EDA on Airbnb Booking Analysis

By

Patan Ismail Alli Khan

Data Science Trainee

AlmaBetter, Bangalore

Abstract:

Airbnb stands for "Air Bed and Breakfast," it is a service that lets property owners rent out their spaces to travelers looking for a place to stay. Travelers can rent a space for multiple people to share, a shared space with private rooms, or the entire property for themselves.

Airbnb is based on a peer-to-peer business model. This makes it simple, easy to use, and tends to be more profitable for both parties. The model also gives you the opportunity to customize and personalize your guest's experience the way they want.

1. Problem Statement:

For this project, we will be analysing Airbnb's New York City (NYC) data of 2019. This dataset contains listings information such as listing name, host name, room types, neighbourhood group (location), neighbourhood (area), minimum night to be paid for, availability of listing in days and number of reviews etc.

Our main objective behind this project is to explore and analyse the data to discover the key understandings. For this, we will explore and visualize the dataset from Airbnb in NYC using basic exploratory data analysis techniques.

2. Data Overview:

The data has 48895 rows and 16 columns. The 16 columns are:

1. **id:** Unique id of listing
2. **name:** Name of the listing
3. **host_id:** Unique id of host
4. **host_name:** Name of the host
5. **neighbourhood_group:** Location of the listing
6. **neighbourhood:** Area of the listing
7. **latitude:** Latitude of listing
8. **longitude:** Longitude of listing
9. **room_type:** Type of rooms.
10. **price**: Price of listing
11. **minimum_nights:** Minimum number of nights to be paid for
12. **number_of_reviews:** Number of reviews given for listing.
13. **last_review:** Date of last review given for the listing.
14. **reviews_per_month:** Number of reviews given per month.
15. **calculated_host_listings_count:** Total Number of listings for host.
16. **availability_365:** N.o of days listing is available.

3. Introduction

Since 2008, guests and hosts have used Airbnb to expand on travelling possibilities and present a more unique, personalised 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.

3. Data Understanding:

After loading the dataset, understanding the data is very important. We need to understand the various features of the dataset and their meaning. We need to identify the numerical and categorical features.

4. Data Cleaning:

Our dataset has large number of null values. Out of 16 features, only 4 features have null values. Features such as “name” and “host_name” had few null values. As we are not using these features we are not handling these null values.

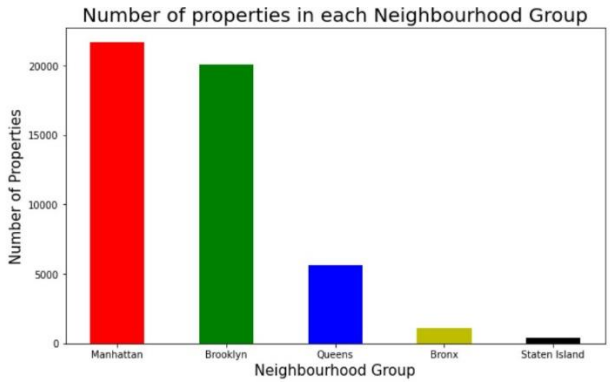
Features named ‘last_review’, ‘reviews_per_month’ has 10052 null values. We replaced reviews_per_month with 0 as these properties might not have been rated at all. As last_review is a date column, we cannot replace with any random date as it would corrupt the dataset. So, we are leaving last_date as it is.

5. Exploratory Data Analysis:

After performing above steps our dataset is ready for Exploratory Data Analysis. I started by exploring key features such as price, number of reviews, minimum nights and their variation with respect to different neighbourhood and neighbourhood groups. I have checked for any correlation between any features using the Heatmap. There is no correlation between any of the features. Then I distinguished the numerical and categorical data and created few categorical features for better visualizing. And then I started to explore categorical features and how these categorical features vary with numerical features and drawn some conclusions from it. And I visualize those observations using bar chart, scatter plot, violin plot, pie chart, multiple bar chart.

Following are some analyses which I have done.

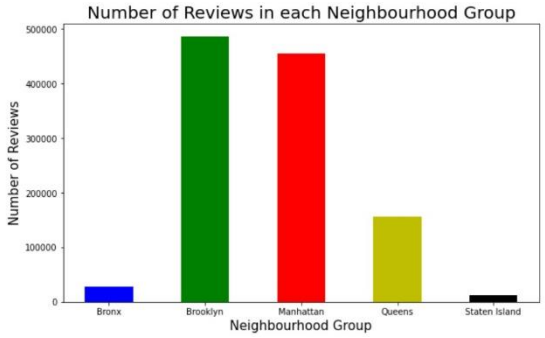
Which location has more number of properties?



There are five locations in NYC namely Manhattan, Brooklyn, Queens, Bronx and Staten Island.

From the above graph we can see that Manhattan has more number of properties. Followed by Brooklyn and Queens. Whereas Bronx and Staten Island have the least number of Properties.

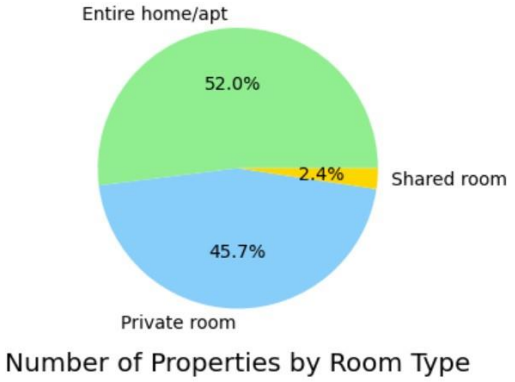
Which location has more Number of Reviews?



We can see that Brooklyn has the most number of Reviews. Followed by Manhattan and Queens.

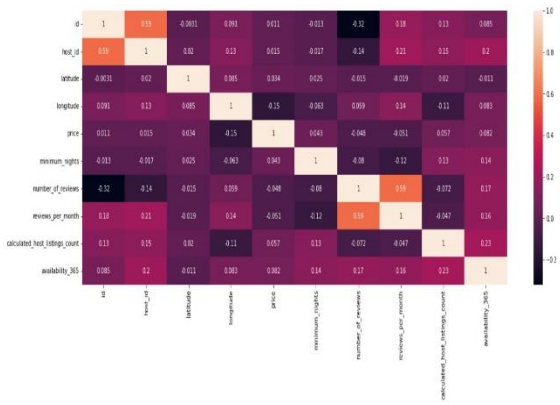
Which Room Type is offered by most Number of Properties?

From There are three room types namely Entire home or apartment, Private room and Shared room type.



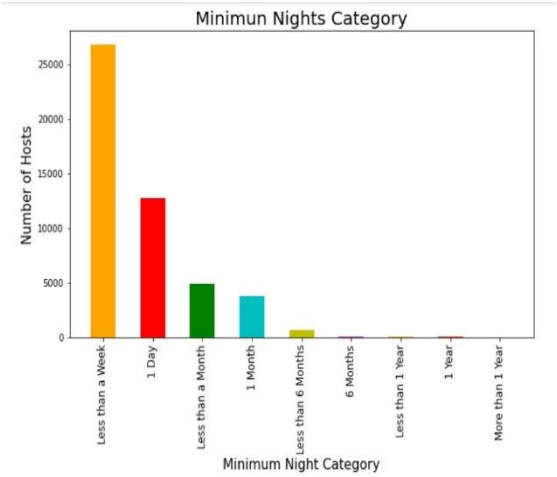
From the above Pie chart we can see that most properties offer Entire home or apartment type of property.

Is there any multicollinearity between any of the features?



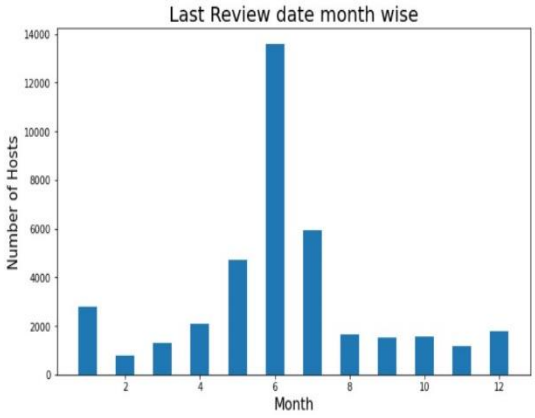
From the above heat map, we can say that there is no correlation between any of the features of dataset.

What is the minimum nights to be paid for in most Number of Properties?



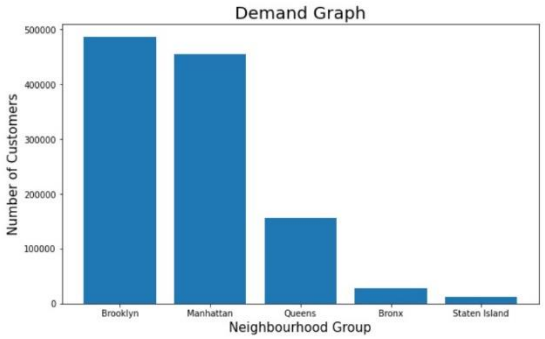
We can see that many properties prefer at least 1 day or less than as week to be paid for.

Which is the month when most people have vacated their properties?

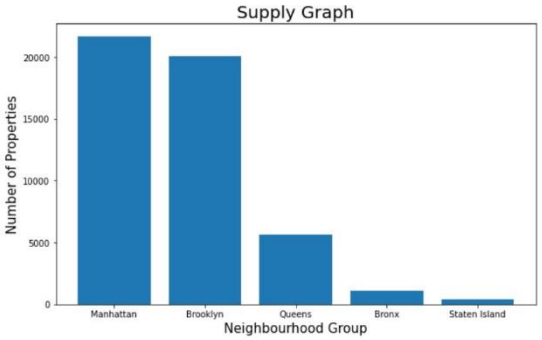


From the above bar graph we can see that many customers have either vacated or changed the property in the 6th month (June Month).

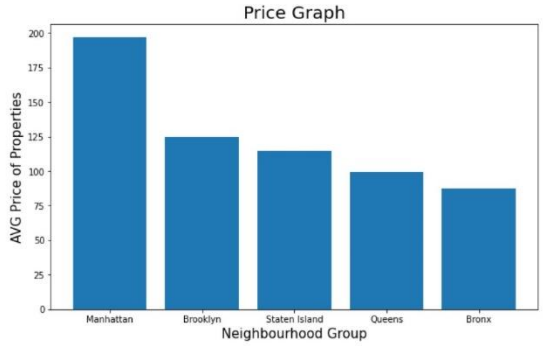
Which location has good demand and which location more properties need to be created to increase profits?



The above graph shows which location has more demand in terms of number of reviews. We can see that Brooklyn has the most demand.



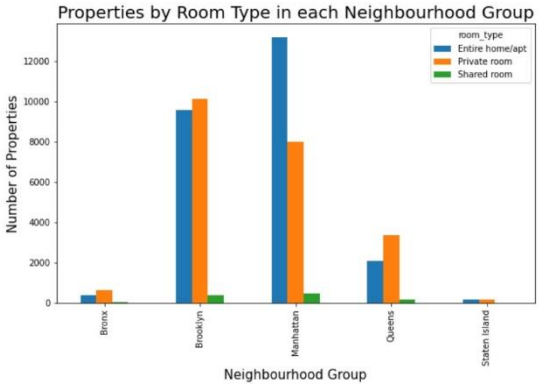
The above graph shows which location has more supply in terms of number of properties. We can see that Manhattan has the most number of Properties.



The above graph shows which location has the highest average price. We can see that Manhattan has the highest average price.

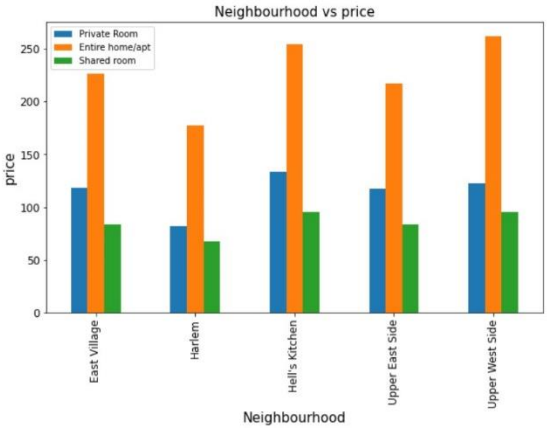
From the above three graphs, since Brooklyn has more demand and less supply we can suggest to increase number of properties in Brooklyn location to get higher Profits.

Which location has highest Number of Properties of each room type?

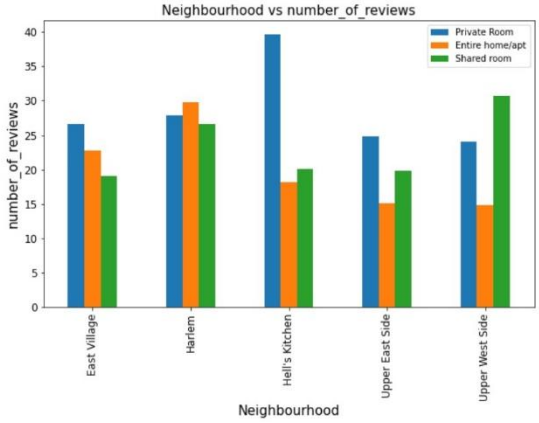


We can see that Manhattan and Brooklyn have the highest number of Entire home/apt Type of properties. There are very few shared rooms in almost all locations.

How Price and number of reviews for each room type vary with the Area of the property in Manhattan?

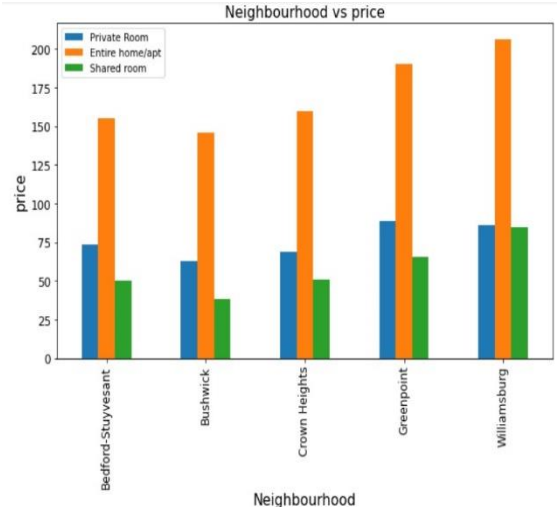


From the graph we can see that Hell's Kitchen and Upper West side Areas in Manhattan have Highest Average Price for all room types.

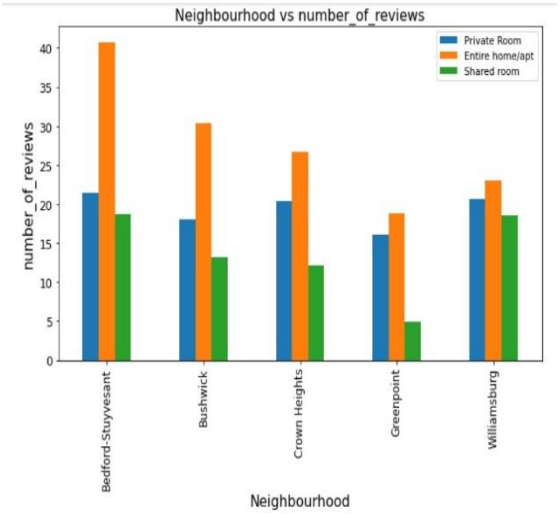


We can see that in the Hell's Kitchen Area of Manhattan most people prefer Private room Type of Properties.

How Price and number of reviews for each room type vary with the Area of the property in Brooklyn?

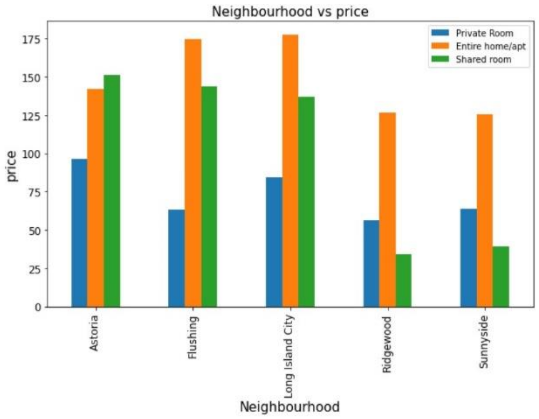


We can see that Williamsburg Area in Brooklyn has the Highest Average Price for all room types.

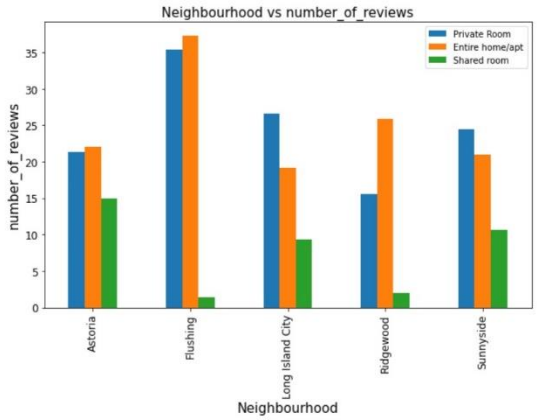


From the graph we can see that most reviews are for the Entire home/apt type in Brooklyn and Williamsburg has the least reviews for Entire home/apt type.

How Price and number of reviews for each room type vary with the Area of the property in Queens?

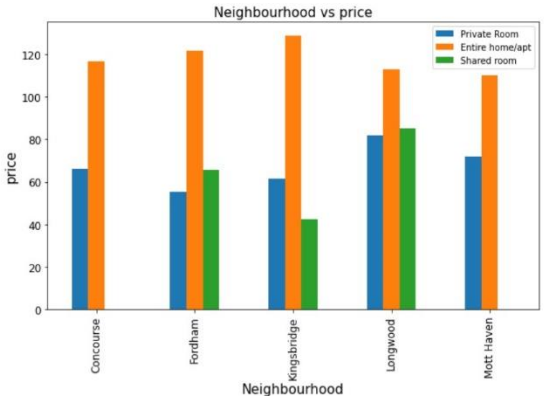


We can see that the maximum average price in Queens is very less compared to Manhattan and Brooklyn.

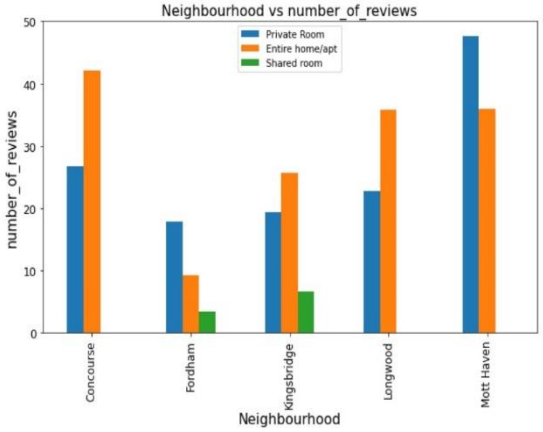


From the above graph we can see that Private Room Type in the Flushing Area has highest number of reviews compared to other Areas.

How Price and number of reviews for each room type vary with the Area of the property in Bronx?

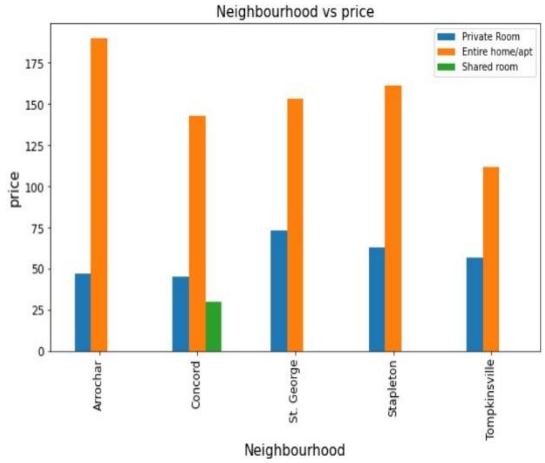


We can see that average price for Entire home/apt Type are high in almost all Areas in Bronx and Prices are even more lower than Queens.

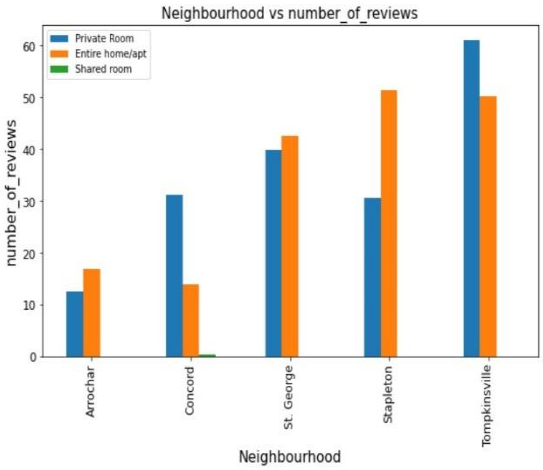


Most of the reviwes in Bronx are for Entire home/apt or Private room type.

How Price and number of reviews for each room type vary with the Area of the property in Staten Island?

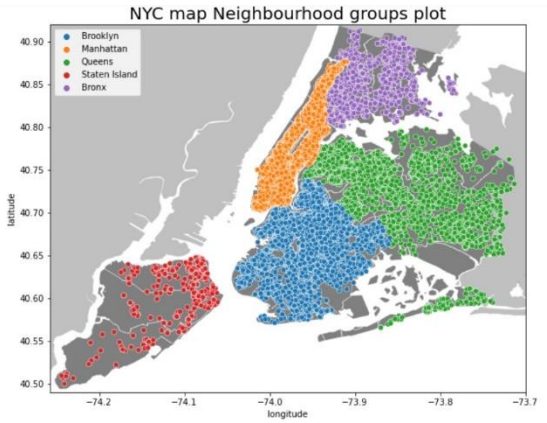


We can observe that except Concord Area, there are no shared rooms in any other Area.



From the graph we can see that Tompkinsville has the highest number of reviews. This might be because of the lower prices in this Area.

Overview of all the properties in the New York City



From the above graph we can see an overview of the properties listed in the New York City differentiated by locations.