The dataset (10_Supermarket Sale) is a record of sales transactions from a retail business with physical store locations in multiple cities. It captures details about each sale, including the products purchased, the customer type, and the transaction value. This type of data is often used to analyze sales performance, customer behavior, and product popularity.

Data Types: The dataset includes a variety of data types:

Categorical Data: branch, city, customer_type, product_name, and product_category.

Numerical Data:

- Discrete Data: quantity is discrete data because you can count the number of items sold.
- Continuous Data: total_price is continuous data because it can have any value.

Identifier: sale_id serves as a unique identifier for each transaction.

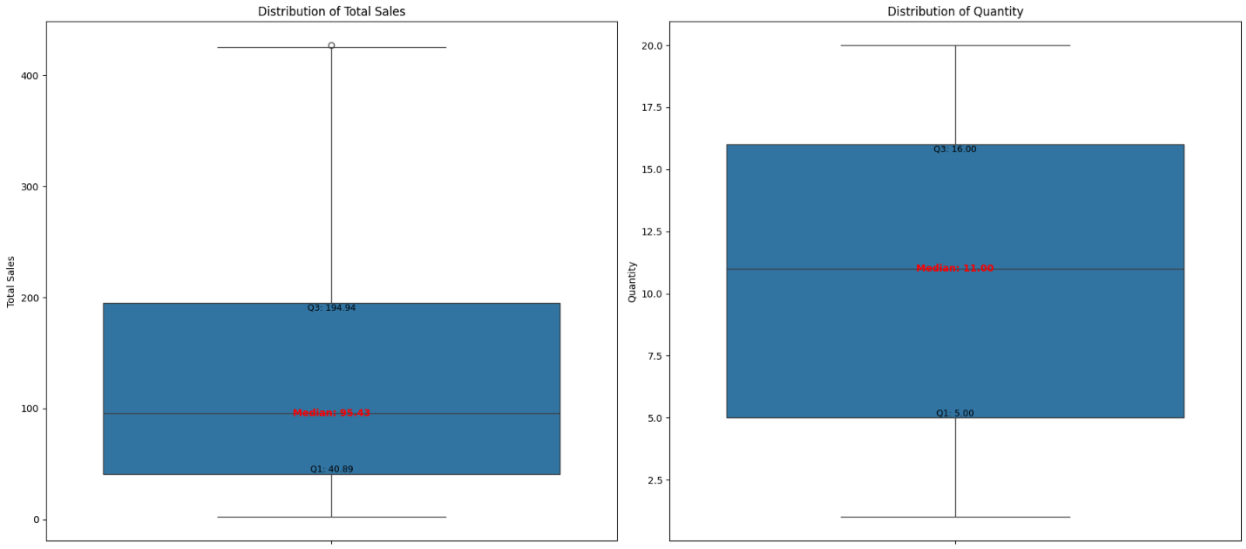
There are three duplicate rows in these dataset, so I will remove it by

To handling Missing customer_type, I will impute using the mode (most frequent value) of customer_type within the same branch and city. It's based on the logical assumption that customers in the same geographic location are likely to have similar characteristics.

To handling Missing product_category, I impute based on the most frequent product_category for the corresponding product_name. It is highly likely that a specific product like Shampoo will always fall under the same category Personal Care.

To etimating Missing quantity, I impute by estimating quantity using the median unit price for the same product_name in the same branch and city, then calculating quantity = total_price / median_unit_price. Because median unit price is less sensitive to extreme prices (outliers) than the mean.

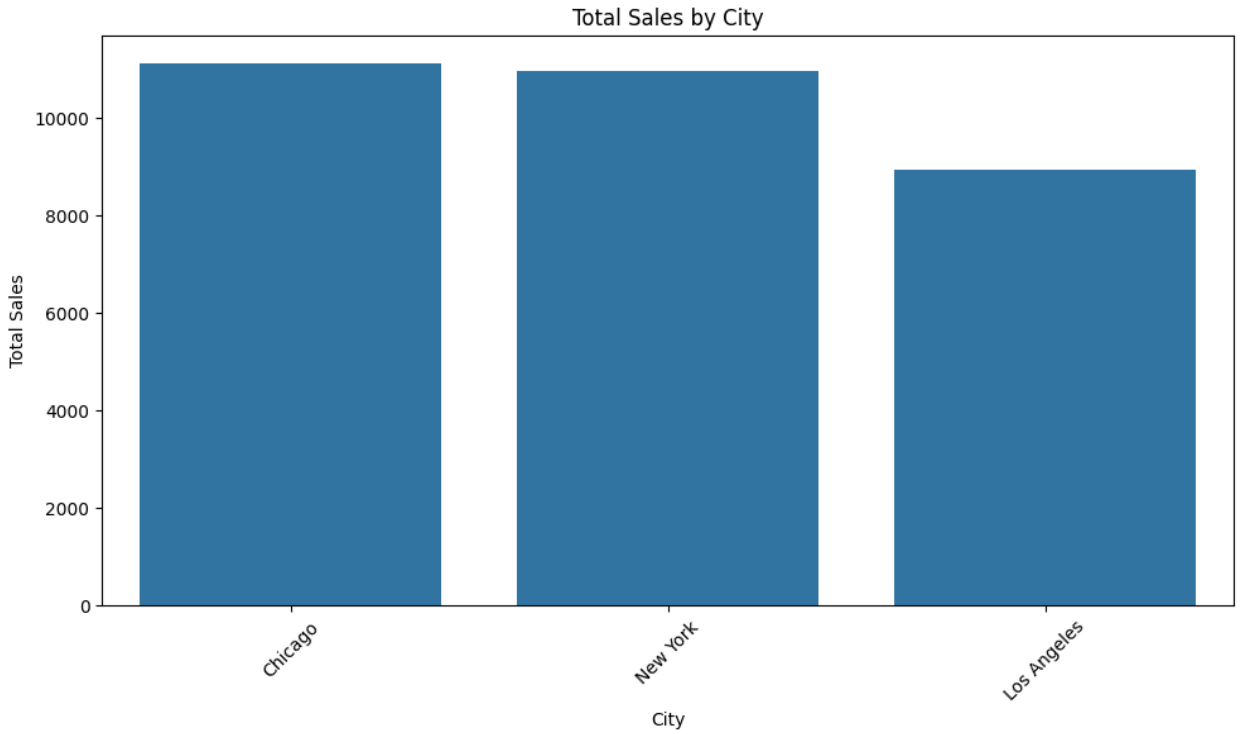
Key descriptive statistics for the quantity and total_price columns.



First, for boxplot of Total sale, Median equal 95.43, this is the middle value of total_sales . Half of all sales are less than 95.43, and half are more. Q1 (First Quartile): 40.89: 25% of the sales are below this value and Q3 (Third Quartile) equal 194.94 which means 75% of the sales are below this value. We have one Outlier above the top whisker represent transactions with unusually high values.

Second, Quantity boxplot has a Median (11.00) means in a typical transaction, 11 items are purchased. Q1 (First Quartile) equal 5.00 as 25% of transactions have 5 or fewer items and Q3 (Third Quartile) equal 16.00 as 75% of transactions have 16 or fewer items. There are no outliers in the quantity column.

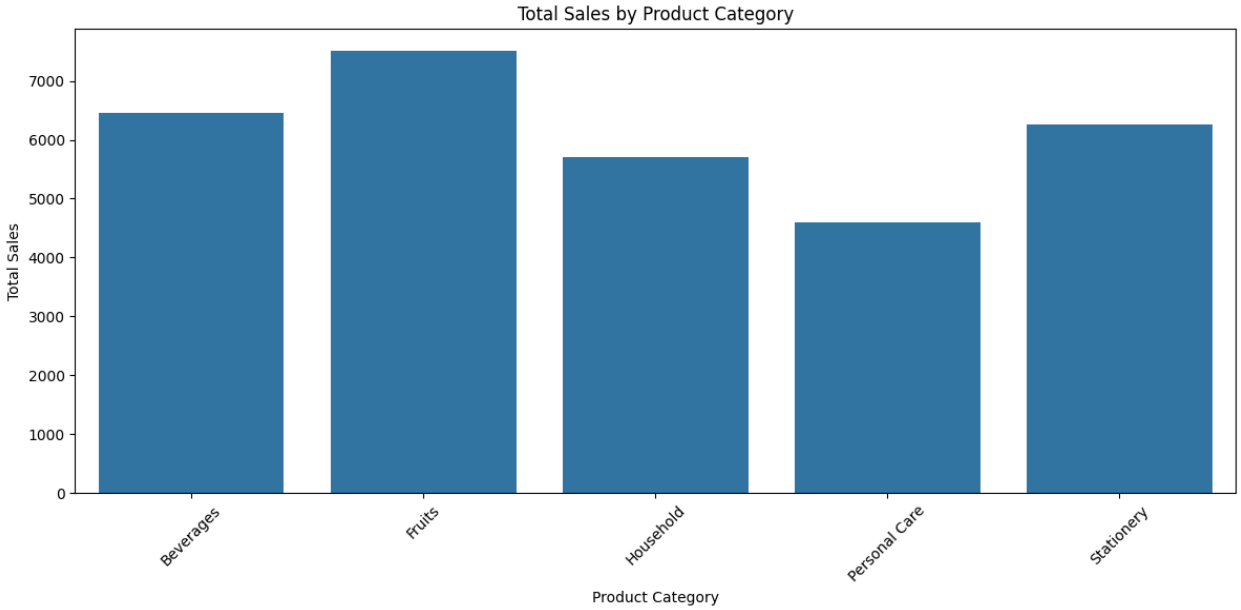
Bar chart to visualize the total sales for each city



The Bar chart (Total sales by City) provides a breakdown of sales performance across different cities.

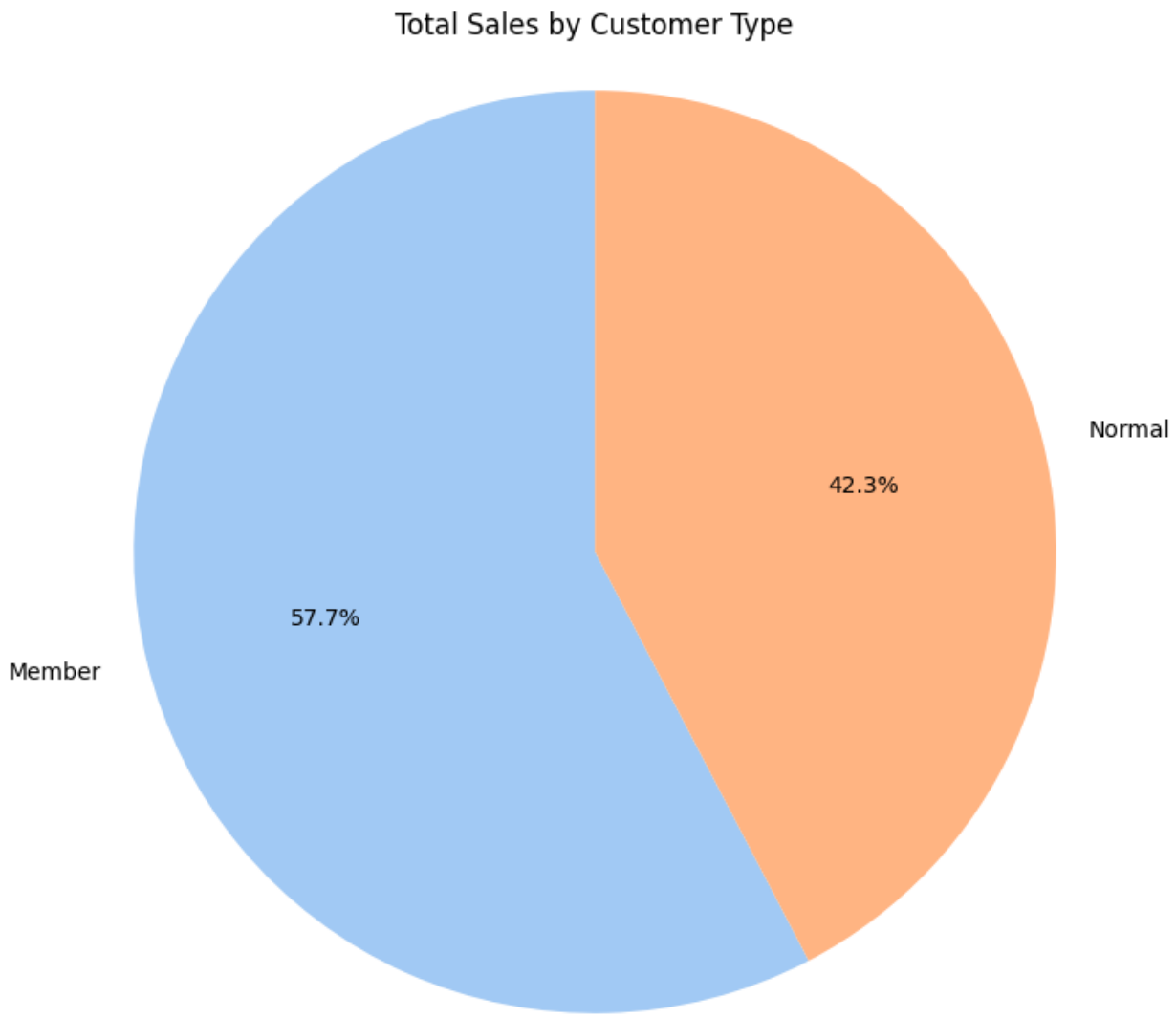
Chicago has the highest total sales (11133.05). New York has total sales (10964.42) which are slightly lower than Chicago's. Los Angeles has the lowest total sales (8948.81) among the three cities. These insights suggest that Chicago is the most profitable city in terms of total revenue.

Bar chart to visualize the total sales for each product category



Fruits have the highest total sales, exceeding 7000, indicating strong demand in this category. Beverages and Stationery follow with sales around 6000 each, showing a solid but slightly lower performance compared to Fruits. Household products have moderate sales, around 5500, suggesting a steady but not leading market presence. Personal Care has the lowest sales, approximately 4500, which may indicate weaker demand or less market penetration in this category. This shows Fruits are the top-performing category.

Pie chart to visualize the total sales for each customer type



The Sales Analysis by Customer Type provides information about the sales performance of different customer segments.

We can see that 'Member' customers contribute significantly more to total sales (17855.87) compared to 'Normal' customers (13106.73). So 'Member' customers are more valuable for the business in terms of both total spending and frequency of visits. Company should focus on Member' customers and provide membership programs.