Analyzing the TripAdvisor Review Data Set

Data Set

The data set that we will be using in this report is the TripAdvisor Review Data Set pulled from Kaggle. It is a collection of 20,491 hotel reviews extracted from TripAdvisor, a travel-based comparison shopping website. The data set itself consists of the full text of 20,491 hotel reviews as well as the rating associated with the specific review. The review text seems to have been processed so that all upper case letters have been converted to lower case. No indication of how these 20,491 hotel reviews were selected was given on the Kaggle website. The assumption of this report's author is that it was randomly crawled.

There are only two attributes in this data set: the Review text itself and the Rating. The Rating is given as an integer value given from 1 to 5, with 1 being a poor rating and 5 being an excellent rating. In examining a statistical summary of all the Rating values, the mean is 3.95, the median is 4, and the 25% quartiles and 75% quartiles are at 3 and 5, respectively. This does seem to indicate that ratings tend to be skewed in favor of higher ratings.

The Reviews themselves are not easy to analyze in themselves. We used a wordcloud visualization to get a sense of the most commonly used words, which, perhaps unsurprisingly are: 'hotel', 'room', 'resort', and 'time'. To do statistical analysis for this report, we analyzed the length of each review and input it as a new column of values. In looking at the review length statistics, we see that the mean length of a review is 725 words, with the average review being 537 words. 25% and 75% percentile are at 339 words and 859 words respectively. This wasn't too relevant for the ratings because the ratings are limited to a range between 1 and 5, but for the review length, the minimum review length was 44, and the maximum was 13501. With the IQR being 780, it's clear that there will be some outliers on the higher end of review lengths that will need to be dealt with in some way.

The Plan

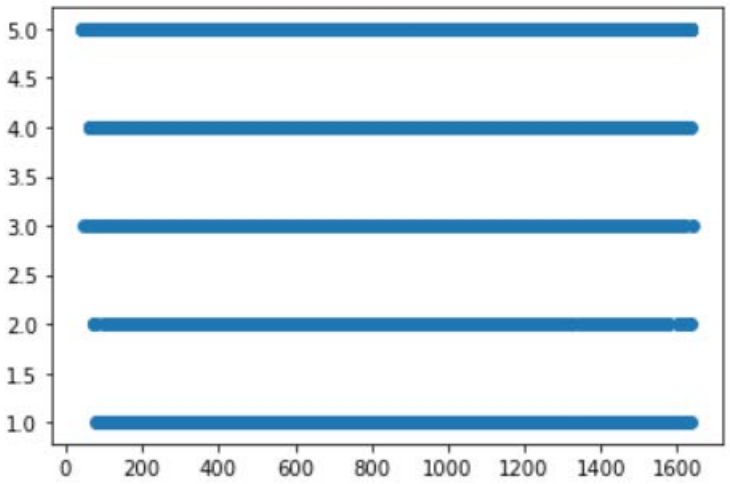
We wonder whether review length has any influence on the rating the review gives to the particular hotel. There are perhaps two ways of thinking about the relationship between these two attributes: it is possible that higher review lengths should be related to higher ratings. Alternatively, perhaps higher review lengths may be related to more extreme scores.

Data Cleaning and Feature Engineering

After converting the CSV into a pandas DataFrame, a drop test was performed to check for any null or missing values and none was found. As previously mentioned, the data set itself only contains review text and ratings, so we started by creating a new column that measures the review length, minus the whitespace present in each review. Also as previously mentioned, when we analyzed the statistical summary for the review length, we noted that there were likely outliers, as median was 725 with 25% and 75% quartiles being at 339 and 859 words, respectively. The IQR range is 1.5* (859-339) = 780, which means the minimum review length of 44 is not an outlier, but values greater than 1639 would be. We therefore drop all entries with a review length 1640 or higher. This removes 1389 entries from the data, leaving us with 19042 reviews to analyze. For further analysis, we also created 5 dataframes, each containing review lengths for each rating (i.e. a dataframe with all reviews with a rating of 1 and so on).

Data Analysis

A look at the scatter plot of review length vs rating is not particularly informative, as shown below:



It doesn't look like we would get much information out of this scatter plot, as the reviews at each rating "level" seem more or less the same. As suspected, linear regression analysis seems to indicate that there is mean squared error of 1.49%, which is not very informative. However, when we look at the mean and median for review length for each rating, there appears to be a slight trend. Mean/Median average review lengths seem to actually decrease as you go from rating 1 to 5. This appears to be the opposite of our original hypothesis.

Possible Hypotheses

Previously we discussed two hypotheses to test, but the analysis has introduced other possibilities:

1. Higher review length is associated with higher ratings
2. Lower review length is associated with higher ratings
3. Review length has no effect on ratings
4. Extreme ratings (i.e. 1 and 5) are associated with higher review lengths.

Hypothesis Analysis and Testing

Based on what we considered earlier from the data, we can probably eliminate both possibilities 1 and 4. If we look at the overall data set, we don't see much of a relationship on the scatter plot. However, if we just look at mean and median of each rating, there is a possibility for relationship, namely that higher ratings are actually associated with lower review scores. We will test to see if this is statistically significant. We will set our null hypothesis as review length has no effect on ratings, and the alternative will be that review length does have an effect, and we will set a significant level of $p = 0.05$.

The relevant statistics for our analysis are as follows (all values are rounded to nearest whole number):

Overall data set: Average Review Length = 584, Standard Deviation = 337

1 Ratings: Average Review Length = 611, Standard Deviation = 356

2 Ratings: Average Review Length = 665, Standard Deviation = 342

3 Ratings: Average Review Length = 620, Standard Deviation = 344

4 Ratings: Average Review Length = 586, Standard Deviation = 338

5 Ratings: Average Review Length = 555, Standard Deviation = 327

Assuming that review length and ratings are not related, then the average review length of each rating should be the same as the overall average review length. We will test to see whether the actual average review length of each rating is statistically significant.

1 Rating: t score = 0.07

2 Rating: t score = 0.24

3 Rating: t score = 0.10

4 Rating: t score = 0.01

5 Rating: t score = -0.09

We could calculate the p-values, but it's probably safe to say that none of these are statistically significant. This seems to support the null hypothesis then that Review Length does not have an effect on Ratings.

Conclusions

The data seems to suggest that review length does not have a particularly strong effect on the rating given for hotel reviews. A more useful analysis, though outside the scope of this particular course, would be to use a sentiment classifier, which assigns a score through machine learning based on the type of words used in the reviews based on the positive or negative connotation of the words. This will likely be something we can explore later in this specialization.

The data set itself is fairly simple, with only two columns worth of data. It was certainly not designed for simple linear regression analysis, but instead for a more complicated machine learning approach. With tools we can learn later in this specialization, I believe we can take a better approach to this data than we can with the tools describe in this course.

Source

Codebook for Report is located here: https://github.com/nanl00127/Exploratory-Data-Analysis/blob/main/TripAdvisor%20Analysis.ipynb

Source of data: https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews