**ASSIGNMENT-1**

**BASIC STATISTICS -1**

**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

**Step-1:**Importing the required libraries to perform the task

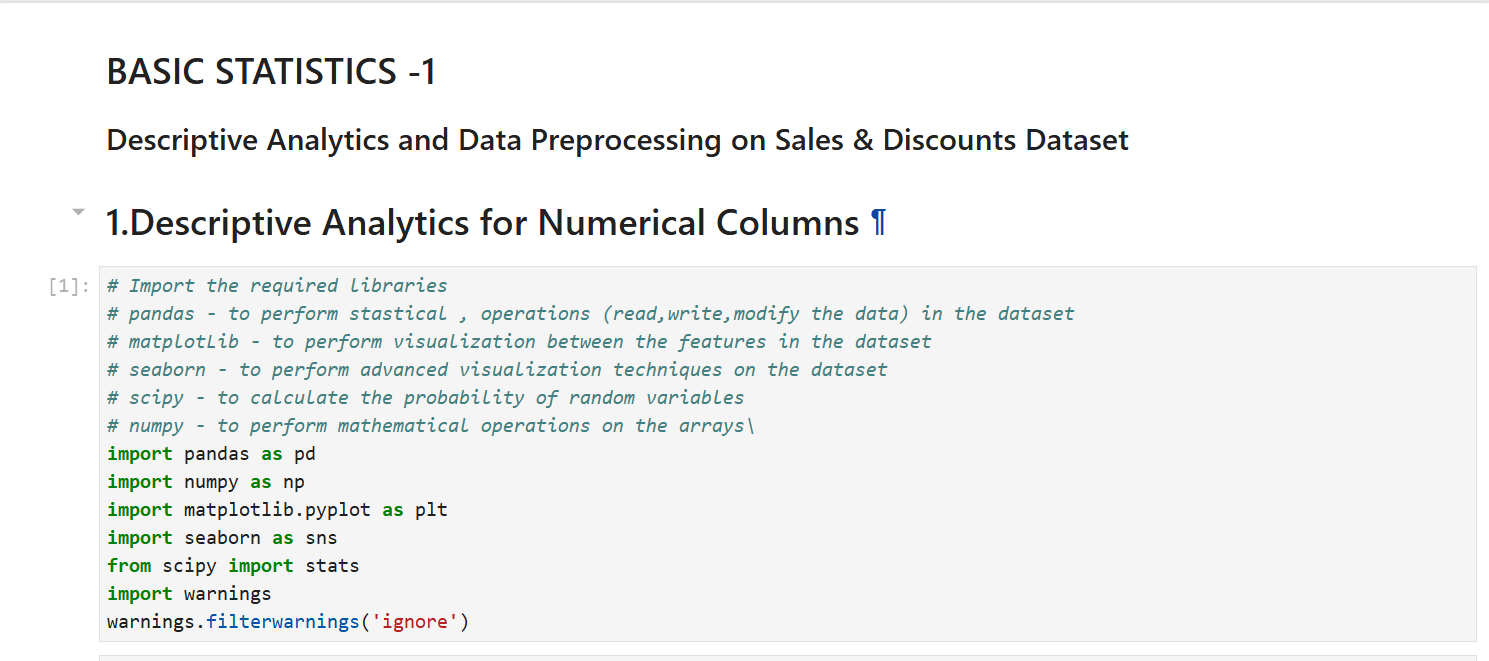
**Pandas:** the pandas are used to perform the statistical operations ,reading dataset, writing, and modifying the values in the dataset .

**Matplotlib.pyplot :** the matplotlib library is used to plot the 2D graphs like boxplot, hist plot, scatter plot etc between the features (columns) of the dataset.

**Seaborn :** the seaborn library is used to plot the advanced graphs containing high graphics and quality graphs like swarm plot, pair plot, violinplot etc.

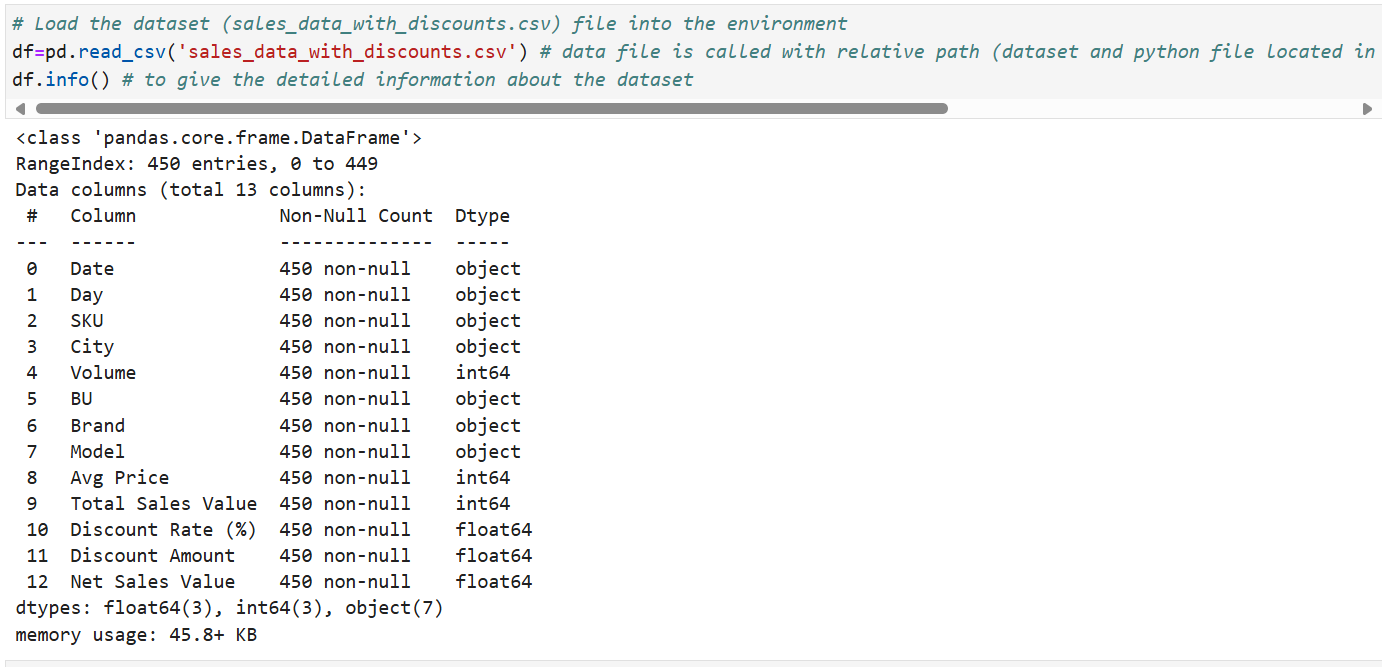
**Numpy :** the numpy library is used to perform mathematical operation on the arrays.

**Scipyt:** It is used to perfom the hypothesis testing and anova test.

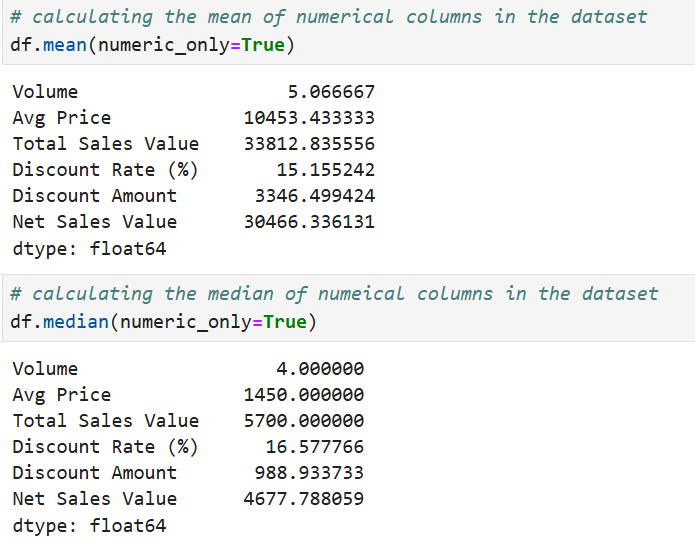
****

**Fig-1: Importing the required libraries**

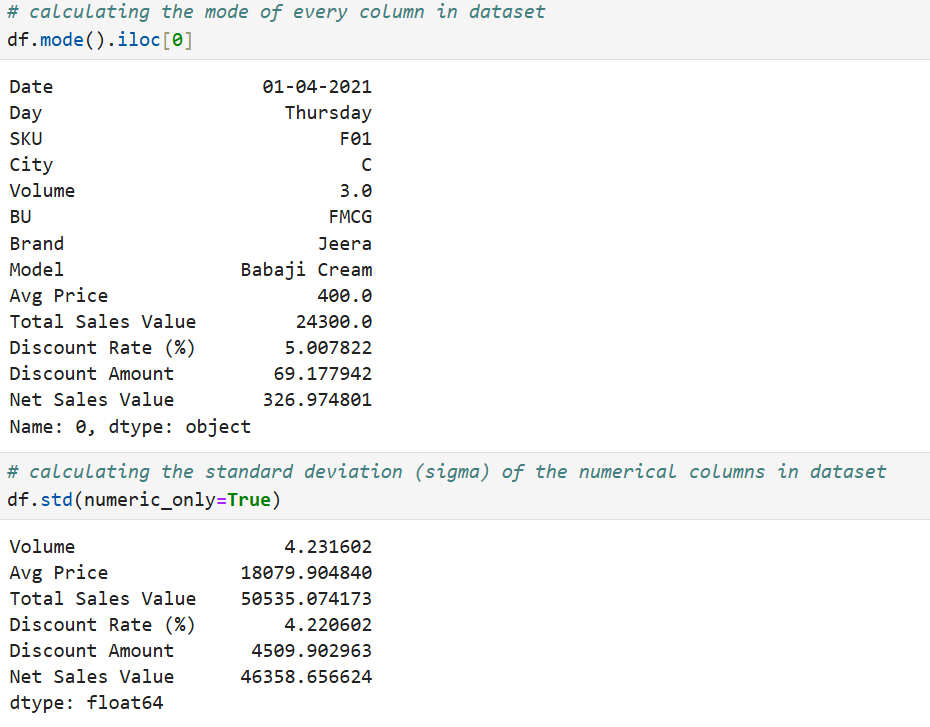
**Step-2:** Loading the dataset into the environment



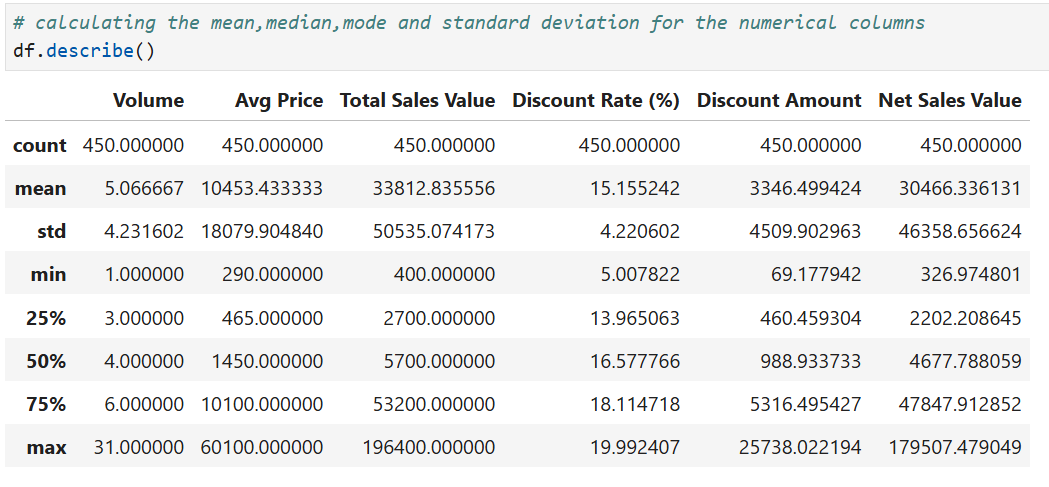
**Fig-2: Loading the dataset into the python environment**

**Step-3 :** calculating the mean, median, mode and standard deviation of the numerical columns in the dataset.

**Fig-3: calculating mean and median of the numerical columns in the dataset**



**Fig-4: Calculating the mode and standard deviation for the numerical columns in the dataset**



**Fig-5: Calculating the basic statistics of the numerical columns using describe function**

**Basic analysis by interpreting the statistical values:**

**Volume:**

1. It has no null values.
2. By comparing the mean (5.066) and median (4) the difference is less so it may be normal distribution.
3. Comparing the 75% value (6) and max value (31) there is lot of difference so in this data there may have the outliers so better to find the outliers and remove .

**Avg Price:**

1. It has no null values.
2. Mean (10453) and median (1450) there is lot of difference between them and mean is greater than median so it may be right skewed data.
3. 75% (quantile-3) value is 10100 and max value is 60100 there is lot off difference so this column has outliers.

**Total Sales Value:**

1. No null values.
2. Mean (33812 ) and median ( 5700) , mean> median , the data is right skewed.
3. Quantile-3 value is 53200 and max value is 196400 so nearly 150000 difference so this column contains outliers.

**Discount Rate:**

1. No null values.
2. Mean (15.15) and median (16.57) the difference is very less so it has normal distribution.
3. Quantile-3 value is18.11 and max value is 19.99 the difference is less and it matches the 25% so in this column there may be no outliers.

**Discount Amount:**

1. No null values.
2. Mean (3346.5) and median (988) , mean > median , the data is right skewed.
3. Quantile-3 value is 5316 and max value is 25736.02 the difference is very high so there may present outliers in this column.

**Net Sales Value:**

1. No null values.
2. Mean (30466) and median (4677) , mean >> median , the data is right skewed.
3. Quantile-3 value is 47847 and max value is 179507 the max value is approximately 3 times more than the quantile-3 value so it may contain the outliers.

**Overall Analysis:**

1. There is no null values in the numerical columns.
2. Few are normally distributed and most of the columns are right skewed .
3. Most of the columns contains the outliers that’s may effecting the distribution of the data so, it is necessary to handle the outliers.

**Data Visualization**

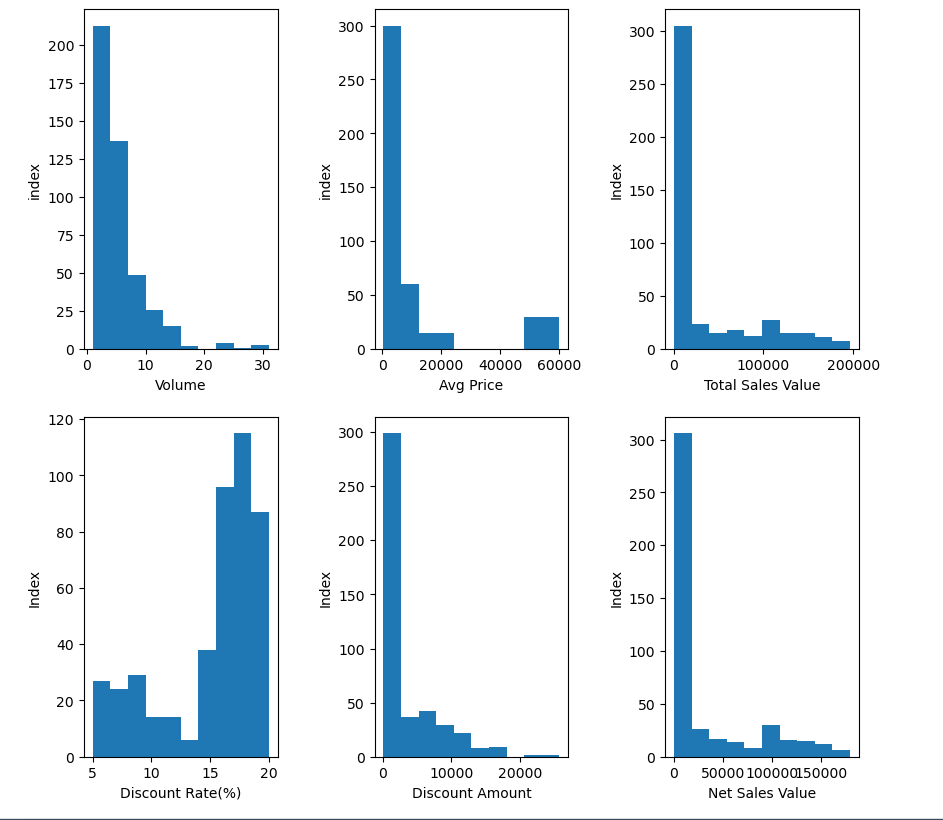
**Objective:** To visualize the distribution and relationship of numerical and categorical variables in the dataset.

**Histogram:**

Histogram is the type of bar plot where the bars are continuous and the width of the bars called as bins, histogram represents the probability distribution curve also.

**Fig-6: Code for plotting Histplot for the numerical columns in the dataset**

The histogram is plotted for each numeric column using the matplotlib and all the plots are grouped by using the concept called subplots, so that we can analysis the each graph clearly.



**Fig-7: Histplots of the numerical groups in the dataset**

**Analysis of histplots:**

1. By observing the above graphs in the volume feature the values are distributed normally except few values , it may contains the outliers.
2. In the Avg price feature the graph is not continuous it has break in between the graph, it may represents that there is no values between 20000 and 60000, it also contains outliers .
3. In the total sale feature the graph is normally distributed except one bin , that bin has highest frequency so excluding the first bin ,rest of the values are normally distributed.
4. In the Discount Rate feature the starting bins are normally distributed but the last bins have the outliers.
5. In the Discount Amount feature it is also similar to the sale Feature ,excluding the first bin all remaining are normally distributed.
6. The Net sales value feature hist plot is also similar to the sale Feature, in this also excepting the first bin all remaining are normally distributed.

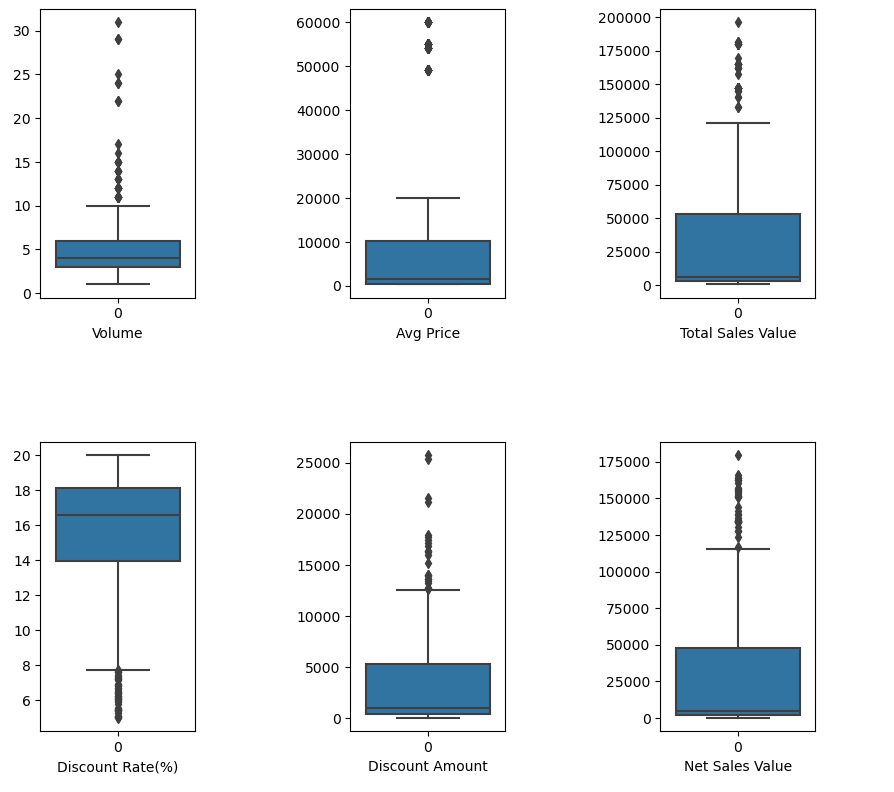
**Box plot:**

The box plot is used for finding the outliers, It contains quantiles, 50% quantile is the median of the data, and it has Q1 25% and Q3 75% by finding the IQR the maximum and minimum quantile ranges are determined , and by utilizing the upper limit and lower limit the outliers are determined.

**Fig-8: Code for Plotting Box plot for the numerical data in the dataset**

**Analysis of Box plot:**

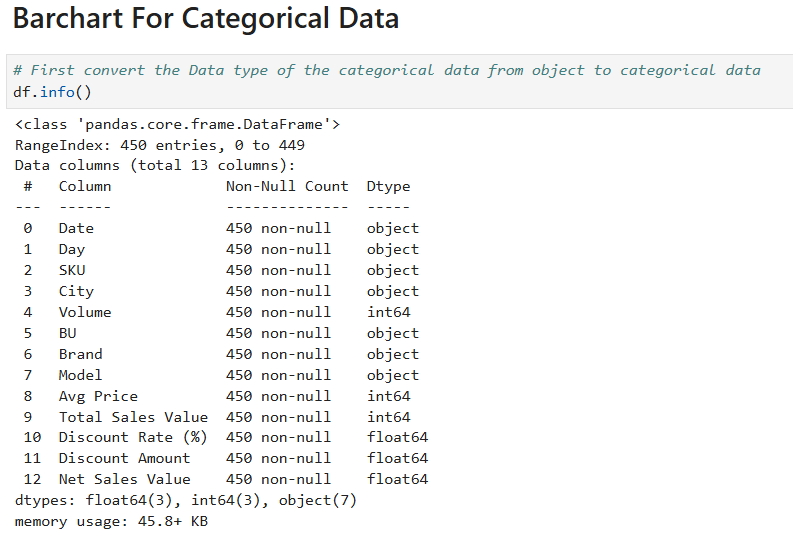
1. **Volume:** From the boxplot drawn for the volume column we can analysis it has more number of outliers that to greater than the upper limit.
2. **Avg Price**: From the boxplot drawn for the Avg Price we can analyse that it has three outliers , and the outliers values are more than the upper limit.
3. **Total Sales Value**: In the boxplot of total Sales value we can analyse that there are more outliers and those values are more than the upper limit value.
4. **Discount Rate:** From the Boxplot of the discount rate we can analyse that there are more number or outliers and those values are less than its lower limit.
5. **Discount Amount:** in the discount amount box plot the distance between the lower limit and the median is very less and most of the data is concentrated on the center , it also consists of outliers and those values are more than the upper limit.
6. **Net Sales Value:** form the boxplot drawn for the net sales value there are outliers and those values are more than the upper limit.



**Fig-9: Box plot for the numerical data in the dataset**

**Bar chart:**

Fig-9: Boxplot for the numerical columns in the dataset

The bar chart is used for the plotting small number of categorical data, it represents a bar for every category on the x-value it indicates the frequency value on the y-axis so that we can identify the outliers in the categorical data.

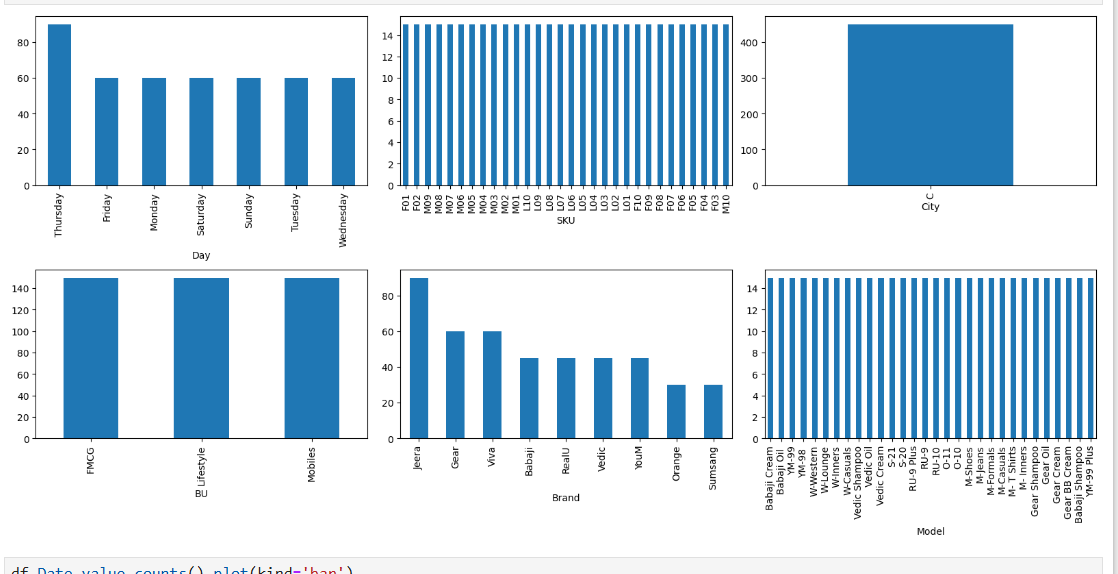
**Fig-10: Category data types before changing its data types**



**Fig-11: Category data types after changing its data types**



**Fig-12: Code for plotting barchart for the categorical data in the dataset**



**Fig-13: Bar chart for the categorical data in the dataset**

**Analysis if Bar chart:**

1. Day: From the bar chart drawn for the Day feature we can conclude that all the days have equal frequency except Thursday , there may contain outliers in the data or there may be some reason for the Thursday for having more frequency.
2. SKU: In the SKU feature all the categories are equally distributed , there are no outliers.
3. City: in the City column there is only one value so in the barchart its showing only one bar.
4. BU: From the bar chart drawn for the BU feature , we can conclude that there are only three categories in It and they are equally distributed.
5. Brands: from the Brands bar chart we can say that the brands are not normally distributed, based on the brands the frequency is changed , among them one brand has high frequency.
6. Model: from the Model bar chart we can analyse that every model has equal frequency so the frequency is constant for all the models.

**Standardization of Numerical Variables**

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

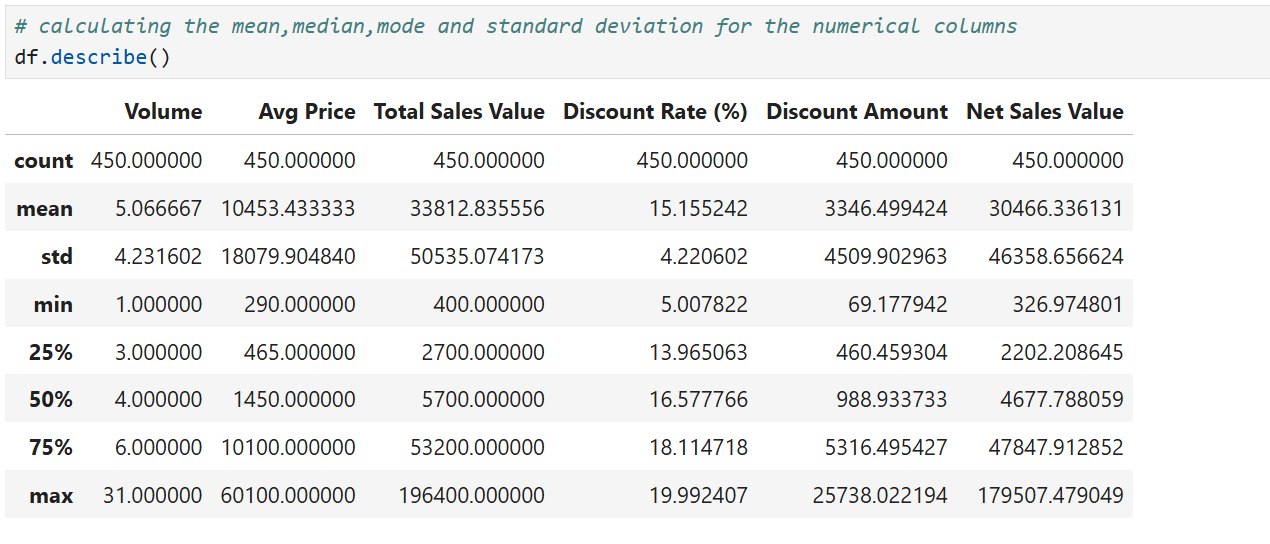
**Standardization:**

Machine learning algorithms are sensitive to the scale, so we convert all the features scale into a particular range, so that the scale is same for all the features (columns).

**Z-Score normalization:**

Replace every value in the feature with their z-score values, so that the range of every feature is from -3.4 to 3.4 .

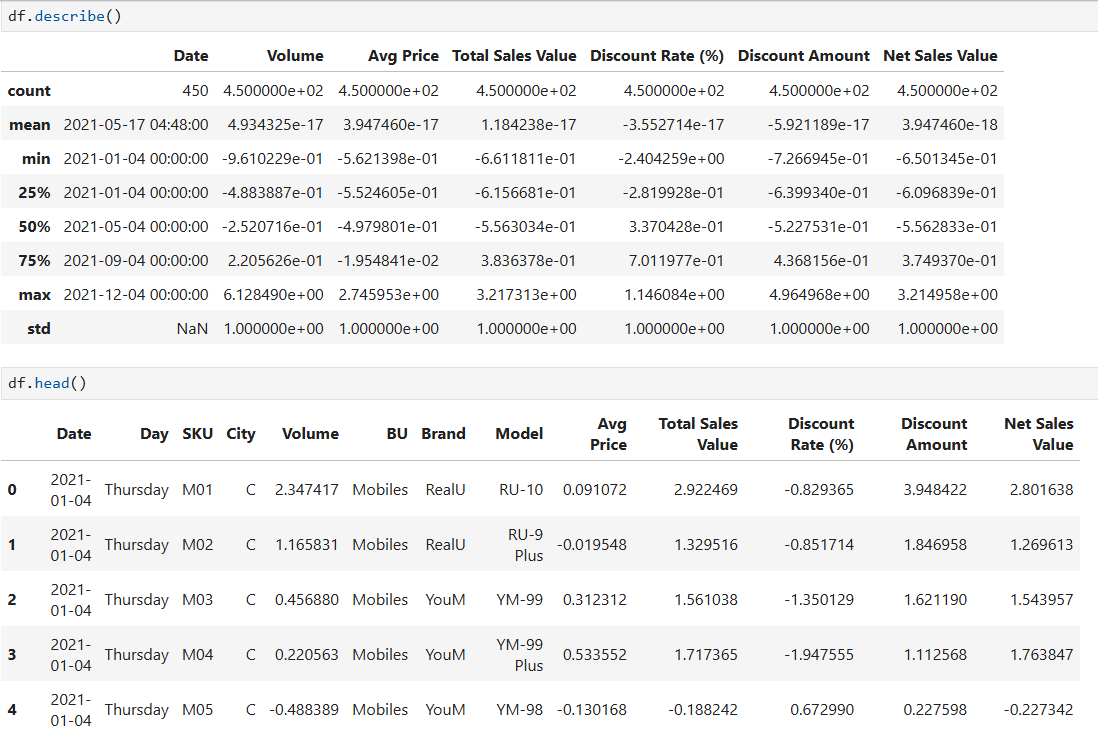
z-score =( (x-x.mean())/x.std())

It can be achieved by using the sklearn library or by defining a function manually we can achieve the same thing.

**Fig-14: Data before Normalizing**



**Fig-15: Function for Normalizing the Data**



**Fig-16: Data After Normalizing**

**Conversion of Categorical Data into Dummy Variables**

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

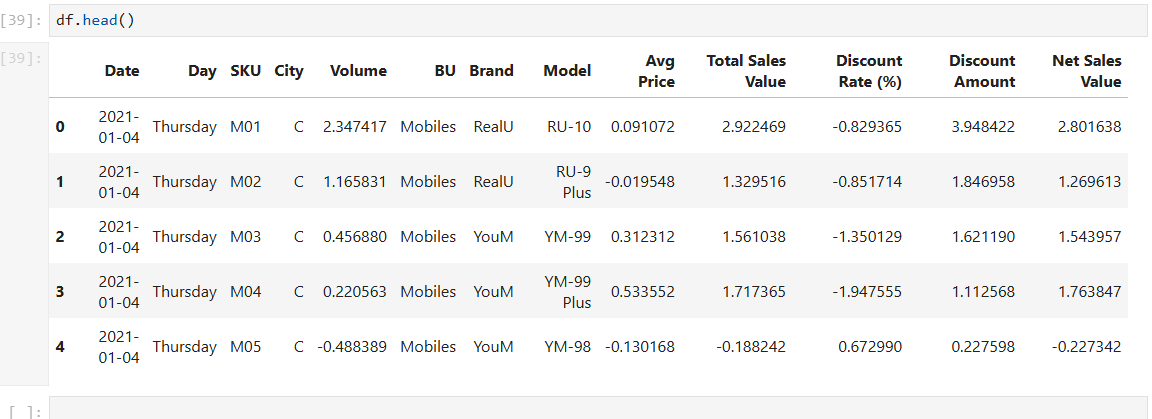
The transformation of data types is also called as feature scaling.

There are mainly two types of transformations are there,

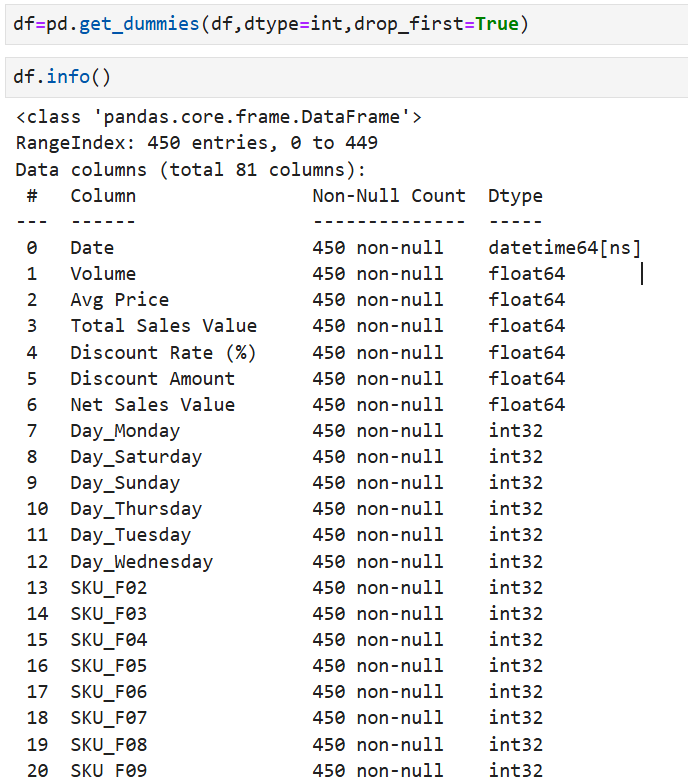
1. **Dummy variables encoding / one-hot encoding :**

Assigning the Boolean values or 0,1 to the data is called as one-hot encoding , this method is used when the dependent variable is numeric in nature.

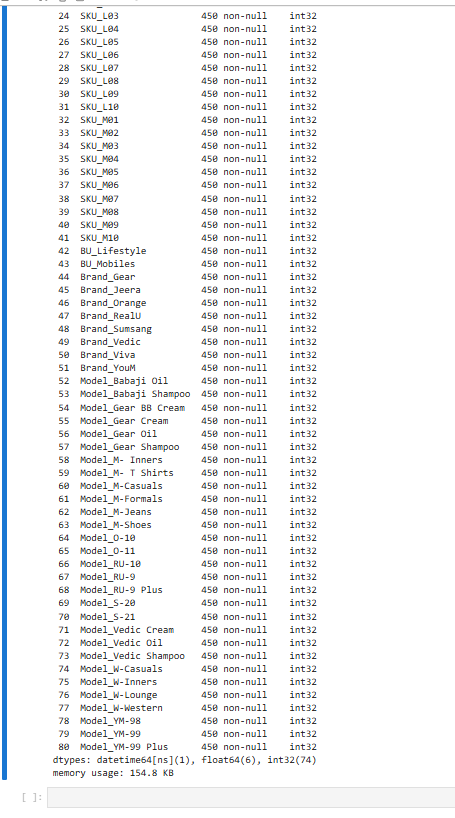
1. **Label encoding:**

Assigning the label to each of the value in the columns, this method of encoding is used when the dependent variable is categorical in nature.

**Fig-17: Data before Transformation**



**Fig-18: Data After Transformation**



**Fig-19: Data After Transformation**

**Conclusion**

**Key Findings from Descriptive Analytics and Data Visualizations**

Descriptive analytics and data visualizations help us understand our data better. They show us the average values, how spread out the data is, and any unusual points. Visuals like charts and graphs make it easy to see patterns, trends, and relationships between different parts of the data.

**Importance of Data Preprocessing Steps like Standardization and One-Hot Encoding**

Data preprocessing steps like standardization and one-hot encoding are important for making our data ready for analysis and machine learning. Standardization scales the data so that different features have similar ranges, which helps the algorithms work better. One-hot encoding turns categories into numbers that the machine can understand, ensuring that the algorithms interpret the data correctly. These steps help in building accurate and reliable models.