Airbnb Review Analysis

There is no doubt that the rapid growth of Airbnb has changed the lodging industry and tourists' behaviors. Airbnb welcomes customers and engages them by creating and providing unique travel experiences to "live like a local" through the delivery of lodging services. With the special experiences that Airbnb customers pursue, more investigation is needed to systematically examine the Airbnb customer lodging experience. Online reviews offer a representative look at individual customers' personal and unique lodging experiences. Moreover, the overall ratings given by customers are reflections of their experiences with a product or service. Since customers take overall ratings into account in their purchase decisions, a study that bridges the customer lodging experience and the overall rating is needed. In contrast to traditional research methods, mining customer reviews has become a useful method to study customers' opinions about products and services. Searching for homes is the primary mechanism guests use to find the place they want to book. Ranking at Airbnb is a quest to understand the needs of the guests and the quality of the hosts to strike the best match possible.

Data spread over more than 191 countries and 81,000 cities, Airbnb listings are in every corner of the planet and the definition of the best deal has a lot of local details to it. The categories of homes include ranging from apartments to igloos, even within a standard category such as apartments, when different variations along price, location, amenities, decor etc. For this project, I just selected data for Seattle, WA. for period of 2014 to 2016. I used the dataset provided by Kaggle, https://www.kaggle.com/airbnb/ seattle, where publicly available information about a Seattle's Airbnb's listings and released for independent, non-commercial use. This includes detailed listing information such as no. of rooms, location, and reviews with text description and also separate dataset specifically for dates.

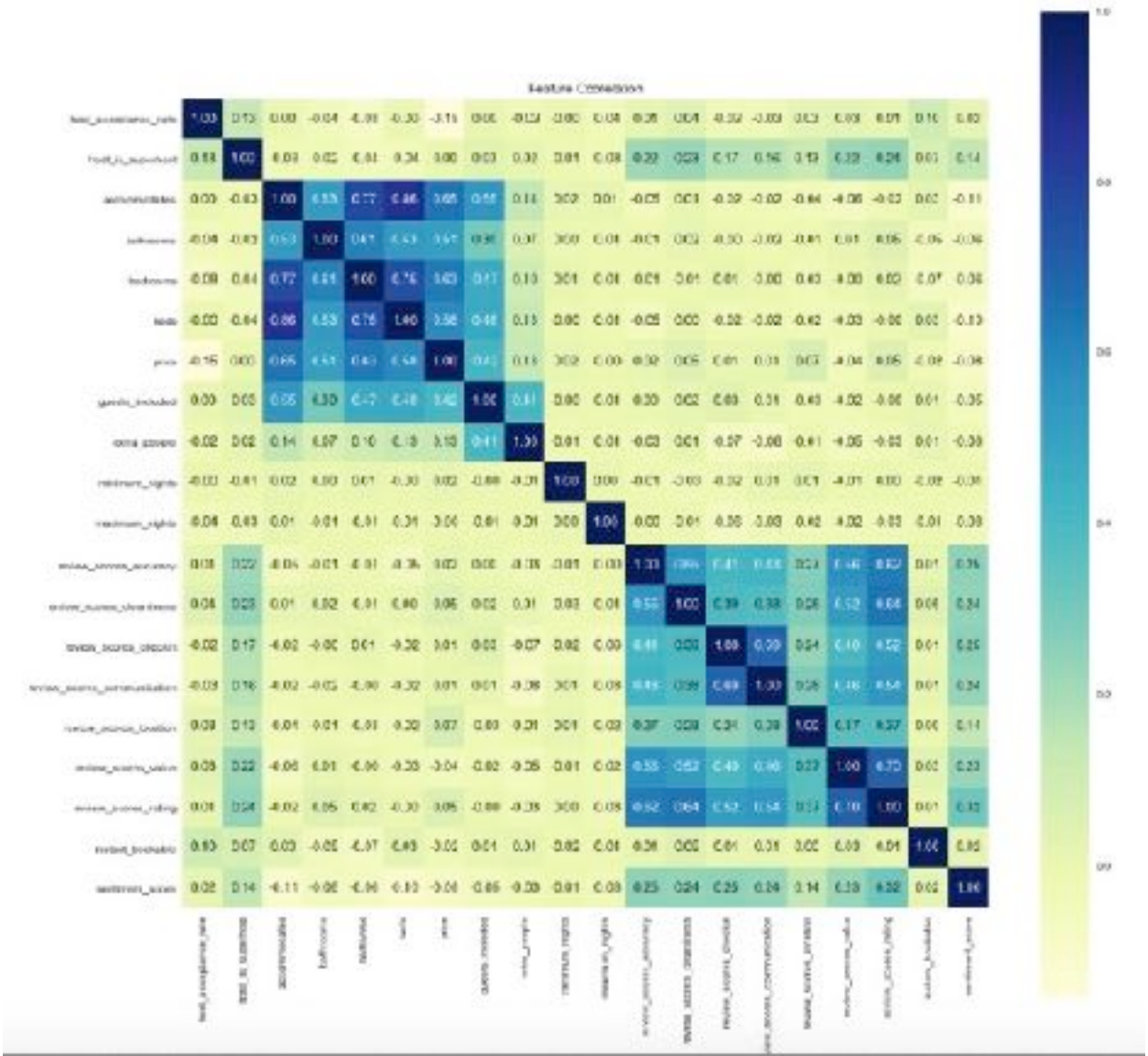
This is a great place to experiment and apply Natural Language Processing (NLP) techniques. Sentiment Analysis, or Opinion Mining, is a sub-field of Natural Language Processing (NLP) that tries to identify and extract opinions within a given text. The aim of sentiment analysis is to gauge the attitude, sentiments, evaluations, attitudes and emotions of a speaker/writer based on the computational treatment of subjectivity in a text.

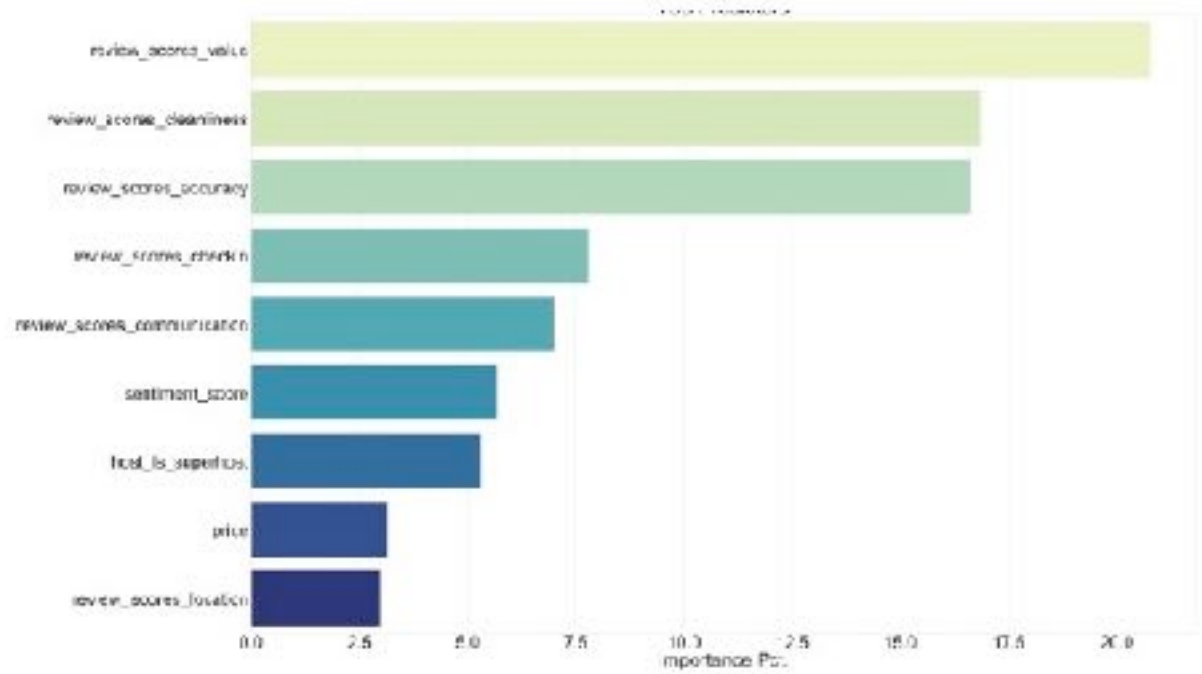
This study also provides practical advice to Airbnb investors and hosts regarding methods to improve the star rating system, emphasizing the importance of addressing the sub-ratings of the various lodging dimensions that more accurately reflect what customers really value and care about.

1. What are the positive and negative reviews?
2. What features impact the reviews waiting?
3. Can we predict the review score based on the reviews?

• **Features Selection**

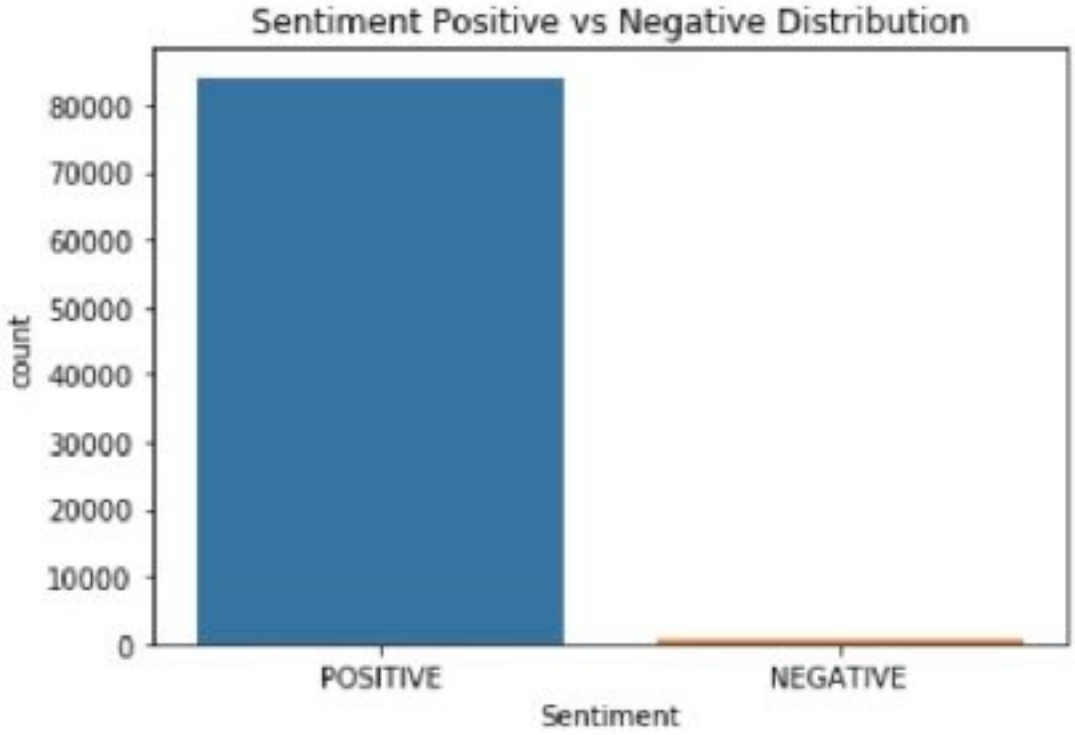
Based on the plot, clients are looking for better highest rating score, cleanliness, and accuracy in reviews. Also, in correlations graph we saw the similar results. Overall it looks like customers are looking for a fairly priced, clean and accurate listing with a smooth check in and open communication throughout the stay. All these factors will play into the sentiment and the host being labeled as a super_host.





- **Text Mining Of Reviews**

I used NLTK package for reviews dataset to calculate sentiment which is positive, negative or neutral. As you can see in the plot, the overwhelming majority of reviews (> 95%) appears to be positive or neutral.



- **Rating Score prediction and Feature Impact**

After selecting features, now it's time to create our prediction model. For this dataset, I chose Random Forest method to predict review score based on historical data. Get a more in depth perspective on the sentiment of these reviews and compare the differences between the two types. For the sake of time and computing power, I decided to only pull a random sample of 80% of the comments from the data. This way we get an accurate interpretation of the comments left by customers in a minimal time frame.

- **Classification Models and Accuracy**

My main focus on this analysis was to distinguish between a good review, bad review and what can be said about both. Also, we are trying to predict the score for the review. So, in order to do so, I had to convert this from a regression problem into a classification problem. By using Random Forest classification, logistic regression, KNN, Decision Tree classifier and SVC models, I was able to train the data and predict the rating score with accuracy of 98.86% percent. The best model for this project is Random Forest Classification. Since this was a classification problem, at the end I used confusion matrix to summarize the result.

- **Conclusion**

Overall it looks like customers are looking for a fairly **priced**, **cleanness** and **accurate** listing information with a smooth **check-in** and open **communication** throughout the stay. All these factors will play into the **sentiment**, and the host being labeled as a **super host** and of course it will effect on the **review score**.