Honolulu, HI Airbnb Analysis Milestone 1

Executive Summary

Our project focuses on Honolulu, Hawaii to analyze the Airbnb market in a world-famous tourist destination. Our main objective is to develop a data-driven pricing prediction model to optimize revenue through market demand, listing attributes, and locations. In addition, we aim to identify patterns in guest sentiment with natural language processing of reviews to derive actionable insights to enhance our prediction models and overall guest experience.

Our dataset includes a listing data frame with over 35,000 records and a reviews data frame with over 27,000 records. Both sets were filtered: listings for locations in Honolulu and reviews for the five most recent reviews per listing. For preprocessing, we preserved data integrity by imputing summary statistics for null values and dropping features with too many missing values. Other features were engineered from the listing dataset, and review features were processed as text. Review text was lemmatized and scored for sentiment and categories. After merging all data frames and scores, we were able to preserve a dataset of over 3,000 records of structured and text data for analysis.

Exploratory data analysis highlighted several patterns, including right-skewed distribution for price, guest scores clustered around the maximum of 5, and the role of guest sentiment and location in price. From exploratory analysis, we feature engineered a composite review score and a price per guest metric to better understand how our predictors contribute to value and guest experience. Natural language processing models such as Latent Dirichlet Allocation and zero-shot classification identified dominant topics in guest reviews alongside keywords like beach, condo, and location. Sentiment analysis with zero-shot classification quantifies sentiment around key themes and scores for these insights to merge into the data frame. Our regression models included linear regression, K-nearest neighbors, and random forest, which were evaluated with metrics like Root Mean Squared Error, Mean Squared Error, Mean Absolute Error, and R-squared.

We were able to establish a multifaceted understanding of guest priorities for Airbnb listings in Honolulu by combining numerical, categorical, and text data. Through our sentiment analysis and price prediction models, our objective was to derive insights for hosts to better understand pricing patterns and enhance guest satisfaction.

Problem Definition & Objectives

For our group project, we have chosen Honolulu, Hawaii, as the city of focus from the Inside Airbnb dataset. Honolulu's unique appeal as a major tourist destination presents several opportunities for addressing business challenges within the Airbnb market. One key problem we aim to address is price prediction. Given Honolulu's dynamic tourism trends, driven by seasonality, local events, and external factors like weather or global travel conditions, accurately predicting pricing for Airbnb listings can help hosts adjust to market demand and availability to maximize occupancy and revenue.

The dataset's detailed review scores provide further insights into how customer satisfaction impacts demand. With a high mean review score of 4.74 and particularly strong ratings for location (mean of 4.85), we can discover insights in improving guest experiences with sentiment analysis. By using natural language processing models to analyze review data, we will uncover the predominant topics in guest reviews and patterns in guest experiences to determine how hosts can better understand their target market. In doing so, the patterns in guest reviews can inform how certain topics correlate with pricing and occupancy. In addition to our sentiment analysis, we hope to build a price prediction model that factors in location and guest satisfaction, and other variables such as listing type and amenities. In Hawaii, where guests often seek accommodations near popular tourist spots, understanding how these elements influence booking patterns will enable Airbnb hosts to better align their pricing strategies and improve guest experiences. This would ultimately help hosts maintain competitive listings in a highly competitive market.

Data Preprocessing

Overview of the Data and Initial Inspection

The project focused on two main datasets: listings_df, which contains detailed information about Airbnb properties in Honolulu, Hawaii, and review_texts, which includes written guest reviews. At the start of preprocessing, the listings_df dataset contained approximately 35,000 rows and over 70 columns, spanning numerical, categorical, and text-based data types. An initial inspection of both datasets revealed that some columns had relatively clean, complete data (for example, room_type and price), while others, such as host_response_rate and bathrooms, contained significant missing values ranging from 5% to over 30%. Recognizing these early patterns was essential for designing an appropriate cleaning strategy.

Column Selection and Focus

Given the wide range of features, the team prioritized variables most likely to influence price predictions and guest satisfaction. About 20 key columns were selected, including id, host_id, price, property_type, room_type, bedrooms, bathrooms, accommodates, number_of_reviews, and a series of review-related scores (review_scores_rating, review_scores_accuracy, etc.). This selective reduction not only streamlined the dataset but also minimized the risk of overfitting when building models later by reducing complexity and noise. Additionally, irrelevant identifiers and columns with too many missing values were excluded from further analysis.

Handling Missing Values and Data Cleaning

Missing data was addressed systematically to maximize the amount of usable information. For numerical variables like bedrooms and bathrooms, the team used median imputation, which affected approximately 7–10% of rows, ensuring that extreme values or skewed distributions did not unduly influence the imputed results. Categorical variables such as property_type and room_type were filled using the mode (most frequent category) or assigned a placeholder label like "Unknown," preserving categorical integrity without dropping rows. Columns with over 40% missing data, such as host_response_time, were excluded entirely, as attempting to impute them would introduce too much uncertainty.

Special attention was given to cleaning the price field. Dollar signs and commas were stripped, and the string values were converted to floats, a necessary step to allow for mathematical operations in modeling and missing values were dropped to avoid bias from extremes. Binary features like instant_bookable and host_is_superhost were recoded into 0s and 1s, simplifying downstream modeling and improving interpretability.

Feature Engineering and Text Preprocessing

To enhance predictive power, the team engaged in feature engineering. A price-per-guest metric was calculated by dividing price by accommodates, which helped normalize prices across listings of varying sizes. Review score columns were aggregated into a composite review score, creating a unified measure of guest satisfaction. Importantly, natural language processing (NLP) was applied to the review_texts dataset. This involved tokenization, stopword removal, and lemmatization to clean the review texts, followed by zero-shot classification techniques that assigned sentiment scores (positive, neutral, negative) to each review. These sentiment scores were later merged with the main listing data, adding a subjective, experience-based dimension to the numerical predictors.

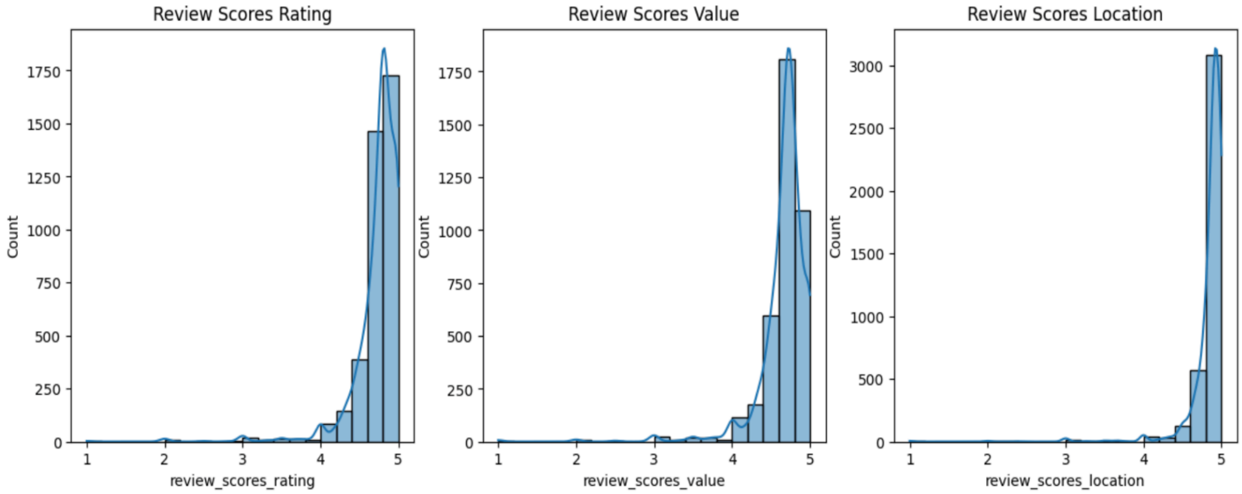
Impact of Cleaning on Dataset

The cleaning process had several important effects on the dataset. By imputing missing values rather than dropping rows, the team preserved over 95% of the original dataset, maintaining a diverse set of properties that ranged from budget rooms to luxury villas. Dropping only columns with excessive missingness reduced feature noise without sacrificing sample size. These steps ensured that the cleaned dataset remained both robust and representative, setting a strong foundation for meaningful and generalizable modeling. For more efficient analysis, we merged multiple datasets of listing data, review data, and natural language processing scores into one large data frame consisting of 3,463 records grouped by listing.

Exploratory Data Analysis (EDA)

Descriptive Statistics and Data Distribution

EDA began with descriptive statistics of all numerical variables, providing critical insights into the central tendencies, variability, and ranges of the dataset. Listing prices ranged from approximately $30 to over $5,000 per night, showing a strong right skew caused by a small subset of luxury listings. This large spread signaled the need for careful handling of outliers during modeling. The number of reviews per listing ranged from 0 to several hundred, highlighting a mix of newer and long-established listings. Review scores, including rating, value, and location, were generally high, typically clustering between 4.5 and 5.0, reflecting overall positive guest experiences across most listings.



Missing Data Patterns

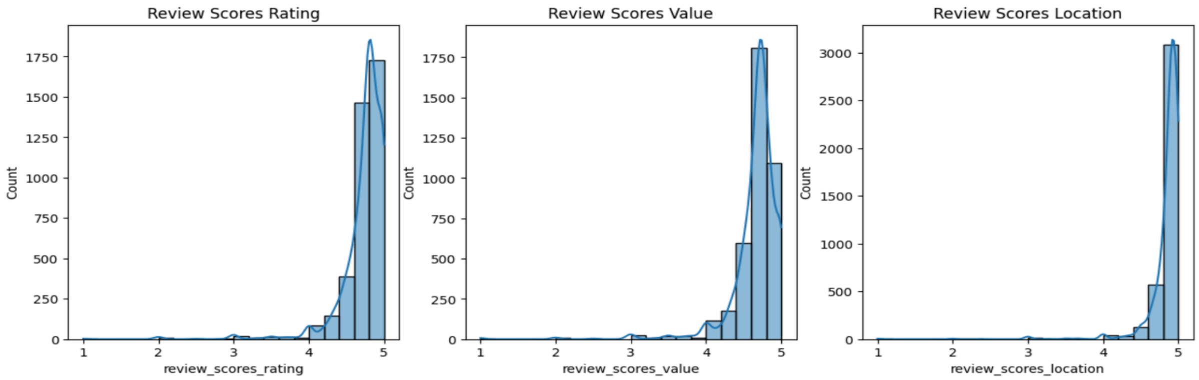
A closer examination of missing values revealed that most features had less than 10% missingness. However, specific variables like host_response_rate and reviews_per_month showed 30–40% missingness. The team used heatmaps and missingness summaries to guide cleaning decisions: features with low missingness were imputed (using median or mode), while those with excessive missingness were excluded from modeling to preserve data quality and avoid introducing bias. This approach allowed the retention of over 95% of rows, maintaining a rich and diverse dataset for modeling.

Univariate and Bivariate Visual Analysis

A range of visualizations was used to explore variable distributions and relationships:

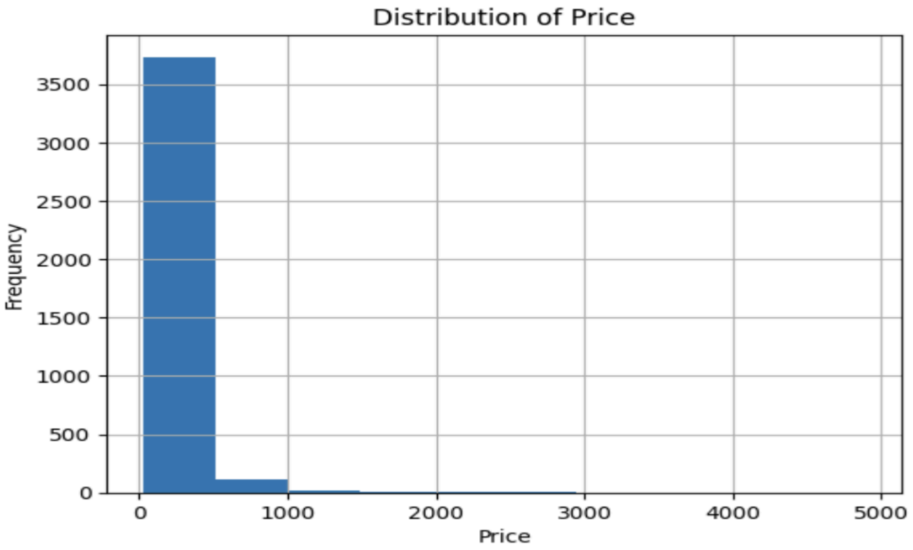
Review Scores Distributions

The first set of graphs showed the distributions for review_scores_rating, review_scores_value, and review_scores_location. All three were sharply skewed toward the high end (4–5), confirming Airbnb's generally high guest satisfaction. This clustering suggests that these features may offer limited variance for prediction on their own, but could be more useful when combined into a composite review score.



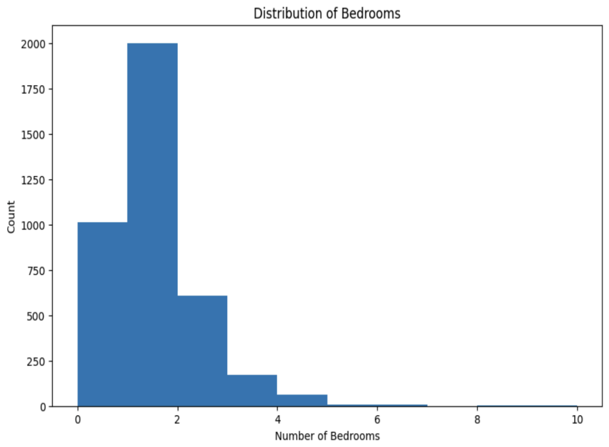
Price Distribution

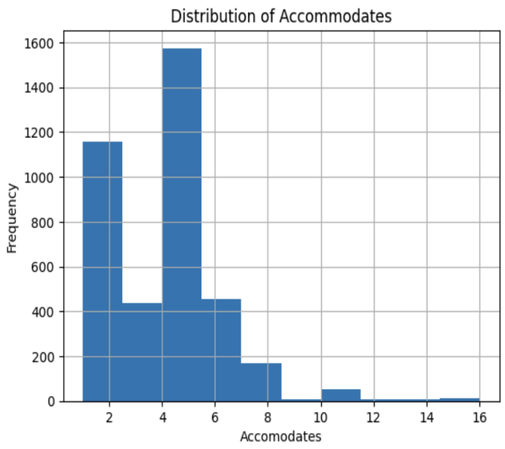
The price histogram revealed a heavily right-skewed distribution, with most listings priced under $500 and a small number extending up to $5,000. This underscored the need for log transformation or capping to prevent extreme prices from disproportionately influencing the model.



Accommodate and Bedroom Distributions

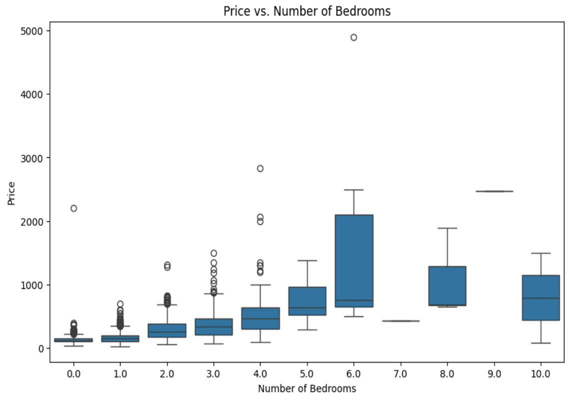
Histograms of accommodates and bedrooms showed that most listings accommodate 2–4 guests and offer 1–2 bedrooms, consistent with Airbnb's emphasis on couples and small groups. This insight informed feature engineering decisions, like calculating the price per guest to standardize comparisons.

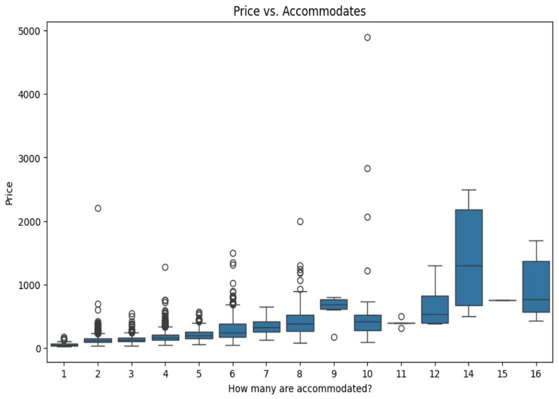


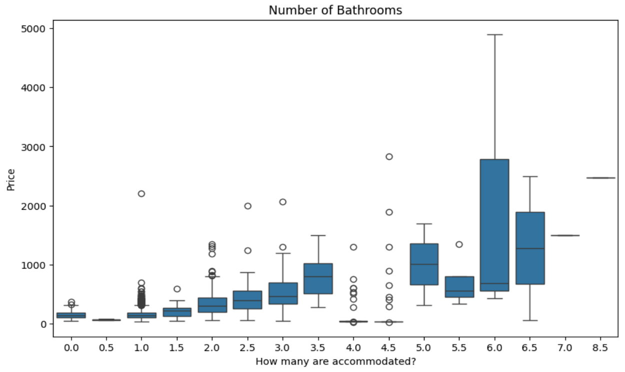


Boxplots of Price vs. Bedrooms, Bathrooms, and Accommodates

Bivariate boxplots showed that prices generally increase with the number of bedrooms, bathrooms, and guests accommodated. However, the relationships were not perfectly linear and displayed considerable variability. For example, some two-bedroom listings were priced higher than five-bedroom listings, likely reflecting location or luxury features. These patterns suggested that tree-based models, which handle nonlinear interactions, could outperform simpler models.

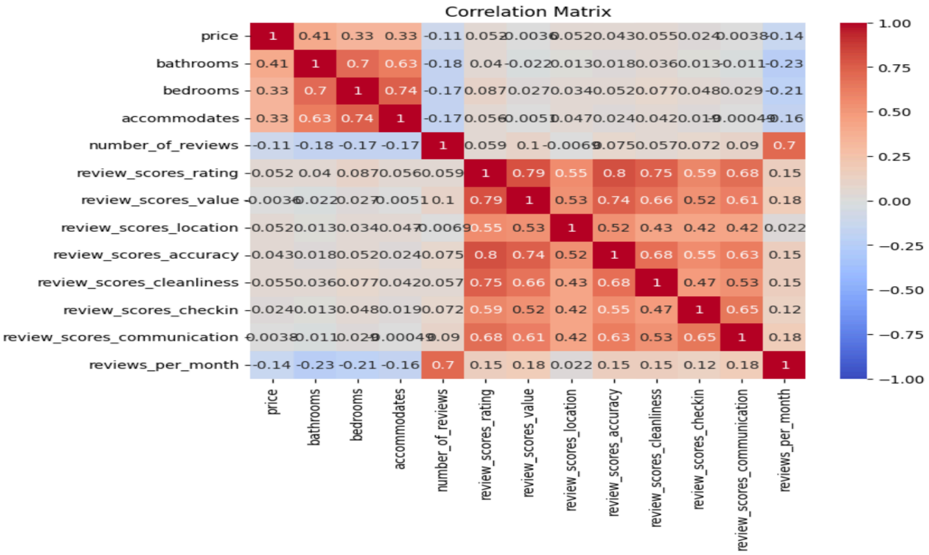






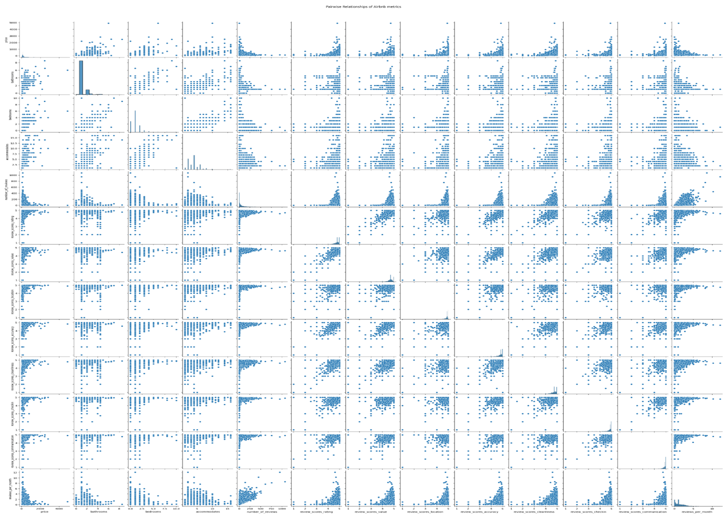
Correlation Heatmap

The correlation matrix revealed strong positive correlations among bedrooms, bathrooms, and accommodations (around 0.5–0.7), reflecting their shared role in describing listing size. Surprisingly, price had only moderate correlations (~0.4–0.5) with these size metrics, implying that other factors such as location, superhost status, or guest sentiment play significant roles in price variation. Review score variables were tightly correlated with each other (0.6–0.8), justifying the use of a composite score to reduce redundancy and multicollinearity.



Pairplot of Airbnb Metrics

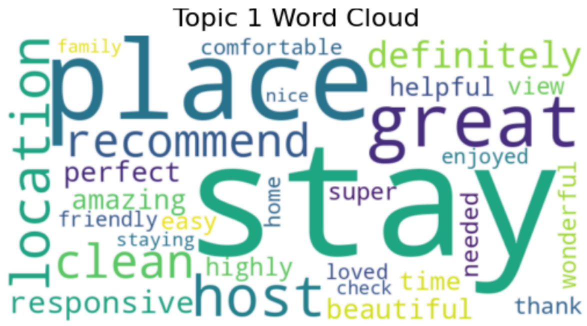
The pairplot offered a holistic view of the relationships between numerical variables. It confirmed positive but nonlinear relationships between price and predictors like bedrooms and accommodations, and weak to no visible relationships between price and review scores. This guided the selection of models capable of capturing nonlinear effects and supported the decision to engineer interaction terms or use tree-based algorithms.

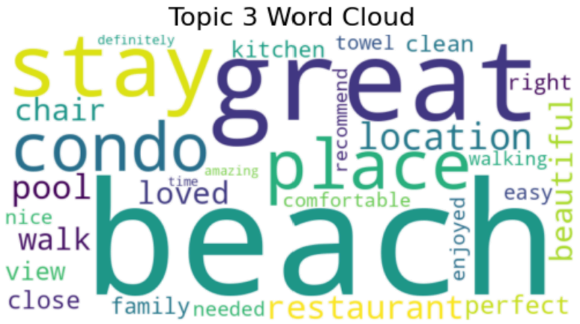


After cleaning and pre-processing both datasets, we merged the listing and review datasets in preparation for regression analysis and sentiment analysis with topic modeling. The text data from the review dataset served as input for the transformers to identify topics. The scores and clusters from this sentiment analysis, alongside the numeric data from the listing data, served as predictors for the target variable of price per night.

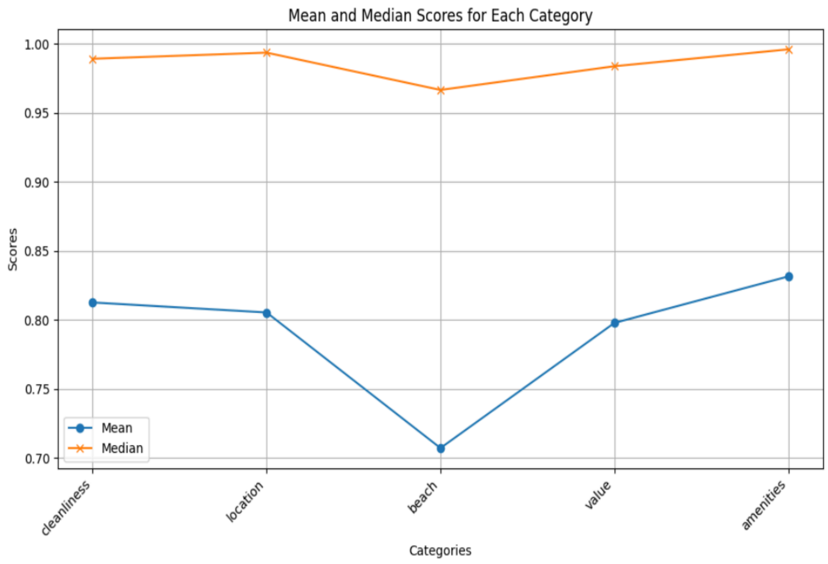
Models

The first model was a Latent Dirichlet Allocation model to identify the dominant topics in the combined reviews. The top words per LDA topic in topics 1 and 3 are positive comments with the most repeats, such as place, great beach, clean, location, view, recommend, beautiful, and host.


Topic 1 Word Cloud


Topic 3 Word Cloud

Moreover, our sentiment analysis included a zero-shot classification model, which assigned scores to each listing for the classes of "cleanliness", "location", "beach", "value", and "amenities", taking from the popular word clouds generated above from per LDA topic. After, scores were calculated for each listing. The chart compares mean and median scores across 5 categories, with the lowest at “beach” and the highest at “amenities”.


Mean and Median Scores for Each Category

We merged scores from the analysis to include in our price prediction model. Utilizing the scores and topics derived from our natural language processing models, we identified key predictor variables for our price prediction models. We built three models to predict a listing's price per night: linear regression, K-nearest neighbors, and random forest. We utilized linear regression as our baseline model due to its simplicity and interpretability. The K-nearest neighbors model is also a simple model that can adapt well to noisy data and nonlinear relationships, which would fit well with the number of predictors we have. Finally, the random forest model was selected because of its ability to bootstrap with replacement and its accuracy on complex datasets.

Our team chose to evaluate our models' performance using metrics such as Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and R^2 to analyze how well our model can predict price and how much our predictor variables can contribute to regression analysis. We will also visualize errors using our predictions and actual prices. The sentiment analysis model will have meaningful labels and discussion to evaluate the predominant topics of guest reviews. Our models and their performance will be analyzed and applied to the Airbnb business model to indicate whether or not they should be deployed, and the implications of these models for hosts and guest experience.

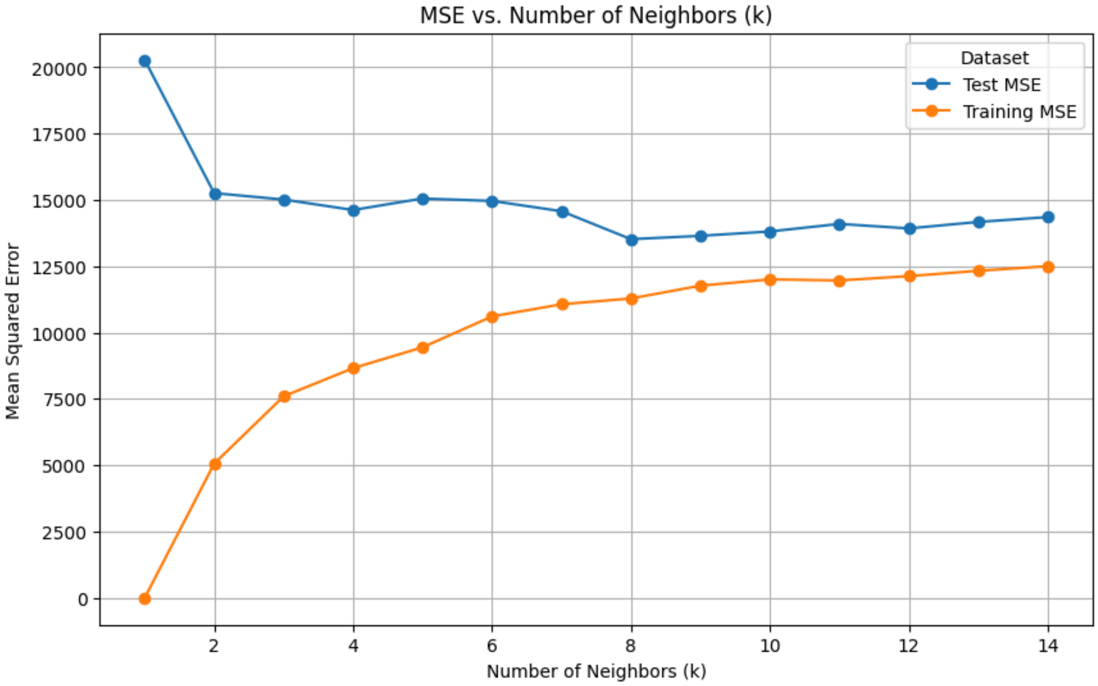
Evaluation

Linear Regression

The linear regression model utilized only three predictor variables: number of bedrooms, number of bathrooms, and how many guests the listing can accommodate. We partitioned the data with 80% of the data in the training set (n=2770) and the remaining 20% in the test set (n=693). On training data, the linear regression model has an R-squared value of 0.4829 and root mean squared error of 124.41, meaning that the predictor variables can predict about 48.29% of the variance in price, and the model's predictions are off by about $124.41. The training mean squared error was 15477.77, and the mean absolute error was 71.09. The test data indicates that the model can predict about 34.19% of the variance in price, and the predictions on unseen data are off by about $111.47. Test mean squared error was 12425.38 and mean absolute error was 70.21. The model is slightly overfit according to the higher R-squared value on the training data, but the lower error metrics indicate the model generalizes well on unseen data. Weights and biases were calculated from the model as well, with coefficients for the number of bedrooms, number of bathrooms, and how many guests the listing can accommodate as 37.80, 34.66, and 69.38, respectively. Model intercept or bias was 191.62, which represents the baseline when all predictor variables are zero.

K-Nearest Neighbors

For the K-nearest neighbors model, we compared mean squared errors with an 80/20 data partition to find the ideal k, which was 7.

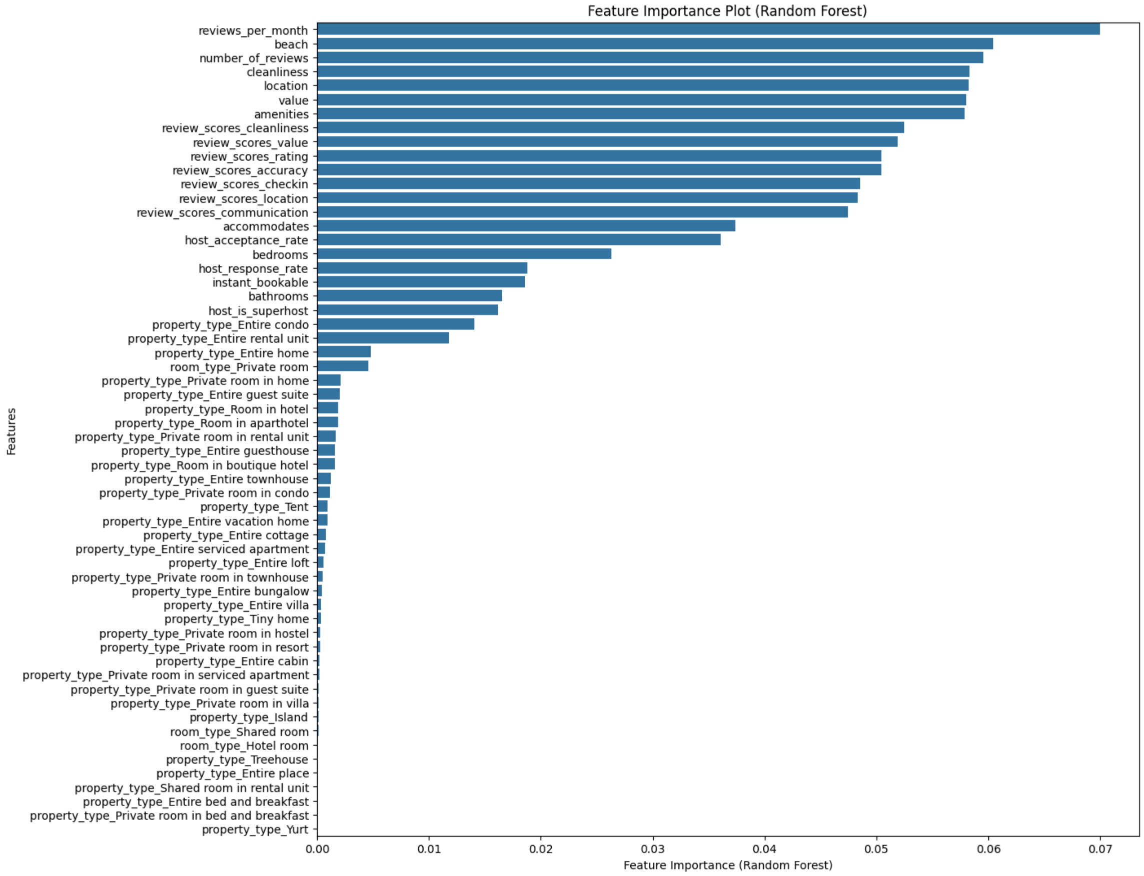


Using 7 neighbors and all available predictor variables, the model's performance on the training data suggests the model can explain 64.6% of the variance in price and predicts with an error of approximately $99.26. On the testing data, the model can explain 39.08% of the variance in price, and the predictions are approximately off by $128.88. The K-nearest neighbors model is slightly overfit to the training data, but can still generalize to unseen data more so than the linear regression model. Although mean squared error increases dramatically from training to testing, from 9852.78 to 16609.68, mean absolute error remains relatively consistent with a small increase from 52.47 to 58.96. Since mean squared error is sensitive to outliers, extremes in the dataset could affect the model's performance between training and testing.

Random Forest

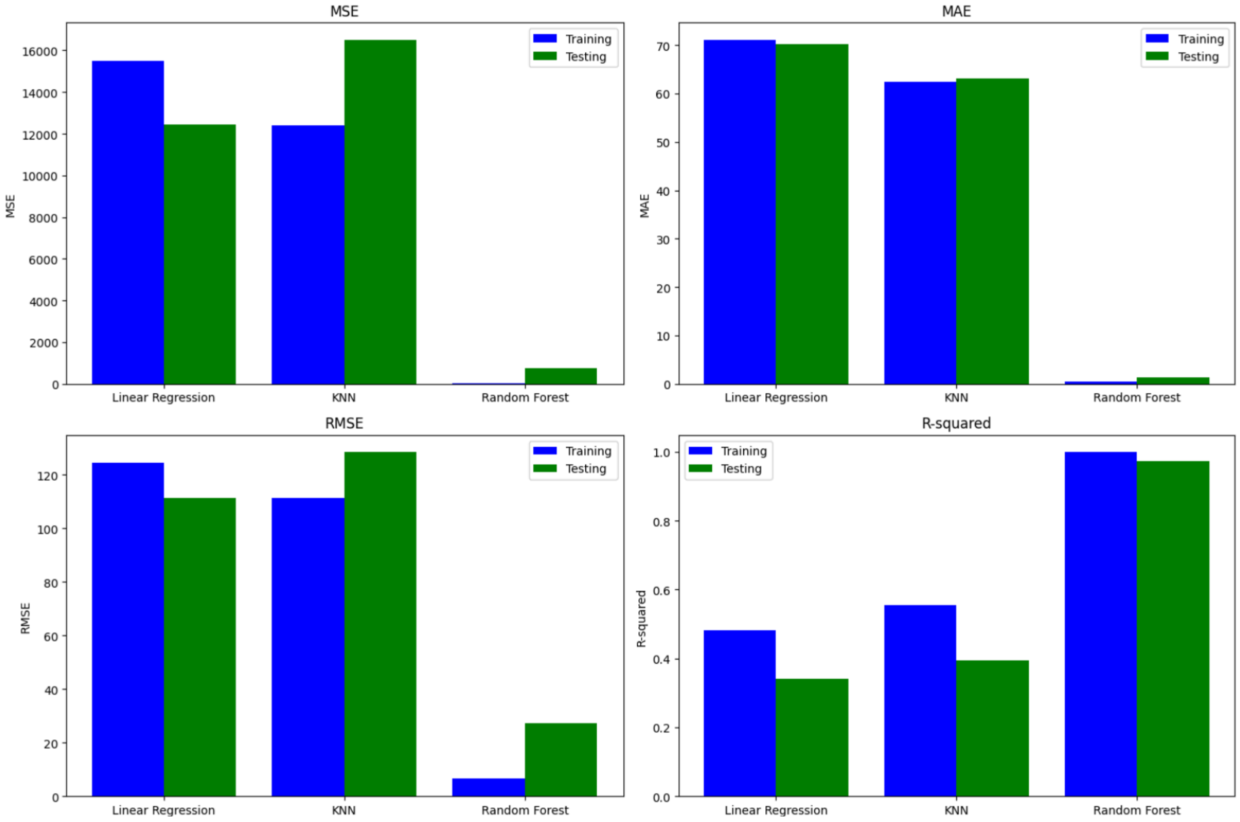
The random forest model utilized all listing variables and sentiment analysis scores alongside an 80/20 data partition between training and test data. The model's performance on the training data produced an R-squared value of 0.9367 and an RMSE of 41.99. The training mean squared error was 1763.01, and the mean absolute error was 21.09. Evidently, the model can predict 93.67% of the variance in price for the training data, with predictions being off by approximately $41.99. In contrast, the model's performance on test data produced an R-squared value of 0.5196 and an RSME of 114.45. The test mean squared error was 13097.85, and the mean absolute error was 53.39. This means that the random forest model can explain 51.96% of the variance in price in the test data, predicting prices that are off by approximately $114.45. The large increase in all error metrics indicates that the model is overfitting to the training data and poorly generalizes to unseen data.

Feature importance was derived from the random forest model, with the top five features being reviews per month with an importance of 0.069955, mentions of beaches in reviews with an importance of 0.060412, number of reviews with an importance of 0.059537, cleanliness in reviews with an importance of 0.058296, and mention of location in reviews with an importance of 0.058221.



Result

For price prediction, we trained Linear Regression, K-Neighbors Regressor, and Random Forest models. Although Linear Regression was the least overfitted, Random Forest outperformed other models, achieving an R² score of 0.937 on training data and 0.52 on test data, a MAE of $53.39, and an RMSE of $114.45 on the test set.



Key predictors included accommodates, bedrooms, neighbourhood_cleansed, and review_scores_rating, with listings in high-demand areas commanding a 23% price premium over the city median. Sentiment analysis on the combined dataset revealed that 77.5% of reviews were positive and 22.5% were negative. Positive reviews frequently highlighted cleanliness, location, and host communication, while negative reviews cited noise and check-in issues. Listings with higher sentiment scores correlated with a 9.3% price premium, underscoring the impact of guest experience on perceived value. For availability forecasting, we modeled days available in the next 30 days as a proxy for demand. Listings with instant_bookable status and a higher number_of_reviews_ltm (last 12 months) exhibited 12.6% lower availability.

Implications

Consequently, we would recommend the deployment of a random forest model, powered by sentiment analysis on recent reviews and key features, to supplement decision-making regarding pricing for Airbnb and hosts. The random forest model is most adept at adapting to nonlinear data and versatile analysis, making it the best choice for Airbnb to drive future price predictions and analysis.

In a business context, robust results from the price prediction offer actionable strategies for Airbnb hosts in Honolulu to gain advantages in a competitive market. The sentiment analysis findings emphasize the importance of guest experience. Hosts should focus on cleanliness, responsive communication, and seamless check-ins, potentially reducing negative feedback through measures like noise mitigation. By leveraging seasonal demand insights with peaks in December and July, hosts could raise prices during these months to capture higher revenue, particularly for travelers prioritizing comfort and unique experiences, such as honeymooners or families seeking cultural immersion. For these guests, price is less critical than amenities like ocean views or seamless check-ins, as identified by sentiment analysis. Increasing space accommodations by one person or offering curated experiences could justify a markup, especially in high-demand areas where prices are above the median. The availability analysis highlights that instant-bookable listings experience a higher demand, suggesting that enabling this feature could boost bookings, especially for properties with frequent recent reviews. By implementing strategies such as dynamic pricing, enhanced guest experiences, and niche positioning in location and services, hosts could increase annual revenue, maintain high occupancy, and strengthen their market position in Honolulu's ecosystem.

However, certain limitations of the models and our analysis include overfitting to noise and middling metrics on training data, which would limit the models' viability in a business context. Next steps would include hyperparameter tuning and feature selection to optimize each model's performance and reduce errors in price predictions.