Objective:
In this session, you will learn to build your first simple linear regression model, to validate your model and interpret the results, and the evaluation metrics. You will also observe the diagnostic plots and validate linear regression assumptions.

Key takeaways:

- Building univariate linear predictive model using lm()
- Reading diagnostic plots to test the linear regression assumptions
- Identifying influential observations and handling them
- Data Pre-Processing
    - Binning- Converting a numeric attribute to categorical
    - Dummies- Converting a categorical variable to numeric
    - Standardization- Effect of scaling the data
    - Imputation- Handling the missing values
    - Train-Test splitting of the data
- Using Correlation plots, correlation values to identify appropriate variables for model building

This lab is split into three sessions:

Session 1: Demonstration of linear regression on a toy day and in class activity

- A data and R code for simple linear regression is shared with you. As we demonstrate this problem, please execute the commands and observe the outputs.

Session 2:

Problem Definition:

- A large child education toy company (company name is confidential and data is masked) which sells edutainment tablets and gaming systems both online and in retail stores in the US wanted to analyze the customer data. They are operating from last 15 years and maintaining all transactional information data. The given data 'CustomerData.csv' is a sample of customer level data extracted and processed for the analysis from various set of transactional files. Using this data, they want us to understand the life time value of each customer (LTV). This will enable them to design marketing strategies and customize the product offerings. The objective of activity is building a regression model to predict the customer revenue based on one factor that influences revenue the most.
    - Read the given "CustomerData.csv" data into R.
    - Are there any missing values in the data? If there are, then impute using central imputation method
    - The target for this problem is "Total Revenue generated"
    - Select one most influencing numeric variable as predictor for predicting the revenue generated. How would you do this
    - Build a linear regression model to predict the target with the selected variable
    - Check the diagnostic plots and report your observations

- How do you plan to improve the predictive capability of the model

Session 3: Data Preprocessing- Why Do we need it

- Previously, we have shared the code for basic preprocessing. These are the common steps which you would frequently use on data.
- Understand these processes and answer the following questions
  - Why/when do we need binning
  - What is the difference between equal width and equal frequency binning?
  - Why/when do you think creating dummies of a variable is required?
  - Why Standardization is necessary?
- Train-Test split
  - It is customary, that whenever we get the data for analysis. We use a part of the data for model building and on remaining data we test our model. Why is this train-test split of the data is necessary?
  - On the train dataset you felt the need for standardization before model building. You have standardized the data and built the model. You received the test data later. Now how would you standardize the test data