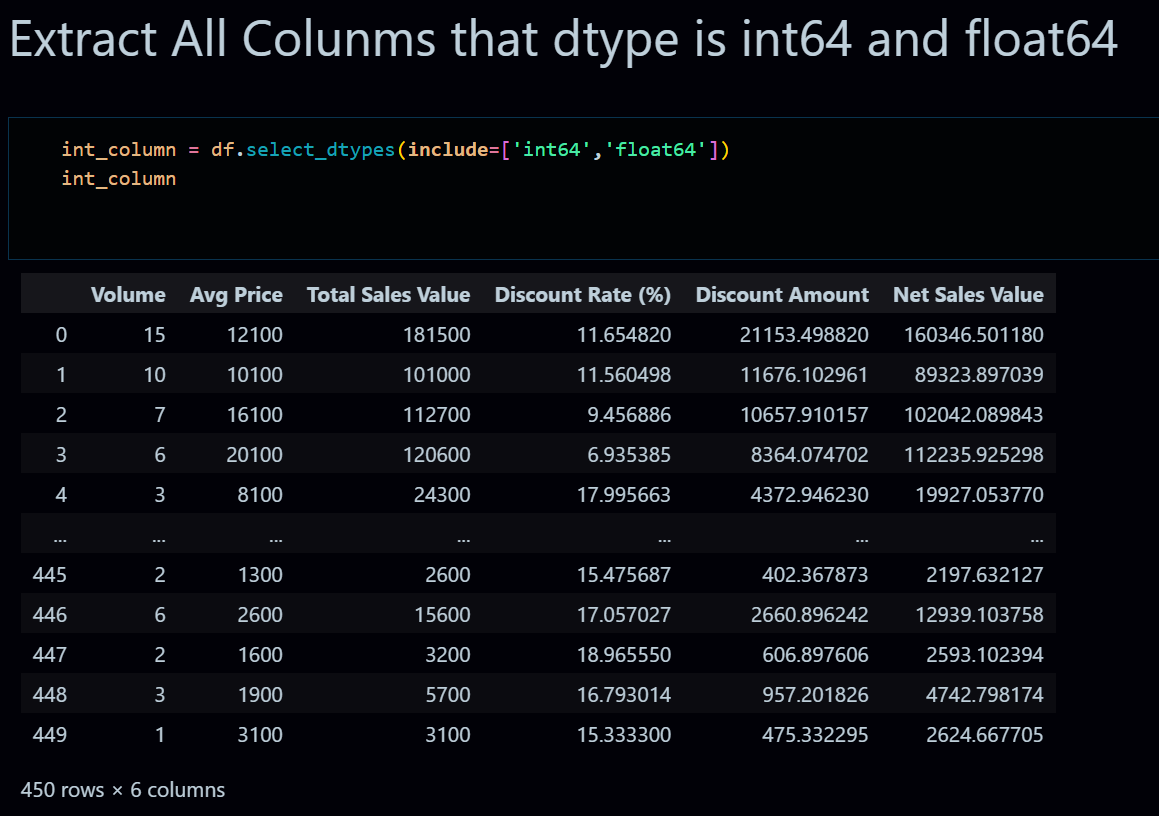
Basic Statistics

### 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.
* Steps:
  + Load the dataset into a data analysis tool or programming environment (e.g., Python with panda’s library).
  + 
  + Identify numerical columns in the dataset.
  + -> *In the given dataset the numerical columns name is Volume, Avg Price, Total Sales Value, Discount Rate (%), Discount Amount, Net Sales Value.*
  + **

Calculate the mean, median, mode, and standard deviation for these columns.

->



Provide a brief interpretation of these statistics.

**Mean:**

The average number of units sold (Volume) is approximately 5.

The average price of products (Avg Price) is about 10,453.

The average total sales value is around 33,813.

The average discount rate applied is roughly 15.16%.

The average discount amount given is about 3,346.

The average net sales value (after discounts) is approximately 30,466.

**Median**:

The median number of units sold is 4, indicating that half the sales transactions involved selling 4 or fewer units.

The median average price is 1,450, which suggests a skew in price distribution since the mean is much higher.

The median total sales value is 5,700, showing that many transactions were relatively smaller in value.

The median discount rate is 16.58%, close to the mean, indicating a relatively consistent discount strategy.

The median discount amount is 988.93, lower than the mean, indicating a few high-discount transactions.

The median net sales value is 4,677.79, suggesting that smaller transactions are more common.

**Mode:**

The most frequent number of units sold in a transaction is 3.

The most common average price is 400, indicating a popular low-cost product.

The most frequent total sales value is 24,300, suggesting certain high-value transactions were repeated.

The most common discount rate is around 5%.

The most common discount amount is 69.18.

The most frequent net sales value is 326.97.

**Standard Deviation:**

There's a moderate variation in the number of units sold (Volume), with a standard deviation of 4.23.

The average price and total sales value show high variability, as indicated by their large standard deviations (18,079.90 and 50,535.07, respectively).

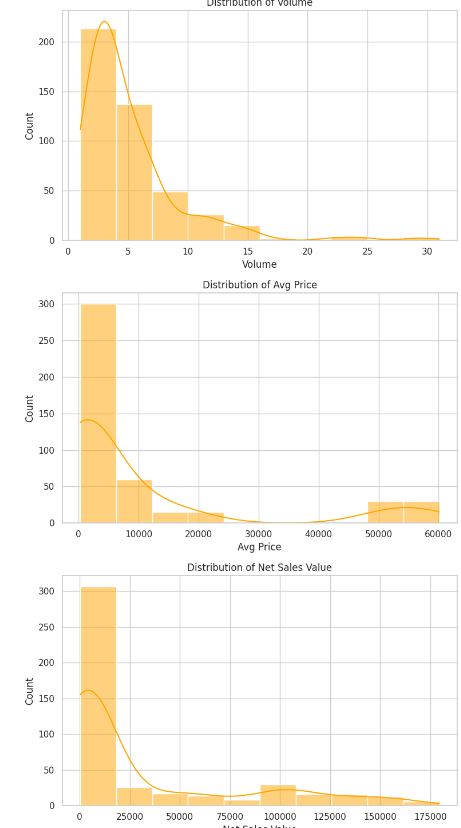
The discount rate has a low variation, suggesting a consistent discount strategy.

Discount amounts and net sales values also show considerable variation, indicating diverse sales scenarios.

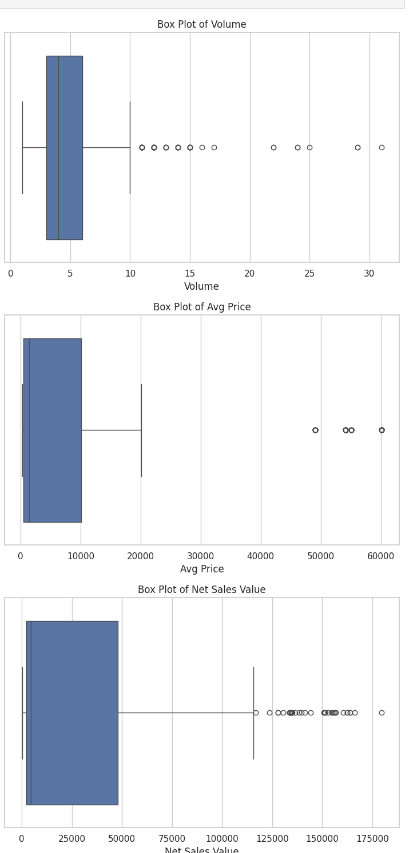
#### Data Visualization

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

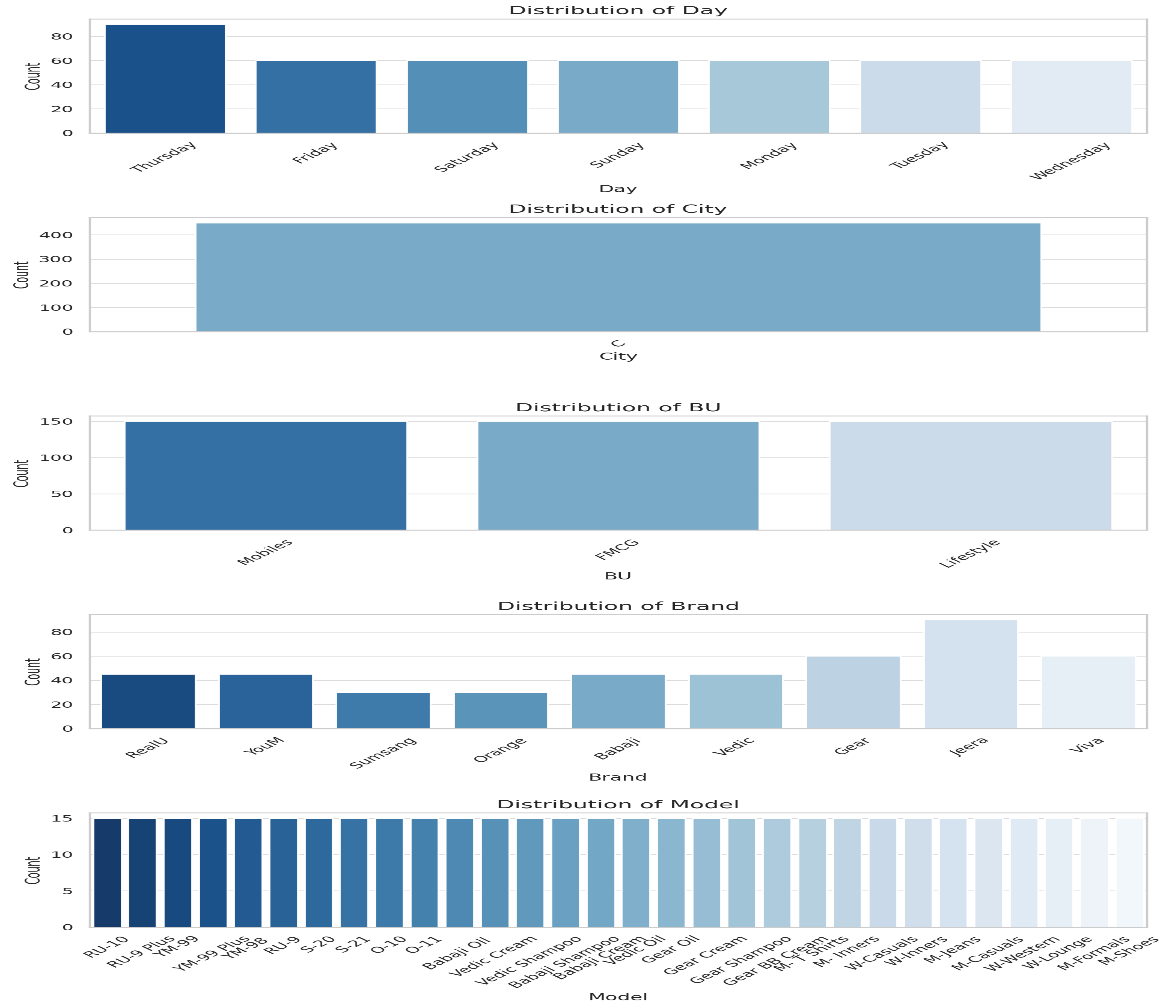
Plot histograms for each numerical column.

* + 

Analyze the distribution (e.g., skewness, presence of outliers) and provide inferences.

* +  All three variables are **right-skewed**, which is common in retail or sales data, where a few high-volume or high-priced items tend to dominate.
  +  Outliers: Each distribution has a long right tail, suggesting the presence of outliers or extreme values, which could be worth further investigation.
  +  Data **Characteristics**: The concentration of data in the lower ranges for all three variables suggests that most of the data points are clustered in lower ranges (small-volume, low-price, and low-sales items), but there are a few that are significantly larger, which might warrant special attention or further analysis.
* **Boxplots**:
  + Create boxplots for numerical variables to identify outliers and the interquartile range.
  + 
  + Discuss any findings, such as extreme values or unusual distributions.
  +  **Outliers**: Each of the three variables exhibits a large number of outliers. These outliers could indicate special cases like premium items, bulk purchases, or highly successful products.
  +  **Skewness**: The data is right-skewed across all variables, meaning the majority of the data points are concentrated at lower values, while a few higher values create long tails.
  +  **Further Investigation**: The presence of so many outliers, especially in net sales and average price, suggests that further investigation may be required. These outliers could represent anomalies, high-end items, or specific market segments that behave differently from the rest of the data.
* **Bar Chart Analysis for Categorical Column:**
  + Identify categorical columns in the dataset.

The dataset contains several categorical columns. Here are the ones that are identified based on the data types and content:

* **Date**: Represents dates (could be treated as categorical if grouped by month/year).
* **Day**: Categorical, representing the day of the week.
* **SKU**: Categorical, representing stock-keeping units.
* **City**: Categorical, representing different cities.
* **BU**: Categorical, representing business units (e.g., product types like "Mobiles").
* **Brand**: Categorical, representing the brand of the products.
* **Model**: Categorical, representing the specific model of a product.
  + Create bar charts to visualize the frequency or count of each category.
  + 
  + Analyze the distribution of categories and provide insights.
  +  **Day Distribution**:
  + **Observation**: The most frequent transactions occur on **Thursday**, with over 80 occurrences, while other days of the week have relatively balanced but lower transaction counts.
  + **Insight**: There may be specific sales promotions or market behavior causing a spike in sales on Thursdays, which could be investigated further.
  1.  **City Distribution**:

**Observation**: All transactions seem to occur in a single city (denoted as **C**).

**Insight**: This suggests the data is either limited to one geographical location, or city data is anonymized or consolidated. Expanding the dataset to multiple cities might provide a better understanding of geographical trends.

* 1.  **BU (Business Unit) Distribution**:

**Observation**: The sales data is evenly distributed across three business units: **Mobiles**, **FMCG**, and **Lifestyle**, each with roughly equal representation (~150 transactions).

**Insight**: This balanced distribution shows a well-rounded dataset with respect to product categories, allowing for a diverse analysis across different industries.

* 1.  **Brand Distribution**:

**Observation**: The **Gear** and **Jeera** brands have the highest frequency of transactions (~60-70 occurrences), while others like **RealU**, **YouM**, and **Sumsang** have lower counts (~20-40).

**Insight**: This suggests that the **Gear** and **Jeera** brands may be more popular or dominant in this dataset, possibly due to higher sales or wider product ranges.

* 1.  **Model Distribution**:

**Observation**: The transaction counts for different models are relatively even, with each model having around 10-15 occurrences.

**Insight**: The even distribution indicates that there is no single dominating model, implying a diversified sales portfolio with similar demand across different models.

#### Standardization of Numerical Variables

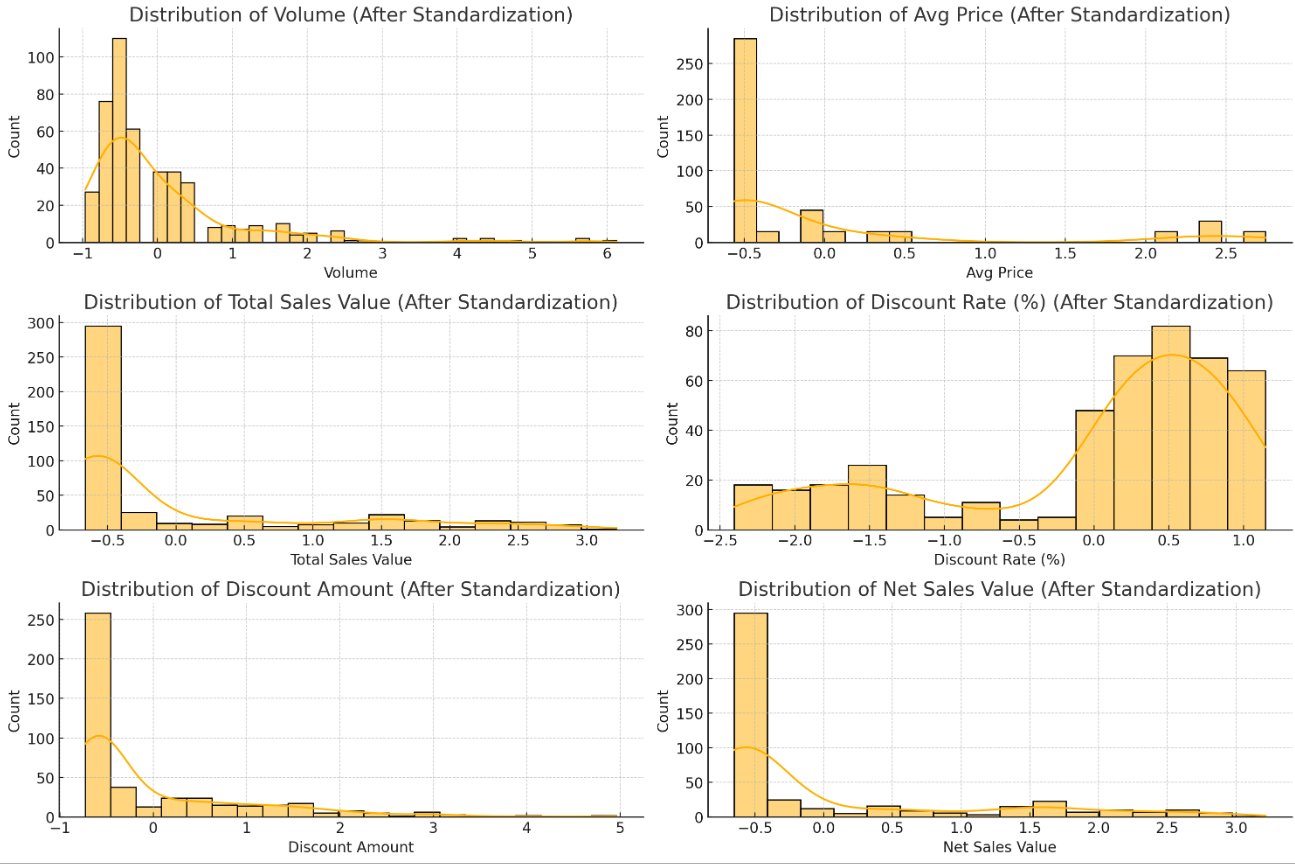
* Objective: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Steps:
  + Explain the concept of standardization (z-score normalization).

A **z-score** (or standard score) measures how many standard deviations a data point xxx is away from the mean μ of the dataset. It provides a way to standardize data across different scales, allowing for comparison.

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

* + Show before and after comparisons of the data distributions.

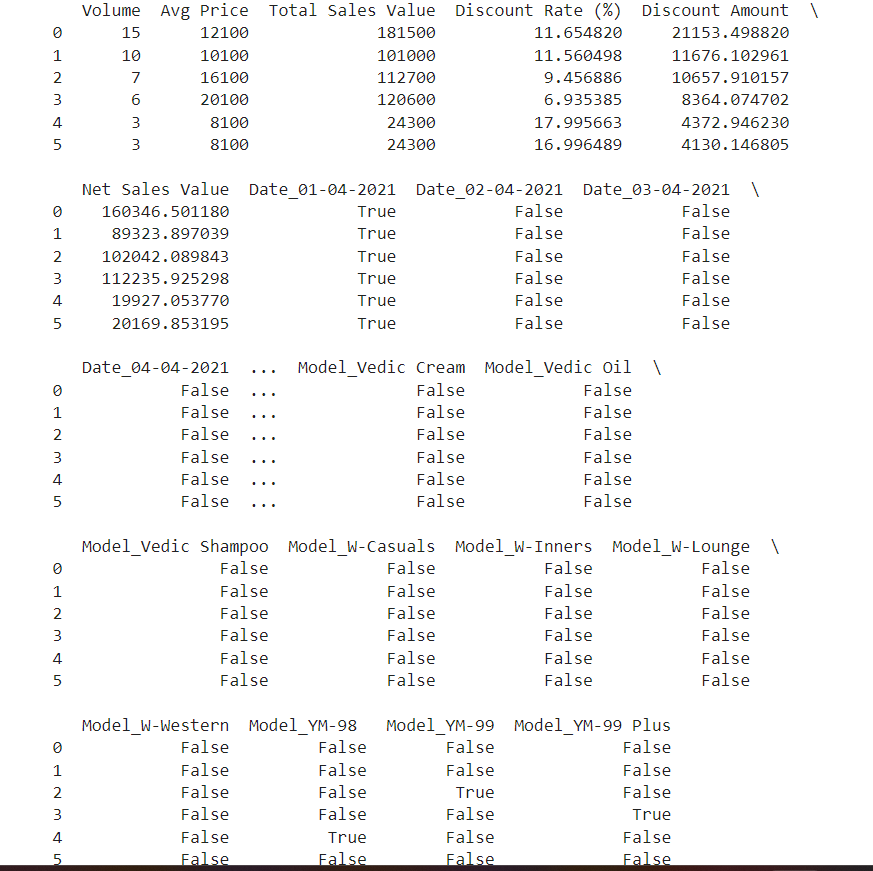
 The plots above show the distributions of the numerical columns before standardization.

 The plots following show the distributions of the numerical columns after applying z- score standardization.

#### 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).
* Machine Learning Models: Most ML algorithms (like linear regression, logistic regression, decision trees, etc.) require numerical input. Categorical variables are nominal (having no intrinsic order) or ordinal (having a clear ordering) and must be converted to numbers.
* Avoiding Misinterpretation: If you assign arbitrary numerical values to categories (e.g., 1, 2, 3), models may interpret these numbers as having an order or relationship that doesn't exist, leading to poor performance.
* Handling Multiple Categories: One-hot encoding creates a new binary column for each unique category level, allowing the model to treat each category independently.
  + Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.

Applying One-Hot Encoding One-hot encoding creates a new binary column for each category level in a categorical variable. For example, if you have a column Color with values Red, Blue, and Green, one-hot encoding would create three new columns: Color\_Red, Color\_Blue, and Color\_Green.

* + 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.

Also Add the repo like where all python work Are done <https://github.com/Rushi-code1/Data_Science_Practice/blob/main/ASSIGNMENT/Basic-Stat.ipynb>