Airbnb Bookings Analysis

Agam Singh, Jyoti Patel,
Data science trainees,
AlmaBetter,
Bangalore

Abstract:

Airbnb (ABNB) began in 2008, it is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.Airbnb offers people an easy, relatively stress-free way to earn some income from their property.

This project aims to analyse the Airbnb dataset using exploratory data analysis techniques to gain insights into the booking trends and patterns of Airbnb listings. The dataset used in this analysis is the publicly available Airbnb dataset that contains information on various aspects of Airbnb listings such as location, availability, pricing, and reviews. The analysis focuses on understanding the relationship between various attributes of the listings and the booking behaviour of the guests. The findings from this analysis provide valuable insights into the factors that influence the booking behaviour of guests on Airbnb, which can be useful for hosts to optimise their listings and for guests to make informed decisions while booking. The results of this analysis are presented in this project documentation along with visualisations that highlight the key insights.
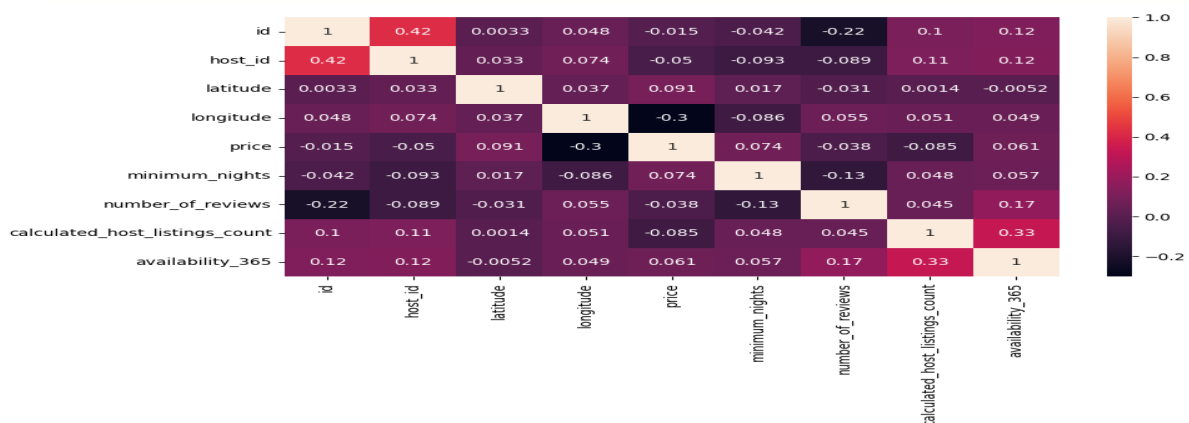
1. Introduction

This dataset has around 48,895 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 neighbourhoods 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
 ➢ What is the correlation between different variables ?

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
id	1	0.42	0.0033	0.048	-0.015	-0.042	-0.22	0.1	0.12
host_id	0.42	1	0.033	0.074	-0.05	-0.093	-0.089	0.11	0.12
latitude	0.0033	0.033	1	0.037	0.091	0.017	-0.031	0.0014	-0.0052
longitude	0.048	0.074	0.037	1	-0.3	-0.086	0.055	0.051	0.049
price	-0.015	-0.05	0.091	-0.3	1	0.074	-0.038	-0.085	0.061
minimum_nights	-0.042	-0.093	0.017	-0.086	0.074	1	-0.13	0.048	0.057
number_of_reviews	-0.22	-0.089	-0.031	0.055	-0.038	-0.13	1	0.045	0.17
calculated_host_listings_count	0.1	0.11	0.0014	0.051	-0.085	0.048	0.045	1	0.33
availability_365	0.12	0.12	-0.0052	0.049	0.061	0.057	0.17	0.33	1

➢ What is the distribution of Airbnb listings by borough in New York City?

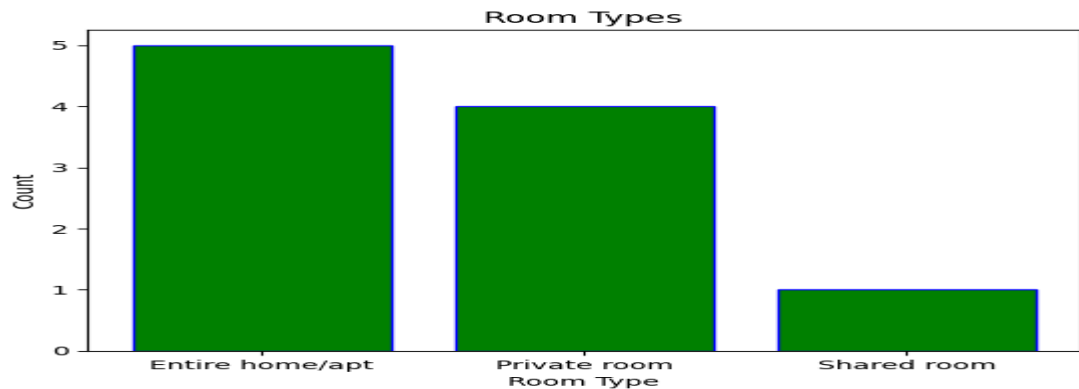|	index	neighbourhood_group
0	Manhattan	21661
1	Brooklyn	20104
2	Queens	5666
3	Bronx	1091
4	Staten Island	373

➢ Which hosts are the busiest and why?
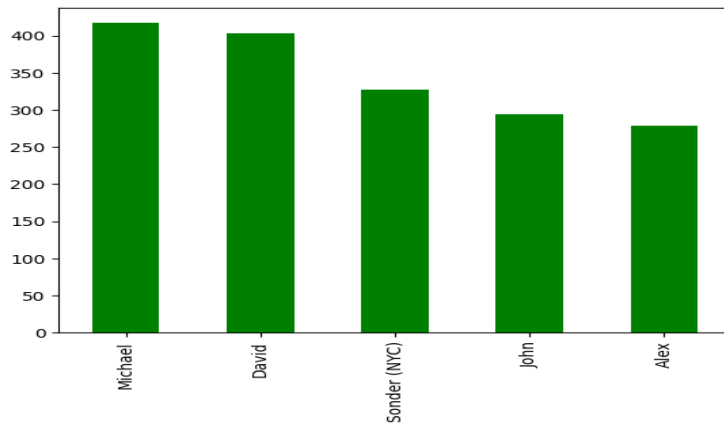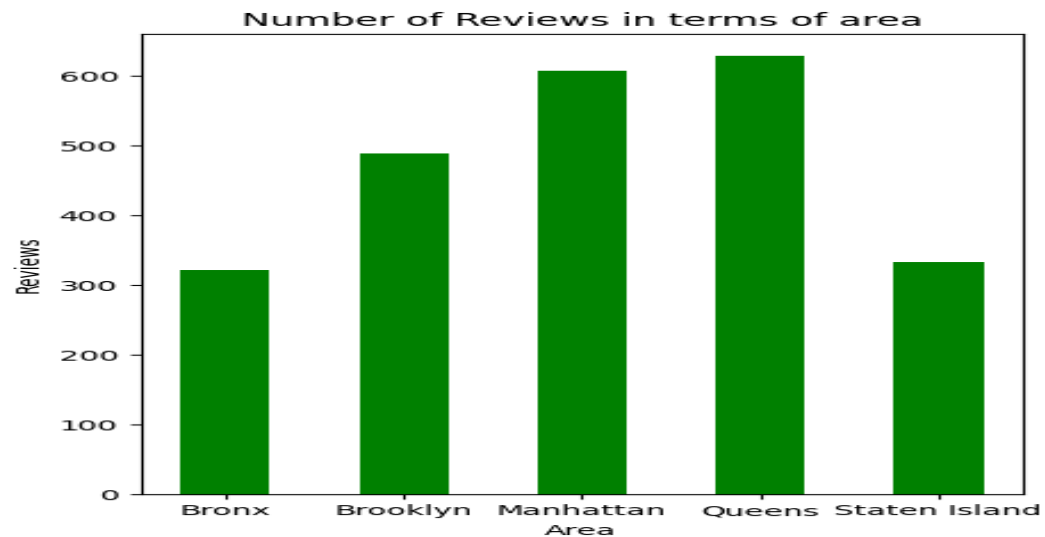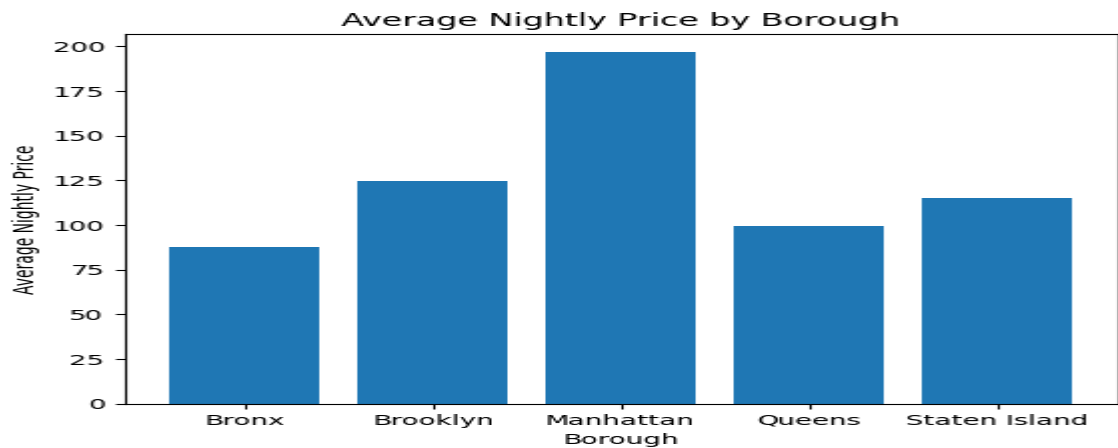

➢ Which Hosts are charging higher prices?
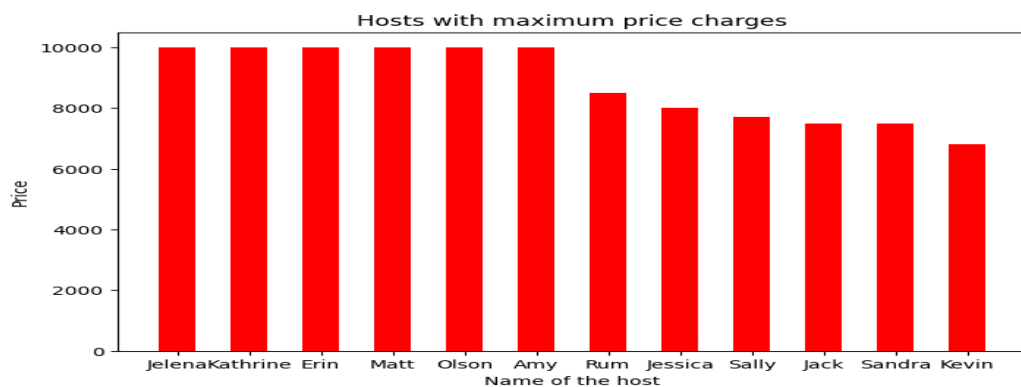
➢ What are the most popular types of properties available on Airbnb in New York City?

Room Types

➢ What can we learn from Data? (ex: locations, prices, reviews, etc)

Number of Reviews in terms of area

Number of Reviews VS Price
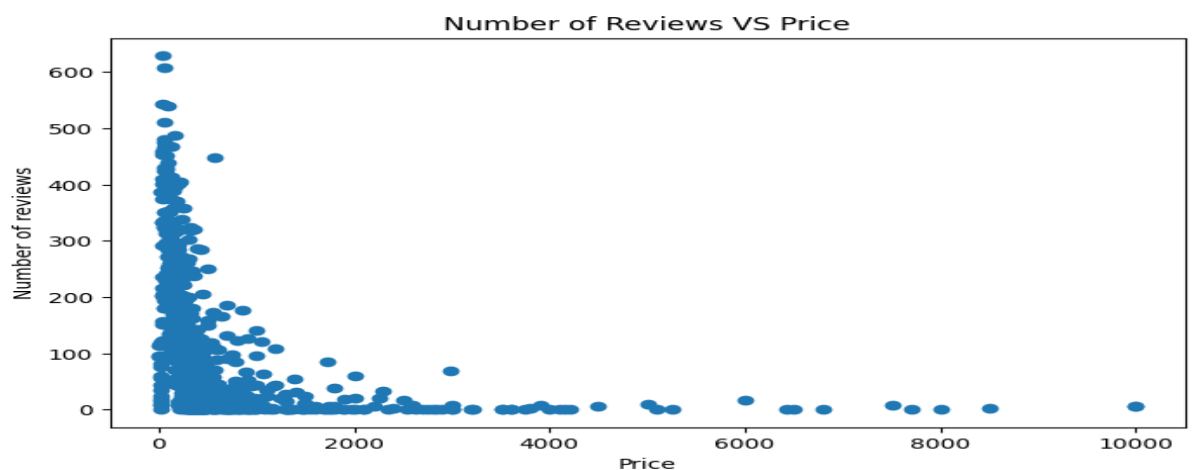
➢ What is the average nightly price of Airbnb listings in New York City, and how does it vary by borough

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.

- **Data Exploration**

  Exploring the dataset to understand the structure of the data, identify any patterns or trends, and detect outliers.

- **Feature Engineering**

  Creating new features from the existing data to gain more insights and improve the predictive power of the model.

- **Data Visualization**

  Creating various charts and graphs to visualise the data and communicate the insights gained from the analysis.

- **Statistical Analysis**

  Conducting statistical tests to validate the insights and hypotheses generated from the data.

● Standardization of features

Our main motive through this step was to scale our data into a uniform format that would allow us to utilise 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:

Throughout this datase, based on the findings of the EDA project on Airbnb booking analysis, we can conclude that the Sonder(NYC) and Blueground hosts have listed the highest number of listings in Manhattan. The most common room type preferred by visitors is Entire home/apt, and prices are relatively high in Brooklyn and Manhattan for this room type. Visitors also tend to prefer staying in rooms with lower prices and higher reviews. The busiest hosts are those who list their room types as Entire home and Private room, and their reviews are generally higher. The top 10 hosts charging the maximum price have properties located in Manhattan, Brooklyn, and Queens. Manhattan has the highest number of Entire home/apt listings, while Brooklyn, Queens, and Bronx have more private room types. This analysis can help inform future strategies for hosts and guests to optimize their Airbnb experiences.