Airbnb Booking Analysis

Nikunj Sonule
EDA Capstone Project
Almabetter, Bangalore

Abstract:

Airbnb data analytics is a qualitative and quantitative processes and techniques used to enhance productivity, marketing strategies, occupancy rates, and yield. The data is first extracted and then categorized to identify and analyze behavioral data and patterns; the techniques used can vary depending on the business requirements. Analyzing booking patterns shows the demand trends which you can use to implement dynamic pricing.

I've performed project right from proper data cleaning to analysis a problem statement. To understand more better I've accomplished project in General and Business Analysis. The task of analysis has helped a lot to understand the technical growth of a company like Airbnb industry. Working on different view of analysis to make it more reliable which is been provided in this technical documentation.

Problem Statement :

I come up with Airbnb Company industry project to work on Airbnb Booking Analysis. The Airbnb company has provided with the past data to come up with the insightful business strategy. The major target of the project is to provide data analysis techniques to draw out key features of the market from the datasets. The initial task is to understand the dataset and discuss the problem. Looking towards the factor how region, room type, price, location etc. can helpful to the problem. The next task is to data analysis parts based on bookings on different view, market segment and many more to draw a proper conclusions and results.

Data Summary :

The dataset provided was cleaned but we found some of the missing and null values in the column. By using info( ) method and fillna() to replace it to '0' with "airbnb_copy.info()" and "Airbnb_copy.fillna()" which shows total 16 data columns. We found that –

1. The dataset has a shape of (48895, 16) and for proper analysis I've taken 48895 columns and 16 rows.
2. Finally, we have dtypes: float64(3), int64(7), object(6)

Following are the columns from the dataset –

- ‘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’

Introduction :

The objective of the project is to make analysis on the Airbnb booking status of the customer of various feature like price variation, booking changes, availability_365, type of the neighbourhood and more. The datasets have data related to region based on neighbourhood group – Brooklyn, Manhattan, Queens, Staten Island and Bronx. The Exploratory Data Analysis is done to get the insight about the data and find out the solutions of the problems. In General and business analysis focusing on the problem with business related strategy. Through this I found analysis related to property like property owned by each neighbourhood groups and distribution on room type.

By taking this issue I also took analysis to it like price exploration (in what price people prefer more). Staten Island is quite away from all those groups and its not much prefer by people too. Airbnb provide with room type – private, entire home/apt and shared room. It’s nearly 2.4 % people like shared room rather private and entire home/apt are on more scale demand in guest. I have work with correlation of data were we can understand each relation of data with each other. These can help us a lot while dealing with certain numerical data.

Steps Involved :

1. **Understanding the data i.e dataset :**
   Before get started with the project I firstly look up at dataset. The dataset was pretty cleaned. The column doesn't have any list and dictionary values. By looking to data I came up with many creative ideas that how I can deal with those column.

2. **Looking up for problem statement :**
   After analyzing the datasets and approaching to every single problem to overcome it. I decided to divide my task and initialized with problem statement. The problem statement were based on target variable we took for analysis. Approaching to a problem with both hosts and area perspective I manage and initialized the problem statement.

3. **Data Cleaning :**
   The next task was data cleaning which was easy with this dataset. As mentioned in above points the data were float64 dtype , int64 dtype, object dtype. Some of the column like 'reviews_per_month' was having null values. So, I decided to keep fill those data space with '0'.

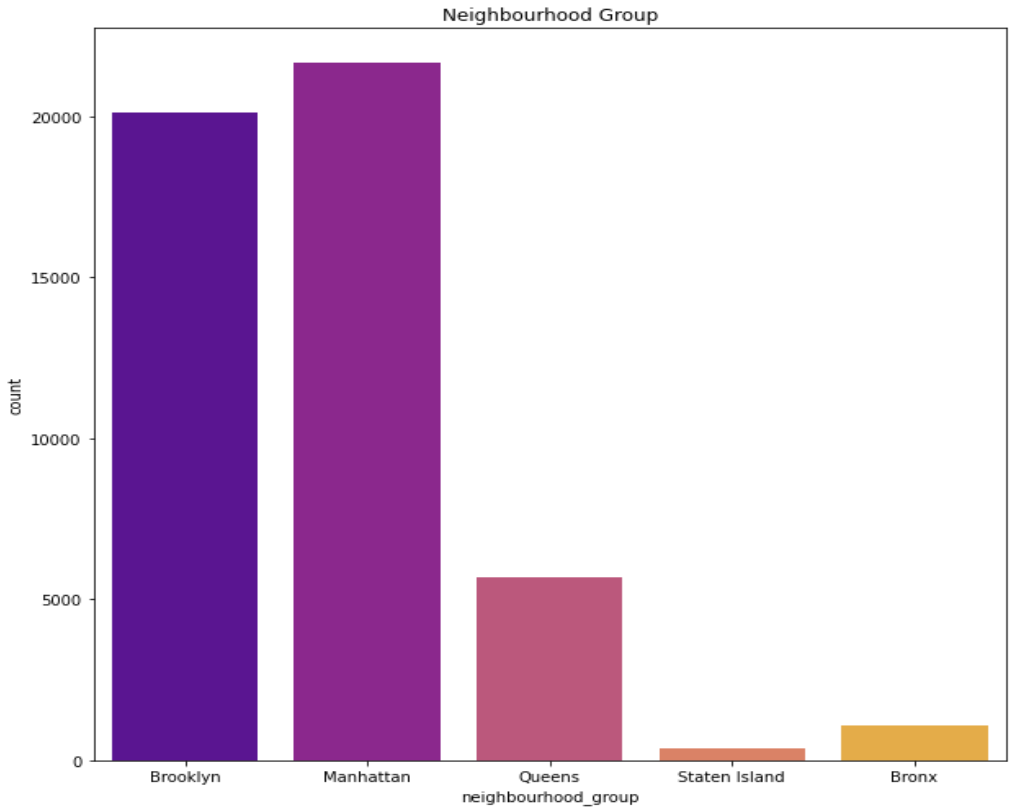
4. **Exploratory Data Analysis :**
   After data cleaning it was sure to target some important columns for Exploratory Data Analysis. Relating data with business is important aspect relationship for any problem. Matching the data with correct suitable problem by python libraries to result some insightful visualization was great task. These also gives us a more information from different charts and graphs.

5. **Visualization of analysis :**
   The EDA parts make more clear about data in a picture and graphical form. Mainly I perform matplotlib and seaborn libraries of python for the data analysis. The libraries helps a lot with bar charts, pie charts, heatmap, scatter plot, box plot. This shows how past industry data can be useful to grow a company like Airbnb and scale up the market segment.

6. **Results and Conclusions :**
   At last I conclude all the analysis with perfect market strategy. I've kept my analysis simple, creative and more problem statement from insightful data. Throughout the project it also helped a lot to understand Airbnb Booking analysis.

General and Business Analysis by Nikunj Sonule

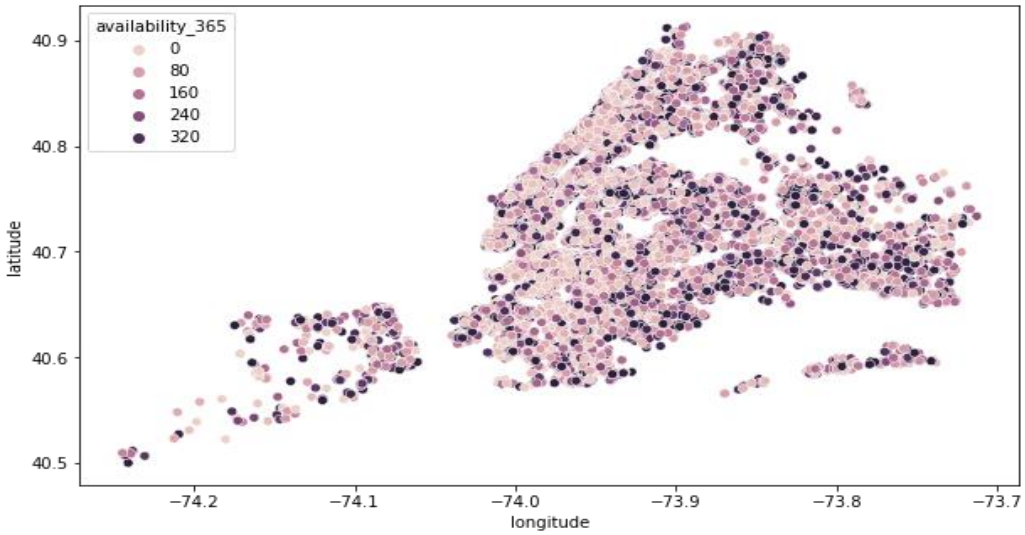
1. Neighbourhoods groups with among most highest count



Insights from analysis :- The above bar chart shows all neighbourhood groups with Manhattan highest number of count. Following this Brookyln is second and others are Queens, Bronx and Staten Island. If we look at the number count, we can observe Manhattan cross maximum with 20k + and Brookyln with exact 20k. Coming to other part groups Queens with 5k + , Bronx with nearly 1.7k and Staten Island with nearly 800.

By looking towards market and business strategy we can say Manhattan and Brooklyn can be more attention to customer. But besides this Queens, Bronx and Staten Island also has to keep on more ideas and business tactic to built an increment in this regions.

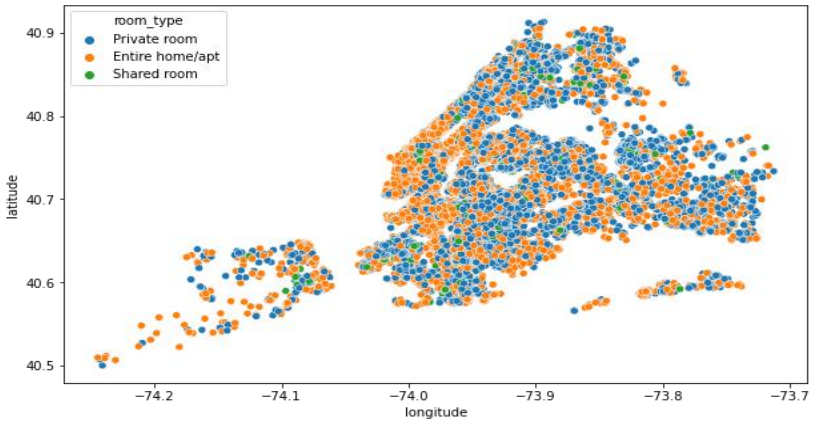
2. Availability of number of days for room when listing is available for bookings.



Insights from analysis:- According to availability_365 we can't see any pattern to understand this scatter plot. Definitely it gives a clear idea of available for bookings according to latitude and longitude.

Availability of number of days for room is quite vary in latitude and longitude. The more numb er of distribution can be observe in latitude above 40.50 to 40.91 and longitude ranging from -74.16 till -73.88.

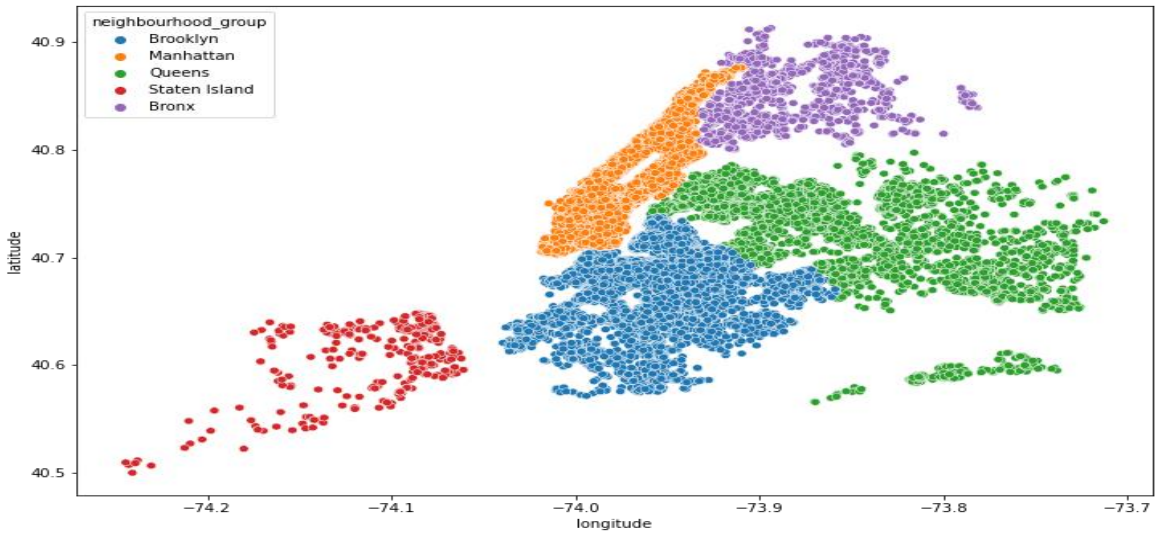
3. Analysis based on room type.



Insights from analysis:- As per the dataset, we can see three types of room. All through the neighbourhood group based on the scatter plot we can see very less number of 'shared room' but 'private room' and 'entire home/apt' are more occupied by people. The guest prefer most as a private room and entire home/apt in all groups. Private rooms are great when you prefer a little privacy, yet still value a local connection. Shared rooms are for when you don't mind sharing a space with others. When you book a shared room, you'll be sleeping in a space that is shared with others, and you'll share the entire space with other people.

Hosts on Airbnb offer a wide variety of spaces, ranging from shared rooms to private islands. By analysis we can say guest and hosts won't happy and prefer with 'shared room'. The more we can do is to make fascinating accommodation and can make discounts with varies business tactic to overcome and many more to do. Shared rooms are popular among flexible travellers looking for new friends and budget-friendly stays.

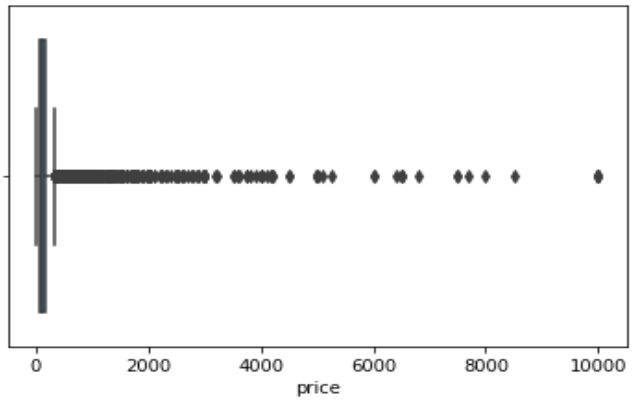
4. Analysis of neighbourhood groups based on latitude and longitude.



Insights from analysis:- Based on the above analysis we can observe different types of 'neighbourhood group' by measuring latitude and longitude. Obviously Staten Island is quite away from all those groups. Airbnb is one of the most prestigious company with customer hotel host and guest.

Airbnb business model is multi-sided marketplace that connects travelers with host and experience providers. It's important to keep data with related to based on latitude and longitude as it shows the host and guest distribution on property way were certain business tactic can be useful when problem arises.

5. Analysis on price exploration.

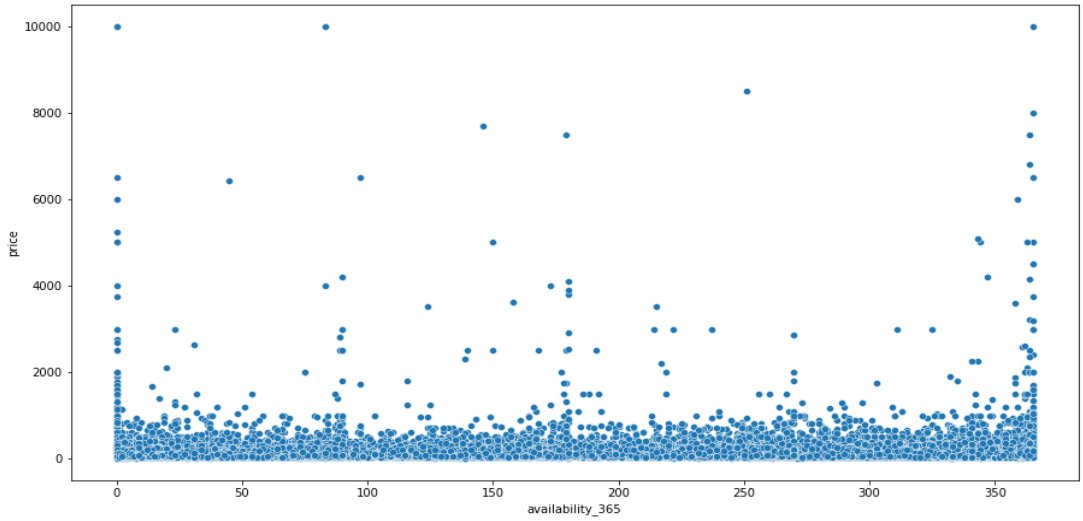


Insights from analysis:- According to data in the dataset, we don't have null values for 'price' column. Every price status is been vary for every room type. The guest and host prefer different accommodation based on room type.

6. Price based on the availability 365.

Insights from analysis :- The question arises is that, "What's The Average Price Of An Airbnb?" Deciding what to charge for your Airbnb is a decision that will have a massive impact on how successful you are in the short-term rental market.

Many fall into the trap of charging too high a price and find the result is below average occupancy rates. On the other end of the scale, many Airbnb hosts leave a lot of money on the table charging a lower price than they could have, particularly during peak season or around dates of popular



local events. There are many other variables to take into account as well; what type of property do you have? what city and neighbourhood is your listing in? how many reviews does your listing have? These and other questions need answering before you can find that perfect price.

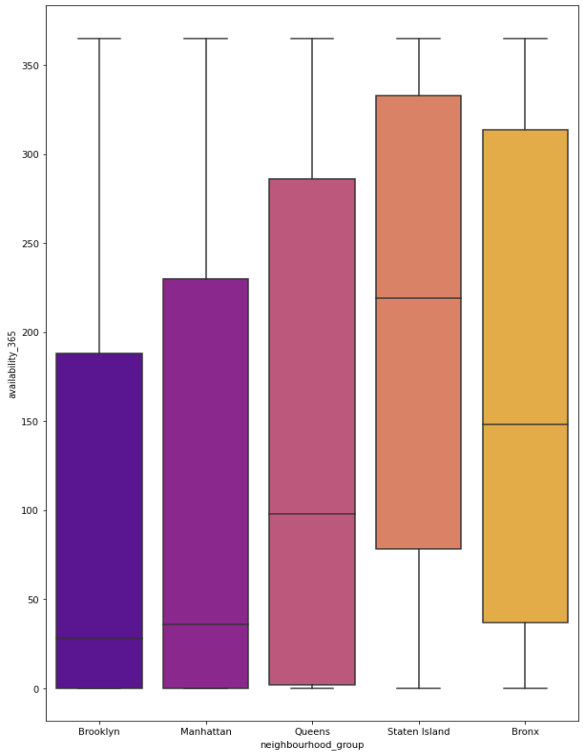
I've used Airbnb database to analyse some factors which can affect –

- Average Airbnb prices by location
- Average Airbnb prices by listing factors (like room count and arrangement type)

Now this are the conditions which can be used to known based on availability365. The prices are ranging from 0 – 10k. There is any price increase based on the availability and looking below the plot its hardly to infer but looks like with availability with 365 the price increases to 10k. A lot of distributions can be seen by scatter plot but guest and host prefer all kind of price tags facility.

7. Relation between neighbourhood group and availability of number of days for room.

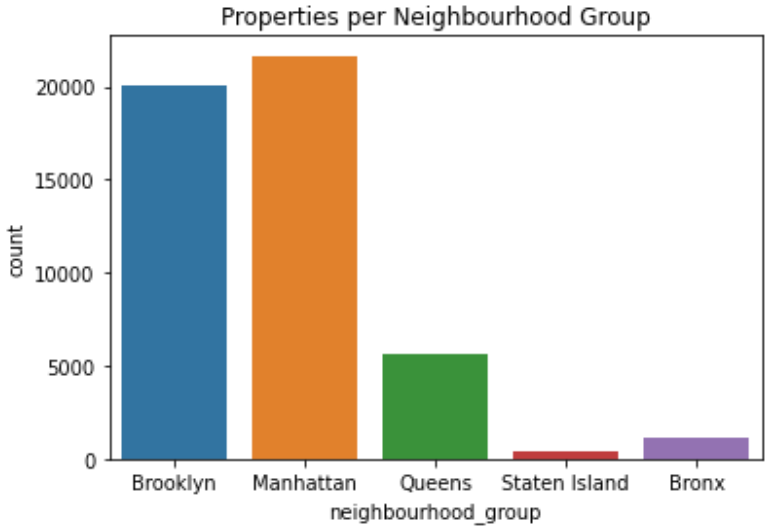
Insights from analysis :- The image above is a boxplot. A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3) and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped and if and how your data is skewed.



- For Brooklyn the plot varies between minimum o to maximum 365. The Q1 and Q3 for that is 0 - 185 where median is 25.
- For Manhattan the plot varies between minimum o to maximum 365. The Q1 and Q3 for that is 0 - 230 where median is 30.
- For Queens the plot varies between minimum o to maximum 365. The Q1 and Q3 for that is 2 - 280 where median is 95.
- For Staten Island the plot varies between minimum 0 to maximum 365. The Q1 and Q3 for that is 75-325 where median is 220.
- For Bronx the plot varies between minimum o to maximum 365. The Q1 and Q3 for that is 30-312 where median is 150.

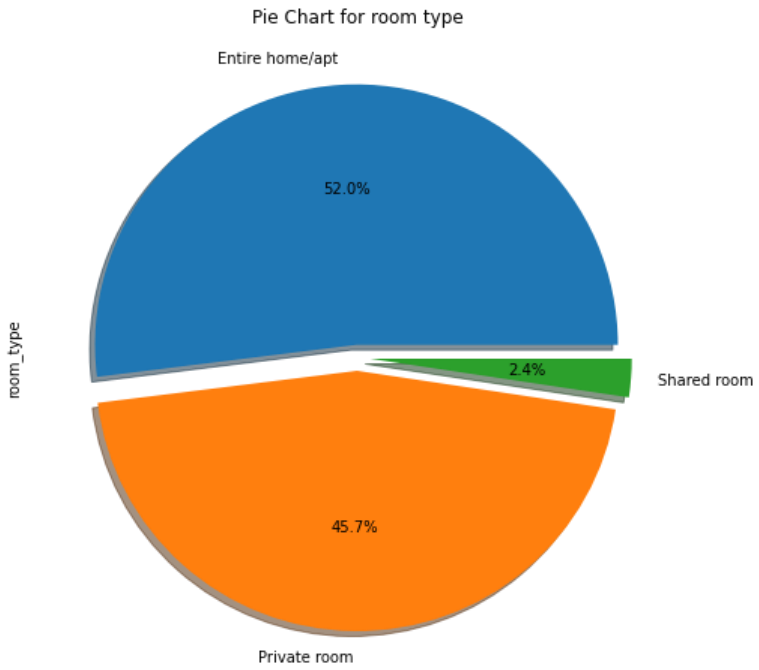
Availability 365 and neighbourhood group shows how the number of days were available for bookings in different five types region. The most minimum available is for Brooklyn and Manhattan which is Q1 of 25% shows of zero value but for Brooklyn Q3 is nearly 185 and for Manhattan Q3 is nearly 230. It also shows vary in median too.

8. Property owned by each neighbourhood groups.

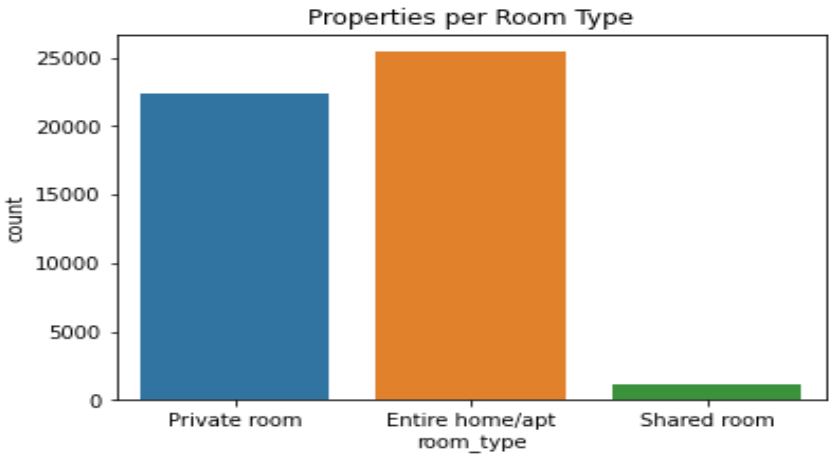


Insights from analysis:- Manhattan has the highest property preferable among the five different neighbourhood groups. Manhattan has the highest above 20k count properties in the 'neighbourhood_group' followed by Brooklyn with exactly 20k.

9. Property distribution on room type.



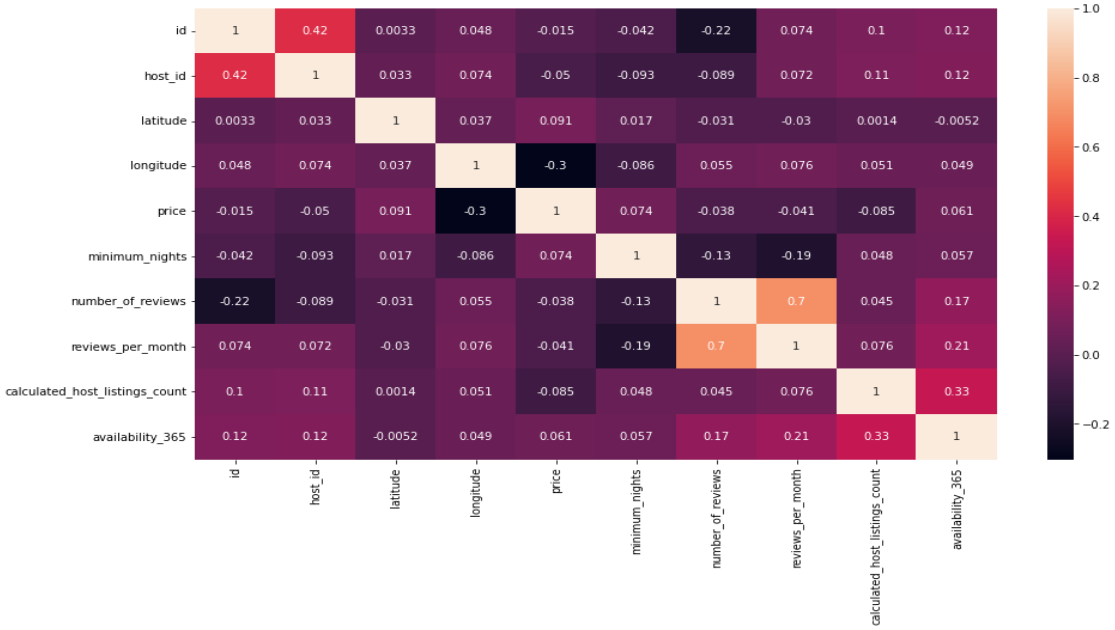
Insights from analysis :- The above pie chart signifies a very clear analysis by data of Airbnb room type. The most used and flexible accommodation is Entire home/apt with 52.0% followed by Private room with 45.7%. Throughout the data collected and used through dataset shows clarity among all room types.
The Shared room which is the least count per cent in room type with 2.4%. Maybe guest and host doesn't want to share their room and private space with other stranger. But some are likely to afford shared room also.



Insights from analysis :- The above bar graph also describes the similar thing as discussed above in the pie chart. The count for all the room types are totally different with their individual perspectives.

As per the properties per room type the count for 'Entire home/apt' goes on up scale with 25k followed by 'Private room' with 20k + count. The 'Shared room' also shows least count with below 5k.

10. Correlation of data.



Firstly, I'm doing a correlation analysis to understand the level of correlation of other int/float variables. Heatmap can be super useful when you want to see which intersections of the categorical values have higher concentration of the data compared to the others.

- The bright white color is highly positive correlated with 1.0 value
- The dark black color shows the highly negative correlated with -0.2 value. Here the values of the variables is going in one direction and corresponding value of other variable is going in another direction (let say in opposite direction). The value of one variable can change with respect to value of another.
- We can see the correlation of data, stated that number_of_reviews and reviews_per_month are highly correlated with 0.7 value. If we want, we can drop one of feature to help for further analysis and prediction.
- The numeric value of 'price' and 'longitude' also shows very important visualization. The value is -0.3 which is negatively correlated.

- It definitely shows ‘id’ and ‘host_id’ are not useful for further use cases (in condition if require).

The Airbnb data is quite enough to visualize as to display a more generalized view of numeric values.

Concluding Statement:-

At the end, we can clearly see that the Airbnb industry is getting through a lot of challenges to incline their company. We solved the problem based on business and host aspects and factors which can transform Airbnb to scale better. So, using the past data we made some insights which would help the management and manager to bring out some better changes to understand the market strategy.