Airbnb Bookings Analysis

Vaibhavkumar Gupta, Bhavik Verma,

Priyanka Pal, Shayan Somanna, Dilkhush Sharma

Data Science Trainees

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.

We were 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 we 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 we are not limited to that, we will also try to 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:

The Airbnb dataset has around 49000 observation and 16 columns.

The dataset has mix data type of numerical, unique, date and categorical type.

Airbnb, as in "Air Bed and Breakfast," is a service that lets property owners rent out their spaces to travellers looking for a place to stay. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves.

Airbnb was started in 2008 by Brian Chesky and Joe Gebbia, Airbnb is based on a peer-to-peer business model. This makes it simple, easy to use, and tends to be more profitable for both parties. The model also gives you the opportunity to customize and personalize your guests' experience the way you want.

The data we are going to analyse is the data of Airbnb NYC (2011-19). Our main objectives of analysis will be above four statements which can be briefed as learnings from hosts, areas, price, reviews, locations etc. but we will not be limited to here. We 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 we have used in this project is defined in the given format-

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

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

3) **Analysis and Visualization**: In this section we have tried to explore all variables which can play an important role for the analysis. In the next parts we have tried to explore the effect of one over the other. In the next part we tried to answers 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, we can find the relation with neighbourhood with these apartments if we could dig much, various micro trends could be unearthed, which we are 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 we found out that 36% of the data has 0 availability in the availability_365 column, which is an extreme case. But we didn't have other relevant required data so we couldn't alter this column.

- Further we found out that there were many listings whose price was 0, which is not normal. So, we filled these values by the

respective median price and updated the price column.

- While getting host_name with highest listings we found out that there are many hosts whose name are same so we 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 we didn't alter this column.

- There were many outliers in price column of some hosts which weren't benefitting the host as well as the customer.

- The biggest challenge that we faced is finding busiest hosts. If we try to find the busiest hosts by only number of reviews then this may be not the correct metric, because we don't the current status of the host having highest number of reviews. For example, if we check that one host are x number of reviews which is highest but when we check the date of last_review and find out that the reviews are very old than the current date, then we can infer that business is currently shutdown so how can we take such hosts into consideration for knowing the busiest hosts. Ideally busiest host should be that one whose occupancy is almost full or full.

10. Conclusions

Airbnb-NYC(2019) is a dataset of house booking having 48895 rows and 16 columns.

In the data Sonder(NYC) is having highest number of houses holding 327 listings.

In neighbourhood_group "Manhattan" having highest number of business or listing.

Williamsburg neighbourhood has the most number of listings among all neighborhoods .

While looking for costliest listing in NYC we got that 'Upper West Side, Astoria and Greenpoint' neighbourhood having the costliest listing.

Bedford-Stuyvesant neighbourhood is having highest number of total reviews and highest number of reviews_per_month also.

Manhattan and Brooklyn neighbourhood_groups are similar to each other in the count of listing and having highest number of listing. Staten Island and Bronx neighbourhood_group have very less numbers of listings according to Manhattan and Brooklyn.

Entire Home/Apartment or Private Rooms are most listed on Airbnb-NYC.

Guests prefer to stay in private rooms are stay for a shorter period of time according to the guests prefer to stay in the entire home/apartment.

Many price is having 0 in the price column, it seems like a glitch which must be rectified by Airbnb.

Maya (host) having 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 all neighbourhood_group.

Average prices of all room_type in Bronx neighbourhood_group is less than all the other neighbourhood_groups.