Airbnb Booking Analysis
(EDA)

Prakshita Agrawal, Deepak Karki Bhatu Sonawane,
Bharatsoni, Sopan Wadekar
Data science trainees,
AlmaBetter.

Introduction

Airbnb is an online marketplace that connects people who want to rent out their homes with people looking for accommodations in that locale. NYC is the most populous city in the United States, and one of the most popular tourism and business places globally.

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Nowadays, Airbnb became one of a kind service that is used by the whole world. Data analysts become a crucial factor for the company that provided millions of listings through Airbnb. These listings generate a lot of data that can be analyzed and used for security, business decisions, understanding of customers' and providers' behavior on the platform, implementing innovative additional services, guiding marketing initiatives, and much more.

Abstract

The data of airbnb we are make goods insights which is infomative. First we load the dataset then we doing some manipulation in the dataset. Check the columns, rows, and null values. After checking the all things we make some changes on it. Drop which some columns which have most null values like we have last_review and reviews_per_month columns which have the most null value. Null values of name and host_name can be filled by fillna method as these are less.

Change the columns which have some null values with NA. Airbnb dataset contains 49,000 observations and 16 columns which specifies about the hosts, customer and places and it is a mix between categorical and numeric values.

Our goal is to explore and analyse the data, provide helpful conclusions through Exploratory Data Analysis build a statistical model that could be used to effectively predict the price for the listings and future decision makings.

We make a strong insights which shows you all the customers reviews and which place is good for customer stay. Also shows you the busiest and most stayed rooms and host. We mae charts like Bar graph, pie-chart, heatmap, etc. Who helps you to easy understanding the dataset and their goals. This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

We learn different host and areas in the datasets. Why Manhattan and Brooklyn are the most popular area in he dataset. In the dataset you see some area occupations by their percentages. We explore and analyze the data to discover key understandings

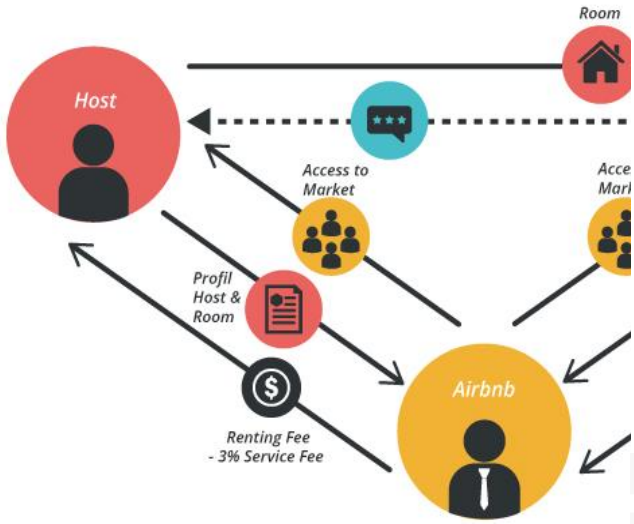
Keywords: Data manipulations, data mining, data describes, Visualizing etc.

Upcoming Content

1. Data description and Data pre-processing
2. Exploratory Data Analysis and Visualization
3. Inferences and conclusions

Business understanding

Airbnb is world wide business airbnb is a community-based online platform for listing and renting local homes. It connects hosts and travelers and facilitates the process of renting without owning any rooms itself. Moreover it cultivates a sharing-economy by allowing property owners to rent out private flats.



On the other side it provides travelers easy access to renting private homes. With over 1,500,000 listings in 34,000 cities and 190

```
top_host_id = airbnb_data['host_id'].value_counts().head(15)
```

```
top_host_id
```

```
219517861    327
107434423    232
30283594     121
137358866    103
16098958      96
12243051      96
61391963      91
22541573      87
200380610     65
7503643       52
1475015       52
120762452     50
2856748       49
205031545     49
190921808     47
Name: host_id, dtype: int64
```

countries, its wide coverage enables travelers to rent private homes all over the world. Personal profiles as well as a rating and reviewing system provide information about the host and what is on offer. Vice versa, hosts can choose on their own who to rent out their space to.

As well as a rating and reviewing system provide information about the host and what is on offer. Vice versa, hosts can choose on their own who to rent out their space to.

In our dataset we do count Top 15 hosts in term of listing counts. We easily check listing counts by doing this code . This total top 15 host ids in term of listing counts .

Data description and Data pre-processing

In the data description and pre-processing we surely make first load the data set and analyzed the data

Import our libs and packages in note book

There are all the packages and lib what we needed in our dataset to manipulate and visualized.

```
# For Manipulations
import numpy as np
import pandas as pd

# For Data Visualization
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

import warnings
warnings.filterwarnings('ignore')
```

We tag comment above the every code. In first for manipulations we import Numpy and Pandas as their alias. Basically the reasons behind to import Numpy and Pandas to pandas help to load the dataframe and numpy helps to create dim in the data.

Second for data visualization we import our libs matplotlib and seaborn as there alias. Matplotlib and seaborn are main helps to make beautifull and understandable charts for you.

%matplotlib inline is use for the print there in their where you write the code its like a magic keyword.

We load our dataset in the colab which help of pandas you easily see the process in the image.

```
#importing the data by mounting the google drive

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

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

```
#Loading the data set
path = "/content/drive/MyDrive/Capstone project/Copy of Airbnb NYC 2019.csv"
```

```
airbnb_data = pd.read_csv(path)
# Succesfully loaded
```

Exploratory Data Analysis and Visualization

After write the codes and doing manipulations and all we move on to the basic and high level data analysis and data visualization. We check the dataset then how much columns and rows we have in dataset. which rows have what columns like.

We load the dataset with their variable name

airbnb_data

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19
...	...	...	...	...	...	...	...	...	...	...	...	...	...
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	NaN
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	NaN
48892	36485431	Sunny Studio at Historical	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire	115	10	0	NaN

This is the big dataset live we have more rows and columns in the dataset. Then we check how many rows and columns we have with the help of the .shape method.

So we have total 48895 rows and 16 columns.

```
# checking the shape of the dataset

print('Shape of the dataset', airbnb_data.shape)
```

```
Shape of the dataset (48895, 16)
```

The data is big so not possible to see full data in one time. So we check top5 and last 5 rows and columns with help of .head() for top 5 row and columns and .tail() for last 5 rows and columns. The default value of this method is it takes 5 as default.

Top 5

```
airbnb_data.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19

Last 5

```
airbnb_data.tail()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853	-73.94995	Private room	70	2	0	NaN
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184	-73.93317	Private room	40	4	0	NaN
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475	-73.94867	Entire home/apt	115	10	0	NaN
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	Hell's Kitchen	40.75751	-73.99112	Shared room	55	1	0	NaN
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	Hell's Kitchen	40.76404	-73.98933	Private room	90	7	0	NaN

We also check the information of the dataset with the help of .info() method. Which shows you same how many columns and rows but they shows you those columns rows which have null or different type of data and also shows you memory informations.

Let’s see how it look like

```
airbnb_data.info()
```

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

The data have much null values in some columns and rows check and see it.

Isnull() is the method who gave you the rows and columns which have null values. We use Sum() method to calculate those values and print them and see clearly.

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

```
The number of missing values before cleaning the dataset are:
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
```

The column name have 16 null values and the column host_name have 21.

The two columns which have most null values last_review and review_per_month both have 10052 null values.

We remove those columns which have most null values with the help of drop() and replace those columns which have some null values with the help of fillna()

Now move on to the visualization part. How we visualize the data. We use different different charts in the visualization part to make data more visualize and more understandale.

Charts and graphs :-

We use bar charts, Heatmap, Histogram, and much more . Beacause different type of the data needs different type of charts and graphs.

.In this we find top 15 values counts of the column host_id . the method for find the top 15 value count is value_count().head(15)

This will show you the top 15 value count of this particular column.

```
top_host_id = airbnb_data['host_id'].value_counts().head(15)
```

```
top_host_id
```

```
219517861    327
107434423    232
30283594     121
137358866    103
16098958      96
12243051      96
61391963      91
22541573      87
200380610     65
7503643       52
1475015       52
120762452     50
2856748       49
205031545     49
190921808     47
Name: host_id, dtype: int64
```

```
from matplotlib import figure
# set the figure size for data visualizations plot using a bar chart
sns.set(rc={'figure.figsize' : (15,8)})
host_bar = top_host_id.plot(kind='bar')
host_bar.set_title('Host with the most listnings')
host_bar.set_xlabel('Host IDs')
host_bar.set_ylabel('Count of listing')
plt.plot(top_host_id)
plt.show()
```

After the above code we make the graphs of the data for better understanding. We write the code now we use matplotlib first. You can give your graphs title with the help of .set_title(----) same as you can set x and y asix like .set_xlabel(---) and .set_ylabel(---)

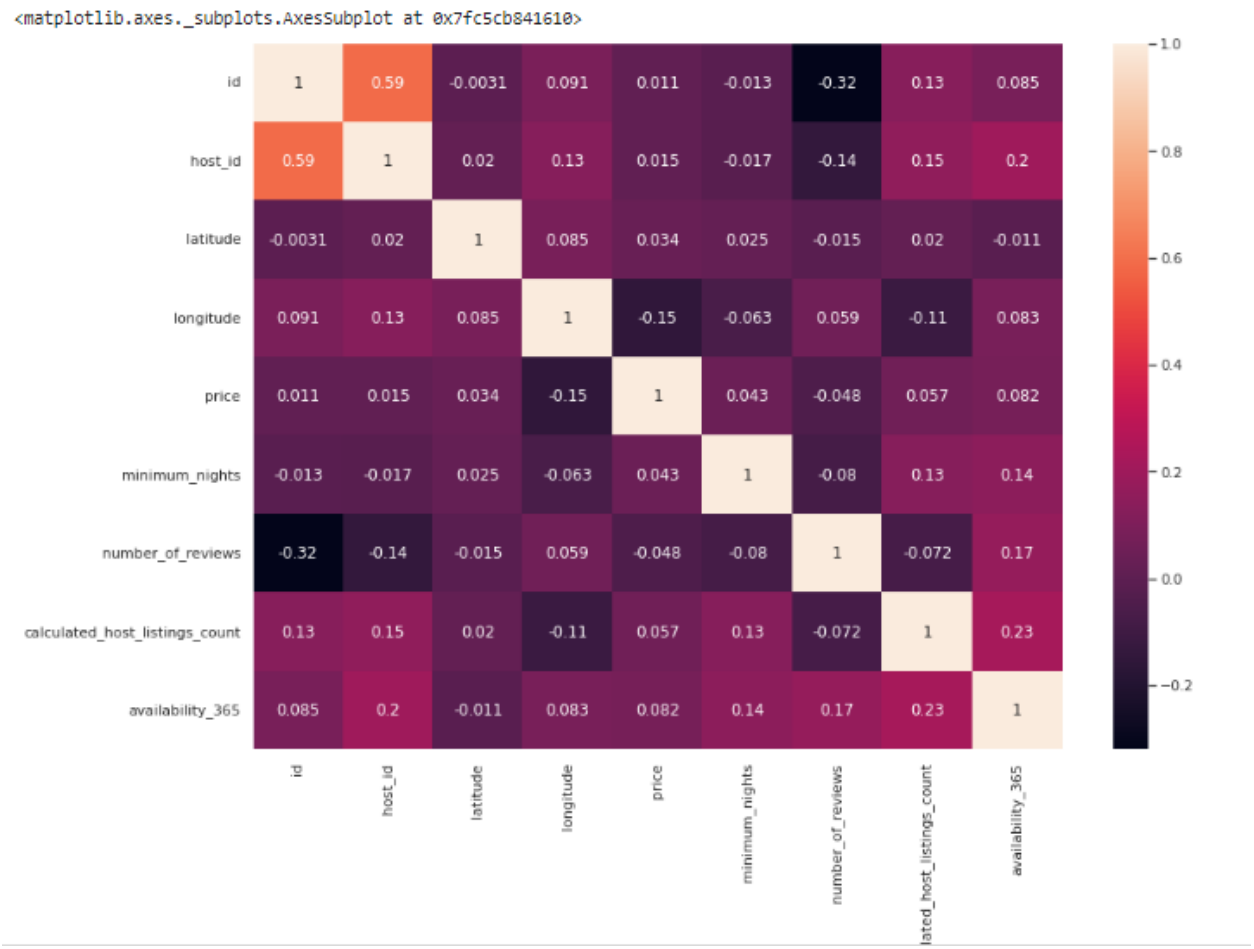
set figure (15,8) for better size.

This time we make correlation matrix same we can metioned the figure 14,10 and write some correct code and run this.

```
from numpy.lib.shape_base import column_stack
# Correlation matrix

plt.figure(figsize=(14,10))
sns.heatmap(airbnb_data.corr(),annot=True)
```

it’s look like this



Heatmap gives a correlation matrix to quantify and summarize the relationships between the variables

Now we check our numerical and categorical columns

```
numeric_features = airbnb_data.describe().columns

categorical_fetures = airbnb_data.describe(exclude=[int, float]).columns

print('Numeric Features:',list(numeric_features))
print('_'*160)
print('Categorical Features:',list(categorical_fetures))
```

```
Numeric Features: ['id', 'host_id', 'latitude', 'longitude', 'price', 'minimum_nights', 'number_of_reviews', 'calculated_host_listings_count', 'availability_365']
________________________________________________________________________________________________________________________________________________________________
Categorical Features: ['name', 'host_name', 'neighbourhood_group', 'neighbourhood', 'room_type']
```

In above the code we make two variable one for categorical_features and second is numeric_features and give their those columns which have numerical and categorical values.

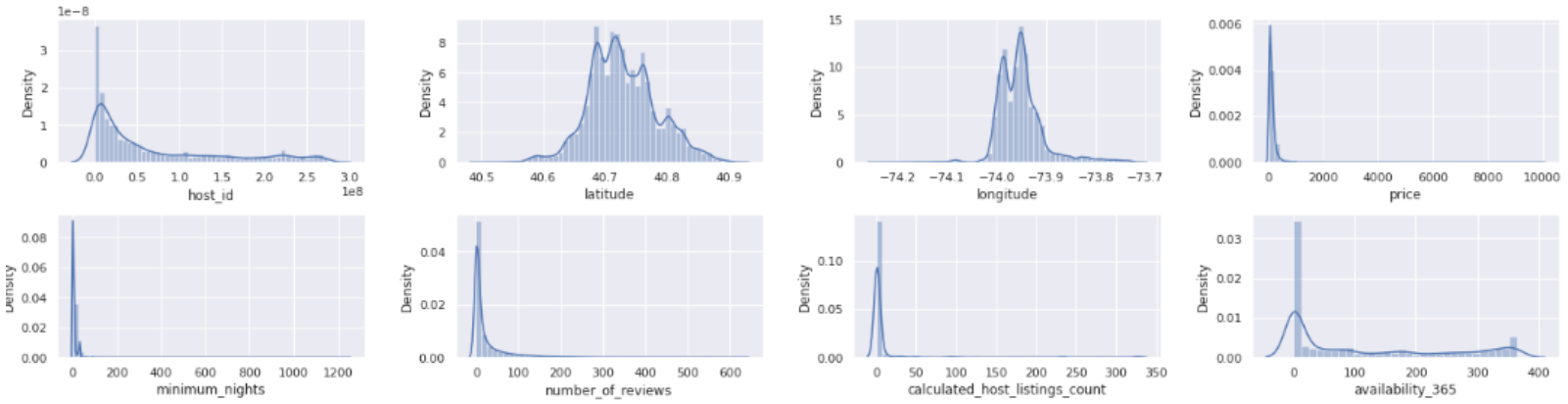
for presenting into the graph we write more codes for graphs. Same as give graphs size 20,5 asix names and much more functions.

```
names = numeric_features.values[1:]  #exclude id column.
ncols = 4
nrows = 2
fig, axes = plt.subplots(nrows,ncols, figsize=(20,5))
fig.tight_layout(h_pad=2, w_pad=4)

for name, ax in zip(names, axes.flatten()):
    sns.distplot(airbnb_data[name], ax=ax)

plt.show()
```

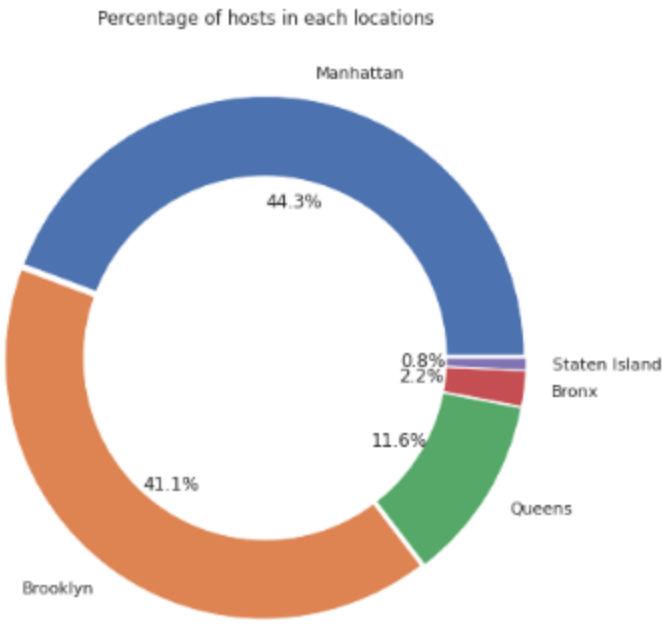
Here the result is



Our busiest host and areas:-

	area	host_count
2	Manhattan	21661
1	Brooklyn	20104
3	Queens	5666
0	Bronx	1091
4	Staten Island	373

Let’s see in graphs for better understanding



It can be observed that majority of the hosts are belong to the locations.

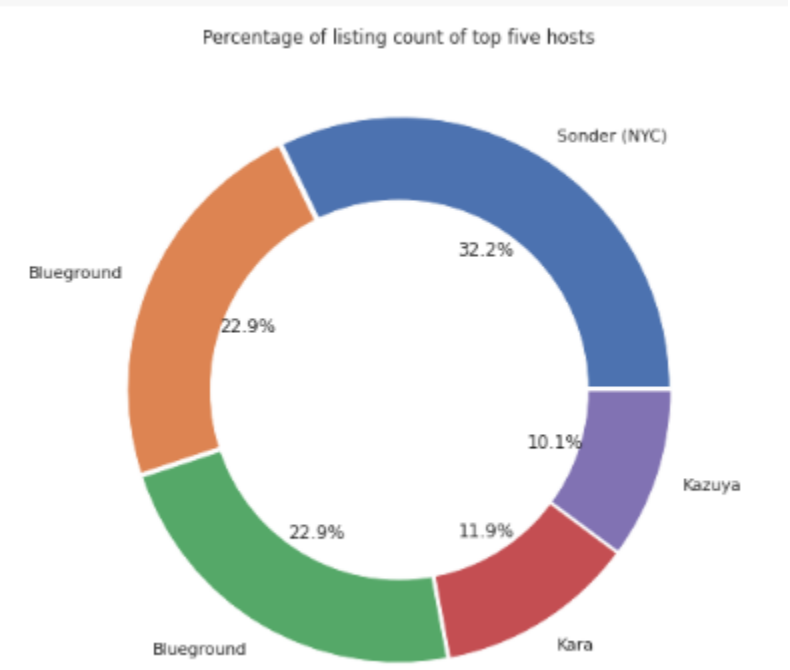
Manhattan and Brooklyn, hence these are the most popular destinations

It is clear that majority of the hosts are belong to the locations Manhattan and Brooklyn, hence these are the most popular destinations

So now we see who has the most listing and which neighborhood.

	host_name	neighbourhood_group	calculated_host_listings_count
13221	Sonder (NYC)	Manhattan	327
1833	Blueground	Brooklyn	232
1834	Blueground	Manhattan	232
7275	Kara	Manhattan	121
7478	Kazuya	Brooklyn	103

You also see this in the graph so you got better understanding about it.

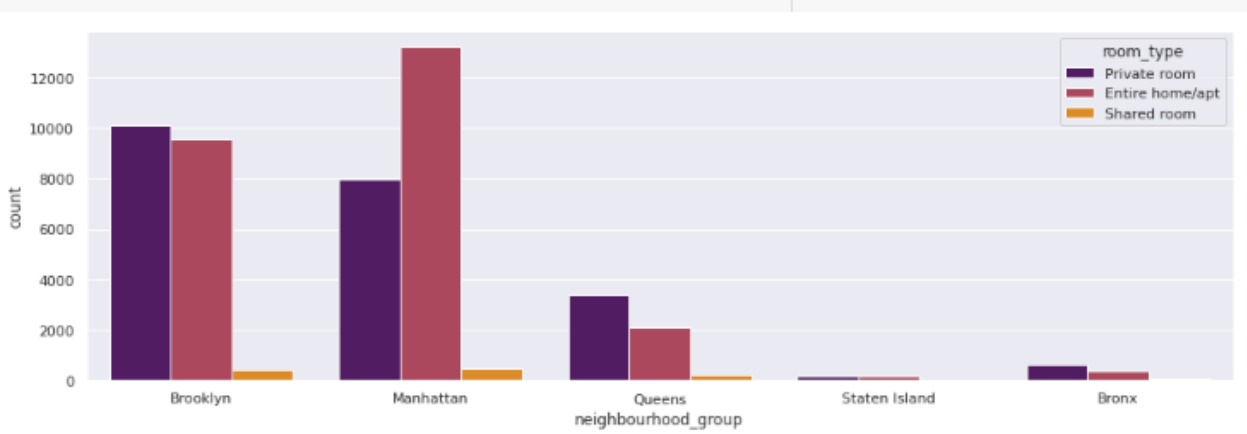


Room type is preferred in most popular neighborhood

```
fig = plt.figure(figsize=(16,5))
sns.countplot(data=airbnb_data, x='neighbourhood_group', hue='room_type', palette='inferno')
plt.show()
```

Set the figure 16,5 the make the graph with seaborn sns.countplot gives you the count plot graph and mention important details in it.

Let's see in the graph.



If people are looking for rooms in these areas of **Manhattan** and **Brooklyn** then hosts are providing either **Private room** or **Entire home/apt**.

Our unique busiest host:-

```
Michael          417
David            403
Sonder (NYC)     327
John             294
Alex             279
Blueground       232
Sarah            227
Daniel           226
Jessica          205
Maria            204
Name: host_name, dtype: int64
```

we can also perform this in the graphs after write the correct codes see below.

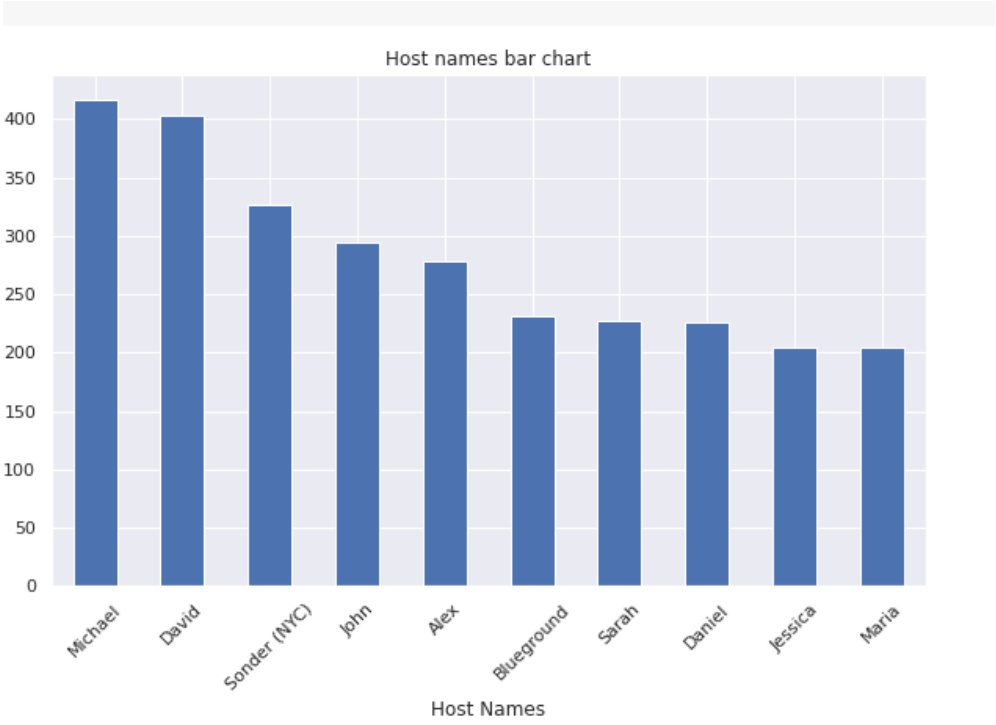
Give the size is figsize 10,6 then set labels and set styles and show it.

It will create a bar graphs which show you the details.

```
from matplotlib import style

fig = plt.figure(figsize= (10,6))
bar_hstName = unique_host_names.plot(kind="bar")
bar_hstName.set_title('Host names bar chart')
bar_hstName.set_xlabel('Host Names')
bar_hstName.set_xticklabels(bar_hstName.get_xticklabels(), rotation=45)
style.use('ggplot')

plt.show()
```



You will see the unique busiest host by looking their unique Ids.

Michale is 400 above, Alex is 250 above, and Maria is same as Jessica which is 200 above.

Top number of nights spent per room type

	neighbourhood_group	number_of_reviews
1	Brooklyn	486574
2	Manhattan	454569
3	Queens	156950
0	Bronx	28371
4	Staten Island	11541

	neighbourhood_group	number_of_reviews	availability_365
0	Brooklyn	486574	100.232292
1	Manhattan	454569	111.979410
2	Queens	156950	144.451818
3	Bronx	28371	165.758937
4	Staten Island	11541	199.678284

Manhattan and Brooklyn receiving the most number of reviews thus making the host busiest

Manhattan and Brooklyn have maximum number of reviews, so offering the most desired room types. Thus for these groups availability of rooms is less

Brooklyn has the maximum number of reviews 486574 and least availability of rooms by average is 100.23.

Inferences and conclusions :-

1. Starting with the dataset loading, we have performed data pre-processing, EDA
    a. .
2. Manhattan and Brooklyn are the two top most popular neighborhood groups in terms of hosts count, number of reviews, number of listing, maximum number of nights spends in these areas. So it might also be reason of traffic and high prices.

3. For other neighborhood groups namely Queens, Bronx and Staten island there aren't as popular as these two, especially on Staten Island.

4. The dataset can be further used for price prediction by building a linear model. The data needs to be treated of outliers and skewness for a linear regression as well as other models.

    a. This Airbnb(NYC 2019) Dataset For The Year 2019 Appeared to be v ery rich Dataset with a variety of columns that allowed us to do deep d ata exploration.

5. In the column name and host_name which have 16 and 21 null value only. Null values are present in last_review and reviews_per_month which can be dropped both have most null values is 10052.
6. From the dist. plots it can be observed that latitude and longitude data seem to be normally distributed and most of the numeric_features are positively skewed.
7. People stay for longer duration of time in private rooms in Brooklyn and Manhattan.
8. More customers preferred Manhattan location for night stay then Brooklyn.
9. Entire home/apt’room type has the highest number of listing of 52% and shared room is the least listed room type at only 2.4% in total.
10. 63.2% costumer spend night in entire home and 1.6% spend night in shared room.