CHRIST (Deemed to be University), Bangalore – 560 029
Department of Computer Science
CIA Component 2 Test – JANUARY 2024
PG II Trimester

CLASS: 2MCA-A&B **Max. Marks: 50**
Course Name: APPLIED STATISTICS USING R **Time: 2 Hrs**
Course Code: MCA232

UScerealDataset

The UScereal data set has been collected from the 1993-ASA Statistical Graphics Exposition and is taken from the mandatory F&DA food label. The recorded variables are:

mfr-Manufacturer, represented by its first initial: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.	
calories - number of calories in one portion.	carbo -grams of complex carbohydrates in one portion.
protein - grams of protein in one portion.	sugars -grams of sugars in one portion.
fat - grams of fat in one portion.	shelf-display shelf (1, 2, or 3, counting from the floor).
sodium - milligrams of sodium in one portion.	potassium - grams of potassium.
fibre - grams of dietary fibre in one portion.	vitamins-vitamins and minerals (none, enriched, or 100%).

R1- Understanding of selected Dataset - 5M

Import the given "UScereal" dataset, understand and UScereal. Find the maximum protein value of each Manufacturer. (5M)

R2-Descriptive Analysis - 10M

1. Investigate the data set for missing/NA values. Look at the distribution of the dataset to replace the missing values. (Normal distribution-MEAN, Left Skew-Min, Right Skew-Max). (6M)
2. Get the summary statistics after handling missing data (mean, median, min, max, 1st quartile, 3rd quartile, and standard deviation). (4M)

R3-Exploratory Analysis - 15 M

Using the *ggplot2* package, generate the suitable plots. Explain your findings on generated plots.

1. Analyze the spread of the data set for the Manufacturer to check how each one has given preference for Fiber. (5M)
2. Create a plot to find the outlier on *calories* for each *shelf*. (5M)
3. Create a plot to explore all numeric variables. (5M)

R4-. Model Building 15M

1. Identify the top-four mean variables and create the data frame GreaterMeanFour.(1M)
2. Find the strength of the relationship of GreaterMean and Plot the relationship. (3M)
3. Create a simple linear regression model using strongly positively correlated variables and plot it. (2+5M)
4. Show the prediction for the given value before and after removing outliers. (2+2M)

R5-Conclusion - 5M

Five concluding points about your analysis.