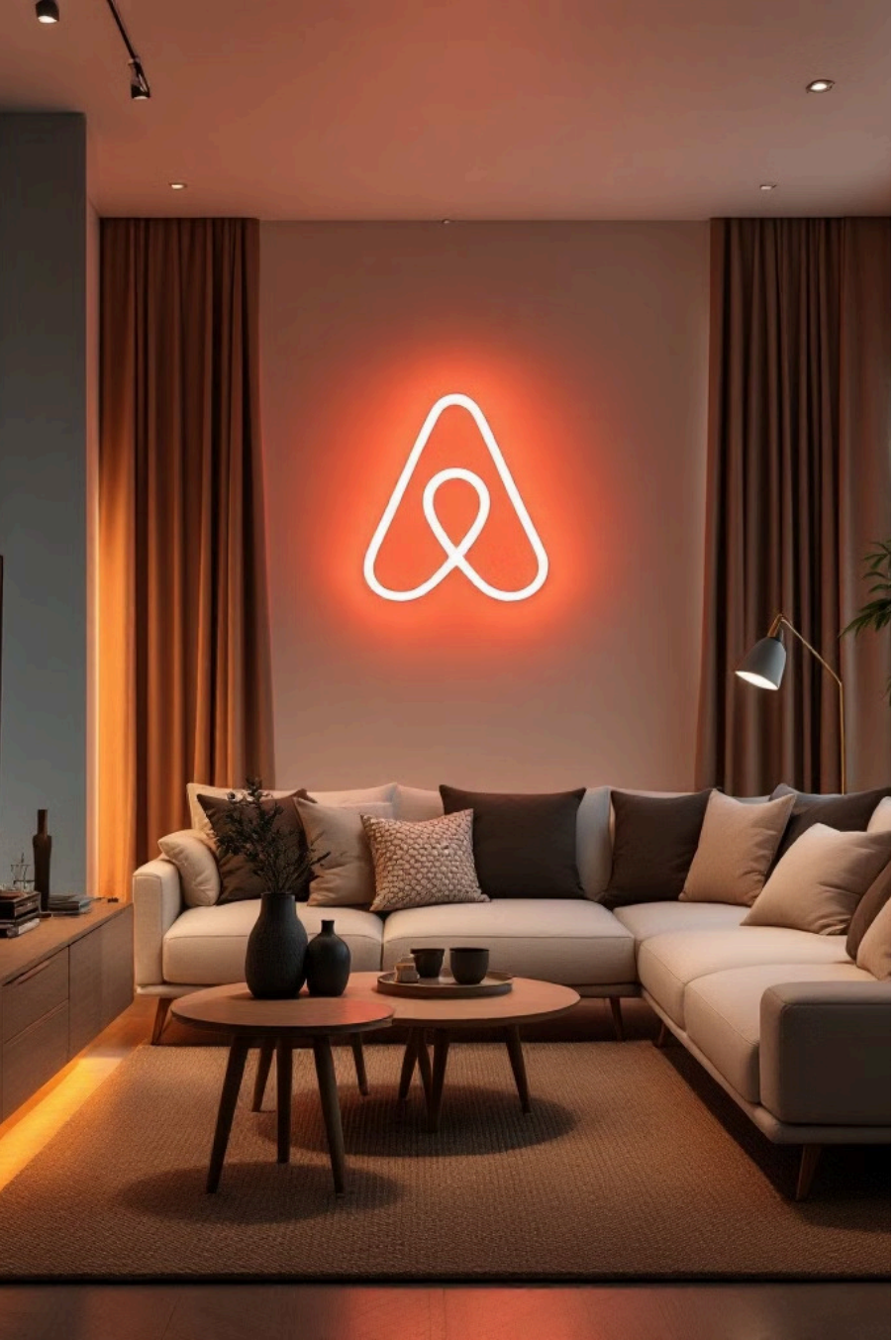
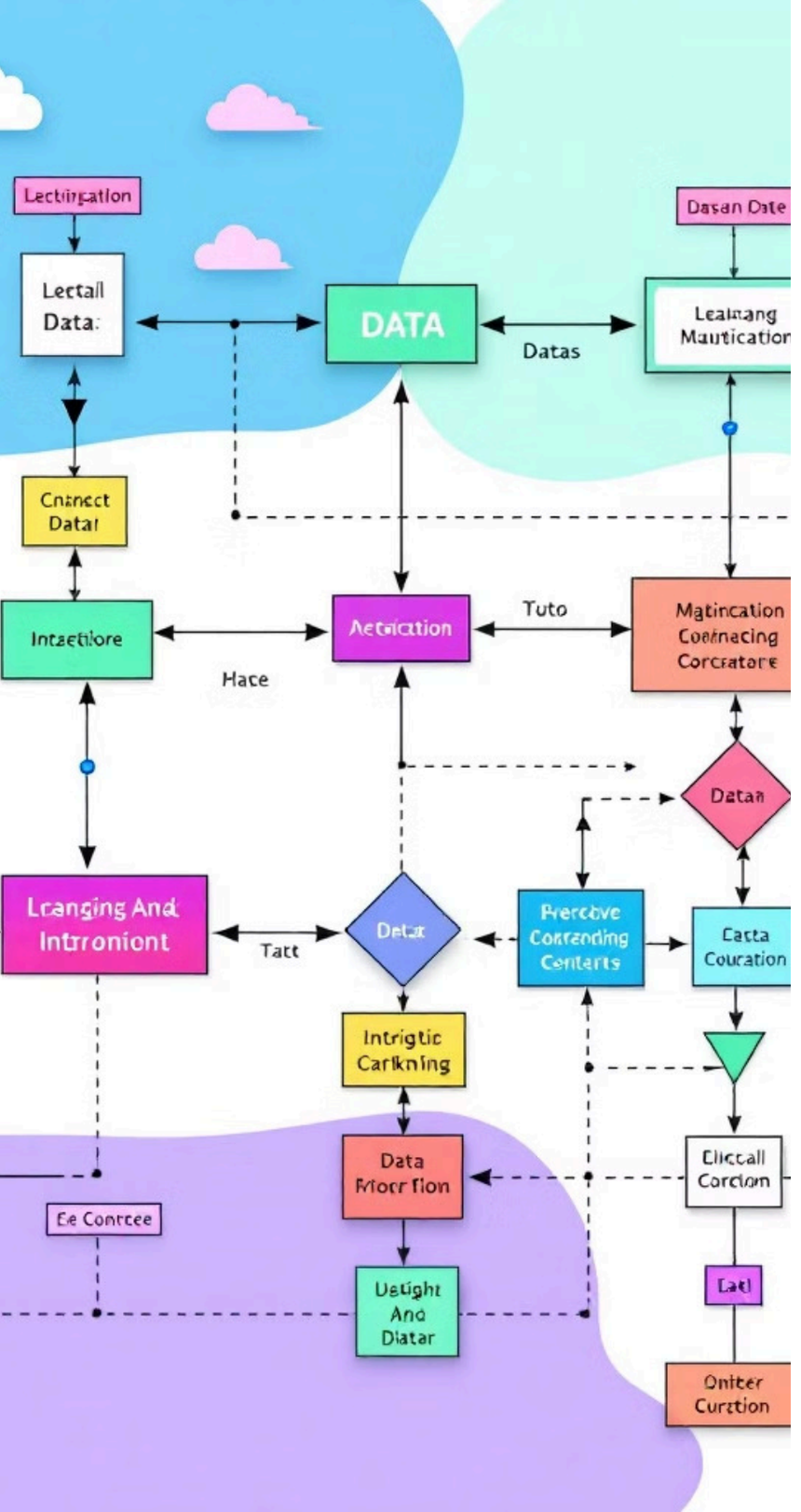


Predictive Modeling for Airbnb Prices: A Case Study

Welcome to our presentation on predicting Airbnb prices using structured data. We'll explore various models to identify the best pricing prediction strategy for improving Airbnb's market approach.



Project Agenda

1

Introduction & Objective

We'll outline our goals and the business implications of our study.

2

Data Preparation

Learn about our data cleaning and enhancement processes.

3

Exploratory Data Analysis (EDA)

Discover insights from our initial data exploration.

4

Model Exploration

We'll examine different predictive models and their performance.

5

Key Insights

End with key insights and recommendations based on the final selected model

Project Objective and Implications

Predict Airbnb Prices

We aim to create accurate models using structured data.

Evaluate Model Performance

We'll use Adjusted R^2, RMSE, and AIC as key metrics.

Improve Market Strategy

Our goal is to enhance Airbnb's pricing approach through data-driven insights.

Data Preparation: Initial Steps

1

Drop Missing Values

We removed NA entries to ensure data quality.

2

Normalize Variables

We used the IQR method to handle outliers effectively.

3

Remove Incorrect Entries

We eliminated data that didn't meet our quality standards.

4

Final Dataset

We ended up with 886 observations across 25 variables.

Data Preparation: Enhancements

One-Hot Encoding

We transformed categorical variables into numeric representations for model compatibility.

- Increased variable count to 36
- Ensured all data was in a suitable format for analysis

Why Not Label Encoding?

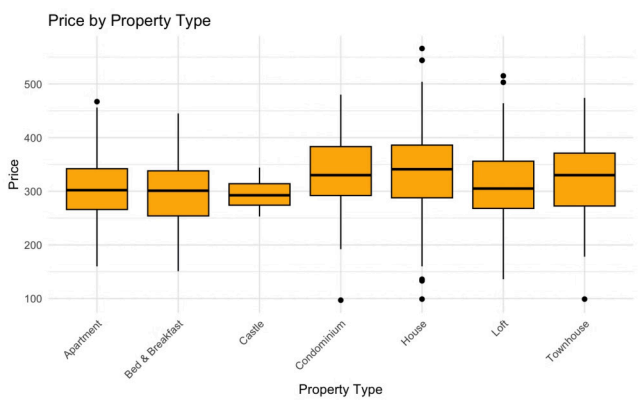
One-hot encoding avoids implying ordinal relationships in categorical data.

- Preserves true nature of categorical variables
- Prevents potential model bias from arbitrary numeric assignments

Exploratory Data Analysis: Key Insights

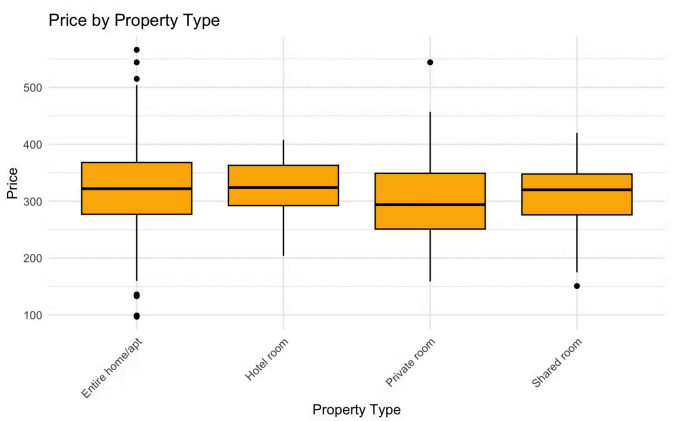
Property Type Impact

Houses and lofts generally command higher prices than apartments or B&Bs.



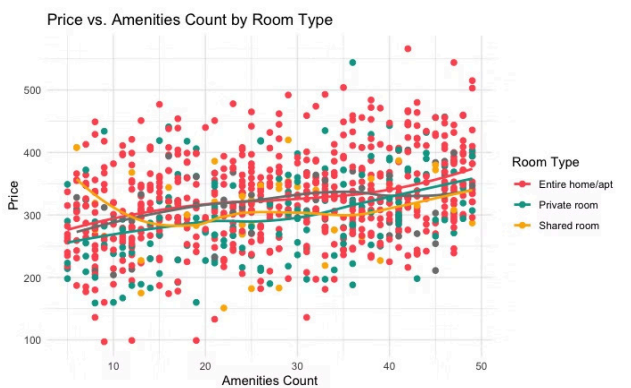
Room Type Influence

Entire homes/apartments have the highest average prices among room types.



Amenities Correlation

More amenities generally correlate with higher prices, especially for entire homes.



Model 1: Full Model

Description

Included all 35 predictors without any selection process.

Metrics

Achieved an Adjusted R2 of 38.95%.

Outcome

Retained as a benchmark but not selected due to lack of optimization.

Model 2: Significant Variables Only

All Variables

Started with 35 predictors

Significant Predictors

Identified 9 significant variables

Final Selection

Retained 8 predictors with p-values < 0.05

Model Performance

Achieved Adjusted R2 of 37.56%

Model 3: Random Test Model

19

Predictors

Iteratively removed high p-value predictors while monitoring Adjusted R^2.

39.75%

Adjusted R^2

Highest performance among the three models.

67.14

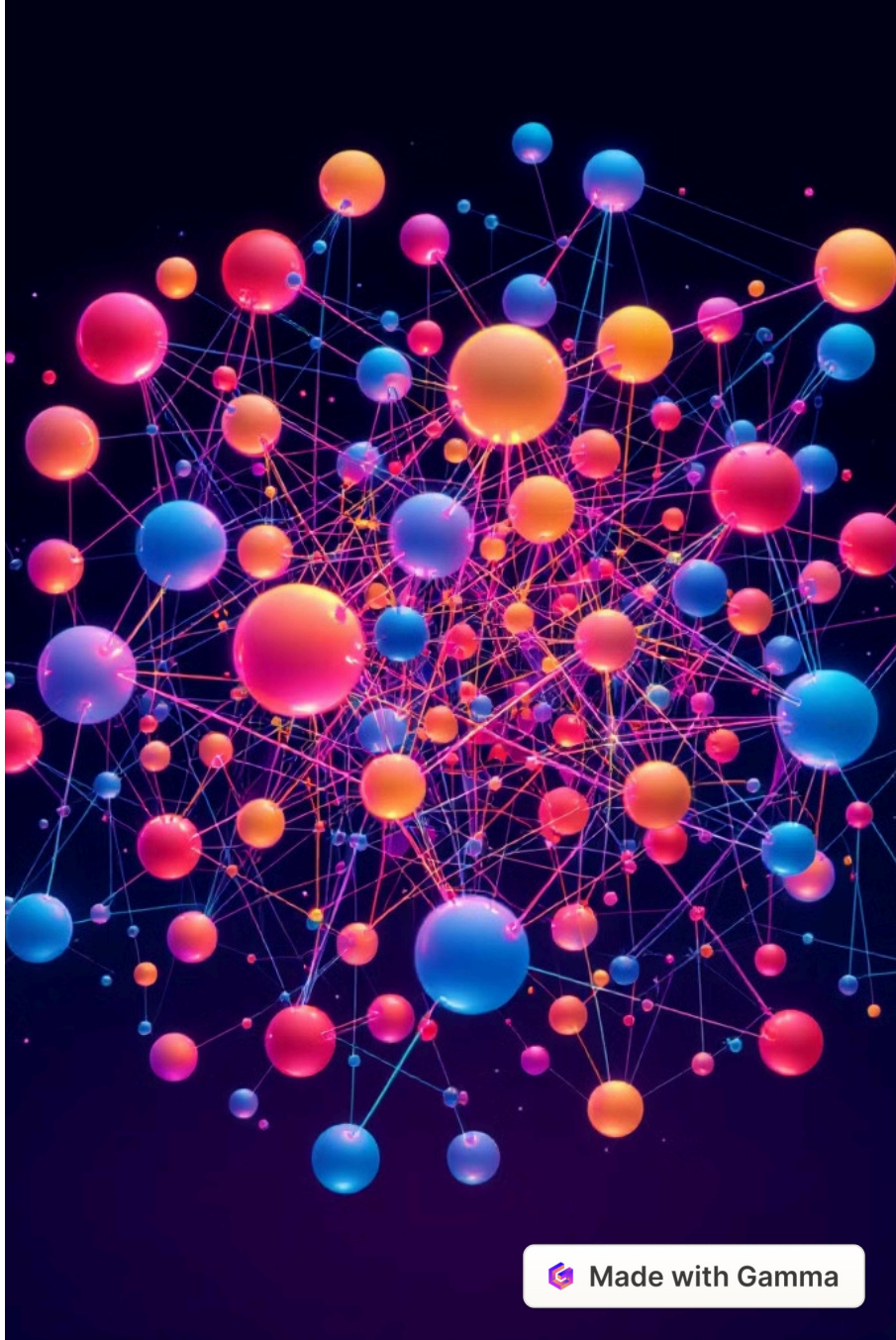
RMSE

Root Mean Square Error, indicating prediction accuracy.

5995.936

AIC

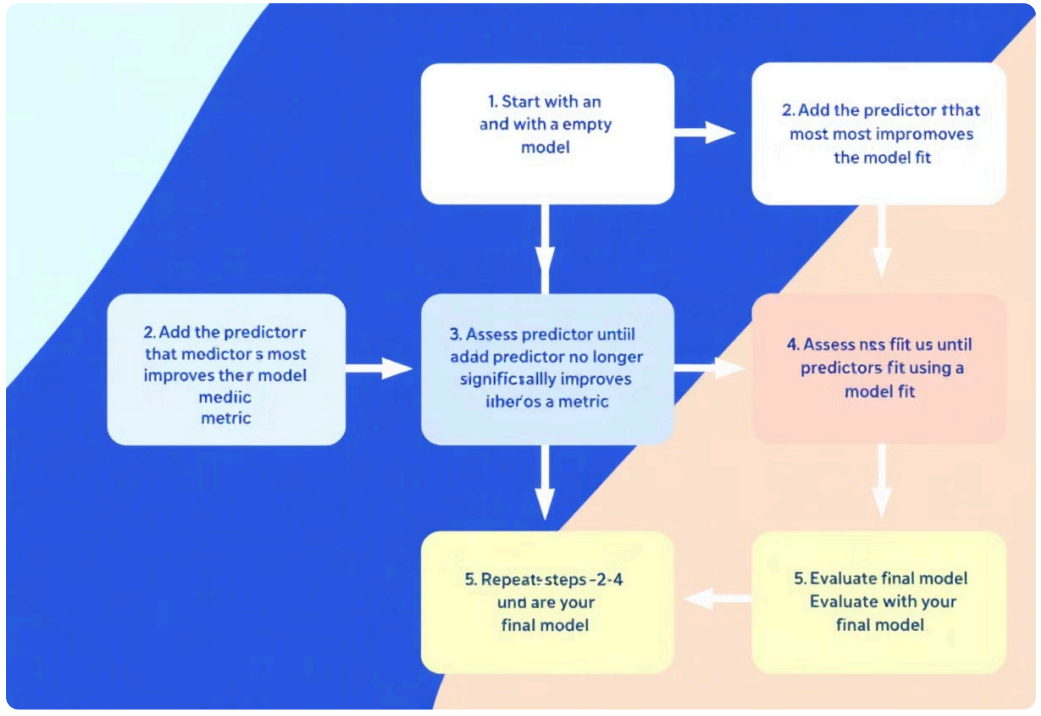
Akaike Information Criterion, balancing goodness of fit and simplicity.



Model 4: Forward Selection

Forward selection yielded a model with 13 predictors, achieving an Adjusted R^2 of 39.84% and a lower AIC (5989.34) than previous models. This resulted in a lower RMSE (67.16), indicating improved Airbnb price prediction accuracy.

This optimized model serves as a valuable benchmark due to its efficient predictor selection and strong performance metrics.



Model 5: Backward Selection

Predictor Removal

Predictors were systematically removed one by one, focusing on optimizing the Akaike Information Criterion (AIC), a metric that balances model fit and complexity.

Optimal Model

The best-performing model identified 15 key predictors, achieving an acceptable balance between prediction accuracy and model simplicity.

Performance Metrics

The backward selection approach yielded an Adjusted R^2 of 39.79%, a Root Mean Squared Error (RMSE) of 67.38, and an AIC of 5991.73.

Comparable Outcomes

The backward selection model exhibited comparable performance to the forward selection model, suggesting the robustness of the identified predictors.

Model 6: Exhaustive Search

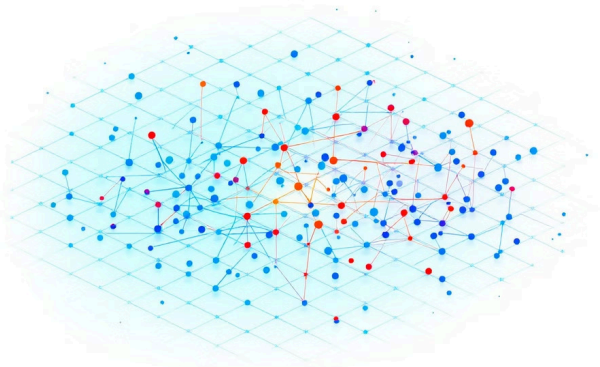
Exhaustive search involves testing all possible combinations of predictors to find the optimal subset.

This comprehensive approach identified a best model with 9 predictors.

The model achieved an Adjusted R^2 of 38.85%, with an RMSE of 67.36 and an AIC of 5994.10.

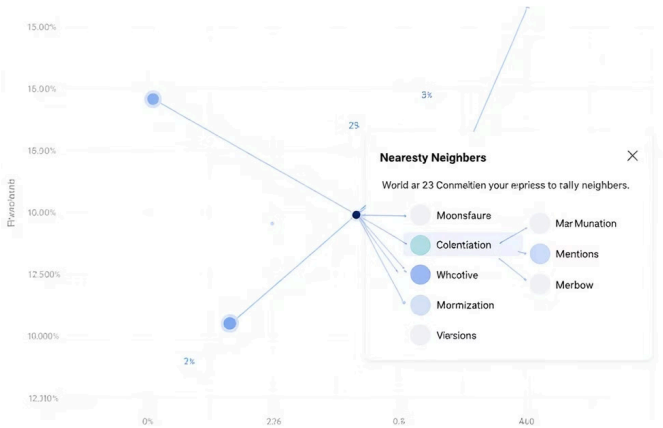
The exhaustive search method resulted in a slightly lower Adjusted R^2 compared to the stepwise methods, suggesting that the simpler models may provide better performance.

Model 7: k-Nearest Neighbors (KNN)



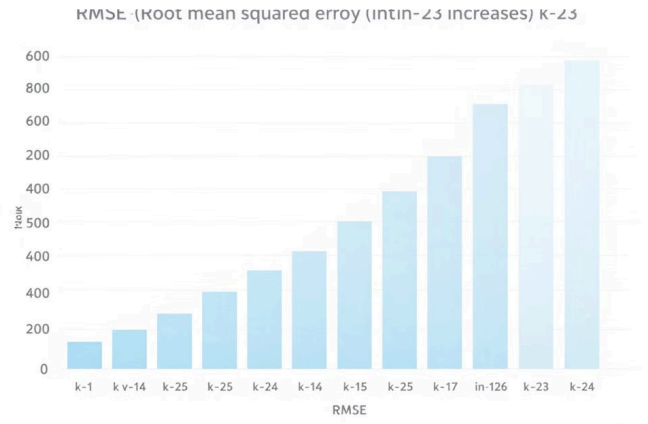
KNN Algorithm

The k-Nearest Neighbors (KNN) algorithm predicts the value of a new data point by identifying its k-nearest neighbors in the training data.



Optimal k Value

The optimal k value was determined to be 23



RMSE Performance

The model achieved an RMSE of 73.09, indicating a lower prediction accuracy compared to other models.

Model 8: Linear Regression (CV) (selected)

Cross-Validation

Applied k-fold cross-validation to evaluate model performance and prevent overfitting.

Generalization

Improved the model's ability to generalize to unseen data by evaluating its performance on multiple data splits.

Metrics

- Adjusted R^2: 39.94%
- RMSE: 67.16

Outcome

Selected as the final model due to its optimal balance of Adjusted R^2 and RMSE.

Model Performance Summary

Linear Regression (CV)

The Linear Regression model, with cross-validation, achieved the highest Adjusted R^2, indicating strong predictive power. It also exhibited a competitive RMSE, suggesting accurate predictions.

Random Test Model

The Random Test Model showcased the lowest RMSE, signifying highly accurate predictions. Despite its high AIC, it demonstrates impressive performance, potentially due to its simplicity.

Forward Selection

The Forward Selection model outperformed in terms of AIC, highlighting its optimal balance between model complexity and predictive accuracy. It also achieved a solid Adjusted R^2.

Other Models

Backward Selection and Exhaustive Search exhibited slightly higher AIC and RMSE compared to Forward Selection and Linear Regression (CV). KNN underperformed with the highest RMSE, indicating lower accuracy.

Final Model Key Findings

Metric	Value	Significance
Residual Standard Error (RSE)	64.99	Typical deviation of observed prices from predicted values
Multiple R-squared	0.4131	41.31% of price variance explained by the model
Adjusted R-squared	0.3984	Slightly less variance explained when adjusted for model complexity
F-statistic	28.15 (p-value < 2.2e-16)	Overall model significance, indicating at least one predictor is significantly related to price

The model achieved a moderate R-squared value, indicating that the identified predictors explain a significant portion of the price variation. The F-statistic further supports the overall model significance, while the RSE suggests that predicted prices deviate from actual prices by an average of $64.99.

The analysis reveals interesting insights into specific predictors. The extra people fee shows a negative correlation with price, suggesting that higher charges for additional guests may slightly reduce price appeal. Strict cancellation policies also exhibit a negative effect on price, though the correlation is not strongly significant.

Key Predictors

The analysis identified several significant predictors of Airbnb price, including **amenities**, **accommodation size**, and **property type**.

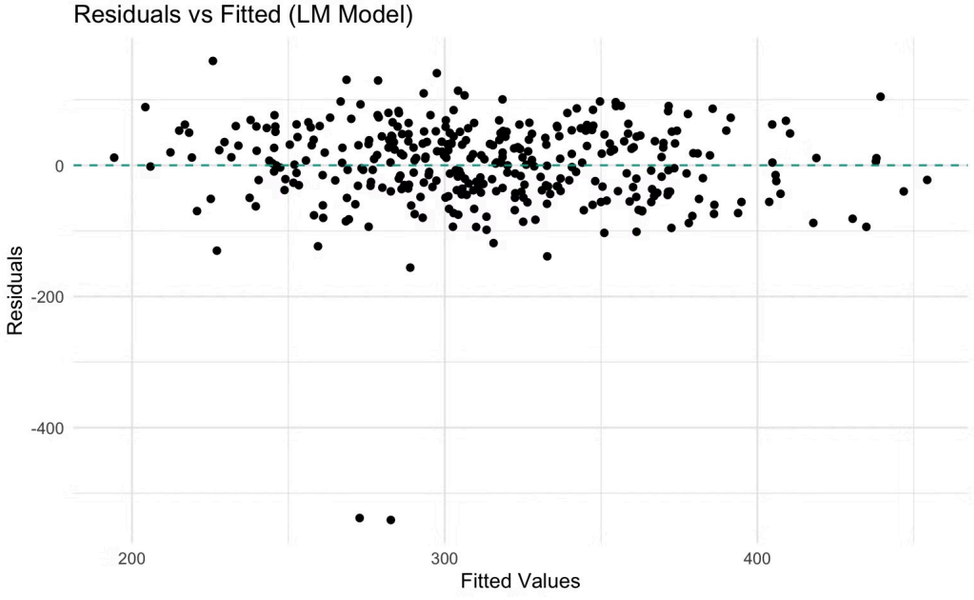
For example, each additional amenity was estimated to increase price by $2.14, while larger accommodations generally commanded higher prices.

Houses and condominiums significantly increased prices compared to the baseline, while bed and breakfasts had a negative impact on price, indicating there are priced lower.

Model Evaluation - Residuals and Fit

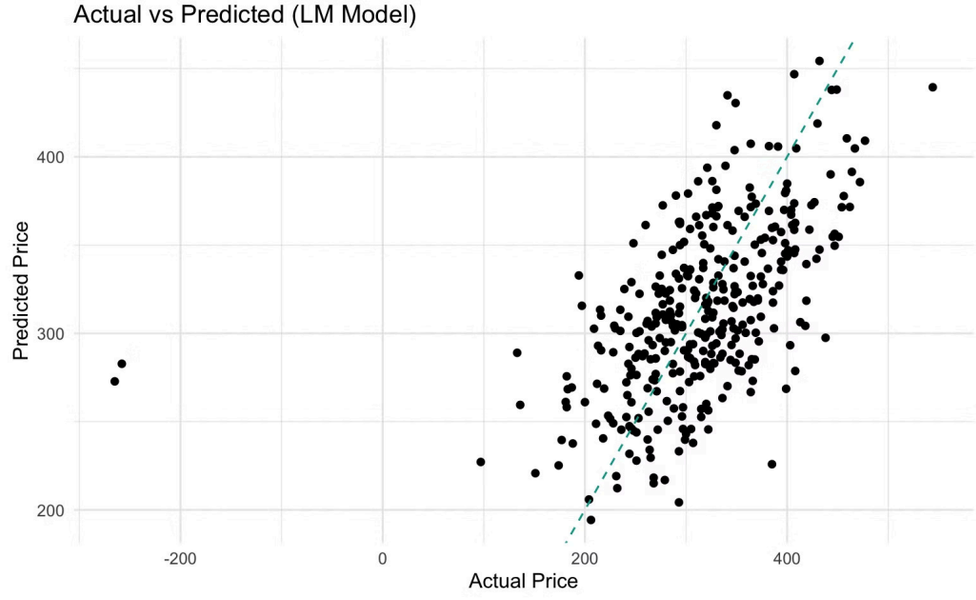
Residuals vs Fitted

Residuals show random scatter around zero, supporting linearity and homoscedasticity. Extreme outliers suggest potential model improvements.



Actual vs Predicted

Predicted values closely track actuals, with points near the diagonal. Higher-price predictions show more spread, highlighting challenges with extreme values.



Strengths:

- The model performs well for most price predictions.
- Residuals meet key assumptions of linear regression.

Limitations:

- Outliers and deviation at extreme values slightly affect performance.

Pricing Strategy Recommendations

1

Maximize Amenities

Add and highlight amenities (kitchens, laundry, parking) to increase perceived value.

2

Increase Capacity

Larger spaces accommodate more guests, justifying higher rates. Optimize for group bookings.

3

Property Features

Emphasize bedrooms, bathrooms, and premium property types. Highlight unique features.

4

Leverage Host Longevity

Market positive reviews and consistent availability.

5

Cancellation Policies

Flexible cancellation policies attract guests and minimize price reductions.

Thank You

We've analyzed your Airbnb data to create actionable insights.