**Airbnb Booking Analysis**

**Sharath Diwakar**

**Data Science Enthusiast**

**AlmaBetter, Bangalore**

**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 on in terms of improving their services. The EDA can also give the customers an idea on 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**

Airbnb generates 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.

Explore and analyze the data to discover key understandings:

* What can we learn about different hosts and areas?
* What can we learn from predictions (ex: locations, prices? reviews)
* 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??

1. **Introduction**

[Airbnb](https://www.airbnb.co.in/) is a $75 Billion online marketplace for renting out homes/villas/ private rooms. The website charges a commission (3 to 20 percent,) for every booking. Even though the prospects are sound, but there are critics who argue that this has driven up rent, and caused damage to the local communities living in the vicinity.

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

How is Airbnb affecting the neighborhoods? [Insideairbnb](http://insideairbnb.com/about.html" \t "_blank)is an activist project, which has curated this dataset, to measure the impact of rentals 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.

1. **Airbnb Booking Dataset Insight**

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**

* name = Description about the listings.
* host\_id = unique id for each listed host.
* host\_name = Hostname for the listings.
* neighbourhood\_group = Location
* neighborhood = Area
* latitude = Latitude coordinates
* longitude = Longitude coordinates
* room\_type = Listing space types
* price = Price in dollars
* minimum\_nights = minimum nights required to stay
* number\_of\_reviews = No. of reviews written for the listing
* last\_review = Last reviewed date for the listing
* reviews\_per\_month = Total review per month for the listings
* calculated\_host\_listings\_count = Total no of listing against the host id
* availability\_365 = Available days of a listing in a year.

1. **Steps involved**

* **Setting up the notebook**

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

* **Cleaning the Dataset**

**Null values:** The dataset contains a good amount of null values. The columns name, host name & reviews contained large number of null values. So, we tried to preserve as many rows as possible by replacing null values with suitable values.

**Data with errors:** These data were considered as erroneous data and removed

.

* **Exploratory Data Analysis**

During the preparation of the dataset for EDA, a basic statistical summarization was done. The information regarding the data types of each column was explored.

This was followed by a detailed exploratory data analysis. The distribution of bookings with respect to different categories of features was visualized. Relationship between variables was observed.

1. **Exploratory Data Analysis**

Throughout the analysis, we tried to answer questions that help us understand the factors determining the booking trends.

1. **What is the price of each room type?**

The mean price of the entire room/apartment type was found to be $211, the mean price of a private room was $89 followed by a shared room which was $72.

1. **Which was the most wished/booked room type?**

Entire home/apt is the most wished/booked, followed by private room and shared room

So we can say that, when people go on vacation with their loved ones they prefer staying in an Entire home/Apartment followed by a private room, who wish for some privacy And people on a budget prefer shared rooms.

1. **Price analysis of each room type in all neighborhood groups.**

All the groups have Entire home/apt as the most expensive commodity. Manhattan and Brooklyn agree with the general price trend completely, while for Bronx and Queens, the price for shared rooms and private rooms are almost equal and we see a trend reversal in the price of shared rooms and private rooms in Staten Island. We can also see that Manhattan is the most expensive for all room types by a considerable margin.

1. **Total number of listings of all neighborhood groups?**

Manhattan has 21661 listings which is the highest of all, followed by Brooklyn which has 20104 listings, Queens having 5666, Bronx having 1091 and at last Staten Island having 373.

1. **What is the density and location of all the properties?**

With less area and the highest number of properties, Manhattan is the densest of all followed by others. we can concur here that Manhattan, despite having the highest number of listings, still has the highest price, shows high demand for this group

1. **Who is the most successful host?**

Sonder (NYC) is the most successful host andhas about 327 listings which is 30.61% of the total listings and these are present only in Manhattan with just private rooms and apartments. They deal with the most expensive aspect of the business, hence possess more money, which in turn increases their number of listings.

1. **Which is the most expensive neighborhood in each neighborhood groups?**

* . The most expensive neighborhood in Brooklyn is Sea Gate.
* The most expensive neighborhood in Manhattan is Tribeca.
* The most expensive neighborhood in Queens is Neponset.
* The most expensive neighborhood in Staten Island is Fort Wadsworth.
* The most expensive neighborhood in Bronx is Riverdale.

**6. Conclusion**

We were able to answer some really important questions about the bookings analysis using this dataset. By studying the trends of the past bookings, we can take steps to satisfy the customer demands and understand where to focus the efforts the most to make the most positive results.

> Entire home/apt is the most expensive

room type and yet gets the most traffic.

> Private rooms are the second highest in

terms of booking, which suggests people

value their privacy.

> Manhattan is the most expensive

 neighborhood group, and Brooklyn gets

the most traffic.

> Price has a moderately negative

correlation with availability.

> Sonder (NYC) is the most successful host, who only caters in Manhattan and only lists apartments and private room and is also located in financial district.

The hotels can increase connections with travel agencies (especially online ones) in order to get access to a wider pool of customers as we know that they’re the most common intermediary.

The customers can also make their choices based on the observations to get the best properties. For instance, they can find out the best property at a good price.

This analysis resulted in a lot of knowledge which helps the hotels to grow their business. In addition, it can potentially help new neighborhood to have a clear understanding of the field while listing their properties.

**References**

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