Exploratory Data Analysis

Final Project

Beyond the Hotel : A Deep Dive into Airbnb in Asheville, NC

Team Airbnb Asheville

Lakshmi Sai Ram Marupudi

Niharika Ganji

Yerramsetti Dharma Teja

Sravya Vujjini

Beyond the Hotel: A Deep Dive into Airbnb in Asheville, NC

Introduction

The way we travel and lodge has changed a lot, thanks to platforms like Airbnb. They offer a variety of distinctive accommodations for travelers worldwide. In our effort to explore the complex dynamics of these Airbnb listings, we decided to center our research on the city of Asheville, North Carolina. Surrounded by scenic beauty, Asheville is a city known for its unique charm with beautiful views captivating hilly Applachian mountain ranges.

Our choice to focus on Asheville goes beyond a simple geographical preference; it originates from a personal connection, as our mini project team is affiliated with the state of North Carolina.

Objective

Our project focuses on understanding the details of Airbnb listings through the exploration of two core questions:

1. What would be the average price of an Airbnb listing based on its location, size, and the availability time of the year? How do these pricings vary with respect to the factors considered?
2. What factors contribute most to the guest's perceived value of a listing – Review Score?

Our primary objective is to determine the average price of Airbnb listings based on the considered factors, aiming to empower both hosts and guests with valuable insights that seek to facilitate more informed decisions during the listing and booking processes. By delving into the key factors influencing the pricing of Airbnb listings, we aim to provide hosts with actionable insights to optimize their offerings.

As part of our efforts to enhance Airbnb's system, we delve into the factors that influence guest's perceived value (review score) of a listing. This exploration unveils the specific aspects that make guests genuinely appreciate a place, so hosts can improve their offerings, and guests can have a better experience.

Benefits:

- Gaining insights into how the pricing works significantly improves the transparency of the Airbnb marketplace in Asheville. This helps hosts by assisting them in establishing competitive prices and allows travelers to make more informed decisions while booking.
- Identifying the key factors that influence price listings not only benefits hosts by maximizing their property's value but also provides valuable insights that contribute to a broader understanding of dynamics within the shared accommodation industry.
- The factors contributing to the guest's perceived value of listing hold significance for hosts aiming to enhance their listings and for Airbnb in refining its criteria for guest evaluations, ultimately improving the quality of stays on the platform.

Description of data

Beyond the Hotel: A Deep Dive into Airbnb in Asheville, NC

Our dataset, acquired from Inside Airbnb (http://insideairbnb.com/get-the-data/), is titled "Airbnb Listings in Asheville, NC". The Asheville dataset comprises 3288 observations, each with 75 variables offering diverse information, ranging from basic listing attributes to host-related details and customer review metrics. Each entry in the dataset corresponds to a unique Airbnb listing in the time frame of 1 year. The data set we used is during the time period 13 sep 2022 – 13 sep 2023.

Data Cleaning and Exploration

In our data cleaning and exploration process, we narrowed our focus to specific variables relevant for our research questions:

Price Determination and Factors:

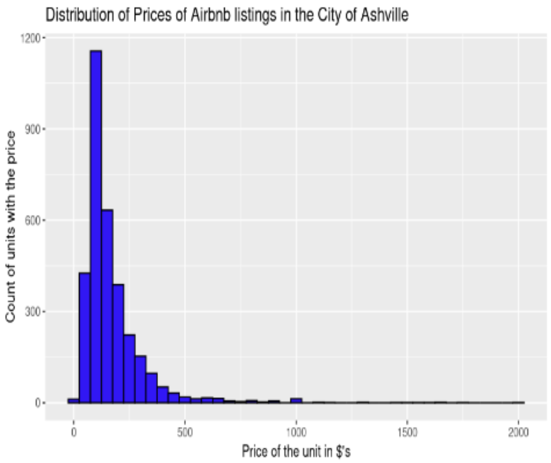
- Latitude
- Longitude
- Accommodates
- Bathrooms
- Bedrooms
- Availability_365
- Price (Response Variable)

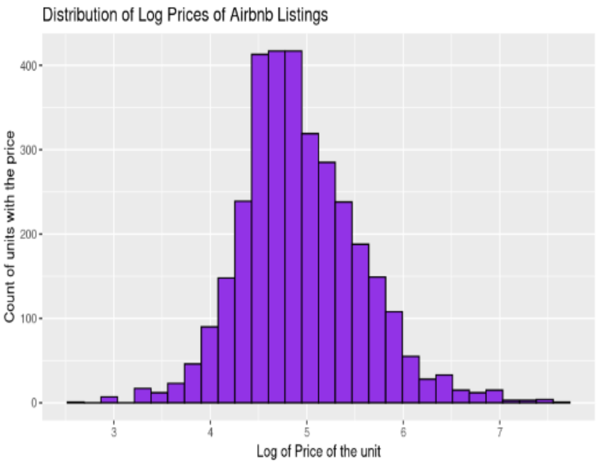
Factors Influencing Review Score:

- Review Score (Response Variable)
- Host Response Time
- Host is Superhost
- Host Identity Verified
- Host Response Rate
- Cleanliness Score
- Check-In Score
- Amenities

We addressed missing values by imputing the median to maintain the central tendency of the data. Additionally, we extracted numeric values and converted categorical variables to numeric format.

During the analysis of distribution of prices, we observed a rightward skewness, indicating a concentration of listings at the lower end. To address this, we applied a log transformation on the Airbnb listing prices, aiming to mitigate the influence of extreme values and achieve a distribution closer to normal. This transformation resulted in a more balanced distribution, to analyze the pricing patterns effectively.





Feature Engineering Weighted Review Score:

In assessing guest satisfaction for Airbnb listings in Asheville, we chose to focus on a more reliable metric: the **weighted review score**. This approach stems from the understanding that while **review scores** provide valuable insights into guest experiences, they can be significantly influenced by the **number of reviews** a listing receives.

Step-1: Log Transformation of Number of Reviews:
The variable **number of reviews** is observed to be a **right skewed distribution** as shown in Figure -1 , which was transformed by applying natural logarithm.

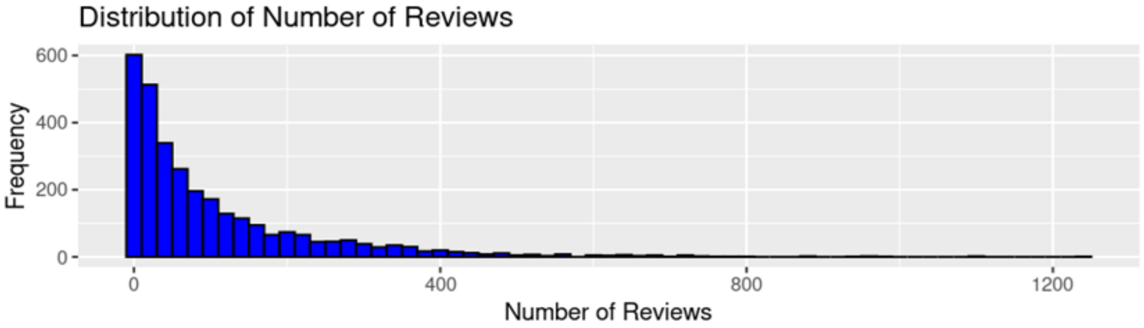


Figure - 1: Distribution of Number of Reviews of each Airbnb Listing

Step-2: Normalization of Log-Transformed Reviews: To avoid potential biases from varying scales, we normalized the logarithmic reviews by dividing each value by the maximum value in the log-transformed reviews column. This step scales the logarithmic reviews between 0 and 1.

Step 3: Calculation of Weighted Review Score: The core of our methodology involves creating a weighted review score. This score is computed as the sum of the normalized logarithmic reviews and the original review scores.

weighted_review_score = (normalized_log_reviews + review_scores_rating)

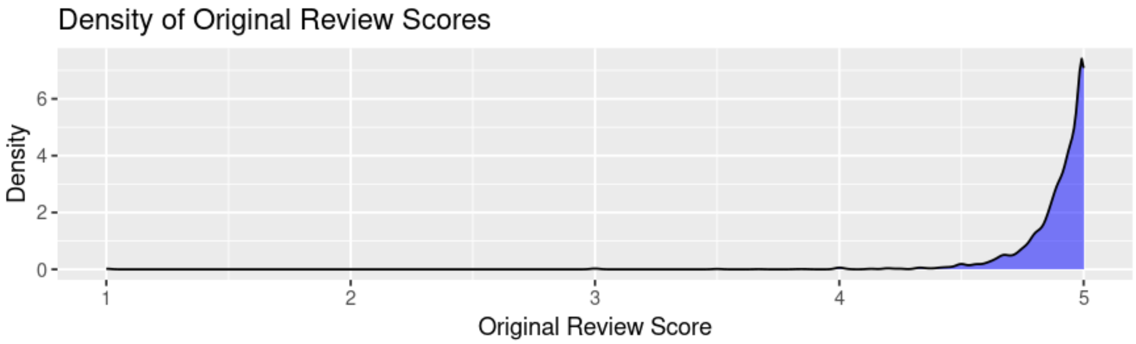
By combining these two factors, we aim to provide a more comprehensive representation of guest satisfaction. This will add to the original review score in the range of [0,1], we term it as '**added incentive'**. Hence, the weighted score now ranges between [1,6], where each score is dragged by a value between [0,1]. The listings with higher number of ratings are dragged a bit more than the listings with a lower number of ratings.

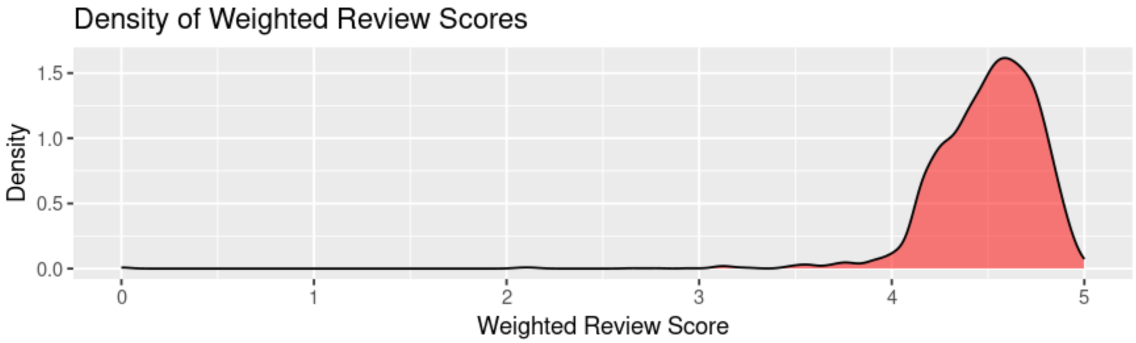
Furthermore, this will not outweigh listings of higher review score (5) with lower ratings (normalized_log_value~0) in comparison to popular ratings (normalized_log_value~1) with average review score (3), as the **added incentive is between [0,1].** This formula is similar to weighted score = review score + log (number of reviews)/scaling factor where the scaling factor is the log of maximum number of ratings.

Step-4: Scaling to a 5-point Scale: For ease of interpretation and comparison, we scaled the resulting weighted review score to a 5-point scale. This was achieved by subtracting the minimum value of the weighted review score, dividing by the range of the scores, and then multiplying by 5.

weighted_review_score - min(weighted_review_score)) / (max(weighted_review_score) - min(weighted_review_score))) * 5

By employing this methodology, we aim to provide a more reliable metric - **weighted reliable metric** for assessing guest satisfaction, one that accounts for the nuances introduced by varying numbers of reviews. The distribution of weighted review score, scaled to a 5-point scale is as follows:



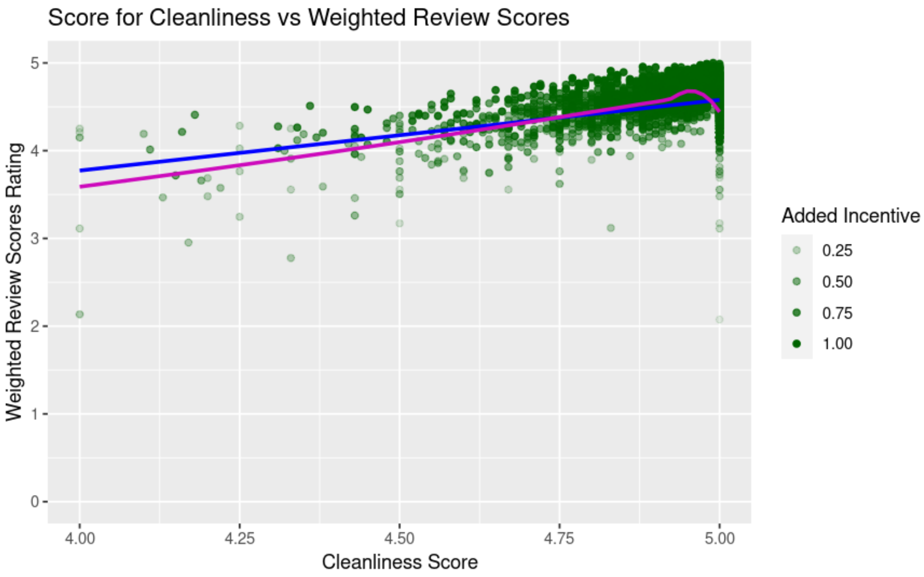


The original review scores exhibit a left-skewed distribution, with a high density of scores clustered towards the higher end, indicating that most listings have high review scores. This skewness could potentially overemphasize the satisfaction level of guests.

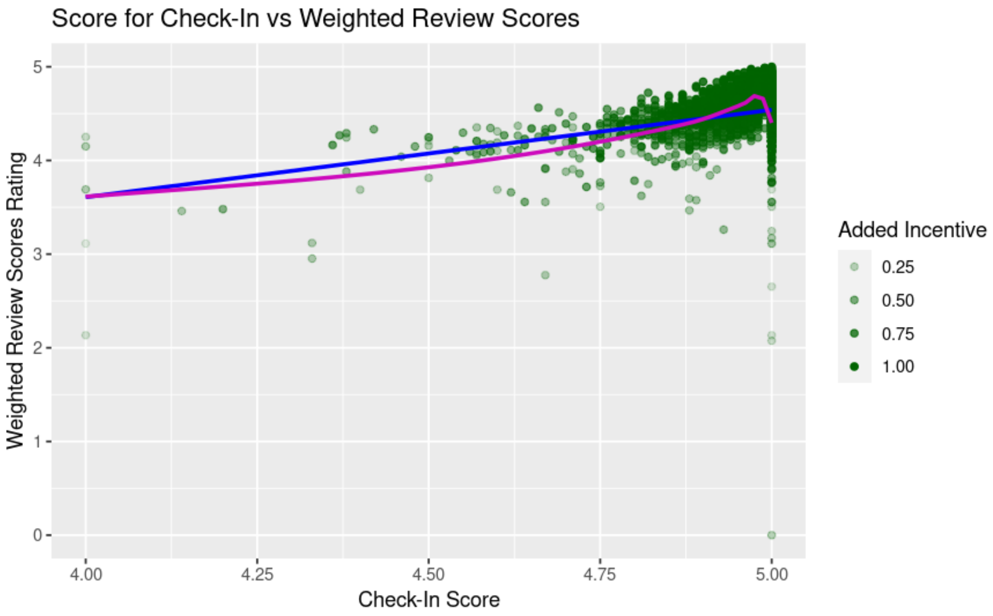
In contrast, the weighted review scores show a more normally distributed pattern, which suggests a more balanced and reliable reflection of guest satisfaction. The application of weighting has seemingly reduced the skew, spreading out the scores more evenly across the spectrum. This adjustment not only accounts for the quality of reviews but also integrates the quantity of feedback received, thereby offering a more comprehensive view of guest experiences.

Exploring Relationships

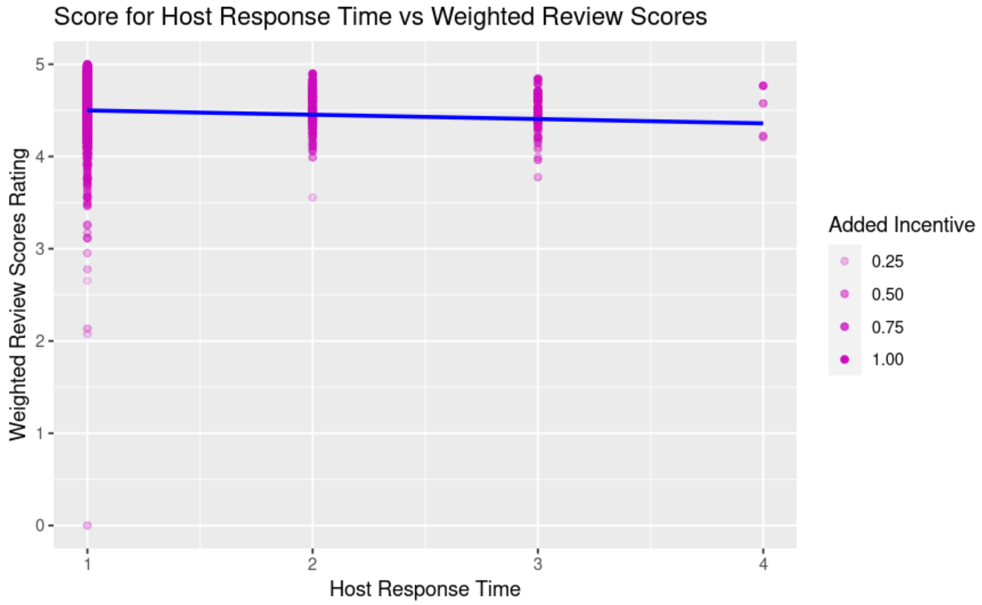
Before proceeding to the modeling of the response variable, let's explore the relationship of the response variable with each of the explanatory variables. As mentioned earlier, our explanatory variables are: cleanliness_score, checkin_score, host_is_superhost, host_is_verified, host_response_time, host_response_rate.



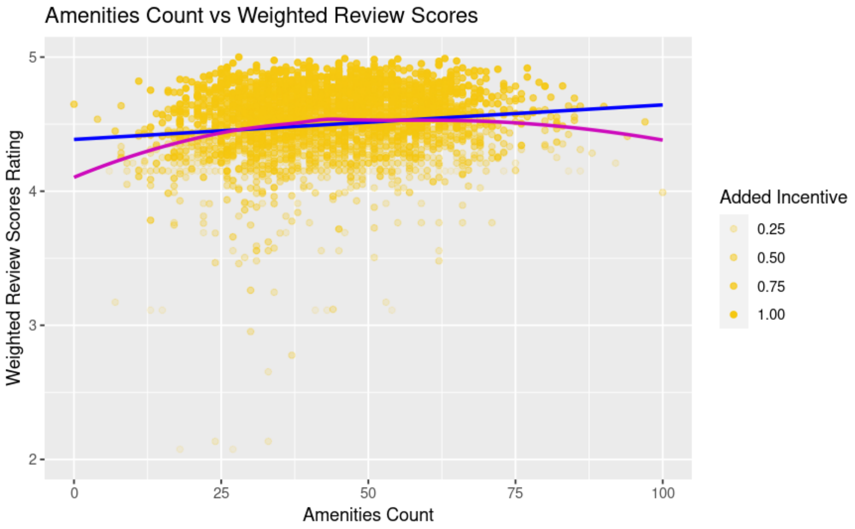
The positive slope in the graph suggests that as cleanliness scores increase, so do the weighted review scores, indicating that guests place high value on cleanliness when rating their stay. The concentration at the higher end shows that listings with exceptional cleanliness are rewarded with higher weighted scores. Although there are a significant number of listings with high cleanliness scores, the lower number of ratings add to their weighted score at the lower end.



Similar to cleanliness, the positive trend line for check-in demonstrates that smooth check-in processes are likely to result in higher weighted review scores. The cluster of higher scores at the top end emphasizes the importance guests place on an easy check-in experience. The decline in weighted review score at the higher end of check-in scores is due to a lower number of ratings for those listings as seen.



The downward trend indicates that longer host response times may lead to lower weighted review scores, suggesting that guests appreciate quick and efficient communication from hosts.



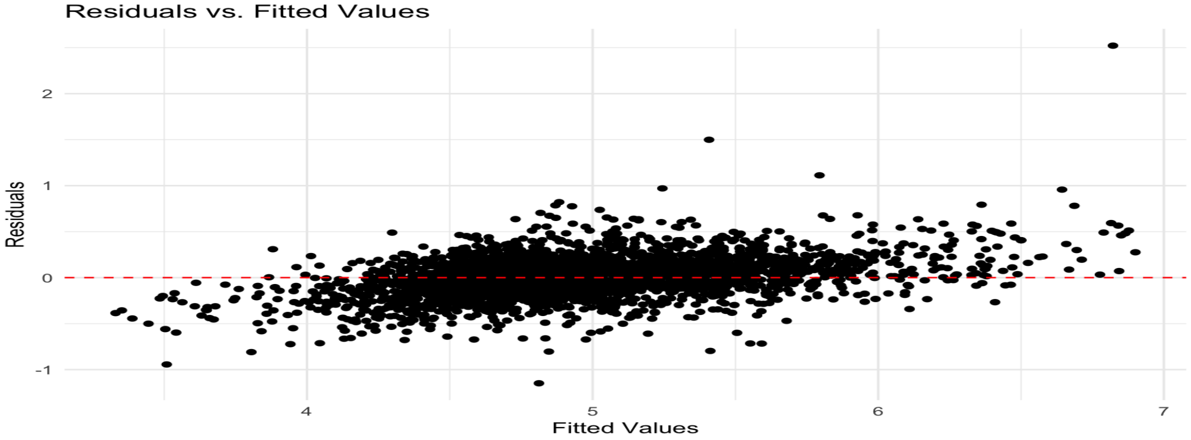
The plot shows a nonlinear relationship; an optimal range of amenity counts correlates with higher weighted review scores, but beyond a certain point, additional amenities do not significantly increase scores.

Modeling

Prices:

The pricing data for our Airbnb listings underwent a thorough analysis, and the Random Forest Regressor emerged as a powerful model for predicting prices. Configured with 500 trees and employing two variables at each split, the model demonstrated impressive performance. The mean of squared residuals, a metric assessing prediction accuracy, was notably low at 0.1562844, indicating a strong fit to the actual prices. Moreover, the model explained 60.62% of the variance in the pricing data, showcasing its robustness in capturing complex relationships.

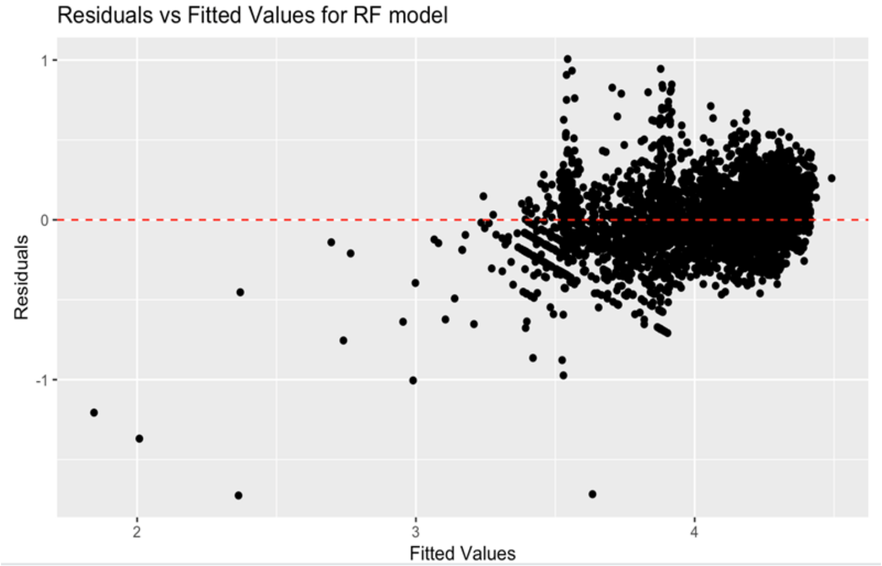
Leveraging features such as latitude, longitude, accommodates, bedrooms, bathrooms, and availability_365, the Random Forest Regressor excelled in providing accurate and reliable predictions. This model's strength lies in its ability to handle intricate patterns within the dataset, making it a valuable tool for precise pricing predictions in the dynamic landscape of Airbnb listings.



Weighted Review Scores:

We have modeled our data with a simple linear regression model , a generalized linear model (GLM), a robust linear model and a random forest model. The Random Forest Regressor operates by constructing a multitude of decision trees during training and outputs the mean prediction of the individual trees for regression tasks. It excels in handling complex relationships within data, reducing overfitting, and providing robust predictions by aggregating the outputs of multiple decision trees

Out of the three, Random forest (regressor) model demonstrated the most significant explained variance, indicating a superior fit to the data. The model explained approximately 57.08% of the variance in the **weighted review scores**, a substantial improvement over the linear approaches.

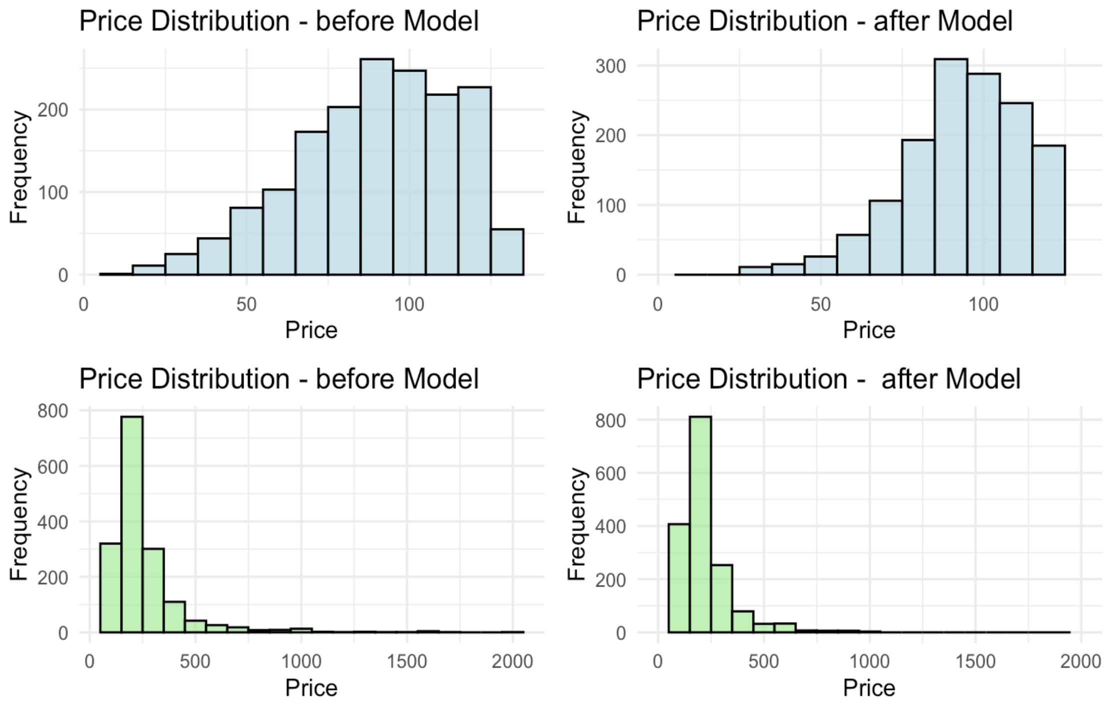


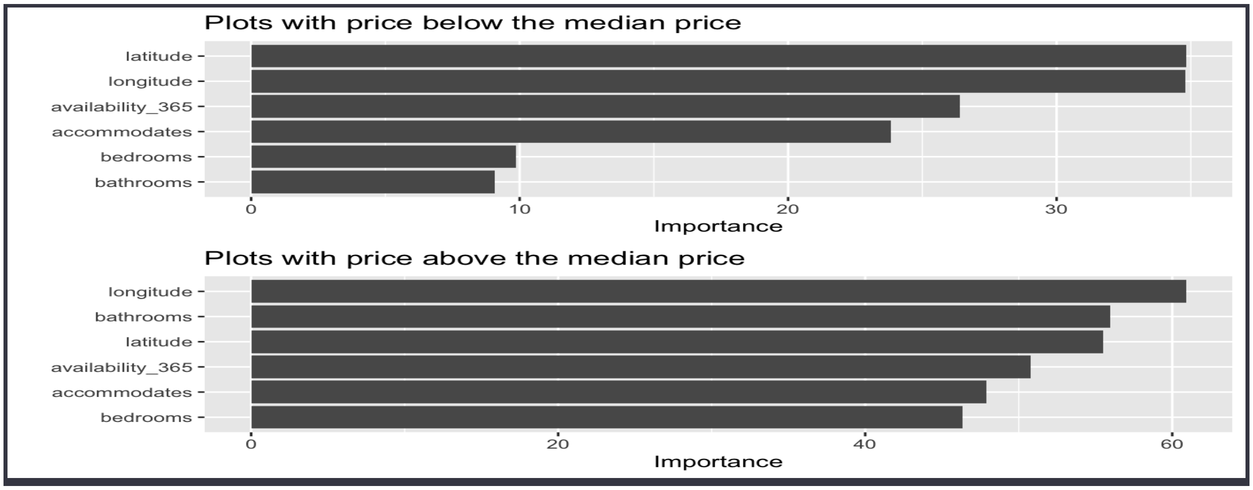
The above plot of residuals vs. fitted values for the Random Forest model revealed a random dispersion of residuals around the zero line, not showing any pattern that would suggest model inadequacies. This randomness in residuals is indicative of a well-fitting model, as it infers that the model's predictions are consistent across the range of fitted values even though there are few outliers.

Results and Observations:

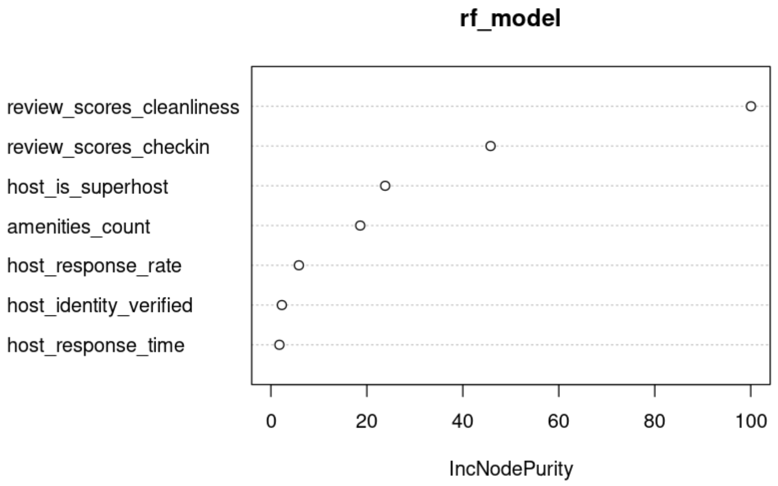
The analysis indicates that the random forest model, adept at considering factors such as location, size, and time of year, outperforms in predicting Airbnb listing prices, offering valuable insights for informed pricing strategies based on factors like accommodates, bedrooms, and bathrooms. Below are the results with comparison on the price bucketing based on the median price with the Random forest model.

Beyond the Hotel: A Deep Dive into Airbnb in Asheville, NC





Our study into the factors affecting Airbnb weighted review scores in Asheville has shown that the Random Forest model, outweighed simple linear model and GLM approaches in explanatory capacity with an explained variance of nearly 58%.

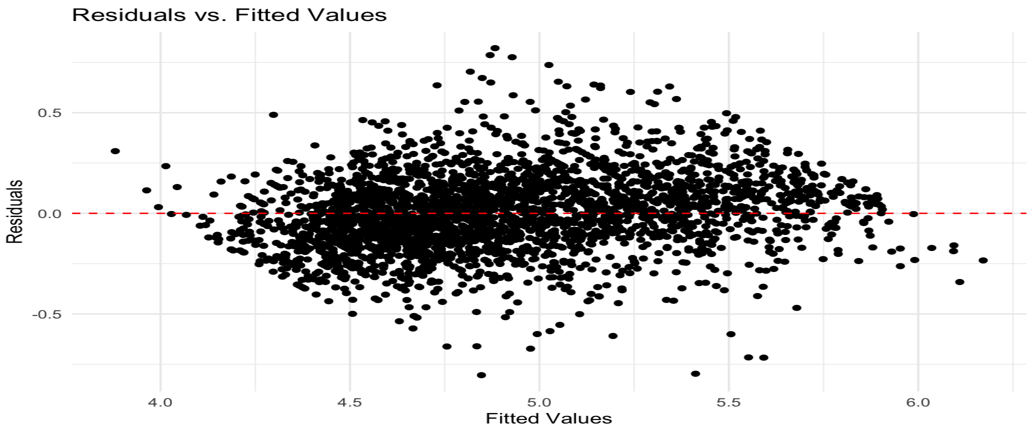


Cleanliness and **check-in** experience turned out to be the most influential factors of positive guest experience. This finding is in line with the hospitality industry standards emphasizing first impressions and core amenities.

The **superhost status** also emerged as a significant variable, suggesting that Airbnb's own classification system is a reliable indicator of quality of guest experience.

Conclusion and Future Work

We have few outliers in the data, rendering some difficulty in fitting the model, as high outliers in a random forest regression model can skew predictions and introduce a positive slope to the residuals plot. To address this issue, we will look into outlier detection and removal, hyperparameter tuning to limit tree complexity and adjust for outliers, exploring other robust regression models with weighted factors, or applying further data transformations and exploring more feature engineering. Experimenting with these techniques and evaluating their impact on model performance will help us to mitigate outlier influence and achieve a more accurate and robust model. Below is the same model, fitted on the data after the outliers limitations.



Beyond the Hotel: A Deep Dive into Airbnb in Asheville, NC

In our analysis of Airbnb listings in Asheville, we identified key factors influencing price of the listing and review scores. **Location (**Both Latitude and Longitude) and **Cleanliness** and **ease of check-in** emerged as pivotal elements shaping guests' perceptions. Properties situated in the north-east regions have high price listings and properties maintaining high cleanliness standards and providing a seamless check-in experience tended to garner more favorable reviews. Additionally, the status of being a **Superhost** was associated with higher review scores, underscoring the quality of judgment given by Airbnb. On the flip side, we observed that **host response time had a negative impact** on review scores, emphasizing the importance of timely communication in guest satisfaction. This initial examination suggests that a combination of tangible factors like cleanliness and intangible elements such as host status and responsiveness plays a crucial role in shaping the Airbnb experience in Asheville.

Further enhancing our analysis, we propose exploring the geographical variation in listing price and review scores to gain deeper insights into location-specific factors influencing guest satisfaction. By examining how these parameters vary across different neighborhoods or areas in Asheville, we can uncover unique patterns and preferences related to location. Additionally, to delve into the qualitative aspects of guest feedback, we recommend leveraging natural language processing (NLP) techniques to analyze the textual content of reviews that can unveil sentiments expressed by guests and extract specific feedback related to amenities, services, or the overall experience.

This advanced approach will not only complement our quantitative analysis but also provide a nuanced understanding of guests' perceptions, allowing for more comprehensive insights into the factors that contribute to positive or negative reviews in the Asheville Airbnb market.