**Problem Statement**

**Company Information**

An automobile consultancy firm “*Mycar* *Dream*” provides assistance to its clients in making appropriate car deals, based on their requirements.

Based on various market surveys, the firm has gathered a large dataset of different types of cars and their attributes across the world. The business model of the company is solely based on consumer interest, aiming to provide the most appropriate car to their clients and hence maximise the customer satisfaction.

**Problem Statement:**

Nowadays, the automobile market has become very dynamic as the buyers have varied preferences. Customers look for various features (brand value, mileage, model\_year etc) in their dream car. In order to fulfil it's customer requirement, *Mycar* *Dream* wants to automate the process of predicting the car mileage which fits the customer preferences, based on the dataset of car features and attributes obtained by market surveys.

The data set contains the following details about cars:

The aim here is to predict the city-cycle fuel consumption in miles per gallon, in terms of 3 multivalued discrete and 5 continuous attributes. Below is the data dictionary to be used for the dataset.

|  |  |  |
| --- | --- | --- |
| **Data Dictionary** | | |
| **Sl.No.** | **Variables** | **Description** |
| 1 | Mpg | Mileage per gallon (continuous variable) |
| 2 | Cylinders | Number of cylinders in car (multi-valued discrete) |
| 3 | Displacement | Volume of fuel inside the engine i.e size of engine (continuous ) |
| 4 | Horsepower | picks up of the car (continuous) |
| 5 | Weight | Weight of car (continuous) |
| 6 | Acceleration | Acceleration of car (continuous) |
| 7 | Model year | Year when the car launched (multi-valued discrete) |
| 8 | Origin | Origin of car (multi-valued discrete) |
| 9 | Car name | Name of car company (unique for each instance) |

**How to start the assignment?**

As we have learnt in the linear regression lectures, there is a 5-step process. You ideally need to follow while solving any regression problem. The chart below will guide you through the process of solving the assignment.

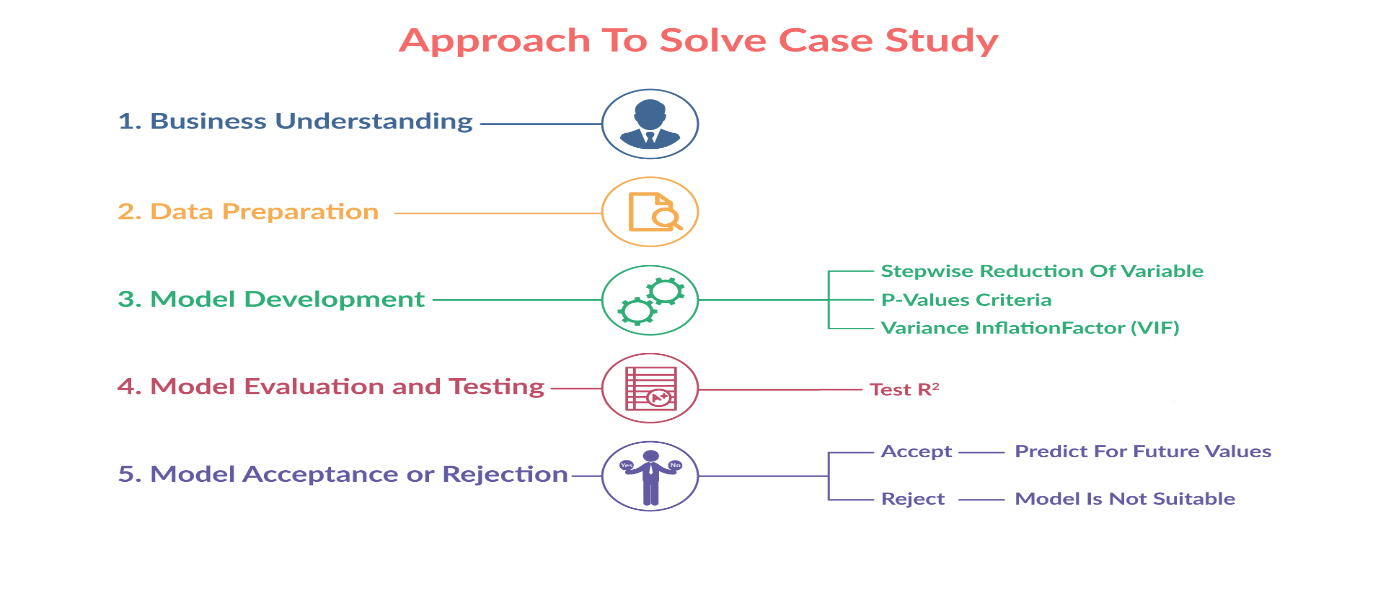


Figure 1: Linear Regression flow chart

**The goal of this assignment ?**

You are required to develop a predictive model which can follow these three constraints thoroughly:

* The model should not contain more than **5 variables.**
* According to the business needs, set the **VIF to 2.**
* The model should be highly predictive in nature i.e it should show **80% (R squared)** of accuracy.

**Downloads**:

You can download the dataset file from the link given below:

[Car Mileage Dataset](https://cdn.upgrad.com/UpGrad/temp/33991f5e-1edb-4e83-a7cf-5165fd417a30/carMPG.csv" \o "carMPG.csv)

*[file\_download](https://cdn.upgrad.com/UpGrad/temp/33991f5e-1edb-4e83-a7cf-5165fd417a30/carMPG.csv" \o "carMPG.csv)*[Download](https://cdn.upgrad.com/UpGrad/temp/33991f5e-1edb-4e83-a7cf-5165fd417a30/carMPG.csv" \o "carMPG.csv)

**Note**:

Please make sure the below points are to be followed strictly for evaluation purpose:

* Store the dataset into “*carmileage*” object.
* Divide the dataset into **70:30** ratio and set seed to 100 for the reference. It should be renamed as “train” and “test” respectively.

**Packages Required:**

You have to install the below packages below starting this assignment.

* install.packages(“MASS”) for StepAIC
* install.packages(“car”) for VIF

**Checkpoints**

**Checkpoint 1: Business Understanding and Data Understanding**

This is the most important stage of solving any problem. In this checkpoint, you are supposed to get well versed with the dataset. For example, you should know the aim of solving this assignment and should also be attentive to explore your data completely.

**NOTE**:- Keep writing every step of commented R codes properly in a new .R file which you have to submit at the end of this assignment for the evaluation purpose.

**Checkpoint 2: Data Cleaning and Preparation**

After Business understanding, the next important step is data preparation.  Data preparation takes the highest amount of time and effort. It is estimated that 60-80% of the time in any model building process is typically spent in preparing and cleaning data.

In our assignment, we have further divided Data preparation into three stages as shown in **figure-1**:

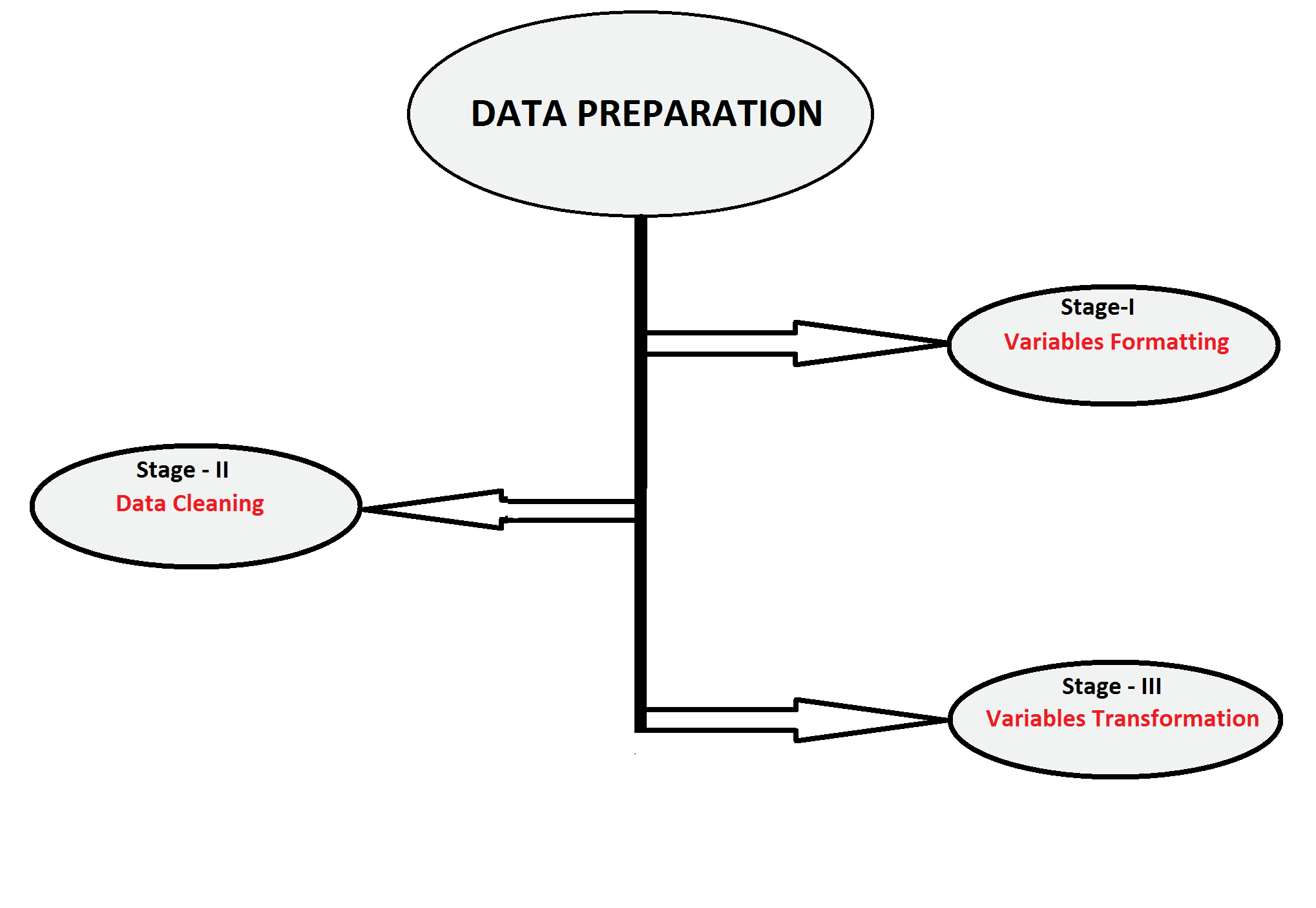


Figure1: Data Preparation Stages

**First stage**: Variables Formatting

* Check the structure of “*carmileage*” dataset and change the type of variable as per the business understanding if required.

**Second Stage**: Data Cleaning

* Clean the data set for any anomalies it might have. Treat the missing values (if any) and the outliers (if any) at this stage.

**Third Stage**: Variables Transformation

* Adding more meaningful variables in the modeling process. These could be transformed variables. If done correctly, these variables can add tremendous power to your analysis.

 (**Hint:** “Car\_Name” variable level should be reduced to under 30 - 35 dummy variables)

**NOTE**:- Don't forget to divide your dataset into the train and test datasets just after data preparation. Keep writing every step of commented R codes properly in a new .R file which you have to submit at the end of the case study for the evaluation purpose.

**Checkpoint 3: Model Development**

Once the data preparation is over. You should move on to the model development phase. In this checkpoint, you are supposed to consider the below points while developing the model.

* Multicollinearity
* p-value
* R square metric

**NOTE**:- Keep writing every step of commented R codes properly in a new .R file which you have to submit at the end of the assignment for the evaluation purpose.

**Checkpoint 4: Model Evaluation and Testing**

Once the model is developed, apply your model to the test dataset. You should calculate the test R-squared to know whether your model has good predictive ability or not.

**Hint**: (correlation between actual and predicted values of test dataset)

**NOTE**:- Keep writing every step of commented R codes properly in a new **.R file** which you have to submit at the end of this assignment for the evaluation purposes.

**Checkpoint 5: Model acceptance or Rejection**

In this checkpoint you should make sure that all the goals of this assignment are met, which are:

* The model should not contain more than **5 variables.**
* The model should be highly predictive in nature i.e it should show **80% (R squared)** of accuracy.
* The model should give **high accuracy (test R-squared )** when tested it on the test dataset.