**Objective**:

**Assignment Task:**

Your task is to perform a multiple linear regression analysis to predict the price of Toyota corolla based on the given attributes.

**Dataset Description:**

The dataset consists of the following variables:

Age: Age in years

KM: Accumulated Kilometers on odometer

FuelType: Fuel Type (Petrol, Diesel, CNG)

HP: Horse Power

Automatic: Automatic ( (Yes=1, No=0)

CC: Cylinder Volume in cubic centimeters

Doors: Number of doors

Weight: Weight in Kilograms

Quarterly\_Tax:

Price: Offer Price in EUROs

**Taskes:**

1.Perform exploratory data analysis (EDA) to gain insights into the dataset. Provide visualizations and summary statistics of the variables. Pre process the data to apply the MLR.

2.Split the dataset into training and testing sets (e.g., 80% training, 20% testing).

3.Build a multiple linear regression model using the training dataset. Interpret the coefficients of the model. Build minimum of 3 different models.

4.Evaluate the performance of the model using appropriate evaluation metrics on the testing dataset.

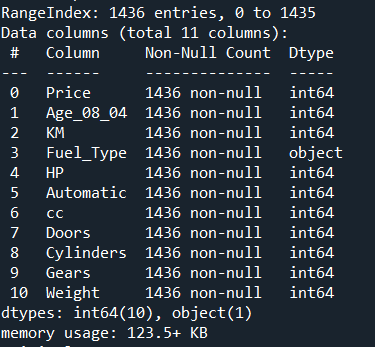
5.Apply Lasso and Ridge methods on the model.

1.Perform exploratory data analysis (EDA) to gain insights into the dataset. Provide visualizations and summary statistics of the variables. Pre-process the data to apply the MLR

Explanation –

**Understanding of Data**

* **Load the dataset**: Read the data into a Dataframe.
* **Inspect structure**: Use info() to check the first few rows and data types.
* **Check dimensions**: Identify the number of rows and columns using shape.



**Data Cleaning**

* **Handle missing values**:
  + Identify missing data with isnull().sum().
  + Decide to fill (fillna()), drop (dropna()), or impute missing values.

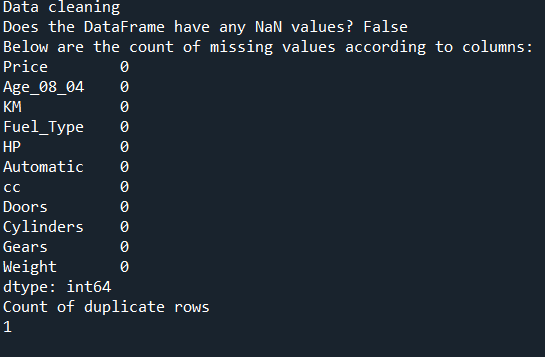
**Note - As there are no missing values into data, no action has been performed**

* **Fix incorrect data types**: Ensure dates, categories, and numerical columns are properly typed.

**Note – All the data type of columns as expected hence no action has been performed**

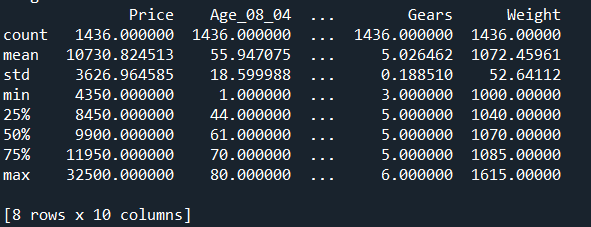
* **Remove duplicates** if needed

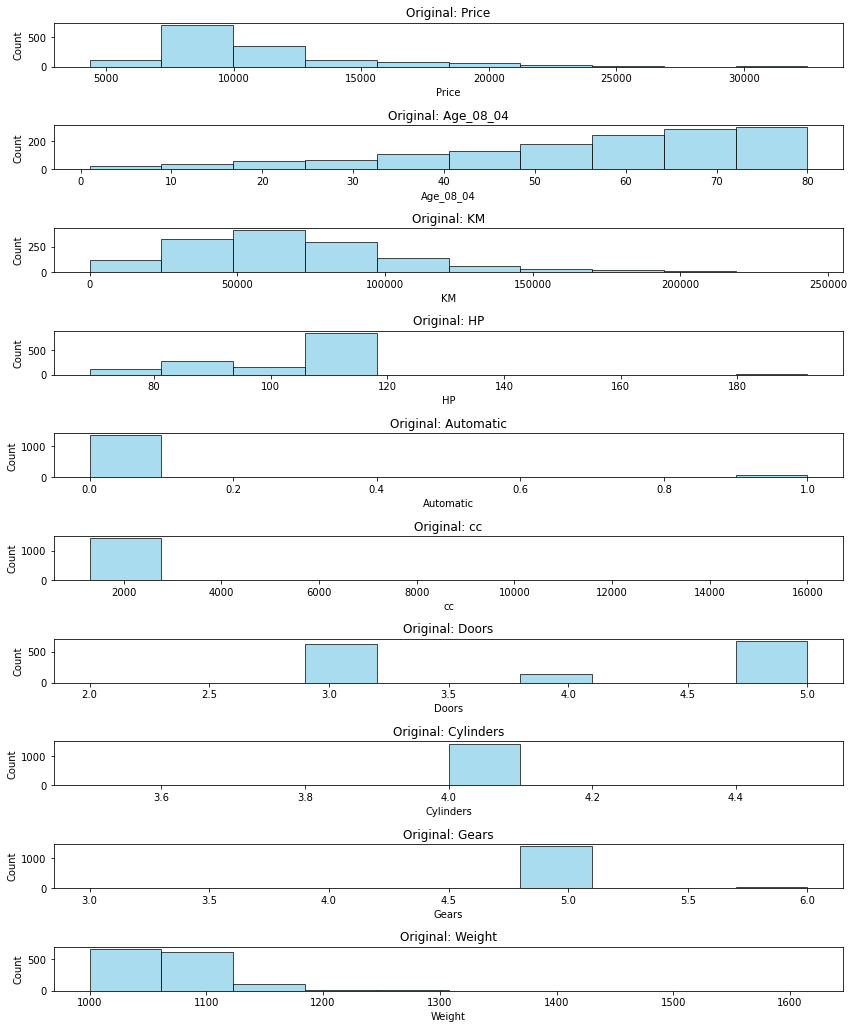
**Note – Count for duplicate as 1, dropped the duplicate column as its not useful for further insight**



### **Summary Statistics**

* Used describe () for a quick overview of mean, standard deviation, min, and max values.



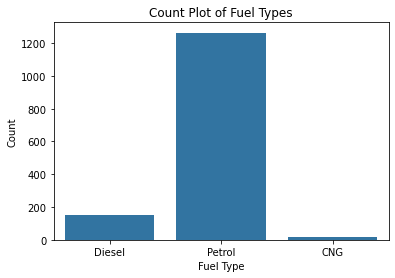
* Distributions of individual variables using:
* **Histograms** for continuous variables. 

**Key Insights:**

* **Price**: Most car prices are concentrated between 5,000 and 15,000 units, with very few cars priced higher than 30,000.
* **Age**: The distribution of car age shows a peak around 60-80 months, indicating a majority of cars are within this age range.
* **KM (Kilometers Driven)**: Most cars have driven less than 100,000 kilometers, with a decreasing frequency for higher values.
* **HP (Horsepower)**: The data peaks around 100 to 120 HP, showing that cars with higher horsepower are less common.
* **Automatic**: Most cars in the dataset are manual (0 represents manual, and 1 represents automatic), showing an imbalance favoring manual transmission.
* **CC (Engine Capacity)**: There are cars mostly clustered around 2,000 to 6,000 cc.
* **Doors**: The distribution suggests most cars have **4** doors, followed by **3-door** and **5-door** cars.
* **Cylinders**: Cylinder distribution appears to be consistent, with most cars having values around **4**.
* **Gears**: The majority of cars have 5 gears.
* **Weight**: Most cars have weights between **1,000** and **1,300** units.

**General Insights**:

* The dataset shows typical attributes of cars on the market.
* Price, kilometers driven, and age could significantly affect car price prediction.
* **Count plots** for categorical variables.



**Key Insights:**

* Petrol cars dominate the dataset.
* Diesel cars are much less frequent than petrol, and CNG cars are rare.
* This imbalance suggests fuel type may need to be treated carefully in regression or classification to avoid bias toward petrol.

### Suggestions for Analysis:

1. **Feature Selection**:
   * Variables like Age, KM, Fuel\_Type, HP, and CC seem important for price prediction.
2. **Imbalanced Data Handling**:
   * Consider balancing fuel types (if important for a specific model) using oversampling/undersampling or applying weight adjustments.

To handle the Imbalance Data for Categorical column applied the one-hot encoding technique.

Data observations:

* Since no missing values are present, all input features are valid without needing further preprocessing steps like imputation.
* Removing duplicates will slightly reduce overfitting, improving your model’s generalization ability.
* The cleaned and updated dataset should be used to refit and re-evaluate all models for more accurate comparisons.

2. Split the dataset into training and testing sets (e.g., 80% training, 20% testing).

Split the data into training and testing Three different linear regression models.

* Build below 3 models:
  1. Simple Linear Regression model

Output –



* 1. Model 2: Linear Regression with selected features (Age, KM, HP, Weight)



* 1. Model 3: Linear Regression without 'Doors' and 'Automatic'



Based on the provided Mean Squared Error (MSE) and R-squared (R²) values for the three models, here are the insights and recommendations:

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error (MSE)** | **R-squared (R²)** |
| **Model 1** | 1,926,044.83 | 0.8395 |
| **Model 2** | 1,934,478.82 | 0.8388 |
| **Model 3** | 1,902,836.13 | 0.8414 |

|  |  |
| --- | --- |
| **Model** | **Performance** |
| **Model 1** | Good, but slightly higher MSE and lower R² compared to Model 3. |
| **Model 2** | Slightly worse than Model 1 in both MSE and R². |
| **Model 3** | Best overall performance with the lowest MSE and highest R². |

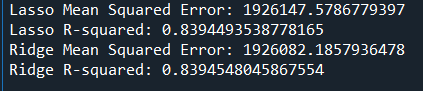
**Key Insight**

* **Model 3** performs **best in both accuracy (lowest MSE) and explanatory power (highest R²)**, making it the **most suitable model** for predicting the price of Toyota Corolla cars.
* **Model 1 and Model 2** are close competitors but do not outperform Model 3. Therefore, they may be considered less optimal.

5. Apply Lasso and Ridge methods on the model.

Lasso and Ridge regressions are **regularization techniques** used in linear regression models to improve prediction accuracy and handle multicollinearity by adding a penalty term to the cost function.

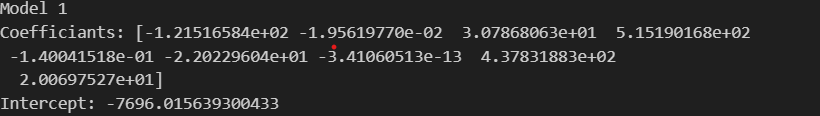
* Both Lasso and Ridge improve linear regression models by preventing overfitting.
* Lasso is ideal for feature selection, while Ridge is better for handling multicollinearity.



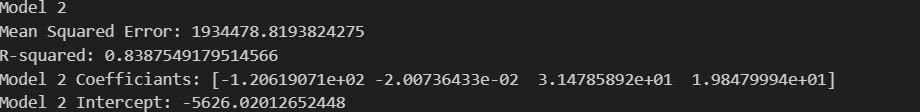
* Ridge regression shows slightly better performance, but both models are good.
* If the primary goal is accuracy with no feature elimination, Ridge is the best choice.
* If the goal is simplicity and interpretability, Lasso might be preferred.

6. Interpret the Coefficient:

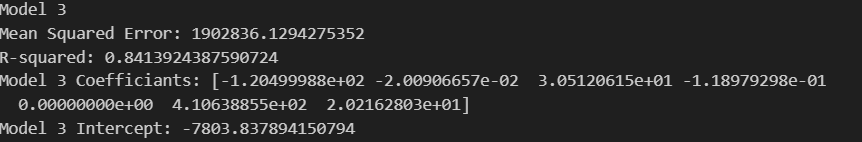
* Model 1:



* **Age**: Each additional month decreases the car price by **€121.52**.
* **KM**: Each kilometer driven reduces the price by **€1.95**.
* **HP**: Each extra horsepower increases price by **€30.79**.
* **Automatic = 1**: Automatic cars are worth **€515.19 more** than manual cars.
* **CC (Engine size)**: Slight negative effect per cc, **€14.04** reduction per cc.
* **Doors**: Not very impactful; more doors slightly reduce value (may vary by demand).
* **FuelType\_Diesel**: Practically 0 — no significant price difference from CNG baseline.
* **Weight**: Each extra kg increases price by **€437.83**, which is substantial.
* **FuelType\_Petrol**: Petrol cars are worth **€20.07 more** than CNG (base category).
* Model 2:



* **Age**: Price drop by €120 per month — very consistent.
* **KM**: Consistent — each km decreases value by **€0.02**.
* **HP**: Stronger influence — price increases **€31.48 per horsepower**.
* **Weight**: Lower than Model 1 — now only adds **€19.84 per kg**, likely due to other features being removed (collinearity reduced).
* Model 3:



* **Age**: Price drops by **€120.50 per month** — very consistent.
* **KM**: Same trend — very minor drop per kilometer.
* **HP**: Each extra HP adds **€30.51** to price.
* **CC**: Slight negative impact — not a strong predictor.
* **FuelType\_Diesel**: No added value over CNG (base fuel type).
* **Weight**: Adds **€410.64 per kg** — high impact!
* **FuelType\_Petrol**: Petrol cars are worth **€20.22 more** than CNG.

**Summary Table**

| **Feature** | **Model 1** | **Model 2** | **Model 3** | **Meaning** |
| --- | --- | --- | --- | --- |
| Age | -121.52 | -120.62 | -120.50 | Price ↓ by €120/month |
| KM | -0.0196 | -0.02007 | -0.02009 | Price ↓ per km driven |
| HP | 30.79 | 31.48 | 30.51 | Price ↑ per horsepower |
| Automatic | 515.19 | — | — | Auto cars ↑ €515 |
| CC | -0.14 | — | -0.1189 | Slight ↓ with bigger engine |
| Doors | -22.02 | — | — | Minor negative effect |
| FuelType\_Diesel | ~0 | — | 0 | Not significant |
| FuelType\_Petrol | 20.07 | — | 20.22 | Petrol cars ↑ ~€20 |
| Weight | 437.83 | 19.84 | 410.64 | Price ↑ per kg |

**Interview Questions:**

1.What is Normalization & Standardization and how is it helpful?

* Both normalization and standardization are essential tools in pre-processing to enhance model accuracy and interpretability. Choosing the right method depends on the type of model and data distribution.
* Normalization is a scaling technique that **rescales the values** of a feature to a **range of [0, 1]** or sometimes [-1, 1]. It adjusts the data to have minimum and maximum values of 0 and 1 respectively.
* Standardization transforms the data to have a **mean of 0 and a standard deviation of 1**. It assumes that the data follows a Gaussian distribution (normal distribution), though it can still be effective even if the data is not perfectly normal.

2.What techniques can be used to address multicollinearity in multiple linear regression?

**Principal Component Analysis (PCA)**:

* PCA transforms the original predictors into a smaller set of uncorrelated components. You can then use these components in the regression model. This is especially useful when you have a large number of predictors.

**Ridge Regression**:

* Ridge regression (L2 regularization) adds a penalty to the size of the coefficients, which can reduce the impact of multicollinearity. It shrinks the coefficients of correlated variables toward zero but does not eliminate them completely.

**Lasso Regression**:

* Lasso regression (L1 regularization) also adds a penalty to the model, but it has the ability to set some coefficients exactly to zero. This can effectively eliminate some variables, helping with multicollinearity.

**Increase Sample Size**:

* Sometimes, multicollinearity arises because the sample size is too small. Increasing the sample size can help reduce the effects of multicollinearity by providing more data for the regression model.