**\*DISCLAIMER**

    -This project is a sole Intellectual property of mine i.e. **GANDHARV GUPTA** .Kindly acknowledge before copy or use of this project. The dataset is    open and is a property of Kaggle which holds all proprietary rights over it

**Data Analysis and Machine Learning-Regression**

**for**

**Price Determination**

DATASET-Car prices data set

**INDEPENDENT VARIABLES:**

* -**Brand**: Specifies the brand of car
* -**Model**: The specific car model of a brand
* -**Year**: Year of manufacturing(The older the cheaper)
* -**Engine size** :Capacity of engine(high capacity implies costlier car)
* -**Fuel type**: Petrol, Diesel ,Electric, Hybrid
* -**Transmission**: Manual, Automatic, Semi-Automatic
* -**Mileage**: Total distance travelled by car till date(higher mileage implies cheaper price)
* -**Doors**: No of doors(2,3,4)
* -**Owner** **count**: Number of Pre-owners of car(higher count implies cheaper car)

**DEPENDENT VARIABLE/TARGET VARIABLE:**

* -**PRICE**

**Problem Statement:**

* -Predicting price of a car based on a model trained over LABELLED data

**Model Selection:**

* -The problem already has labelled data and requires prediction of a **CONTINOUS** **NUMERICAL** **Variable**,Hence,We will make a **SUPERVISED** **LEARNING** based **REGRESSION** **MODEL**

**Data Source:**

* -A **Kaggle** based **UNCLEEANED** Car-Price dataset is pre-loaded in **MYSQL** **database**
* -File Type: A FLAT FILE(.**CSV**)

**Libraries :**

* Pandas
* Numpy
* Mysql.connector
* Matplotlib
* Seaborn
* Scipy.stats
* Sklearn.preprocessing.StandardScaler
* Sklearn.linear\_model.LinearRegression
* Sklearn.metrics.mean\_squared\_error,r2\_score
* Sklearn.model\_selection.train\_test\_split
* Sklearn.model\_selection.cross\_val\_score,cross,val,predict
* Sklearn.linear\_model.Ridge
* Warnings

**Procedure:**

* Data Collection
  + Connecting to MYSQL data base with mysql.connector
  + Running queries to fetch requisite data
  + Convert the fetched data into pandas.Dataframe
* Data Wrangling, Data cleaning, Data Preprocessing, Data Transformation:
  + Handling Missing Values
  + Handling Duplicate and Outlying Values
  + Data Standardization
  + Data normalization and scaling using z-score
  + Data Binning
  + One-Hot-Encoding: Converting categorical to numerical variable
* Exploratory Data Analysis(EDA):
  + Continuous numerical variables:
    - Finding correlation coefficients
    - Finding p-values
    - Creating correlation matrix
    - Plotting Heatmap of correlation
    - Plotting Regplot
  + Categorical variables:
    - Plotting Box plots
    - Plotting Bar plots
* Model Building and Predicting:
  + SLR(Simple Linear Regression)
  + MLR(Multiple Linear Regression)
  + Polynomial Regression
* In-Sample Evaluation:
  + MSE(Mean squared error)
  + R2(R-squared)
  + Regression plot for SLR
  + Residual Plot
  + Distribution plot for MLR
* Model Testing(using test-data)
  + Train\_test\_split
* Out-Sample Evaluation(Using test-data)
  + MSE
  + R2
  + Regression plot
  + Residual Plot
  + Distribution plots
* Model Refinement:
  + Ridge regression
  + Hyper-Parameter Tuning(Alpha)
  + Cross validation with
    - Cross\_val\_score
    - Cross\_val\_predict
* Making Final Predictions by taking input from User and translating results

**Final Remarks and Conclusions:**

* -Given the data source, Supervised Learning with Regression was adopted to predict a Continuous Numerical Value.
* -Among the 3 Models built(SLR MLR and Polynomial),SLR performed best in Evaluation Metrics Like MSE and R2.
* -The Regression plots and Residual Plots of SLR show a STRONG POSITIVE LINEAR RELATIONSHIP among Year and Price.
* -The Final Predictions also testify the same trend.
* -The newer the car, the costlier it gets.
* -The model estimates not only the relationship among Year and Price, but also the STRENGTH of such a relationship that exists.

**\*DISCLAIMER**

    -This project is a sole Intellectual property of mine i.e. **GANDHARV GUPTA** .Kindly acknowledge before copy or use of this project. The dataset is    open and is a property of Kaggle which holds all proprietary rights over it

END.