Logistic Regression Exercise on "CentralAir"

In this exercise, we'll predict whether a house has central air conditioning (variable "CentralAir") using logistic regression. This analysis will help us understand the influence of features like **"GrLivArea"** (above-grade living area) and **"OverallQual"** (overall quality) on this binary outcome.

Context:

Logistic regression will help us predict the probability of a house having central air conditioning, which is a binary classification problem. Alongside model building, we'll evaluate its performance using metrics like accuracy, confusion matrix, ROC, and AUC.

Questions:

1. **Exploratory Data Analysis**
   - What is the distribution of the "CentralAir" variable? Check for balance between houses with and without central air.
   - Use a correlation matrix or plot to identify which features (e.g., "GrLivArea," "OverallQual") may influence "CentralAir."
2. **Data Preparation**
   - Why is it necessary to encode "CentralAir" as 1 (Yes) and 0 (No) for logistic regression?
   - Check for missing values in the chosen predictors and explain how you would handle them.
3. **Introduction to Logistic Regression**
   - Briefly explain why logistic regression is suitable for binary classification. Describe how logistic regression models probabilities between 0 and 1.
   - Define the logistic regression equation for predicting "CentralAir" based on features like "GrLivArea" and "OverallQual."
4. **Model Training**
   - Write code to fit a logistic regression model using predictors such as "GrLivArea" and "OverallQual."
   - What is the logistic regression equation after training? Interpret the coefficients briefly.
5. **Hypothesis Testing**
   - Conduct a hypothesis test on the coefficient for "GrLivArea" at a 5% significance level. What does this test indicate about the relationship between living area and the likelihood of having central air?
   - If the "GrLivArea" coefficient is statistically significant, interpret its effect on the probability of a house having central air.
6. **Model Interpretation**
   - Interpret the coefficient of "OverallQual." What does a one-unit increase in "OverallQual" suggest about the odds of having central air conditioning?
   - Discuss the implication of positive or negative coefficient values for understanding relationships in the model.
7. **Introduction to Confusion Matrix and Accuracy**

- **New Concept**: The confusion matrix is a table that shows how well the model classifies each class. Each entry represents counts of True Positives, True Negatives, False Positives, and False Negatives.
- Accuracy is the proportion of correct predictions out of total predictions, giving an overall sense of model performance.
- Why are confusion matrices and accuracy important for evaluating binary classification models?

8. **Model Evaluation with Confusion Matrix and Accuracy**
    - Generate a confusion matrix for the model's predictions and interpret each element in the matrix.
    - Calculate the model's accuracy based on the confusion matrix. What does this metric suggest about model performance?
9. **Introduction to ROC and AUC**
    - **New Concept**: ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) are metrics used to evaluate the model's ability to distinguish between classes. The ROC curve shows the trade-off between True Positive Rate and False Positive Rate. AUC summarizes the ROC curve; values closer to 1 indicate a strong classifier.
    - Why might ROC and AUC be valuable additions to accuracy in evaluating logistic regression performance?
10. **ROC and AUC Analysis**
    - Plot the ROC curve for the logistic regression model and compute the AUC score.
    - Based on the AUC score, how well does the model distinguish between houses with and without central air conditioning?