Analysis of Seattle Airbnb Dataset

TABLE OF CONTENTS

1. INTRODUCTION

Airbnb is an online marketplace and hospitality service that connects people seeking accommodation with hosts who have space to rent.
Mission: “To create a world where anyone can belong anywhere.”
Airbnb was founded in 2008 by Brian Chesky, Joe Gebbia, and Nathan Blecharczyk in San Francisco, California. The platform offers a wide range of lodging options, including homes, apartments, rooms, and unique accommodations like boats, treehouses, and RVs. As of 2023, there are 1.5+ billion guests, 7.7+ million listings, and 5+ million hosts worldwide.

2. PROJECT PURPOSE

Seattle is known for its booming tech industry, vibrant culture, and stunning landscapes that attract millions of visitors each year. Understanding the dynamics of the Airbnb dataset in this city provides significant insights that can guide strategic choices and enhance the productivity and effectiveness of researchers, investors, and Airbnb hosts. One can learn more about the variables affecting listing pricing, booking rates, and customer preferences by examining the relationships, trends, and patterns within the dataset. This data can help determine possibilities to improve the Airbnb experience overall, as well as optimize pricing strategies, listing descriptions, and high-demand periods. The ultimate objective is to use data-driven insights into the dynamic and swiftly evolving Airbnb market to boost revenue, customer satisfaction, and competitiveness.

3. DATA DESCRIPTION

We have considered 2016 Airbnb Seattle data from Kaggle. This dataset has three CSV files namely listings, reviews, and calendar. The 3,818 entries in the listings file provide a comprehensive overview of the accommodations that are available by detailing the properties and room types. The 84849 reviews help us understand the guest's views about the listing and host offerings. Additionally, the calendar entries, reflect daily availabilities and allow us to explore trends in listing availability and pricing strategies. By utilizing these data sources, we will be able to provide a comprehensive overview of the Airbnb market, notifying both hosts and guests in the future.

3.1 LISTING DATASET

- id: Unique identification code for the listing.
- description: Detailed description of the listing/property with the amenities.
- neighborhood_overview: Brief explanation of the neighborhood attractions.
- summary: Listing summarized by the host.
- latitude: Geographical location of the listing.
- longitude: Geographical location of the listing.

- property_type: Listing type like apartment, Boat, Cabin, RV, House, etc.
- room_type: Type of room in the listing that is an entire home/apartment, private room, or shared room.
- bedrooms: Number of bedrooms in the listing.
- amenities: Amenities of the listing.
- price: Price of the listing per night.
- review_scores_rating: Overall rating given by the guest.
- review_scores_accuracy: Rating for the accuracy of the listing.
- review_scores_cleanliness: Rating for the cleanliness of the listing.
- review_scores_checkin: Rating for the check-in convenience.
- review_scores_communication: Rating for the communication of the host.
- review_scores_location: Rating for the listing location.
- review_scores_value: Rating for value for money.
- number_of_reviews: Total number of reviews for the listing.
- zipcode: Numeric code that is used for grouping mailing addresses.

3.2 REVIEW DATASET

- listing_id: Unique identification code for the listing.
- date: Date of the review.
- comments: Review by the customer.

3.3 CALANDER DATASET

- listing_id: Unique identification code for the listing.
- date: Date in 2016.
- available: Listing availability on a specific date.
- price: Price in the specific date.

3.4 DASHBOARD

The Tableau dashboard in Figure 1 is a complete analysis of Airbnb 2016 listings data, geared to provide insights into numerous parameters relevant to property management and rental pricing in particular locations. It is divided into four main visualizations:

- Top Regions Based on Ratings
- Top Regions Based on Room type
- Average Prices Based on Property Type
- Average Prices Based on Zip code

Overall, the dashboard helps property managers and investors make strategic decisions by displaying major trends.

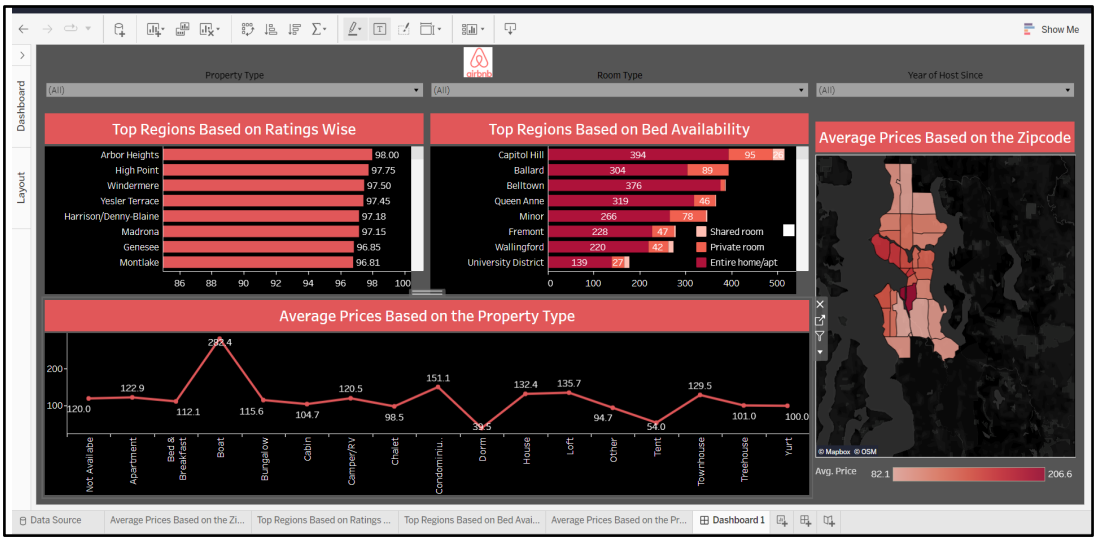

Figure 1. Tableau Dashboard

Top Regions Based on Ratings: This section ranks regions based on average user ratings, highlighting Arbor Heights and High Point as top performers with scores around 98. This graphic can help identify regions with the highest visitor satisfaction, which is critical for focused marketing and improvement activities.

Top Regions Based on Room type: It breaks down room type by region, with Capitol Hill and Ballard topping the way, indicating that these regions have the greatest number of available room types, which may correlate with larger visitor capacity or popular tourist destinations.

Average Prices by Property Type: This graph depicts the average rental price for various types of properties, such as apartments, houses, and dorms, with noteworthy peaks in guest houses and lofts, indicating that these property types may be more profitable or in-demand.

Average prices Based on the Zip code: This section uses a heat map to illustrate typical rental costs across different zip codes, providing a geographical view on pricing trends that could be useful for adjusting pricing strategies based on location-specific demand.

4. METHODOLOGY

The data analysis methodology is broken down into the following steps:

- Preprocessing - This initial step involves collecting and preparing the data for analysis. This includes cleaning the data by removing punctuations, formatting it consistently,

handling any missing values, normalizing data, and encoding categorical variables to make it suitable for further steps.

- Exploratory Data Analysis (EDA) & Word Cloud Creation - Exploratory data analysis (EDA) is a statistical approach used to get a foundational understanding of the data and identify patterns or trends. A word cloud is a visual representation of the frequency of words in a text dataset. It is useful for spotting common terms and trends. Here, EDA help identify frequently used words in reviews, which could be positive words like “clean” or “comfortable” or negative words like “noisy” or “disappointed”. A word cloud could then visually depict these terms, with larger words representing those used more often.

- VADER Sentiment Analysis - VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool used to determine the sentiment of text data. It classifies text into positive, negative, or neutral categories. Sentiment analysis help Airbnb understand overall guest sentiment towards their listings and the hosting experience. This is valuable for identifying areas for improvement and informing marketing strategies.

- Topic Modeling - This is a machine learning technique used to automatically identify hidden thematic groups within a collection of documents. Text data is often categorized into topics based on the words that appear together most frequently. For Airbnb, topic modeling is used to uncover common patterns.

- Classification - This is the final stage where a machine learning model is trained on the data. The trained model is used to classify new data points. We have implemented these models for predicting price of a new listing and review of a listing.

Methodology for Price Prediction Model:

- Data Collection - Gathered data on past listings that include the final sale price. The data features like the number of bedrooms, individual amenities offered, and property type, etc.

- Data Preprocessing - For Listings dataset, preprocessing involved filling missing values (often with 0s), encoding categorical features (like label encoding for "instant_bookable"), handling textual inconsistencies (e.g., replacing empty property types with "Not Mentioned"), transforming boolean columns (potentially to binary representations), cleaning price data (removing symbols and converting to floats), and feature scaling for numerical features. Feature engineering by transforming the original

data (a single string column with amenity lists) into a new format (multiple binary columns) using hot-one encoding.

- Model Selection and Training - Choosing an appropriate machine learning model for price prediction among these models.
  1. Linear Regression: Establishes a linear relationship between features and price.
  2. Ridge Regression: Similar to linear regression, but addresses overfitting by penalizing models with high coefficient values.
  3. Lasso Regression: Another variation that encourages sparsity by shrinking some coefficients to zero.
  4. Random Forest Regression: Creates an ensemble of decision trees, improving prediction accuracy through averaging.
  5. XGBoost: A powerful tree boosting algorithm known for handling complex relationships.

  Trained the model(s) on the dataset, allowing the model to learn the underlying relationships between features and prices.

- Model Evaluation - Evaluating the model's performance. This helps assess how well the model generalizes to data. Common metrics for evaluation included:
  Mean Squared Error (MSE): Measures the average squared difference between predicted and actual prices (lower MSE indicates better performance).
  Root Mean Squared Error (RMSE): Square root of MSE, representing the standard deviation of the prediction errors.
  Compared the performance of the models and chose the one with the lowest MSE/RMSE.

- Model Deployment and Prediction - Deployed the Ridge Regression model along with Tree interpreter to make predictions on listings. Provide the model with the features of a listing, and it will predict the corresponding sale price.

Methodology for Predicting Review Score Rating:

- Data Acquisition and Preprocessing - Gathered the Reviews dataset and Listings dataset. For Listings dataset, preprocessing involved filling missing values (often with 0s), encoding categorical features (like label encoding for "instant_bookable"), handling textual inconsistencies (e.g., replacing empty property types with "Not Mentioned"), transforming boolean columns (potentially to binary representations), cleaning price data (removing symbols and converting to floats),and feature scaling for numerical features. For reviews dataset, preprocessing involved involves text cleaning steps like

removing non-English reviews, converting text to lowercase, removing punctuation and potentially stop words, and optionally lemmatization.

- Sentiment Analysis - Applying Vader sentiment analysis to the review text in the reviews dataset. Vader is a lexicon-based sentiment analysis tool that assigns a sentiment score (positive, neutral, or negative) and sentiment type (compound score) to each review.

- Data Merging - Merge the reviews dataset (containing listing ID, sentiment score, and sentiment type) with the listings dataset (containing listing ID, price, and number of reviews) based on the common listing ID. This creates a combined dataset with features relevant to review score prediction.

- Model Selection and Training - Chose a Linear regression model to predict review_score_rating of any listing. Train the chosen model on the merged dataset, allowing it to learn the relationships between features (price, sentiment score, sentiment type, number of reviews) and the target variable (review_score_rating).

- Model Evaluation - Evaluated the model's performance on the testing set. This assesses how well the model generalizes to unseen data. Common metrics for regression include:
  Mean Squared Error (MSE): Measures the average squared difference between predicted and actual review scores (lower MSE indicates better performance).
  Root Mean Squared Error (RMSE): Square root of MSE, representing the standard deviation of the prediction errors.

- Model Deployment and Prediction - Provide the model with the features of a listing (price, sentiment analysis results, number of reviews), and it will predict the corresponding review score rating.

Methodology for Predicting Seasonal Price and Availability of Listings:

- Data Preparation: For Listings dataset, preprocessing involved filling missing values (often with 0s), encoding categorical features (like label encoding for "instant_bookable"), handling textual inconsistencies (e.g., replacing empty property types with "Not Mentioned"), transforming boolean columns (potentially to binary representations), cleaning price data (removing symbols and converting to floats),and feature scaling for numerical features. For Calendars dataset, preprocessing involved filling missing values (often with 0s), cleaning price data (removing symbols and converting to floats. Determined the seasonality period (monthly) based on the date.

- Model Selection and Training - Selected a suitable seasonal forecasting model like SARIMA to incorporate seasonality into their forecasting process. SARIMA (Seasonal Autoregressive Integrated Moving Average) is a statistical model used for forecasting time series data with seasonality. Fit the chosen model to your time series data using the identified internal and external parameters.

- Forecasting - Used it to predict future values for the time series. Specified the desired forecast horizon (e.g., next 12 months). Used the SARIMA model to predict seasonal prices for 2017 based on month, average price of the listings and number of listings per month of 2016. Then further used SARIMA model to predict the number of available listings for each month in 2017 based on month, average price of the listings and number of listings per month of 2016. Furthermore, used SARIMA model to predict availability of a particular listing for each month in 2017 based on the number of available days of each month, average price and review score rating.

- Interpretation and Visualization - Displayed the forecasted values and plotted the historical data and the generated forecasts to visualize the predicted trend and seasonality.

5. RESEARCH QUESTIONS

1. **How do Airbnb listing prices vary among Seattle neighborhoods, and what variables impact these differences?**

Initially, we started our analysis by plotting a bar chart that depicts a comparative analysis of average listing prices across various neighborhoods in Seattle. The x-axis represents the mean price and the y-axis represents the neighborhood.

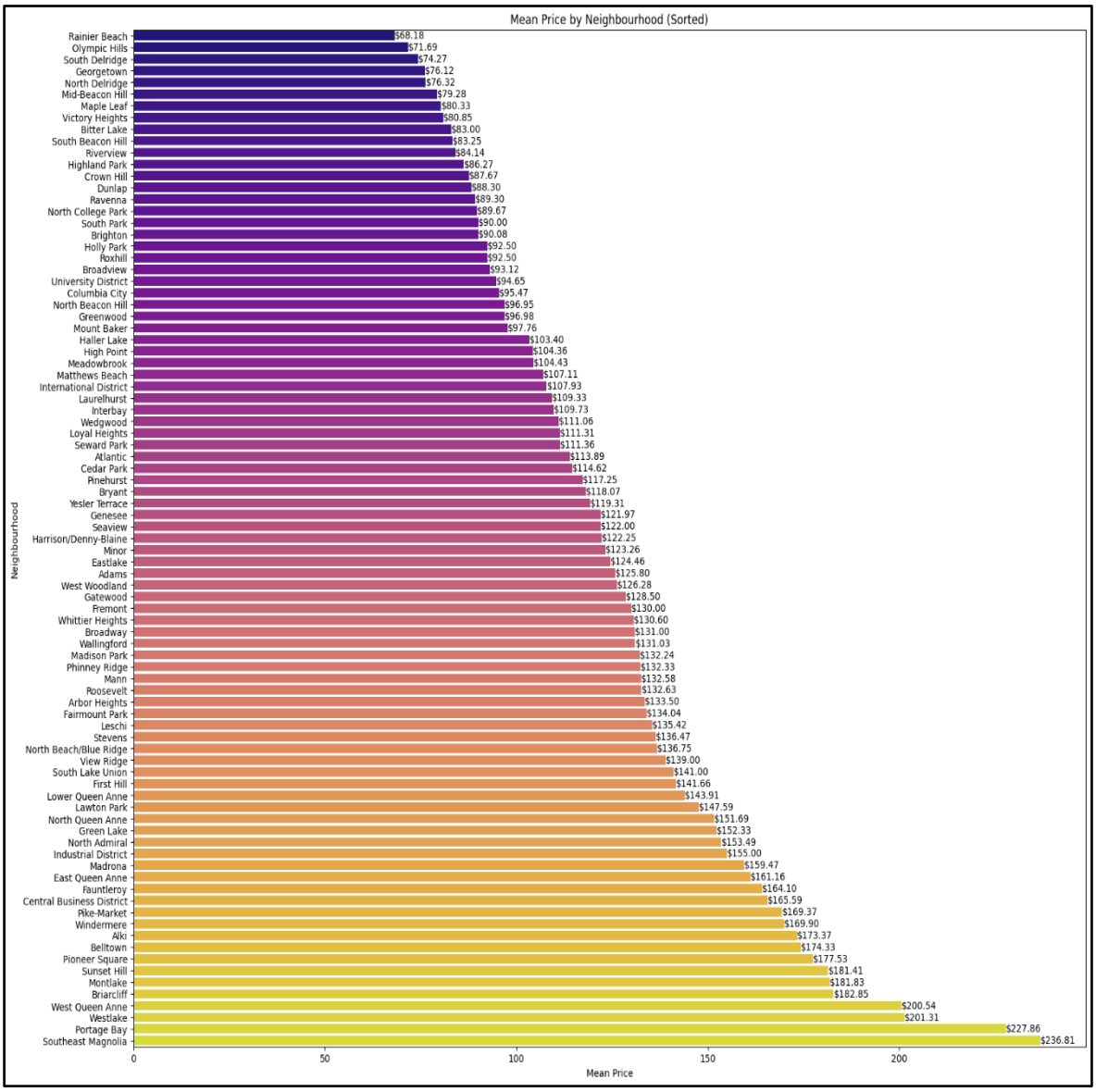

Figure 2. Bar plot of Mean price by Neighborhood (sorted)

From Figure 2, we can see a significant discrepancy in listing affordability across the Seattle neighborhood. The data ranges from an average price with a high of $236.81 in Southeast Magnolia and a low of $68.18 in Rainier Beach.

Further in our analysis, we look into the factors contributing to the price variations across the neighborhood. Here we have analyzed through the means of a heatmap. In this map, the mean price is overlapped over the density of the listings in the neighborhood considering the price percentile over a region. The heatmap employs a color gradient scheme to represent price percentiles, with lighter colors indicating lower percentiles (more affordable) and darker colors signifying higher percentiles (more expensive). From

the zoomed area in the heatmap Figure 3, we can see that Belltown has the most evident cluster because it was close to downtown and had more convenient amenities nearby like parks, shops, restaurants, etc. making it a popular area for guests to stay.

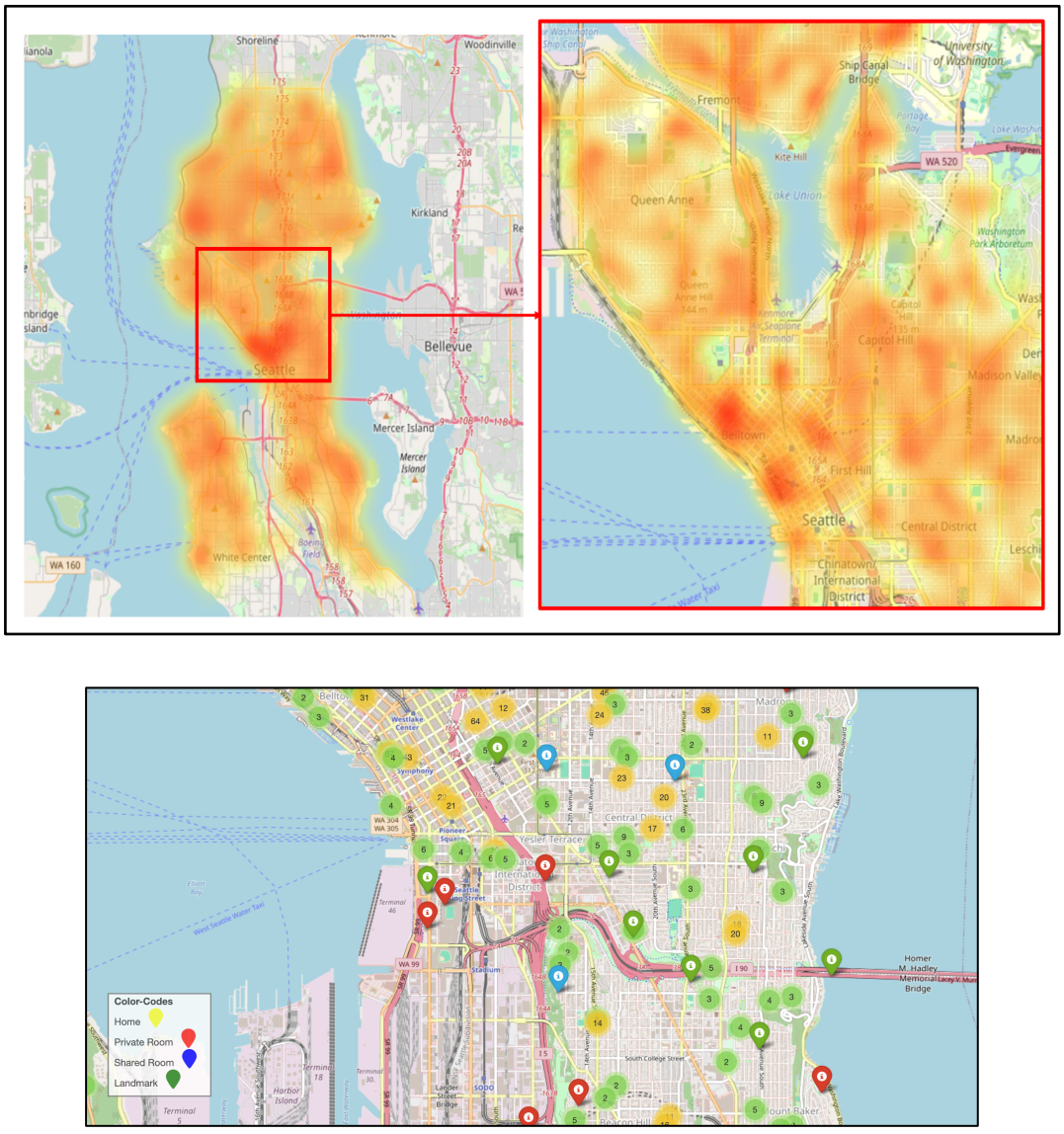

Figure 3. Heatmap of the price percentile

For an in-depth analysis, we have considered 3 individual factors that affect the mean price of a listing. The first factor was property type, in figure 4 we observe how the mean price varies through different kinds of property types.

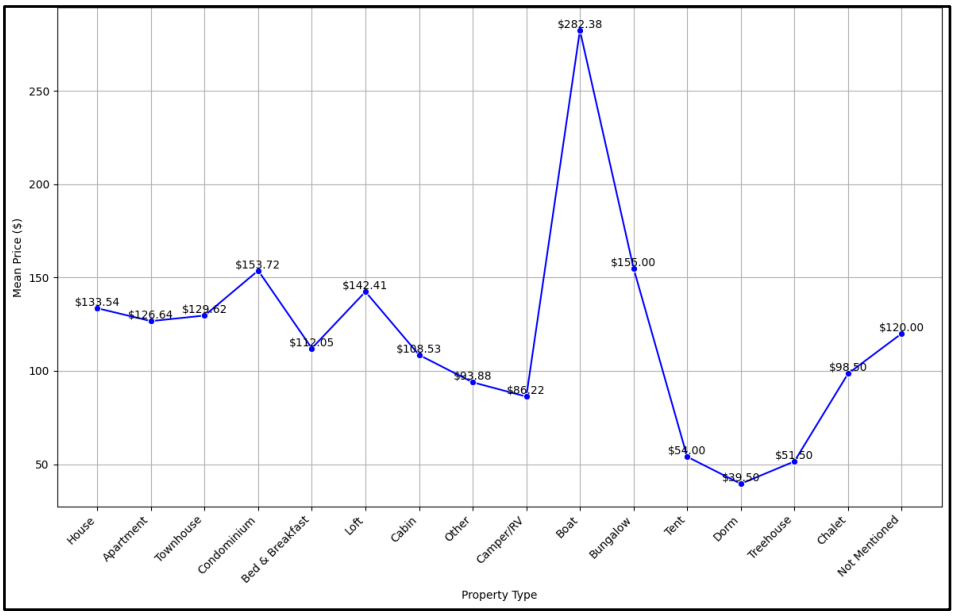

Figure 4. Line plot of Mean price by property type

The second factor was the number of bedrooms, in figure 5 we observe how the mean price varies when the number of bedrooms varies across all the property types.

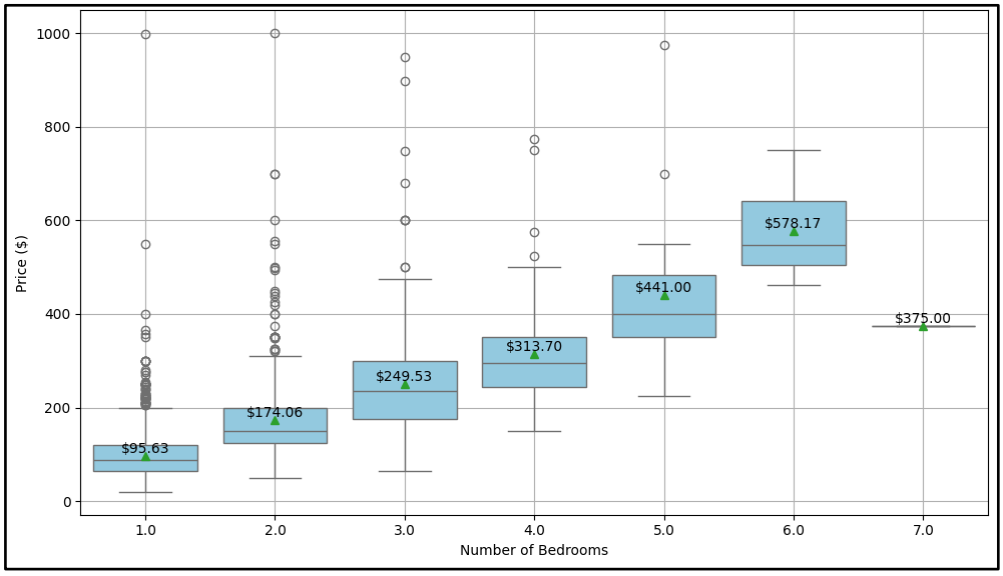

Figure 5. Box plot of Mean price by number of bedrooms

We can see that as the number of bedrooms increases, the mean price increases that is it is directly proportional. This is up to 6 bedrooms. We have only one listing with 7 bedrooms in our dataset which is the reason leading to a drop in the price for the property with 7 bedrooms. Considering our plot we can notice outliers; this is because we have a wide

range of listings with expensive price ranges. These outliers cannot be removed and play an important part throughout our analysis.

The third factor is the room type, in figure 6 we can see the percentage distribution of the three types of rooms. Most of the listings are entire home or apartments with 63% with the highest mean of $165.

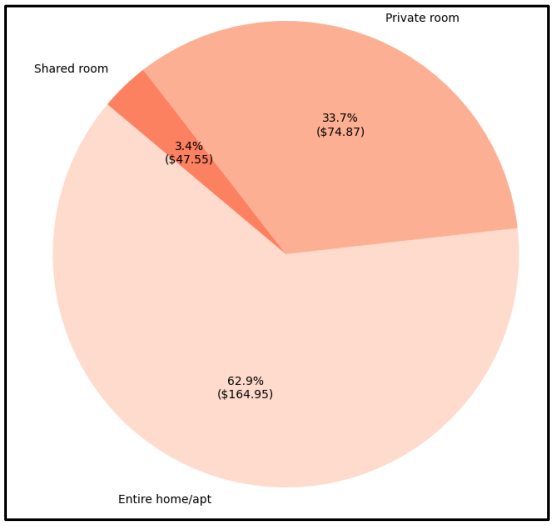

Figure 6. Pie chart of Distribution of listings by room type

After looking into the three factors individually further we combined all these factors to observe variations. From Figure 7 we can see how there are variations in mean price. We can see that the boat rentals have the highest mean price of $447. A shared tent or loft has the lowest mean price of $25.

Figure 8 shows the variations of mean price with respect to the property type and number of bedrooms. Here we can see that the Boat property type has the highest mean price of $775. We also notice that as the number of bedrooms increases the mean price increases across all the property types.

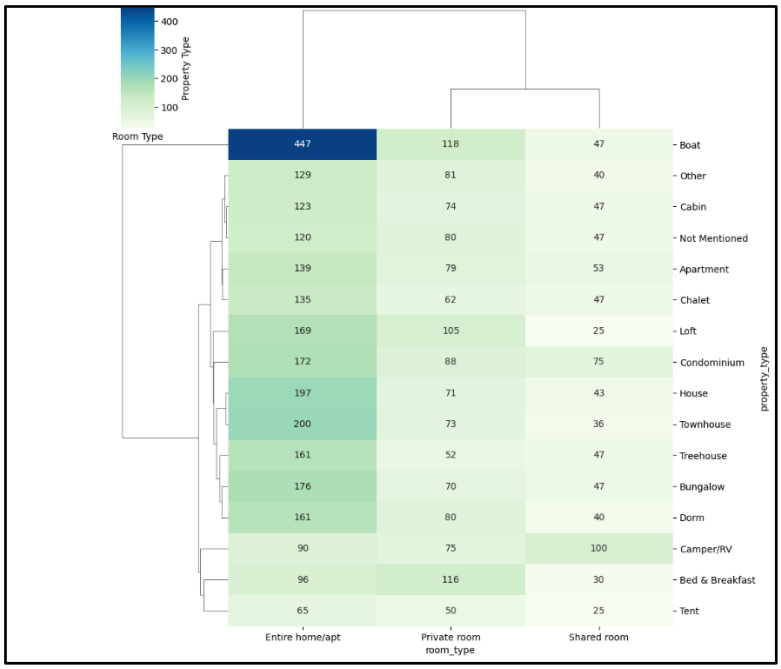

Figure 7. Heatmap of mean price by property type and room type

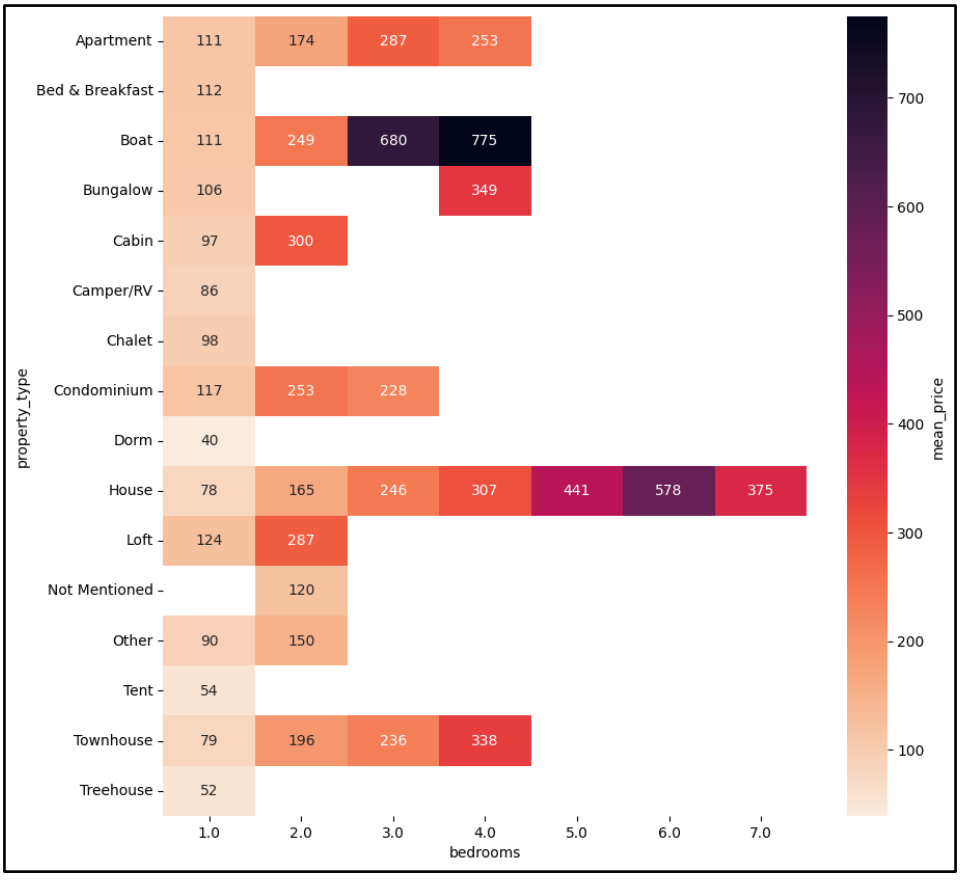

Figure 8. Heatmap of mean price by property type and number of bedrooms

Looking into our analysis of mean price with the help of heatmap of property type and neighborhood in figure 9. We have considered Camper or RV property type for our analysis since this type of property does not have multiple number of bedrooms, which is one of the key factors that affect the price of the listing. So, from the heatmap we can see variations in mean price for this particular property type across all the neighborhood starting with lowest mean price of $50 in Maple Leaf gradually increasing through $52 in West Queen Anne, $75 in Loyal Heights, $82 in Fermont, $99 in Mount Baker, $100 in South Lake Union and highest mean price of $145 in Lower Queen Anne.

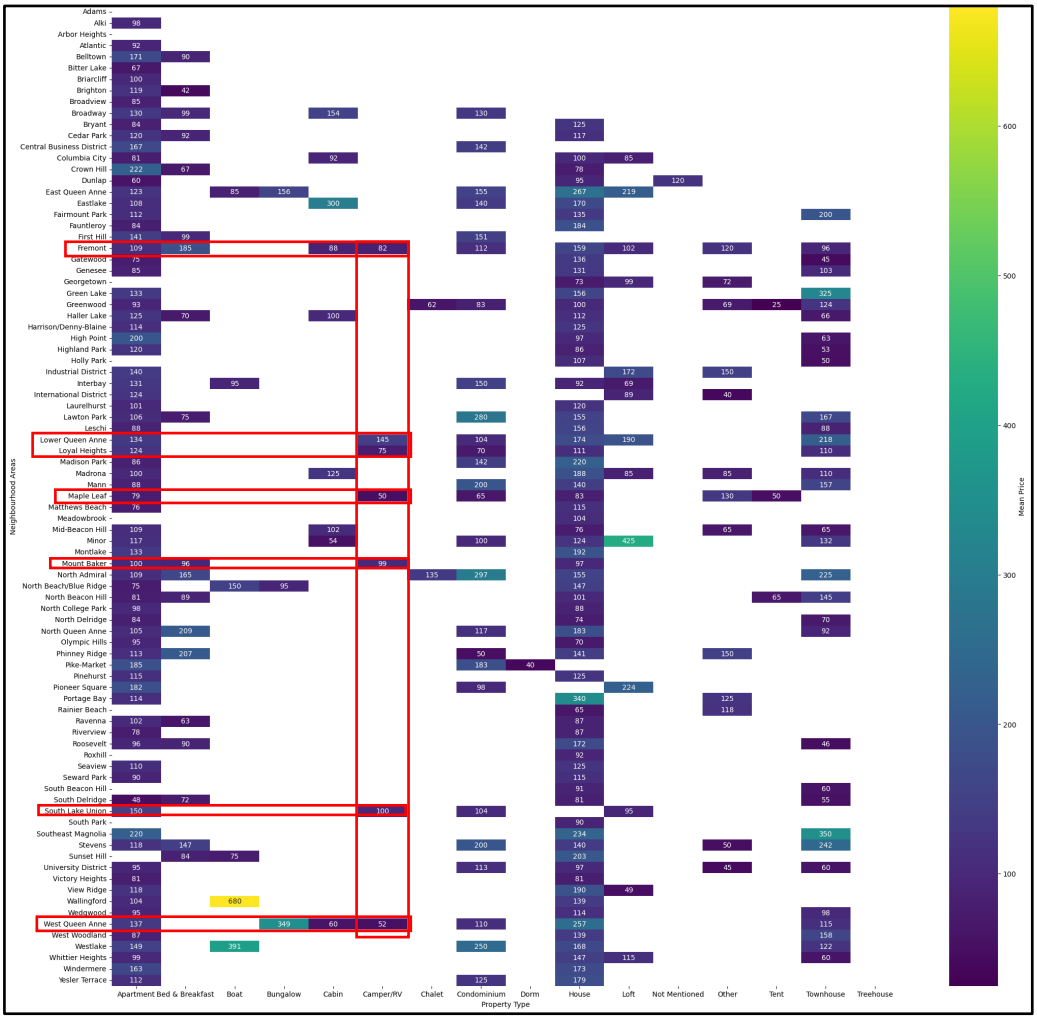

Figure 9. Heatmap of mean price by property type and neighborhood

Therefore we can conclude that all the above factor that we have analysed play a crucial role in affecting the price of a listing and these vary throughout different neighborhood based on the aminities avilable in those regions.

2. What are the essential qualities and characteristics of Seattle's most highly rated Airbnb rentals, and how do they affect customer satisfaction?

To answer this question, we have considered the listings with listing price above $500 and review rating score above 75 out of 100. For these set of listings, we have created two set of word clouds.

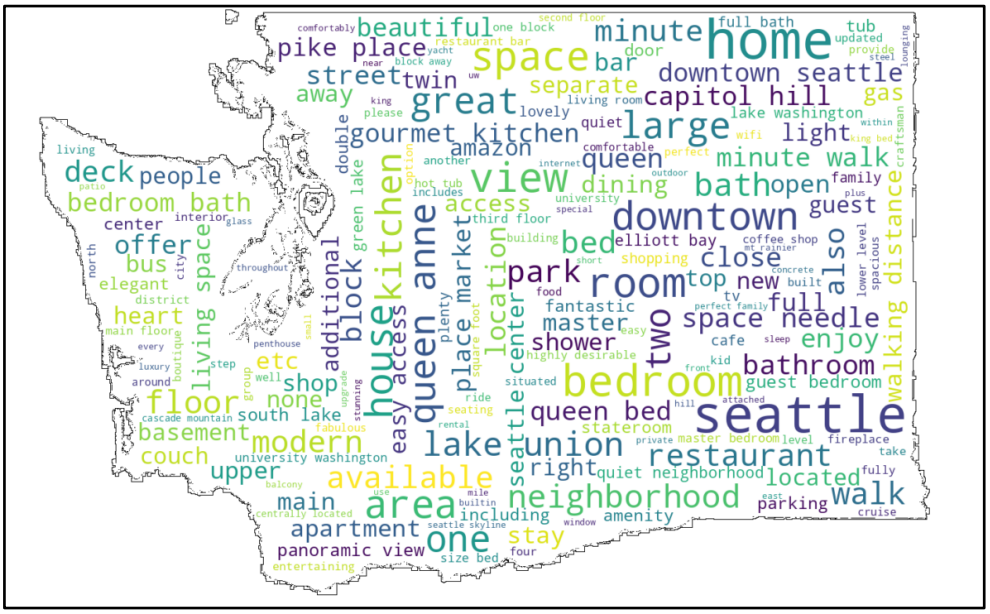

Fig 10. Word cloud based on Host's description of the listings

Fig 11. Word cloud based on Guest's reviews of the listings.

A word cloud is a visual representation of text data, where the most frequently occurring words are displayed in larger font sizes and are usually emphasized by color or other graphical elements. It is a way to visualize the most common words or themes in a body of text. The first word cloud as seen in figure 10 is created by using the host's description about a listing. The second word cloud as seen in figure 11 is based on the guest's reviews of those listings.

By visualizing these word clouds we can see some common words like downtown, comfortable, parking, neighborhood, walking distance, restaurant, view, home, space, easy access, etc. these help us understand the qualities that contribute to the customer satisfaction. We can conclude that higher ratings are due to the conveniences and experiences of guests in these listings.

3. Can we uncover essential elements and patterns that lead to a great visitor experience, as seen by reviews and ratings, to help host improve their offerings?

Considering our dataset, we have seven parameters that strongly or weakly impact the overall review rating score of the listings. These parameters are namely

- Review score rating
- Review score accuracy
- Review score cleanliness
- Review score check-in
- Review score communication
- Review score location
- Review score value

	review_scores_rating	review_scores_accuracy	review_scores_cleanliness	review_scores_checkin	review_scores_communication	review_scores_location	review_scores_value
count	2874.000000	2874.000000	2874.000000	2874.000000	2874.000000	2874.000000	2874.000000
mean	93.996868	9.543145	9.471468	9.689979	9.736952	9.517049	9.368824
std	9.934330	1.175387	1.180854	1.144726	1.023758	1.101160	1.173890
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	93.000000	9.000000	9.000000	10.000000	10.000000	9.000000	9.000000
50%	96.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
75%	99.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
max	100.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000

Figure 12. Analytics of all the seven review parameters.

We can consider these seven parameters to check how they affect the number of reviews received for a particular listing. In figure 12 we can see detailed analytics of all these parameters by observing their mean, standard deviation, minimum rating, and maximum ratings.

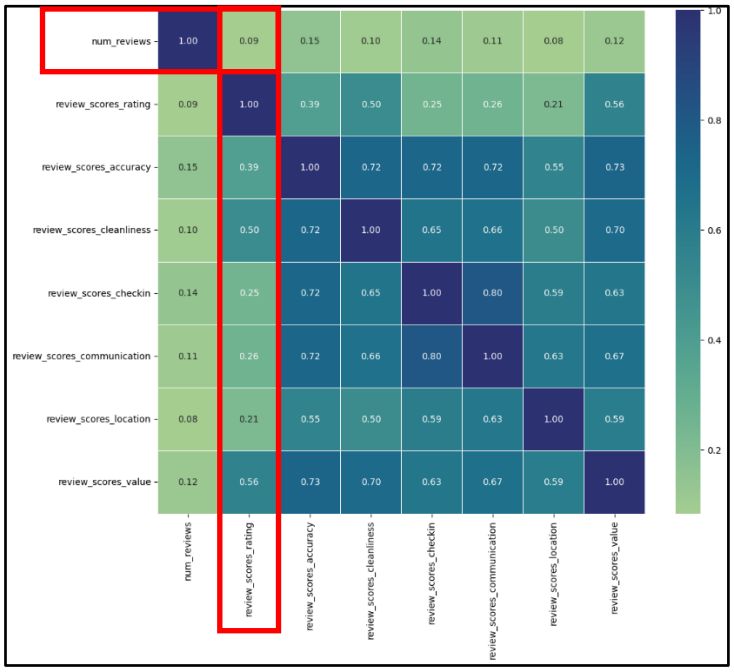

Figure 13. Correlation matrix of the review parameters

Further depicting the analytics of the review parameters using a correlation matrix from figure 13 we can see that there is a weak positive correlation of 0.09 between the number of reviews and review score rating. This means that listings with more reviews tend to have slightly higher average ratings but the effect is weak.

In our dataset we have majority of the listings which are priced in the range $100 - $120. For more detailed analysis we considered the listings in this price range and compared the highest and the lowest reviewed rating listing against all the seven parameters.

From the correlation matrix in figure 15 we can see a stronger positive correlation of 0.66 between the review score for cleanliness and the value of the listings. This means that reviews who find a listing to be clean also tend to find that it offers good value for the price.

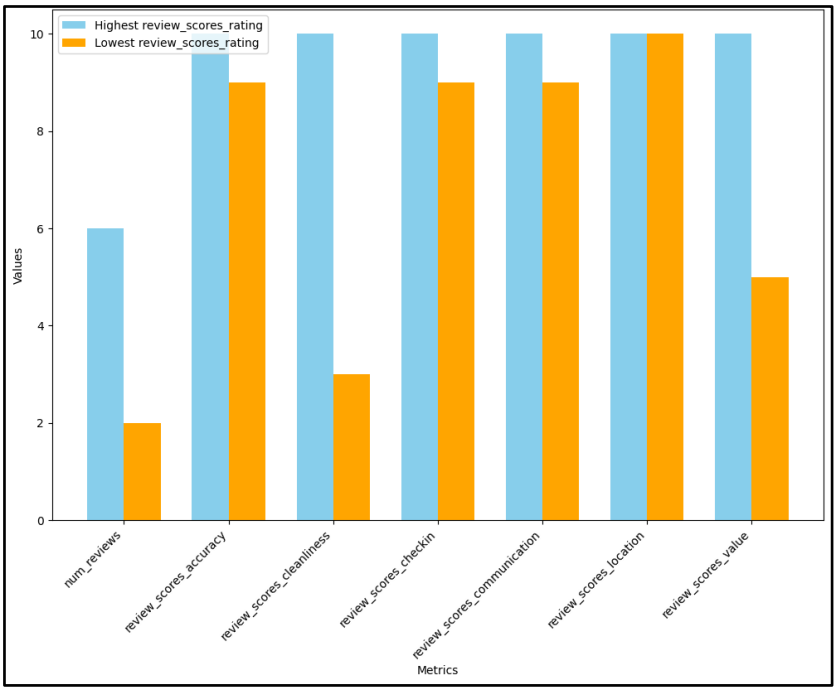

Figure 14. Comparison of matrix between highest and lowest review score rating listings

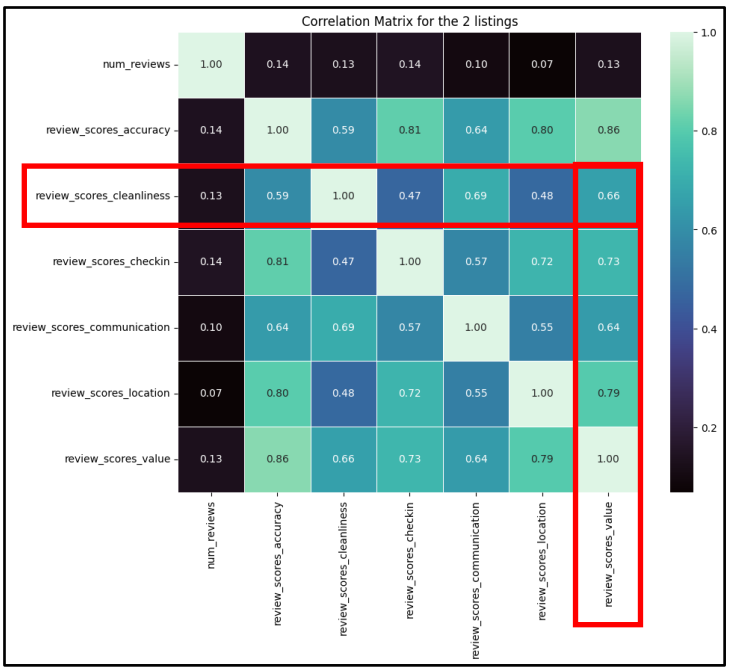

Figure 15. Correlation matrix of highest and lowest review score listings

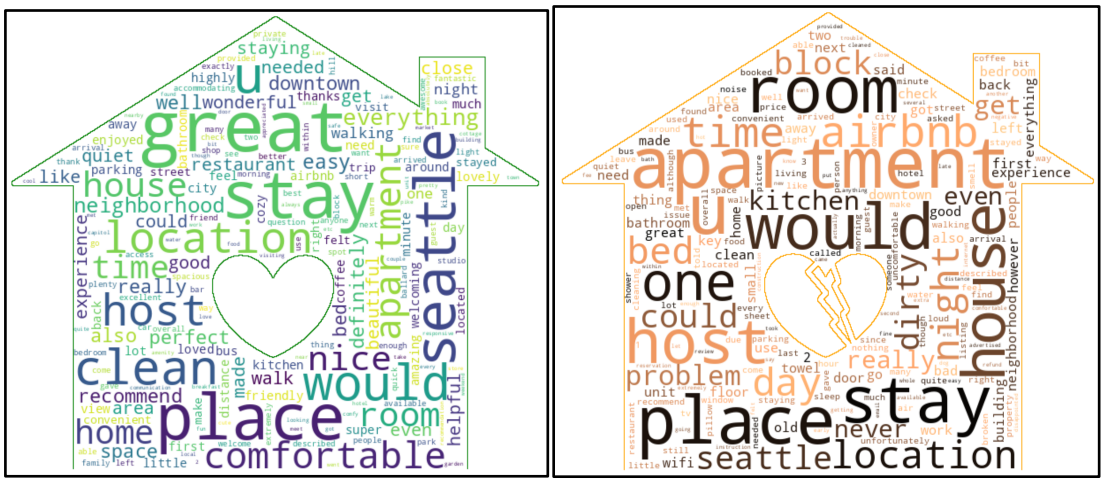

Figure 16. Highest rated listing and lowest rated listing word clouds

After analyzing the correlation matrices, we have visualized the word clouds of the highest review rated listings and the lowest review rated listings. We can see words like clean, great, comfortable, wonderful, easy, quiet, convenient, etc. from the highest reviewed listing and can see contradicting words like dirty, Wi-Fi, noise, small, uncomfortable, smell, reservation problem, etc. from lowest reviewed listing.

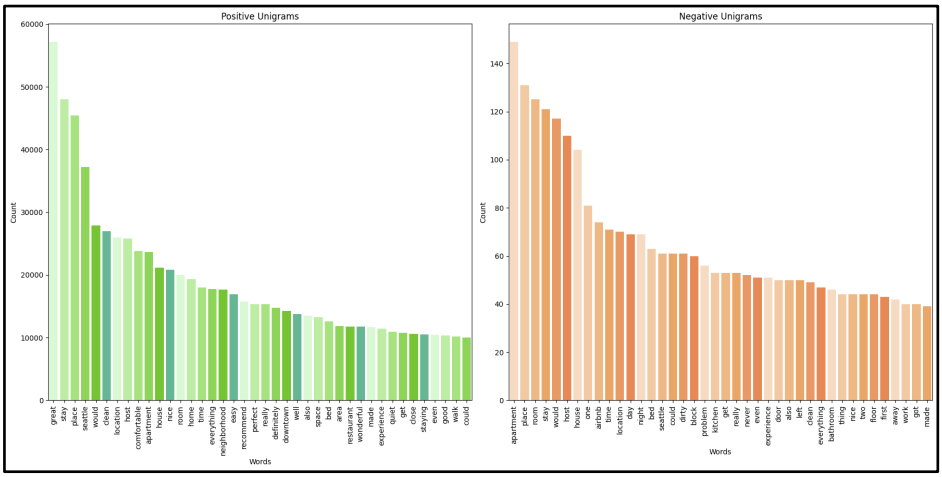

Figure 17. Guest Review Unigrams

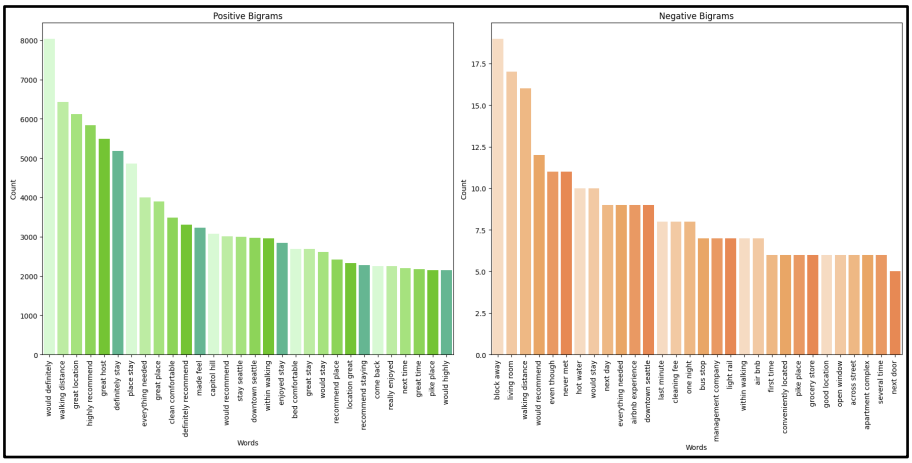

Figure 18. Guest review Bigrams

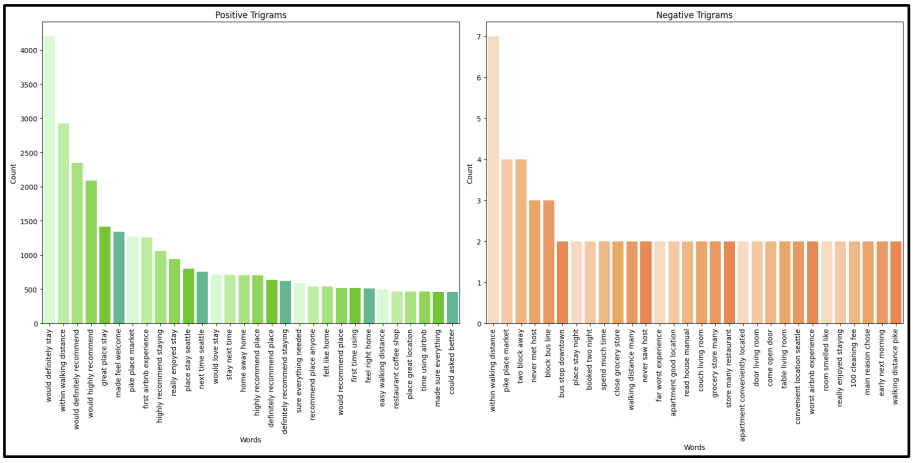

Figure 19. Guest review Trigrams

Figure 17, 18 and 19. Indicate the unigrams, bigrams and trigrams based on the highest and lowest reviewed listings based on all the key occurrences of positive and negative words provided by customer reviews which will help the host to understand the customers and guest perspectives. From all the above analysis and visualizations, we can conclude that the host has to improve on the keywords which occur in the lowest review rated listing to increase customer satisfaction to lead to great visitor experience.

4. What is the seasonal pattern of booking rates in Seattle, and are there any peak periods of demand or limited availability that hosts should be aware of?

Seattle, like many cities, experiences fluctuations in booking rates throughout the year as seen in figure 20. Summer from June to August is typically the peak season for tourism in Seattle. The weather is generally mild and sunny during these months, making it ideal for

outdoor activities. Consequently, booking rates tend to be highest during this time, especially in popular neighborhoods.

Spring from March to May sees an increase in bookings as the weather starts to improve, and the city begins to bloom with cherry blossoms and other flowers. The demand may not be as high as in summer.

Conversely, the winter months, particularly from November to February, typically see lower booking rates and more availability. However, this period may still attract visitors due to holiday events and activities, as well as business travel.

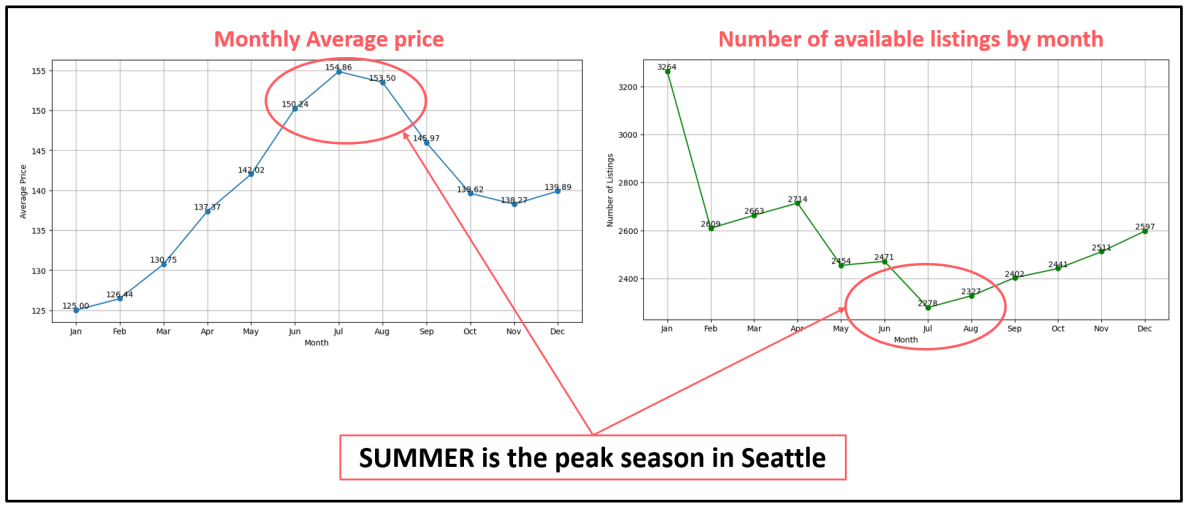

Figure 20. Seasonal pattern of listing price and number of available listings in Seattle

Hosts should be aware of these seasonal patterns and peak periods of demand when setting their prices and managing their availability. Adjusting prices accordingly and planning for potential spikes in demand during peak periods can help hosts optimize their bookings and maximize their revenue.

From figure 21 we can see the calculated correlation coefficient between the average price and number of available listings is -0.7951. The correlation matrix clearly indicates that the average price and the number of available listings is inversely proportional to each other throughout the year.

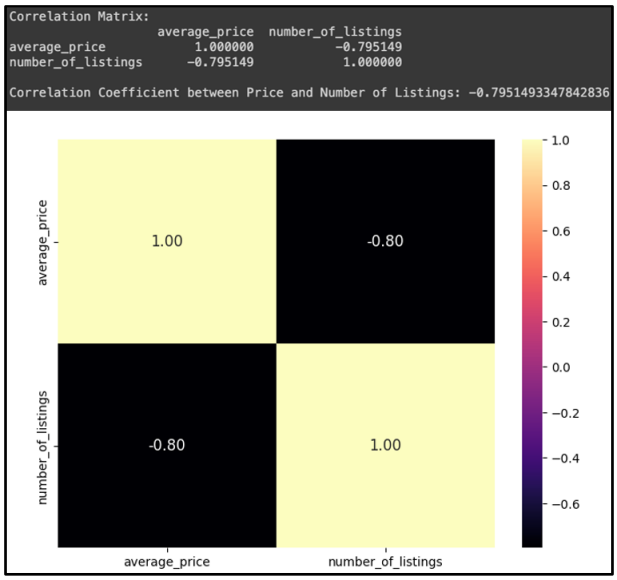

Figure 21. Correlation matrix and coefficient between average price and available listings

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a time series forecasting model used to predict future values based on past observations. It is an extension of the ARIMA model with the ability to handle seasonality.

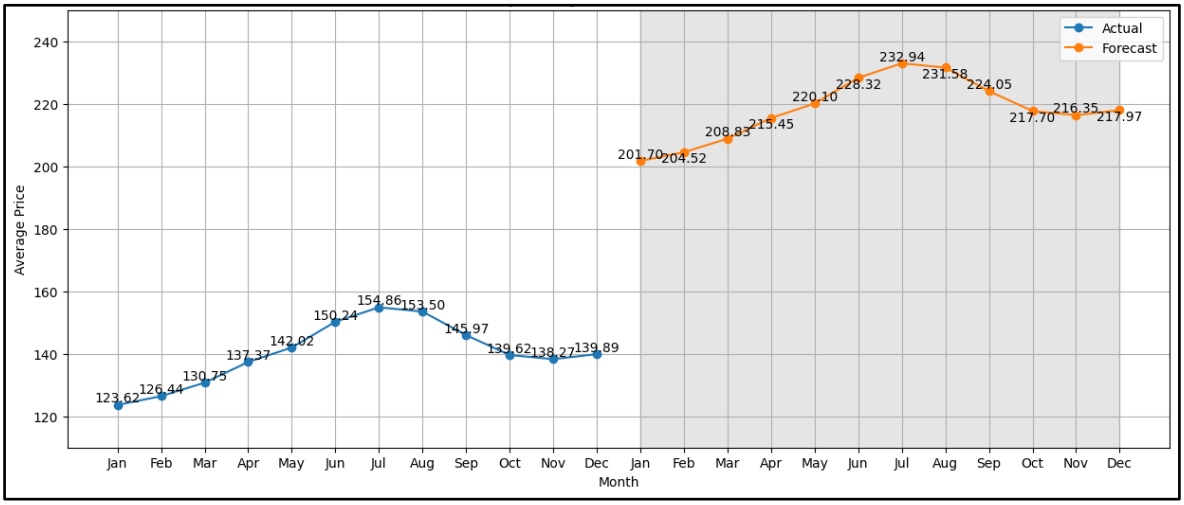

Figure 22. Monthly average price forecast in Seattle

Using SARIMA for price prediction involves fitting a model to our price data and then using that model to forecast future prices. Here by implementing this, we can see in figure 22 that there is a price increase predicted for the next year indicating that there will be more demand for the number of available listings.

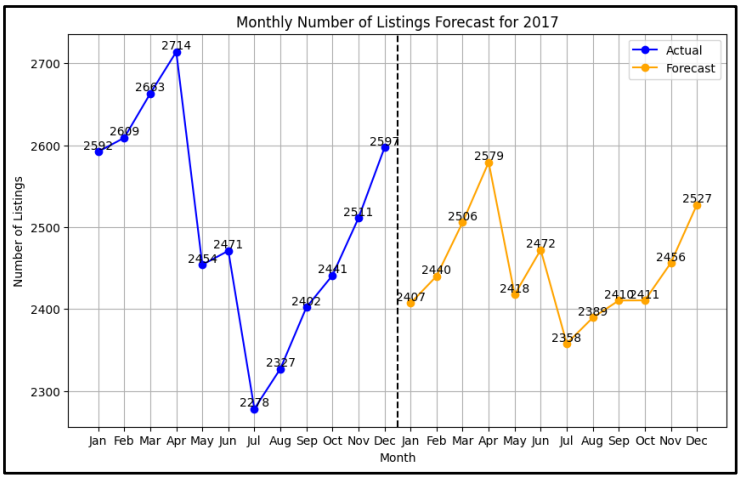

Figure 23. Predicting number of available listings for the next year

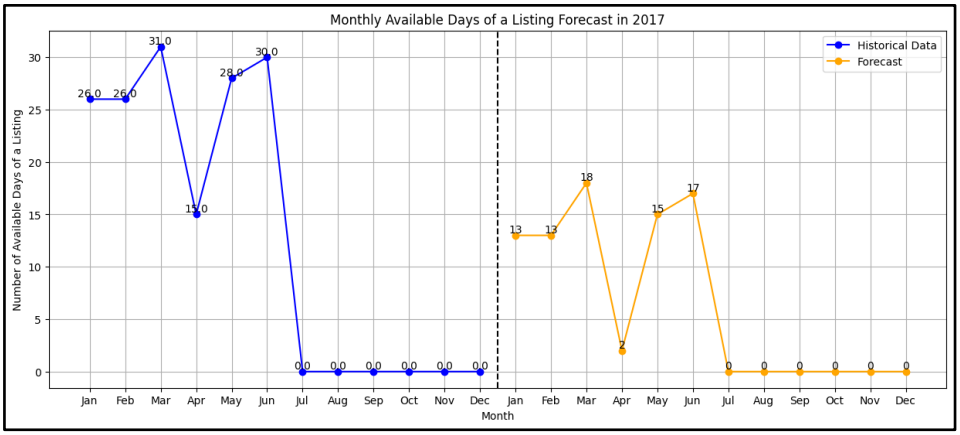

Figure 24. Predicting number of available listings of a particular listing id for the next year

Based on the inputs given by the professor, we have developed a SARIMA model to predict the number of available listings throughout the year for the next year assuming the number of availability of listings for the next year. Figure 23 depicts a generic version of the analysis. For deeper analysis we have considered a listing id '6315435' to see how the predictions vary for the next year considering the availability of the listing for the current year.

5. How can one ascertain the appropriate pricing of a new Airbnb listing based on its features and amenities?

One approach to pricing a new Airbnb listing involves using machine learning models trained on data about the existing listings. We compiled data on the existing Airbnb listings, including features like the number of bedrooms, individual amenities offered, and property type. We also included the corresponding listing prices. Several machine learning models

were trained on this data. These models included Linear Regression, Ridge Regression, Random Forest Regression, Lasso Regression, and XGBoost as seen in figure 25 - 29.

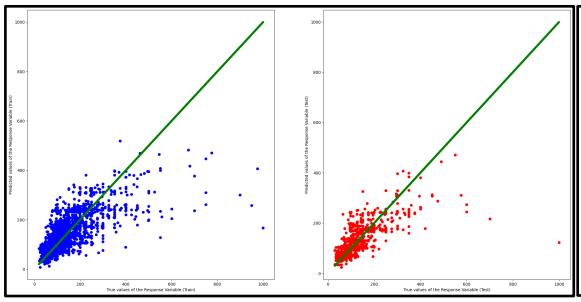

Figure 25. Liner Regression model

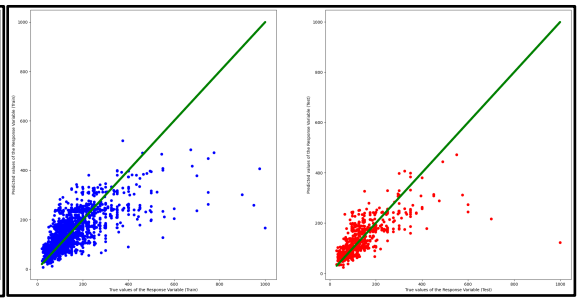

Figure 26. Ridge Regression model

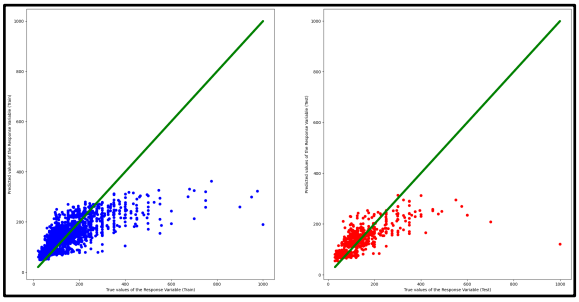

Figure 27. Random Forest Model

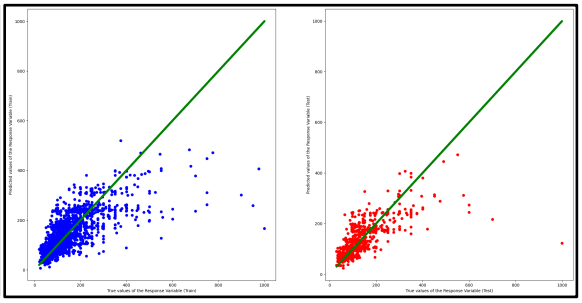

Figure28.Lasso Regression Model

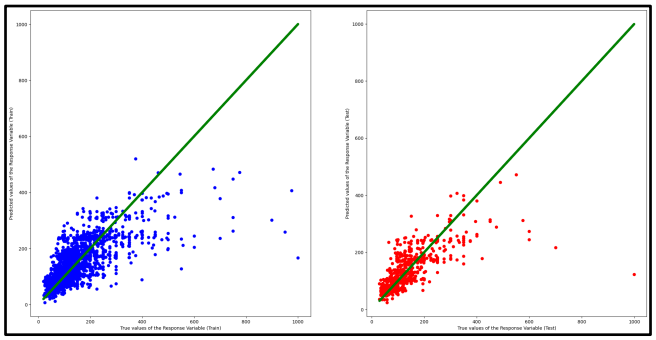

Figure 29. XG Boost model

	Linear	Ridge	Random Forest	Lasso	XG Boost
MSE	4198.610566407986	4198.478343041446	4803.63188550323	4198.7198533414885	4578.906163741651
RMSE	64.79668638447482	64.7956660822423	69.30823822247417	64.79752968548638	67.66761532477446

Table 1. MSE and RMSE scores for the 5 models

To assess how well each model performed in table 1, we evaluated the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of their predictions. An ANOVA test was conducted on the MSE scores, and a Kruskal-Wallis's test was performed on the RMSE

scores to check if there was any significant difference between these predictions. Since neither test revealed a significant difference between these models, we chose Ridge Regression based on it having the lowest MSE and RMSE scores.

After selecting Ridge Regression, we used the model in figure 30 to predict the price for a random listing and displayed our predicted price with the actual pricing.

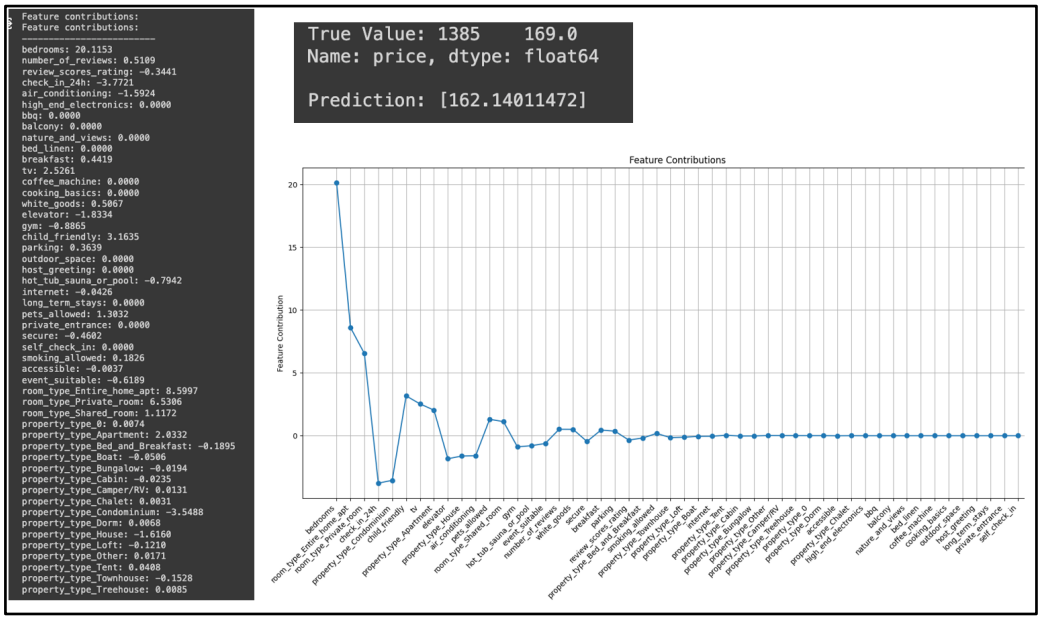

Figure 30. Price prediction model and feature contributing for change of price

The Ridge Regression model would analyse the features of the listing and assigned weights to them based on how they influence the pricing in the training data. The sum of these weighted features would result in the predicted price. Given a new listing with details about their features, amenities and property type in the same format, our model can be used to predict the appropriate pricing. Furthermore, by combining the machine learning predictions with market research and ongoing adjustments based on the personal interests of the host, one can establish an appropriate pricing strategy for their new Airbnb listing.

6. Is the price of a listing influenced by the type and the number of reviews it receives?

The correlation matrix in figure 31 revealed a weak positive correlation between the number of reviews and the review score rating (0.26). This means that listings with more reviews tend to have slightly higher scores on average. There is a very weak negative correlation between price and review score rating (-0.02). This means there is little to no relationship between price and how a listing is scored. The correlation between factors like sentiment score and review score, is positive and strong (around 0.9). This suggests these

factors have a much stronger influence on the review score rating than the number of reviews or price of the listing.

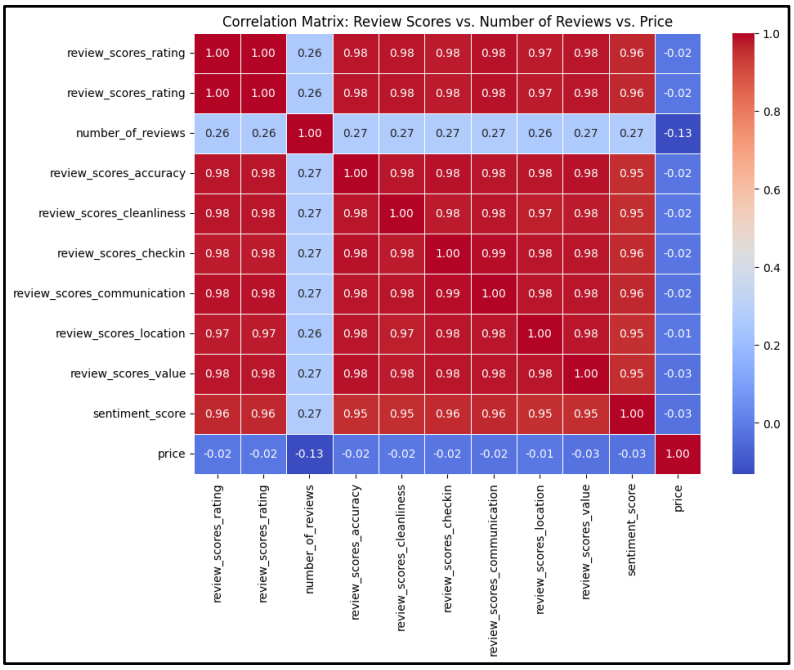

Figure 31. Correlation matrix for review scores vs number of reviews vs price

The number of reviews vs. price scatter plot in figure 32 does not show a clear relationship between them. The scatter plot aligns with the weak negative correlation coefficient described earlier between price and number of reviews. There is not a clear linear relationship between the two. While there are some listings with high numbers of reviews that are also expensive, there are also many that are inexpensive. This suggests price has little to no effect on the number of reviews a listing receives.

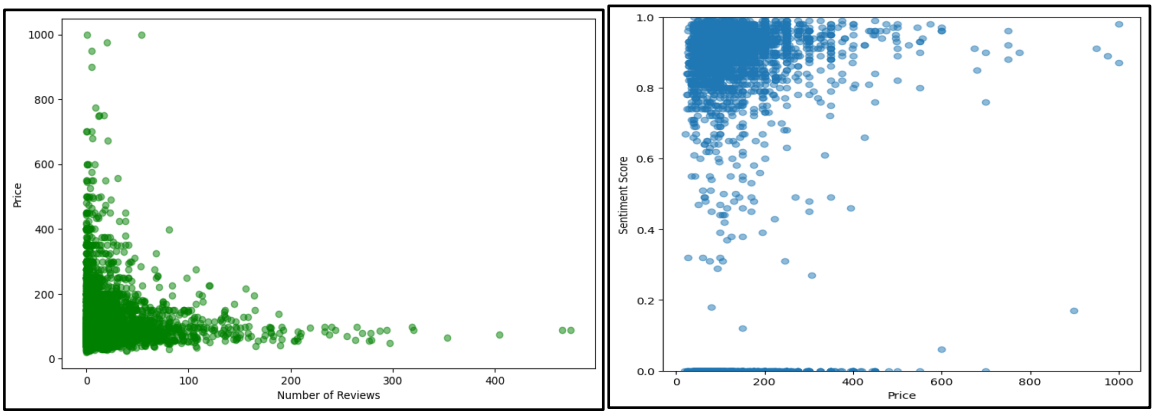

Figure 32. Number of reviews vs price scatter plot

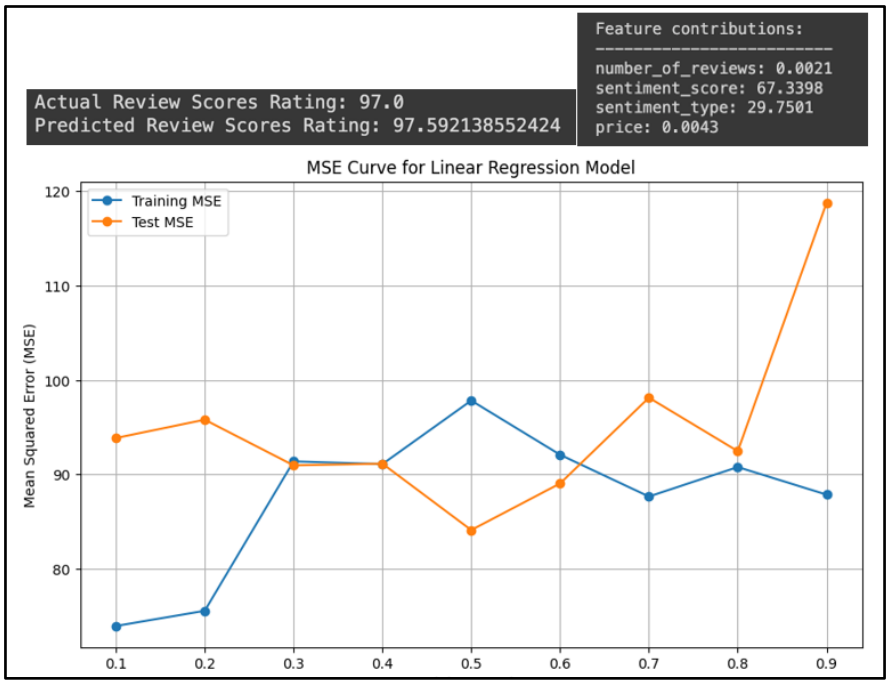

Figure 33. Review rating prediction model and feature contributing

The review score rating vs. sentiment score scatter plot indicated a positive correlation. Listings with higher sentiment scores, meaning more positive reviews, also had higher review scores. There are 2 ways to interpret this relationship:

- Sentiment reflects satisfaction: Listings with more positive sentiment in their reviews likely reflect a higher level of customer satisfaction, which translates into higher review scores.
- Review score influences sentiment: Customers who are happy with their stay are more likely to leave positive reviews, inflating the sentiment score.

6. CONCLUSION

The analysis of this Airbnb data has provided a wide range of meaningful insights for both guests and hosts. Hosts must be aware of the multiple variables that affect the price of a listing, like property type, number of bedrooms, room type, and the neighborhood. These must be taken into consideration for price prediction of a new listing. From the analysis of review scores, it is evident that communication and cleanliness are the most significant attributes that contribute to guest experience.

Hosts must be aware that guests tend to prefer listings that are clean, comfortable, quiet, close to downtown, with good internet, good neighborhood, parking, and good space. Lastly, Seasonal analysis shows that the peak demand is during summer when the availability is lower and the mean price is higher.

7. TAKEAWAY

From our analysis, we have come up with three takeaways that can be implemented by Airbnb to help hosts enhance their offerings.

- Hosts can focus on improvising the key insights using the guest reviews. This will improve the review score in turn increasing the listing price of the accommodation.
- Host can emphasize the listing's proximity by including the surrounding neighborhood in details like parks, downtown area, and the convenience of accessing attractions, restaurants, etc. in the description. This will enhance the chance of conversion rate of a guest.
- For a new listing, Airbnb must emphasize the details provided by the host and can use our ridge regression model to predict the price of the new listing.
- Airbnb can also provide suggestions or recommendations for a new host regarding the price of the new listing using price-prediction models.

8. FUTURE SCOPE

Future improvements could benefit from additional deeper analysis that includes numerical data on amenities. Clarity on availability i.e., whether amenities are disabled by the host or unavailable because of a booking, will provide greater insights. Additionally, data from the application or website on views, clicks, and conversions can be utilized for analysis and improvements. Tracking the number of bookings, cancellations, and reasons for cancellations will provide valuable insights into user behavior and satisfaction. Furthermore, integrating numerical values for neighborhood characteristics will enhance the accuracy of price prediction for a new listing, offering more tailored and competitive pricing strategies.

9. REFERENCES

[1] Airbnb. (2018, June 26). *Seattle Airbnb Open Data*. Kaggle. https://www.kaggle.com/datasets/airbnb/seattle

[2] *About Us*. Airbnb Newsroom. (2024, April 16). https://news.airbnb.com/about-us/

[3] Theerthala, A. (2022, October 21). *Investigating the Seattle Airbnb dataset*. Medium. https://akhiltvsn.medium.com/investigating-the-seattle-airbnb-dataset-dd564fa30ab8

[4] Moubine, R. (2023, November 8). *Seattle Airbnb Dataset: An exploratory data analysis*. Medium. https://medium.com/@moubine.rabab/seattle-airbnb-dataset-467fabd5fb76

[5] itsmeSamrat. (2023, April 29). *Exploratory Data Analysis on Seattle Airbnb Data*. Medium. https://medium.com/@itsmeSamrat/exploratory-data-analysis-on-seattle-airbnb-data-77eb4b2be0d1