Airbnb Dataset Exploration and Analysis

Azhar Ali, Mohd Taufique, Aishwary Sharma, Pushpam Raghuvanshi
Data science trainees,
Alma Better, Bangalore

Abstract:

Airbnb is an online marketplace that connects people around the world who want to rent out their homes (i.e., hosts) with people who are looking for accommodations. We were given a dataset of Airbnb booking of New York city (NYC)-2019 that contains the columns like host name, host id property name, location, price, reviews etc. Data analysis on thousands of listings provided through Airbnb is a crucial factor for the company.

1. Problem Statement

Data provided by the international marketplace for renting and hosting properties. Hosts and people looking for accommodations both are their customers.

However, in this particular dataset, we will be analysing data only for hosts. Hosts can list their properties either on the app or on a website.

Airbnb has been in operation since 2008 and its presence across the world is well known. A study on their platform in 2019 reports that they provide accommodation in more than 100,000 cities and 191 countries.

The main objective is to find out the key metrics that influence the listing of properties on the platform, which could help them understand hosts better, give them more user-friendly suggestions, and help them expand their business further.

Description of columns present in the dataset:

- id: ID for property
- name: Names of properties
- host_id: ID for hosts
- host_name: Names of hosts
- neighbourhood_group: Neighbourhood of NYC like Manhattan, Brooklyn etc.
- neighbourhood: Neighbourhoods inside neighbourhood groups like Uptown neighbourhoods and West Harlem are in Manhattan.
- latitude & longitude: Latitude and longitude (Coordinate) of listed properties.
- room type: Types of room.
- price: Price per night in USD
- minimum nights: Minimum number of nights customers must book.
- number_of_reviews: Total number of reviews a property has.
- last review: Date of the last review.
- reviews_per_month: Total number of reviews upon the total number of months the property had been listed.
- calculated_host_listing_count: Total number of listings a host has on Airbnb.
- availability_365: Number of days a property is available for booking.

2. Introduction

Airbnb is a service that allows people to rent their homes to those looking for affordable accommodation around the world. The company also provides its users with what they call experiences — paid activities designed and led by locals — such as surfing at sunset, hiking with rescue dogs, and food tours.

New York City — one of the world's most famous cities — plays an essential part in international finance, politics, entertainment, and culture. It is not surprising that the home of Central Park and such wonderful museums, skyscrapers, and stores attract countless tourists all year long. And just like in other big cities, travellers can pick and choose from several accommodation options, including Airbnb.

We will dive into an Airbnb dataset from 2019 to learn about rental options and how they are distributed around the city's five boroughs: Manhattan, Brooklyn, Bronx, Queens, and Staten Island. In this project, I used Python to clean, analyse and visualise the data.

3. Workflow

We will divide our workflow into the following three steps:

- Data Collection and Understanding
- Data Cleaning and Manipulation.
- Exploratory Data Analysis (EDA).

4. Data collection and understanding

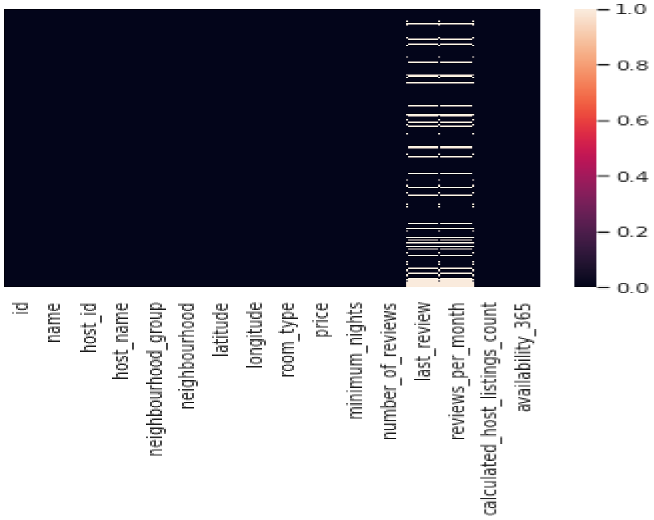
Data was obtained from the **Alma better** portal and stored as a CSV file in Google drive. This was later read in order to obtain the data. There were in total 16 columns wherein 3 columns were of float data type, 7 columns were of integer data type and 6 were of string data type.

5. Data Cleaning and Manipulation

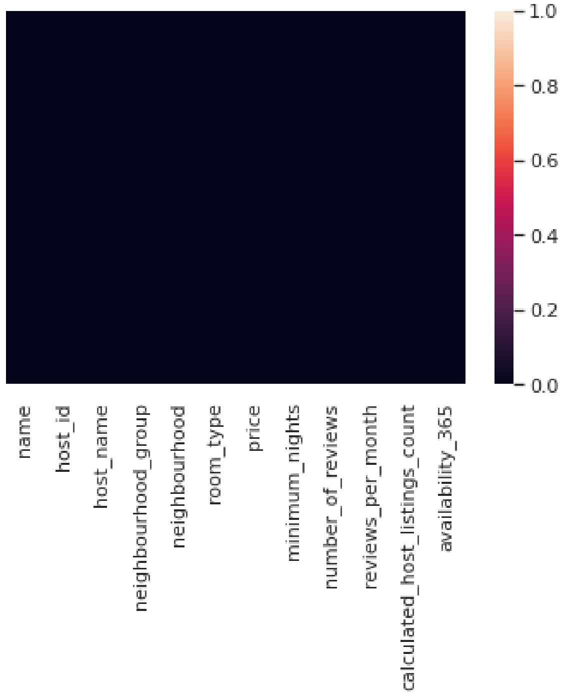
- Null values Treatment

The column-like review per month and last review has more null values. If there is no review then we can easily replace them with zero which is possible and valid as well. Since these are unnecessary columns, hence we dropped them at the beginning of our project in order to get a better result.

Data Visualisation using Heatmap before removing null values.



Data Visualisation using Heatmap after removing null values and dropping insignificant columns.



- There were almost 11 listings which have zero price listed. It is an anomaly because Airbnb does not provide free stays. So, replace this zero-price error with the mean of all listings with prices less than 100$.
- We also drop unnecessary columns like latitude and longitude.
- We checked for duplicates and removed them.
- There were a few 16 and 21 missing values in column name and host_name respectively. These are of string data type which means they store textual information. We imputed these missing values with the word ‘missing’.
- Extracts the most frequently used keywords from the ‘name’ column and stores them in a list. Removed the stop words and converted them into a data frame prior to analysis and visualisation.

6. Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) was then performed on the clean dataset to obtain certain observations and derive key insights. If a person coming to the Airbnb platform to book a listing/property for stay/rent, he/she will look for the following factors at the time of booking: -
name, host_name, neighbourhood_group, neighbourhood, room type, price, minimum nights, number_of_reviews. These are the important variables/features in our analysis and visualisation. These are the necessary columns and are a mix of categorical and numerical attributes.

- **Steps involved in our EDA-**

1. Importing Necessary Libraries – NumPy, Pandas, Matplotlib, Seaborn.
2. Understanding the Data
3. Data Wrangling
4. Data Analysis and Visualisation

- **Categorical features: -** Room type, neighbourhood_group, neighbourhood, name, hostnames

- **Numerical Features: -** price, number of reviews, availability_365, calculated_host_listing_count, minimum_night.

- **Univariate analysis (analysis of single variable)**

1. Most preferred room types
2. The list of busiest hosts
3. The most expensive area
4. Average night spends per neighbourhood group
5. The most popular host
6. The most popular areas/boroughs
7. The top 20 most frequently used keywords in listings.
8. Most preferred price by the guest.
9. Most affluent neighbourhood.

- **Multivariate/Bivariate analysis (analysis on two or more variables)**

1. Number of reviews vs price
2. The number of reviews for different neighbourhood groups.
3. Types of accommodation provided by top hosts.
4. Number of listings in a different neighbourhood group
5. Top 10 neighbourhoods across all neighbourhood groups with the highest listings.
6. Number of room types in different neighbourhood groups.
7. Percentage of properties in different neighbourhood groups.
8. Total number of minimum nights spent per room type.
9. Density and price distribution across different neighbourhood groups.
10. Availability_365 across all room types.
11. The average price of room type in different neighbourhood groups.

7. Correlation of all the Columns

Correlation shows the relationship between variables. Either it is positive, negative or zero. Its range is from -1 to +1. IF positive then directly proportional while negative values show inverse relation. Zero correlation means no association between variables.

- The number of reviews and reviews per month has a high correlation.
- The price column has a very low correlation with other attributes/features.
- There is no significant correlation between any other two variables.

8. Challenges Faced

1. To answer some of the questions we had to understand the business model of Airbnb and how they work.
2. Designing multiple visualisations to summarise the information in the dataset and successfully communicate the results and trends to the reader.
3. understanding the meaning of some columns.
4. Dealing with null values and duplicates.

9. Scope of Improvement

As the dataset has few qualifying attributes to value a property, more features can be added like bedroom, bathroom, property age (it might be one of the most important ones), the tax rate applicable, distance to nearest airport, hospital, or schools. This will be helpful in extracting more valuable insights which will be more helpful in efficient business decision-making. It will ultimately lead to the increment of company revenue as we know that intelligence is everything.

10. Conclusions/Analysis Summary

That’s it! Now, it is time to conclude the observations. The key takeaways/insights are mentioned below: -

- Manhattan and Brooklyn are the two most popular, expensive & posh areas of NY city. The average price variation is maximum for Manhattan
- Dona and Ji are the busiest hosts with the maximum number of reviews from the visitors/guests
- Most visitors don't prefer shared rooms, they tend to visit a private room or entire home/apt.
- Sonder (NYC) has the highest number of listings (327) in the entire NYC.
- 'room', 'private', and 'studio' are used most of the time by the hosts. Neighbourhood groups like Manhattan and Brooklyn are also present in this list among the top 20 keywords.
- Around 44.3% of properties are listed in Manhattan followed by 41.1% in Brooklyn. Staten Island has a minimum number of listed properties.
- Most guests prefer a budget price for their stay.
- Williamsburg is the neighbourhood with the highest number of listings in NYC.
- Percentage of nights spent is maximum for the entire home/apt with 62.8% followed by a private room with 35%.
- The average number of nights spent is maximum for Manhattan Island which is more than 8 nights and the minimum is 4 nights for the Bronx**.**
- Listings by top 10 hosts is almost 2.5% (1270 listings) of the whole dataset.

References: -

1. Kaggle.com
2. Stackoverflow.com
3. Medium.com