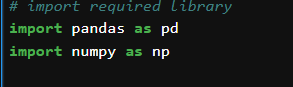
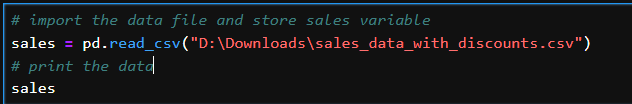
**Descriptive Analytics for Numerical Columns**

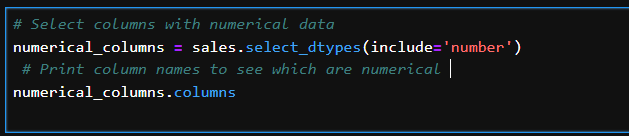
**Objective**: To compute and analyze basic statistical measures for numerical columns in the dataset

**Import pandas and NumPy libraries.** These libraries help in data manipulation and numerical operations.



**Load the data from a CSV file** located at the specified path and store it in the variable named sales

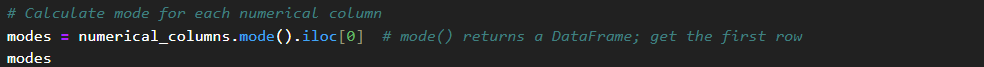


Identify Numerical Columns: Check which columns contain numerical data (e.g., sales amounts, discounts)

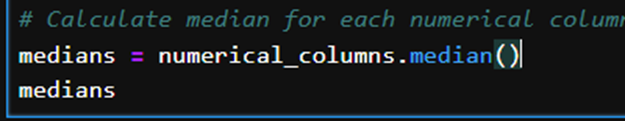
2

Calculate Basic Statistics:

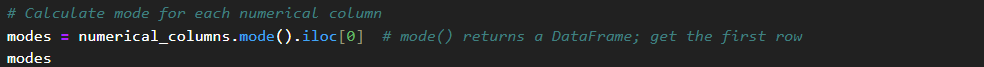
**Compute the average (mean) for each numerical column and store it in means.**



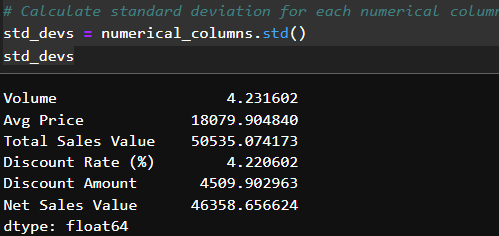
**Calculate the middle value (median) for each numerical column and store it in medians.**



**Find the most common value (mode) for each numerical column. Since mode() returns a DataFrame, take the first row and store it in modes.**



**Calculate the standard deviation (which measures spread) for each numerical column and store it in std\_devs.**



**Interpretation:**

**Mean:** Provides the central value around which other values cluster.

**Median:** Gives a better idea of the center when data is skewed.

**Mode:** Shows the most common value, useful for categorical data.

**Standard Deviation:** Indicates how spread out the values are.

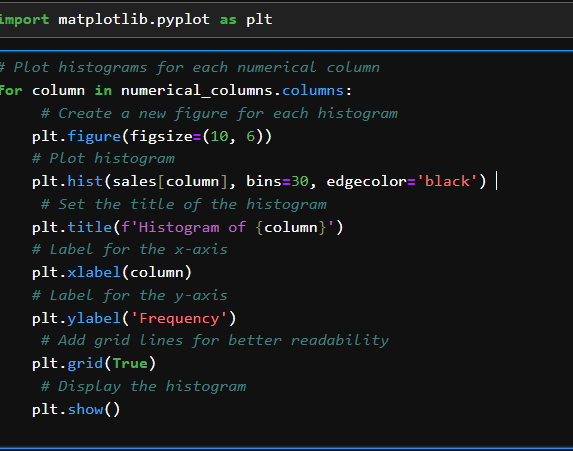
**Data Visualization**

Objective: To visualize the distribution and relationship of numerical

and categorical variables in the dataset

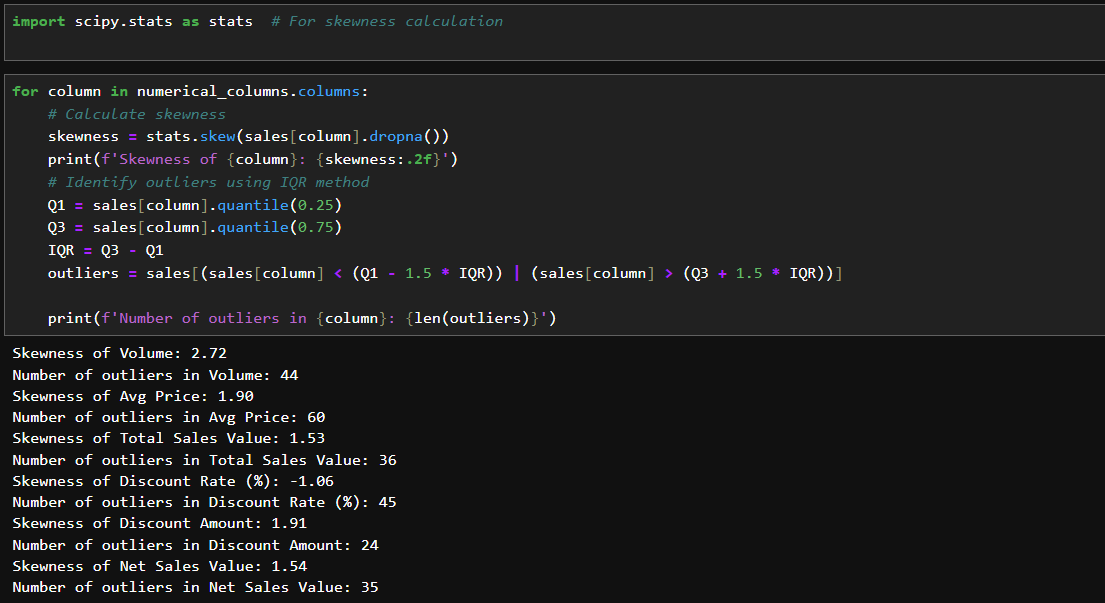
**Histograms:**

**Create histograms for each numerical column to show the distribution of values. Each histogram will be plotted in a separate figure with labels and grid lines.**

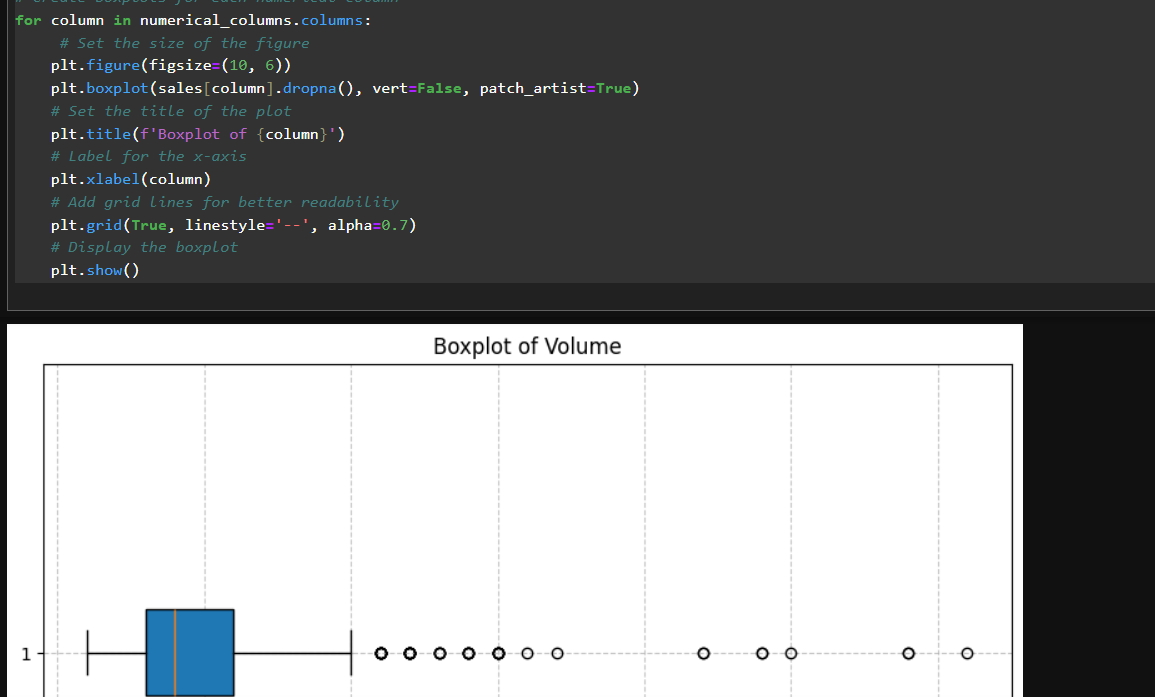


**Analysis:** Look for skewness (asymmetry) or outliers (values far from the majority)

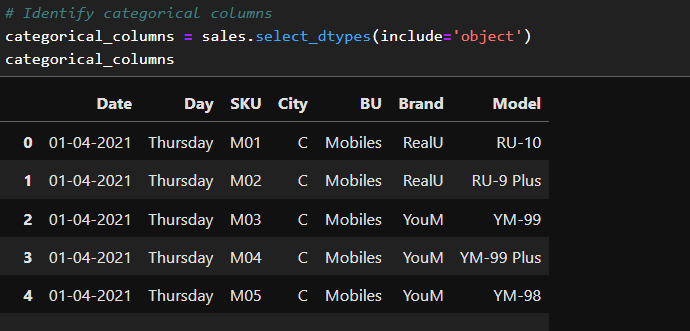
**Calculate skewness** (asymmetry) for each numerical column and print it. Also, find and count outliers using the Interquartile Range (IQR) method.



Boxplots:



**Create boxplots** for each numerical column. Boxplots show the spread and detect outliers, with each plot displayed in a separate figure.

**Select columns that contain categorical (non-numerical) data and store them in categorical\_columns.**

**Create bar charts** for each categorical column to show the frequency of each category. Rotate x-axis labels for better readability and add grid lines.

**Purpose:** Visualize the frequency of each category.



**Analysis:** Understand which categories are most or least frequent.

**Standardization of Numerical Variables**

**Objective:** Scale numerical variables for uniformity to improve model performance.

**Explanation:**

**Standardization:** Converts data to a common scale with a mean of 0 and a standard deviation of 1.



**Comparison:**

**Before:** Raw data with potentially different scales.

**After:** Data with standardized values, making it easier for machine learning algorithms to process.

**Conversion of Categorical Data into Dummy Variables**

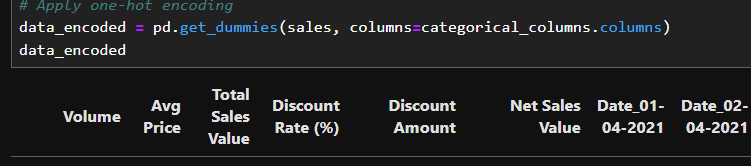
**Objective:** Prepare categorical variables for machine learning by converting them into numerical format.

**Explanation:**

**One-Hot Encoding:** Converts categorical variables into binary (0 or 1) columns.

**Purpose:** Ensures that categorical data can be used in algorithms that require numerical input.

Display Transformed Dataset:



**Conclusion**

**Summary of Key Findings:**

**Descriptive Statistics and Visualizations:** We analyzed the data to understand patterns and characteristics. For instance, we looked at how the numerical data is distributed, identified any outliers (extreme values), and noted any skewness (asymmetry) in the data. We also explored categorical data to see which categories are most common.

**Significant Discoveries:** We found key insights such as whether the data is skewed or if there are prevalent categories. These findings help us understand the data's structure and distribution better.

**Importance of Data Preprocessing:**

**Standardization:** By scaling the data so that all variables have the same mean and standard deviation, we ensure that each variable contributes equally to the analysis. This is important for creating fair and balanced models.

**One-Hot Encoding:** This process converts categorical data into numerical format, which is necessary for machine learning models to process and analyze the data effectively.

Top of Form

Bottom of Form

