
## CAPSTONE PROJECT - I

# AIRBNB BOOKING ANALYSIS

By:

Shrinidhi Choragi

Data Science Trainee

Almabetter

# Contents

# Introduction

Since 2008 Airbnb has  expand on travel possibilities and presented a more unique and personalised way of experiencing the world.
It is an American company that facilitates an online marketplace for lodging, primarily homestays for vacation, rentals, and tourism activities.

# Problem Statement

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

Explore and analyze the data to discover  the following key understandings.

- 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 density distribution among neighbourhood groups.

# Dataset

This dataset describes the listing activity and metrics in NYC, NY for 2019.

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

The features of the dataset are

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:  Number of reviews per month

Calculated_host_listings_count: Number of listings done by particular host

Availability_365: Number of days the property is available

# Data Cleaning

The Percentage of missing values in the dataset is found to be: 2.57%

Handling missing values:

- *host* and *host name*:

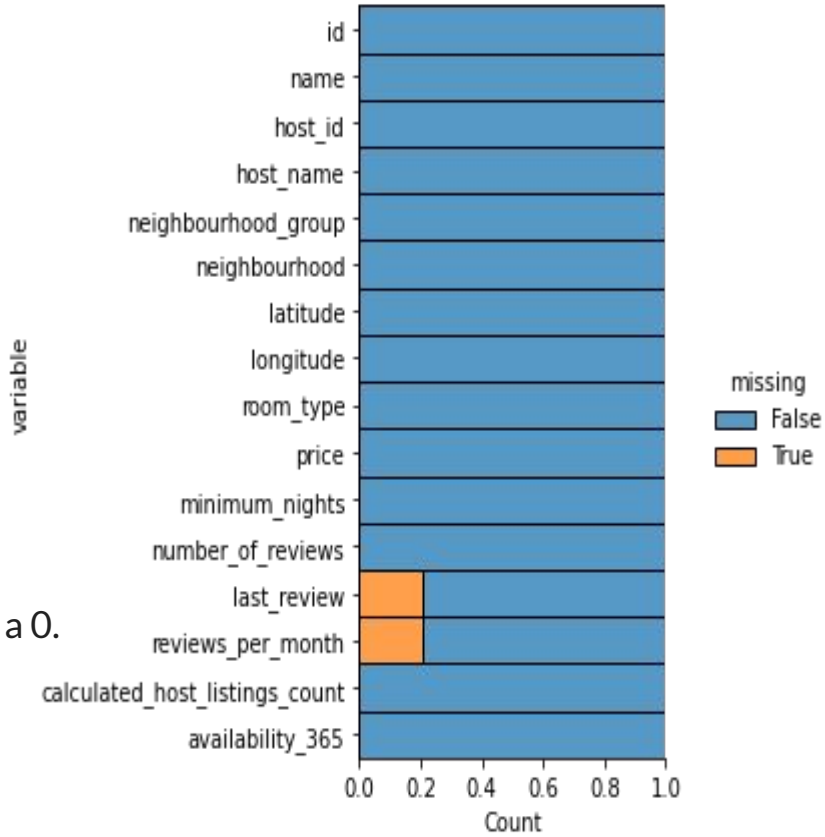
imputed with a dummy variable.

- *"review_per_month"*:

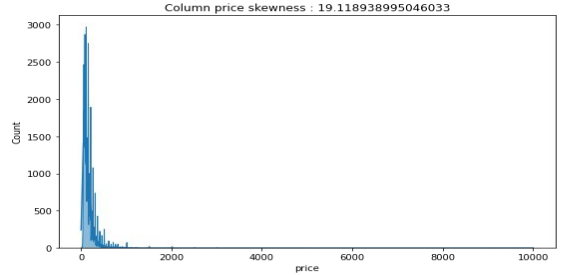
imputed with 0.0 for missing values since "*number_of_review*" of the corresponding column has a 0.

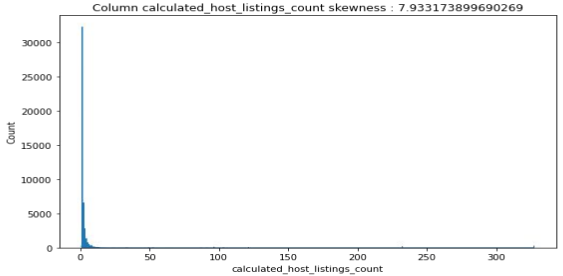
- "*last_review*" :

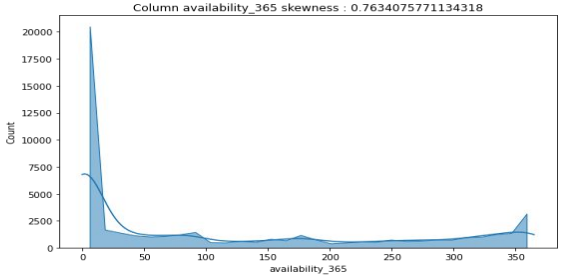
It's dropped.

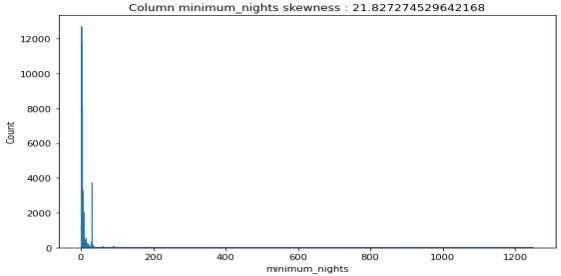


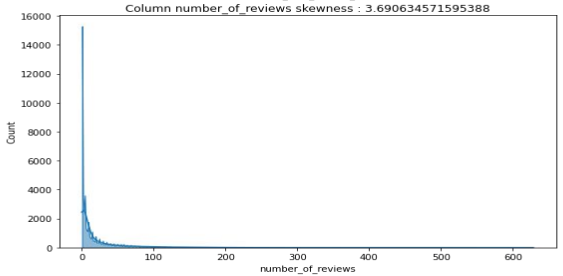
# Descriptive Statistics

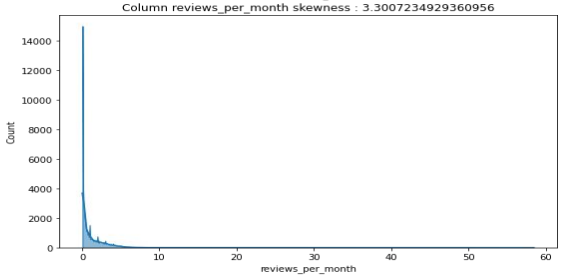


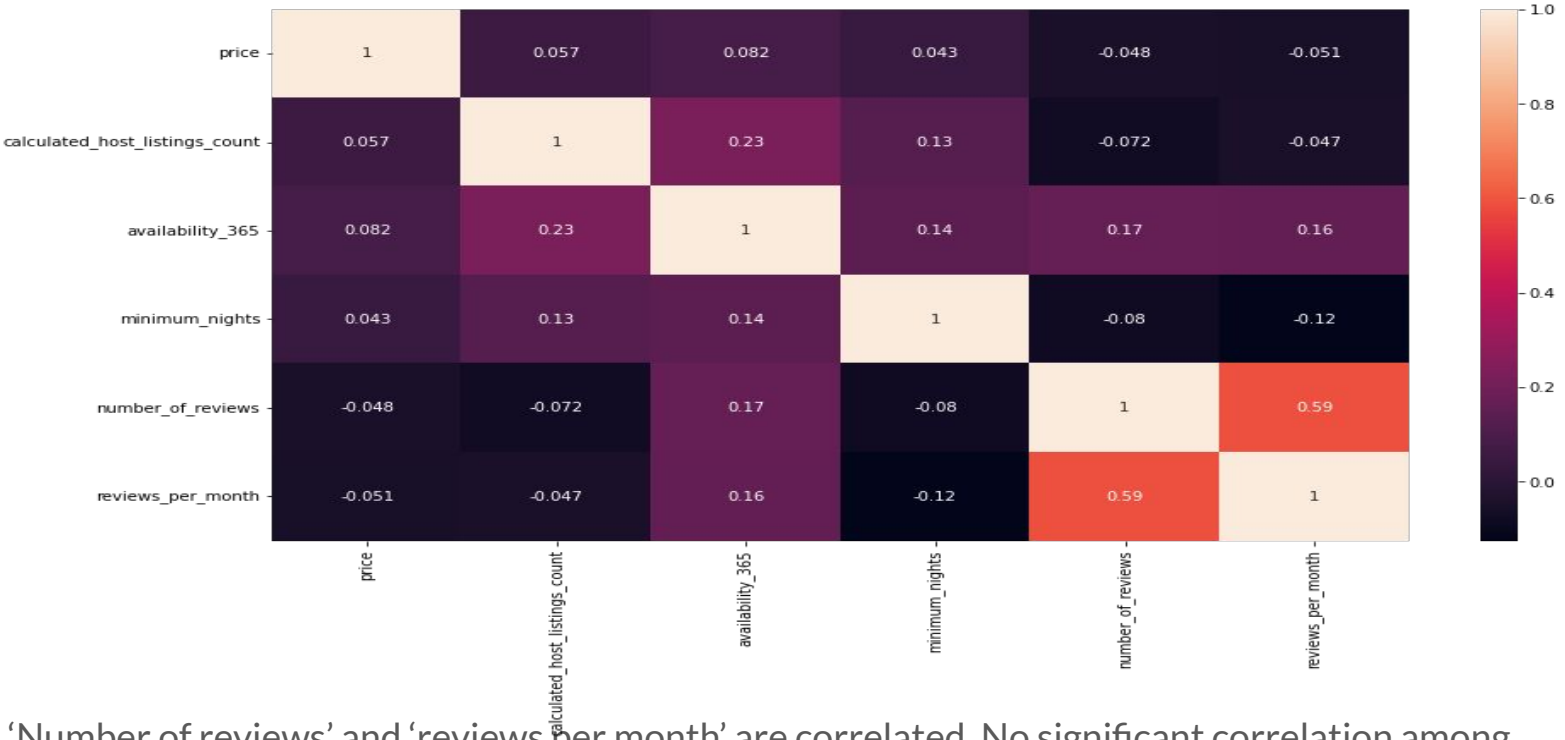




Most of the features are positively skewed

# Correlation Analysis

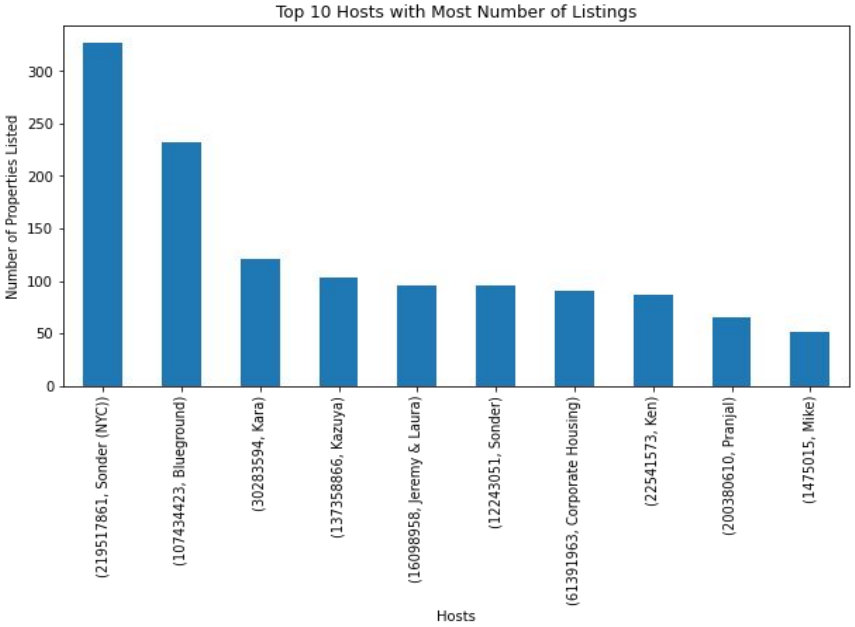



‘Number of reviews’ and ‘reviews per month’ are correlated. No significant correlation among others.

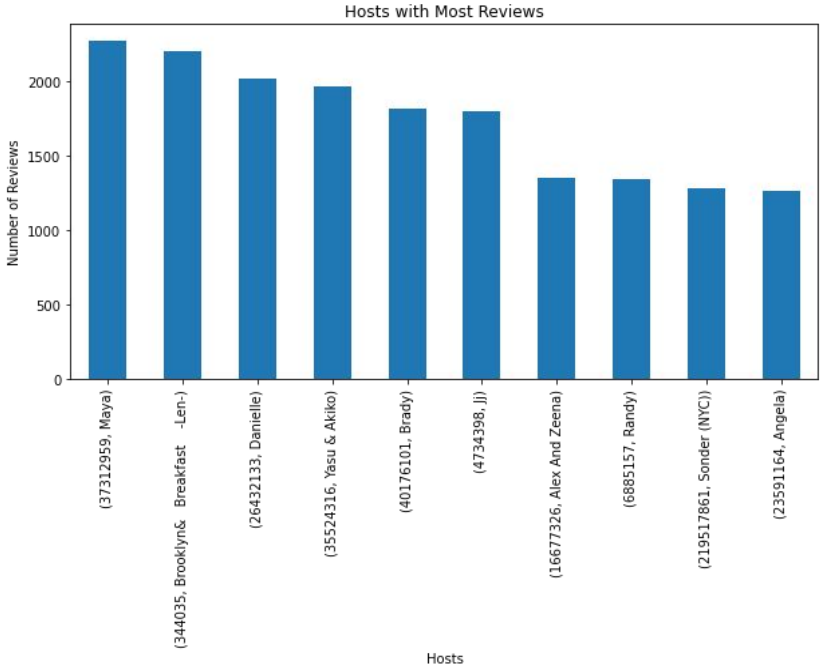
# Key Understandings :

- 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.

# Hosts

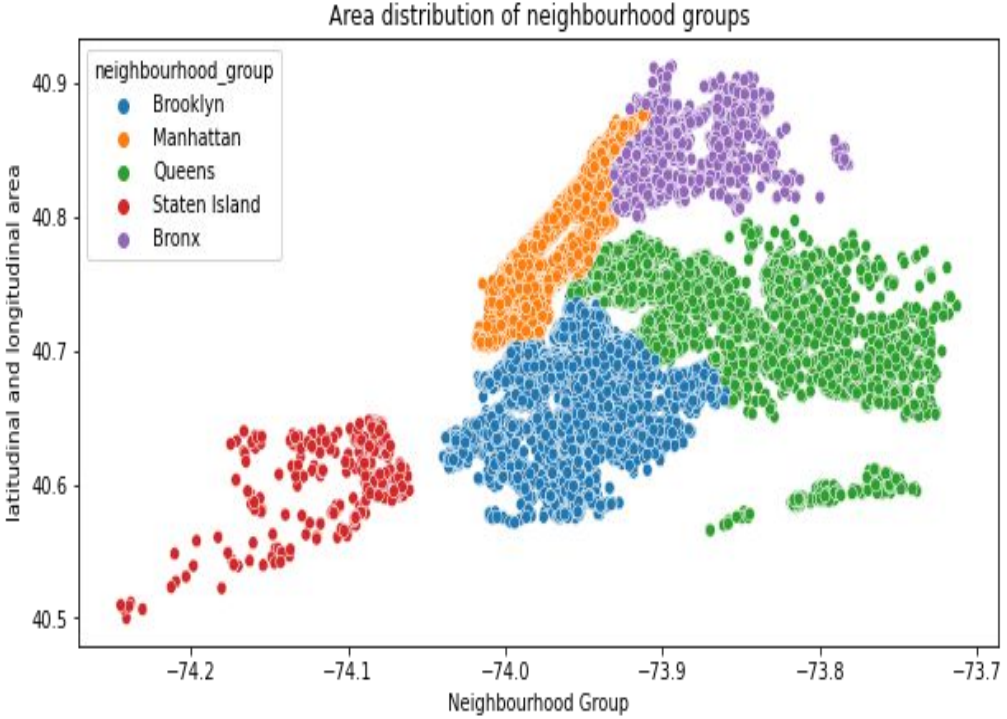


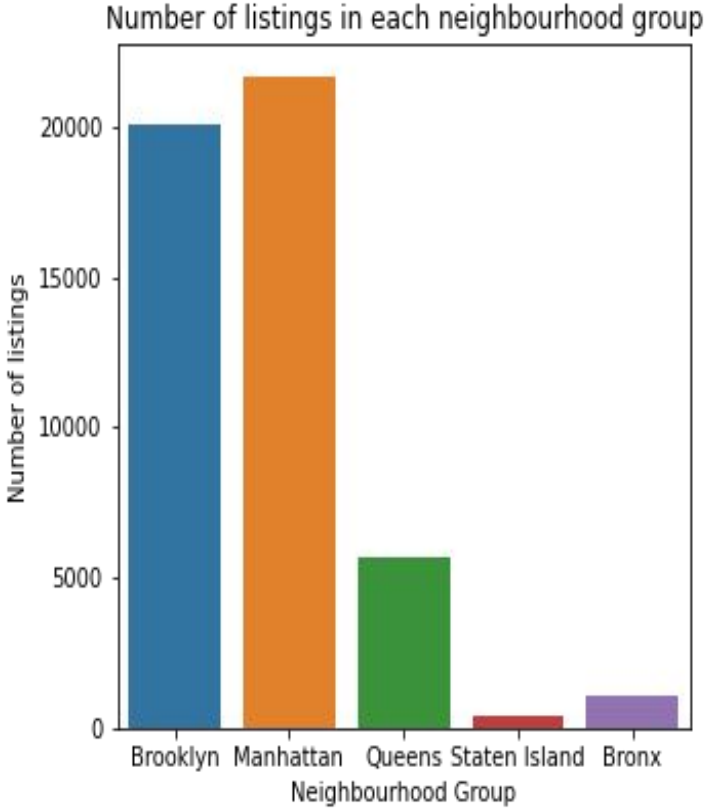

Sonder(NYC) has the most listings




(Maya,37312959) has the most reviews

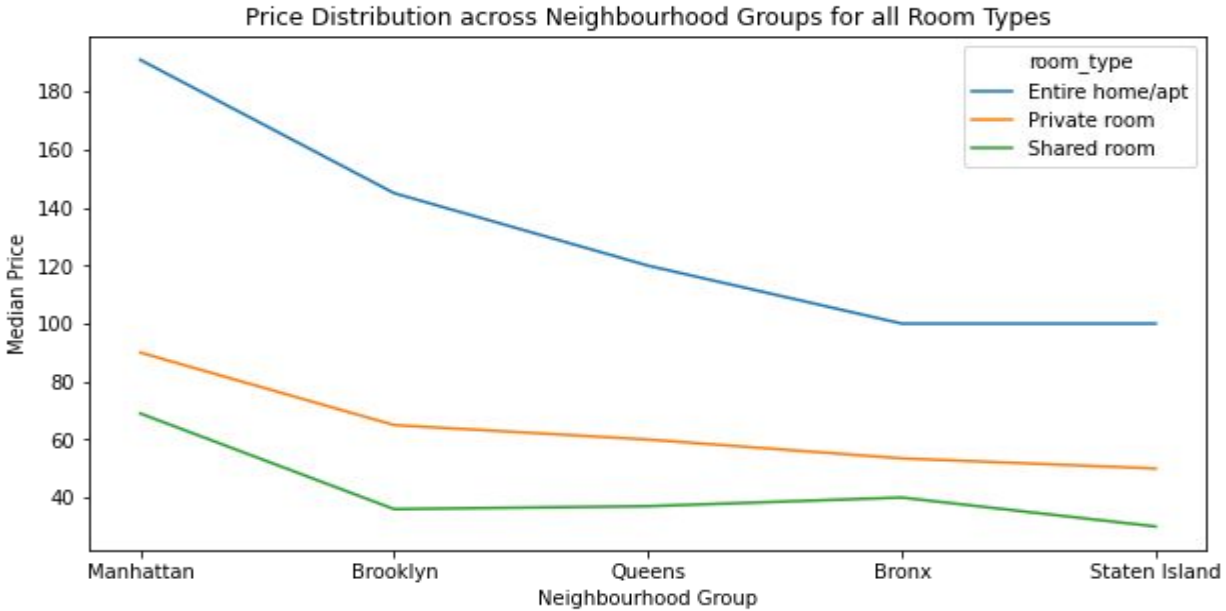
# Neighbourhood groups

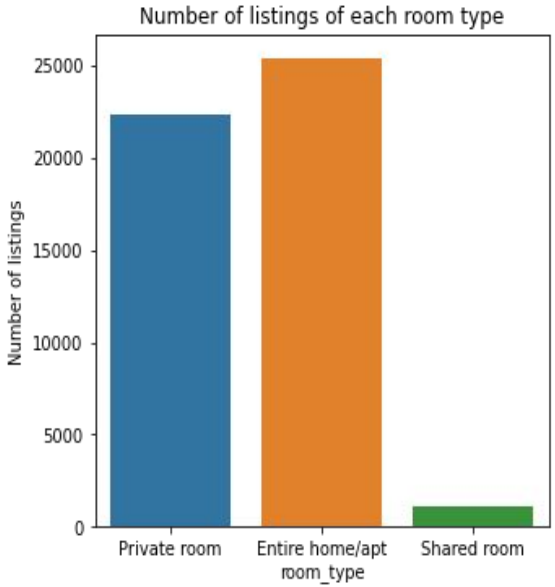



Manhattan has the highest listings.

# Price Distribution

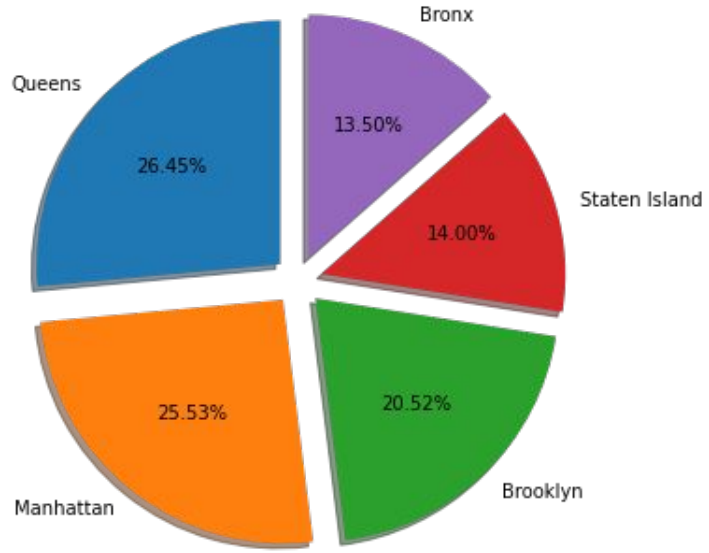


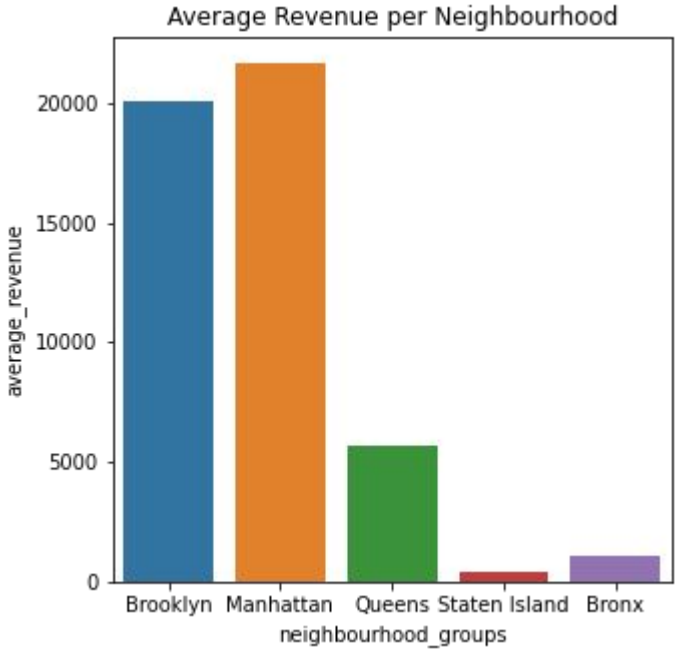

- Manhattan has the highest median price for all room types (Private, Apt, Shared).
- The entire home/apt room type has the highest average price in all neighbourhood groups.
- Shared rooms are not preferred by many

# Reviews

Number of reviews in each neighbourhood group

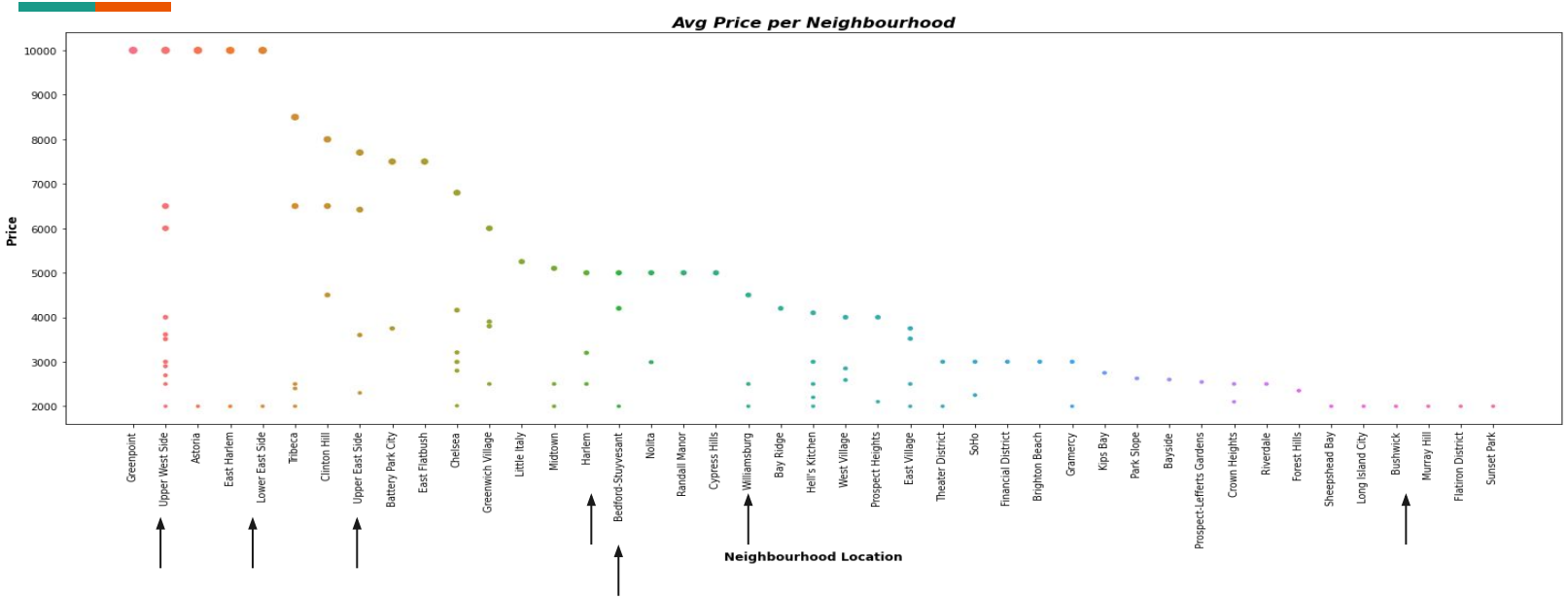



# Average revenue




Queens has the highest reviews

# Distribution of price among neighbourhoods

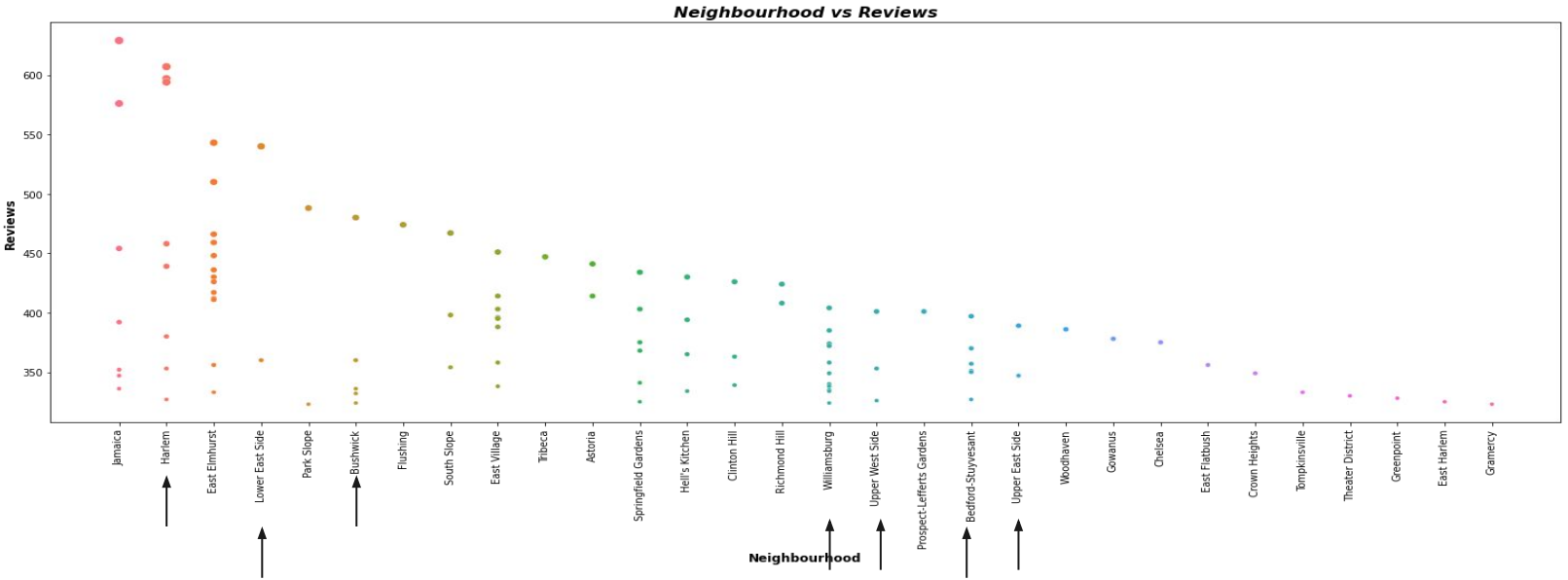



Most neighbourhoods with high reviews are of high average price like-

(Harlem, Bushwick, East Village, Williamsburg, Upper East Side, and Upper West Side )

This implies most highly-priced neighbourhoods are highly reviewed.

# Distribution of reviews among neighbourhoods

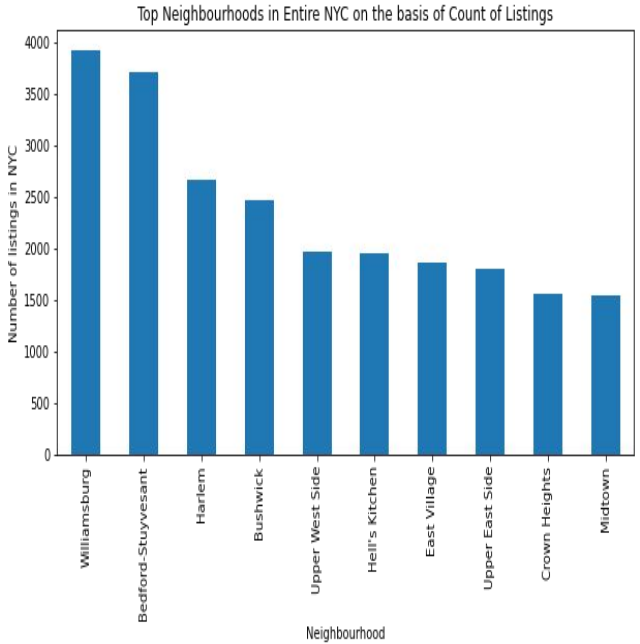



Most neighbourhoods with high reviews are of high average price like-

(Harlem, Bushwick, East Village, Williamsburg, Upper East Side, and Upper West Side )

This implies most highly-priced neighbourhoods are highly reviewed.

# Busy Host Analysis




| | host_name | host_id | neighbourhood_group | room_type | number_of_reviews |
|---|---|---|---|---|---|
| **10310** | Dona | 47621202 | Queens | Private room | 629 |
| **17755** | Jj | 4734398 | Manhattan | Private room | 607 |
| **25626** | Maya | 37312959 | Queens | Private room | 543 |
| **6259** | Carol | 2369681 | Manhattan | Private room | 540 |
| **8973** | Danielle | 26432133 | Queens | Private room | 510 |
| **3966** | Asa | 12949460 | Brooklyn | Entire home/apt | 488 |
| **37848** | Wanda | 792159 | Brooklyn | Private room | 480 |
| **22556** | Linda | 2680820 | Queens | Private room | 474 |
| **8651** | Dani | 42273 | Brooklyn | Entire home/apt | 467 |
| **2953** | Angela | 23591164 | Queens | Private room | 466 |

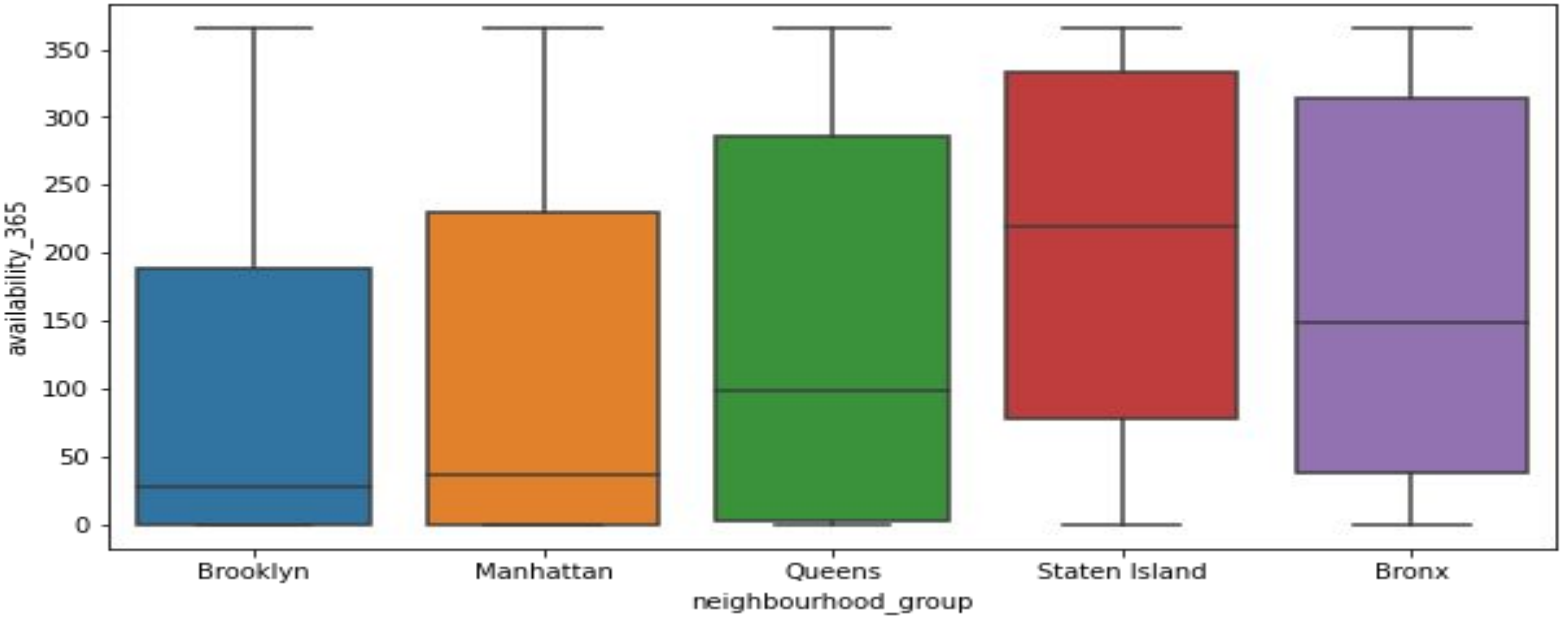
# Observation

Busy  Host Analysis

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

- Neighbourhood groups Queens and Manhattan have the maximum number of reviews {assuming the reviews to be positive} as seen in the pie chart.  Hence busiest hosts belong to these.
- Most listings are of type private room or entire home/apt. Again private rooms are often chosen due to their affordable prices.
- Dona from Jamaica, Queens is the busiest host according to the number of reviews. (assuming the reviews to be positive)

# Availability of Rooms

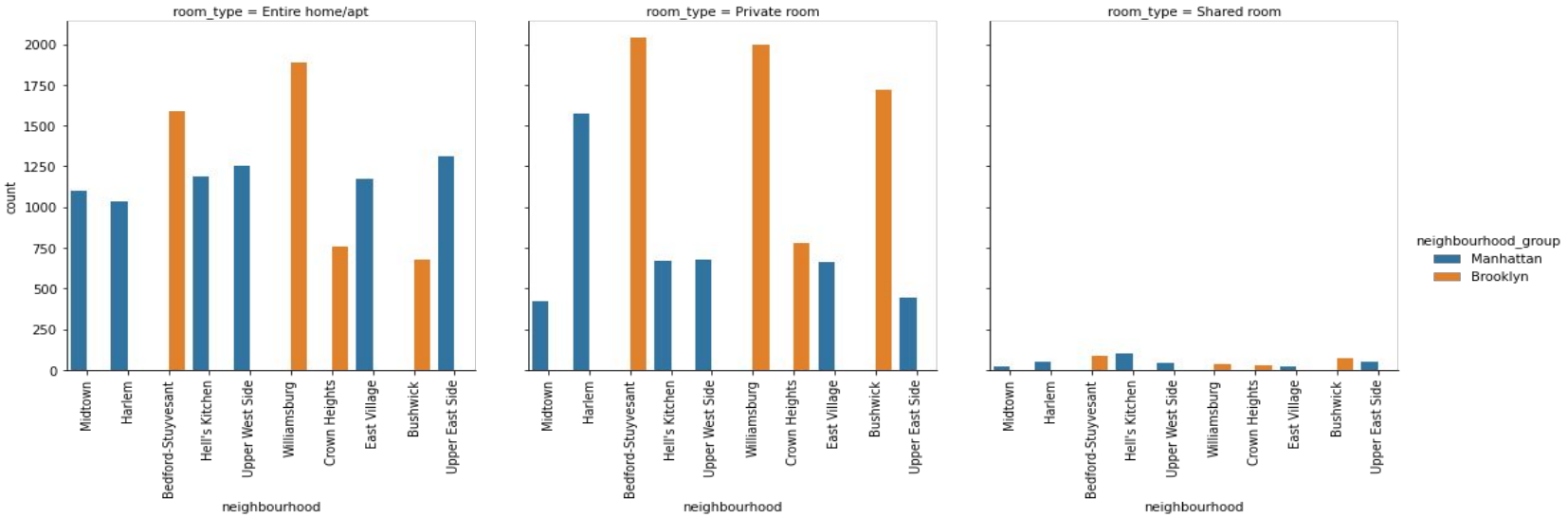



The mean availability of Brooklyn shows that it Is the least available and hence, the busiest

# Traffic in different areas

| | neighbourhood | neighbourhood_group | calculated_host_listings_count |
|---|---|---|---|
| **214** | Williamsburg | Brooklyn | 3920 |
| **13** | Bedford-Stuyvesant | Brooklyn | 3714 |
| **94** | Harlem | Manhattan | 2658 |
| **28** | Bushwick | Brooklyn | 2465 |
| **202** | Upper West Side | Manhattan | 1971 |
| **95** | Hell's Kitchen | Manhattan | 1958 |
| **64** | East Village | Manhattan | 1853 |
| **201** | Upper East Side | Manhattan | 1798 |
| **51** | Crown Heights | Brooklyn | 1564 |
| **127** | Midtown | Manhattan | 1545 |

# Listings In Top Neighbourhoods




For top 10 neighbourhoods Manhattan and Brooklyn are the busiest destinations with most listings.  'Shared room' type listing is barely available among the 10 most listing-populated neighbourhoods.

# Observation

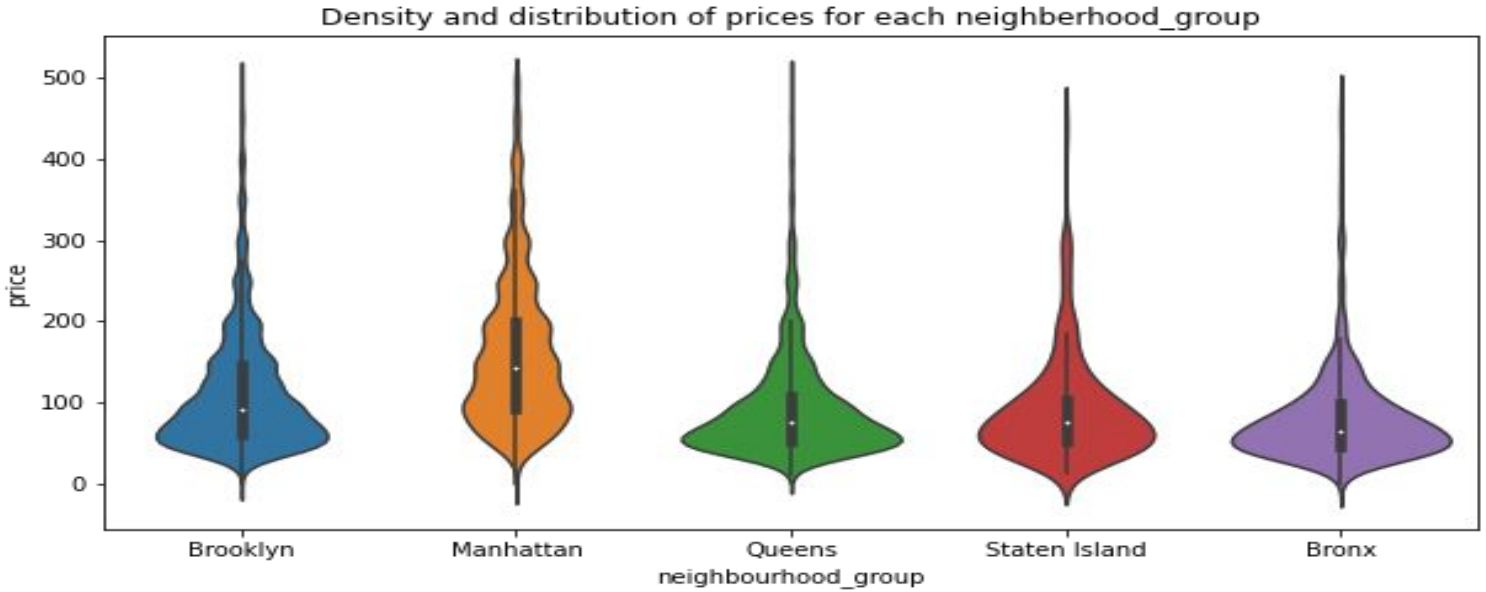
Analysis of noticeable traffic among different areas

- Most neighbourhoods with high listings are present in Manhattan making it a busy place.
- The private rooms are mostly of average price, which makes them affordable to a wide range of people making  them prone to traffic.
- The neighbourhoods with the most listings like- Bedford-Stuyvesant, Williamsburg in Brooklyn, and Harlem in Manhattan, have the most number of private rooms explaining the traffic there.
- Though the shared rooms have the least average price, due to lesser listings of this type and its least preference, the traffic cannot be expected.
- Brooklyn and Manhattan having the least availability are expected to be prone to traffic.

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

We can conclude that Manhattan is most prone to traffic.

# Price Density Distribution




- Brooklyn has an avg price of 80$,.
- Manhattan has the highest range of price with avg price around 150$.
- The Bronx, Staten Island and Queens have prices concentrated towards the median price.

# Conclusion

The exploratory data analysis for Airbnb dataset has been successfully done and the following inferences have been made from the obtained visualizations and also from the dataset.

- Data cleaning, data preparation is done  and correlation of features is checked..
- The problem objectives is met and the key observations are discussed and elaborated with help of multivariate analysis.
- The analysis would have been more systematic if the reviews were scaled according to the emotion specified in it (like(0-5)). However, throughout this analysis, we have assumed reviews to be positive only.
- The analysis would definitely help in making better business decisions.


# Thank you