**Data Science Project Report on Automobile Price Prediction**

**Project Team ID:** PTID-CDS-AUG-23-1607B

**Project ID:** PRCP-1017-AutoPricePred

**BUSINESS CASE: BASED ON THE GIVEN FEATURE OF DATASET WE NEED TO PREDICT THE AUTOMOBILE PRICE**

**Abstract:**

This project aims to develop a robust automobile price prediction model using machine learning techniques. The primary objective is to create a tool that can accurately estimate the market value of automobiles based on various features and specifications.

Methodology involves data preprocessing, exploratory data analysis (EDA), feature engineering, and the application of machine learning algorithms. We used regression models, including linear regression, decision trees, to build and evaluate our price prediction models.

**Device Project In-to Multiple Steps:**

1. Data Collection
2. 2.Loading data
3. 3.Domain Analysis
4. 4.Basic Checks of data
5. EDA (Univariate, Bivariate, Multivariate Analysis)
6. Data Pre-processing
7. Feature Selection
8. Building ML Model
9. Training & Model Evaluation
10. Model Savings

**Data Collection:**

The dataset received from the Datamites educational platform and provided in a spreadsheet format, specifically as a CSV File.The CSV file contains the necessary fields and columns related to automobiles, including features such as make, Engine type, fuel type, and other relevant specifications.

**Loading data:**

load data in python using pandas library

**DOMAIN ANALYSIS**

Understanding the meaning of each feature and check the relationship of independent feature to target feature.

**EDA (Univariate, Bivariate, Multivariate Analysis)**

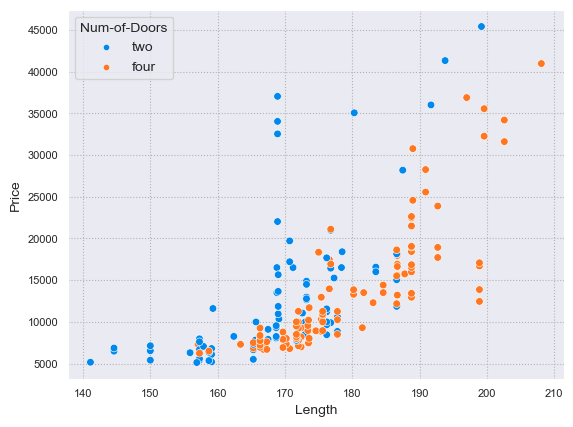
1. **Univariate Data Analysis**

Use sweetviz library and generate a html report of all feature to do univariate analysis, In that we get the Minimum, Maximum, Some statistical information of the particular feature.

1. **Bivariate Data Analysis**

In Bivariate analysis we check the relation of independent feature with respect to target variable

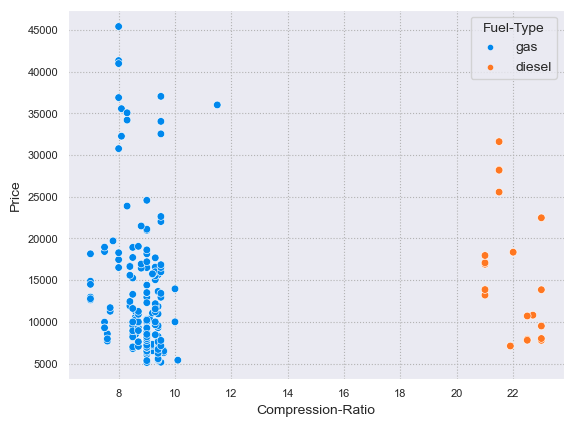
1. **Multivariate Data Analysis¶**
2. **Relation Between length & Price with respect to Num-of-Doors**



**Observation/Insights**

• Maximum number of vehicles have four doors and its length is more and price is high as compared to vehicles which have two doors and its length is less and price is also less.

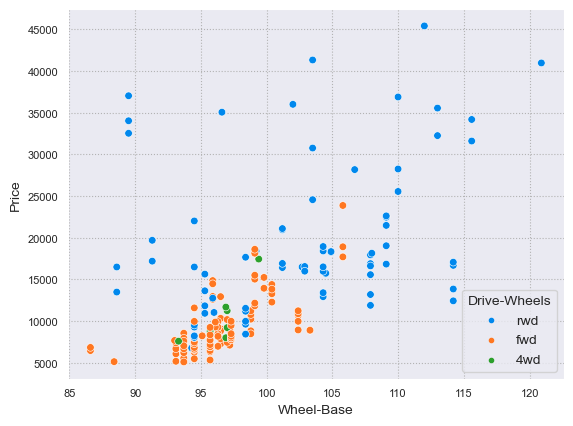
1. **Relation Between Compression Ratio & Price with respect to Fuel-type**



**Observation/Insights**

• If the Fuel Type is Gas there comparession ratio is low and most of the time price is low

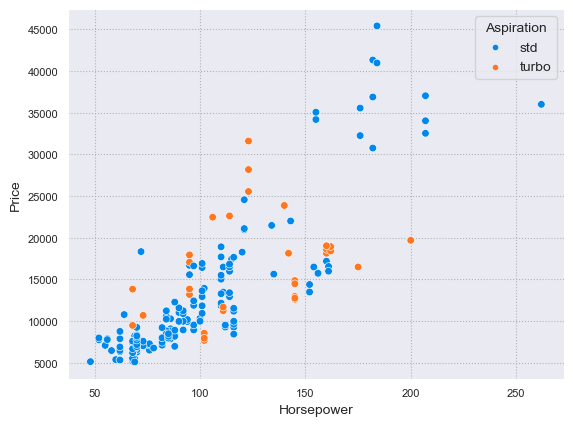
1. **Relation Between Wheel Base & Price with respect to Drive-Wheels**



**Observation/Insights**

• If the wheel base is high at that time drive wheels with RWD are high price.

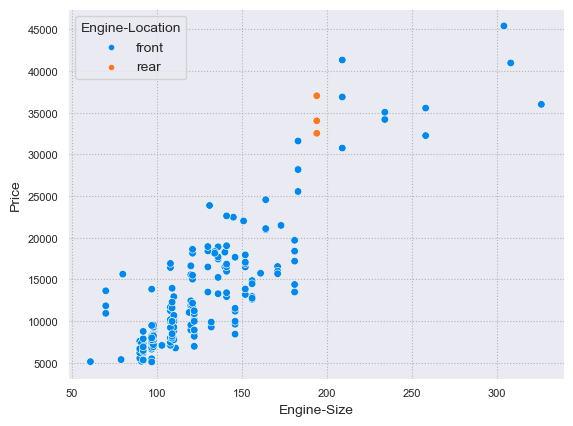
1. **Relation Between Horsepower & Price with respect to Aspiration**



**Observation/Insights**

• Maximum number of vehicles have std aspiration and its horsepower is incresing and price is high as compared to vehicles which have turbo aspiration.

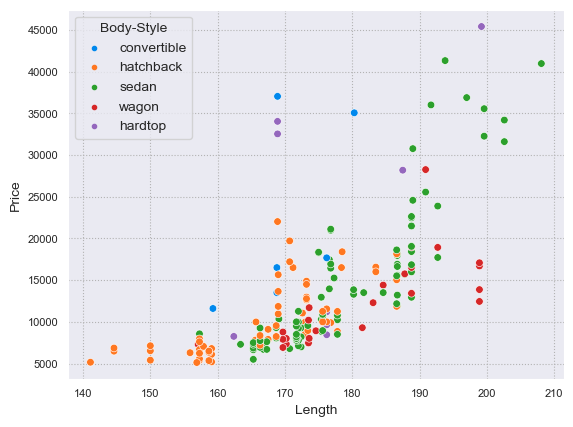
1. **Relation Between Engine-Size & Price with respect to Engine-Location**



**Observation/Insights**

• Maximum number of vehicles have front engine location and its engine size is more and price is high as compared to vehicles which have rear engine location.

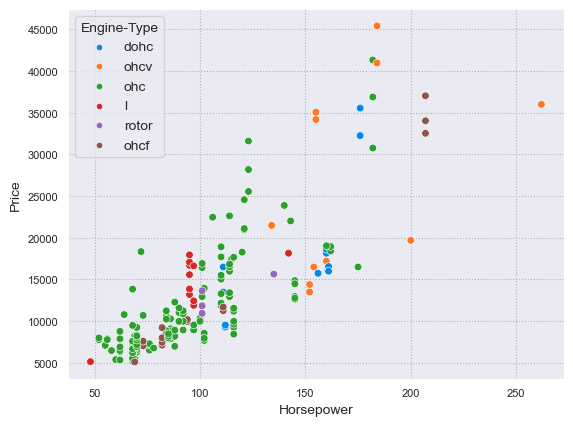
1. **Relation Between Length & Price with respect to Body-Style**



**Observation/Insights**

• In this graph there are more "sedan" body styles and the length of the vehicles of that body style is also more and accordindly the price is also increasing

1. **Relation Between HorsePower & Price with respect to Engine-Type**



**Observation/Insights**

• If the horsepower is low at that time engine type is ohc most of the time as well as car price also less.

**Data Preprocessing**

* First we check the missing values, and we seen that the 5-6 feature has contain missing value and impute them with median and mode.
* Handle categorical data and use Manual encoding and frequency encoding. Because features has contain lots of label.
* In this data I’m Clearly seen that some feature has lots of outlier & we impute them, for that first we check the distribution of all feature and plot the box plot and decide the technique.
* Scale the numerical independent feature with the help of standard scalar and scale the feature. Because standard scaling give me the best result of ML model.

**Feature Scaling**

* Check the correlation with the help of heatmap and seen the two feature has highly correlated with each other, and we decide to drop one feature i.e City-mpg
* The dataset not contain any duplicates

**Model Creation & Evaluation**

* Define Independent and dependant variable and split the data into training and testing.
* For Training 80% & Testing 20% data
* Use linear regression ,decision tree XGBoost regression algorithm to get a best result and XGB regression is give best result. i.e R2 Score 92.28% and Adjusted R2 Score is 80.72% with 2424 RMSE

**Model Saving**

* Save the model using pickle file