Descriptive Analytics and Data Preprocessing on Sales & Discounts Dataset

# Introduction

This document outlines descriptive analytics and data preprocessing steps performed on the sales and discounts dataset. It includes basic statistical measures, visualizations, standardization, and conversion of categorical variables using one-hot encoding.

# Descriptive Analytics for Numerical Columns

## Statistics Summary

The following are the descriptive statistics for numerical columns:

Mean values:  
Volume 5.066667  
Avg Price 10453.433333  
Total Sales Value 33812.835556  
Discount Rate (%) 15.155242  
Discount Amount 3346.499424  
Net Sales Value 30466.336131  
dtype: float64

Median values:  
Volume 4.000000  
Avg Price 1450.000000  
Total Sales Value 5700.000000  
Discount Rate (%) 16.577766  
Discount Amount 988.933733  
Net Sales Value 4677.788059  
dtype: float64

Mode values:  
Volume 3.000000  
Avg Price 400.000000  
Total Sales Value 24300.000000  
Discount Rate (%) 5.007822  
Discount Amount 69.177942  
Net Sales Value 326.974801  
Name: 0, dtype: float64

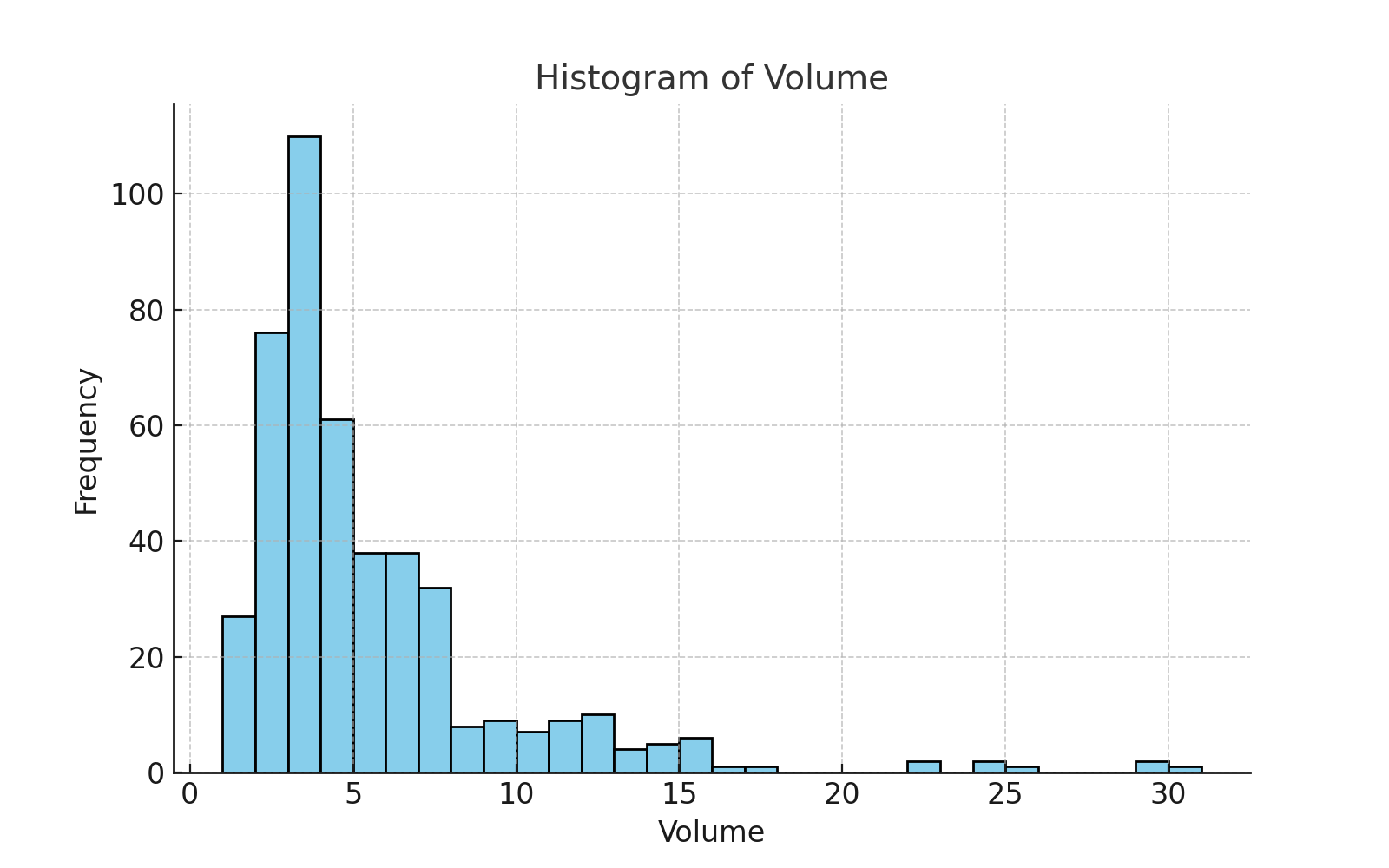
Standard Deviation values:  
Volume 4.231602  
Avg Price 18079.904840  
Total Sales Value 50535.074173  
Discount Rate (%) 4.220602  
Discount Amount 4509.902963  
Net Sales Value 46358.656624  
dtype: float64

# Data Visualization

## Histograms of Numerical Columns

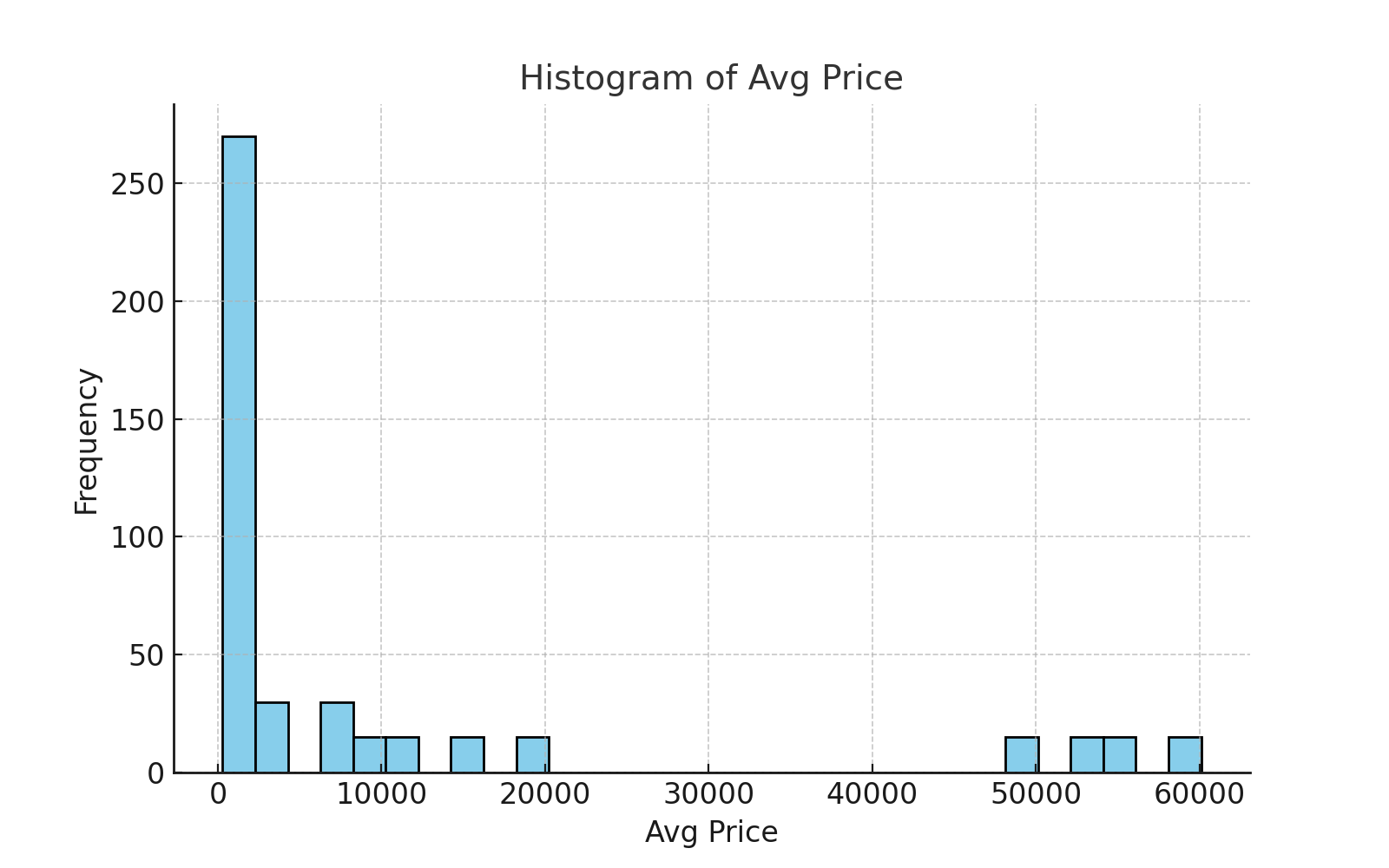
Histograms provide a graphical representation of the distribution of numerical data. They help in identifying the shape of the distribution (e.g., normal, skewed) and potential outliers.

### Histogram of Volume



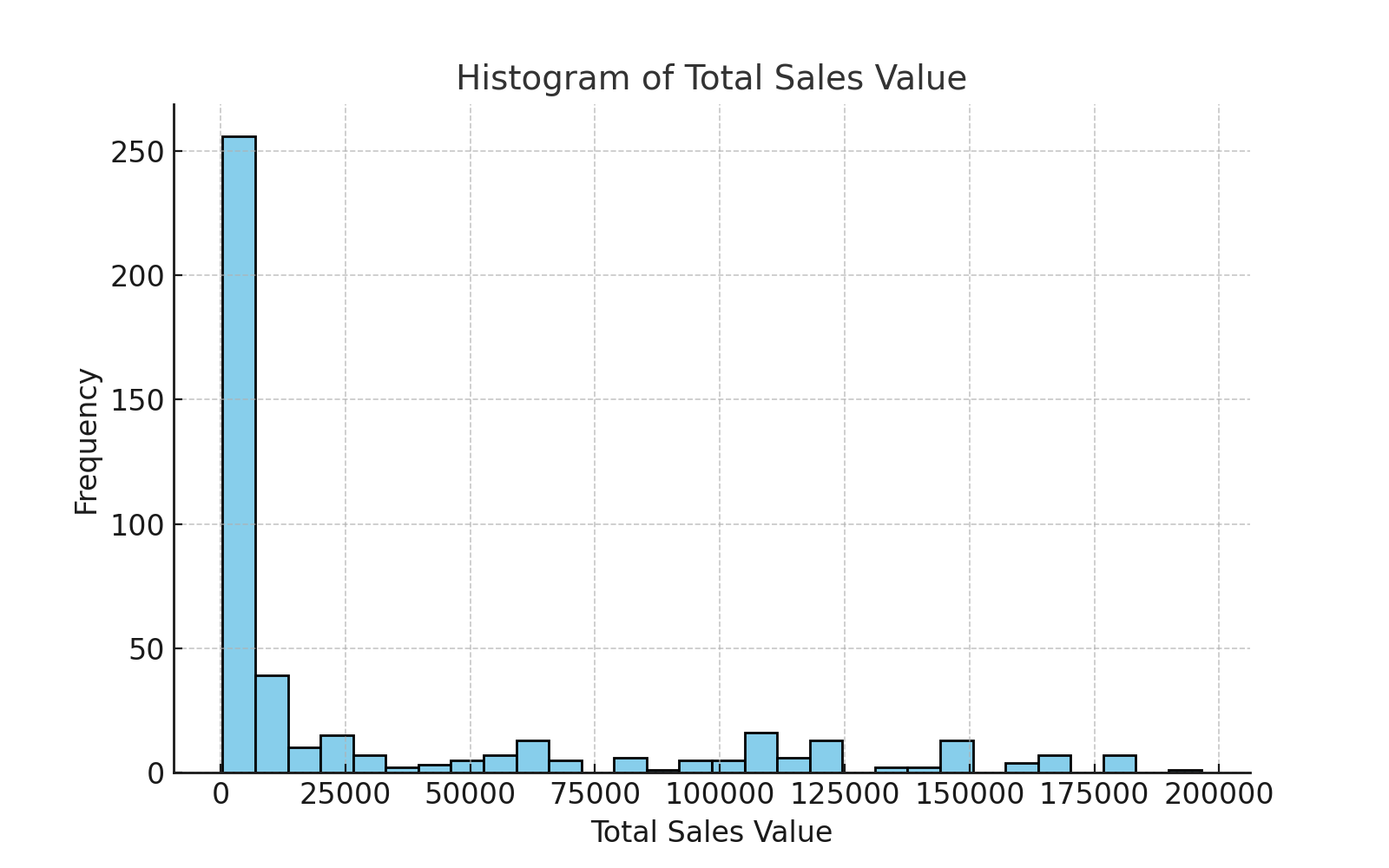
This histogram shows the distribution of Volume. Observations regarding skewness or outliers can be made from the visual representation.

### Histogram of Avg Price



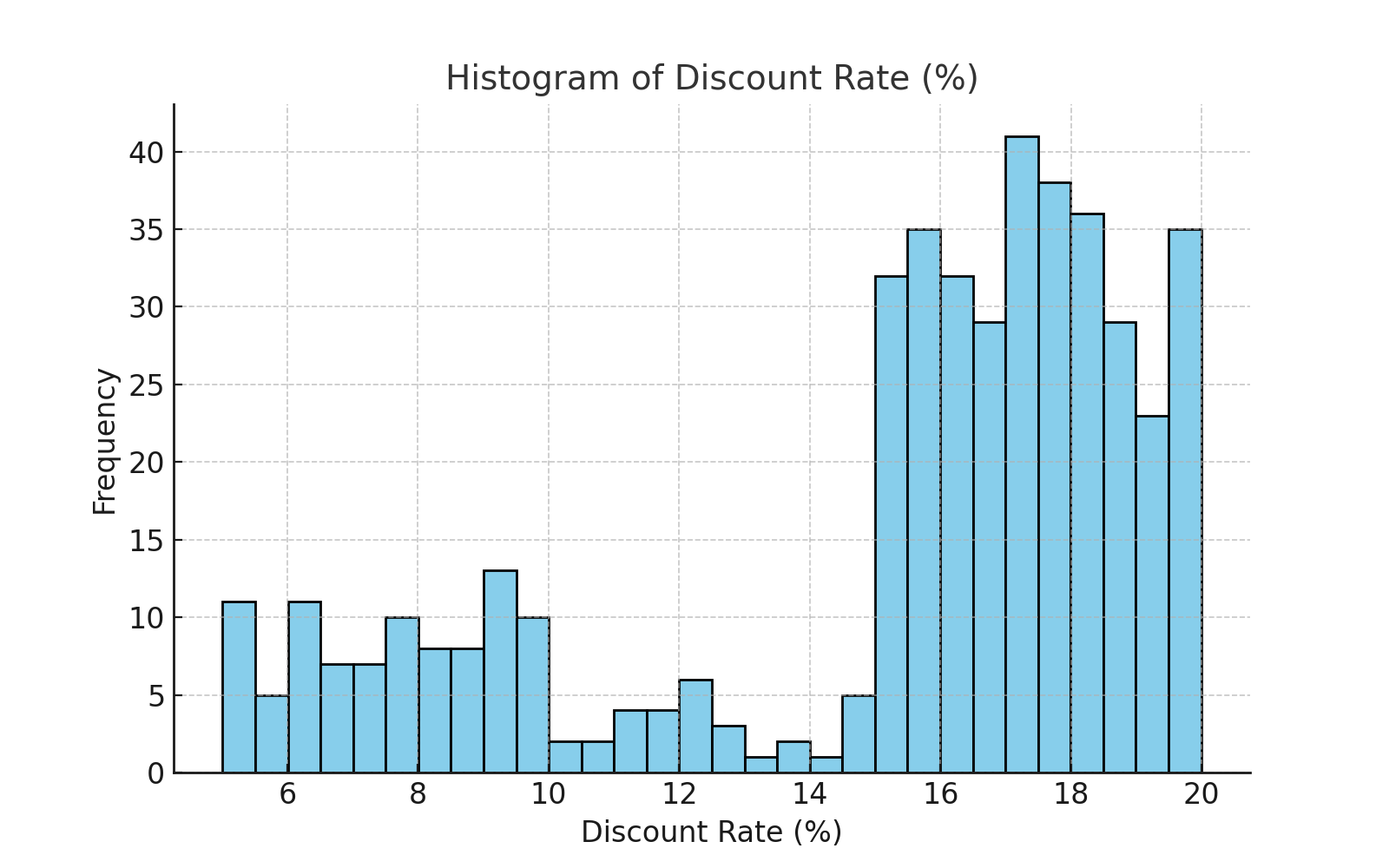
This histogram shows the distribution of Avg Price. Observations regarding skewness or outliers can be made from the visual representation.

### Histogram of Total Sales Value



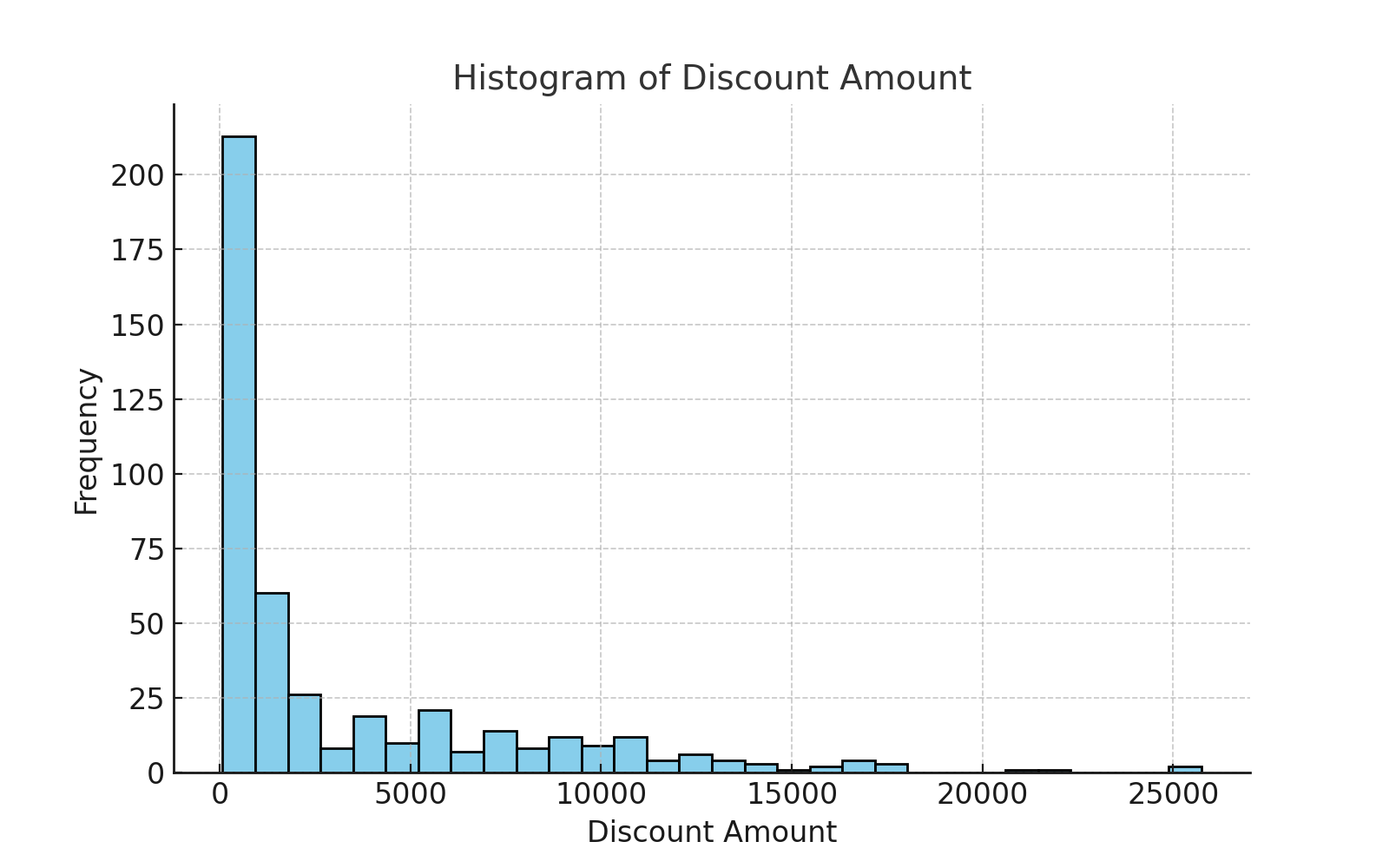
This histogram shows the distribution of Total Sales Value. Observations regarding skewness or outliers can be made from the visual representation.

### Histogram of Discount Rate (%)



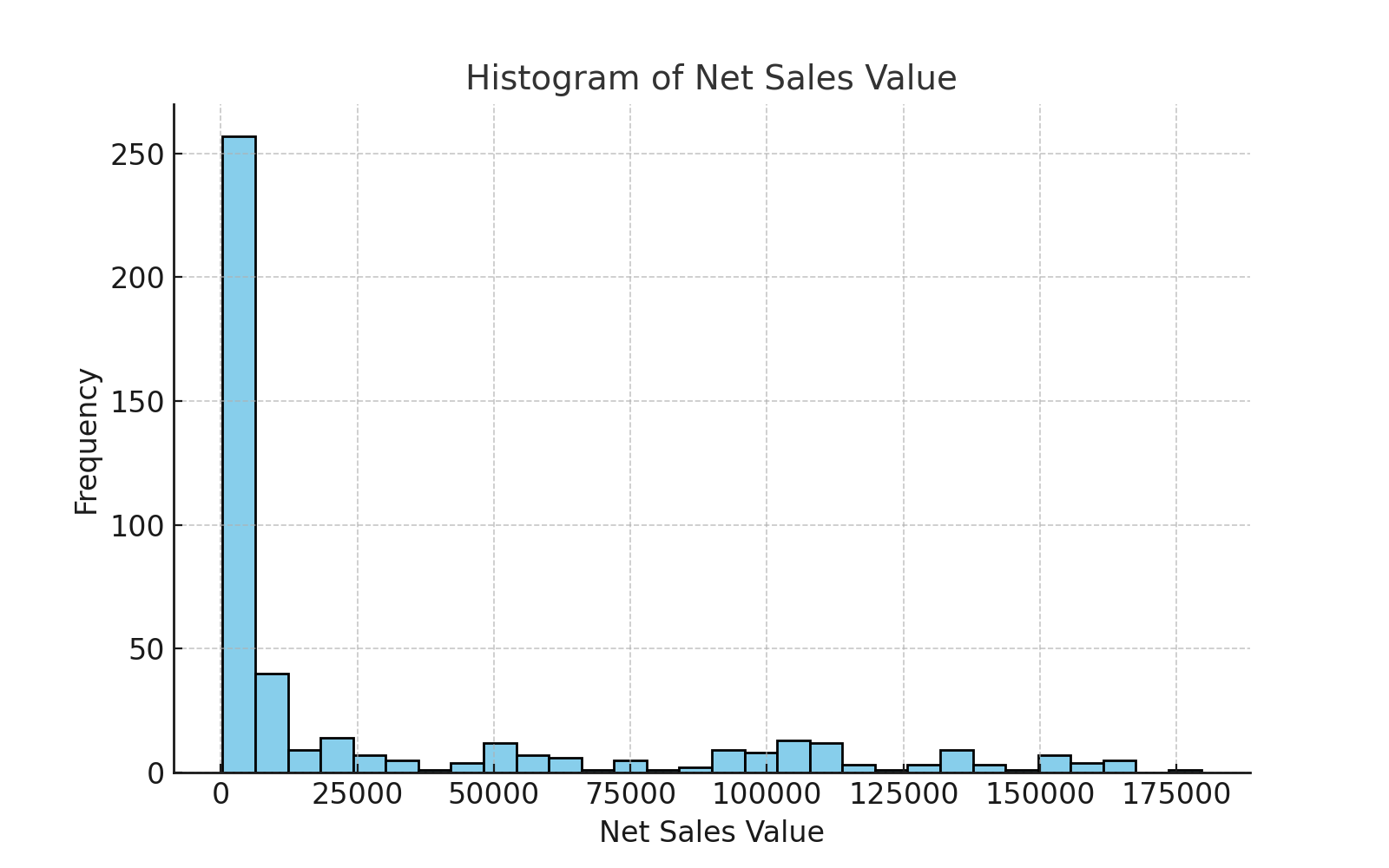
This histogram shows the distribution of Discount Rate (%). Observations regarding skewness or outliers can be made from the visual representation.

### Histogram of Discount Amount



This histogram shows the distribution of Discount Amount. Observations regarding skewness or outliers can be made from the visual representation.

### Histogram of Net Sales Value

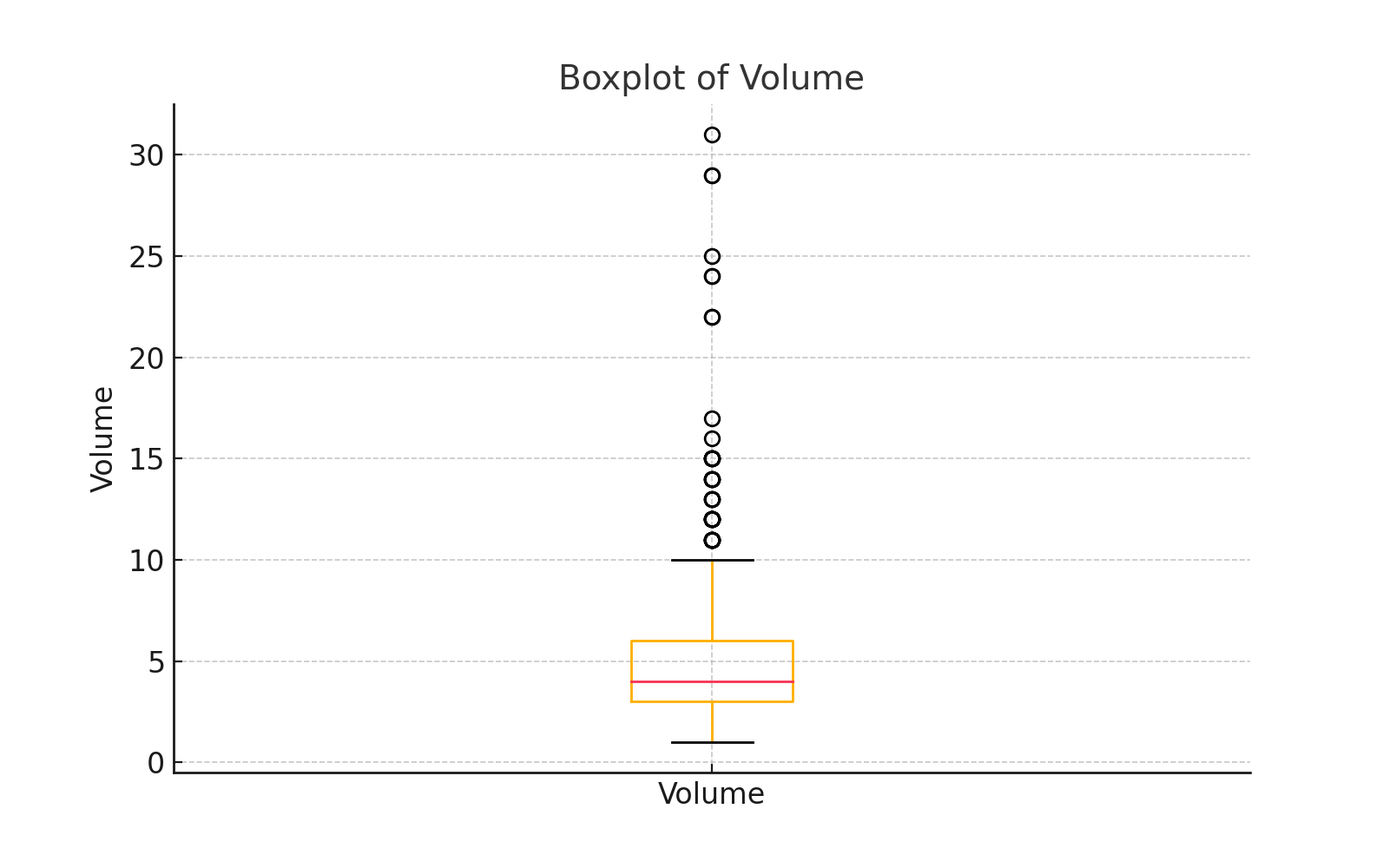


This histogram shows the distribution of Net Sales Value. Observations regarding skewness or outliers can be made from the visual representation.

## Boxplots of Numerical Columns

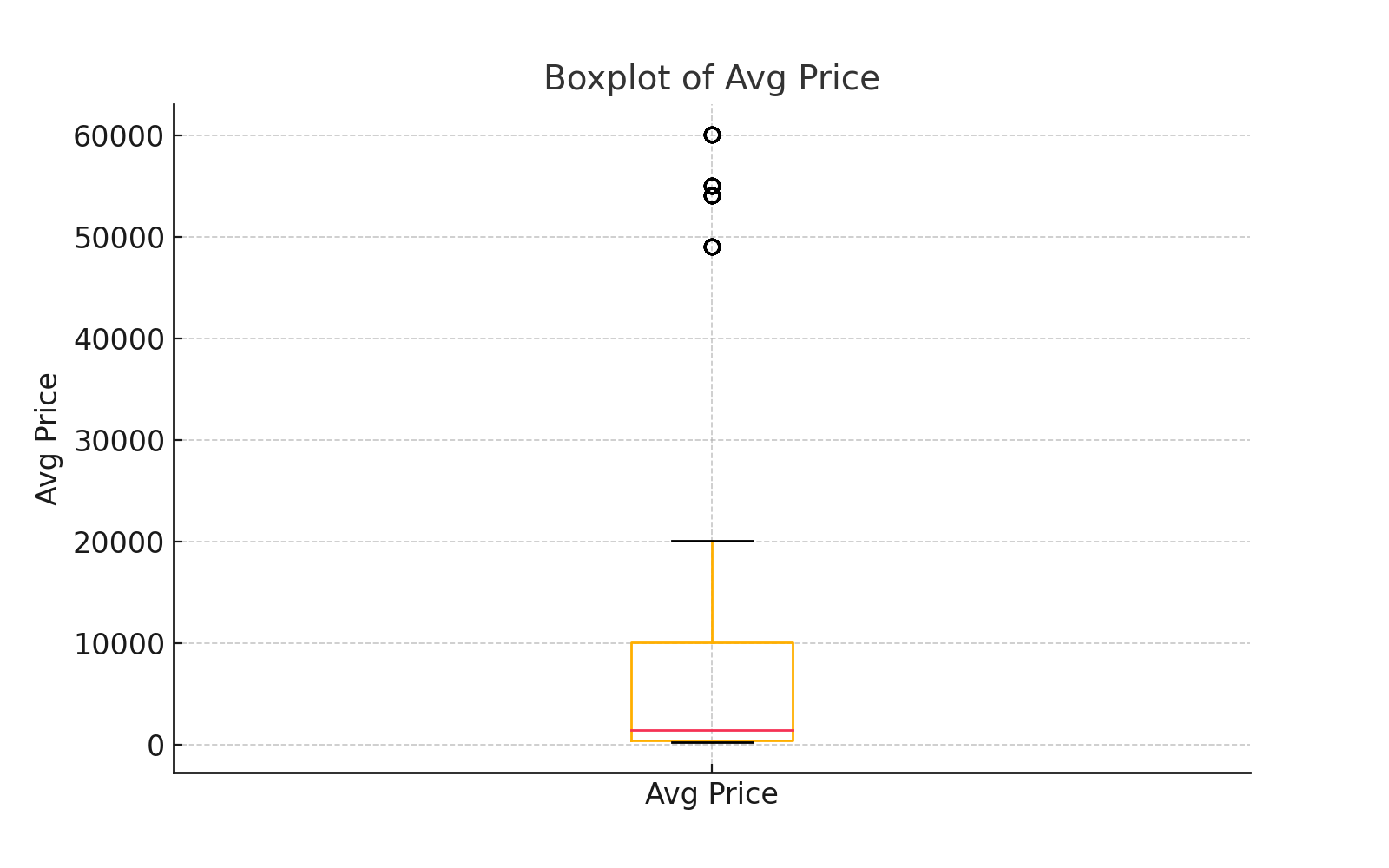
Boxplots provide a summary of the distribution of numerical data, showing the median, quartiles, and potential outliers. They are useful for identifying the spread of data and the presence of extreme values.

### Boxplot of Volume



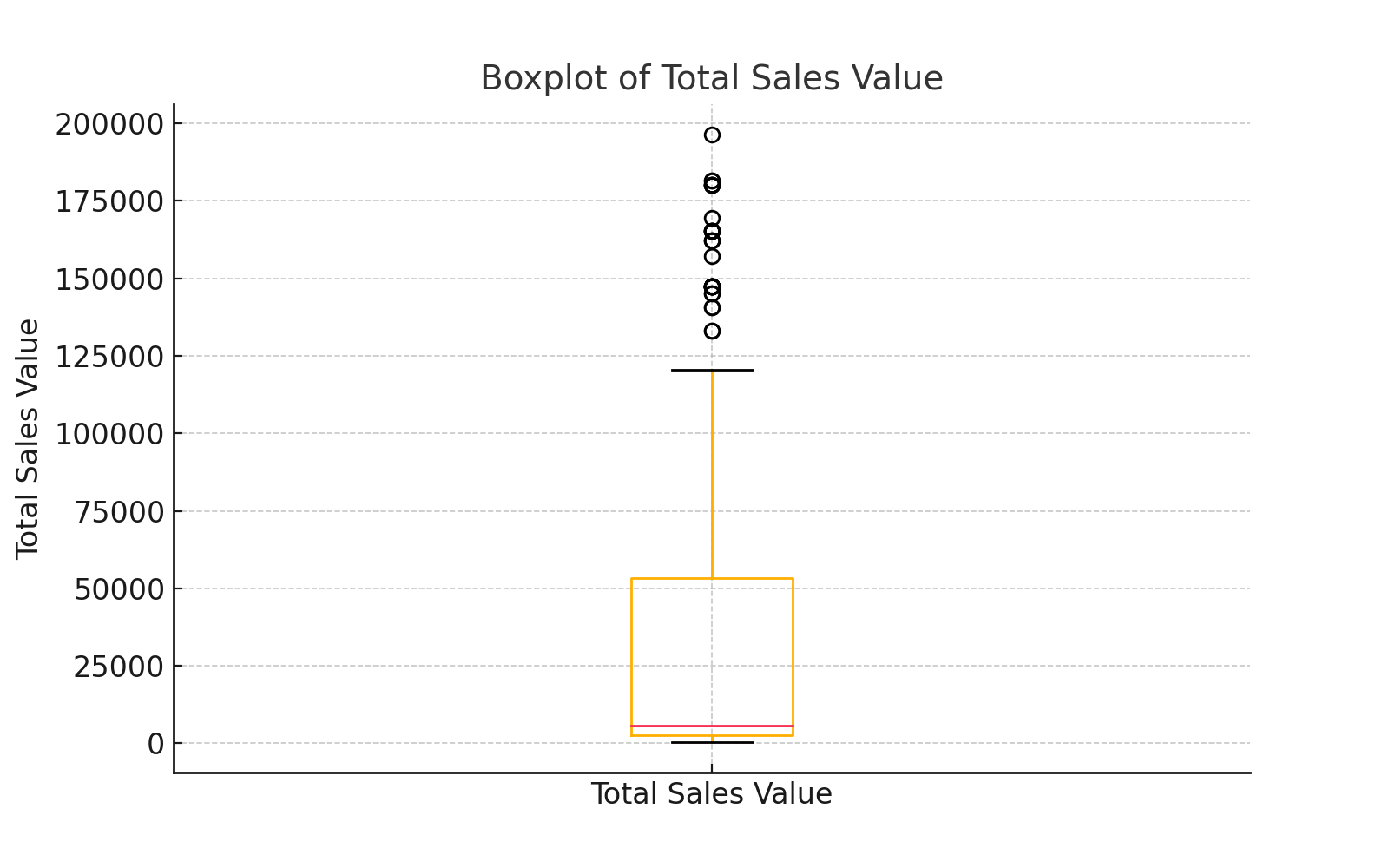
The boxplot of Volume shows the interquartile range and identifies any outliers in the dataset.

### Boxplot of Avg Price



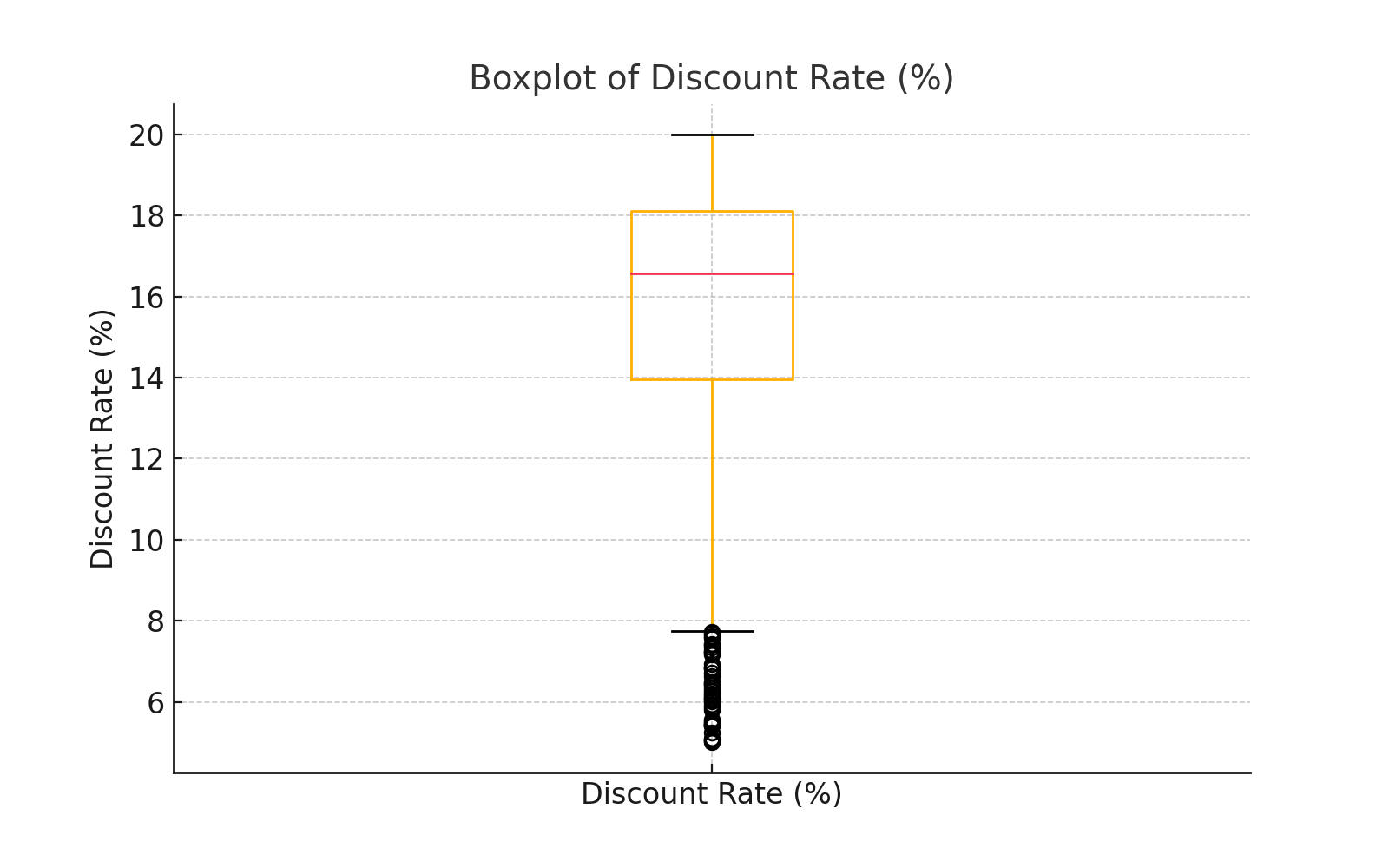
The boxplot of Avg Price shows the interquartile range and identifies any outliers in the dataset.

### Boxplot of Total Sales Value



The boxplot of Total Sales Value shows the interquartile range and identifies any outliers in the dataset.

### Boxplot of Discount Rate (%)



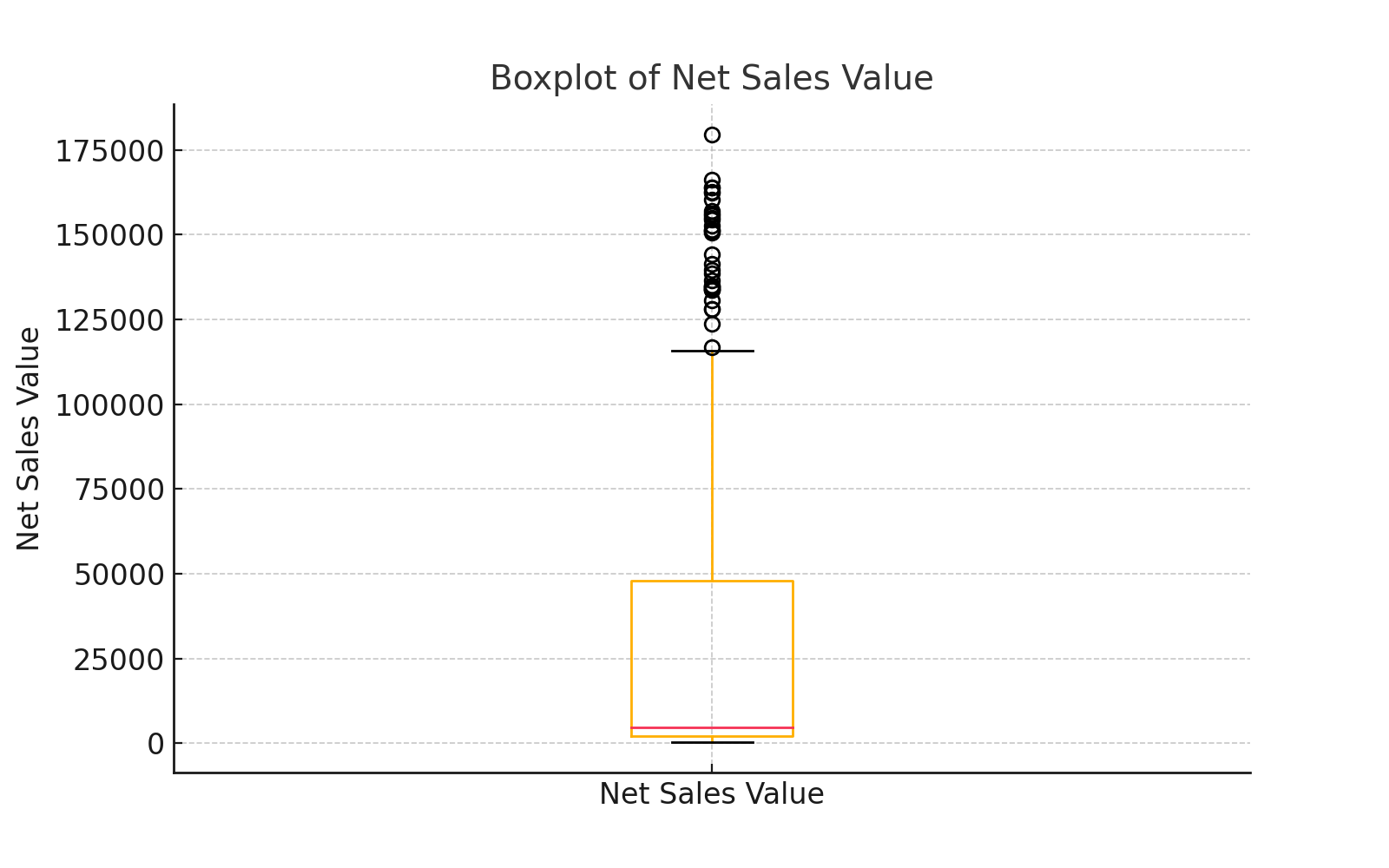
The boxplot of Discount Rate (%) shows the interquartile range and identifies any outliers in the dataset.

### Boxplot of Discount Amount



The boxplot of Discount Amount shows the interquartile range and identifies any outliers in the dataset.

### Boxplot of Net Sales Value

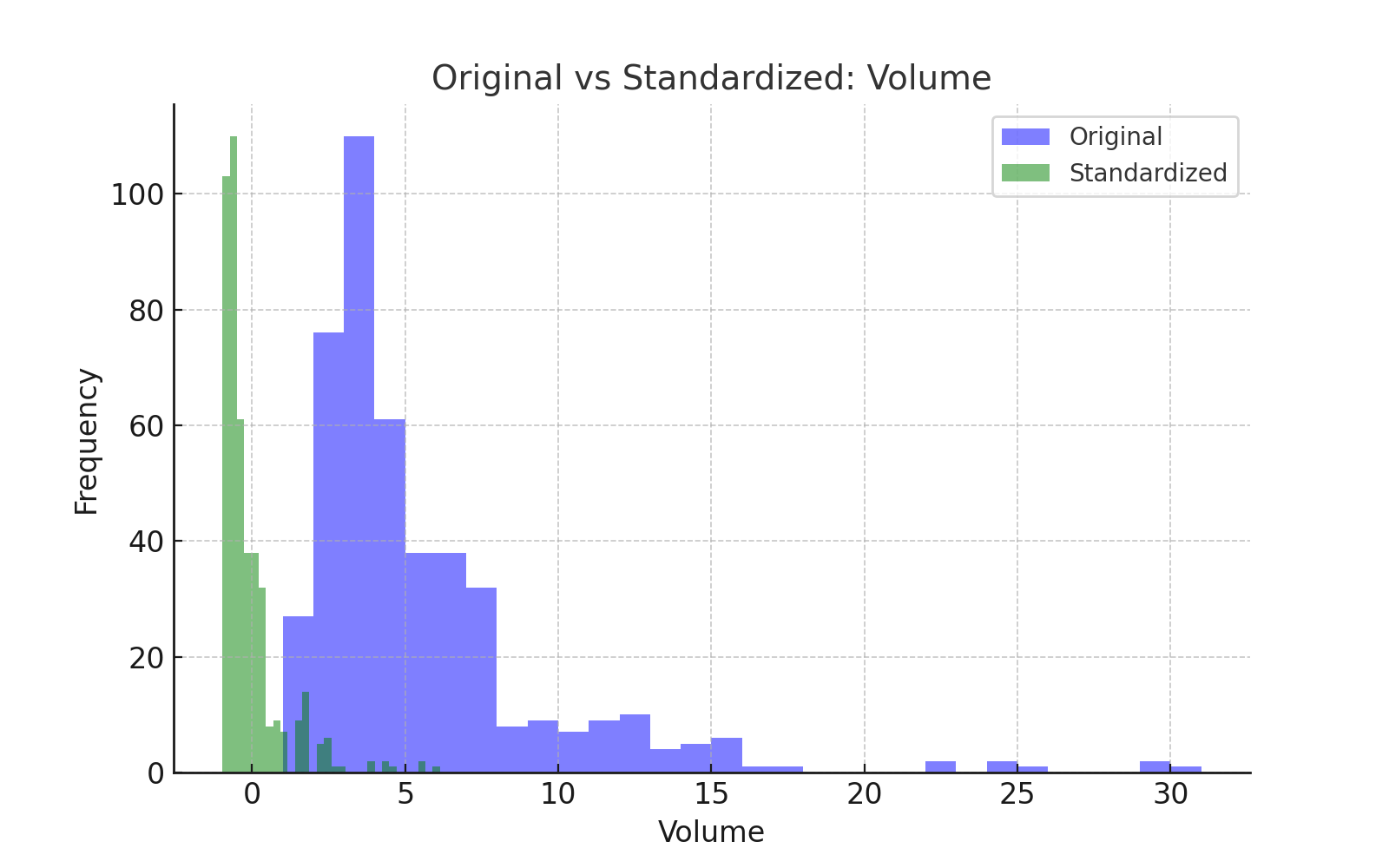


The boxplot of Net Sales Value shows the interquartile range and identifies any outliers in the dataset.

# Standardization of Numerical Variables

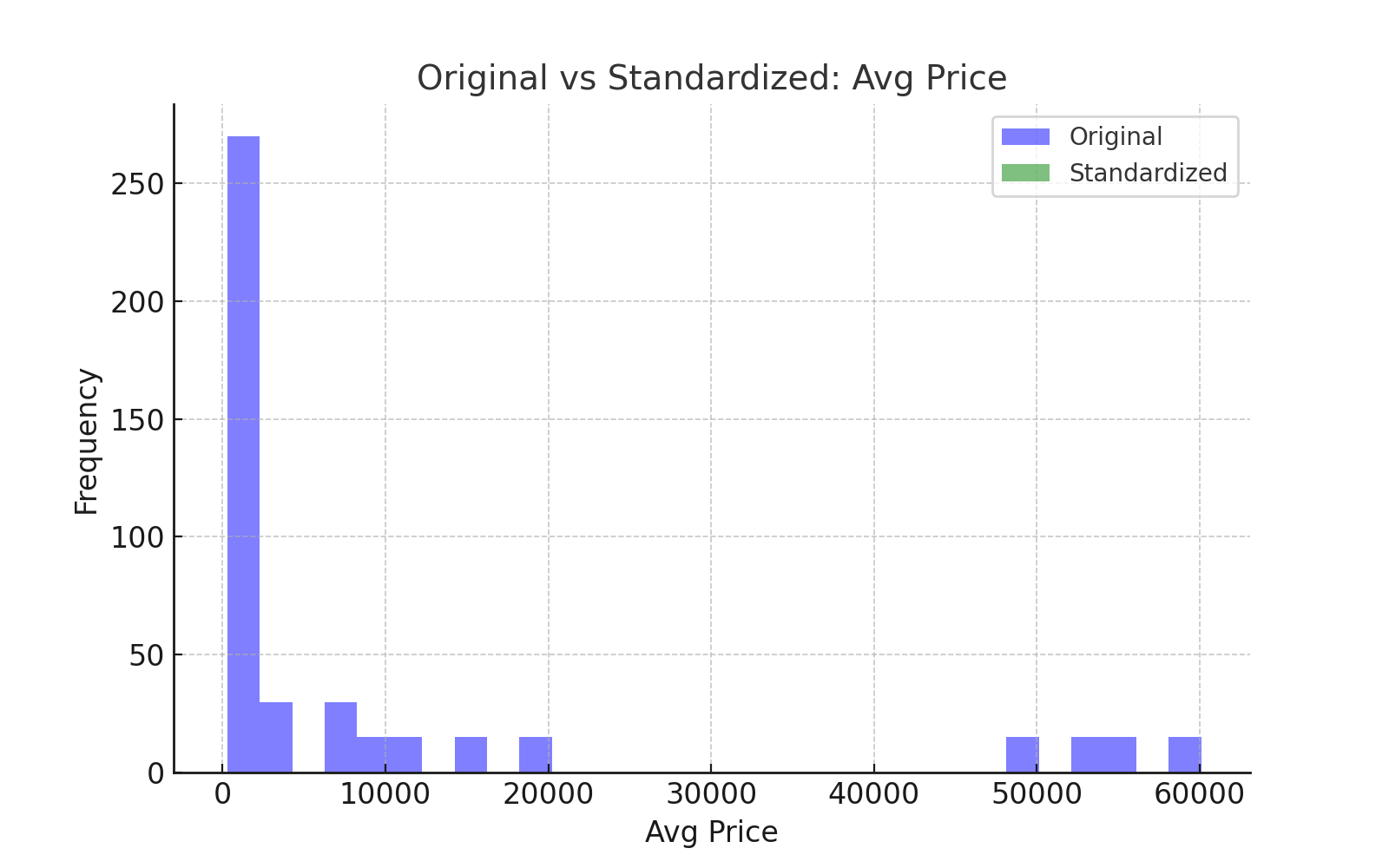
Standardization (z-score normalization) was applied to numerical columns. The data was transformed so that each numerical feature has a mean of 0 and a standard deviation of 1, making it easier for machine learning algorithms to perform accurately.

### Original vs Standardized: Volume



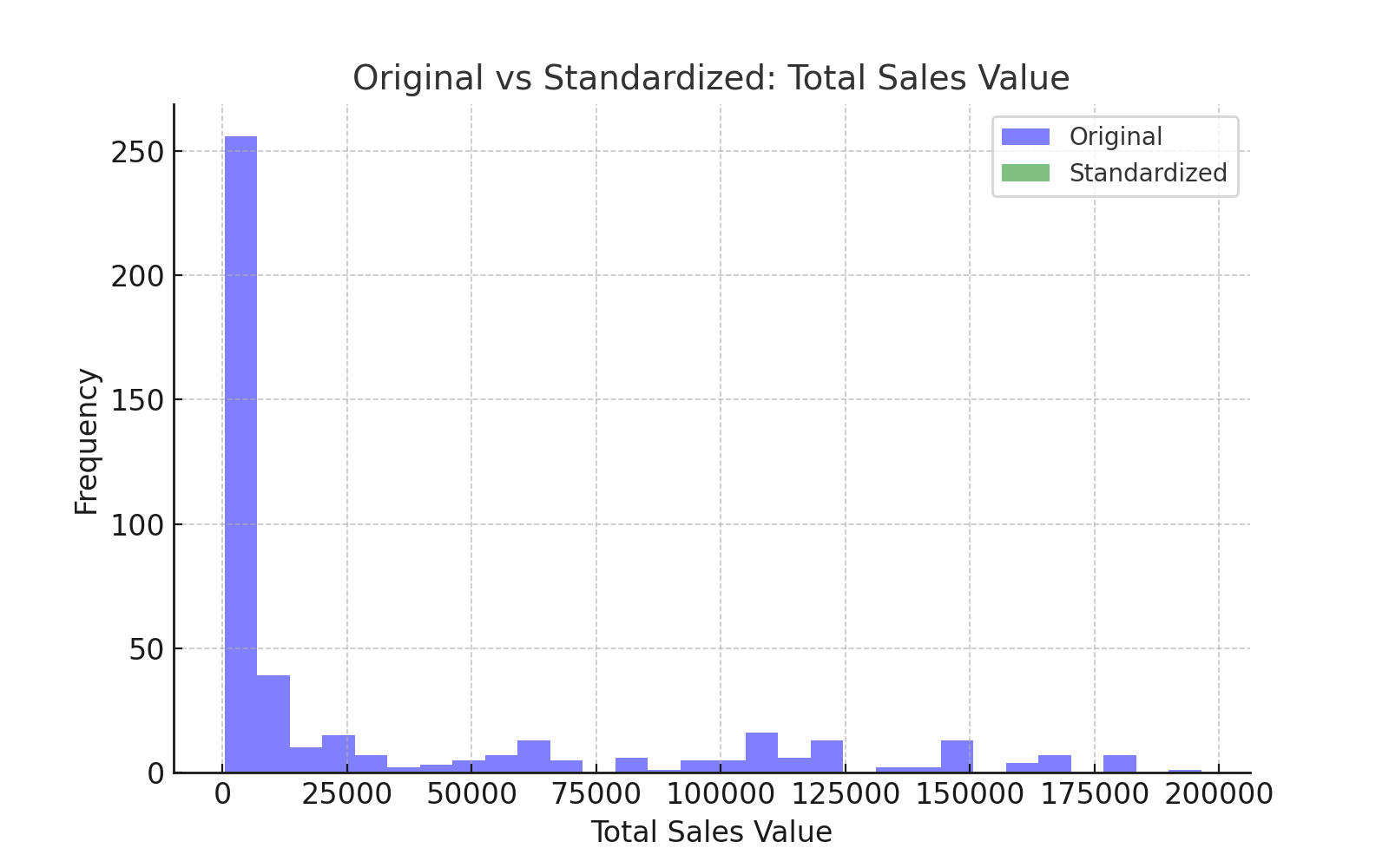
This comparison shows the original and standardized distribution of Volume. Standardization helps in scaling the data uniformly, making it easier to interpret in models.

### Original vs Standardized: Avg Price



This comparison shows the original and standardized distribution of Avg Price. Standardization helps in scaling the data uniformly, making it easier to interpret in models.

### Original vs Standardized: Total Sales Value



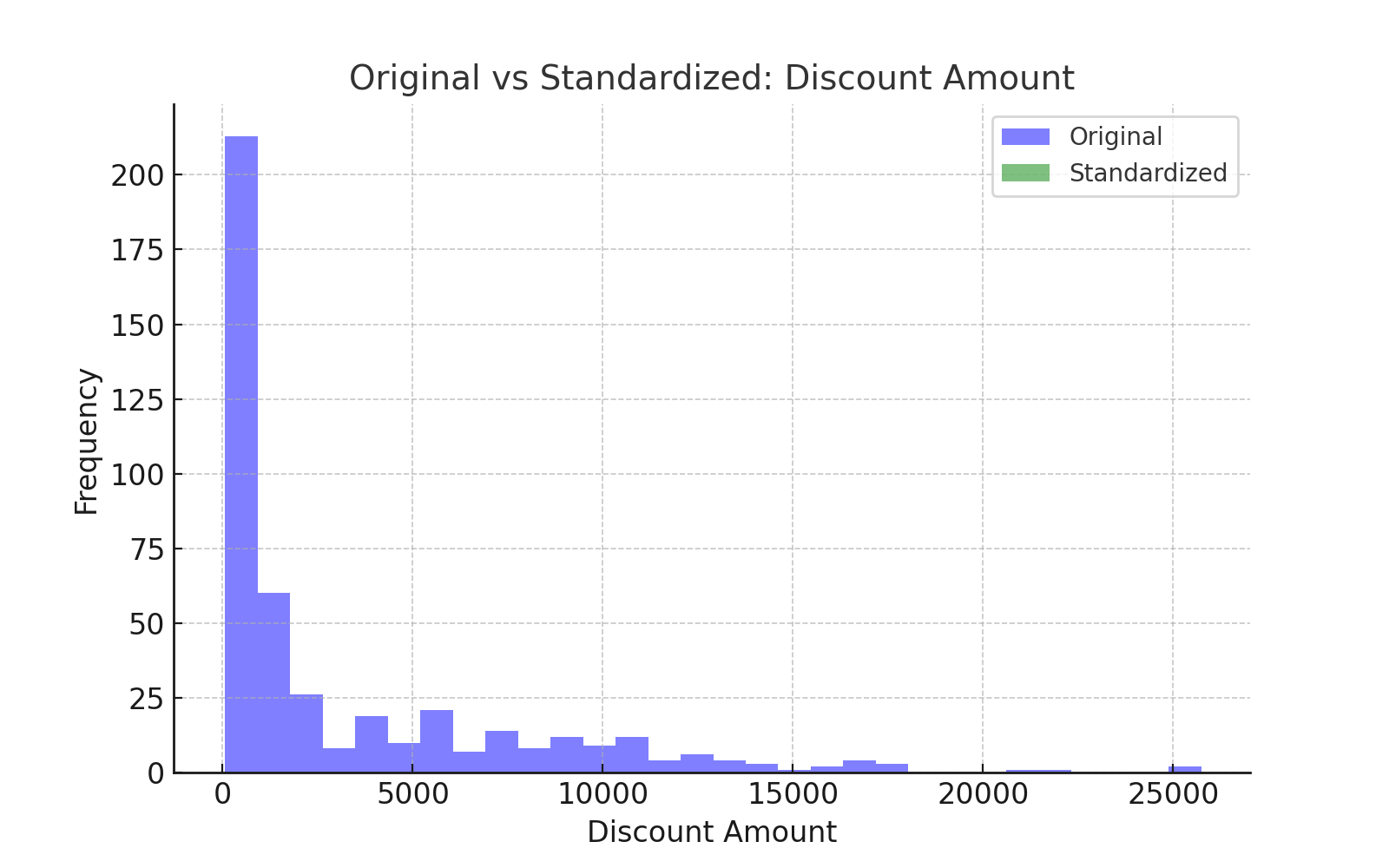
This comparison shows the original and standardized distribution of Total Sales Value. Standardization helps in scaling the data uniformly, making it easier to interpret in models.

### Original vs Standardized: Discount Rate (%)



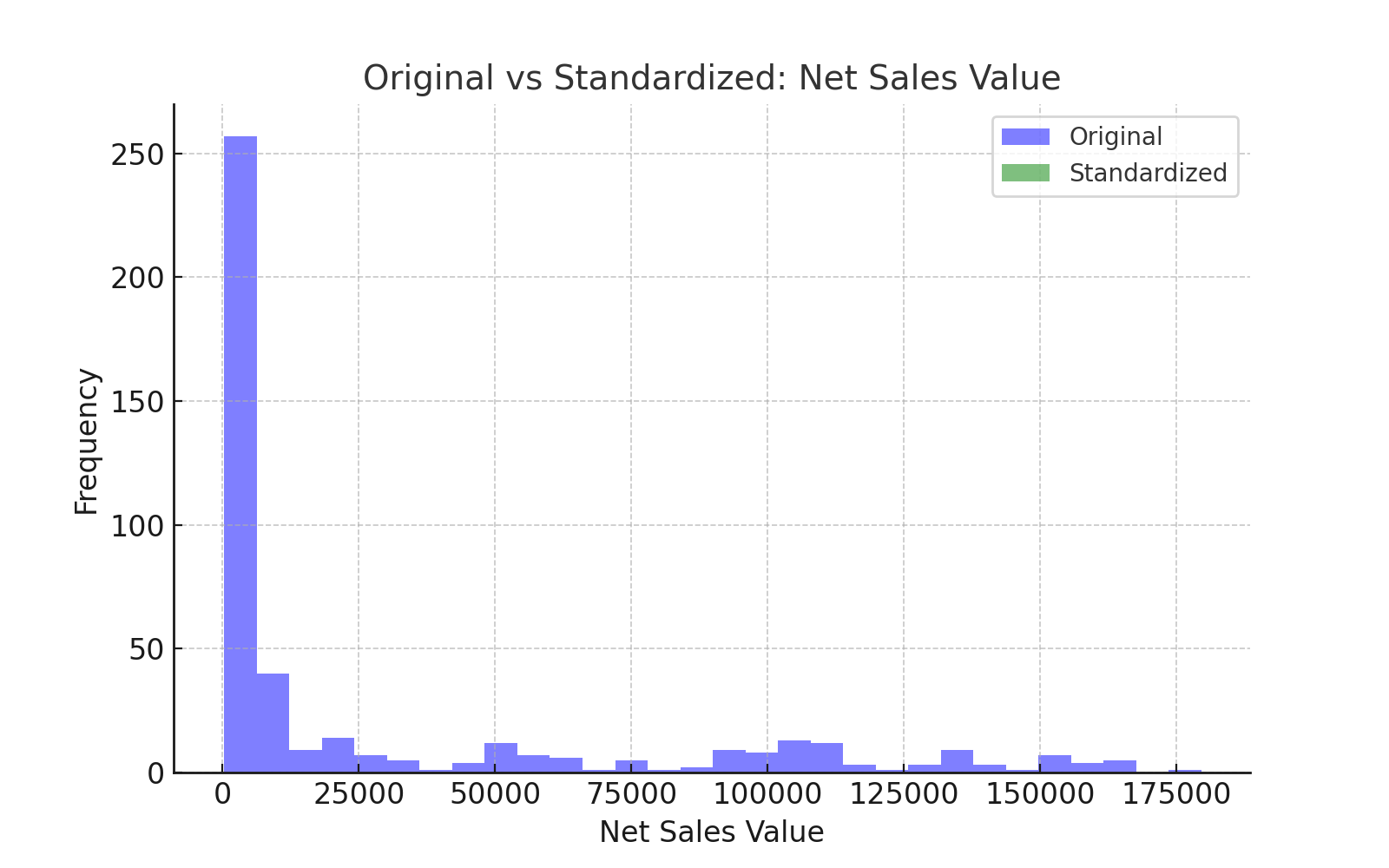
This comparison shows the original and standardized distribution of Discount Rate (%). Standardization helps in scaling the data uniformly, making it easier to interpret in models.

### Original vs Standardized: Discount Amount



This comparison shows the original and standardized distribution of Discount Amount. Standardization helps in scaling the data uniformly, making it easier to interpret in models.

### Original vs Standardized: Net Sales Value



This comparison shows the original and standardized distribution of Net Sales Value. Standardization helps in scaling the data uniformly, making it easier to interpret in models.

# Conversion of Categorical Data into Dummy Variables

Categorical variables in the dataset were converted to dummy variables using one-hot encoding. This transformation allows categorical data to be included in machine learning algorithms by creating binary columns.

# Conclusion

This analysis provided key insights into the sales and discounts data. Descriptive analytics helped in understanding the data distribution, while data preprocessing steps like standardization and one-hot encoding prepared the dataset for further analysis in machine learning models.