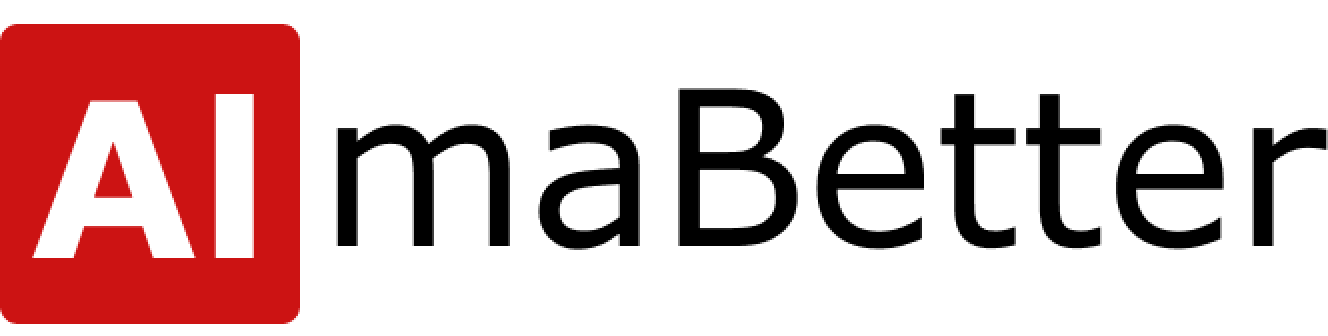
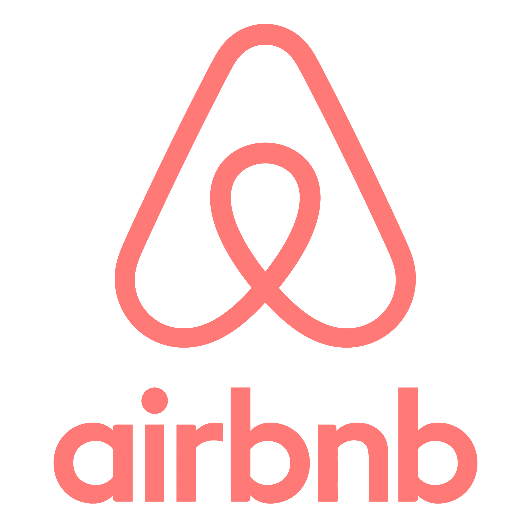
****

**Exploratory Data Analysis on Airbnb Bookings Analysis**

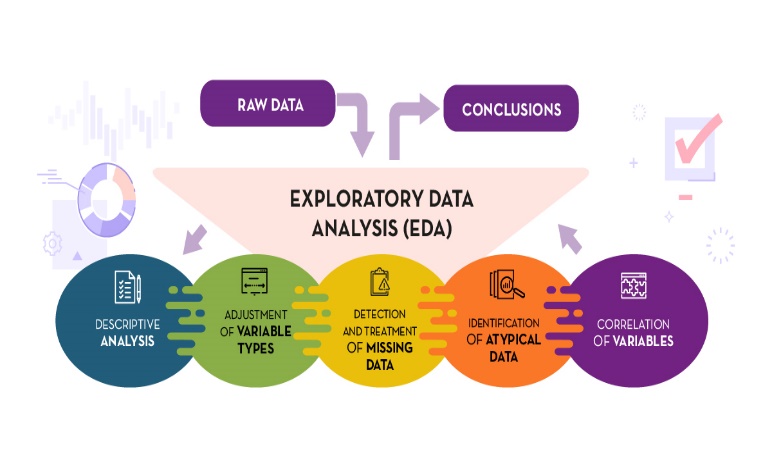
****

**By – Sachin Yallapurkar**

1. **About Exploratory Data Analysis.**

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

**Steps in Exploratory Data Analysis**



1. **Descriptive analysis**

  **What is it?** Synthesis of the information provided by the dataset, extracting its most representative characteristics.

 Why **is it necessary?** To understand the types of data, discover patterns and prepare the data for future analysis.

1. **Adjustment of Variable types**

  **What is it?** Verify that the variables have been stored with the appropriate corresponding value type.

 Why **is it necessary?** Bad coding of variables can negatively influence the grouping of data or the results of the analysis.

1. **Detection and treating of Missing Data**

  **What is it?** Identify some of the missing data in the variable.

 Why **is it necessary?** Missing data can create problems when applying machine learning techniques, building predictive models, performing statistical analysis or generating graphical representations.

1. **Detection and treatment of atypical Data**

  **What is it?** To identify data with values significantly different from those of the variable.

 Why **is it necessary?** They can modify the results and reduce the power of the statistical analysis or machine learning techniques applied.

1. **Correlation of variables**

  **What is it?** Analysing the relationship between two or more variables.

 Why **is it necessary?** Among the other reasons, to discard possible variables that provide redundant information in the dataset, causing noise in the analysis.

1. **Introduction to Airbnb Bookings Analysis.**

**What is Airbnb?**

Airbnb is an online marketplace for short-term homestays and experiences. Founded in 2008, it acts as a broker and charges a commission from each booking. Airbnb is a shortened version of its original name AirBedandBreakfast.com.

 It is available in 65000 cities and over 191 countries around the world.

 In 2021, Airbnb generated $5.9 billion revenue.

 Airbnb has 150 million users.

 There are over seven million listings on Airbnb, run by four million hosts.

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 analysed 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.

**Explore and analyse the data to discover key understandings (not limited to these) such as:**

1. What can we learn about different hosts and areas?
2. What can we learn from predictions?

**2.1** Type of room

**2.2** locations,

**2.3** prices,

**2.4** reviews

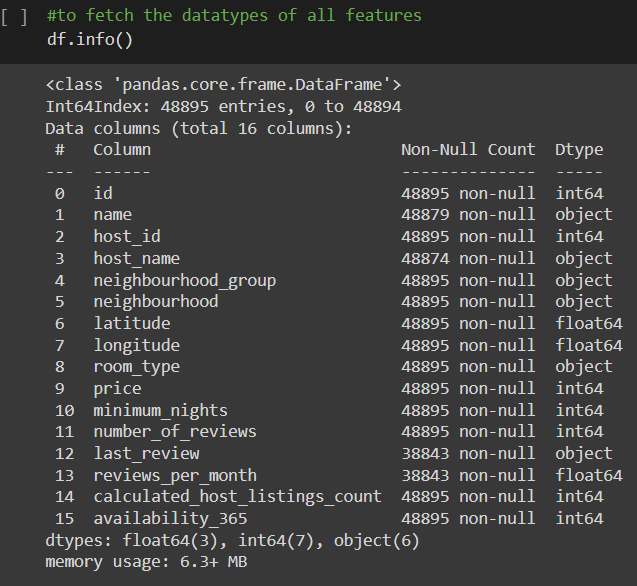
1. Which hosts are the busiest and why?
2. Is there any noticeable difference of traffic among different areas and what could be the reason for it?
3. What is the percentage of listings owned by Airbnb in different neighbourhoods?
4. **Data Cleaning of Airbnb data frame**

**Step 1 - Descriptive Analysis**

The given dataset has 48,895 observations and 16 different features. Let us look what each feature is all about

1. id id given to listings
2. name name of the listing
3. host\_id unique host ids
4. host\_name Gives host name
5. neighbourhood\_group It contains 5 neighbourhood groups namely : Brooklyn,Manhattan,Queens,Staten island,Bronx.
6. neighbourhood There are total of 221 different neighbourhoods.
7. latitude It gives the latitude of house listing.It helps in getting the location.
8. longitude It gives the longitude of the house listing. It helps in getting the location.
9. room\_type There are total of 3 different types of rooms available on Airbnb i.e. Private room, Entire home or apartment and shared room.
10. price It tells about the price of each listing
11. minimum\_nights It tells about minimum nights spent by people in listing
12. number of reviews It gives the total number of reviews
13. last\_review It tells about when the last review was given
14. reviews\_per\_month It tells about review got by listing per month
15. calculated\_host\_listings It tells about the number of times a host was listed or booked by people
16. availability\_365 It tells about availability of listing out of 365 days

Step 2 - **Adjustment of Variable types**



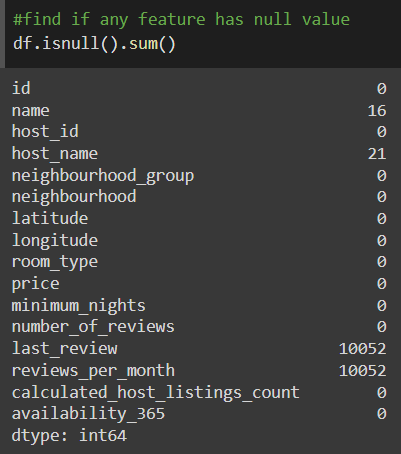
We can see that datatype of columns is same as what is expected.

**For example**

* **id,host\_id,price,number\_of\_reviews,calculated\_host\_listings\_count,availability\_365** are supposed to be integer datatypes and they are in actual int64.So they are compliant.
* **name,host\_name,neighbourhood\_group, neighbourhood,room\_type,last\_review** are supposed to be characters and they are "object" datatype.
* **latitude, longitude** are supposed to be floats and in the given dataset they are float64.

So the dataset doesn't need to be adjusted for variable types.

Step 3 - **Detection and treating of missing data.**



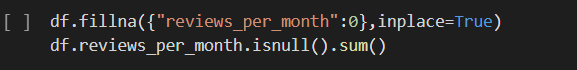
**name** and **host\_name** column have 16 and 21 null values respectively.

**last\_review** and **reviews\_per\_month** each has 10,052 observations as null.

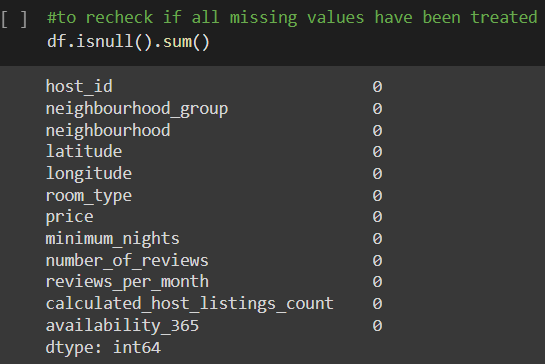
* In this case we observe that **"id", "name "** and "host\_name" are redundant for us as we will be referring to listings based on unique **host\_id**. So we will be dropping "name " and "host\_name" features.
* "last\_review" feature depicts the date on which last review was given for the listing, it is irrelevant here. So we will be getting rid of this feature.



Few of the listings have missing value for "reviews\_per\_month”, as the value is unknown, we will be using zero (0) so that missing values don't intervene with our analysis.



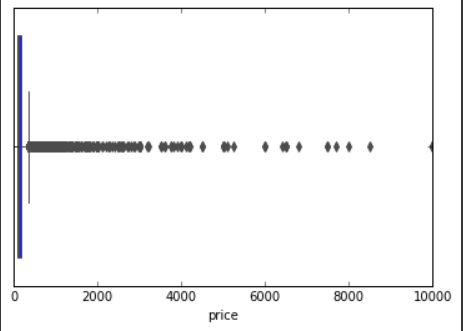
We can see that there are no columns with any missing values now,all have been properly handled.



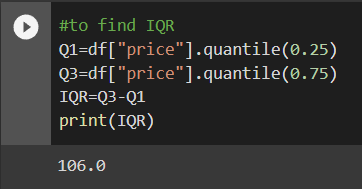
Step 4 - **Detection and treatment of atypical data**

The interquartile range (IQR) measures the spread of the middle half of your data. It is the range for the middle 50% of your [sample](https://statisticsbyjim.com/glossary/sample/). Use the IQR to assess the variability where most of your values lie. Larger values indicate that the central portion of your data spread out further. Conversely, smaller values show that the middle values cluster more tightly.

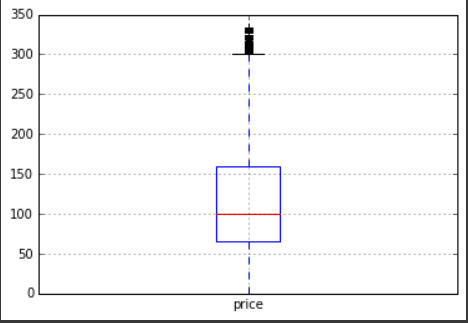
**Boxplot of this price column to have a feel of the presence of outliers.**

****

**Now we will use Interquartile Range(IQR) to handle the outliers.**

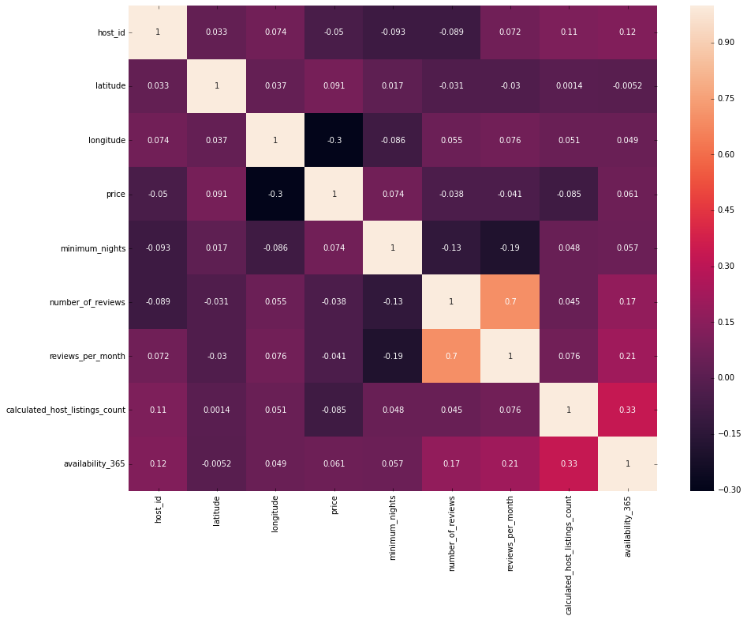


**Boxplot for price after removing outliers**

****

Step 5 - **Correlation of variables**

checking the **correlation** matrix to understand how are the features interrelated with each other. I have plotted using seaborn heatmap to understand the strength between the variables used.



**Observations:**

Features 'reviews\_per\_month' and 'number\_of\_reviews' have a positive correlation with a value of 0.7. So they

almost give the same information. So analysis can be carried out with any of the two variable.

Features 'reviews\_per\_month' and 'minimum\_nights' have a negative correlation with a value of -0.19.

**Time to visualize data**

After cleaning data, now it's time to explore and visualize data to get some insights and get answers to the following questions:

1. What can we learn about different hosts and areas?
2. What can we learn from predictions?

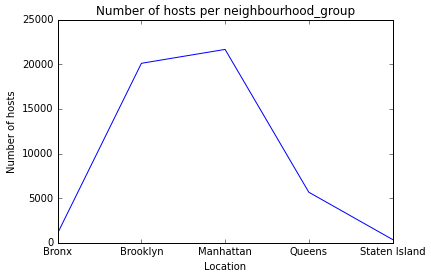
2.1 Type of room

2.2 locations,

2.3 prices,

2.4 reviews

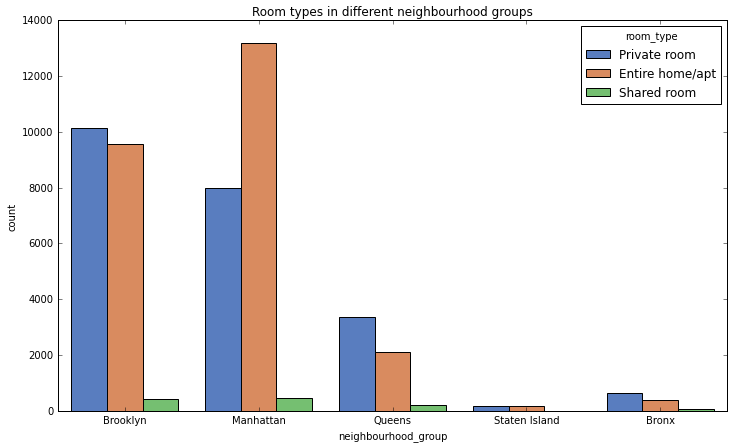
1. Which hosts are the busiest and why?
2. Is there any noticeable difference of traffic among different areas and what could be the reason for it?
3. What is the percentage of listings owned by Airbnb in different neighbourhoods?
4. **What can we learn about different hosts and areas?**



**Observation**

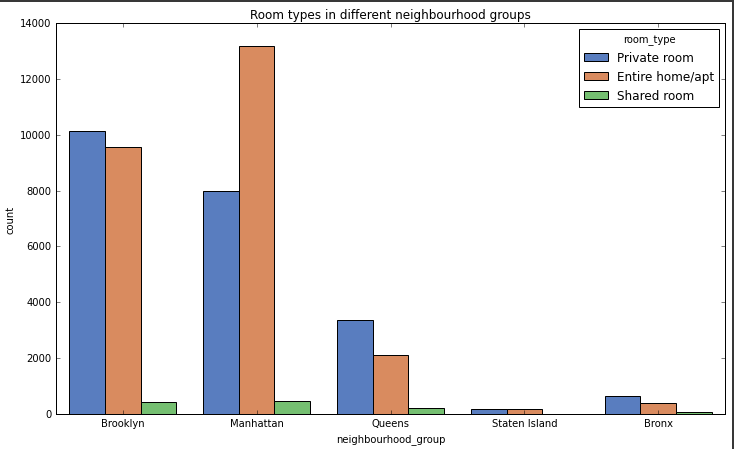
* Most of the hosts are located in Manhattan.i.e.,about 21661 hosts.
* Least number of hosts are in Staten Island i.e., about 373 hosts.

1. **What can we learn from predictions? (ex: locations, prices, reviews, type of room)**



**Observation** :

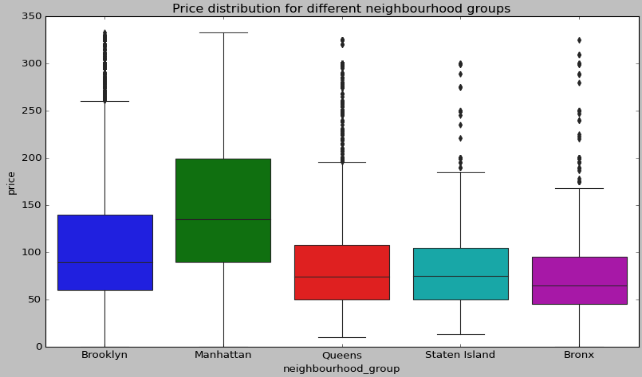
* On Airbnb 3 different types of rooms are available for booking. They are **Private room, Entire home/apartment and Shared rooms**
* Most people opt for Entire home/apartment type of listing.
* Shared rooms are the least sought out option on Airbnb.
* In particular Manhattan most sought out option is Entire home or apartment, contrary to this in Brooklyn most sought option is private rooms.
  1. **Prediction based on type of room**



**Observation**:

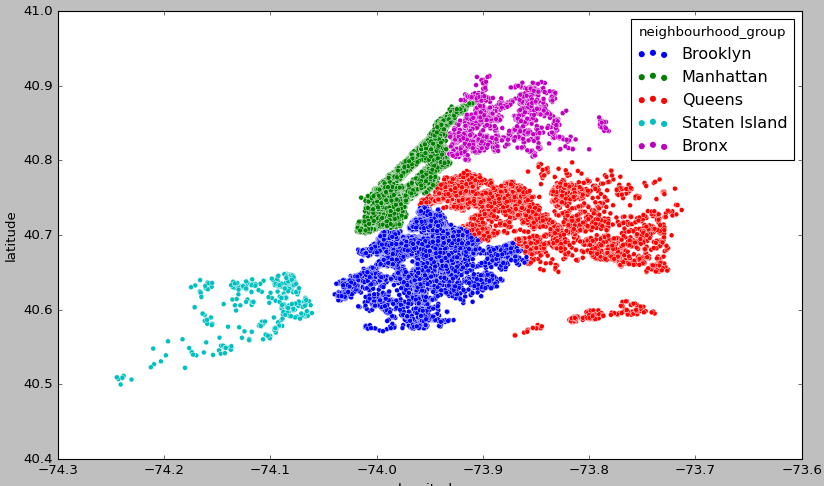
On Airbnb 3 different types of rooms are available for booking. They are Private room, Entire home/apartment and Shared rooms.

* Most people opt for Entire home/apartment type of listing.
* Shared rooms are the least sought out option on Airbnb.
* In particular Manhattan most sought out option is Entire home or apartment, contrary to this in Brooklyn most sought option is private rooms.
  1. **Prediction based on Price**



**Observation**:

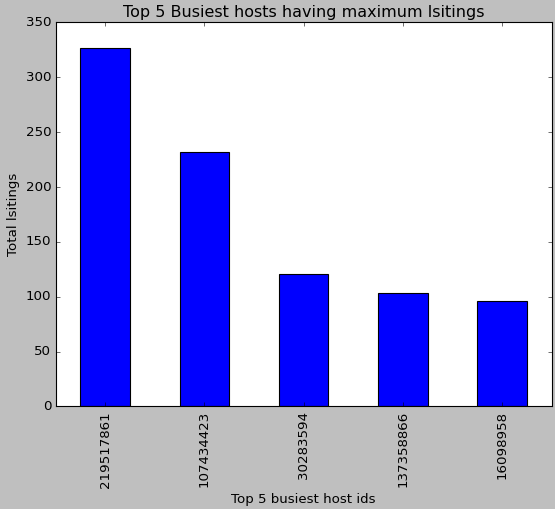
* Average price is highest for Entire home or apartment in Manhattan.
* Among all 5 neighbourhood\_groups, highest price is for Entire home or apartment.
* Among all 5 neighbourhood\_groups, lowest price is for Shared rooms.
  1. **Prediction based on location.**



**Observation:**

* From the location scatterplot we can see that area occupied by Airbnb in Queens is highest and Manhattan is lowest. But still maximum of hosts are located in Manhattan.
* Using scatterplot for latitude and longitude we can map how the listings are located.

1. **Which hosts are the busiest and why?**



**Observation** :

**host\_id 219517861** is the busiest host with total of **327 listings** in Manhattan.

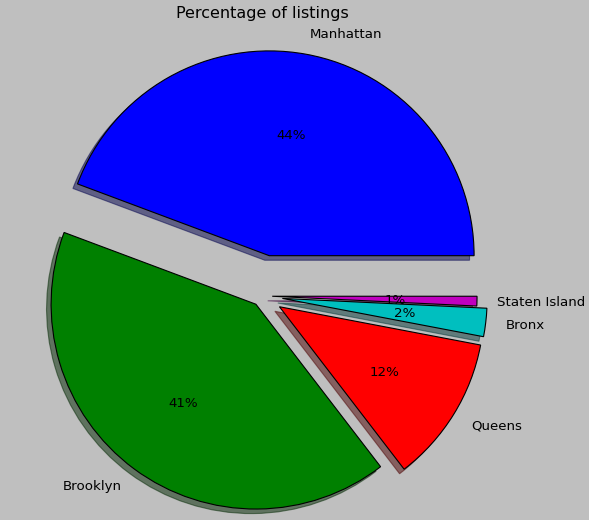
1. **Is there any noticeable difference of traffic among different areas and what could be the reason for it?**



**Observations**

* Traffic is mainly in Manhattan and Brooklyn. As Bronx and Staten Island are away from city centre, we see less traffic over there.
* People are preferring mainly Entire home or Private room than shared rooms. It is due to privacy preference. People are ready to pay more for this.
* People who are staying in Apartment or entire home are staying for longer duration compared to people staying in private room and shared rooms.

1. **What is the percentage of listings owned by Airbnb in different neighbourhoods?**



**Observation:**

* Percentage of listings in Manhattan is 44% and then followed by Brooklyn 41%
* Percentage of listings in in Staten Island (1%)

**Conclusion**

**Insights**:

* Most of the hosts are located in Manhattan.i.e.,about., about 21661 hosts.
* Least number of hosts are in Staten Island i.e., about 373 hosts. Average price is highest for Entire home or apartment in Manhattan.
* Among all 5 neighbourhood\_groups, highest price is for Entire home or apartment.
* Among all 5 neighbourhood\_groups, lowest price is for Shared rooms. On Airbnb 3 different types of rooms are available for booking. They are **Private room, Entire home/apartment and Shared rooms**
* Most people opt for Entire home/apartment type of listing.
* Shared rooms are the least sought out option on Airbnb.
* In particular Manhattan most sought out option is Entire home or apartment, contrary to this in Brooklyn most sought option is private rooms. -
* Traffic is mainly in Manhattan and Brooklyn. As Bronx and Staten Island are away from city centre, we see less traffic over there.
* People are preferring mainly Entire home or Private room than shared rooms. It is due to privacy preference. People are ready to pay more for this.
* host\_id 219517861 is the busiest host with total of 327 listings in Manhattan.
* From the location scatterplot we can see that area occupied by Airbnb in Queens is highest and Manhattan is lowest. But still maximum of hosts are located in Manhattan.
* People who are staying in Apartment or entire home are staying for longer duration compared to people staying in private room and shared rooms.

**Business suggestions**:

* Manhattan and Brooklyn are the high demand areas, owning few more listings by Airbnb helps in their business acquisition.
* As shared rooms are less preferred in these areas, Airbnb can look for converting few percent of shared rooms to entire home or private room.
* As Staten Island and Bronx are the least sought options by customers, by giving more discounts customers can be attracted here.

**References:**

* Almabetter study materials
* Medium
* Wikipedia
* datos.gob.es