Contents

INTRODUCTION

In the realm of data analysis, uncovering valuable insights from complex datasets has become an essential practice. This project embarks on a journey through the vibrant city of **Chicago, Illinois**, leveraging Airbnb's detailed listings data to unravel patterns and predict the variable that matters most to both hosts and guests — the "**Price**."

Unlike conventional projects guided by specific stakeholder demands, this endeavor embraces a purely data-driven modeling approach. Free from predefined assumptions or targeted outcomes, the objective here is to explore the dataset with an unbiased lens, allowing the data to reveal its inherent patterns and relationships, using SAS as the analytical tool to navigate the dataset's intricacies. This exploration encapsulates the essence of a "Data-Driven Modeling" project, where the focus is on deriving meaning directly from the data itself.

The dataset, meticulously curated from Inside Airbnb's wealth of information, serves as the cornerstone for our analysis. By selecting the Detailed Listings data (*listings.csv.gz*), we ensure a granular examination of the factors influencing Airbnb prices in Chicago. To enhance the relevance and efficiency of our analysis, the dataset has been streamlined in MS Excel, retaining only the columns pertinent to our exploration.

EXPLORATORY DATA ANALYSIS

1. THE DATA

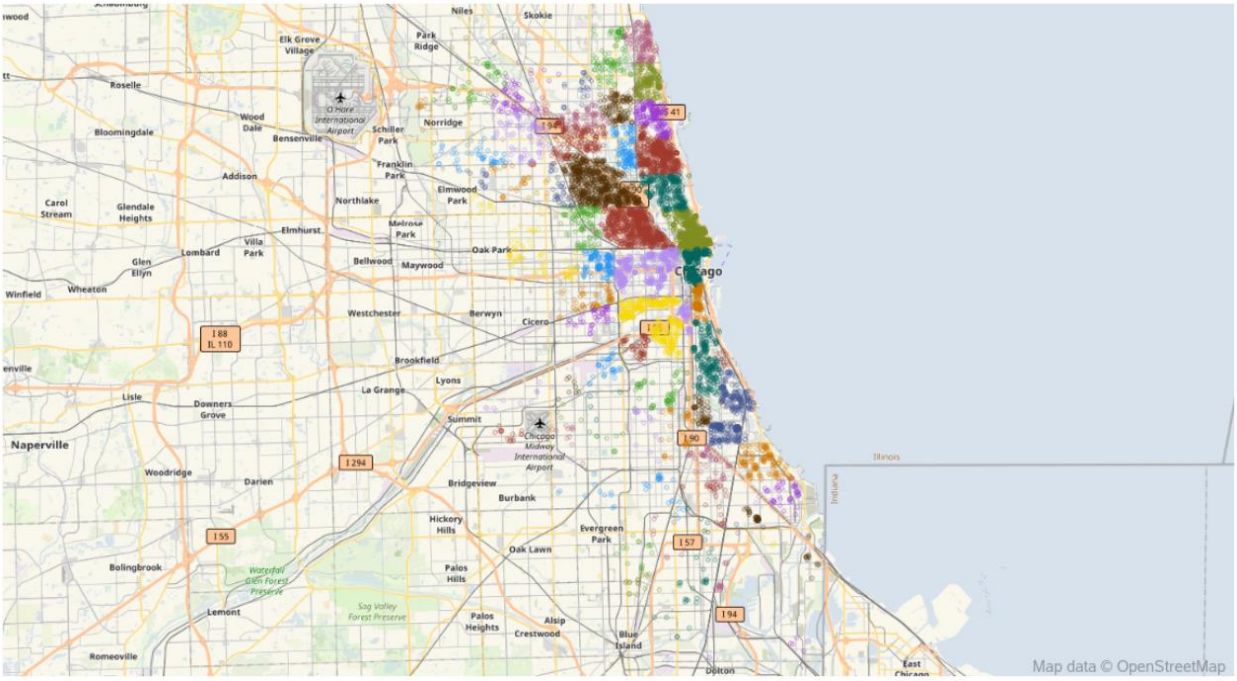
Source : Link to the data used for the analysis can be accessed from the link "Chicago-Listings"

Data Import

- The PROC IMPORT procedure was employed to bring the data into SAS from the Excel file.
- The dataset was stored in the WORK library under the name "Import."
- The GETNAMES=YES option allowed SAS to retrieve variable names from the Excel file.

Map View

PROC SGMAP is used to generate a map view of the listings grouped by neighborhood which can be seen below. However, as the data does not have a grouping information about these neighborhoods, it is difficult to understand which areas in Chicago have relatively higher number of listings.



Data Overview

- To gain insights into the structure of the imported data, the PROC CONTENTS procedure was executed.
- The summary revealed that the dataset comprises a total of 8,528 rows and 36 variables.
- Among these variables, 23 are numeric, and 13 are character variables.

2. INVESTIGATING THE DEPENDENT VARIABLE

In this, the focus is on understanding the distribution and characteristics of the target variable, which is the 'Price' of Airbnb listings. PROCE UNIVARIATE is used to understand the intricacies of 'Price.'

Descriptive Statistics: The summary statistics provide a snapshot of the 'Price' variable's key characteristics:

Variable (Price)			
N	8528	**Coeff Variation**	134.311202
Mean	199.163813	**Std Error Mean**	2.89667034
Std Deviation	267.499311	**Variance**	71555.8816
Skewness	8.02807258	**Kurtosis**	120.741661

- **Mean:** The average price is approximately $199.16.
- **Standard Deviation:** The price exhibits considerable variability with a standard deviation of $267.50. We can see that the Standard Deviation is larger than mean which inducates potential outliers.
- **Skewness:** The skewness value of 8.03 indicates a right-skewed distribution.
- **Kurtosis:** The high kurtosis value (120.74) suggests heavy tails and a peaked distribution.
- **No Missing observations:** The total number of price observations (8528) matches with the total number of rows in the dataset indicating that there are no missing observations.

Extreme Observations: Identifying extreme observations helps in understanding the range and potential outliers within the data.

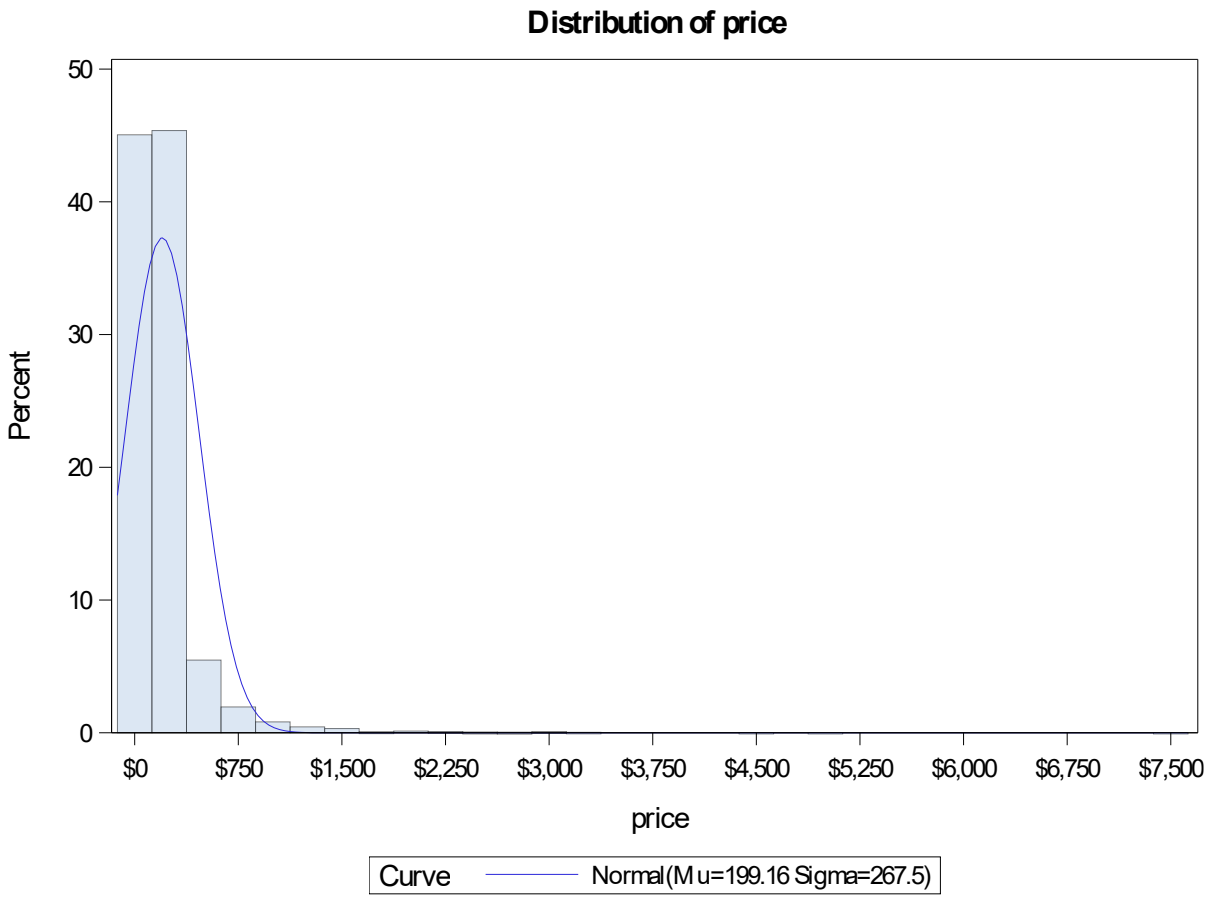
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
12	2890	3357	7719
13	3121	3357	7720
16	8462	4500	2124

Notable extreme observations include listings with prices as low as $12 and as high as $3357.

Goodness-of-Fit Tests: The goodness-of-fit tests assess whether the 'Price' variable follows a normal distribution:

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.25163	Pr > D	<0.010
Cramer-von Mises	W-Sq	210.78333	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1106.68891	Pr > A-Sq	<0.005

We can see that the p-value is less than 0.05 in all the above tests which suggest a significant departure from normality. This can be further evidenced from the below distribution:



To conclude, we can say that the 'Price' variable exhibits a right-skewed distribution with notable extreme values. The goodness-of-fit tests reject the hypothesis of normality, indicating that the 'Price' distribution deviates significantly from a normal distribution.

Data Transformation & Eliminating Outliers

To enable our predictive modeling, a log transformation was applied to address the price variable's right-skewed distribution. To address the presence of extreme values. A decision has been made to exclude values below 27 and above 1248 based on an analysis of quantiles, which revealed that 98% of the data falls within this range. This range was chosen to eliminate potential outliers while retaining a significant portion of the dataset.

Quantiles (Price)	
Level	**Quantile**
100% Max	7585
99%	1248
50% Median	135
1%	27
0% Min	12

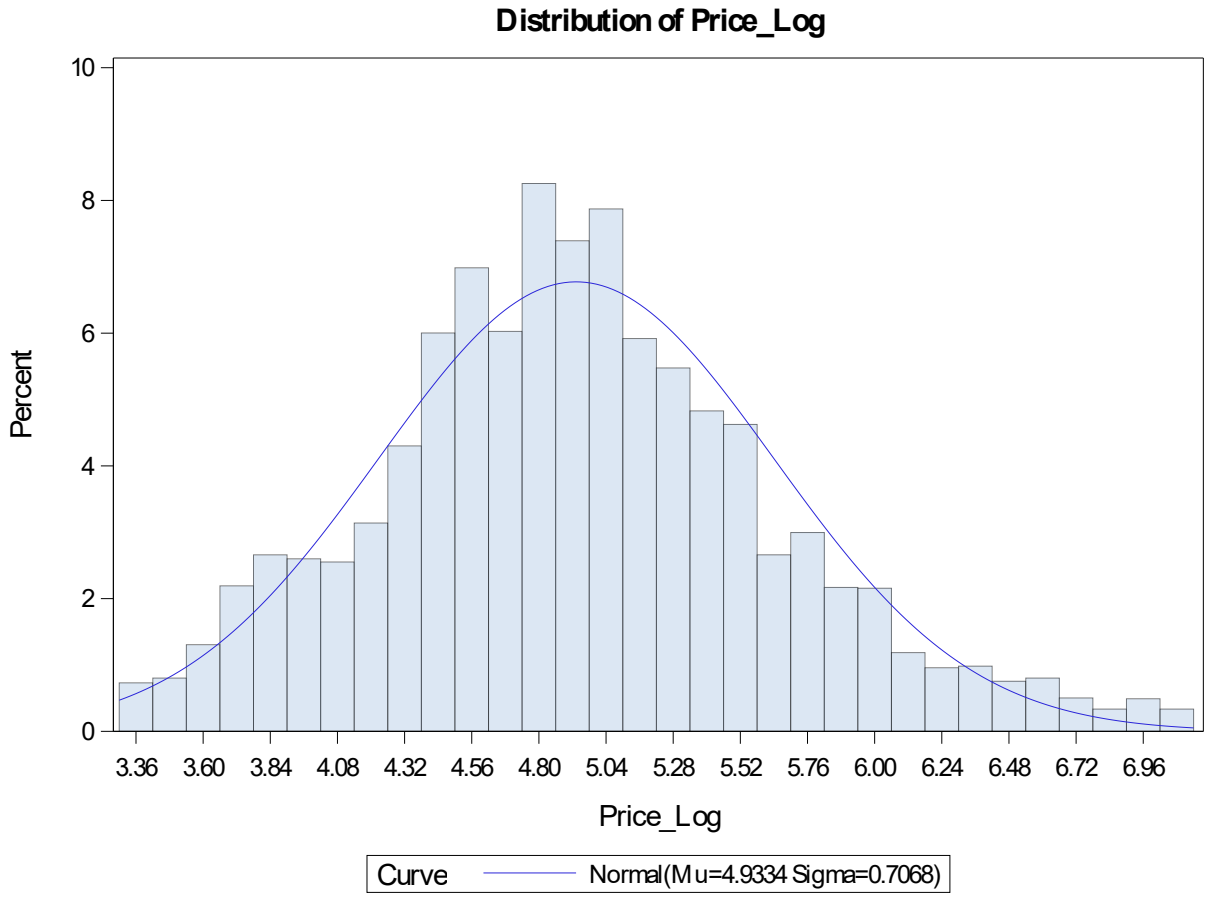
As you can see from above, this interval aligns with the 1% to 99% quantiles, excluding extreme values observed at the 0% and 100% quantiles.

PROC SQL is used to generate a list of extreme observations based on the above criteria. This list can be viewed in the "List of Outliers" of the appendices section. From the list, we can understand that the listings in the areas like "West Town" and "Near North Side" are exorbitantly priced whereas, areas like "South Chicago" and "West Ridge" displayed the most economical side of the city.

Below are the insights from the summary statistics of distribution of the log-transformed 'Price':

Variable (Price_Log)			
N	8346	**Sum Weights**	8346
Mean	4.9334291	**Sum Observations**	41174.3993
Std Deviation	0.70678975	**Variance**	0.49955175
Skewness	0.31624392	**Kurtosis**	0.14223247
Uncorrected SS	207299.739	**Corrected SS**	4168.75936
Coeff Variation	14.3265412	**Std Error Mean**	0.00773662

- **Mean:** The average log-transformed price is approximately 4.93.
- **Standard Deviation:** The spread of log-transformed prices is 0.71.
- **Skewness:** The skewness of 0.32 indicates a slight rightward skew, suggesting that the log-transformed prices are more symmetrically distributed than the original prices.
- **Kurtosis:** The kurtosis of 0.14 implies a relatively normal distribution, with fewer extreme values compared to a perfectly normal distribution.



We can see from the above figure that the log transformation has successfully addressed the right-skewed distribution of the 'Price' variable, providing a more symmetric and normalized distribution for further analysis.

3. INVESTIGATING NUMERIC VARIABLES

Variables Excluded from Analysis: The following variables have been excluded from the analysis for specific reasons:

Excluded Variable(s)	Reasoning
• number_of_reviews_ltm • number_of_reviews_l30d	Focus is placed on the broader variable "number_of_reviews."
• availability_60 • availability_90 • availability_365	By excluding the availability of 60, 90, and 365 variables, we can still retain the necessary information that we get from the single “availability_30” variable.

Handling Missing Values: Missing values are identified using PROC MEANS procedure. The following table summarizes how these missing values are handled for our analysis.

Variable	N	N Miss	% of Missing	Handling Criteria
host_response_rate	7480	866	10.38%	Replace with Mean
host_acceptance_rate	7620	726	8.70%	Replace with Mean
Bathrooms	8335	11	0.13%	Set to “0”
Bedrooms	6224	2122	25.43%	Replace with Minimum
Beds	8301	45	0.54%	Replace with Minimum
review_scores_rating	6760	1586	19.00%	Replace with Mean
review_scores_accuracy	6749	1597	19.13%	Replace with Mean
review_scores_cleanliness	6748	1598	19.15%	Replace with Mean
review_scores_checkin	6749	1597	19.13%	Replace with Mean
review_scores_communication	6748	1598	19.15%	Replace with Mean
review_scores_location	6749	1597	19.13%	Replace with Mean
review_scores_value	6749	1597	19.13%	Replace with Mean
reviews_per_month	6760	1586	19.00%	Replace with Mean

- host_response_rate missing values are replaced with mean as per generally accepted principles. With regards to host_acceptance_rate, Instances with "N/A" were verified against the variable "instant_bookable." As more instances of instant_bookable were false (591) than true (158), it was determined that there is no significant relation between "N/A" values of acceptance rate and the instant booking feature. Therefore, "N/A" values were replaced with the mean.
- It is assumed that absence of information about the number of bathrooms mean that there is no bathroom facility in the unit.
- When a listing does not have any value for “Bedrooms” (or) “Number of Beds”, it is assumed that there will be at least 1 bedroom and 1 bed in such facilities. “1” is the minimum value for the said variables in the summary data.
- Missing values in the various ‘review scores’ are replaced with their means as per generally accepted principles.

A full output of the PROC MEANS procedure can be seen in the “Summary – Numeric Variables” Section of the appendices.

Categorization of Certain Numeric Variables

The variables under scrutiny include accommodates, bathrooms, bedrooms, and beds. The approach involved setting thresholds for extreme values, examining the levels within these variables, and further categorizing them for a more insightful analysis.

From the results of PROC MEANS, we can see that the 99th percentile values for the specified predictors.

Variable	1%	50%	99%
accommodates	1	4	16
bathrooms	1	1	4
bedrooms	1	1	5
beds	1	2	9

- ***Setting Thresholds***
  To manage extreme values and enhance the stability of the analysis, specific thresholds are applied to the specified numeric variables. By capping the values at the 99th percentile, we ensure that any outliers beyond these thresholds are brought within a reasonable range, preventing potential distortions in our analysis.
- ***Checking Levels***
  PROC FREQ procedure is used to generate a summary of levels for each variable.

Variable	Levels
accommodates	16
availability_30	31
bathrooms	9
bedrooms	5
beds	9

  Full results of the PROC FREQ procedure can be seen in the 'FREQ Numeric Variables' section of appendices.

- ***Categorization***
  Based on the frequency distribution of each variable from the above procedure, categorization is made. The criteria used for categorization are provided below:

Variable : accommodates New Variable: accom_cat		
No of People	**Category**	**Distribution**
1 to 2	Very Small	Around 38%
3 to 4	Small	Around 29%
5 to 6	Medium	Around 19%
more than 6	Large	Around 14%

Variable : availability_30 New Variable: avail_cat		
No of Days	**Category**	**Distribution**
Zero	Booked	Around 23%
1 to 10	Low	Around 22%
11 to 20	Medium	Around 29%
> 20	High	Around 26%

Variable : bedrooms New Variable: bedrooms_cat		
No of Bedrooms	**Category**	**Distribution**
1	Single	Around 28%
2	Double	Around 52%
More than 2	Extra	Around 20%

Variable : beds New Variable: beds_cat		
No of Beds	**Category**	**Distribution**
1	Single	Around 43%
2	Double	Around 26%
More than 2	Extra	Around 31%

Variable : bathrooms New Variable: bath_cat		
No of bathrooms	**Category**	**Distribution**
Zero	Nil	Around 0.38%
0.5, 1, 1.5	Normal	Around 72%
2, 2.5	Extra	Around 23%
3, 3.5, 4	Luxury	Around 5%

Feature Engineering

To enrich the dataset, two new variables, “beds_per_person” and “bath_per_person”, were created. These new variables capture the number of beds and bathrooms per person, providing additional dimensions for analysis.

Checking Predictor Multicollinearity

To assess multicollinearity among numeric predictor variables, a stepwise regression model was constructed using the PROC REG procedure. The variables included in the initial model were:

host_response_rate	host_acceptance_rate	minimum_nights
maximum_nights	beds_per_person	bath_per_person
number_of_reviews	review_scores_rating	review_scores_accuracy
review_scores_cleanliness	review_scores_checkin	review_scores_communication
review_scores_location	review_scores_value	reviews_per_month

The results of the above procedure as far as VIFs are concerned is produced below:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.08228	0.14181	28.79	<.0001	0
host_response_rate	1	-0.53155	0.06478	-8.20	<.0001	1.17066
host_acceptance_rate	1	0.34059	0.04427	7.69	<.0001	1.19885
minimum_nights	1	-0.00090796	0.00017549	-5.17	<.0001	1.06688
maximum_nights	1	0.00006633	0.00001523	4.36	<.0001	1.03606
beds_per_person	1	0.03987	0.03174	1.26	**0.2090**	1.35019
bath_per_person	1	-0.83077	0.02452	-33.89	<.0001	1.37779
number_of_reviews	1	-0.00072918	0.00009154	-7.97	<.0001	1.39206
review_scores_rating	1	0.14260	0.03078	4.63	<.0001	3.33161
review_scores_accuracy	1	-0.05226	0.04007	-1.30	**0.1922**	4.16412
review_scores_cleanliness	1	0.29271	0.02989	9.79	<.0001	2.66767
review_scores_checkin	1	-0.08166	0.03307	-2.47	0.0135	2.33193
review_scores_communication	1	-0.05490	0.03529	-1.56	**0.1198**	2.83720
review_scores_location	1	0.32590	0.02481	13.13	<.0001	1.47790
review_scores_value	1	-0.28043	0.03197	-8.77	<.0001	3.35340
reviews_per_month	1	0.00453	0.00405	1.12	**0.2637**	1.40925

From the above results, we can see that "beds_per_person," "review_scores_accuracy," "review_scores_communication," and "reviews_per_month" exhibited relatively higher p-values. As they are statistically not significant to our analysis, they were consequently excluded from the model.

Though there are no VIFs above the threshold of 5, individual review scores (cleanliness, check-in, location, and value) are considered as more granular aspects of the overall rating. In order to simplify the analysis, these variables are excluded, and focus has been made on the broader sentiment expressed by reviewers. Consequently, the variable "review_scores_rating" is considered a comprehensive indicator and is chosen as a representative variable to capture the essence of reviews in a more consolidated manner.

Polynomial Terms

While considering the inclusion of polynomial terms, it was observed that introducing higher-degree polynomials led to increased VIFs. So, they were no polynomial terms introduced in the model.

Transformations of Numeric Predictors

An assessment of numeric predictor variables was conducted to determine if any transformations were necessary for modeling purposes. Firstly, a univariate analysis was performed using PROC UNIVARIATE, focusing on the remaining numeric predictors namely:

host_response_rate	host_acceptance_rate	minimum_nights	review_scores_rating
maximum_nights	bath_per_person	number_of_reviews	

Histograms were generated to visualize the distribution of each variable. The examination revealed notable skewness and kurtosis in several variables, and indications revealing departures from a normal distribution.

Variable	Skewness	Kurtosis
host_response_rate	-6.49	45.67
host_acceptance_rate	-3.66	14.53
minimum_nights	18.99	486.83
maximum_nights	0.18	-1.71
number_of_reviews	8.97	243.82
review_scores_rating	-6.06	51.69
bath_per_person	4.2	29.69

host_response_rate and review_scores_rating exhibited substantial negative skewness, implying a concentration of values on the higher end. Conversely, minimum_nights displayed pronounced positive skewness, suggesting a concentration of values on the lower end. The kurtosis values for these variables indicated heavy-tailed distributions. Even though maximum_nights have relatively low magnitude, the "Goodness-of-fit" tests revealed a failure of normality assumptions.

To address the non-normality, log (or) sqrt transformations were applied to the variables.

Variable	Transformation Applied	New Variable
minimum_nights	Log	min_nights_tf
maximum_nights	Log	max_nights_tf
bath_per_person	Square Root	bpp_tr
number_of_reviews	Log	num_reviews_tf

The above transformations aim to enhance the suitability of these variables for inclusion in regression models and facilitate a more robust analysis of their relationship with the price.

However, despite these efforts, 3 variables (host_response_rate, host_acceptance_rate, and review_scores_rating) continued to exhibit non-normality, failing goodness-of-fit tests. Scatter plots of these three variables illustrated an unclear relationship with the target variable (Price_Log). A scatterplot matrix of these three variables with the 'Price_Log' can be accessed in the appendices, 'Scatterplot Matrix : Non-normal Predictors'.

Categorization of Numeric Predictors that were Not Normal

Given the challenges in achieving normality through transformations, a decision was made to categorize the aforementioned variables. PROC FREQ was employed to generate frequency tables and assess the number of levels in each variable. The criteria used for categorization are provided in the below tables. Even though we can see that the contributions of some categories are insignificant, the categorization is applied as such, as these are ordinal variables.

Variable : host_response_rate New Variable: resp_cat		
Percentage	**Category**	**Distribution**
100%	Perfect	Around 75%
>= 97%, < 100%	Great	Around 16%
>= 80%, < 97%	Good	Around 6%
< 80%	Bad	Around 2%
ZERO	Worst	Around 1%

Variable : host_acceptance_rate New Variable: acc_cat		
Percentage	**Category**	**Distribution**
100%	Absolute	Around 32%
>= 90%, < 100%	High	Around 52%
>= 70%, < 90%	Moderate	Around 8%
< 70%	Low	Around 6.5%
ZERO	Zero	Around 1.5%

Variable : review_scores_rating New Variable: rating_cat		
Rating	**Category**	**Distribution**
5	Perfect	Around 18.67%
>=4, <5	Great	Around 79%
>=3, <4	Good	Around 1.65%
>=2, <3	Okay	Insignificant
>=1, <2	Bad	Insignificant
>=0, <1	Terrible	Insignificant
ZERO	Disgusting	Insignificant

A correlation analysis was conducted for the four numeric predictors that were chosen for the model and the results are given below:

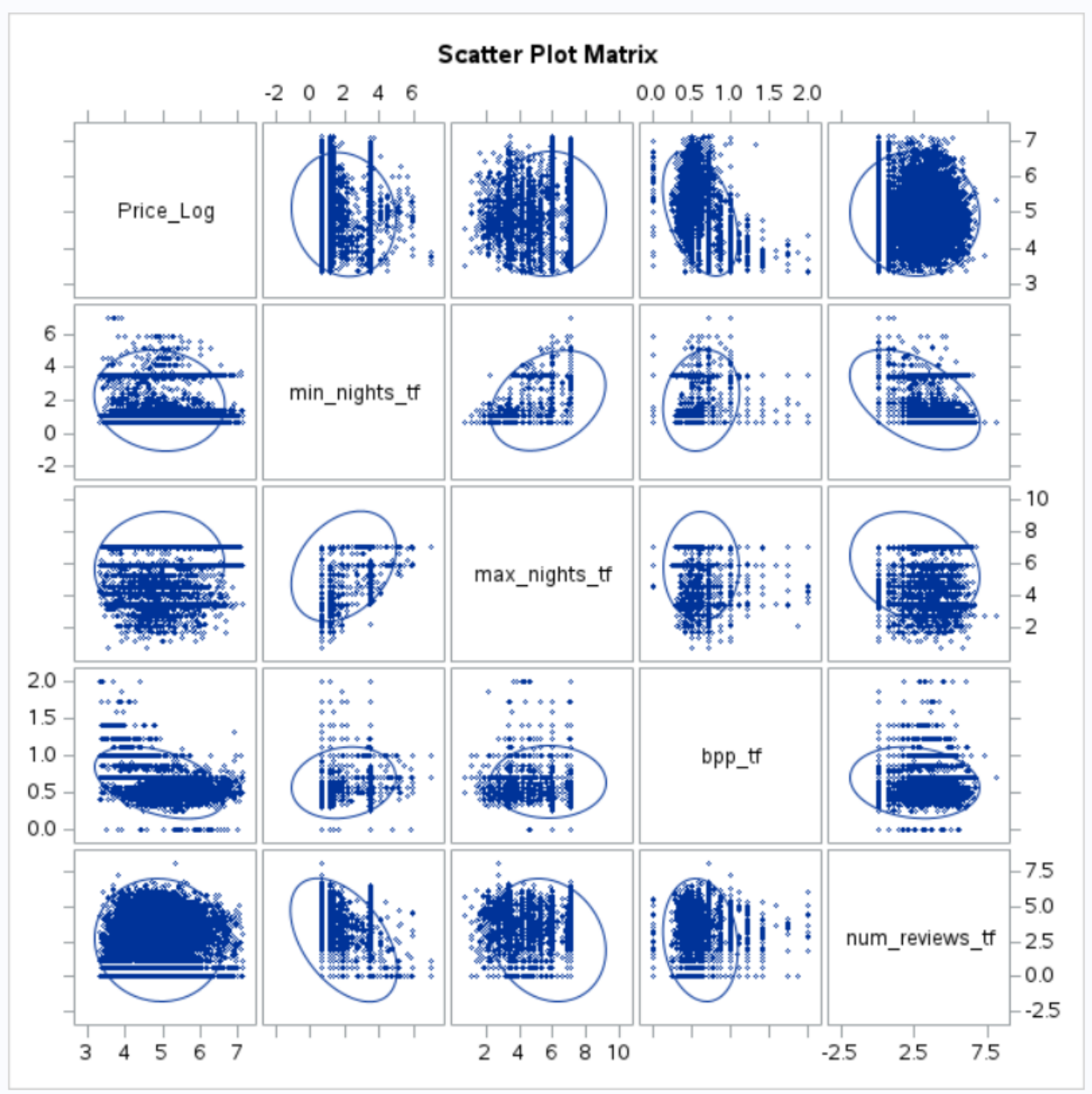
Pearson Correlation Coefficients, N = 8346 Prob > |r| under H0: Rho=0				
	min_nights_tf	**max_nights_tf**	**bpp_tf**	**num_reviews_tf**
min_nights_tf	1.00000	0.31975 <.0001	0.15112 <.0001	-0.49112 <.0001
max_nights_tf	0.31975 <.0001	1.00000	0.00265 0.8091	-0.18467 <.0001
bpp_tf	0.15112 <.0001	0.00265 0.8091	1.00000	-0.15535 <.0001
num_reviews_tf	-0.49112 <.0001	-0.18467 <.0001	-0.15535 <.0001	1.00000

We can see that the minimum nights is significantly correlated with the maximum nights. However, both variables are included in the model as the previous steps ensure that there will be no multi-collinearity. Similarly, any other correlations that can be observed above are ignored as there are no multi-collinearity issues and such relationships also do not have a practical significance.

Variable	Skewness	Kurtosis
min_nights_tf	0.54	-1.22
max_nights_tf	-0.96	-0.12
bpp_tf	1.62	6.55
num_reviews_tf	-0.02	-1.2

From the above table, we can see that the transformations applied on the variables had significantly altered the skewness and kurtosis and these are much better when compared to the original variables, with an exception being maximum_nights. However, the transformed variable ‘max_nights_tf’ is included in the final model to maintain the consistency.

A scatterplot matrix has been generated for the above numeric variables in the model.



Despite the above variables displaying a slightly categorical behavior, they are included in the model as they predominantly capture non-linear relationships with the Price_Log.

4. INVESTIGATING CHARACTER VARIABLES

Three-character variables are identified and a PROC FREQ is implemented on those. A summary of the results is shown below:

Variable	Levels
neighbourhood_cleansed	77
property_type	48
room_type	4

These variables are categorized based on their percentage of frequency in the total distribution. A brief about the criteria used is shown in the below tables.

Variable : neighbourhood_cleansed New Variable: hood_cat		
Criteria	**Category**	**Distribution**
List of neighborhoods that contribute to the top 50%	High-Demand	Around 50%
List of neighborhoods that contribute to next 35%	Moderate-Demand	Around 35%
Remaining neighborhoods	Low-Demand	Around 15%

Variable : property_type New Variable: property_cat		
Criteria	**Category**	**Distribution**
List of types that comprise the top 50%	Primary	Around 52%
List of types that comprise the next 40%	Secondary	Around 40%
Remaining Property types	Other	Around 8%

Variable : room_type New Variable: room_cat		
Room Type	**Category**	**Distribution**
Entire home/apt	Home	Around 77%
Hotel Room Shared Room Private Room	Room	Around 33%

5. DATA PARTITIONING

The dataset was divided into two distinct sets: TRAIN and TEST. This partitioning was performed using the PROC SURVEYSELECT procedure with a sampling rate of 20% (80/20 split) and a random seed value for reproducibility.

TRAIN Dataset:

- This subset, constituting 80% of the original data, is utilized for model training.
- The chosen machine learning or statistical models learn patterns and relationships within this dataset.

TEST Dataset:

- Comprising the remaining 20% of the original data, this subset serves as an unseen dataset for model evaluation.
- After the models are trained on the TRAIN dataset, their performance is assessed on the TEST dataset to gauge their predictive accuracy and generalization to new, unseen data.

By segregating the data into training and testing sets, we ensure that the models are not evaluated on the same data used for training. This practice helps prevent overfitting, providing a more realistic assessment of model performance on new, unseen observations. The seed value ensures reproducibility, enabling consistent results across multiple iterations of the analysis.

PREDICTIVE MODELING

The following are the list of variables that are considered in the model:

Numeric Variables	Categorical Variables
min_nights_tf max_nights_tf bpp_tf num_reviews_tf	instant_bookable accom_cat avail_cat bath_cat bedrooms_cat beds_cat resp_cat acc_cat rating_cat hood_cat property_cat room_cat

Before modeling, a set of macro variables were created to ease the coding process of the modelling. The details are given below:

Macro Variable	Description	Comprises
num_vars	Numerical Predictors	min_nights_tf , max_nights_tf, bpp_tf, num_reviews_tf
cat_vars	Categorical Predictors	instant_bookable, accom_cat, avail_cat, bath_cat, bedrooms_cat, beds_cat, resp_cat, acc_cat, rating_cat, hood_cat, property_cat, room_cat
nom_cats	Nominal Categorical Predictors	instant_bookable, hood_cat, property_cat, room_cat
ord_cats	Ordinal Categorical Predictors	accom_cat, avail_cat, bath_cat, bedrooms_cat, beds_cat, resp_cat, acc_cat, rating_cat

A designated path, specified as "/home/u63048816/DSCI 519/Project," was established in SAS for downloading and storing scoring codes. This path served as a centralized location for accessing relevant files related to the scoring process.

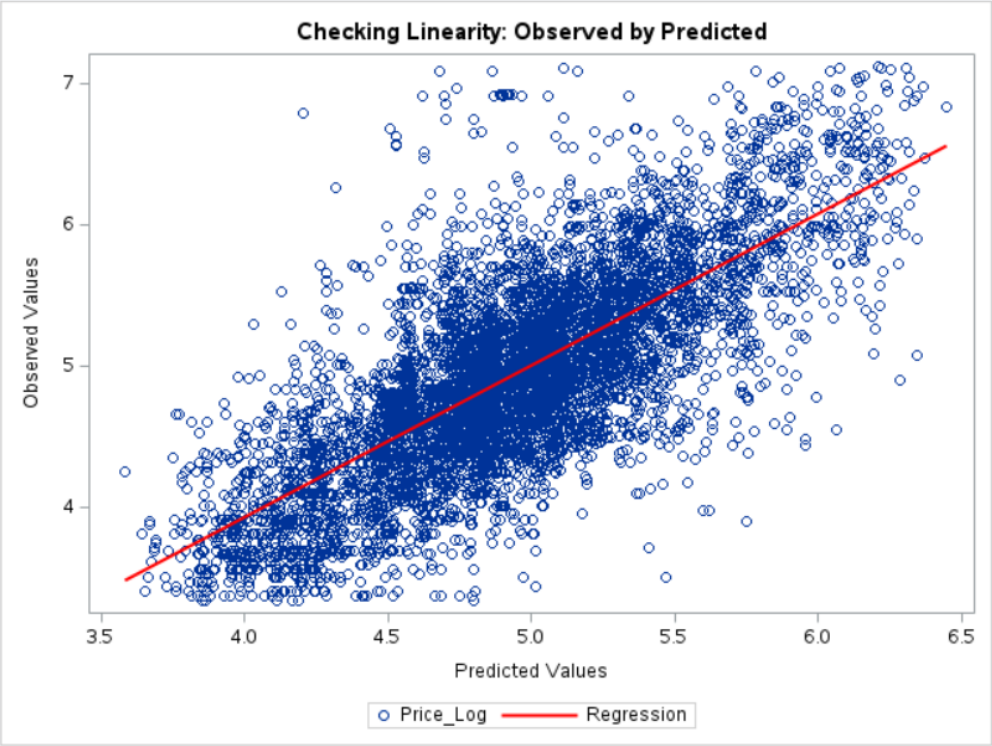
1. LINEAR REGRESSION (With LASSO Selection)

Application on Training Dataset

The training dataset was utilized to fit a linear regression model using the PROC GLMSELECT procedure. With 'Price_Log' as the response variable and both numeric and categorical variables as predictors, a LASSO selection was applied to optimize model performance. Lasso regression involves selecting a subset of features. Cross-Validation (CV) is used as a criterion to stop and choose the selection as it is not only widely recommended but also helps to ensure that the selected features generalize well to new data, improving the model's predictive performance. Once the model is applied, the linear regression assumptions are verified by generating a series of plots.

Verifying Linear Regression Assumptions

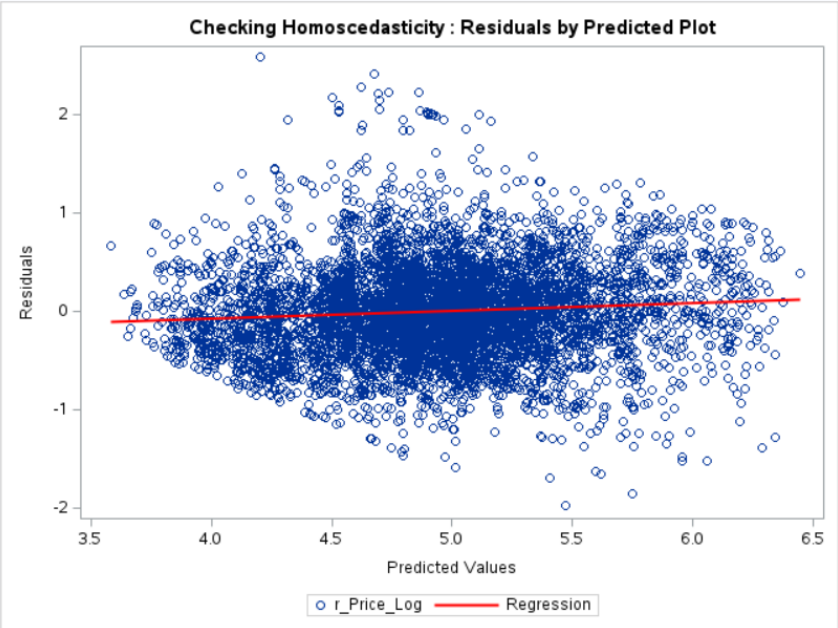
Linearity: We can say that the linearity assumption is met because of the below reasons.



From the above plot, we can see that the:

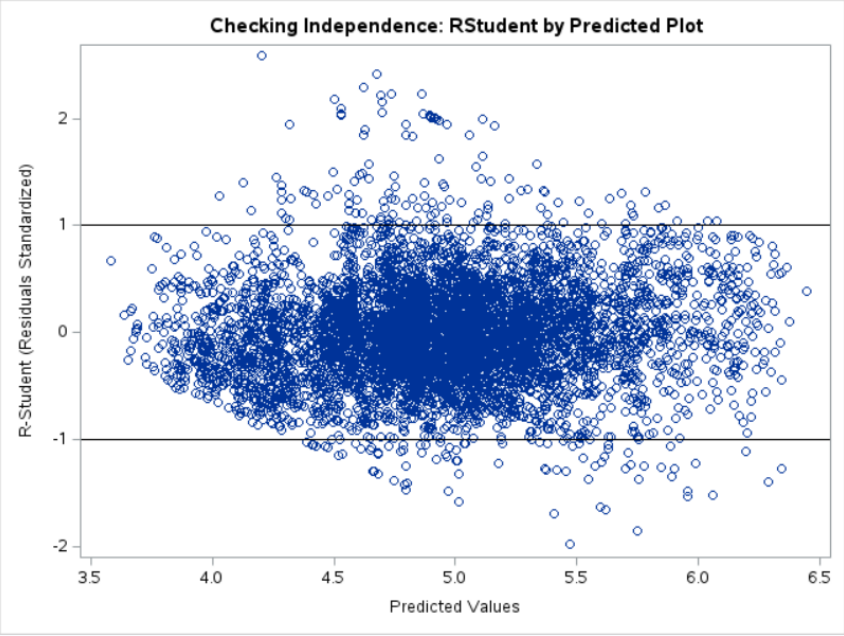
- The points form a roughly straight line from the bottom-left to the top-right of the plot, which suggests a linear relationship between the observed and predicted values.
- The spread of points is roughly consistent across the range of predicted values.

Homoscedasticity (Constant Variance):



A scatterplot of residuals vs predicted is generated to check this assumption of constant variance. From the below plot, we can see that there is no clear trend (or) pattern in the spread of the residuals. We can also see that the spread is roughly the same across the range of predicted values barring a few. Hence, we can say that the constant variance assumption is met.

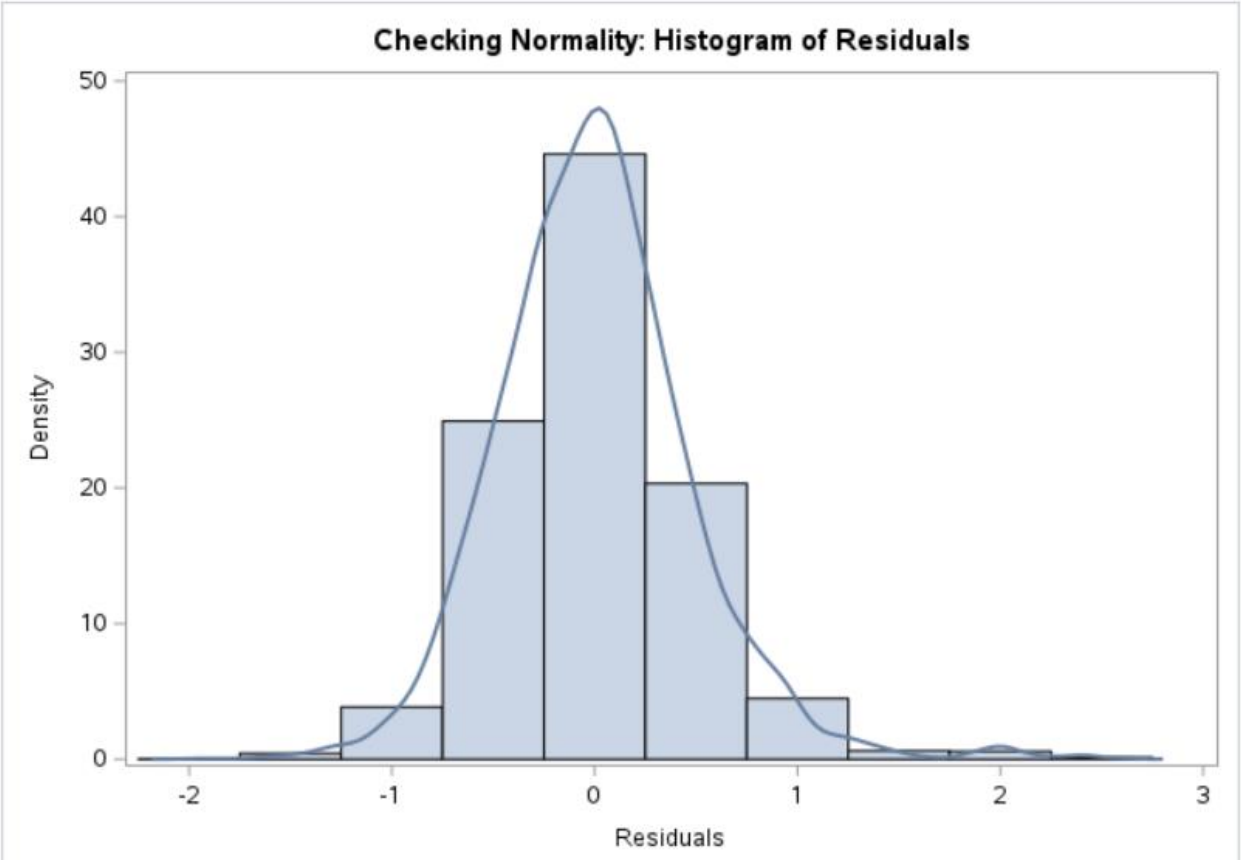
Independence



A R-Student by Predicted plot is used to check the independent assumption. From the above plot, we can say that there is no clear pattern in the spread of the predicted values as they are randomly distributed around zero (the horizontal axis). Though we can see some outliers, the spread of points is relatively constant across the range of predicted values.

Hence, we can say that the independence assumption is met.

Normality:



From the above plot, we can say that the normality assumption is met as we can see a relatively smooth bell-shaped curve and is roughly symmetric. But we can see a little peak in the distribution, which might indicate a slight deviation from normality. However, this slight deviation is not considered a concern for relying on the model's statistical inferences and estimates due to the following reasons:

- Linear regression models are quite robust to violations of normality, especially when the sample size is large.
- As all other assumptions are met, the performance of the model can be considered quite reliable.

Evaluating the model performance

From the below ANOVA table, we can see that the F-statistic of 373.51 with an associated p-value less than 0.0001 indicates that the model is statistically significant, suggesting that at least one of the predictors is contributing significantly to explaining the variance in the Airbnb price.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	21	1823.56278	86.83632	373.51	<.0001
Error	6654	1546.95489	0.23248		
Corrected Total	6675	3370.51767			

Root MSE	0.48217
Dependent Mean	4.93110
R-Square	0.5410
Adj R-Sq	0.5396

The R-Square value of 0.54 indicates that the model explains approximately 54% of the variance in the dependent variable. The Adjusted R-Square 0.54 suggests that the model is still effective when considering the number of predictors.

The RMSE of 0.48 suggests that, on average, the predicted values from the model deviate by approximately 0.48 units from the actual observed values. Comparing the RMSE with the dependent mean (4.93), we can confidently say that the RMSE is relatively small and that the model's predictions are close to the actual values.

In summary, the linear regression model with Lasso selection appears to perform well, as indicated by the low RMSE, significant F Value, and a reasonable amount of variance explained (R-Square).

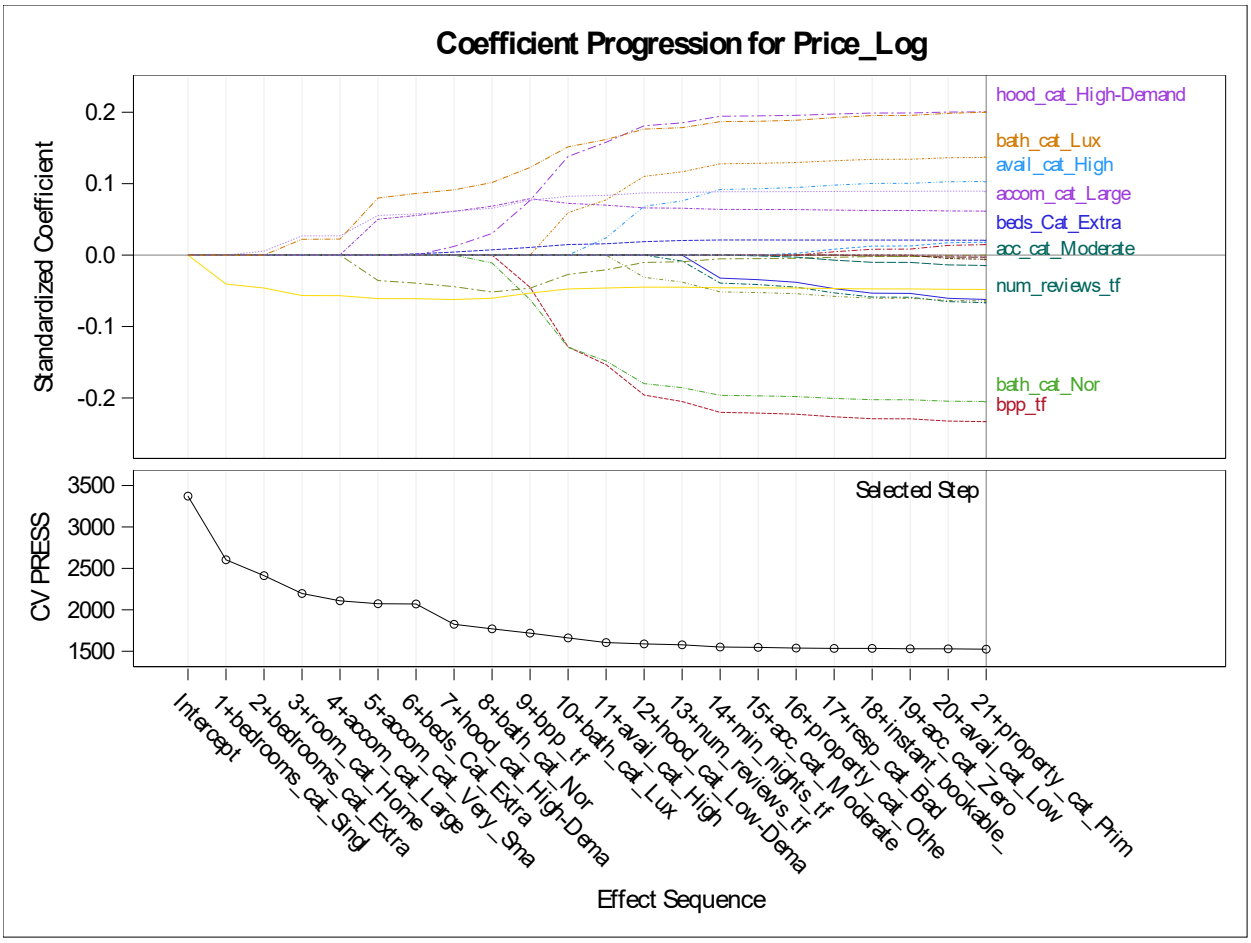
Interpreting the results

The LASSO algorithm identified 22 steps in the variable selection process. From the summary of the LASSO selection process, we can say that all the variables considered for our analysis are influential in affecting the listings' prices with two exceptions – 'max_nights_tf' and 'rating_cat'.

LASSO Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	3371.9272
1	bedrooms_cat_Single		2	2603.6541
2	bedrooms_cat_Extra		3	2412.3691
3	room_cat_Home		4	2196.8303
4	accom_cat_Large		5	2108.5918
5	accom_cat_Very_Small		6	2073.3408
6	beds_Cat_Extra		7	2070.3171
7	hood_cat_High-Demand		8	1825.3130
8	bath_cat_Nor		9	1770.3921
9	bpp_tf		10	1718.4597
10	bath_cat_Lux		11	1660.9235
11	avail_cat_High		12	1604.7789
12	hood_cat_Low-Demand		13	1588.3716
13	num_reviews_tf		14	1577.4322
14	min_nights_tf		15	1550.8852
15	acc_cat_Moderate		16	1546.6202
16	property_cat_Others		17	1538.4908
17	resp_cat_Bad		18	1534.3174
18	instant_bookable_f		19	1534.1595
19	acc_cat_Zero		20	1529.6648
20	avail_cat_Low		21	1529.0534
21	property_cat_Primary		22	1524.4979*
* Optimal Value of Criterion				

We already have a basic idea that the price of a listing may not depend on the maximum number of days it can be booked up to. However, the interesting insight from the analysis is that the ratings of a listing do not seem to have any impact on the listings' prices. From this we can say that the

hosts are not particularly considering boosting the prices of their listings even when they consistently get positive reviews from their customers. Below is a visual representation of how the estimated coefficients change as more variables are added to the model during the lasso selection process.



We know that the lasso method introduces regularization by penalizing the absolute values of the coefficients, encouraging some coefficients to shrink towards zero. This regularization is apparent in the plot as coefficients are constrained, and the model stroked a balance between including variables for predictive power and penalizing excessive complexity. That is, you can see a coefficient, like **hood_cat High-Demand,** is positively contributing to the model, as it is moving higher than 0. At the same time, another coefficient, like **bpp_tf**, is decreasing, and is negatively impacting the model. This suggests a trade-off or a balancing act between the variables.

As we have log transformed response variable, it is a difficult task to interpret the parameter estimates. However, an attempt has been made to transform these variables with anti-log to gain insights from them. However, no statistical interpretation is made as it is complex and challenging, given the time constraints.

Insights from the analysis of parameter estimates.

- Units featuring over 3 bathrooms command higher prices in their listings.
- Listings categorized as entire homes are associated with higher estimated prices.
- Properties situated in high-demand neighborhoods tend to be associated with inflated prices.
- Listings boasting additional bedrooms tend to have higher estimated prices.
- Larger capacity in terms of the number of guests is linked to higher estimated prices.
- Units offering extra beds are associated with higher estimated prices.

A full list of parameter estimates of the model can be found in 'Parameter Estimates – LASSO' section in appendices.

Application on Test Dataset

The trained model was then applied to the test dataset, where both predicted values (p_Price_Log) and residuals (r_Price_Log) were captured and stored in a separate dataset (test_lasso). The performance of the model was assessed on the generated scores using the **PROC REG** procedure. The ANOVA results suggest that the applied model demonstrates a robust fit to the test data. The model exhibits a significant fit to the test data, as indicated by a highly significant F-value of 1943.64 This suggests that at least one predictor variable in the model has a meaningful effect on the dependent variable.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	429.48401	429.48401	1943.64	<.0001
Error	1668	368.57658	0.22097		
Corrected Total	1669	798.06059			

Comparing Metrics of TRAIN and TEST

RMSE: The model performs slightly better on the test data (RMSE = 0.47007) compared to the training data (RMSE = 0.48217), suggesting good generalization.

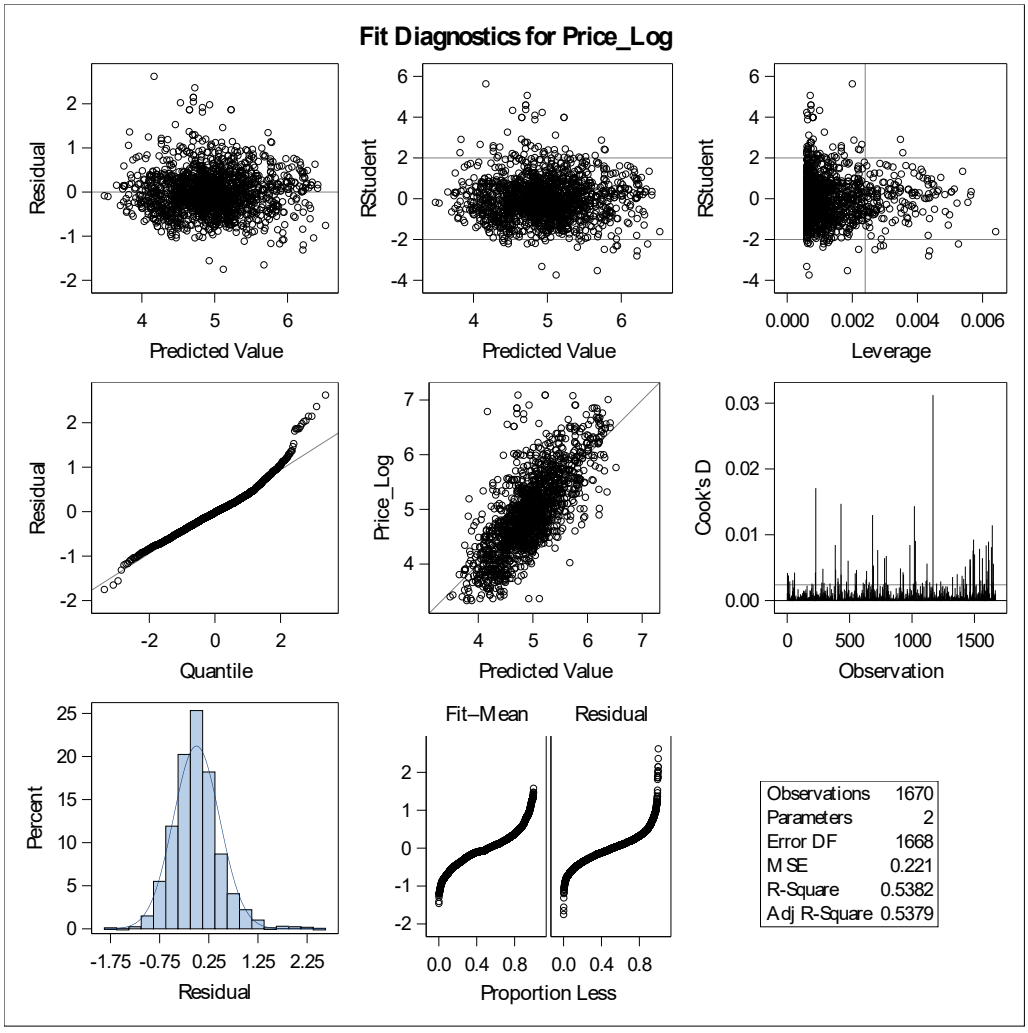
R-Square: The test R-Square (0.5382) is slightly lower than the training R-Square (0.5410), indicating that the model explains a slightly lower percentage of the variance in the test data.

Adjusted R-Square: The values are very close between test (0.5379) and training (0.5396), suggesting that the model is not overfitting the data.

Particulars	TRAIN	TEST
Root MSE	0.48	0.47
Dependent Mean	4.93	4.94
R-Square	0.541	0.5382
Adj R-Sq	0.54	0.538

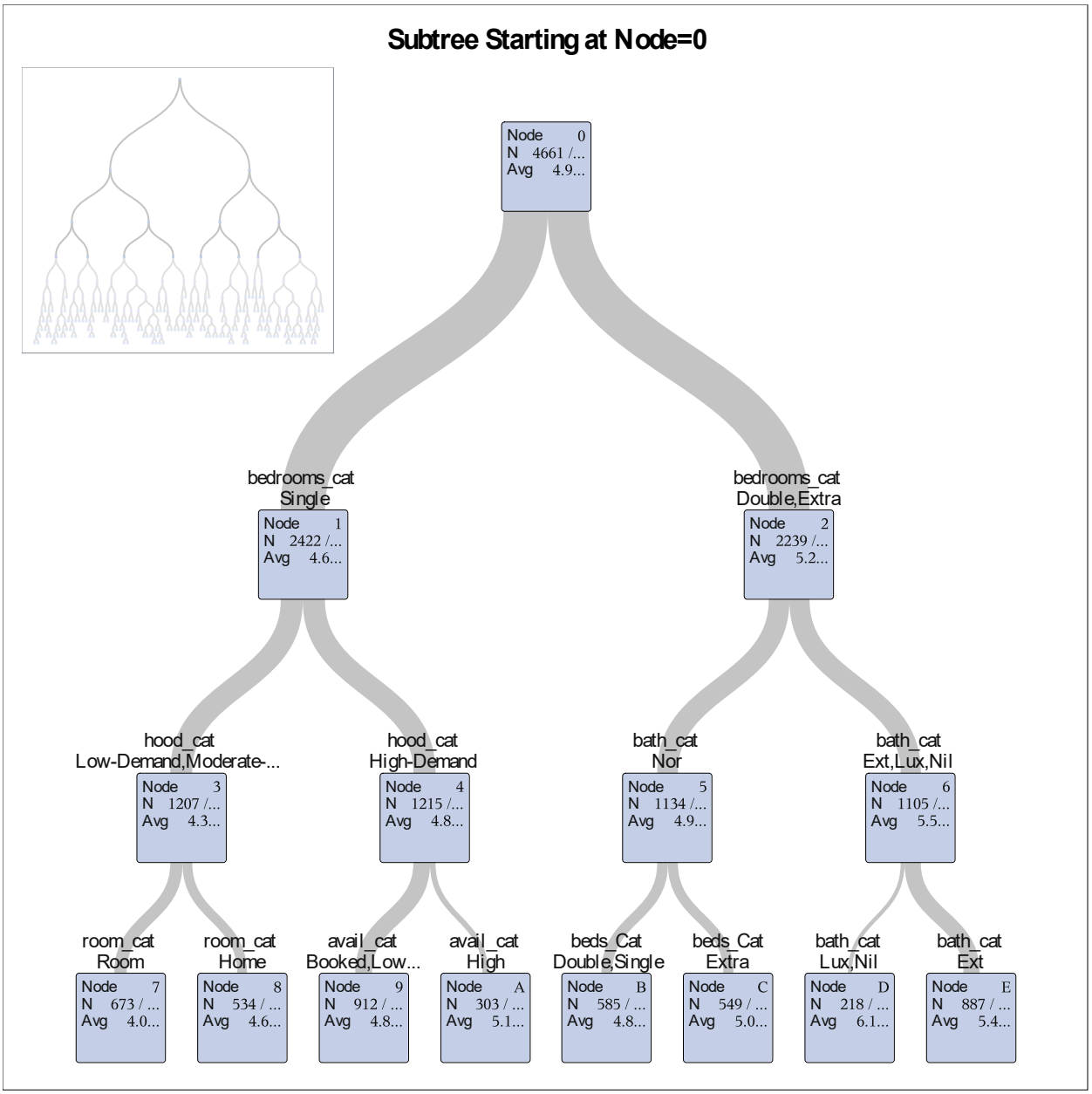
Overall, the model performs well on the test data, demonstrating good generalization from the training set. The small differences observed in performance metrics between the test and training datasets suggest that the model is robust and effective in making predictions on new, unseen data.

The assumptions of normal distribution are met for the TEST data set as well as indicated by the below visualizations.



2. DECISION TREE

PROC HPSLIP has been used to fit a decision tree on the training dataset with both numeric and categorical variables included in the model. The results of the model are as follows:



From the above tree, we can understand the following:

- The number of bedrooms plays a pivotal role in determining the price of a unit.
- If a unit offers a single bedroom, then the neighborhood in which the property is located might inflate the price of the unit.
- If a listing has 2 (or) more bedrooms, then the number of bathrooms it facilitates per guest plays a significant role in fixing the price.

These are similar insights we obtained from the linear regression model. These can also be evidenced in the variable importance table below:

Variable Importance							
		Training		Validation			
Variable	Variable Label	Relative	Importance	Relative	Importance	Relative Ratio	Count
bedrooms_cat		1.0000	23.1052	1.0000	15.3880	1.0000	3
bath_cat		0.7397	17.0917	0.7415	11.4105	1.0024	2
hood_cat		0.6694	15.4671	0.6051	9.3111	0.9039	11
room_cat		0.4927	11.3847	0.4795	7.3788	0.9732	6
bpp_tf		0.3104	7.1728	0.2815	4.3325	0.9069	11
avail_cat		0.3210	7.4178	0.2691	4.1404	0.8381	11
min_nights_tf		0.2950	6.8171	0.2627	4.0428	0.8904	17
accom_cat		0.2655	6.1336	0.2212	3.4031	0.8331	9
resp_cat		0.3412	7.8836	0.2048	3.1512	0.6002	12
property_cat		0.2648	6.1178	0.1623	2.4975	0.6130	7
num_reviews_tf		0.3140	7.2547	0.1622	2.4960	0.5166	17
beds_Cat		0.1840	4.2519	0.1555	2.3923	0.8448	5
instant_bookable	instant_bookable	0.2384	5.5088	0.1550	2.3848	0.6500	2
acc_cat		0.2546	5.8820	0.1087	1.6728	0.4270	11
max_nights_tf		0.2429	5.6115	0.1042	1.6041	0.4292	18
rating_cat		0.1216	2.8097	0.0509	0.7831	0.4185	4

Later, the generated scoring code from the decision tree model was applied to the hold-out test sample. Then, RMSE is calculated for both TRAIN and TEST samples and the results are given in the table.

RMSE	
TRAIN	0.412504
TEST	0.458771

From the results, we do not see similar performance between the training and test datasets. The training RMSE is much lower than the test RMSE, which indicates us overfitting. This suggests that the model might be too complex and has memorized the training data rather than learning underlying patterns. This can lead to poor generalization of new data.

3. RANDOM FOREST

PROC HPFOREST procedure is used to build a Random Forest model on the training dataset. The results are discussed below.

Loss Reduction Variable Importance					
Variable	Number of Rules	MSE	OOB MSE	Absolute Error	OOB Absolute Error
bedrooms_cat	3941	0.100101	0.09752	0.056329	0.054703
bath_cat	2189	0.061537	0.06142	0.047829	0.047645
hood_cat	9202	0.044795	0.04144	0.032050	0.028457
room_cat	1679	0.034345	0.03332	0.035060	0.034703
accom_cat	6144	0.026316	0.02261	0.017677	0.014623
bpp_tf	18869	0.019923	0.00880	0.018272	0.008723
instant_bookable	9874	0.011104	0.00615	0.007886	0.003662
min_nights_tf	23732	0.014127	0.00220	0.015236	0.004648
resp_cat	9815	0.008701	0.00194	0.006325	0.001483
property_cat	9650	0.006494	0.00128	0.006348	0.001762
beds_Cat	4462	0.002521	0.00019	0.002334	0.000243
avail_cat	19888	0.008072	-0.00054	0.007677	-0.000331
acc_cat	14027	0.005595	-0.00082	0.006316	0.000385
rating_cat	5753	0.001543	-0.00089	0.001500	-0.000754
max_nights_tf	31990	0.013897	-0.00096	0.015120	0.001159
num_reviews_tf	75428	0.026502	-0.00543	0.024188	-0.006584

We can gain the following insights from the above table:

- The number of bedrooms in a listing contributes significantly to determining the price.
- The number of bathrooms per person is the second most influential indicator.
- The neighborhood in which the property is located is also vital.
- Whether the unit is a complete apartment (or) not is crucial.
- The number of guests a listing allows plays the next most influential parameter.

These are again similar to what we concluded from the previous models.

Later, PROC HP4SCORE procedure is employed to score the hold-out test sample using the previously trained Random Forest model. Then, RMSE is calculated for both TRAIN and TEST samples and the results are given in the table.

RMSE	
TRAIN	0.335414
TEST	0.40008

Like decision tree, we can see that the test RMSE is significantly higher than the training RMSE, which suggests us that the model is overfitting the training data. Random Forest has learned the training data too well but struggles with new, unseen data.

MODEL COMPARISON AND RECOMMENDATIONS

RMSE	Linear Regression	Decision Tree	Random Forest
TRAIN Data	0.48217	0.412504	0.335414
TEST Data	0.47007	0.458771	0.40008

RMSE is used to compare and assess the three models' predictive performances. Here are the conclusions we can draw from the comparison:

- The Random Forest model has the lowest RMSE on the TRAIN data, suggesting that it fits the training data better than the other models.
- The Random Forest model also has the lowest RMSE on the TEST data, indicating superior generalization performance compared to both Linear Regression and Decision Tree models.
- However, as pointed out earlier in the report, Both Decision Tree and Random Forest models suffer from 'Overfitting' as they poorly generalize to the unseen data, which is evident from their TEST data RMSEs that are significantly higher than their TRAIN data counterparts.
- While the Linear Regression model has a slightly higher RMSE compared to the other models, it performs better than them on the TEST data.
- Hence, the Linear Regression model is recommended for its good balance between training and test performance. It shows better generalization to unseen data compared to the Decision Tree and Random Forest models.

CONCLUSION

In the pursuit of developing a robust predictive model for Airbnb listing prices, a comprehensive analysis was conducted, comparing Linear Regression, Decision Tree, and Random Forest models. The goal is to unravel patterns and predict Airbnb listing prices.

The dataset has 8,528 rows and 36 variables, with 23 numeric and 13 character variables. The Dependent Variable (Price) exhibited right-skewed distribution with extreme values. A log transformation is applied to address skewness and outliers. Transformation and categorization are applied to numeric variables to enhance the analysis. Multicollinearity is assessed, and non-normal variables are transformed or categorized. 3 character variables are analyzed and categorized based on their cumulative frequencies.

The dataset is then split into TRAIN (80%) and TEST (20%) sets for model training and evaluation, ensuring realistic assessment on unseen data. Variables considered include transformed/categorized numeric and categorical variables.

The Linear Regression model emerged as the optimal choice, showcasing superior generalization on new, unseen data. Notably, variables like 'maximum nights' and 'rating' were excluded from the final model, indicating their insignificant influence on listing prices. Intriguingly, host ratings were found to have no discernible impact on prices, suggesting hosts might not actively leverage positive reviews to adjust their pricing strategy.

Parameter estimates, though challenging to interpret directly due to log transformations, provided valuable insights. Listings with over 3 bathrooms, Entire homes for rent, extra bedrooms, higher number of guests, and those located in high-demand neighborhoods tended to command higher prices. On the other end, Decision Tree and Random Forests displayed potential overfitting with poor performance on the test data.

In conclusion, the Linear Regression model, enriched by the insights from the LASSO variable selection process, proves to be a reliable tool for predicting Airbnb listing prices. The model's successful application on the test dataset affirms its generalization capabilities and underscores its practical utility for stakeholders in the Airbnb ecosystem.

APPENDICES

LIST OF OUTLIERS

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
42933476	South Shore	Private room	1	1	.	1	$12.00
45487005	Logan Square	Entire home/apt	1	0.5	.	1	$13.00
43414161	South Chicago	Private room	1	1	.	1	$16.00
41911629	South Chicago	Private room	2	1	.	1	$16.00
42471401	South Chicago	Private room	2	1	.	1	$16.00
43055338	South Shore	Private room	2	1	.	1	$16.00
9.6880953E17	Near North Side	Entire home/apt	4	1	1	2	$16.00
9.2417447E17	South Chicago	Private room	2	1	.	1	$16.00
41911779	South Chicago	Private room	2	1	.	1	$17.00
54359408	South Shore	Shared room	1	1	.	1	$17.00
9.7651989E17	South Chicago	Private room	2	1	1	1	$18.00
42366362	South Chicago	Private room	1	1	.	1	$18.00
26317308	Beverly	Private room	1	0	1	1	$19.00
6039038	West Ridge	Shared room	2	2	.	1	$19.00
41528983	South Chicago	Entire home/apt	4	1	1	1	$19.00
23122592	West Ridge	Shared room	2	2	.	1	$19.00
46282113	West Ridge	Shared room	1	2	.	1	$20.00
9.2418202E17	South Chicago	Private room	2	1	.	1	$20.00
23123158	West Ridge	Shared room	1	2	.	1	$20.00
5.5363775E17	West Ridge	Shared room	1	2	.	1	$20.00
3742513	West Ridge	Shared room	1	2	.	1	$20.00
9.7564026E17	South Chicago	Private room	2	1	1	1	$20.00
22478199	West Ridge	Shared room	1	2	.	1	$20.00
39870266	Avondale	Private room	1	1	.	1	$20.00
8.782086E17	West Ridge	Shared room	1	2	.	1	$20.00

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
1461451	West Ridge	Shared room	1	2	.	1	$20.00
40201619	South Shore	Shared room	1	1	.	1	$20.00
21062555	Lake View	Shared room	1	1	.	1	$21.00
4365466	West Ridge	Shared room	1	2	.	1	$21.00
7.4590194E17	East Side	Private room	1	1	.	1	$21.00
1562331	Portage Park	Private room	1	1	.	1	$21.00
7.8910751E17	Lincoln Park	Entire home/apt	4	1	2	2	$21.00
41620912	South Chicago	Private room	2	1	.	2	$21.00
51710533	Belmont Cragin	Private room	1	1	.	1	$21.00
34081264	South Shore	Private room	2	1	.	1	$21.00
40423941	South Shore	Private room	2	1	.	1	$21.00
9.6162585E17	West Pullman	Shared room	1	1	.	1	$22.00
8.5657806E17	South Chicago	Private room	2	1	.	1	$22.00
37865528	West Garfield Park	Private room	1	3	.	1	$22.00
9.1622242E17	New City	Private room	2	1	.	1	$23.00
6.5745044E17	New City	Private room	2	1	.	1	$23.00
6.5302124E17	Grand Boulevard	Private room	2	1	.	1	$23.00
7.4589307E17	East Side	Private room	2	1	.	1	$24.00
6.7047387E17	West Englewood	Private room	1	1.5	.	1	$24.00
54385917	South Shore	Private room	1	1	.	1	$24.00
9.7774661E17	West Englewood	Entire home/apt	4	2	3	3	$24.00
9.6619256E17	West Pullman	Shared room	1	1	.	1	$24.00
7.7703479E17	Pullman	Private room	2	1	.	2	$24.00
9.6333295E17	West Englewood	Private room	1	1.5	.	1	$24.00
42315805	South Chicago	Private room	2	1	.	1	$24.00
16972979	Garfield Ridge	Private room	2	1	1	1	$25.00
5283285	Lower West Side	Private room	2	1	.	1	$25.00
8.9007967E17	New City	Private room	2	1	.	1	$25.00
39696721	Englewood	Shared room	4	2	.	3	$25.00
39277763	South Deering	Private room	1	1	.	1	$25.00

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
38394654	Greater Grand Crossing	Private room	2	1	.	1	$25.00
37866160	West Garfield Park	Private room	1	3	.	1	$25.00
31058338	Calumet Heights	Shared room	1	4	.	1	$25.00
6.5386659E17	New City	Private room	2	1	.	1	$25.00
23467251	West Englewood	Shared room	4	2	.	.	$25.00
6.4719749E17	Woodlawn	Private room	1	1	.	1	$25.00
34137266	South Shore	Shared room	2	1	.	1	$25.00
26162767	Calumet Heights	Shared room	1	4	.	1	$25.00
45565646	South Shore	Private room	2	1	.	1	$25.00
49876778	Calumet Heights	Shared room	1	4	.	2	$25.00
9.0609857E17	New City	Private room	2	1	.	1	$25.00
39697278	West Englewood	Shared room	1	2	.	4	$25.00
27168613	Fuller Park	Entire home/apt	4	1	2	2	$25.00
41275135	South Lawndale	Shared room	1	0.5	.	1	$25.00
9.529657E17	Woodlawn	Private room	1	1	.	1	$26.00
27979141	Calumet Heights	Shared room	1	3	.	1	$26.00
8.6317703E17	Grand Boulevard	Private room	1	1	1	1	$26.00
8.5438728E17	South Chicago	Private room	1	1.5	.	1	$26.00
26628029	Calumet Heights	Shared room	1	4	.	1	$26.00
44259231	South Shore	Private room	4	1	.	1	$26.00
9.0680212E17	Brighton Park	Private room	2	1	.	1	$26.00
25971516	Calumet Heights	Shared room	1	1	.	1	$26.00
5.7261228E17	New City	Private room	2	1	.	1	$26.00
9.6447755E17	Woodlawn	Private room	1	1	.	1	$26.00
33928454	Calumet Heights	Shared room	1	3	.	1	$26.00
51252382	Calumet Heights	Shared room	1	4	.	1	$26.00
9.1623658E17	New City	Private room	2	1	.	1	$26.00
22629758	Uptown	Private room	3	1	.	2	$26.00
7.3605478E17	East Side	Private room	2	1	.	1	$26.00

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
37056214	Irving Park	Entire home/apt	14	3	7	7	$1,254.00
18876938	West Town	Entire home/apt	12	2	3	4	$1,286.00
30811588	Rogers Park	Entire home/apt	14	7	7	10	$1,286.00
8.7258365E17	West Town	Entire home/apt	14	3.5	4	7	$1,290.00
8.0131187E17	Loop	Entire home/apt	16	6	5	24	$1,294.00
7.0785592E17	Lower West Side	Entire home/apt	10	2.5	3	6	$1,328.00
45719938	West Town	Entire home/apt	8	3.5	4	3	$1,333.00
3172794	Logan Square	Entire home/apt	16	2.5	6	9	$1,350.00
28115625	West Town	Entire home/apt	3	1	.	2	$1,352.00
7.9070741E17	Near West Side	Entire home/apt	16	8	8	8	$1,378.00
50112403	West Town	Entire home/apt	12	4	5	5	$1,383.00
9.0505434E17	Near North Side	Private room	8	2	2	2	$1,393.00
9.0505465E17	Near North Side	Private room	8	2	2	2	$1,393.00
9.0505457E17	Near North Side	Private room	8	2	2	2	$1,393.00
13884932	Lake View	Entire home/apt	16	4.5	6	9	$1,422.00
9.6385799E17	Loop	Private room	16	0	.	8	$1,428.00
28238637	West Town	Entire home/apt	2	1	.	1	$1,429.00
28238620	West Town	Entire home/apt	2	1	.	1	$1,429.00
28238665	West Town	Entire home/apt	2	1	.	1	$1,429.00
28238649	West Town	Entire home/apt	2	1	.	1	$1,429.00

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
8.7341852E17	Lake View	Entire home/apt	2	1	1	1	$1,500.00
15216226	North Center	Entire home/apt	10	3.5	5	5	$1,500.00
48680437	Near North Side	Entire home/apt	2	1	1	1	$1,500.00
11456952	Douglas	Entire home/apt	8	3.5	4	4	$1,500.00
9.4283918E17	Loop	Entire home/apt	8	3.5	4	4	$1,503.00
36200479	Logan Square	Entire home/apt	16	3	4	6	$1,528.00
21321497	Logan Square	Entire home/apt	16	4	8	8	$1,550.00
9.0506128E17	Near North Side	Private room	8	2	2	2	$1,559.00
9.0506107E17	Near North Side	Private room	8	2	2	2	$1,559.00
9.0506095E17	Near North Side	Private room	8	2	2	2	$1,559.00
9.1349104E17	Near North Side	Private room	8	2	2	2	$1,566.00
9.1349112E17	Near North Side	Private room	8	2	2	2	$1,566.00
9.1349099E17	Near North Side	Private room	8	2	2	2	$1,566.00
9.4282655E17	Loop	Entire home/apt	6	3	3	3	$1,581.00
49399182	Near North Side	Entire home/apt	16	6	6	12	$1,607.00
33376719	Logan Square	Entire home/apt	16	5.5	8	12	$1,613.00
6.4997494E17	Near North Side	Entire home/apt	12	7.5	5	5	$1,628.00
29898010	West Town	Entire home/apt	10	3.5	5	8	$1,650.00
30718928	West Town	Entire home/apt	3	1	.	2	$1,653.00
30718918	West Town	Entire home/apt	3	1	.	2	$1,653.00
9.1349058E17	Near North Side	Private room	8	2	2	2	$1,750.00

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
9.1349055E17	Near North Side	Private room	8	2	2	2	$1,750.00
9.1349052E17	Near North Side	Private room	8	2	2	2	$1,750.00
6.9989344E17	Near North Side	Entire home/apt	14	6.5	5	6	$1,943.00
7.595811E17	Loop	Entire home/apt	16	10	10	10	$1,960.00
6.2091654E17	Loop	Entire home/apt	16	9.5	10	15	$1,977.00
21900672	Logan Square	Entire home/apt	12	2	4	5	$1,999.00
34336556	Mckinley Park	Entire home/apt	6	1	3	3	$2,000.00
9.0505283E17	Near North Side	Private room	12	3	3	3	$2,018.00
9.0505273E17	Near North Side	Private room	12	3	3	3	$2,018.00
9.0504621E17	Near North Side	Private room	12	3	3	3	$2,018.00
7.4075797E17	West Town	Entire home/apt	16	7.5	9	19	$2,051.00
7.8102779E17	Near North Side	Entire home/apt	12	5	6	6	$2,057.00
6.4548437E17	West Town	Entire home/apt	8	3.5	4	4	$2,070.00
34890754	West Town	Entire home/apt	14	3.5	4	7	$2,249.00
9.0505382E17	Near North Side	Private room	12	3	3	3	$2,267.00
9.050537E17	Near North Side	Private room	12	3	3	3	$2,267.00
9.0505329E17	Near North Side	Private room	12	3	3	3	$2,267.00
9.1349411E17	Near North Side	Private room	12	3	3	3	$2,278.00
9.1349391E17	Near North Side	Private room	12	3	3	3	$2,278.00
9.1349373E17	Near North Side	Private room	12	3	3	3	$2,278.00
35351226	West Town	Entire home/apt	16	7	8	9	$2,358.00
50890121	West Town	Entire home/apt	16	6	6	4	$2,429.00
9.1349164E17	Near North Side	Private room	12	3	3	3	$2,553.00

id	Area	Description	Capacity	bathrooms	bedrooms	beds	price
9.134915E17	Near North Side	Private room	12	3	3	3	$2,553.00
9.1349145E17	Near North Side	Private room	12	3	3	3	$2,553.00
33940403	Lincoln Park	Entire home/apt	12	4.5	6	7	$2,599.00
9.0504533E17	Near North Side	Private room	16	4	4	4	$2,643.00
9.0504526E17	Near North Side	Private room	16	4	4	4	$2,643.00
9.050452E17	Near North Side	Private room	16	4	4	4	$2,643.00
7.4061693E17	Near North Side	Entire home/apt	12	6	5	7	$2,743.00
9.6385374E17	Near North Side	Private room	3	0	.	1	$2,973.00
9.0504561E17	Near North Side	Private room	16	4	4	4	$2,975.00
9.0504552E17	Near North Side	Private room	16	4	4	4	$2,975.00
9.1349816E17	Near North Side	Private room	16	4	4	4	$2,990.00
9.1349794E17	Near North Side	Private room	16	4	4	4	$2,990.00
9.1349768E17	Near North Side	Private room	16	4	4	4	$2,990.00
6.9992751E17	Lincoln Park	Entire home/apt	12	3	5	6	$3,000.00
6.4147913E17	Bridgeport	Entire home/apt	16	4	5	8	$3,000.00
9.1349667E17	Near North Side	Private room	16	4	4	4	$3,357.00
9.1349655E17	Near North Side	Private room	16	4	4	4	$3,357.00
9.134964E17	Near North Side	Private room	16	4	4	4	$3,357.00
35060034	West Town	Entire home/apt	16	6	9	13	$4,500.00
5185790	West Town	Entire home/apt	8	3.5	3	4	$5,000.00
14167586	West Town	Private room	4	1	.	1	$7,585.00

Summary – Numeric Variables

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev
host_response_rate	host_response_rate	7480	866	0	1.0000000	0.9725227	1.0000000	0.1210753
host_acceptance_rate	host_acceptance_rate	7620	726	0	1.0000000	0.9198648	0.9800000	0.1776646
accommodates	accommodates	8346	0	1.0000000	16.0000000	4.2776180	4.0000000	2.9082589
bathrooms	bathrooms	8335	11	0	12.5000000	1.4017397	1.0000000	0.7534040
bedrooms	bedrooms	6224	2122	1.0000000	13.0000000	2.0562339	2.0000000	1.1359690
beds	beds	8301	45	1.0000000	21.0000000	2.2308156	2.0000000	1.6497447
availability_30	availability_30	8346	0	0	30.0000000	12.6925473	12.0000000	10.3246700
number_of_reviews	number_of_reviews	8346	0	0	3332.00	47.2592859	15.0000000	88.4601828
review_scores_rating	review_scores_rating	6760	1586	0	5.0000000	4.7406686	4.8600000	0.4522300
review_scores_accuracy	review_scores_accuracy	6749	1597	1.0000000	5.0000000	4.7905275	4.9000000	0.3886814
review_scores_cleanliness	review_scores_cleanliness	6748	1598	1.0000000	5.0000000	4.7408002	4.8600000	0.4171127
review_scores_checkin	review_scores_checkin	6749	1597	1.0000000	5.0000000	4.8397200	4.9400000	0.3524817
review_scores_communication	review_scores_communication	6748	1598	1.0000000	5.0000000	4.8511841	4.9500000	0.3642932
review_scores_location	review_scores_location	6749	1597	1.0000000	5.0000000	4.7574470	4.8800000	0.3739335
review_scores_value	review_scores_value	6749	1597	1.0000000	5.0000000	4.6702874	4.7800000	0.4371942
reviews_per_month	reviews_per_month	6760	1586	0.0100000	108.1000000	1.9465902	1.6200000	2.2340776

FREQ Numeric Variables

accommodates				
accommodates	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	481	5.76	481	5.76
2	2704	32.40	3185	38.16
3	549	6.58	3734	44.74
4	1830	21.93	5564	66.67
5	490	5.87	6054	72.54
6	1091	13.07	7145	85.61
7	188	2.25	7333	87.86
8	415	4.97	7748	92.83
9	59	0.71	7807	93.54
10	193	2.31	8000	95.85
11	17	0.20	8017	96.06
12	134	1.61	8151	97.66
13	14	0.17	8165	97.83
14	49	0.59	8214	98.42
15	23	0.28	8237	98.69
16	109	1.31	8346	100.00

availability_30				
availability_30	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1923	23.04	1923	23.04
1	140	1.68	2063	24.72
2	150	1.80	2213	26.52
3	161	1.93	2374	28.44
4	169	2.02	2543	30.47
5	186	2.23	2729	32.70
6	179	2.14	2908	34.84
7	171	2.05	3079	36.89
8	230	2.76	3309	39.65
9	195	2.34	3504	41.98
10	239	2.86	3743	44.85
11	277	3.32	4020	48.17
12	289	3.46	4309	51.63
13	233	2.79	4542	54.42
14	237	2.84	4779	57.26
15	238	2.85	5017	60.11
16	228	2.73	5245	62.84
17	212	2.54	5457	65.38
18	224	2.68	5681	68.07
19	249	2.98	5930	71.05
20	212	2.54	6142	73.59
21	184	2.20	6326	75.80
22	184	2.20	6510	78.00
23	201	2.41	6711	80.41
24	147	1.76	6858	82.17
25	156	1.87	7014	84.04
26	132	1.58	7146	85.62
27	147	1.76	7293	87.38
28	135	1.62	7428	89.00

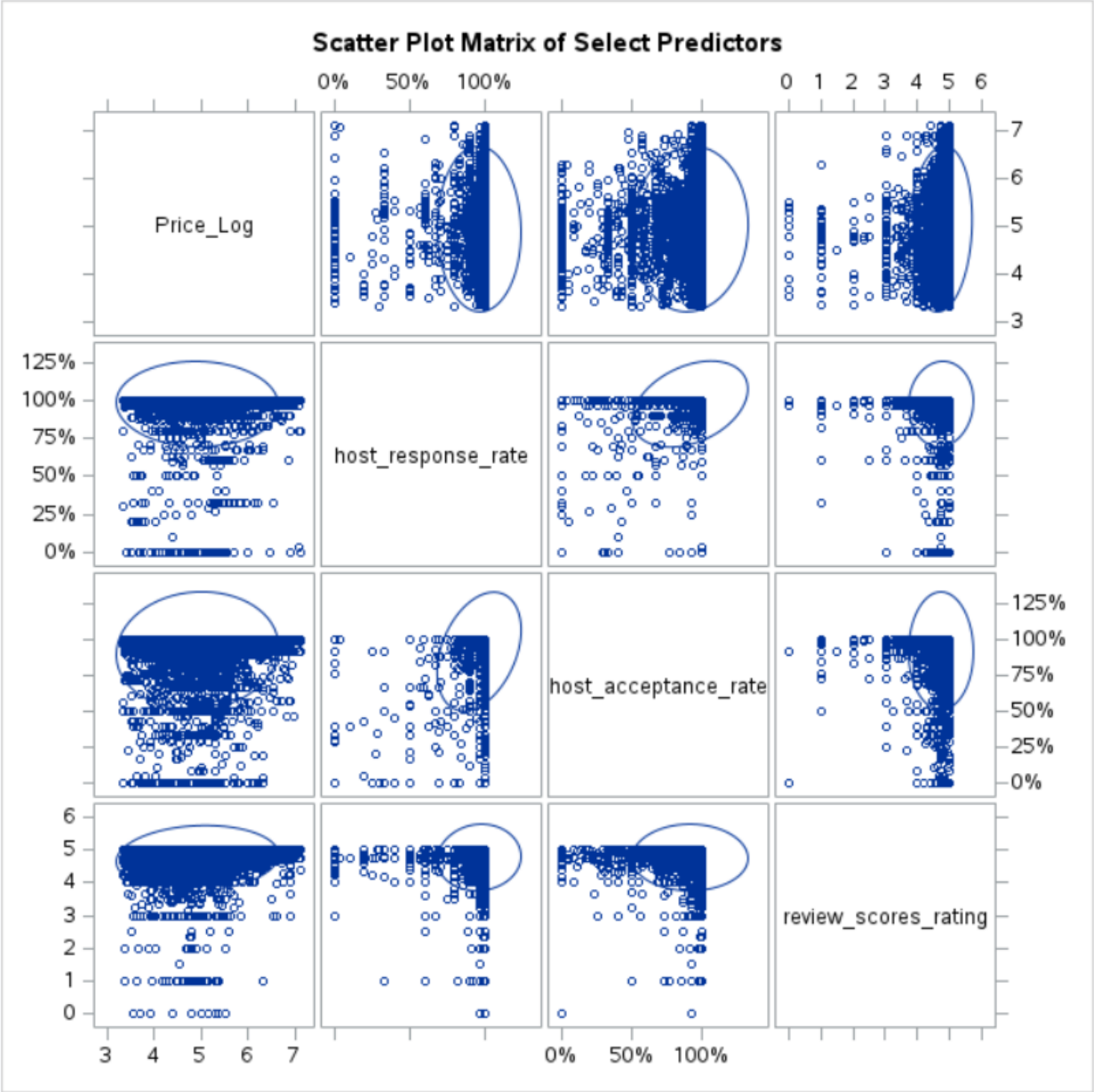
availability_30				
availability_30	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
29	224	2.68	7652	91.68
30	694	8.32	8346	100.00

bathrooms				
bathrooms	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	43	0.52	43	0.52
0.5	5	0.06	48	0.58
1	5581	66.87	5629	67.45
1.5	380	4.55	6009	72.00
2	1634	19.58	7643	91.58
2.5	244	2.92	7887	94.50
3	230	2.76	8117	97.26
3.5	117	1.40	8234	98.66
4	112	1.34	8346	100.00

bedrooms				
bedrooms	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	4426	53.03	4426	53.03
2	2198	26.34	6624	79.37
3	1134	13.59	7758	92.95
4	382	4.58	8140	97.53
5	206	2.47	8346	100.00

beds				
beds	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
1	3610	43.25	3610	43.25
2	2132	25.55	5742	68.80
3	1313	15.73	7055	84.53
4	637	7.63	7692	92.16
5	292	3.50	7984	95.66
6	158	1.89	8142	97.56
7	63	0.75	8205	98.31
8	55	0.66	8260	98.97
9	86	1.03	8346	100.00

Scatterplot Matrix : Non-normal Predictors



Parameter Estimates – LASSO

Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	5.338822
min_nights_tf	1	-0.035243
bpp_tf	1	-0.821920
num_reviews_tf	1	-0.026113
instant_bookable_f	1	-0.004713
accom_cat_Large	1	0.124073
accom_cat_Very_Small	1	0
avail_cat_High	1	0.166501
avail_cat_Low	1	-0.005042
bath_cat_Lux	1	0.429367
bath_cat_Nor	1	-0.322385
bedrooms_cat_Extra	1	0.156548
bedrooms_cat_Single	1	-0.068519
beds_Cat_Extra	1	0.031877
resp_cat_Bad	1	0.060667
acc_cat_Moderate	1	-0.037877
acc_cat_Zero	1	-0.033755
hood_cat_High-Demand	1	0.284941
hood_cat_Low-Demand	1	-0.128116
property_cat_Others	1	0.047461
property_cat_Primary	1	-0.002212
room_cat_Home	1	0.340427