Airbnb Bookings Analysis

Raja Chowdhury
Data Science Trainees
Alma Better, Bangalore

Abstract:

New York city has been one of the most popular cities for travel and the hottest market for Airbnb. Airbnb is an online-based marketing company that connects people looking for accommodation (Airbnb guests) to people looking to rent their properties (Airbnb hosts) on a short-term or long-term basis. The dataset contains the real-world data of Airbnb of New York city. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

1. Problem Statements

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Explore and analyze the data to discover key understandings (not limited to these) such as:

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

2. Introduction

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, New York City, the platform is accessible via website and mobile app. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking. In general, Airbnb is cheaper than hotels because Airbnb does not have to pay for the overhead costs of a hotel or the general management of such a large operation.

3. Steps performed in EDA

Handling this dataset with the fundamental steps:

- ❖ Imports Libraries, Dataset and Other modules.
- ❖ Understanding Our Dataset.
  - Analysing Data: Reading & Inspection of Data.
  - Data Structure (Head, Tail, Shape)
  - Data information (Basic Schema)

- Arithmetic measurement (Statistical Values)

❖ Data Preparation & Processing.
- Cleaning Null Values
- Removing Zero Price Values
- Handling Outliers
- Finding relation and dependency in Data

❖ Exploratory Analysis
- Categorical Data Visualisation.
- Visualization using different plots in relation with milestone questions.
- Other important Visualisation.

❖ Key Findings and Conclusions

4. Data Understanding

As the objective is clear the data needs to be analysed and this process starts with understanding our dataset of Airbnb NYC(New York City) and we come to understand the following things.

Our dataset contains 16 variables as follows:

1. id- refers to the identity number of the property listed by a particular host.

2. name- refers to name of the listed property in Airbnb.

3. host_id- refers to the identity number of the host who registered on AirBnb website.

4. host_name- refers to name of the hosts, who listed their properties.

5. room_type- represents the various types of room in the listed property.
- Entire home/apt
- Private room
- Shared room

6. price- refers to the cost of the room per night in USD.

7. minimum_nights- refers to the minimum number of nights stayed by the customer

8. number_of_reviews- refers to the number of customers reviewed the property.

9. last_review- refers to the date when the listed property was last reviewed.

10. availability_365- refers to the availability of the listed property out of the total 365 days of a year.

11.reviews_per month- refers to the count of reviews per month the property received

12. longitude- these represents the longitude coordinates of the property listed.

13. latitude- these represents the latitude coordinates of the property listed.

14. neighbourhood_group- refers to the names of the neighbourhood groups present in the NYC.
- Manhattan
- Brooklyn
- Queens
- Bronx
- Staten Island

15. Neighbourhood- refers to the names of the neighbourhood present in NYC.

16. calculated_host_listings_count- refers to the number of properties listed under a particular host.

5. Data Processing & Preparation

After exploring our dataset, we can say that the dataset needs some cleaning before going to visualisation. We observed that there were some null values present in few of our columns and there were some outliers too. Other columns such as number_of_reviews and calculated_host_listings_count areskewed toward right, so we need to transfer them into categorical variables.

Cleaning Null Values:

last_review: last_review column has more than 20% of the null values and this is quite irrelevant column for our analysis so we can simply drop this column.

host_name & name: Missing values are 21 & 16 in host_name and name columns respectively and both are less than 0.5% so we will simply drop the rows corresponding to the missing values as they are very less in numbers and that won't affect our visualisation.

reviews_per_month: We found some null values in reviews_per_month column and from analysis point of view it is an important column, so we will replace all the Null values of review_per_month column with zero ('0')

Removed Rows Having Zero Price Values:

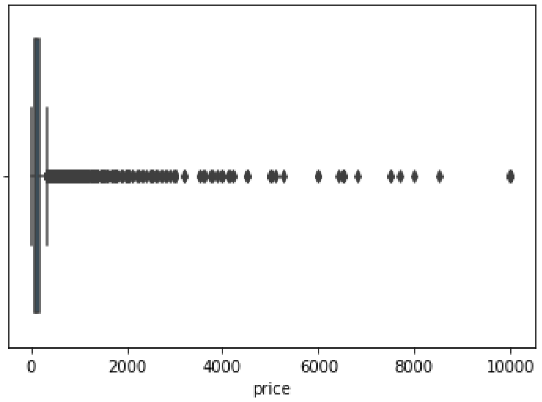
We clearly saw that there were eleven entries having price equal to zero (0) which needed to be drop in order to get meaningful analysis.

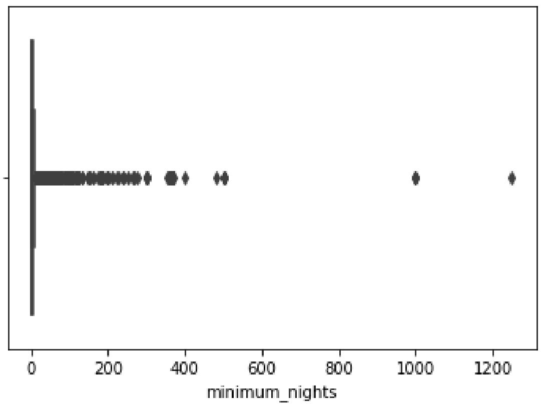
Finding Relation and Dependency In Data:

Using Heatmap we checked the correlation between the numerical data however we did not find any potential correlation between the numerical variables except number of reviews and reviews per month.

Handling Outliers:

Two column price & minimum_nights are having Outliners confirmed using Boxplot chart. We use the z-score method to handle our outliners, in this method we add two more columns for z-score “z_price, z_min_nights” then we filter our data set by putting conditions that only entries having z-score less than 3 will be considered.





6. Data Analysis

Exploring the information present in the dataset by analyzing it and visualizing the values of various columns and correlations between different columns. Basically we have two categorical columnsin our Airbnb dataset and on the basis which, we are going to do the further analysis.

1. Neighbourhood Group
2. Room Type

Neighborhood Groups:

Manhattan has the highest number of bookings followed by Brooklyn and they together hold more than 85% of the total booking while Staten Island holds the least percentage of booking in the neighborhood group.

Room Types:

Entire home/apt holds the highest number of bookings followed by Private room and they together hold around 97% of the total booking while Shared room holds the least percentage of booking.

Room Type Demand In Neighborhood Groups:

- In almost all neighborhood groups, private rooms have been more preferred followed by Entire Home/apt.
- Except in Manhattan, where Entire Home/Apt is more preferred than Private rooms.

Hosts and Areas:

Manhattan is the most densely populated neighborhood group by hosts and the reason for it can be the presence of Manhattan Island, bounded by the Hudson, East and Harlem rivers, and others tourist place and therefore it has highest number of hosts within lesser area space than any other neighborhood group.

Location and Price:

Manhattan has higher prices and the reason for it can be the presence of Manhattan Island, bounded by the Hudson, East and Harlem rivers, and other tourist places.

Prices In Neighbourhood Groups:

- Firstly, the price of the Entire Home/Apt is very high in all Neighborhood Groups followed by Private Rooms.

- Also, the price in Manhattan is relatively high in all its room types followed by Brooklyn and almost similar for Queens, Staten Island and Bronx.

- We can think of it as demand for Entire home/apartment is higher, especially in Manhattan and Brooklyn thus pushing the price to the higher side.

Reviews in Neighborhood Groups:

Brooklyn got the maximum percentage of reviews followed by Manhattan despite of Manhattan having the highest number of bookings and they together hold the 82% of the reviews. While Staten Island has the least percentage of total reviews

Availability Of Room In Neighborhood Groups:

Brooklyn and Manhattan have higher number of bookings compared to the other neighborhood groups therefore they both have less availability of rooms. Similarly Staten Island and Br onx have least number of booking therefore they have more availability of rooms compare to others.

Average Night Stay In Room Types:
Entire Home/Apartment has higher average night stay and reason for it, we can assume is that mostly Entire Home/Apt are booked for family vacations, group tours or for events so they tend to stay longer. While Private room and Shared room are most likely to be booked for official purpose or for limited time stay.

Host Listing In Neighborhood Groups:
The hosts of Manhattan neighborhood group have high number of properties listed in Airbnb and this could be because Manhattan has higher demand as well as higher price compared to any other neighborhood groups location which encourage hosts to list more of his properties.

Most Busiest Host :
Michael is the most busiest host in the Airbnb NYC dataset as it has the most number of bookings. The reason for it can be its location as it is situated in a Manhattan neighborhood group and has Entire Home/apartment where demand is at peak with higher price that we have already seen in the previous visualizations.

Host With Most Listing:
Sonder (NYC) is the host with the most number of listed properties in Airbnb NYC. Sonder (NYC) belongs to the Manhattan neighborhood group and most of its listings have Entire home/apartment.

Most Demanded Neighborhoods:
Williamsburg is the neighbourhood with the most number of bookings in our Airbnb dataset and seven out of ten top neighbourhoods belong from Manhattan neighbourhood groups with Entire home/apartment room type.

7. Challenges faced

- Studying the dataset, understanding the columns & observed exploration tasks.
- To answer a number of the questions we needed to understand the business model of Airbnb and its works.
- Dealing with data size & handling missing values, null values and duplicates.
- Designing multiple visualizations to summarize the statistics in the dataset and correctly communicate the results and traits to the reader.
- How to choose which variables we need to explore?

8. Key Findings

1. Manhattan and Brooklyn neighbourhood groups hold around 85% of the total booking while Staten Island has the least number of bookings.

2. Manhattan neighborhood hosts have higher number of properties listed in Airbnb.

3. Entire home/apt is the most preferred room type followed by private room and they hold around 97% of the total booking.

4. Queens has significantly less host listings than Manhattan. So, we should take enough steps to encourage host listings in Queens.

5. Seven out of ten top neighborhood belongs from Manhattan neighborhood group with Entire home/apartment room type.

9. Conclusions

❖ We can conclude from the analysis that Manhattan is the top neighborhood group with the highest number of bookings and host listings. Seven out of ten top hosts are from Manhattan. One of the probable reasons for being the most preferred Neighborhood Group is that Manhattan is world-famous for its museums, stores, parks, music and cultures. It has high number of tourist places where Entire Home/Apartment are more preferred as stay options and for longer periods. These factors increased the demand and led to higher prices compared to any other neighbourhood groups.

❖ Brooklyn has significant number of bookings because Brooklyn also has some famous bridges, parks, museums, islands and other tourist places but with more affordable prices as compared to Manhattan. It also received the maximum number of reviews followed by Manhattan.

❖ Rest 3 neighborhood groups namely Queens, Bronx and Staten Island are observing very less number of bookings and hosts, especially on Staten Island. Considering that these are less popular areas, it is possible that some guests choose these locations to save up money or for official purposes who want to stay for limited time periods. And these neighborhood groups have higher room availability than Brooklyn and Manhattan.