Customer personality analysis

About the Dataset

- **Context**
  Customer Personality Analysis helps businesses **understand and target** customer segments effectively. By analyzing customer behavior, companies can tailor products and marketing efforts to specific groups, saving resources and boosting effectiveness. For example, instead of marketing to all customers, a business can focus on segments most likely to purchase.

About the Dataset

- **Dataset Attributes**

1. **Demographics**:
   - ID, Year_Birth, Education, Marital_Status, Income
   - Household composition: Kidhome, Teenhome
2. **Engagement**:
   - Dt_Customer (enrollment date), Recency (days since last purchase), Complain

About the Dataset

- **Dataset Attributes**

3. **Spending**:
   - Amount spent on Wines, Fruits, Meat, Fish, Sweets, Gold (last 2 years)**Promotions**:
   - Deals and Campaign Responses (AcceptedCmp1-5, Response)

4. **Purchasing Channels**:
   - Web, Catalog, Store purchases, Web visits

About the Dataset

- **Goal**:
  Cluster customers into segments to enable targeted marketing and product optimization.

Sections

- Data Preprocessing(Feature Engineering, Data Cleaning and Exploratory Data Analysis)
- Standardization
- Clustering
- PCA
- Selecting a model
- Visualizing different features based on clusters
- Observations

Preprocessing

- check number of unique values for each feature

```
df.nunique()

ID                     2240
Year_Birth               59
Education                 5
Marital_Status            8
Income                 1974
Kidhome                   3
Teenhome                  3
Dt_Customer             663
Recency                 100
MntWines                776
MntFruits               158
MntMeatProducts         558
MntFishProducts         182
MntSweetProducts        177
MntGoldProds            213
```

```
NumDealsPurchases      15
NumWebPurchases        15
NumCatalogPurchases    14
NumStorePurchases      14
NumWebVisitsMonth      16
AcceptedCmp3            2
AcceptedCmp4            2
AcceptedCmp5            2
AcceptedCmp1            2
AcceptedCmp2            2
Complain                2
Z_CostContact           1
Z_Revenue               1
Response                2
dtype: int64
```

Preprocessing

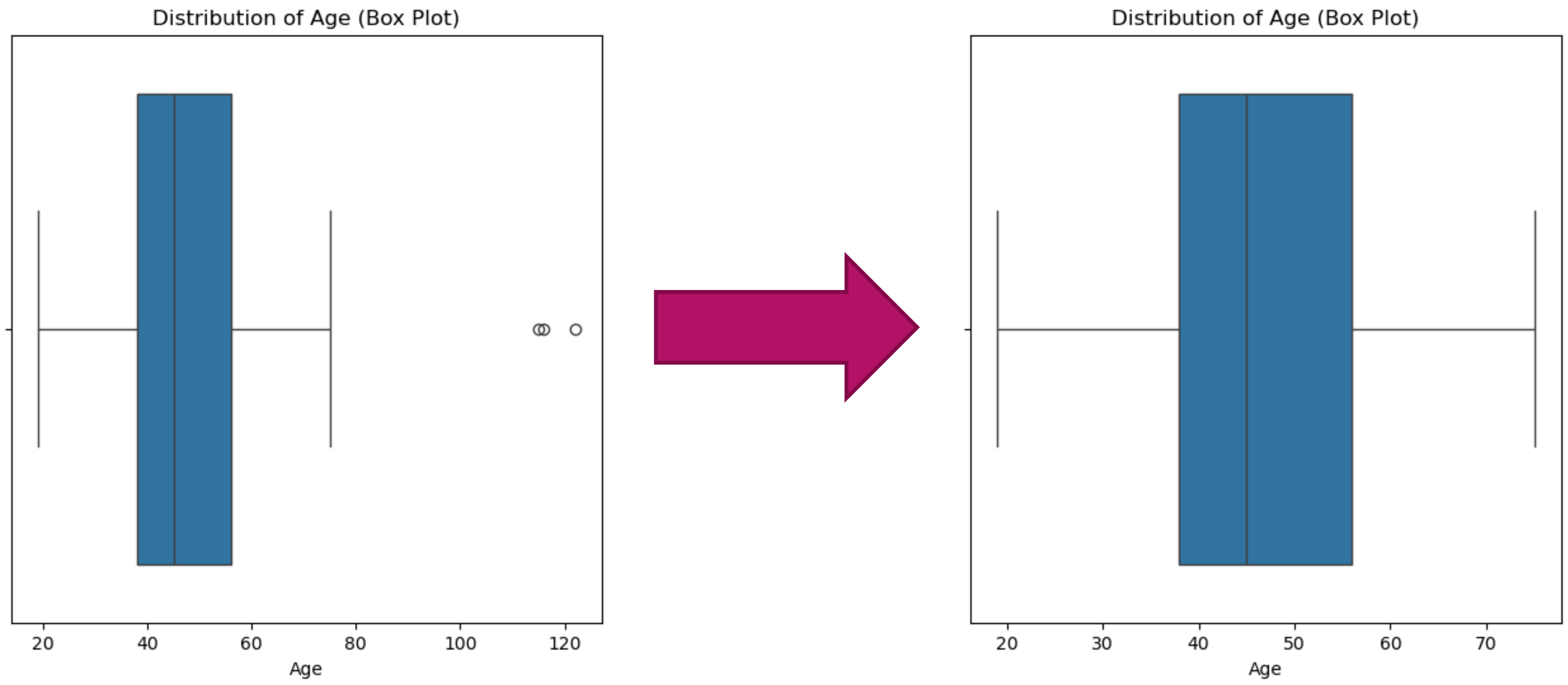
- Replaced 24 missing income values with the mean.
- Converted **Dt_Customer** to **date format** and calculated customer tenure in months (*assuming 2016 as the reference year*).
- Calculated **Age** from Year_Birth.
- **TotalSpent**: Sum of all spending categories.
- **TotalAcceptedCmp**: Total offers accepted across campaigns.
- **NumTotalPurchases**: Total purchases across all channels.
- **Children**: Sum of Kidhome and Teenhome.

Preprocessing

- Education: Hierarchical encoding (Basic → 0, ..., PhD → 3).
- Marital_Status: Encoded as numeric categories (Married → 0, ..., Widow → 4).
- Complain column has only 20 complaints out of 2230 records, making it unsuitable for clustering due to potential noise.

Preprocessing

- Since only **three datapoints** are far from the normal range, we **drop** them to avoid complexities and potential biases



Preprocessing

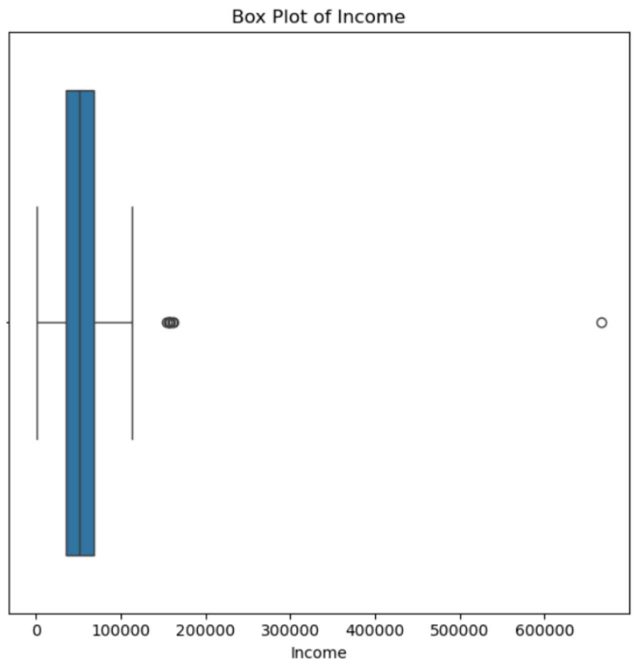
```
col_drop = ["AcceptedCmp1" , "AcceptedCmp2", "AcceptedCmp3" , "AcceptedCmp4","AcceptedCmp5", "Response", # used them to create TotalAcceptedCmp
           "NumWebVisitsMonth", "NumWebPurchases","NumCatalogPurchases","NumStorePurchases", "NumDealsPurchases", # used them to create NumTotalPurchases
           "Kidhome", "Teenhome", # used them to create children
           "MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", "MntGoldProds", # used them to create TotalSpent
           "Year_Birth","Dt_Customer", # used to obtain Age and Months_Since_Registration
           "Complain",
           "ID" # irrelevant for clustering
           ]
```

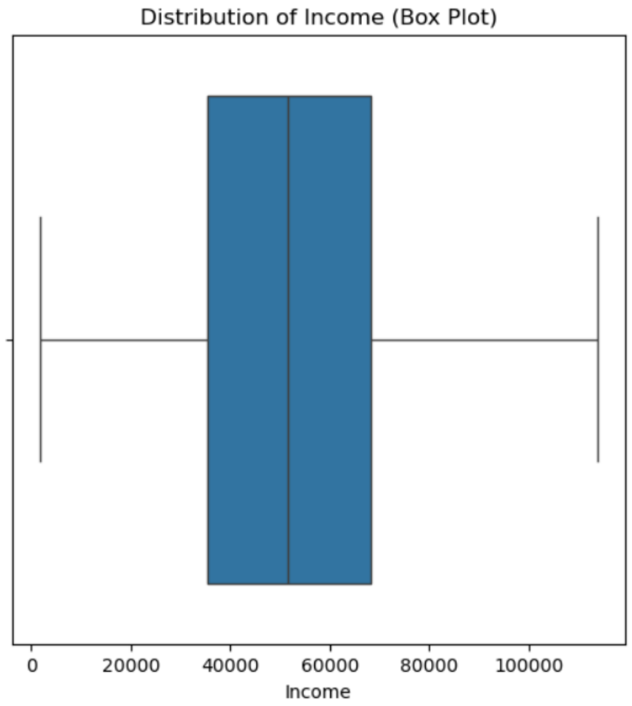
Preprocessing

	Education	Marital_Status	Income	Recency	Months_Since_Registration	Age	TotalSpent	TotalAcceptedCmp	TotalPurchases	Children
0	1	2	58138.0	58	45	58	1617	1	25	0
1	1	2	46344.0	38	17	61	27	0	6	2
2	1	1	71613.0	26	29	50	776	0	21	0
3	1	1	26646.0	26	15	31	53	0	8	1
4	3	0	58293.0	94	24	34	422	0	19	1

Preprocessing

- Since only **Eight datapoints** are far from the normal range, we **drop** them to avoid complexities and potential biases





Standardization

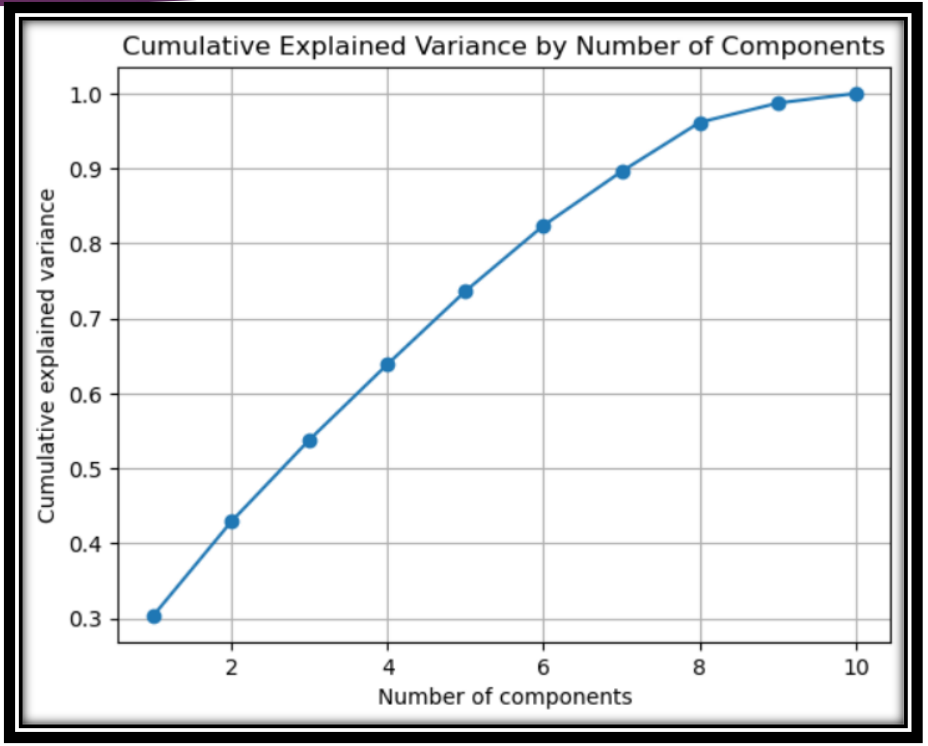
- **Scaling** features before clustering ensures that all variables contribute equally to the analysis. Clustering algorithms like K-Means use distance metrics (e.g., Euclidean distance) to group data points, so large-valued features(like income) can dominate smaller ones if not scaled.
- By standardizing (scaling to mean 0 and variance 1), we make the features comparable, improving the clustering's accuracy and interpretability.

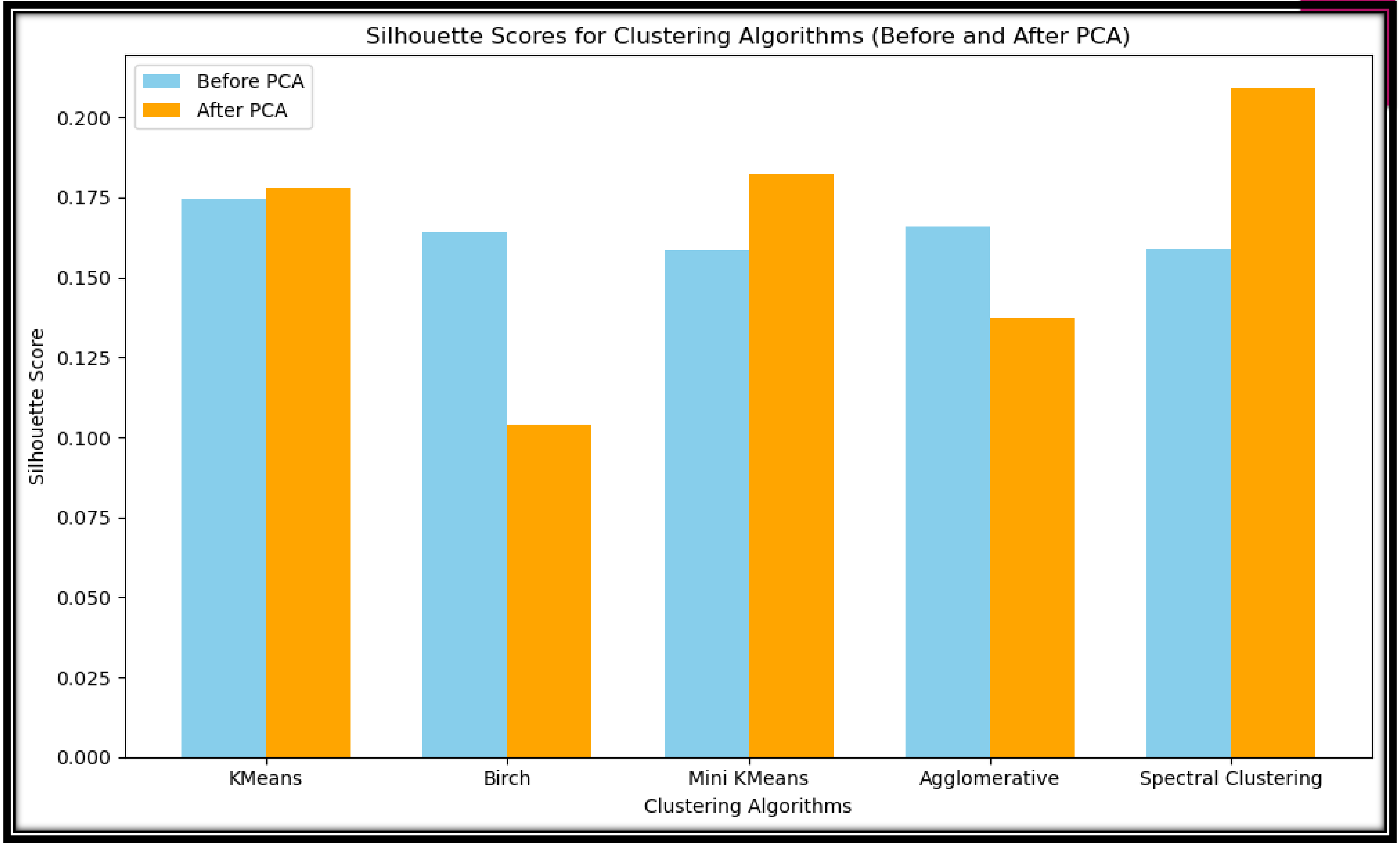
Clustering and PCA

- In this project, we will use various clustering algorithms to group customers based on their behaviors.
- These algorithms include **KMeans**, **Birch**, **Mini KMeans**, **Agglomerative Clustering**, and **Spectral Clustering**. For each algorithm, we will evaluate performance with and without dimensionality reduction using PCA.
- While we assume we want three clusters for this analysis, we still use the elbow method with KMeans to explore the optimal number of clusters.

Clustering and PCA

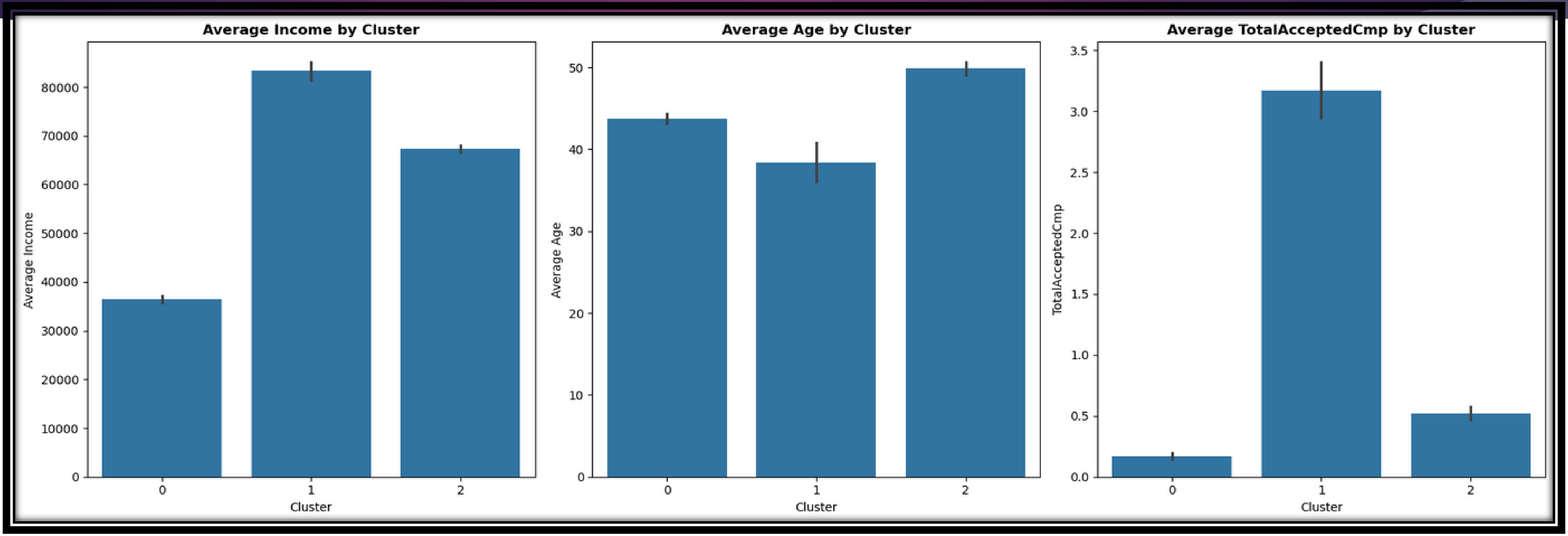
- **Component Analysis (PCA)** is a technique used to reduce the number of features in a dataset while preserving as much of the original information as possible.
- PCA is often used for simplifying data, visualizing high-dimensional data, and improving the performance of machine learning algorithms.



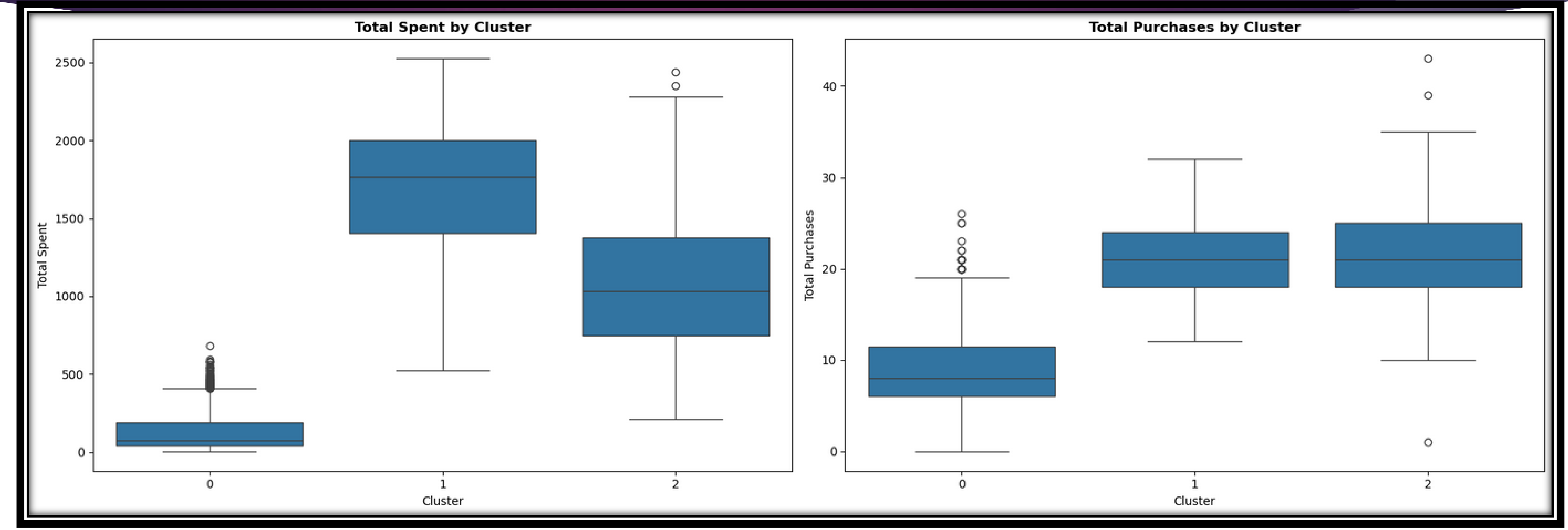


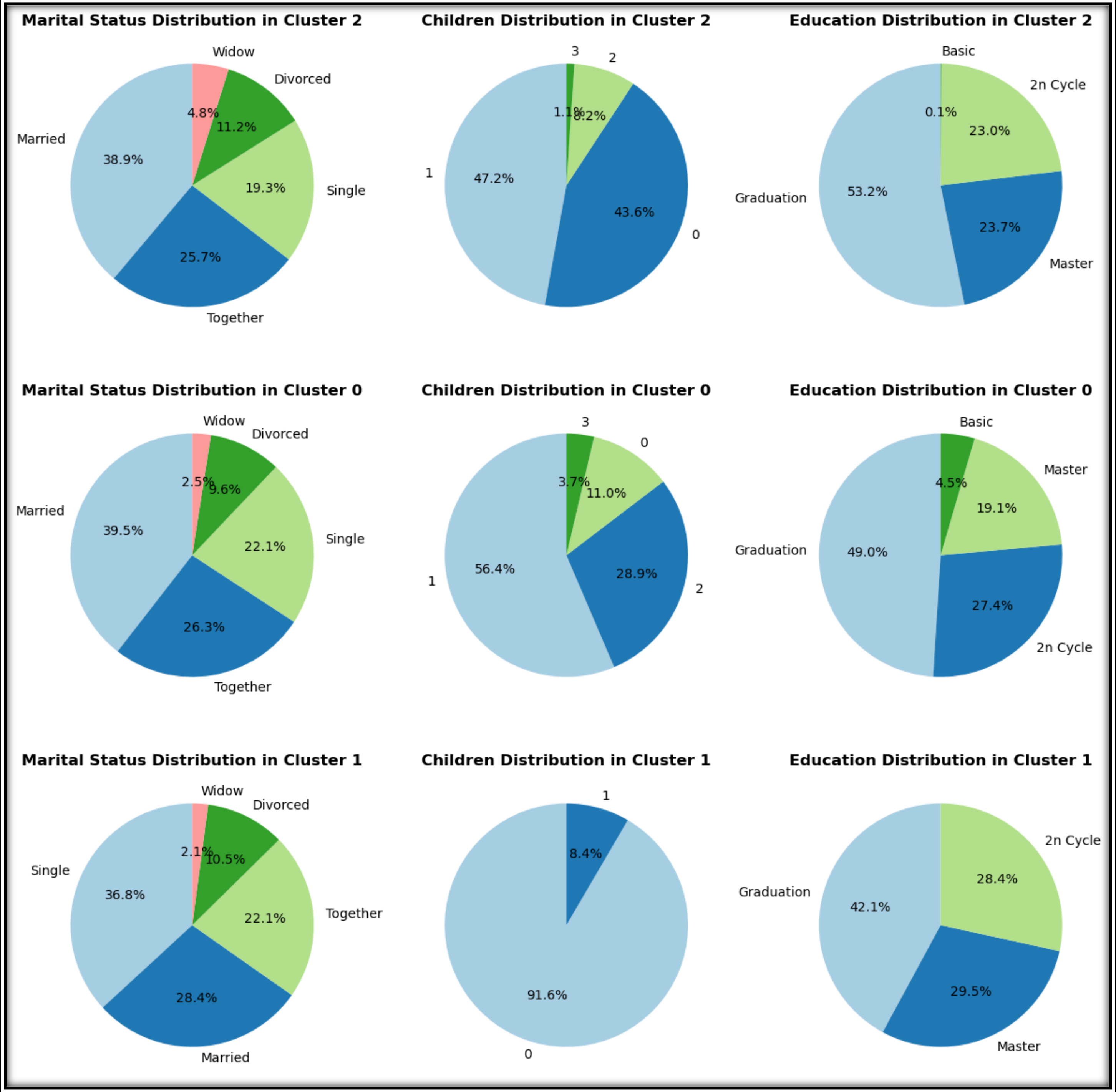
Silhouette Scores for Clustering Algorithms (Before and After PCA)
Before PCA
After PCA
Silhouette Score
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
KMeans
Birch
Mini KMeans
Agglomerative
Spectral Clustering
Clustering Algorithms

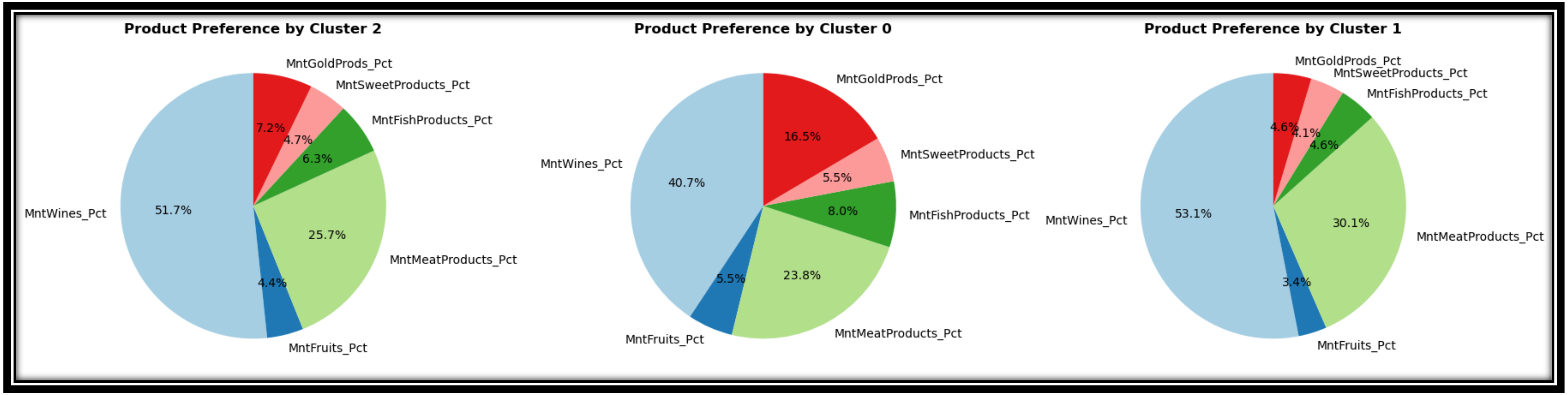
Visualization



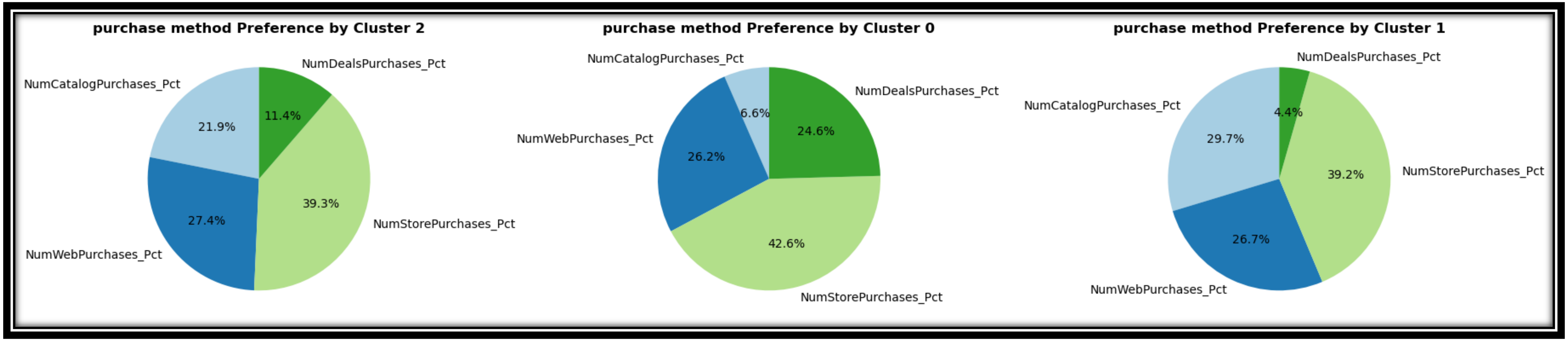
Visualization




Marital Status Distribution in Cluster 2
Widow
Divorced
Married
Single
Together
4.8%
11.2%
38.9%
19.3%
25.7%
Children Distribution in Cluster 2
3
2
1
0
1.1%
8.2%
47.2%
43.6%
Education Distribution in Cluster 2
Basic
2n Cycle
Graduation
Master
0.1%
23.0%
53.2%
23.7%
Marital Status Distribution in Cluster 0
Widow
Divorced
Married
Single
Together
2.5%
9.6%
39.5%
22.1%
26.3%
Children Distribution in Cluster 0
3
0
1
2
3.7%
11.0%
56.4%
28.9%
Education Distribution in Cluster 0
Basic
Master
Graduation
2n Cycle
4.5%
19.1%
49.0%
27.4%
Marital Status Distribution in Cluster 1
Widow
Divorced
Single
Together
Married
2.1%
10.5%
36.8%
22.1%
28.4%
Children Distribution in Cluster 1
1
0
8.4%
91.6%
Education Distribution in Cluster 1
2n Cycle
Graduation
Master
28.4%
42.1%
29.5%



Product Preference by Cluster 2
MntGoldProds_Pct
MntSweetProducts_Pct
MntFishProducts_Pct
MntWines_Pct
MntMeatProducts_Pct
MntFruits_Pct
7.2%
4.7%
6.3%
51.7%
25.7%
4.4%
Product Preference by Cluster 0
MntGoldProds_Pct
MntSweetProducts_Pct
MntFishProducts_Pct
MntWines_Pct
MntMeatProducts_Pct
MntFruits_Pct
16.5%
5.5%
8.0%
40.7%
23.8%
5.5%
Product Preference by Cluster 1
MntGoldProds_Pct
MntSweetProducts_Pct
MntFishProducts_Pct
MntWines_Pct
MntMeatProducts_Pct
MntFruits_Pct
4.6%
4.1%
4.6%
53.1%
30.1%
3.4%



purchase method Preference by Cluster 2
NumDealsPurchases_Pct
NumCatalogPurchases_Pct
11.4%
21.9%
39.3%
27.4%
NumStorePurchases_Pct
NumWebPurchases_Pct
purchase method Preference by Cluster 0
NumCatalogPurchases_Pct
NumDealsPurchases_Pct
6.6%
24.6%
NumWebPurchases_Pct
26.2%
42.6%
NumStorePurchases_Pct
purchase method Preference by Cluster 1
NumDealsPurchases_Pct
NumCatalogPurchases_Pct
4.4%
29.7%
39.2%
NumStorePurchases_Pct
26.7%
NumWebPurchases_Pct

Observations

- **Cluster 0** (Low income, Family-oriented):

- **Income**: Lowest income; price-sensitive.
- **Age**: Middle-aged; high number of children.
- **Spending**: Low; prefers gold, fish, and sweets; shops on deals.
- **Education**: Least educated.
- **Engagement**: Low campaign response.

Observations

- **Cluster 1** (High income, Young professionals):

- **Income**: Wealthiest; less price-sensitive.
- **Age**: Youngest; few/no children.
- **Spending**: High; prefers wine and meat; active catalog shoppers.
- **Education**: Most educated.
- **Engagement**: Highest campaign response.

Observations

- **Cluster 2** (Mid income, Older demographic):

  - **Income**: Middle-income; balanced spending.
  - **Age**: Oldest; widows/divorced common.
  - **Spending**: Moderate; wine and meat preferred.
  - **Education**: Well-educated.
  - **Engagement**: Low campaign response.

The End