ASSIGNMENT 6

MULTI - LINEAR REGRESSION

Multi-Linear Regression Assignment Report:

1. Libraries Used

The following libraries were imported and used in the notebook:

- numpy
- pandas
- statsmodels.api
- matplotlib.pyplot
- seaborn
- lasso and ridge from sklearn.linear_model
- train_test_split
- statsmodel.formula.api
- influence plot from statsmodel.graphics.regressionplots

2. Data Loading

- The dataset ToyotaCorolla - MLR.csv was loaded into a DataFrame.

3. Exploratory Data Analysis (EDA)

- Renamed column Age_08_04 to Age.
- Converted the categorical variable Fuel_Type into dummy variables and dropped the original Fuel_Type column.
- Placed the target variable Price as the last column.
- Removed duplicated columns.
- Provided a description of the dataset's features.

4. Data Cleaning

- Verified that there were no null values.
- Identified and removed duplicate rows.
- Stored a cleaned version of the data for further analysis.

5. Descriptive Statistics

- Displayed summary statistics for the cleaned dataset.

6. Data Visualization

- Created histograms and Box-Plots for understanding the distribution of the data.

7. Feature Selection

- Selected features for the model.

8. Model Training

- Started off with assumptions of multi linear regression and
- linearity check
- feature selection
- in feature selection removed some features that were insignificant to the model building or happens to influence the prediction
- calculated the **VARIANCE INFLATION FACTOR** and interpreted the values of it respectively with the obtained values of the features

Assumption For Influential Observations

- removed some outliers according to the cooks distance metrics
- went on with the visualization of the influence plot

Assumptions Of The Error,

  - check the model with the error factor and satisfied the condition of mean error to be 0
- plotted q-q plot and identified skewness in the data
- Split the data into train and test (80-20)
- Trained the models using the training data.
- built three final models and predicted the price and finalized the final model based on the f statistic score of the model
- Implemented both Lasso and Ridge regression models.

9. Model Prediction

- Made predictions on the test set using both Lasso and Ridge regression models.

10. Additional Insights

- Discussed normalization and standardization techniques for feature scaling.
- Explained multicollinearity and the Variance Inflation Factor (VIF).