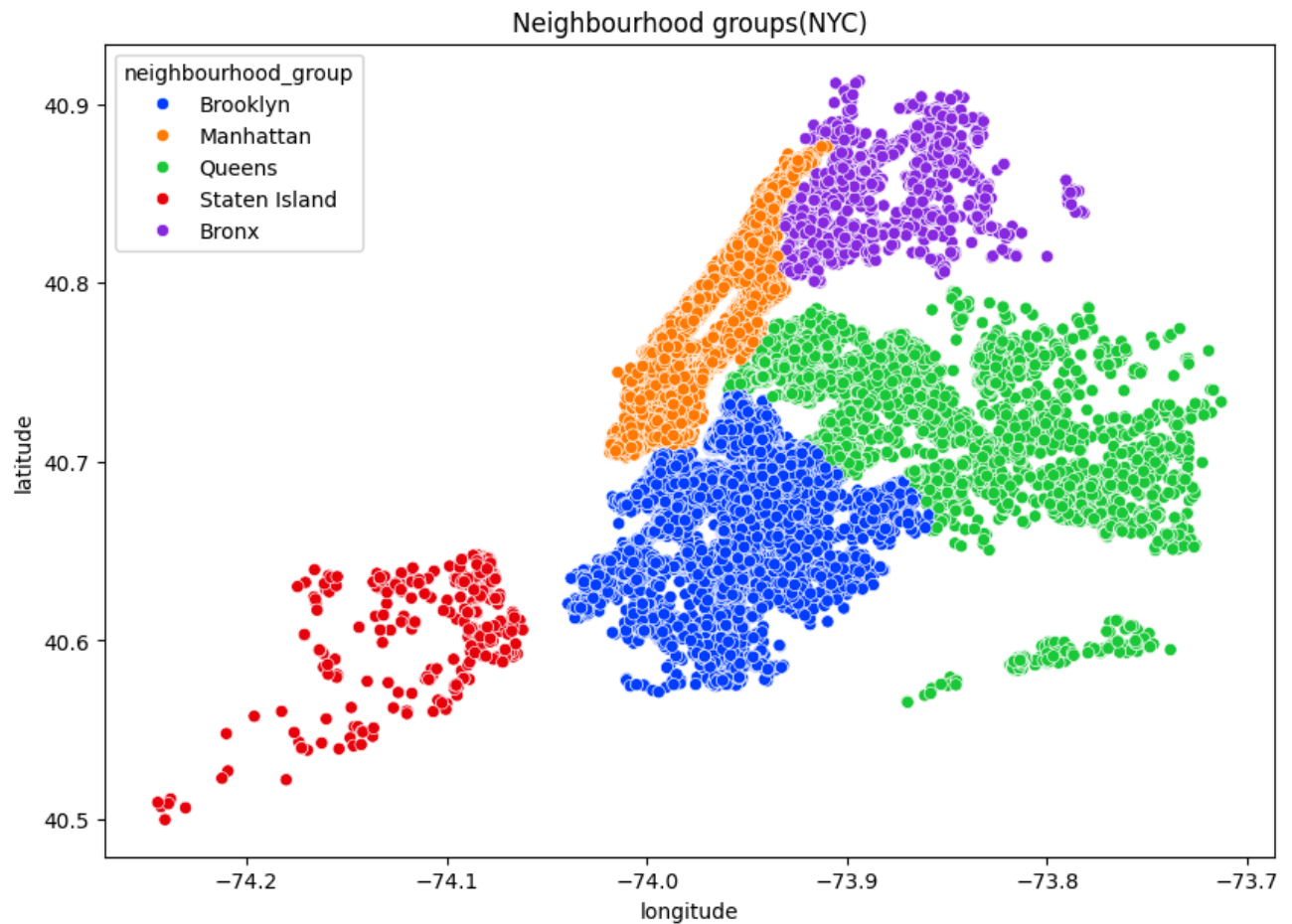
✓ Chart - 5

Using latitude and longitude in scatterplot map to find neighbourhood_groups

```
# Chart - 5 visualization code
#Set figure size
plt.figure(figsize=(10,7))

#Create scatter plot
sns.scatterplot(data=data,y='latitude',x='longitude',palette='bright',hue='neighbourhood_

#Customization
plt.title("Neighbourhood groups(NYC)")
plt.show()
```

✓ 1. Why did you pick the specific chart?

The scatterplot map was chosen because it allows for a clear geographic distribution of neighborhoods using latitude and longitude coordinates. It visually represents the location-based grouping of neighborhoods, while the color coding by neighborhood group makes distinctions easily identifiable.

✓ 2. What is/are the insight(s) found from the chart?

The chart reveals the geographical spread of different neighborhood groups in NYC. Certain neighborhood groups cluster in specific areas of the city, while others are more spread out, providing a spatial understanding of where various Airbnb listings are concentrated.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, the insights can help optimize business strategies by identifying high-density areas for targeted marketing or resource allocation. However, if too many listings are concentrated in a few neighborhoods, it could lead to market saturation, making it more competitive and potentially reducing individual host profitability. Expanding focus to less crowded areas could help avoid this.



New york actual map!

```
# This is formatted as code
```

✓ Chart - 6

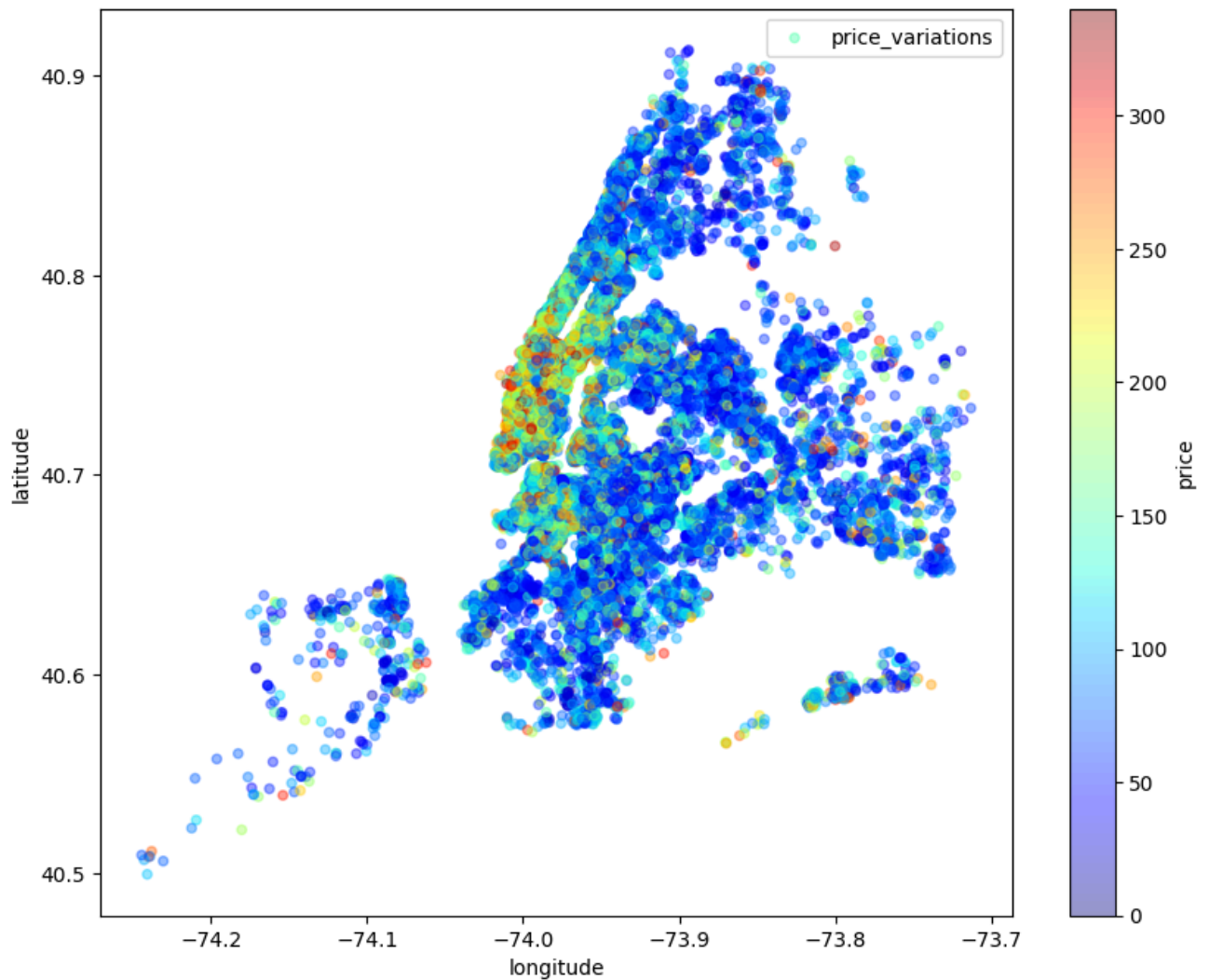
Price variations in NYC Neighbourhood groups using scatter plot *

```
#Find the max and min value of price to set the range in the plot
_min=data['price'].min()
_max=data['price'].max()
print(_min)
print(_max)
```

```
↔ 0
   334
```

```
# Chart - 6 visualization code
# create a scatter plot that displays the longitude and latitude of the listings in the A
lat_long = data.plot(kind='scatter', x='longitude', y='latitude', label='price_variations
                        cmap=plt.get_cmap('jet'), colorbar=True, alpha=0.4, figsize=(10, 8),vmi

# add a legend to the plot
```



✓ 1. Why did you pick the specific chart?

The scatter plot was chosen because it visualizes the geographic distribution of Airbnb listings while also representing the price variations through color coding. This allows for an intuitive understanding of both location and price differences across the city.

✓ 2. What is/are the insight(s) found from the chart?

The chart indicates that Manhattan and Brooklyn have higher-priced listings, as shown by the color gradient representing higher price points. This suggests that these neighborhoods are more expensive areas for Airbnb stays, which aligns with their popularity and centrality in NYC.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, understanding price variations can help businesses target marketing and pricing strategies effectively, focusing on high-demand, high-price areas like Manhattan and Brooklyn. However, excessive concentration on high-priced listings could limit market growth, especially if demand is price-sensitive. Expanding offerings in more affordable areas could attract a broader customer base, leading to a more sustainable business model.

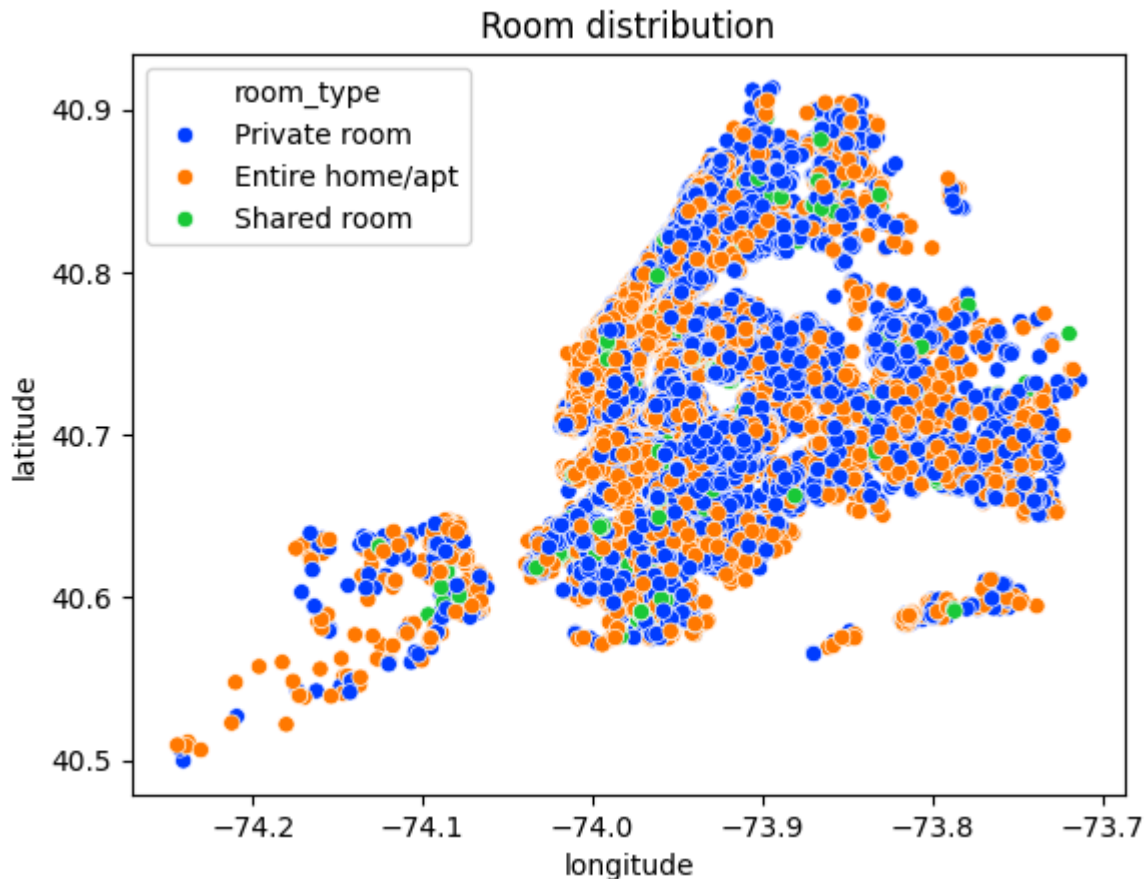
✓ Chart - 7

Room type analysis in different neighbourhood group using scatter plot

```
# Chart - 7 visualization code
sns.scatterplot(data=data,x='longitude',y='latitude',palette='bright',hue='room_type')

#Assigning title to the plot
plt.title("Room distribution")

#Display the plot
plt.show()
```



✓ 1. Why did you pick the specific chart?

The scatter plot was chosen because it allows for a visual distribution of room types across different neighborhoods, showing not only the location but also the variety of room types (private, entire home, etc.) in different parts of the city. Color coding by room type makes these distinctions easy to interpret.

✓ 2. What is/are the insight(s) found from the chart?

The chart reveals that in Manhattan, most listings are either "Private rooms" or "Entire homes/apt," with a preference for "Private rooms." In Brooklyn, all three room types—"Private room," "Entire home/apt," and "Shared room"—are well represented, indicating a diverse range of accommodation options. In Queens, "Private rooms" and "Entire homes/apt" dominate, with few listings for "Shared rooms." This suggests that different neighborhoods cater to different preferences in terms of room types.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights can help tailor business strategies to specific areas. For example, focusing on "Private rooms" in Manhattan and Queens may be a good strategy, while expanding "Entire home/apt" offerings in Brooklyn could meet diverse customer demands. However, if "Shared rooms" are less popular, businesses may want to reconsider focusing on that category in areas where it's not well-represented. Ignoring local preferences could hinder growth in certain neighborhoods.

✓ Chart - 8

Total Reviews by Each Neighborhood Group using Pie Chart

Chart - 8 visualization code

```
_grouped_reviews=data.groupby('neighbourhood_group')['number_of_reviews'].sum().reset_index()
_grouped_reviews.columns=['neighbourhood_group','total_count_of_reviews']
_grouped_reviews
```



	neighbourhood_group	total_count_of_reviews
0	Bronx	28185
1	Brooklyn	475936
2	Manhattan	428143
3	Queens	155719
4	Staten Island	11536

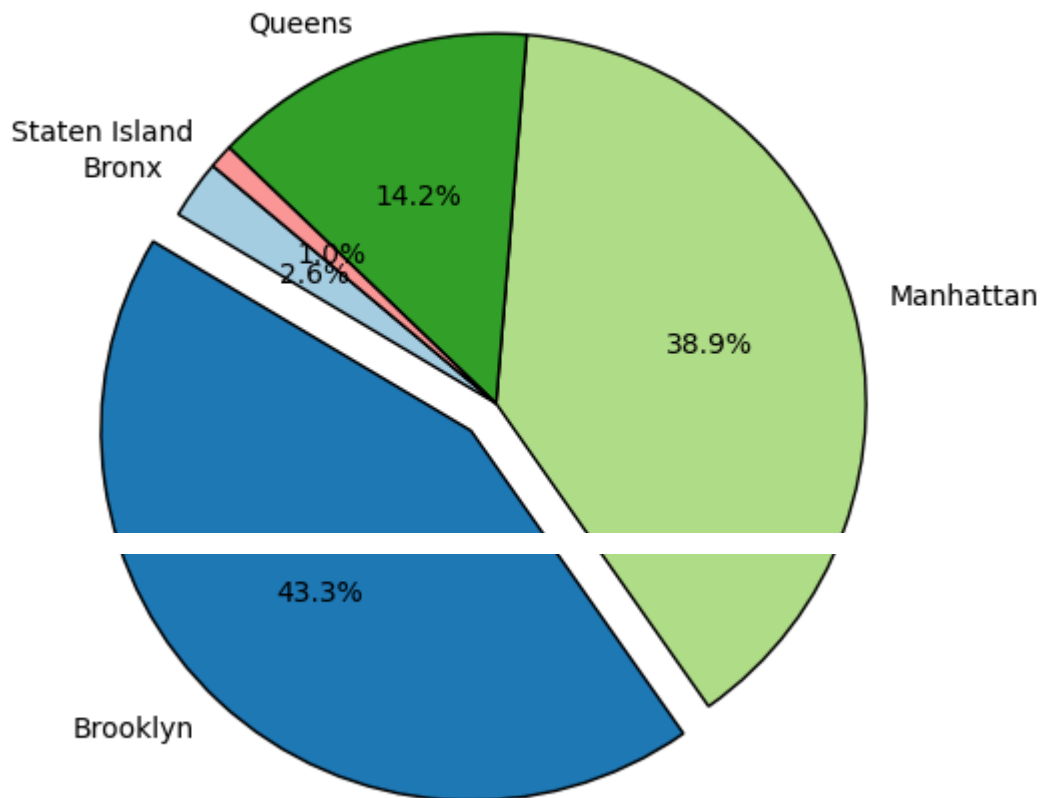


```
sizes = _grouped_reviews['total_count_of_reviews']
labels = _grouped_reviews['neighbourhood_group']
# Create a customized pie chart
plt.figure(figsize=(6, 6)) # Set the figure size
colors = plt.cm.Paired(range(len(labels))) # Use a colormap for distinct colors
explode = [0.1 if size == max(sizes) else 0 for size in sizes] # Highlight the largest slice
plt.pie(
    sizes,
    labels=labels,
    autopct='%1.1f%%', # Display percentages
    startangle=140, # Rotate to start at a specific angle
    colors=colors, # Assign colors
    explode=explode, # Highlight the largest slice
    wedgeprops={'edgecolor': 'black'} # Add a black border for better visibility
)
#Add a title to the chart
plt.title("Distribution of Reviews by Neighbourhood Group", fontsize=14, weight='bold')
```

```
# Display the pie chart  
plt.show()
```



Distribution of Reviews by Neighbourhood Group



✓ 1. Why did you pick the specific chart?

The pie chart was chosen because it effectively represents proportional data, showing the percentage of total reviews for each neighborhood group. Highlighting the group with the maximum reviews makes the comparison even clearer, allowing for quick and intuitive insights.

✓ 2. What is/are the insight(s) found from the chart?

The chart reveals that Brooklyn has the highest percentage of reviews, followed by Manhattan, Queens, Bronx, and Staten Island. This indicates that Brooklyn and Manhattan are the most active areas for Airbnb stays, reflecting their popularity among guests.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights can positively impact the business by focusing efforts on maintaining and improving services in Brooklyn and Manhattan, which dominate guest reviews. However, the significantly lower activity in Staten Island might indicate limited demand or awareness. Businesses should evaluate whether investing in Staten Island could attract more guests or if it's more strategic to concentrate resources in higher-performing areas.

✓ Chart - 9

Neighbourhood count using Point Plot

```
data.neighbourhood.nunique()
```

↗ 219

```
#Create dataframe for the count of neighbourhood
_neighbourhood_count=data['neighbourhood'].value_counts().reset_index()
```

```
#reset column name
_neighbourhood_count.columns=['neighbourhood','count']
```

```
#display
_ten_n_c=_neighbourhood_count.head(10)
_ten_n_c
```

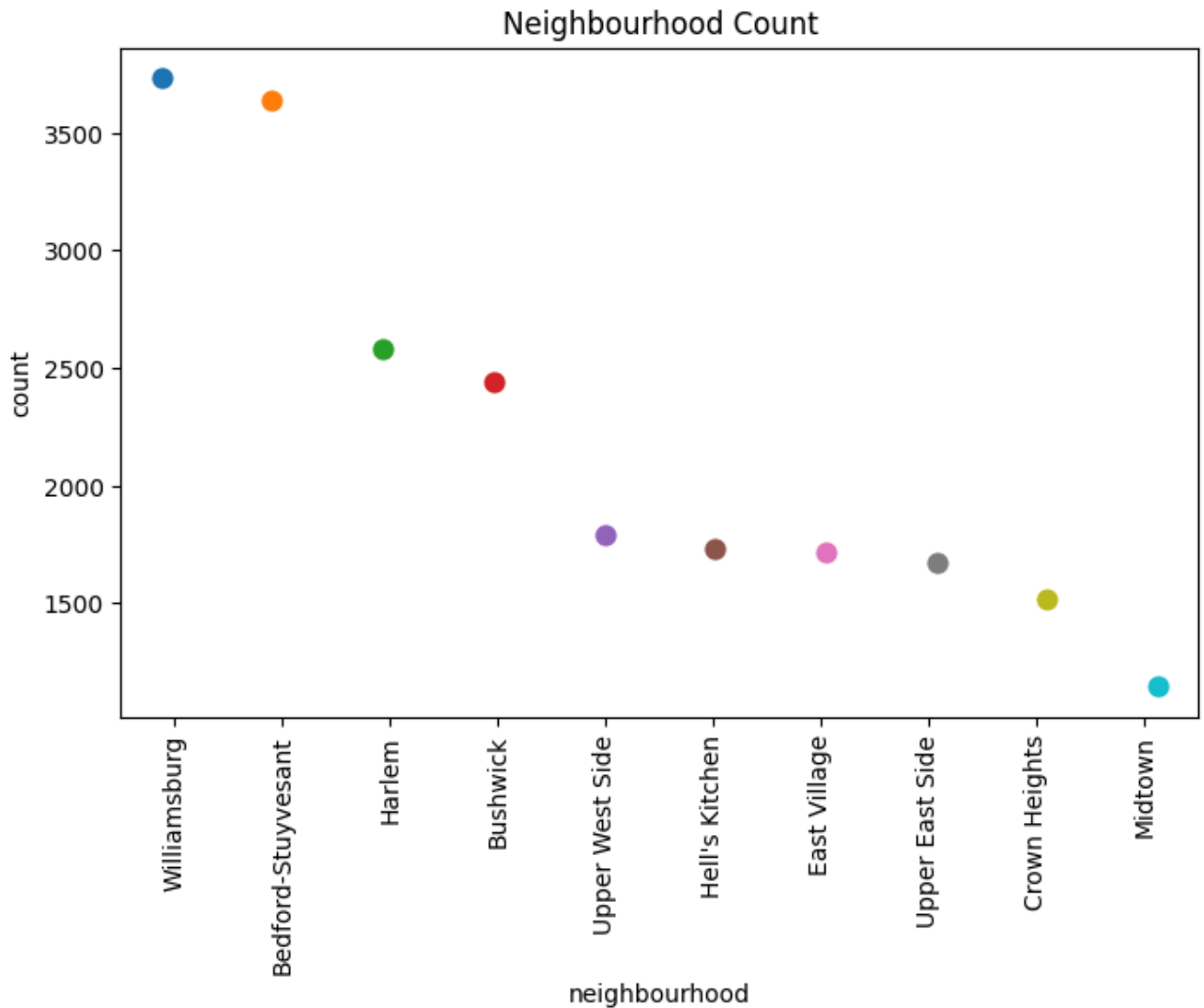
↗

	neighbourhood	count	
0	Williamsburg	3732	
1	Bedford-Stuyvesant	3638	
2	Harlem	2585	
3	Bushwick	2438	
4	Upper West Side	1788	
5	Hell's Kitchen	1731	
6	East Village	1714	
7	Upper East Side	1670	
8	Crown Heights	1519	
9	Midtown	1144	

```
# Create a point plot
plt.figure(figsize=(8,5))
sns.pointplot(
```

```
data=_ten_n_c,  
x='neighbourhood', # Column name  
y='count',         # Column name  
hue='neighbourhood', # Optional, if there's a grouping variable  
dodge=True         # Adjust points for clarity  
)
```

```
plt.title("Neighbourhood Count")  
plt.xticks(rotation=90) # Rotate labels if needed  
plt.show()
```



✓ 1. Why did you pick the specific chart?

The point plot was chosen because it provides a clear view of neighborhood counts, highlighting the most frequent neighborhoods. It allows for easy comparison of values while keeping the visualization simple and readable, especially when dealing with a large dataset.

✓ 2. What is/are the insight(s) found from the chart?

The chart shows that Williamsburg has the highest count of listings, followed by Bedford-Stuyvesant. This suggests that Williamsburg is a popular area for Airbnb listings, which could be due to its high demand or attractiveness to travelers. Other neighborhoods also show varying levels of activity.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, the insights can help guide marketing and operational strategies in Williamsburg, where there is a higher concentration of listings. By targeting popular areas for promotions and support, the business can maximize engagement. However, the high concentration in Williamsburg could lead to market saturation, increasing competition among hosts and possibly reducing profitability unless managed carefully. It may also be worthwhile to promote underrepresented neighborhoods to diversify the offering.

✓ Chart - 10

Availability Across Neighborhoods

```
# Filter data to remove extreme outliers for better visualization
# Calculate the total availability for each neighbourhood group
total_availability_by_group = data.groupby('neighbourhood_group')['availability_365'].sum

# Plot
plt.figure(figsize=(10, 6))
sns.barplot(data=total_availability_by_group, x='neighbourhood_group', y='availability_365')

# Customization
plt.title("Total Availability by Neighbourhood Group", fontsize=16)
plt.xlabel("Neighbourhood Group", fontsize=14)
plt.ylabel("Total Availability (Days)", fontsize=14)

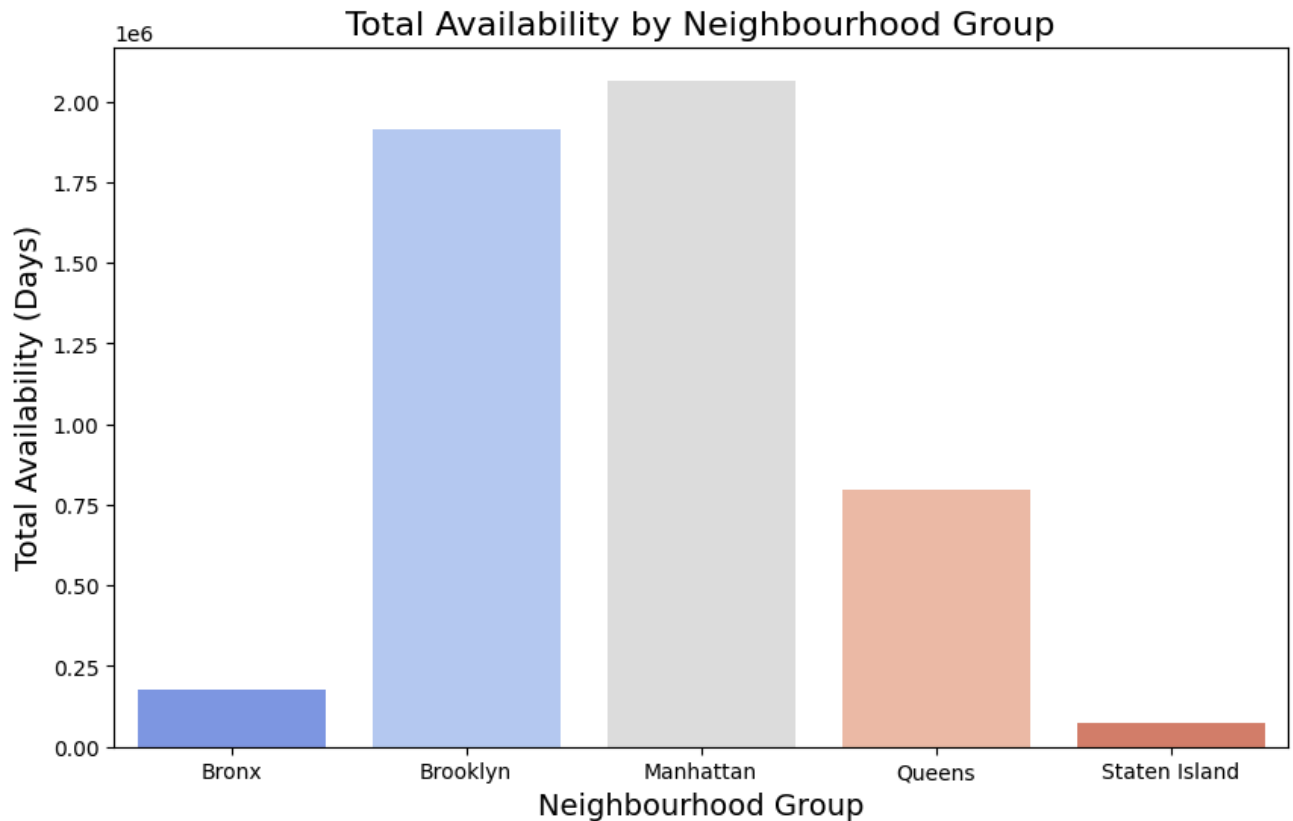
# Display
plt.show()
```



```
<ipython-input-163-9ba69921e57a>:7: FutureWarning:
```

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.

```
sns.barplot(data=total_availability_by_group, x='neighbourhood_group', y='availabil
```



✓ 1. Why did you pick the specific chart?

The bar chart was selected because it effectively compares the total availability of listings across different neighborhood groups. It provides a clear, quantitative view of availability, making it easy to identify the neighborhood with the highest values.

✓ 2. What is/are the insight(s) found from the chart?

The chart shows that Manhattan has the highest total availability of listings, followed by Brooklyn. This suggests that Manhattan has the most active and available listings for guests year-round, likely due to its high demand and popularity among travelers.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights can help businesses focus on maximizing operations in Manhattan, where availability is highest. This could include increasing marketing efforts or supporting hosts to maintain high availability. However, consistently high availability might indicate over-saturation, which could reduce individual host earnings in the long term. Diversifying efforts into other neighborhoods like Brooklyn or Queens could help balance demand and supply.

✓ **Chart - 11**

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

```
# Chart - 11 visualization code
#calculate total listing
_grouped_data=data['neighbourhood_group'].value_counts()

#set column name
_grouped_data.columns=['neighbourhood_group','total_listing']

#display
_grouped_data
```



	count
neighbourhood_group	
Manhattan	19506
Brooklyn	19415
Queens	5567
Bronx	1070
Staten Island	365

dtype: int64

✓ 1. Why did you pick the specific chart?

Answer Here.

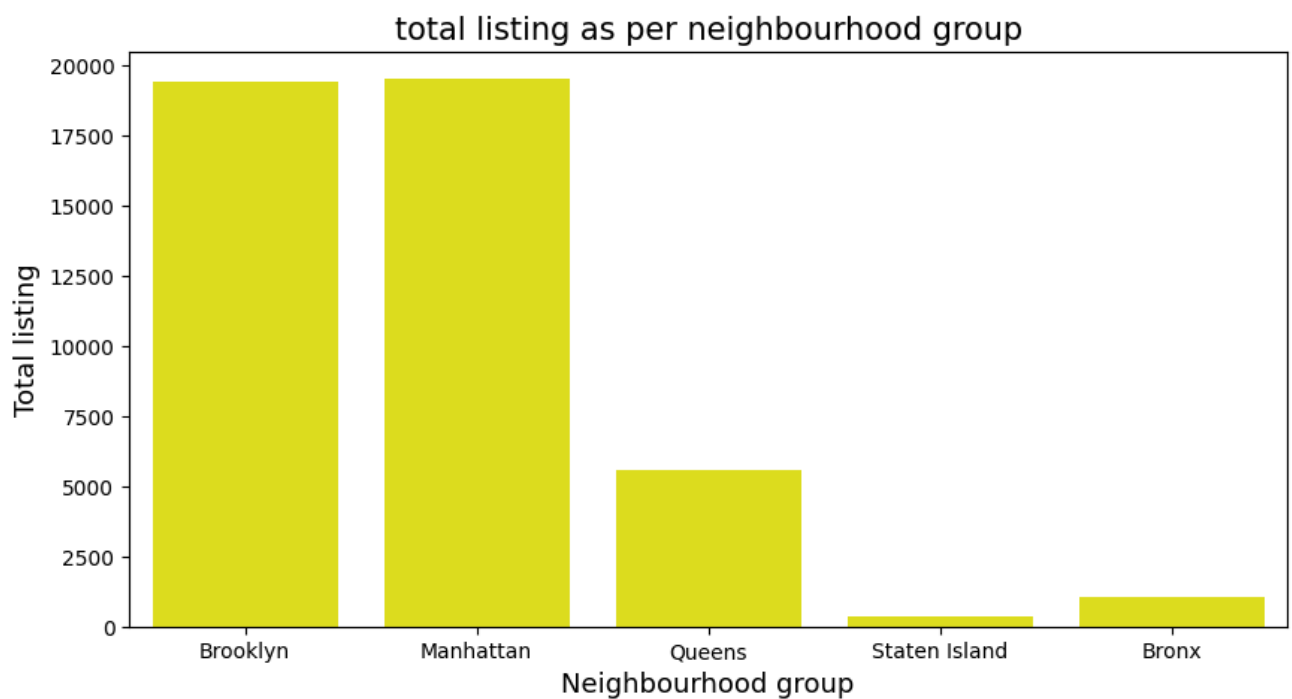
```
#Display the chart
plt.figure(figsize=(10,5))

#Plot
sns.countplot(x=data['neighbourhood_group'],color='yellow')

#Customization
plt.title("total listing as per neighbourhood group", fontsize=15) #adding title
plt.xlabel('Neighbourhood group',fontsize=13) # Set the x-axis label
plt.ylabel('Total listing',fontsize=13) #Set the y-axis label

# Set y-axis range
plt.ylim(300,20000)

#Display
plt.show()
```



> 2. What is/are the insight(s) found from the chart?

↳ 1 cell hidden

➤ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

✓ Chart - 12


Stay Requirement counts by Minimum Nights using Bar chart



```
# Group the DataFrame by the minimum_nights column and count the number of rows in each g
min_nights_count = data.groupby('minimum_nights').size().reset_index()
```

```
#rename
min_nights_count.columns=['minimum_nights','count']
```

```
# Sort the resulting DataFrame in descending order by the count column
min_nights_count = min_nights_count.sort_values('count', ascending=False)
```

```
#display
min_nights_count.head(10)
```



	minimum_nights	count	
0	1	12067	
1	2	11080	
2	3	7375	
29	30	3493	
3	4	3066	
4	5	2821	
6	7	1951	
5	6	679	
13	14	539	
9	10	462	

Next
steps:

[Generate code with min_nights_count](#)

[View recommended plots](#)

[New interactive shee](#)

```
#bar chart
#set figure size
plt.figure(figsize=(12,8))
```

```
#Chart function
```

```
plt.bar(min_nights_count['minimum_nights'],min_nights_count['count'], color='purple')
```

```
#Customization
```

```
plt.title("Stay Requirement by Minimum Nights", fontsize=15) #adding title
```

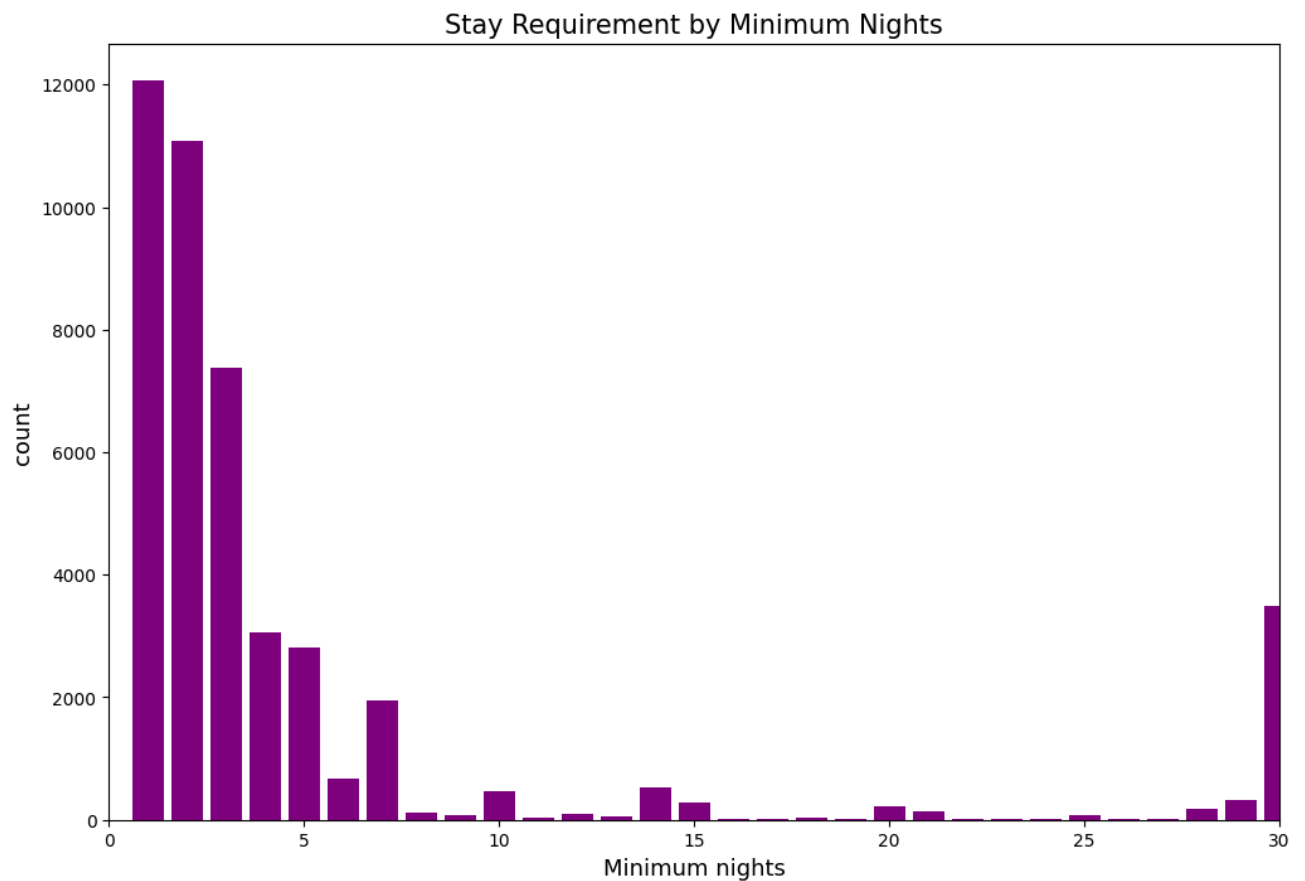
```
plt.xlabel('Minimum nights',fontsize=13) # Set the x-axis label
```

```
plt.ylabel('count',fontsize=13) #Set the y-axis label
```

```
plt.xlim(0,30)
```

```
#Display
```

```
plt.show()
```



✓ 1. Why did you pick the specific chart?

The bar chart was selected because it effectively visualizes the frequency distribution of stay requirements across minimum nights. It provides an intuitive way to observe the most common stay requirements and their frequency within a specified range.

✓ 2. What is/are the insight(s) found from the chart?

The chart reveals that shorter minimum night requirements, particularly those close to one or two nights, dominate the dataset. This suggests that most hosts are flexible and cater to short-term stays, which aligns with typical traveler preferences for short visits.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, the insights can help Airbnb and hosts optimize their offerings to match customer demand for short-term stays, potentially increasing bookings. However, if too many hosts adopt very short minimum stays, it may lead to operational inefficiencies, such as frequent turnover and cleaning costs, which could reduce profitability. Encouraging longer stays in certain neighborhoods may help balance this.

✓ Chart - 13

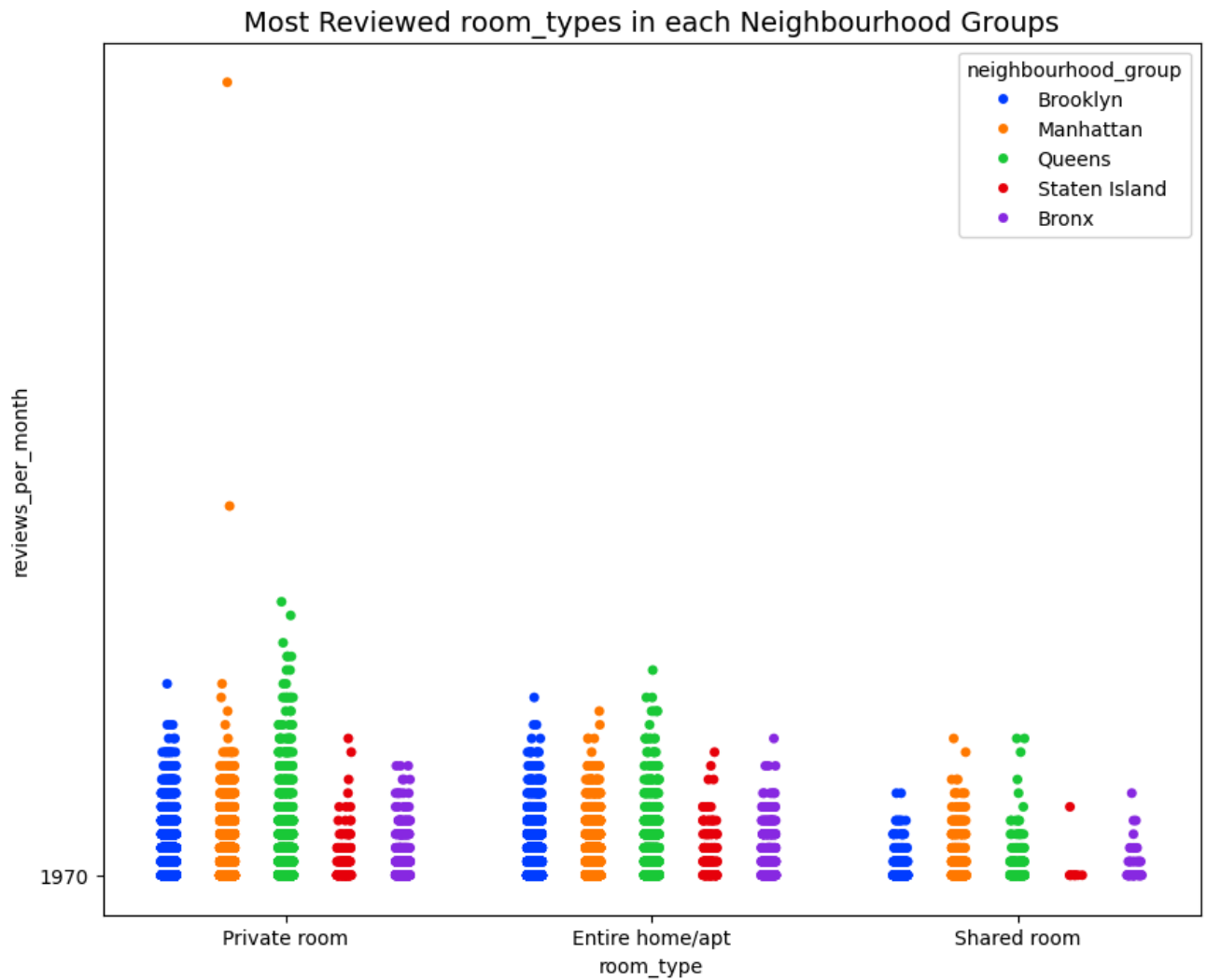
Most reviewed room type per month in neighbourhood groups using strip plot

```
# Chart - 13 visualization code
# create a figure with a default size of (10, 8)
f, ax = plt.subplots(figsize=(10, 8))

# create a stripplot that displays the number of reviews per month for each room type in
ax = sns.stripplot(x='room_type', y='reviews_per_month', hue='neighbourhood_group', dodge)

# set the title of the plot
ax.set_title('Most Reviewed room_types in each Neighbourhood Groups', fontsize='14')

#Display
plt.show()
```



✓ 1. Why did you pick the specific chart?

The strip plot was chosen because it effectively visualizes the distribution and frequency of reviews per month for each room type, differentiated by neighborhood groups. It allows for easy identification of trends and outliers within the dataset.

✓ 2. What is/are the insight(s) found from the chart?

The plot reveals that "Private rooms" and "Entire home/apt" receive the most reviews across all neighborhood groups, with Manhattan leading in reviews per month. This indicates higher guest

engagement in these categories and highlights neighborhood-specific preferences for room types.

✓ 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights can help focus efforts on popular room types like "Private rooms" and "Entire home/apt" in high-demand neighborhoods such as Manhattan. This could boost profitability and customer satisfaction. However, neglecting room types or neighborhoods with fewer reviews might miss out on untapped market segments, leading to uneven growth.

✓ Chart - 14 - Correlation Heatmap

```
# Correlation Heatmap visualization code
# calculate correlation between columns

num_coll=data.select_dtypes(include='number')
num_coll
_correlation=num_coll.corr()
_correlation
```



	id	host_id	latitude	longitude	price	minimum_nights
id	1.000000	0.581460	-0.008046	0.101337	-0.017973	0.032040
host_id	0.581460	1.000000	0.015976	0.144296	-0.034706	0.136545
latitude	-0.008046	0.015976	1.000000	0.091325	0.068841	0.012520
longitude	0.101337	0.144296	0.091325	1.000000	-0.307061	0.053874
price	-0.017973	-0.034706	0.068841	-0.307061	1.000000	0.027600
minimum_nights	-0.013732	-0.017921	0.025881	-0.064218	0.031454	1.000000
number of reviews	0.320440	0.136545	0.012520	0.053874	0.027600	0.027600