**Proposal Template**

**Team Name (Choose a name to represent your team)**

Guest

**Team Members**

201400944 산업경영공학과 김홍범...

201402322 산업경영공학과 이새결

201804554 산업경영공학과 박세윤

201803603 독일어통번역학과 윤다영

**Project Objective**

By using regression analysis, we aim to statistically estimate the relationship among number of accommodates or amenities, reviews and price using parameter model

The Association Rule aims to see the relationship among preferred room type on the countries we went on a trip.

Clustering - We use the data to know the relevance of accommodation price, review scores, and type of property by grouping prices. Review scores is written by customers and type of property is apartment, villas, loft and so on.

The detailed assessment items of the Airbnb review are taken as predictors and, based on this, the categories of total ratings, the target variables, are classified into several classes. This also predicts the category of total ratings for new data.

**Project Scope**

Regression- To analyze the relationship between price and any other variables about Airbnb., the price in the data set is the target variable. We use numerical data and categorical data, to do regression analysis. Such as we analyze accomdates, Bathrooms, Bedrooms, or number of reviews using parameter model(linear regression, multiple linear regression). And then we do the simple text mining and count the number of Amenities. By using this data we can also do regression analysis. The presence or absence of a superhost is a binary variable so we need dummy transformation. Based on this, I will use logistic regression analysis.

Association - The rules of association are intended to identify the rules in which the event occurred simultaneously and to predict according to the conditions of the event. The goal is to find out which type of room you prefer depending on the country you traveled to. So I will use the country and room type data. Using this, we identify the minimum support and minimum confidence, and we can predict the preferred type of room depending on the destination we choose.

Clustering - Prices will be segmented into specific ranges by using K-means clustering method and hierarch algorithms ,because price is numerical data. There isn’t specific target variables. Through this, we can know the relevance between review score and property type by group price.

Classification - Because the target variables are discrete, the classification tree uses the appropriate algorithm classification and variable selection criteria to refine the categories of total ratings, which are the target variables, using six predictors (accuracy, location, cleanliness, communication, etc.) and to predict the total rating of the new data. In addition, we will select the variables that have the greatest influence on the target variable by dimension reduction and selection of variables. If necessary, merge the categories of target variables into a few. If possible, combine several predictors to identify the interaction effects that act on the target variable.

**Dataset Explanation**

This dataset is about Airbnb rating score and price in recent years.

**Listing ID** – numerical data

**Name** – categorical data

**Host ID** – numerical data

**Host Name** – categorical data

**Host Response Rate** – numerical data(percentage). Host response rate according to message.

**Host Is Superhost** -categorical data. This variable tells us whether it is a superhost or not.

**Host total listings count** – numerical data. How many times host upload their house on Airbnb.

**Street** -catrgorical data – it contains so many various language.

**City** – catrgorical data

**Neighbourhood cleansed –** categorical data

**State** -categorical data

**Country** – categorical data

**Latitude** – numreical data, Show the

**Longitude** - numreical data

**Property type** – categorical data, the type of room that the host provides to the customer

**Room type** - categorical data

**Accommodates**

**Bathrooms** – numerical data. How many baths are exist in house

**Bedrooms -**numerical data.

**Amenities** -categoorical data, provided option by host, categorical data

**Price –** numerical data

**Minimum** nights – numerical, aviable minimum Accommodation date

**Maximum** nights - numerical, aviable maximum Accommodation date

**Availability 365 –** numerical data, possible accommodation date

**Calendar last scraped –** numerical data, The data for the last scraped date.

**Number of reviews –** numerical data

**Last Review Date –** numerical data

**- Review Scores Rating :** numerical data type. the final score scored by evaluating the sum of the detailed assessments by traveler

**- Review Scores Accuracy :** numerical data type. Evaluation of the accuracy of the description written in the room

**- Review Scores Cleanliness :** numerical data type. Evaluation of room cleanliness

**Review Scores Check in :** numerical data type. Evaluation of convenience or satisfaction of check-in

**- Review Scores Communication :** numerical data type. evaluation of communication with hosts

**- Review Scores Value :** numerical data type. Price/Satisfaction Assessment for RoomReviews per month

Dataset : **About Airbnb's Accommodation rating data**

This is Airbnb's latest accommodation data for cities around the world.

It shows the host response, location, property, room type, accommodates, amenities, price, rating, number of reviews.

**Role of team members:**

|  |  |
| --- | --- |
| Students | Tasks |
| 201400944 김홍범 | Do the regression Analysis |
| 201402322 이새결 | Do the Classification Analysis |
| 201804554 박세윤 | Do the Clustering Analysis |
| 201802366 윤다영 | Do the Association Analysis |

**Conclusion**

**Clustering**

we checked the number of clusters using the elbow method. We can see that the graph is being bent in the part where k is three, therefore the number of clusters is three.

Create algorithm for clustering. The variables used are 'price', 'number of reviews' and 'review scores rating'. K-means++ is used.

Visualized using Mglean. First is the relationship between price and number of reviews. As a result, clusters were created based on prices of 0~180, 180~420, and 420~1125. In Cluster0, we can see that the number of reviews is relatively high. The lower the price, the higher the number of reviews.

Next, find the relationship between price and review score ratings. Prices are clustered at the same interval, and it can be seen that review score ratings are more distributed relatively lower in Cluster0. The lower the price, the lower the review score ratings.

In conclusion, the lower the price, the higher the review, and the lower the review score.

**Association**

We can get two meaningful information.

About 87 percent of those who traveled to France used "Entire Room/apt" on Airbnb, and 37 percent of those who traveled to the United States used "Private Room" on Airbnb.

The reason why results in numbers 2 and 5 cannot be considered significant information is that the value of Lift is less than 1. If the value of Lift is less than 1, it is not meaningful because it can include the possibility of chance.

In the rest of the world except for the United States and France, rule is not found.

Because their preference for a particular room type may be influenced by variables other than the "country" variable.

**Regression**

We are first using scatter plot to choose proper features. Based on the results, regression analysis was performed between price and other fatures.

First, we analysis prices between bedrooms, bathrooms, and reviewscoresrating using simple linear regression on. As a result, the three possible variables are all values have p-value lowe than 0.05 so statistically significant. However, lowe r- square value is a problem.  
To solve these problems, multiple linear regression to go ahead. Seven selected on the Review of the variables are selected.  
The r-square is 0.021 that is not still valid but the score is higher than before analysis.  
For better results, categorical variables are also changed to numerical, and we analyzed them after dummy transformations.

As we proceeded with regression, we realized the importance of the result values as a result of variable selection. If we randomly select and analyze variables without proper methods, the results were not good.  
When proceeding with the next regression analysis, we will need to proceed with a regression analysis (ridge, lasso) that matches the characteristics of the variable, or we will need to be careful in selecting the variable.

**PCA**

Through confusion matrix and Naïve Bayes Classifier, we correctly predicted 1,444 cases. And it correctly predicted respectively 2415, 2565, 2565, 2585, 2596, 2634 cases when the existing eight variables were reduced to 1 to 7 variables. However, the total number of samples in CSV file (‘airbnb10’) is 10000. Both the existing and the new potential variables have significantly lower predictability. In fact, this is because the deviation of each detailed score in the review is not significant. Of course, we may know intuitively before making these measurements, but we've gone through PCA for accurate mathematical measurements, and we can once again confirm that this dataset was not very suitable for doing PCA. As a result, it was regrettable, but I think it was meaningful in the process of carrying it out.

**References**