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 a car depends. Specifically, they want to understand the factors affecting the pricing of cars in the American marketing, since those may be very different from the Chinese market. Essentially, the company wants to know:

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

Based on various market surveys, the consulting firm has gathered a dataset of different types of cars across the Americal market.

Goal of this assignment

You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Data Preparation Hint

- There is a variable named **CarName** which is comprised of two parts - the first word is the name of '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 company name as the independent variable for the model building.

- Create new variables from given variables (like in mutate function) if new variables seem more practical and useful for analysis and modelling.

Criteria To Meet expectations

All data quality checks are performed, and all data quality issues are addressed in the right way (missing values, removing duplicate data and other kinds of data redundancies, etc.). Explanations for data quality issues are clearly mentioned in comments.

Dummy variables are created properly wherever applicable.

The data is converted to a clean format suitable for analysis in R.

Model parameters are tuned using correct principles and the approach is explained clearly. Both technical and business aspects are considered while building the model.

Correct variable selection techniques are used. A reasonable 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 at par with the best possible model on the dataset.

The model is interpreted and explained correctly. The commented code includes a brief explanation of the important variables and the model in simple terms.

Appropriate comments are written wherever applicable.

If new variables are created, the names are descriptive and unambiguous.

The code is written concisely wherever possible.

Overall, code readability is good with appropriate indentations.