**Final Project**

Dhruvang Pravinkumar Patel (002195090)

Parva Pareshbhai Patel (002195186)

CPS Analytics, Northeastern University

ALY 6040 Data Mining and Applications

Prof. Hootan Kamran Habibkhani

October 30, 2022

**Introduction**

This project is based on the performing data mining technique and find the insights from the Airbnb dataset. The dataset has the information related to the Airbnb of the different cities of USA. The dataset has more than 25 columns and more than 70000 rows in it. It was taken from the Kaggle website. Various model are created from the data to answer the various business question that would be helpful for the Airbnb business.

Business Questions:

How we can predict the Airbnb rental price?

Which features are important for price prediction?

What changes one should made in Airbnb so it can become more popular?

Does the host and its ratings matter while booking the Airbnb?

**Data Cleaning**

The initial stage of the data exploration is to understand the variable of the dataset.

Variables Description:

* Target variable: log price (Numeric)
* id (numeric) | Unique identifier for each listing
* property\_type (categorical) | Type of the property (e.g. Apartment, house, condo)
* room\_type (categorical) | Type of rooms (e.g. Entire home/apt, private room)
* amenities (text) | Available amenities. Unstructured list separated by commas (e.g. tv, kitchen).
* accommodates (numeric) | Number of people the rental fits
* bathrooms (numeric) | Number of full and/or half baths
* bed\_type (categorical) | Types of bed (e.g. futon, real bed)
* cancellation\_policy (categorical) | (e.g. Flexible, moderate, strict)
* cleaning\_fee (boolean) | TRUE/FALSE
* city (categorical) | (e.g. Boston, NYC, LA)
* description (text) | Basic description of the room. Unstructured and up to the host how to populate.
* first\_review (Character) | How long ago the first review was left
* host\_has\_profile\_pic (boolean) | TRUE/FALSE (no link to picture)
* host\_identity\_verified (boolean) | TRUE/FALSE (via email verification)
* host\_response\_rate (numeric) | How often the host replies to inquiries (%)
* host\_since (Character) | Date that they opened their account
* instant\_bookable (boolean) | TRUE/FALSE
* last\_review (Character) | Date of most recent hosting
* latitude (numeric)
* longitude (numeric)
* name (text) | Name of rental property.
* neighbourhood (categorical) | Informal description of neighbourhood (e.g. Brooklyn Heights, Downtown)
* number\_of\_reviews (numeric) | Total number of reviews given by guests
* review\_scores\_rating (numeric) | Mean rating of reviews given by guests
* thumbnail\_url (numeric) | Link to primary photo of rental property.
* zipcode (numeric; likewise) | Zip code. Candidate for bringing in additional data.
* bedrooms (numeric) | Number of bedrooms in rental
* beds (numeric) | Number of beds in rental

First the dataset was imported into the R using the read\_csv() function. After importing the dataset, str() function was used to get the structure of the dataset which also gives information about the data as in how many rows and columns are there what are the datatypes and first five observations of each variable to get a basic idea of how a dataset looks like.



Figure 1 : Structure of the dataset.

After getting familiar with the dataset were there any null values or not was checked.



Figure 2 : Null values in the dataset

Figure 2 shows the number of null values available in the dataset. As there are more than 70000 entries in the dataset we can remove all the null values which would be the best option. It was important to deal with the null values because they can create inaccurate results in the further analysis.

With the help of the duplicate() function it can be checked whether there were any duplicate values in the dataset. After running the duplicate function the result shows no duplicate values which can be seen in the figure 3.



Figure 3 : Duplicate values

From the figure 1 it can be seen that the datatype of the log price is numeric but it is the log value of price as mention in the documentation of the dataset which is available on the Kaggle website from where the dataset was taken. So, it is necessary for the analysis to convert the price in the normal dollar price. It can be converted from log to dollar using the exponential.



Figure 4 : Log to Exponential

The target variable of the project is the price of the Airbnb rooms so it was necessary to remove the outliers from the price columns. The summary() function was used to see the summary of the price.

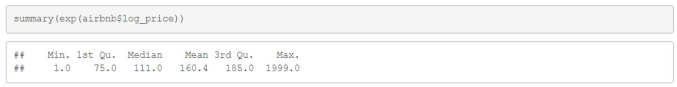


Figure 5 : Summary of price

From figure 5, it is seen that there was huge difference between the mean and the maximum value of the price. This means there are lot of outliers in the dataset for price which were either exception of the house price or mistaken while filling the data. To remove the outlier first box plot was created to know the outlier present in the priceperday column of the dataset.

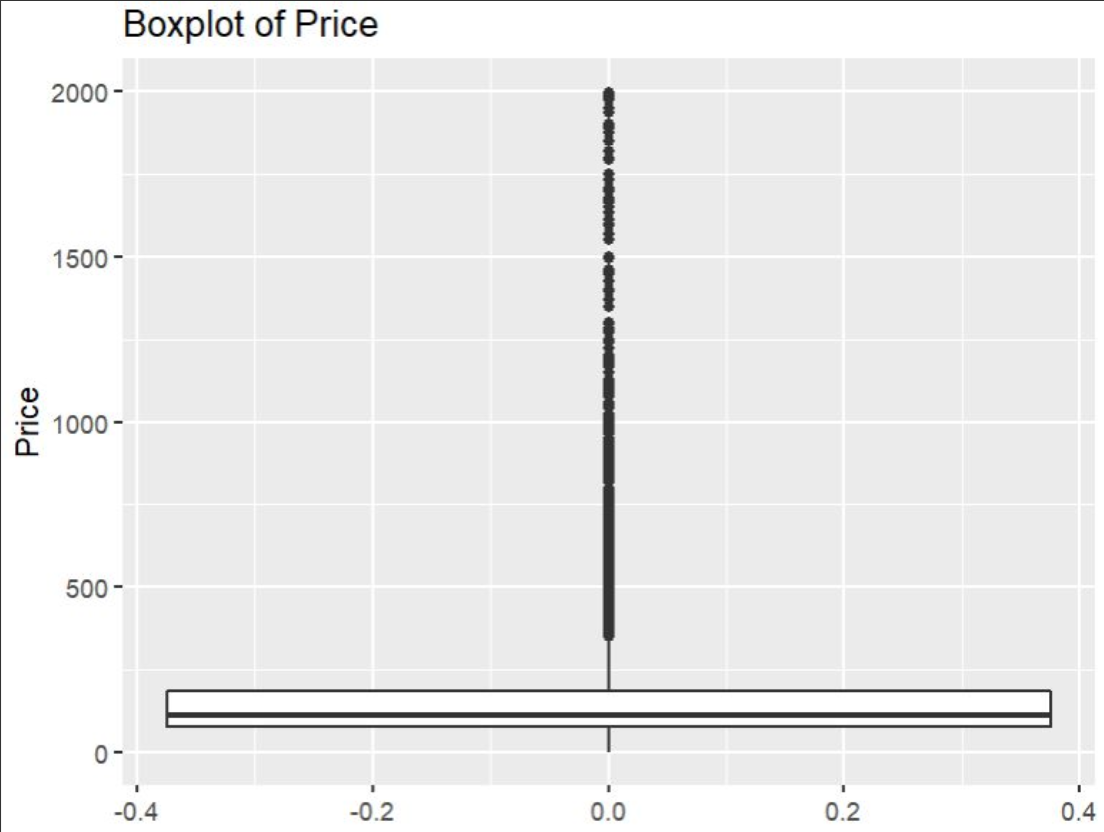


Figure 6 : Box plot of price

Figure 6, shows the boxplot of the price. As it can be seen that there are lots of outliers, so to remove these outliers, quantiles method was used. Here quantile() function is used and then subset is made for upper quartiles and lower quartiles for determining the outliers.



Figure 7 : Removing Outliers

From the figure 6 which is the boxplot of price it can be seen that the outliers values are of greater than 350 and in the quartile method also the upper quartile also show 326.5. So, all the values of price greater than 350 are removed.

The new summary of the price after removing the outliers is shown in figure 7. As seen in figure 4 the maximum price was $1999 which has been reduced to $350.



Figure 8 : Summary of updated price without Outliers

The boxplot after removing the outliers is shown in the below figure.

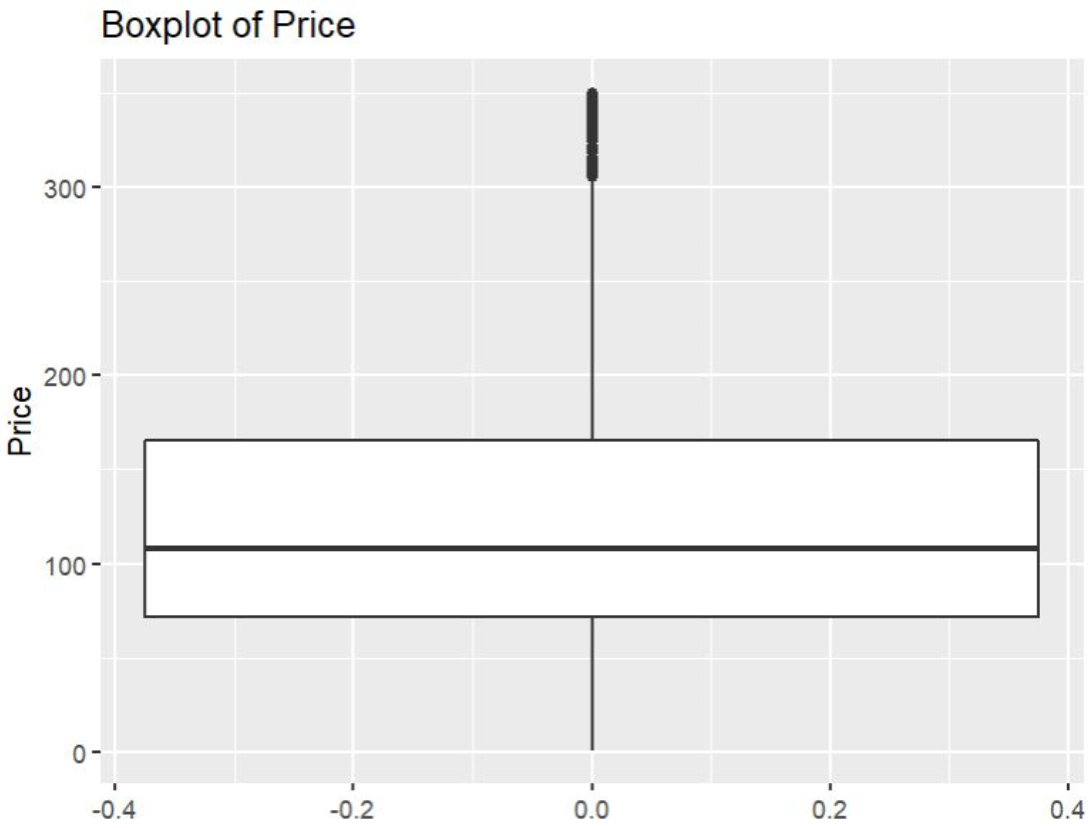


Figure 9 : Boxplot of price after removing the outliers

Removing the outlier is the important part of the data cleaning process but the Airbnb dataset which was used in this project have only three numeric variable which can be useful in the analysis and creating the different model in further analysis.

In the property\_type column there are many different type of property which have minimal number of entries. So, the property type which have the less number of entries are combined into one property type other. If all the property type whose entries was less were not combined then they would create the confusion and complexity in the further analysis.



Figure 10 : Combining the different property type to one

The below bar plot is of the updated property type after collapsing the types into other and reducing it into only five major types.

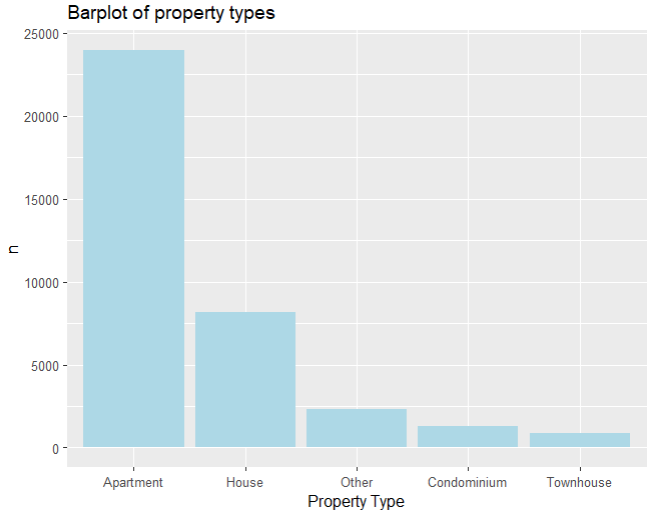


Figure 11 : Bar plot of the updated property type

Now the dataset is clean and in the proper format where there are no null, duplicate and suspicious data so data mining technique can be applied for the analysis and to built the various models that can be used to answer the different business question.

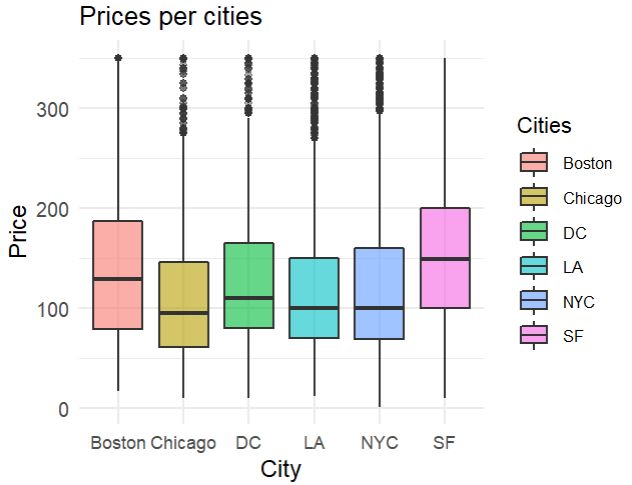


Figure 12: Boxplot of prices per cities

Figure 12 depicts the boxplot of prices per cities mentioned in the dataset. As it can be seen that San Francisco(SF) has the highest mean of prices followed by Boston, DC, and New York.



Figure 13: Boxplot of prices per room type

Figure 13 displays the prices of Airbnb for room types. From the figure it can be said that if the Airbnb is listed for entire house, then the price of that house would be much higher than private or shared room.

**K-Means Clustering (**Does the host and its ratings matter while booking an Airbnb?)

The cluster analysis can only be applied on the numeric data only and the library need for clustering analysis are stats, dplyr, ggplot2 and ggfortify. For making cluster the columns selected were priceperday, number of reviews and review score rating. Before applying the K-mean clustering the optimal number of cluster should be founded. To do so the wssplot is used. Elbow method is used to select the optimal number of cluster from the plot made by wssplot.

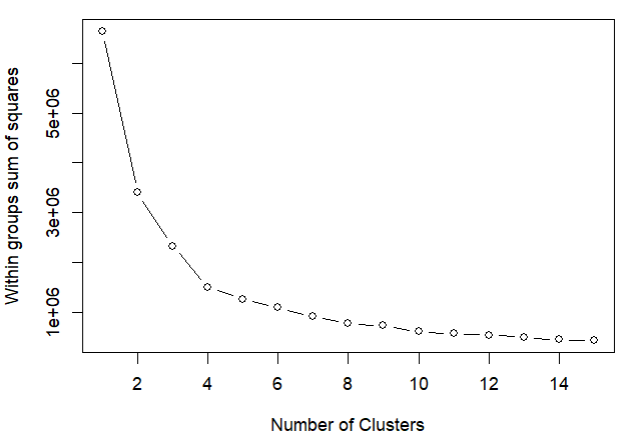


Figure 14 : wssplot to know optimal number of cluster

The preceding figure 14 shows a dramatic decline between the points 1 and 3, which suggests that 3 is the ideal number of clusters. Now, the k-mean() function may be used to form a cluster. Cluster plots can be made using the autoplot() method.

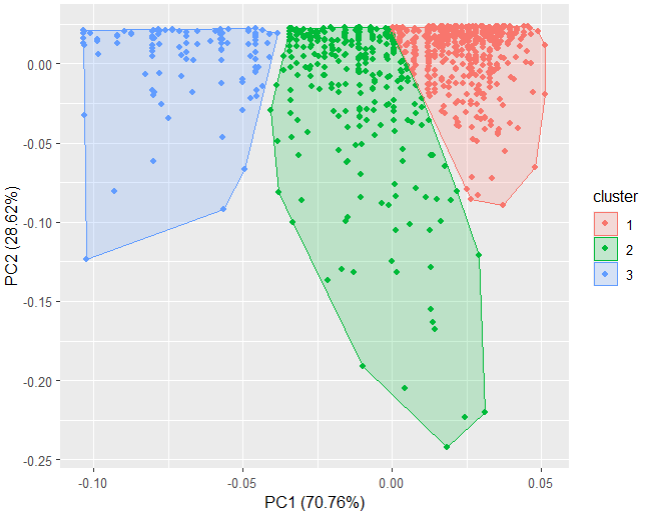


Figure 15 : Plot of cluster

It is clear from the figure above that each cluster is distinct from the others and is therefore autonomous. It is also possible to learn more about the clusters, including their centres and sizes.

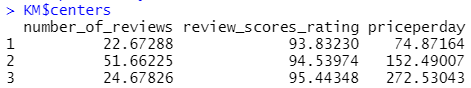


Figure 16 : Cluster Definition

**ANALYSIS:**

By constructing the cluster, it is possible to ascertain how much the popularity of an Airbnb and how much its rating influence its pricing. The cluster's centre can be seen in figure 16. The cluster's centre, which symbolises the cluster's data, is the location from which all of the cluster's points are not far away. By examining the cluster's centres, it can be concluded that the price of the Airbnb rises as the rating does because the price in cluster 1 is 74.87 and the rating is 93.83, while the price in cluster 2 is 152.49 and the rating is 94.53.In cluster 3 the price is 272.53 and its rating is highest which is 95.44. So, it can be said that the rating matter for predicting the price of the Airbnb.

**Decision Tree:**

**(**How can we predict Airbnb Rental Price? And Which features are important for predicting the rental price?)

Decision trees are easy to understand. The majority of individuals won't have any trouble comprehending them because of their structure, which follows the development of the human mind naturally. The model is very straightforward to visualise and makes it possible for you to fully comprehend the decisions being made.

The optimal split would now divide the dataset into two sets, training data and testing data. The model would then be run using training data, and its accuracy would be evaluated using a test dataset.

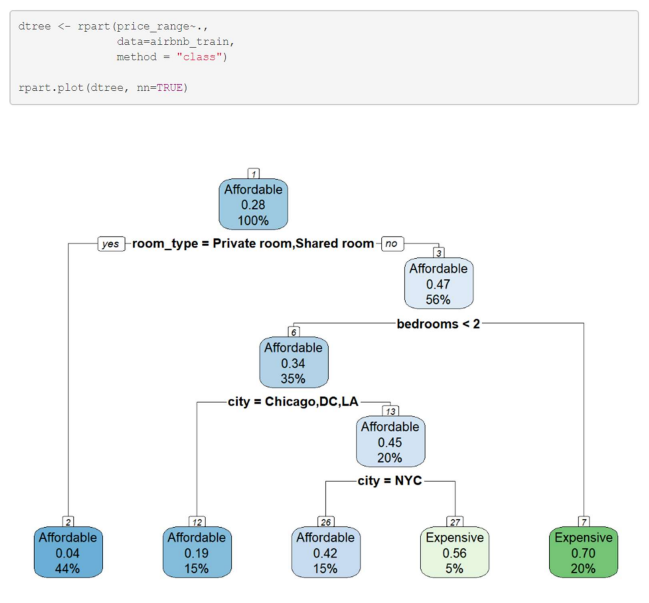


Figure 17 : Decision Tree

As can be seen in Figure 17, the first affordable price range is selected as a root node. After that, it is divided based on the type of room. Private or shared rooms are reasonably priced, and if the full home or apartment is available, there is a further division of bedrooms. An apartment or home is pricey if it has two or more bedrooms; otherwise, it is divided into cities. Homes with affordable prices fall between $10 and $150 while those with costly prices exceed $151.

The prune() method is now used to prune the model, which reduces overfitting.

Confusion matrix is now utilised to verify the accuracy of the decision tree model. For this, a data frame of predicted values is first created using the predict() function, and then the confusionMatrix() function is used to depict the matrix and assess accuracy.

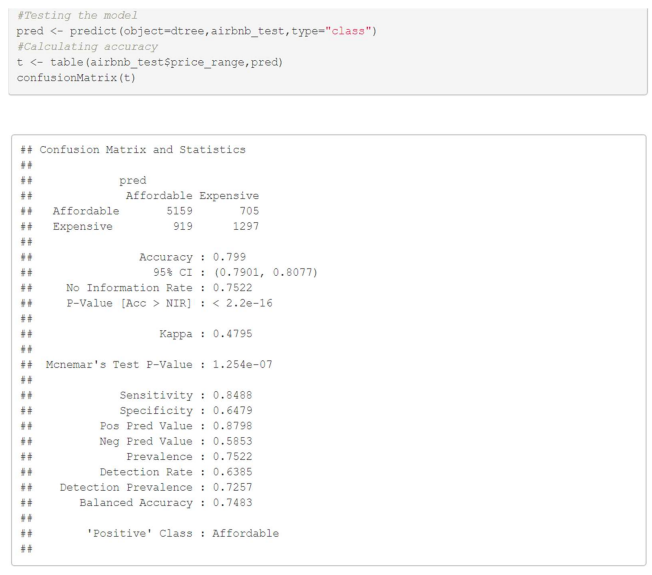


Figure 18 : Confusion Matrix

Figure 18 demonstrates that the model's accuracy is 0.799, indicating that the decision tree is 80% correct. 1297 true negative values and 5159 true positive values are present.

**ANALYSIS:**

Both Airbnb guests and hosts may benefit from this dataset by using it to forecast the value of their properties. Customers may use the model to discover what is within their price range, while hosts could use it to determine which regions are pricey or reasonable. Customers could afford it whether they preferred a communal or a private room. The host should choose a high price range for pricing the property if it has two or more bedrooms and is offered as a whole on the website.

**Text Mining: (**What changes one should made in Airbnb so it can become more popular?)

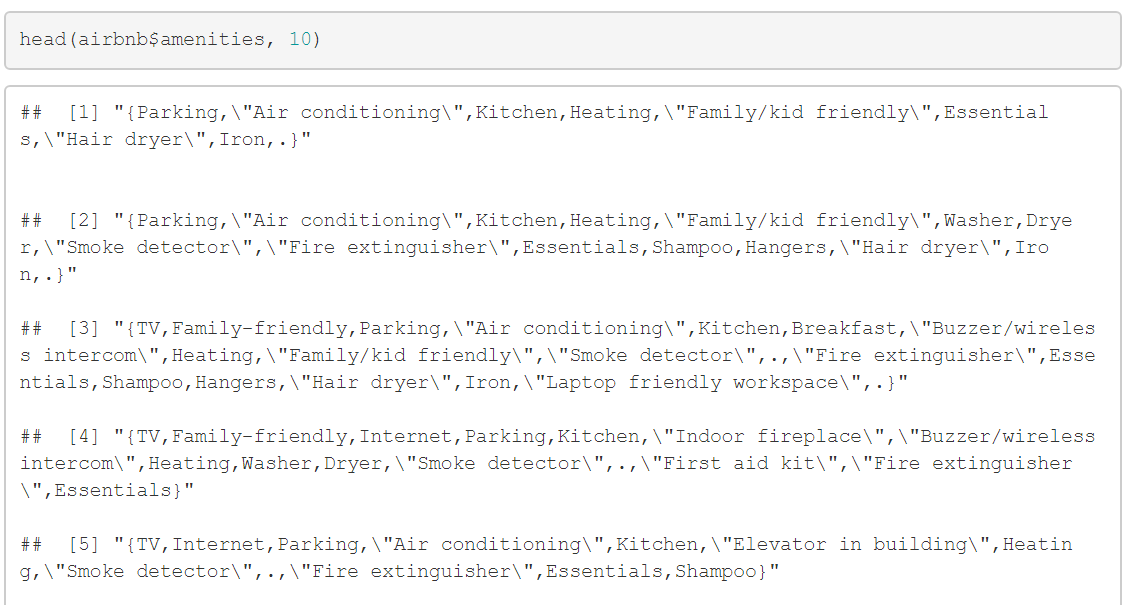


Figure 19 : Amenities

Figure 19 demonstrates how poorly organised and imprecise the description of amenities is, with erroneous symbols and columns. Therefore, text mining would be used on this variable to determine which amenity the host or the consumers preferred.

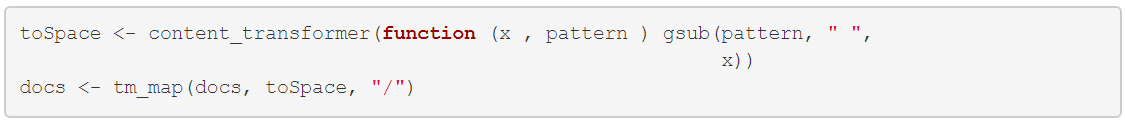






Figure 20: Replacing symbols.

The text has to be cleaned up for appropriate mining of the text once the data has been entered into the corpus. Figure 20 illustrates how the symbols have been replaced by spaces. We can change the content of a R object by using the function contenttransformer(). When using the Tm\_map() method, transformations are applied to the provided objects, such as the strange symbols into empty space.



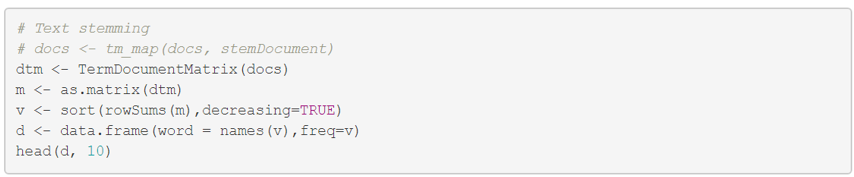






Figure 21: Cleaning text

Figure 21 demonstrates how the data was cleaned to improve the accuracy of the analysis by removing all number formats and transforming all text to lower case. The data may have been erratic once extra spaces and punctuation were removed.



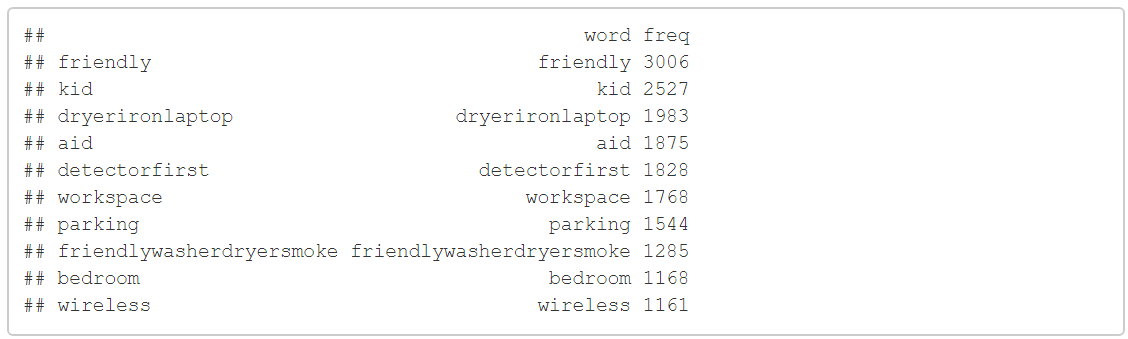


Figure 22 : Text stemming

A matrix called the Document-matrix lists the word frequencies. While column names are words, row names are actual documents. After making the documentation matrix, top 10 were ran and from the figure it can be said that the top amenities which were in the Airbnb houses were friendly, kid, dryer – iron – laptop and aid.

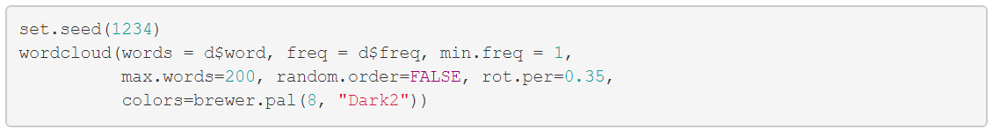




Figure 23 : Word Cloud

Following the generation of the documentation matrix, a word cloud plot is made. Using the wordcloud() method, the displayed figure—a visual representation of text data—was produced. The font size or colour of each tag denotes its importance. Tags are frequently single phrases.

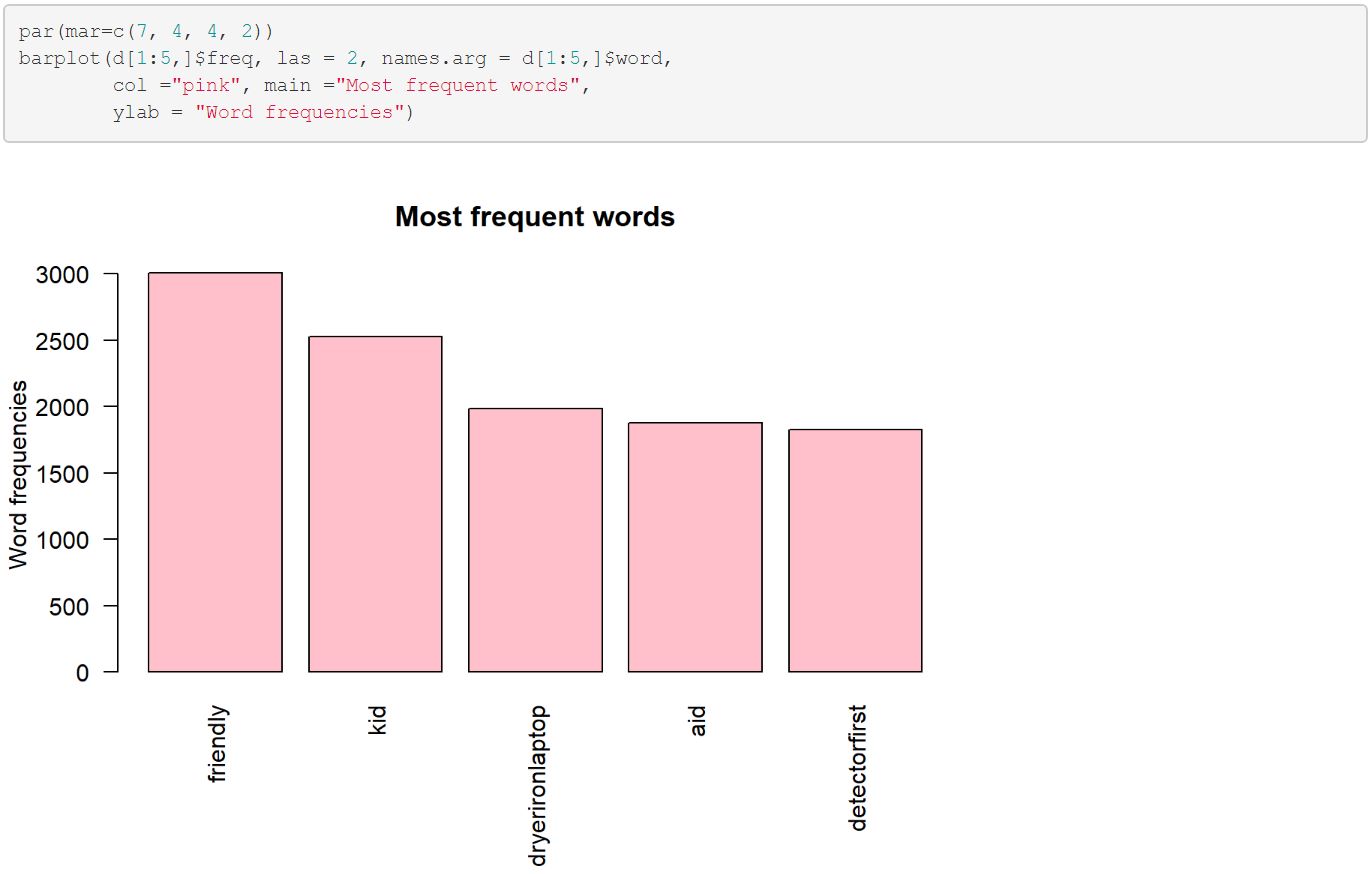


Figure 24: Barplot of frequently used amenities

Bar plot is used to understand properly the number of occurrence of highest number of amenities mentioned in the dataset by barplot() function.

**ANALYSIS:**

From figure 23, it can be understood that friendly is the most frequently used amenities in this dataset. This means that all the host as well as customers of Airbnb wants to have a kid - friendly and a better workspace.

The words' frequency of recurrence is shown in descending order in Figure 24's bar plot. As a result, friendly, kid, dryer-iron-laptop and aid are the most used amenities mentioned in the dataset.

**Interpretation:**

It has been found that Airbnb's excellent review rating justifies its premium pricing. The owner must thus raise their review score if they wish to obtain a high rent from Airbnb. They may do this by providing better service and attending to their clients' demands.

In addition, it is clear that the Airbnb with the moderate price has the most guests when compared to the least and costly Airbnb. The owners must thus make sure that their price is fair for a variety of clients in addition to the review rating if they want to increase revenue because more customer would be attracted by good rating and affordable price.

Figure 17 shows that the factors utilised to forecast the goal variable, price range, from the Airbnb dataset include room type, bedroom, and cities. A host can predict its price by just knowing the room type and bedrooms of the house and the accuracy would be 80% for that price prediction.

An important factor in selling or renting a home is its amenities. Customers of Airbnb typically travel there or are there on business. The facilities of the home are the main thing visitors notice. In order to determine which facilities are most frequently used across all of the listed homes and what kinds of amenities guests like throughout their stay, text mining might be beneficial for Airbnb owners. The text mining revealed that a family-friendly, kid-friendly, and office-friendly apartment is what a consumer mostly seeks. From there, the host may determine which facilities the guests appreciate and what they might do to improve accessibility to their home.

**References :**

K-Means Cluster Analysis in R. (2018, December 14). YouTube. Retrieved October 16, 2022, from <https://www.youtube.com/watch?v=KmYUE7Of5rU&list=PL_jahYUFcXYxgQN2S9o7gURe_28tIaBz&index=10&ab_channel=EasyML>

Elbow Method — Yellowbrick v1.5 documentation. (n.d.). Retrieved October 16, 2022, from <https://www.scikityb.org/en/latest/api/cluster/elbow.html#:~:text=The%20elbow%20method%20runs%20k,point%20to%20its%20assigned%20center>.

Anonymous. (2019). Decision tree pruning - Wikipedia. Decision Tree Pruning - Wikipedia. Retrieved October 6, 2022, from <https://en.wikipedia.org/wiki/Decision_tree_pruning>

Team. (2021, November 17). Introduction to Decision Trees: Why Should You Use Them? | 365 Data Science. 365 Data Science. Retrieved October 16, 2022, from <https://365datascience.com/tutorials/machine-learning-tutorials/decision-trees/>