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.
- **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

- **Splitting Data**: Split the data into training and testing sets.
- **Applying Regression Models**:
    - **Linear Regression**: Train and evaluate a Linear Regression model.
    - **Decision Tree Regression**: Train and evaluate a Decision Tree Regressor.
    - **Random Forest Regression**: Train and evaluate a Random Forest Regressor.

5. Model Evaluation

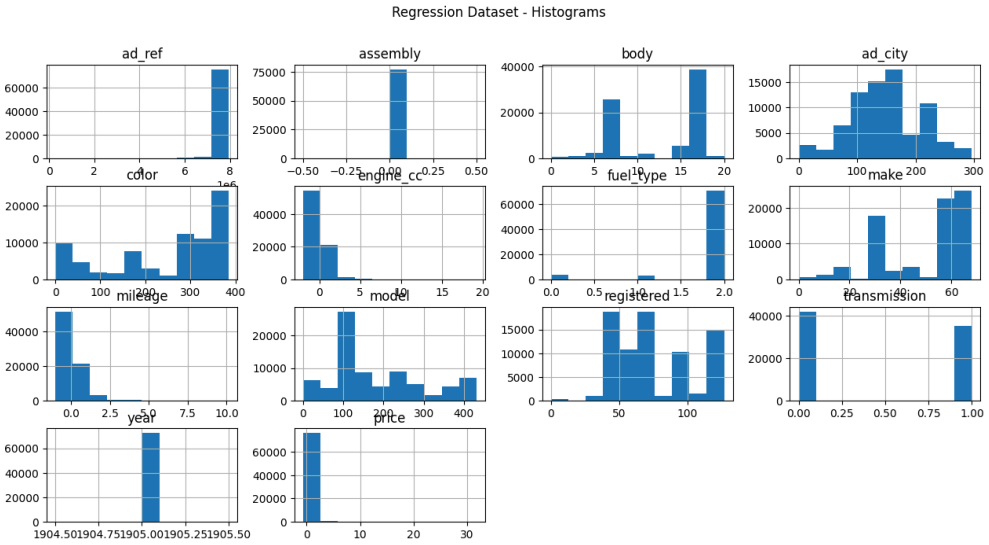
- **Performance Metrics**: Evaluate and compare the performance of each regression model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
- **Performance Comparison**:

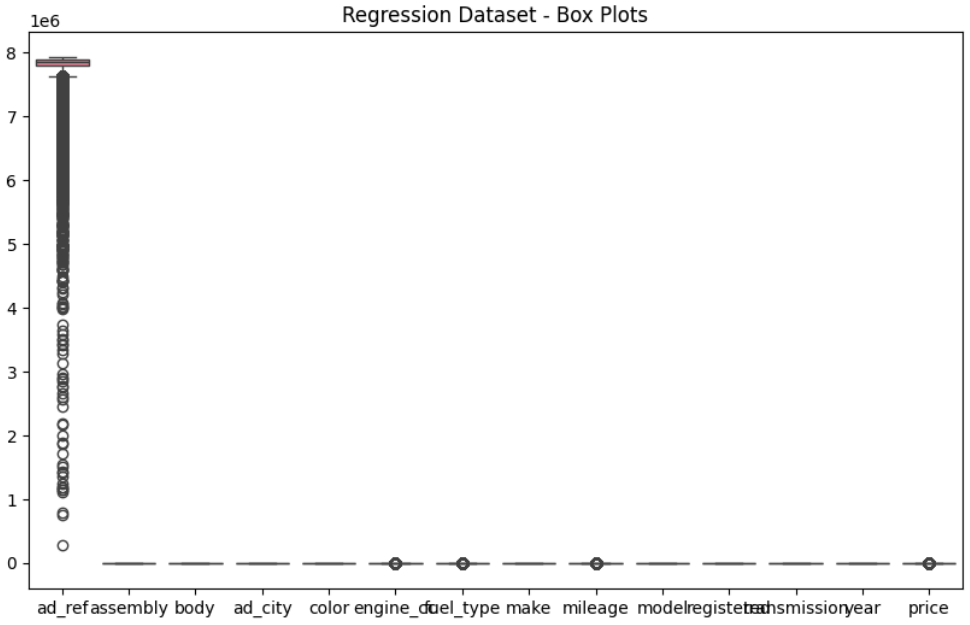
- Create a DataFrame to compare the performance of the regression models.
- Plot a bar chart to visualize the performance comparison.

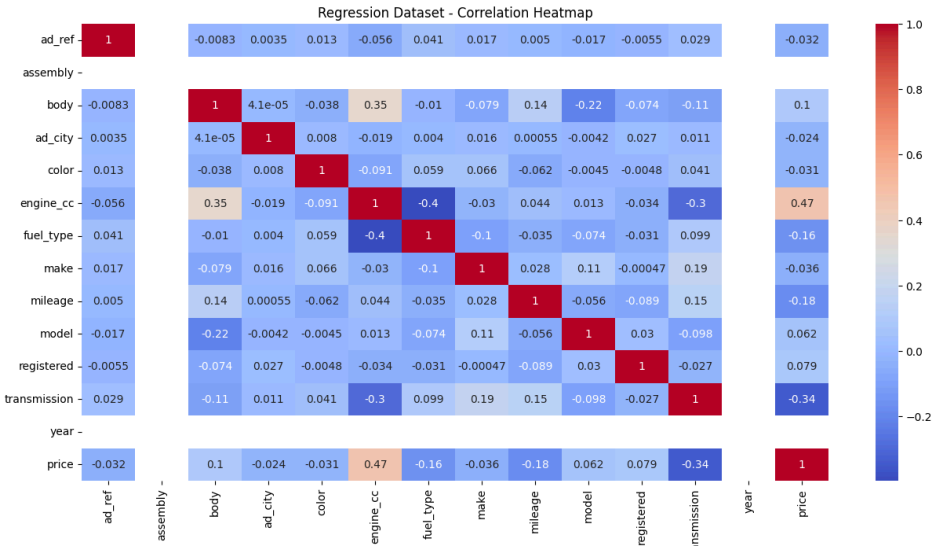
6. Conclusion

- **Summary of Findings**: Summarize the findings from the regression models, highlighting which model performed best and why.
- **Future Work**: Suggest areas for improvement, such as tuning model parameters, incorporating more features, or using advanced regression techniques.

The graphs for pakwheels regression:




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 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
ad_ref
assembly
body
ad_city
color
engine_cc
fuel_type
make
mileage
model
registered
transmission
year
price
1.0
0.8
0.6
0.4
0.2
0.0
−0.2