# Task 5

## Insights:

1. **Dataset Overview**:
    - The dataset was successfully loaded and basic exploratory data analysis (EDA) was performed.
    - Key statistics (mean, median, missing values) were computed.
    - Data types were checked for each column.
2. **Missing Values**:
    - Missing data was identified and visualized.
    - Imputation strategies (mean/median filling) were discussed/applied where necessary.
3. **Exploratory Data Analysis**:
    - Several visualizations were created:
        - Histograms to see feature distributions
        - Box plots for outlier detection
        - Correlation heatmaps to find relationships among variables
    - Important patterns:
        - Certain features had strong positive or negative correlations.

        - A few outliers were present that might affect modeling later.

4. **Feature Engineering**:
    - New features were created based on the existing ones (e.g., ratios, combined features).
    - Categorical variables were encoded.
5. **Modeling Preparation**:
    - Data was split into training and testing sets.
    - Feature scaling (StandardScaler/MinMaxScaler) was applied.
6. **Initial Modeling**:
    - Basic models (Linear Regression, Decision Trees, Random Forests) were trained.
    - Performance metrics like accuracy, RMSE, and $R^2$ were evaluated.
    - Random Forest outperformed basic models based on preliminary results.
7. **Conclusion**:
    - Data cleaning, visualization, and basic modeling were completed.
    - Suggestions were made for hyperparameter tuning and advanced models (like XGBoost).

## Interview Questions

**1. What is EDA and why is it important?**

- **EDA (Exploratory Data Analysis)** is the process of analyzing data sets to summarize their main characteristics, often using visual methods.
- **Importance:**
    - Understand the structure, patterns, and relationships in data.
    - Detect outliers, missing values, and anomalies.
    - Guide feature selection, preprocessing, and modeling strategies.

**2. Which plots do you use to check correlation?**

- **Heatmap** (especially with a correlation matrix, `sns.heatmap()`)
- **Pairplot** (scatterplots for each pair of features, `sns.pairplot()`)
- **Scatter plots** (for two continuous variables)
- **Correlogram** (correlation visualization

**3. How do you handle skewed data?**

- **Apply transformations**:
    - Log transformation (`np.log1p(x)`)
    - Square root transformation
    - Box-Cox transformation
    - Yeo-Johnson transformation
- **Use robust models** that are less sensitive to skewed data.
- **Remove outliers** if appropriate.

---

**4. How to detect multicollinearity?**

- **Correlation Matrix**: High correlation between features indicates potential multicollinearity.
- **Variance Inflation Factor (VIF)**:
  - VIF > 5 (or sometimes > 10) indicates high multicollinearity.

**5. What are univariate, bivariate, and multivariate analyses?**

- **Univariate Analysis**:
  - Analyze a single variable. (e.g., histogram, boxplot)
- **Bivariate Analysis**:
  - Analyze the relationship between two variables. (e.g., scatter plot, correlation)
- **Multivariate Analysis**:
  - Analyze more than two variables simultaneously. (e.g., multiple regression, PCA)

**6. Difference between heatmap and pairplot?**

| Aspect | Heatmap | Pairplot |
|---|---|---|
| Purpose | Show correlation or intensity in matrix form | Show pairwise relationships with plots |
| Data Type | Numeric correlations (or matrix-like data) | Multiple variable distributions |
| Output | Color-coded matrix | Grid of scatterplots and histograms |
| Library | `seaborn.heatmap()` | `seaborn.pairplot()` |

**7. How do you summarize your insights?**

- Identify key findings (e.g., important correlations, outliers, feature distributions).
- Highlight issues (missing data, skewness, multicollinearity).
- Link findings to next steps (feature engineering, modeling decisions).
- Present summaries visually (charts, graphs) and textually (bullet points or brief reports).