Report

Table of contents

1.Data Analysis

Our initial data exploration will involve analyzing each column of the dataset to understand its characteristics. We will utilize R's summary function to obtain a statistical overview of each variable.

```
corrplot 0.92 loaded


-- Attaching core tidyverse packages ---------------------- tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
head(data)
```

```
  X   ID Year_Birth  Education Marital_Status Income Kidhome Teenhome
1 1 5524       1957 Graduation         Single  58138       0        0
2 2 2174       1954 Graduation         Single  46344       1        1
3 3 4141       1965 Graduation       Together  71613       0        0
4 4 6182       1984 Graduation       Together  26646       1        0
5 5 5324       1981        PhD        Married  58293       1        0
6 6 7446       1967     Master       Together  62513       0        1
  Dt_Customer Recency MntWines MntFruits MntMeatProducts MntFishProducts
1  2012-09-04      58      635        88             546             172
2  2014-03-08      38       11         1               6               2
3  2013-08-21      26      426        49             127             111
4  2014-02-10      26       11         4              20              10
5  2014-01-19      94      173        43             118              46
6  2013-09-09      16      520        42              98               0
  MntSweetProducts MntGoldProds NumDealsPurchases NumWebPurchases
1               88           88                 3               8
2                1            6                 2               1
3               21           42                 1               8
4                3            5                 2               2
5               27           15                 5               5
```

6	42	14	2	6

	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3
1	10	4	7	0
2	1	2	5	0
3	2	10	4	0
4	0	4	6	0
5	3	6	5	0
6	4	10	6	0

	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact
1	0	0	0	0	0	3
2	0	0	0	0	0	3
3	0	0	0	0	0	3
4	0	0	0	0	0	3
5	0	0	0	0	0	3
6	0	0	0	0	0	3

	Z_Revenue	Response
1	11	1
2	11	0
3	11	0
4	11	0
5	11	0
6	11	0

```
dim(data)
```

```
[1] 2216    30
```

```
summary(data)
```

```
      X                ID           Year_Birth     Education
 Min.   :   1.0   Min.   :    0   Min.   :1893   Length:2216
 1st Qu.: 554.8   1st Qu.: 2815   1st Qu.:1959   Class :character
 Median :1108.5   Median : 5458   Median :1970   Mode  :character
 Mean   :1108.5   Mean   : 5588   Mean   :1969
 3rd Qu.:1662.2   3rd Qu.: 8422   3rd Qu.:1977
 Max.   :2216.0   Max.   :11191   Max.   :1996
 Marital_Status        Income          Kidhome          Teenhome
 Length:2216       Min.   :  1730   Min.   :0.0000   Min.   :0.0000
 Class :character  1st Qu.: 35303   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character  Median : 51382   Median :0.0000   Median :0.0000
                   Mean   : 52247   Mean   :0.4418   Mean   :0.5054
```

```
                        3rd Qu.: 68522   3rd Qu.:1.0000   3rd Qu.:1.0000
                        Max.    :666666  Max.    :2.0000  Max.    :2.0000
 Dt_Customer            Recency           MntWines         MntFruits
 Length:2216        Min.    : 0.00   Min.    :   0.0   Min.    :  0.00
 Class :character   1st Qu.:24.00   1st Qu.:  24.0   1st Qu.:  2.00
 Mode  :character   Median :49.00   Median : 174.5   Median :  8.00
                    Mean    :49.01   Mean    : 305.1   Mean    : 26.36
                    3rd Qu.:74.00   3rd Qu.: 505.0   3rd Qu.: 33.00
                    Max.    :99.00   Max.    :1493.0   Max.    :199.00
 MntMeatProducts   MntFishProducts   MntSweetProducts   MntGoldProds
 Min.    :   0.0   Min.    :  0.00   Min.    :  0.00   Min.    :  0.00
 1st Qu.:  16.0   1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00
 Median :  68.0   Median : 12.00   Median :  8.00   Median : 24.50
 Mean    : 167.0   Mean    : 37.64   Mean    : 27.03   Mean    : 43.97
 3rd Qu.: 232.2   3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00
 Max.    :1725.0   Max.    :259.00   Max.    :262.00   Max.    :321.00
 NumDealsPurchases NumWebPurchases   NumCatalogPurchases NumStorePurchases
 Min.    : 0.000   Min.    : 0.000   Min.    : 0.000   Min.    : 0.000
 1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.: 3.000
 Median : 2.000   Median : 4.000   Median : 2.000   Median : 5.000
 Mean    : 2.324   Mean    : 4.085   Mean    : 2.671   Mean    : 5.801
 3rd Qu.: 3.000   3rd Qu.: 6.000   3rd Qu.: 4.000   3rd Qu.: 8.000
 Max.    :15.000   Max.    :27.000   Max.    :28.000   Max.    :13.000
 NumWebVisitsMonth  AcceptedCmp3       AcceptedCmp4      AcceptedCmp5
 Min.    : 0.000   Min.    :0.00000   Min.    :0.00000   Min.    :0.0000
 1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
 Median : 6.000   Median :0.00000   Median :0.00000   Median :0.0000
 Mean    : 5.319   Mean    :0.07356   Mean    :0.07401   Mean    :0.0731
 3rd Qu.: 7.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.    :20.000   Max.    :1.00000   Max.    :1.00000   Max.    :1.0000
  AcceptedCmp1       AcceptedCmp2        Complain         Z_CostContact
 Min.    :0.00000   Min.    :0.00000   Min.    :0.000000   Min.    :3
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:3
 Median :0.00000   Median :0.00000   Median :0.000000   Median :3
 Mean    :0.06408   Mean    :0.01354   Mean    :0.009477   Mean    :3
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:3
 Max.    :1.00000   Max.    :1.00000   Max.    :1.000000   Max.    :3
   Z_Revenue        Response
 Min.    :11   Min.    :0.0000
 1st Qu.:11   1st Qu.:0.0000
 Median :11   Median :0.0000
 Mean    :11   Mean    :0.1503
```

```
3rd Qu.:11     3rd Qu.:0.0000
Max.    :11     Max.     :1.0000
```

1.1 Food Item Analysis

This section delves into the exploration of four food item categories in our dataset: wine, meat, fish, and fruit. We focus on identifying outliers and data distribution within these categories using boxplots.

No. of fish purchases   **No. of fruit purchases**
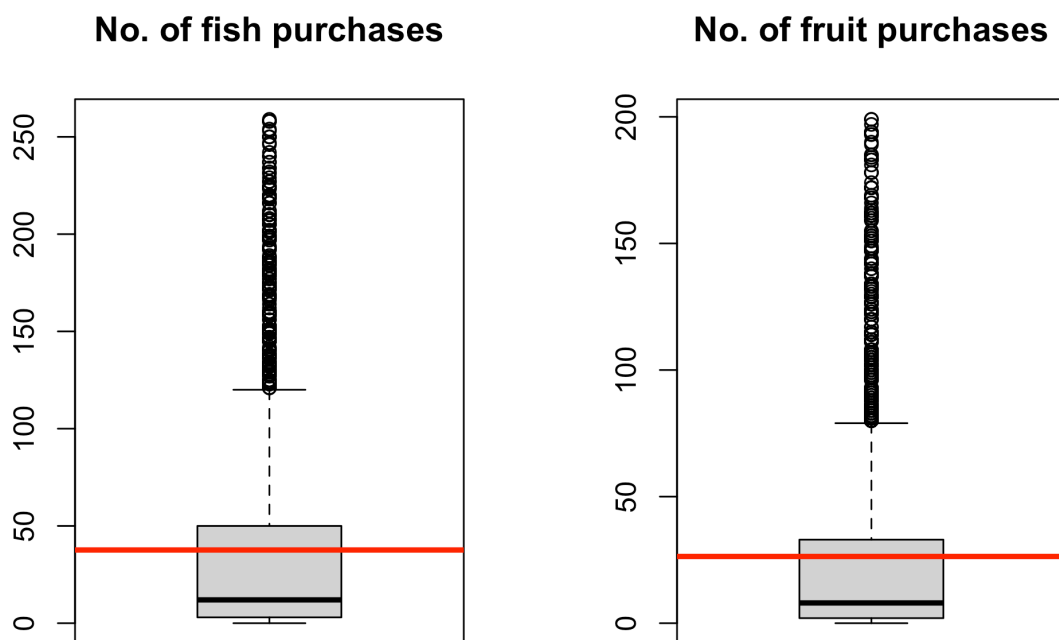
Our analysis revealed a significant presence of outliers in all four food item categories based on the boxplots. The boxes within the plots represent the interquartile range (IQR), encompassing the middle 50% of the data. Values falling outside the whiskers extending from the boxes are considered potential outliers. **We have made a horizontal red line along mean and which clearly shows that our mean and median differ from each other quite a bit in each food items.**

No. of wine purchases   **No of meat purchases**
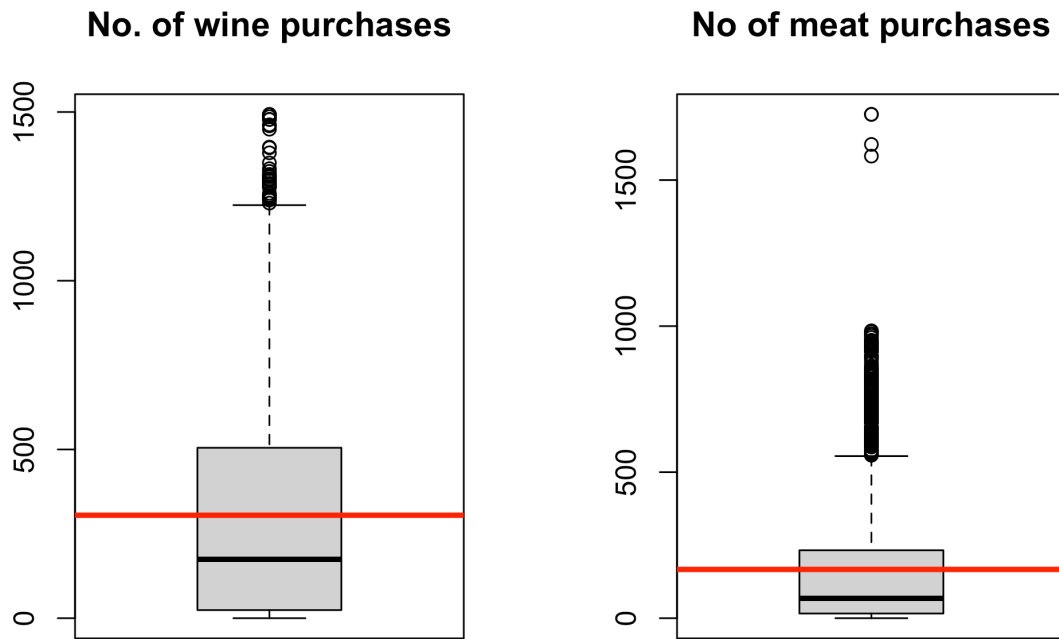
1.2 Mosaic Plots and Our Data

In this case, the mosaic plot will depict the proportion of individuals within each education level category (e.g. 2nd cycle, basic, graduation, masters, PhD) segmented by their marital status (e.g., married, single, together, divorced). The size of each rectangle will visually represent the percentage of people in that specific education level and marital status combination.

	Absurd	Alone	Divorced	Married	Single	Together	Widow	YOLO
2n Cycle	0	0	23	80	36	56	5	0
Basic	0	0	1	20	18	14	1	0
Graduation	1	1	119	429	246	285	35	0
Master	1	1	37	138	75	102	11	0
PhD	0	1	52	190	96	116	24	2

Mosaic plot Education vs Marital status
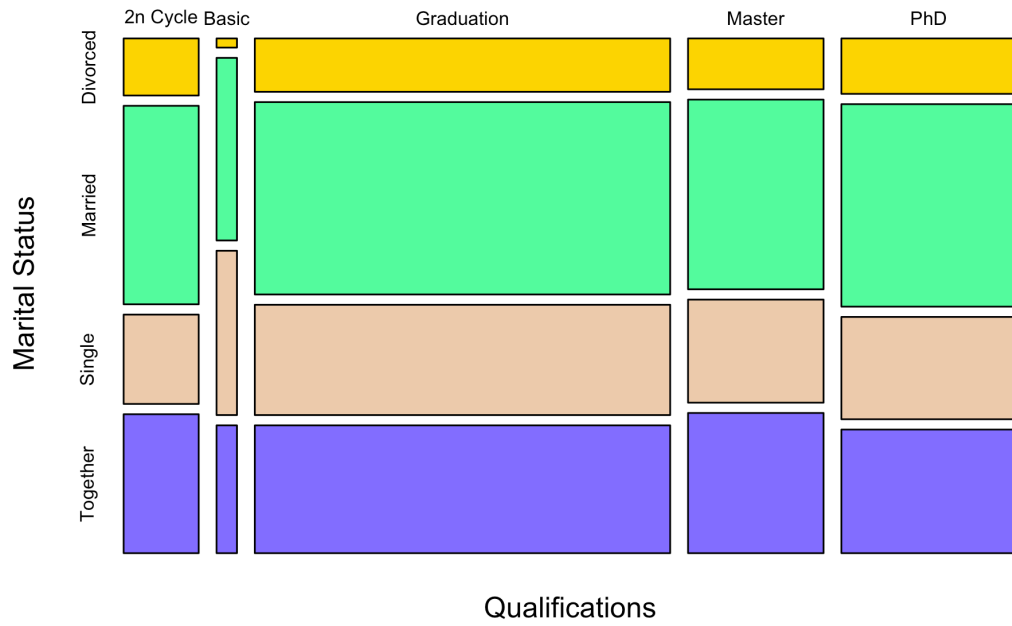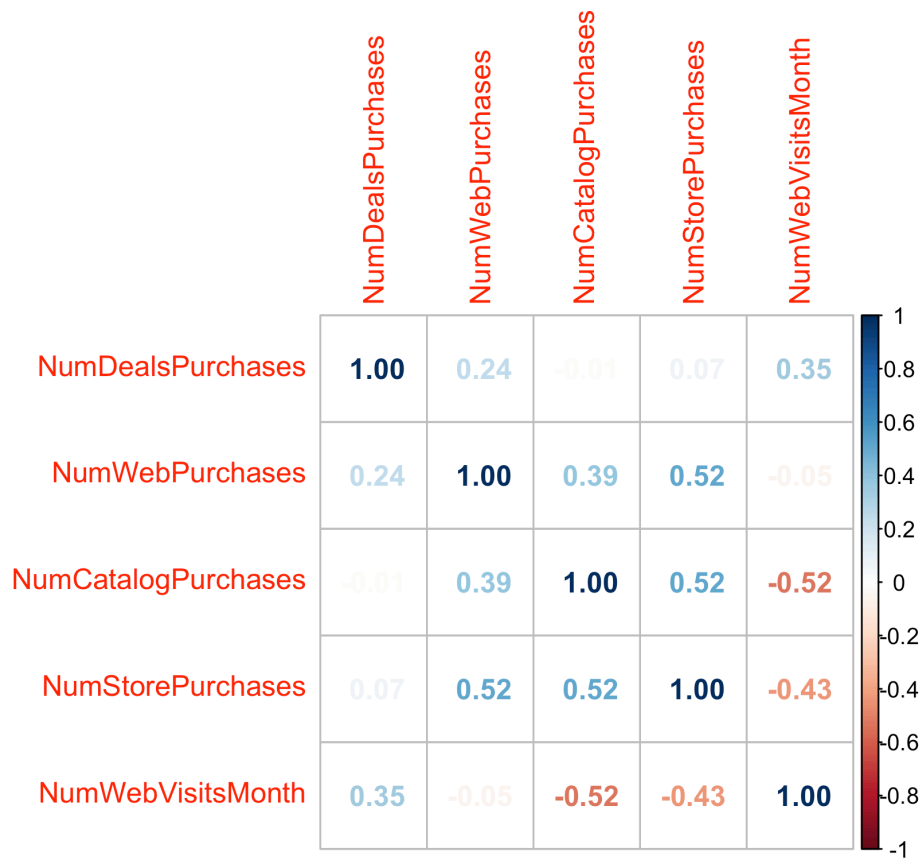
1.3 Correlation Plot

1.3.1 Plot 1: Correlations Between All Products

This correlation plot reveals a positive correlation (likely depicted by blue color) between all the products, including seats, fruits, wine, meat, fish, and gold. Positive correlation signifies that when the value of one product increases, the values of other products tend to increase as well, and vice versa.

	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
MntWines	1.00	0.39	0.57	0.40	0.39	0.39
MntFruits	0.39	1.00	0.55	0.59	0.57	0.40
MntMeatProducts	0.57	0.55	1.00	0.57	0.54	0.36
MntFishProducts	0.40	0.59	0.57	1.00	0.58	0.43
MntSweetProducts	0.39	0.57	0.54	0.58	1.00	0.36
MntGoldProds	0.39	0.40	0.36	0.43	0.36	1.00
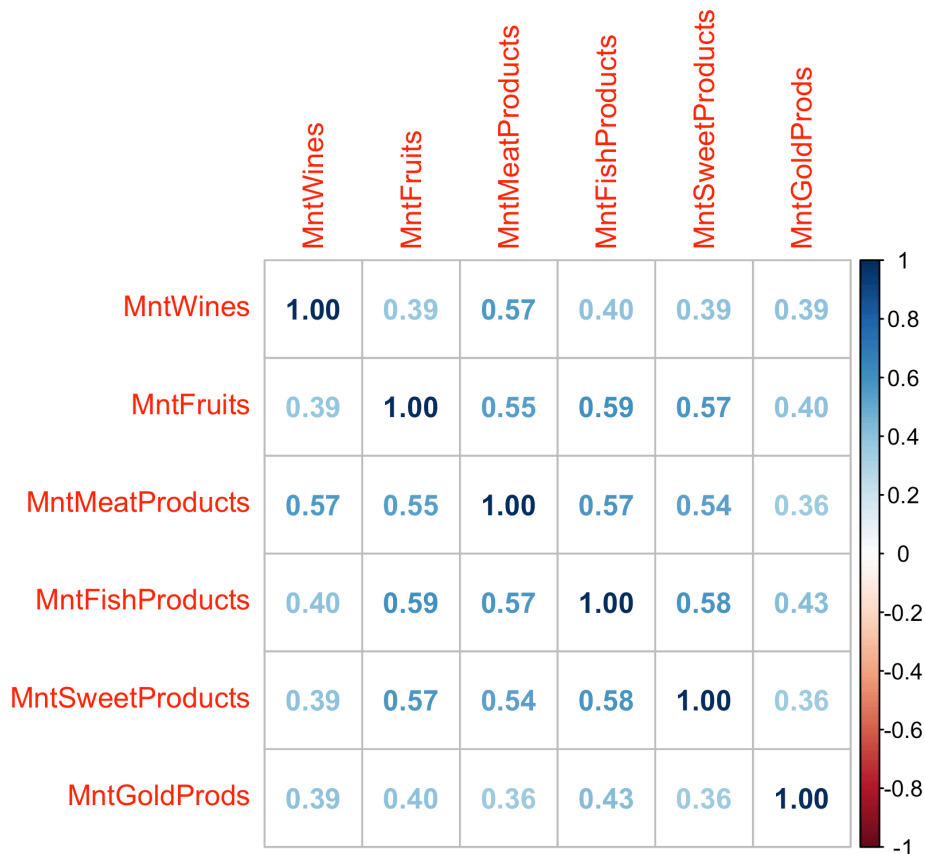
1.3.2 Plot 2: Correlations Between Purchase Sources

This correlation plot examines the relationship between purchase sources, such as web, catalog, store, and potentially others. **We can see a negative correlation between the Number of web visits and (Number of store and catalog purchases) and we can infer from it that tech savvy people prefer less to go physically to stores**

	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
NumDealsPurchases	1.00	0.24	-0.01	0.07	0.35
NumWebPurchases	0.24	1.00	0.39	0.52	-0.05
NumCatalogPurchases	-0.01	0.39	1.00	0.52	-0.52
NumStorePurchases	0.07	0.52	0.52	1.00	-0.43
NumWebVisitsMonth	0.35	-0.05	-0.52	-0.43	1.00

1.3 Scatter plot between premium products

Another analysis is using scatter plot where we can see that the buyers of the premium products of company like sweets and Wines have more income (Black color). And people with less income(Yellow color) didn't buy these products.
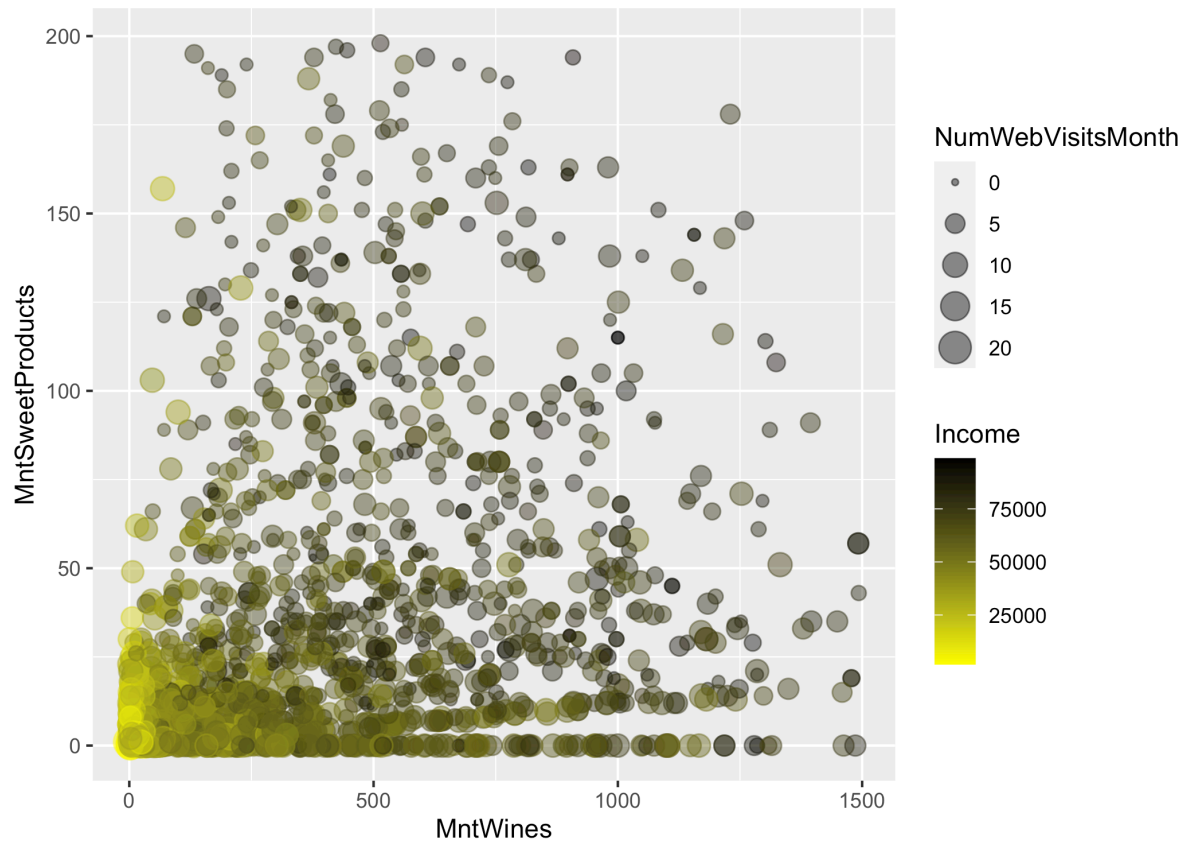
1.5 Amount spend on products (Kid vs No Kid)

his analysis show the total expenditure per customer for different products and with category kid vs no kid. It can be clearly seen that customer who have kids at their home spend more in every product category.

Kid vs No Kids
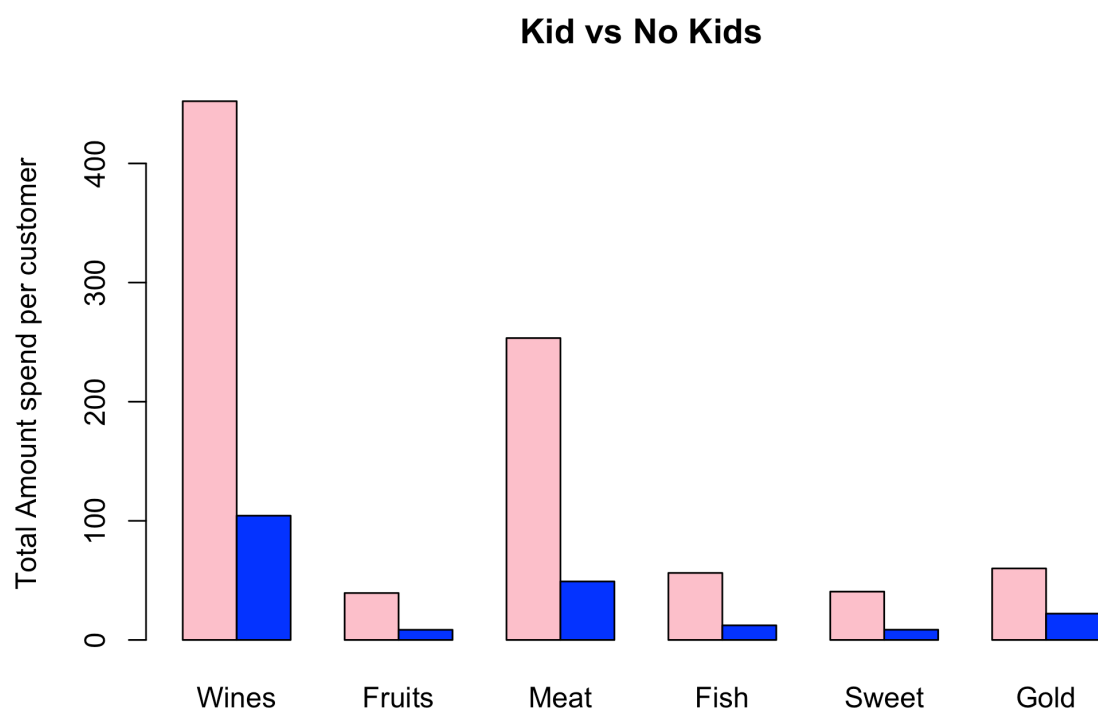
2. Maximum Likelihood Estimator

Now we tried to estimate our one of the variable NumStorePurchases using MLE. First we saw it's frequency shape using a histogram. And since it's a discrete variable. We assumed that they follow poisson distribution which is a good guess. Then using the optim function we found the best value of lambda. Below bar plot shows the density of our columns and red line shows our MLE.
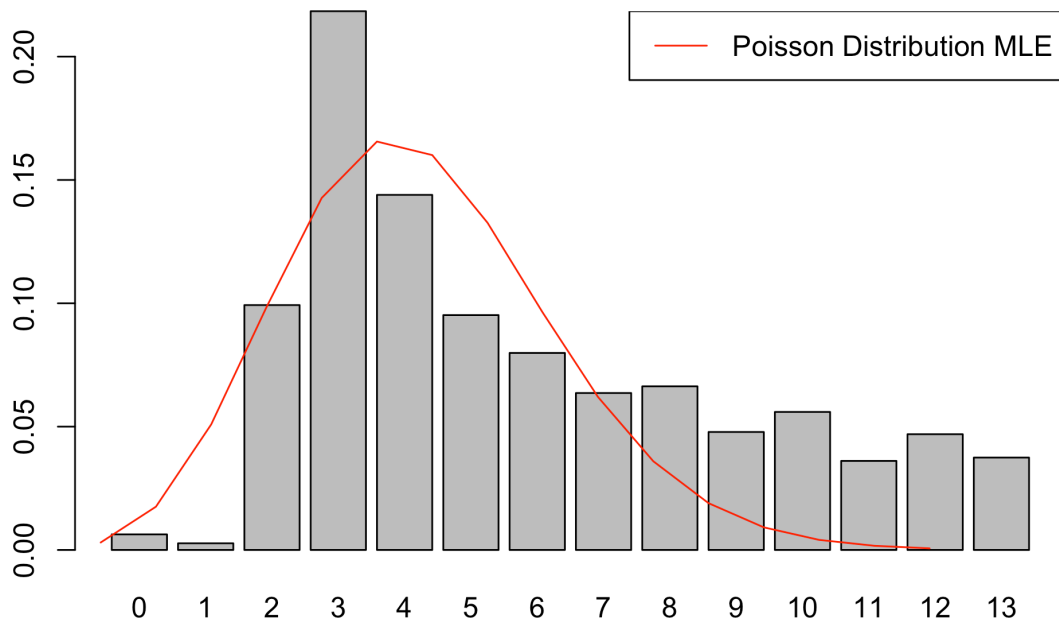
We can check the quality of our estimator using the qqplots. Below we have plotted a QQplot of our data wrt data from MLE. And we can see that mostly points lie on the unit slope line from which we can infer that they are similar in distribution. Whereas when we take the parameter of our distribution randomly let's say 10 then the points doesn't lie on the unit slope line

Using MLE

Random lambda=10

Quantiles of our actual data

Theoretical quantiles of Poiss(lam)
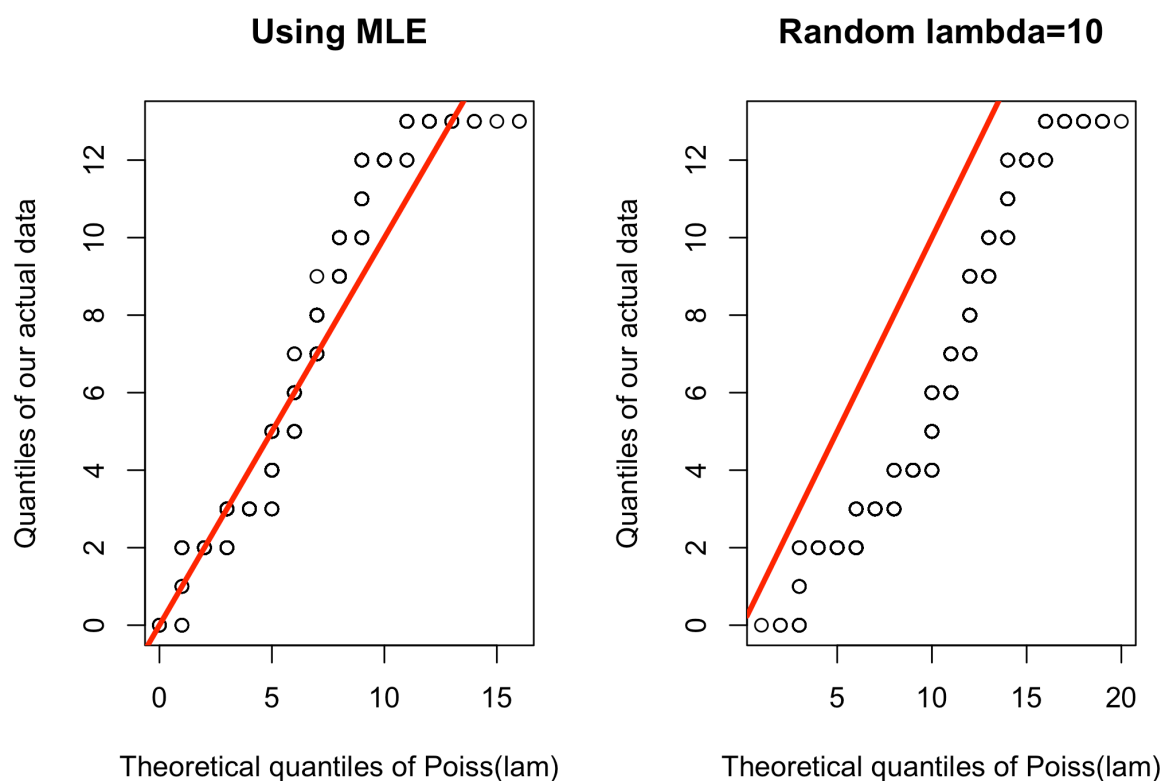
3. Principal Component Analysis

The data we have was high dimensional data with multiple columns which led to difficulty in organizing and analyzing the various components. We thus used PCA to reduce the dimensionalty of our data.

First step was to clean and filter out the data which was not required. We removed the categorical variables as they were increasing the complexity of the data. Further, we removed the columns with zero variance and the rows with missing values. Following is the head of the cleaned data:

```
    X    ID Year_Birth Income Kidhome Teenhome Recency MntWines MntFruits
1 1 5524       1957  58138       0        0      58      635        88
2 2 2174       1954  46344       1        1      38       11         1
3 3 4141       1965  71613       0        0      26      426        49
4 4 6182       1984  26646       1        0      26       11         4
5 5 5324       1981  58293       1        0      94      173        43
  MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds
1             546             172               88           88
2               6               2                1            6
3             127             111               21           42
```

4	20	10	3	5
5	118	46	27	15

	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
1	3	8	10	4
2	2	1	1	2
3	1	8	2	10
4	2	2	0	4
5	5	5	3	6

	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1
1	7	0	0	0	0
2	5	0	0	0	0
3	4	0	0	0	0
4	6	0	0	0	0
5	5	0	0	0	0

	AcceptedCmp2	Complain	Response	age_at_enroll	days_customer_for
1	0	0	1	55	663
2	0	0	0	60	113
3	0	0	0	48	312
4	0	0	0	30	139
5	0	0	0	33	161

As we can see, there are 27 columns even after cleaning the data, which shows the high dimensionalty.

Then, we performed PCA which gave the following result:

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5643 1.57840 1.4001 1.33304 1.13183 1.03342 1.02668
Proportion of Variance 0.2435 0.09227 0.0726 0.06582 0.04745 0.03955 0.03904
Cumulative Proportion  0.2435 0.33581 0.4084 0.47423 0.52168 0.56123 0.60027
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     1.00059 0.97605 0.96592 0.92873 0.8757 0.85637 0.78433
Proportion of Variance 0.03708 0.03528 0.03456 0.03195 0.0284 0.02716 0.02278
Cumulative Proportion  0.63735 0.67263 0.70719 0.73914 0.7675 0.79470 0.81749
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.77386 0.75817 0.75284 0.71983 0.67202 0.65333 0.63793
Proportion of Variance 0.02218 0.02129 0.02099 0.01919 0.01673 0.01581 0.01507
Cumulative Proportion  0.83967 0.86096 0.88195 0.90114 0.91786 0.93367 0.94875
                          PC22    PC23    PC24    PC25    PC26    PC27
Standard deviation     0.61110 0.55780 0.51989 0.48397 0.4410 0.01660
Proportion of Variance 0.01383 0.01152 0.01001 0.00867 0.0072 0.00001
Cumulative Proportion  0.96258 0.97410 0.98411 0.99279 1.0000 1.00000
```

Below is the plot of the PCA analysis and the cumulative variance of the components:

pc
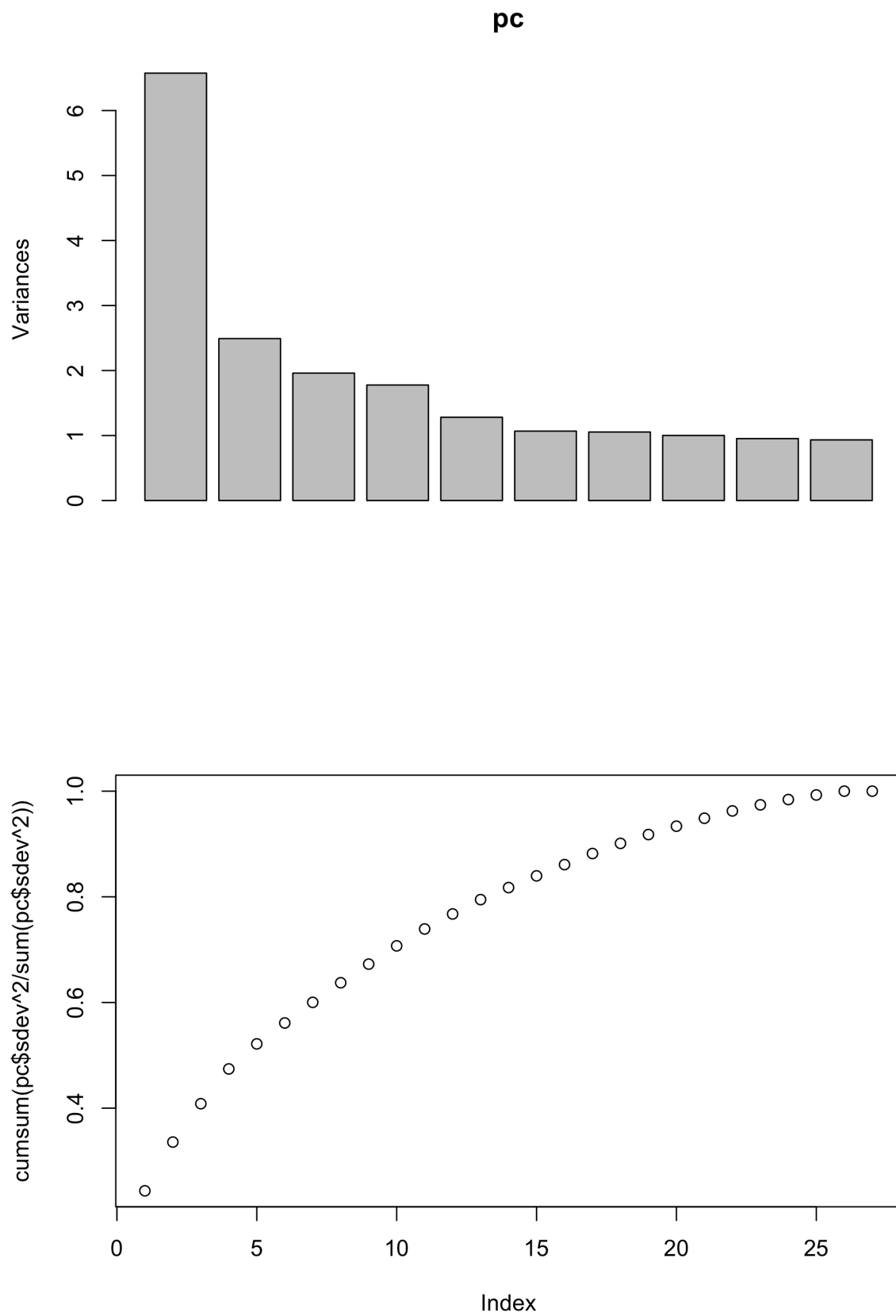
Thus, around 13 components are able to explain 80% variability in the data.

Below is a plot of the relationship between the first two components after PCA.
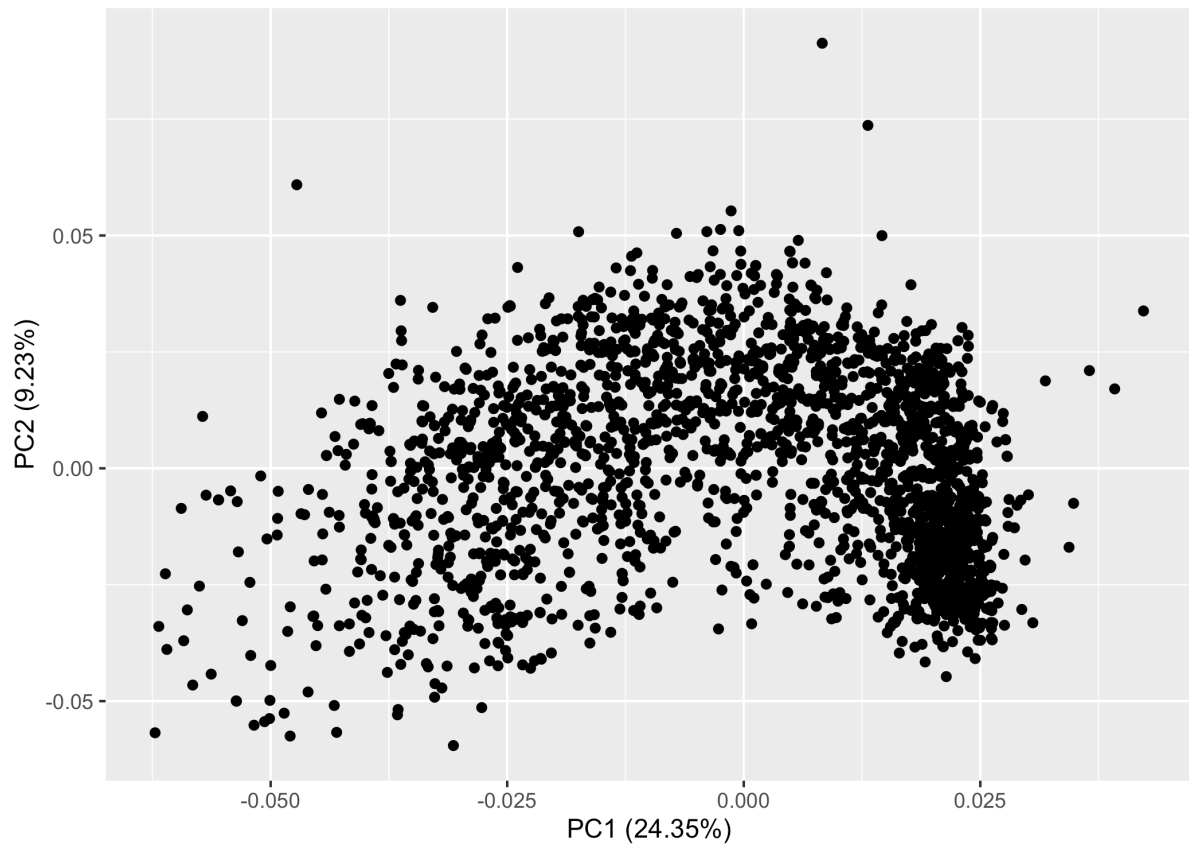

Thus, with the help of PCA we can reduce the data with 27 columns to upto 13 columns and still explain 80% variability in the data.

4. K-Means Clustering

After reducing the dimensionality, we moved on to clustering of the data to segment the customers according to their spending behaviors. For clustering, we considered the first 3 components from the PCA reduction.

We first applied the elbow method to find the optimum number of clusters

Elbow Method
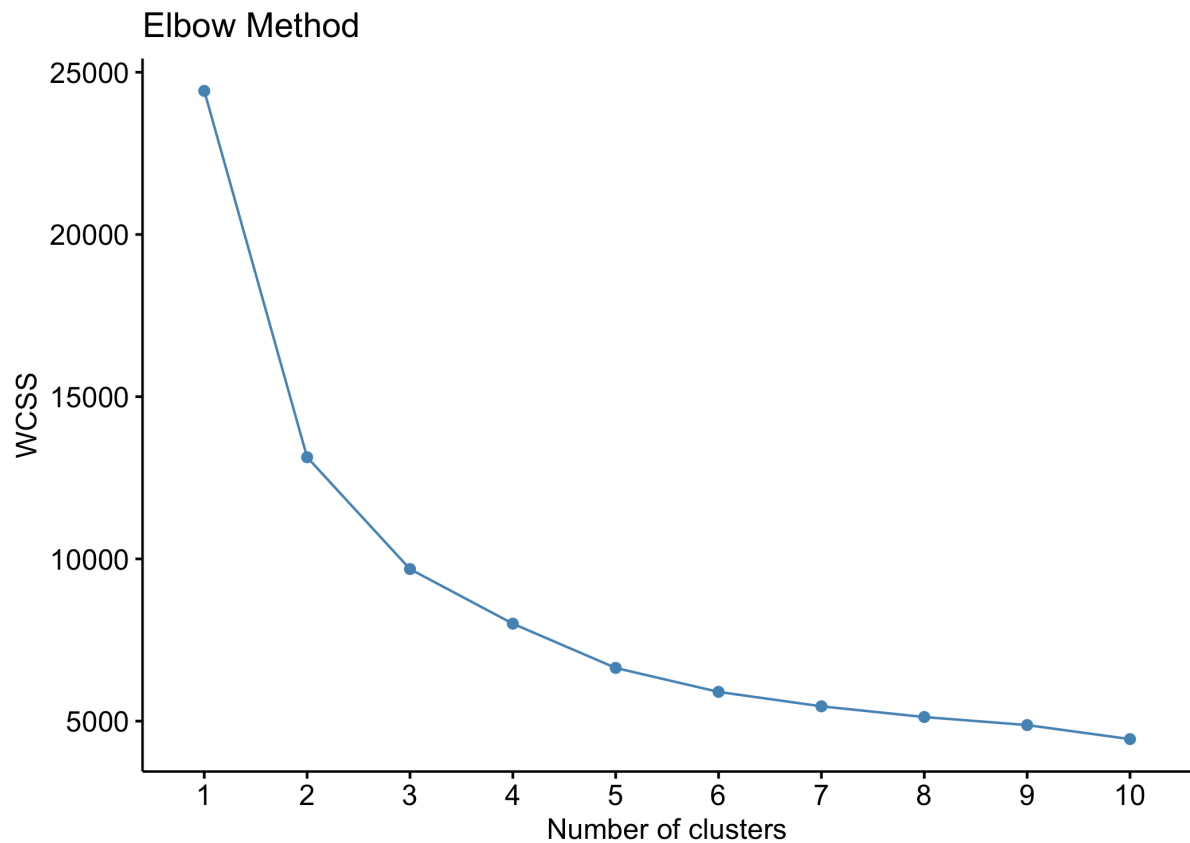
As is evident from the graph, the optimal number of clusters is 4, as after that there is not a significant reduction in the WCSS even when we increase the number of clusters.

Clusters of customers
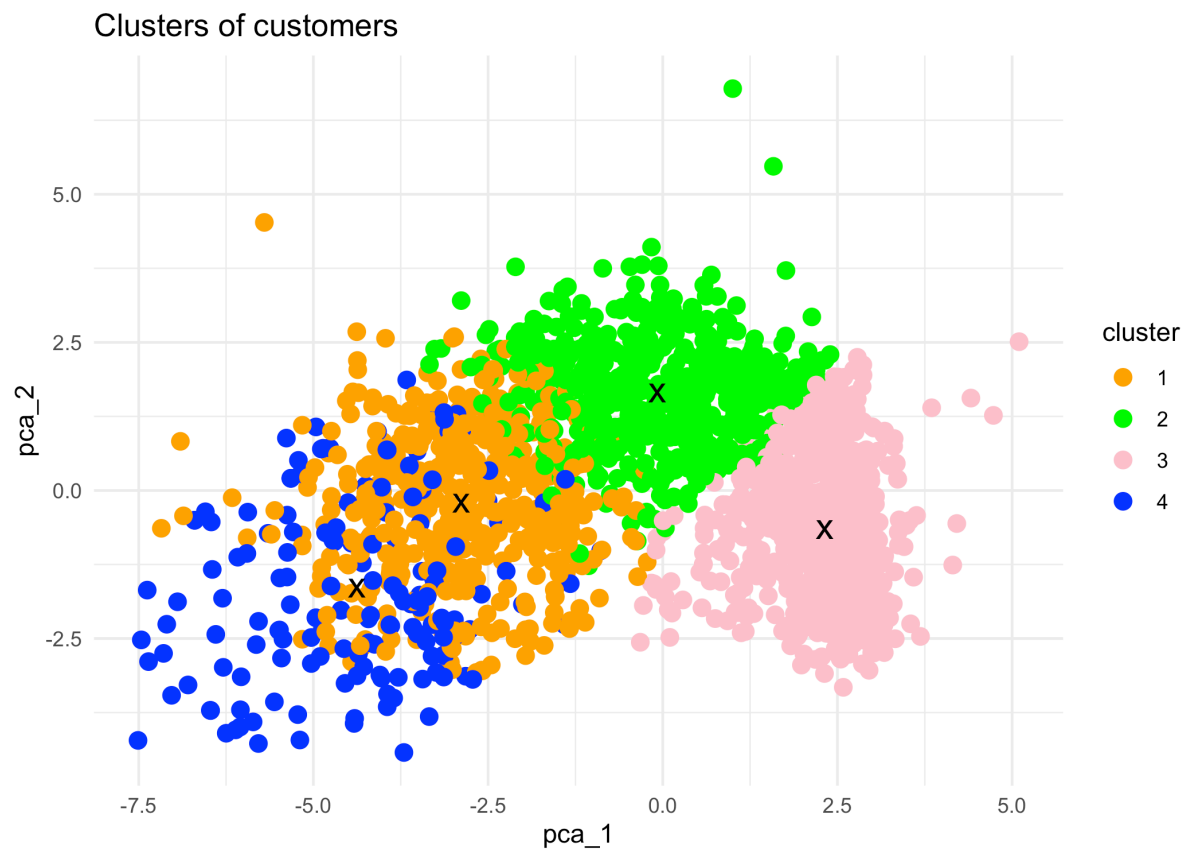
The figure above represents the different clusters of customers based and their relationship with the first and second PCA components.

Below we have a summary of the different characteristics of all the different clusters:

	V1	V2	V3	V4
cluster	1	2	3	4
Income	72832.08	55567.16	33837.31	80414.98
num_children	0.3299595	1.2112211	1.2286617	0.2108434
Kidhome	0.04858300	0.24752475	0.83667018	0.06024096
Teenhome	0.2813765	0.9636964	0.3919916	0.1506024
num_total_purchases	23.78138	24.80033	14.55427	25.04217
total_Mnt	1228.2146	614.9373	100.0875	1631.5904

4.1 Inference on the basis 0f Clusters

4.1.1 Plot 1: Income and Spending Patterns

Income by Cluster

Total Spending by Cluster
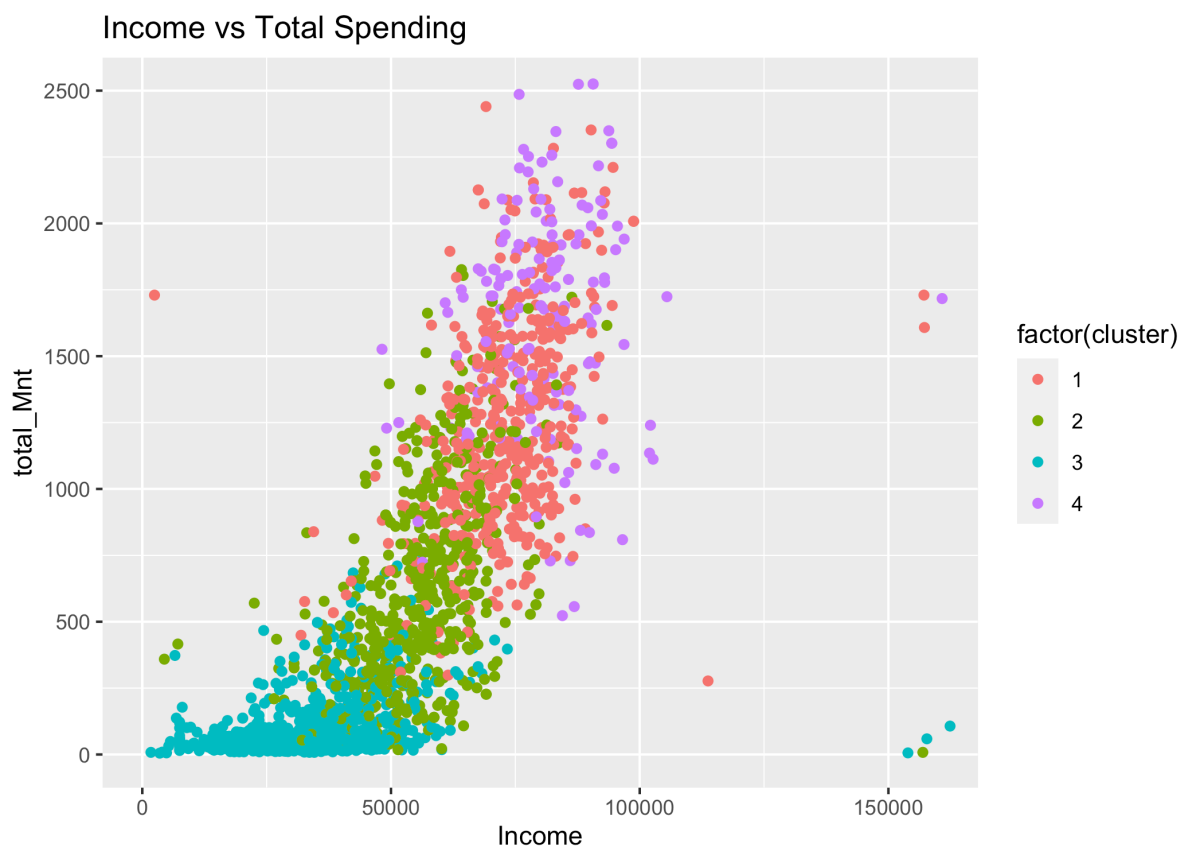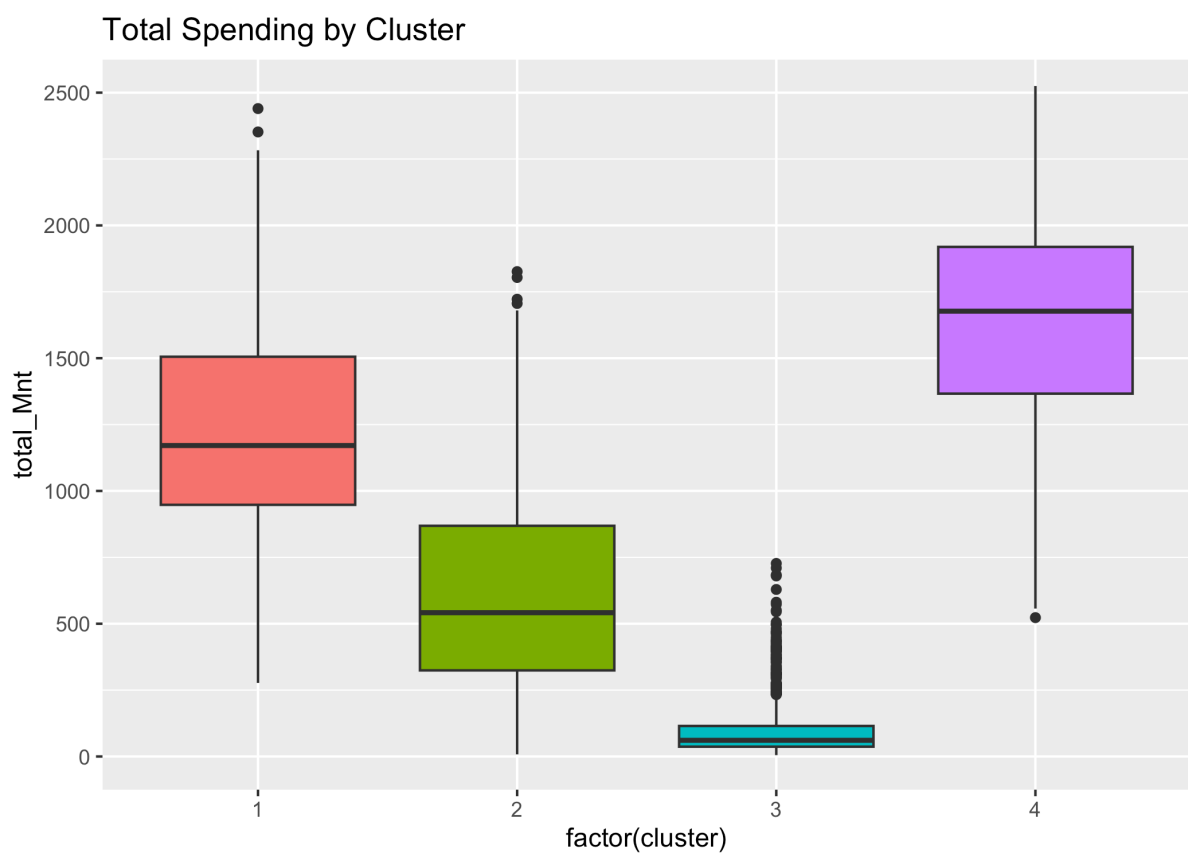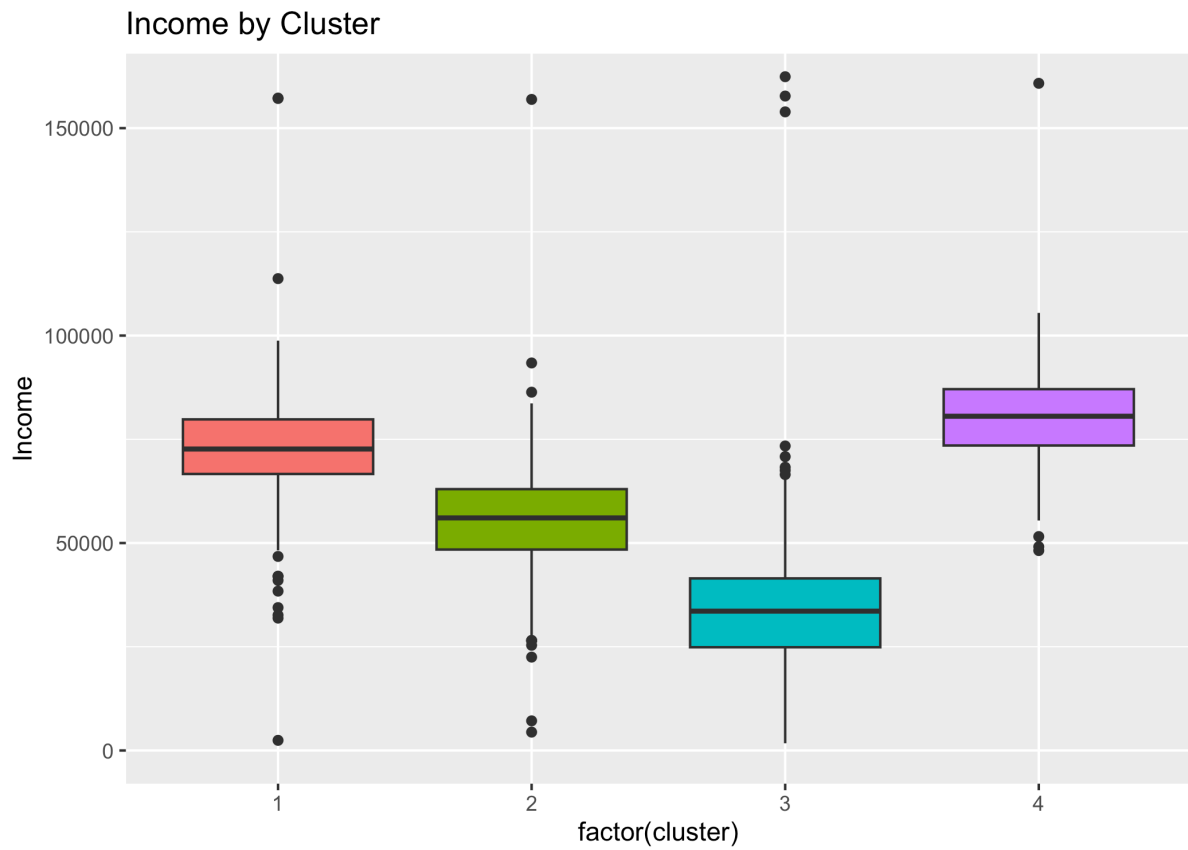
As is evident by the plots:

Cluster 1: This cluster is characterised by high income and high spending pattern customers. Their spending ranges from 1000 - 1500 units.

Cluster 2: This cluster is characterized by middle income and middle spending pattern customers. They also have a high variability in spending ranging from around 250 - 800 units.

Cluster 3: This cluster contains customers with low income and low spending patterns. They have small variability in spending ranging from just around 10 - 100 units.

Cluster 4: This cluster is characterized by customers with highest income and highest spending patterns. They also have a large variability in spending patterns ranging from around 1400 - 1900 units.

4.1.2 Plot 2: Number of Children

Count of Clusters by Number of Children

Count of Clusters by Kidhome

Count of Clusters by Teenhome
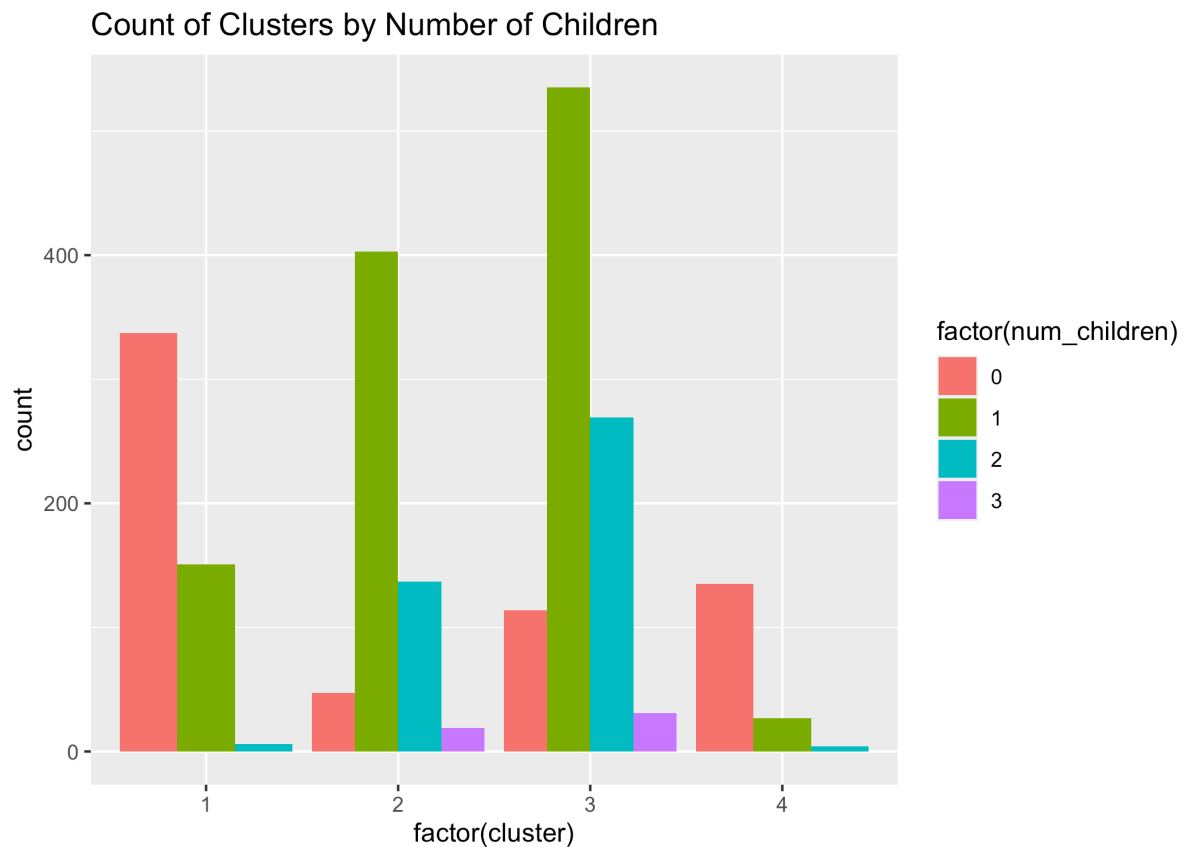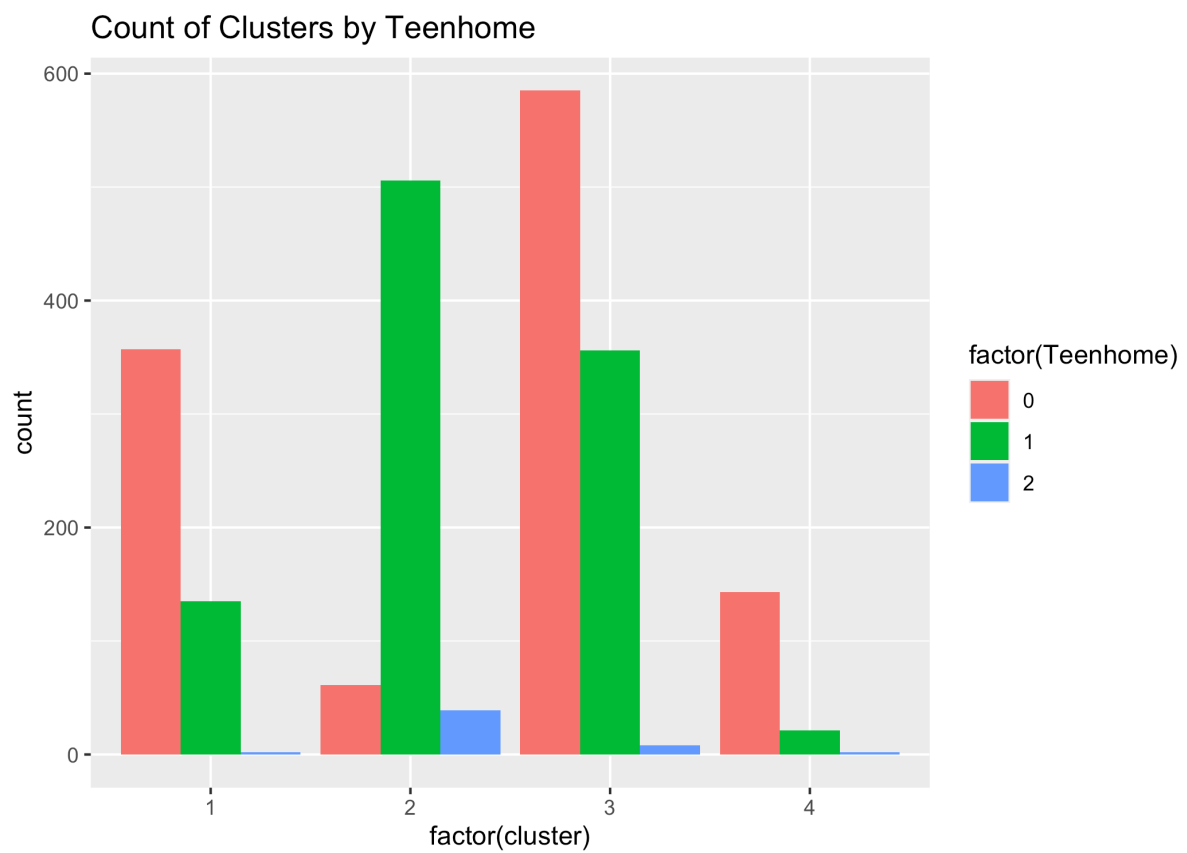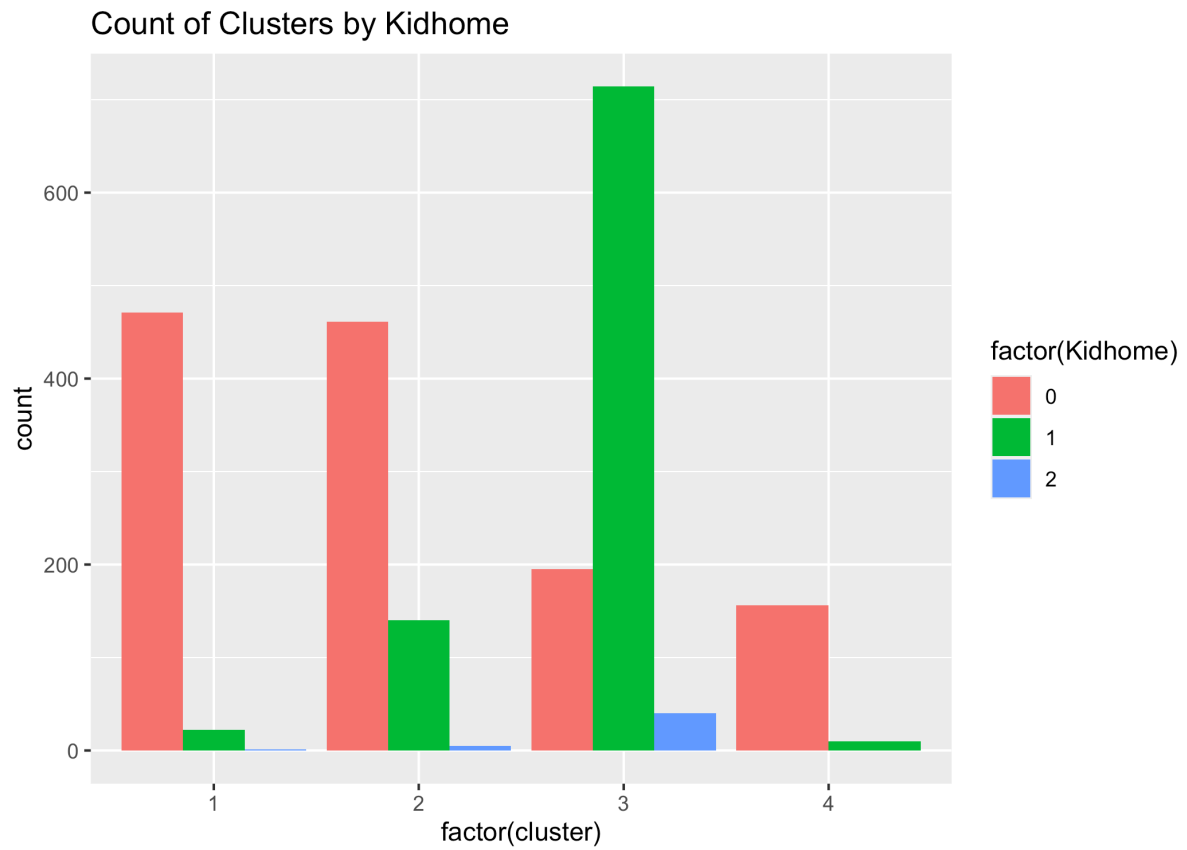
Our analysis reveals distinct clusters based on income, spending habits, and the presence of children within each group:

Cluster 1: Characterized by high income and high spending, with fewer children on average (0.13), predominantly having one child.

Cluster 4: Represents middle-income individuals with moderate spending patterns. This group averages 1.16 children, with a higher proportion of teenagers (0.95) than younger children (0.21).

Cluster 2: Comprises individuals with low income and spending, yet the highest number of children overall. With an average of 1.73 children, this group primarily consists of teenagers (1.02) with fewer younger children (0.71).

Cluster 3: Exhibits the lowest income and spending levels, with a moderate number of children (0.81). Notably, most children in this group are teenagers.

Understanding these clusters aids in tailoring marketing strategies and product offerings to better meet the diverse needs of customers across income levels and family demographics.

4.1.3 Plot 3: Age Analysis
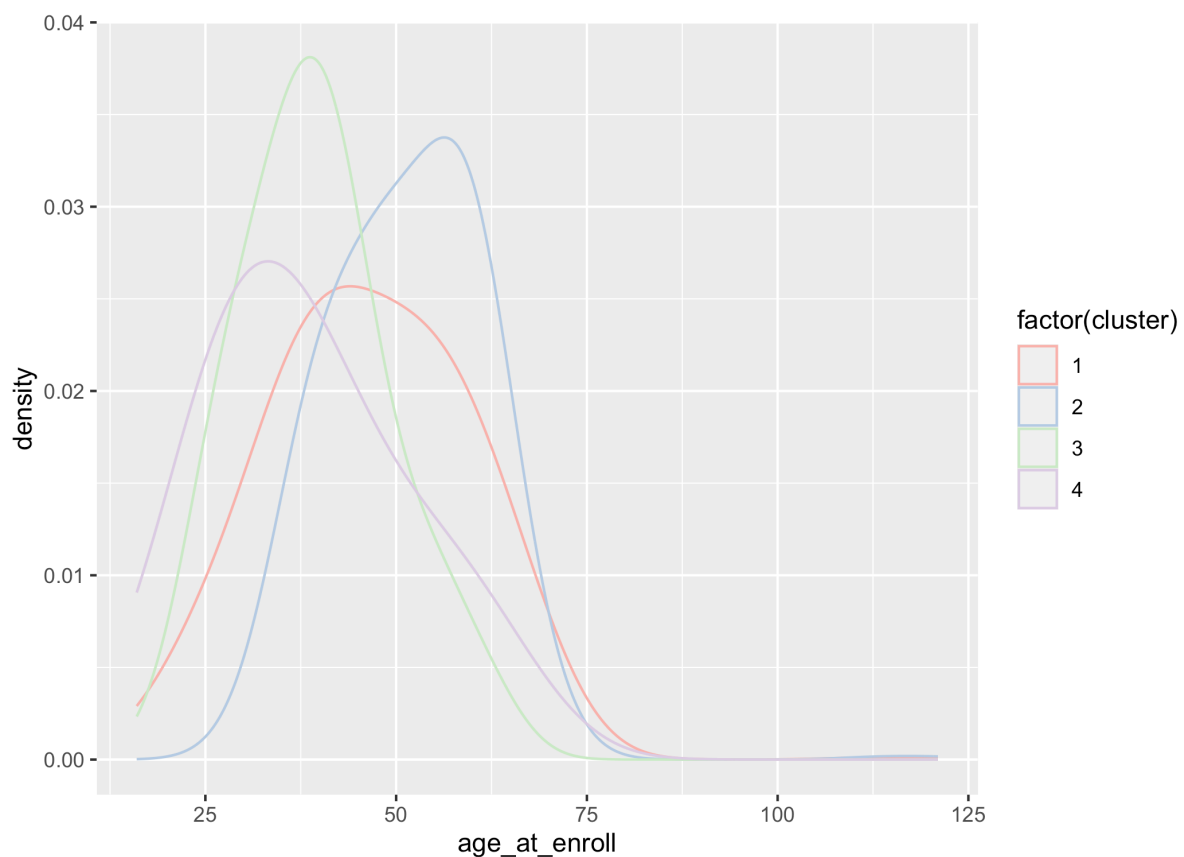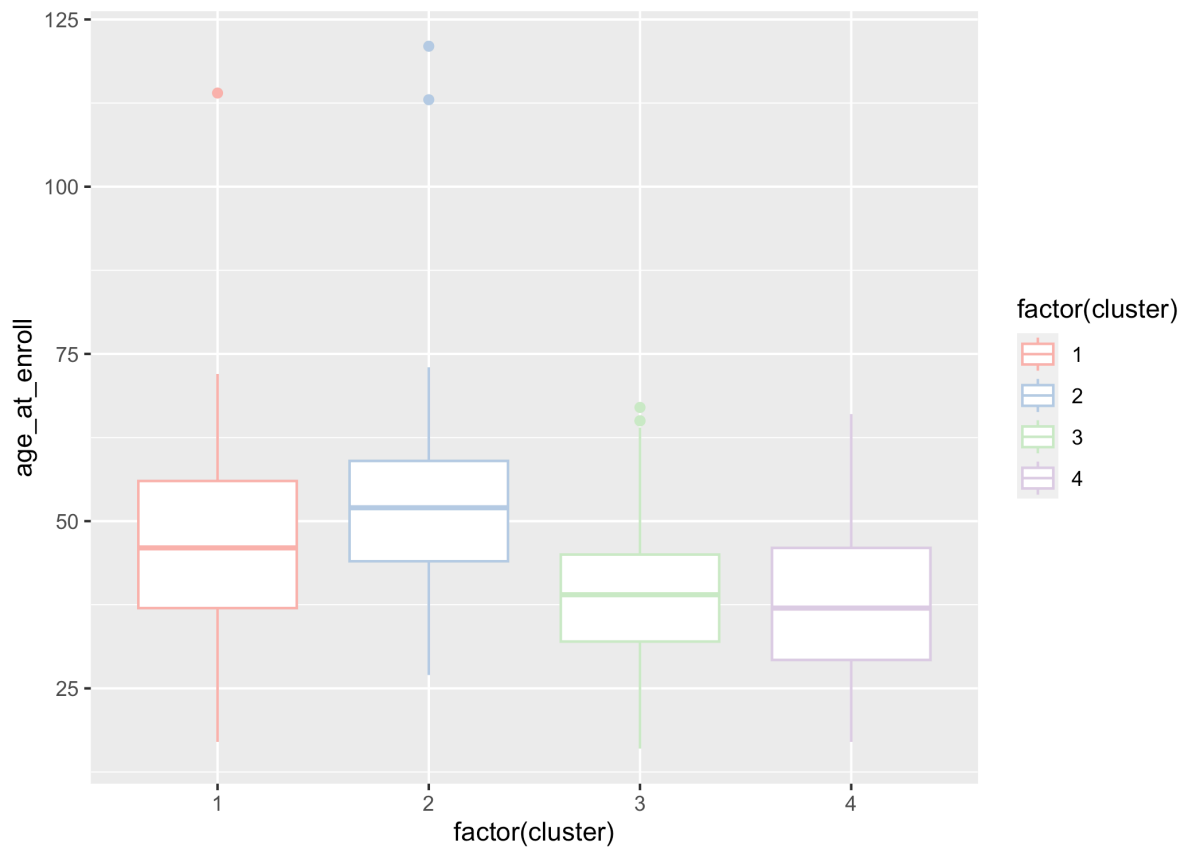
Based on our analysis, we've identified distinct age-income dynamics among different customer segments:

High-Income Group: Typically falls within the age range of 40-45 years old.

**Lowest and Middle-Income Groups**: Predominantly consist of individuals aged 30-35 years old.

Second Lowest Income Group: Shows the highest average age, centered around 50 years old.

Middle Income Customers: Mostly comprise youngsters above the age of 27.

By aligning strategies with these insights, businesses can optimize their approach to engage with customers across varying income levels and age groups.

4.1.4 Plot 4: Spending Pattern on different Products

Percentage of Spending by Category and Cluster
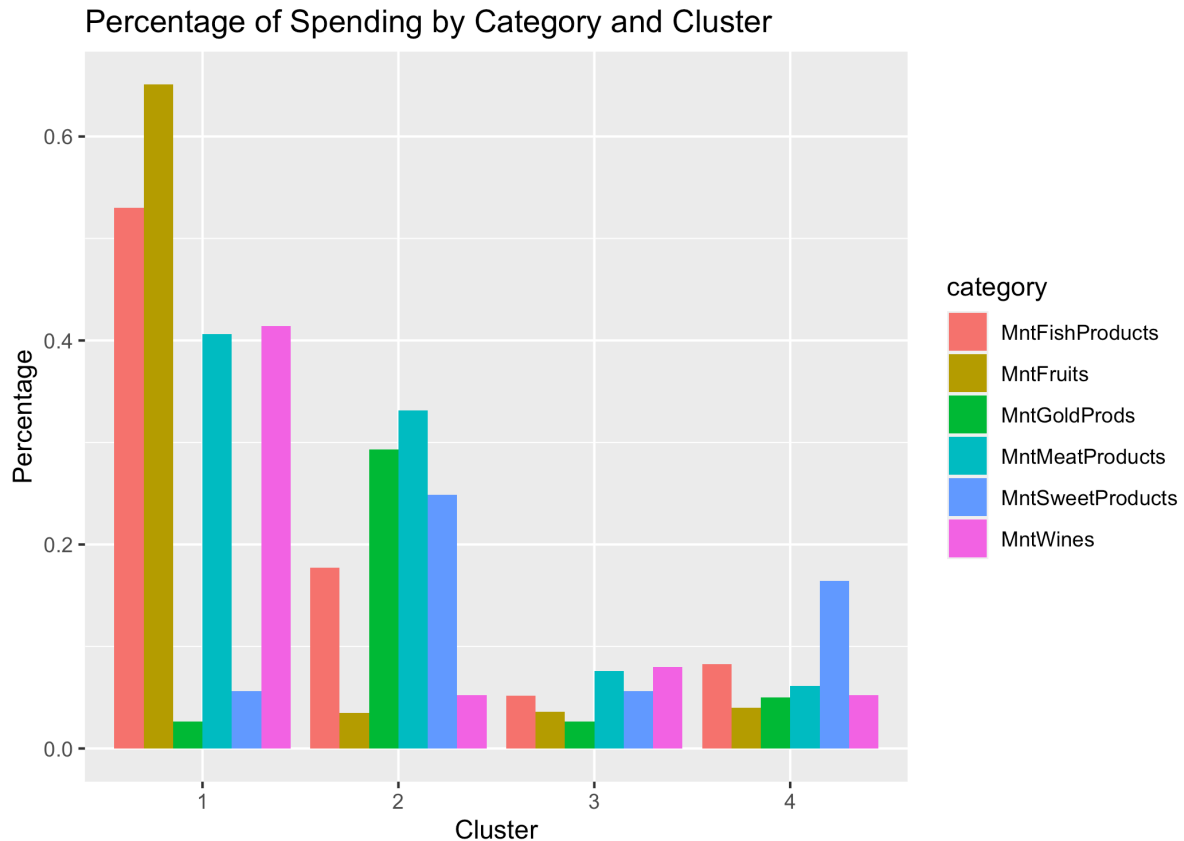
For food category, we can see that:

The clusters spend on the category pretty much follows their overall spend. Cluster 1 spends the most on food, followed by cluster 2, cluster 3 and cluster 0.

All clusters spend more than wines, followed by meat products.

Cluster 3 and cluster 1 spend significantly more (over half of their total spend) on wine compared to other categories, then around 20% on meat products

Cluster 4 spend more balanced than other clusters, this cluster spend more on gold than any other clusters. Interestingly, this cluster and cluster 0 are both lower income and lower spend clusters, but they spend more on gold than other clusters.

**Cluster 2** spend around 45% on wine and 35% on meat products, they spend more than meat than other clusters

4.1.5 Plot 5: Purchase Preferences

	cluster	NumDealsPurchases	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4
1	1	752	34	0	15	19
2	2	2200	14	6	35	82

| 3 | 3 | 1980 | 1 | 2 | 76 | 2 |
| 4 | 4 | 213 | 93 | 22 | 37 | 61 |

	AcceptedCmp5	Response
1	42	50
2	5	73
3	0	86
4	115	124
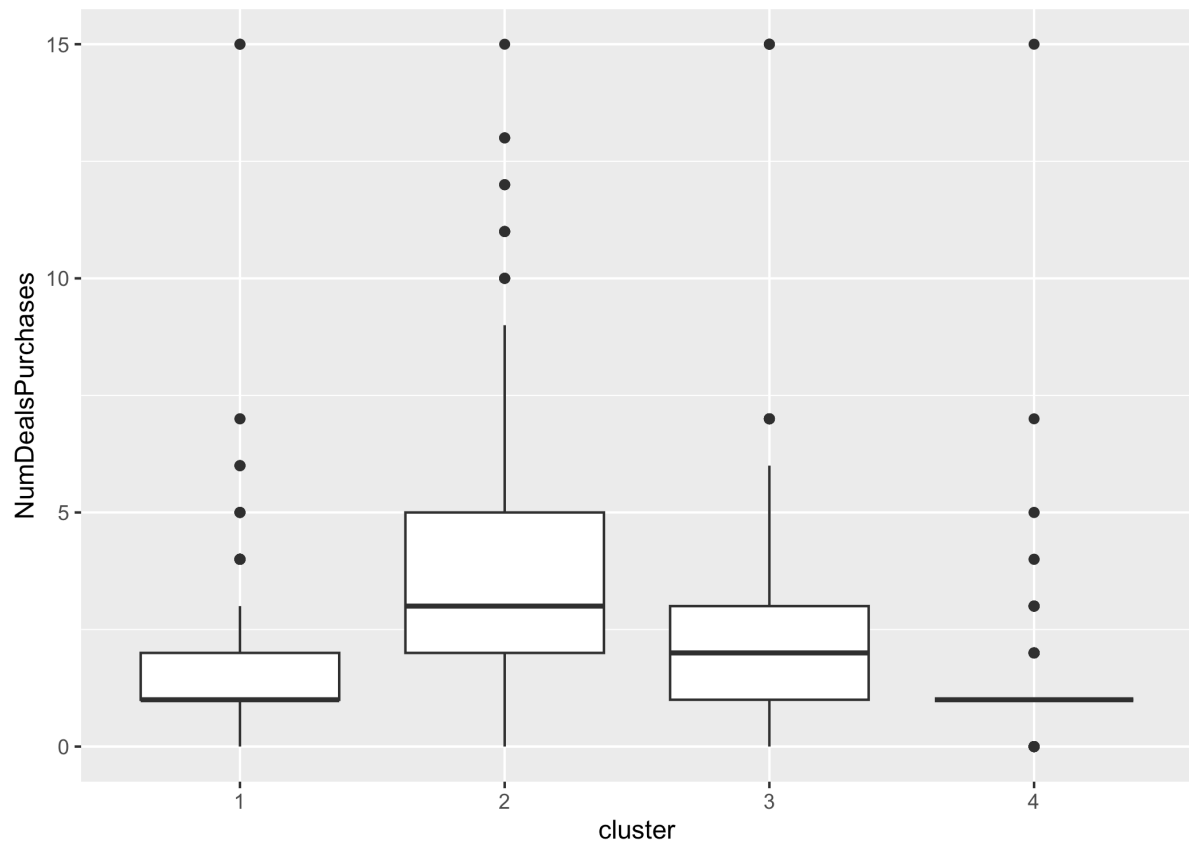
Our analysis indicates that low-income customers demonstrate the highest inclination towards purchasing discounted products, followed by the lowest income group. Conversely, the highest and middle income brackets show less interest in discounted deals. This suggests that discount promotions should primarily target low-income customers to optimize sales opportunities.
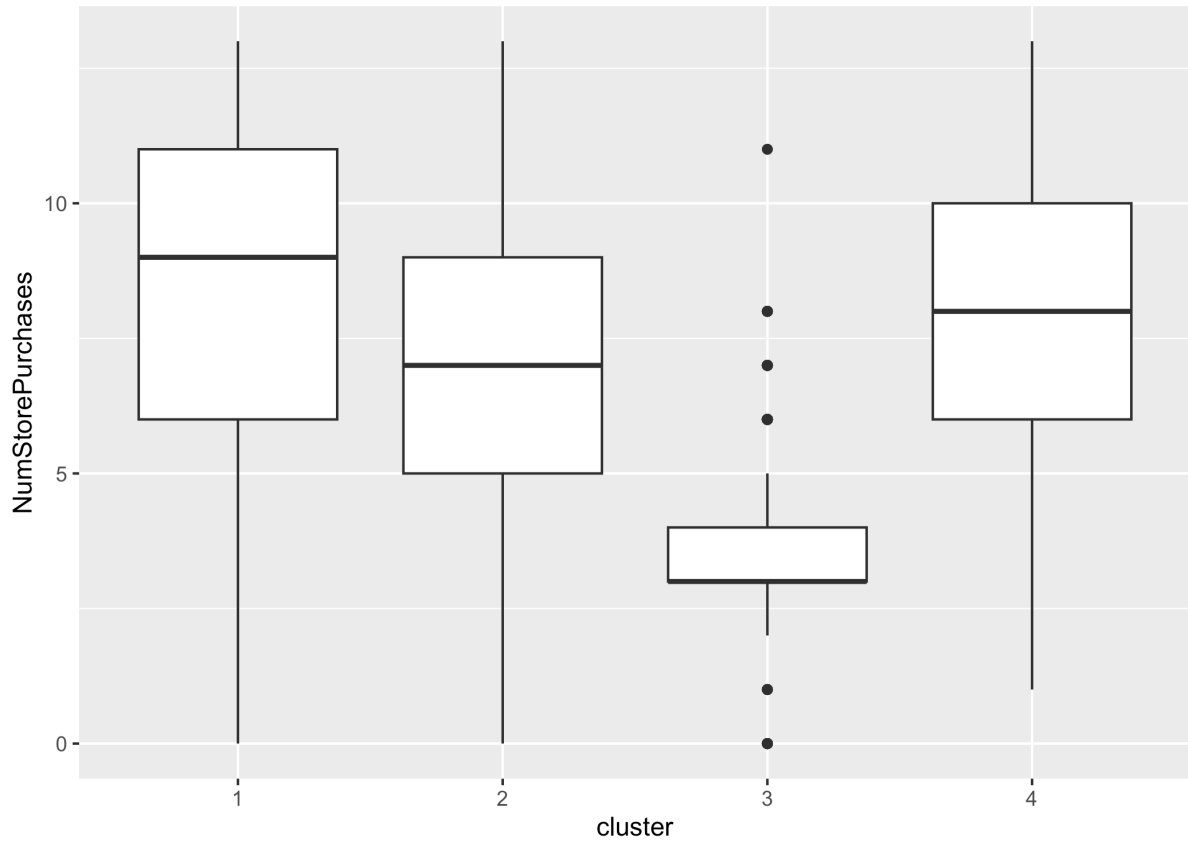
This analysis reveals significant disparities in store purchase behavior across different income clusters:

High-Income Cluster: Exhibits the highest frequency of store purchases, indicating a strong preference for in-store shopping among affluent customers.

Low and Lowest Income Customers: Display similar levels of store purchases, suggesting that these income groups also favor traditional brick-and-mortar retail experiences.

Middle-Income Customers: Show a comparatively lower preference for store purchases, indicating a potential inclination towards alternative shopping channels or consumption patterns.
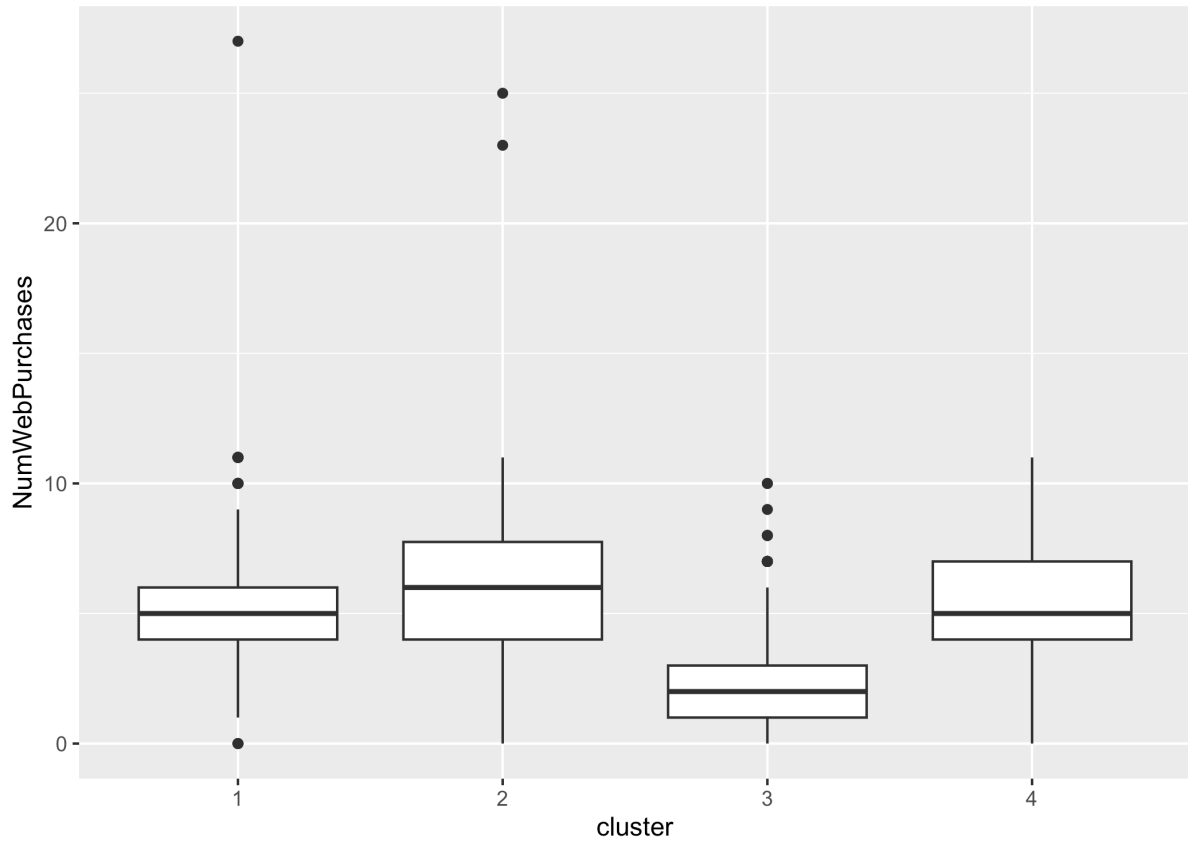
Lowest Income Cluster (Cluster 3): Displays the lowest frequency of web purchases, indicating limited engagement with online shopping platforms among individuals in this income bracket.

Low-Income Cluster: Shows the highest level of web purchases, suggesting a significant reliance on online channels for shopping among this demographic.

High and Middle-Income Clusters (Clusters 1 and 4): Exhibit similar levels of web purchases, indicating comparable engagement with online shopping platforms among affluent and middle-income customers.
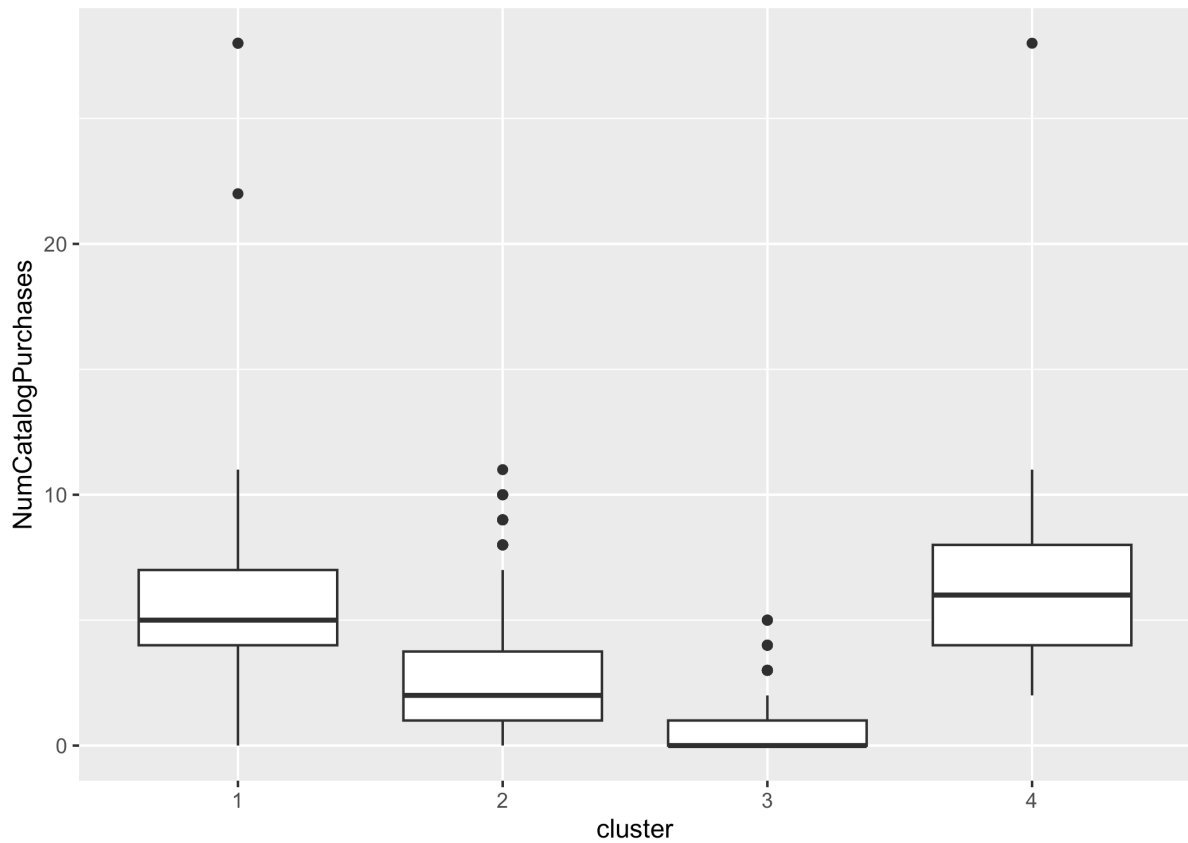
Catalog purchases are most prevalent among middle-income customers, indicating a strong preference for this shopping channel. Conversely, individuals in the lowest income bracket show minimal engagement with catalogs, followed by moderate usage among high and low-income clusters.

|	cluster	NumWebVisitsMonth
1	1	1411
2	2	3530
3	3	6253
4	4	587
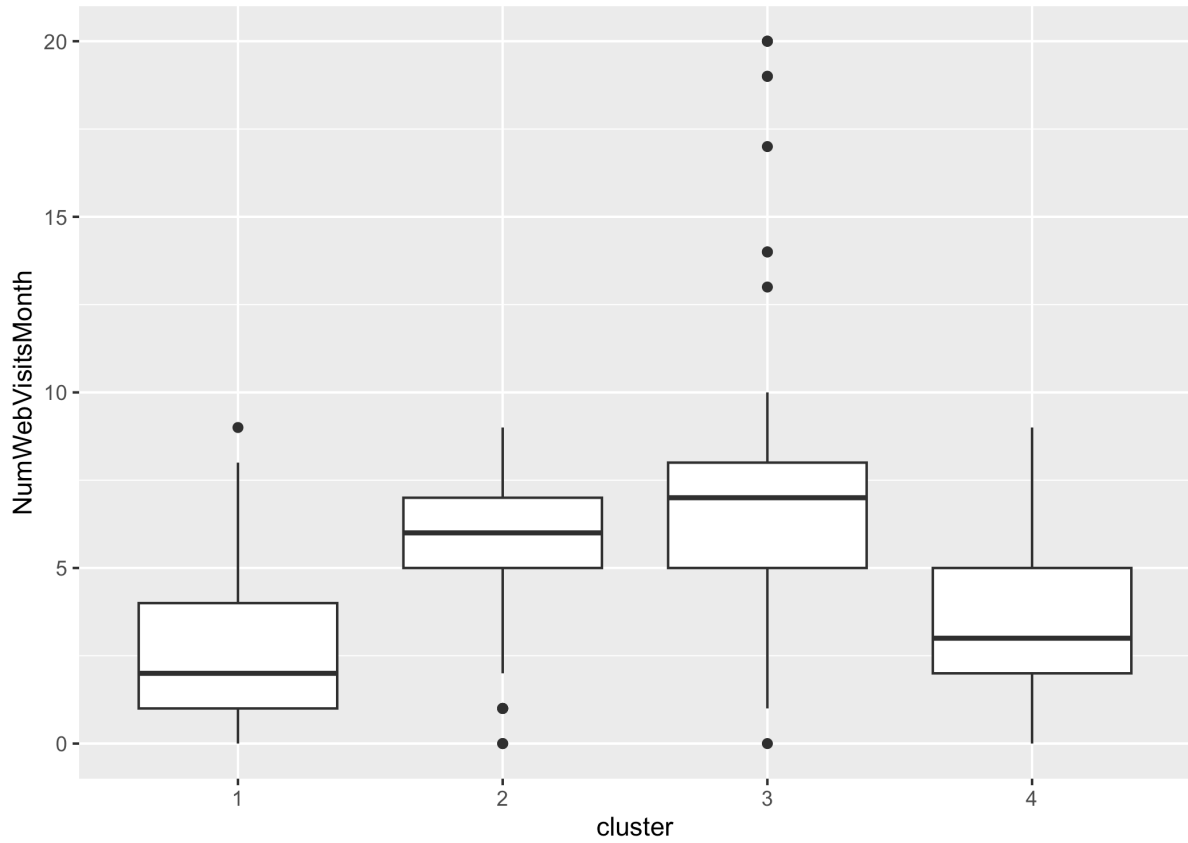
Web visits are most frequent among the lowest income group, but their conversion to purchases is the lowest. Conversely, high and middle-income individuals visit the web less frequently but exhibit higher purchase conversion rates. This suggests a more purposeful and efficient online shopping behavior among affluent and middle-income segments.

4.1.6 Plot 6: Recency Analysis
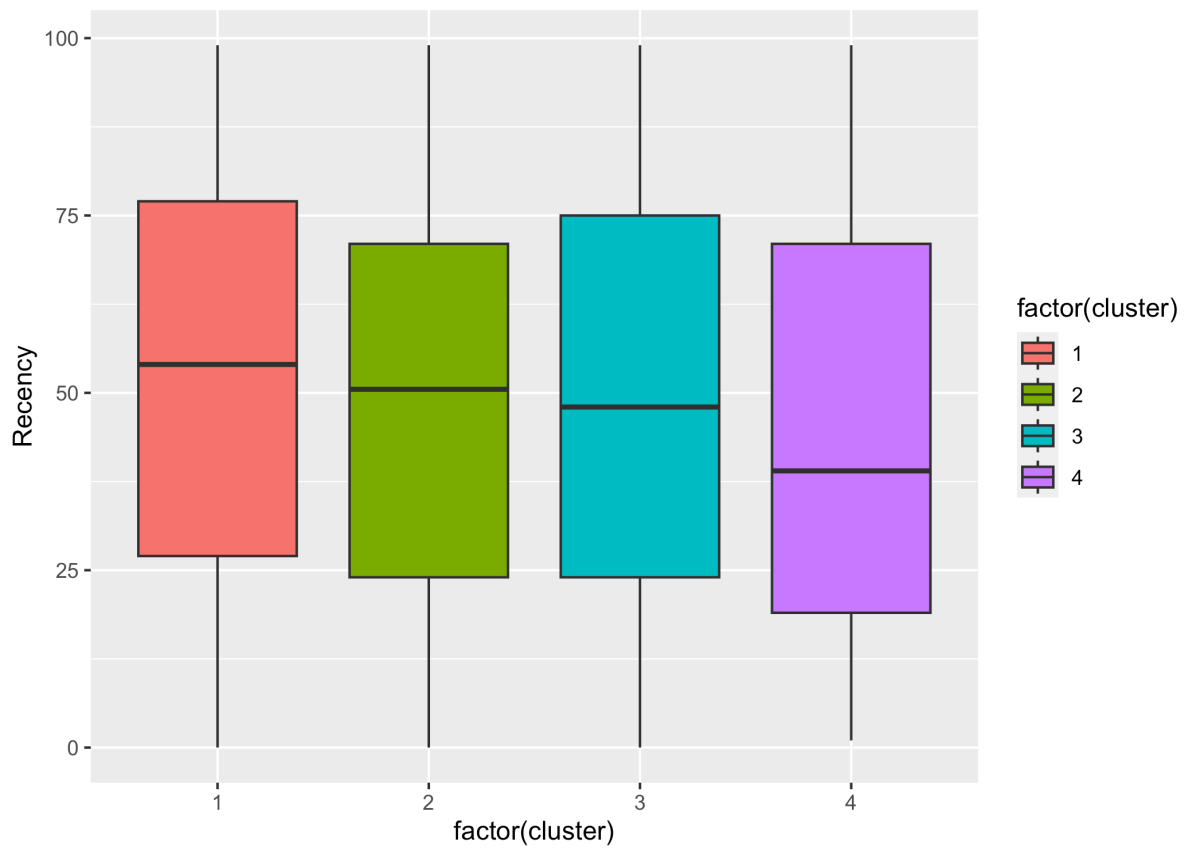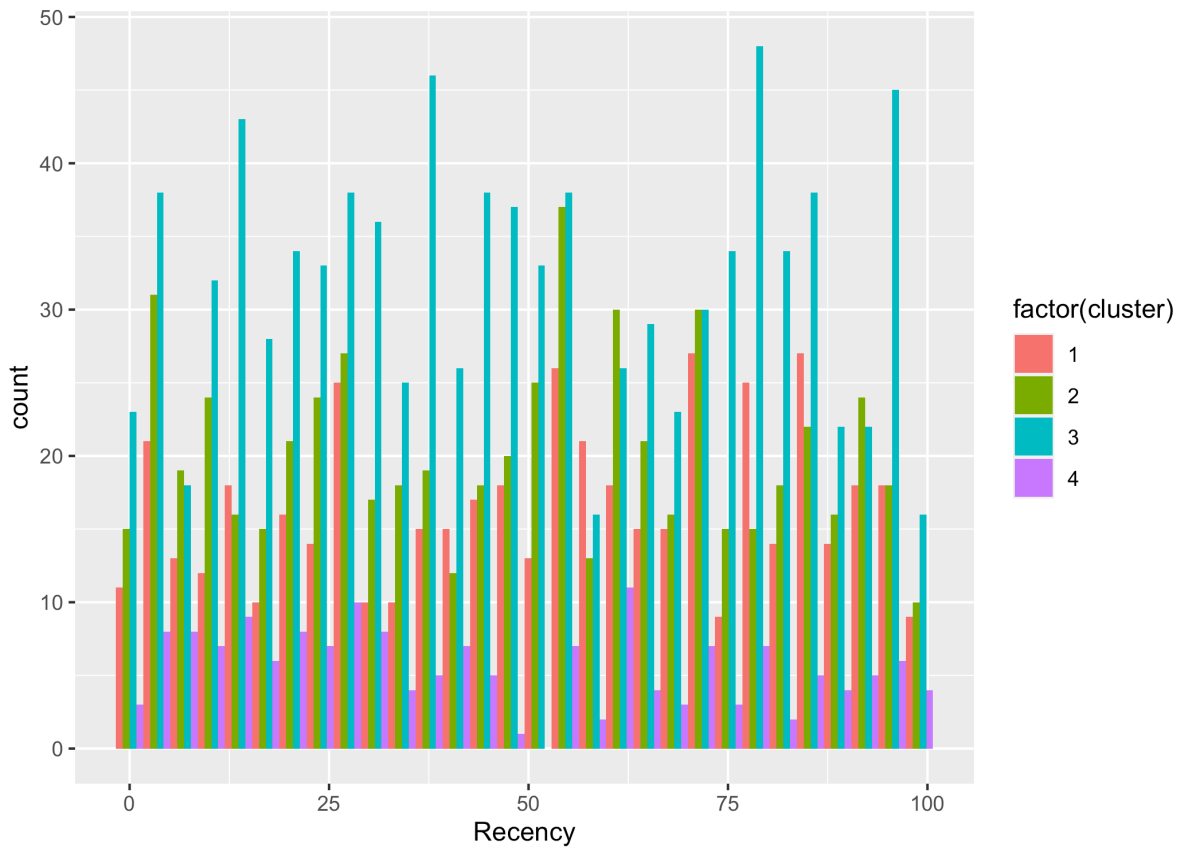
Recency analysis reveals the middle-income group with the least recent engagements, contrasting with the high-income group, which exhibits the highest recency. Interestingly, the low and lowest income clusters show recency levels akin to the high-income group, despite their lower income status. These findings emphasize the need for tailored engagement strategies across income segments to optimize customer retention and satisfaction.

4.1.7 Plot 7: Retention Analysis
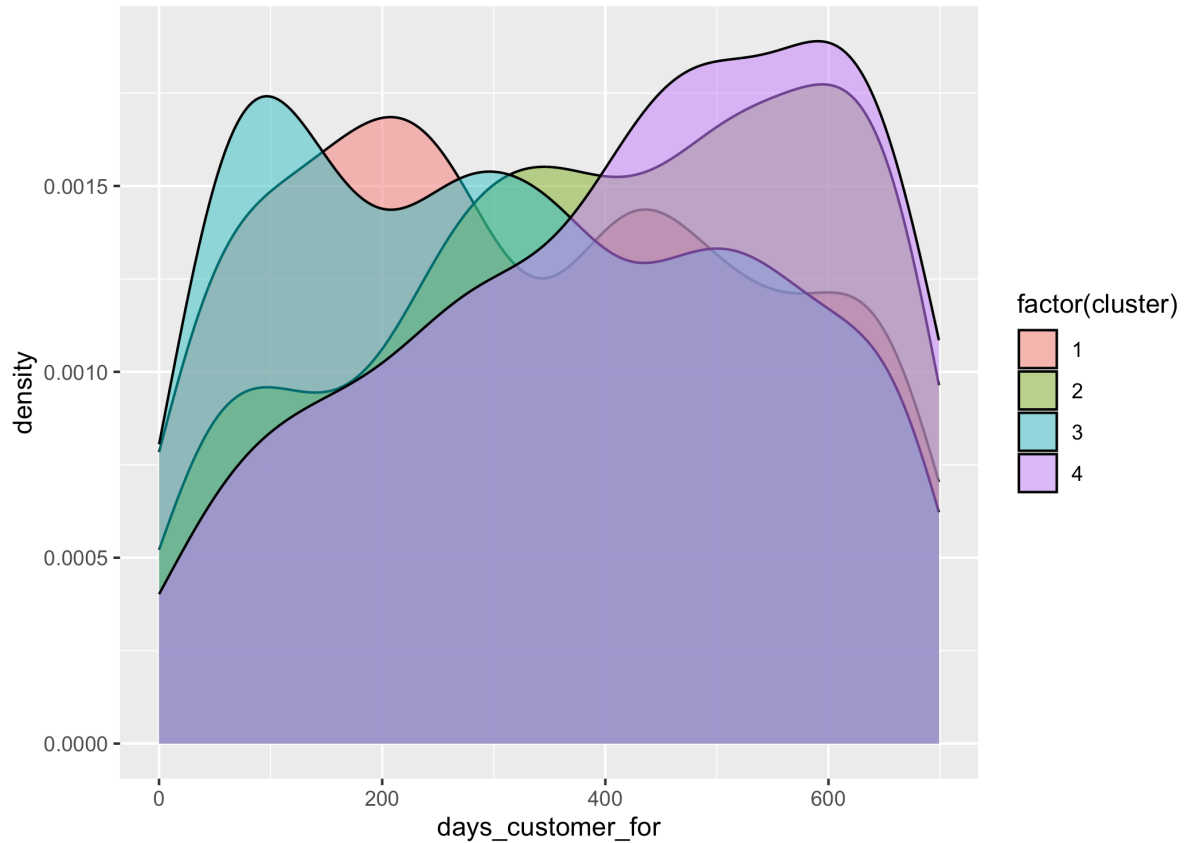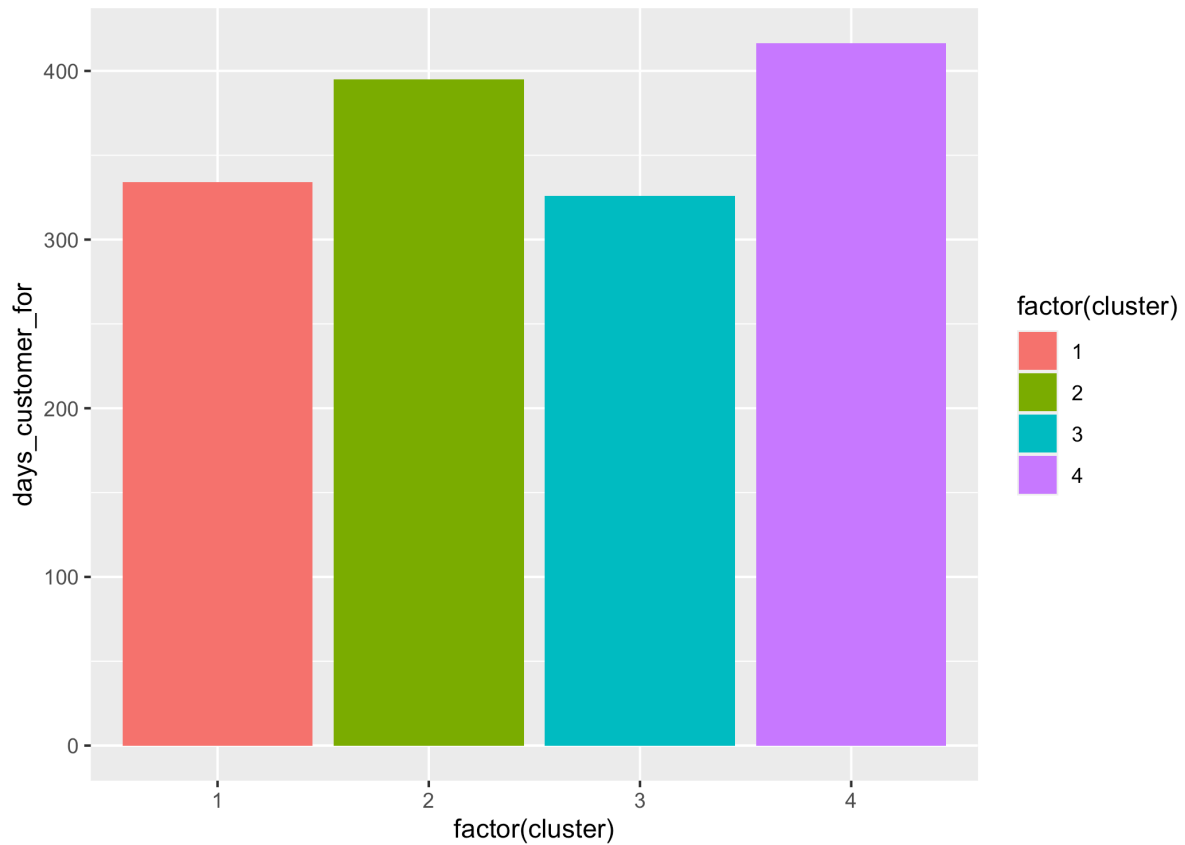
Customer retention rates vary across income groups, with middle-income individuals showing the highest retention, followed closely by low-income customers. Interestingly, both high and lowest income groups exhibit similar, lower retention rates, indicating a propensity for changing preferences among these demographics. Tailored strategies are crucial to enhance loyalty across diverse income segments and maximize long-term customer value.

4.2 Conclusion

After conducting data analysis and clustering using K-means, four distinct customer segments emerged with unique characteristics and spending behaviors:

Cluster 4 - High Affluent and High Spender:

Family Portrait: Typically childless or with only one child; diverse age distribution. Spending Behavior: Prefers in-store and catalog shopping over online; shows a preference for wine and meat products; less inclined towards discounted items and prefer traditional shopping channels.

Cluster 1 - Middle Income and Middle Spend:

Family Portrait: More likely to have children, especially teenagers. Spending Behavior: Displays a strong affinity for wine purchases and discounts; exhibits higher online spending compared to other clusters showing a balance between online and offline shopping preferences.

Cluster 2 - Low Income, Low Spend:

Family Portrait: Likely to have both young children and teenagers. Spending Behavior: Despite low income, allocates a significant portion of spending to wine; also shows interest in gold products.

Cluster 3 - Lowest Income, Low Spend:

Family Portrait: Skews towards families with more young children than teenagers; relatively younger demographic. Spending Behavior: Balanced spending across categories; displays a preference for gold products despite limited overall spending; frequent website visits with low purchase conversion rates. Our analysis illustrates the challenges faced by lower-income families, highlighting their limited spending capacity despite preferences for certain product categories.

Understanding these distinct segments allows businesses to tailor marketing strategies and product offerings to better meet the needs and preferences of each customer group, ultimately enhancing customer satisfaction and engagement.

5. Hypothesis Testing

5.1 Test 1

In our first testing let Null hypothesis be - Mean of the income of people who have opted for higher studies is same as to those who have not.

```
    Two Sample t-test

data:  graduation.data and higherstudies.data.income
t = -1.773, df = 1960, p-value = 0.07638
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4280.3840   215.7034
sample estimates:
mean of x mean of y
 52720.37  54752.71
```

Based on the p-value (0.0764) being slightly higher than 0.05, we *fail to reject the null hypothesis* at the 95% confidence level. This implies that the observed difference in mean income (-1.773) between the graduation and higher studies groups might be due to chance and not a true population difference.

5.2 Test 2

In our second test our **Null hypothesis is- The proportion of customers who accept the offer is the same for both graduation and Master's degree holders.** This essentially suggests that education level (graduation vs. Master's) has no influence on a customer's decision to accept the offer.

Alternative Hypothesis (H): The proportion of customers who accept the offer differs between graduation and Master's degree holders.

```
          Response
Education     0   1
  Graduation 964 152
  Master     309  56


   2-sample test for equality of proportions without continuity correction

data:  c(964, 309) out of c(1116, 365)
X-squared = 0.6759, df = 1, p-value = 0.411
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02487068  0.05931856
sample estimates:
   prop 1    prop 2
0.8637993 0.8465753
```

In this case, with a p-value of 0.411, we fail to reject the null hypothesis, indicating that the observed difference in offer acceptance proportions might be due to chance and not a true difference between the education groups.