DSE 2141– Data Analytics Lab

Lab 2 – Date: 9th August 2023

Exercise 1 – Data Preprocessing

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?

Exercise 2 – Data Preprocessing and Visualization

Using the given **BENGALURU HOUSE PRICES DATASET**, perform data preprocessing and answer the following questions.

Find the missing values in the DataFrame and replace them with the right missing values. However, if a column has more than 15% missing values then drop (or remove) the column from the DataFrame except for the location, size and total_sqft columns because the house prices are directly dependent on these three crucial parameters. In fact, the rate of a house is reported as a cost per unit area of the house.

Drop (or remove) the rows containing the missing values in the location, size, and total_sqft columns. Additionally, drop any row which contains nonsensical values in the context of houses.

1. Compute the total number of missing values in the DataFrame.
2. Compute the percentage of missing values in the DataFrame. If a column has more than 15% missing values then drop (or remove) the column from the DataFrame.
3. Which column has the most number of missing values?
4. drop (or remove) the rows containing the missing values in the *location, size,* and *total_sqft* columns
5. drop all the rows in the bath column containing more than 5 bathrooms.
6. Find the houses available in each area.
7. Find the top five areas where the large number of houses available
8. Visualize the house price based on build up area of top five locations where the large number of houses available (Visualize by independent charts.)