Project Name -

AirBnb Bookings Analysis

Project Type - EDA

Name Amit Singh

Project Summary -

Introduction Airbnb has transformed the travel industry ,offering of millions of listings worldwide.The project analyse the dataset of 49000 listings to extract a key insighs,focusing on user behaviour,pricing and host performance.

Objective The goal is to identify trends and patterns in Airbnb listings to inform business strategies,enhance user experience,and optimise pricing.

Data oberview The dataset includes both categorical(property type,neighbourhood) and numeric variables(price,reviews), offering a snapshot of airbnbs global presence.

Analysis Approach

1.**Data Cleaning:** Handle missing values,outliers,and standardize formats.

2.**Exploratory Data Analysis(EDA):** Examine statistics,distributions,and categorical frequencies.

3.**Trends and Patterns:** Analyse pricing ,host performance,and customer preferences.

GitHub Link -

https://github.com/amit-singh-tech

▯ Problem Statement

The goal is to analyse Airbnb dataset to uncover thr key patterns that inform strategic decisions. The focus areas are:

1.**Key Pricing Factors**: Idenbtify how property types , location, and amenities effect prices and optimise revenue.

2.**Host Perfomance**: Evaluates hostv ratings,response times, and listing to support or improve performance.

3.**Customer Preferences**: Analyse booking patterns and understand user satisfaction and property popularity.

The analysis will offer actionable recommentations to improve service,host performance and pricing strategies.

Define Your Business Objective

The project aims tob use AirBnb listing data to enhance decision making and operational efficiency by:

<1> **Optimise Pricing**: recommended data_driven pricing strategies based on property type,location, and amenities to maximise host revenue.

<2> **Improve Host Perfomance**: Provide insights to enhance host ratings , response times,and service quality.

<3> **Understanding Customer Preferences**: Analyse booking patterns to tailor offerings and marketing strategies.

<4> **Forcastinf Trends**: Build models to predict pricing and demands for strategic planning. The goal is

to boost AirBnb competitiveness ,revenue,user satisfaction and market growth.

▯ Let's Begin !

▯ 1. Know Your Data

▯ Import Libraries

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

Next steps: View recommended plots New interactive sheet

▯ Dataset Loading

```
from google.colab import drive drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
# Load Dataset data=pd.read_csv("/content/Airbnb
Dataset.csv") data
```

	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.
...	...	...	...	...	...	...	
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	Hell's Kitchen	40.
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	Hell's Kitchen	40.

48895 rows × 16 columns

⍰ Dataset First View

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

0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-7
1	2595	Skylit Midtown 2845 Jennifer Castle	Manhattan	Midtown	40.75362	-7		
		THE VILLAGE						

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	lon
2	3647	OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-7
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-7
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-7

Dataset Rows & Columns count

Next steps: View recommended plots New interactive sheet

```
# Dataset Rows & Columns count print('number of rows in the dataset are',data.shape[0]) print('number of columns in the dataset are',data.shape[1])
```

```
number of rows in the dataset are 48895 number
of columns in the dataset are 16
```

Dataset Information

```
# Dataset Info
print('dataset completer information',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 dataset completer information
None
```

? Duplicate Values

```
# Dataset Duplicate Value Count
print('no. of duplicates are',data.duplicated().sum())
```

```
no. of duplicates are 0
```

? Missing Values/Null Values

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

```
20141
```

```
# Visualizing the missing values
print('percentage wise missing value',round(data.isnull().sum()/len(data)*100))
```

```
percentage wise missing value id                                    0.0
 name                         0.0 host_id                         0.0
 host_name                                 0.0  neighbourhood_group
 0.0  neighbourhood                                    0.0  latitude
 0.0 longitude                  0.0 room_type                     0.0
 price                       0.0 minimum_nights                   0.0
 number_of_reviews            0.0 last_review                    21.0
 reviews_per_month              21.0 calculated_host_listings_count
 0.0 availability_365                   0.0 dtype: float64
```

? 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	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000

	id	host_id	latitude	longitude	price	minimum_nights	numb
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	

Double-click (or enter) to edit

? Check Unique Values for each variable.

```
# Check Unique Values for each variable. #unique
value for variable "name" data['name'].unique()
```

```
array(['Clean & quiet apt home by the park', 'Skylit Midtown Castle',
       'THE VILLAGE OF HARLEM....NEW YORK !', ...,
       'Sunny Studio at Historical Neighborhood',
       '43rd St. Time Square-cozy single bed',
       "Trendy duplex in the very heart of Hell's Kitchen"], dtype=object)
```

```
data['host_id'].unique()
```

```
array([      2787,       2845,       4632, ..., 274321313,  23492952,
 68119814])
```

```
data['host_name'].unique()
```

? 3. **Data Wrangling** ?

Data Wrangling Code

```
#we will  neglect the data where price=0 data=data[data['price']>0]
```

Therefore 11 data has been removed from dataset where price=0

```
# in order to fill missing values firstly we need to check
# weather the data followed a normal distribution or it is skewed
#select the column with missing values missing_values=
data[['last_review','reviews_per_month','name','host_name']] for i in
missing_values:   if data[i].dtype != 'object':       skewness = data[i].skew()
print(f'skewness of {i} is :{skewness:.2f}')   else:       print(f'skewness of
{i} is not applicable (non-numeric column)')
```

```
skewness of last_review is not applicable (non-numeric column)
 skewness of reviews_per_month is :3.13 skewness of name is
 not applicable (non-numeric column) skewness of host_name is
 not applicable (non-numeric column)
```

```
#imputing the numerical column  with skewed data-------------->median
#imputing the non numerical column --------->mode from
sklearn.impute import SimpleImputer impute_median =
SimpleImputer(strategy='median') impute_mode=
SimpleImputer(strategy='most_frequent')

data[['reviews_per_month']] = impute_median.fit_transform(data[['reviews_per_month']])
data[['last_review','name','host_name']]=impute_mode.fit_transform(data[['last_review','name','host_name']])


# changing last_review data type from object to date
data['last_review']=pd.to_datetime(data['last_review'])
```

What all manipulations have you done and insights you found?

Filtering out 0 in price column

<> Upon discovering the **price** column had a minimum value of **0**, which is not plausible for rental price.I applied a filter to remove these entries.The filter **df[df['price']>0]** was used to exclude records where the price was **0**, ensuring the dataset reflects the only valid active listings .

Imputation of Missing Values

For numerical column with skeqed distributions, such as <> For numerical column with skewed distributions, such as **reviews_per_month**, missing values are limputed using the median . The approach helps address skewness and provide a central measure of the data.

<> For categorical columns **(last_review,name,host_name)**, missing values were imputed using the mode.This strategy replaces the missing values with the most frequently occuring values in each column, ensuring a common value is used to fill gap.

Datatype conversion:

<> The **last_review** column , initially of type object , was converted to datetime . This conversion allows for more accurate data-based operations and analysis, such as time series analysis or date comparisons.

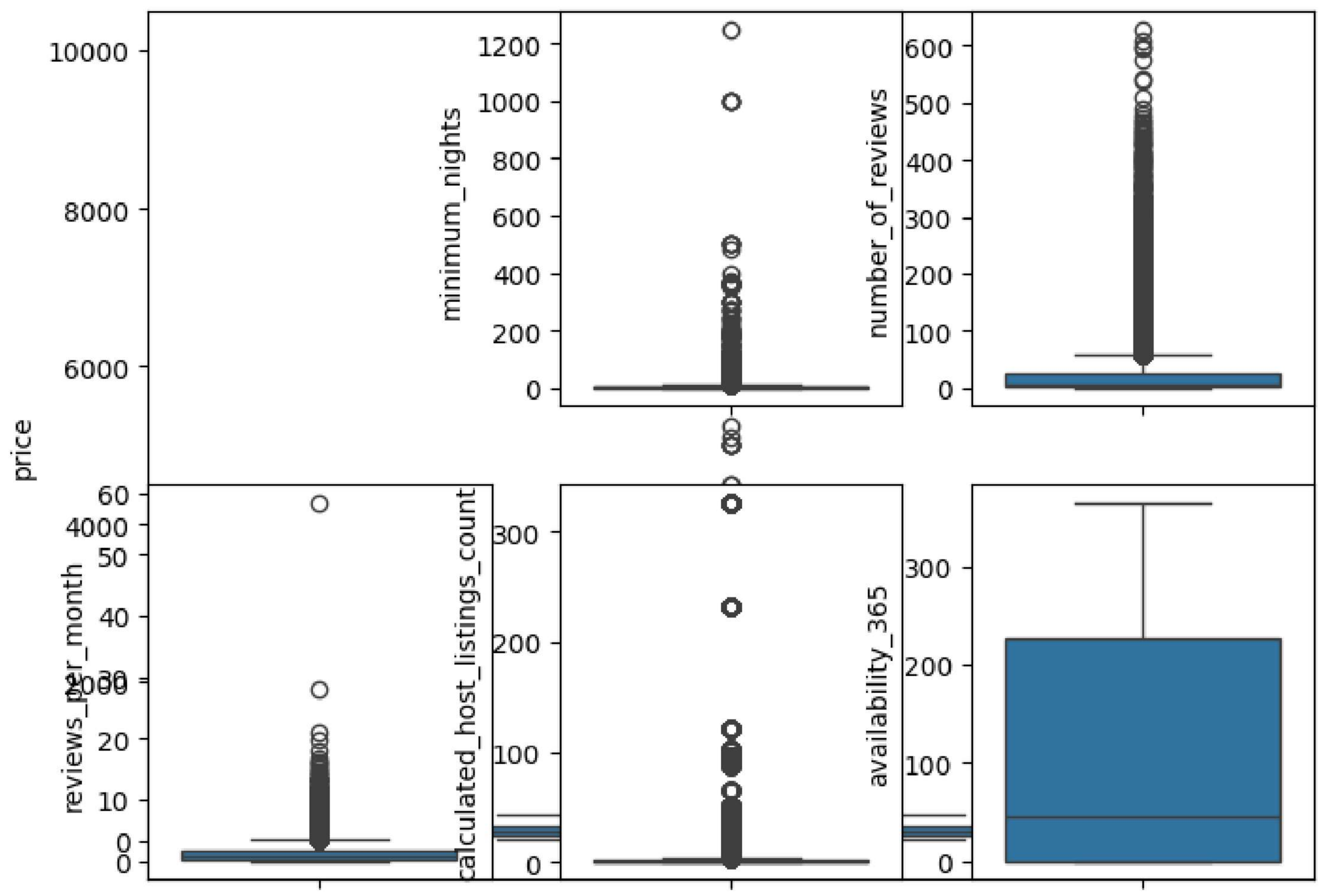
4. Data Vizualization, Storytelling & Experimenting with charts :

Understand the relationships between variables

Chart - 1

```
# Chart - 1 visualization code
f, ax =plt.subplots(figsize=(8,6))
sns.boxplot(data['price'])

plt.subplot(2,3,2)
sns.boxplot(data['minimum_nights'])

plt.subplot(2,3,3)
sns.boxplot(data['number_of_reviews'])

plt.subplot(2,3,4)
sns.boxplot(data['reviews_per_month'])

plt.subplot(2,3,5)
sns.boxplot(data['calculated_host_listings_count'])

plt.subplot(2,3,6)
sns.boxplot(data['availability_365'])
```

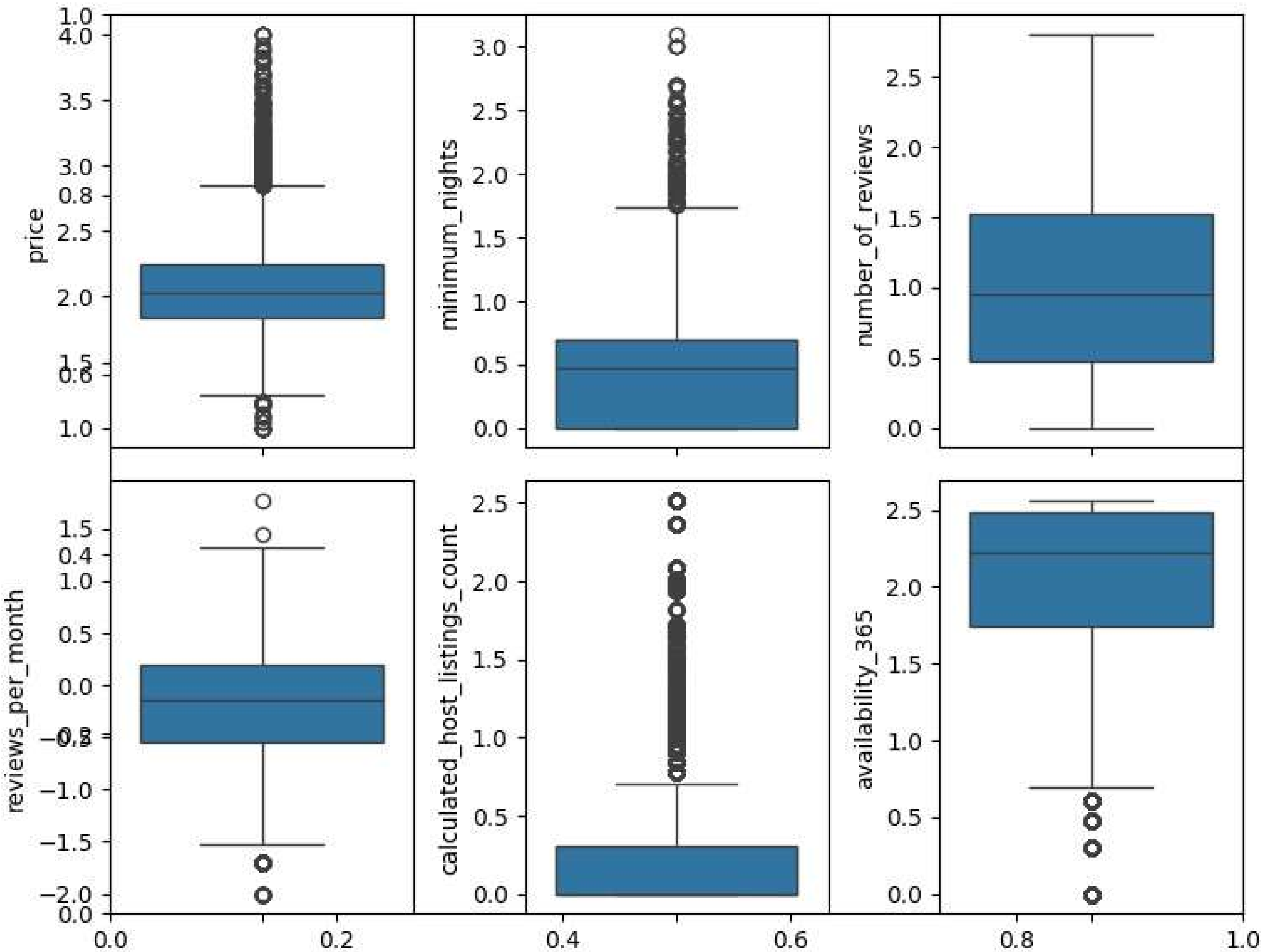
```
<Axes: ylabel='availability_365'>
```




```
# log transformations of variables
# chart 1.1 visualisation code
f , ax = plt.subplots(figsize =(8,6))

plt.subplot(2,3,1)
sns.boxplot(np.log10(data['price']))

plt.subplot(2,3,2)
sns.boxplot(np.log10(data['minimum_nights']))

plt.subplot(2,3,3)
sns.boxplot(np.log10(data['number_of_reviews']))

plt.subplot(2,3,4)
sns.boxplot(np.log10(data['reviews_per_month']))

plt.subplot(2,3,5)
sns.boxplot(np.log10(data['calculated_host_listings_count']))

plt.subplot(2,3,6)
sns.boxplot(np.log10(data['availability_365']))
```

```
plt.tight_layout() plt.show()
```

```
/usr/local/lib/python3.11/dist-packages/pandas/core/arraylike.py:399: RuntimeWarning: divide by
 result = getattr(ufunc, method)(*inputs, **kwargs)
 /usr/local/lib/python3.11/dist-packages/pandas/core/arraylike.py:399: RuntimeWarning: divide by
 result = getattr(ufunc, method)(*inputs, **kwargs)
```



? 1. Why did you pick the specific chart?

The initial box plot was created to visualize the distribution of the selected numeric variables (price,minimum_nights,number_of_reviews,reviews_per_month).Boxplot are ideal for identifying outliers and understanding the spread and central tendency of the data.

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

<1>. *Low reviews per month : * The reviews per month for each host are generally very low, including either a low engagement from guests or a potentially small number of bookings.

<2>. **Median Availabolity :** The median value of availability_365 , is around 50, suggesting that many properties are only available for 50 days a year. this could imply that a significant portion of host are not truly time renters.

<3>. **Price Outliers :** The price column contain many outliers , which could indicate a wide range of pricing strategies among hosts or thee presence of extremely high price listings that may distortthe overall data listings.

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

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

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

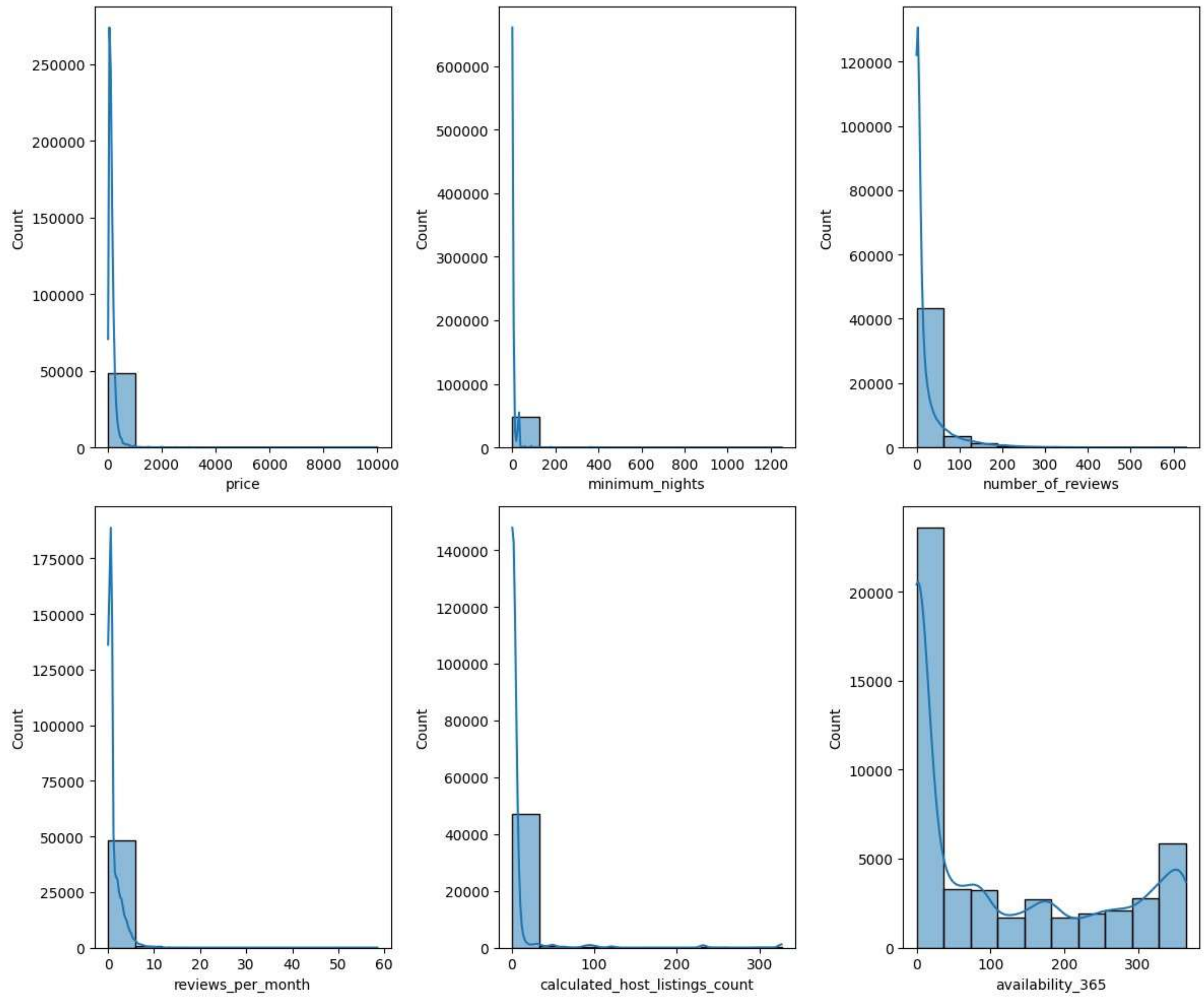
<1> **Low reviews per month: opportunity** : Encourage guest reviews through follow ups or incentives to boost engagement. **impact** : More reviews enhance creadibility , driving bookings and revenue growth.

<2> **Median Availibility**(50 Days) : **Opportunity** : Encourage hosts to increase availibility with targeted campaigns. **Impact** : More availibility leads to increased bookings and revenue.

<3> **Price Outliers** : **Opportunity** : Offer pricing tools to help hosts set competitive rates. **Impact** : Optimized pricing boosts occupancy and revenue for the hosts and airbnb.

Chart - 2

```
# Chart - 2 visualization code f, ax =
plt.subplots(figsize =(12,10), nrows=2, ncols=3)

plt.subplot(2,3,1) sns.histplot(data['price'], kde = True, bins
= 10, ax=ax[0,0])

plt.subplot(2,3,2)
sns.histplot(data['minimum_nights'], kde = True, bins = 10, ax=ax[0,1])

plt.subplot(2,3,3)
sns.histplot(data['number_of_reviews'], kde = True, bins = 10, ax=ax[0,2])

plt.subplot(2,3,4) sns.histplot(data['reviews_per_month'], kde =True, bins =
10 , ax=ax[1,0])

plt.subplot(2,3,5)
sns.histplot(data['calculated_host_listings_count'],kde=True,bins=10,ax=ax[1,1])

plt.subplot(2,3,6)
sns.histplot(data['availability_365'],kde=True,bins=10,ax=ax[1,2])

plt.tight_layout() plt.show()
```



1. Why did you pick the specific chart?

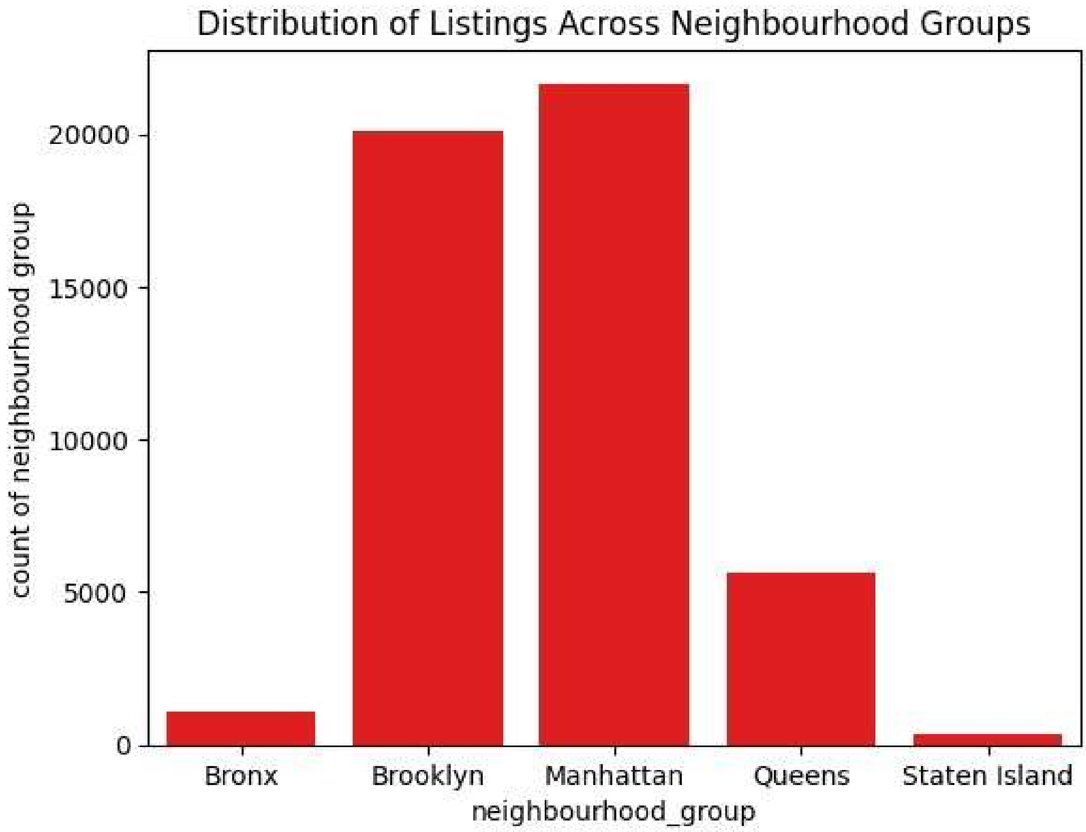
A histogram is used to visualize the distribution of a single numeric variable by showing the frequency of data points with specified bins.It is particularly useful for understanding the distribution of the data, including the shape (e.g normal distribution), central tendency and spread. In this case , the histogram helps in assessing how frequently different values occur or weather the data follow a normal distribution or exhibits skewness. The insights is valueable for making decisions about data transformation and understanding the underlying patterns in your dataset.

Chart - 3

```
a = data.groupby('neighbourhood_group').count().reset_index()
```

```
# Chart - 3 visualization code sns.barplot(x =
a['neighbourhood_group'], y=a['id'], color ='red')
plt.xlabel('neighbourhood_group') plt.ylabel('count of neighbourhood
```

```
group') plt.title('Distribution of Listings Across Neighbourhood
Groups') plt.show()
```



1. Why did you pick the specific chart?

The bar plot was specifically choosen to highlight the distribution of AirBnb listings across different neighbouring groups . By visualizing the number of listing in each group, we can quickly identify which neighbourhoods have a high concentration of listings and which one have fewer.This helps in understranding the popularity of saturation of listings in various areas , providing insights that are crucial for market analysis and decision making

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

The barplot reveals the Manhattan and brooklyn dominate the airbnb market with over 20000 listings each, making them most popular neighbourhood for hosts. In contrast queens has a moderate number of listings, with around 5500, while the Bronx and Staten island are the least popular, with approximately 1000 and 300 listings , respectively. These insight suggests that hosts and traveller alike favor certain neighbourhoods , with Manhatten and Brooklyn being the clear leaders in term of Airbnb presence.

Double-click (or enter) to edit

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

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

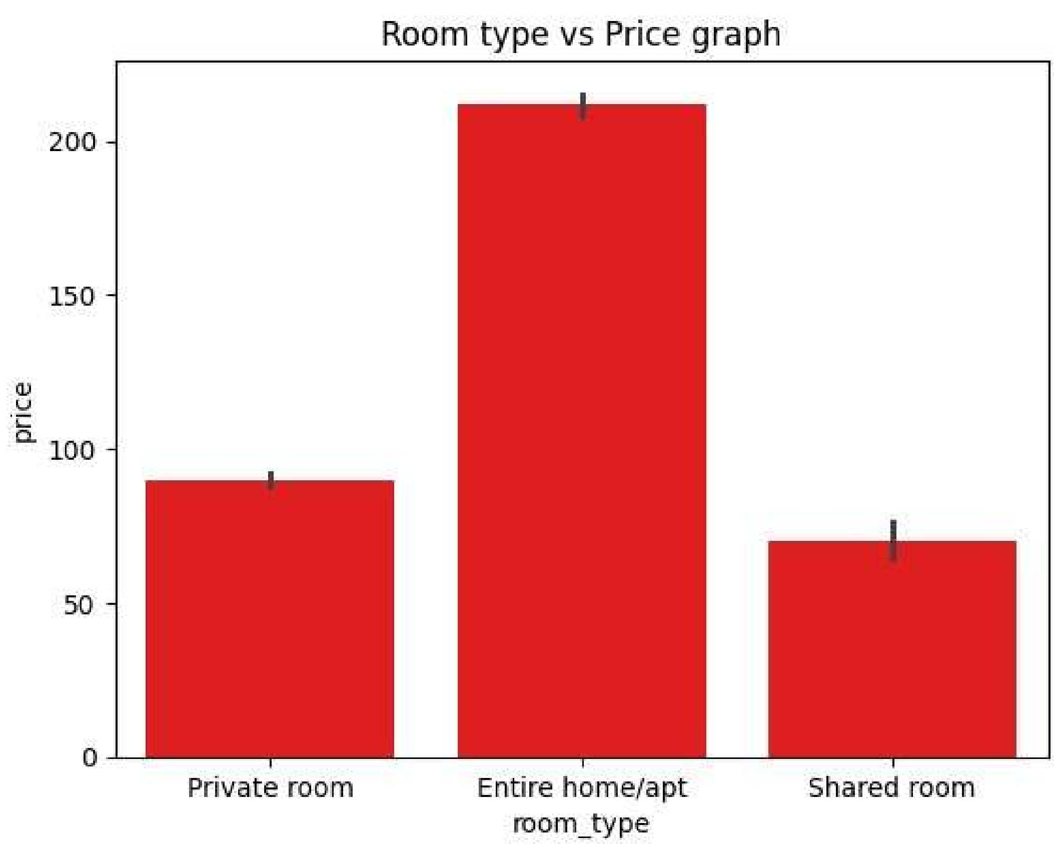
Positive business impact The gain insights can indeed contribute to a positive business impact . Understanding the Manhatten and Brooklyn have the higher number of airbnb listings can help property owners,hosts,and business maked informed decisions about where to invest and expand their operation .

By focussing in these high demand areas business can target a large market and potentially increase their revenue.Additionally making strategies can be tailored to attract more guesits to these popular neighbourhood, further boosting business opportunities.

Potential insights leading to negetive growth The insights also indicate that the Bronx and Staten island have significantly fewer listings compared to other neighbourhoods, with only about 1000 and 300 listings respectively. This could signal a lack of demand in these areas , potentially leading to a negetive growth if resources are invested here without propoer market analysis.The lower number of listings might due to factor such as lower tourist interest , less desirable locations or inadequate infrastructure . Investing in these areas without addressing these underlying issues could result in poor returns and business stagnation.

⍰ Chart - 4

```
# Chart - 4 visualization code sns.barplot(x=data['room_type'], y =
data['price'], color = 'red') plt.title('Room type vs Price graph')
plt.show()
```



1. Why did you pick the specific chart?

I choose the barplot between room type and price to effectively showcase the price ranges of different room types avai;lable on airbnb. This visualization allows for a clear comparison of how prices varies across various room categories, such as entire homes, private rooms and shared spaces .By using this chart, we can easily identify which room types command higher price and which ones are more budget friendly, providing valueable insights into pricing trends across different accomodations options.

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

The bar plot reveals that average pricing of private rooms and shared rooms hovers around Dollar 100, making them more budget friendly options for travellers .In contrast, the average price for an entire home or apartment, is significantly higher , at around dollar 200. This indicates that entire homes and apartment are priced at a premium compared to other room types, likely due to the added privacy and space they offer.

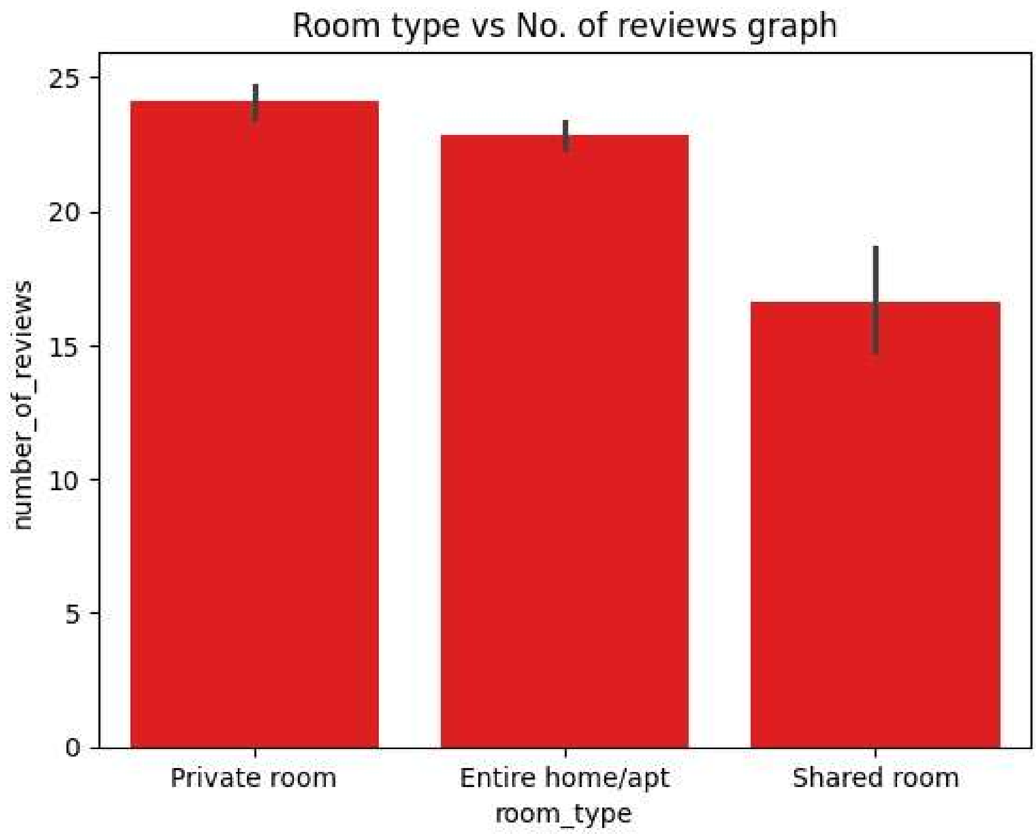
 3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason. The insights gained from the chart can lead to a positive business impact by informing pricing strategies and market positioning . Lnowing that private rooms and shared rooms are generally priced around dollar 100, business can target budget conciopus travellers by offering competitive rates or value added services within this price range. On mthe other hand recognizing,that entire homes or apartments are priced higher around dollar 200, allows hosts and property managers to caters to travellers seeking more privacy and space. By adjusting more pricing and marketing strategies accordingly,business can bettter needs of different customer segments , thereby increasing occupancy rates and profitability.

The insights gain from the chart can lead to a positive business impact by informing pricing strategies and markrt positioning. Knowing that private rooms and shared rooms are generally priced around dollar 100, business can target budget concious travelers by offering competitive rates and value added services within thios price range. On the other hand, recogniged that entire homes and apartment are priced higher around dollar 200, allows host and party managers to cater to traveller seeking more privacy and space. By adjusting pricing and marketing strategies accordingly, business can better meet the needs of different customers segments , thereby increasing the occupancy rate and profitability.

Chart - 5

```
# Chart - 5 visualization code sns.barplot(y =data['number_of_reviews'], x =
data['room_type'], color = 'red') plt.title('Room type vs No. of reviews graph')
plt.show()
```



1. Why did you pick the specific chart?
   A bar plot was choosen to compare the number of reviews across different room types . The type of chart is effective for visualizing categorical data,as it allows for straight forward comparison between distinct categories- in thus case the different room types. By displaying the number of reviews for each room types as bars, the chart clearly illustrates how review counts vary among the various types of accomodations . This comparison helps ius to understanding whicgh room types are more frequently reviewed, potentially reflecting their popularity or the level of guest engagement.

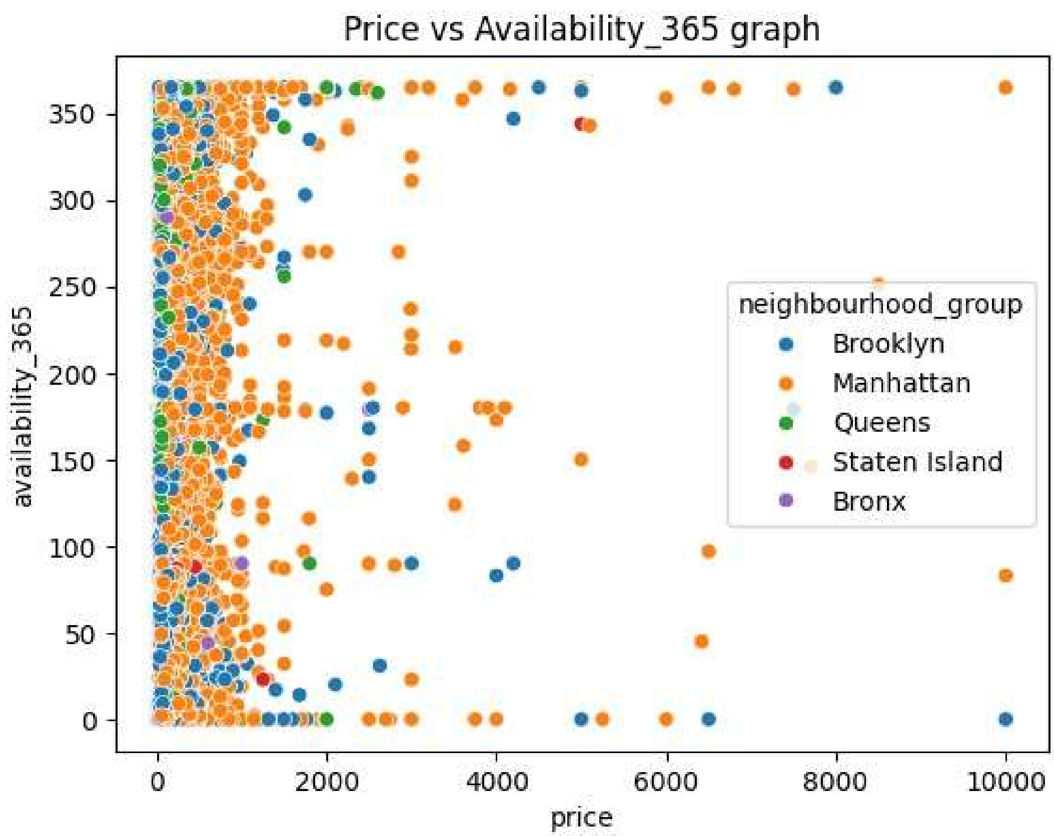
2. What is/are the insight(s) found from the chat? The bar plots indicate that : <> Private rooms : These have the highest number of reviews compared to the other room types .This suggest that the private room are the most popular of frequently looked type of accomodations, possibly due to their balance of cost and privacy.

<> Entire room/apartment : This room type follows with a specific number of reviews . The highest review counts for entire homes/apartments indicate that they are also popular, likely among guests seekinng more space and privacy for longer stays.

<>Shared rooms : This have the fewest reviuew among the three categories . The lower number of reviews could reflect the less popularity or a differentmarket segments, such as budget travellers who preferred shared accomodations.

⍰ Chart - 6

```
# Chart - 6 visualization code sns.scatterplot(x = data['price'], y = data['availability_365'], hue =
data['neighbourhood_group']) plt.title('Price vs Availability_365 graph')
plt.show()
```



⍰ 1. Why did you pick the specific chart?

Ascatteer plot is choosen to analyse the relation between price and availability_365 for listings .Scatter plot are particularly effective for identifying potential correlations between two numerical variables.

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

No significant corelations exists between botyh variables. Aditionally the scatter plot reveals that Manhatten has the highest number of available days.

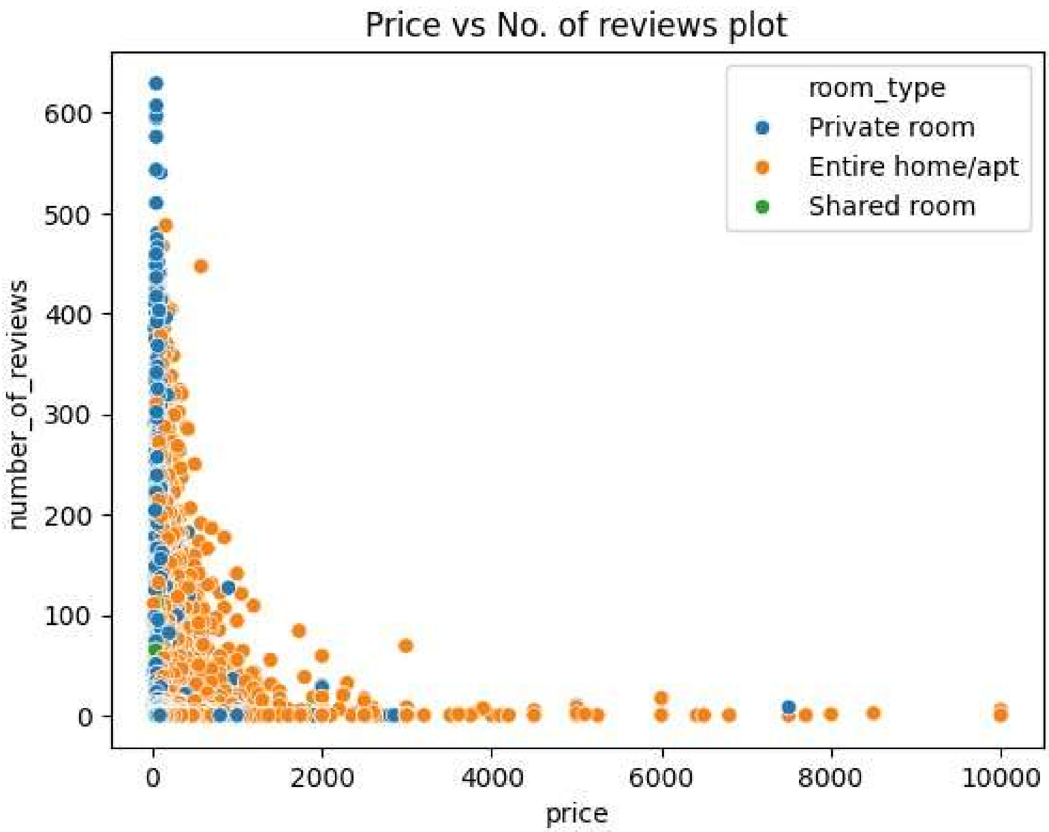
3. Will the gained insights help creating a positive business impact?

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

With manhatten having the highest number of available days, business should condider focussing marketing eddorts or adjusting strategies for this area.

? Chart - 7

```
# Chart - 7 visualization code sns.scatterplot(x = data['price'], y =
data['number_of_reviews'], hue = data['room_type']) plt.title('Price vs No. of reviews plot')
plt.show()
```



1. Why did you pick the specific chart?

A scatter plot was choosen to analyse the relationship between price and number_of_reviews.This chart is ideal for examining how changes in price might correlate with the number of reviews a listings receives.

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

The scatter plot shows no clear correlations between the price of listings and the number_of_reviews it receives.Entire home apartment apper to receive the largest number of reviews. This indicate that the guests are more inclined to book and review entire homes or apartments.

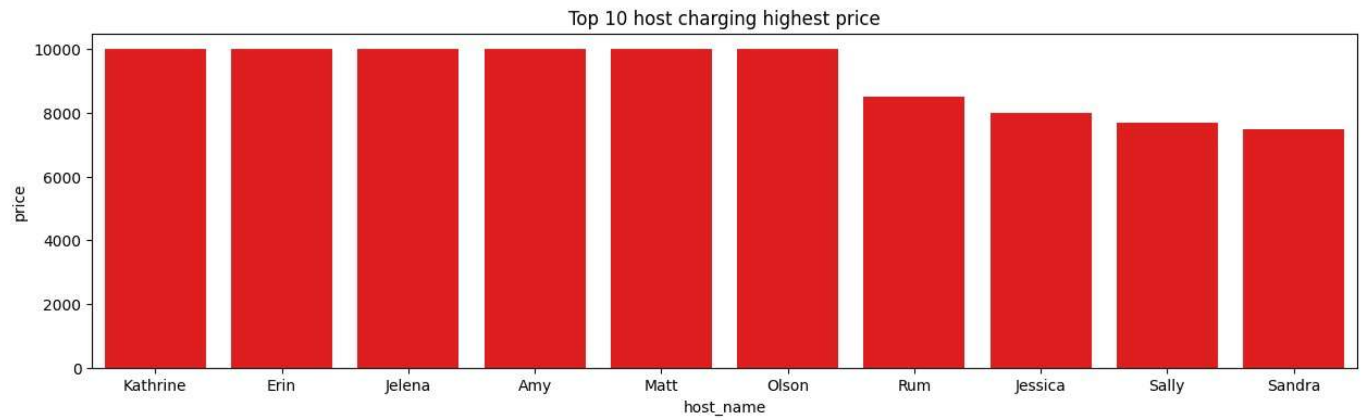
3. Will the gained insights help creating a positive business impact?

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

The insight that entire home and apartments recieves the largest number of review suggests a strong preference among guests for this type of accomodations. This can inform the business strategies to increase the availability or improve the quality of nentire ,house/apartments listings. By catering this demand business can enhance guests satisfactions and potentially increase bookings.

? Chart - 8

```
# Chart - 8 visualization code b= data.sort_values('price', ascending =
False)[['host_name','price']].reset_index().head(10) plt.figure(figsize=(15,4)) sns.barplot(x =
b['host_name'], y = b['price'], color = 'red') plt.title('Top 10 host charging highest price')
plt.show()
```



1. Why did you pick the specific chart?

A bar plot was choosen to visualize the top 10 Data charging the highest price.Parplot are particularly effective for comparing the prices charged by different hosts because they provide a clear and straightforward way to rank and display the relative values across categories.

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

Top high priced hosts : The chart reveals the host such as Jelena,Erin,Kathrine,Amy and Matt Olson charge in the highest price range, approximately around $10000.This indicates that these host are positioned at the premium end of the market potentially offering high end of luxury accomodations .

Secondary high priced group : Hosts like Rum,Jessica,Sally and Jack charges in the $8000 range . While slightly lower than the top group, these hosts are command high prices,suggesting that they are positioned at high value options within the market.

Price range segmentation : The distinct price ranges help categorize the hosts based on their pricing strategies . The clear sepration between the top group(10000 range) and the secondary group ($8000 range) suggests different tier of high end offerings.

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

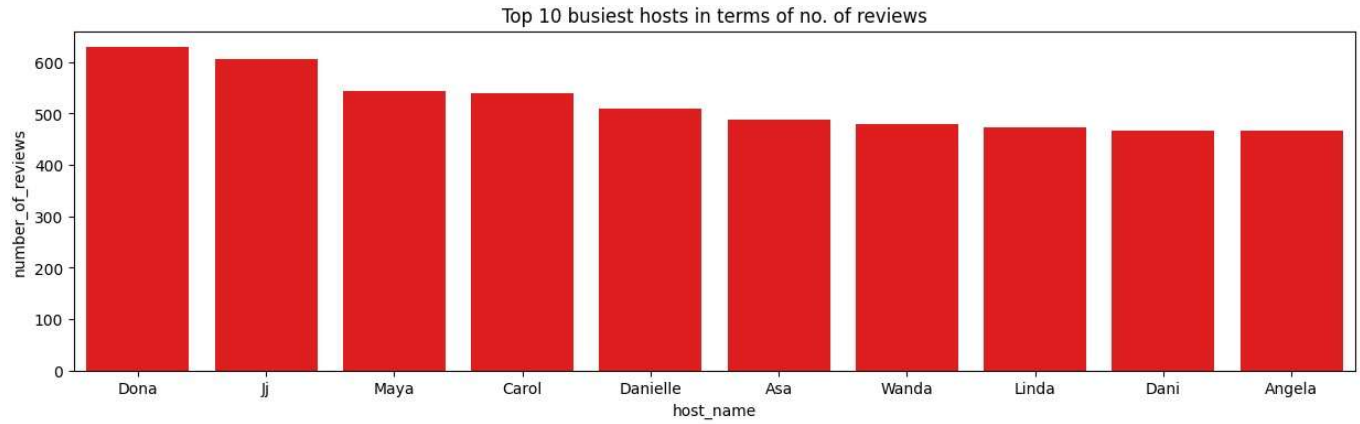
Are there any insights that lead to negative growth? Justify with specific reason.

The insight reveal that the hosts are positioned at the top end of the price range. This can help the other business and hosts to understanmd the marketfor high end accomodations and potentially develop strategies to compete or differentiate themselves in the premium segments.

Chart - 9

```
# Chart - 9 visualization code c =
data.groupby(['host_id','host_name'])['number_of_reviews'].max().reset_index() c =
c.sort_values('number_of_reviews',ascending = False).head(10)
plt.figure(figsize=(15,4)) sns.barplot(x = c['host_name'], y =
c['number_of_reviews'], color = 'red') plt.title('Top 10 busiest hosts in terms of
no. of reviews') plt.show()
```

1. Why did you pick the specific chart?



A bar plot was choosen to visualize the top 10 busiests hosts in terms of number of reviews .Bar plot are particularly effective for ranking and comparing categories , making them ideal for highlighting the hosts who have received the most guest reviews.

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

The barplot reveals that the Dona is the busiest host, with the highest number of reviews.This suggest donas listings are very popular among guests, possible due to factors like exceptional service ,desirable locations, or competitive pricings .Angela completes the list of the top 10 busiest hosts. While still within the top ranks.

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

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

The insights from the bar plot identifying the top 10 busiest hosts can help business initiate targeted loyalty programs. By encouraging repeat bookings with these high reviewed hosts,business can strengthen customers relationships, increase guests satisfactions, and boosts overall bookings rates. For instance,offering discounts or special perks for returning guests could drive more bookings for these popular hosts , enhancing their revenue and guests engagement.

? Chart - 10

```
# Chart - 11 visualization code e =
data.groupby('neighbourhood_group')['price'].mean().reset_index() e =
e.sort_values('price', ascending = False) sns.barplot(x =e['neighbourhood_group'], y
= e['price'], color = 'red') plt.title('Average price by Neighbourhood Group')
plt.show()
```

? 1. Why did you pick the specific chart?

A bar plot was choosen to compare the average price of listings across different neighbourhood_group categories . Bar plots are particularly effective for illustriating the differences in categorical data, makingb them ideal for viusualizing how average price varies across different neighbourhood.

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

<> The barplpot reveals that Manhatten has the highest average price, around $200.This indicates that the listings in manhatten are generally more expensive compared to other neighbourhood, which can be attributed to the area in high demand,prime locations, and premium amenities.

<> The average price in Brooklyn,Staten island,Queensand the Bronx are relatively similar and fall within the lower price range compared to Manhatten. This suggests that these neighbourhood are more affordable for guests,which might be due to lower demand, different property type and varing local market conditions.

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

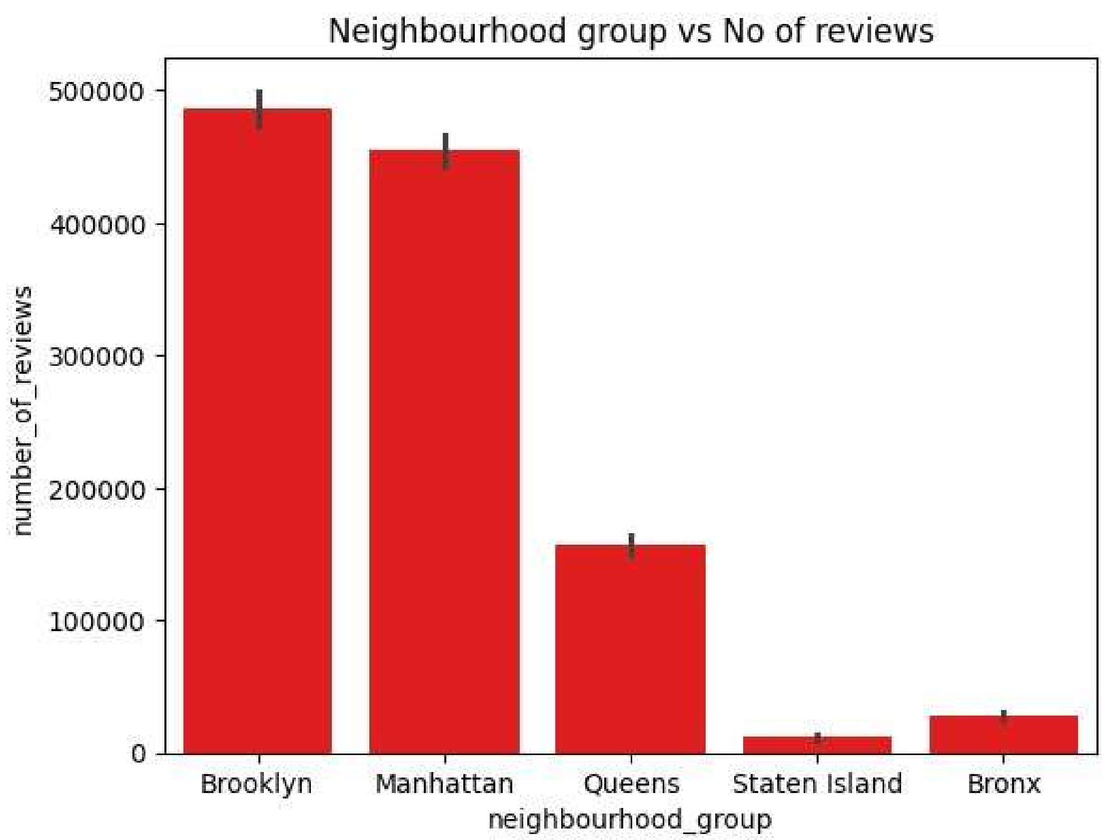
Are there any insights that lead to negative growth? Justify with specific reason.

Understanding that Manhatten has the highest average price can help the business tailor their pricing strategies to maximize the revenue.For listings in manhatten,business might consider premium pricing for upscale offerings to allign with the high demand and willingne3ss of the guest to pay more in this area.

Cpmversaly, for listings in more affordable neighbourhoods like brooklyn and queens,business can attract pricing to attract price sensitive travellers , potentially increasing occupancy rates.

Chart - 11

```
# Chart - 12 visualization code sns.barplot(y = data['number_of_reviews'], x = data['neighbourhood_group'],
estimator = sum, color ='red') plt.title('Neighbourhood group vs No of reviews')
plt.show()
```



1. Why did you pick the specific chart?

Bar plot is choosen to realise the number of reviews across different neighbourhood_group categories.Bar plot are effective for comparing quantities among discrete categories.

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

<> Brooklyn Leads in Reviews : The bar plot shows that the brooklyn have the highest number of reviews,approximately 500000 .This suggests that Brooklyn is the most popular neighbourhood group among guests,potentially due to the diverse attraction,accomodations options, tob overall appeal.

<> Manhatten Close Behind : Manhatten is mjust below brooklyn in terms in the number of revies.With the significant numbe of reviews,manhatten also attract ahigh volume of guests,which align with its ststus as a major travel destination.

<> Queens and Moderate reviews :Queens has around 150000 reviews,indicate a moderate level of guests activity compared to Brooklyn and Manhatten. It is less popular than top of two neighbourhoods but still attracts a noteable amount of visitors.

<> Bronx and Staten island Lowest : The Bronx and Staten island have the lowest number of reviews,with very few reviews as compared to the other neighbourhood .This suggests that these areas may be less frequented by guests,potentially due to fewer attractions,less assessibility,or other factors influencing their popularity.

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

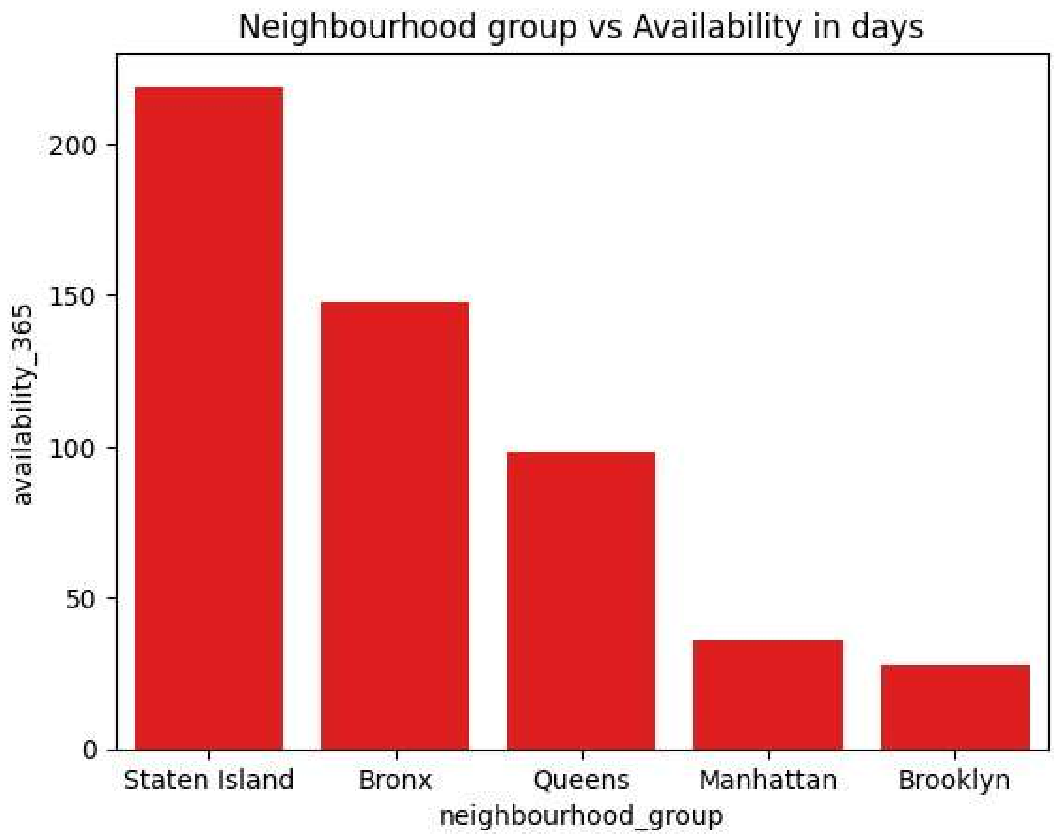
Are there any insights that lead to negative growth? Justify with specific reason.

The insights that Brooklyn and Manhatten have the highest number of reviews can guide targeted marketing efforts .Business can focus their potential activities on these high traffic areas to attract more guests.For example special deals and exclusive offer in brooklyn could capiutalize on its popularity and further boosts booking.

□ Chart - **12**

```
f= data.groupby('neighbourhood_group')['availability_365'].median().reset_index() f=
f.sort_values('availability_365', ascending = False) sns.barplot(x=
f['neighbourhood_group'], y = f['availability_365'], color = 'red')
plt.title('Neighbourhood group vs Availability in days')
```

```
plt.show()
```



1. Why did you pick the specific chart?

Barploits are ideal for comparing a categorical variable(neighbourhood_group) against a summary of numerical variable (available_365).It visually conveys how availability varies across different neighbourhoods.

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

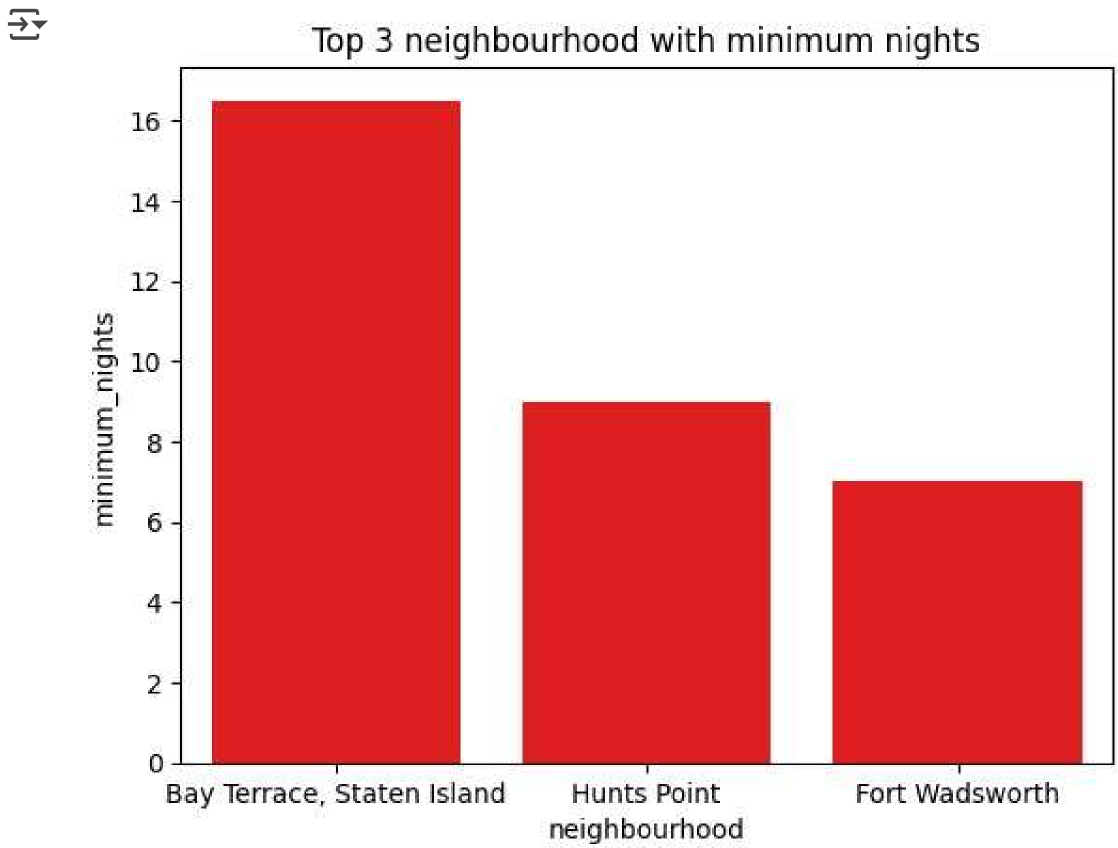
Staten island stand with out with over 200 days of availability, indicating that properties in this neighbourhood are generally available for booking much longer than in other areas.This could suggests lower demand for short term rentals, or perhaps hosts keep their property open for more extended periods

Bronx and queens shows around 150 days and 100 days of availability,respectively indicating moderate availability.These neighbourhood may have moderate demand for rentals,with some properties being booked while other remain available for longer periods.

Manhatten and Brooklyn both have less than 50 days of availability,which could indicate for a high demand for short term rentals.Properties in this area are likely booked frequently,leading to fewer daus of availability throughout the year.

? CHART - 13

```
data.columns y =
data.groupby('neighbourhood')['minimum_nights'].median().reset_index() y =
y.sort_values('minimum_nights',ascending = False).head(3) sns.barplot(x =
y['neighbourhood'], y = y['minimum_nights'], color = 'red') plt.title('Top 3
neighbourhood with minimum nights')
plt.show()
```



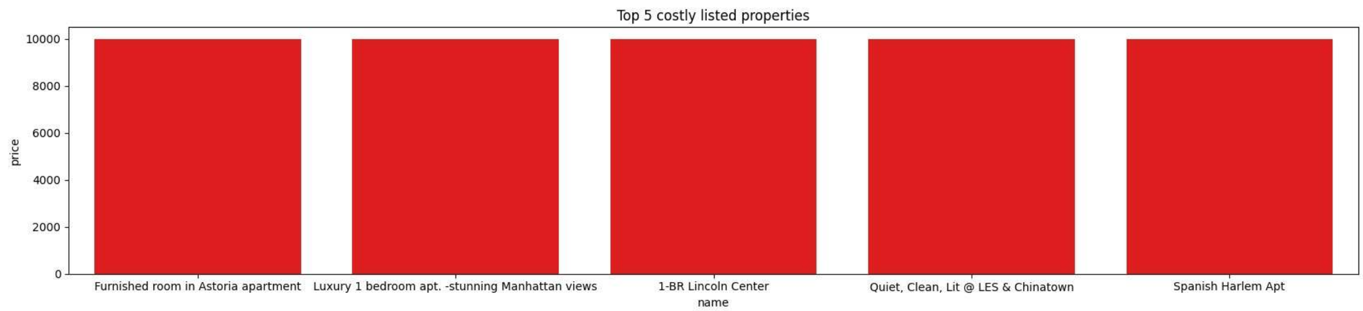
What are the insights found from the data ?

Bay Terrace,State island has the highest median minimum nights of 16 nights, suggestingb that the property in the neighbourhood typically require guests to book longer stays .This could indicate either lower demand for a sort stays or a strategy by hosts to focus on longer term renters.

? CHART - 14

```
data.columns z =
data.groupby('name')['price'].max().reset_index()
plt.figure(figsize = (17,4)) z =
z.sort_values('price',ascending = False).head(5)
plt.title("Top 5 costly listed properties") sns.barplot(x
= z['name'],y = z['price'], color = 'red')
plt.tight_layout()
```

```
plt.show()
```

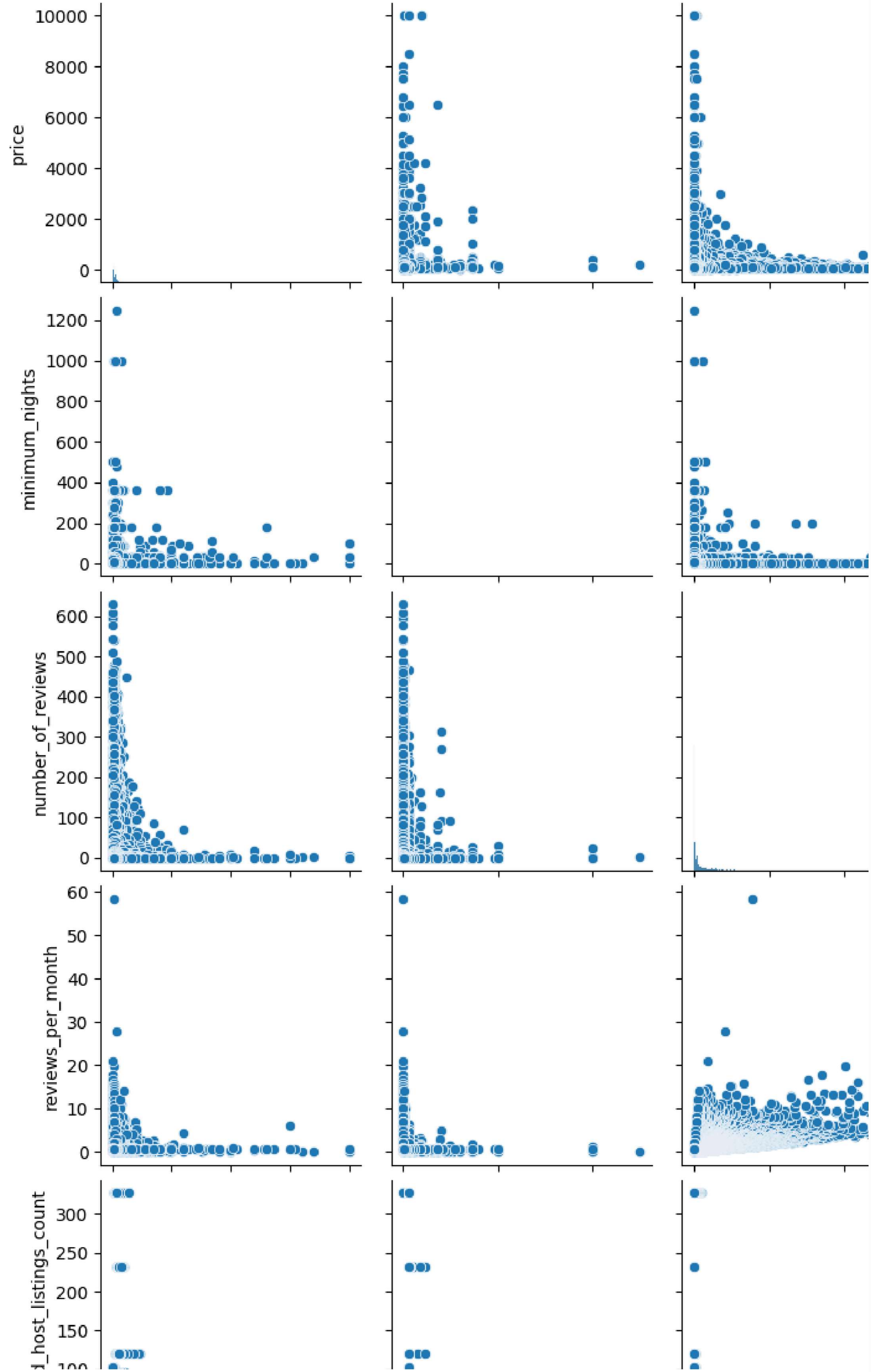


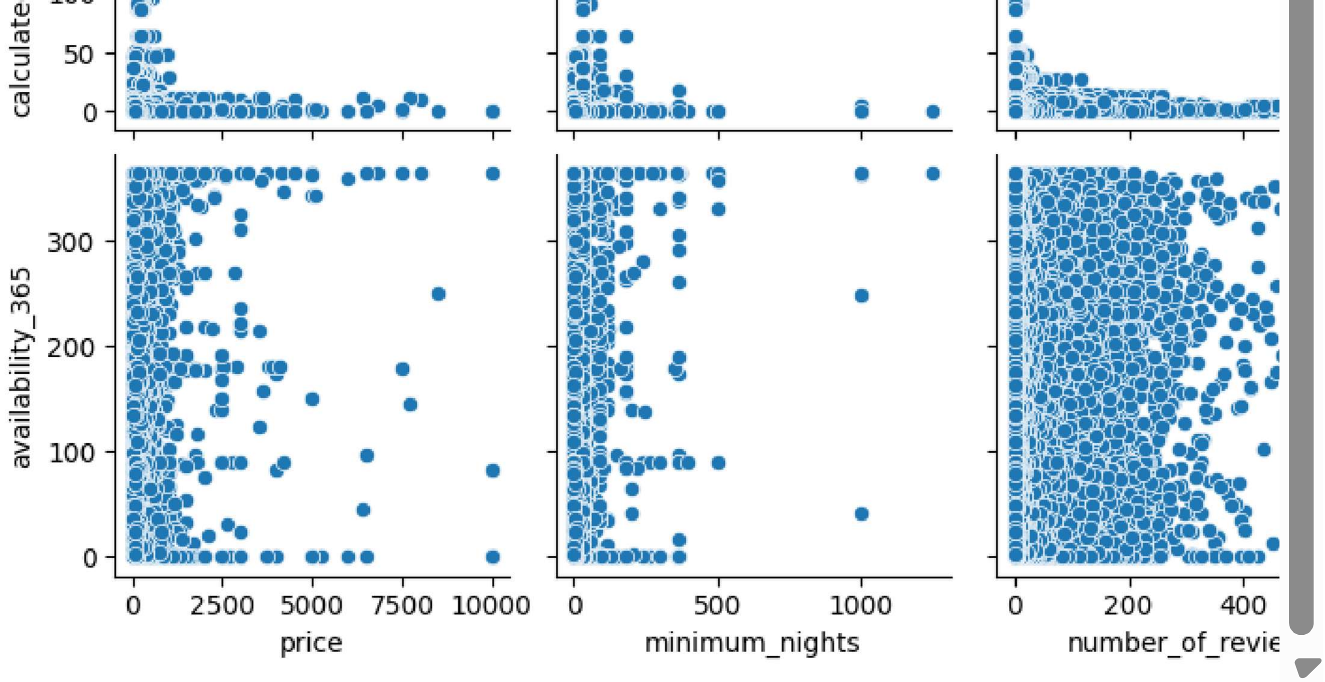
What are the insights found from the data ?

The Bar plot graph reveals that these 5 properties are 5 top costly and high demand. They are costly maybe they are in high posh area or business area

? CHART - 15

```
data.columns num =
data[['price','minimum_nights','number_of_reviews','last_review','reviews_per_month','calculated_host_listings_count','availability_365']]
sns.pairplot(num) plt.show()
```


price
minimum_nights
number_of_reviews
reviews_per_month
d_host_listings_count
10000
8000
6000
4000
2000
0
1200
1000
800
600
400
200
0
600
500
400
300
200
100
0
60
50
40
30
20
10
0
300
250
200
150
100



Why did you pick the specific chart?

A pair polt was choosen to visualize the relationship between multiple numerical variables simultaneously.Pair plot are particularly useful for exploring how variables interact with each other and identifying patterns,correlations, and potential outliers.

Solution to Business Objective**

What do you suggest the client to achieve the business objective?

Strategic Recommendations

<> **Enhance Pricing Tools** : Introduce dynamic pricing market based on marketn trends to help host optimise rates.

<> **Boost Reviews And Bookings** : Encourage guests reviews and promote properties with shorter minimum stays for higher engagement.

<> **Leverage Popular Neighbourhoods** : Focus marketing on high demand areas like Manhatten and Brooklyn for growth.

<> **Improve less popular areas** : Promote Staten Island,the Bronx and the Queens through targeted campaign and offers.

<> **Optiimise for longer stays** : Offer discount for extended stays and promote properties with higher availability.

<> **Differentiate Rooms Types** : Tailormarketing for entire homes as premium options and emphasise private/share rooms for affordability.

<> **Learn from top hosts** : Share best practices from high review hosts to improve service and satisfaction. <> **Always stay market responsive** : Regularly adjust strategies bases on pricing,booking patterns,and guest prefences.

These action help in enhancing AirBnb offerings,boost host performance and drive growth.

CONCLUSION

The explarotary data analysus revealed the the key opportunities for airbnb to enhance business performance

1. **Pricing Tools :** Address price outliers by implementing dynamic pricing tools to help hosts set competitive rates and maximise revenue
2. **Guest engagement :** Increase reviews and booking through strategies like incentivizing reviews and promoting flexible booking options.
3. **Neighbourhood Focus :** Target high demand areas like Manhattan and Brooklyn for marketing , while boosting visibility in lower engagement areas like staten island and The Bronx.
4. **Stay Duration :** Promote properties with longer stays and offer discounts to attract extended stays guests.
5. **Room Type Differentiation :** Tailor strategies for entire homes for privacy and shared rooms for affordibility to meet diverse guests preferences.
6. **Successful Hosts :** The practices of highly successful hosts like Dona should be analyzed and shgared to eleverate the performance of other listings. Learning from top performer can enhance service quality and guests satisfaction across the platform.

By leverging these insights and ijmplementing targeted strategies, the airbnb platform can enhance its market position, drive higher engagement and optimize both host and giests experiences. Continuous monitoring and adoptations tob market trenda will ensure sustain growth and competitiveness to the evolving short term rental market.

The chart reveals the reviews_per_month and number_of_reviews are highly positively correlated . This suggests that listings with more reviews per month tend to have a higher total number of reviews ,indicating that frequent reviews are associated with total higher review count.

All other variables doesnot show any corelation with each other.