EDA on Airbnb NYC Dataset

Akshat Mishra
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales. Airbnb offers people an easy, relatively stress-free way to earn some income from their property.
I am provided with the dataset which has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

This EDA can help understand when and where bookings are higher than normal.? what could be the reason for the less bookings and what are the reason with which airbnb can increase the booking. And many other factors like which type of rooms are preferred by the clients.

Keywords: Exploratory Data Analysis, Booking analysis, Airbnb NYC Dataset

1.Problem Statement

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-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 - 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.

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?

Information of Data Set

- Id - ID is the unique ID given to each listing in the dataset
- Name - Name of listings
- host_id - IDs associated with hosts
- host_name - Name of the hosts who listed their properties on AirBnB
- neighbourhood_group - List of neighbouring cities from where people listed their properties.
- Neighbourhood - It is the sub division of neighbourhood_group
- latitude - latitude of the listing
- longitude - longitude of listings
- room_type - Different types of rooms which are listed on AirBnB

- price - Price of rooms per night
- minimum_nights - Minimum number of nights people can stay.
- number_of_reviews - Number of reviews associated to the listings.
- last_review - Last review given by users
- reviews_per_month - Number of reviews per month
- calculated_host_listings_count - Number of a listings of hosts
- availability_365 - Rooms availability

2. Introduction

Exploratory data analysis (EDA) is an approach using different methods from statistics through pandas and graphical tools like – matplotlib and seaborn, to better understand data. It is used mainly to maximize insight into a dataset, detect and test underlying assumptions.

The NYC Airbnb dataset contains 49000 observations with 16 columns which describes information about room_type, neighborhood_group, availability, host_id, name, reviews etc.

6. Steps involved:

- **Exploratory Data Analysis**
    - **Understanding the data**

      After loading the dataset, I used **AirBnB_Data.info()** to get the basic info about the data.

- **Finding Null values and cleaning data**

  Using **isna()** function, I found out the null values so that I can change the information later when I need it.

- **Working on problem statements to find out the desired information.**

  There are 8 questions for which I analyzed the data to get the desired result.

- **Features used**

  I used the pandas for calculations. And to visualize the data I used mostly bar plots from matplotlib and seaborn.

- **Findings and Conclusion**
    - The “entire home / apt category” has the highest number of rooms followed by private rooms. Shared room is the least preferred category.
    - In Manhatten (neighborhood group), we have highest "Entire home / apt" category whereas Brooklyn has the highest number of private rooms. Staten Island is having the least number of listings in all three types of rooms category.
    - After calculating we got to know that top 10

neighborhoods belong to only 2 neighborhood_groups which are Manhattan " and "Brooklyn".

- The most expensive category of rooms is the "Entire home/apt" followed by "Private room" and "shared room" is the least expensive comparing to others.
- Highest room price is in " Manhatten" and the lowest is in “Bronx"
- Manhatten, Brooklyn and Queens are the top 3 neighborhood groups in terms of total number of listings associated with them.
- Among all the neighborhoods under neighborhood groups "Fort Wadsworth" is the most expensive neighborhood where average per day price of rooms is highest in the whole data set.

Few more points can be drawn from the exploration

On the basis of above exploration and analysis, following points can be considered

Why “Manhattan” is Most preferred location.?

- In the whole Airbnb data set, most preferred room type is "Entire home/ apt". And the highest number of this room category is the found in Manhattan. Which means people have more option for their desired room category in Manhattan that's why they prefer Manhattan.
- Average room price on the basis of neighborhood group, is highest in Manhattan. But if look at the bigger picture and try to figure out the most expensive rooms on the basis of neighborhood, we found that the most expensive rooms are in "Fort Wadsworth" which is in Staten Island. And that is least preferred place.

Why “Staten Island” is Least preferred place.?

- Most preferred room category is "Entire home/apt" and Staten Island has the least number of rooms of this category.
- The average per day price on the basis of neighborhoods, in Staten island there are very expensive neighborhoods.

So, that’s is all whatever information I gather using this Airbnb NYC Dataset.