Objective:

Data preprocessing is an essential in any machine learning process, apart from the learning algorithm. It is critical for us as data scientists to understand data, clean it and bring it in a format with which we can work. It is also important for us to realize that there no particular method is guaranteed to work and hence, this activity is designed to achieve the same.

Key takeaways:

- Given a dataset and its description, understand the objective.
- Read data which can be in any format
- Data preprocessing steps
    - Handling missing values row & column wise
    - How to handle factor variables? What if some levels occur only once or twice?
    - How to approach the above scenario
    - Train, Validation & Test datasets? Why three sets?
- Building Linear and Logistic models
- Model diagnostics and Model evaluations
- ***Listen to the data, understand it, perform as many trails as possible and follow the evidences.***

Problem 1

There are two files given. ‘Automobiles’ is the dataset. ‘Automobiles_Description’ is the description of the dataset and the problem statement. Our objective is to predict ‘price’ of the automobile based on other variables.

Problem 2

There are two files given. ‘horse-colic_data’ is the dataset. ‘horse-colic_names’ is the description of the dataset and the problem statement. Our objective is to classify ‘surgical-lesion’ for the horse based on other variables. ‘horse-colic’ is the final test dataset.