Almabetter Capstone Project: 01

Exploratory Data Analysis on Airbnb Bookings Dataset

Siddhartha Pasayat , Pankaj Beldar
Data Science Trainee, Almabetter, Bengaluru

Abstract - Airbnb has transfigured the hospitality industry forever by introducing a peer to peer hospitality company. Prior to 2008, travelers would have likely booked a traditional hotel or lodging for their stays. Nowadays, many of these same people are opting for Airbnb where they love the idea of staying at a residence of a stranger giving a very human touch to the living experience and an authentic one where they have the liberty to book any type of imaginable lodging possible from tree houses, boats,castles,igloos to private rooms.The idea behind Airbnb is quite straightforward where you are encouraging locals to rent out their extra space or room to people visiting the area for some extra money. This not only gives the customers a holistic and homely experience but also allows the hosts to connect with a lot of people across the globe helping in building a community based on trust, empathy and experiences. Hosts use this platform to list and advertise their rentals to millions of people worldwide, with the reassurance that the company will handle payments and offer support when needed. And for guests, Airbnb can offer a homely place to stay that has more character, perhaps even with a kitchen to avoid dining out, often at a lower price than what hotels charge.The basic phenomena of renting lodging to guests visiting the area have existed for centuries but the presence of internet and other digital gizmos has scaled the business model drastically ensuring a strong trust building network between hosts and guests.

Keyword - exploratory data analysis,price, minimum nights, review,neighbourhood

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) behaviour 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 between categorical and numeric values By performing Data Analysis on Airbnb dataset, we are going to figure out following questions-

1. What is the preferred location according to the average best price?
2. Where are most of the hosts located?
3. The highest and lowest rent paying locations by customers

4. Most popular/demanded host based on reviews and calculated host listings count
5. Establishing relation between neighbourhood group and availability of rooms
6. Which are the top hosts, neighbourhoods, neighbourhood groups based on their turnover?
7. Room type selection based on price, availability on 365 days
8. Top ten neighbourhood based on listing price
9. Distribution of properties based on mandatory stays.
10. Type of visit based on mandatory stay allowed for a single visit.

2. Introduction -

What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is an approach to **analyze the data using visual techniques.** It is the very first step done in the data analysis process before moving to any advanced modeling.It is used to unravel what data can reveal beyond the formal modeling or hypothesis testing tasks to understand various relations between the dataset features by understanding trends, patterns and their graphical representations. EDA is a thorough examination meant to unveil the underlying structure of a data set and is important for a company because it helps to **scrutinize data** much before any assumptions are made.It can help identify erroneous informations, outliers, anomalous informations and establish relations between features which can be useful in advanced modeling.

Types of exploratory data analysis-

The four types of EDA are-

1.Univariate non-graphical

Dealing with a single variable and main purpose is to describe data and find patterns. Doesn't work with establishing relations or causes.

2.Univariate graphical

Graphical methods are used to provide a full picture of the data. Eg - Stem and leaf plots,Histograms,Barplot,Boxplot etc.

3.Multivariate non- graphical

Dealing with more than a single variable and main purpose is to establish relations or causes between the variables through cross-tabulations and statistics.

4.Multivariate graphical

Uses graphical aid to express relation between two or more datasets.

Ex - Scatterplot, Heatmap, Bubble Chart etc.

Techniques and Tools:

There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques.

Basic tools used are -

1. Python - An interpreted, OOP language combined with dynamic typing and binding making it useful for handling missing values and other exploratory analysis of the data set.

2. R - An open source programming language used for statistical computing and graphical observations.

Typical graphical techniques used in EDA :
Box plot
Histogram
Multi-vari chart
Run chart
Pareto chart
Odds ratio
Heat map
Bar chart
Horizon graph
Multidimensional scaling
Principal component analysis (PCA)
Multi-linear PCA
Iconography of correlations
Non-Graphical Exploratory Data Analysis
Scatter plot (2D/3D)
Stem-and-leaf plot
Parallel coordinates
Dimensionality reduction

Non-graphical exploratory data analysis is the first step when beginning to analyze your data as part of the general data analysis approach.
Measures of central tendency
i.e. the mean, the media and mode
Measures of spread,
i.e. variability, variants and standard deviation, the shape of the distribution, and the existence of outliers.

3. Data Loading-

Data is mounted on the Google drive. In the above dataset following are the column distribution

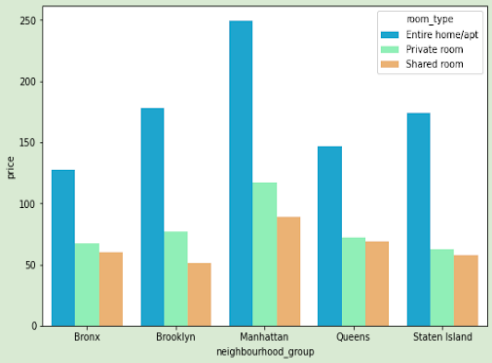
1. **id** : a unique id identifying an airbnb listing
2. **name** : name representing the accommodation
3. **host id** : a unique id identifying an airbnb host
4. **hostname** : name under whom host is registered
5. **neighbourhood group** : a group of area
6. **neighbourhood** : area falls under neighbourhood group
7. **latitude** : coordinate of listing
8. **longitude** : coordinate of listing
9. **room type** : type to categorize listing rooms
10. **price** : price of listing
11. **minimum nights** : the minimum nights required to stay in a single visit
12. **number of reviews** : total count of reviews given by visitors
13. **last review** : date of last review given
14. **reviews per month** : rate of reviews given per month
15. **calculated host listings count** : total no of listing registered under the host
16. **Availability 365**: the number of days for which a host is available in a year.

Columns like name, host name, last review and reviews per month have null values, we need to clean these missing values. From the descriptive data we can see that the minimum price is 0, which is not possible. Also the maximum value of minimum nights is 1250, which is also not possible. We need to set a minimum price of 100$, and maximum of minimum nights can’t be greater than 365 days.

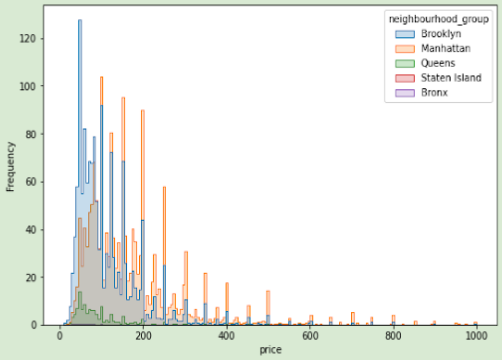
4. Exploratory Data Analysis -

A **bar plot** shows categorical data as rectangular bars with heights proportional to the value they represent. It is often used to compare between values of different categories in the data.

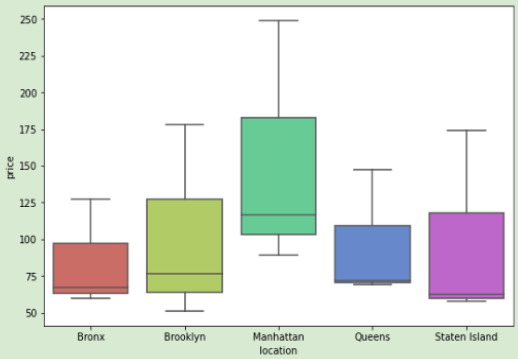
	location	room_type	price
6	Manhattan	Entire home/apt	249.246685
3	Brooklyn	Entire home/apt	178.338006
12	Staten Island	Entire home/apt	173.846591
9	Queens	Entire home/apt	147.050573
0	Bronx	Entire home/apt	127.506596



A **histogram** is a graph showing frequency distributions. It is a graph showing the number of observations within each given interval.

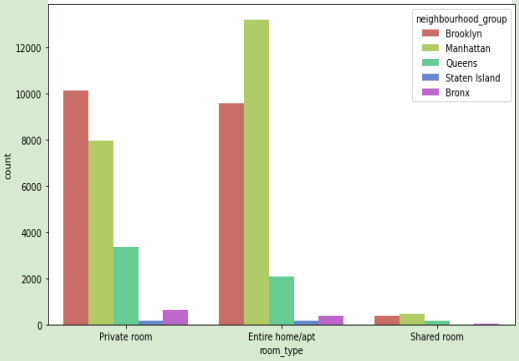


Boxplots are a measure of how well distributed the data in a data set is. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data set.



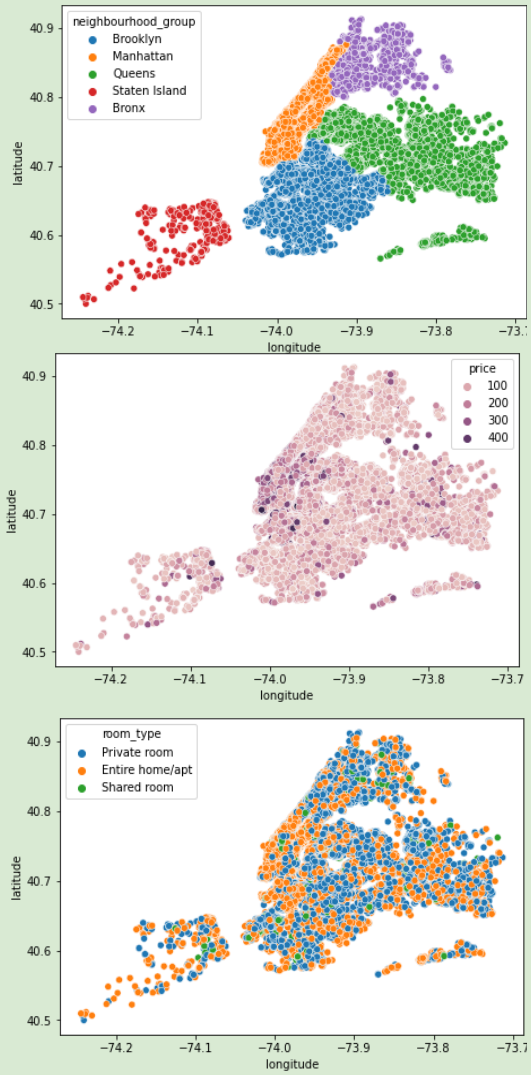
The distribution of Room Type is as follows

```
Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```



- Manhattan is the preferred location based on pricing and frequency of customer visit.
- Manhattan is preferred in all types of rooms.
- Prices in Manhattan are higher as compared to other neighbourhoods.
- Entire Home/Appt. Are higher in number in Manhattan.
- Private rooms are higher in Brooklyn.

- Shared rooms are higher in Manhattan.

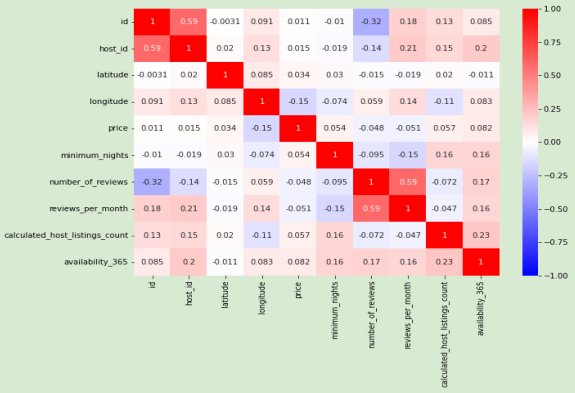


- Here we have plotted the scatter plot based on hue of location and price. We can clearly see that visitors prefer to visit the Manhattan location as mostly the high price region is saturated near Manhattan.
- In second plot, we have plotted the scatter plot over the price less than 500 $. we can clearly see that customers has high frequency to visit near Manhattan region

- In the third plot, we have a scatter plot over room type. We can see that shared rooms are least preferred by all customers in all neighbourhood groups. Private rooms and entire homes/apartments are highly preferred.

In the Manhattan , maximum hosts are located

A **heatmap** contains values representing various shades of the same colour for each value to be plotted. Usually the darker shades of the chart represent higher values than the lighter shade. For a very different value a completely different colour can also be used. In our case , we have plotted a heatmap for plotting correlation between two variables.



- High correlation number represents high correlation between two variables eg. Number of reviews and reviews per month has a correction factor as 0.59 which represents they are highly correlated.
- Low correlation number represents less correlation between two variables. eg. Host id and minimum nights have correlation factor -0.019

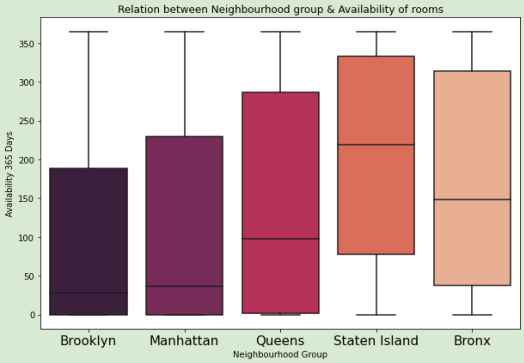
which represents they are not much dependent on each other.

Following table shows the top neighbourhood location where customers pay maximum and minimum rent.

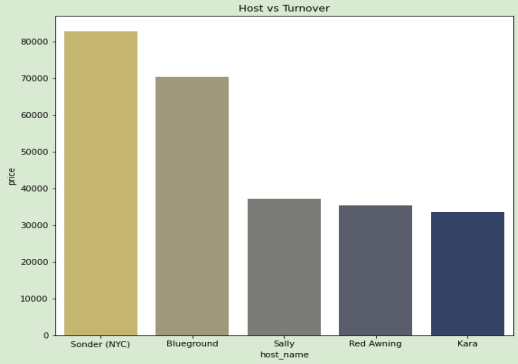
	Location	Maximum price	Minimum price
0	Brooklyn	10000	10
1	Manhattan	10000	10
2	Queens	10000	10
3	Staten Island	5000	13
4	Bronx	2500	10

Customers are paying the highest rent price of 10000 and lowest rent price of 10 at Manhattan, Brooklyn and Queens locations.

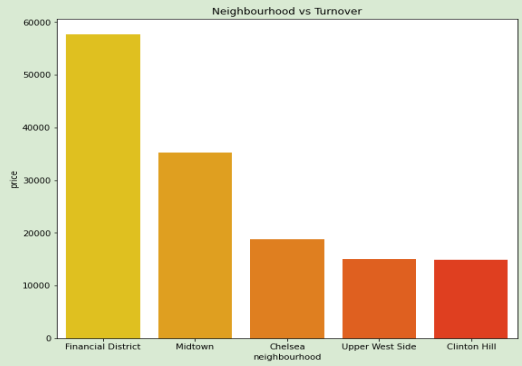
Most Popular/demanded host based on reviews and calculated host listings count are Sonder, Blueground, Kara, Kazuya, Jeremy & Laura



1. Staten Island has the highest availability of rooms over 365 days followed by Bronx.
2. Brooklyn and Manhattan has least availability of rooms



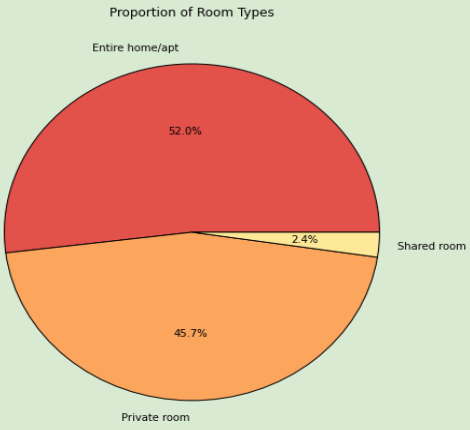
- Top 5 hosts based on turnover are- Sonder, Blue ground, sally, Red Awning, Kara

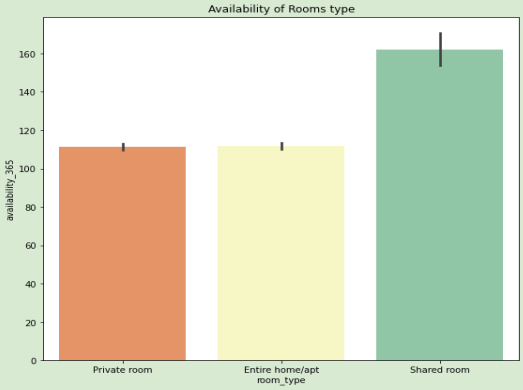


- It's clear that financial District, Midtown, Chelsa, Upper West Side, Clinton Hill are the top 5 neighbourhood groups respectively

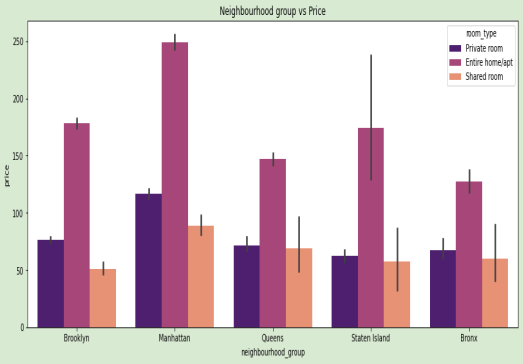
Shared rooms are available for most of the times as compared to private and entire home/ apartment

A **Pie Chart** is a circular statistical plot that can display only one series of data. The area of the chart is the total percentage of the given data. The area of slices of the pie represents the percentage of the parts of the data. The slices of pie are called wedges.



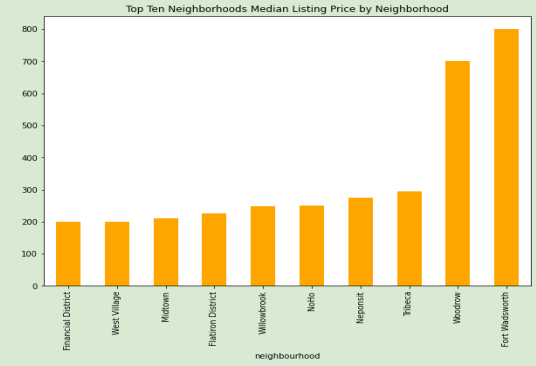


- Around **52%** of room types are entire home/apt, followed by **45.7%** as private rooms and **2.4%** as shared rooms.

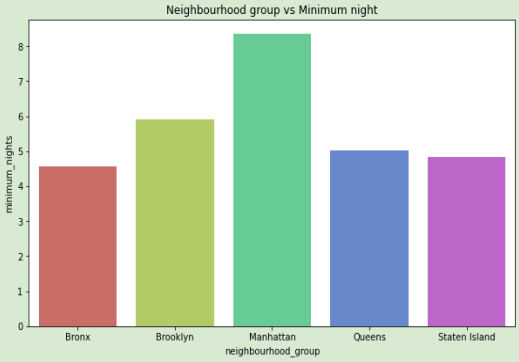


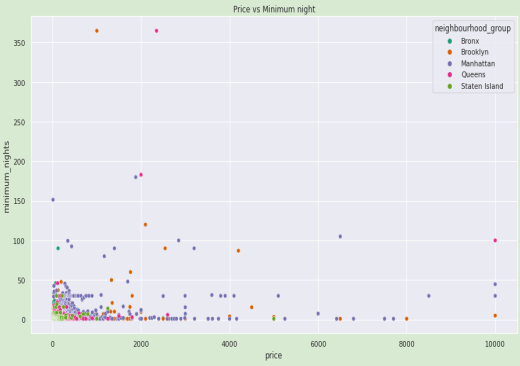
- Room type Entire home/apt has maintained a higher price range in almost all neighbourhoods groups.

- Manhattan and Brooklyn being the posh areas with more high end properties.



- Fort Wadsworth and Woodrow are the top neighbourhood in terms of median listing prices belonging to Staten Island.





- Most of the Hosts allow less than 5 days Mandatory stay for a single booking.

- Manhattan, Brooklyn and Queens have properties where the average of mandatory stay required is more than 5 days.
- Customers prefer to stay in properties where mandatory stay is minimum and budget friendly.

5.Limitations and Scope

Datasets have limiting attributes to classify various categories of properties. Customer experiential and Category wise ratings for Hosts seemed to be missing which could have played an important role in identifying Star Hosts. A lot of guest information was missing like Purpose of Visit, Number of Guests, which could have given a sense of understanding about the relation of customer footfall and neighbourhoods. Key attributes of properties like Number of Beds, Closets, Gyms,Bathrooms, Property Age, distances from nearest Hospitals, Shopping Complexes, Airport, and Station were missing.

6.Conclusion

Manhattan and Brooklyn are the posh areas in NYC as there is maximum footfall and properties based on prices and listings are on the higher side. They also have the highest number of hosts. Manhattan has the highest number of Private rooms and Entire House/Appt. in culmination followed by Brooklyn.
Rooms type distribution are as follows:
Entire Home/App. - 52 %
Private Rooms - 45.7%
Shared Rooms - 2.4%

Highest accommodations of 10,000 USD are available in Manhattan, Brooklyn and Queens. Most popular hosts are Sonder, Blueground, Kara to name a few based on number of reviews received, properties listed and possible turnover generated.

Staten Island seems more available for booking throughout the year compared to other neighbourhoods.Though it has fewer properties but the median listing price is more in its neighbourhood Fort Wadsworth and Woodrow.

Financial District which is the buzz of Manhattan has possibilities of high turnover based on its listed properties.

Most of the Hosts allow less than 5 days Mandatory stay for a single booking.Manhattan, Brooklyn and Queens have properties where the average of mandatory stay required is more than 5 days.Customers prefer to stay in properties where mandatory stay is minimum and budget friendly.

We have also come up with certain marketing campaigns and creative initiatives focused on neighbourhood groups like festive discounts, loyalty coupons, frequent check-in cards and more to increase the traction and business viability.

References-

1. GeeksforGeeks
2. Analytics Vidhya
3. Medium