Basic Statistics

### Descriptive Analytics and Data Preprocessing on Sales & Discounts Dataset

#### Descriptive Analytics for Numerical Columns

Numerical columns in the dataset are

['Volume', 'Avg Price', 'Total Sales Value', 'Discount Rate (%)', 'Discount Amount', 'Net Sales Value']

Mean, median, mode, and standard deviation for these columns.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Std Deviation |
| Volume | 5.066 | 4 | 3 | 4.23 |
| Avg Price | 10453.43 | 1450 | 400 | 18079.9 |
| Total Sales Value | 33812.83 | 5700 | 24300 | 50535.07 |
| Discount Rate | 15.15 | 16.58 | 5 | 4.22 |
| Discount Amount | 3346.49 | 988.93 | 69.17 | 4509.9 |
| Net Sales Value | 30466.33 | 4677.79 | 326.97 | 46358.65 |

1. Volume:

- The average volume of sales is approximately 5 units.

- The majority of sales cluster around 3 units, as it's the mode.

- There's a moderate amount of variability in the volume of sales, indicated by the standard deviation of 4.23.

2. Average Price:

- The mean average price is significantly higher than the median and mode, suggesting that there might be some outliers or a skewed distribution towards higher prices.

- There's a considerable amount of variability in prices, as indicated by the large standard deviation of 18079.9.

3. Total Sales Value:

- The mean total sales value is substantially higher than both the median and mode, indicating potential skewness towards higher sales values.

- There's a wide range in the total sales values, as evidenced by the large standard deviation of 50535.07.

4. Discount Rate:

- The average discount rate is around 15.15%.

- The most common discount rate is 5%, as it's the mode.

- There's relatively low variability in discount rates, with a small standard deviation of 4.22.

5. Discount Amount:

- The average discount amount is approximately 3346.49 units.

- The most common discount amount is 69.17 units, as it's the mode.

- There's moderate variability in the discount amounts, indicated by the standard deviation of 4509.9.

6. Net Sales Value:

- The mean net sales value is substantially higher than both the median and mode, indicating potential skewness towards higher sales values.

- There's a wide range in the net sales values, as evidenced by the large standard deviation of 46358.65.

#### Data Visualization

* **Histograms**:

1. **Volume:**

Most of the data is distributed toward the left side, and the longer tail is on the right side. It is a right-skewed distribution. And outliers are present on the right side.

1. **Avg Price:**

Most of the data is distributed toward the left side, and the longer tail is on the right side. It is a right-skewed distribution. And outliers are present on the right side.

1. **Total Sales Value:**

Most of the data is distributed toward the left side, and the longer tail is on the right side. It is a right-skewed distribution. And outliers are present on the right side.

1. **Discount Rate:**

Most of the data is distributed toward the right side, and the longer tail is on the left side. It is a left-skewed distribution. Outliers are present on the left side.

1. **Discount Amount:**

Most of the data is distributed toward the left side, and the longer tail is on the right side. It is a right-skewed distribution. And outliers are present on the right side.

1. **Net Sales Value:**

Most of the data is distributed toward the left side, and the longer tail is on the right side. It is a right-skewed distribution. And outliers are present on the right side.

* **Boxplots**:

**Volume:**

The median volume is 4, indicating that half of the observations have a volume of 4 or less.

The first and third quartiles are 3 and 6, respectively, meaning that 25% of the observations have a volume of 3 or less, and 75% have a volume of 6 or less.

The minimum volume is 1, and the maximum volume is 31, suggesting the presence of some outliers at the higher end of the distribution.

**Avg Price:**

The median average price is 1,450, meaning that half of the observations have an average price of 1,450 or less.

The first and third quartiles are 465 and 10,100, respectively, indicating that 25% of the observations have an average price of 465 or less, and 75% have an average price of 10,100 or less.

The minimum average price is 290, and the maximum is 60,100, suggesting the presence of some outliers at the higher end of the distribution.

**Total Sales Value:**

The median total sales value is 5,700, meaning that half of the observations have a total sales value of 5,700 or less.

The first and third quartiles are 2,700 and 53,200, respectively, indicating that 25% of the observations have a total sales value of 2,700 or less, and 75% have a total sales value of 53,200 or less.

The minimum total sales value is 400, and the maximum is 196,400, suggesting the presence of some outliers at the higher end of the distribution.

**Discount Rate (%):**

The median discount rate is 16.58%, meaning that half of the observations have a discount rate of 16.58% or less.

The first and third quartiles are 13.97% and 18.11%, respectively, indicating that 25% of the observations have a discount rate of 13.97% or less, and 75% have a discount rate of 18.11% or less.

The minimum discount rate is 5.01%, and the maximum is 20.00%, suggesting the presence of some outliers at the higher end of the distribution.

**Discount Amount:**

The median discount amount is 988.93, meaning that half of the observations have a discount amount of 988.93 or less.

The first and third quartiles are 460.46 and 5,316.50, respectively, indicating that 25% of the observations have a discount amount of 460.46 or less, and 75% have a discount amount of 5,316.50 or less.

The minimum discount amount is 69.18, and the maximum is 25,738.02, suggesting the presence of some outliers at the higher end of the distribution.

**Net Sales Value:**

The median net sales value is 4,677.79, meaning that half of the observations have a net sales value of 4,677.79 or less.

The first and third quartiles are 2,202.21 and 47,847.91, respectively, indicating that 25% of the observations have a net sales value of 2,202.21 or less, and 75% have a net sales value of 47,847.91 or less.

The minimum net sales value is 326.97, and the maximum is 179,507.48, suggesting the presence of some outliers at the higher end of the distribution.

**Bar Chart Analysis for Categorical Column**

Categorical columns:

['Day', 'SKU', 'City', 'BU', 'Brand', 'Model']

Bar Chart Analysis:

* Day:

Distribution of sales on all days is the same except on Thursday.

Thursday has a little more sales comparatively other days.

* SKU:

All SKUs have the same distribution of sales.

* City:

All sales are happening in only one city.

* BU:  
  There are three brand units Mobiles, FMCG, and Lifestyle has the same frequency.
* Brand:

some brands sell more than others. This could be because they're better at marketing or because people prefer their products.

* Model:

All Models are distributed on the same frequency.

**Standardization of Numerical Variables**

We are converting the entire data towards zero as a mean and data has been distributed. This is called “**standardization**”.

While standardizing the data all the x variables will lose their units.

Standard normal distribution formula:

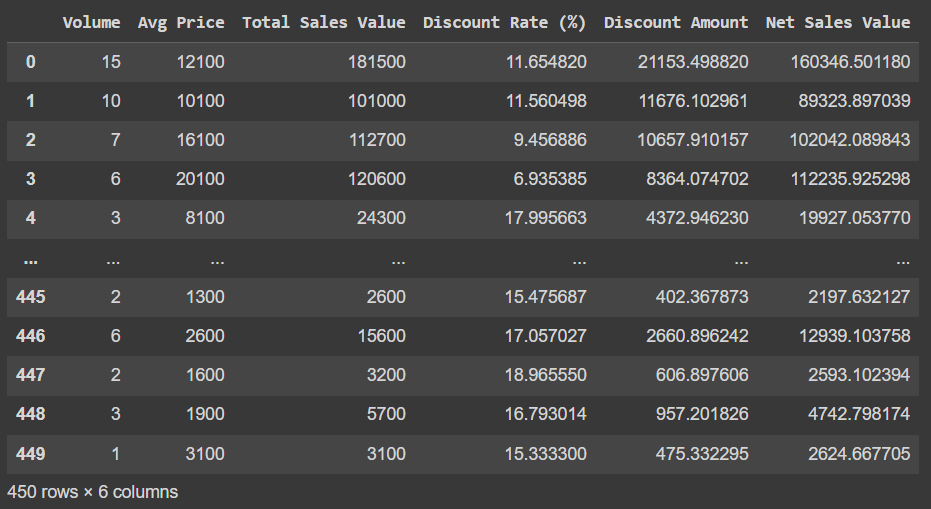
Z=(x-mu)/sigma

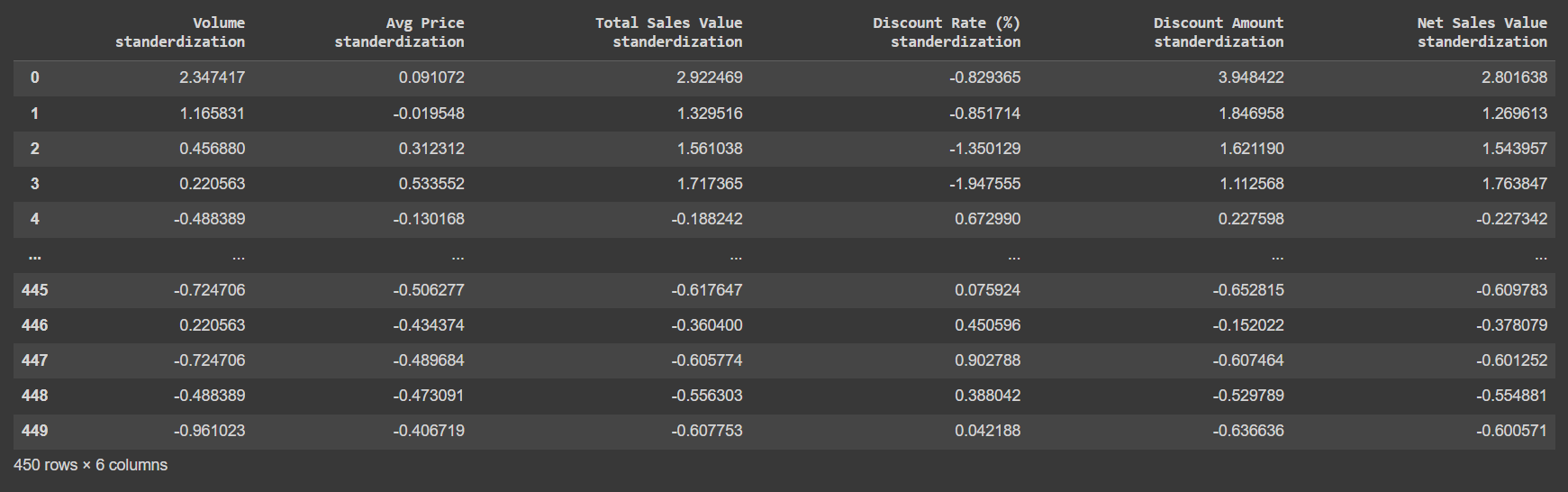
Where,

X – variable

Mu – mean

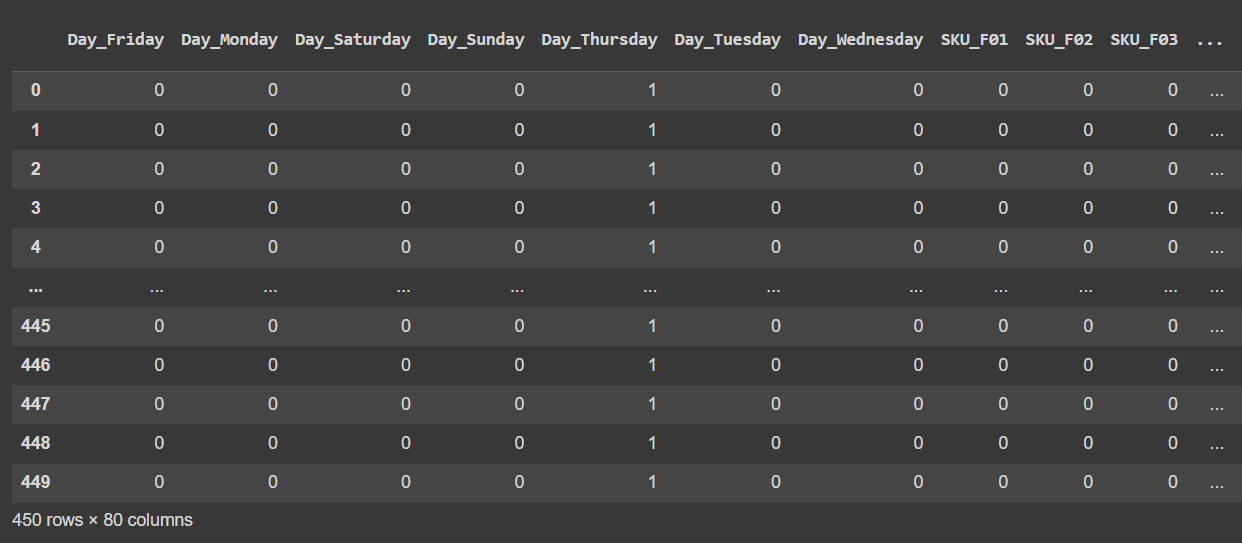
Sigma – standard deviation.

 **Before standardization**

** After standardization**

#### **Conversion of Categorical Data into Dummy Variables**

One hot encoding technique transforms the categorical data into Integer data (0 and 1). So, it can be used in machine learning algorithms.



**Conclusion:**

**2:**

* Standardization in data science and machine learning involves scaling features with a mean of 0 and a standard deviation of 1. It ensures all features contribute equally, aiding fair model training and better algorithm performance**.**
* Many machine learning algorithms cannot directly handle categorical data. One-hot encoding transforms categorical variables into a format that can be provided to ML algorithms more effectively. Each category becomes a separate binary feature, where 1 indicates the presence of the category and 0 is its absence. This prevents the model from assigning false ordinality to categorical variables and ensures that all categories are treated equally.