Analysis of Prices Movement of

*Airbnb* Listings in New York City

Subal Bhattarai, Prabal Chhatkuli, [Stephanie Waterhouse](mailto:swaterho@ramapo.edu)

December 12, 2020

**1. Executive Summary**

Airbnb is an online marketplace that provides accommodations by acting as an intermediary between hosts who want to rent out their homes and the customers who want to stay. Airbnb spans over 100,000 cities across 220 countries, and there are 2.9 million hosts with over 7 million listings. The data is collected from an independent, non-commercial website launched in 2016 called Inside Airbnb. The website reports and visualizes scraped data on the property rental marketplace called Airbnb focusing on highlighting illegal renting on the site. By analyzing publicly available information about a city’s Airbnb listings, Inside Airbnb provides filters and key metrics so we can predict the price of an Airbnb listing and how various variables affect the price. Our target city is New York for this project and the last date that the data was compiled was on October 5, 2020. The dataset for New York City can be accessed under:

<http://insideairbnb.com/get-the-data.html>. In this paper, we observe how several factors like neighbourhood, type of room, minimum night policy, reviews per month, and availability of the Airbnb listing affect the price. To capture some of these ideas, we make extensive use of exploratory data analysis and finally various machine learning approaches like linear regression, ridge regression, principal component regression, partial least squares, boosting, and random forest. We observe that Manhattan has the most number of airbnb listings and on average, is the most expensive neighbourhood for Airbnb listings. Contrary to that, Staten Island has the least number of listings and is the cheapest. Based on this information, a host can maximize his profit by renting an Airbnb in Manhattan and a consumer can save his money by renting an Airbnb in Staten Island. However, there are additional factors like availability, and minimum number of nights policy that should also be taken into consideration. Another important thing to note is that we did not find any specific relationship between price and number of reviews. The data does not distinguish between positive and negative sentiment of the reviews, however, we assume that it is positive based on our exploratory data analysis. In general, private rooms and listings with less minimum-nights-to-stay policy tend to have higher number of reviews. This information suggests that potential Airbnb hosts should rent their listing as private rooms and remove the constraint on renters to book the Airbnb for a certain period of time if their goal is to maximize profit. In our attempt to predict prices of Airbnb listings, we use various machine learning approaches and judge a model based on the total mean squared error(MSE). To achieve this, we split our data into training and test in the ratio 7:3 and observe that Random Forest technique gives us the least MSE, and hence, is the superior model.

**2. Data Set and Variables**

Our data initially consisted of 46 columns and 47,840 rows for each listing of Airbnbs. For our project, we only selected a few relevant variables that could potentially affect our target price, namely price.

**2.1 Variables in the data**

There are 9 variables in our dataset, including our target variable, price. “host\_name”, “neighbourhood\_group”, and “room\_type” are categorical variables and the rest are numeric continuous variables. Below are the variables with their respective definitions:

* **host\_name** - Name of Host
* **neighbourhood\_group** - Borough that contains listing
* **latitude** - latitude of listing
* **longitude** - longitude of listing
* **room\_type** - Type of public space that is being offered
* **price** - price per night, USD
* **minimum\_nights** - minimum number of nights required to book listing
* **reviews\_per\_month** - total number of reviews divided by the number of months the listing is active
* **availability\_365** - number of days per year the listing is active

**2.2 Missing data**

Checking the missing data, we observed that there were 10,052 missing data for “reviews\_per\_month” column. Apparently, most of the Airbnb renters do not provide reviews of Airbnb during their stay. This is completely fine and we changed all the NA values to 0, meaning that that Airbnb listing did not have any reviews.

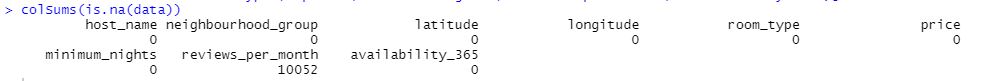


Fig 1. Missing data

**2.2 Anomalies in the data**

We observed some big outliers for our target variable, price, in the dataset. Some of the listings were priced at $0, and some were outlandishly expensive. We should eliminate all the listings priced at $0, but it was completely possible that high priced Airbnbs do exist. In order to confirm our skeptical data, we looked for expensive listings for New York City on the Airbnb website and observed that listings above $5000 were almost non-existent. Also, we plotted the latitude and longitude position for the listings from our dataset and confirmed that such high priced listings do not exist. In our project, we only considered Airbnb listings that are below $500 primarily because listings with high prices were mostly faulty data and secondarily, to keep our analysis budget-friendly.

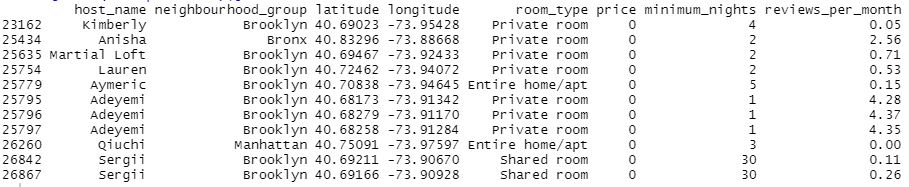


Fig 2.. Listings priced at $0

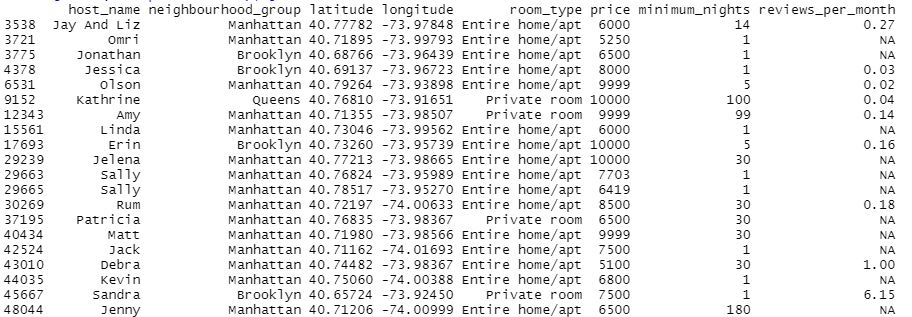
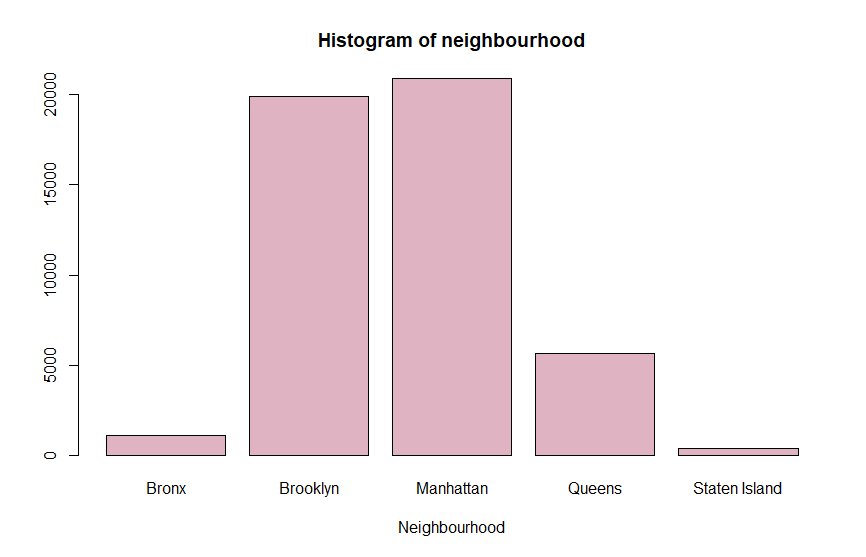
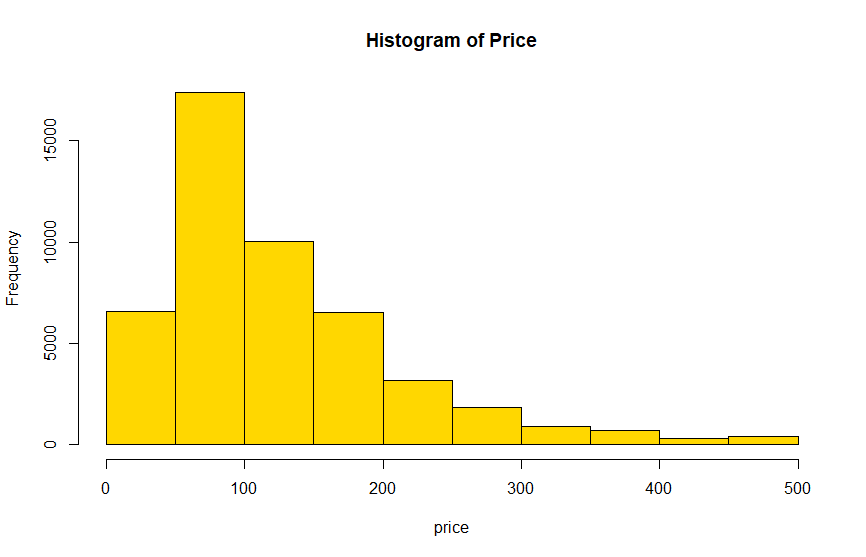
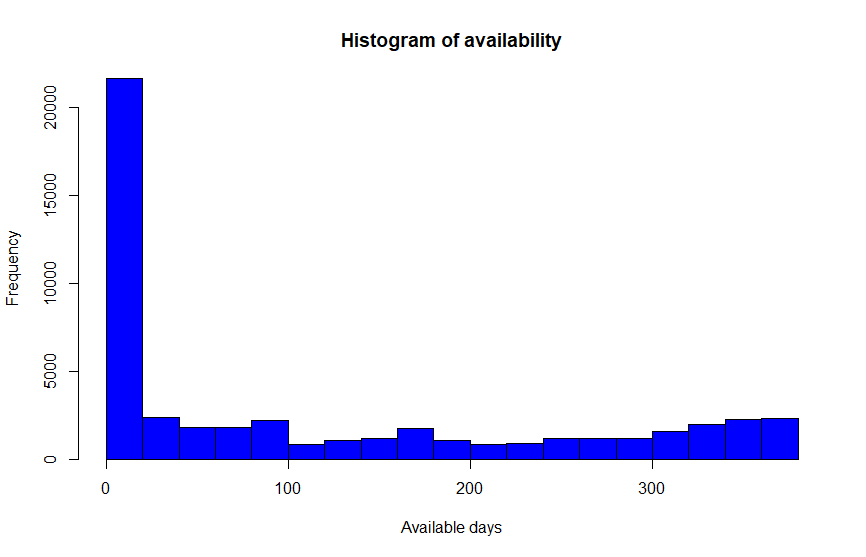
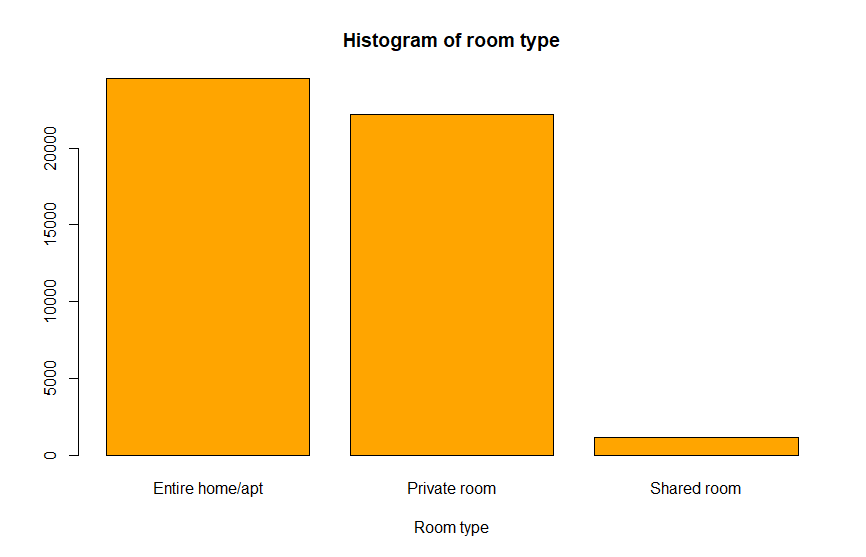
****

Fig 3. Listings above $5000

**2.3 Exploratory data analysis**

We look at the histograms of various variables, including price, to gain some insights on the spread of data across factors affecting price and the general trend in Airbnb listings.



****

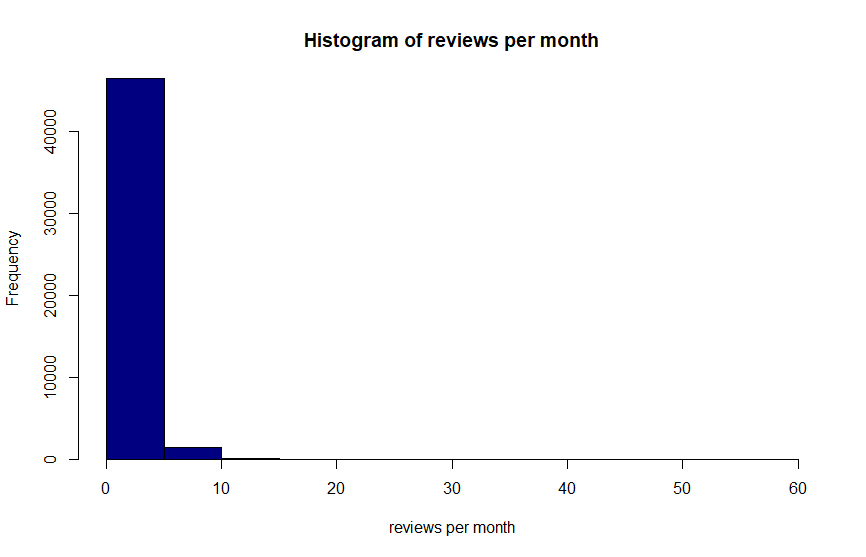
****

Fig 4. Histogram of price, neighbourhood, room type, availability and reviews per month

From Figure 4, we can draw some important conclusions:

* Most of the Airbnb listings are priced between $50 to $100 per night.
* Manhattan has the highest number of Airbnb listings and Staten Island has the lowest.
* Most of the Airbnb listings consider renting their entire home/apartment. Only a few listings have a “shared room” policy.
* Surprisingly, most of the listings (almost 40%) were available below 10 days in a year.
* 96% of the Airbnb listings had below 10 reviews in a month.

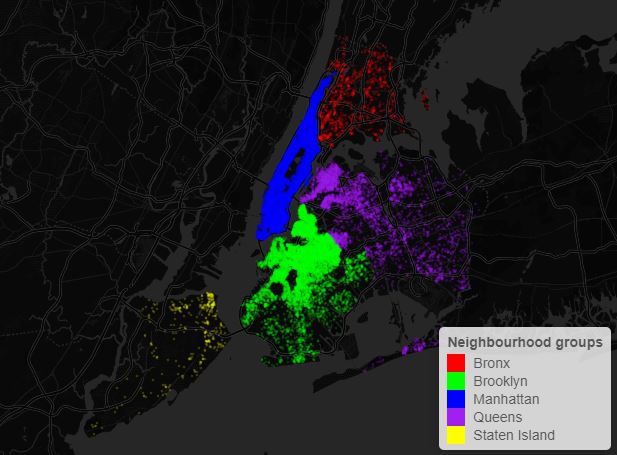
An important thing to observe would be the features of an Airbnb listing that has a high number of reviews. We consider 15 as the threshold and observe some trends in highly reviewed Airbnbs:



Fig 5. Listings with more than 15 reviews per month

* Private rooms tend to have higher number of reviews per month
* Queens neighbourhood has a higher number of reviews compared to Manhattan and other neighbourhood boroughs of NYC.
* Airbnb listings with less minimum-nights-to-stay policy tend to have higher number of reviews per month.

Finally, Figure 5 shows the concentration of Airbnb listings in New York City:



**Fig 5. Airbnbs in New York City based on neighborhood**

As expected, Manhattan is the most densely populated and Staten Island is sparsely populated.

**3 Methods:**

The following machine learning approaches were undertaken for this study:

1. **Multiple Linear Regression**

* This method uses a linear approach to model the relationship between different predictors to predict the value of a certain response variable. It is based on ordinary least-squares (OLS) regression that involves more than one predictor variable. It is very helpful when there are multiple factors that affect the outcome of an event or a response.

1. **Ridge Regression**

* Ridge Regression is a modelling technique useful for data that have multicollinearity. All Ridge regression variables are based on standardized variables. The final result is in the original scale but the ridge trace is in a standardized scale. This method provides improved efficiency in parameter estimation problems in exchange for an increase in bias.

1. **Lasso**

* This variable performs both variable selection and regularization to provide better accuracy in variable predictions. Hence, it helps in feature selection and reducing overfitting of the model. Like the other regression methods it also minimizes the usual sum of squared errors.

1. **PCR**

* It is a regression analysis technique that is based on principal component analysis (PCA). PCR is used for estimating the unknown regression coefficients in a standard linear regression model. By adding a degree of bias to the regression estimates, PCR reduces the standard errors.

1. **PLS**

* PLS reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components, instead of on the original data.

1. **Boosting**

* It produces a prediction model in the form of a collection of shallow decision trees. Boosting reduces error mainly by reducing bias in the prediction.

1. **Random Forest**

* Unlike Boosting Methods, it is based on fully grown decision trees and it works by reducing the variance in the predicted model.

**4. Data Processing:**

First step in the data processing was to normalize the data, so a logarithm was applied to the “price” predictor.

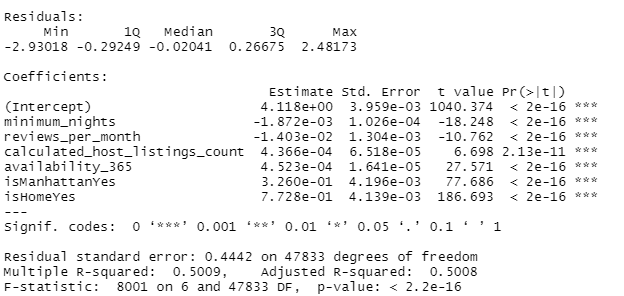
Two of the predictors in the initial dataset were processed into two different predictors. The data had many categorical data, for example, the neighborhood\_category predictor contained 5 different boroughs of NYC in string format. According to our initial EDA, it was seed that most of the data was cluttered in Manhattan, so the neighborhood\_group was changed to a isManhattan predictor which would be 0 if true 1 if false. Similarly, the room\_type predictor was changed to isHome predictor and the values were set to be factors 0 if true and 1 if false.

**5. Results:**

The following results were obtained with various machine learning approaches applied:

1. **Multiple Linear regression**

* The result of the Multiple Linear regression were as follows:



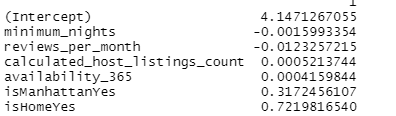
The Mean Squared Error(M.S.E) we received from this was: 0.1972572

1. **Ridge Regression**

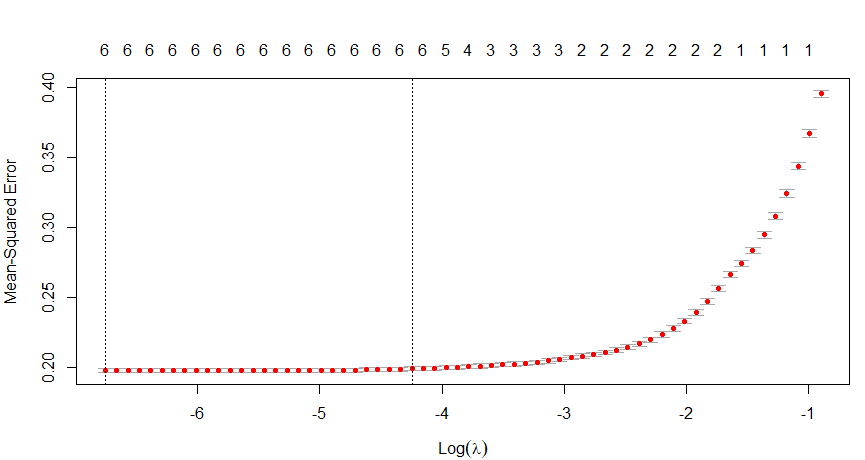
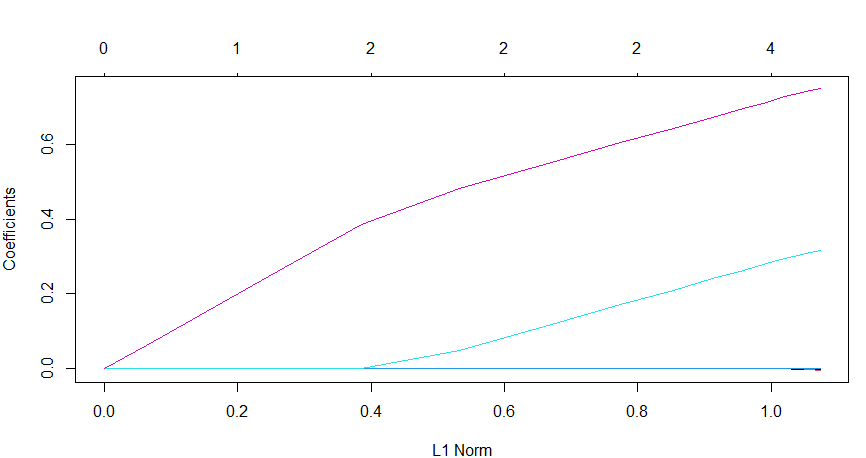
* For Ridge Regression, we receive the following:

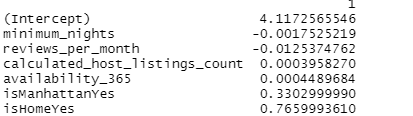
Best Lambda: 0.04080627

MSE: 0.1968215

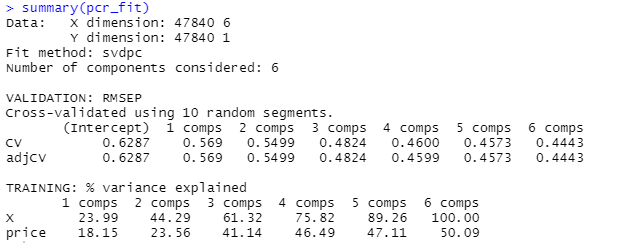
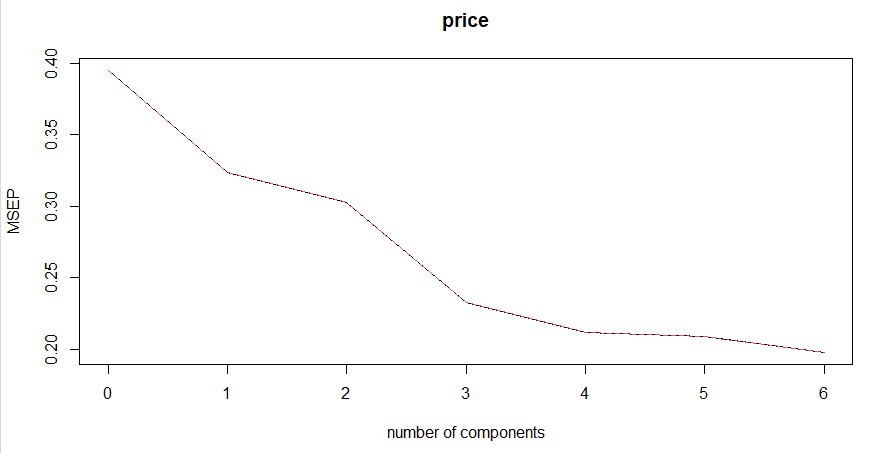
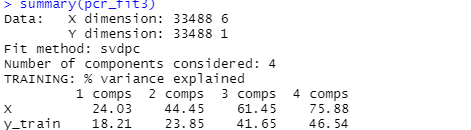


1. **LASSO**

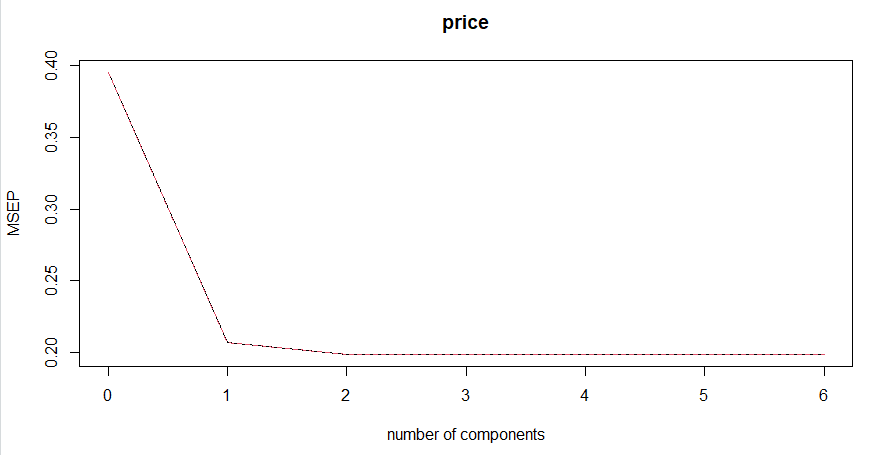
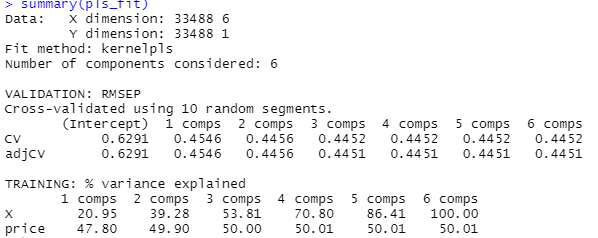
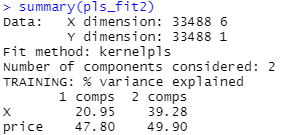
* 
* The Best Lambda was: 0.001162177
* The MSE we received was 0.1966722
* The coefficient of the lasso are:



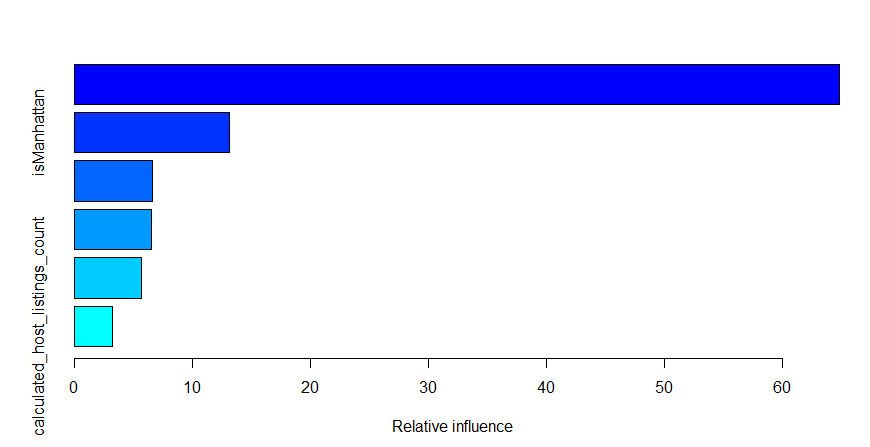
1. **PCR**

* 6 components were taken for the PCR and the PLS.
* 
* The validation plot of the pcr showed:
* 
* MSE for PCR was: 0.1966722
* Finally, the following results were obtained for the test data:
* 

1. **PLS**

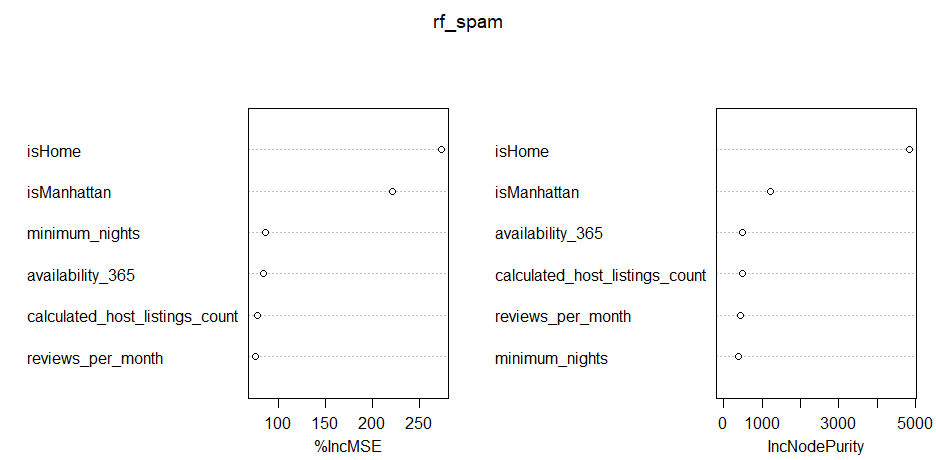
* The PLS gave the following validation plot:
* 
* The following summary was obtained from the training.
* 
* An MSE of 0.2103112 was obtained from this model.
* The testing data gave the following results for the test data
* 

1. **Boosting**

* Boosting provided the following influence plot for the model:
* 
* An MSE of 0.1857905 was found.

1. **Random Forest**

Random Forest Method provided the following results:

* MSE obtained is 0.1766407
* The following results were obtained from the influence plot:
* 

**6. Conclusion:**

Overall, this study provided a general outlook on various factors that affect the price of an Airbnb listing in New York City. Random Forest Method provided the lowest MSE for the test dataset. From the importance plot, we observed that the most important predictors are whether the listing is a home or not, and whether it is in Manhattan. From the linear regression, lasso, and ridge regression, it could be inferred that more number of reviews and the minimum number of days of listing were involved with listings with low prices.