EDA
Capstone Project
Airbnb Booking Analysis

Contents

Introduction

- Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other, Airbnb enables travelers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.

Dataset Overview

```
airbnb_df.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 given Airbnb dataset has total 16 columns.
- There are total 48895 entries.
- Data types are float64(3 columns),int64(7 columns) and object(6 columns).

By analysing datas and corresponding data types it is found that all datas are availiable in correct dtype except last review. Last review should be in datetime format but we are not going to use this column for our current EDA so lets leave it as object dtype.

Dataset Overview(continue..)

```
airbnb_df.describe()
```

	id	host_id	latitude	longitude	price	minimum_nights
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
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

- Average price for the room is 152.7 $
- On an average people stay 7 days in a room
- Maximum price of the room is 10000 $
- Mean review given to the room/apartment is 23.

Data Wrangling and data Cleaning

```
airbnb_df.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
```

- Name and host name has 12 null values
- Last review and reviews per month has 10052 null values.

There we have total 4 columns with null values but "last_review" has no significance for this current EDA project.The missing values present in "reviews_per_month" can be replaced by 0.

Now lets take a look at "availability_365" column, there we can see even though availability is zero, there are some corresponding values available in "reviews_per_month" column.So that means that airbnb reviewed some rooms having no guests yet.We can drop those values

Main Objectives

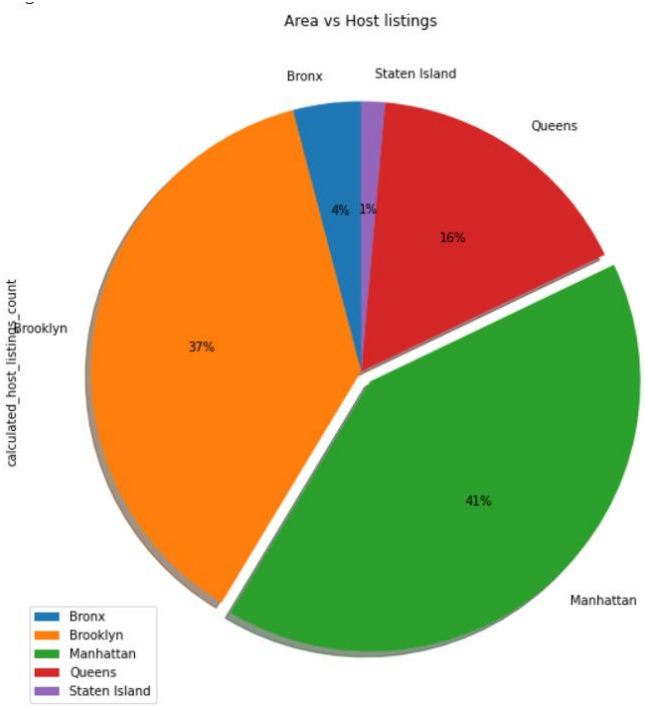
- What can we learn about different hosts and areas?
- Which hosts are the busiest and why?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Data Analysis And Data Visualisation

	host_name	neighbourhood_group	calculated_host_listings_count
0	Sonder (NYC)	Manhattan	327
1	Blueground	Brooklyn	232
2	Blueground	Manhattan	232
3	Kara	Manhattan	121
4	Kazuya	Brooklyn	103
5	Kazuya	Manhattan	103
6	Kazuya	Queens	103
7	Jeremy & Laura	Manhattan	96
8	Sonder	Manhattan	96
9	Corporate Housing	Manhattan	91

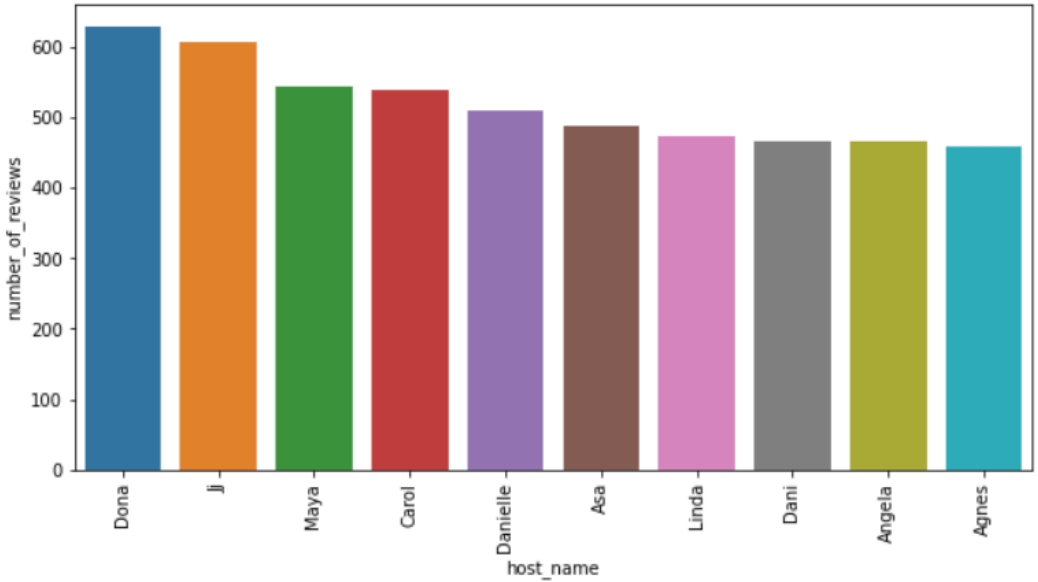
From the top 10 observations accordings to highest calculated_host_listings_count, we can found that 7 results are from Manhattan area,2 from Brooklyn and 1 from Queens.So it is clear that airbnb is a popular businesss model in manhattan.The host who has most host_listings are sonder(NYC).

Data Analysis And Data Visualisation(cont..)



Manhattan has maximum host listings which is 41% of the entire listings,then brooklyn has 37%,Queens has 16%,Bronx has 4% and at last staten island has only 1% listings.

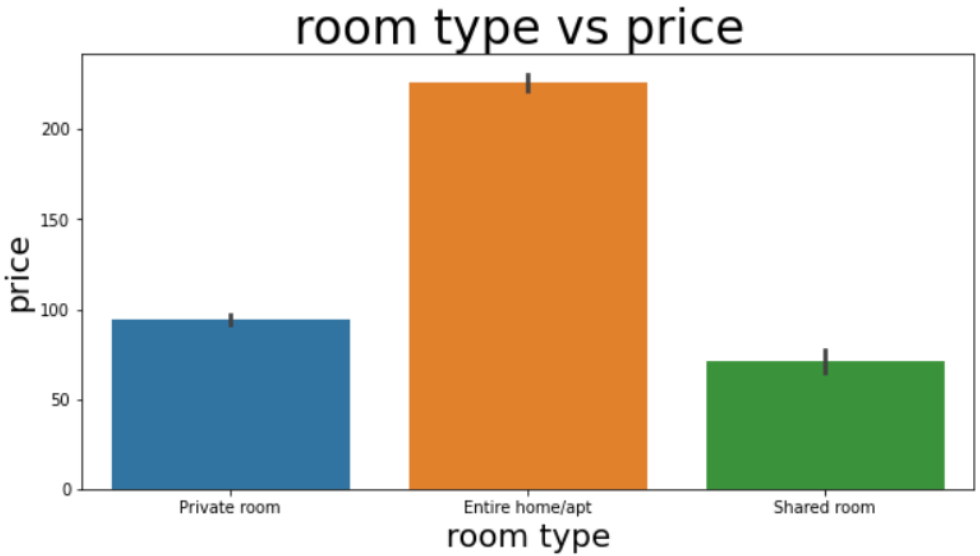
Data Analysis And Data Visualisation(cont..)



To find the busiest host we have to make an analysis on host_name versus number_of_reviews.The logic is very simple those hosts who has most number of reviews would be having most number of bookings.

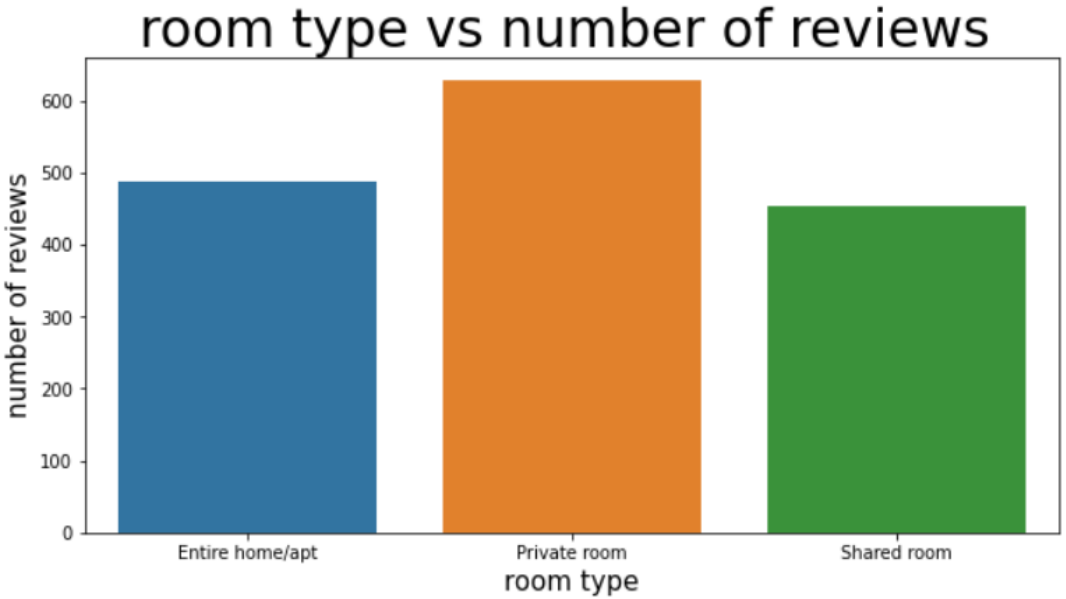
- From top 10 list of busiest hosts we can see that 8 results are from private room type.
- The name of the busiesthost is Dona having 629 review.

Data Analysis And Data Visualisation(cont..)



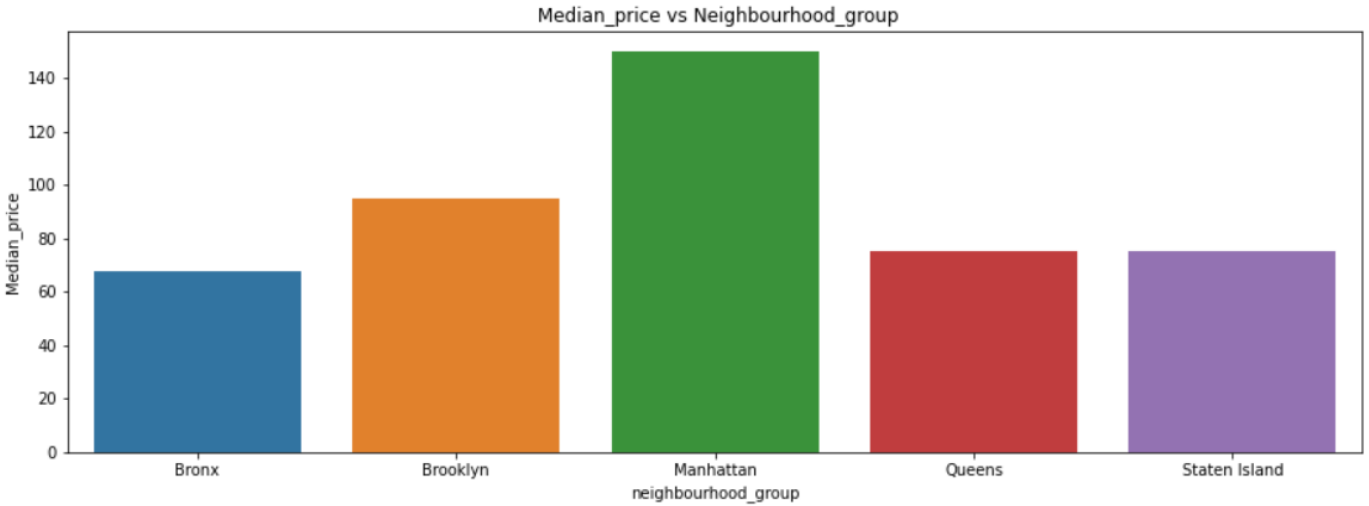
1. Shared rooms are very cheap as compared to Entire home/apt
2. Private rooms are little costlier than shared rooms
3. Entire home/apt is very expensive than all others.

Data Analysis And Data Visualisation(cont..)



Even though shared rooms are much cheaper as compared to private rooms most reviews are for private rooms which means that most number of people prefer private rooms over other room types.Probably couples or small group of peoples who need more privacy and prefer budget friendly stay will go for private rooms,so we can say that most of the airbnb customers are from this category.

Data Analysis And Data Visualisation(cont..)



'Manhattan' is the area having highest median price,followed by Brooklyn.So we can say that most costly hosts are situated in "manhattan".

Data Analysis And Data Visualisation(cont..)

Data Analysis And Data Visualisation(cont..)

Some Observations From graph of “Neighbourhood group vs availability”

1. Room availability is high in "Staten Island" ,the average value shows that rooms are available for 246 days in each year.

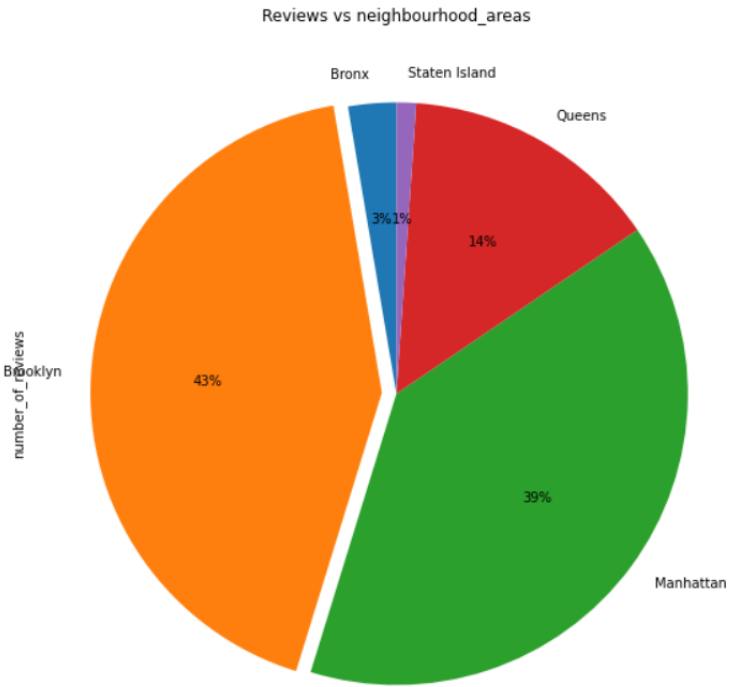
2. Room availability is very low in "Brooklyn", as per available data rooms are available only for 93 days in each year.

3. From the previous observations we found that maximum host listings are in "manhattan area" but room availability is very low as compared to other areas that means number of people visiting "manhattan" is more than available rooms. Which creates high demands for rooms and thats the reason for high price.

Data Analysis And Data Visualisation(cont..)

- Number of reviews are more at low price and reviews decreasing when price increases.
- From the scatterplot it is clear that most number of people prefer budge t friendly rooms.

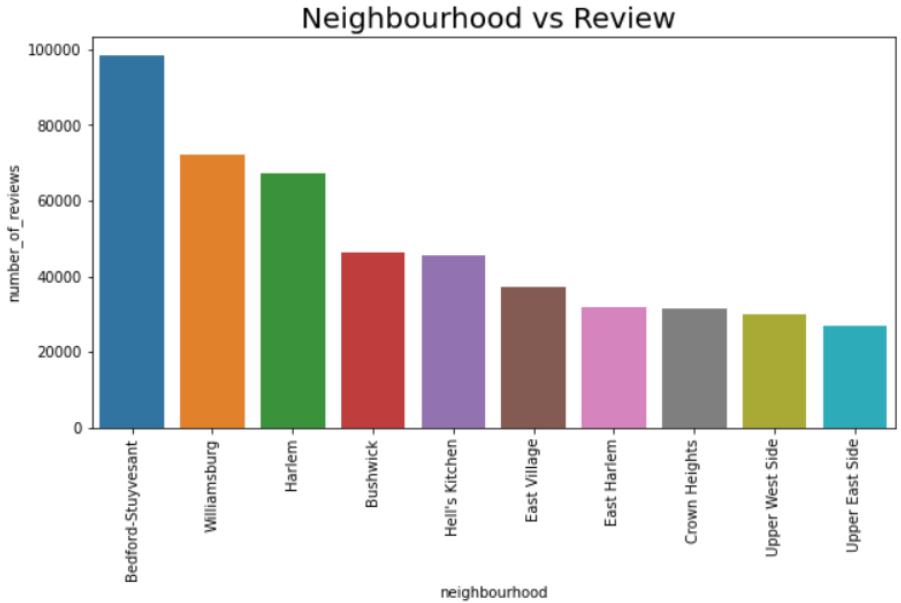
Data Analysis And Data Visualisation(cont..)



From the previous observations we seen that maximum host listings are in "manhattan area" and "brooklyn" but room availability is very low in both these areas that means number of people visiting these areas are high. To strengthen this assumption lets do some more analysis based on number of reviews in different neighbourhood_areas.

From the diagram it is clear that both "manhattan" and "brooklyn" has got most number of reviews which strengthen our assumption that these areas are having high traffic.

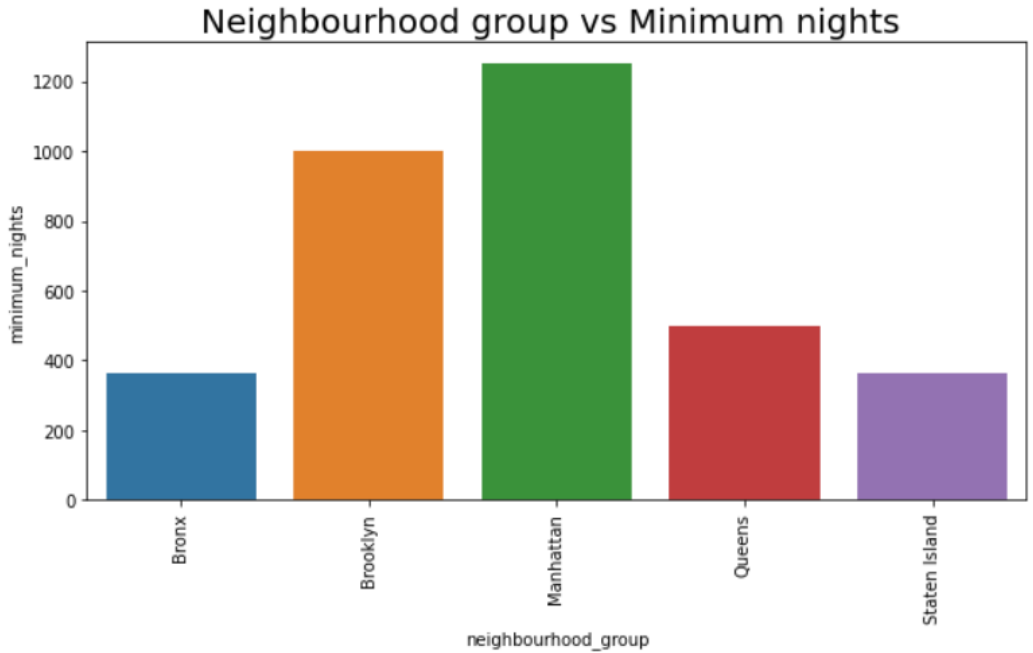
Data Analysis And Data Visualisation(cont..)



From this top 10 results it is also very clear that neighbourhoods belongs to "brooklyn and "manhattan" has got more reviews.

“Bedford-Stuyvesant” belongs to “Brooklyn” has got more reviews which means more traffic in this particular area.

Data Analysis And Data Visualisation(cont..)



From minimum nights data,it is clear that people like to spend more days in "manhattan" and "brooklyn".

From all these last three observations based on review,minimum nights and availability, we can say in “Manhattan” and “Brooklyn” has more traffic

Conclusions

1. From the top 10 observations accordings to highest calculated_host_listings_count, we can found that 7 results are from Manhattan area,2 from Brooklyn and 1 from Queens.So it is clear that airbnb is a popular businesss model in manhattan.The host who has most host_listings are sonder(NYC).

2. From top 10 list of busiest hosts we can see that 8 results are from private room type.

3. The name of the busiest host is Dona from queens area having 629 reviews.

4. Even though shared rooms are much cheaper as compared to private rooms most reviews are for private rooms which means that most number of people prefer private rooms over other room types.probably couples or small group of peoples who need more privacy and prefer budget friendly stay will go for private rooms,so we can say that most of the airbnb customers are from this category.

5. 'Manhattan' is the area having highest median price,followed by Brooklyn.So we can say that most costly hosts are situated in "manhattan".

Conclusions(cont..)

6. Room availability is high in "Staten Island" ,the average value shows that rooms are available for 246 days in each year.

7. Room availability is very low in "Brooklyn", as per available data rooms are available only for 93 days in each year.

8. From the previous observations we found that maximum host listings are in "manhattan area" but room availability is very low as compared to other areas that means number of people visiting "manhattan" is more than available rooms. Which creates high demands for rooms and thats the reason for high price.

9. From the scatterplot it is clear that most number of people prefer budget friendly rooms.

10. by analysing available datas of "number_of_reviews","minimum nights" we found "manhattan" and "brooklyn" are the high traffic areas.

Questions :

THANK YOU