CUSTOMER SEGMENTATION

1. **Context**

"I'm a data analyst, and the Chief Marketing Officer has told me that previous marketing campaigns have not been as effective as they were expected to be. I need to analyze the data set to understand customer behavior and propose data-driven solutions."

2. **Objective**

Analyze the provided dataset containing customer information and purchasing behavior to make informed decisions to identify patterns, trends, and correlations that will help to understand customer needs better and reach the right customer with right messaging.

3. **About the dataset**

Data description is as follows;

1. Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise
2. ID - Unique ID of each customer
3. Year_Birth - Age of the costumes
4. Complain - 1 if the customer complained in the last 2 years
5. Dt_Customer - date of customer's enrollment with the company
6. Education - customer's level of education
7. Marital - customer's marital status
8. Kidhome - number of small children in customer's household
9. Teenhome - number of teenagers in customer's household
10. Income - customer's yearly household income
11. MntFishProducts - the amount spent on fish products in the last 2 years
12. MntMeatProducts - the amount spent on meat products in the last 2 years
13. MntFruits - the amount spent on fruits products in the last 2 years
14. MntSweetProducts - amount spent on sweet products in the last 2 years
15. MntWines - the amount spent on wine products in the last 2 years
16. MntGoldProds - the amount spent on gold products in the last 2 years
17. NumDealsPurchases - number of purchases made with discount
18. NumCatalogPurchases - number of purchases made using catalog (buying goods to be shipped through the mail)
19. NumStorePurchases - number of purchases made directly in stores
20. NumWebPurchases - number of purchases made through the company's website
21. NumWebVisitsMonth - number of visits to company's website in the last month

22. Recency - number of days since the last purchase

4. **Analysis Process**

   Task 1: Data Validation

   Task 2: Exploratory Data Analysis

   Task 3: K-Mean clustering

   Task 4: Insights and Customer Segmentation

   Task 5: Conclusion and Recommendation

5. **Data Validation (**Clean, preprocess and transform the dataset (handling missing values, data types, etc.))

- The "Income" column has 24 missing values → Fill null data with mean income value

```
mean_income = data['Income'].mean()
data['Income'].fillna(mean_income, inplace = True)
```

- Dt_Customer's datatype is string → Convert to datetime and extract year to the Dt_Customer column

```
data['Dt_Customer'] = pd.to_datetime(data['Dt_Customer'])
```

```
data['Dt_Customer'] = data['Dt_Customer'].astype(str).str[:4]
```

- Add 2 new column to calculate total amount spent and total purchase into the data

```
data['Total_Purchases'] = (data['NumCatalogPurchases'] + data['NumDealsPurchases']
                           + data['NumStorePurchases'] + data['NumWebPurchases'])
data['Total_Amount_Spent'] = (data['MntSweetProducts'] + data['MntFishProducts'] + data['MntFruits'] +
                              data['MntGoldProds'] + data['MntMeatProducts'] + data['MntWines'])
```

- Transform new feature 'In_Relationship' from 'Marital_Status'

```
def get_relationship(row):
    if row['Marital_Status'] == 'Married':
        return 1
    elif row['Marital_Status'] == 'Together':
        return 1
    else:
        return 0
data['In_Relationship'] = data.apply(get_relationship, axis=1)
data.head()
```

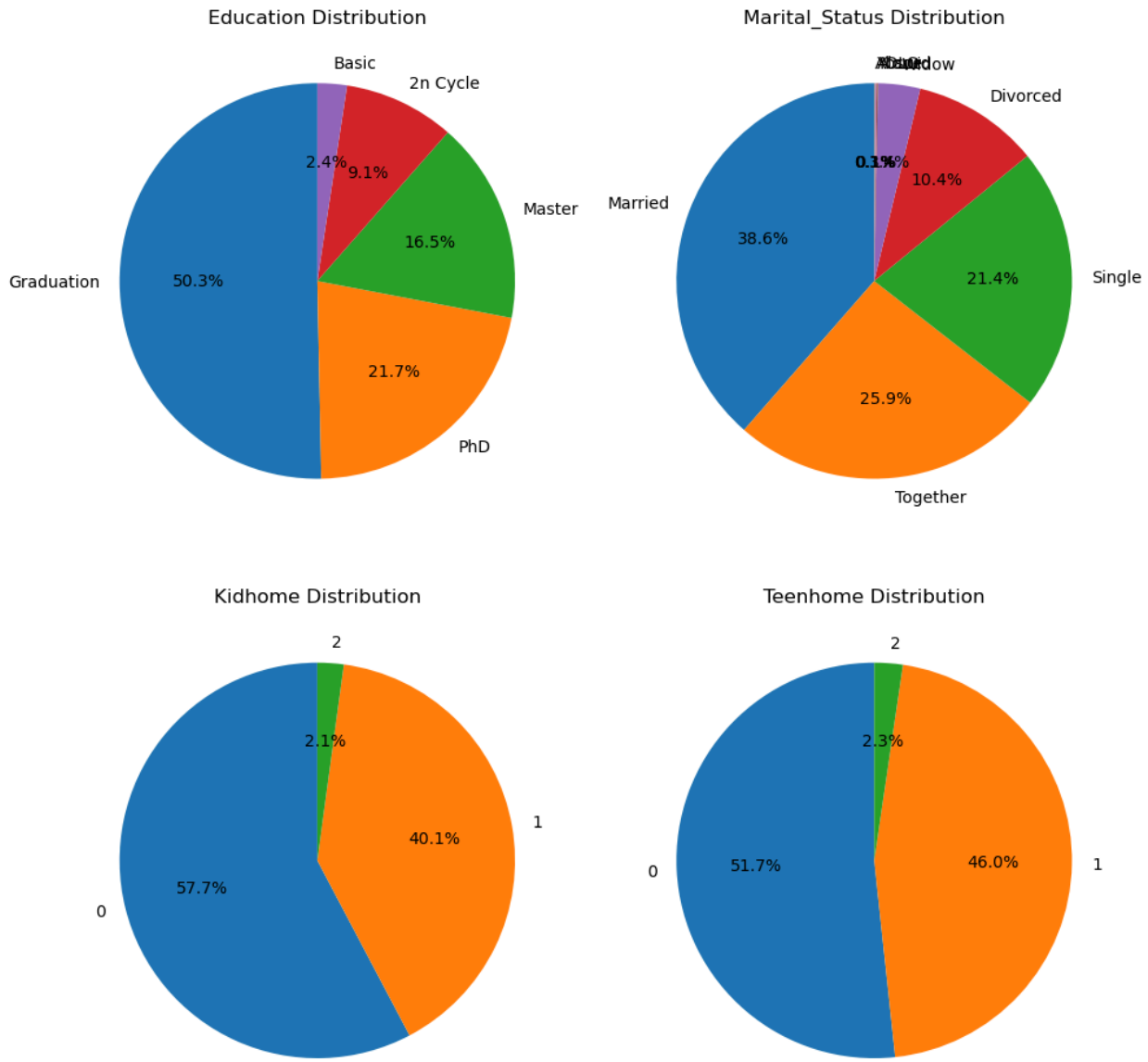
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 24 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Id                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2240 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
...
 22  Total_Purchases      2240 non-null   int64
 23  Total_Amount_Spent   2240 non-null   int64
dtypes: float64(1), int64(20), object(3)
memory usage: 420.1+ KB
```

Metadata after preprocessing

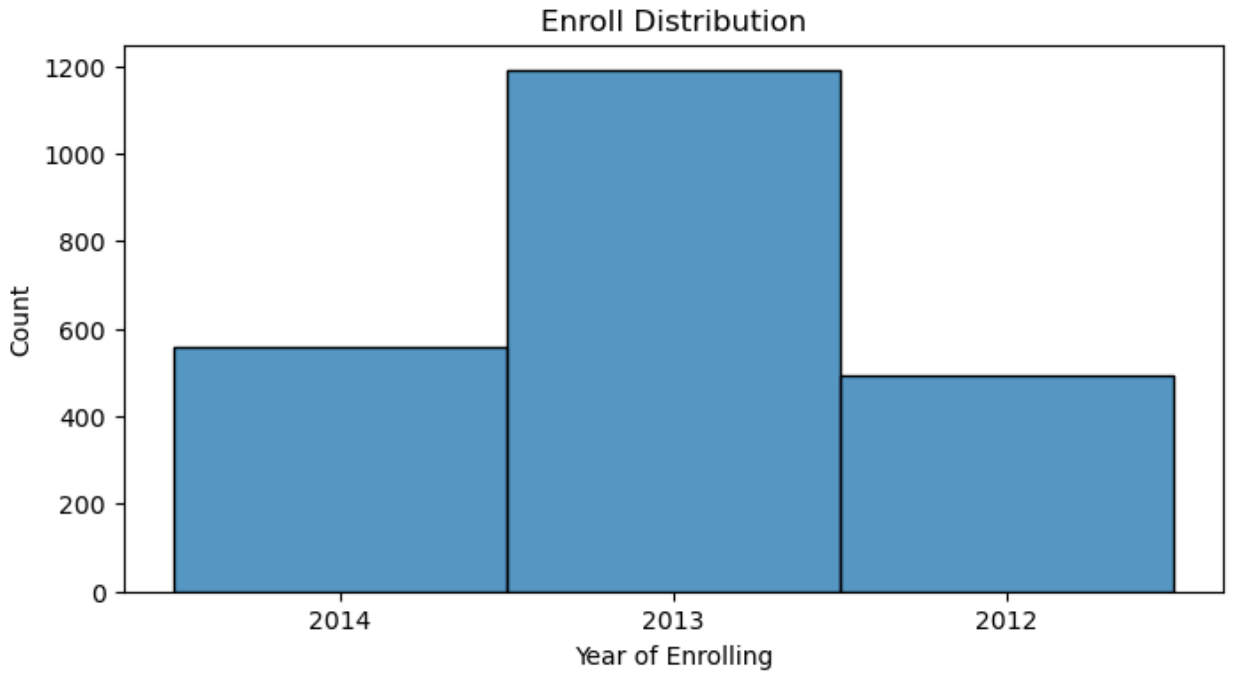
6. Exploratory Data Analysis

a. Identify Outliers

- **Object features**

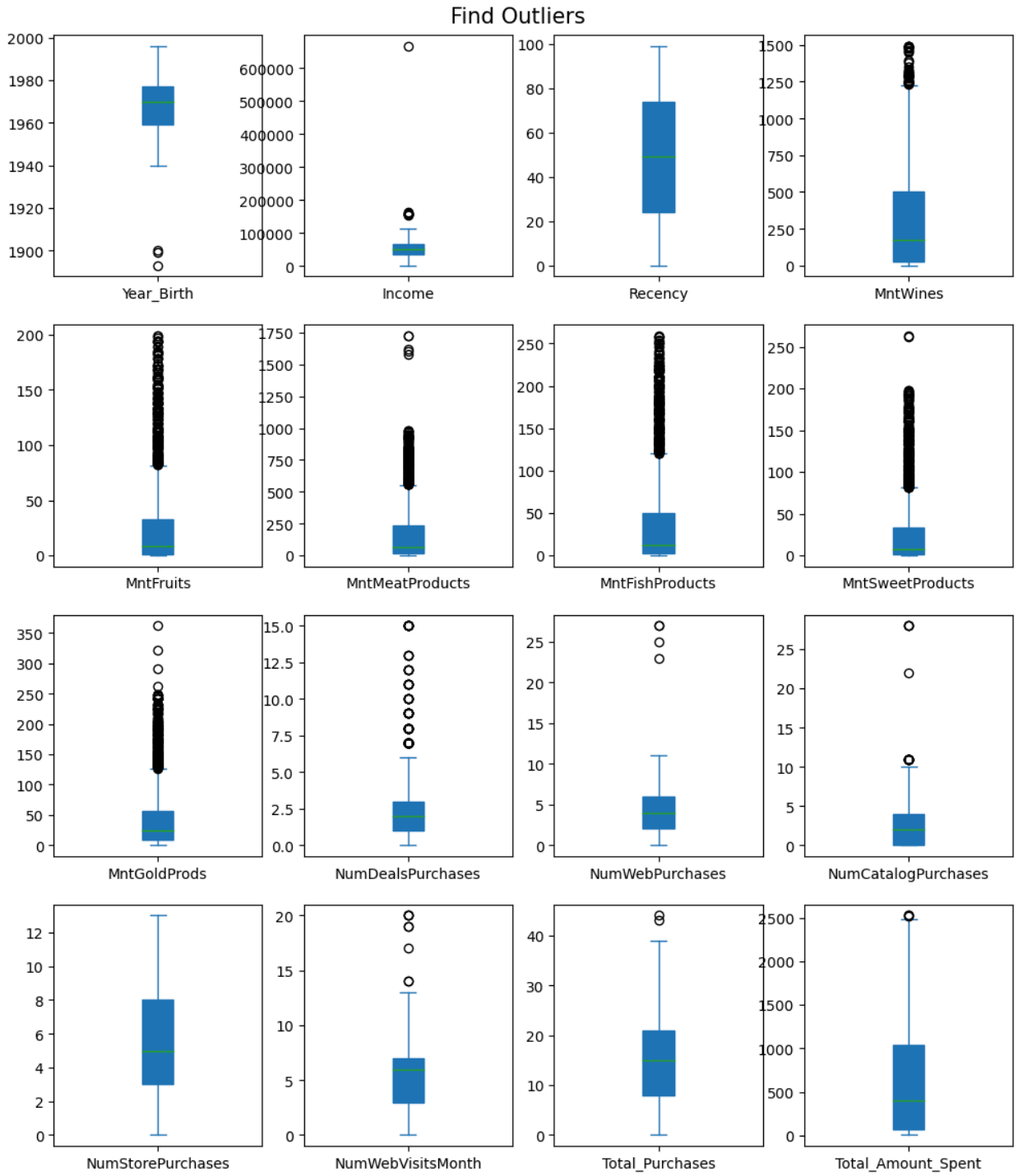


+ The majority of customers have a graduation degree, accounting for more than 50.3%, followed by 21.7% of customers with a PhD, and 16.5% with a master's degree, showing that the marketing campaign's customers have a high level of education.
+ The majority of customers are people in relationships, accounting for 64.5%, of which 38.6% are married and 25.9% are living together.
+ More than 50% of the population has no kid and 40% has 1 kid while the proportion with teenagers is similar, showing a low fertility rate in the customer group.



+ 2013 was the year with an outstanding number of new customers, nearly double that of the other two years, showing that the 2013 marketing strategy was extremely effective.

- **Numerical features**



+ The outliers in Year_birth seem like entry errors since it's impossible that people who were born before 1900 are still alive. Therefore, I will remove the outliers in Year_birth. User mean and standard variation.

```
new_data = data[data['Year_Birth'] >= (data['Year_Birth'].mean()-3*data['Year_Birth'].std())]
data['Year_Birth'].describe()
```

+ The outliers in Total Amount Spent and Income impacts on K-Means clustering so I need to remove it. User IQR method.

```
Q1 = data['Total_Amount_Spent'].quantile(0.25)
Q3 = data['Total_Amount_Spent'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = data[(data['Total_Amount_Spent'] < lower_bound) | (data['Total_Amount_Spent'] > upper_bound)]
outliers.head()
```

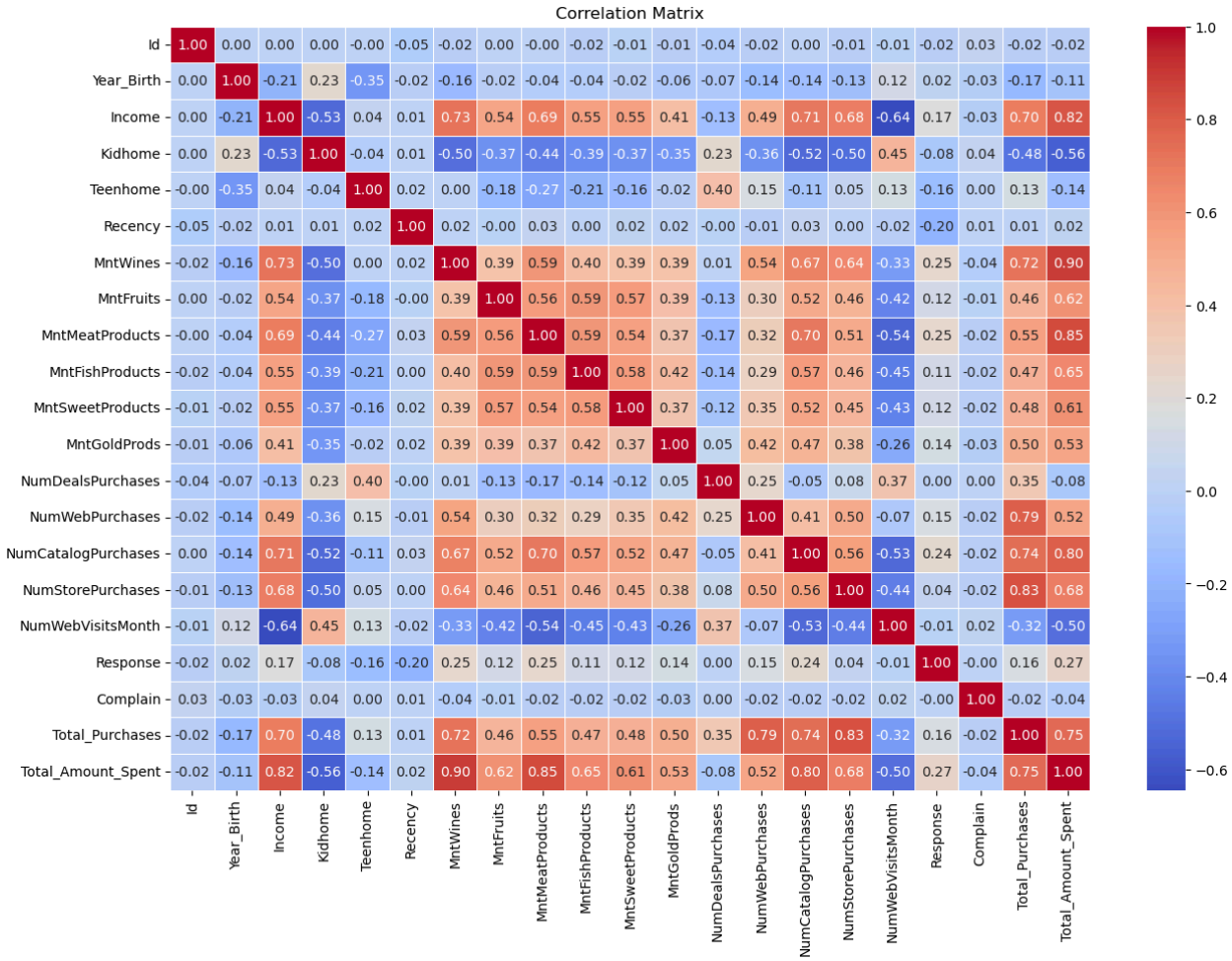
```
data = data[(data['Income'] > lower_bound) & (data['Income'] < upper_bound)]
data.describe()
```

```
Q1 = data['Income'].quantile(0.25)
Q3 = data['Income'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = data[(data['Income'] < lower_bound) | (data['Income'] > upper_bound)]
outliers.head()
```

```
data = data[(data['Total_Amount_Spent'] > lower_bound) & (data['Total_Amount_Spent'] < upper_bound)]
data.describe()
```

b. Identify Patterns or Anomalies

- Use a heatmap to see the correlations between each variable. When it gets redder, they are more positively correlated, and when it gets bluer, they are more negatively correlated.



- **Patterns**

High Income People

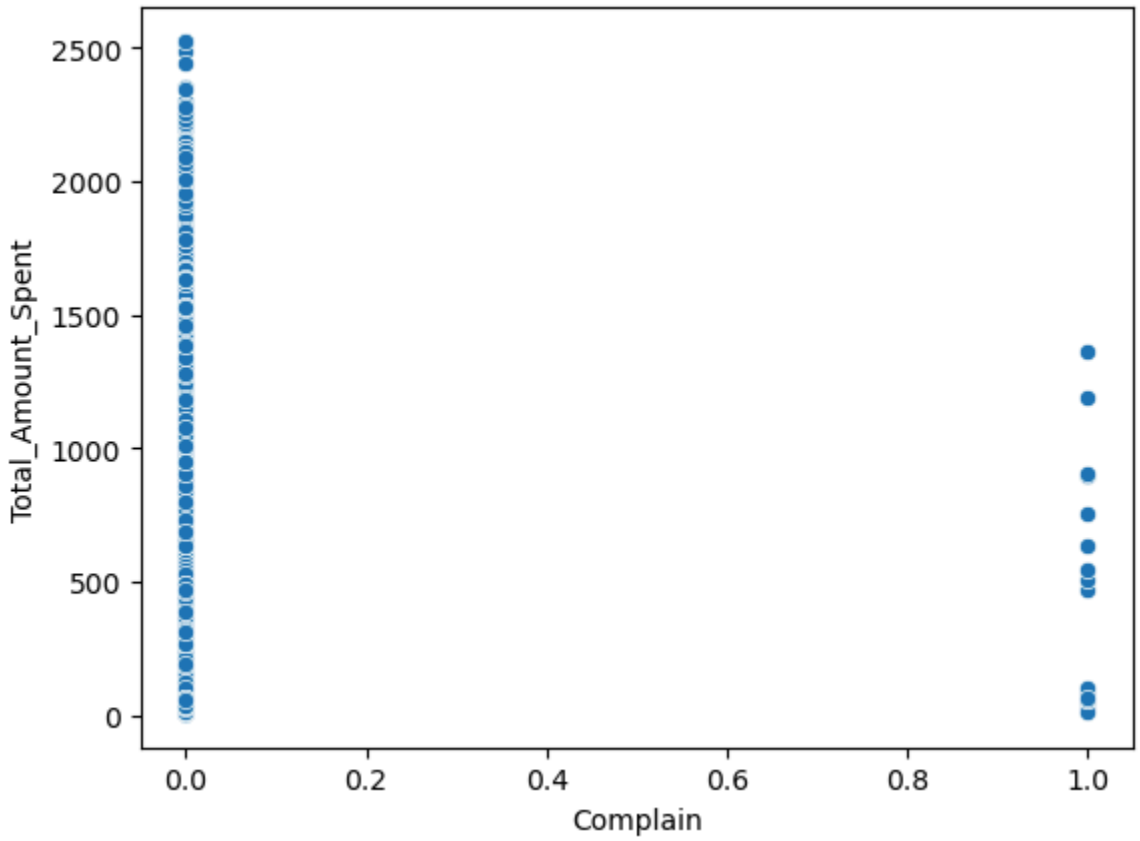
+ tend to spend more and purchase more.
+ tend to visit the company's website less frequently than other people.
+ tend to has few number of purchases made with a discount

People having kids at home

+ tend to spend less and purchase less.
+ tend to visit the company's website most frequently than other people
+ tend to has high number of purchases made with a discount

- **Anomalies**

+ The number of complaints in the last two years has almost no correlation with the total amount spent in the last two years => After further investigating the data, I found that it is because we only have 21 customers who complained in the last two years, but we have 2240 customers in total. The customer service in the company has done a wonderful job in the last two years.




```
Pearson correlation (r):  -0.03714246943718898
Pearson p-value:  0.07936491333673967
```

-> There is a weak negative correlation between the two variables. The correlation results were not statistically significant.

7. K-Means clustering

- K-means clustering is an unsupervised machine learning algorithm used to cluster data based on similarity.

a. Standardizing data

K-means clustering algorithm is based on the calculation of distances between data points to form clusters. When features have different scales, features with larger scales can disproportionately influence the distance calculation. There are various ways to standardize features, we will use standard scaling.

	Income	Total_Amount_Spent	In_Relationship
count	2.232000e+03	2.232000e+03	2.232000e+03
mean	2.415431e-16	2.183638e-17	-1.270390e-16
std	1.000224e+00	1.000224e+00	1.000224e+00
min	-2.423163e+00	-9.988141e-01	-1.344451e+00
25%	-7.867831e-01	-8.923797e-01	-1.344451e+00
50%	-3.604195e-03	-3.477350e-01	7.437983e-01
75%	8.000023e-01	7.294974e-01	7.437983e-01
max	3.014678e+00	3.192040e+00	7.437983e-01

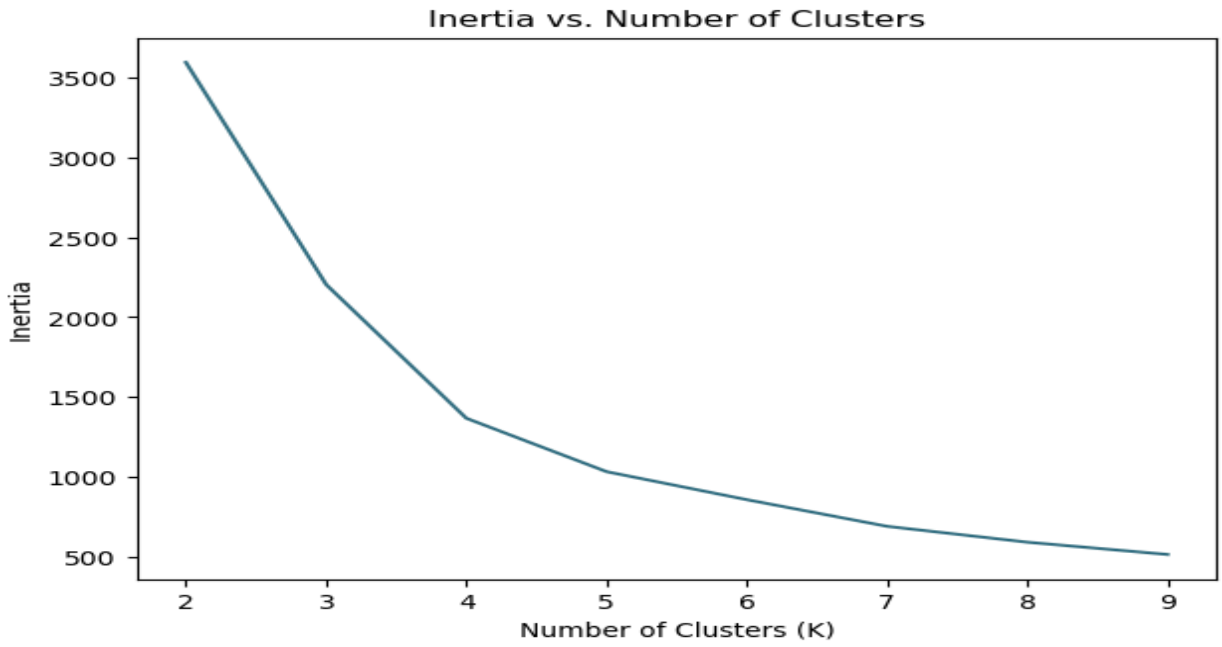
Data after standardizing

b. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in various fields, including data analysis, machine learning, and image processing. I use this method to simplify complex datasets by reducing the number of variables while preserving most of the essential information.

c. Elbow method

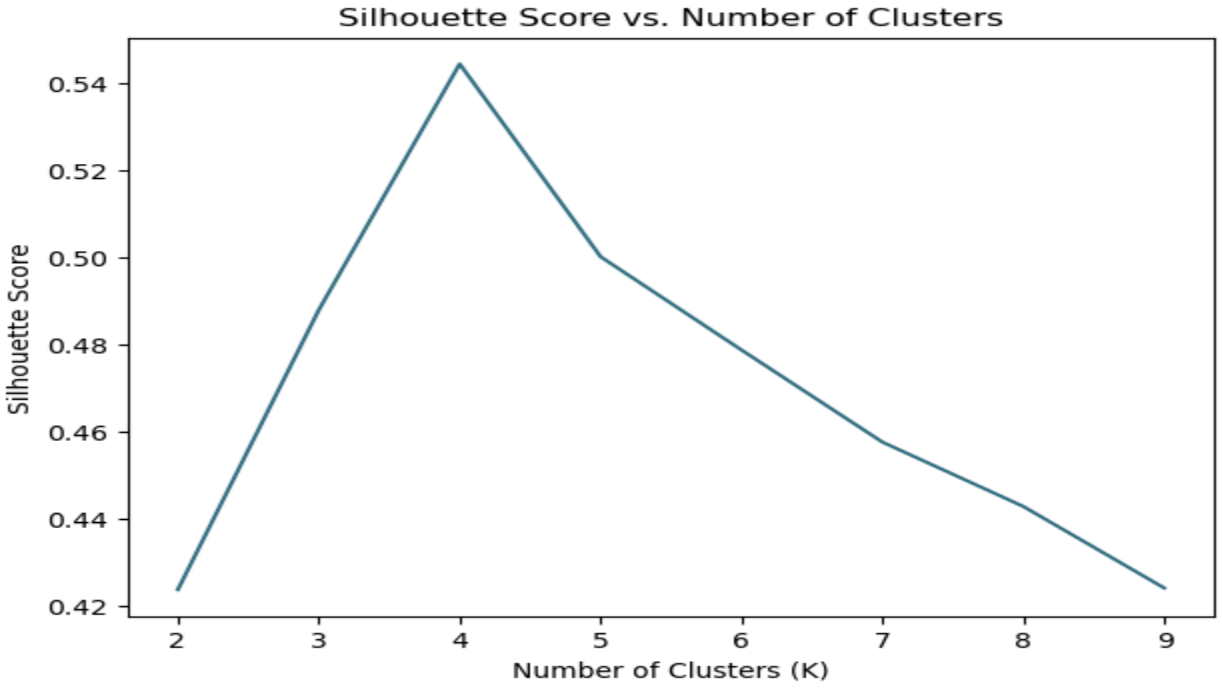
The elbow method is a heuristic technique used to determine the optimal number of clusters (k) in k-means clustering. It's a visual approach based on the distortion or inertia within each cluster, which represents the sum of squared distances between data points and their respective cluster centroid.



Elbow method suggests 4 or 5 clusters. Let's check the silhouette score.

d. Silhouette score analysis

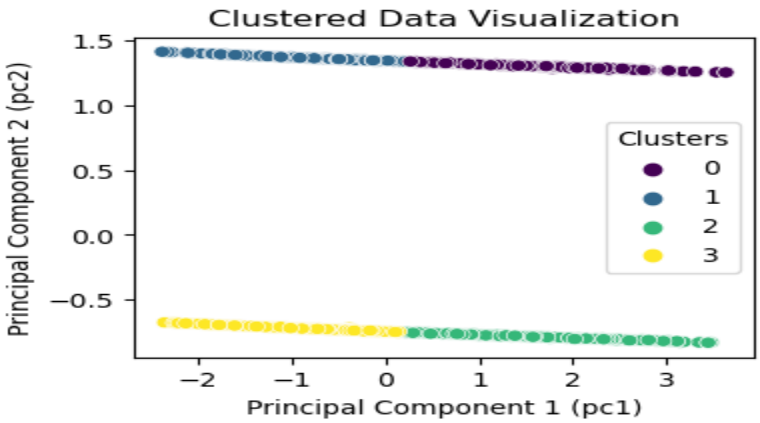
Silhouette score analysis is a technique used to evaluate the quality of clustering in unsupervised machine learning. It measures the average silhouette coefficient for all data points, which represents how well-placed a data point is within its assigned cluster.



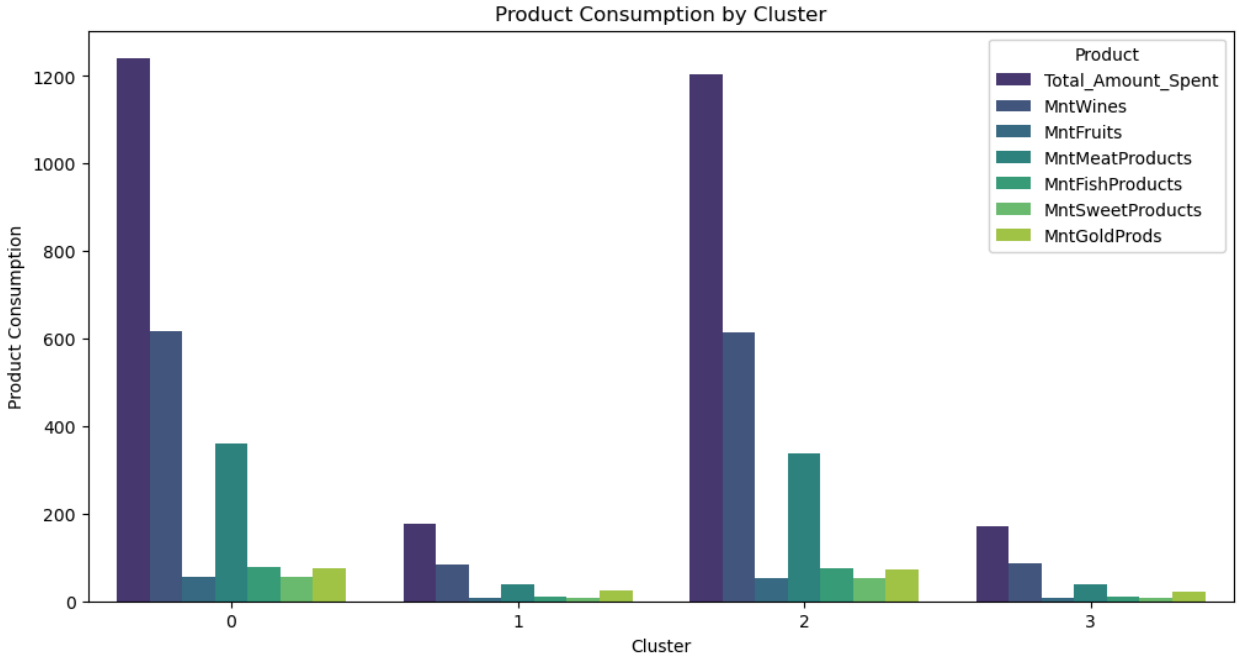
-> The Elbow Method and Silhouette Analysis suggested 4 clusters (k=4). The elbow method highlighted the number of 4 or 5 clusters as a reasonable number of clusters. The silhouette score analysis revealed a peak silhouette score for k=4.

8. Insights and Customer Segmentation

a. Visualization of clusters

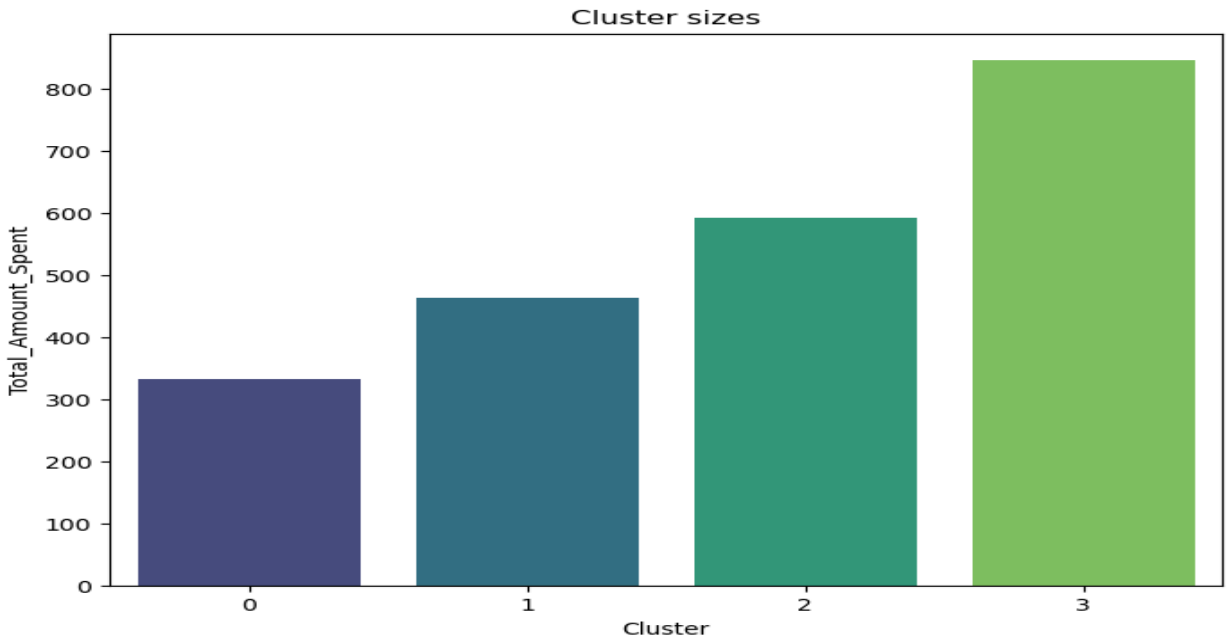


b. Mean consumption of different product types by cluster



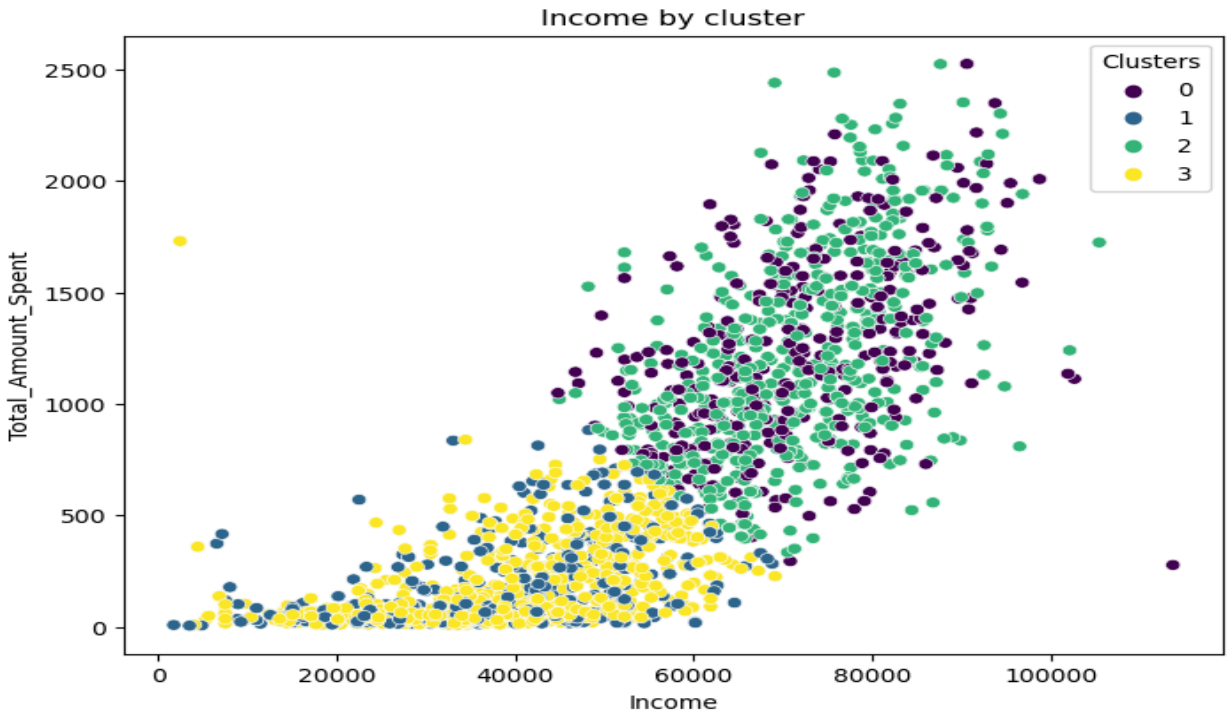
Cluster 0 and cluster 2 are the two clusters that spend the most, in which wines and meat are the two types of products that these two clusters consume the most and outperform the other two clusters.

c. Cluster sizes



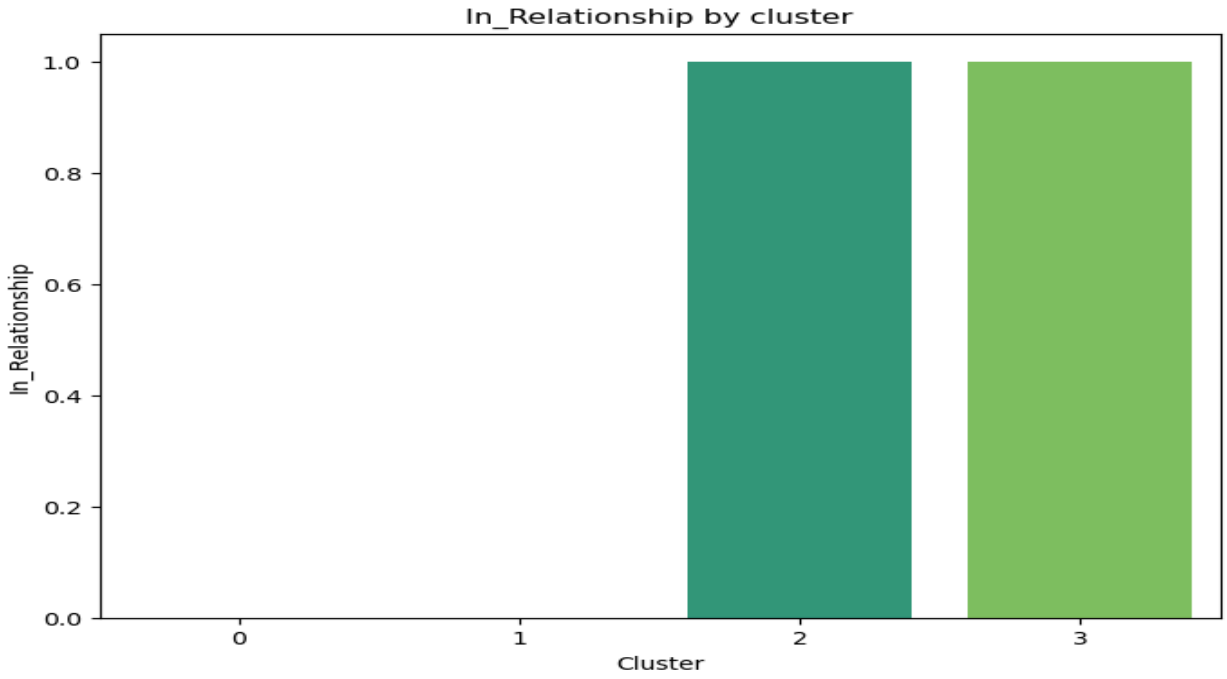
Cluster 2 and cluster 3 are the clusters with the largest number of customers in the company's customer base

d. Income by cluster



Clusters 0 and 2 are clusters with a much higher average income than the other two clusters

e. In_relationship feature by cluster



Clusters 2 and 3 are 2 clusters that are in a relationship

9. Conclusion and Recommendations

This section contains the results of the K-means clustering analysis, which aimed to identify distinct customer segments based on the total amount of purchases they made (Total_Amount_Spent). The analysis utilized 'Income' and 'In_Relationship' features.

a. Cluster Characteristics

- Cluster 0: High Income single customers

+ This cluster are customers have high income and they are single, they account for 15% of the customer base

- Cluster 1: Low Income single customers

+ These customers have low income and they are single and account for 21% of the customer base

- Cluster 2: High Income customers in relationship (either married or together)

+ These customers have high income and they are in a relationship, and they account for 26% of the customer base

- Cluster 3: Low Income customers in relationship

+ These customers have low income and they are in a relationship and represent 38% of the customer base

b. Recommendations

Based on the clusters, tailored marketing strategies can be created. Customers from these segments will have different interests and product preferences.

Marketing Strategies for Each Cluster

- Cluster 0: High Income single customers

+ These customers buy a lot of wines and fruits.

+ This cluster contains single customers. Promo images with friends, parties or single trips may be more efficient for single customers

- Cluster 1: Low Income single customers

+ Promos with discounts and coupons may bring good results for this targeted group.

+ Loyalty programs may stimulate these customers to purchase more often.

- Cluster 2: High Income customers in relationship (either married or together)

+ Preliminary analysis showed that high income customers buy more wines and fruits.

+ A tailored campaign to promote high quality wines may bring good results.

+ This cluster contains customers in relationships, family-oriented promo-images should be quite effective for this audience.

- Cluster 3: Low Income customers in relationship

+ This cluster has the highest percentage of our customers (38%).

+ Family offers and discounts may influence these customers to make more purchases