Supermarket Sales Data Analysis Report

1. Introduction

This report dives into the analysis of a **Supermarket Sales dataset**, with a spotlight on **data cleaning** and **exploratory data analysis (EDA)**. The dataset includes both **numerical** attributes like sales, price, and quantity, and **categorical** attributes such as customer type, payment method, and branch.

2. Data Cleaning

Initially, the dataset had some issues, including missing values, duplicates, and outliers. Here's how we tackled those:

- **Missing Values:** We dropped these since they were minimal.
- **Duplicates:** We eliminated duplicates to maintain the integrity of the data.
- **Outliers:** We identified and removed outliers using the **Interquartile Range (IQR)** method to enhance the accuracy of our analysis.
- **Standardization of Categorical Variables:** We corrected formatting inconsistencies, like case sensitivity and typos.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- **Distribution of Total Sales:** Our histogram showed that most transactions were around **moderate total sales**, with some higher extremes.
- We noticed skewness in the sales distribution, suggesting that normalization might be needed for machine learning applications.
- **Summary Statistics:** The mean and median for total sales indicated a slight right skew.
- There was high variance in numerical features like quantity and unit price, pointing to a diverse range of transactions.

3.2 Bivariate Analysis

- **Correlation Matrix:** We found a strong positive correlation between **Total Sales** and both **Quantity** and **Tax**.
- **Gross Income** showed a major correlation with **Tax**, meaning tax plays a big role in income calculations.
- **Box Plot (Customer Type vs. Total Sales):** Customers labeled as "Member" had **higher total sales** than those categorized as "Normal".
- **Scatter Plots:** These illustrated a **linear relationship** between unit price and total sales.

3.3 Multivariate Analysis

- **Pair Plots:** These showed clear **positive trends** among tax, total sales, and gross income.
- **Heatmap:** This backed up our earlier findings from the correlation matrix, emphasizing strong connections among sales-related variables.

4. Key Insights

- **Total Sales is considerably affected by Quantity and Tax.**
- **Members typically spend more** than Normal customers.
- **Outliers in high-value transactions** may indicate bulk purchases.
- **Categorical variables influence sales behavior**, with different payment methods revealing distinct spending trends.

5. Conclusion

In this analysis, we cleaned up and explored the **Supermarket Sales dataset**, discovering key trends and relationships. This dataset can be further used for **predictive modeling** (like forecasting customer spending) or for knowledgeable **business decision-making** (such as optimizing product pricing).