**Assignment- Basic Statistics-I**

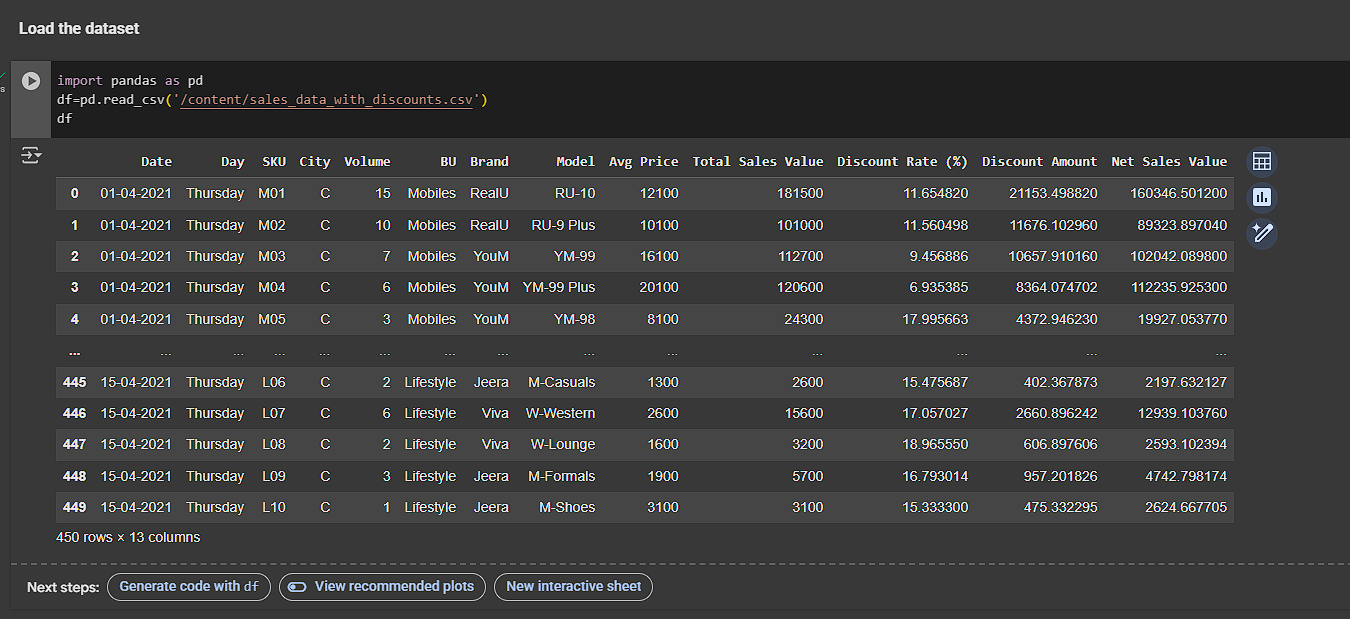
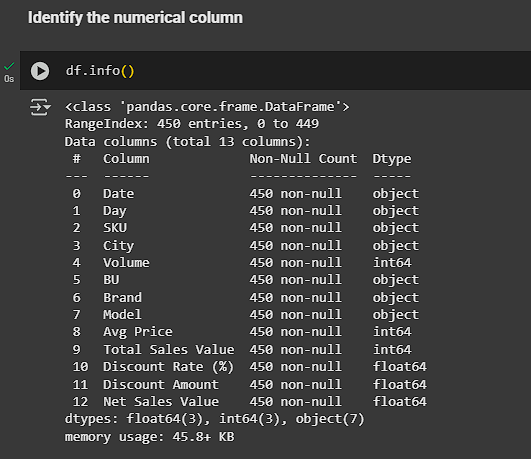
**Descriptive Analytics and Data Preprocessing on Sales & Discounts Dataset**

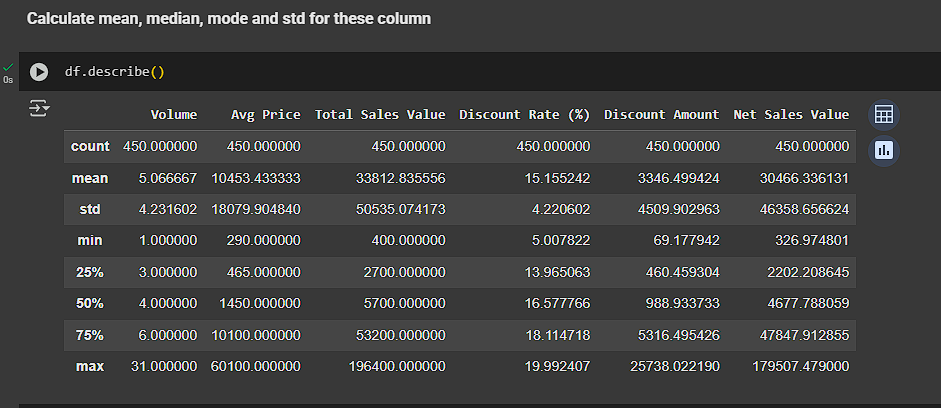
***Introduction***

* To perform descriptive analytics, visualize data distributions, and preprocess the dataset for further analysis.

***Descriptive Analytics for Numerical Columns***

* **Objective:** To compute and analyze basic statistical measures for numerical columns in the dataset.
* **Steps:**

1. Load the dataset into a data analysis tool or programming environment (e.g., Python with pandas library).
2. Identify numerical columns in the dataset.
3. Calculate the mean, median, mode, and standard deviation for these columns.



1. Provide a brief interpretation of these statistics.

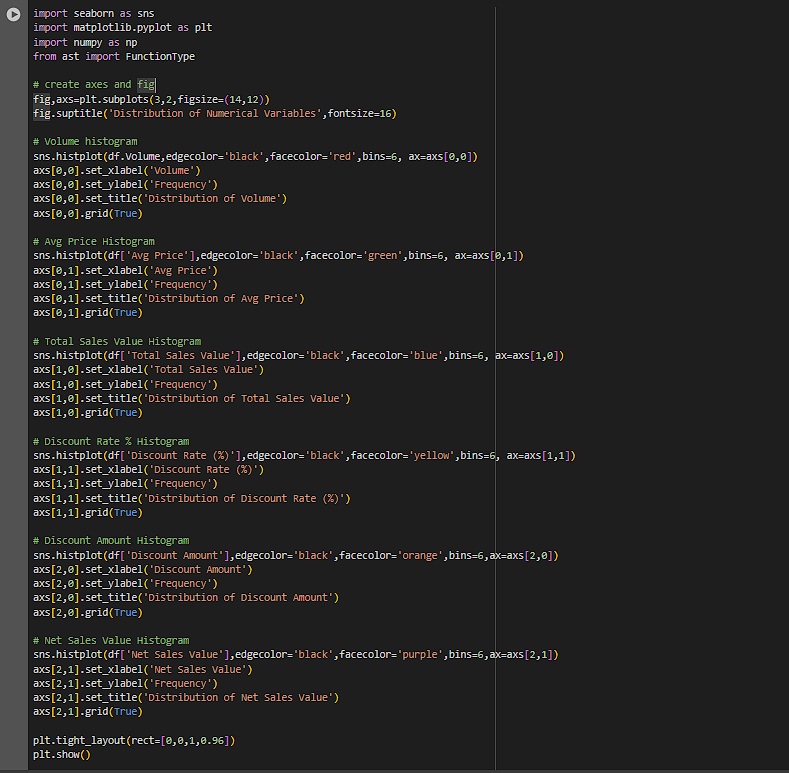
Ans:- Here's a brief interpretation of the descriptive statistics:

* **Volume:** The average number of units sold is around 5, with a standard deviation of 4.23. This indicates that the sales volume for individual products can vary quite a bit. The maximum volume sold for a single product is 31, while the minimum is 1.
* **Avg Price:** The average price of a product is about 10,453. However, the standard deviation is very high (18,080), which suggests a wide range of product prices. This is confirmed by the minimum price of 290 and a maximum price of 60,100. The median price is 1,450, which is much lower than the mean, indicating that the distribution of prices is skewed to the right by a few very expensive items.
* **Total Sales Value:** The average total sales value is approximately 33,813, but like the average price, it has a large standard deviation (50,535). This again points to a wide range in sales values across different products.
* **Discount Rate (%):** The average discount rate is about 15.16%, with a relatively small standard deviation of 4.22. This suggests that the discount rates are somewhat consistent across products, mostly falling between 11% and 19%.
* **Discount Amount & Net Sales Value:** These values follow a similar pattern to the 'Total Sales Value' and 'Avg Price', with large standard deviations indicating significant variability.

#### **Data Visualization**

* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:

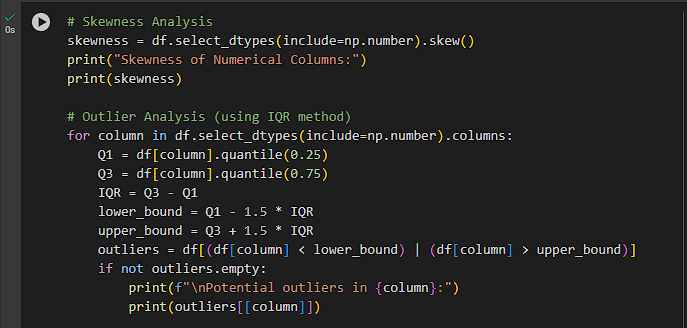
1) Plot histograms for each numerical column.



A graph of a number of percents

Description automatically generated with medium confidenceA group of graphs with different colored bars

Description automatically generated with medium confidence

2) Analyze the distribution (e.g., skewness, presence of outliers).

3) Provide Inferences

Ans:- Inferences from the analysis

1. **Volume, Avg Price, Total Sales Value, Discount Amount, and Net Sales** Value are all highly positively skewed. This indicates that a majority of the sales involve lower values, with a few transactions having exceptionally high values. These high-value transactions could be considered outliers.

2. **Discount Rate (%)** is negatively skewed, suggesting that most of the discounts offered are on the higher side, with fewer instances of lower discounts.

3. The outlier analysis confirms the presence of high-value outliers in the skewed distributions, which significantly affect the mean and standard deviation.

* **Boxplots**:

A screenshot of a computer program

Description automatically generated1) Create boxplots for numerical variables to identify outliers and the interquartile range.

A close-up of a graph

Description automatically generatedA group of graphs showing different colors

Description automatically generated

2) Discuss any findings, such as extreme values or unusual distributions.

Ans:- Inferences from Boxplots

Boxplots visually confirm the presence of numerous outliers, especially in 'Volume', 'Avg Price', 'Total Sales Value', 'Discount Amount', and 'Net Sales Value'.

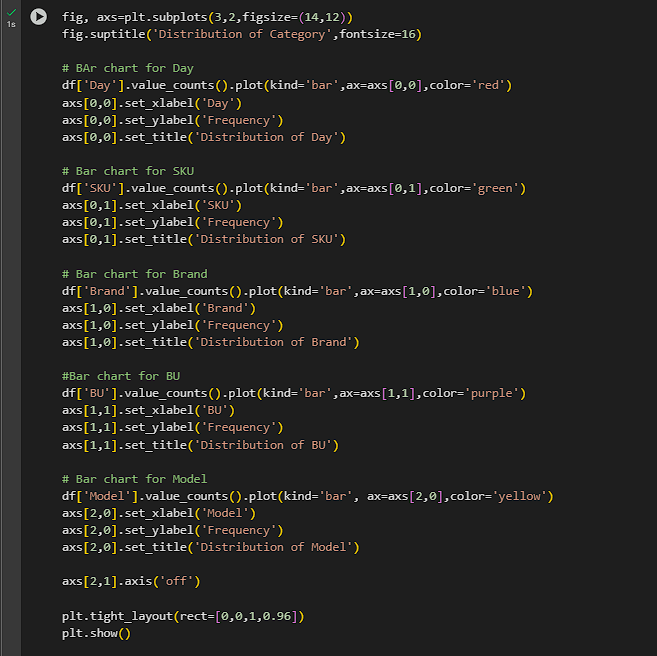
These outliers are clearly visible as individual points beyond the whiskers of the boxplots.

The boxplots for these variables are also highly skewed, with the median line closer to the bottom of the box and long upper whiskers, indicating the positive skewness observed in the skewness analysis.

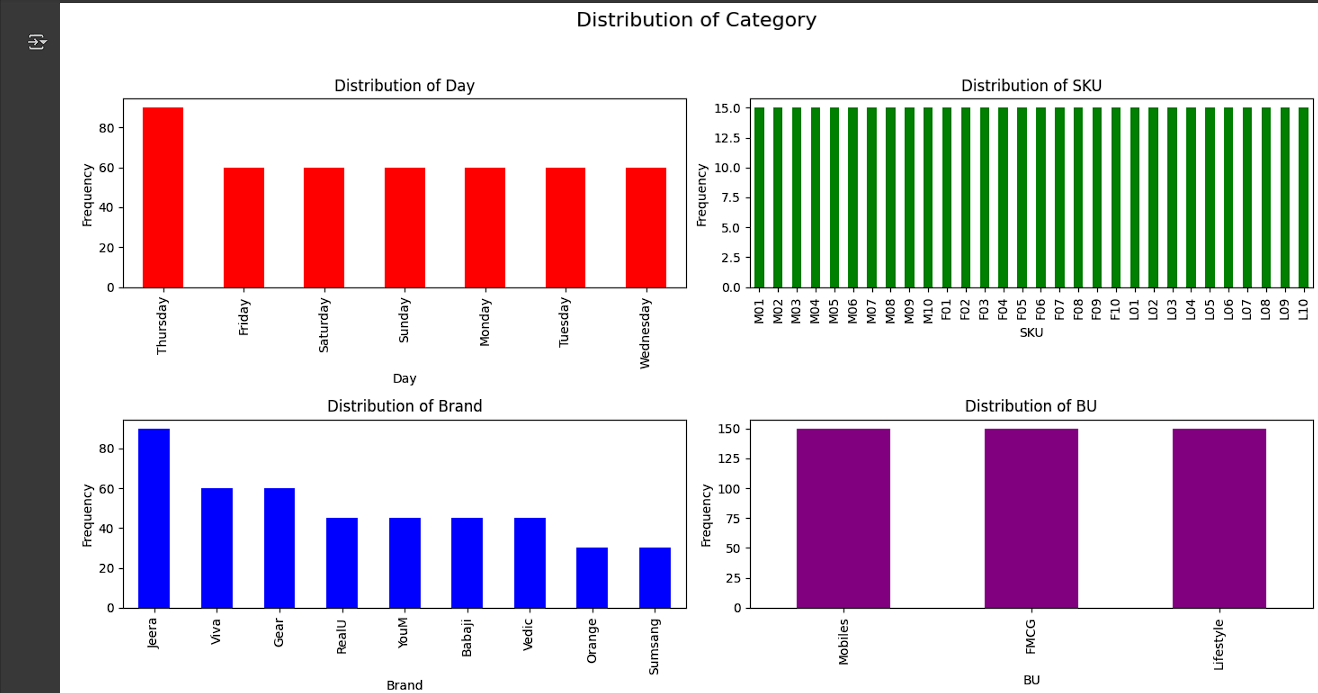
Discount Rate (%)' shows a more symmetrical distribution compared to the other numerical variables, although it still has some outliers.

The interquartile range (the box itself) is relatively small for 'Volume', 'Avg Price', 'Total Sales Value', 'Discount Amount', and 'Net Sales Value' compared to the overall range of the data, further highlighting the influence of outliers on the spread.

* **Bar Chart Analysis for Categorical Column:**

1. Identify categorical columns in the dataset.
2. Create bar charts to visualize the frequency or count of each category.

A yellow and black text

Description automatically generated

1. Analyze the distribution of categories and provide insights.

Ans:- Analyze the distribution of categories and provide insights

1. **Day**: The sales are distributed relatively evenly across the days of the week, with Thursday having the highest frequency.

2. **SKU and Model**: There is a large number of unique SKUs and Models, and the bar charts show that the sales frequency is somewhat evenly distributed among them, with no single SKU or Model dominating the sales.

3. **Brand:** 'Jeera' is the most frequent brand in the dataset, followed by 'Viva' and 'Gear'. This suggests that these brands have a higher volume of transactions compared to others.

4. **BU (Business Unit):** The 'Mobiles' and 'Lifestyle' business units have the highest frequencies, indicating that these are the primary categories of products sold.

#### **Standardization of Numerical Variables**

* **Objective:** To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* **Steps:**

1) Explain the concept of standardization (z-score normalization).

Ans:- Standardization, also known as Z-score normalization, is a data preprocessing technique used to scale numerical features so that they have a mean of 0 and a standard deviation of 1.

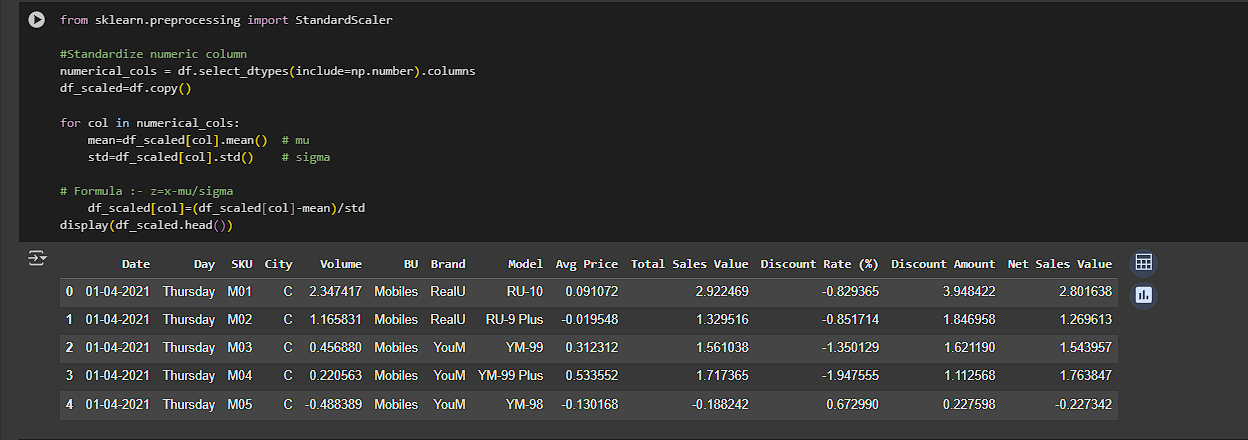
The formula for standardization is:

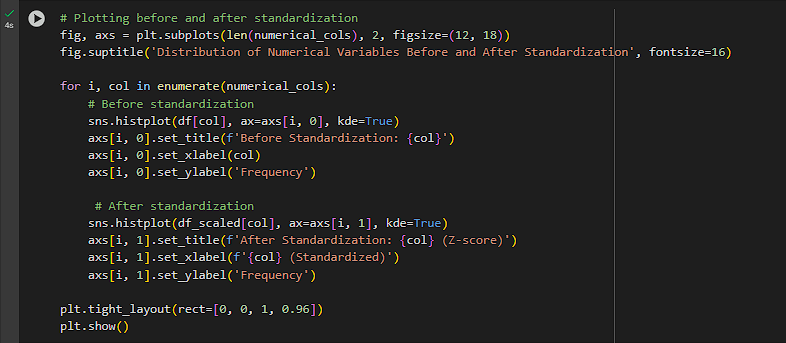
z=(x−μ)σ

Where:

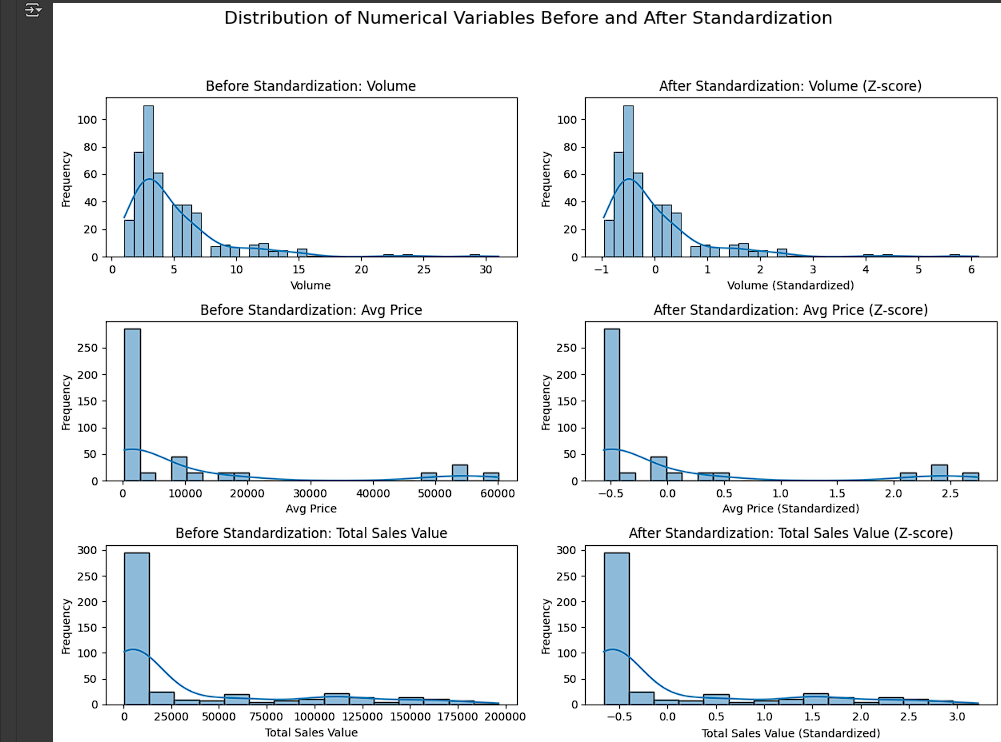
* x is the original value
* μ is the mean of the column
* σ is the standard deviation of the column
* z is the standardized value

This technique is particularly useful for algorithms that are sensitive to the scale of features, such as gradient descent-based algorithms

2) Standardize the numerical columns using the formula: z=x-mu/sigma.

3) Show before and after comparisons of the data distributions.

A group of graphs showing different types of value

Description automatically generated with medium confidence

**Conversion of Categorical Data into Dummy Variables**

* **Objective**: To transform categorical variables into a format that can be provided to ML algorithms.
* **Steps:**

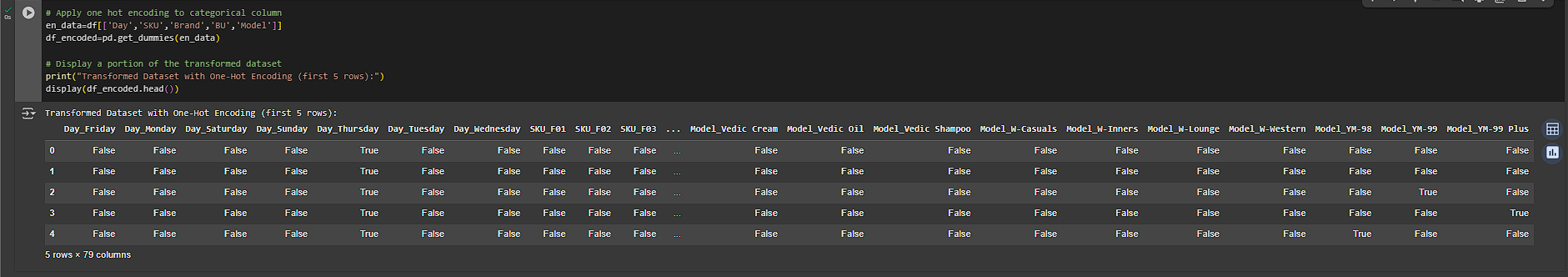
1. Discuss the need for converting categorical data into dummy variables (one-hot encoding).

Ans:- Need for Converting Categorical Data into Dummy Variables (One-Hot Encoding):

Machine learning algorithms typically work best with numerical data. Categorical variables, which represent categories or labels, need to be converted into a numerical format before being fed into most algorithms. One-hot encoding is a common technique for this conversion.

One-hot encoding creates new binary columns for each unique category in a categorical variable. For example, if a 'Color' column has categories 'Red', 'Blue', and 'Green', one-hot encoding will create three new columns: 'Color\_Red', 'Color\_Blue', and 'Color\_Green'. For each row, the corresponding color column will have a value of 1, and the others will have a value of 0.

This process avoids assigning arbitrary numerical values to categories that might imply an unintended ordinal relationship (e.g., assigning 1 to 'Red', 2 to 'Blue', and 3 to 'Green' could imply that 'Green' is 'more' than 'Red', which is not the case).

1. Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category and display a portion of the transformed dataset.

**Conclusion**

* **Numerical Variables:** Features like *Volume, Avg Price, Total Sales Value, Discount Amount,* and *Net Sales Value* are highly positively skewed with significant outliers, meaning a few high-value transactions heavily impact averages. *Discount Rate (%)* is more symmetrical with fewer extreme values.
* **Categorical Variables:** Sales are fairly steady across weekdays. Top brands are *Jeera, Viva,* and *Gear*, while *Mobiles* and *Lifestyle* dominate as business units.
* **Pre-processing Steps:**
  1. ***Standardization (Z-score normalization):*** Scales numerical features to handle differences in ranges, crucial for algorithms sensitive to magnitude.
  2. ***One-hot encoding****:* Converts categorical variables into numerical form without implying order, making them usable for ML models.