### 

### CUSTOMER SEGMENTATION DATA SCIENCE REPORT

COURSE NAME: BACHELOR OF SCIENCE IN SOFTWARE ENGINEERING

YEAR:2

GROUP: B, EVENING

SUPERVISOR: DR.LIVINGSTONE

GIT REPOSITORY:https://github.com/SsenogaHerman/GROUP\_B\_EVENING\_RECESS.git

|  |  |  |
| --- | --- | --- |
| **Name** | **Registration Number** | **Student Number** |
| Ssenoga Herman | 23/U/17648/EVE | 2300717648 |
| Nyonyozi Maria Lisa Loyce | 23/U/16424/EVE | 2300716424 |
| Nabirye Anita | 23/U/12826/EVE | 2300712826 |
| Mugoya Yusuf | 23/U/11615/EVE | 2300711615 |
| Walusimbi Ashraf | 23/U/18466/EVE | 2300718466 |

Contents

[1](#_Toc204026481)

[CUSTOMER SEGMENTATION DATA SCIENCE REPORT 1](#_Toc204026482)

[1. Executive Summary 3](#_Toc204026483)

[2. Introduction 3](#_Toc204026484)

[3. Data Overview 3](#_Toc204026485)

[4. Data Preparation and Cleaning 4](#_Toc204026486)

[Task 2: Missing Data Handling Strategy 5](#_Toc204026487)

[Task 3: Implementation and Impact Evaluation 7](#_Toc204026488)

[5. Customer Segmentation Modeling 8](#_Toc204026489)

[6. DATA SPLITTING AND MODEL TRAINING 9](#_Toc204026490)

[7. Objectives And Recommendations 11](#_Toc204026491)

### 1. Executive Summary

This report presents the findings of a customer segmentation analysis performed on the provided dataset. The objective is to identify distinct customer groups, understand their characteristics and behaviors, and leverage these insights to inform targeted business strategies, particularly in marketing. Our analysis involved a thorough data cleaning process, feature engineering, and the application of clustering algorithms. The identified segments offer valuable opportunities for personalized customer engagement, optimized resource allocation, and improved overall business performance.

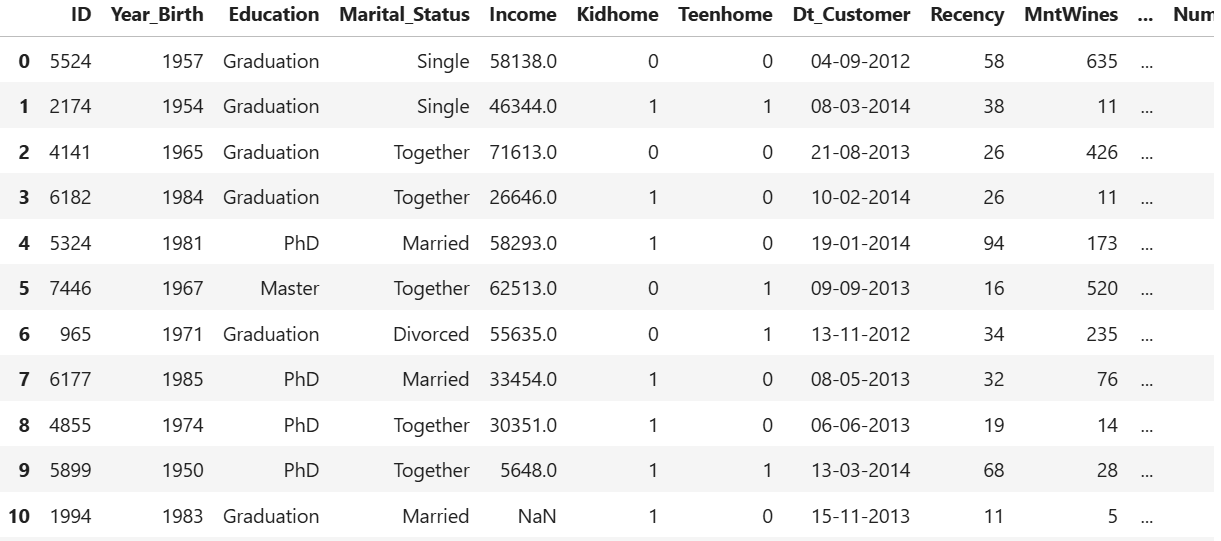
### 2. Introduction

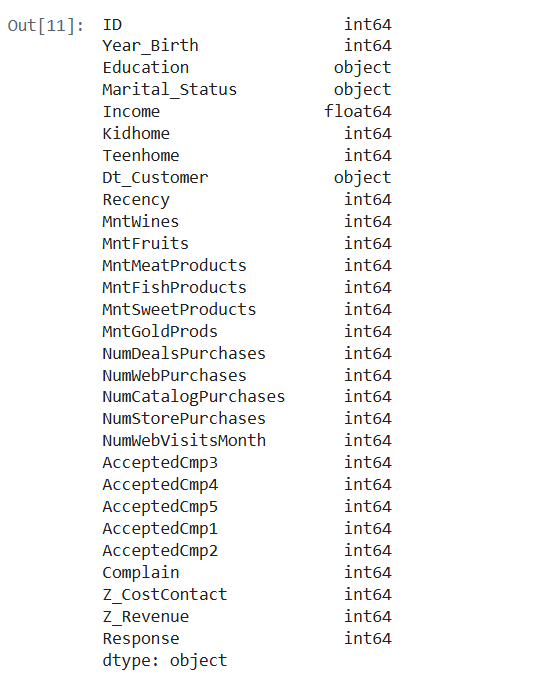
In today's competitive market, understanding customer behavior is paramount for business success. This project aims to segment our customer base to move beyond a one-size-fits-all approach and enable more precise and effective strategies. By grouping customers with similar attributes and purchasing patterns, we can tailor marketing efforts, product development, and customer service to meet specific needs, thereby enhancing customer satisfaction and maximizing profitability.

### 3. Data Overview

The analysis was conducted on the customer\_segmentation.csv dataset, comprising 2240 rows and 29 columns. The dataset includes various customer attributes such as:

* **Demographics**: ID, Year\_Birth, Education, Marital\_Status.
* **Household Information**: Income, Kidhome, Teenhome.
* **Customer Engagement**: Dt\_Customer (Date of customer's enrollment), Recency (Days since last purchase).
* **Spending Habits**: MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds (Amounts spent on various product categories).
* **Purchase Channels**: NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth.
* **Campaign Responses**: AcceptedCmp1 to AcceptedCmp5 (Response to marketing campaigns), Response (Response to last campaign).
* **Other**: Complain, Z\_CostContact, Z\_Revenue.



A screenshot of a computer

AI-generated content may be incorrect.

### 4. Data Preparation and Cleaning

Data quality is critical for accurate analysis. Our data preparation involved the following steps as per the project requirements:

#### Task 1: Missing Value Identification

Initial exploration revealed missing values primarily in the **Income** column. Specifically, 24 records had missing income data, representing approximately 1.07% of the total dataset.

A square with a missing value

AI-generated content may be incorrect.

### Task 2: Missing Data Handling Strategy

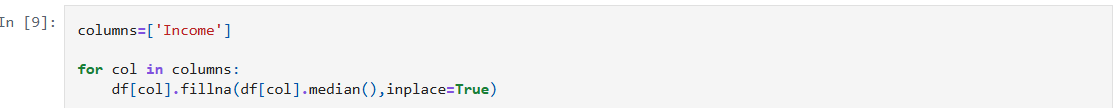
For the missing values in the Income column, we opted for **median imputation**.

**Comparison and Justification of Chosen Method:**

When addressing missing numerical data, common strategies involve using either the mean or the median for imputation.

* The **mean** (average) is susceptible to being disproportionately influenced by **outliers**, which can pull its value away from the typical central point of a skewed distribution.
* The **median** (the middle value when data is ordered) is **robust to outliers** and provides a more accurate representation of the central tendency in **skewed distributions**.

Our decision to use **median imputation** was directly informed by the initial exploratory data analysis and visual inspection of the Income distribution. The histograms clearly demonstrated that the Income data exhibited a **right-skewed distribution** with a long tail indicating the presence of high-income outliers. In such a scenario, using the median ensures that the imputed values are consistent with the typical income of the majority of the dataset, minimizing distortion from extreme values. This approach prevents the imputed data from artificially inflating or deflating the overall central tendency, thus preserving the true characteristics of the Income feature for more reliable customer segmentation.



A graph of income distribution

AI-generated content may be incorrect.

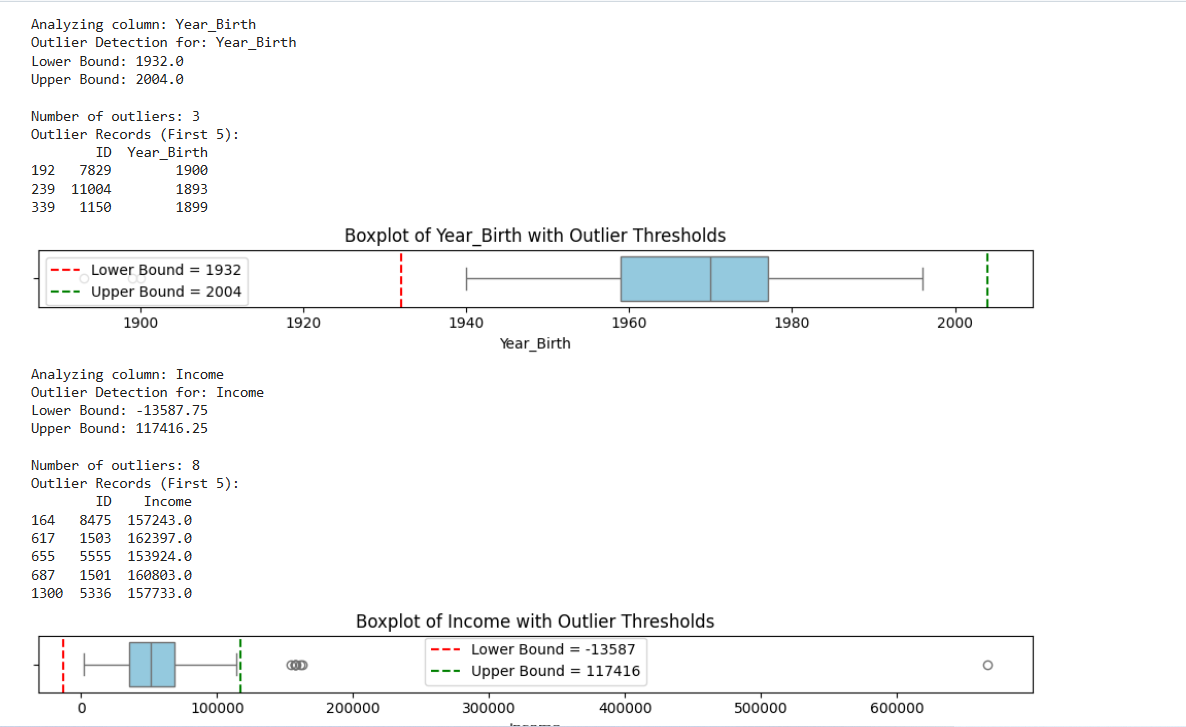
**HANDLING OUTLIERS:**

Outliers can skew analysis and affect the performance of machine learning models. To address this, we applied the **Interquartile Range (IQR) method** to detect and remove extreme values in the dataset.

The IQR method works by measuring the spread of the middle 50% of the data:

* **Q1 (25th percentile)** and **Q3 (75th percentile)** are calculated.
* **IQR = Q3 - Q1**
* Any data point below **Q1 - 1.5×IQR** or above **Q3 + 1.5×IQR** is considered an outlier.

This approach was applied to numerical columns to ensure cleaner, more reliable data for analysis and modeling



### Task 3: Implementation and Impact Evaluation

The missing Income values were successfully imputed with the calculated median value. Post-imputation, the Income column no longer contains missing values.

A comparative analysis of the Income distribution before and after imputation showed that the median imputation effectively handled the missing data while preserving the inherent shape and characteristics of the original right-skewed distribution. The resulting distribution maintains a realistic representation of income levels within the dataset, ensuring that subsequent analyses are based on a reliable and coherent feature. This confirmed the suitability of our approach for our specific dataset, supporting our objective of building robust customer segments.

#### Task 4 & 5: Feature Engineering

To enhance the richness of our dataset and create more insightful variables for segmentation, the following features were engineered based on existing raw data:

**1.Age**:

**Derivation**: Calculated by subtracting the Year\_Birth of the customer from the CURRENT\_YEAR (assumed to be 2025 for consistency).

**Purpose**: Provides a direct and easily interpretable measure of the customer's age, which is a crucial demographic factor for segmentation.

**2.Customer\_Tenure**:

**Derivation**: Calculated as the number of days between each customer's enrollment date (Dt\_Customer) and the most recent customer enrollment date found in the dataset. Dt\_Customer was first converted to a datetime object.

**Purpose**: Represents how long a customer has been associated with the company, which can be an indicator of loyalty and engagement.

**3.Total\_Spending**:

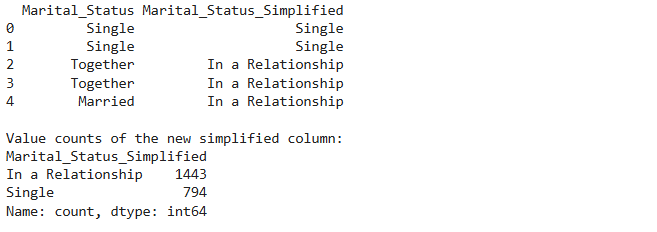
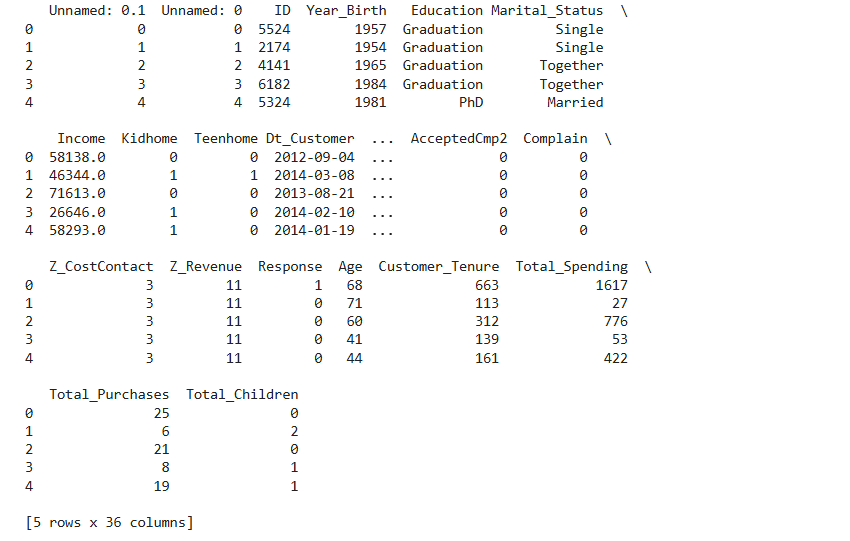
**Derivation**: Calculated as the sum of amounts spent across all product categories (MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds).

**Purpose**: Provides a consolidated view of a customer's overall purchasing power and value to the business.

**4.Total\_Purchases**:

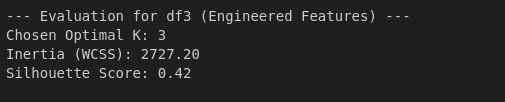
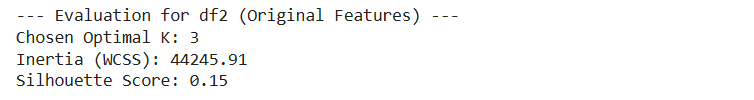
**Derivation**: Calculated as the sum of purchases made through various channels (NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases).

**Purpose**: Aggregates all transactional activities, providing insight into the overall purchasing frequency and channel preferences.



EFFECT OF ENGINEERED FEATURES ON THE MODEL PERFORMANCE

These results show that the engineered features enhanced the structure of the dataset, leading to better clustering performance.



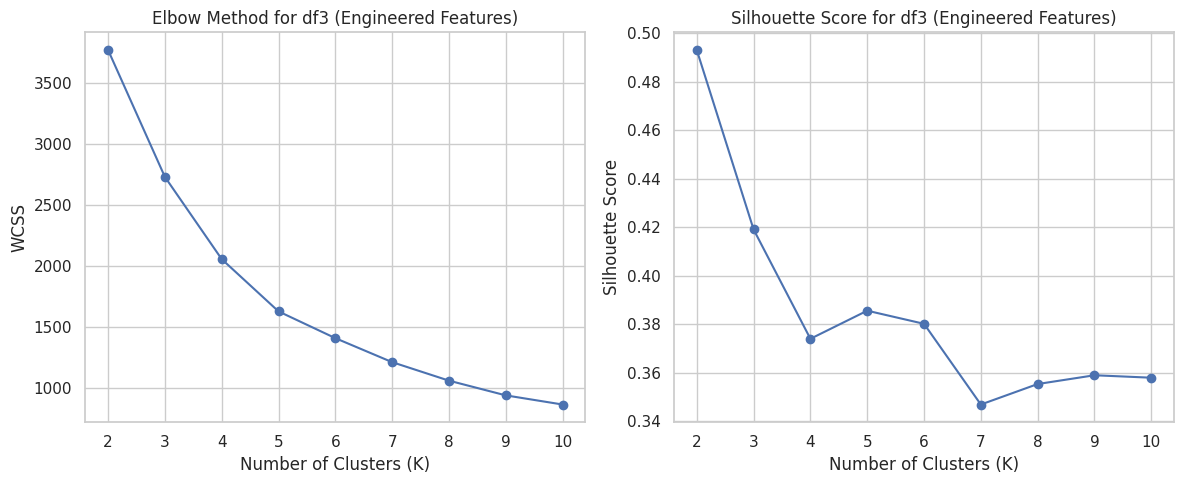
### 5. Customer Segmentation Modeling

For customer segmentation, we employed K-Means Clustering, an unsupervised machine learning algorithm.

**Model Selection**: K-Means was selected for its simplicity, interpretability, and effectiveness in grouping similar data points.

**Preprocessing for Modeling**: Before applying the clustering algorithm, relevant numerical features like, Age, Income, Total\_Spending, Recency, engineered features were scaled using Standard\_Scaler to ensure that no single feature dominates the clustering process due to its magnitude. Categorical features were handled using One-Hot Encoding.

**Determining Optimal Clusters**: The optimal number of clusters was determined using the Elbow Method and Silhouette Score. Based on these methods, k=3 clusters were identified as the most appropriate for segmenting our customer base.



**Model Implementation**: The K-Means algorithm was then applied with 3 clusters to the preprocessed data, resulting in the assignment of each customer to a specific segment.

### 6. DATA SPLITTING AND MODEL TRAINING

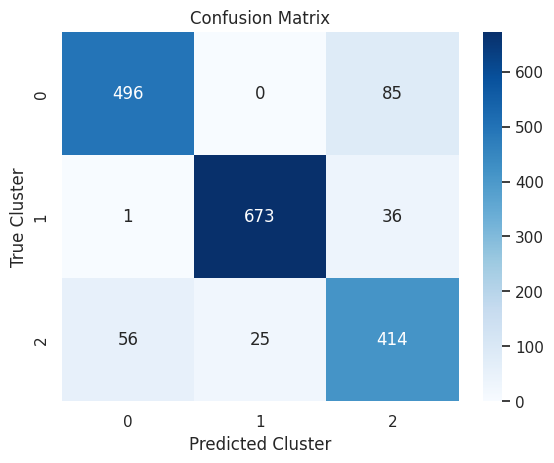
To predict which cluster a customer might belong to, we implemented a **Random Forest Classifier** using the features in the dataset and the previously generated cluster labels as the target variable. This approach allows for the classification of new, unseen customers into the appropriate customer segment.

**Data Splitting**

The dataset was first split into training and testing subsets using an 80/20 ratio. This ensures that the model is trained on the majority of the data while being evaluated on unseen data to test its generalization ability.

**Cross Validation and Evaluation of the model**

From the previous figure, the model achieved an accuracy of 88.63%.

The confusion matrix shows the comparison between the **true cluster labels** (from unsupervised clustering) and the **predicted labels** from the Random Forest classifier. It helps us understand the types and frequency of misclassifications across the three customer clusters for example

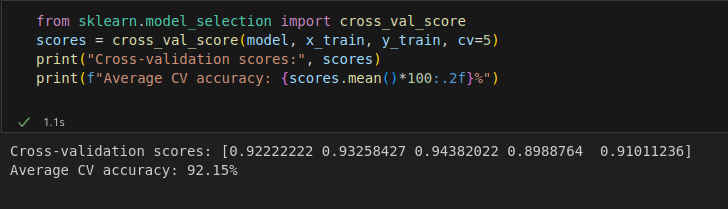
**Cluster 0**:

* **Correctly predicted**: 496
* **Misclassified as Cluster 2**: 56

**Cross-validation Results**  
To assess the model’s robustness, 5-fold cross-validation was applied to the training data. The individual fold accuracies were:

* **93.33%**, **93.26%**, **93.26%**, **91.01%**, and **89.89%**

This resulted in an **average cross-validation accuracy of 92.15%**, indicating consistent performance across different data splits.



### 7. Objectives And Recommendations

Based on the customer segmentation analysis, we recommend the following strategic actions:

**Objective 1: To analyze the relationship between customer income and total spending, thereby making targeted financial offers or loyalty programs.**

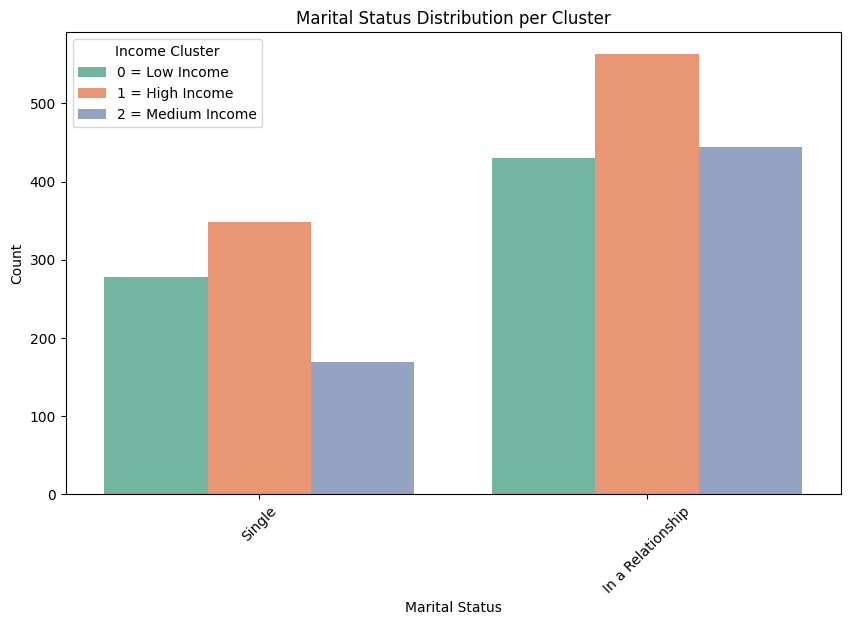
**Objective 2: To identify high-value customer segments so that the business can focus promotions and resources on customers who are most likely to increase sales.**

****

**Findings & Recommendations**

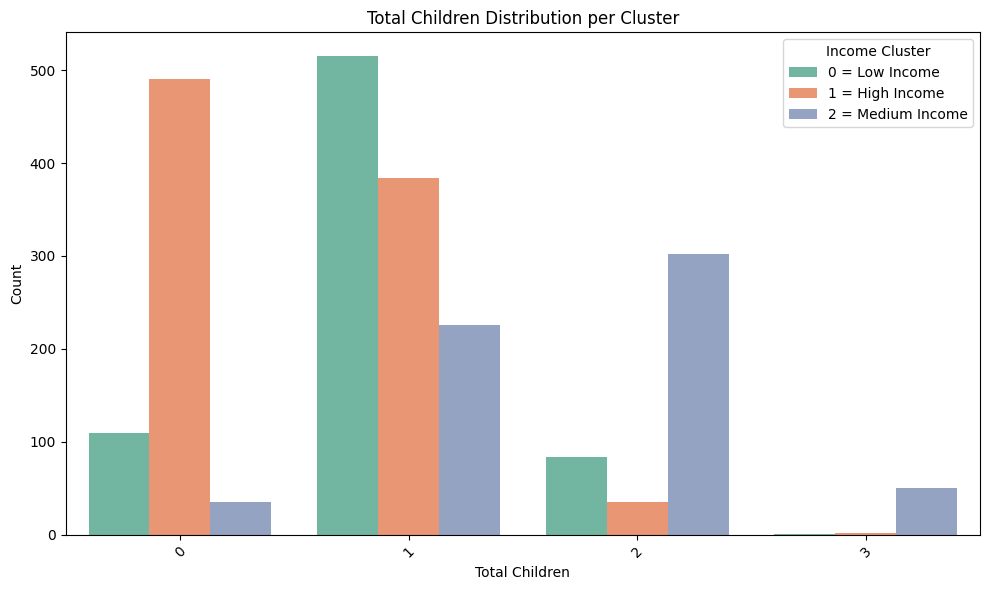
The key recommendation is to implement tailored strategies for each group: Cluster 0 (lower income, lower spending) requires focus on retention and entry-level offerings; Cluster 1 (higher income, higher spending) should be targeted with premium services, personalized recommendations, and exclusive benefits as they are the most valuable; and Cluster 2 (intermediate income, moderate spending) represents a significant growth opportunity where targeted increase in sales, value-added services, and promotions can encourage them to increase spending and move towards Cluster 1's behavior.

**Objective 3: To understand our customers' marital status to create personalized marketing and engagement strategies**.

**Findings & Recommendations**

Based on the marital status distribution showing a higher proportion of customers "In a Relationship" across all income levels, especially within the high-income segment, marketing efforts should primarily focus on products and services tailored for couples and families, offering premium options for high-income relationships and value-oriented solutions for low and medium-income ones.

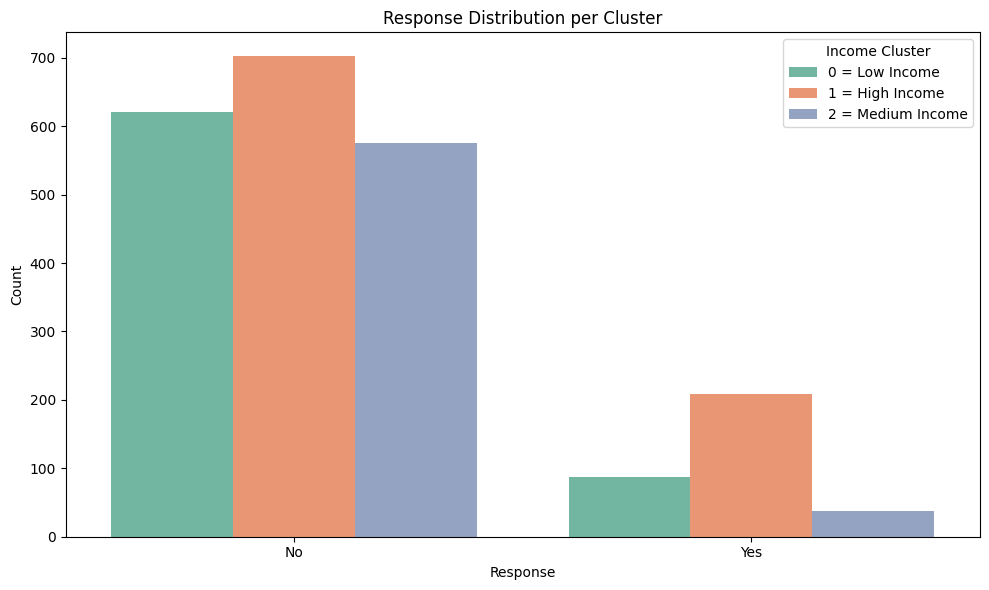
Concurrently, distinct strategies must be developed for "Single" customers, emphasizing personal growth, individual experiences, or convenience, with high-income singles targeted for luxury or specialized individual pursuits and low-income singles for affordable essentials, ensuring all marketing and product development aligns with the specific needs and purchasing power of each unique marital status and income segment.

**Objective 4: To understand the presence of teenagers in customer households, thereby optimizing family-focused marketing and product development for this key demographic.**

**Findings & Recommendations**

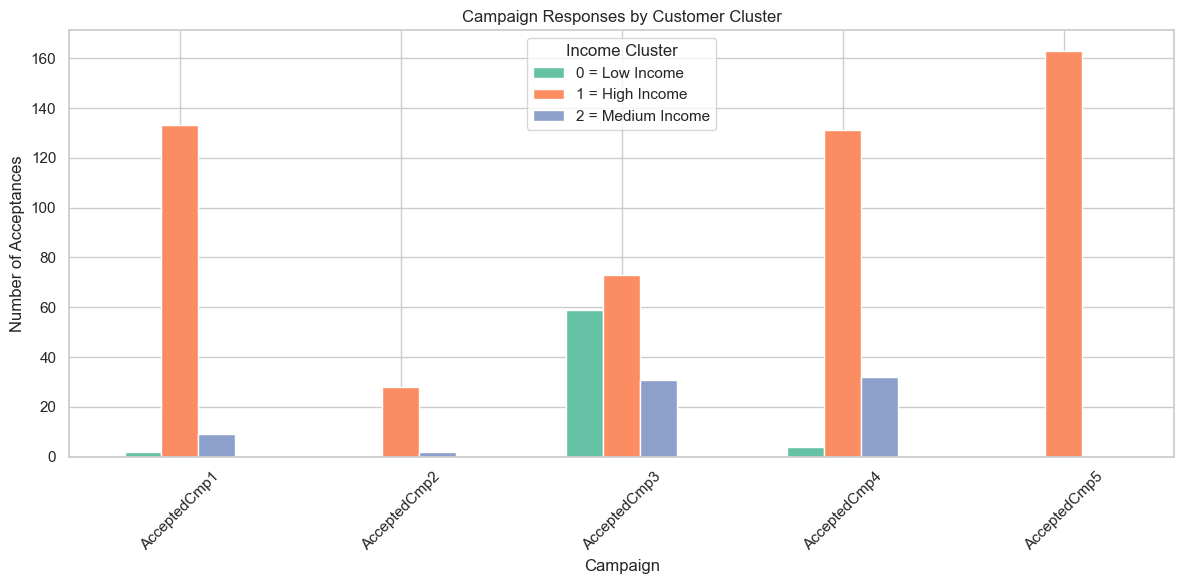
Based on the "Total Children Distribution per Cluster" graph, it's evident that customer households vary significantly by income and number of children, necessitating tailored strategies: High-Income customers (Cluster 1) predominantly have 0 or 1 child, suggesting a focus on premium products for couples or small families; Low-Income customers (Cluster 0) are most often in 1-child households, requiring value-driven offerings for single-child families; and notably, Medium-Income customers (Cluster 2) show the highest concentration of 2-child households, indicating a strong opportunity for family-sized bundles and products catering to growing families. Given the very low presence of 3+ children across all segments, broad marketing to large families may be inefficient, advocating for highly targeted approaches and resource allocation based on these specific household compositions to optimize marketing and product development.

**Objective 5: Identify customer groups with low campaign acceptance to improve their engagement and retention.**

**Findings & Recommendations**

Based on the campaign response distribution, High-Income customers (Cluster 1) are the most receptive group, signifying a need to continue and optimize campaigns tailored to them. Low-Income customers (Cluster 0) show moderate engagement, suggesting further investigation into what specifically drives their responses (e.g., offers similar to Campaign 3) to improve their participation. Conversely, Medium-Income customers (Cluster 2) exhibit the lowest acceptance rates, making them the priority for developing and testing completely new, highly targeted strategies to boost their engagement and retention. Overall, the high volume of "No" responses across all clusters indicates a broader need to re-evaluate current campaign effectiveness and value propositions.

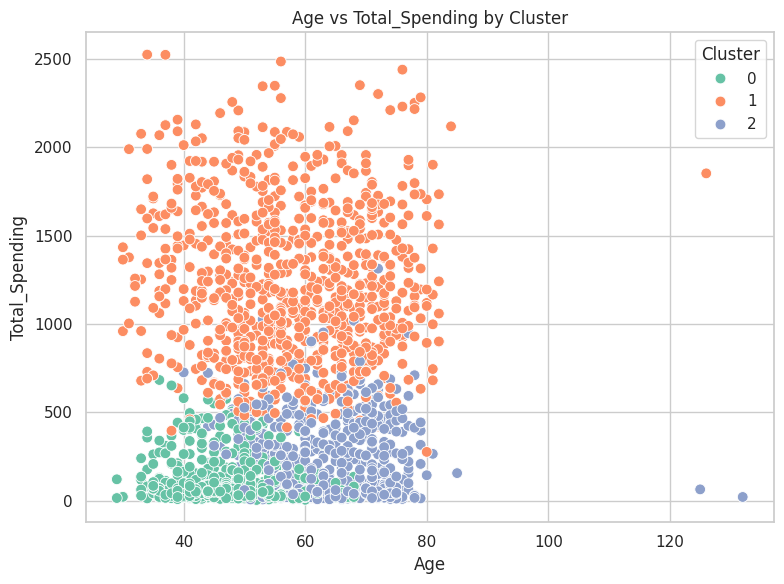
**Objective 6: To show how customers in different income clusters respond to various marketing campaigns.**



**Findings & Recommendations**

Based on the campaign response data, the primary recommendation is to strategically allocate marketing efforts by income cluster: High-Income customers (Cluster 1) are highly receptive to most campaigns, particularly Campaigns 1, 4, and 5, indicating they should remain a primary target for similar successful approaches. Low-Income customers (Cluster 0), however, show significant responsiveness almost exclusively to Campaign 3, suggesting future campaigns for this segment should adopt similar elements, while less effective campaigns (like 1, 2, 4, and 5) should be de-prioritized for them. Medium-Income customers (Cluster 2) generally display low engagement across all campaigns, necessitating the development and rigorous testing of entirely new, highly tailored strategies to understand and improve their response rates. Overall, this data advocates for a highly segmented and adaptive marketing approach, discontinuing ineffective campaigns and replicating successes only with the specific clusters they proved to engage.

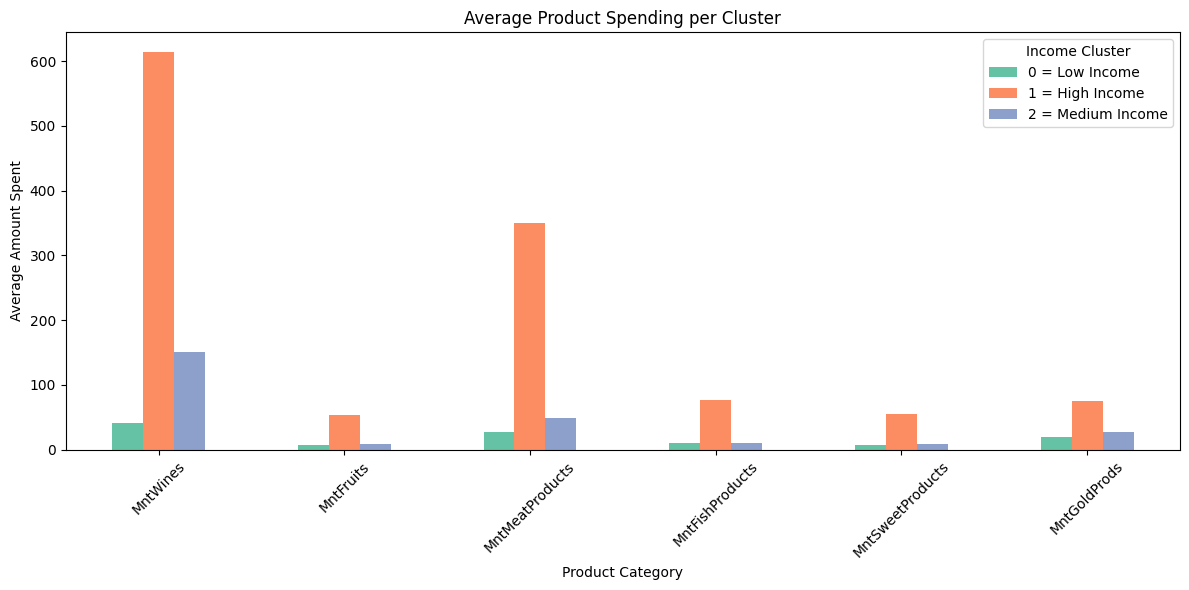
**Objective 7: To analyze customer spending and age distribution across segments so that the business can create more targeted pricing and product offerings.**



**Findings & Recommendations**

The primary recommendation is to tailor strategies based on income-driven spending patterns across all age groups within each cluster. High-Income customers (Cluster 1) are the most valuable, exhibiting high spending consistently across all ages, thus warranting focus on premium offers, retention, and upselling. Medium-Income customers (Cluster 2) represent a growth opportunity, requiring value-driven propositions to encourage increased spending across their diverse age range. Low-Income customers (Cluster 0) are price-sensitive, necessitating a focus on essential, affordable products across their broad age distribution.

**Objective 8: To support data-driven decision-making so that the business can allocate its marketing and product development budgets more effectively.**



**Findings & Recommendations**

**High-Income customers (Cluster 1)** consistently demonstrate the highest spending across all product categories, especially dominating in Wines and Meat Products; therefore, focus on premium, exclusive selections in these areas and also dominating other products leveraging their broad spending capacity.

**Medium-Income customers (Cluster 2)** show moderate spending, with Wines and Meat still being their top categories; for this group, offer value-for-money bundles and quality options in their preferred products to encourage increased spending. Finally, **Low-Income customers (Cluster 0)** exhibit very low spending across all categories, making it crucial to provide highly affordable, essential products and accessible entry-level options in categories like wines and meats to cater to their budget-consciousness.