Executive Summary

This report presents a detailed segmentation analysis for SmartFresh Retail, employing Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), Cluster Analysis, and Independent Sample T-Tests to refine marketing strategies. The findings indicate that SmartFresh primarily serves a financially stable, middle-aged customer base, with considerable variation in spending behaviour and engagement with promotions. While a small segment of high-spending elite consumers exists, overall responsiveness to promotional offers is inconsistent across different customer groups.

PCA on spending behaviour identifies three distinct purchasing patterns: a health-conscious segment that prioritises organic food and wellness products, a traditional shopper group focused on wine and meat, and a luxury-oriented segment that favours premium goods. Cluster analysis segments customers into three groups: Elite-Consumers, who are affluent, demonstrate low engagement with promotions, and favour high-end products; Economical-Consumers, who have moderate incomes and the highest engagement with promotional offers; and Budget-Consumers, who are price-sensitive and primarily purchase essential goods. T-Tests further confirm statistically significant differences in promotional responsiveness across these segments, reinforcing the validity of the classification.

The report recommends targeted marketing strategies, including bundled essentials for Budget-Consumers, optimised personalised promotions for Economical-Consumers, and enhanced premium loyalty initiatives for Elite-Consumers. By aligning marketing efforts with these consumer behaviours, SmartFresh can strengthen customer retention, increase revenue, and improve the effectiveness of promotional campaigns.

Introduction

SmartFresh Retail, a high-end omnichannel retailer, seeks to construct a roadmap to refine its approach to customer segmentation. In order to achieve this, the company has gathered a wealth of customer data, including demographics, purchasing habits, and responses to past promotional campaigns. This report investigates the ways in which SmartFresh can gain insights into various segments and ultimately, create better marketing campaigns, optimise promotions, and personalise customer interactions.

Employing data-driven segmentation can allow SmartFresh to improve retention strategies, allocate resources more efficiently, and enhance its market position (Uddin *et al.*, 2024). The insights gained will help the company design customer-centric marketing strategies that maximise loyalty, satisfaction, and profitability. Figure 1 demonstrates that segmented and well-targeted promotional campaigns deliver a staggering 77% of marketing return on investment (Gherca, 2023).

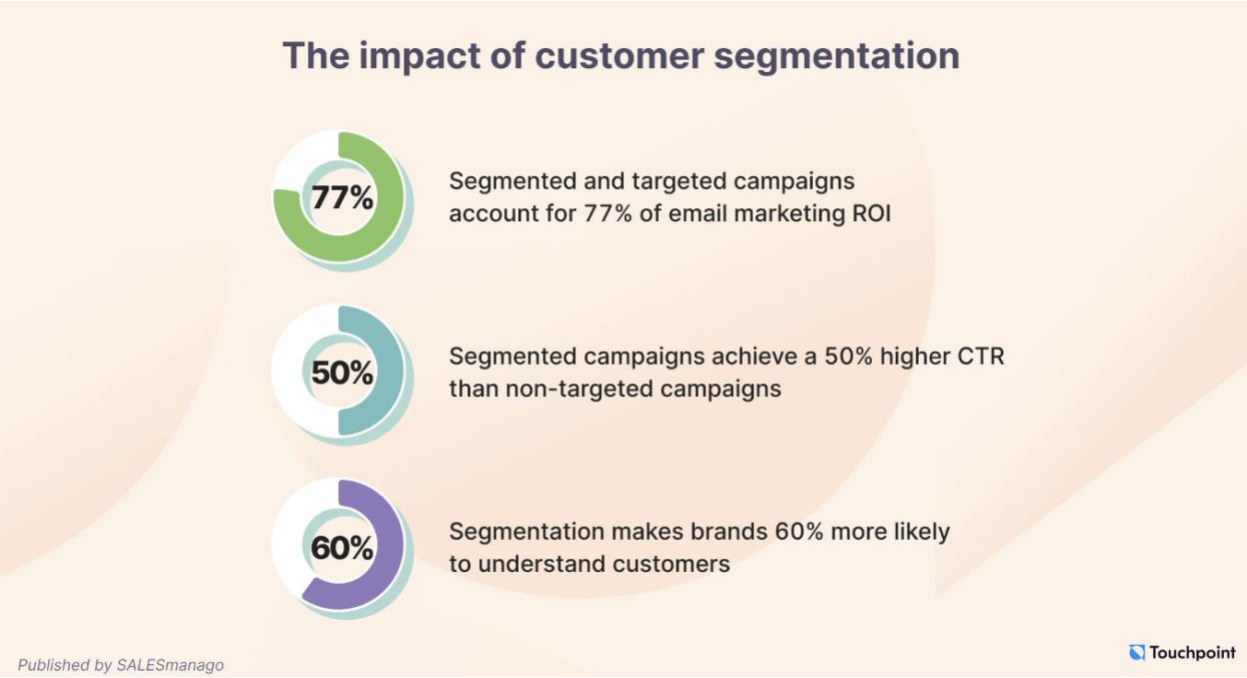


Figure 1: Impact of Customer Segmentation (Gherca, 2023)

Exploratory Data Analysis (EDA)

EDA involves summarising and visualising data to reveal trends, detect anomalies, and explore relationships between variables (Komorowski *et al.*, 2016). It provides a foundational framework for evaluation of customer behaviours, spending patterns, and promotional engagement by identifying trends that shape SmartFresh's marketing strategies (G *et al.*, 2024).

Categorical Feature Analysis

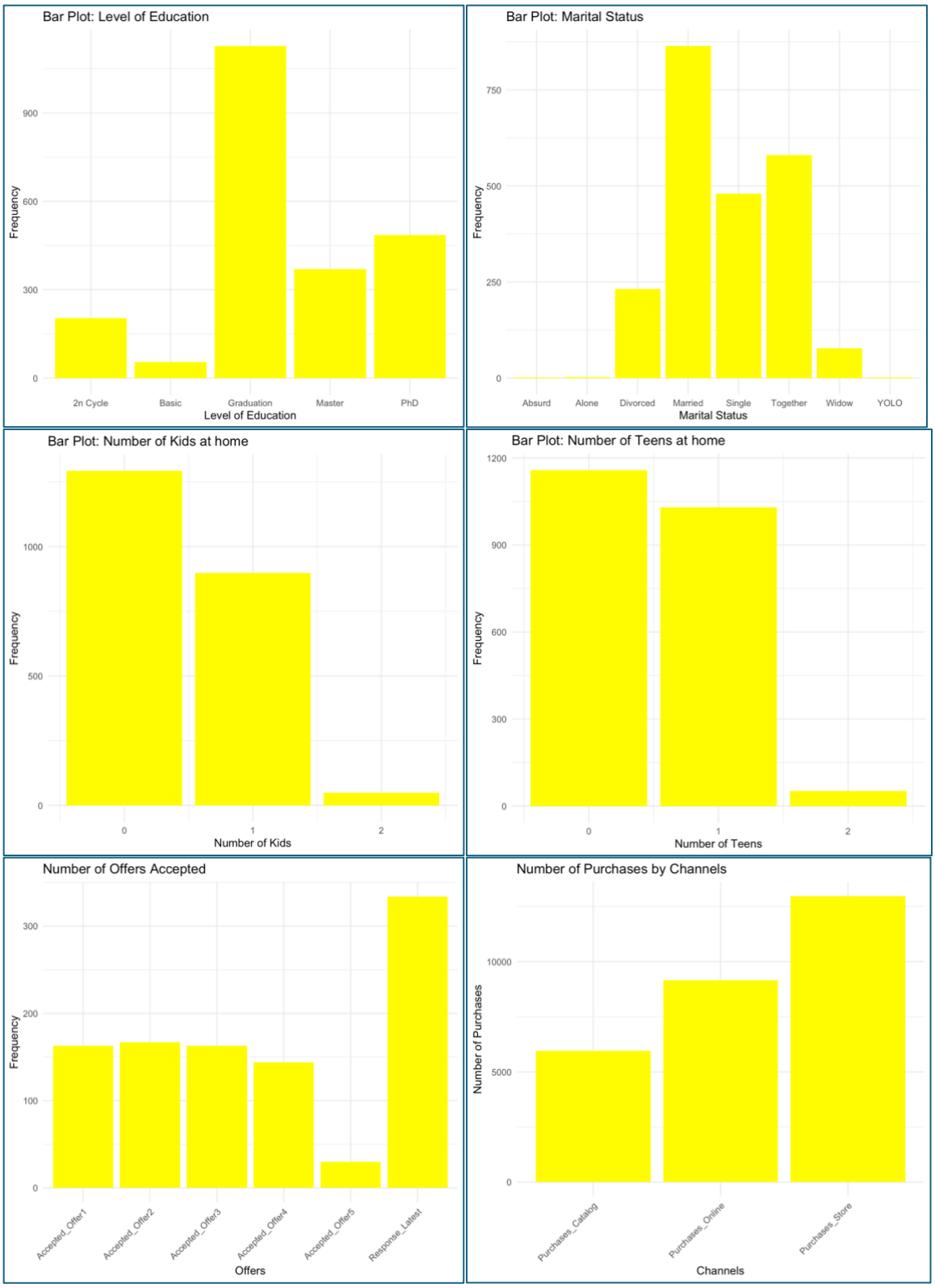


Figure 2: Bar Plots of Categorical Variables

Figure 2 shows most customers are graduates and married, indicating a financially stable consumer base. Households generally have few or no children, making broad family-focused promotions less effective. Shopping behaviour reveals a preference for in-store shopping, while online engagement remains limited.

Numerical Feature Analysis

	Mean	Variance	Std_Dev	Skewness	Kurtosis
Age	56.19	143.62	11.98	0.35	0.71
Annual_Income	52247	633683789	25173	6.76	159
Spend_Wine	304	113298	337	1.17	0.59
Spend_OrganicFood	26.30	1582	39.77	2.10	4.04
Spend_Meat	167	50947	226	2.08	5.50
Spend_WellnessProducts	37.53	2984	54.63	1.92	3.09
Spend_Treats	27.06	1704	41.28	2.13	4.36
Spend_LuxuryGoods	44.02	2721	52.17	1.88	3.54
Spend_Total	606	362704	602	0.86	-0.34
Purchases_Online	4.08	7.72	2.78	1.38	5.69
Purchases_Catalog	2.66	8.54	2.92	1.88	8.03
Purchases_Store	5.79	10.57	3.25	0.70	-0.62
Purchases_Total	12.54	51.92	7.21	0.30	-1.12
Promo_Purchases	2.33	3.73	1.93	2.42	8.91
Response_Latest	0.15	0.13	0.36	1.97	1.88

Table 1: Descriptive Statistics

Table 1 reveals that age and spending distributions are right-skewed, with luxury spending exhibiting high skewness (2.1) and kurtosis (21.5). This indicates that only a small fraction of customers engages with premium products. Annual income shows high variance, suggesting an economically diverse customer base. Promotional purchases have a variance of 3.73, reflecting inconsistent engagement across customer segments.

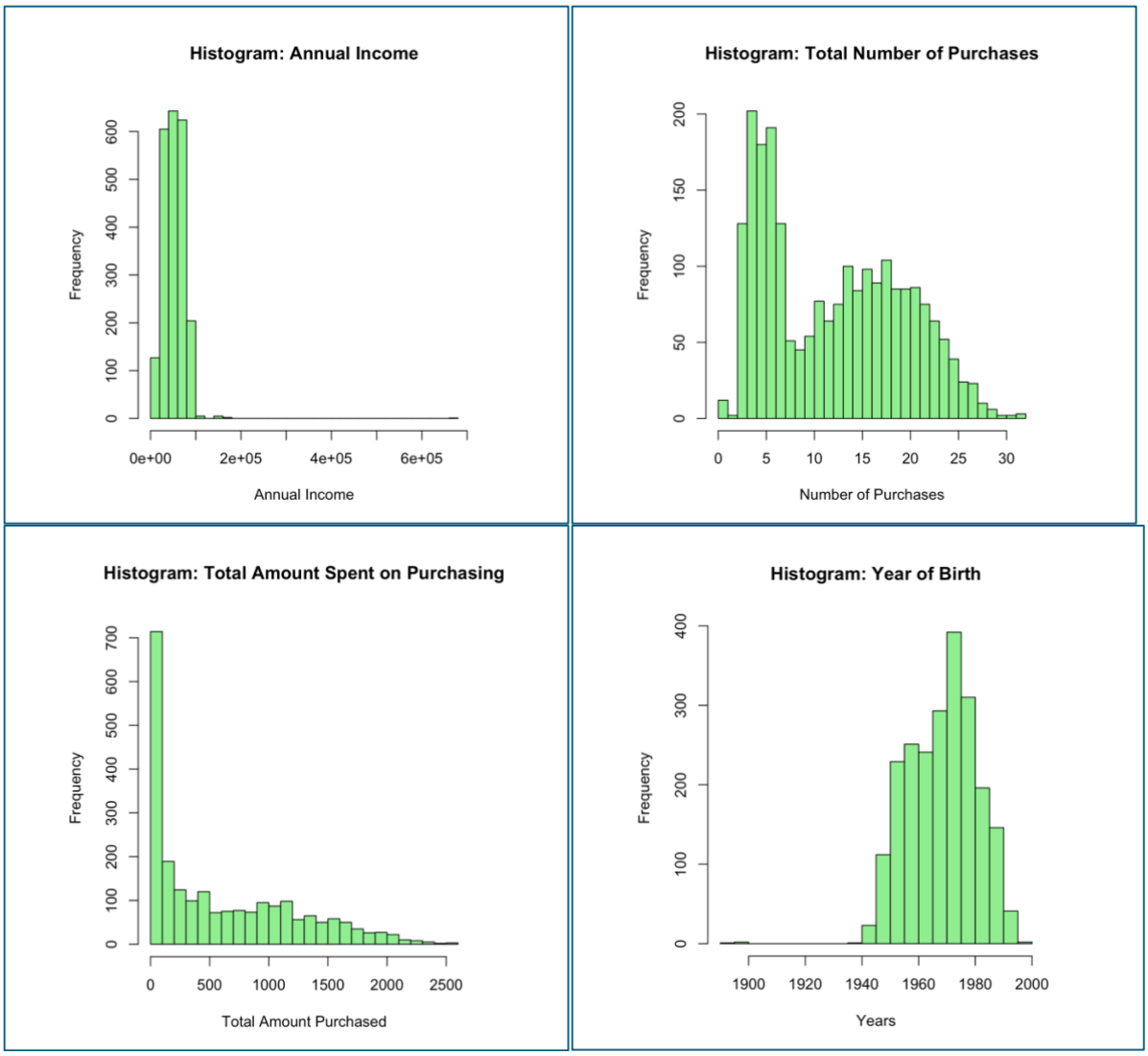


Figure 3: Histograms of Numerical Variables

Figure 3 Age distribution skews toward 40-65-year-olds, with a mean age of 56 years indicating that SmartFresh primarily serves a mature audience, which may influence product preferences and marketing strategies. Annual income is right-skewed, with a mean income of £52,237, driven by a few high earners.

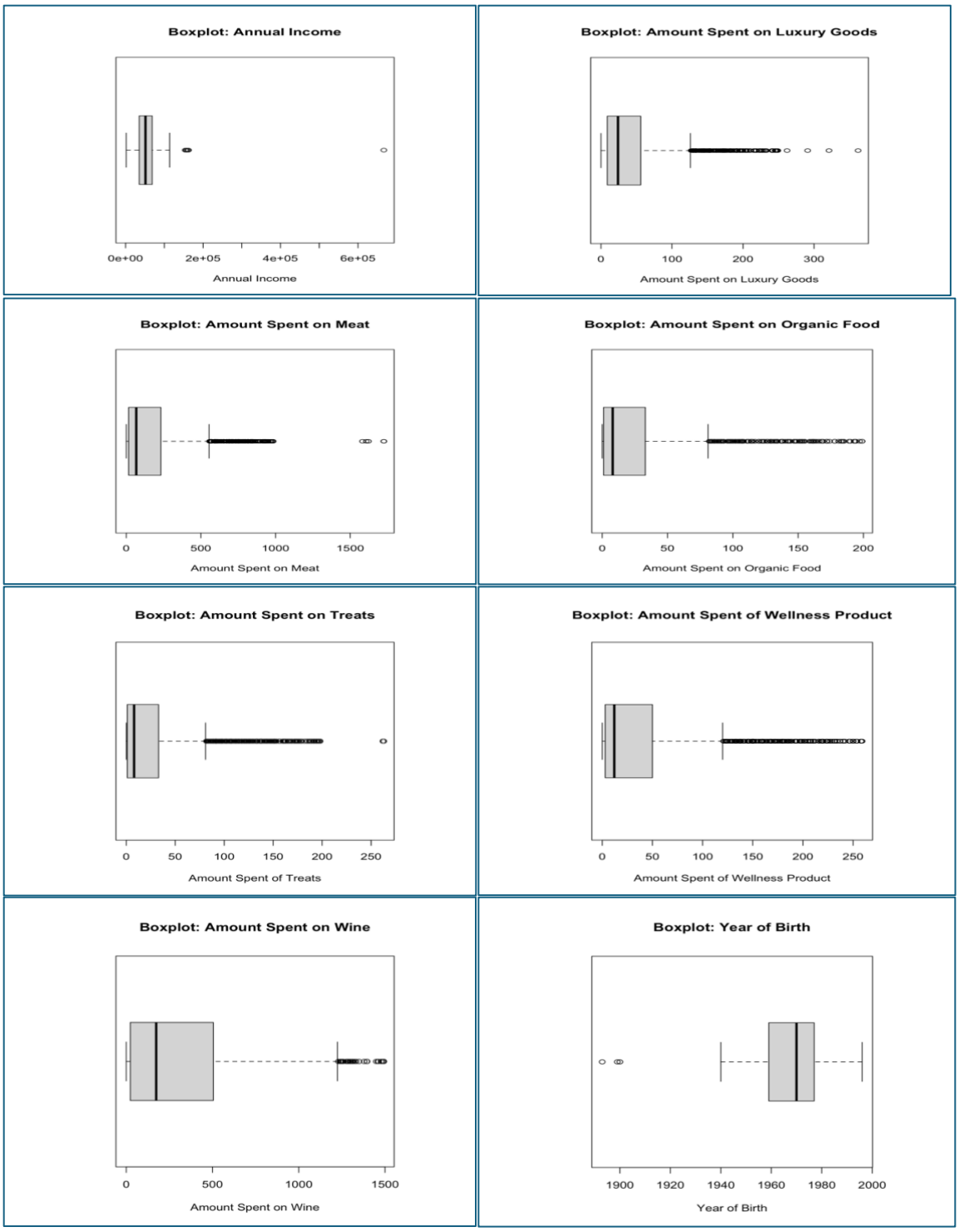


Figure 4: Boxplots of Numerical Variables

Figure 4 highlights high spending variability, especially in wine and meat, the store's top-selling categories. Wine spending shows a variance of 113,311, while meat has a variance of 50,939, indicating a diverse purchasing pattern. Outliers in income and spending suggest a niche ultra-wealthy segment.

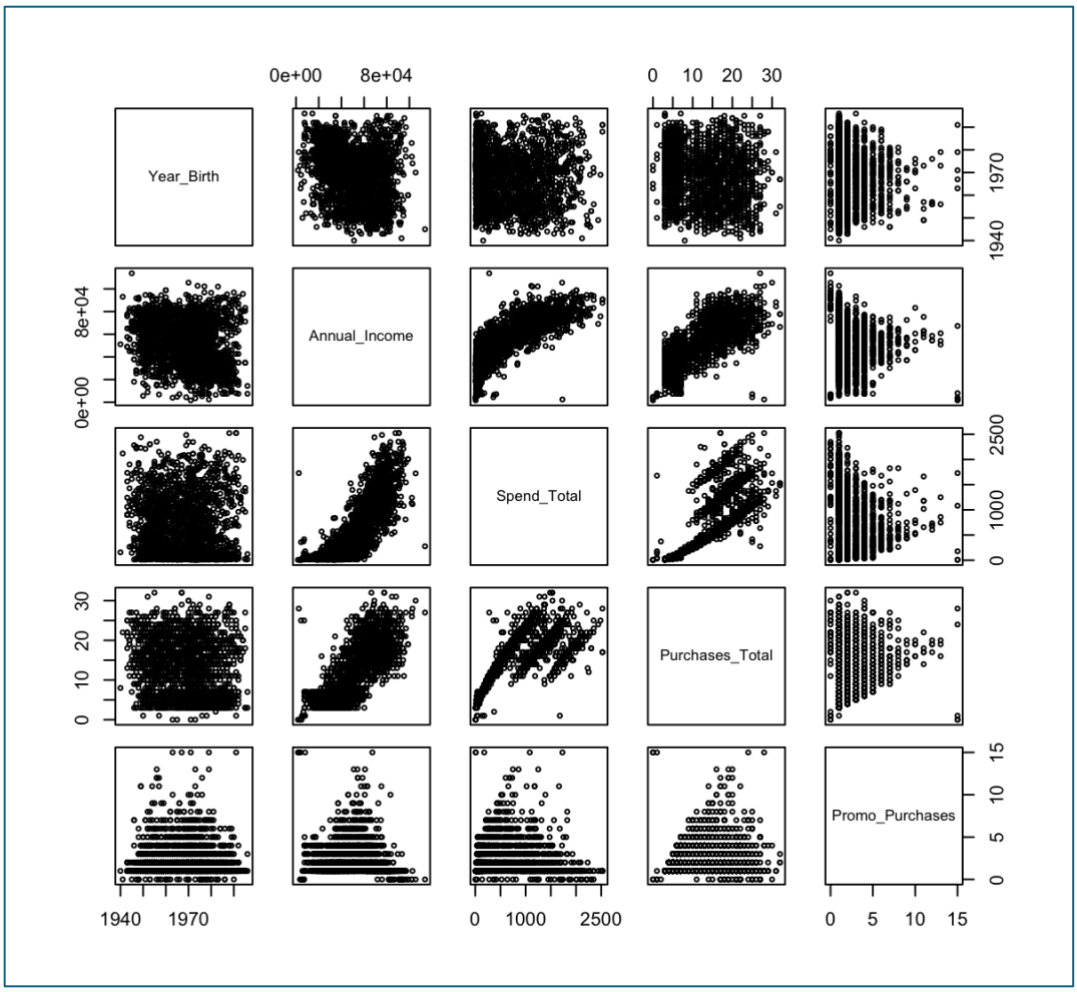


Figure 5: Pair Plot of Numerical Variables

Figure 5 reveals that middle-aged consumers engage most in promotions, particularly those with moderate incomes. Promotional engagement has a variance of 3.73, indicating that responses are inconsistent across this group. Additionally, moderate-income earners show higher purchase frequency, while high-income consumers do not proportionally increase spending.

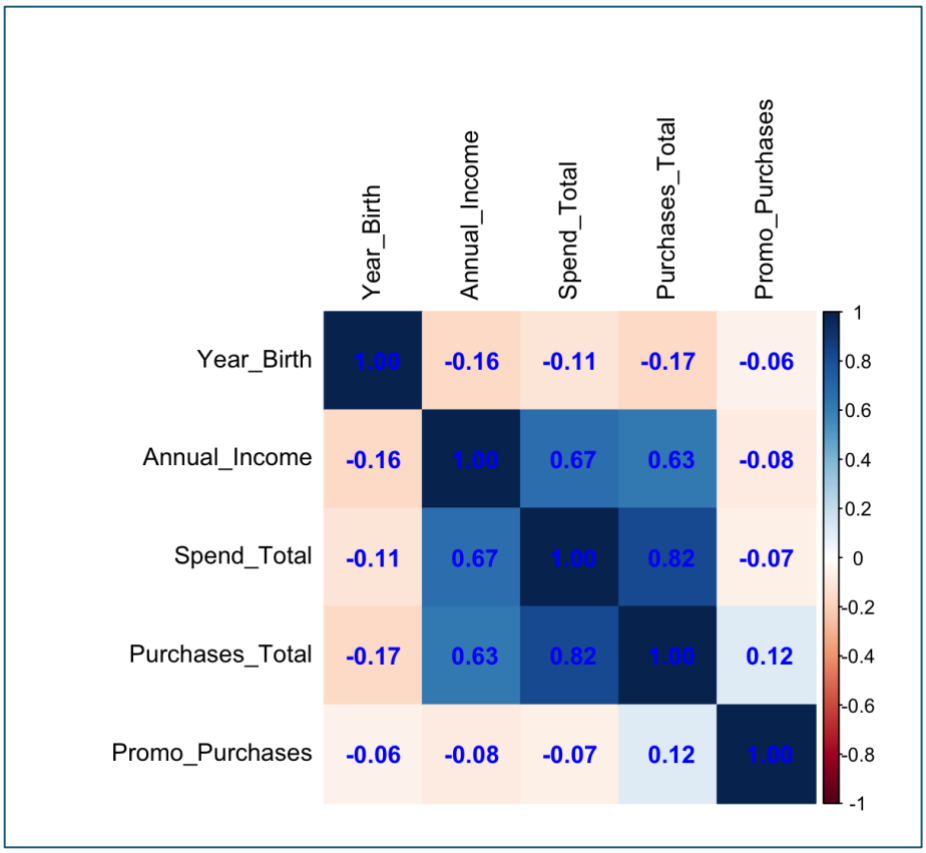


Figure 6: Correlation Heatmap of Numerical Variables

Figure 6 confirms that higher income correlates with increased spending, but luxury product purchases remain low despite affluence. This misalignment may allude to high-income customers who do not perceive luxury offerings as valuable or relevant. Additionally, spending correlations indicate that customers prioritise essential and mid-tier products over high-end purchases.

Principal Component Analysis (PCA)

Capturing trends in customers' purchasing patterns and promotional responsiveness is key to segmentation (Nur and Siregar, 2024). However, it can be complex due to multiple interrelated variables. PCA simplifies this by reducing data dimensions while preserving important insights and robustness (Kalantan *et al.*, 2025).

PCA on Spending Behavior

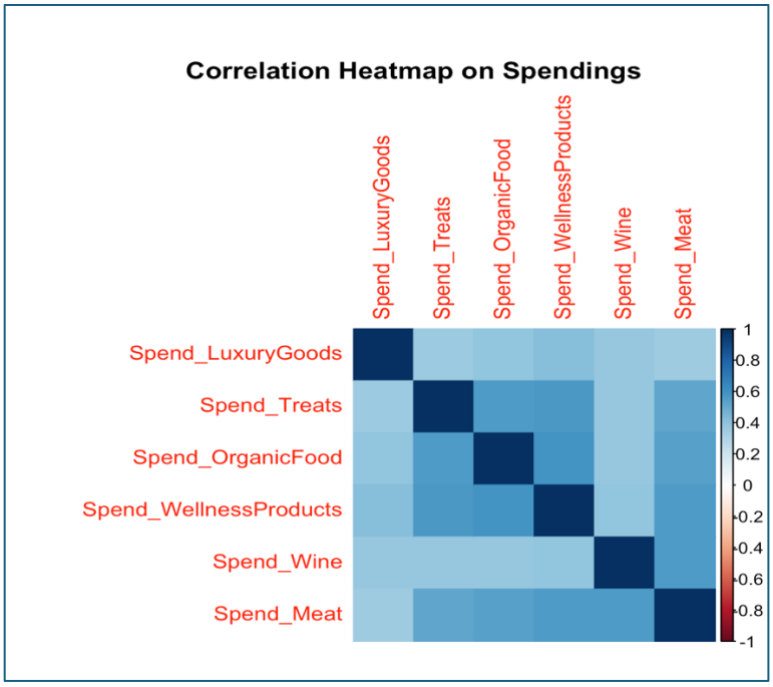
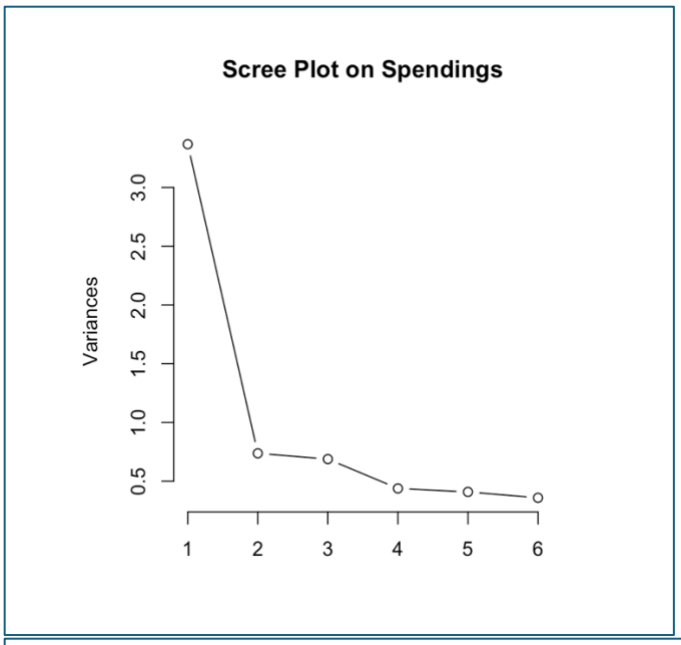


Figure 7: Correlation Heatmap on Spendings

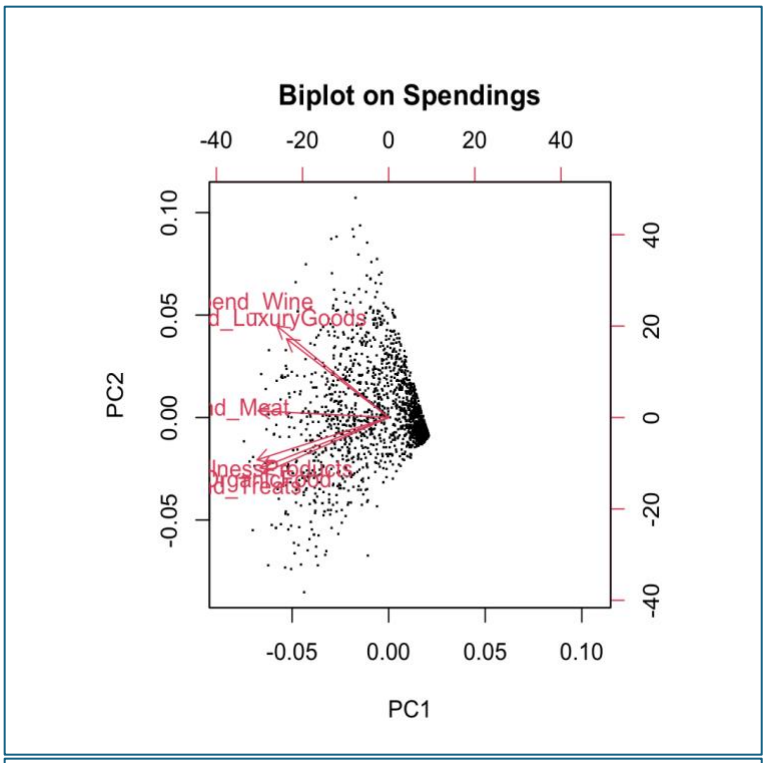
PCA was conducted on spending behaviours across Wine, Meat, and all other items to uncover dominant purchasing patterns by identifying components that explain the highest variance in spending behaviour. Figure 7 highlights strong associations between Organic Food, Wellness Products, and Treats, suggesting shared purchasing behaviour. Luxury Goods spending shows weaker correlations, indicating a separate spending pattern.




```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.8352 0.8587 0.8296 0.6623 0.63931 0.59910
Proportion of Variance 0.5614 0.1229 0.1147 0.0731 0.06812 0.05982
Cumulative Proportion  0.5614 0.6842 0.7990 0.8721 0.94018 1.00000
```

Figure 8: Scree Plot and Importance of Components

Figure 8 shows that three principal components explain 79.9% of the variance, with a steep drop afterward, confirming they optimally summarise the dataset.




```
Loadings:
                         RC1   RC2   RC3
Spend_Wine               0.168 0.910 0.225
Spend_OrganicFood        0.800 0.198 0.180
Spend_Meat               0.569 0.667
Spend_WellnessProducts   0.786 0.221 0.229
Spend_Treats             0.805 0.193 0.135
Spend_LuxuryGoods        0.250 0.184 0.943

                  RC1   RC2   RC3
SS loadings     2.319 1.431 1.044
Proportion Var  0.386 0.239 0.174
Cumulative Var  0.386 0.625 0.799
```

Figure 9: Biplot and Cumulative Variance

A biplot and the rotated component loadings (Figure 9) reveal that RC1 is strongly associated with Organic Food (0.80), Wellness Products (0.79), and Treats (0.81), suggesting a consumer segment that prioritises health-conscious and indulgent purchases. RC2 is driven by Wine (0.91) and Meat (0.67), indicating a preference for traditional grocery essentials with a focus on premium beverages. RC3 is linked to Luxury Goods (0.94), reflecting a distinct consumer segment that prioritises high-end purchases over necessity-based spending.

PCA on Offer Acceptance

Customer responsiveness to promotions varies, impacting SmartFresh Retail's marketing effectiveness. PCA on five past offers and the latest campaign identified engagement patterns.

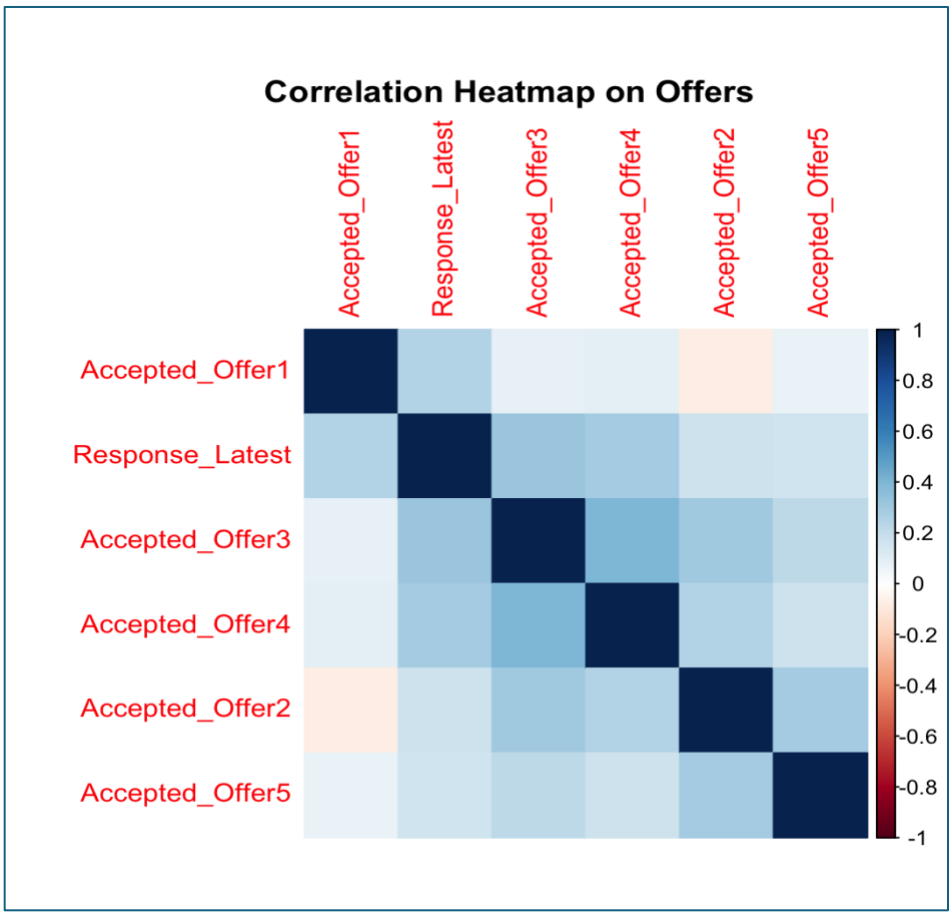
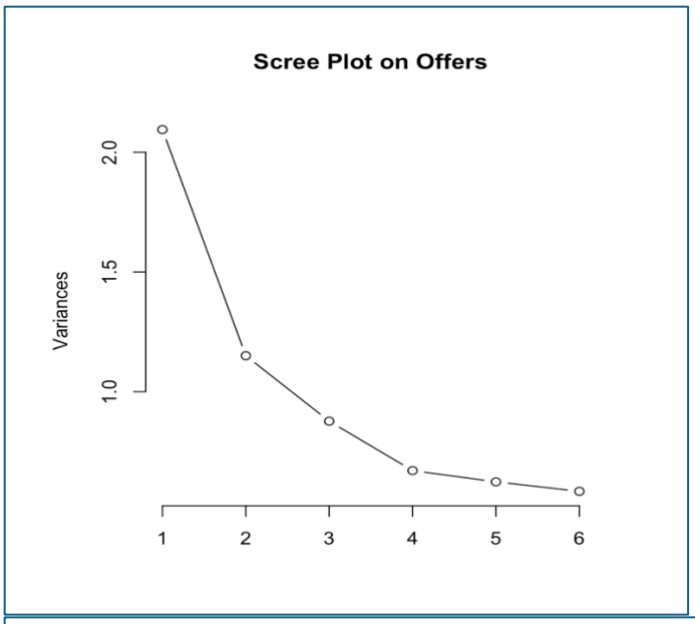


Figure 10: Correlation Heatmap on Offers

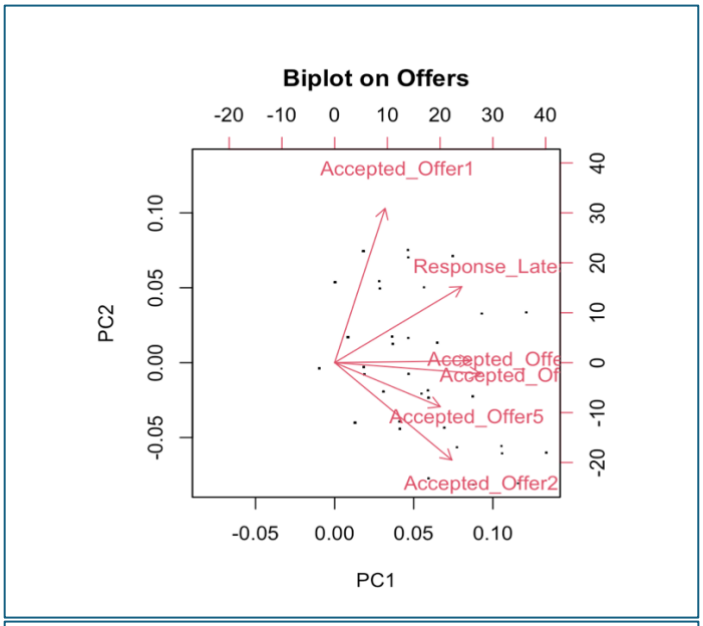
Figure 10 indicates moderate correlations, suggesting selective rather than uniform response to campaigns.




```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.4473 1.0725 0.9365 0.8189 0.7897 0.76414
Proportion of Variance 0.3491 0.1917 0.1462 0.1118 0.1039 0.09732
Cumulative Proportion  0.3491 0.5408 0.6870 0.7987 0.9027 1.00000
```

Figure 11:Scree Plot and Importance of Components

Figure 11 confirms that four components explaining 68.7% of the variance are optimal, as the sharp drop after the third component justifies their selection.




```
Loadings:
                RC1    RC3    RC4    RC2
Accepted_Offer1         0.102  0.205  0.882
Accepted_Offer2  0.181  0.470  0.508 -0.488
Accepted_Offer3  0.703  0.142  0.333
Accepted_Offer4  0.905
Accepted_Offer5  0.121  0.946
Response_Latest  0.211         0.849  0.264

                 RC1   RC3   RC4   RC2
SS loadings    1.407 1.151 1.133 1.101
Proportion Var 0.234 0.192 0.189 0.184
Cumulative Var 0.234 0.426 0.615 0.799
```

Figure 12: Biplot and Variances

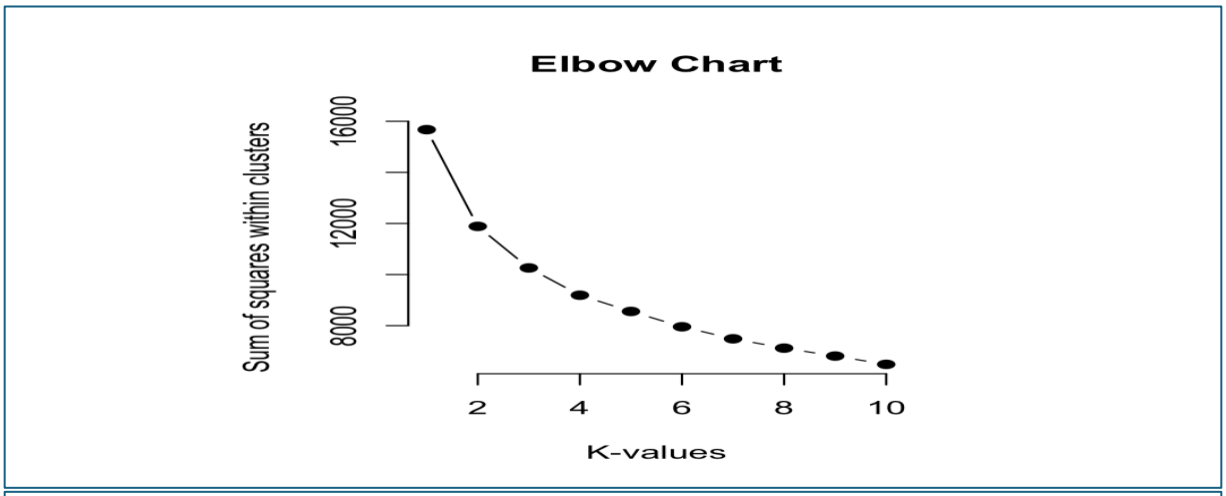
The biplot and the rotated component loadings (Figure 12) reveal RC1 is associated with high engagement in Offers 3 (0.70) and 4 (0.90), indicating a segment of customers who actively respond to promotions with broad appeal. RC2 is linked to Offer 2 (0.50) and the Latest Offer (0.85), reflecting customers may respond based on specific incentives or timing. RC3 is strongly associated with Offer 5 (0.95), suggesting preference for specific exclusive deals. RC4 has the highest loading on Offer 1 (0.88), indicating a segment with sporadic, inconsistent engagement in promotions.

Summary

Customers often purchase related product categories, reflecting overlapping spending behaviours. Similarly, multiple customer segments respond to the same promotions, highlighting the interconnected nature of purchasing tendencies and marketing effectiveness. These findings provide deeper insights into consumer behaviour and strategic targeting.

Cluster Analysis

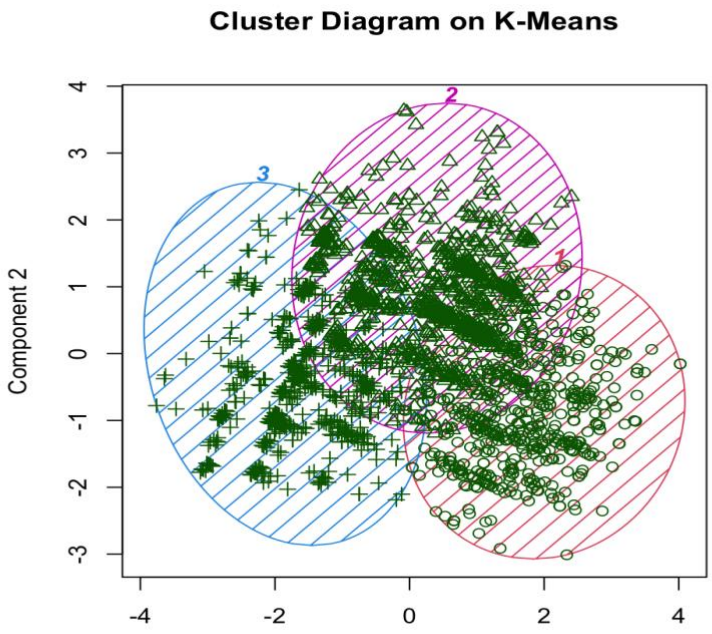
Cluster analysis is a rigorous approach that leverages numerous attributes to create relevant customer segments as demographics will be inferred (Yuan *et al.*, 2020). This is crucial for optimising campaign effectiveness and refining targeted marketing efforts. Incorporating total spending in this analysis enhances differentiation, helping identify customer loyalty levels and enabling more personalised engagement strategies (Thakur, 2016).




```
> WSS
 [1] 15673.000 11886.034 10259.701  9191.387  8556.389  7955.940  7485.403
 [8]  7120.047  6809.681  6485.161
```

Figure 13: Elbow Chart and WSS

After running multiple iterations of K-Means clustering, Figure 13 illustrates a sharp decline until K=3, after which the reduction slows. This suggests that three clusters provide the best balance between complexity and interpretability. The final WSS value is 10,259.99, with a between-cluster sum of squares of 5,413.01, reinforcing the validity of three clusters.




```
> clusters$tot.withinss
[1] 10259.99
> clusters$betweenss
[1] 5413.008
> clusters$size
[1] 572 944 724
> clusters$withinss
[1] 2655.070 4266.224 3338.698
> clusters$centers
  Year_Birth Education_Level Marital_Status Annual_Income    Kidhome   Teenhome
1  0.2237583      0.05405015     0.06535998     0.8918898 -0.7568436 -0.7852133
2  0.2446913      0.21030158    -0.02262662     0.2309815 -0.1718088  0.9709284
3 -0.4958265     -0.31690797    -0.02213588    -1.0058114  0.8219641 -0.6456000
  total_purchase_spent
1            1.2742627
2           -0.1688682
3           -0.7865562
```

Figure 14: Cluster Diagram and withins-betweens

Choosing three clusters (K=3) results in a well-defined segmentation, with Cluster 2 (944 customers) forming a central group in Figure 14, while Clusters 1 (572) and 3 (724) are more distinct. The WSS values indicate tighter clustering in Cluster 1 (2,655) and Cluster 3 (3,338), whereas Cluster 2 (4,266) is more dispersed.

Cluster 1 consists of affluent consumers with the highest income (0.89), fewer household responsibilities (Kidhome: -0.75, Teenhome: -0.78), and moderate education (0.05), skewing younger (Year_Birth: 0.22) with more single or divorced individuals (-0.06). In contrast, Cluster 3 represents lower-income consumers (-1.00), higher household responsibilities (Kidhome: 0.82), lower education (-0.31), and an older demographic (-0.49), with more married or widowed individuals (-0.02). Cluster 2 falls in between, with a balanced income (0.23) and mixed household structures.

Customer Profile	Demographics	Behavioral Factors
Elite-Consumers	High-income, professionals	High spending across all categories, brand loyalty
Economical-Consumers	Middle-income, working-class	Balanced spending, selective offer acceptance
Budget-Consumers	Low-income, price-conscious	Minimal spending, highly price-sensitive

Table 2: Customer Profiles

To gain deeper insights, Table 2 shows that customers were segmented based on Annual Income (demographics) and Total Expenditures (behaviour), forming distinct customer profiles with personalised marketing strategies.

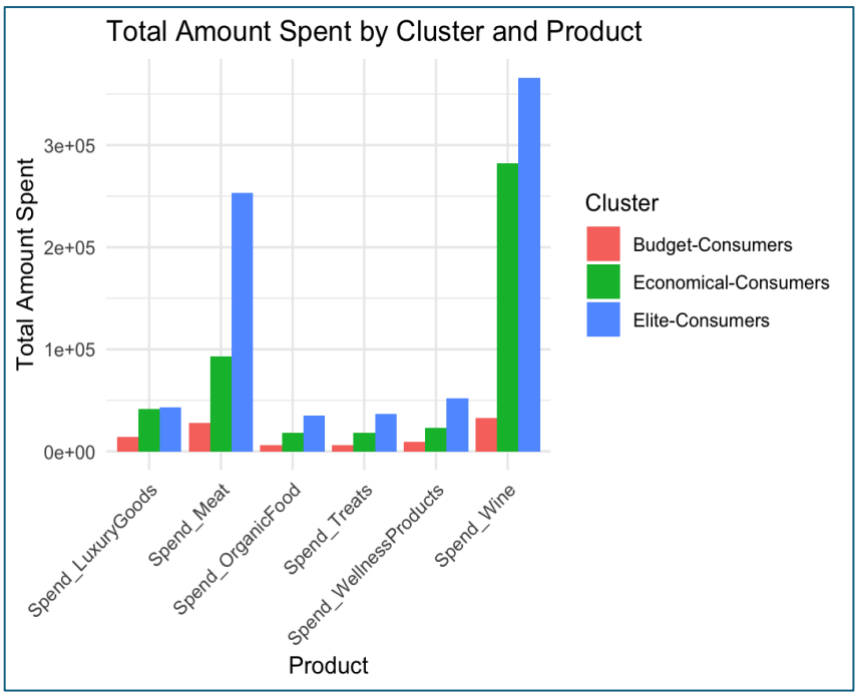


Figure 15: Clustered Bar Plots

Figure 15 further strengthens these segmentations. Elite-Consumers have the highest expenditure

on premium goods, particularly wine and meat, emphasising their affinity for luxury products (Figure 5). Economical-Consumers maintain balanced spending, distributing their purchases more evenly across all product categories other than wine. In contrast, Budget-Consumers focus on essentials like meat and wellness products, reflecting their cost-conscious purchasing decisions.

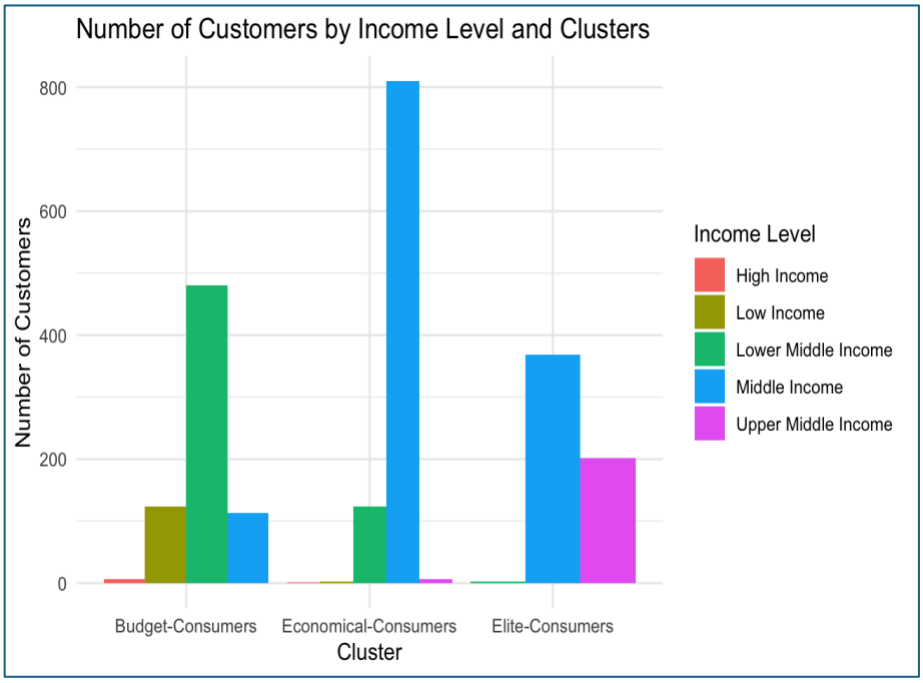


Figure 16: Clustered Bar Plots

Figure 16 supports these distinctions. Elite-Consumers predominantly belong to high-income and upper-middle-income groups, reinforcing their stronger purchasing power. Economical-Consumers are primarily found in middle and lower-middle-income brackets, aligning with their selective spending behaviour. Meanwhile, Budget-Consumers are largely from low-income segments, further confirming their price-sensitive approach to shopping.

Independent Sample T-Test

An independent sample T-test was conducted to examine whether significant differences exist in promotional purchase behaviour across customer clusters. Since segmentation was based on spending and demographic patterns, this test helps validate differences in promotional responsiveness among the groups (Wolf, Sandner and Welpe, 2014). Identifying variations in promotional engagement is critical for understanding consumer behaviour and refining discount-driven marketing strategies (Gonen, Tavor and Spiegel, 2024).

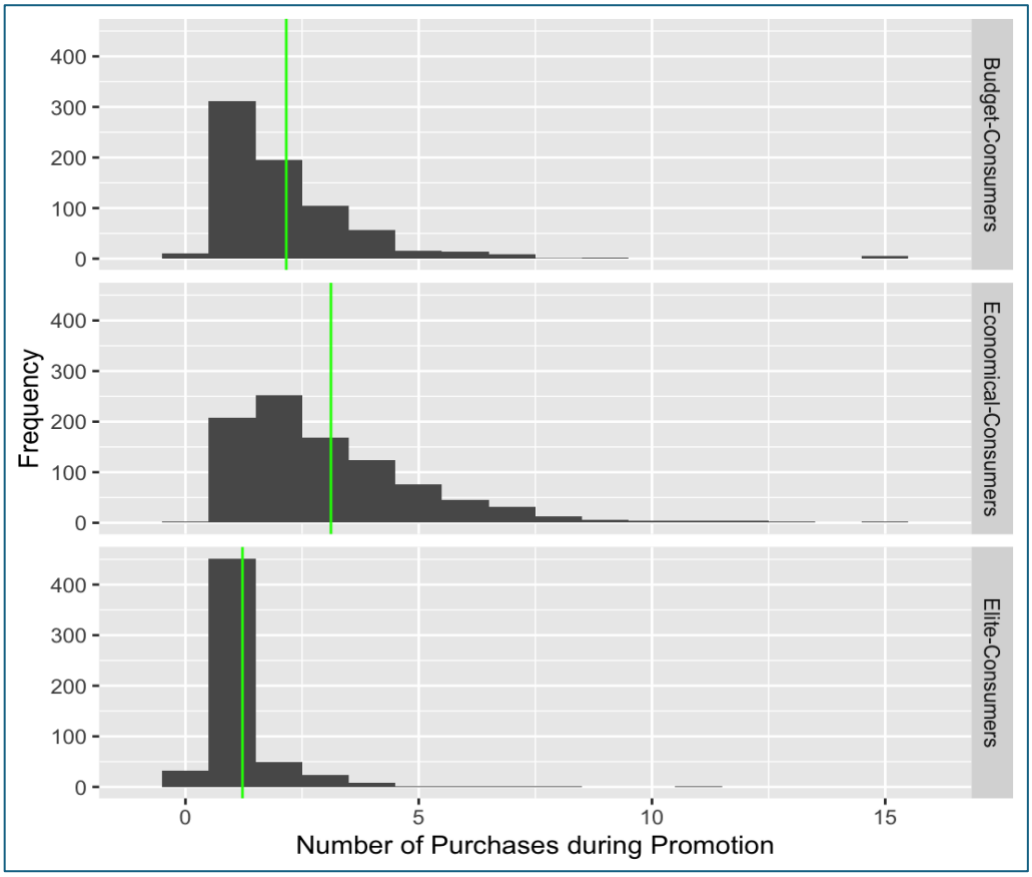


Figure 17: Distributions

Figure 17 shows that Budget-Consumers engage in promotional purchases more frequently than other groups, indicating higher price sensitivity. Economical-Consumers participate in promotions but maintain a balance between discounts and regular spending. In contrast, Elite-Consumers make the fewest promotional purchases, reinforcing their lower sensitivity to discounts and preference for premium-priced products.

Elite-Consumers vs Economical-Consumers

```
        Welch Two Sample t-test

data:  Promo_Purchases by Cluster
t = 23.756, df = 1390.8, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Economical-Consumers and gr
oup Elite-Consumers is not equal to 0
95 percent confidence interval:
 1.740002 2.053230
sample estimates:
mean in group Economical-Consumers      mean in group Elite-Consumers
                          3.118644                           1.222028
```

Figure 18: First test values

To quantify these differences, three independent sample T-tests were conducted. The first test compared Elite-Consumers and Economical-Consumers (Figure 18), revealing a statistically significant difference ($t = 23.76$, $p < 0.001$). The higher mean promotional purchases of Economical-Consumers (3.12) compared to Elite-Consumers (1.22) suggest that discounts are significantly less effective for the latter group.

Budget-Consumers vs Economical Consumers

```
        Welch Two Sample t-test

data:  Promo_Purchases by Cluster
t = -10.024, df = 1660.1, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Budget-Consumers and group
Economical-Consumers is not equal to 0
95 percent confidence interval:
 -1.1443073 -0.7697765
sample estimates:
    mean in group Budget-Consumers mean in group Economical-Consumers
                          2.161602                           3.118644
```

Figure 19: Second test values

The second test compared Budget-Consumers and Economical-Consumers (Figure 19), also showing a significant difference ($t = -10.02$, $p < 0.001$). While Budget-Consumers (2.16) respond to promotions, their engagement is lower than that of Economical-Consumers.

Budget-Consumers vs Elite-Consumers

```
        Welch Two Sample t-test

data:  Promo_Purchases by Cluster
t = 12.434, df = 1141.7, p-value < 2.2e-16
alternative hypothesis: true difference in means between group Budget-Consumers and group
Elite-Consumers is not equal to 0
95 percent confidence interval:
 0.7913093 1.0878392
sample estimates:
mean in group Budget-Consumers  mean in group Elite-Consumers
                      2.161602                       1.222028
```

Figure 20: Third test values

The final test compared Budget-Consumers and Elite-Consumers (Figure 20), with results ($t = 12.43$, $p < 0.001$) confirming that Budget-Consumers engage in significantly more promotional purchases (2.16) than Elite-Consumers (1.22).

The results confirm distinct behavioural patterns across clusters, reinforcing the segmentation model. Economical-Consumers exhibit the highest engagement with promotional offers, Budget-Consumers participate selectively, and Elite-Consumers remain the least responsive.

Recommendations

The segmentation analysis of SmartFresh's customers reveals clear differences in purchasing behavior, promotional engagement, and preferred shopping channels. By aligning marketing strategies with these insights, SmartFresh can optimize revenue and customer retention across all segments (Ali and Shabn, 2024).

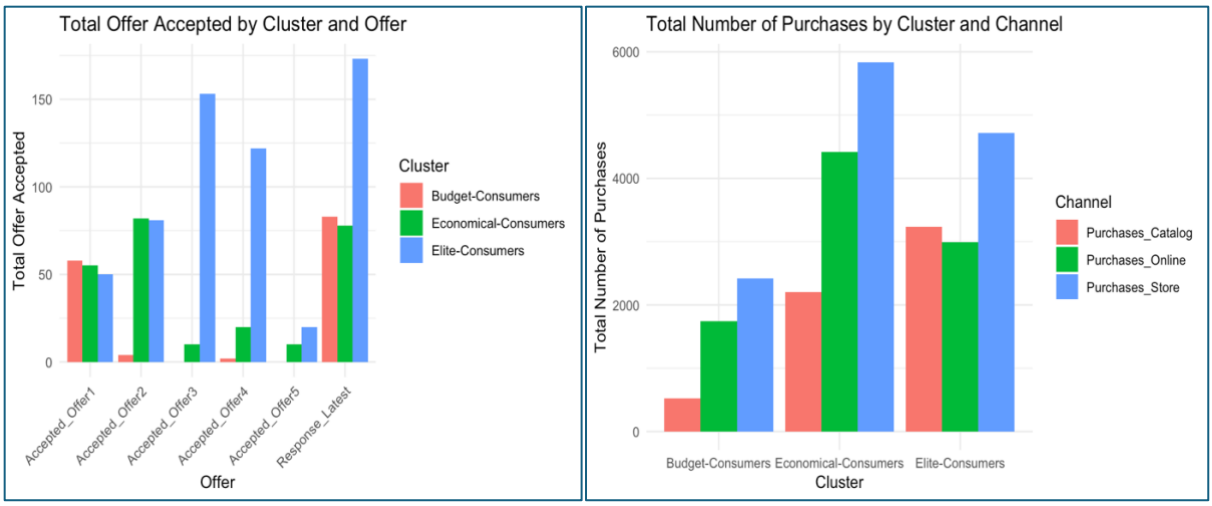


Figure 21: Clustered Bar Plots

Maximising Volume Sales Through Bundling for Budget-Consumers

Budget-Consumers prioritise essential products, particularly during promotions, as indicated by the T-Test. Their price sensitivity suggests that bundled pricing on staples like treats and meat would encourage bulk purchases (Kopalle *et al.*, 2009). Since this segment primarily shops in-store, SmartFresh should leverage limited-time in-store promotions and flash sales to drive immediate conversions (Biricevski, 2025).

Enhancing Offer Personalisation for Economical-Consumers

Economical-Consumers exhibit diverse spending habits and the highest engagement with promotions according to the T-Test results. However, PCA shows they do not spend equally across all categories. SmartFresh should refine promotional campaigns by targeting high-response segments, such as organic food and wellness products, where they show the strongest interest (Dwivedi *et al.*, 2021). Analysing total purchases per offer campaign, promotions should be optimised for peak shopping periods. Furthermore, implementing an omnichannel loyalty system, integrating in-store and digital purchases, will maximise engagement across shopping channels (Muthaffar, Vilches-Montero and Bravo-Olavarria, 2024).

Strengthening Premium Loyalty Among Elite-Consumers

PCA and Cluster Analysis confirm that Elite-Consumers allocate most of their spending to wine and meat, showing low engagement with promotional discounts but a preference for high-end product quality (Mapes, 2020). Given their inclination towards catalogue shopping, SmartFresh should enhance its catalogue offerings by showcasing premium selections and introducing exclusive subscription-based packages for high-value products like meat and wine. Personalised concierge services and invitation-only wine-tasting events would reinforce exclusivity and brand prestige while fostering long-term loyalty (McKinsey & Company, 2020).

Optimising Promotional Timing Across Segments

T-Test results show clear variance in promotional response rates, indicating that not all campaigns drive equal engagement. SmartFresh should analyse conversion patterns from past offers to optimise promotion frequency and timing (Ologunebi and Taiwo, 2024). This ensures that campaigns target the right consumers at the right time, increasing ROI while reducing ineffective discounting.

By implementing data-driven, segment-specific marketing strategies, SmartFresh can enhance customer retention, optimise promotional spending, and drive revenue growth.

Bibliography

Ali, N. and Shabn, O.S. (2024) 'Customer lifetime value (CLV) insights for strategic marketing success and its impact on organizational financial performance', *Cogent Business & Management*, 11(1), p. 2361321. Available at: https://doi.org/10.1080/23311975.2024.2361321.

Biricevski, D. (2025) *Limited-time offers: The good, the bad, and 21 examples*, *Personizely - Website Widgets and Personalization*. Available at: https://www.personizely.net/blog/limited-time-offers (Accessed: 17 March 2025).

Dwivedi, Y.K. *et al.* (2021) 'Setting the future of digital and social media marketing research: Perspectives and research propositions', *International Journal of Information Management*, 59, p. 102168. Available at: https://doi.org/10.1016/j.ijinfomgt.2020.102168.

G, S.P. *et al.* (2024) 'Customer Segmentation using Clustering Techniques: Data-Driven Approach to Enhance Marketing Strategy', in *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*. *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pp. 1127–1134. Available at: https://doi.org/10.1109/ICSCNA63714.2024.10864053.

Gherca, I. (2023) 'Mastering customer segments: Strategies to boost your ROI', *Touchpoint*, 5 July. Available at: https://www.touchpoint.com/blog/customer-segments/ (Accessed: 17 March 2025).

Gonen, L.D., Tavor, T. and Spiegel, U. (2024) *Unlocking Market Potential: Strategic Consumer Segmentation and Dynamic Pricing for Balancing Loyalty and Deal Seeking*. Available at: https://www.mdpi.com/2227-7390/12/21/3364 (Accessed: 17 March 2025).

Kalantan, Z.I. *et al.* (2025) 'Robust Dimensionality Reduction: A Bootstrap-Based Evaluation of PCA with Applications in Nutritional and Environmental Sciences', *Contemporary Mathematics*, pp. 923–942. Available at: https://doi.org/10.37256/cm.6120256016.

Komorowski, M. *et al.* (2016) 'Exploratory Data Analysis', in MIT Critical Data (ed.) *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 185–203. Available at: https://doi.org/10.1007/978-3-319-43742-2_15.

Kopalle, P. *et al.* (2009) 'Retailer Pricing and Competitive Effects', *Journal of Retailing*, 85(1), pp. 56–70. Available at: https://doi.org/10.1016/j.jretai.2008.11.005.

Mapes, G. (2020) 'Marketing elite authenticity: Tradition and terroir in artisanal food discourse', *Discourse, Context & Media*, 34, p. 100328. Available at: https://doi.org/10.1016/j.dcm.2019.100328.

McKinsey & Company (2020) 'Perspectives on retail and consumer goods', Number 8, p. 210.

Muthaffar, A., Vilches-Montero, S. and Bravo-Olavarria, R. (2024) 'From digital touchpoints to digital journeys: How shopping mindsets influence appraisal of omnichannel journeys', *International Journal of Information Management*, 77, p. 102778. Available at: https://doi.org/10.1016/j.ijinfomgt.2024.102778.

Nur, M.F. and Siregar, A. (2024) 'Exploring the Use of Cluster Analysis in Market Segmentation for Targeted Advertising', *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, 5(2), pp. 158–168. Available at: https://doi.org/10.34306/itsdi.v5i2.665.

Ologunebi, J. and Taiwo, E. (2024) *Personalized ad Content and Individual User Preference: A boost for Conversion Rates in the UK E-commerce Business*. Available at: https://mpra.ub.uni-muenchen.de/120595/ (Accessed: 17 March 2025).

Thakur, R. (2016) 'Understanding Customer Engagement and Loyalty: A Case of Mobile Devices for Shopping', *Journal of Retailing and Consumer Services*, 32, pp. 151–163. Available at: https://doi.org/10.1016/j.jretconser.2016.06.004.

Uddin, M.A. *et al.* (2024) 'Data-driven strategies for digital native market segmentation using clustering', *International Journal of Cognitive Computing in Engineering*, 5, pp. 178–191. Available at: https://doi.org/10.1016/j.ijcce.2024.04.002.

Wolf, F., Sandner, P. and Welpe, I.M. (2014) 'Why Do Responses to Age-Based Marketing Stimuli Differ? The Influence of Retirees' Group Identification and Changing Consumption Patterns', *Psychology & Marketing*, 31(10), pp. 914–931. Available at: https://doi.org/10.1002/mar.20743.

Yuan, Y. *et al.* (2020) ‘A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand’, *IEEE Transactions on Power Systems*, 35(5), pp. 4026–4035. Available at: https://doi.org/10.1109/TPWRS.2020.2979943.

Appendix

```
# I will first load tidyverse package to load the pipe operator
install.packages('tidyverse')
library(tidyverse)

# I will now load the dataset given by our professor on Canvas
smartfresh <- read.csv('/Users/jawadzaarif7/Desktop/38159/SmartFresh Retail.csv')

# I will now view at the dataset
summary(smartfresh)
str(smartfresh)

# Since there were some missing values in Annual Income, I will fill them up by grouping
the Education Level
# And taking a mean of it so that my mean values don't vary too much for same Education
Level groups
smartfresh <- smartfresh %>%
  group_by(Education_Level) %>%
  mutate(Annual_Income=
           ifelse(is.na(Annual_Income),
                  mean(Annual_Income, na.rm=TRUE),
                  Annual_Income
  ))

# Since there were some nonsense Marital Status like "Absurd" and "Yolo", I will turn them
all into Single and finally have 5 different Marital Status
smartfresh$Marital_Status <-
  ifelse(smartfresh$Marital_Status %in%
           c('Divorced',
             'Married',
             'Single',
             'Together',
             'Widow'),
         smartfresh$Marital_Status, 'Single')

# I will now turn the Year_Birth into age and create bins to group them into 4 for better
understanding when we go deeper into the analysis
smartfresh <- smartfresh %>%
  mutate(Year_Birth = case_when(
    Year_Birth >= 1990 ~ 'Youngsters',
    Year_Birth >= 1975 ~ 'Adults',
    Year_Birth >= 1955 ~ 'Matured-Adults',
    TRUE ~ 'Seniors'
  ))

# I will now create bins for Annual Income for better understanding further into the
analysis
smartfresh <- smartfresh %>%
  mutate(Annual_Income = case_when(
    Annual_Income <= 20000 ~ 'Low Income',
    Annual_Income <= 40000 ~ 'Lower Middle Income',
    Annual_Income <= 80000 ~ 'Middle Income',
    Annual_Income <= 120000 ~ 'Upper Middle Income',
    TRUE ~ 'High Income'
  ))

# Since the values are in strings, I will turn them into factors or else they would be
treated as individuals instead of groupings further in the analysis
smartfresh$Education_Level <- as.factor(smartfresh$Education_Level)
smartfresh$Marital_Status <- as.factor(smartfresh$Marital_Status)
smartfresh$Year_Birth <- as.factor(smartfresh$Year_Birth)
smartfresh$Annual_Income <- as.factor(smartfresh$Annual_Income)
```

```
# I will first load corrplot package to generate correlation heatmap between the selected
variables.
library(corrplot)
# I will also install psych for performing varimax rotation on the PCA
install.packages("psych")
library(psych)

# I will first scale the dataset so that the variables are in a comparable scale
smartfresh.scaled <- smartfresh
factors.pca <- c('Spend_Wine', 'Spend_OrganicFood', 'Spend_Meat',
'Spend_WellnessProducts', 'Spend_Treats', 'Spend_LuxuryGoods')

# I will now standardise the dataset
smartfresh.scaled[,factors.pca] <- data.frame(scale(smartfresh[,factors.pca]))
summary(smartfresh.scaled)

# Now I will be creating a correlation heatmap to see how strongly are each variables
related before performing PCA
corrplot(cor(smartfresh.scaled[, factors.pca]), method = 'color', order="hclust")
title(main = "Correlation Heatmap on Spendings", line = 2)

# Since the correlation was determined, let's now perform PCA
spending.pca <- prcomp(smartfresh.scaled[,factors.pca], scale=TRUE)

# Let's see the summary of the PCA and see which variables has more variance that's
getting explained
summary(spending.pca)

# Time to create the scree plot
screeplot(spending.pca, type = "lines", main = "Scree Plot on Spendings")

# Time to create the biplot
biplot(spending.pca,
       xlabs = rep(".", nrow(spending.pca$x)),
       main='Biplot on Spendings')

# Since the plots have been created, to get better insights, we will now run the
cummulative variance
cumsum(summary(spending.pca)$importance[2,])
summary(spending.pca)$importance[3,]

# From the insights, I saw that 3 variables were only important for my further
explanations, so I will create a new dataset with only those 3 components
spendingfinal.pca <- spending.pca$x[,1:3]
head(spendingfinal.pca)

# As the new dataset has been created with the 3 components, I will now view the PCA
values
spending.pca$rotation

# Done! Final step would be to run varimax rotation and get the final insights
spendingrotated.pca <- principal(smartfresh.scaled[, factors.pca],
                          nfactors = 3, rotate = "varimax")
spendingrotated.pca$loadings
```

```
# I will first load corrplot package to generate correlation heatmap between the selected
variables.
library(corrplot)
# I will also install psych for performing varimax rotation on the PCA
install.packages("psych")
library(psych)

# I will first scale the dataset so that the variables are in a comparable scale
smartfresh.scaled <- smartfresh
factors.pca <- c('Accepted_Offer1', 'Accepted_Offer2', 'Accepted_Offer3',
'Accepted_Offer4', 'Accepted_Offer5', 'Response_Latest')

# I will now standardise the dataset
smartfresh.scaled[,factors.pca] <- data.frame(scale(smartfresh[,factors.pca]))
summary(smartfresh.scaled)

# Now I will be creating a correlation heatmap to see how strongly are each variables
related before performing PCA
corrplot(cor(smartfresh.scaled[, factors.pca]), method = 'color', order="hclust")
title(main = "Correlation Heatmap on Offers", line = 2)

# Since the correlation was determined, let's now perform PCA
spending.pca <- prcomp(smartfresh.scaled[,factors.pca], scale=TRUE)

# Let's see the summary of the PCA and see which variables has more variance that's
getting explained
summary(spending.pca)

# Time to create the scree plot
screeplot(spending.pca, type = "lines", main = "Scree Plot on Offers")

# Time to create the biplot
biplot(spending.pca,
       xlabs = rep(".", nrow(spending.pca$x)),
       main='Biplot on Offers')

# Since the plots have been created, to get better insights, we will now run the
cummulative variance
cumsum(summary(spending.pca)$importance[2,])
summary(spending.pca)$importance[3,]

# From the insights, I saw that 3 variables were only important for my further
explanations, so I will create a new dataset with only those 3 components
spendingfinal.pca <- spending.pca$x[,1:4]
head(spendingfinal.pca)

# As the new dataset has been created with the 4 components, I will now view the PCA
values
spending.pca$rotation

# Done! Final step would be to run varimax rotation and get the final insights
spendingrotated.pca <- principal(smartfresh.scaled[, factors.pca],
                          nfactors = 4, rotate = "varimax")
spendingrotated.pca$loadings
```

```
# First I will introduce a new variable where I will take the total amount spent on all
the products for interesting analysis
smartfresh
smartfresh$total_purchase_spent <- rowSums(smartfresh[, c("Spend_Wine",
"Spend_OrganicFood", "Spend_Meat", "Spend_WellnessProducts", "Spend_Treats",
"Spend_LuxuryGoods")], na.rm = TRUE)

# Now I will be turning the features that are in string into integers.
smartfresh$Year_Birth <- as.integer(smartfresh$Year_Birth)
smartfresh$Education_Level <- as.integer(smartfresh$Education_Level)
smartfresh$Marital_Status <- as.integer(smartfresh$Marital_Status)
smartfresh$Annual_Income <- as.integer(smartfresh$Annual_Income)

# These will be my final variables for clustering
variables_clus <- c('Year_Birth', 'Education_Level', 'Marital_Status', 'Annual_Income',
'Kidhome', 'Teenhome', 'total_purchase_spent')
smartfresh.clus <- smartfresh[variables_clus]

# I will now standardise the dataset
clus.scaled <- as.data.frame(lapply(smartfresh.clus,scale))
summary(clus.scaled)

# Now I will use a randomiser to train my dataset
set.seed(2500)

# I have decided to go with 10 possible iterations for clustering
k.max <- 10

# Since there will be different values of k now, I will have to run multiple iterations
wss <- sapply(1:k.max,
              function(k){kmeans(clus.scaled,
                                 k, nstart=50,
                                 iter.max = 15 )$tot.withinss})

# To determine the optimal k value, I will be generating an elbow chart now
wss
plot(1:k.max, wss, type="b", pch = 19, frame = FALSE, xlab="K-values", ylab="Sum of
squares within clusters", main = "Elbow Chart")

# As my elbow chart showe 3 as the optimal value of k, I will run it for k=3
clusters <- kmeans(clus.scaled, 3)
clusters

# Time to look at all the measures in clustering
clusters
clusters$tot.withinss
clusters$betweenss
clusters$size
clusters$withinss
clusters$centers

# The measures looked good to me, I will proceed with generating the cluster chart
library(cluster)
clusplot(clus.scaled, clusters$cluster, color = TRUE, shade = TRUE, labels = 4, lines = 0,
main = "Cluster Diagram on K-Means")

# I will now create cluster columns and give them cluster values so that I can connect
these cluster values at the very end of my discussions
smartfresh_cluster <- smartfresh
smartfresh_cluster$Cluster <- clusters$cluster
```

```
# I will first load tidyverse package to load the pipe operator
library(tidyverse)

# I will first factor Annual Income and Year Birth as they are currently in strings
smartfresh_cluster <- smartfresh_cluster %>%
  mutate(Year_Birth = recode(Year_Birth, `1` = "Adults", `2` = "Matured-Adults", `3` =
"Seniors", `4` = "Youngsters"))
smartfresh_cluster <- smartfresh_cluster %>%
  mutate(Annual_Income = recode(Annual_Income, `1` = "High Income", `2` = "Low Income",`3`
= "Lower Middle Income", `4` = "Middle Income", `5` = "Upper Middle Income "))
smartfresh_cluster <- smartfresh_cluster %>%
  mutate(Cluster = recode(Cluster, `1` = "Elite-Consumers", `2` = "Economical-Consumers",
`3` = "Budget-Consumers"))

# I will now bring total spending based on products and clusters for interpretations
product.clus <- smartfresh_cluster %>%
  pivot_longer(cols = c(Spend_Wine, Spend_OrganicFood, Spend_Meat, Spend_WellnessProducts,
Spend_Treats, Spend_LuxuryGoods),
               names_to = "Product",
               values_to = "Spending") %>%
  group_by(Cluster, Product) %>%
  summarise(Total_Spending = sum(Spending), .groups = "drop")

# Let's view the bar chart now
ggplot(product.clus,
       aes(x = Product, y = Total_Spending, fill = as.factor(Cluster))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total Amount Spent by Cluster and Product",
       x = "Product",
       y = "Total Amount Spent",
       fill = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# I will now bring total offer accepted based on offers and clusters for interpretations
offer.clus <- smartfresh_cluster %>%
  pivot_longer(cols = c(Accepted_Offer1, Accepted_Offer2, Accepted_Offer3,
Accepted_Offer4, Accepted_Offer5, Response_Latest),
               names_to = "Offer",
               values_to = "Acceptance") %>%
  group_by(Cluster, Offer) %>%
  summarise(Total_Acceptances = sum(Acceptance), .groups = "drop")

# Let's view the bar chart now
ggplot(offer.clus,
       aes(x = Offer, y = Total_Acceptances, fill = as.factor(Cluster))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total Offer Accepted by Cluster and Offer",
       x = "Offer",
       y = "Total Offer Accepted",
       fill = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# I will now bring total purchases based on channels and clusters for interpretations
channel.clus <- smartfresh_cluster %>%
  pivot_longer(cols = c(Purchases_Online, Purchases_Catalog, Purchases_Store),
               names_to = "Channel",
               values_to = "Purchase") %>%
  group_by(Cluster, Channel) %>%
  summarise(Total_Purchases = sum(Purchase), .groups = "drop")

# Let's view the bar chart now
```

```
ggplot(channel.clus,
       aes(x = as.factor(Cluster), y = Total_Purchases, fill = Channel)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total Number of Purchases by Cluster and Channel",
       x = "Cluster",
       y = "Total Number of Purchases",
       fill = "Channel") +
  theme_minimal()

# I will now bring customer counts based on age group and clusters for interpretations
age.clus <- smartfresh_cluster %>%
  group_by(Cluster, Year_Birth) %>%
  summarise(Customer_Count = n(), .groups = "drop")

# Let's view the bar chart now
ggplot(age.clus,
       aes(x = as.factor(Cluster), y = Customer_Count,
           fill = as.factor(Year_Birth))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Customers by Age Group and Cluster",
       x = "Cluster",
       y = "Number of Customers",
       fill = "Age Group") +
  theme_minimal()

# I will now bring customer counts based on income level and clusters for interpretations
income.clus <- smartfresh_cluster %>%
  group_by(Cluster, Annual_Income) %>%
  summarise(Customer_Count = n(), .groups = "drop")

# Let's view the bar chart now
ggplot(income.clus,
       aes(x = as.factor(Cluster), y = Customer_Count,
           fill = as.factor(Annual_Income))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Customers by Income Level and Clusters",
       x = "Cluster",
       y = "Number of Customers",
       fill = "Income Level") +
  theme_minimal()
```

```r
# Load Packages
library(tidyverse)
library(lubridate)
options(width=100)

# Labels
# 1-"Elite-Consumers", 2-"Economical-Consumers", 3-"Budget-Consumers"

# Mean of Promo Purchases by Clusters
plot_test <- smartfresh_cluster
plot_test_test <- plot_test %>%
  group_by(Cluster) %>%
  summarise(mean=mean(Promo_Purchases), n=n())
plot_test_test

# Plot the Promo Purchases Distribution
ggplot(plot_test) +
  geom_histogram(aes(x=Promo_Purchases), binwidth=1) +
  facet_grid(Cluster~.) + xlim(-1,16) +
  labs(x="Number of Purchases during Promotion",
       y="Frequency") +
  geom_vline(data=plot_test_test,
             mapping=aes(xintercept=mean),
             col="green")

# Filter the Dataset (Cluster 1 and 2)
main.data.cluster1n2 <- smartfresh_cluster %>%
  filter(Cluster %in% c("Elite-Consumers", "Economical-Consumers"))

# Mean of Promo Purchases by Clusters
promo.by_cluster1n2 <- main.data.cluster1n2 %>%
  group_by(Cluster) %>%
  summarise(mean=mean(Promo_Purchases), n=n())
promo.by_cluster1n2

# Summary of Mean Differences
promo.by_cluster1n2 %>% summarise(difference=diff(mean))

# Plot the Cluster-wise Promo Purchases Distribution
ggplot(main.data.cluster1n2) +
  geom_histogram(aes(x=Promo_Purchases), binwidth=1) +
  facet_grid(Cluster~.) + xlim(-1,16) +
  labs(x="Number of Purchases during Promotion",
       y="Frequency") +
  geom_vline(data=promo.by_cluster1n2,
             mapping=aes(xintercept=mean),
             col="green")

# Performing T-test
t.test(Promo_Purchases~Cluster, data=main.data.cluster1n2)

# Filter the Dataset (Cluster 3 and 2)
main.data.cluster3n2 <- smartfresh_cluster %>%
  filter(Cluster %in% c("Budget-Consumers", "Economical-Consumers"))

# Mean of Promo Purchases by Clusters
promo.by_cluster3n2 <- main.data.cluster3n2 %>%
  group_by(Cluster) %>%
  summarise(mean=mean(Promo_Purchases), n=n())
promo.by_cluster3n2

# Summary of Mean Differences
promo.by_cluster3n2 %>% summarise(difference=diff(mean))
```

```
# Plot the Cluster-wise Promo Purchases Distribution
ggplot(main.data.cluster3n2) +
  geom_histogram(aes(x=Promo_Purchases), binwidth=1) +
  facet_grid(Cluster~.) + xlim(-1,16) +
  labs(x="Number of Purchases during Promotion",
       y="Frequency") +
  geom_vline(data=promo.by_cluster3n2,
             mapping=aes(xintercept=mean),
             col="green")

# Performing T-test
t.test(Promo_Purchases~Cluster, data=main.data.cluster3n2)

# Filter the Dataset (Cluster 3 and 1)
main.data.cluster3n1 <- smartfresh_cluster %>%
  filter(Cluster %in% c("Budget-Consumers", "Elite-Consumers"))

# Mean of Promo Purchases by Clusters
promo.by_cluster3n1 <- main.data.cluster3n1 %>%
  group_by(Cluster) %>%
  summarise(mean=mean(Promo_Purchases), n=n())
promo.by_cluster3n1

# Summary of Mean Differences
promo.by_cluster3n1 %>% summarise(difference=diff(mean))

# Plot the Cluster-wise Promo Purchases Distribution
ggplot(main.data.cluster3n1) +
  geom_histogram(aes(x=Promo_Purchases), binwidth=1) +
  facet_grid(Cluster~.) + xlim(-1,16) +
  labs(x="Number of Purchases during Promotions",
       y="Frequency") +
  geom_vline(data=promo.by_cluster3n1,
             mapping=aes(xintercept=mean),
             col="green")

# Performing T-test
t.test(Promo_Purchases~Cluster, data=main.data.cluster3n1)
```

```
# Load Packages
library(tidyverse)
library(corrplot)
library(ggplot2)
library(dplyr)
library(tidyr)

# Create a Dataset for Visualization
main.eda <- read_csv('/Users/jawadzaarif7/Desktop/38159/SmartFresh Retail.csv')
main.eda$Spend_Total <- rowSums(main.eda[, c("Spend_Wine", "Spend_OrganicFood",
"Spend_Meat", "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")], na.rm =
TRUE)
main.eda$Purchases_Total <- rowSums(main.eda[, c("Purchases_Online", "Purchases_Catalog",
"Purchases_Store" )], na.rm = TRUE)
main.eda$Education_Level <- as.factor(main.eda$Education_Level)
main.eda$Marital_Status <- as.factor(main.eda$Marital_Status)

# Plot the Year_Birth Box Plot
boxplot(main.eda$Year_Birth, xlab="Year of Birth", main="Boxplot: Year of Birth",
horizontal=TRUE)

# Plot the Annual_Income Box Plot
boxplot(main.eda$Annual_Income, xlab="Annual Income", main="Boxplot: Annual Income",
horizontal=TRUE)

# Plot the Spend_Wine Box Plot
boxplot(main.eda$Spend_Wine, xlab="Amount Spent on Wine", main="Boxplot: Amount Spent on
Wine", horizontal=TRUE)

# Plot the Spend_OrganicFood Box Plot
boxplot(main.eda$Spend_OrganicFood, xlab="Amount Spent on Organic Food", main="Boxplot:
Amount Spent on Organic Food", horizontal=TRUE)

# Plot the Spend_Meat Box Plot
boxplot(main.eda$Spend_Meat, xlab="Amount Spent on Meat", main="Boxplot: Amount Spent on
Meat", horizontal=TRUE)

# Plot the Spend_WellnessProducts Box Plot
boxplot(main.eda$Spend_WellnessProducts, xlab="Amount Spent of Wellness Product",
main="Boxplot: Amount Spent of Wellness Product", horizontal=TRUE)

# Plot the Spend_Treats Box Plot
boxplot(main.eda$Spend_Treats, xlab="Amount Spent of Treats", main="Boxplot: Amount Spent
on Treats", horizontal=TRUE)

# Plot the Spend_LuxuryGoods Box Plot
boxplot(main.eda$Spend_LuxuryGoods, xlab="Amount Spent on Luxury Goods", main="Boxplot:
Amount Spent on Luxury Goods", horizontal=TRUE)

# Plot the Year of Births Histogram
hist(main.eda$Year_Birth,
     main="Histogram: Year of Birth",
     xlab="Years",
     ylab="Frequency",
     breaks=30,
     col="lightgreen")

# Plot the Total Annual Income Histogram
hist(main.eda$Annual_Income,
     main="Histogram: Annual Income",
     xlab="Annual Income",
     ylab="Frequency",
     breaks=30,
```

```
     col="lightgreen")

# Plot the Total Spending Histogram
hist(main.eda$Spend_Total,
     main="Histogram: Total Amount Spent on Purchasing",
     xlab="Total Amount Purchased",
     ylab="Frequency",
     breaks=30,
     col="lightgreen")

# Plot the Total Purchases Histogram
hist(main.eda$Purchases_Total,
     main="Histogram: Total Number of Purchases",
     xlab="Number of Purchases",
     ylab="Frequency",
     breaks=30,
     col="lightgreen")

# Pair Plot
filtered_data <- main.eda %>%
  filter(!Customer_ID %in% c('7829','1150','11004','9432',
                             '8475','1503','5555','1501',
                             '5336','4931','11181'))
pairs(formula = ~Year_Birth + Annual_Income + Spend_Total + Purchases_Total +
Promo_Purchases, data=filtered_data, cex=0.5)

# Plot the Correlation Heatmap
cor_matrix <- cor(main.eda[, c('Year_Birth', 'Annual_Income', 'Spend_Total',
'Purchases_Total','Promo_Purchases')],
  use = "complete.obs")
corrplot(cor_matrix,
         method = 'color',
         addCoef.col = "blue",
         tl.col = "black"
         )

# Plot the Education Level Bar Plot
ggplot(main.eda, aes(x = Education_Level)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Level of Education",
       x = "Level of Education",
       y = "Frequency") +
  theme_minimal()

# Plot the Marital Status Bar Plot
ggplot(main.eda, aes(x = Marital_Status)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Marital Status",
       x = "Marital Status",
       y = "Frequency") +
  theme_minimal()

# Plot the Kid Home Bar Plot
ggplot(main.eda, aes(x = Kidhome)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Number of Kids at home",
       x = "Number of Kids",
       y = "Frequency") +
  theme_minimal()

# Plot the Teen Home Bar Plot
ggplot(main.eda, aes(x = Teenhome)) +
  geom_bar(fill = "yellow") +
  labs(title = "Bar Plot: Number of Teens at home",
```

```
       x = "Number of Teens",
       y = "Frequency") +
  theme_minimal()

# Plot the Offers Bar Plot
offer_summary <- main.eda %>%
  summarise(across(c('Accepted_Offer1', 'Accepted_Offer2', 'Accepted_Offer3',
'Accepted_Offer4', 'Accepted_Offer5', 'Response_Latest'),
                   sum, na.rm = TRUE)) %>%
  pivot_longer(cols = everything(),
               names_to = "Offers",
               values_to = "Total")
ggplot(offer_summary, aes(x = Offers, y = Total)) +
  geom_bar(stat = "identity", fill = 'yellow') +
  labs(title = "Number of Offers Accepted",
       x = "Offers",
       y = "Frequency",
       ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Plot the Purchases Bar Plot
purchase_summary <- main.eda %>%
  summarise(across(c('Purchases_Online', 'Purchases_Store', 'Purchases_Catalog'),
                   sum, na.rm = TRUE)) %>%
  pivot_longer(cols = everything(),
               names_to = "Offers",
               values_to = "Total")
ggplot(purchase_summary, aes(x = Offers, y = Total)) +
  geom_bar(stat = "identity", fill = 'yellow') +
  labs(title = "Number of Purchases by Channels",
       x = "Channels",
       y = "Number of Purchases",
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
# Load necessary library
library(tidyverse)

# Load dataset
main.eda <- read_csv('/Users/jawadzaarif7/Desktop/38159/SmartFresh Retail.csv')

# Create new calculated columns
main.eda$Age <- 2025 - main.eda$Year_Birth
main.eda$Spend_Total <- rowSums(main.eda[, c("Spend_Wine", "Spend_OrganicFood",
"Spend_Meat",
                                             "Spend_WellnessProducts", "Spend_Treats",
"Spend_LuxuryGoods")],
                                na.rm = TRUE)
main.eda$Purchases_Total <- rowSums(main.eda[, c("Purchases_Online", "Purchases_Catalog",
"Purchases_Store")],
                                    na.rm = TRUE)

# Select key variables for central tendency and variance
variables <- c("Age", "Annual_Income", "Spend_Wine", "Spend_OrganicFood", "Spend_Meat",
               "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods",
"Spend_Total",
               "Purchases_Online", "Purchases_Catalog", "Purchases_Store",
"Purchases_Total",
               "Promo_Purchases", "Response_Latest")

# Function to compute mean, variance, and standard deviation while handling NA values
calculate_summary <- function(x) {
  data.frame(
    Mean = mean(x, na.rm = TRUE),
    Variance = var(x, na.rm = TRUE),
    Std_Dev = sd(x, na.rm = TRUE)
  )
}

# Apply function to selected variables
summary_table <- main.eda %>%
  summarise(across(all_of(variables), calculate_summary))

# Transpose the table for better readability
summary_table <- summary_table %>%
  pivot_longer(everything(), names_to = c("Variable", ".value"), names_sep = "_")

# Print the table
print(summary_table)
```