

**MAKERERE UNIVERSITY**

**COLLEGE OF COMPUTING AND INFORMATION**

**SCIENCES**

**DEPARTMENT OF NETWORKS**

**CUSTOMER SEGMENTATION ANALYSIS REPORT**

**PROJECT TITLE**: Customer Segmentation for Enhanced Marketing Strategies

**YEAR** : RECESS YEAR TWO

**REASON**: DATASCIENCE PROJECT

**DATE OF SUBMISSION: July 13, 2025**

**GROUP B EVENING**

* **Ssenoga Herman 23/U/17648/EVE**
* **Mugoya Yusuf 23/U/11615/EVE**
* **Walusimbi Ashraf 23/U/18466/EVE**
* **Nabirye Anitah 23/U/12826/EVE**
* **Nyonyozi Maria Lisa Loyce 23/U/16424/EVE**

### 1. Executive Summary

This report presents the findings of a customer segmentation analysis performed on the provided dataset. The objective was 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.

### 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.

### 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.

### Task 3: Implementation and Impact Evaluation

The missing Income values were successfully imputed with the calculated median value of [Median Value, e.g., 51,381.50]. 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).

CURRENT\_YEAR = 2025

df2['Age'] = CURRENT\_YEAR - df2['Year\_Birth']

**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.

# Convert 'Dt\_Customer' to datetime objects

df2['Dt\_Customer'] = pd.to\_datetime(df2['Dt\_Customer'], format='%d-%m-%Y')

# Find the most recent date in the dataset to calculate tenure relative to it

most\_recent\_date = df2['Dt\_Customer'].max()

# Calculate tenure in days

df2['Customer\_Tenure'] = (most\_recent\_date - df2['Dt\_Customer']).dt.days

**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).

spending\_cols = [

'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts',

'MntSweetProducts', 'MntGoldProds'

]

df2['Total\_Spending'] = df2[spending\_cols].sum(axis=1)

**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.

purchase\_cols = [

'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases'

]

df2['Total\_Purchases'] = df2[purchase\_cols].sum(axis=1)

### 5. Customer Segmentation Modeling

For customer segmentation, we employed [Chosen Model, e.g., K-Means Clustering], an unsupervised machine learning algorithm.

**Model Selection**: [Briefly explain why you chose this model, e.g., "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 (e.g., Age, Income, Total\_Spending, Recency, engineered features) were scaled using [e.g., StandardScaler] to ensure that no single feature dominates the clustering process due to its magnitude. Categorical features were handled using [e.g., One-Hot Encoding].

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

**Model Implementation**: The [e.g., K-Means] algorithm was then applied with [e.g., X] 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 [e.g., X] distinct customer segments, each with unique characteristics and behaviors. Here's a detailed profile of each segment:

#### Segment 1: [Give a descriptive name, e.g., "Affluent Wine Enthusiasts"]

* **Key Characteristics**: [Describe average age, income, family size, marital status, education level for this segment].
* **Spending Habits**: [Detail their spending patterns on different products, e.g., "Highest spending on wines and meat products, moderate on fruits."].
* **Engagement**: [Describe their purchase channels, web visits, and campaign responses, e.g., "Frequent online and catalog shoppers, high response rate to Campaign 1 and 5."].
* **Insights**: [Summarize key insights, e.g., "This segment represents our most loyal and high-value customers, prioritizing quality over deals."].

#### Segment 2: [Give a descriptive name, e.g., "Young Families, Value Seekers"]

* **Key Characteristics**: [Describe average age, income, family size, marital status, education level for this segment].
* **Spending Habits**: [Detail their spending patterns, e.g., "Lower overall spending, but higher spending on essential items or products for children. More responsive to deals."].
* **Engagement**: [Describe their purchase channels, e.g., "Prefer in-store purchases, less frequent web visits, respond well to discount campaigns."].
* **Insights**: [Summarize key insights, e.g., "This segment is price-sensitive and family-focused, looking for value."].

#### Segment 3: [Give a descriptive name, e.g., "Digital Gold Shoppers"]

* **Key Characteristics**: [Describe average age, income, family size, marital status, education level for this segment].
* **Spending Habits**: [Detail their spending patterns, e.g., "Significant spending on gold products, minimal on other categories."].
* **Engagement**: [Describe their purchase channels, e.g., "Primarily web purchasers, low in-store activity."].
* **Insights**: [Summarize key insights, e.g., "A niche segment focused on specific high-value items, preferring convenience."].

### 7. Insights and Useful Predictions

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

* **Behavioral Insights**: We can now clearly see that different demographics and household structures correlate with distinct spending habits and campaign responsiveness. For instance, older, single customers without children tend to spend more on luxury items like wine, while younger, married customers with kids are more budget-conscious.
* **Campaign Optimization**: We have a clear understanding of which campaign types resonate with which segments. This allows for highly optimized future marketing efforts, reducing wasted ad spend and increasing conversion rates.
* **Product Development Insights**: The spending patterns of each segment directly inform product development. If a segment with significant market share is underserved in a particular product category, it presents a clear opportunity for new offerings.

**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 (e.g., "Affluent Wine Enthusiasts") based on their initial demographic and early engagement data.
* **Churn Risk Prediction**: By monitoring changes in spending or engagement within segments (e.g., a "Value Seeker" suddenly reducing purchases), 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:

* **Tailored Marketing Campaigns**:
  + For "Affluent Wine Enthusiasts," focus on premium product launches, exclusive tasting events, and loyalty programs.
  + For "Young Families, Value Seekers," emphasize family-sized discounts, essential product bundles, and loyalty points.
  + For "Digital Gold Shoppers," continue to enhance the online shopping experience and highlight new high-value or unique gold items.
* **Personalized Product Development**: Utilize segment-specific spending insights to guide new product development and inventory management.
* **Optimized Customer Service**: Develop segment-specific customer service protocols, such as dedicated support channels for high-value segments.
* **Targeted Outreach for Campaign Response**: Leverage the insights from AcceptedCmp columns to design future campaigns that are more likely to resonate with each identified segment.
* **Data Collection Enhancement**: Continuously collect data on customer interactions and preferences to refine existing segments and identify emerging ones.

**Future Steps**:

* **Implement A/B Testing**: Conduct A/B tests on targeted marketing campaigns to validate the effectiveness of segment-specific strategies.
* **Develop Predictive Models**: Build and deploy supervised learning models to automate the prediction of new customer segments and key behaviors like CLTV and churn.
* **Integrate with CRM**: Integrate segmentation insights into the Customer Relationship Management (CRM) system for real-time personalized interactions.
* **Monitor Segment Evolution**: Regularly revisit and re-evaluate the customer segments to account for changes in market trends and customer behavior.