
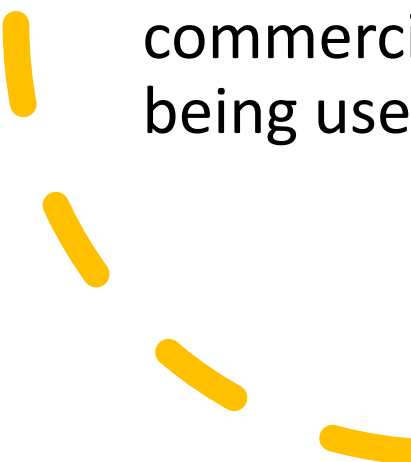


AIR BNB

- GROUP -5
- SWETHA POTHU
- UDAY NAIDU GADWALA
- ALEKYA JALADI
- PROFESSOR: JIN FANG

INTRODUCTION

- Airbnb is an online community marketplace that connects people looking to rent their homes with people who are looking for accommodations.
- Airbnb users include hosts and travelers: hosts list and rent out their unused spaces, and travelers search for and book accommodations in 192 countries worldwide.
- It's free to create a listing, and hosts decide how much to charge per night, per week or per month.
- Each listing allows hosts to promote properties through titles, descriptions, photographs with captions and a user profile where potential guests can get to know a bit about the hosts.
- Travelers (or "guests*") search the available database of properties by entering details about when and where they'd like to travel. Travelers can further refine searches by making selections for: Room type, size, price, amenities etc.

- 
- The dataset contains information about Airbnb listings in New York City.
 - It includes data from 2019 such as price, location, number of reviews, availability, and more.
 - The dataset consists of multiple files, including listings, reviews, and neighborhood data.
 - There are over 49.000 listings in the dataset.
 - The data was sourced from Inside Airbnb, an independent, non-commercial set of tools and data that allows you to explore how Airbnb is being used in different cities around the world.
- 

DATA DESCRIPTION

| Table name | Table description |
|--------------------------------|---|
| id | listing ID |
| name | name of the listing |
| host_id | host ID |
| host_name | name of the host |
| neighbourhood_group | location of the listing by borough |
| neighborhood | location of the listing by neighborhood |
| latitude | latitude coordinates |
| longitude | longitude coordinates |
| room_type | type of room |
| price | price in dollars per night |
| minimum_nights | minimum number of nights to stay |
| number_of_reviews | number of reviews for the listing |
| last_review | date of last review |
| reviews_per_month | number of reviews per month |
| calculated_host_listings_count | number of listings by the same host |
| availability_365 | number of days when listing is available for booking within the next 365 days |

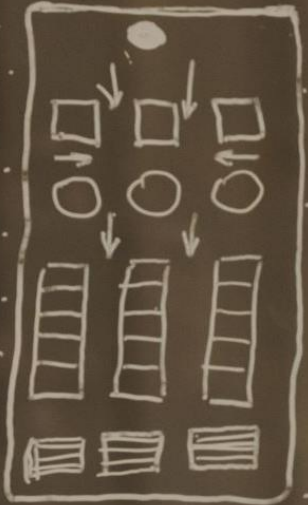


PROBLEM STATEMENT

- To predict price base on selected independent variables like reviews_per month, number_of_reviews, availability_365,etc.
- To investigate the critical factors which are mostly affecting dependent variable that is price.
- To build different models like random forest and multiple linear regression and find out which model can predict our dependent variable more accurately.

DATA PREPROCESSING

1. Checking the dimensions of the dataset using `dim ()`
2. Checking for null values in the dataset using `colSums (is.na())`
3. Replacing the null values in the 'reviews_per_month' column with zero.
4. Creating dummy variables for the 'room_type', 'neighbourhood', and 'neighbourhood_group' columns using `model.matrix()*`
5. Combining the original dataset with the dummy variables using `bind()*`
6. Removing unnecessary columns from the dataset.
7. Standardizing the data by taking the logarithm of certain columns.
8. Removing null and infinite values from the standardized data using `is. finite()`



METHODS



The methods included in this dataset are linear regression and random forest models



The objective of linear regression is to find the best linear relationship between the dependent variable and the independent variable(s), such that the sum of the squares of the residuals (i.e., the difference between the actual Y value and the predicted Y value) is minimized.



The output of the random forest algorithm is the average (for regression) or the mode (for classification) of the predictions made by the individual decision trees.



MODEL BUILDING

- Splitting the data into training and testing sets using `createDataPartition()`
- Removing statistically insignificant and highly correlated columns from the training data using p-values and VIF (Variance Inflation Factor) values.
- Training a linear regression model using the pre-processed training data.
- Using the trained model to predict the prices of the test data.
- Plotting the predicted prices against the actual prices to evaluate the model's performance.



LINEAR REGRESSION MODEL:



The code that we have performed, first split the data into training and testing datasets using the **createDataPartition ()**° function from the caret package.



It then trained a linear regression model using the **train()** function from the 'caret' package on the training data.



The ***lm()** function was used to fit the linear regression model to the training data. Once the model was trained, the code used the **predict ()** function to make predictions on the testing data.



The **lmse()** function from the Metrics package was used to calculate the root mean squared error (RMSE) of the model's predictions.

- The model has a residual standard error of 0.2209, meaning the average difference between the predicted values and actual values is 0.2209.
- The multiple R-squared value of 0.4528 indicates that 45.28% of the variance in the target variable (price) can be explained by the independent variables in the model.
- The adjusted R-squared value of 0.4519 is similar to the multiple R-squared value but takes into account the number of variables in the model.
- The F-statistic of 487.8 and its associated p-value of $< 2.2 \times 10^{-16}$ suggest that the overall model is significant and that at least one of the independent variables is a significant predictor of the target variable.

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2209 on 18272 degrees of freedom

Multiple R-squared: 0.4528, Adjusted R-squared: 0.4519

F-statistic: 487.8 on 31 and 18272 DF, p-value: $< 2.2 \times 10^{-16}$

RANDOM FOREST

- A random forest model is created using the `randomForest()` function.
- The dependent variable is the `price` column in the training data and the independent variables are all other columns except `id`, `name`, `host_id`, `host_name`, and `last_review`.
- The `ntree` parameter is set to 200 to specify the number of trees to grow in the forest.
- The `predict()` function is used to make predictions on the `test_data` using the random forest model created in step 1.
- The predicted prices are then combined with the actual prices from the `test_data` using the `cbind()` function.
- The `mean absolute error` (MAE) of the predictions is calculated using the `mean()` and `abs()` functions.
- The accuracy of the random forest model is calculated by subtracting the MAE from the mean of the `test_data$price` and dividing the result by the mean of the `test_data$price`.
- The accuracy of the random forest model is printed to the console using the `print()` and `paste()` functions.

- The random forest model was used to predict the price of Airbnb listings.
- The model used 200 trees and tried 77 variables at each split.
- The mean squared residual, which measures the difference between the predicted and actual values of the target variable, was 0.0325.
- The percent variance explained by the model was 63.51%, indicating that the model explained a significant proportion of the variation in the target variable.
- The output suggests that the random forest model performed well in predicting the prices of Airbnb listings, with an accuracy of 93.66%.

"Accuracy of the random forest model is: 93.66 %"

```
Call:
randomForest(formula = price ~ ., data = train_data, ntree = 200)
      Type of random forest: regression
      Number of trees: 200
No. of variables tried at each split: 77

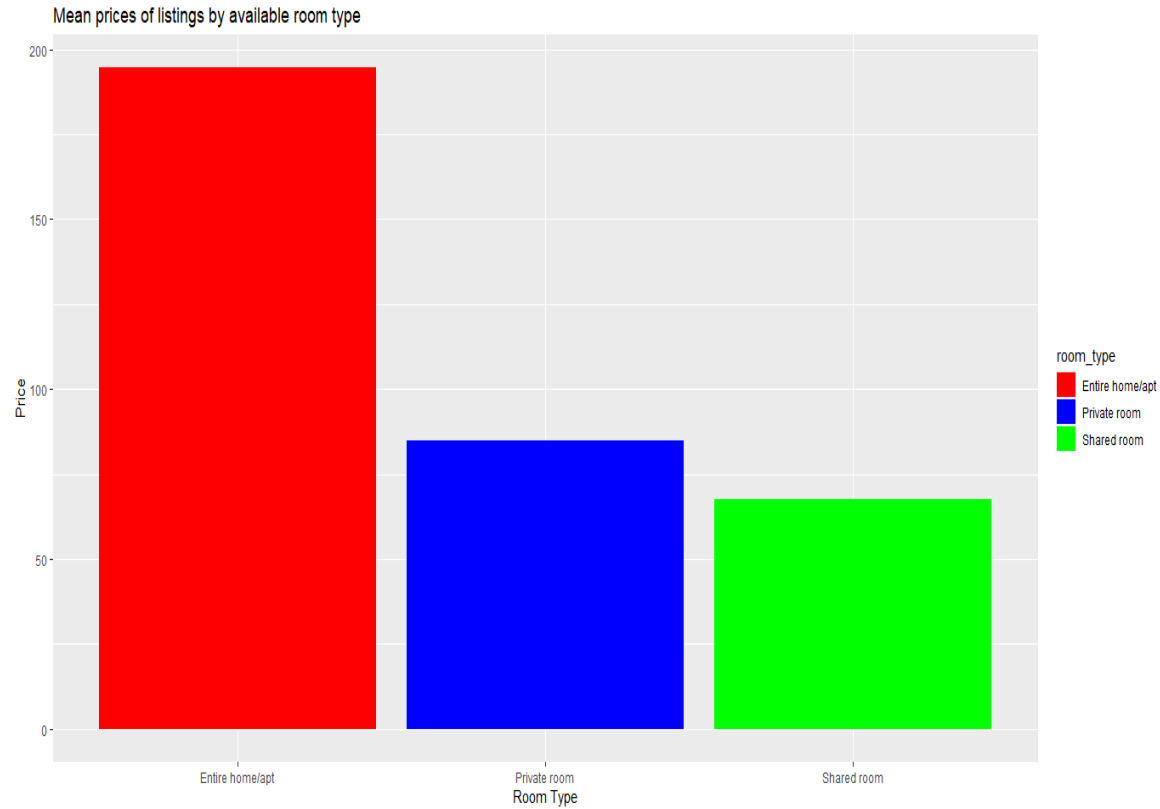
      Mean of squared residuals: 0.03248177
      % var explained: 63.51
```

>

| | | | | | | | | | | | | | | | | | |
|-----------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| room_type_Private.room | 0.04 | 0.06 | 0 | 0.18 | -0.24 | -0.07 | 0.02 | -0.11 | -0.01 | 0.04 | 0.08 | -0.16 | 0.1 | 0.01 | -0.95 | 1 | -0.14 |
| room_type_Entire.home.aprt | -0.05 | -0.08 | -0.01 | -0.19 | 0.26 | 0.07 | -0.01 | 0.11 | -0.01 | -0.05 | -0.07 | 0.16 | -0.11 | -0.01 | 1 | -0.95 | -0.16 |
| neighbourhood_group_Staten.Island | 0.02 | 0.03 | -0.19 | -0.29 | -0.01 | -0.01 | 0.02 | -0.01 | 0.06 | -0.01 | -0.07 | -0.08 | -0.03 | 1 | -0.01 | 0.01 | 0 |
| neighbourhood_group_Queens | 0.09 | 0.13 | 0.02 | 0.62 | -0.08 | -0.03 | 0.04 | -0.03 | 0.09 | -0.05 | -0.3 | -0.32 | 1 | -0.03 | -0.11 | 0.1 | 0.03 |
| neighbourhood_group_Manhattan | -0.02 | 0 | 0.59 | -0.43 | 0.16 | 0.07 | -0.05 | 0.15 | -0.01 | -0.13 | -0.75 | 1 | -0.32 | -0.08 | 0.16 | -0.16 | -0.01 |
| neighbourhood_group_Brooklyn | -0.06 | -0.12 | -0.67 | 0.02 | -0.1 | -0.04 | 0.02 | -0.12 | -0.08 | -0.13 | 1 | -0.75 | -0.3 | -0.07 | -0.07 | 0.08 | -0.02 |
| neighbourhood_group_Bronx | 0.05 | 0.07 | 0.33 | 0.22 | -0.04 | -0.02 | 0.01 | -0.02 | 0.06 | 1 | -0.13 | -0.13 | -0.05 | -0.01 | -0.05 | 0.04 | 0.03 |
| availability_365 | 0.09 | 0.2 | -0.01 | 0.08 | 0.08 | 0.14 | 0.17 | 0.23 | 1 | 0.06 | -0.08 | -0.01 | 0.09 | 0.06 | -0.01 | -0.01 | 0.06 |
| calculated_host_listings_count | 0.13 | 0.15 | 0.02 | -0.11 | 0.06 | 0.13 | -0.07 | 1 | 0.23 | -0.02 | -0.12 | 0.15 | -0.03 | -0.01 | 0.11 | -0.11 | -0.01 |
| reviews_per_month | | | | | | | | 1 | | | | | | | | | |
| number_of_reviews | -0.32 | -0.14 | -0.02 | 0.06 | -0.05 | -0.08 | 1 | -0.07 | 0.17 | 0.01 | 0.02 | -0.05 | 0.04 | 0.02 | -0.01 | 0.02 | -0.02 |
| minimum_nights | -0.01 | -0.02 | 0.02 | -0.06 | 0.04 | 1 | -0.08 | 0.13 | 0.14 | -0.02 | -0.04 | 0.07 | -0.03 | -0.01 | 0.07 | -0.07 | 0 |
| price | 0.01 | 0.02 | 0.03 | -0.15 | 1 | 0.04 | -0.05 | 0.06 | 0.08 | -0.04 | -0.1 | 0.16 | -0.08 | -0.01 | 0.26 | -0.24 | -0.05 |
| longitude | 0.09 | 0.13 | 0.08 | 1 | -0.15 | -0.06 | 0.06 | -0.11 | 0.08 | 0.22 | 0.02 | -0.43 | 0.62 | -0.29 | -0.19 | 0.18 | 0.03 |
| latitude | 0 | 0.02 | 1 | 0.08 | 0.03 | 0.02 | -0.02 | 0.02 | -0.01 | 0.33 | -0.67 | 0.59 | 0.02 | -0.19 | -0.01 | 0 | 0 |
| host_id | 0.59 | 1 | 0.02 | 0.13 | 0.02 | -0.02 | -0.14 | 0.15 | 0.2 | 0.07 | -0.12 | 0 | 0.13 | 0.03 | -0.08 | 0.06 | 0.07 |
| id | 1 | 0.59 | 0 | 0.09 | 0.01 | -0.01 | -0.32 | 0.13 | 0.09 | 0.05 | -0.06 | -0.02 | 0.09 | 0.02 | -0.05 | 0.04 | 0.06 |

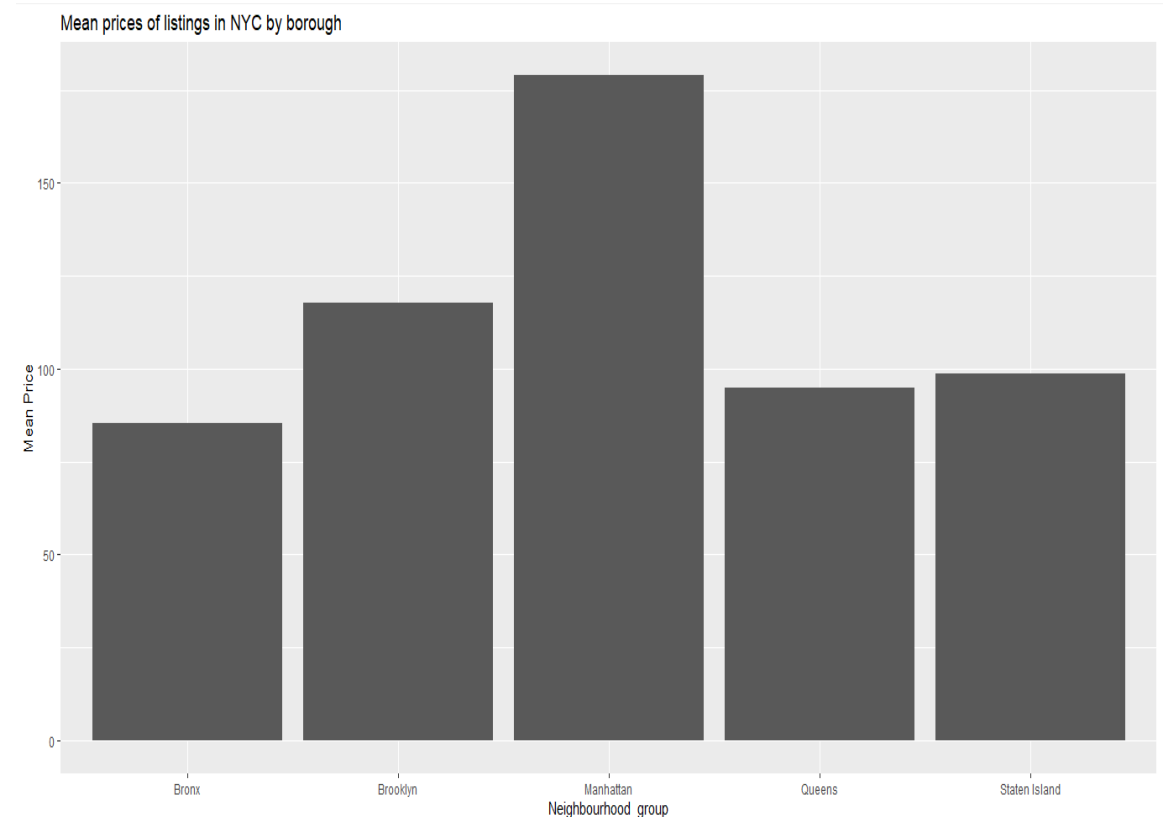
CORRELATION BETWEEN ALL THE VARIABLES IN THE DATASET

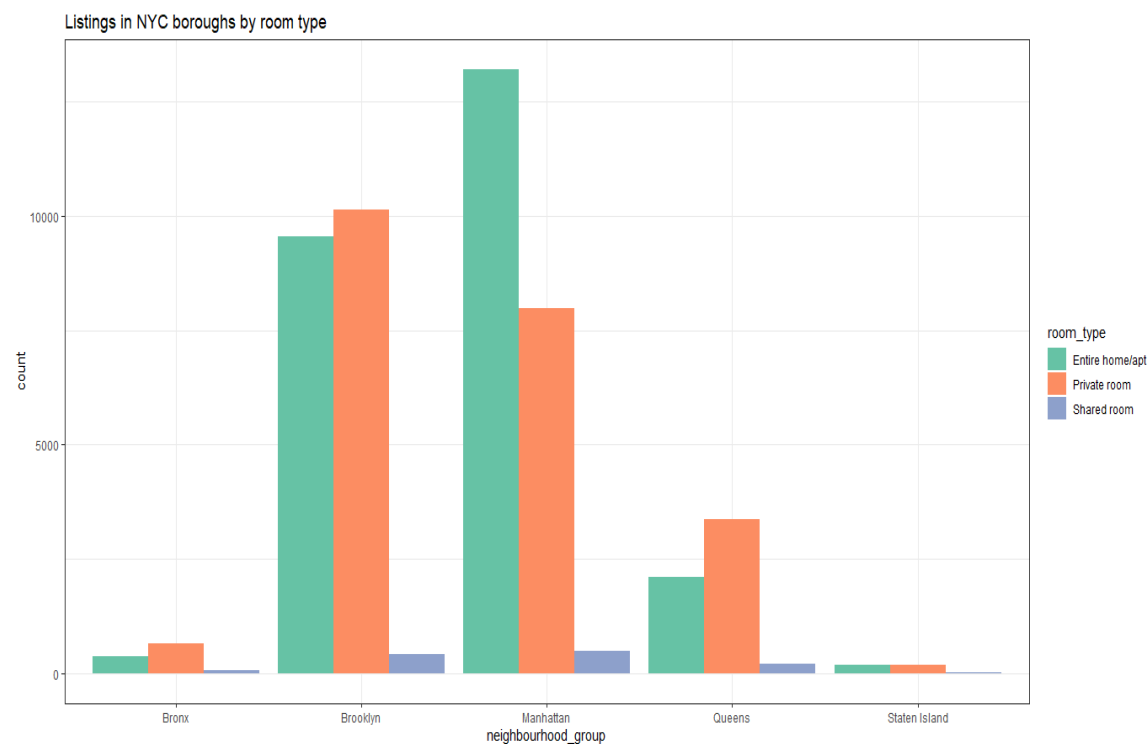
- Number_of_reviews
- Number_reviews_permonth
- Host-id
- Availability_365
- Calculating_host_listing_counts
- Minimum_nights
- Neighborhood
- room_type
- Are highly correlated



From the graph we can say that people are opting for the entire room even though price is high

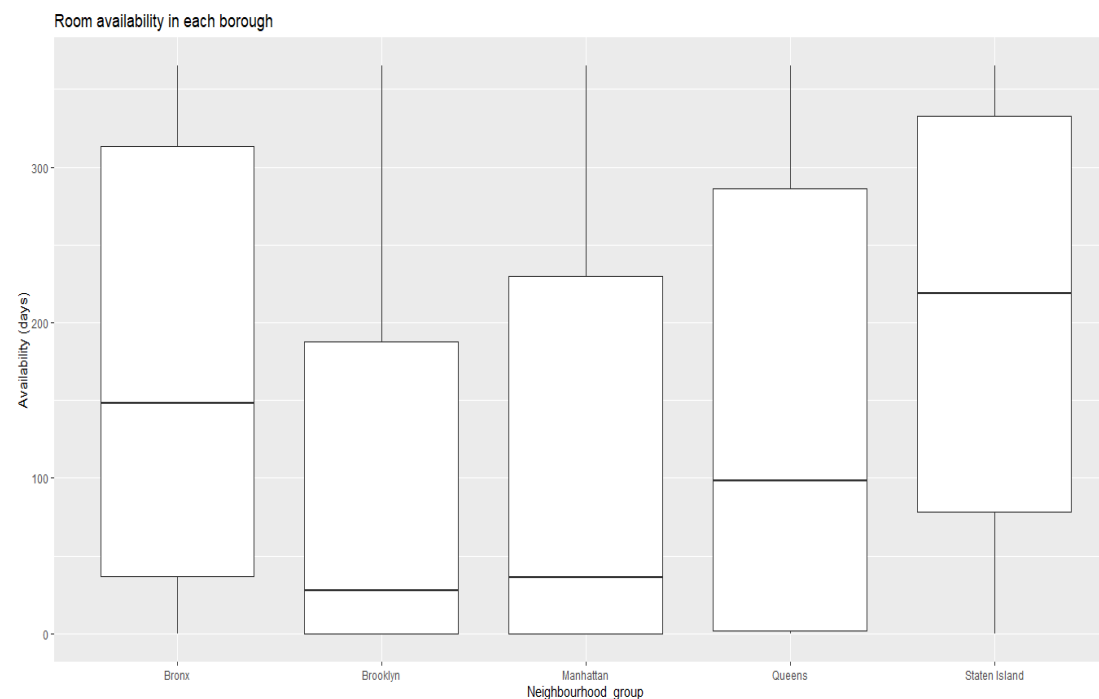
The graph say that from the mean price the Manhattan neighborhood group has more visits throughout the year even the mean price is high





Manhattan has more occupancy in entire and shared room and Staten island has very less occupancy for the all types of rooms.

From the box plot the room availability for the Staten island is more than other states due to faster reservation



CONCLUSION

- Location plays a significant role in determining the price of a listing, with certain neighborhoods and boroughs being more expensive than others.
- Property type is also an important factor, with entire homes/apartments being generally more expensive than private rooms or shared spaces.
- Other important predictors of price include the number of bedrooms and bathrooms, the availability of amenities such as air conditioning and internet, and the overall rating of the listing.
- Linear regression and random forest models can be used to accurately predict the price of an Airbnb listing, with the random forest model achieving an accuracy of 93.66% on the test data.
- Host-related variables, such as host response rate and the number of listings a host has, did not appear to be significant predictors of price in the models.
- Further analysis could be done to explore the relationships between variables in more depth and potentially identify additional predictors of price.