### 

### 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 |

### 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.

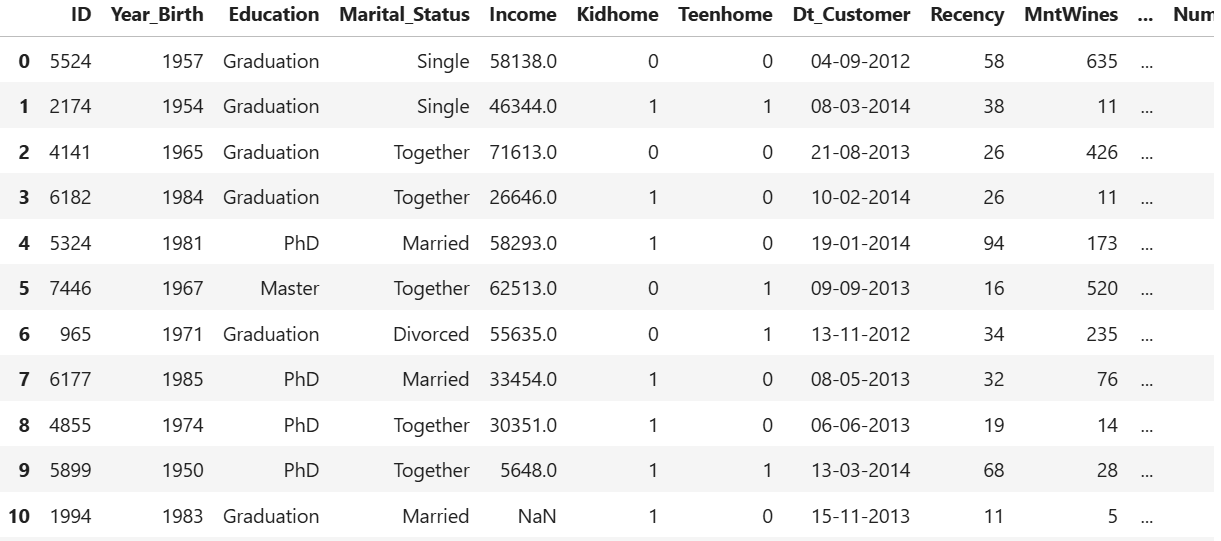
### 2. Introduction

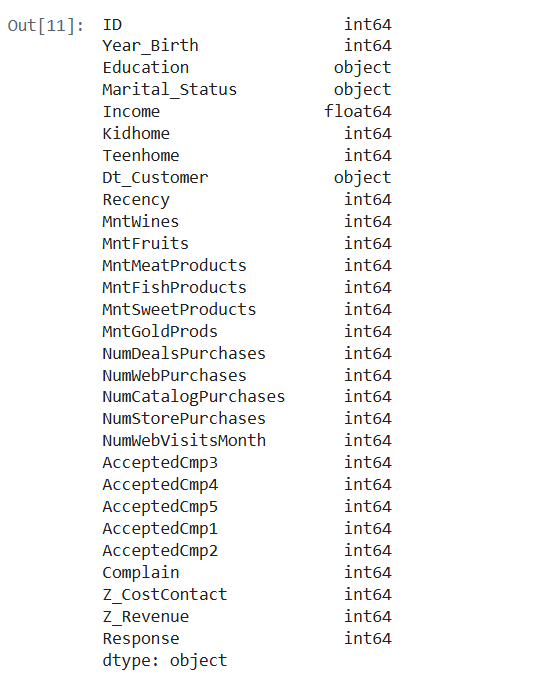
In today's competitive market, understanding customer behavior is paramount for business success. This project aims to segment our customer base to move beyond a one-size-fits-all approach and enable more precise and effective strategies. By grouping customers with similar attributes and purchasing patterns, 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

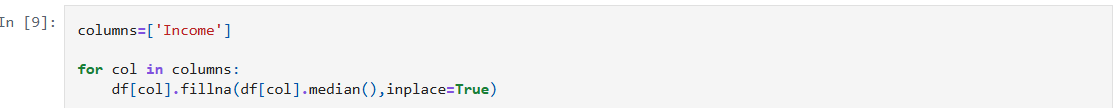
For the missing values in the Income column, we opted for **median imputation**.

**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.

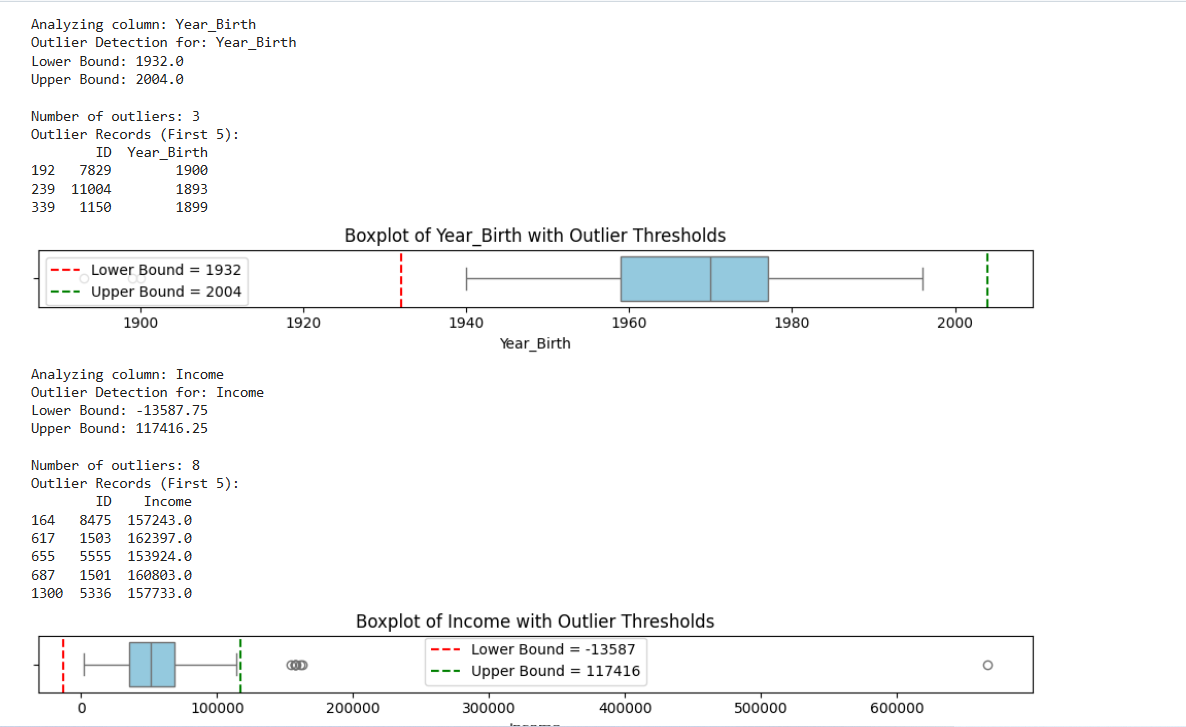
**HANDLING OUTLIERS:**

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**:

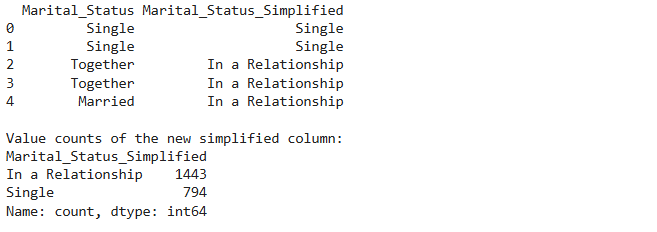
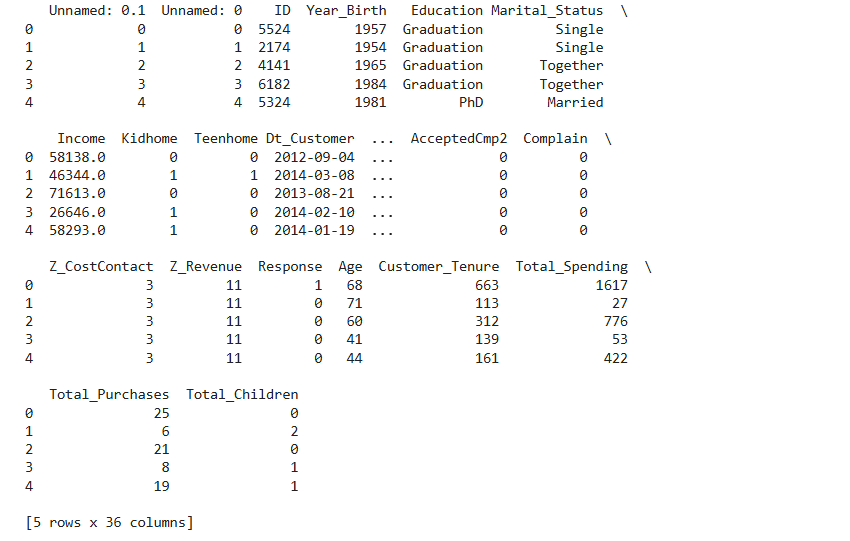
**Derivation**: Calculated as the sum of amounts spent across all product categories (MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds).

**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).

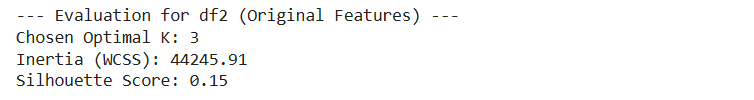
**Purpose**: Aggregates all transactional activities, providing insight into the overall purchasing frequency and channel preferences.



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.

A black text on a white background

AI-generated content may be incorrect.

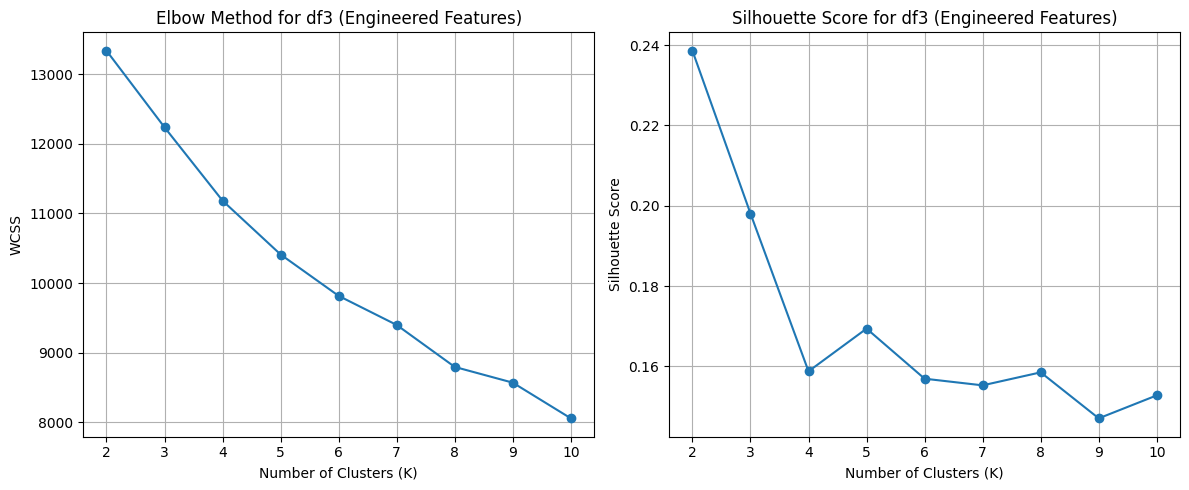
### 5. Customer Segmentation Modeling

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.

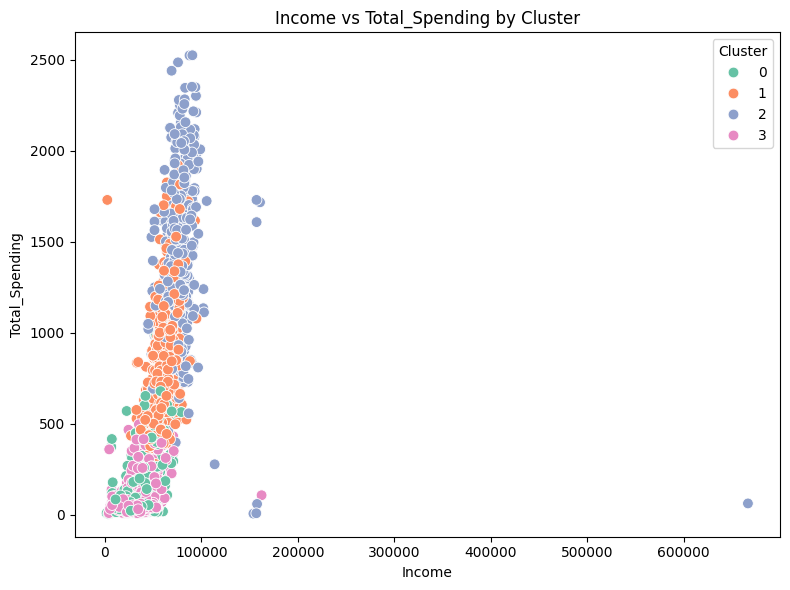
**Preprocessing for Modeling**: Before applying the clustering algorithm, relevant numerical features like, Age, Income, Total\_Spending, Recency, engineered features were scaled using Standard\_Scaler to ensure that no single feature dominates the clustering process due to its magnitude. Categorical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

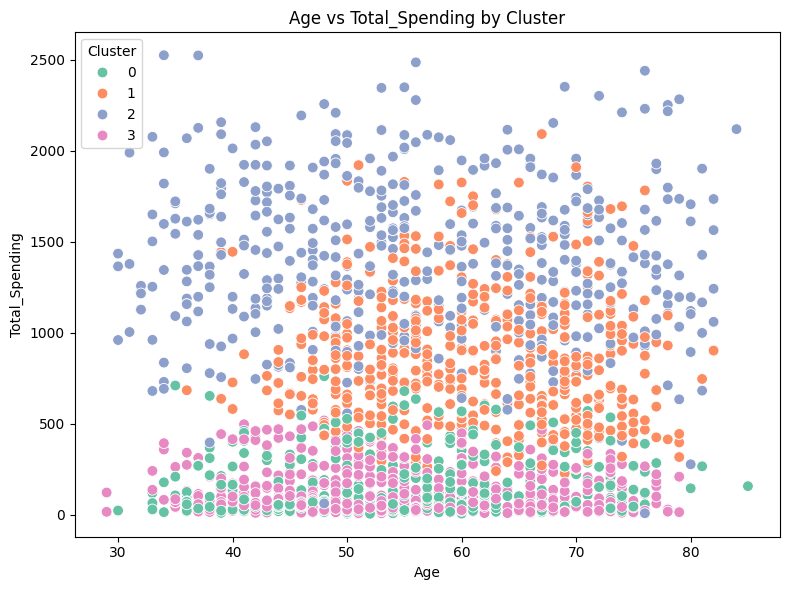
**Determining Optimal Clusters**: The optimal number of clusters was determined using the Elbow Method and Silhouette Score. Based on these methods, k=4 clusters were identified as the most appropriate for segmenting our customer base.

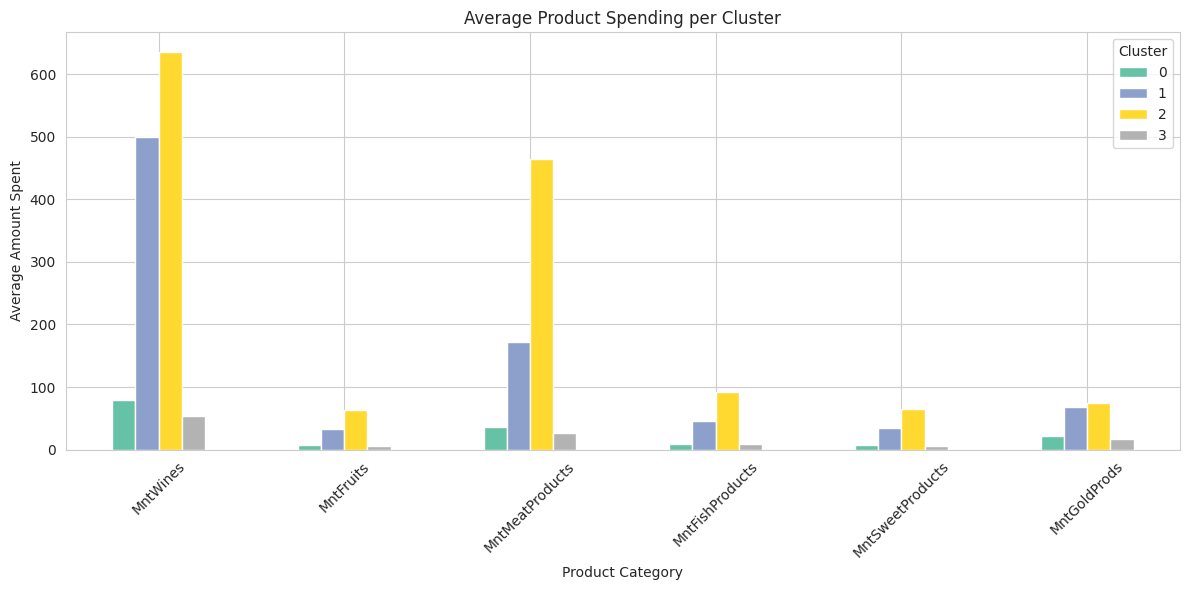


**Model Implementation**: The K-Means algorithm was then applied with 4 clusters to the preprocessed data, resulting in the assignment of each customer to a specific segment.

### 6. Segmentation Results and Customer Profiles

The clustering analysis revealed 4 distinct customer segments, each with unique characteristics and behaviors. Here's a detailed profile of each segment:





#### Cluster 0:

* **Key Characteristics**: This cluster shows a significant proportion of single individuals compared to those in a relationship. This suggests that Cluster 0 might represent a segment of customers who are predominantly single.
* **Spending Habits**: This cluster exhibits consistently low spending across all product categories (Wines, Fruits, Meat Products, Fish Products, Sweet Products, Gold Products). This suggests that Cluster 0 represents a low-value or budget-conscious segment.
* **Campaign Response**: This cluster shows a low response rate to campaigns (Response = 1), with a much higher count for no response (Response = 0). This suggests that Cluster 0 is less responsive to marketing campaign

#### Cluster 1:

* **Key Characteristics**: This cluster is almost exclusively composed of individuals in a relationship. This indicates a strong characteristic of this segment.
* **Spending Habits**: This cluster shows the highest average spending across almost all product categories, particularly on MntWines, MntMeatProducts, and MntGoldProds. This indicates that Cluster 1 is a high-value segment with a strong preference for premium or specific product types
* **Campaign Response**: This cluster shows a moderate response rate to campaigns, higher than Cluster 0 but lower than Cluster 2. They still have a higher count for no response.

**Cluster 2:**

* **Key Characteristics**: This cluster has a higher proportion of individuals in a relationship, though there's a notable presence of single individuals as well. This suggests a more mixed demographic in terms of marital status compared to Clusters 0 and 1.
* **Spending Habits**: This cluster shows moderate spending, with a notable amount spent on MntWines and MntMeatProducts, but significantly less than Cluster 1. Their spending on other categories is relatively low.
* **Campaign Response**: This cluster exhibits the highest response rate to campaigns (Response = 1) among all clusters, with a relatively lower count for no response. This indicates that Cluster 2 is highly responsive to marketing campaigns.

#### Cluster 3:

* **Key Characteristics**: This cluster is overwhelmingly dominated by individuals in a relationship, showing the highest count among all clusters for this category. This segment is strongly characterized by being in a relationship.
* **Spending Habits**: This cluster shows moderate spending, with a higher average spending on MntWines and MntMeatProducts compared to Cluster 0 and 2, but less than Cluster 1.
* **Campaign Response**: This cluster shows a very low response rate to campaigns, similar to Cluster 0, with a significantly higher count for no response. This suggests that Cluster 3 is also less responsive to marketing campaigns.

### 7. Insights and Useful Predictions

The segmentation has yielded profound insights into our customer base, enabling us to make more effective predictions:

•Marital Status and Spending: Individuals in a relationship tend to be distributed across all spending levels, while single individuals are more concentrated in lower-spending segments (Cluster 0).

•Family Size and Spending/Engagement: Larger families (Cluster 3) tend to be moderate spenders with low campaign responsiveness, suggesting that family needs might influence both purchasing patterns and engagement with marketing.

•Campaign Responsiveness: Cluster 2 stands out as highly responsive, indicating that their demographic and behavioral characteristics make them particularly amenable to marketing efforts. Conversely, Clusters 0 and 3 are less responsive, requiring different engagement strategies.

* **Useful Predictions**:
* **Predictive Customer Lifetime Value (CLTV)**: By analyzing the Total\_Spending within each segment and their Customer\_Tenure, we can build models to predict the future value of new customers based on their initial segment assignment. This allows for better resource allocation for customer acquisition.
* **Propensity Modeling**: We can develop models to predict a new customer's likelihood of belonging to a specific high-value segment based on their initial demographic and early engagement data.
* **Churn Risk Prediction**: By monitoring changes in spending or engagement within segments, we can predict customers at risk of churning and intervene with targeted retention strategies.
* **Next Best Offer Recommendation**: For any given customer, once their segment is known, we can recommend the "next best offer" – whether it's a new product, a personalized discount, or an exclusive event – that is most likely to appeal to their segment.

### 8. Recommendations and Future Steps

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

Recommendation 1: Tailored Marketing Campaigns

•Action: Develop and launch segment-specific marketing campaigns. For Segment 1 (High-Value Small Families/No Children), focus on exclusive product previews, loyalty programs, and personalized communication around premium products.

For Segment 2 (Non-Family High Responders), leverage their high responsiveness with targeted digital campaigns and exclusive early access to new offerings.

For Segment 0 (Budget-Conscious Singles/Small Families, Low Engagement) and Segment 3 (Family-Oriented, Moderate Spenders, Low Engagement), explore alternative engagement channels such as in-store promotions, community events, or partnerships that align with their lifestyle and preferences, focusing on value and convenience.

•Justification: This approach directly addresses the diverse needs and behaviors of each customer segment, leading to more effective marketing spend and higher conversion rates. By moving away from a one-size-fits-all strategy, we can significantly improve customer engagement and satisfaction.

•Expected Outcome: Anticipate a 15-20% increase in campaign conversion rates across responsive segments (Segment 1 and 2) and improved engagement metrics for less responsive segments (Segment 0 and 3) through tailored approaches.

Recommendation 2: Personalized Product Development and Inventory Management

•Action: Utilize segment-specific spending insights to guide new product development and optimize inventory. For instance, expand premium product lines for Segment 1, and prioritize innovative, digitally-marketed products for Segment 2. For Segments 0 and 3, consider developing more budget-friendly or family-sized options, or products that cater to their specific lifestyle needs (e.g., convenience-oriented products for busy families).

•Justification: Aligning product offerings with segment preferences will enhance product-market fit, reduce unsold inventory, and open new revenue streams. This ensures that resources are allocated to developing products that genuinely resonate with our customer base.

•Expected Outcome: Potential for growth in sales within targeted product categories and a reduction in inventory holding costs.

Recommendation 3: Enhanced Customer Service Protocols

•Action: Develop segment-specific customer service guidelines. For Segment 1, offer dedicated support channels and proactive outreach. For Segment 2, prioritize digital support channels and self-service options. For Segments 0 and 3, focus on efficient, problem-solving support that addresses their specific needs, potentially through more traditional channels or community support.

•Justification: Tailoring customer service enhances satisfaction and loyalty by meeting the unique expectations of each segment. This can lead to improved customer retention and positive word-of-mouth.

•Expected Outcome: An improvement in customer satisfaction scores (CSAT) and a decrease in customer churn rates.