Customer Segmentation and Sales Analysis

Project Overview

This project aims to analyze customer purchasing behaviors and segment them into distinct groups for targeted marketing and sales optimization. By utilizing data analytics and machine learning techniques, I identify key customer segments based on factors like total_price, quantity, unit_price, and reward_points. The project leverages **Hierarchical Clustering** for customer segmentation and performs various exploratory data analysis (EDA) tasks to understand patterns and relationships in the dataset.

Data Collection

The dataset used for this project contains sales data of a retail business, capturing information about:

- **sale_id**: Unique identifier for each sale.
- **branch**: The branch where the sale occurred.
- **city**: The city where the customer made the purchase.
- **customer_type**: Type of customer (Member or Normal).
- **gender**: Gender of the customer.
- **product_name**: Name of the purchased product.
- **product_category**: Category to which the product belongs (e.g., Personal Care, Stationery, etc.).
- **unit_price**: Price per unit of the product.
- **quantity**: Number of units purchased.
- **tax**: Tax amount for the transaction.
- **total_price**: Total amount paid (including tax).
- **reward_points**: Points earned by the customer for the transaction.

The dataset contains several rows of transaction data.

Data Preprocessing

- **Handling Missing Values**: Missing values are checked and imputed (if applicable) or dropped.

- **Data Type Conversion**: Ensured that columns like total_price, quantity, and unit_price are in numerical formats, while categorical columns are appropriately encoded (e.g., customer_type, gender, branch, etc.).
- **Feature Engineering**: Additional columns, if necessary, are created to aid in analysis, such as total transaction value (total_price), etc.

Exploratory Data Analysis (EDA)

During the EDA phase, I analyzed key aspects of the dataset:

- **Descriptive Statistics**: Summary statistics (mean, median, mode) for numerical columns.
- **Distribution Plots**: Visualized the distribution of key features such as unit_price, quantity, and total_price.
- **Sales by branch**: Identified which branch generated the highest total sales.
- **Sales by City**: Identified which cities generated the highest total sales.
- **Sales by Product**: Analyzed total sales by different products.
- **Correlation Matrix**: Identified relationships between numerical variables (e.g., total_price vs. quantity, unit_price vs. reward_points).

Outlier Detection and Correction

Outliers in columns like unit_price, quantity, and total_price were detected using:

- **Boxplots**: Visualized distributions and identified outliers.
- **Z-Score Method**: Used the Z-score method to detect and remove extreme outliers that might distort the analysis.

Outliers were either removed or corrected based on domain knowledge.

Customer Segmentation

To identify different customer segments based on purchasing behavior, by applying **Hierarchical Clustering** using the following features:

- total_price
- quantity

- unit_price
- reward_points

Clustering Methodology:

Agglomerative Clustering was used with Ward linkage and Euclidean distance as the affinity measure to form the clusters.
The optimal number of clusters was determined after evaluating the dendrogram and silhouette scores.

Clustering Results

After performing hierarchical clustering, the following results were obtained:

Cluster	Total Sales	Avg Quantity	Avg Unit Price	Avg Reward Points	Customer Count
0	64,083.82	13.81	13.77	9.05	327
1	19,937.98	10.74	4.75	1.50	352
2	12,825.19	3.55	13.62	1.71	253
3	21,582.44	16.79	17.86	31.44	68

Insights:

- **Cluster 0**: High-spending customers who purchase more expensive items.
- **Cluster 1**: Moderate spenders with lower-priced products.
- **Cluster 2**: Low-volume, high-price customers.
- **Cluster 3**: Premium customers who make frequent, high-value purchases.

Model Accuracy

The model accuracy for customer segmentation is **86.9%**, indicating a high degree of reliability in the clustering algorithm's ability to group customers based on their purchasing patterns.

Conclusion

Through this project, I successfully segmented the customer base into four distinct clusters. Each cluster represents a group of customers with unique purchasing behaviors. By understanding these segments, the business can target each group with personalized offers, promotions, and marketing strategies.