Lab 4 – SECTION A, BATCH 2 Date: 17th Sept. 2022

Exercise 1 – Data Preprocessing, Regression

Using the given **CEREALS** dataset, perform data preprocessing and answer the following questions.

1) Create a table with the 5-number summary of all the numeric attributes.
2) For each of the numeric attributes (proteins up to vitamins), identify and replace all missing data (indicated with -1) with the arithmetic mean of the attribute.
3) Create a table with the 5-number summary of all the numeric attributes after treating missing values. Do you think the strategy used in dealing with missing values was effective?
4) For each of the numeric attributes (proteins up to vitamins), identify and replace all noisy data with the median of attribute.
5) Create a table with the 5-number summary of all the numeric attributes after treating noisy values. Do you think the strategy used in dealing with noisy values was effective?

Use the prepared or preprocessed data to answer the following:

6) Cross tabulate the type of cereal (hot vs cold) against the manufacturer
7) Which is the cereal with the best rating, worst rating?
8) Plot a side-by-side boxplot comparing the consumer rating of hot vs. cold cereals.
9) Is there a relation between sugars, calories, carbs, and fat?
10) Which manufacturers produce cereal with highest calories?
11) Use correlation tests and visualization to identify if the two variables calories and consumer rating associated?
12) Use correlation tests and visualization to identify if the two variables shelf and consumer rating associated?
13) Is there a relation between manufacturer and rating?
14) Which nutrients are essential for a good rating for a cereal?
15) Design a Linear regression model to predict the rating of a cereal based on top 3 related nutrients. Tabulate the accuracy of the model using an 80, 20 split.