Airbnb Bookings Analysis

Shantanu Houzwala, Harshiv Bhatt and
Suloy Kumar Mandal
Data Science Trainees,
AlmaBetter, Bangaluru

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, the platform is accessible via website and mobile app.

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 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. New York is the most populous city in the United States, and one of the most popular tourism and business places globally.

I. Problem Statement

The main objective of this project is to explore and visualize the dataset from Airbnb in New York City using basic exploratory data analysis techniques. This will be done by finding out the distribution of every

Airbnb listing based on their location, including their price range, room type, listing name, and other related factors.

- **Understanding the data:**

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

1. id: Listing id
2. name: name by which the place is listed
3. host_id: ID for the host
4. host_name: person's name who has listed the property
5. neighborhood_group: list of all NYC regions
6. neighborhood: list of areas within different regions
7. latitude: coordinates for east-west direction
8. longitude: coordinated for north-south direction
9. price: price of rooms per night
10. minimum_nights: average of minimum nights booked
11. number_of_reviews: total number of reviews for each listing
12. last_review: date on which the listing was last reviewed
13. reviews_per_month: average reviews per month
14. calculated_host_listings_count: total number of listings by each host
15. availability_365: count of number of days the property is vacant.
16. room_type: Airbnb have 3 room types.

- Based on the information on the Airbnb website, the definition of each room type are:

- **Private room**

  Guests have exclusive access to the bedroom/sleeping area of the listing. Other parts area such as the living room, kitchen, and bathroom are likely open either to the host even to other guests.

- **Entire home/apt**

  Guests have the whole place for themselves. It usually includes a bedroom, bathroom, and kitchen.

- **Shared Room**

  Guest sleep in a bedroom or a common area that could be shared with others.

➢ **Map of New York City**

II. Steps Involved

1) Loading Data

For this project, we are using Google colab notebook IDE with a python programming language to write our script.

To get the data, we are using Airbnb data that shared by AlmaBetter for this project.

Before loading the data into IDE, first we need to import various external libraries/modules that are needed for visualization and analysis.

a. Load python libraries

- **Pandas** and **Numpy** library used for data analysis
- **Matplotlib**, **Seaborn and Ploty** library used for data visualization

b. Load Dataset

To load the dataset, we first need to mount our google drive. Next we use pandas library and function to read the CSV file pd.read_csv(file_path)

2) Cleaning Dataset

The next step is cleaning up the data, oftentimes the data that we load have various faults, such as duplicates, missing value, incomplete data, etc. By doing cleaning up, the data quality will have better quality to be used for further analysis.

a. Removing duplicates if any.

b. Dropping null observations.

3) Analyzing and Visualizing Data

After we clean up the data, the next step is exploring the data by visualizing and analyzing the values of the features, explaining the process and the results.

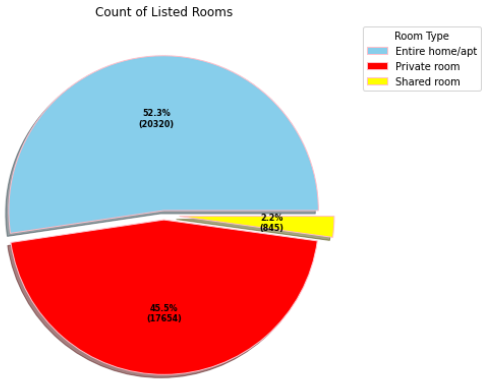
In my case, I looked over distributions of the data and the value counts for the various categorical variables.

Below are a few highlights from the analysis.

a. Total count of each room types

The total count of each room type suggests the kinds of property hosts are willing to rent.

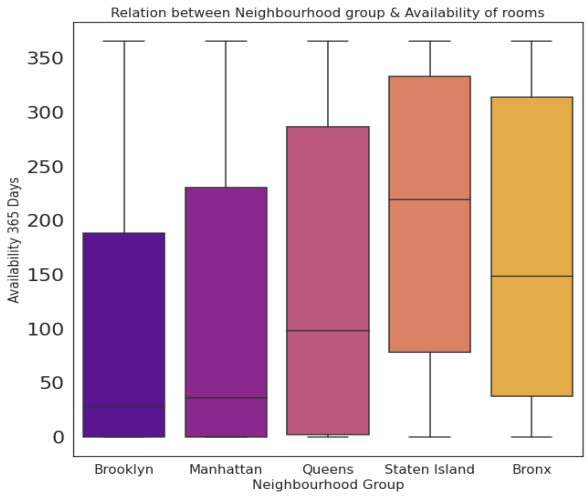
For this, I found unique values from column 'room_type'. Then I counted the values of each room type.



b. Room types and their relation with availability in different neighbourhood groups

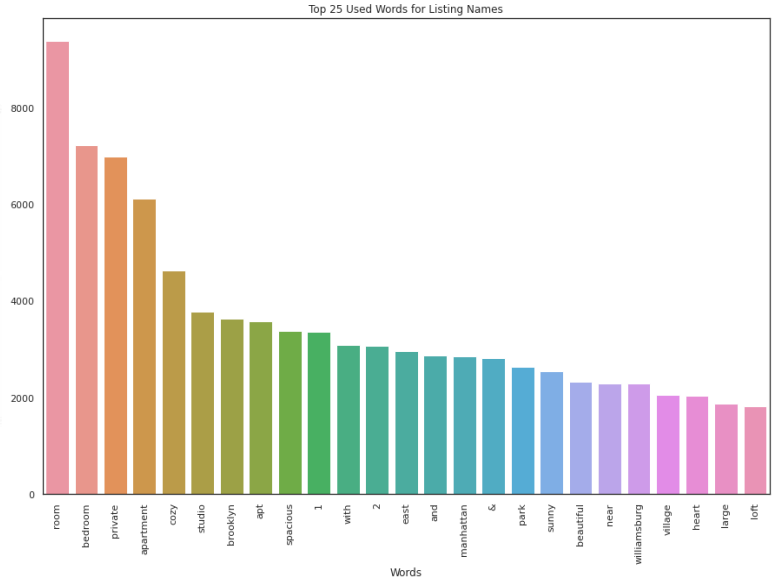
For this analysis I used plotly library for graph visualization. With the help of this I plotted room type proportion on all neighbourhood groups.

After this I used boxplot from Seaborn library to find the correlation between neighbourhood group and availability.



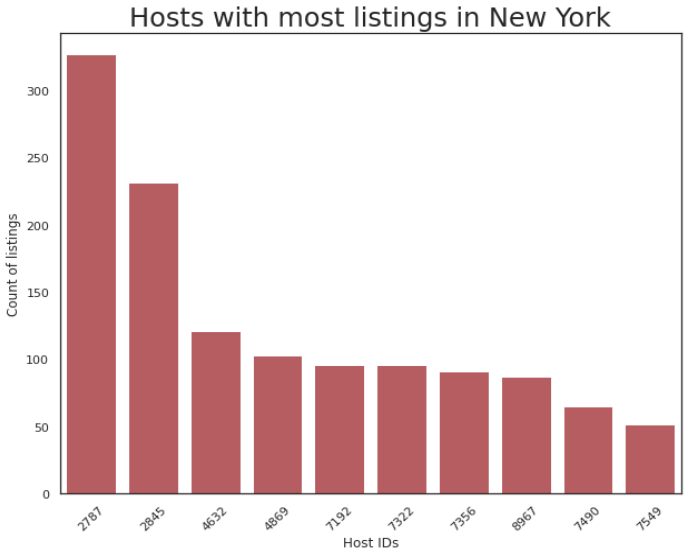
c. Top 25 most common words used in listing names

The most used words in listing names could represent the value of the property for the prospective guests. For this I used **counter** library to count and generate raw data which contains the top 25 words used by the host. Then I used WordCloud library for visualization purpose.



d. Top 10 hosts with most listings.

This is calculated by counting total number of values from the host_id column of each host. These insights can be used for making future predictions



III. Conclusion

Simply by performing EDA on the dataset, we can identify various new insights of the Airbnb New York business.

We now have the following insights from the analysis:

- **'Entire home/apt' room type has the highest number of listing of 52% and 'Shared Room' is the least listed room type at only 2.4% in total.**
- **People stay for longer duration of time in Private rooms in Brooklyn and Manhattan as compared to other regions.**
- **Words such as 'bedroom', 'cozy', 'private', 'apartment' and 'spacious' are used more frequently than words such as 'park', 'near', 'village' and 'heart'.**
- **Count of listing by top 10 hosts is almost 2.5% (1270 listings) of the whole dataset.**

These insights can further be used to make future predictions and make data driven decisions that are beneficial for the business.

Reference:

1. Wikipedia
2. Python Module
3. GeeksforGeeks
4. Stackoverflow