DATA-DRIVEN CUSTOMER SEGMENTATION FOR SMART FRESH RETAIL

A PCA and K-Means Clustering Approach to optimize Retail Marketing Strategies

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

In the competitive retail market, understanding consumer preferences and purchasing behaviors is essential for optimizing marketing efforts. *SmartFresh Retail*, an omnichannel retailer specializing in high-end lifestyle products, aims to enhance customer segmentation and improve targeted marketing strategies. Through the analysis of customer profiles and transactional behaviors, the company seeks to deliver personalized experiences, improve retention, and drive revenue growth.

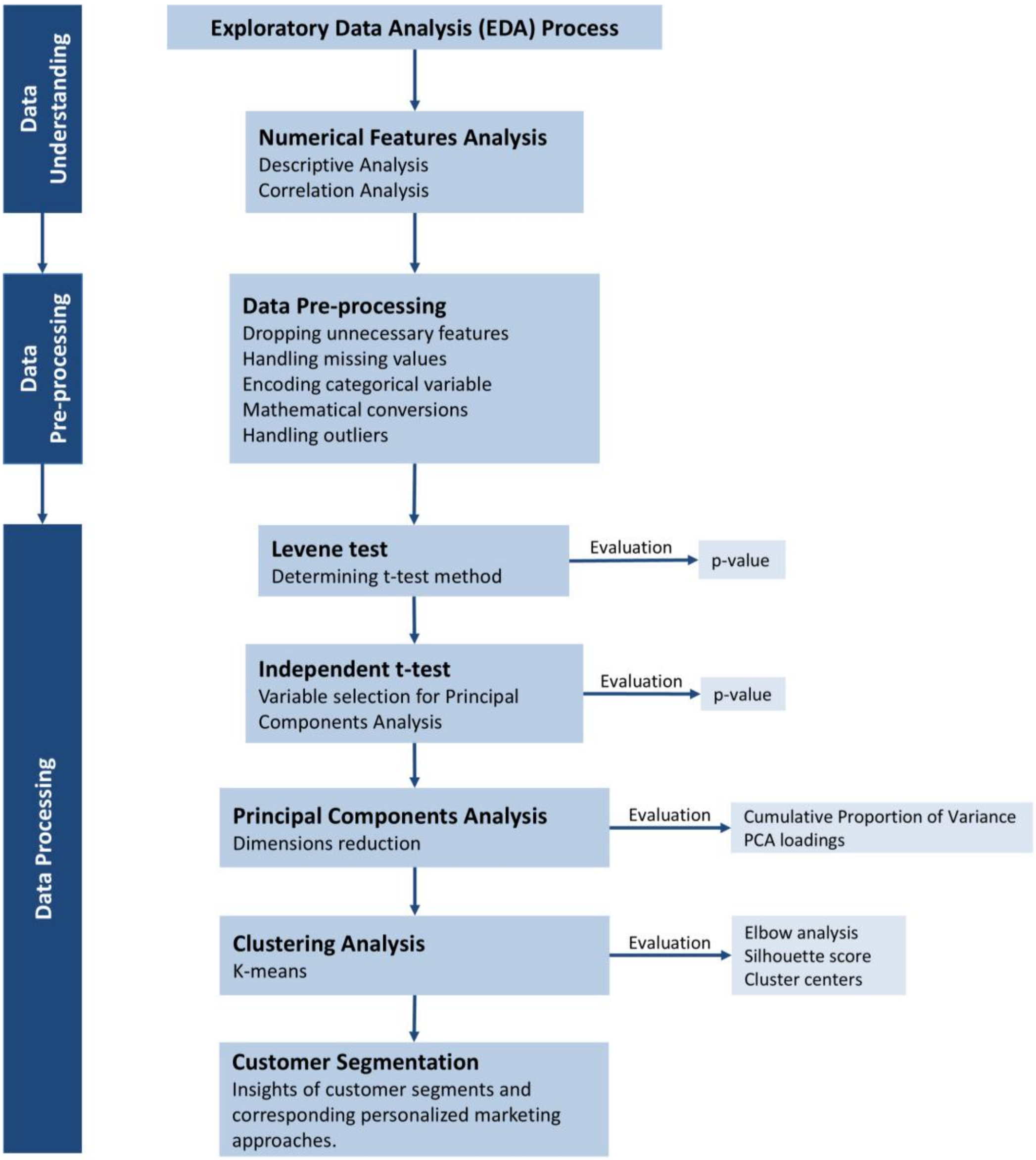


Figure 1: Data Analysis Framework

To achieve these objectives, a data-driven approach was implemented using a comprehensive dataset, processed according to the above Data Analysis Framework. As shown in Figure 1, this structured framework begins with Exploratory Data Analysis (EDA) to identify spending trends, outliers, and correlations. Subsequent data pre-processing ensures analytical integrity by addressing missing values, encoding categorical variables, and mitigating outliers.

Levene's test and independent t-tests guide the selection of variables for Principal Component Analysis (PCA), which reduces dimensionality while preserving 82.2% of data variance. The first four components capture income-driven spending, promotional sensitivity, luxury preferences, and online versus offline shopping behaviors.

After reducing data complexity through PCA, K-means clustering segments consumers into three distinct groups: high-income broad spenders, budget-conscious promotion-driven consumers, and selective online luxury shoppers. This segmentation framework enables SmartFresh to enhance marketing efficiency by ensuring that each consumer group receives tailored experience, optimizing engagement and profitability.

I. EXPLORATORY DATA ANALYSIS PROCESS

1. Customer Demographic Variables

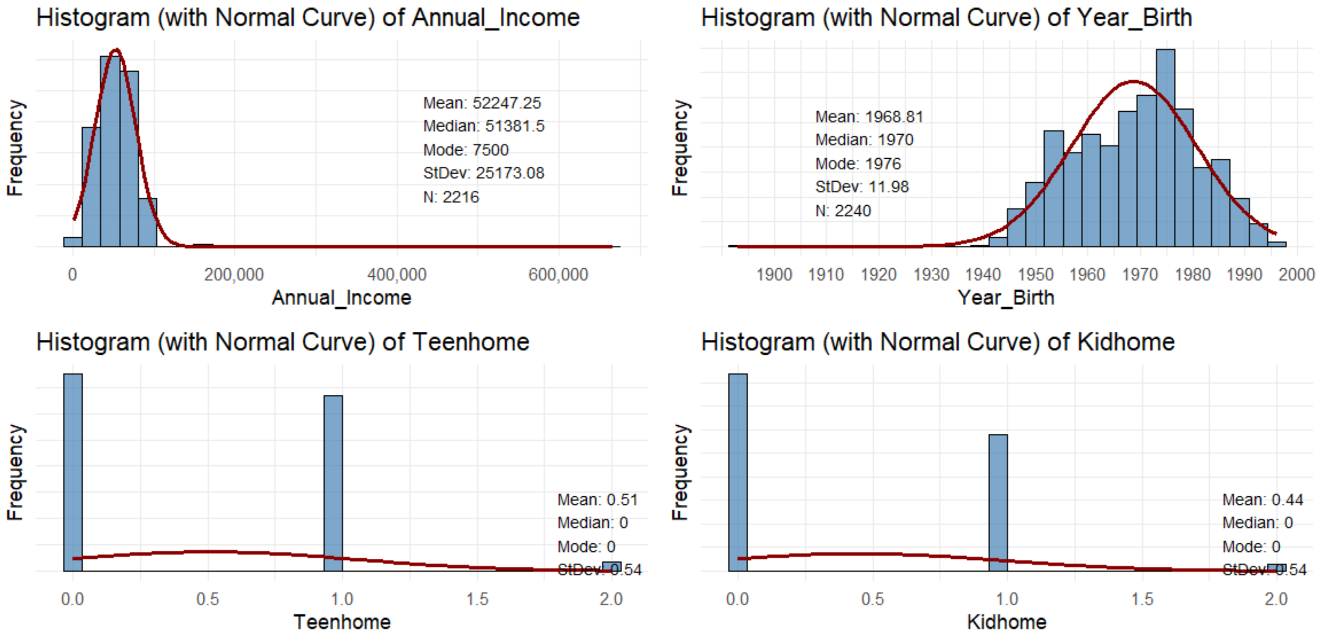


Figure 2: Histograms of Customer Demographic Variables
R Code

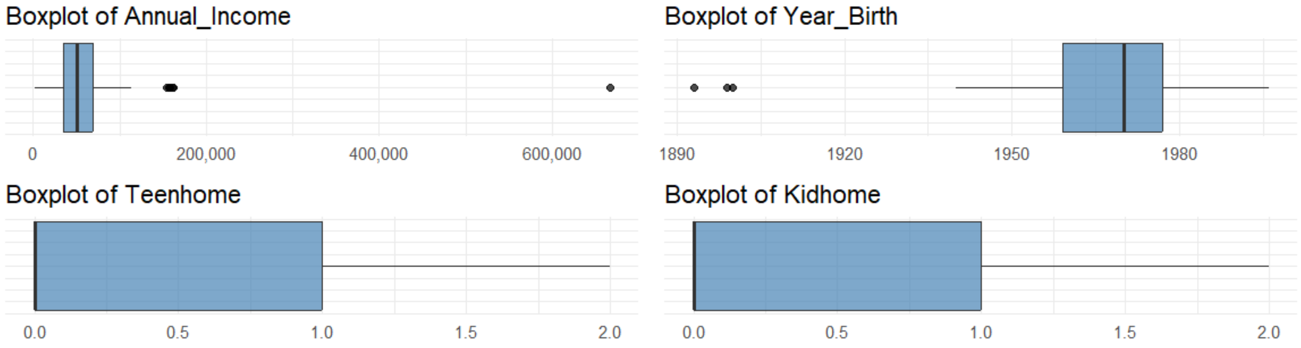


Figure 3: Boxplots of Customer Demographic Variables
R Code

Demographic variables such as Annual_Income, Year_Birth, and household composition (Teenhome and Kidhome) are crucial in marketing analysis as they segment consumers based on purchasing power, preferences, and life stage (Kotler and Keller, 2021). Annual income influences spending capacity, while birth year reflects generational behavior (Solomon, 2020). Household composition affects product demand, particularly for family-related goods (Homburg and Wielgos, 2022). The histograms illustrate skewness in income, normality in the birth year, and categorical distribution for family structure. Boxplots highlight outliers in income and birth year, indicating outliers and diverse consumer groups.

2. Customer Spending Pattern Variables

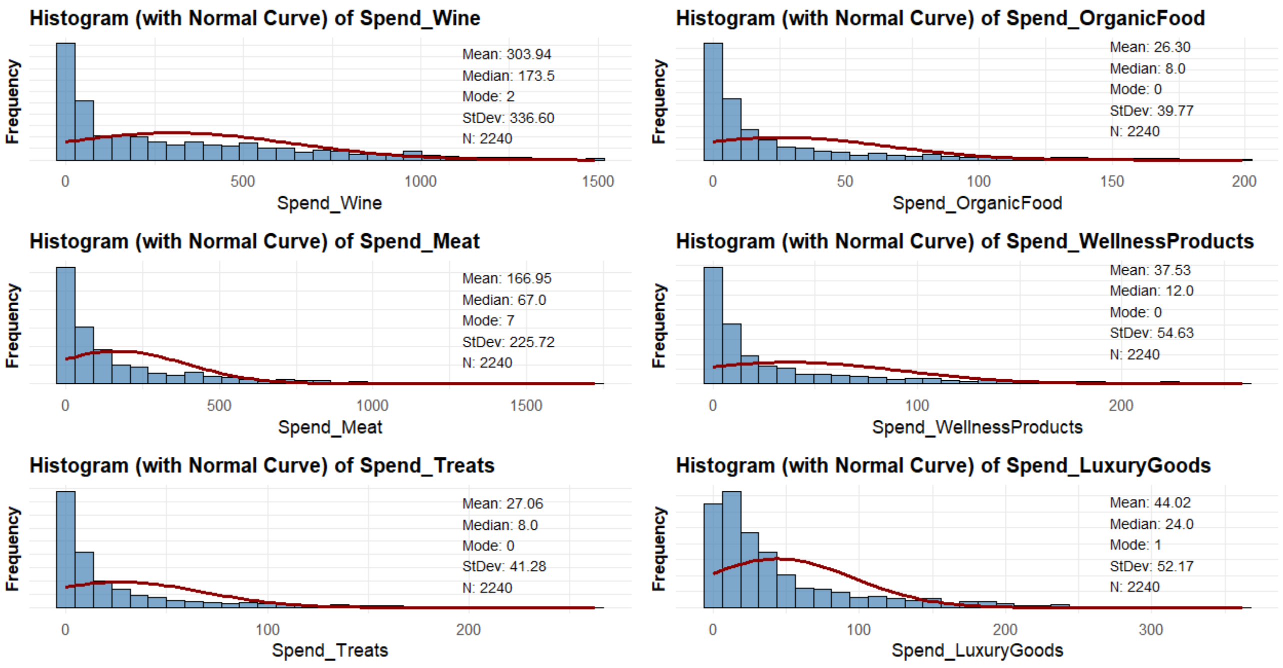


Figure 4: Histograms of Customer Spending Pattern Variables
R Code

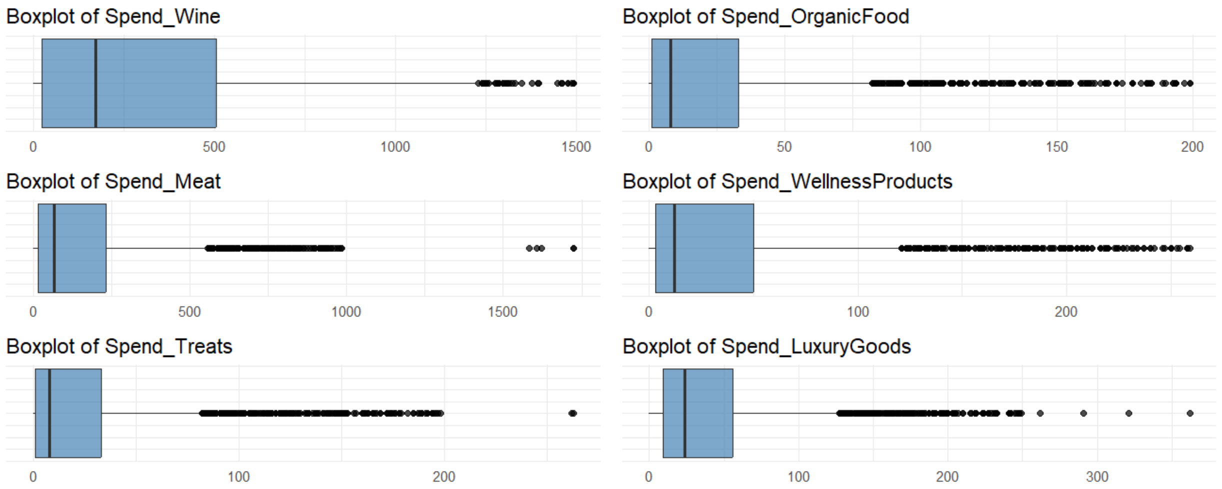


Figure 5: Boxplots of Customer Spending Pattern Variables
R Code

The spending pattern variables, including wine, organic food, meat, wellness products, treats, and luxury goods, exhibit highly skewed distributions, with most customers spending minimally and a few significantly increasing the mean values. The right-skewed histograms highlight that spending is concentrated at lower values, with a long tail extending towards higher expenditures. The boxplots further confirm the presence of extreme outliers, particularly in spending on wine, meat, and luxury goods, suggesting a small group of high spenders disproportionately affecting the distribution. The variability in spending patterns indicates heterogeneous consumer behavior, requiring proper scaling and outlier treatment for accurate analysis.

3. Customer Shopping Channel Variables

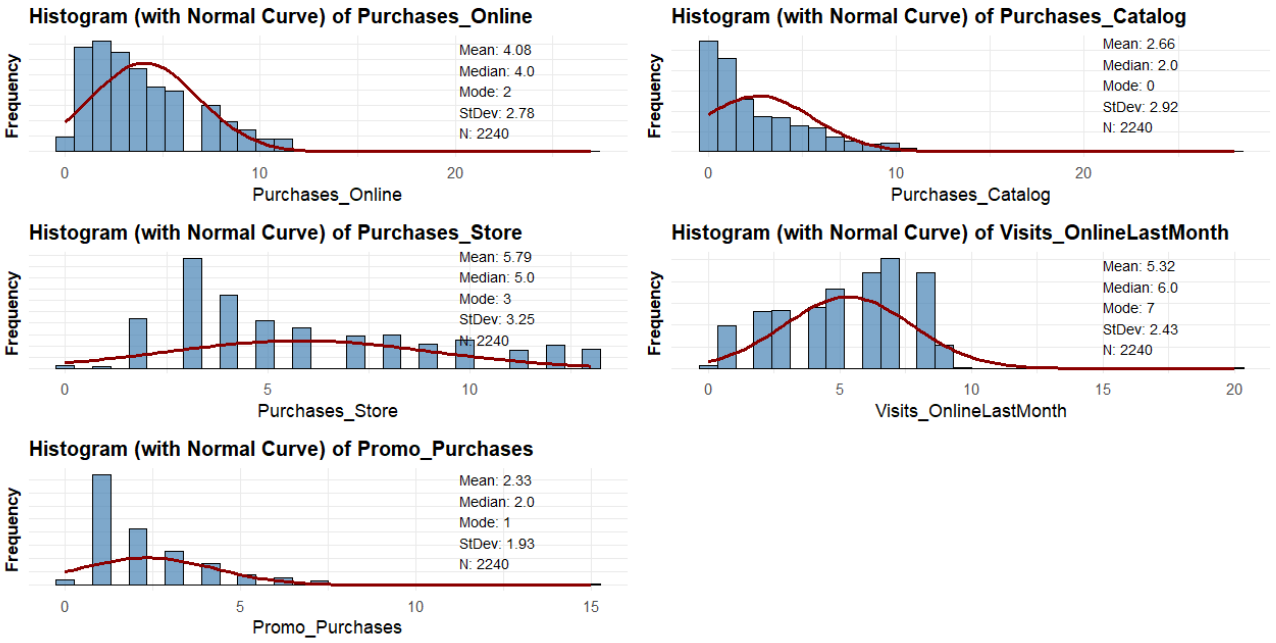


Figure 6: Histograms of Customer Shopping Channel Variables
[R Code]

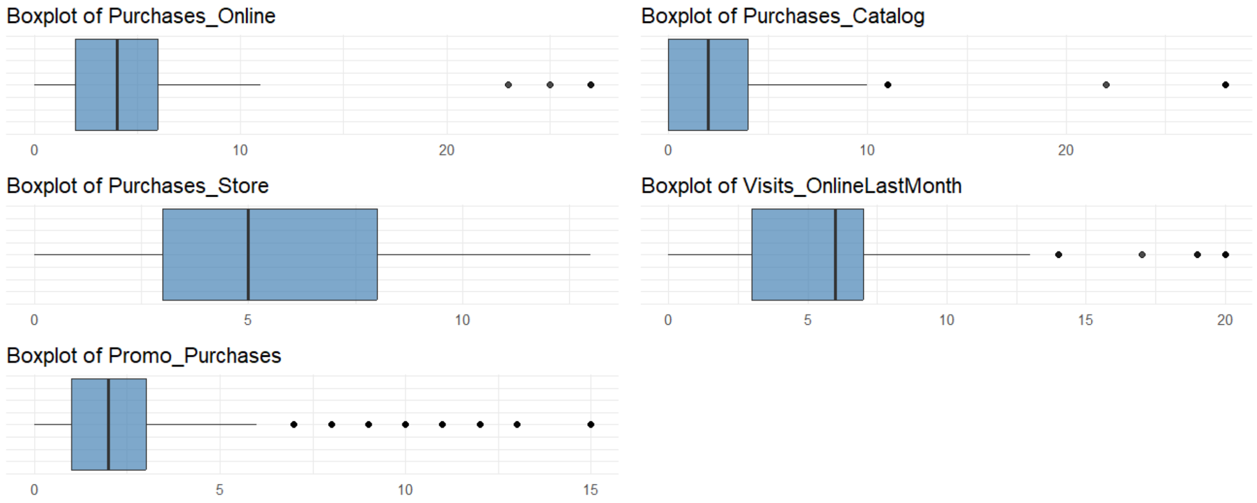


Figure 7: Boxplots of Customer Shopping Channel Variables
[R Code]

The right-skewed histograms indicate that most customers make few purchases, while a smaller segment exhibits high purchase frequency, particularly in-store and online purchases. Catalog and promo purchases show lower median values, suggesting they are used less frequently. The boxplots highlight extreme outliers, especially in online and store purchases, indicating a subset of highly engaged shoppers. The visits to online platforms show a multimodal distribution, suggesting distinct browsing behaviors among customer groups. These variations emphasize the need for tailored data scaling and potential segmentation for identifying frequent and occasional buyers based on their channel preferences.

4. Correlation Analysis

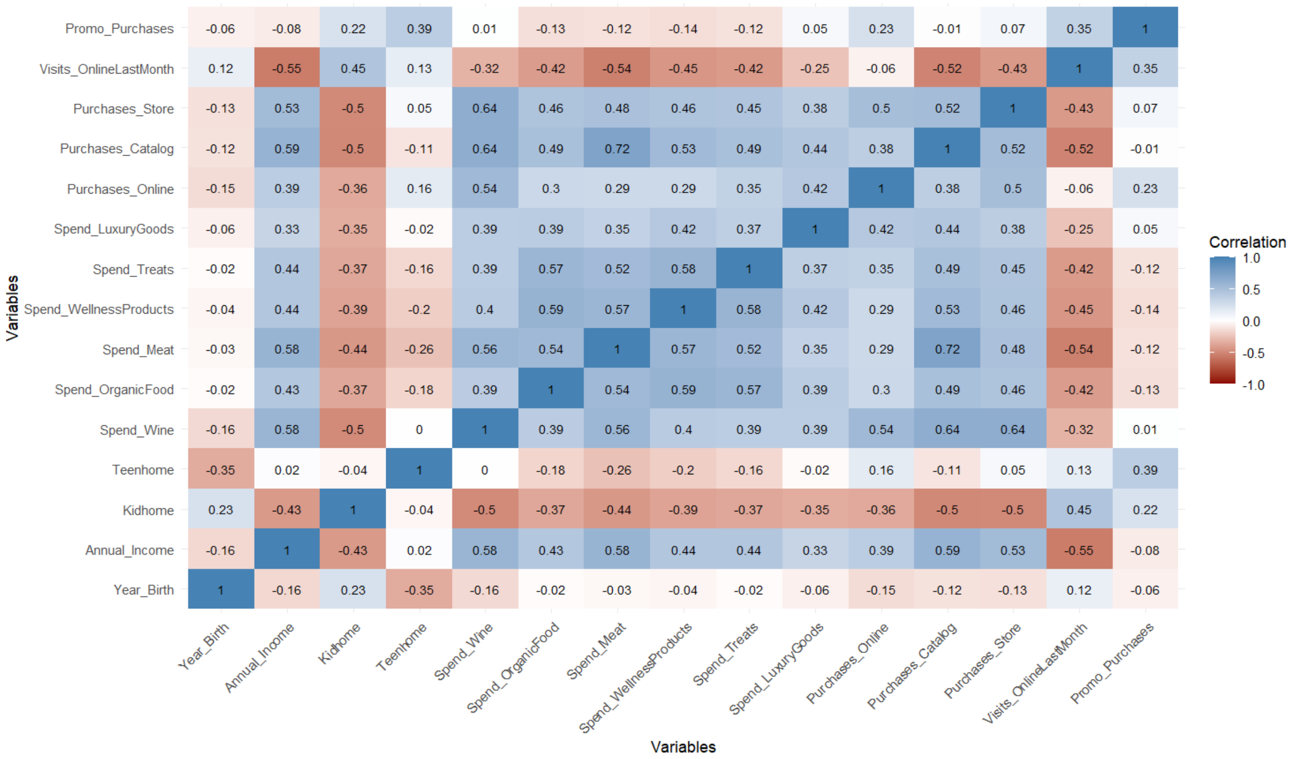


Figure 8: Correlation Matrix of key variables
R Code

The correlation heatmap reveals key consumer behavior trends. A **negative correlation** between **Year_Birth** and **spendings** suggests that **younger** customers allocate **fewer resources to purchases** across various categories. Conversely, **Annual_Income positively correlates** with spending on **luxury goods**, wine, and organic food, reinforcing **income-driven purchasing patterns**. Additionally, **Kidhome** and **Teenhome negatively correlate** with discretionary **spending**, indicating that households **with children** prioritize **essential expenses**. However, these variables **positively correlate** with **online shopping**, implying that **families with children** favor **digital channels** for convenience and accessibility.

II. DATA PRE-PROCESSING

The pre-processing steps involved several key transformations to **enhance data quality and interpretability**. Unnecessary features, such as Customer_ID and Marital_Status, were dropped. **Marital_Status** was removed due to its **lack of significant ordinality**, unlike Education Level, which was numerically encoded. Additionally, Num_Dependents was created by combining Kidhome and Teenhome, offering a compact view of household structure and social status without the need for Marital_Status.

Missing values in Annual Income were handled by **imputing the median**, a robust approach to minimize skewness caused by extreme values. Mathematical conversions were applied to derive Age from Year of Birth, ensuring consistent age representation. Furthermore, Dt_Customer was transformed into Days_Engaged, calculating the number of days since the initial enrollment.

To address **outliers**, values were **capped at the 1st and 99th percentiles**, retaining essential variations while mitigating extreme distortions. This step **preserves meaningful consumer patterns without** allowing outliers to **bias** clustering or segmentation models (Hair et al., 2019). Outlier capping ensures stable statistical modeling, a necessity in marketing analytics, where segmentation insights must reflect real customer behavior (Iglewicz and Hoaglin, 1993; Kotler and Keller, 2021).

III. DATA PROCESSING

1. Levene Test

The Levene test assesses the homogeneity of variances between groups, determining whether a standard independent t-test or Welch's t-test should be used for the subsequent independent t-test.

Variable groups	Variables	p-value					
		Accepted Offer1	Accepted Offer2	Accepted Offer3	Accepted Offer4	Accepted Offer5	Response Latest
Customer Information	Annual_Income	0.250	0.000	0.000	0.000	0.080	0.000
	Age	0.856	0.763	0.000	0.000	0.427	0.201
	Num_Dependents	0.389	0.541	0.000	0.000	0.861	0.000
	Education_Level	0.942	0.945	0.717	0.264	0.400	0.633
	Days_Engaged	0.171	0.656	0.590	0.901	0.832	0.195
	Promo_Purchases	0.380	0.010	0.000	0.000	0.159	0.011
Customer Spending Pattern	Spend_Wine	0.000	0.000	0.081	0.071	0.085	0.000
	Spend_OrganicFood	0.509	0.890	0.000	0.000	0.444	0.000
	Spend_Meat	0.264	0.005	0.000	0.000	0.395	0.000
	Spend_Treats	0.711	0.185	0.000	0.000	0.624	0.000
	Spend_LuxuryGoods	0.001	0.536	0.000	0.000	0.417	0.001
Customer Shopping Channel	Purchases_Online	0.012	0.493	0.000	0.010	0.865	0.638
	Purchases_Catalog	0.000	0.295	0.656	0.438	0.561	0.000
	Purchases_Store	0.543	0.232	0.443	0.108	0.550	0.353
	Visits_OnlineLastMonth	0.202	0.945	0.000	0.022	0.910	0.000

Table 1: Levene test results matrix
R Code

A p-value > 0.05 indicates **equal variances** (highlighted cells), allowing the use of the **standard t-test**, while a p-value < 0.05 suggests **unequal variances**, necessitating **Welch's t-test**, which adjusts for variance differences and provides more reliable results (Ruxton, 2006). Welch's t-test is preferred when heteroscedasticity is present, as it reduces Type I errors and enhances robustness in hypothesis testing

(Delacre et al., 2017). In the table, several variables show significant p-values (< 0.05), indicating that Welch's t-test should be applied for those cases.

2. Independent t-test

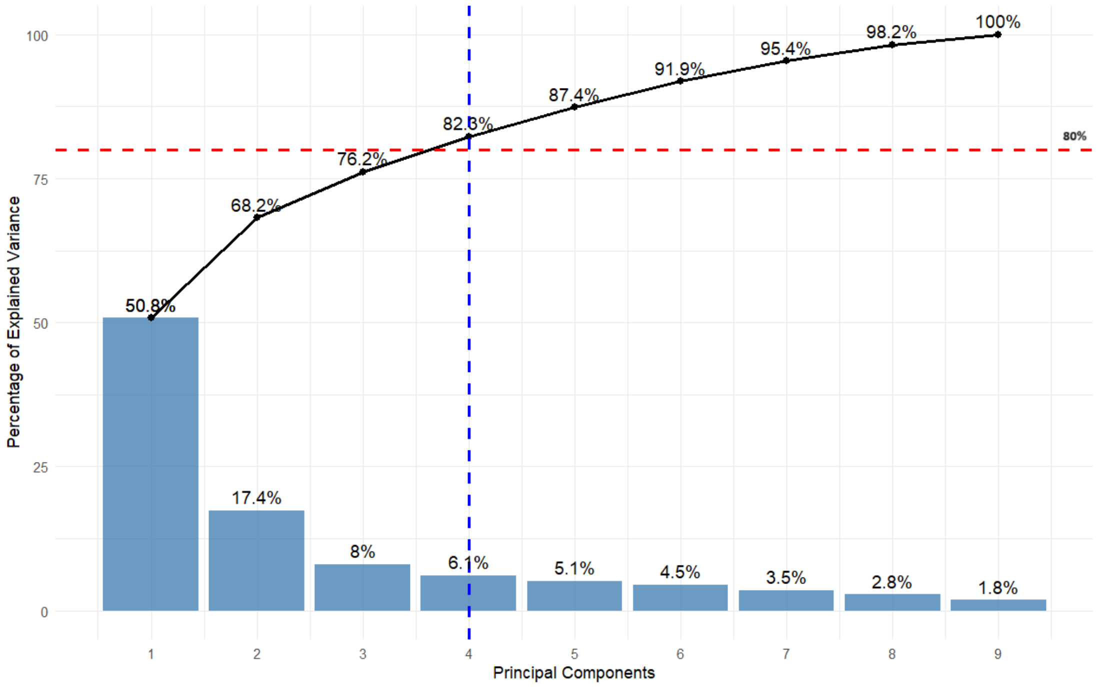
Variable groups	Variables	p-value					
		Accepted Offer1	Accepted Offer2	Accepted Offer3	Accepted Offer4	Accepted Offer5	Response Latest
Customer Information	Annual_Income	0.535	0.000	0.000	0.000	0.000	0.000
	Age	0.004	0.003	0.608	0.752	0.705	0.362
	Num_Dependents	0.334	0.000	0.000	0.000	0.001	0.000
	Education_Level	0.942	0.611	0.717	0.264	0.400	0.270
	Days_Engaged	0.171	0.656	0.590	0.901	0.832	0.195
	Promo_Purchases	0.324	0.458	0.000	0.000	0.066	0.794
Customer Spending Pattern	Spend_Wine	0.013	0.000	0.000	0.000	0.000	0.000
	Spend_OrganicFood	0.518	0.600	0.000	0.000	0.654	0.000
	Spend_Meat	0.312	0.000	0.000	0.000	0.031	0.000
	Spend_Treats	0.916	0.202	0.000	0.000	0.616	0.000
	Spend_LuxuryGoods	0.000	0.254	0.000	0.000	0.026	0.000
Customer Shopping Channel	Purchases_Online	0.026	0.000	0.000	0.000	0.079	0.000
	Purchases_Catalog	0.000	0.000	0.000	0.000	0.000	0.000
	Purchases_Store	0.001	0.000	0.000	0.000	0.000	0.066
	Visits_OnlineLastMonth	0.005	0.158	0.000	0.000	0.773	0.976

Table 2: Independent t-test results matrix
R Code

The independent t-test results provide insights into how variables differ based on offer acceptance. Highlighted cells of variables **with consistently high p-values** (> 0.05) across multiple offers indicate that they **do not significantly contribute to distinguishing** between customer segments. Including such variables in Principal Component Analysis (PCA) can dilute meaningful patterns, reducing the efficiency of dimensionality reduction (Jollife and Cadima, 2016). From the table, *Age, Num_Dependents, Days_Engaged, Promo_Purchases, Spend_OrganicFood, Spend_Treats, and Visits_OnlineLastMonth* frequently show non-significant p-values, meaning they do not substantially differentiate customer responses. These variables may lack variability in the dataset or not be directly relevant to purchasing behaviors linked to promotions. Removing them ensures that the subsequent PCA focuses on highly discriminative variables, improving the interpretability of principal components.

In contrast, variables with **consistent statistical significance** (< 0.05) across multiple offers - such as *Annual_Income, Spend Wine, Spend_Meat, Spend Luxury Goods, and other shopping channel variables* - **indicate more substantial differentiation** in customer behavior, making them more suitable for PCA. The exclusion of weakly contributing variables enhances pattern recognition in later clustering analysis, leading to more precise and more actionable customer segmentation (Hair et al., 2019). Ultimately, refining variable selection before PCA ensures that principal components effectively capture the most influential drivers of consumer behavior, optimizing segmentation for marketing strategies and targeted campaigns.

3. Principal Components Analysis



Statistics	Principal Components								
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.137	1.250	0.847	0.741	0.674	0.635	0.558	0.500	0.428
Proportion of Variance	0.508	0.174	0.080	0.061	0.051	0.045	0.035	0.028	0.020
Cumulative Proportion	**0.508**	**0.681**	**0.761**	**0.822**	0.872	0.917	0.952	0.980	1.000

Figure 9: Scree plot of the principal components
R Code

The Principal Component Analysis (PCA) results show that the first four principal components (PCs) capture **82.2%** of the variance, **exceeding the 80% sufficiency threshold** (Jollife and Cadima, 2016). PC1 alone explains **50.8%**, followed by **17.4%**, **8%**, and **6.1%** for PC2–PC4. Variance **beyond PC4 is minimal**, making these four PCs optimal for dimensionality reduction while retaining essential information. This enhances efficiency and interpretability in the subsequent k-means clustering analysis.

To derive the principal component (PC) loadings, Principal Component Analysis (PCA) was applied, a statistical technique commonly used in customer segmentation to reduce dimensionality while retaining key behavioral patterns (Hair et al., 2019). PCA identifies latent variables — principal components — that explain the maximum variance in customer data, transforming correlated variables into a set of uncorrelated factors (Jollife and Cadima, 2016). The loadings in the following table indicate the correlation between the original variables and each principal component, allowing interpretation of underlying behaviors.

Variables	Principal Components' loadings			
	PC1	PC2	PC3	PC4
Annual_Income	-0.4067	0.0135	-0.2501	0.2332
Num_Dependents	0.2576	-0.4964	-0.2106	0.4818
Spend_Wine	-0.3955	-0.0990	-0.2334	-0.0696
Spend_Meat	-0.3777	0.1905	-0.1183	0.3945
Spend_LuxuryGoods	-0.2803	-0.1210	0.8878	0.1729
Promo_Purchases	0.0348	-0.7138	0.0198	0.1505
Purchases_Online	-0.3050	-0.3950	0.0238	-0.5688
Purchases_Catalog	-0.4019	0.0519	-0.0044	0.3313

Table 3: Principal Components' Loadings
R Code

PC1: Income-Driven Shopping and General Spending Behavior

PC1 reflects shopping behavior **influenced by income levels**, indicated by strong **negative loadings** on **Annual_Income (-0.4067)**, **Spend_Wine (-0.3955)**, **Spend_Meat (-0.3777)**, and **Purchases_Catalog (-0.4019)**. **Higher-income** consumers exhibit **selective shopping habits**, spending less across multiple categories and channels. In contrast, **lower-income** consumers allocate **more of their budget to discretionary spending** and engage in **broader retail activity**. The **negative correlation** with shopping channels implies that **lower-income groups shop more frequently** across different formats, while **higher-income consumers purchase selectively**, prioritizing essential needs and premium goods (Kotler and Keller, 2021).

PC2: Promotional Sensitivity and Household Influence

PC2 is associated with **promotional sensitivity** and **household-driven** spending, as indicated by **negative loadings** on **Promo_Purchases (-0.7138)** and **Num_Dependents (-0.4964)**. Lower PC2 values suggest that **budget-conscious** households actively **engage with promotions and discounts**, possibly due to financial constraints or family-related shopping behaviors. In contrast, **higher PC2** values indicate consumers who are **less responsive** to promotional campaigns, implying **more financial stability** and a lower reliance on discounts and sales events.

PC3: Luxury Preferences and Status-Oriented Spending

PC3 captures **luxury-oriented spending behaviors**, driven by a **strong positive** loading on **Spend_LuxuryGoods (0.8878)**. Consumers with **higher PC3** values prioritize **premium and status-driven** purchases, aligning with research indicating that luxury consumption signals wealth and social standing (Kapferer and Bastien, 2009). **Lower PC3** values indicate spending **focused on essentials** rather than indulgence, making cost-effective offerings more appealing.

PC4: Online vs. Offline Shopping Preference

PC4 differentiates **digital and traditional shopping behaviors**, as seen in the negative loading on **Purchases_Online (-0.5688)** and the positive loading on **Num_Dependents (0.4818)**. **Larger households** tend to **favour in store shopping**, where bulk purchases and physical interactions with products are preferred, while **smaller households** gravitate toward **e-commerce convenience**. This aligns with research suggesting that family size influences shopping preferences (Grewal and Roggeveen, 2017).

By leveraging these principal components, SmartFresh can **identify key behavioral trends** that **influence consumer spending**. For instance, customers with high PC3 scores may exhibit a stronger inclination toward discretionary purchases, whereas high PC2 scores could indicate higher responsiveness to promotions. These insights help **refine downstream segmentation**, enabling **more tailored marketing strategies** after clustering is applied.

4. K-means Clustering Analysis

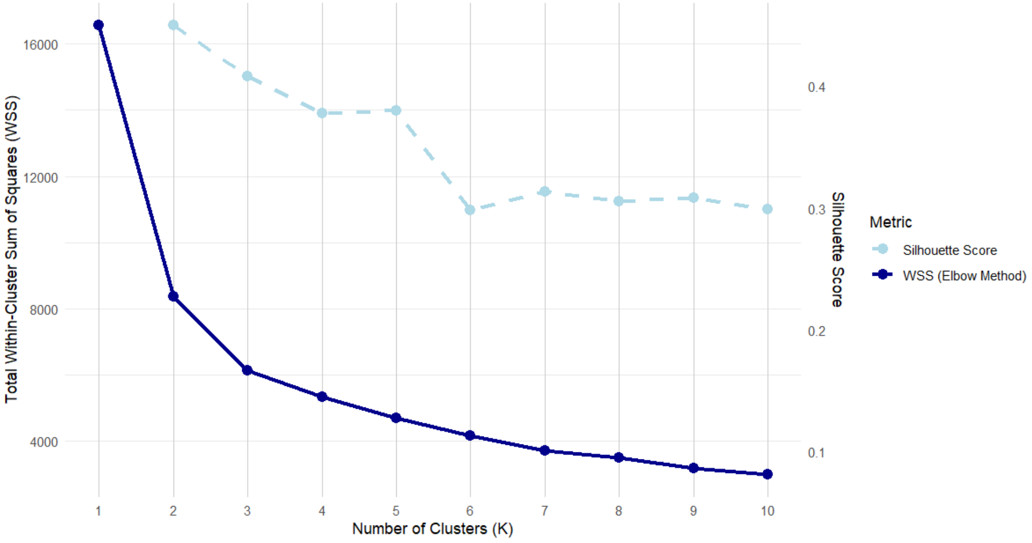


Figure 10: Elbow Analysis and Silhouette Score of k-means Clustering Analysis
R Code

The first four principal components (PCs) from PCA were selected for K-means clustering, capturing sufficient variance while reducing dimensionality. The elbow method suggests that both $k = 3$ and $k = 4$ are reasonable choices, as they are near the inflection point. However, the silhouette score, which measures how similar points are within clusters versus other clusters, is higher for $k = 3$, indicating more cohesive and well-separated clusters (Rousseeuw, 1987). A higher silhouette score suggests that data points are better grouped within their assigned clusters and more distinct from other clusters. Therefore, $k = 3$ is optimal, ensuring meaningful customer segmentation with minimal overlap between groups for targeted marketing strategies.

5. Customer Segmentation

Principal Components	Clusters and Clusters' centers		
	1 #579	2 #1043	3 #618
PC1	1.9430	-2.7678	-0.6860
PC2	0.3758	0.8042	-1.3877
PC3	0.0213	-0.0540	0.0146
PC4	0.0267	0.2073	-0.2393

Table 4: Clusters' centers corresponding to principal components
R Code

The k-means cluster centers analysis was conducted for the optimal $k = 3$. From the table above, the k-means clustering results indicate three distinct customer segments, each with unique characteristics based on their principal component (PC) scores. These distinct segments confirm that $k = 3$ was a statistically justified choice that also aligns with meaningful customer differentiation.

Cluster 1: High-Income, Broad Spending Consumers (579 customers)

Cluster 1 has a **strong positive loading on PC1 (1.9430)**, indicating **high-income, broad-spending customers** across multiple categories and channels. Their **moderate PC2** score **(0.3758)** suggests **mild promotion sensitivity**, though lower than Cluster 2. A **low but highest PC3** score **(0.0213)** indicates **slight discretionary spending**, but luxury shopping is not their defining trait. Their neutral PC4 score (0.0267) implies no strong preference for online or offline shopping.

Marketing Approach: A **premium engagement strategy** is suitable, including **VIP programs**, **early access** to luxury launches, and personalized experiences to strengthen brand loyalty. However, over-reliance on spending-based targeting **risks excluding lower-income but highly engaged customers** (Emma, 2024). Research suggests alternative engagement metrics, such as purchase frequency, loyalty behavior, and brand interaction, help create a fairer strategy (Wedel and Kannan, 2016). By integrating these insights, businesses can balance personalization with accessibility, fostering stronger customer relationships while ensuring segmentation remains both ethical and effective in driving long-term loyalty.

Cluster 2: Budget-Conscious, Promotion-Driven Consumers (1043 customers)

Cluster 2 has a **strong negative PC1 score (-2.7678)**, a **high positive PC2 score (0.8042)**, and a **moderate positive PC4 score (0.2073)**. This indicates that these consumers are **lower-income** but highly **responsive to promotions and discounts**. The positive PC4 score suggests they are more **likely to shop offline**, possibly in traditional brick-and-mortar stores rather than e-commerce platforms.

Marketing Approach: Since these customers are highly price-sensitive and responsive to promotions, a **discount-based marketing strategy** is **most effective**. Tactics such as coupon-based promotions, seasonal discounts, bulk purchase deals, and in-store loyalty programs will enhance engagement (Grewal and Roggeveen, 2017). Additionally, leveraging in-store experiences with personalized discounts for frequently purchased items can increase retention. Research suggests that limited-time offers and bundled deals are highly effective in engaging this segment (Ailawadi et al., 2001).

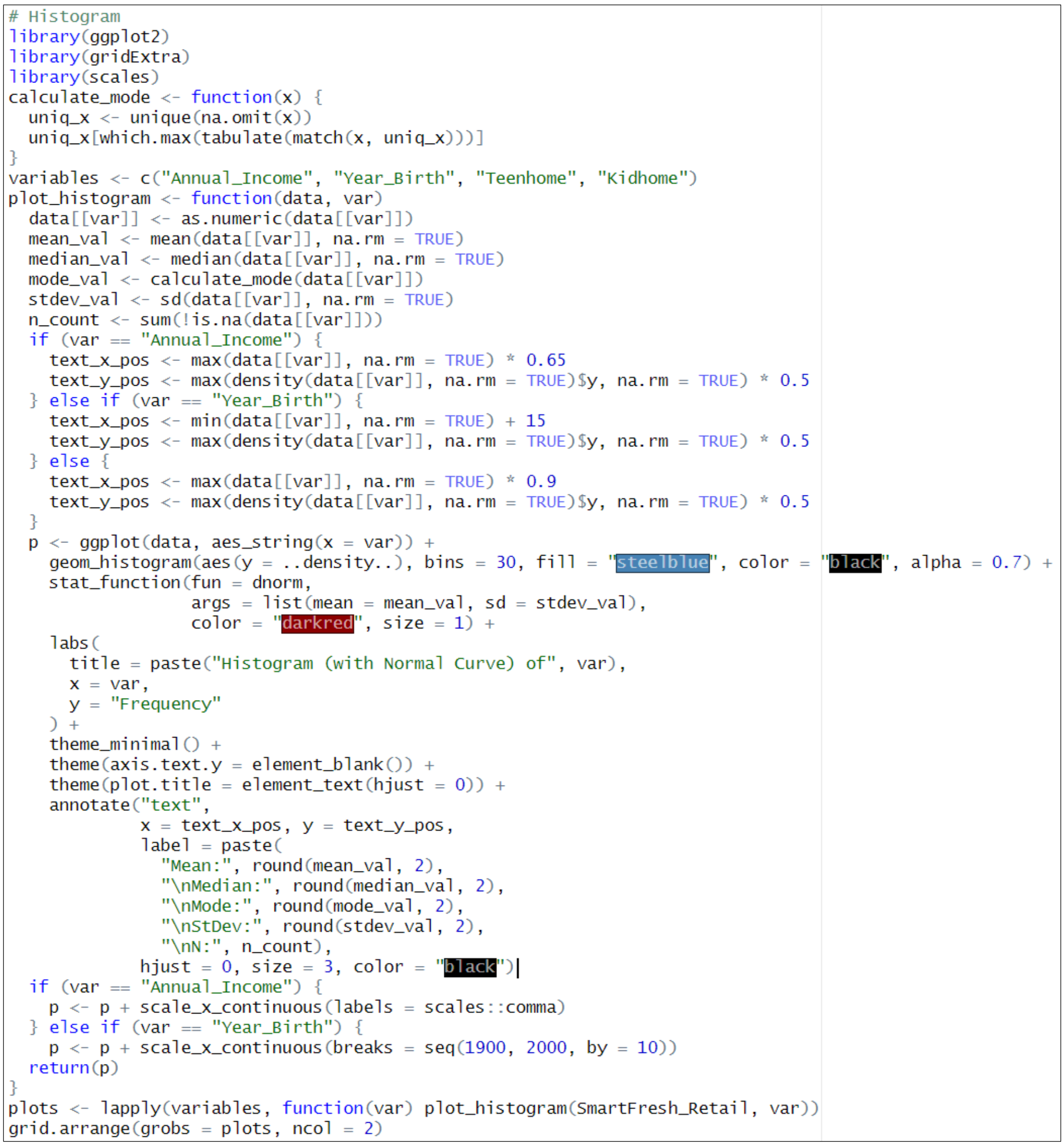
Cluster 3: Selective Shoppers with a Preference for Luxury (618 customers)

Cluster 3 exhibits **a negative PC1 score (-0.6860)**, a **strong negative PC2 score (-1.3877)**, a **neutral PC3 score (0.0146)**, and a **negative PC4 score (-0.2393)**. This suggests that they are **not broad spenders**, **nor** are they particularly **promotion-sensitive**. The negative PC4 score implies a **preference for online shopping**, meaning they may **engage more in e-commerce** purchases rather than in-store visits.

Marketing Approach: This group consists of **digital-savvy**, selective shoppers who **prioritize quality over quantity**. A focus on luxury product positioning, personalized **online experiences**, and premium product bundling will work best (Kapferer and Bastien, 2009). **Targeted marketing through social media ads**, influencer collaborations, and personalized email campaigns can effectively reach them. Additionally, exclusive **online-only offers** and high-end product experiences will appeal to this segment (Nadine et al., 2024). Luxury e-commerce brands have successfully engaged such customers through customized recommendations and seamless omnichannel experiences (Chandon et al., 2016).

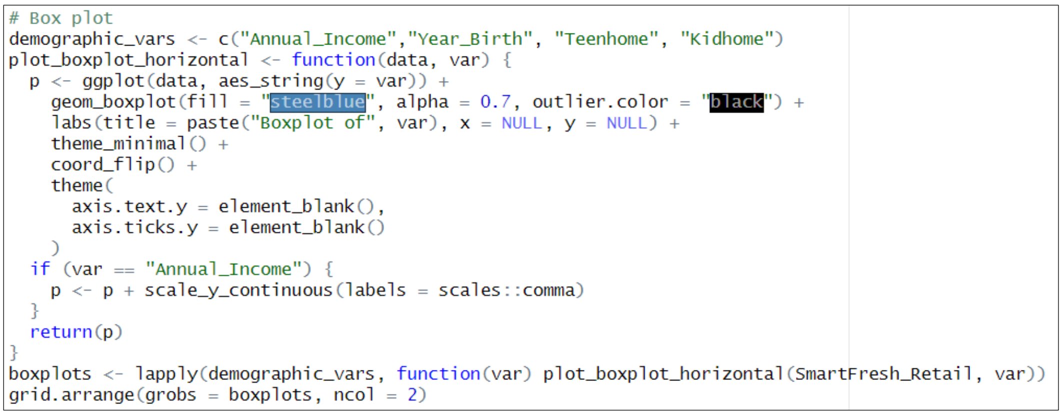
APPENDIX

I. EXPLORATORY DATA ANALYSIS (EDA)



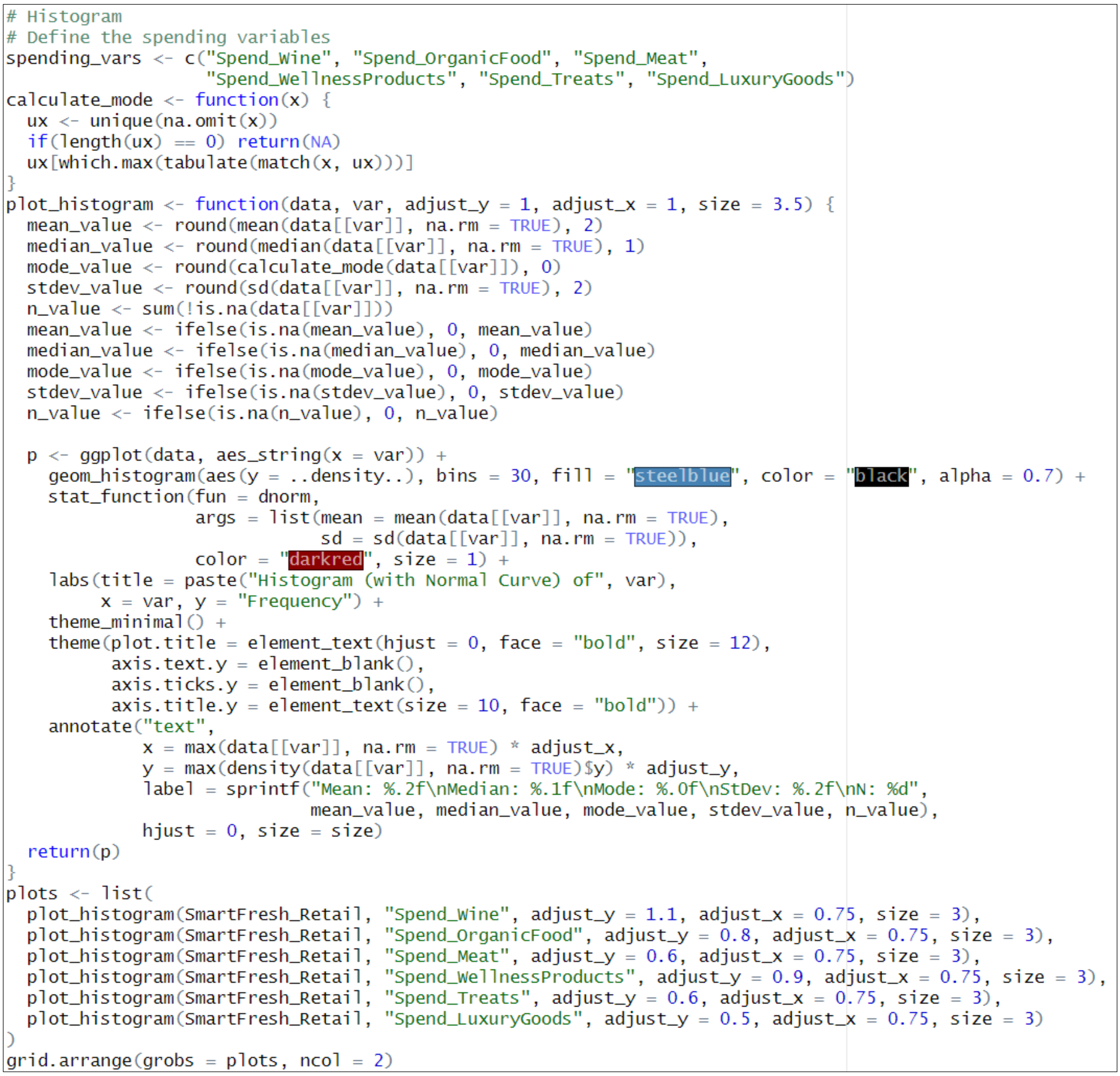
```r
# Histogram
library(ggplot2)
library(gridExtra)
library(scales)
calculate_mode <- function(x) {
  uniq_x <- unique(na.omit(x))
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}
variables <- c("Annual_Income", "Year_Birth", "Teenhome", "Kidhome")
plot_histogram <- function(data, var)
  data[[var]] <- as.numeric(data[[var]])
  mean_val <- mean(data[[var]], na.rm = TRUE)
  median_val <- median(data[[var]], na.rm = TRUE)
  mode_val <- calculate_mode(data[[var]])
  stdev_val <- sd(data[[var]], na.rm = TRUE)
  n_count <- sum(!is.na(data[[var]]))
  if (var == "Annual_Income") {
    text_x_pos <- max(data[[var]], na.rm = TRUE) * 0.65
    text_y_pos <- max(density(data[[var]], na.rm = TRUE)$y, na.rm = TRUE) * 0.5
  } else if (var == "Year_Birth") {
    text_x_pos <- min(data[[var]], na.rm = TRUE) + 15
    text_y_pos <- max(density(data[[var]], na.rm = TRUE)$y, na.rm = TRUE) * 0.5
  } else {
    text_x_pos <- max(data[[var]], na.rm = TRUE) * 0.9
    text_y_pos <- max(density(data[[var]], na.rm = TRUE)$y, na.rm = TRUE) * 0.5
  }
  p <- ggplot(data, aes_string(x = var)) +
    geom_histogram(aes(y = ..density..), bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
    stat_function(fun = dnorm,
                  args = list(mean = mean_val, sd = stdev_val),
                  color = "darkred", size = 1) +
    labs(
      title = paste("Histogram (with Normal Curve) of", var),
      x = var,
      y = "Frequency"
    ) +
    theme_minimal() +
    theme(axis.text.y = element_blank()) +
    theme(plot.title = element_text(hjust = 0)) +
    annotate("text",
             x = text_x_pos, y = text_y_pos,
             label = paste(
               "Mean:", round(mean_val, 2),
               "\nMedian:", round(median_val, 2),
               "\nMode:", round(mode_val, 2),
               "\nStDev:", round(stdev_val, 2),
               "\nN:", n_count),
             hjust = 0, size = 3, color = "black")
  if (var == "Annual_Income") {
    p <- p + scale_x_continuous(labels = scales::comma)
  } else if (var == "Year_Birth") {
    p <- p + scale_x_continuous(breaks = seq(1900, 2000, by = 10))
  return(p)
}
plots <- lapply(variables, function(var) plot_histogram(SmartFresh_Retail, var))
grid.arrange(grobs = plots, ncol = 2)
```

Figure 11: Histograms - Customer Demographic Variables



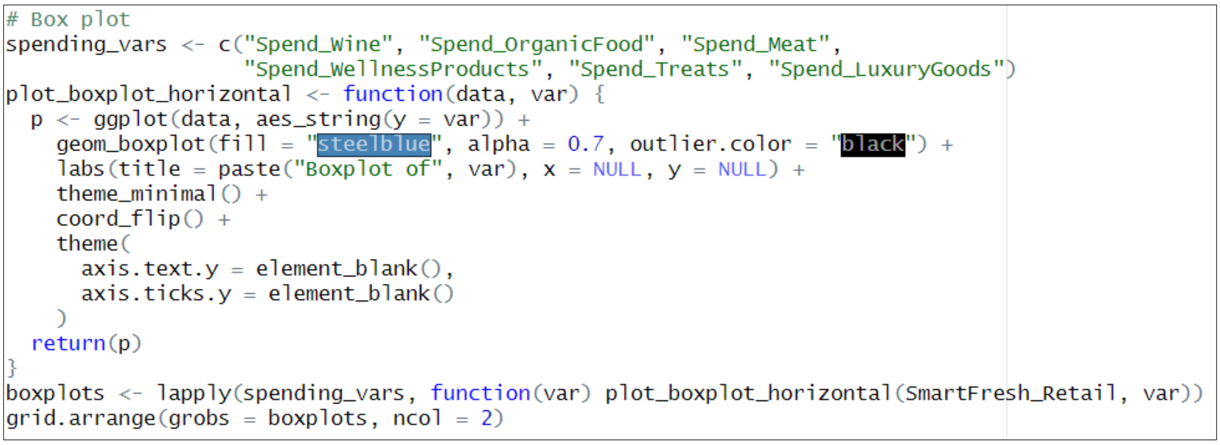
```r
# Box plot
demographic_vars <- c("Annual_Income","Year_Birth", "Teenhome", "Kidhome")
plot_boxplot_horizontal <- function(data, var) {
  p <- ggplot(data, aes_string(y = var)) +
    geom_boxplot(fill = "steelblue", alpha = 0.7, outlier.color = "black") +
    labs(title = paste("Boxplot of", var), x = NULL, y = NULL) +
    theme_minimal() +
    coord_flip() +
    theme(
      axis.text.y = element_blank(),
      axis.ticks.y = element_blank()
    )
  if (var == "Annual_Income") {
    p <- p + scale_y_continuous(labels = scales::comma)
  }
  return(p)
}
boxplots <- lapply(demographic_vars, function(var) plot_boxplot_horizontal(SmartFresh_Retail, var))
grid.arrange(grobs = boxplots, ncol = 2)
```

Figure 12: Boxplots - Customer Demographic Variables



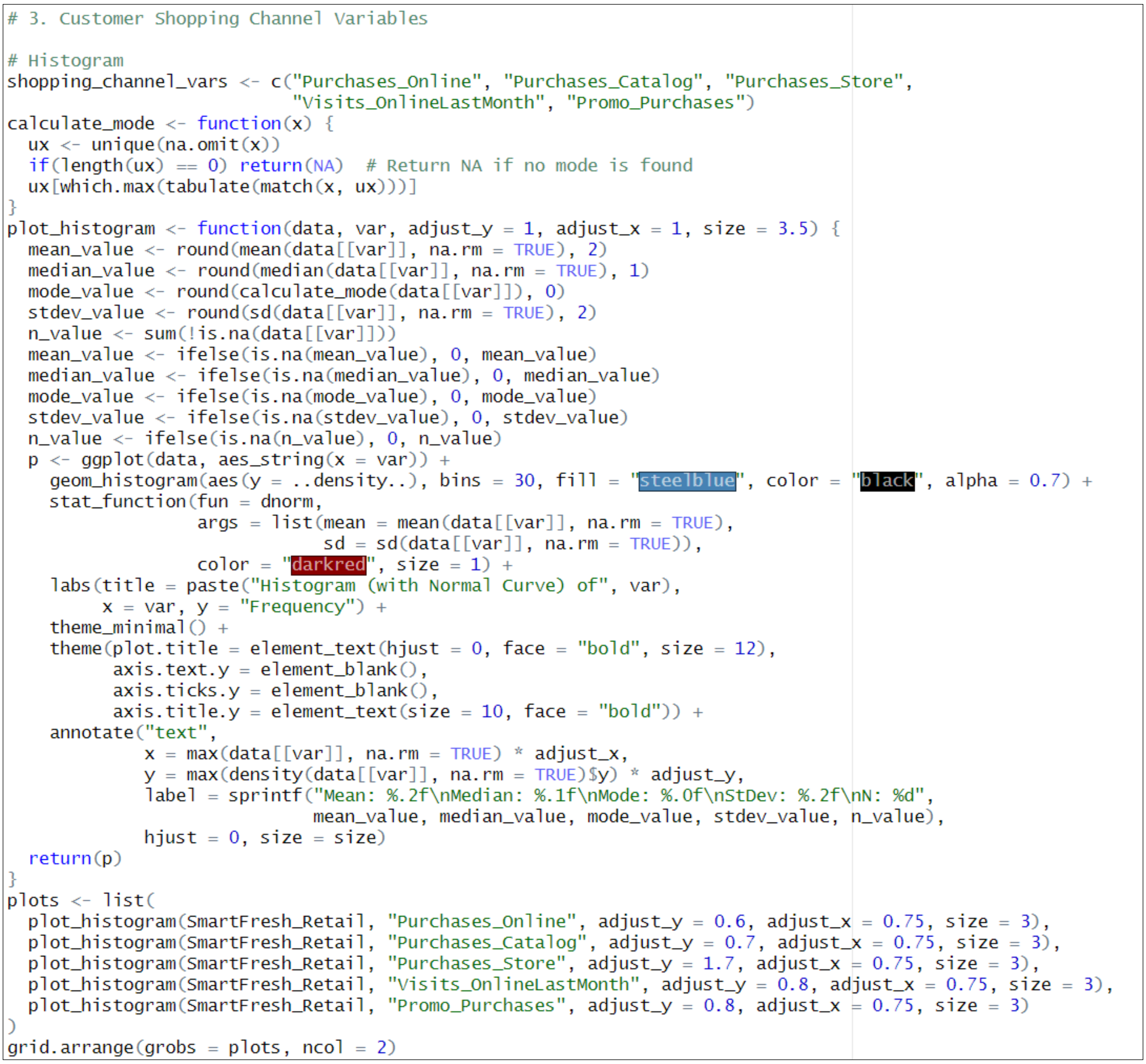
```r
# Histogram
# Define the spending variables
spending_vars <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat",
                   "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")
calculate_mode <- function(x) {
  ux <- unique(na.omit(x))
  if(length(ux) == 0) return(NA)
  ux[which.max(tabulate(match(x, ux)))]
}
plot_histogram <- function(data, var, adjust_y = 1, adjust_x = 1, size = 3.5) {
  mean_value <- round(mean(data[[var]], na.rm = TRUE), 2)
  median_value <- round(median(data[[var]], na.rm = TRUE), 1)
  mode_value <- round(calculate_mode(data[[var]]), 0)
  stdev_value <- round(sd(data[[var]], na.rm = TRUE), 2)
  n_value <- sum(!is.na(data[[var]]))
  mean_value <- ifelse(is.na(mean_value), 0, mean_value)
  median_value <- ifelse(is.na(median_value), 0, median_value)
  mode_value <- ifelse(is.na(mode_value), 0, mode_value)
  stdev_value <- ifelse(is.na(stdev_value), 0, stdev_value)
  n_value <- ifelse(is.na(n_value), 0, n_value)

  p <- ggplot(data, aes_string(x = var)) +
    geom_histogram(aes(y = ..density..), bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
    stat_function(fun = dnorm,
                  args = list(mean = mean(data[[var]], na.rm = TRUE),
                              sd = sd(data[[var]], na.rm = TRUE)),
                  color = "darkred", size = 1) +
    labs(title = paste("Histogram (with Normal Curve) of", var),
         x = var, y = "Frequency") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0, face = "bold", size = 12),
          axis.text.y = element_blank(),
          axis.ticks.y = element_blank(),
          axis.title.y = element_text(size = 10, face = "bold")) +
    annotate("text",
             x = max(data[[var]], na.rm = TRUE) * adjust_x,
             y = max(density(data[[var]], na.rm = TRUE)$y) * adjust_y,
             label = sprintf("Mean: %.2f\nMedian: %.1f\nMode: %.0f\nStDev: %.2f\nN: %d",
                             mean_value, median_value, mode_value, stdev_value, n_value),
             hjust = 0, size = size)
  return(p)
}
plots <- list(
  plot_histogram(SmartFresh_Retail, "Spend_Wine", adjust_y = 1.1, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Spend_OrganicFood", adjust_y = 0.8, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Spend_Meat", adjust_y = 0.6, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Spend_WellnessProducts", adjust_y = 0.9, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Spend_Treats", adjust_y = 0.6, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Spend_LuxuryGoods", adjust_y = 0.5, adjust_x = 0.75, size = 3)
)
grid.arrange(grobs = plots, ncol = 2)
```

Figure 13: Histograms - Customer Spending Pattern Variables



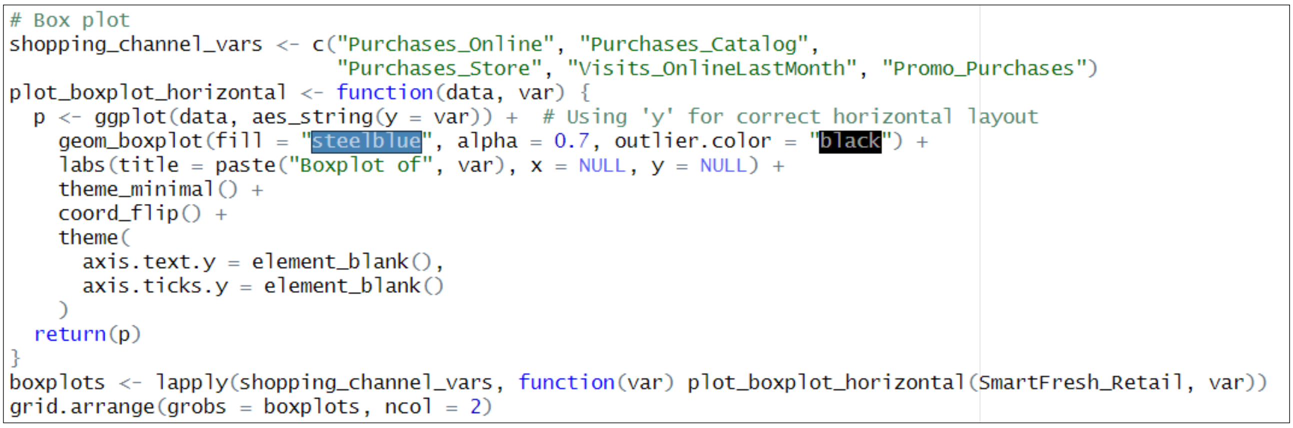
```r
# Box plot
spending_vars <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat",
                   "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")
plot_boxplot_horizontal <- function(data, var) {
  p <- ggplot(data, aes_string(y = var)) +
    geom_boxplot(fill = "steelblue", alpha = 0.7, outlier.color = "black") +
    labs(title = paste("Boxplot of", var), x = NULL, y = NULL) +
    theme_minimal() +
    coord_flip() +
    theme(
      axis.text.y = element_blank(),
      axis.ticks.y = element_blank()
    )
  return(p)
}
boxplots <- lapply(spending_vars, function(var) plot_boxplot_horizontal(SmartFresh_Retail, var))
grid.arrange(grobs = boxplots, ncol = 2)
```

Figure 14: Boxplots - Customer Spending Pattern Variables



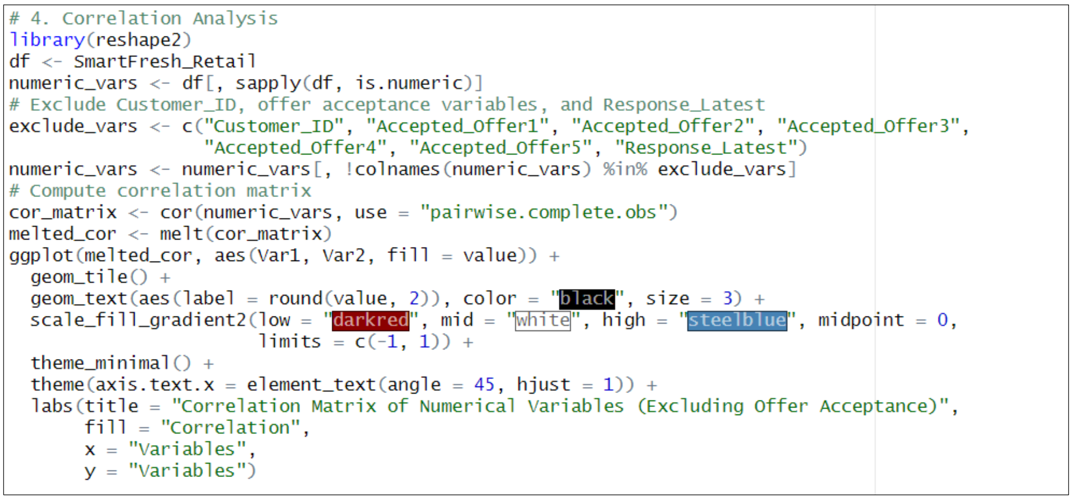
```
# 3. Customer Shopping Channel Variables

# Histogram
shopping_channel_vars <- c("Purchases_Online", "Purchases_Catalog", "Purchases_Store",
                           "Visits_OnlineLastMonth", "Promo_Purchases")
calculate_mode <- function(x) {
  ux <- unique(na.omit(x))
  if(length(ux) == 0) return(NA)  # Return NA if no mode is found
  ux[which.max(tabulate(match(x, ux)))]
}
plot_histogram <- function(data, var, adjust_y = 1, adjust_x = 1, size = 3.5) {
  mean_value <- round(mean(data[[var]], na.rm = TRUE), 2)
  median_value <- round(median(data[[var]], na.rm = TRUE), 1)
  mode_value <- round(calculate_mode(data[[var]]), 0)
  stdev_value <- round(sd(data[[var]], na.rm = TRUE), 2)
  n_value <- sum(!is.na(data[[var]]))
  mean_value <- ifelse(is.na(mean_value), 0, mean_value)
  median_value <- ifelse(is.na(median_value), 0, median_value)
  mode_value <- ifelse(is.na(mode_value), 0, mode_value)
  stdev_value <- ifelse(is.na(stdev_value), 0, stdev_value)
  n_value <- ifelse(is.na(n_value), 0, n_value)
  p <- ggplot(data, aes_string(x = var)) +
    geom_histogram(aes(y = ..density..), bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
    stat_function(fun = dnorm,
                  args = list(mean = mean(data[[var]], na.rm = TRUE),
                              sd = sd(data[[var]], na.rm = TRUE)),
                  color = "darkred", size = 1) +
    labs(title = paste("Histogram (with Normal Curve) of", var),
         x = var, y = "Frequency") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0, face = "bold", size = 12),
          axis.text.y = element_blank(),
          axis.ticks.y = element_blank(),
          axis.title.y = element_text(size = 10, face = "bold")) +
    annotate("text",
             x = max(data[[var]], na.rm = TRUE) * adjust_x,
             y = max(density(data[[var]], na.rm = TRUE)$y) * adjust_y,
             label = sprintf("Mean: %.2f\nMedian: %.1f\nMode: %.0f\nStDev: %.2f\nN: %d",
                             mean_value, median_value, mode_value, stdev_value, n_value),
             hjust = 0, size = size)
  return(p)
}
plots <- list(
  plot_histogram(SmartFresh_Retail, "Purchases_Online", adjust_y = 0.6, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Purchases_Catalog", adjust_y = 0.7, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Purchases_Store", adjust_y = 1.7, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Visits_OnlineLastMonth", adjust_y = 0.8, adjust_x = 0.75, size = 3),
  plot_histogram(SmartFresh_Retail, "Promo_Purchases", adjust_y = 0.8, adjust_x = 0.75, size = 3)
)
grid.arrange(grobs = plots, ncol = 2)
```

Figure 15: Histograms - Customer Shopping Channel Variables



```
# Box plot
shopping_channel_vars <- c("Purchases_Online", "Purchases_Catalog",
                           "Purchases_Store", "Visits_OnlineLastMonth", "Promo_Purchases")
plot_boxplot_horizontal <- function(data, var) {
  p <- ggplot(data, aes_string(y = var)) +  # Using 'y' for correct horizontal layout
    geom_boxplot(fill = "steelblue", alpha = 0.7, outlier.color = "black") +
    labs(title = paste("Boxplot of", var), x = NULL, y = NULL) +
    theme_minimal() +
    coord_flip() +
    theme(
      axis.text.y = element_blank(),
      axis.ticks.y = element_blank()
    )
  return(p)
}
boxplots <- lapply(shopping_channel_vars, function(var) plot_boxplot_horizontal(SmartFresh_Retail, var))
grid.arrange(grobs = boxplots, ncol = 2)
```

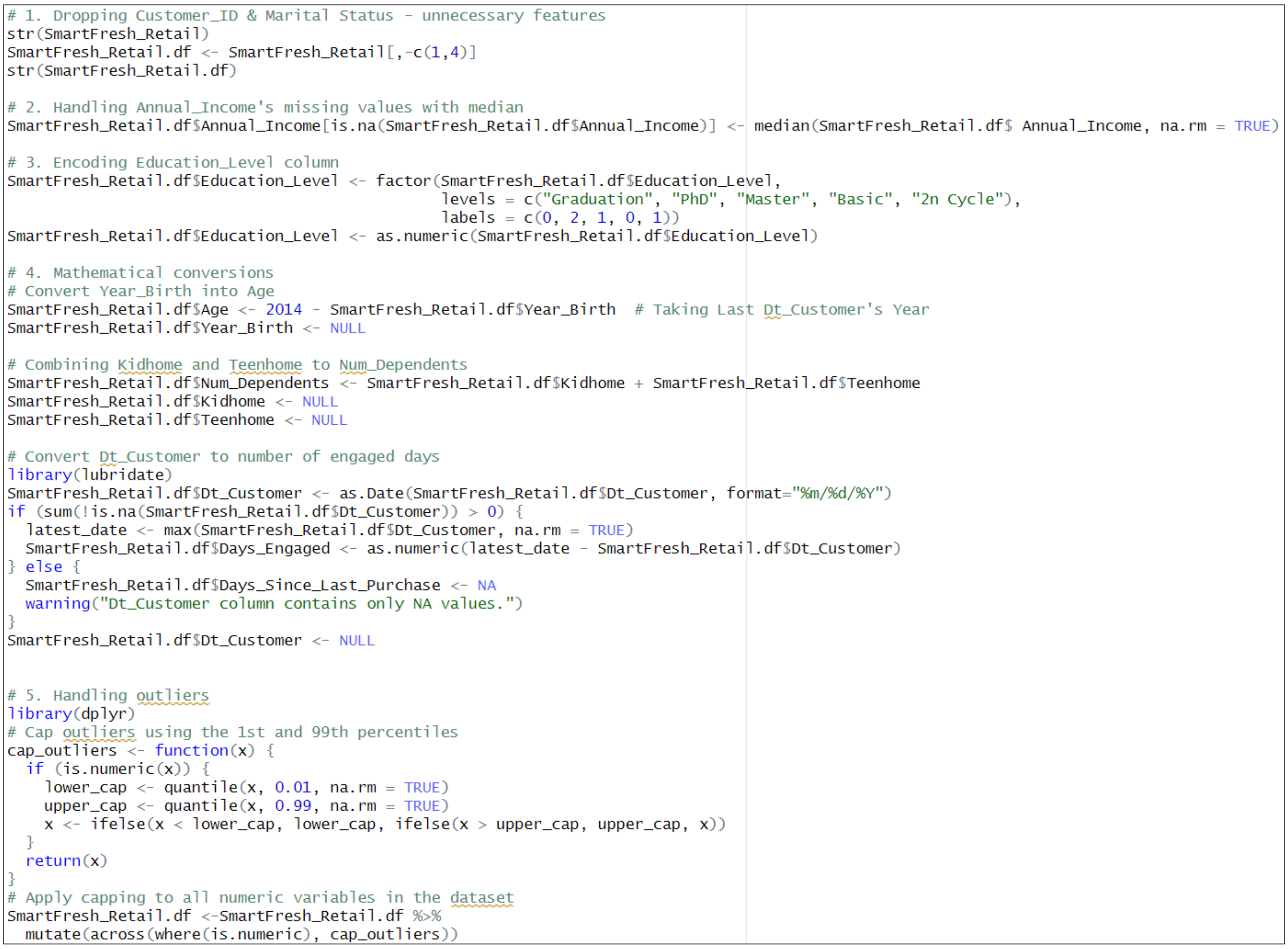
Figure 16: Boxplots - Customer Shopping Channel Variables



```r
# 4. Correlation Analysis
library(reshape2)
df <- SmartFresh_Retail
numeric_vars <- df[, sapply(df, is.numeric)]
# Exclude Customer_ID, offer acceptance variables, and Response_Latest
exclude_vars <- c("Customer_ID", "Accepted_Offer1", "Accepted_Offer2", "Accepted_Offer3",
                  "Accepted_Offer4", "Accepted_Offer5", "Response_Latest")
numeric_vars <- numeric_vars[, !colnames(numeric_vars) %in% exclude_vars]
# Compute correlation matrix
cor_matrix <- cor(numeric_vars, use = "pairwise.complete.obs")
melted_cor <- melt(cor_matrix)
ggplot(melted_cor, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  scale_fill_gradient2(low = "darkred", mid = "white", high = "steelblue", midpoint = 0,
                       limits = c(-1, 1)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Correlation Matrix of Numerical Variables (Excluding Offer Acceptance)",
       fill = "Correlation",
       x = "Variables",
       y = "Variables")
```

Figure 17: EDA - Correlation Analysis

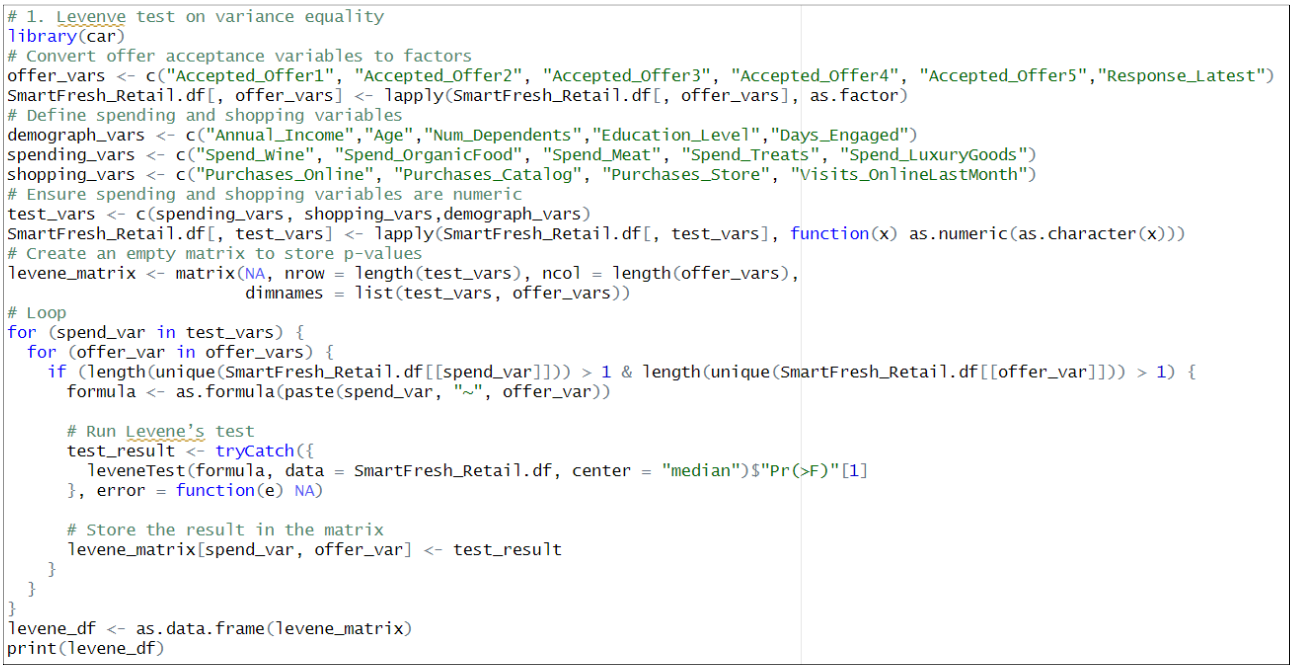
II. DATA PRE-PROCESSING



```r
# 1. Dropping Customer_ID & Marital Status - unnecessary features
str(SmartFresh_Retail)
SmartFresh_Retail.df <- SmartFresh_Retail[,-c(1,4)]
str(SmartFresh_Retail.df)

# 2. Handling Annual_Income's missing values with median
SmartFresh_Retail.df$Annual_Income[is.na(SmartFresh_Retail.df$Annual_Income)] <- median(SmartFresh_Retail.df$ Annual_Income, na.rm = TRUE)

# 3. Encoding Education_Level column
SmartFresh_Retail.df$Education_Level <- factor(SmartFresh_Retail.df$Education_Level,
                                               levels = c("Graduation", "PhD", "Master", "Basic", "2n Cycle"),
                                               labels = c(0, 2, 1, 0, 1))
SmartFresh_Retail.df$Education_Level <- as.numeric(SmartFresh_Retail.df$Education_Level)

# 4. Mathematical conversions
# Convert Year_Birth into Age
SmartFresh_Retail.df$Age <- 2014 - SmartFresh_Retail.df$Year_Birth  # Taking Last Dt_Customer's Year
SmartFresh_Retail.df$Year_Birth <- NULL

# Combining Kidhome and Teenhome to Num_Dependents
SmartFresh_Retail.df$Num_Dependents <- SmartFresh_Retail.df$Kidhome + SmartFresh_Retail.df$Teenhome
SmartFresh_Retail.df$Kidhome <- NULL
SmartFresh_Retail.df$Teenhome <- NULL

# Convert Dt_Customer to number of engaged days
library(lubridate)
SmartFresh_Retail.df$Dt_Customer <- as.Date(SmartFresh_Retail.df$Dt_Customer, format="%m/%d/%Y")
if (sum(!is.na(SmartFresh_Retail.df$Dt_Customer)) > 0) {
  latest_date <- max(SmartFresh_Retail.df$Dt_Customer, na.rm = TRUE)
  SmartFresh_Retail.df$Days_Engaged <- as.numeric(latest_date - SmartFresh_Retail.df$Dt_Customer)
} else {
  SmartFresh_Retail.df$Days_Since_Last_Purchase <- NA
  warning("Dt_Customer column contains only NA values.")
}
SmartFresh_Retail.df$Dt_Customer <- NULL


# 5. Handling outliers
library(dplyr)
# Cap outliers using the 1st and 99th percentiles
cap_outliers <- function(x) {
  if (is.numeric(x)) {
    lower_cap <- quantile(x, 0.01, na.rm = TRUE)
    upper_cap <- quantile(x, 0.99, na.rm = TRUE)
    x <- ifelse(x < lower_cap, lower_cap, ifelse(x > upper_cap, upper_cap, x))
  }
  return(x)
}
# Apply capping to all numeric variables in the dataset
SmartFresh_Retail.df <-SmartFresh_Retail.df %>%
  mutate(across(where(is.numeric), cap_outliers))
```

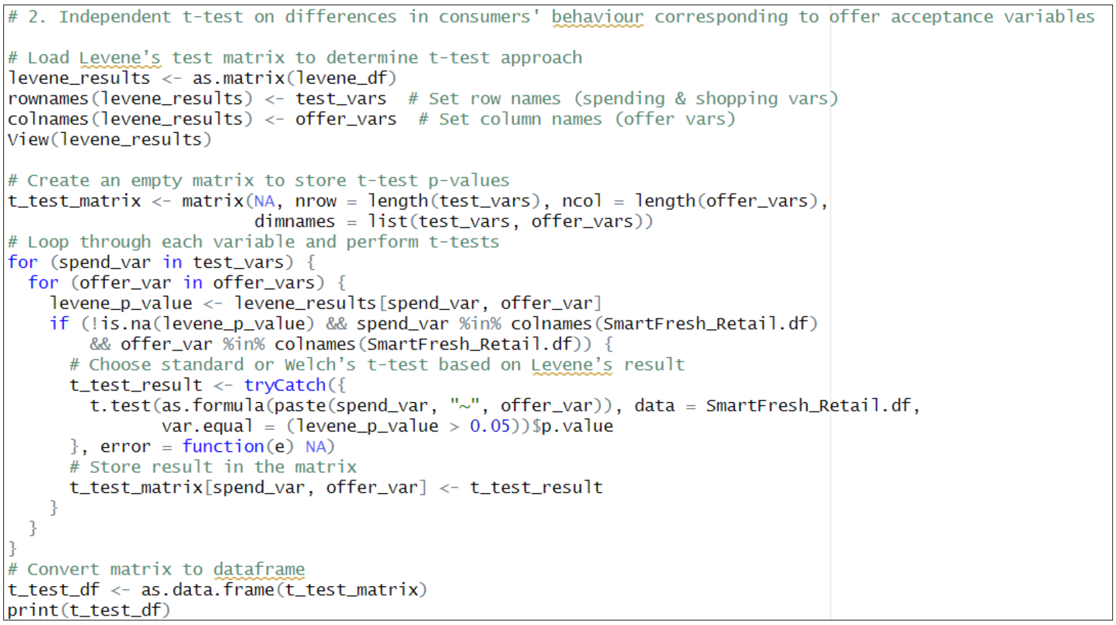
Figure 18: Data - Preprocessing Steps

III. DATA PROCESSING



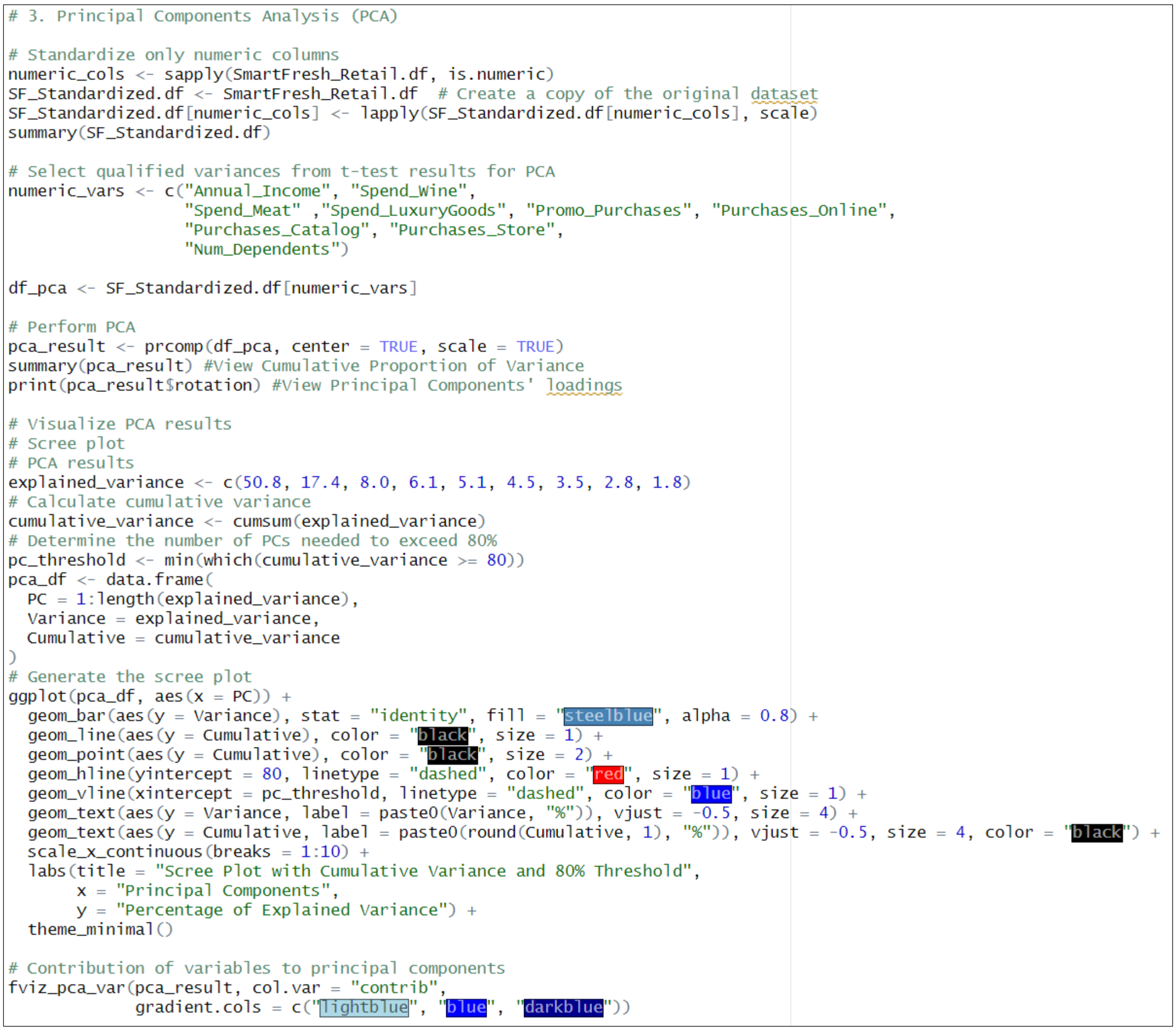
```
# 1. Levenve test on variance equality
library(car)
# Convert offer acceptance variables to factors
offer_vars <- c("Accepted_Offer1", "Accepted_Offer2", "Accepted_Offer3", "Accepted_Offer4", "Accepted_Offer5","Response_Latest")
SmartFresh_Retail.df[, offer_vars] <- lapply(SmartFresh_Retail.df[, offer_vars], as.factor)
# Define spending and shopping variables
demograph_vars <- c("Annual_Income","Age","Num_Dependents","Education_Level","Days_Engaged")
spending_vars <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat", "Spend_Treats", "Spend_LuxuryGoods")
shopping_vars <- c("Purchases_Online", "Purchases_Catalog", "Purchases_Store", "Visits_OnlineLastMonth")
# Ensure spending and shopping variables are numeric
test_vars <- c(spending_vars, shopping_vars,demograph_vars)
SmartFresh_Retail.df[, test_vars] <- lapply(SmartFresh_Retail.df[, test_vars], function(x) as.numeric(as.character(x)))
# Create an empty matrix to store p-values
levene_matrix <- matrix(NA, nrow = length(test_vars), ncol = length(offer_vars),
                        dimnames = list(test_vars, offer_vars))
# Loop
for (spend_var in test_vars) {
  for (offer_var in offer_vars) {
    if (length(unique(SmartFresh_Retail.df[[spend_var]])) > 1 & length(unique(SmartFresh_Retail.df[[offer_var]])) > 1) {
      formula <- as.formula(paste(spend_var, "~", offer_var))

      # Run Levene’s test
      test_result <- tryCatch({
        leveneTest(formula, data = SmartFresh_Retail.df, center = "median")$"Pr(>F)"[1]
      }, error = function(e) NA)

      # Store the result in the matrix
      levene_matrix[spend_var, offer_var] <- test_result
    }
  }
}
levene_df <- as.data.frame(levene_matrix)
print(levene_df)
```

Figure 19: Levene test on variance equality



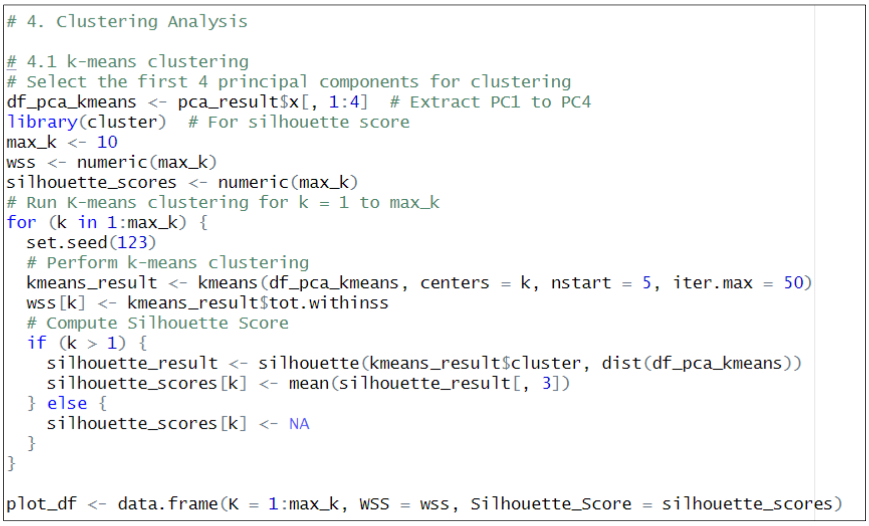
```
# 2. Independent t-test on differences in consumers' behaviour corresponding to offer acceptance variables

# Load Levene’s test matrix to determine t-test approach
levene_results <- as.matrix(levene_df)
rownames(levene_results) <- test_vars  # Set row names (spending & shopping vars)
colnames(levene_results) <- offer_vars  # Set column names (offer vars)
View(levene_results)

# Create an empty matrix to store t-test p-values
t_test_matrix <- matrix(NA, nrow = length(test_vars), ncol = length(offer_vars),
                        dimnames = list(test_vars, offer_vars))
# Loop through each variable and perform t-tests
for (spend_var in test_vars) {
  for (offer_var in offer_vars) {
    levene_p_value <- levene_results[spend_var, offer_var]
    if (!is.na(levene_p_value) && spend_var %in% colnames(SmartFresh_Retail.df)
        && offer_var %in% colnames(SmartFresh_Retail.df)) {
      # Choose standard or Welch’s t-test based on Levene’s result
      t_test_result <- tryCatch({
        t.test(as.formula(paste(spend_var, "~", offer_var)), data = SmartFresh_Retail.df,
               var.equal = (levene_p_value > 0.05))$p.value
      }, error = function(e) NA)
      # Store result in the matrix
      t_test_matrix[spend_var, offer_var] <- t_test_result
    }
  }
}
# Convert matrix to dataframe
t_test_df <- as.data.frame(t_test_matrix)
print(t_test_df)
```

Figure 20: Independent t-test based on Levene test's results



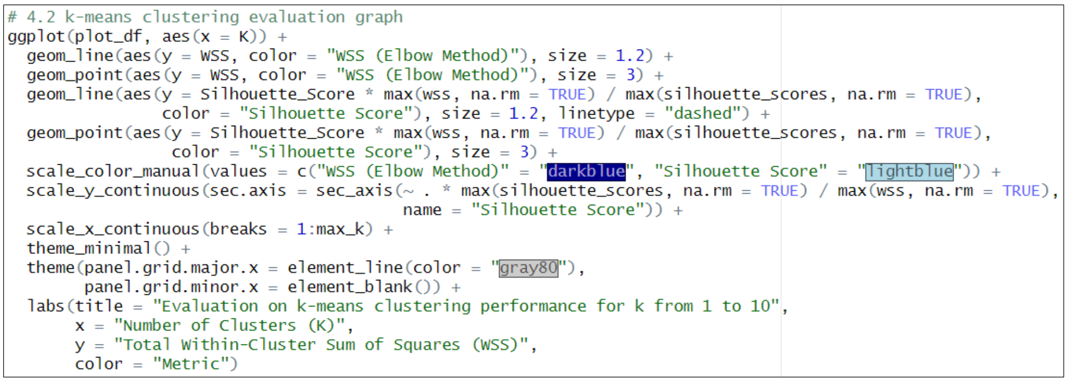
```
# 3. Principal Components Analysis (PCA)

# Standardize only numeric columns
numeric_cols <- sapply(SmartFresh_Retail.df, is.numeric)
SF_Standardized.df <- SmartFresh_Retail.df  # Create a copy of the original dataset
SF_Standardized.df[numeric_cols] <- lapply(SF_Standardized.df[numeric_cols], scale)
summary(SF_Standardized.df)

# Select qualified variances from t-test results for PCA
numeric_vars <- c("Annual_Income", "Spend_Wine",
                  "Spend_Meat" ,"Spend_LuxuryGoods", "Promo_Purchases", "Purchases_Online",
                  "Purchases_Catalog", "Purchases_Store",
                  "Num_Dependents")

df_pca <- SF_Standardized.df[numeric_vars]

# Perform PCA
pca_result <- prcomp(df_pca, center = TRUE, scale = TRUE)
summary(pca_result) #View Cumulative Proportion of Variance
print(pca_result$rotation) #View Principal Components' loadings

# Visualize PCA results
# Scree plot
# PCA results
explained_variance <- c(50.8, 17.4, 8.0, 6.1, 5.1, 4.5, 3.5, 2.8, 1.8)
# Calculate cumulative variance
cumulative_variance <- cumsum(explained_variance)
# Determine the number of PCs needed to exceed 80%
pc_threshold <- min(which(cumulative_variance >= 80))
pca_df <- data.frame(
  PC = 1:length(explained_variance),
  Variance = explained_variance,
  Cumulative = cumulative_variance
)
# Generate the scree plot
ggplot(pca_df, aes(x = PC)) +
  geom_bar(aes(y = Variance), stat = "identity", fill = "steelblue", alpha = 0.8) +
  geom_line(aes(y = Cumulative), color = "black", size = 1) +
  geom_point(aes(y = Cumulative), color = "black", size = 2) +
  geom_hline(yintercept = 80, linetype = "dashed", color = "red", size = 1) +
  geom_vline(xintercept = pc_threshold, linetype = "dashed", color = "blue", size = 1) +
  geom_text(aes(y = Variance, label = paste0(Variance, "%")), vjust = -0.5, size = 4) +
  geom_text(aes(y = Cumulative, label = paste0(round(Cumulative, 1), "%")), vjust = -0.5, size = 4, color = "black") +
  scale_x_continuous(breaks = 1:10) +
  labs(title = "Scree Plot with Cumulative Variance and 80% Threshold",
       x = "Principal Components",
       y = "Percentage of Explained Variance") +
  theme_minimal()

# Contribution of variables to principal components
fviz_pca_var(pca_result, col.var = "contrib",
             gradient.cols = c("lightblue", "blue", "darkblue"))
```

Figure 21: Principal Components Analysis



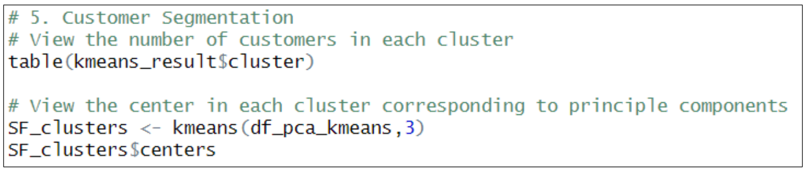
```
# 4. Clustering Analysis

# 4.1 k-means clustering
# Select the first 4 principal components for clustering
df_pca_kmeans <- pca_result$x[, 1:4]  # Extract PC1 to PC4
library(cluster)  # For silhouette score
max_k <- 10
wss <- numeric(max_k)
silhouette_scores <- numeric(max_k)
# Run K-means clustering for k = 1 to max_k
for (k in 1:max_k) {
  set.seed(123)
  # Perform k-means clustering
  kmeans_result <- kmeans(df_pca_kmeans, centers = k, nstart = 5, iter.max = 50)
  wss[k] <- kmeans_result$tot.withinss
  # Compute Silhouette Score
  if (k > 1) {
    silhouette_result <- silhouette(kmeans_result$cluster, dist(df_pca_kmeans))
    silhouette_scores[k] <- mean(silhouette_result[, 3])
  } else {
    silhouette_scores[k] <- NA
  }
}

plot_df <- data.frame(K = 1:max_k, WSS = wss, Silhouette_Score = silhouette_scores)
```

Figure 22: K-means Clustering Analysis



```
# 4.2 k-means clustering evaluation graph
ggplot(plot_df, aes(x = K)) +
  geom_line(aes(y = WSS, color = "WSS (Elbow Method)"), size = 1.2) +
  geom_point(aes(y = WSS, color = "WSS (Elbow Method)"), size = 3) +
  geom_line(aes(y = Silhouette_Score * max(wss, na.rm = TRUE) / max(silhouette_scores, na.rm = TRUE),
                color = "Silhouette Score"), size = 1.2, linetype = "dashed") +
  geom_point(aes(y = Silhouette_Score * max(wss, na.rm = TRUE) / max(silhouette_scores, na.rm = TRUE),
                 color = "Silhouette Score"), size = 3) +
  scale_color_manual(values = c("WSS (Elbow Method)" = "darkblue", "Silhouette Score" = "lightblue")) +
  scale_y_continuous(sec.axis = sec_axis(~ . * max(silhouette_scores, na.rm = TRUE) / max(wss, na.rm = TRUE),
                                         name = "Silhouette Score")) +
  scale_x_continuous(breaks = 1:max_k) +
  theme_minimal() +
  theme(panel.grid.major.x = element_line(color = "gray80"),
        panel.grid.minor.x = element_blank()) +
  labs(title = "Evaluation on k-means clustering performance for k from 1 to 10",
       x = "Number of Clusters (K)",
       y = "Total Within-Cluster Sum of Squares (WSS)",
       color = "Metric")
```

Figure 23: K-means Clustering Evaluation



```
# 5. Customer Segmentation
# View the number of customers in each cluster
table(kmeans_result$cluster)

# View the center in each cluster corresponding to principle components
SF_clusters <- kmeans(df_pca_kmeans,3)
SF_clusters$centers
```

Figure 24: Customer segmentation

LIST OF REFERENCES

1. Ailawadi, K., Neslin, S. and Gedenk, K. (2001) Pursuing the Value-Conscious Consumer: Store Brands Versus National Brand Promotions. *Journal of Marketing - J MARKETING*, 65: 71–89. doi:10.1509/jmkg.65.1.71.18132.
2. Chandon, J.L., Laurent, G. and Valette-Florence, P. (2016) Pursuing the concept of luxury: Introduction to the JBR Special Issue on “Luxury Marketing from Tradition to Innovation.” *Journal of Business Research*, 69 (1): 299–303. doi:10.1016/j.jbusres.2015.08.001.
3. Delacre, M., Lakens, D. and Leys, C. (2017) Why Psychologists Should by Default Use Welch’s t-test Instead of Student’s t-test. *International Review of Social Psychology*. doi:10.5334/irsp.82.
4. Emma, J. (2024) *The ethics of market segmentation*. Available at: https://www.internationalbunch.com/post/the-ethics-of-market-segmentation (Accessed: 16 March 2025).
5. Grewal, D. and Roggeveen, A. (2017) The Future of Retailing. *Journal of Retailing*, 93. doi:10.1016/j.jretai.2016.12.008.
6. Hair, J.F., Black, W.C., Babin, B.J., et al. (2019) *Multivariate Data Analysis Eighth Edition*. Available at: www.cengage.com/highered.
7. Homburg, C. and Wielgos, D. (2022) The value relevance of digital marketing capabilities to firm performance. *Journal of the Academy of Marketing Science*, 50: 1–23. doi:10.1007/s11747-022-00858-7.
8. Iglewicz, B. and Hoaglin, D.C. (1993) Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques. *Technometrics*, 36 (3): 315.
9. Jollife, I.T. and Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374 (2065). doi:10.1098/RSTA.2015.0202.
10. Kapferer, J.-N. and Bastien, V. (2009) *The Luxury Strategy: Break the Rules of Marketing to Build Luxury Brands*.
11. Kotler, P. and Keller, K.L. (2021) Marketing Management 16th Global Edition Chernev, A. (ed.). *Pearson Education Limited*, pp. 1–832. Available at: https://online.fliphtml5.com/xzfda/pkef/ (Accessed: 15 March 2025).
12. Nadine, H., Klaus-Peter, W., Christiane, K., et al. (2024) *The Concept of Luxury: A Global Phenomenon with Local Implications - The European Financial Review*. Available at: https://www.europeanfinancialreview.com/the-concept-of-luxury-a-global-phenomenon-with-local-implications/ (Accessed: 15 March 2025).
13. Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 (C): 53–65. doi:10.1016/0377-0427(87)90125-7.
14. Ruxton, G.D. (2006) The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test. *Behavioral Ecology*, 17 (4): 688–690. doi:10.1093/beheco/ark016.
15. Solomon, M.R.. (2020) *Consumer behavior : buying, having, and being*. Pearson.
16. Wedel, M. and Kannan, P.K. (2016) Marketing analytics for data-rich environments. *Journal of Marketing*, 80 (6): 97–121. doi:10.1509/JM.15.0413.