Exploratory Data Analysis of Airbnb NYC Bookings

Rubina Fathima
Data science trainee,
AlmaBetter, Hyderabad

Abstract:

The Airbnb Data set of New York City is a collection of customers who make online hotel bookings.Neighbourhood in NYC rent their houses and apartments to make extra income.Tourists and people who are in need to stay in a home like environment opt for such bookings.

My analysis can help to understand what could be the reason for less reviews,high bookings.Price of the different room types and reviews of them.

Keywords: Data analysis, availability,reviews,room type,price.

Problem Statement

Data provided has customer information who make bookings of hotels.It has data spread in the form of rows and columns.

Rows are the various hosts and columns represent their features like calculated number of days stayed,reviews per month,name of the host and their span of stay.

Dataset

Airbnb dataset consists of
48895 rows,
16 columns.

Features of the dataset are explained as follows

Id: column which is unique to every observation

Name: gives name of the hotel of residency place for booking

Host_id: Unique id given to the host

Host_name:Name of host

Neighbourhood_group:There are places which are grouped together into neighborhood group

Neighborhood: Column has city names Latitude and Longitude: These describe the location of the hotel.

Room_type:This again has three categorizations like single room, shared and entire home for booking

Price: Charges for every room type

Minimum_nights:Number of days stayed.

Number_of_reviews:The reviews given by hosts after the stay.

Last_review:This gives date when the review was recorded

Reviews_per_month:This gives the average of reviews for every host.

Calculated_host_listings_count:column has count of hosts in the dataset.

Availability_365:Hotels or room types availability is given here.

Every column has a datatype to describe the content inside fields.Datatype of the columns is explained in the figure below.

Exploratory Data Analysis

Airbnb Dataset consists of hotel booking observations of hosts in NYC.
I have done some exploratory data analysis using python to gain insights of the dataset.

Null Values: Dataset has few columns whose fields are empty, filling such fields to perform analysis.

Room types of the hotel bookings dataset has Private room ,Shared room and Entire home/apartment.

```
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

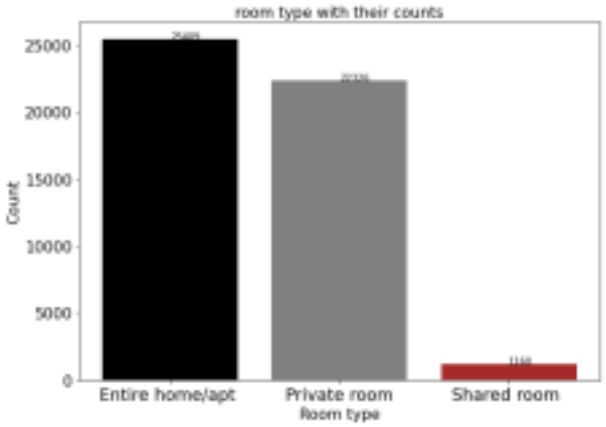

Figure:1

Observation-1:

Number of room types is given on the top of the bar graph to have a clear view of type and count of rooms in hotels.
Analysis shows shared rooms are the least of the bookings.
Entire room bookings of the dataset are very close to the Private room bookings.Hence the Shared room types are rendered with less bookings.

Reviews of each room type are very crucial to the online booking platforms.New users will start using hotels with high ratings from past customers.

Observation-2:

The hotels with high bookings sometimes face less average ratings.
This could be due to low maintenance of the hotels.

Figure:2

Top 5 Hotel Names who have high bookings but below average reviews.Among these
Hillside Hotel is the top and others are given in the figure.

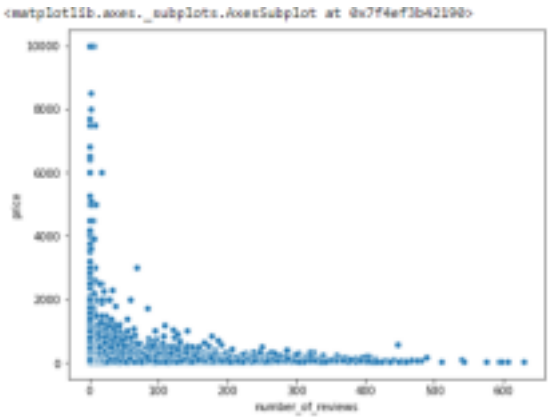

Figure:3

```
Hillside Hotel                                    18
Artsy Private BR in Fort Greene Cumberland        10
IN MINT CONDITION-STUDIOS EAST 44TH/UNITED NATIONS    6
A CLASSIC NYC NEIGHBORHOOD-EAST 86TH/5TH AVENUE    5
Are You Ready for Central Park State of Mind?     4
Modern, Cozy Apt Share for Young Professionals    4
❤ of Manhattan | Fantastic 1 Bedroom              4
Sonder | The Biltmore | Spacious 1BR + Kitchen    4
Sonder | The Biltmore | Stunning 1BR + Sofa Bed   4
Cozy apartment by Central Park                    3
Name: name, dtype: int64
```

Observation-3:

Scatterplot is used to show the distribution of reviews per month and price.
Here I have tried to find out the relation between price and reviews.
Figure shows that most of the reviews are awarded to the room prices which are low.

```
room_type
Entire home/apt    215.746181
Private room        86.903289
Shared room         78.978102
Name: price, dtype: float64
```

Average price of the room types is given above.

Observation-5:

Analysis over availability 365 column is Below observations states that
availability “0” has an average price of 136
As the availability increases the price increases of the room type.

```
availability_365
0    136.032111
1    129.509804
2    145.944444
3    136.486928
4    135.759657
5    134.179412
6    123.612245
7    135.123288
8    131.047210
9    146.445596
Name: price, dtype: float64
```

Observations-5:

Here I have explored the columns minimum nights and count of the nights the host has stayed.
This observation gives a glance that single day bookings are at high compared to more days bookings.

```
minimum_nights
1       12720
2       11696
3        7999
4        3303
5        3034
        ...
480         1
500         5
999         3
1000        1
1250        1
Name: number_of_reviews, Length: 109, dtype: int64
```

Observation-6:

Here I have tried to show correlation among features.
Reviews column and longitude have high correlation compared to other features.

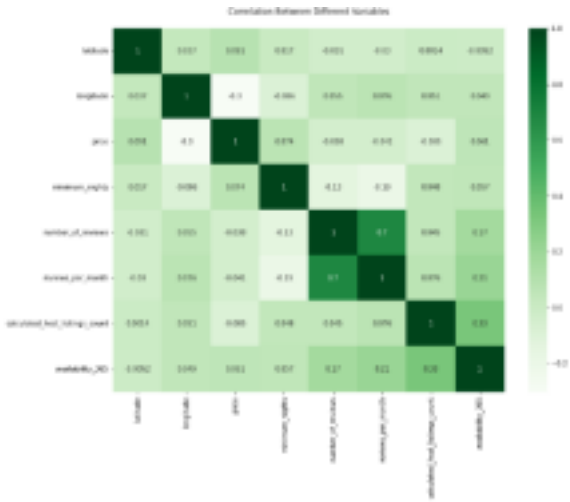

Figure-4

Observation-7:

Here I have shown the relationship between availability and price column.
High availability is shown in the neighborhood group brooklyn.
It conveys that the reason behind Brooklyn's availability could be that the place is less among tourist attractions.
Low number of visitors to brooklyn could also be the reason.
Other neighborhood groups like Manhattan, Queensland, States Island and Bronx.

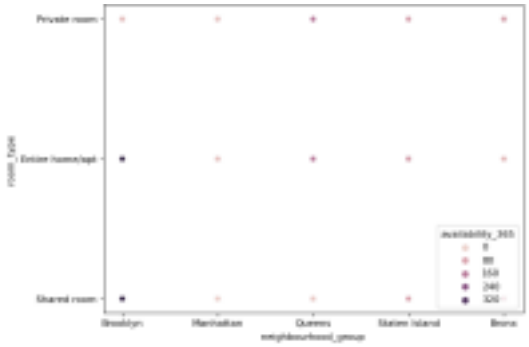

Figure-5

Conclusion:

After all the observations I have collected points to explain

- Airbnb is an online booking platform where the locals earn through giving their houses on rent.
- Reviews from customers give a lot of information on the room types and their price.
- Availability of the room type is more important because it has a direct impact on price.
- Reviews from single bookings and multiple bookings are different.
- Some neighborhood groups like Brooklyn show high availability of room types like entire homes and shared rooms

References:

1. Analytics Vidhya
2. Medium platform