Customer Personality Analysis

0.1 Background

In today’s highly competitive market, understanding customer behavior and preferences is crucial for businesses aiming to enhance customer satisfaction, loyalty, and profitability. Traditional demographic-based segmentation often fails to capture the complexity of customer needs and behaviors. A more sophisticated approach, leveraging machine learning and clustering techniques, can provide deeper insights by analyzing various customer attributes to identify distinct customer personas.

0.2 Objective

The objective of this project is to develop an unsupervised machine learning model to cluster customers based on their purchasing behavior, demographics, and engagement metrics. These clusters will help the marketing team to tailor strategies for each segment, ultimately aiming to increase customer retention, satisfaction, and lifetime value.

1. Exploratory Data Analysis

I have obtained the data set and now it is time to perform an Exploratory data Analysis (EDA) to gain insight about the dataset and prepare the data for modeling purposes.

1.1 About the Dataset

The dataset is the historical sales data that contains sales from 2010-12-01 to 2011-12-09 which is a total of 699 days. It has 29 features and 2,240 observations. The following are a short information of our feature :

People:
ID: Customer's unique identifier
Year_Birth: Customer's birth year
Education: Customer's education level
Marital_Status: Customer's marital status
Income: Customer's yearly household income
Kidhome: Number of children in customer's household
Teenhome: Number of teenagers in customer's household
Dt_Customer: Date of customer's enrollment with the company
Recency: Number of days since customer's last purchase
Complain: 1 if the customer complained in the last 2 years, 0 otherwise

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

Promotion:
NumDealsPurchases: Number of purchases made with a discount
AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place of Purchase:
NumWebPurchases: Number of purchases made through the company's website
NumCatalogPurchases: Number of purchases made using a catalog
NumStorePurchases: Number of purchases made directly in stores
NumWebVisitsMonth: Number of visits to company's website in the last month

1.2 Data Preprocessing And Feature Engineering

1.2.1 Data Preprocessing

The data preprocessing includes the following steps:

- Cleaning and handling missing values.
- Capping Outlier using IQR technique
- Dropping unnecessary features

1.2.2 Feature Engineering:

During the feature engineering process I use the original features to create some new relevant features that capture customer behavior and engagement. These features were obtained by using existing customer data and accurately reflect the values that these features should contain for the given customer. These features add a whole new and better dimensionality by which a customer cluster can be identified. A total of 7 features were derived and the following are these features:

- **Age** : The Age of the customer derived from Year_Birth
- **Total_Spent** : The total spent by the customer which is a sum of all features in the "products" category.
- **Children** : The total number of children the customer has.

- **living_with** : The value is 0 if the customer lives alone and is 1 if the customer lives with others
- **family_size** : The size of the customer's family.
- **Is_Parent** : The value is 0 if the customer is not a parent and is 1 if the customer is a parent
- **Married** : The value is 0 if the customer is not married and is 1 if the customer is a married

2. Data distribution and Descriptive statistic Visualization

The following diagrams show the distribution of the four major features "Age", "Income", "Total_Spent", and "Education" through pair plots and histograms. These visualizations are crucial for understanding our customer base.

2.1 Histogram

The diagram below shows the distribution of our three features with education as the hue.

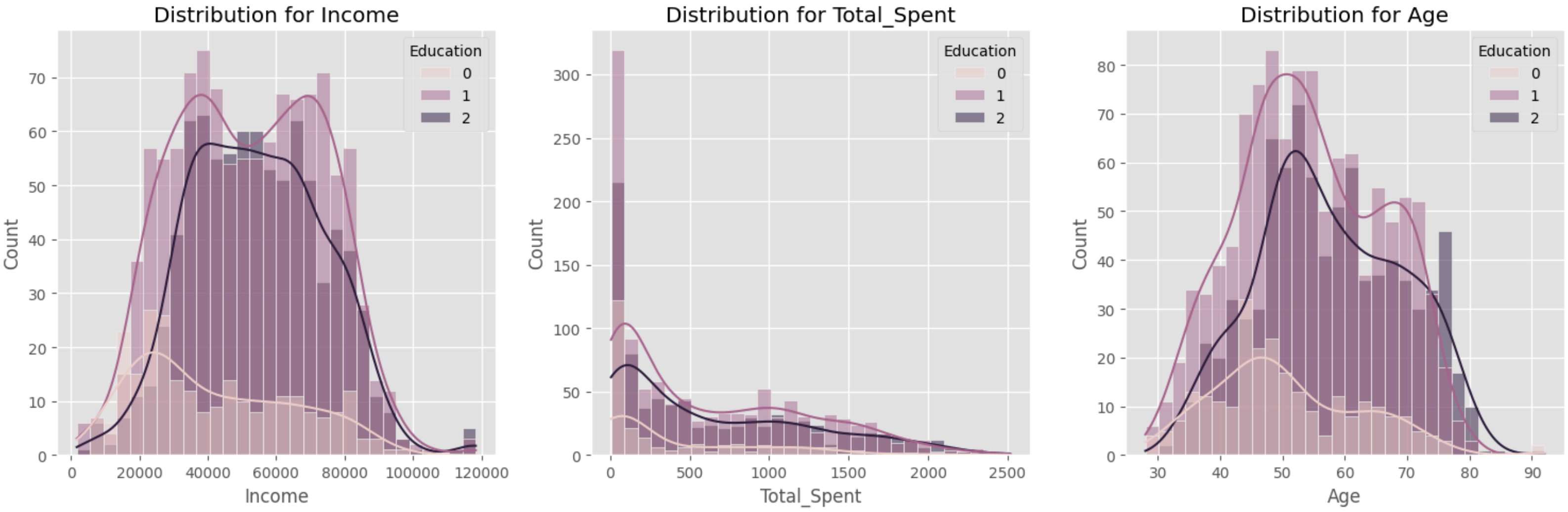

Figure 1. Histogram

2.2 Pair plot

The diagram below shows the pair plot of our three features with education as the hue. There are KDE plots and scatter plots for each feature against the other. This helps us observe the relationship between these features.

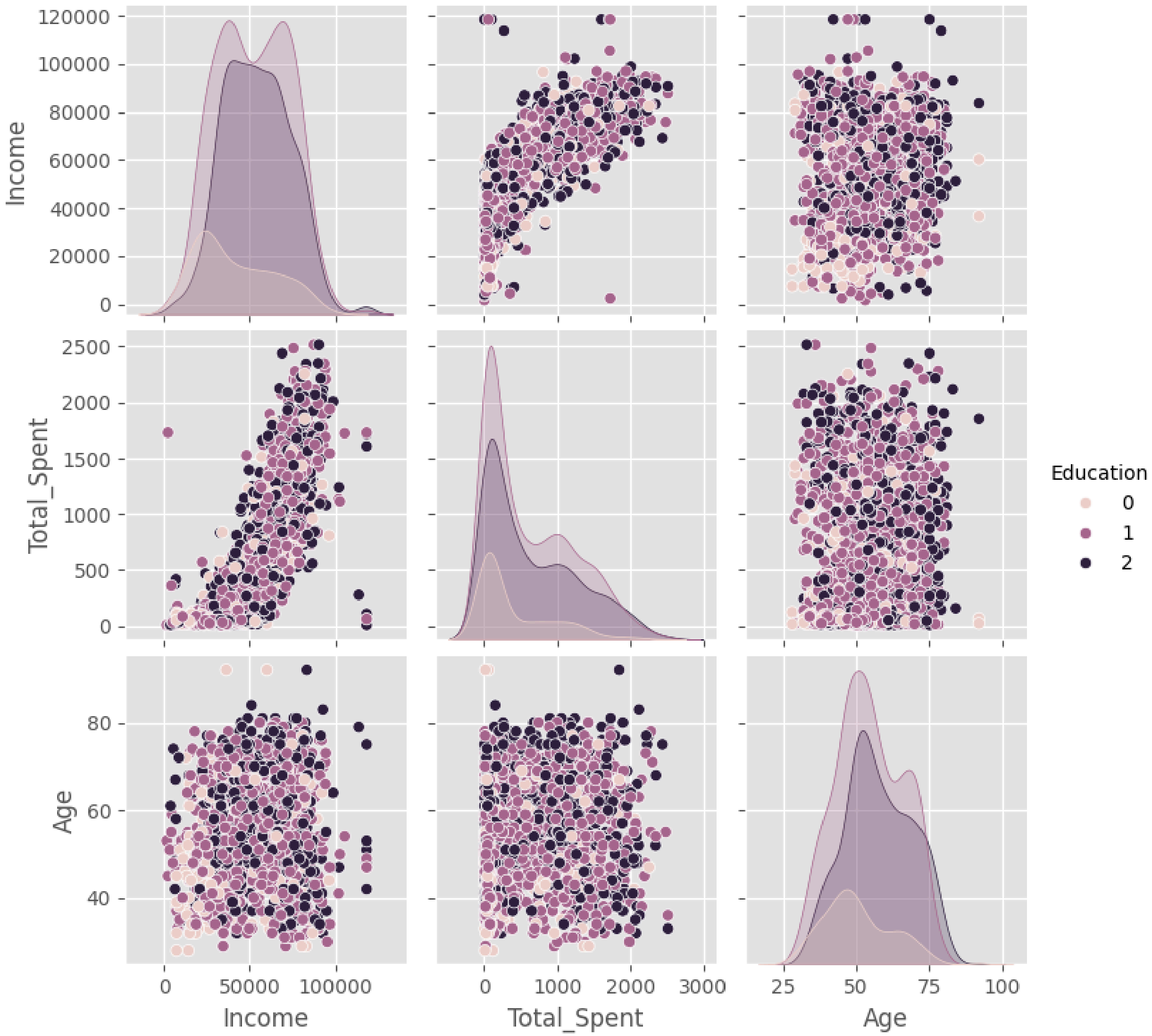

Figure 2. Pair Plot

3. Machine Learning Approach

I will be implementing and comparing K-means, DBSCAN, and K medoid clustering to cluster our dataset. I have determined the optimal number of clusters using methods like the Elbow method, Silhouette score

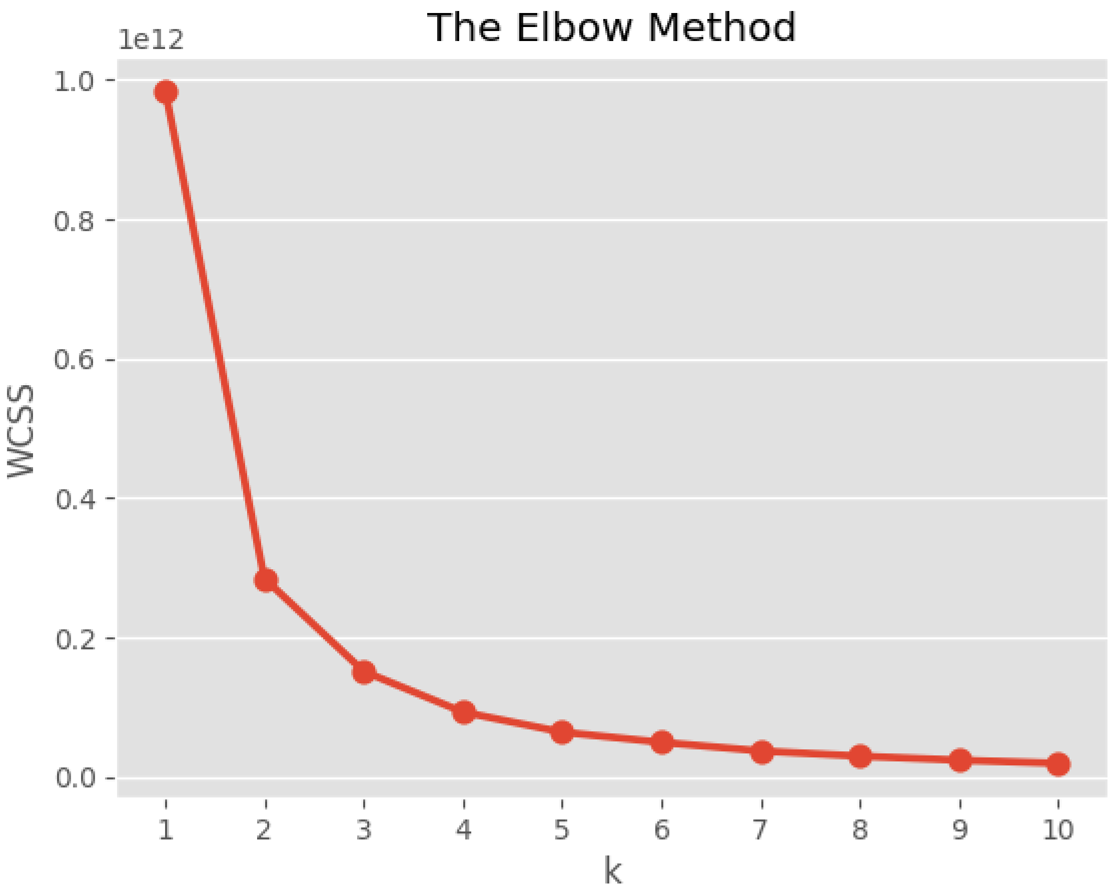

Figure 3. The Elbow Method

4. Evaluation

The following are the Silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Index for our models:

4.1 Metrics Overview:

Silhouette Score:

The Silhouette score is a metric used to evaluate the quality of a clustering algorithm. It provides a measure of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, with higher values indicating better-defined clusters.

Silhouette Coefficient Calculation:

- **a(i):** The average distance from the i^{th} point to all other points in the same cluster. This represents how well the point is clustered with its own cluster (lower is better).

- **b(i):** The average distance from the i^{th} point to all points in the nearest cluster that it is not a part of. This represents how well the point is separated from the nearest other cluster (higher is better).

The Silhouette score for a single point is given by:

$$s(i) \ = \ \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Overall Silhouette Score: The average of the Silhouette scores for all points in the dataset. A higher average score indicates better-defined and well-separated clusters.

Davies-Bouldin Index (DBI):

The Davies-Bouldin Index (DBI) is another metric used to evaluate the quality of a clustering algorithm. It is an internal evaluation scheme, where a lower value indicates a better clustering result. Specifically, the DBI is a measure of the average "similarity" ratio of each cluster with the cluster that is most similar to it.

- **Inter-cluster Distance:** It measures the distance between clusters. Smaller distances between clusters indicate that the clusters are not well-separated.

- **Intra-cluster Distance:** It measures the distance within each cluster. Smaller intra-cluster distances indicate that the clusters are more compact.

The DBI takes the ratio of intra-cluster distances to inter-cluster distances and then averages these ratios over all clusters. A lower DBI value implies:

- **Tighter Clusters:** Each cluster is more compact (low intra-cluster distance).
- **Better Separation:** Clusters are well-separated from each other (high inter-cluster distance).

Calinski-Harabasz Index (CH):

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is another metric used to evaluate the quality of a clustering model. This index assesses the ratio of the sum of between-cluster dispersion and within-cluster dispersion for all clusters.

- **Between-Cluster Dispersion (B):** This measures the variance between different clusters. Higher values indicate that clusters are well-separated from each other.
- **Within-Cluster Dispersion (W):** This measures the variance within each cluster. Lower values indicate that clusters are more compact.

$$CH = \frac{B}{W} \times \frac{(N-k)}{(k-1)}$$

where,

N = The total number of data points
k = The number of cluster

4.2 Model Scores:

The following are the scores for each of the models I used for clustering:

- **K means Clustering:**

```
Silhouette score: 0.5442126866165884
Davies-Bouldin Index: 0.5482239239539933
Calinski-Harabasz Index: 6099.403519951906
```

- **DBSCAN:**

```
Silhouette score: -0.7580272402830124
Davies-Bouldin Index: 1.0304901525291839
Calinski-Harabasz Index: 4.249458998733949
```

- **K-Medoids :**

```
Silhouette score: 0.5399221408024186
Davies-Bouldin Index: 0.547363762370909
Calinski-Harabasz Index:6025.901674737833
```

4.3 Analysis:

- **Silhouette Score:** K-Means has the highest Silhouette Score (0.5442), indicating better-defined clusters compared to K-Medoids (0.5399) and DBSCAN (-0.7580, which is very poor as negative values indicate that samples may have been assigned to incorrect clusters).
- **Davies-Bouldin Index:** K-Medoids has a slightly lower Davies-Bouldin Index (0.5474) compared to K-Means (0.5482), which suggests slightly better clustering quality. DBSCAN has the highest value (1.0305), indicating poor clustering.
- **Calinski-Harabasz Index:** K-Means has the highest Calinski-Harabasz Index (6099.40), suggesting better cluster separation and compactness compared to K-Medoids (6025.90) and DBSCAN (4.25, which is extremely low).

4.4 Conclusion:

Given the metrics, K-Means Clustering is the best model among the three. It has the highest Silhouette Score and Calinski-Harabasz Index, and although its Davies-Bouldin Index is marginally higher than K-Medoids, the differences are minimal and compensated by the other two metrics indicating superior performance.

5. Cluster Quality and Interpretation through Visualization

I will now access cluster quality and interpretability of the cluster through various visualizations with regards to the machine learning cluster. I have used the clustering results from K-means clustering as this model provided the best performance.

5.1 Cluster Visualization

The following are the visualization of our data set to determine the nature of each of our clusters.

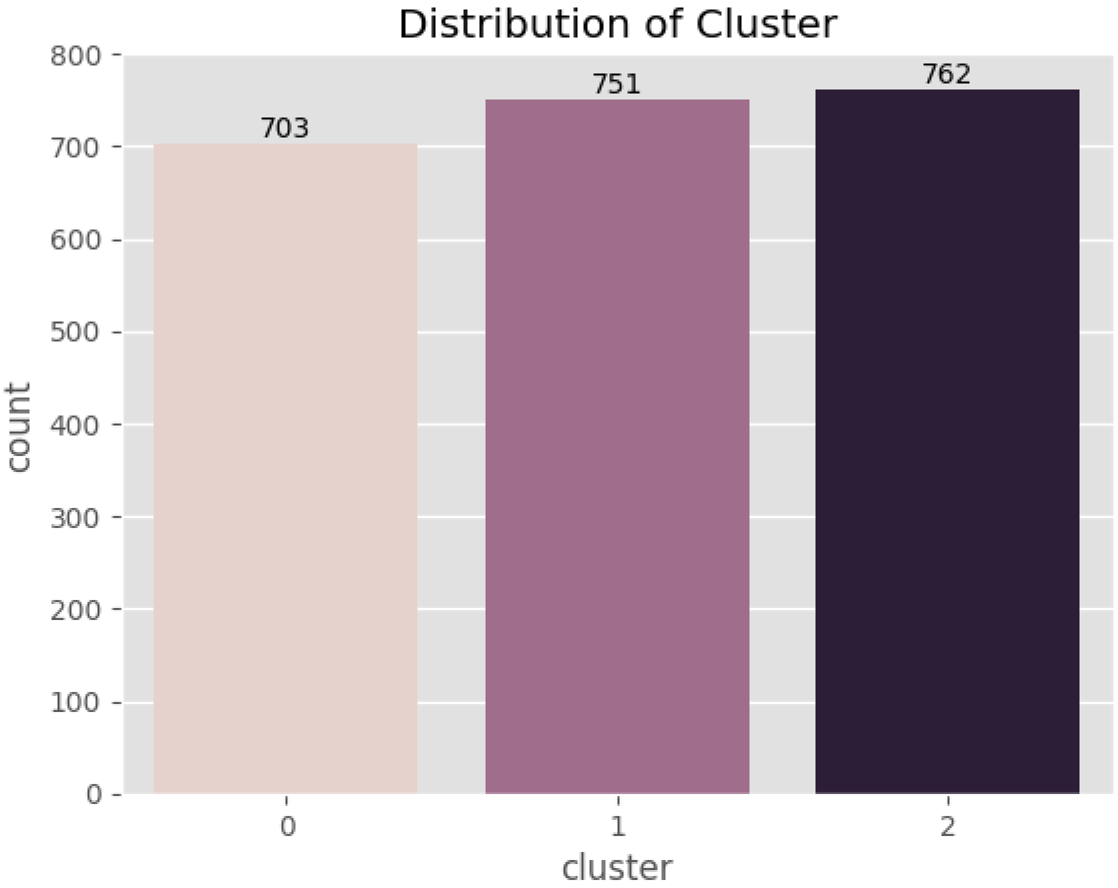

Figure 4. Cluster Countplot

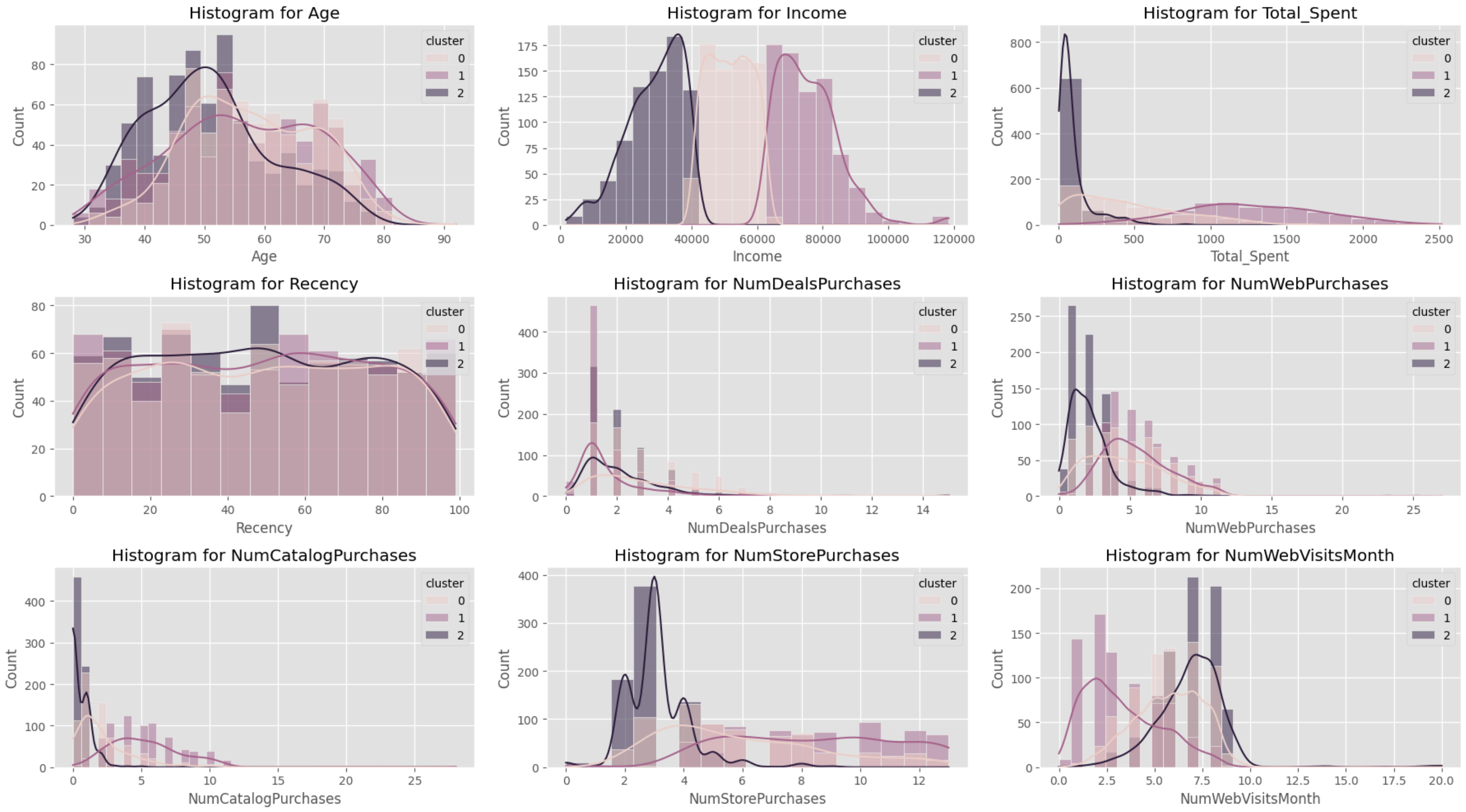

Figure 4. Distribution of Cluster by feature

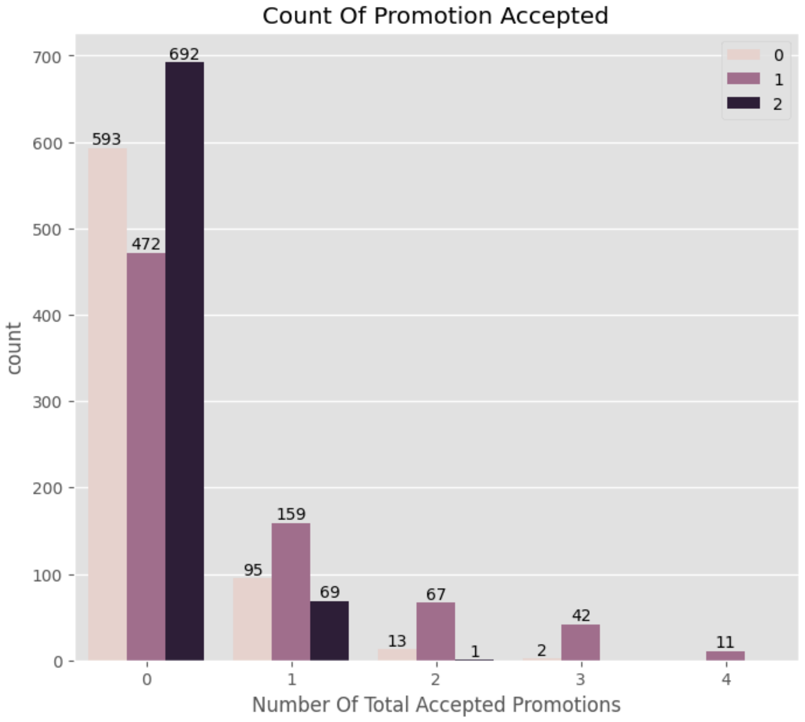

Figure 5. Countplot of Promotion Accepted by cluster

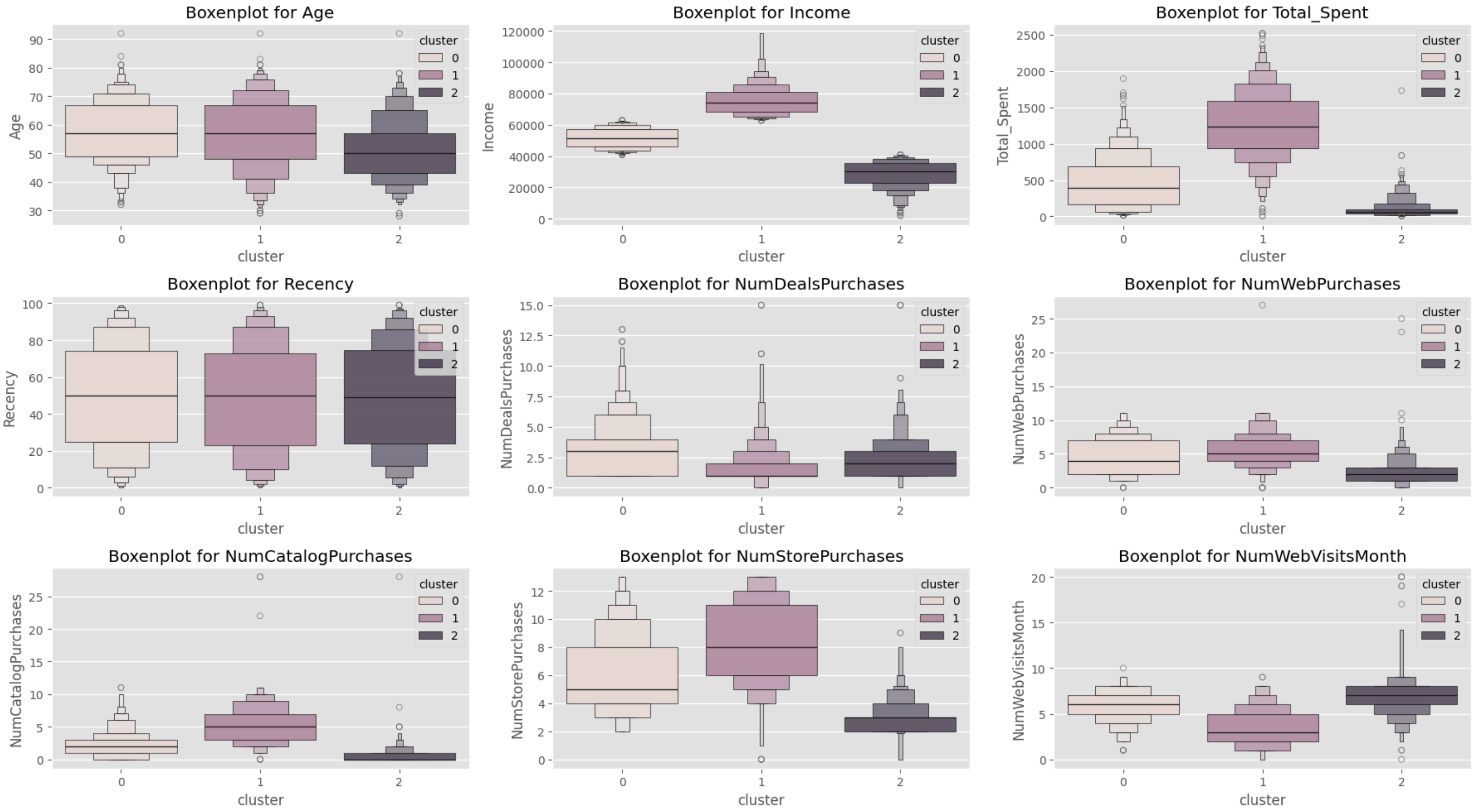

Figure 6. Boxenplot of features by cluster

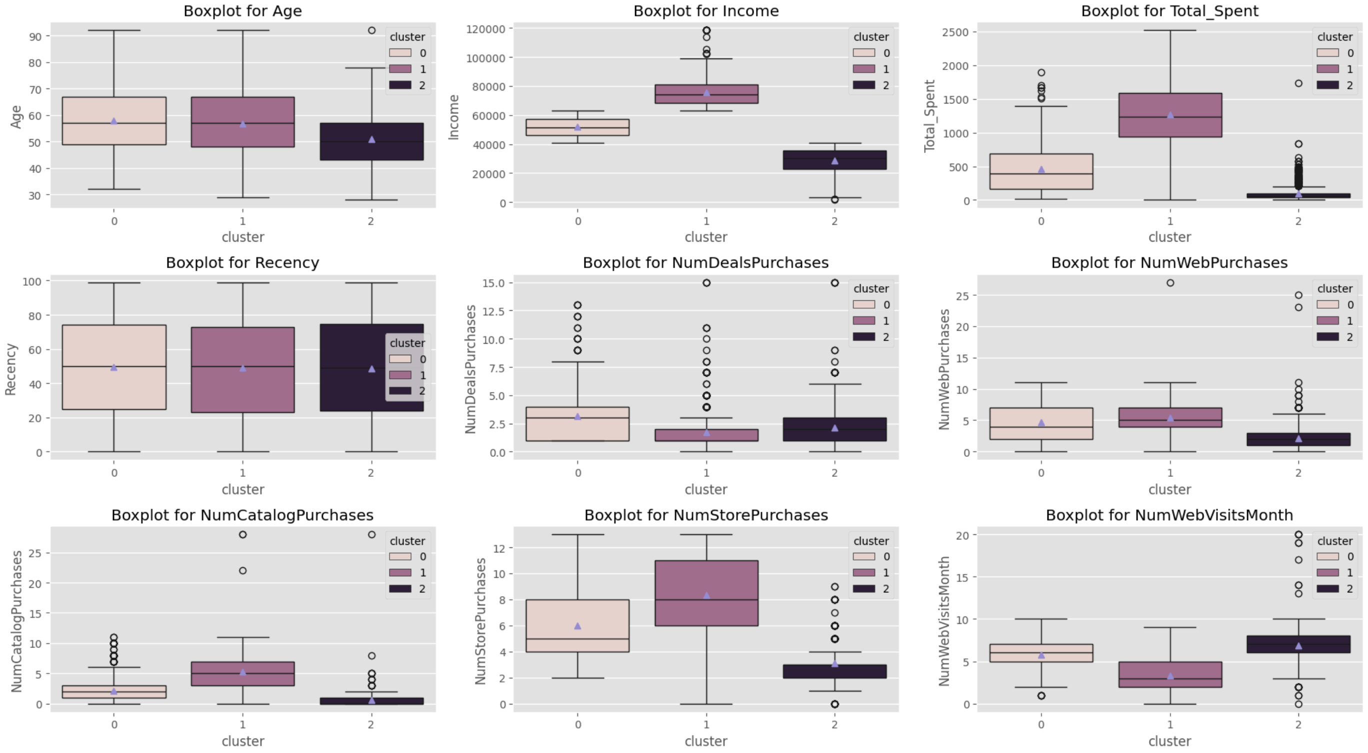

Figure 7. Boxplot of features by cluster

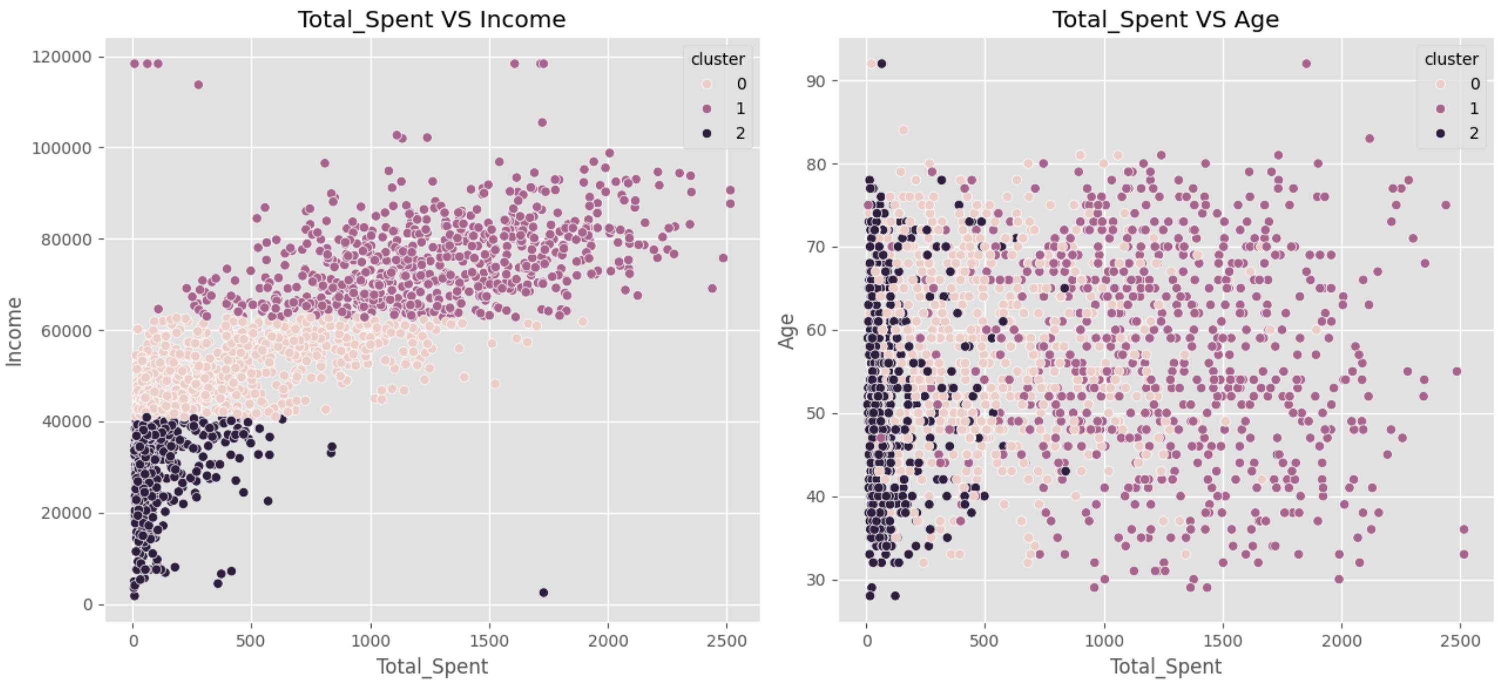

Figure 8. Scatter Plot by cluster

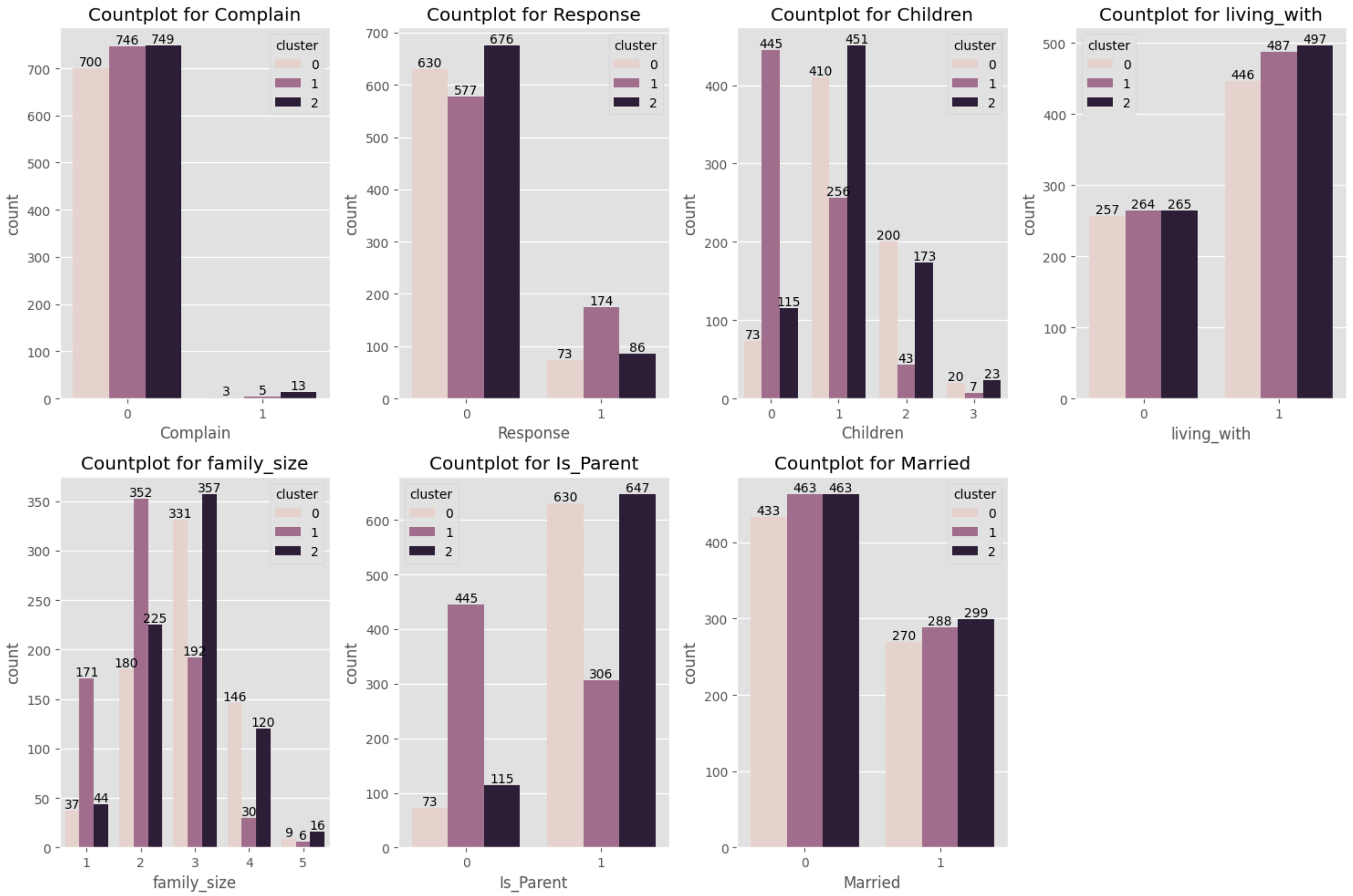

Figure 7. Countplot for feature by cluster

5.2 Cluster Observation

The following observations can be made about each of the cluster:

1. **Cluster 0**

- There are 730 unique customers in this cluster
- This cluster is in the middle range when in come to customer income and amount spent
- On average the number of deals purchase is higher in this cluster
- `Age`:
  - The average age in this cluster is about 57 years old
- `Income`:
  - The average yearly household income of customers in this cluster is about $50,000
- `Spent`:
  - Total spending for the last two years of customers in this cluster is on average $490
- `Complain`:
  - This cluster has the lowest number of complain i.e. 3
- `Response`:
  - This cluster has the lowest response to campaigns compared to other clusters
- `Is_Parent`:
  - This cluster has a really small group of non parent customers i.e. 73.
- `Promotion Accepted`:
  - This cluster has a fair amount of accepted promotions with 127 promotion accepted

2. **Cluster 1**

- There are 751 unique customers in this cluster
- This cluster is the top tier customers with high income and amount spent
- On average the number of deals purchase in this cluster is lowest
- `Age`:
  - The average age in this cluster is about 55 years old
- `Income`:
  - The average yearly household income of customers in this cluster is about $750,000
- `Spent`:
  - Total spending for the last two years of customers in this cluster is on average $1250
- `Complain`:
  - This cluster also has fairly few number of complain i.e. 5
- `Response`:

    - This cluster responds very well to campaign as 174 converted to buying at the last campaign
- `Is_Parent`:
    - This cluster has the largest group of non parent customers i.e. 445.
- `Promotion Accepted`:
    - This cluster has the highest amount of accepted promotions with 463 promotion accepted

3. **Cluster 2**

- There are 762 unique customers in this cluster
- This cluster is in the lowest range when in come to customer with the lowest income and amount spent
- On average the number of deals purchased is higher in this cluster so incentivising this cluster with more deals can be a good strategy.
- `Age`:
    - The average age in this cluster is lowest i.e. 50 years old
- `Income`:
    - The average yearly household income of customers in this cluster is about $30,000
- `Spent`:
    - Total spending for the last two years of customers in this cluster on average is only slightly above $0
- `Complain`:
    - This cluster has the largest number of complain i.e. 13
- `Response`:
    - This cluster has a fair response to campaigns compared to other clusters
- `Is_Parent`:
    - This cluster has the largest group of parent customers i.e. 647.
- `Promotion Accepted`:
    - This cluster has the highest amount of rejected promotions with 692 promotion rejected