# Scania AB’s US Car Business

## 1. Business Context

Scania AB is a major Swedish manufacturer focusing on automobiles. It also manufactures diesel engines for heavy vehicles as well as marine and general industrial applications. The company is planning to enter the American Market by setting up their manufacturing unit there.

## 2. Problem Statement

To identify the factors that influence the price of cars in America, based on which they can manipulate the design of their cars and can take a good start in the business.

### Objective of Analysis :

1. Variables that affect the price of cars in America
2. Analyze the impact of each feature on the price
3. Suggest possible requirements for Scania AB to work on

Dependent Variable : Price

## 3. Solution Developed

Steps followed :-

1. Data Cleaning : This section involves checking for inconsistencies in values that could hinder the efficiency of the model.
    - Handle Missing Values
    - Convert to relevant Data Types
    - Clean format for all columns
    - Split CarName to get Company Name
    - Delete CarName column (has too many unique values and less likely to contribute to the analysis)
    - Convert all values to lowercase since python is case-sensitive.

2. EDA Analysis : It's time to analyze the cleaned dataset to see if we can identify any basic patterns between different variables.
   Steps taken are :
   - Checking Correlation amongst the numerical variables
   - Analyzing the categorical variables
   - Deriving variables
3. Data Preparation : The point of this section is to convert all object values to numerical. This is necessary before moving on to modelling since statsmodel and sklearn only accept numerical values.
   2 possible ways of dealing with the categorical variables is to :
   - Aspiration & Door Number: convert to 0 & 1 values in a single column (no need for dummy variables)
   - Rest of the categorical variables: these could be converted to dummy variables since they don't have that many levels plus they are not ordinal values.
4. Splitting into train and test set : Now we are ready to start building a model, our dataset is all ready. Before moving to that step we divide the dataset into training and test dataset. Considering the fact that there are only 205 entries, they could be divided in the ratio of 90:10.
5. Feature Scaling : This step is to basically normalize all the numerical variables, which helps speed up the process of calculation in modelling.
6. Model : 2 different approaches, with and without car companies are taken since in Model 1, the final features are dominated by car company names. While this gives clear indication that Car Company is important, Scania AB, initially, may not be able to have control over their Brand image in America. Therefore, we need to identify other possible features as well.
   - RFE: to bring down the number of feature to a more reasonable quantity
   - Manualing dropping features
     1. Deleting columns by looking for features with a high p-value (>0.05)
     2. Looking at the VIF (Variance Inflation Factor) to looking for highly correlated features (deleting ones with VIF>5)
   - Metrics : R-Squared, Adj. R-Squared, F-statistic, AIC
7. Residual Analysis
8. Predicting the model
9. Model Evaluation

## 4. Suggestions

Final solution for Scania AB Business plan: One of the factors affecting the price is company name. Brand holds a certain value that influences the price. While this is out of Scania AB's control, they have other features that they can control.

- **Engine Size**: this is one of the most important features in determining the price. Greater, the better
- **Engine Type**: avoid OHCV, one better option could be OHCF.
- **Fuel System**: IDI is a good suggestion. It does not affect the price much though.
- **No.of Cylinders**: avoid 12, choose either 2 or 5. Although looking at the 2 models it seems that 2 has more impact on the price.
- **Peak RPM**: both the models show that it has some significant on the price of a car
- **Car length**: along with this, during the analysis it was evident that curb weight, horsepower & car width are also important features. Although they were highly correlated to the other features, it is important to note the significance of these features as well.

## 5. Improvements

While our model has given a good insight on the factors that affect the price of cars in the USA, we can always improvise on the model by trying out different regression models such as Linear Regression. Additionally it is clear that the existing features were more than useful, it could also be enhanced further by adding more derived variables.

## 6. Link to the prototype