MGMT 59000 – WEB DATA ANALYTICS – FINAL PROJECT SUMMARY

Introduction and Problem Statement

The hospitality industry has been affected immensely due to Covid 19. Considering this, any hotel chain that would like to open a branch would prefer to do a data driven research before finalizing on the location and features of the hotels, and what kind of customers the hotel should focus on to maximize its performance. This analysis could also be used by underperforming hotels. The hotel rank has been used as a proxy for the hotel's performance in the analysis.

Data Collection and Variable Selection

The Tripadvisor data was scraped for 3262 hotels in California over 22 different parameters. The reviews data was also scraped for the top 10 hotels in California. To determine which variables have an impact on the rank of hotels, Linear Regression was used on the data and the following were significant (based on p-value at 0.05 significance level) – Rating, Number of Reviews, Having Covid Safety Measures, Location, Description Length, Having Non-Smoking Rooms. This regression was done after eliminating the highly correlated variables using a correlation matrix.

Hypothesis Testing and Analysis

Based on the significant variables, the following hypotheses were tested on the data at a 0.05 significance level –

Impact of Having Covid Safety Measures – A t-test was performed on the average ranking of hotels that have and do not have covid safety measures implemented. The results indicated that hotels that have such measures implemented have lower average ranking than those that do not.

Impact of Having Non-Smoking Rooms – A z-test was performed on the proportion of hotels that were rated 4 and 5 between those that had non-smoking rooms and those that did not. It was found that the proportion of hotels that had non-smoking rooms was greater.

Impact of Description – A t-test performed on ranking based on description length showed that the hotels that have a description of more than 100 word had significantly better average ranking.

Impact of Number of Attractions – Surprisingly, the t-test showed that the number of nearby attractions caused no statistically significant difference in rankings.

Impact of Number of Restaurants – However, the t-test performed on number of restaurants nearby showed that hotels having more than 40 restaurants nearby had a significantly smaller value of ranking than those that do not.

Impact of Traveler Location – Conducting a z-test for proportions showed that the percentage of negative reviews given by Californians is lower than those by people from other locations for the top 10 hotels. The review sentiment was calculated by sentiment analysis using Textblob.

Impact of Trip Type – Similarly, reinforcing the above result, a z-test showed that the % of business users rating 4 and 5 was higher for people in California for the top 10 hotels.

Topic Modeling – Topic modeling was performed using genism and lda on the top 100 hotel descriptions to find out the key words to find out what ranks them better. Based on this, we were able to discern that travelers prefer hotels with pool view, near restaurants and beach area.

Recommendations

- New or existing hotels should have COVID safety measures implemented and have it listed on their Tripadvisor page. Also, they should have some rooms as Non-Smoking rooms.
- Provide a detailed description of hotel for guests to make an informed decision on their stay.
- Have the hotels near restaurants rather than attractions for better performance. New upcoming hotels should not spend too much money on premiere locations near attractions.
- Focus more on travelers from other locations as these customers are not as satisfied, especially business travelers.