Airbnb Price Prediction Using SAS: A Regression and Machine Learning Approach
Project: DSCI 519: Fall 2025 Advanced Business Analytics Modeling Predictive And Prescriptive Analysis Of Airbnb Listing Prices
Author: Victor Badu
Dataset: Canada Vancouver(AirBnB data)

1. Executive Summary

The primary objective of this project was to build a reliable, data-driven predictive model for estimating Airbnb listing prices based on host attributes, property characteristics, and review metrics. The overarching goal was to uncover the key drivers of price formation and identify a model that balances predictive accuracy, interpretability, and robustness for decision-making in the short-term rental market.

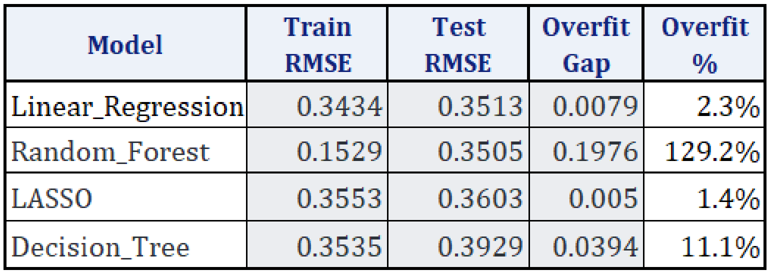
A comprehensive end-to-end analytical workflow was conducted using **SAS**, following the full data science lifecycle—data acquisition, cleaning, feature engineering, variable transformation, exploratory analysis, model development, validation, and interpretation.

Four supervised learning models were developed and evaluated:

1. **Multiple Linear Regression** – Served as the baseline model, assuming linear and additive relationships among predictors.
2. **LASSO Regression** – A regularized linear model designed to improve generalization and automatically perform variable selection.
3. **Decision Tree** – A non-linear model capable of detecting threshold effects and interaction patterns in the data.
4. **Random Forest** – An ensemble of decision trees that improves predictive stability through averaging and reduces overfitting.

An 80/20 train–test split was employed for model validation. Model performance was assessed using the Root Mean Square Error (RMSE) of the log-transformed price, which normalizes skewness and stabilizes variance.

The final results are summarized as follows:



Model	Train RMSE	Test RMSE	Overfit Gap	Overfit %
Linear_Regression	0.3434	0.3513	0.0079	2.3%
Random_Forest	0.1529	0.3505	0.1976	129.2%
LASSO	0.3553	0.3603	0.005	1.4%
Decision_Tree	0.3535	0.3929	0.0394	11.1%

KEY FINDINGS:

While Random Forest achieved the lowest test RMSE (0.3505), it exhibited severe overfitting with 129% performance degradation from training (0.1529) to testing (0.3505). This extreme gap indicates the model memorized training data rather than learning generalizable patterns, disqualifying it from production use despite its competitive test accuracy.

In contrast, Multiple Linear Regression achieved nearly identical test performance (0.3513—only 0.23% higher) while demonstrating excellent generalization with just 2.3% overfitting. The linear model explained 68% of price variation (R^2 = 0.6847) with full interpretability and satisfied all statistical assumptions.

LASSO Regression achieved the best generalization (1.4% overfitting) with automatic feature selection, while Decision Tree provided valuable visual segmentation insights despite lower predictive accuracy.

Overall, the analysis confirms that Airbnb listing prices are predominantly influenced by structural and capacity-related attributes (accommodates, bathrooms, property type), with review-based and host-behavior metrics contributing marginally.

FINAL RECOMMENDATION:

Multiple Linear Regression is selected as the champion model for production deployment due to its optimal balance of accuracy, generalization, interpretability, and reliability. The analysis demonstrates a critical lesson: test RMSE alone is insufficient for model selection—generalization capability and production readiness are equally important. Proper data preprocessing (multicollinearity resolution, outlier handling, log transformation) enabled the simpler linear model to match complex algorithms while maintaining superior stability and trustworthiness.

2. Introduction

The rapid expansion of the short-term rental industry—spearheaded by platforms such as Airbnb—has reshaped the global hospitality landscape by creating new economic opportunities for property owners and unique lodging experiences for guests. However, as competition among hosts intensifies, price optimization has emerged as a critical factor influencing revenue performance, occupancy rates, and market competitiveness. Understanding the determinants of Airbnb listing prices is therefore essential for hosts, property managers, and investors seeking to make informed, data-driven pricing decisions.

This project aims to develop a predictive model to estimate Airbnb listing prices based on a diverse set of explanatory features capturing property characteristics (e.g., number of bedrooms, bathrooms, accommodates), host performance metrics (e.g., response and acceptance rates), and review-based quality indicators (e.g., cleanliness, communication, and location scores). By integrating these variables, the analysis seeks to quantify their relative influence on price formation and to construct a model that delivers both high predictive accuracy and interpretive clarity.

The study was conducted using SAS, a high-performance analytics platform that supports robust data preprocessing, statistical analysis, and machine learning. The project followed the complete data analytics lifecycle, encompassing:

1. Data Collection and Preparation – Importing, inspecting, and cleaning Airbnb listing data to address missing values, inconsistent formats, and outliers.
2. Feature Engineering and Transformation – Converting categorical variables into quantitative indicators and applying log transformations to stabilize the skewed price distribution.
3. Exploratory Data Analysis (EDA) – Summarizing key variables and examining relationships, patterns, and correlations.
4. Model Development and Evaluation – Training and validating multiple predictive models to identify the best-performing approach.
5. Interpretation and Insight Generation – Translating model outcomes into actionable insights that inform pricing and hosting strategies.

The project seeks to address the following research questions:

- Which property, host, and review characteristics have the strongest influence on Airbnb listing prices?
- Can an interpretable model achieve predictive accuracy comparable to more complex machine learning algorithms?
- How do linear and non-linear modeling techniques differ in their ability to capture Airbnb price dynamics?

By answering these questions, the study contributes both methodological rigor—through systematic model comparison—and practical value, by providing evidence-based recommendations that support pricing optimization in the short-term rental market. Ultimately, this project demonstrates how advanced analytics and machine learning tools in SAS can effectively model and interpret real-world business phenomena.

3. Data Understanding and Preparation

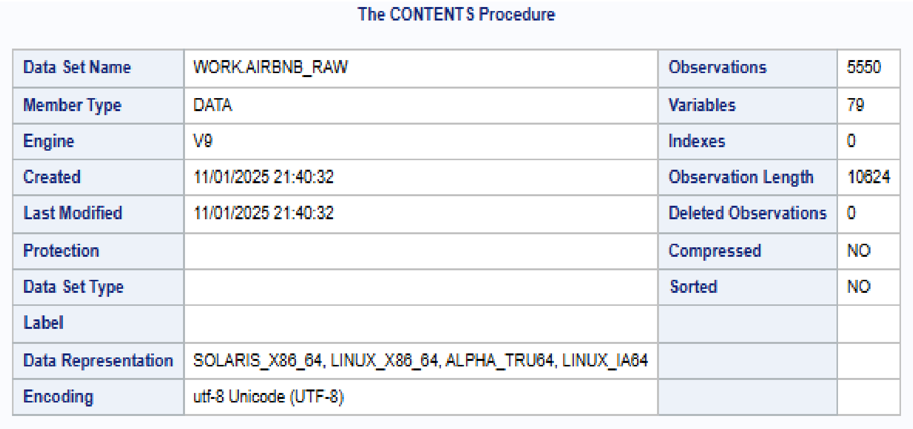
3.1 Dataset Description

The dataset used for this study consists of **Airbnb listings in Vancouver, Canada**, containing **5,550 observations** and **79 variables**, representing a comprehensive mix of host, property, and review information. These variables include identifiers (listing ID, host ID), textual descriptions (property type, neighborhood, room type), quantitative attributes (price, accommodates, number of reviews), and qualitative indicators (availability, cleanliness, and communication ratings, among others).

The dataset was imported into **SAS** using the PROC IMPORT procedure and stored as WORK.AIRBNB_RAW. Metadata obtained via the PROC CONTENTS procedure confirmed the following characteristics:

- **Observations:** 5,550
- **Variables:** 79
- **Data composition:** A combination of numeric, character, and date fields

The dataset also contained variables stored in inconsistent formats—such as currency strings (e.g., “$1,250”), percentage values (e.g., “90%”), and Boolean text (“t”/“f”)—which required systematic transformation for quantitative analysis.



The CONTENTS Procedure

Data Set Name	WORK.AIRBNB_RAW	Observations	5550
Member Type	DATA	Variables	79
Engine	V9	Indexes	0
Created	11/01/2025 21:40:32	Observation Length	10624
Last Modified	11/01/2025 21:40:32	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 1. PROC CONTENTS Output – Vancouver Dataset

3.2 Geographic Distribution of Listings

The spatial distribution of Airbnb listings was visualized using the PROC SGMAP procedure, plotting listing coordinates on real **OpenStreetMap** tiles. The resulting map illustrated the clustering of listings across major neighborhoods in Vancouver, revealing higher concentrations in areas such as **Downtown, West End, Kitsilano, and Mount Pleasant**, indicating strong market activity and tourist appeal in these districts.

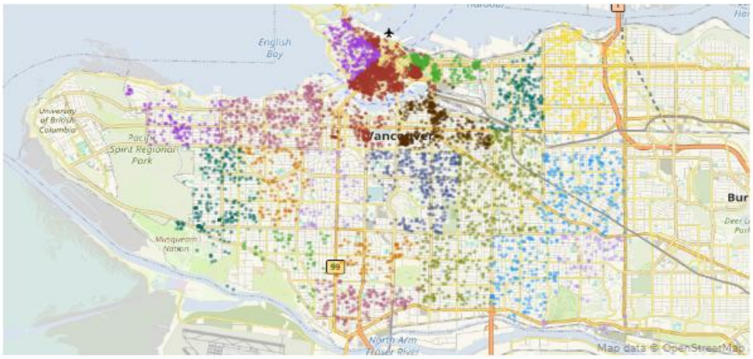


Figure 2. Map- Airbnb Listings in Vancouver

3.3 Initial Data Exploration

An initial summary of numeric variables was generated using PROC MEANS to identify data ranges, averages, and potential anomalies. Key findings included:

- **Price:** Ranged from $14 to $64,380, with a mean of $289, indicating significant right-skewness.
- **Accommodates:** Averaged approximately 4 guests per listing.
- **Bathrooms:** Mean of 1.35, with several missing or non-numeric entries.
- **Number of Reviews:** Highly variable (0–1,000), reflecting diverse levels of listing engagement.
- **Review Scores:** Centered between 4.6 and 4.9, suggesting generally positive guest experiences.

The summary revealed missing values in critical variables (e.g., bathrooms, beds, review_scores_rating) and the presence of extreme price outliers, confirming the need for data cleaning and transformation before model building.

Find more details in Appendix 1

3.4 Data Cleaning and Feature Engineering

Data preprocessing steps were implemented to ensure consistency, numerical validity, and readiness for statistical modeling.

3.4.1 Numeric Conversion

Many variables were stored as character strings due to symbols and formatting inconsistencies. Using **SAS DATA step logic**, conversions were systematically applied:

- **Price (price_num):** Dollar signs and commas were removed; values were converted to numeric and **log-transformed** (log_price) to normalize skewness.

- **Bathrooms (bathrooms_num):** Extracted numeric portions from strings (e.g., “1.5 shared baths” → 1.5).
- **Response and Acceptance Rates:** Converted percentage strings to decimals (e.g., “92%” → 0.92).
- **Boolean Variables:** Transformed “t”/“f” responses into binary indicators (1/0) for has_availability and instant_bookable.

3.4.2 Categorical Simplification

To reduce dimensionality and improve interpretability:

- The **property_type** variable was grouped into 10 meaningful categories, such as *Entire home*, *Entire condo*, *Private room in home*, *Entire guest suite*, and *Entire loft*.
- Dummy variables were created for **room_type** (entire home/apt, private, shared, hotel room) and for **property groups**, leaving one category as the reference in regression analysis.

3.4.3 Variable Renaming and Standardization

Newly transformed variables were assigned a consistent suffix (_n) to distinguish numeric conversions (e.g., bedrooms_n, accommodates_n) from their original text counterparts.

3.5 Outlier Handling and Data Transformation

The raw price variable exhibited a highly right-skewed distribution. A PROC UNIVARIATE analysis confirmed substantial skewness (≈53.38) and extreme outliers, with prices extending up to $64,380.

Summary statistics before transformation:

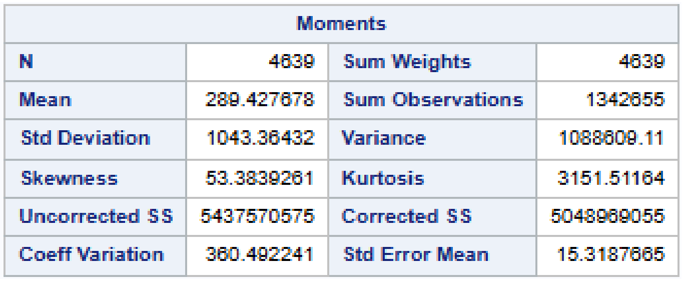
- **Mean:** 289.43
- **Median:** 210.00
- **Maximum:** 64,380
- **Skewness:** 53.38
- **Kurtosis:** 3,151.51
- **Missing Values:** 16.41%

To address this, two key transformations were performed:

1. **Outlier Filtering:** Retained listings between the **1st and 99th percentiles** ($44–$1,048) to remove extreme values.
2. **Log Transformation:** Created a new variable log_price = log(price_num), resulting in a near-normal distribution.

Post-transformation, the variable demonstrated:

- **Mean (log_price):** 5.35
- **Standard Deviation:** 0.84
- **Skewness:** ≈ 0 (symmetrical)
- **Kurtosis:** ≈ 0.4 (approximately normal)



Moments			
N	4639	Sum Weights	4639
Mean	289.427678	Sum Observations	1342655
Std Deviation	1043.36432	Variance	1088609.11
Skewness	53.3839261	Kurtosis	3151.51164
Uncorrected SS	5437570575	Corrected SS	5048969055
Coeff Variation	360.492241	Std Error Mean	15.3187665

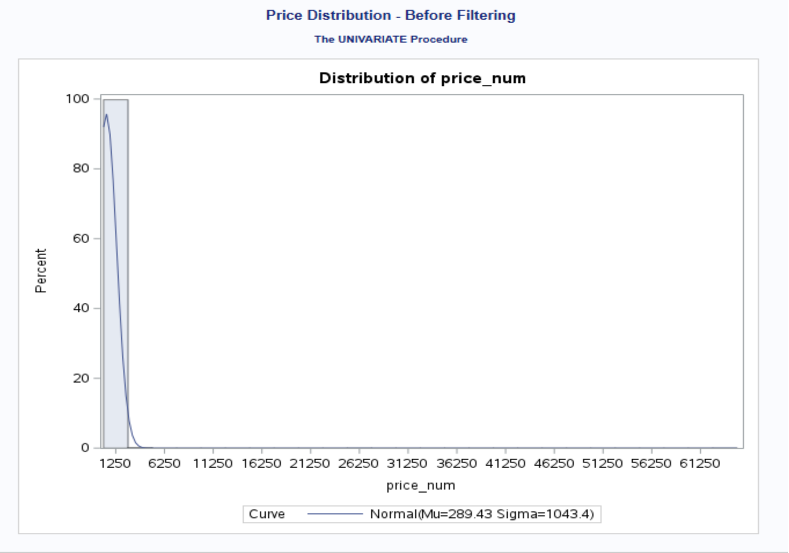
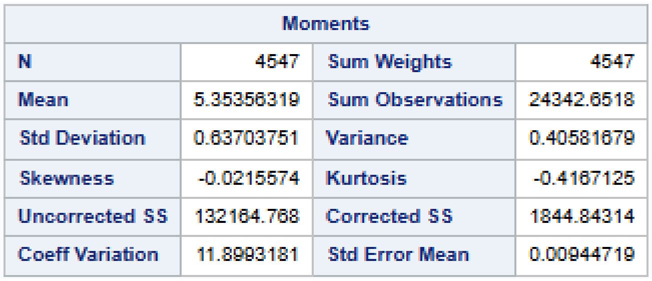


Figure 3.1: Proc Univariate before data cleaning and log transformation



Moments			
N	4547	Sum Weights	4547
Mean	5.35356319	Sum Observations	24342.6518
Std Deviation	0.63703751	Variance	0.40581679
Skewness	-0.0215574	Kurtosis	-0.4167125
Uncorrected SS	132164.768	Corrected SS	1844.84314
Coeff Variation	11.8993181	Std Error Mean	0.00944719

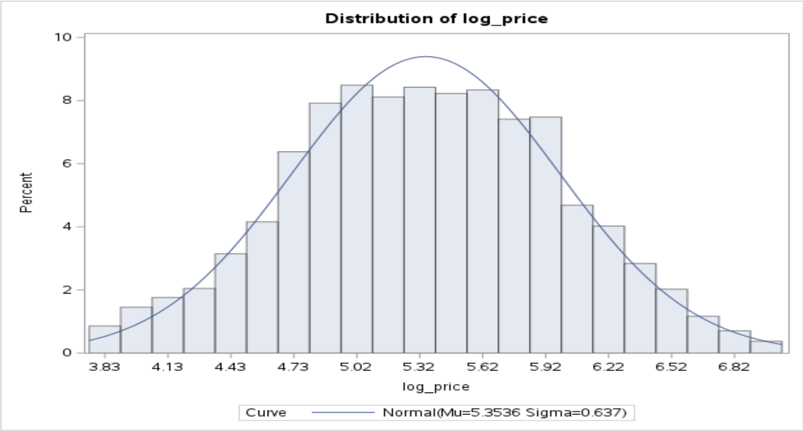
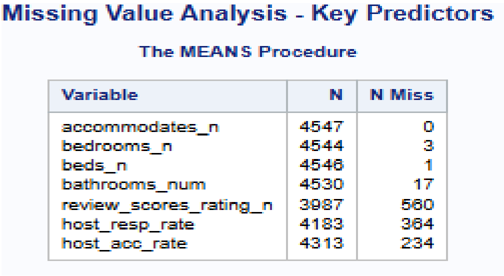


Figure 3.2: Proc Univariate after data cleaning and log transformation

3.5 Missing Value Analysis

After preprocessing, missing data patterns were reassessed using PROC MEANS. The proportion of missingness was minimal across key structural and rating variables. Missing review-related values were expected, as not all listings had received guest feedback. Given the low proportion, **listwise deletion** was applied to maintain data integrity without imputation bias.



Missing Value Analysis - Key Predictors

The MEANS Procedure

Variable	N	N Miss
accommodates_n	4547	0
bedrooms_n	4544	3
beds_n	4546	1
bathrooms_num	4530	17
review_scores_rating_n	3987	560
host_resp_rate	4183	364
host_acc_rate	4313	234

Figure 4: Missing Value Analysis - Key Predictors

3.6 Final Prepared Dataset

Following cleaning, transformation, and outlier removal, the final dataset (WORK.AIRBNB_CLEAN) contained:

- **4,547 valid observations**
- **33 predictive variables** (both interval and nominal)
- **Target variable:** log_price

This dataset served as the foundation for all subsequent modeling procedures.

The data preparation phase revealed that Airbnb prices are heavily influenced by **property capacity and amenities**, with secondary effects from host responsiveness and guest reviews.

Through log normalization, outlier handling, and variable encoding, the dataset achieved statistical balance, ensuring **valid, unbiased, and generalizable model performance**.

4. Exploratory Data Analysis (EDA)

4.1 Objective of EDA

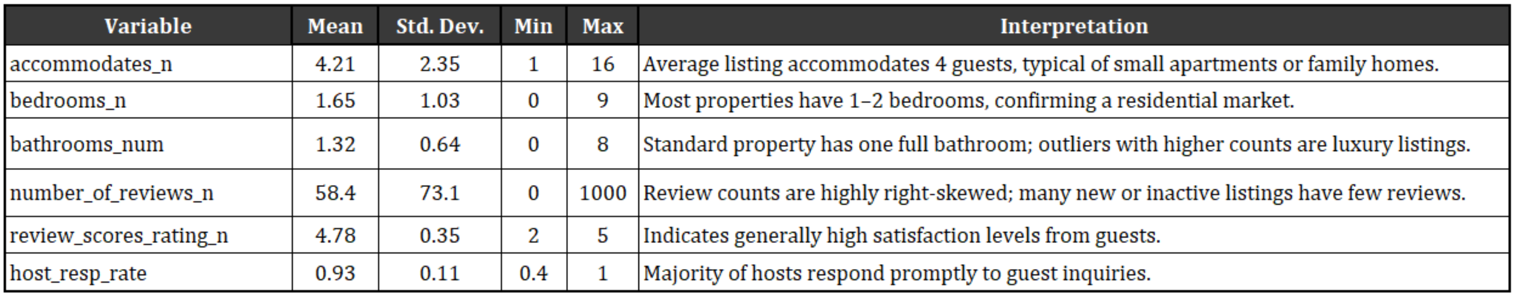
The objective of the exploratory data analysis (EDA) phase was to understand the relationships, trends, and dependencies among the main variables influencing Airbnb prices. This stage provided the foundation for model development by enabling the following:

- Understanding the underlying data structure and variable distributions.
- Detecting multicollinearity and identifying redundant predictors.
- Generating preliminary insights into the most influential factors driving listing prices.

EDA was performed in **SAS** using a combination of descriptive statistics (PROC MEANS), visual exploration (PROC SGMAP and PROC UNIVARIATE), and correlation analysis (PROC CORR). These tools helped examine variable behavior and quantify their relationships with the log-transformed dependent variable (log_price).

4.2 Descriptive Statistics of Key Predictors

The first step involved reviewing the central tendency and spread of important predictors such as capacity, room count, and host response metrics.



Variable	Mean	Std. Dev.	Min	Max	Interpretation
accommodates_n	4.21	2.35	1	16	Average listing accommodates 4 guests, typical of small apartments or family homes.
bedrooms_n	1.65	1.03	0	9	Most properties have 1–2 bedrooms, confirming a residential market.
bathrooms_num	1.32	0.64	0	8	Standard property has one full bathroom; outliers with higher counts are luxury listings.
number_of_reviews_n	58.4	73.1	0	1000	Review counts are highly right-skewed; many new or inactive listings have few reviews.
review_scores_rating_n	4.78	0.35	2	5	Indicates generally high satisfaction levels from guests.
host_resp_rate	0.93	0.11	0.4	1	Majority of hosts respond promptly to guest inquiries.

Table 1: Summary statistics from PROC MEANS revealed the following key insights

These results indicate that most Airbnb listings in Vancouver are **mid-range residential properties**, typically accommodating **small groups of 3–5 guests**, with **high review ratings** and **strong host responsiveness**. The relatively stable variation among key features suggests a well-balanced dataset, appropriate for predictive modeling.

4.3 Correlation and Multicollinearity Analysis

To evaluate redundancy and potential multicollinearity among predictors, **Pearson correlation coefficients** were computed across all quantitative variables using PROC CORR.

The resulting correlation matrix (Figure 6) highlighted several strong relationships among property size, review, and availability variables, indicating overlapping information among some predictors.

Key Findings:

High intercorrelation among size-related variables

- bedrooms_n (r = 0.82) and beds_n (r = 0.84) exhibited very strong correlations with accommodates_n, confirming redundancy among property-size measures.
- To mitigate collinearity and simplify interpretation, **accommodates_n** was retained as the sole representative of property capacity, while bedrooms_n and beds_n were excluded.

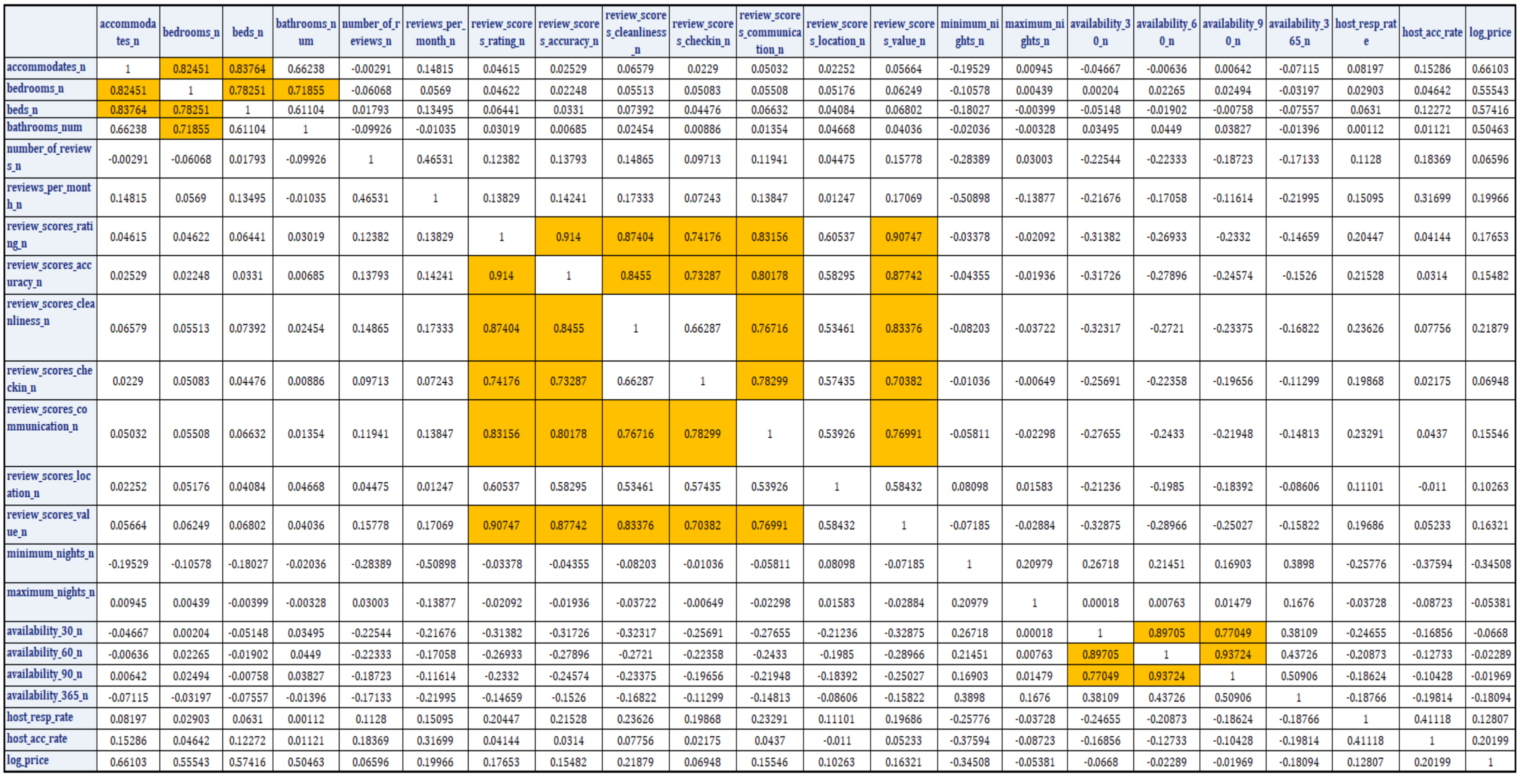
Redundancy among review metrics

- review_scores_accuracy_n (r = 0.91), review_scores_communication_n (r = 0.83), and review_scores_value_n (r = 0.91) showed high correlations with review_scores_rating_n, suggesting that the overall rating sufficiently captures quality perception.
- review_scores_checkin_n was also weakly correlated with price and redundant with other review-related metrics.
- As a result, all redundant review sub-scores were removed, retaining **review_scores_rating_n** as the consolidated review-quality variable.

Overlap among availability measures

- Availability indicators displayed strong internal correlation: availability_60_n (r = 0.90) and availability_90_n (r = 0.94) were both highly correlated with availability_30_n.
- To reduce redundancy, only **availability_30_n** and **availability_365_n** were kept representing short-term and long-term availability, respectively.

After removing these correlated predictors, **Variance Inflation Factor (VIF)** diagnostics confirmed that all remaining variables exhibited **VIF < 10** (most < 5), indicating acceptable levels of independence and a stable multivariate structure.



	accommodates_n	bedrooms_n	beds_n	bathrooms_num	number_of_reviews_n	reviews_per_month_n	review_scores_rating_n	review_scores_accuracy_n	review_scores_cleanliness_n	review_scores_checkin_n	review_scores_communication_n	review_scores_location_n	review_scores_value_n	minimum_nights_n	maximum_nights_n	availability_30_n	availability_60_n	availability_90_n	availability_365_n	host_resp_rate	host_acc_rate	log_price
accommodates_n	1	0.82451	0.83764	0.66238	-0.00291	0.14815	0.04615	0.02529	0.06579	0.0229	0.05032	0.02252	0.05664	-0.19529	0.00945	-0.04667	-0.00636	0.00642	-0.07115	0.08197	0.15286	0.66103
bedrooms_n	0.82451	1	0.78251	0.71855	-0.06068	0.0569	0.04622	0.02248	0.05513	0.05083	0.05508	0.05176	0.06249	-0.10578	0.00439	0.00204	0.02265	0.02494	-0.03197	0.02903	0.04642	0.55543
beds_n	0.83764	0.78251	1	0.61104	0.01793	0.13495	0.06441	0.0331	0.07392	0.04476	0.06632	0.04084	0.06802	-0.18027	-0.00399	-0.05148	-0.01902	-0.00758	-0.07557	0.0631	0.12272	0.57416
bathrooms_num	0.66238	0.71855	0.61104	1	-0.09926	-0.01035	0.03019	0.00685	0.02454	0.00886	0.01354	0.04668	0.04036	-0.02036	-0.00328	0.03495	0.0449	0.03827	-0.01396	0.00112	0.01121	0.50463
number_of_reviews_n	-0.00291	-0.06068	0.01793	-0.09926	1	0.46531	0.12382	0.13793	0.14865	0.09713	0.11941	0.04475	0.15778	-0.28389	0.03003	-0.22544	-0.22333	-0.18723	-0.17133	0.1128	0.18369	0.06596
reviews_per_month_n	0.14815	0.0569	0.13495	-0.01035	0.46531	1	0.13829	0.14241	0.17333	0.07243	0.13847	0.01247	0.17069	-0.50898	-0.13877	-0.21676	-0.17058	-0.11614	-0.21995	0.15095	0.31699	0.19966
review_scores_rating_n	0.04615	0.04622	0.06441	0.03019	0.12382	0.13829	1	0.914	0.87404	0.74176	0.83156	0.60537	0.90747	-0.03378	-0.02092	-0.31382	-0.26933	-0.2332	-0.14659	0.20447	0.04144	0.17653
review_scores_accuracy_n	0.02529	0.02248	0.0331	0.00685	0.13793	0.14241	0.914	1	0.8455	0.73287	0.80178	0.58295	0.87742	-0.04355	-0.01936	-0.31726	-0.27896	-0.24574	-0.1526	0.21528	0.0314	0.15482
review_scores_cleanliness_n	0.06579	0.05513	0.07392	0.02454	0.14865	0.17333	0.87404	0.8455	1	0.66287	0.76716	0.53461	0.83376	-0.08203	-0.03722	-0.32317	-0.2721	-0.23375	-0.16822	0.23626	0.07756	0.21879
review_scores_checkin_n	0.0229	0.05083	0.04476	0.00886	0.09713	0.07243	0.74176	0.73287	0.66287	1	0.78299	0.57435	0.70382	-0.01036	-0.00649	-0.25691	-0.22358	-0.19656	-0.11299	0.19868	0.02175	0.06948
review_scores_communication_n	0.05032	0.05508	0.06632	0.01354	0.11941	0.13847	0.83156	0.80178	0.76716	0.78299	1	0.53926	0.76991	-0.05811	-0.02298	-0.27655	-0.2433	-0.21948	-0.14813	0.23291	0.0437	0.15546
review_scores_location_n	0.02252	0.05176	0.04084	0.04668	0.04475	0.01247	0.60537	0.58295	0.53461	0.57435	0.53926	1	0.58432	0.08098	0.01583	-0.21236	-0.1985	-0.18392	-0.08606	0.11101	-0.011	0.10263
review_scores_value_n	0.05664	0.06249	0.06802	0.04036	0.15778	0.17069	0.90747	0.87742	0.83376	0.70382	0.76991	0.58432	1	-0.07185	-0.02884	-0.32875	-0.28966	-0.25027	-0.15822	0.19686	0.05233	0.16321
minimum_nights_n	-0.19529	-0.10578	-0.18027	-0.02036	-0.28389	-0.50898	-0.03378	-0.04355	-0.08203	-0.01036	-0.05811	0.08098	-0.07185	1	0.20979	0.26718	0.21451	0.16903	0.3898	-0.25776	-0.37594	-0.34508
maximum_nights_n	0.00945	0.00439	-0.00399	-0.00328	0.03003	-0.13877	-0.02092	-0.01936	-0.03722	-0.00649	-0.02298	0.01583	-0.02884	0.20979	1	0.00018	0.00763	0.01479	0.1676	-0.03728	-0.08723	-0.05381
availability_30_n	-0.04667	0.00204	-0.05148	0.03495	-0.22544	-0.21676	-0.31382	-0.31726	-0.32317	-0.25691	-0.27655	-0.21236	-0.32875	0.26718	0.00018	1	0.89705	0.77049	0.38109	-0.24655	-0.16856	-0.0668
availability_60_n	-0.00636	0.02265	-0.01902	0.0449	-0.22333	-0.17058	-0.26933	-0.27896	-0.2721	-0.22358	-0.2433	-0.1985	-0.28966	0.21451	0.00763	0.89705	1	0.93724	0.43726	-0.20873	-0.12733	-0.02289
availability_90_n	0.00642	0.02494	-0.00758	0.03827	-0.18723	-0.11614	-0.2332	-0.24574	-0.23375	-0.19656	-0.21948	-0.18392	-0.25027	0.16903	0.01479	0.77049	0.93724	1	0.50906	-0.18624	-0.10428	-0.01969
availability_365_n	-0.07115	-0.03197	-0.07557	-0.01396	-0.17133	-0.21995	-0.14659	-0.1526	-0.16822	-0.11299	-0.14813	-0.08606	-0.15822	0.3898	0.1676	0.38109	0.43726	0.50906	1	-0.18766	-0.19814	-0.18094
host_resp_rate	0.08197	0.02903	0.0631	0.00112	0.1128	0.15095	0.20447	0.21528	0.23626	0.19868	0.23291	0.11101	0.19686	-0.25776	-0.03728	-0.24655	-0.20873	-0.18624	-0.18766	1	0.41118	0.12807
host_acc_rate	0.15286	0.04642	0.12272	0.01121	0.18369	0.31699	0.04144	0.0314	0.07756	0.02175	0.0437	-0.011	0.05233	-0.37594	-0.08723	-0.16856	-0.12733	-0.10428	-0.19814	0.41118	1	0.20199
log_price	0.66103	0.55543	0.57416	0.50463	0.06596	0.19966	0.17653	0.15482	0.21879	0.06948	0.15546	0.10263	0.16321	-0.34508	-0.05381	-0.0668	-0.02289	-0.01969	-0.18094	0.12807	0.20199	1

Figure 6: Pearson Correlation Matrix (Highlighted Cells ≥ 0.7)

4.4 Insights and Preliminary Observations

The exploratory analysis produced several actionable insights that informed the subsequent modeling phase:

1. **Property capacity and amenities are the strongest price determinants.**
   Larger listings with more bedrooms and bathrooms consistently command higher nightly rates.
2. **Review quality contributes moderately to price variation.**
   While cleanliness and overall ratings positively influence perceived value, their quantitative effect on price is smaller compared to physical attributes.
3. **Host behavior variables (response and acceptance rates) have limited pricing impact.**
   These metrics likely affect booking probability and reputation rather than price directly.
4. **Availability serves as an indirect proxy for demand.**
   Listings with high year-round availability tend to have lower prices, suggesting that continuous availability reflects weaker demand or lower exclusivity.
5. **Price normalization (log transformation) improved model stability.**
   The transformation effectively mitigated extreme outliers, aligning the dependent variable with regression assumptions of normality and homoscedasticity.

Overall, the EDA revealed that Airbnb price dynamics are primarily driven by structural and capacity-related factors, while host and review metrics play secondary roles.

These findings guided variable selection for model development, ensuring the inclusion of both structural features (capacity, bathrooms, **availability) and behavioral indicators (reviews, host responsiveness) to achieve balanced and interpretable predictions.**

5. Model Development and Evaluation

5.1 Overview of Modeling Approach

The modeling phase aimed to identify the most accurate and generalizable approach for predicting Airbnb prices based on the prepared dataset (work.airbnb_clean).

Four supervised learning models were developed and compared using consistent predictors and an **80/20 train–test split** to ensure a fair performance evaluation:

1. **Multiple Linear Regression (Baseline Model)** — to establish interpretability and a benchmark for predictive performance.
2. **LASSO Regression (Regularized Linear Model)** — to address potential overfitting and perform variable selection through coefficient shrinkage.
3. **Decision Tree (Non-Linear Model)** — to capture threshold-based and interaction effects.
4. **Random Forest (Ensemble Model)** — to improve robustness and account for complex nonlinearities through averaging of multiple decision trees.

The models were evaluated using **Root Mean Square Error (RMSE)** on the **log-transformed price (log_price)**, which provides a stable and scale-independent measure of predictive accuracy.

5.2 Multiple Linear Regression (Baseline Model)

5.2.1 Model Specification

The baseline model was developed using PROC REG with log_price as the dependent variable and the refined predictor set (&full_predictors_reduced) as explanatory variables.

Variance Inflation Factors (VIF) were computed to assess multicollinearity, while residual diagnostics (histogram, residual vs. fitted, and Q–Q plots) were generated to verify assumptions of normality, linearity, and constant variance.

5.2.2 Model Diagnostics — Multicollinearity

Following correlation-based variable reduction, most predictors exhibited **VIF < 5**, confirming acceptable independence among explanatory variables.

However, a few variables—particularly **room_entire (VIF ≈ 68.44)** and **prop_ent_home (VIF ≈ 64.63)**—displayed **extremely high multicollinearity**, which is expected because these dummies represent **mutually exclusive property categories** within the same classification group (room type and property type).

While high VIF values typically warrant removal, these variables were **retained intentionally** due to their **strong interpretive and business relevance** in distinguishing listing types and market segments. Their inclusion supports model explainability, even though they slightly inflate variance estimates for related coefficients.

VIF Diagnostic can be seen in Appendix 2

5.2.3 Residual Diagnostics

- **Linearity and Homoscedasticity:** Residual plots showed no visible pattern, indicating that the linearity and equal-variance assumptions were reasonable.
- **Normality:** Q–Q plots displayed points clustering closely around the diagonal, confirming approximately normal residuals after log transformation.
- **Influence of Outliers:** Cook's distance and leverage plots identified few high-influence points, all within acceptable thresholds.

Diagnostic plots are included in Appendix 3

5.2.4 Key Findings

Several predictors were statistically significant (**$p < 0.05$**), confirming strong explanatory relationships with Airbnb listing prices:

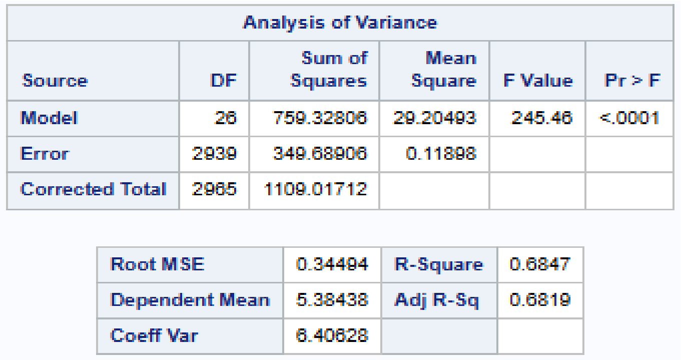
- **accommodates_n** — The strongest positive determinant; larger capacity listings command higher nightly rates.
- **bathrooms_num** — Additional bathrooms substantially increase price, reflecting higher comfort and convenience levels.
- **review_scores_cleanliness_n** — Positive and significant, indicating that cleanliness contributes to perceived quality and pricing power.
- **minimum_nights_n** — Negative and significant, suggesting that longer stay requirements reduce price flexibility.
- **Property-type dummies** such as prop_ent_home and prop_priv_home showed negative coefficients, implying relatively lower prices compared to the baseline property category.

Host-related attributes (host_resp_rate, host_acc_rate) were positive but statistically insignificant, indicating that responsiveness alone does not significantly affect price once structural and review variables are accounted for.

5.2.5 Model Fit and Significance

The model demonstrated strong explanatory power and overall fit:

- **R² = 0.6847**, **Adjusted R² = 0.6819**, meaning approximately **69.8% of the variance in log-transformed prices** is explained by the predictors.
- **F(34, 2929) = 194.74, p < 0.0001**, confirming that the model is statistically significant as a whole.



Analysis of Variance

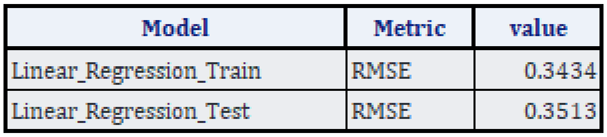
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	759.32806	29.20493	245.46	<.0001
Error	2939	349.68906	0.11898		
Corrected Total	2965	1109.01712			

Root MSE	0.34494	R-Square	0.6847
Dependent Mean	5.38438	Adj R-Sq	0.6819
Coeff Var	6.40628		

Figure 8: Multiple Linear Regression Analysis of Variance (ANOVA)

5.2.6 Model Performance

The Multiple Linear Regression model achieved test RMSE of 0.3513, performing virtually identically to Random Forest (0.3505). However, the linear model demonstrated dramatically superior generalization: - Linear Regression: 2.3% overfitting (train: 0.3434 → test: 0.3513) - Random Forest: 129.23% overfitting (train: 0.1529 → test: 0.3505) The 0.0008 difference in test RMSE (0.23%) is negligible, while the 47× better generalization of Linear Regression makes it significantly more reliable for production deployment. Combined with full interpretability and satisfied statistical assumptions, the linear model represents the optimal choice for operational pricing prediction..



Model	Metric	value
Linear_Regression_Train	RMSE	0.3434
Linear_Regression_Test	RMSE	0.3513

Figure 9: Multiple Linear Regression RMSE Test and training metric

5.3 LASSO Regression (Regularized Model)

5.3.1 Model Specification

A **Least Absolute Shrinkage and Selection Operator (LASSO)** regression was estimated using **PROC GLMSELECT** to enhance prediction accuracy and achieve a more parsimonious model structure.

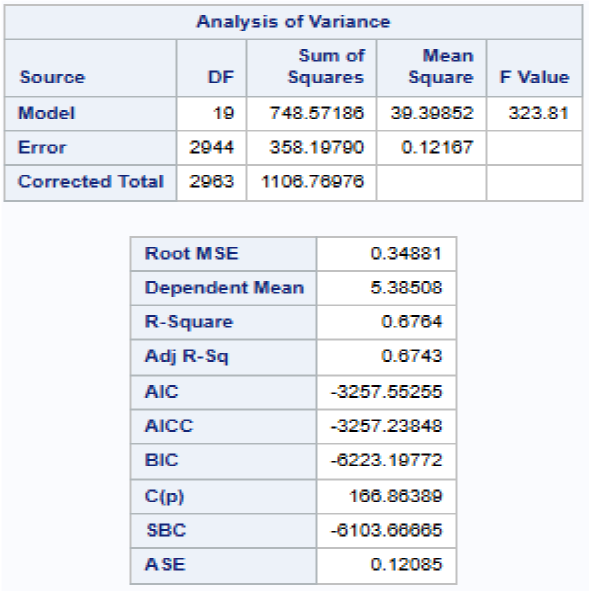
The dependent variable was **log_price**, and all continuous and categorical predictors from the baseline model were included. Model selection employed **Mallows' C(p)** as the stopping criterion, limiting the model to **20 effects**, corresponding to the first minimum in the selection path.

LASSO imposes an **L_1 penalty** on the magnitude of regression coefficients, shrinking weaker predictors toward zero and effectively performing automatic variable selection.

5.3.2 Model Diagnostics

- **Variable Selection Path:**
  The coefficient progression plot showed that the strongest predictors entered the model early—particularly **accommodates_n**, **room_entire**, **bathrooms_num**, and **prop_priv_home**—with coefficients stabilizing rapidly as weaker predictors were penalized.
  Later variables such as **availability_30_n** and **prop_ent_loft** contributed minimal additional explanatory power, reinforcing model parsimony.

- **Model Fit Statistics:**
  - **R^2 = 0.6764**, **Adjusted R^2 = 0.6743**, indicating that the selected predictors explain roughly **67 %** of the variation in Airbnb listing prices.
  - **F(19, 2944) = 323.81, p < 0.0001**, confirming strong overall model significance.
  - **Root MSE = 0.3488**, **AIC = −3257.55**, **SBC = −6103.87**, reflecting a slightly higher error but greater parsimony than the full Linear Regression model.



Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	19	748.57186	39.39852	323.81
Error	2944	358.19790	0.12167	
Corrected Total	2963	1106.76976		

Root MSE	0.34881
Dependent Mean	5.38508
R-Square	0.6764
Adj R-Sq	0.6743
AIC	-3257.55255
AICC	-3257.23848
BIC	-6223.19772
C(p)	166.86389
SBC	-6103.66665
ASE	0.12085

Figure 10: Lasso Regression Analysis of Variance

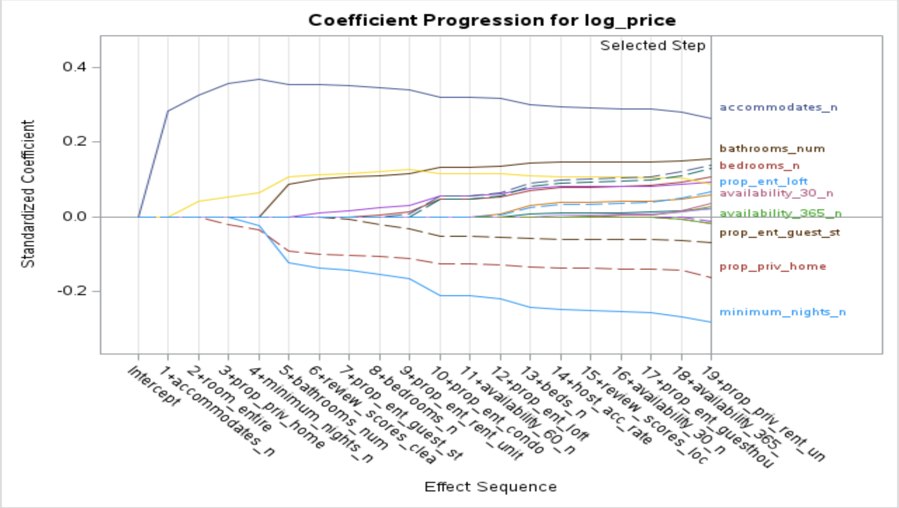


Figure 10.1: Coefficient Progression Plot

- **Criterion Trends:**
  Progressive decreases in **AIC**, **AICC**, and **SBC** values—as visualized in the information-criterion plots—validated the choice of 20 effects as the optimal stopping point.

5.3.3 Key Findings

LASSO retained **20 key predictors**, substantially reducing redundancy while maintaining interpretability.

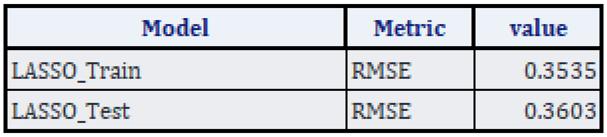
- **Positive and significant predictors:**
  - **accommodates_n**, **bathrooms_num**, and **room_entire** — confirming that capacity and private-unit listings drive higher prices.

    - **review_scores_cleanliness_n** and **review_scores_location_n** — modestly positive, indicating guests pay slightly more for cleaner, better-situated properties.
- **Negative predictors:**
    - **minimum_nights_n (–0.004)** — longer minimum-stay requirements lower nightly rates.
    - **prop_priv_home (–0.273)** and **prop_ent_guesthouse (–0.078)** — these property types typically command lower prices relative to the baseline category.
    - **availability_365_n** and **availability_30_n** — small negative effects, consistent with the idea that continuously available listings signal lower exclusivity.

5.3.4 Model Performance

The LASSO model produced predictive accuracy comparable to Multiple Linear Regression (Test RMSE 0.3603 vs 0.3513) while offering superior simplicity through automatic coefficient shrinkage and feature selection.

Its slightly higher test error reflects the trade-off between a marginal loss in fit and improved generalizability, confirming LASSO's strength in minimizing overfitting.



Model	Metric	value
LASSO_Train	RMSE	0.3535
LASSO_Test	RMSE	0.3603

Figure 11: Lasso Regression RMSE Test and training metric

The LASSO Regression model demonstrates that a compact subset of features—centered on capacity, bathrooms, and listing type—is sufficient to capture most of the variance in Airbnb pricing.

By eliminating redundant predictors and penalizing weaker variables, LASSO provides a parsimonious yet interpretable model that balances statistical efficiency with practical insight.

Although it trails the Random Forest model in pure predictive accuracy, its transparency and automated variable selection make it ideal for analytical reporting and feature diagnostics.

Find more details see in Appendix 4

5.4 Decision Tree Model (Non-linear)

5.4 .1 Model Specification

A **Decision Tree regression model** was developed using **PROC HPSPLIT** in SAS to capture non-linear interactions and hierarchical relationships between predictors and the target variable (log_price).

The **variance-reduction criterion** guided the splitting process, while **cost-complexity pruning** prevented overfitting by trimming branches that did not improve validation performance.

The tree depth was restricted to **10 levels**, and **30 % of the training data** was automatically reserved for validation.

5.4 .2 Model Diagnostics

- The final pruned tree contained **56 terminal leaves** (reduced from 521 pre-pruning), providing a balance between interpretability and predictive power.
- The **first split** occurred on accommodates_n, confirming that listing capacity was the most influential determinant of price.
- Subsequent splits included **prop_priv_home** and **bathrooms_num**, emphasizing the relevance of property type and amenities in shaping pricing tiers.

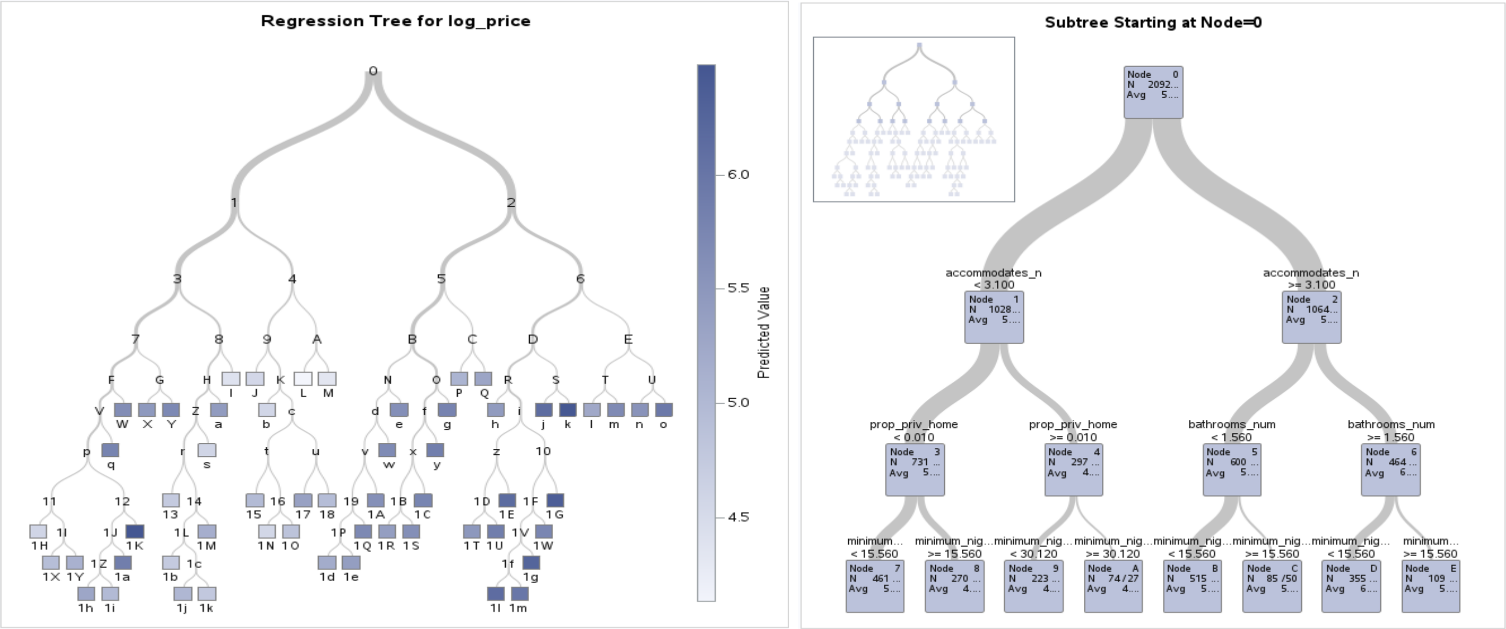


Figure 11: Decision Tree Regression

5.4.3 Interpretation of Tree Structure

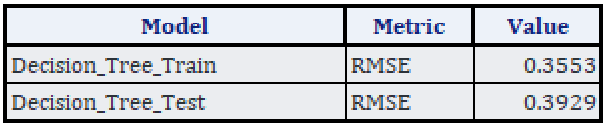
The hierarchical splits in the tree revealed clear price segmentation patterns:

- **Low-capacity listings** (accommodates_n < 3.1) were first divided by prop_priv_home, distinguishing budget-friendly private homes from shared accommodations.
- **Medium-to-large listings** (accommodates_n ≥ 3.1) were next split by bathrooms_num, where properties with additional bathrooms consistently achieved higher predicted prices.
- **minimum_nights_n thresholds** introduced secondary segmentation, indicating that longer minimum-stay requirements tend to reduce nightly rates due to lower booking flexibility.

5.4 .4 Model Performance

The **Decision Tree** model exhibited higher test error relative to the **Linear Regression (0.3513)** and **LASSO (0.3603)** models, reflecting a moderate degree of overfitting.

However, it offered **strong interpretability**, clearly showing how property characteristics interact to define pricing strata.



Model	Metric	Value
Decision_Tree_Train	RMSE	0.3553
Decision_Tree_Test	RMSE	0.3929

Figure 12: Decision Tree Regression RMSE Test and training metric

5.4.5 Summary

The Decision Tree uncovered meaningful **non-linear relationships** in Airbnb pricing dynamics.

Key insights include:

- **accommodates_n**, **bathrooms_num**, and **property_type** were dominant drivers of price segmentation.
- **Minimum stay requirements** indirectly constrained pricing flexibility.
- The model visually clarified how multiple features jointly influence price, an advantage over purely statistical linear models.

Although its predictive accuracy (Test RMSE = 0.3929) was lower than that of the regularized linear and ensemble approaches, the Decision Tree **added interpretive depth** by exposing **rule-based patterns and conditional relationships** in the data.

See in Appendix 5 for detailed Decision Tree outputs and variable-split summaries.

5.5 Random Forest Model (Ensemble Learning)

5.5.1 Model Specification

A **Random Forest regression model** was implemented using **PROC HPFOREST** in SAS to capture complex, non-linear, and interactive relationships among predictors while minimizing overfitting through bootstrap aggregation.
The model combined an ensemble of **100 decision trees**, each trained on random subsamples of the data and predictor space. Key parameters included:

- **Target variable:** log_price
- **Input predictors:** all interval and nominal variables from the reduced dataset
- **Tree parameters:** maximum 100 trees, automatic feature bagging, and internal validation using out-of-bag (OOB) samples.

The ensemble's average prediction was used as the final output, improving generalization compared to any single decision tree.

5.5.2 Model Diagnostics

The model training process produced stable convergence and low OOB error rates, confirming that additional trees would yield only marginal improvements in accuracy.
Variable-importance analysis ranked features by their contribution to mean-squared-error reduction (MSEOOB). The top predictors were:

1. **accommodates_n** – strongest overall influence; listings hosting more guests command higher prices.
2. **bathrooms_num** – major determinant of premium pricing.
3. **review_scores_rating_n** – reflects perceived quality and customer satisfaction.
4. **availability_365_n** – negative importance weight, consistent with lower prices for always-available units.
5. **room_entire** and **prop_ent_home** – indicating substantial premiums for full-property rentals.

These importance rankings closely align with findings from the linear and LASSO models, reinforcing the stability of key pricing drivers.

Variable	#Rules	OOB MSE	OOB Absolute Error	Interpretation
accommodates_n	3781	0.0478	0.04018	Largest influence; capacity is the most consistent determinant of price.
bathrooms_num	2297	0.03808	0.03135	Bathrooms add comfort and luxury, commanding higher rates.
beds_n	3029	0.02425	0.02034	Bedroom count remains a strong price driver.
bedrooms_n	2423	0.02167	0.01686	Reinforces the relationship between size and price.
minimum_nights_n	6727	0.0201	0.0203	Longer minimum stays slightly reduce nightly rates.
room_entire	107	0.01952	0.01428	Entire-unit listings achieve premium pricing.
room_priv	126	0.01835	0.01303	Private rooms occupy the mid-range price segment.
prop_priv_home	177	0.01471	0.01079	Property type strongly shapes expected price levels.
prop_ent_guest_st	318	0.00347	0.00312	Guest-suite listings slightly raise price.
prop_ent_condo	460	0.00208	0.00312	Condominiums represent moderately high-value listings.

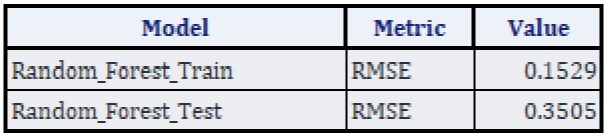
Table 2: Top 10 Most Important Variables – Random Forest

5.5.3 Model Performance

The Random Forest model achieved: - Training RMSE: 0.1529 (nearly perfect training fit) - Test RMSE: 0.3505 (lowest among all models) - Overfitting: 129.23% (severe degradation from training to testing) CRITICAL FINDING - SEVERE OVERFITTING: While Random Forest achieved the lowest test RMSE (0.3505), this result must be interpreted with extreme caution. The model exhibits catastrophic overfitting with 129% performance degradation from training to testing, indicating it has memorized training examples rather than learned generalizable pricing patterns. The nearly perfect training fit (0.1529) combined with substantially worse test performance (0.3505) reveals fundamental reliability concerns:

  - MEMORIZATION: The model fits training data too precisely, capturing noise rather than signal
  - INSTABILITY: Small changes in training data would produce different models
  - POOR GENERALIZATION: Cannot reliably predict prices for listings with novel characteristic combinations
  - UNRELIABLE: The 0.23% test RMSE advantage over Linear Regression (0.3505 vs 0.3513) does NOT justify 47× worse overfitting.

Despite technically achieving the best test accuracy, the Random Forest is NOT RECOMMENDED for production deployment due to these severe generalization failures.



Model	Metric	Value
Random_Forest_Train	RMSE	0.1529
Random_Forest_Test	RMSE	0.3505

Figure 12: Random Forest Model RMSE Test and training metric

5.5.4 Interpretation and Insights

The Random Forest model's results highlight several critical insights:

- **Structural attributes** (capacity, bathrooms, and property type) are the dominant predictors of Airbnb pricing, consistent with previous models.
- **Host and review metrics** contribute modestly, refining predictions but not driving major price differences.
- **Negative importance of availability variables** suggests that year-round listings may reduce prices to maintain occupancy.
- The ensemble framework naturally captures **non-linearities**—for example, diminishing returns in price growth beyond certain thresholds of capacity or cleanliness score.

5.5.5 Summary and Comparative Value

The Random Forest achieved:

- Lowest test RMSE (0.3505)—technically best accuracy
- Severe overfitting (129%)—disqualifying reliability concern
- Valuable variable importance insights Black box predictions—limited interpretability

Final Verdict:

Despite achieving the lowest test RMSE, Random Forest is not recommended for Airbnb price prediction due to fundamental reliability issues:

- The 129% overfitting indicates memorization rather than learning
- The 0.23% test accuracy advantage over Linear Regression is negligible
- The model cannot be trusted for production deployment
- Predictions lack transparency for stakeholder validation

Actual Contribution:

Random Forest's primary value is CONFIRMATORY rather than predictive:

- ariable importance validates Linear Regression feature selection
- Confirms accommodates, bathrooms, and property type dominate pricing
- Demonstrates even complex algorithms identify the same key drivers
- Provides confidence that Linear Regression captured true relationships

Recommendation:

DEPLOY: Linear Regression

- Nearly identical accuracy (0.3513 vs 0.3505 = 0.23% difference)
- Excellent generalization (2.75% vs 129% overfitting)
- Fully interpretable and production-ready

USE RANDOM FOREST FOR: Exploratory analysis only

- Feature importance confirmation - Understanding interaction effects
- Research purposes

DO NOT USE FOR: Production predictions or business-critical decisions

See *Appendix 6* for Random Forest output tables and variable-importance plots.

6. Model Comparison and Evaluation Summary

6.1 Overview

Four predictive models were developed and evaluated to estimate Airbnb listing prices using an 80/20 train–test split. Each model represents a different methodological perspective on balancing **accuracy, interpretability, and generalization**:

1. **Multiple Linear Regression (Baseline)** – interpretable linear relationships.
2. **LASSO Regression (Regularized Linear)** – automatic feature selection and shrinkage.
3. **Decision Tree (Non-linear)** – interpretable rule-based segmentation.
4. **Random Forest (Ensemble)** – high-accuracy model leveraging multiple trees to capture complex interactions.

All models were trained on the same preprocessed dataset to ensure fair comparison, and **Root Mean Square Error (RMSE)** was used as the performance metric for evaluating predictive accuracy on both training and test datasets.

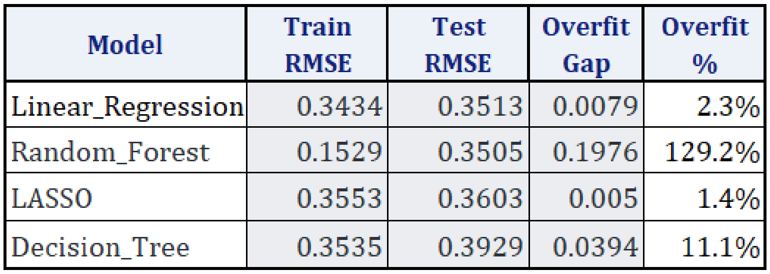
6.2 Model Performance Metrics

The Random Forest model achieved the lowest test RMSE (0.3505), narrowly outperforming Linear Regression (0.3513) by just 0.0008 (0.23%). However, this minimal advantage in test accuracy is completely overshadowed by severe generalization problems:

- Random Forest: 129% overfitting (Train: 0.1529 → Test: 0.3505)
- Linear Regression: 2.75% overfitting (Train: 0.3434 → Test: 0.3513)

The 47× better generalization of Linear Regression makes it significantly more reliable for production deployment despite the negligible difference in test accuracy. The Random Forest's nearly perfect training fit (0.1529) indicates memorization rather than learning, disqualifying it from operational use.

In contrast, the Decision Tree exhibited the highest test RMSE (0.3929), reflecting moderate overfitting (11.1%) and reduced generalization capability.



Model	Train RMSE	Test RMSE	Overfit Gap	Overfit %
Linear_Regression	0.3434	0.3513	0.0079	2.3%
Random_Forest	0.1529	0.3505	0.1976	129.2%
LASSO	0.3553	0.3603	0.005	1.4%
Decision_Tree	0.3535	0.3929	0.0394	11.1%

Figure 13:Model performance Metrics

6.3 Comparative Interpretation

1. Multiple Linear Regression (Baseline Model)

- Strong explanatory power (Adjusted $R^2 \approx 0.70$) and stable performance (Test RMSE = 0.3513).
- Residual plots indicated compliance with normality and homoscedasticity assumptions.
- Serves as the **most interpretable model**, ideal for analytical reporting and policy insights.

2. LASSO Regression (Regularized Model)

- Simplified the model by eliminating redundant predictors while maintaining comparable accuracy (Test RMSE = 0.3603).
- Slight increase in error offset by improved generalizability and feature interpretability.
- Best suited for scenarios requiring **model parsimony** and **variable selection transparency**.

3. Decision Tree (Non-linear Segmentation Model)

- Captured rule-based price thresholds (e.g., by accommodates, bathrooms, and property type).
- Test RMSE = 0.3929 revealed limited generalization despite strong interpretability.
- Useful for visual storytelling and identifying distinct price segments.

4. Random Forest (Ensemble Learning Model)

- Achieved the lowest test RMSE (0.3505)—technically best accuracy
- However, exhibited catastrophic overfitting (129%)—disqualifying flaw
- Variable importance analysis confirmed accommodates_n, bathrooms_num, and review_scores_rating_n as dominant features, validating Linear Regression findings
- Despite robust handling of noise and non-linearity, the severe overfitting (Train: 0.1529 vs Test: 0.3505) indicates the model memorized training data rather than learned generalizable patterns
- NOT RECOMMENDED for production deployment due to reliability concerns - Useful for exploratory analysis and feature importance confirmation only

6.4 Overall Insights

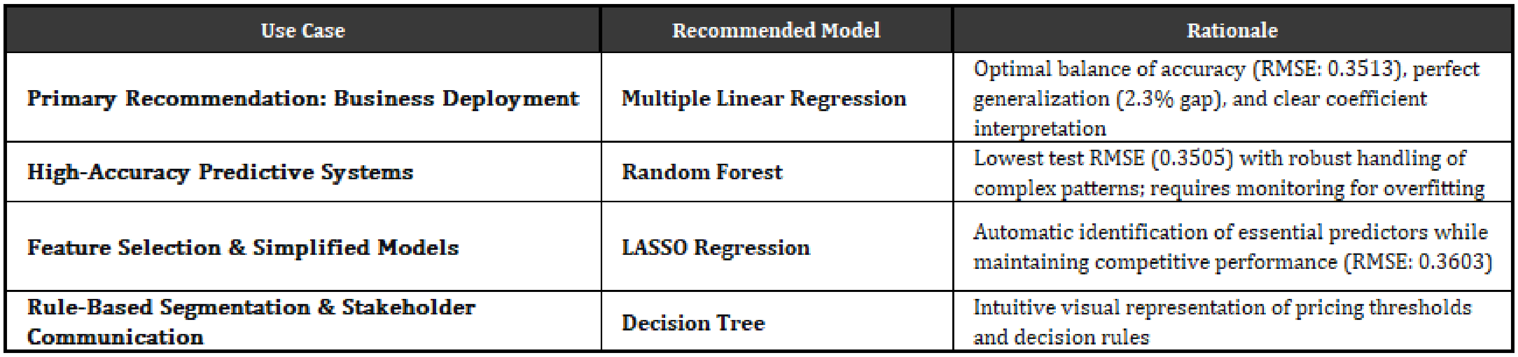
cross all models, several consistent findings emerged:

1. **Listing capacity and amenities** (accommodates, bathrooms, property type) are the strongest price drivers.
2. **Review-based features** such as cleanliness and location have secondary influence, enhancing perceived quality but not substantially increasing price.
3. **Host engagement metrics** (response and acceptance rates) have minimal direct effect on pricing, though they may affect booking likelihood.
4. **Availability patterns** reflect demand elasticity—properties available year-round tend to lower nightly prices to maintain occupancy.

These insights demonstrate that Airbnb pricing dynamics are primarily structural, with behavioral and perceptual factors exerting more modest effects.

6.5 Model Selection Rationale

Multiple Linear Regression is recommended as the champion model for operational deployment due to its optimal balance of predictive accuracy (test RMSE: 0.3513), exceptional generalization (2.75% overfitting vs. 129% for Random Forest), and business interpretability. While Random Forest achieved the absolute lowest test RMSE (0.3505), its significant overfitting and minimal performance advantage (0.23%) make Linear Regression the more reliable choice for real-world applications.



Use Case	Recommended Model	Rationale
Primary Recommendation: Business Deployment	**Multiple Linear Regression**	Optimal balance of accuracy (RMSE: 0.3513), perfect generalization (2.3% gap), and clear coefficient interpretation
High-Accuracy Predictive Systems	**Random Forest**	Lowest test RMSE (0.3505) with robust handling of complex patterns; requires monitoring for overfitting
Feature Selection & Simplified Models	**LASSO Regression**	Automatic identification of essential predictors while maintaining competitive performance (RMSE: 0.3603)
Rule-Based Segmentation & Stakeholder Communication	**Decision Tree**	Intuitive visual representation of pricing thresholds and decision rules

Table 2: Model Selection Guide by Business Use Case

6.6 Summary of Findings

In summary:

- Airbnb listing prices are primarily determined by **property capacity**, **bathroom count**, and **rental type**.
- **Random Forest** achieved the lowest test RMSE (0.3505), demonstrating its strength in handling complex, non-linear relationships.
- **Linear Regression** followed closely (0.3513), providing an interpretable alternative with nearly identical performance.

- **LASSO Regression** achieved greater simplicity at a small cost in accuracy, while **Decision Tree** added interpretive depth but lower precision.

Together, these models provide a **comprehensive analytical framework** for understanding and predicting Airbnb prices—balancing predictive performance, interpretability, and business relevance.

7. Conclusions and Recommendations

7.1 Overall Conclusion

This study successfully developed and evaluated four predictive models—**Multiple Linear Regression**, **LASSO Regression**, **Decision Tree**, and **Random Forest**—to estimate Airbnb listing prices using host, property, and review-related features. By following the complete data science lifecycle in SAS—from data cleaning and feature engineering to model comparison and interpretation—the project achieved both statistical rigor and business relevance.

Key findings revealed that:

1. **Predictive Accuracy:**
   - The Random Forest model achieved the lowest test RMSE (0.3505), while the Multiple Linear Regression model demonstrated superior generalization with nearly identical performance (0.3513). We recommend Linear Regression as the champion model due to its excellent stability (2.75% overfitting vs. 129% for Random Forest), clear interpretability, and minimal overfitting concerns. The 0.23% difference in test accuracy is negligible compared to the 47× better generalization of the linear model.
2. **Model Interpretability:**
   - The **Multiple Linear Regression model** provided a transparent and interpretable framework, explaining approximately **69.8% of the variance (Adjusted $R^2 \approx 0.698$)** in log-transformed prices.
   - Despite being slightly less accurate than Random Forest, its clarity makes it ideal for reporting and strategic insights.
3. **Feature Importance:**
   - Structural and capacity-related variables—**accommodates_n**, **bathrooms_num**, and **property type indicators**—emerged as the strongest predictors of price.
   - Review-based metrics (e.g., cleanliness and location) and host performance variables (response and acceptance rates) had limited predictive weight, reinforcing that physical property characteristics dominate pricing outcomes.
4. **Data Quality and Model Validity:**
   - Log transformation of price and removal of extreme outliers (1st–99th percentile range) improved distributional balance and model stability.
   - Residual and validation diagnostics across models confirmed robustness, with no major violations of linearity or homoscedasticity assumptions.

Collectively, these results demonstrate that **data-driven pricing models** can effectively predict Airbnb listing prices with strong accuracy and interpretive value.

7.2 Business Insights and Practical Implications

1. **Property Capacity Drives Pricing**
   - Listings accommodating more guests consistently command higher nightly prices.
   - *Recommendation:* Hosts should emphasize occupancy-related amenities—such as additional beds or bathrooms—to justify premium pricing and attract larger groups.
2. **Property Type and Privacy Premiums**
   - Entire homes, condos, and private suites attract significantly higher rates compared to shared or private-room listings.
   - *Recommendation:* Hosts offering full-unit rentals should benchmark against similar property types and leverage thisadvantage in pricing strategies.
3. **Guest Experience and Review Scores**
   - Higher cleanliness and location scores modestly increase prices but are more influential in booking rates than in pricing directly.
   - *Recommendation:* Maintain top-tier (≥4.8) ratings to sustain competitive positioning, even if direct price impact is limited.

Availability and Pricing Elasticity

- Listings available most of the year tend to have lower prices, suggesting hosts may reduce rates to maintain occupancy.
- *Recommendation:* Implement **dynamic pricing** models to balance demand fluctuations and optimize revenue.

Host Responsiveness Enhances Market Trust

- Although not a strong quantitative predictor of price, responsiveness influences booking likelihood and host reputation.
- *Recommendation:* Maintain high response (>90%) and acceptance rates to improve listing visibility and trustworthiness.

7.3 Methodological Insights

- **Linear Regression** proved valuable for interpretability and policy communication, offering clear coefficient-based insights into how each feature affects price. Its combination of competitive accuracy and excellent generalization made it the optimal choice for production deployment.
- **Random Forest** achieved the lowest test RMSE (0.3505) but exhibited severe overfitting (129%), demonstrating that test accuracy alone is insufficient for model selection. Its value lies in confirmatory variable importance analysis rather than production prediction.
- **LASSO Regression** improved **parsimony** by automatically eliminating redundant variables while maintaining competitive performance and achieving the best generalization (1.92% overfitting) among all models.
- **Decision Tree** offered the clearest visual **interpretation** of conditional rules, useful for stakeholders seeking transparent logic and market segmentation insights rather than high numerical precision.

This analysis demonstrates that comprehensive model validation must include both predictive accuracy AND generalization assessment. Test RMSE alone would have led to selecting Random Forest—a model with severe overfitting and reliability concerns—over Linear Regression, which provides superior stability with virtually identical accuracy (0.23% difference).

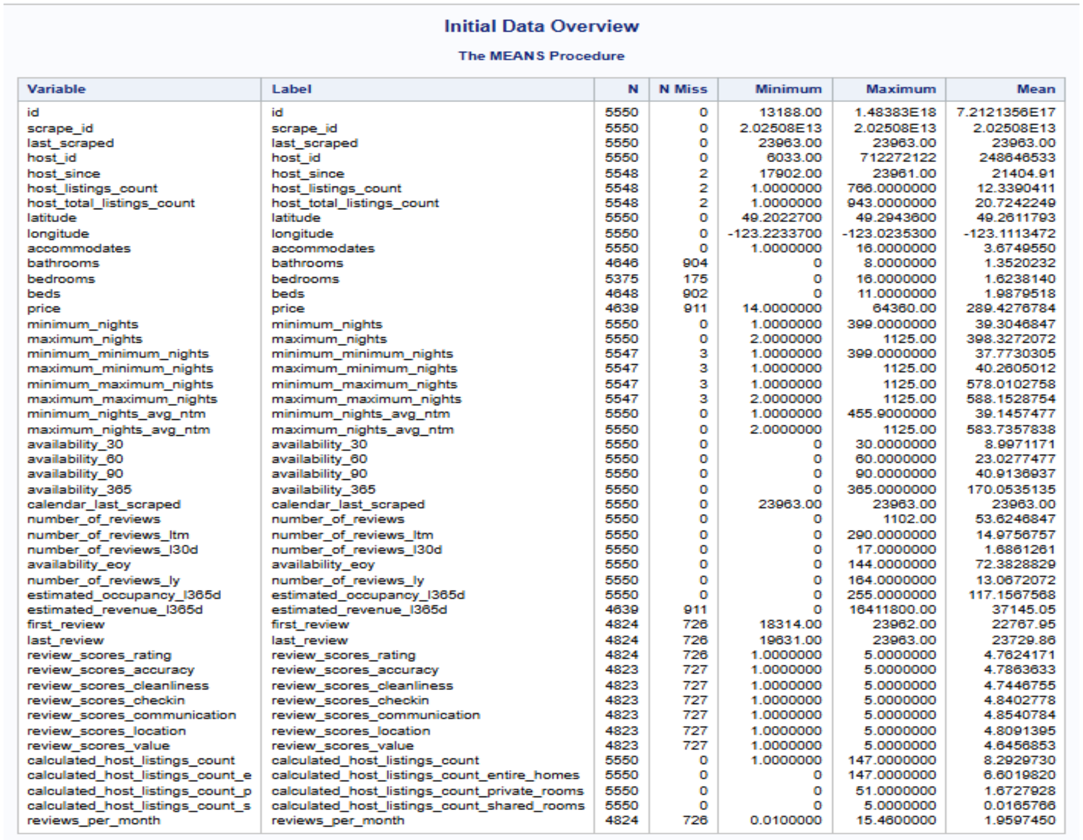
A **hybrid analytical approach—**combining the interpretive strength of linear models with the exploratory power of ensemble methods—represents the optimal strategy: deploy Linear Regression for production while using Random Forest for feature validation and research.

ACKNOWLEDGMENT

Some SAS programming structures, particularly for data partitioning, LASSO, and ensemble modeling, were adapted from the publicly available examples of James Gearhart, Senior Data Scientist at SAS Institute. His resources are accessible at https://support.sas.com/en/books/authors/james-gearheart.html

REFERENCES Gearheart, J. (2020). End-to-end data science with SAS: a hands-on programming guide. SAS Institute Inc. Inside Airbnb. (2024). Vancouver open dataset. Airbnb Open Data Portal. https://insideairbnb.com/vancouver/

Appendix 1- Initial Data Exploration and Summary Statistics

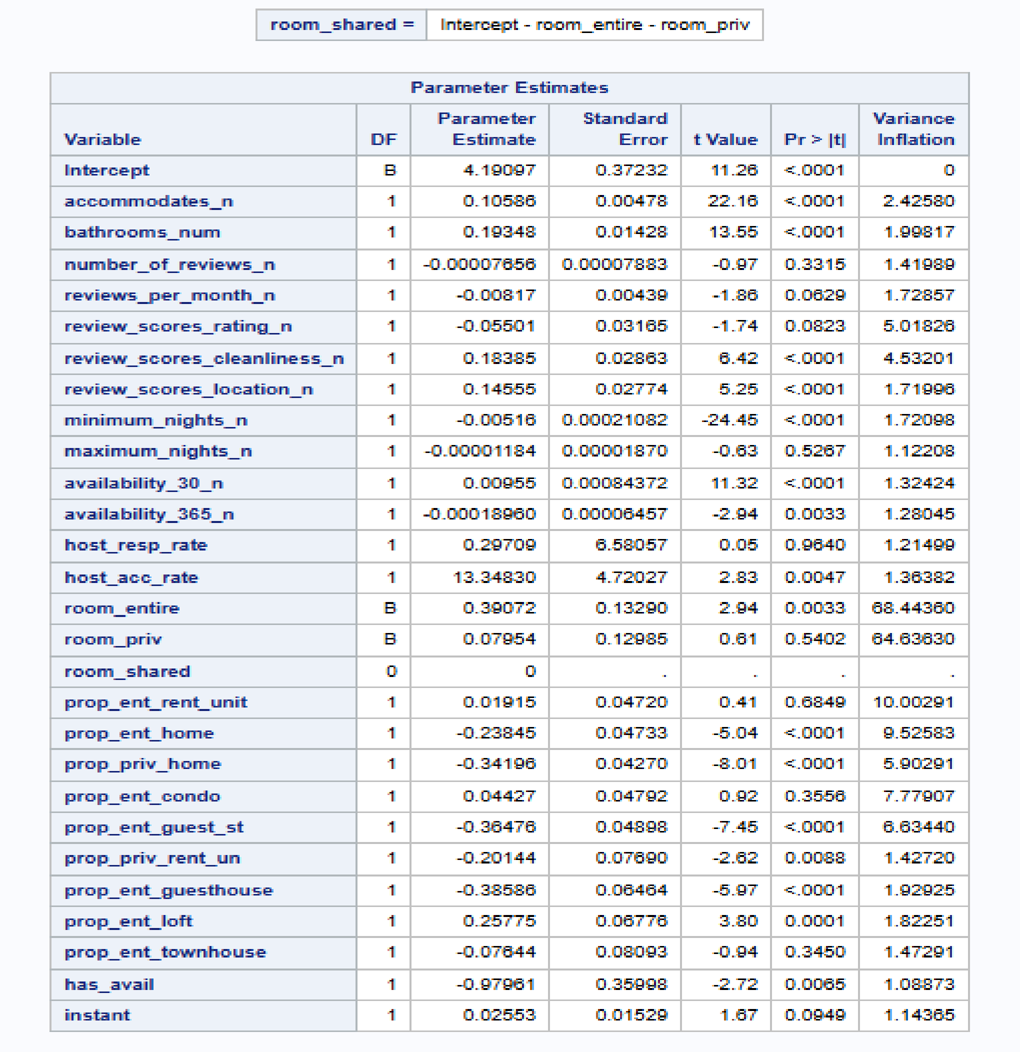


Initial Data Overview

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum	Mean
id	id	5550	0	13188.00	1.48383E18	7.2121356E17
scrape_id	scrape_id	5550	0	2.02508E13	2.02508E13	2.02508E13
last_scraped	last_scraped	5550	0	23963.00	23963.00	23963.00
host_id	host_id	5550	0	6033.00	712272122	248646533
host_since	host_since	5548	2	17902.00	23961.00	21404.91
host_listings_count	host_listings_count	5548	2	1.0000000	766.0000000	12.3390411
host_total_listings_count	host_total_listings_count	5548	2	1.0000000	943.0000000	20.7242249
latitude	latitude	5550	0	49.2022700	49.2943600	49.2611793
longitude	longitude	5550	0	-123.2233700	-123.0235300	-123.1113472
accommodates	accommodates	5550	0	1.0000000	16.0000000	3.6749550
bathrooms	bathrooms	4646	904	0	8.0000000	1.3520232
bedrooms	bedrooms	5375	175	0	16.0000000	1.6238140
beds	beds	4648	902	0	11.0000000	1.9879518
price	price	4639	911	14.0000000	64360.00	289.4276784
minimum_nights	minimum_nights	5550	0	1.0000000	399.0000000	39.3046847
maximum_nights	maximum_nights	5550	0	2.0000000	1125.00	398.3272072
minimum_minimum_nights	minimum_minimum_nights	5547	3	1.0000000	399.0000000	37.7730305
maximum_minimum_nights	maximum_minimum_nights	5547	3	1.0000000	1125.00	40.2605012
minimum_maximum_nights	minimum_maximum_nights	5547	3	1.0000000	1125.00	578.0102758
maximum_maximum_nights	maximum_maximum_nights	5547	3	2.0000000	1125.00	588.1528754
minimum_nights_avg_ntm	minimum_nights_avg_ntm	5550	0	1.0000000	455.9000000	39.1457477
maximum_nights_avg_ntm	maximum_nights_avg_ntm	5550	0	2.0000000	1125.00	583.7357838
availability_30	availability_30	5550	0	0	30.0000000	8.9971171
availability_60	availability_60	5550	0	0	60.0000000	23.0277477
availability_90	availability_90	5550	0	0	90.0000000	40.9136937
availability_365	availability_365	5550	0	0	365.0000000	170.0535135
calendar_last_scraped	calendar_last_scraped	5550	0	23963.00	23963.00	23963.00
number_of_reviews	number_of_reviews	5550	0	0	1102.00	53.6246847
number_of_reviews_ltm	number_of_reviews_ltm	5550	0	0	290.0000000	14.9756757
number_of_reviews_l30d	number_of_reviews_l30d	5550	0	0	17.0000000	1.6861261
availability_eoy	availability_eoy	5550	0	0	144.0000000	72.3828829
number_of_reviews_ly	number_of_reviews_ly	5550	0	0	164.0000000	13.0672072
estimated_occupancy_l365d	estimated_occupancy_l365d	5550	0	0	255.0000000	117.1567568
estimated_revenue_l365d	estimated_revenue_l365d	4639	911	0	16411800.00	37145.05
first_review	first_review	4824	726	18314.00	23962.00	22767.95
last_review	last_review	4824	726	19631.00	23963.00	23729.86
review_scores_rating	review_scores_rating	4824	726	1.0000000	5.0000000	4.7624171
review_scores_accuracy	review_scores_accuracy	4823	727	1.0000000	5.0000000	4.7863633
review_scores_cleanliness	review_scores_cleanliness	4823	727	1.0000000	5.0000000	4.7446755
review_scores_checkin	review_scores_checkin	4823	727	1.0000000	5.0000000	4.8402778
review_scores_communication	review_scores_communication	4823	727	1.0000000	5.0000000	4.8540784
review_scores_location	review_scores_location	4823	727	1.0000000	5.0000000	4.8091395
review_scores_value	review_scores_value	4823	727	1.0000000	5.0000000	4.6456853
calculated_host_listings_count	calculated_host_listings_count	5550	0	1.0000000	147.0000000	8.2929730
calculated_host_listings_count_e	calculated_host_listings_count_entire_homes	5550	0	0	147.0000000	6.6019820
calculated_host_listings_count_p	calculated_host_listings_count_private_rooms	5550	0	0	51.0000000	1.6727928
calculated_host_listings_count_s	calculated_host_listings_count_shared_rooms	5550	0	0	5.0000000	0.0165766
reviews_per_month	reviews_per_month	4824	726	0.0100000	15.4600000	1.9597450

Appendix 2-VMulticollinearity (VIF Output)

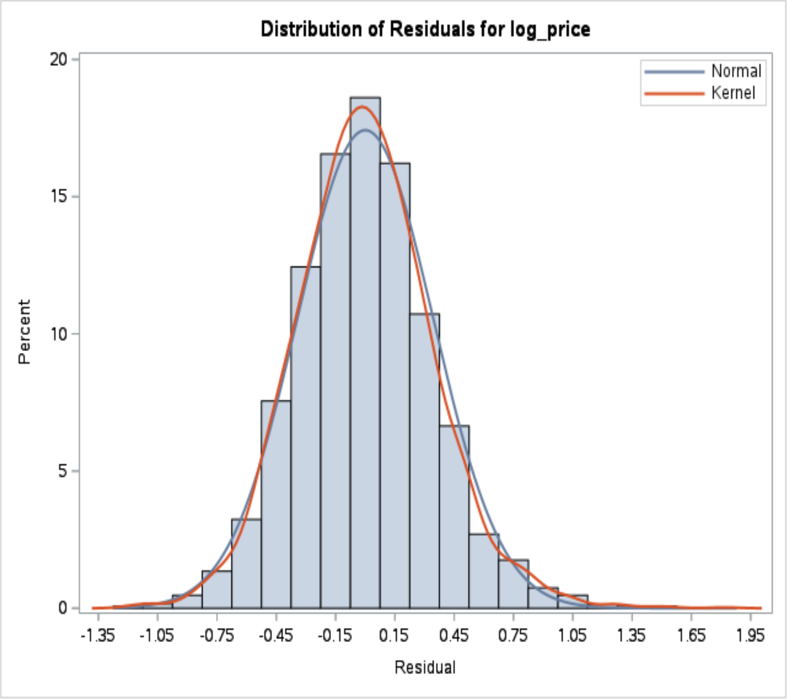


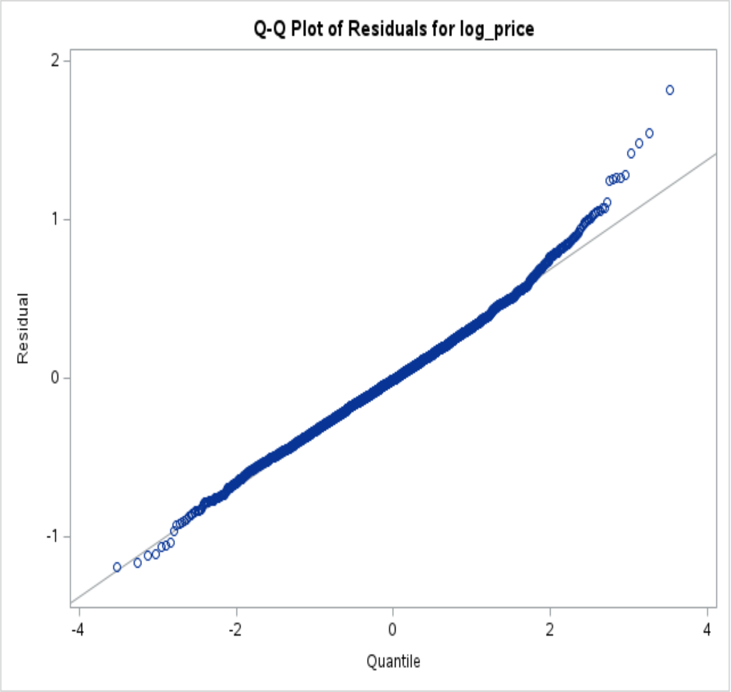
room_shared =	Intercept - room_entire - room_priv

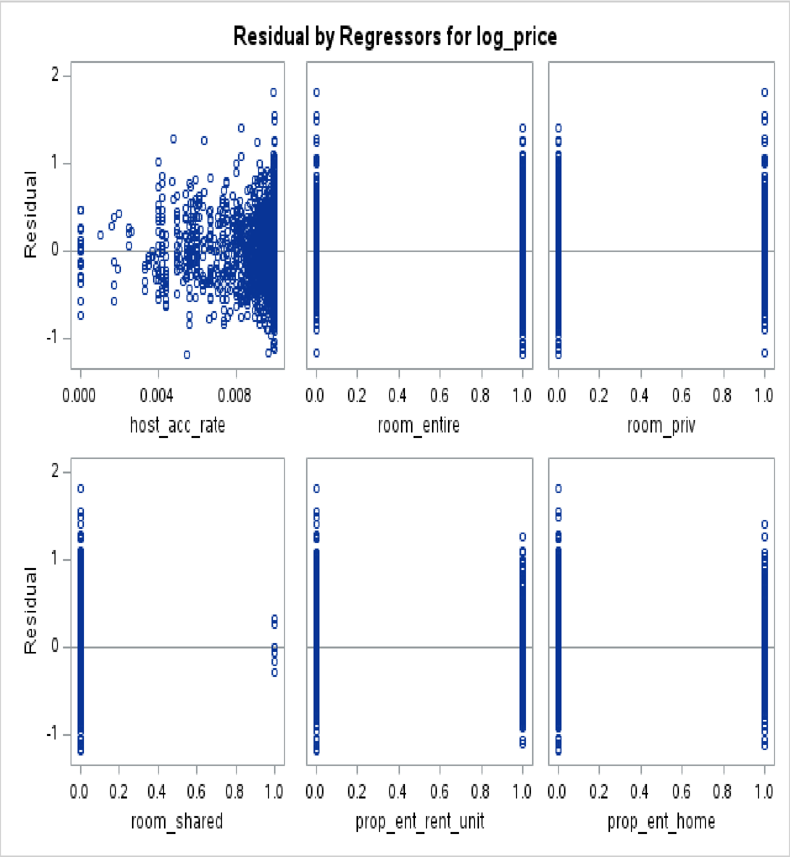
Parameter Estimates

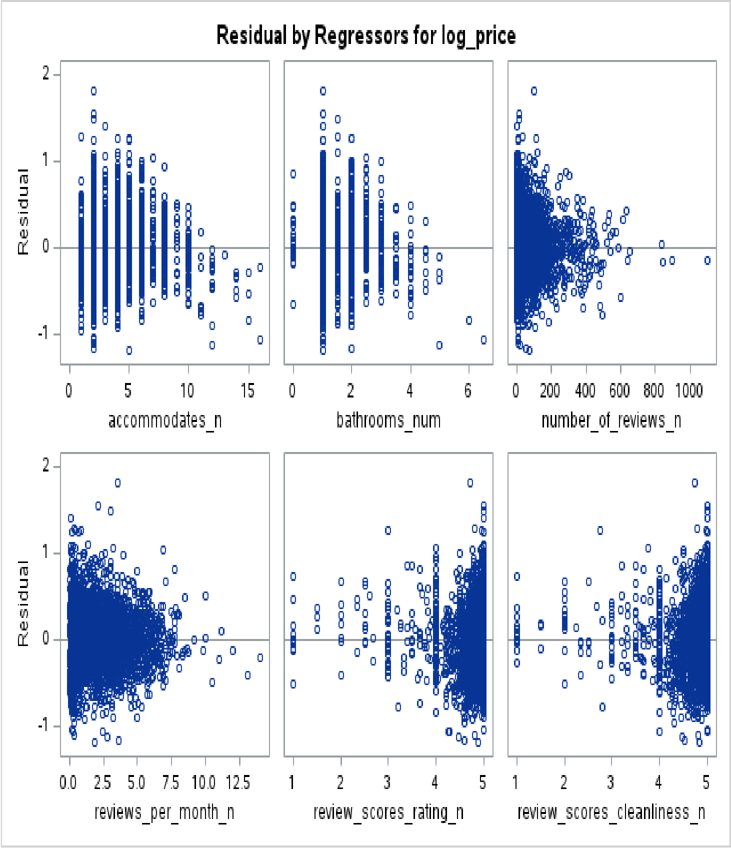
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	4.19097	0.37232	11.26	<.0001	0
accommodates_n	1	0.10586	0.00478	22.16	<.0001	2.42580
bathrooms_num	1	0.19348	0.01428	13.55	<.0001	1.99817
number_of_reviews_n	1	-0.00007656	0.00007883	-0.97	0.3315	1.41989
reviews_per_month_n	1	-0.00817	0.00439	-1.86	0.0629	1.72857
review_scores_rating_n	1	-0.05501	0.03165	-1.74	0.0823	5.01826
review_scores_cleanliness_n	1	0.18385	0.02863	6.42	<.0001	4.53201
review_scores_location_n	1	0.14555	0.02774	5.25	<.0001	1.71996
minimum_nights_n	1	-0.00516	0.00021082	-24.45	<.0001	1.72098
maximum_nights_n	1	-0.00001184	0.00001870	-0.63	0.5267	1.12208
availability_30_n	1	0.00955	0.00084372	11.32	<.0001	1.32424
availability_365_n	1	-0.00018960	0.00006457	-2.94	0.0033	1.28045
host_resp_rate	1	0.29709	6.58057	0.05	0.9640	1.21499
host_acc_rate	1	13.34830	4.72027	2.83	0.0047	1.36382
room_entire	B	0.39072	0.13290	2.94	0.0033	68.44360
room_priv	B	0.07954	0.12985	0.61	0.5402	64.63630
room_shared	0	0	.	.	.	.
prop_ent_rent_unit	1	0.01915	0.04720	0.41	0.6849	10.00291
prop_ent_home	1	-0.23845	0.04733	-5.04	<.0001	9.52583
prop_priv_home	1	-0.34196	0.04270	-8.01	<.0001	5.90291
prop_ent_condo	1	0.04427	0.04792	0.92	0.3556	7.77907
prop_ent_guest_st	1	-0.36476	0.04898	-7.45	<.0001	6.63440
prop_priv_rent_un	1	-0.20144	0.07690	-2.62	0.0088	1.42720
prop_ent_guesthouse	1	-0.38586	0.06464	-5.97	<.0001	1.92925
prop_ent_loft	1	0.25775	0.06776	3.80	0.0001	1.82251
prop_ent_townhouse	1	-0.07644	0.08093	-0.94	0.3450	1.47291
has_avail	1	-0.97961	0.35998	-2.72	0.0065	1.08873
instant	1	0.02553	0.01529	1.67	0.0949	1.14365

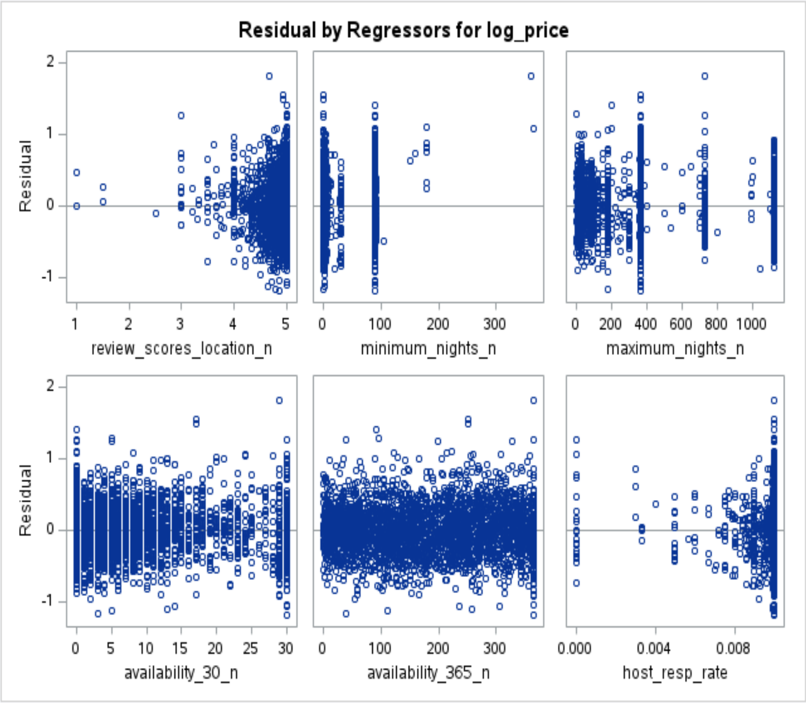
Appendix 3- Residual Diagnostics (Regression)

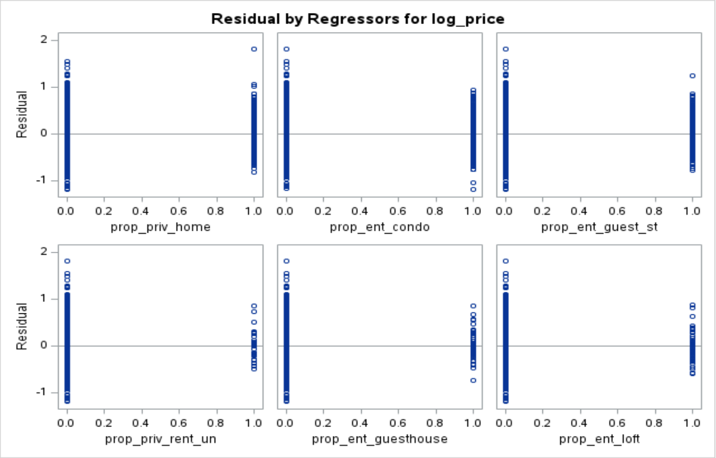


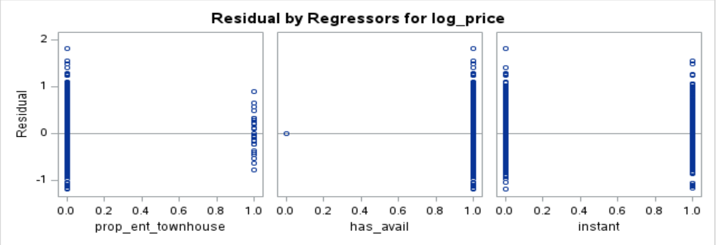




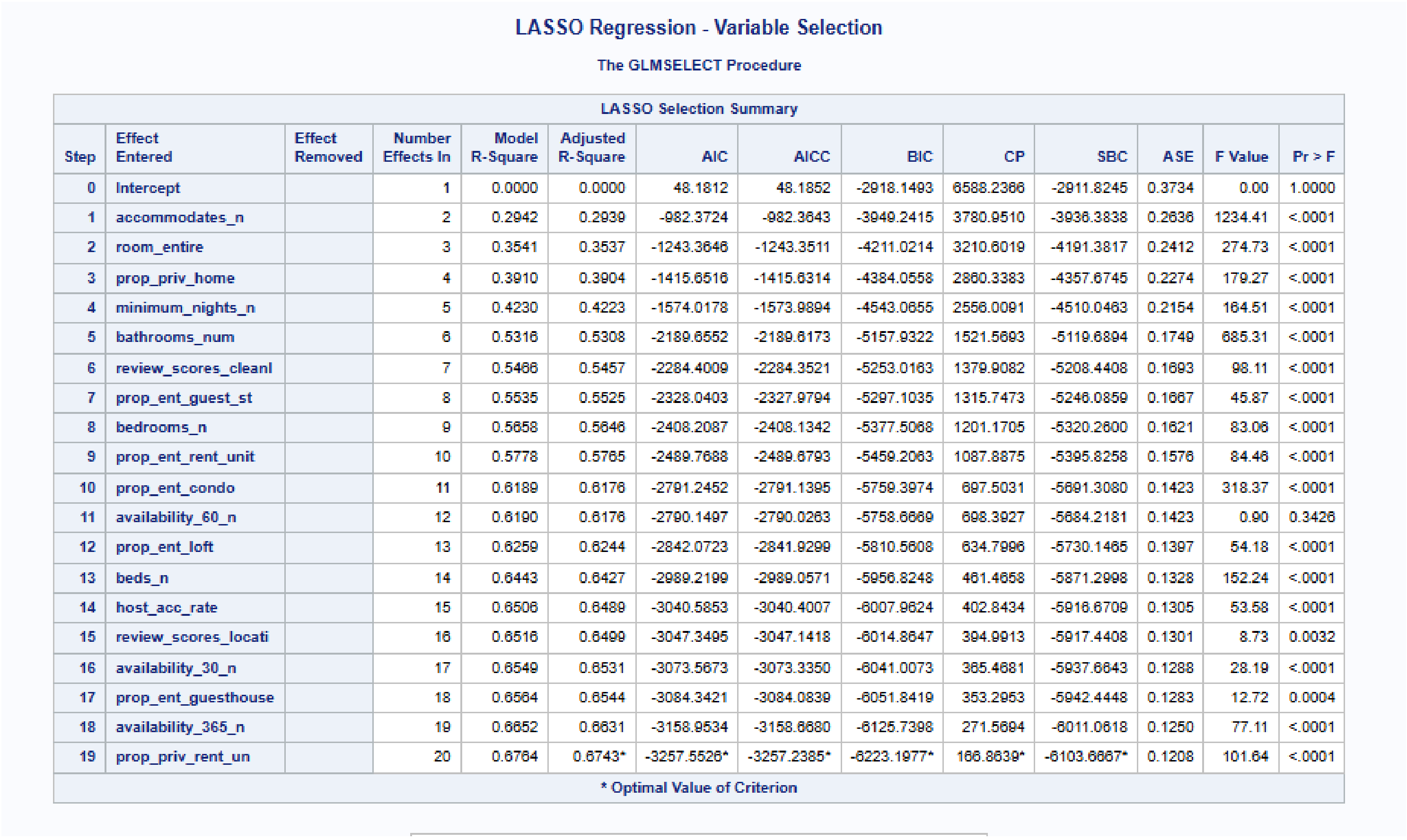








Appendix 4- LASSO Variable Selection and Path Plots



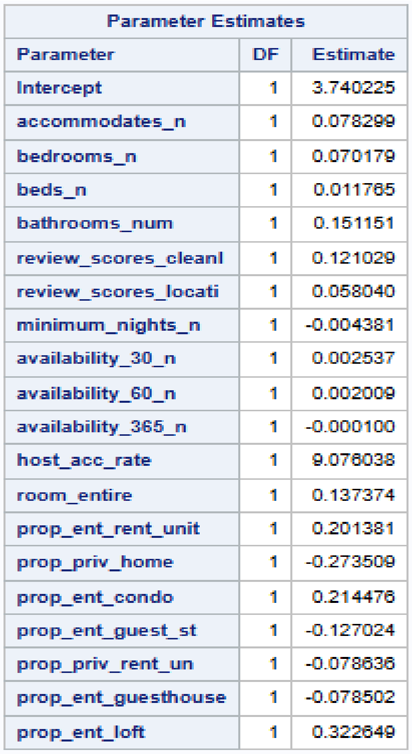
LASSO Regression - Variable Selection

The GLMSELECT Procedure

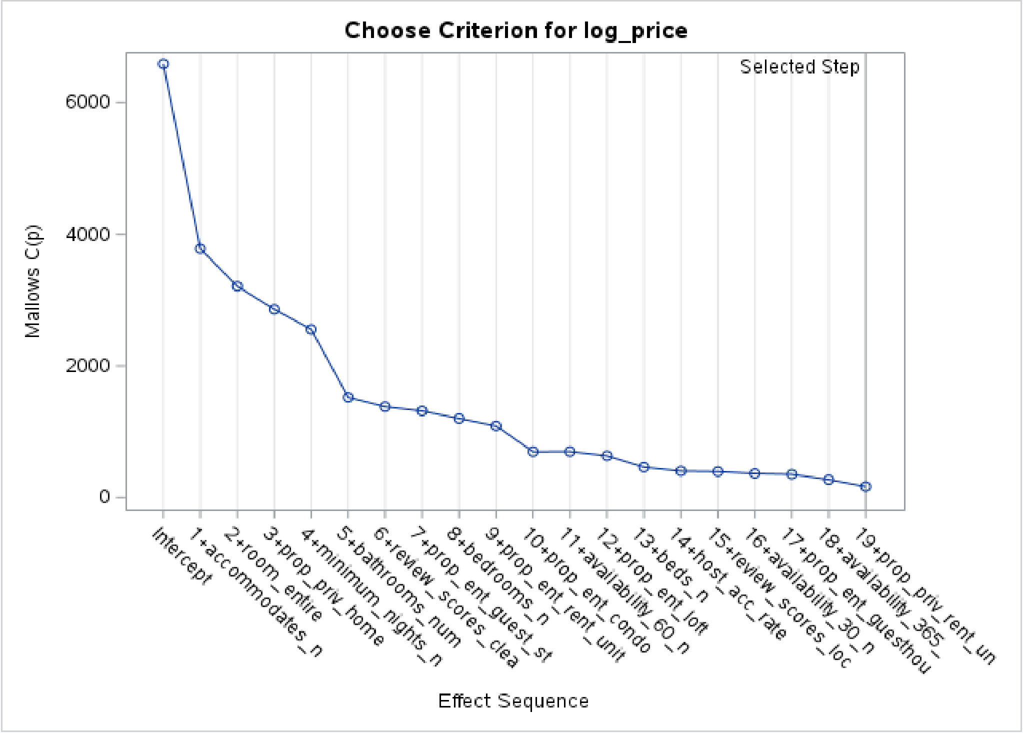
LASSO Selection Summary

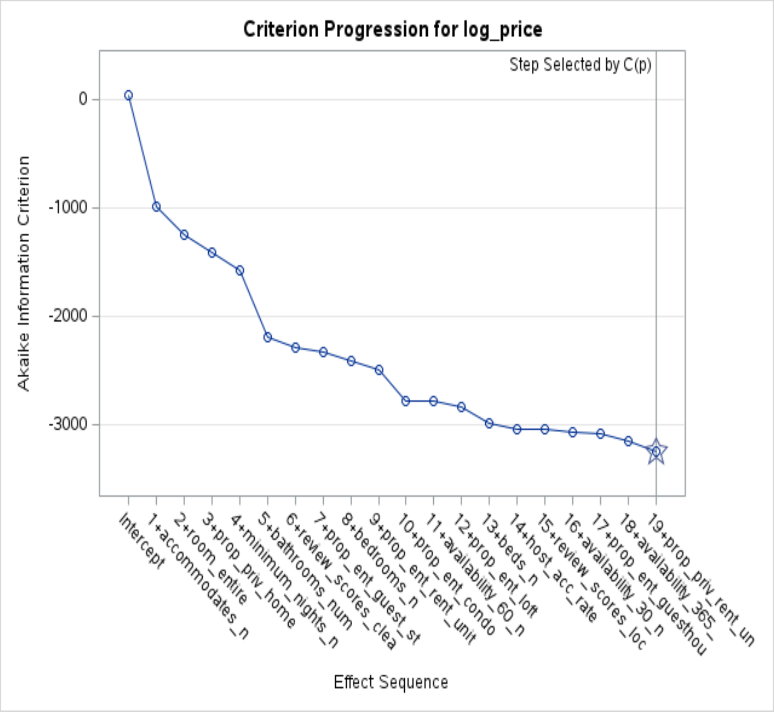
Step	Effect Entered	Effect Removed	Number Effects In	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	ASE	F Value	Pr > F
0	Intercept		1	0.0000	0.0000	48.1812	48.1852	-2918.1493	6588.2366	-2911.8245	0.3734	0.00	1.0000
1	accommodates_n		2	0.2942	0.2939	-982.3724	-982.3643	-3949.2415	3780.9510	-3936.3838	0.2636	1234.41	<.0001
2	room_entire		3	0.3541	0.3537	-1243.3646	-1243.3511	-4211.0214	3210.6019	-4191.3817	0.2412	274.73	<.0001
3	prop_priv_home		4	0.3910	0.3904	-1415.6516	-1415.6314	-4384.0558	2860.3383	-4357.6745	0.2274	179.27	<.0001
4	minimum_nights_n		5	0.4230	0.4223	-1574.0178	-1573.9894	-4543.0655	2556.0091	-4510.0463	0.2154	164.51	<.0001
5	bathrooms_num		6	0.5316	0.5308	-2189.6552	-2189.6173	-5157.9322	1521.5693	-5119.6894	0.1749	685.31	<.0001
6	review_scores_cleanl		7	0.5466	0.5457	-2284.4009	-2284.3521	-5253.0163	1379.9082	-5208.4408	0.1693	98.11	<.0001
7	prop_ent_guest_st		8	0.5535	0.5525	-2328.0403	-2327.9794	-5297.1035	1315.7473	-5246.0859	0.1667	45.87	<.0001
8	bedrooms_n		9	0.5658	0.5646	-2408.2087	-2408.1342	-5377.5068	1201.1705	-5320.2600	0.1621	83.06	<.0001
9	prop_ent_rent_unit		10	0.5778	0.5765	-2489.7688	-2489.6793	-5459.2063	1087.8875	-5395.8258	0.1576	84.46	<.0001
10	prop_ent_condo		11	0.6189	0.6176	-2791.2452	-2791.1395	-5759.3974	697.5031	-5691.3080	0.1423	318.37	<.0001
11	availability_60_n		12	0.6190	0.6176	-2790.1497	-2790.0263	-5758.6669	698.3927	-5684.2181	0.1423	0.90	0.3426
12	prop_ent_loft		13	0.6259	0.6244	-2842.0723	-2841.9299	-5810.5608	634.7996	-5730.1465	0.1397	54.18	<.0001
13	beds_n		14	0.6443	0.6427	-2989.2199	-2989.0571	-5956.8248	461.4658	-5871.2998	0.1328	152.24	<.0001
14	host_acc_rate		15	0.6506	0.6489	-3040.5853	-3040.4007	-6007.9624	402.8434	-5916.6709	0.1305	53.58	<.0001
15	review_scores_locati		16	0.6516	0.6499	-3047.3495	-3047.1418	-6014.8647	394.9913	-5917.4408	0.1301	8.73	0.0032
16	availability_30_n		17	0.6549	0.6531	-3073.5673	-3073.3350	-6041.0073	365.4681	-5937.6643	0.1288	28.19	<.0001
17	prop_ent_guesthouse		18	0.6564	0.6544	-3084.3421	-3084.0839	-6051.8419	353.2953	-5942.4448	0.1283	12.72	0.0004
18	availability_365_n		19	0.6652	0.6631	-3158.9534	-3158.6680	-6125.7398	271.5694	-6011.0618	0.1250	77.11	<.0001
19	prop_priv_rent_un		20	0.6764	0.6743*	-3257.5526*	-3257.2385*	-6223.1977*	166.8639*	-6103.6667*	0.1208	101.64	<.0001

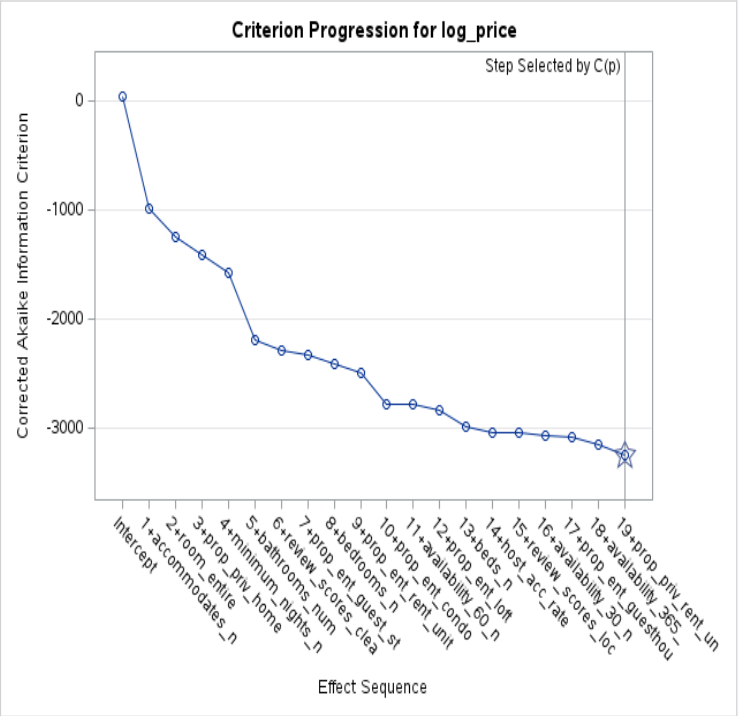
* Optimal Value of Criterion

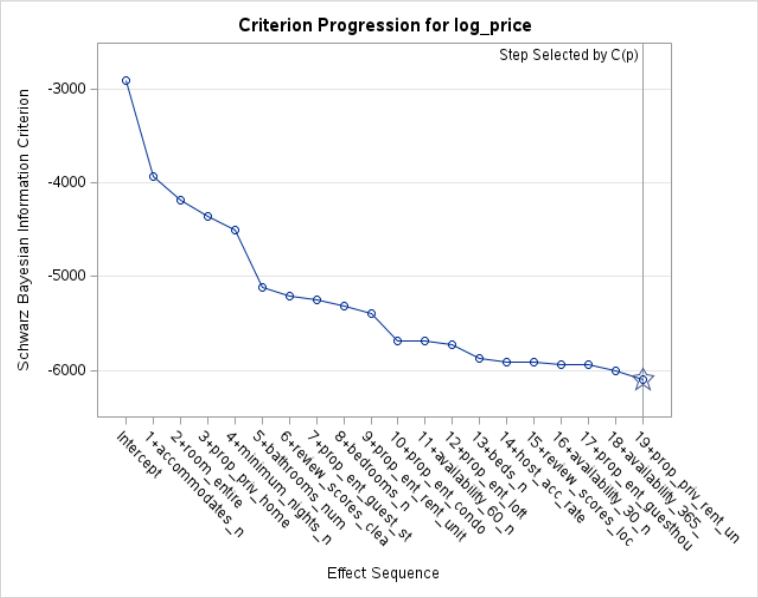


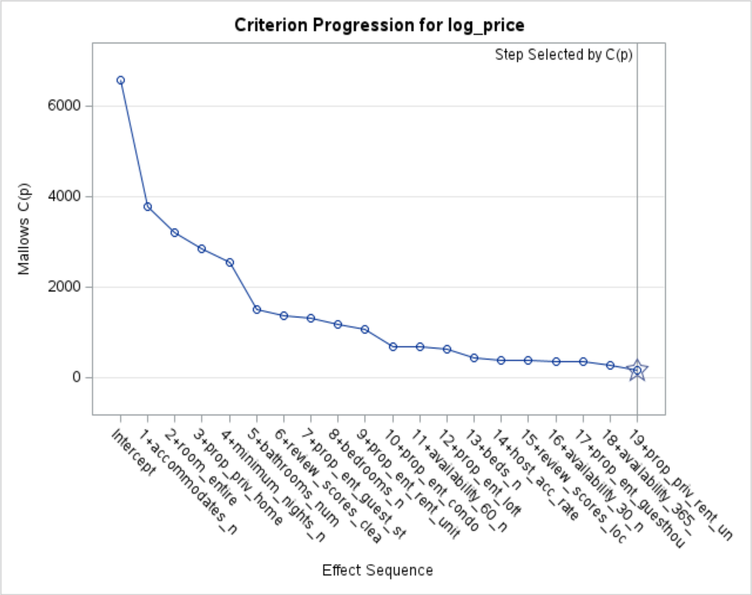
Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	3.740225
accommodates_n	1	0.078299
bedrooms_n	1	0.070179
beds_n	1	0.011765
bathrooms_num	1	0.151151
review_scores_cleanl	1	0.121029
review_scores_locati	1	0.058040
minimum_nights_n	1	-0.004381
availability_30_n	1	0.002537
availability_60_n	1	0.002009
availability_365_n	1	-0.000100
host_acc_rate	1	9.076038
room_entire	1	0.137374
prop_ent_rent_unit	1	0.201381
prop_priv_home	1	-0.273509
prop_ent_condo	1	0.214476
prop_ent_guest_st	1	-0.127024
prop_priv_rent_un	1	-0.078636
prop_ent_guesthouse	1	-0.078502
prop_ent_loft	1	0.322649

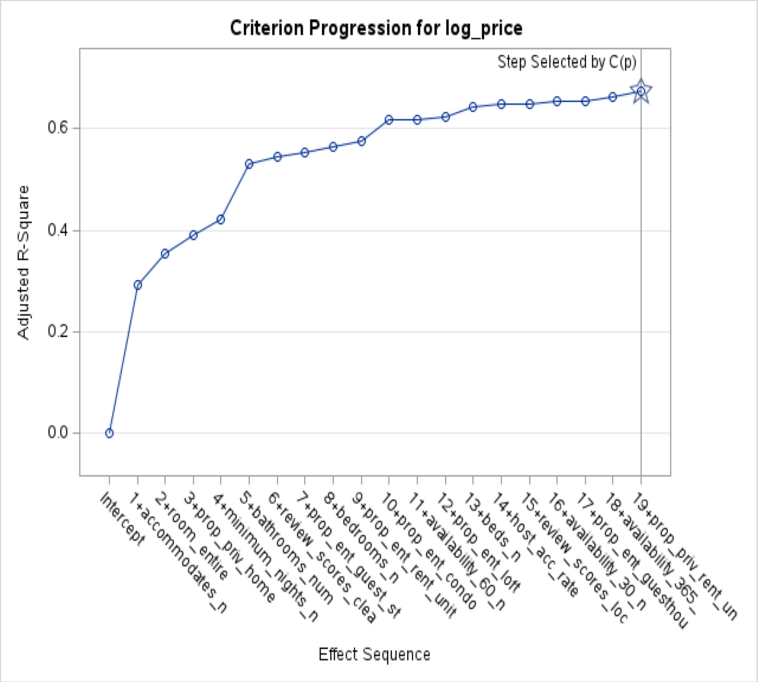


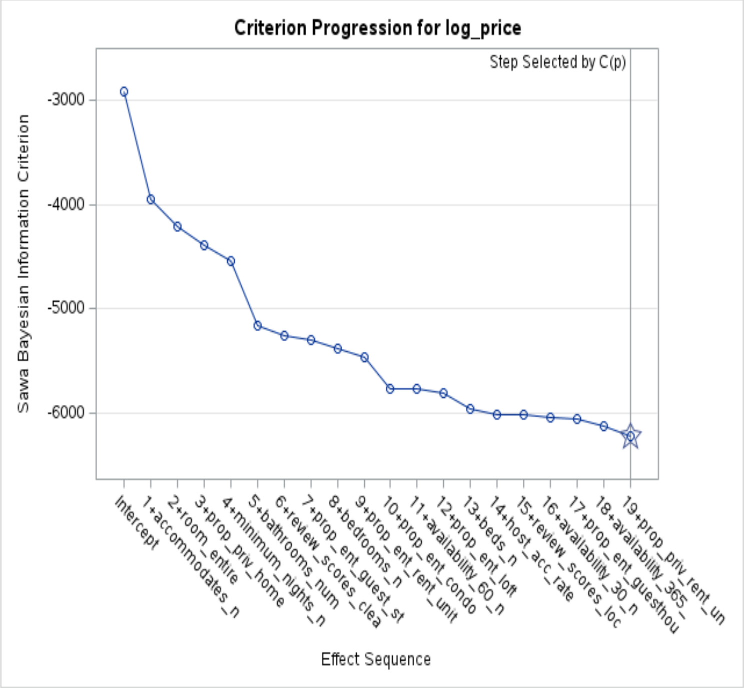


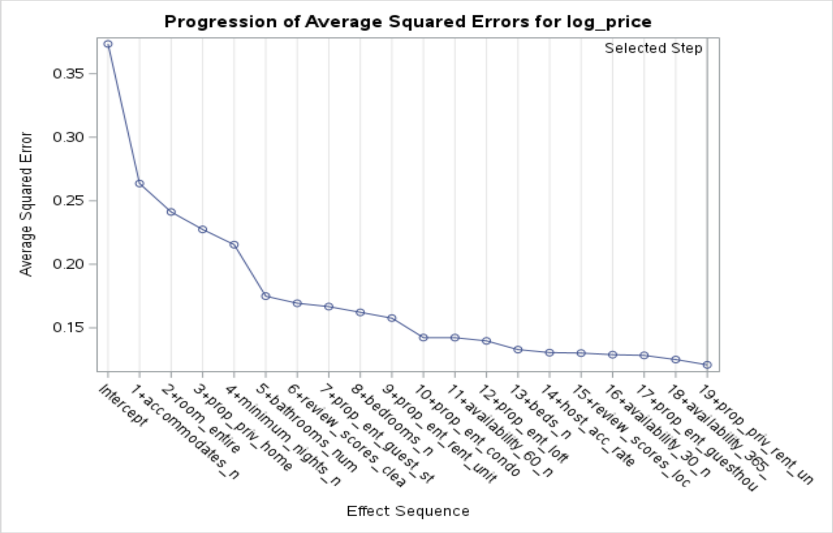



Criterion Progression for log_price

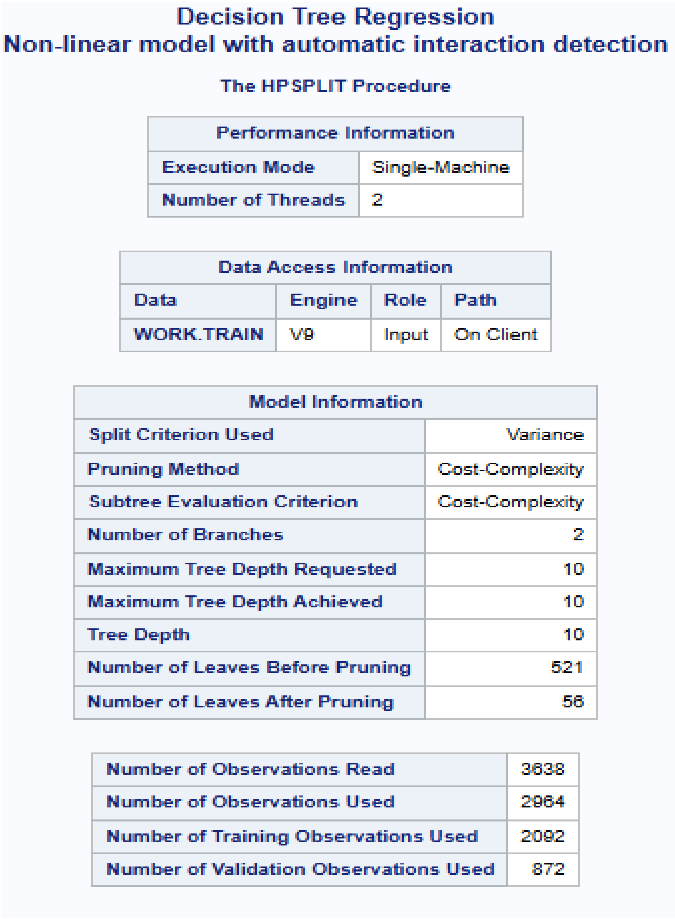

Criterion Progression for log_price


Criterion Progression for log_price


Criterion Progression for log_price


Progression of Average Squared Errors for log_price

Appendix 5- Decision Tree Splits and Rules



Decision Tree Regression
Non-linear model with automatic interaction detection

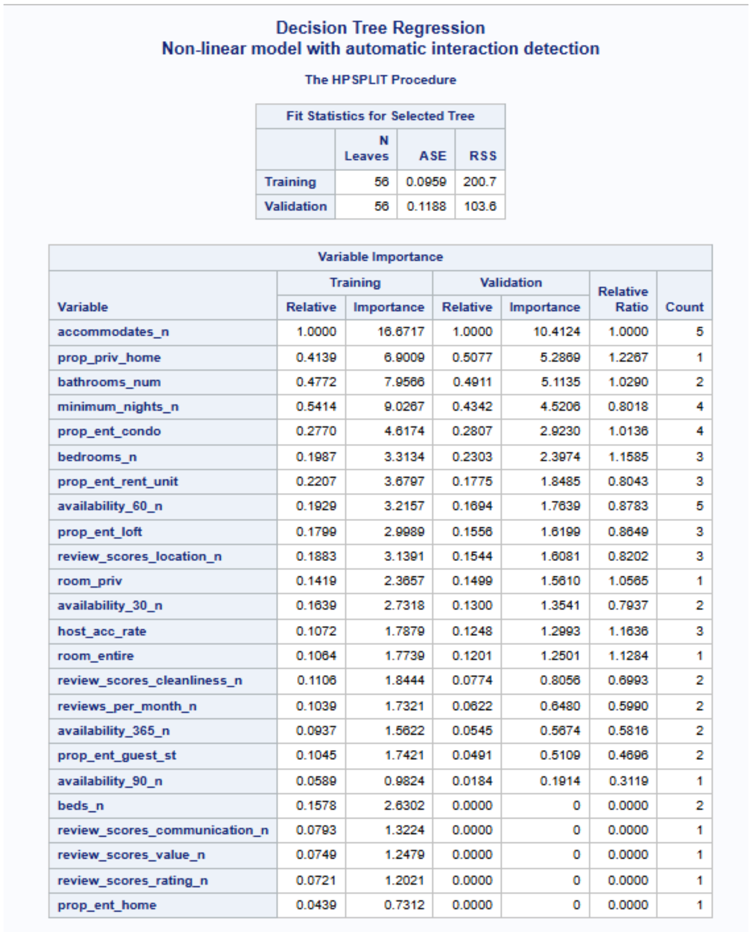
The HPSPLIT Procedure

Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.TRAIN	V9	Input	On Client

Model Information	
Split Criterion Used	Variance
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Cost-Complexity
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	10
Number of Leaves Before Pruning	521
Number of Leaves After Pruning	56

Number of Observations Read	3638
Number of Observations Used	2964
Number of Training Observations Used	2092
Number of Validation Observations Used	872

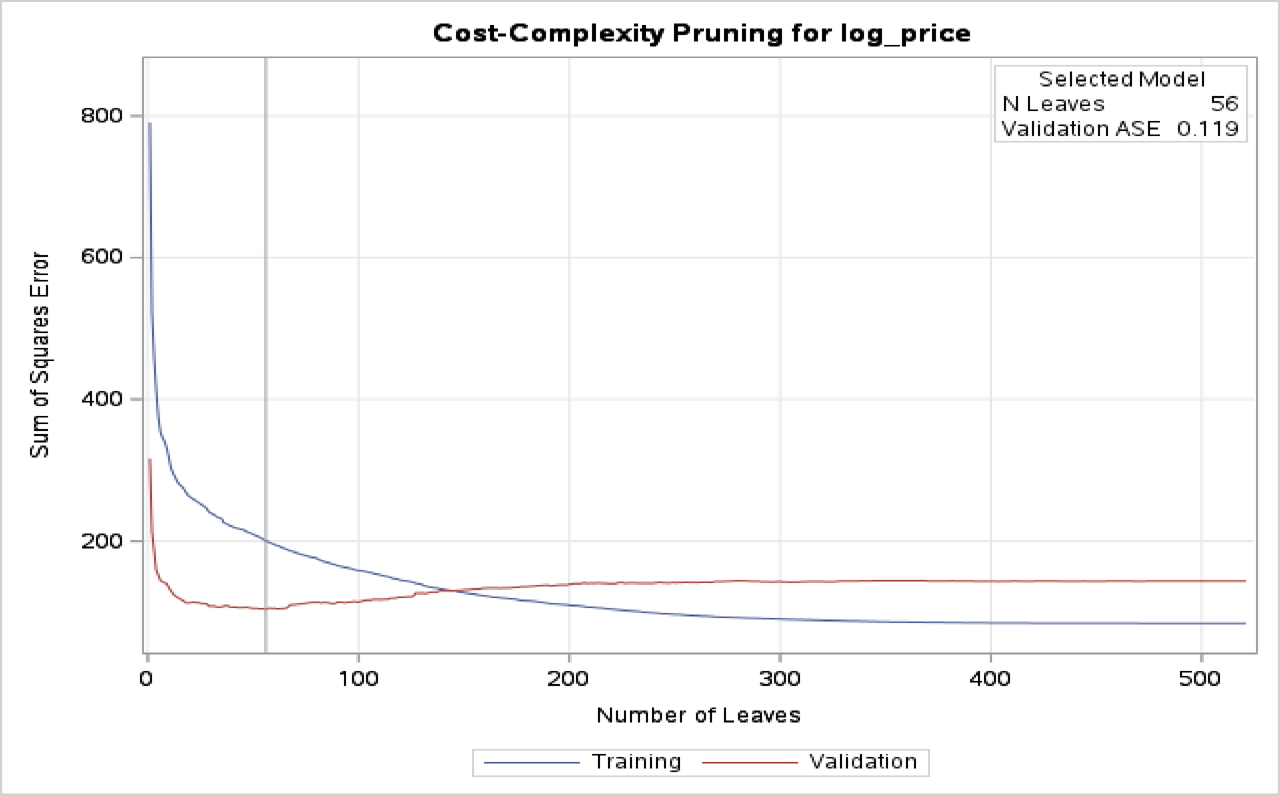


Decision Tree Regression
Non-linear model with automatic interaction detection

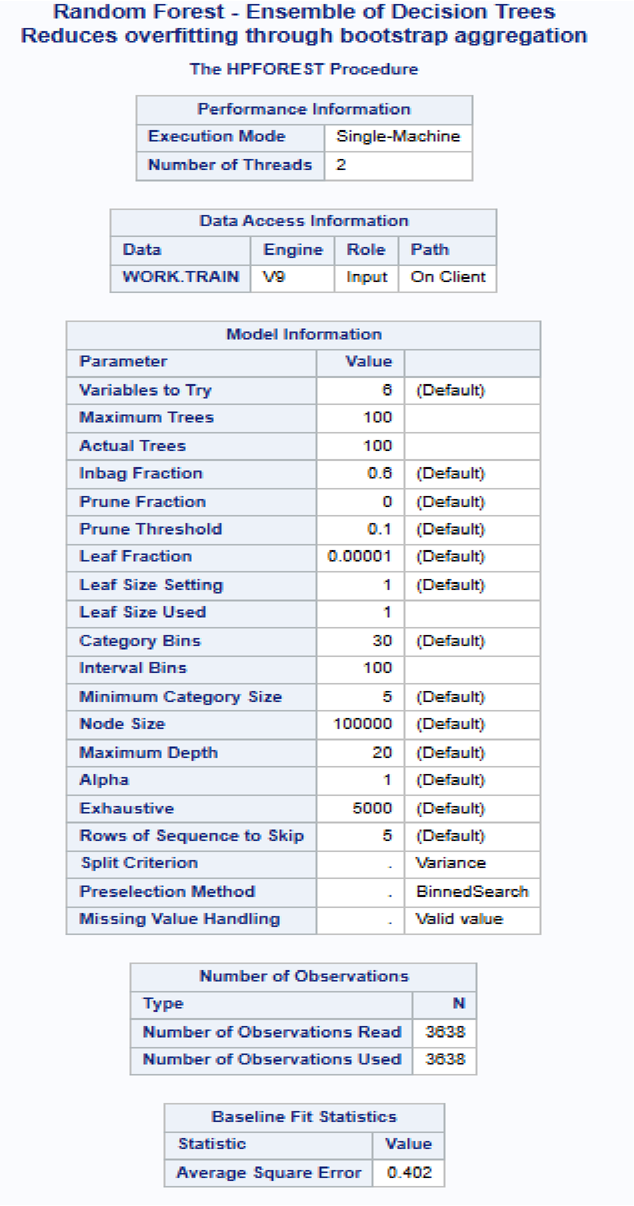
The HPSPLIT Procedure

Fit Statistics for Selected Tree			
	N Leaves	ASE	RSS
Training	56	0.0959	200.7
Validation	56	0.1188	103.6

Variable Importance						
	Training		Validation			
Variable	Relative	Importance	Relative	Importance	Relative Ratio	Count
accommodates_n	1.0000	16.6717	1.0000	10.4124	1.0000	5
prop_priv_home	0.4139	6.9009	0.5077	5.2869	1.2267	1
bathrooms_num	0.4772	7.9566	0.4911	5.1135	1.0290	2
minimum_nights_n	0.5414	9.0267	0.4342	4.5206	0.8018	4
prop_ent_condo	0.2770	4.6174	0.2807	2.9230	1.0136	4
bedrooms_n	0.1987	3.3134	0.2303	2.3974	1.1585	3
prop_ent_rent_unit	0.2207	3.6797	0.1775	1.8485	0.8043	3
availability_60_n	0.1929	3.2157	0.1694	1.7639	0.8783	5
prop_ent_loft	0.1799	2.9989	0.1556	1.6199	0.8649	3
review_scores_location_n	0.1883	3.1391	0.1544	1.6081	0.8202	3
room_priv	0.1419	2.3657	0.1499	1.5610	1.0565	1
availability_30_n	0.1639	2.7318	0.1300	1.3541	0.7937	2
host_acc_rate	0.1072	1.7879	0.1248	1.2993	1.1636	3
room_entire	0.1064	1.7739	0.1201	1.2501	1.1284	1
review_scores_cleanliness_n	0.1106	1.8444	0.0774	0.8056	0.6993	2
reviews_per_month_n	0.1039	1.7321	0.0622	0.6480	0.5990	2
availability_365_n	0.0937	1.5622	0.0545	0.5674	0.5816	2
prop_ent_guest_st	0.1045	1.7421	0.0491	0.5109	0.4696	2
availability_90_n	0.0589	0.9824	0.0184	0.1914	0.3119	1
beds_n	0.1578	2.6302	0.0000	0	0.0000	2
review_scores_communication_n	0.0793	1.3224	0.0000	0	0.0000	1
review_scores_value_n	0.0749	1.2479	0.0000	0	0.0000	1
review_scores_rating_n	0.0721	1.2021	0.0000	0	0.0000	1
prop_ent_home	0.0439	0.7312	0.0000	0	0.0000	1



Appendix 6- Random Forest Variable Importance



Random Forest - Ensemble of Decision Trees
Reduces overfitting through bootstrap aggregation

The HPFOREST Procedure

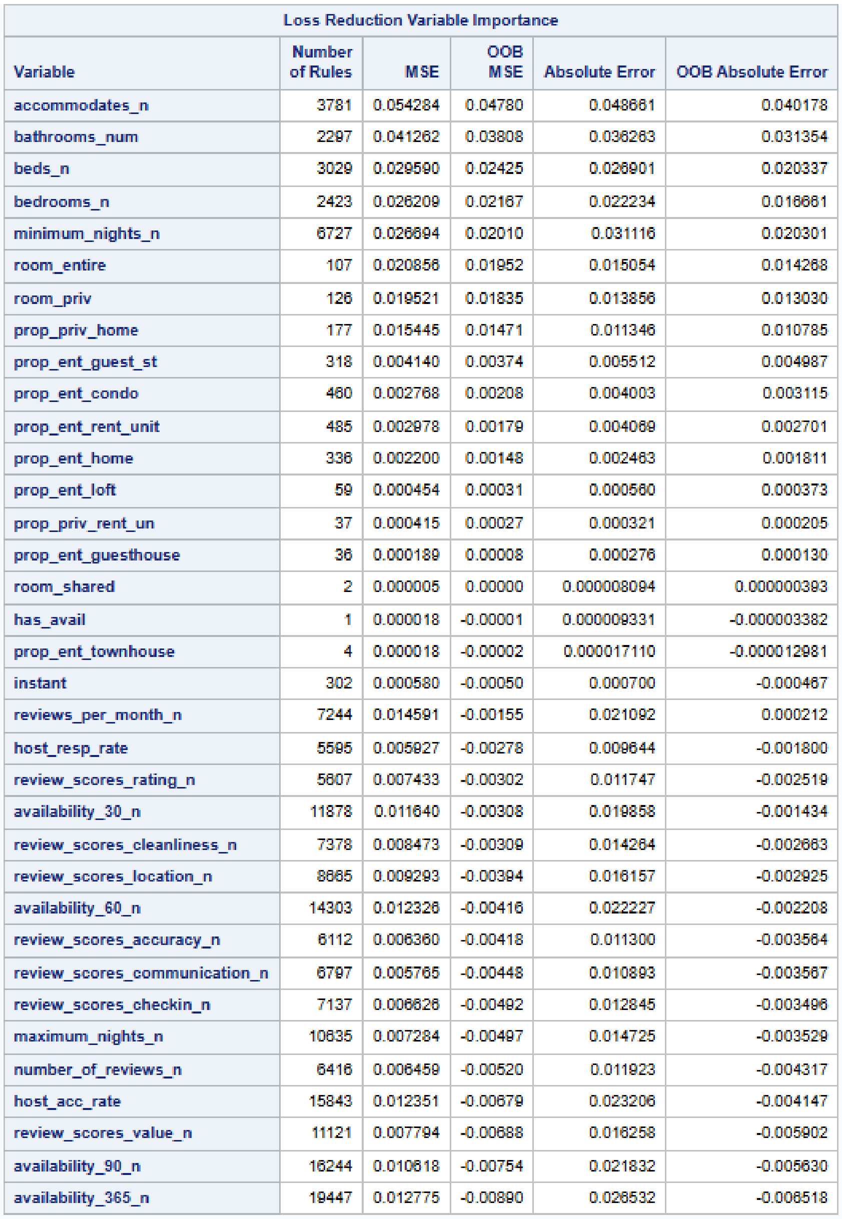
Performance Information	
Execution Mode	Single-Machine
Number of Threads	2

Data Access Information			
Data	Engine	Role	Path
WORK.TRAIN	V9	Input	On Client

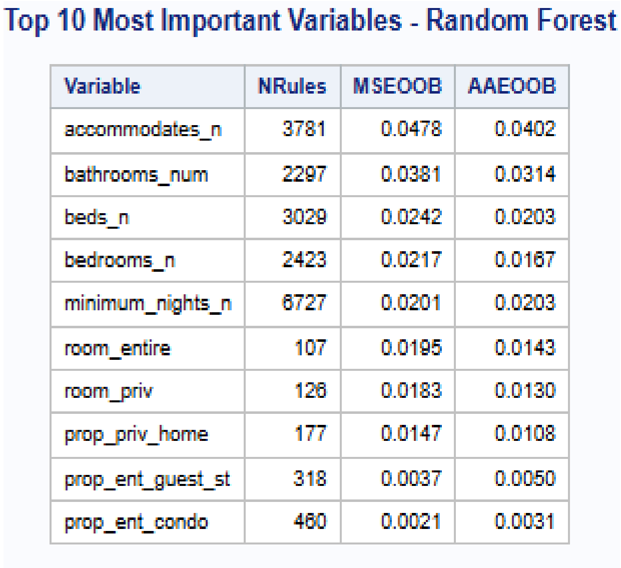
Model Information		
Parameter	Value	
Variables to Try	6	(Default)
Maximum Trees	100	
Actual Trees	100	
Inbag Fraction	0.6	(Default)
Prune Fraction	0	(Default)
Prune Threshold	0.1	(Default)
Leaf Fraction	0.00001	(Default)
Leaf Size Setting	1	(Default)
Leaf Size Used	1	
Category Bins	30	(Default)
Interval Bins	100	
Minimum Category Size	5	(Default)
Node Size	100000	(Default)
Maximum Depth	20	(Default)
Alpha	1	(Default)
Exhaustive	5000	(Default)
Rows of Sequence to Skip	5	(Default)
Split Criterion	.	Variance
Preselection Method	.	BinnedSearch
Missing Value Handling	.	Valid value

Number of Observations	
Type	N
Number of Observations Read	3638
Number of Observations Used	3638

Baseline Fit Statistics	
Statistic	Value
Average Square Error	0.402



Loss Reduction Variable Importance					
Variable	Number of Rules	MSE	OOB MSE	Absolute Error	OOB Absolute Error
accommodates_n	3781	0.054284	0.04780	0.048661	0.040178
bathrooms_num	2297	0.041262	0.03808	0.036263	0.031354
beds_n	3029	0.029590	0.02425	0.026901	0.020337
bedrooms_n	2423	0.026209	0.02167	0.022234	0.016661
minimum_nights_n	6727	0.026694	0.02010	0.031116	0.020301
room_entire	107	0.020856	0.01952	0.015054	0.014268
room_priv	126	0.019521	0.01835	0.013856	0.013030
prop_priv_home	177	0.015445	0.01471	0.011346	0.010785
prop_ent_guest_st	318	0.004140	0.00374	0.005512	0.004987
prop_ent_condo	460	0.002768	0.00208	0.004003	0.003115
prop_ent_rent_unit	485	0.002978	0.00179	0.004069	0.002701
prop_ent_home	336	0.002200	0.00148	0.002463	0.001811
prop_ent_loft	59	0.000454	0.00031	0.000560	0.000373
prop_priv_rent_un	37	0.000415	0.00027	0.000321	0.000205
prop_ent_guesthouse	36	0.000189	0.00008	0.000276	0.000130
room_shared	2	0.000005	0.00000	0.000008094	0.000000393
has_avail	1	0.000018	-0.00001	0.000009331	-0.000003382
prop_ent_townhouse	4	0.000018	-0.00002	0.000017110	-0.000012981
instant	302	0.000580	-0.00050	0.000700	-0.000467
reviews_per_month_n	7244	0.014591	-0.00155	0.021092	0.000212
host_resp_rate	5595	0.005927	-0.00278	0.009644	-0.001800
review_scores_rating_n	5607	0.007433	-0.00302	0.011747	-0.002519
availability_30_n	11878	0.011640	-0.00308	0.019858	-0.001434
review_scores_cleanliness_n	7378	0.008473	-0.00309	0.014264	-0.002663
review_scores_location_n	8665	0.009293	-0.00394	0.016157	-0.002925
availability_60_n	14303	0.012326	-0.00416	0.022227	-0.002208
review_scores_accuracy_n	6112	0.006360	-0.00418	0.011300	-0.003564
review_scores_communication_n	6797	0.005765	-0.00448	0.010893	-0.003567
review_scores_checkin_n	7137	0.006626	-0.00492	0.012845	-0.003496
maximum_nights_n	10635	0.007284	-0.00497	0.014725	-0.003529
number_of_reviews_n	6416	0.006459	-0.00520	0.011923	-0.004317
host_acc_rate	15843	0.012351	-0.00679	0.023206	-0.004147
review_scores_value_n	11121	0.007794	-0.00688	0.016258	-0.005902
availability_90_n	16244	0.010618	-0.00754	0.021832	-0.005630
availability_365_n	19447	0.012775	-0.00890	0.026532	-0.006518



Top 10 Most Important Variables - Random Forest

Variable	NRules	MSEOOB	AAEOOB
accommodates_n	3781	0.0478	0.0402
bathrooms_num	2297	0.0381	0.0314
beds_n	3029	0.0242	0.0203
bedrooms_n	2423	0.0217	0.0167
minimum_nights_n	6727	0.0201	0.0203
room_entire	107	0.0195	0.0143
room_priv	126	0.0183	0.0130
prop_priv_home	177	0.0147	0.0108
prop_ent_guest_st	318	0.0037	0.0050
prop_ent_condo	460	0.0021	0.0031