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 pandas library).
  + Identify numerical columns in the dataset.
  + Ans:

| **Volume** | **Avg Price** | **Total Sales Value** | **Discount Rate (%)** | **Discount Amount** | **Net Sales Value** |
| --- | --- | --- | --- | --- | --- |

* + Calculate the mean, median, mode, and standard deviation for these columns.
  + Ans:
  + Mean median mode
  + Volume 5.066667 4.000000 3.000000
  + Avg Price 10453.433333 1450.000000 400.000000
  + Total SalesValue 33812.835556 5700.000000 24300.000000
  + Discount Rate(%) 15.155242 16.577766 5.007822
  + Discount Amount 3346.499424 988.933733 69.177942
  + Net Sales Value 30466.336131 4677.788059 326.974801
  + Provide a brief interpretation of these statistics.
  + Ans:
  + This sales data shows that, on average, 5 units are sold daily at a price of ₹10,453, with total sales of ₹33,812. Discounts average 15.16%, with a discount amount of ₹3,346. After discounts, the net sales value averages ₹30,466. Most days, 3 units are sold, with frequent prices around ₹400 and net sales often around ₹327

#### 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.
  + Ans:
  + Volume : right skewed (in the left ),outliers are in the right
  + Avg price : right skewed (in the left ),outliers are in the right
  + Discount Rate (%) : left skewed (in the right),outliers are in the left
  + Discount Amount: right skewed (in the left ),outliers are in the right
  + Total Sales Value: right skewed (in the left ),outliers are in the right
  + Net Sale Value : right skewed (in the left ),outliers are in the right
* **Boxplots**:
  + Create boxplots for numerical variables to identify outliers and the interquartile range.
  + Discuss any findings, such as extreme values or unusual distributions.

**Ans**:

Volume : Outliers are in upper extreme,most datas are in between 2 and 6

Avg price : Outliers above ₹40,000 ,most prices are between ₹0 and ₹10,000

Discount rate : Outliers below the lower ,most rates are between 12% and 18%

Discount amount: Outliers in upper extreme , above 10000,

Most values from 0 to 5000

**Total Sales Value**: Outliers above ₹100,000 , upper extreme most values fall between ₹0 and ₹50,000.

**Net Sales Value**: Outliers above ₹100,000 suggest very high net sales; most values range from ₹0 to ₹40,000, with a median of ₹5,000.

* **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.
  + Ans:
    - We can see the Thursday are most in the data comparing to other
    - Most sold brand is jira

#### Standardization of Numerical Variables

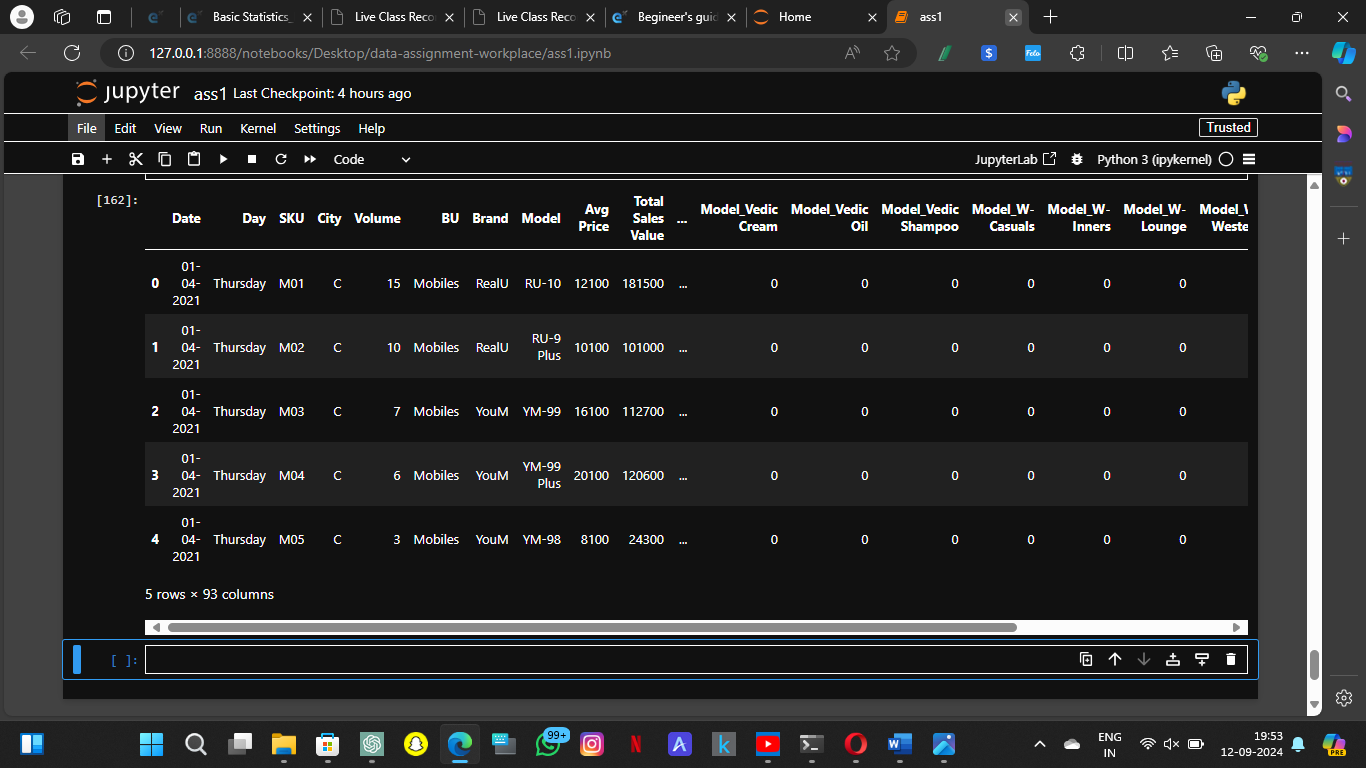
* Objective: To scale numerical variables for uniformity, improving the dataset’s suitability for analytical models.
* Steps:
  + Explain the concept of standardization (z-score normalization).
  + Ans:
    - Standardization scales data so that it has a mean of 0 and a standard deviation of 1. This makes it easier to compare features with different ranges. It’s done by subtracting the mean from each value and dividing by the standard deviation.
  + Standardize the numerical columns using the formula: z=x-mu/sigma
  + ​Show before and after comparisons of the data distributions.
  + **Before:**
  + Mean median mode
  + Volume 5.066667 4.000000 3.000000
  + Avg Price 10453.433333 1450.000000 400.000000
  + Total SalesValue 33812.835556 5700.000000 24300.000000
  + Discount Rate(%) 15.155242 16.577766 5.007822

**After :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Volume** | 6.118562e-17 | -0.252352 | -0.488932 |
| **Avg Price** | 3.157968e-17 | -0.498534 | -0.556675 |
| **Total Sales Value** | 1.578984e-17 | -0.556923 | -0.188452 |
| **Discount Rate (%)** | -3.552714e-17 | 0.337418 | -2.406935 |
| **Discount Amount** | -5.526444e-17 | -0.523335 | -0.727503 |
| **Net Sales Value** | 1.973730e-17 | -0.556902 | -0.650858 |

#### 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).
  + **Ans**:
    - This process involves creating binary columns for each category in the original categorical variable
  + Apply one-hot encoding to the categorical columns, creating binary (0 or 1) columns for each category.
  + Display a portion of the transformed dataset.
  + **Ans:**



#### 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.

Ans:

From the descriptive analytics we can see that sales data is skewed in some , there are some outliers.

preprocessing steps like standardization and one-hot encoding are essential for preparing the data for accurate machine learning.