Big Apple Stays

A data-driven analysis of Airbnb listings in NYC

Intent

The motivation for this project stems from New York City's reputation as the busiest place in the world, and the desire to personally investigate how tourists utilize Airbnb accommodations.

Goal

This project aims to analyze a dataset related to Airbnb in New York City to understand the booking trends of tourists and gain insights into the preferences of visitors to different types of properties. Through data analysis, this project seeks to establish relationships between the various attributes and draw insights that could inform future improvements which could be valuable for stakeholders in tourism industry.

Duration

2 months

Objective

The objective of this study is to analyze the dataset and investigate which Airbnb properties have the highest demand, the preferred room types, and the availability of Airbnb properties throughout the year in different neighborhoods. The research questions guiding this study are:

1. **What is the distribution of Airbnb properties among different neighborhood groups?**
2. **Which type of room is more popular among Airbnb guests?**
3. **How does the price of Airbnb properties vary across different neighborhoods?**
4. **How does the availability of Airbnb properties fluctuate throughout the year?**

Airbnb as a potential competitor

As a platform for short-term rentals, Airbnb offers travelers unique and often cheaper accommodation options that hotels cannot match. This shift in consumer behavior has led to a decrease in hotel occupancy rates and a need for the hotel industry to adapt to the changing market. The impact of Airbnb on the hotel industry is a significant trend to explore, and this topic remains an area of ongoing research and analysis.

Performance of Airbnb listings in New York

Apart from price, vacationers take into account several other factors when selecting accommodations. Research by Guttentag suggests that location and amenities are practical considerations that attract customers to Airbnb. Additionally, the influence of online reviews has become increasingly important. In 2019, Trip Advisor conducted an independent study with Ipsos MORI, which surveyed over 23,000 Trip Advisor users across 12 markets. The study found that 81% of the participating travelers always or frequently read reviews before booking a place to stay. These findings highlight the significance of online reviews in the decision-making process of vacationers when selecting their accommodations.

Data Pre-processing

To get accurate results from our data before we apply any operations, I pre-processed the data. By pre-processing the data, we can address issues such as inconsistencies, outliers, and errors that can impact the accuracy and reliability of our results. This can improve the performance of our models and analyses and help ensure that we are drawing valid conclusions from our data.

The chosen dataset contains many columns. I used the .shape attribute to understand the size of the data, and check the inconsistencies.

To ensure that the data types of each column are accurately assigned, I used .dtypes attribute.

```
kr.shape
(38277, 18)

kr.dtypes
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
number_of_reviews_ltm               int64
license                            object
dtype: object
```

Checking for Null Values

```
[ ] kr.isnull().sum()

id                                    0
name                                 13
host_id                               0
host_name                            34
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                        9504
reviews_per_month                  9504
calculated_host_listings_count        0
availability_365                      0
number_of_reviews_ltm                 0
license                           38276
dtype: int64
```

After pre-processing the data, the next step I took was to check for null values in the entire dataset using .isnull() .

Upon examination, I discovered that the last_review and number_of_reviews columns had 9504 null values. Additionally, I observed a column named license which appeared to have been accidentally included in the dataset.

Data Cleaning

After identifying the null values in the dataset, the subsequent step I took was to remove any columns that were irrelevant to my research questions. By doing so, I aimed to simplify the dataset and focus solely on the columns that would be helpful for my analysis.

```
kr.drop('license',axis=1,inplace=True)
kr.drop('last_review',axis=1,inplace=True
kr.isnull().sum()

id                                    0
name                                 13
host_id                               0
host_name                            34
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
reviews_per_month                  9504
calculated_host_listings_count        0
availability_365                      0
number_of_reviews_ltm                 0
dtype: int64
```

```
data=kr.dropna()

[ ] data.isnull().sum()

id                                0
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
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
number_of_reviews_ltm             0
dtype: int64
```

After removing the irrelevant columns from the dataset, I removed any null values from the reviews_per_month, host_name, and name columns. By doing so, I aimed to ensure the dataset only contained complete and accurate data for the remaining relevant columns.

Following the data cleaning process, we can observe that the dataset now has 28747 rows and 16 columns. This means that we have removed any null values and irrelevant columns, resulting in a streamlined dataset that contains complete and accurate data for our analysis.

```
[ ] data.shape

(28747, 16)
```

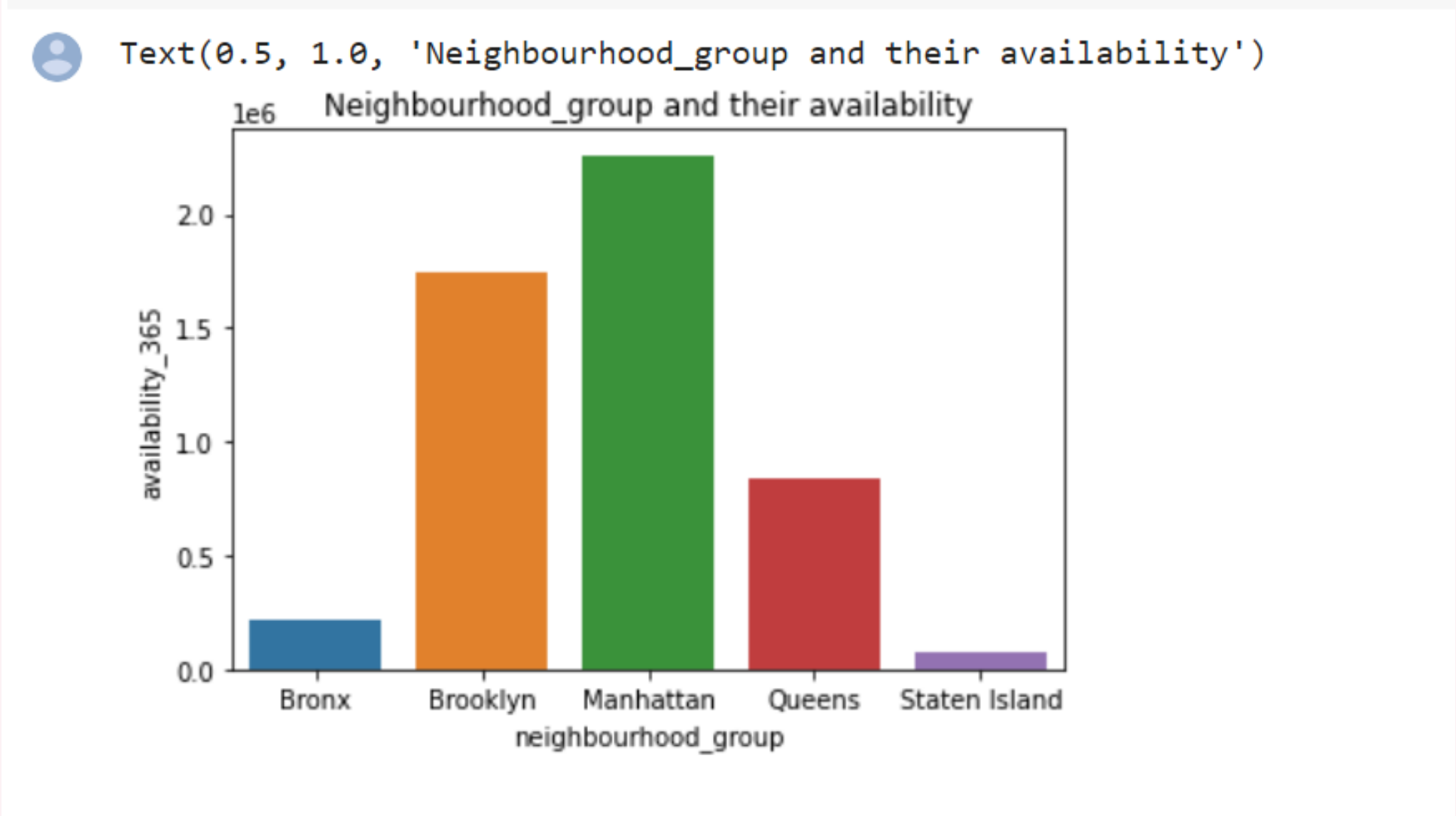
Results

1. **What is the distribution of Airbnb properties among different neighborhood groups?**

Based on the figure presented, it is apparent that Manhattan in New York has the highest Airbnb availability, which is represented by the color green. In contrast, Staten Island has the least Airbnb availability, and is represented by the color pink.

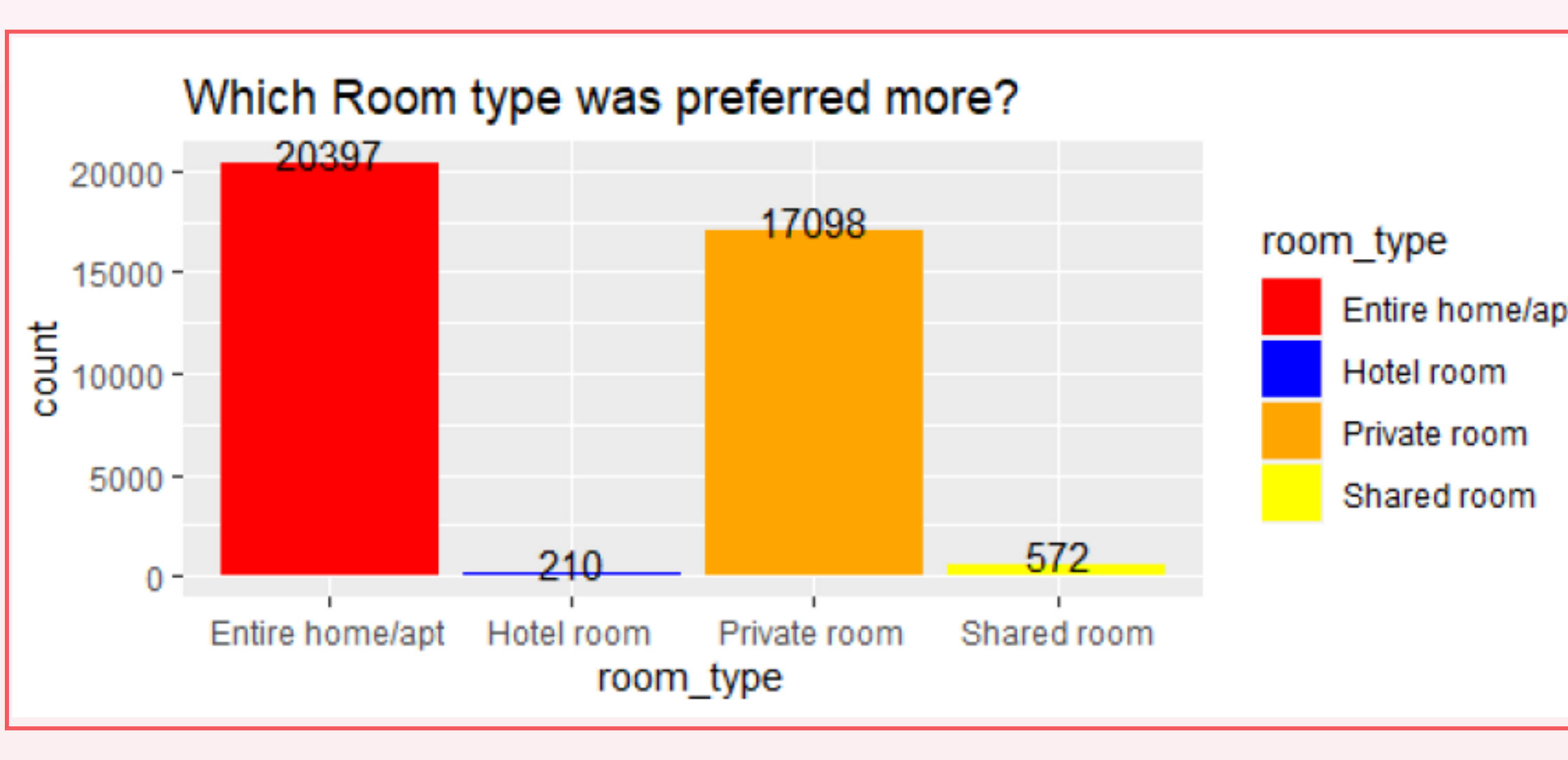
```
k=sns.barplot(x='neighbourhood_group',y='availability_365',data=kr1)
k.set_title('Neighbourhood_group and their availability')

Text(0.5, 1.0, 'Neighbourhood_group and their availability')
```



2. **Which type of room is more popular among Airbnb guests?**

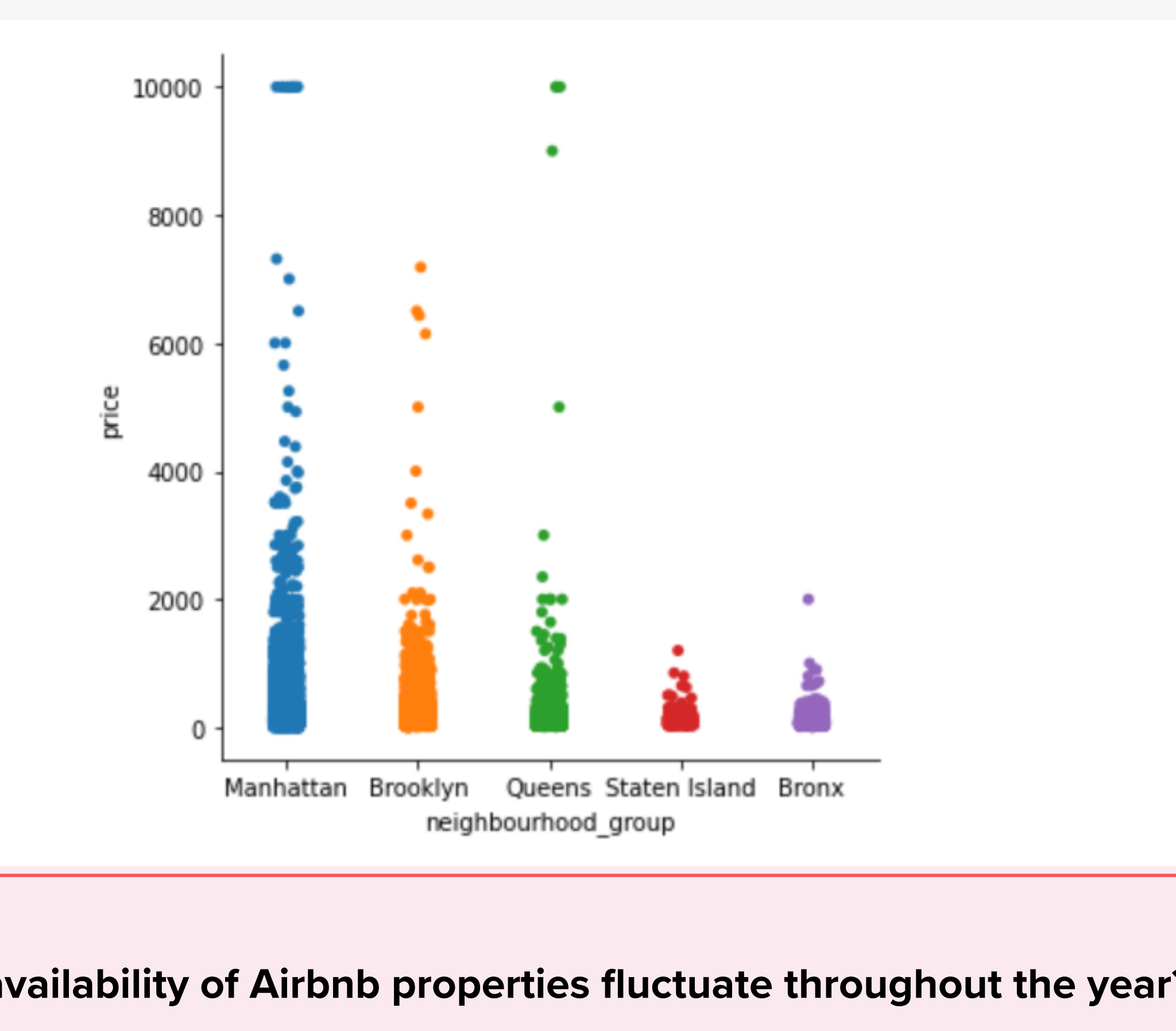
After visualizing the data using RStudio, it is evident that the preferred type of Airbnb rental was Entire home/apt, with a value of 20397. On the other hand, the least preferred type was Hotel room, with a value of 210. To improve the clarity and aesthetics of the visualization, I used the scale_fill_manual function to change the default colors.



3. **How does the price of Airbnb properties vary across different neighborhoods?**

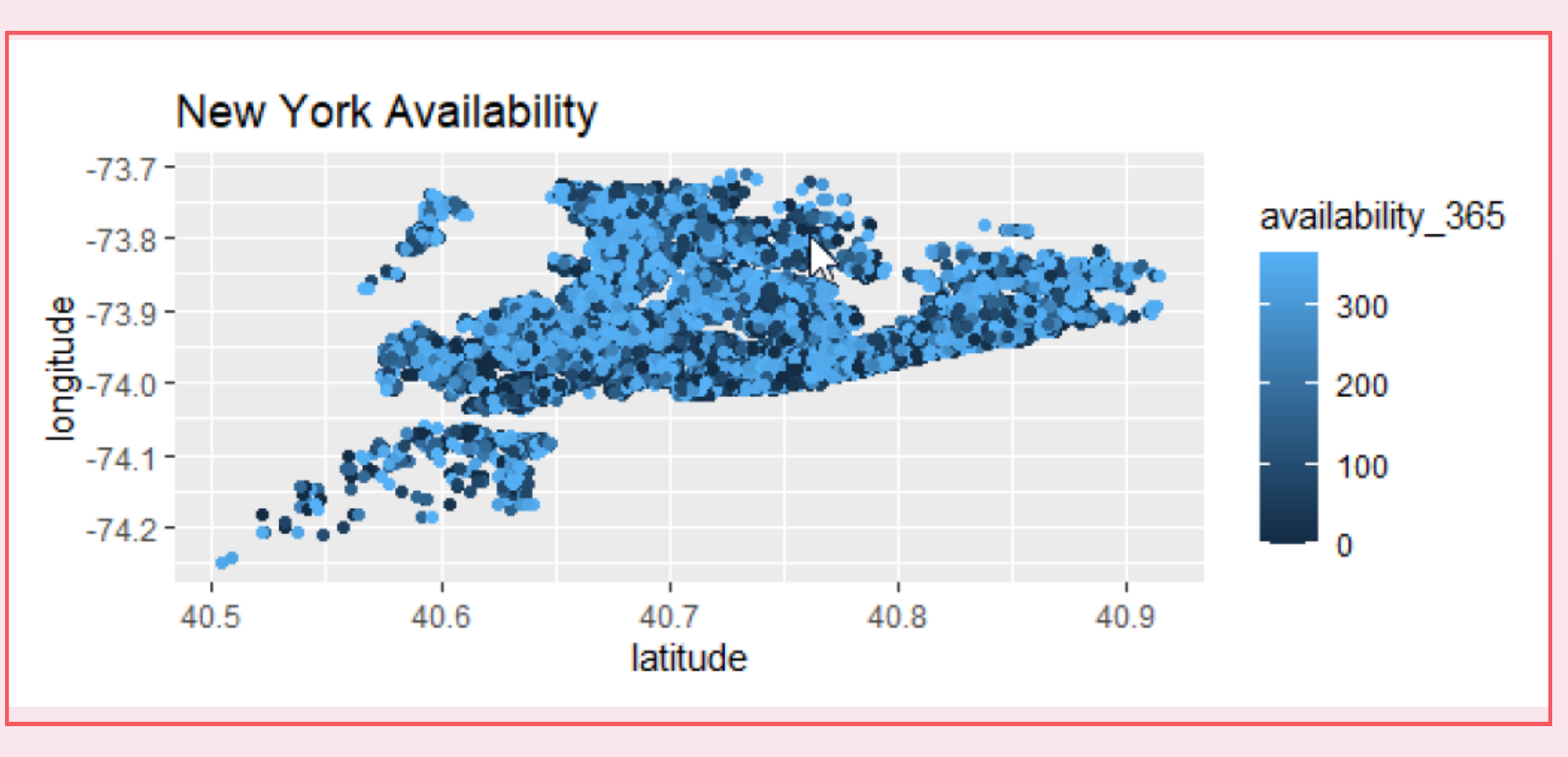
This plot was generated using Python to illustrate the variation in Airbnb prices across different areas. I utilized the seaborn library to create the plot. Upon examining the plot, we can observe that in Manhattan, Airbnb prices are scattered across every price range, implying that this area is affordable for everyone. In Brooklyn, the Airbnb prices can go up to 8000, indicating that it is more expensive compared to Manhattan. In contrast, Staten Island is the most affordable area, with Airbnb prices below 2000.

```
[ ] g = sns.catplot(x="neighbourhood_group", y="price", data=kr)
```



4. **How does the availability of Airbnb properties fluctuate throughout the year?**

The dataset includes columns for latitude and longitude, which I utilized to create a plot that displays the availability of Airbnb across different locations in New York. To represent the varying levels of availability, I utilized a color range that goes from dark blue to light blue. This enables users to easily identify areas with higher availability of Airbnb, and make informed decisions accordingly.



Conclusion

In conclusion, the airbnb dataset obtained from Inside AirBnb's NYC page is a comprehensive dataset with numerous columns that enabled us to perform an in-depth analysis on each significant column. We started by identifying our objectives and formulating research questions that could be answered through data analysis. Furthermore, we conducted a brief analysis on the competition and performance of Airbnb by reviewing relevant literature.

Before commencing with data exploration, we pre-processed the data to eliminate any inconsistencies and issues. Subsequently, we checked for null values and eliminated unwanted columns to clean the data. Finally, we utilized graphs in Python and R to visualize the data.

Through this analysis, we discovered several compelling relationships between features and elaborated on each step of the process. This data analysis process is emulated at a higher level in Airbnb's Data team to facilitate better business decisions, improve platform control, drive marketing initiatives, implement new features, and more.