EDA for Synthetic Dataset: Online Sales Data

**Introduction:**

In the vast and ever-changing world of e-commerce, online sales data serves as a treasure trove of information. It's not merely a collection of numbers; it's a treasure trove of insights waiting to be unearthed. As digital transactions weave a tapestry of interactions, each click, purchase, and feedback creates a narrative within the expansive dataset. In this blog post, we embark on an exploratory journey into the rich landscape of online sales data, delving deep into the intricacies and revelations it holds. This dataset, brimming with a plethora of attributes, serves as a dynamic canvas, each attribute a thread intricately weaving the story of unique transactions in the digital marketplace. Join us as we unravel the hidden patterns, decode customer behaviors, and navigate the labyrinth of online sales data to uncover the keys to strategic success in the ever-evolving e-commerce ecosystem.

What is Synthetic data?

* Synthetic data is artificially generated information that emulates the characteristics of real-world data. Here to generate this we have used Faker module (python method)

Importing libraries:

We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Seaborn, Missingno, Matplotlib

import numpy as np

import pandas as pd

import seaborn as sns

import missingno as msno

import matplotlib.pyplot as plt

from numpy.lib.npyio import DataSource

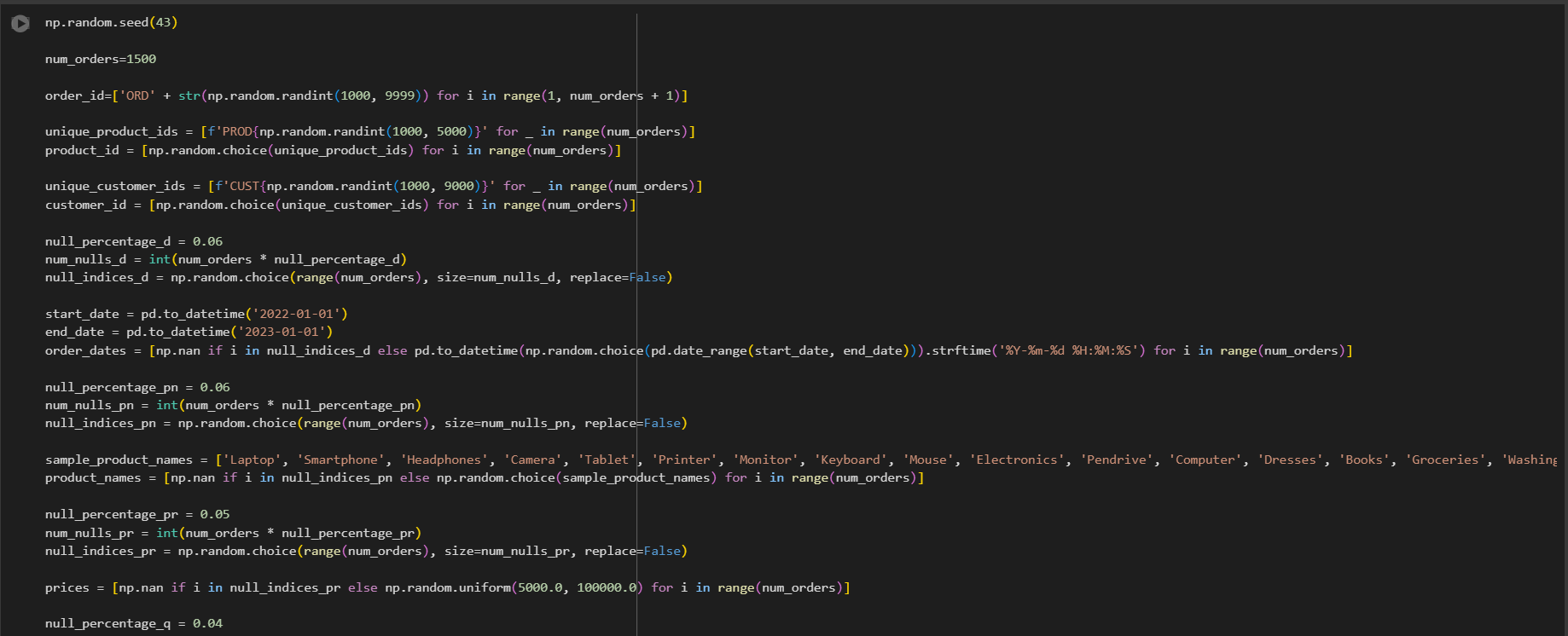
from sklearn.preprocessing import MinMaxScaler

import random

import datetime

Synthetic data generation:

Using faker in python



Total rows are 1500 and 13 columns

**Feature Analysis:**

Order ID:

The journey begins with the unique identifier for each transaction, the Order ID. It acts as our compass, guiding us through the labyrinth of orders, revealing the chronology and diversity within the dataset.

Order Date:

Time is of the essence, and the Order Date provides a temporal dimension to our exploration. Understanding when transactions occur allows us to identify trends, seasonality, and patterns that shape the e-commerce landscape.

Product ID and Product Name:

Products are the lifeblood of e-commerce, and their identifiers and names bring clarity to the array of offerings. The combination of Product ID and Product Name creates a comprehensive inventory map, showcasing the diversity of goods available.

Price and Quantity:

The financial aspects of transactions come into focus with Price and Quantity. These attributes lay the groundwork for understanding the monetary value of products and the volume of goods exchanged, forming the economic backbone of our analysis.

Total Sales:

Total Sales encapsulates the culmination of each transaction, representing the financial success tied to a combination of product prices and quantities sold. It serves as a key metric for evaluating overall performance.

Discount:

Discounts are powerful influencers in the world of e-commerce. The Discount attribute sheds light on the strategies employed to attract customers, impact sales, and foster loyalty.

Customer ID:

Customers are the heartbeat of any business. The Customer ID links transactions to individual customers, enabling the exploration of purchase histories, preferences, and the cultivation of lasting relationships.

Shipping Address:

E-commerce transcends digital realms to the tangible world through Shipping Address. It provides insights into geographic trends, shipping logistics, and customer demographics.

Payment Method:

Payment Method unravels the diverse ways customers choose to complete transactions. It's a reflection of convenience, security, and evolving preferences in the digital payment landscape.

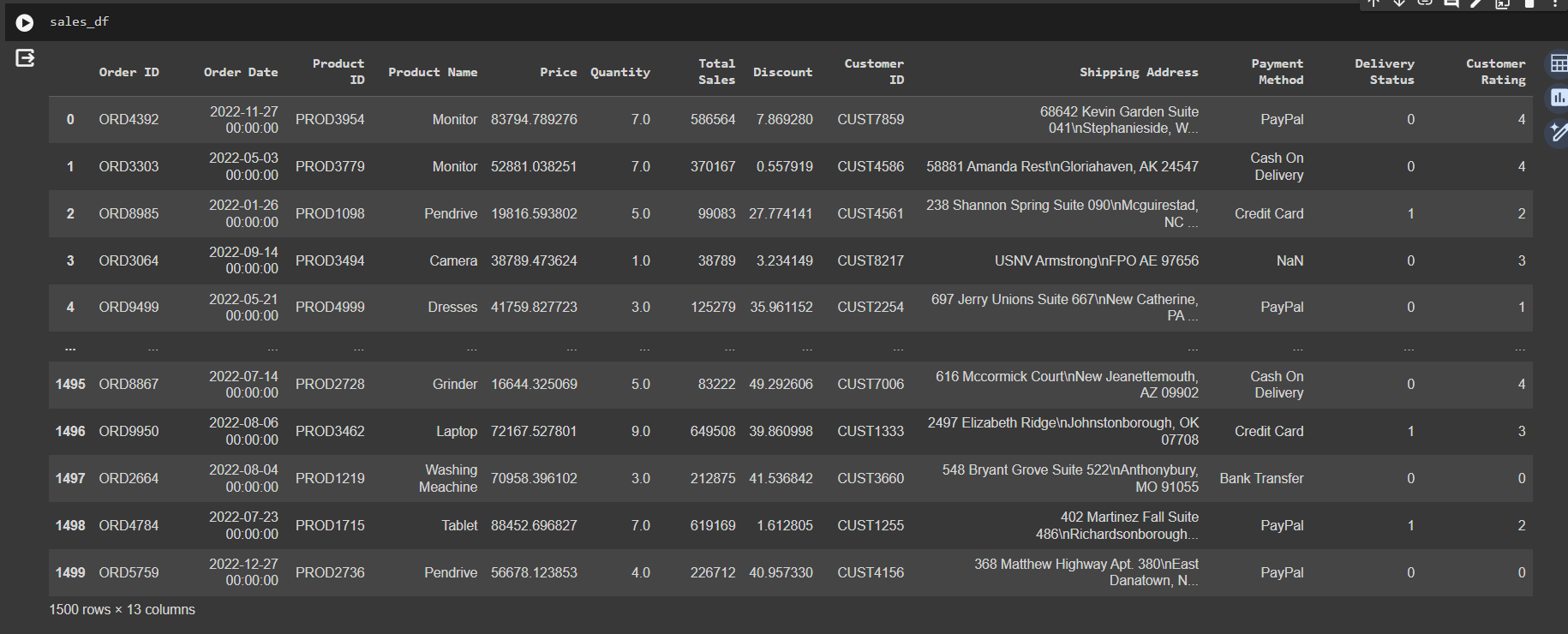
Delivery Status:

The journey concludes with Delivery Status, revealing the culmination of the customer experience. It provides a snapshot of order fulfillment, shipping efficiency, and the commitment to customer satisfaction.

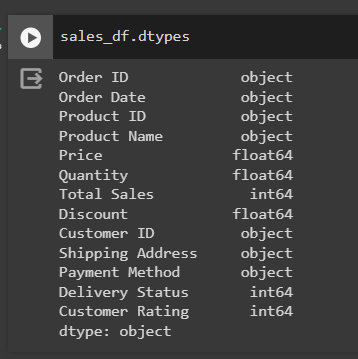
Customer Rating:

As a final note, the Customer Rating attribute encapsulates the voice of the customer. Their satisfaction levels, expressed through ratings, offer invaluable feedback for businesses to refine and optimize their operations.

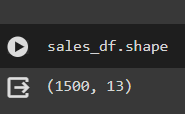
**Synthetic data:**

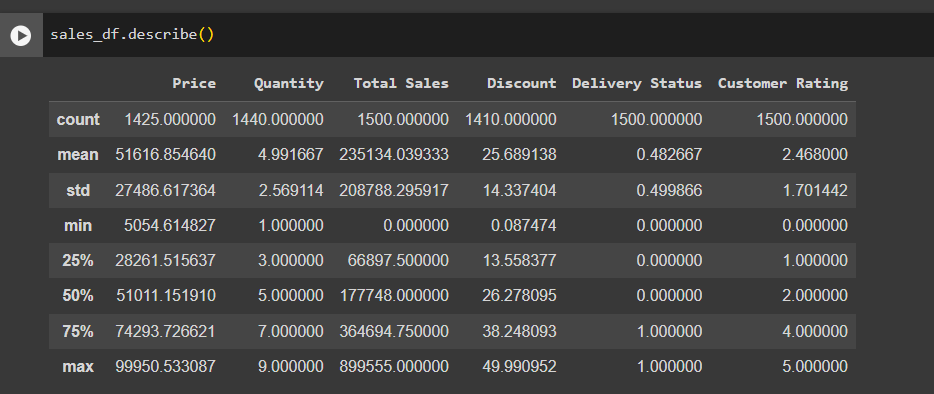


Data types:

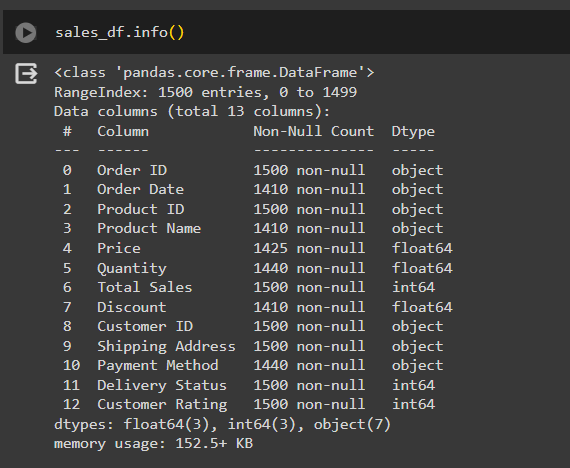


**Statistical Information:**

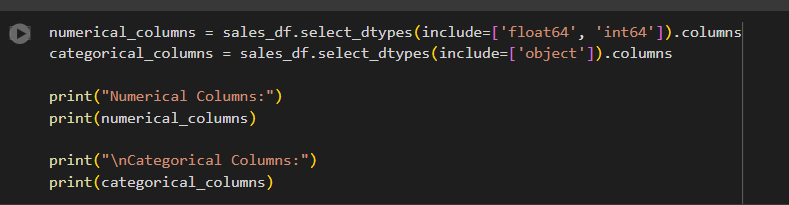
* df.shape is indeed an attribute of a Pandas DataFrame, and it returns a tuple representing the dimensions of the DataFrame.

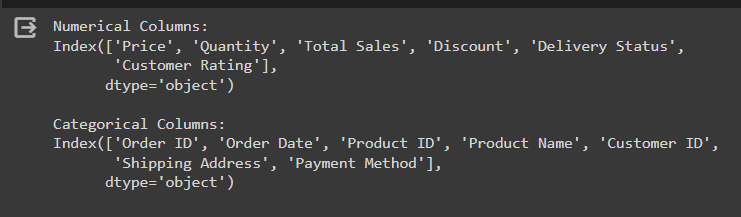


* The df.describe() function provides statistical summary measures of the numerical columns in the dataset, such as mean, standard deviation, minimum, maximum, and quartiles, offering insights into the central tendency and dispersion of the data.



The df.info() provides a concise overview of the dataset, including the data types, non-null counts, and memory usage. Together, these functions assist in a quick understanding of both the statistical properties and structural characteristics of the dataset.



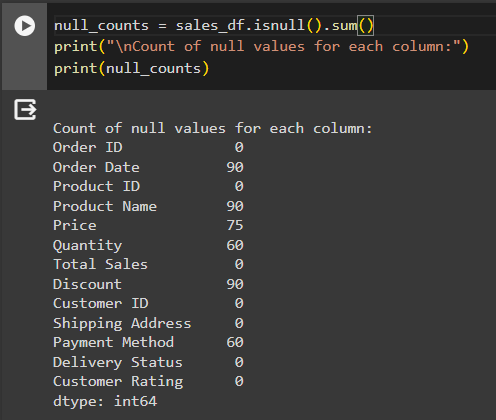


* Distinguish between numerical and categorical attributes within a DataFrame. The 'numeric\_columns' list is generated by including columns with data types 'float64' and 'int32,' representing numerical features. Simultaneously, the 'non\_numeric\_columns' list is created by excluding columns of numeric data types, encompassing the categorical attributes in the dataset. The subsequent print statements display these two lists, aiding in the clear identification of numeric and non-numeric columns in the DataFrame.

**Data Pre-Processing**

Detecting Null values:

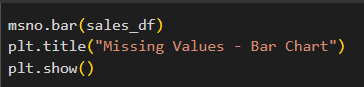
The presence of missing values within a dataset poses significant challenges to data analysis. Not only does it introduce the potential for biased results, but it also undermines the accuracy of predictions and complicates the interpretation of the overall dataset. Properly addressing missing values is essential to prevent the derivation of misleading conclusions and to guarantee the reliability of insights drawn from the data. Mitigating the impact of missing values is paramount in ensuring robust and trustworthy analyses, facilitating accurate decision-making processes in various domains.

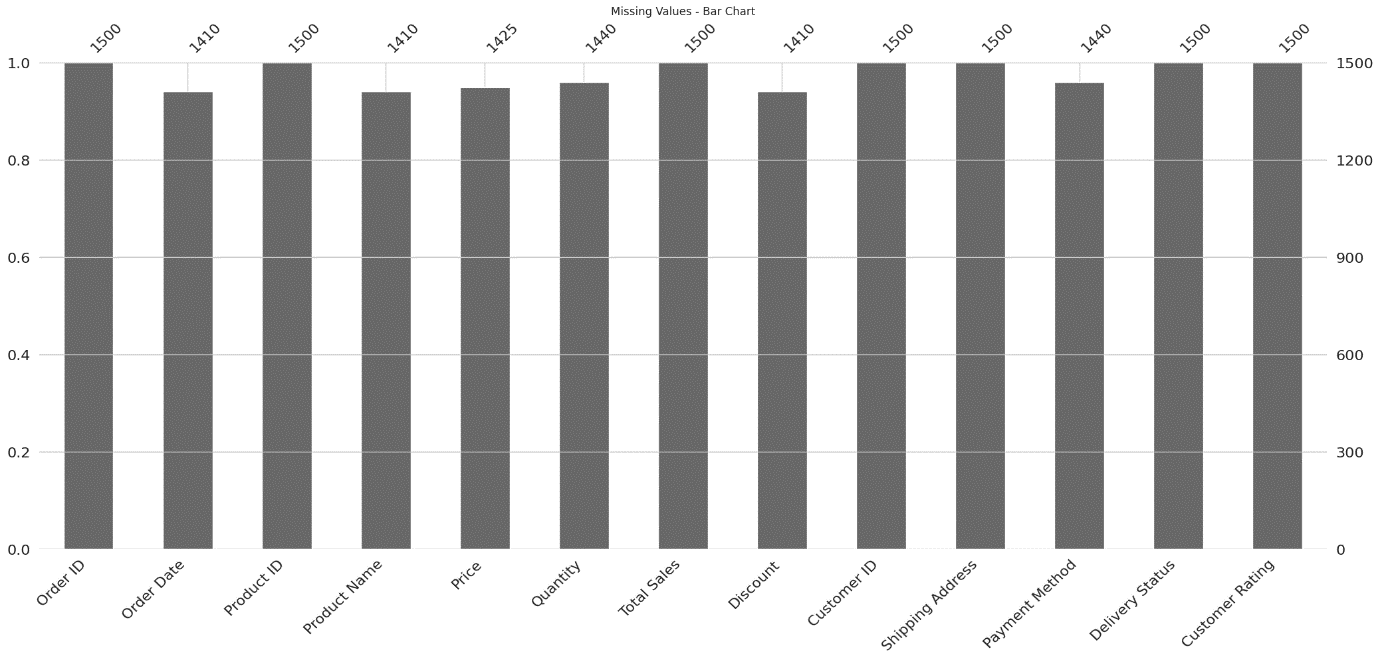
What happens if we have Missing values in the Data set?

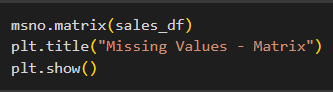
* Having missing values in a dataset can make data analysis tricky. It might lead to biased results, reduce the accuracy of predictions, and make it challenging to understand the overall picture. Dealing with missing values properly is crucial to avoid drawing incorrect conclusions and ensure reliable insights from the data.

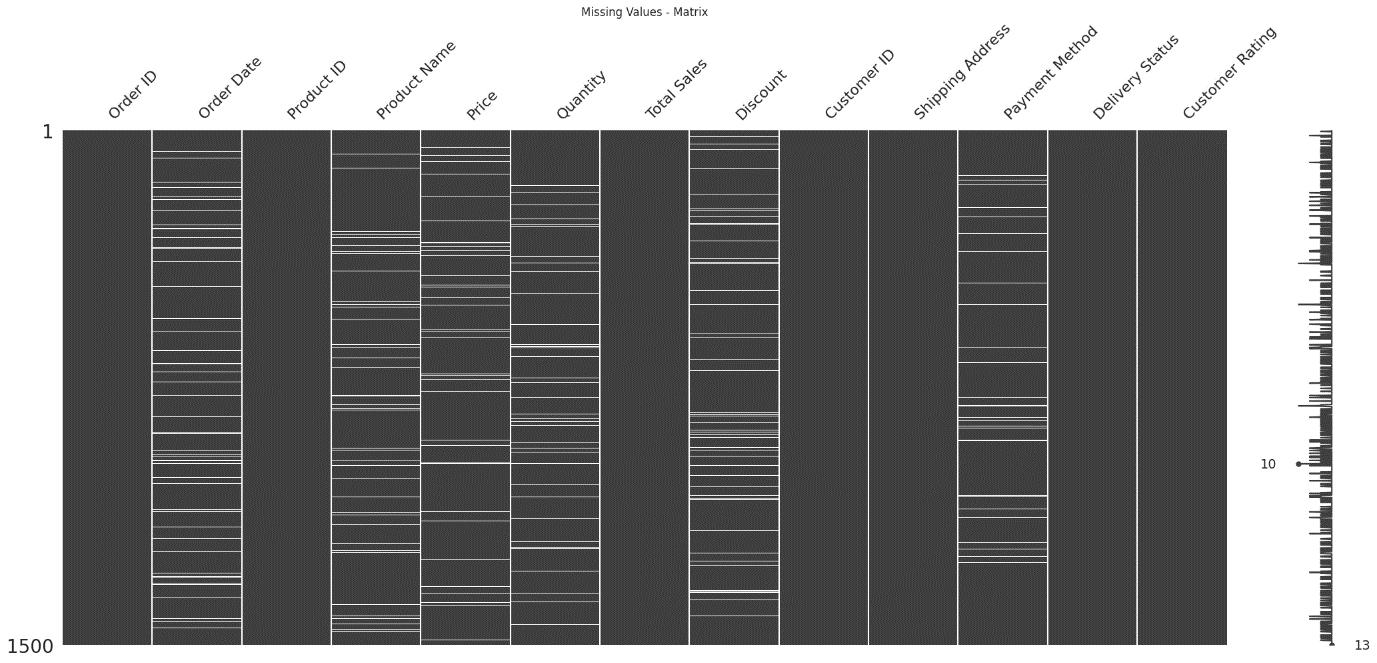
**Data Quality:**

Using missing Library:









**NULL Value Imputation Methods:**

Imputing missing values is a crucial step in data preprocessing, ensuring the completeness and reliability of the dataset. There are various strategies for handling missing data

Mean/Median Imputation:

* Replace missing values with the mean or median of the observed values in the variable.
* Suitable for numerical data with a normal distribution.

Mode Imputation:

* Replace missing values with the most frequently occurring value in the variable.
* Suitable for categorical variables.

Forward Fill (or Previous Value) and Backward Fill (or Next Value):

* Use the last known value to fill missing values (forward fill) or the next available value (backward fill).
* Suitable for time series data where values are often sequential.

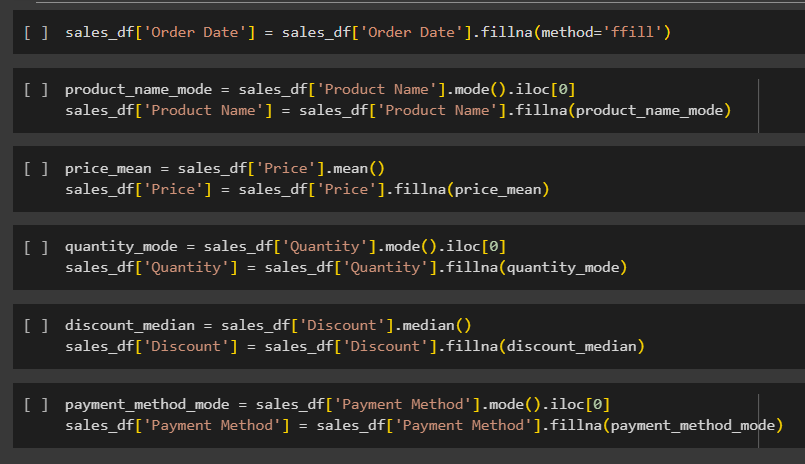
K-Nearest Neighbors (KNN) Imputation:

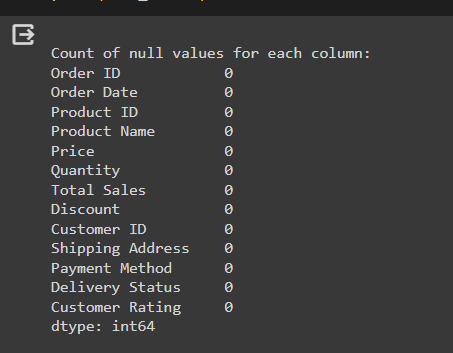
* Impute missing values by considering the values of k-nearest neighbors.
* Suitable for datasets where similar observations are likely to have similar values.

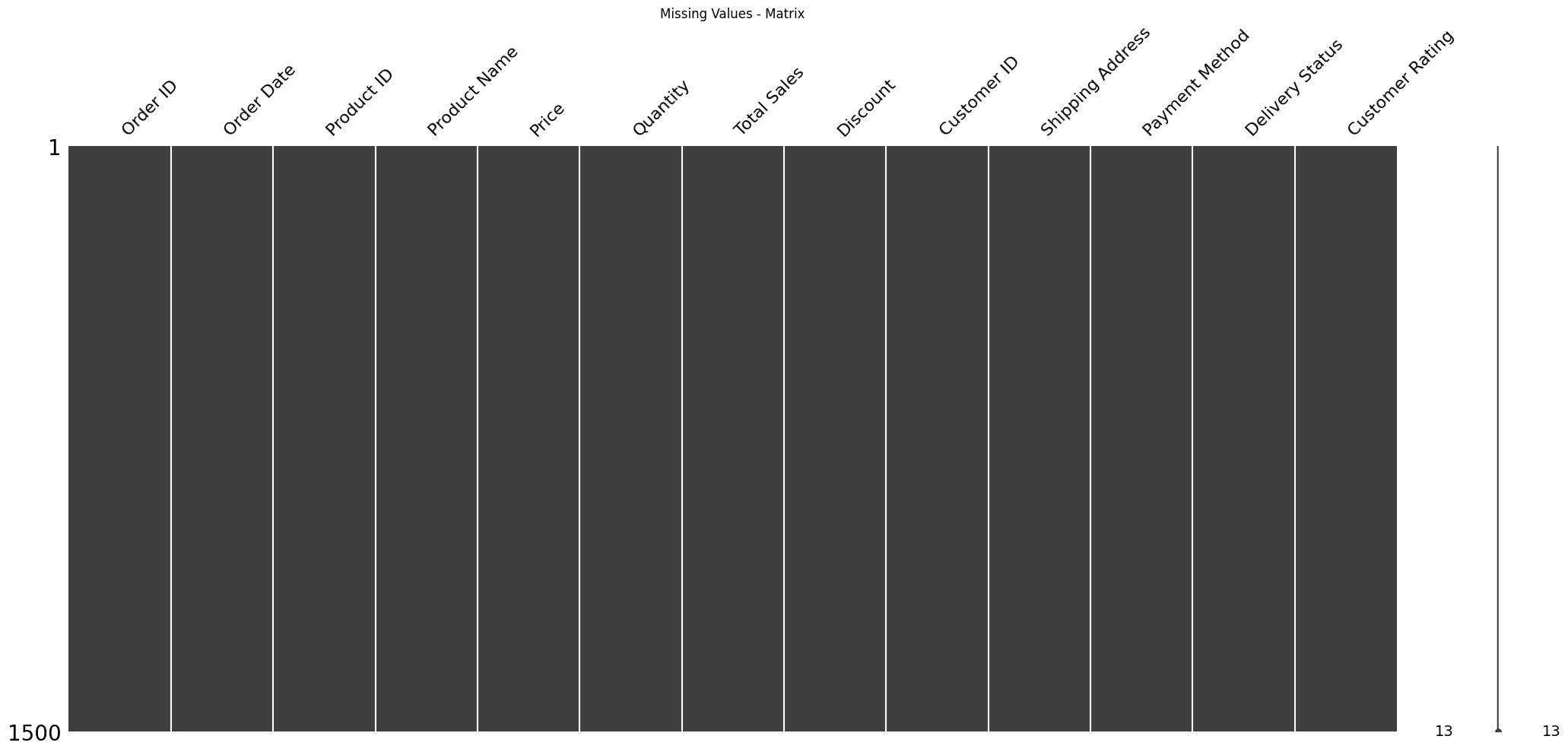
Iterative Imputation:

* An iterative process where each variable is modelled as a function of other variables, iteratively updating missing values.
* Beneficial when variables are interdependent.

After imputing null values:







**Outlier Detection:**

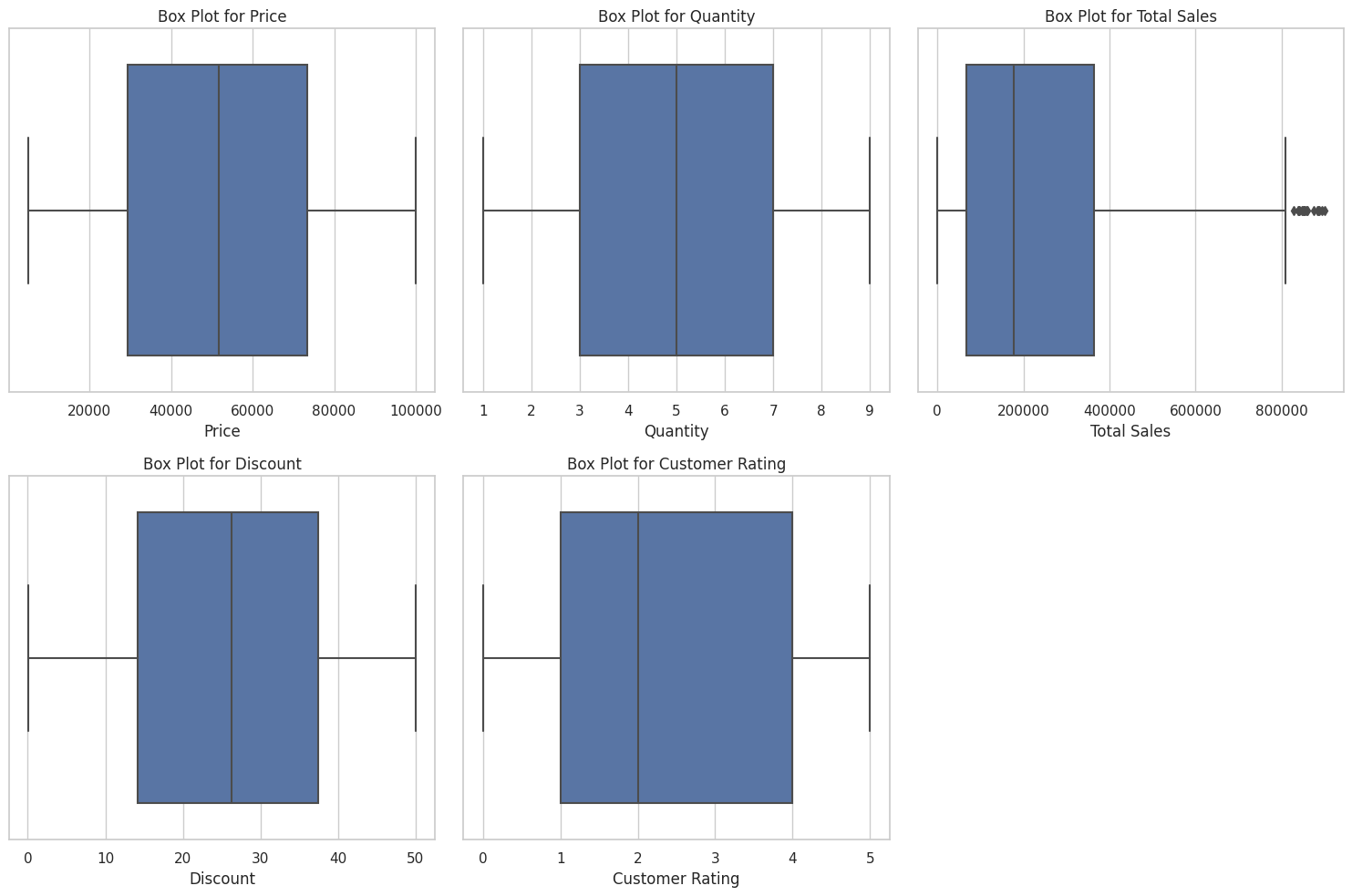
* Outliers are data points that deviate significantly from the rest of the data in a dataset.
* They are observations that lie at an abnormal distance from other values in a random sample from a population.
* In other words, outliers are data points that are unusually high or low compared to the majority of the data.
* Outliers can be the result of errors in data collection, or measurement variability, or they may indicate a real and important pattern in the data.

What are Outliers?

* Outliers are data points that deviate significantly from the rest of the data in a dataset.
* They are observations that lie at an abnormal distance from other values in a random sample from a population.
* In other words, outliers are data points that are unusually high or low compared to the majority of the data.
* Outliers can be the result of errors in data collection, or measurement variability, or they may indicate a real and important pattern in the data.

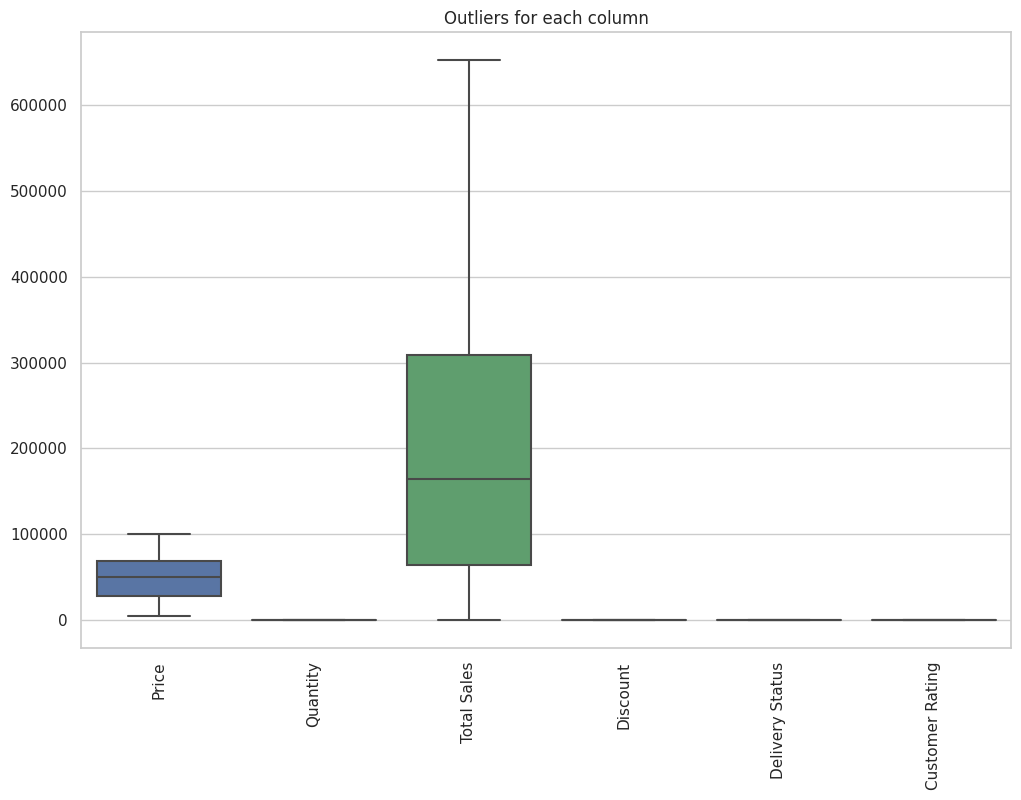
Outlier detection using Box-plot:

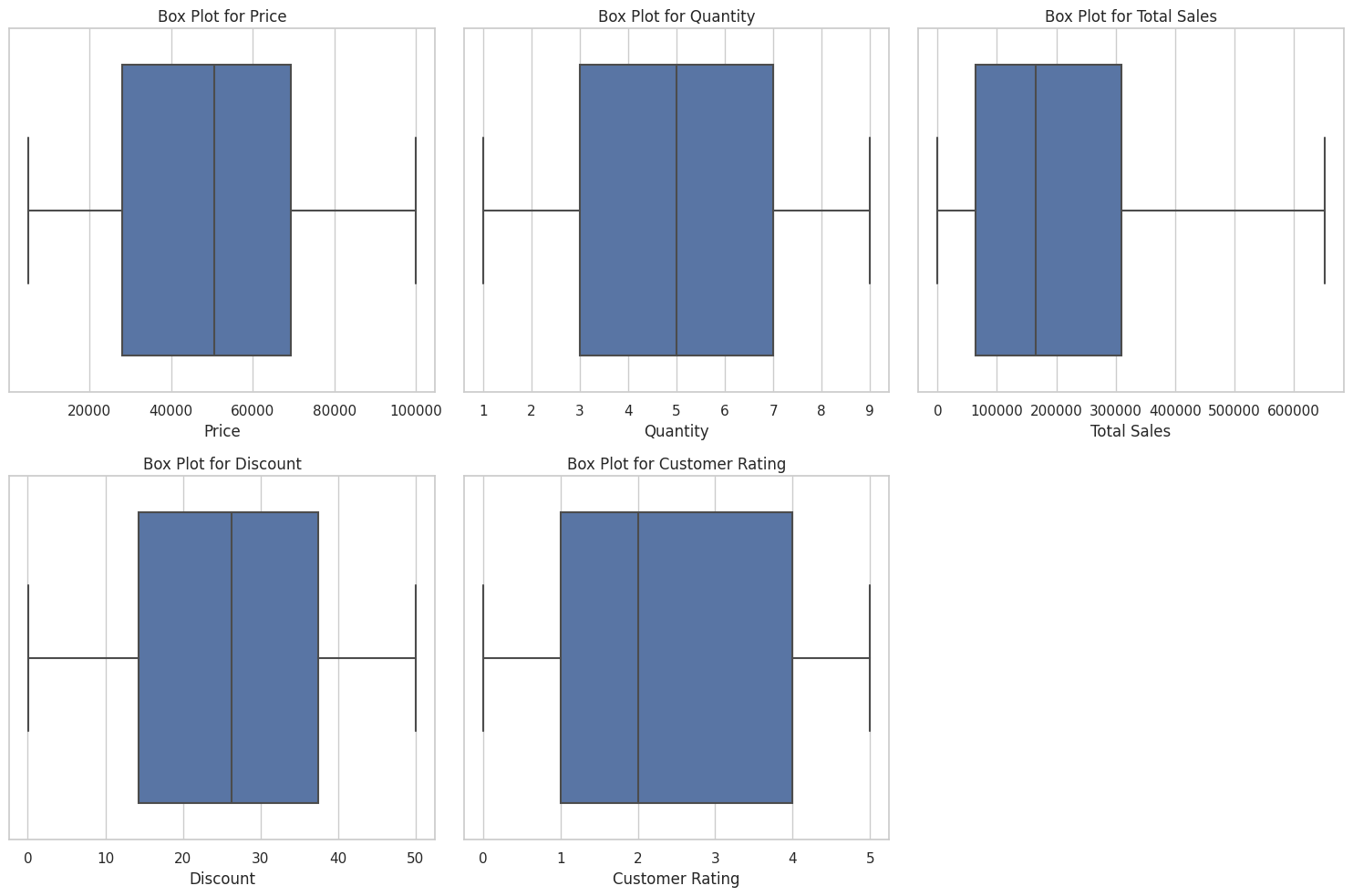
* A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It provides a visual summary of the central tendency, spread, and skewness (asymmetry) of the data. Box plots are particularly useful for identifying the presence of outliers and comparing the distributions of different groups.



**Removing outliers:**

* Removing outliers from a dataset is crucial for several reasons. Outliers can distort statistical measures, such as the mean and standard deviation, leading to inaccurate representations of central tendency and spread. In the context of machine learning, outliers can disproportionately influence algorithms, impacting model performance and generalization to new data. Violating assumptions of normal distribution can affect the validity of statistical tests.

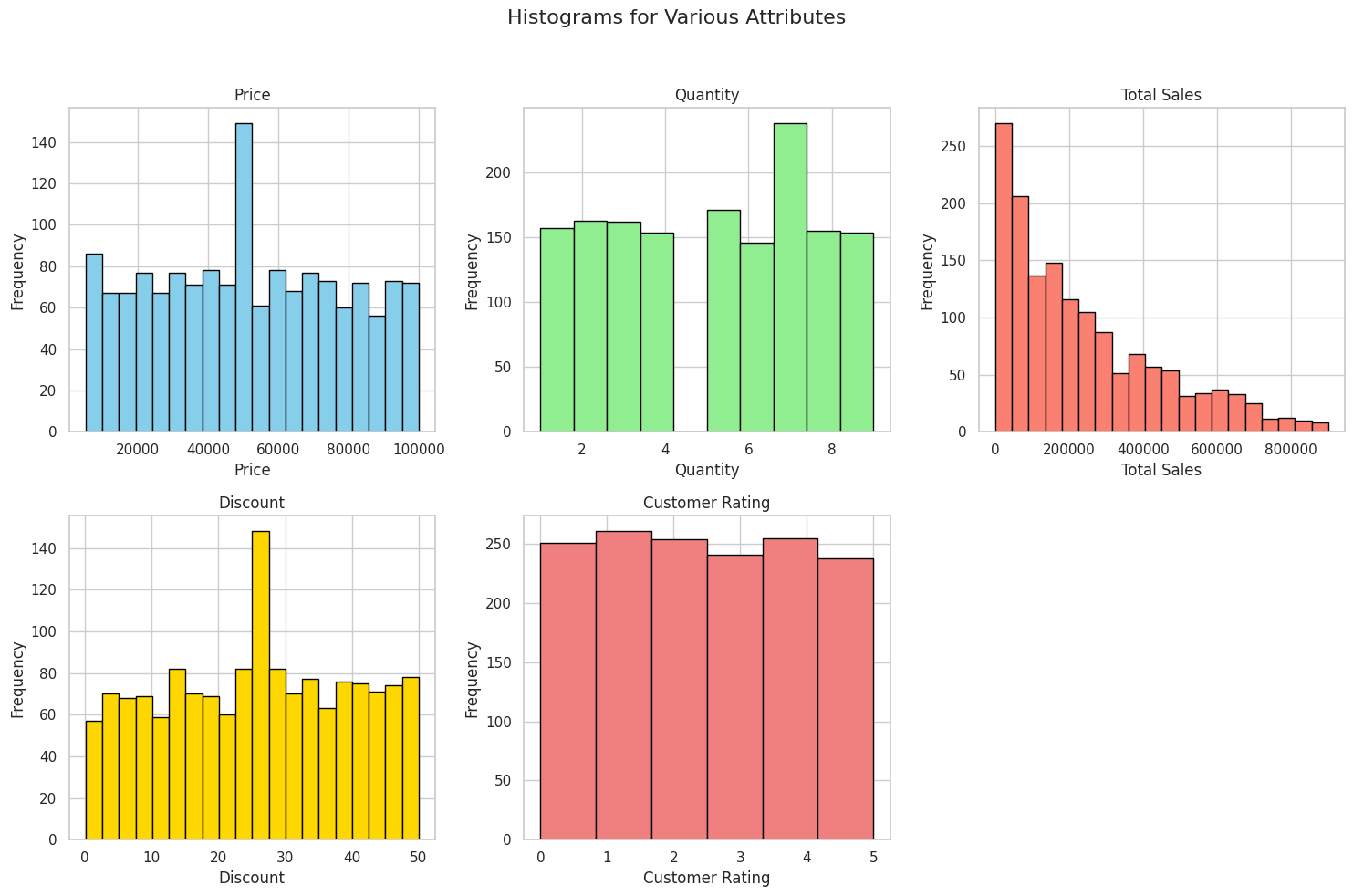




**Exploratory Data Analysis (EDA) Data visualization:**

* Exploratory Data Analysis (EDA) is a crucial phase in the data analysis process that involves visually exploring and summarizing datasets to gain insights, identify patterns, and detect outliers. Through the use of various graphical and statistical techniques, EDA helps analysts and data scientists understand the underlying structure of the data, assess its distribution, and uncover potential relationships between variables. Visualizations such as histograms, scatter plots, box plots, and heatmaps are commonly employed to represent data patterns and trends effectively. EDA not only aids in formulating hypotheses but also guides subsequent steps in the data analysis pipeline, facilitating informed decision-making and problem-solving. The primary goal of EDA is to provide a comprehensive and intuitive understanding of the dataset before more advanced modeling or hypothesis testing is undertaken.

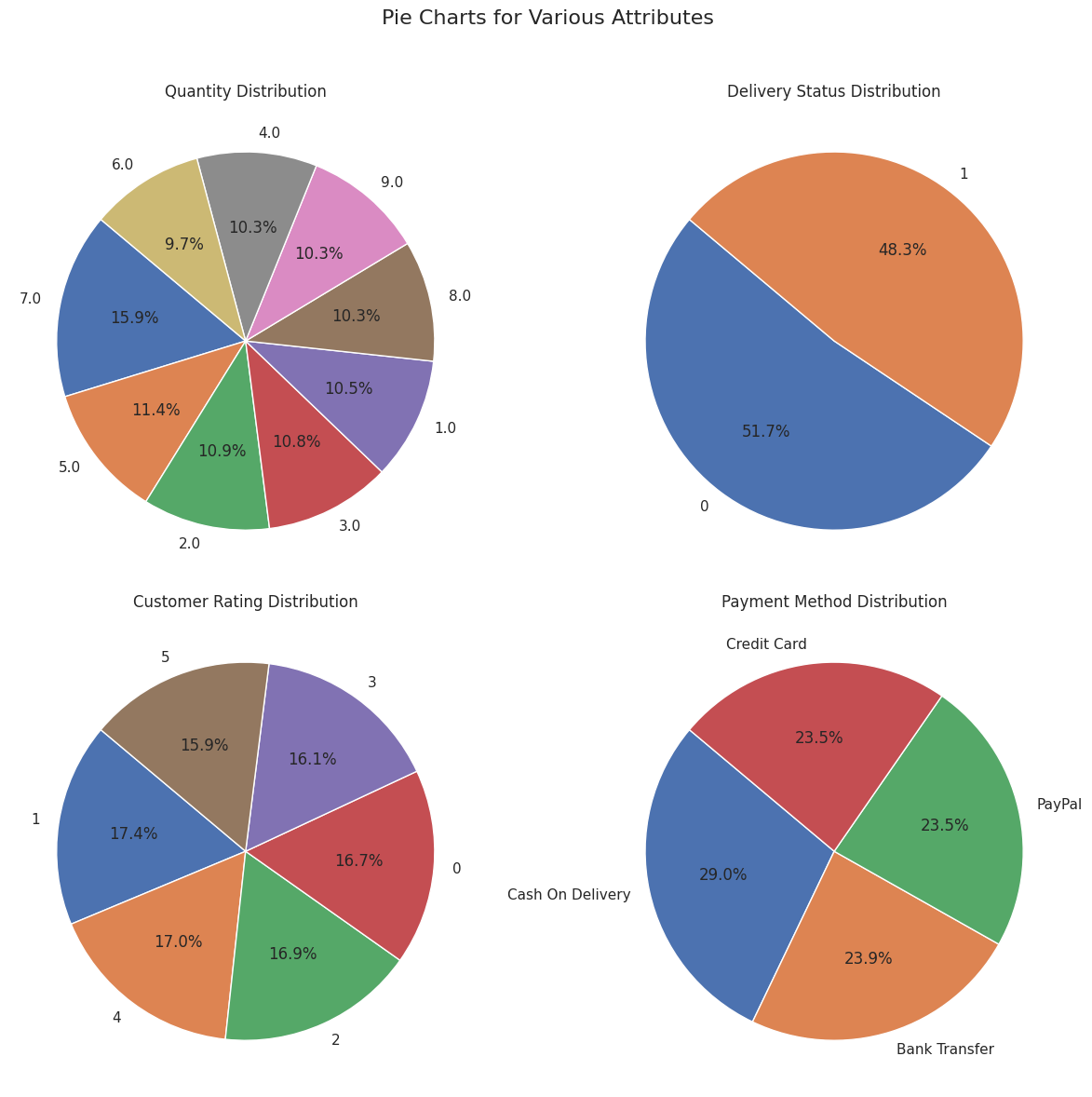
**Histogram:**



* The histogram for 'Price' reveals that the majority of products fall within lower price ranges, with a peak around $20,000. This skewed distribution suggests a concentration of relatively lower-priced items.
* For 'Quantity,' the histogram displays a relatively uniform distribution, indicating that the quantity of products sold is evenly spread across different ranges. Most transactions involve quantities between 2 and 6.
* Analyzing 'Total Sales' unveils a positively skewed distribution, implying that a significant portion of transactions results in lower total sales. A notable peak occurs around $50,000.
* The 'Discount' histogram illustrates that discounts are prevalent, with a peak in the 10-20% range. This information is valuable for understanding the impact of discounts on sales.
* Examining the 'Customer Rating' histogram showcases a nearly uniform distribution of ratings. Most customers provide ratings between 3 and 5, indicating overall satisfaction.

Histograms provide a visual snapshot of the distribution of key attributes in our synthetic sales data. Understanding these distributions is essential for identifying trends, making informed decisions, and optimizing business strategies. The positively skewed 'Total Sales' distribution and concentrated 'Price' distribution are noteworthy findings that warrant further investigation. As we continue to explore and analyze the data, these visualizations serve as a foundation for more in-depth insights.

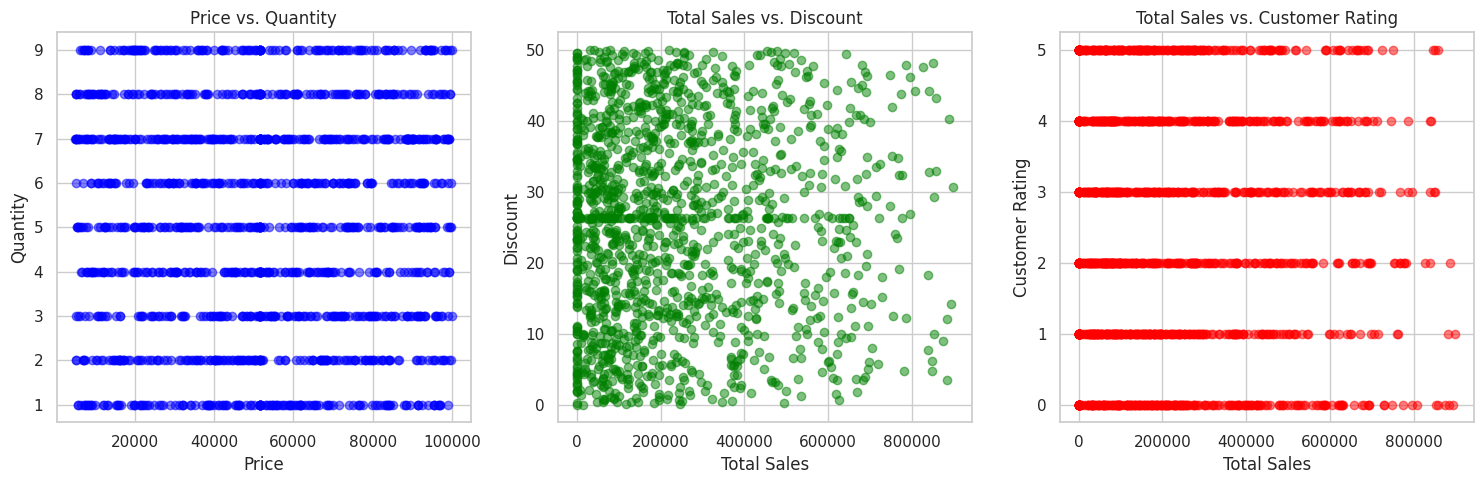
**Pie Chart:**



* Our first foray into the data landscape unveils the 'Quantity Distribution' through a pie chart. The visual narrative paints a picture of diverse transactional quantities, with the majority falling within the 2 to 6 range. This suggests a dynamic array of product quantities contributing to the overall sales scenario.
* Shifting our gaze to the 'Delivery Status Distribution,' the second pie chart illuminates the efficiency of order fulfillment. Dominated by 'Not Delivered' and 'Delivered' categories, this visualization offers insights into the success of the delivery process, providing a snapshot of operational efficacy.
* The third pie chart captures the essence of customer satisfaction with the 'Customer Rating Distribution.' Ratings predominantly cluster between 3 and 5, indicating a high level of contentment among customers. This positive sentiment bodes well for the product and service quality.
* Our final exploration dives into the diverse landscape of 'Payment Method Distribution.' Common modes, such as 'Credit Card,' 'PayPal,' 'Cash On Delivery,' and 'Bank Transfer,' come to life. This pie chart serves as a valuable tool for discerning customer preferences in payment modes.

Pie charts emerge as compelling storytellers, distilling attribute distributions into visually digestible insights. Their efficacy in conveying categorical data proportions enables stakeholders to swiftly grasp the intricate dynamics of sales data. The diverse spread in quantities, efficient delivery processes, high customer satisfaction, and varied payment methods underscore the adaptability and resilience of the business.

**Scatter Plot:**



Price vs. Quantity:

* The first scatter plot navigates the terrain of product pricing and quantity sold. Each point on the plot represents a transaction, with the x-axis showcasing the product price and the y-axis depicting the corresponding quantity. The scatter of points provides insights into whether there's a correlation or clustering pattern between price and quantity.

Total Sales vs. Discount:

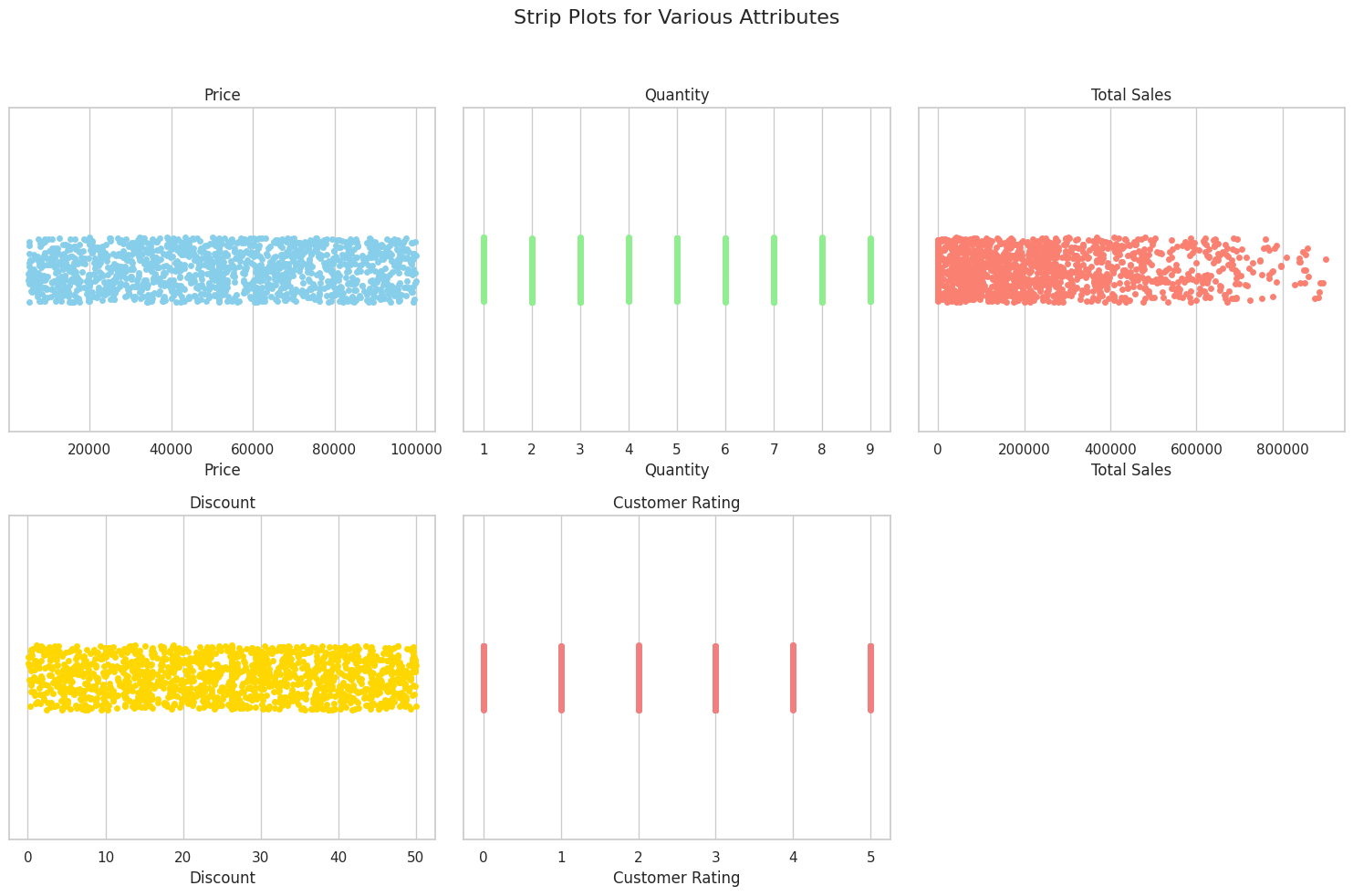
* Moving forward, we pivot to the dynamics of total sales and discounts. This scatter plot juxtaposes the total sales amount against the applied discount for each transaction. A closer look at the spread of points allows us to discern any discernible trends, such as the impact of discounts on the overall sales figures.

Total Sales vs. Customer Rating:

* The final frontier explores the relationship between total sales and customer ratings. By plotting total sales on the x-axis and customer ratings on the y-axis, we aim to uncover whether higher sales correlate with better customer ratings. The scatter plot provides a snapshot of the distribution, offering insights into the customers' satisfaction levels concerning different sales volumes.

As we navigate through these scatter plots, the intricate dance of data points reveals underlying stories within the sales dataset. From pricing strategies to the influence of discounts on sales and the intersection of customer satisfaction, these visualizations serve as a compass for businesses, guiding strategic decisions and fostering a deeper understanding of their sales dynamics.

**Strip Plot:**



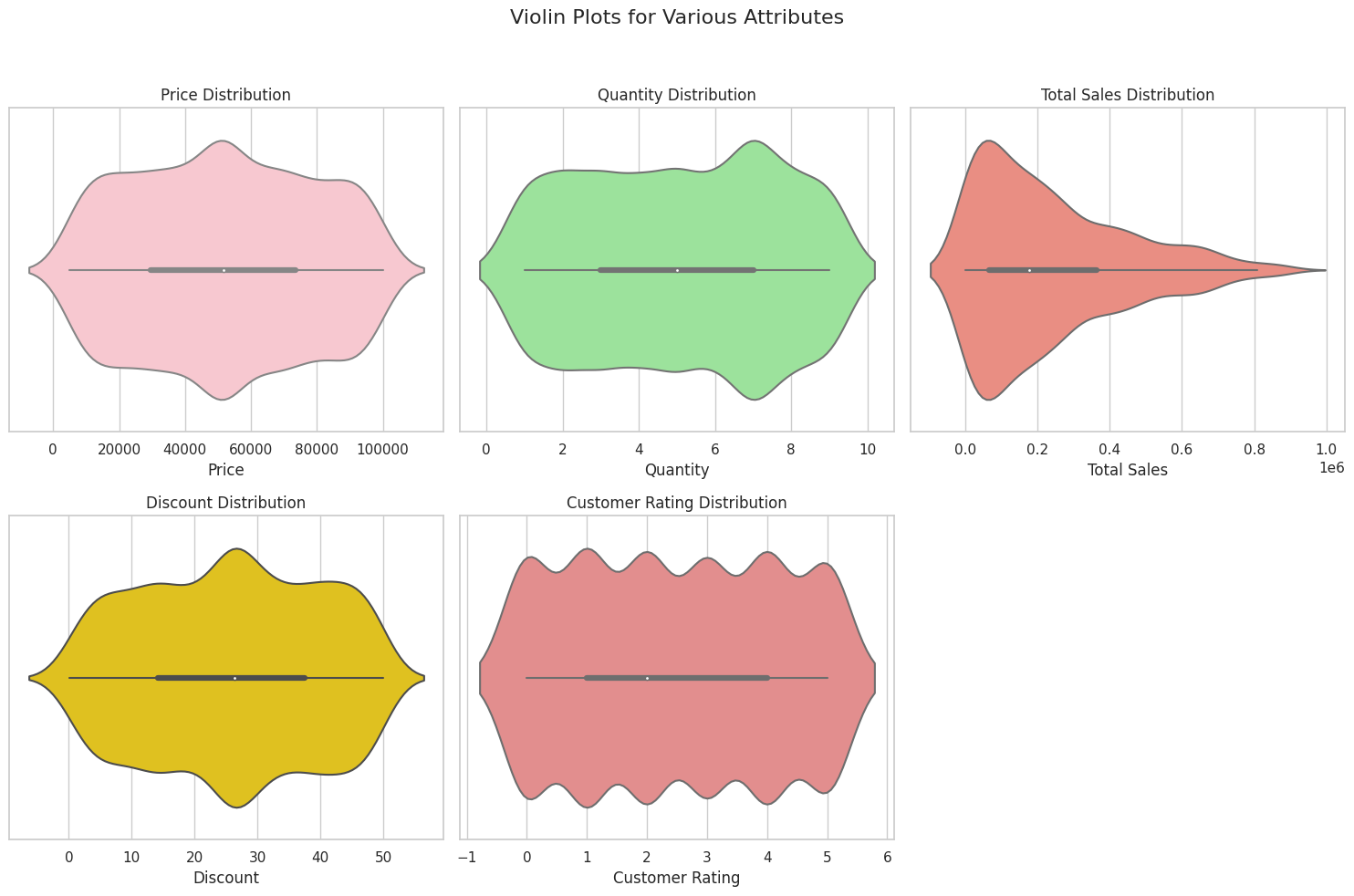
* Sky-blue dots depict the concentration of product prices, offering a quick overview of price ranges.
* A light-green strip plot showcases the uniform distribution of product quantities sold, aiding in identifying common transaction sizes.
* Salmon-colored dots in the strip plot for 'Total Sales' highlight the diversity in transaction values, providing insights into sales patterns.
* A gold-themed strip plot for 'Discount' illustrates the distribution of discount percentages, aiding in understanding the prevalence and impact of discounts.
* Light-coral dots in the 'Customer Rating' strip plot reveal the distribution of ratings, indicating a generally positive customer sentiment.

The strip plots have provided us with a visually rich exploration of key attributes within our sales dataset. Each strip plot, whether unveiling the spread of prices, distribution of quantities, variation in total sales, impact of discounts, or the landscape of customer ratings, contributes unique insights into the dynamics of our business.

These visualizations are instrumental in identifying patterns, outliers, and trends that might be overlooked in raw numerical data. The nuanced understanding gained through strip plots empowers businesses to optimize pricing strategies, streamline inventory management, tailor discount offerings, and enhance customer satisfaction.

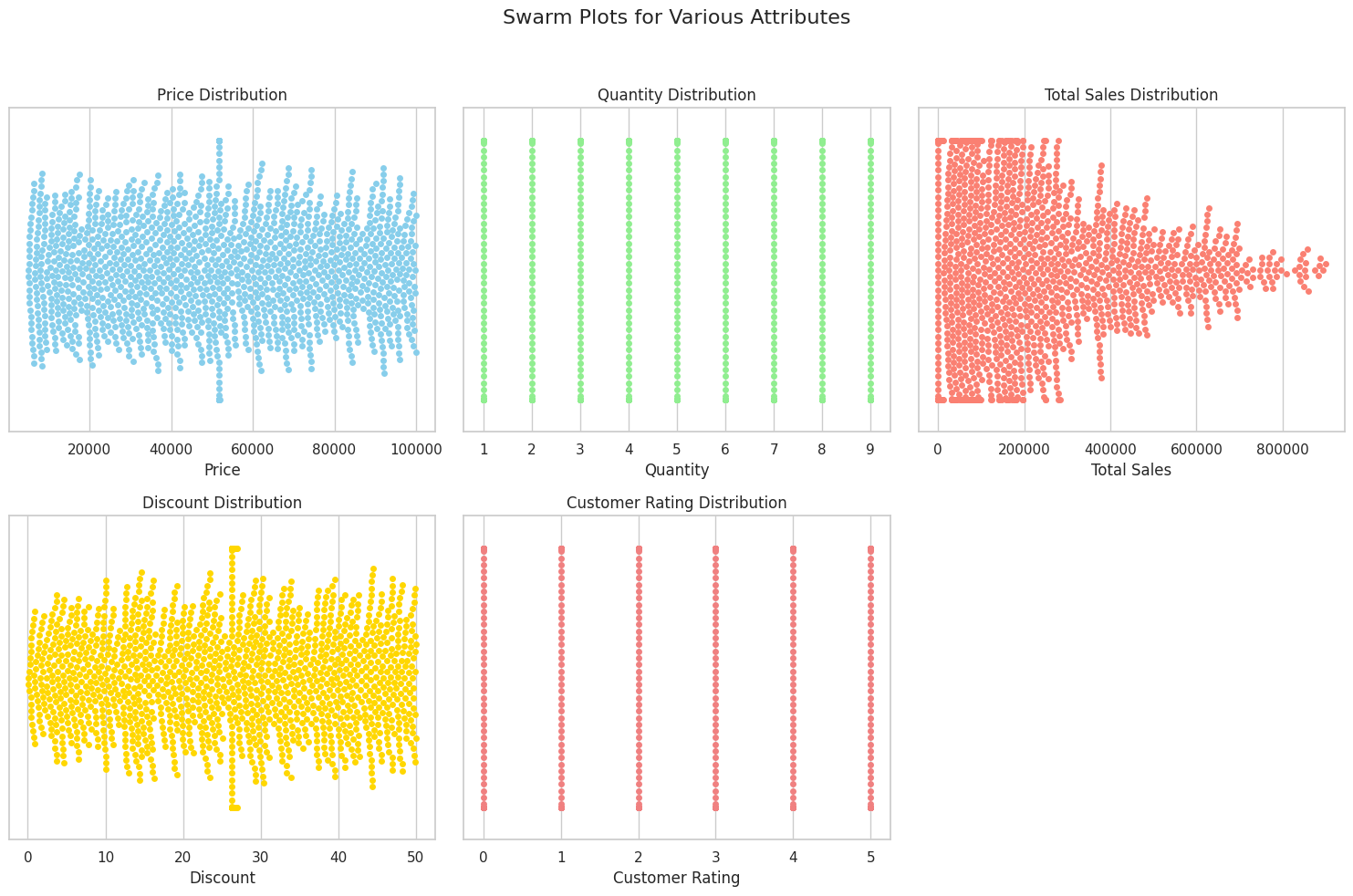
As we navigate the intricate landscape of sales data, the utilization of visualization tools like strip plots proves invaluable for decision-makers. By harnessing the power of visual storytelling, businesses can adapt strategies, refine operations, and ultimately foster a more responsive and customer-centric approach.

**Violion Plot:**



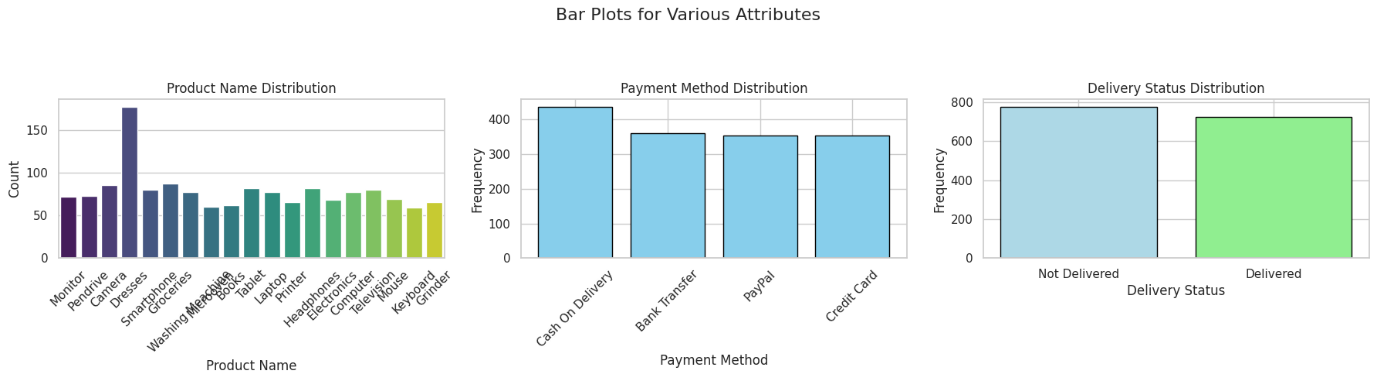
* The exploration of our sales dataset through violin plots has provided valuable insights into the distribution and variation of key attributes, namely 'Price,' 'Quantity,' 'Total Sales,' 'Discount,' and 'Customer Rating.' Each violin plot has contributed to a deeper understanding of the dataset's nuances, enabling businesses to make informed decisions.
* The visualizations have unveiled diverse pricing clusters, highlighted popular purchase quantities, and depicted the spread of revenue generation. Understanding discount distributions and customer satisfaction scores is essential for refining pricing and promotional strategies. The collective information derived from these violin plots equips businesses with actionable insights for optimizing operations and enhancing customer relations.
* Violin plots, with their ability to showcase data distributions effectively, serve as powerful tools for data exploration and decision-making in the realm of sales analytics. Through these visualizations, businesses can identify patterns, outliers, and concentration areas, paving the way for strategic adjustments and sustainable growth.

**Swarm Plot:**



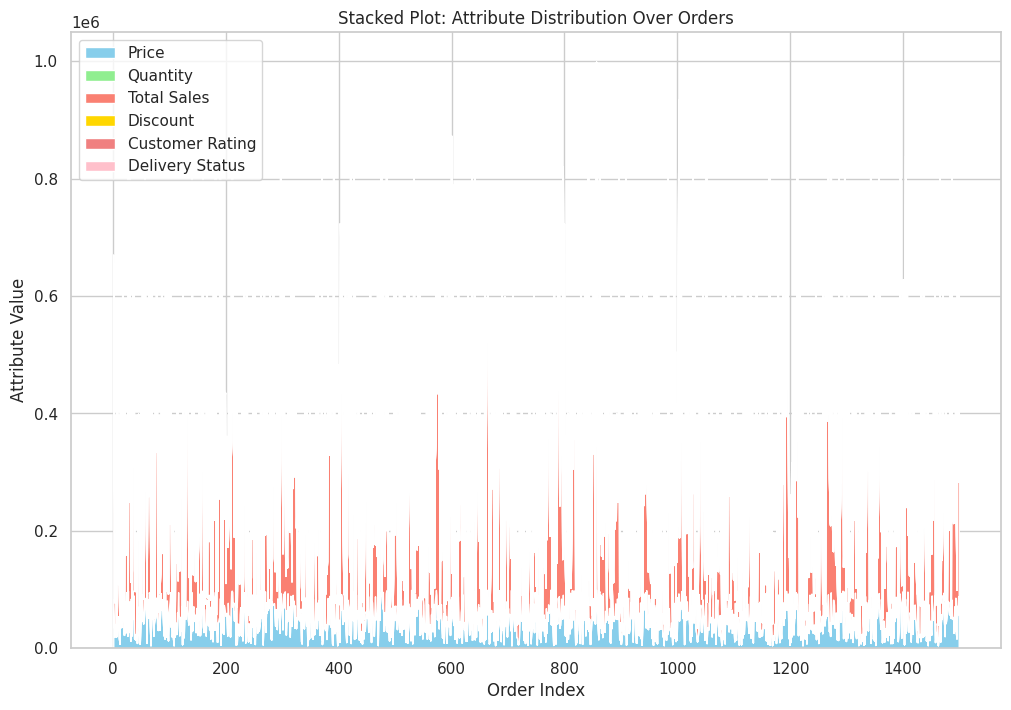
* The swarm plots provide a detailed view of the distribution of various attributes in our sales dataset. Each attribute, including 'Price,' 'Quantity,' 'Total Sales,' 'Discount,' and 'Customer Rating,' is visualized individually, allowing for a closer inspection of data points.
* Observing the swarm plots, businesses can identify patterns and clusters within each attribute's distribution. For instance, in the 'Price' distribution, the swarm plot showcases how different price points are populated across transactions. Similarly, the 'Customer Rating' swarm plot reveals the density of ratings, highlighting areas where customers tend to rate products.
* Swarm plots are particularly effective in displaying the granularity of data, making them a valuable tool for identifying outliers and understanding the density of attribute values. This level of detail can aid businesses in making informed decisions related to pricing strategies, inventory management, and customer satisfaction initiatives.
* The swarm plots contribute to a nuanced understanding of attribute distributions, providing valuable insights that can guide strategic actions for businesses aiming to optimize their operations and enhance customer experiences.

**Bar Plot:**



* The bar plots for 'Product Name,' 'Payment Method,' and 'Delivery Status' offer valuable insights into key aspects of the synthetic sales dataset. Each visualization provides a clear representation of the distribution of these attributes, aiding in understanding customer preferences, payment method popularity, and order fulfillment status.
* From the 'Product Name' distribution, we can observe the frequency of each product, helping identify popular items and potential areas for marketing focus. The 'Payment Method' bar plot illustrates the preferred methods of payment among customers, providing essential information for optimizing the payment process.
* The 'Delivery Status' distribution sheds light on the efficiency of order fulfillment, distinguishing between delivered and undelivered orders. This insight can guide improvements in the delivery process to enhance customer satisfaction.
* These visualizations contribute to a comprehensive understanding of the synthetic sales dataset, facilitating data-driven decision-making for business strategies and improvements.

**Stack Plot:**



Preparing the Canvas:

* Before diving into the stacked plot, let's ensure our data is ready for this visual spectacle. We've filled in any missing values for the selected attributes to guarantee a seamless representation.

Crafting the Stacked Plot:

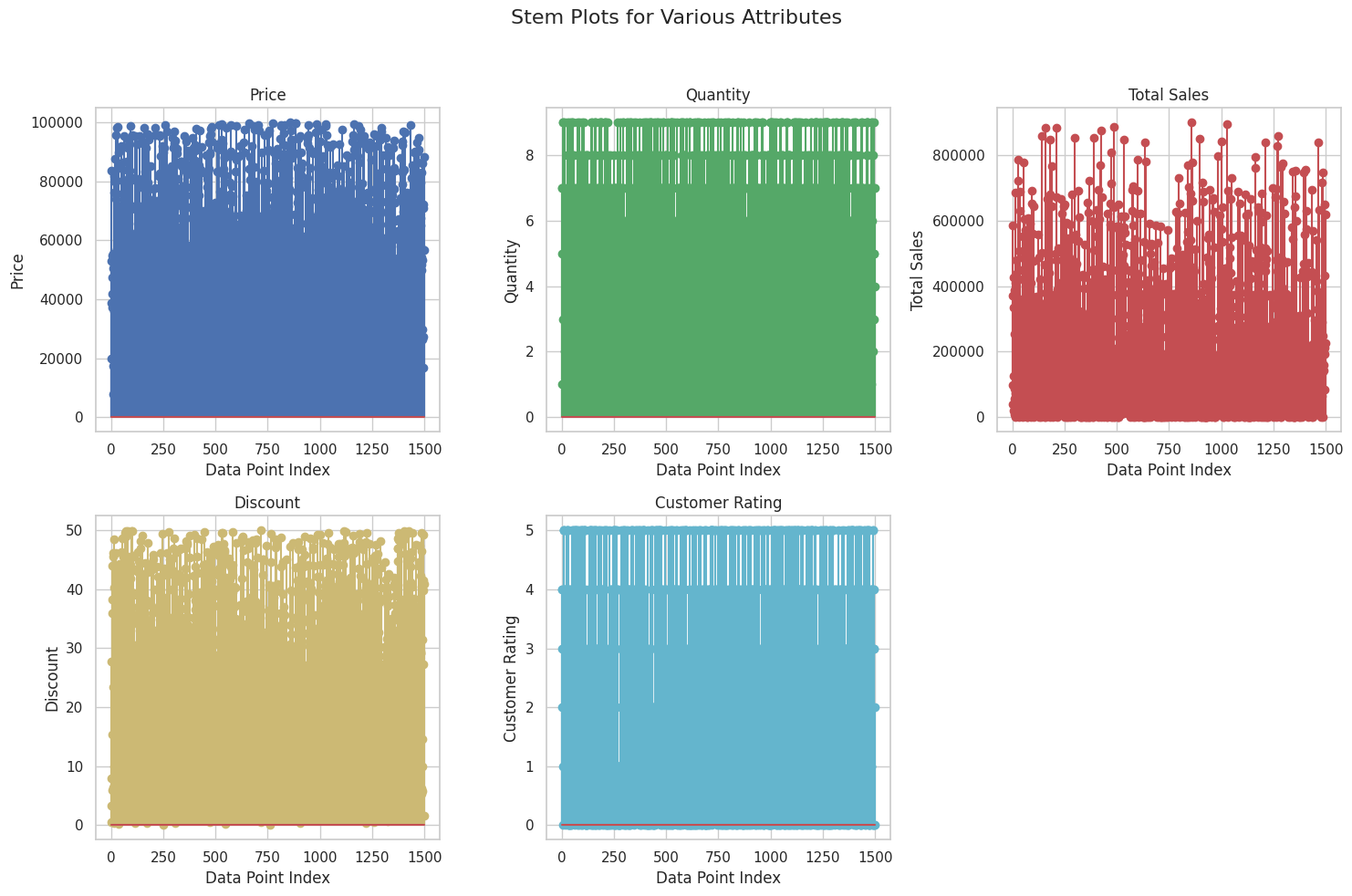
* With our data primed, we create a visually stunning stacked plot. Each layer represents a different attribute, and the sum of these layers forms a unique pattern for each order index.

Insights from the Stacked Plot:

* This stacked plot encapsulates the evolution of various attributes across orders. As you traverse the x-axis (Order Index), observe the interplay of colors to discern patterns and trends. The stacking allows for a quick comparison of attribute distributions, offering valuable insights into the dynamics of your sales data.

In conclusion, the stacked plot serves as a visual symphony, harmonizing diverse attributes into a cohesive narrative. Use this visual feast to identify trends, outliers, and correlations that drive informed decision-making in your sales strategy.

**Stem Plot:**

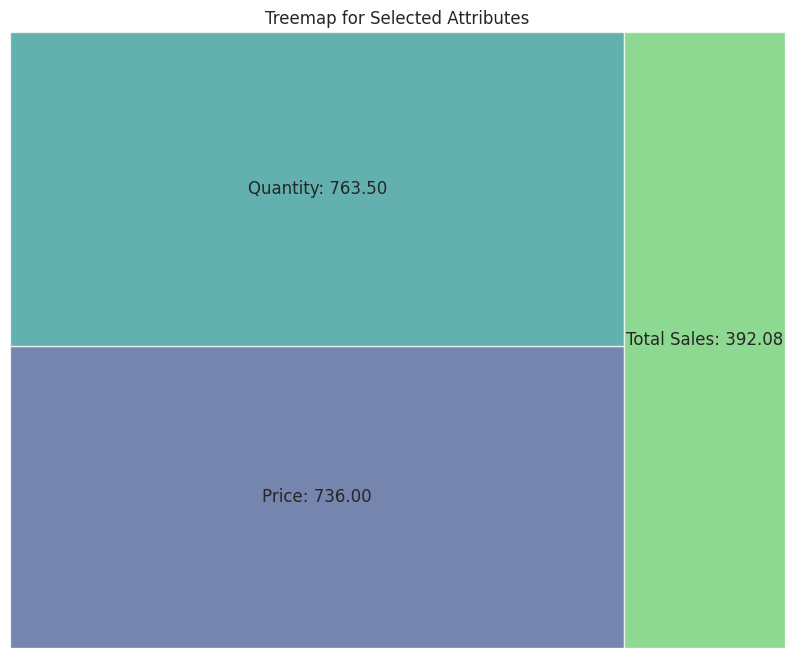


The stem plots for various attributes—'Price,' 'Quantity,' 'Total Sales,' 'Discount,' and 'Customer Rating'—provide a unique and detailed perspective on the distribution of these values across the synthetic sales dataset.

In the stem plots, each data point is represented by a vertical line, allowing for a visual assessment of the spread and density of the attribute values.

* The 'Price' stem plot indicates the pricing distribution, showcasing individual price points and their occurrences.
* The 'Quantity' stem plot visually represents the distribution of quantities in transactions, providing insights into the frequency of different quantity values.
* The 'Total Sales' stem plot illustrates the spread of total sales values, offering a comprehensive view of sales performance across orders.
* The 'Discount' stem plot reveals the distribution of discount values, highlighting the prevalence of specific discount levels.
* The 'Customer Rating' stem plot displays the distribution of customer ratings, helping identify patterns and concentrations of ratings.

**Tree plot:**



Setting the Stage:

* Before diving into the treemap magic, let's introduce the selected attributes – 'Price,' 'Quantity,' and 'Total Sales.' The data is normalized for a more effective visual representation, ensuring each attribute's contribution is accurately portrayed.

Crafting the Treemap:

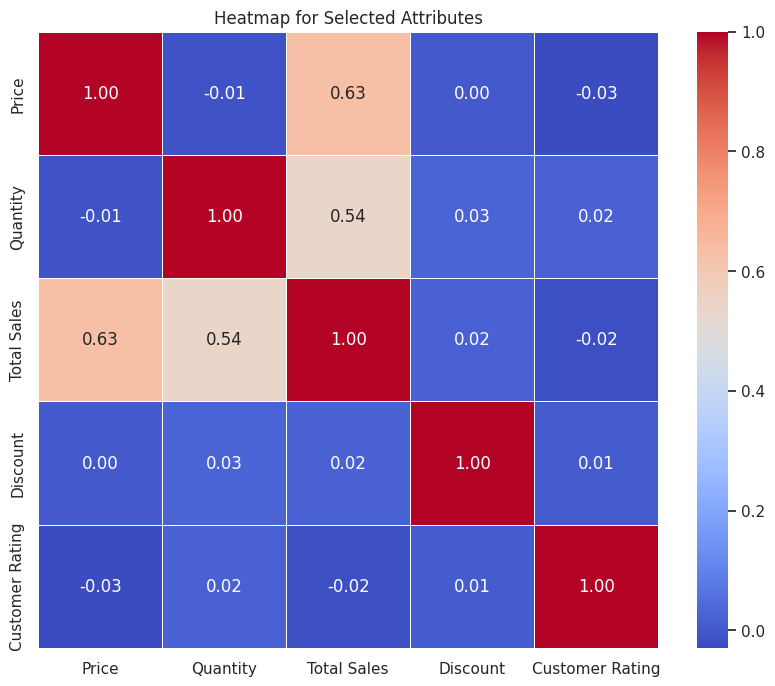
* The treemap itself is a captivating display of data proportions. Each square represents an attribute, and the size of the square corresponds to its contribution. The color palette adds a visual flair, making it easy to distinguish between attributes.

Decoding the Treemap:

* As we immerse ourselves in this visual feast, the treemap unveils the relative weight of each attribute. The juxtaposition of colors and sizes allows for a quick assessment of their significance in shaping the overall sales landscape. Dive into the treemap to decode the intricate dance of 'Price,' 'Quantity,' and 'Total Sales' within your dataset.

Treemaps offer a dynamic and visually engaging approach to understanding attribute contributions. The interplay of colors and sizes provides a nuanced perspective that transcends traditional numerical analysis. Armed with this treemap insight, businesses can prioritize and tailor strategies based on the relative impact of each attribute. The journey through data visualization continues, and treemaps have undoubtedly left an indelible mark on our quest for insights.

**Heat map:**



Selecting Attributes for Analysis:

* Before we dive into the heatmap, let's set the stage by selecting the attributes we want to scrutinize. 'Price,' 'Quantity,' 'Total Sales,' 'Discount,' and 'Customer Rating' take center stage in our analytical spotlight.

Unveiling Patterns: The Correlation Matrix:

* The correlation matrix becomes our compass, guiding us through the relationships between selected attributes. This matrix quantifies the degree and direction of correlation, laying the groundwork for a comprehensive understanding of the sales landscape.

Crafting the Heatmap:

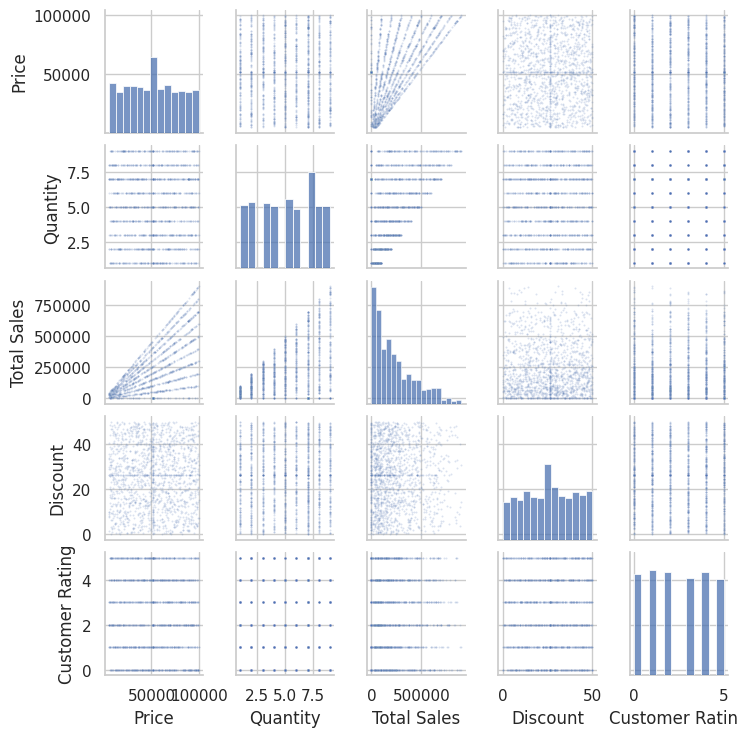
* The heatmap transforms the correlation matrix into a visual masterpiece. Warm hues signify positive correlations, while cool tones highlight negative correlations. Annotations provide numerical insights, offering a holistic view of how each attribute interacts with others.

Decoding the Heatmap:

* As we immerse ourselves in the colorful mosaic of the heatmap, patterns and connections come to life. A positive correlation between 'Quantity' and 'Total Sales,' or perhaps a negative correlation between 'Discount' and 'Customer Rating,' may emerge. Decoding these relationships is key to making informed decisions that can drive business strategies.

The heatmap is a powerful ally in unraveling the intricate relationships within sales data. Armed with this visual insight, businesses can identify areas of focus, potential optimizations, and hidden opportunities. The journey through data analysis continues, and the heatmap stands as a testament to the beauty of uncovering patterns in the seemingly complex world of sales.

**Pair plot:**



Price and Quantity:

* The scatter plots reveal a nuanced interplay between 'Price' and 'Quantity.' While a concentration of points near lower prices and moderate quantities is evident, some outliers showcase high-priced items with varying quantities. This dynamic hints at potential market segments demanding diverse product ranges.

Total Sales and Discount:

* Moving to 'Total Sales' and 'Discount,' a scatter of points unfolds, showcasing the impact of discounts on overall sales. The concentration of points around lower discounts and moderate sales suggests a balanced approach, while outliers with higher discounts hint at promotional strategies and their potential influence on sales volume.

Total Sales and Customer Rating:

* Exploring the relationship between 'Total Sales' and 'Customer Rating,' the scatter plots showcase a scattered yet positive trend. While a cluster of points with high ratings and varying sales exists, the general direction suggests a positive correlation between customer satisfaction and overall sales.

Discount and Customer Rating:

* The intersection of 'Discount' and 'Customer Rating' provides a canvas reflecting customer sentiment. The spread of points across different discount levels and ratings implies that discounts alone may not be the sole driver of high or low customer ratings. Other factors might contribute to the overall customer experience.

Price and Customer Rating:

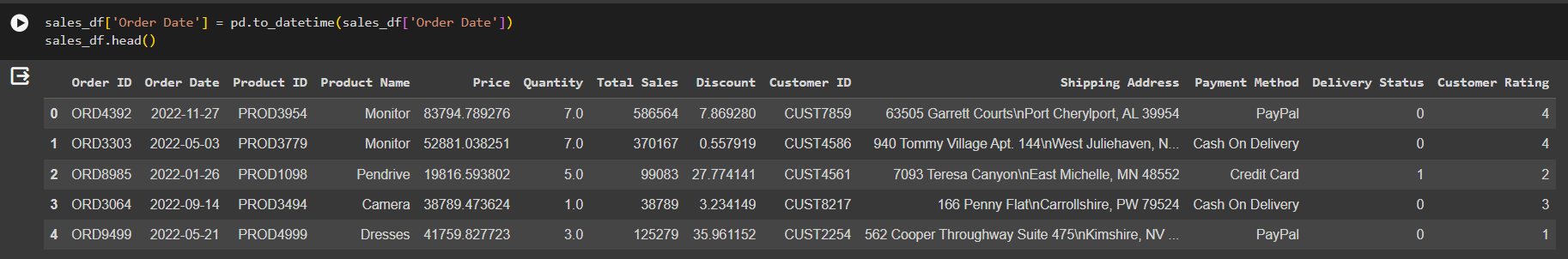
* Lastly, the pair plot delves into 'Price' and 'Customer Rating.' The majority of points cluster around moderate prices and high ratings, showcasing a positive correlation. However, outliers with high prices and varying ratings emphasize the need for a nuanced understanding of customer preferences at different price points.

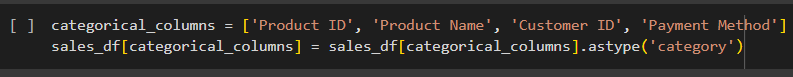
In this visual symphony, each scatter plot plays a unique note, contributing to the overarching melody of our sales data. As we decipher these visual cues, we unearth valuable insights that will guide our exploration into the intricacies of customer behavior, pricing strategies, and the dynamics of our synthetic sales landscape.

**Data Transformation:**

why we use data transformation for big data?

* data transformation is essential for preparing big datasets for analysis and modeling. It involves normalizing scales, handling missing values, encoding categorical variables, detecting and handling outliers, creating relevant features through feature engineering, and optimizing data for efficient processing in distributed computing environments. Proper data transformation improves model performance and ensures accurate and meaningful insights from large datasets.

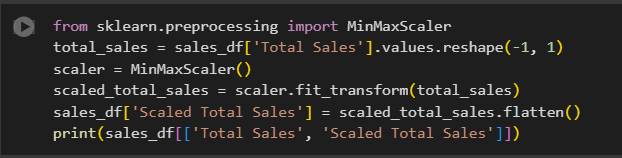


A Pandas DataFrame named sales\_df is created, and the 'Order Date' column is converted to a datetime format using pd.to\_datetime(). This operation is crucial for handling time-related data. The head() method then displays the initial rows of the DataFrame, showcasing the modified 'Order Date' column. This preprocessing step is essential for time-series analyses and ensures proper handling of date values in subsequent data exploration and modeling tasks.

The column names 'Product ID', 'Product Name', 'Customer ID', and 'Payment Method'. The Pandas DataFrame sales\_df is then updated using the astype('category') method for the specified columns. This operation converts the selected columns to categorical data types, which can be beneficial for memory efficiency and can also enhance performance in certain analyses.

**Normalization:**

* Normalization refers to the process of transforming or scaling data to a common scale, often to bring values within a specific range or to adhere to a standard distribution. The goal of normalization is to eliminate the influence of varying scales among different features in a dataset, ensuring fair comparisons and preventing certain features from dominating others during analysis. One common normalization technique involves transforming numerical values to a standard normal distribution with a mean of 0 and a standard deviation of 1, a process known as z-score normalization. This method is particularly useful when dealing with machine learning algorithms that are sensitive to the scale of input features, such as gradient-based optimization algorithms. Another normalization approach involves scaling values to a predefined range, such as [0, 1], using min-max scaling. Normalizing data is crucial in scenarios where features have different units or scales, ensuring that each feature contributes proportionately to the overall analysis, promoting robust and accurate modeling or statistical inferences.



A DataFrame **sales\_df** is assumed to contain a column named 'Total Sales,' and the values from this column are extracted and reshaped into a one-dimensional NumPy array. Subsequently, the MinMaxScaler from scikit-learn is employed to scale the 'Total Sales' values between 0 and 1, preserving the relative proportions of the original data. The scaled values are then added to the DataFrame as a new column named 'Scaled Total Sales.' The resulting DataFrame, displayed using the print statement, presents both the original 'Total Sales' and the corresponding scaled values. This preprocessing step, using MinMax scaling, is beneficial when dealing with machine learning algorithms that are sensitive to the scale of input features, ensuring that all values lie within a standardized range for improved model performance.

**Conclusion:**

* The exploration of synthetic sales data through various visualizations has provided valuable insights into the distribution and relationships of key attributes. Histograms illuminated the pricing concentration around $20,000 while 'Quantity' displayed a more evenly spread distribution across different ranges. The positively skewed 'Total Sales' distribution hinted at a substantial number of transactions resulting in lower sales, with a prominent peak around $50,000.
* Analyzing the 'Discount' histogram revealed a prevalence of discounts, particularly in the 10-20% range. This information is crucial for understanding the impact of discounts on sales. The 'Customer Rating' histogram showcased a nearly uniform distribution of ratings between 3 and 5, indicating overall customer satisfaction.
* These visualizations serve as a foundation for identifying trends and making informed decisions. The positively skewed 'Total Sales' distribution and concentrated 'Price' distribution are noteworthy findings that warrant further investigation. As we delve deeper into the data, these visualizations will guide us towards more in-depth insights, helping optimize business strategies for improved performance.