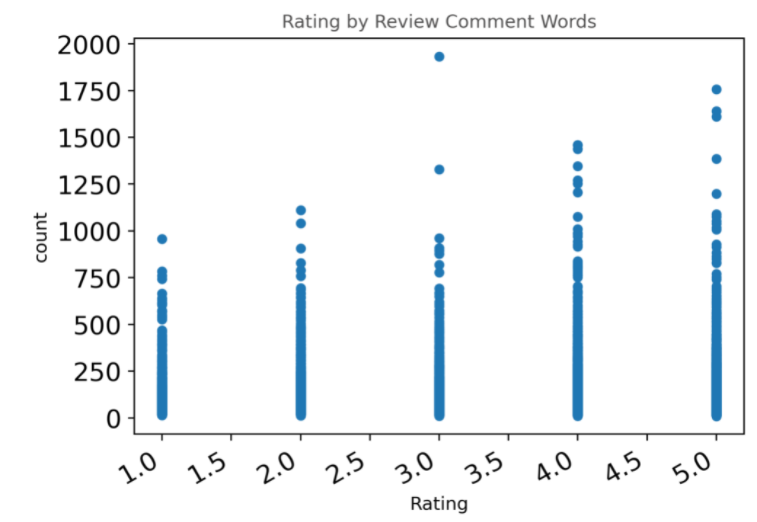
Selection of Model based on Hotel Reviews

Hotel Review Analysis is done by analyzing the review comments, as well as the ratings, which are given on a scale of 1-5. The data is from Kaggle, an extract from Tripadvisor. There are 20491 records. The review comment sizes vary, the least being 9 words, and the highest is 1933 words per comment.

The dataset consists only of two columns – Review and Rating. But I have derived more columns from these like count, which is the number of words to form the review; including punctuations, senti\_score - which is calculated based on the review comment, polarity which is the first field from the senti\_score, and Review\_Status - which is Positive, Negative and Neutral based on the polarity value.

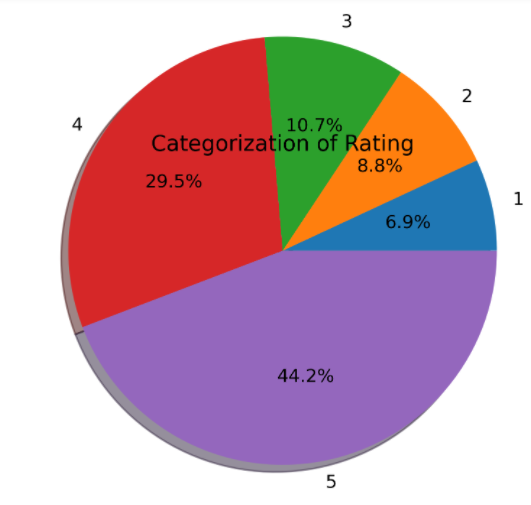
In the future, multiple models will be built and then it will be trained to give results based on the review comment as input. The analysis will be done on the ratings and reviews, and a conclusion will be drawn. Below are the four graphs plotted along with the explanation required per Milestone 1.

Graph1:



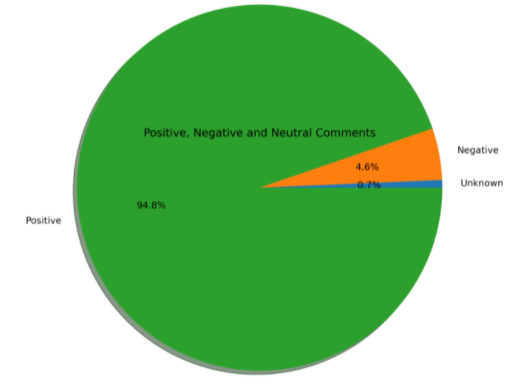
The Scatter graph is used to plot the number of words in the review comments against the Rating. As per the observation, the number of words has increased gradually when the ratings have increased. When the customer is happy and satisfied, they provide detailed feedback about their experience.

Graph2:



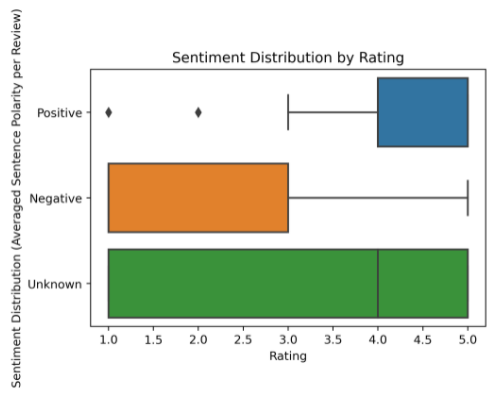
The pie chart is used to represent the Ratings. Categorization of Rating is the title of the pie chart. As per the picture Rating, 5 is the highest with 44.2%, and the least is of Rating 1 of 6.9%. As per the ratings, most of the comments should be positive, and few comments should be negative.

Graph 3:



I have used a pie chart again because it is best to describe when the data falls into different categories. As predicted in our previous plot, we have 94.8% of the comments as positive and only 4.6 % of the comments as negative. When we compare this char and the previous one, a rating of 2 and higher has likely provided positive feedback.

Graph4:



Boxplot is a great chart to represent. I tried a bar chart earlier to represent the results, it worked fine for a smaller set of data but did not make much sense when we had a bigger set of data. Also, the observations in the previous graph may not be accurate, and I may want to correct them after seeing the boxplot, the rating of 3 and higher have a positive comment.

As I kept changing my observations, I may be doing more as I build the model and do more analysis. The more analysis we do, the more observations may be accurate.

As I progressed, I tried to clean the data. I reduced the records that did not make much sense. I had to drop rows that had a rating of 1, but the review was positive, as well as the rating was 5, and the review comment was negative.

I created the cloud of words for the most common words used for best and worst hotel ratings, as shown below:

Text

Description automatically generated with low confidence

Text

Description automatically generated

Also, I tried to calculate the most common words used along with the frequency for both best and the worst rating.

A screenshot of a computer

Description automatically generated with low confidence

Graphical user interface, application

Description automatically generated

As I progressed and discussed with my professor, I decided to build and train a model for the same kind of data. I wanted to try different models and gain experience and check the accuracy of all of them. Below are the models I built and trained:

- Random Forest Classifier

- Logistic Regression

- Decision Tree Classifier

- Linear Regression

- Neural Network

The accuracy was highest with the **Logistic Regression**.

Probably spending more time in the initial phase to look at different kinds of data is a good idea. Prepare a draft plan with a goal and discuss it with the professor or peers. There may be scenarios that you would have never thought of, and the other person may be able to advise.

The word cloud and frequency identify if a restaurant is the best or the worst based on rating and reviews. You can use the existing logic and predict the kind of restaurant.

I like to build more models and train them to see the results. More analysis is possible. If given an opportunity, I would want to start from scratch with a new dataset that includes more fields to do the research.

Overall, it was a great experience and learning. It was an excellent course and a project to accomplish.