📄 Table of Contents

✅ Suitability

The dataset comprises a mix of continuous, categorical, and binary variables, making it appropriate for linear regression analysis, provided that categorical variables are properly encoded .

Initial assessment reveals data quality issues such as duplicate entries, missing values, outliers, and skewed distributions. These anomalies necessitate thorough data cleaning procedures, which will be detailed subsequently.

Importantly, the dataset does not exhibit significant multicollinearity among independent variables, as indicated by acceptable Variance Inflation Factor (VIF) scores . This characteristic aligns well with the assumptions underlying linear regression models.

Caution is advised to avoid common pitfalls such as presuming linear relationships across all variables and overfitting the model by including irrelevant features. These concerns will be addressed through correlation analyses and data visualization techniques.

🗺️ Plan

1. Data Collection

- Utilize the "Medical Cost Personal Datasets" as the primary data source, deemed relevant for predicting insurance charges.

2. Data Cleaning

- Handling Missing Values: Identify and remove any records with missing data .
- Removing Duplicates: Detect and eliminate duplicate records to prevent data redundancy .
- Encoding Categorical Variables: Transform categorical variables (e.g., gender, smoker status, region) into numerical formats using label encoding or one-hot encoding .
- Outlier Treatment: Identify outliers using z-scores and apply log transformations or quantile-based clipping to mitigate their impact .

3. Exploratory Data Analysis (EDA)

- Descriptive Statistics: Compute mean, median, standard deviation, quartiles, and range for each variable.
- Distribution Analysis: Visualize distributions of continuous variables using histograms and boxplots to detect skewness and outliers.
- Categorical Analysis: Use bar charts to display the frequency distribution of categorical variables.
- Relationship Exploration: Examine relationships between independent variables and the target variable ('charges') using:
    - Boxplots for binary variables (e.g., smoker status, gender).
    - Boxplots for discrete variables (e.g., number of children, region).
    - Scatter plots with regression lines for continuous variables (e.g., BMI, age).
- Multicollinearity Check: Calculate VIF scores to assess multicollinearity among predictors .
- Correlation Analysis: Determine the strength and direction of relationships using appropriate correlation coefficients:
    - Point-Biserial Correlation for binary and continuous variable pairs .

    - Spearman's Rank Correlation for ordinal or non-normally distributed variables .
    - Pearson's Correlation for continuous variables with linear relationships .
    - ANOVA for assessing differences in means across categorical groups.

4. Model Training

- Data Splitting: Partition the dataset into training (80%) and testing (20%) subsets using random sampling to ensure reproducibility.
- Model Development: Develop two models:
    - A standard multiple linear regression model.
    - A Lasso regression model for feature selection and regularization.
- Model Comparison: Evaluate and compare the performance of both models to determine the most effective approach.

5. Evaluation

- Performance Metrics: Assess model accuracy using:
    - Coefficient of Determination (R^2).
    - Root Mean Square Error (RMSE).
    - Comparison of actual versus predicted values through residual plots .