**Data Analytics Capstone Topic Approval Form**

**Student Name:**  Laycee Glass

**Student ID:**  11896026

**Capstone Project Name:** Multiple Linear Regression Analysis of Airbnb Pricing

**Project Topic**: Regression Modeling of Airbnb Rentals in Pennsylvania

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** What listing characteristics most influence nightly Airbnb rental prices in Pennsylvania, and how accurately can a multiple linear regression model predict those prices?

**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).

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).

**Context:** This study will contribute to the field of data analytics and the MSDA program by developing a statistically grounded, reusable predictive model to estimate nightly Airbnb rental prices in Pennsylvania. The model will also identify which listing characteristics most strongly influence pricing outcomes. These insights are valuable for understanding rental pricing dynamics and guiding strategic decisions related to listing optimization, competitive positioning, and market segmentation.   
The study will employ a Generalized Linear Model (GLM), specifically Multiple Linear Regression (MLR), to examine the relationship between a set of independent variables (e.g., number of bedrooms, review score, location type) and a continuous dependent variable (nightly price). MLR enables the quantification of each variable's impact through regression coefficients and generates a predictive equation used to estimate price based on input features. Prior research supports this method: Liu (2022) demonstrated that MLR is effective in predicting real estate prices even in volatile housing markets, while Wang (2023) concluded that MLR provides a strong balance between interpretability and predictive accuracy in Airbnb pricing tasks.

**Data:** The dataset used in this study was generated using the Faker package in Python. According to Awan (2022), synthetic data is widely used in machine learning due to its ability to preserve privacy, improve model performance, and reduce operational costs. This approach was used to simulate a realistic sample of Airbnb listings in Pennsylvania for predictive modeling purposes without relying on proprietary or personally identifiable information. The dataset contains approximately **150,000 records** and **13 variables**, which exceeds the recommended minimum for regression modeling. Brooks and Barcikowski (2012) recommend a minimum of 10 observations per predictor variable, and this dataset provides a strong statistical foundation for building and evaluating an MLR model.

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Role** |
| city | Categorical | Independent Variable |
| room\_type | Categorical | Independent Variable |
| accommodates | Numeric (Discrete) | Independent Variable |
| bedrooms | Numeric (Discrete) | Independent Variable |
| bathrooms | Numeric (Continuous) | Independent Variable |
| beds | Numeric (Discrete) | Independent Variable |
| minimum\_nights | Numeric (Discrete) | Independent Variable |
| review\_score\_rating | Numeric (Continuous) | Independent Variable |
| instant\_bookable | Categorical (Binary) | Independent Variable |
| superhost | Categorical (Binary) | Independent Variable |
| wifi | Categorical (Binary) | Independent Variable |
| kitchen | Categorical (Binary) | Independent Variable |
| price | Numeric (Continuous) | Dependent Variable |

A full data dictionary describing generation logic and value rules is available upon request. As the data is entirely synthetic, there is no risk of exposing sensitive or identifying information.  
Limitations: The primary limitation is that the dataset is artificially generated and does not reflect actual Airbnb behavior or bookings. Although variables were generated with realistic distributions and interdependencies, the insights may not fully generalize to real-world data.  
Delimitations: Duplicate listings that may arise from random generation will be removed to maintain data integrity. Because such duplicates are minimal in number, they are not expected to meaningfully impact results. The dataset retains more than enough observations for robust model fitting and evaluation.

**Data Gathering:** The Treatment of the Data: The dataset was generated using the Faker package in Python, simulating realistic Airbnb listings in Pennsylvania. During data generation, some duplicate records are inadvertently created. These duplicates will be identified and removed to maintain data integrity. Retaining duplicates can significantly bias regression estimates and reduce their efficiency. Sarracino and Mikucka (2017) found that even a small proportion of duplicates can lead to a high probability of biased estimates. They recommend removing duplicates to mitigate these issues. The synthetic dataset is complete with no missing or null values, and it includes both qualitative and quantitative variables. All categorical variables that are not binary will be encoded as dummy variables for modeling purposes. Given the data’s completeness and the encoding approach, data sparsity is expected to be minimal and not a concern for this analysis.

**Data Analytics Tools and Techniques**: The Design of the Study: This study will employ a GLM, specifically MLR, to explore the relationship between listing characteristics and nightly rental prices in Pennsylvania. Because GLMs do not require the raw input data to be normally distributed, no preliminary normality tests will be performed. As Midway and White (2025) explain, testing for normality on raw data instead of the residuals can reduce model performance, especially in larger datasets.  
During the modeling process, it is important to address potential multicollinearity. Variance inflation factor (VIF) scores will be calculated to identify and eliminate variables that show a high degree of multicollinearity. The general recommendation for VIF cut off is 10, but a study conducted by Jeng (2023) suggests that this may not be strict enough and enforcing a universal cut off value is impractical. This analysis will adopt a VIF cut off of 5 to eliminate variables. Further, backward elimination will be applied to iteratively remove statistically insignificant variables. This will ensure that the final model retains only variables that contribute meaningfully to the response variable.  
After the final model is constructed, the dataset will be split into training and testing subsets to evaluate predictive performance. Performance metrics such as R-squared (R2) and root mean square error (RMSE) will be calculated to quantify model accuracy and reliability.  
To assess model assumptions, residual diagnostics will be performed. A Q-Q (quantile-quantile) plot will be generated to evaluate whether the residuals are approximately normally distributed. As Waples (2024) notes, if the residuals fall along a straight line, it indicates that the model is generally valid for hypothesis testing and p-value interpretation. A histogram of the residuals will also be created for visual inspection, and a scatter plot of fitted versus actual values will be used to assess how closely the model’s predictions align with observed outcomes.  
Visual outputs including univariate/bivariate graphs, diagnostic plots, and model summaries will be compiled and presented using PowerPoint slides.

**Justification of Tools/Techniques:** Python will be used for the creation of this regression model due to its widespread use in data science and machine learning. According to Elhalid et al. (2023), Python is a highly accessible language because of its English-like syntax, making it easy to learn and apply across analytics tasks. Additionally, Python includes a wide range of libraries, such as statsmodels and scikit-learn, which offer effective implementations of statistical models, including MLR.  
While languages such as R and SAS are also strong options, Python will be used for its versatility, reproducibility, and compatibility with synthetic data generation tools such as Faker. Furthermore, Hill et al. (2024) found that Python performs comparably to R in estimation tasks for univariate statistics and analysis of variance. Python’s overall ecosystem and ease of integration into broader data pipelines makes it the preferred tool for this study.

**Project Outcomes**:This study will seek to develop a MLR model to estimate the nightly rental price of Airbnb listings in Pennsylvania using a synthesized dataset containing 13 predictor variables. The goal is twofold: (1) to identify which listing characteristics most strongly influence rental pricing, as determined by the magnitude of their regression coefficients, and (2) to construct a model that can predict prices with reasonable accuracy, defined as an R2 value greater than 0.60.   
The resulting model will be reusable for pricing simulations or applied analysis in similar real estate or rental market contexts. It may also assist in informing pricing decisions through an interpretable equation-based framework. Camatti et al. (2024) demonstrated that, despite the rise of advanced machine learning techniques, traditional linear models like MLR remain valuable for determining the significance of predictors in Airbnb pricing.

**Projected Project End Date**: 6/15/2025

**Sources**:  **SORT ALPHABETICALLY BY LAST NAME**

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

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.

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.

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.

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.

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.

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.

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.

To be filled out by a course mentor:

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Mentor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by: Click here to enter text.

Comments: Click here to enter text.