

How to get higher income on airbnb

Exploratory data analysis on airbnb dataset

Name/Member: xxx

Identity number: xxx

Date: xxx

[Executive Summary 3](#_Toc43537536)

[Business Problem 3](#_Toc43537537)

[Data Mining Technology 3](#_Toc43537538)

[Model Outcomes 3](#_Toc43537539)

[Insights & Suggestions 3](#_Toc43537540)

[1. Business understanding 3](#_Toc43537541)

[1.1. Research object 3](#_Toc43537542)

[1.2. Target Audience 3](#_Toc43537543)

[1.3. Specific business issues 4](#_Toc43537544)

[1.4. Further research on price 4](#_Toc43537545)

[1.5. Further research on demand 5](#_Toc43537546)

[2. Data overview and preprocessing 5](#_Toc43537547)

[2.1. which variables to use 5](#_Toc43537548)

[2.2. Handling special columns 6](#_Toc43537549)

[2.3. Handling missing values 6](#_Toc43537550)

[2.4. Further processing 6](#_Toc43537551)

[2.5. Variable overview 7](#_Toc43537552)

[3. Modelling & Evaluation 7](#_Toc43537553)

[3.1. Linear regression 7](#_Toc43537554)

[3.1.1. Model introduction 7](#_Toc43537555)

[3.1.2. Regression result 7](#_Toc43537556)

[3.1.3. Result analysis 8](#_Toc43537557)

[3.1.4. Model diagnosis 9](#_Toc43537558)

[3.2. Decision tree 10](#_Toc43537559)

[3.2.1. Model introduction 10](#_Toc43537560)

[3.2.2. Feature Engineering 10](#_Toc43537561)

[3.2.3. Main results of the model 11](#_Toc43537562)

[3.3. Comprehensive model evaluation 11](#_Toc43537563)

[3.3.1. ROC curve 11](#_Toc43537564)

[3.3.2. AUC value 12](#_Toc43537565)

[3.3.3. Contrast diagnosis 12](#_Toc43537566)

[3.4. Clustering 12](#_Toc43537567)

[3.4.1 Cluster analysis 12](#_Toc43537568)

[3.4.2. Principal component analysis 13](#_Toc43537569)

[3.4.3. Analysis results 13](#_Toc43537570)

[4. Insights and Implications 14](#_Toc43537571)

[4.1. Evaluate Business Problem 14](#_Toc43537572)

[Problem1: What factors affect airbnb prices 14](#_Toc43537573)

[Problem2: What factors can be improved by the landlord to increase the price of airbnb 14](#_Toc43537574)

[4.2. Recommendations & Suggestion 15](#_Toc43537575)

[Appendix A：Descriptive statistics of selected variables 15](#_Toc43537576)

[References 17](#_Toc43537577)

# Executive Summary

## Business Problem

In response to this airbnb data, the business problem we are going to solve is "how to let airbnb landlords get higher profits". Based on our business problems, we will analyze what factors affect the price of airbnb, and put forward how to let the landlord take airbnb in combination with the actual situation. Sell ​​better prices.

## Data Mining Technology

The data mining techniques that can help us solve business problems are regression analysis, decision tree analysis and cluster analysis. We first use regression analysis to analyze the relationship between the selected variables and airbnb prices, and study the direction and significance of the impact. Second, we use the decision tree method to pick out some of the variables that have the greatest impact on airbnb prices, and use the model to quantify this impact. Finally, we use a clustering method to classify airbnb into different categories based on airbnb prices, facility conditions and service levels, and explore the differences between the different categories.

## **Model Outcomes**

Regression analysis, decision tree analysis and cluster analysis yielded similar results. The biggest factor in deciding the price of airbnb is the basic condition of airbnb, that is, the type and size of the room; secondly, the equipment attached to the room such as washing machine, dryer, elevator Etc. can also affect the price of airbnb, and finally, the service level of airbnb also has a certain impact on its price

## I**nsights & Suggestions**

After analyzing each data analysis model, we provide three directions on how to improve the income and profit of airbnb landlords: first, focus on improving basic conditions, second, provide as many facilities as possible, and finally, provide better Service Level.

# 1. Business understanding

## 1.1. Research object

Airbnb, as a leader in short-term rental trends, successfully attracted imitation of many products around the world, and also triggered the rise of professional short-term rental companies. Moreover, it successfully expanded the sharing economy model from the original basic industry to the real estate market. Through continuous efforts, Airbnb has successfully created a fresh and elegant product image (for example, I found in the research that in the early stages of development, some merchants found that beautifully photographed rooms are more attractive to users, so they asked photographers to take pictures and gradually Formed the brand image of the product) This is a business operation case worth analyzing.

## 1.2. Target audience

There are three main categories of Airbnb participants: travelers, landlords and other experience providers.

**Travelers**

Basically the same as the users of ordinary travel products. They are short-term travelers who need to rent a house at the destination for a short time. Specially, airbnb travelers have higher demand and knowledge about accommodation than ordinary travelers, and they are tired of the sense of mission, strong purpose and standard services of traditional hotels or hotels. The leisure, from the original exploratory travel (curiosity for things I haven’t seen) to experiential travel (curiosity for unexperienced lifestyles), from a business analysis point of view this is a This kind of consumption upgrade is reflected in tourism.

**Landlord**

A group who owns an idle house or room and wants to make a profit by renting out the idle space. Due to the high requirements of Airbnb traveler users and the limited energy of individual landlords, at this stage, a group of companies focused on short-term rentals have been created worldwide to standardize the design of homestays.

**Experience provider**

Refers to local residents who provide experience services. They have some local cultural special skills or are familiar with local characteristics. They can bring users an interesting experience of local culture, and use it to earn profits and make friends.

**Target Audience**

Obviously, selecting the target audience needs to stand from the perspective of data. The data provided by the course should be data get from the airbnb website. There are 99 variables, which are cross-sectional data for airbnb, so the question must be raised from the standpoint of these airbnb owners, the landlord mentioned above. To raise business issues on the landlord's standpoint, the first thing we noticed was how to increase the income brought to them by airbnb short-term rental housing.

## 1.3. Specific business issues

Research on this issue is divided into two angles here: price and demand. That is to say, to make an airbnb short-term rental house get the most benefit for the landlord, the answer to this question should be determined by the airbnb price and the traveler's demand for airbnb. Although there may be other influencing factors in this process, it is clear that these two are the main factors.

## 1.4. Further research on price

It is relatively easy to study the impact of price, because the data clearly provides the price of each airbnb, including various service fees that may be incurred. There are many theories and methods to study the influence of other variables on prices or to explore the relationship between them. For example, two methods that are feasible and will be adopted in this article are: using regression methods and decision tree methods to study the correlation between price variables and other variables.

These two methods are used in this article to model and analyze the relationship between variables. A more important question is how to combine price with other variables and produce specific results together. Since most of them are numeric variables or can be converted into numeric variables, the linear regression method is more appropriate. This article will use a linear model to create a model of the relationship between multiple independent input variables (characteristic variables) and output dependent variables . The model remains linear, and the output is a linear combination of input variables.

For the decision tree method, we choose to divide the price variable according to the interval into different ordinal variables. Due to the large amount of data in a given data set and the low information density of some data, we choose the CART algorithm in the algorithm, because the algorithm can choose the right time to end. In other words, ideally, the classification should stop when there is only one category in each leaf node, but a lot of data is not easy to completely divide, or complete division requires many splits, which will inevitably cause a long running time, so CART can Analyze the mean variance of the data in each leaf node. When the variance is less than a certain value, the split can be terminated in exchange for a reduction in computational cost.

## 1.5. Further research on demand

It is more difficult to study the impact of demand. The main reason is that the data provided does not contain many variables related to demand. Another important reason is that the demand is flexible. The revenue and price change in the opposite direction, that is, the total revenue decreases with the increase of the price and decreases with the decrease of the price; for goods with inelastic demand, the sales revenue and the price change in the same direction, that is, the total revenue varies with the price Increase with the increase and decrease with the price.

Therefore, the demand itself and its related factors are a problem that is difficult to obtain results using fixed data research methods. Fortunately, for this problem, we have some methods dedicated to solving this type of problem, such as clustering methods As an important method of unsupervised learning, the idea of ​​clustering is to classify samples with similar attributes into one class. For each data point, we can classify it into a specific class, and at the same time all data points between each class have a certain degree of commonality. In the algorithm logic, it may be that the spatial position is close to the other characteristics. In business logic, there is convergence between certain variables, which can be sorted out with the help of algorithms to highlight the influence of the main variables.

# 2. Data overview and preprocessing

In the data conversion stage, we perform sampling processing, type conversion, and normalization on the data. Sampling is the process of extracting sample points from a specific probability distribution. Sampling has a very important application in machine learning: simplifying complex distributions into discrete sample points; re-sampling can adjust the sample set to better adjust and adapt to later model learning; used for random simulation for complex Approximate solution or reasoning of the model. An important role of sampling is to deal with unbalanced data sets.

But before performing these tasks, manually selecting the required data columns is an important task. The processing method of special columns and the processing method of missing values will also significantly affect the quality of the analysis results

## 2.1. which variables to use

According to the analysis method we selected, combined with the specific situation of the airbnb data set, we will not analyze the variables, such as "ID\_airbnb", "title\_airbnb", before entering the programming language processing, delete them.

## 2.2. Handling special columns

Data types can be simply divided into numeric and non-numeric types. Numerical types are continuous and discrete. Non-numeric types include categorical types and non-categorical types. Among the categorical features, if the sorting problem exists in the category, it is the ordered type. If there is no sorting problem, it is the fixed type. The non-category type is the string type.

For non-numeric types, we need to perform category conversion, that is, convert non-numeric types to numeric types to facilitate subsequent processing by machine learning algorithms.

For the sequenced type, we can use the serial number encoding, such as the "refund\_rule" variable, which can be assigned a numeric ID according to the corresponding relationship of: 엄격：3、일반：2、유연：1, and the level relationship is still retained after conversion.

For date variables, such as: "4월2013", the actual meaning of the variable is: registered in April 2014. To make the variable meaningful, we calculate the difference between this date and January 2019. such as 4월2013, it is: (2018-2013)\*12+(13-4)=69, the actual meaning is that the registration time is 69 months ago.

## 2.3. Handling missing values

The missing value here needs to be understood from two perspectives. The first is the missing value of the variables. Some variables have many missing values in the statistical process. We believe that the limit of this problem is 80%. If the missing value of a variable reaches 80%, then analyzing the relationship between this variable or other variables and it may not get the desired effect, so we should delete these variables.

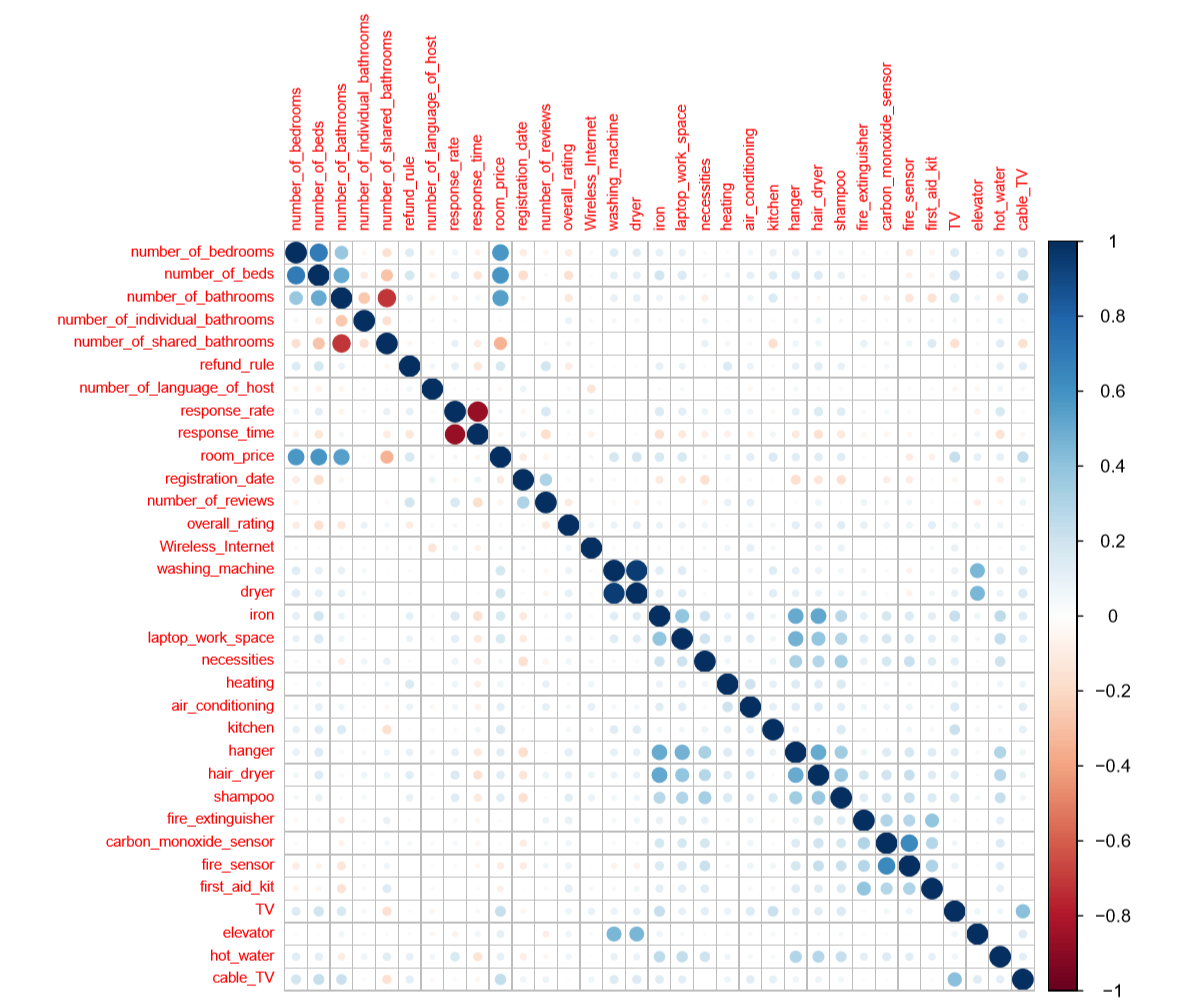
The second is the missing value of the observation, that is, some variables of each airbnb may be missing. Obviously, these missing in the algorithm will cause this observation to be unable to be combined, and it will also cause the overall performance of the algorithm to decrease. Therefore, our strategy for this missing value is that if an observation still has missing values after the previous step, we delete the observation.

## 2.4. Further processing

After careful observation of the data set provided, we found that there is a problem with the data set processed in the above two steps. The variable ranked lower in the data set is, according to its actual meaning, a variable that measures whether the airbnb provides some additional functions or equipment, but after a preliminary analysis of variance, we found such a problem that some of the variables exceed 90% The observations of are all 0, which means that there are very few airbnb providing the service or equipment. Obviously such variables will not have any positive effect on the model (mainly may cause overfitting, or do not meet statistical significance Requirements), so we imitate the first step and set the threshold to 80%, which means that if the proportion of airbnb providing these devices or services reaches 80%, this variable can participate in the comprehensive analysis.

## 2.5. Variable overview

The descriptive statistics of each variable are attached to Appendix A. Below is the correlation statistics between each variable.



# 3. Modelling & Evaluation

## 3.1. Linear regression

### 3.1.1. Model introduction

The multiple linear regression model is usually used to describe the random linear relationship between the variables y and x, that is:

In the formula, x and y are the independent and dependent variables of the regression analysis, β is the regression coefficient, and ζ is the random error variable.

### 3.1.2. Regression result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Pr(>|t|) |  | Estimate | Pr(>|t|) |
| number\_of\_bedrooms | 104977.2 | < 2e-16 | dryer | 14605.9 | 0.66068 |
| number\_of\_beds | 38049.9 | 2.38E-08 | necessities | 21872.7 | 0.36092 |
| number\_of\_bathrooms | 166792.2 | < 2e-16 | heating | 27223.9 | 0.37728 |
| number\_of\_individual\_bathrooms | 97187.1 | 7.42E-06 | refund\_rule | 25133.1 | 0.00224 |
| number\_of\_shared\_bathrooms | 28174.2 | 0.08555 | kitchen | 21596.2 | 0.2529 |
| fire\_extinguisher | 5238.2 | 0.66329 | hanger | -4404.2 | 0.8047 |
| number\_of\_language\_of\_host | 2154.9 | 0.61343 | hair\_dryer | 28837.0 | 0.08632 |
| response\_rate | -18055.2 | 0.7621 | shampoo | 16075.5 | 0.2698 |
| response\_time | -238.4 | 0.74512 | overall\_rating | 71864.2 | 3.92E-06 |
| registration\_date | -66.9 | 0.78808 | iron | -5707.2 | 0.70335 |
| number\_of\_reviews | -233.1 | 0.0243 | fire\_sensor | -40615.4 | 0.0274 |
| air\_conditioning | 28658.3 | 0.14753 | first\_aid\_kit | -13622.7 | 0.25908 |
| Wireless\_Internet | 4086.8 | 0.93758 | TV | 22721.1 | 0.08468 |
| washing\_machine | -7725.0 | 0.81678 | elevator | 39257.0 | 0.00198 |
| laptop\_work\_space | 18670.6 | 0.15657 | hot\_water | 16314.1 | 0.15229 |
| carbon\_monoxide\_sensor | 11306.5 | 0.44175 | cable\_TV | 21171.1 | 0.08573 |
| (Intercept) | -541662 | 1.70E-06 |  |  |  |

In the regression results, the *Estimate* value can be regarded as the coefficient corresponding to the independent variable. The positive or negative value of this value indicates the positive or negative influence of the variable on the price, that is, under the statistical caliber of the sample data, if the Estimate is positive, then there is a positive correlation between this variable and the dependent variable. if the Estimate is negative, and there is a negative correlation between this variable and the dependent variable.

Pr's understanding involves hypothesis testing theory. In a relationship in which many independent variables affect a dependent variable, hypothesis testing theory can determine which (or which) independent variables have significant effects and which independent variables have insignificant effects. Then, in the final model, independent variables with significant impact can be added to the model, and variables with insignificant impact can be eliminated. According to relevant data, it is generally considered that the pr value of less than 0.05 indicates that the impact of the variable corresponding to the variable is significant.

### 3.1.3. Result analysis

Combining the above, we find that the variables that have a significant impact on airbnb prices are:

* number\_of\_bedrooms,
* number\_of\_beds
* number\_of\_bathrooms
* number\_of\_reviews
* refund\_rule, elevator
* overall\_rating
* number\_of\_individual\_bathrooms

Obviously, some basic conditions(such as the room type and area of arirbnb) can significantly affect the price of airbnb, but the impact of these variables on the price has no reference value for the landlord, because the landlord cannot change the basic conditions of this room under general conditions. Therefore, excluding the influence of these variables, the influence of the remaining variables on prices is our concern.

**Refund\_rule**

It can be seen that the refund policy has a strong correlation with hotel prices. This may be because a reasonable refund policy may stimulate users' interest in booking hotels and facilitate transactions.

**Overall\_rating**

Comprehensive ratings can affect the price of airbnb, and the two have a positive correlation, which seems not difficult to understand. The high rating may be due to better infrastructure and services. But this also gives the landlord some hints. Although the basic conditions of airbnb cannot be improved in a short time, through the service and care of customers, the score can be improved to a certain extent, which can be translated into an increase in price.

**Elevator**

Regarding the strong positive correlation between elevators and prices, we believe that there are some mysteries. Obviously, the airbnb elevator gives guests the most direct accommodation feeling. If the taste in an elevator is too unpleasant, then the guest’s impression of the hotel will be greatly reduced, and the question of whether the hotel’s hygiene is really clean will be left in mind. It can be seen that there are many optional devices in the variable list, such as iron, hair\_dryer, shampoo, but the importance of elevators is among the best.

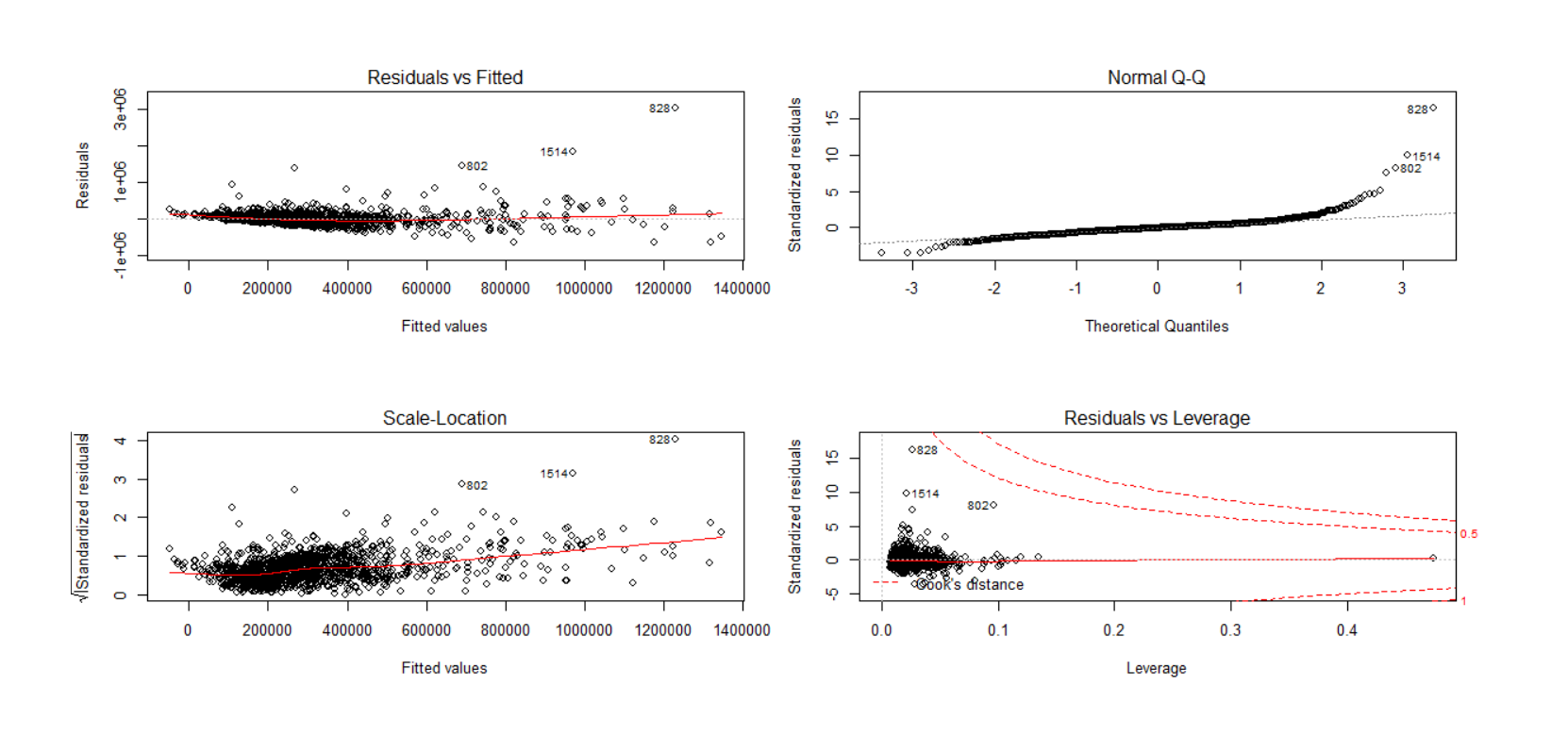
**Number\_of\_reviews**

This is the only variable that has a negative correlation among several highly correlated variables. We think this may be due to the fact that a hotel with too high a price will have fewer occupants, so there will not be too many reviews. This also enlightens us: For some high-end hotels with few occupants, try to let every occupant comment, which is a good window to display high-end airbnb, which may greatly increase the hotel’s exposure and thus affect price.

### 3.1.4. Model diagnosis

Linear regression can also be used when the independent variable is categorical and the dependent variable is continuous. Only more often, this type of analysis is more concerned with differences between groups than linear regression prediction, usually using analysis of variance or t test, especially when there are only one independent variable. The model assumes that different groups are sampled from the same population, the residuals of each group (strictly speaking, should be various variables) follow the same normal distribution, and the residuals of different groups all follow the same mean and the standard deviation is σ ^2 Normal distribution. In actual investigations, we often directly investigate whether the dependent variable values corresponding to fixed independent variable values (different variables) are normally distributed.

Of course, we can also use linear regression for analysis. In order to eliminate the error caused by the assignment, the multi-class independent variables need to be set as dummy variables in the linear regression model, and the results are consistent with the analysis of variance. For the regression model obtained in the previous step, we made a regression diagnosis as follows:



**Normal distribution hypothesis**

Continuous variables often take many values. Even if our sample size is large enough, there are multiple values ​​for each fixed value of the independent variable. In this case, it is easier to return from the investigation of normality to the investigation of all residuals. Based on this assumption, the validity of the airbnb data is guaranteed.

**Outliers**

Because too large or too small data may affect the analysis results, especially when doing regression, we need to deal with those outliers. In the basic processing, we use a standardized method, which is used when multiple indicators of different magnitudes need to be compared with each other. It can be seen from the figure that the overall of airbnb data has fewer outliers.

**Linear model rationality**

Quantile regression can be understood as a weighted least squares estimate. The general least squares objective is to estimate the parameter expectation (mean), while quantile regression is to directly estimate the value of the quantile. The reason for quantile regression is because we do not want to study the expectations of the parameters only, but want to explore the complete distribution of the parameters, or in some cases we may want to know a certain quantile of the parameters. Observing the QQ chart, we can see that in the main area, the quantiles are expected to obey the population, and the regression model is reasonable.

## 3.2. Decision tree

### 3.2.1. Model introduction

Decision tree is a basic classification and regression method. It is a tree structure describing the classification of instances. When using the classification tree for classification, start from the root node, test a certain feature of the instance, and assign the instance to its child nodes according to the test results. At this time, each sub-node corresponds to a value of the feature. Recursively test and assign instances until the leaf node is reached. Finally, classify the instances into leaf nodes.

### 3.2.2. Feature Engineering

In the data set, the dependent variable is a continuous attribute as price. According to the division conditions of the decision tree, it is to select a most suitable characteristic attribute, and then divide the set into multiple sub-sets according to the different values ​​of this characteristic attribute, and repeat the process of this operation continuously. The appearance of continuous attributes obviously affects this process. But we also want the price attribute to participate in the establishment of the decision tree, and we process it to meet the needs of the decision tree

For example, in a rectangular coordinate system, there are countless points on the x-axis. These points are like the corresponding density values. We cannot calculate the total number, so we cannot use this point to divide. However, there are also y-axes in the rectangular coordinate system. The points on the left of the y-axis are all negative, and the points on the right of the y-axis are all positive, so it can be said that the y-axis divides the x-axis into positive and negative parts. Then we can use this division method to divide the density value into multiple sub-regions, so that the sub-region of the density value is limited, so we can calculate it.

According to this theory, we divide the price variable, we calculate the quartile of the price variable, divide the overall interval into four parts, and convert it into a categorical variable for further calculation.

### 3.2.3. Main results of the model

|  |  |
| --- | --- |
| Variable | importance |
| number\_of\_bedrooms | 40 |
| number\_of\_bathrooms | 25 |
| number\_of\_beds | 19 |
| registration\_date | 4 |
| number\_of\_shared\_bathrooms | 3 |
| washing\_machine | 3 |
| dryer | 2 |
| elevator | 1 |
| number\_of\_reviews | 1 |
| number\_of\_individual\_bathrooms | 1 |
| hot\_water | 1 |

It can be found that the decision tree analysis and regression analysis are roughly the same. The decisive role is the basic condition of airbnb, followed by additional facilities and some service levels.

## 3.3. Comprehensive model evaluation

### 3.3.1. ROC curve

The ROC curve is a curve drawn according to a series of different two classification methods (cutoff value or decision threshold), with the true positive rate (sensitivity) as the ordinate and the false positive rate (1-specificity) as the abscissa. Traditional diagnostic test evaluation methods have a common feature, and the test results must be divided into two categories before statistical analysis. The evaluation method of the ROC curve is different from the traditional evaluation method, without this limitation, but according to the actual situation, an intermediate state is allowed, and the test results can be divided into multiple ordered categories, such as normal, roughly normal, suspicious, roughly abnormal and Statistical analysis will be conducted for five abnormal levels. Therefore, the scope of ROC curve evaluation method is more extensive.

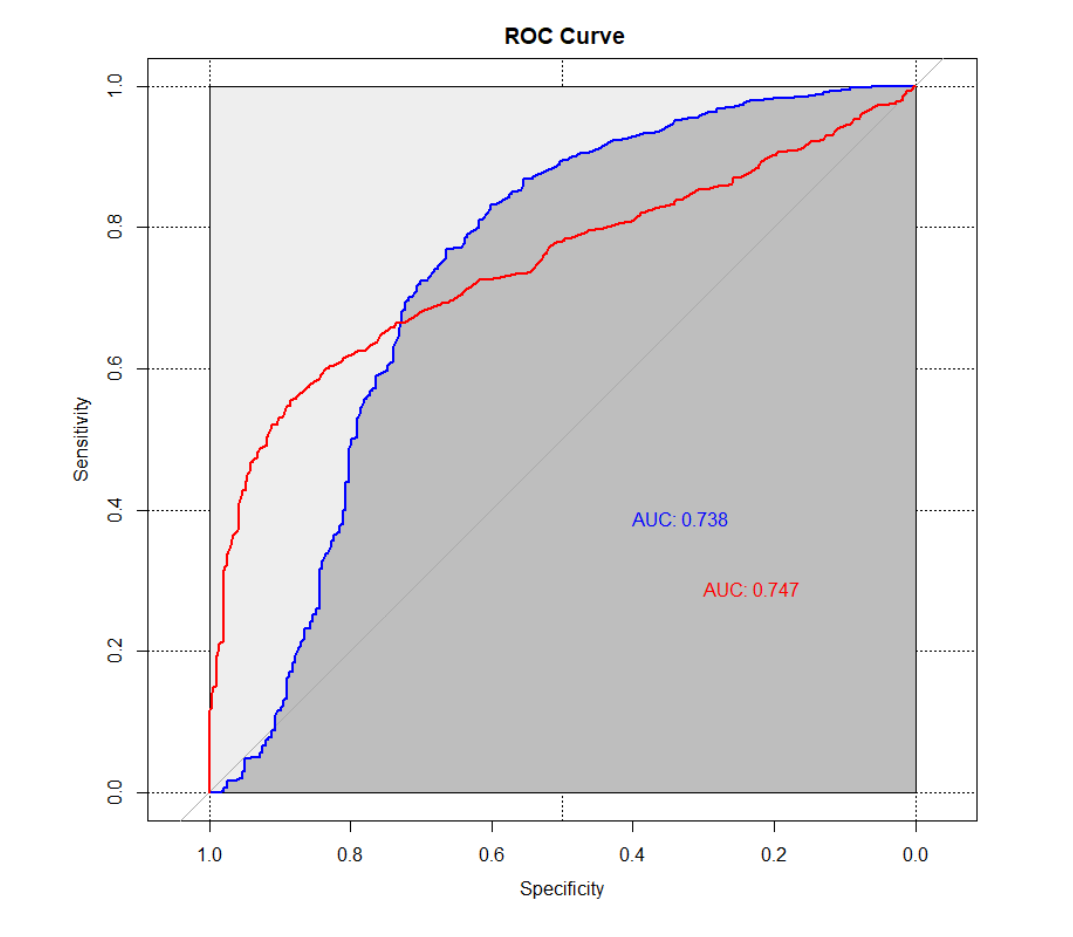
### 3.3.2. AUC value

AUC is a measure similar to the consistency or c-statistic of a binary model. This is the probability that observations with a positive category will have a greater predicted probability than observations in a negative category. If AUC = 1, it means that the model has a perfect prediction. If AUC = 0.5, it means that the model cannot be classified. Its behavior in logistic regression is similar to r^2, because adding more predictors increases AUC. Therefore, it is very important to include cross-validation or verification of external data in the analysis.

Our goal of building the model is higher and higher AUC

### 3.3.3. Contrast diagnosis

The ROC curves of the two models are as follows, where the blue curve represents the linear regression algorithm and the red curve represents the decision tree algorithm



It can be seen that the AUC values of the two algorithms are in the range of 0.73-0.75, indicating that the performance of the two algorithms is similar, and overall the performance is good.

## 3.4. Clustering

### 3.4.1 Cluster analysis

Cluster analysis is to group data objects according to the information describing the objects and their relationships found in the data. The purpose is that the objects in the group are similar to each other (related), while the objects in different groups are different (unrelated). The greater the similarity within the group and the greater the gap between the groups, the better the clustering effect.

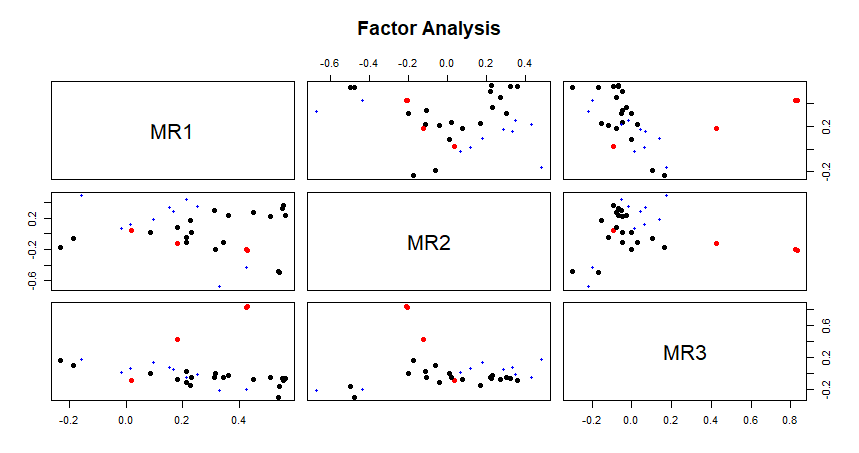
Cluster analysis can naturally group variables to discover natural categories and extract the characteristics of variables, so that we can further understand the relationship between observations.

### 3.4.2. Principal component analysis

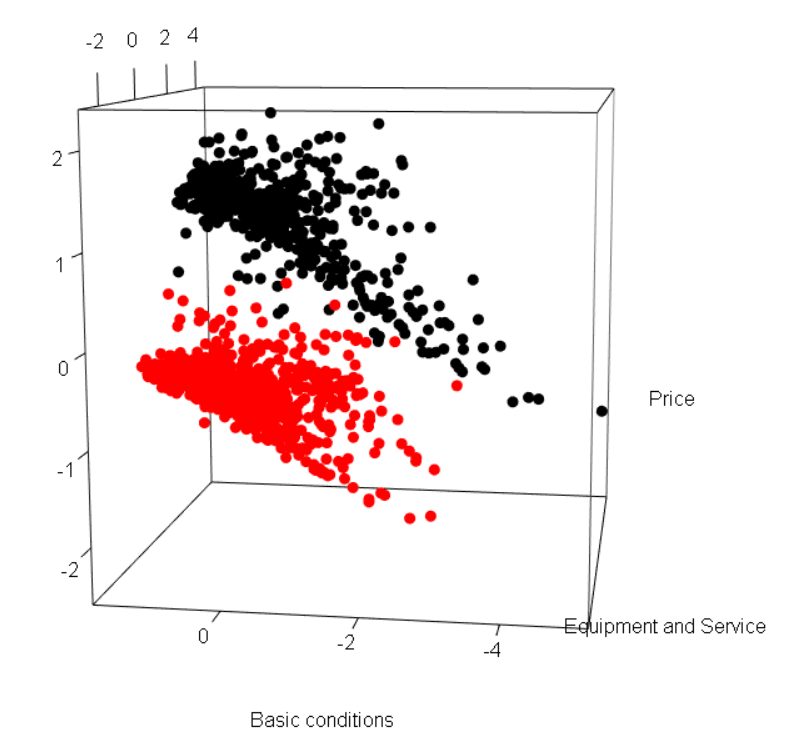
Principal component analysis is a multivariate statistical method. It is one of the most commonly used dimensionality reduction methods. It uses orthogonal transformation to convert a set of variable data that may be related to a set of linearly uncorrelated variables. The converted variables It is called the principal component. Principal component analysis can play a role in reducing dimensions, which is very suitable for the current data set (with more than thirty variables)

### 3.4.3. Analysis results

Extract the principal components and use different principal components to fit different observations, and finally obtain three observation value sequences, and select the factor with the highest proportion in each feature value sequence as the representative factor of the sequence.



According to our selection rules, the representative factors of the three principal components are "Equipment and Service", "Basic conditions", and "Price".Finally, the three eigenvalue sequences are used as the raw materials for clustering, and cluster analysis is performed.



According to the clustering results on the way, we found that different airbnb are significantly divided into two parts according to price factors , Basic conditions and Equipment and Service in these two parts may have the same level, but the price gap is larger, which also confirms our previous point: some soft power can significantly affect the price of airbnb.

# 4. Insights and Implications

## 4.1. Evaluate Business Problem

### Problem1: What factors affect airbnb prices

* Basic conditions determine prices
* Additional facilities affect prices
* Service level disturbance price

### Problem2: What factors can be improved by the landlord to increase the price of airbnb

**Basic conditions determine price**

Airbnb's operations cannot raise prices based on temporary turnover. It is necessary to set a reasonable price level based on the basic conditions of airbnb (such as room type, room area). This can not only ensure a comprehensive profit margin, but also maintain a good airbnb brand image to a certain extent. Let airbnb prices meet the basic expectations of customers.

**Additional facilities affect prices**

The operating conditions of airbnb with the same basic conditions may not be the same. Due to the impact of more additional facilities, airbnb needs to set different goals according to different additional facilities. Airbnb, which lacks additional facilities, can reduce target profits to lower house prices and increase occupancy. On the contrary, in room types with limited basic conditions but sufficient additional facilities, target profits can be increased, house prices can be increased, and room operating income can be increased. Although the low-income facilities rooms can be used to pay lower fixed-income income, it will have a positive impact on offsetting the fixed costs of operating activities; in high-additional facilities rooms, this offset will increase.

**Service level disturbance price**

Airbnb's pricing is based on airbnb's own brand, business purpose, season and passenger flow. For high-end airbnb with high service level, the price is increased within a certain range, which is in line with the airbnb brand and local airbnb prices, and improves the profit of airbnb under the premise of a good customer experience. When the service level is low, it cannot be discounted blindly. The discount should be compatible with the basic conditions of airbnb and the level of additional facilities to ensure airbnb's own interests.

## 4.2. Recommendations & Suggestion

1. Maintain the advantage area

Maintaining an advantage area is a factor that determines the price level of airbnb. The factors located in this area mainly include the basic indicators such as the number of rooms, the number of bathrooms, the room area, and the number of beds. Explain that these factors are the factors that travelers take very seriously when choosing airbnb, and are also factors that feel more satisfactory in the actual accommodation experience. Therefore, airbnb should continue to maintain the service level of these projects, and pay attention to the investment of these projects in the future.

2. Centralized improvement area

The centralized improvement zone is a factor that can further affect the airbnb price level. Factors located in this area include various additional facilities provided by airbnb, such as dryers, washing machines, vacuum cleaners, TV sets, shampoos, etc. The importance scores of these factors are higher than average, indicating that travelers pay more attention to these factors. Airbnb hosts should face the needs of customers and travelers and take effective measures to improve the deficiencies.

3. The last supplementary area

The last supplementary area is the last factor that can raise the price level of airbnb. Mainly include response time, refund rule and so on. Airbnb hosts can consider improving these factors when basic conditions and additional facilities cannot be improved.

# Appendix A：Descriptive statistics of selected variables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| number\_of\_bedrooms | | | | number\_of\_beds | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 5.00 | 1.17 | 0.00 | 1.00 | 8.00 | 1.74 |
| number\_of\_bathrooms | | | | number\_of\_individual\_bathrooms | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 4.00 | 0.71 | 0.00 | 0.00 | 3.00 | 0.07 |
| number\_of\_shared\_bathrooms | | | | refund\_rule | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 0.00 | 4.00 | 0.35 | 1.00 | 3.00 | 3.00 | 2.60 |
| number\_of\_language\_of\_host | | | | response\_rate | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 1.00 | 1.00 | 31.00 | 1.49 | 0.00 | 1.00 | 1.00 | 0.92 |
| response\_time | | | | type\_airbnb | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 1.00 | 1.00 | 72.00 | 7.30 | 1.00 | 3.00 | 35.00 | 3.11 |
| room\_price | | | | registration\_date | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 3.E+04 | 2.E+05 | 4.E+06 | 3.E+05 | 10.00 | 48.00 | 124.00 | 49.32 |
| number\_of\_reviews | | | | overall\_rating | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 2.00 | 30.00 | 378.00 | 50.54 | 3.00 | 4.50 | 5.00 | 4.67 |
| Wireless\_Internet | | | | washing\_machine | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.99 | 0.00 | 0.00 | 1.00 | 0.40 |
| hot\_water | | | | cable\_TV | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 0.00 | 1.00 | 0.49 | 0.00 | 0.00 | 1.00 | 0.35 |
| TV | | | | elevator | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.71 | 0.00 | 0.00 | 1.00 | 0.31 |
| fire\_sensor | | | | first\_aid\_kit | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.82 | 0.00 | 0.00 | 1.00 | 0.37 |
| fire\_extinguisher | | | | carbon\_monoxide\_sensor | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 0.00 | 1.00 | 0.37 | 0.00 | 1.00 | 1.00 | 0.69 |
| hair\_dryer | | | | shampoo | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.80 | 0.00 | 1.00 | 1.00 | 0.78 |
| kitchen | | | | hanger | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.91 | 0.00 | 1.00 | 1.00 | 0.82 |
| heating | | | | air\_conditioning | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.97 | 0.00 | 1.00 | 1.00 | 0.92 |
| laptop\_work\_space | | | | necessities | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 1.00 | 1.00 | 0.68 | 0.00 | 1.00 | 1.00 | 0.94 |
| dryer | | | | iron | | | |
| Minimum | Median | Maximum | Mean | Minimum | Median | Maximum | Mean |
| 0.00 | 0.00 | 1.00 | 0.40 | 0.00 | 1.00 | 1.00 | 0.74 |
|  |  |  |  |  |  |  |  |

# References

[Principal component analysis-From Wikipedia, the free encyclopedia]

<https://en.wikipedia.org/wiki/Principal_component_analysis>

[K-means Cluster Analysis-UC Business Analytics R Programming Guide]

<https://uc-r.github.io/kmeans_clustering>

[Decision tree-From Wikipedia, the free encyclopedia]

<https://en.wikipedia.org/wiki/Decision_tree>

[Decision Tree in R with Example]

<https://www.guru99.com/r-decision-trees.html>