Airbnb Bookings Analysis

Rudrajit Bhattacharyya
Cohort Zanskar
AlmaBetter

Abstract:

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. Hosts and travelers use Airbnb platform to list their properties and book their stay respectively.

The dataset consists of 49000 records of Airbnb listings in New York City and my task was to explore and find insights about hosts, neighborhoods, and room types which can help the company, hosts, and travelers to make better decisions. The main purpose of EDA is to detect any errors, outliers as well as understand patterns in the data. EDA on the Airbnb listings data will help us understand many insights about the hosts, neighborhood groups, room types, prices etc.

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 hosts' behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

The dataset has around 49000 observations and 16 columns in it which is a mix of categorical and numerical variables. The main objective here is to explore and analyze the data to find out insights such as:

- What can we learn about different hosts and areas?
- What can we learn from predictions?
- 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?

EDA will help us generate such insights not limited to these which will help the company, hosts and travelers to make better decisions in future.

The different features the dataset contains:

- **id:** It is an unique id given to the property listed in airbnb NYC which is a numerical variable.
- **name:** It represents the name of the airbnb listed property which is a categorical variable.
- **host_id:** This is an unique id given to the host of the property which is a numerical variable.
- **host_name:** The name of the host of the property listed which is a categorical variable.
- **neighbourhood_group:** This represents a big neighborhood inside which there are many mini neighborhoods which is a categorical variable. There are 5 neighborhood groups in the data:
    - Manhattan
    - Brooklyn
    - Staten Island
    - Queens
    - Bronx
- **neighbourhood:** This represents all the mini neighborhoods present in NYC which is another categorical variable.
- **latitude:** latitude coordinates
- **longitude:** longitude coordinates
- **room_type:** This represents the type of room in the listed property which is a categorical variable. There are three room types available in the dataset:
    - Entire Home/Apt
    - Private Rooms
    - Shared Rooms
- **price:** Represents the price per day of stay in the respective listed property which is a numerical variable.

- **minimum_nights**: This represents the minimum number of nights a person has to pay for or stay in the property which is a numerical variable.
- **number_of_reviews**: The number of reviews given to the property and the host which is a numerical variable.
- **last_review**: The date of the last review given which is a datetime object.
- **reviews_per_month**: The number of reviews given over a month to a property or the host which is a numerical variable.
- **calculated_host_listings_count**: How many listings a particular host has in NYC which is another numerical variable.
- **availability_365**: Availability of the property out of 365 days which is also a numerical variable.

Introduction:

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. The company has come a long way since 2008, when its co-founders first came up with the idea to invite paying guests to sleep on an air-mattress in their living room. According to Airbnb's latest data it has in excess of 5.6 million listings covering more than 100000 cities and towns and 220+ countries worldwide.

The idea behind Airbnb is simple:

- Hosts list out their property details on Airbnb along with other factors like pricing, amenities provided etc.
- Airbnb sends a professional photographer (if available) to the property location in order to take high quality photographs.
- Travelers search for a property in the city where they wish to stay and browse available options according to the price, amenities etc.
- Booking is made through Airbnb where travelers pay the amount mentioned by the host and some additional money as transaction charges.
- Host approves the booking, the traveler stays there and finally Airbnb pays the amount to the host after deducting their commissions.

Airbnb's business model is quite profitable. The company like Uber, Lyft and others has capitalized on the sharing economy, essentially making money renting out property that it doesn't own. Every time a reservation is made, Airbnb takes a cut.

Exploratory Data Analysis is a set of techniques that were developed by Tukey, John Wilder in 1970. The philosophy behind this approach was to examine the data before building a model. Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques.

The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. It allows analysts to understand the data better before making any assumptions. The outcomes of EDA helps businesses to know their customers, expand their business and take decisions accordingly.

My goal here is to discover insights which will help

- The hosts know which areas receive more traffic, what kind of rooms are preferred by travelers, how many nights travelers generally book for etc,
- The travelers know which hosts receive the most reviews, which neighborhood costs how much, what room types are available, which properties are available out of 365 days for booking etc,
- The company figure out which areas need to be optimized, what amenities need to be included etc.

Approach:

To accomplish the above task, it was divided into two parts as follows:

- **Data Cleaning:**
  When it comes to data, there are many different sorts of quality issues, which is why data cleaning is one of the most time consuming aspects of data analysis. Formatting issues, missing values, duplicated rows, spelling discrepancies, and so on could all be present. These difficulties make data analysis difficult, resulting in inaccuracies or inappropriate results.

  The first step was to look for duplicate values in the dataset. Thankfully, there were no duplicate values present.

  The second step was to look for missing values and there were 4 columns which had a good number of missing values present. Missing values are

usually represented in the form of NaN or NULL or NONE in the dataset. Missing values can be dealt with any of the following techniques:

1. Deleting the columns with missing data
2. Deleting the rows with missing data
3. Imputing the missing data with an appropriate value
4. Imputing the missing data with an additional column

There were few columns with missing values which were irrelevant for our analysis so I decided to drop those columns and imputed a few columns with an appropriate value.

- **EDA:**
  EDA is used to take insights from the data and the outcomes help businesses to know their customers, expand their business and make decisions accordingly.

  The third step was to explore and perform some univariate, bivariate and multivariate analysis to discover insights from the data. This step helped me figure out that the 'price' column was heavily skewed and there were few observations with 'price' listed as 0 which cannot be true. Assuming that no one is giving Airbnb stays for free, the data containing 'price' as 0 was excluded from the analysis.

  The process of EDA helped uncover a lot of insights about different things which were not known by looking at the data.

Results:

A lot of insights were generated through the help of EDA which will definitely assist in better decision making. Let us summarize a few of the important insights that were discovered.

Insights about the different hosts:

- There are 37455 unique hosts in NYC.
- Sonder (NYC) has the most multiple property listings in NYC. Despite having the most number of properties she isn't the busiest host in NYC.

- Maya who receives the most number of reviews is the one who is the busiest of all in NYC. She has 5 properties listed in the same neighborhood and there are multiple reasons acting in favor of receiving the most number of customers.
    - The price at which she offers her properties is less than the average price of all neighborhood groups.
    - The condition for minimum nights is one, which is way less than many others.
    - There is enough availability of her properties out of 365 days.
    - High number of reviews will definitely help customers make a booking.
- There are many hosts who don't receive a huge number of customers but still stay occupied for the whole year. This could be due to the reason that the customers they are getting are staying for a longer period of time keeping their properties occupied.

Insights about the different areas:

- Manhattan has the most number of properties listed followed by Brooklyn. These two neighborhoods are the most expensive and receive the most number of customers or traffic as compared to others.
- Staten Island being mostly a residential area has the least number of properties listed and receives the least number of customers. Interestingly, the average prices of Private Rooms in Staten Island are the least expensive and stay available for a good number of days which makes a good choice for customers seeking low cost accommodations.
- Manhattan contains the most number of Entire Home/Apt and in general they are costlier than any other room type across all neighborhoods.
- Bronx is the least expensive neighborhood and less preferred by customers.

Insights about the different room types:

- There are 25407 Entire Home/Apt and 22319 Private Rooms which is way higher than the number of Shared Rooms available.
- Entire Home/Apt is the most expensive room type and the most preferred by customers.
- Entire Home/Apt and Private Rooms receive way more customers than Shared Rooms across all neighborhoods.
- Shared Rooms are the least preferred room type and remain the most available out of 365 days.

Challenges:

There were no as such huge problems faced with the dataset given as it was mostly a cleaned dataset ready for analysis apart from a few things like:

- Missing values which had to be dropped or imputed.
- Manipulating the price column to exclude the records which had price as 0.

Conclusion:

We have reached the end of our analysis of Airbnb listings in NYC. We started from looking out for duplicate values, then missing value treatment and finally used EDA to discover many insights from the dataset. To summarize few of the important insights we gathered:

- Host Maya is the busiest host in NYC and there are multiple reasons in favor of it like price, minimum nights, availability and number of reviews. She has a total of 5 properties listed in the same neighborhood.
- Manhattan and Brooklyn are the most expensive neighborhoods and they receive the most traffic as well. Due to many tourist attractions and the number of properties available, people tend to visit these two areas comparatively more than other ones.
- Entire Home/Apt is the costliest room type available but still the most preferred ones for the customers. Entire Home/Apt and Private Rooms receive way more traffic than Shared Rooms and as a result Shared Rooms stay available for most of the time out of the 365 days.

These insights generated can definitely help everyone make better decisions in future to enhance their experience of staying in an Airbnb in NYC.

References:

1. Analytics Vidhya
2. Investopedia
3. Stack Overflow