Project Name - AirBnb NYC Analysis

Project Type - EDA

Contribution - Team

Team Member 1 - Priya Debrani

Project Summary -

Write the summary here within 500-600 words.

This Airbnb ('AB_NYC_2019') dataset for the 2019 year appeared to be a very rich dataset with a variety of columns that allowed us to do deep data exploration on each significant column presented. First, we have found hosts that take good advantage of the Airbnb platform and provide the most listings; we found that our top host has 327 listings. After that, we proceeded with analyzing boroughs and neighborhood listing densities and what areas were more popular than another. Next, we put good use of our latitude and longitude columns and used to create a geographical heatmap color-coded by the price of listings. Further, we came back to the first column with name strings and had to do a bit more coding to parse each title and analyze existing trends on how listings are named as well as what was the count for the most used words by hosts. Lastly, we found the most reviewed listings and analyzed some additional attributes. For our data exploration purposes, it also would be nice to have couple additional features, such as positive and negative numeric (0-5 stars) reviews or 0-5 star average review for each listing; addition of these features would help to determine the best-reviewed hosts for NYC along with 'number_of_review' column that is provided. Overall, we discovered a very good number of interesting relationships between features and explained each step of the process. This data analytics is very much mimicked on a higher level on Airbnb Data/Machine Learning team for better business decisions, control over the platform, marketing initiatives, implementation of new features and much more.

GitHub Link -

PriyaDebrani/Airbnb-NYC-2019-Analysis-: EDA_PROJECT (github.com)

Problem Statement

Write Problem Statement Here.

1) What can we learn about different Hosts and Areas?

2) How are the rentals distributed among the 5 boroughs?

3) What is the price distribution and what's the range of the fair prices available?

4) Which Hosts are busiest?

5) Which are the top 10 hosts on the basis of review?

6) Which are the top 10 hosts on the basis of count of listing?

Define Your Business Objective?

The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data.

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 executable in one go
          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

```
# Import Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
import missingno as msno
```

Dataset Loading

```
# Load Dataset
from google.colab import drive
drive.mount('/content/drive')
airbnb='/content/Airbnb NYC 2019.csv'
df_airbnb=pd.read_csv(airbnb)
```

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

Dataset First View

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

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	pric
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	14
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	22
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	15
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	8
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	8

Dataset Rows & Columns count

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

```
(48895, 16)
```

Dataset Information

```
# Dataset Info
```

```
df_airbnb.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
df_airbnb.duplicated().sum()
```

```
0
```

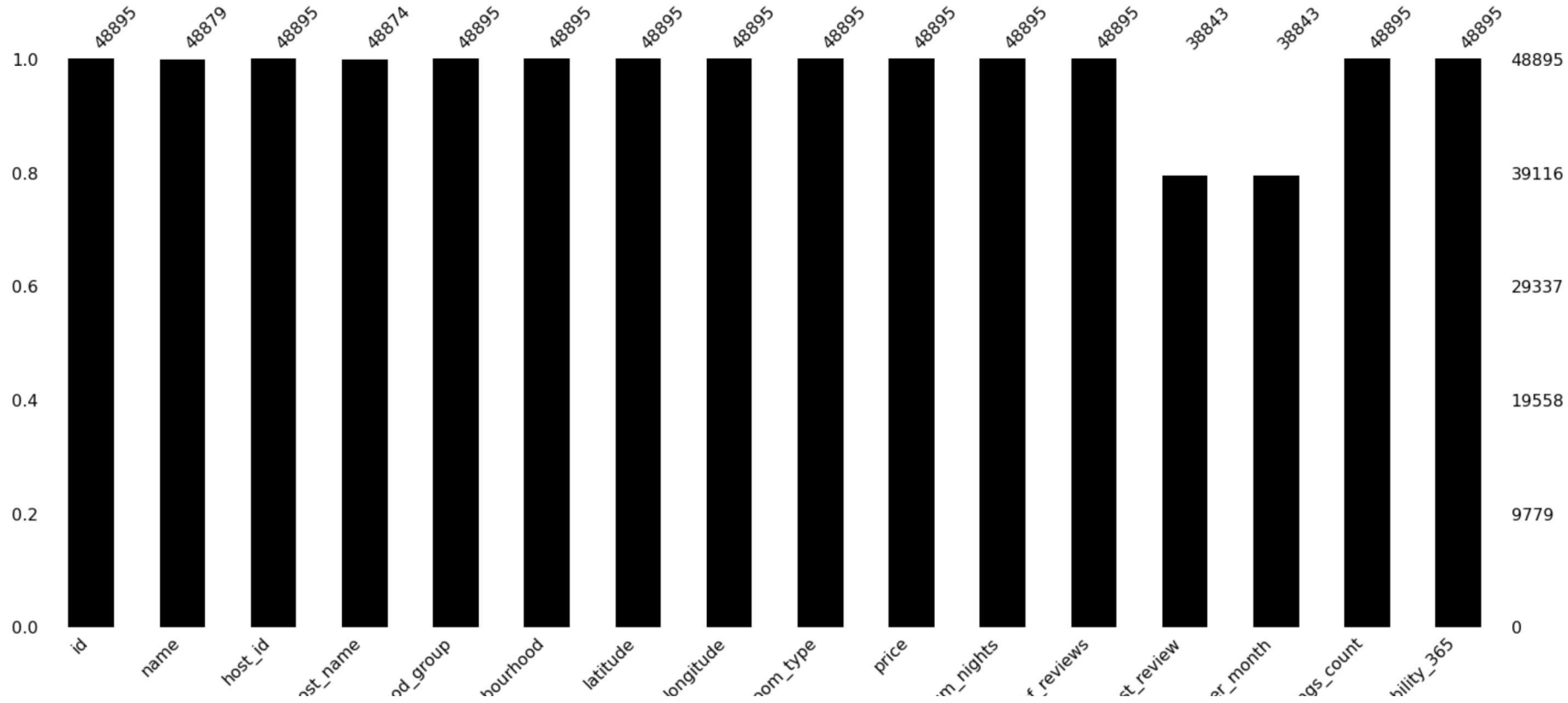
Missing Values/Null Values

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

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

```
# Visualizing the missing values
msno.bar(df_airbnb, color='Black')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7935cfa1f0>
```



What did you know about your dataset?

Following are the various columns in the present in the dataset :-

1-id:- This column contains the room booking id. It contains unique values.

2-name:- This column contains the name of the lodge where room is booked.

3-host_id:- This column contains the host id as registered on the airbnb platform.

4-host_name:- This column contains the host name as registered on the airbnb platform.

5-neighbourhood_group:- This column contains the name of the neighbourhood group where the room is booked

6-neighbourhood:- This column contains the neighbourhood name to narrow down the location precision provided by the neighbourhood_group column. For example if my neighbourhood group is manhattan then the neighbourhood column will basically tell - where in manhattan.

7-latitude:- This column stores the geographical latitude value of the location of lodging place.

8-longitude:- This column stores the geographical longitude value of the location of lodging place.

9-room_type:- This column stores the room type which is booked i.e single room, shared room, apartment etc.

10-price:- This column stores the price in USD of the room booked.

11-minimum_nights:- This column stores the minimum number of nights one must book for the particular room.

12-number_of_reviews:- This column shows the total number of reviews a room has recieved so far.

13-last_review:- This column shows the date when the last review for a particular room was published.

14-reviews_per_month:- This column shows the average number of reviews a room recieves per month.

15-calculated_host_listings_count:- This column shows the total number of listings by a particular host.

16-availability_365:- his column stores the number of days a particular room is avaiable for booking throught the year.

2. Understanding Your Variables

```
# Dataset Columns
df_airbnb.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
df_airbnb.describe(include='all')
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	r
count	4.889500e+04	48879	4.889500e+04	48874	48895	48895	48895.000000	48895.000000	
unique	NaN	47905	NaN	11452	5	221	NaN	NaN	
top	NaN	Hillside Hotel	NaN	Michael	Manhattan	Williamsburg	NaN	NaN	
freq	NaN	18	NaN	417	21661	3920	NaN	NaN	
mean	1.901714e+07	NaN	6.762001e+07	NaN	NaN	NaN	40.728949	-73.952170	
std	1.098311e+07	NaN	7.861097e+07	NaN	NaN	NaN	0.054530	0.046157	
min	2.539000e+03	NaN	2.438000e+03	NaN	NaN	NaN	40.499790	-74.244420	
25%	9.471945e+06	NaN	7.822033e+06	NaN	NaN	NaN	40.690100	-73.983070	
50%	1.967728e+07	NaN	3.079382e+07	NaN	NaN	NaN	40.723070	-73.955680	
75%	2.915218e+07	NaN	1.074344e+08	NaN	NaN	NaN	40.763115	-73.936275	
max	3.648724e+07	NaN	2.743213e+08	NaN	NaN	NaN	40.913060	-73.712990	

Variables Description

By basic inspection i figured out that a particular property name will have one particular host name hosted by that same individual but a particular host name can have multiple properties in an area. So, host_name is a categorical variable here. Also neighbourhood_group (comprising of Manhattan, Brooklyn, Queens, Bronx, Staten Island), neighbourhood and room_type (private,shared,Entire home/apt) fall into this category.

While id, latitude, longitude, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365 are numerical variables.

Check Unique Values for each variable.

```
# Check Unique Values for each variable.
df_airbnb.nunique()
```

```
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.
df_airbnb.fillna(0, inplace=True)
```

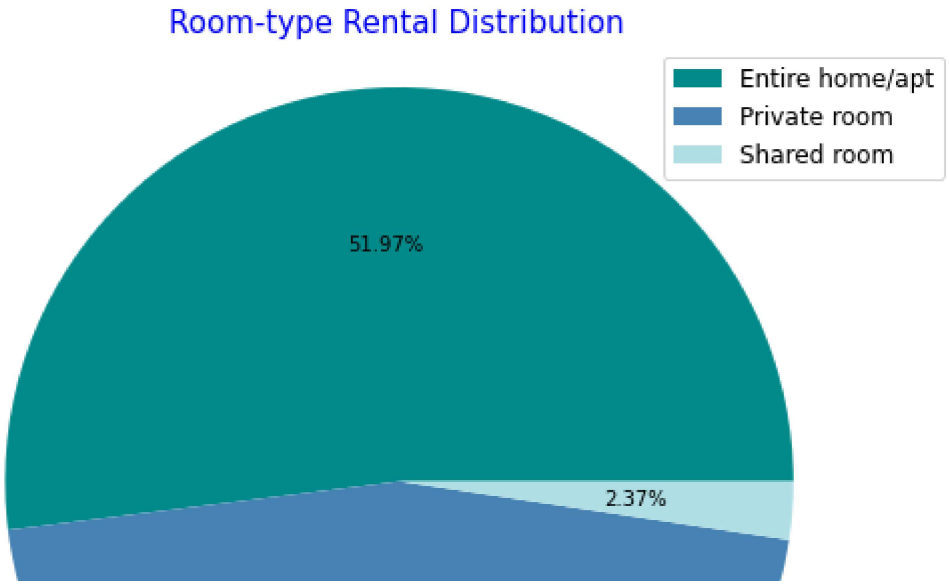
What all manipulations have you done and insights you found?

Here in this dataset total 48895 rows are present and 16 coloumns are there.

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

Chart - 1

```
# Chart - 1 visualization code
room_type = df_airbnb.groupby('room_type')['latitude'].count().reset_index()
room_type.rename(columns={'latitude':'n_rooms'},inplace=True)
plt.figure(figsize=(10,8))
plt.pie(room_type['n_rooms'],autopct='%1.2f%%', colors=['darkcyan', 'steelblue','powderblue'])
plt.axis('equal')
plt.legend(labels=room_type['room_type'],loc='best',fontsize='12')
plt.title('Room-type Rental Distribution', fontsize='15',color='b')
plt.show()
plt.close()
```



- 1. Why did you pick the specific chart?

  I took this pie chart to Know, What proportion of the rentals correspond to each room type?

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

  From the whole rentals available in the dataset, 52% of them correspond to entire-home apartments, 46% to private-room rentals and the minority remaining corresponds to shared rooms with a 2% of the sample.

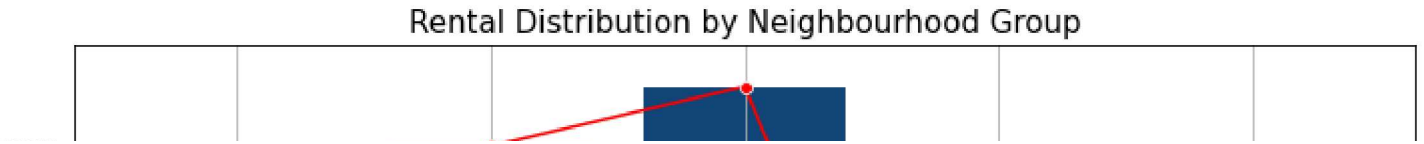
- 3. Will the gained insights help creating a positive business impact?

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

  No, there is no insight which can lead us to negative growth.

- Chart - 2

```
# Chart - 2 visualization
neighbourhood = df_airbnb.groupby('neighbourhood_group')['neighbourhood'].count().reset_index()
fig,ax = plt.subplots(figsize=(12,8))
sns.barplot(x=neighbourhood[neighbourhood.columns[0]],
y=neighbourhood[neighbourhood.columns[1]],color='#004488',ax=ax)
sns.lineplot(x=neighbourhood[neighbourhood.columns[0]],y=neighbourhood[neighbourhood.columns[1]],color='r',marker='o',ax=ax)
plt.ylabel('Rentals', fontsize='15')
plt.xlabel('Borough',fontsize='15')
plt.title('Rental Distribution by Neighbourhood Group',fontsize='15')
plt.grid('x')
plt.show()
sns.set()
```



1. Why did you pick the specific chart?

How are rentals distributed among the five boroughs that I mentioned in the introduction?

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

As it's reflected in the visualization, Brooklyn and Manhattan boroughs concentrate the majority of the listed rentals on Airbnb, adding up more than 40,000 rentals between the two of them. This means that the bulk of visitors of New York stay in properties, rooms or residencies located in these areas.

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

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

No, there is no insight which can lead us to negative growth.

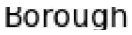


Chart - 3

```
# Chart - 3 visualization code
price = df_airbnb.loc[:,['neighbourhood','price']].set_index('neighbourhood')
price_stats = df_airbnb['price'].describe().reset_index()
price_counts = price.price.value_counts().reset_index()
price_counts.rename(columns={'index':'price','price':'count'},inplace=True)
fig2,ax = plt.subplots(figsize=(12,8))
fig2.patch.set_facecolor('lightgray')
ax.set_facecolor('lightgray')
```

```
plt.hist(price_counts['price'],bins=30,color='#004488',edgecolor='salmon')
ax.set_xticks(range(0,10000,500))
for tick in ax.get_xticklabels():
  tick.set_rotation(45)
plt.xlabel('Price',fontsize='15')
plt.ylabel('Rentals', fontsize='15')
plt.xlim((-0.5,10000))
plt.title('New York Price-Rental Distribution',fontsize='15')
plt.show()
```

1. Why did you pick the specific chart?

What's the price distribution and what's the range of fair prices available?

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

Price distribution is focused around the $ 300–400 prices with few observations that present higher prices.

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

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

No, there is no insight which can lead us to negative growth.

Chart - 4

```
# Chart - 4 visualization code
sns.countplot(x="neighbourhood_group",data=df_airbnb)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f79315ee280>
```


- 1. Why did you pick the specific chart?

Here, I chose Countplot to show neighbourhood vs count data.

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

This graph below is a plot between Neighbourhood groups and the number of airbnbs that are in that area.From this graph we infer that Manhattan has the maximum number of airbnbs in the whole of NYC

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

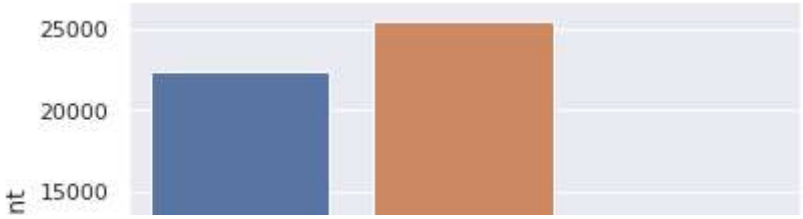
Are there any insights that lead to negative growth? Justify with specific reason.

Yes, there can be negative impact of such groups which have less counts such as Staten Island, Bronx etc.

- Chart - 5

```
# Chart - 5 visualization code
sns.countplot(x="room_type",data=df_airbnb)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f793143d400>
```



1. Why did you pick the specific chart?

Here, I chose Countplot to show room_type vs count data.

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

This graph below is a plot between room types and the number of airbnbs that are of that type.From this graph we infer the maximum number of airbnbs in the whole of NYC are that of entire room type.

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

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

No, there is no such data which can lead to negative growth.

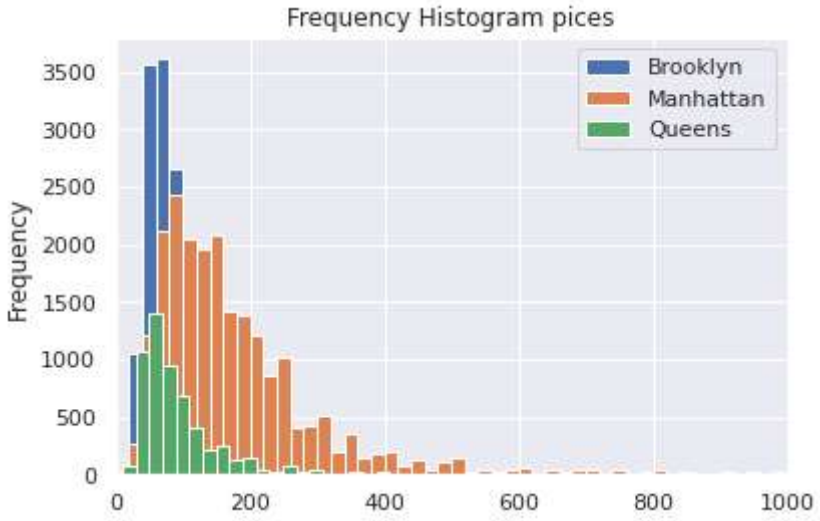
Chart - 6

Explore one or more columns by plotting a graph below, and add some explanation about it

```
pricedist= df_airbnb[df_airbnb['price'] <= 2000]
x1 = pricedist.loc[pricedist.neighbourhood_group =='Manhattan', 'price']
x2 = pricedist.loc[pricedist.neighbourhood_group =='Brooklyn', 'price']
x3 = pricedist.loc[pricedist.neighbourhood_group =='Queens', 'price']
```

```
x4 = pricedist.loc[pricedist.neighbourhood_group =='Staten Island', 'price']
x5 = pricedist.loc[pricedist.neighbourhood_group =='Bronx', 'price']
```

```
# Chart - 6 visualization code
kwargs = dict(alpha=0.5, bins=100)
kwargx = dict(alpha=0.5, bins=100)
plt.hist(x2, bins=100,label='Brooklyn')
plt.hist(x1,bins=100,label='Manhattan')
plt.hist(x3,bins=100, label='Queens')
plt.gca().set(title='Frequency Histogram pices', ylabel='Frequency')
plt.xlim(0,1000)
plt.legend();
```



- 1. Why did you pick the specific chart?

Here , I created 3 histograms and gcs() is used to get the current Axes instance on the current figure matching the given keyword args.

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

Frequency of Brooklyn is highest.

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

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

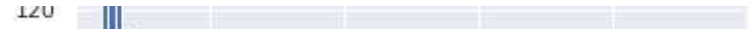
No, there is no insight which will lead us to negative growth.

Chart - 7

Explore one or more columns by plotting a graph below, and add some explanation about it

```
# Chart - 7 visualization code
plt.hist(x5, bins=100,label='Bronx')
plt.hist(x4,bins=100,label='Staten Island')
plt.gca().set(title='Frequency Histogram pices', ylabel='Frequency')
plt.xlim(0,1000)
plt.legend();
```

- 1. Why did you pick the specific chart?



Here , I created 2 histograms and gcs() is used to get the current Axes instance on the current figure matching the given keyword args.

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

Bronx has the highest frequency between Bronx and Staten Island.

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

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

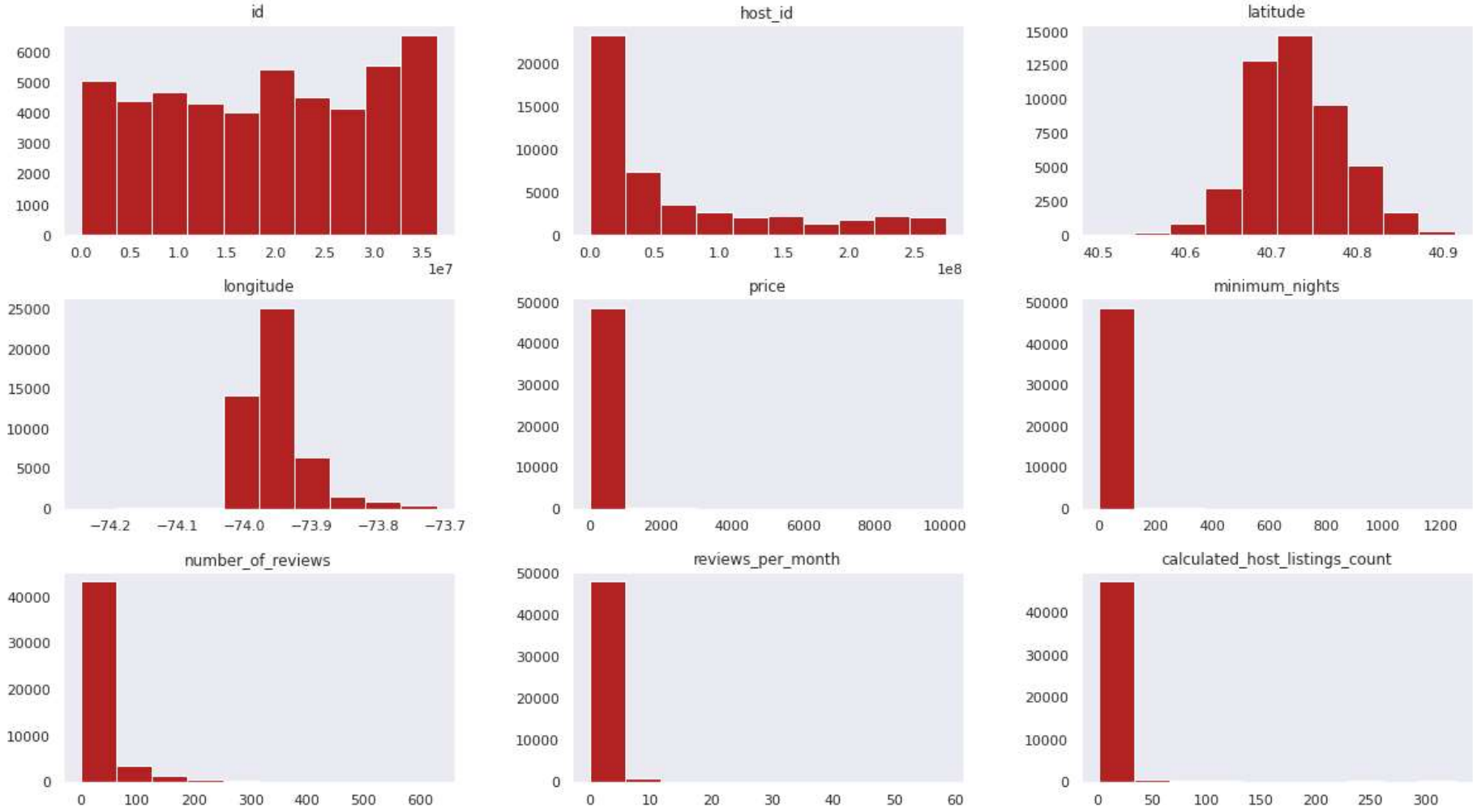
No , there is no such data which will lead us to negative growth.

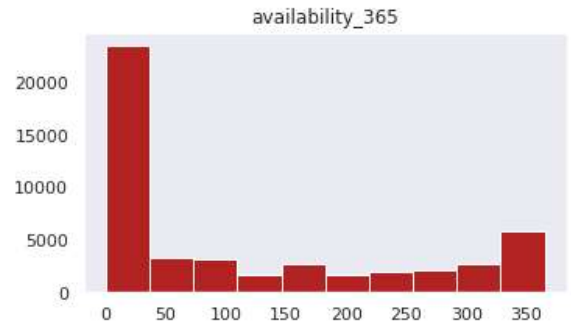
- Chart - 8

```
# Chart - 8 visualization code

df_airbnb.hist(figsize=(20,15), grid = False, color = 'firebrick')
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f7930f19e20>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7930ec8d00>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7935b1d3d0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7930fae9d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7931013940>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f79310884f0>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f79310885e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f79310a2c70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f793140a730>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f7930ecb2b0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7930e1a5e0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f7930e31b80>]],
      dtype=object)
```





1. Why did you pick the specific chart?

Here, I chose this graph to View distribution of numeric data.

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

Some of the numerical values are positively skewed; Therefore, for the analysis, I chose to focus on the median values, because it is less susceptible to skewed data.

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

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

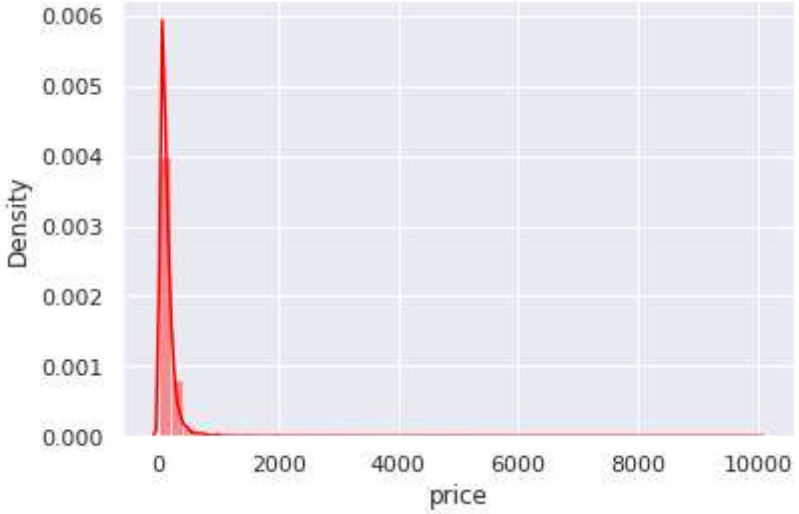
No

Chart - 9

```
# Chart - 9 visualization code
# check the distribution of price
```

```
sns.distplot(df_airbnb['price'], color ="Red")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f792fa0a490>
```



▼ 1. Why did you pick the specific chart?

Here, I chose distplot to show the density and price data range.

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

The price column is heavily skewed with most of the price ranging between 10 to 200$.

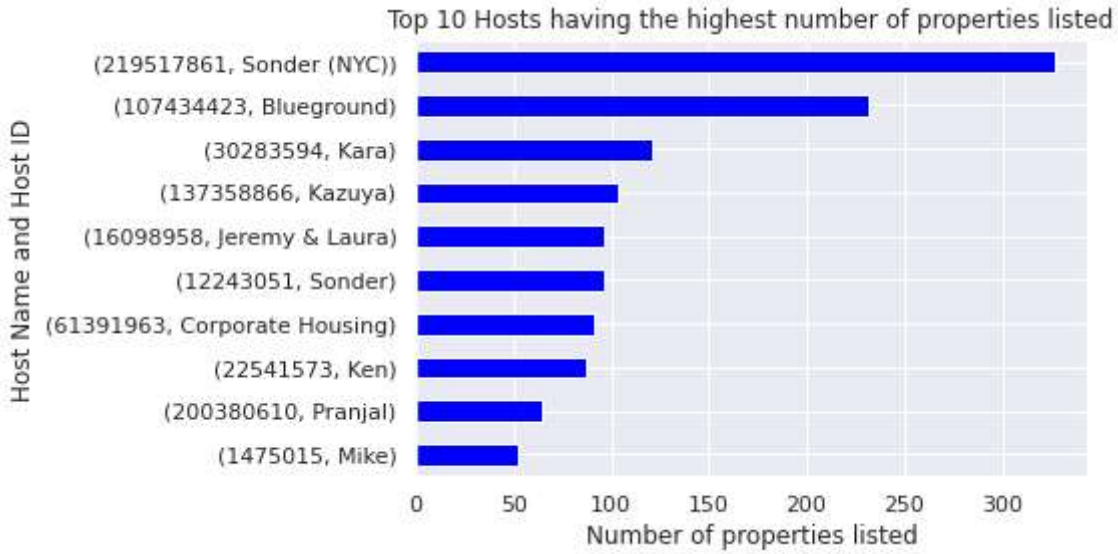
▼ 3. Will the gained insights help creating a positive business impact?

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

There are a very few observations where the minimum nights is very high and thus for those very few observations the price could go high upto 10000$. Hence, they aren't considered as outliers here.So it can lead us to negative growth.

Chart - 10

```
# Chart - 10 visualization code
# top 10 hosts on the basis of count of listings
top_hosts_listings = df_airbnb.groupby(['host_id','host_name'])['host_id'].count().sort_values(ascending=False)[:10]
top_hosts_listings.plot.barh(color="Blue").invert_yaxis()
plt.xlabel('Number of properties listed')
plt.ylabel('Host Name and Host ID')
plt.title('Top 10 Hosts having the highest number of properties listed')
plt.show()
```



1. Why did you pick the specific chart?

Here, I chose barplot to show top 10 hosts on the basis of count of listings.

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

Sonder(NYC) has the highest number of properties.

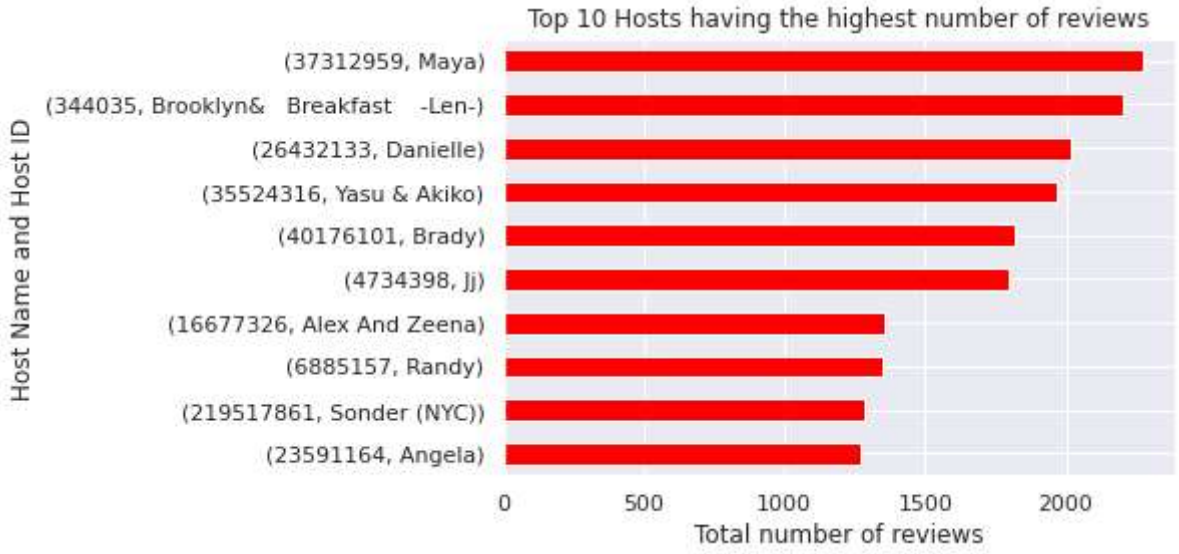
3. Will the gained insights help creating a positive business impact?

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

No , there is no such data which will lead us to negative growth.

Chart - 11

```
# Chart - 11 visualization code
# top 10 hosts on the basis of reviews
top_hosts_reviews = df_airbnb.groupby(['host_id','host_name'])['number_of_reviews'].sum().sort_values(ascending=False)[:10]
top_hosts_reviews.plot.barh(color="Red").invert_yaxis()
plt.ylabel('Host Name and Host ID')
plt.xlabel('Total number of reviews')
plt.title('Top 10 Hosts having the highest number of reviews')
plt.show()
```



1. Why did you pick the specific chart?

Here, I chose barplot and inverted it to show host name,ID and total number of reviews.

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

We can clearly see that Sonder (NYC) is not the top host who has received the most number of reviews.

As host Maya receives the most number of reviews we can infer that she gets the most number of customers.

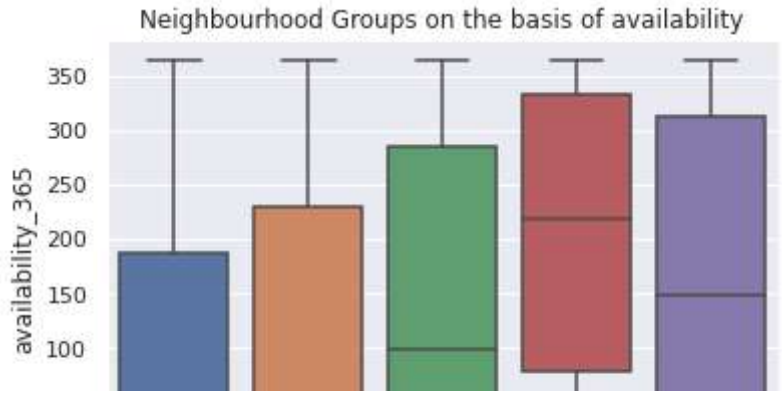
3. Will the gained insights help creating a positive business impact?

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

Yes, these insights can lead to negative growth for the hosts which have less reviews as it can lead customers to pre-assume negative thought about host.

Chart - 12

```
# Chart - 12 visualization code
# top neighbourhood groups on the basis of availability
plt.figure()
plt.title('Neighbourhood Groups on the basis of availability')
sns.boxplot(data=df_airbnb, x='neighbourhood_group', y='availability_365')
plt.show()
```



1. Why did you pick the specific chart?

Here , I chose Boxplot to show neighbourhood on the basis of availibility.

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

We can infer that the listings in Staten Island seems to be more available throughout the year to more than 300 days. On an average, these listings are available to around 210 days every year followed by Bronx where every listings are available for 150 on an average every year.

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

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

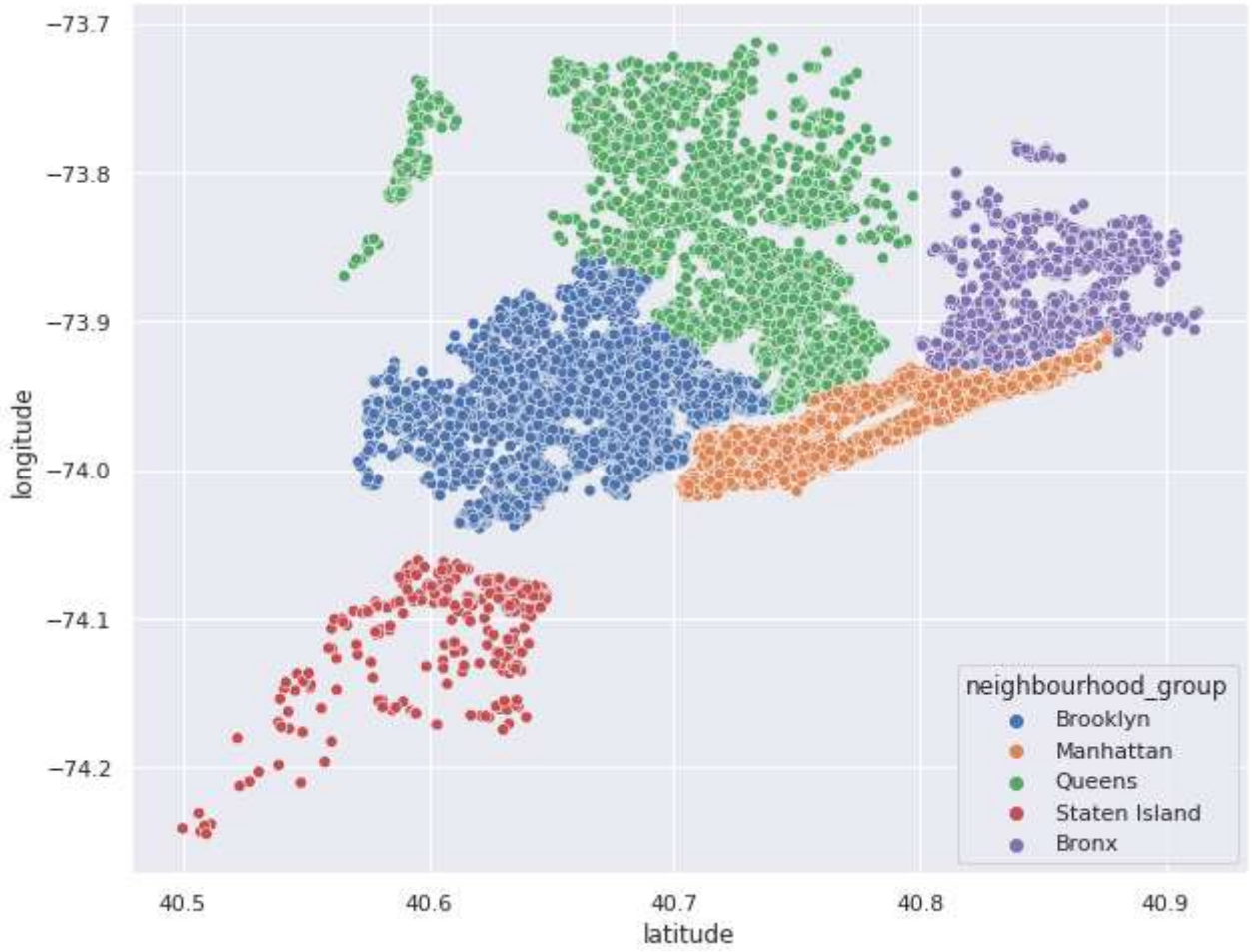
No , there is no such data which will lead us to negative growth.

Chart - 13

```
# Chart - 13 visualization code
#neighborhood group based on the latitude and longitude
```

```
plt.figure(figsize=(10,8))
sns.scatterplot(df_airbnb.latitude,df_airbnb.longitude, hue='neighbourhood_group', data=df_airbnb)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7926978850>
```



▼ 1. Why did you pick the specific chart?

Here, I chose scatterplot to show neighbourhood group on latitude and longitude.

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

The above resemble the map of NYC and shows the various neighbourhoods and the properties listed in each neighbourhood.

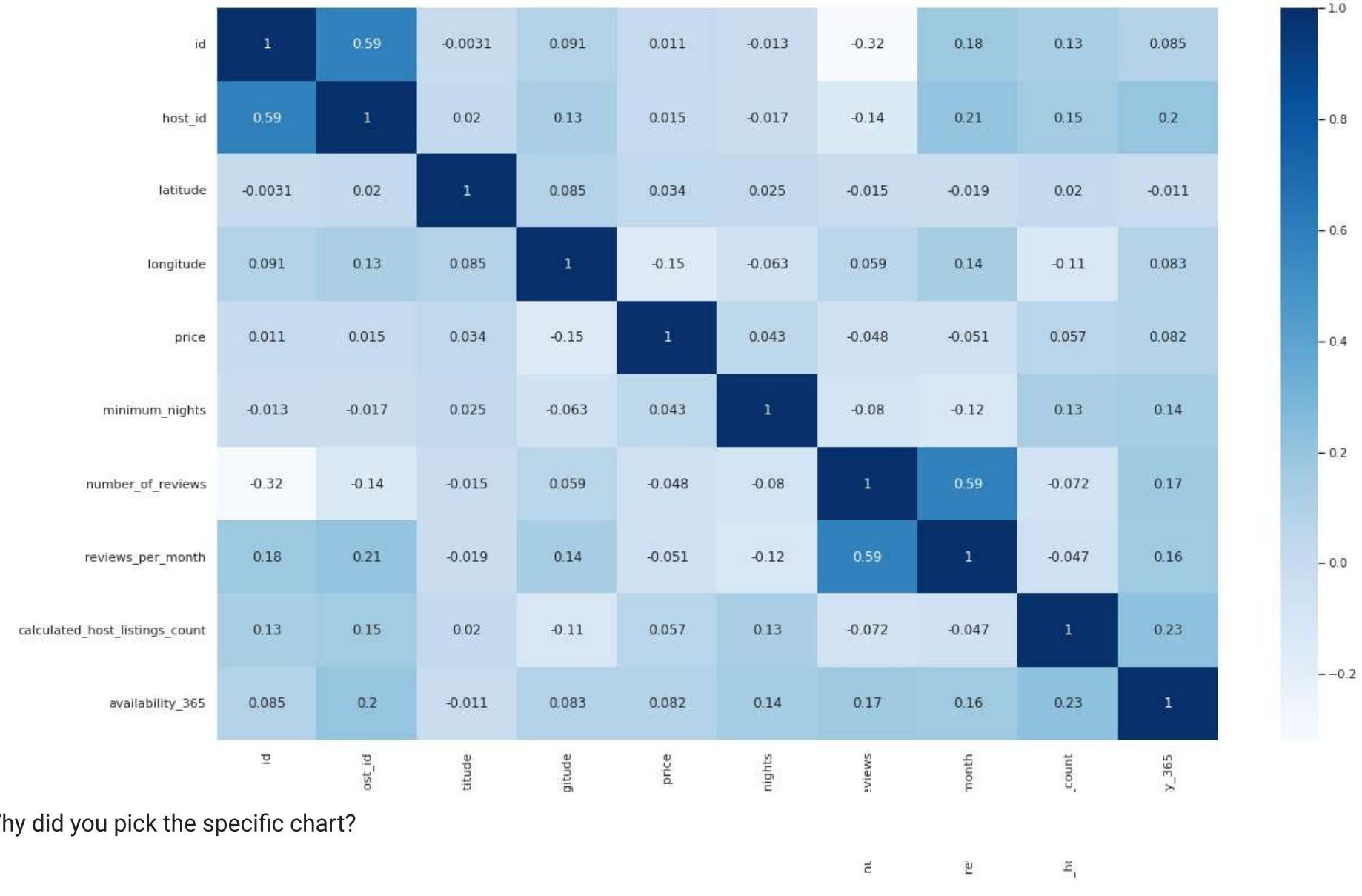
3. Will the gained insights help creating a positive business impact?

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

No , there is no such data which will lead us to negative growth.

Chart - 14 - Correlation Heatmap

```
# Correlation Heatmap visualization code
#correlation of the numerical values
plt.figure(figsize=(20,12))
abnb_corr = df_airbnb.corr()
plt1 = sns.heatmap(abnb_corr ,cbar=True,annot=True, cmap="Blues")
```



▾ 1. Why did you pick the specific chart?

Here,I chose heatmap to see the correlation between different features that can affect a airbnb listing.

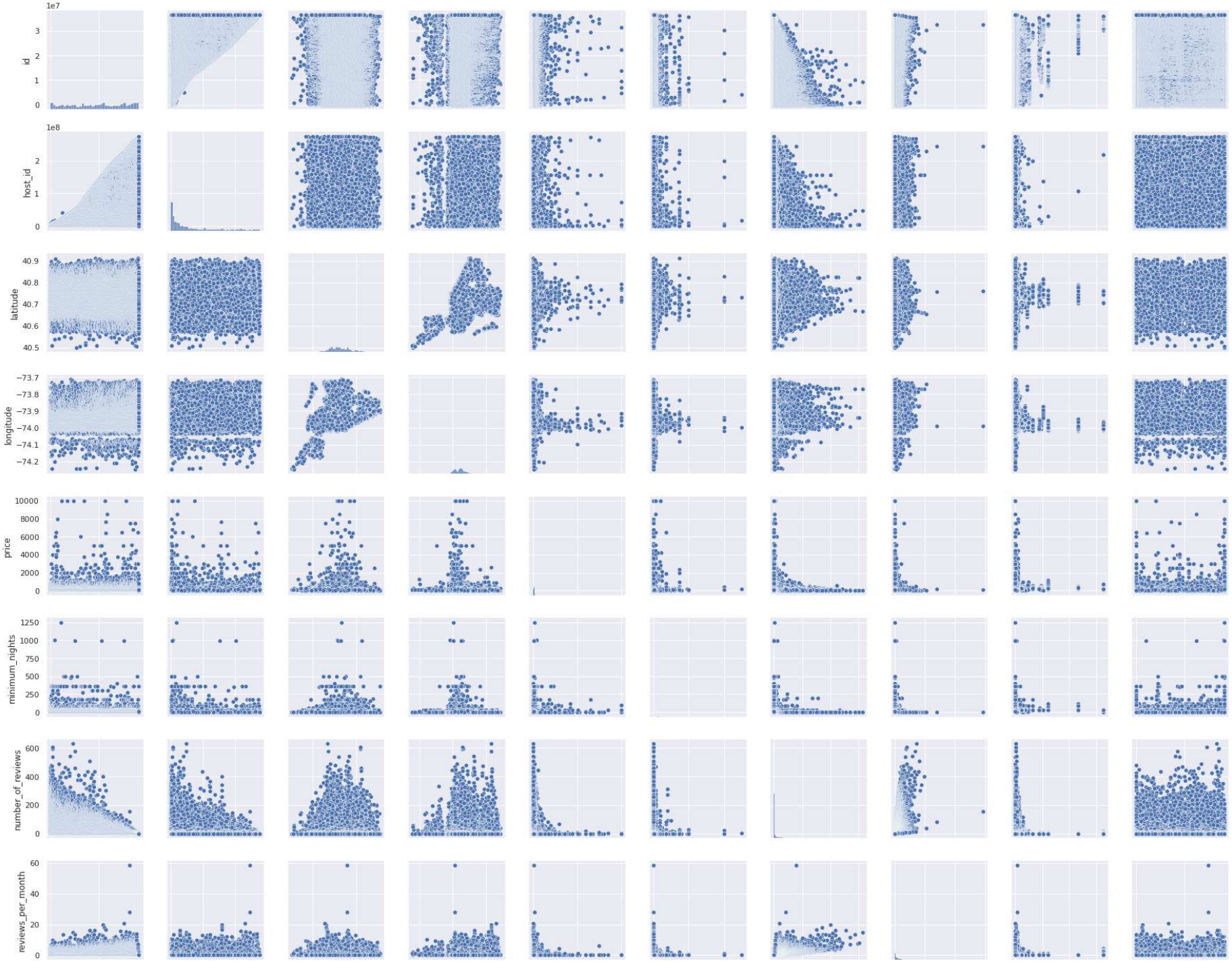
▾ 2. What is/are the insight(s) found from the chart?

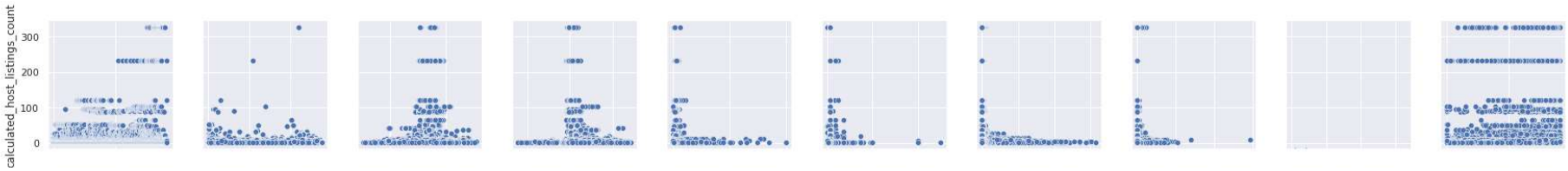
There's correlation among host_id to reveiws_per_month & availability_365. Also there's noticiable correlation between min_nights to no_of_listings_count & availability_365. Price also shows some correlation with availability_365 & host_listings_count.

Chart - 15 - Pair Plot

```
# Pair Plot visualization code
sns.pairplot(df_airbnb)
```

<seaborn.axisgrid.PairGrid at 0x7f79313b84f0>





1. Why did you pick the specific chart?

Here, I chose this scatterplot to show relationships between many quantities like latitude,longitude,price etc.

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

From the above graph maximum prices in longitude and latitude lie between 0 to 4000 dollars.Very few prices are went upto 8000 to $10000. In latitude there is more availability_365.

5. Solution to Business Objective

My first task was to understand the problem statement and the variables, and then look out for any duplicate values, missing values or errors in the data. There were a few missing values which were dropped or imputed with a value depending upon the variable at hand. After dealing with missing values, I performed some univariate, bivariate and multivariate analysis to uncover some insights about hosts, neighborhoods, room types etc. Through EDA, I found out that there are many hosts with multiple properties listed in NYC, hosts who do not have the most number of reviews being occupied for the whole year, hosts who are the busiest and what could be the reason behind it. Exploring the neighborhoods gave me insights like which area is the most expensive, which area gets the most traffic, what type of rooms are available, what number of reviews each area gets, what is the availability of rooms in each area etc.

What do you suggest the client to achieve Business Objective ?

Explain Briefly.

These insights generated can definitely help everyone make better decisions in future to enhance their experience of staying in an Airbnb in NYC. So,please refer it.

Conclusion

The Conclusions we gathered from the EDA are:-

Manhattan and Brooklyn are the most expensive neighborhoods and they receive the most traffic as well. Due to many tourist attractions and the number of properties available, people tend to visit these two areas comparatively more than other ones.

Host Maya is the busiest host in NYC and there are multiple reasons in favor of it like price, minimum nights, availability and number of reviews. She has a total of 5 properties listed in the same neighborhood.

Entire Home/Apt is the costliest room type available but still the most preferred ones for the customers. Entire Home/Apt and Private Rooms receive way more traffic than Shared Rooms and as a result Shared Rooms stay available for most of the time out of the 365 days.

The average price for Private Rooms in Staten Island is the least and has a good availability out of 365 days which makes a good choice for customers seeking low cost accommodations.

Bronx is the least expensive neighbourhood and very less preferred by customers.

Hurrah! You have successfully completed your EDA Capstone Project !!!