Capstone Project-1
Airbnb Bookings Analysis

Team Project by

Prakshita Agrawal

Deepak Karki

Bhatu Sonawane

Bharat Soni

Sopan Wadekar

TABLE OF CONTENTS

1. Introduction
2. Data description and Data pre-processing
3. Exploratory Data Analysis and Visualization
4. Inferences and conclusions.

Introduction:

Airbnb

Airbnb is **a service that lets property owners rent out their spaces to travelers looking for a place to stay**. It simply provides a platform where travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves.

How Airbnb Works:

Find your room to stay → Check the details → Make a booking → Pay the rent

Objective:

- Airbnb dataset contains 49,000 observations and 16 columns which specifies about the hosts, customer and places and it is a mix between categorical and numeric values.
- Our goal is to explore and analyse the data, provide helpful conclusions through Exploratory Data Analysis build a statistical model that could be used to effectively predict the price for the listings and future decision makings.

Numerical

- id
- price
- Reviews per month
- longitude
- latitude
- Minimum nights
- No of reviews
- calculated_host_listings_count
- Availability 365
- Host_id

Categorical

- neighbourhood
- name
- Room type
- Neighbourhood_group
- Host name

EXPLORATORY DATA ANALYSIS

What is Exploratory Data Analysis?

EDA is an approach to analyze the data using visual techniques. It is used to identify trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

A detailed analysis and pre-processing are done in the dataset. It gave us a better idea of contribution of features towards the target variable.

Why is EDA important?

- Explore data
- Helps to identify patterns,
- Visualize the data,
- Understand the features.

Checking for Null values:

- It's always better to handle the null values before starting with further analysis in order to get best results.
- Columns **last_review** and **reviews_per_month** contain many null values. So removed both columns from the dataset
- Null values of name and **host_name** can be filled by **fillna** method as these are less.

Before treating

name	16
host_id	0
host_name	21
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
last_review	10052
reviews_per_month	10052

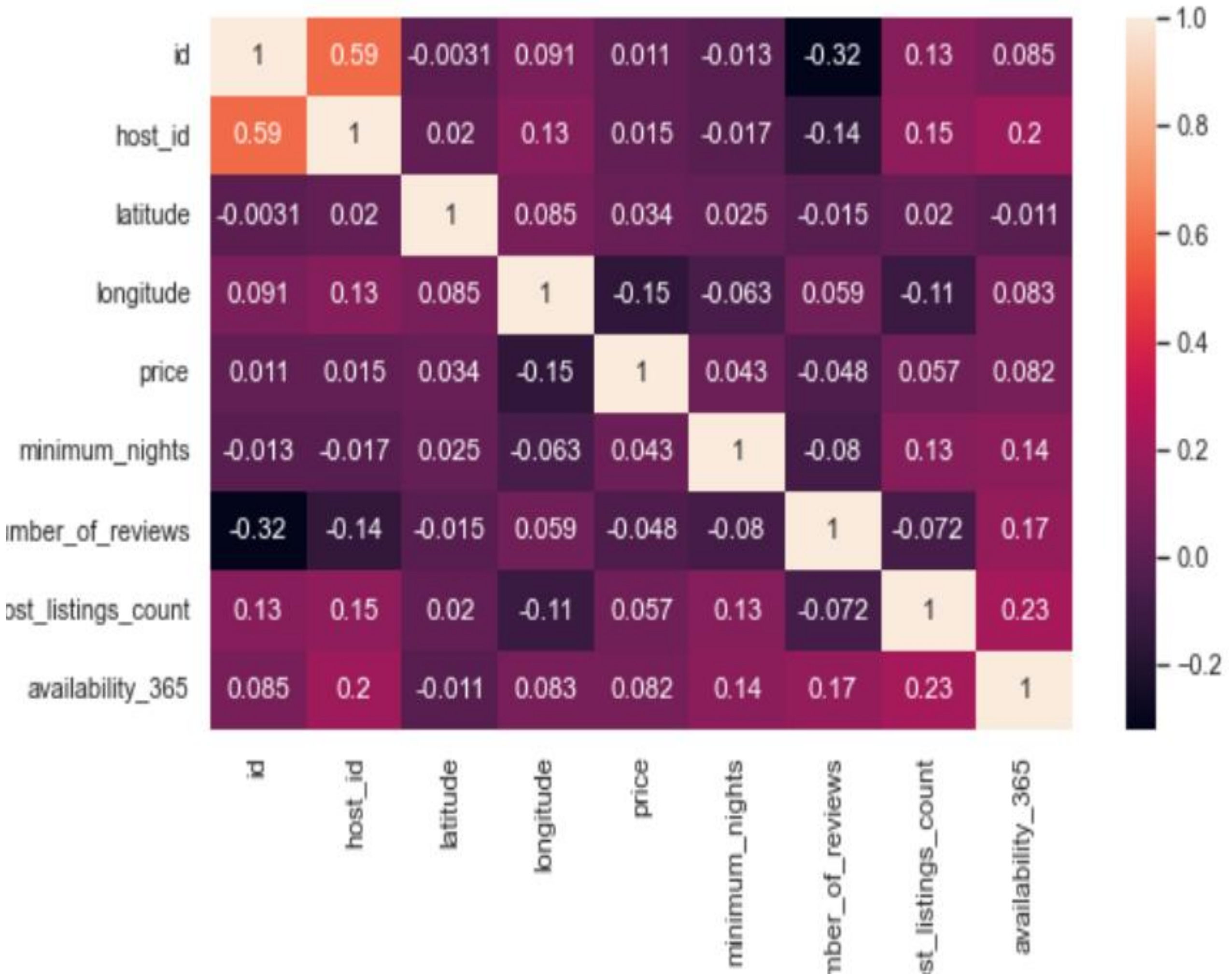
After treating

name	0
host_id	0
host_name	0
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
calculated_host_listings_count	0
availability_365	0

Correlation and Heatmap of all variables:

- Heatmap gives a correlation matrix to quantify and summarize the relationships between the variables

- From the above plot we can see that there is not much observable correlation between variables



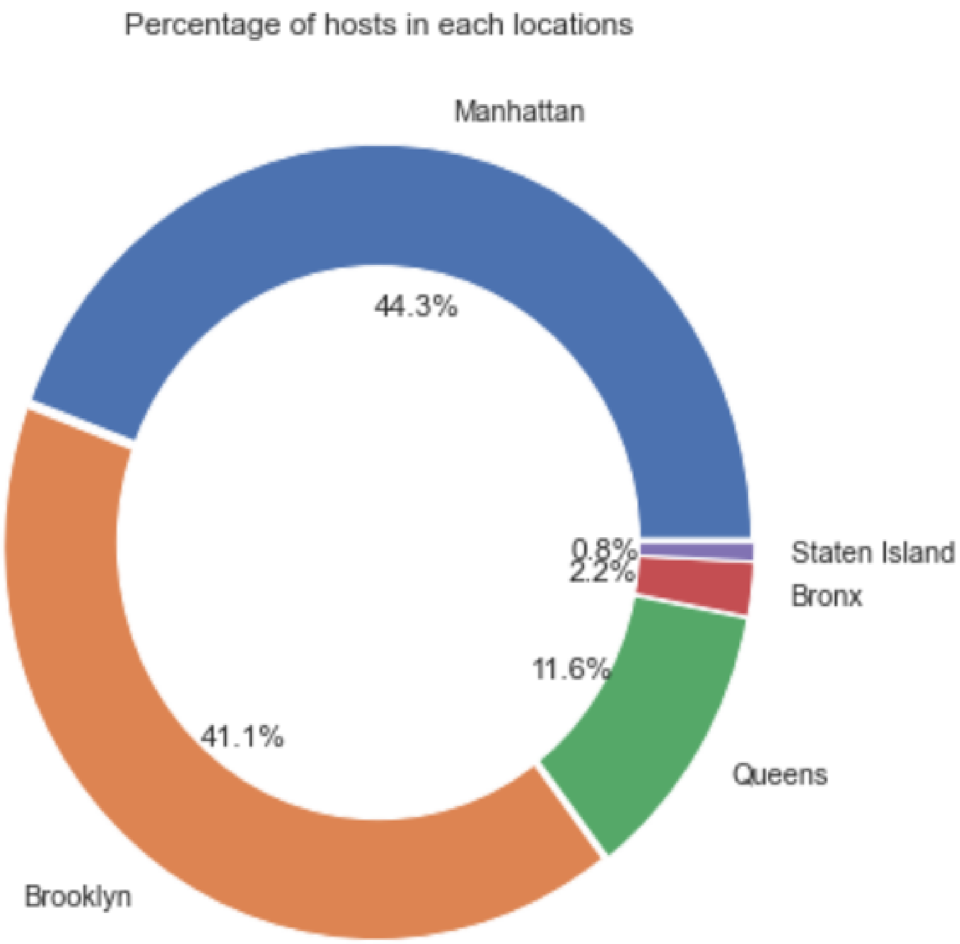
What can we learn about different hosts and areas?

Following points can be understood from the pie chart:

✓Manhattan and Brooklyn are home to 85.4% of the hosts followed by Queens with 11.6%.

✓Bronx and Staten Islands are occupied by only 3.0% of the hosts.

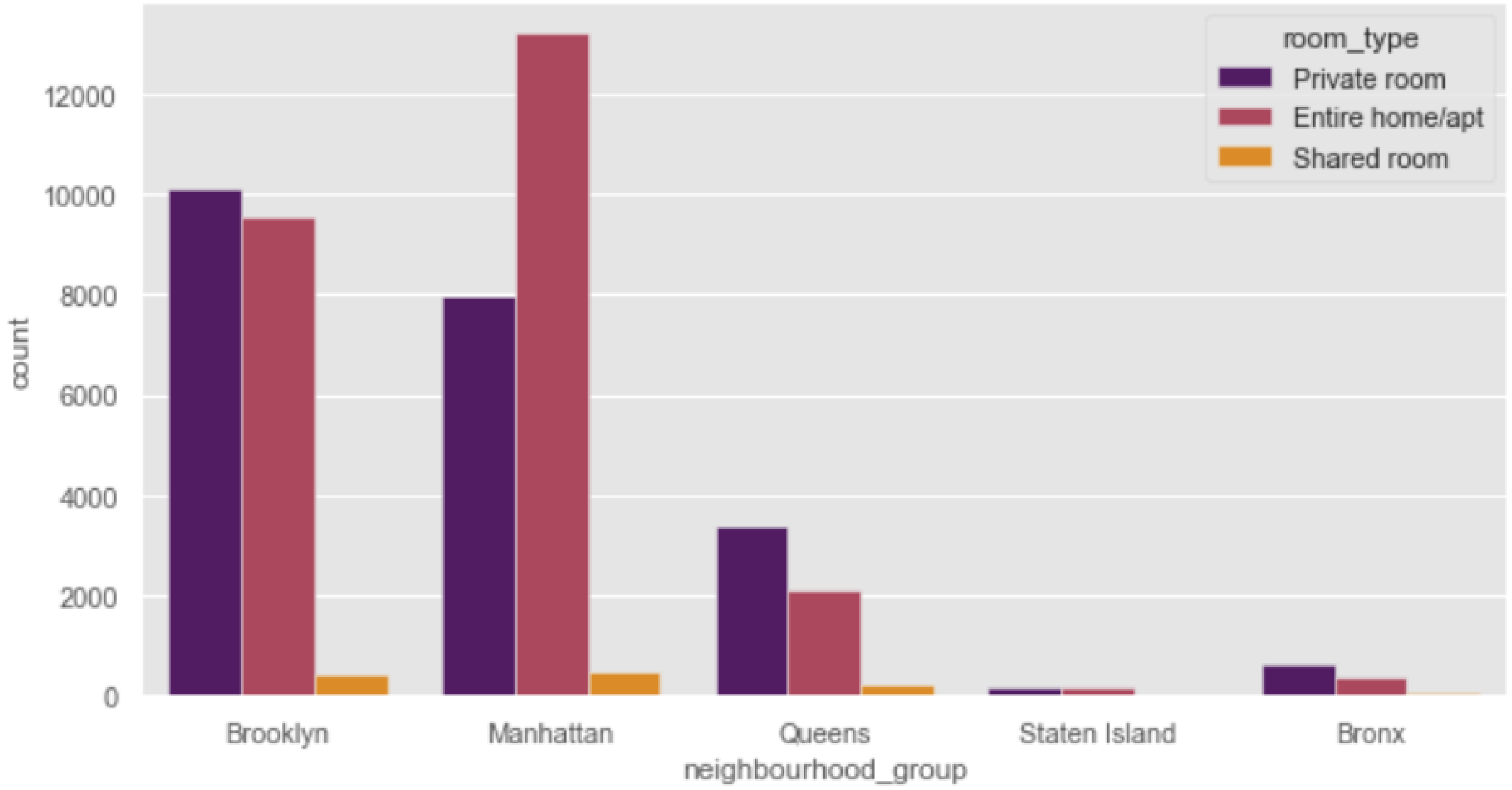
✓ It is clear that majority of the hosts are belong to the locations **Manhattan** and **Brooklyn**, hence these are the most popular destinations.



Preferred room type in most popular neighbourhood

If people are looking for rooms in these areas of Manhattan and Brooklyn then hosts are providing either Private room or Entire home/apt.

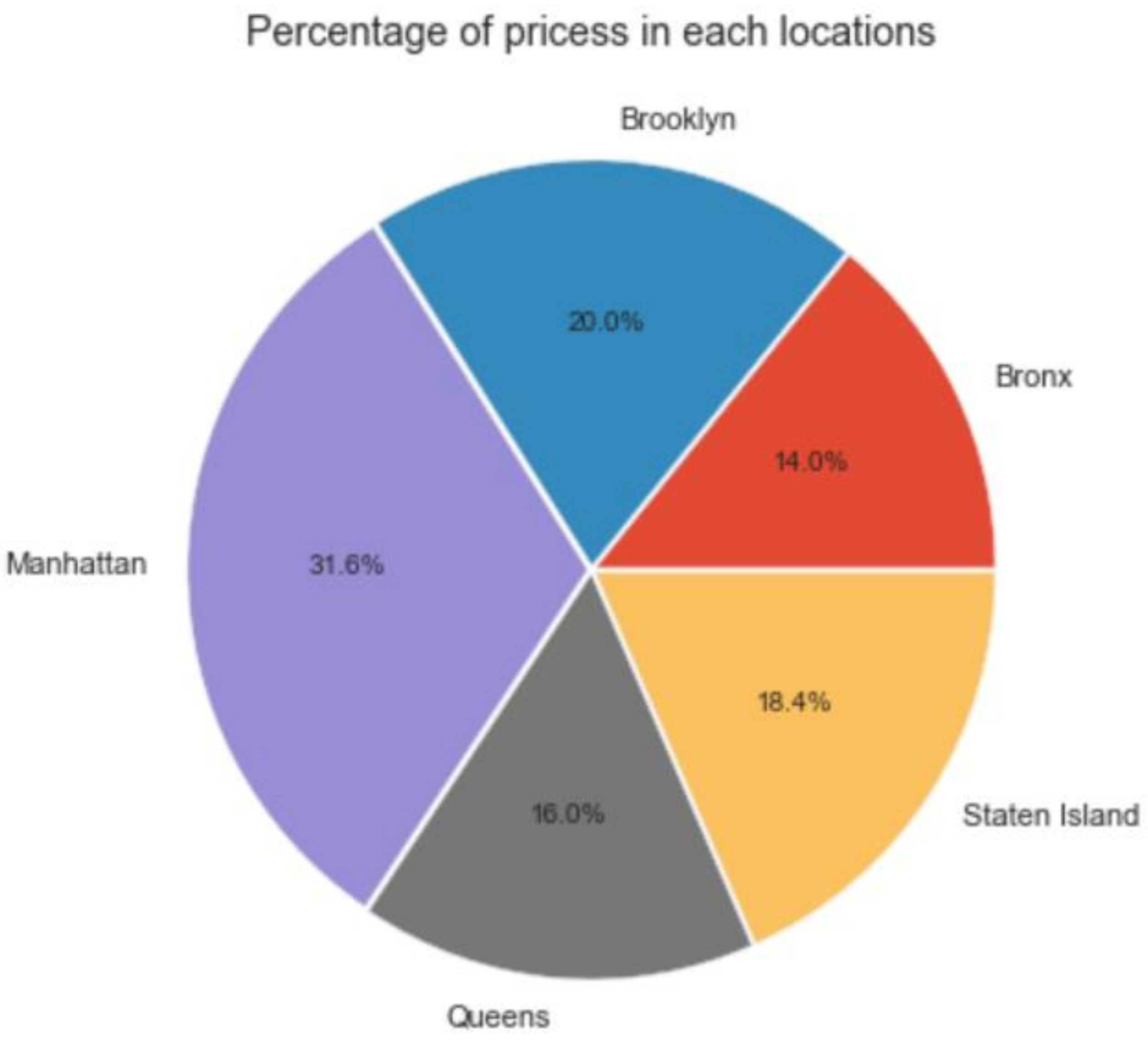
Shared rooms are least preferred in these areas.



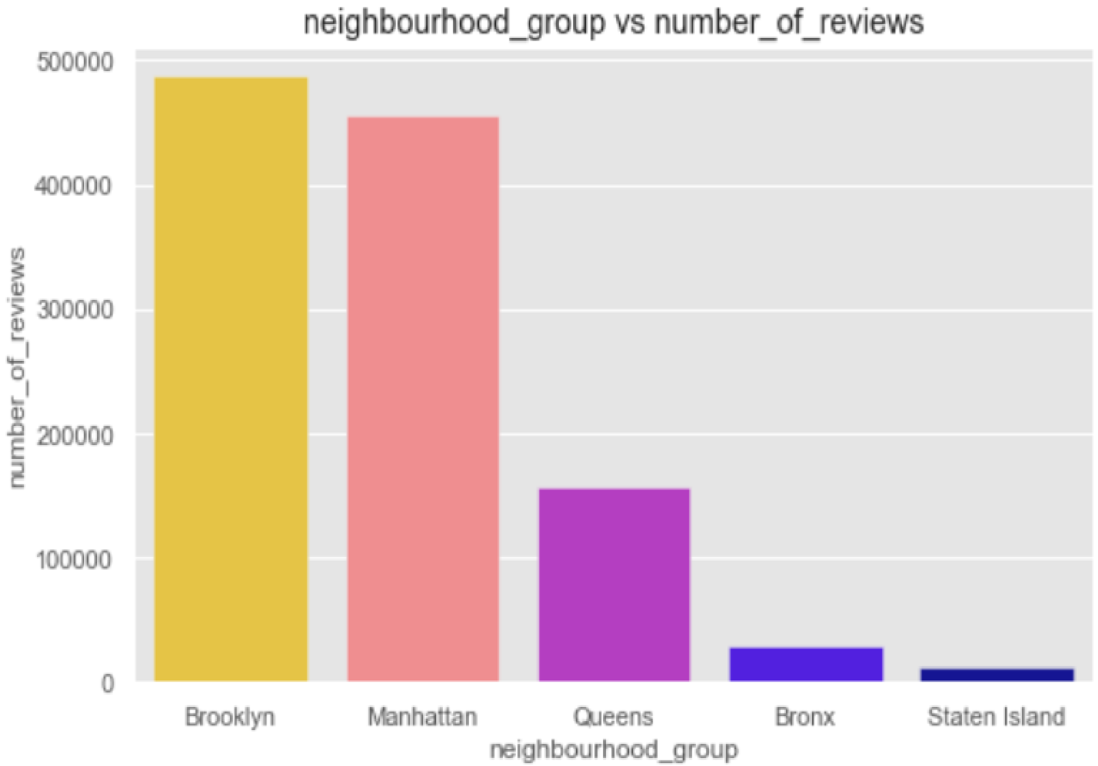
What can we learn from predictions?

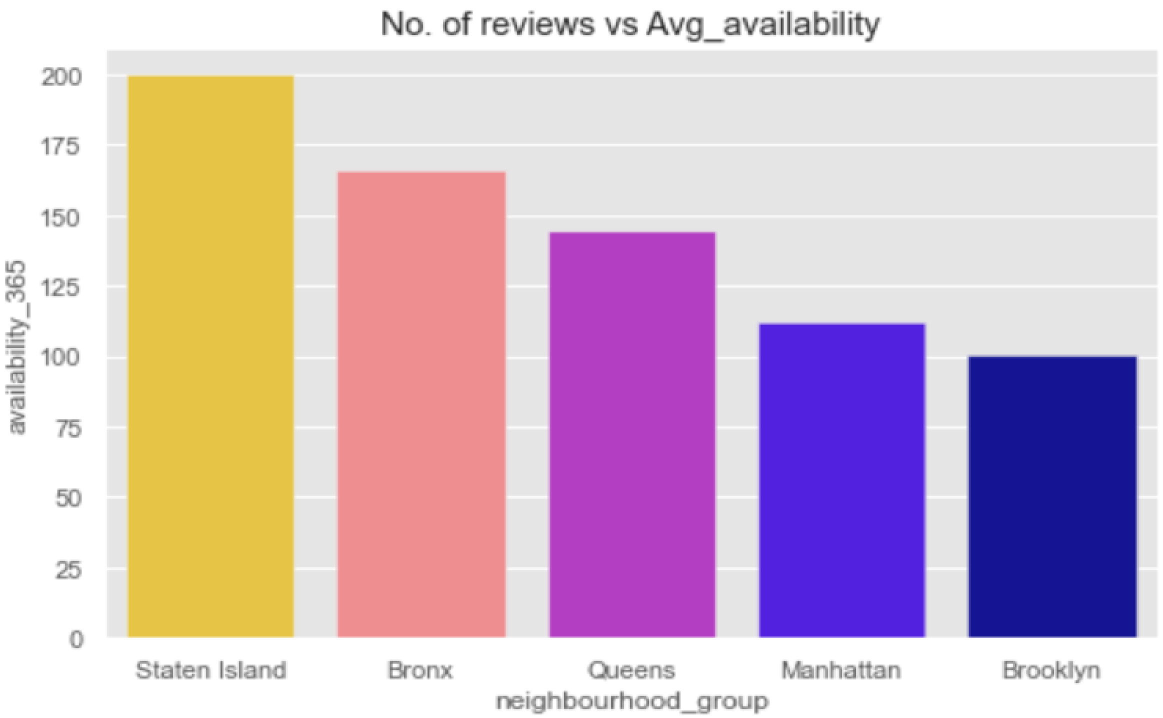
- Mean price is highest for Manhattan followed by Brooklyn and other locations.

- The higher number of hosts present in these areas might be the reason for these high prices.



Which hosts are the busiest and why?



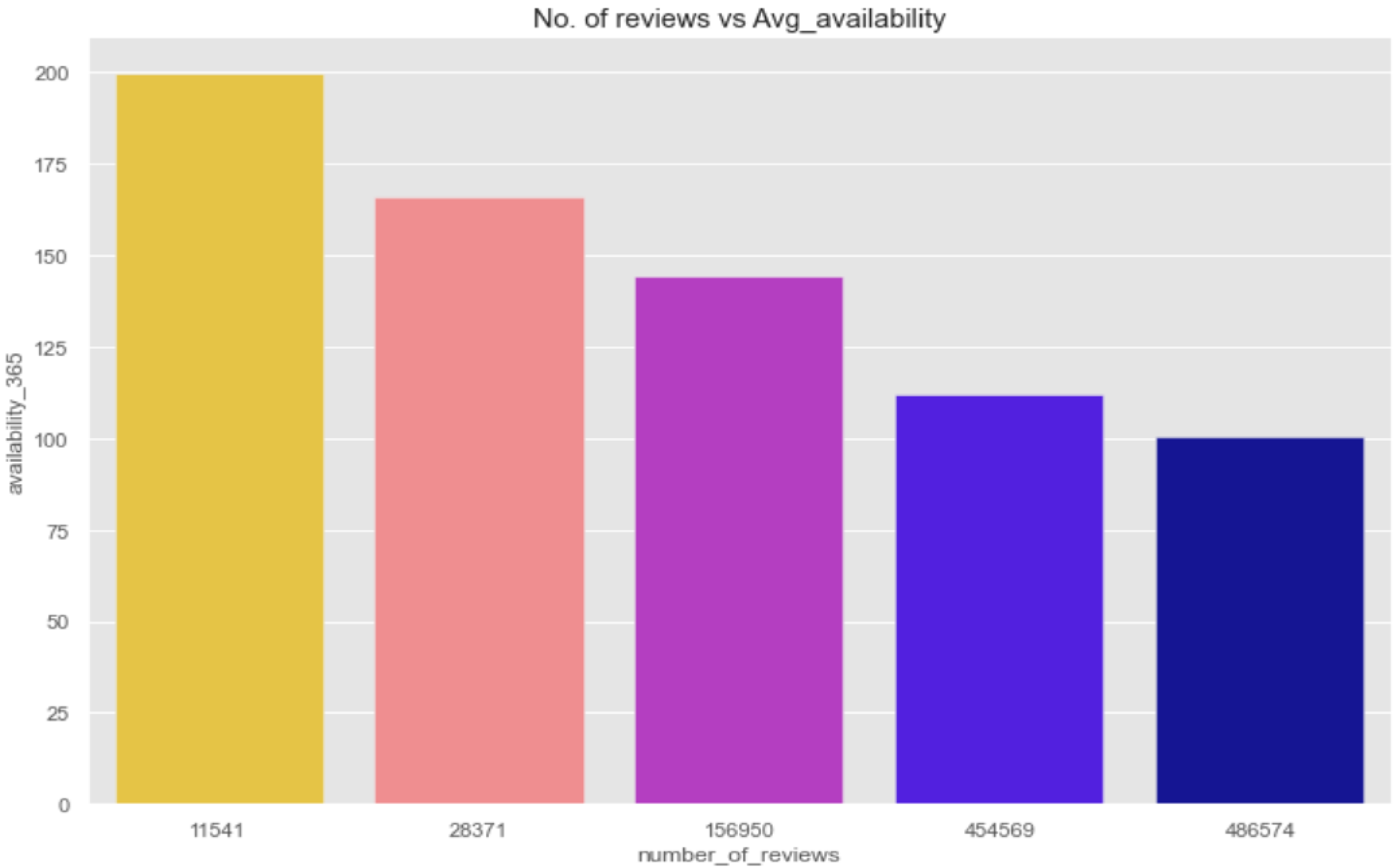


Manhattan and **Brooklyn** receiving the most number of reviews thus making the host busiest

Manhattan and **Brooklyn** have maximum number of reviews, so offering the most desired room types. Thus for these groups availability of rooms is less.

The bar plot shows the relation of number of reviews with the availability of these rooms throughout the year.

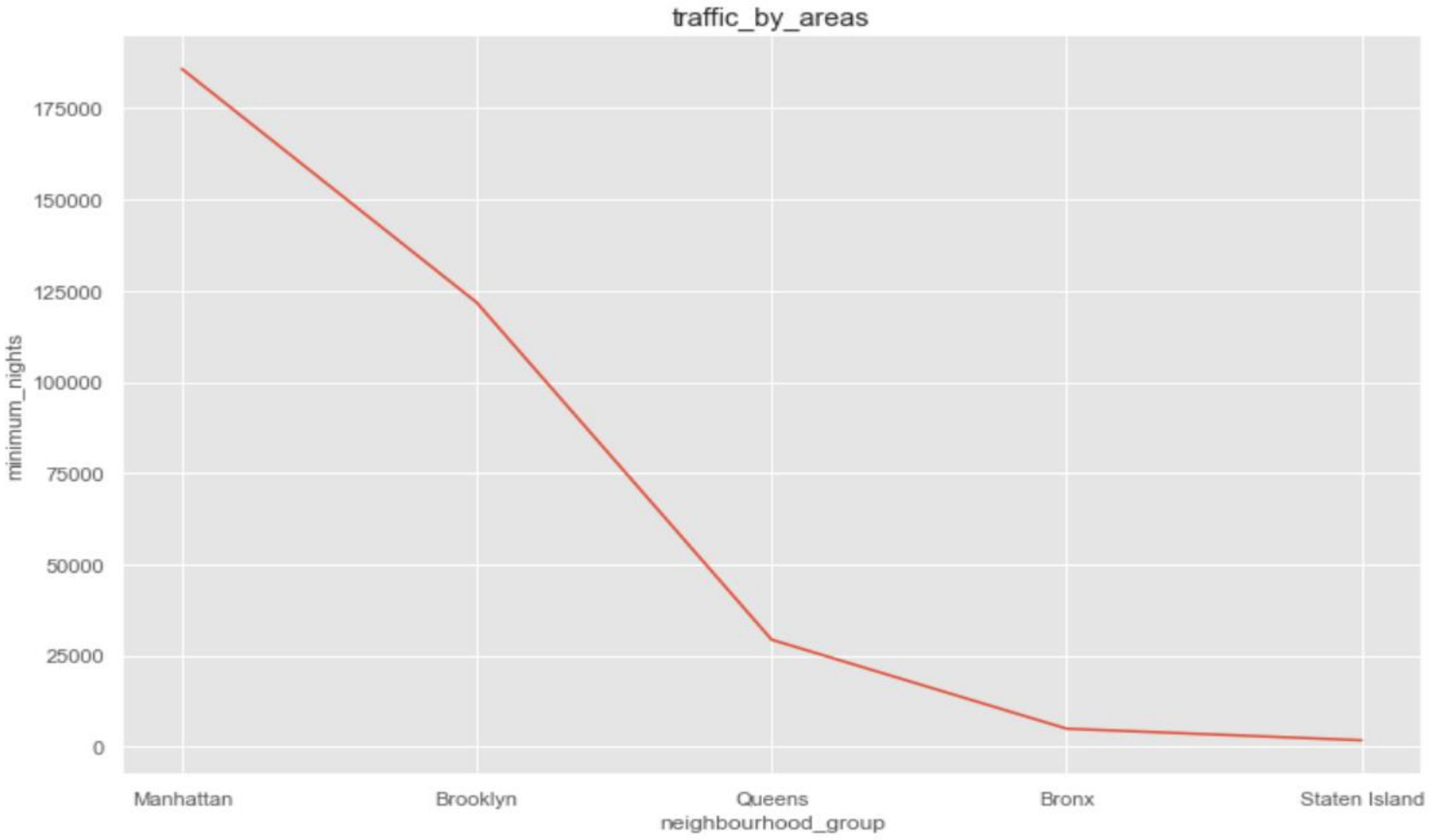
This relation shows that as the number of reviews increase the availability decreases.



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

A huge difference in the traffic between the different locations. Reasons may be:

- **Maximum no of reviews**
- **Most preferred room types**
- **More host count**



Inferences and conclusions

- This Airbnb (NYC 2019) Dataset For The Year 2019 Appeared to be very rich Dataset with a variety of columns that allowed us to do deep data exploration.
- In the column name and host_name which have 16 and 21 null value only. Null values are present in **last_review** and **reviews_per_month** which can be dropped both have most null values is 10052.
- From the dist plots it can be observed that latitude and longitude data seem to be normally distributed and most of the numeruc_features are positively skewed.
- People stay for longer duration of time in private rooms in **Brooklyn** and **Manhattan**.
- More customers preferred **Manhattan** location for night stay then **Brooklyn**.
- Entire **home/apt'room** type has the highest number of listing of 52% and shared room is the least listed room type at only 2.4% in total.
- 63.2% costumer spend night in entire home and 1.6% spend night in shared room.

- As the number of reviews go higher the availability decreases indicating that the **busiest hosts** are the people receiving the most reviews.
- **Manhattan** and **Brooklyn** are the two top most popular **neighbourhood groups** in terms of **hosts count, number of reviews ,number of listing, maximum number of nights spends** in these areas. So it might also be reason of traffic and high prices.
- For other **neighbourhood groups** namely **Queens, Bronx** and **Staten island** there aren't as popular as these two, especially on **Staten Island**.
- Reviews obtained by **Manhattan** and **Brooklyn** locations contribute towards the traffic difference, as people tend to book stays with higher number of reviews.
- The dataset can be further used for price prediction by building a linear model. The data needs to be treated of outliers and skewness for a linear regression as well as other models.

Thank You