ASSIGNMENT 1 – BASIC STATS – 1

1. Numerical columns in Sales\_data\_with\_discounts dataset are Volume, Price, Total Sales value, Discount Rate, Discount Amount, Net Sales value.
2. The mean of each numerical column:

Volume: 5.06

Avg Price: 10453.4

Total Sales Value: 33812.835556

Discount Rate (%): 5.155242

Discount Amount: 3346.499424

Net Sales Value: 30466.336131

1. The Median of each numerical column:

Volume: 4.000000

Avg Price: 1450.000000

Total Sales Value: 5700.000000

Discount Rate (%): 16.577766

Discount Amount: 988.933733

Net Sales Value: 4677.788059

1. The Standard Deviation in each column:

Volume: 4.231602

Avg Price: 18079.904840

Total Sales Value: 50535.074173

Discount Rate (%): 4.220602

Discount Amount: 4509.902963

Net Sales Value: 46358.656624

1. Using Histogram plot on numerical columns:

Volume: Positively Skewed

Avg Price: Positively Skewed

Total Sales Value: Positively Skewed

Discount Rate (%): Negatively Skewed

Discount Amount: Positively Skewed

Net Sales Value: Positively Skewed

1. Box Plot for Numerical Columns to identify Outliers:

Volume: Positively Skewed: 8

Avg Price: Positively Skewed: 0

Total Sales Value: 1

Discount Rate (%): 0

Discount Amount: 9

Net Sales Value: 1

1. The Categorical Columns in the dataset are Date, Day, SKU, City, BU, Brand, Model.
2. Standardization (also called Z-score normalization) is a method used to scale data so that it has a mean of 0 and a standard deviation of 1. This helps to bring all features of the data onto a common scale, especially when they have different units or ranges.

Volume Avg Price Total Sales Value Discount Rate (%)

0 2.347417 0.091072 2.922469 -0.829365 – Standardized data

Volume Avg Price Total Sales Value Discount Rate (%)

15 12100 181500 11.654820 – Normal Data

1. Categorical data consists of values that represent different categories or groups, but these

values are not inherently numerical, making them difficult for machine learning algorithms to process directly. Most algorithms, like linear regression, decision trees, or neural networks, require numerical inputs to perform mathematical operations. Therefore, we need to convert these categorical variables into a numerical format.

Machine Learning Algorithms Require Numeric Input: Many algorithms can’t process categorical data directly, so we need to represent the categories numerically. One-hot encoding creates binary columns (0s and 1s) for each category.

Simplicity: It's simple to implement and interprets categorical features more explicitly as separate binary flags.

