Data Analysis Report

Introduction

The provided Python code performs data analysis on a supermarket sales dataset. The analysis covers various aspects such as data manipulation, summary statistics, handling outliers, sales analysis, pricing and profitability, customer analysis, time analysis, payment analysis, and includes a simple linear regression model for sales prediction.

Importing Libraries and Data

The code begins by importing essential libraries such as pandas, numpy, matplotlib, and seaborn. It then reads the dataset, 'supermarket\_sales.csv', using pandas and displays the initial data overview.

Data Manipulation with Pandas

The analysis starts with exploring the DataFrame using methods like head(), info(), and isna(). Missing values are identified and visualized using a bar plot. The code also demonstrates how to handle missing values by dropping them and provides insights into numerical and categorical data using various pandas functions.

Summary Statistics

The code calculates summary statistics, including mean, standard deviation, median, and quartiles for numerical columns. It explores the distribution of numerical data through histograms and provides information about unique values, columns, and index.

Handling Outliers

Outliers in the 'Total' column are identified using the interquartile range (IQR) method, and a boxplot is used for visualization. The code then removes the outliers from the dataset.

Summaries by Group

The code performs group-wise analysis, including average gross income by city, branch, gender, customer type, and payment method. It also explores the distribution of product lines and provides insights into the correlation between different numerical features.

Sales Analysis

Sales analysis involves examining the total sales revenue, branch-wise sales, city-wise sales distribution, and top-selling product lines in terms of quantity and revenue. The code also explores the variation in sales based on customer type and gender.

Pricing and Profitability

This section analyzes average unit prices, gross margin percentages across product lines, product lines with the highest gross income, and correlations between unit price and quantity sold.

Customer Analysis

The distribution of customer types is visualized, and patterns in the time or day of purchases are explored. The average rating for the supermarket and its correlation with other factors are also analysed.

Time Analysis

The code investigates seasonal trends or patterns in sales, sales variations on different days of the week or months of the year, and specific time slots where sales peak.

Payment Analysis

Payment preferences among customers, differences in payment methods between customer types or genders, and correlations between payment methods and total sales or ratings are explored.

Sales Prediction Model

A simple linear regression model is built to predict sales based on features such as unit price, quantity, and tax. The model is trained and evaluated, and scatter plots and regression plots are provided to visualize relationships between features and the target variable.

Conclusion

The analysis provides a comprehensive understanding of the supermarket sales dataset, covering various aspects of data exploration, summarization, and modeling. Insights gained from this analysis can be valuable for making data-driven business decisions and improving overall performance.