BANA 277 LEC B: CUST & SOCIAL ANLYT
Final Project

Enhancing Airbnb Hosting Strategies Through Sentiment Analysis of Customer Reviews

Group: Johnson & Johnson Cerenovus section B

Yiwei (David) Lu
Yi-En (Ivy) Liu
Sorasak Joshi
Wan-Lun (Ivy) Tsai
Dennis Wu
Eunhye Kim

Table of contents

1. Introduction

The travel industry is experiencing large amounts of growth. “The global tourism sector grew to about $1.9 trillion in 2023, a significant increase from the previous year and expected to continue to grow in the upcoming years” (https://www.statista.com/statistics/1220218/tourism-industry-market-size-global/). This spike in demand has created a highly competitive short term stay market (comprising home stays, hotels, rentals and hostels). Airbnb has been able to set itself apart from the competition by offering trustworthy service, worldwide access, and ease of use. The platform heavily relies on customer reviews to influence booking decisions of travelers. Our project aims to utilize sentiment analysis in order to analyze Airbnb customer reviews in order to find patterns in guest satisfaction and dissatisfaction. Our goal is to understand customer sentiment in order for hosts to enhance guest experiences and optimize business strategies (such as service quality, pricing strategies and customer retention strategies).

2. Objectives

For this customer and social analytics project using the Airbnb Listings 2016 dataset, our primary focus is to develop insights and recommendations for Airbnb hosts in order to help them improve the experience they provide to their customers. In order to do this we analyze booking patterns, price sensitivity, and review trends. Additionally, we will conduct a review sentiment analysis and train a model to develop a review deception detection application for hosts to utilize. This application will allow the hosts to import their latest reviews and quickly determine whether they are positive or negative, helping improve customer insights and make the necessary changes in order to provide their customers with a better experience.

The key objectives of this analysis are:

- Customer Segmentation – This involves categorizing customers based on their booking behaviors. These insights will allow Airbnb hosts to understand the various target markets they are catering to and optimize their listings/improvement and price efforts based on who they intend to target.
- Booking Pattern Analysis – Examining trends in customer reservations. This includes seasonal demand trends, location, room type preference. This will help us provide

insights into demand periods, preferred property types and booking habits, allowing the hosts to understand their customers better.

- Price Sensitivity Analysis – Understanding how price influences booking decisions. This will help hosts = determine optimal pricing strategies, to maximize their profits
- Review & Rating Analysis – Exploring review trends and their impact on listings in terms of satisfaction drivers and so that hosts can understand what can be improved in order to improve guest experience. Additionally, we will generate word clouds to visualize common themes in reviews, helping hosts quickly grasp recurring feedback trends.
- Review Sentiment Analysis – Developing a model to classify and detect deceptive reviews for hosts.

3. Dataset Description

We utilized the **Airbnb Listings 2016** dataset from Kaggle, which comprises three data files: Listings, Reviews, and Calendar. The Listings file contains 3,818 rows and 92 columns, providing a comprehensive summary of Airbnb postings, including details such as room type, location, price, and host information.

The Calendar file consists of 1,048,575 rows and 4 columns, recording the availability and pricing of each listing over time. Lastly, the Reviews file includes 84,849 rows and 6 columns, linking each listing ID with user reviews and reviewer IDs, offering insights into customer experiences and feedback.

- Listing (3818, 92)
- Reviews (84849, 6)
- Calendar (1048575, 4)

4. Methodology

To conduct an analysis of the Airbnb Listings 2016 dataset, our team followed a structured approach which included data preprocessing, exploratory data analysis (EDA), and sentiment analysis modeling to uncover key insights related to customer behavior, pricing and review sentiment.

4.1 Data Preprocessing

Before proceeding with the analysis, we first checked for missing values and ensure data integrity. Reviews with missing text were removed to maintain the quality of the sentiment analysis.

4.2 Exploratory Data Analysis (EDA)

We using several analytical techniques to explore the relationships between different variable in the dataset:

- Scatter Plots
- Clustering Analysis
- Correlation Heatmaps
- Histograms with normal distribution
- Word Cloud Analysis

4.3 Sentiment analysis model

To focus on building a sentiment analysis model, we first applied TF-IDF (Term Frequency-Inverse Document Frequency) to convert text data into numerical features, capturing important words while reducing noise.

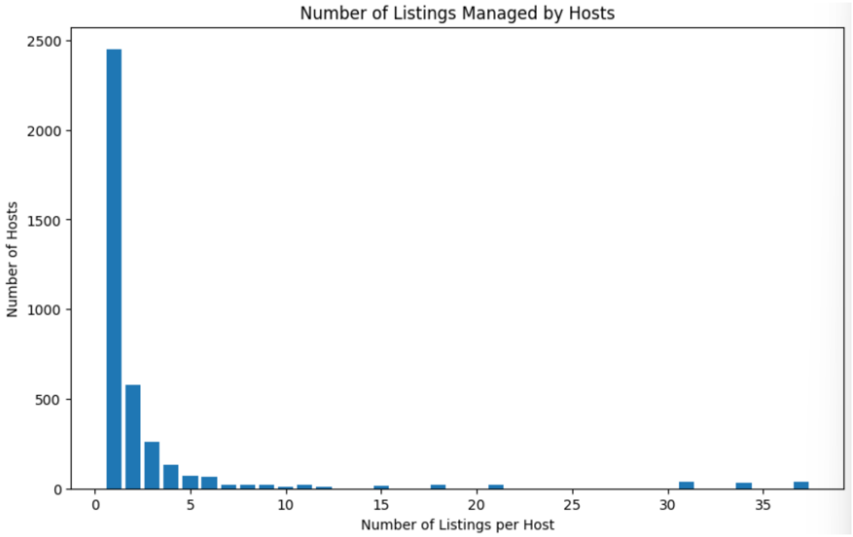
For model training and selection, we trained and compared two classifiers: XGBoost and RandomForest, assessing their performance on the dataset. After training and testing, we evaluated the models using classification metrics to determine their accuracy in predicting positive and negative reviews. Once the model was built, we implemented it in PyCharm to develop an application that allows users to quickly analyze and classify Airbnb customer reviews.

5. Sentiment Analysis and Review Classification

5.1 Host Segmentation

5.1.1 Host-Listing Relationships: Clustering hosts by properties management style

Before we can curate recommendations for our hosts, it is important to understand the different types of hosts we are catering towards. In order to do this we were able to draw from trends in the data. The histogram below helps visualize the distribution of listings that each host has.



As we can see, most of the hosts have a single listing with a sharp drop off as the number of listings decrease. This indicates that the majority of hosts are individual/small scale hosts. There are a few outliers present in the data with some individuals managing over 30 properties which could indicate these are commercial hosts. This helps us understand that there are two types of hosts we are catering towards, the individual/small hosts who mostly have 1-5 listings and large commercial hosts who have a larger number of listings. Although this gave us two distinct types of hosts, we believed that it would be beneficial to have a more deeper understanding of the characteristics of each host.

5.1.2 Host Clustering: Labeling distinct host communities

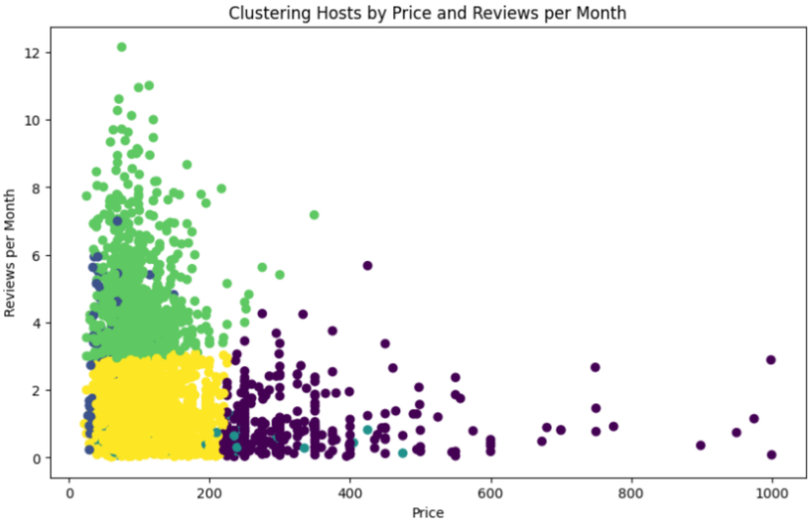
In order to get a deeper understanding of different types of hosts, we used clustering to distinguish them based on the number of listings they manage, price, review and review scores. Our analysis highlights two key observations:

(1) Cluster Distribution

- Green Cluster: Might represent budget listings where hosts are more engaged with their guests and provide more frequent interactions, potentially leading to higher review frequency.
- Yellow Cluster: Hosts with medium prices and moderate reviews per month. These could be more balanced listings, perhaps targeting a middle-range market
- Purple Cluster: Represents luxury listing where hosts manage fewer listings or focus on high-end experiences with less frequent guest interactions, resulting in fewer reviews per month.

(2) Clustering Pattern

- Higher-priced (luxury) listings tend to receive fewer reviews per month, suggesting lower guest turnover or less frequent interactions.
- Lower-priced (budget) listings tend to receive more reviews per month, indicating higher guest volume and engagement.



Budget-friendly hosts (Green Cluster) can enhance guest satisfaction through personalized engagement, leveraging their frequent interactions. Meanwhile, luxury hosts (Purple Cluster) may benefit from encouraging more guest reviews to build credibility. Airbnb can use these insights to personalize marketing strategies, catering to different guest preferences.

The scatter plot analysis reveals key insights into the relationship between the number of listings a host manages and their review frequency. We categorize the following three key observations.

(1) Inverse Correlation:

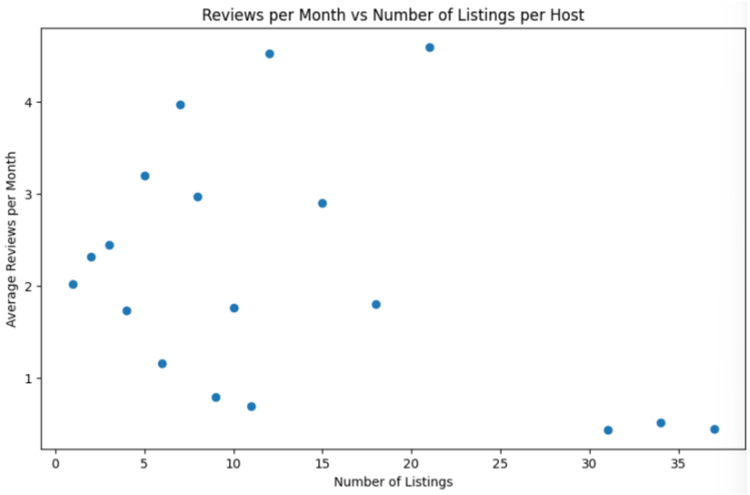
- There is an inverse correlation between the number of listings a host manages and the average reviews per month.
- Hosts with fewer listings tend to receive more frequent reviews.
- Hosts with many listings receive fewer reviews per listing on average.

(2) Interpretation of Review Patterns:

- Hosts with fewer listings are likely more engaged with guests, leading to higher review frequency
- Hosts with many listings often manage properties across various locations, reducing their direct interaction with guests. They tend to focus more on operational efficiency rather than personalized engagement, leading to a lower review frequency.

(3) Potential Impact on Influence

- Smaller-scale hosts (fewer listings) may have a greater influence on guest experience due to higher engagement and more frequent reviews.
- Larger-scale hosts may be less connected with individual guests, leading to lower engagement and fewer reviews per listing.



This scatter plot helps us understand these patterns, which can help tailor strategies for different host types. For the larger-scale hosts, we encourage them to improve guest engagement by incorporating more personalized services. And for small-scale hosts, we advise leveraging the strong guest interactions for targeted marketing and customer retention.

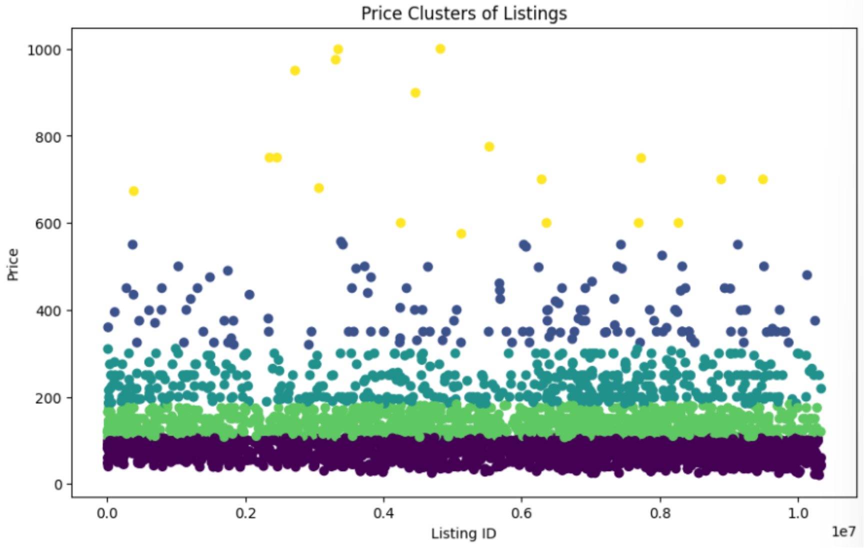
5.2 Booking Pattern Analysis

5.2.1 Host Interaction Based on Price Similarity: Identifying market strategies

We use scatter plots to visualize clusters of listings based on price, with each cluster represented by a different color. The yellow cluster, representing the highest price range around 600-1000, likely indicated luxury listings. Other clusters such as green, blue and purple correspond to lower-price listings.

Price Clusters:

- Lower Price Range: Purple and green clusters likely represent budget listings (around $100 to $250), offering more affordable accommodations.
- Mid-Range Price: Blue and green clusters represent mid-range priced listings targeting a broader audience.
- Higher Price Range: Yellow cluster indicates luxury listings with higher price tags.

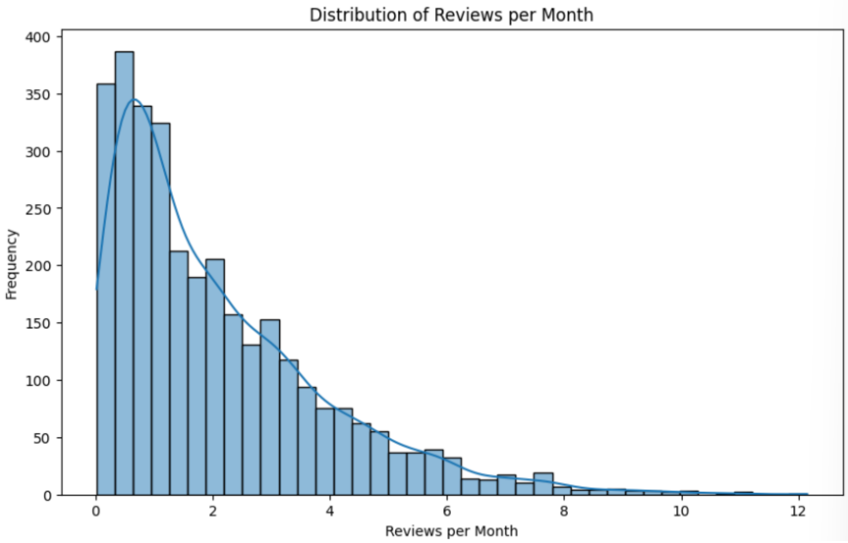


Hosts could be segmented based on the price range of listing. Luxury listings may attract customers willing to spend more comfort, while budget listings appeal to price sensitive travelers.

Therefore, by analyzing price clusters, we can identify different market segments and understand how hosts with similar pricing strategies may be competing or cooperating in the same space.

5.2.2 Review Frequency Analysis: Understanding host activity and booking rates

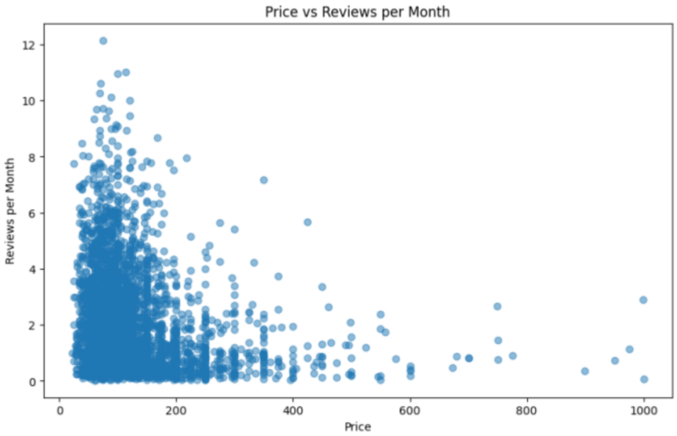
To gain insights into host activity and booking rates, we analyze the distribution of reviews per month. The histogram below shows the frequency of monthly reviews received by the host. This distribution is right-skewed, indicating that most hosts receive few reviews per month, while a smaller subset of hosts accumulates a significantly higher number of reviews.



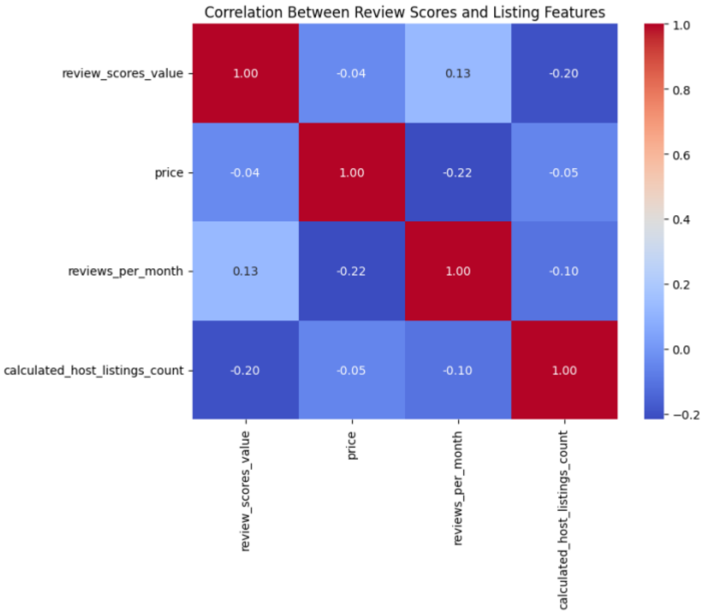
This pattern suggests that a majority of hosts have lower engagement levels, possibly due to fewer bookings or less frequent guest interactions. Conversely, hosts who receive reviews more frequently may be more active on the platform, offering higher availability or maintaining stronger guest relationships, which could contribute to increased booking rates.

5.3 Understanding what drives customer reviews

Customer satisfaction is an important factor in driving the success of a host's listings. Positive reviews help increase visibility and increase the likelihood of a guest booking. When hosts receive an increase in positive reviews, this may increase the demand of their properties. Through our analysis our team has uncovered the different factors that contribute towards guest satisfaction.



An interesting finding from our study showed that there is a negative correlation between price and the number of reviews per month and the price. This shows that high priced listings receive fewer reviews while lower priced listings have higher engagement rates. We can infer that budget friendly hosts tend to attract more guests leading to more reviews. Another reason could be that the higher priced premium listings may offer greater amenities and target a different type of traveler (who may be less likely to leave feedback).

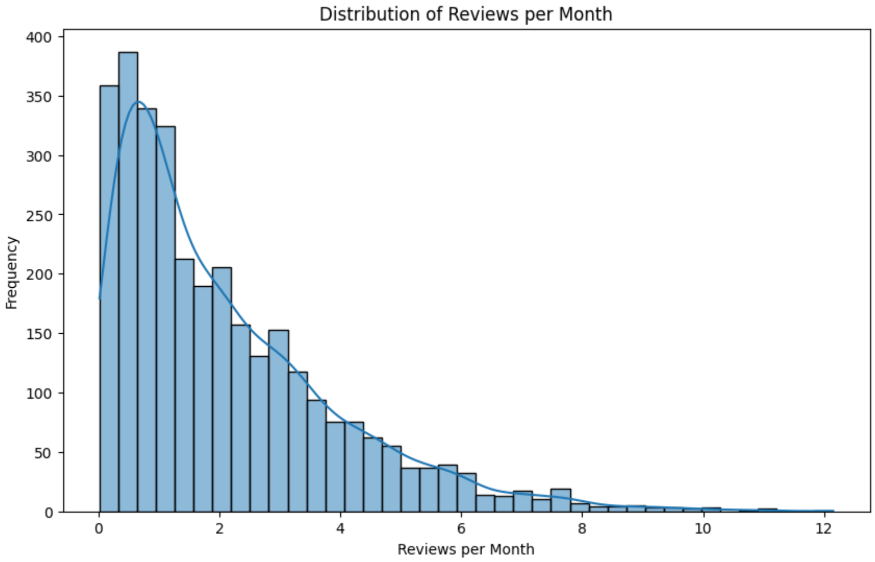


Important drivers of review scores

One of the most significant findings our analysis uncovered was the weak negative correlation (-0.2) that is between the number of listings a host manages and their review scores. This suggests that hosts that manage multiple listings tend to receive slightly lower ratings than hosts that manage lesser listings. This may be due to the fact that smaller scale hosts are more engaged with their guests and are able to provide personalized experiences, quicker responses and have more direct communication. This may be difficult when a host has multiple listings to manage which may lead to a slightly lower guest satisfaction level.

An important insight we found was the negative correlation (-0.22) between price and the number of reviews per month. This indicates that higher priced listings receive fewer reviews while cheaper listings receive more reviews. This could be due to the fact that budget listings attract more travelers while luxury listings cater to a niche audience.

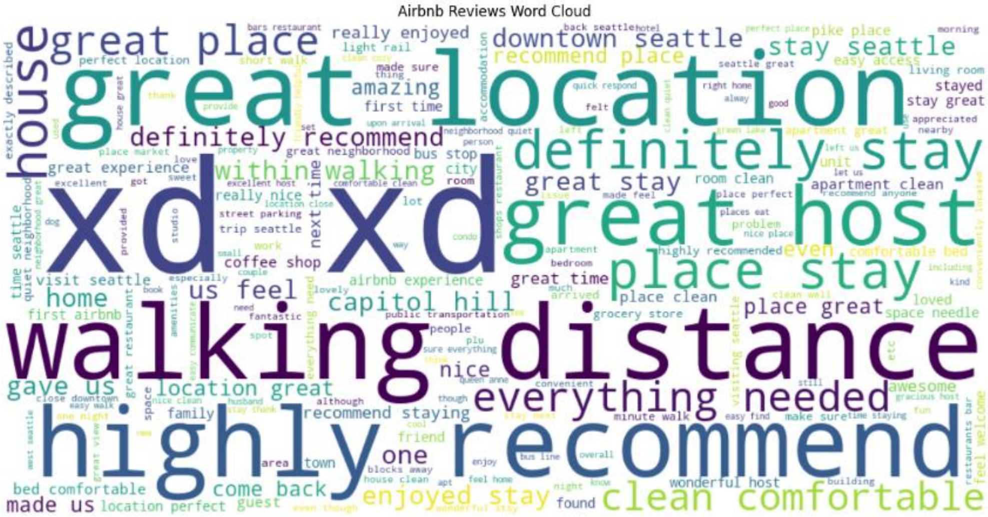
The correlation matrix also showed a small positive correlation of 0.13 between review scores and the number of reviews hosts received per month. This suggests that frequent reviews tend to be associated with slightly higher ratings. This helps show that active engagement with guests will contribute to better satisfaction.



5.4 Review Analysis

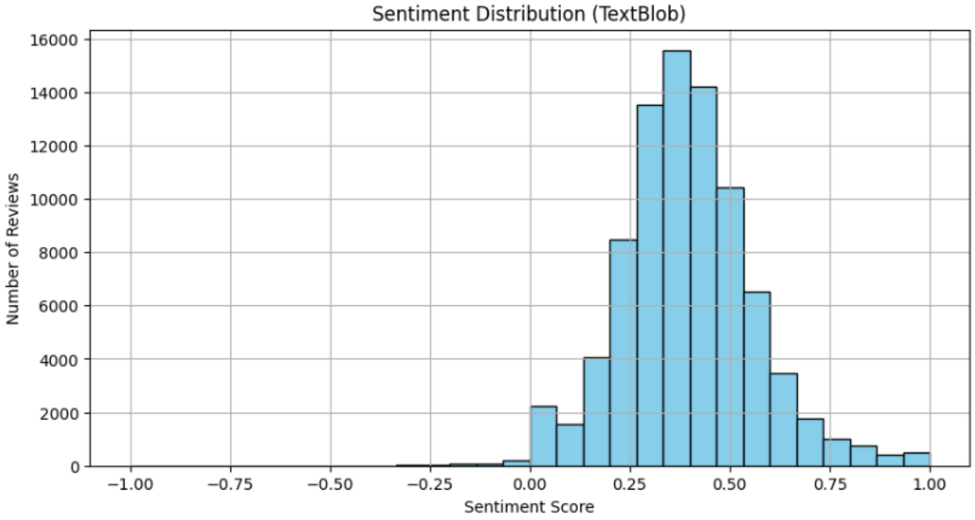
Customer reviews were preprocessed using tokenization, stop words removal and sentiment scoring. Sentimental analysis was conducted using VADER and TextBlob to classify customer reviews into positive and negative sentiments. By evaluating sentiment trends, we aim to identify key themes in customer feedback that reflect guest satisfaction.

The word cloud visualization highlights the most frequently mentioned words in customer reviews, providing insights into common topics. Larger words indicate higher frequency, with prominent terms including “location,” “walking distance,” and “great host.” This suggests that customers using Artbnb highly value the location of the property, whether it is within an appropriate walking distance, and appreciate a clean and comfortable environment.

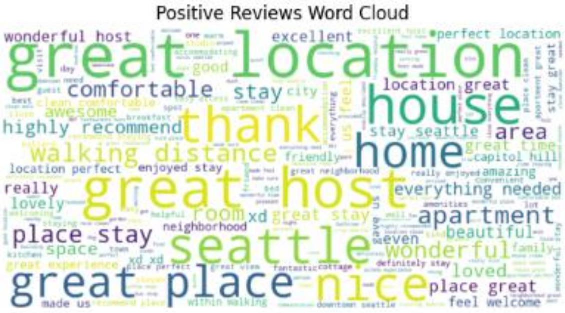


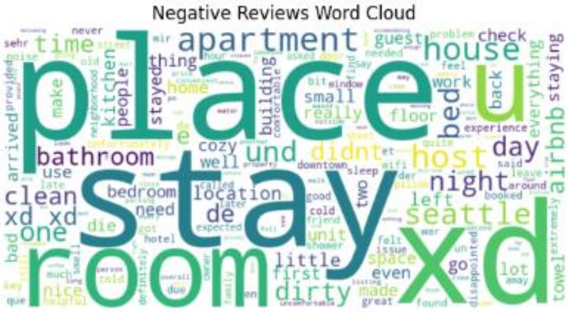
To further analyze sentiment distribution, we used a bar chart to visualize the sentiment scores obtained from TextBlob. The overall trend indicates that customer sentiment tends to be positive. Specifically, we filtered reviews as follows:

- good_reviews = df[df['sentiment_blob'] > 0.5]['cleaned_comments']
- bad_reviews = df[df['sentiment_blob'] < 0]['cleaned_comments']

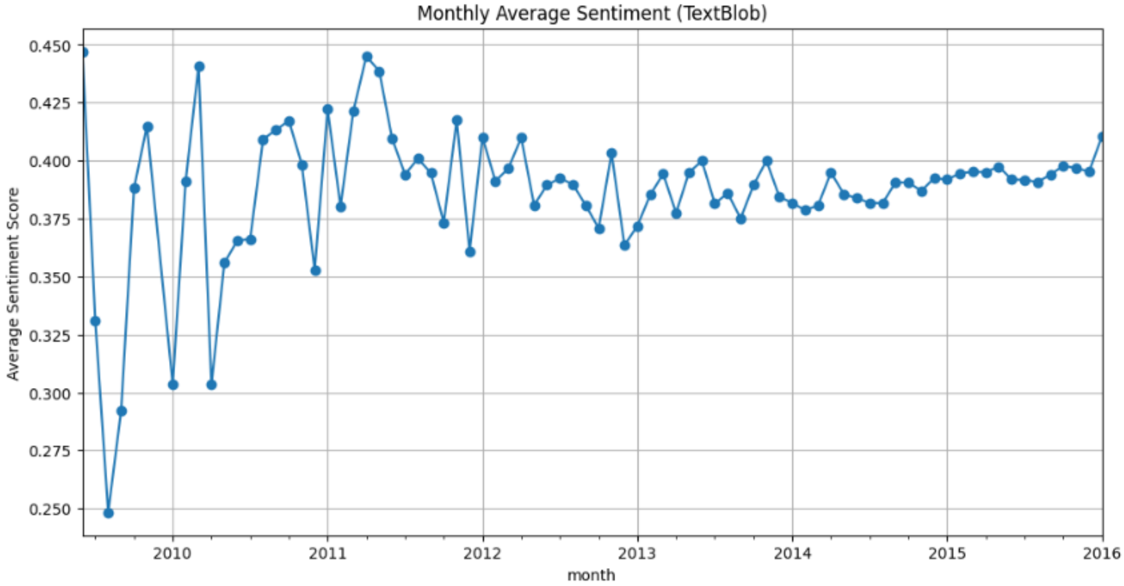


Next, we compared positive and negative reviews using separate word clouds. The negative reviews were more focused on issues related to cleanliness and the hosting experience, indicating dissatisfaction when the property did not meet cleanliness expectations or lacked proper hosting. In contrast, positive reviews emphasized a good location and a great host, reinforcing the importance of these factors in guest satisfaction.





Finally, we conducted a time-series analysis of sentiment trends by month. The results show that between 2009 and 2010, sentiment scores exhibited noticeable fluctuations. However, from 2012 onward, sentiment scores became more stable, suggesting that hosts gradually adapted to customer preferences and were able to provide a more consistent and satisfactory experience.



6. Strategic Business Recommendations

Based on the insights fathered from the sentiment analysis, price sensitivity, review scores and understanding review trends. Our team has developed recommendations to help Airbnb hosts improve their listings, increase guest satisfactions and increase profitability.

Our first recommendation is to improve guest engagement and personalization. Hosts that manage a smaller number of listings are seen to receive more reviews and higher ratings which may be due to higher interactions between the hosts and guests. With the negative correlation between the number of listings and review scores, we would recommend larger hosts to implement automation tools for guest engagement. Personalized messages such as check in instructions and follow ups that can be automated can create a personalized experience for the guests even when the host is managing multiple listings. These personalized experiences may lead customers to provide reviews in order to improve credibility.

Our second recommendation is based on the relationship of pricing and reviews. Budget friendly hosts should take advantage of the frequent reviews they receive, they should ensure that they maintain a consistent quality, have fast response times, and keep their listings well maintained in order to continue benefiting from this trend. Moreover, if a host is looking to increase their overall review ratings, providing improved service will definitely lead to an increase.

On the other hand, hosts that have high end listings would benefit from encouraging guests to leave reviews of their stays. This can include follow up messages, or small appreciation gestures such as a welcome gift in order to increase the likelihood of receiving reviews. This will lead to improved listing rankings and visibility helping the hosts in the long run.

For our third recommendation, we would like to recommend hosts to take advantage of the correlation between review scores and the number of reviews per month. Hosts should ensure that they are improving guest interactions and encouraging their guests to leave feedback. This can be done by sending reminders to guests to leave reviews and providing small value added perks to enhance guest satisfaction and experience. The hosts should ensure that they are responding to all the reviews that are good/bad and offer seamless guest experiences.

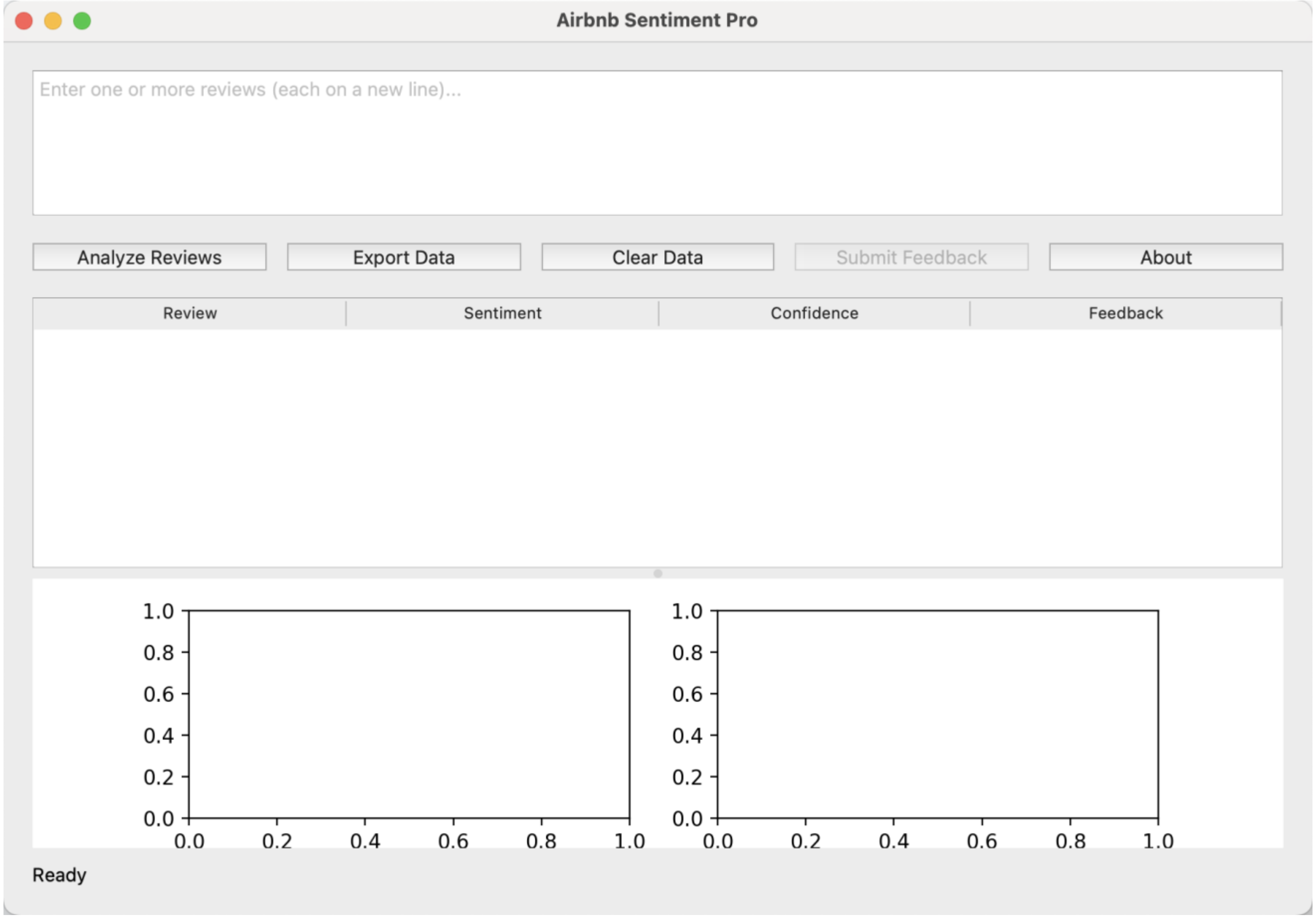
Our fourth recommendation is based on analysis on the frequency of words from good/bad reviews. As seen in the word cloud, negative reviews are linked to cleanliness and unavailability of the host. Positive reviews are based on well prepared locations and responsiveness of hosts. Based on these results, we can recommend hosts to ensure they are keeping their properties clean and inspect them well before guests arrive and hosts should also be available and provide prompt responses to their guests to ensure maximum guest satisfaction.

Our final recommendation is based on the pricing of locations. Different types of listings attract different segments of customers. For budget travelers hosts should prioritize affordability, for mid range travelers hosts should look to provide properties that balance affordability and service for high budget travelers hosts should prioritize additional amenities. It is important for the host to understand which customer segment they are looking to target rather than having a one size fits all approach (which may lead to an increase in negative reviews due to them not meeting standards of a customer).

In order for hosts to better understand their customers and categorize their reviews in an efficient manner, our group has created a review Sentiment Application that provides an intuitive interface for Airbnb hosts.

App Interface :

- **Top Section:** An input box where users can enter multiple reviews simultaneously.
- **Middle Section:** Displays sentiment analysis results, including review sentiment, confidence score, and feedback options.
- **Bottom Section:** Visual representations, featuring a word cloud and a bar chart summarizing the distribution of review sentiments.

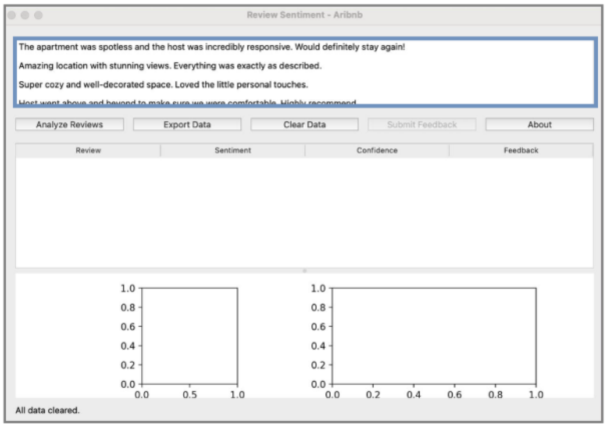


How It Works:

1. **Input Reviews (Testing Purpose):**
   a. Generate 100 mixed positive and negative sample reviews using the prompt: *"Generate 100 mixed positive and negative sample reviews."*
   b. Copy and paste the reviews into the input box.
2. **Analyze Sentiment:**
   a. Click the **"Analyze Reviews"** button to process the input.
   b. The sentiment analysis results will appear in the output table.
3. **Output Table Includes:**
   a. **Review Content:** Original text of the review.
   b. **Sentiment:** Classification as *Positive* or *Negative.*
   c. **Confidence Percentage:** Model's confidence in the classification.
   d. **Feedback Column:** Users can manually correct classifications by selecting *Correct, Unsure, or Incorrect.*

4. **Visualization & Feedback:**

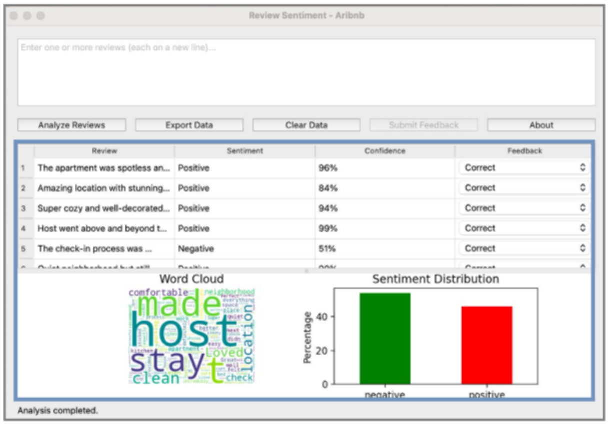
   a. The bottom section provides an intuitive summary using a word cloud and bar chart to display distribution of review sentiments.



The apartment was spotless and the host was incredibly responsive. Would definitely stay again!
Amazing location with stunning views. Everything was exactly as described.
Super cozy and well-decorated space. Loved the little personal touches.
Host went above and beyond to make sure we were comfortable. Highly recommend

Review	Sentiment	Confidence	Feedback

All data cleared.



	Review	Sentiment	Confidence	Feedback
1	The apartment was spotless an...	Positive	96%	Correct
2	Amazing location with stunning...	Positive	84%	Correct
3	Super cozy and well-decorated...	Positive	94%	Correct
4	Host went above and beyond t...	Positive	99%	Correct
5	The check-in process was ...	Negative	51%	Correct
6	Quiet neighborhood but still	Positive	90%	Correct

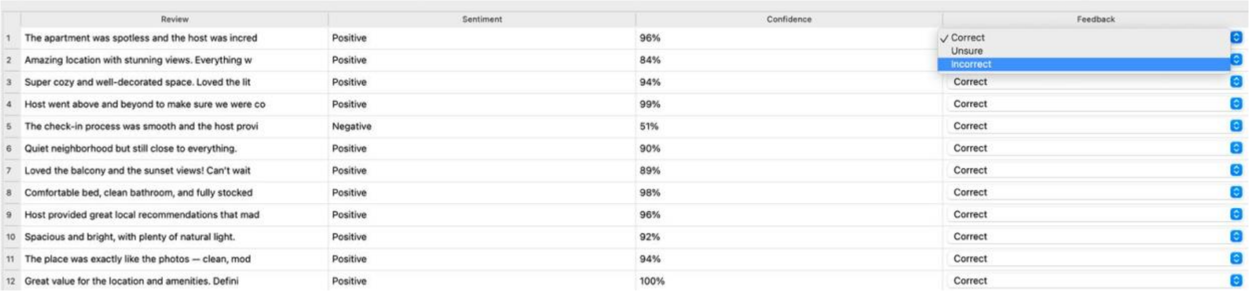
Analysis completed.

5. **Supervised Learning Optimization:**

   a. If users change the Feedback to *Incorrect* or *Unsure,* the **"Submit Feedback"** button becomes active.

   b. Clicking **"Submit Feedback"** generates an Excel file containing corrected labels, which can be used for supervised learning to improve model accuracy over time.

   c. This interactive feedback loop allows the model to learn from misclassified reviews, refining sentiment predictions for better future performance.



	Review	Sentiment	Confidence	Feedback
1	The apartment was spotless and the host was incred	Positive	96%	✓ Correct Unsure Incorrect
2	Amazing location with stunning views. Everything w	Positive	84%	
3	Super cozy and well-decorated space. Loved the lit	Positive	94%	Correct
4	Host went above and beyond to make sure we were co	Positive	99%	Correct
5	The check-in process was smooth and the host provi	Negative	51%	Correct
6	Quiet neighborhood but still close to everything.	Positive	90%	Correct
7	Loved the balcony and the sunset views! Can't wait	Positive	89%	Correct
8	Comfortable bed, clean bathroom, and fully stocked	Positive	98%	Correct
9	Host provided great local recommendations that mad	Positive	96%	Correct
10	Spacious and bright, with plenty of natural light.	Positive	92%	Correct
11	The place was exactly like the photos — clean, mod	Positive	94%	Correct
12	Great value for the location and amenities. Defini	Positive	100%	Correct

7. Reference

Freberg, A. (2016). *Airbnb Listings 2016 Dataset* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/alexanderfreberg/airbnb-listings-2016-dataset