**Airbnb Price Prediction Data Analysis**

1. **Introduction and Statement of the Problem**

Airbnb is a social marketplace for accommodations that matches hosts up with guests. You can rent a place to stay through Airbnb or you can rent out your own home for making cash. It is one of the fastest growing companies with rapidly increasing user base. One challenge that Airbnb hosts are facing is that to determine the optimal nightly rent. This amount can be determined by considering some features like number of bedrooms, bath, room type, and more over the price should be linked to the dynamics of the market place. If the price is higher than the market place then renters may prefer the affordable options and on the other hand if the price is set too low then the potential revenue will be missed.

Statement of the project problem is to determine the price and build a predictive model for the price estimation. Also to find out any patterns in the data and correlation between different variables. This project helps me to get hands on experience on handling real world data and gain knowledge on building prediction models for any new data.

1. **Review of Literature**

Some existing literature works made studies on analyzing and predicting the spatial distribution of the Airbnb in the U.S cities. In this study [3] the distribution of the Airbnb is made in relation with both geographic, socio-demographic, and economic information.

Another studies which are more related to our study, inspect the hotels and sharing economy rental prices. In a recent work, Wang and Nicolau [1] have studied price determinants of sharing economy by analyzing Airbnb listings using ordinary least squares and quantile regression analysis. In a similar study, Masiero et al. [2] use quantile regression model to analyze the relation between travel traits and holiday homes as well as hotel prices. In a simpler work, Yang et al. [4] applied linear regression to study the relationship between market accessibility and hotel prices in Caribbean.

1. **Objectives of the Study**

The objective of this project is “To predicting competitive price for a listing based on the characteristics of the listing which helps the hosts when they enter the market”. Extend this study to predict the profits based on the location. To build a predictive model for the price estimation using different regression techniques. Visualize different patterns in the variables and represent the locations of the hotels using Heat maps. Through this project I would like to check which attributes have effect on the target variable price. I want to check whether reviews affect the price and reviews has any effect on price trough visualization using Tableau software.

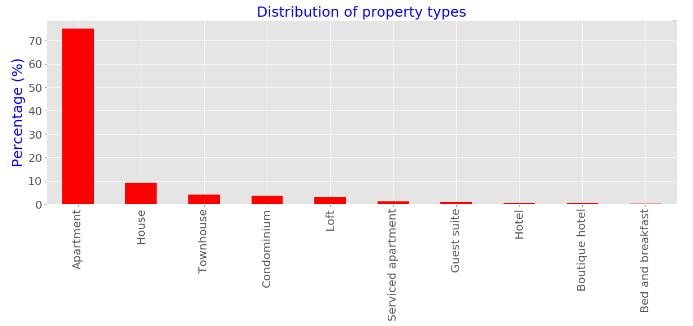
1. **Data Collection**

The data used in this project has been collected from the Inside Airbnb website. Airbnb website has a group named inside Airbnb in which the data on the listings of the different major cities is extracted. In this project I have worked on the listings from New York City a popular and crowded city in the United States. The dataset gives us the information regarding the Airbnb available in New York City and their metrics. This data file includes all needed information to find out more about hosts, geographical availability, and necessary metrics to make predictions and draw conclusions.

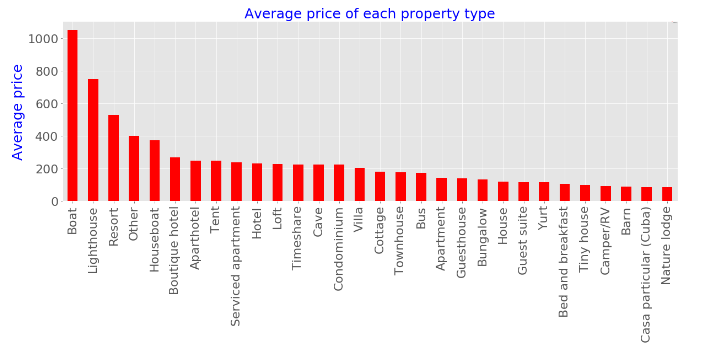
This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values. Each of the dataset is a listing available in the New York City for renting on Airbnb. Each row can be termed as an observation and each observation is described using different features called as variables. This dataset is sufficient to build a model to predict the prices. In order to build a model for the profit prediction I need to collect more data related to the expenditure in the New York and synchronize these two data sets.

1. **Exploratory Data analysis**

EDA is performed to gain insights about the various features of the dataset. This analysis is made using tableau software where it gives visual representations.



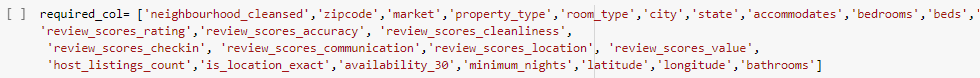
Above figure is the frequency of the rentals where apartment are the highest in number.



Above figure shows the average price per each rental properties in New York City. We can see that Boat, Lighthouse and Resort are the expensive rentals.

1. **Data Exploration and Cleaning**

There is a continuous process of cleaning the data throughout the project. The dataset has about 106 attributes some of which are numerical and some are categorical data. As the main goal of the project is to predict the price, we consider only those attributes which are related to the price attribute. Some of the columns which are not helpful for this particular analysis were deleted by manually searching and only some columns were considered. A new data frame is created with only required columns in it.



After the creating a data frame with required item, price attribute is taken into consideration as it is the main attribute in the dataset. In the dataset, the price column has the $ symbol for each and every entry. This symbol will not allow to make the analysis hence this symbol is removed using replace function.



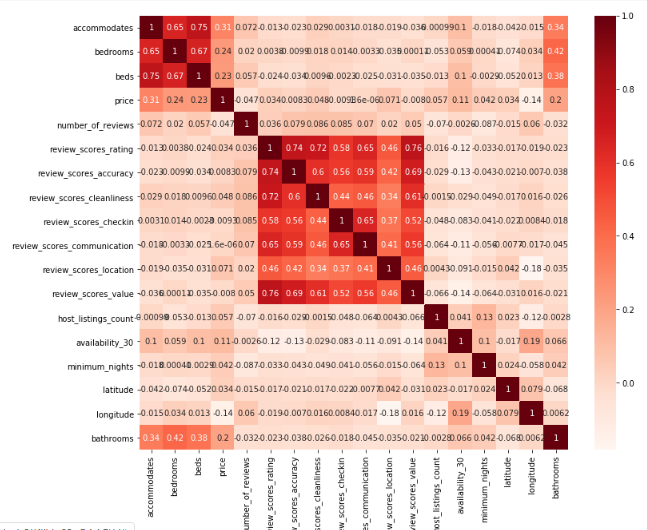
As the prices column has values which are at different, I standardized the values using the log function. This is the technique usually applied to make the numerical values to a common scale to obtain better results. Whenever we apply log function to any numerical data it is necessary to check for the zero values in that data. Because log of zero would tend to infinity. Infinity values in the dataset will not allow any analysis techniques. In the dataset I found 10 null values and these values were filled with the mean value. Once all the zeros are replaced with the numerical values then the log function is applied to the entire column.



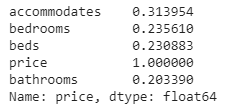


**Correlation Matrix**

Correlation matrix is created using pandas. From this correlation matrix we can know the relationship between the multiple attributes. From this correlation matrix we can find the attributes which are highly related to each other. All the attributes which are highly correlated with each other are eliminated from the dataset.



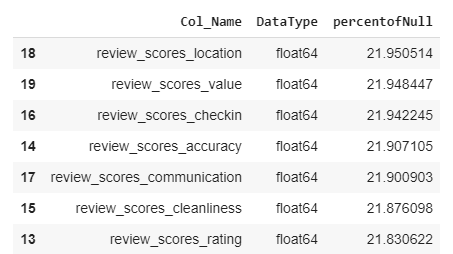
In order to know which attributes are more related to the target variable a numerical tabular value is determined.



From the above values we can see that number of accommodates attribute has high effect on the price.

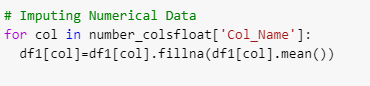
**Handling Missing values:**

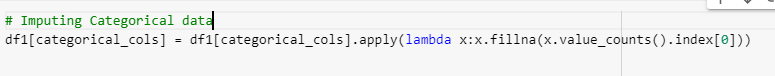
The dataset has many missing values in different column. These missing values were handled step by step using different functions. Firstly, null percentage value is calculated in order to know how many values from total number of records in each column. If the null percentage is more than 50 it better to delete the column than imputing it.



**Imputing Missing values:**

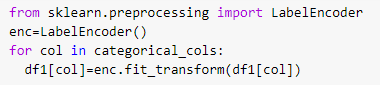
In this process two different data frames were created one with float datatype variables and one with int datatype variables. Since we need different methods to handle numerical and categorical data types. Numerical data has been imputed with the mean values. Whereas the categorical missing data has been filled with the highly repeated words (which can be termed as mode)





**Label Encoding**

Label encoding is the process of converting the categorical data into numerical data so as to make it suitable for machine to read it. For this process I have used label encoder from the sklearn library. From this all the categorical data has been encoded into numerical labels.



1. **Data Analytics and model building**

The main aim of this project is to develop a price prediction model. In this project I have used different types of regression models to predict the price. Linear regression, KNN model, ridge, Lasso regression and random forest regression. Mean square error and R2 score of each model is calculated and the results were compared. All these models were implemented from the Scikit-learn library.

Scikit-Learn library is the machine learning library in python. It contains all the built-in functions for the major machine learning algorithms. All these functions are simple and easy to implement on any data. Models which are used to predict any numerical data like price from Airbnb dataset are known as the regression model.

For the model to build, data need to split into training and testing data. The training data consists of the inputs and the known outputs. The model will learn and get trained on this data and then it is used to generate outputs on the other new data. The test data is used in order to test our model prediction. This testing data will be given to the model and output will be predicted using this model and compared to the actual outputs. In this analysis the training and testing data are split into 80% and 20% respectively.



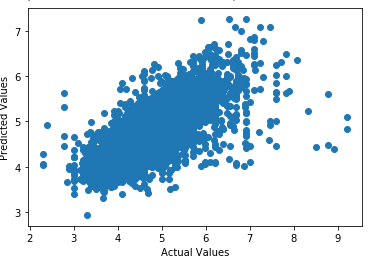
Where X is the input data and Y is the target variable (Price)

**Linear regression**

This is the simple regression type where the dependent variables and target variables are considered to be in liner nature. WE performed multi linear regression in this analysis.

The Mean square errors and R2 score of this model are as follows





Above is the scatter plot draw between actual values and the predicted values using the linear repression model.

**KNN regression**

KNN refers to the KneighborsRegressors class where Kneighbors are the K-nearest neighbors. This also comes into regression model types. By default n-neighbors are set to 5, algorithm is set to auto, and p value is set to 2 which corresponds to Euclidean distance. In this analysis I have set the algorithm to Brute type. After fitting this model and making the predictions the results are as follows.





**Ridge regression**

In ridge regression (also known as shrinkage regression) we add a constraint on the sum of squares of the regression coefficients. Thus in ridge regression our objective function is:



Here λ is the regularization parameter which is a non-negative number. Here we do not assume normality in the error terms.



**Lasso Regression**

Lasso stands for Least Absolute Shrinkage and Selection Operator. It makes use of L1 regularization technique in the objective function. Thus the objective function in LASSO regression becomes:



λ is the regularization parameter and the intercept term is not regularized. We do not assume that the error terms are normally distributed.

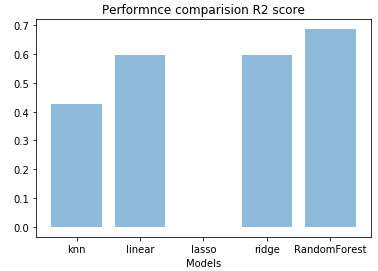
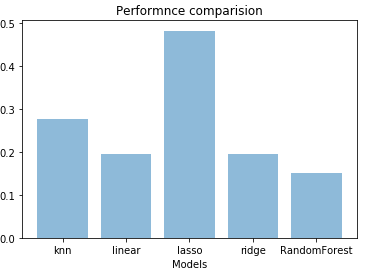
Results are as follows





**Results**

Predictions models were built using different regression types. Accuracy and mean square errors of all the models were compared and showed in the bar chart format. By analyzing the results of all the prediction model, the Random forest model has the less mean square error value and the more R2 score. This mean Random forest model is the best fit to this Airbnb dataset for predicting the price for any rental house.



**References**

[1] D. Wang and J. L. Nicolau, “Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com,” International Journal of Hospitality Management, vol. 62, pp. 120–131, 2017.

[2] L. Masiero, J. L. Nicolau, and R. Law, “A demand-driven analysis of tourist accommodation price: A quantile regression of room bookings,” International Journal of Hospitality Management, vol. 50, pp. 1–8, 2015.

[3] Quattrone, G., Greatorex, A., Quercia, D. et al. Analyzing and predicting the spatial penetration of Airbnb in U.S. cities. EPJ Data Sci. 7, 31 (2018) doi:10.1140/epjds/s13688-018-0156-6

[4] Y. Yang, N. J. Mueller, and R. R. Croes, “Market accessibility and hotel prices in the caribbean: The moderating effect of quality-signaling factors,” Tourism Management, vol. 56, pp. 40–51, 2016.