NYC Airbnb Listing Price Prediction

Kien Le, Tony Nguyen

2023-05-25

1. PROJECT OVERVIEW

Airbnb is an online marketplace and hospitality service platform that allows individuals to rent out their properties or spare rooms to guests seeking short-term accommodations. The Airbnb platform connects hosts and guests, providing a user-friendly interface where hosts can list their available spaces, set their own prices, and provide detailed descriptions and photos of their accommodations. Guests can then search for listings based on their desired location, travel dates, and specific preferences. The main purpose of this report is to analyze and predict the prices of Airbnb listings in New York City (NYC). This project is intended to provide valuable insights for both hosts and guests, enabling them to make informed decisions regarding pricing and booking accommodations in NYC.

Guiding questions:

1. What are the key factors that influence the pricing of Airbnb listings in NYC?
2. How accurate are the predictions of our model, and what are the potential limitations or uncertainties associated with the predictions?

Domain: Housing Price. The goal is to predict a continuous target variable (in this case, the price) based on a set of input variables (the other columns in the dataset). We will split the data set into a training set and a testing set, where the training set is used to train the model, and the testing set is used to evaluate its performance. Once we have built and evaluated the model, we can use it to predict the price of new Airbnb listings in the future based on their features. These are our initial thoughts and they might change later on

2. DATA & RESOURCES USED.

The data set we used is on Kaggle website and it is called the "New York City Airbnb Open Data".The New York City Airbnb Open Data set contains information about Airbnb listings in New York City. Here is the list of the variables that we will mostly focus on:

1.`neighbourhood_group`: The borough or area of New York City where the listing is located.

Data Type: Categorical (string)

Example: "Brooklyn"

2.`room_type`: The type of room or accommodation.

Data Type: Categorical (string)

Example: "Private room"

3.`price`: The price per night to rent the listing.

Data Type: Numeric (integer)

Example: 149

4.`minimum_nights`: The minimum number of nights required for a stay.

Data Type: Numeric (integer)

Example: 1

5.`number_of_reviews`: The number of reviews received for the listing.

Data Type: Numeric (integer)

Example: 9

6.`availability_365`: The number of days the listing is available within a year.

Data Type: Numeric (integer)

Example: 365

7.`calculated_host_listings_count`: The total number of listings by the host.

Data Type: Numeric (integer)

Example: 6

3. DATA PREPROCESSING & ANALYSIS

Checking duplicated values:

```
sum(duplicated(data))
```

```
## [1] 0
```

Checking missing values:

```
sum(is.na(data))
```

```
## [1] 7675
```

Dropping unnecessary columns which does not generate useful insights. The columns will be dropped are: `id`, `name`, `host_id`, `host_name`, `neighbor_hood`, `last_review`, `longtitude`, and `latitude`

```
data <- data [, -c(1,2,3,4,6,7,8,13)]
```

Check how many levels in column `room_type` and `neighborhood_group`, and turn them into factors:

```
unique(data$room_type)
```

```
## [1] "Private room"    "Entire home/apt" "Shared room"
```

```
unique(data$neighbourhood_group)
```

```
## [1] "Queens"        "Bronx"         "Brooklyn"      "Manhattan"
## [5] "Staten Island"
```

```
data$neighbourhood_group <- factor(data$neighbourhood_group)
data$room_type <- factor(data$room_type)
```

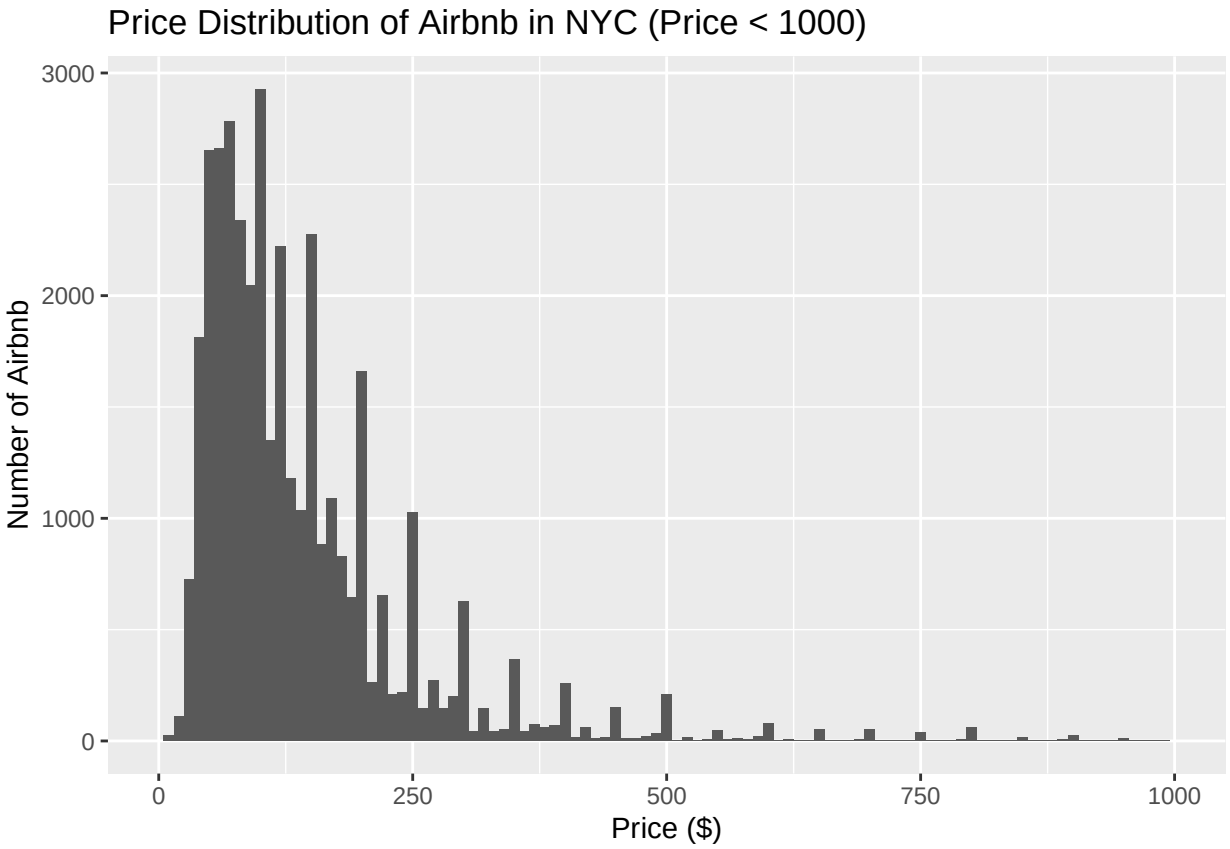
Fill out missing values and check to see if there is any missing values again:

```
data$reviews_per_month <- ifelse(is.na(data$reviews_per_month), 0, data$reviews_per_month)
sum(is.na(data))
```

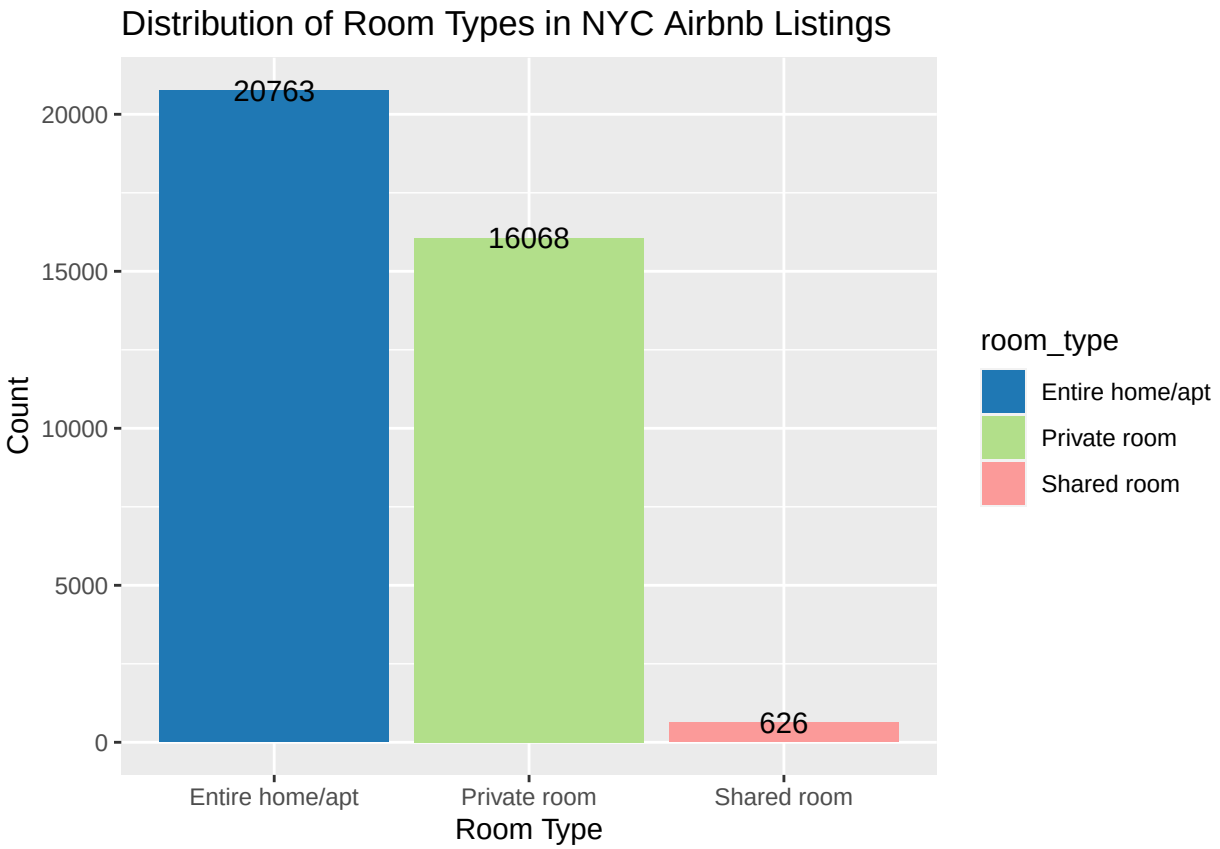
```
## [1] 0
```

Price distribution

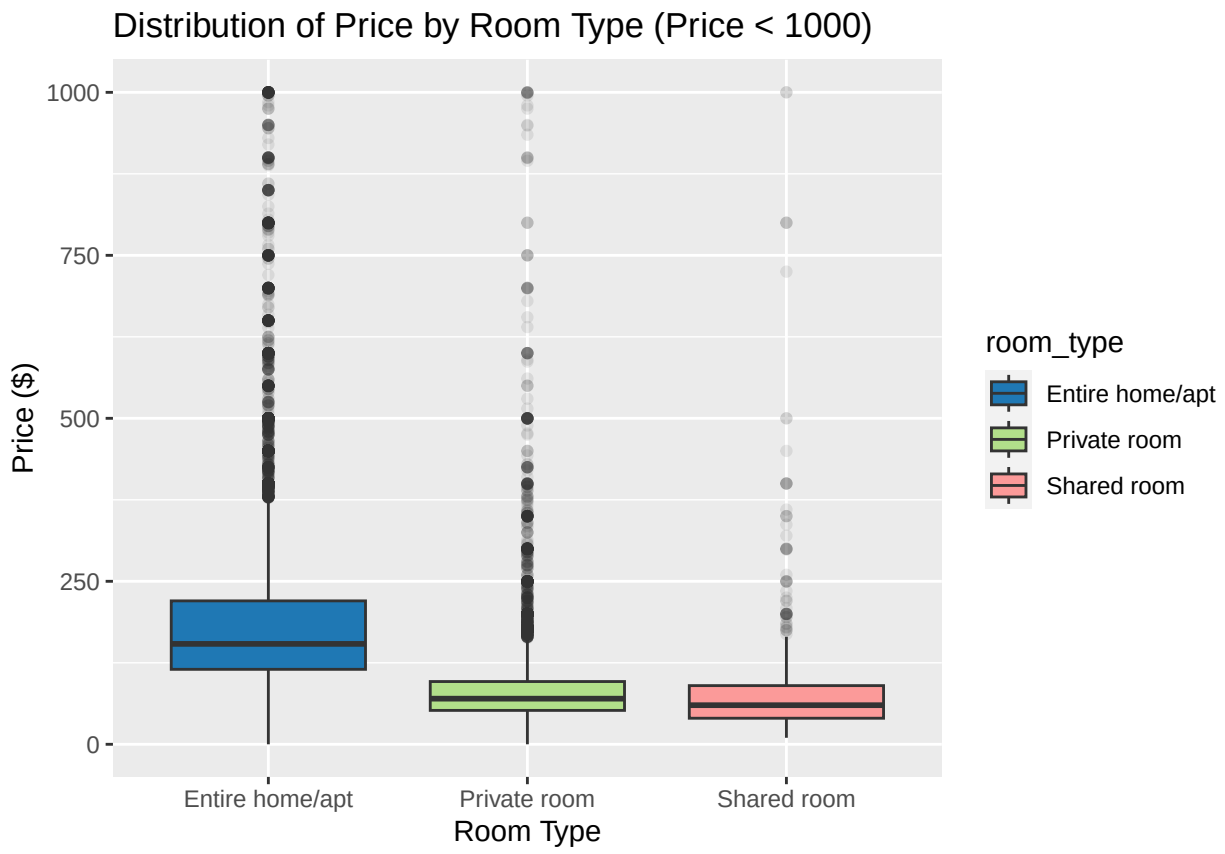
The price range in NYC for airbnb is from 20 to more than 1000 per night, with the peak of about 100 dollars per night. I set the range of price to smaller than 1000 because the dataset is right-skewed so by doing that, it is easier to see the distribution.



In this dataset, most of the room type are “entire home/apt” and “private room”.



There are a lot of outliers so I set the limit of y-axis to 1000 to have a better visualization of the dataset. We could see that the Shared room does not have much that is above 2000. However, the other have several that are approximately $10000



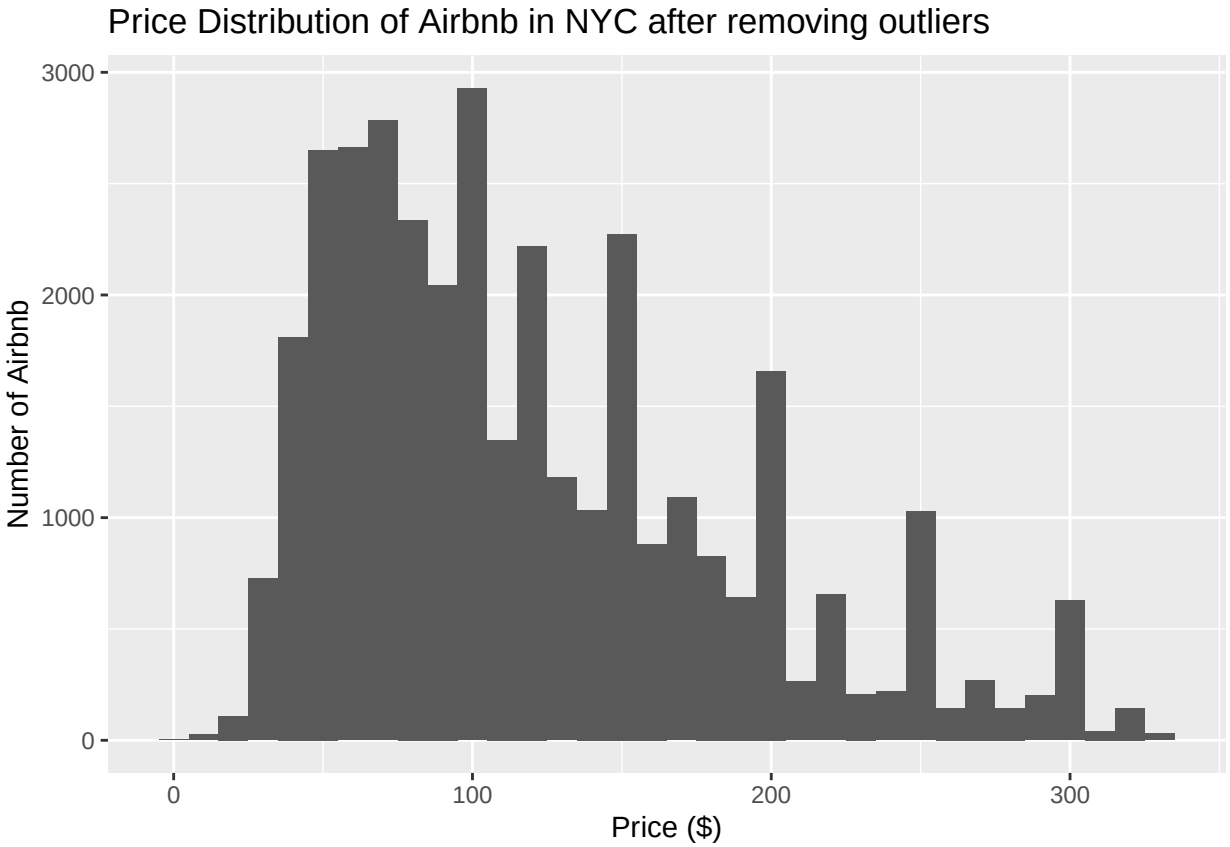
Remove outliers

Calculate IQR and quantile to remove outliers:

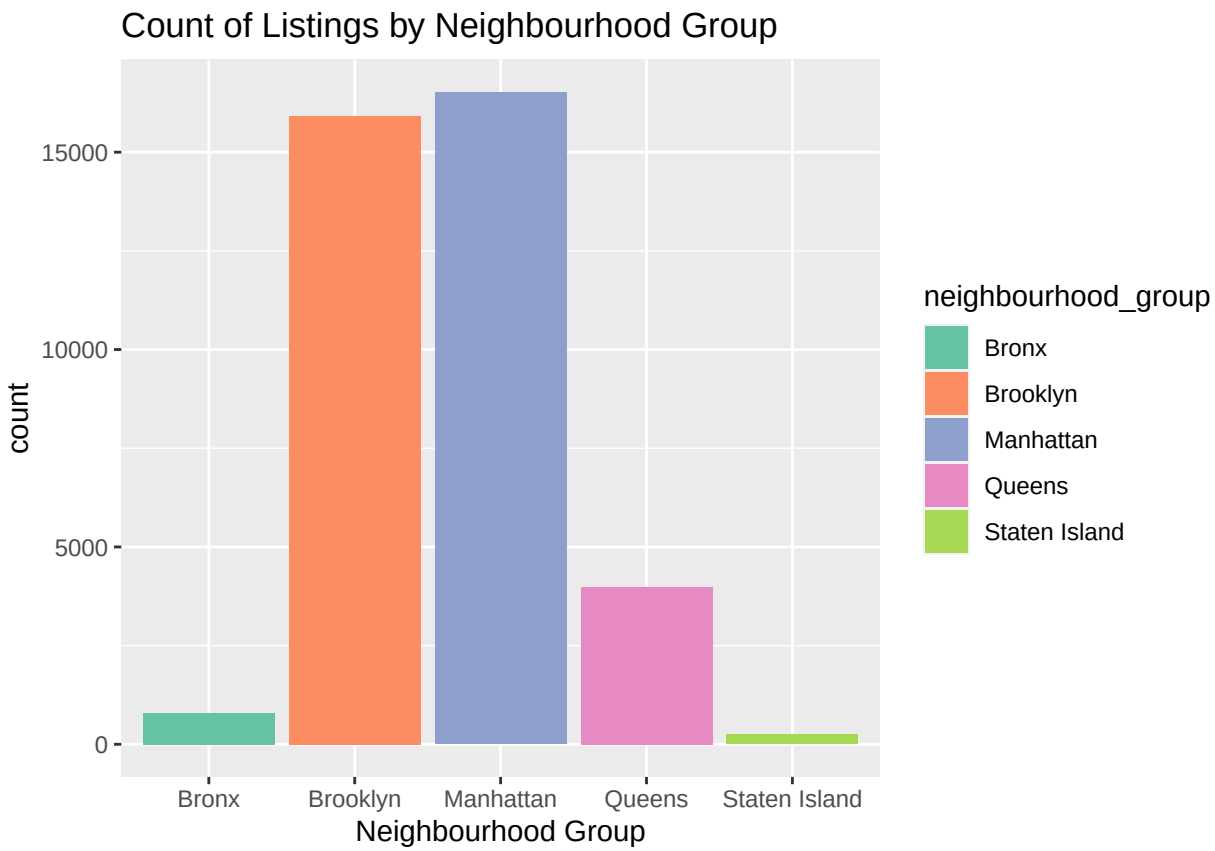
```
quantile <- quantile(data$price, probs=c(.25, .75))
iqr <- IQR(data$price)

#New dataframe without outliers
data_no_OL <- data %>% filter(price > (quantile[1] - 1.5*iqr) & price < (quantile[2] + 1.5*iqr))
```

Now let's plot the data set with our new dataframe after handling outliers.



Also take a look at the count of listings that each cities have in Airbnb platform. We can infer that Manhattan and Brooklyn seems to be a very attractive place to visit or an ideal place to stay!



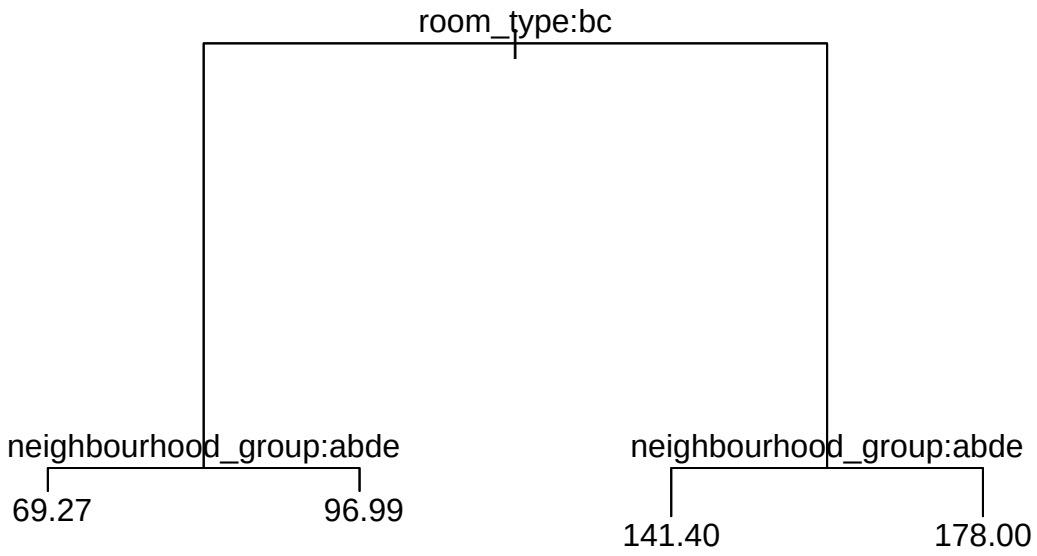
Decision Tree

The very first statistical model we want to apply our dataset is the Decision Tree model which we have learned in class. We decided to choose this model since we think it is the most suitable to our dataset and what it has. We will be splitting the data into training and testing set and also set the "price" column - which is our dependent variable, assign to true.val.

Afterwards, we can plot the decision tree here. The initial split was chosen by the predictor "room_type" and second partition was decided by "neighborhood_group" (which include Bronx, Queens, Manhattan, Brooklyn). And therefore, predict the potential price per night for listings.

```
# Fitting the model to training data
tree.airbnb <- tree(price ~ ., data = train.data)

# Plotting the fit object
plot(tree.airbnb)
text(tree.airbnb)
```



We also examine the RMSE of the Decision Tree Model and it has the value of 99.23, which is a bit high for us when predicting the price for Airbnb listings.

```
## [1] 99.23069
```

Multiple Linear Regression Model

The linear regression model on data without outliers has an R-squared value of 0.4825, which means that the model explains approximately 48.25% of the variability in the target variable (price). Some of the coefficient can be interpreted as follow: Airbnb listings with a private room have an expected price that is approximately $75.63 lower than listings with the reference room type (entire home/apt), holding all other variables constant. Airbnb listings with a shared room have an expected price that is approximately $100.90 lower than listings with the reference room type (entire home/apt), holding all other variables constant.

Output:

Residual standard error: 50.44 on 35210 degrees of freedom

Multiple R-squared: 0.4309, Adjusted R-squared: 0.4307

F-statistic: 2666 on 10 and 35210 DF, p-value: $<$ 2.2e-16

Now I decided to take the Stepwise function to see if the model after step explains the variable “Price” better. It turns out that the R-saqured value was not improved compared to the initial model we put out.

Assign the new model after using Stepwise.

```
# Assign to a new model after stepwise
new_model <- lm(price ~ neighbourhood_group + minimum_nights + room_type + number_of_reviews + availabil
```

Predicting and Comparing MSEs (Regression)

First, we create a training set and test set. The training set has 80% of the data and the test set has 20%. THen we Create predictions fr the model. The printed result is the root mean squared error (RMSE) of the initial model and it is 50.00326, which means that the average difference between the predicted price and the actual price on the testing set is around $49. This indicates that the model is relatively accurate in predicting the prices of Airbnb listings based on the given features. We also print out the RMSE for the model after using Stepwise, and the RMSE turns out to be slightly smaller but not significant (50.00046)

```
## [1] 50.00326
```

```
## [1] 50.00046
```

Interaction Linear Regression

We decided to add an interaction term to the model to see if it fits the model and explains the dependent variable "price" better. We consider 2 questions to answer whether the interaction term contribute in a meaningful way to the explanatory power of the model. And in fact, the multiple R-squared did Increase but not by much, and 5 out of 8 of the interaction term are statistically significant (***). The output below shows that our multiple R-squared is larger (43.39%) than the Multiple linear regression Rsquared value (43%)

Ouput:

Residual standard error: 50.31 on 35203 degrees of freedom

Multiple R-squared: 0.4339, Adjusted R-squared: 0.4336

F-statistic: 1587 on 17 and 35203 DF, p-value: $<$ 2.2e-16

```
new_model2 <- lm(price ~ minimum_nights + number_of_reviews + availability_365 + neighbourhood_group * 

summary(new_model2)
```

The RMSE of the interaction model is also lower (with 49.8) than that of the multiple linear regression's.

```
predictions2 <- predict(new_model2, newdata = test)

# Calculate the root mean squared error (RMSE) to evaluate the performance of the model
rmse_inter <- sqrt(mean((test$price - predictions2)^2))
print(rmse_inter)
```

```
## [1] 49.89207
```

4. SUMMARY & CONCLUSION

In conclusion, our initial goal of this analysis is to see how pricing per night on Airbnb platform is affected by other factors such as their availability, number of reviews, room type, or locations. For our preliminary analysis, we figured out that our dataset has a lot of outliers and we handled it by caculating IQR to eliminate them. The price per night has a high distribution from the range from approximately $60 to more than $100. In addition, Manhattan and Brooklyn are highly popular tourist destinations in New York City, attracting a large number of visitors throughout the year. The high demand for accommodations in these areas encourages property owners to list their spaces on Airbnb to cater to the influx of tourists.

In terms of Statistcal Models, also we have used Mulitple Linear Regression, Interaction Linear Regression and also attempted to put in Decision Tree to examine how well these models predict our dependent variable (price) in this dataset and we observe that the model with the lowest RMSE is the Interaction Linear Regression Model

Key Factors Influencing Pricing: Our analyses revealed that several factors play a significant role in determining Airbnb prices. These factors include the neighborhood group, minimum nights required for booking, room type, number of reviews, and availability throughout the year. The interaction between neighborhood group and room type was also found to have an impact on pricing.

5. REFERENCES

Link to the original dataset obtained on Kaggle: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data