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
        - R^2 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
        - R^2 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
        - R^2 Score: 0.758

Model Performance Comparison

Based on the Mean Squared Error (MSE) and R^2 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 R^2 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

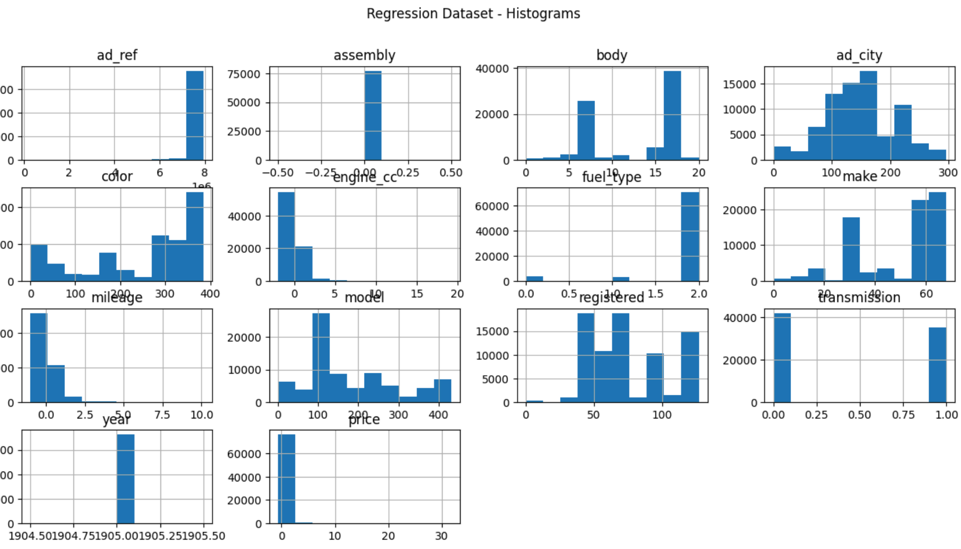
1. **Data Preparation:**
    - The dataset was successfully loaded and initial inspection showed the presence of missing values.
    - Missing values in numeric columns were handled by filling them with the mean, while categorical columns were filled with the mode.
    - Categorical variables were converted 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. **Model Training and Evaluation**:
    - Three regression models were used: Linear Regression, Decision Tree Regression, and Random Forest Regression.

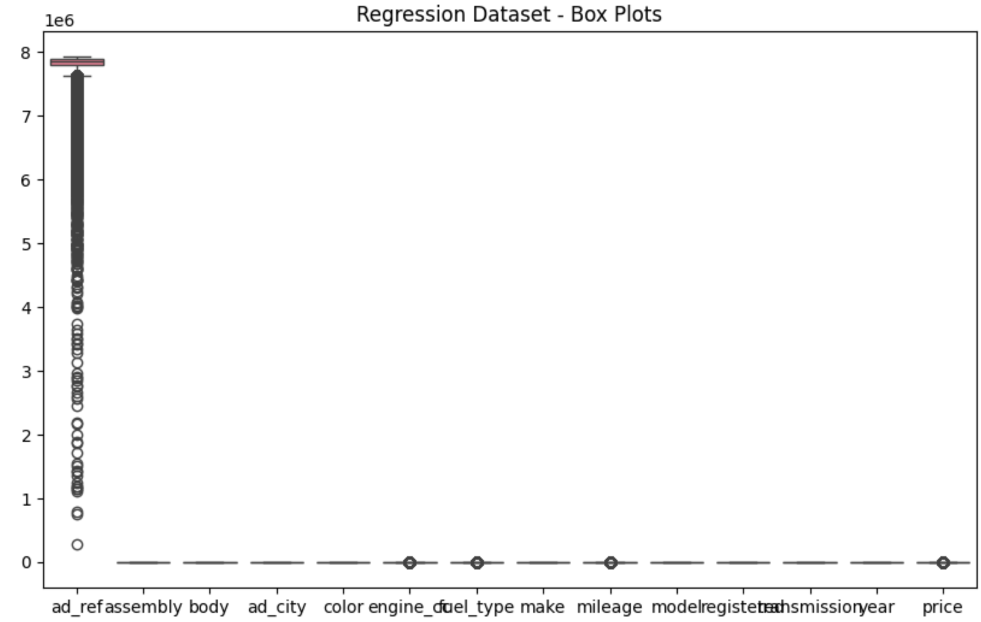
- Performance metrics (Mean Squared Error and R^2 Score) were used to evaluate and compare the models.
- Random Forest Regression emerged as the best model with the lowest Mean Squared Error (0.494) and highest R^2 Score (0.758), indicating superior predictive performance.
- Linear Regression provided a reasonable baseline, while Decision Tree Regression showed higher error and lower accuracy.

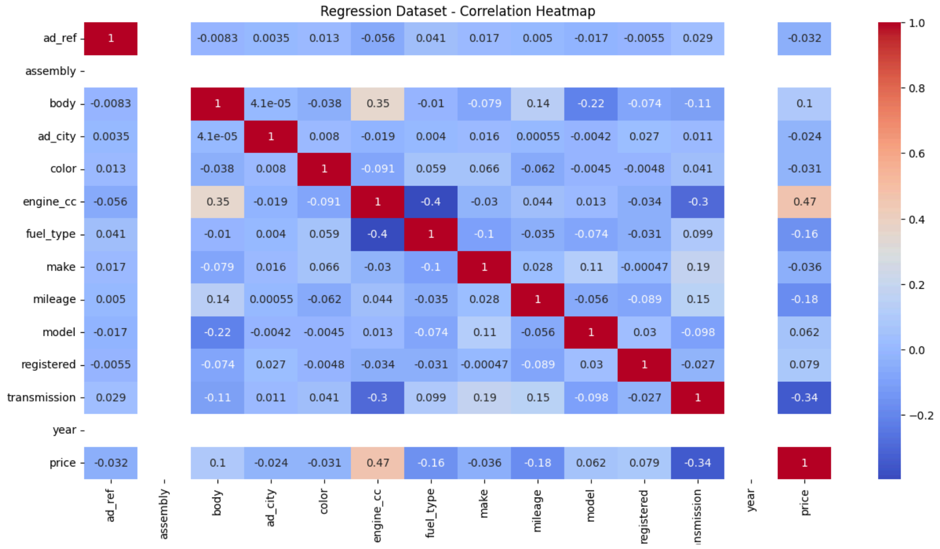
Possible Future Work:

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:


Regression Dataset - Histograms
ad_ref
assembly
body
ad_city
color
engine_cc
fuel_type
make
mileage
model
registered
transmission
year
price
1e6
0
2
4
6
8
20000
40000
60000
75000
50000
25000
−0.50
−0.25
0.00
0.25
0.50
5
10
15
20
100
200
300
400
0.0
0.5
1.0
1.5
2.0
2.5
5.0
7.5
10.0
0.25
0.75
1.00
1904.50
1904.75
1905.00
1905.25
1905.50
30


Regression Dataset - Box Plots
1e6
8
7
6
5
4
3
2
1
0
ad_ref
assembly
body
ad_city
color
engine_cc
fuel_type
make
mileage
model
registered
transmission
year
price



Regression Dataset - Correlation Heatmap

	ad_ref	assembly	body	ad_city	color	engine_cc	fuel_type	make	mileage	model	registered	nsmission	year	price
ad_ref	1		-0.0083	0.0035	0.013	-0.056	0.041	0.017	0.005	-0.017	-0.0055	0.029		-0.032
assembly														
body	-0.0083		1	4.1e-05	-0.038	0.35	-0.01	-0.079	0.14	-0.22	-0.074	-0.11		0.1
ad_city	0.0035		4.1e-05	1	0.008	-0.019	0.004	0.016	0.00055	-0.0042	0.027	0.011		-0.024
color	0.013		-0.038	0.008	1	-0.091	0.059	0.066	-0.062	-0.0045	-0.0048	0.041		-0.031
engine_cc	-0.056		0.35	-0.019	-0.091	1	-0.4	-0.03	0.044	0.013	-0.034	-0.3		0.47
fuel_type	0.041		-0.01	0.004	0.059	-0.4	1	-0.1	-0.035	-0.074	-0.031	0.099		-0.16
make	0.017		-0.079	0.016	0.066	-0.03	-0.1	1	0.028	0.11	-0.00047	0.19		-0.036
mileage	0.005		0.14	0.00055	-0.062	0.044	-0.035	0.028	1	-0.056	-0.089	0.15		-0.18
model	-0.017		-0.22	-0.0042	-0.0045	0.013	-0.074	0.11	-0.056	1	0.03	-0.098		0.062
registered	-0.0055		-0.074	0.027	-0.0048	-0.034	-0.031	-0.00047	-0.089	0.03	1	-0.027		0.079
transmission	0.029		-0.11	0.011	0.041	-0.3	0.099	0.19	0.15	-0.098	-0.027	1		-0.34
year														
price	-0.032		0.1	-0.024	-0.031	0.47	-0.16	-0.036	-0.18	0.062	0.079	-0.34		1