**CIS9660 Group Project**

**NYC Airbnb Price Prediction**

**Group Rent**

Oreofe Adenekan

Pengnan Chen

Brigitte Gonzalez

Kenny Li

**Introduction**

Since its inception in Aug 2008, Airbnb has provided a platform for short-term rental to millions of hosts around the world. As of Aug 31, 2015, there were 30,461 active listings on Airbnb in New York City. Among the thousands of listings, what are the factors that determine the price of a property? What is a fair price to the property you’re about to list or thinking about renting. At the end of our analysis, you will find answers to these questions which provide some insights to both the hosts on Airbnb and potential guests.

**EDA and Data Cleaning**

The dataset we used to perform the analysis below covers NYC Airbnb listings from 2008-06-26 to 2015-08-31. It can be found in Tableua.com <https://public.tableau.com/en-us/s/resources>

After some data exploration, we decided to exclude the following variables from the dataset. These variables are the name of the property, number of records, zip code, review rating bins, host id and host since. The analysis focuses on structured data, therefore unstructured data such as the name of the property is not on our list of variables. Number of records has the value of “1” throughout the dataset. Neighbourhood is the smallest unit we used for property location, as a result, zip code was removed. Review rating bins have a perfect positive correlation to review ratings and we only need one of these predictors. Lastly, host id is simply an identifier given to a host and it doesn’t help explain the property price. Other than the 5 columns above, we also removed duplicate listings from the dataset.

After the initial cleaning, we noticed a considerable number of missing values in the column review rating, and some null values under the number of beds. Rows with null values were deleted. Next, we removed the outliers in response variable Price, and shortened the list of property types to 5 levels. The cleaned dataset consists of 21,979 rows and 7 columns.

*Figure 1* in the appendix is the pairwise scatterplot for all 3 quantitative variables. Other than the number of reviews, price doesn’t have a clear relationship with the number of beds and review ratings. *Figure 2-4* demonstrates the price distribution based on where the property is located at, room types and property type. We can tell that the majority of the rental properties are located in Brooklyn or Manhattan. The average and maximum price of Manhattan is higher than Brooklyn. Apartment is the most popular property on Airbnb. And as we would expect, the average price to rent the entire home is substantially higher than the mean price to pay for a private room.

**Models**

1. **Regression Tree**

The first model we implemented to predict the price is regression tree. We used the resampling method of 80-20-holdout with 80% of the original data in the training set, and the remaining 20% for testing.

The regression tree model contains all 7 predictors. We then performed a 10-fold cross validation to find the optimal size of the tree which turned out to be 6. The result of the pruned tree is in *figure 6*. The plot depicts the important factors in determining the price. Among these factors, room type is the most important. The highest price for a private or shared room is $101.50 which is $44 less than the cheapest entire home rental. Property location split the data into 4 branches. In Manhattan, to rent an apartment with less than 1.5 beds costs $177.30. Spending $30 less on a unit outside Manhattan most likely will get you 2 beds.

We also calculated the mean squared error for the predicted price using the regression tree. The result will be evaluated and compared to it of different models under the section model evaluation.

**MSE=2409.353**

1. **Random Forest**

The second model we implemented is the Random Forest model. To build this model, we used our previous test data, training data and the optimal number=6 (that we got when calculating the regression tree. To run our random forest, we used 3 predictors to construct a random forest model.(*total predictor number is 7, 7/3 =3).* After we tested the MSE of this model by comparing the predicted values with the true values, we decided to run the importance() function which calculates the mean decrease of accuracy in predictions for each predictor variable (*figure 7)*. In other words, *figure 7*  ranks the usefulness of the predictor variables. Making **Room\_Type** the most useful/significant variables for this model

**MSE=2287.082**

1. **Linear Regression**Lastly, we will implement our linear models. After we performed EDA on our dataset, we plotted the correlation for all variables. We were not able to visually choose significant predators for **Price,** as you see in *figure 1.* 
   1. Our first linear regression model (LModel1) has all variables as predictors where 4/14 variables were not significant (all factors).
   2. Our second linear regression model (LModel2) has the top 4 predictor variables from our Random Forest Model (Room\_Type, Neighbourhood, Beds, and Review\_Score\_Rating)
   3. For our last and third linear regression model (LModel3), we decided to exclude the insignificant predictor variables from our LMModel1 and compare it against LModel2

**LMModel1 MSE=2356.286**

**LModel2 MSE=2317.272**

**LModel3 MSE=2327.079**

**Models Evaluation**

**Conclusion**

**Appendix**

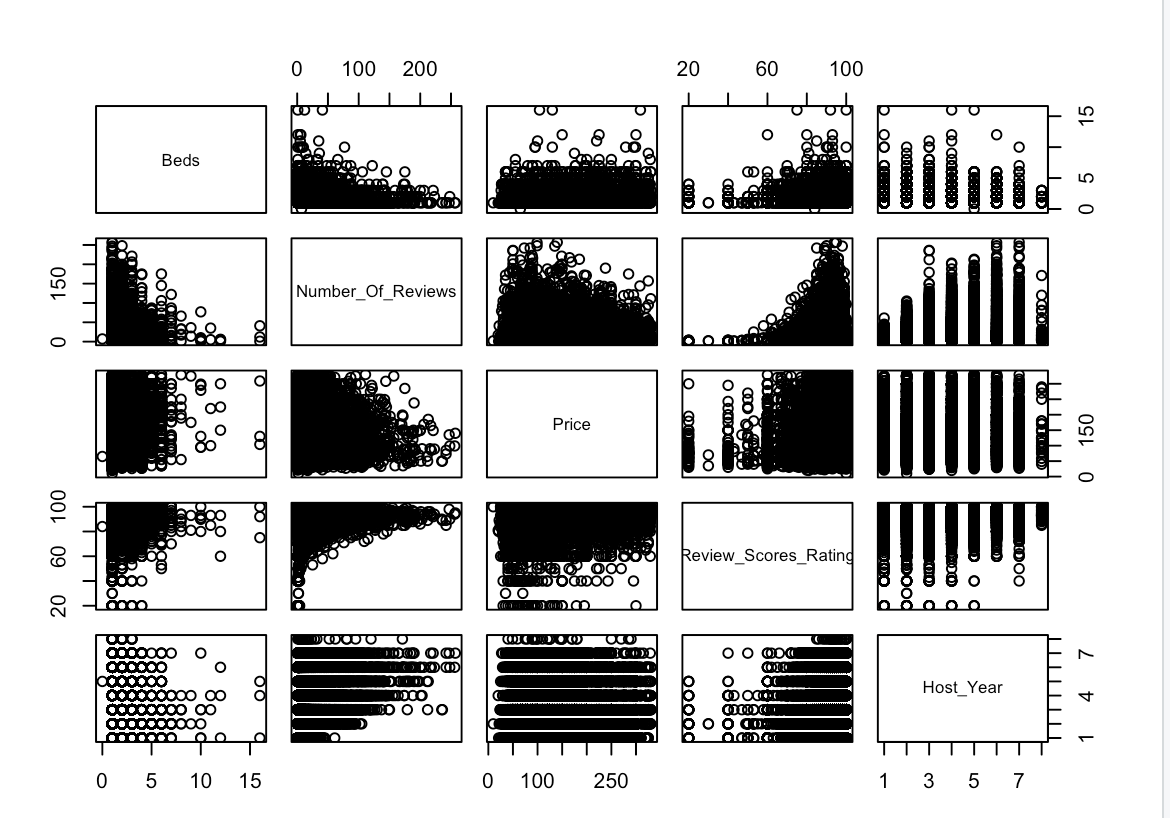


Figure1. Pair() Function for correlation Analaysis



Figure 2. Boxplot of Property Type vs. Price (to find possible outliers)

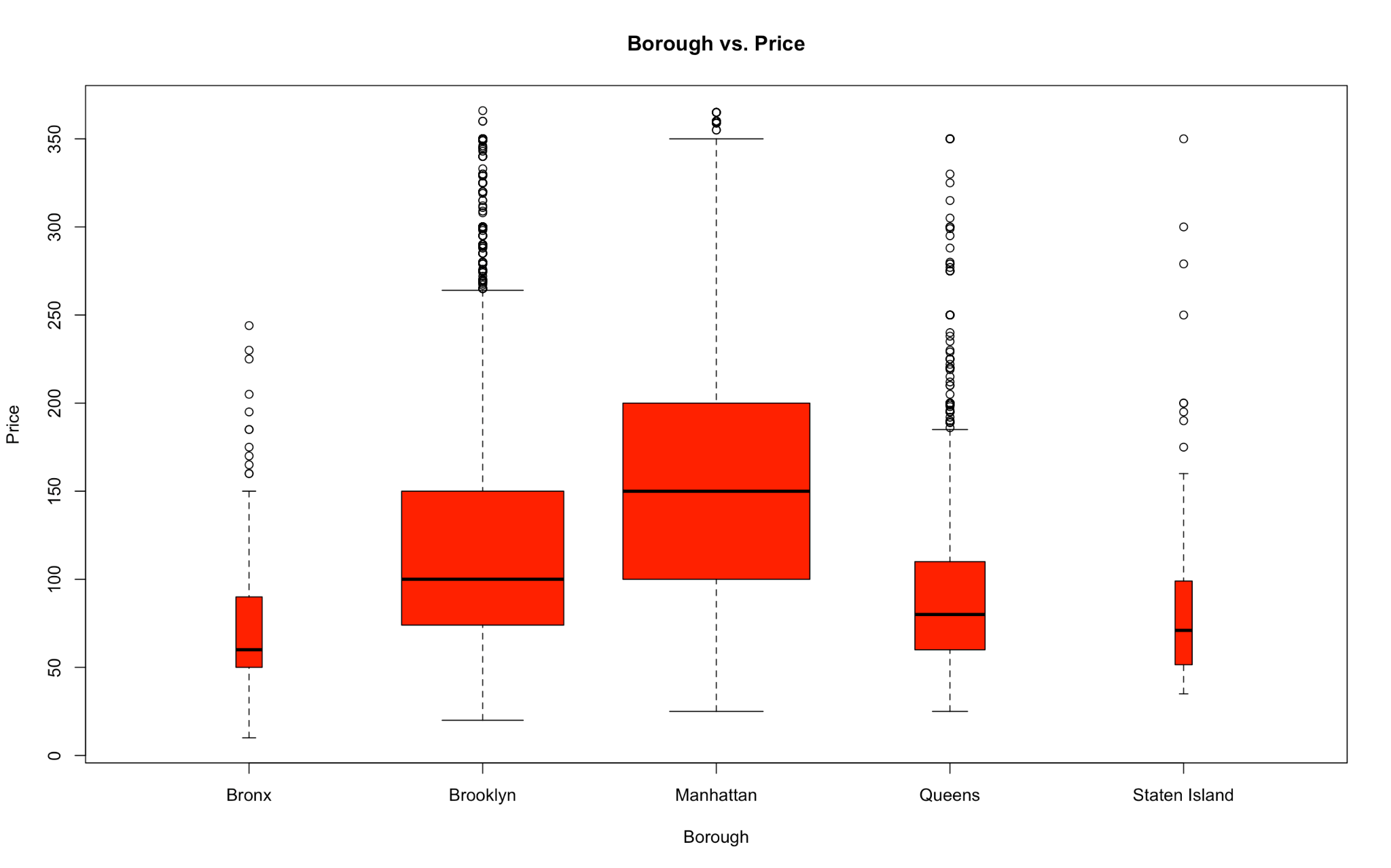


Figure 3. Boxplot of Borough/Neighbourhood vs. Price (to find possible outliers)

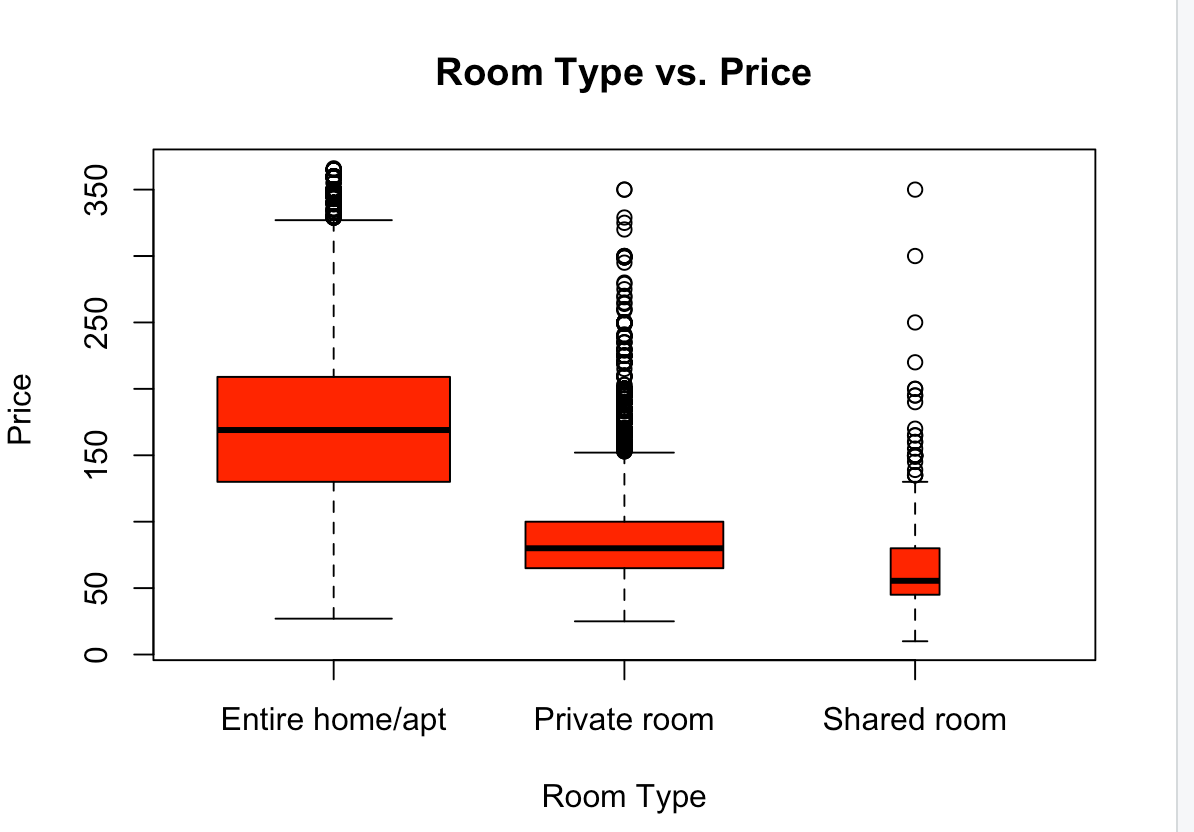


Figure 4. Boxplot of Room Type vs. Price (to find possible outliers)

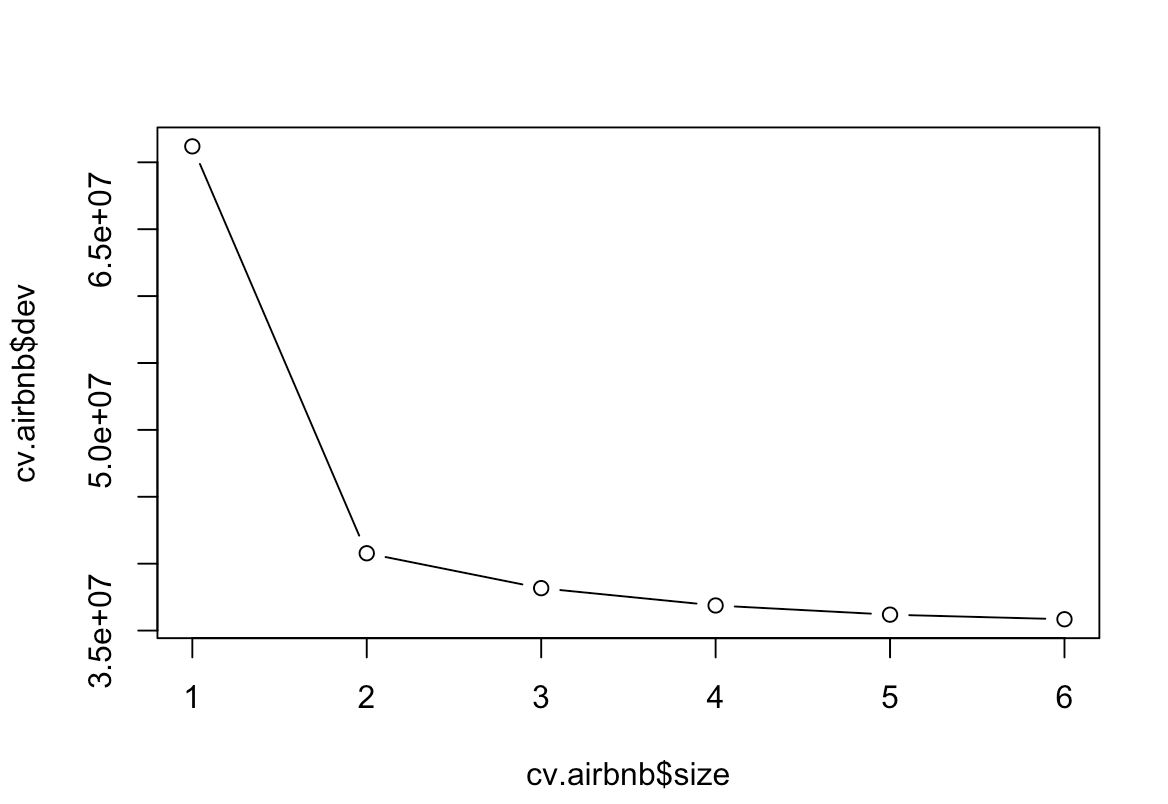


Figure 5. plot to find the optimal size of of tree

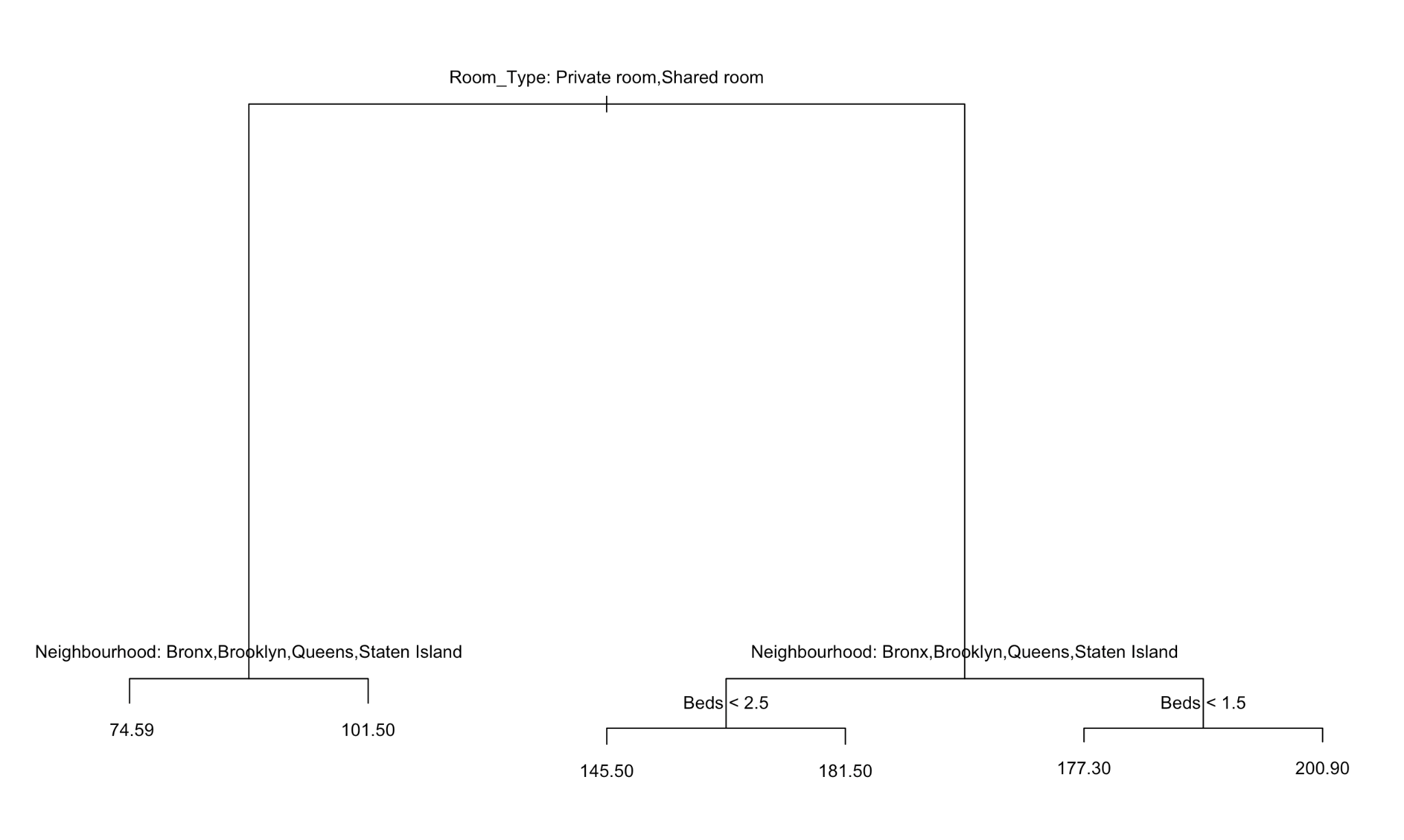


Figure 6. Tree Model

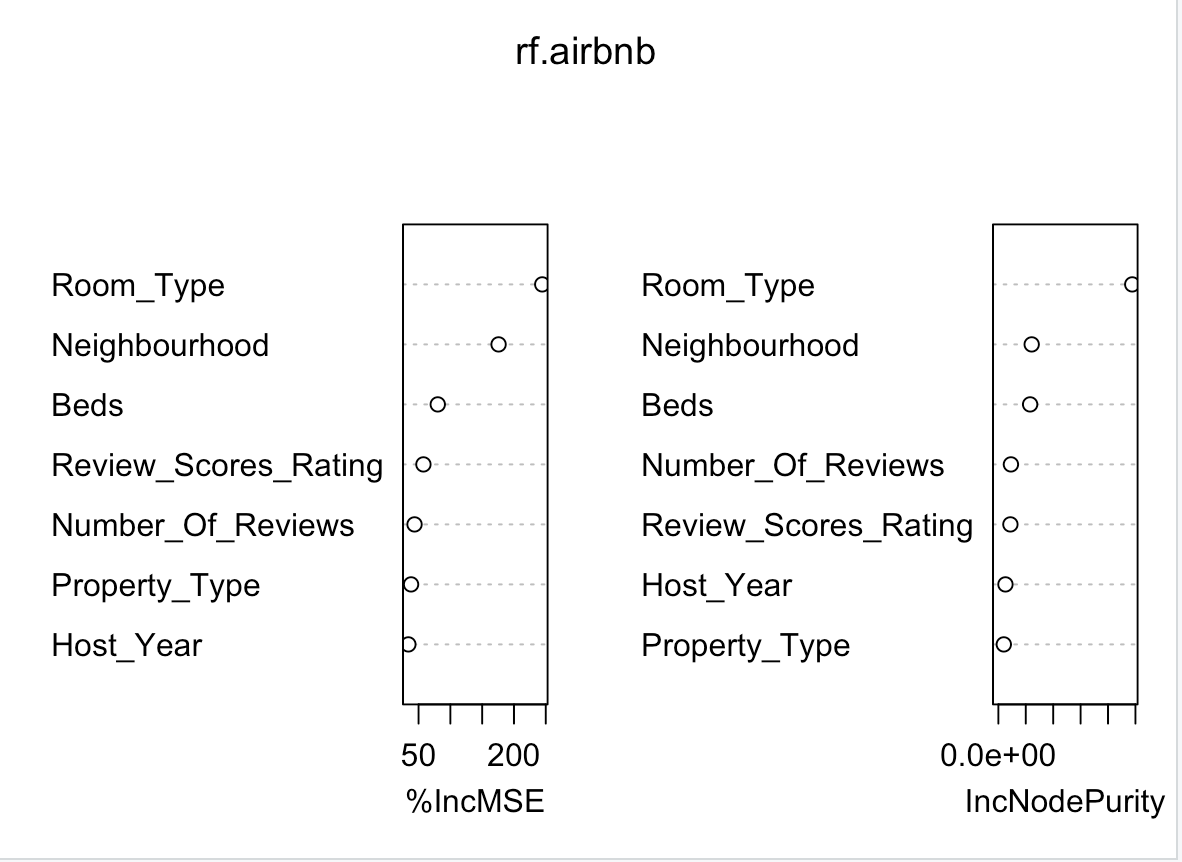


Figure 7. Importance() for Random Forest Model

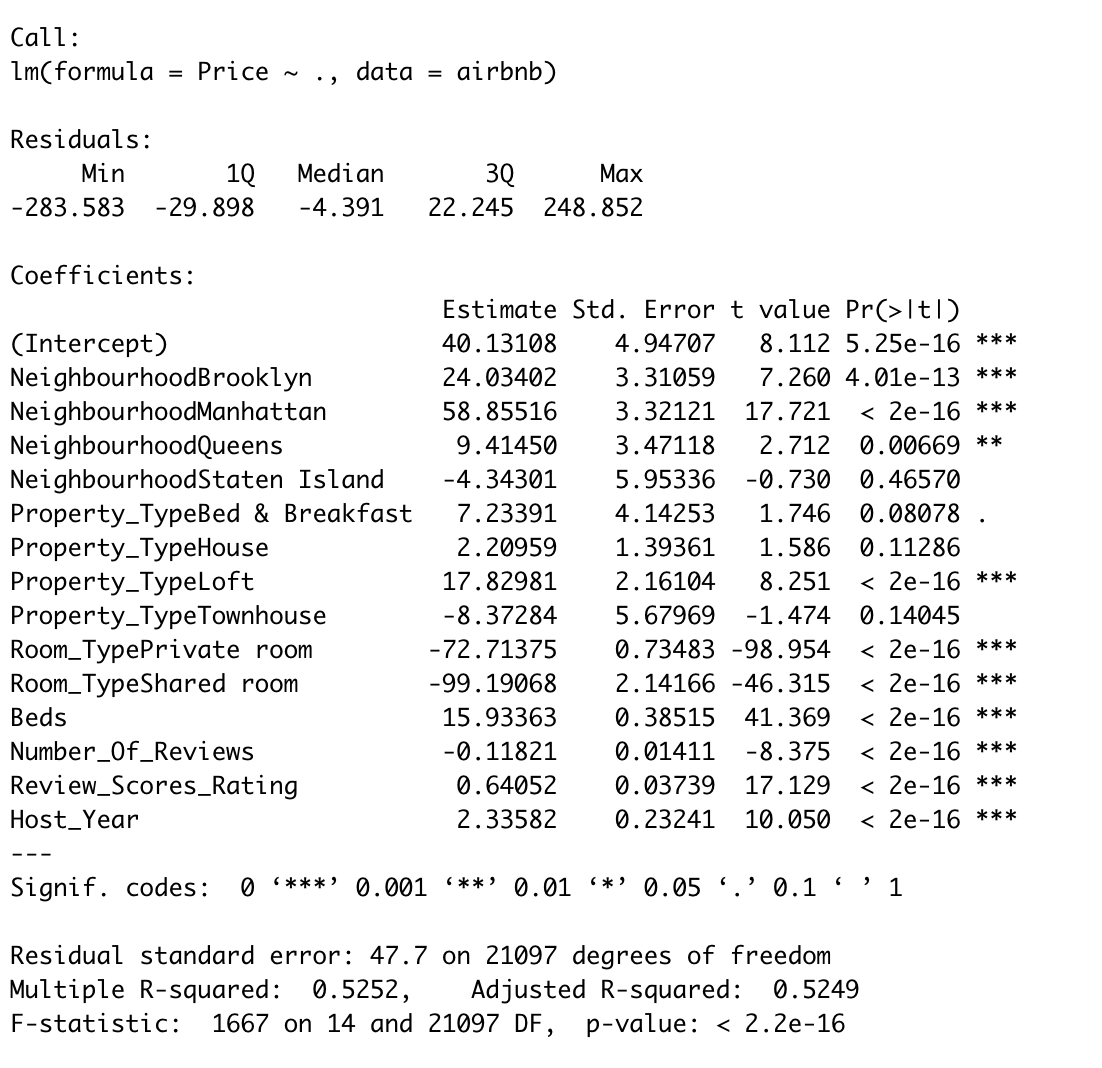
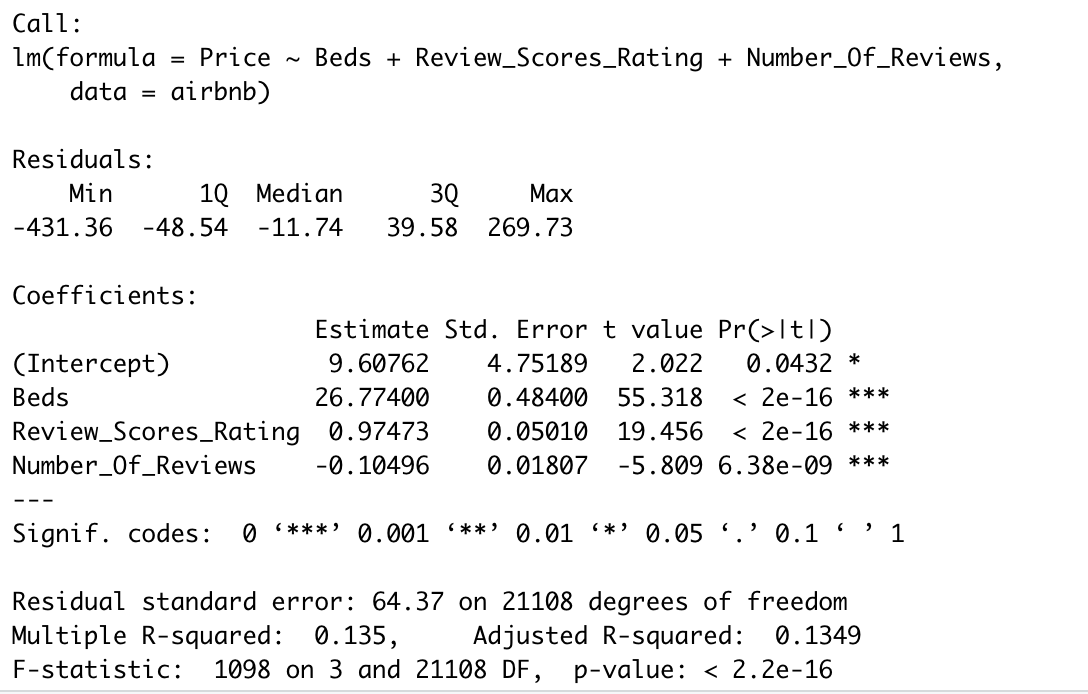


Figure 7. Linear Model (LModel1) Summary



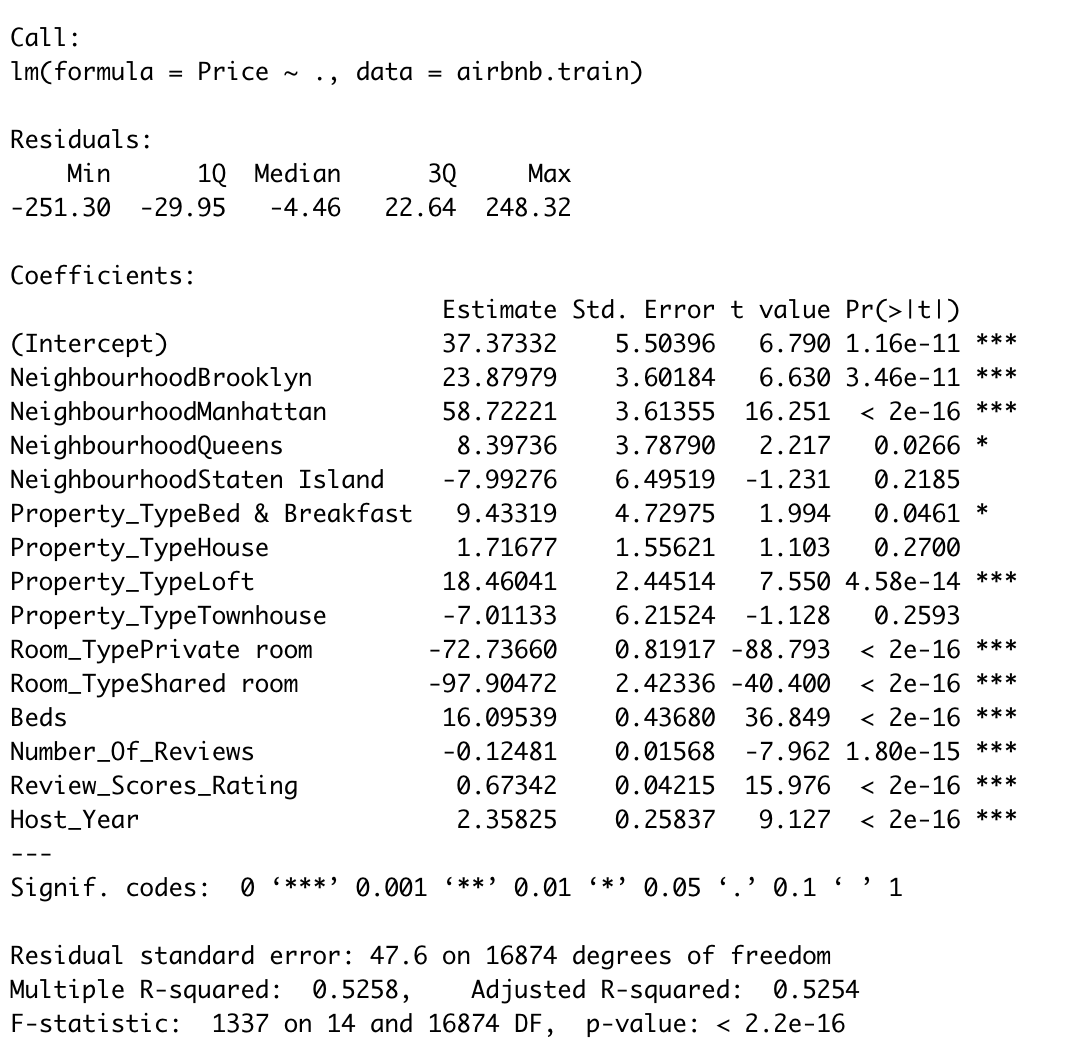


Figure 8. Linear Model (LModel2) Summary

