**Airbnb Booking Analysis**

**G MOHAMED LUQMAN**

**Data Science Enthusiast, Cohort - Berlin**

**Alma Better, Bangalore**

**Abstract:**

Airbnb is an online marketplace that  
connects people who want to rent out their  
homes with people looking for  
accommodations in that locale. We were  
provided the Airbnb dataset which has large  
number of observations in New York City.  
Analysis was done on this data set based on  
different parameters

Our EDA can make us understand about the  
famous place to live, the famous host, the  
affordable place to live based on different  
parameters.

We’ve performed project right from proper data cleaning to analysis a problem statement. For better understanding we’ve accomplished project in General and Business Analysis. The task of analysis has helped a lot to understand the technical growth of a company like Airbnb industry. Working on different view of analysis to make it more reliable which is been provided in this technical documentation.

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 conclusions from this EDA can benefit who want to do business or who want to market their product. Important inferences have been provided throughout analysis in the collab notebook. This EDA will also help common people or customer to make choice decision which room to take according to their price, availability etc

***Keywords: Airbnb, Data Cleaning, Exploratory Data Analysis***

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.

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 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* 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.

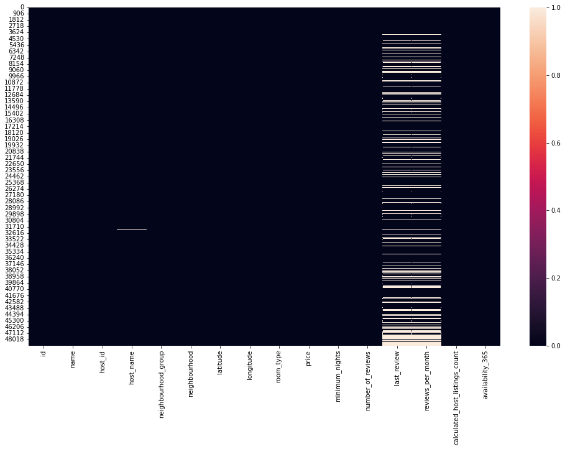


Fig.1 Heatmap of missing values

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

* **Data Cleaning**

Here in the below table, we can  
easily identify the presence of  
outliers for column 'Price'. There we  
can see that the min price is 0$ and  
maximum price is 10000$.

After doing some background  
research it is admissible that Airbnb  
customers mostly searched for rent  
between 20$ to 800$.

Removal of outliers if possible will  
give new ways to look into the data.  
Two graphical techniques for  
identifying outliers are scatter plots  
and box plots initially.

* **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. **Price Analysis:** The first focus we put on is ‘Price’. Here first we looked in to average price of different room type across New York. From this we get to know that costly room type is ‘Entire home’. Then we looked at average price of room type according to different neighborhood group. From this analysis we made inference that if a salaried employee wants to increase his saving then he will prefer to work in Bronx. Then we did few more on analysis on finding cheapest neighborhood and the cheapest listing throughout New York.

**2. Listing Analysis:** Here we focused on different listing. In this we take a look at listing according to their neighborhood group. From the result we made inference that is someone want to do advertisement or marketing he should focus on Manhattan and Brooklyn. Then we deep dive in data and looked different listing according to neighborhood group. We get to know that in Manhattan, entire room type is highly listed.

**3. Availability analysis:** Here we focused on availability of different room according to their neighborhood group and then the average availability of different room type. We get to know that private room has highest availability and entire home has least availability. The inference which come out from this result if host is having entire room then he will be making good money. But from customer point of view private room is the best as half of the year it is available.

1. **Profitability analysis:** To confirm the above inference we created some new columns which talks about revenue generated by different host and their property. First we did analysis on revenue generation of different room type by their neighborhood group. Seeing the result, we were astonished that irrespective of any neighborhood group, entire home is making way ahead revenue then other room types. Then we deep dived and look in to which host is having highest no. of property, which neighborhood group has highest review etc.
2. **Question & Answer**

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

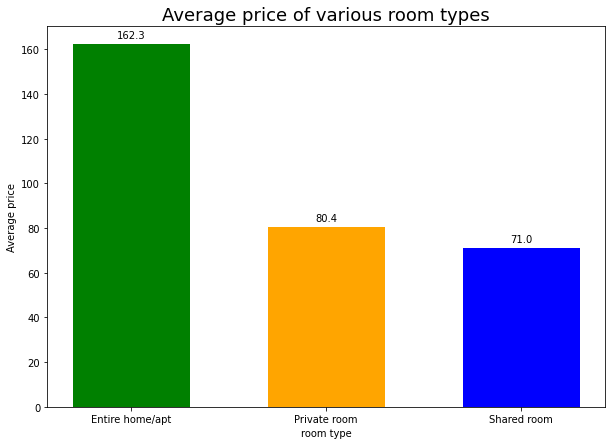


Fig 2. Most wished room type

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.

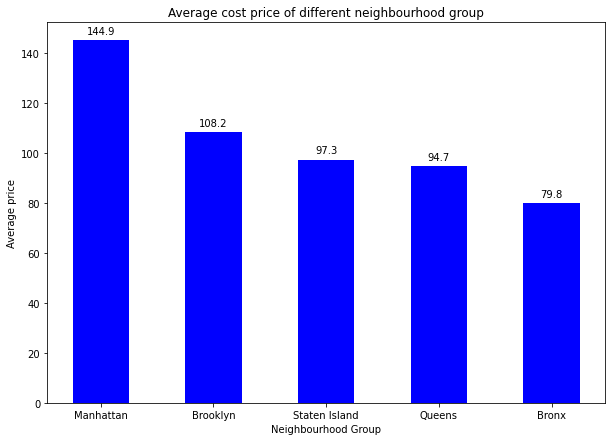


Fig 3. Price analysis

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.

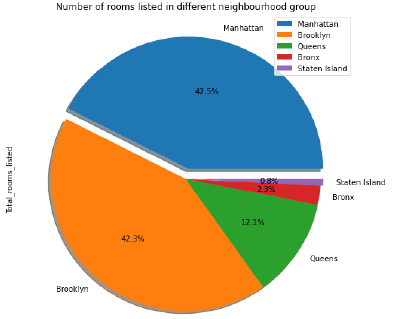


Fig 4. Total no of listing

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

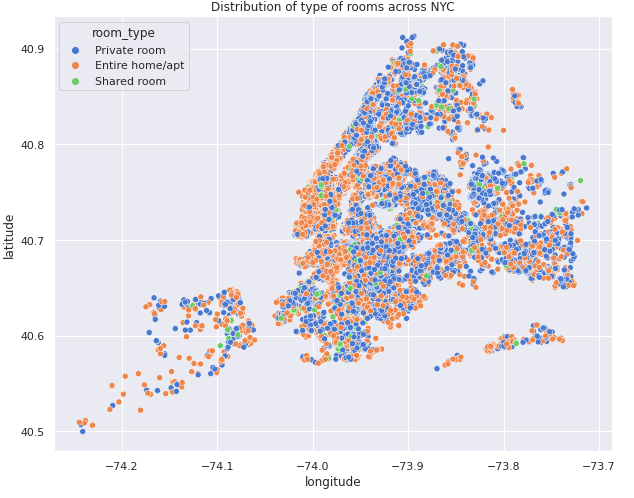


Fig 5. Density of properties

1. **How to find out which neighbourhood is costlier for a stay?**

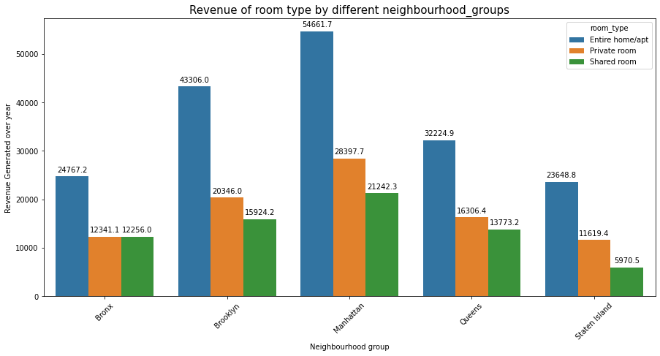


Fig 6. Top host

By finding out the mean price for each neighborhood group, anyone can compare neighborhood groups. In the given data set we found that Manhattan is costlier and Brooklyn is cheapest neighborhood group as per the mean price of various room type.

1. **How to know which room type is mostly available?**

It can be done by grouping room type and find the mean of the availability365 according to the room type. In the data set we could find it that;

1. Private room has highest mean of

availability.

1. Entire home has least mean of availability.

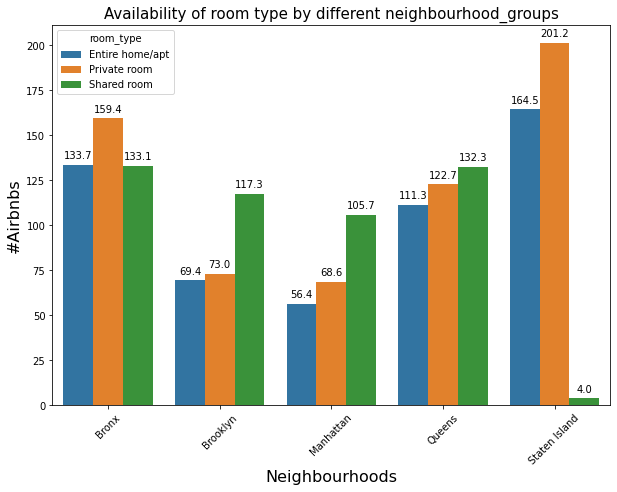


Fig 7. Most expensive in neighborhood

**7. Conclusion**

Finally 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.

We were able to answer some really important questions about the bookings analysis using this dataset.

1. Entire home/apt is highly expensive.
2. Manhattan living cost is highest, Bronx living cost is lowest.
3. Cheapest neighborhood is Bulls head.
4. Cheapest listing is Bronx apart.
5. Manhattan have highest no. of listing.
6. In Manhattan entire home is mostly preferred but in Brooklyn ratio between entire home and private room is 50:50.
7. Private room has highest availability; Entire home has least availability.
8. Revenue generated by Entire home is highest irrespective of neighborhood group.

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.

**References**

1. [*https://towardsdatascience.com/*](https://towardsdatascience.com/)
2. [*https://stackoverflow.com/*](https://stackoverflow.com/)
3. *https://medium.com/*