**MGMT 590 Web Data Analytics**

**Executive Summary**

We went ahead to analyze the Airbnb listings in Seattle, Washington, to gauge the impact of various parameters on the rating received from the customers.

**Business Hypotheses:**

This model could help Airbnb owners achieve higher ratings, which eventually would lead to more visibility on the Airbnb portal. A higher profile would mean more revenue for the users\*.

\*Assumption:

1. At a given price point, a user would book Airbnb with a higher rating.
2. We are assuming a few parameters will remain constant, and the owner will not/ cannot make these changes, i.e., Number of Rooms, Number of guests, square foot, and locality.

**Process:**

1. **Natural Language Processing:** For each review, we calculated a review sentiment score using Google NLP API and took into consideration only the reviews which were in supported languages.
2. **Merging Data Sources:** The reviews sentiment score we generated from Google NLP API and listing data was available from different data sources. We had to join the data and get rid of Airbnb listings with null reviews.
3. **Data Cleaning:** Regarding columns for price, cleaning fee, security deposit, and extra cost per person- we had to do string manipulation to convert them to continuous variables.
4. **Linear Regression:** After cleaning the data, we removed the highly correlated variables. Keeping the sentiment score as our target variable, we did variable selection using the p-value threshold (0.05) and removed insignificant variables. Post that, we ran the linear regression model.

**Findings:**

1. We see that increasing the price will have a negative impact on the review rating. However, price clubbed with a higher cleanliness score will have a positive effect beyond a certain cleanliness score.
2. The coefficient of (extra people\*cleaning fee) is negative, which suggests that if the number of additional guests in the booking increases as well as the cleaning fee, it negatively impacts the rating.
3. Having a security deposit instills a sense of genuineness for the host, hence the rating increases with an increase in security deposit, as corroborated by the positive coefficient in the regression equation.

**Limitations:**

1. There were cases with no apparent review for the hosts. Hence our data size got reduced when merging between the listing and the review file.
2. Only reviews in languages supported by Google NLP API were considered as the NLP software is not trained for comparing rating across languages (non-supported reviews, Danish and Dutch, <1%).
3. Since we do not have actual Airbnb sales data, we have considered sales to be proportional to the sentiment score.