California State University, Los Angeles

Airbnb Price Prediction and Business Insights Through R and Machine Learning

Prepared by:

Ayushi Porwal

Zalak Patel

Snehil Sarkar

Sapan Shah

CIS 5550-01: Data Mining for Business Analytics

Dr. Lusi Li

1 Dec. 2024

TABLE OF CONTENTS

INTRODUCTION

Airbnb, an abbreviation for “Airbed and Breakfast,” has revolutionized the way people find accommodations. As an online platform connecting property owners with travelers, Airbnb has transformed tourism and business travel experiences by offering unique and flexible rental options. The company itself does not own property but acts as an intermediary, earning commissions from bookings made on its platform. This innovative business model aligns with the dynamic consumer-centric trends of the 21st century, offering economic, convenient, and customizable lodging options.

For travelers, Airbnb provides a range of accommodations, from single rooms and apartments to entire homes and boutique hotels, often at a fraction of the cost of traditional hotels. For hosts, it creates an avenue to monetize their unused properties or spare rooms. However, the platform also introduces challenges for both parties.

In a competitive market like Los Angeles (LA) and its surrounding regions, understanding the factors that influence pricing, demand, and customer preferences is critical. This project addresses these challenges by analyzing publicly available Airbnb listing data. The dataset includes diverse attributes such as room types, number of reviews, review ratings, and regional segmentation (e.g., LA City, Outside of LA City, and Unincorporated Areas of LA County). The aim is to leverage historical data from the past three years to understand market trends, predict optimal pricing, and provide actionable insights for both hosts and travelers.

The research focuses on uncovering the dynamics of Airbnb’s marketplace by examining factors such as seasonality, neighborhood characteristics, traveler preferences, and pricing strategies. By using machine learning algorithms and data visualization techniques, this study aims to:

1. Predict prices for Airbnb listings based on various influencing factors, enabling hosts to optimize their strategies.
2. Identify trends in guest preferences, such as the most popular property types and travel durations.
3. Analyze booking patterns to offer recommendations that enhance customer satisfaction and host competitiveness.

By providing data-driven recommendations, this study helps bridge the gap between hosts and guests, fostering a better understanding of the market. For hosts, it offers guidance on pricing strategies to remain competitive. For guests, it aids in identifying ideal accommodations based on their preferences and budgets.

Conclusively, this project not only contributes to enhancing customer satisfaction but also provides hosts with tools to maximize their profitability. By applying insights drawn from Airbnb data, this study highlights trends that improve decision-making, ensuring a win-win outcome for all stakeholders involved in Airbnb's ecosystem.

DATASET OVERVIEW

The dataset used for this project is the Airbnb dataset.

Dataset & Code Link: https://github.com/snehilsarkar97/Airbnb-Price-Prediction-and-Business-Insights-Through-R-and-Machine-Learning

Unit of Analysis: Listing, Host, Pricing, Reviews and Neighborhood.

Here's a distribution of the columns based on the type of data:

Column Name	Type of Data	Description	Example
Host Response Time	Ordinal	Host reply speed.	"within an hour"
Host Name	Nominal	Host's name.	"Yoga Priestess"
Neighbourhood Group	Nominal	Broad area of the listing.	"Other Cities"
Neighbourhood	Nominal	Specific area of the listing.	"Santa Monica"
License	Nominal	License number.	"228269"
Host Is Superhost	Nominal	Superhost status (Yes/No).	"FALSE"
Room Type	Nominal	Type of room offered.	"Private Room"
Instant Bookable	Nominal	Allows instant booking (Yes/No).	"FALSE"
Id	Discrete	Listing ID.	2732
Host Id	Discrete	Host ID.	3041
Accommodates	Discrete	Number of guests allowed.	1

Beds	Discrete	Number of beds.	1
Minimum Nights	Discrete	Minimum nights to book.	7
Number Of Reviews	Discrete	Total reviews.	24
Number Of Reviews L30D	Discrete	Reviews in the last 30 days.	0
Number Of Reviews Ltm	Discrete	Reviews over the listing's lifetime.	0
Host Acceptance Rate	Continuous	Booking acceptance rate.	42%
Host Response Rate	Continuous	Response rate to messages.	100%
Host Since	Continuous	Hosting start date.	"9/17/2008"
Latitude	Continuous	Latitude of the listing.	34.0044
Longitude	Continuous	Longitude of the listing.	-118.48095
Price	Continuous	Price per night.	179
First Review	Continuous	Date of first review.	"6/6/2011"
Last Review	Continuous	Date of last review.	"8/21/2022"
Reviews Per Month	Continuous	Avg. reviews per month.	0.16
Review Scores Rating	Continuous	Overall rating score.	4.41
Review Scores Accuracy	Continuous	Accuracy score.	4.26
Review Scores Checkin	Continuous	Check-in score.	4.39
Review Scores Cleanliness	Continuous	Cleanliness score.	4.58
Review Scores Communication	Continuous	Communication score.	4.48
Review Scores Location	Continuous	Location score.	4.91
Review Scores Value	Continuous	Value score.	4.22

TEAM CONTRIBUTIONS

- **Data Cleaning, Summary Statistics, and Predictive Modeling**

Jointly performed by the team, covering data preprocessing, exploratory analysis, and predictive modeling tasks to clean and prepare the dataset for further analysis.

- **Listing Analysis**

Sapan: Focused on exploring different property types, visualizing their distributions, and identifying trends in listings across various categories.

- **Host & Neighborhood Analysis**

Zalak: Analyzed Superhosts, explored neighborhood distribution patterns, and visualized the variation in listing types across different regions.

- **Pricing Analysis**

Snehil: Conducted an analysis on price trends across various room types and neighborhoods, identifying factors influencing price variations.

- **Review Analysis**

Ayushi: Investigated the relationship between review scores and various factors such as cleanliness, check-in process, and accuracy of listing descriptions, using correlation matrices and visualizations.

ANALYTICAL TOOLS

For this project, the analysis was carried out using **R Studio**, leveraging the following packages to perform data processing, visualization, and predictive modeling:

1. **Data Handling**:

   - **readxl**: This package was used to import Excel files into R, enabling seamless access to structured datasets for analysis.
   - **tidyverse**: A collection of tools such as dplyr for data manipulation and ggplot2 for visualization, used extensively to clean, transform, and summarize the data.

2. **Visualization**:

   - **leaflet:** Used to create interactive maps for geospatial analysis, helping visualize location-based data trends effectively.
   - **corrplot:** Assisted in generating correlation plots to identify and interpret relationships among numerical variables.

3. **Predictive Modeling**:

- **broom:** Utilized for tidying up model outputs, making it easier to interpret the results of predictive models.
- **caret, random forest, rattle, rpart:** Employed for building and evaluating machine learning models, including Linear Regression, Decision Trees, and Random Forest, to identify patterns and predict outcomes.

These tools allowed for efficient analysis and ensured the delivery of meaningful insights through robust modeling and visualization techniques.

DATA CLEANING

1. Renaming Columns

The original column names were inconsistent, with spaces and mixed capitalization, making them harder to work with in R. Renaming columns to a standardized format (snake_case) improved code readability and usability. For example, Host Acceptance Rate became host_acceptance_rate. This step ensured consistent column naming, simplified coding, and set a clean foundation for analysis.

Before:

```
 [1] "Id"                          "Host Id"
 [3] "Host Name"                   "Host Is Superhost"
 [5] "Host Acceptance Rate"        "Host Response Rate"
 [7] "Host Response Time"          "Host Since"
 [9] "Neighbourhood Group"         "Neighbourhood"
[11] "Latitude"                    "Longitude"
[13] "Room Type"                   "Accommodates"
[15] "Beds"                        "Price"
[17] "Instant Bookable"            "First Review"
[19] "Last Review"                 "License"
[21] "Reviews Per Month"           "Minimum Nights"
[23] "Number Of Reviews"           "Number Of Reviews L30D"
[25] "Number Of Reviews Ltm"       "Review Scores Rating"
[27] "Review Scores Accuracy"      "Review Scores Checkin"
[29] "Review Scores Cleanliness"   "Review Scores Communication"
[31] "Review Scores Location"      "Review Scores Value"
```

After:

```
 [1] "id"                          "host_id"
 [3] "host_name"                   "host_is_superhost"
 [5] "host_acceptance_rate"        "host_response_rate"
 [7] "host_response_time"          "host_since"
 [9] "neighbourhood_group"         "neighbourhood"
[11] "latitude"                    "longitude"
[13] "room_type"                   "accommodates"
[15] "beds"                        "price"
[17] "instant_bookable"            "first_review"
[19] "last_review"                 "license"
[21] "reviews_per_month"           "minimum_nights"
[23] "number_of_reviews"           "number_of_reviews_l30d"
[25] "number_of_reviews_ltm"       "review_scores_rating"
[27] "review_scores_accuracy"      "review_scores_checkin"
[29] "review_scores_cleanliness"   "review_scores_communication"
[31] "review_scores_location"      "review_scores_value"
```

2. Handling Missing Review Scores

Missing review scores were imputed with the average value of their respective columns. This approach preserved the dataset's integrity, ensuring no rows were discarded due to missing data. By filling in missing values with the column averages, the analysis remained consistent and reliable, avoiding potential bias or distortion in model results. This method ensured a complete dataset, critical for producing accurate insights and maintaining the robustness of subsequent analyses.

Before:

```
                    last_review                         license
                              0                           22704
             reviews_per_month                  minimum_nights
                              0                               0
              number_of_reviews          number_of_reviews_l30d
                              0                               0
           number_of_reviews_ltm            review_scores_rating
                              0                               0
       review_scores_accuracy           review_scores_checkin
                             12                              20
  review_scores_cleanliness review_scores_communication
                             13                              13
       review_scores_location            review_scores_value
                             22                              26
```

After:

```
                    last_review                         license
                              0                           22704
             reviews_per_month                  minimum_nights
                              0                               0
              number_of_reviews          number_of_reviews_l30d
                              0                               0
           number_of_reviews_ltm            review_scores_rating
                              0                               0
       review_scores_accuracy           review_scores_checkin
                              0                               0
  review_scores_cleanliness review_scores_communication
                              0                               0
       review_scores_location            review_scores_value
                              0                               0
```

3. **Addressing Outliers in Pricing**

The `price` column had extreme values that could distort analysis, such as listings priced unusually high or low. These outliers were removed using the IQR method, which calculates boundaries to identify and filter such values. By removing these outliers, the dataset became more representative of typical pricing, improving the accuracy of insights.

Before:

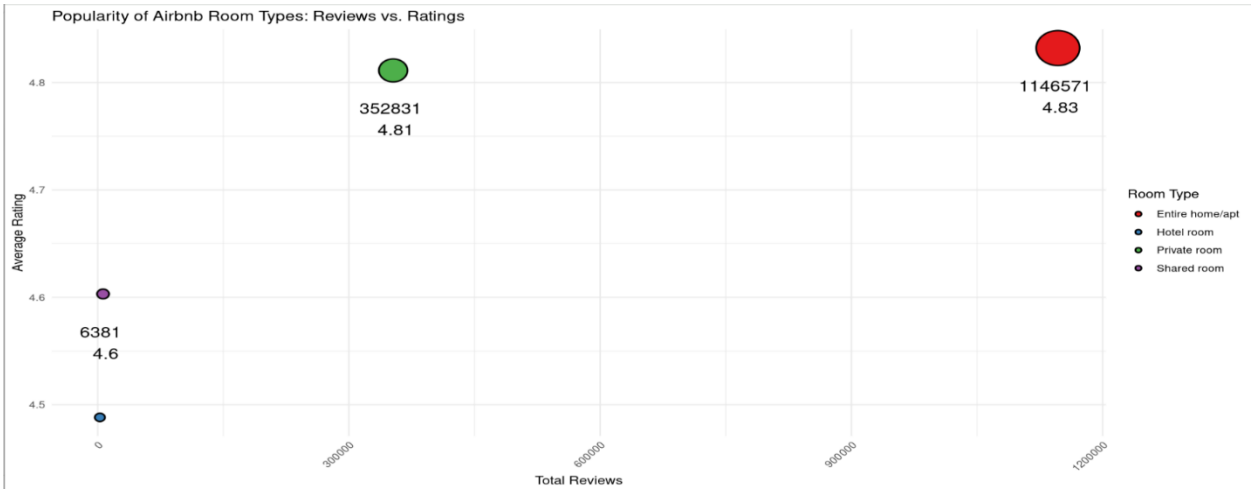
```
# A tibble: 4 × 6
  room_type       Average_Price Median_Price Min_Price Max_Price SD_Price
  <chr>                   <dbl>        <dbl>     <dbl>     <dbl>    <dbl>
1 Entire home/apt         268.           170         5     99999     777.
2 Hotel room              798.           100.       22      9999    2439.
3 Private room            118.            69        10     99999    1204.
4 Shared room              53.7           35        12      1200      95.3
> |
```

After:

```
# A tibble: 4 × 6
  room_type       Average_Price Median_Price Min_Price Max_Price SD_Price
  <chr>                   <dbl>        <dbl>     <dbl>     <dbl>    <dbl>
1 Entire home/apt         176.           155         5       440     85.0
2 Hotel room              155.            82        22       413    105.
3 Private room             84.2           69        10       440     54.8
4 Shared room              42.3           35        12       325     30.6
```

DATA ANALYSIS RESULTS & INTERPRETATION

Analysis Question 1: Which type of Airbnb properties garner the most reviews, indicating popularity?

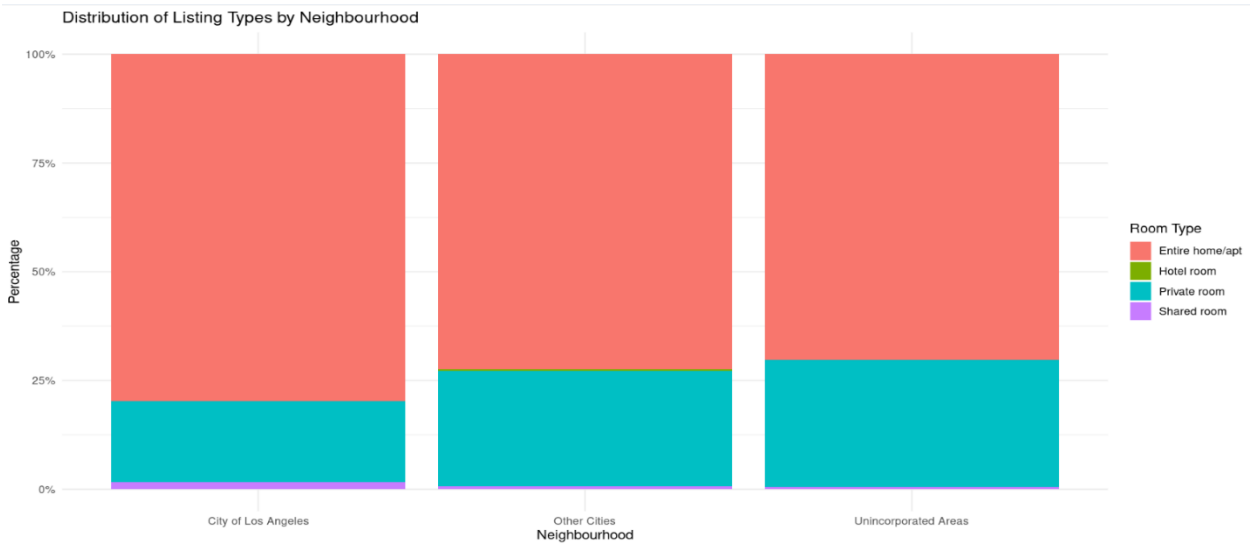


The plot visualizes the popularity of different Airbnb room types based on the total number of reviews and average rating.

Entire home/apt is the most popular room type, with a high number of reviews (1,146,571) and a very good average rating (4.83). Private rooms also have a significant number of reviews (352,831) and a good average rating (4.81). Hotel room and Shared room have fewer reviews but still maintain decent average ratings.

Overall, the plot suggests that entire homes/apartments and private rooms on Airbnb are almost equally popular by considering their average ratings.

Analysis Question 2: How does the distribution of listing types vary across different neighborhoods or regions?



This stacked bar plot shows the distribution of different Airbnb listing types (e.g., entire home, private room, shared room, hotel) across various regions, with each bar representing a region. Since it's a percentage-based (100%) stack, the heights of each colored segment within a bar reflect the proportion of each listing type in that region, summing up to 100% per region.

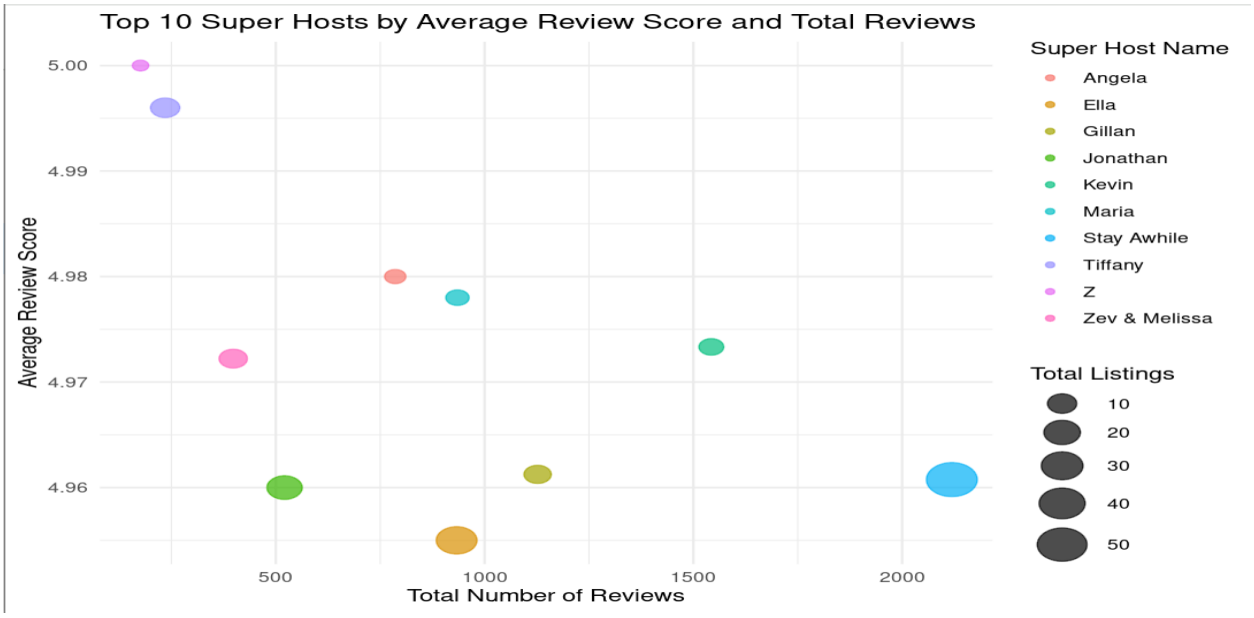
This visualization makes it easy to compare the relative availability of listing types across regions at a glance.

· **Entire home/apt** is the most dominant listing type in all three neighborhoods, comprising over 75% of the listings in each.

· **Private room** listings follow, making up around 20-25% of the total listings in each neighborhood.

· **Hotel room** and **Shared room** listings are relatively rare, accounting for less than 10% of the listings in each neighborhood.

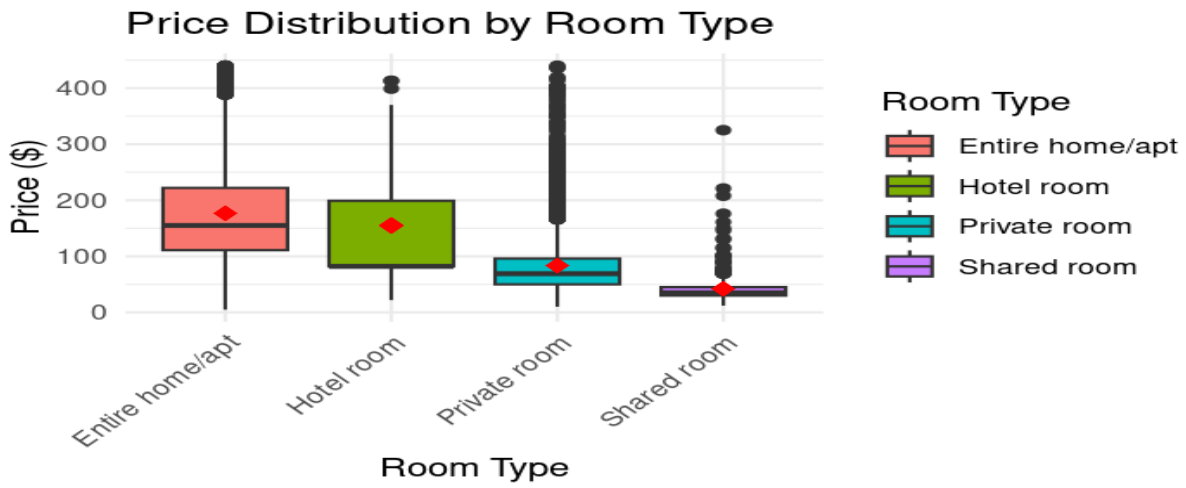
Analysis Question 3: Who are the top 10 Super hosts based on listings, review scores, and number of reviews? How do their listing and review score distributions vary?



Interpretation:

This scatter plot showcases the top 10 Super hosts based on their average review score, total number of reviews, and the number of listings they have. All Super hosts boast exceptionally high average review scores, ranging from 4.96 to 5.00, indicating consistent guest satisfaction with their accommodations and services. The number of reviews varies widely, from a few hundred to over 2,000, indicating differing levels of experience and popularity. Hosts with more reviews and listings tends to have slightly lower review scores, while those with fewer listings generally maintain higher review scores, possibly due to the challenges of maintaining consistency across a larger number of properties and guests. In contrast, hosts with fewer listings can focus more on providing personalized experiences, leading to higher review scores.

Analysis Question 4: What is the overall price trend for different room types on Airbnb?

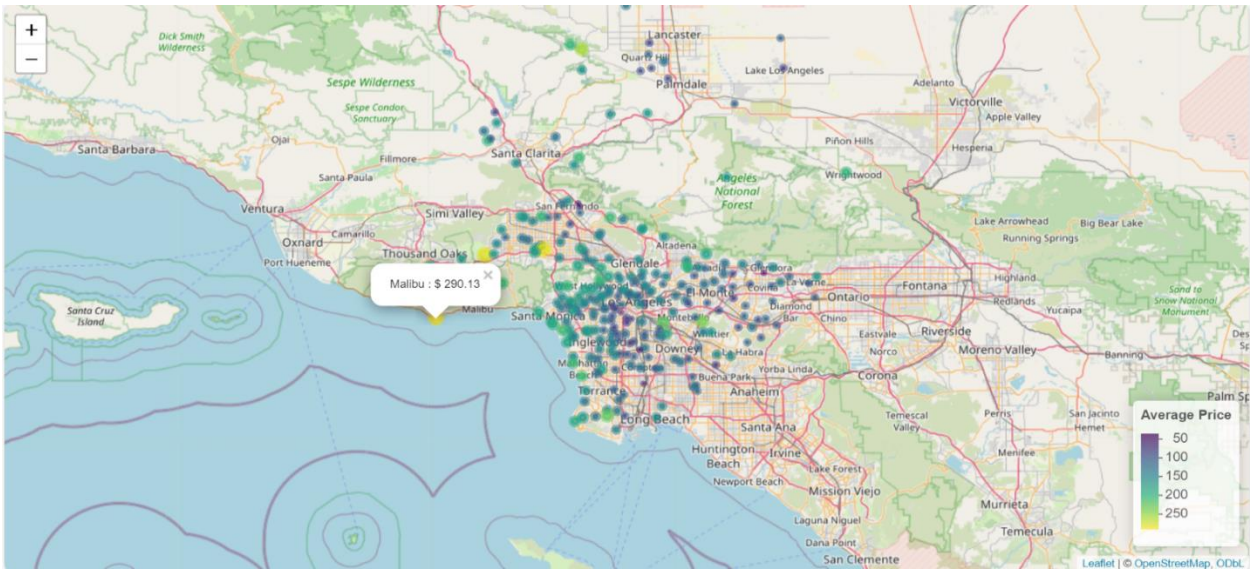


Interpretation:

The box plot provides a clear visual representation of the price distribution for different room types on Airbnb. We can observe that entire homes/apartments have the highest median price, with some outliers exceeding $400 per night. While the mean price for this category is also relatively high, it is lower than the median, suggesting the presence of a few very expensive listings. Hotel rooms have a more consistent pricing structure, with the median and mean prices being similar. Private rooms offer a more affordable option, with the median and mean prices significantly lower than entire homes/apartments and hotel rooms. Shared rooms are the most budget-friendly option, with the lowest median and mean prices.

Overall, the plot emphasizes the significant price disparities between room types on Airbnb. Guests can select the room type that aligns best with their budget and preferences**.**

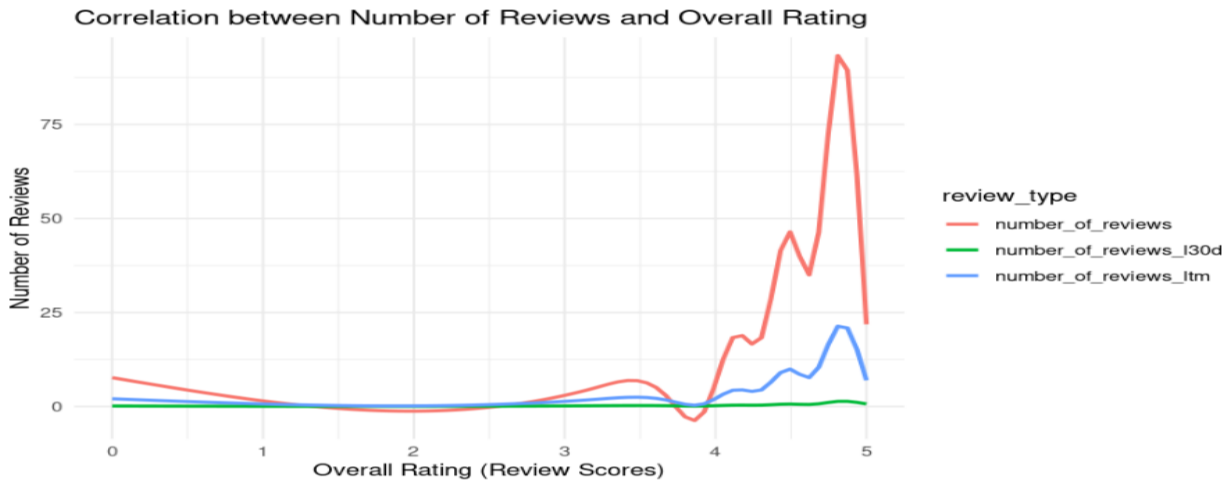
Analysis Question 5: How does the average price of Airbnb listings vary across different neighborhoods in Los Angeles?



Interpretation:

The analysis reveals a strong correlation between location and Airbnb pricing in Los Angeles. Coastal neighborhoods like Malibu, Santa Monica, and Venice Beach command premium prices due to their stunning views, beach access, and proximity to popular attractions. Conversely, inland neighborhoods, especially those further from the city center, offer lower average prices due to lower demand, fewer amenities, and reduced accessibility to tourist destinations. It also reveals that neighborhood locations like Hidden Hills with celebrity residents and places closer to national parks consistently rank among the most expensive areas in Los Angeles County. Understanding these pricing dynamics can help both hosts and guests make informed decisions regarding pricing strategies and accommodation choices.

Analysis Question 6: Is there a correlation between the number of reviews and overall ratings? Do hosts with more reviews tend to have better ratings?



Interpretation:

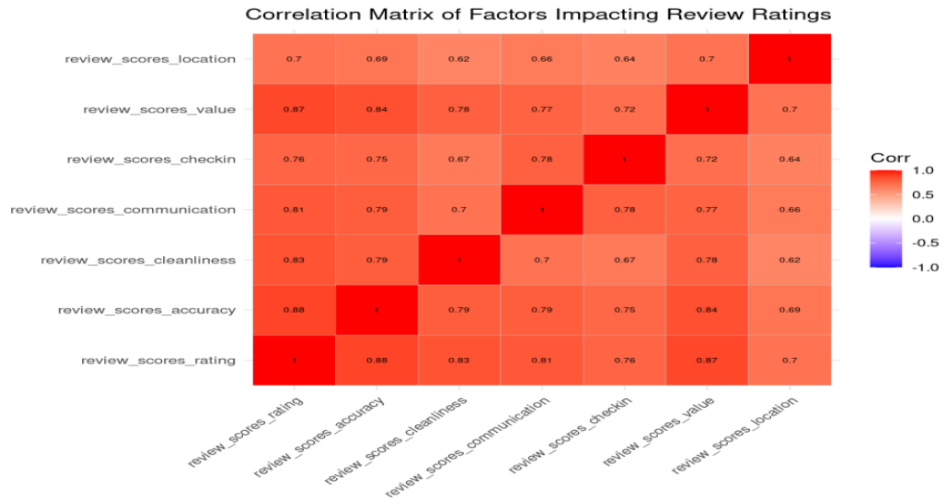
This plot shows the relationship between overall rating and number of reviews across different timeframes (total, last 30 days, and last 12 months).

Trend: A positive slope indicates that higher ratings correlate with more reviews in that timeframe, suggesting that well-rated hosts get more reviews.

Comparison: By comparing the slopes for different review periods (overall, last 30 days, last 12 months), you can see if the recency of reviews influences the relationship with rating.

Relationship Strength: A consistently strong positive trend across all timeframes suggests that better-rated hosts generally receive more reviews, regardless of review timing.

Analysis Question 7: Which factors, such as the check-in process, cleanliness, accuracy of listing descriptions, etc., most significantly impact review ratings?



Interpretation:

The correlation matrix and visualization generated by corrplot help identify which factors are most closely associated with the overall review score rating.

Review Scores Correlation: Higher correlations between review_scores_rating and factors like review_scores_accuracy and review_scores_value suggest these aspects significantly impact guest satisfaction.

Significant Factors: Stronger correlations indicate factors with a higher influence on overall ratings.

Visualization: Darker or more intense shading shows stronger correlations, making it easy to spot key influences on review_scores_rating.

SUMMARY STATISTIC

Summary Statistics for Listing Capacity across Room Type

	room_type	mean	median	min	max	sd	mode	Inter_quertile	count
1	Entire home/apt	4.660515	4	1	16	2.8073463	2	4	24493
2	Hotel room	2.338710	2	1	6	0.9222842	2	0	62
3	Private room	1.989907	2	1	16	1.1148208	2	1	7530
4	Shared room	2.236264	1	1	16	2.3270368	1	1	364

The table above provides a statistical summary of listing capacities across various room types.

Mean: Entire homes average 4.66, suggesting on average, around 5 people can be accommodated, while hotel rooms average indicate about 3 people can be accommodated. Private and shared rooms average 1.99 and 2.24, respectively, accommodating around 2 people.

Median: The median capacity is 4 for Entire homes or apartments, indicating half of these listings can accommodate 4 or fewer people. Similarly, hotel and private rooms 2 people or fewer, and shared rooms 1 person or fewer.

Mode: The mode is 2 for entire homes, hotel rooms, and private rooms, indicating most of these listings accommodate up to 2 people. For shared rooms, the mode is 1, indicating most shared rooms accommodate up to 1 person.

Spread: Entire homes show the largest variation (SD of 2.80, IQR of 4), while hotel rooms have an IQR of 0, suggesting very little variation in the middle 50% of values, implying a

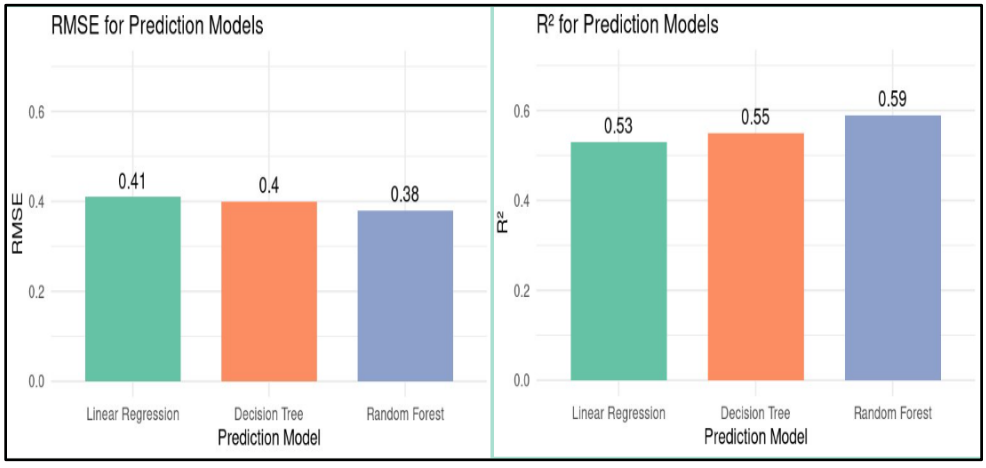
concentration around a single capacity (primarily 2 people). Private and shared rooms show moderate variation.

Overall, entire homes or apartments have more listing and offer more diverse capacities, other room types, particularly hotel and private rooms, tend to have more consistent capacities around 1 or 2 people.

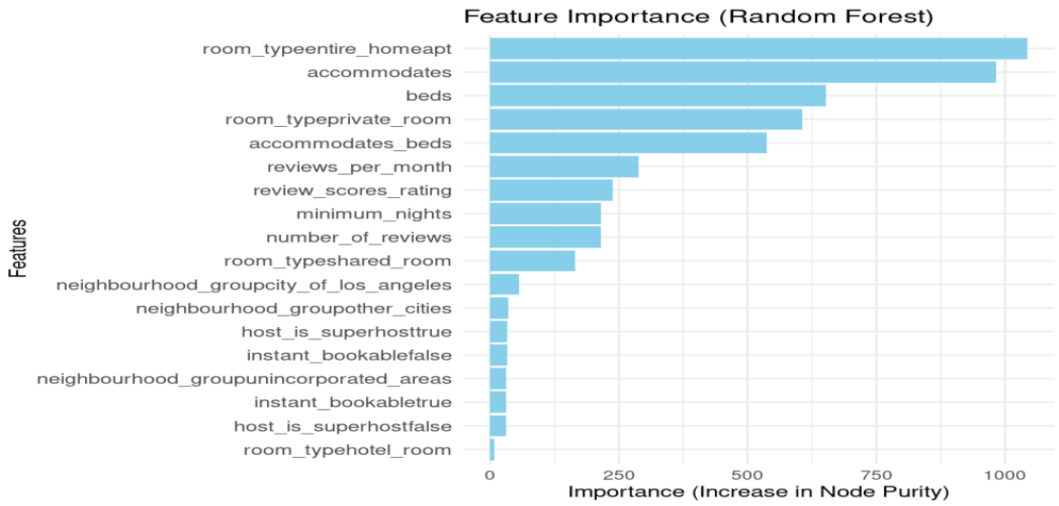
PREDICTIVE DATA MODELING

Predictive data modeling for Airbnb price prediction involves building a regression model to estimate the price of a listing based on features such as accommodates, room type, beds, reviews, and location. Initially, outliers in the price label were removed to improve model accuracy. Categorical variables like room_type, neighbourhood_group, and instant_bookable were transformed into numerical values using one-hot encoding, while numerical features such as accommodates and beds were scaled using robust scaling to handle variability. Feature engineering was applied to create meaningful interactions (e.g., accommodates × beds) and transformations to reduce skewness in the price label. The data was then split into training, validation, and test sets to support model building and evaluation.

Three regression models—Linear Regression, Decision Tree, and Random Forest Regression—were developed and evaluated using key performance metrics such as Root Mean Squared Error (RMSE) and R2 (R-squared) to assess model accuracy. Among the models, Random Forest demonstrated the highest R2 score of 0.59 and the lowest RMSE of 0.38, outperforming the other models in predictive performance.



Additionally, feature importance analysis was conducted using the Random Forest model, which identified the most influential features. This analysis provided valuable insights for prioritizing features, enabling the addition or removal of variables to enhance the model's accuracy and effectiveness.



Overfitting checks were performed to assess the robustness of the models. The Decision Tree model exhibited consistent R2 and RMSE values across both training and test datasets, indicating a balanced fit without significant overfitting. In contrast, the Random Forest model showed slight overfitting, with a training accuracy (R2 = 0.70) notably higher than the test accuracy (R2 = 0.59). Although hyperparameter tuning methods, such as parameter grids and cross-validation, could not be applied due to RAM limitations, these techniques are expected to mitigate the overfitting issue and significantly improve the model's R2 score and predictive accuracy.

KEY FINDINGS

Key observations from the analysis reveal several important trends in Airbnb listings.

1. Entire homes are the most popular and profitable listing types, consistently performing better in terms of both occupancy and pricing.
2. Price and location also play crucial roles, with coastal areas and luxury neighborhoods commanding premium prices.

3. Super Hosts stand out by maintaining high ratings, which significantly contribute to attracting more guests.
4. Furthermore, positive reviews are a key driver of success, as they lead to more bookings and increased income for hosts.
5. Guests prioritize accurate listings and good value for money. Hosts who provide detailed descriptions and fair pricing are more likely to receive positive reviews.
6. In terms of predictive modeling, the Random Forest price prediction model outperforms others, achieving the lowest RMSE (0.38) and the highest R2 (0.59), demonstrating its superior accuracy in estimating listing prices.

LIMITATIONS

Missing Data Challenges:

1. High prevalence of missing values in key metrics such as host_response_rate and host_acceptance_rate impacted the ability to analyze host behavior comprehensively.
2. Missing booking dates hindered the ability to explore seasonal trends in pricing and demand effectively.

Model Constraints:

1. The Random Forest model showed slight overfitting due to limited hyperparameter tuning capabilities caused by RStudio's RAM constraints.
2. Inability to perform cross-validation or advanced optimization techniques restricted the model's predictive performance.

Dataset Limitations:

1. The dataset is static and does not account for temporal trends or shifts in market dynamics, such as changes in demand over time or external factors like economic conditions.
2. Sparse representation of certain listing types (e.g., shared rooms, hotels) limited insights for less popular categories.

Generalization:

1. Analysis focused on Los Angeles Airbnb listings, which may limit the generalizability of findings to other cities or regions with different market dynamics

FUTURE DIRECTIONS

Improved Data Collection:

1. Incorporate more dynamic datasets, including time-series data for seasonal trend analysis and long-term demand forecasting.
2. Address missing data by integrating external data sources or adopting advanced imputation techniques such as multiple imputation.

Enhanced Modeling Techniques:

1. Explore advanced machine learning models like Gradient Boosting (e.g., XGBoost, LightGBM) to improve predictive accuracy.
2. Perform hyperparameter tuning using grid search or random search to optimize model performance.

Feature Engineering:

1. Develop more nuanced interaction features, such as dynamic pricing adjustments based on demand trends or competitor pricing.
2. Include external factors like local events, weather patterns, and economic indicators to enrich the predictive model.

Geospatial and Temporal Insights:

1. Leverage GIS tools for detailed spatial analysis, identifying hyper-local pricing trends.
2. Examine temporal shifts in guest preferences and demand patterns to guide adaptive pricing strategies.

Scalability and Automation:

1. Implement scalable workflows for continuous analysis, enabling real-time insights and updates.

2. Automate data pipelines for periodic updates, improving efficiency and relevance of recommendations.

WORKS CITED

Snowflake. "Data Mining for Business Analytics." *Snowflake*, 2023, https://www.snowflake.com/trending/data-mining-for-business-analytics/. Accessed 1 Nov. 2024.

Kumari, Archana, and Mohan Kumar. S. "Dynamic Pricing: Trends, Challenges and New Frontiers." 2024 IEEE International Conference on Contemporary Computing and Communications (InC4), vol. 1, 2024, pp. 1-7. Accessed 1 Nov. 2024.

Lektorov, Alexander, Eman Abdelfattah, and Shreehar Joshi. "Airbnb Rental Price Prediction Using Machine Learning Models." 2023 IEEE 13th Annual Computing Conference, IEEE, 2023. Accessed 14 Nov. 2024.

Jasleen D., Nandana & others. "Analysis of Airbnb Prices using Machine Learning Techniques" 2021 IEEE 11th Annual Computing and Communication workshop and Conference (CCWC).