**BASIC STATISTICS**

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

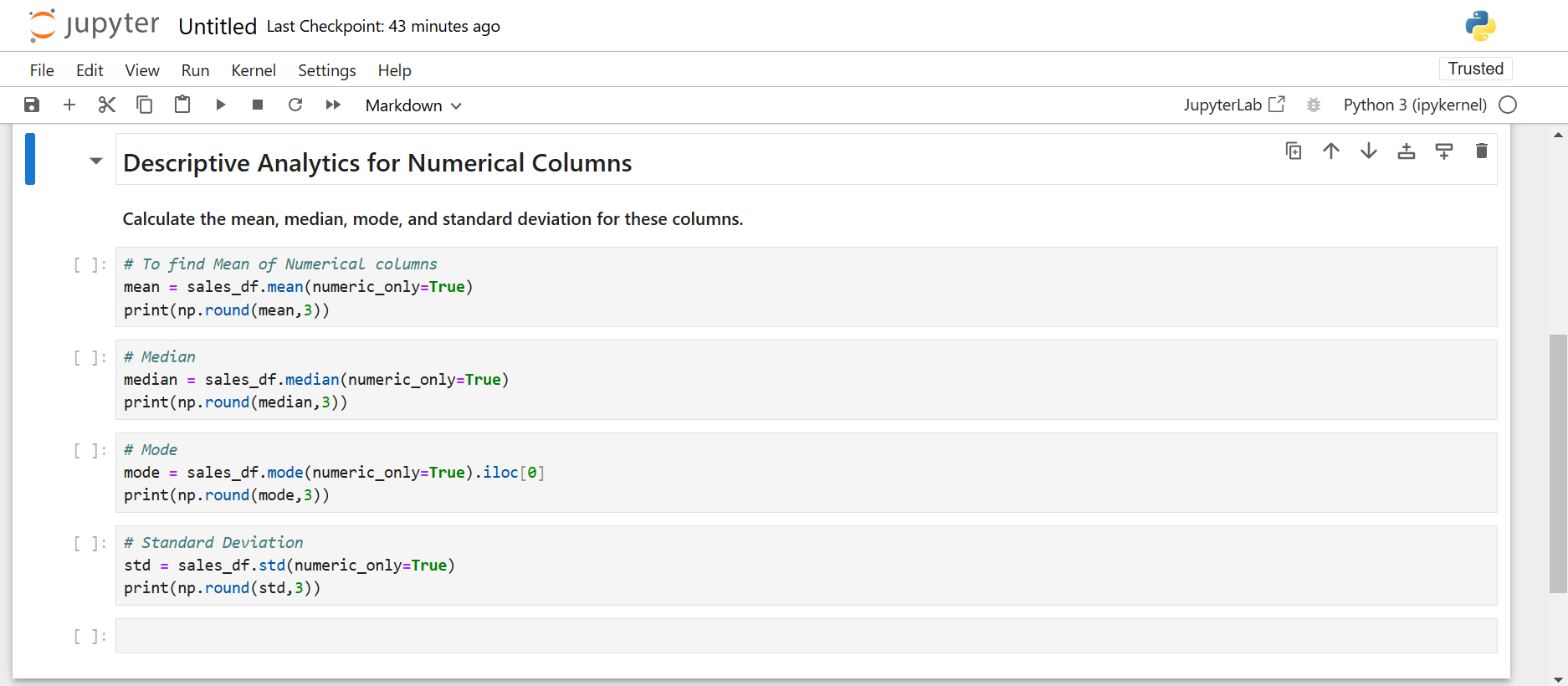
*Descriptive Analytics for Numerical Columns:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column | Mean | Median | Mode | Standard Deviation |
| Volume | 5.067 | 4.000 | 3.000 | 4.232 |
| Avg Price | 10453.433 | 1450.000 | 400.000 | 18079.905 |
| Total Sales Value | 33812.836 | 5700.000 | 24300.000 | 50535.074 |
| Discount Rate (%) | 15.155 | 16.578 | 5.008 | 4.221 |
| Discount Amount | 3346.499 | 988.934 | 69.178 | 4509.903 |
| Net Sales Value | 30466.336 | 4677.788 | 326.975 | 46358.657 |

**Interpretation:**

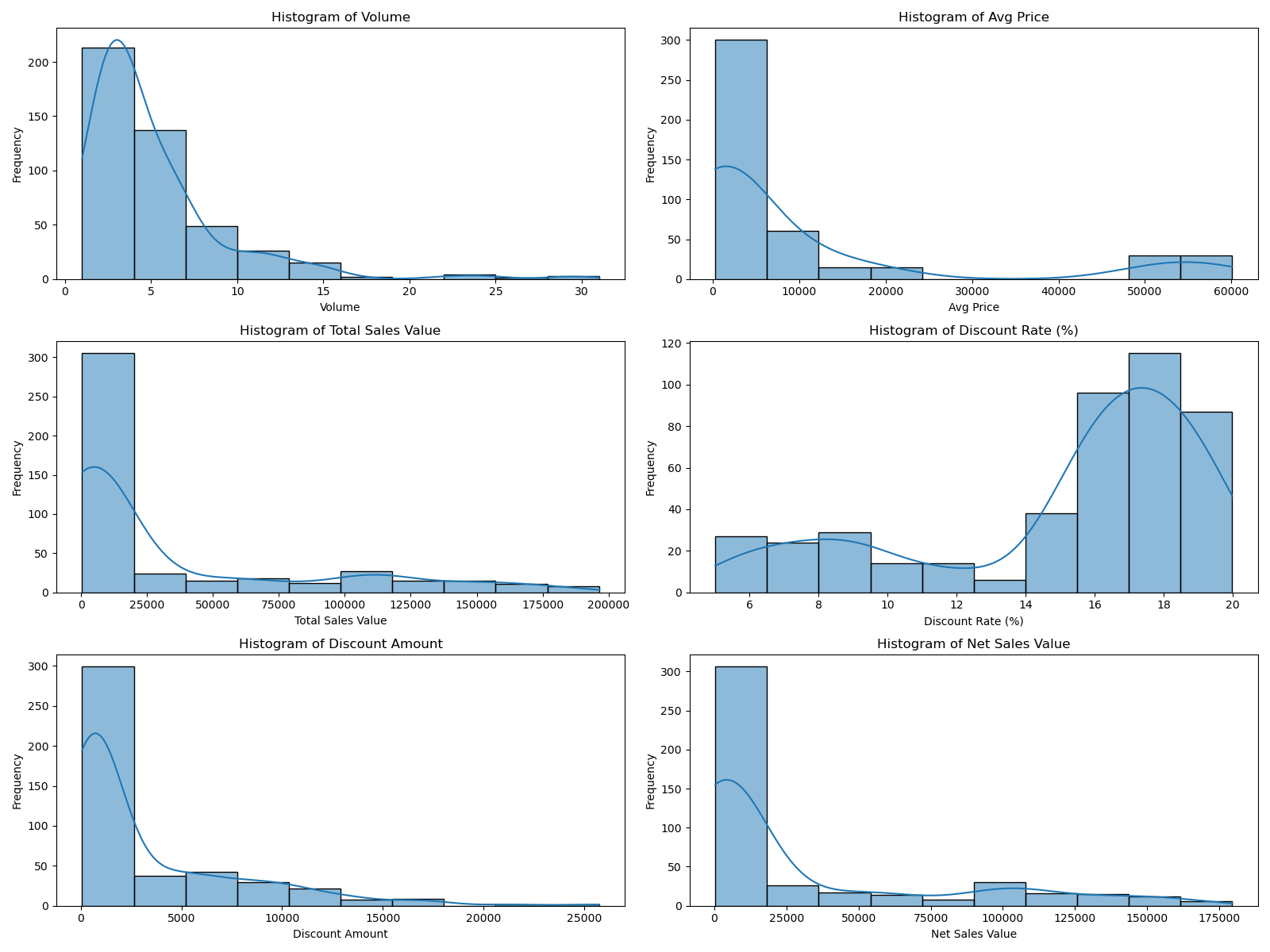
* **Mean**: Represents the average value for each column. For instance, the average price of an item is approximately 10,453.43 units.
* **Median**: Indicates the middle value when the data is sorted. For example, the median discount amount is 988.93 units, which means half of the discounts are below this value.
* **Mode**: The most frequently occurring value in the data. For example, the most common Volume is 3.
* **Standard Deviation**: Measures the dispersion of data from the mean. A higher standard deviation, like in Total Sales Value, indicates more variability in sales figures.

**Python Code:**

****

*Data Visualization:*

**Histograms:**



**Interpretation:**

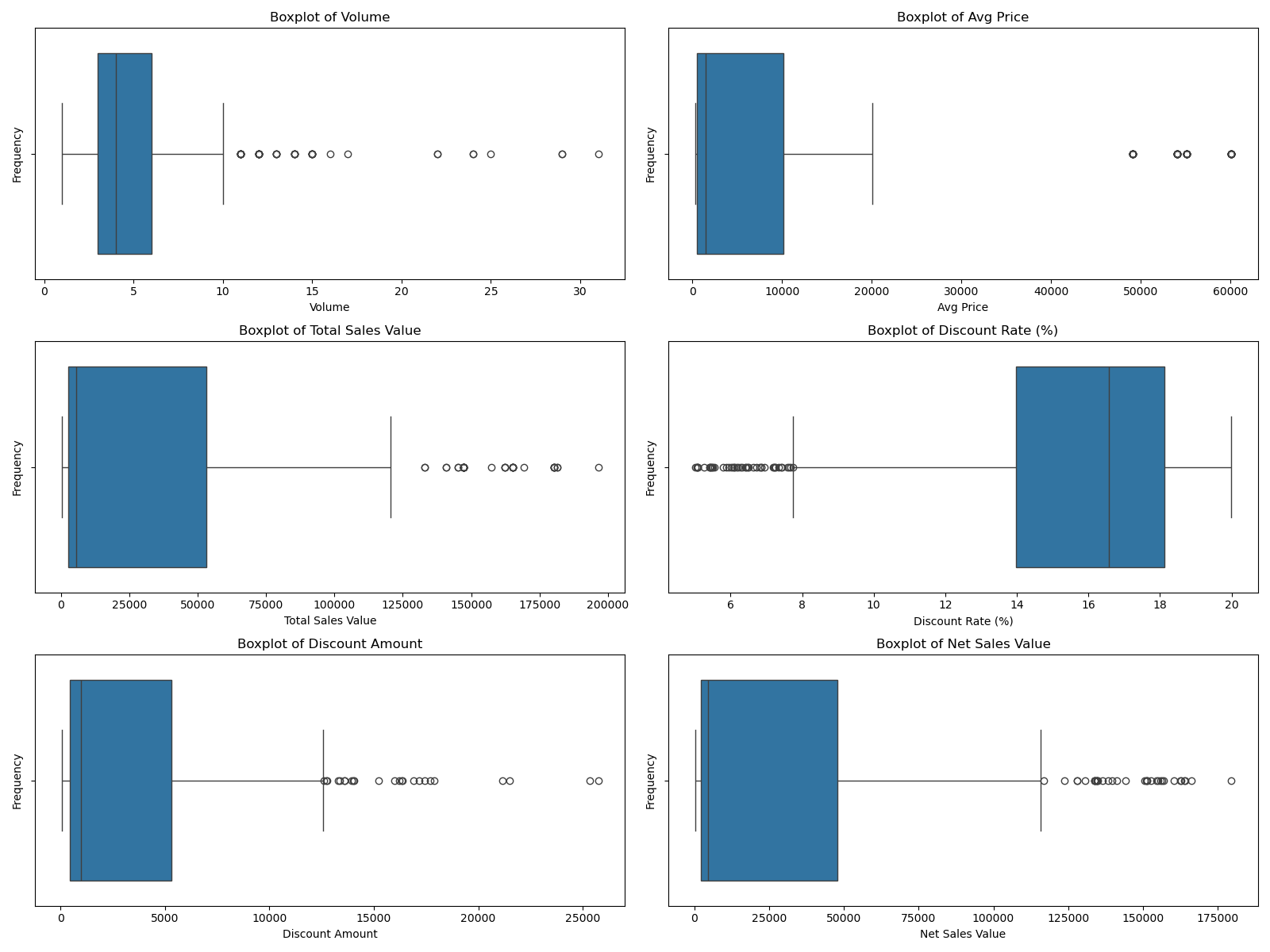
The distribution of columns ‘Volume’, ‘Avg Price’, ‘Total Sales Value’, ‘Discount Amount’, ‘Net Sales Value’ are right-skewed and ‘Discount Rate (%)’ is left-skewed.

All these column distributions contain outliers

**Python Code:**

****

**Boxplot:**

****

**Interpretation:**

Volume: There are numerous outliers in the volume data. many values exceed this limit, indicating potential extreme values. Significant outliers suggest variability in volume data.

Avg Price: The average price has several extreme values well beyond the upper bound. Repeated high values likely indicate data entry errors or unusual pricing.

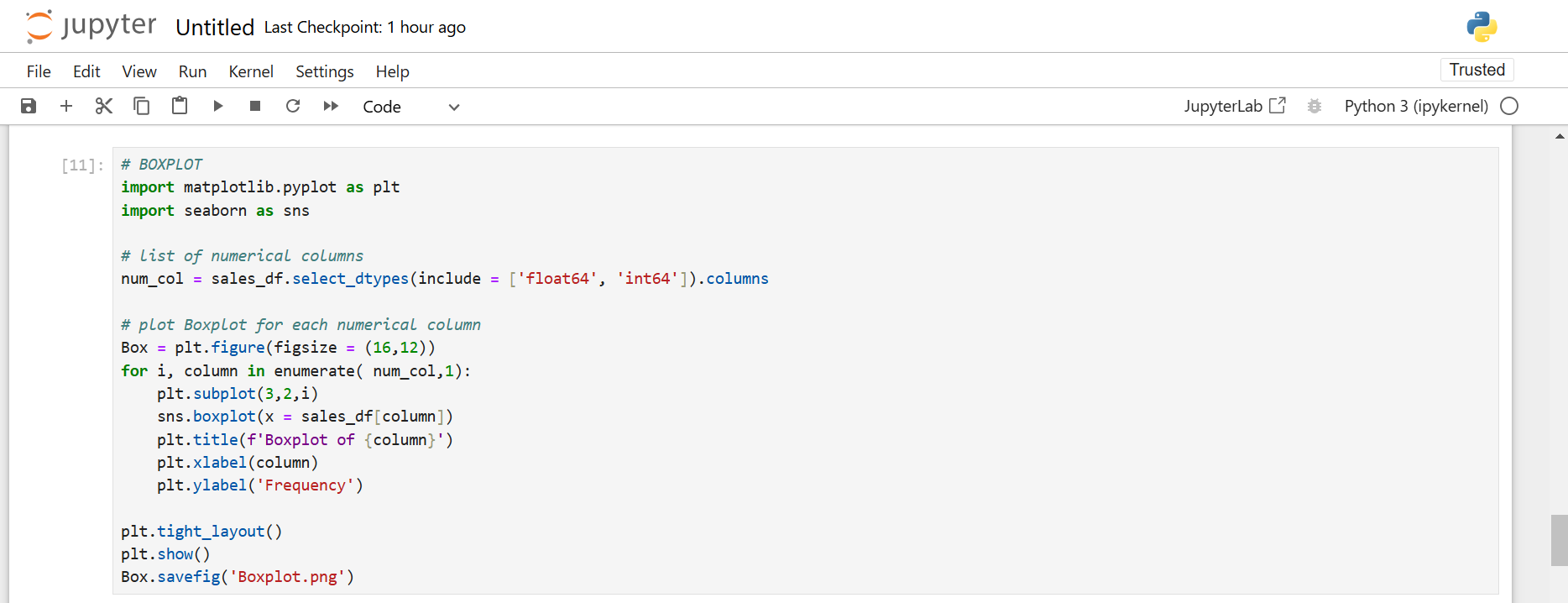
Total Sales Value: The total sales value has several outliers, indicating extreme sales figures. Several high sales values suggest variability in sales performance.

Discount Rate (%): There are several outliers in the discount rate, mainly on the lower end, indicating discounts below the usual range. This suggests that while most discounts fall within a typical range, some are significantly lower.

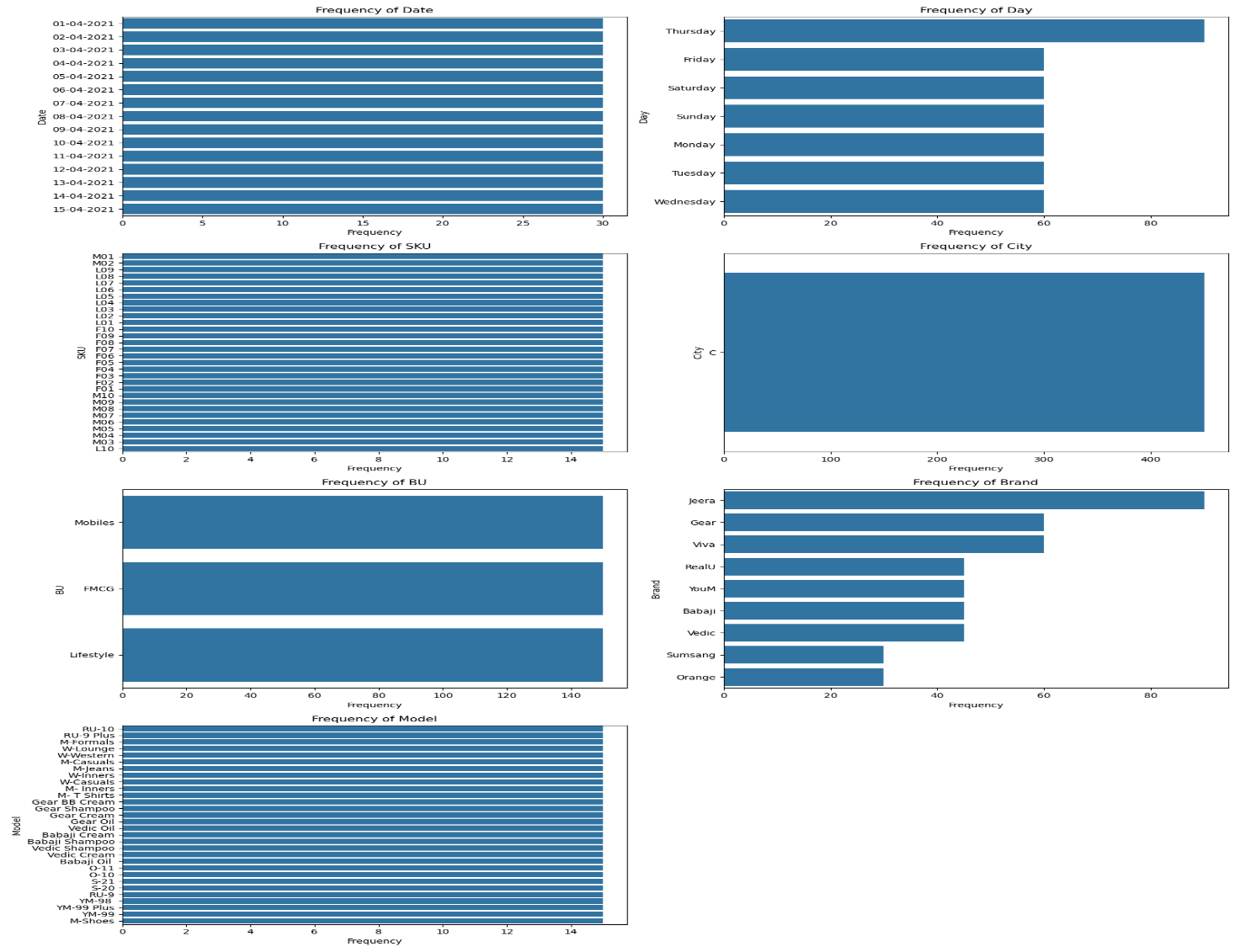
Discount Amount: The discount amount has numerous high outliers, indicating several large discounts. These high values suggest that while most discount amounts are within a reasonable range, there are instances of significantly higher discounts.

Net Sales Value: The net sales value column has several high outliers, indicating large sales figures. This suggests that while most sales values are within a typical range, there are instances of significantly higher sales.

**Python Code:**



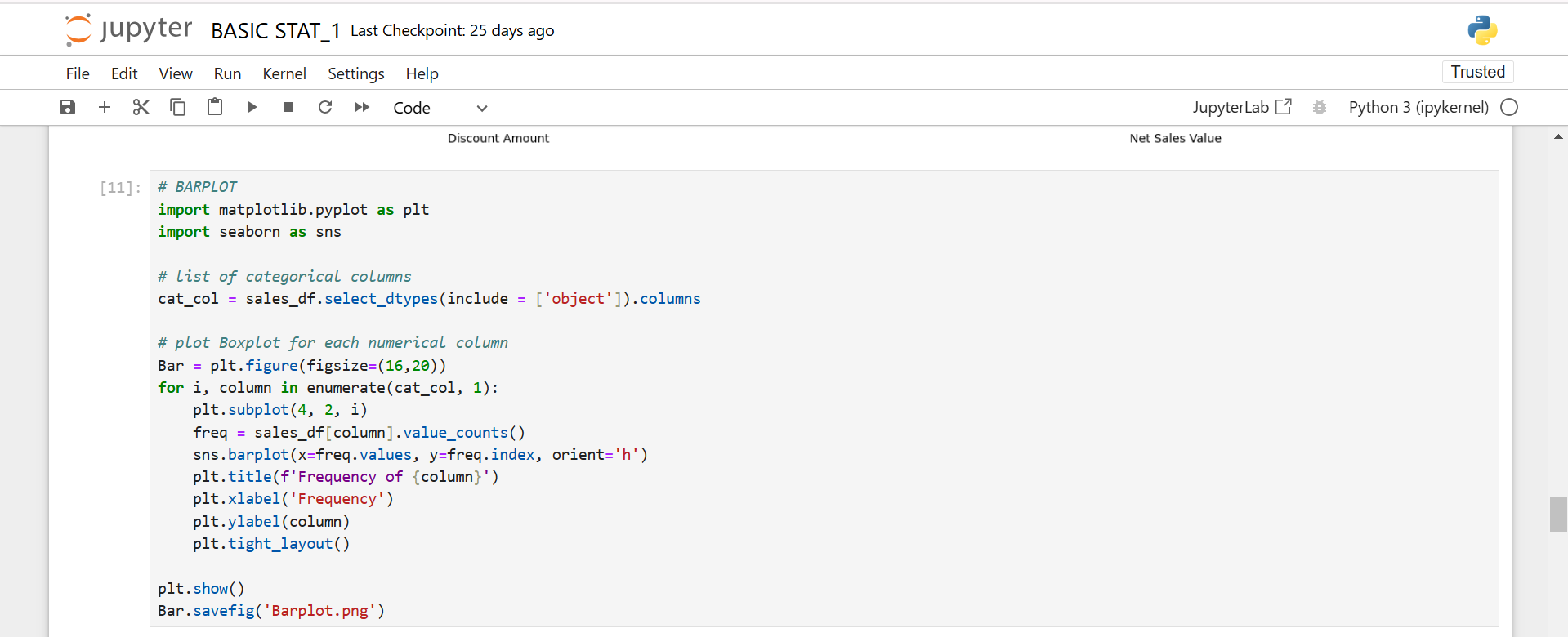
**Bar plot:**

****

**Interpretation:**

* Activity is concentrated towards the end of the week, especially on Thursdays.
* SKUs and Models are evenly distributed, suggesting a balanced product or model offering.
* The "C" city dominates the dataset, possibly indicating that only one location's data is being analyzed.
* Business Units related to Mobiles and FMCG are more active than Lifestyle.
* Jeera is the most frequent brand, with Sumsang and Orange being less frequent.

**Python Code:**

****

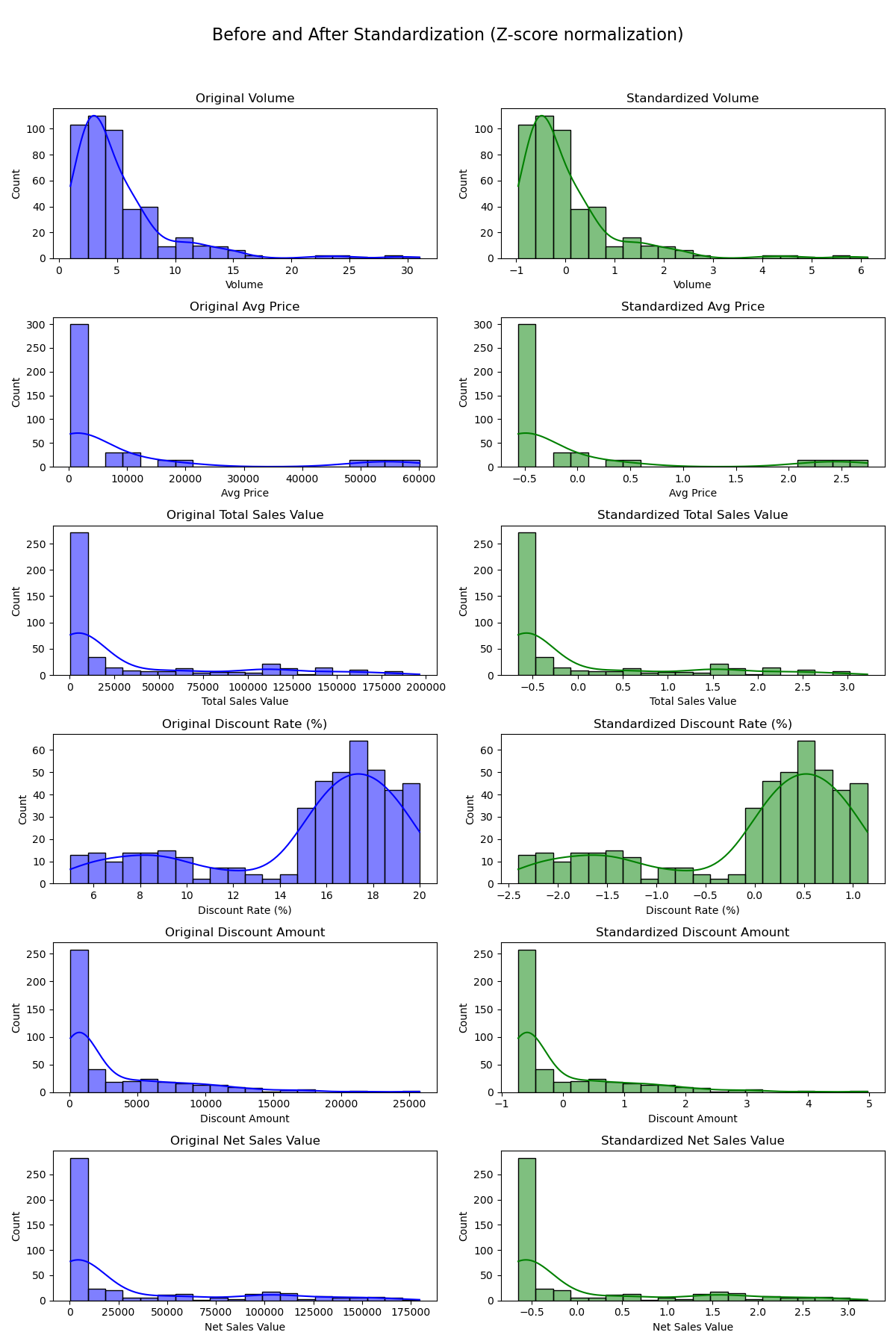
*Standardization of Numerical Variables:*

Data standardization (Z-Score Normalization) is a preprocessing technique used to rescale features in a dataset so that they have a common scale. The goal is to ensure that all features contribute equally to a machine learning model, especially when they have different units or scales.

**Formula:**

Where:

* X is the original data point (feature value).
* μ is the mean of the feature.
* σ is the standard deviation of the feature.
* Z is the standardized (or z-score) value.



#### Conversion of Categorical Data into Dummy Variables:

Converting categorical data into dummy variables (one-hot encoding) is necessary because many machine learning algorithms require numerical input and cannot handle categorical data directly. One-hot encoding transforms each category into a binary column (0 or 1), ensuring that no unintended ordinal relationship is implied between categories. This method makes categorical data suitable for models like regression, decision trees, and neural networks.

A white background with black text

Description automatically generated