**Understanding Price Variability for Airbnb Rentals in the Seattle Area**

**Business Problem**

Airbnb aims to create a model that predicts how much a residency should charge per night based on attributes of the unit and amenities the property provides. This will allow Airbnb to maximize profits by potentially increasing the cost per nightly stay at certain places and create a more equitable environment for Airbnb users.

**Linear Regression Model**

At first a general linear regression model was built using nightly price as the response variable and many other variables as the explanatory variables to see how a general model would perform

price ~ accommodates + bathrooms + beds to accommodates +

Apartment + House + Townhouse + Condominium + review scores\_ rating + cleaning fee

pct + Days Booked + Family kid friendly + Hot Tub + Safety Card + washer dryer + Indoor

Fireplace + Pets Allowed

This model consisted of 17 parameters with an *R2* = 0.541 and an AIC = 25,520. See Appendix 1 for model summary and estimated coefficients. Unsurprisingly the Prob (F-statistic) value was 0 implying that at least one variable has a significant effect on the response.

**A grid of blue and white squares

Description automatically generatedDetermining the key variable**

* Accommodates, Bathrooms, and Bedrooms are highly correlated numerical variables. We will focus on Accommodates as the main explanatory variable with the highest correlation (0.67)

**Key Consideration in Modeling**

* A graph of a graph

  Description automatically generated with medium confidenceChecking for Heteroskedasticity: the relationship between accommodates and price appear to display heteroskedasticity when visualized (Figure 1). Variance between accommodation and price increases as the number of people a property can accommodate increase. To correct for this variable, a transformation was performed by taking the log of accommodates as well as price. A similar analysis was performed for bathrooms against price as well and the log of bathrooms was taken (Figure 2). Results from this transformation can be found in the appendix (Figures 3 and 4)

Figure 1

A graph of a graph

Description automatically generated with medium confidenceFigure 2

* To check for multicollinearity VIF scores were obtained from the original model to see if any variables received a score equal to or higher than 2. Accommodates is slightly correlated with other variables but not overly correlated where removal is required (Table 3).

**Model Selection**

* The amenities categorical features (Family kid friendly, hot tub, safety card, washer dryer, indoor fireplace, and pets allowed) were excluded to create a smaller model in which a F test was run against the larger model. The scores were compared to see if this smaller model performed better than the larger one. From the results, the P-Value of the model that omitted the amenities = 0.017. If we are to use 0.01 as the significance level than we would fail to reject the null hypothesis that the smaller model is insufficient. Due to these results the amenities factors were removed due to insufficient impact on the model (Table 2). However, based on the P>|t| score from the original model, the Hot Tub variable (0.011) as well as the beta coefficient greater than 2x the standard error (4.61) this amenity feature was not omitted form the model (Table 1)
* A similar test was run for the property type (House, Townhouse, Condominium, and Apartment). These features were excluded from the primary model and an F test was performed. The Pr(>F) score obtained from the smaller model was 1.26e-17. Falling well below the threshold of 0.01 we would reject the null hypothesis that the smaller model is sufficient (Table 2).
* A similar test was run for Days Booked. When this feature was removed, and an F test was performed the Pr(>F) value was 0.94. We fail to reject the null hypothesis that the smaller model is insufficient so Days Booked is removed (Table 2).

**Final Model**

* After removing all necessary features from the original model and taking the log of price, accommodates, and bathrooms we produced the below final model.

np.log(price) ~ np.log(accommodates) + np.log(scaled\_bathrooms) + beds to

accommodates + Apartment + House + Townhouse + Condominium +

review scores rating + cleaning fee pct + Hot Tub

* The modified model produced an R2 = 0.683, compared to the original model’s score of 0.541, and an AIC score of 514, compared to the original model’s score of 25,520. All variables had a significant P>|t| score except for the Hot Tub variable of 0.589.

**Appendix**

A screenshot of a computer

Description automatically generatedTable 1: Original Model

Table 2: Results from three F tests performed. Each index is the features that are removed from the model to form the smaller model.

A screenshot of a computer

Description automatically generated

Table 3: VIF scores of each variable to check for multi collinearity.

A screenshot of a computer program

Description automatically generated

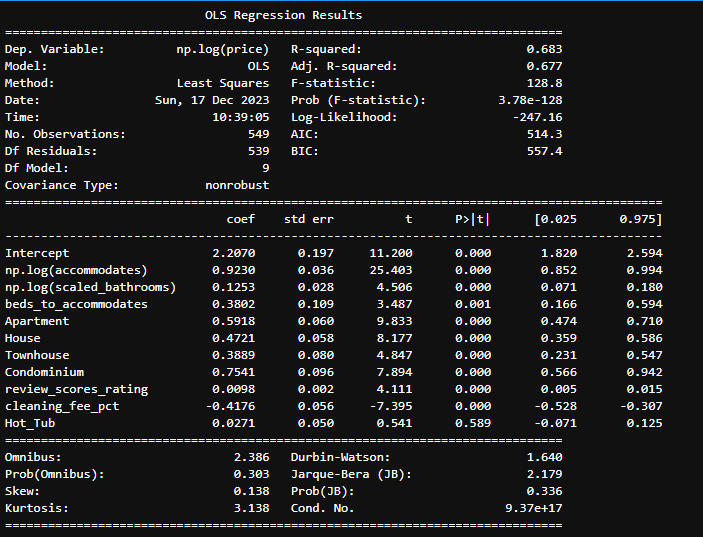
Table 4: Optimized Final Model

Figure 3

A comparison of graphs with numbers

Description automatically generated with medium confidence