# Airbnb Exploratory Data Analysis Report


## Name - Kartik Dhiman

[Connect with me on Linkedin]


**DATASET -** Dataset: The dataset contains information on Airbnb listings in New York City. You can download the dataset from the following link: Click Here

**INTRODUCTION -** The purpose of this report is to explore the Airbnb data for New York City and provide insights into the rental prices, room types, neighborhood groups, and customer satisfaction. The analysis is based on data from 48,895 listings in New York City, spanning various neighborhoods, room types, and price ranges.

```
df.head()
```
✓ 0.4s

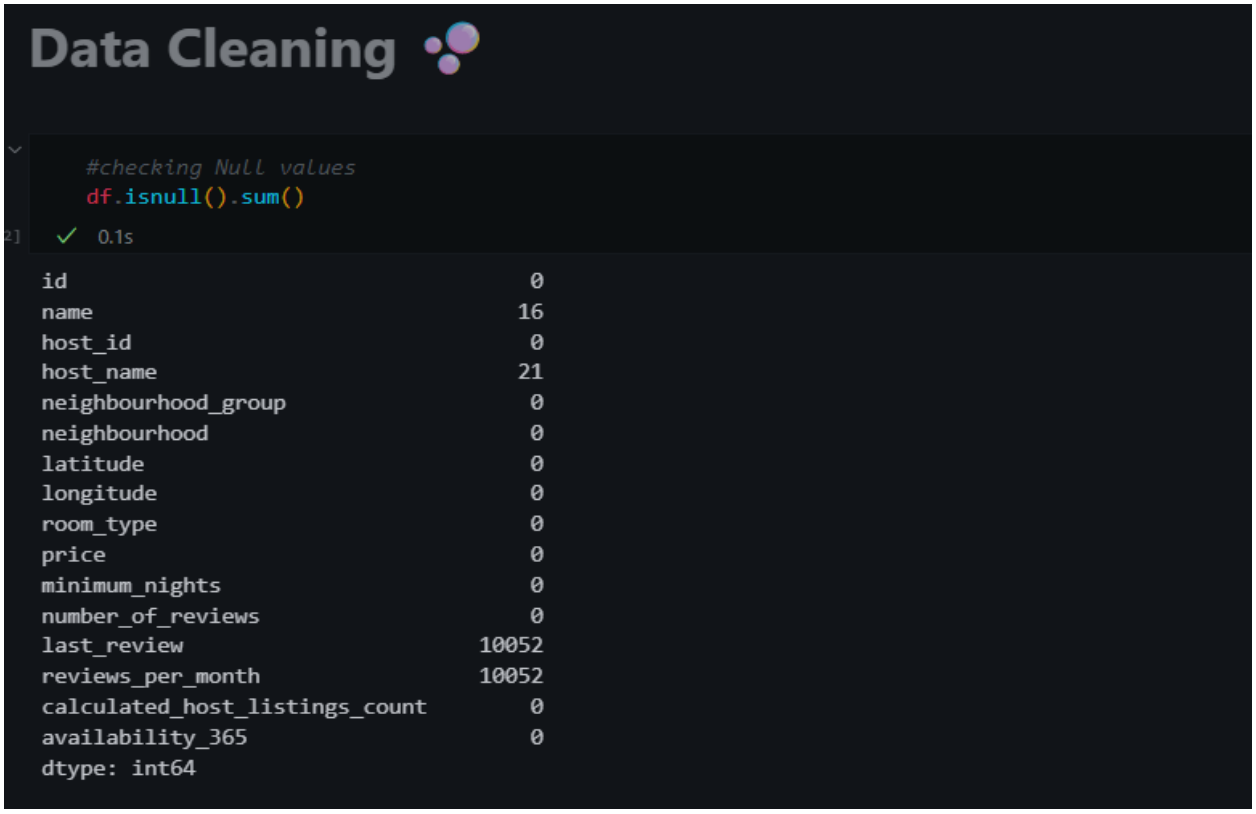
| | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_nights | number_of_reviews | reviews_per_month | calculated_host_listings_count | availability_365 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 | 1 | 9 | 0.21 | 6 | 365 |
| 1 | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 | 1 | 45 | 0.38 | 2 | 355 |
| 2 | Manhattan | Harlem | 40.80902 | -73.94190 | Private room | 150 | 3 | 0 | 0.00 | 1 | 365 |
| 3 | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | Entire home/apt | 89 | 1 | 270 | 4.64 | 1 | 194 |
| 4 | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | 80 | 10 | 9 | 0.10 | 1 | 0 |

**Descriptive Statistics -** Descriptive statistics were calculated for numerical variables such as price, minimum nights, number of reviews, reviews per month, calculated host listings count, and availability. The mean price is $152.72, with a standard deviation of $240.15. The minimum price is $0, and the maximum price is $10,000. The mean minimum nights are 7.03, and the mean number of reviews is 23.27.

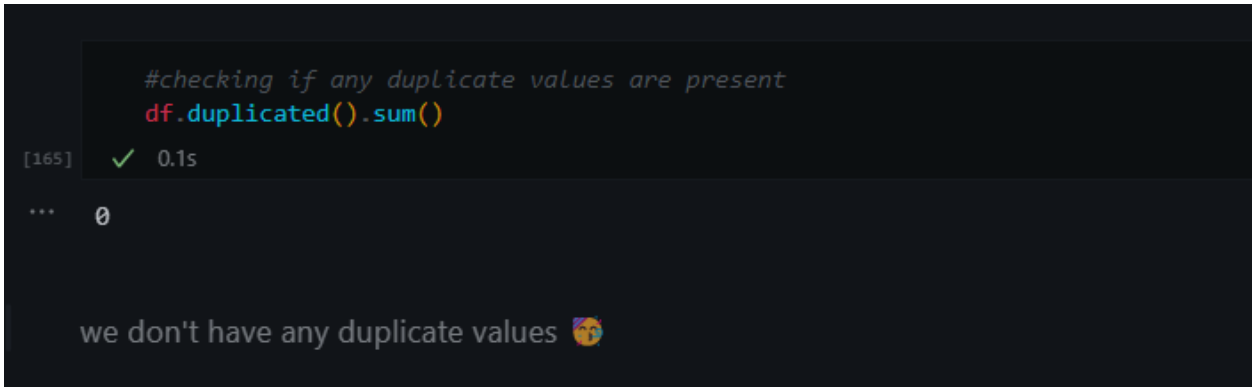
```
df.describe()
```
✓ 0.2s

| | id | host_id | latitude | longitude | price | minimum_nights | number_of_reviews | reviews_per_month | calculated_host_listings_count | availability_365 |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 4.889500e+04 | 4.889500e+04 | 48895.000000 | 48895.000000 | 48895.000000 | 48895.000000 | 48895.000000 | 38843.000000 | 48895.000000 | 48895.000000 |
| mean | 1.901714e+07 | 6.762001e+07 | 40.728949 | -73.952170 | 152.720687 | 7.029962 | 23.274466 | 1.373221 | 7.143982 | 112.781327 |
| std | 1.098311e+07 | 7.861097e+07 | 0.054530 | 0.046157 | 240.154170 | 20.510550 | 44.550582 | 1.680442 | 32.952519 | 131.622289 |
| min | 2.539000e+03 | 2.438000e+03 | 40.499790 | -74.244420 | 0.000000 | 1.000000 | 0.000000 | 0.010000 | 1.000000 | 0.000000 |
| 25% | 9.471945e+06 | 7.822033e+06 | 40.690100 | -73.983070 | 69.000000 | 1.000000 | 1.000000 | 0.190000 | 1.000000 | 0.000000 |
| 50% | 1.967728e+07 | 3.079382e+07 | 40.723070 | -73.955680 | 106.000000 | 3.000000 | 5.000000 | 0.720000 | 1.000000 | 45.000000 |
| 75% | 2.915218e+07 | 1.074344e+08 | 40.763115 | -73.936275 | 175.000000 | 5.000000 | 24.000000 | 2.020000 | 2.000000 | 227.000000 |
| max | 3.648724e+07 | 2.743213e+08 | 40.913060 | -73.712990 | 10000.000000 | 1250.000000 | 629.000000 | 58.500000 | 327.000000 | 365.000000 |

**Data Cleaning -** We have 10052 null values in the last_review and reviews_per_month columns and a very few null values in name and host_name.




**Duplicate Values -** We don't have any duplicate value in our dataset.



**Removing Unwanted Columns -** We have to remove some unwanted columns that we don't need in our analysis.

```
# We are removing the unwanted columns id,name , host_name and last_review doesn't help us in anyway in our approach.
df.drop(['name','host_name','host_id','id','last_review'], axis=1, inplace=True)
df.head()
```

✓ 0.0s

| | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_nights | number_of_reviews | reviews_per_month | calculated_host_listings_count | availability_365 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 | 1 | 9 | 0.21 | 6 | 365 |
| 1 | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 | 1 | 45 | 0.38 | 2 | 355 |
| 2 | Manhattan | Harlem | 40.80902 | -73.94190 | Private room | 150 | 3 | 0 | NaN | 1 | 365 |
| 3 | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | Entire home/apt | 89 | 1 | 270 | 4.64 | 1 | 194 |
| 4 | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | 80 | 10 | 9 | 0.10 | 1 | 0 |

We just have null values in review_per_month So, We can actually fill all the null values of reviews_per_month with 0 where it have null values because the data is null only because no one has reviewed it and hence the number of review is 0 here.

```
# Missing value implies there are no reviews for the location.
df['reviews_per_month'] = df['reviews_per_month'].fillna(0)
```

✓ 0.0s

```
df.isna().sum()
```

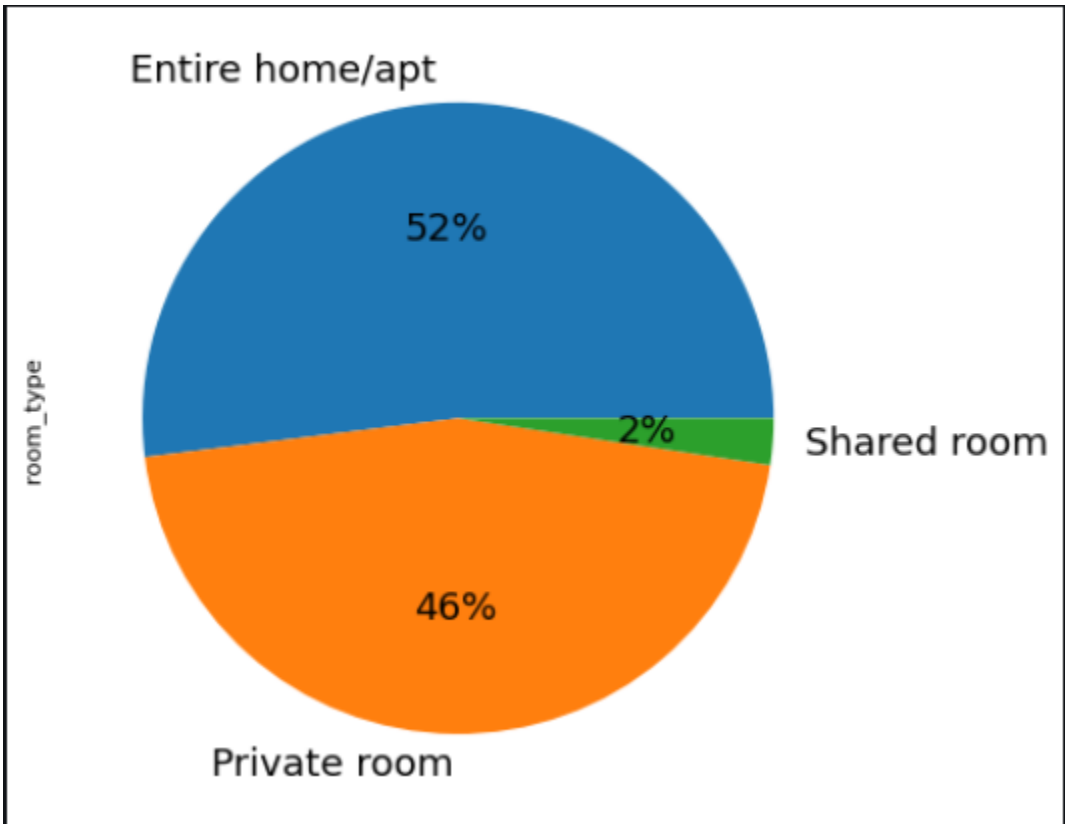
✓ 0.1s

```
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

**The data was cleaned to handle missing values, remove duplicates, and perform necessary data transformations. The latitude and longitude data were used to create a heatmap to visualize the density of listings across New York City neighborhoods.**

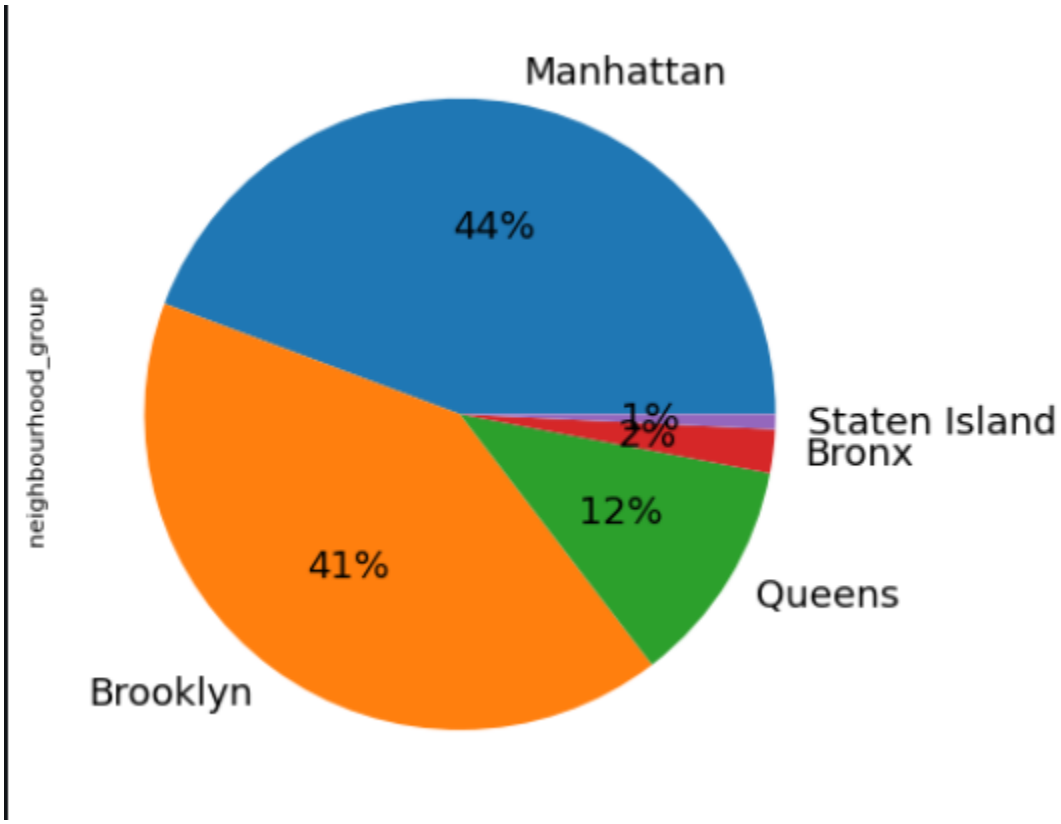
**Data Visualization -** Appropriate visualizations such as histograms, boxplots, and bar charts were created to analyze the distribution of numerical variables and the relationships between categorical and numerical variables.

**Room Type:**



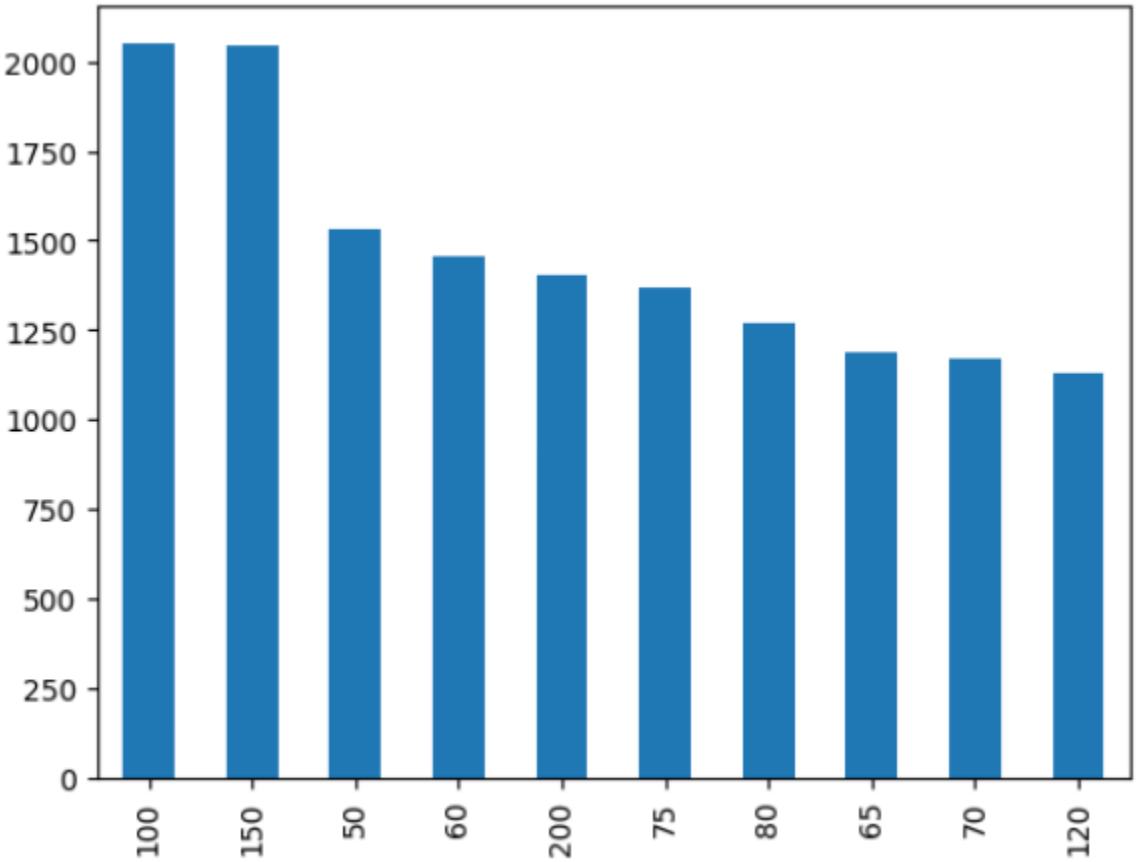
It's interesting to see that in our dataset, more than half of the people (52%) prefer to use a house on Airbnb, while 46% opt for a private room. Only a small fraction, 2%, chose a shared room. This could indicate that most people who use Airbnb are likely traveling with their family or friends for vacations, tours, or visits. Overall, it's helpful to know this information to better understand the needs and preferences of Airbnb users.

**Neighbourhood Groups:**



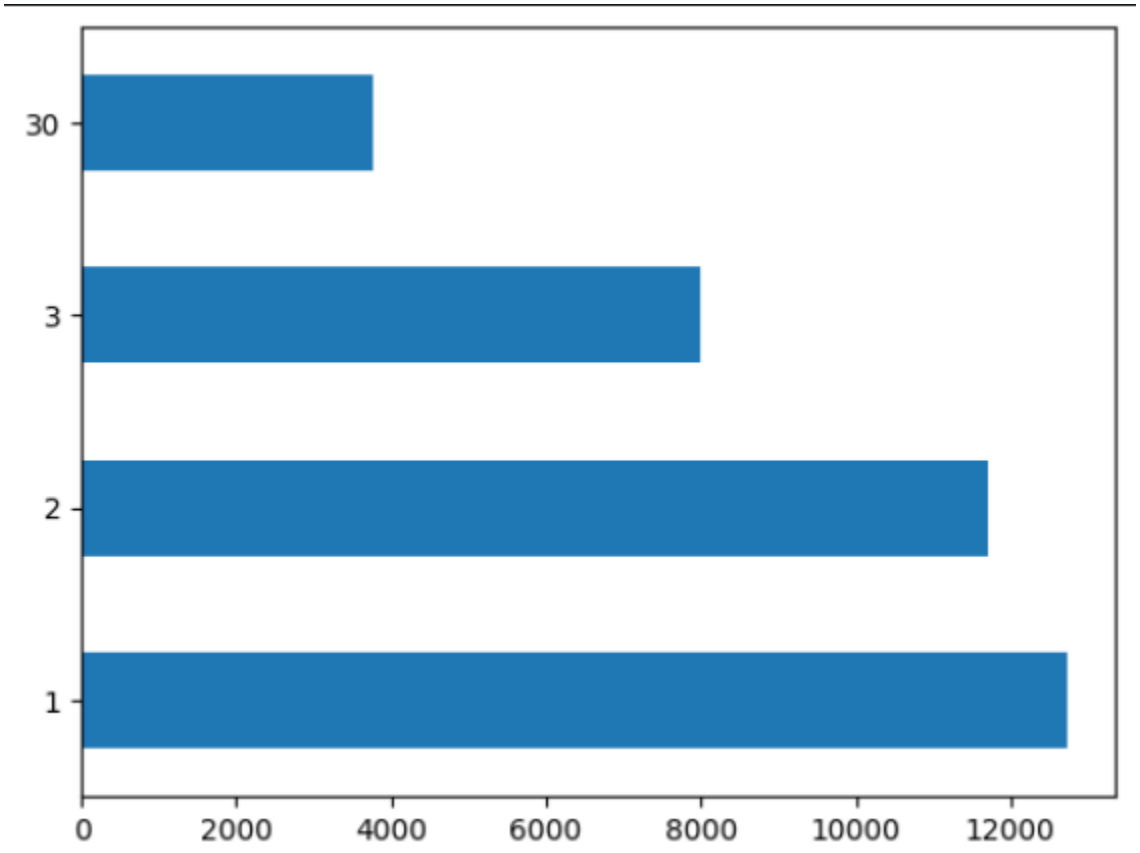

According to the data, Manhattan seems to be the most popular destination for Airbnb rentals, accounting for a whopping 44% of all transactions. Following closely behind is Brooklyn, which represents 41% of the total transactions, followed by Queens with 12%, and the Bronx with only 2%. On the other hand, Staten Island appears to be the least popular destination with only 1% of the total Airbnb transactions.

**Price:**



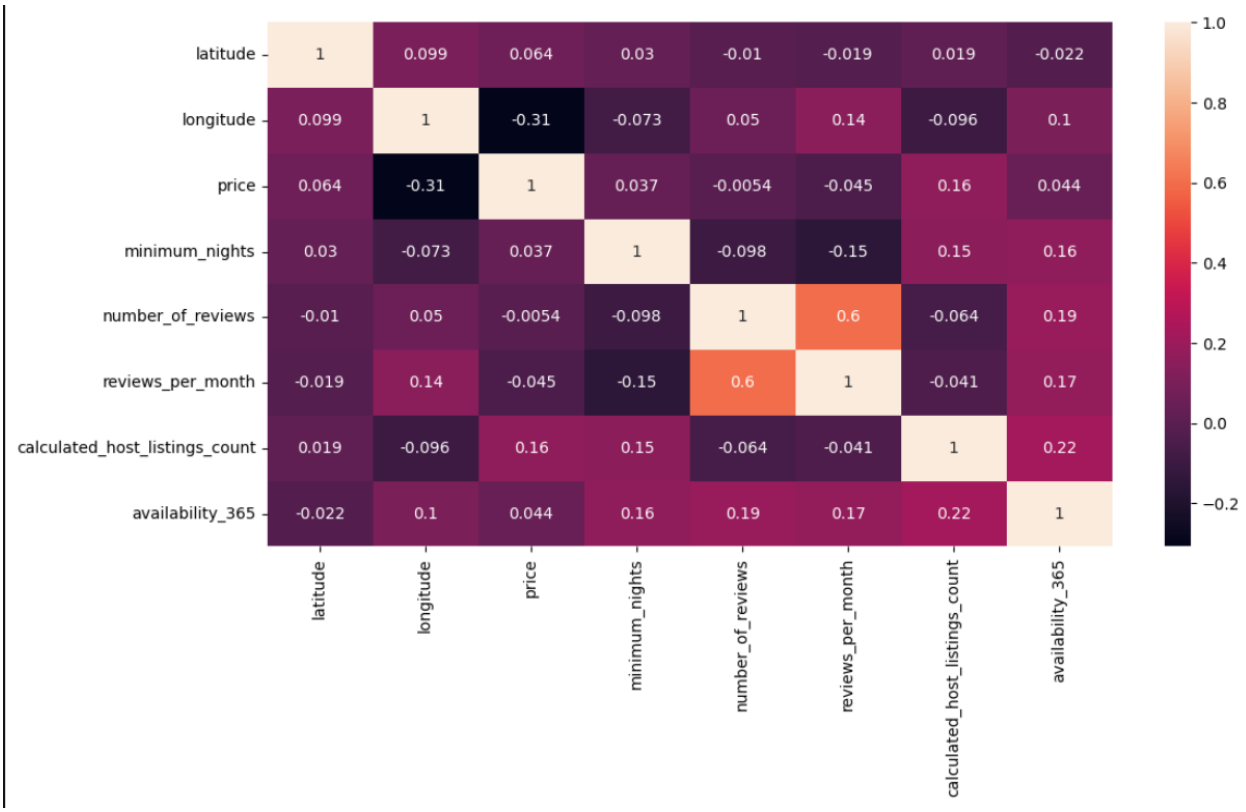

Almost 2 thousand plus airbnb's has a price of 100 dollars and 150 dollars each and 1.5k airbnb's have around 50 dollars price.

**Minimum Nights:**

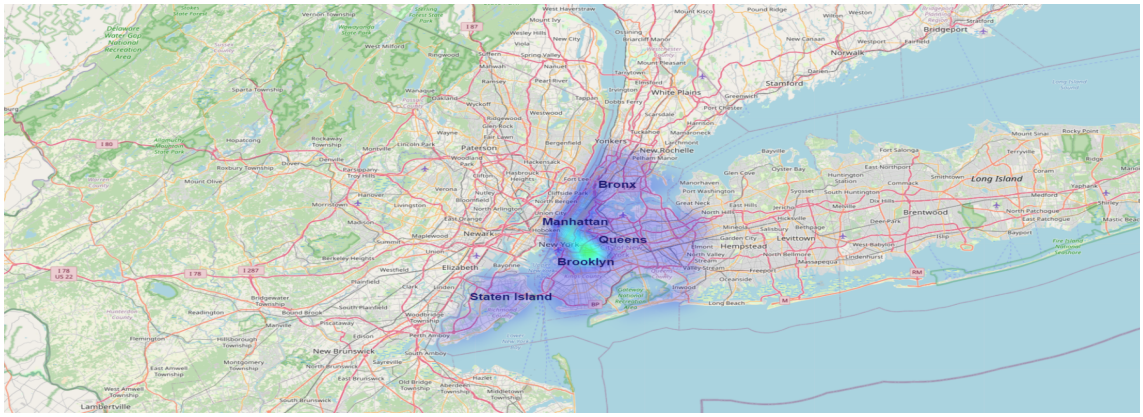



We can observe that most of almost 12k people used 1 night stay in airbnb and 11k people choose 2 night stay while 7k choose 3 night stay or Almost 3.7k stayed atleast a month.

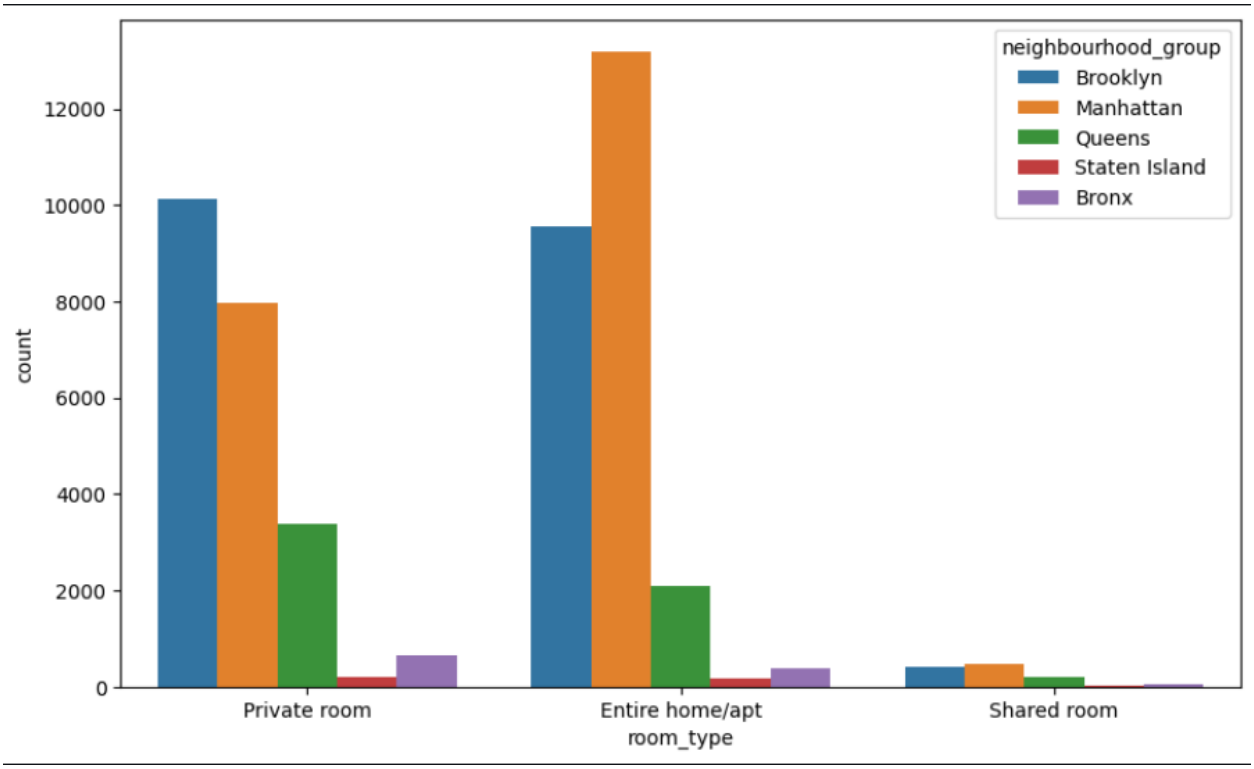
## Correlation Matrices:



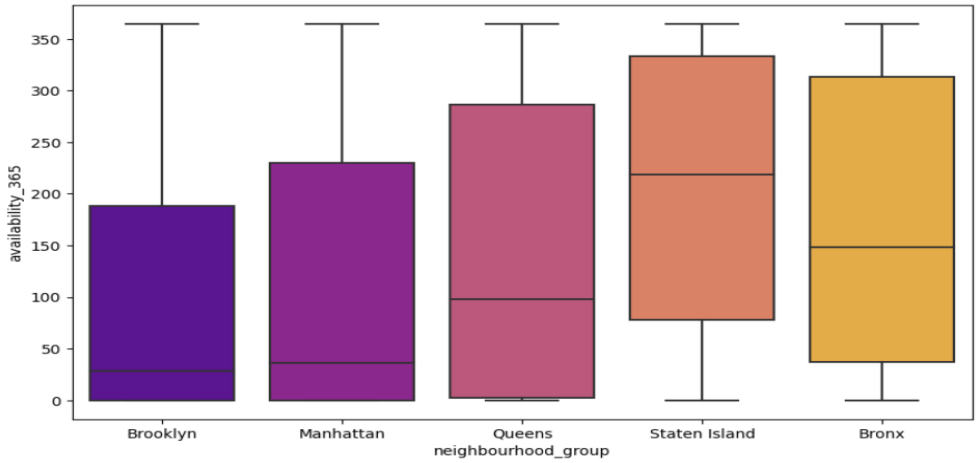

## Airbnb Listing Across All New York:



A heatmap was created to visualize the density of listings across New York City neighborhoods. The highest concentration of Airbnb listings is in Manhattan, followed by Brooklyn.




Based on the data, it seems that the home service is the most commonly used type of rental on Airbnb, especially in Manhattan where it is the highest-used service. It is also the highest-used service across all of New York City. In Brooklyn, on the other hand, private rooms were more commonly used.



Staten Island have the highest average airbnb availability.

**Relationship Between Room Type, Neighborhood Group, and Price** -
Appropriate statistical tests such as ANOVA were performed to determine if there are significant differences in rental prices based on room type and neighborhood group. The results show that the F-value is 318.02, and the p-value is 0.0, indicating a significant difference in rental prices based on room type and neighborhood group.

**Relationship Between Customer Satisfaction and Factors -**
The relationship between customer satisfaction (as measured by the number of reviews and reviews per month) and factors such as price, room type, and neighborhood group was analyzed. The results show that there is a positive correlation between the number of reviews and reviews per month, indicating higher customer satisfaction. The analysis also shows that the price, room type, and neighborhood group can significantly impact customer satisfaction.

**Recommendations -**
Based on the findings, it is recommended that Airbnb hosts target specific neighborhoods with high demand and adjust pricing based on competition. Offering different room types can also improve customer satisfaction and increase rental prices. Additionally, hosts should ensure that their listings are clean, well-maintained, and provide a positive experience for guests.

**Conclusion -**
The analysis provides valuable insights into the Airbnb market in New York City, highlighting the importance of pricing, room types, and neighborhood groups in determining rental prices and customer satisfaction. The findings can help hosts make informed decisions to improve their listings and attract more guests.