# **Automobile Dataset**

Automobile dataset is taken from kaggle.

Datast link:

https://github.com/Sakil786/Data\_Science\_Online\_Course/blob/master/Data\_Science\_Online\_Course/Automobile\_data.csv

### **Context:**

This dataset consist of data From 1985 Ward's Automotive Yearbook. Here are the sources

Sources:

1. 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
2. Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038
3. Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

### **Content**

This data set consists of three types of entities: (a) the specification of an auto in terms of various characteristics, (b) its assigned insurance risk rating, (c) its normalized losses in use as compared to other cars. The second rating corresponds to the degree to which the auto is more risky than its price indicates. Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling". A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

The third factor is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc…), and represents the average loss per car per year.

Note: Several of the attributes in the database could be used as a "class" attribute.

Tasks:

1. Perform EDA on the dataset:

* Finding number of datapoints and number of features in the dataset.
* Replacing special characters(?) to NaN
* Getting missing values
* Finding numerical features and categorical features from the dataset
* Finding correlation among the features(You can use heatmap here)
* Finding outliers (You can use different graphs like box plot,etc)
* Find the distribution of the data(whether the data is left skewed,rigth skewed,normal distribution)
* You can use different graphs for visualization to find more insights from the dataset.

1. **Feature Engineering**

* Handle missing values in the dataset
* Finding feature importance
* Select important feature for model development
* Handle outliers if available in the feature
* Handle categorical features

1. **Model development**

## **Objective:** Develop the model to predict price of the Automobile

**Steps to develop the model:**

* This is a regression problem
* Feature ‘price’ is the target variable
* Divide the datset into train and test split by taking 80% as training data and 20% as test data
* Use Linear regression to train the model

1. **Model Evaluation**

* Use metric for the evaluation of the model(You can use MSE,RMSE,R square)
* Finally save the model in pickle format in your local system

**Note: You can refer the following notebook for the reference:**

**<https://www.kaggle.com/code/ganeshlokare/automobile-price-prediction>**