# **ABSTRACT**

Airbnb is an online platform that allows homeowners to rent out their extra space to travelers looking for a place to stay. The "hosts" are able to set their own prices at will. Airbnb data from Kaggle was used to see if nearby listings have an impact on Airbnb rental prices. To model the rental pricing, exploratory data analysis was done to pick useful features. A model was created using the multiple regression method to forecast listing prices based on relevant factors. After creating an initial model with all of the variables that were deemed to be significant, models were built using four different variable selection approaches: stepwise regression methods (using AIC and BIC both), regression subset method, and Lasso regression methods. The MSE, R2, and Adj R2 of these 5 models were then compared in order to determine which model was the best. The resultant model was then subjected to in-sample validation, out-of-sample prediction, and cross-validation procedures. Validation revealed that the MSE of the Test Dataset, 1650.6, is nearly identical to the MSE of the Full Data, 1645.83. As a result, it was determined that the model used for study accurately predicts the variable.

# **CHAPTER ONE**

## **INTRODUCTION**

Airbnb has increasingly increased in popularity and is now the most popular online service for securing short-term lodging. Whether you're a tourist, visiting friends and family, or looking for a place to stay on your trip(Alin, 2020).This new rental platform innovation has wrought havoc on the hospitality business; nonetheless, this sector has thrived in city centers in particular (Kunwar, 2020). Airbnb is a privately owned rental website that provides a peer-to-peer platform for users to rent rooms, flats, apartments, villas, and other temporary lodgings at a variety of pricing in countries all over the world.

People enjoy how efficient and cost effective the service is. However, not all Airbnb properties are economical, in terms of location, design, accommodating capacity, pricing, and facilities, there is a wide range of options. One of the most important factors that impacts a client's choice of hotel is price. Previous study has looked into a variety of elements that influence hotel prices, establishing a set of pricing drivers such as brand name, star rating, location, hotel age, room count, and amenities, among others(Zhang *et al.*, 2017).

The goal of this study is to see if surrounding Airbnb listings have an impact on a property's rental pricing in New York City. Because the city is the most visited in the United States, pricing a rental property on Airbnb is challenging for the owner because the number of clients on the premises is fixed. Customers, on the other hand, must evaluate the given price with only a rudimentary understanding of the property's true worth. In order to assist both the landlord and the consumer in estimating prices. Predictors will be included in owner property and customer reviews, and regression techniques will be used, as well as variable and model selection approaches such as sub-set regression and step-wise regression using AIC/BIC.

In comparison with conventional hotels, Airbnb standards are quite unique in terms of pricing systems. When it comes to setting rates that are fairly high while not losing popularity, both new and established hosts of updated lists experience obstacles(Washington, 2019). Even though consumers can compare prices across relevant listings, it is still necessary for them to know whether the present pricing is reasonable and whether it is a good time to reserve the rooms.

Furthermore, In order to define the dataset, Airbnb rentals are subjected to numerous criteria each night, dividing the data into four groups. From the dataset, 16 features have been retrieved for further analysis and prediction.

For future jobs, only a few properties have been categorized and categorized. locality, latitude, longitude, neighborhood group, reviews room type private room, shared room, complete room/apt locale, latitude, longitude, neighborhood group number of reviews, most recent review, number of reviews per month, estimated host listings, availability.

## **CHAPTER TWO**

## **LITERATURE REVIEW**

(Alin, 2020) in his research, developed a model that can accurately forecast nightly Airbnb room and property costs, as well as other factors. He also had a secondary purpose of experimenting with new features, such as review comments, text-sentiment, and listing photography quality rankings, as well as new techniques to training data utilizing sets of numerous cities rather than individual cities than the data from a single market. Random Forests and XG-Boost were used as models. The outcome revealed that XG-Boost proven to be the best regression approach, ranking the highest in several tests techniques, with an R2 value of 0.64 as the best result.

The impact of Airbnb on Beijing's neighborhood housing prices was analyzed by (Li and Biljecki, 2019) using machine learning and geographic information systems (GIS). The relationship between Airbnb activity in a neighborhood and historical housing prices were analyzed using machine learning models. The best-fit model is then used to investigate neighborhood price sensitivity in light of rising Airbnb activity. The results reveal that price sensitivity varies by neighborhood, with some likely to be more price sensitive to Airbnb activity and others likely to be more price robust.

The study done by (Quattrone *et al.*, 2018) which investigated the spatial distribution of Airbnb in eight U.S. cities in connection to topographical, socio-demographic, and economic data. Regardless of their differences in demography, size, and wealth, all eight cities showed the same trend, indicating that regions with a strong Airbnb presence are those occupied by the "skilled and creative" classes, as well as those near city centers. Forecast of Airbnb’s spatial penetration showed an accuracy of 0.725.

## **CHAPTER THREE**

## **METHODOLOGY AND DATA**

### **3.1 Dataset Description**

Kaggle was used to collect the data. Kaggle is a Google LLC-owned data science and machine learning online community that allows users to find and submit data sets, as well as analyze and construct models in a web-based data science environment. Exploratory data analysis, visualization, and modeling of New York Airbnb data were carried out as part of this research effort.

Airbnb, Inc. is a website that allows anyone to book or sell lodging, mostly homestays and tourism experiences. Airbnb earns money from both visitors and hosts: hosts are paid 3% of the booking value, while guests are charged between 6% and 12% depending on the type of the booking. The Airbnb market in New York City (NYC) is booming, with over 48,000 listings as of December.

Airbnb listings in New York City is utilized to conduct an in-depth study and answer the research question "Do nearby Airbnb listings affect a property's rental price in New York City?" which contain about 48,895 observations and 16 variables containing continuous variables, numeric variables and characters.

The column variables include id,name, host\_name, neighbourhood\_group, latitude, longitude, minimum\_nights,room\_type, price, minimum\_night,number of reviews, last review, review per month, availablity\_365,calculated\_host\_listings\_count.

**3.2 METHODOLOGY**

### **3.2.1 Variable Selection**

The selection of variables is a critical stage in ensuring the model's accuracy, make predictions about the response variable. R-squared and RMSE are useful methods for assessing data quality. Adding a predictor to a model, on the other hand, will not make these values worse. backward and forward. Backward selection was used to avoid goodness-of-fit in order to create a smaller model that would still fit, well-fitting which end up fitting noise that has nothing to do with the response variable when fitting a larger model(Cassotti and Grisoni, 2012). Backward selection entails fitting the model with all available predictors.

This operation was completed using R's step function. At each stage, it assesses how removing a predictor will affect the outcome variable. The AIC, a measure of model quality, is calculated for each variable throughout this time. Each phase removes the variable with the lowest AIC one by one until no variable improves AIC. Forward selection is the polar opposite of reverse selection(KNBS, 2021).

The starting model has no predictors. R adds a variable to the model at each step until it reaches the maximum number of variables or finds a good model. Any variables utilized in one of these two selection processes were kept.

### **3.2.2 Modeling**

On the predictors of the variable selection results, machine learning models were applied.  These algorithms will forecast the output variable, the daily price of an Airbnb, based on the data and certain assumptions. The dataset was sampled and divided into train and testing data in a 70/30 split.

These samples were put through their paces in a Multiple Linear Regression model to see how well they performed in terms of R2 and root mean square error (RMSE)

### **3.2.3 Multiple Linear Regression**

Multiple regression, often known as multiple linear regression, is a statistical method for predicting the outcome of a response variable based on the use of two or more explanatory variables. This is a more complicated sort of regression than linear regression, which provides linear mathematical correlations between one independent variable and one dependent variable. Multiple linear regression is used to represent the linear connection between the explanatory variables and a single response variable. There can only be one response variable and two or more explanatory variables. The explanatory variables are the independent variables, whereas the response variable is the dependent variable.

Here is the formula for multiple regression:

Y = ß0 + ß1x1 + ß2x2 + ... + ßpxp

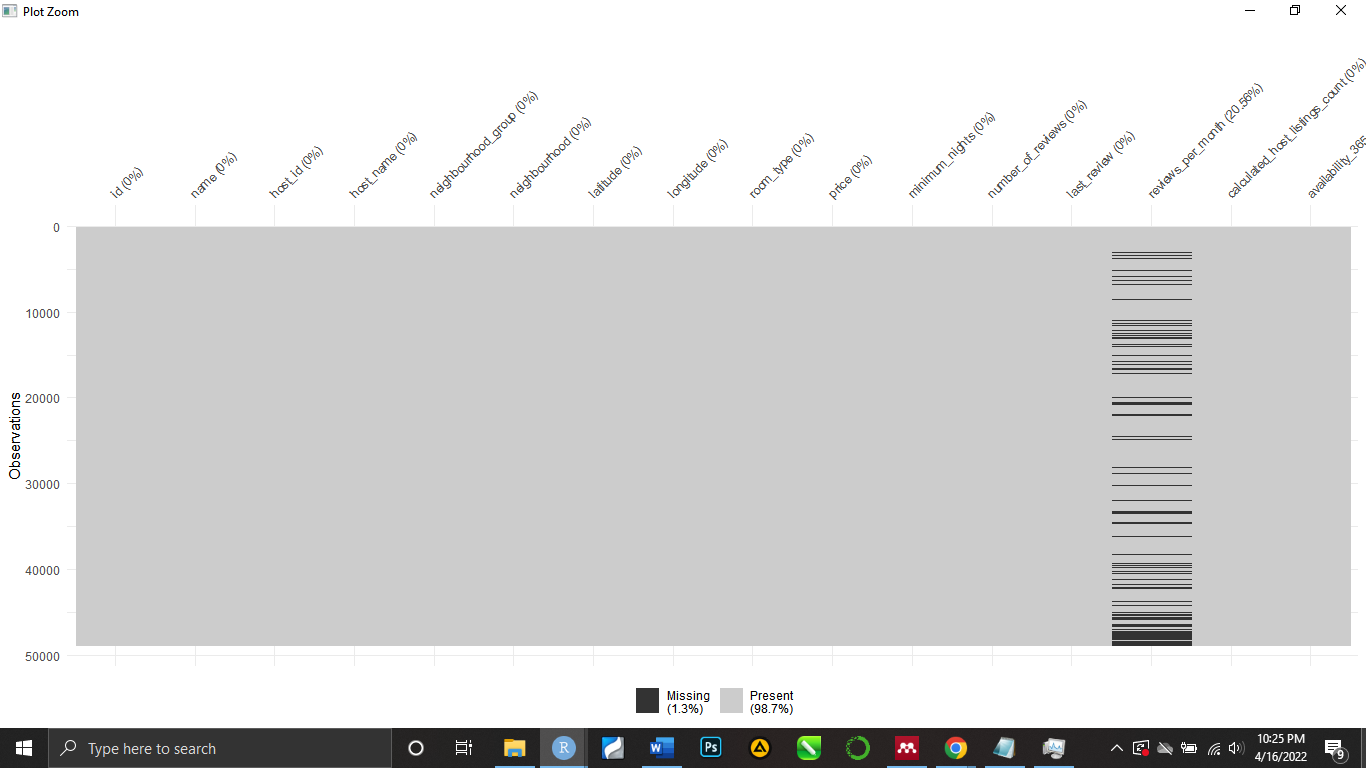
The variables in this equation are as follows:

The projected or expected value of the dependent variable is Y.x1, x2, and xp are three independent or predictive variablesß0 is the value of Y when all of the independent variables are equal to zero.The regression coefficients computed are ß1, ß2, and ßp. When the independent variable is adjusted by one unit, each regression coefficient shows the change in Y.

## **CHAPTER FOUR**

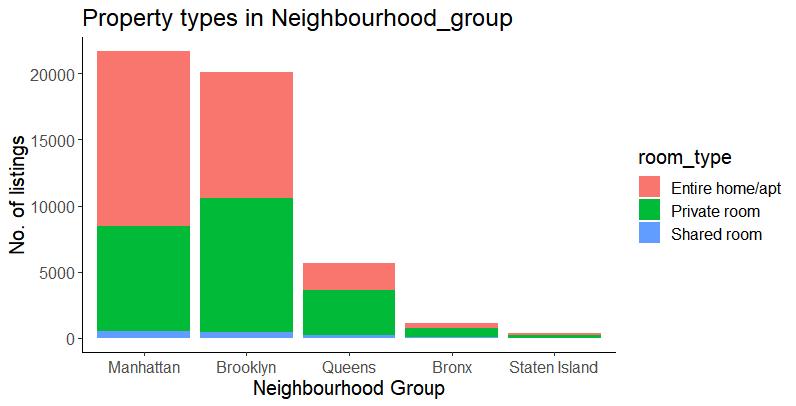
## **DATA ANALYSIS AND RESULT**

### **4.1 Exploratory Data Analysis**

The data is evaluated and the values of features and interactions between features were visualized. Data analysis gives details in graphs and charts that are easily understood in order to learn about various hosts and areas.

The summary statistics of the data were examined, followed by a search for missing values, which revealed 1.3% missing values in one category. The NA values were omitted in our analysis in other to fix the NA problem.

### **Total Property type listings In Neighborhood**



**Figure 1 shows Property types across neighbourhood groups**

### **Price Estimate by Neighborhood group**

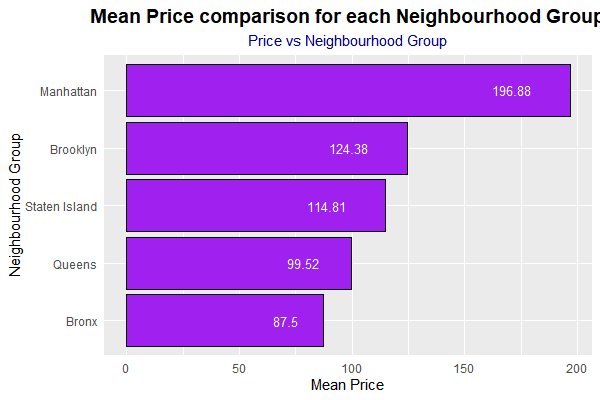
**Figure 2 showing average price across each neighborhood group**

### **Average Price by Room Type**

**Figure 3 shows Average price by room type**

The first graph illustrates the different types of listings available in NYC, with private rooms being the most popular and shared rooms being the least prevalent. The average pricing across neighborhood groups in NYSC is depicted in the second graph. The most expensive listing is in Marhantta, followed by Brooklyn and then Staten Island. The average price by room type is shown in the third graph. The most expensive option is to buy the entire house/apartment, which is followed by the least expensive options are a private room and a shared room.

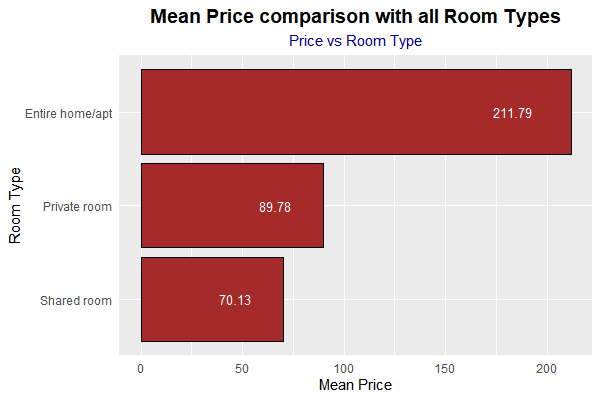
### **Mean Price Comparison for each Neighborhood Group**

****

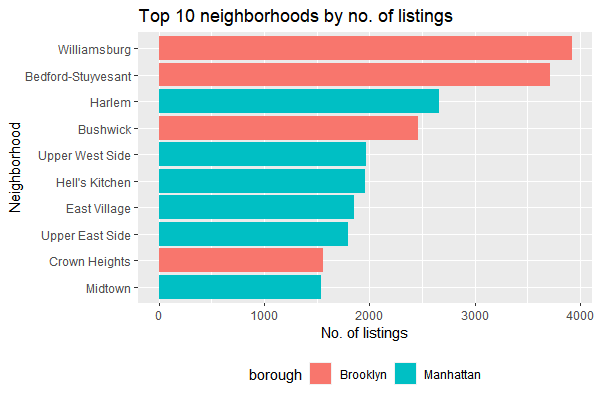
**Figure 4 showing the mean price across neighborhood group**

Manhattan has the highest average listing price (196.88 USD), followed by Brookyln (124.38). The fact that full apartments/homes are the most popular form of listing in Manhattan could be one factor for the high average price. The lowest offerings are in Bronx, with an average price of 87.5 USD.

### **Mean Price comparison with all Room Types**



Entire home has the highest average price. Apartment is followed by a private room and a shared room, as is customary.

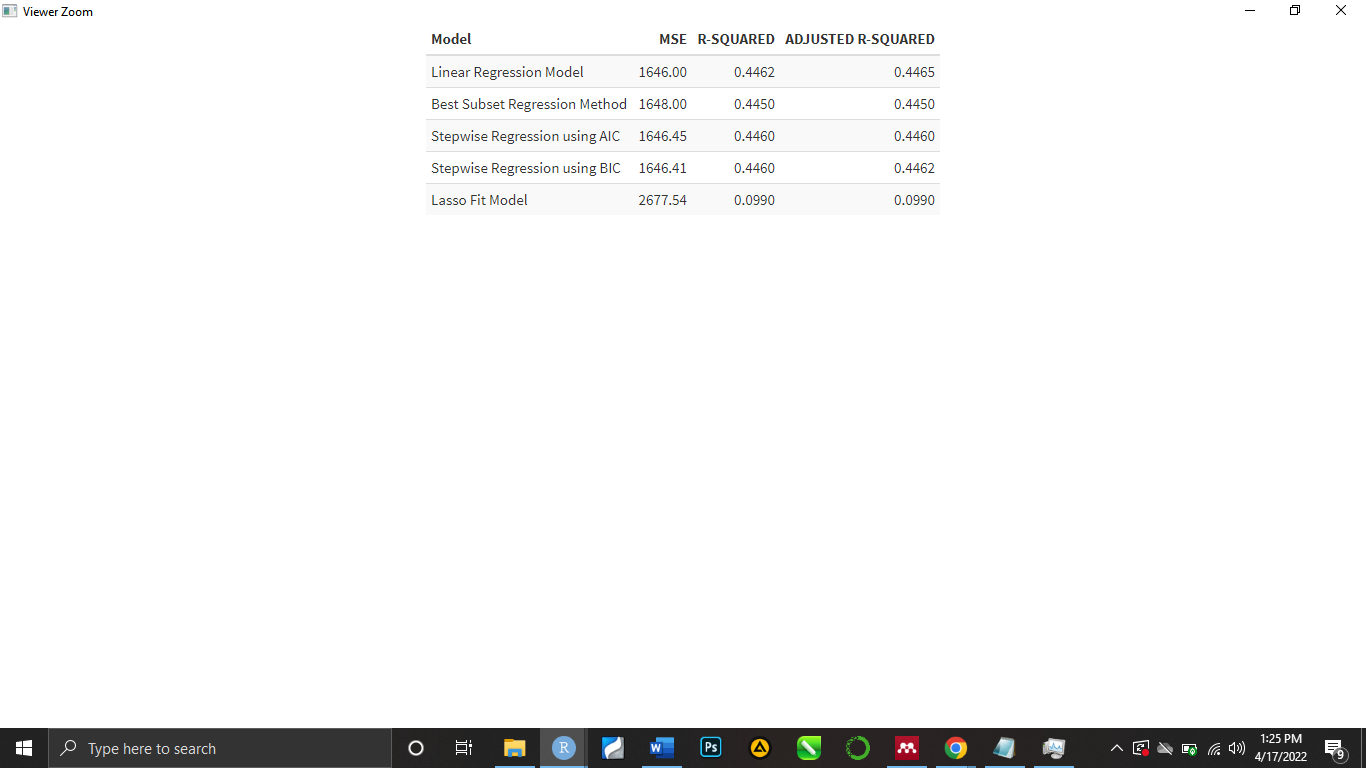


**Figure 6 showing the most popular Neighborhood by listings**

In terms of listings, Williamsburg is the most popular, followed by Bedford, while Midtown is the least popular.

# **RESULT**

Following data analysis and visualization, five distinct Machine Learning Models using Regression were compared and deployed on the testing and training datasets, with the findings given below.

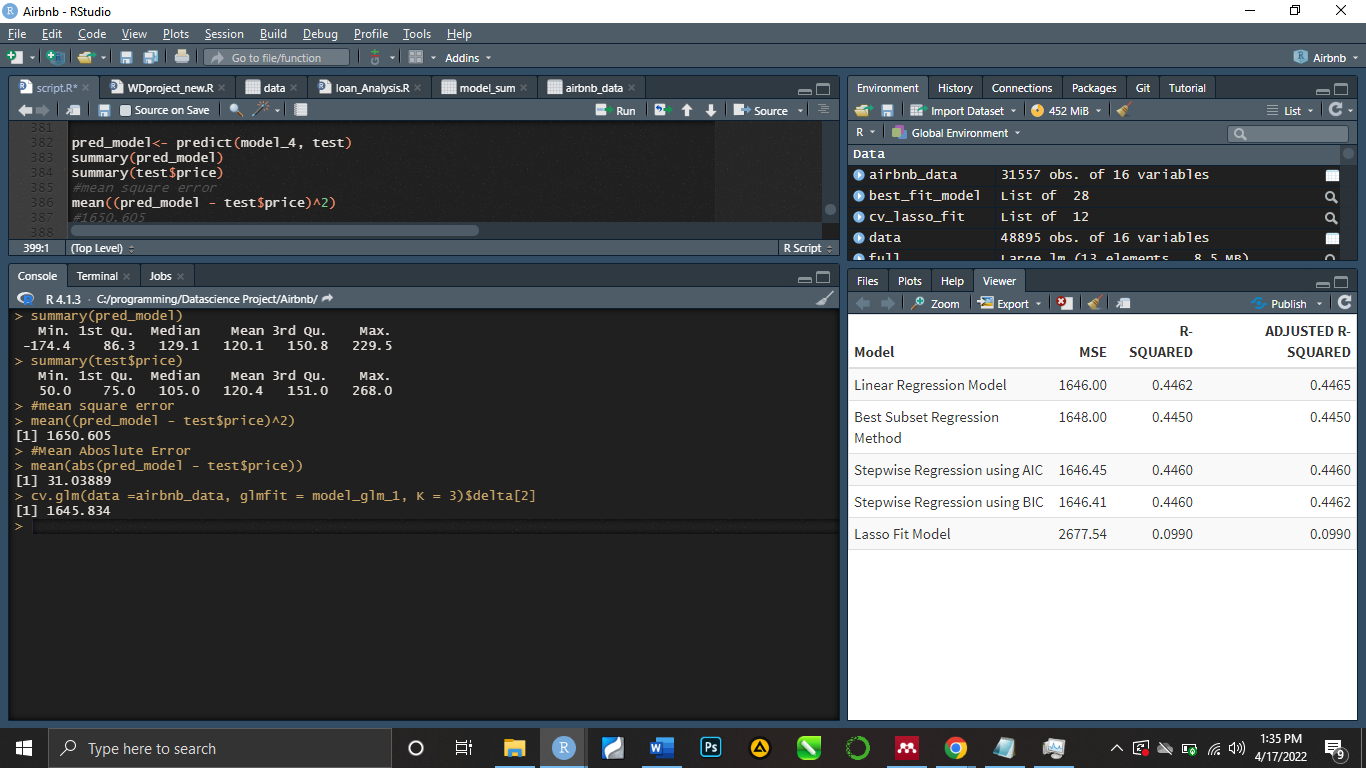
Root Mean Square, r2 Score(Regression score function. ), Mean Absolute Error(is a difference measurement between two constant variables) and the respected model predictions were calculated to determine the accuracy.

**Fig7 Showing comparison of different models with there parameters**

Because it has the lowest adjusted R-squared and the greatest MSE value, the model created using Lasso Regression is the least ideal.

The optimal combination of MSE and Adjusted R-squared is found in the linear model generated using stepwise regression utilizing both AIC and BIC, and these values are the same for both. However, Model 4 is chosen (stepwise regression with BIC) because BIC is thought to be more conservative, making the model more understandable.

## **4.1 Prediction and Cross Validation**



We can observe that the MSE of the Test Dataset, which is 1650.6, and the MSE of the Full Data, which is 1645.83, are nearly identical and also the compared summary of the prediction and price variable are identical. As a result, the variables chosen for the model work well as estimators of our dependent variable.

# **CHAPTER FIVE**

**CONCLUSION**

The price per night for NYC Airbnb listings was predicted in this study, and all the variables affecting the price of listings were investigated using exploratory analysis.

The factors that impact listing prices are: room\_type, neighbourhood\_group, longitude, availability\_365, minimum\_nights, calculated\_host\_listings\_count, latitude, number\_of\_reviews, reviews\_per\_month

The data yielded the following insights, which can help us answer the research question.The type of property and the type of neighborhood are crucial considerations. In comparison to other neighborhood categories, Manhattan has the most expensive properties. Private room is the most common listing type in all neighborhoods except Manhattan, where Entire Home/Apartment is the most common kind., Brooklyn is a decent pick with average price and quantity of reviews. The least prevalent form of listing is a shared room. Manhattan has the highest average listing price, followed by Brooklyn. The lowest offerings are in Bronx, with an average price of 87.5 USD.

Entire home/apartment has the highest average price, followed by private room and shared room. In conclusion the Model selected for analysis, predicts the variable with a good fit.

# **REFRENCES**

Alin (2020) ‘Predicting Airbnb nightly prices A regression problem using machine learning’.

Cassotti, M. and Grisoni, F. (2019) ‘T6\_moleculardescriptors\_variable\_selection’, *Molecular Descriptors*, pp. 1–11. Available at: https://www.math.arizona.edu/~hzhang/.

KNBS (2021) ‘No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title’, p. 6.

Kunwar, R. R. (2020) ‘Airbnb: Understanding the Concept, Recognizing the Values’, *Journal of Tourism and Hospitality Education*, 10, pp. 132–169. doi: 10.3126/jthe.v10i0.28764.

Li, J. and Biljecki, F. (2019) ‘THE IMPLEMENTATION of BIG DATA ANALYSIS in REGULATING ONLINE SHORT-TERM RENTAL BUSINESS: A CASE of AIRBNB in Beijing’, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(4/W9), pp. 79–86. doi: 10.5194/isprs-annals-IV-4-W9-79-2019.

Quattrone, G. *et al.* (2018) ‘Analyzing and predicting the spatial penetration of Airbnb in U.S. cities’, *EPJ Data Science*, 7(1). doi: 10.1140/epjds/s13688-018-0156-6.

Washington, D. C. (2019) ‘in Washington D.C.’

Zhang, Z. *et al.* (2017) ‘Key factors affecting the price of Airbnb listings: A geographically weighted approach’, *Sustainability (Switzerland)*, 9(9), pp. 1–13. doi: 10.3390/su9091635.