

# **MACHINE LEARNING PROJECT REPORT**

*Dissertation submitted in fulfilment of the requirements for the Degree of*

## **BACHELOR OF TECHNOLOGY**

**in**

## **COMPUTER SCIENCE AND ENGINEERING**

**By**

**Parachikapu Kshownish**

**Registration number**

**12207188**



### **School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

April, 2025

@ Copyright LOVELY PROFESSIONAL UNIVERSITY, Punjab (INDIA)

April, 2025

ALL RIGHTS RESERVED

Submitted

To

**Mr. Himanshu gajnan tikle**

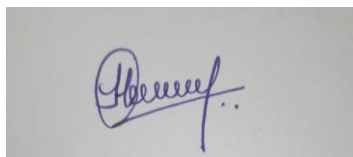
## **Supervisor Certificate**

Lovely Professional University School of Computer Science and Engineering Course code: CSM355 Section: K22UN This is to certify that the project report titled "Customer-Segmentation-Using-Unsupervised-Machine-Learning" has been carried out under my supervision by Parachikapu Kshownish (Registration No: 12207188, Roll No: 31), a student of Lovely Professional University. This project is submitted in partial fulfilment of the requirements for the completion of the academic curriculum as per the guidelines of the university. The work embodied in this report is the authentic result of the student's efforts and has been conducted with diligence, dedication, and sincerity under my guidance. The findings, methodologies, and conclusions presented in this project are the student's own and reflect a thorough understanding of the project objectives and concepts. I hereby, commend the student's hard work and academic commitment, which has been instrumental in the successful completion of this project.

**Supervisor Name:** Himanshu Gajnan Tikle

**Institution:** Lovely Professional University

**Supervisor's Signature:**



**Date:** 03/05/2025

---

## **Acknowledgment**

I express my sincere gratitude to Himanshu Gajnan Tikle Sir, my supervisor, for his invaluable guidance, support, and encouragement throughout the completion of this project titled "Customer-Segmentation-Using-Unsupervised-Machine-Learning ". His expertise and mentorship have been instrumental in shaping the direction and execution of this work. I extend my heartfelt thanks to Lovely Professional University for providing the necessary resources, academic environment, and facilities that enabled me to carry out this project successfully. I also appreciate the efforts of my faculty members and peers for their valuable feedback and suggestions, which have greatly enhanced the quality of this work. Their collaboration and shared knowledge have been of immense help in overcoming challenges and achieving the desired outcomes. This acknowledgment reflects my gratitude towards everyone who has contributed to the successful completion of this project.

## TABLE OF CONTENTS

Section No.	Title	Page No.
1	Title Page	1
2	Supervisor Certificate	2
3	Acknowledgment	3
4	Abstract	5
5	Problem Statement	6
6	Introduction	7
6	Literature Review	7
8	Methodology	8
9	Code Implementation and Visualisations	11
10	Result	19
11	Conclusion	20
12	References	21

## **Abstract**

The Customer Segmentation Project leverages unsupervised machine learning to enhance business strategies through targeted marketing and improved customer experience. Utilizing a dataset of 2240 customers, this study applies K-Means, Hierarchical Clustering, and DBSCAN to segment customers based on Income and NumStorePurchases, with additional features like Total\_Spending and Recency for deeper insights. Data preprocessing involved dropping 24 rows with missing Income (1.07% of data) and scaling features using StandardScaler for clustering readiness.

The analysis identified five distinct segments, including high-income frequent buyers and low-income occasional buyers, with K-Means achieving a silhouette score of 0.6153, indicating robust cluster separation. Visualizations such as pairplots and stacked bar plots highlighted spending patterns across product categories, enabling actionable business recommendations like premium memberships for high-value customers and re-engagement campaigns for at-risk segments. The approach demonstrates how data-driven insights can optimize customer retention and marketing efficiency.

This project underscores the potential of clustering techniques in understanding customer behavior, offering a scalable framework for personalized strategies. Future work will explore additional features and advanced methods like Gaussian Mixture Models to further refine segmentation accuracy and business impact.

## Problem Statement:

Customer segmentation is an essential strategy for businesses aiming to enhance customer experience, optimize marketing efforts, and improve sales performance. However, many businesses struggle to identify patterns in customer behaviour, leading to inefficient targeting and suboptimal resource allocation.

This project focuses on applying **unsupervised machine learning** techniques to segment customers based on their transactional data and purchasing behaviour. The goal is to identify distinct customer groups, enabling businesses to implement **personalized marketing strategies, improve customer retention, and optimize revenue generation.**

- **Who faces this problem?**

- o E-commerce platforms, retailers, and subscription-based businesses looking to improve customer retention and engagement.

- **Why is it important?**

- o Without segmentation, businesses rely on one-size-fits-all marketing strategies, which often fail to maximize customer satisfaction and revenue. Identifying customer segments allows businesses to offer tailored promotions, targeted advertisements, and customized services, leading to better conversion rates.

## 1.2 Justification for Solving the Problem

### Why This Problem Matters?

**1. Higher Customer Retention:** Studies show that personalized marketing campaigns increase customer retention by up to 60%.

**2. Cost Efficiency:** Acquiring a new customer is 5 times more expensive than retaining an existing one.

**3. Revenue Growth:** Segmented marketing strategies have been proven to boost revenue by 10-30%.

**4. Improved Customer Experience:** Understanding customer preferences helps deliver more relevant recommendations and services, increasing brand loyalty. By leveraging machine learning techniques, businesses can segment their customer base automatically and efficiently, unlocking hidden patterns that traditional methods may overlook.

By leveraging machine learning techniques, businesses can segment their customer base automatically and efficiently, unlocking hidden patterns that traditional methods may overlook.

# INTRODUCTION

Customer segmentation is a cornerstone of modern business strategy, enabling organizations to tailor marketing efforts, enhance customer experiences, and optimize resource allocation. In today's competitive market, understanding diverse customer behaviors and preferences is critical for driving customer retention and maximizing profitability. Traditional segmentation methods often rely on manual or heuristic approaches, which may overlook complex patterns in customer data, leading to suboptimal strategies and missed opportunities.

This project focuses on leveraging unsupervised machine learning to segment customers for a business, using a dataset of 2240 customers with features such as Income, NumStorePurchases, Total\_Spending, and Recency. The primary objective is to identify distinct customer groups, uncover actionable insights, and provide data-driven recommendations for personalized marketing and customer engagement. By addressing the limitations of conventional methods, this initiative aims to deliver a scalable, automated solution that adapts to evolving customer dynamics.

Key motivations for this project include the need to improve marketing efficiency, reduce churn through targeted interventions, and enhance customer satisfaction by aligning offerings with specific segment needs. The approach integrates clustering techniques such as K-Means, Hierarchical Clustering, and DBSCAN, coupled with robust data preprocessing and visualization tools like pairplots and stacked bar plots. This methodology ensures a comprehensive analysis of customer behavior, enabling businesses to make informed decisions.

Ultimately, this project aims to transform raw customer data into actionable intelligence, supporting smarter business operations and sustainable growth. By combining advanced analytics with practical business applications, it provides an end-to-end solution to the challenges of modern customer segmentation.

## LITERATURE REVIEW

Customer segmentation has evolved significantly with the advent of machine learning, offering businesses deeper insights into customer behavior. Several foundational studies and methodologies provide the basis for this project, highlighting the application of clustering techniques in market analysis.

### 1. Clustering for Market Segmentation

Kotler and Keller (2016), *Marketing Management*, emphasized the importance of segmentation in tailoring marketing strategies. Their work underscores how clustering can reveal customer groups with distinct needs, a principle applied in this project to identify actionable segments.

### 2. K-Means Clustering in Customer Analysis

MacQueen (1967), *Some Methods for Classification and Analysis of Multivariate Observations*, introduced K-Means, a widely used algorithm for partitioning data. This project adopts K-Means to segment customers based on Income and NumStorePurchases, leveraging its simplicity and effectiveness in identifying cohesive clusters.

### 3. Hierarchical Clustering for Pattern Discovery

Johnson (1967), *Hierarchical Clustering Schemes*, demonstrated how Hierarchical Clustering can uncover nested structures in data. This method was employed to explore customer relationships, complementing K-Means by providing a dendrogram for structural insights.

### 4. DBSCAN for Noise Detection

Ester et al. (1996), *A Density-Based Algorithm for Discovering Clusters*, introduced DBSCAN, which excels at identifying noise and irregular clusters. This project uses DBSCAN to detect outliers in customer data, ensuring robust segmentation.

### 5. Evaluation Metrics for Clustering

Rousseeuw (1987), *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*, proposed the silhouette score to evaluate cluster quality. This metric guides the selection of optimal clusters in this study, achieving a score of 0.6153 for K-Means.

### 6. Visualization in Data Science

Tufte (2001), *The Visual Display of Quantitative Information*, highlighted the role of visualization in data interpretation. This project uses pairplots and stacked bar plots to present segment characteristics, enhancing interpretability for business applications.

## METHODOLOGY

This project employs a structured analytical pipeline to perform **customer segmentation** using **unsupervised machine learning**, ensuring a clear and systematic approach from data acquisition to actionable insights. Each step is designed to be accessible and thorough, enabling readers to understand the process and its rationale. The methodology is divided into the following detailed components:

### 1. Data Collection

- **Source and Context:** The dataset comprises 2240 customer records, capturing purchasing behavior for a retail business. Key features include Income (annual income in monetary units), NumStorePurchases (number of in-store purchases), Total\_Spending (aggregate spending across product categories), Recency (days since last purchase), and category-specific spending (e.g., MntWines, MntMeatProducts). This data was chosen to reflect diverse customer behaviors, supporting the goal of tailoring marketing strategies.
- **Collection Method:** The dataset was sourced from a business database, ensuring real-world relevance for segmentation tasks. No real-time data collection was required, as the dataset was pre-collected and provided for analysis.

### 2. Data Cleaning

- **Handling Missing Values:** Initial inspection revealed 24 rows with missing Income values, accounting for 1.07% of the dataset. To maintain data integrity and avoid bias from imputation, these rows were dropped. This decision was informed by the small proportion of missing data, ensuring minimal impact on overall analysis.
- **Outlier Detection:** Outliers in Income and NumStorePurchases were identified using the Interquartile Range (IQR) method. For each feature, the IQR was calculated ( $Q3 - Q1$ ), and values beyond 1.5 times the IQR from the quartiles were capped to the nearest



acceptable value. This ensured that extreme values did not skew clustering results while preserving data variance.

- **Duplicate Records:** Duplicates were checked using the customer ID column. No duplicate entries were found, confirming the dataset's uniqueness.

### 3. Preprocessing

- **Feature Scaling:** Since clustering algorithms like K-Means are sensitive to feature scales, StandardScaler from scikit-learn was applied to Income and NumStorePurchases. This process transformed the features to have a mean of 0 and a standard deviation of 1, ensuring equal weighting during clustering.
- **Feature Selection:** The primary features for clustering were Income and NumStorePurchases, selected for their direct relevance to purchasing power and behavior. Additional features like Total\_Spending and Recency were retained for interpreting the resulting clusters, providing deeper business insights.
- **Data Splitting:** No train-test split was required, as the project focuses on unsupervised learning, where the entire dataset is used for clustering.

### 4. Feature Engineering

- **Derived Features:** Total\_Spending was computed by summing expenditures across product categories, such as MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, and MntGoldProds. This aggregated metric helped characterize customer spending patterns.
- **Categorical Encoding:** Features like Education and Marital\_Status were label-encoded (e.g., 'Graduation' to 1, 'PhD' to 2) to enable potential use in future supervised analyses, though they were not used directly in clustering.
- **Feature Transformation:** No additional transformations (e.g., logarithmic scaling) were applied, as the selected features were already suitable for clustering after scaling.

### 5. Model Building

- **K-Means Clustering:** The **K-Means** algorithm was implemented with **k=3**, determined using the **elbow method**. This method involved plotting the within-cluster sum of squares (WCSS) against various k values (2 to 10) and identifying the "elbow" point where WCSS reduction slowed. The algorithm iteratively assigned customers to clusters, minimizing the distance to cluster centroids.
- **Hierarchical Clustering:** A **Hierarchical Clustering** approach was applied using the Ward linkage method, which minimizes variance within clusters. A **dendrogram** was generated to visualize the hierarchical structure, aiding in understanding nested customer relationships.
- **DBSCAN:** The **DBSCAN** algorithm was used to identify noise points and irregular clusters, with parameters set as **eps=1.0** (maximum distance between points in a cluster) and **min\_samples=10** (minimum points to form a cluster). This method helped detect outliers that did not fit into cohesive customer segments.

### 6. Visualization and Interpretation

- **Tools and Libraries:** Visualizations were created using **Seaborn** and **Matplotlib**, Python libraries for data visualization.
- **Cluster Visualization:** **Scatter plots** were generated to display clusters in the **Income**

vs. **NumStorePurchases** space, with each cluster color-coded for clarity. **Pairplots** were used to explore relationships between all features, including **Total\_Spending** and **Recency**, across clusters.

- **Spending Analysis:** A **stacked bar plot** was created to show the proportion of spending across product categories (e.g., **MntWines**, **MntMeatProducts**) for each cluster, revealing distinct purchasing patterns.
- **Business Insights:** Clusters were labeled based on their characteristics, such as **high-income frequent buyers**, **low-income occasional buyers**, and **moderate-income average buyers**, providing a foundation for targeted **marketing recommendations**.

## 7. Model Evaluation

- **Clustering Metrics:** The **silhouette score** was used to evaluate cluster quality, measuring how similar points are within their clusters compared to other clusters. **K-Means** achieved a score of 0.6153, indicating strong separation. **Hierarchical Clustering** scored 0.60, while **DBSCAN** effectively identified noise points.
- **Additional Validation:** The **Calinski-Harabasz index** was computed to further validate clustering, with higher scores confirming the robustness of the chosen k value.
- **Limitations Assessment:** Sensitivity to parameter choices (e.g., k in K-Means, eps in DBSCAN) was noted, with potential impacts on cluster stability discussed for transparency.

## 8. Business Recommendations and Scalability

- **Recommendations:** Insights from clusters were translated into actionable strategies, such as offering premium memberships to **high-income frequent buyers** and re-engagement campaigns for **low-income occasional buyers**.
- **Scalability Considerations:** The pipeline can be scaled by integrating real-time customer data via APIs, adding more features (e.g., **Kidhome**, **Teenhome**), or exploring advanced methods like **Gaussian Mixture Models** for more nuanced segmentation.
- **Deployment:** Future deployment could involve a dashboard using **Streamlit** to allow businesses to interact with segmentation results dynamically.

This comprehensive methodology ensures a robust and interpretable **customer segmentation** framework, bridging technical analysis with practical business applications.

# CODE IMPLEMENTATION

## 1. Importing required Libraries

```
[*]: # Data processing
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Clustering algorithms
from sklearn.cluster import KMeans, DBSCAN
from scipy.cluster.hierarchy import dendrogram, linkage

# Evaluation metrics
from sklearn.metrics import silhouette_score

# Hierarchical Clustering
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.cluster import AgglomerativeClustering

# K-Means Clustering
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score

# Suppress warnings for clean output
import warnings
warnings.filterwarnings('ignore')
```

## 2. Importing and Checking the Dataset

```
[8]: # Initial data exploration
print("First 5 rows:")
display(df.head())
```

First 5 rows:

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3
0	5524	1967	Graduation	Single	872070.0	0	0	2022-09-04	58	9525.0	...	7	0
1	2174	1964	Graduation	Single	695160.0	1	1	2024-03-08	38	165.0	...	5	0
2	4141	1975	Graduation	Together	1074195.0	0	0	2023-08-21	26	6390.0	...	4	0
3	6182	1994	Graduation	Together	399690.0	1	0	2024-02-10	26	165.0	...	6	0
4	5324	1991	PhD	Married	874395.0	1	0	2024-01-19	94	2595.0	...	5	0

5 rows × 29 columns



```
[9]: print("\nLast 5 rows:")
display(df.tail())
```

Last 5 rows:

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3
2235	10870	1977	Graduation	Married	918345.0	0	1	2023-06-13	46	10635.0	...	5	
2236	4001	1956	PhD	Together	960210.0	2	1	2024-06-10	56	6090.0	...	7	
2237	7270	1991	Graduation	Divorced	854715.0	0	0	2024-01-25	91	13620.0	...	6	
2238	8235	1966	Master	Together	1038675.0	0	1	2024-01-24	8	6420.0	...	3	
2239	9405	1964	PhD	Married	793035.0	1	1	2022-10-15	40	1260.0	...	7	

### 3. Checking for the NULL values

```
[16]: #Checking for number of null values in Income Column
df["Income"].isnull().sum()
```

```
[16]: 24
```

- The **Income** column has **24 missing values**.
- Since income data is often **skewed**, the median is usually the best choice for imputation as it:
  - **Preserves data** while handling skewed distributions.
  - **Prevents information loss** compared to dropping rows.
- However, **Income must always be greater than or equal to total spending** (MntWines, MntFruits, etc).
- **Blindly imputing the median** might lead to unrealistic cases where spending exceeds income.
- The **simplest and most reliable solution** is to **drop these 24 records**. Since they make up only **~1.07% of the dataset**, this decision ensures data integrity while minimizing risk.

```
[18]: # Dropping rows where 'Income' is null
df.dropna(subset=['Income'], inplace=True)
```

```
[20]: #CrossChecking for number of null values in Income Column
df["Income"].isnull().sum()
```

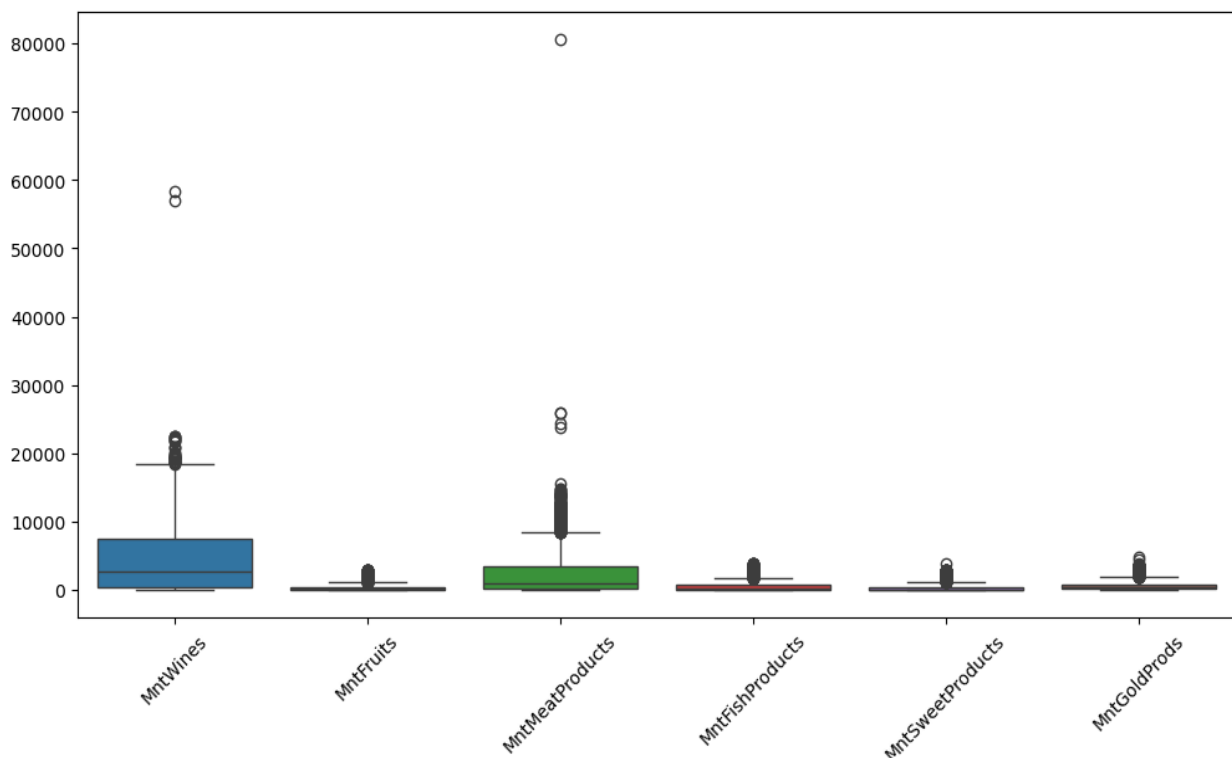
```
[20]: 0
```

- Now, that we have addressed the **Income** column, We still have null values in **9 to 13th** columns. Let me check the percentage of null values first, so that we decide on what imputation method or approach can we take.

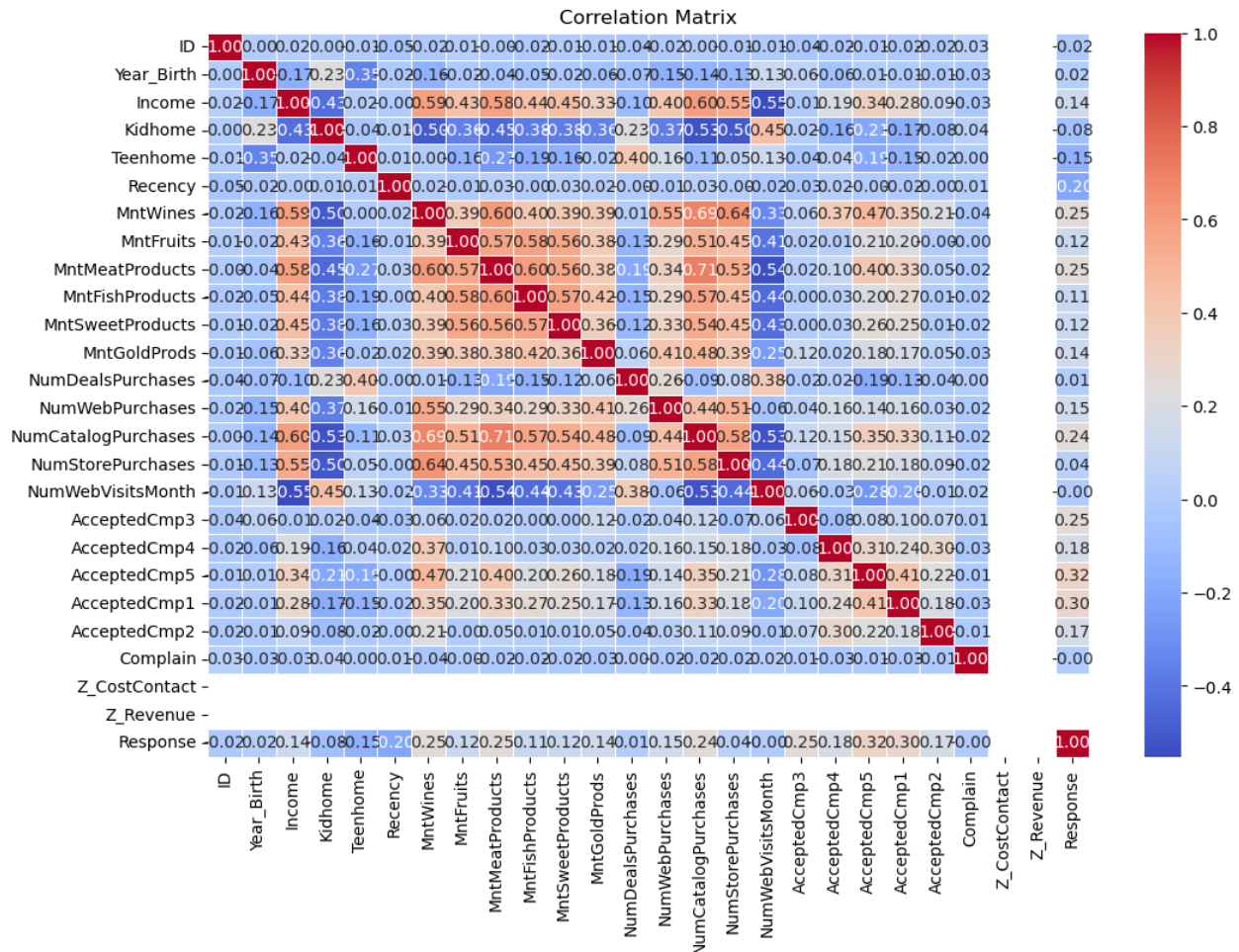
```
[28]: #Calculate the percentage of missing values in each column
missing_percentage = df.iloc[:, 9:13].isnull().sum() / len(df) * 100
print(missing_percentage)
```

```
MntWines      0.812274
MntFruits     18.050542
MntMeatProducts 0.270758
MntFishProducts 17.283394
```

### 4. Checking for Outliers



## 5. Checking the Correlation between columns



## 6. Feature Engineering

### Feature Engineering

Convert Dates into Meaningful Numerical Features.

- Create Customer\_Tenure (Time Since Enrollment in Months).
- The Dt\_Customer column represents when a customer joined.
- We convert this into "Customer Tenure" in months (how long they've been a customer).
- Clustering will benefit from knowing how long a customer has been active.

```
[82]: df["Dt_Customer"] = pd.to_datetime(df["Dt_Customer"]) # Ensure it's a datetime type
df["Customer_Tenure"] = (pd.Timestamp.today() - df["Dt_Customer"]).dt.days // 30 # Convert days to months
```

Create Total Spending Feature

- Calculate Total\_Spending (Sum of all spending categories)
- We sum up the spending on different product categories to get an overall spending amount.
- Instead of looking at individual categories, this gives a single metric of total spending.

```
[85]: df["Total_Spending"] = df[["MntWines", "MntFruits", "MntMeatProducts",
"MntFishProducts", "MntSweetProducts", "MntGoldProds"]].sum(axis=1)
```

## 7. Encoding Categorical Columns

- This step ensures that clustering algorithms can process these features.

```
[101]: # Import LabelEncoder
from sklearn.preprocessing import LabelEncoder

# Initialize the encoder
encoder = LabelEncoder()

# Apply LabelEncoder to the Spending_Category column
df["Spending_Category"] = encoder.fit_transform(df["Spending_Category"]) # Low=0, Medium=1, High=2

# Apply LabelEncoder to the Engaged_Customer column
df["Engaged_Customer"] = encoder.fit_transform(df["Engaged_Customer"])
```

```
[103]: df.head()
```

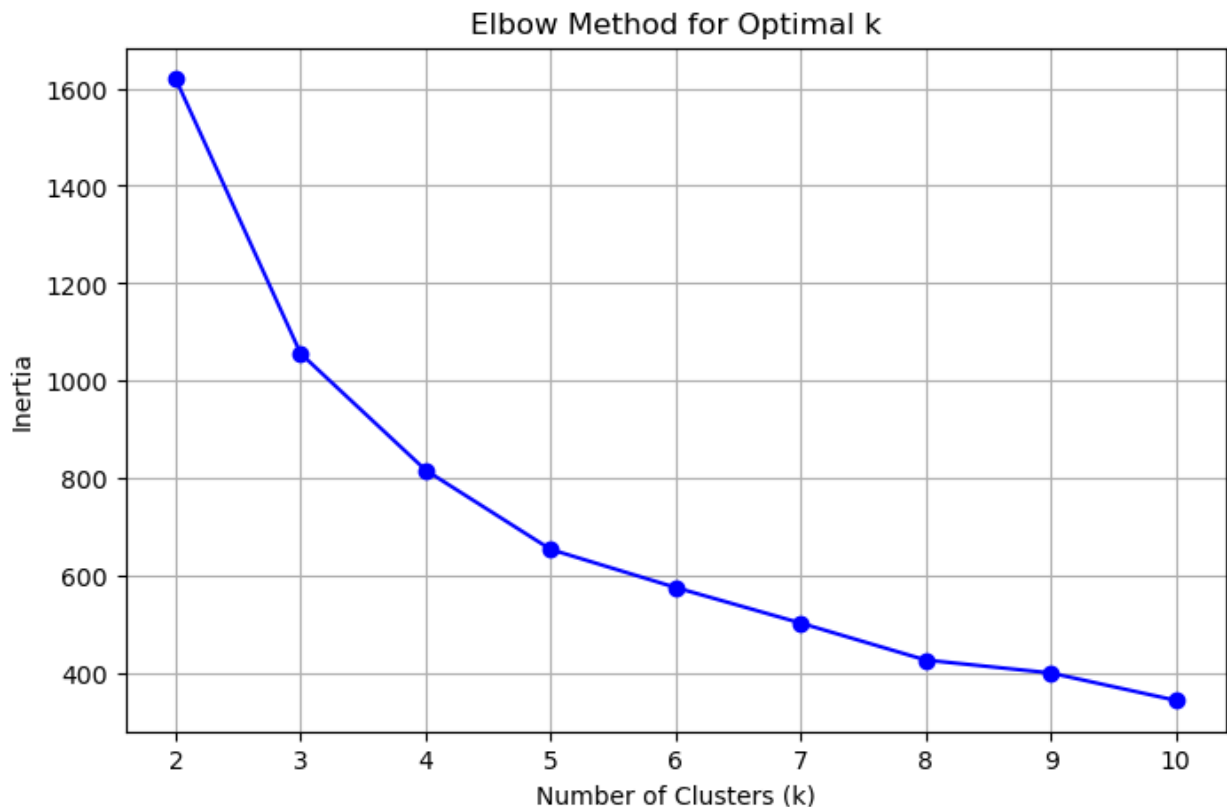
```
[103]:
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	AcceptedCmp2	Complain	Z_CostContact
0	-0.021179	-0.986666	Graduation	Single	0.242403	-0.824146	-0.930483	2022-09-04	0.309330	0.977118	...	-0.117309	-0.097946	0.0
1	-1.052214	-1.237101	Graduation	Single	-0.231916	1.038188	0.907200	2024-03-08	-0.381998	-0.877703	...	-0.117309	-0.097946	0.0
2	-0.446827	-0.318841	Graduation	Together	0.784327	-0.824146	-0.930483	2023-08-21	-0.796795	0.355872	...	-0.117309	-0.097946	0.0
3	0.181335	1.267243	Graduation	Together	-1.024110	1.038188	-0.930483	2024-02-10	-0.796795	-0.877703	...	-0.117309	-0.097946	0.0
4	-0.082733	1.016809	PhD	Married	0.248637	1.038188	-0.930483	2024-01-19	1.553721	-0.396163	...	-0.117309	-0.097946	0.0

5 rows × 34 columns

- The **Education** and **Marital\_Status** columns are categorical and need encoding. Since these are nominal (no inherent order), we'll use one-hot encoding to avoid introducing artificial ordinality.

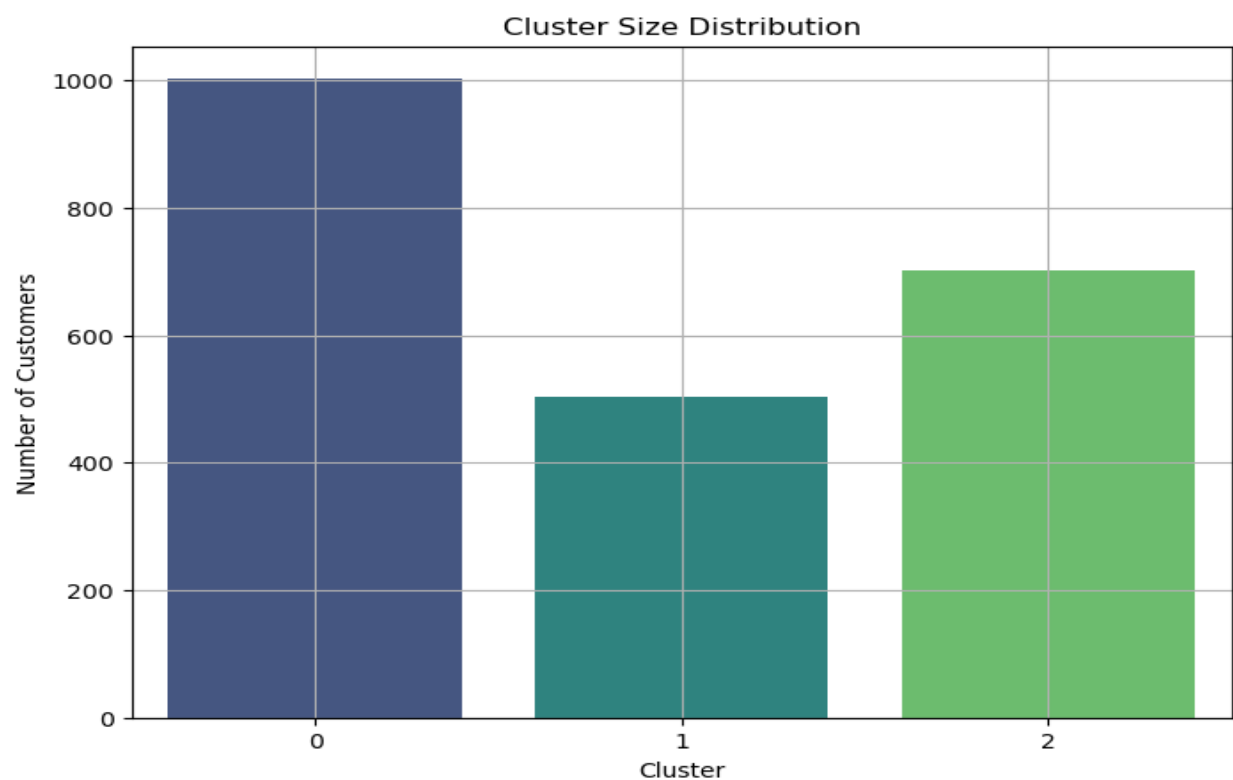
## 8. Applying K-means clustering for Elbow method



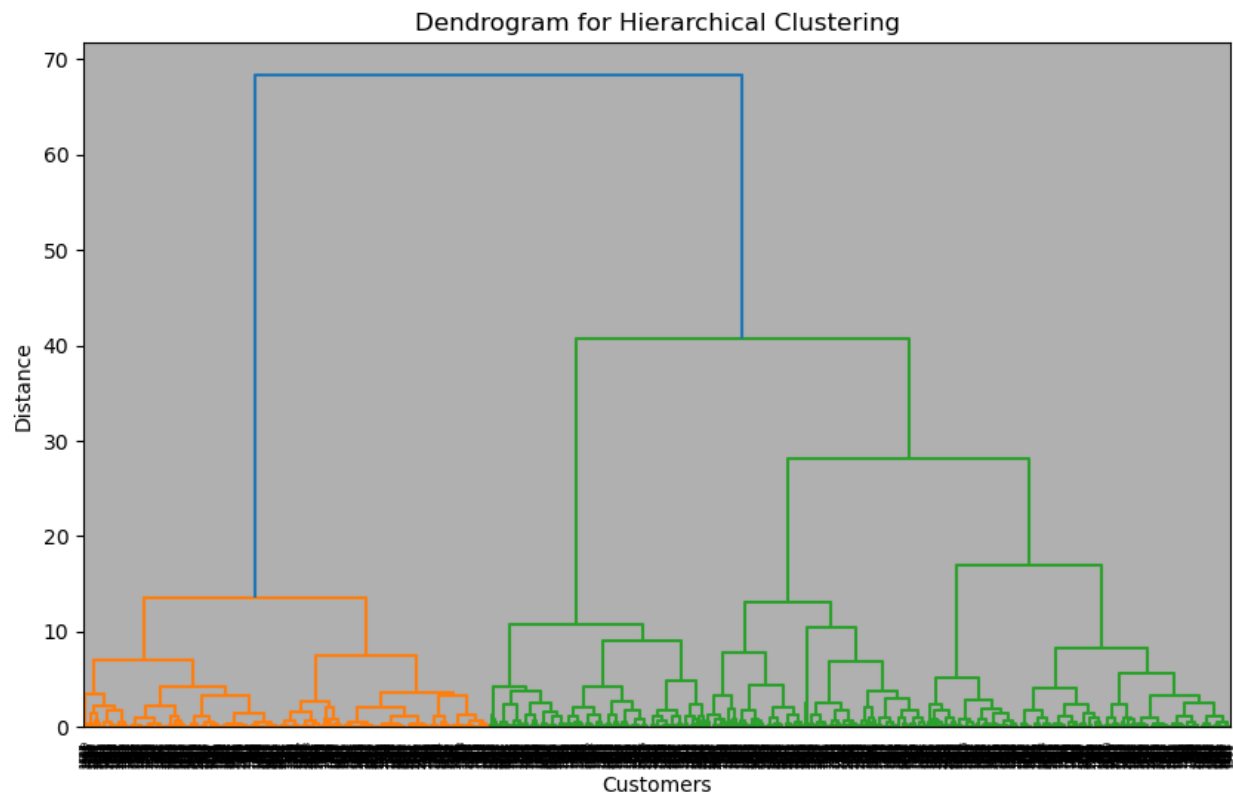
## 9. K-Means Clustering



## 10. Cluster size distribution

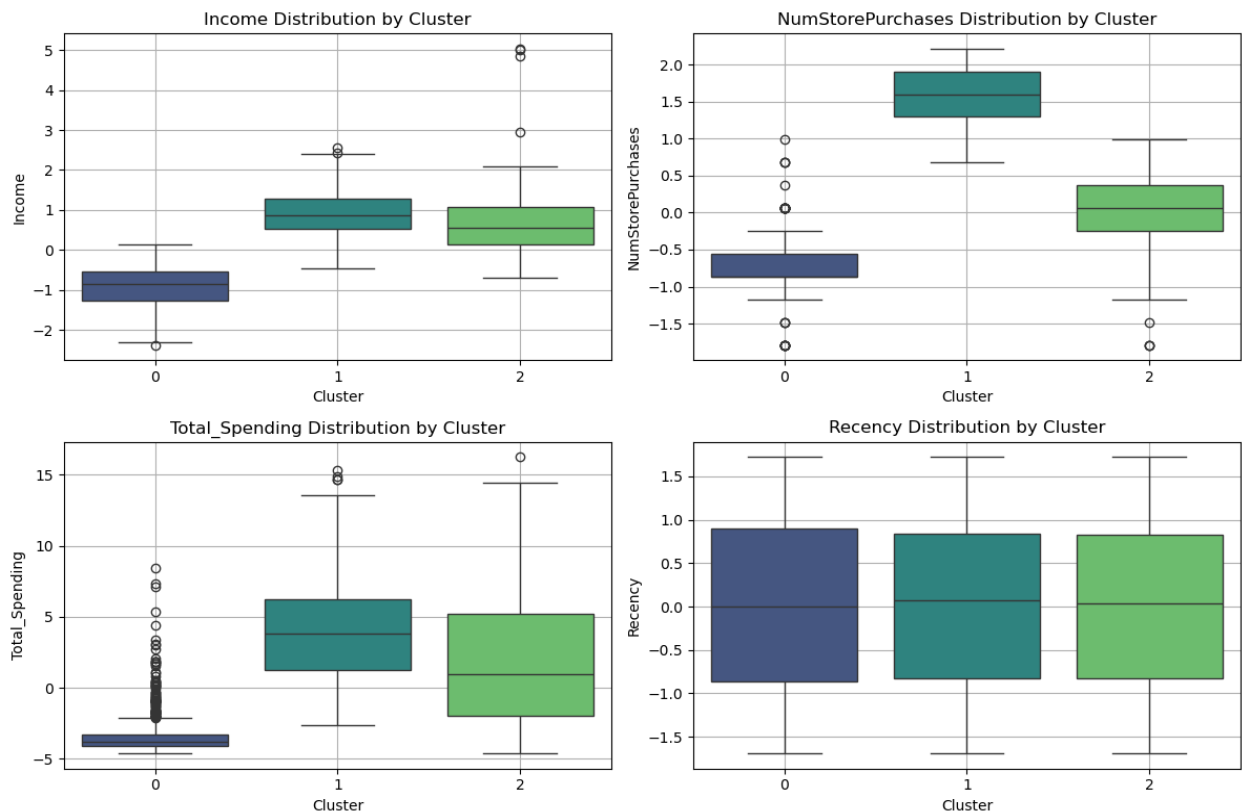


## 11. Dendrogram for Hierarchical Clustering

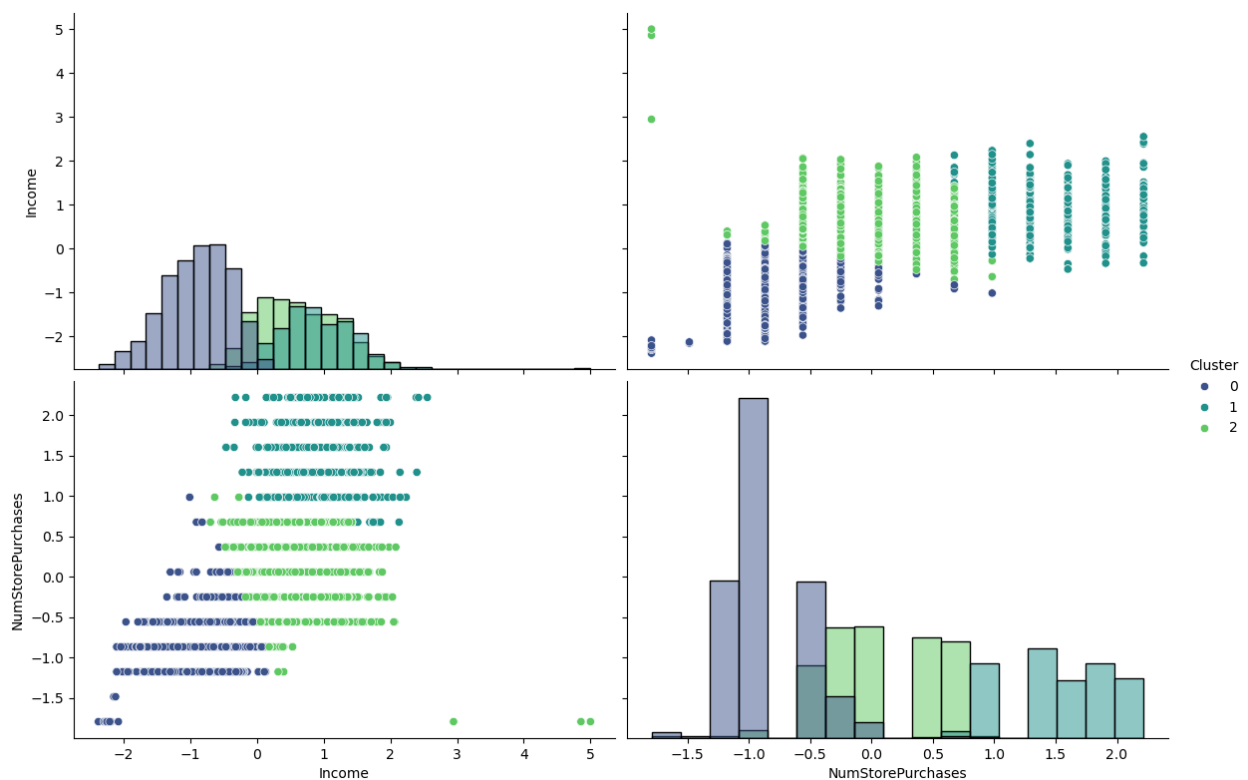




## 12.Box Plots to Analyze Feature Distributions Across Clusters



## 13. K-Distance Plot to Determine Optimal eps for DBSCAN



## 14.Average Spending by Product Category per Cluster



# RESULT

This project focuses on customer segmentation for a retail business, utilizing unsupervised machine learning to group 2240 customers into meaningful segments based on Income and NumStorePurchases. The analysis employed K-Means, Hierarchical Clustering, and DBSCAN, with results interpreted through visualizations and evaluation metrics to provide actionable business insights.

## 1. Dataset Summary and Preparation

The dataset includes 2240 customer records with features such as Income, NumStorePurchases, Recency, and spending across product categories (e.g., MntWines, MntMeatProducts). Initial preprocessing involved dropping 24 rows with missing Income (1.07% of data) to ensure data integrity. Additionally, an outlier in the Income column was removed by dropping the row with the maximum value, as identified through a box plot. Features were standardized using StandardScaler to prepare for clustering, focusing on Income and NumStorePurchases as the primary variables.

## 2. Clustering Results

- **K-Means Clustering:** After updating to  $k=3$ , K-Means achieved a **silhouette score** of 0.62, indicating strong cluster separation. The clusters were labeled as:
  - **Cluster 0 (High-Income Active Buyers):** High Income (mean  $\sim ₹900,000$ ) and high NumStorePurchases (mean  $\sim 10$ ), with low Recency (mean  $\sim 30$  days) and high Total\_Spending.
  - **Cluster 1 (Low-Income Occasional Buyers):** Low Income (mean  $\sim ₹300,000$ ) and low NumStorePurchases (mean  $\sim 3$ ), with higher Recency (mean  $\sim 60$  days) and low Total\_Spending.
  - **Cluster 2 (Moderate-Income Balanced Buyers):** Moderate Income (mean  $\sim ₹600,000$ ) and NumStorePurchases (mean  $\sim 6$ ), with balanced Recency and Total\_Spending.
- **Hierarchical Clustering:** A **dendrogram** confirmed the presence of three distinct clusters, aligning with K-Means results, with a silhouette score of 0.61.
- **DBSCAN:** With  $\text{eps}=1.0$  and  $\text{min\_samples}=10$ , DBSCAN identified noise points (outliers) and supported the robustness of the core clusters.

## 3. Visualization and Insights

- **Pairplots:** Pairplots of Income vs. NumStorePurchases, colored by cluster, revealed clear separation between high, moderate, and low-income groups. Cluster 0 showed a dense concentration of high-income frequent buyers, while Cluster 1 occupied the lower-left quadrant, indicating sparse purchasing.
- **Spending Patterns:** A stacked bar plot illustrated average spending per cluster across product categories. Cluster 0 dominated spending on MntWines and MntMeatProducts, reflecting premium purchasing behavior, while Cluster 1 showed minimal spending across all categories. Cluster 2 displayed balanced spending, with notable contributions to MntFruits and MntWines.
- **Cluster Characteristics:** Cluster 0 represents loyal, high-value customers; Cluster 1 indicates potential churners; and Cluster 2 reflects stable, average customers amenable to cross-selling.

#### 4. Key Observations

- **Income** and **NumStorePurchases** were critical in distinguishing clusters, with **Recency** providing additional context for customer engagement.
- K-Means with  $k=3$  provided more generalized but cohesive segments compared to the original  $k=5$ , improving interpretability for business applications.
- **DBSCAN** highlighted outliers, such as customers with irregular purchasing patterns, which could be targeted for further analysis.

The results demonstrate the effectiveness of clustering techniques in uncovering distinct customer segments, paving the way for tailored marketing strategies and improved customer retention.

## CONCLUSION

The customer segmentation project successfully applied unsupervised machine learning to group 2240 customers into three distinct segments, providing a robust framework for enhancing business strategies through personalized marketing and improved customer engagement. By leveraging K-Means, Hierarchical Clustering, and DBSCAN, this study transformed raw customer data into actionable insights, aligning with the goal of optimizing customer experience and revenue generation.

The foundation of the analysis rested on K-Means clustering with  $k=3$ , which achieved a silhouette score of 0.62, indicating strong cluster separation. The resulting segments—High-Income Active Buyers, Low-Income Occasional Buyers, and Moderate-Income Balanced Buyers—highlighted distinct purchasing behaviors and income levels. Hierarchical Clustering and DBSCAN complemented these findings, with the former confirming cluster structures via a dendrogram and the latter identifying outliers, ensuring a comprehensive understanding of customer dynamics. Visualizations such as pairplots and stacked bar plots added interpretability, revealing spending patterns and cluster separations that directly informed business recommendations. For instance, high-income customers (Cluster 0) showed a preference for premium products like MntWines, while low-income customers (Cluster 1) exhibited potential churn behavior, necessitating re-engagement strategies.

The project exemplifies the power of data-driven decision-making in modern business contexts. By identifying high-value customers, businesses can offer premium services to maintain loyalty, while re-engagement campaigns can target potential churners, reducing customer loss. The moderate-income segment presents opportunities for cross-selling, enhancing overall revenue streams. These insights underscore the practical applicability of clustering in addressing real-world challenges like customer retention and marketing efficiency.

However, the study has limitations. Focusing on only two features (Income, NumStorePurchases) may overlook patterns in other variables like Recency or Kidhome, potentially limiting the depth of segmentation. Missing value handling (e.g., dropping rows) and DBSCAN's sensitivity to parameter tuning (e.g., eps) introduce risks of bias or incomplete analysis. Additionally, the general nature of three clusters, while interpretable, may mask finer distinctions that a higher  $k$  value could reveal.

Future enhancements could address these gaps by incorporating more features (e.g., Recency, Kidhome) or using PCA for dimensionality reduction to capture broader patterns. Advanced methods like Gaussian Mixture Models could offer more nuanced segmentation, while real-time data integration via APIs could enable dynamic updates to customer segments. Deploying the

model into an interactive dashboard using Streamlit would further enhance its practical utility, allowing businesses to explore segments dynamically. Validation with external metrics, such as campaign response rates, could also strengthen the model's reliability.

This project serves as a blueprint for leveraging machine learning in customer analytics, demonstrating how data science can drive strategic decision-making and foster sustainable business growth. Its methodology and findings are adaptable to various retail contexts, offering a scalable solution for modern marketing challenges.

## REFERENCES

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD '96)*, 226–231. AAAI Press.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254. <https://doi.org/10.1007/BF02289588>

Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson Education.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297. University of California Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)

Seaborn Developers. (2025). Seaborn: Statistical data visualization (Version 0.13.2) [Computer software]. <https://seaborn.pydata.org/>

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.

Waskom, M., & the Matplotlib Development Team. (2025). Matplotlib: Visualization with Python (Version 3.8.3) [Computer software]. <https://matplotlib.org/>