# Basic Statistics - 1

Bollina Jaya Chandra

November, 2024

**Abstract**

To provide descriptive analytics, visualize data distributions, preprocessing on sales and discount dataset for further analysis and make a summary out of it.

## 1 Descriptive Analytics on numerical columns

### 1.1 Table providing required values for descriptive analytics

| | Volume | Avg Price | Total Sales Value | Discount Rate (%) | Discount Amount | Net Sales Value |
|---|---|---|---|---|---|---|
| mean | 5.066667 | 10453.433333 | 33812.835556 | 15.155242 | 3346.499424 | 30466.336131 |
| median | 4.000000 | 1450.000000 | 5700.000000 | 16.577766 | 988.933733 | 4677.788059 |
| std | 4.231602 | 18079.904840 | 50535.074173 | 4.220602 | 4509.902963 | 46358.656624 |
| min | 1.000000 | 290.000000 | 400.000000 | 5.007822 | 69.177942 | 326.974801 |
| 25% | 3.000000 | 465.000000 | 2700.000000 | 13.965063 | 460.459304 | 2202.208645 |
| 50% | 4.000000 | 1450.000000 | 5700.000000 | 16.577766 | 988.933733 | 4677.788059 |
| 75% | 6.000000 | 10100.000000 | 53200.000000 | 18.114718 | 5316.495427 | 47847.912852 |
| max | 31.000000 | 60100.000000 | 196400.000000 | 19.992407 | 25738.022194 | 179507.479049 |

Table 1: Descriptive statistics on sales and discount data

### 1.2 Interpretation

Most of the columns has multiple mode i.e., every observation is a mode. Let us interpret the data using mean, median and standard deviation.

- For column Volume, Average is around 5, median around 4 and standard deviation around 4 suggests moderate variability in Volume column.
- For column Avg Price, mean >>> median, makes it right - skewed distribution. Higher standard deviation shows high variability in the column. It depicts some high priced values present in dataset pulling the average up. Same explanation goes with columns like Total Sales Value, Discount Amount and Net Sales Volume.
- For column Discount Rate (%) , mean is around 15, median around 16 and standard deviation around 4 suggests moderate variability in the column.

# 2 Data Visualization

## 2.1 Histograms

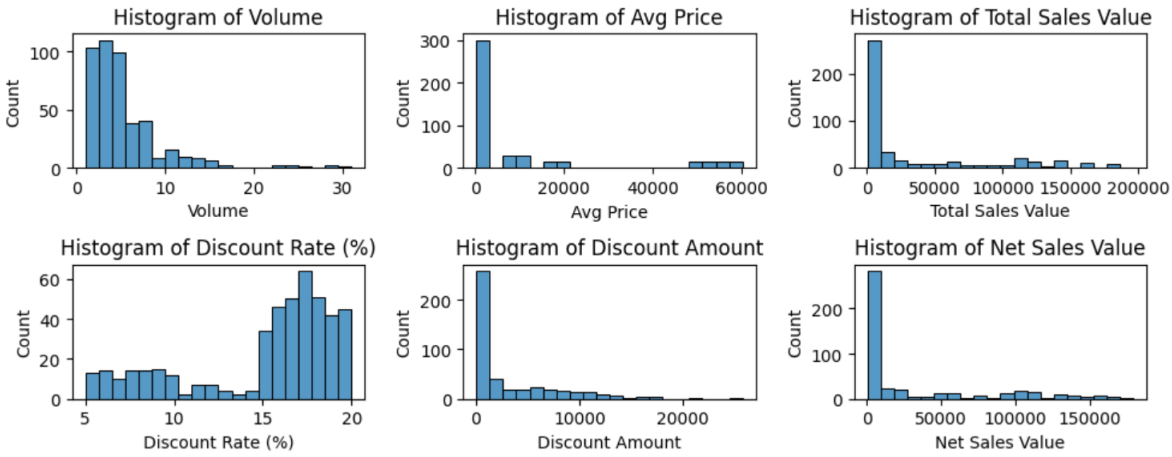


Figure 1: Histograms of all numerical columns in the dataset.

- The above histograms (right - skewed) shows that customers tendency to make low cost transactions more, along with some high cost transactions (extending tails (mostly outliers) in most histograms) which has significant impact on overall revenue.
- The only histogram that is different from others is Discount Rate (%), It is mainly concentrated in two regions, that shows the policy of sellers in offering discounts.

## 2.2 Boxplots

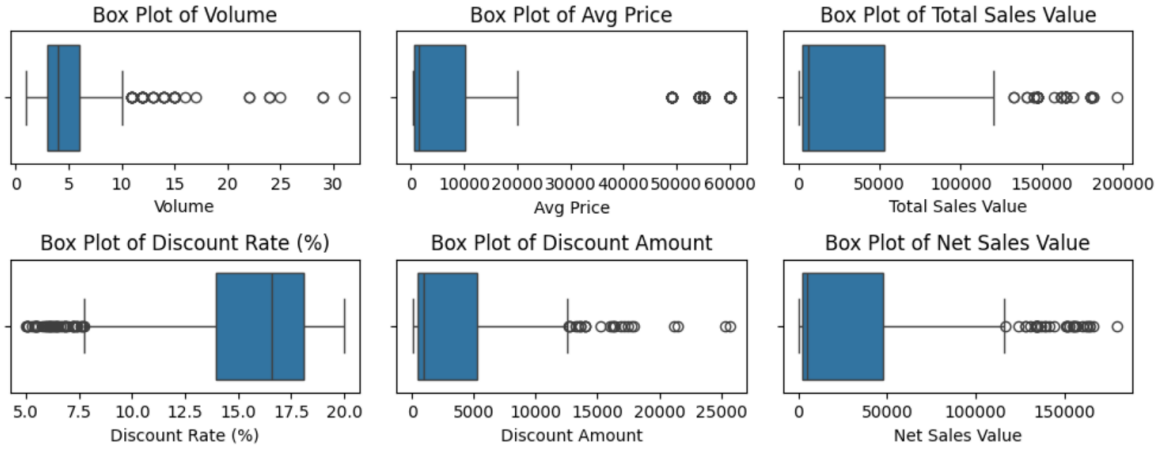


Figure 2: Box-plots of all numerical columns in the dataset.

- Extreme Outliers: Box plot of Volume has many outliers to its right, that might be case of bulk purchases, Similarly outliers in Avg Price depicts high valued products, outliers in Sales value depicts either bulk purchases or high valued product purchase and discount amount depicts discount on high valued purchases.
- Box plot of discount rate has aligned towards its right (15-18) region and more predictable than others.

### 2.3 Bar Charts

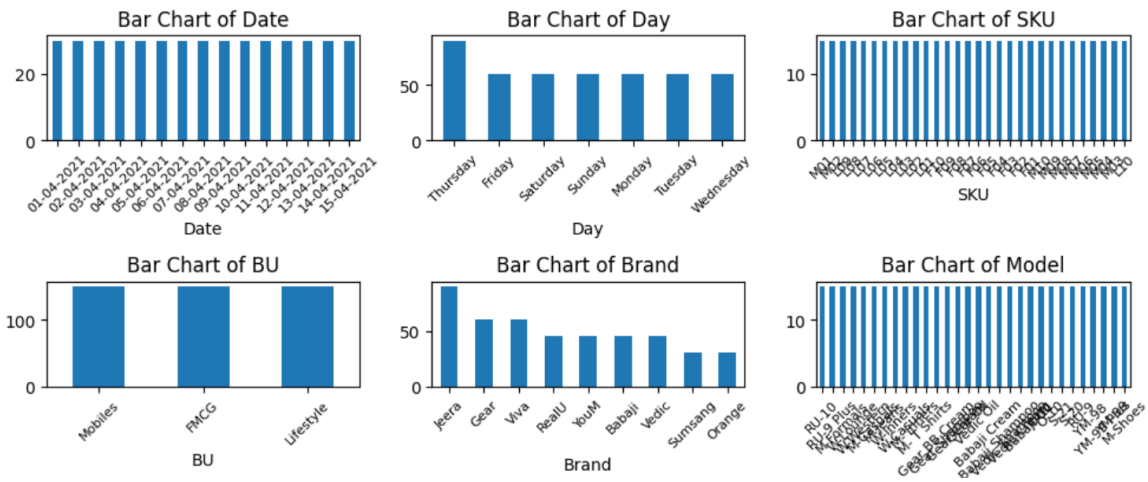


Figure 3: Bar Charts of categorical columns in the dataset.

- The uniformity across bar charts depicts the dataset is well structured in most of the aspects like BU, Model, SKU and date wise.
- There is a change in day wise chart and it favors to Thursday. In Brand chart there is some favors towards life style needs compared to mobile brands.

## 3 Standardization

### 3.1 Z-score normalization

It is a method to scale numerical data, with mean of 0 and standard deviation of 1.

$$z = \frac{x - \mu}{\sigma}$$

Where:

- z is standardized value
- x is original value
- $\mu$ is mean of dataset
- $\sigma$ is standard deviation of dataset.

### 3.2 Standard Distribution of numerical columns

- Before standard distribution, numerical columns has wide range of scale, and every columns has its own scale.
- After standard distribution, every numerical column has same scale which makes it easy for analysis,and has mean close to 0 and standard deviation close to 1.

Before Standard Distribution

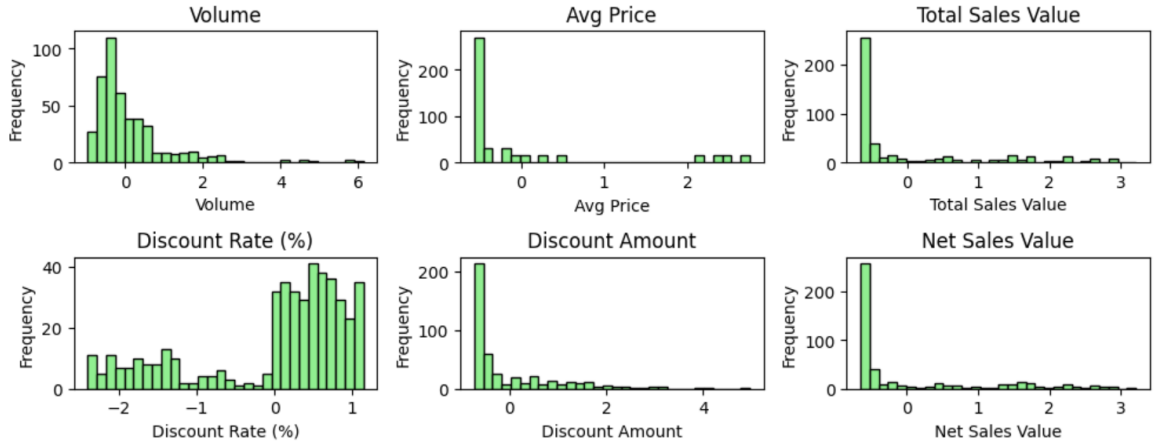


Figure 4: Standard Distribution of Numerical columns

# 4 Categorical columns to dummy variables

Machine Learning algorithms work on numerical inputs, To take categorical inputs into consideration for building a model, we need to convert them into numerical values. One-hot encoding is a method to convert each category to a binary column of present or absent.

| | Discount Amount | Net Sales Value | Day_Monday | Day_Saturday | Day_Sunday | \ |
|---|---|---|---|---|---|---|
| 0 | 21153.498820 | 160346.501180 | False | False | False | |
| 1 | 11676.102961 | 89323.897039 | False | False | False | |
| 2 | 10657.910157 | 102042.089843 | False | False | False | |
| 3 | 8364.074702 | 112235.925298 | False | False | False | |
| 4 | 4372.946230 | 19927.053770 | False | False | False | |

| | ... | Model_Vedic Cream | Model_Vedic Oil | Model_Vedic Shampoo | \ |
|---|---|---|---|---|---|
| 0 | ... | False | False | False | |
| 1 | ... | False | False | False | |
| 2 | ... | False | False | False | |
| 3 | ... | False | False | False | |
| 4 | ... | False | False | False | |

| | Model_W-Casuals | Model_W-Inners | Model_W-Lounge | Model_W-Western | \ |
|---|---|---|---|---|---|
| 0 | False | False | False | False | |
| 1 | False | False | False | False | |
| 2 | False | False | False | False | |
| 3 | False | False | False | False | |
| 4 | False | False | False | False | |

Figure 5: Portion of one-hot encoding on given dataset.

# 5 Conclusion

Descriptive analysis and visualizations interprets the right skewed distributions, which depicts the presence of high value transactions is significantly affecting the overall distribution.

Standardization is helpful in bringing every numerical variable into a single scale to avoid studying wide spread of value over the visualizations.

One - hot encoding makes categorical columns to numerals that ensures machine learning algorithms to understand and enhance the performance.