Problem Statement

A Chinese automobile company, **Geely Auto**, aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an **automobile consulting company** to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting car pricing in the American market, as they may differ from the Chinese market.

The company wants to know the following things:

- Which variables are significant in predicting the price of a car?
- How well do those variables describe the price of a car?

Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the American market.

Business Goals

You are required to model the price of cars with the available independent variables. The management will use this model to understand exactly how the prices vary with the independent variables. Accordingly, they can change the design of the cars, the business strategy, etc., to meet certain price levels. Further, the model will allow the management to understand the pricing dynamics of a new market.

Data Preparation

There is a variable named **CarName** that comprises two parts: the first word is the name of the car company, and the second is the car model. For example, **Chevrolet Impala** has ‘Chevrolet’ as the car company name and ‘Impala’ as the car model name. You need to consider only the company name as the independent variable for model building.

Model Evaluation

When you are done with model building and residual analysis and have made predictions on the test set, ensure that you need to calculate the R2-score of the model.

Evaluation Rubrics

Criteria	**Meets Expectations**	**Does Not Meet Expectations**
Data understanding, preparation and EDA (40%)	All data quality checks are performed, and all data quality issues are addressed in the right way (missing value imputation, removing duplicate data and other kinds of data redundancies, etc.). Explanations for data quality issues are clearly provided in the comments. Categorical variables are handled appropriately.	All quality checks are not done, data quality issues are not addressed correctly to an appropriate level. Categorical variables are not handled appropriately where required. Dummy variables are not created properly. New metrics are not derived or are not used for analysis. The data is not converted to a clean format that is suitable for analysis or is not cleaned using commands.

	Dummy variables are created properly wherever applicable. New metrics are derived, if applicable, and are used for analysis and modelling. The data is converted to a clean format suitable for analysis.	
Model Building and Evaluation (60%)	Model parameters are tuned using correct principles and the approach is explained clearly. Both the technical and business aspects are considered while building the model. Correct variable selection techniques are used. A reasonable	Parameters are not tuned enough or tuned incorrectly. Relevant business aspects are not considered while model building. Variable selection techniques are used incorrectly/not conducted. A variety of models are not considered, or a sub-optimal one is finalised.

	number of different models are attempted, and the best one is chosen based on key performance metrics. Model evaluation is done using the correct principles, and appropriate evaluation metrics are chosen. The results are on par with the best possible model on the data set. The model is interpreted and explained correctly. The commented code includes a brief explanation of the important variables and the model in simple terms.	The evaluation process deviates from the correct model selection principles; inappropriate metrics are evaluated or are incorrectly evaluated. The results are not on par with the best possible model on the dataset. The model is not interpreted and explained correctly.