You are a data scientist at **HealthAnalytics Inc.**, responsible for developing a linear regression model to predict **medical insurance costs** for individuals based on their personal attributes. Your goal is to create a predictive model that can accurately estimate the insurance charges given a set of features. The dataset provided includes the following variables for several individuals: **age**, **sex**, **BMI** (Body Mass Index), **number of children**, **smoking status**, **region**, and **medical insurance charges**.

Your Tasks:

1. **Data Preprocessing:**
    - **Handle Missing Data**: Identify and treat any missing values in the dataset by either removing them or imputing appropriate values.
    - **Encode Categorical Variables**: Convert categorical features such as 'sex', 'smoker', and 'region' into numerical formats using techniques like one-hot encoding or label encoding.
    - **Scaling/Normalizing Features**: Apply feature scaling (if necessary) to continuous variables like age, BMI, and children for better model performance.
2. **Linear Regression Model Development:**
    - **Feature Selection**: Select relevant features from the dataset that will be used to predict medical insurance costs (age, sex, BMI, children, smoker, region).
    - **Model Building**: Implement a **linear regression** model to predict medical insurance charges (target variable) using the features selected.
    - **Multicollinearity Check**: Perform checks for multicollinearity (e.g., using the Variance Inflation Factor (VIF)) and eliminate highly correlated features, if needed.
    - **Model Training**: Fit the linear regression model on the training data to establish a relationship between the features and the target variable (insurance costs).
3. **Model Evaluation:**
    - **Data Splitting**: Divide the dataset into **training and testing sets** (e.g., 80% training and 20% testing) to evaluate the model's generalization performance.
    - **Performance Metrics Calculation**: After training the model on the training set, evaluate its performance on the testing set by calculating:
        - **Mean Absolute Error (MAE)**: Measures the average magnitude of the errors.
        - **Mean Squared Error (MSE)**: Measures the average squared difference between actual and predicted values.
        - **Root Mean Squared Error (RMSE)**: The square root of MSE, providing an error estimate in the same units as the target variable.
        - **R-squared (R2) Score**: Indicates the proportion of variance in the target variable explained by the model.

        - **Adjusted R-squared**: Adjusts the R2 score based on the number of predictors, penalizing models that include irrelevant features.
        - **Residual Sum of Squares (RSS)**: Quantifies the total squared error between the predicted and actual values.
        - **Explained Variance Score**: Evaluates how much of the variance in the target variable is captured by the model.
4. **Feature Importance Analysis:**
    - **Coefficient Interpretation**: Examine the coefficients of the linear regression model to determine the impact of each feature (age, BMI, smoking status, etc.) on medical insurance costs.
    - **Feature Ranking**: Identify the most important features based on their contribution to the model, with a particular focus on whether smoking status or BMI significantly increases costs.
5. **Visualization:**
    - **Scatterplot for Model Performance**: Create a scatterplot showing **actual vs. predicted insurance charges**. This visual will help in assessing how well the model fits the data, and it will highlight any discrepancies between predicted and actual values (e.g., overfitting or underfitting).
6. **Residual Analysis:**
    - **Residual Plot**: Visualize the residuals (difference between predicted and actual values) to check for patterns, ensuring that errors are randomly distributed, a key assumption in linear regression.