

# DS3030: Data Analytics Lab

## Assignment 3

Date: Aug 18, 2025

Timing: 2:00 to 5:00 PM

Max mark: - 20

---

### Instructions

- Write separate R scripts to implement each of the parts
  - Submit three .r files containing all answers named as [studentname]\_lab[assignmentnumber]\_part[partnumber].r
  - Write **justifications/comments** as required.
- 

### CEREALS DATASET : cereals.csv

This dataset contains the nutrition data on various cereal products.

<https://www.kaggle.com/datasets/crawford/80-cereals>

## Part I : Fundamentals, Measures Of Central Tendency

1. Load the dataset into a dataframe, print the shape and top 7 rows of the dataset  
(HINT: Use “read.csv”, “dim”, “head” built-in functions.) (2)
2. What is the average calorie value of all the cereals?  
(HINT: Use “mean” function and column “calories”.) (1)
3. Which cereals have the maximum and minimum ratings?  
(HINT: Use the columns “rating”, “name”; Use “max” and “min” built-in functions; Use logical operation “==” for filtering rows.) (2)
4. Which company is the most common (frequent) cereal manufacturer?  
(HINT: Use “function” to define a new function; Use “table” and “names” built-in functions; Use the column “mfr”.) (3)

## Part II

**Total Marks:** 9

5. **Range Calculation (Manual + Built-in)** Create a numeric vector containing the ages of 15 students in a class.

1. Calculate the range manually ( $\max - \min$ ) without using the `range()` function.
2. Verify your answer using the `range()` function in R.

6. **Standard Deviation (Manual + Function)** Create a dataset of daily temperatures for a week:

```
temp <- c(28, 30, 29, 31, 33, 32, 30)
```

- (a) Calculate the standard deviation manually using the formula:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

- (b) Confirm your answer using the `sd()` function in R.

7. **Interquartile Range (IQR)** A shop recorded the daily sales amount (in Rs) for 20 days:

```
sales <- c(1200, 1150, 1180, 1300, 1250, 1400, 1500,  
         1450, 1350, 1380, 1420, 1550, 1600, 1650,  
         1700, 1750, 1800, 1850, 1900, 2000)
```

- (a) Find Q1 (25th percentile) and Q3 (75th percentile) using the `quantile()` function.
- (b) Calculate the IQR manually ( $Q3 - Q1$ ).
- (c) Verify your answer using the `IQR()` function in R.

## Part III

**Total Marks:** 3

8. Using the given cereal dataset:

1. Find the Pearson correlation coefficient between calories and rating.
2. Interpret whether the correlation is positive, negative, or none.
3. Test if the correlation is statistically significant at the 5% level.