**SCH-MGMT 661**

**Application of Artificial Intelligence (AI) in Business – SCH-MGMT 661**

# Assignment 1: Price Prediction using Linear Regression

## 1.**Data Exploration and Cleaning**

**1.A Ensuring Listings Belong to Asheville, NC**

The dataset was filtered to include only Airbnb listings where host\_location contains "Asheville, NC" to ensure that only relevant data was used.

**1.B Descriptive statistics of 5 variables**

The following table summarizes key descriptive statistics for selected features:

A screenshot of a computer screen

AI-generated content may be incorrect.

**1.C Handling Missing Values**

* **Price**: Dropped rows where price was missing.
* **Accommodates, Bathrooms, Bedrooms, Minimum Nights**: Missing values were replaced with the **median**.
* **Review Scores Rating**: Missing values were replaced with **0**, and a new variable has\_review\_scores was created:
  + has\_review\_scores = 1 if review\_scores\_rating > 0
  + has\_review\_scores = 0 if review\_scores\_rating == 0

**1.D Discussion and Reflection**

* Handling missing values prevents unnecessary data loss and ensures a complete dataset for modeling. By filling missing values in accommodates, bathrooms, and bedrooms using median imputation and setting review\_scores\_rating to 0 for missing entries, we maintain data integrity while allowing the model to recognize differences between reviewed and non-reviewed listings.
* Potential improvements: Additional cleaning, such as removing extreme outliers in minimum\_nights, could improve model performance. Some properties have abnormally high minimum\_nights values (e.g., 365+ days), which likely represent long-term rentals rather than short-term Airbnb stays. Filtering these outliers would prevent them from skewing price predictions and leading to higher error rates.
* Applying log transformation to price can stabilize variance and improve model accuracy. Since price distribution is highly skewed, using log(price) reduces the impact of extreme values, making it easier for the model to detect meaningful pricing patterns.
* Dropping highly correlated features can reduce redundancy and improve model efficiency. Since accommodates, bedrooms, and bathrooms provide similar information, removing one of them prevents multicollinearity issues, ensuring that each predictor contributes unique value to the model.

## 2. **Exploratory Data Analysis (EDA)**

Here are all the visualizations for the 5 variables

**A collage of graphs

AI-generated content may be incorrect.**

**A graph of a number of guests

AI-generated content may be incorrect. A graph of a number of bedrooms per listing

AI-generated content may be incorrect.A graph of a number of bathrooms per listing

AI-generated content may be incorrect.**

**A graph with red line

AI-generated content may be incorrect. A graph of a number of nights

AI-generated content may be incorrect.**

**Histograms of Key Variables**

* Accommodates: Most listings accommodate 2–4 guests, with smaller rentals being the most common, while larger properties are less frequent.
* Bathrooms: The majority have 1–2 bathrooms, with larger homes having more, but they are relatively rare.
* Bedrooms: One-bedroom listings dominate, catering to solo travelers and couples, while larger homes form a smaller portion
* Review Scores Rating: Most listings have ratings above 90, indicating overall high guest satisfaction, with few low-rated properties.
* Minimum Nights: The distribution has a long tail, with most listings allowing 1–3 nights, but some have extreme outliers (100+ nights). Correlation Analysis

**A correlation matrix was generated to explore relationships between variables:**

* **Accommodates, Bedrooms, and Bathrooms are Highly Correlated:** These features show strong positive correlations (**>0.79**), indicating redundancy. Including all may introduce **multicollinearity**, which can distort regression coefficients. Dropping one (e.g., bedrooms) could improve model efficiency.
* **Moderate Correlation Between Price and Accommodates/Bathrooms/Bedrooms (~0.49):** Larger listings tend to have higher prices, but this relationship is not very strong, suggesting other factors (e.g., location, amenities) also influence pricing.
* **Weak or No Correlation for Review Scores Rating (-0.06) and Minimum Nights (0.01):** These features have almost no relationship with price, meaning they may not contribute significantly to the prediction model. Removing them or transforming them into categorical variables (e.g., "short-term vs. long-term stays") could improve performance.
* **Review Scores Rating and Has Review Scores are Almost Identical (0.99):** Since has\_review\_scores is derived from review\_scores\_rating, one of them should be removed to avoid redundancy.
* **Negative Correlation of Minimum Nights with Review Scores (-0.29):** This suggests that properties requiring longer stays tend to receive lower review scores, possibly due to guest dissatisfaction with strict booking policies.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Discussion and Insights**

* More bedrooms and bathrooms typically indicate a higher price.
* Listings with no reviews (review score = 0) might negatively impact predictions.

Outliers in minimum\_nights should be considered for removal.

## **3. Model Building and Evaluation**

**Train-Test Split**

* **80% training data**, **20% testing data**.

**Linear Regression Model**

The model was trained using:

* accommodates
* bathrooms
* bedrooms
* review\_scores\_rating
* minimum\_nights
* has\_review\_scores

**Performance Evaluation**

* **Mean Squared Error (MSE)**: *Evaluated and computed in the notebook*
* **R-squared (R²)**: *Measured to assess variance explanation*

**Findings & Suggested Improvements**

Discussion After Model Evaluation

After running the model evaluation, the results show a Mean Squared Error (MSE) of 20656.70 and an R-squared (R²) value of 0.18. These values indicate that the model currently has high error and low explanatory power, meaning it does not accurately predict price.

Key Observations:

* The high MSE suggests that predictions have a large variance from actual prices, possibly due to outliers, missing key features, or multicollinearity.
* The low R² (0.18) means that the selected features explain only 18% of the variance in price, indicating that many important pricing factors are missing.

Potential Reasons for Poor Performance:

* Presence of outliers in price and minimum\_nights may be distorting the model's ability to learn meaningful patterns.
* Multicollinearity between accommodates, bedrooms, and bathrooms can reduce model stability by introducing redundant information.
* Weak features like review\_scores\_rating and minimum\_nights show little to no correlation with price, making them ineffective predictors.
* Missing key features such as neighborhood, property\_type, and amenities likely play a major role in pricing but are not included in the model.
* Linear assumption may not capture complex price variations, suggesting that non-linear models might perform better.

Next Steps for Improvement:

* Handling outliers by applying log transformation on price and filtering extreme values in minimum\_nights could improve model stability.
* Feature selection should be refined by removing redundant predictors, such as dropping bedrooms due to its high correlation with accommodates.
* Including additional features like neighborhood, property\_type, and amenities can provide a better understanding of pricing determinants.
* Exploring non-linear models like Random Forest or Gradient Boosting might improve predictive accuracy by capturing complex relationships.
* Implementing cross-validation will ensure that model performance is evaluated consistently across different data splits.

These improvements should lead to lower MSE, higher R², and better price predictions.

**Conclusion**

This report analyzed Airbnb listings in Asheville, NC using **linear regression** to predict price based on key features such as accommodates, bathrooms, bedrooms, review\_scores\_rating, and minimum\_nights. The initial data exploration revealed **outliers, missing values, and multicollinearity**, which were addressed through data cleaning and feature selection. Histograms and correlation matrices provided deeper insights into the distribution and relationships among variables.

The model evaluation showed a **high Mean Squared Error (MSE) and low R² (0.18)**, indicating that the chosen features did not fully explain price variations. Further improvements, such as **removing redundant features, adding categorical variables (e.g., neighborhood and property type), and applying log transformation** to stabilize price distribution, were suggested. Additionally, experimenting with **non-linear models like Random Forest or Gradient Boosting** could enhance predictive accuracy.

In summary, while the current model provides a **baseline understanding of Airbnb pricing**, additional feature engineering and advanced modeling techniques could **significantly improve performance** and **deliver more accurate price predictions**.