# Airbnb Bookings Analysis

**Shrinidhi Choragi**
**Nakshith D N**
Data science trainees,
AlmaBetter, Bangalore

## Abstract

Airbnb, as in “Air Bed and Breakfast,” is a service that lets property owners rent out their spaces to travellers looking for a place to stay. It is an American company that facilitates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. It basically connects travellers with local hosts who want to rent out their homes to people who are looking for accommodations in that locality. On the other hand, this platform enables hosts to list their available space and earn extra income in the form of rent and it also enables travellers to book unique homestays from local hosts, saving them money and giving them a chance to interact with locals.

In the world of rising new technology and innovation, the Travel industry is advancing with the role of Data Science and Analytics. Data analysis can help them to understand their business in a quite different manner and helps to improve the quality of the service by identifying the weak areas of the business. This study demonstrates how different analyses help out to make better business decisions and help analyse customer trends and satisfaction, which can lead to new and better products and services.

Our experiment can help understand and explore this model using Exploratory Data Analysis to get insights from this data based on which business decisions will be taken.

***Keywords: Exploratory Data Analysis, Business Analysis.***

## Problem Statement

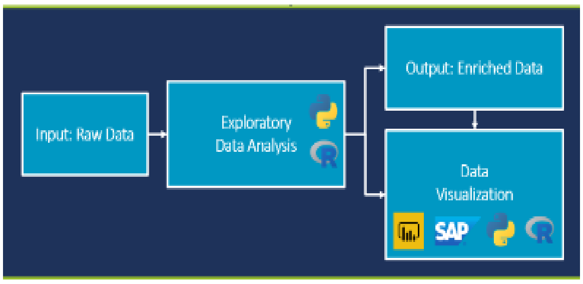
The objective of the project is to perform exploratory data analysis, data pre-processing, data cleaning & imputation, and in the end, apply different Data Visualisation techniques to get meaningful insights from the given data.

Explore and analyse the data to discover key understandings.
The key understandings explored in our analysis include:

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference in traffic among different areas and what could be the reason for it?
- What is average revenue per host and how does it vary for different neighbourhood groups
- Depict the price distribution among neighbourhood groups.

## Introduction

Travel industries are having an important reflection on the economy over the past few decades, and Airbnb housing price ranges are of great interest for both Hosts and Travelers. In this project, we are analysing the various aspects which cover many aspects of Airbnb listings. It helps in understanding the meaningful relationships between attributes and allows us to do our own research and come up with our findings.



## Exploratory Data Analysis (EDA)

"Exploratory Data Analysis" (EDA) is a "Data Exploration" step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

Understanding the Dataset can refer to a number of things including but not limited to extracting important "variables", identifying "outliers", "missing values", or "human error". Ultimately, maximising our insights of a dataset and minimising potential "errors" that may occur later in the process.

## Dataset

Since 2008, guests and hosts have used Airbnb to expand on travelling possibilities and present a more unique, personalised way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

This public dataset is part of Airbnb, and the original source can be found on this *website*.

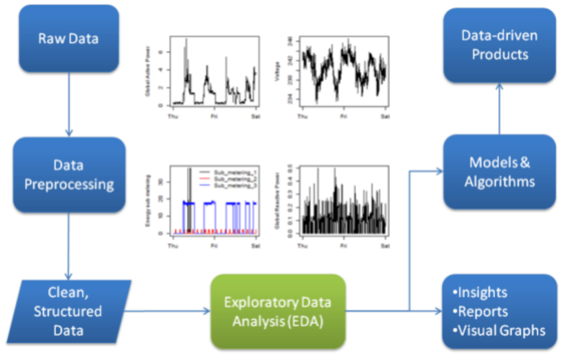
## Data Exploration

This dataset has around 48895 observations in it with 16 columns and it is a mix of categorical and numeric values.

The features description include

| | |
|---|---|
| id | Identity number of the property listed |
| name | Name of the property |
| host_id | Id number of hosts registered on Airbnb |
| host_name | Name of the host registered |
| neighbourhood_group | Names of neighbourhood groups in NYC |
| neighbourhood | Names of neighbourhood present in neighbourhood groups |
| Latitude | Coordinate of latitude of the property listed |
| Longitude | Coordinate of longitude of the property listed |
| room_type | Type of room listed by host |
| price | Rent of the property listed |
| Minimum_nights | the minimum number of nights customer can rent the property |
| number_of_reviews | Number of customers that have reviewed the property |
| last_review | Date when the property was last reviewed. |
| reviews_per_month | |
| Calculated_host_listings_count | Number of listings done by particular host |
| Availability_365 | Number of days the property is available |

## Architecture



## Data Pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for our analysis purpose, where we have to do a lot of Data Cleaning and handle the missing values by using appropriate imputation techniques and based on that variable nature i.e. either of categorical & numerical variable.

Substitution/imputation of missing values using either mean, median, mode or zero according to the nature of those variables. Here, in this project, we have imputed with zero. Moreover, we also removed the columns which do not participate in our analysis.

## Data Preparation

Data Preparation includes loading the dataset into a data frame, exploring the number of rows & columns, ranges of values, data types, descriptive summary of numerical features, correlation and distribution of features etc.

## Data Cleaning

It includes identifying the missing values in the dataset and handling the same.

The Percentage of missing values in the dataset is found to be: 2.57% from features namely: *name, host_name, last_review and reviews_per_month.*

Where missing values are handled as follows:

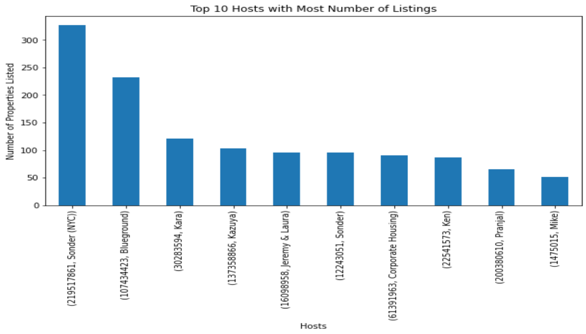
- Missing values of categorical columns can be filled with a dummy variable. The features *host* and *host name* were imputed with a dummy variable.
- Missing values of numerical columns are imputed with suitable numbers. In our case, column "*review_per_month*" missing values can be imputed with 0.0 for missing values since "*number_of_review*" of the corresponding column has a 0.
- Column "*last_review*" is of type date; For rows with no reviews of the listing, the date doesn't exist. Hence considering this column irrelevant, It's dropped.

**Analysis of crucial understandings explored includes**

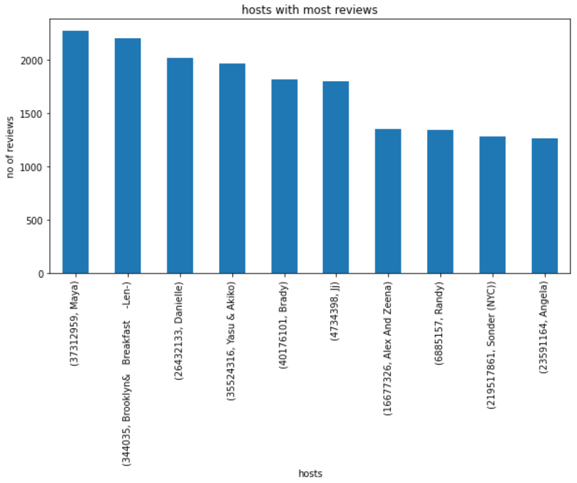
1. **What can we learn about different hosts and areas?**

- <u>Hosts with the most number of listings</u>:

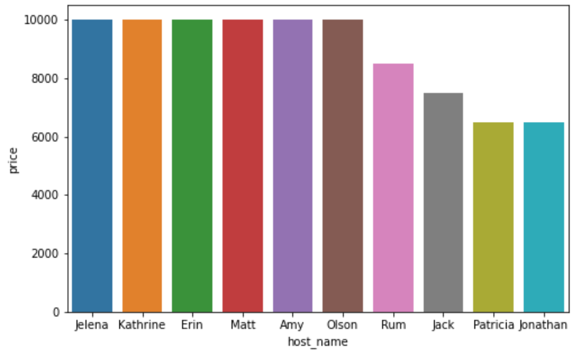
Various hosts in the dataset have similar names which leads to conflict if the dataset is grouped based on host names only. Hence to distinctively differentiate the listings the dataset is grouped based on host id and hostname.
And top hosts with the most number of listings can be obtained.
Similarly with reviews and earnings.



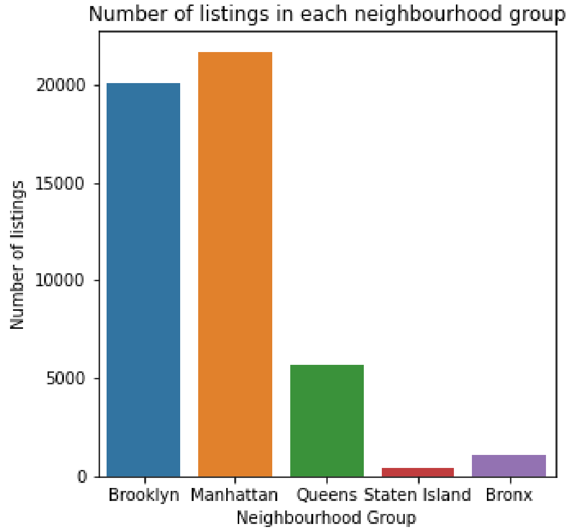
## Hosts with the highest number of reviews



## Host with the highest earnings

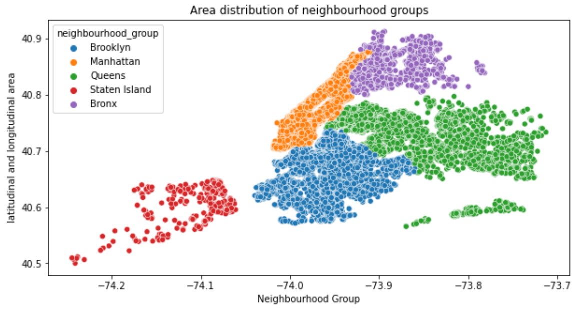


## Number of listings in each neighbourhood group



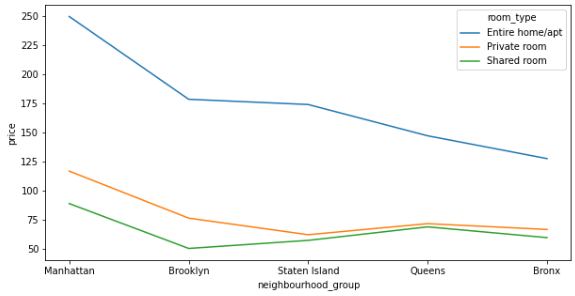
Manhattan has the highest number of listings compared to other neighbourhood groups. Staten Island has the lowest number of listings.

## Area distribution of neighbourhood groups



## 2. What can we learn from predictions? (ex: locations, prices, reviews, etc)

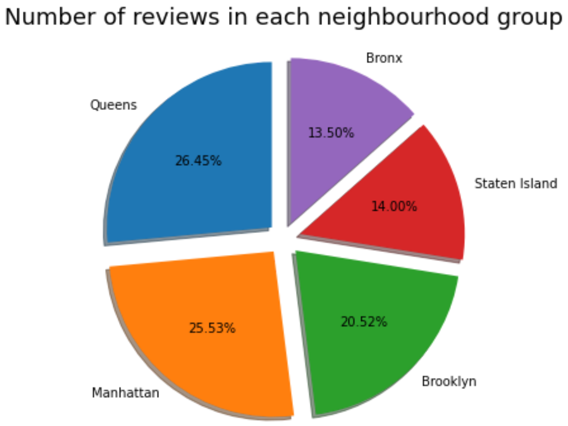
## Variation of price across the neighbourhood groups for all room types.



Most private rooms are of medium price, hence being affordable and opted by many Manhattan has the highest avg price for all room types

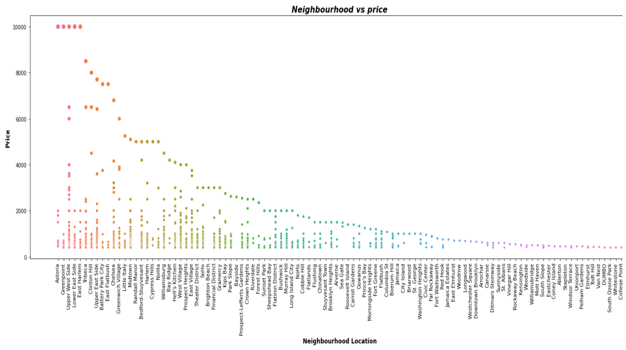
Entire home/apt room type has the highest avg price in all neighbourhood groups.

Number of reviews per neighbourhood groups

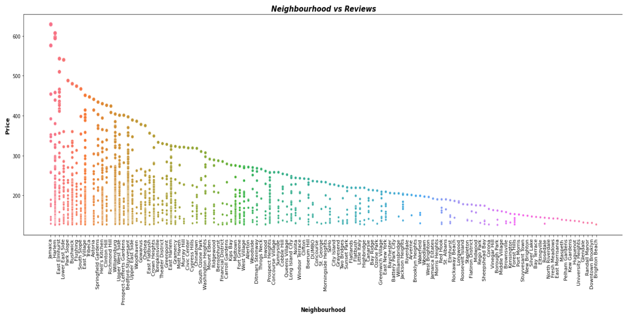


Queens and Manhattan are the most likeable neighbourhood groups with the most reviews (assuming reviews to be positive only )

Distribution of price among neighbourhoods



Distribution of reviews with respect to neighbourhoods



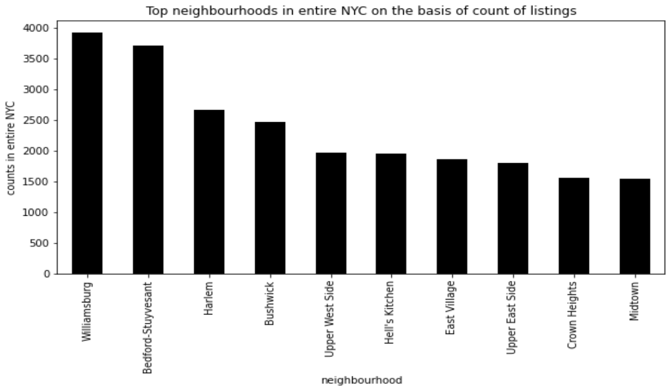
Some of the highly reviewed places like Jamaica, Harlem, Flushing have moderate or low priced listings. Explaining affordability.

While some places like Bushwick, and Williamsburg constitute the maximum number of listings leading to higher reviews.

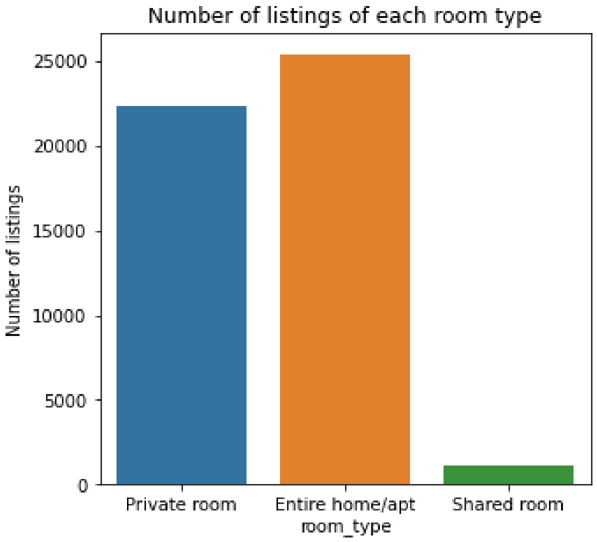
Thus most reviewed places can come from many factors such as affordability, more number of listings in the area, and economical importance.

### 3. Which hosts are the busiest and why?

Neighbourhoods with the most number of listings



Number of listings of each room type



Entire home/apt are the most preferred room_types

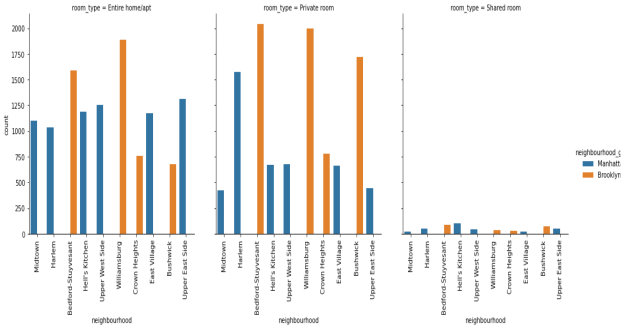
Conclusion: The above-mentioned hosts are the busiest due to the following reasons:

The private room and Entire home/apt are the most preferred room types according to the above histogram plot.

Neighbourhood groups Queens and Manhattan have the maximum number of reviews {assuming the reviews to be positive} as seen in the pie chart.

Maya is the busiest host according to the number of reviews. (assuming the reviews to be positive)

### 4. Is there any noticeable difference in traffic among different areas and what could be the reason for it



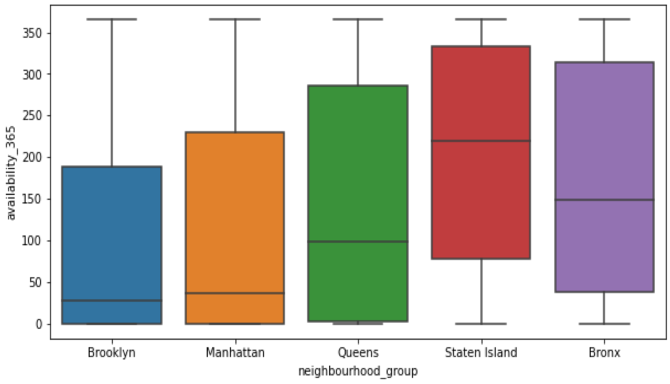
Observation:

For these top 10 neighbourhoods since Manhattan and Brooklyn are the busiest destinations, also with the most number of listings.

Bedford-Stuyvesant and Williamsburg are the most popular for Brooklyn, and Harlem for Manhattan.

'Shared room' type Airbnb listing is barely available among the 10 most listing-populated neighbourhoods.

Availability among neighbourhood groups



The mean availability of Brooklyn shows that it has the least number of availability. Hence it can be said that the listings in Brooklyn are seldom available and mostly booked. This implies there is a high possibility of the group being prone to traffic.

Followed by Manhattan, Queens, Bronx, and Staten Island.

Conclusion: Why there is noticeable traffic in most areas

Most of the neighbourhoods with a high number of listings are present in the Manhattan neighbourhood group making it a busy place.

The private rooms are mostly of average price, which makes them affordable to a wide range of people.

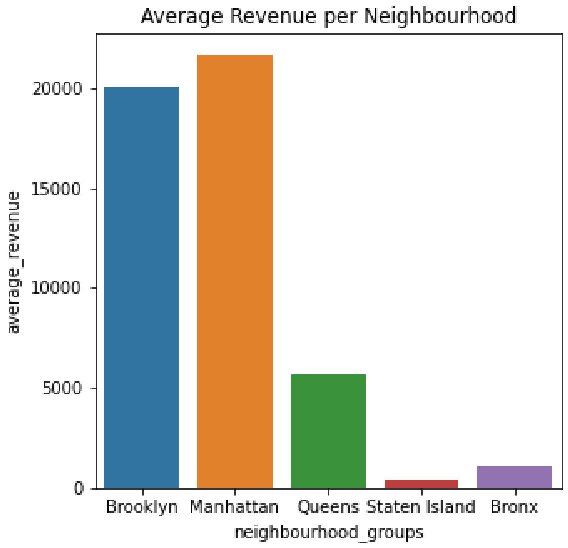
The neighbourhoods with the most number of listings like- Bedford-Stuyvesant, Williamsburg from Brooklyn and Harlem from Manhattan, have the most private rooms explaining the traffic there.

Highly-reviewed places mostly have average/ low prices. Thus high traffic is seen.

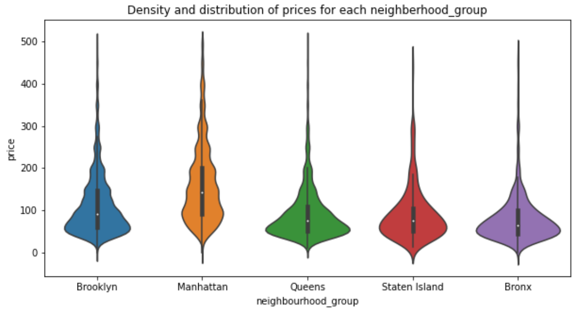
Though the shared rooms have the least average price, due to the lesser number of listings of type shared room, the traffic cannot be expected.

Thus, the number of listings, number of reviews, affordability etc contribute to traffic in certain areas

Calculation of Average Revenue



Distribution of price among neighbourhood groups



Price distribution for neighbourhood groups can be observed. Brooklyn has an avg price of $80. Manhattan has the highest range of price with an average of around $150. Queens and Staten Island have almost similar distributions with an average price of around $50.The  Bronx is the cheapest neighbourhood group.

## CONCLUSION

This Airbnb ('AB_NYC_2019') dataset allows us to perform data exploration on most of the columns. Starting with the data cleaning, handling missing values, distribution and correlation, we have moved on to answer a few of the key understandings.

Talking about hosts with the most number of listings, most number of reviews and top-earning hosts includes the analysis on feature-host. Next, we have explored the neighbourhood groups with the most listings and their geographical distribution.

We explored various relations to learn from predictions.

Variation of price across the neighbourhood groups for all room types. The number of reviews per neighbourhood group, Describing the distribution of price and number of reviews  among neighbourhoods.

Analysed the busiest host based on neighbourhoods with the most listings, number of listings of each room type and most number of reviews.

Explained the reason for traffic in some areas using the relation between the number of reviews and price, neighbourhood group and room type and availability among different neighbourhood groups using boxplot.

Calculated the average revenue per host and plotted it for different neighbourhood groups

Depicted the price distribution among neighbourhood groups using a Violin plot.

It would have been better if the reviews were broken down into positive and negative or scaled (0-5). Which helps in making the analysis stronger. However, throughout this analysis, we have assumed reviews to be positive only.