**Introduction**

**Analysis**

**Data information**

Dataset named “HotelReviews” and consists of 35912 observations and 19 variables. Variables are of different types of Factor, Integer, Logic and String. There are 1000 hotels in the dataset, and it includes the data about hotel location, name, rating, review data, title, username and more.

**Pre-Processing**

**Missing value detection and handling**

As in every model development, it is important to manage missed values before model development. According to the data provider, there is no missed values in the dataset but to make sure, missing value detection code included in the model.

**Removing non-English text**

The review texts are not entirely in English; hereby, in pre-processing these reviews were removed from the analysis. In consideration of the dataset’s size and robustness, removal of non-English reviews has minimal impact on the overall analysis. The review text in the “reviews.text” variable has been tested for any non-English words using “textcat” function.

**Exploratory text analysis**

Exploratory analysis involved the main variable that was the reviews text. Word count, special character count, punctuation count has been analyzed. Comparison of reviews text from users of different facilities, from different cities and especially between users with overall positive and negative reviews based on the “reviews.doRecommend” variable.

**Charts of most frequent words**

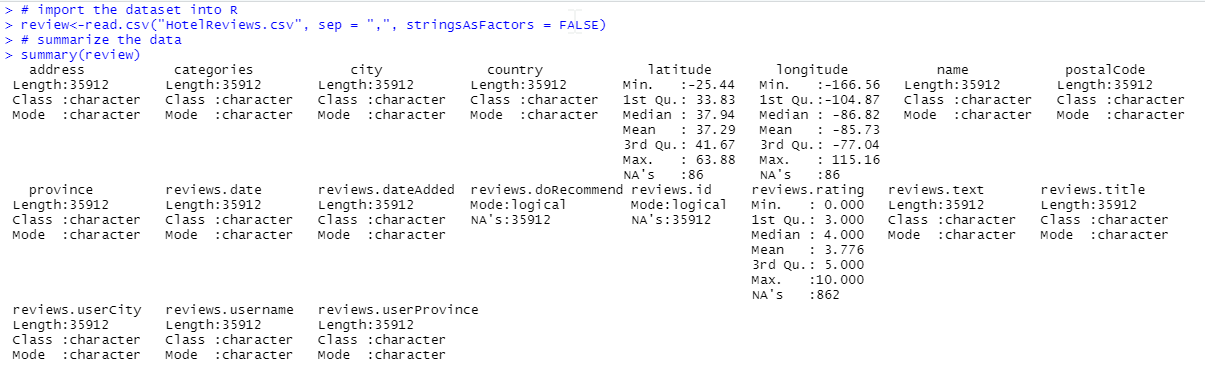
Bar charts and word clouds have been used as the most common ways to visualize a text corpus and compare the distribution of frequently used words. The word cloud varies the size of the words based on the frequency.

**Building classification model**

Classification models have been built based on the review text analysis as the independent variable to predict if the customer recommends the hotel or not.

**Results**

**Dataset Summary**

****

**Figure-1:** Summary of the dataset.

**Classification Model**

**Conclusion**