Airbnb Booking Analysis

Surabhi Mali & Ruchi Yadav

Abstract:

Airbnb Bookings Analysis is based on discovering key understandings about hosts, areas, and their traffic.

We will gain information about factors affecting booking demands by observing distributions of booking across various features as well as the relationship between features.

The conclusions from this EDA can benefit both the company and the customers. Understanding the dynamics of bookings allows them to know the needs of the customers and where to focus in terms of improving their services. The EDA can also give the customers an idea of what to expect in terms of services and understand what choices to make in order to get the best deals.

Keywords: Airbnb, Data Cleaning, Exploratory Data Analysis

1. Problem Statement

Since 2008, guests and hosts have used Airbnb to expand possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb become of kind 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 decision, understanding of customers and providers (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services, and much more. This dataset has around 49,000 observations in it with 16 columns and it is a mix of categorical and numerical values.

How do explore and analyze the data to discover key understandings so that Airbnb business can expand?

2. Introduction

The data used in this analysis is the outcome of the quest to answer the question

How is Airbnb affecting the neighborhoods? Inside Airbnb is an activist project, which has curated this dataset, to measure the impact of rental housing on neighborhoods and communities.

We will explore and visualize the dataset from Airbnb in New York using basic exploratory data analysis techniques. We will find out the distribution of every Airbnb listing based on their location, including their price range, room type, listing name, and other related factors.

The goal here is to explore the data and find useful insights from the data and find out different relations between the columns.

3. Airbnb Booking Dataset Summary

This dataset has around 49,000 observations in it with 16 columns and it is a mix of categorical and numeric values It contains different hosts, the neighborhood group the

properties are located in and the type of property customers most wish for. Exploring them will definitely help us have a very good understanding of the booking trends.

Column Information

- Host id: It is an id given to specific host in the given dataset
- Neighbourhood group: It represents the location in the given dataset.
- There are 5 different neighborhood groups:- 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island' and 'Bronx'
- Neighbourhood: It represents specific areas where the listings are located in 5 different neighborhood groups.
- Room type: It represents category of room type being listed as:-
- 'Private room', 'Entire home/apt', 'Shared room'.
- Minimum nights: It represents the number of nights spent by the customer in given listing.
- Number of reviews: It represents the number of reviews for listings.
- Availability 365: It represents a number of days in a year for which a given property is available for rent.
- Price: It represents the rate for a given room type in a given location for one night.

4. Steps involved

- **Setting up the notebook**

  The notebook is set up in the Google Collaboratory platform. The Google drive containing the dataset is mounted in the notebook and it is loaded as a pandas data frame. The necessary libraries such as NumPy, Pandas (for working on the dataframe), seaborn and matplotlib (for visualization) are imported.

- **Cleaning the Data**

  Null values and missing values: The dataset contains missing and null values.Host_name,last_review, and reviews_per_month have some missing values. So, we tried to preserve as many rows as possible by replacing null values with suitable values.

  Dropping unnecessary Data: Some data were considered erroneous data and removed. Here we dropped unnecessary columns called ‘id’ and ‘name’ to make a clean dataset. Now we have 48895 rows and 14 columns in the new dataset.

5. Exploratory Data Analysis(EDA)

Throughout the analysis, we tried to explore data and tried to find answers to questions that help us understand the factors determining the booking trends.

1. Exploring Neighbourhood-Groups and Room-types

There are 3 types of rooms:

1. Private room
2. Entire home/apt
3. Shared room

The majority of entire home/apartments are located in Manhattan and the majority of private rooms are located in Brooklyn.

There are 5 Neighbourhood-Groups called

1. Manhattan
2. Brooklyn
3. Queens
4. Bronx

5. 5.Staten Island

2.Which neighbourhood group has most number of grouping?

we can conclude that we have highest number of listing i.e. 44.3%

3.How many minimum_nights people stayed in each room_type?

If someone is booking Entire home/apt , They tend to stay for longer duration on an average 8 to 9 days. For private room on an average of 5 to 6 days. For shared room on an average of 6 to 7 days.

4. Explore the price insights

According to the statistics, it is clear that 75% of the listing's Price ranges from 0−200. But there are also 3 Airbnb with a maximum price of $10000.There are 11 values with the price of $0, which can be due to dynamic pricing or the willingness of not to share the price with the Airbnb.

5.Which neighborhood has the highest and lowest price?

We have plotted the most expensive and least expensive neighbourhood, and we will plot only Top 15 neighbourhood and Bottom 15 with respect to average price. This will help a traveller to choose the appropriate neighbourhood based on his budget and the number of reviews.Fort Wadsworth is the most expensive in terms of neighbourhood with 0 number of reviews on an avarage. Whereas Bull's head locality is the least expensive to stay with 15 number of reviews on an avarage.

6.How many number of locality listed and how price is getting affected by listings?

Williamsburg has most number of listing count which is '3920' whereas Fort Wadesworth,Rossville,Richmondtown, Willowbrook,FortWadsworth,NewDorp,Woodrow has one of the least listing which is '1'.

We can see this neighbourhood(Fort Wadsworth ,Woodrow) are one of the highest stay based on price the reason that the price is high in this neighborhood is due to the less number of listings.

7.Which Neighborhood Group have good number of reviws?

Review gives lot of insights about a particular place for tourist when it comes to online booking. A cheap place with bad review can drive a tourist for not booking and an expensive place with nicest review can attaract a tourist more. So we tried to compare the review of each neighbourhood-group and number of listings to figure out how they are related?

We have considered reviews more than 50, so that we can have an good insight of the data.So we can see from the below plot, Brooklyn got most review around 3000 in comparison to Manhattan around 2700, even if the listings for Manhatten is morewhich is 21661 than Brooklyn is 20104. Also Staten Island which is cheaper has less review than the other neighbourhood group also less number of listings.From this we can say that if we have more listings in neighbourhood_groups, the tourists can have more options to try which will leads to more number of reviws after visiting the place.

8.Which 5 host has most number of reviews?

Based on the review score(Minimum 50) we will see who is our top 5 Host, this increases the confidence of tourist before booking.

We have total 48874 hosts out of them 7075 hosts have review count greater than 50.Now we will get Top 5 host who have most number of reviews.
We can say that Michael has received most number of reviews after David.

6. Conclusion

The given dataset appear to be very rich dataset with a variety of columns that allowed us to do deep exploration on each significant column presented.After cleaning the data we had 48895 rows and 14 columns in the new dataset. There are 3 different types of rooms and 5 different Neighbourhood-Groups. we can conclude that the highest number of listing i.e. 44.3% in Manhatten. To conclude we can say, people, stay in private room for an average of 5 to 6 days and shared rooms on an average of 6 to 7 days. Statistics state that the 75% of the listing's Price ranges from 0− 200. But there are also 3 Airbnb with a maximum price of $10000.Fort Wadsworth is the most expensive in terms of neighborhood with 0 number of reviews on an avarage. Whereas Bull's head locality is the least expensive to stay with 15 number of reviews on an avarage.
We can see this neighbourhood(Fort Wadsworth,Woodrow) are one of the highest stay based on price the reason that the price is high in this neighborhood is due to the less number of listings.Brooklyn got most review around 3000 in comparison to Manhattan around 2700, even if the listings for Manhatten is morewhich is 21661 than Brooklyn is 20104. Staten Island which is cheaper has less review than the other neighbourhood group also less number of listings.From this we can say that if we have more listings in neighbourhood_groups, the tourists can have more options to try which will leads to more number of reviws after visiting the place.Michael has received most number of reviews after David.

7.References

1. Towards Data Science
2. Stack overflow
3. Medium
4. Kaggle