USING NATURAL LANGUAGE PROCESSING TO ANALYSE AIRBNB REVIEWS FOR SEATTLE

DATASET; https://www.kaggle.com/datasets/airbnb/seattle

BELINDA NYAMAI – TEAM LEADER

KINYANJUI KAMAU

RYAN MUSASIA

MARJORIE OPIYO

ABDUBA GALGALO

1.1 BUSINESS UNDERSTANDING

An Airbnb is a community based platform for listing and renting local homes that connects hosts and travellers by facilitating the process of renting without owning rooms. It cultivates a sharing-economy since it allows property owners to rent out private flats. This research aims to better understand what factors are considered when an individual chooses to book an Airbnb and what features contribute most to their experience. Established in 2008, Airbnb has experienced growth in the number of rental listings available and it continues to disrupt the hospitality industry with its service offerings. It has helped guests and hosts to travel in a more unique and personalized way. The company went from a single air mattress for rent to global cooperation valued at more than 30 billion dollars all thanks to its energetic founder- Brian Chesky.

Sentiment analysis is extremely important because it helps businesses quickly understand the overall opinions of their customers. By automatically sorting the sentiment behind reviews, businesses can effectively gauge brand reputation, understand customers and make faster and more accurate decisions. Reviews are extremely important on Airbnb as customers are generally wary of airbnbs with bad reviews, while good reviews will increase the number of bookings you get as a host. This study will build from the data to identify a set of broad themes that characterize the attributes that influence Airbnb users' experience in Seattle.

1.2 PROBLEM STATEMENT

When choosing an Airbnb, apart from the obvious requirements like price and location, customers tend to spend time reading through guest reviews to understand more about the host and the experience they can expect while staying there. The only problem is that this manual effort can be very time consuming.

The main goal of this project is to come up with a way guests can get a concise understanding of prior guests experience without having to read through pages of reviews. Customers are not only interested in knowing whether most reviews were positive they are also interested in knowing what most guests have said about their experience. With this problem framed, the study aims to approach the problem by relevant keyword extraction using TF-IDF (Term Frequency — Inverse Document Frequency) and Text summarizations.

1.3 SPECIFIC OBJECTIVES

- To identify accommodation attributes Airbnb guests use to rate their experience
- To extract sentiments from unstructured customer review texts.
- To build a word cloud with key word attributes customers use in their reviews.

1.4 BUSINESS SUCCESS CRITERIA

Perform sentiment analysis on reviews of comments left by customers and predicting the given scores based on the reviews displayed in the dataset. Produce snapshots (word cloud) of feedback for airbnbs to allow travellers to compare different options at a glance and make the best choice in no time. Recommend solutions that can benefit hotel owners, online travel agencies, booking sites and travel review platforms seeking ways to put their customers in more relaxed mood. Building a model that has an accuracy score of not least than 75%.

2. DATA UNDERSTANDING

2.1 EXPLORING DATA

The data set includes fields that represent features influencing people decision to book an Airbnb. The Airbnb Seattle dataset comprised of three parts; listings and their details, availability information at a day-by-day level and reviews for each listing.

2.2 DATA PREPARATION

The following steps were followed in preparing the data;

- Importing the necessary libraries
- Loading the dataset from the CSV format it was stored in
- Creating a new data frame with the necessary columns for our research
- Cleaning the data
  - Checking for missing values and replacing them with mode
  - Merging columns for effective analysis

2.3 DATA CLEANING

This process will be conducted in three steps and involves;

- Clean text to remove the excess unnecessary part of the text
- Tokenize to split text into smaller pieces
- Document Term Matrix to put into a matrix so a machine can effectively read it.

Clean Text

This will be done in the following steps using Python’s regular expressions library; removing punctuation, changing the text to lowercase letters, removing numbers and removing common non-sensical text.

Tokenize

This is splitting a text into smaller pieces also known as token. The most common token size is a word but it can also be a sentence as is the case in our dataset. This process is mainly accomplished by removing words with little meaning, known as stop words and is achieved using a text representation known as a bag of words model. In this stage we will perform lemmatization which involves grouping together words with the same root or lemma but with different inflections or derivatives of meaning so they can be analyzed as one item. The aim is to take away inflectional suffixes and prefixes to bring out the word’s dictionary form.

Document-Term Matrix

The reason for placing data into a matrix is to store the information into multiple documents. This can be done using scikit-learn's countvectorizer, where every row will represent a different document and every column will represent a different word.

In addition, to countvectorizer, we can remove stop words which are common words that add no additional meaning to text such as 'a', 'the', etc. Generally, the goal for this whole process which we’ve actualized is to get our data in a clean and organized standard format for further analysis. Corpus is a collection of text while Document-Term-Matrix is word counts in matrix format.

3. EXPLORATORY DATA ANALYSIS

This process is conducted to summarize the main characteristics of the dataset, often by using visual methods.

Outputs

They determine the main trends in the data and whether it attempts to convey the intended message. Some of the ways we used to explore our data set include;

Most common words where the most frequently used words were identified and visualized using word clouds and frequency distribution graph.

Document-term matrix was used to aggregate the data by sorting across the rows and finding the words with the highest count. The process involved exploring the dataset to;

- find the top most used words,
- identify the most frequently used two words combination
- identify the most frequently used three word combinations

Finding the top 40 most used words in the dataset

A graph that highlights the frequency of the most used single word in the data was plotted and a unigram ngram range of (1,1) where our output will be the most frequently used word occurring concurrently in our dataset.

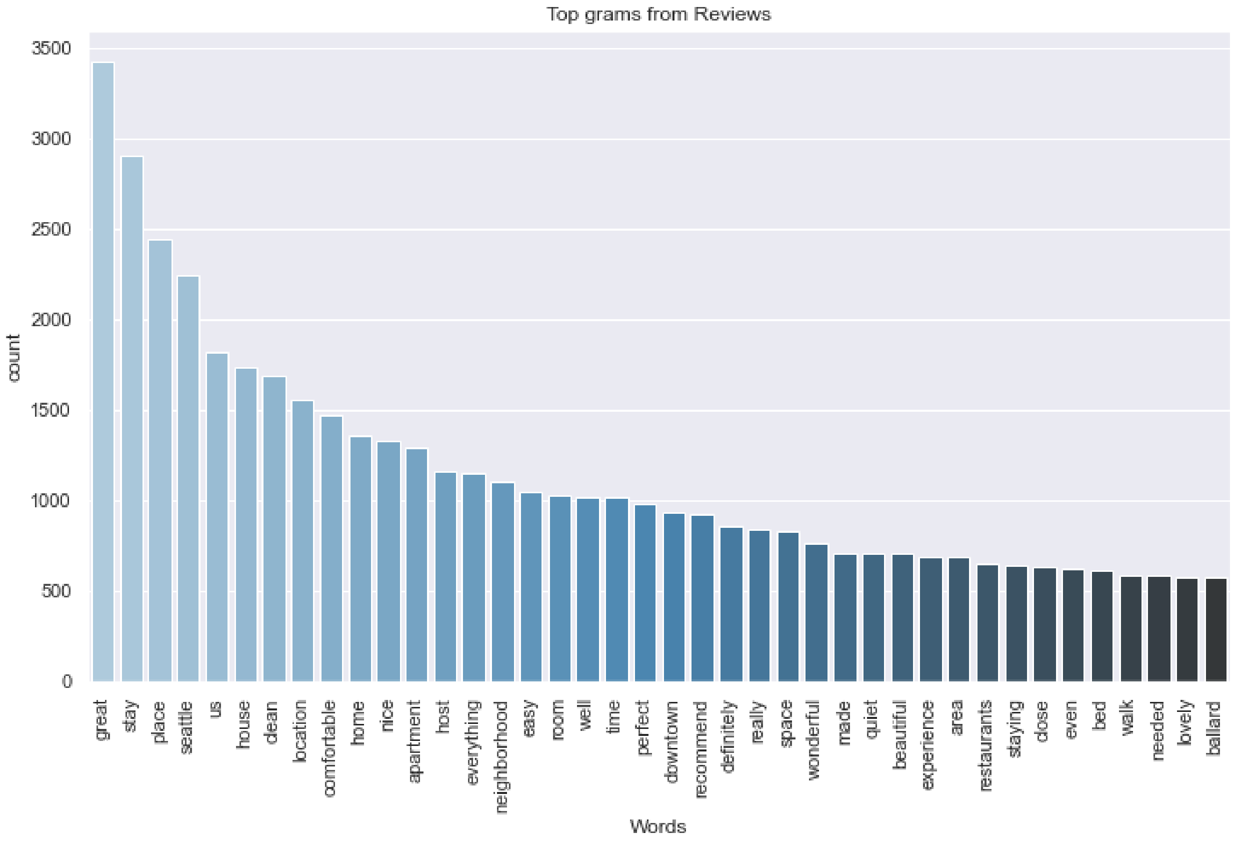
Identifying the most frequently used two word combinations

A bigram ngram range was used in this process, where our output was the most frequently used two-word combinations occurring concurrently in our data.

Identifying the most frequently used three word combinations

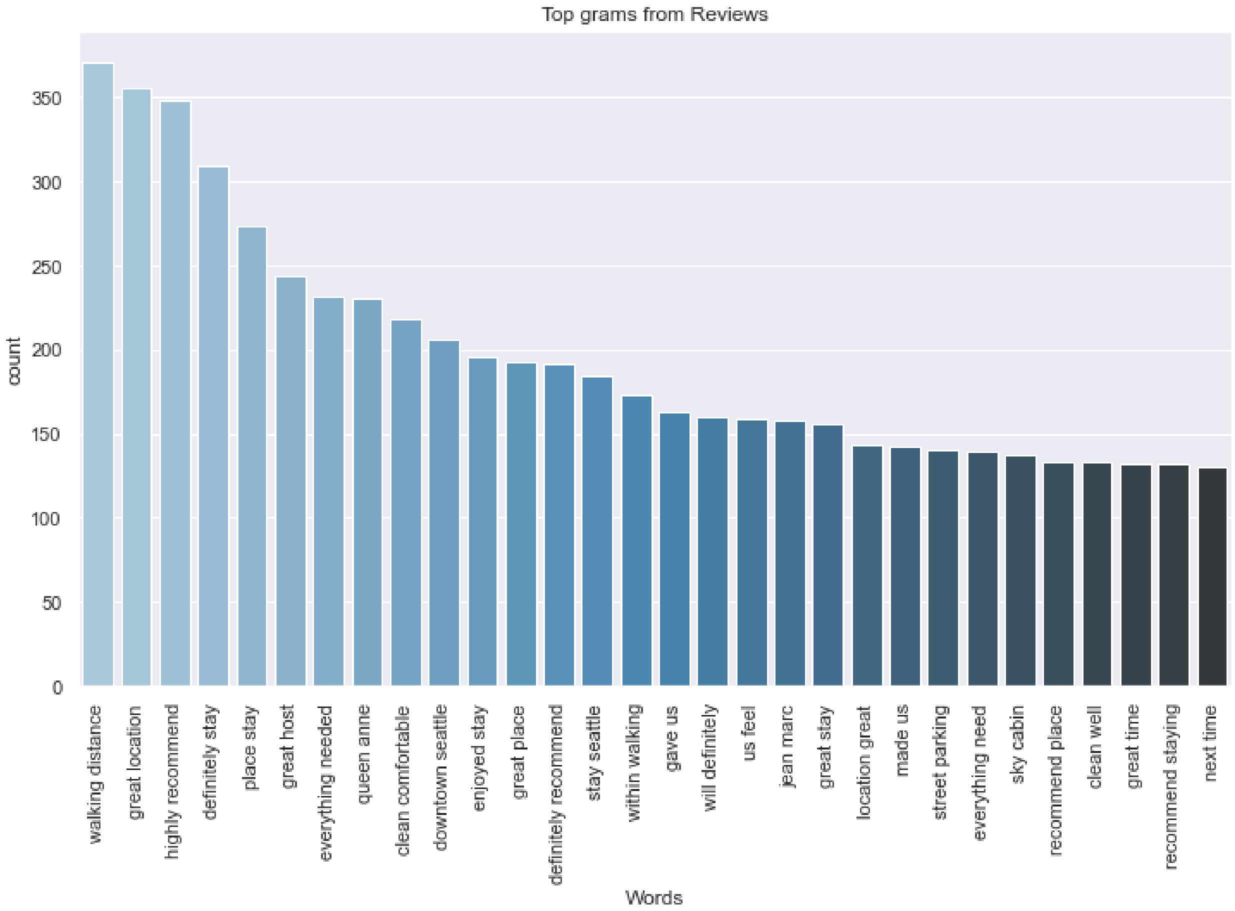
A trigram ngram range was used in this process, where our output will be the most frequently used three-word combinations occurring concurrently in our data.

Top 40 most frequently used words in the dataset



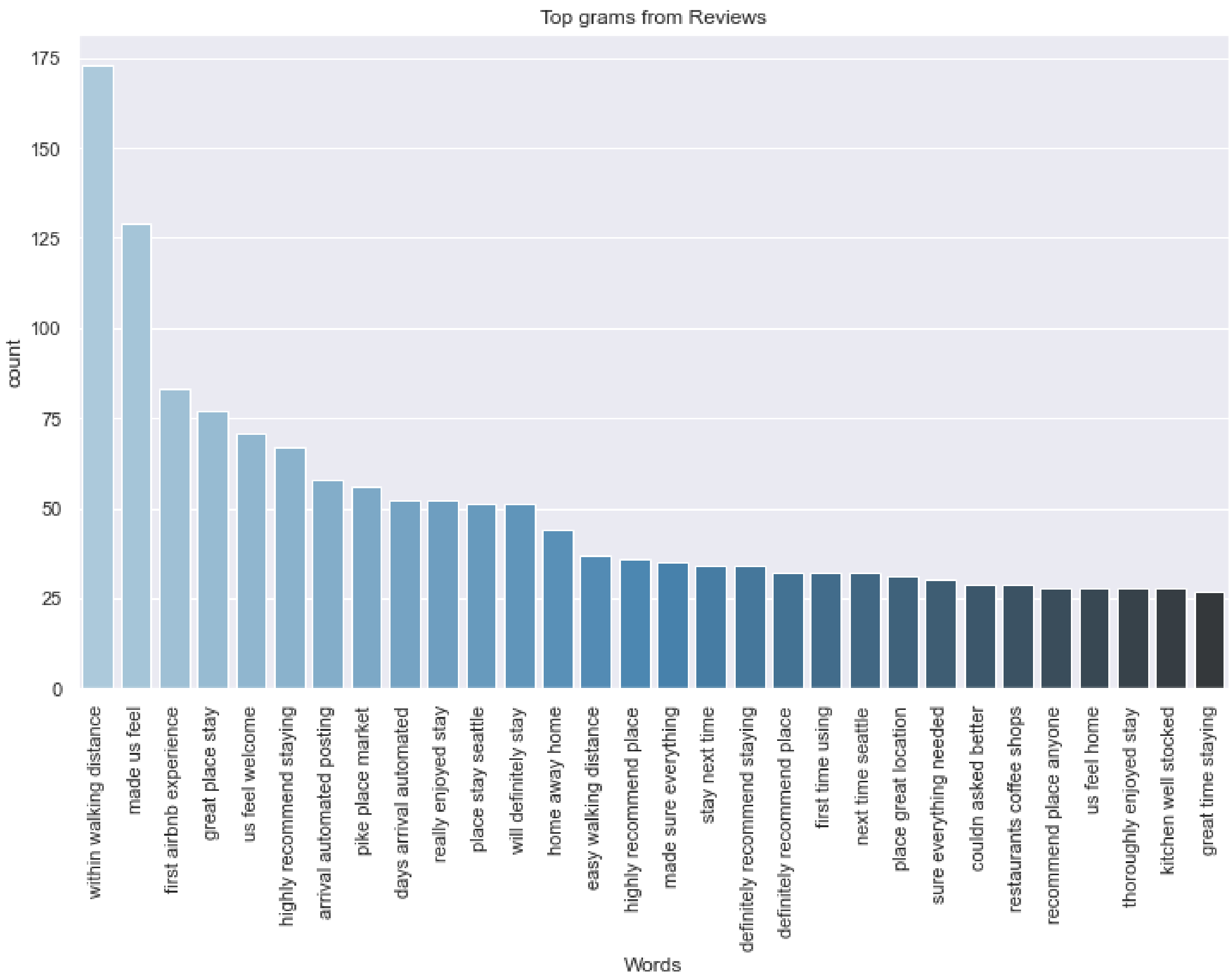
The most frequently used words are great, stay, place and Seattle. This implies that people who go booking for an Airbnb are mainly concerned with the services rendered and the surroundings since they are frequently mentioned when giving reviews.

Most frequently used two word combinations (bigram)



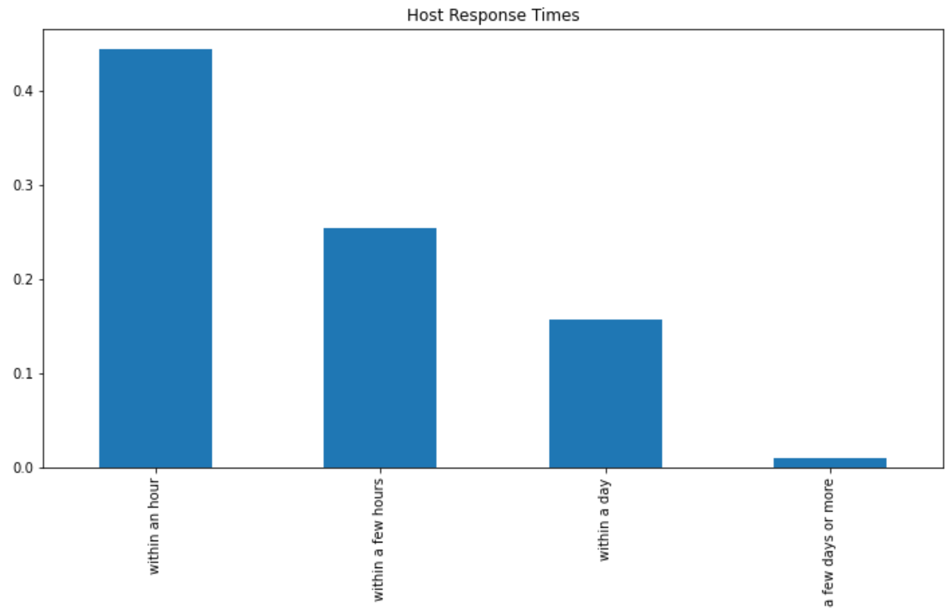
The commonly used two combinations of words are walking distance, great location, highly recommended, definitely stay and place stay. This evidently shows that people staying in Airbnbs are mainly concerned with the location and how comfortable they are when staying there.

Most commonly used three word combinations (trigram)



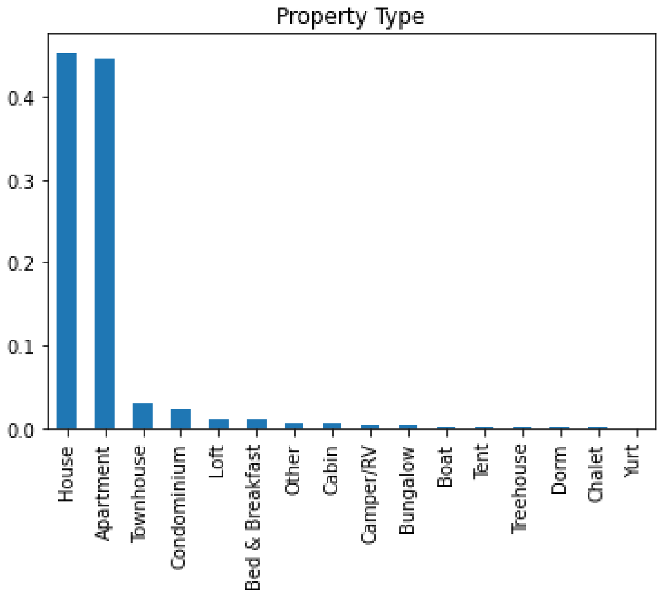
The most commonly used three words combinations are within walking distance, made us feel and first Airbnb experience, great place to stay this emphasizes that people are mainly concerned with the distance covered and the services received when staying at the Airbnb.

Host Response Times



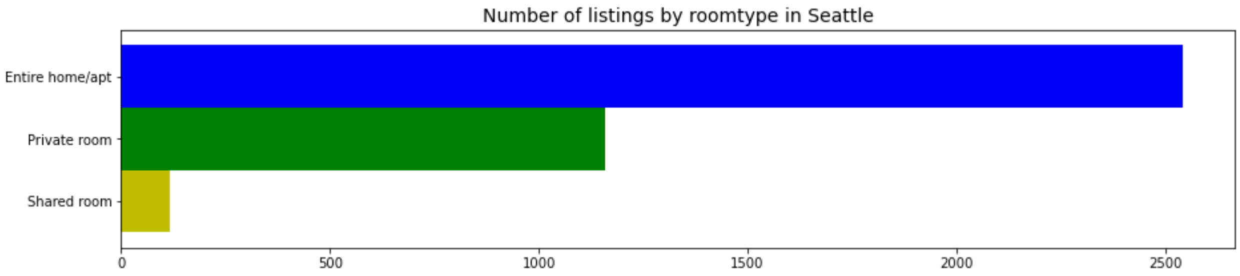
A good number of hosts respond to complaints and enquiries from customers within an hour.

Property type



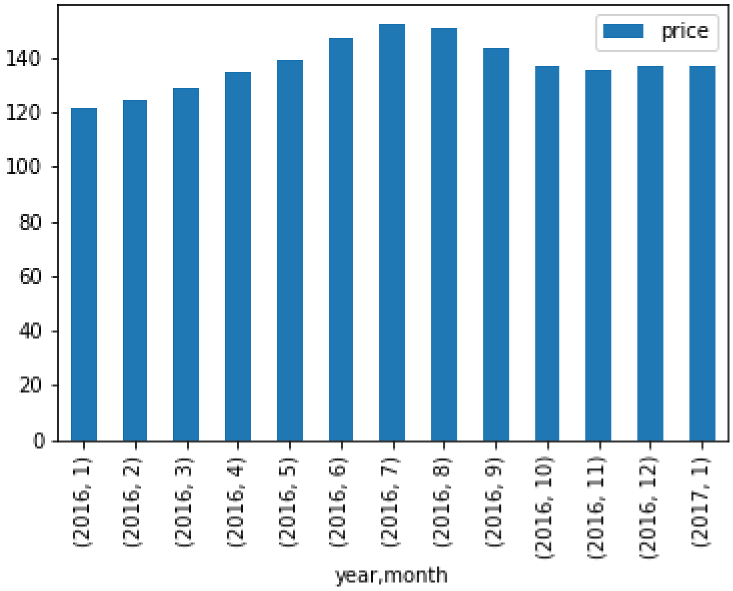
Most people prefer Airbnb that mainly comprise of houses or apartments.

Room types



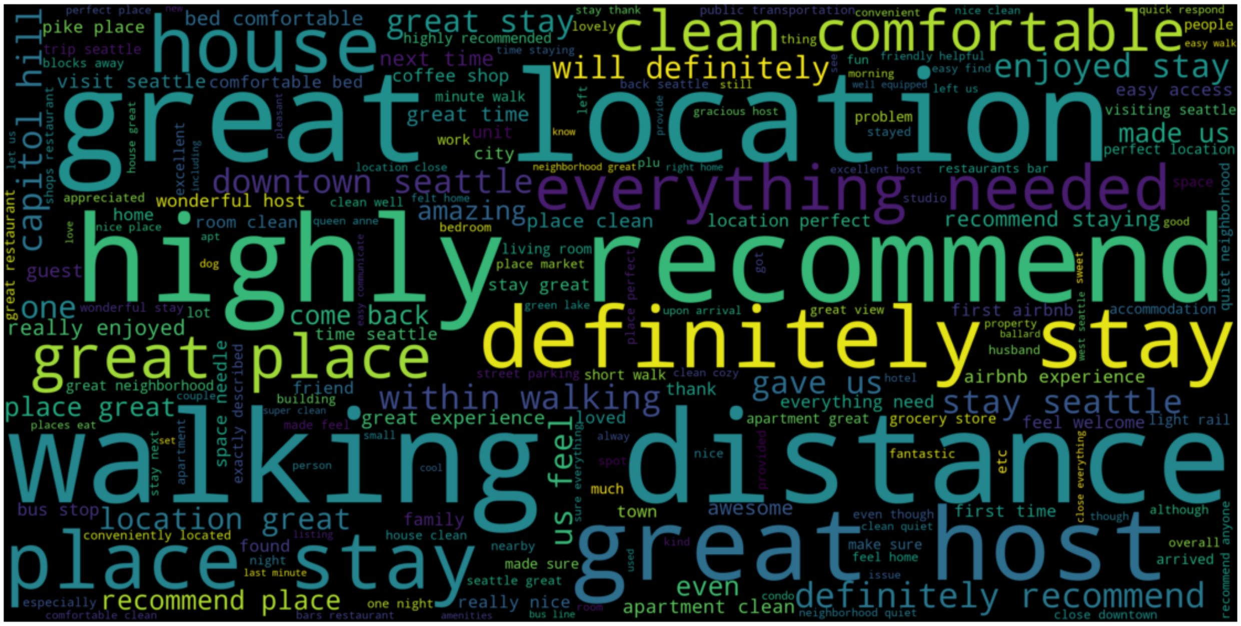
It is evident that people are concerned with their privacy and adequate space since they prefer having an entire home as an Airbnb compared to a shared or private room.

Average price per month

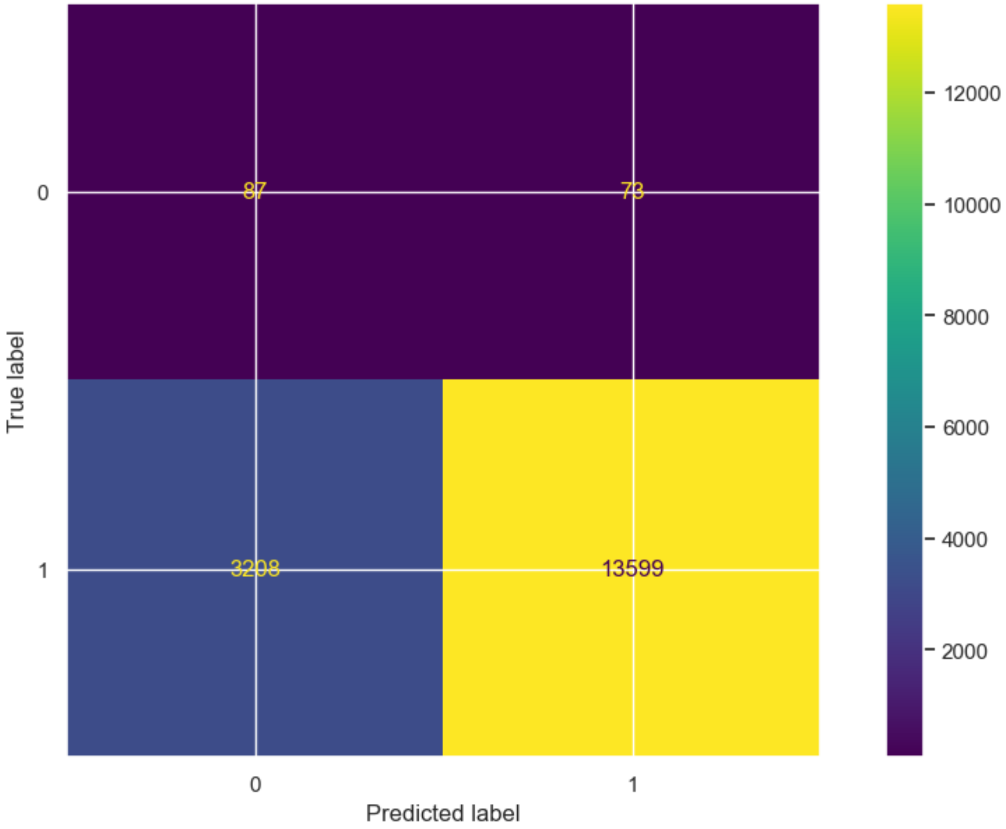


Based on the above chart, it is evident that June through August are the peak months, with July being the highest. A quick google search confirms this since these months have the best weather in Seattle with summer in full swing and low chances of rain.

Word cloud for the most commonly used three word combinations in reviews



Confusion matrix

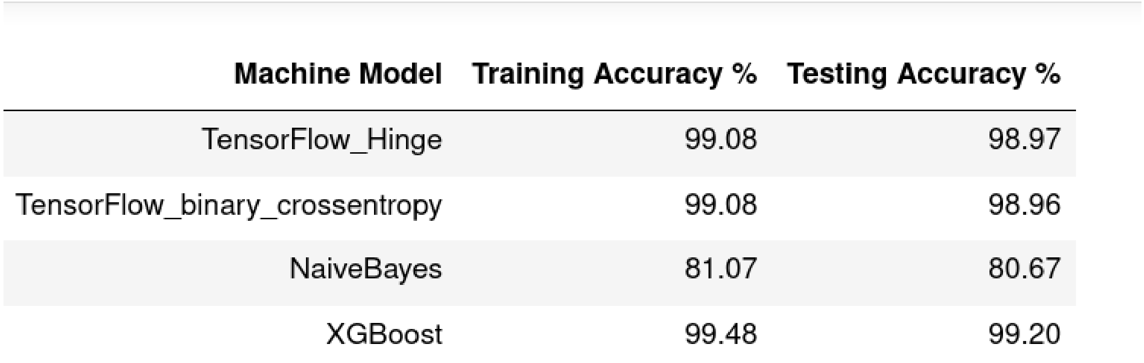


3. DATA PREPROCESSING

Since the major part of data was text, some preprocessing was required. Some of the techniques applied include:

- Cleaning reviews text by extracting only alphabets and numbers, removing stop words.
- Filling null values where columns with missing values were dropped since they were irrelevant to the analysis.
- Cleaning some columns where those with special characters like %, $ were removed and values changed to float.

The model is trained with the training features (X_train) and training labels (y_train) and is given some new data it hasn't seen before (X_test) to evaluate how well it classifies the new data. The addition of the Naive Bayes classifier provided an extra layer of analysis by providing an indication of the percentage of listings that are less expensive than the average overnight hotel room cost. We also incorporated text analysis, which assisted in tying our information together. By identifying the key terms, we were able to determine what customers are looking for and generate ideas on how to better market Airbnb and its listings hence increase customer base and generate more bookings, ultimately generating more revenue. Also, the low price of listings will continue to give Airbnbs an edge over hotels while being able to cater to specific customer needs is likely to raise listing ratings and improve overall customer satisfaction.



Machine Model	Training Accuracy %	Testing Accuracy %
TensorFlow_Hinge	99.08	98.97
TensorFlow_binary_crossentropy	99.08	98.96
NaiveBayes	81.07	80.67
XGBoost	99.48	99.20

The above are the results of models created all of which have passed the success criteria 75%. Hence they can effectively predict reviews upto the accuracy displayed hence fit for deployment.

4. CONCLUSIONS

Beyond price, there are many other factors people consider when booking accommodations. For instance, location and amenities are other practical considerations that attract customers to an Airbnb. Most importantly, online reviews that have consistently grown in importance over the years also determine the rate at which an Airbnb gets booked. The occupancy rate in Seattle tends to be higher during summer where super hosts tend to rank higher compared to regular hosts. To utilize the information gathered from reviews, an appropriate method was selected to analyze the words used within the reviews to parse any data useful for better understanding user behavior, as well as, past and future experiences. Additionally, natural language processing techniques were applied to interpret user review comments associated with the listings. This method of analysis highlighted the text of considerable importance as well as attributed a measure of sentiment which is a dynamic element that provided meaning to the text in addition to significance. The significant elements identified in our model provided justification in selection of which listing characteristics to highlight in our new campaign efforts to increase the reach of Airbnb promotions and capture a wider audience.

5. RECOMMENDATIONS

The hosts of Airbnbs should take considerable interest in better understanding the user and travel data provided through its bookings. Evaluating who their customers are, formulating profiles, and knowing what they take into consideration when booking travel accommodations is a valuable tool that can be used to better target and market to future travelers. The results of the text analysis, illustrated a lack of consistency between key terms identified from the reviews in comparison to the description of Airbnb listings. This discrepancy illustrates a need to execute quality control by adjusting the listing descriptions to incorporate key terms identified from the reviews. This will improve how Airbnb markets each listing in accordance with what customers and potential customers are looking for hence have a positive impact on the listing review scores and increase in the number of bookings.