Airbnb Bookings Analysis

Aniket Raj, Kushal Kishor, A Balaji,
Akhilesh Bhardwaj, Aswathaman,
Data science trainees,
AlmaBetter, Bangalore

Abstract:

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.

1. Introduction

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

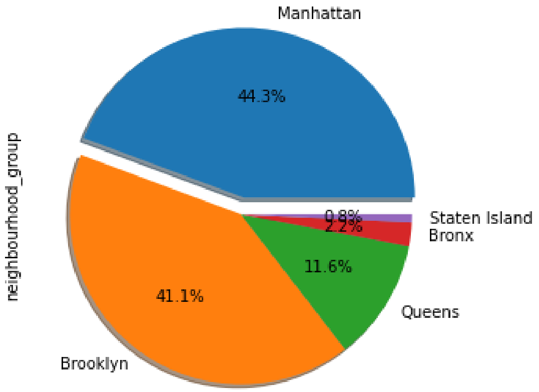
- **Id:** This variable consists of numerical values and all values are unique.
- **Name:** This variable consists of categorical values and each value represents the name of the property/apartment.
- **Host id**: This variable consists of numerical values and all values are unique.
- **Host name**: This variable consists of categorical values and each value represents the name of the host of property/apartment.
- **neighbourhood group:** This variable consists of categorical values and each value represents Boroughs of New York City.
- **neighbourhood:** This variable consists of categorical values and each value comes under the neighborhoods of Boroughs of New York City.
- **latitude:** This variable consists of numerical values and each value represents the

latitude of the specific location of NYC

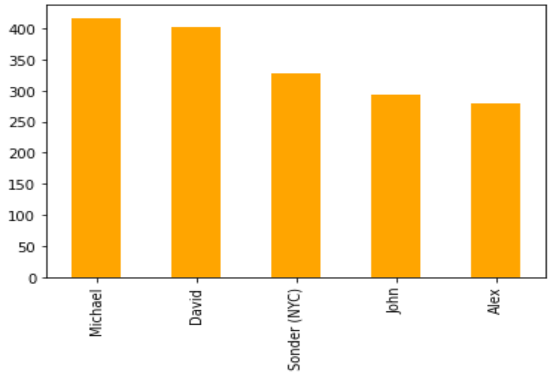
- **longitude:** This variable consists of numerical values and each value represents the latitude of the specific location of NYC
- **room type:** This variable consists of categorical values and each value represents the types of rooms
- **price:** This variable consists of numerical values and each value represents the price of property/apartment and it is a dependent or main variable
- **minimum nights:** This variable consists of numerical values and each value represents the minimum nights for stay or booking purpose
- **number of reviews:** This variable consists of numerical values and each value represents the count of reviews
- **last review:** This variable consists of numerical values and each value represents the date of last review
- **reviews per month:** This variable consists of numerical values and each value represent the total number of reviews per month
- calculated host listing count: This variable consists of numerical values and each value represents the count of calculated host listing
- **availability 365:** This variable consists of numerical values and each value represents the number of days availability

2. Problem Statement

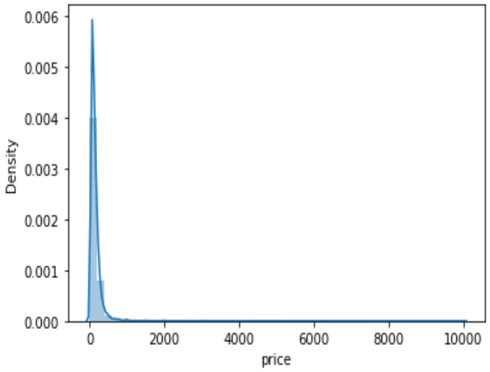
- Check how many host ids belongs to which neighbourhood group.



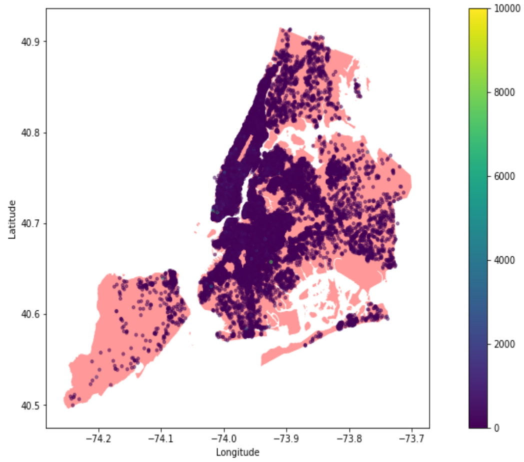
- Which hosts are the busiest and why?



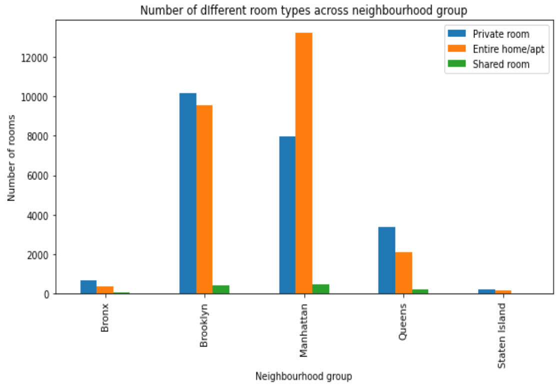
- We concluded the overall descriptive summary of the main variable('Price')



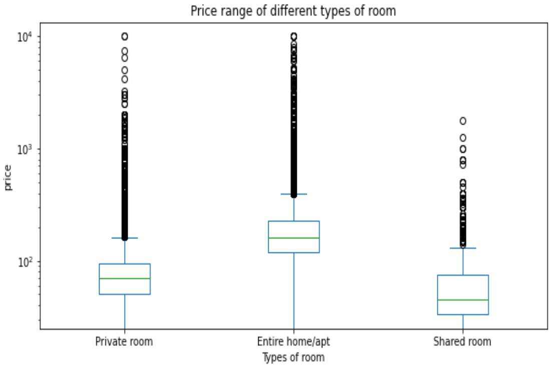
- We are analyzing 'Price' variable with 'latitude' variable and 'longitude' variable of the data:



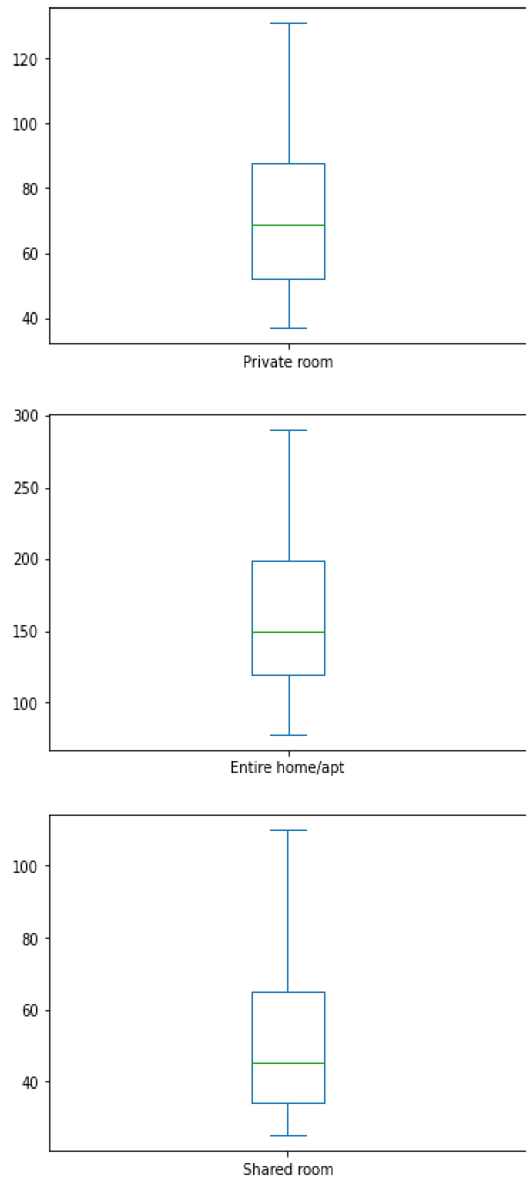
- How many types of room and number of different types of room in each neighbourhood group?



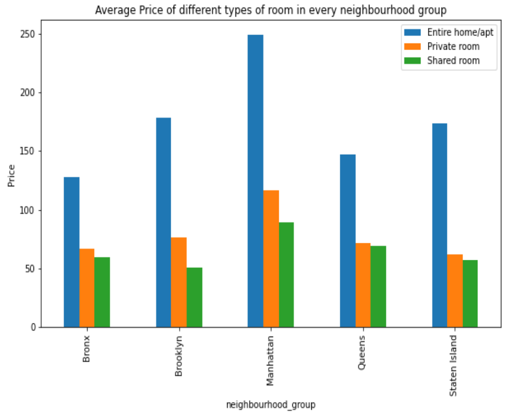
- We concluded the overall descriptive summary of the main variable ('Price') along different types of rooms.



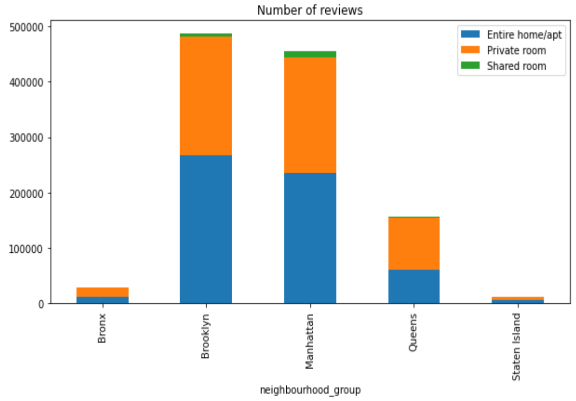
Eliminating the outliers and increasing the accuracy we have got the below graphs



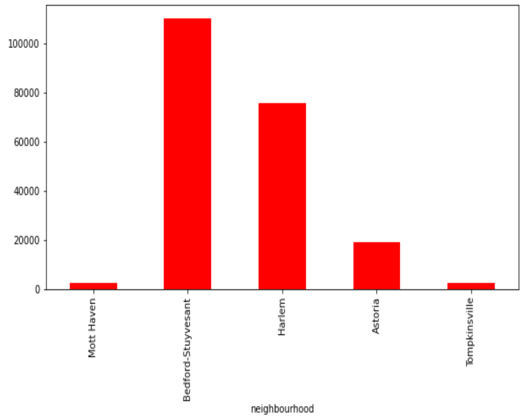
- What is the average price of different types of rooms in each neighbourhood group?



- What are the numbers of visits in different types of rooms in each neighbourhood group?



- On the basis of number of reviews, we concluded the maximum number of visits in which neighbourhood in each neighbourhood group.



3. Steps involved:

- **Exploratory Data Analysis**

  After loading the dataset we performed this method by comparing our target variable that is Price with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- **Null values Treatment**

  Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

- **Outliers Treatment**

  Outliers is also something that we should be aware of. Why? Because outliers can markedly affect our models and can be a valuable source of information, providing us insights about specific behaviors. Outliers is a complex subject and it deserves more attention.

- **Standardization of features**

  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

  The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

4. Conclusion:

The findings from the EDA can be informative to either hosts or tourists or both. Write down a few findings from the EDA so that hosts can gain better ratings or higher prices. For tourists, findings can be around the best time to visit, or easy commute options, etc. Another effective way to communicate results is to visualize data, histograms to visualize the spread, bar graphs to compare different neighborhoods, and room types.