The ID column was dropped and so were the 24 NAs that were in Income. Upon visualising the univariate distribution of the features, Income and Year_Birth seemed to have outliers, which were removed. As for the discrete variable Marital_Status, there were 3 unintuitive categories with 1 observation in each; these categories were all combined into one as ‘other’. It can be inferred that the vast majority of customers have an income between $25,000 and $75,000. Recency approximates a uniform distribution implying that there are many customers with different days since last purchase. The other purchase related features appear positive and continuous with much concentration towards the left. The Dt_Customer variable was converted to reflect the number of years the customer has been with the company and Year_Birth was changed to reflect the customer age.

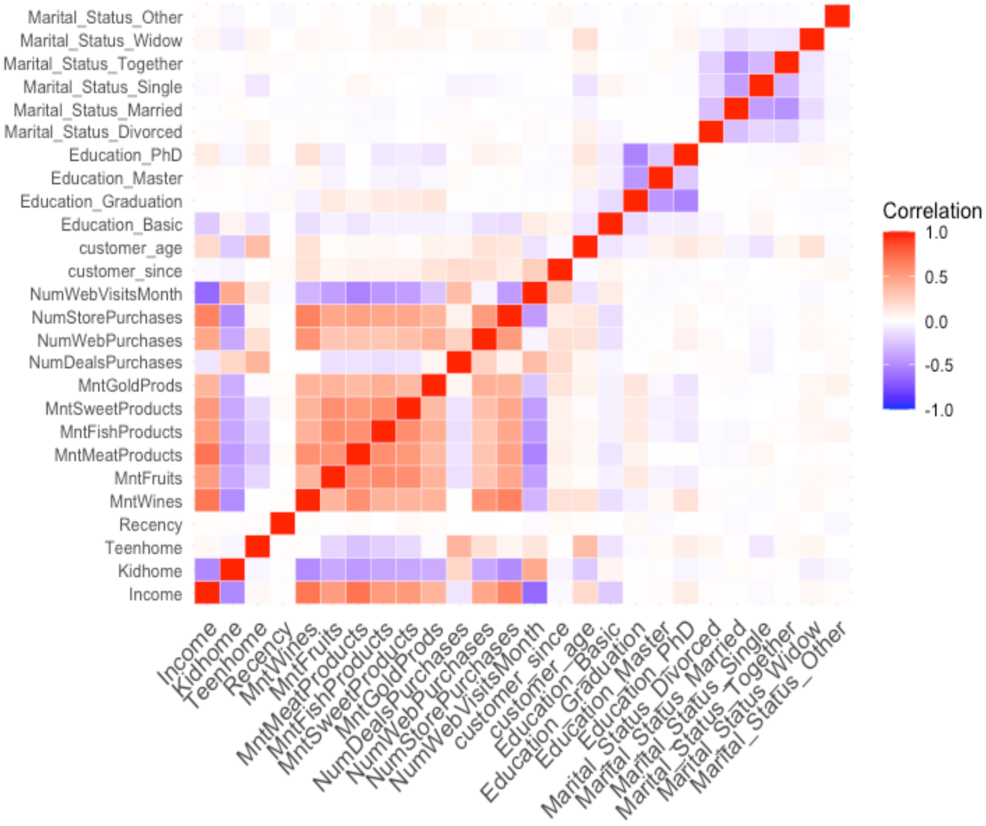


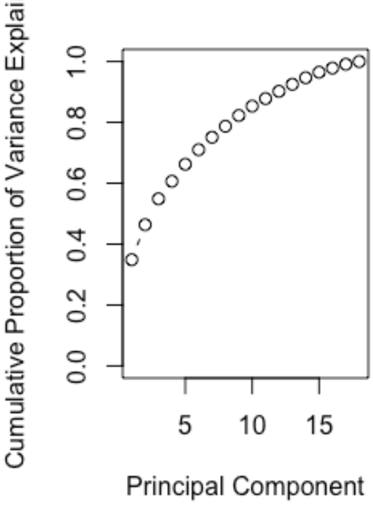
fig 1: Correlation Matrix

The correlation matrix (fig1) shows strong correlation of Income with the purchase-related features such as the amounts spent on wines, meat, fish, etc. and NumDealsPurchases is also positively correlated with customer_age and Teenhome. Marital_Status variables appear quite uncorrelated except with age or if children in the household. Overall, there is correlation between purchasing behaviours, but the features are too many to feasibly understand the customers.

To reduce the dimensions and facilitate the understanding of customers, Principal Component Analysis (PCA) is explored. Before the PCA, the Education variable is ordinally encoded to reflect the natural hierarchy of the education levels; and Marital_Status, being the only categorical variable without much correlation and link between levels was dropped for PCA. This resulted in a higher variance explained by the principal components (PCs) and 5 were chosen with ~66% explained variance (fig2). Before PCA, scaling was done since there were many skewed features.

Upon examining the feature loadings of the 5 PCs and considering only those with greater than 0.25 correlation, the following composition was achieved:

fig 2: PC selection



	Features (loadings>0.25)	**Interpretation**
PC1	Income, Kidhome, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProducts, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth	*"Income and purchase activity"* Customer high on this component will show high purchase-related activities with higher income levels.
PC2	Teenhome, NumWebPurchases, NumDealsPurchases	*"Home-bound and deal preference"* Customer high on this component will priorities savings and may have busier lifestyles due to teens and web purchases.
PC3	customer_since	*"Duration of customer relation"* This component captures a customer's duration of relation with the company.
PC4	Education	*"Education level"* High on this component means more educated is a customer.
PC5	Recency	*"Brand attraction of customers"* This component explains when the last purchase was made by a customer.

The PCs explain different behaviours of the customers and as noted in the correlation matrix earlier, Income and purchase-related features were correlated and PC1 has captured that direction and same is for PC2. For example, if a customer buys meat, then he/she is also likely to buy other associated products like wine. PCs 3,4, and 5 do not explain a lot of features but collectively add to distinguish customers in a five-dimensional space instead of the original 20.

To understand the categories of customers, clustering was explored on the reduced 5-dimensional feature space to segment customers based on similar purchasing behaviours. The number of clusters are determined by the within sum of squares and gap statistics method. The total within sum of squares decreases significantly until k=3, and then the decreases flattens-out. The elbow method suggests that 3 clusters are optimal for segmentation, and the same is confirmed by the gap statistics plot where the gap is maximized for clusters. Upon running the k-means clustering with 3 clusters and comparing them on PC1 and PC2, it is observed that the clusters are distinct from each other and that some clusters are sparser than the others. For instance, clusters 1 & 3 are more dispersed than cluster 2 and 1020 observations fall into group 2 when compared to 595 & 597 observations in cluster 1 & 3.

From fig3 it can be inferred that cluster 1 scores highest on PC2, which means that they are mid-income families who have teenagers at home and have a high preference for deals/discounted products. They are mostly middle-aged and educated customers who are frequent shoppers, and do not have a clear preference for online or in-store shopping. Thus, they are comfortable with both channels of shopping and do so on a convenience basis. On the other hand, cluster 2 scores extremely low on PC1, meaning it is a low to mid income group with kids at home, and comparatively shop less frequently for the products, and have an average level of education. They are not likely to spend on items such as gold, sweets, and other indulgence products, and do not seem to be purchasing any deals/discount products. In cluster 3, it

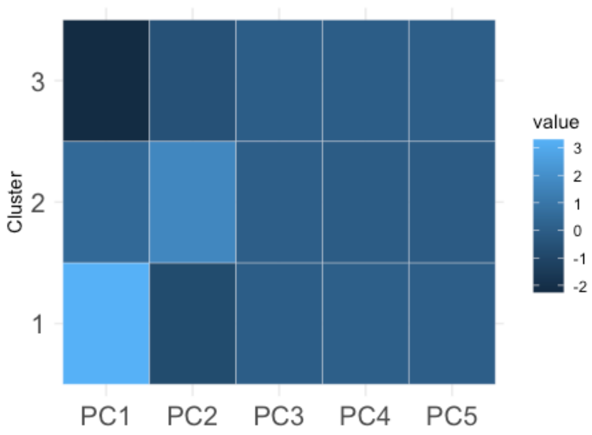


fig 3: Cluster-wise performance on Principal Components

can be observed that the customers are a high-income group, who are regular shoppers with a taste for indulgent products such as wines, chocolates etc. as these groups most likely have kids at home and do not prefer buying discounted products.

Almost all observations in all three groups seem to have started shopping at the store at the same time and do not seem to have a clear preference for the channel of purchase. They also seem to be all have the same level of education.

Since PC1 and PC2 have shown the greatest distinction between clusters, fig4 visualises these clusters to observe this distinction in consumer behaviour. The actual means of the main features grouped by the clusters validate the outcomes of clustering using the PCs.

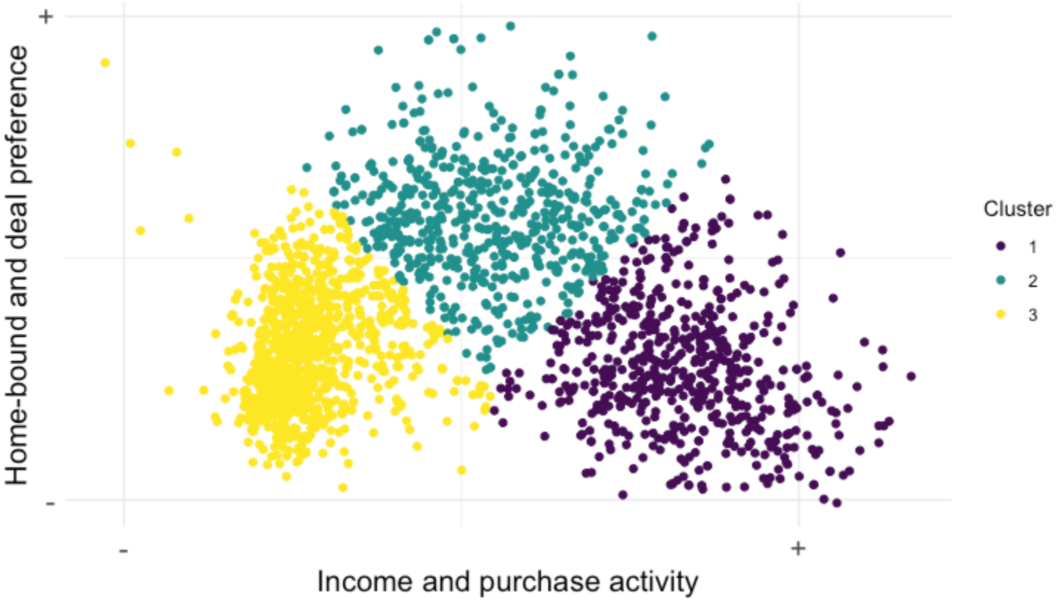


fig 4: Cluster-specific purchase behaviour