Exploratory Data Analysis on Airbnb Booking Data set

Amani chivilkar, Abdul Rahaman,

Shivam, Vinay Kumar

Data science trainees,

Alma Better, Bangalore

Contents:

1. Abstract:

This EDA project includes "Airbnb booking Analysis" database.In this work, we present analytical results obtained by data mining on Airbnb booking dataset. The main objective is to visualize airbnb data and gather useful insights related to host's, properties, average price and availability of room in different neighbourhood groups.

For each property , widespread information is available, including the details such as host i'd ,host name, neighbourhood ,room types, price of rooms, duration of stay, number of reviews, availability and host listing count.

keywords: Exploratory Data Analysis, Data Visualization, Airbnb booking.

2.Problem Statement:

The Airbnb defines the listing done by the hosts for their property, in different neighbourhood groups, namely Bronx, Brooklyn, Manhattan, Queens and Staten Island. It also defines their room types , reviews per month , but we don't have detail about number of booking so we consider the number of reviews as the minimum number of bookings a particular a host could receive and with this we will do our data analysis

Objectives:

- Exploring and Cleaning the Dataset
- To establish relationship between various features of the Dataset
- Present these relationships using various Data Visualization Techniques
- Draw the useful insights from it

3. Introduction and Summary of the dataset

- Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app.
- Airbnb stands for "Air Bed and Breakfast" Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com.
- Airbnb booking analysis dataset contains 16 columns and 48895 rows.

Understanding the columns of the Data Frame

- **Id-** It is listing id. Whenever a host list its property on Airbnb, a unique id is created, since every id is unique there are 458895 unique id. As per my observation ,every listing is a listing of unique properties.(means the same property is not listed more than 1)
- **Name-** It is the property description. Two properties can have the same name, but they are not the same property, every property in the data are unique, as all of them are having different longitude and latitude, that means every properties are situated in different location.
- **Host id-** Its the identity id given to individual host (For example- two different host cant have the same host_id). So if we want to count the number of host, we count the unique host_id.
- **Neighbourhood group-** This are the 5 Boroughs of New York city, in which the property is located.

- **Neighbourhood-** This are the towns and villages in New York
- **Room type-** Types of properties
- **Price-** Property price per day
- **Minimum nights-** It is the minimum number of night you can book a property,(example-1 year rental contract)
- **Number of reviews-** Total number of reviews
- **Calculated host listing count-** The number of times the host has listed its property.
- **Availabilitty_365-** the number of days the property were available for booking.

4. Steps involved:

The operations on the dataset are done by python scripts on Google Colab.

A. Importing Libraries:

Following Libraries are used in this analysis

i. **NumPy:** NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of arrays

ii. **Pandas:** Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured and time series data both easy and intuitive. It aims to be the fundamental highlevel building block for doing practical, real world data analysis in Python.

iii. **Matplotlib:** Matplotlib is a visualization library in Python for 2D plots of arrays iv. Seaborn: Seaborn is a library for making statistical graphics in Python

iv. **Seaborn :** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data. One has to be familiar with Numpy and Matplotlib and Pandas to learn about Seaborn

v. **Datetime:** Python Datetime module comes built into Python, so there is no need to install it externally. Python Datetime module supplies classes to work with date and time ,datetime – Its a combination of date and time along with the attributes year, month, day, hour, minute, second, microsecond.

vi. **Plotly:** The plotly express module contains functions that can create entire figures at once, and is referred to as Plotly Express or PX. Plotly Express is a built-in part of the Plotly library, and is the recommended starting point for creating most common figures.

5. Preprocessing data

A. Data preparation and cleaning:

Drop the column which are not required,like i'd

Null values Treatment: Our dataset contains a large number of null values which might tend to disturb our accuracy hence we replace them at the beginning of our project in order to get a better result.

- Fill the NaN values in 'host name' by 'no name'
- Replace NaN values in 'name' column by 'no description'
- Replace NaN values in 'reviews per month' column by 0
- If price is 0 , replace it by the average price of all similar 'room type' with similar 'neighbourhood group'
- **Handling Outliers of price column:** By plotting boxplot, we see some outliers. We have used mean for removing outliers.Seaborn boxplot are used for visualizing the outliers in the present data.
- Plotting different features against one another:
- After doing the data cleaning we are left with 45923 rows and 16 columns

B. Creating the visualizations:

- Using the various functions of above mentioned libraries, we have created various types of charts like Correlation heatmap, Pie chart, Bar Plot, Scatter Plot, Scatter Map Box, etc
- To establish meaningful relationships between the variables of the Data set. You can see those charts in the attached code file.

6. Exploratory Data Analysis (EDA)

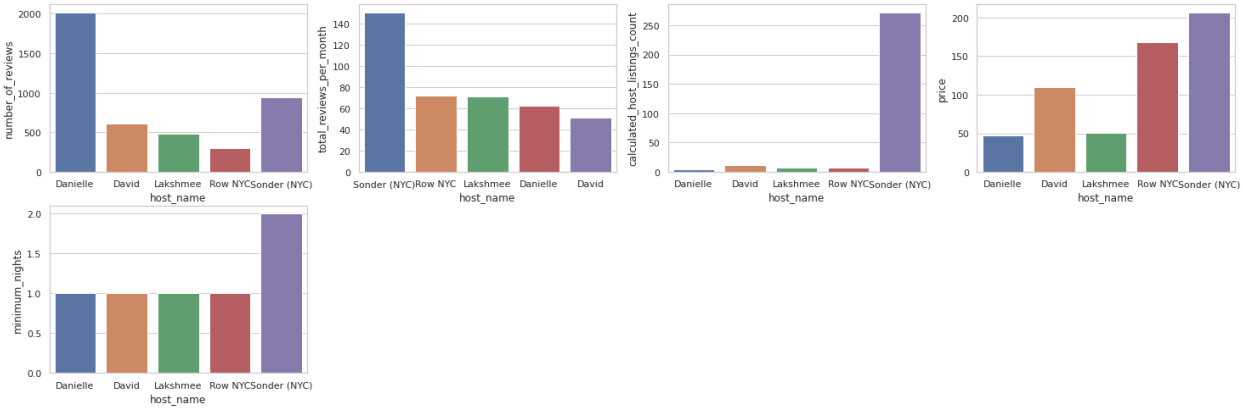
- **Collection of useful insights:**

1. Count the number of unique hosts that have listed properties on Airbnb

- Since 2 different hosts can have the same host_name, we use host_id.
- host id is unique for an individual host
- with the help of value_count(), we can count the total number of unique hosts that have listed their property on Airbnb. There are three 335398 unique hosts.
- We apply length function to above data frame to get total number of unique host ids

	host_id	listing_counts
0	219517861	272
1	107434423	180
2	137358866	103
3	12243051	95
4	30283594	95
...	...	...
35384	95485067	1
35385	167560332	1
35386	263742622	1
35387	205706382	1

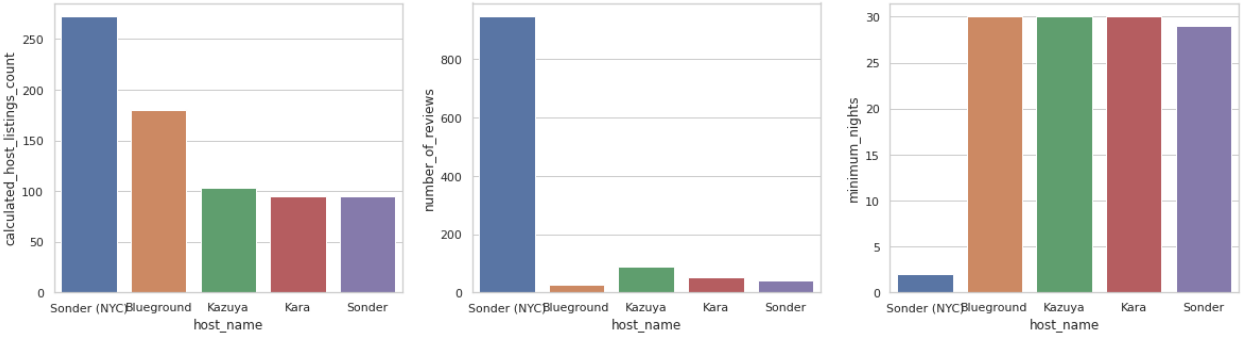
2. Find the rate of reviews per month, received by each host for all its properties (combined) listed on Airbnb



- Grouping by the host ID and summing up the reviews per month is not the correct way to find the rate of review per month of a particular host on all his property.
- For this problem, we will find the total number of reviews received by a particular host for all its properties and we will also find the time interval(in a month), in which the host has received all these reviews.

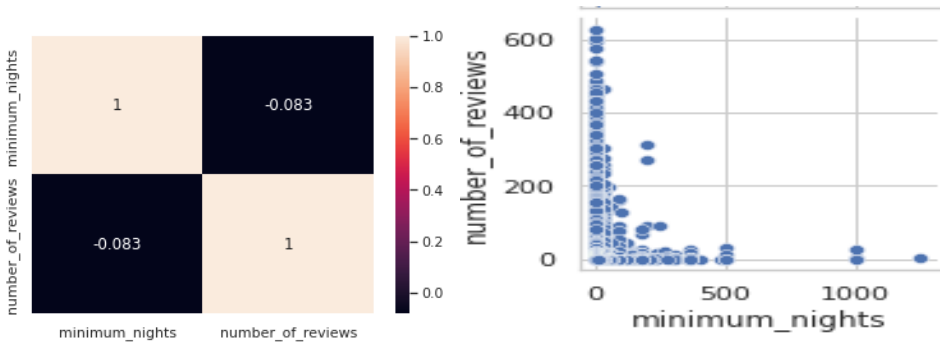
- Then we will find the rate of review per month by dividing the total number of reviews by the time interval (is in a month).
- Sonder has the highest review per month, highest listing count, and most of his properties are in Manhattan which is the most densely populated among the 5 boroughs of New York city. Hence Sonder is the busiest host.
- Most of his properties are entire home/apt, with an average price of 251$ and the minimum_night stay is 2 nights.

3. Top 5 host with heights listing counts



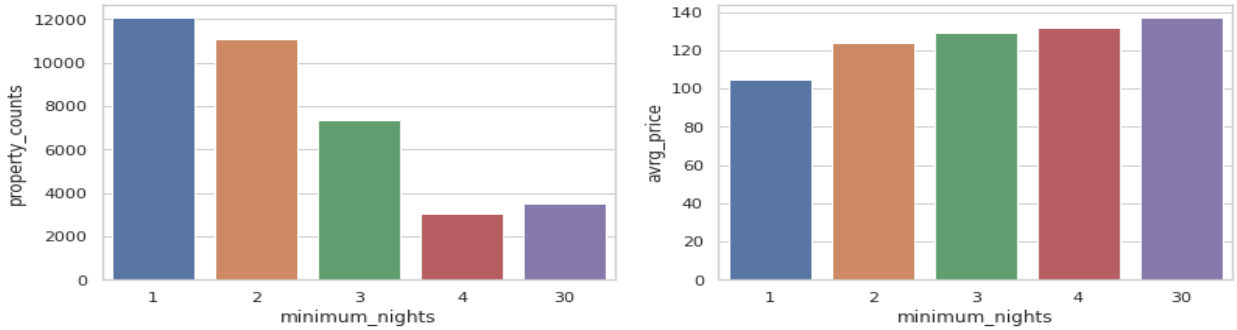
- Group by the host ID and count all the calculated_host_listings_count of the host ID
- Sort the host_id in descending order of their calculated_host_listings_count
- Get the top 5 hosts that have the maximum number of calculated_host_listings_count.
- Sunder has the most number of listing with (272)followed by Blueground (170)
- In spite of Blueground having the second highest listing count, he has got very few reviews among all, that is because the minimum night required to book Blueground's property is 30 nights. whereas for Sonder it is 2 nights.
- Scatterplot will shows the relation between minimum_nights and total_reviews of all the host

4. The relation between minimum nights and total_reviews.



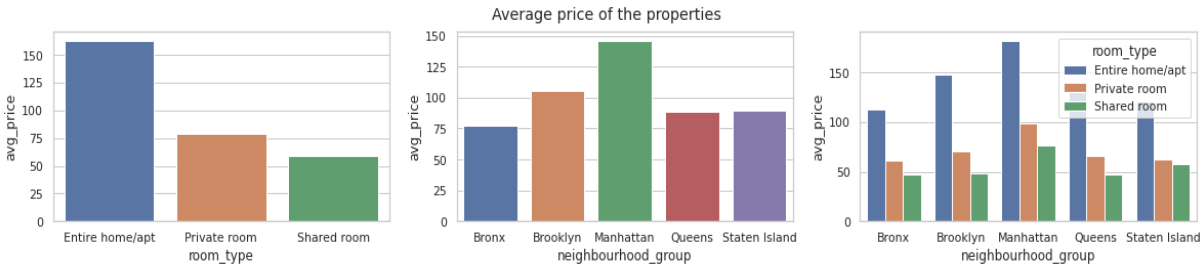
- This can be shown with the help of a heat map
- We see that are negatively correlated, that is lower minimum nights, higher is the number of reviews, this is because if a host gets rental for his properties for 60 days, then he will not get booking for at least 60 days.
- Hence he gets 1-0 review within 60 days. And the host who has set minimum_nights as 1 or 2 will get more than 1 booking within the period of 60 days, hence he will get more than 1 review within this 60 days.

5. What is the minimum nights stay set by most of the hosts on Airbnb?



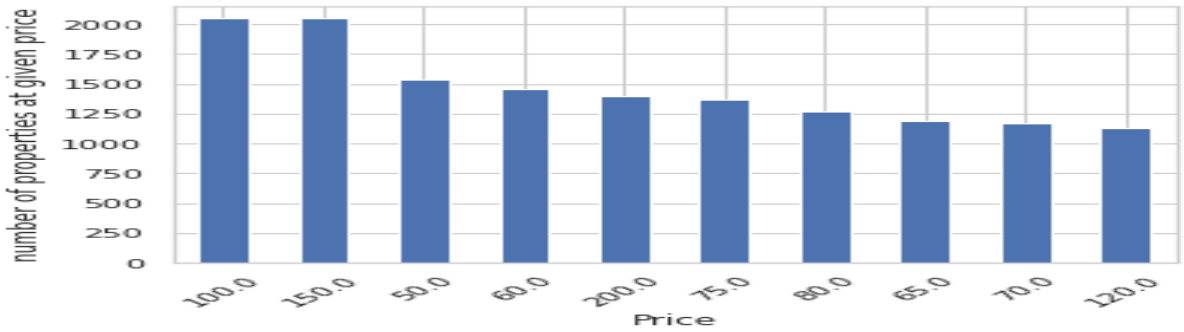
- For this, we can apply the value_counts() method On the mimimum_nights column and sort the counts in descending order
- We see that most of the hosts on Airbnb have set the minimum_nights stay as 1, 2 and 3 nights.

6. The average price of Properties.



- Group by the room type and take out the mean of their price.
- Group by the neighbourhood group and take out the mean of their price
- Group by the room type and neighbourhood group and take out their corresponding mean price
- Entire home/apt is costly than other room_type and shared_room is cheaper.
- Properties in Manhattan are costly as compared to other neighbourhood groups.

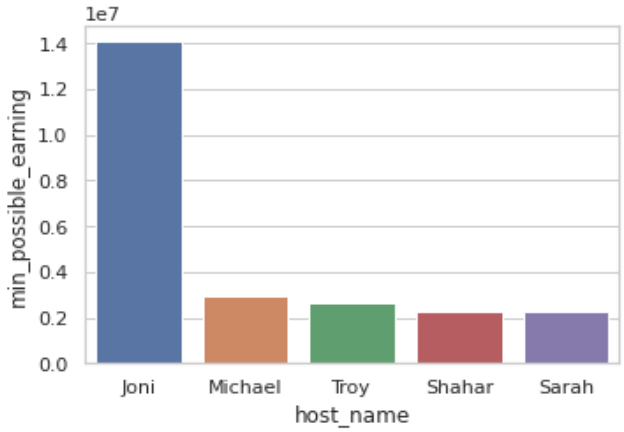
- **7. Price at which most of the properties are available.**

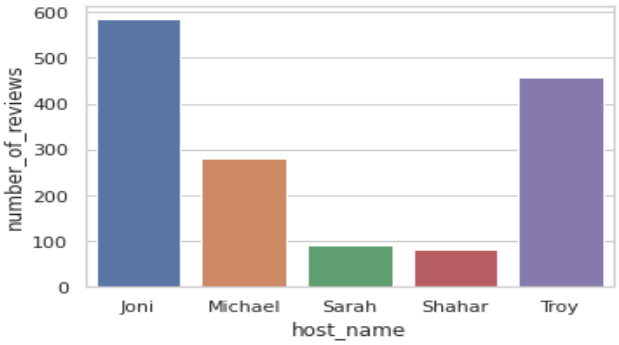


- We use the value count method to count the number of unique prices.
- Sort count of the prices in descending order and plot the top 10 Prices at which most of the properties are available.
- Most of the properties listed on Airbnb cost between 100 to 200 dollars per day.

8 Top 5 hosts with minimum possible earnings.

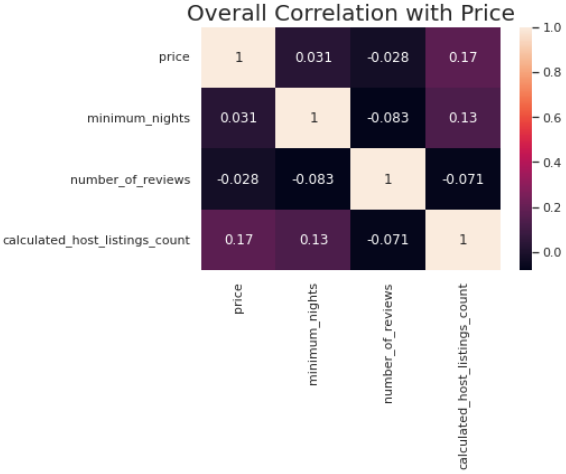
- number of reviews received by the host.
- Joni's min_possible_earning is maximum (1.4 billion $) compared to other hosts.



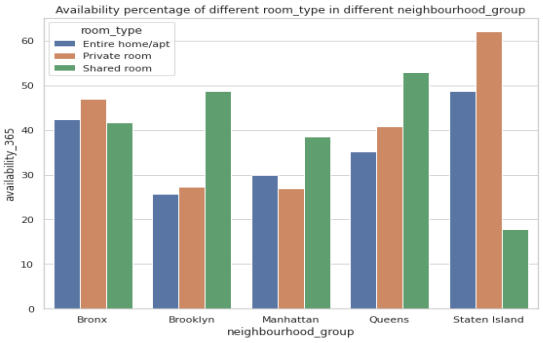


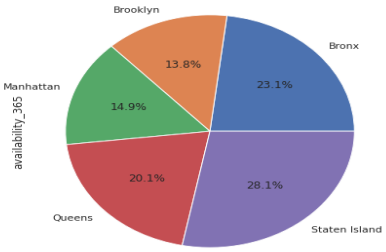
9. Overall Correlation with Price.

- We will find the correlation of different numeric columns with the price, with the help of a heat map.
- Most of the properties with cheaper prices often get more reviews as compared to the properties with higher prices.
- That is maybe because people prefer renting cheaper property, which leads to more reviews.



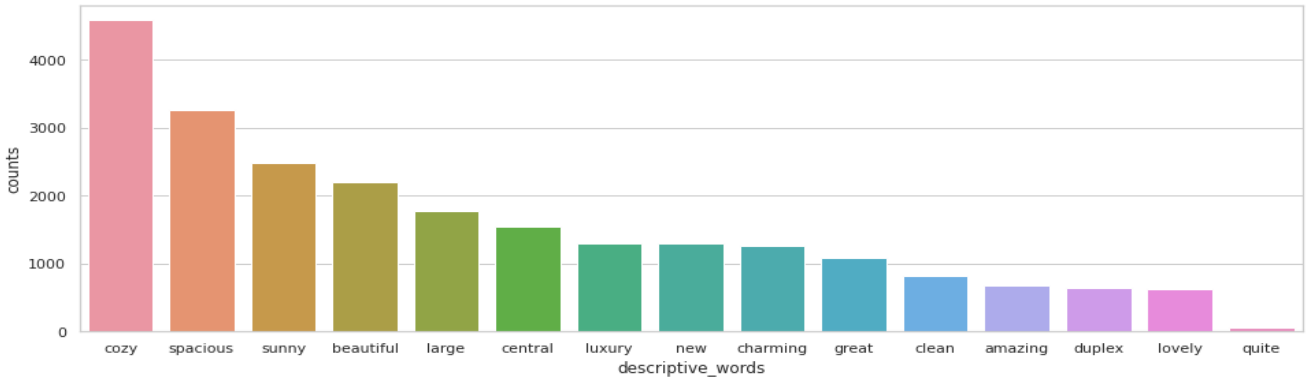
10. Availability percentage of properties in different neighbourhood group





- We have groupby by the neighbourhood group and room type and apply lambda function on the availability 365 column. Lambda function defined as
- lambda x:(x.sum()*100)/(365*x.count())
- One of the factors on which the availability of property depends is booking.

11. Most Popular Descriptive words are used by the host to describe their properties



- Cozy, spacious, sunny, beautiful, large are the few most popular descriptive words used by the host to describe their properties

Conclusion:

- Entire room/apt is the highest property type listed on airbnb
- Most of the properties listed on Airbnb are situated in Manhattan
- Brooklyn has the highest number of private rooms.
- Manhattan has the highest number of entire room home or apartment and shared room
- Most of the property types have one or two minimum night stay
- Apartment in Manhattan are more expensive whereas apartment in Bronx and Queen are cheaper
- The avg price of private room and shared room are almost same, so people will prefer renting private room over shared room
- Most of the properties listed on Airbnb cost between 100 to 200 dollars per day.
- Sonder is the busiest host among all, with 397 reviews per month, highest listing count of 272 listings on Airbnb
- minimum_nights and total _reviews are negatively correlated
- Manhattan is the smallest neighborhood group but has about 42.5% of the properties listed on Airbnb, that's because it is a densely populated city of New York.
- Staten(0.8%) and Bronx(2.3%) have a very low number of properties listed on Airbnb.
- Availability of properties for booking in Manhattan and in the north part of Brooklyn is less as compared to other neighbourhood groups. This shows that the properties in this region are often booked, hence People prefer renting an apartment in Manhattan and in the northern part of Brooklyn.
- The average price of private rooms and shared rooms are almost the same, so people will prefer renting private rooms over shared rooms.

Reference:

1. Alma Better
2. Stack Overflow
3. GeeksforGeeks