Customers’ Preferences Analysis

Team names and Ids

|  |  |
| --- | --- |
| **Name** | **id** |
| **محمد أحمد سعد** | **23011453** |
| محمد ايمن السيد | **23011465** |
| **محمد حسام الدين محمد** | **23011471** |
| **محمد ياسر محمد سليمان** | **23011507** |
| **محمد فتحي عبد العزيز** | **23011485** |
| **محمد جهاد محمد أحمد** | **23011467** |

Tools Used: Python, Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, SciPy

# Objective

The goal of this project was to analyze customer-related retail data to identify purchasing behavior patterns using clustering and then to build a decision tree model to predict customer preferences. This analysis is crucial for helping retail businesses understand their target audience and optimize their marketing strategies.

# Data Loading and Initial Inspection

we started by loading the dataset using pandas.read\_excel() from a sheet named "Stores-Data". An initial inspection using .head() and .info() revealed that the dataset had no missing values, which was great

Some columns, especially categorical ones like "Store No.", "Location", "State", "Sundays", "Mng-Sex", and "HomeDel" were found to be redundant or already encoded. To simplify the analysis and focus on meaningful numeric features, we dropped these columns early on. A screenshot of a computer

AI-generated content may be incorrect.A black screen with white text

AI-generated content may be incorrect.

# Data Preprocessing

Before applying any MachineLearning techniques, we standardized the dataset using StandardScaler to bring all variables to a scale. Label encoding wasn’t necessary post column drops

A screen shot of a computer program

AI-generated content may be incorrect.

# Clustering Analysis

For unsupervised learning, we performed Agglomerative Clustering: A screenshot of a computer program

AI-generated content may be incorrect.

- we visualized the dendrogram using scipy.cluster.hierarchy.dendrogram to determine the optimal number of clusters.

- we chose a cluster count based on dendrogram spacing and validated this with the Silhouette Score.

- The clustering results were added to the dataset and visualized to ensure interpretability.

The purpose of clustering was to group similar stores/customers and later use this segmentation in the supervised model. A graph of a city

AI-generated content may be incorrect. **Why Agglomerative Clustering with Ward linkage?**  
We used Ward’s method in Agglomerative Clustering because it minimizes the variance within each cluster, leading to more compact and meaningful groupings. This is particularly suitable for customer segmentation, as it ensures that stores/customers within the same cluster are as similar as possible based on their standardized features. Ward’s linkage also integrates well with the dendrogram visualization, helping us clearly identify the optimal number of clusters.

# Classification with Decision Tree

To predict customer segment (from clustering), I split the dataset using train\_test\_split() and applied a Decision Tree Classifier:

- Visualized the decision tree using plot\_tree()

- Evaluated model performance using confusion\_matrix, classification\_report, and accuracy metrics.

- The model demonstrated clear rules and interpretability, making it easy to explain to non-technical stakeholders.

**-Why Decision Tree?**  
We chose the Decision Tree Classifier because of its interpretability and simplicity. After segmenting the data using clustering, we needed a model that could predict cluster membership based on feature values. Decision trees are ideal for this task as they generate clear, rule-based outputs that are easy to visualize and explain to stakeholders. This transparency supports business decisions and helps highlight the key factors that influence customer behavior.

# Results Summary

- No missing data or null values—clean preprocessing.

- Hierarchical clustering identified distinct customer groups with a decent silhouette score.

- The Decision Tree model successfully classified customers with interpretable rules and good accuracy. A diagram of a graph

AI-generated content may be incorrect.A graph with multiple columns

AI-generated content may be incorrect.A graph with green and blue dots

AI-generated content may be incorrect.

# Conclusions and Recommendations

This analysis confirmed that:

- Retail customers can be effectively grouped based on standardized purchasing and store-related metrics.

- Decision tree models are well-suited for explaining which features (e.g., sales, footfall, etc.) are most important in determining customer behavior.

- Businesses can use these insights to tailor strategies for each cluster (e.g., promotions, store layouts, etc.)