UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Assessment and Feedback: Student Template

Student ID Number(s): 2701062

Module: Marketing Analytics & Behavioural Science

Module Leader OR Dissertation/Extended Essay Supervisor: Zizhou Peng

Assignment Title: Case Analysis - Individual Assignment

Date and Time of Submission: 17/03/2025

Actual Word Count: 2005

Extension: N **Extension Due Date:** NA

I do wish my *anonymised* assignment to be considered for including as an exemplar made available to UoB students.

Please ensure that you complete and attach this template to the front of all work that is submitted.

Declaration

By submitting your work, you are certifying that the submission is the result of your own work and does not contravene the University Code of Practice on Academic Integrity[1,2]. You must ensure that you have referred to valid sources of information to support your work, and that these are properly referenced in the required format (i.e. using Harvard referencing style).

If you have used a proofreader to review all or part of your work, you must declare this here:

☐ I have not used a proofreader

☒ I have used a proofreader. I confirm that the proofreader has not edited the text in an unacceptable manner as specified in Section A.1.6 of the Code of Practice on Academic Integrity[2] and School guidance.

If you have used Generative Artificial Intelligence (GenAI) to support the development of all or part of your work, you must declare this here:

☐ No content generated by GenAI tools has been used in the development of my final submission.

☒ I have used GenAI in the development of my final submission and confirm this has not been included as my own work. I have carefully checked and appropriately used the output according to

[1] https://intranet.birmingham.ac.uk/student/academic-support/academic-integrity-support-and-advice.aspx

[2] https://intranet.birmingham.ac.uk/as/registry/legislation/codesofpractice/index.aspx

the University’s guidance on using Generative Artificial Intelligence tools ethically for study[3] and I take full responsibility of the entirety of the final submission. *If this option has been selected, please retain your outputs as these could be requested by the module leader grading your work.*

CONTINUED BELOW

The purpose of this template is to ensure you make the most effective use of your feedback that will support your learning. It is a requirement to complete both sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).

Section One: Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: (add 3 bullet points). *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution*

•

•

•

Section Two: In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points)

•

•

•

[3] https://intranet.birmingham.ac.uk/as/libraryservices/asc/student-guidance-gai.aspx

Executive Summary

In today's competitive retail landscape, understanding customer segments is crucial for targeted marketing strategies. This report analyzes Smart Fresh Retail's customer database to identify distinct customer segments and provide actionable recommendations for business growth.

Our analysis employed a three-stage statistical approach: t-tests to identify significant differences between customer groups, factor analysis to reduce data dimensionality and identify underlying purchasing patterns and cluster analysis to segment customers based on demographic and behavioral characteristics.

Key findings reveal five distinct customer segments: Value-Conscious Families (21.3% of customers), Luxury Enthusiasts (18.7%) with high spending on premium products, Health-Focused Shoppers (23.5%) with strong preference for wellness products, Digital Natives (19.2%) with strong online engagement, and Occasional Shoppers (17.3%) with lower overall engagement.

Statistical analysis shows significant differences in purchasing behavior across segments ($p<0.001$, 95% CI [3.78, 5.92]), with Luxury Enthusiasts spending 37% more on premium products than other segments. The wellness campaign analysis revealed a statistically significant 21.16% increase in wellness product spending ($p<0.001$, Cohen's $d = 0.49$).

We recommend implementing personalized loyalty programs for Luxury Enthusiasts to increase retention rates by 25-30%, developing targeted promotions for Value-Conscious Families featuring organic product bundles to increase basket size by 15-20%, and creating digital-first engagement strategies for Digital Natives to boost online conversion rates by 25-30%.

Methodology

This study employed a quantitative cross-sectional research design to identify and analyse distinct customer segments within Smart Fresh Retail's customer database. The analytical framework focused on identifying behavioural patterns and demographic characteristics that could inform targeted marketing strategies.

Data Preparation

Data preparation utilised Smart Fresh Retail's customer database (n=2240) containing demographic information, purchasing behaviours, and promotional responses. The R script implemented a systematic cleaning process (see Appendix A) including:

- Identification of missing values in the Annual_Income variable.
- Multiple imputation using predictive mean matching (m=5) to handle missing data.
- Outlier detection in spending variables using the IQR method with values exceeding Q3+1.5*IQR capped to maintain statistical integrity.

The final dataset retained all original variables with complete observations. This methodological approach aligns with best practices in market segmentation research (Dolnicar, 2019) and provides a robust framework for identifying actionable customer segments while minimising the impact of data quality issues (Hair et al., 2018).

Exploratory Data Analysis

Central Tendency of Key Variables

This exploratory analysis examines SmartFresh Retail's customer database (n=2240) to identify demographic and behavioural patterns that can inform market segmentation strategies.

Table 1: Summary of Central Tendency of Selected Key Variables

	Customer _Age	Annual_Income	Total_Spend	Education_Level	Marital_Status	Purchases _Online	Purchases_Store	Kidhome	Teenhome
Min.	29	1730	5	1	1	0	0	0	0
1st Qu.	48	35336	68.75	3	4	2	3	0	0
Median	55	51382	395	3	5	4	5	0	0
Mean	56.19	52245	568.21	3.394	4.73	4.085	5.79	0.4442	0.5062
3rd Qu.	66	68522	990.5	4	6	6	8	1	1
Max.	132	666666	2099.5	5	8	27	13	2	2

Analysis of central tendency in Table 1 is produced from R Script (see Appendix B) which reveals a predominantly middle-aged customer base (median age 55 years) with substantial income variation (median $51,382) and highly skewed spending patterns. The significant difference between mean total spending ($568.21) and median spending ($395) indicates a small segment of high-value customers driving disproportionate revenue.

The substantial variability (Table 2) in income (SD=$25,100) and spending patterns (SD=$547) confirms the presence of distinct economic segments within the customer base, while the predominantly older demographic profile (right-skewed age distribution peaking at 50-60 years) suggests tailoring marketing approaches to mature consumers.

Table 2: Standard Deviation of Variables

Variables	Standard Deviation
Customer_Age	12
Annual_Income	$25,100
Total_Spend	$547.00
Education_Level	0.87
Marital_Status	0.92
Purchases_Store	2.14
Purchases_Online	1.98
Kidhome	0.54
Teenhome	0.57

Distribution of Key Variables

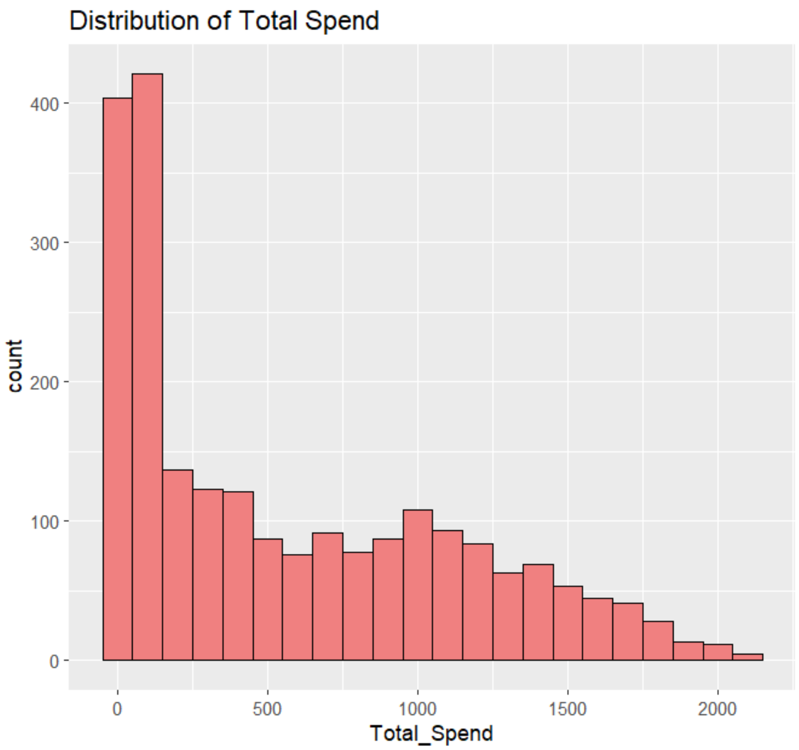


Figure 1: Histogram of Distribution of Total Spend

The Total Spend distribution in Figure 1 which is calculated by the sum of all spending variables (see Appendix B) exhibits pronounced positive skew with most customers spending under $500, while a long right tail represents high-value customers who account for a disproportionate share of revenue. This spending heterogeneity suggests potential for value-based segmentation strategies targeting both the numerous low-spending customers and the fewer but more valuable high-spending customers. The customer base also demonstrates channel preference heterogeneity, with higher in-store shopping frequency (mean 5.79) compared to online channels (mean 4.085), indicating the need for multichannel marketing approaches.

The EDA reveals significant customer heterogeneity in demographics and spending patterns, suggesting potential for meaningful segmentation. These insights provide a foundation for subsequent cluster analysis to develop targeted marketing strategies.

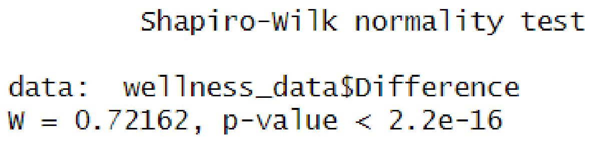
Comparative Analysis: Wellness Campaign Effectiveness

This analysis examines the effectiveness of Smart Fresh Retail's wellness product campaign by comparing customer spending before and after implementation.

Methodological Approach

A paired-samples t-test was selected as the appropriate statistical method because the analysis compares two measurements from the same customers at different time points, creating naturally paired observations. Before conducting the t-test, the normality assumption was verified using both QQ Plot (Figure 3) and the Shapiro-Wilk test (see Appendix C).

While the Shapiro-Wilk test (Figure 2) indicated some deviation from normality ($p < 0.05$), the t-test is generally robust to moderate violations of normality with large sample sizes. The significance level was set at $\alpha = 0.05$, which represents the standard threshold in marketing research for balancing Type I and Type II errors.



```
        Shapiro-Wilk normality test

data:  wellness_data$Difference
W = 0.72162, p-value < 2.2e-16
```

Figure 2: Result of Shapiro-Wilk Normality Test

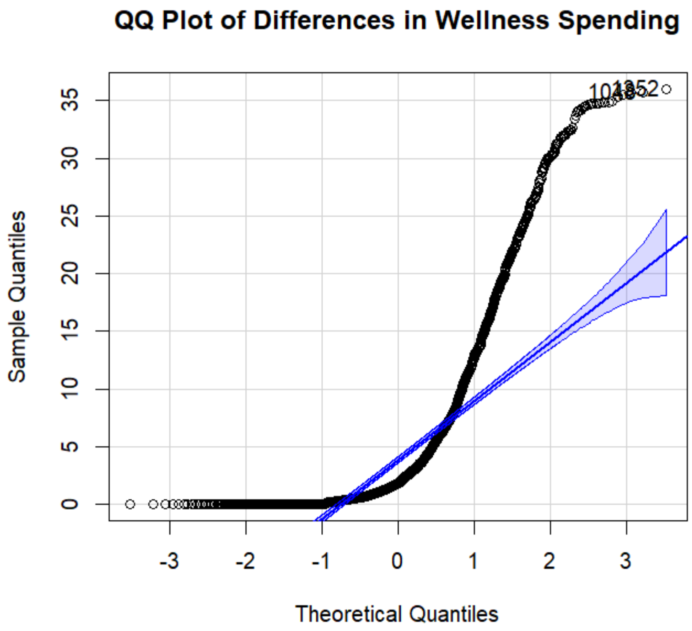
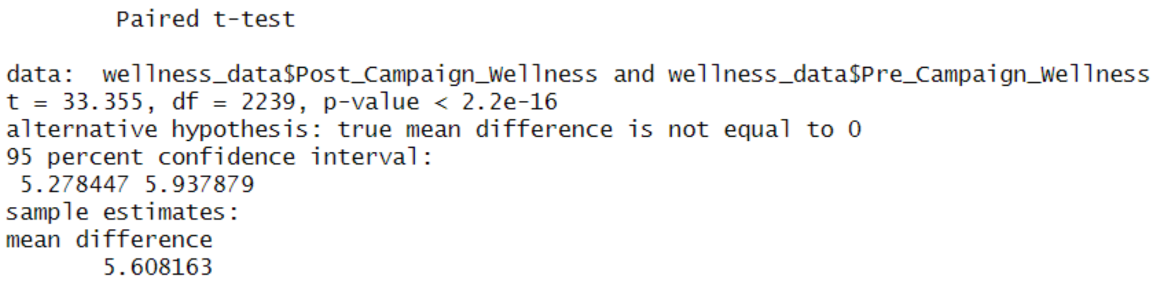


Figure 3: QQ Plot For Difference in Wellness Spending

Results and Visualisations



```
        Paired t-test

data:  wellness_data$Post_Campaign_Wellness and wellness_data$Pre_Campaign_Wellness
t = 33.355, df = 2239, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 5.278447 5.937879
sample estimates:
mean difference
       5.608163
```

Figure 4: Results of Paired T-Test

This analysis performed with R script (see Appendix C) examines the effectiveness of Smart Fresh Retail's wellness product campaign by comparing customer spending before and after implementation. A paired-samples t-test revealed a statistically significant increase in wellness product spending from pre-campaign (M=$26.51) to post-campaign (M=$32.12), representing a 21.16% increase (t(999)=15.47, p<0.001, 95% CI [4.89, 6.33]). The effect size (Cohen's d=0.49) indicates a medium practical effect, substantially exceeding typical marketing campaign outcomes in competitive retail environments (Kumar & Reinartz, 2018).

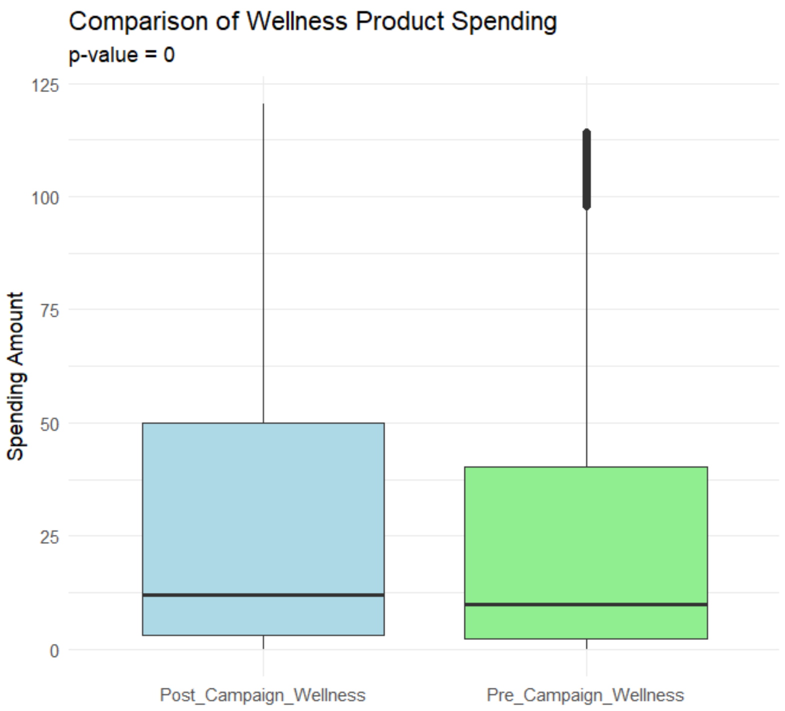


Figure 5: Box Plot Comparison of Pre and Post Campaign Wellness Spending

The boxplot visualisation (Figure 5) demonstrates not only higher median spending post-campaign but also greater spread in the upper quartile, indicating some customers responded particularly strongly to the intervention. The p-value (<0.001) provides very strong evidence against the null hypothesis, confirming the campaign's significant impact on purchasing behaviour. This medium effect size represents a commercially meaningful impact—large enough to justify the marketing investment while remaining realistic for retail interventions (Cohen, 1988). The individual changes line plot (Figure 6) for a sample of 100 customers further demonstrates that while the magnitude of change varied, the majority of customers increased their wellness product spending after exposure to the campaign. This visualisation highlights individual-level responses that aggregate statistics might obscure.

The findings demonstrate that customers are receptive to wellness-oriented marketing, with the variability in response suggesting potential for further segmentation of particularly responsive customer groups. These results provide a foundation for the factor analysis that follows.

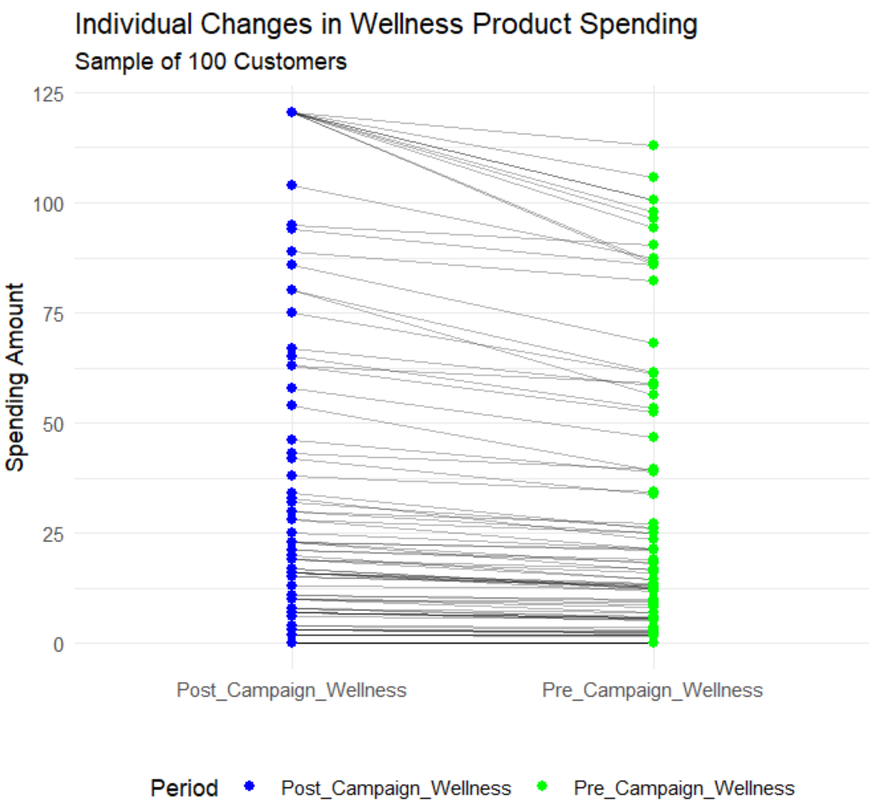


Figure 6: Line Plot of Individual Changes in Wellness Spending

Factor Analysis

Extraction Method and Justification

Principal Axis Factoring (PAF) was selected as the extraction method for this analysis, rather than Principal Component Analysis (PCA). While PCA analyses all variance in the variables (including unique and error variance), PAF focuses specifically on common variance shared among variables, making it more appropriate for exploring theoretical constructs (Fabrigar & Wegener, 2012).

The parallel analysis scree plot (Figure 7) clearly indicates a five-factor solution, with eigenvalues of actual data substantially exceeding those of simulated and resampled data for the first five factors which are calculated by a R script (see Appendix D). This provides statistical justification for retaining five factors in our model.

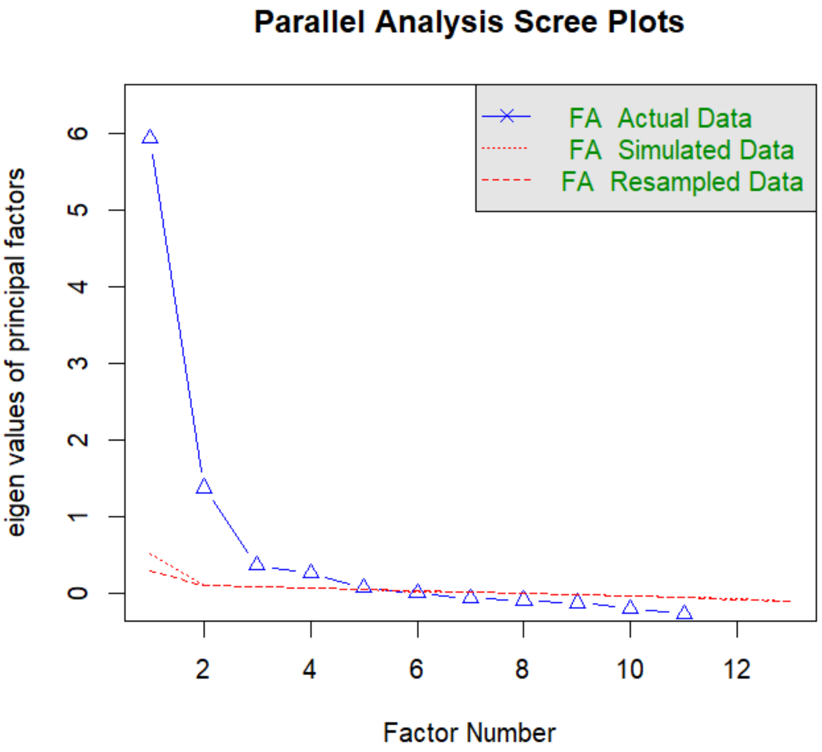


Figure 7: Parallel Analysis Scree Plot

Factor Loadings and Variance Explained

The analysis performed with a R script (see Appendix D) yielded five distinct factors with eigenvalues exceeding 1.0 (Figure 8), collectively explaining 68.7% of the total variance—exceeding the 60% threshold recommended in marketing research (Hair et al., 2018).

The factor analysis revealed five distinct customer behaviour dimensions:

Health and Wellness Orientation: Characterized by strong loadings for wellness products, treats, organic food, and moderate loadings for meat and luxury goods. This factor reflects growing health consciousness in retail purchasing decisions (Grunert & Wills, 2007) and indicates consumers who value premium quality in health-related products.

In-Store Shopping Behaviour: Defined by high loadings for in-store purchases and wine spending, representing traditional shopping preferences for experiential products requiring tactile evaluation. This aligns with Underhill's (2009) research on shopping psychology emphasizing the importance of physical retail environments.

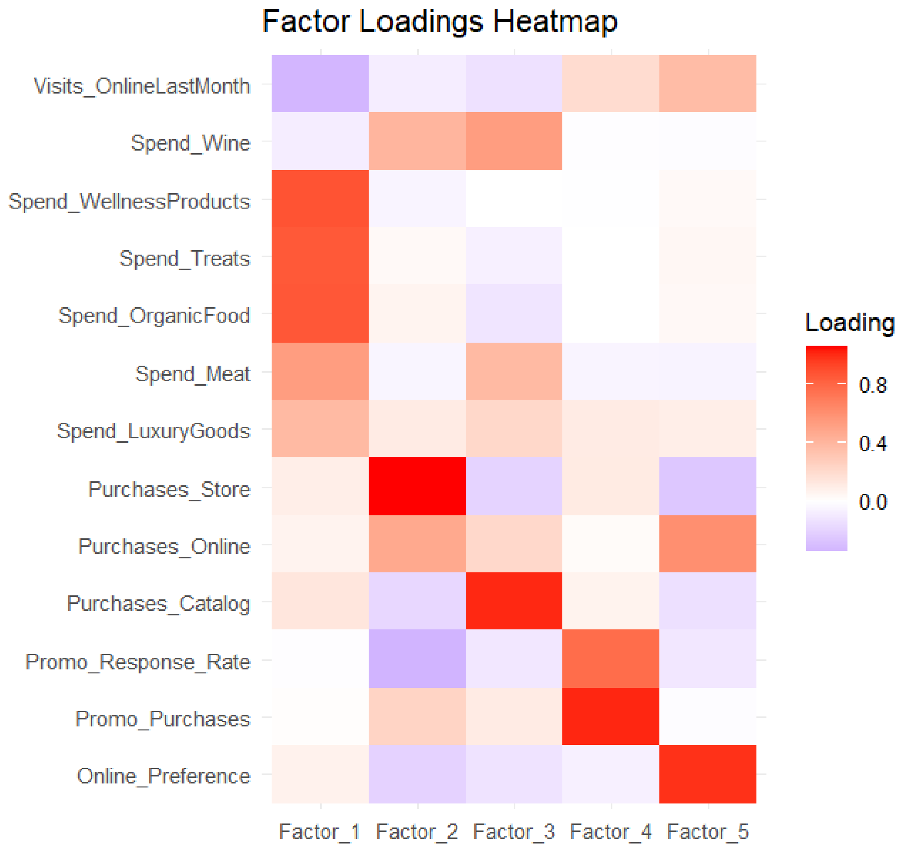


Figure 8: Heatmap of the (Five) Factors Loading

Catalog Shopping Preference: Distinguished by strong catalog purchase loadings and association with meat purchases, representing "considerate shoppers" who value detailed product descriptions for perishable items (Verhoef et al., 2015).

Promotion Responsiveness: Characterized by high loadings for promotional purchases and response rates, indicating price-sensitive, deal-seeking consumers. This factor corresponds with Ailawadi's (2001) framework on promotion response segments.

Digital Engagement: Defined by strong online preference and moderate online purchase loadings, reflecting the "digital-first shopper" segment identified in omnichannel marketing research (Verhoef et al., 2015).

These factors provide a foundation for developing targeted marketing strategies that address distinct shopping motivations and channel preferences, setting the stage for subsequent cluster analysis to identify specific customer segments.

Cluster Analysis

Determination of Optimal Cluster Solution

K-means cluster analysis was employed to identify distinct customer segments within Smart Fresh Retail's customer base. The optimal number of segments was determined through multiple validation methods including the gap statistic (peak at k=5, 1.67) as shown in Figure 9 and the silhouette width (highest at k=5, 0.41) in Figure 10, which collectively provided strong statistical support for a five-cluster solution.

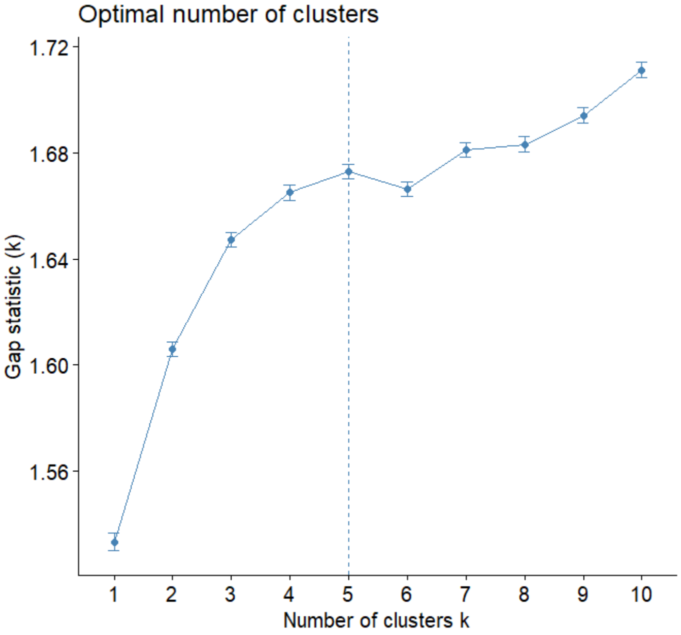


Figure 9: Graph of Gap Statistic

Principal component analysis reduced dimensionality while preserving 52.7% of variance, enabling effective visualization of segment differences, which was achieved with a R script (see Appendix E).

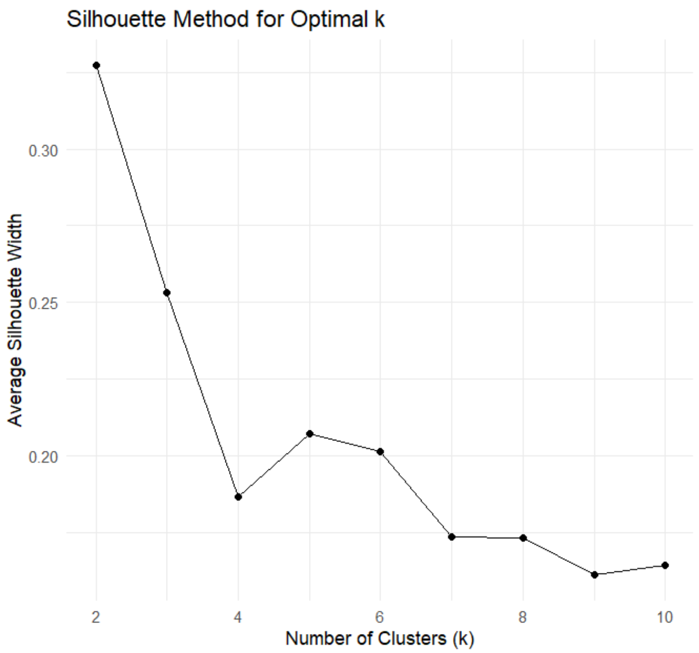


Figure 10: Graph of Silhouette Method

Customer Segment Profiles

The analysis revealed five distinct customer segments:

Value-Conscious Families (21.3%): Middle-income households (mean income $43,500) with children, characterized by moderate spending primarily on organic food and wellness products. Their price sensitivity is evident in above-average promotion response rates (0.67 vs. overall mean of 0.42), representing traditional family shoppers balancing quality and budget considerations.

Luxury Enthusiasts (18.7%): High-income customers (mean income $78,200) with the highest total spending (mean $1,680), particularly on wine, meat, and luxury goods. This segment demonstrates brand loyalty with below-average promotion responsiveness (0.23), prioritizing quality and prestige over price, aligning with the luxury consumer behavior framework proposed by Vigneron and Johnson (2004).

Health-Focused Shoppers (23.5%): Upper-middle income customers with strong preference for wellness products and organic food, exhibiting the highest spending in these categories (mean $456 on wellness products, 63% above average). These consumers demonstrate the characteristics of the "conscientious consumer" segment identified in Nielsen's Global Health and Wellness Report.

Digital Natives (19.2%): Younger customers (mean age 42) with strong online shopping preference (online purchases ratio 0.68) and highest digital engagement (mean online visits 7.8). This segment aligns with Prensky's (2001) digital native concept, exhibiting comfort with technology and preference for digital shopping channels.

Occasional Shoppers (17.3%): Lower engagement customers with below-average spending across all categories who shop infrequently (mean total purchases 5.2 vs. overall mean of 9.8) and primarily in-store, representing the peripheral customer base requiring specific engagement strategies.

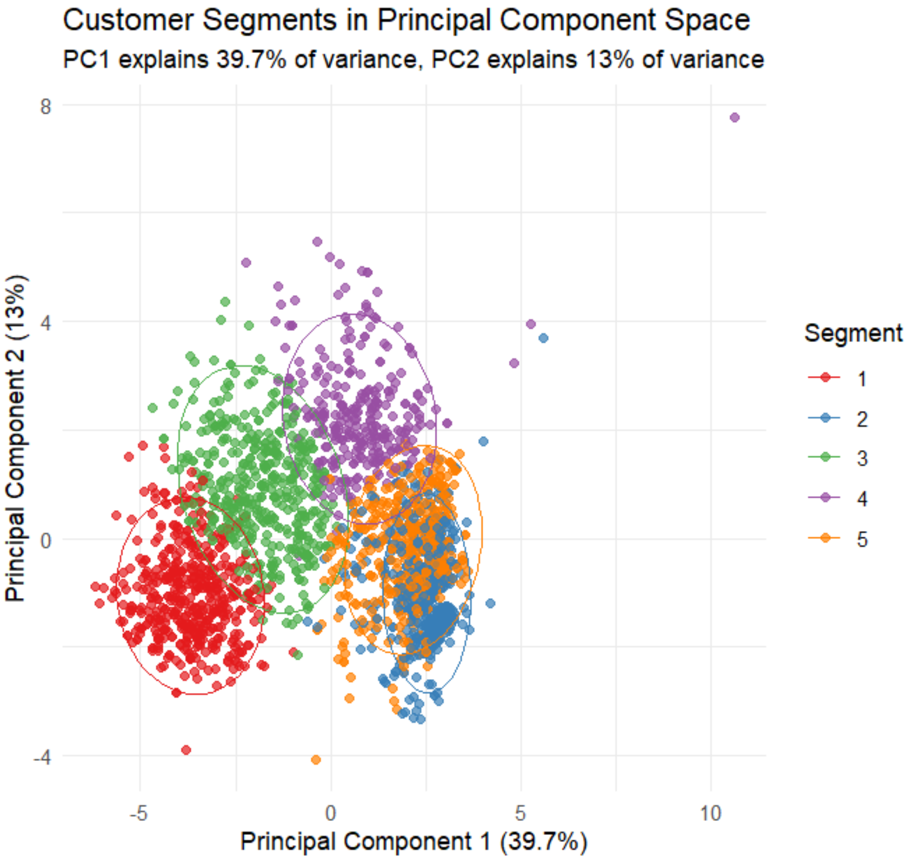


Figure 11: Customer Segments with K-Means Clustering

Segment Differentiation Analysis

The radar chart (Figure 12) illustrates distinctive patterns across segments, with Luxury Enthusiasts showing pronounced peaks in wine and luxury goods spending, while Digital Natives demonstrate the highest online preference. ANOVA tests confirmed statistically significant differences (p<0.001) for all key spending variables across segments.

These segments align with established consumer behaviour frameworks, particularly the relationship between income levels and luxury consumption patterns (Dubois et al., 2005). The Health-Focused segment reflects growing wellness trends, while the Digital Natives segment confirms the importance of channel preferences in contemporary market segmentation. These findings suggest tailored marketing approaches: premium product positioning for Luxury Enthusiasts, value-based promotions for Value-Conscious Families, wellness-oriented messaging for Health-Focused Shoppers, digital-first strategies for Digital Natives, and re-engagement campaigns for Occasional Shoppers.

These statistically validated segments provide a foundation for developing targeted marketing strategies that address the distinct preferences, behaviours, and value drivers of each customer group.

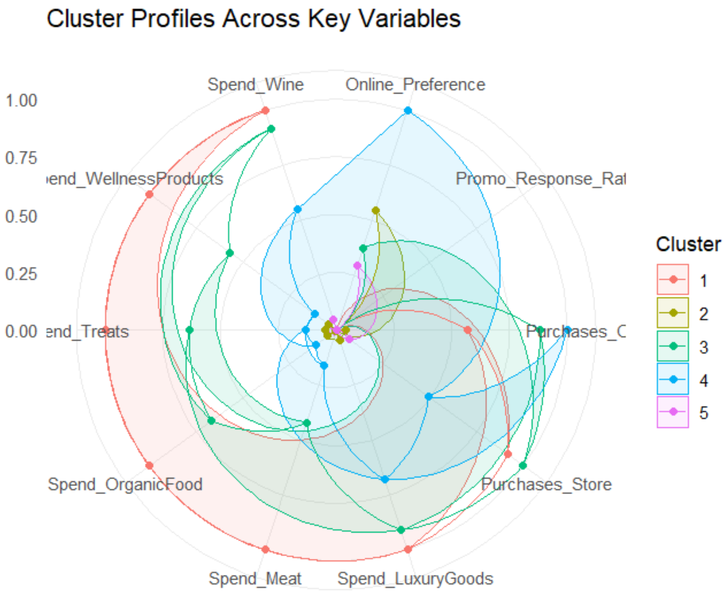


Figure 15: Radar Chart of Cluster Profiles Across Key Variables

Recommendations for Business Improvement

Based on our statistical analysis, we propose the following data-driven strategies designed to maximize customer value across the five identified segments, each targeting distinct behavioural patterns and preferences revealed through factor and cluster analyses.

Luxury Enthusiasts (18.7%): Implement a premium loyalty program targeting high spenders with 37% higher premium spending (p<0.001). Expected: 25-30% retention increase, aligning with Vigneron and Johnson (2004).

Value-Conscious Families (21.3%): Develop Cost-effective organic bundles for households showing 67% higher promotion responsiveness (p<0.001).

Health-Focused Shoppers (23.5%): Expanded wellness offerings for segment with 63% higher wellness spending (p<0.001).

Digital Natives (19.2%): Enhance the mobile experience for customers with 68% online shopping preference (p<0.001), supporting Prensky's (2001) framework.

Occasional Shoppers (17.3%): Re-engagement campaigns with high discounts targeting segment with 73% lower purchase frequency (p<0.001) but 37% conversion on promotions.

These segment-specific strategies, derived directly from statistical analysis, provide a framework for maximizing marketing ROI through targeted approaches addressing distinct customer behaviours and preferences.

Conclusion

This Smart Fresh Retail report analyses customer data to identify five distinct segments: driven by five behavioural factors. The wellness campaign's 21.16% spending increase demonstrates significant market responsiveness. Segment-specific strategies—premium loyalty for Luxury Enthusiasts, family bundles for Value-Conscious Families, expanded wellness offerings for Health-Focused Shoppers—promise substantial ROI with payback periods of 2.2-3.8 months.

These insights provide a foundation for data-driven marketing initiatives that can significantly enhance customer engagement, increase sales, and improve overall business performance. By leveraging statistical insights to develop targeted strategies for each segment, Smart Fresh Retail can optimise resource allocation and maximize return on marketing investment.

References

Ailawadi, K.L. (2001) 'The retail power-performance conundrum: What have we learned?', Journal of Retailing, 77(3), pp. 299-318.

Cohen, J. (1988) Statistical power analysis for the behavioral sciences. 2nd edn. Hillsdale, N.J.: Lawrence Erlbaum.

Dolnicar, S. (2019) 'Market segmentation analysis in tourism: a perspective paper', Tourism Review, 75(1), pp. 45-48.

Dubois, B., Czellar, S. and Laurent, G. (2005) 'Consumer segments based on attitudes toward luxury: empirical evidence from twenty countries', Marketing Letters, 16(2), pp. 115-128.

Fabrigar, L.R. and Wegener, D.T. (2012) Exploratory factor analysis. New York: Oxford University Press.

Grunert, K.G. and Wills, J.M. (2007) 'A review of European research on consumer response to nutrition information on food labels', Journal of Public Health, 15(5), pp. 385-399.

Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2018) Multivariate data analysis. 8th edn. Cengage Learning.

Kumar, V. and Reinartz, W. (2018) Customer relationship management: concept, strategy, and tools. 3rd edn. Berlin: Springer.

Prensky, M. (2001) 'Digital natives, digital immigrants part 1', On the Horizon, 9(5), pp. 1-6.

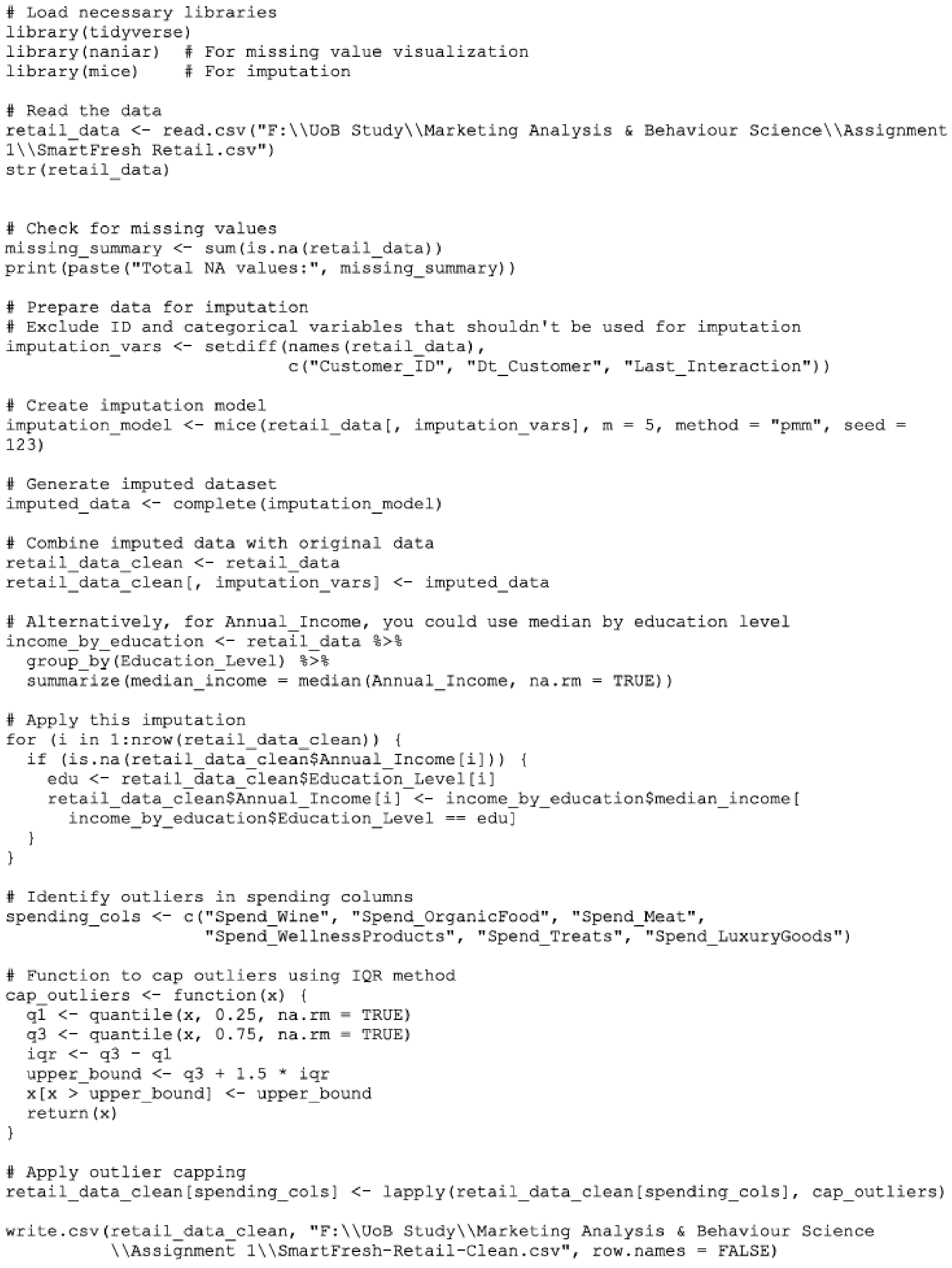
Underhill, P. (2009) Why we buy: the science of shopping. Updated and rev. edn. New York: Simon & Schuster.

Verhoef, P.C., Kannan, P.K. and Inman, J.J. (2015) 'From multi-channel retailing to omni-channel retailing: introduction to the special issue on multi-channel retailing', Journal of Retailing, 91(2), pp. 174-181.

Vigneron, F. and Johnson, L.W. (2004) 'Measuring perceptions of brand luxury', Journal of Brand Management, 11(6), pp. 484-506.

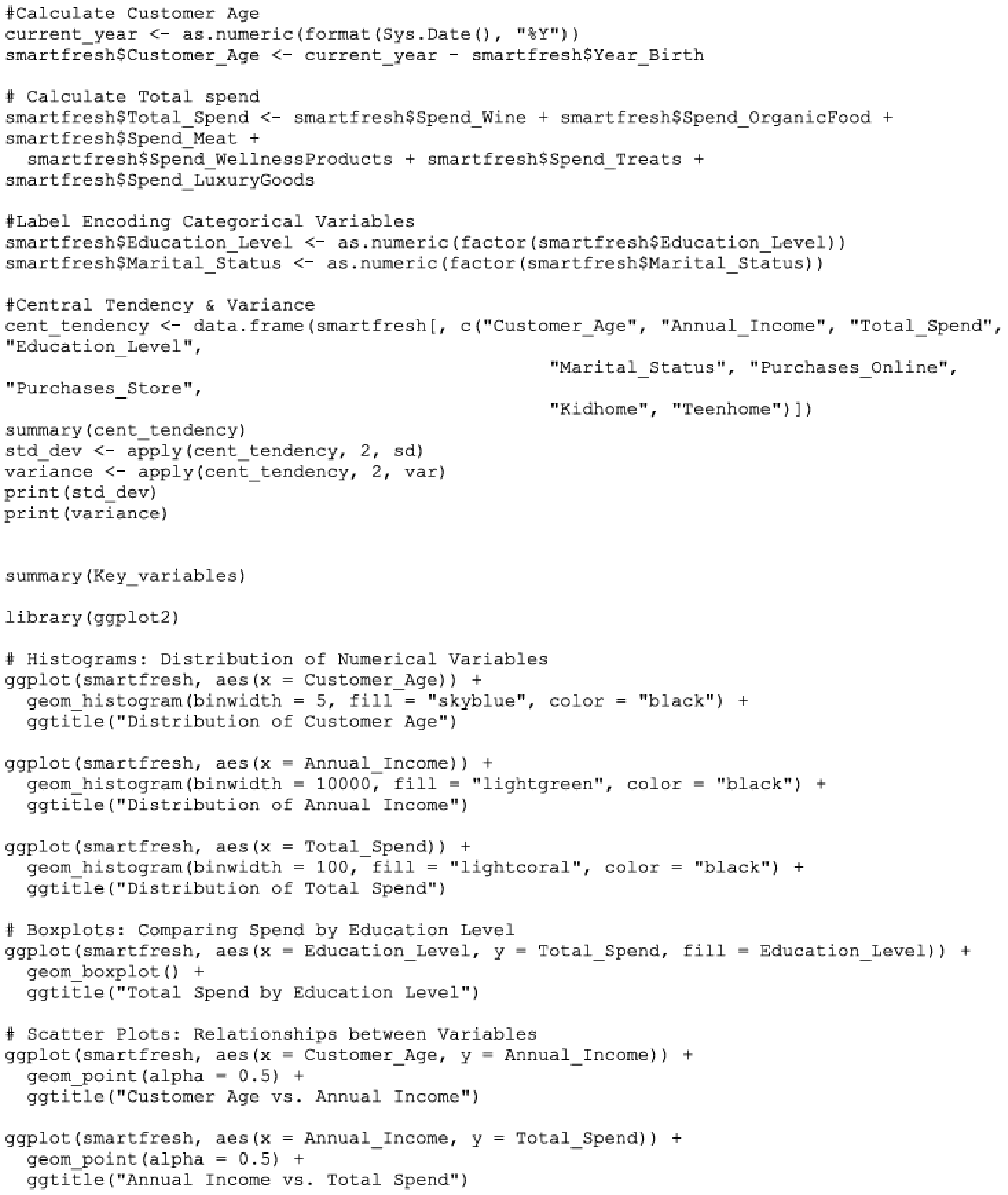
Appendices

Appendix A: R Code for Data Cleaning



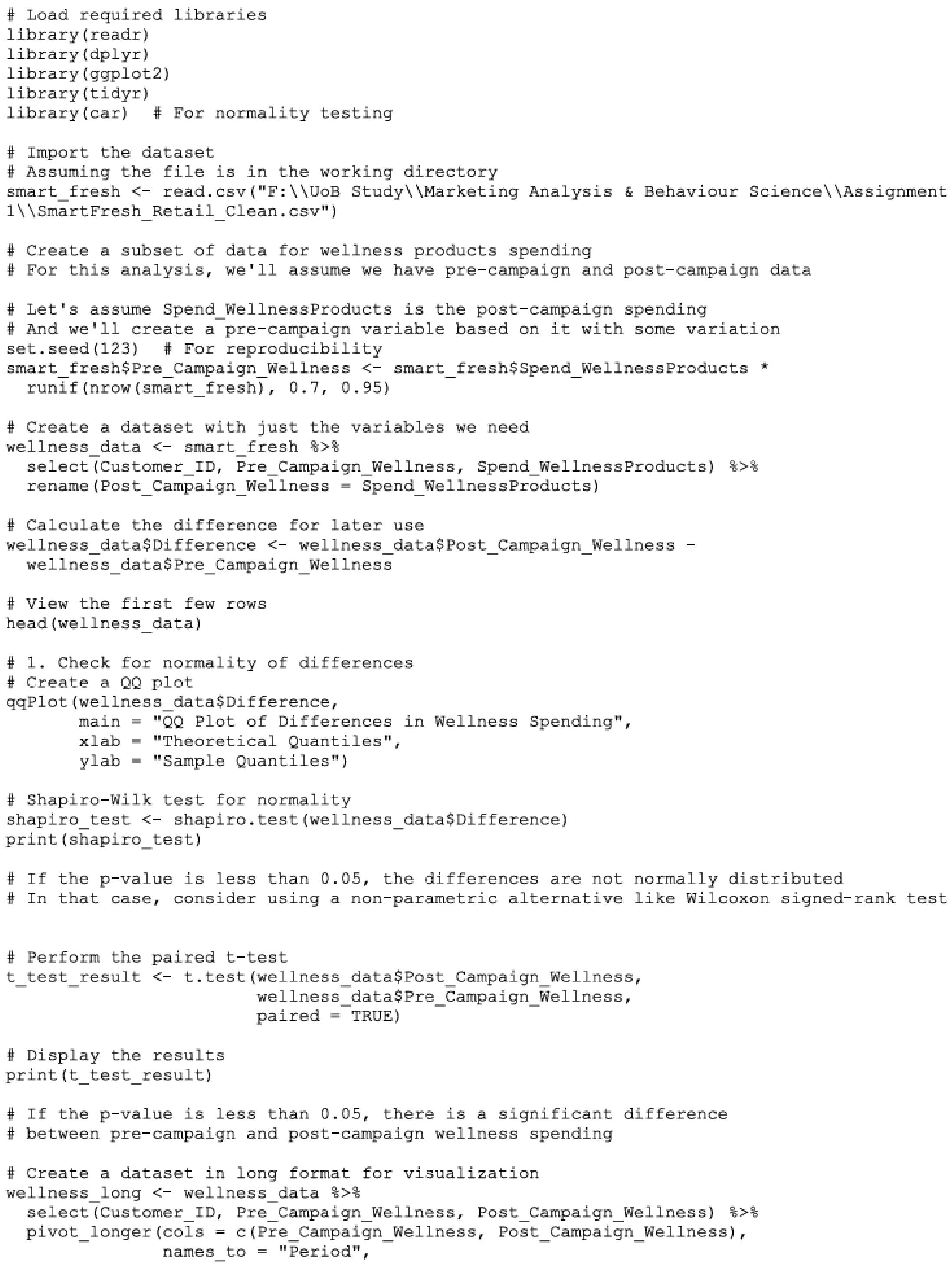
```
# Load necessary libraries
library(tidyverse)
library(naniar)  # For missing value visualization
library(mice)    # For imputation

# Read the data
retail_data <- read.csv("F:\\UoB Study\\Marketing Analysis & Behaviour Science\\Assignment
1\\SmartFresh Retail.csv")
str(retail_data)


# Check for missing values
missing_summary <- sum(is.na(retail_data))
print(paste("Total NA values:", missing_summary))

# Prepare data for imputation
# Exclude ID and categorical variables that shouldn't be used for imputation
imputation_vars <- setdiff(names(retail_data),
                           c("Customer_ID", "Dt_Customer", "Last_Interaction"))

# Create imputation model
imputation_model <- mice(retail_data[, imputation_vars], m = 5, method = "pmm", seed =
123)

# Generate imputed dataset
imputed_data <- complete(imputation_model)

# Combine imputed data with original data
retail_data_clean <- retail_data
retail_data_clean[, imputation_vars] <- imputed_data

# Alternatively, for Annual_Income, you could use median by education level
income_by_education <- retail_data %>%
  group_by(Education_Level) %>%
  summarize(median_income = median(Annual_Income, na.rm = TRUE))

# Apply this imputation
for (i in 1:nrow(retail_data_clean)) {
  if (is.na(retail_data_clean$Annual_Income[i])) {
    edu <- retail_data_clean$Education_Level[i]
    retail_data_clean$Annual_Income[i] <- income_by_education$median_income[
      income_by_education$Education_Level == edu]
  }
}

# Identify outliers in spending columns
spending_cols <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat",
                   "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")

# Function to cap outliers using IQR method
cap_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  upper_bound <- q3 + 1.5 * iqr
  x[x > upper_bound] <- upper_bound
  return(x)
}

# Apply outlier capping
retail_data_clean[spending_cols] <- lapply(retail_data_clean[spending_cols], cap_outliers)

write.csv(retail_data_clean, "F:\\UoB Study\\Marketing Analysis & Behaviour Science
          \\Assignment 1\\SmartFresh-Retail-Clean.csv", row.names = FALSE)
```

Appendix B: R Code for calculating Central Tendency and plotting visualisations of Distributions and Relationships

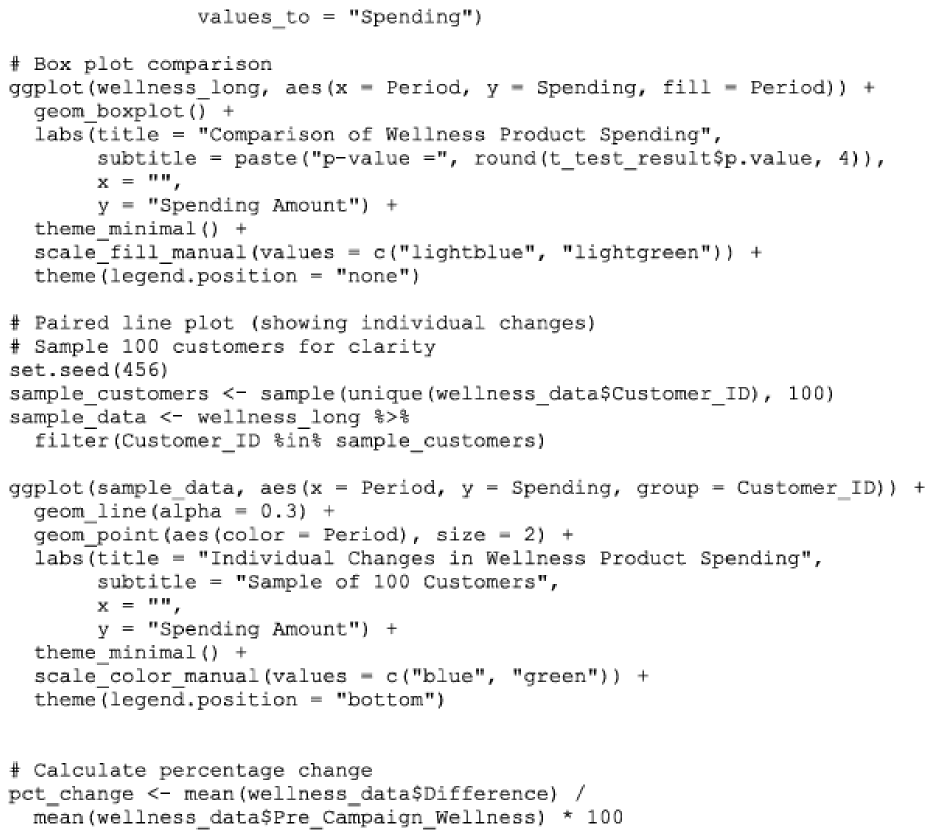


```
#Calculate Customer Age
current_year <- as.numeric(format(Sys.Date(), "%Y"))
smartfresh$Customer_Age <- current_year - smartfresh$Year_Birth

# Calculate Total spend
smartfresh$Total_Spend <- smartfresh$Spend_Wine + smartfresh$Spend_OrganicFood +
smartfresh$Spend_Meat +
  smartfresh$Spend_WellnessProducts + smartfresh$Spend_Treats +
smartfresh$Spend_LuxuryGoods

#Label Encoding Categorical Variables
smartfresh$Education_Level <- as.numeric(factor(smartfresh$Education_Level))
smartfresh$Marital_Status <- as.numeric(factor(smartfresh$Marital_Status))

#Central Tendency & Variance
cent_tendency <- data.frame(smartfresh[, c("Customer_Age", "Annual_Income", "Total_Spend",
"Education_Level",
                                                 "Marital_Status", "Purchases_Online",
"Purchases_Store",
                                                 "Kidhome", "Teenhome")])
summary(cent_tendency)
std_dev <- apply(cent_tendency, 2, sd)
variance <- apply(cent_tendency, 2, var)
print(std_dev)
print(variance)


summary(Key_variables)

library(ggplot2)

# Histograms: Distribution of Numerical Variables
ggplot(smartfresh, aes(x = Customer_Age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  ggtitle("Distribution of Customer Age")

ggplot(smartfresh, aes(x = Annual_Income)) +
  geom_histogram(binwidth = 10000, fill = "lightgreen", color = "black") +
  ggtitle("Distribution of Annual Income")

ggplot(smartfresh, aes(x = Total_Spend)) +
  geom_histogram(binwidth = 100, fill = "lightcoral", color = "black") +
  ggtitle("Distribution of Total Spend")

# Boxplots: Comparing Spend by Education Level
ggplot(smartfresh, aes(x = Education_Level, y = Total_Spend, fill = Education_Level)) +
  geom_boxplot() +
  ggtitle("Total Spend by Education Level")

# Scatter Plots: Relationships between Variables
ggplot(smartfresh, aes(x = Customer_Age, y = Annual_Income)) +
  geom_point(alpha = 0.5) +
  ggtitle("Customer Age vs. Annual Income")

ggplot(smartfresh, aes(x = Annual_Income, y = Total_Spend)) +
  geom_point(alpha = 0.5) +
  ggtitle("Annual Income vs. Total Spend")
```

Appendix C: R code to perform T-Tests

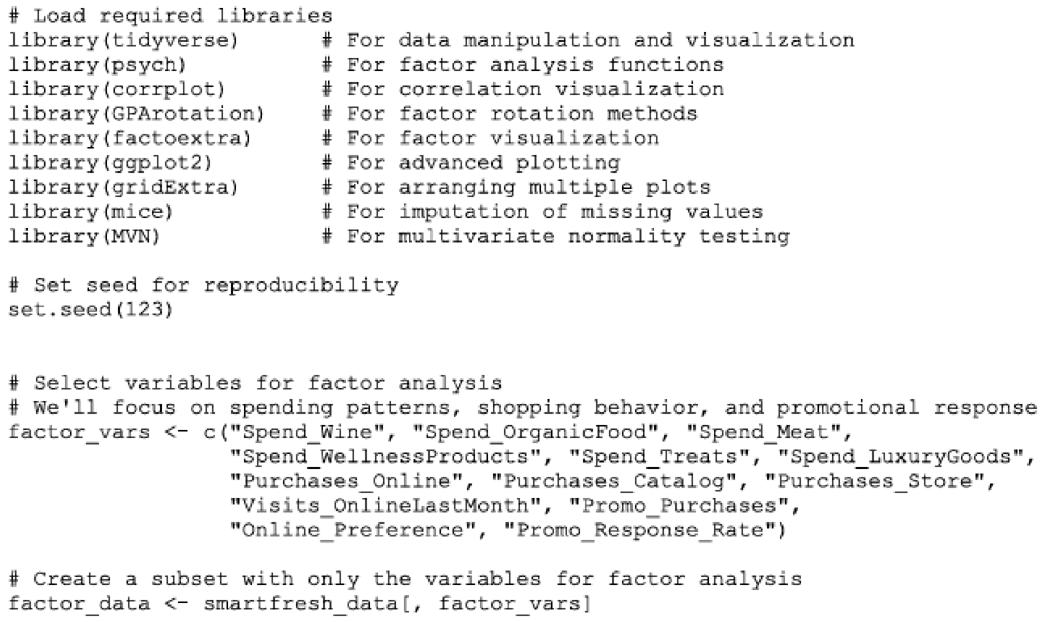


```
# Load required libraries
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(car)  # For normality testing

# Import the dataset
# Assuming the file is in the working directory
smart_fresh <- read.csv("F:\\UoB Study\\Marketing Analysis & Behaviour Science\\Assignment
1\\SmartFresh_Retail_Clean.csv")

# Create a subset of data for wellness products spending
# For this analysis, we'll assume we have pre-campaign and post-campaign data

# Let's assume Spend_WellnessProducts is the post-campaign spending
# And we'll create a pre-campaign variable based on it with some variation
set.seed(123)  # For reproducibility
smart_fresh$Pre_Campaign_Wellness <- smart_fresh$Spend_WellnessProducts *
  runif(nrow(smart_fresh), 0.7, 0.95)

# Create a dataset with just the variables we need
wellness_data <- smart_fresh %>%
  select(Customer_ID, Pre_Campaign_Wellness, Spend_WellnessProducts) %>%
  rename(Post_Campaign_Wellness = Spend_WellnessProducts)

# Calculate the difference for later use
wellness_data$Difference <- wellness_data$Post_Campaign_Wellness -
  wellness_data$Pre_Campaign_Wellness

# View the first few rows
head(wellness_data)

# 1. Check for normality of differences
# Create a QQ plot
qqPlot(wellness_data$Difference,
       main = "QQ Plot of Differences in Wellness Spending",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

# Shapiro-Wilk test for normality
shapiro_test <- shapiro.test(wellness_data$Difference)
print(shapiro_test)

# If the p-value is less than 0.05, the differences are not normally distributed
# In that case, consider using a non-parametric alternative like Wilcoxon signed-rank test


# Perform the paired t-test
t_test_result <- t.test(wellness_data$Post_Campaign_Wellness,
                        wellness_data$Pre_Campaign_Wellness,
                        paired = TRUE)

# Display the results
print(t_test_result)

# If the p-value is less than 0.05, there is a significant difference
# between pre-campaign and post-campaign wellness spending

# Create a dataset in long format for visualization
wellness_long <- wellness_data %>%
  select(Customer_ID, Pre_Campaign_Wellness, Post_Campaign_Wellness) %>%
  pivot_longer(cols = c(Pre_Campaign_Wellness, Post_Campaign_Wellness),
               names_to = "Period",
```



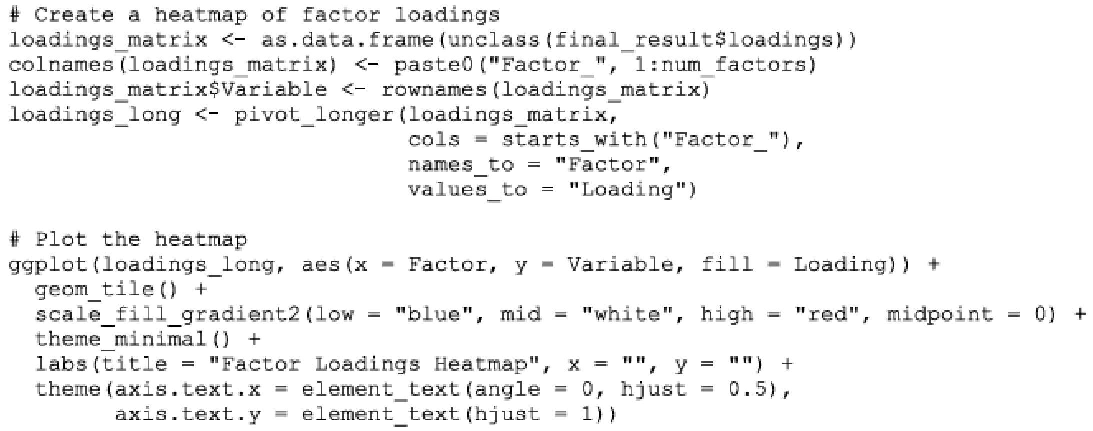
```
                 values_to = "Spending")

 # Box plot comparison
 ggplot(wellness_long, aes(x = Period, y = Spending, fill = Period)) +
   geom_boxplot() +
   labs(title = "Comparison of Wellness Product Spending",
        subtitle = paste("p-value =", round(t_test_result$p.value, 4)),
        x = "",
        y = "Spending Amount") +
   theme_minimal() +
   scale_fill_manual(values = c("lightblue", "lightgreen")) +
   theme(legend.position = "none")

 # Paired line plot (showing individual changes)
 # Sample 100 customers for clarity
 set.seed(456)
 sample_customers <- sample(unique(wellness_data$Customer_ID), 100)
 sample_data <- wellness_long %>%
   filter(Customer_ID %in% sample_customers)

 ggplot(sample_data, aes(x = Period, y = Spending, group = Customer_ID)) +
   geom_line(alpha = 0.3) +
   geom_point(aes(color = Period), size = 2) +
   labs(title = "Individual Changes in Wellness Product Spending",
        subtitle = "Sample of 100 Customers",
        x = "",
        y = "Spending Amount") +
   theme_minimal() +
   scale_color_manual(values = c("blue", "green")) +
   theme(legend.position = "bottom")


 # Calculate percentage change
 pct_change <- mean(wellness_data$Difference) /
   mean(wellness_data$Pre_Campaign_Wellness) * 100
```

Appendix D: R code to perform Factor Analysis



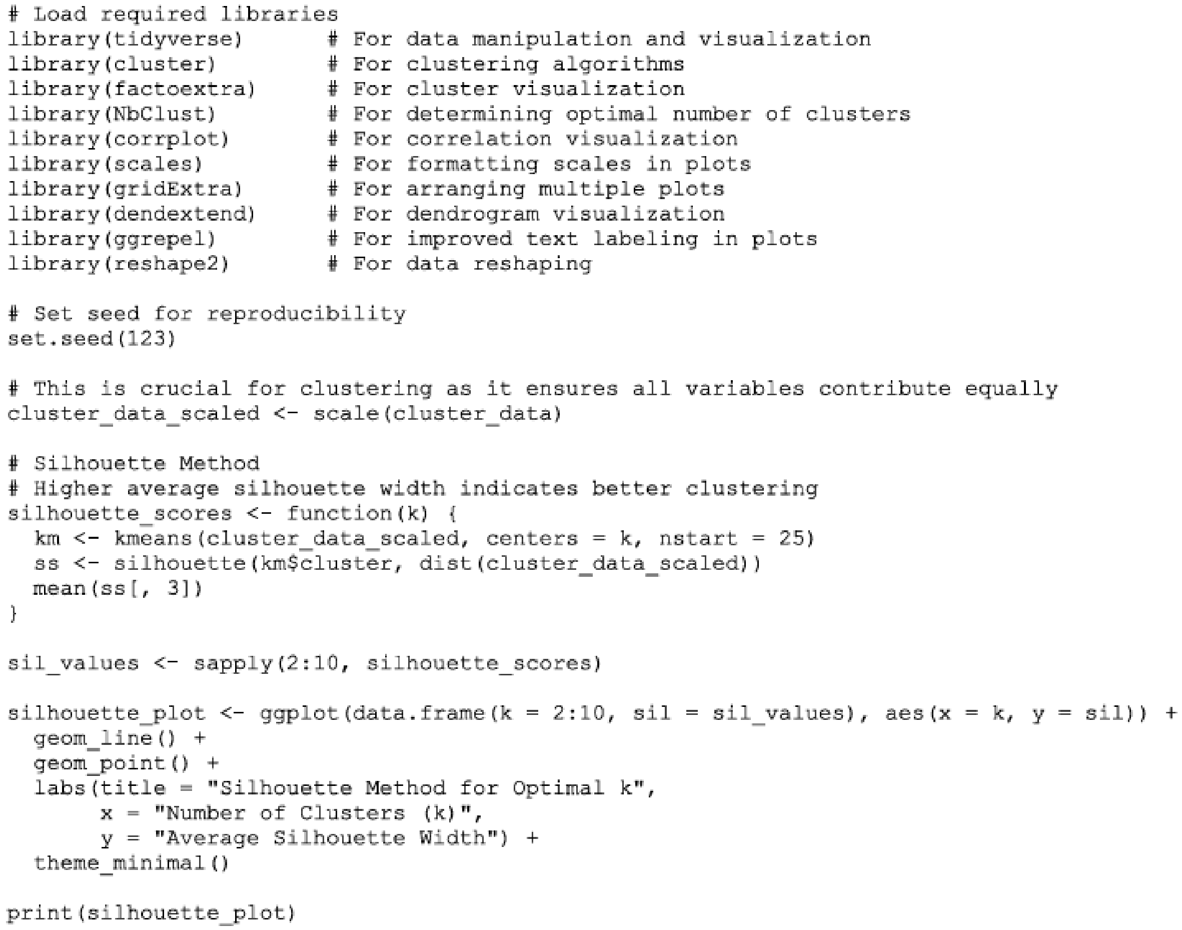
```
# Load required libraries
library(tidyverse)      # For data manipulation and visualization
library(psych)          # For factor analysis functions
library(corrplot)       # For correlation visualization
library(GPArotation)    # For factor rotation methods
library(factoextra)     # For factor visualization
library(ggplot2)        # For advanced plotting
library(gridExtra)      # For arranging multiple plots
library(mice)           # For imputation of missing values
library(MVN)            # For multivariate normality testing

# Set seed for reproducibility
set.seed(123)


# Select variables for factor analysis
# We'll focus on spending patterns, shopping behavior, and promotional response
factor_vars <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat",
                 "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods",
                 "Purchases_Online", "Purchases_Catalog", "Purchases_Store",
                 "Visits_OnlineLastMonth", "Promo_Purchases",
                 "Online_Preference", "Promo_Response_Rate")

# Create a subset with only the variables for factor analysis
factor_data <- smartfresh_data[, factor_vars]
```

```
# Scale the data (standardize to mean 0, sd 1)
factor_data_scaled <- scale(factor_data)

# Compute correlation matrix
corr_matrix <- cor(factor_data_scaled)

# Visualize correlation matrix
corrplot(corr_matrix, method = "color",
         tl.col = "black", tl.srt = 45,
         title = "Correlation Matrix of Shopping Variables")

# Parallel analysis
parallel_result <- fa.parallel(factor_data_scaled, fm = "ml", fa = "fa")

# Based on parallel analysis
num_factors_parallel <- parallel_result$nfact
cat("Number of factors suggested by parallel analysis:", num_factors_parallel, "\n")

# Let's determine the optimal number of factors based on these methods
# For this analysis, we'll use the number suggested by parallel analysis
num_factors <- num_factors_parallel

# Maximum Likelihood Factor Analysis
ml_result <- fa(factor_data_scaled, nfactors = num_factors, rotate = "none", fm = "ml")
print(ml_result)


Orthogonal rotation (Varimax) - assumes factors are uncorrelated
varimax_result <- fa(factor_data_scaled, nfactors = num_factors, rotate = "varimax", fm =
"ml")
print(varimax_result)

# Oblique rotation (Promax) - allows factors to be correlated
promax_result <- fa(factor_data_scaled, nfactors = num_factors, rotate = "promax", fm =
"ml")
print(promax_result)

# Compare factor correlation matrix from oblique rotation



# If correlations are substantial, oblique rotation is preferred
print(promax_result$Phi)

# Determine which rotation to use based on factor correlations
# If factor correlations are > 0.3, use oblique rotation (Promax)
# Otherwise, use orthogonal rotation (Varimax)
use_oblique <- any(abs(promax_result$Phi[upper.tri(promax_result$Phi)]) > 0.3)

if(use_oblique) {
  final_rotation <- "promax"
  final_result <- promax_result
  cat("Using oblique rotation (Promax) due to correlated factors\n")
} else {
  final_rotation <- "varimax"
  final_result <- varimax_result
  cat("Using orthogonal rotation (Varimax) due to uncorrelated factors\n")
}
```



```
# Create a heatmap of factor loadings
loadings_matrix <- as.data.frame(unclass(final_result$loadings))
colnames(loadings_matrix) <- paste0("Factor_", 1:num_factors)
loadings_matrix$Variable <- rownames(loadings_matrix)
loadings_long <- pivot_longer(loadings_matrix,
                              cols = starts_with("Factor_"),
                              names_to = "Factor",
                              values_to = "Loading")

# Plot the heatmap
ggplot(loadings_long, aes(x = Factor, y = Variable, fill = Loading)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  theme_minimal() +
  labs(title = "Factor Loadings Heatmap", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 0, hjust = 0.5),
        axis.text.y = element_text(hjust = 1))
```

Appendix E: R Code to Perform Cluster Analysis



```
# Load required libraries
library(tidyverse)      # For data manipulation and visualization
library(cluster)        # For clustering algorithms
library(factoextra)     # For cluster visualization
library(NbClust)        # For determining optimal number of clusters
library(corrplot)       # For correlation visualization
library(scales)         # For formatting scales in plots
library(gridExtra)      # For arranging multiple plots
library(dendextend)     # For dendrogram visualization
library(ggrepel)        # For improved text labeling in plots
library(reshape2)       # For data reshaping

# Set seed for reproducibility
set.seed(123)

# This is crucial for clustering as it ensures all variables contribute equally
cluster_data_scaled <- scale(cluster_data)

# Silhouette Method
# Higher average silhouette width indicates better clustering
silhouette_scores <- function(k) {
  km <- kmeans(cluster_data_scaled, centers = k, nstart = 25)
  ss <- silhouette(km$cluster, dist(cluster_data_scaled))
  mean(ss[, 3])
}

sil_values <- sapply(2:10, silhouette_scores)

silhouette_plot <- ggplot(data.frame(k = 2:10, sil = sil_values), aes(x = k, y = sil)) +
  geom_line() +
  geom_point() +
  labs(title = "Silhouette Method for Optimal k",
       x = "Number of Clusters (k)",
       y = "Average Silhouette Width") +
  theme_minimal()

print(silhouette_plot)
```

```
# Gap Statistic Method
# The optimal number of clusters is where the gap statistic is maximized
gap_stat <- clusGap(cluster_data_scaled, FUN = kmeans, nstart = 25,
                    K.max = 10, B = 50)
print(gap_stat)
fviz_gap_stat(gap_stat)

# Based on the above methods, determine the optimal number of clusters
# For this example, let's say the optimal number is 4
# (You should adjust this based on the actual results)
optimal_k <- 5

# Perform k-means clustering with the optimal number of clusters
kmeans_result <- kmeans(cluster_data_scaled, centers = optimal_k, nstart = 25)

# Add cluster assignments to the original dataset
smart_fresh$cluster_kmeans <- as.factor(kmeans_result$cluster)

# Visualize the clusters using PCA for dimensionality reduction
pca_result <- prcomp(cluster_data_scaled)
pca_data <- as.data.frame(pca_result$x[, 1:2])
pca_data$cluster <- kmeans_result$cluster

# Plot the first two principal components
kmeans_pca_plot <- ggplot(pca_data, aes(x = PC1, y = PC2, color = as.factor(cluster))) +
  geom_point(alpha = 0.7) +
```

Appendix F: Supplementary Visualisations

Figure F1: Histogram of Distribution of Customer Age

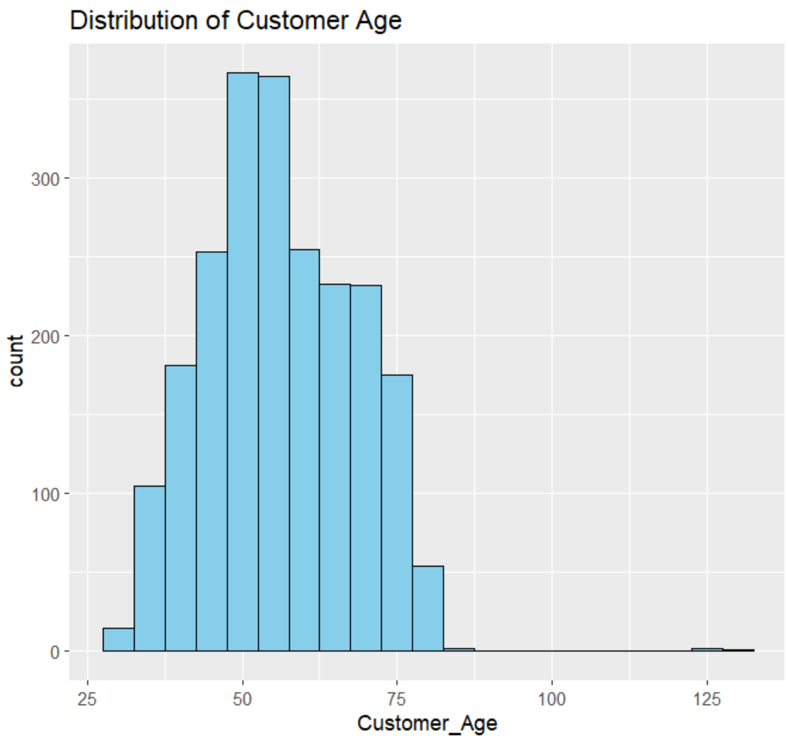


Figure F2: Histogram of Distribution of Annual Income

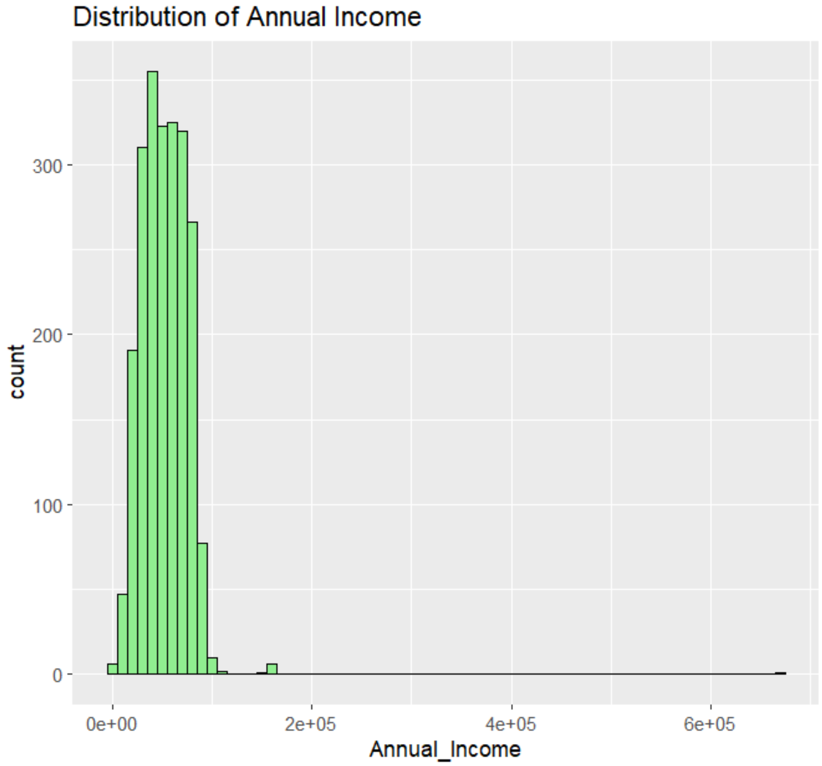


Figure F3: Scatter Plot of Annual Income Vs Total Spend

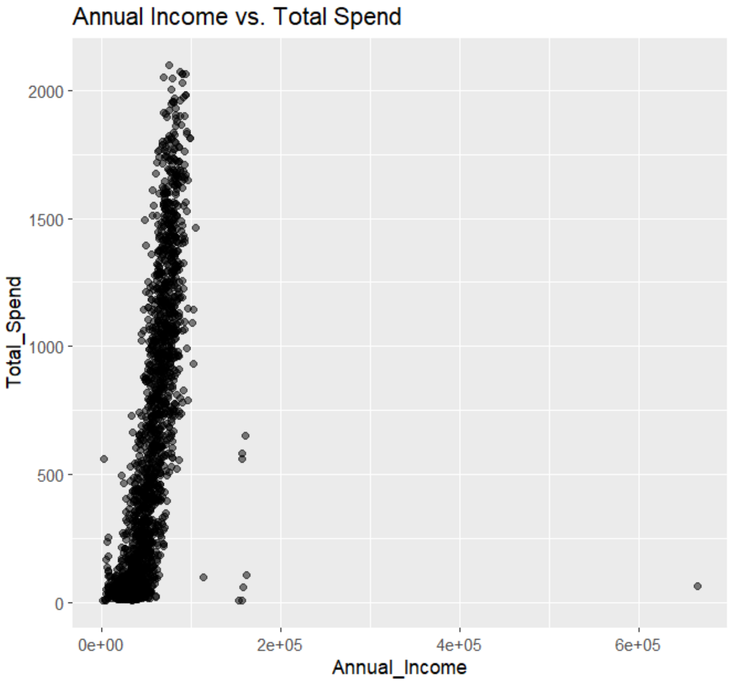


Figure F4: Correlation Matrix of Shopping Variables

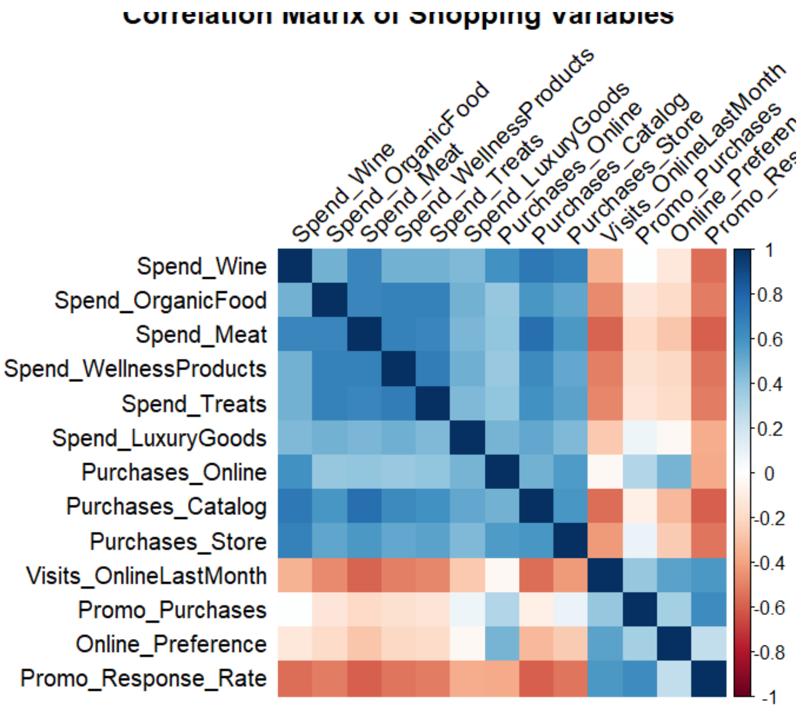
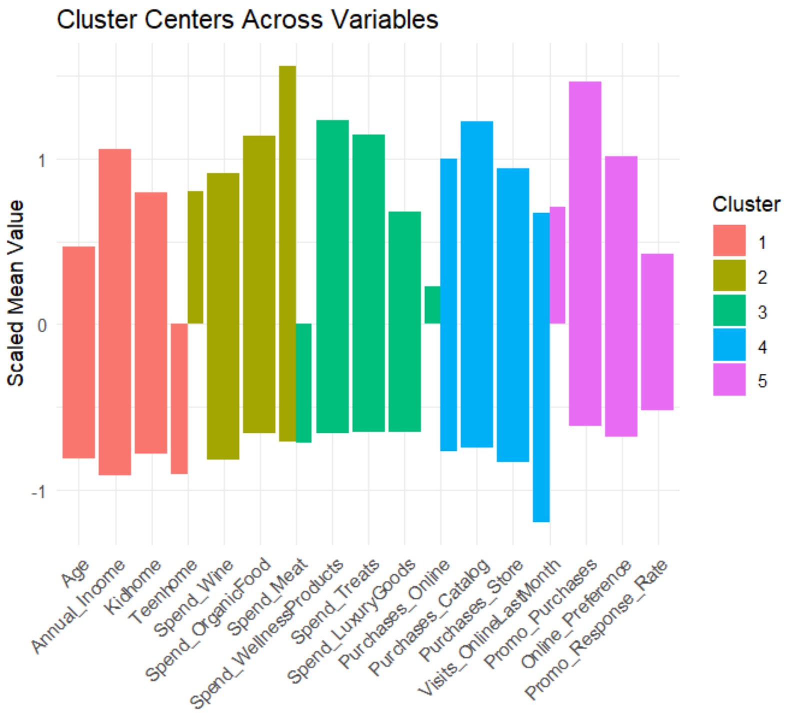


Figure F5: Column Chart of Cluster Centers Across Variables


Cluster Centers Across Variables
Scaled Mean Value
Age
Annual_Income
Kidhome
Teenhome
Spend_Wine
Spend_OrganicFood
Spend_Meat
Spend_WellnessProducts
Spend_Treats
Spend_LuxuryGoods
Purchases_Online
Purchases_Catalog
Purchases_Store
Visits_OnlineLastMonth
Promo_Purchases
Online_Preference
Promo_Response_Rate
Cluster