Liana Cramb

ECO 520

**Assignment 5**

Table of Contents

[Missing Values 1](#_Toc190000060)

[Potential Outliers 2](#_Toc190000061)

[Nonlinear Relationships with Price 5](#_Toc190000062)

[Running the Independent Variables Squared: 5](#_Toc190000063)

[Running the Independent Variables Logarithmic: 7](#_Toc190000064)

[Machine Learning with Regression Analysis 10](#_Toc190000065)

[Adjusted R-Squared Regression Model: 11](#_Toc190000066)

[Stepwise Regression Model: 11](#_Toc190000067)

[My Regression Model: 12](#_Toc190000068)

[Performance of the Models: 15](#_Toc190000069)

# Missing Values

**The MEANS Procedure**

| **Variable** | **N** | **N Miss** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- |
| ListingMonth  host\_total\_listings  accommodates  bathrooms  bedrooms  beds  guests\_included  minimum\_nights  number\_of\_reviews  review\_scores\_rating  reviews\_per\_month  PricePerNight  hostclass  popular\_host  big\_unit  longterm | 2000  2000  2000  2000  2000  2000  2000  2000  2000  2000  2000  2000  1890  2000  2000  2000 | 0  0  0  0  0  0  0  0  0  0  0  0  110  0  0  0 | 4.4043500  57.1670000  4.6530000  1.3835000  1.7570000  2.3895000  2.4830000  3.7640000  51.2640000  95.3760000  2.3874150  143.5700000  1.7333333  0.2605000  0.0795000  0.0630000 | 2.2910108  206.5244738  3.0031817  0.7203730  1.1254263  1.8202875  2.1151709  9.0724560  63.6786761  5.8313486  1.9447124  122.3724299  0.8008730  0.4390172  0.2705852  0.2430237 | 0.3000000  0  1.0000000  0  0  0  1.0000000  1.0000000  1.0000000  20.0000000  0.0200000  6.0000000  1.0000000  0  0  0 | 11.6000000  1283.00  24.0000000  11.0000000  9.0000000  20.0000000  16.0000000  180.0000000  528.0000000  100.0000000  12.5500000  979.0000000  3.0000000  1.0000000  1.0000000  1.0000000 |

The dummy variable hostclass had 110 missing values because the SAS code originally classified hosts into three categories: those with 1 to 2 listings as hostclass = 1, those with 3 to 14 listings as hostclass = 2, and those with 15 or more listings as hostclass = 3. However, hosts with 0 listings did not fit into any of these categories, resulting in missing values for hostclass in the output dataset. Since these observations only had missing values for hostclass but contained valuable information for other variables, removing them would not be beneficial. To address this, I modified the SAS code to assign hostclass = 0 for hosts with 0 listings, which eliminated the missing values for the hostclass variable.

## Potential Outliers

**Box Plot for Variables with Little Outliers**

A chart of a graph

AI-generated content may be incorrect.

**Box Plots for Variables with Significant Outliers**

A screen shot of a computer

AI-generated content may be incorrect.

A blue and black drawing of a rectangular object

AI-generated content may be incorrect.

A diagram of a line

AI-generated content may be incorrect.

After running the box plots to identify outliers, the first five variables showed some outliers, but not in such large quantities that would suggest they should be removed. The variables review\_scores\_rating and reviews\_per\_month had more outliers, while PricePerNight exhibited significant outliers. However, based on the nature of these variables, wide variations are expected and are not necessarily problematic.

For reviews\_per\_month, some listings receive many reviews, especially those in popular and touristy neighborhoods, while others may only get a few or none. As a result, it is normal for there to be a wide range of values. Similarly, review\_scores\_rating can vary greatly due to factors like the host’s quality of service, property conditions, guest expectations, and the subjective nature of reviews. Both very positive and very negative reviews can skew the ratings, which is expected in this context. For PricePerNight, variability is also anticipated due to factors such as location, amenities, seasonal price changes, and the host’s discretion in setting the price. Since the price is determined by the individual host, it’s reasonable to expect a wide range in values.

Given the context of these variables, although the outliers should be acknowledged, there is no need to remove them. They are not unusual and likely represent valid data points that reflect the natural variability in Airbnb listings. Therefore, it is important to keep these outliers in the dataset, as they can provide valuable insights for the analysis.

# Nonlinear Relationships with Price

## Running the Independent Variables Squared:

**Price Per Night based on # of Bedrooms Squared**

**A graph of blue and white dots

AI-generated content may be incorrect.**

**Price Per Night based on # of Beds Squared**

**A graph of a square bed

AI-generated content may be incorrect.**

**Price Per Night based on the # of Bathrooms Squared**

A graph of blue circles

AI-generated content may be incorrect.

When squaring the variables bedrooms, bathrooms, and beds, the scatter plots did not reveal a clear linear relationship. However, they also failed to provide meaningful insight into potential non-linear relationships. This was due to the x-axis scale, where the squared values caused the data points to be tightly clustered, making it difficult to observe any distinct patterns.

**The MEANS Procedure**

| **Variable** | **N** | **Mean** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- |
| bedrooms  beds  bathrooms | 2000  2000  2000 | 1.7570000  2.3895000  1.3835000 | 0  0  0 | 9.0000000  20.0000000  11.0000000 |

Squaring the three variables to identify a non-linear relationship is not the best approach due to the way these variables are distributed. Since the average values for bedrooms, bathrooms, and beds range between 1 and 2, squaring them does not significantly alter their scale. As a result, the scatter plots do not reveal a clear non-linear trend.

## Running the Independent Variables Logarithmic:

**# of Bathrooms Impact on Price Per Night**

A graph of blue dots

AI-generated content may be incorrect.

**# of Beds Impact on Price Per Night**

A graph of blue dots

AI-generated content may be incorrect.

**# of Bedrooms Impact on Price Per Night**

A graph of blue dots

AI-generated content may be incorrect.

The scatter plots showing the log-transformed values of each independent variable against log(PricePerNight) reveal some degree of non-linearity. The slight curve near the top right suggests diminishing returns— as the number of bedrooms, beds, or bathrooms increases, their impact on price decreases. Using log transformations instead of squaring helps better represent the non-linear relationships, making the model more balanced and easier to interpret.

Log transformation is useful when small changes in a variable have a larger impact at lower values but a smaller impact at higher values. For example, an Airbnb listing increasing from 1 to 2 bedrooms typically leads to a significant price increase, as many renters would see this as a major upgrade. However, going from 4 to 5 bedrooms might raise the price by a smaller amount, since it is likely that fewer Airbnb guests require that many rooms. This same pattern can be applied to the number of beds and bathrooms, where additional units have a diminishing impact on price as the quantity increases. As a result, hosts are more likely to adjust prices more aggressively when the change appeals to a broader customer base.

**# of Guests Included in the Price Impact on Price Per NightA graph of blue dots

AI-generated content may be incorrect.**

Out of the other numerical variables, the variable guests\_included shows a somewhat non-linear relationship with PricePerNight. There is a slight curve near the right side, where we observe the highest prices are slowly decreasing as the number of guests included increases. This could be explained by several factors. For instance, smaller groups (e.g., 1-2 guests) typically experience a larger price increase per additional guest, as hosts tend to cater more to solo travelers or couples. In contrast, larger groups (e.g., 4+ guests) may see a smaller price increase because there is a limit to how much hosts can charge based on the unit's space and amenities, which can cause the price increase to plateau. Additionally, hosts may offer discounts for larger groups to attract more bookings, leading to diminishing returns for additional guests.

# Machine Learning with Regression Analysis

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -181.49855 | 38.26050 | -4.74 | <.0001 |
| **ListingMonth** | 1 | -1.74550 | 1.01968 | -1.71 | 0.0871 |
| **hostclass** | 1 | 23.97444 | 2.66778 | 8.99 | <.0001 |
| **popular\_host** | 1 | -13.14492 | 6.10125 | -2.15 | 0.0313 |
| **big\_unit** | 1 | 37.32201 | 10.82925 | 3.45 | 0.0006 |
| **longterm** | 1 | -26.31842 | 9.50402 | -2.77 | 0.0057 |
| **bathrooms** | 1 | 18.09907 | 3.80956 | 4.75 | <.0001 |
| **bedrooms** | 1 | 22.61266 | 3.40519 | 6.64 | <.0001 |
| **beds** | 1 | 11.28289 | 2.13921 | 5.27 | <.0001 |
| **guests\_included** | 1 | 4.36821 | 1.28686 | 3.39 | 0.0007 |
| **review\_scores\_rating** | 1 | 2.22851 | 0.38904 | 5.73 | <.0001 |
| **reviews\_per\_month** | 1 | -8.17723 | 1.40490 | -5.82 | <.0001 |

After running the regression analysis, I found that all variables were statistically significant except for ListingMonth, which had a p-value greater than 0.05, so I removed it. I then re-examined the remaining variables to check for correlations. I decided to remove beds because it closely correlates with bedrooms—a listing with an additional bedroom generally implies that it includes an additional bed. Additionally, I removed popular\_host as I found it to overlap conceptually with reviews\_per\_month. I opted to keep reviews\_per\_month instead, as I believe it provides a clearer interpretation of its impact on price. Popular\_host divided hosts into two groups: those with 65 or fewer reviews and those with 66 or more. I found that this closely correlates with reviews\_per\_month. In the context of how each could impact PricePerNight, I believe reviews\_per\_month provides a better view.

## Adjusted R-Squared Regression Model:

**The REG Procedure**

**Model: MODEL1**

**Dependent Variable: y**

**Adjusted R-Square Selection Method**

|  |  |
| --- | --- |
| **Number of Observations Read** | 2000 |
| **Number of Observations Used** | 1600 |
| **Number of Observations with Missing Values** | 400 |

| **Number in Model** | **Adjusted R-Square** | **R-Square** | **Variables in Model** |
| --- | --- | --- | --- |
| **11** | 0.3700 | 0.3743 | ListingMonth hostclass popular\_host big\_unit longterm bathrooms bedrooms beds guests\_included review\_scores\_rating reviews\_per\_month |
| **10** | 0.3698 | 0.3737 | hostclass popular\_host big\_unit longterm bathrooms bedrooms beds guests\_included review\_scores\_rating reviews\_per\_month |

Based on the Adjusted R-Square regression model, after evaluating all possible variable combinations for the best fit, the model concluded that using all variables resulted in the highest Adjusted R-Square of 0.3700, or 37%. This suggests that approximately 37% of the variance in PriceperNight is explained by the independent variables. I was initially surprised by the inclusion of ListingMonth, as it had a p-value greater than 0.05. However, after reviewing the analysis, I noticed that excluding it and using only ten variables resulted in a lower Adjusted R-Square value of 0.3698.

## Stepwise Regression Model:

**The REG Procedure**

**Model: MODEL2**

**Dependent Variable: y**

|  |  |
| --- | --- |
| **Number of Observations Read** | 2000 |
| **Number of Observations Used** | 1600 |
| **Number of Observations with Missing Values** | 400 |

**All variables left in the model are significant at the 0.1500 level.**

**No other variable met the 0.1500 significance level for entry into the model.**

| **Summary of Stepwise Selection** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | bedrooms |  | 1 | 0.2725 | 0.2725 | 250.475 | 598.52 | <.0001 |
| **2** | beds |  | 2 | 0.0314 | 0.3039 | 172.733 | 72.08 | <.0001 |
| **3** | reviews\_per\_month |  | 3 | 0.0183 | 0.3222 | 128.211 | 43.16 | <.0001 |
| **4** | hostclass |  | 4 | 0.0183 | 0.3405 | 83.7551 | 44.27 | <.0001 |
| **5** | review\_scores\_rating |  | 5 | 0.0103 | 0.3509 | 59.5095 | 25.39 | <.0001 |
| **6** | big\_unit |  | 6 | 0.0081 | 0.3590 | 40.9917 | 20.09 | <.0001 |
| **7** | bathrooms |  | 7 | 0.0061 | 0.3650 | 27.6014 | 15.20 | 0.0001 |
| **8** | guests\_included |  | 8 | 0.0050 | 0.3701 | 16.8217 | 12.72 | 0.0004 |
| **9** | popular\_host |  | 9 | 0.0023 | 0.3723 | 13.0258 | 5.78 | 0.0163 |
| **10** | longterm |  | 10 | 0.0014 | 0.3737 | 11.5366 | 3.49 | 0.0620 |

The stepwise model resulted in ten independent variables. This is not surprising, as stepwise regression evaluates each variable based on its p-value and includes only those predictors that are statistically significant. As we already know, ListingMonth did not meet the statistical significance threshold, so it was excluded from the model.

## My Regression Model:

**The REG Procedure**

**Model: MODEL3**

**Dependent Variable: y**

|  |  |
| --- | --- |
| **Number of Observations Read** | 2000 |
| **Number of Observations Used** | 1600 |
| **Number of Observations with Missing Values** | 400 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 8 | 8783189 | 1097899 | 112.42 | <.0001 |
| **Error** | 1591 | 15537806 | 9766.06282 |  |  |
| **Corrected Total** | 1599 | 24320995 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 98.82339 | **R-Square** | 0.3611 |
| **Dependent Mean** | 143.83625 | **Adj R-Sq** | 0.3579 |
| **Coeff Var** | 68.70548 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | -187.87514 | 41.89685 | -4.48 | <.0001 |
| **hostclass** | 1 | 22.74814 | 2.93909 | 7.74 | <.0001 |
| **big\_unit** | 1 | 63.70186 | 11.66163 | 5.46 | <.0001 |
| **bathrooms** | 1 | 20.93543 | 4.01683 | 5.21 | <.0001 |
| **bedrooms** | 1 | 33.88532 | 3.20692 | 10.57 | <.0001 |
| **guests\_included** | 1 | 5.55564 | 1.45266 | 3.82 | 0.0001 |
| **longterm** | 1 | -22.20171 | 10.67794 | -2.08 | 0.0378 |
| **review\_scores\_rating** | 1 | 2.20106 | 0.42576 | 5.17 | <.0001 |
| **reviews\_per\_month** | 1 | -8.91157 | 1.31423 | -6.78 | <.0001 |

As mentioned earlier in the report, I only included variables that were statistically significant and removed two variables that I felt were closely correlated with others to eliminate redundancy. My regression model had an Adjusted R-Squared value of 0.3579 (35.79%), suggesting that approximately 35.79% of the variance in PricePerNight is explained by the independent variables. Additionally, the F-value of 112.42 confirms that the model is statistically significant, indicating that at least one independent variable has a nonzero effect on the dependent variable.

Only the numeric variable reviews\_per\_month and the dummy variable longterm had a negative impact, while all the other variables had a positive impact on price per night. This can be expected for variables like bedrooms and bathrooms, as it is natural that when you add another bedroom and bathroom the price increases too. The dummy variable hostclass likely reflects a trend where hosts managing multiple listings tend to price them higher, possibly due to greater experience and pricing flexibility. Similarly, big\_unit, representing listings that accommodate large groups (8+ people), would be associated with higher prices. Guests\_included also logically contributes to price increases, as listings that include more guests in the base price will generally be priced higher. Lastly, review\_scores\_rating, which represents the average rating score for the listing, has a very small but positive impact. A host who sees that their listing has a positive rating average may consider raising their price slightly, as the good reviews could attract more interested customers, allowing the host to test the waters and increase the price.

The dummy variable longterm shows a significant negative impact on PricePerNight due to several factors. Long-term listings, a stay with 8 nights minimum, appeal less to Airbnb’s primary customer base, who likely tends to seek short-term accommodations. This reduces demand and potentially lowers the price. Additionally, long-term listings often compete with traditional rental platforms like Zillow or Apartments.com, prompting hosts to adjust their prices to stay competitive. Similarly, the variable reviews\_per\_month also had a negative impact, likely because a high volume of reviews indicates frequent bookings, often at a lower price to attract more guests. Hosts with frequent reviews may also adjust prices based on customer feedback, which can lead to price reductions in response to dissatisfaction or to remain competitive.

## Performance of the Models:

**The MEANS Procedure**

| **Variable** | **N** | **Mean** |
| --- | --- | --- |
| mse1  mse2  mse3  rmse1  rmse2  rmse3  mpe1  mpe2  mpe3  mae1  mae2  mae3 | 400  400  400  400  400  400  400  400  400  400  400  400 | 10399.61  10434.42  10533.98  67.4173406  67.2944977  68.2345836  0.7413820  0.7371848  0.7525266  67.4173406  67.2944977  68.2345836 |

\* Model 1 = Adjusted R-Squared, Model 2 = Stepwise, and Model 3 = My Own Model

* **Mean Square Error (MSE):** Model 1 was the best with the lowest value of 10399.61.
* **Root Mean Square Error (RMSE):** Model 2 was the best with the lowest value of 67.2944977.
* **Mean Percentage Error (MPE):** Model 2 was the best with the lowest value of .7371848.
* **Mean Absolute Error (MAE):** Model 2 was the best with the lowest value of 67.2944977.

Based on the statistics we measured, the overall best model was Model 2, the stepwise regression model. This is not too surprising since it removed the only variable that was not statistically significant and all three models had similar adjusted R-squared values: Model 1 (.3700), Model 2 (.3698), and Model 3 (.3579). Since the difference in adjusted R-squared values is minimal, it is better to use the one with the best overall performance in terms of other error metrics. Model 2, in this case, also showed the best performance in terms of RMSE, MPE, and MAE.

Lastly, all the regression models show that the independent variables have some impact on PricePerNight, and that they explain less than 40% of its variance. This makes sense, as pricing on Airbnb is highly influenced by the host's personal decisions, and Airbnb does not impose strict pricing rules. Each host likely has their own rationale for determining the price of their listing.