Customer Segmentation Analysis: Identifying Distinct Customer Profiles Using K-Means Clustering

Table of Contents

1 Introduction

1.1 Overview of the Project

This project focuses on customer segmentation through unsupervised machine learning. It follows a structured data science workflow that includes data preprocessing, feature engineering, dimensionality reduction, and clustering. Starting with raw customer data containing 29 features, the analysis moves through key steps: cleaning to fix quality issues, creating behavior-focused variables, removing redundant features through multicollinearity checks, applying Principal Component Analysis (PCA) to reduce complexity, and using K-Means clustering to uncover distinct customer groups. The businesses can use the resulting set of clear, actionable customer profiles to better understand different types of customers and create more personalized marketing strategies.

1.2 Objectives

The main objectives of this project are:

1. **Data Preprocessing & Feature Engineering:** To clean and transform the raw customer data by handling missing values, outliers, and categorical variables, while creating new behavioral features that capture spending patterns and channel and product preferences.
2. **Dimensionality Reduction:** To address the curse of dimensionality using PCA by choosing the optimal number of components that preserve maximum variance and focus on the most important patterns
3. **Cluster Identification:** To determine the optimal number of customer segments using both elbow method and silhouette analysis, and apply K-Means clustering to place customers into distinct groups.
4. **Cluster Profiling:** To analyze and translate the statistical results into clear profiles to understand who these customers are and how they behave
5. **Strategic Implementation:** To provide insights that can support personalized marketing campaigns, strengthen customer retention strategies, and optimize resource allocation based on segment behaviors and preferences.

1.3 Dataset Description

The dataset comes from a marketing campaign and includes detailed information about customers. It contains the information of 2,240 customers (rows), described across 29 features capturing their demographic, financial, and behavioral characteristics.
Link: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

1.3.1 Variables

Some of the key variables in the dataset include:

- **People**
  - **ID:** Unique identifier for each customer.
  - **Year_Birth:** The year the customer was born
  - **Education:** Education level of the customer (Graduation, PhD, Master, 2n Cycle, Basic)
  - **Marital_Status:** Marital status (Married, Together, Single, Divorced, Widow, Alone, Absurd, YOLO)
  - **Income:** Customer's yearly household income
  - **Kidhome:** Number of children living in the household.
  - **Teenhome:** Number of teenagers living in the household.
  - **Dt_Customer:** Date of customer's enrollment with the company
  - **Recency:** Number of days since the last purchase.
  - **Complain:** Whether the customer has lodged a complaint (1 = yes, 0 = no).
- **Products**
  - **MntWines:** Amount spent on wine in last 2 years
  - **MntFruits:** Amount spent on fruits in last 2 years
  - **MntMeatProducts:** Amount spent on meat in last 2 years
  - **MntFishProducts:** Amount spent on fish in last 2 years
  - **MntSweetProducts:** Amount spent on sweets in last 2 years
  - **MntGoldProds:** Amount spent on gold in last 2 years
- **Promotions**
  - **NumDealsPurchases:** Number of purchases made with a discount
  - **AcceptedCmp1:** Whether the customer accepted the offer in the 1st campaign (1 = yes, 0 = no).
  - **AcceptedCmp2:** Whether the customer accepted the offer in the 2nd campaign (1 = yes, 0 = no).
  - **AcceptedCmp3:** Whether the customer accepted the offer in the 3rd campaign (1 = yes, 0 = no).
  - **AcceptedCmp4:** Whether the customer accepted the offer in the 4th campaign (1 = yes, 0 = no).
  - **AcceptedCmp5:** Whether the customer accepted the offer in the 5th campaign (1 = yes, 0 = no).
  - **Response:** Whether the customer accepted the offer in the latest campaign (1 = yes, 0 = no).
- **Place**
  - **NumWebPurchases:** Number of purchases made through the company’s website
  - **NumCatalogPurchases:** Number of purchases made using a catalogue
  - **NumStorePurchases:** Number of purchases made directly in stores

    - **NumWebVisitsMonth:** Number of times the customer visited the company's website in the last month.
- **Miscellaneous**
    - **Z_CostContact:** Constant value (= 3 for all records).
    - **Z_Revenue:** Constant value (= 11 for all records).

2 Preprocessing Data

2.1 Data Cleaning

The first step of preprocessing the data was to address the quality issues and remove irrelevant information.

- **Date Conversion:** The Dt_Customer column, which recorded the customer's sign-up date, was converted from a text format into a proper datetime format to allow for time-based analysis.
- **Handling Missing Values:** The dataset was checked for missing information. Only the Income column was found to have a small number of missing entries (24 rows). These rows were removed from the dataset to ensure accuracy, as imputing (estimating) income could have introduced bias.
- **Removing Duplicates:** The data was checked for and found to have no duplicate customer records.
- **Removing Constant Columns:** Two columns, Z_CostContact and Z_Revenue, were found to contain only a single, constant value for every customer. These columns provided no useful information for distinguishing between customers and were therefore removed.

2.2 Recoding Categorical Variables

To simplify the analysis and create more meaningful customer groups, the Marital_Status and Education categories were simplified into broader groups that better represented the categories.

- Marital_Status: The original column contained eight categories. These were grouped into two simpler, more general categories: Partner (for customers indicating a relationship) and Single (for all others).
- Education: The original education levels were also grouped into broader, more standardized categories. Basic was recoded as Undergraduate, Graduation became Graduate, Master and 2n Cycle were combined into Master, and PhD remained as PhD.

This regrouping makes the data less complicated, while also making sure that there are enough customers in each group to analyze effectively.

2.3 Handling Outliers

An analysis of outliers was conducted to prevent them from distorting clusters and reducing overall accuracy and coherence. The process was as follows:

- **Investigation:** Outliers were identified using the Interquartile Range (IQR) method applied to key numerical variables including Income, spending categories, and purchase frequencies. Customers were flagged as outliers if they exceeded the IQR thresholds in any single variable.
- **Finding Errors:** One customer was found to have an implausible yearly income of $666,666 while having near-zero spending. This was determined to be a clear data entry error, and this record was removed from the dataset.
- **Retention of True High-Value Customers:** Customers with very high income and spending across categories were flagged as statistical outliers but retained, since they represent real, valuable segments of interest.
- **Decision:** The legitimate outlier records were kept in the dataset to make sure that these critical customers would be identified in our segmentation.

2.4 Feature Engineering

New features were created from the existing data to provide more powerful and interpretable inputs for the clustering algorithm. The goal was to use raw transaction counts and amounts to create variables that directly describe customer behavior and characteristics. This process also involved removing the original, redundant columns used to create them. The following are the new engineered features:

- **New Demographic Features:**
    - **Age** was calculated from the Year_Birth column to better represent life stage than birth year alone.
    - **Tenure_Years** was calculated from the Dt_Customer column to measure customer loyalty and longevity with the company.
    - **Total_Children and Family_Size** were created to provide a clear picture of household composition, which can be a key driver of purchasing habits.
    - **Is_Parent** was created as a simple yes/no (1/0) flag to easily segment customers based on parental status.
- **New Behavioral Features:**
    - **Total_Spent** summed all individual product spending (e.g., wines, meats, sweets) into a single metric of customer value.
    - **Total_Purchases** summed all purchases across different channels (web, catalog, store).
    - **Total_Accepted_Cmp** summed the total number of marketing campaigns a customer accepted. This can be used to measure a customer's overall responsiveness to marketing, which is more useful than looking at individual campaign responses.

- New Preference Features (Ratios):
    - **Spending ratios** were created to understand a customer's product preferences, independent of how much they spend. This can help identify ,for example, whether a customer is a "Wine Lover" or "Meat Fan".
    - **Purchase channel ratios** (Web_Purchase_Ratio, Catalog_Purchase_Ratio, Store_Purchase_Ratio) were created to understand a customer's channel preference, independent of how often they buy. This helps identify "Online Shoppers" vs. "Catalog Shoppers".
- **Removal of Redundant Features:**
  The original columns used to create these new features (e.g., Year_Birth, Kidhome, MntWines, NumWebPurchases) were dropped from the dataset. This prevents redundancy and makes sure that the model focuses on the most insightful data, which will lead to cleaner and more interpretable segments.

2.5 Addressing Multicollinearity

It's important to identify and avoid highly correlated features. While some correlation is expected, very high multicollinearity can distort the results of clustering algorithms by giving undue weight to overlapping concepts. This can confuse the clustering algorithm and make it harder to find clear, unique customer segments. Below is the process taken to addressing multicollinearity:

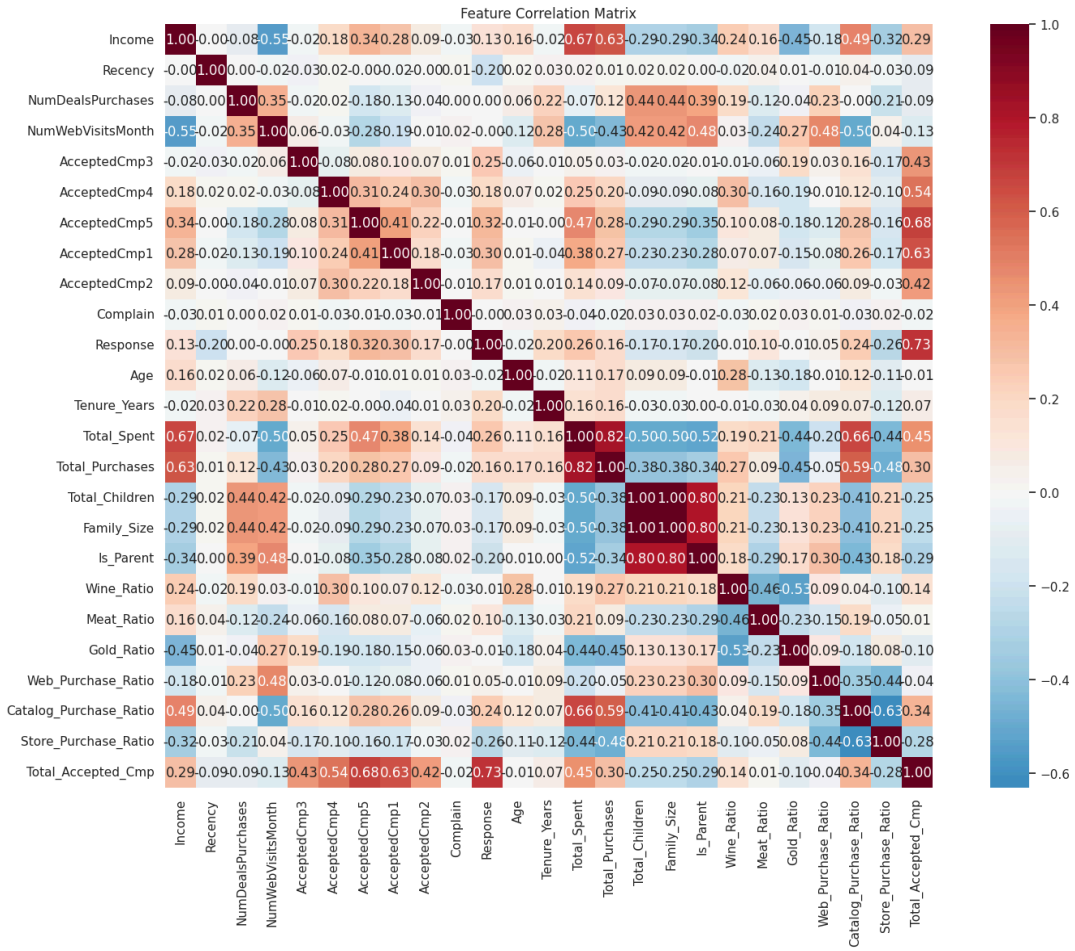

Figure 1 Correlation Heatmap of Customer Features: Correlation heatmap identifying multicollinear features. Redundant variables (campaign responses, Total_Children) were removed to optimize clustering feature selection.

- **Method:** A correlation matrix was calculated and visualized as a heatmap (Figure 1) to see the relationships between all numerical features.
- **Critical Findings:**
    - The feature Total_Children was found to be perfectly correlated (r = 1.0) with Family_Size, which is understandable seeing as one was derived from the other.
    - Total_Accepted_Cmp is highly correlated with many campaign columns (ex. AcceptedCmp5: 0.68, AcceptedCmp1: 0.63, Response: 0.73). This is expected since it's their sum.
- **Other Findings:**
    - Total_Spent and Income are positively correlated (r = 0.67), which makes perfect sense).
    - Total_Spent and Catalog_Purchase_Ratio are also positively correlated (r = 0.66), this suggests that big spenders enjoy buying from catalogs

    - NumWebVisitsMonth and Total_Spent have a negative correlation (r = -0.50). This suggests that the more someone visits the website, the less they spend. These customers might be browsers or deal-seekers.
- **Action Taken:** The following redundant or highly multicollinear features were dropped:

  ['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'Response', 'Total_Children']

  The Total_Accepted_Cmp feature was kept as a summary of overall marketing campaign engagement.

By addressing multicollinearity, the final feature set is non-redundant and focused on unique information of customer behavior, which is optimal clustering results.

2.6 Cleaned and Engineered Dataset Description

The dataset has been cleaned, preprocessed, and feature-engineered to create 20 powerful variables that are ideal for customer segmentation. The information of 2,215 customers are described over features that can be grouped into five intuitive categories that describe the customer's demographic, engagement, spending habits, product preferences, and channel behavior.

1. Demographic & Status Features

- **Education** (object): The highest education level of the customer (Undergraduate, Graduate, Master, PhD).
- **Marital_Status** (object): The family status of the customer (Single, Partner).
- **Income** (float64): The customer's yearly household income ($).
- **Age** (int64): The age of the customer (derived from Year_Birth).
- **Family_Size** (int64): The total number of adults and children in the customer's household.
- **Is_Parent** (int64): A binary flag (1/0) indicating whether the customer has children living at home.

2. Engagement & Tenure Features

- **Recency** (int64): The number of days since the customer's last purchase.
- **Tenure_Years** (float64): The number of years the customer has been enrolled with the company.
- **Complain** (int64): A binary flag (1/0) indicating if the customer has lodged a complaint in the last 2 years.
- **Total_Accepted_Cmp** (int64): The total number of marketing campaigns (out of 6) the customer accepted and responded to.

3. Overall Spending & Activity Features

- **Total_Spent** (int64): The total amount of money the customer has spent across all product categories in the last 2 years.
- **Total_Purchases** (int64): The total number of purchases made across all channels (Web, Catalog, Store).
- **NumDealsPurchases** (int64): The number of purchases the customer made using a discount or deal.
- **NumWebVisitsMonth** (int64): The number of times the customer visited the company's website in the last month.

4. Product Preference Features (Ratios)

- **Wine_Ratio** (float64): The proportion of the customer's total spending dedicated to wine. (0 to 1)
- **Meat_Ratio** (float64): The proportion of the customer's total spending dedicated to meat products. (0 to 1)
- **Gold_Ratio** (float64): The proportion of the customer's total spending dedicated to gold products. (0 to 1)

5. Channel Preference Features (Ratios)

- **Web_Purchase_Ratio** (float64): The proportion of the customer's total purchases made through the company's website. (0 to 1)
- **Catalog_Purchase_Ratio** (float64): The proportion of the customer's total purchases made using a catalogue. (0 to 1)
- **Store_Purchase_Ratio** (float64): The proportion of the customer's total purchases made directly in physical stores. (0 to 1)

2.7 Data Encoding and Scaling

The final prepared dataset contained mixed data types: one ordinal categorical feature (Education), one nominal categorical feature (Marital_Status), and numerical features with different scales (ex. Income in the tens of thousands vs. Wine_Ratio between 0 and 1).

To prepare this data for clustering:

- **Education (Ordinal Encoding):** The Education feature, which has a natural hierarchy, was converted using ordinal encoding with the mapping: 'Undergraduate' = 0, 'Graduation' = 1, 'Master' = 2, 'PhD' = 3. This preserves the ordinal relationship between education levels.
- **Marital_Status (One-Hot Encoding):** The Marital_Status feature, being a nominal categorical variable without inherent order, was converted using one-hot encoding. This created separate binary columns for each category (ex. Marital_Status_Partner, Marital_Status_Single, etc.), preventing false ordinal relationships from being implied.

- **Standardization:** All numerical features were then standardized using the StandardScaler. This process transforms each feature to have a mean of 0 and a standard deviation of 1, ensuring that no single feature dominates the clustering algorithm's distance calculations due to scale differences.

The result is a fully numerical dataset (df_scaled) where all 20 features are on a common, comparable scale, making it the ideal input for the distance-based clustering algorithms to follow. The final dimensions of this processed dataset are 2,215 customers x 21 features.

3 Data Modeling: Clustering

3.1 Dimensionality Reduction with PCA

Before applying clustering algorithms, it is important to address the high dimensionality of the dataset. The preprocessed data contains 20 features. While all are informative, high-dimensional spaces present a significant challenge for distance-based algorithms like K-Means, which is often referred to as the "Curse of Dimensionality." In high dimensions, the concept of distance becomes less meaningful, as data points tend to be equally far apart, making it difficult for the algorithm to form tight, distinct clusters.

To mitigate this, Principal Component Analysis (PCA) was used. PCA is a dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables called Principal Components (PCs). These PCs are ordered so that the first few retain most of the variation present in the original dataset, which effectively denoises the data

The process was as follows:

1. **Initial PCA Fit:** A PCA model was fitted to the entire scaled dataset (df_scaled) without specifying the number of components. This was done to analyze the cumulative explained variance, which is the proportion of the dataset's total variance that is explained by each additional component.
2. **Explained Variance Analysis:** Below is a plot of the cumulative explained variance versus the number of components (Figure 2). This plot was used to determine the optimal number of components to retain. It shows the trade-off between dimensionality reduction and information retention.

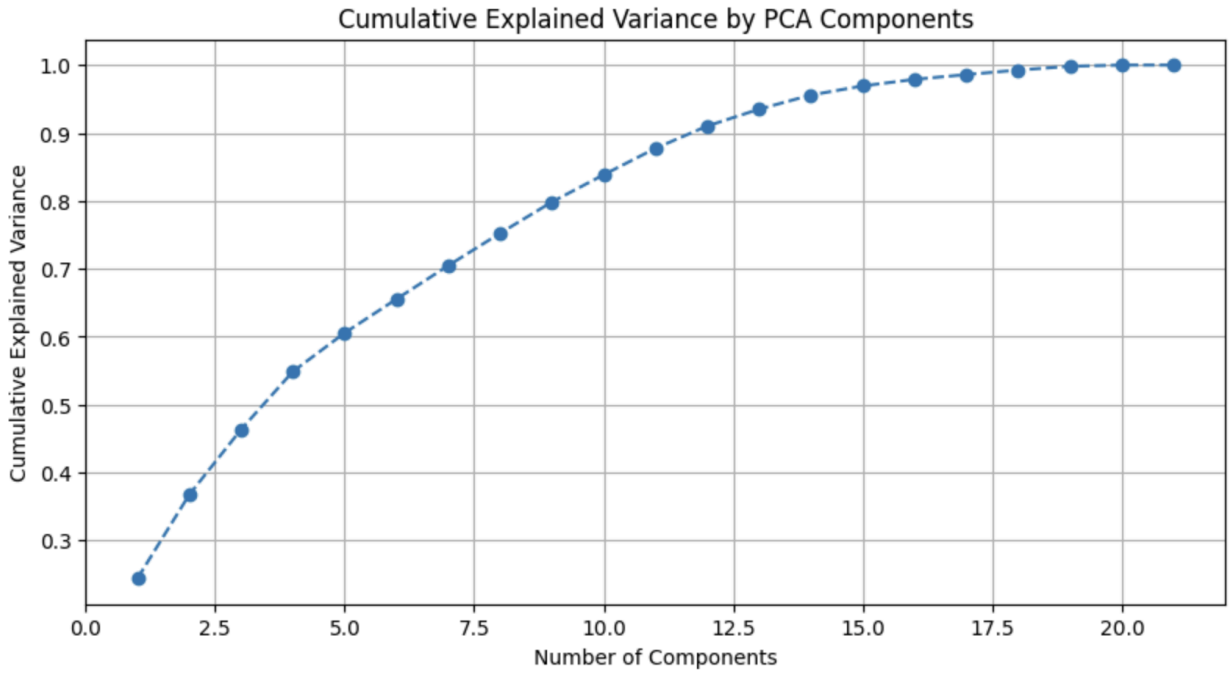

Figure 2 PCA Cumulative Explained Variance Ratio: The first 10 principal components were selected as they capture between 80-90% of the total variance in the original 20-feature dataset, providing an optimal balance between dimensionality reduction and information preservation for clustering analysis.

3. **Determining the Number of Components:** The standard approach is to choose a number of components that capture a sufficiently large amount of the variance (typically between 70% and 90%) while significantly reducing the number of dimensions. Based on the plot, the first 10 principal components were selected, as they collectively explain between 80% and 90% of the total variance in the data. This provides an excellent balance by preserving the majority of the information for clustering while reducing the computational complexity and mitigating the curse of dimensionality.
4. **Dimensionality Reduction:** The original 20-dimensional data was projected onto a new 10-dimensional subspace defined by the first ten principal components for the final clustering analysis.

While 10 dimensions are used for clustering, the data can be projected into 2 or 3 dimensions for visualization purposes only. This allows for a preliminary inspection of the data structure. The 3D and 2D plots below (Figure 3) show the projection onto the first three and two components, respectively. These visualizations, while not containing all the information, can possibly show initial patterns and potential groupings within the customer base.

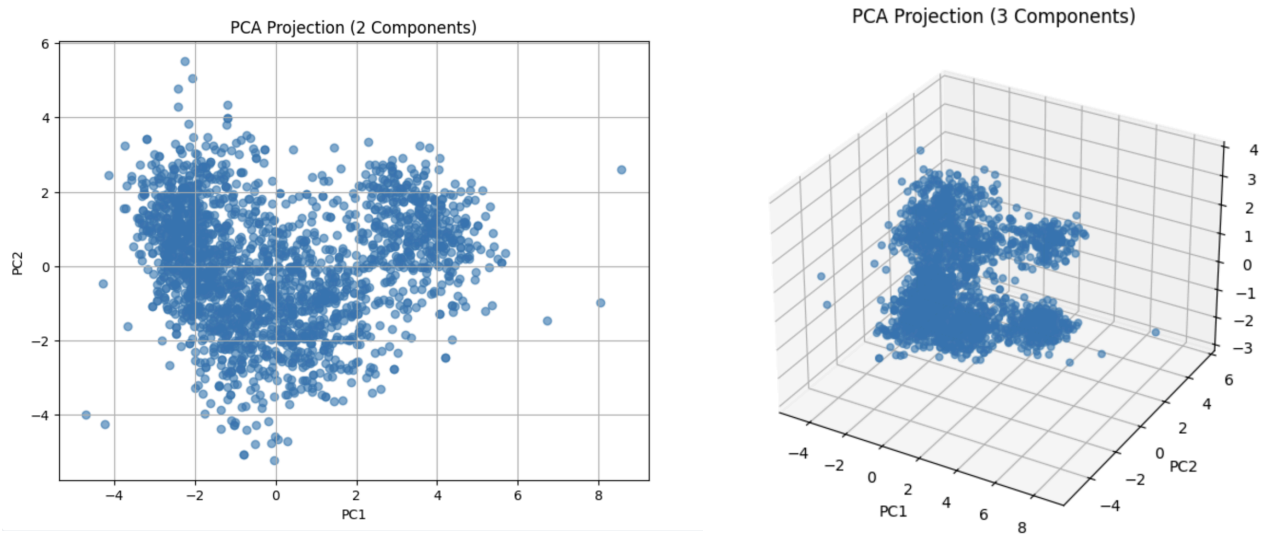

Figure 3 PCA Projection for Data Visualization: Two-dimensional (left) and three-dimensional (right) PCA projections of the customer data using the first principal components. While clustering was performed using 10 components, these visualizations provide an intuitive overview of potential groupings and data structure in reduced dimensional space.

3.2 Cluster Analysis with K-Means

With the data transformed into a lower-dimensional, de-noised space through PCA, the next step is to apply the K-Means clustering algorithm. The K-Means algorithm was selected for this customer segmentation task because of its computational efficiency, interpretability, and suitability for this project's objectives. K-Means is a centroid-based algorithm that aims to partition n observations into k clusters, where each observation belongs to the cluster with the nearest mean. It’s efficient and highly scalable, which makes it well-suited for datasets of this size (2,215 customers). This allows for rapid iteration when determining the optimal number of clusters. The algorithm produces easily interpretable results where each cluster is defined by its centroid, which is the mean of all points in the cluster. This makes profiling the resulting segments very straightforward. K-Means also performs well on spherical clusters of similar size, which is an assumption that becomes more valid after PCA preprocessing, which helps to normalize the structure of the data and mitigate the curse of dimensionality.

The upcoming steps in the modeling process will involve:

1. **Determining the Optimal Number of Clusters (k)**: Methods such as the Elbow Method (based on within-cluster sum of squares) and Silhouette Analysis are used to find the value of k that results in the most coherent and well-separated clusters.
2. **Fitting the K-Means Model:** Once k is determined, the K-Means algorithm will be run on the 10-dimensional PCA-transformed data to assign each customer to a cluster.

3. **Profiling and Interpreting Clusters:** The final and most important step will be to analyze the characteristics of each cluster by examining the original features of the customers within them. This will help create detailed profiles (e.g., "High-Income Big Spenders," "Budget-Conscious Parents," "Deal-Seeking Browsers") that can provide actionable insights for marketing strategies.

3.2.1 Determining the Optimal Number of Clusters

To identify the most appropriate number of customer segments (k), the Elbow Method and Silhouette Analysis were used.

- **Elbow Method:** This technique plots the Within-Cluster Sum of Squares (Inertia) against different values of k. The optimal k is typically found at the "elbow" of the curve, which is the point where the rate of decrease in inertia sharply changes. As shown in Figure 4, an elbow is observed at k=4.

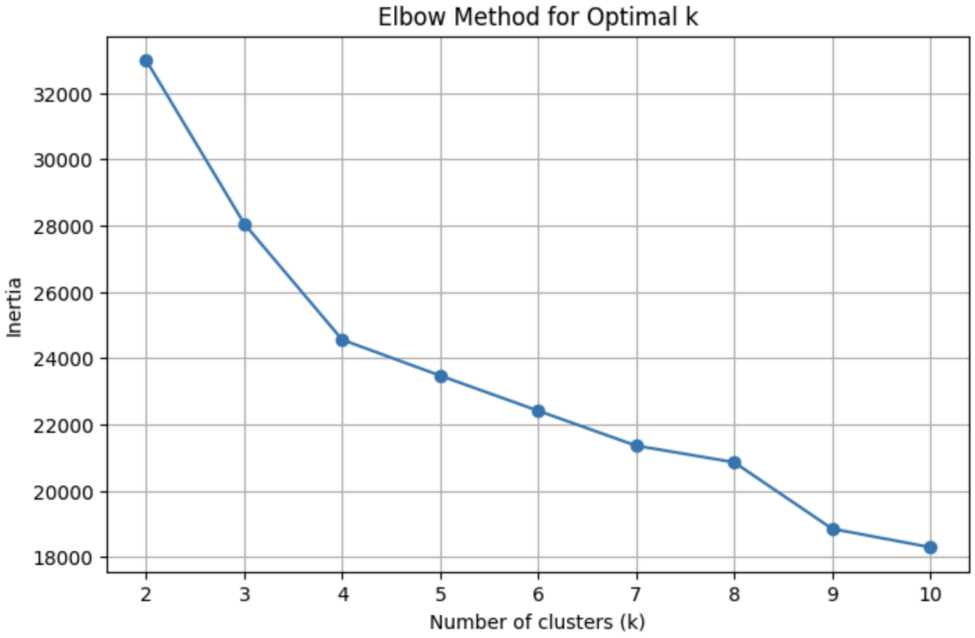

Figure 4 Elbow Method for Optimal Cluster Selection: Plot of inertia values for k ranging from 2 to 10 clusters. The optimal number of clusters was determined to be k=4, corresponding to the point where the rate of inertia reduction substantially decreases (the 'elbow').

- **Silhouette Analysis:** This method measures how similar a customer is to their own cluster compared to other clusters. Scores range from -1 to 1, where higher values indicate better-defined clusters. The analysis (Figure 5) showed that k=3 achieved the highest score (0.22). However, k=4 yielded a nearly identical score (approximately 0.215), with a difference of less than 0.01.

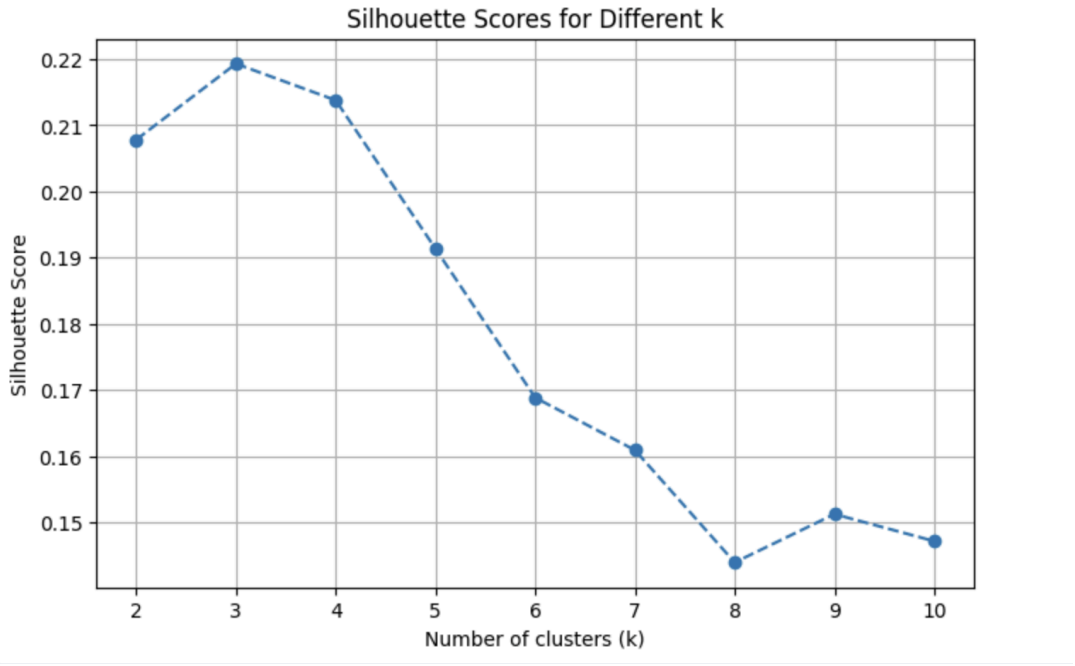

Figure 5 Silhouette Analysis for Cluster Validation: Silhouette analysis comparing cluster quality across k values from 2 to 10. Notice that k=3 achieved the highest score (0.22), but k=4 yielded a nearly identical score (0.215) with a difference of less than 0.01. The comparable performance at k=4, combined with the elbow method results, supported the selection of four customer segments.

Given the elbow at k=4 and the negligible difference in silhouette score, k=4 was selected as the optimal number of clusters. This choice makes sure that there will be well-defined clusters while allowing for a more granular and actionable segmentation of the customer base.

It is important to note that the overall silhouette scores obtained (ranging from 0.21 to 0.22) are considered relatively low. This indicates a weak clustering structure according to conventional interpretation guidelines, where scores above 0.5 are typically desirable. This suggests that the natural customer segments in the data are not extremely well-separated and may have overlapping characteristics. In the context of customer behavior analytics, this is often expected and even reasonable, because purchasing habits and demographics frequently exist on a spectrum rather than falling into perfectly distinct groups. The low scores reinforce that the identified clusters should be interpreted as useful, data-based groupings for strategic segmentation rather than as absolute, natural categories. The primary value lies in the actionable business insights these segments can provide, even with some degree of overlap.

3.2.2 Fitting the K-Means Model

After determining the optimal number of clusters, the K-Means algorithm was run on the 10-dimensional PCA-transformed data with n_clusters=4 and a fixed random state for reproducibility. Each of the 2,215 customers was assigned to one of the four distinct segments. The resulting clusters are visualized in the reduced 2D and 3D PCA spaces in Figure 6. While these visualizations only show

the first two or three principal components, they confirm that the algorithm identified well-separated groups within the dominant dimensions of the data.

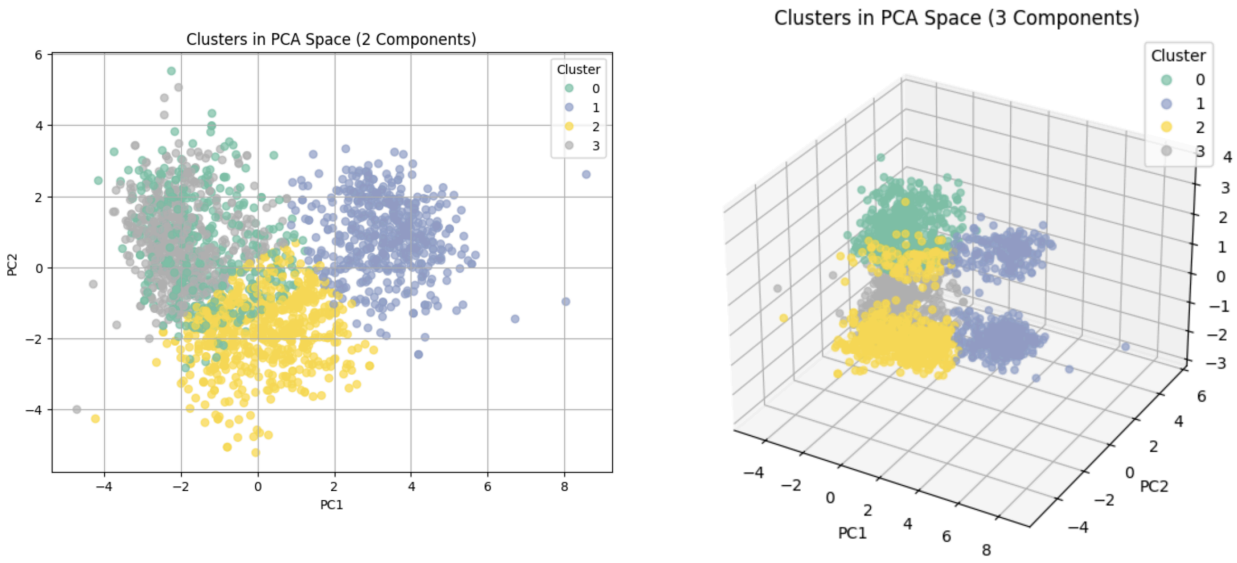

Figure 6 Final K-Means Clustering Results: Final customer segments projected onto 2D (left) and 3D (right) PCA space. The K-Means algorithm identified four distinct clusters using 10 principal components.

4 Customer Profiling

4.1 Introduction to Customer Segments

After fitting the K-Means model to the pca data, four distinct customer segments were identified. The main goal of this segmentation is to understand the unique behavioral and demographic patterns that define each group. Profiling these clusters turns raw data into actionable business insights, which can be used to inform targeted marketing strategies and personalized communication.

The table below provides a summary of the characteristics that define each cluster. This profile serves as a foundational reference that compares the segments across different characteristics such as demographics, spending habits, product and channel preferences, and overall engagement.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Demographics	**Income** ≈ $39K **Age** ≈ 44 **Family Size** ≈ 2.2, ~89% parents	**Income** ≈ $77K (highest) **Age** ≈ 46 **Family Size** ≈ 1.1 (mostly singles/couples), only 6% parents	**Income** ≈ $59K **Age** ≈ 49 (oldest) **Family Size** ≈ 2.3, 97% parents	**Income** ≈ $33K (lowest) **Age** ≈ 41 (youngest) **Family Size** ≈ 2.2, 87% parents
Spending/Activity	**Low spenders** ≈ $211 avg **Few purchases** ≈ 8.1 **High website visits** ≈ 6.3 web visits/month	**Highest spenders** ≈ $1,367 avg **Most purchases** ≈ 19 **Lowest website visits** ≈ 2.6 web visits/month	**Strong spenders** ≈ $753 avg **Above avg. purchases** ≈ 16.7 **Moderate website visits** ≈ 5.8 web visits/ month .	**Lowest spenders** ($101 avg) **Fewest purchases** (5.9) **Highest website visits** ≈ 6.5 web visits/month (high browsing, low conversion)
Product Preference	**Wine** ~40%, fairly balanced with meat, low gold .	**High wine share** (43%), modest gold (~12%)	**Wine-dominant** (64% of spend, highest), very low meat/gold .	**Wine** ~34% Higher share in meat & "other" (essentials), very low gold
Channel Preference	**Dominated by store purchases** (≈60%) .	**Catalog dominates** (~32%)	**Balanced** mix of web (37%), store (44%), some catalog .	**Store-heavy** (≈60%)
Engagement	**Very recent purchasers** (lowest recency, most engaged)	**Average tenure** **Responsive to campaigns** **Low complaint rate**	**Longest tenure** **Least recent purchases** (loyal but less active)	**Short tenure** **Low campaign response**
Interpretation	Budget-conscious families shopping in-store, moderate engagement, low luxury spending.	Wealthy, non-parent professionals; prefer catalogs, spend heavily, value wine; prime for premium offers.	Older, established families; loyal wine buyers; multichannel users; lower campaign response but stable base.	Younger budget-conscious families; shop mostly in-store; browse online but rarely buy; deal-oriented.

4.2 Cluster 0: The Family Store Shoppers

This cluster is made up of middle-income families, often in their forties, who typically have children at home. They prefer to shop in stores rather than online or through catalogs, and while they buy fairly regularly, their overall spending is modest compared to other groups. Their purchases cover a broad mix of products, with little focus on luxury items. Despite their lower spending, these customers are very engaged and tend to shop recently. Although their high engagement and recent activity make them a reliable customer base, their cost-consciousness suggests a high sensitivity to price and value.

Businesses can keep their loyalty by offering in-store promotions, family discounts, and bundled deals that appeal to their cost-conscious nature. The targeted in-store promotions can utilize point-of-sale discounts and weekly specials to incentivize purchases. Family-oriented bundles and discounts can offer perceived savings and meet household needs. Loyalty rewards for frequent visits can also be used to reinforce their shopping habits and increase lifetime value.

Businesses can keep the loyalty of this cost-conscious segment by tailored to their preference for value and in-store shopping. Recommended actions to look further into include:

- Offer **in-store promotions** and **family discounts** to appeal to their price sensitivity.
- Create **bundled deals** (e.g., family meal packs) that maximize value.
- Use **loyalty programs** with small, frequent rewards to maintain engagement.
- Promote **basic and everyday essentials** over luxury products.

By focusing on these value-centric strategies, businesses can effectively nurture this engaged and reliable customer base.

4.3 Cluster 1: The Catalog Connoisseurs

This cluster represents the wealthiest group in the dataset. These customers tend to be singles or couples without children in their mid-forties. They spend the most money overall and show a particular interest in wine and occasional luxury items. Unlike other clusters, they prefer catalog shopping over online or in-store purchases, which sets them apart. They are also more responsive to marketing campaigns and rarely complain, which makes them an attractive audience.

To maximize value from this group, businesses can focus on quality and convenience. Recommended actions to look further into include:

- Send **exclusive premium** catalogs with personalized product suggestions.
- Promote **wine collections and luxury bundles** that match their tastes.
- Introduce **VIP loyalty perks** (priority shipping, early access to promotions)
- Experiment with **direct-mail** offers since they engage strongly with catalog marketing.

By focusing on these premium and personalized strategies, businesses can effectively nurture this highly profitable and responsive customer base.

4.4 Cluster 2: The Loyal Wine Enthusiasts

Cluster 2 consists mostly of older, established families in their late forties and fifties who almost always have children at home. These customers are long-tenured and loyal, though their purchases tend to be less recent compared to others. They spend a significant amount, with wine making up the majority of their purchases, while meat, gold, and other luxury categories play a much smaller role. They use a balanced mix of shopping channels; this shows they are comfortable across platforms.

To keep this valuable group engaged, businesses should focus on efforts to encourage them to purchase more frequently. Recommended actions to look further into include:

- Introduce **wine club memberships** that reward recurring purchases.
- Offer **subscription services** (monthly curated wine boxes, loyalty tastings).
- Design **reactivation campaigns** (special offers if inactive for 60+ days).
- Highlight **cross-selling opportunities** (e.g., cheese or premium food pairings with wine).

4.5 Cluster 3: The Young Budget Browsers

This cluster represents the youngest segment, averaging around forty years old, with lower incomes and families to support. They are heavy store shoppers who focus on essentials like meat and basic goods, spending very little on luxury items such as wine or gold. Interestingly, they visit the company's website often but rarely make online purchases, suggesting they like to browse but are hesitant to commit. Their short tenure and low responsiveness to campaigns make them harder to engage, but not impossible.

Businesses can target this group with efforts to turn their browsing into actual buying.

- Launch **web-to-store coupons** to convert browsing into physical purchases.
- Use **flash sales and time-limited discounts** to drive quick action.
- Target with **affordable bundles** (meat packs, family staples).
- Test **digital promotions through mobile apps or emails** that highlight value.

4.6 Conclusion

Overall, the clusters reveal four very different types of customers, each with unique needs and shopping behaviors. Family Store Shoppers are dependable but price-sensitive, responding best to in-store value offers. Catalog Connoisseurs represent the wealthiest segment, with high spending habits and a preference for premium catalog experiences. Loyal Wine Enthusiasts are long-term, stable customers whose purchases are dominated by wine, making them perfect candidates for specialized loyalty programs. Finally, Young Budget Browsers are the most cost-conscious and digitally curious, browsing online but mainly purchasing essentials in-store. By tailoring strategies to each group, whether that is through family promotions, premium catalogs, wine subscriptions, or digital discounts, the company can engage customers more effectively and maximize long-term value.

5 Conclusion

5.1 Conclusion

This project applied clustering techniques to segment customers into meaningful groups based on their demographics, spending behavior, and channel preferences. After preprocessing, feature engineering, and dimensionality reduction with PCA, K-Means clustering revealed four distinct customer profiles. These included family-oriented store shoppers, affluent catalog-driven spenders, wine-focused loyalists, and younger budget-conscious browsers.

Although the silhouette scores indicated moderate overlap between clusters, the resulting segmentation still provides clear, actionable insights. Each cluster aligns with real-world behavioral patterns that businesses can use to design personalized marketing campaigns, loyalty programs, and promotional strategies. Overall, this project shows the value of data-driven segmentation in turning raw customer data into actionable insights.

5.2 Limitations

While the project achieved its objectives, several limitations should be acknowledged:

- **Cluster Quality:** The silhouette scores (0.21–0.22) indicate that clusters are not strongly separated. This suggests that customer behaviors exist on a spectrum, and the clusters should be viewed as practical groupings rather than perfectly distinct categories.
- **Feature Selection Bias:** The engineered features may emphasize certain behaviors (e.g., spending ratios, purchase channels) more heavily than others. This focus may have downplayed other potential behavioral patterns. Different transformations could lead to slightly different segments.
- **Temporal Limitations:** The dataset is static and reflects customer behavior at one point in time. Customer preferences and spending habits may change over time, requiring re-segmentation.
- **Algorithm Choice:** Only K-Means was applied. While effective and interpretable, alternative clustering methods, such as hierarchical clustering, DBSCAN and Gaussian Mixture Models might be able to capture more nuanced relationships.

These limitations highlight that clustering is an exploratory tool, best used for guiding strategic thinking rather than producing rigid categories.