Basic Statistics

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

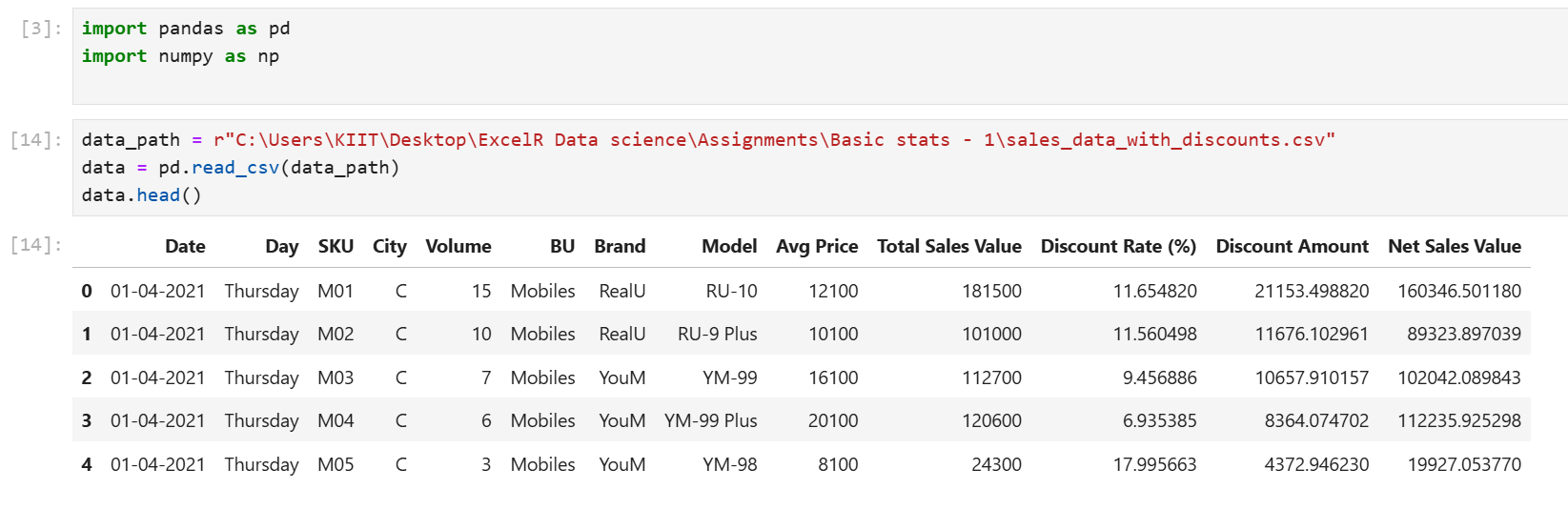
#### Introduction

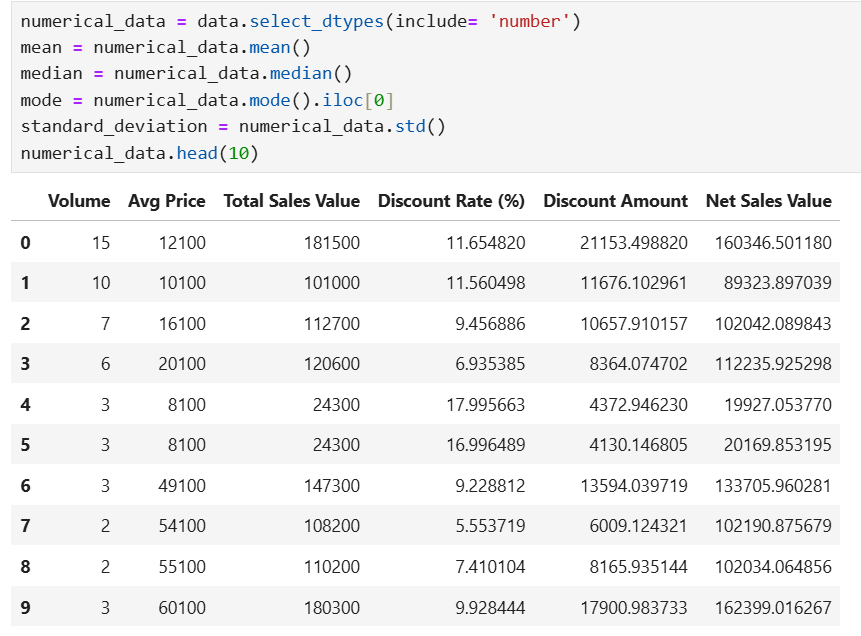
* To perform descriptive analytics, visualize data distributions, and preprocess the dataset for further analysis.

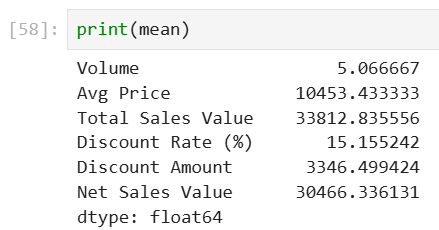
#### Descriptive Analytics for Numerical Columns

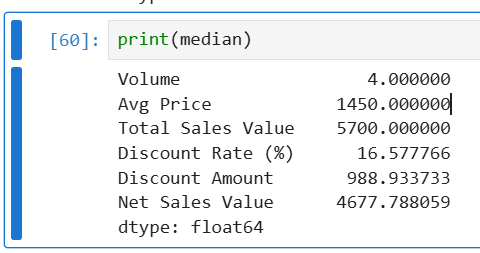
* Objective: To compute and analyze basic statistical measures for numerical columns in the dataset.

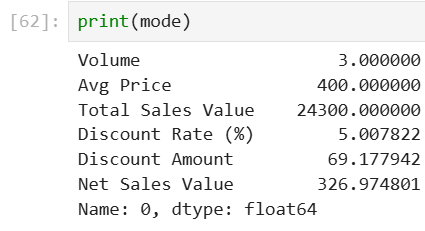
I have loaded the dataset using pandas and have calculated Mean, Mode, Median and Standard deviation for the numerical columns.

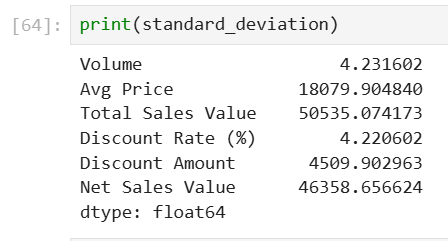












**Interpretations:**

1. Volume column:
2. Mean : 5.06
3. Median : 4
4. Mode : 3
5. Standard deviation: 4.23

* Since the mean is greater than median, we can say that the volume column is right skewed. More data points are greater than 4 which is causing the mean to be slightly higher than median.
* The standard deviation is closer to mean so the spread of the data points is not much.
* In the ordered data, the mode value is 3, which means that most of the volume value is below the median.
* Since the Volume column represents the quantity of goods sold, we can say that on an average a customer purchases 5 products, with most of the customers purchasing 3 units of products.

1. **Avg price:**
2. Mean: 10543.43
3. Median: 1450
4. Mode: 400
5. Standard deviation: 18079.9

* The large gap between mean and median shows a strong right skew, indicating that while many products are lower-priced, a few high-priced items are pulling the mean up substantially.
* The mode being lower than both the median and mean suggests that the majority of products are priced on the lower end, with prices clustering around 400.
* The very high standard deviation shows significant price variability, meaning that product prices fluctuate widely, from low-cost items to very expensive ones.

1. **Total sales Value:**
2. Mean: 33,812.44
3. Median: 5,700
4. Mode: 24,300
5. Standard deviation: 50535.07­

* The high difference between Mean and Median indicates a very strong right skew. This suggests that the Total sales value are upper – priced that is the reason for the high mean value.
* The mode is higher than the Median suggests that majority of Total sales Value is on the upper end with Values clustering around 24,300.
* The Standard deviation being higher indicates that there is a large variability in the total sales due to some high value transactions.

1. **Discount rate:**
2. Mean: 15.16%
3. Median: 10.58%
4. Mode: 5.08%
5. Standard deviation: 4.22%

* The mean being higher than Median indicates a slight right skew, meaning that maximum transactions have a higher discount rate.
* The Mode is lower than median, which indicates that mostly the discount rate was 5.08%.
* The standard deviation is moderately high indicating variability in the dataset.

1. **Discount amount:**
2. Mean: 3,346.5
3. Median: 988.93
4. Mode: 69.18
5. Standard deviation: 4,509.29

* A large gap between the mean and median points to a right-skewed distribution, with a few high discount amounts inflating the average.
* The mode being much lower than both the mean and median suggests that most discount amounts are small, with a few larger discounts skewing the data.
* The high standard deviation indicates significant variability in discount amounts.

1. **Net sales value:**
2. Mean: 30,466.34
3. Median: 4,677.78
4. Mode: 326.97
5. Standard deviation: 46,358.66

* The mean is significantly higher than the median, indicating a strong right skew. This suggests that while most transactions have a moderate net sales value, a few very high-value transactions inflate the mean.
* The mode being very low relative to the median and mean suggests that many transactions have small net sales values, with only a few high-value transactions.
* The very high standard deviation reflects a wide range in net sales values, showing that some transactions contribute a much larger amount to total sales.

#### Data Visualization

* **Objective**: To visualize the distribution and relationship of numerical and categorical variables in the dataset.
* **Histograms**:
  + Plot histograms for each numerical column.
  + Analyze the distribution (e.g., skewness, presence of outliers) and provide inferences.
* **Boxplots**:
  + Create boxplots for numerical variables to identify outliers and the interquartile range.
  + Discuss any findings, such as extreme values or unusual distributions.
* **Bar Chart Analysis for Categorical Column:**
  + Identify categorical columns in the dataset.
  + Create bar charts to visualize the frequency or count of each category.
  + Analyze the distribution of categories and provide insights.

#### Standardization of Numerical Variables

* Objective: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Steps:
  + Explain the concept of standardization (z-score normalization).
  + Standardize the numerical columns using the formula: z=x-mu/sigma
  + ​Show before and after comparisons of the data distributions.

#### Conversion of Categorical Data into Dummy Variables

* Objective: To transform categorical variables into a format that can be provided to ML algorithms.
* Steps:
  + Discuss the need for converting categorical data into dummy variables (one-hot encoding).
  + Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.
  + Display a portion of the transformed dataset.

#### Conclusion

* Summarize the key findings from the descriptive analytics and data visualizations.
* Reflect on the importance of data preprocessing steps like standardization and one-hot encoding in data analysis and machine learning.