Capstone Project-1

Airbnb Bookings Analysis

Shashwat Kumar Pani

W’s To Answer

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

Exploratory Data Analysis (EDA)

EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects.
These aspects include:

– main features of data
– variables and relationships that hold between them
– identifying which variables are important for our problem

Various EDA methods include:

- **Descriptive Statistics, which is a way of giving a brief overview of the dataset we are dealing with, including some measures and features of the sample.**
- **Grouping data [Basic grouping with *group by*]**
- **ANOVA, Analysis Of Variance, which is a computational method to divide variations in an observations set into different components.**
- **Correlation and correlation methods.**

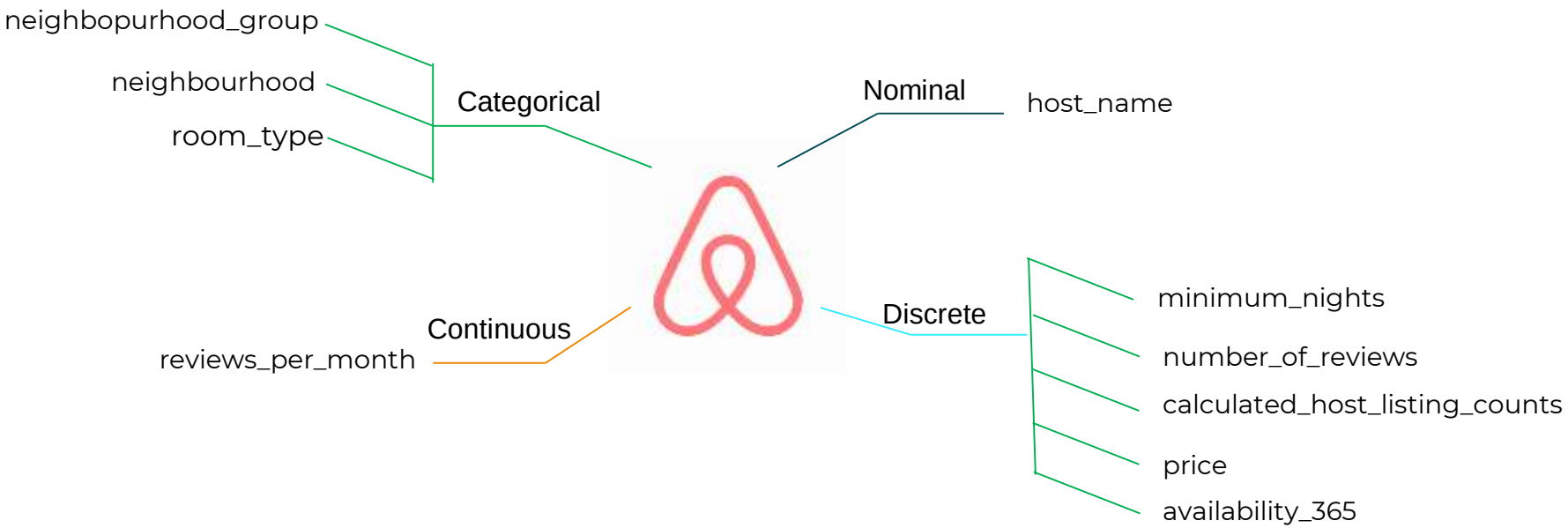
Data Pipeline

Data Processing 1: In this part we have removed all the unnecessary columns and took only the columns that helps with the scope of the project.

Data Processing 2: In this second part we replaced all null values by there appropriate central tendency measures.

EDA: In this part, we do the exploratory data analysis of the data processed from the above two steps.

Data Summary



Data Summary (Continued)

HOST_NAME: Name of the Host

NEIGHBOURHOOD_GROUP: Neighborhood Group in which the property is located. It is also a categorical variable and is divided among Brooklyn, Bronx, Manhattan, Queens, and Staten Island.

NEIGHBOURHOOD: Neighborhood name in which the property is located.

ROOM_TYPE: Types of rooms which are categorized into private, shared, and entire home/apt.

Data Summary (Continued)

REVIEWS_PER_MONTH: Average reviews per month.

CALCULATED_HOST_LISTINGS_CO UNT: Number of time property is listed.

AVAILABILITY_365: Number of days rooms available in a year

PRICE: Price of home per night.

MINIMUM_NIGHTS: Minimum nights guest stayed.

NUMBER_OF_REVIEWS: Count of customer reviews. last

Handling NULL Values

‘reviews_per_month’ has more than 20% Missing values which are replaced by Median Value.

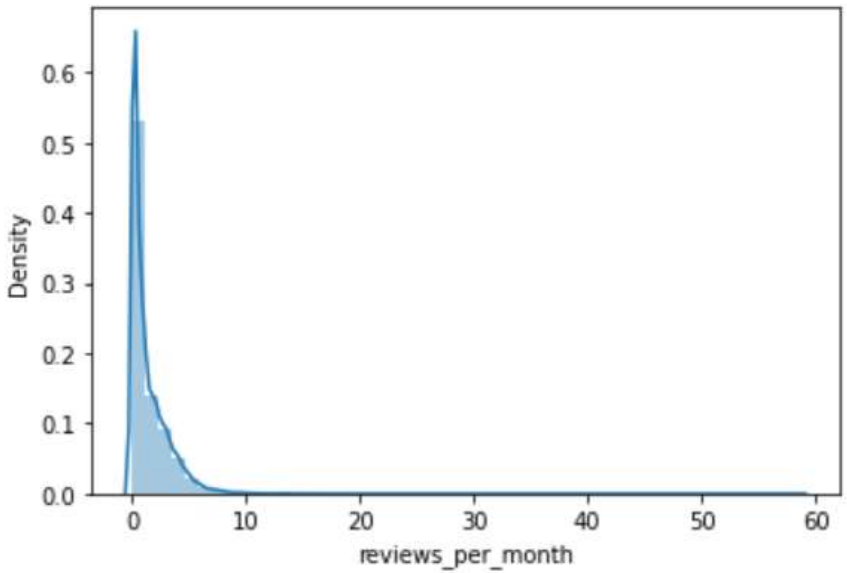
‘host_name’ has less than 4% of data missing, Such observations are dropped.

```
Missing Data Count
reviews_per_month    10052
host_name               21
dtype: int64
----------------------------
Missing Data Percentage
reviews_per_month    20.56
host_name             0.04
dtype: float64
```

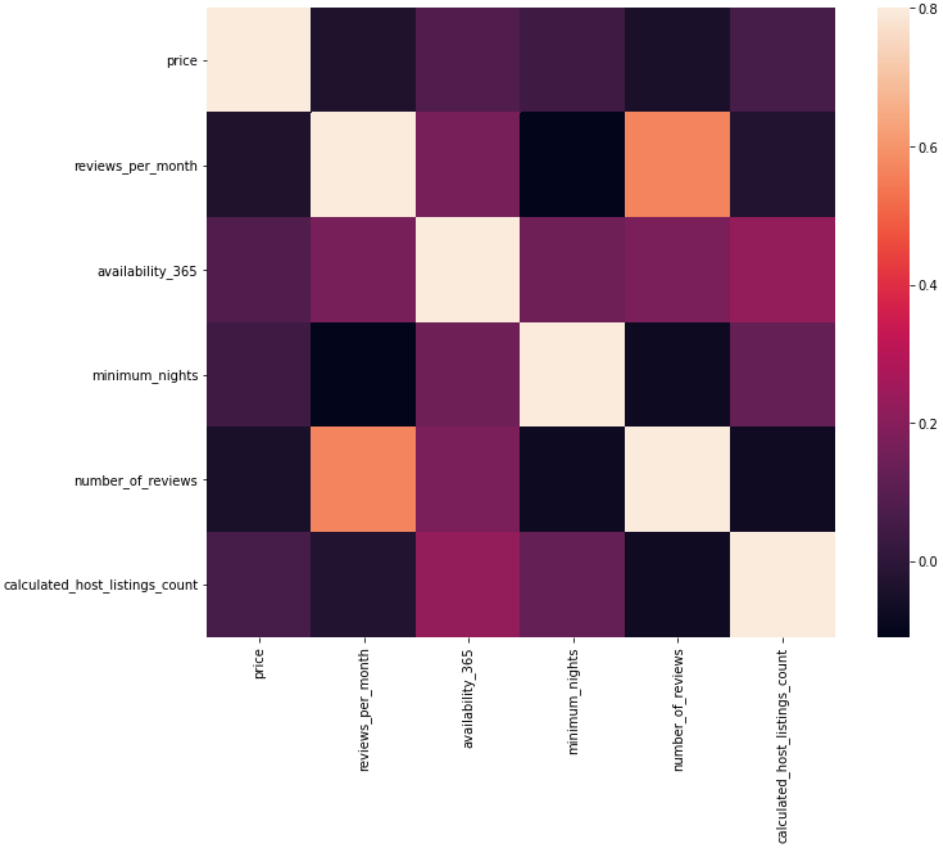
Measure of Central Tendency (Median)

- Positive/Right skewness.
- MEAN > MEDIAN > MOD

Missing discrete numerical data type can be replaced by median as the best option among measure of central tendency.

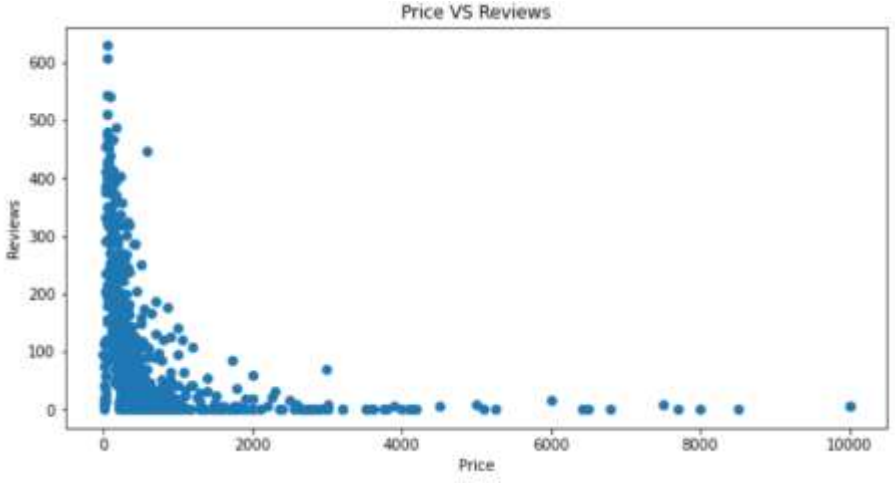


HEATMAP

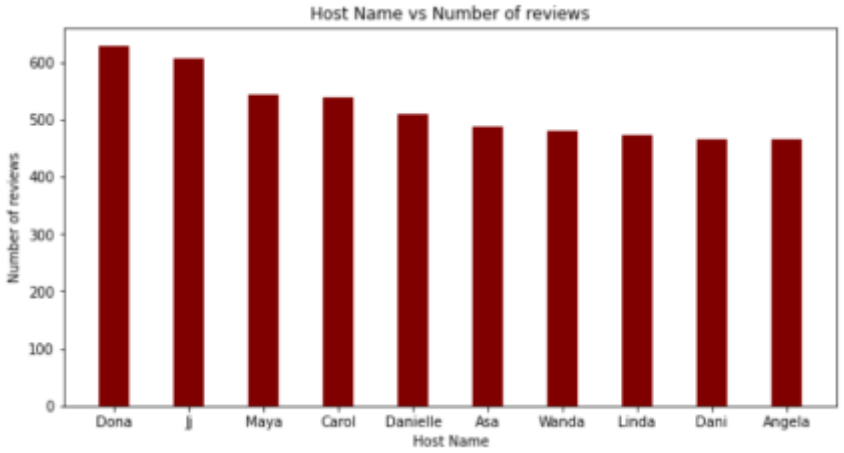

price
reviews_per_month
availability_365
minimum_nights
number_of_reviews
calculated_host_listings_count
0.8
0.6
0.4
0.2
0.0

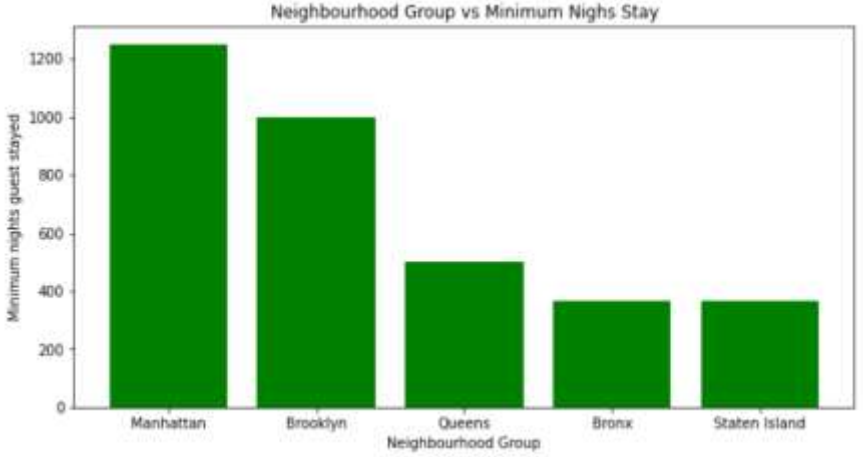
EDA


Neighbourhood Group VS Price
price
10000
8000
6000
4000
2000
0
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Neighbour_group

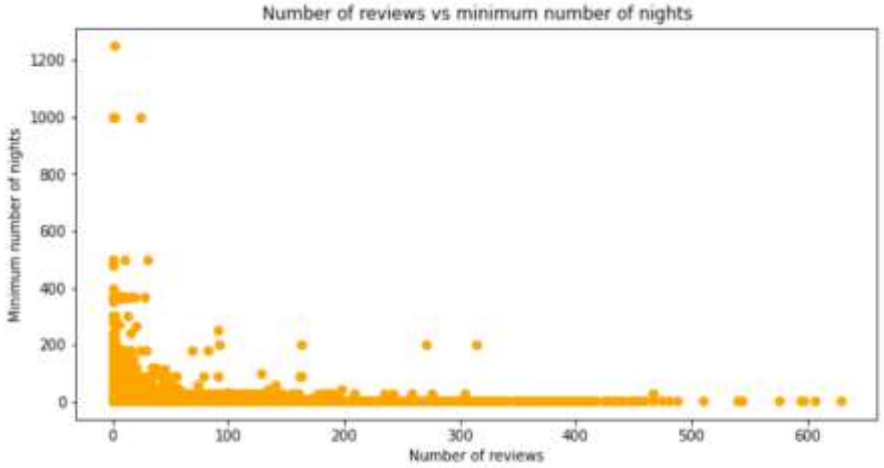

Price VS Reviews
Reviews
600
500
400
300
200
100
0
0
2000
4000
6000
8000
10000
Price

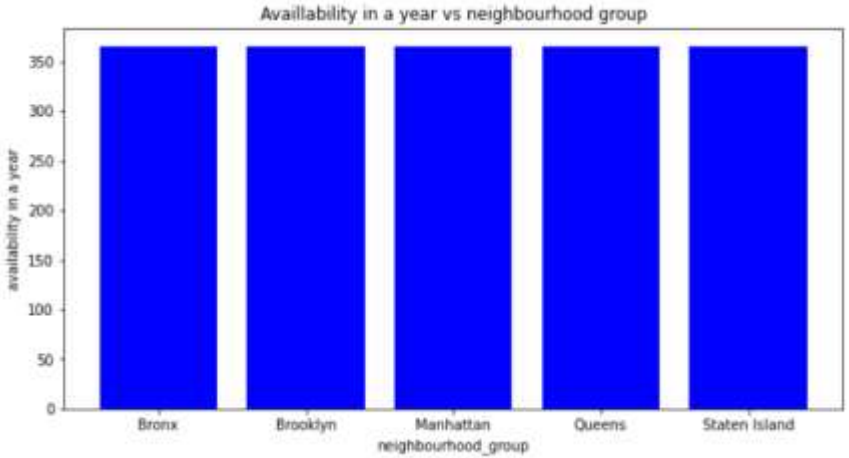
EDA (Continued)



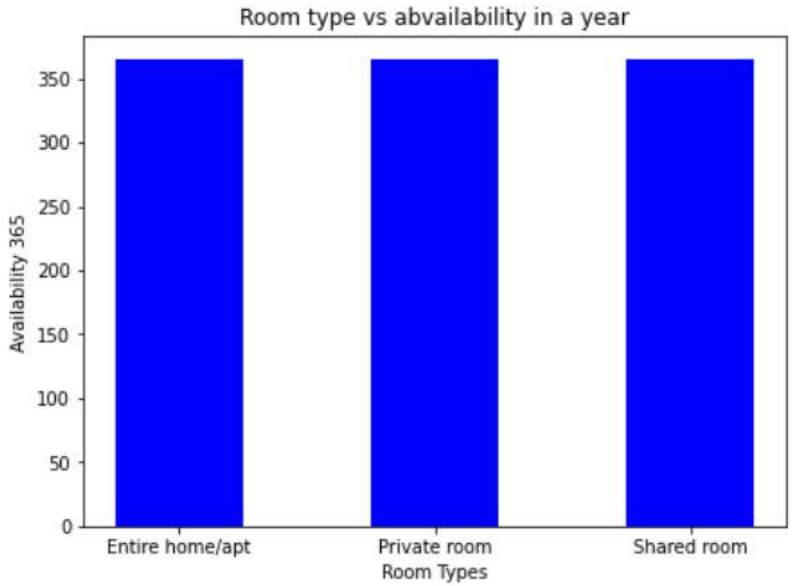


EDA (Continued)





EDA (Continued)



Conclusion

- Most guests are preferring cheaper stays than the expensive ones.
- Guests can book stays in any neighborhood group, as rentals are available 365days a year at all places.
- Brooklyn, Manhattan and Queens are most expensive neighborhood group options and, Bronx and Staten island are some of the relatively cheaper options.

Conclusion (Continued)

- Brooklyn and Manhattan are the busiest places with large number of guests staying for multiple years because of more preference to private rooms and apartments in these areas and cheap options are also available.

- The demand for private room, entire room and apartment is maximum.

- All types of rooms are available round the clock irrespective of neighborhood group.

Challenges

- Huge chunks of data was to be handled keeping in mind not to miss anything which is even of little relevance.
- Finding the relationships between various features and drawing conclusions.
- Computation time.

Q & A