#### **1.Descriptive Analytics for Numerical Columns:**

* Volume

Mean of volume is: 5.066666666666666

Standard deviation of Volume is: 4.231602391213926

Median of Volume is: 4.0

Mode of Volume is: 0 3

Name: Volume, dtype: int64

Analysis:

If mean > median → data is **right-skewed** (few large sales).

Standard Deviation=4.23-> Moderate variation in quantities sold

Mode=3->Most frequent volume sold units

* AvgPrice

Mean of Average price is: 10453.433333333332

Standard deviation of Average Price: 18079.90483993645

Median of Average price is: 1450.0

Mode of Average price is: 0 400

1 450

2 500

3 1300

4 8100

Name: Avg Price, dtype: int64

If Mean>Median- data and High standard deviation -> Strong **right-skewed** (few large sales)

Mode =400 shows many low cost product while some very expensive ones inflate the average

* Total Sales

Mean of Total Sales Value: 33812.83555555555

Standard deviation of Total sales value: 50535.07417255328

Median of Total Sales Value is: 5700.0

Mode of Total Sales Value is: 0 24300

Mean is much grater than Median ->Heavy outliers(bulk/High Value Sales)

Mode=24300 is common sales value,possibly for specific period

Stddev=50535.07->data is widespread not concentrated

* Discount Rate%

Mean of Discount Rate % is : 15.155241895330914

Standard deviation of Discount rate % is: 4.220602153325511

Median of Discount Rate% is: 16.57776564929597

Mode of Discount Rate% is: 0 5.007822

1 5.055218

2 5.059801

3 5.072124

4 5.084108

...

445 19.940610

446 19.965340

447 19.976879

448 19.982257

449 19.992407

Mean=15.15 and Median=16.57->Reasonable Symmetry

Mode=5.01% suggests many products have small fixed discount

Std dev=4.22%->fair variation in discounting strategy

* Discount Amount

Mean of Discount Amount is: 3346.499424342415

Standard deviation of Discount amount is: 4509.9029633539985

Median of Discount is: 988.9337332382784

Mode of Discount is: 0 69.177942

1 73.025199

2 93.649161

3 94.682735

4 102.705757

...

445 17900.983733

446 21153.498820

447 21496.675367

448 25328.224204

449 25738.022194

Large difference between **mean (3,346)** and **median (988)** → High-end discounts exist.

**Std Dev = 4,509** → Big fluctuations, likely tied to Avg Price.

* Net Sale Value

Mean of Net sales value is: 30466.336131270138

Standard deviation of Net sales volume is: 46358.656624066934

Median of Net sales value is : 4677.788059301853

Mode of Net sales value is : 0 326.974801

1 330.822058

2 466.350839

3 485.317265

4 496.607104

...

445 162603.186379

446 163915.971000

447 163967.080070

448 166263.161348

449 179507.479049

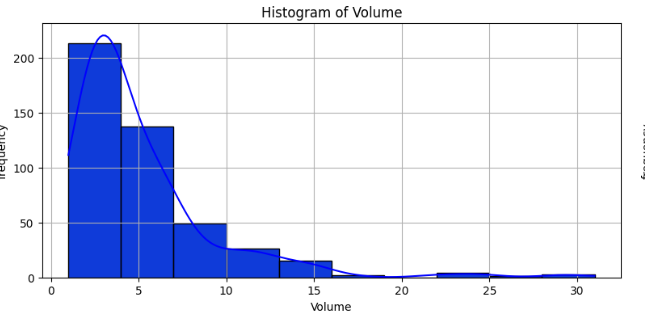
High **mean (30K)** and **very high Std Dev (46K)** → Substantial variability.

**Median is much lower (4.7K)** → Most sales are low value; few very high-value sales pull the mean up.

**Mode = 326.97** → A common post-discount price point for one or more SKUs.

#### Data Visualization

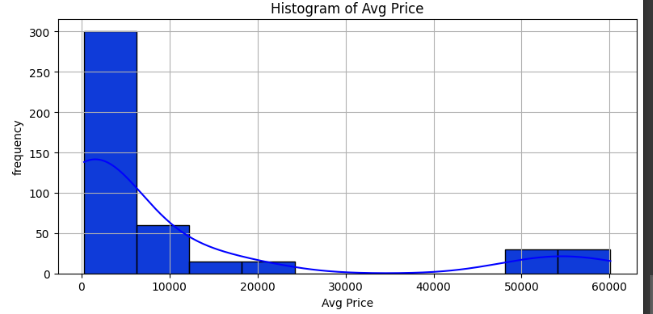
* **Histograms**:
* **Graphs**:



1.Volume

* Right Skewed or Positive Skewness Present
* Outliers: few High volumes of sales from the rest
* Interference: Most Products are sold in low volume only few have significantly high sales

2. Avg Price

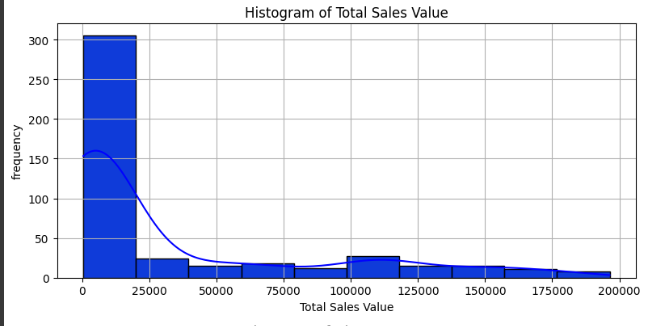


Slightly Right Skewed

Outliers: Higher Price less frequent

Interference: Majority of products are Moderate price, few Items are in high price.

3.Total Sales Value

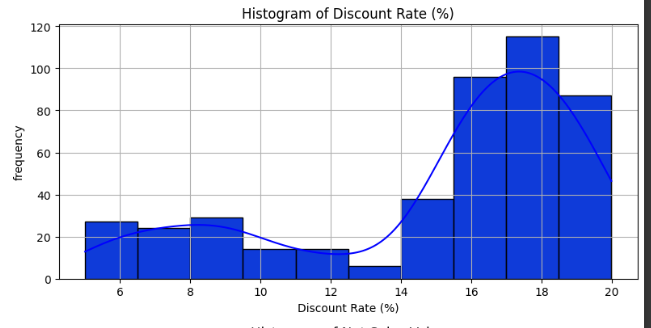


Right Skewed

Outliers: Several High Value Transaction

Interference: Most of the Sales are in range where Higher sales may be because of Higher volume or Higher Price.

4. Discount Rate%

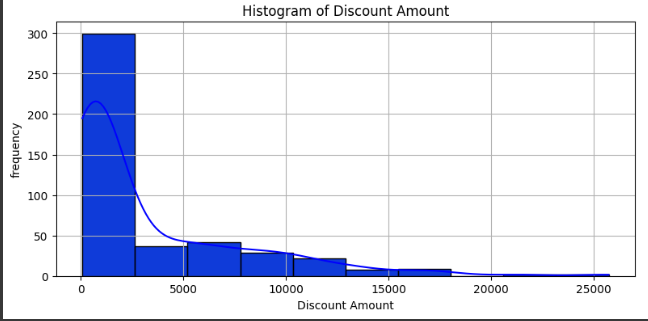


Left Skewed

Outliers: Some has very low Discount rate

Interference: Discounts are common. generally falling in mid-range and very low discounts are rare.

5.Discount Amount:

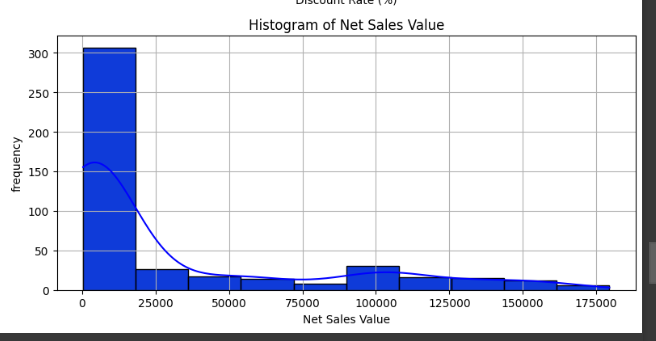


Strongly Right Skewness

Outliers: A few Extreme high discount amount

Interference: Most discounts are small value .High discounts are applied in high transaction

6. Net Sales Value



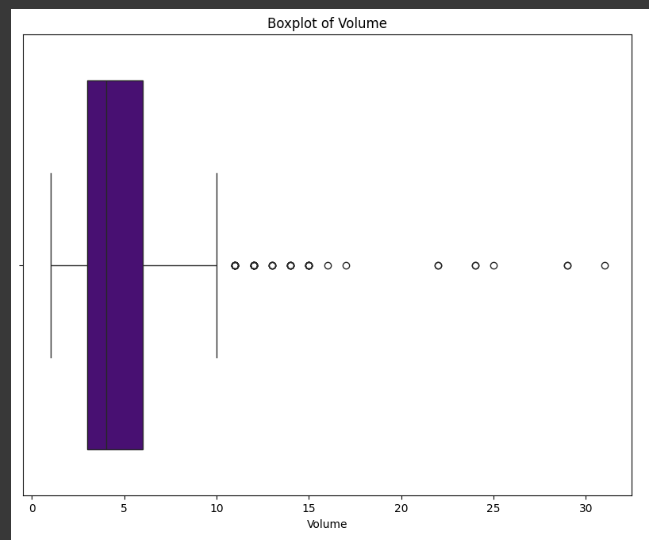
Right Skewness

Outlier: High Volume Sales

Interference: Majority of Sales Fall into lower range may be because of Lower Volume or Price

* BOX PLOT

1.Volume

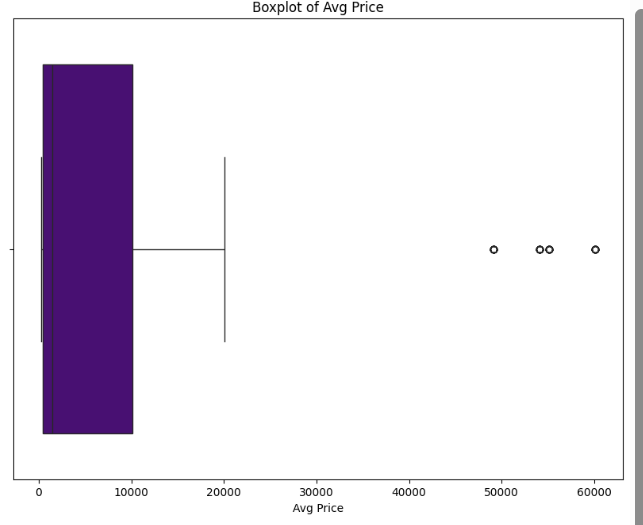


Inter Quartile Range: The central 50 % of data lies in narrow band of small volume.

Outlier: several extreme value above upper whiskers.

Insight: Majority of product sell in low value. few have exceptionally high volume may be due to bulk order or SKUS.

Avg Price:

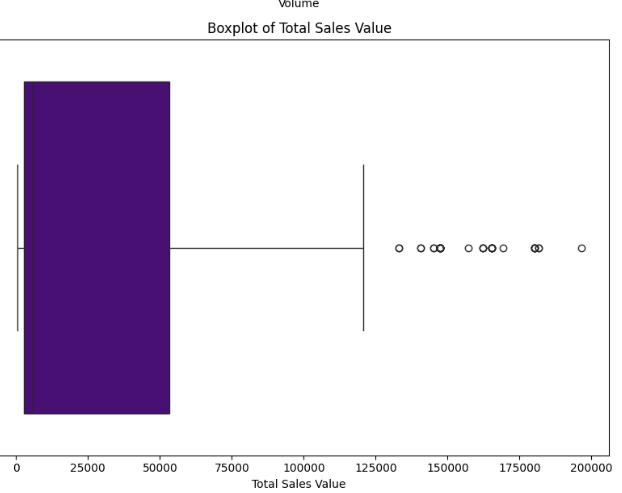


Inter Quartile Range: Price are mostly concentrated in mid-range bucket.

Outlier: High end price are outliers

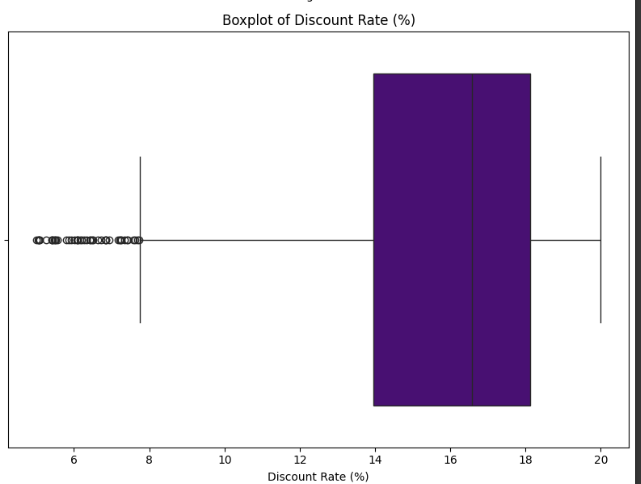
Insight: Most Product are affordably priced. Few high pried products differs significantly from norm.

Total SalesValue:



Right Skewed Distribution with High value of outliers, reflecting big transaction

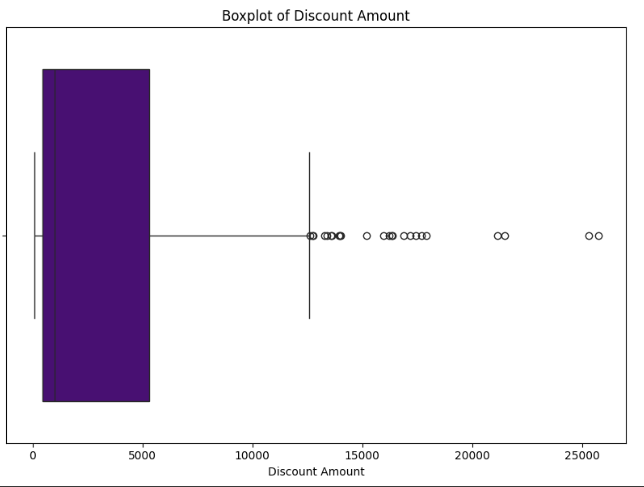
Discount Rate %:



Most Discount fall with in tight Range

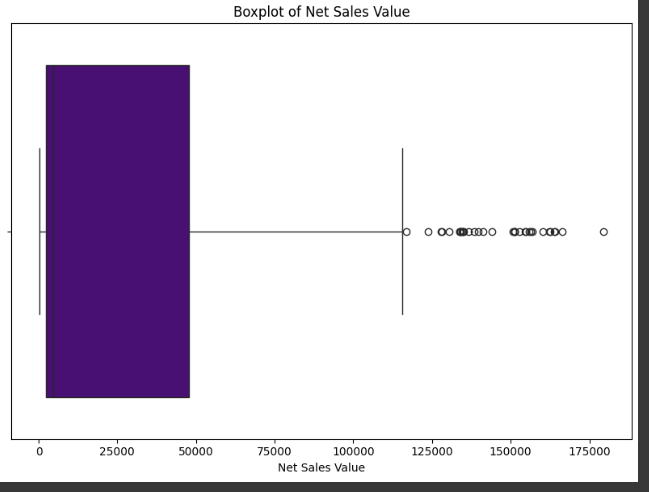
Several high discount rate(outliers) found in Left Skewed.

Discount Amount:

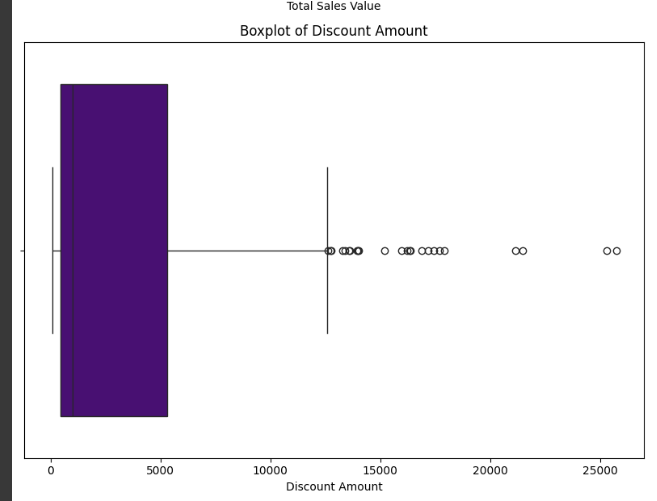


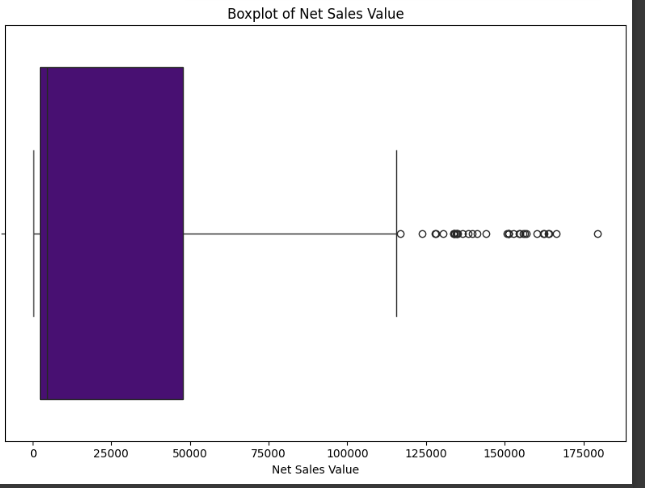
Right Skewed with Significant outliers-likely linked to high priced Products

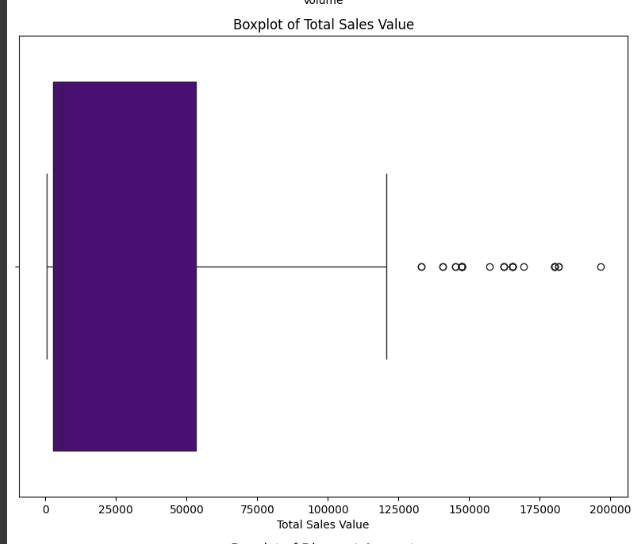
**Net Sales Value**::



Most sales are moderate with high outliers due to expensive transaction



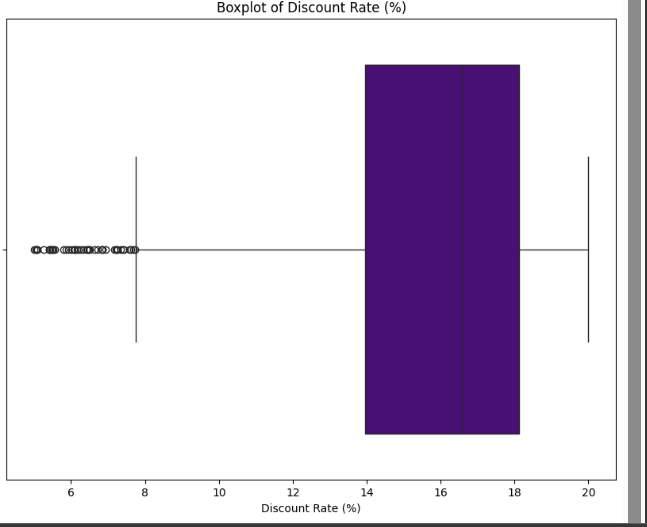




IQR: middle 50 % of sales are tightly grouped

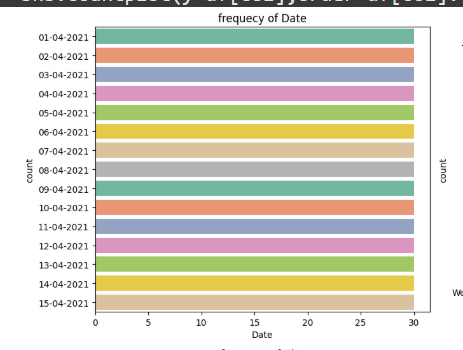
Outliers: many values are extremely high

Insight: A small number of transactions contribute dipropionate to revenue -either due to large volume, high price or both

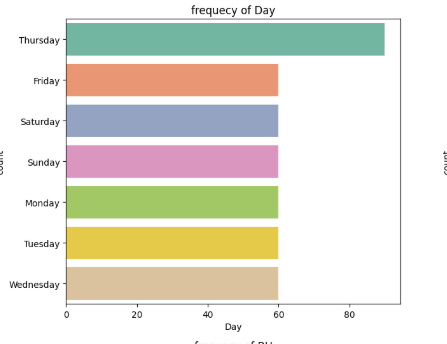


Barchart:

Date: Data is Evenly distributed across Date



2.Day:



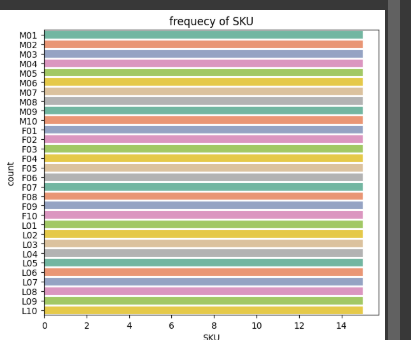
Mode: Most Frequent day (Thursday)

Data is uniformly distributed across category

3.City:

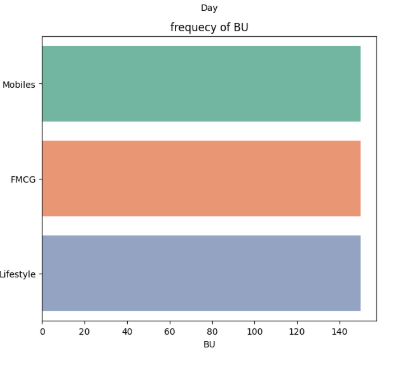
City C has highest frequency. Potential bias towards C

4.SKU:



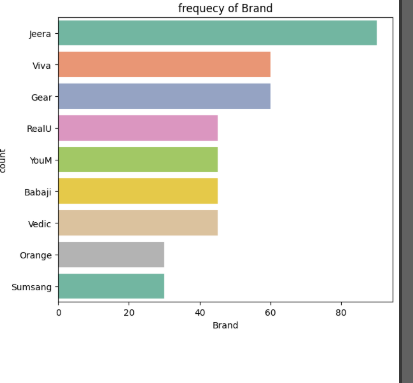
Data is evenly Distributed for SKU

5.BU



Data is evenly distributed among three Bus that is FMCG, Mobiles and Lifestyle

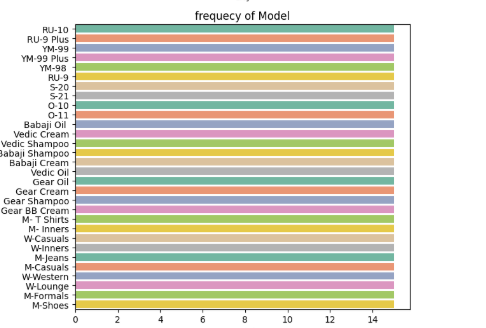
Brand:



Top brands: **Jeera followed by Viva and Gear**

**Jeera** is slightly more frequent than **others Brands**, indicating a possibly stronger presence.

Models:



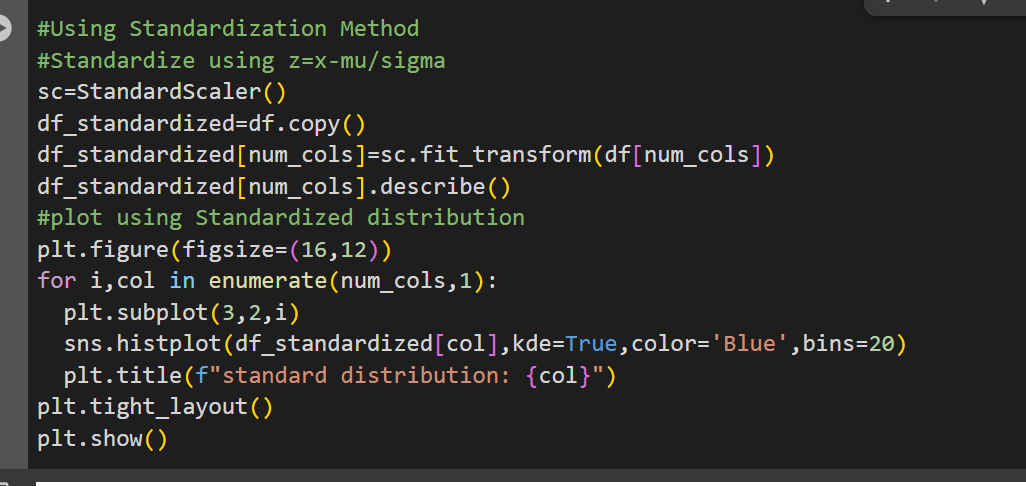
All Models are evenly distributed across the graph

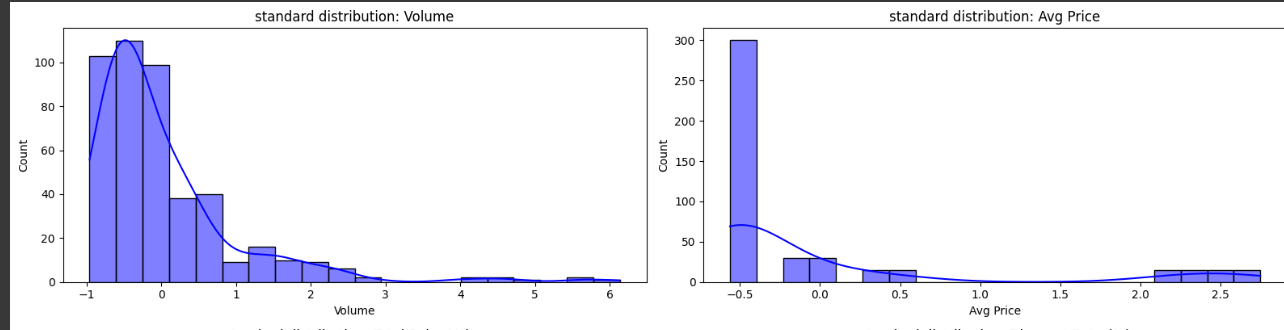
Standardization Method: uses Z Score for Scaling in such a way that mean=0 and Standard deviation=1.This helps analytical models like regression-means and PCA to treat all figure equally regardless of all scale

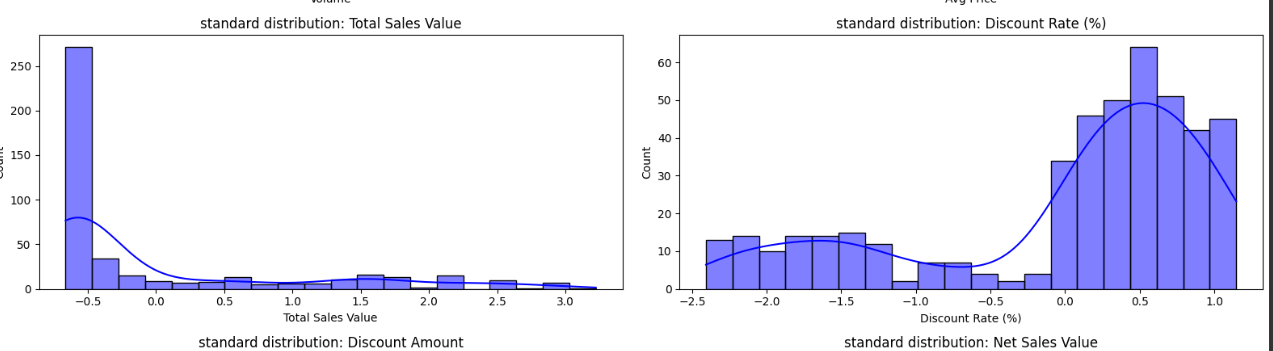
Z-Score=x-Mu/Sigma

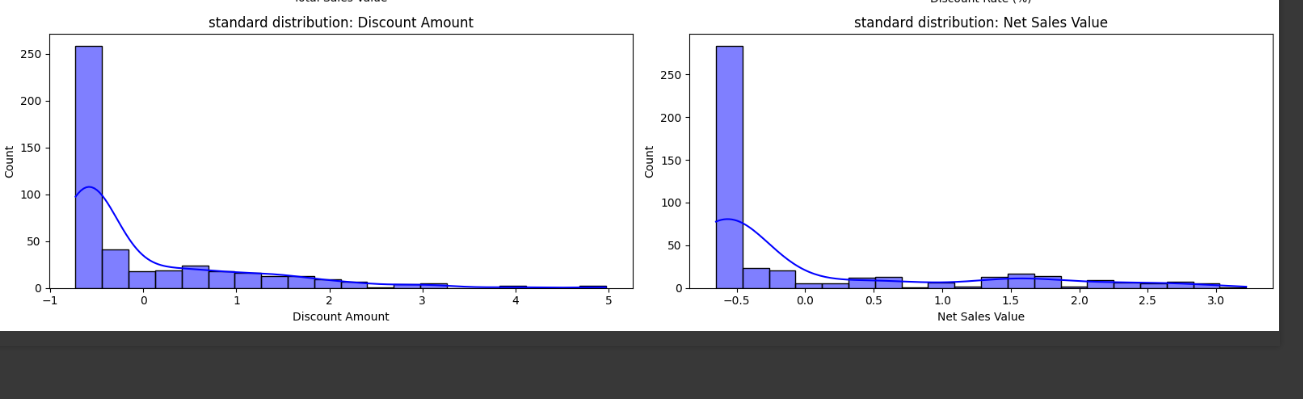
Mu= mean of the column

Sigma=Standard deviation of column









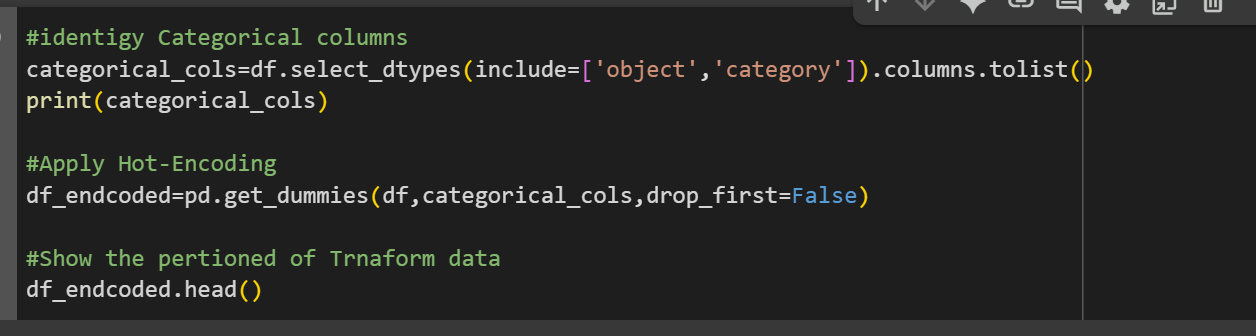
* Why Hot Encoding is Required

Machine Learning typically require numerical input. Categorical data such as

Brand,SKU,City etc are converted into a format where

Each category becomes a separate column

Each row has 1 in the column corresponding to category else o elsewhere



Conclusion:

**Numerical Columns**

1. **Volume**
   * Mean = 5.06, Median = 4.0 → *Right-skewed distribution*
   * Moderate variation; mode = 3 units
   * Most products sell in low volume; few bulk sales inflate mean
2. **Average Price**
   * Mean = ₹10,453, Median = ₹1,450 → *Strong right-skew*
   * High standard deviation → vast price differences
   * Most products are moderately priced; a few expensive ones skew average
3. **Total Sales Value**
   * Mean = ₹33,813, Median = ₹5,700 → *Right-skewed*
   * High variation; large transactions skew data
   * Most transactions are moderate; few large ones dominate sales totals
4. **Discount Rate %**
   * Mean ≈ 15.15%, Median ≈ 16.58% → *Reasonably symmetrical*
   * Most discounts fall in a moderate range
   * Frequent fixed discount levels (e.g., ~5%)
5. **Discount Amount**
   * Mean = ₹3,346, Median = ₹989 → *Strong right-skew*
   * Large discounts typically linked to high-value products
6. **Net Sales Value**
   * Mean = ₹30,466, Median = ₹4,678 → *Heavy right-skew*
   * Few large transactions (volume or price driven) pull mean higher
   * Majority of sales are smaller in value

**Visualizations**

**Histograms & Boxplots Insights**

* All monetary columns show **right-skewed distributions**, indicating **outliers and unequal distributions**.
* **Discount Rate %** shows **left-skewness**, suggesting a norm around moderate discounts but a few very low ones.
* **Boxplots** confirm:
  + Outliers are common in volume, price, sales, and discount amounts.
  + Most values are tightly grouped in the lower or mid-range with exceptions.

**Bar Charts**

* **City**: City C dominates — possible market concentration.
* **Day**: Sales fairly uniform; peak on Thursdays.
* **Brand**: "Jeera" leads — indicating strong brand preference.
* **BU & SKU**: Fairly balanced across categories and items.

**Importance of Data Preprocessing Steps**

**1. Standardization (Z-Score Scaling)**

* Formula: Z = (x - mu) / sigma
* Transforms data to mean = 0, standard deviation = 1.
* Ensures **all features contribute equally** regardless of original scale.
* Critical for algorithms like **KNN**, **PCA**, and **linear regression** where scale impacts outcome.

**2. One-Hot Encoding**

* Converts categorical columns (like *City*, *Brand*, *SKU*) into numerical format.
* Each category becomes a **binary feature column**.
* Prevents models from assuming **ordinal relationships** where there are none.
* Necessary for **most ML models**, especially tree-based and regression models.