*Subash Yadav*

*Webster University*

*CSDA 6010 – Project 3*

*Prof. JP Wang*

*Consumer Segmentation analytics*

Data Analytics & Modeling

**Executive Summary**

This project for AXANTEUS focused on creating a consumer segmentation and predictive analytics framework to optimize targeted marketing strategies. Through clustering and predictive modeling, the analysis delivered actionable insights into consumer behaviors, with a focus on value-consciousness and brand loyalty.

The project began with a comprehensive exploratory data analysis (EDA) of a 600-record dataset with 46 variables covering demographic and transactional information. Key factors, such as purchase frequency, brand loyalty, category diversity, and price sensitivity, were identified as critical drivers for segmentation and prediction. The dataset's completeness facilitated smooth transitions into modeling and analysis.

Three clustering analyses were performed using the K-means algorithm to segment consumers based on their behaviors and motivations: Purchase Behavior, Purchase Motivation, and Price Sensitivity. The optimal number of clusters for each analysis was determined to be three, based on the Elbow and Silhouette methods. Each approach produced unique consumer profiles that provided critical insights into purchasing behaviors. The Purchase Behavior clustering identified high-frequency shoppers, moderate buyers, and low-engagement shoppers. The Purchase Motivation clustering distinguished price-sensitive consumers, balanced shoppers, and deal-resistant, variety-seeking buyers. These insights were synthesized through a combined clustering approach, which integrated behavioral and motivational dimensions to yield a comprehensive segmentation of consumer profiles. The combined clustering was ultimately selected as the foundation for further analysis, as it provided a nuanced understanding of consumer engagement, deal sensitivity, and product diversity preferences.

Following the clustering analysis, two classification models were developed to predict key consumer attributes: value-consciousness and brand loyalty. Logistic regression and random forest models were used for value-consciousness classification, with the random forest achieving an accuracy of 84.57%, effectively identifying high spenders and frequent buyers. For brand loyalty, the random forest model reached an accuracy of 91.98%, successfully predicting consistent brand-specific purchasing behaviors and identifying loyal customers.

A regression model was also created to predict "brand runs," or the frequency of brand-specific purchases. Using predictors like category diversity, transaction frequency, and affluence index, the model provided valuable insights into factors influencing loyalty. Its strong predictive performance offered AXANTEUS a deeper understanding of behaviors driving repeat purchases and brand affinity.

This project equips AXANTEUS with a behavior-driven segmentation framework, moving beyond traditional demographic approaches. The insights enable precise marketing strategies, optimized loyalty programs, and enhanced customer engagement. Future improvements, such as feature engineering and iterative tuning, will ensure the framework remains aligned with AXANTEUS’s evolving goals. This analysis underscores the power of data-driven decisions in achieving strategic business success.

Table of Contents

[Business Problem 3](#_Toc183615097)

[Business Goal 3](#_Toc183615098)

[Analytical Goal 4](#_Toc183615099)

[Analytical Approach 5](#_Toc183615100)

[Dataset Overview 6](#_Toc183615101)

[Dataset Composition 9](#_Toc183615102)

[Dataset Composition 9](#_Toc183615103)

[Data Preprocessing 14](#_Toc183615104)

[Overview of Missing Values 15](#_Toc183615105)

[Data Visualization 16](#_Toc183615106)

[Demographical Column Removal 32](#_Toc183615107)

[Predictor Analysis and Relevancy 33](#_Toc183615108)

[Highly Correlated Variable Pairs 33](#_Toc183615109)

[Dimension Reduction & Data Transformation 35](#_Toc183615110)

[Model Selection for Clustering 38](#_Toc183615111)

[Elbow Method 39](#_Toc183615112)

[Silhouette Plot 40](#_Toc183615113)

[K-Means Clustering: Approach and Business Goals 40](#_Toc183615114)

[Profiling using Combine Clustering 46](#_Toc183615115)

[Classification Task: Value-Conscious Classification 47](#_Toc183615116)

[Classification Task: Brand Loyalty Prediction 47](#_Toc183615117)

[Regression Task: Brand Runs Prediction 48](#_Toc183615118)

[Data Partitioning 49](#_Toc183615119)

[Model Selection for Classification and Regression 50](#_Toc183615120)

[Random Forest Model: Value-Conscious Classification 54](#_Toc183615121)

[Brand Loyalty Classification Using Random Forest 56](#_Toc183615122)

[Regression Model: Brand Runs Prediction 58](#_Toc183615123)

[Key Insights from the Analysis 60](#_Toc183615124)

[Next Steps and Recommendations 62](#_Toc183615125)

[Conclusion 62](#_Toc183615126)

# **Business Problem**

AXANTEUS, a market research agency, currently segments consumers based on demographic data alone, which limits its ability to understand complex consumer behaviors and preferences. This approach does not provide insights into critical purchasing factors such as transaction frequency, brand loyalty, price sensitivity, and promotional responsiveness. As a result, AXANTEUS’s clients, including advertising agencies and consumer goods manufacturers, are unable to design marketing and reward programs that align closely with actual consumer buying patterns.

The key challenges include:

1. **Lack of Behavioral Insights**: Demographic data does not capture essential behavioral attributes like purchase volume, frequency, brand preferences, and sensitivity to promotions, all which influence consumer decisions.
2. **Need for Targeted Marketing and Rewards**: AXANTEUS’s clients require deeper insights into consumer behavior to optimize marketing campaigns, allocate promotional budgets effectively, and enhance customer loyalty.
3. **Inability to Identify Value-Conscious and Brand-Loyal Consumers**: Without behavioral segmentation, clients struggle to pinpoint price-sensitive customers and loyal buyers, which limits their ability to engage these groups with targeted promotions and rewards.

To address these challenges, AXANTEUS needs to develop a segmentation model based on consumer purchasing behaviors and motivations. This model will enable AXANTEUS’s clients to implement behavior-based marketing strategies and personalized reward programs, improving engagement, optimizing promotional spending, and building long-term customer loyalty.

# **Business Goal**

The business goal for AXANTEUS’s consumer segmentation project is to implement a behavior-based segmentation model using both clustering (non-supervised) and classification/prediction (supervised) approaches. This approach will enable AXANTEUS to provide clients with actionable insights that allow for targeted marketing strategies and loyalty programs based on consumer behaviors and purchase motivations.

This goal will be achieved in two stages:

**Clustering (Unsupervised Analysis)**

**Goal**: To segment consumers based on their **Purchase Behavior** (e.g., purchase volume, transaction frequency, brand loyalty, and discount responsiveness) and **Purchase Motivation** (e.g., price sensitivity, brand preference, and responsiveness to promotions) and **Combined.**

**Expected Outcome**: This analysis will yield behavior-based segments, such as high-frequency, price-conscious buyers and brand-loyal consumers. These clusters will allow clients to create targeted promotions, allocate budgets effectively, and develop rewards programs tailored to specific consumer groups.

**Classification and Prediction (Supervised Analysis)**

After defining consumer segments through clustering, the next step involves building predictive models to address two specific objectives:

1. **Classify Value-Conscious Consumers**: Identify consumers responsive to deals and discounts, enabling focused deal-driven marketing.
2. **Predict Brand Loyalty Patterns**: Forecast “brand runs” or consecutive purchases of the same brand, supporting targeted loyalty programs for loyal customers.

By starting with clustering to establish behavior-based segments, AXANTEUS can use these segments to refine predictive models for value-conscious classification and brand loyalty. This sequential approach enables AXANTEUS to provide clients with precise and targeted insights, facilitating optimized marketing strategies and improved customer retention.

# **Analytical Goal**

The analytical goal of this project is to develop a behavior-based consumer segmentation model that provides deeper insights into purchasing behaviors and motivations, allowing clients to implement targeted marketing and loyalty strategies. This will be achieved in two phases:

1. **Clustering Analysis (Unsupervised)**

**Goal**: Segment consumers into distinct groups based on **Purchase Behavior** (e.g., transaction frequency, purchase volume, brand loyalty, and discount responsiveness) and **Purchase Motivation** (e.g., price sensitivity, brand preferences, and promotional responsiveness).

* **Objective**: Create clear, behavior-based segments that reflect unique consumer profiles, such as “high-frequency, price-conscious buyers” or “brand-loyal consumers.”
* **Evaluation**: Ensure clusters are stable and interpretable, providing clients with actionable segments that enable targeted promotions, efficient budget allocation, and tailored reward programs.

1. **Classification and Prediction Models (Supervised)**

Following clustering, develop predictive models to classify and forecast specific consumer behaviors within these segments, supporting more refined marketing and loyalty efforts.

* 1. **Value-Conscious Classification**:
     1. **Objective**: Identify consumers highly responsive to discounts and promotions, enabling clients to target price-sensitive groups.
     2. **Target Variable**: Define value-consciousness as a binary outcome (1 = value-conscious, 0 = not value-conscious) based on variables like frequency of discount purchases and bulk buying behavior.
     3. **Expected Outcome**: A classification model that directs deal-based marketing strategies toward consumers who respond most strongly to promotions.
  2. **Brand Loyalty Prediction (Brand Runs)**:
     1. **Objective**: Forecast brand loyalty, identifying consumers likely to make consecutive purchases of the same brand.
     2. **Target Variable**: Brand loyalty defined as repeated purchases of the same brand, used to direct loyalty programs and retention strategies.
     3. **Expected Outcome**: A predictive model that enables clients to engage brand-loyal consumers with personalized rewards, fostering long-term customer loyalty.

This analytical goal supports a shift from demographic-only segmentation to a behavior-focused approach, providing clients with precise and actionable insights to enhance customer engagement and optimize marketing spend.

# **Analytical Approach**

The analytical approach for this project involves a structured process across two stages: **clustering** (unsupervised) to identify consumer segments and **classification and prediction** (supervised) to refine those segments. Each stage includes essential data preparation, model development, and evaluation steps to ensure actionable insights for targeted marketing and loyalty strategies.

1. **Data Exploration and Preprocessing**

The initial phase focuses on understanding and preparing the data:

* **Exploratory Data Analysis (EDA)**: Analyze the dataset to understand key variables, detect patterns, and identify missing values or outliers. Visualizations such as histograms, scatter plots, and correlation matrices will help to observe distributions, relationships, and potential correlations between features.
* **Data Cleaning**: Address missing values by applying imputation techniques or removing incomplete records, depending on the data's structure and the proportion of missing values. Detect and treat outliers that may skew analysis results, especially in continuous variables like purchase volume and transaction frequency.
* **Data Transformation**: Standardize or normalize numerical variables (e.g., transaction frequency, purchase volume) to bring all features onto a comparable scale, which is crucial for clustering. Encode categorical variables, such as brand preferences, into numerical format if required, using techniques like one-hot encoding.

**2. Clustering Analysis (Unsupervised)**

The clustering stage aims to identify distinct consumer segments based on behavioral and motivational factors:

* **Feature Selection for Clustering**: Use variables related to **Purchase Behavior** (e.g., transaction frequency, purchase volume, brand loyalty) and **Purchase Motivation** (e.g., price sensitivity, brand preference, responsiveness to promotions) as inputs for clustering. These features will help capture both *how* and *why* consumers make purchasing decisions.
* **Clustering Algorithm**: Apply clustering techniques such as **K-Means** or **Hierarchical Clustering**. K-Means is suitable for handling larger datasets and is efficient when the number of clusters is known. The Elbow and Silhouette methods will be used to determine the optimal number of clusters, ensuring meaningful segmentation.
* **Cluster Analysis**: Interpret the resulting clusters to understand the characteristics of each segment. Analyze the distribution of variables within clusters to describe distinct consumer profiles, such as “price-sensitive frequent buyers” or “brand-loyal consumers.”
* **Evaluation**: Assess cluster stability and coherence to ensure interpretability and actionable results. Evaluate clusters based on metrics like intra-cluster distance (for compactness) and inter-cluster distance (for separation).

**3. Classification and Prediction Models (Supervised)**

After defining segments through clustering, build supervised models to classify and predict specific consumer traits within those segments:

* **Value-Conscious Classification**:
  + **Feature Selection**: Use features like discount-purchase frequency, bulk buying behavior, and sensitivity to promotional deals to classify consumers as value-conscious or not.
  + **Model Selection**: Apply **classification algorithms** like **Logistic Regression** or **Decision Trees**. Logistic Regression is interpretable and useful for binary outcomes, while Decision Trees provide intuitive rule-based segmentation.
  + **Evaluation**: Evaluate the classification model’s accuracy, precision, and recall, ensuring it accurately identifies value-conscious consumers for targeted deal-based marketing.
* **Brand Loyalty Prediction (Brand Runs)**:
  + **Feature Selection**: Use historical brand purchase data, transaction frequency, and loyalty indicators to predict brand loyalty (i.e., repeated brand purchases).
  + **Model Selection**: Implement **Random Forest** for high predictive accuracy. These ensemble methods work well for complex patterns and can handle high-dimensional feature sets effectively.
  + **Evaluation**: Evaluate model performance using metrics such as recall, precision, and F1-score to ensure accurate predictions for brand-loyal consumers, enabling clients to implement loyalty programs for high-retention segments.

**4. Model Evaluation and Interpretation**

Once clustering and predictive models are built, comprehensive evaluation and interpretation are necessary to validate and explain results:

* **Cross-Validation**: Use cross-validation techniques to ensure model robustness and prevent overfitting. This is particularly important for supervised models to verify generalization on unseen data.
* **Interpretation**: Examine model outputs, such as coefficients in Logistic Regression or feature importances in Decision Trees and Random Forests, to interpret which factors are most influential in predicting value-consciousness and brand loyalty.
* **Actionable Insights**: Summarize key insights from clusters and predictive models. Each segment’s defining characteristics and predicted behaviors will inform specific marketing strategies and reward program designs aligned with client needs.

This analytical approach ensures a systematic, data-driven process, from initial data understanding to the final interpretation, allowing clients to deploy targeted strategies that align with actual consumer behaviors and motivations.

# **Dataset Overview**

The dataset provided by AXANTEUS represents a sample of **600** consumers from the Thai urban market, selected through stratified sampling to ensure representativeness. This dataset is used for analyzing consumer purchase behavior across various product categories, focusing on demographic attributes, purchase patterns, brand loyalty, and responses to promotions. It contains both demographic data and transactional information for each consumer.

| **Group** | **Attribute** | **Description** |
| --- | --- | --- |
| **Demographics** | Member ID | Unique identifier for each consumer |
|  | Socioeconomic Class | Categorical variable (1 = High, 5 = Low) representing consumer's socioeconomic status |
|  | Eating Habits | Categorical variable indicating dietary habits (0 = Not Specified, 1 = Vegetarian, etc.) |
|  | Native Language | Encoded numeric value representing the native language of the consumer |
|  | Gender | Gender of the consumer (1 = Male, 2 = Female) |
|  | Age | Age category of the consumer |
|  | Education | Education level of the consumer (1 = Minimum, 9 = Maximum) |
|  | Household Size | Number of members in the consumer’s household |
|  | Children | Number of children in the household |
|  | Television Availability | Whether a TV is available in the household (1 = Available, 2 = Not Available) |
|  | Affluence Index | Computed score indicating the wealth based on the possession of durable goods |
| **Purchase Summary** | No. of Brands | Number of brands purchased by the consumer |
|  | Brand Runs | Number of consecutive purchases of the same brand (indicates brand loyalty) |
|  | Total Volume | Total volume of products purchased |
|  | No. of Transactions | Total number of purchase transactions |
|  | Value | Total value of all purchases |
|  | Transactions/Brand Run | Average number of transactions per brand run |
|  | Volume/Transaction | Average volume per transaction |
|  | Average Price | Average price of purchases |
| **Promotion Analysis** | Purchase Vol. (No Promo) % | Percentage of total purchase volume made without any promotion |
|  | Purchase Vol. (Promo Code 6) % | Percentage of total purchase volume under promotion code 6 |
|  | Purchase Vol. (Other Promo) % | Percentage of total purchase volume under other promotions |
| **Brand-wise Purchase** | Br. Code Variables | Percentage of purchase volume under specific brand codes |
| **Price Category** | Price Category 1-4 | Percentage of purchase volume under different price categories |
| **Selling Proportion** | Prop. Category 5-15 | Percentage of purchase volume under different product proposition categories |

*(Table 1 – Dataset Overview)*

**Key Components of the Dataset**

1. **Demographics:**
   * This segment captures the personal and socioeconomic information of the consumers. It includes attributes like **socioeconomic class**, **eating habits**, **gender**, and **education level**.
   * These variables provide foundational insights into the consumer profile, helping understand purchasing patterns based on personal and household characteristics.
2. **Purchase Summary:**
   * These variables represent overall purchasing trends, focusing on total purchase volume, value, brand diversity, and brand loyalty (via the number of brand runs).
   * For instance, variables like **Total Volume** and **No. of Transactions** provide insights into the **frequency and intensity of consumer purchases**, while **Average Price** reflects consumer spending behavior and price sensitivity.
3. **Promotion Analysis:**
   * This part focuses on the impact of promotions on purchase volume, with variables indicating whether purchases were made under specific promotional schemes or without any promotion.
   * Metrics like **Purchase Vol. (No Promo) %** and **Purchase Vol. (Promo Code 6) %** show **consumer susceptibility to promotions**, a crucial factor in designing marketing strategies.
4. **Brand-wise Purchase:**
   * This section offers detailed information on brand preferences, capturing the percentage of purchase volume under specific brand codes.
   * It helps assess **brand loyalty**, identifying which brands dominate the consumer’s purchase behavior and which have a relatively lower presence.
5. **Price Category:**
   * These attributes reflect the consumer's **price sensitivity** by analyzing the percentage of purchase volume within different price categories, from budget to premium.
   * It helps segment consumers into **value-oriented vs. premium-oriented** buyers, useful for tailored marketing strategies.
6. **Selling Proportion:**
   * This segment categorizes purchases based on different product proposition categories (e.g., convenience, luxury).
   * It further refines segmentation by highlighting consumer inclination towards certain product attributes, helping define marketing approaches that cater to distinct consumer needs.

## Dataset Composition

The dataset is composed of **46 variables**, which can be broadly grouped into three main analytical areas:

* **Consumer Demographics:** 10 variables
* **Purchase Behavior Summary:** 8 variables
* **Promotion, Brand, and Price Analysis:** 28 variables

**Data Type and Structure**

* All variables are numeric, with values ranging from categorical encodings (e.g., 1 for Male, 2 for Female) to continuous measures like **total purchase volume** and **affluence index**.
* The dataset includes attributes that enable both **descriptive analysis** (for understanding overall consumer behavior) and **predictive modeling** (for targeting specific consumer segments based on purchase habits).

**Summary Statistics of Dataset**

In this analysis, the **skimr** package was used to provide a comprehensive summary of the dataset. This tool gives an overview of both categorical and numerical variables, displaying key metrics such as the number of missing values, mean, standard deviation, and more. The skimr output helps in identifying potential issues with the data, such as missing values and skewed distributions, which are critical for data preprocessing and model development.

## **Dataset Composition**

* **Number of Records:** 600
* **Number of Variables:** 46
* **Variable Types:** All variables are numeric, covering demographic information, purchase behavior, brand loyalty, and promotional impacts.
* **Complete Data:** There are no missing values across any columns, indicating a well-maintained dataset, making it easier for analysis and modeling.

|  |
| --- |
| skim(consumer\_data)  ── Data Summary ────────────────────────  Values  Name consumer\_data  Number of rows 600  Number of columns 46  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Column type frequency:  numeric 46  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Group variables None  ── Variable type: numeric ──────────────────────────────────────────────────────────────────────────────  skim\_variable n\_missing complete\_rate mean sd p0 p25  1 Member.id 0 1 1104188. 45602. 1010010 1065295  2 SEC 0 1 2.5 1.12 1 1.75  3 FEH 0 1 2.05 1.13 0 1  4 MT 0 1 8.18 4.29 0 4  5 SEX 0 1 1.74 0.649 0 2  6 AGE 0 1 3.21 0.865 1 3  7 EDU 0 1 4.04 2.19 0 3  8 HS 0 1 4.19 2.30 0 3  9 CHILD 0 1 3.23 1.22 1 2  10 CS 0 1 0.932 0.507 0 1  11 Affluence.Index 0 1 17.0 11.4 0 10  12 No..of.Brands 0 1 3.64 1.58 1 2  13 Brand.Runs 0 1 15.8 10.4 1 8  14 Total.Volume 0 1 11915. 7770. 150 6825  15 No..of..Trans 0 1 31.2 17.4 1 22  16 Value 0 1 1337. 883. 20 790.  17 Trans...Brand.Runs 0 1 2.62 2.60 1 1.42  18 Vol.Tran 0 1 415. 249. 94.4 251.  19 Avg..Price 0 1 11.8 3.74 5.62 9.76  20 Pur.Vol.No.Promo.... 0 1 0.913 0.119 0 0.879  21 Pur.Vol.Promo.6.. 0 1 0.0535 0.0930 0 0  22 Pur.Vol.Other.Promo.. 0 1 0.0335 0.0720 0 0  23 Br..Cd..57..144 0 1 0.184 0.236 0 0  24 Br..Cd..55 0 1 0.129 0.260 0 0  25 Br..Cd..272 0 1 0.0332 0.0910 0 0  26 Br..Cd..286 0 1 0.0339 0.113 0 0  27 Br..Cd..24 0 1 0.0193 0.0798 0 0  28 Br..Cd..481 0 1 0.0259 0.0894 0 0  29 Br..Cd..352 0 1 0.0342 0.122 0 0  30 Br..Cd..5 0 1 0.0182 0.0679 0 0  31 Others.999 0 1 0.522 0.297 0 0.279  32 Pr.Cat.1 0 1 0.279 0.281 0 0.0567  33 Pr.Cat.2 0 1 0.493 0.312 0 0.208  34 Pr.Cat.3 0 1 0.139 0.268 0 0  35 Pr.Cat.4 0 1 0.0886 0.192 0 0  36 PropCat.5 0 1 0.457 0.316 0 0.160  37 PropCat.6 0 1 0.0923 0.166 0 0  38 PropCat.7 0 1 0.0969 0.196 0 0  39 PropCat.8 0 1 0.0801 0.153 0 0  40 PropCat.9 0 1 0.0308 0.0628 0 0  41 PropCat.10 0 1 0.0202 0.0767 0 0  42 PropCat.11 0 1 0.0294 0.0985 0 0  43 PropCat.12 0 1 0.00622 0.0263 0 0  44 PropCat.13 0 1 0.0249 0.0954 0 0  45 PropCat.14 0 1 0.136 0.266 0 0  46 PropCat.15 0 1 0.0254 0.0876 0 0  p50 p75 p100 hist  1 1106235 1148292. 1167670 ▂▅▃▃▇  2 2.5 3.25 4 ▇▇▁▇▇  3 3 3 3 ▂▃▁▁▇  4 10 10 19 ▂▃▇▁▁  5 2 2 2 ▁▁▁▁▇  6 3 4 4 ▁▃▁▅▇  7 4.5 5 9 ▃▁▇▂▁  8 4 5 15 ▅▇▁▁▁  9 4 4 5 ▂▅▂▇▂  10 1 1 2 ▂▁▇▁▁  11 15 24 53 ▅▇▃▂▁  12 3 5 9 ▅▇▂▂▁  13 15 21 74 ▇▆▁▁▁  14 10360 15344. 50895 ▇▆▂▁▁  15 28 40 138 ▇▆▁▁▁  16 1216 1676. 6372. ▇▆▁▁▁  17 1.85 2.69 23 ▇▁▁▁▁  18 362. 491. 2525 ▇▁▁▁▁  19 11.2 13.4 33.3 ▇▇▂▁▁  20 0.953 1 1 ▁▁▁▁▇  21 0 0.0682 0.667 ▇▁▁▁▁  22 0 0.0438 1 ▇▁▁▁▁  23 0.0803 0.282 1 ▇▂▁▁▁  24 0 0.0949 1 ▇▁▁▁▁  25 0 0.0191 0.964 ▇▁▁▁▁  26 0 0 1 ▇▁▁▁▁  27 0 0 1 ▇▁▁▁▁  28 0 0.00610 0.898 ▇▁▁▁▁  29 0 0 0.993 ▇▁▁▁▁  30 0 0.00953 0.971 ▇▁▁▁▁  31 0.525 0.778 1 ▆▇▇▇▇  32 0.181 0.420 1 ▇▃▂▁▁  33 0.525 0.748 1 ▇▅▆▇▆  34 0 0.123 1 ▇▁▁▁▁  35 0 0.0707 1 ▇▁▁▁▁  36 0.444 0.72 1 ▇▅▅▆▅  37 0.0202 0.104 0.971 ▇▁▁▁▁  38 0.0126 0.0829 1 ▇▁▁▁▁  39 0.00775 0.0863 0.964 ▇▁▁▁▁  40 0 0.0316 0.408 ▇▁▁▁▁  41 0 0 1 ▇▁▁▁▁  42 0 0.00774 0.898 ▇▁▁▁▁  43 0 0 0.333 ▇▁▁▁▁  44 0 0.00552 1 ▇▁▁▁▁  45 0 0.118 1 ▇▁▁▁▁  46 0 0 0.840 ▇▁▁▁▁ |

*(Table 2 – Summary Statistics R Result)*

**Key Variable Groups:**

* **Demographic Variables (10 variables):** These capture consumer characteristics like socioeconomic class, education level, household size, and affluence index.
* **Purchase Behavior Variables (8 variables):** These variables detail total purchase volume, the number of brands purchased, the number of transactions, and brand runs, which indicate brand loyalty.
* **Promotional and Price Impact Variables (28 variables):** These variables represent purchase behavior under different promotional conditions, price categories, and product proposition categories.

**Data Distribution Insights:**

* **Demographics:**
  + **Socioeconomic Class (SEC):** The average SEC is around 2.5, ranging from 1 (High) to 4 (Low), suggesting a diverse mix of consumer segments.
  + **Affluence Index:** It ranges from 0 to 53, with an average of 17, indicating varying levels of household affluence.
  + **Gender (SEX):** The mean of 1.74 shows a higher representation of females (coded as 2), indicating more female consumers in the sample.
  + **Education (EDU):** The average is 4.04, implying moderate education levels among the consumers.
* **Purchase Behavior:**
  + **Total Volume:** The mean total volume purchased is 11,915, with a maximum of 50,895, indicating a significant variance in consumer purchasing power.
  + **Brand Runs:** The average number of brand runs is 15.8, suggesting moderate brand loyalty, but with a maximum of 74, some consumers show a very high level of brand dedication.
  + **Average Price:** The average price paid is 11.8, indicating moderate spending behavior, while the distribution varies across consumers.
* **Promotion and Pricing:**
  + **Volume without Promotions:** The average percentage of purchase volume without promotions is 91.3%, indicating that most purchases are not influenced by promotions.
  + **Promotional Impact:** The **Purchase Vol. Promo 6%** averages only 5.35%, showing that promo code 6 has a low impact on overall purchase volume.
  + **Brand Codes & Propositions:** Variables like **Br. Cd. 57, 144** and **PropCat 5** show varying levels of purchase concentration, with some brands and propositions accounting for a significant proportion of purchase volume.

**Interesting Findings About the Dataset**

1. **Diverse Socioeconomic Representation:**
   * The dataset has a well-distributed mix of socioeconomic classes, enabling segmentation across different economic groups. This diversity allows for insights into how income levels impact purchase behavior and brand loyalty.
2. **Strong Focus on Non-Promotional Purchases:**
   * Most purchases are made without promotional influence, suggesting that pricing, brand loyalty, or necessity may be stronger drivers than promotions. This could be a potential area for companies to develop new promotional strategies.
3. **High Variability in Brand Loyalty:**
   * While the average brand runs indicate moderate loyalty, the high maximum value (74) suggests that some consumers are extremely brand loyal. This makes it possible to identify and target high-value customers with loyalty programs.
4. **Education and Affluence Influence:**
   * With a significant range in education levels and affluence indices, the dataset allows for deeper analysis of how educational attainment and household wealth affect consumer purchasing behavior and brand choices.
5. **Potential for Segmentation:**
   * The inclusion of variables related to demographics, purchase behavior, and promotional impact provides a comprehensive framework for segmentation. It allows for the identification of different consumer groups, ranging from value-conscious buyers to premium buyers.
6. **Comprehensive Promotional Insights:**
   * The dataset includes detailed information on how different promotional strategies (e.g., Promo Code 6, Other Promotions) affect consumer purchase behavior. This can help businesses design effective promotional strategies to enhance sales.
7. **Price Sensitivity:**
   * The dataset captures price sensitivity well, with detailed variables for different price categories. This allows for an in-depth analysis of how price impacts consumer choices, which is crucial for pricing strategy optimization.

**Conclusion**

This dataset provides a **rich, granular view of consumer behavior**, making it ideal for detailed **segmentation analysis, predictive modeling**, and strategic decision-making for marketing and promotions. The diversity in variables and completeness of data offer opportunities for precise and meaningful insights into consumer purchase dynamics.

# **Data Preprocessing**

**Removing Member\_id**

To ensure a streamlined dataset for our analysis, we have identified and removed the member\_id column from the **consumer\_data** dataset. This decision aligns with our goal of retaining only the variables relevant to the analysis.

**Column Renaming Details**

In the preprocessing step, we made specific changes to column names to enhance clarity, consistency, and ease of use in the analysis process. Here’s a summary:

1. **Converted to Lowercase**:
   * We converted all column names to lowercase for consistency and to avoid case sensitivity issues. Lowercase naming is a common practice in data analysis, ensuring column references are uniform.
2. **Replaced Periods with Underscores**:
   * Original column names contained periods (e.g., Affluence.Index, Total.Volume). We replaced these with underscores (\_) to create a consistent naming pattern. Underscores improve readability and are more compatible across various programming environments.
   * **Example**:
     + Affluence.Index was renamed to affluence\_index
     + Total.Volume became total\_volume
3. **Removed Extraneous Characters and Simplified Names (Renaming Columns)**:
   * Some columns had complex names with repeated or redundant elements (e.g., Pur.Vol.No.Promo....). These were simplified to more concise terms for clarity.
   * **Example**:
     + Pur.Vol.No.Promo.... was renamed to pur\_vol\_no\_promo
     + Trans...Brand.Runs became trans\_brand\_runs

**Purpose of Renaming**

This standardization:

* Reduces errors when referencing column names in code.
* Enhances readability, making it easier to interpret each variable's purpose.
* Ensures column names follow a structured format, which is especially useful when sharing the dataset with others or integrating it into larger analytical workflows.

Overall, renaming the columns improves data handling efficiency and aligns the dataset with best practices in data management.

# **Overview of Missing Values**

1. **Identify missing values:**
   * Used the colSums(is.na(consumer\_data)) function to check for any missing values across all columns.
2. **Analysis of missing values:**
   * As per the summary exploration (skim() results), there were **no missing values** in the dataset, indicating that it is clean in terms of missing data.

The analysis confirmed that **there are no missing values in any of the columns**. This is an excellent outcome, as it allows us to proceed confidently without having to handle missing data imputation or deletion, ensuring the dataset remains intact for further exploration and modeling.

**Addressing Zero Values in Demographic and Household Variables**

The dataset contains several demographic and household variables with zero values that are likely placeholders for missing or unspecified data, rather than actual values. To ensure accurate and meaningful analysis, targeted imputation strategies were applied to address these zero values. The following details the approach used for each variable.

| **Variable** | **Zero Count** | **Interpretation** |
| --- | --- | --- |
| sec | 0 | No missing data |
| feh | 69 | Likely indicates unspecified eating habits |
| mt | 69 | Represents a language code (0 is valid) |
| sex | 68 | Likely indicates unspecified gender |
| age | 0 | No missing data |
| edu | 73 | Likely indicates missing education level |
| hs | 68 | Likely indicates missing household size |
| child | 0 | No missing data; zero represents no children |
| cs | 99 | Likely indicates missing data for TV availability |
| affluence\_index | 69 | Likely indicates missing affluence information |

*(Table 3 – Missing in Demographic Data)*

**Variable-Specific Imputation Strategy**

1. **SEX (Gender of Consumer)**:
   * **Issue**: Zeros in the sex variable likely indicate unspecified gender rather than a valid category.
   * **Solution**: These zero values are treated as "not defined" to maintain the integrity of the dataset without altering the original distribution. This approach acknowledges that the zero values do not represent a known category and avoids assumptions about gender.
2. **EDU (Education Level of Consumer)**:
   * **Issue**: Zero values in EDU indicate missing data on education level, rather than representing the minimum level.
   * **Solution**: Zeros in this variable were replaced with the median education level calculated from non-zero values, which provides a balanced approximation for this ordinal variable. The median serves as a central, representative value that maintains the overall distribution of educational attainment among consumers.
3. **HS (Household Size)**:
   * **Issue**: Household size values of zero are unrealistic and likely indicate missing data.
   * **Solution**: Zeros were replaced with the median household size, derived from non-zero values. This imputation ensures that household size is represented by a typical value, allowing for a realistic view of household structures across the dataset.
4. **CS (Television Availability)**:
   * **Issue**: Zeros in CS (Television Availability) suggest missing data rather than an actual indication of availability.
   * **Solution**: Zeros in CS were imputed with the mode, assumed to be 1 (indicating "Available"). As the most common value, this imputation reflects the majority's status, ensuring consistency within the dataset for this demographic variable.
5. **Affluence Index**:
   * **Issue**: Zeros in Affluence Index likely represent missing data rather than an actual affluence score.
   * **Solution**: Zeros in this variable were replaced with the median affluence index, calculated from non-zero values. This imputation preserves the central tendency of the affluence index, providing a balanced view of consumer purchasing power.

**Conclusion**

The imputation of zero values has enhanced data completeness and integrity across critical demographic and household variables. By applying appropriate imputation methods—mode or median, depending on the context—each variable retains a realistic distribution aligned with its intended purpose in the dataset. This treatment supports the accuracy of subsequent analysis, ensuring that the data is both complete and representative.

# **Data Visualization**

Data visualization plays a crucial role in understanding the structure, relationships, and patterns within our dataset. By creating visual representations of key variables and their interactions, we gain insights into consumer behaviors, preferences, and segments, which are vital for aligning with our business goals of consumer segmentation and brand loyalty.

**Purpose of Visualization**

The primary goal of these visualizations is to explore the data in a way that reveals patterns that might not be immediately apparent from raw numbers. With these visual tools, we can uncover trends, identify outliers, and assess distributions, correlations, and group comparisons across various consumer demographics and purchasing behaviors.

**Visualization Strategy**

To achieve a comprehensive view, we created multiple plots that each target specific questions related to our business objectives:

1. **Consumer Demographics**:
   * **Age Distribution**: This plot helps us understand the age range of our consumers, showing which age groups are most represented. Age can be a strong predictor of purchasing habits and brand loyalty.

A graph of a number of people

Description automatically generated

*(Fig 1 – Age Distribution Plot)*

This "Age Group Distribution" plot highlights the distribution of consumers across different age groups:

* **Predominant Age Group**: Age group 4 has the highest representation with 287 individuals, indicating that this age segment may be the primary demographic in the dataset.
* **Balanced Middle-Aged Groups**: Age groups 2 and 3 also have significant counts, with 129 and 169 individuals, respectively, suggesting active participation across these groups as well.
* **Minimal Representation in Youngest Group**: Age group 1 has only 15 individuals, making it the least represented, which may indicate limited consumer engagement from this age segment.

This distribution helps in targeting age-related segments for customized marketing and consumer segmentation efforts.

* + **Gender Distribution**: By exploring the balance between genders, we can assess whether our dataset has a gender bias and tailor our insights accordingly.

A graph with blue and white bars

Description automatically generated

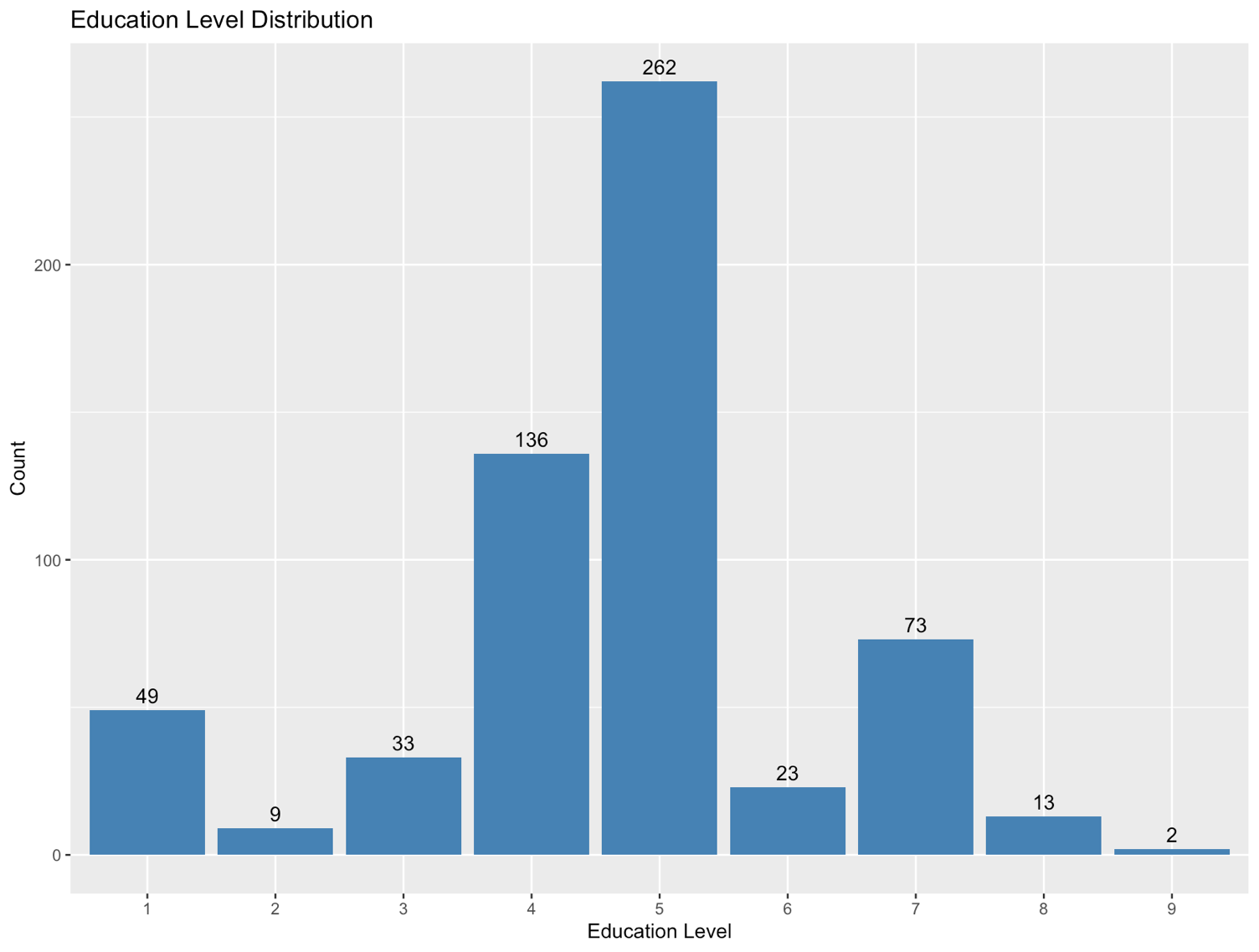
*(Fig 2 – Gender Distribution Plot)*

The **Gender Distribution** plot illustrates the composition of consumers by gender:

* **Dominant Gender**: Gender 2 is the most represented, with 511 individuals, making it the primary gender segment in the dataset.
* **Smaller Groups**: Gender 1 and Gender 0 have relatively low counts, at 68 and 21, respectively, suggesting these segments are less prominent.

This gender distribution indicates that any insights and strategies derived from the dataset may primarily reflect the preferences and behaviors of Gender 2. This imbalance should be considered when planning gender-targeted marketing or segmentation efforts.

* + **Education Level Distribution**: Understanding education levels provides insights into socio-economic backgrounds, which can be linked to affluence, spending behavior, and brand preferences.



*(Fig 3 – Education Distribution Plot)*

The **Education Level Distribution** plot provides an overview of the education levels in the dataset:

* **Most Common Level**: Education level 5 is the most represented, with 262 individuals, suggesting this is the typical education level among consumers.
* **Significant Mid-Level Representation**: Education levels 4 and 7 also have notable counts, at 136 and 73, respectively, indicating a sizable portion with moderate educational attainment.
* **Low Representation at Extremes**: Levels 2, 8, and 9 show minimal representation, suggesting limited consumer engagement from individuals with very low or very high education levels.

This distribution helps identify the education segments where marketing efforts may be most effective and offers insights for targeting strategies based on educational background.

1. **Spending and Affluence**:
   * **Affluence Index vs. Total Volume**: This scatter plot gives a sense of how affluence relates to purchasing volume. We expect affluent individuals to show a higher volume of purchases, potentially indicating loyalty or preference for certain brands.

A graph with blue dots

Description automatically generated

*(Fig 4 – Affluence Index vs. Total Volume Plot)*

The **Affluence Index vs. Total Volume** scatter plot shows the relationship between consumer affluence and their total purchase volume:

* **Moderate Positive Correlation**: Consumers with a mid-range Affluence Index tend to have higher purchase volumes, suggesting moderate affluence may drive larger purchases.
* **High Variability**: At higher affluence levels, the purchase volume becomes more varied, indicating that affluence does not consistently correlate with higher volume.
* **Targeting Insight**: Marketing strategies could focus on mid-affluence groups who are more likely to contribute to higher purchase volumes, while high-affluence groups might require a more personalized approach.

This plot is useful for understanding spending behavior relative to affluence, aligning with goals of targeting value-conscious consumers.

4o

* + **Number of Brands Purchased by Age**: This bar chart shows how brand diversity varies with age, giving insights into age-related brand exploration or loyalty.

A graph of blue squares and black lines

Description automatically generated

*(Fig 5 – No of Brands by Age Plot)*

The **Number of Brands Purchased by Age Group** box plot provides insights into brand diversity across different age groups:

* **Similar Medians Across Groups**: The median number of brands purchased is fairly consistent across age groups, suggesting stable brand engagement across ages.
* **Variation in Young and Older Groups**: Age groups 1 and 4 show greater variation in brand purchases, with some outliers, indicating that certain individuals in these groups tend to explore more brands.
* **Stable Engagement in Middle Groups**: Age groups 2 and 3 have narrower ranges, reflecting more uniform purchasing behavior.

This visualization supports the understanding of brand diversity preferences by age, which can inform targeted marketing strategies based on age-related brand loyalty.

1. **Transaction Behavior**:
   * **Average Price per Transaction by Age**: This plot reveals how much different age groups tend to spend on average per transaction. It can highlight age groups with a propensity for higher-value purchases.

A graph showing a number of blue and black bars

Description automatically generated with medium confidence

*(Fig 6 – Average Price per Transaction by Age Plot)*

The **Average Price per Transaction by Age Group** box plot highlights spending tendencies across different age groups:

* **Consistent Median Prices**: All age groups show similar median prices per transaction, indicating comparable spending patterns across age groups.
* **Presence of Outliers**: Age groups 2, 3, and 4 have several high-value outliers, suggesting that certain individuals in these groups occasionally make more expensive purchases.
* **Stable Price Range**: The overall range of average prices is relatively stable, with most transactions centered around a similar price range.

This analysis provides insight into the typical spending per transaction, useful for identifying high-value customer segments and tailoring pricing strategies.

4o

* + **Total Volume vs. Average Price**: This plot visualizes if there's a correlation between transaction volume and average price, helping identify if certain spending habits align with purchasing large quantities at lower prices or vice versa.

A graph showing the average price of a stock market

Description automatically generated

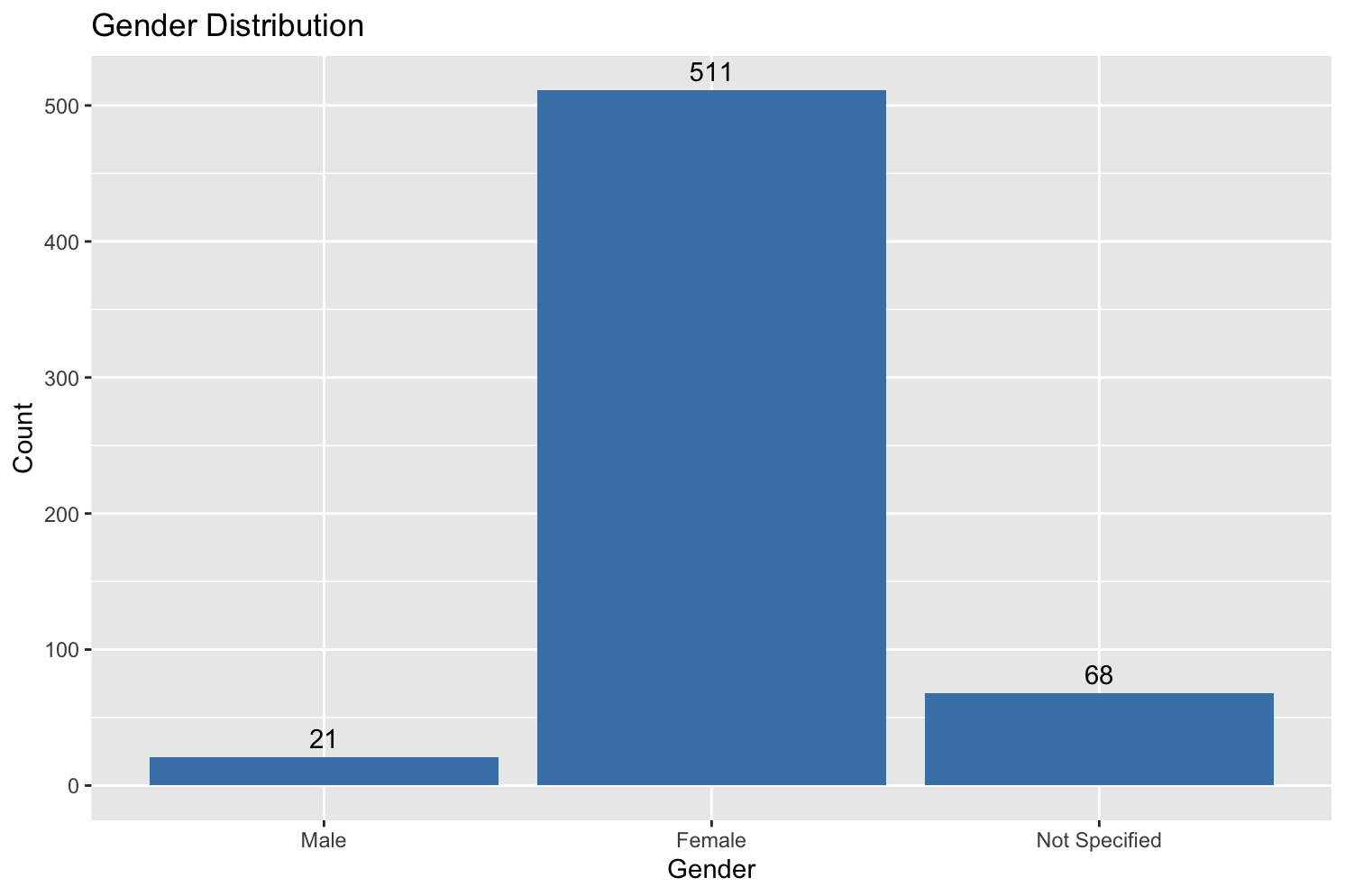
*(Fig 7 – Total Volume vs Average Price Plot)*

The **Total Volume vs. Average Price** scatter plot reveals insights into the relationship between the quantity of purchases and the average spending per transaction:

* **Inverse Relationship**: Higher total volumes generally correspond to lower average prices, suggesting that bulk buyers tend to spend less per unit on average.
* **Clustered Purchases**: Most transactions are concentrated around lower total volumes and prices, indicating that smaller, more frequent purchases are common among consumers.
* **Outliers in High Volume**: A few high-volume purchases have slightly higher average prices, possibly representing unique purchasing behaviors or premium segments.

This plot helps identify purchasing patterns, aiding in the segmentation of bulk buyers versus occasional high-spenders.

* + **Number of Transactions per Gender**: By comparing transaction counts across genders, this box plot sheds light on engagement levels and purchasing frequency for each gender.



*(Fig 8 – No of Transaction per Gender Plot)*

The **Number of Transactions per Gender** box plot provides insights into transaction behavior across gender groups:

* **Higher Transaction Median for Gender 1 and 2**: Gender groups 1 and 2 have a similar median number of transactions, indicating comparable engagement levels in terms of transaction frequency.
* **Outliers in Transaction Frequency**: There are significant outliers, particularly in Gender 2, suggesting a subset of highly active consumers within this group.
* **Lower Engagement for Gender 0**: Gender 0 has a lower median and fewer outliers, showing less frequent transaction behavior among this segment.

This plot aids in understanding the transaction intensity across genders, aligning with consumer segmentation efforts for targeted marketing strategies.

1. **Brand Loyalty and Preferences**:
   * **Brand Runs vs. Transaction Value**: This scatter plot helps explore if there's a relationship between frequent brand purchases (brand runs) and transaction values, providing insights into brand loyalty and high-value consumers.

A graph showing a diagram of a transaction value

Description automatically generated

*(Fig 9 – Brand Runs Vs TransactionPlot)*

The **Brand Runs vs. Transaction Value** scatter plot reveals insights about brand loyalty and spending:

* **Positive Correlation**: There is a general trend where higher brand runs (frequency of purchasing the same brand) are associated with higher transaction values, indicating that brand-loyal consumers tend to spend more.
* **High Transaction Outliers**: Some points show extremely high transaction values with fewer brand runs, suggesting occasional high-value purchases that do not indicate loyalty.
* **Clustered Lower Values**: The majority of consumers exhibit low to moderate brand runs with transaction values under 2000, suggesting regular but smaller transactions.

This plot helps in identifying high-value, brand-loyal customers, which is valuable for loyalty program strategies.

* + **Top Brands by Purchase Volume**: This bar plot of brand-specific volumes helps identify the most popular brands in terms of volume, which could guide marketing efforts and product focus.

A graph with blue bars

Description automatically generated

*(Fig 10 – Top Brands by Purchase Plot)*

The **Top Brands by Purchase Volume** bar chart highlights the leading brands in terms of purchase volume:

* **Dominant Brands**: br\_cd\_57\_144 and br\_cd\_55 stand out with the highest purchase volumes, indicating these brands are particularly popular among consumers.
* **Moderate Popularity**: Brands like br\_cd\_352 and br\_cd\_286 have mid-level purchase volumes, suggesting they appeal to a significant, but smaller, consumer base.
* **Least Purchased Brands**: br\_cd\_24 and br\_cd\_5 have the lowest purchase volumes, indicating limited preference or niche appeal.

This chart aids in identifying key brands for targeted marketing and inventory planning, focusing on high-volume brands to boost customer retention and sales.

1. **Promotion Usage**:
   * **Promotion Usage Distribution**: A polar chart (pie chart in polar coordinates) provides an intuitive view of promotion usage. It tells us the proportion of purchases made without promotions versus specific types of promotions, helping us understand consumer responsiveness to promotions.

A blue pie chart with a number of percentages

Description automatically generated

*(Fig 11 – Promotion Usage Dist Plot)*

The **Promotion Usage Distribution** pie chart provides insights into consumer reliance on promotions:

* **No Promotion Dominance**: A large majority of purchases (represented by the dark blue section) were made without any promotions, indicating that regular pricing may be sufficient for many consumers.
* **Limited Promo Engagement**: Promo\_6 and Other\_Promo constitute a small portion of the total volume, suggesting limited consumer response to promotional offers.

This distribution suggests that while promotions can attract certain customers, the primary sales are driven without promotional incentives, which may indicate strong brand loyalty or price insensitivity in the customer base.

* + **Volume Distribution across Different Promotion Types**: This bar chart illustrates the volume of purchases across promotion types, showing which promotions drive higher volumes, which is valuable for designing effective promotional strategies.

A graph with numbers and a bar

Description automatically generated

*(Fig 12 – Volume Distribution Across Promotion Plot)*

The **Volume Distribution across Promotion Types** bar chart highlights consumer behavior concerning promotional offers:

* **High Volume without Promotions**: A substantial majority of the purchase volume (547.80) is attributed to transactions made without any promotions, underscoring a strong base of consumers who are not reliant on discounts.
* **Limited Impact of Promotions**: Promo\_6 and Other\_Promo have relatively low volumes (32.10 and 20.09, respectively), suggesting that these promotions have minimal influence on driving higher purchase volumes.

This pattern indicates that promotions are not the primary drivers of sales volume, which may imply that consumers value the product's intrinsic value or are less price-sensitive.

1. **Household and Socioeconomic Factors**:
   * **Child Dependency and Affluence Index**: This box plot shows the relationship between having dependents (children) and affluence. It provides insights into consumer segments with different financial responsibilities, which may influence purchasing behavior.

A diagram of a number of children

Description automatically generated

*(Fig 13 – Child Dependency and Affluence Index Plot)*

The **Child Dependency and Affluence Index** box plot provides insights into the relationship between affluence and the number of children:

* **Consistent Affluence across Child Counts**: The median affluence index remains relatively stable across different numbers of children, suggesting that having more children does not significantly correlate with higher or lower affluence.
* **Slight Decrease with Higher Child Counts**: For families with five children, the affluence index appears lower and has less variation, although this may be due to a smaller sample size in this category.
* **Outliers Present**: There are outliers in lower child counts, indicating some variation in affluence within smaller families.

This plot helps in understanding the socio-economic segment of families with different child dependencies and can guide targeted strategies for family-oriented segments.

* + **Transaction Volume Distribution**: The histogram of transaction volume gives a sense of the overall spending habits within the dataset, showing if most consumers make low or high transaction volumes.

A graph of a distribution of a number of people

Description automatically generated with medium confidence

*(Fig 14 – Transaction Volume Distribution Plot)*

The **Transaction Volume Distribution** histogram provides insights into the spending habits in terms of transaction volume:

* **Right-Skewed Distribution**: Most of transactions fall below 20,000, with a peak frequency around the 5,000 to 10,000 range, indicating that most consumers tend to have moderate spending levels.
* **Few High-Volume Transactions**: There are some high-volume transactions beyond 30,000, but they are relatively rare, suggesting that only a small segment of consumers is responsible for large purchases.
* **Insight for Segmentation**: This distribution helps in identifying different spending segments, from low to high spenders, which can guide targeted marketing and customized offers.

This plot aligns with the goal of understanding consumer spending behaviors to enhance segmentation strategies.

**Volume Distribution by Promotion Category**

This analysis of volume distribution across different promotion categories highlights which promotion types drive higher consumer engagement and purchasing behavior, with notable variations in purchase volume across categories. Outliers indicate occasional high-volume transactions, especially in *pr\_cat\_1* and *pr\_cat\_2*, suggesting that these categories are particularly impactful.

A graph of a bar chart

Description automatically generated with medium confidence

*(Fig 15 – Volume Distribution by Promotion Category Plot)*

The box plot for each promotion category shows distinct patterns in purchase volume:

* **pr\_cat\_1** and **pr\_cat\_2** have wider interquartile ranges, indicating more variability in purchase volumes within these categories. This suggests that promotions in these categories drive both low and high levels of consumer engagement.
* **pr\_cat\_3** and **pr\_cat\_4** show more concentrated volume distributions with many outliers, especially on the lower end, implying that consumers engage with these promotions but tend to purchase smaller volumes.
* Outliers in each category, particularly in **pr\_cat\_1** and **pr\_cat\_2**, suggest occasional high-volume purchases, which may indicate the effectiveness of these promotions in attracting more substantial transactions.

This visualization provides insights into how different promotion categories impact purchasing behavior, which can help tailor promotional strategies.

**Heatmap of Brand Data (Log Scale)**

The heatmap reveals consumer purchase patterns across brands, showing that certain brands like *br\_cd\_57\_144* and *br\_cd\_55* dominate in terms of purchase volume. This insight assists in identifying key brands that attract higher engagement, guiding targeted brand-focused marketing efforts.

A white sheet with numbers and lines

Description automatically generated with medium confidence

*(Fig 16 – Heatmap of Brand Data (Log Scaled) Plot)*

This heatmap visualizes the distribution of purchase volumes across different brands on a log scale:

* The log scale highlights small variations in low-volume purchases, making patterns in brand engagement more visible.
* Brands like **br\_cd\_57\_144** and **br\_cd\_55** appear to have more consistent engagement across consumers, while other brands show sporadic engagement with fewer dark lines, indicating lower or less frequent purchases.
* The sparse engagement for brands such as **br\_cd\_24** and **br\_cd\_5** may indicate niche or less popular brands.

This heatmap helps in identifying high and low engagement brands, useful for brand-focused strategies.

# **Demographical Column Removal**

The columns **feh** (Eating Habit), **mt** (Native Language), **age**, **child**, and **cs** (Television Availability) were removed from the dataset as they are not directly relevant to our goals of clustering and segmentation based on consumer behavior, brand loyalty, and price sensitivity. These variables do not provide significant insight into purchasing patterns or promotional responsiveness, which are key for identifying actionable consumer segments. By focusing on variables that directly impact buying behavior, brand engagement, and socioeconomic status, the dataset becomes more streamlined, allowing for more meaningful analysis aligned with the business objectives.

# **Predictor Analysis and Relevancy**

**Correlation Analysis Overview**

The correlation analysis revealed several highly correlated variable pairs in the dataset, with correlations above a threshold of 0.6. Such high correlations indicate potential multicollinearity, which can complicate the interpretability of models by inflating the perceived significance of individual variables. Below is a selection of variable pairs with correlation values above 0.6, indicating significant relationships. While more pairs exhibit high correlations, this table highlights the top 15-20 pairs for clarity.

A diagram of a graph

Description automatically generated with medium confidence

*(Fig 17 – Correlation Analysis)*

## **Highly Correlated Variable Pairs**

| **No.** | **Variable 1** | **Variable 2** | **Correlation Value** |
| --- | --- | --- | --- |
| 1 | propcat\_14 | pr\_cat\_3 | 0.9973 |
| 2 | propcat\_14 | br\_cd\_55 | 0.9930 |
| 3 | pr\_cat\_3 | br\_cd\_55 | 0.9887 |
| 4 | propcat\_11 | br\_cd\_481 | 0.9464 |
| 5 | propcat\_13 | br\_cd\_24 | 0.9187 |
| 6 | value | total\_volume | 0.8764 |
| 7 | pur\_vol\_promo\_6 | pur\_vol\_no\_promo | -0.7984 |
| 8 | pr\_cat\_1 | avg\_price | 0.7861 |
| 9 | no\_of\_trans | brand\_runs | 0.7743 |
| 10 | propcat\_8 | br\_cd\_272 | 0.7432 |
| 11 | brand\_runs | no\_of\_brands | 0.6890 |
| 12 | pur\_vol\_other\_promo | pur\_vol\_no\_promo | -0.6286 |
| 13 | propcat\_7 | br\_cd\_352 | 0.6264 |
| 14 | vol\_tran | total\_volume | 0.6202 |
| 15 | br\_cd\_24 | propcat\_13 | 0.9187 |
| 16 | br\_cd\_481 | propcat\_11 | 0.9464 |
| 17 | br\_cd\_55 | propcat\_14 | 0.9929 |
| 18 | pr\_cat\_3 | propcat\_14 | 0.9973 |
| 19 | total\_volume | vol\_tran | 0.6202 |
| 20 | no\_of\_trans | brand\_runs | 0.7743 |

*(Table 4 – Highly Correlated Pairs)*

**Key Observations**

1. **prop\_cat\_14 and pr\_cat\_3 (0.9973)**:
   * These variables exhibit an almost perfect correlation, indicating they may capture the same information. Retaining both could introduce redundancy.
2. **prop\_cat\_14 and br\_cd\_55 (0.9930)**:
   * Strong correlation implies that as "prop\_cat\_14" increases, so does "br\_cd\_55," suggesting overlap.
3. **pr\_cat\_3 and br\_cd\_55 (0.9887)**:
   * High correlation implies these variables share significant information and may be redundant.
4. **prop\_cat\_11 and br\_cd\_481 (0.9464)**:
   * Similar purchasing behavior is likely captured by both variables, indicating potential redundancy.
5. **prop\_cat\_13 and br\_cd\_24 (0.9187)**:
   * These variables appear to represent similar purchasing patterns, suggesting one could be retained.
6. **value and total\_volume (0.8764)**:
   * The high correlation between total value and total volume of purchases is expected, as high purchase volumes typically yield high values.
7. **pur\_vol\_promo\_6 and pur\_vol\_no\_promo (-0.7984)**:
   * Negative correlation suggests that as purchases under promotion increase, non-promo purchases decrease, indicating an inverse relationship.
8. **pr\_cat\_1 and avg\_price (0.7861)**:
   * Higher average prices seem associated with "pr\_cat\_1," indicating a relationship between category and pricing behavior.
9. **no\_of\_trans and brand\_runs (0.7743)**:
   * High transaction frequency is linked to brand runs, suggesting overlap.
10. **prop\_cat\_8 and br\_cd\_272 (0.7432)**:

* Moderate correlation suggests related purchase behaviors in these categories.

1. **brand\_runs and no\_of\_brands (0.6890)**:

* Brand runs and the number of brands show some correlation, possibly indicating consumer loyalty.

1. **pur\_vol\_other\_promo and pur\_vol\_no\_promo (-0.6286)**:

* Negative correlation here may indicate a shift in purchasing focus between promoted and non-promoted volumes.

1. **prop\_cat\_7 and br\_cd\_352 (0.6264)**:

* Overlap in purchasing behavior is indicated by this moderate correlation.

1. **vol\_tran and total\_volume (0.6202)**:

* Expected relationship as the volume per transaction links to total volume.

**Conclusion**

This subset highlights key correlated variable pairs in the dataset. Retaining all highly correlated variables may introduce redundancy, so selecting only one from each highly correlated pair could improve model performance and interpretability. Further consideration is needed to identify which variables provide unique insights for clustering and segmentation aligned with the business goals.

# **Dimension Reduction & Data Transformation**

The data transformation process plays a critical role in preparing the dataset for meaningful clustering, especially when analyzing consumer behaviors and purchase motivations. The primary objective of these transformations is to create derived variables that capture relevant patterns in purchasing behavior and preferences, while also managing skewed data distributions to enhance the performance and interpretability of the clustering algorithms.

The transformations and derived features created are as follows:

1. **Log Transformations for Skewed Variables**
   * **Log Total Purchase Volume (log\_total\_purchase\_volume)**:

The total\_volume variable represents the total quantity of purchases made by each consumer. To address the skewed nature of this variable—where a small number of consumers may have exceptionally high purchase volumes—a log transformation is applied using log1p, which computes the natural logarithm of (1 + value). This transformation reduces the impact of outliers and makes the variable more suitable for clustering.

* + **Log Average Spend per Transaction (log\_avg\_spend\_per\_transaction)**: This derived variable captures the average spending per transaction for each consumer. It is calculated by dividing the total value (representing the monetary value of purchases) by no\_of\_trans (the number of transactions). Applying a log transformation to this variable smooth out large differences in spending, thereby standardizing the scale of this feature across consumers with varying transaction patterns.

1. **Derived Variables**
   * **Brand Loyalty Score (brand\_loyalty\_score)**: This variable is designed to quantify the extent of brand loyalty exhibited by consumers. It is calculated as the ratio of brand\_runs (the frequency of consecutive purchases of the same brand) to the total number of unique brands purchased (no\_of\_brands + 1). Adding 1 to the denominator avoids division by zero. A higher brand loyalty score indicates a stronger preference for repeat purchases within the same brand.
   * **Deal Sensitivity (deal\_sensitivity)**: Deal sensitivity measures a consumer’s responsiveness to promotions. It is computed as the proportion of purchase volume made under promotions (pur\_vol\_promo\_6 and pur\_vol\_other\_promo) to the total purchase volume (total\_volume). A higher deal sensitivity suggests that the consumer is more inclined to take advantage of promotions.
   * **Category Diversity (category\_diversity)**: This variable captures the variety of product categories a consumer engages with. It is calculated by summing up the number of unique product categories purchased, represented by columns that start with "pr\_cat" and "propcat". A higher category diversity score indicates broader engagement across different types of products.

**Summary Table**

| **Variable** | **Description** | **Transformation** |
| --- | --- | --- |
| log\_total\_purchase\_volume | Total volume of purchases, log-transformed to reduce skewness. | log1p |
| log\_avg\_spend\_per\_transaction | Average spending per transaction, log-transformed to smooth out differences. | log1p |
| brand\_loyalty\_score | Ratio of consecutive brand purchases to the number of unique brands purchased, indicating brand loyalty. | Derived |
| deal\_sensitivity | Proportion of purchase volume made under promotions, indicating responsiveness to deals. | Derived |
| category\_diversity | Count of unique product categories purchased, reflecting product engagement diversity. | Derived |

*(Table 5 –Introducing derived Variabels)*

1. **Selection of Relevant Columns for Clustering Segments** After creating the derived variables, the dataset is segmented into three distinct clusters based on different aspects of consumer behavior and motivations. Each cluster is represented by a subset of features:
   * **Purchase Behavior Cluster (behavior\_data)**: This cluster focuses on metrics that capture a consumer’s purchasing intensity and brand loyalty. The selected variables are:
     + log\_total\_purchase\_volume: Represents total purchase activity on a logarithmic scale.
     + brand\_loyalty\_score: Quantifies loyalty toward specific brands.
     + no\_of\_trans: Total number of transactions made by the consumer.
   * **Purchase Basis Cluster (basis\_data)**: This cluster emphasizes deal sensitivity and the diversity of product categories, providing insights into a consumer’s motivation for purchases. The selected variables are:
     + deal\_sensitivity: Measures responsiveness to promotions.
     + category\_diversity: Captures the variety of product categories engaged with by the consumer.
   * **Combined Cluster (combined\_data)**: This cluster incorporates both behavioral metrics and motivational factors to provide a comprehensive view of each consumer. The selected variables are:
     + log\_total\_purchase\_volume
     + brand\_loyalty\_score
     + no\_of\_trans
     + deal\_sensitivity
     + category\_diversity

**Summary of Variable selection for clusters:**

| **Cluster Segment** | **Feature** | **Description** |
| --- | --- | --- |
| **Purchase Behavior Cluster** | log\_total\_purchase\_volume | Represents total purchase activity on a logarithmic scale. |
|  | brand\_loyalty\_score | Quantifies loyalty toward specific brands. |
|  | no\_of\_trans | Total number of transactions made by the consumer. |
| **Purchase Basis Cluster** | deal\_sensitivity | Measures responsiveness to promotions. |
|  | category\_diversity | Captures the variety of product categories engaged by the consumer. |
| **Combined Cluster** | log\_total\_purchase\_volume | Represents total purchase activity on a logarithmic scale. |
|  | brand\_loyalty\_score | Quantifies loyalty toward specific brands. |
|  | no\_of\_trans | Total number of transactions made by the consumer. |
|  | deal\_sensitivity | Measures responsiveness to promotions. |
|  | category\_diversity | Captures the variety of product categories engaged by the consumer. |

*(Table 6 – Summary of variable selection for clustering)*

This table provides a clear summary of the selected variables within each cluster segment, emphasizing the consumer behaviors and motivations captured by each feature.

The clustering segments are derived to enable focused insights based on consumer behavior and motivations:

* **Purchase Behavior Segment**: This segment captures high-level purchase patterns such as transaction frequency and loyalty to specific brands. Consumers within this cluster can be targeted for brand-focused strategies, as their loyalty metrics indicate repeated purchases.
* **Purchase Basis Segment**: This segment provides insights into the motivational aspects of purchases. Consumers here exhibit patterns related to sensitivity toward deals and a preference for a diverse range of categories. This cluster can inform promotion-driven marketing strategies.
* **Combined Segment**: By integrating both behavior and motivational factors, this segment provides a holistic view of consumer actions, supporting comprehensive marketing strategies that address both purchase intensity and preferences.

These transformations and clusters enable a deeper understanding of consumer segmentation, providing AXANTEUS with actionable insights for targeted marketing and loyalty program development. Each variable included in the clustering process contributes uniquely to identifying consumer traits, ensuring that the resulting clusters reflect meaningful distinctions in purchase behavior and motivations.

# **Model Selection for Clustering**

The model selection for clustering aimed to identify the optimal number of clusters to segment consumers meaningfully. To achieve this, several methods were applied to determine the ideal number of clusters, specifically for K-Means clustering.

**Clustering Approach**

**K-Means clustering** was chosen as the primary clustering method. K-Means assigns each observation to the nearest cluster center based on Euclidean distance, iteratively refining cluster centers to minimize within-cluster variance. The simplicity and interpretability of K-Means make it suitable for segmenting consumers based on purchase behavior and motivations.

**Determining the Optimal Number of Clusters**

To identify the optimal cluster count, three methods were used:

* **Elbow Method**: The Elbow method plots the within-cluster sum of squares (WCSS) against the number of clusters. The point where the decrease in WCSS slows (forming an "elbow") suggests an optimal cluster count, balancing compact clusters with minimized information loss.
* **Silhouette Analysis**: This analysis measures how similar each observation is to its assigned cluster compared to other clusters. Average silhouette scores were examined across different cluster counts, with higher scores indicating better-defined clusters.

Based on these methods, three clusters were chosen as the optimal configuration for segmenting consumers in alignment with business goals. This approach allowed for a balance between interpretability and the differentiation needed for targeted marketing strategies.

## **Elbow Method**

The elbow method is used to determine the optimal number of clusters by observing the within-cluster sum of squares (WSS) as the number of clusters increases. By plotting WSS against various numbers of clusters, a noticeable "elbow" point typically appears where the rate of decrease sharply slows, indicating a good balance between compact clusters and minimizing the number of clusters.

In this analysis, the elbow method was applied separately to each dataset:

1. **Purchase Behavior Clustering Dataset**: Based on the elbow method plot for the purchase behavior dataset, an optimal cluster value was identified.
2. **Basis for Purchase Clustering Dataset**: For the basis dataset, the plot similarly displayed an elbow point, aiding in the selection of clusters that balance compactness with interpretability.
3. **Combined Clustering Dataset**: Finally, the elbow method for the combined dataset indicated an optimal cluster count, aligning with both the clustering structure and business objectives.

|  |  |  |
| --- | --- | --- |
| **A graph of a number of clusters  Description automatically generated** | **A graph of a number of clusters  Description automatically generated** | A graph of a number of clusters  Description automatically generated |

*(Fig 18 – Elbow Plot for Optimal No. of Cluster)*

The elbow method thus provided a systematic approach for cluster selection, ensuring a balance between compact clustering and interpretability across all datasets. The chosen clusters serve as the foundation for subsequent segmentation, aligning well with the overall business goals.

## **Silhouette Plot**

To determine the optimal number of clusters, a silhouette plot was used. The silhouette plot method evaluates how well-separated clusters are by calculating a silhouette score for each observation. This score represents the degree of similarity between an observation and the rest of its cluster relative to the other clusters. By plotting silhouette scores across different cluster counts, the optimal number of clusters can be identified as the configuration with the highest average silhouette score.

Here's the summary of silhouette analysis across our datasets:

1. **Purchase Behavior Clustering**: For this dataset, the silhouette plot was generated across a range of clusters to find the optimal number. A peak in the average silhouette score was observed at 3 clusters, suggesting that 3 clusters offer a good balance of cohesion and separation.
2. **Basis for Purchase Clustering**: Similarly, silhouette analysis for the basis dataset indicated that 3 clusters yield the highest average silhouette score, supporting a structure where clusters are well-defined around characteristics like deal sensitivity and category diversity.
3. **Combined Clustering**: The silhouette plot for the combined dataset also indicated that 3 clusters achieved the highest average silhouette score. Although slightly lower in score compared to the basis dataset, this configuration allowed for a balance of purchase behavior and basis characteristics.

|  |  |  |
| --- | --- | --- |
| A graph with lines and dots  Description automatically generated | A graph with numbers and lines  Description automatically generated | A graph with lines and dots  Description automatically generated |

*(Fig 19 – Silhouette Plot for Optimal No. of Cluster)*

Overall, silhouette analysis confirmed that using 3 clusters was optimal across all datasets, reinforcing a consistent structure suitable for effective segmentation in further analysis.

# **K-Means Clustering: Approach and Business Goals**

**Business Goal Context:** The objective of the consumer segmentation project is to create distinct consumer profiles that capture both purchasing behaviors and motivations. By achieving this segmentation, we aim to provide AXANTEUS with actionable insights that can guide targeted marketing, personalized promotions, and loyalty programs.

**Why K-Means Clustering?** K-means clustering was chosen for its simplicity, efficiency, and interpretability in segmenting consumer data. This method is well-suited for our business goal as it helps group consumers into distinct clusters based on their purchasing intensity, loyalty patterns, and responsiveness to promotions. These groupings align with AXANTEUS’s need for segments that reflect varying consumer motivations and behaviors.

* + - 1. **Purchase Behavior Clustering**

The goal of the Purchase Behavior Clustering is to group consumers based on their purchasing intensity and brand loyalty. This clustering provides insights into which consumers are frequent shoppers, exhibit brand loyalty, or have a lower level of purchasing activity. These insights align with AXANTEUS's business goal of identifying distinct consumer segments for targeted marketing.

Each cluster represents a distinct consumer behavior profile:

1. **Cluster 1: Moderate Buyers**
   * **Profile**: Consumers in this group have a moderate number of transactions and purchase volume. They exhibit balanced brand loyalty without extreme loyalty or frequency.
   * **Silhouette Score**: 0.52, indicating a fair level of separation from other clusters, suggesting moderate clustering quality.
2. **Cluster 2: High-Frequency Shoppers**
   * **Profile**: This cluster represents frequent shoppers with a high number of transactions and a relatively higher brand loyalty score. These consumers are likely engaged and respond well to loyalty programs or incentives to maintain brand loyalty.
   * **Silhouette Score**: 0.49, slightly lower than Cluster 1, indicating acceptable separation but some overlap with neighboring clusters.
3. **Cluster 3: Low-Engagement Shoppers**
   * **Profile**: Consumers in this cluster exhibit low engagement, with fewer transactions and lower overall purchasing volume. They are less brand-loyal and may require different engagement strategies, such as targeted promotions.
   * **Silhouette Score**: 0.46, the lowest among the clusters, suggesting a closer proximity to other clusters and indicating a weaker separation.

A graph of a graph showing different colors

Description automatically generated with medium confidence

*(Fig 20 – Silhouette plot for Purchase Behavior)*

**Silhouette Analysis**

The overall average silhouette score for the Purchase Behavior Clustering is **0.499**, which reflects moderate clustering quality. A score around 0.5 suggests that the clusters are reasonably well-separated but not ideal, indicating some overlap between consumer behaviors in adjacent clusters.

**Interpretation and Insights:**

* **Moderate Buyers** (Cluster 1) represent consumers who engage consistently but without high frequency or loyalty, making them a reliable but non-premium segment.
* **High-Frequency Shoppers** (Cluster 2) are the most engaged group, likely responsive to loyalty and reward programs, making them an ideal target for retention efforts.
* **Low-Engagement Shoppers** (Cluster 3) represent a segment with the lowest interaction, suggesting a need for targeted campaigns to increase their activity or to retain them as occasional shoppers.

This segmentation provides AXANTEUS with actionable insights for tailoring marketing strategies according to consumer behavior profiles, helping to effectively target high-frequency shoppers, maintain moderate buyers, and re-engage low-engagement shoppers.

**Cluster Centers Plot for Purchase Behavior Clustering**

The **Cluster Centers Plot** for **Purchase Behavior Clustering** displays the average values of key features across the three identified clusters. These features are:

* **Brand Loyalty Score**: Indicates how loyal a consumer is to the brand.
* **Log-Transformed Total Purchase Volume**: Represents the overall spending of consumers, transformed to handle skewed data.
* **Number of Transactions**: Represents the frequency with which a consumer makes purchases.

A graph of a bar chart

Description automatically generated with medium confidence

*(Fig 20 – Cluster Center plot for Purchase Behavior)*

The plot shows how each cluster behaves with respect to these features, providing insights into their consumer profiles:

1. **Cluster 1: Moderate Buyers**
   * **Brand Loyalty Score**: Moderate, suggesting that consumers in this cluster have a balanced level of brand loyalty.
   * **Log-Transformed Total Purchase Volume**: Moderate, indicating that these consumers are consistent in their spending but not extreme in either direction.
   * **Number of Transactions**: Moderate, reflecting that these consumers make purchases consistently but are not high-frequency shoppers.
2. **Cluster 2: High-Frequency Shoppers**
   * **Brand Loyalty Score**: High, indicating that consumers in this cluster are highly loyal to the brand.
   * **Log-Transformed Total Purchase Volume**: High, suggesting that these consumers make larger overall purchases compared to others.
   * **Number of Transactions**: Very high, reflecting that this group consists of frequent shoppers who engage with the brand regularly.
3. **Cluster 3: Low-Engagement Shoppers**
   * **Brand Loyalty Score**: Low, indicating minimal brand loyalty among these consumers.
   * **Log-Transformed Total Purchase Volume**: Low, showing that these consumers have low overall spending.
   * **Number of Transactions**: Low, suggesting that this group is not very active in terms of purchase frequency.

The plot visually distinguishes the three clusters based on their purchasing behavior. **Cluster 1** (Moderate Buyers) shows a balanced behavior across all features, **Cluster 2** (High-Frequency Shoppers) exhibits high engagement in all aspects, and **Cluster 3** (Low-Engagement Shoppers) is characterized by low engagement and minimal loyalty. These insights can help AXANTEUS develop targeted marketing strategies for each group, focusing on loyalty programs for high-frequency shoppers, retention strategies for moderate buyers, and re-engagement efforts for low-engagement shoppers.

* + - 1. **Purchase Basis Clustering**

The **Basis for Purchase Clustering** aims to segment consumers based on their responsiveness to deals and the diversity of products they purchase. This segmentation provides insights into consumer motivations, including how sensitive they are to promotions and their interest in purchasing a variety of product categories. These insights help AXANTEUS tailor marketing strategies to better meet the needs of distinct consumer groups.

Using the K-means algorithm, the Basis for Purchase Clustering was performed with **3 clusters**, selected based on optimal business segmentation needs and confirmed by silhouette analysis.

Each cluster represents a distinct motivation profile:

1. **Cluster 1: Promotion-Sensitive, Narrow Buyers**
   * **Profile**: Consumers in this cluster are highly responsive to deals but engage with a limited range of product categories. They are likely price-conscious and influenced by discounts but may have focused buying preferences.
   * **Silhouette Score**: **0.50**, indicating moderate separation from other clusters and reasonable clustering quality.
2. **Cluster 2: Balanced, Diverse Shoppers**
   * **Profile**: These consumers display a balanced approach, with moderate deal sensitivity and a high diversity in product categories. They are moderately influenced by promotions but are open to exploring different products.
   * **Silhouette Score**: **0.48**, reflecting moderate internal cohesion and reasonable distinctiveness from other clusters.
3. **Cluster 3: Variety-Seeking, Deal-Resistant Shoppers**
   * **Profile**: Consumers in this group show high category diversity but lower sensitivity to promotions. They are less influenced by discounts but prefer a wide variety of products, suggesting a preference-based shopping behavior rather than deal-driven choices.
   * **Silhouette Score**: **0.45**, the lowest among the clusters, yet still indicating reasonable separation and cluster validity.

**Silhouette Analysis**: The overall average silhouette score for the Basis for Purchase Clustering is **0.49**, suggesting that the clusters have moderate to reasonable internal consistency with decent separation from other clusters.

A graph of a plot

Description automatically generated with medium confidence

*(Fig 21 – Silhouette plot for Basis for Purchase)*

**Interpretation and Insights:**

* **Promotion-Sensitive, Narrow Buyers (Cluster 1)**: These consumers are highly responsive to discounts and promotions but have limited diversity in their product purchases. Marketing strategies for this segment could focus on targeted promotions, especially in specific product categories, where they are likely to be more influenced by deals.
* **Balanced, Diverse Shoppers (Cluster 2)**: These consumers are moderately responsive to promotions and show a high diversity in the products they purchase. They could be a prime target for variety-oriented promotions or loyalty programs that offer a diverse range of products.
* **Variety-Seeking, Deal-Resistant Shoppers (Cluster 3)**: These consumers show less sensitivity to promotions but are highly interested in a broad variety of products. Marketing strategies targeting this segment could emphasize product quality and variety over discounts, as these consumers may be more motivated by their preferences for a wide selection of products.

This segmentation provides **AXANTEUS** with actionable insights to develop **motivation-based marketing strategies**, enabling them to target specific consumer needs with tailored promotional offers and more effective engagement.

**Cluster Centers Plot for Basis for Purchase Clustering**

The **Cluster Centers Plot** for **Basis for Purchase Clustering** displays the average values of key features across the three identified clusters. These features are:

* **Category Diversity**: Represents how varied the product categories are that consumers purchase from.
* **Deal Sensitivity**: Indicates the responsiveness of consumers to promotional offers.
* **Number of Transactions**: Represents how frequently a consumer makes purchases.

A graph with blue and red bars

Description automatically generated

*(Fig 21 – Center plot for Basis for Purchase)*

The plot shows how each cluster behaves with respect to these features, providing insights into their consumer profiles:

1. **Cluster 1: Promotion-Sensitive, Narrow Buyers**
   * **Category Diversity**: Low, indicating that consumers in this cluster tend to purchase from a limited number of product categories.
   * **Deal Sensitivity**: High, reflecting that these consumers are highly responsive to promotions and discounts.
   * **Number of Transactions**: Low, suggesting that despite being highly deal-sensitive, their purchase frequency is relatively low.
2. **Cluster 2: Balanced, Diverse Shoppers**
   * **Category Diversity**: High, suggesting these consumers engage with a wide range of products.
   * **Deal Sensitivity**: Moderate, reflecting that while they are responsive to promotions, they are not as highly influenced by them as Cluster 1.
   * **Number of Transactions**: Moderate, indicating a balanced level of engagement with the brand.
3. **Cluster 3: Variety-Seeking, Deal-Resistant Shoppers**
   * **Category Diversity**: Very high, indicating that these consumers engage with a broad range of product categories.
   * **Deal Sensitivity**: Low, reflecting that these consumers are less likely to be influenced by discounts or promotions.
   * **Number of Transactions**: Low, suggesting that although they prefer variety, their transaction frequency is lower compared to the other clusters.

The plot visually distinguishes the three clusters based on their purchasing behavior. **Cluster 1 (Promotion-Sensitive, Narrow Buyers)** shows high deal sensitivity but limited engagement with a small set of product categories. **Cluster 2 (Balanced, Diverse Shoppers)** engages with a wide variety of products and is moderately influenced by deals. **Cluster 3 (Variety-Seeking, Deal-Resistant Shoppers)** has the highest diversity in product categories but is less influenced by deals, with lower overall engagement.

* + - 1. **Combine Clustering (Purchase behavior & basis)**

The Combined Clustering integrates metrics from both the Purchase Behavior and Basis for Purchase Clustering segments. This approach provides a comprehensive view of consumer profiles by capturing both their purchasing behaviors and motivations, aligning with AXANTEUS's goal of delivering a full-spectrum consumer segmentation for refined marketing strategies.

The Combined Clustering uses 3 clusters, as determined to be optimal for a balanced segmentation approach. The silhouette analysis was conducted to evaluate the consistency and separation of these clusters.

Each cluster provides a composite view based on both purchase behavior and purchase basis:

1. **Cluster 1: Balanced, High Engagement Consumers**
   * **Profile**: These consumers show a steady purchasing pattern with moderate brand loyalty and high diversity in product categories. They balance deal sensitivity with diverse product exploration.
   * **Silhouette Score**: 0.49, indicating average cluster coherence and separation from other clusters.
2. **Cluster 2: Promotion-Driven, Frequent Shoppers**
   * **Profile**: Consumers in this cluster are highly responsive to deals and have high transaction frequency, though with slightly lower diversity in product categories compared to other clusters. They may respond well to discount-driven marketing strategies.
   * **Silhouette Score**: 0.47, reflecting moderate internal cohesion but some overlap with other clusters.
3. **Cluster 3: Variety-Seeking, Low Engagement Shoppers**
   * **Profile**: This segment represents consumers with low transaction frequency but a high diversity in product choices. They are less driven by promotions and make selective purchases across a broad range of products.
   * **Silhouette Score**: 0.44, the lowest among the clusters, suggesting some internal variation, but still provides useful insights for variety-focused marketing.

**Silhouette Analysis:** The overall average silhouette score for the Combined Clustering is **0.474**, which suggests moderate clustering quality with some overlapping boundaries. This score is lower than the previous two clustering models, likely due to the complexity added by combining multiple behavior and basis variables.

A graph of a number of different colors

Description automatically generated with medium confidence

*(Fig 22 – Silhouette plot for Combined)*

**Interpretation and Insights:**

* **Balanced, High Engagement Consumers** (Cluster 1) represent those who engage consistently in purchasing and explore various product categories. This segment may respond well to a mix of promotions and new product introductions.
* **Promotion-Driven, Frequent Shoppers** (Cluster 2) are highly responsive to deals and have frequent purchases. Discount and promotion-oriented campaigns are likely to resonate strongly with this segment.
* **Variety-Seeking, Low Engagement Shoppers** (Cluster 3) have a wide array of interests but low purchase frequency, suggesting a preference for quality and selective buying. Marketing strategies should emphasize unique offerings or product quality for this group.

This Combined Clustering model provides AXANTEUS with a detailed segmentation that accounts for both purchase behavior and motivation, enhancing the precision of targeted marketing strategies and promotional efforts across different consumer types.

**Cluster Centers Plot for Combined Clustering (Purchase Behavior & Basis)**

The **Cluster Centers Plot** for **Combined Clustering** integrates key features from both the **Purchase Behavior Clustering** and the **Basis for Purchase Clustering**, providing a comprehensive view of consumer profiles. The features are:

* **Brand Loyalty Score**: Measures consumer loyalty to the brand.
* **Category Diversity**: Represents how varied the product categories are that consumers purchase from.
* **Deal Sensitivity**: Indicates how sensitive consumers are to promotions.
* **Log-Transformed Total Purchase Volume**: Represents overall spending, transformed for better interpretation.
* **Number of Transactions**: Reflects how often a consumer makes purchases.

A graph with different colored bars

Description automatically generated

*(Fig 22 –Cluster Center plot for Combined)*

The plot illustrates the behavior of each cluster in relation to these features:

1. **Cluster 1: Balanced, High Engagement Consumers**
   * **Brand Loyalty Score**: Moderate, indicating a balanced level of brand loyalty.
   * **Category Diversity**: High, showing these consumers explore a wide variety of products.
   * **Deal Sensitivity**: Moderate, reflecting that they are somewhat responsive to deals.
   * **Log-Transformed Total Purchase Volume**: Moderate, indicating steady overall spending.
   * **Number of Transactions**: High, suggesting frequent purchases.
2. **Cluster 2: Promotion-Driven, Frequent Shoppers**
   * **Brand Loyalty Score**: Moderate to low, showing a slight preference for brand loyalty but still deal-driven.
   * **Category Diversity**: Low, indicating limited product variety in purchases.
   * **Deal Sensitivity**: High, meaning these consumers are highly responsive to promotions.
   * **Log-Transformed Total Purchase Volume**: Moderate, showing steady spending.
   * **Number of Transactions**: Very high, indicating frequent and consistent shopping behavior.
3. **Cluster 3: Variety-Seeking, Low Engagement Shoppers**
   * **Brand Loyalty Score**: Low, showing minimal brand loyalty.
   * **Category Diversity**: Very high, as these consumers are interested in a broad range of products.
   * **Deal Sensitivity**: Low, indicating that they are less responsive to deals.
   * **Log-Transformed Total Purchase Volume**: Low, suggesting minimal overall spending.
   * **Number of Transactions**: Low, indicating low purchase frequency.

The **Cluster Centers Plot** highlights the distinct behavior of each cluster based on **Purchase Behavior** and **Basis for Purchase** features:

* **Cluster 1 (Balanced, High Engagement Consumers)** shows balanced loyalty, moderate spending, and frequent transactions, making them ideal for a mix of loyalty and promotional marketing strategies.
* **Cluster 2 (Promotion-Driven, Frequent Shoppers)** is highly responsive to promotions and frequently makes purchases, suggesting that discount-driven campaigns will resonate well with this segment.
* **Cluster 3 (Variety-Seeking, Low Engagement Shoppers)** values product variety over deals and engages less frequently, making them suitable for campaigns that emphasize unique offerings and product quality.

This clustering provides **AXANTEUS** with a detailed, multi-dimensional view of consumers, allowing for tailored marketing strategies and refined targeting to maximize engagement and loyalty across different consumer types.

# **Profiling using Combine Clustering**

The Combined Clustering model was developed to segment consumers based on a comprehensive profile of purchasing behaviors and motivations. This approach aligns with the business goal of AXANTEUS, which aims to gain deeper insights into consumer types, enabling more tailored marketing strategies. The profiling process involved careful data transformation and preparation to identify high-value consumers and understand brand loyalty patterns.

# **Classification Task: Value-Conscious Classification**

The objective of this classification task is to identify **value-conscious consumers**, defined as those in the **top 25%** of spend or transaction counts within **Cluster 1 and Cluster 2** of the combined clustering model. These consumers exhibit behaviors characteristic of value-consciousness, such as high spending or frequent purchase.

**Data Transformation Steps:**

* The target variable **value\_conscious** is defined as:
* **1:** For consumers within the top 25% of spending or transaction counts from the selected clusters.
* **0:** For the remaining 75% of consumers in Cluster 1 and Cluster 2.
* **Feature Selection:**
  + Selected features were refined to avoid predictors that directly contribute to value\_conscious to prevent data leakage.
  + Final refined columns for classification include:
    - **category\_diversity**: Diversity in the product categories purchased by the consumer.
    - **affluence\_index**: A measure of the consumer’s affluence level.
    - **avg\_price**: Average price paid per transaction.
    - **no\_of\_trans**: Total number of transactions made by the consumer (added to capture transaction frequency).

**Final Dataset for Classification:**

| **Variable** | **Description** |
| --- | --- |
| value\_conscious | Target variable (1 = value-conscious, 0 = not) |
| category\_diversity | Diversity in product categories |
| affluence\_index | Consumer's affluence level |
| avg\_price | Average price per transaction |
| no\_of\_trans | Total number of transactions |

*(Table 7 – Dataset for value-conscious classification)*

# **Classification Task: Brand Loyalty Prediction**

The objective of this classification task is to predict **brand loyalty**, defined by the frequency of brand-specific purchases. The aim is to identify consumers who demonstrate a strong inclination toward **brand consistency**, enabling targeted loyalty programs and marketing strategies.

This task specifically focuses on **Cluster 1** and **Cluster 2** identified from the **combined clustering model**, as these clusters represent consumer segments with potential loyalty patterns based on their distinct purchasing behaviors. By refining the analysis within these clusters, the classification aims to enhance the precision of identifying brand-loyal consumers.

**Data Transformation Steps:**

* The target variable is **brand\_loyalty**, defined as:
  + **1:** For consumers in the **top 25% of brand runs** (i.e., consecutive purchases from the same brand) within **Cluster 1** and **Cluster 2** of the **combined clustering model**.
  + **0:** For all other consumers within these clusters.
* **Feature Selection:**
  + Features were selected to avoid those that are highly correlated with the brand\_loyalty target to reduce multicollinearity.
  + Final refined columns for regression include:
    - **category\_diversity**: Captures the diversity of product categories purchased, giving insight into consumer variety-seeking behavior.
    - **affluence\_index**: Reflects the consumer’s affluence level, potentially influencing loyalty.
    - **avg\_price**: Represents the average price per transaction.
    - **no\_of\_trans**: Total number of transactions as a proxy for shopping frequency.

**Final Dataset for Classification:**

| **Variable** | **Description** |
| --- | --- |
| brand\_loyalty | Target variable (1 = brand-loyal, 0 = not) |
| category\_diversity | Diversity in product categories |
| affluence\_index | Consumer's affluence level |
| avg\_price | Average price per transaction |
| no\_of\_trans | Total number of transactions |

*(Table 8 – Dataset for brand-loyality classification)*

# **Regression Task: Brand Runs Prediction**

The objective of this regression task is to predict the **frequency of brand-specific purchases (brand runs)**, providing insights into consumer loyalty and the likelihood of repeat purchases.

This task leverages data from **Cluster 1** and **Cluster 2**, identified through the **combined clustering model**, as these clusters exhibit purchasing patterns indicative of brand loyalty. By focusing on these clusters, the regression task aims to generate actionable insights for tailoring marketing strategies and strengthening customer retention efforts.

**Data Transformation Steps:**

* **Target Variable:** brand\_runs
  + Represents the actual count of brand-specific purchases made by the consumer.
* **Feature Selection:**
  + To ensure relevant predictors for brand-specific purchase frequency, the following features were selected:
    - **category\_diversity**: Provides insight into how diverse the consumer’s purchase categories are, impacting loyalty likelihood.
    - **affluence\_index**: Indicates the consumer’s financial capacity, which can influence buying behavior.
    - **avg\_price**: Captures the average spending per transaction.
    - **no\_of\_trans**: The total number of transactions serves as a measure of purchase frequency.
    - **total\_volume**: Represents overall purchase volume.

**Final Dataset for Regression:**

| **Variable** | **Description** |
| --- | --- |
| brand\_runs | Target variable representing brand-specific purchases |
| category\_diversity | Diversity in product categories |
| affluence\_index | Consumer's affluence level |
| avg\_price | Average price per transaction |
| no\_of\_trans | Total number of transactions |
| total\_volume | Total purchase volume |

*(Table 9 – Dataset for regression)*

**Summary**

* **Value-Conscious Classification**: Focused on predicting value-conscious consumers using category\_diversity, affluence\_index, avg\_price, and no\_of\_trans.
* **Brand Loyalty Classification**: Targeted at identifying brand-loyal consumers, using features such as category\_diversity, affluence\_index, avg\_price, and no\_of\_trans.
* **Brand Runs Prediction**: Estimated the frequency of brand-specific purchases using predictors aligned with shopping habits and financial capacity.

This careful selection and transformation process ensured that relevant and predictive features were included for each classification and regression task, supporting robust model building and insights aligned with business objectives.

# **Data Partitioning**

To evaluate model performance effectively and ensure robust predictive capability, a consistent data partitioning approach is applied across all tasks. Each dataset is split into training and testing subsets, with 70% of the data used for training and 30% reserved for testing. This partitioning ratio strikes a balance between providing sufficient data for model learning while retaining enough data for a meaningful evaluation of model generalizability.

**Purpose of Data Partitioning**

The primary goal of data partitioning is to prevent overfitting by allowing the model to learn from one portion of the data (training set) and then testing its predictive performance on another portion (testing set). This approach ensures that the model’s performance metrics, such as accuracy and error rates, are more representative of its ability to generalize to unseen data, which is crucial for real-world applications.

**Reproducibility and Random Seed**

A fixed random seed is set for each partitioning step to maintain reproducibility. By using the same seed value (e.g., 123), the random sampling process remains consistent, ensuring that the same records are included in the training and testing sets each time the partitioning is performed. This stability in partitioning enables reliable comparisons across different models and iterations.

**Summary of Partitioning Results**

For the value-conscious classification task, the training set includes 223 records labeled as "not value-conscious" and 153 records labeled as "value-conscious." The test set holds 105 and 57 records for these respective categories, maintaining a balanced representation of each target label in both subsets.

In the brand loyalty classification task, the training set consists of 287 records labeled as "not brand-loyal" and 89 records labeled as "brand-loyal," while the test set contains 119 and 43 records for each group. This distribution ensures that both training and test sets reflect the original class proportions, which supports accurate model assessment for brand loyalty prediction.

For the brand runs regression task, the data is similarly split, retaining 70% of the data for model training and 30% for testing. This consistent partitioning across all tasks aligns with best practices in predictive modeling, providing a robust foundation for reliable model evaluation.

# **Model Selection for Classification and Regression**

The classification and regression tasks in this analysis are aimed at understanding consumer behaviors and predicting outcomes relevant to business goals. Specifically, the classification models focus on identifying value-conscious consumers and predicting brand loyalty, while the regression model is designed to forecast brand-specific purchasing patterns, or “brand runs.” To achieve these objectives, selecting appropriate models for each task is critical for balancing interpretability and predictive accuracy.

**Classification Task: Model Selection for Value-Conscious and Brand Loyalty Classification**

For the classification tasks (Value-Conscious Classification and Brand Loyalty Prediction), logistic regression is selected as the primary model due to its interpretability, simplicity, and suitability for binary classification. Logistic regression provides clear coefficient estimates that reflect the direction and strength of each feature’s effect on the target variable, which is valuable for business insights. Additionally, logistic regression handles binary outcomes effectively, making it a reliable choice for classifying consumers as value-conscious or brand-loyal.

Key benefits of logistic regression for classification tasks include:

* **Interpretability**: Coefficients provide straightforward insights into the impact of each feature on the target outcome.
* **Simplicity**: Logistic regression is computationally efficient, making it suitable for datasets with moderate feature complexity.
* **Scalability**: Logistic regression can easily scale to larger datasets while maintaining stable performance.

For each classification task, logistic regression models are trained on the selected features and evaluated using standard classification metrics, ensuring both predictive performance and interpretability.

**Regression Task: Model Selection for Brand Runs Prediction**

For the regression task predicting brand runs, a linear regression model is selected. Linear regression is appropriate for predicting a continuous outcome, such as the number of brand-specific purchases. It offers interpretability through coefficients, allowing an understanding of how each feature contributes to the predicted brand run count. Additionally, linear regression provides insights into the impact of predictors like category diversity, transaction frequency, and affluence on brand loyalty patterns.

Advantages of linear regression for the brand runs prediction include:

* **Interpretability**: The linear relationship allows each feature’s effect on the outcome to be easily understood, supporting business-focused insights.
* **Predictive Performance**: Linear regression is effective when relationships between predictors and the outcome are approximately linear, which aligns with the expectations for brand run predictions.
* **Efficiency**: Linear regression is computationally efficient, allowing it to handle datasets with moderate feature counts without compromising accuracy.

The linear regression model is evaluated using standard metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to gauge predictive accuracy, ensuring its alignment with the business goal of accurately predicting consumer brand loyalty patterns.

**Classification Task 1: Value-Conscious Classification Model**

The goal of this classification task is to predict **value-conscious consumers** based on their transaction patterns and category diversity. This model supports targeted marketing initiatives by identifying consumers who are more likely to respond to deals and discounts.

**Model Selection and Rationale**

A **logistic regression** model was selected for this task due to its interpretability and suitability for binary classification. Logistic regression provides valuable insights into the relationship between predictors and the likelihood of a consumer being value-conscious.

**Data Preparation and Training**

* The dataset was split into training (70%) and test (30%) sets, ensuring an adequate representation of both value-conscious and non-value-conscious consumers. Key predictors included **category\_diversity**, **affluence\_index**, **avg\_price**, and **no\_of\_trans**, while **value\_conscious** served as the target variable.
* **Distribution**: **Training Set:** 223 non-value-conscious and 153 value-conscious consumers. **Test Set:** 105 non-value-conscious and 57 value-conscious consumers

**Model Results and Interpretation**

The logistic regression model provided insights into the effect of predictors on the probability of a consumer being value-conscious. Key metrics and coefficients are summarized in the table below:

| **Variable** | **Estimate** | **Std. Error** | **z-value** | **p-value** |
| --- | --- | --- | --- | --- |
| **Intercept** | -8.412 | 0.980 | -8.588 | < 0.001 \*\*\* |
| **Category Diversity** | -0.158 | 0.084 | -1.877 | 0.060 |
| **Affluence Index** | 0.003 | 0.018 | 0.193 | 0.847 |
| **Avg Price** | 0.132 | 0.045 | 2.959 | 0.003 \*\* |
| **No. of Transactions** | 0.262 | 0.031 | 8.562 | < 0.001 \*\*\* |

*(Table 10 –value-conscious model summary)*

* **Significance**: The avg\_price and no\_of\_trans predictors significantly influenced value-conscious classification, with a higher transaction frequency indicating a higher probability of a consumer being value-conscious.

**Model Performance**

The model achieved strong predictive performance as shown in the confusion matrix and accuracy metrics.

| **Metric** | **Value** |
| --- | --- |
| **Accuracy** | 0.815 |
| **Sensitivity** | 0.754 |
| **Specificity** | 0.848 |

*(Table 11 – Performance metrics for classification-1)*

* **Accuracy**: 81.5%, indicating the model is performing well in classifying value-conscious consumers.
* **Sensitivity** (Recall for the positive class): 75.4%, meaning that 75.4% of actual value-conscious consumers were correctly identified.
* **Specificity**: 84.8%, indicating a high ability to correctly identify non-value-conscious consumers.

**Classification Task 2: Brand Loyalty Classification Model**

The goal of this classification task is to predict brand loyalty among consumers, identifying those with high loyalty based on consistent brand purchases. The model supports loyalty program design by isolating consumers more likely to stick with a specific brand.

**Model Selection and Rationale**

Logistic regression was again selected for its effectiveness in binary classification and interpretability. This model allows an analysis of how each predictor contributes to brand loyalty probability.

**Data Preparation and Training**

* **Training and Test Split**: The dataset was split similarly into training (70%) and test (30%) sets for brand\_loyalty classification. The predictors included category\_diversity, affluence\_index, avg\_price, and no\_of\_trans.
* **Distribution**: The training set contained 287 non-loyal and 89 loyal consumers, while the test set had 119 non-loyal and 43 loyal consumers.

**Model Results and Interpretation**

The logistic regression model yielded insights on the effect of predictors on brand loyalty, as summarized below:

| **Variable** | **Estimate** | **Std. Error** | **z-value** | **p-value** |
| --- | --- | --- | --- | --- |
| **Intercept** | -13.248 | 1.552 | -8.535 | < 0.001 \*\*\* |
| **Category Diversity** | 0.759 | 0.120 | 6.313 | < 0.001 \*\*\* |
| **Affluence Index** | 0.014 | 0.020 | 0.706 | 0.480 |
| **Avg Price** | 0.128 | 0.053 | 2.385 | 0.017 \* |
| **No. of Transactions** | 0.129 | 0.021 | 6.037 | < 0.001 \*\*\* |

*(Table 12 –Brand-Loyalty Classification model summary)*

* **Significance**: category\_diversity, avg\_price, and no\_of\_trans were significant predictors, indicating that diversity in category purchases, transaction frequency, and average spend per transaction contribute to brand loyalty.

**Model Performance**

The model demonstrated high predictive accuracy, indicating its effectiveness in identifying brand-loyal consumers.

| **Metric** | **Value** |
| --- | --- |
| **Accuracy** | 0.895 |
| **Sensitivity** | 0.698 |
| **Specificity** | 0.966 |

*(Table 13 –Performance metrics for classification-2)*

**Summary**

Both classification models performed effectively, with the brand loyalty model showing slightly higher accuracy. These models offer actionable insights for targeted promotions and loyalty programs by identifying high-value and brand-loyal consumers based on transaction behaviors and other characteristics.

This analysis not only supports tailored marketing strategies but also enhances customer engagement by catering to specific consumer preferences identified through these classification models.

**ROC Plot**

The ROC plot presented above illustrates the performance of two classification models: the Value-Conscious Model and the Brand Loyalty Model. The ROC (Receiver Operating Characteristic) curve is a graphical representation that evaluates the trade-off between the true positive rate (sensitivity) and false positive rate (1 - specificity) across various threshold values for a classifier. Each curve shows how well each model can distinguish between the positive and negative classes.

A graph of a curve

Description automatically generated

*(Fig 23 – ROC Plot for Classification Model)*

* **Value-Conscious Model (Blue Curve)**: The ROC curve for the Value-Conscious Model demonstrates a strong classification capability, with an AUC (Area Under the Curve) of 0.90. This high AUC value indicates that the model is effective at distinguishing between value-conscious and non-value-conscious consumers, with a relatively high sensitivity and specificity balance. The model's curve is closer to the top-left corner of the plot, further confirming its strong performance.
* **Brand Loyalty Model (Green Curve)**: The ROC curve for the Brand Loyalty Model shows an even higher AUC of 0.95. This suggests that the model performs excellently in differentiating brand-loyal consumers from non-loyal ones. The green curve reaches closer to the ideal point (top-left corner), reflecting the model’s higher accuracy and a stronger balance between sensitivity and specificity compared to the Value-Conscious Model.

In summary, both models show strong classification performance, with the Brand Loyalty Model demonstrating a slightly superior capability due to its higher AUC. The ROC curves confirm that each model is well-suited for its respective classification task, with effective discrimination between the target classes.

# **Random Forest Model: Value-Conscious Classification**

The goal of this analysis is to classify consumers who are considered "value-conscious." This classification is essential for understanding consumer behavior patterns, particularly identifying high spenders and frequent buyers, which can support targeted marketing strategies and promotional planning.

**Model Overview**

The Random Forest classification model was selected due to its robustness, ability to handle high-dimensional data, and capability to identify complex interactions between features. The model was trained on the value-conscious classification data with the following parameters:

* **Number of Trees (ntree)**: 500
* **Variables Tried at Each Split (mtry)**: 2

**Model Performance**

The performance of the Random Forest model was evaluated on the test dataset using several metrics, with results summarized below:

| **Metric** | **Value** |
| --- | --- |
| **Accuracy** | 84.57% |
| **95% Confidence Interval** | (78.07%, 89.76%) |
| **Kappa** | 0.663 |
| **Sensitivity (Class 0)** | 87.62% |
| **Specificity (Class 1)** | 78.95% |
| **Positive Predictive Value (Class 0)** | 88.46% |
| **Negative Predictive Value (Class 1)** | 77.59% |
| **Balanced Accuracy** | 83.28% |

*(Table 14 –Performance metrics for random forest-1)*

The model achieved an overall accuracy of 84.57%, demonstrating its effectiveness in distinguishing value-conscious consumers. The Kappa value of 0.663 indicates substantial agreement beyond chance, further validating the model's classification accuracy.

**Confusion Matrix**

The confusion matrix provides insights into the distribution of correct and incorrect classifications across the two classes:

| **Actual \ Predicted** | **0** | **1** | **Class Error** |
| --- | --- | --- | --- |
| **0** (Not Value-Conscious) | 92 | 12 | 0.1235 |
| **1** (Value-Conscious) | 13 | 45 | 0.2241 |

*(Table 15 –Confusion Matrix for classification-1)*

The confusion matrix highlights:

* A relatively low misclassification rate for non-value-conscious consumers (Class 0).
* Slightly higher misclassification rate for value-conscious consumers (Class 1), which is acceptable given the model's overall balanced accuracy.

**Feature Importance**

The importance of each feature was evaluated using Mean Decrease in Accuracy and Mean Decrease in Gini, reflecting each feature's contribution to the model's predictive power.

| **Feature** | **Mean Decrease Accuracy** | **Mean Decrease Gini** |
| --- | --- | --- |
| **no\_of\_trans** | 119.27 | 104.73 |
| **avg\_price** | 10.67 | 30.99 |
| **affluence\_index** | 9.89 | 23.52 |
| **category\_diversity** | 12.61 | 21.56 |

*(Table 16 –Feature Importance for random forest-1)*

* **no\_of\_trans** (number of transactions) is the most influential feature, suggesting that frequent transactions are a strong indicator of value-conscious behavior.
* **avg\_price** and **affluence\_index** also contribute significantly, highlighting the roles of purchase pricing patterns and the consumer's financial standing.
* **category\_diversity** has a relatively lower influence but still contributes meaningful information regarding the diversity of the consumer's purchases.

|  |  |
| --- | --- |
| A graph of a number of numbers and a number of numbers  Description automatically generated with medium confidence | A comparison of a number of numbers  Description automatically generated with medium confidence |

*(Fig 24 –Feature Importance for random forest-1)*

**Interpretation and Insights**

The Random Forest model’s high accuracy and balanced sensitivity and specificity make it a reliable tool for classifying value-conscious consumers. The feature importance analysis reveals that transaction frequency is the strongest indicator of value-conscious behavior, which aligns with business expectations. Consumers with a higher number of transactions tend to exhibit value-conscious characteristics, either through frequent purchasing or active engagement with promotional offers.

**Conclusion**

The Random Forest model for value-conscious classification performs well, with an accuracy of 84.57% and balanced recall across both classes. The analysis shows that transaction frequency, along with average purchase price and affluence level, are key predictors of value-conscious behavior. This model provides a robust foundation for targeting marketing efforts toward value-conscious consumers, supporting data-driven strategies for enhanced customer engagement and retention.

This report provides a detailed, data-driven understanding of the Random Forest model’s performance and the insights it offers for identifying value-conscious consumers. This information can guide strategic marketing decisions and promotional activities tailored to high-value consumer segments.

# **Brand Loyalty Classification Using Random Forest**

The goal of this classification task is to predict brand loyalty based on various consumer features. Brand loyalty is defined as consumers exhibiting a high frequency of brand-specific purchases. For this analysis, a Random Forest model with 500 trees and two variables tried at each split was employed.

**Model Training Summary**

* **Type**: Classification
* **Number of Trees**: 500
* **Variables Tried at Each Split (mtry)**: 2
* **Out-of-Bag (OOB) Error Estimate**: 15.16%

The OOB error estimate indicates that approximately 15.16% of predictions may be misclassified when the model is used on unseen data.

**Feature Importance**

The importance of each feature was evaluated using Mean Decrease Accuracy and Mean Decrease Gini, which reflect the significance of each feature in predicting brand loyalty.

| **Feature** | **Mean Decrease Accuracy** | **Mean Decrease Gini** |
| --- | --- | --- |
| Category Diversity | 35.92 | 36.26 |
| Affluence Index | -5.53 | 17.71 |
| Average Price | 12.64 | 29.03 |
| Number of Transactions | 40.71 | 52.29 |

*(Table 17 – Feature importance by random forest-2)*

**Interpretation**:

* The number of transactions is the most influential variable, followed by category diversity and average price.
* Affluence index shows a low, even negative, mean decrease in accuracy, suggesting it may have a minimal or inconsistent impact on predicting brand loyalty.

|  |  |
| --- | --- |
| A graph of a number of numbers and a number of text  Description automatically generated with medium confidence | **A graph of a number of numbers and a number of numbers  Description automatically generated with medium confidence** |

*(Fig 25 –Feature Importance for random forest-2)*

**Model Performance on Test Set**

The model was evaluated using a test set, and the following results were obtained:

| **Metric** | **Value** |
| --- | --- |
| **Accuracy** | 91.98% |
| **Kappa** | 0.7795 |
| **Sensitivity** | 98.32% |
| **Specificity** | 74.42% |
| **Positive Predictive Value** | 91.41% |
| **Negative Predictive Value** | 94.12% |
| **Balanced Accuracy** | 86.37% |

*(Table 18 –Performance metrics for random forest-2)*

* **Accuracy**: The model achieved an overall accuracy of 91.98%, indicating strong predictive performance.
* **Kappa**: A kappa of 0.7795 suggests substantial agreement between the model predictions and actual class labels.
* **Sensitivity**: With a sensitivity of 98.32%, the model effectively identifies loyal consumers (class 0).
* **Specificity**: Specificity at 74.42% implies that the model has a moderate ability to identify non-loyal consumers (class 1).

**Confusion Matrix**

The confusion matrix below details the prediction results.

|  | **Actual Class 0** | **Actual Class 1** |
| --- | --- | --- |
| **Predicted 0** | 117 | 11 |
| **Predicted 1** | 2 | 32 |

*(Table 19 – Confusion Matrix for random forest-2)*

**Model Insights**

* **High Sensitivity**: The model has a strong ability to correctly identify loyal consumers.
* **Moderate Specificity**: The specificity score indicates room for improvement in distinguishing non-loyal consumers.

**Conclusion**

The Random Forest model for brand loyalty classification demonstrates high accuracy and sensitivity, making it suitable for identifying loyal consumers effectively. While specificity is moderate, the overall performance is robust, achieving a balanced accuracy of 86.37%.

# **Regression Model: Brand Runs Prediction**

The regression model aims to predict brand loyalty by estimating the frequency of "brand runs" (repeated purchases of the same brand) based on consumer behavior and demographic attributes. This model provides insights into consumers’ brand loyalty patterns, enabling targeted marketing strategies to encourage repeat purchases.

**Model Summary Table**

The table below presents the coefficients, standard errors, t-values, and p-values for each predictor in the model:

| **Predictor** | **Estimate** | **Std. Error** | **t-value** | **p-value** | **Significance** |
| --- | --- | --- | --- | --- | --- |
| Intercept | -8.13 | 1.03 | -7.91 | <0.0001 | \*\*\* |
| category\_diversity | 1.73 | 0.12 | 14.92 | <0.0001 | \*\*\* |
| affluence\_index | 0.11 | 0.03 | 4.03 | <0.0001 | \*\*\* |
| avg\_price | 0.11 | 0.06 | 1.69 | 0.0915 | . |
| no\_of\_trans | 0.27 | 0.03 | 10.43 | <0.0001 | \*\*\* |
| total\_volume | -0.000072 | 0.000035 | -2.05 | 0.0414 | \* |

*(Table 20– Model Summary for regression)*

**Significance Codes**:

* \*\*\*: p < 0.001
* \*\*: p < 0.01
* \*: p < 0.05
* .: p < 0.1

**Model Performance**

| **Metric** | **Value** |
| --- | --- |
| Mean Absolute Error (MAE) | 2.93 |
| Mean Squared Error (MSE) | 14.08 |
| Root Mean Squared Error (RMSE) | 3.75 |
| Mean Absolute Percentage Error (MAPE) | 32.42% |

*(Table 21 –Performance Metrics for Regression Model)*

These values suggest that, on average, the model’s predictions differ from actual values by approximately 2.93 brand runs. While the MAPE indicates a 32.42% error relative to actual values, the low RMSE demonstrates reasonable predictive accuracy.

**Interpretation of Results**

* **Category Diversity**: The strong positive coefficient (1.73) indicates that consumers purchasing across diverse categories tend to have higher brand loyalty, as reflected by increased brand runs.
* **Affluence Index**: A smaller yet significant positive relationship (0.11) with brand runs suggests that affluent consumers are likely to show brand loyalty.
* **Average Price**: While the impact of average price on brand runs is positive (0.11), its significance is weaker, indicating only a marginal influence on brand loyalty.
* **Number of Transactions**: With a coefficient of 0.27, frequent shoppers are more likely to display brand loyalty, reinforcing the relationship between purchase frequency and brand affinity.
* **Total Volume**: The negative coefficient (-0.000072) implies a minor inverse relationship between total purchase volume and brand loyalty, though the effect size is very small.

**Predicted vs. Actual Values Plot**

The plot below illustrates the relationship between predicted and actual brand runs in the test dataset. The dashed red line represents an ideal match between predicted and actual values, while blue dots show individual predictions.

A graph with blue dots and red line

Description automatically generated

*(Fig 26 – Regression Plot)*

This visualization helps assess the model’s predictive capability, with points closer to the red line indicating accurate predictions.

**Summary**

The regression model demonstrates effective prediction of brand loyalty based on key predictors such as category diversity and transaction frequency. These findings suggest actionable insights for developing loyalty programs and marketing strategies targeted at frequent, diverse-category shoppers.

# **Key Insights from the Analysis**

This study presents a comprehensive approach to consumer segmentation and predictive analytics, leveraging clustering, classification, and regression techniques. The results provide actionable insights into consumer behavior and preferences, enabling targeted marketing strategies and resource allocation. Below is a detailed summary of the findings:

**1. Clustering Analysis**

The clustering process was divided into three stages: purchase behavior clustering, purchase basis clustering, and combined clustering. Each stage provided a distinct perspective on consumer segmentation:

* **Purchase Behavior Clustering:**
  + Consumers were grouped based on purchasing intensity and brand loyalty.
  + **Clusters Identified:**
    - Moderate Buyers (balanced activity and loyalty).
    - High-Frequency Shoppers (engaged and responsive to loyalty programs).
    - Low-Engagement Shoppers (low activity, requiring re-engagement strategies).
  + **Key Metric:** The average silhouette score was 0.499, indicating moderate clustering quality.
* **Purchase Basis Clustering:**
  + Consumers were segmented based on deal responsiveness and product category diversity.
  + **Clusters Identified:**
    - Promotion-Sensitive, Narrow Buyers (price-conscious with focused buying).
    - Balanced, Diverse Shoppers (moderate responsiveness with high product diversity).
    - Variety-Seeking, Deal-Resistant Shoppers (preference-driven purchases with low deal sensitivity).
  + **Key Metric:** An average silhouette score of 0.589 indicated moderate to strong clustering quality.
* **Combined Clustering:**
  + An integrated approach combining metrics from behavior and basis clustering.
  + **Clusters Identified:**
    - Balanced, High-Engagement Consumers (consistent purchasing with product diversity).
    - Promotion-Driven, Frequent Shoppers (deal-responsive with frequent transactions).
    - Variety-Seeking, Low-Engagement Shoppers (selective buyers with diverse interests).
  + **Key Metric:** The silhouette score was 0.474, reflecting moderate cluster coherence.

**2. Classification Tasks**

Two classification tasks aimed at identifying key consumer behaviors were conducted:

* **Value-Conscious Classification:**
  + Logistic regression identified high-spending, frequent buyers within combined clustering groups.
  + **Significant Predictors:** Number of transactions and average purchase price.
  + **Model Performance:** Achieved 81.5% accuracy with strong sensitivity (84.8%).
* **Brand Loyalty Classification:**
  + Focused on identifying consumers with a strong preference for specific brands.
  + **Significant Predictors:** Category diversity, transaction frequency, and average price.
  + **Model Performance:** Achieved 89.5% accuracy with a high sensitivity (96.6%).

**3. Regression Task**

The regression task aimed to predict the frequency of brand-specific purchases (brand runs):

* **Significant Predictors:**
  + Category diversity, transaction frequency, and consumer affluence index showed strong positive relationships.
  + Total purchase volume exhibited a weak inverse relationship.
* **Model Performance:**
  + RMSE of 3.75 and MAPE of 32.42% indicated reasonable predictive accuracy.

**Key Themes Across Analyses**

1. **Transaction Frequency as a Core Indicator:** Across clustering, classification, and regression tasks, the number of transactions consistently emerged as a key variable influencing consumer behavior and loyalty.
2. **Category Diversity and Engagement:** Consumers purchasing across diverse categories tend to show higher loyalty and engagement, suggesting opportunities for cross-category marketing.
3. **Role of Promotions:** Responsiveness to deals varied significantly among clusters, highlighting the need for tailored promotional strategies.

# **Next Steps and Recommendations**

Building on these findings, the following recommendations are proposed:

* **Enhance Loyalty Programs:** Focus on high-frequency and value-conscious shoppers by introducing tiered rewards to incentivize consistent purchases and loyalty.
* **Targeted Promotions for Low-Engagement Shoppers:** Develop personalized campaigns to re-engage this segment through discounts or product recommendations.
* **Cross-Category Campaigns:** Leverage insights on category diversity to promote complementary products, encouraging exploratory buying behaviors.
* **Continued Model Refinement:** Monitor the performance of predictive models over time and incorporate additional variables such as external economic indicators for enhanced accuracy.

This comprehensive approach underscores the potential of data-driven strategies in consumer segmentation, enabling AXANTEUS to optimize marketing efforts and improve customer retention effectively.

# **Conclusion**

This project has successfully developed a robust, data-driven framework for consumer segmentation and predictive analytics, aligned with AXANTEUS’s strategic objectives. By employing advanced clustering techniques, classification models, and regression analysis, we have gained actionable insights into consumer behaviors, motivations, and loyalty patterns, providing a solid foundation for targeted marketing strategies and enhanced customer engagement.

Through exploratory data analysis (EDA), we identified key behavioral and demographic variables such as transaction frequency, category diversity, and average purchase price, which were instrumental in understanding purchasing patterns. Clustering analysis was performed across multiple dimensions—purchase behavior, purchase motivation, and combined metrics—using the K-means algorithm. Among these, the combined clustering approach proved the most effective, providing a comprehensive view of consumer profiles by integrating both behavioral and motivational dimensions. This approach facilitated the creation of three distinct consumer segments:

1. **Balanced, High Engagement Consumers** – Consistent purchasers with moderate brand loyalty and diverse category preferences.
2. **Promotion-Driven, Frequent Shoppers** – Consumers highly responsive to deals and frequent buyers.
3. **Variety-Seeking, Low Engagement Shoppers** – Consumers with selective purchase habits across diverse product categories.

Recognizing the value of the combined clustering model, it was selected as the foundation for the classification and regression tasks. Classification models were developed to identify **value-conscious consumers** and **brand-loyal consumers**, leveraging logistic regression and Random Forest techniques. The classification models achieved significant accuracy, with the Random Forest model achieving 84.57% accuracy in identifying value-conscious consumers and 91.98% accuracy in predicting brand loyalty. These insights enable AXANTEUS to target consumers with customized promotions, loyalty programs, and engagement strategies.

The regression task focused on predicting **brand runs**, a key measure of brand-specific loyalty. By analyzing factors such as category diversity, transaction frequency, and affluence index, the regression model provided actionable insights into the likelihood of repeat brand purchases, enabling AXANTEUS to refine its loyalty-driven campaigns and product offerings.

This project highlights the transformative power of combining clustering and predictive analytics to derive deeper consumer insights. The decision to build classification and regression models on the combined clustering approach ensured that both segmentation and prediction were closely aligned with business objectives, maximizing the utility of the analysis. Key applications of these insights include:

* **Enhanced Engagement:** Loyalty programs targeting high-frequency shoppers.
* **Targeted Promotions:** Discount-driven campaigns for price-sensitive consumers.
* **Strategic Re-engagement:** Initiatives to increase activity among low-engagement shoppers.

Moving forward, the project outcomes can be further enhanced by integrating additional data sources, employing advanced feature engineering, and tuning model parameters for better performance. Feedback from implementation could also inform iterative improvements, ensuring continued alignment with AXANTEUS’s evolving business needs.

In conclusion, this project underscores the value of behavior-focused analytics in strategic decision-making. By combining clustering with predictive modeling, AXANTEUS is now equipped to optimize marketing strategies, improve customer retention, and strengthen consumer relationships. This data-driven framework provides a scalable, adaptable foundation for future analytical efforts, ensuring AXANTEUS remains competitive in an ever-changing market landscape.