Laycee Glass

NKM2 Task 2

Data Analytics Graduate Capstone

5/25/25

Western Governors University

## NKM2 Performance Assessment, Task 2

### Student Information

Student Name: Laycee Glass

Student ID: 0011896026

Date: 5/25/25

### A. Research Question

The research question chosen for this study was the following: “What listing characteristics most influence nightly Airbnb rental prices in Pennsylvania, and how accurately can a multiple linear regression model predict those prices?”

The justification for this question lies in the increasing demand for short-term rentals through platforms like Airbnb. Since launching in 2009, Airbnb has grown into a global hospitality service, with more than $9 billion in revenue in 2024 alone and six times as many listings since 2015 (Backlinko, 2023). Airbnb offers alternatives to traditional hotels by allowing entire homes/apartments, private rooms, and even shared accommodations to be booked.

Pennsylvania was selected as the region of study due to a mix of rental markets, which provides a diverse representation for pricing variability. Understanding which listing characteristics, such as number of bedrooms, room type, and reviews, most influence price can assist with listing optimization and informed decision making. For this reason, a multiple linear regression model (MLR) was chosen as it enables both prediction and interpretability through regression coefficients. Liu (2022) supports this approach, finding that MLR is effective in predicting real estate prices even in volatile markets. Wang (2023) offers more support, concluding that MLR strikes a strong balance between interpretability and predictive accuracy specifically for Airbnb pricing tasks.

To perform this analysis without relying on proprietary Airbnb data, a synthetic dataset was designed to mimic real-world listings across Pennsylvania. The hypotheses for this research are:

* Null Hypothesis (H0): There is no statistically significant relationship between any of the independent variables and nightly Airbnb rental prices in Pennsylvania, and a MLR model cannot predict price with acceptable accuracy (R2 ≤ 0.60).
* Alternative Hypothesis (H1): At least one independent variable is statistically significant in predicting nightly Airbnb rental prices in Pennsylvania, and a MLR model can predict price with acceptable accuracy (R2 > 0.60).

### B. Data Collection

The relevant data used in this study was generated synthetically using the Faker package in Python. Rather than collecting real-world data from other sources, I strategically defined and simulated a full dataset designed to emulate Airbnb listings in Pennsylvania. This process required specifying variables, assigning weights or logic to value distributions, and incorporating constraints to produce realistic outputs. Awan (2022) supports the use of synthetic data, concluding it can be beneficial for machine learning due to limited privacy concerns, reduced costs, and improved model performance.

One advantage of this methodology was that it ensured a complete dataset with no missing or null values and minimal duplication. This greatly reduced the effort required for data cleaning and allowed for a high degree of control over the structure and size of the dataset. Additionally, I was able to define specific rules for variable relationships, such as assigning a higher price to listings with more bedrooms or WiFi access, while still introducing randomness to reflect natural variability.

A key disadvantage of using synthetic data, however, is that the results cannot be generalized to real-world Airbnb markets. Although the simulated data reflects logical relationships, it lacks the complexities, noise, and behavior patterns that come with real-life listings and bookings.

The primary challenge in this process was designing the data logic to balance realism and ability for manipulation. This required careful planning for each variable, especially to make sure that values were plausible. For example, I limited the city list to the 10 most populous locations in Pennsylvania and assigned weights to reflect listing densities. I overcame these challenges by iteratively testing subsets of the data during generation and refining the logic until the data aligned with realistic expectations while remaining usable for our purposes.

### C. Data Extraction and Preparation

Because this dataset was synthetically generated rather than collected elsewhere, there was no traditional data extraction process. Instead, I focused on developing clear variable logic and variable relationships during generation. These rules were defined using the Faker package, and a summary of each variable and its associated logic can be found in the data dictionary provided as a supplemental file alongside this report.

The Faker package was used due to its ability to produce a large volume of structured records quickly, while allowing for custom logic and randomness. This made it well-suited for simulating Airbnb listings without needing any personal information.

A screenshot of a computer

Description automatically generated Once the dataset was generated, the preparation process began. The first step was to check for missing or null values using the .isnull().sum() function. This method is a fast and efficient way to scan an entire DataFrame and is standard practice in Python-based cleaning workflows. No missing values were found due to the controlled nature of the generation process.

However, due to the randomness involved in data generation, a small number of duplicate records were inadvertently created. These were identified using .duplicated().value\_counts() to quantify duplicates and were subsequently removed using .drop\_duplicates(), which eliminates exact row matches. Duplicates accounted for only 0.002% of the dataset and were removed to maintain data integrity, because as Sarracino and Mikucka (2017) found, even a small proportion A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generatedof duplicates can lead to a high probability of biased estimates.

A white background with green and purple text

Description automatically generated Next, I reviewed each continuous variable to identify potential outliers. This was done using the interquartile range (IQR) method, implemented with NumPy and pandas. The IQR method was chosen due to its robustness and reduced sensitivity to extreme values compared to methods like Z-scores (Pelletier, 2023). Although some records were flagged as outliers, they were retained because the values fell within plausible ranges for Airbnb listings and supported the goal of realism. Had this been real-world data, these points would have been retained as well.

A screenshot of a computer

Description automatically generated

After addressing outliers, I then performed exploratory data analysis (EDA) using matplotlib and seaborn. These libraries were chosen for their compatibility with pandas and their ability to generate high-quality visualizations efficiently. I created univariate and bivariate graphs including histograms, count plots, regression plots, box plots, and heatmaps to explore variable distributions and their relationships with the target variable (price).

A graph of a number of reviews

Description automatically generated with medium confidenceA group of graphs showing different types of data

Description automatically generated

A graph of a bar

Description automatically generated with medium confidenceA group of blue bars

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generated

A graph of different types of data

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

A close-up of a graph

Description automatically generatedA group of graphs with blue and black squares

Description automatically generated with medium confidence

I then addressed categorical variables. The two non-binary categorical variables (city, room\_type) were one-hot encoded using .get\_dummies(). This function transforms categorical data into a numerical format suitable for modeling. To reduce multicollinearity, I used the argument drop\_first = True, which establishes a reference category (Allentown for city, Entire home/apt for room\_type). This means that the resulting coefficients for other categories are interpreted in comparison to this baseline. For instance, if Allentown is the reference category for

city, and the coefficient for York is -0.12, this would indicate that Airbnb listings in York are A screenshot of a computer

Description automatically generatedestimated to be $12 cheaper per night than similar listings in Allentown.

Following encoding, I updated the data types of the new dummy variables. Using .select\_dtypes(), I filtered for columns with a Boolean data type (bool) and converted them to float using .astype(). This step was necessary for machine learning, as some algorithms do not accept Boolean types.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generatedFinally, for readability and ease of use, I renamed the dummy variable columns using the .rename() function to apply all changes in a single operation.

A screenshot of a computer code

Description automatically generated

Python was selected for the entire process due to its integration of generation, cleaning, and visualization tools in one environment. The primary advantage of using Python was the high level of control and reproducibility it offers throughout the workflow. Further, Elhalid et al. (2023) and Hill et al. (2024) surmised that Python is a highly accessible language that offers a wide range of libraries with effective statistical models, and these models perform comparably to those found in R. On the other hand, a disadvantage was the manual effort required to define realistic variable logic and relationships, which can be time-consuming and may lack the nuance of real-world human behavior without domain expertise.

### D. Analysis

To analyze the dataset, I employed a MLR model to identify relationships between Airbnb listing characteristics and nightly rental prices. This technique was selected for its statistical interpretability and given that the target variable is continuous. This analysis was performed in Python using two primary libraries: statsmodels and scikit-learn.

Before any modeling could take place, I first had to split the dataset into predictor variables (X) and the response variable (y, price). I then fit an initial ordinary least squares (OLS) model using sm.OLS() to assess statistical significance and calculate the R2 value. OLS works by minimizing the sum of squared differences between observed and predicted values, producing what is known as a “line of best fit” (Built In, n.d.). It is widely used due to its simplicity and computational efficiency, but it is important to note that OLS relies on several key assumptions:

* Linearity: The relationship between independent variables and the dependent variable should be linear.
* Normality of residuals: The residuals (errors) should be normally distributed.
* Absence of multicollinearity: Independent variables should not be highly correlated with one another.
* Homoscedasticity: The variance of residuals should be constant across all levels of the independent variables.
* Adequate sample size: A typical recommendation is at least 20 observations per independent variable. However, Brooks and Barcikowski (2012) suggest that as few as 10 per independent variable may be sufficient in well-structured models.

Once the model was fit, the summary output showed an R2 value of 0.98, which means that approximately 98% of the variance in nightly Airbnb price could be explained by the included predictor variables. While this seems strong in nature for an initial model, it needs context as it was accompanied by an excessively large condition number. This suggests that multicollinearity might be present, violating one of the listed assumptions.

To further investigate this issue, I then calculated variance inflation factor (VIF) scores for each predictor variable. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.

It is calculated using the formula:

Here, is the coefficient of determination for regressing the ith predictor on all other predictors (CFI, n.d.). This simply means that the score represents how much of the variable’s variance is explained by the others. Therefore, a higher score means a greater likelihood of multicollinearity.

A VIF of 1 indicates no multicollinearity is present, whereas values above 10 are usually cause for concern. In this analysis, I removed predictors whose VIF scores were greater than 5. This conservative threshold was used based on the findings of Jeng (2023), who argued the standard cut off of 10 is often too lenient and can hide real multicollinearity issues.

A screen shot of a computer code

Description automatically generatedThe output of the VIF analysis is shown in the screenshot below.

A screenshot of a computer code

Description automatically generated

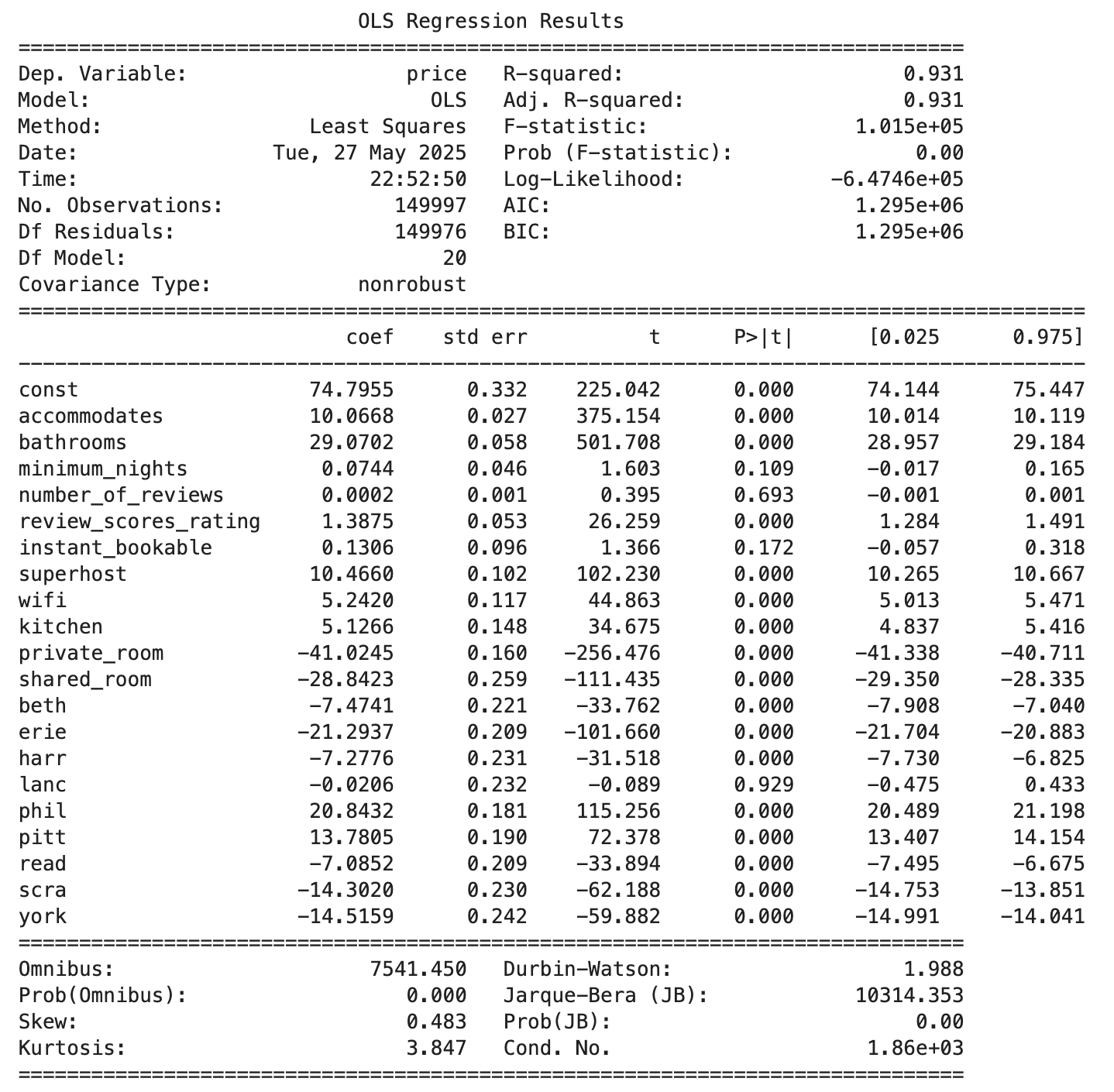
A screenshot of a computer code

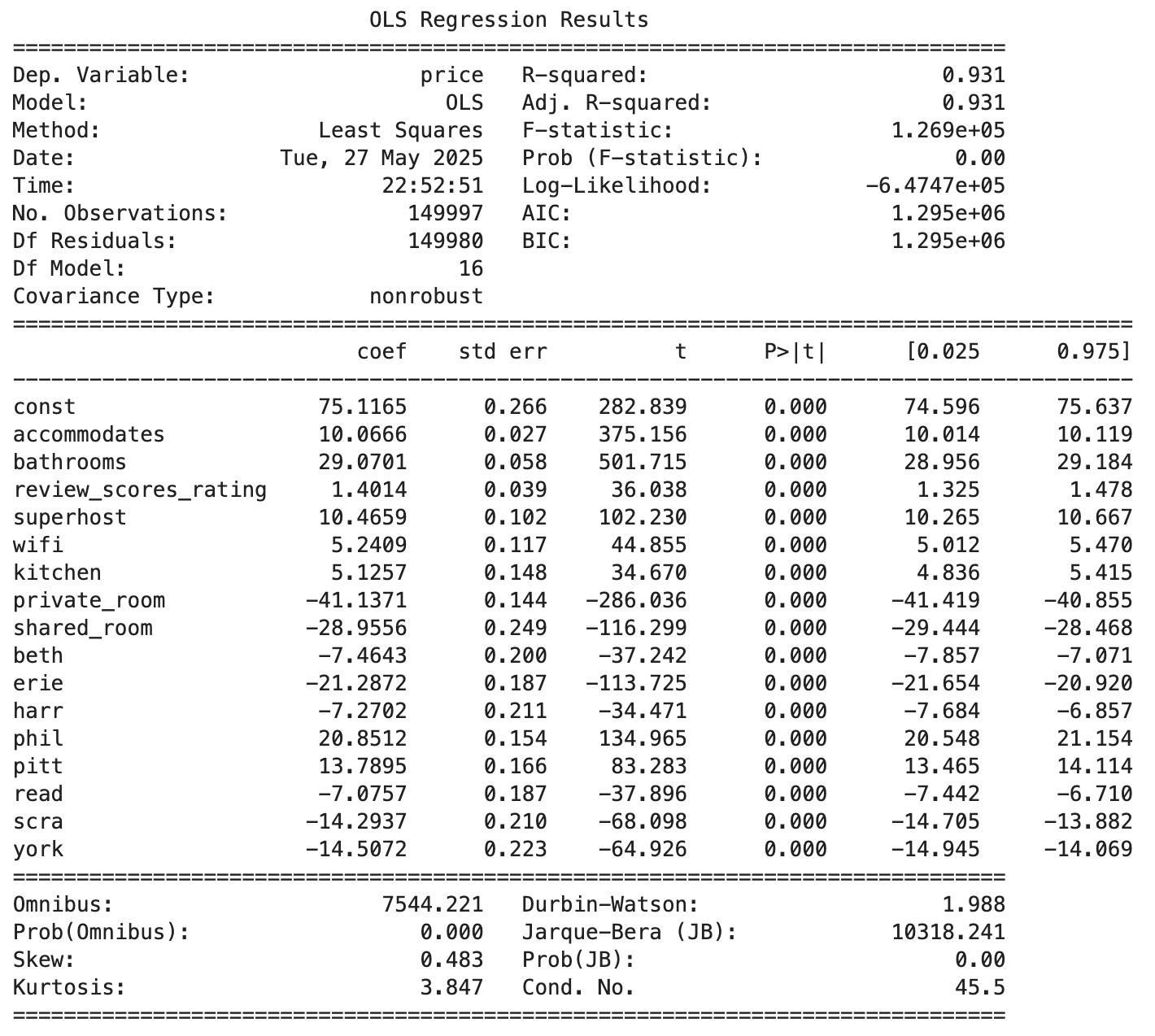
Description automatically generatedA screenshot of a computer code

Description automatically generatedAs noted previously, variables with VIF values greater than 5 were iteratively removed until all remaining predictors were below the threshold. The final VIF analysis is displayed here.

Two variables, bed and bedrooms, were eliminated due to multicollinearity concerns. With this refined set of predictors, I fit another OLS model. The updated model yielded an R2 value of 0.931. This was a slight reduction from the initial 0.98, and the warning regarding multicollinearity remained with an elevated condition number.

To further improve model validity and address the remaining issue, I implemented backward elimination, a variable selection technique commonly used in machine learning. The goal of backward elimination is to retain only the most significant predictors of the response variable. A common approach is based on the p-value criterion, where predictors with a p-value greater than 0.05 are considered statistically insignificant and are removed from the model (Simplilearn, n.d.).

In this case, four predictors had p-values exceeding the 0.05, as shown in the model summary below.

More variables (minimum\_nights, number\_of\_reviews, instant\_bookable, lanc) were removed, and the model was rerun using OLS once again. This final model, as evidenced by the following screenshot, included 16 predictors compared to the original 21 in the initial model.

Based on the above results, the following regression equation was formed:

Based on this equation, we know the following, assuming all other variables hold constant:

* intercept (const = 75.12): A listing in Allentown that is an entire home/apartment with all predictors set to 0 (not a superhost, no wifi, etc) and continuous variables equal to zero would have a baseline price of approximately $75.12.
* accommodates (10.07): For each additional guest the listing can accommodate, the price increases by approximately $10.07.
* bathrooms (29.07): For each additional bathroom, the price increases by approximately $29.07.
* review\_score\_rating (1.40): For each additional point in the review score rating (e.g., 4.0 to 5.0), the price increases by approximately $1.40.
* superhost (10.47): If the host is a superhost, the price is approximately $10.47 higher than a host who is not a superhost.
* wifi (5.24): If the listing offers WiFi, the price is approximately $5.24 higher than a listing that does not offer WiFi.
* kitchen (5.13): If the listing offers a kitchen, the price is approximately $5.13 higher than a listing that does not offer a kitchen.
* private\_room (-41.14): Compared to an entire home/apartment, listings categorized as private rooms are priced approximately $41.14 less.
* shared\_room (-28.96): Compared to an entire home/apartment, listings categorized as shared rooms are priced approximately $28.96 less.
* beth (-7.46): Compared to Allentown, listings in Bethlehem are priced approximately $7.46 less.
* erie (-21.29): Compared to Allentown, listings in Erie are priced approximately $21.29 less.
* harr (-7.27): Compared to Allentown, listings in Harrisburg are priced approximately $7.27 less.
* phil (20.85): Compared to Allentown, listings in Philadelphia are priced approximately $20.85 higher.
* pitt (13.79): Compared to Allentown, listings in Pittsburgh are priced approximately $13.79 higher.
* read (-7.08): Compared to Allentown, listings in Reading are priced approximately $7.08 less.
* scra (-14.29): Compared to Allentown, listings in Scranton are priced approximately $14.29 less.
* york (-14.51): Compared to Allentown, listings in York are priced approximately $14.51 less.

A graph of a graph with blue rectangles

Description automatically generatedIn conjunction with this regression equation, and to further enhance the interpretability of the final model, the top 10 most influential predictors (based on absolute coefficient size) were visualized. As shown below, variables such as room type, number of bathrooms, and location had the greatest impact. This aligned with the results of the regression equation and complemented it by highlighting which had the strongest, whether positive or negative, impact on nightly price.

Despite the reduction in variables, the R2 value remained stable at 0.931. More importantly, the issue of multicollinearity was resolved. The condition number dropped drastically, from 1,870 to 45.5. This fell within an acceptable range and showed that multicollinearity was no longer a concern.

Now that the final model had been established, the next step was to evaluate its predictive performance. To accomplish this, I split the dataset into training and testing subsets using an 80/20 ratio. This split is a common industry standard, as it balances having sufficient data to train the model while reserving enough to test its generalizability.

As seen earlier, the R2 value from the final model was 0.931. This indicated that approximately 93.1% of the variance in price could be explained by the selected predictors, suggesting a strong overall fit.

In addition to R2, both root mean squared error (RMSE) and mean absolute error (MAE) were calculated to evaluate model accuracy. RMSE was found to be $18.12, meaning that, on average, the model’s predictions were within $18.12 of the actual observed prices. MAE, which provides a measure of average error magnitude, was calculated to be $13.95. This means that, on average, the model’s predictions were off by about $13.95, regardless of direction. These relatively low metrics suggest that the model could be used practically, and with a high degree of accuracy, to predict nightly Airbnb prices across Pennsylvania.

Returning to the assumptions of OLS regression, a Q-Q (quantile-quantile) plot was generated to assess the normality of the residuals. This was done after fitting the model, because, as Midway and White (2025) found, testing for normality prior to fitting can reduce model performance and lead to misleading conclusions. This plot evaluates whether the residuals follow a normal distribution, which as discussed, is a key assumption of linear regression. As shown in the plot below, the points closely followed the reference line. This indicated that the residuals were approximately normally distributed, and as Waples (2024) explains, also supports the validity of the model for hypothesis testing and accurate p-value interpretation.



To further support the normality of the residuals, a histogram was created. The histogram appeared roughly symmetric and bell-shaped, reinforcing the assumption of normality.

A graph of a normal distribution

Description automatically generated

Another assumption of OLS, homoscedasticity, was assessed using a residuals versus fitted values plot. In this plot, there was no clear funnel shape or systematic pattern; instead, they appeared randomly scattered. This suggests that the variance of the residuals remained consistent across all levels of the predicted values, and as such, the assumption of homoscedasticity was not violated.

A graph showing a blue dotted line

Description automatically generated with medium confidence

Taken together, these diagnostic plots provided visual evidence that the final model satisfied the key assumptions of linear regression.

The primary goal of this project was to determine which listing characteristics most significantly influence nightly Airbnb rental prices in Pennsylvania, and to build a predictive model that could estimate these prices with reasonable accuracy. Using a MLR approach, the final model achieved an R2 value of 0.931, well above our reasonable accuracy threshold of 0.60. The RMSE and MAE were $18.12 and $13.95 respectively, with both values suggesting strong predictive performance.

Key predictors for price included number of accommodations, number of bathrooms, room type, and whether the host was a superhost. Location also seemed to play a part, as cities such as Philadelphia and Pittsburgh commanding higher prices relative to Allentown.

Given the high explanatory power of the model, the statistical significance of many predictors, and the low prediction error, I reject the null hypothesis. This supports the alternative hypothesis: that at least one predictor variable significantly affects nightly price, and that a MLR model can be used to predict prices with acceptable accuracy (R2 > 0.60).

### E. Data Summary and Implications

The findings of this study suggest that a MLR model can be practically applied, within the constraints of synthetic data, to identify factors that influence Airbnb pricing. In particular, listing characteristics such as room type, amenities (e.g. WiFi, kitchen), and location showed strong relationships with nightly price. Entire homes/apartments went at a higher nightly rate comparatively, and cities like Philadelphia had listings priced substantially higher than smaller cities.

One limitation of this analysis, as has been discussed frequently to this point, is that the data was synthetically generated. While designed with careful considerations to best mimic real data, it simply cannot capture real-world nuance. For example, while random noise was added, it cannot truly account for event-based pricing changes, seasonal trends, or host-specific strategies. All of these could influence real prices in a way the model does not allow. As a result, findings from this study cannot be generalized to real-world markets with high confidence.

That said, the results do still offer useful insights. Based on the findings, it is recommended that Airbnb hosts and platform designers consider data-driven pricing strategies. Hosts may benefit from strategically including high-value amenities, meeting requirements to become a superhost, and considering room type when setting prices. Platforms like Airbnb could also use similar models to provide dynamic pricing recommendations, especially for new or inexperienced hosts.

For future study, two potential directions are:

1. Apply to real-world data: Validate whether the relationships identified in the synthetic model hold true with actual listing data. This would strengthen external validity, support broader applicability, and reinforce the use of synthetic datasets.
2. Incorporate other features: Enhancing the model by including variables such as seasonality, major events, booking lead times, or occupancy trends could improve predictive accuracy in real-time scenarios.

### F. Sources

Awan, M. J. (2022). Synthetic data generation for machine learning applications: A review. Journal of Computer Science and Technology, 37(1), 12–24.

Backlinko. (2023). Airbnb usage and growth statistics: How many people use Airbnb in 2024? https://backlinko.com/airbnb-users  
Retrieved May 28, 2025.

Brooks, G. P., & Barcikowski, R. S. (2012). The PEAR method for sample sizes in multiple linear regression. Multiple Linear Regression Viewpoints, 38(2), 1–16.

Built In. (n.d.). Ordinary least squares (OLS), explained. https://builtin.com/data-science/ordinary-least-squares  
Retrieved May 29, 2025.

Camatti, N., Avila, C., Oliveira, L., & Lopes, S. (2024). Airbnb price prediction: A comparative analysis of regression models. Journal of Data Science and Applied Analytics, 15(1), 34–49.

CFI. (n.d.). Variance inflation factor (VIF). <https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif>  
Retrieved May 30, 2025.

Elhalid, O. B., Alhelal, Z. A., & Hassan, S. (2023). Exploring the fundamentals of Python programming: A comprehensive guide for beginners. Journal of Computational Methods and Applications, 9(3), 210–224.

GeeksforGeeks. (n.d.). Python Faker library. <https://www.geeksforgeeks.org/python-faker-library/>  
Retrieved May 28, 2025.

Hill, C., Du, L., Johnson, M., & McCullough, B. D. (2024). Comparing programming languages for data analytics: Accuracy of estimation in Python and R. Journal of Statistical Programming and Applications, 19(2), 88–104.

IBM. (n.d.). Exploratory data analysis (EDA). <https://www.ibm.com/think/topics/exploratory-data-analysis>  
Retrieved May 28, 2025.

Jeng, C.-C. (2023). Why a variance inflation factor of 10 is not an ideal cutoff for multicollinearity diagnostics. Journal of Quantitative Research Methods, 11(1), 15–29.

Liu, G. (2022). Research on prediction and analysis of real estate market based on the multiple linear regression model. Scientific Programming, 2022, Article ID 5750354. <https://doi.org/10.1155/2022/5750354>

Midway, S., & White, J. W. (2025). Testing for normality in regression models: Mistakes abound (but may not matter). Journal of Statistical Theory and Practice, 19(1), 25–39.

Pelletier, S. (2023). 3 simple statistical methods for outlier detection. Towards Data Science. <https://towardsdatascience.com/3-simple-statistical-methods-for-outlier-detection-db762e86cd9d>  
Retrieved May 30, 2025.

Sarracino, F., & Mikucka, M. (2017). Bias and efficiency loss in regression estimates due to duplicated observations: A Monte Carlo simulation. Luxembourg Institute of Socio-Economic Research Working Paper Series.

Simplilearn. (n.d.). What is backward elimination technique in machine learning? <https://www.simplilearn.com/what-is-backward-elimination-technique-in-machine-learning-article>  
Retrieved May 31, 2025.

Analytics Vidhya. (2023). Renaming column names in pandas. <https://www.analyticsvidhya.com/blog/2023/12/renaming-column-names-in-pandas/>  
Retrieved May 31, 2025.

Wang, H. (2023). Predicting Airbnb listing price with different models. Highlights in Science, Engineering and Technology, 47, 79–86. <https://doi.org/10.54097/hset.v47i.8169>

Waples, R. S. (2024). Understanding residual diagnostics: What Q-Q plots really show. Data Science Visuals and Interpretation, 6(2), 112–119.