ANLY 501 Project Report  
Airbnb Price Analysis

short line

Yunjia Zeng, Weile Chen, Tuo Wu, Lechuan Qiu, Prerna Kaul

Advised by: Dr. Ami Gates

11th November 2018

# 

[**Introduction**](#_vc8fzo7x5u1x) **2**

[**Basic Statistical Analysis and Data Cleaning Insight**](#_ovawxjoi07ia) **3**

[**Histograms and Correlations**](#_lzr2si3oshbv) **8**

[**Cluster Analysis**](#_i85ws8se9wc5) **11**

[K Means Clustering](#_sjuknko6jlzv) 11

[Agglomerative Ward Clustering](#_j3vsq58btqu1) 15

[DBSCAN Clustering](#_oswjm46bpy) 17

[**Association Rules Mining Analysis**](#_2bqgcvln9o5l) **20**

[**Hypothesis Testing**](#_pm33ha4fyknl) **21**

[Hypothesis 1: To test whether the property type apartment and house show difference in Airbnb prices](#_q4b2bmy2jgh2) 21

[Hypothesis 2: To test if different room types differ from each other significantly in Airbnb prices](#_dajsoxh8fv3) 22

[Hypothesis 3: To test there is a significant linear relationship between some features and Airbnb prices.](#_93aa3gibfv88) 24

[**Predictive Models**](#_cmyvxpi80pbl) **26**

[Preprocessing](#_uj2fa8dzpsek) 26

[Decision Tree Predictive Model](#_4t9r06r0b0xb) 27

[K Nearest Neighbor Predictive Model](#_hginpkdnokaz) 29

[Support Vector Machine Predictive Model](#_t9jt4gg8wjzj) 31

[Naive Bayes Predictive Model](#_7d5cs5n33qkq) 33

[Random Forest Predictive Model](#_9r6b4xrbsyug) 36

# Introduction

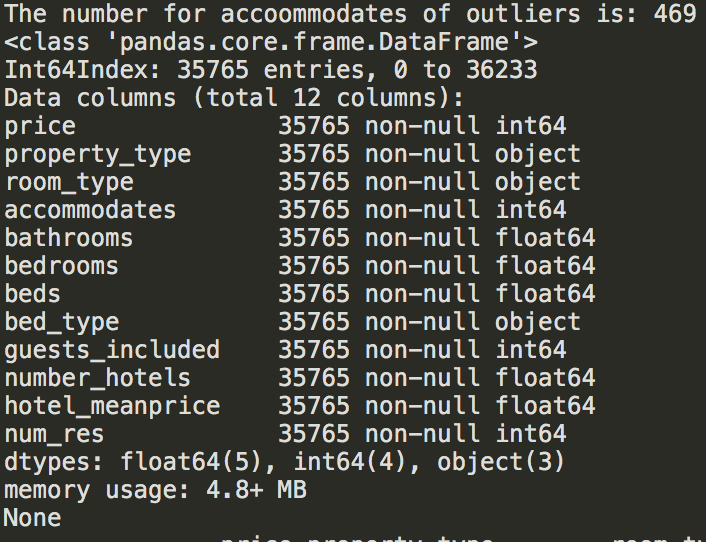
The aim of the analysis is to explore the Airbnb listings in the city of New York to better understand that how different factors like bedrooms, hotel’s price, house types among other factors can be used to predict the price of a new listing that is optimal in terms of the host lucrativeness and still is affordable to the guests.

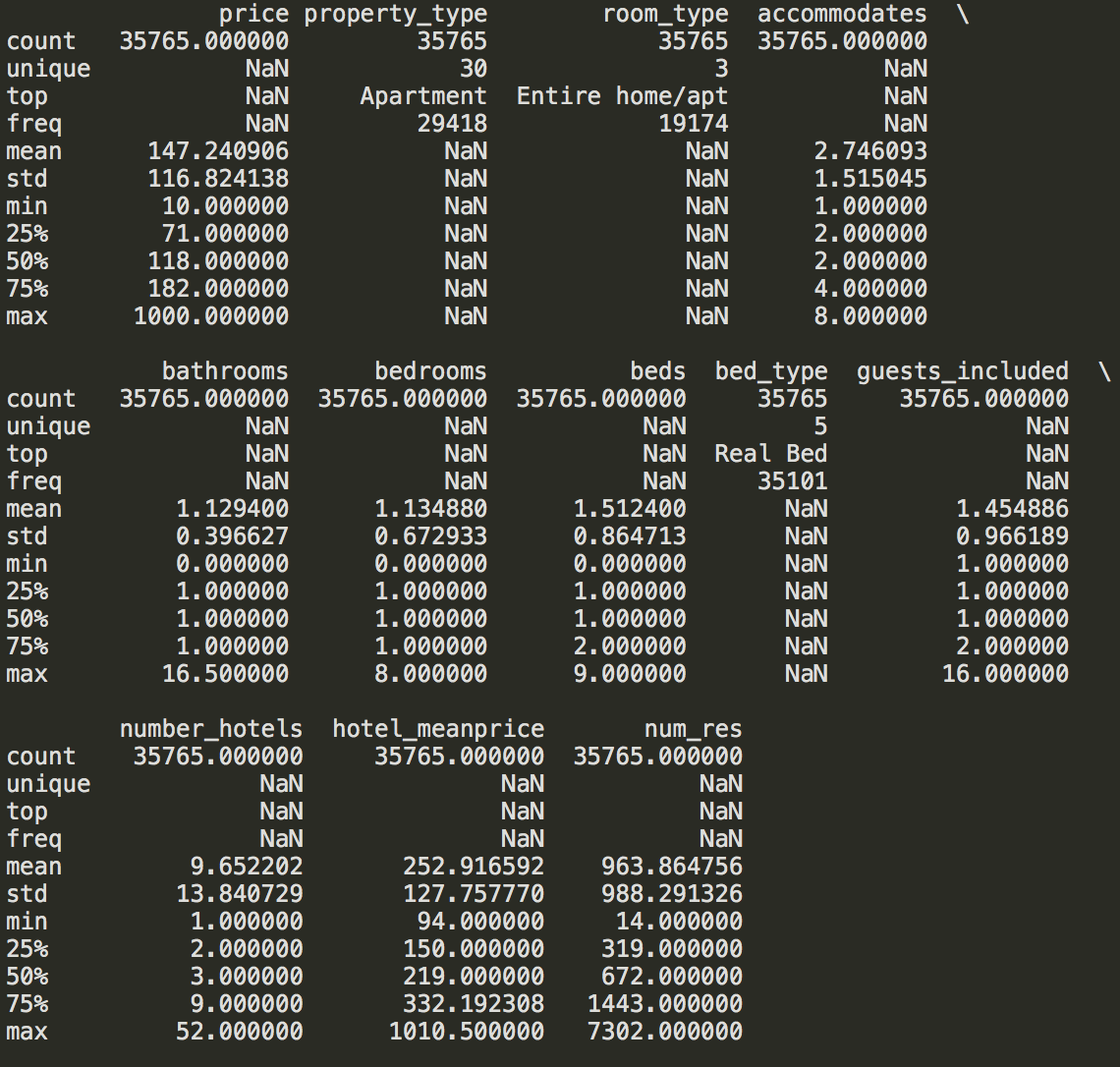
Airbnb is a well-known online marketplace and hospitality service, enabling people to list or rent short-term lodging. It is a platform that connects people looking to rent out their homes to earn income with people who are looking for convenient and affordable accommodations. People use Airbnb for the purposes of traveling, business, and homestay. However, many hosts and guests have less knowledge about how to fix a fair rental price. The purpose of the project is to help hosts set a smarter price to attract more customers.

In this part of the project, various methodologies were applied like clustering, association rule mining, histogram depictions, correlation analysis, and some basic statistical analysis to explore and summarize the main characteristics. After coming up with three hypotheses during the analysis, some models like the Decision tree, Naive Bayes Predictive Model etc. were applied to test the three hypotheses that were arrived at. The analysis and the work done is supported by graphs for easy visual understanding for both professional and non-professional audience.

# Basic Statistical Analysis and Data Cleaning Insight

The following figure shows the “type” of each variable.

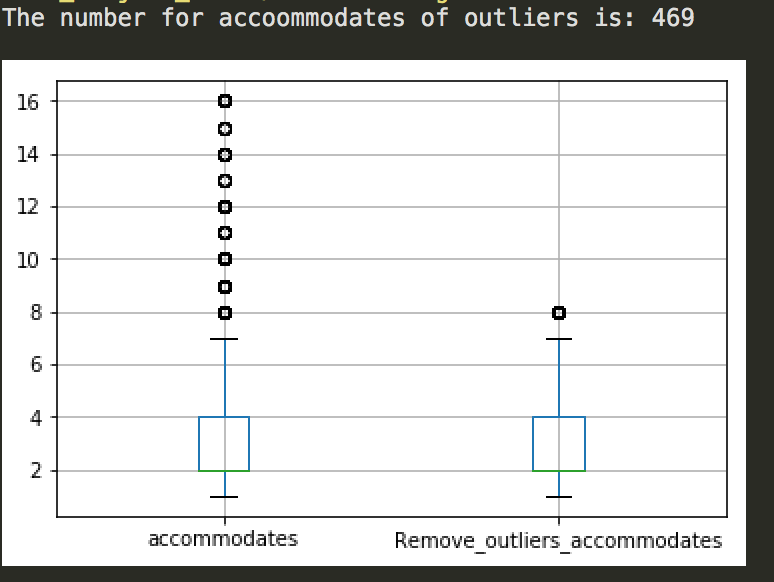




Above summary statistics offers the mean, standard deviation, minimum, maximum, and other statistical measures for numeric variables, and also provides the number of values for categorical variables. Box plots below also show some results.

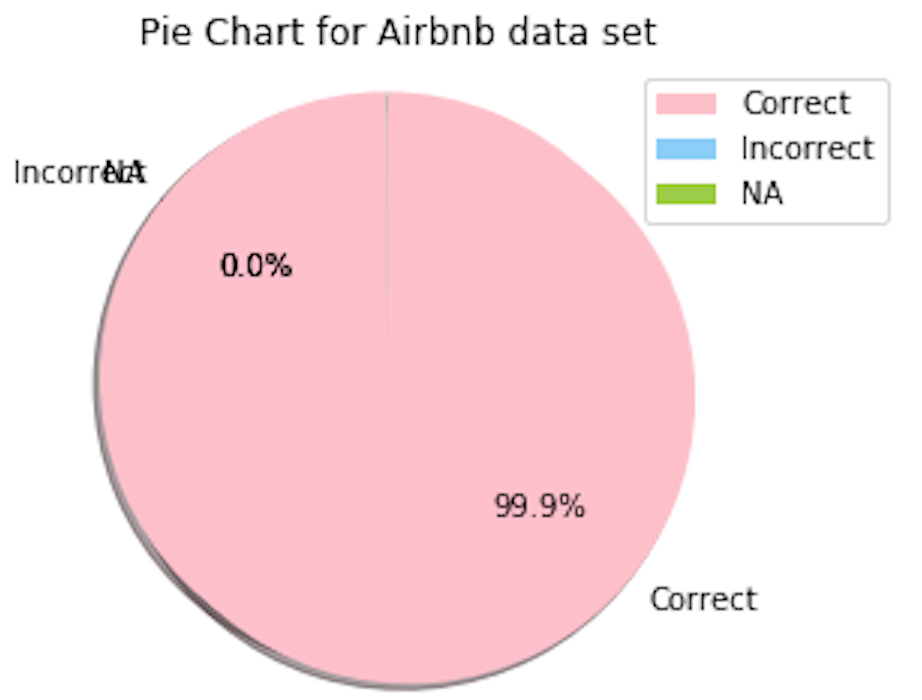


Here, mean and standard deviation method is used to detect and remove outliers. Specifically, if rows contain value of each attributes exceed the range from mean-3\*standard deviation to mean+3\*standard deviation for each variable, they are removed. Removing outliers may make the prediction more accurate. For example, below box plot shows the range of accommodates before and after removing outliers.

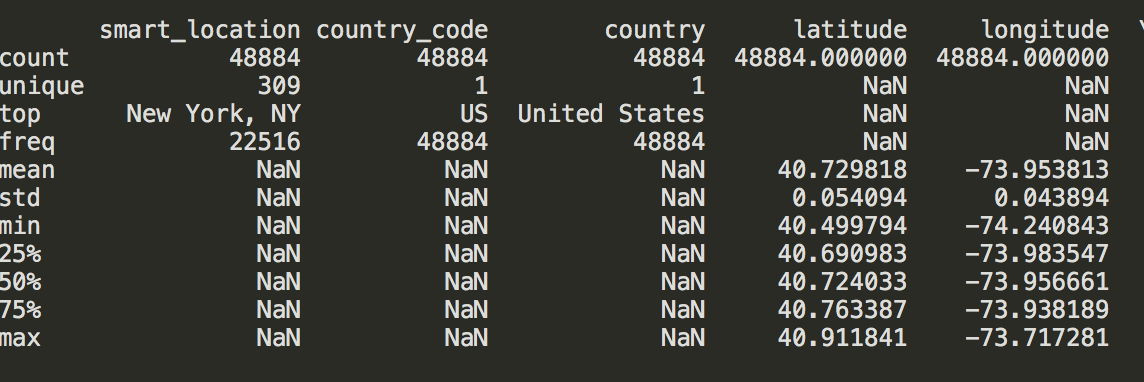


In terms of hotel prices data set, the hotel prices are binned together based on the zip code and taken mean. The new variable is the mean of hotel prices based on zip code.

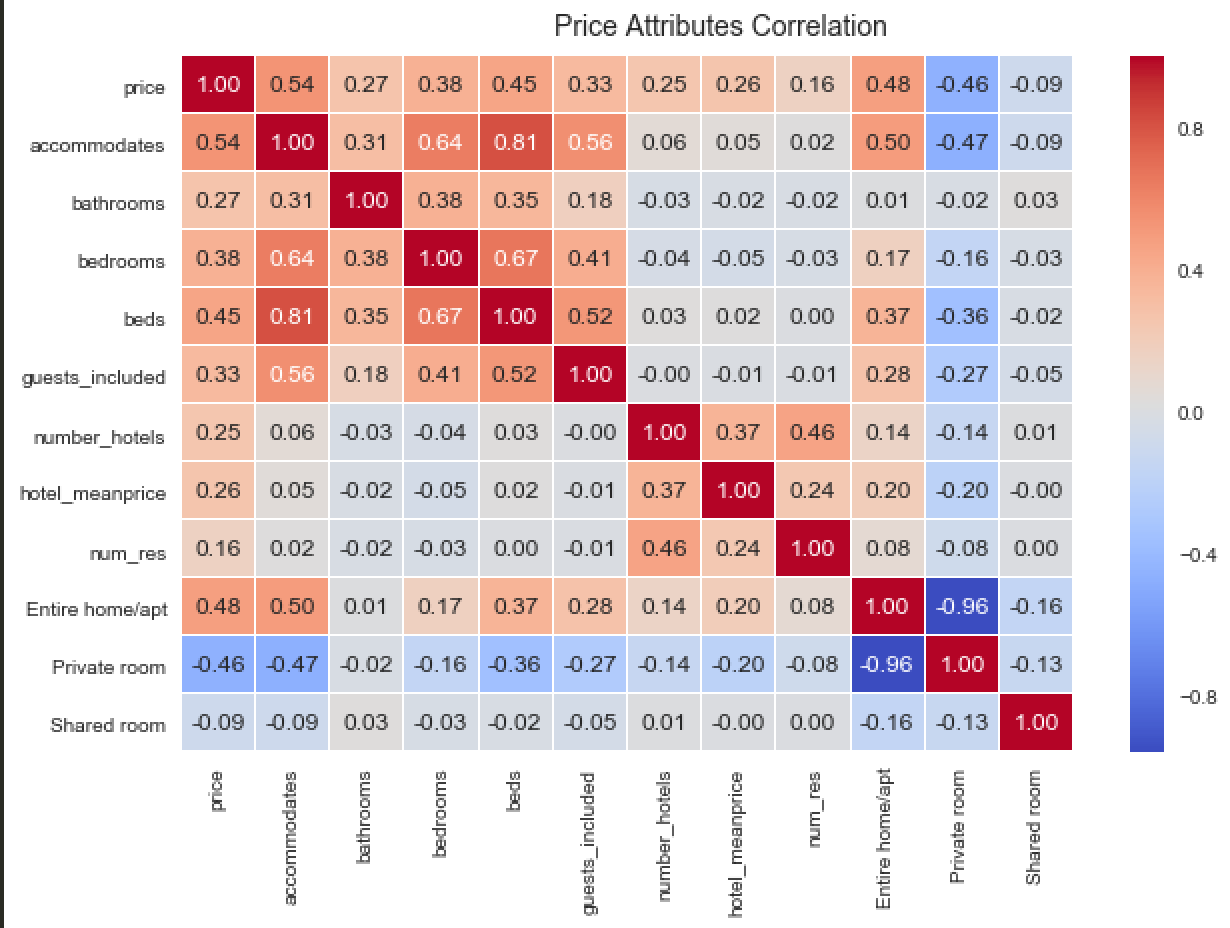
Given that ‘NA’ and incorrect values in all datasets only account for a minor part, removing the rows that contain ‘NA’ and incorrect values will not negatively influence the final result for all three data sets. For example, in the Airbnb data set, there are more than 36,000 rows, but only 227 rows contain ‘NA’ values and 16 rows contain incorrect values. The results is shown in the following figure.



Some variables or columns offering little information are also removed, e.g, the variables: ‘country’ and ‘country\_code’ have one 1 unique value, which is shown by the following figure, since all data concentrates on the United States. Also, some variables contain duplicate information, e.g., smart\_location, and latitude and longitude also contain the geographic information, but latitude and longitude contain more details and are essential to build up the Geographic Information System. In this case, the feature ‘smart\_location’ is dropped.



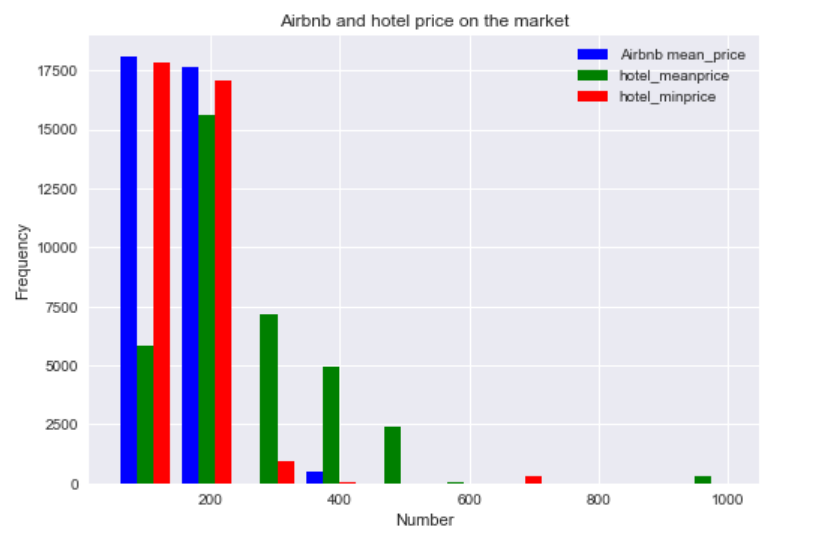
To analyze the categorical variables, first they are converted into dummies variables. Then, a heat map was made that shows the correlation coefficients between these variables. If the variables showed strong correlation with the price, they were kept. For instance, the variable ‘room\_type’ has 3 values, which are Entire home/apt, Private room, and Shared room respectively. The figure below shows the correlation between them and the price.

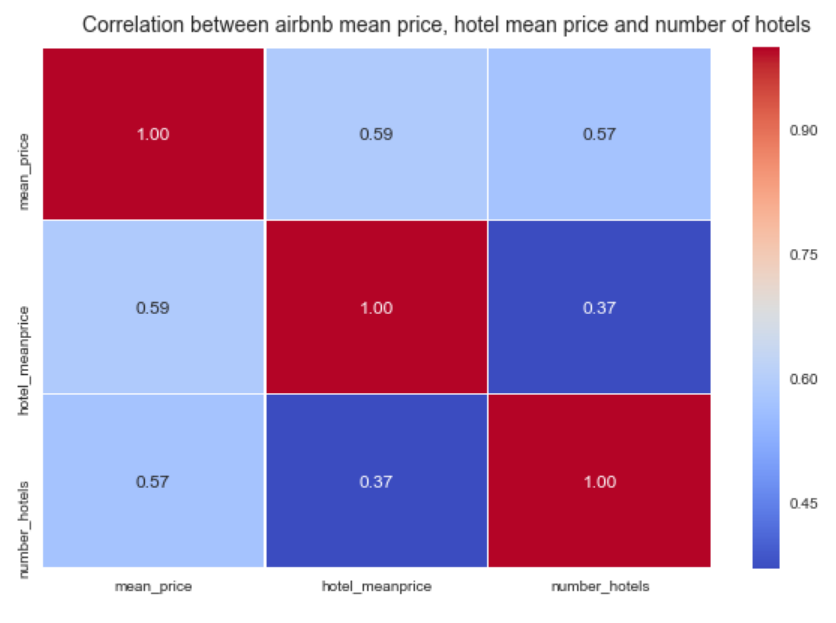


From the figure above, Entire home/apt with value 0.48 has a positive relation with the price, whereas the Private room with value -0.46 has a negative relation with the price. In this case, this categorical variable is kept. On the contrary, other categorical variables that have a little impact on the price are dropped.

# Histograms and Correlations

To figure out the smart Airbnb’s price strategy, it is important to compare Airbnb’s price with the hotel price in an area and to know when Airbnb has price advantages. In the following graph, Airbnb’s mean price in a certain region, hotel’s minimum price and mean price in a certain region were picked as variables, since we want to learn the price distribution of Airbnb and hotels. Observed from the graph, around half of Airbnb’s price is nearly $100, but the number of hotels which set their price near $100 is ⅓ of Airbnb’s. Also, hotel’s minimum price in a certain region has a close price to Airbnb’s price. When looking at a price near $200, it can be seen that the frequency of hotel’s mean price rapidly increases. At this price level, Airbnb still has the highest frequency in most of the region. However, when price keeps going up, there are few of Airbnb houses available, but there are still lots of hotels setting their price at a high level. In conclusion, compared with the hotel, Airbnb’s price concentrates more near $100 and $200. It shows that one of the reasons that customers choose Airbnb is its price advantages. When price going up, customers who can afford high price accommodations might choose a hotel, since it could provide better amenities and services.







When exploring the correlation between different variables, the number of nearby hotels, hotel’s mean price and Airbnb’s mean price were chosen as targets. From the above graph, it is shown that the hotel’s mean price has a relatively strong positive relation with Airbnb’s mean price, since their correlation coefficient is 0.59. This means that Airbnb’s owner will take look at nearby hotels’ price to set the price for their property. In the first subplot as well this positive correlation is shown. In addition, Airbnb’s mean price also has a strong positive relation with the number of hotels and the correlation coefficient is 0.57. It indicates that when there are more hotels in a certain region, the Airbnb’s price will be relatively higher. At last, the hotel’s mean price and the number of hotels have relatively weak coefficient comparing with the other two pairs which is 0.37, but it is still a positive correlation between two variables.

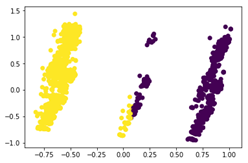
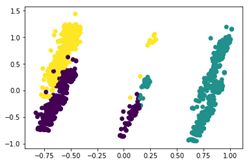
# Cluster Analysis

There were 3 cluster analysis conducted on the dataset - K Means clustering, Agglomerative Ward clustering, and DBSCAN clustering. Before conducting the analysis, the categorical values were converted to dummy variables and then the dataset was normalized. In general, it was noticed as the number of clusters were increased in the testing for best-fit clusters the silhouette score, which tells the degree of goodness of clustering, decreased. When there was a large number of clusters it could be seen from the graphs that the clusters started overlapping each other. Since there was a high dimensionality in the dataset the PCA projection of clusters was plotted.

The observations and conclusions for different clustering methods are as follows:

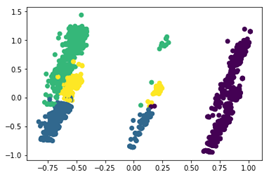
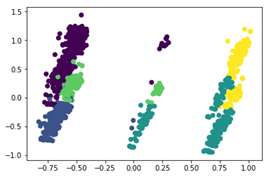
## K Means Clustering

The K Means was carried out for different k values ranging from 2 to 10. From observing the graphs and the silhouette score for different cluster parameters it was inferenced that the best clustering occurred when the number of clusters = 3 and the average silhouette score = 0.4188064504340923. It was also seen that silhouette score for higher k values (like 10) was also about 0.4 but in the graph there were overlapping clusters and hence, it was not considered a good clustering.

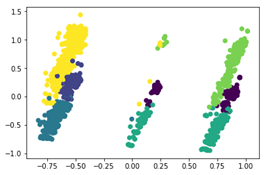
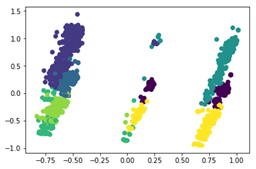
Number of clusters = 2, Number of clusters = 3,

average silhouette\_score is : 0.4140038748224665 average silhouette\_score is : 0.4188064504340923

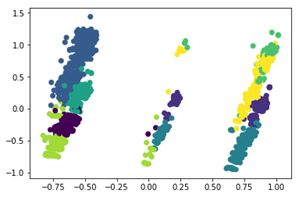
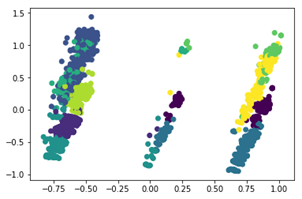
Number of clusters = 4, Number of clusters = 5,

average silhouette\_score is : 0.34738506204337805 average silhouette\_score is : 0.3301044126639676

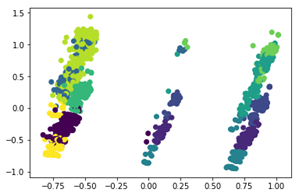
Number of clusters = 6, Number of clusters = 7,

average silhouette\_score is: 0.3581382738932519 average silhouette\_score is: 0.36500807627807125

Number of clusters = 8, Number of clusters = 9,

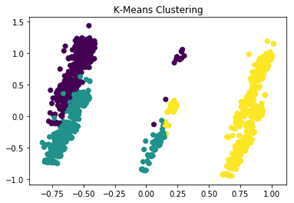
average silhouette\_score is: 0.37618610723528373 average silhouette\_score is: 0.3907518377014031



Number of clusters = 10,

average silhouette\_score is: 0.39805964453599835

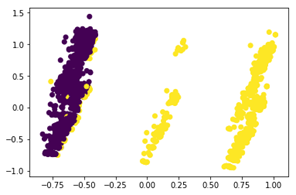
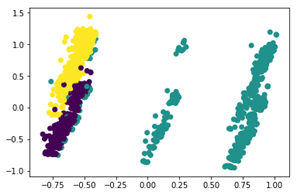
Best K-Means Clustering: Number of clusters = 3, average silhouette\_score is: 0.4188064504340923



## 

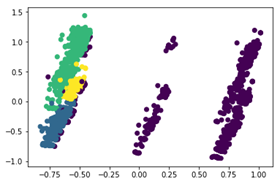
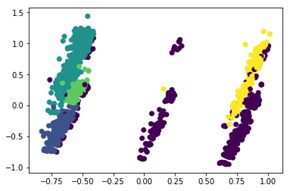
## Agglomerative Ward Clustering

The agglomerative clustering was carried out for different n\_clusters i.e. the number of cluster values ranging from 2 to 10. From observing the graphs and the silhouette score for the different number of clusters it was inferenced that the best clustering occurred when the number of clusters = 3 and the average silhouette score was = 0.40560405413301365. It was also seen that silhouette score for higher k values (like 10) was also about 0.4 but in the graph, there were overlapping clusters and hence again, it was not considered a good clustering.

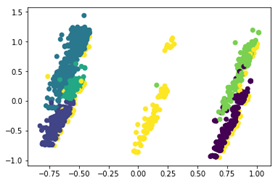
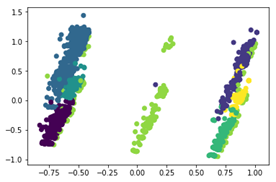
Number of clusters = 2, Number of clusters = 3,

average silhouette\_score is : 0.4053959900808611 average silhouette\_score is : 0.40560405413301365

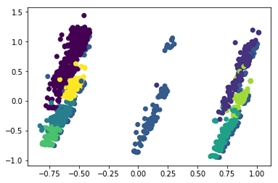
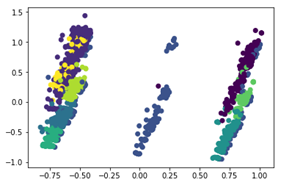
Number of clusters = 4, Number of clusters = 5,

average silhouette\_score is : 0.3341187979366904 average silhouette\_score is : 0.36107969679360713

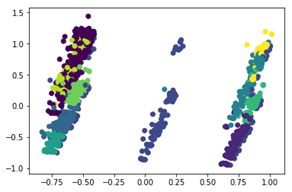
Number of clusters = 6, Number of clusters = 7,

average silhouette\_score is : 0.3830961280504269 average silhouette\_score is : 0.3645299986916324

Number of clusters = 8, Number of clusters = 9,

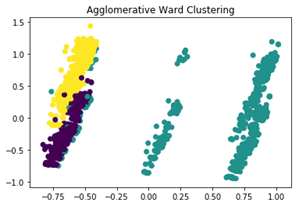
average silhouette\_score is: 0.36896898481055157 average silhouette\_score is: 0.38364938785207836



Number of clusters = 10,

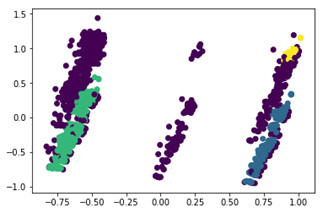
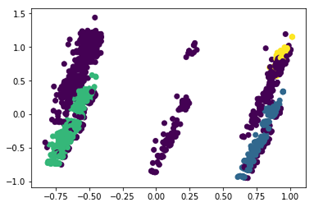
average silhouette\_score is: 0.39749121241716795

Best Ward Clustering: Number of clusters = 3, average silhouette\_score is: 0.40560405413301365



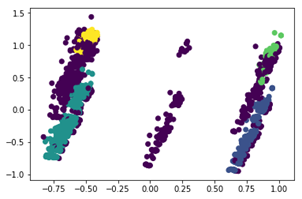
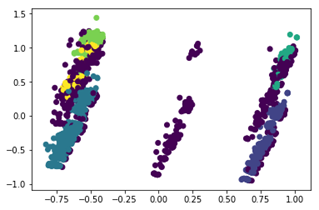
## DBSCAN Clustering

The DBSCAN clustering was performed for different values of ‘eps’ i.e. the radius of a cluster, keeping the number of minimum samples same. The number of clusters was estimated using the cluster labels that were generated by performing this clustering. It was seen that as the eps increased the number of clusters slightly increased and also the silhouette score increased. But again, with large number of clusters formed the graph showed overlapping clusters which is not considered good clustering measure and hence after carefully observing the graphs and the silhouette scores for different estimated number of clusters it was inferenced that the best fit clustering occurred when the estimated number of clusters = 4 and the average silhouette\_score for the estimated number of clusters = 0.4002130195933568.

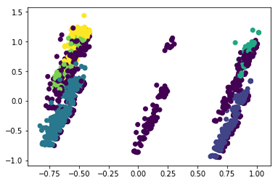
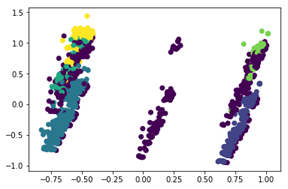
Estimated number of clusters: 4 Estimated number of clusters: 4

average silhouette\_score is: 0.3970906600035595 average silhouette\_score is: 0.4002130195933568

Estimated number of clusters: 5 Estimated number of clusters: 6

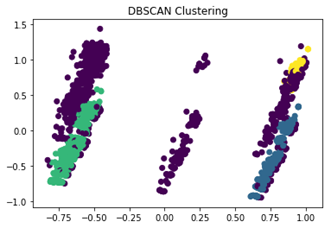
average silhouette\_score is: 0.40886867189807563 average silhouette\_score is: 0.43123375418437526

Estimated number of clusters: 6 Estimated number of clusters: 6

average silhouette\_score is: 0.43966352875699444 average silhouette\_score is: 0.4416355362233816

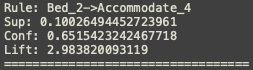
Best DBSCAN Clustering: Estimated number of clusters = 4, average silhouette\_score is: 0.4002130195933568



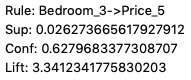
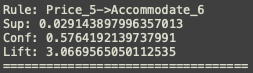
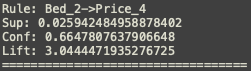
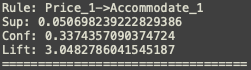
Plotting the PCA projection of clusters gave a better insight into the clusters that were formed by these various methods. If only silhouette score was to be considered for deciding the goodness of clustering it would not give a proper and correct result because the silhouette scores were similar for 3 clusters and 10 clusters formed. But because by plotting the graphs for clusters it was seen that for a high value of the number of clusters i.e k, the clusters were overlapping each other and hence were not considered good clustering, even though the silhouette score was high. By seeing the clusters visually, a better and correct result was reached.

# Association Rules Mining Analysis

Before creating an accurate predictive model, Association Rules Mining Analysis is a very useful tool for exploring the data and looking for hidden relationships. To perform the analysis, the Apriori Algorithm was utilized. Initially, a subset of 7 variables was selected, which included Price, Property types, Number of Accommodates, number of bathrooms/bedrooms, etc. Surprisingly, the strongest rule is between “2 beds” (Offering 2 beds for guests) and “4 Accommodates” (Available for 4 guests). “Sup” with the value “0.1002,” tells that from about 40000 listings, 10% of the hosts are offering 2 beds for 4 people. Besides this, the rest of the strong rules are all about room type. This exposes the fact that, in the dataset almost every single row has room type, and it is either “Entire house” or “Private room”.



Therefore, the variables Property type and Room type were dropped from the subset. For the new subset, three different support levels were selected: 0.25, 0.5, and 0.75. With this new subset of data, association rules relating to price range were shown in the results. The price column is binned into 5 different groups based on the value of price. “Price\_1” has the lowest price range and “Price\_5” has the highest price range. The following rules imply a positive relationship between the price and the house capacity i.e. a house with a larger capacity will have a higher price range.



Also, since all the results come from the same dataset, a different “sup” value did not make a big difference on the strongest rules, such as “Price\_1” and “Accommodate\_1”, “Bed\_2” and “Bathroom\_1”, “Bed\_2” and ”Bedroom\_2”. The appearance of these frequent rules and patterns mostly result from the nature of houses and human behavior. Most houses have similar features and construction standards, this makes the listings show similar patterns. Also, listings’ information is edited by the hosts, so they can control the listings information to make it more attractive to guests. The rule like “Bed\_2” and “Accommodate\_4” is a good example of how hosts try to attract more guests.

# Hypothesis Testing

## Hypothesis 1: To test whether the property type apartment and house show difference in Airbnb prices

Firstly, the figure below shows the box plot on prices for apartment and house.



Then, the hypothesis can be set up as follows:

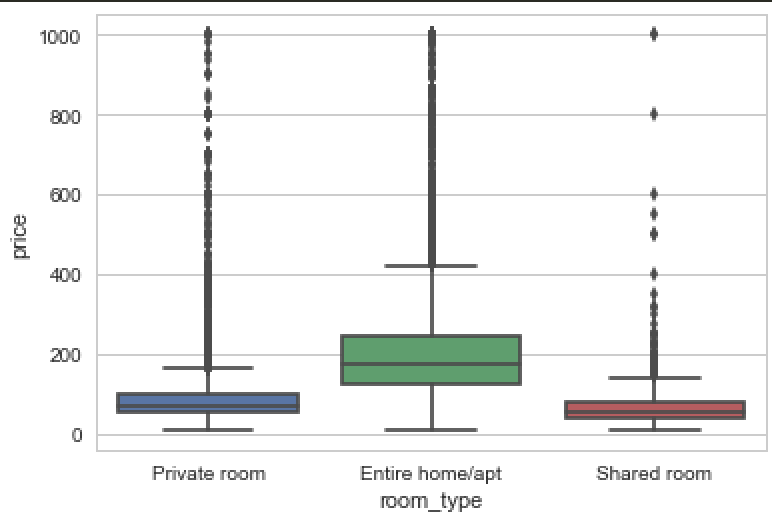
* Null hypothesis: there is no difference between the mean prices of the apartment and that of house
* Alternative hypothesis: the mean prices of these two property types are the different

## 

From above result, the p-value<0.05 provides evidence to reject the null hypothesis that there is no difference between the mean prices of apartment and that of house.

## Hypothesis 2: To test if different room types differ from each other significantly in Airbnb prices

First, the following figure shows the range of prices based on different room types by making box plots.



Next, it is necessary to set up a hypothesis to see whether different room types differ from each other in prices. So,

* Null hypothesis: the mean of prices from different room types are the same
* Alternative hypothesis: the mean of prices from different room types are the different

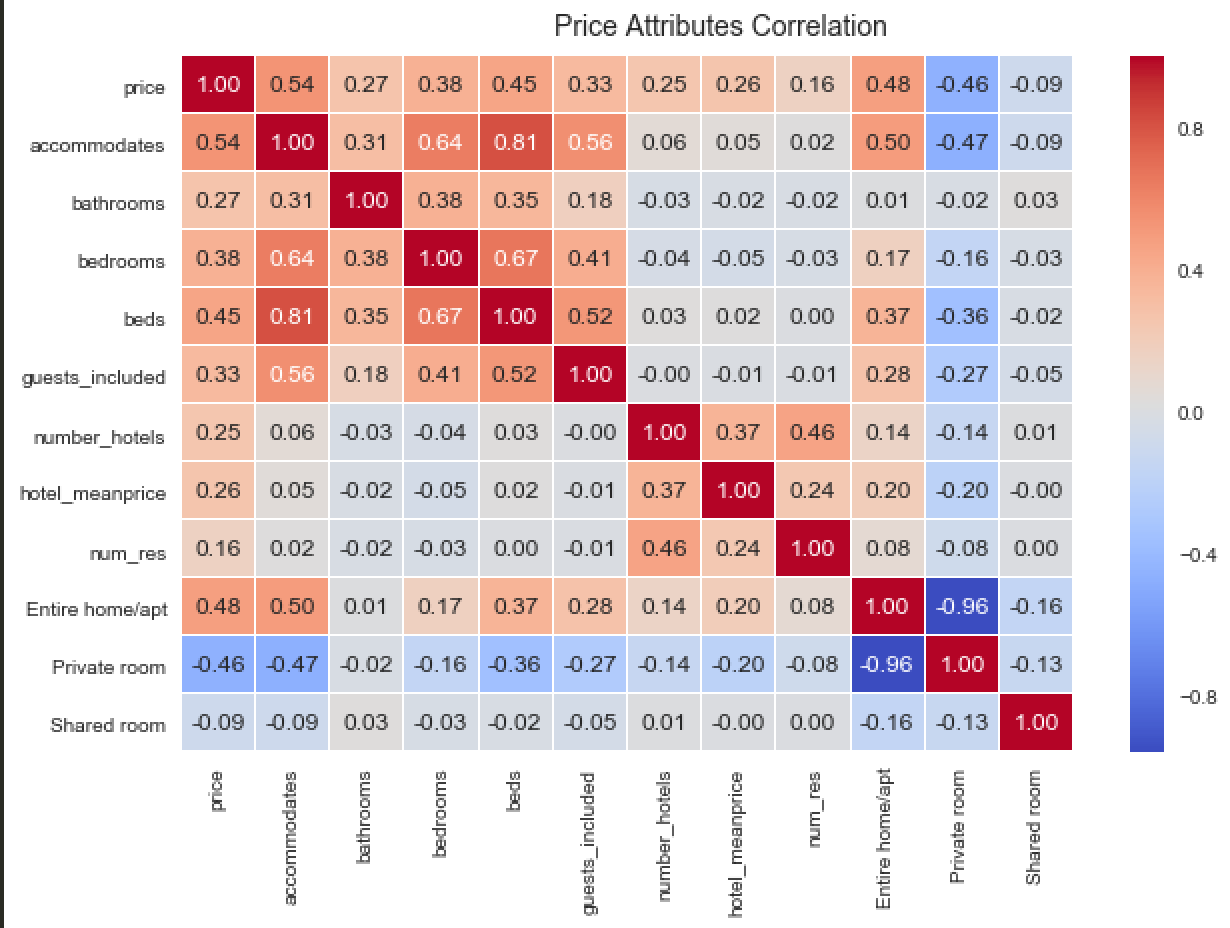


From above result, the p-value < 0.5 provides evidence to reject the null hypothesis that various room types in prices are independent.

## 

## Hypothesis 3: To test there is a significant linear relationship between some features and Airbnb prices.

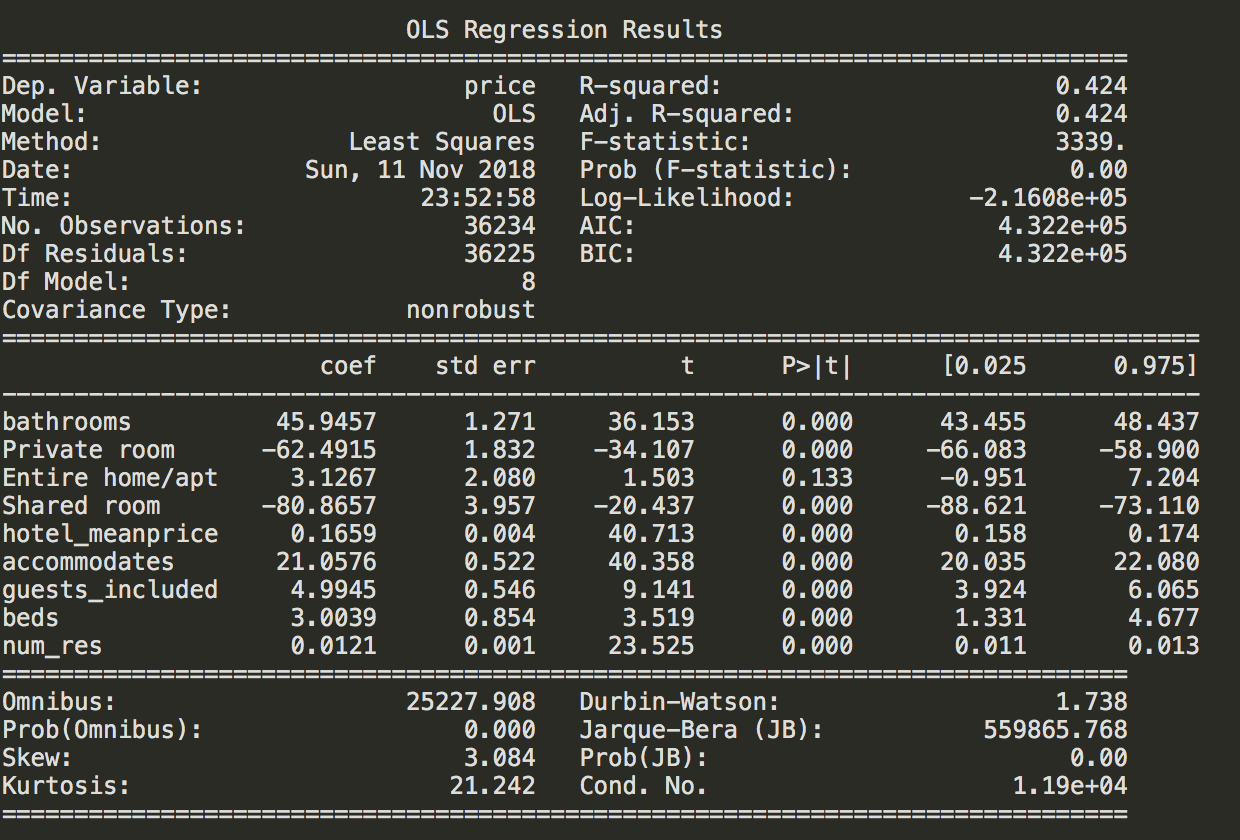
From hypothesis 2 and below heatmap, the attributes are selected based on the correlation coefficients.



Then, the hypothesis can be set up as follows:

* Null hypothesis: there exists a linear relationship between the Airbnb price and other attributes
* Alternative hypothesis: the Airbnb price cannot be presented by a linear regression model

From the below results, R-squared is less than 0.5 which means that the data is not very close to the fitted regression line.

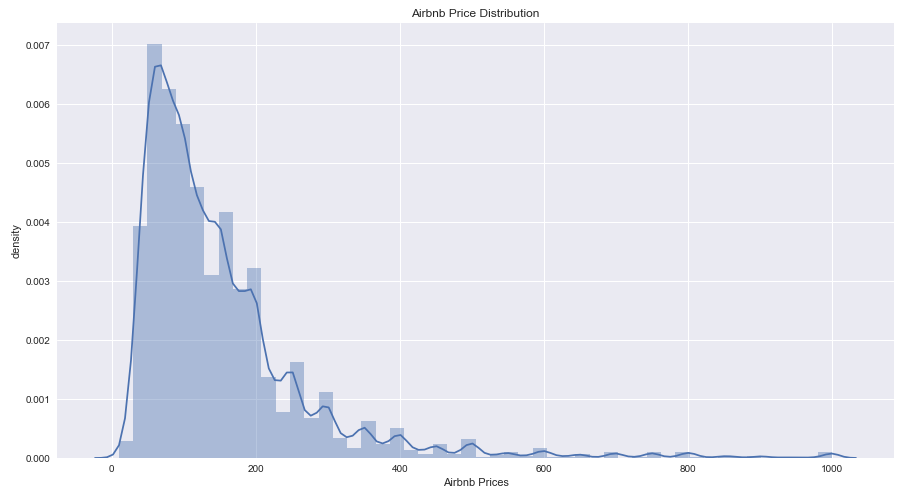


# Predictive Models

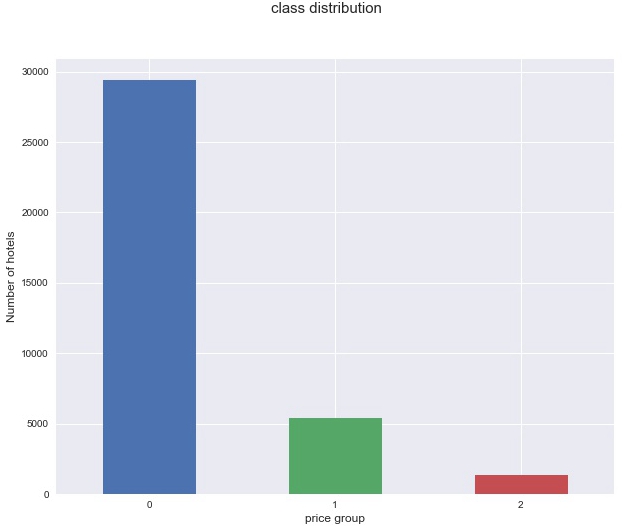
In this analysis, the predictive models are constructed by the five most commonly used classification methods, in order to predict the price group of Airbnb listings. The methods are Gaussian Naive Bayes, Decision Tree, K Nearest Neighbor, Support Vector Machine, and Random Forest. The classification results are in the following subsections.

## Preprocessing

In order to perform the classification methods on Airbnb dataset, preprocessing has been done to the dataset for dividing it into groups. The distribution of Airbnb prices is illustrated in the following figure.

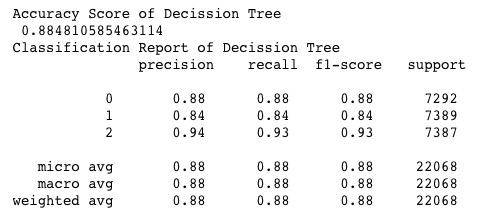


Based on the distribution, the prices are divided into three price groups, price group 0 with label A of price range from $0 to $200, price group 1 with label B of price range from $201 to $400, and price group 2 with label C of price range from $401 to the maximum price in listings. The distribution of the classes is illustrated in the following figure after grouping the prices.

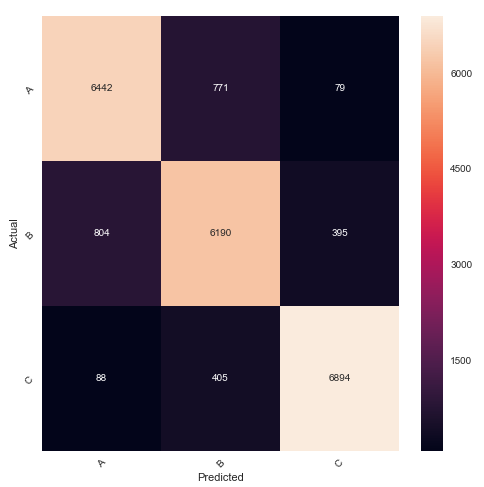


## Decision Tree Predictive Model

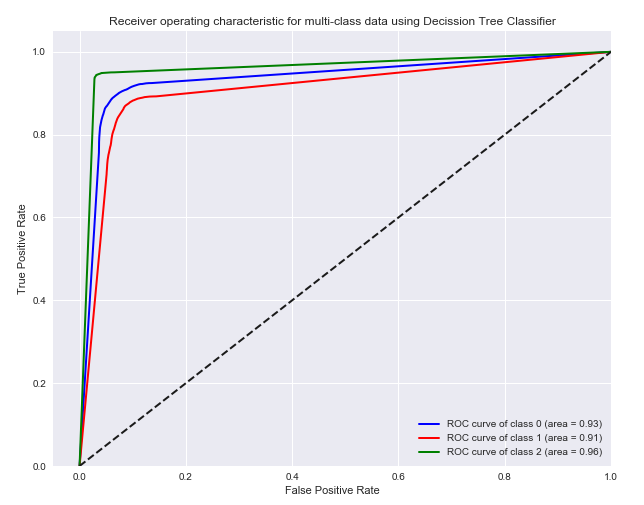
The Decision Tree algorithm builds a classifier that is a directional tree structure. The tree consists of root nodes, internal nodes, and leaf nodes, where the leaf nodes represent the class labels. The tree is created in a recursive fashion by continuing to partition the training features into purer subsets. Decision Tree is a non-parametric supervised classifier, which makes classification inexpensive to construct and it is extremely fast to classify the unknown records. It is also robust for missing values and does not get adversely affected by any redundant attributes. The performance of Decision Tree classifier to the Airbnb data is fairly good, with an accuracy score of 0.8849. By the classification report in the following figure, Decision Tree classifier classifies all three classes good and performs excellently on classifying class 2, with corresponding label C.



The confusion matrix and ROC curve plot of the Decision Tree classifier support the classification report. In the confusion matrix, we can clearly tell that price group C has been excellently classified by Decision Tree classifier, with 6894 out of 7837 samples. Price group A has 6442 out of 7292 samples been successfully classified, and price group B has 6190 out of 7389 samples been successfully classified.



The AUC of price group A, B, and C are 0.93, 0.91, and 0.96, which supports the interpreted results. The performance of Decision Tree classifier is a litter bit off by comparing with Random Forest classifier in the later subsection, but better than the other three classifiers that we will use for predictive models.

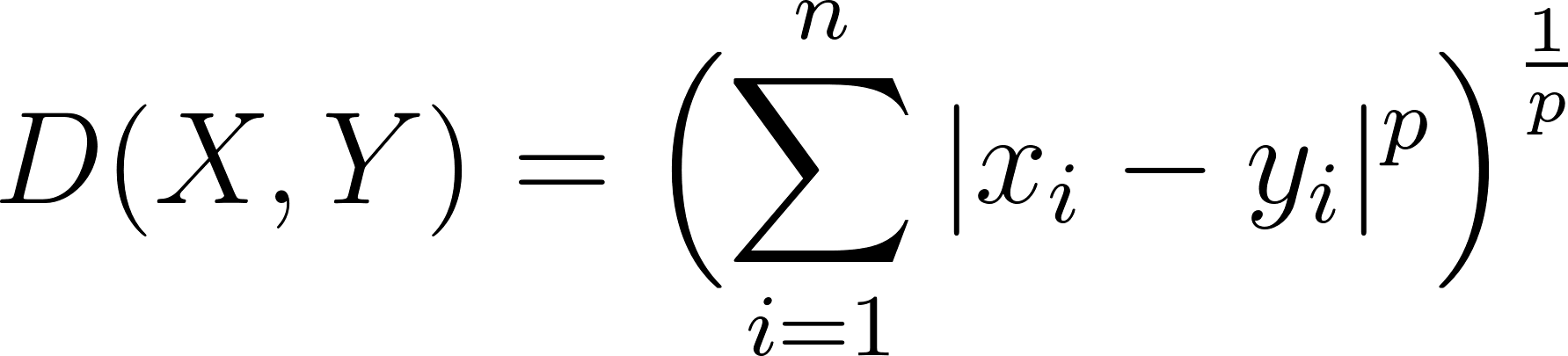


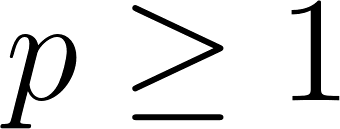
## K Nearest Neighbor Predictive Model

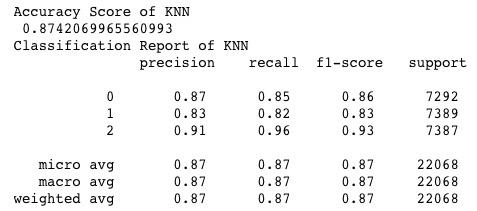
K Nearest Neighbor (KNN) classifier is one of the instance-based classifiers. It is a type of Lazy Learner methods, which does not attempt to construct a general internal model, but simply stores instances of the training data. KNN classifier uses k closest points for performing classification, the basic idea is to choose k of the nearest records out of the training records and then compute the distance of them to the test records. It requires the set of stored records, the metric of distance, and the value of k to perform the classification. To classify an unknown record, KNN classifier computes the distance of the unknown record to other training records, then identifies k nearest neighbors of the unknown record, and at last uses class labels of nearest neighbors to vote the class label of this unknown record. The KNN predictive model for Airbnb dataset uses the Minkowski distance as a metric. The Minkowski distance of order [](about:blank) between two points

[](about:blank)

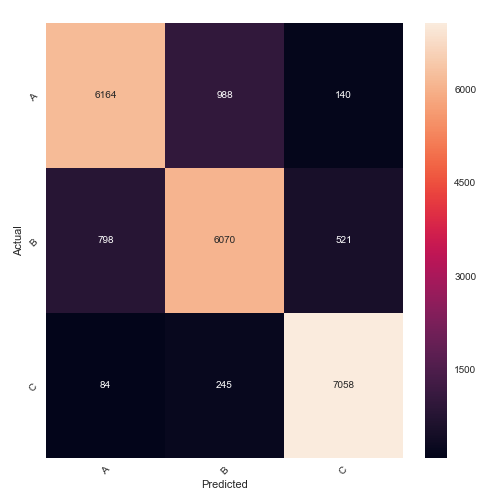
is defined by

[](about:blank).

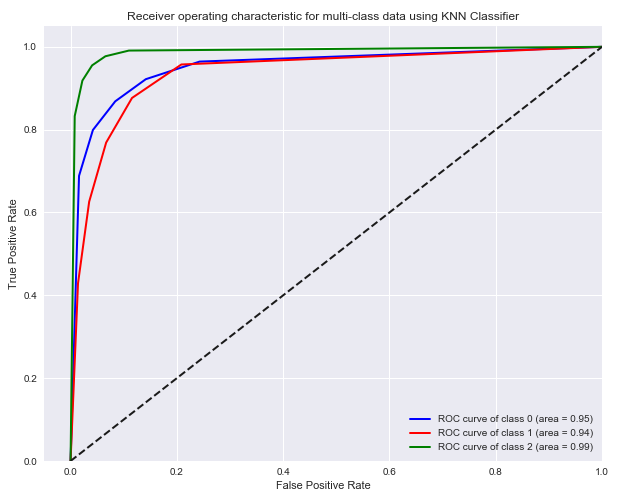
The Minkowski metric uses Minkowski distance with [](about:blank). KNN classifier performs well on Airbnb dataset, with an accuracy score of 0.8742. From the classification report in the figure below, it can be seen that the KNN classifier predicts good results for class 2, which is the price group C, with a 0.91 accuracy.



The confusion matrix of KNN predictive model shows that it classifies 7058 out of 7387 testing samples of price group C successfully, 6164 out of 7292 testing samples of price group A and 6070 out of 7389 testing samples of price group B successfully.



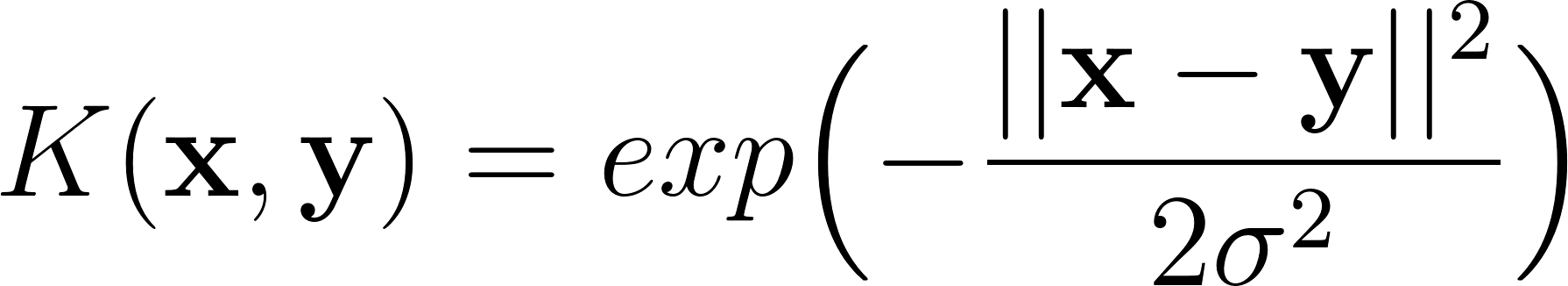
The ROC curve shows that the classification to price group C has an AUC of 0.99, to price group A has an AUC of 0.95, and to price group B has an AUC of 0.94.



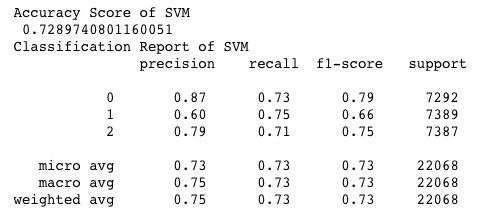
To the unknown records, KNN classifier works in a relatively expensive manner. Comparing with all five classification methods that were used, KNN classifier performs good but is relatively time consuming. The accuracy of KNN is a little bitter lower than Decision Tree, but much higher than the Support Vector Machine classifier in the next subsection.

## Support Vector Machine Predictive Model

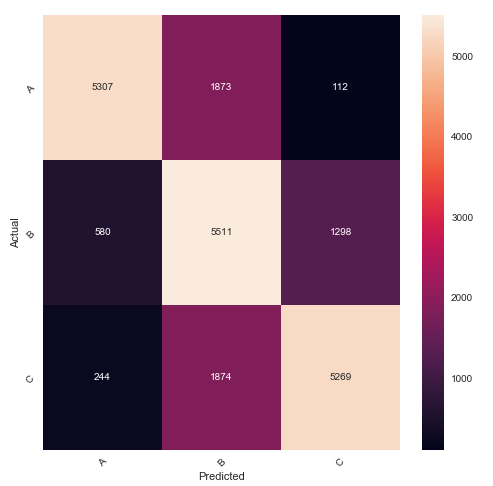
Support Vector Machine (SVM) is a popular supervised machine learning algorithm. It performs feature transformation into infinite Hilbert Space, therefore the input vectors can be separated by hyperplanes. The goal of SVM is to maximize the margin between the classes. To predict the price group of the Airbnb listings, we use the radial basis function (RBF) kernel. The RBF kernel on two samples [](about:blank) and [](about:blank) that represent the feature vectors is defined as

[](about:blank),

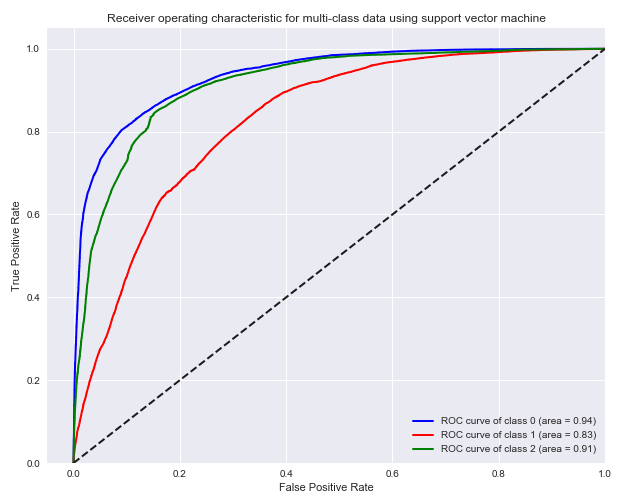
where [](about:blank) is a free parameter. The SVM classifier performs ordinary on the Airbnb dataset, with an accuracy score of 0.7290. SVM classifies class 0 relatively good compare to class 1 and class 2. The results are illustrated in the following figure.



The confusion matrix of SVM predictive model shows that it classifies 5307 out of 7297 testing samples of price group A successfully, 5511 out of 7389 testing samples of price group B and 5269 out of 7387 testing samples of price group B successfully.



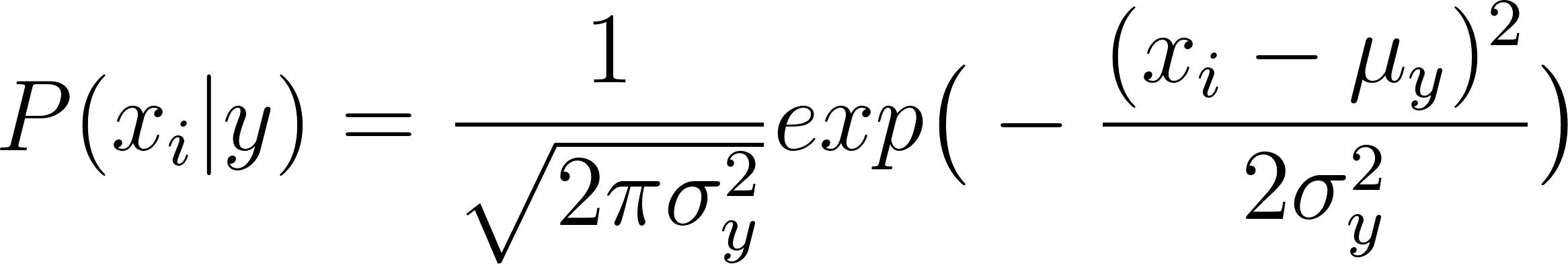
The ROC curve shows that the classification to price group A has an AUC of 0.94, to price group B has an AUC of 0.83, and to price group C has an AUC of 0.91.



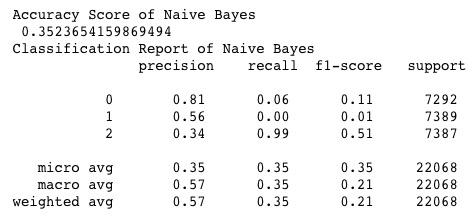
SVM works effectively in high dimensional spaces, especially when the dimension is greater than the number of samples. But if the features are far greater than samples, it performs poorly. Since SVM has no probability estimates, the probability must be computed using at least 5-fold cross-validation. This makes the training for SVM predictive model on Airbnb dataset extremely time consuming and is far slower than all the other methods that are used.

## Naive Bayes Predictive Model

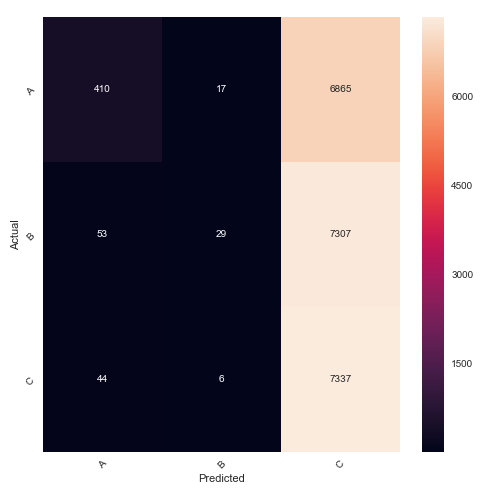
Naive Bayes is a family of algorithms that assume applies Bayes’ theorem with the assumption of conditional independence between every pair of features, in other words, assumes the features of the predictive model are independent to each other. It is a set of supervised learning algorithms, such that it needs predefined labels to the samples. The Gaussian Naive Bayes algorithm for classification, as the name implies, assumes the likelihood of the features to be Gaussian. In the following equation, the probability of [](about:blank) given [](about:blank) is computed by estimating y and y with maximum likelihood.

[](about:blank)

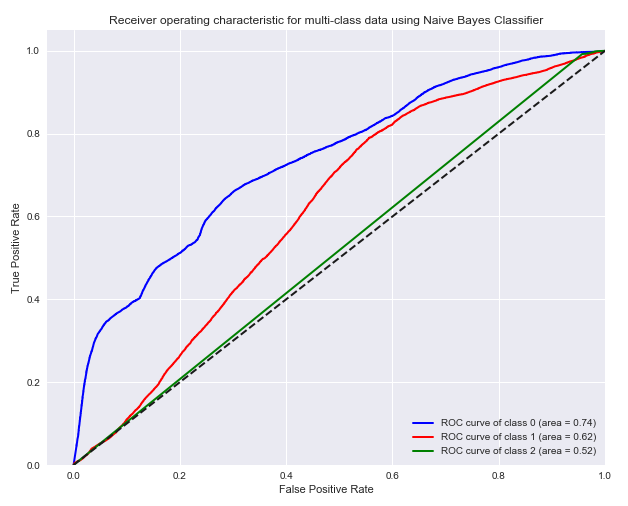
The Gaussian Naive Bayes classifier is robust to isolated noise points, and irrelevant attributes. It handles missing values by ignoring the instance during probability estimate calculation. But for some data, the independence assumption may not hold for all attributes. The performance of the Gaussian Naive Bayes classifier on Airbnb price dataset is disappointing, with an accuracy score of 0.3524. From the classification report in the figure below it can be seen, Gaussian Naive Bayes successfully classifies class 0, which is the lower price range group in Airbnb dataset, with 0.81 precision; it poorly classifies class 1 and class 2, which are two higher price range Airbnb listings, with only 0.56 and 0.34 precisions.



The confusion matrix of predicting results of Gaussian Naive Bayes classifier illustrated in the following figure gives much clearer sense of how it miss-classified the testing data. The corresponds labels to class 0, 1, 2 are price group C, A, and B.

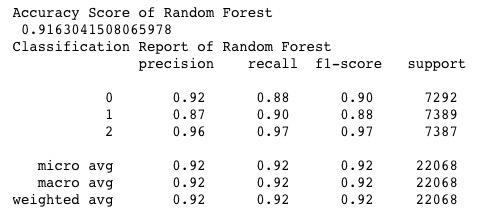


The ROC curve shows that the classification to price group A has an AUC of 0.74, to price group B has an AUC of 0.62, and to price group C has an AUC of 0.52. The Gaussian Naive Bayes classifier is the the one with worst performance on Airbnb dataset.

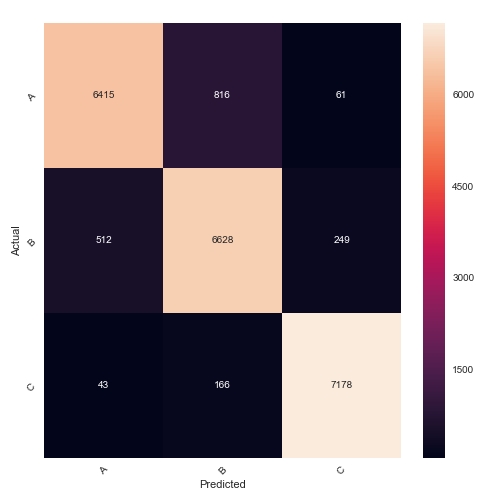


## Random Forest Predictive Model

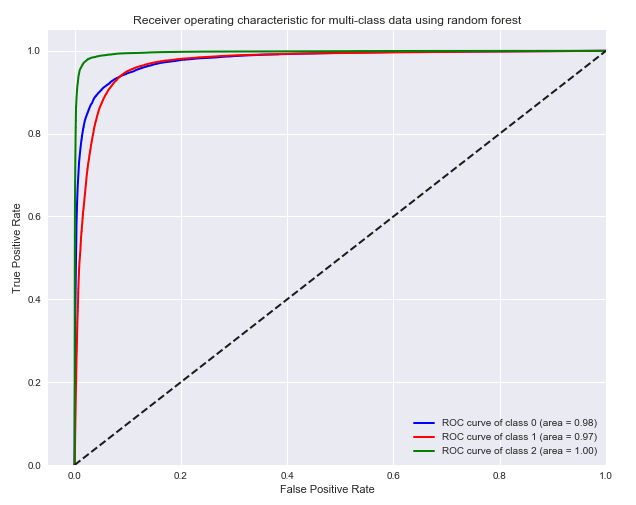
The Random Forest classifier grows multiple trees based on training data. It classifies a new record by offering a classification from each tree, then output the results based on the majority classification. Recall the Decision Tree methods, Random Forest creates a forest of decision trees and each of them votes on the classification, and the output of each testing sample is the class that wins most of the votes. Similar to SVM, Random Forest can handle large datasets with high dimensionality, but with much better training and testing time. It uses bootstrap sampling, also known as bagging, to the training data, which selected the whole training set with replacement. By averaging the various sub-samples of the training set, Random Forest improves the predictive accuracy and controls overfitting problem. Bagging reduces variance and it is very efficient with strong learner systems with low bias.



The confusion matrix of Random Forest predictive model shows that it classifies 6415 out of 7292 testing samples of price group A successfully, 6628 out of 7389 testing samples of price group B and 7178 out of 7387 testing samples of price group C successfully.



The ROC curve shows that the classification to price group A has an AUC of 0.98, to price group B has an AUC of 0.97, and to price group C has an AUC of 1.00. The Random Forest classifier is the one with outstanding accuracy score among all five classifier that used on the Airbnb dataset.



While training the predictive model, Random Forest classifier computes the feature importance scores for all features. The results in the following figure illustrate the most important features that contribute to the classification results. The room type of a listing mainly dominates the price, even more important than the hotel prices around in the same area. Another interesting important feature is “accommodates”, since it implies the area of a listing. By selecting important features, there can be simpler predictive model that performance the same but with less expensive training cost. The full table of feature importance and thresholds of accuracy are stored for possible further exploration in later part of this project.

