# Analyzing the Impact of Reviews on Airbnb Booking Rates

Edith Tache, Cali Hugelen, Jacob Brooks, Remington Webb

Summer, 2024

Table of Contents

# Introduction

# Business Problem/Hypothesis

# Methods/Analysis

# Results

# Recommendations/Ethical Considerations

Limitations and Future Work

# Conclusion

# References

# Appendix

Introduction

In today's highly competitive hospitality industry, customer feedback is a crucial determinant of a listing's success on platforms like Airbnb. Reviews significantly influence booking rates, as potential guests often rely on previous experiences shared by others to make their decisions. Understanding how these reviews impact booking behavior is essential for hosts looking to optimize their listings and attract more guests.

This project aims to analyze the impact of reviews on Airbnb booking rates in California. Using a dataset sourced from Kaggle, we will explore the relationship between number of reviews, review scores and booking rates. Our analysis will involve several steps, including data cleaning and preparation, exploratory data analysis (EDA), and building a predictive model. By employing various statistical and machine learning techniques, we aim to uncover the key factors in reviews that contribute to higher booking rates and provide actionable insights for hosts to enhance their listings' appeal and performance.

This research will not only shed light on the significant predictors of booking success but also guide hosts in implementing data-driven strategies to improve their overall booking rates and customer satisfaction.

**Dataset Sourcing:** The dataset being examined is from Kaggle, titled "AirBnB Listings in California."

**Link to Dataset:** [AirBnB Listings in California (kaggle.com)](https://www.kaggle.com/datasets/setseries/airbnb-listings-in-california)

**Description:** The dataset includes various attributes crucial for analyzing the impact of reviews on booking rates. Relevant columns include:

* **price** - The price of the listing.
* **availability\_365** - The number of available days within the 365 days (one year) time periods
* **number\_of\_reviews** - Total number of reviews the listing has received.
* **number\_of\_reviews\_ltm** - Number of reviews in the last 12 months.
* **number\_of\_reviews\_l30d** - Number of reviews in the last 30 days.
* **review\_scores\_rating** - Overall rating score.
* **review\_scores\_accuracy** - Rating score for accuracy.
* **review\_scores\_cleanliness** - Rating score for cleanliness.
* **review\_scores\_checkin** - Rating score for check-in.
* **review\_scores\_communication** - Rating score for communication.
* **review\_scores\_location** - Rating score for location.
* **review\_scores\_value** - Rating score for value.
* **reviews\_per\_month** - Average number of reviews per month.
* **host\_response\_time** - How quickly the host responds to inquiries.
* **host\_response\_rate** - Host's response rate.
* **host\_acceptance\_rate** - Host's acceptance rate.
* **host\_is\_superhost** - Whether the host is a superhost.
* **instant\_bookable** - Whether the listing is instantly bookable.
* **room\_type** - Type of room (e.g., Entire home/apt, Private room).
* **minimum\_nights** - Minimum number of nights a guest can stay.
* **maximum\_nights** - Maximum number of nights a guest can stay.
* **accommodates** - Number of guests the listing accommodates.

# Business Problem/Hypothesis

The core of our project revolves around understanding how customer reviews affect booking rates. Specifically, we seek to answer the following research questions:

1. How do the number of reviews impact booking rates for Airbnb listings in California?
2. How do the quality of reviews impact booking rates for Airbnb listings in California?
3. What are the key factors in reviews that contribute to higher booking rates?

#### By focusing on these questions, we aim to provide actionable insights for Airbnb hosts to optimize their listings. Understanding the influence of reviews will enable hosts to enhance their properties' appeal, thereby improving booking rates and customer satisfaction. Additionally, we will develop a predictive model for booking rates, helping hosts make data-driven decisions to maximize their success on the platform.

#### **Hypothesis**

Based on preliminary research, we hypothesize that:

1. The number of reviews significantly influence booking rates.
2. Higher review scores correlate with higher booking rates

These hypotheses will guide our analysis and model development, helping us identify the critical factors influencing booking rates and provide actionable recommendations for optimizing Airbnb listings to enhance their appeal and booking performance.

# Methods/Analysis

**Data Cleaning and Preparation**

* **Dropped unnecessary columns:** Focused on relevant columns like price, number\_of\_reviews, review scores, etc.,.
* **Filled missing values:** Used the median for numerical columns and the mode for categorical columns.
* **Created new features:** Calculated the booking rate
  + **Calculated booked\_days\_365**: Derived by subtracting availability\_365 from 365.
  + **Calculated Booking Rate**: Using the formula: Booking Rate = booked\_days\_365 /365

**Exploratory Data Analysis (EDA)**

EDA was conducted to understand the distributions and relationships within the data, guiding the subsequent steps in feature engineering, model building, and evaluation.

**Outliers**:

* **Identification and Handling**: Outliers were identified and handled by removing rows with review scores of 0. This step ensures the integrity of the data, as review scores of 0 may indicate missing or incorrect entries.
* **Thresholds for Minimum and Maximum Nights**: To handle extreme values, thresholds were set for minimum\_nights (capped at 30) and maximum\_nights (capped at 365). This prevents the model from being influenced by unusually high or low values that do not represent typical booking patterns.

**Relationships Between the Number of Reviews and Booking Rates**:

* A box plot was used to visualize booking rate distribution across different review count ranges. This helped understand how the number of reviews impacts booking rates and identify any trends or patterns.

**Relationships Between Review Scores and Booking Rates**:

* **Scatter Plot**: Plotted the overall review scores against booking rates to visualize their relationship. This helped identify whether higher review scores correlate with higher booking rates.
* **Line Plot of Average Booking Rate vs. Review Scores**: This provided a clearer view of how average booking rates change with different review scores.

**Key Factors in Reviews That Contribute to Higher Booking Rates**:

* Analyzed the relationship between individual review scores (e.g., accuracy, cleanliness, check-in, communication, location, value) and booking rate using scatter plots. This helped pinpoint specific review aspects that may influence booking decisions.
* **Correlation Analysis**: Created a heatmap to show the correlation between individual review scores and booking rates. This visual representation helps to quickly identify which review factors are most strongly associated with booking rates.

This thorough EDA provided valuable insights into the relationships between review characteristics and booking rates, guiding subsequent steps in feature engineering, model building, and evaluation.

**Feature Engineering**

* **Encoded Categorical Variables**: Converted categorical variables to numerical format using one-hot encoding.
* **Normalized Features**: Applied standardization to numerical features to ensure uniform scaling.

**Model Training and Evaluation**

**Model Selection**

**Random Forest Regressor**: The Random Forest Regressor was chosen for its robustness to overfitting and its ability to handle complex interactions between features. Additionally, it provides feature importance scores, which helped understand the impact of different variables on the target outcome, the booking rate.

**XGBoost Regressor**: XGBoost was also utilized due to its efficiency and performance, particularly in handling large datasets. It offers strong predictive power and the ability to tune various hyperparameters for optimal performance.

**Model Training**

Both models were trained to predict the continuous booking rate values.

**For the Random Forest Regressor:**

* The model was trained using 100 estimators, and its performance was evaluated on the test set.

**For the XGBoost Regressor:**

* The model was trained with 100 estimators, a max depth of 3, and a learning rate of 0.1.

**Hyperparameter Tuning**

**Random Forest**:

* Hyperparameter tuning was performed using RandomizedSearchCV, exploring various combinations such as the number of estimators, max depth, min samples split, min samples leaf, and max features ('sqrt' and 'log2'). The best hyperparameters were identified and used to retrain the model.

**XGBoost**:

* Hyperparameter tuning was conducted using GridSearchCV, exploring parameters like the number of estimators, max depth, learning rate, subsample, and colsample\_bytree. The best hyperparameters were identified and applied to the model.

**Evaluation Metrics**

To evaluate the models, the following metrics were used:

* **Mean Squared Error (MSE)**: Measures the average squared difference between the actual and predicted values. A lower MSE indicates a better fit.
* **R² Score**: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R² score signifies better predictive performance.

**Feature Importance**

* Feature importance scores were calculated, providing insights into which features had the most significant impact on predicting the booking rate. These scores were visualized in a bar plot.

# Results

**Model Performance**

**Random Forest Regressor Model:** The Random Forest Regressor model was the primary focus due to its known robustness and ability to handle complex data interactions. Initially, the model exhibited overfitting with a near-perfect Mean Squared Error (MSE) and R² score of 2.73×10^-30 and an R² score of 1.0, respectively. To address this, the feature names and the shape of the training data were examined to ensure a proper setup for model training. The features availability\_365 and booked\_days\_365 were then dropped from the dataset since they directly relate to the calculation of the target variable.

After addressing overfitting, the model's performance was recalibrated.

* **Before Hyperparameter Tuning:**
  + **Mean Squared Error:** 0.0778
  + **R² Score:** 0.4356

The initial evaluation indicated reasonable performance, but further improvements were sought through hyperparameter tuning.

* **After Hyperparameter Tuning:**
  + **Best Parameters:** {'max\_depth': None, 'max\_features': 'log2', 'min\_samples\_leaf': 1, 'min\_samples\_split': 3, 'n\_estimators': 137}
  + **Mean Squared Error:** 0.0794
  + **R² Score:** 0.4238

**XGBoost Regressor Model:** The XGBoost model was tested as an alternative to assess whether it could provide better predictive performance.

* **Before Hyperparameter Tuning:**
  + **Mean Squared Error:** 0.1115
  + **R² Score:** 0.1913
* **After Hyperparameter Tuning:**
  + **Best Parameters:** {'colsample\_bytree': 1.0, 'learning\_rate': 0.1, 'max\_depth': 5, 'n\_estimators': 100, 'subsample': 0.8}
  + **Mean Squared Error:** 0.1050
  + **R² Score:** 0.2380

**Comparative Analysis: Random Forest Regressor vs. XGBoost Regressor**

1. **Superior Model Performance:**
   * The Random Forest Regressor model Demonstrated better performance with higher R² scores compared to XGBoost Regressor model, suggesting it captured more variance in booking rates. However, both models indicated that room for improvement remains, as the R² scores were moderate, highlighting the complexity of accurately predicting booking rates based solely on the features available.
2. **Hyperparameter Tuning Outcomes:**
   * Hyperparameter tuning had a notable impact, particularly for XGBoost, where the R² score improved from 0.1913 to 0.2380, indicating better model fitting. However, the Random Forest model experienced a slight decline post-tuning, suggesting the possibility of slight overfitting even after adjustments.
3. **Key Findings:**
   * The Random Forest model consistently outperformed XGBoost in both pre-tuning and post-tuning phases, underscoring its robustness and ability to handle the dataset's complexity. The differences in MSE and R² scores suggest that while Random Forest is a more reliable model for this task, significant challenges remain in achieving highly accurate predictions, likely due to data quality and the nuanced nature of the factors affecting booking rates.

**Analysis of the Relationship Between the Number of Reviews and Booking Rate**

The analysis utilized a box plot to explore the distribution of booking rates across different ranges of review counts. As depicted in the box plot, there is no clear trend indicating that a higher number of reviews consistently leads to a higher booking rate. Instead, booking rates exhibit a wide range of values across all review count categories, suggesting that the sheer volume of reviews alone does not directly correlate with booking rates.

A graph of blue rectangular objects

Description automatically generatedFigure 1 Booking Rate Distribution for Different Review Count Ranges

**Analysis of the Relationship Between the Quality of Reviews and Booking Rate**

To examine the impact of review quality on booking rates, a scatter plot was generated, illustrating the relationship between overall review scores and booking rates. Additionally, a line plot was used to depict the average booking rate as a function of overall review scores. The scatter plot reveals a weak positive relationship, indicating that higher review scores slightly correlate with higher booking rates. The line plot further supports this, showing an upward trend in average booking rates as review scores improve. However, the variability within the data suggests that factors other than review scores also play significant roles in determining booking rates.

A graph of a review score

Description automatically generated with medium confidence

Figure 2: Overall Review Scores vs Booking Rate

A graph of blue lines

Description automatically generated

Figure 3: Average Booking Rate vs. Overall Review Scores

**Analysis of Key Factors in Reviews That Contribute to Higher Booking Rates**

A correlation heatmap was employed to analyze the relationships between individual review aspects (e.g., accuracy, cleanliness, communication, location, value) and booking rates. The heatmap reveals generally weak correlations between these factors and the booking rate, with review\_scores\_value showing the strongest (though still modest) correlation. This indicates that while certain aspects of reviews are more influential than others, none of the review scores individually has a strong impact on booking rates. The analysis suggests that the overall impression of a listing, possibly influenced by a combination of factors, is more critical than individual review aspects in determining booking rates.

A screenshot of a graph

Description automatically generatedFigure 4: Correlation Between Review Scores and Booking Rate

These analyses collectively highlight that while the number and quality of reviews are related to booking rates, they are not the sole determinants. The weak correlations suggest that other factors, perhaps including listing attributes, pricing strategies, or marketing efforts, may play more substantial roles in influencing booking rates.

**Feature Importance Analysis**

The feature importance analysis using the Random Forest model revealed that **price** is the most influential factor affecting booking rates. Other significant factors include the **host acceptance rate**, **maximum nights allowed**, and **reviews per month**, which highlight the importance of pricing strategy, host responsiveness, and listing activity. The **number of reviews** and review scores for **value** and **cleanliness** also play roles but are less impactful.

A graph with blue and white bars

Description automatically generated

Figure 5: Feature Importance of Random Forest Model

Recommendations/Ethical Considerations

Based on the analysis, the following strategies are recommended to optimize booking rates and enhance guest experiences on Airbnb:

1. **Optimizing Pricing Strategies:**
   * **Dynamic Pricing:** Hosts should implement dynamic pricing strategies that adjust based on demand, seasonality, and competition. Since price was identified as the most significant factor influencing booking rates, optimizing pricing can directly impact the number of bookings.
   * **Discounts and Promotions:** Offering discounts for extended stays or last-minute bookings can attract more guests, especially during low-demand periods.
2. **Improving Host Acceptance Rate and Responsiveness:**
   * **Fast Response Time:** Increasing the host response rate and reducing response time can significantly enhance the likelihood of securing bookings. Prompt responses to inquiries make listings more appealing to potential guests.
   * **Flexible Cancellation Policies:** Providing flexible cancellation policies can increase booking rates by offering potential guests greater peace of mind.
3. **Enhancing Property Features:**
   * **Quality of Accommodation:** Emphasizing high cleanliness standards, accurate descriptions, and good value standards can improve review scores and attract more bookings as these review aspects are influential in guest decision-making.
4. **Leveraging Reviews and Guest Feedback:**
   * **Encouraging Positive Reviews:** Hosts should encourage guests to leave reviews and provide feedback. A high number of positive reviews can increase booking rates by building trust and credibility.
   * **Addressing Negative Feedback:** Actively addressing and resolving issues raised in negative reviews can improve overall ratings and future guest experiences.

**Ethical Considerations**

1. **Transparency and Honesty:**
   * **Accurate Listings:** Hosts must ensure that their listings are accurate and truthful. Misleading descriptions or photos can lead to negative reviews and damage reputation.
   * **Clear Communication:** Providing clear and transparent communication regarding booking policies, prices, and amenities is crucial in maintaining trust with guests.
2. **Guest Privacy and Data Security:**
   * **Data Protection:** Protecting guest data and privacy is paramount. Hosts should comply with data protection regulations and ensure that guest information is secure and not misused.
   * **Ethical Use of Reviews:** While leveraging reviews for marketing, it is important to respect guest privacy and not manipulate or misrepresent guest feedback.

Limitations and Future Work

**Limitations:**

1. **Data Quality and Completeness:** The dataset has limitations in data quality and completeness, potentially impacting the accuracy of analysis and model predictions.
2. **Generalizability:** The findings from this study, based on a specific region (California), may not be applicable to other locations with different market dynamics and guest preferences.
3. **Temporal Dynamics:** The analysis provides a snapshot based on a specific timeframe without accounting for temporal changes in market trends and guest behavior.

**Future Work:**

1. **Broader Data Collection:** Future studies should consider including data from multiple regions and a more diverse range of property types to enhance the generalizability of the findings.
2. **Dynamic Analysis:** Incorporating time-series analysis and monitoring trends over time can provide deeper insights into changing market dynamics and guest preferences.
3. **Advanced Predictive Models:** Exploring more advanced machine learning models and techniques, such as deep learning and natural language processing, can improve the accuracy of predictions and uncover nuanced insights from reviews.

By addressing these limitations and focusing on continuous improvement, Airbnb hosts can better understand the factors influencing booking rates and leverage these insights to optimize their listings, ultimately leading to higher guest satisfaction and increased bookings.

# Conclusion

This study explored the impact of Airbnb reviews on booking rates, with a focus on listings in California. The findings reveal that the number of reviews does not influence booking rates. However, the quality of reviews, as indicated by higher review scores, shows a slight positive correlation with booking rates. This suggests that while good reviews can slightly boost a listing's attractiveness, they are not the primary drivers of booking success.

Price emerged as the most significant determinant of booking rates, underscoring the importance of strategic pricing. Other critical factors include host acceptance rate and maximum nights. These elements are crucial in shaping guest perceptions and influencing their booking decisions.

The machine learning models employed, particularly the Random Forest Regressor, displayed moderate predictive capabilities. This indicates that factors beyond the dataset's scope, such as unique property features, location-specific attractions, and individual guest preferences, may also play a significant role in determining booking rates.

To achieve a comprehensive understanding and optimization of booking rates, Airbnb hosts should adopt a multi-faceted approach. This includes setting competitive prices, maintaining high property standards, ensuring prompt and effective communication with guests, and encouraging positive reviews.

Future research should focus on incorporating more diverse datasets and exploring advanced predictive techniques to better capture the complex factors affecting booking rates. By expanding on these insights, hosts can more effectively position their listings to attract a broader audience and increase their booking success.

# References

1. Valeria G. (2023, September 6). Understanding the process: How do Airbnb reviews work? Airbnb. Retrieved from [Understanding the Process: How Do Airbnb Reviews Work? (reputationsciences.com)](https://www.reputationsciences.com/how-do-airbnb-reviews-work/) Explains how Airbnb's review system works, providing insights into the review process, rating criteria, and how reviews impact listings.
2. Filieri, R., & McLeay, F. (2014). E-WOM and accommodation: An analysis of the factors that influence travelers' adoption of information from online reviews. Journal of Travel Research, 53(1), 44-57. <https://doi.org/10.1177/0047287513481274> Analyzes the factors that influence travelers' adoption of information from online reviews, highlighting the importance of reviews in decision-making.
3. Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In P. O'Connor, W. Höpken, & U. Gretzel (Eds.), Information and communication technologies in tourism 2008 (pp. 35-46). Springer. <https://doi.org/10.1007/978-3-211-77280-5_4> Discusses the use and impact of online travel reviews, emphasizing how reviews affect consumer behavior and perceptions.
4. Kaggle. (n.d.). Airbnb listings in California. Retrieved from [AirBnB Listings in California (kaggle.com)](https://www.kaggle.com/datasets/setseries/airbnb-listings-in-california) Provides the dataset used for this analysis, which includes comprehensive information on Airbnb listings in California.
5. Liu, Y., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. Tourism Management, 47, 140-151. <https://doi.org/10.1016/j.tourman.2014.09.020> Investigates the elements that make online reviews useful for consumers, offering insights into the characteristics of effective reviews.
6. Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Quarterly, 34(1), 185-200. <https://doi.org/10.2307/20721420> Explores the factors that contribute to the helpfulness of online reviews, based on a study of customer reviews on Amazon.
7. Pan, B., MacLaurin, T., & Crotts, J. C. (2007). Travel blogs and the implications for destination marketing. Journal of Travel Research, 46(1), 35-45. <https://doi.org/10.1177/0047287507302378> Examines the role of travel blogs in destination marketing, providing insights into how user-generated content influences travel decisions.
8. Racherla, P., Connolly, D. J., & Christodoulidou, N. (2013). What determines consumers’ ratings of reviews? An exploratory study of online traveler reviews. Journal of Hospitality Marketing & Management, 22(2), 135-161. Retrieved from [What Determines Consumers’ Ratings of Service Providers? An Exploratory Study of Online Traveler Reviews | Request PDF (researchgate.net)](https://www.researchgate.net/publication/254364123_What_Determines_Consumers'_Ratings_of_Service_Providers_An_Exploratory_Study_of_Online_Traveler_Reviews)Investigates the determinants of consumer ratings for online reviews, highlighting key factors that influence review evaluations.
9. Sparks, B. A., Perkins, H. E., & Buckley, R. (2013). Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. Tourism Management, 39, 1-9. <https://doi.org/10.1016/j.tourman.2013.03.007> Analyzes how different aspects of online travel reviews, such as content type and source, affect consumer behavior and decision-making.
10. Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? International Journal of Hospitality Management, 44, 120-130. <https://doi.org/10.1016/j.ijhm.2014.10.013> Uses big data and text analytics to explore hotel guest experiences and satisfaction, providing valuable insights into customer feedback

# Appendix

1. **Detailed Model Evaluation Metrics Scores**

**Random Forest Regressor**

**Before Handling Overfitting**:

* **Mean Squared Error**: 2.733634179936952e-30
* **R² Score**: 1.0

**After Handling Overfitting**:

* **Mean Squared Error**: 0.07779803799834155
* **R² Score**: 0.4356080943532512

**After Hyperparameter Tuning**:

* **Mean Squared Error**: 0.07942994122304631
* **R² Score**: 0.4237693257349241

**XGBoost Regressor**

**Before Hyperparameter Tuning**:

* **Mean Squared Error**: 0.11146918747390884
* **R² Score**: 0.19133812679652373

**After Hyperparameter Tuning**:

* **Mean Squared Error**: 0.10504300858804198
* **R² Score**: 0.23795734034916638

1. **Booking Rate Calculation**

A number of numbers and lines

Description automatically generated with medium confidence