ABNB Data Analysis

Mahbuba Siddiqua Jyoti

2023-12-02

Introduction

- **Main Goal**
  - ABNB, as a current investment target, is attracting a lot of attention. Many individuals are seek- ing a stable income independent of their regular jobs through ABNB. What factors are commonly focused on by investors and affect the earnings of ABNB? Common factors include room rate, occupancy rate, location, room type, customer reviews, minimum stay etc., with the room rate often regarded as a primary consideration.
  - In this project, we explore a detailed data analysis of Airbnb listings in New York City, leveraging various statistical and visualization techniques. The analysis encompasses data preprocessing, exploratory data analysis (EDA), and the construction of a linear regression model for predicting listing prices. In this narrative, we will delve into the key components of the analysis, elucidating the rationale behind each step and interpreting the findings.
- **Data Dictionary**
  - https://docs.google.com/spreadsheets/d/1iWCNJcSutYqpULSQHlNyGInUvHg2BoUGoNRIGa6Szc4/edit#gid=1938308660

Step 1: Exploratory Data Analysis (EDA)

Data Overview

The analysis begins with a comprehensive overview of the dataset, detailing fundamental statistics such as sample size, the number of variables, and the types of variables. This information provides a foundational understanding of the dataset's structure and informs subsequent analytical decisions.

```
ExpData(data=data,type=1)
```

##	Descriptions	Value
## 1	Sample size (nrow)	38792
## 2	No. of variables (ncol)	18
## 3	No. of numeric/interger variables	11
## 4	No. of factor variables	0
## 5	No. of text variables	7
## 6	No. of logical variables	0
## 7	No. of identifier variables	1
## 8	No. of date variables	0
## 9	No. of zero variance variables (uniform)	0

```
## 10               %. of variables having complete cases 77.78% (14)
## 11   %. of variables having >0% and <50% missing cases  16.67% (3)
## 12 %. of variables having >=50% and <90% missing cases      0% (0)
## 13          %. of variables having >=90% missing cases   5.56% (1)
```

```
ExpData(data=data,type=2)
```

```
##    Index                  Variable_Name Variable_Type Sample_n Missing_Count
## 1      1                             id       numeric    38792             0
## 2      2                           name     character    38792             0
## 3      3                        host_id       integer    38792             0
## 4      4                      host_name     character    38787             5
## 5      5            neighbourhood_group     character    38792             0
## 6      6                  neighbourhood     character    38792             0
## 7      7                       latitude       numeric    38792             0
## 8      8                      longitude       numeric    38792             0
## 9      9                      room_type     character    38792             0
## 10    10                          price       integer    38792             0
## 11    11                minimum_nights       integer    38792             0
## 12    12              number_of_reviews       integer    38792             0
## 13    13                    last_review     character    28440         10352
## 14    14              reviews_per_month       numeric    28440         10352
## 15    15 calculated_host_listings_count       integer    38792             0
## 16    16               availability_365       integer    38792             0
## 17    17          number_of_reviews_ltm       integer    38792             0
## 18    18                        license     character     2939         35853
##    Per_of_Missing No_of_distinct_values
## 1           0.000                 38792
## 2           0.000                 12050
## 3           0.000                 23811
## 4           0.000                  8829
## 5           0.000                     5
## 6           0.000                   223
## 7           0.000                 23362
## 8           0.000                 21020
## 9           0.000                     4
## 10          0.000                  1184
## 11          0.000                   114
## 12          0.000                   465
## 13          0.267                  2925
## 14          0.267                   822
## 15          0.000                    70
## 16          0.000                   366
## 17          0.000                   144
## 18          0.924                   279
```

The dataset consists of 38,792 rows (observations) and 18 columns (variables) which gives an initial understanding of the data's volume and dimensionality. There are 11 numeric/integer variables, indicating quantitative data, 7 text variables, suggesting qualitative or categorical information, and the absence of factor variables implies that there are no explicitly defined categorical factors in the dataset. Approximately 77.78% of the variables have complete cases, meaning there are no missing values for these variables, about 16.67% of the variables have between 0% and 50% missing cases, and no variables with missing cases exceeding 50%, ensuring a relatively small amount of missing data in the dataset.

```
df <- subset(data, select = c("neighbourhood_group", "room_type", "minimum_nights", "number_of_reviews"
```

Step 2: Geoplot of Data

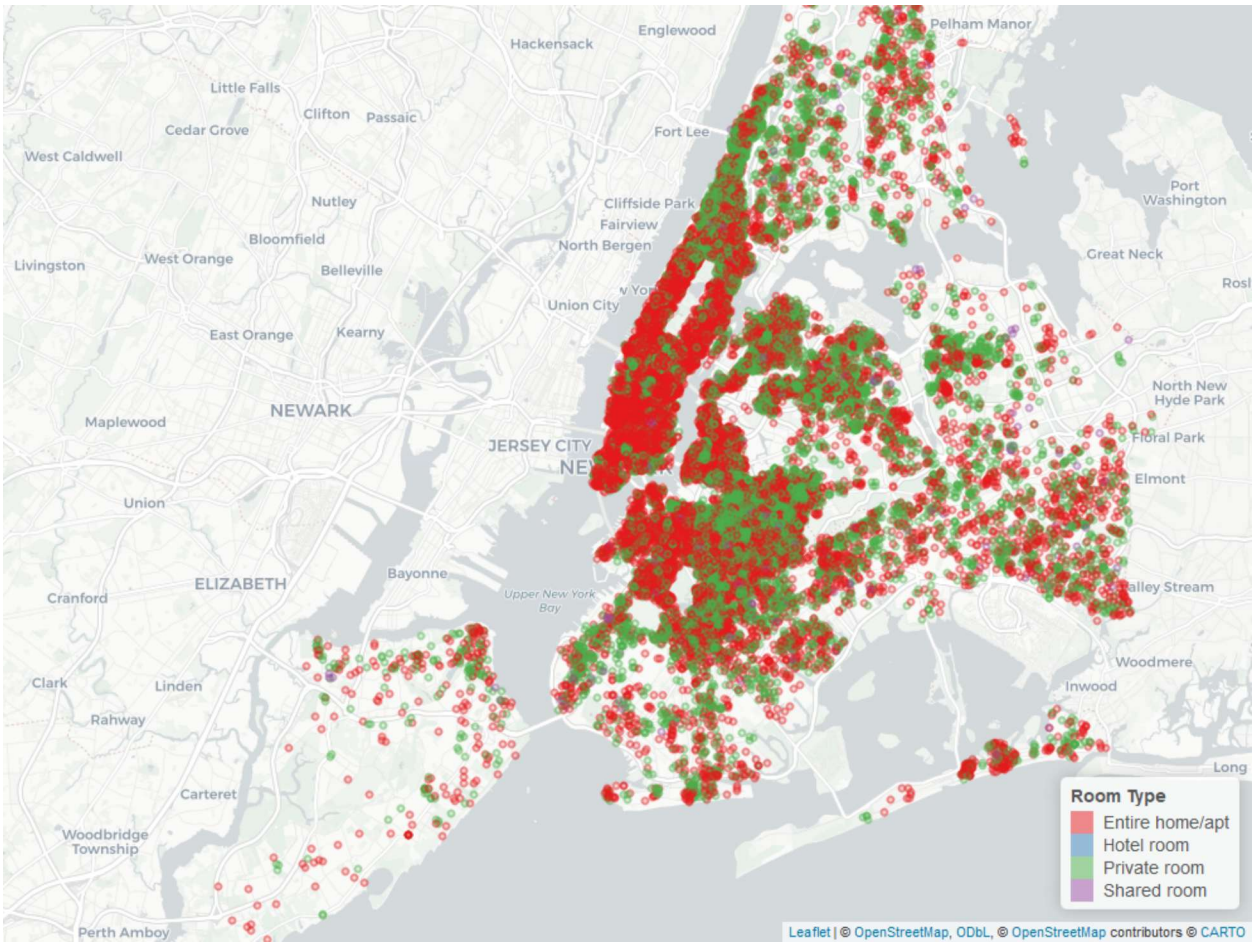
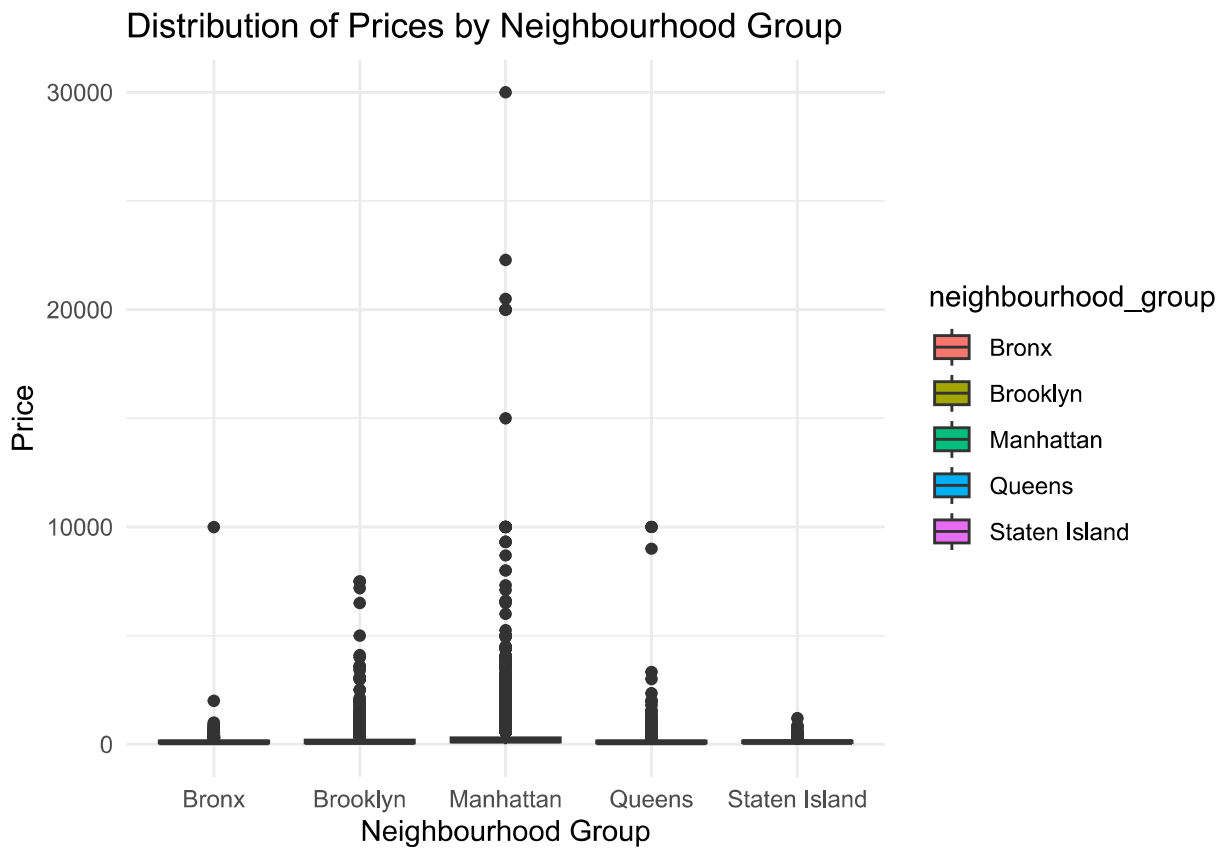


Figure 1: Airbnb locations throughout NYC

Outlier Handling

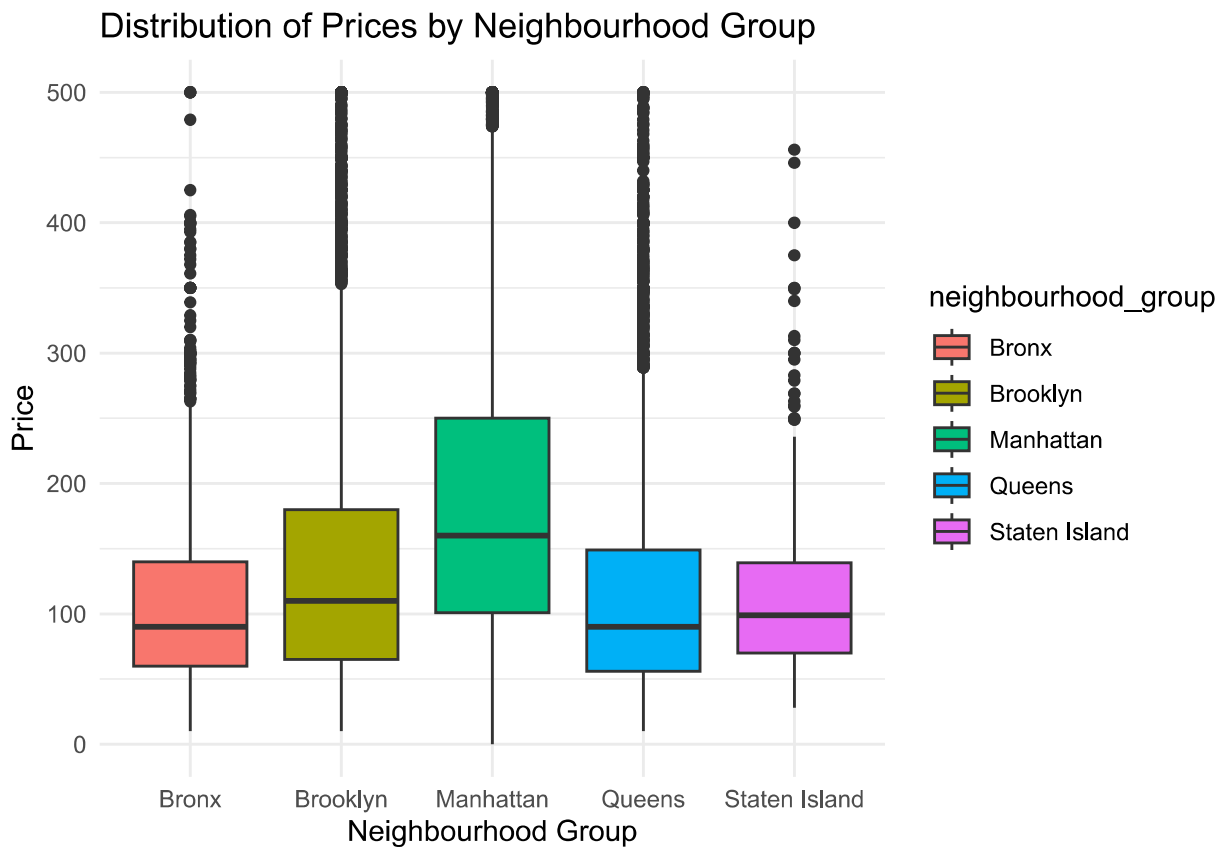
A crucial step in the analysis involves addressing outliers in the dataset, specifically in the context of listing prices. Extreme outliers are identified and filtered to enhance the robustness of subsequent analyses.

```
ggplot(df, aes(x = neighbourhood_group, y = price, fill = neighbourhood_group)) +
  geom_boxplot() +
  labs(title = "Distribution of Prices by Neighbourhood Group",
       x = "Neighbourhood Group",
       y = "Price") +
  theme_minimal()
```



The box plot illustrates the distribution of listing prices across different neighborhood groups. It becomes evident from the plot that extreme outliers are present, making it challenging to discern the detailed distribution due to their impact. To enhance the clarity of the distribution, listings with prices exceeding $500 are identified as potential outliers and targeted for removal.

```
df_filtered <- df %>% filter(price <= 500)
ggplot(df_filtered, aes(x = neighbourhood_group, y = price, fill = neighbourhood_group)) +
  geom_boxplot() +
  labs(title = "Distribution of Prices by Neighbourhood Group",
       x = "Neighbourhood Group",
       y = "Price") +
  theme_minimal()
```




```
num_rows_filtered <- nrow(df_filtered)
cat("Filtered Data:", num_rows_filtered, '\n')
```

```
## Filtered Data: 36417
```

```
cat("Unfiltered Data:", nrow(df), '\n')
```

```
## Unfiltered Data: 38792
```

```
cat("Remaining Data:", nrow(df_filtered)/nrow(df), '\n')
```

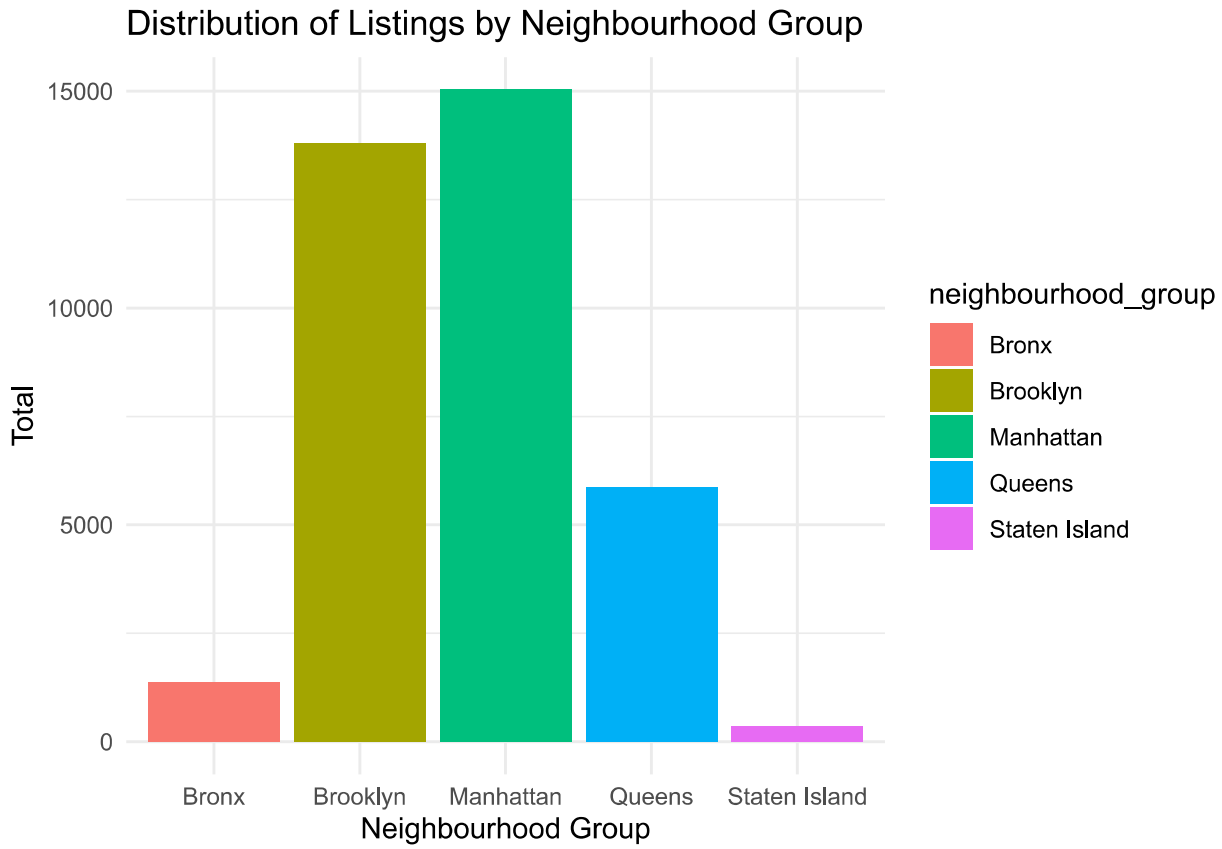
```
## Remaining Data: 0.938776
```

The code demonstrates the practical step of filtering the dataset to exclude listings with prices exceeding $500. The box plot for the filtered data provides a clearer representation of price distribution within a more reasonable range, it enhances the visibility of patterns and trends while mitigating the impact of extreme outliers. Extreme prices can disproportionately influence statistical analyses, and this filtering step contributes to a more accurate representation of typical pricing. About 94% of the data was retained after filtering the extreme price outliers which will better support modeling relationships between variables without the noise introduced by extreme values. The filtered dataset will be used as the primary dataset.

Variable Inspection

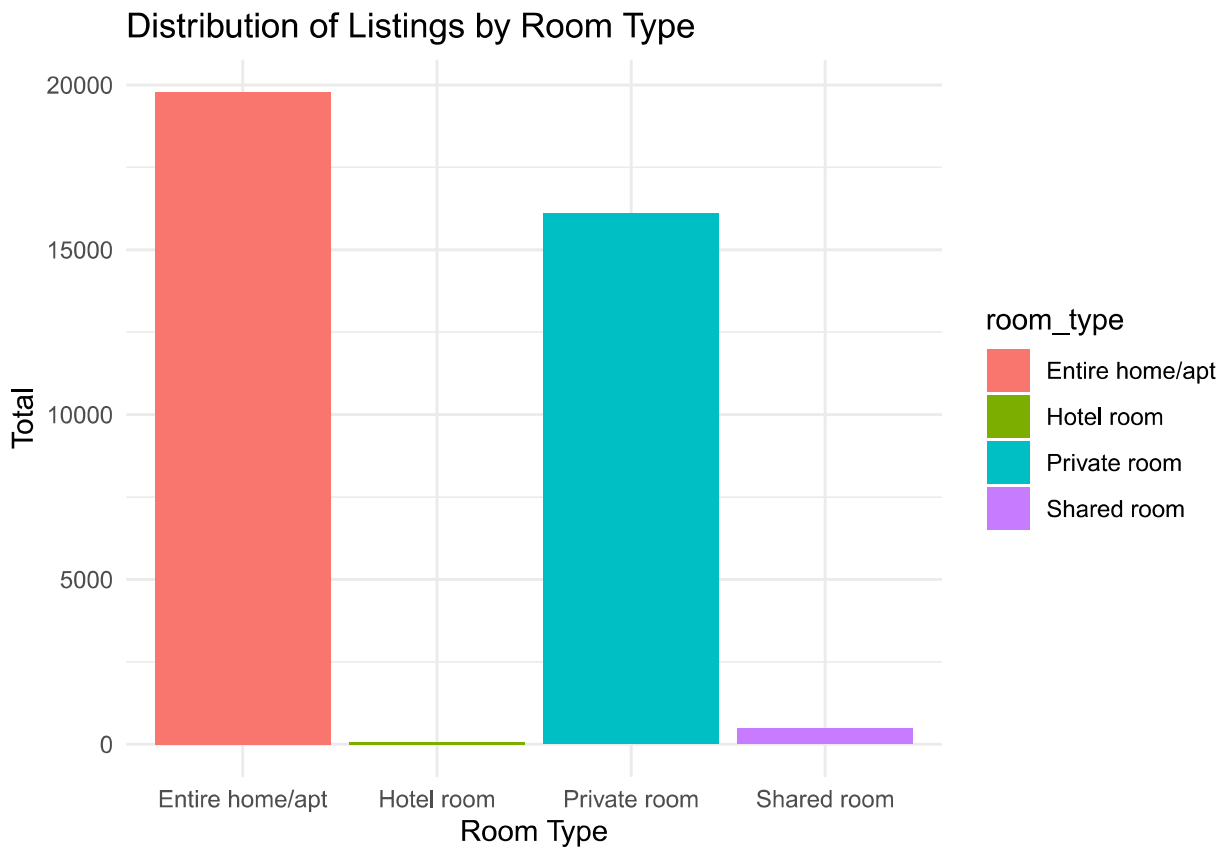
The inspection of individual variables is conducted to ascertain their characteristics, such as data type, number of missing values, and the distribution of unique values. Visualization tools like bar plots and box plots are employed to present a clearer picture of categorical and numerical variables, respectively.

```
# Barplot for 'neighbourhood_group'
ggplot(df, aes(x = neighbourhood_group, fill = neighbourhood_group)) +
  geom_bar() +
  labs(title = "Distribution of Listings by Neighbourhood Group",
       x = "Neighbourhood Group", y = "Total") +
  theme_minimal()
```



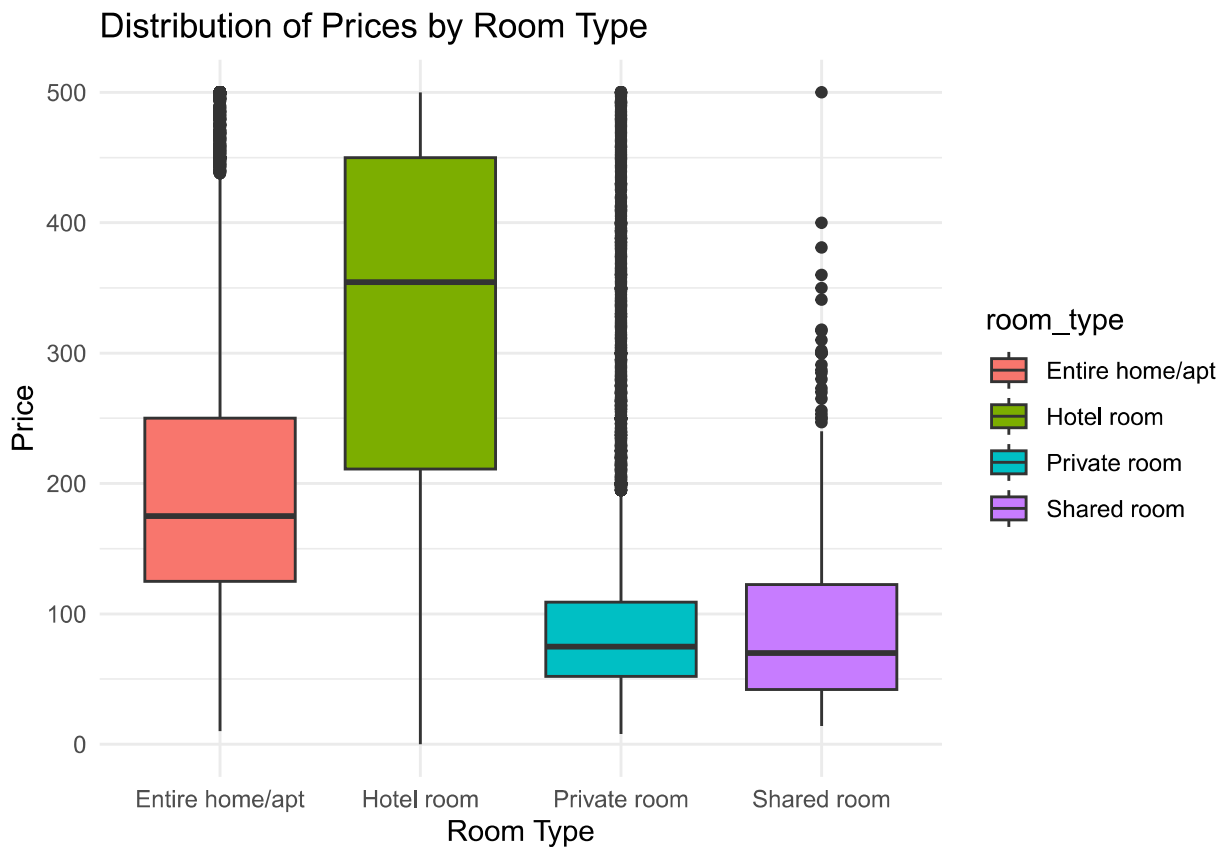
The visualization allows for a quick comparison of listing distribution across different neighborhood groups. Manhattan and Brooklyn are highlighted as the primary areas with a high density of Airbnb listings. This aligns with expectations, considering these regions are popular among tourists due to attractions, cultural sites, and vibrant neighborhoods.

```
# Barplot for 'room_type'
ggplot(df, aes(x = room_type, fill = room_type)) +
  geom_bar() +
  labs(title = "Distribution of Listings by Room Type",
       x = "Room Type",
       y = "Total") +
  theme_minimal()
```



The bar plot illustrates the count of Airbnb listings categorized by different room types. A significant majority of listings fall into the categories of "Entire home/apt" or "Private room", revealing a clear preference among users for private accommodations. The data suggests that the architectural landscape of New York City may influence the distribution of room types as instances of "Shared room" listings are comparatively lower, indicating that shared accommodations might be less common or favored due to spatial constraints. The low number of available hotel room listings suggests that hotel room listings, which may incur additional fees on platforms like Airbnb, might be strategically managed through the companies' private booking websites.

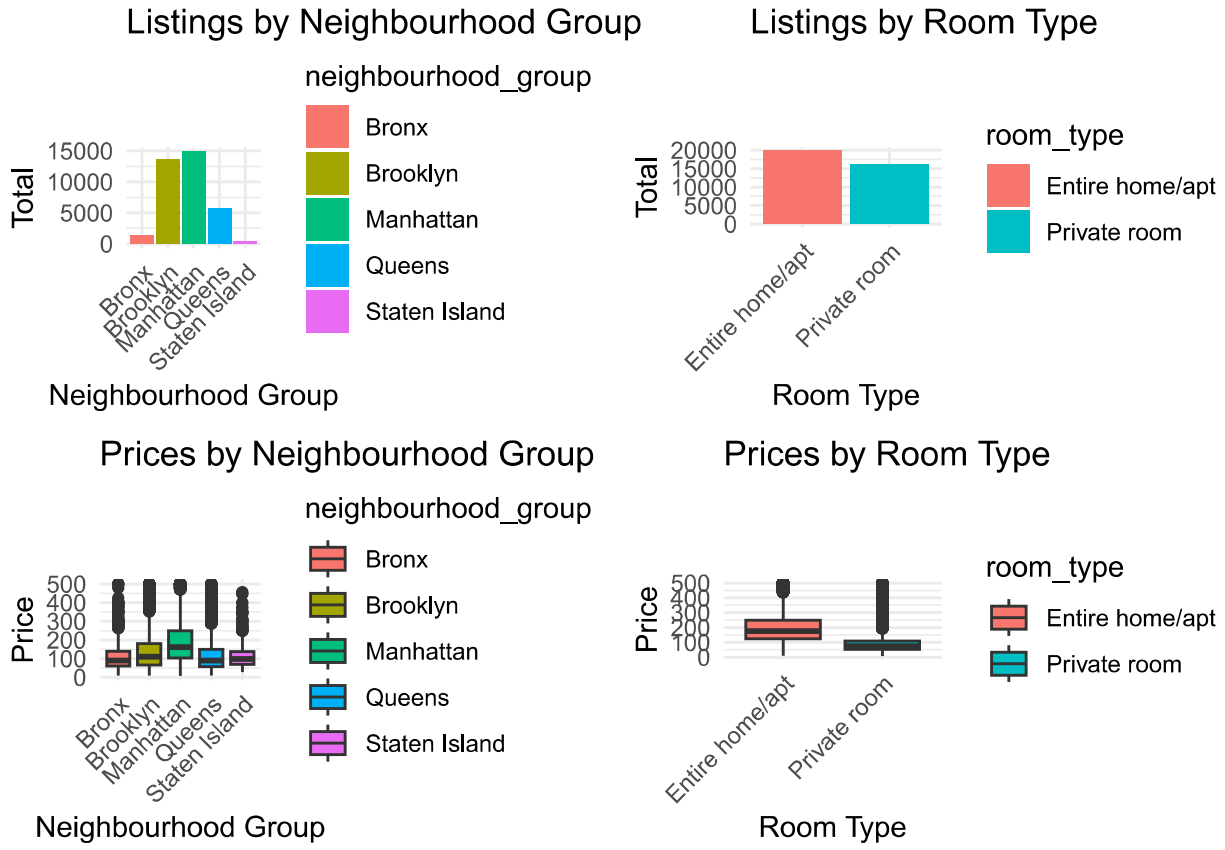
```
# Boxplot for 'room_type' versus 'price'
ggplot(df, aes(x = room_type, y = price, fill = room_type)) +
  geom_boxplot() +
  labs(title = "Distribution of Prices by Room Type",
       x = "Room Type",
       y = "Price") +
  theme_minimal()
```



The box plot illustrates the distribution of listing prices across different room types, including entire home/apartment, private rooms, and shared rooms. Shared rooms and private rooms tend to have lower prices, while entire home/apartment listings show a broader range higher of prices. Hotel prices show the highest range of prices, as hotels might increase prices to offset the impact of fees associated with platforms like Airbnb.

Since Hotel room and Shared room take up very little proportion, and the mean price is significantly different from other two main room types, we will exclude them from our dataset.

```
# exclude Hotel room and Shared room
df <- df[df$room_type %in% c('Entire home/apt', 'Private room'), ]
```



Step 3: Hypothesis Test

While boxplots can provide visual insights into the distribution of data and highlight differences in means between groups, statistical tests like ANOVA (Analysis of Variance) serve a specific purpose in hypothesis testing.

Hypothesis 1: Except for Manhattan and Brooklyn, the mean price for other 3 boroughs are at the same level.

Solution: when there are more than two groups to compare, ANOVA can assess whether there are any statistically significant differences in means among multiple groups. In this case, we select Bronx, Queens and State Island, then perform ANOVA test to see any significant different of their mean prices.

H0: no significant difference among the mean prices of selected boroughs *Ha*: at least one selected borough is different in mean price

```
# Fit an ANOVA model for occupancy rate over boroughs
df_aov <- df[df$neighbourhood_group %in% c('Bronx', 'Queens', 'Staten Island'), ]
lm_aov <- aov(price ~ neighbourhood_group, data = df_aov)
summary(lm_aov)
```

```
##                       Df   Sum Sq Mean Sq F value Pr(>F)
## neighbourhood_group    2    20653   10326   1.579  0.206
## Residuals           7447 48711604    6541
```

The p-value is 0.206, which is greater than 0.05. Therefore, it would likely fail to reject the null hypothesis, suggesting that there is no significant difference in the means of the groups represented by the variable "neighbourhood_group". **Therefore, we can combine the 3 boroughs for afterward analysis, to reduce categories and model complexity**

```
# group Bronx, Queens and Staten Island as BQSI
df$neighbourhood_group[df$neighbourhood_group %in% c("Bronx", "Queens", "Staten Island")] <- "BQSI"
```

Hypothesis 2: the mean price for Entire home/apt and Private room are significant different

Solution: use t-test to check any significant difference in the mean prices between the two room types. t-test is used to compare the means of two groups when the assumption of equal variances is violated. In this case, null hypothesis (H0) is no significant difference in mean price between the two room types, and the alternative hypothesis (Ha) is that there is a significant difference.

H0: no significant difference in mean price of the two room types Ha: there is significant difference in mean price of the two room types

```
entire_home <- df$price[df$room_type == "Entire home/apt"]
private_room <- df$price[df$room_type == "Private room"]
t_test <- t.test(entire_home, private_room)
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  entire_home and private_room
## t = 112, df = 35795, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  100.1245 103.6915
## sample estimates:
## mean of x mean of y
## 199.09420  97.18617
```

The p-value is less than 2.2e-16 (a very small number, essentially zero). The very small p-value (less than 0.05) suggests that we would reject the null hypothesis. In a word, there is a statistically significant difference in means between the "entire_home" and "private_room"

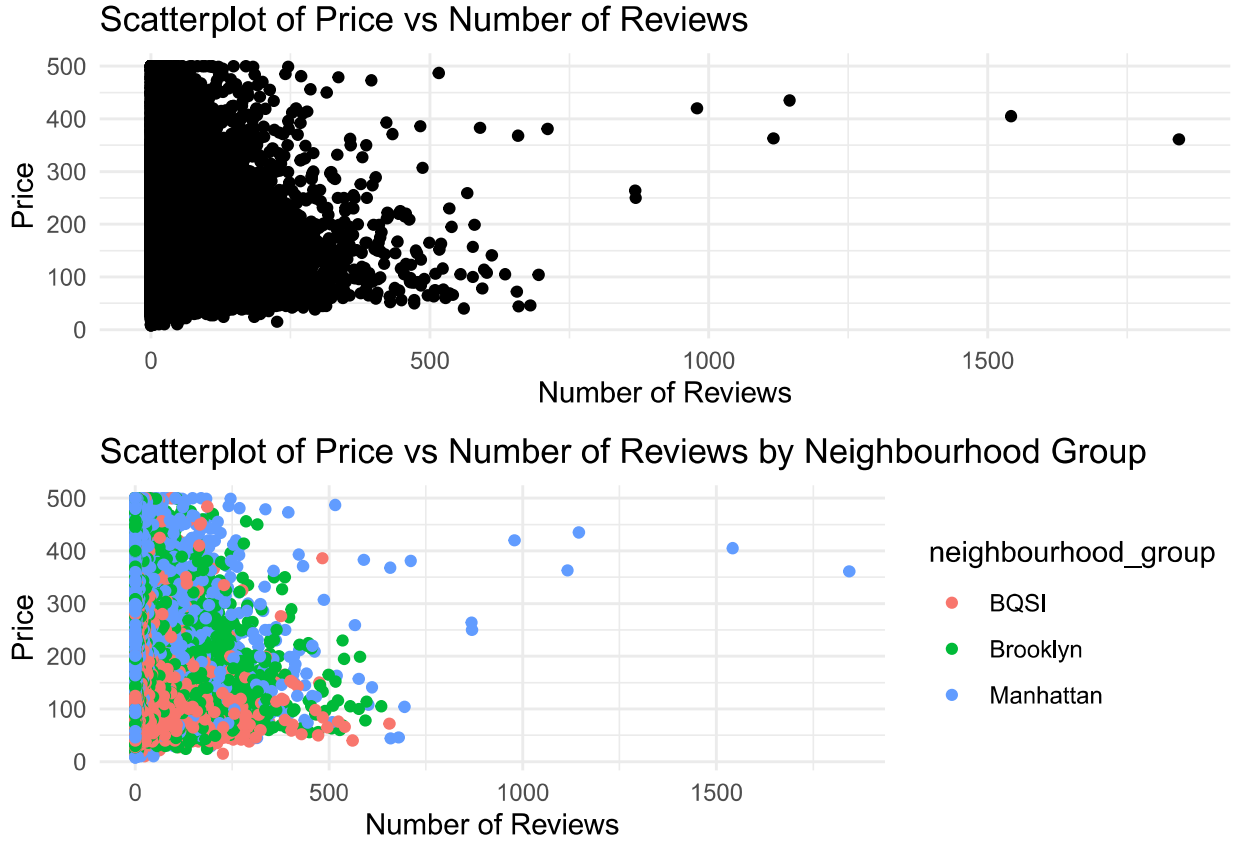
Step 4: Regression Analysis

Start from Simple Linear Regression

```
# Scatterplot of price vs num_of_reviews
scat_p1 <- ggplot(df, aes(x = number_of_reviews, y = price)) +
  geom_point() +
  labs(title = "Scatterplot of Price vs Number of Reviews",
       x = "Number of Reviews",
       y = "Price") +
  theme_minimal()
```

```r
# Scatterplot of price vs num_of_reviews with color for neighbourhood_group
scat_p2 <- ggplot(df, aes(x = number_of_reviews, y = price, color = neighbourhood_group)) +
  geom_point() +
  labs(title = "Scatterplot of Price vs Number of Reviews by Neighbourhood Group",
       x = "Number of Reviews",
       y = "Price") +
  theme_minimal()

grid.arrange(scat_p1, scat_p2,
             ncol = 1)
```




```r
# Scatterplot of price vs availability_365
ggplot(df, aes(x = availability_365, y = price, color = neighbourhood_group)) +
  geom_point() +
  labs(title = "Scatterplot of Price vs Availability 365",
       x = "Availability 365",
       y = "Price") +
  theme_minimal()
```