Final Project - BTRY 6020

Introduction

The dataset used for this project is the “U.S. Airbnb Open Data,” available on Kaggle. This dataset contains detailed information on Airbnb listings across multiple cities in the United States, including price, room type, minimum nights, number of listings per host, city where located, and more. It provides a comprehensive review of the Airbnb market in the U.S. with both numeric and categorical variables. The goal was to build a linear regression model to understand how these factors affect listing price. After cleaning and preparing the data, this analysis aimed to predict the log-transformed Airbnb listing price (log_price) through the variables provided in the dataset.

Methodology

All steps described below were implemented in R. The complete code is provided in the submitted R Markdown file. These steps include exploratory data analysis, regression assumptions verification, variable selection & hypothesis testing, and feature impact analysis. I began by exploring the structure of the Airbnb dataset through the summary statistics to understand the distribution and range of each variable. I visualized the distribution of price, the response variable, and other key numeric variables, including minimum nights, number of reviews, reviews per month, number of host listings, and year-round availability, to assess their shapes and identify potential skewness or the need for transformation later on. To examine potential relationships with price, I created scatterplots for the numeric variables mentioned and boxplots for categorical variables, like room type and city. These visualizations helped guide later decisions on which variables to include in the baseline model. I then identified the missing values in the dataset and handled them by removing rows with missing prices and replacing missing values in ‘Reviews per Month’ with 0. I used boxplots and the 1.5xIQR rule for key numeric variables to detect outliers. I removed those observations that exceeded these

determined thresholds (by 1.5xIQR) to reduce the influence of extreme values. For preprocessing, I converted relevant categorical variables, like room type, city, and neighborhood, into factors to ensure they would be correctly interpreted in the regression model. Finally, I removed non-informative columns, like ID, name, host ID, and host name, to prepare the dataset for modeling.

I then continued to verify the assumptions of linear regression by checking for linearity, normality of residuals, homoscedasticity, independence of observations, and multicollinearity. I plotted the residuals against fitted values to check for linearity to ensure no clear pattern indicated non-linearity. The normality of residuals was assessed using a histogram and QQ plot, as an assumption for linear regression is that residuals are normally distributed. I then examined homoskedasticity through the same residuals vs fitted values plot used prior, where a consistent spread of residuals would satisfy this assumption. I also checked for independence of observations, making sure each Airbnb listing was treated as an independent unit. Lastly, multicollinearity was evaluated using the generalized variance inflation factor (GVIF), as the model includes categorical predictors on multiple levels.

After assessing the assumptions, I handled assumption violations, given that not all assumptions were met. Linearity, homoscedasticity, and normality of residuals appeared violated based on the plots. These violations were likely due to skewed variables. So, I applied appropriate transformations to improve model validity. I used log transformations to reduce skewness and stabilize variance for heavily right-skewed variables, including price, minimum nights, and host listings, assessed by histograms made in prior sections. For those variables with moderate skew, like number of reviews, reviews per month, and year-round availability, I applied a square root transformation. Then, I compared the models before and after corrections by re-running the diagnostic plots and comparing them.

After transformations, I continued onto variable selection and hypothesis testing. I implemented two variable selection techniques based on AIC to identify the most relevant

predictors: backward selection and forward selection. Backward selection starts with a full model and removes predictors that are not statistically significant or do not contribute meaningfully to the model. On the other hand, forward selection begins with a model that only includes the intercept and adds predictors one at a time based on improvement of the AIC. Both of these methods help avoid overfitting while still retaining explanatory power. After selecting the model, I conducted hypothesis tests on each coefficient by assessing the summary output of the linear model. I evaluated the p-values to determine the statistical significance of each variable at a 0.05 significance level. This step helped determine which predictors have a meaningful linear association with the log-transformed price after adjusting for other variables. Then, I assessed the model's performance by evaluating the reported R^2 and adjusted R^2, which quantified how much variation in the response was explained by the model. I also calculated the Root Mean Squared Error (RMSE), a metric used to measure the prediction error of a model. Finally, I performed a 10-fold cross-validation to validate the model. I partitioned the data into 10 subsets, then trained the model on 9 subsets and evaluated it on the last one. This process was repeated for all folds. I then calculated the RMSE for each fold and the mean RMSE. This cross-validation was used to check how well the model would perform on new data.

Finally, I interpreted each predictor's impact on the response. For each significant predictor, I assessed the size and direction of its effect on the log-transformed price. I also reported these coefficients' 95% confidence intervals, which help assess the statistical and practical significance. Finally, I explained the results in the dataset context to provide real-world meaning to the statistical findings.

Results

To begin the analysis, I reviewed the summary statistics for the Airbnb dataset, which included over 230,000 listings and 18 variables. The summary statistics revealed that the response variable, price, had a median of $149 and a maximum value exceeding $100,000,

suggesting a right-skewed distribution and the presence of extreme outliers. Similar skewed distributions were observed in predictors like minimum nights, number of reviews, reviews per month, host listings, and year-round availability. The visualization of relationships, plotted price vs variable, showed mostly weak, nonlinear relationships between price and the numeric predictors. At the same time, boxplots highlighted apparent price differences across room types and cities, where whole houses had higher prices than shared or private rooms, and listings in Santa Cruz, San Francisco, and others trended in higher prices. One variable, reviews per month, contained many missing values, which aligned with listings without prior review; these were converted into zero. Outliers were identified using standard boxplot thresholds and removed to improve model robustness.

Diagnostic checks were performed to evaluate if the linear regression met the assumptions: linearity, normality of residuals, homoscedasticity, independence of observations, and multicollinearity. The linearity assumption was assessed by plotting residuals vs fitted values. The resulting plot showed a funnel-shaped pattern, suggesting a violation of linearity. Homoscedasticity was also evaluated through this plot. The spread of residuals appeared to increase with fitted values (funnel), suggesting heteroscedasticity. The normality of residuals was checked using a histogram and a QQ plot. The histogram displayed a right-skewed distribution, and the QQ plot revealed deviation from the diagonal line on both ends, suggesting that residuals did not follow a normal distribution. The assumption of independence of observations was satisfied, given that each observation in the dataset was a different Airbnb listing. Lastly, multicollinearity was evaluated using the generalized variance inflation factor (GVIF), to allow for proper VIF comparison, the adjusted metric was also reported ($GVIF^{1/(2*df)}$). All adjusted values were below 2, indicating no concern about multicollinearity.

To address the violations of the regression assumptions, I transformed the predictor variables and refitted the model. The initial diagnostic plots showed violations in linearity, normality of residuals, and homoscedasticity. I applied log transformations on price, minimum

nights, and host listings, and applied square root transformations on number of reviews, reviews per month, and year-round availability. After refitting the model with the transformed variables, the residual plots showed more randomness and improved symmetry, improving linearity and homoscedasticity of the model. The QQ plot showed a reduced deviation from the diagonal line, suggesting improved normality of residuals. After comparing the models before and after corrections, the AIC significantly decreases from 1,417,347 for the original model to 174,931 for the transformed model, indicating that the transformations led to a better model fit.

I implemented backward and forward selection using AIC as the criterion to identify the most relevant predictors for Airbnb price. Both models converged on the same final model, which included the log of minimum nights (log_min_nights), the square root of reviews (sqrt_reviews), the square root of reviews per month (sqrt_reviews_month), the square root of year-round availability (sqrt_availability), the log of host listings (log_listings), room type, and city as significant predictors. After selecting the model, I conducted hypothesis tests on the coefficients. Nearly all the predictors were statistically significant at the 0.001 level, indicating strong evidence of their association with the response variable of log_price, in particular, log_min_nights, room types, and cities. Model performance was then assessed using R^2, adjusted R^2, and RMSE. The R^2 value was 0.3872, and the adjusted R^2 was 0.387, indicating that the model explained approximately 38.7% of the variation in log_price. The RMSE was 0.4987. To evaluate generalizability, I performed a 10-fold cross-validation. The average RMSE across the folds was 0.4988, closely aligning with the in-sample RMSE.

Finally, to evaluate the influence of each predictor on the response variable, I examined the coefficients from the final regression model. Variables like log_min_nights, cities, and room type showed notable effects on price. Then, I calculated the 95% confidence interval for each coefficient. These intervals provide a range within which the true effect is likely to lie, allowing me to assess whether each variable positively or negatively impacts the log-transformed price. For example, the interval for log_min_nights ranged from approximately -0.134 to -0.127,

suggesting a negative association with price. Conversely, the city of Santa Cruz County had a confidence interval of approximately 0.386 to 0.476, indicating a strong positive influence on log_price.

Discussion

This analysis sought to understand which factors influence Airbnb pricing the most. A linear regression model was developed and the final model included the log of minimum nights (log_min_nights), the square root of reviews (sqrt_reviews), the square root of reviews per month (sqrt_reviews_month), the square root of year-round availability (sqrt_availability), the log of host listings (log_listings), room type, and city as significant predictors.

Most predictors were found to influence the price significantly ($p<0.05$). Based on the final model, the log of minimum nights (log_min_nights) has a negative coefficient of approximately -0.130. This indicates that for each 1% increase in minimum nights, the expected price decreases by about 0.13%, holding other variables constant, suggesting that listings with more extended minimum stays are priced lower. In addition to coefficient estimates, 95% confidence intervals were calculated to assess the precision of these estimates. The interval for log_min_nights remained entirely below zero ([-0.134, -0.127]), confirming a consistent negative relationship with price. CI for all variables was evaluated, and the coefficient estimate was confirmed. Similarly, the log of host listings and the square root of reviews per month also had a negative coefficient (β= -0.020 and –0.143, respectively). Where a 1% increase in the number of listings by the same host leads to an estimated 0.02% decrease in price, and a one-unit increase in the square root of reviews per month is associated with an estimated 14.3% decrease in price. This would suggest that hosts managing many properties may price their listings competitively. Additionally, more monthly reviews may indicate frequently booked lower-priced listings. On the other hand, the square root of total reviews and the square root of year-round availability had a positive estimated coefficient (β= 0.004 and 0.009, respectively).

Where a one-unit increase in the square root of the total number of reviews is associated with an estimated 0.42% increase in price, this suggests that more popular listings with more reviews are priced slightly higher. Likewise, a one-unit increase in the square root of annual availability is associated with an estimated 0.9% increase in price, where listings with more open availability are less frequently booked and may set higher prices to compensate for lower occupancy.

Room type was a very influential predictor in this model. As expected, all other room types, hotel, private, and shared rooms, were significantly less expensive than entire homes (reference level), with strong negative coefficients of -0.528, -0.796, and -1.246, respectively. To interpret these coefficients within the context of the log-linear model, I translated the coefficients into percent changes (percent change = 100 x (e^{β} - 1)). Based on this, hotel rooms were estimated to be about 41% less expensive, private rooms were about 55% less expensive, and shared rooms were about 71% less expensive than entire homes. These price differences were expected, as entire homes offer more space and privacy, so travellers are willing to pay more for this comfort. In contrast, shared or smaller accommodations are more budget-friendly options.

Finally, the city where the listing is located was also a key predictor in this model. The reference level used for this variable was the city of Asheville. Cities like San Francisco (β=0.334), Pacific Grove (β= 0.723), Cambridge (β= 0.492), and Boston (β= 0.271) all had positive coefficients. When translated into percent changes, the listings in these cities were estimated to be approximately 40%, 106%, 63%, and 31% more expensive than listings in Asheville, respectively. Other cities with notable estimated increases include San Diego (29%), Broward County (27%), and New Orleans (24%). In contrast, cities like Portland (β= -0.194), Salem (β= -0.120), and Columbus (β= -0.154) were estimated to be significantly less expensive than Asheville, with prices approximately 18%, 11%, and 14% lower, respectively. Other cities estimated to have cheaper listings than Asheville include Chicago (β= -0.077) and Twin Cities

(β= -0.057). Cities with no significant differences from Asheville listing prices included Denver and Jersey (p-values were not significant).

The final linear regression model demonstrated a reasonable fit to the data. Its R^2 value was 0.3872, and its adjusted R^2 was 0.387. These values indicate that approximately 38.7% of the variability in the log price is explained by the predictors included in the model. The F-statistic of 2246 and p-value of less than $2.2e^{-16}$ indicate that the overall model is highly statistically significant. The RMSE was 0.4987, which reflects the typical prediction error in log price units. After performing a 10-fold cross-validation, this yielded an average RMSE of 0.4988. This value is closely aligned with the in-sample RMSE. This suggests that the model generalizes well to new data without overfitting.

While this model does provide insightful information on factors that influence Airbnb pricing, several limitations should be considered. First, this model only explains around 38.7% of the variation in log price. This suggests that the analysis may be missing influential factors, such as amenities, seasonal demand, location desirability within the city, and more. However, this explanatory power level is not unusual for a real-world data set. Second, although the transformations performed improved the regression assumptions, some assumptions were still not fully satisfied post-transformation. There was still some mild non-normality of residuals and heteroskedasticity, which could have affected the results. However, since the dataset had a large sample size, these violations are less likely to impact the overall conclusions. Finally, this analysis assumes linear relationships and includes no interaction effects, which may exist in the data but were not explored.

Conclusion

In this project, I explored the factors influencing Airbnb listings using the “U.S. Airbnb Open Data” dataset on Kaggle. After thorough data cleaning and assumption checks, I fit a multiple linear regression model with several predictors to see how they affected my response

variable, log_price. Key factors like room type, city, minimum nights, and total reviews on the listing were found to have statistically and practically significant effects on price. The model explained 38.7% of the variability in log_price, and through cross-validation techniques, its performance was stable across different data subsets. Future work could incorporate additional predictors like amenities, proximity to attractions, or seasonality, to account for more variability in price. I could also explore non-linear relationships and interaction effects to capture more complex relationships within this dataset.

References

(1) Seth, K. (2023, April 14). *U.S. Airbnb Open Data*. Kaggle. https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data/data

(2) Bettache, N. (2025). *Lecture notes from BTRY 6020 Statistical Methods II.* Department of Statistics and Data Science, Cornell University.