SmartRetail: Customer Segmentation for Micro-Targeting

ST309 - Group Project Report

22529 [⅓] // 24369 [⅓] // 28538 [⅓]

08/02/2024

Introduction

Businesses gather data from various sources, and with the development of cloud architecture, business intelligence tools and statistical models, they aim to get actionable insights from the datasets. The retail industry is one of the active markets where the firms under it implement the aforementioned tools to optimise their operations. Amazon and Walmart coming at the top, the retail companies supply a huge variety of goods and services to cater to a large volume of customers all around the world. The corollary of it is that the customer datasets are high-dimensional and retail firms take necessary digitising actions to increase their marketing and sales operations. The challenge we would like to tackle revolves around the inadequacy of effective customer segmentation within the retail industry, hindering the optimisation of marketing strategies and personalisation efforts.

The retail industry is highly competitive, suppliers produce similar or nearly homogeneous goods in substantial volumes, resulting in market equilibrium price that sets a competitive outlook. To stand out as a business in this environment, retailers should go beyond the conventional approaches to appeal to customers and optimise customer lifetime value effectively.

With the price pressure from other competitors, market disruption from emerging e-commerce businesses, one effective strategy to be implemented by businesses is to provide a more personalised customer shopping experience, as well as yielding more price transparency for shoppers. In fact, a survey conducted by Epsilon and GBH Insights on 1,000 US adults found that 80% of respondents want personalization from retailers. (Lindecrantz et al, 2020)

With digitalisation and loyalty cards in retail businesses, firms have a large amount of data on customer activities which plays an integral role in the industry. The data possesses a high volume of customers followed by a high number of features. In this regard, data analytics techniques, particularly unsupervised clustering, can be employed to uncover the hidden patterns in customer behaviour, accurately segment similar customers into groups and potentially find similarities and correlations within each customer cluster. This allows businesses to make effective data-driven marketing decisions to target customers with personalised offers and incentives to meet their wants and preferences. This is pivotal for staying competitive and preserving competitive advantage in the retail business, ultimately yielding a data analytic problem.

We analyse a high-dimensional customer dataset to provide a framework to marketing departments by leveraging the power of analytics tools that will allow them to take strategic actions to increase long-term profitability. The goals of the study are listed below:

1) To segment customers using clustering algorithms based on various features in the dataset, including demographics and their spending behaviour.
2) To construct a differentiated segmentation strategy (Art, 2004) to develop a descriptive customer profile for the key segments provided by selected algorithmic learnings to understand their demographic information and their purchasing motivations, to create strong micro-targeting campaigns without wasting marketing budgets.
3) To identify the segment that will yield the retail business the most profits by observing customers' attributes.
4) To devise a classification algorithm model to refine the profile and personality description of each segment and understand what type of customers are likely to accept marketing campaigns.
5) To conduct an in-depth analysis using the association rules method of the resulting segments based on chosen questions concerning their behaviour across sales channels and purchase of deals.

Customer segmentation and personality analysis have been established practices within firms since the 1920s. (Tedlow, 1990) The earliest attempt at customer segmentation using data analysis focused on demographic segmentation. Before the digitalisation era, the data was primarily collected from surveys and interviews, and businesses grouped their customers based on factors such as age, sex, education, income and geography. This allowed firms to send out marketing messages more tailored towards a specific demographic group rather than a mass marketing campaign. However, this type of segmentation analysis did not take into account the variations in the shopping behaviours and needs within the broadly defined demographic group. (Mehta, 2023)

The clustering algorithms used in our analysis are mainly the K-means and Hierarchical clustering. These algorithms were developed in the 1950s and 60s respectively. With the advances in data collection methods and analytics tools, businesses use those tools to collect and analyse more behavioural data from shopping platforms and other sales channels. In fact, with the usage of modern tools, neuro-marketing plays a significant role along with analytical tools to enhance customer segment analysis.

Data Description

Our data is sourced from the Kaggle "Marketing_Campaign" data file (Kaggle, 2022), which contains 2,240 ideal customers and 29 columns, including all attributes listed in the features table. The file contains a detailed analysis of the retail customers to explore their personality based on their characteristics, which past campaigns are accepted and the amount spent on products, on each sales channel.

The sample size is large, covering a two-year catalogue of customer responses to each attribute. This will be sufficient to observe consumers' purchasing patterns and tendencies to respond to campaigns and deals. Therefore, this enables us to conduct a microeconomic analysis to observe the behaviours of the economic agents to make effective marketing decisions.

The attributes, included in the data file, can be categorised into the following subsets as shown in the table below.

Demographics	**Spending by Categories**	**Sales Channels**	**Marketing Campaigns Performance**
Year_Birth: Customer's Birth Year	*MntWines:* Amount spent on wine in last 2 years	*NumWebPurchases:* Number of purchases made through the company's website	*NumDealsPurchases:* Number of purchases made with a discount
Education: Customer's education level	*MntFruits:* Amount spent on fruits in last 2 years	*NumCatalogPurchases:* Number of purchases made using a catalogue	*AcceptedCmp1:* 1 if customer accepted the offer in the 1st campaign, 0 otherwise
Marital_Status: Customer's marital status	*MntMeatProducts:* Amount spent on meat in last 2 years	*NumStorePurchases:* Number of purchases made directly in stores	*AcceptedCmp2:* 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
Income:	*MntMeatProducts:*	*NumWebVisitsMonth:*	*AcceptedCmp3:*

Customer's yearly household income	Amount spent on meat in last 2 years	Number of visits to company’s website in the last month	1 if customer accepted the offer in the 3rd campaign, 0 otherwise
Kidhome: Number of children in customer's household	*MntFishProducts:* Amount spent on fish in last 2 years	x	*AcceptedCmp4:* 1 if customer accepted the offer in the 4th campaign, 0 otherwise
Teenhome: Number of teenagers in customer's household	*MntSweetProducts:* Amount spent on sweets in last 2 years	x	*AcceptedCmp5:* 1 if customer accepted the offer in the 5th campaign, 0 otherwise
Dt_Customer: Date of customer's enrollment with the company	*MntGoldProds:* Amount spent on gold in last 2 years	x	*Response:* 1 if customer accepted the offer in the last campaign, 0 otherwise
Complain: if the customer complained in the last 2 years	x	x	x

Data Cleaning and Feature Engineering

The initial Customer data included all the customer information including the amount spent on each selected item (Meat, Fish, Wine, Fruits, Sweets, and Gold), the amount spent on each sales channel (Web, Catalogue, and Store) and the number of web visits per month, along with the binary response to promotions.

Firstly, we wanted to check whether there were any duplicate values in our data. Each customer is characterised by their unique ID, and we observed no duplicate values in the ‘ID’ variable.

Furthermore, checking the missing values in the dataset, we observed 24 missing values all belonging to the *Income*. As can be seen in the Figure 1, the missing values only correspond to 1.07% of our entries thus we omit the 24 missing observations as this will not affect our dataset as it is still significant and valuable to observe.

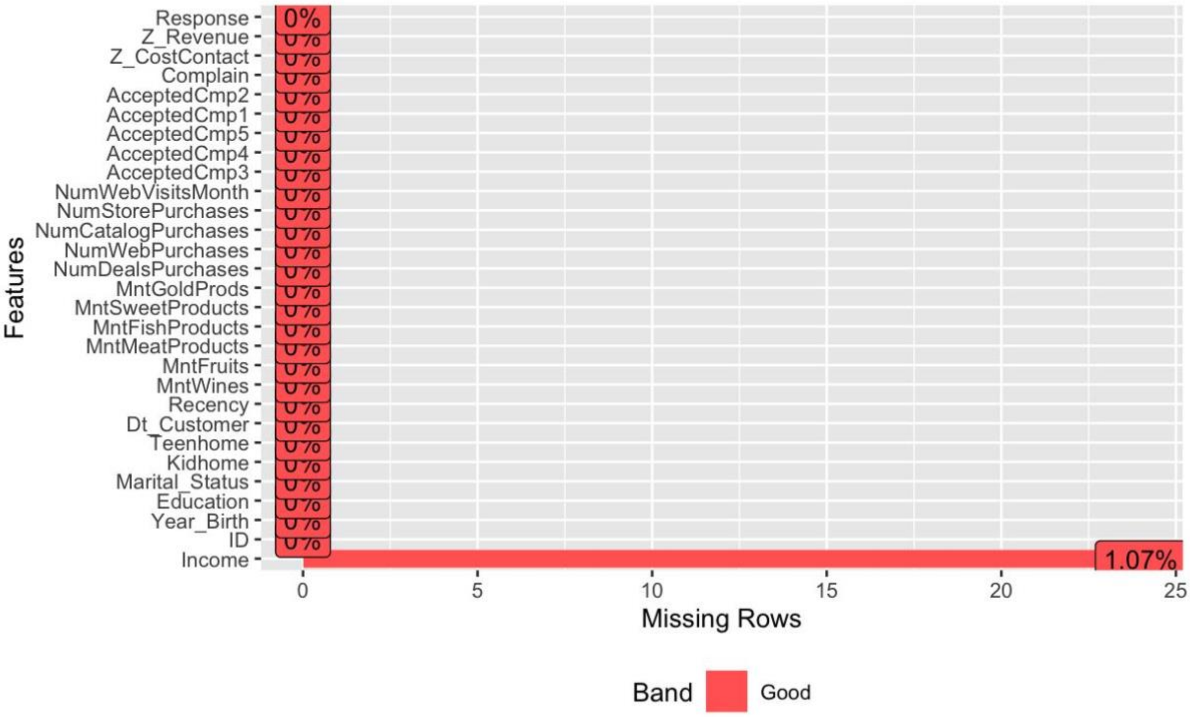


Figure 1: Missing Values

Additionally, we observe some potential outliers from our summary statistics, and we further examine them using the histogram of *Age* (engineered by 2021-Customers$Year_Birth) and *Income* for all the customers. Hence, we filter out the individuals aged above 80 and have an income level above 100,000 as outliers.

We observed that the minimum and maximum ages are 25 and 78, respectively. Therefore, the customers were grouped into three age categories for the "*AgeCategory*" variable, which are "<40", "41-56", and ">57", supporting the generational types from Millennials to Baby Boomers + Silent Generation in this The Standard article (Mata et al, 2024).

The year customers joined were calculated by date formatting the *Dt_Customer* variables in terms of year and measuring them against the year 2021. The results were then assigned to the column variable "*YearsJoined*".

We created new variables, including the *TotalAcceptedCmp*, *NumChildren* and *Spending*. Different products, such as Meat, Fish and Gold, are sold in the retail store. These variables were created because they capture similar results with the original variables without combining the total observations, and are highly correlated to other features.

Therefore, we decided to investigate the total amount of goods spent by summing up the amount spent for each product by a customer and assign it to a variable called "*Spending*". Then, the total accepted campaigns ("*TotalAcceptedCmp*" variable) were summed using the five accepted campaigns and the response. We decided to include the response in the summation of the total campaigns, assuming that it is a response to one of the deals. Lastly, we combined the number of children, including "*KidHome*" and "*TeenHome*", to produce the total number of children in the household for each observation.

The skewness of our data could adversely affect the performance of the supervised learning models. In that regard, the total spending and the amount spent on each product variable were log-transformed using log(1+**Customer_cleaned**$*variable*), to mitigate the impact of the skewness in our data. It is also favourable to implement log-transformation as it does not affect the unsupervised learning algorithms.

Afterwards, we created a histogram to observe the skewness of all the variables in Figure 2. It was observed that the amount spent on each product before the log transformation was highly skewed to the right. According to Feng et al, log transformation can be applied to improve the validity of the associated statistical analysis when the distribution of continuous variables in the data is considered non-normal, like *Spending*. If the original data follows a log-normal distribution or approximately so, then the log-transformed data follows a normal or near-normal distribution. In this case, the log transformation removes or reduces skewness. Compared to the amount spent on each product after log transformation, the distributions were as close to being normally distributed. For example, it was observed that the amount spent on Gold products was highly positively skewed before the log transformation and then, became close to being a normal distribution after the transformation.

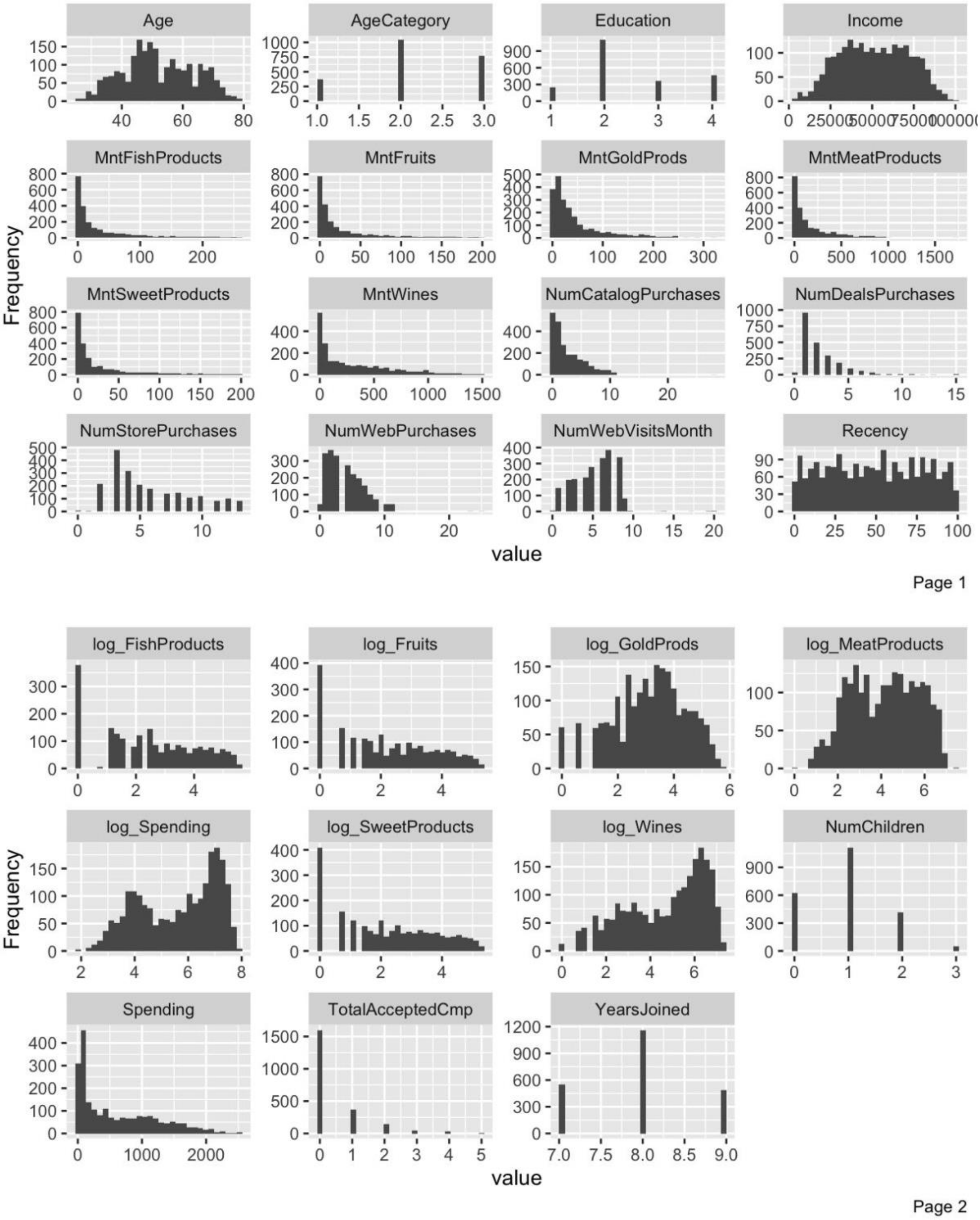


Figure 2: Histograms of Variables

The limitation of log transformation arises as shown in the histograms of the logarithm of the amount spent on each variable as there appear to be zero variables, leading to a potential misrepresentation of the actual quantity of goods purchased, particularly when some customers spend a uniform amount for a good. West (2021) noted that logarithmic transformation is

appropriate for selected variables in the data frame that have no zero values. Alternatively, if zero values are present, it is recommended to use square root transformation.
For the implementation of the classification algorithm and association rule analysis, the *Marital Status* were transformed into the *Relationship* variable. We classified the variable into nominal binary categories: "Not Partnered" and "Partnered". The "Not Partnered" includes all the other categories except for those that are categorised under "Married" and "Together". Furthermore, the *Education* variable was divided into four ordinal categorical variables: **"**Bachelor**"**, Graduate", "Master", and "PhD". Onwards, we assigned the categories in Relationship and Education variables into factor variables.

After the data transformation and creation of new variables in the **Customers_cleaned**, redundant columns were removed to avoid duplicates and unnecessary results in the cleaned dataset, which includes *ID*, *Z_CostContract*, *Z_Revenue*, *Year_Birth*, *Maritial_Status*, *Dt_Customer, Teenhome, KidHome.*

In the pre-processing stage, the label-encoding method was used to transform all the categorical variables into numerical labels based on the alphabetical order of the categories in the **Customers_cleaned**. This enables the statistical learning algorithms to work effectively through the processing of data. For example, the *Education* variables were encoded accordingly into 1 for "Bachelors", 2 for "Graduate", 3 for "Masters", and 4 for "PhD".

Data Analysis

Principal Component Analysis (PCA)

In our preliminary exploratory data analysis, we plotted the correlation matrix using the **Customers_cleaned** data to analyse the relationship between variables. From the plot, we have observed the following.

Firstly, we found that the spending variables are highly correlated with each other. Secondly,the income is highly correlated with our spending variables. Moreover, the number of web visits is highly negatively correlated with many variables in our dataset. Therefore, we decided to use Principal Components Analysis (PCA) to capture the co-movements in correlated variables using only a few factors to denoise the data. (Hastie et al, 2013)

Before the PCA is applied, the redundant variables that are captured in the other variables were removed which we had for our **Customers_cleaned**. This includes *Spending, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Complain, Response*, *AgeCategory*.

If PCA is implemented using unscaled variables in the Customers_cleaned data frame, the variables with higher variance, such as *Income*, will have higher factor loadings, and this adds more weight to the variable. Then we standardised the variables to have a mean of 0 and a variance of 1 as they have different means and variances.

In order to perform PCA, the eigenvalues need to be obtained using the **Customers_cleaned**, which is crucial in observing the largest eigenvalue and determining the number of components to explain a desired cumulative proportion of variance. The PCA graph in Figure 3 showcases that the first component explains 40% of the variance. Then, the next two components further explain 18% of the variance. Twenty components in total capture all of the variances in the **Customers_PCA** dataset. We decided to use 3 components to explain 58% of the total variability in the data frame.

Additionally, with the use of principal components, we are able to produce a sophisticated visualisation of our clusters and use it for our clustering analysis by preventing the overlapping of the observations within each cluster compared to using the raw data for clustering analysis.

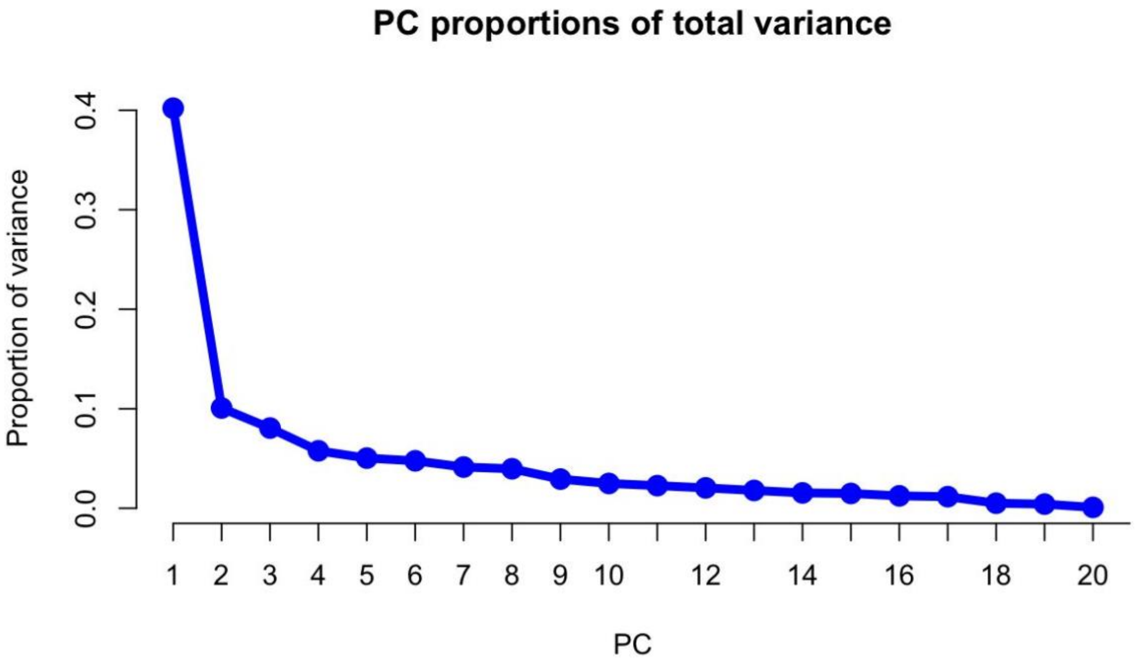


Figure 3: PCA Variance

Clustering

The goal of our study is to perform market segmentation. Therefore, clustering algorithms aid us to partition customers into clusters where individuals within each group are quite similar to each other while individuals across different groups are different.

For our clustering analysis, we used the 3 principal components that capture most of our customer features.

The first clustering algorithm we used was the hierarchical clustering. The hierarchical clustering algorithm uses dissimilarity measures and follows an iterative approach to produce a tree-like structure called dendrogram. Firstly, each observation is assigned to a different cluster. Then we group two individuals with the smallest distance together, defined by our dissimilarity measure. The process of fusion from bottom to top is iterated until all observations are put into a single cluster.

The default distance measure between objects is Euclidean distance, as the data is almost evenly distributed. Then we decided to use the complete linkage method as it tends to produce more compact clusters, which enables us to group similar and clearly distinct customer segments.

In terms of deciding the optimal number of clusters, we used the Elbow and Silhouette method as a reference. The elbow joint of the graph showed that k=4 is the optimal cluster after plotting the distance against the k values and the calculated sum of the squared distance (inertia). Meanwhile, the silhouette method showed that k=2 is the optimal cluster after estimating the sum of the squared distance from each point to the assigned centre for each cluster value. (Kassambara, 2017) We calculated the average of the optimal clusters using both methods, which resulted in k=3. From there, the customers were then split into 3 different clusters.

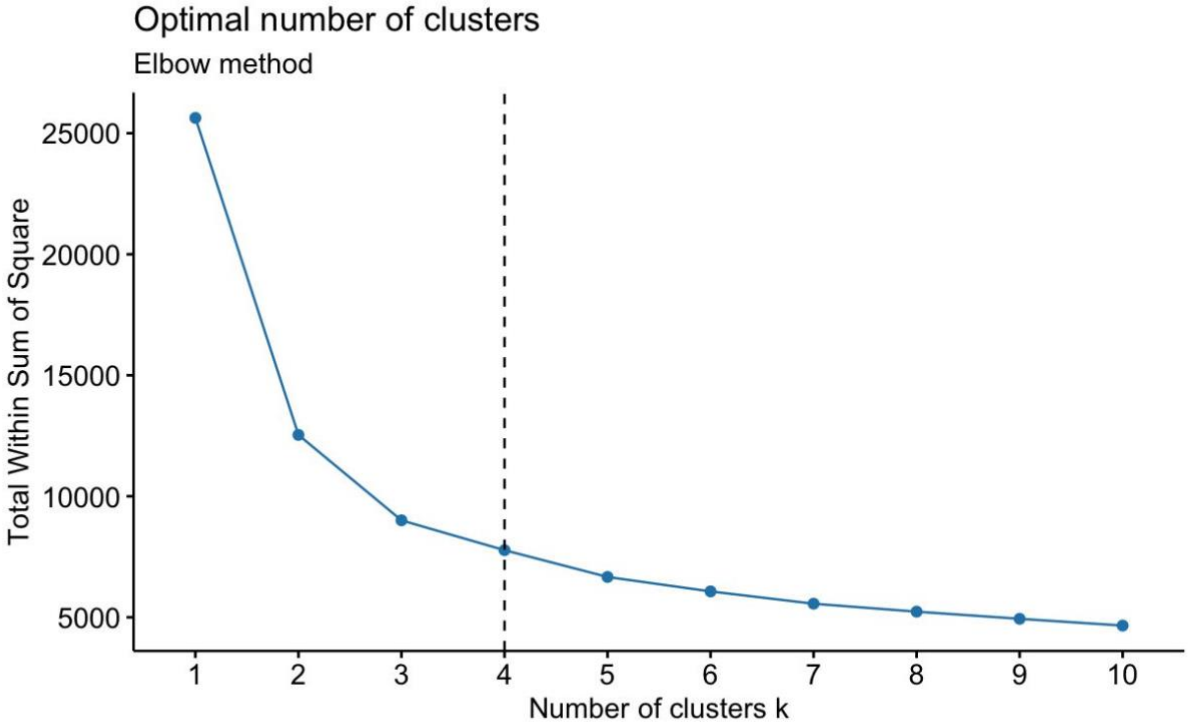


Figure 4: Elbow method for optimal cluster number

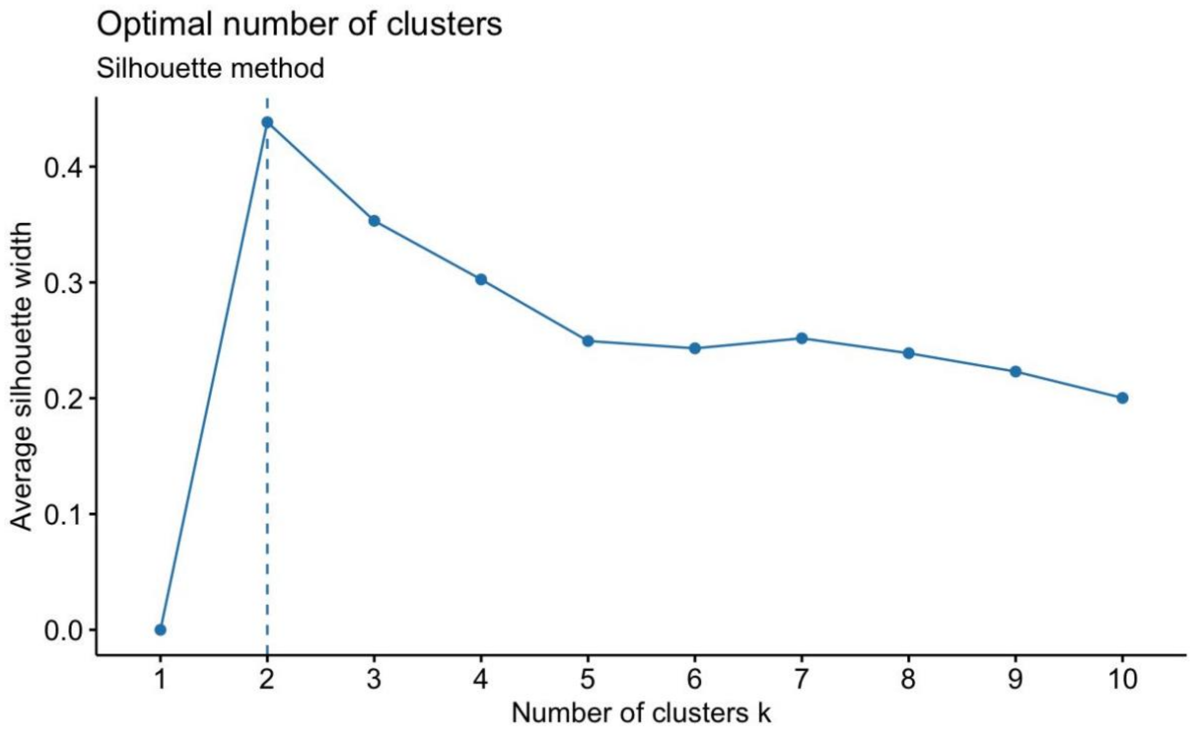


Figure 5: Silhouette method for optimal cluster number

Based on the above method and the plotted dendrogram from hierarchical clustering, we cut the dendrogram line into 3 groups. It is noted from the dendrogram that the higher the height of the plot, the less similar the objects are.

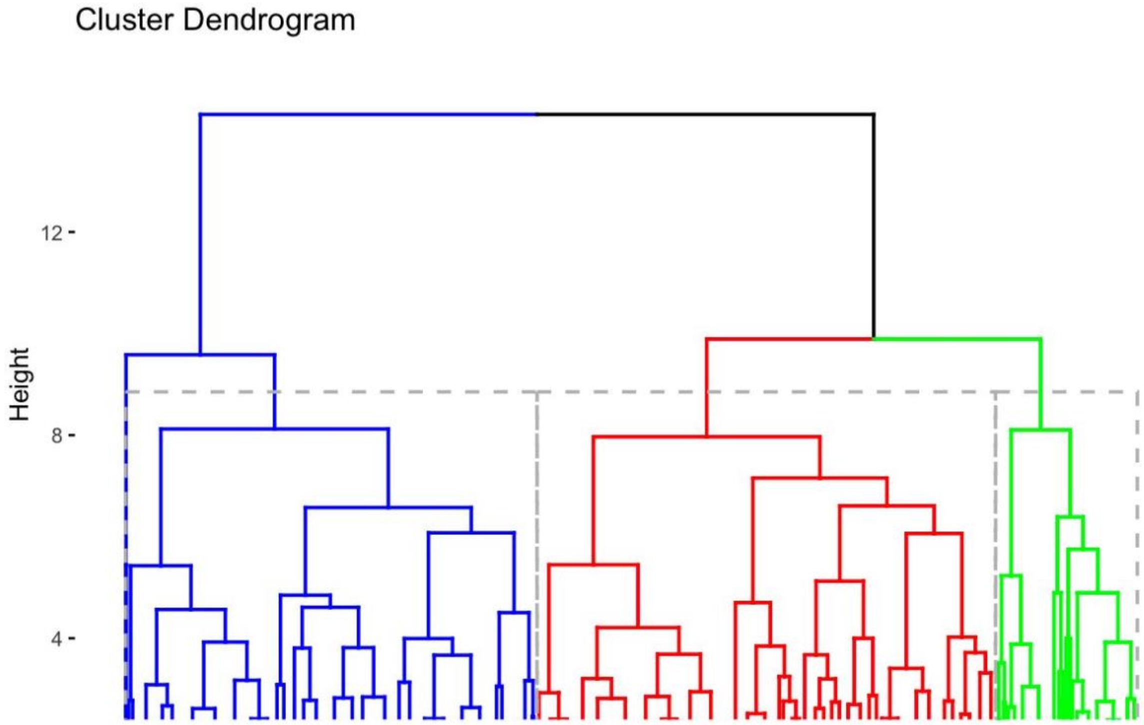


Figure 6: Dendrogram plot with K = 3

A three-dimensional scatter plot, rather than a two-dimensional plot, is plotted to observe the hierarchical clusters against the three principal components. It is evidently shown that PCA is an effective use to visualise the cluster groups as they don’t overlap with each other.

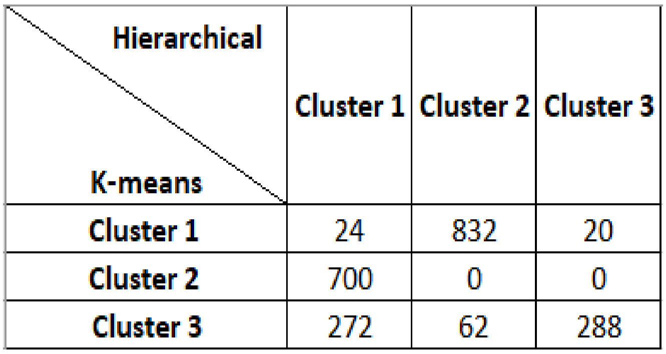
Then we were able to produce the hierarchical clustered groups by using the hclust() method. We observed that each clustered group has a more balanced and large number of individuals ranging from 308 to 996 customers, and this will be beneficial for effective further data analysis.

Cluster 1	**Cluster 2**	**Cluster 3**
996	894	308

Table 1: Hierarchical clustering result with complete linkage

Then we do k-means clustering with the same number of clusters as our pre-defined cluster size. The k-means algorithm divides the observation into clusters, minimising the within-cluster variation. This makes the observations under each cluster as homogeneous as possible. The algorithm achieves this iteratively, starting from a random state until it converges. From this aspect, the k-means algorithm is a stochastic algorithm.

To compare the results from both clustering methods in a table format. (Hastie et al, 2013) Based on the table, we can observe that both methods yield similar clustering results. For example, a large proportion of observations in cluster 1 in K-means clustering are assigned to cluster 2 by hierarchical clustering. Likewise, a high percentage of observations in K-means cluster 2 are contained in hierarchical cluster 1. Cluster 3 looks somewhat different for both methods, but most of the observations in hierarchical cluster 3 are in the K-means cluster 3.



Hierarchical K-means	Cluster 1	Cluster 2	Cluster 3
Cluster 1	24	832	20
Cluster 2	700	0	0
Cluster 3	272	62	288

Table 2: Comparing Hierarchical and K-Means Clustering Results

Therefore, we decide to only use the hierarchical clustering result for further analysis involving supervised learning methods.

	Cluster 1	Cluster 2	Cluster 3
Number of Customers	Most customers against other clusters.		Least number of customers against other clusters.
Education	Highest number of customers who earned Graduate degrees. Similar proportion with Cluster 3 when Bachelor's degree or PhD is earned.	Highest number of those who earned Bachelor's degree.	Similar proportion with Cluster 1 when Bachelor's degree or PhD is earned. Higher proportion of customers with Masters degree.
Mean Customer Household Income	Highest mean household income is estimated around 68,000 in income.	Lowest mean household income is estimated around 34,000 in income.	Mean household income is estimated around 49,500 in income.
Income vs Total Spending	Highest household income and total amount spent on all goods against other clusters.	Lowest household income and total amount spent on all goods against other clusters.	
Amount of Goods Spent	Spends more on Meat	Spends more on Fruits, Fish, Sweets, and Gold	Spends more on Wine
Average Total of Promotion Accepted	Highest average total number of accepting campaigns.		Lowest average total number of accepting campaigns.
Average Number of Deals Accepted	Lowest average number of deals accepted against other clusters.		Highest average number of deals accepted against other clusters.
Average Counts of Purchases in each Channel	Highest average number of store purchases against other clusters. Least	Lowest average number of purchases across all sales channels. Least favoured	Highest average counts of web purchases. Least favoured sales channel for purchase is catalogue.

	favoured sales channel for purchase is catalogue.	sales channel for purchase is catalogue.	
Number of Children, on average	Lowest number of children, on average		High number of children, on average

Table 3: Clustering Results Summary

Classification to generate cluster description

Upon the segmentation of the customer groups, clusters were labelled and put into supervised learning algorithms to generate better cluster descriptions, adding more interpretability to the clustering output. The labels were named *InCluster1*, *InCluster2* and *InCluster3*. For example, *InCluster1* variable is equal to 1 if an observation is in hierarchical cluster 1 and 0 otherwise. Then they were fitted into the classification tree to discover the intragroup commonalities within each cluster. Thus, here, we made three separate classification problems with each classifying one cluster from the others.

Having fit the classification trees for all our labelled clusters, we observed that the interpretability of the tree for *InCluster1* was not ideal. This issue was addressed through tree pruning. The cross-validation method was initially applied to determine the optimal size of the tree and it was then pruned accordingly. The cross-validation suggested nine leaves, corresponding to the lowest value of Cross Validation Value. However, considering the marginal impact of Cross Validation Value, we have determined the optimal number of leaf nodes as six, providing a further interpretability at the cost of a minimal incremental misclassification error.

The classification tree for Cluster 1 depicts that there is a relatively lower income group with higher spending on average, a lower number of accepted deals and a lower spending amount for fish compared to other cluster groups. Those who are high-income customers under Cluster 1, purchase a lower amount of deals compared to other cluster groups. The underlying conclusion from this observation is that the richer quantile under Cluster 1, is less responsive to deals.

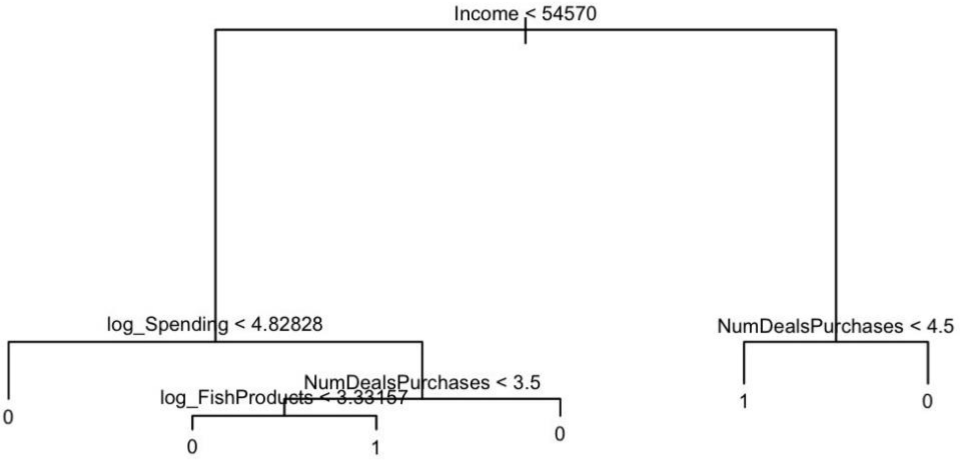


Figure 7: Classification tree for cluster 1

The customers under the Cluster 2 have lower spending tendencies in general. They have a strict budget and need to control their expenditures.

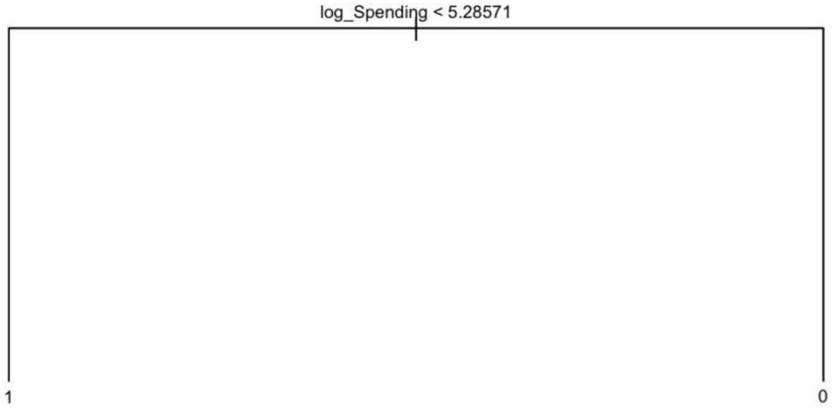


Figure 8: Classification tree for cluster 2

In the Cluster 3 decision tree, it is evident that consumers who spend more on deals exhibit a high frequency of web visits or consumers who spend less on deals are more likely to be wine spenders.

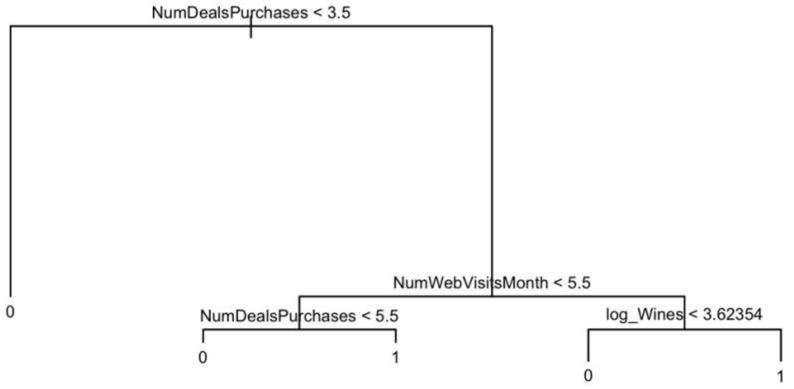


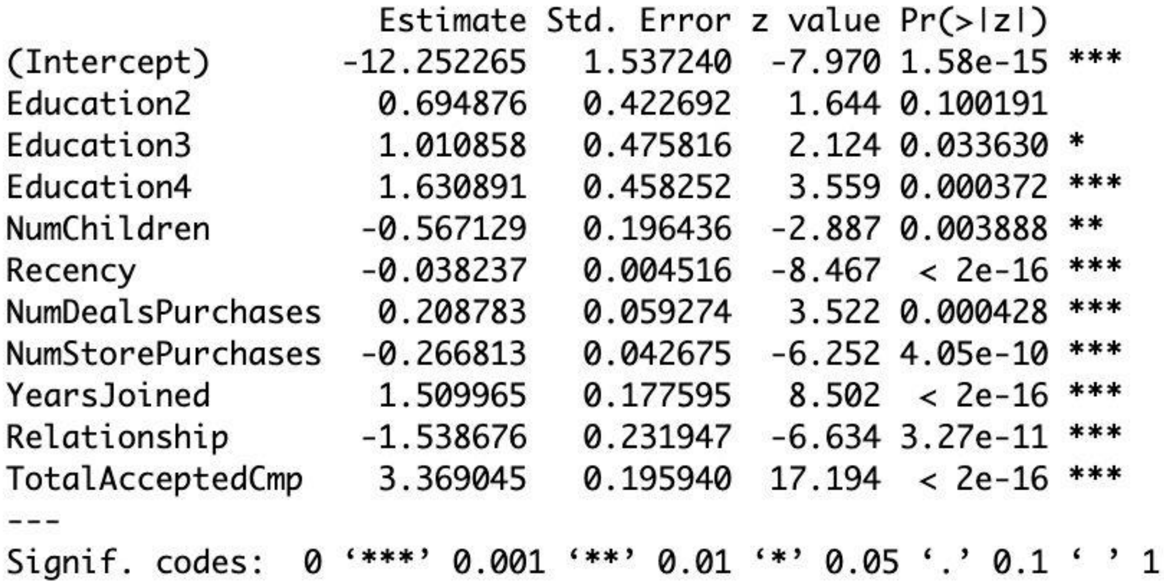
Figure 9: Classification tree for cluster 3

Logistic regression

Logistic regression is another supervised learning model used to analyse the marginal effect of independent variables on a binary dependent variable, such as, whether a customer will respond or not to a promotional advertisement.

For our predictive modelling, we constructed a logistic regression model to predict the *Response*, where 1 indicates the customer accepted an offer in the previous campaign, and 0 otherwise. We used features including *Education, Income, NumChildren, Recency, NumDealsPurchases, NumWebPurchases, NumStorePurchases, YearsJoined, Relationship, Age, Spending, and TotalAcceptedCmp*. In summary of the logistic regression model, *Age, Income, and Spending* variables have insignificant p-values (i.e. 0.922872, 0.958093, 0.232139, respectively) and therefore, we removed them in the model as we iterated the process of finding the insignificant variables. The refined logistic regression model predicts the response variable with an accuracy of 94.3% on all data.

The number of observations in the **Customers_cleaned** where the *Response* is equal to 1 is highly unbalanced compared to the number of observations where the *Response* is equal to 0. To address this imbalance, we subset the data by selecting 329 observations with *Response* == 1. Then, we randomise a sample from the portion of the data where *Response* == 0, ensuring that it matches the size of the subset data (329 rows). They are then merged together into a new data "**train_ind**", containing a total of 658 observations. The combined data is apportioned into training and testing datasets, with a 75%-25% split. We used the same set of selected independent variables, excluding *Age*, *Income,* and *Spending*. Then we made a binary prediction based on a probabilistic threshold of 0.5. When calculating the accuracy score, we observed that the logistic regression model predicts the response variable 93.7% on testing data. Therefore, there is no significant evidence that there is overfitting in the test and training.



```
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -12.252265   1.537240  -7.970 1.58e-15 ***
Education2         0.694876   0.422692   1.644 0.100191    
Education3         1.010858   0.475816   2.124 0.033630 *  
Education4         1.630891   0.458252   3.559 0.000372 ***
NumChildren       -0.567129   0.196436  -2.887 0.003888 ** 
Recency           -0.038237   0.004516  -8.467  < 2e-16 ***
NumDealsPurchases  0.208783   0.059274   3.522 0.000428 ***
NumStorePurchases -0.266813   0.042675  -6.252 4.05e-10 ***
YearsJoined        1.509965   0.177595   8.502  < 2e-16 ***
Relationship      -1.538676   0.231947  -6.634 3.27e-11 ***
TotalAcceptedCmp   3.369045   0.195940  17.194  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Logistic Regression Result

The main aim of this model methodology is to capture whether there exists sufficient evidence that the selected features influence customers' response to a promoting campaign in the **Customers_cleaned** with the confidence level set up to a 99%. The variables are as follows:*TotalAcceptedCampaign, YearsJoined, NumStorePurchases*, *Recency, Relationship, NumChildren*. For example, a longer period that the customer has been with us positively contributes to the probability of campaigns accepted. Also, lower number of store purchases seem to be an indication of higher probability of accepting campaigns.

Association Rules

Association rules (MacKenzie et al., 2013) is a popular method used in data mining to discover undiscovered relationships between variables in a large dataset. In our context, we are interested in using the association rules to answer the question, “Which type of customers are more active under each specific shopping channel?” By understanding the customer demographics under each sales channel, businesses can effectively target the right groups for promotions, thereby maximising return on investment in marketing initiatives.

To find the association rules, the apriori algorithm was applied. The variables selected for analysis were the *AgeCategory*, *NumChildren*, *Education*, *Relationship*, *Recency* and *YearsJoined*. *Income* was not selected because a high-income group is likely to have a high number of purchases across all sales channels, which does not yield any valuable insight for the analysis. The numerical variables were divided into categories based on quantiles, and the categorical variables were relabelled for the association rule table.

We examined each sales channel separately. The support was selected to be at least 5% to derive more interesting insights for customer subset and the confidence to be at least 30%, 40% and 60% for web, store and catalogue channels respectively for interpretability. The interesting rules were then summarised into the table below.

LHS	RHS	Support	Confidence	Lift
{Education=PhDs}	{NumWebPurchases=High}	0.07051865	0.3283898	1.171755
{AgeCategory=BabyBoomer}	{NumWebPurchases=High}	0.11692448	0.3346354	1.194040
{NumChildren=YesChild, YearsJoined=>8 years}	{NumWebPurchases=High}	0.05914468	0.3757225	1.340646
{NumChildren=NoChild, Relationship=Not-Partnered}	{NumStorePurchases=High}	0.05414013	0.4817814	1.654618
{AgeCategory=BabyBoomer, NumChildren=NoChild, YearsJoined=<8 years}	{NumCatalogPurchases=High}	0.06505914	0.7009804	2.255864

Table 4: Association Rules (selected)

The analysis of association rules revealed that the customers born in the 'Baby Boomer' generation and those without children are strongly associated with a high number of purchases across all sales channels. Moreover, customers with a PhD and those who have been with us for more than 8 years showed a preference for web purchases. For both store and catalogue channels, customers without no children are most associated with high purchases.

Evaluation and Interpretation

One of the most important objectives of a marketing department is to retain the most profitable customer segment. This is because this customer group has a more elastic demand and higher profit margin, and they are not constrained so much by their budget constraints compared to the poorer groups which might spend a large proportion of their money on necessity goods rather than high-end products such as meat and fish.

Moreover, the marketing team should also focus marketing efforts on the customer group who are the most responsive to marketing campaigns. This is because their Return on Investment (ROI) would be higher.

While prioritising the retention of high-income and spending customers, it is also important not to overlook other customer groups. The marketing department could target campaigns such as discounts, bundles, and deals to attract the corresponding customer group. This is where our analysis brings value.

From the models, Cluster 1 is affluent consumers more responsive to campaigns and spend more on meat than other cluster groups. This conclusion supports real-time research conducted by PYMNTS' ((2022) data, where 77% of high-income earners buy more prepared food and fresh meat and vegetables than other income groups due to being accustomed to cooking more home-cooked meals. The logistic regression also concludes that individuals, who have been responsive to past campaigns, are likely to accept further campaigns.

Therefore, we advise the business to cater to a more personalised customer shopping experience for Cluster 1, by offering and advertising high-quality meat products. As there are more responses to campaigns, the business can comfortably have higher advertising expenses without the expense of experiencing wastage in the campaigns. Furthermore, they can sell their products, especially meat, at slightly higher competitive prices by restructuring the retail format to a supermarket as they perceive shopping in supermarkets as a status symbol or a way to indulge in luxury experiences of high-quality meat and other products, and also high-quality customer services (Roques, 2023).

Cluster 2 is the lowest-income consumers that are less responsive to campaigns against other cluster 1. Since they are mainly Bachelor's degree earners, they are considered the least spenders and the youngest among the other cluster groups. The classification tree also shows that they have significantly lower spending tendencies.
Therefore, the business should consider other marketing strategies, other than spending on advertising campaigns to them. One credible general marketing strategy is the use of Electronic Word of Mouth (eWOM), for example, admired social media influencers could help persuade young consumers to try the retail products by incorporating the shopping experience in the business as a positive lifestyle in their videos and online conversations (Glucksman, 2017).

Cluster 3 is the oldest generation and middle-income consumers that are highly responsive to deals and the highest number of children on average as they may be price-sensitive. They have a similar proportion of past campaign acceptance with cluster 1. Interestingly, the web purchase channel seems to be their most preferred against other sales channels. However, they have lower web visits, which shows that businesses need to take this into account in their marketing strategy to improve their web visit rate. Furthermore, they spend more on wine, meaning they are more likely to be heavy wine drinkers or party organisers.

The business can create a new web page on their website to be more family-oriented, for example, a picture of the happy family shopping for the products on the website will attract cluster 3 as it reflects a positive retail experience. Furthermore, we can personalise the coupons, for example, wine discounts, kid birthday discounts and "holiday discounts", to target them. The survey poll, conducted by Marketo, Inc. (2015) found that 79% of consumers are likely to engage with brands offering discounts that are tailored to their past shopping interactions with the brand. Therefore, we can learn more about the older generation and collect higher-quality customer data via their response and purchasing history to personalised coupons.

Utilising hierarchical and k-means clustering methods, we have successfully segmented the customers in the dataset into three relatively distinct groups. We further refined the defining characteristics in each segment using a supervised classification tree. These were followed by the association rule analysis to enrich our understanding of customer behaviour, revealing the sales channels each customer group frequently spends on.

The major strength of our machine learning workflow model is that it offers a comprehensive view of consumer behaviour, allowing us to extract insights and turn them into actionable business plans. These results are crucial for the marketing department for the businesses to roll out micro-targeted marketing strategies for different customer groups.

An additional strength of our analysis is the anonymisation of customer data. By tagging customers with their IDs rather than their names, we have addressed the ethical concerns and respected the privacy of the agents.

Employing PCA to our clustering model helps to denoise the data to retain only the important personal information that distinguishes each cluster and enhances the quality of our analysis. Careful considerations needed to be taken when conducting the clustering analysis which involves standardisation of our variables, where to cut the dendrogram, which linkage and the dissimilarity measure to use, and the number of clusters to select. These considerations, including choosing the optimal k number of clusters, were tackled based on our intuition and with the help of Elbow and Silhouette methods.

To improve the interpretability in each of our models, we use the tree pruning method to enhance interpretability in our classification tree which is a common flaw when working with large datasets. Moreover, in the logistic regression model, we avoided the use of highly-correlated features and retained only the significant ones to report as the main factors that determine the acceptance of the last campaign (Byanjankar et al, 2023).

We compared the regression model with a predictive model to address the imbalance issue in our outcome variable in the dataset and employed a test-train split. The logistic regression model achieved a predictive accuracy of 94%.

The association rules have varied support and confidence levels, and have inconsistency associations between each channel. For example, we saw that there is a low confidence level, around 30%, when finding the associations to *NumWebPurchases* and a high confidence level, around 60%, when finding the associations to *NumCataloguePurchases*. Therefore, we selected ones that have the best associations that are relevant and interpretable to our business customer segmentation problem.

A flaw in our dataset is that the absence of country-specific data limited our analysis. Since each country has their own unique market dynamics, the lack of this information would hinder the external validity of our findings to a multinational retail business.

Conclusion

Our study provides credible insights on the customer segmentation of the marketing departments operating under the retail business. The analysis showed the potential efficiency of data analytics tools in enhancing marketing campaigns. By employing several algorithms on customer dataset, we were able to segment customers based on features and develop micro-targeting strategies for the retail business in a highly-competitive retail industry.

Furthermore, our data analysis highlighted the customer data as a ‘real’ asset to the retail business. The large flow of data being supplied by the customers can be gathered and turned into actionable data points to optimise marketing and sales operations, and in turn the long-term profitability of the business.

Finally, the use of data analytics tools reduces the informational asymmetry, allowing businesses to better understand their customers. This is achieved while considering ethical considerations and respecting privacy of agents.

The results obtained from our analysis have proven to be highly beneficial in addressing the practical problem of customer segmentation within the retail industry. Our three advanced analytic approaches allowed us to identify the unique differentiating characteristics in different customer segments, enabling businesses to design targeted, refined and comprehensive marketing strategies.

Incorporating the insights from McKinsey article (2023), it is clear that advanced analytics can yield data-driven insights that create value across the full business. For example, Macy’s has undertaken significant marketing efforts to attract millennials, introducing 13 segment-specific brands and a marketing mix that includes social media and a new blog.

Our analysis framework would enhance the cost-effectiveness of marketing campaigns. By identifying and targeting specific customer segments, the marketing team can now allocate their resources more efficiently, aiming to improve the customer satisfaction and improve the market share and ultimately achieve long-term profitability in this fiercely competitive industry.

Although our analysis provides a comprehensive approach for understanding the customer behaviour, further improvements can be made, particularly considering the applicability to all industries without the concern of dynamics, and region. However, the randomness of human nature should also be kept in mind at all times.

Firstly, we can expand the geographical scope of data collection, which will provide a more comprehensive understanding of the customer behaviour across different countries and cultures.

Secondly, we can expand the horizon for the data collection period for the spending variables as the dataset only covers a period of two years. This will better capture a more reliable picture of the customer shopping behaviour and patterns over time.

Moreover, the analysis can be enriched by incorporating additional features that could influence customer behaviour, such as their lifestyles, attitudes and other preferences factors, which can enable us to further understand our customers for personalisation effort.

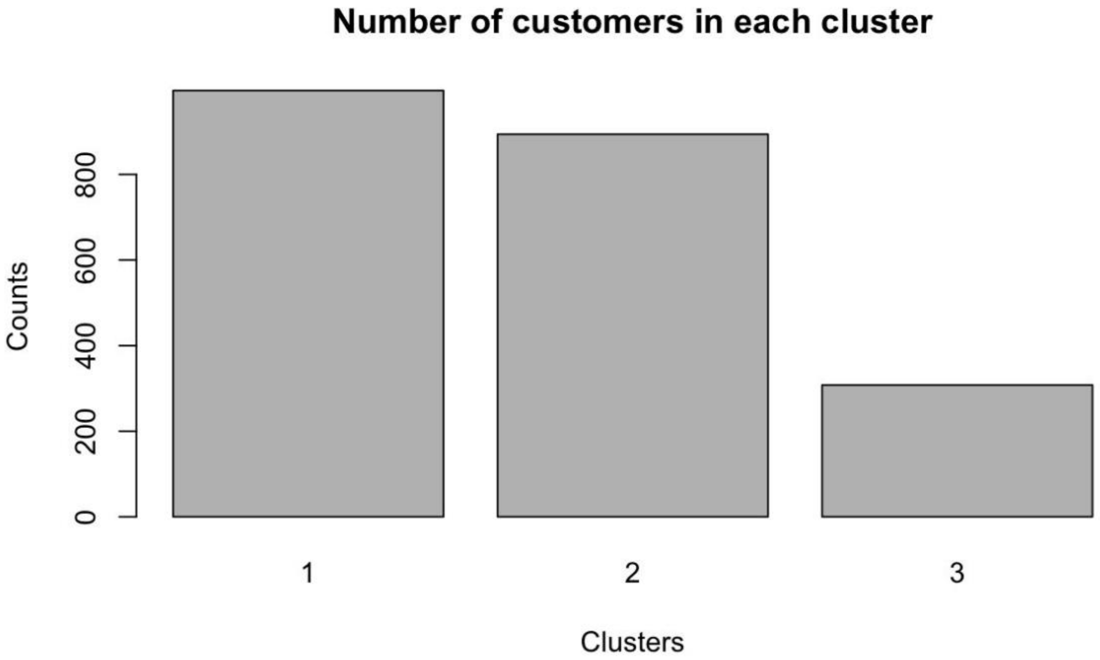
Finally, the needs and preferences of customers are likely to change dynamically, and hence our segmentation result can suffer from instability. Therefore, we should incorporate the dynamic, time-varying aspect in the future analysis to prevent outdated results.

Bibliography

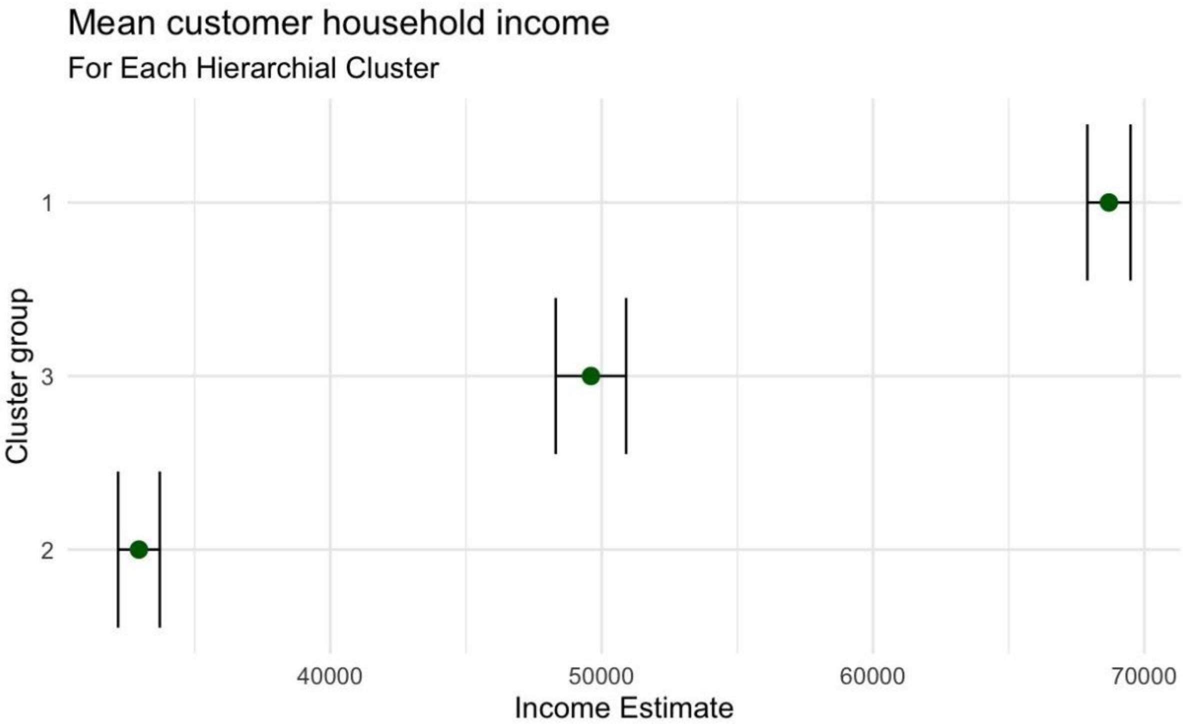
1. Lindecrantz, E., Gi, M. T. P., & Zerbi, S. (2020, April 28). *Personalizing the customer experience: Driving differentiation in retail*. McKinsey & Company. Retrieved from: https://www.mckinsey.com/industries/retail/our-insights/personalizing-the-customer-experience-driving-differentiation-in-retail
2. Weinstein, A. (2004, January 1). *Handbook of Market Segmentation*. Psychology Press, pp.12. Retrieved from: http://books.google.ie/books?id=vFpZcc-nZG0C&pg=PA228&dq=Weinstein,+A.,+Market+Segmentation+Handbook:+Strategic+Targeting+for+Business+and+Technology+Firms,+3rd+ed.,+Haworth+Press,+Binghamton,+N.Y.,+2004,+p.+12&hl=&cd=3&source=gbs_api
3. Mehta, J. (2023, November 17). *The Evolution of Customer Segmentation: From Mass Marketing to Individualized Approaches*. Abmatic AI. Retrieved from: https://abmatic.ai/blog/evolution-of-customer-segmentation-from-mass-marketing-to-individualized-approaches
4. Kapoor, K. (2022). *Customer Segmentation: Clustering.* Kaggle. Retrieved from: https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/notebook
5. Mata, W. (2024, Feb 1). *Which generation are you? Age ranges from Gen Z to Baby Boomers explained.* The Standard. Retrieved from: https://www.standard.co.uk/news/uk/which-age-range-who-generation-z-millennial-boomer-zoomer-b1073540.html
6. Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). *Log-transformation and its implications for data analysis. Shanghai archives of psychiatry, 26(2), 105–109.* Retrieved from: https://doi.org/10.3969/j.issn.1002-0829.2014.02.009.

7. West, R.M. (2022). *Best practice in statistics: The use of log transformation. Annals of Clinical Biochemistry.* 59(3):162-165. doi:10.1177/00045632211050531. Retrieved from: https://journals.sagepub.com/doi/10.1177/00045632211050531
8. James G., Witten D., Hastie T., Tibshirani R. (2013). *An introduction to statistical learning : with applications in R.*, pp.(127-137,303-315,373-399), New York :Springer.
9. Kassambara, A. (2017): *Practical Guide To Cluster Analysis in R Unsupervised Machine Learning, Chapter 12*
10. MacKenzie, I .,Meyer, C., Noble, S. (2013, October 1). *How retailers can keep up with consumers.* McKinsey & Company. Retrieved from: https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers/medium.com/@rishabhchandaliya/market-basket-analysis-using-association-rule-mining-6e8947900b21
11. PYMNTS. (2022, August). *Consumers Buy In To Food Bargains.* Retrieved from: https://www.pymnts.com/study/digital-economy-payments-consumer-finance-grocery-restaurants-inflation/
12. Roques, P. (2023, May 1). *How to ruin a pricing strategy in food retail.* LinkedIn. Retrieved from: https://www.linkedin.com/pulse/how-ruin-pricing-strategy-food-retail-philippe-roques/
13. Glucksman, M. (2017). *The Rise of Social Media Influencer Marketing on Lifestyle Branding: A Case Study of Lucie Fink.* Elon University. Retreievd from: https://www.media-education-portal.com/uploads/1/2/4/7/124735657/08_lifestyle_branding_glucksman.pdf
14. Marketo. (2015, June 22). *Consumers to Brands: The Louder You Scream, the Less We Care.* PR Newswire. Retrieved from: https://www.prnewswire.com/news-releases/consumers-to-brands-the-louder-you-scream-the-less-we-care-300102426.html

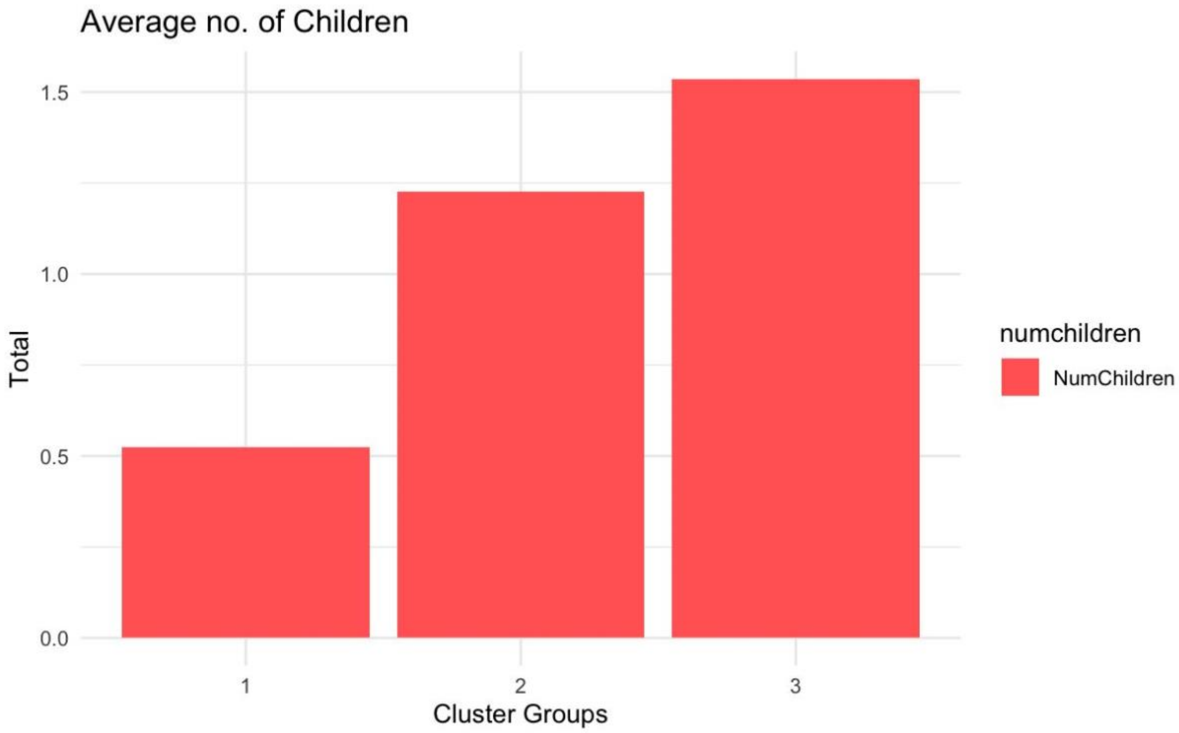
Appendix



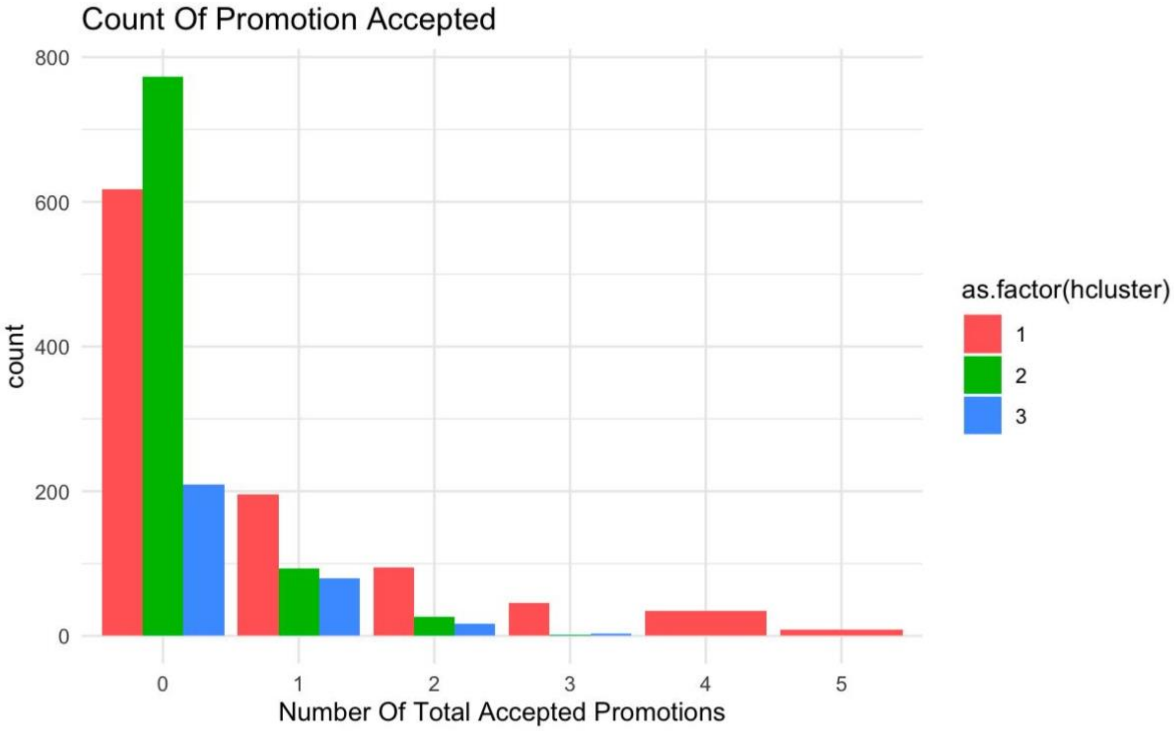
Appendix 1: Number of customers in each cluster



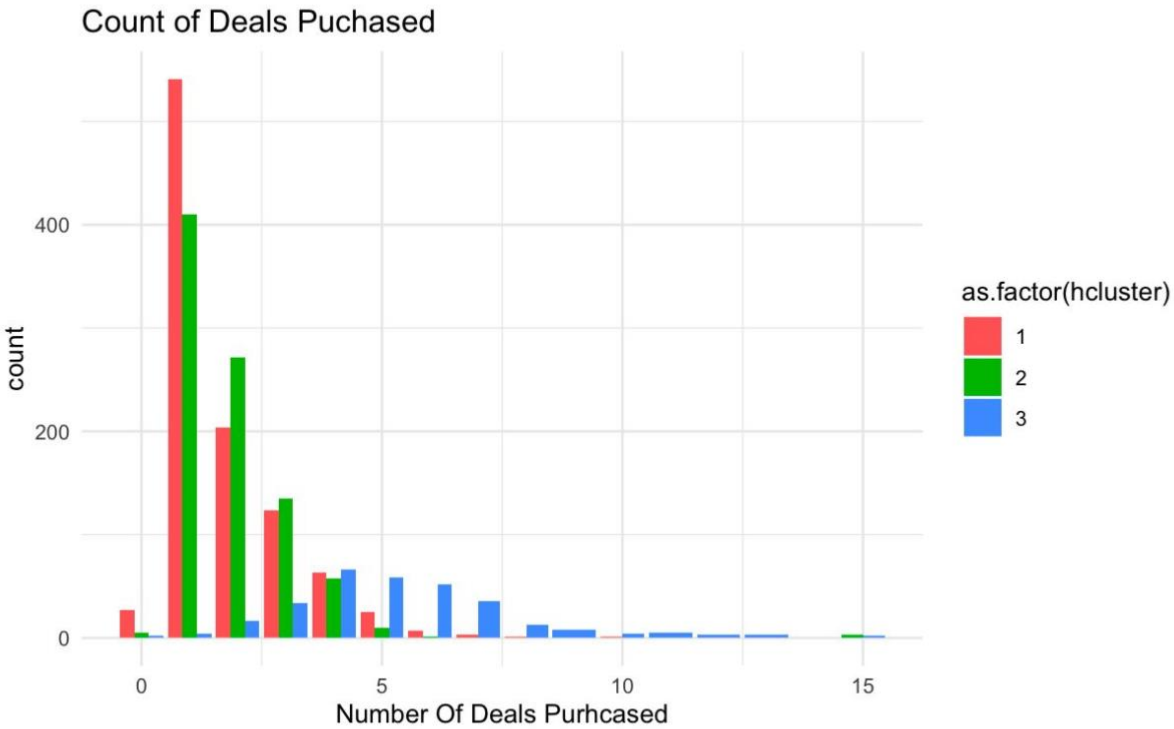
Appendix 2: Mean customer household income



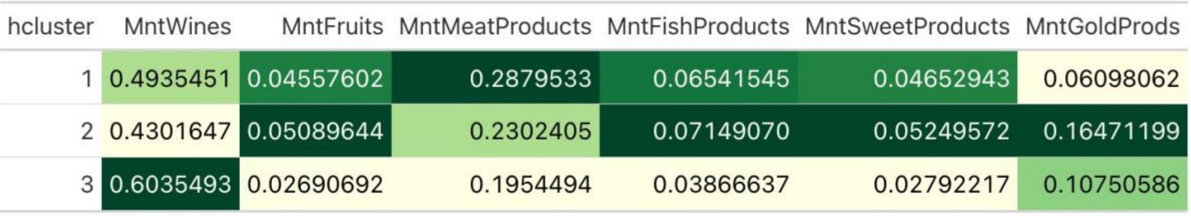
Appendix 3: Average number of children



Appendix 3: Count of Promotions Accepted

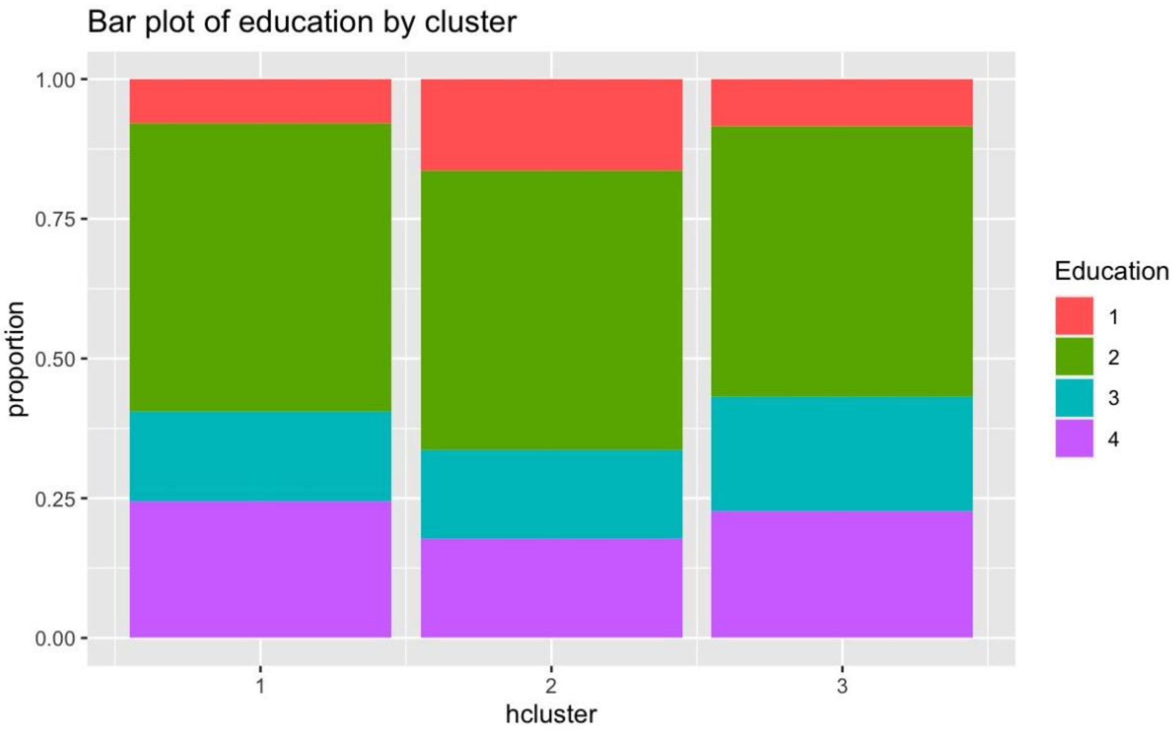


Appendix 4: Count of Deals Purchased

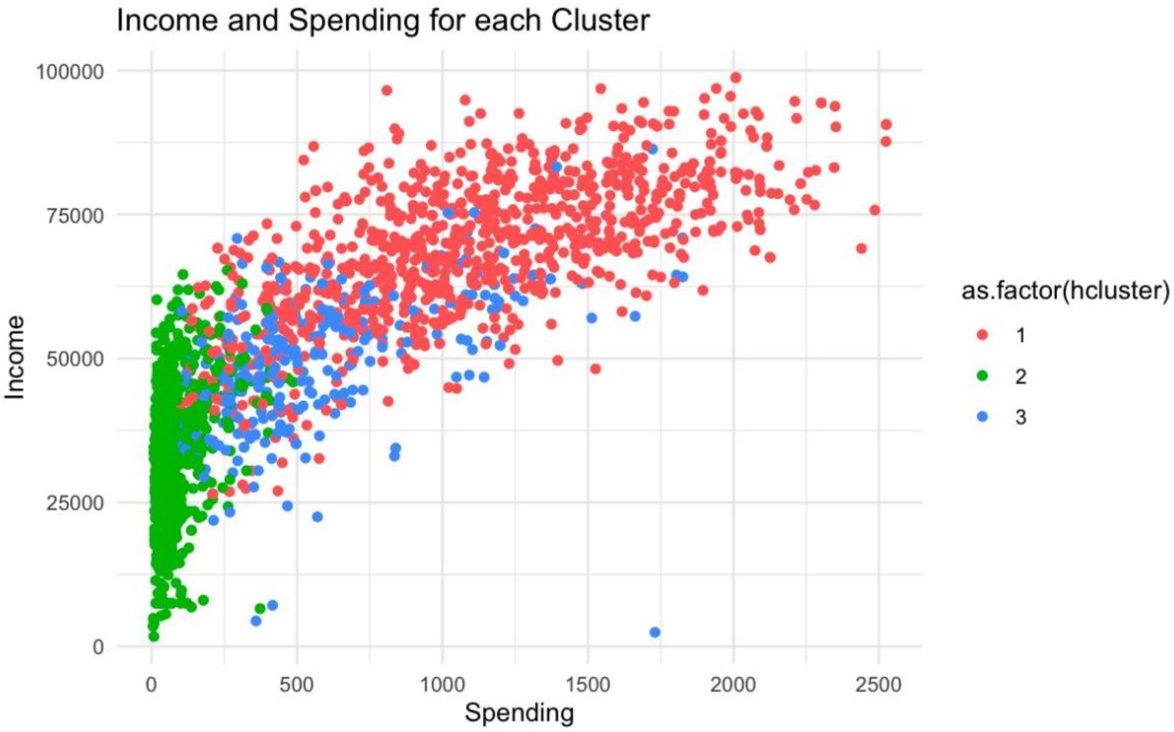


hcluster	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
1	0.4935451	0.04557602	0.2879533	0.06541545	0.04652943	0.06098062
2	0.4301647	0.05089644	0.2302405	0.07149070	0.05249572	0.16471199
3	0.6035493	0.02690692	0.1954494	0.03866637	0.02792217	0.10750586

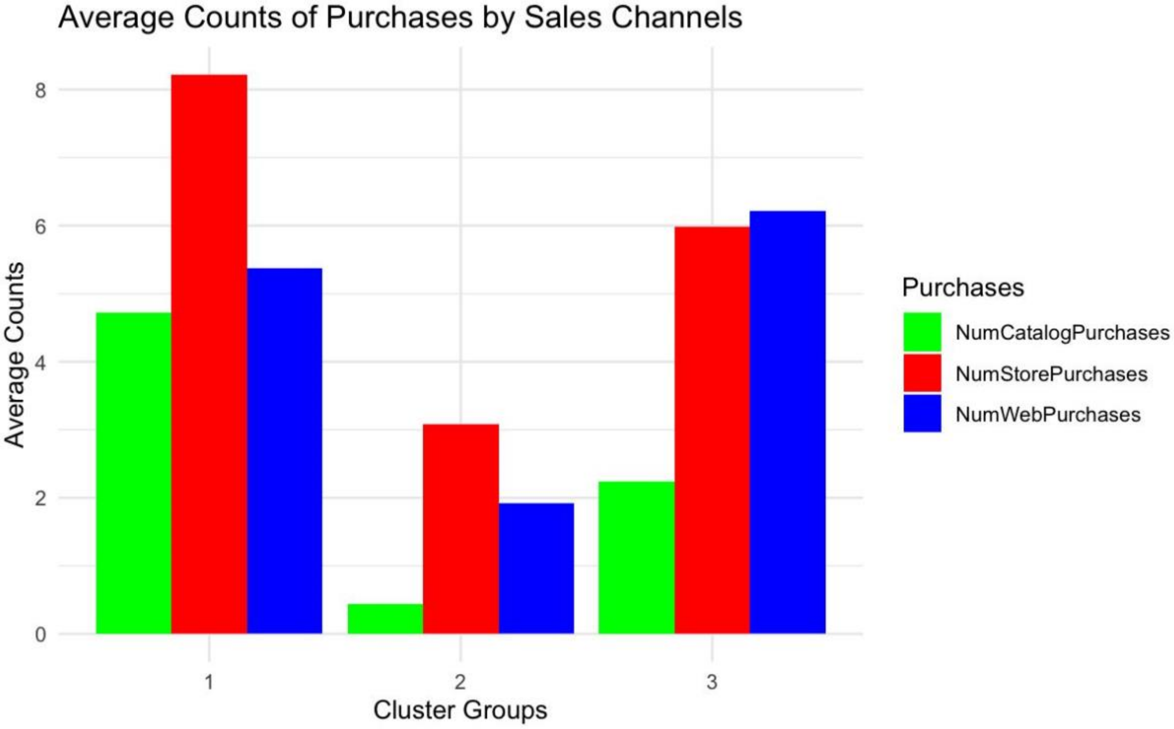
Appendix 5: Spending patterns



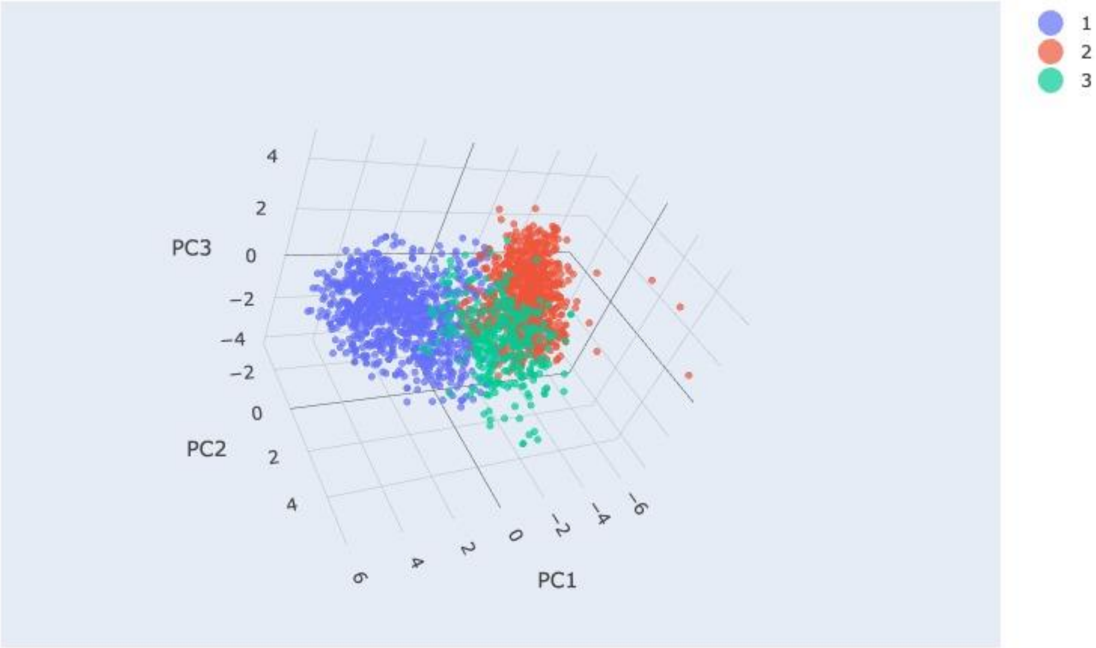
Appendix 6: Education level by cluster



Appendix 7: Income vs Spending



Appendix 8: Sales channels by cluster



Appendix 9: Cluster visualisation with PCA