**PREDICTION MODEL FOR PREDICTING PRICE OF THE TOYOTA COROLLA**

**Prepared By: Tarun Mondal**

**Table of Contents**

[Objective](#_f8xuu84rntyl) 3

[Methodology](#_lawtkz5gttgp) 3

[End-to-end process with solution architecture](#_oxeuszxc3eyv) 3

[Importing libraries in Python](#_6gw5xamkfmch) 3

[Data exploration](#_vrry4e2catd3) 3

[Data preprocessing](#_cjljhxdvpnvd) 5

[Exploratory data analysis (EDA)](#_fn26gc9dt6ak) 6

[Model development](#_45f5uvq36y54) 9

[Time Taken](#_utqrhv1x0iid) 12

[Challenges faced](#_2p2nh574k1j9) 12

[Complexity level](#_1eoqksut98rb) 13

# Objective

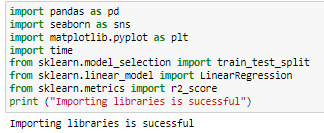
The objective of this study is to predict the price of Toyota cars based on different attributes such as age, power, doors, gears, quarterly\_Tax and weight.

# Methodology

A machine learning approach (multiple linear regression model) has been incorporated in this study to predict the price of Toyota cars. Multiple linear regression models (70-30% splitting and 80-20% split) have been performed using Python in Jupyter Notebook.

# End-to-end process with solution architecture

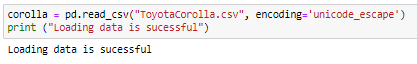
## Importing libraries in Python

****

**Figure 1: Python libraries**

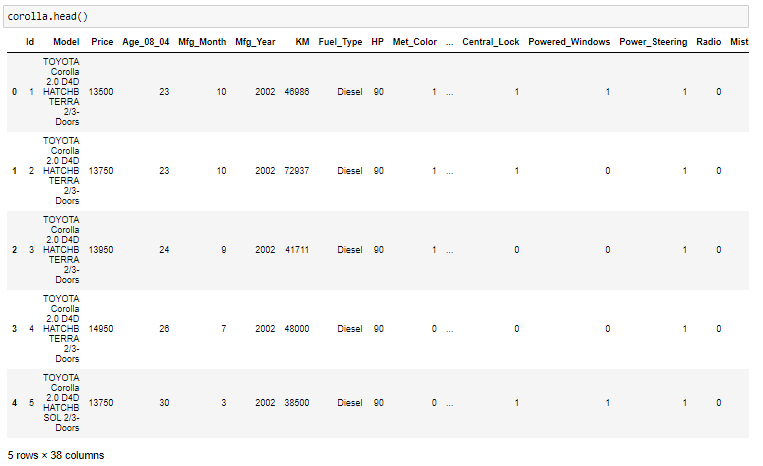
Pandas library has been imported into Python for data loading and manipulation, whereas Seaborn and Matplotlib libraries have been used for data visualisations. For developing machine learning model (Multiple Linear Regression), ‘LinearRegression’ module has been imported from scikit-learn framework, whereas the evaluation metric (R2-score) has been imported from ‘sklearn.metrics module (**Refer to Figure 1**). For performing train-test splitting, from sklearn.model\_selection module, train\_test\_split has been imported into the Python environment.

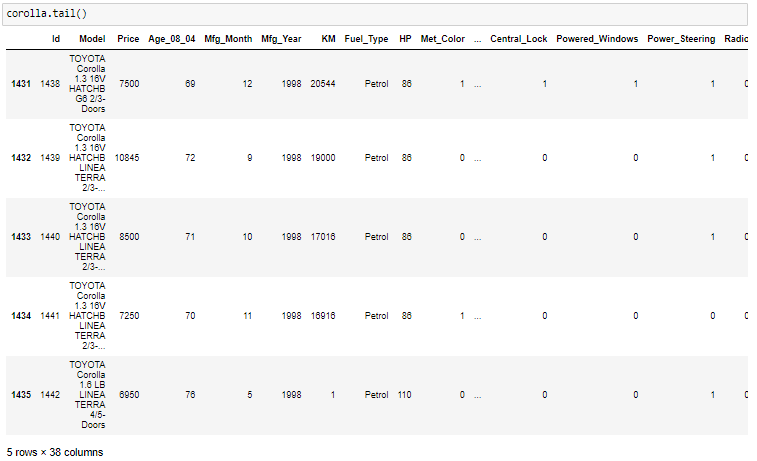
## Data exploration

****

**Figure 2: Data Loading**

Through the use of the ‘read’ function from the ‘Pandas’ library, the ‘ToyotaCoolla’ dataset has been loaded.

****

****

**Figure 3: Head and tail of the dataset**

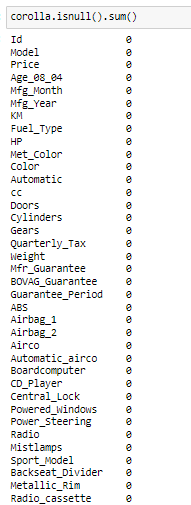
The first few rows and last few rows of the dataset have been shown using ‘head ()’ and ‘tail ()’ functions respectively from pandas library, from which it can be observed that the dataset contains 38 variables.

****

**Figure 4: Shape of the dataset**

Shape of the dataset has been evaluated using the ‘shape’ method in Python, from which it can be observed that the dataset contains 1436 observations and 38 variables (**Refer to Figure 4**).

## Data preprocessing

****

**Figure 5: Null values in the dataset**

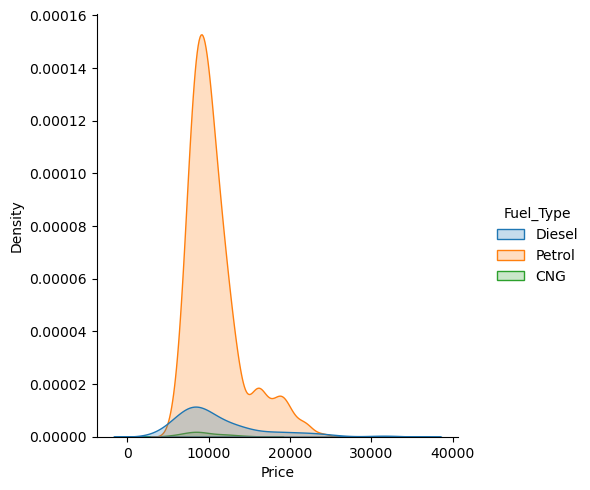
Missing values in the dataset have been checked using ‘isnull ().sum ()’ function, from which it can be observed that the dataset contains 0 missing values for both variables (**Refer to Figure 5**). This indicates that data errors are absent in the dataset, thus, data cleaning steps like dropping null values or filling the null values are not required in this project.

****

**Figure 6: Duplicate values in the dataset**

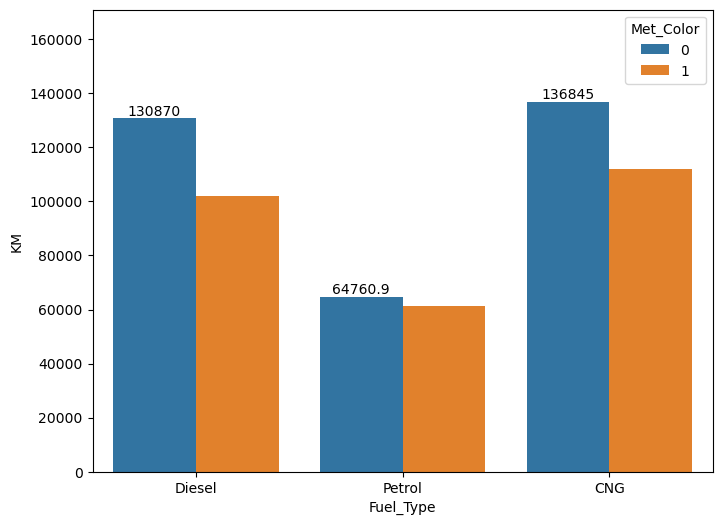
Duplicate values in the dataset have been checked using the ‘duplicated ().sum()’ function, from which total observed duplicate values are 0, indicating the non-existence of repetitive observations in the dataset (**Refer to Figure 6**).

## Exploratory data analysis (EDA)

****

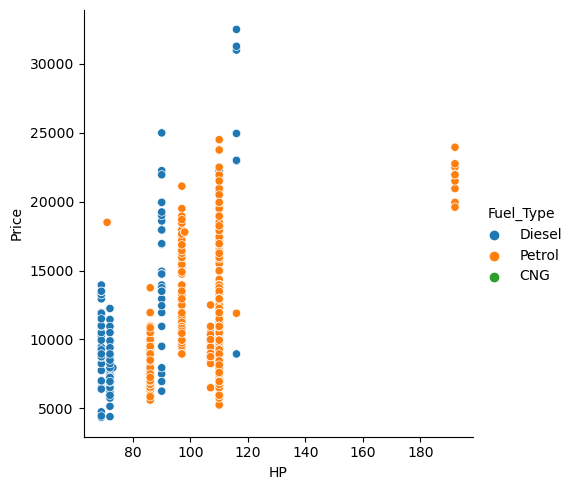
**Figure 7: Distribution of Price for different fuel types**

The density plot shows the price distribution for different fuel types. Petrol vehicles have the highest density at lower prices, while diesel and CNG vehicles have a more spread-out distribution across various price ranges.

****

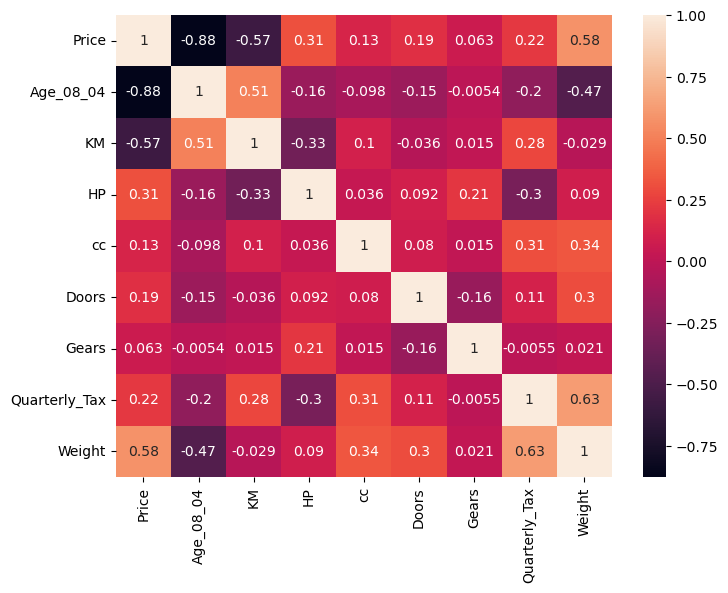
**Figure 8: Price of cards for different fuel types**

The bar chart shows that diesel vehicles have the highest average kilometers driven (130,870 km), followed by CNG vehicles (136,845 km). Petrol vehicles have the lowest average kilometers driven, around 64,760.9 km.

****

**Figure 9: Scatter plot between Horse Power (HP) and Price**

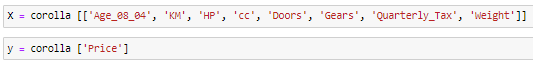
The scatter plot shows the relationship between horsepower (HP) and price for different fuel types. Petrol vehicles dominate higher HP ranges, while diesel and CNG are more concentrated in lower HP ranges with varied prices.

****

**Figure 10: Correlation heatmap**

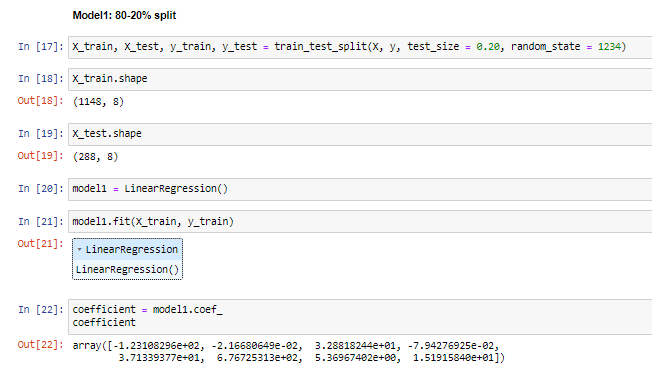
The correlation heatmap shows the association between the features, from which it can be observed that variables like age (r = 0.88) and weight (r = 0.58) have a strong positive correlation with Price. On the other hand, features like mileage (KM) have a negative association (r = -0.57) with price (**Refer to Figure 10**).

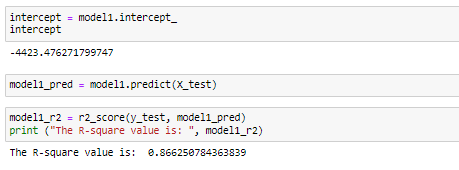
## Model development

****

**Figure 11: Data splitting**

The target variable in this study is ‘Price’, while the features are Age\_08\_04, KM, HP, cc, Doors, Gears, Quarterly\_Tax, and Weight (**Refer to Figure 11**).

****

****

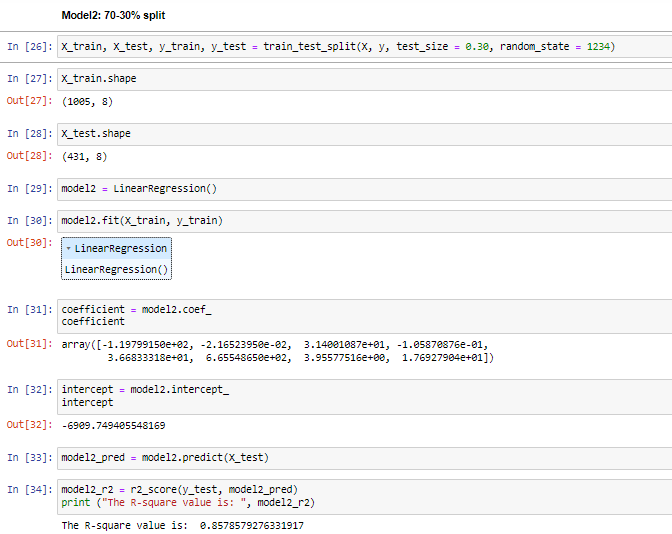
**Figure 12: Multiple Linear regression Model (80-20% split)**

From the multiple linear regression model (with 80-20% split), the obtained coefficients are respectively Age\_08\_04 (-123.10), KM (-0.021), HP (32.88), cc (-0.079), Doors (37.13), Gears (676.72), Quarterly\_Tax (5.36), and Weight (15.19). The intercept obtained from the model is -4423.47.

The equation for Price is:

**Price = -4423.47 + Age\_08\_04 \* (-123.10) + KM \* (-0.021) + HP \* (32.88) + cc \* (-0.079) + Doors \* (37.13) + Gears \* (676.72) + Quarterly\_Tax \* (5.36) + Weight \* (15.19) …. (Equation 1)**

The obtained R-square value of the multiple linear regression model is 0.8662, indicating the model has the capability to explain 86.62% of the variability in Profit (**Refer to Figure 12**).

****

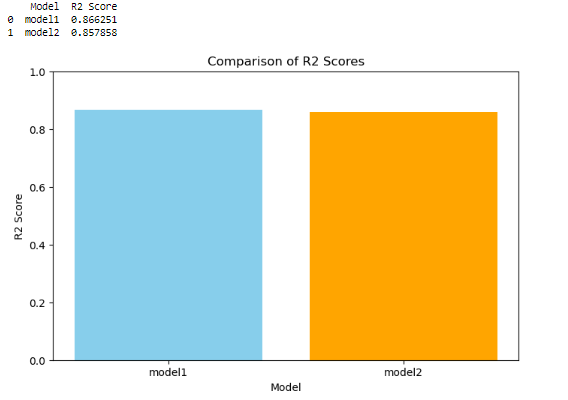
**Figure 13: Multiple linear regression model (with 70-30% split)**

From the multiple linear regression model (with 80-20% split), the obtained coefficients are respectively Age\_08\_04 (-119.19), KM (-0.021), HP (31.40), cc (-0.010), Doors (36.68), Gears (665.54), Quarterly\_Tax (3.95), and Weight (17.692). The intercept obtained from the model is -6909.749.

The equation for Price is:

**Price = -6909.749 + Age\_08\_04 \* (-119.19) + KM \* (-0.021) + HP \* (31.40) + cc \* (-0.010) + Doors \* (36.68) + Gears \* (665.54) + Quarterly\_Tax \* (3.95) + Weight \* (17.692) …. (Equation 2)**

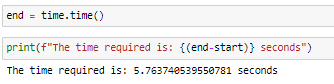
The obtained R-square value of the multiple linear regression model is 0.8578, indicating the model has the capability to explain 85.78% of the variability in Profit (**Refer to Figure 13**).

****

**Figure 14: Model comparison**

The obtained R-square value of a model with an 80-20% split is slightly higher than the model with a 70-30% split (**Refer to Figure 14**).

# Time Taken

****

The total time required for execution of this project is 5.76 seconds, indicating the high speed of the system in estimating delivery time.

# Challenges faced

* Identifying suitable features that can effectively estimate the price of Toyota cars for different models.
* This project faced challenges in identifying the appropriate data-splitting ratio.

# Complexity level

The dataset is very small with no data errors (such as missing values, duplicate values and outliers), reflecting the project is very simple.