Introduction

This report investigates unsupervised clustering of a dataset of customer information to understand customers and their behavior. These insights are valuable for various business use-cases such as personalized marketing, customer segmentation, the development of new marketing strategies, the improvement of product offerings, the enhancement of customer relationships, and more. Personalized marketing helps the business to market the product only to a particular segment that is most likely to buy that product, hence, not wasting its resources to target all customers of all types. By segmenting customers, we can optimize the significance of each customer to the business. The purpose of this segmentation is to divide the customer base into distinct groups with similar behaviors to target marketing strategies more effectively.

Dataset Description

The data includes demographic characteristics like age, marital status, and education, as well as spending patterns, details about purchases in various categories, and marketing campaign response rates. The full list of data columns (total of 29 columns) is as follows (In addition to these pre-existing features, several new features have been engineered from the existing data to provide more granular insights):

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, Teenhome: Number of children\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 customer complained in the last 2 years, 0 otherwise

- Z\_CostContact: The cost associated with contacting or reaching out to customers

- Z\_Revenue: metrics such as revenue generated from customer transactions, sales, etc.

Products  
- MntWines, MntFruits : Amount spent on wine\fruits in last 2 years  
- MntMeatProducts, MntFishProducts: Amount spent on meat\fish in last 2 years  
- MntSweetProducts, MntGoldProds: Amount spent on sweets\gold in last 2 years  
Promotion  
- NumDealsPurchases: Number of purchases made with a discount  
- AcceptedCmp1-5: 1 if customer accepted the offer in the 1st-5th campaign, 0 otherwise  
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise   
Place  
- NumWebPurchases: Number of purchases made through the company’s web site  
- NumCatalogPurchases: Number of purchases made using a catalogue  
- NumStorePurchases: Number of purchases made directly in stores  
- NumWebVisitsMonth: Number of visits to company’s web site in the last month

The data is collected from 2240 customers. By removing rows with missing values, 2216 rows are remained.

1. PRELIMINARY EDA AND FEATURE ENGINEERING

**1.1 Preliminary Feature Engineering**

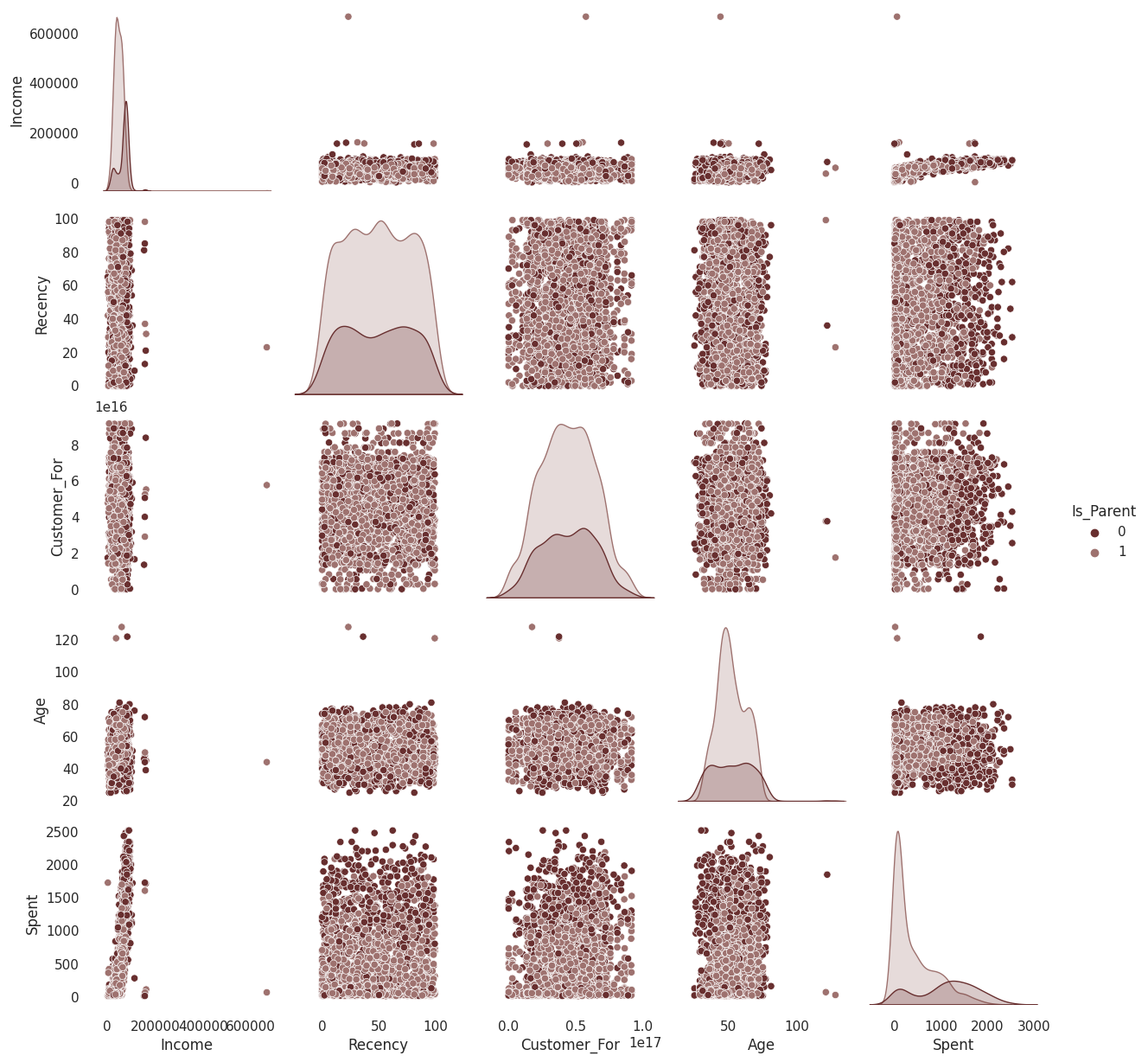
The objective of feature engineering is to create new variables or modify existing ones to enable efficient machine learning model training. Multiple new variables were derived to better understand the characteristics and behavior of the customers.

1. Age: Instead of using the year of birth, it was more insightful to calculate the age of the customer as of 2021.
2. Spent: To represent the total amount spent by the customer across different categories.
3. Living\_With: By categorizing the marital status of the customer into 'Partner' or 'Alone'.
4. Children: The total count of kids and teens at home.
5. Family\_Size: The total number of members in the household, considering the living situation and number of children.
6. Is\_Parent: To indicate whether the customer is a parent or not.
7. Education: Education levels were simplified and segmented into 'Undergraduate', 'Graduate', and 'Postgraduate'.

Some original features such as 'Marital\_Status', 'Dt\_Customer', 'Z\_CostContact', 'Z\_Revenue', 'Year\_Birth', 'ID' were dropped as they were either redundant or not informative for the analysis.

**1.2 Preliminary Exploratory Data Analysis (EDA)**

EDA was performed to understand the structure, relationships, and trends within the dataset. I chose the following variables for this analysis: "Income", "Recency", "Customer\_For", "Age", "Spent", "Is\_Parent".



The distributions on the diagonal show how each of these variables is distributed and allow us to understand the individual variable characteristics. Here are some insights derived from the data:

1. Income: Income distribution is somewhat skewed. Most customers have an income in the mid-range, with fewer customers having very high or very low incomes.
2. Recency: The recency distribution shows that most customers have recent interactions with the brand. A small portion of customers have not interacted recently, indicating a possible churn risk.
3. Customer\_For: This variable represents the duration of the customer relationship. Most customers have a relatively long-lasting relationship with the company, showing good customer retention.
4. Age: The age distribution indicates that most customers are middle-aged. There is a smaller proportion of younger and older customers.
5. Spent: Customers' total spending shows a wide range. While most customers have a moderate spending amount, some customers have high total spending, which might indicate a loyal or premium customer base.
6. Is\_Parent: The number of customers who are parents is less than those who are not. This difference might be due to a variety of factors such as the nature of the products or the customer attraction strategy of the company.

The generated pair-plots provide valuable insights into how these variables interact with each other. Each subplot provides a comparative view of two variables, with the color representing whether the customer is a parent or not. Here are some notable trends and patterns from the pair-plots:

1. Income vs. Spent: There seems to be a positive correlation between Income and Spent. Customers with higher incomes tend to spend more, indicating their higher purchasing power. However, consumers with very high incomes are not spending much on the products of this company.
2. Age vs. Spent: There is not a clear trend between age and spending, suggesting that spending is likely influenced more by other factors.
3. Recency vs. Spent: This pair shows a somewhat scattered relationship, suggesting that recent engagement does not necessarily translate into higher spending. The company might need to look into their engagement strategies to convert recent interactions into sales.
4. Customer\_For vs. Spent: There seems to be a slight positive trend here, indicating that customers who have been with the company for a longer time tend to spend more, possibly reflecting customer loyalty.

As can be seen from the above figure, there are some outliers in the dataset which should be cleaned first to proceed.

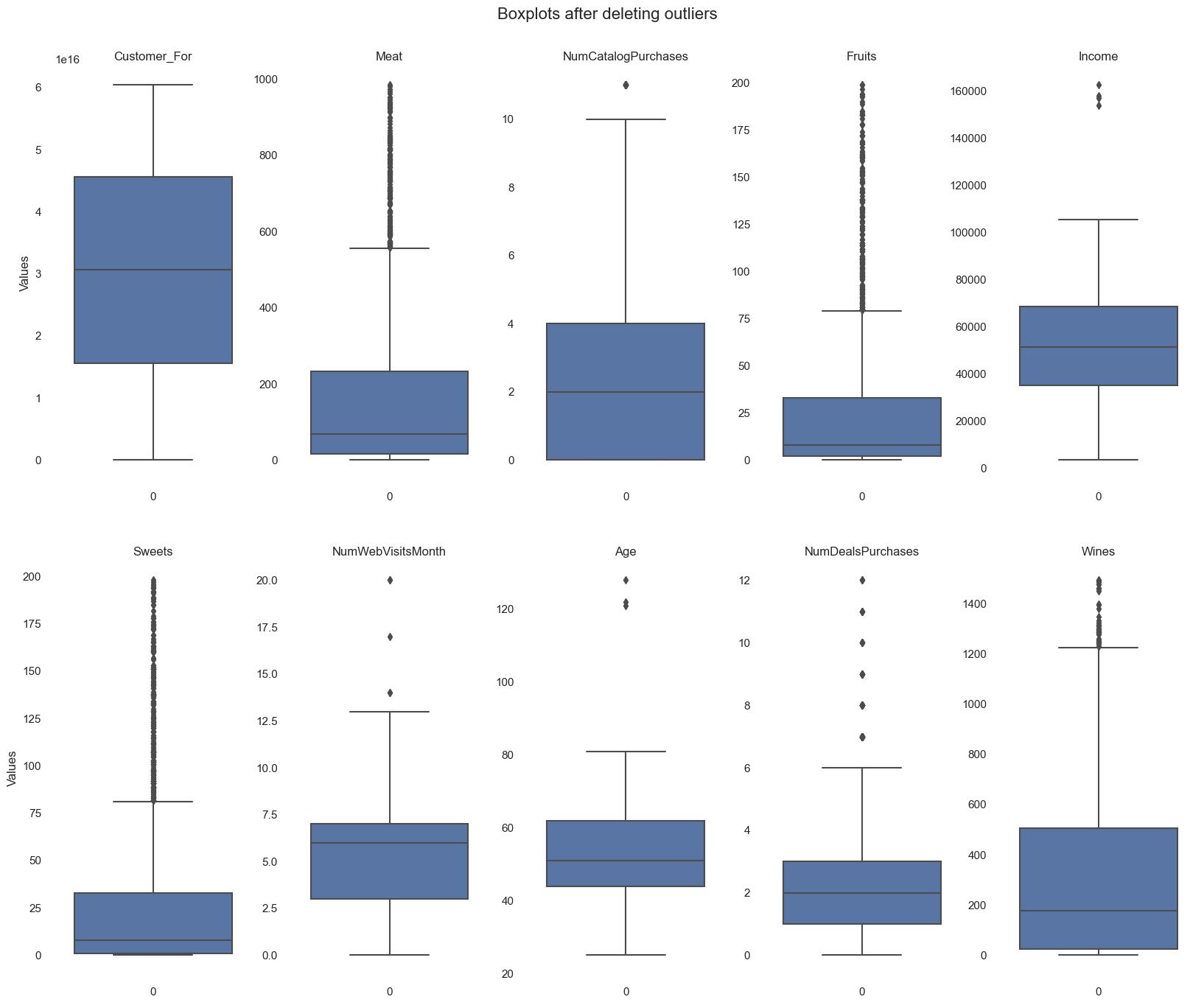
2. Data Cleaning

In the data cleaning phase, I focused on identifying and handling outliers. I used boxplots to visualize the distributions of selected numerical features and identified potential outliers. The features included in the boxplots were "Customer\_For", "Meat", "NumCatalogPurchases", "Fruits", "Income", "Sweets", "NumWebVisitsMonth", "Age", "NumDealsPurchases", and "Wines".

A picture containing text, diagram, screenshot, parallel

Description automatically generated

From the boxplots, it is evident that certain features exhibit outliers. To remove these outliers, I calculated the interquartile range (IQR) for each feature. Any value below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR was considered an outlier.



However, an examination of the "Age" boxplot still reveals the presence of outliers. To manage this, additional filtering is done on the "Age" column, removing any values above 90. After this operation, a boxplot of "Age" demonstrates that the outliers have been effectively removed.

A picture containing rectangle, screenshot, diagram, line

Description automatically generated

The resulting dataset contained 2203 rows and 30 columns, with outliers effectively removed.

3. Detailed EDA

The objective of this process was to understand the underlying patterns, relationships, or anomalies that might exist in our data, in detail.

Education Distribution

The pie chart represents the distribution of customers based on their education level. Each section of the pie chart denotes the percentage of customers with a specific level of education.

A picture containing circle, screenshot, diagram

Description automatically generated

Living Condition Distribution

The bar graph illustrates the number of customers living either alone or as a partner. Each bar represents a unique living condition with its height corresponding to the number of customers living under that condition.

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

Age Distribution

The Kernel Density Estimation (KDE) plot showcases the distribution of customers' ages. The x-axis represents the age, while the y-axis represents the estimated density of that age group.

A picture containing diagram, plot, line, text

Description automatically generated

Education vs Kidhome and Education vs Income

The distribution plot (displot) between Education and Kidhome shows the distribution of customers' education level and the number of kids at home.

A picture containing text, screenshot, font, line

Description automatically generated

Similarly, the displot between Education and Income provides a distribution of the income levels for each education group.

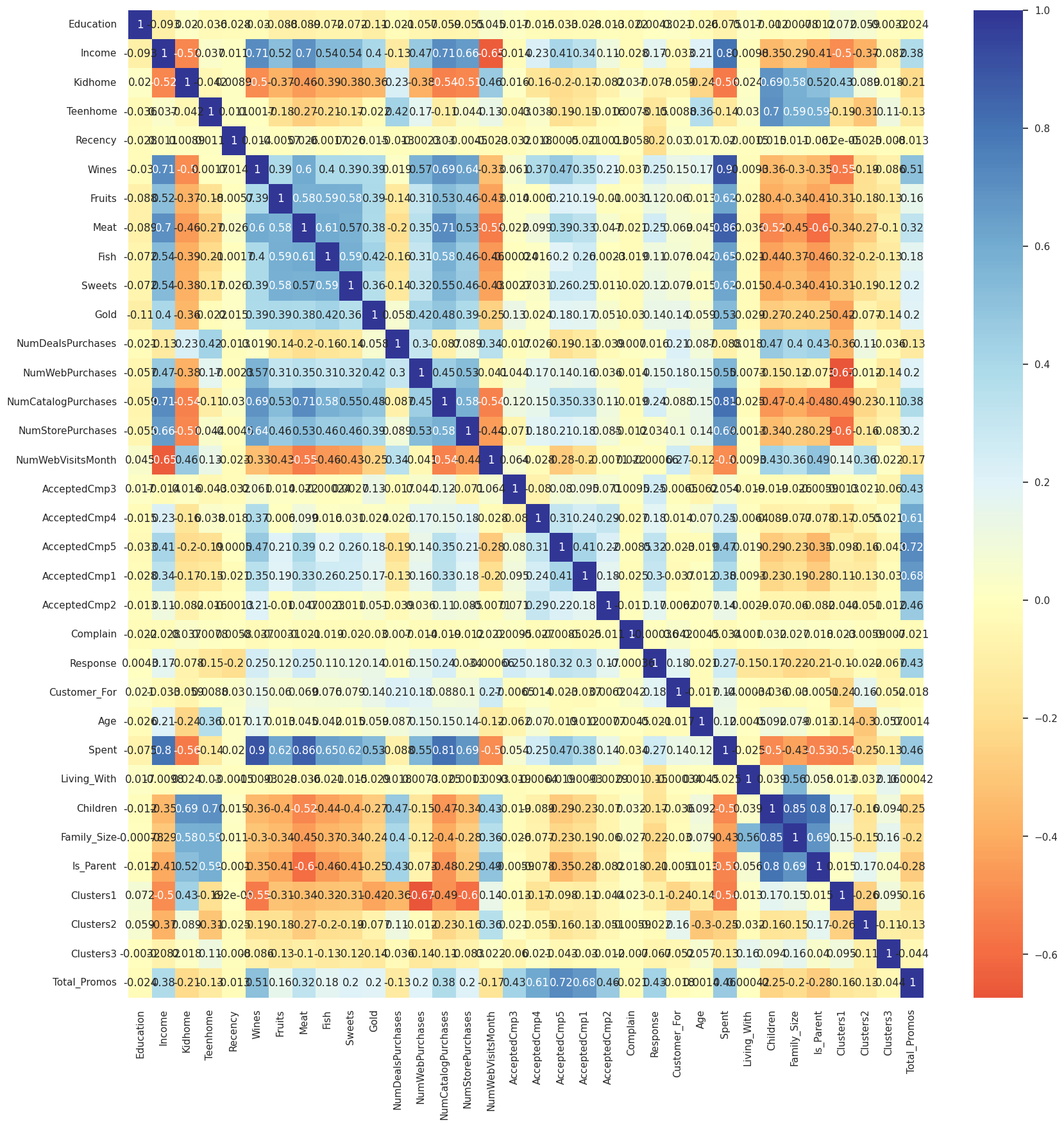
A picture containing text, screenshot, diagram

Description automatically generated

Correlation Matrix

The heatmap represents the correlation between various features. A stronger positive or negative color denotes a stronger positive or negative correlation, respectively. A strong correlation (either positive or negative) between two variables suggests that they might contain similar information from the perspective of predictive modeling. Consequently, we may not need both variables in our model. We can observe that a high spent is correlated to the amount of wines and meat purchased. Also is correlated to the purchases by catalog!

Also, there is a high correlation between if a person is parent and the number of children and the family size, and also between income and spent money, which are obvious correlations that should naturally exist.



4. FEATURE ENGINEERING (ENCODING, SCALING, DIMENSIONALITY REDUCTION)

In the feature engineering stage of the analysis, encoding, scaling, and dimensionality reduction were carried out to prepare the dataset for machine learning models.

**4.1 Encoding**

The first step was encoding, which is a process to convert categorical data into numerical data. Label Encoding was used on binary features like 'Is\_Parent'. This type of encoding assigns each unique category value to a unique integer. On the other hand, One-Hot Encoding was applied to multi-class features like 'Education' and 'Living\_With'. This creates a separate binary (0 or 1) column for each category in the original feature.

**4.2 Scaling**

After encoding, the data was scaled. Scaling modifies the range of data to ensure that all the features are on a similar scale. Here, Min-Max Scaler was used, which transforms the features by scaling each feature to a specified range, often between zero and one.

**4.3 Dimensionality Reduction**

Lastly, a dimensionality reduction technique was applied to the dataset. In this case, Principal Component Analysis (PCA) was used. It was applied to the dataset after encoding and scaling to reduce the number of dimensions without losing much information, and to make the machine learning models more efficient.

As is seen in the figure, the data is reduced to only three dimensions.

A picture containing text, screenshot, design

Description automatically generated

After all these steps, the data was successfully transformed and ready to be used for machine learning models.

5. TESTING DIFFERENT CLUSTERING METHODS

In this section of the analysis, the aim is to identify groups, or clusters, within the data that share common characteristics. Different clustering methods are tested to identify the best performing model for the given dataset.

**5.1 K-Means Clustering**

The K-Means algorithm was employed first, dividing customers into K distinct groups. A critical step in the application of K-Means was the determination of the optimal number of clusters, K. The Elbow method was utilized for this purpose, which involved plotting the distortion score as a function of the number of clusters. The Elbow Method resulted in the choice of four clusters. Optimum K is 4. Then, K-means clustering was performed.

A picture containing text, diagram, line, plot

Description automatically generated

**5.2 Agglomerative Hierarchical Clustering:**

Next, Agglomerative Hierarchical Clustering was performed, providing a tree-like visualization, or dendrogram, of customer grouping. Agglomerative hierarchical clustering starts with each observation as its own cluster and pairs the closest clusters together iteratively. Each level of the dendrogram represented a possible clustering of the data, with the optimal number of clusters determined by examining the largest vertical distance that did not intersect any clusters in the dendrogram.4 clusters were used again here

**5.3 DBScan:** This algorithm does not take specific number of clusters as input and when I applied this it found many clusters. However, this much clustering was not a good way to perform customer segmentation so I am not using this method.

**Evaluation of best performing model**

The performance of a clustering model is usually evaluated based on its coherence and separation, meaning that items in the same cluster should be as similar as possible (coherence), while items in different clusters should be as dissimilar as possible (separation). I evaluated the clustering performance using silhouette score. Kmeans performed best with 0.4 score.

To examine the clusters formed see the distribution of the clusters. The 3D scatter plot illustrates the data points in reduced dimensionality colored based on the cluster they belong to, as per KMeans clustering.

A picture containing text, screenshot, powder, design

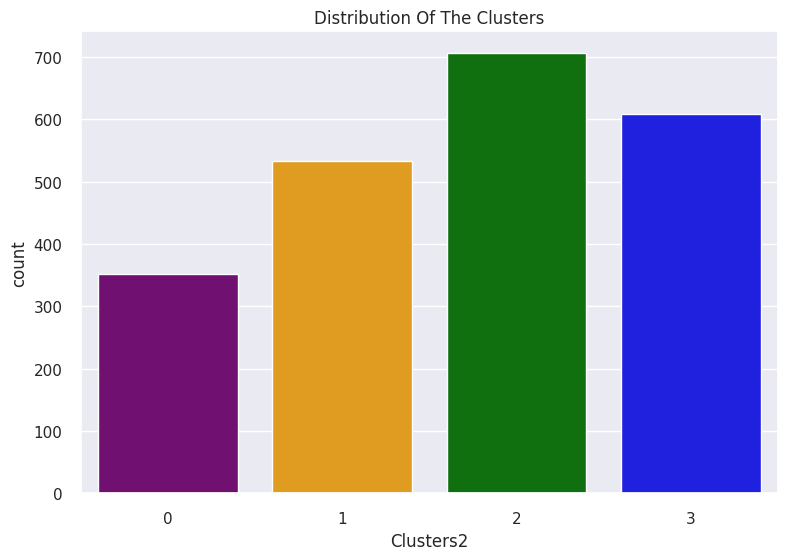
Description automatically generated

6. CUSTOMER SEGMENTATION AND PROFILING

Since this is an unsupervised clustering, labeled data to evaluate or score the model is not available. The purpose of this section is to study the patterns in the clusters formed and determine the nature of the clusters' patterns.

**5.1 Distribution of the Clusters**

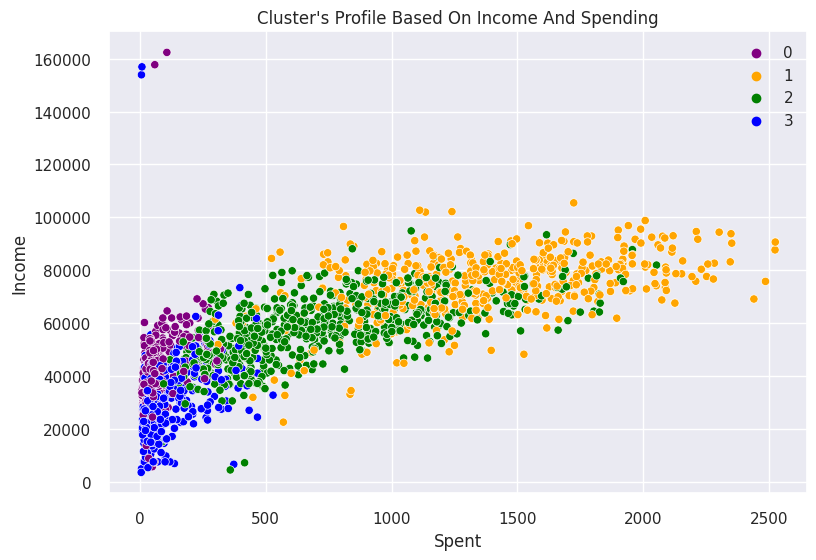
The count plot shows the number of customers in each cluster. Each color represents a different cluster.



**5.2 Cluster Profile based on Income and Spending**

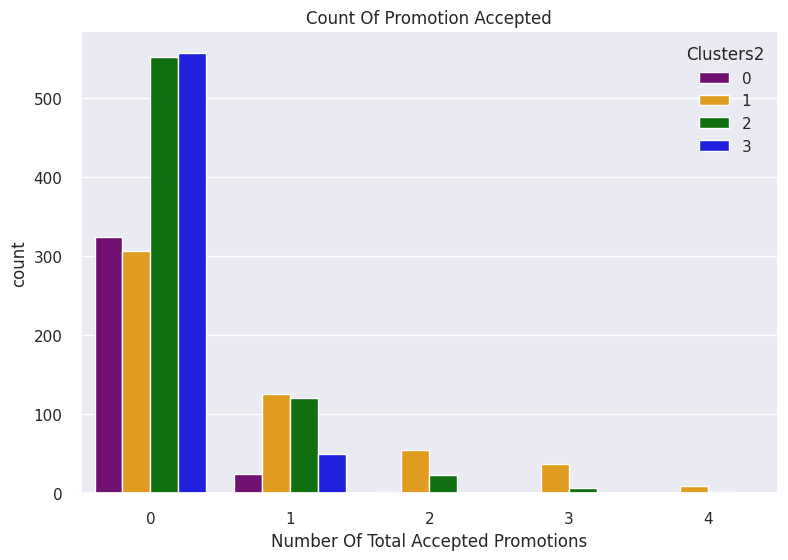
The scatter plot showcases the profile of each cluster based on income and spending habits of the customers. Each color denotes a unique cluster.

* group 0: low spending & average income
* group 1: high spending & high income
* group 2: average spending & average income
* group 3: low spending & low income



**5.3 Count of Total Campaign Accepted**

The count plot represents the number of campaigns accepted by customers in each cluster. It provides insight into the campaign preferences of different clusters and how did the campaign do in the past.



As can be seen, there has not been a good response to the campaigns so far. Perhaps better-targeted and well-planned campaigns are required to boost sales.

**5.4 Spending by Clusters**

The swarm and boxen plot combined provides a distribution and spread of spending habits across the clusters. Each cluster's spending patterns can be observed, which can guide marketing strategies.

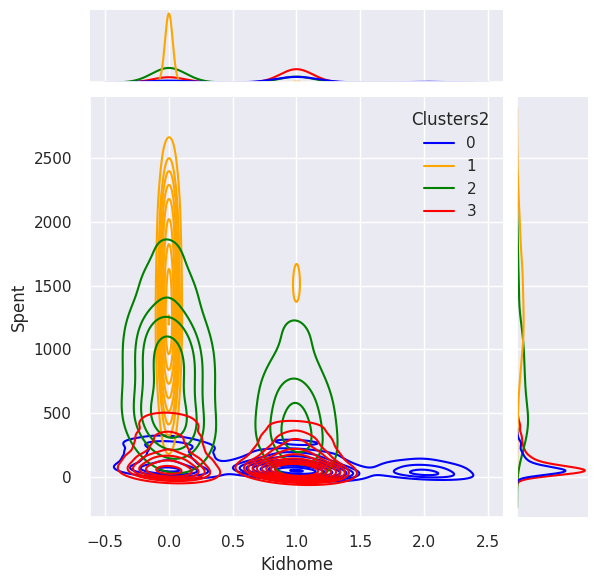
We can observe that clusters 1 and 2 are the biggest set of clusters. Hence, more attention is required for these clusters in marketing and future planning.

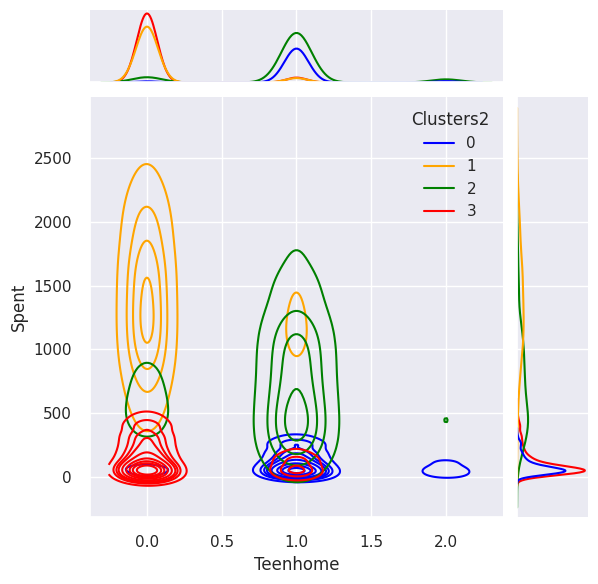
A picture containing diagram, screenshot

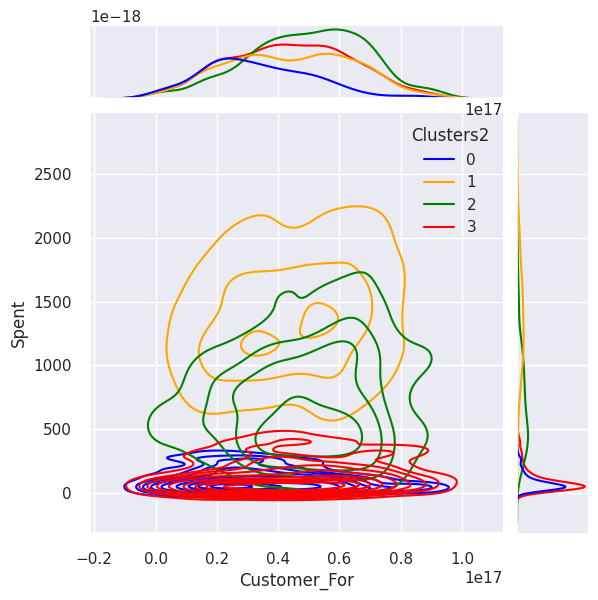
Description automatically generated

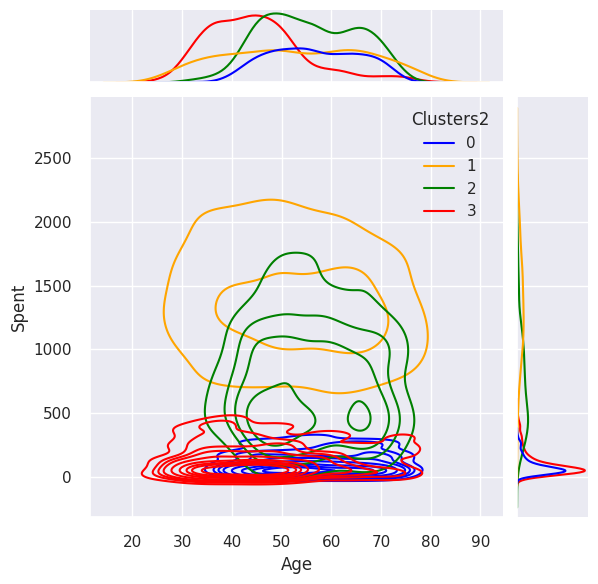
**5.5 Personal Attribute Distribution**

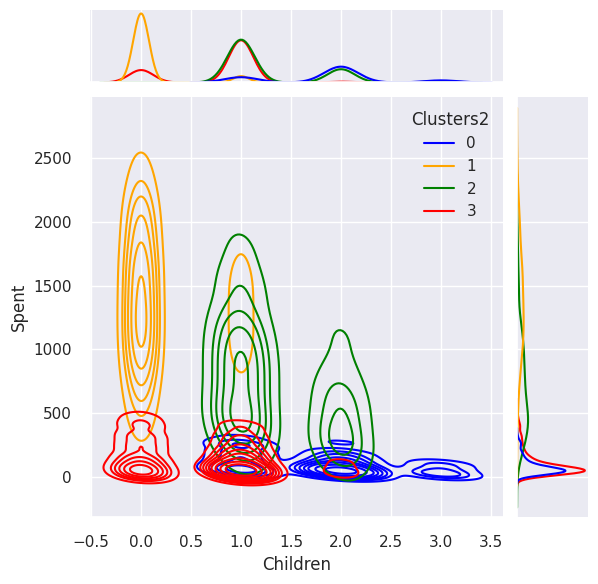
Joint KDE plots were generated for various personal attributes like 'Kidhome', 'Teenhome', 'Customer\_For', 'Age', 'Children', 'Family\_Size', 'Is\_Parent', 'Education', and 'Living\_With' with respect to the spending, color-coded by clusters. These plots depict how these personal attributes vary with the amount spent, and the variance within each cluster. I will be profiling the clusters formed and conclude about who is our star customer and who needs more attention from the marketing team.

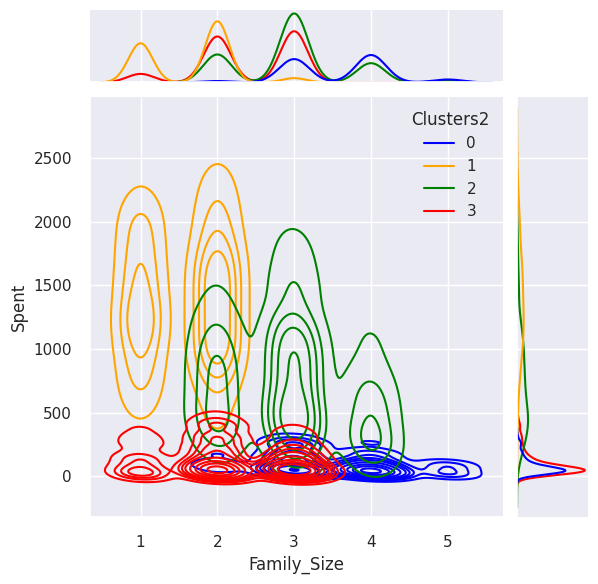


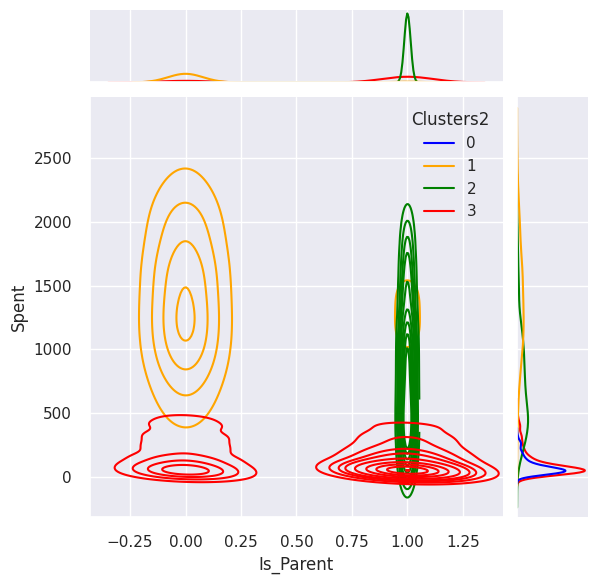


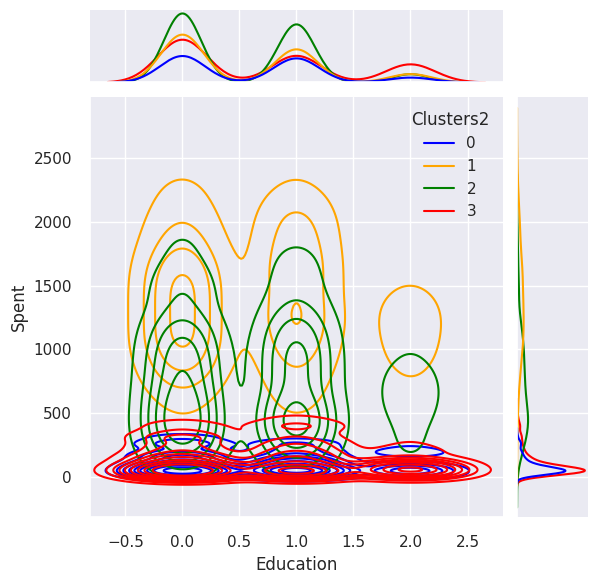


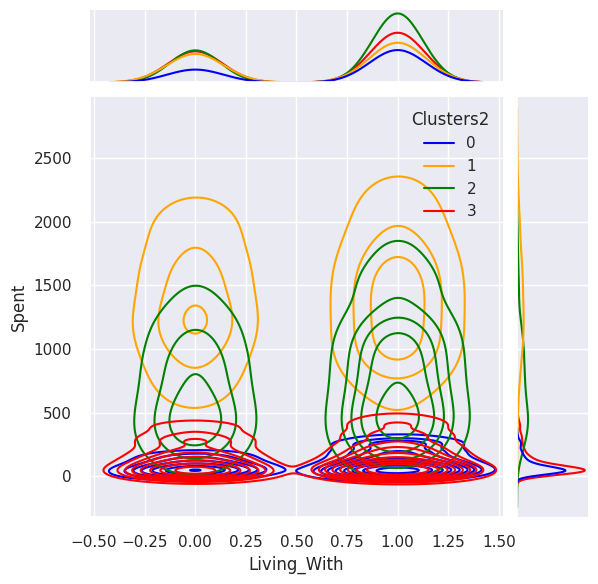












**Cluster number 0:** definitely a parent, max 4 members in the family and at least 2, some single parents as a subset in the group, most of them have a teenager at home and they are relatively older.

**Cluster number 1:** not a parent, amx 2 people in the family, more couples than single people, from all ages and high income group

**Cluster number 2:** majority of people are parents, max 3 people in the family, most of them have one kid and relatively younger

**Cluster number 3:** definitely a parent, max 5 people and at least 2 in the family, majority has a teenager at home and they are relatively older, a lower income group.

Conclusion

**Key Findings**

This analysis served as a demonstration of how machine learning can be used to segment and profile customers, providing useful insights that can guide marketing strategies.

By reducing the dimensionality of our data with PCA, we managed to retain most of the valuable information while making it easier to process. Subsequently, we applied various clustering techniques, of which KMeans appeared to provide the most cohesive customer segments.

The resulting customer segments allow us to see distinct groups of customers characterized by various traits. For example, we could identify clusters that respond more positively to promotional campaigns or spend more on average. These insights could potentially guide the tailoring of marketing campaigns to different customer segments, increasing their effectiveness and improving customer satisfaction. The main trends of the clusters are mentioned just above this section.

**Suggestions**

Other clustering algorithms such as DBscan can be attempted to be used, but trying to optimize the clustering number using the hyper parameter selection. Moreover, a larger dataset with more features can be used for a better clustering separation.