Data Analysis and Visualization

Comprehensive Report

1. Introduction

This report provides an in-depth overview of the processes undertaken to integrate, preprocess, analyze, and visualize data from multiple sources. The objective was to combine data from various CSV files, handle missing values, detect and remove outliers, normalize the data, and perform principal component analysis (PCA). The analysis also included Customer Lifetime Value (CLV) calculations and what-if scenarios to understand the impact of pricing changes.

2. Data Integration

Three datasets, 1.csv, 2.csv, and 3.csv, were combined into a single integrated dataset named tashi.csv. The integration process involved aligning columns and removing unnecessary columns that did not contribute to the analysis. After combining the datasets, the resultant file was used for further preprocessing and analysis.

3. Data Preprocessing

3.1 Handling Missing Values

Missing values in the dataset were addressed by filling them with the meaning of the respective columns. This approach ensured that the dataset remained complete and usable for subsequent analysis.

3.2 Removing Unnecessary Columns

Columns that were deemed irrelevant or redundant were removed from the dataset. This step was crucial to streamline the data and focus on the columns that contributed to the analysis.

3.3 Outlier Detection and Removal

Outliers were identified using the Z-score method. Data points with Z-scores greater than 2.5 were classified as outliers and removed from the dataset. This step aimed to enhance the quality of the data and ensure the analysis was not skewed by extreme values.

3.4 Normalization

Numerical columns were normalized using Min-Max scaling. This technique scaled the values to a range of [0, 1], facilitating comparisons and analyses by ensuring all numerical features were on a similar scale.

4. Data Transformation

4.1 Label Encoding

Categorical data was converted into numerical format using Label Encoding. This transformation was necessary for applying machine learning algorithms and statistical analyses that require numerical input.

4.2 Feature Engineering

Three additional features were created to enrich the dataset:

- **Total Sales**: Calculated as the product of UnitPrice and Quantity.
- **Discount Effectiveness**: Computed as the ratio of discount_amount to Total Sales.
- **Sales per Customer**: Aggregated total sales per customer.

5. Principal Component Analysis (PCA)

PCA was performed to reduce the dimensionality of the dataset while retaining 80% of the variance. The transformed data was saved and visualized using scatter plots, heatmaps, and box plots to understand the distribution and variance of the principal components.

6. Customer Lifetime Value (CLV) Calculation

CLV was computed using the formula: CLV=APV×PF. where APV is the average purchase value, PF is the purchase frequency, and the retention rate was assumed to be 0.8. The CLV was used to segment customers into different categories, providing insights into the value each customer brings over their lifetime.

7. What-If Analysis

A what-if analysis was conducted to examine the impact of varying unit prices on CLV. The analysis showed how different pricing strategies could influence average CLV and overall sales. This scenario analysis was visualized through line plots, bar plots, histograms, and heatmaps to illustrate the effects of price changes.

8. Conclusion

The comprehensive analysis of the integrated dataset provided valuable insights into customer behavior, data distribution, and the impact of pricing strategies. The visualizations and statistical analyses offer actionable information that can guide business decisions and strategic planning.