Airbnb Bookings Analysis

Pranil Savale

Data Science Trainee
Almabetter, Bangalore

Abstract:

Airbnb is an American enterprise that operates an internet market for lodging, in most cases homestays for holiday rentals, and tourism activities.

I was provided with a data set of Airbnb NYC (2019). Our analytics can be used to make better business decisions, understand customer and host behaviour and performance on the platform, target your marketing initiative, implement innovative add-on services, and more.

1. Problem Statement

Since 2008, guests and hosts have used Airbnb to expand travel options and present a more unique and personal way of experiencing the world. Today, Airbnb has become a unique service, used and recognized around the world. Analysing data from millions of Airbnb listings is a critical enabler for the company. These millions of entries create a huge amount of data.

The data I will analyse is from Airbnb NYC (2019). Our main analysis goals are across four propositions that can be summarized as Host Learnings, Areas,

Price, Ratings, Locations, etc. but I am not limited to that, I will also tryto explore some more ideas.

2. Understanding the variables

- **id**: Unique listing ID
- **name**: Name of the listing
- **host_id**: Unique host ID
- **host_name**: Name of the host
- **neighbourhood_group**: Location
- **neighbourhood**: Area
- **latitude**: Latitude coordinates
- **longitude**: Longitude coordinates
- **room_type**: Listing space type
- **price**: price in dollars
- **minimum_nights**: Amount of nights minimum
- **number_of_reviews**: Number of reviews
- **last_review**: Latest review
- **reviews_per_month**: Number of reviews per month
- **calculated_host_listings_count**: Amount of listing per host

- **availability_365**: Number of days when listing is available for booking

3. Introduction

About the dataset:

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

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.

The data I am going to analyse is the data of Airbnb NYC (2019). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas,price, reviews, locations etc. but not limited to. I will also try to explore some more insights.

4. Python Libraries used

- Pandas
- Matplotlib
- Seaborn
- Folium
- Klib

5. Graphs used for data visualization

- Count Plot
- Bar graph
- Heatmap
- Box Plot
- Maps

6. How pricing works?

Airbnb reservation will cost you the nightly rate indicated by the host plus any additional fees or expenses decided by the host or by Airbnb.

7. Type of Fees

Airbnb service fee: Airbnb charges a guest service fee that funds round-the-clock community assistance and general smooth operation.

Some hosts charge a cleaning fee to cover the cost of maintaining their place.

Some hosts charge an extra guest fee for each visitor over a predetermined number.

Security deposit: Using Airbnb offline fees function, hosts who manage their listings using software that is connected to the API can set a security deposit.

Value Added Tax (VAT, JCT, and GST) is levied against visitors from certain nations.

Local taxes are assessed according on where the host's property is located.

Final price mostly includes all the type of fees that are mentioned above.

8. Approach Used

The approach I have used in this project is defined in the given format-

1) Loading our data: In this section, I just loaded our dataset in colab notebook and read the csv file.

2) Data Cleaning and Processing: In this section I have tried to remove the null values and for some of the columns I have replaced the null values with the appropriate values with reasonable assumptions.

3) Analysis and Visualization: In this section I have tried to explore all variables which can play an important role for the analysis. In the next parts, I have tried to explore the effect of one over the other. In the next part I tried to answer our hypothetical questions.

4) Future scope of Further Analysis: There are many apartments having availability as 0 and date of last_review is very old, which can mean that they must have stopped their business, I can find the relation with neighbourhood with these apartments if I could dig much, various micro trends could be unearthed, which I am not able to cover during this short duration efficiently. There are various columns which can play an important role in further analysis such as number of reviews and reviews per month finding its relation with other factors or other grouped factors can play an important role.

9. Challenges Faced

- While doing the analysis I found out that 36% of the data has 0 availability in the availability_365 column, which is an extreme case. But I didn't have other relevant required data so I couldn't alter this column.
- Further I found out that there were many listings whose price was 0, which is not normal. So, I filled these values by the respective median price and updated the price column.

- While getting host_name with highest listings I found out that there are many hosts whose names are the same so I went by host_id as this is unique, host_name is not unique.
- There are many listings whose date of last review is very old. This can mean that they must've stopped their business then those listings are of no use to us for doing analysis in present. But this assumption can also be wrong so I didn't alter this column.
- There were many outliers in the price column of some hosts which weren't benefitting the host as well as the customer.
- The biggest challenge that I faced is finding the busiest hosts. If I try to find the busiest hosts by only the number of reviews then this may not be the correct metric, because I don't know the current status of the host having the highest number of reviews. For example, if I check that one host are x number of reviews which is highest but when I check thedate of last_review and find outthat the reviews are very old than the current date, then I can infer that business is currently shut down so how can I take such hosts into consideration for knowing the busiest hosts. Ideally the busiest host should be that one whose occupancy is almost full or full.

10. Conclusions

This Airbnb-NYC(2019) dataset is a very informative dataset having 48895 rows and 16 columns. I found that SONDER(NYC) has the highest number of listings i.e 327 listings .I found that the highest number of listings in any neighbourhood group is Manhattan. The Williamsburg neighbourhood has the most number of listings among all neighborhoods .Upper West Side, Astoria and Greenpoint neighbourhoods have the costliest listing in NYC. Bedford-Stuyvesant neighbourhood has the highest number of total reviews and highest number of reviews_per_month. And Maximum listings are listed on Manhattan and Brooklyn neighbourhood_groups. Staten Island and Bronx neighbourhood_group have very less numbers of listings. Most of the listings on Airbnb in NYC are either Entire Home/Apartment or Private Room. The people who prefer to stay in the entire home/apartment are likely going to stay longer, whereas people who prefer to stay in private_room are likely to stay for a shorter period of time than the people who prefer to stay in entire home/apartment. Many rows are having values as 0 in the price column, so this seems like an error which must be rectified by Airbnb. Keeping the high price of the listing and having 0 availability isn't benefiting the host as the consumer is ready to pay the price but even after that there are no available rooms then what's the benefit of paying such a premium. Maya (host) has the

heighest total number_of_reviews. Average prices of all the room_types in Manhattan are more than the average price of each room_type in other neighbourhood_group. Average prices of all the room_type in Bronx neighbourhood_group is less than all the other neighbourhood_groups.