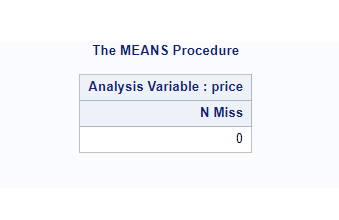
Interpretation of Results:

1. Import the data
2. Fit a linear regression model using price as the dependent variable.
   1. Investigate the dependent variable:
      1. Are there any missing or nonsensical values?



Answer: The dependent variable has 0 missing variables.

* + 1. If you identified any outliers, include the corresponding output and consider removing them from the dataset (see Program 2.7 on 52 for further information)

Answer:

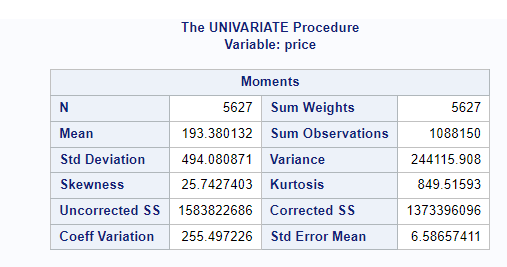
No outliers were identified in the filtered dataset (30 to 750) based on the 1.5 interquartile range rule. The UNIVARIATE output indicates all values within this range, with no extreme observations beyond this threshold. The data appears to be free of outliers after the filtering and log transformation.

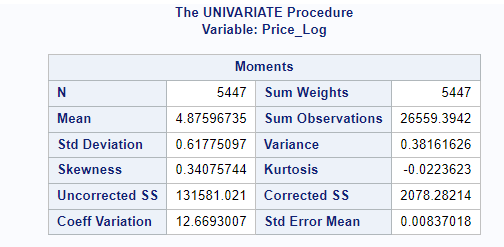
* + 1. PROC UNIVARIATE will also provide you with information about the shape of the distribution of the PRICE variable. Ideally, we would like it to be normally distributed. Different transformations are available for significant departures from normality. Our book uses the log transformation, but other common options include:

Answer: Price\_Log = LOG(Price); I opted this log transformation.

* + 1. If you find the data to be skewed, below are some common guidelines on which transformation to choose for a given skewness level. Program 2.7 includes an example of a log transformation in action.

Answer:





I have opted Price\_Log = LOG(Price); which reduced the Skewness to 0.34

* 1. Investigate the predictors:
     1. Numeric variables:
        1. Is predictor multicollinearity an issue? If so, make sure that you address the issue.

Answer: Yes, predictor multicollinearity appears to be an issue in the model. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. One way to assess multicollinearity is by examining the Variance Inflation Factor (VIF) for each predictor variable.

In the output, the VIF values are provided for each variable. Generally, a VIF value above 10 is considered high, indicating a potential issue with multicollinearity. Looking at the results:

"availability\_30" has a VIF of 8.64636.

"availability\_60" has a very high VIF of 39.00894.

"availability\_90" also has a relatively high VIF of 22.42568.

"beds" and "accommodates" have VIF values above 5.

These high VIF values suggest that there is a high degree of multicollinearity among these variables. Multicollinearity can make it challenging to interpret the individual coefficients of the predictors and can lead to unstable estimates. It's often advisable to address multicollinearity by either removing one of the correlated variables, combining them, or using regularization techniques in your modeling approach.

Regularization methods could help address the multicollinearity.

* + - 1. Categorization – are there any categorical variables with many levels (for instance, more than 5-6)?

Answer: I have taken all the categorical variables and identified the variables with many levels, I have cleaned the Text variables and Replaced the Problematic variables with actual variables.

And Finally Collapsed 3 Categorial variables like neighbourhood\_cleansed, property\_type, host\_response\_time.

* + - 1. Missing values – do any continuous variables include missing values?

Answer: Variables without Missing Values:

latitude, longitude, accommodates, bedrooms,

minimum\_nights, maximum\_nights, availability\_30,

availability\_60, availability\_90, availability\_365, number\_of\_reviews,

Considering the guideline of less than 5% missing values, it seems

that beds is the only variable that might be considered for Impitation.

Feature Engineering:

* Generates higher-order terms (squared terms) for certain numeric variables.
* Standardizes numeric variables using mean and standard deviation.

Scatter Plot Visualization:

* Creates a scatter plot to visualize the relationship between accommodates and Price.

Variable Information:

* Provides information about numeric variables in the dataset.

Frequency Tables for Categorical Variables:

* Generates frequency tables for various categorical variables.

Checking for Missing Values:

* Checks for missing values in a specified range of variables.

Data Transformation and Splitting:

* Collapses categorical levels for the Neighbourhood\_Group variable.
* Splits the data into training and testing datasets at an 80/20 ratio using simple random sampling.

LASSO Regression (PROC GLMSELECT):

Model Selection: LASSO with 10 selected effects.

Selected Effects: Intercept, accommodates, bedrooms, beds, minimum\_nights, number\_of\_reviews\_l3, review\_scores\_locati, and others.

Fit Statistics: R-Square: 0.5299, Adjusted R-Square: 0.5290, Root MSE: 0.38136.

Assessment: The model explains about 52.99% of the variance in the dependent variable. The RMSE provides a measure of the model's predictive accuracy.

Decision Tree (PROC HPSPLIT):

Tree Depth: 10, Number of Leaves After Pruning: 82.

Fit Statistics: Average Square Error for Training: 0.000420, Validation: 0.000906.

Variable Importance: Indicates the importance of different variables in predicting the target variable.

Random Forest (PROC HPFOREST):

Number of Trees: 500.

Fit Statistics: Average Square Error for Training and Out-of-Bag (OOB) data presented for multiple trees.

Variable Importance: Relative importance of variables.

Model Comparison and Recommendations:

A sparse model with chosen characteristics is produced via the LASSO regression method. Excellent for handling multicollinearity and interpretability.

Decision Tree: May be prone to overfitting, yet it captures non-linear interactions. Interpretability has its advantages.

Random Forest: An ensemble technique with high prediction precision. fights overfitting by using several trees.

Suggestions:

Interpretability: LASSO or Decision Tree may be chosen if interpretability is important.

Predictive Accuracy: Random Forest might work well if this is the main objective.