Airbnb Bookings Analysis

Somnath Patnaik, Subham Choudhary,

Akshada Dani, Pritesh Ashok Lonkar

Data Science Trainees,

AlmaBetter, Bangalore

**Abstract:**

Airbnb is a social business. Hosts offer their homes to guests in exchange for money. Visitors document their experience online. Social media has allowed stories to spread. And Airbnb grown through word of mouth. In this project, we analyzed 49,000 listings around New York city and drew some conclusions which might help Airbnb improve their business model. The dataset provided contains information about hosts or guest’s behaviour like which hotel room they chose, what was their review during their stay and for how long they stayed and also about hotel rooms like room types, availability throughout year, etc.

Before the exploration of data, few problems had to be dealt with. The dataset was cleaned such as handling of null and missing values. The cleaned dataset was used to explore various variables and relationship among variables were established. Many plots and charts were used to explain and visualize the relationship between variables. With this exploratory analysis, we came across few conclusions that might help Airbnb increase their customer base and also retain their previous customers by facilitating to their needs.

*Keywords: EDA, Airbnb, data exploration, data visualization, Manhattan, Brooklyn*

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 entire 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.

This dataset has around 49,000 observations in it with 16 columns and it is a mix of categorical and numeric values. Explore and analyze the data to discover key understandings.

1. **Introduction**

We live in an era where lots of data is generated and circulated among various parties. Raw and unprocessed data isn’t of much use unless we derive insights from it. “Torture the data, and it will confess to anything.” – Ronald Coase. If someone is familiar with data science, one can understand the power of the above statement. EDA is the process of summarizing important characteristics of data in order to get better understanding of dataset. However, as the number of data increases, we need to visualize the data to help us in conducting data analysis. By using visualization tools, we can deliver a message to our audience and inform them about our findings.

Airbnb is a unique business. People offer to rent their homes, apartments, or rooms to strangers. It relies heavily on technology. Visitors will book rooms on the app or through the website. Visitors will document their experience online.

In this project, we will explore and visualize the dataset from Airbnb in New York city using basic exploratory data analysis techniques. We shall be looking for distribution of every Airbnb listing based on their location, including their price range, room type, listing name, and other related factors.

Our goal here is to explore the data and find useful insights from the data and find out different relations between the columns.

Basic descriptions of columns after univariate analysis are:

* id: Unique serial number
* name: Description given to each accommodation.
* host\_id: Unique serial number given to each host.
* host\_name: Name of every host.
* neighbourhood\_group: Various boroughs(town/district) within New York city.
* neighborhood: Various divisions within each neighbourhood\_group.
* latitude and longitude: Geographic coordinates that specify the position of a particular location.
* room\_type: Variety of rooms depending on the size.
* Price: room price in dollars
* minimum\_nights: minimum nights customer stayed.
* 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. **Methods involved**
   1. **Acquire and Loading Data**

This project is done using Google Colaboratory notebook with python programming language to write down the code script. Colab notebooks are Jupyter notebooks that run in the cloud and are highly integrated with Google Drive, making them easy to set up, access, and share.

Airbnb dataset was made available to us from Almabetter EDA capstone project section. The dataset contains information about the hosts’ and their behaviour, various hotel rooms and apartments, its location and their reviews. A total of around 49,000 observations was in it and 16 columns with a mix of numerical and categorical variables. These information was enough for us to make predictions and draw conclusions.

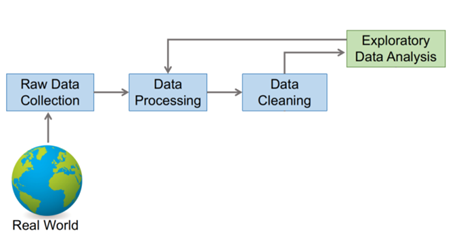


Fig. 1: Flow chart of Exploratory Data Analysis

* 1. **Data Cleaning**
* **Null and missing values** – Using the code isna().sum(), we found that there are few null values in host\_name and name columns. These null values were removed using the dropna() function in the respective columns. It was also found that 1/3rd of observations in last\_review and reviews\_per\_month had null values. Instead of removing all such observations, we replaced it with zero value. We also looked for infinity value, but there was none. Some columns were dropped because it didn’t contribute much to exploration of data.
* **Outlier detection** – The distribution of price column varied from 0 to 10000, whereas the mean price was around 150. Also, some other columns like minimum\_nights displayed similar skewed distribution. We used boxplot to clearly visualize the values outside inter-quartile range. Here, we don’t need to remove outlier cause in EDA we don’t need a decision boundary to distinguish between various categories.
* **Cleaned dataset –** The dataset after removing the null and missing values contains around 49,000 observations and 13 columns. Thus, there won’t be much change in the behaviour of dataset.
  1. **Data exploration and visualization**
* **Host name –** The first feature to be analyzed is host\_name. Using nunique() code, we found there are 11452 unique hosts. Then, we used groupby() with price to find out top 10 most affluent hosts. We also found that among these 10 hosts what are their preferred destinations in New York city and also what room type these chose to spend their stay with.

From the below figure, it was sufficiently understood that Sonder(NYC) is the most affluent host and his most favored destination is Manhattan. His preferred choice of stay is mainly Entire home/apt.

Blueground is the second most affluent host. Its taste is quite similar to Sonder(NYC).

Michael, David, Alex, Jessica and John are the next 5 most affluent hosts. They had visited almost every neighbourhood group, in which most of their visit is in Manhattan and Brooklyn. Each one of them spend their stay mostly in Private room and Entire home/apt.

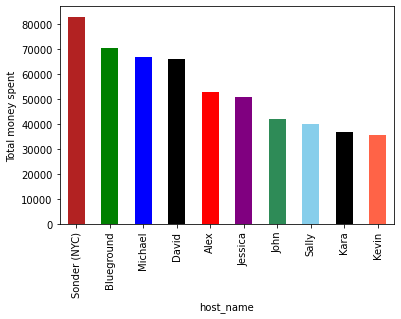


Fig. 2: Top 10 most affluent hosts

* **Neighbourhood\_group –** In NY city, there are 5 neighbourhood\_group (Manhattan, Brooklyn, Queens, Bronx, Staten Island). We tried finding out which group has most number of accommodation and what type.

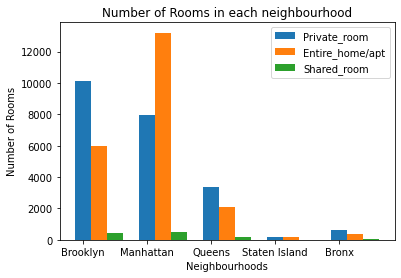
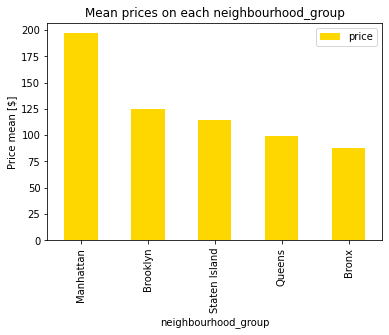
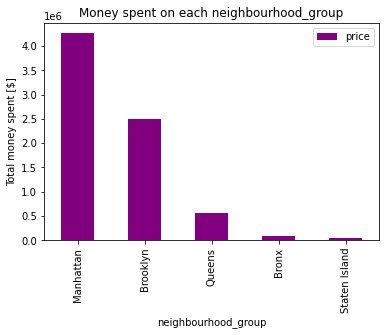
****

Fig. 3: Number of types of room in each neighbourhood group

From the above plot, we can see that most number of accommodations are concentrated in Manhattan and Brooklyn.

Using groupby() with total price and mean price we found that Manhattan is the most costliest, followed by Brooklyn, Queens, Bronx and Staten Island respectively. We also found that in each neighbourhood\_group Entire home/apt is most famous, which also shows most hosts comes with their family to stay or spend vacation together.



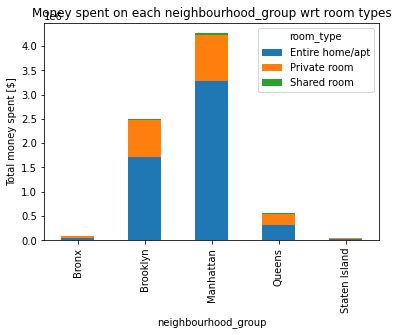


Fig. 4 and 5: total and average money spent on each neighbourhood\_group, Fig. 6: Distribution of money in each neighbourhood\_group wrt room\_type

* **Neighbourhood** – There are total of 221 neighbourhood in NY city. We would be exploring about this variable by first considering which neighbourhood got most number of listings. Using value\_counts(), we found that Williamsburg is the most sought out destination with 8% of listings (i**.e. 3919 bookings)**. We also found top 10 neighbourhood wrt number of listings.

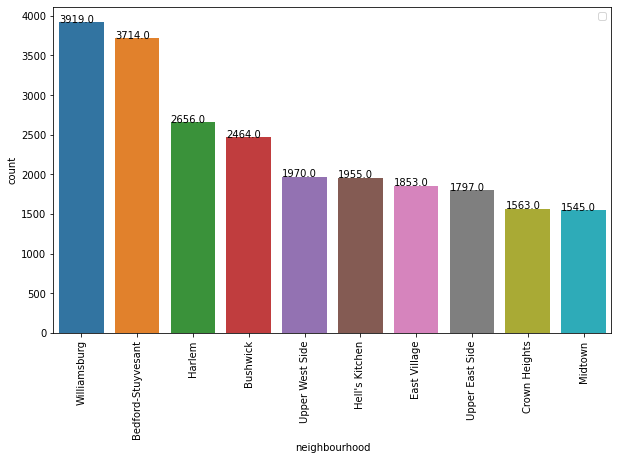


Fig. 7: Top 10 neighbourhood wrt number of listings

As we got to know from previous observations that Manhattan and Brooklyn is the most sought out destinations. So, lets see in both destinations what are the most favored neighbourhood. From the below plot, we can see that Harlem in Manhattan and Williamsburg in Brooklyn are the neighbourhoods with most number of listings.

We also did some comparative analysis between Williamsburg vs other neighbourhood’s mean prices wrt room types and we found that there is almost no difference between the prices of rooms in Williamsburg with respect to other neighbourhoods.

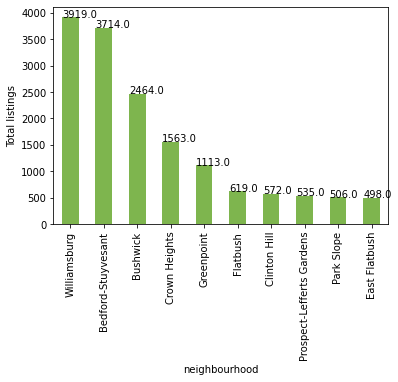
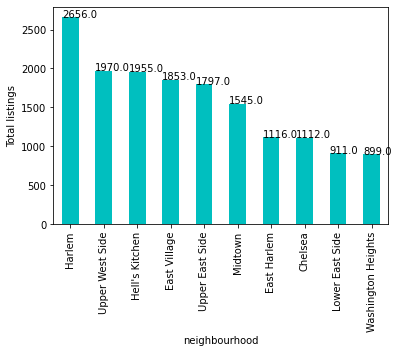


Fig. 8 and 9: Top 10 neighbourhood in Manhattan and Brooklyn in terms of number of listings respectively.

* **Prices –** Prices of rooms varies from 0 to 10000. In order to visualize the number of bookings occur in which price range, we divided the price in 3 groups: prices below 100 is one group, prices between 100 and 200 is in another and above 200 is the third group. The results was fascinating as almost 80% of the bookings is done below 200 of the price range. Here’s a pie chart to visualize it clearly.

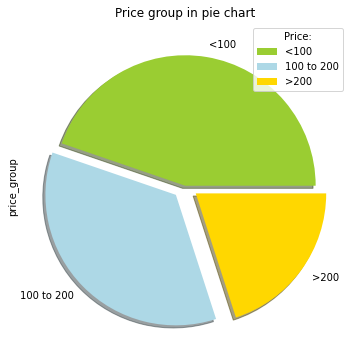
****

Fig. 10: Distribution of prices

We also saw an interesting plot where we saw at what price bracket people like spending more days in the hotels. People spent more nights at hotels at prices below 200 whereas people checkout early in expensive hotels.

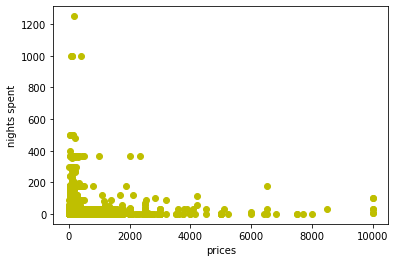


Fig. 11: Distribution of prices wrt nights spent

We again tried exploring more on Manhattan and Brooklyn. We found the most and least expensive neighbourhood in both destinations. Tribeca and Sea Gate are the most expensive neighbourhood in Manhattan and Brooklyn respectively, whereas Inwood and Borough Park are the least expensive neighbourhood in Manhattan and Brooklyn respectively.

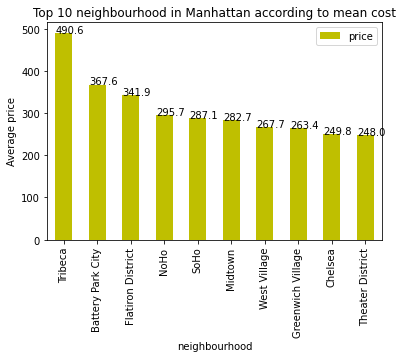


Fig. 12: Top 10 most expensive neighbourhood in Manhattan

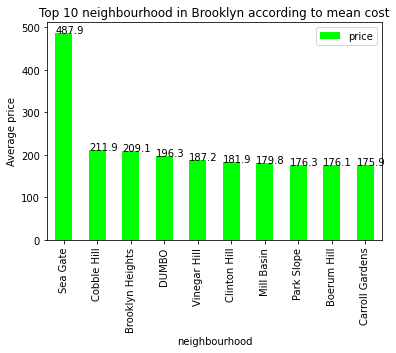


Fig. 13: Top 10 most expensive neighbourhood in Brooklyn

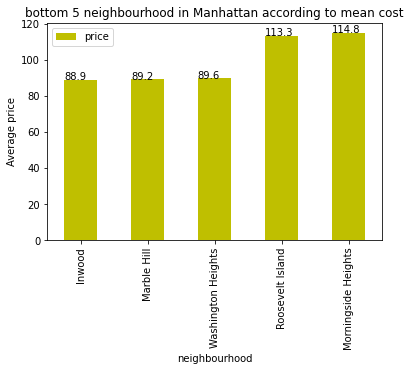
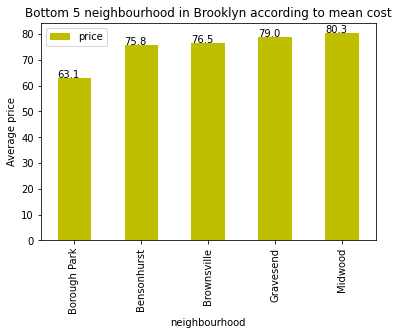


Fig. 14 and 15: 5 least expensive neighbourhood in Manhattan and Brooklyn respectively

We analyzed the footfall by number of rooms booked in most and least expensive neighbourhood in both Manhattan and Brooklyn. The results were similar, as in both destination people staying is less expensive neighbourhood.

1. **Conclusion**

We can finally conclude our data exploration with our final concluding points:

* Sonder(NYC) and other nine is our top priority customer.
* Most preferred destination is Manhattan and Brooklyn.
* Most booking is done in Williamsburg. From our research, we found that Williamsburg is famous for beaches and museums, so its quite evident that most people come here as recreational purposes.
* From our analysis, people more likely to live in less expensive neighbourhood.

1. **References**
2. Geeks for geeks
3. Analytics Vidhya
4. Stack overflow
5. <https://pandas.pydata.org>
6. Data science made simple
7. Code gripper