### Regression Report

#### **1. Introduction**

* Dataset Description: Describe the pakwheels\_used\_cars.csv dataset, which includes features like engine\_cc, mileage, and the target variable price.
* Pakwheels\_used\_cars.csv dataset is present in the same directory as the report.
* Objective: Explain the goal of predicting car prices using various regression algorithms based on the given features.

#### **2. Data Cleaning and Preparation**

* Loading the Data: Load the dataset and display the first few rows.
* Handling Missing Values:
  + Identify missing values and print the results.
  + Fill missing values in numerical columns with the mean.
  + For categorical columns, fill missing values with the mode.
* Encoding Categorical Variables: Convert categorical features into numerical values using Label Encoding.
* Scaling Numerical Features: Normalize numerical features using StandardScaler.

#### **3. Data Analysis and Visualization**

* Summary Statistics: Generate and display summary statistics for the dataset.
* Histograms: Create histograms to visualize the distribution of numerical features.
* Scatter Plots: Generate scatter plots to explore relationships between features and the target variable.
* Box Plots: Use box plots to visualize the distribution and detect outliers in the features.
* Correlation Heatmaps: Create heatmaps to show the correlation between numerical features and the target variable.

#### 4. **Model Building**

**Models Used:**

1. **Linear Regression:**
   * Reason for Use: Linear regression is a fundamental model for understanding the relationship between the target variable and predictors. It provides a baseline for comparison with more complex models.
   * Performance:
     + Mean Squared Error: 0.629
     + R² Score: 0.685
2. **Decision Tree Regression:**
   * Reason for Use: Decision trees can capture non-linear relationships and interactions between features without requiring feature scaling. They are easy to visualize and interpret.
   * Performance:
     + Mean Squared Error: 1.107
     + R² Score: 0.448
3. **Random Forest Regression:**
   * Reason for Use: Random Forests are an ensemble learning method that reduces overfitting by averaging multiple decision trees. They generally provide better performance and robustness.
   * Performance:
     + Mean Squared Error: 0.494
     + R² Score: 0.758

### **Model Performance Comparison**

Based on the Mean Squared Error (MSE) and R² Score, the Random Forest Regression model performs the best in this scenario.

* Random Forest Regression has the lowest Mean Squared Error (0.494) and the highest R² Score (0.758), indicating that it predicts the car prices more accurately than the other models.
* Linear Regression performs moderately well, but not as good as Random Forest.
* Decision Tree Regression has the highest Mean Squared Error (1.107) and the lowest R² Score (0.448), making it the least accurate model among the three.

Best Model: Random Forest Regression

* **Reason**:

It provides the best balance between bias and variance, capturing complex patterns in the data and reducing overfitting through ensemble learning. This results in better predictive performance for the given dataset.

#### **Summary of Findings**

1. **Data Preparation:**
   * The dataset was successfully loaded and initial inspection showed the presence of missing values.
   * Missing values in numeric columns were handled by filling them with the mean, while categorical columns were filled with the mode.
   * Categorical variables were converted to numerical format using Label Encoding.
   * Numerical features were standardized using StandardScaler to ensure better model performance.
2. **Exploratory Data Analysis (EDA):**
   * Histograms and box plots provided insights into the distribution and potential outliers in the dataset.
   * A correlation heatmap identified relationships between different features and the target variable (price).
3. Mo**del Training and Evaluation**:
   * Three regression models were used: Linear Regression, Decision Tree Regression, and Random Forest Regression.
   * Performance metrics (Mean Squared Error and R² Score) were used to evaluate and compare the models.
   * Random Forest Regression emerged as the best model with the lowest Mean Squared Error (0.494) and highest R² Score (0.758), indicating superior predictive performance.
   * Linear Regression provided a reasonable baseline, while Decision Tree Regression showed higher error and lower accuracy.

#### **Possible Future Work:**

1. **Hyperparameter Tuning:**
   * Perform hyperparameter optimization for each model to further improve performance. Techniques such as Grid Search or Random Search could be used to find the best parameters.
2. **Feature Engineering:**
   * Explore and create new features that might better capture the underlying patterns in the data. For instance, combining related features or deriving new ones from existing features.
3. **Handling Outliers:**
   * Investigate and handle outliers in the dataset more rigorously. Outliers can significantly affect the performance of some models.

The graphs for pakwheels regression:





