Objective:
In this session, you will learn to build multivariate linear regression model, to validate your model and interpret the results, and the evaluation metrics. We will also look at how to go about building a model and how to look for evidences which can guide us in building a better model.

Key takeaways:

- Building multivariate linear predictive model using lm()
- Power of variable transformation.
- Interpreting the diagnostic plots to check for the linear regression assumptions
- Identifying influential observations and handling them.
- Checking for multicollinearity through VIF.
- Check if R square is the only reliable measure in evaluating model performance
- Check for multiple metrics like VIF, R square, p values, F statistic, Error metrics (RMSE/MAPE). What does each one of them mean and how to use them in model building.
- Data Pre-Processing
    - Type conversions
    - Standardization- Effect of scaling the data
    - Train-Test splitting of the data

Problem 1: *Effect of variable transformation on model performance*

- The dataset 'credit.csv' into R.
- The task is to predict the variable 'Score' given other attributes.
- First build a naïve linear regression model using all the attributes as it is. Check the performance of this model.
- Perform log transformation of some of the independent variables. Build a linear regression model using these new variable and check for improvement in the performance.

Problem 2: *How to build a multivariate linear regression model*

Linear regression is such a model where there are no hyper-parameters. The model totally depends on how the data is. So better the quality of data, better the performance of the model. Remember, 'Garbage in → Garbage out'.

- Read the 'CustomerData.csv' data into R. Our objective here is to predict 'TotalRevenueGenerated'.
- Perform the preprocessing steps. This includes imputation/binning/ removing undesirable variables etc. Split the data into train/test.
- Start with a naïve model by using all the attributes as it is.
- Perform stepAIC, VIF. Check p values for unimportant variables & diagnostics plots to check for linear regression assumptions.
- Check for the error metrics.
- See how we can fix the evident issues in the data. Can we come up with transformed variables that can improve the performance?