UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Assessment and Feedback: Student Template

Student ID Number(s): 2653531

Module: Marketing Analytics and Behavioral Science

Module Leader OR Dissertation/Extended Essay Supervisor: Dr Zizhou Peng

Assignment Title: Case Analysis - Individual Assignment

Date and Time of Submission: 17 March 11:56 AM

Actual Word Count: 1999

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

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

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 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.*

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

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* • Tried to maintain coherence • Worked on new ideas •
Section Two: In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points) • Provided new branding ideas • I tried to follow a storytelling approach through data •

Table of Contents

Executive Summary

SmartFresh Retail specializes in offering six categories of products: wine, meat, organic foods, wellness products, treats, and luxury goods. It has a customer base mostly around the age 45 to 65 years old and is slightly targeted towards high-income earners. A detailed customer segmentation analysis revealed that customers primarily spend on wine and meat, with less interest in organic, wellness, treats, and luxury products. Thus, SmartFresh Retail's brand is perceived as specializing in wine and meat rather than a holistic retail chain and has the option to leverage this perception.

Principal Component Analysis (PCA) identified key customer spending behavior along with common campaign characteristics. It was found that customers spend on meat and wine together while considering treats, organic foods, and wellness products together. Furthermore, luxury items are a standalone consideration for customers.

Through K-means cluster analysis, customers have been divided into five categories: Loyal II, Loyal I, Regular II, Regular I, and Irregular customers. The analysis of promotional campaigns showed that offers significantly increase spending. T-tests further confirmed the effectiveness of targeted campaigns in boosting customer expenditure.

Based on these findings, the report recommends repositioning SmartFresh's brand to enhance its image in non-specialized product categories while retaining its strength in wine and meat. Strategic promotions, loyalty programs, and bundle offers can be employed to target different customer segments effectively, particularly focusing on the Loyal and Regular groups.

SmartFresh Retail earned revenue of £1,356,988 from its 2,241 distinct customers in the former years. The customer segment varies demographically in age, marital status, annual income, etc. Besides, it also varies on distinctive purchase behavior; some are influenced by promotions and offers, while others prefer routine purchases with unique psychographic behavior.

Part A: Customer Segment and Distribution Analysis

Age:

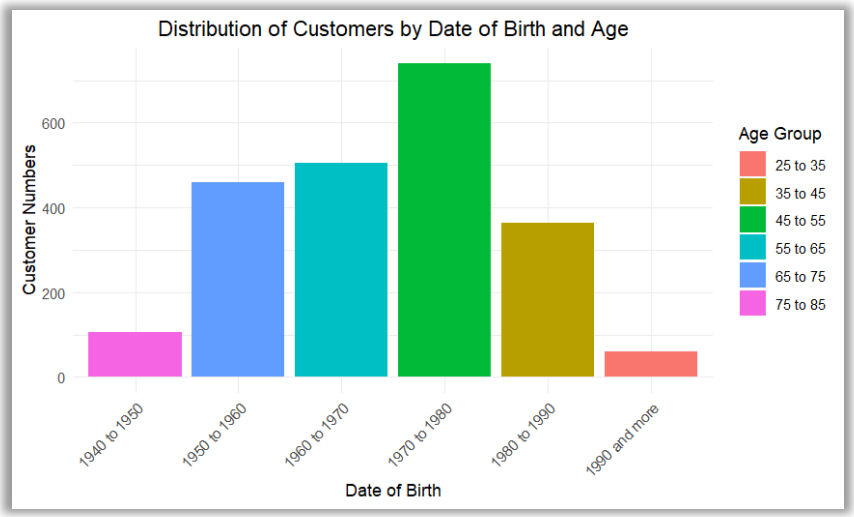

Figure 1 Distribution of Customers by DOB and Age

Till the current year (2025), the present customers can be generally grouped into five major age groups. Most customers are aged 45 to 55 years old, born between 1970 and 1980, also known as Generation X, followed by late baby boomers (1960-70) and early baby boomers (1950-1960)(Brosdahl and Carpenter, 2011). Overall, the customer segment is mostly right-skewed, indicating a customer concentration of age 45 and greater contributes high revenue.

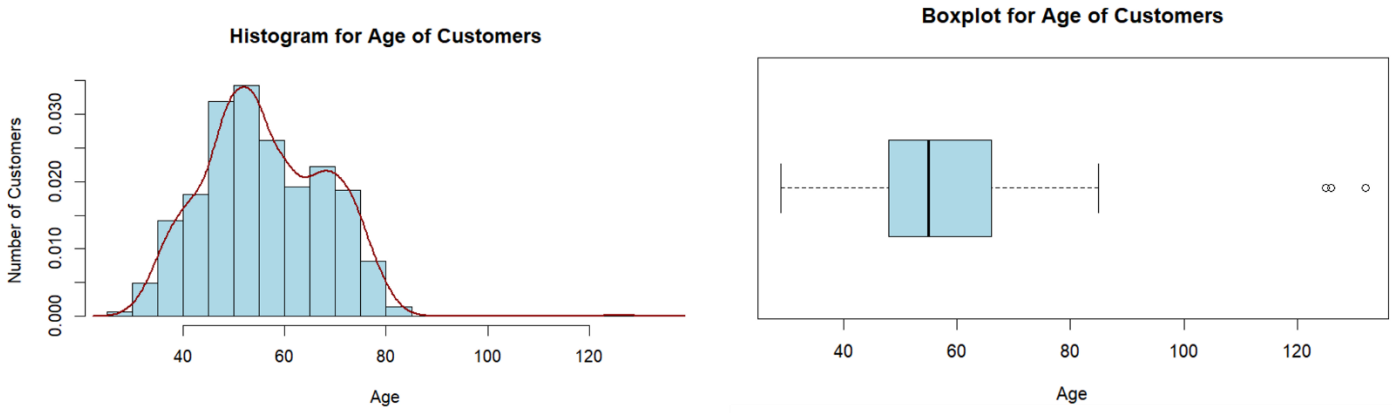

Figure 2 Histogram and Boxplot of Age of Customers

The mean age is 56.18, with a median of 55, indicating a symmetrical distribution with a slightly right-skewed, highlighting more higher-age customers. The standard deviation is nearly 12 years old,

indicating a 12-year spread around the mean. Besides, a Kurtosis of 0.73 indicates few outliers, also visible in the box plot. The higher IQR is due to outliers, which are negligible.

Table 1 Statistical Summary of Age of Customers

Age	
Description	**Values**
Mean	56.18
1st Quartile	48.00
Median	55
3rd Quartile	66
Mode	49
Standard Deviation	11.99
Sample Variance	143.65
Kurtosis	0.73
Skewness	0.35
Range	103
Minimum	29
Maximum	132
Sum	124494
Count	2216

From a customer segmentation point of view, SmartFresh's customer base is from 45 to 65 years old (late baby boomers and Gen X).

Education and Income:

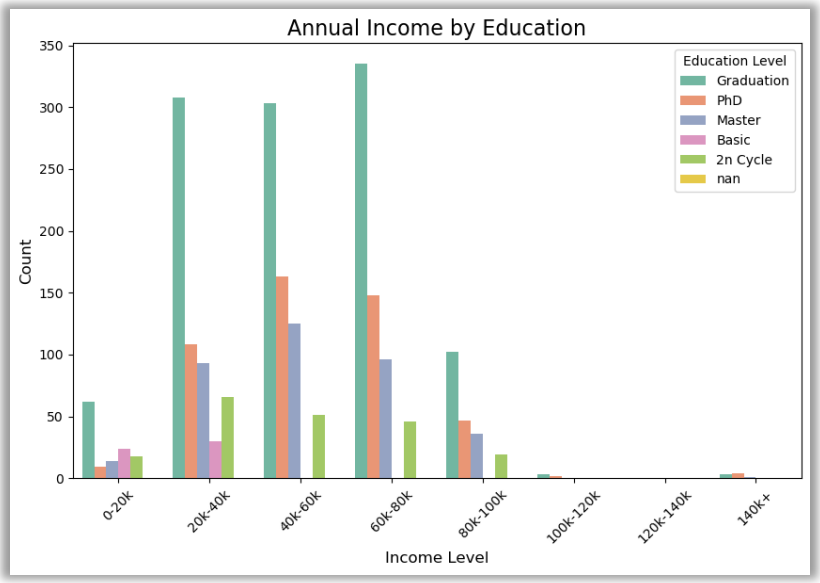

Figure 3 Annual Income by Education

Education has a direct impact on income distribution; as such, customers were segmented in terms of income and educational background (Abdullah, Doucouliagos and Manning, 2015). Graduates have

the highest income in each annual income bracket, followed by PhD and Masters students. In some age brackets, basic education performs over 2n Cycle, and there are few customers with no information on education and income.

Table 2 Statistical Summary of Annual Income

Annual Income	
Description	**Values**
Mean	52247.25
1st Quartile	35303.00
Median	51381.5
3rd Quartile	68522
Mode	7500
Standard Deviation	25173.08
Sample Variance	633683788.58
Kurtosis	159.64
Skewness	6.76
Range	664936
Minimum	1730
Maximum	666666
Sum	115779909
Count	2216

The mean income of £52,247 is slightly higher than the median of £51,382. The dataset is highly right-skewed by 6.76 due to extreme outliers confirmed by the boxplot. Here, 25% of the customers earn below £35,000, and another earns above £ 68,522, and 50-50 customer split at £51,000, indicates a diverse customer range confirmed by a standard deviation of £25,173.

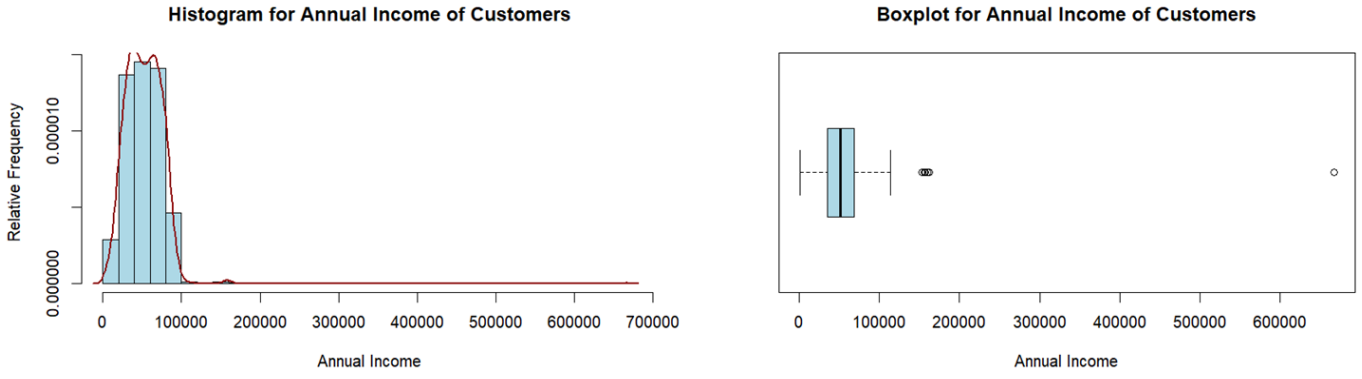

Figure 4 Histogram and Boxplot of Annual Income

Thus, SmartFresh requires a wide range of products and promotions to cater to its customers.

Spending behavior:

SmartFresh Retail has six general categories of products, and each product has a different demand. The customers have spent the most on wine, more than £676,083, followed by meat, luxury, wellness products, treats, and organic foods respectively.

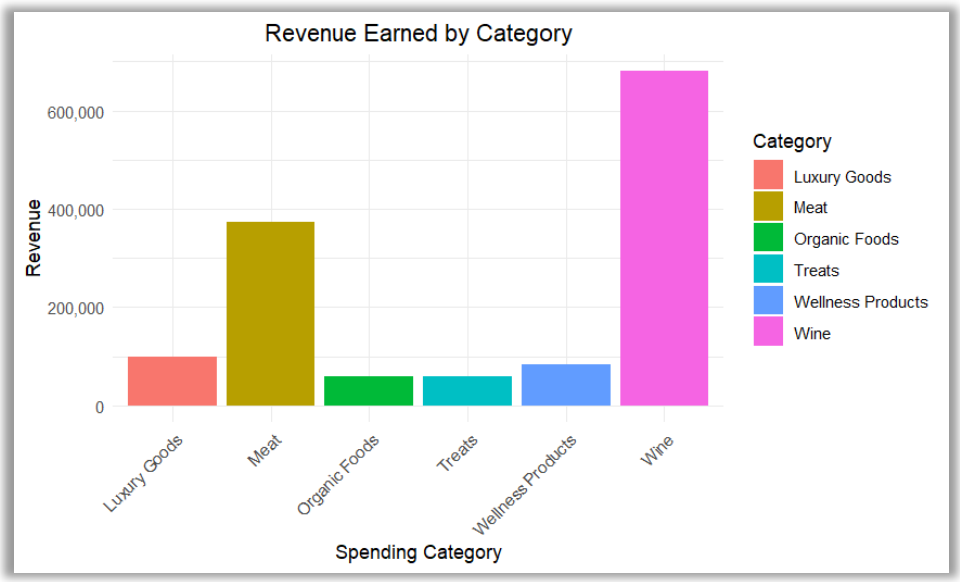

Figure 5 Revenue by Customer Spending Categories

The spending behavior for each of the categories shows a critical difference between the mean and median, emphasizing a diverse array of customers, further proven by the 1st and 3rd quartile difference. Apart from wine and meat, a repeated mode of 0/1 shows many people didn't spend on organic, wellness, treats, and luxury products.

Table 3 Statistical Summary of Spending Behavior

Description	*Wine*	*Organic Food*	*Meat*	*Treats*	*Wellness Products*	*Luxury Goods*
Mean	305.09	26.36	167.00	27.03	37.64	43.97
1st Quartile	24.00	2.00	16.00	1.00	3.00	9.00
Median	174.5	8	68	8	12	24.5
3rd Quartile	505	33	232.2	33	50	56
Mode	2	0	7	0	0	1
Standard Deviation	337.33	39.79	224.28	41.07	54.75	51.82
Sample Variance	113790.13	1583.56	50302.99	1686.91	2997.79	2684.84
Kurtosis	0.58	4.05	5.06	4.11	3.08	3.16
Skewness	1.17	2.10	2.03	2.10	1.92	1.84
Range	1493	199	1725	262	259	321
Minimum	0	0	0	0	0	0
Maximum	1493	199	1725	262	259	321
Sum	676083	58405	370063	59896	83405	97427
Count	2216	2216	2216	2216	2216	2216

Median values of wine and meat suggest that Smartfresh has, in general, average spending customers but some extremely high spenders, proven from box-plot and highly right-skewed data.

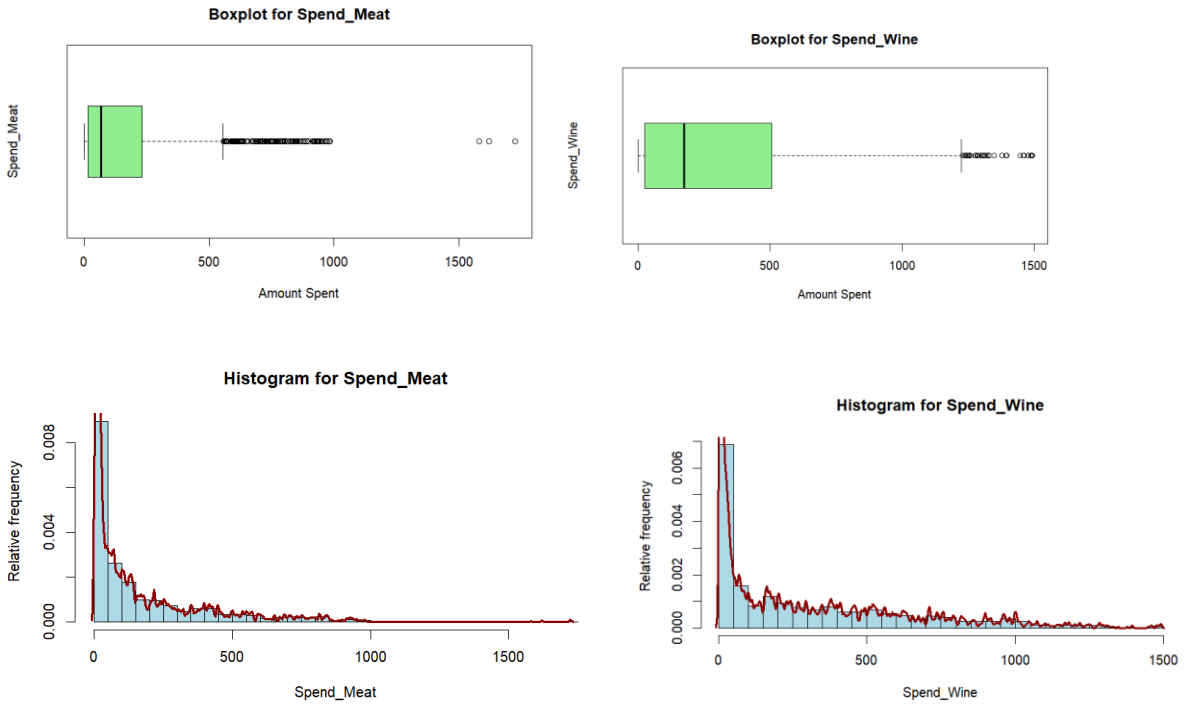

Figure 6 Boxplot and histogram of Meat and Wine Spending

On the contrary, the extreme outliers for organic, wellness, treats, and luxury goods show that Smartfresh has a niche market, with a low average buyers.

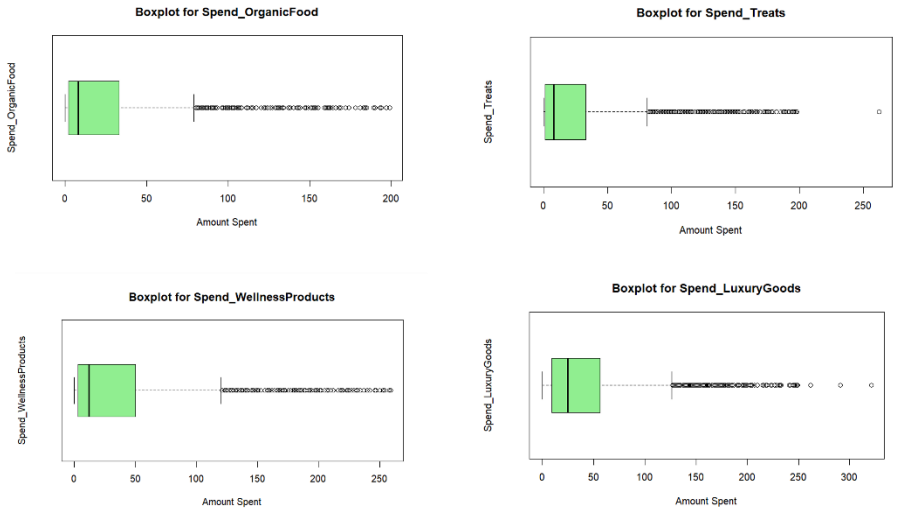

Figure 7 Boxplot of Organic, Treats, Wellness, and Luxury Products spending

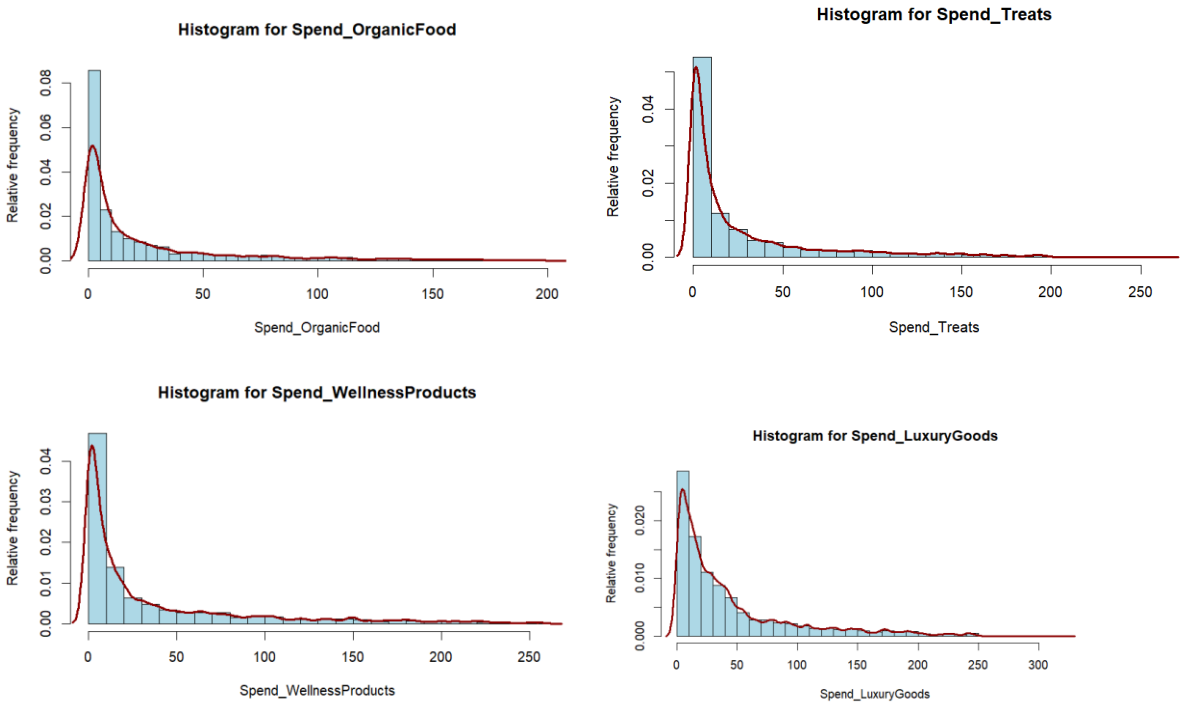

Figure 8 Histograms for Organic, Treats, Wellness, and Luxury Spending Categories

Current Brand Perception:

Based on the spending, SmartFresh is perceived as a specialized wine and meat retail company selling other product categories.

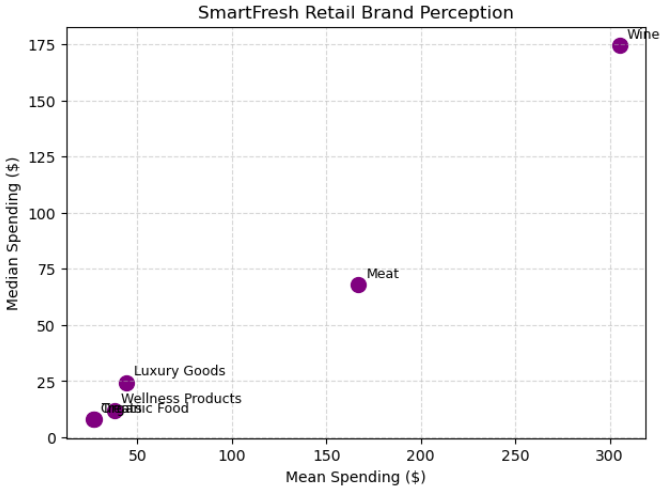

Figure 9 Current Brand Perception of SmartFresh Retail

As seen from the graph, wine is possibly at the top of customers' minds, followed by meat, while all other categories are perceived similarly.

(630 words)

PART B: PCA for grouping spending behavior and similar promotional offers

Principal component analysis (PCA) is used to find principal factors behind existing variables by grouping them, which is often utilized to analyze customer behavior (Liu, 2021). Two separate PCAs were run for spending behavior and promotional offers.

PCA I : Six Spending Categories

An initial correlation matrix was run within the six spending categories to conduct PCA. It was found that wine and meat have a moderate correlation. Besides, organic, wellness, and treats have a moderately higher correlation. However, luxury has almost no correlation with other products.

Figure 10 Correlation Matrix of Spending Categories

Result:

It was found that Principal Component 3 (PC3) explains 80% of the variance of the spending categories, though the scree plot shows PC2.

Description	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.8352	0.8587	0.8296	0.6623	0.63931	0.5991
Proportion of Variance	0.5614	0.1229	0.1147	0.0731	0.06812	0.05982
Cumulative Proportion (%)	56%	68%	80%	87%	94%	100%

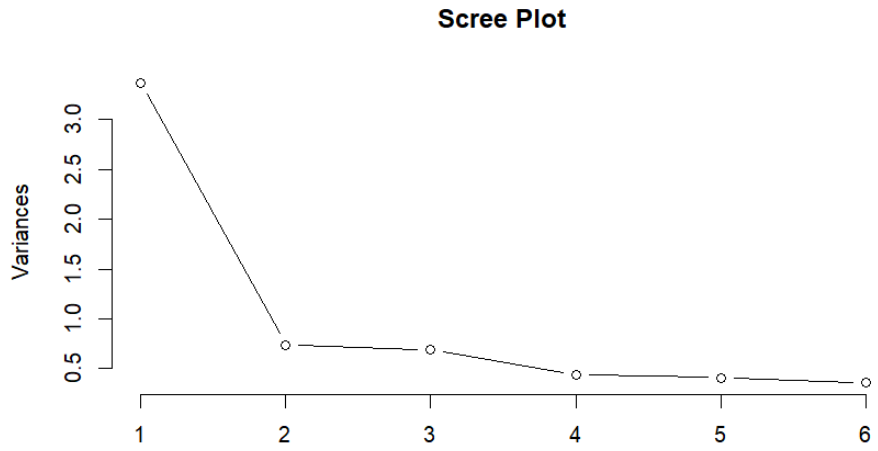

Figure 11 Scree plot for spending categories

By rotating the loadings, it has been found that organic food, wellness products and treats fall under the same factor RC1, whereas wine and meat fall together under RC2. Luxury good stands separately under RC3. Finally, RC3 is able to explain 80% of the spending categories variances.

Loadings	RC1	RC2	RC3
Spend_Wine	0.168	**0.91**	0.225
Spend_OrganicFood	**0.8**	0.198	0.18
Spend_Meat	0.569	**0.667**	
Spend_WellnessProducts	**0.786**	0.221	0.229
Spend_Treats	**0.805**	0.193	0.135
Spend_LuxuryGoods	0.25	0.184	**0.943**

Table 4 Final results after factor rotation

Description	RC1	RC2	RC3
SS loadings	2.319	1.431	1.044
Proportion Var	0.386	0.239	0.174
Cumulative Var (%)	39%	63%	80%

Insights:

From the result, it can be interpreted that customers who buys meat are prone to spend on wine. Similarly, customers of organic foods, wellness products and treats behave similarly and can be target

together in any omnichannel marketing campaign. On the contrary, luxury goods have standalone customers who requires separate strategy for omnichannel campaigns.

PCA II: Grouping 6 Campaigns of similar characteristics

PCA was done on Five campaigns and one recent campaign, and three principal factors were found, which is able to explain approximately 70% of the variances.

Table 5 PCA Result for Campaigns

Description	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.4485	1.0726	0.9369	0.8214	0.7874	0.76082
Proportion of Variance	0.3497	0.1917	0.1463	0.1124	0.1033	0.09648
Cumulative Proportion	0.3497	0.5414	0.6877	0.8002	0.9035	1

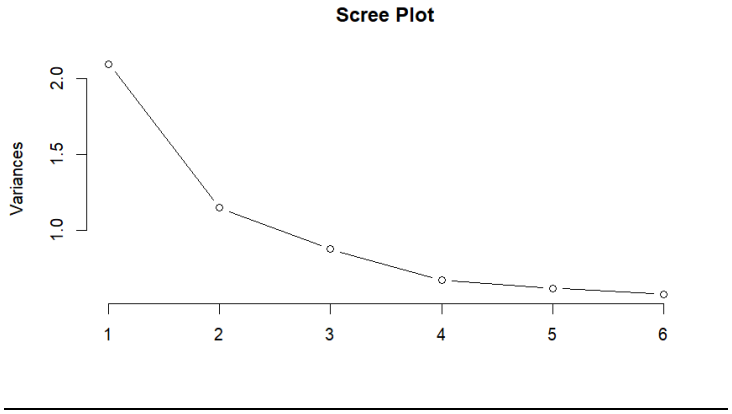

Figure 12 Scree Plot of PCA for Campaigns

After factor rotation, six campaigns were reduced to the three following groups:

Campaigns	RC1	RC3	RC2
Campaign 1			**0.908**
Campaign 2	0.404	**0.626**	-0.268
Campaign 3	**0.765**	0.196	
Campaign 4	**0.794**		
Campaign 5		**0.903**	0.177
Latest Campaign	0.559	0.106	**0.496**

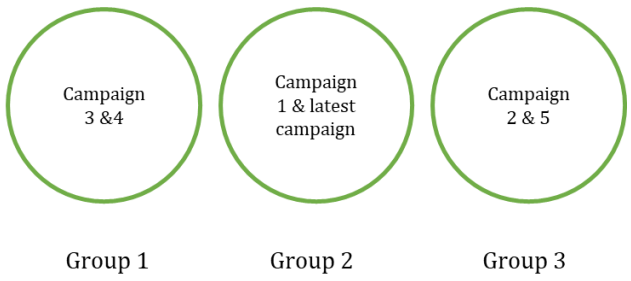

Figure 13 Final results after factor rotation for Campaigns

Each group of campaigns had similar traits and differences from the other group, which would have impacted the spending of the customers. A further hypothesis is discussed in the T-Test section.

Part C: Clustering the similar customers:

For clustering the customers, only relevant features were selected to get practical insights and to reduce complex interpretation. Feature selection is one of the popular techniques for data mining, used during processing and pre-processing part (Liu and Yu, 2005). As such, age, educational level, selected marital status, and annual income were considered. Besides, spending category-wise and purchase channels were selected.

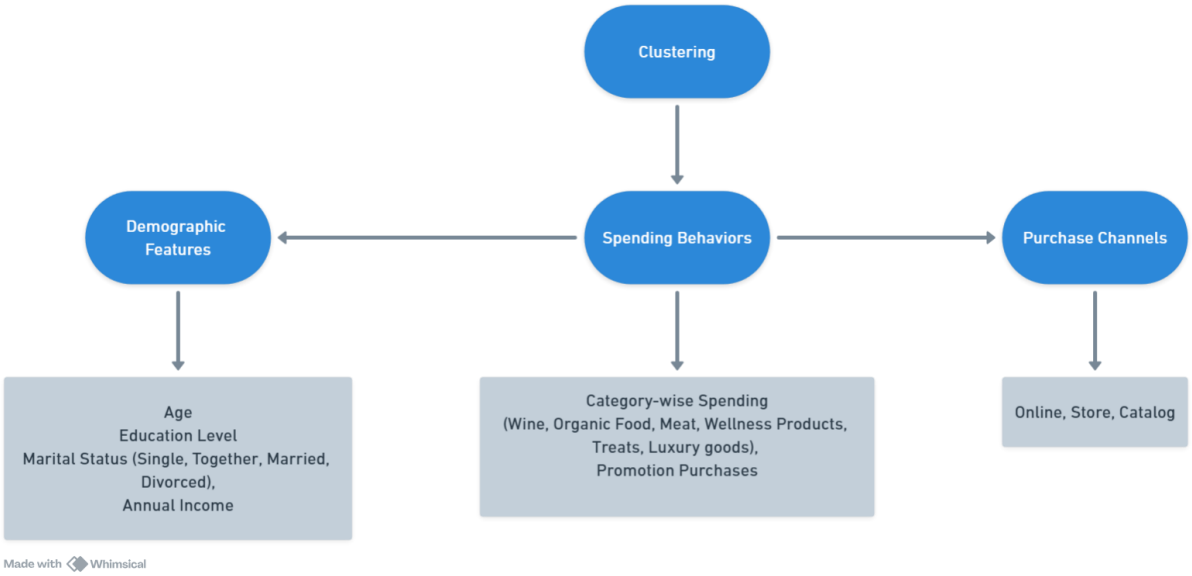

Figure 14 Clustering Feature Selection

Clustering result:

In total 1940 customers were clustered into 5 broader categories. Although the optimal elbow was showing two, but for a diversely segmented customer base, only two clusters is not feasible. Besides, the clustering results were compared with within sum of squares and other matrices.

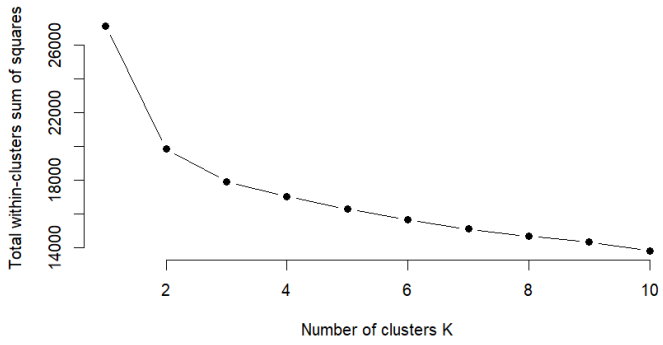

Figure 15 Elbow diagram for Clustering

Table 6 Clustering Result

Cluster	1	2	3	4	5
Count	191	891	223	370	265
Within Cluster Sum of Squares	3868.79	4199.52	1877.82	1892.06	4444.54
between_SS / total_SS	40%				

The within-cluster sum of squares of clusters 3 and 4 are the least, explaining they are closely clustered compared, whereas clusters 2 and 5 have the highest, which means they are loosely clustered. There is a 40% variability between the clusters for K=5, higher than K= 2,3 and 4.

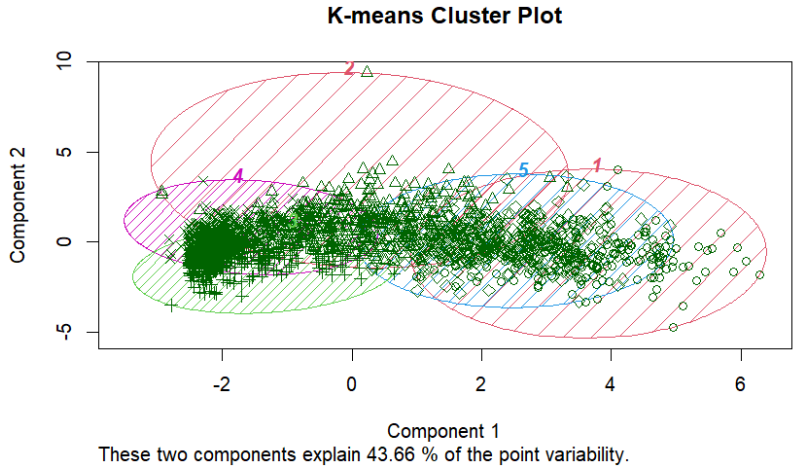

Figure 16K-means Clustering Plot

Cluster Insights:

The five clusters can be divided into five categories: Irregular (Price sensitive), Regular I, Regular II, Loyal I, and Loyal II. The mean age of the customers is quite similar between 55 to 57, and on average, most are graduates with marital status either together or married. An interesting fact is that, the average income of irregular and regular I customers are higher than some higher than rest of the categories. One of the reasons can be they are price sensitive, or they are sporadic customers preferring other retails.

Table 7 Clusters

Customer Demography					
Description	**Irregular (Price Sensitive)**	**Regular I**	**Regular II**	**Loyal I**	**Loyal II**
Count of Cluster	891	191	370	223	265
Age	56	58	56	55	57
Annual Income	51836	58511	49589	52403	54745
Education Level	4	4	3	4	4
Marital Status	2	2	3	2	2

Spending Behavior					
Description	**Irregular (Price Sensitive)**	**Regular I**	**Regular II**	**Loyal I**	**Loyal II**
Total Spending	98	516	814	1381	1449
Total Purchase Counts	6	14	18	20	20
Average Purchase Baskets (£)	17	37	44	70	73

Category-wise Spending Behavior					
Description	**Irregular (Price Sensitive)**	**Regular I**	**Regular II**	**Loyal I**	**Loyal II**
Wine	43.04	309.13	512.44	528.29	709.38
Meat	22.79	107.74	150.31	415.04	506.60
Organic Food	4.97	12.73	27.86	106.87	39.04
Wellness Products	7.35	19.57	34.06	126.09	84.45
Treats	5.17	14.52	27.26	105.78	45.83
Luxury Goods	14.95	52.56	61.58	98.66	63.92

Purchase Channel					
Description	**Irregular (Price Sensitive)**	**Regular I**	**Regular II**	**Loyal I**	**Loyal II**
Online	2	6	6	6	5
Catalog	1	2	3	6	7
Store	3	6	9	8	8

The customers can easily differ in their spending behaviors, with Loyal II toppling the other categories.

Loyal II:

They are the most loyal and highest spending customers; however, they focus mostly on wine and meat and buy a decent number of products from the other categories, preferring catalog and store purchases over online.

Loyal I:

This is a high-spending category with a preference for wine and meat, respectively, but with an overall balanced purchase basket. On average, they purchased 70 times in the last few years and prefer stores but similar to online and catalog.

Regular II:

They spend almost similar to Loyal I customers when it comes to wine, and they have a higher purchase basket compared to Regular I customers.

Regular I:

Though they have a higher income and slightly older customers, they may prefer other retail shops.

Irregular:

This category consists of the largest group who purchase irregularly, having the lowest order values in every category. They are mostly price-sensitive customers and may only purchase during discounts and promotions.

Spend_Meat	Spend_Wine	Spend_WellnessProducts	Spend_Treats	Spend_LuxuryGoods	Spend_OrganicFood
-0.26357451	0.0105175	-0.33315035	-0.311565608	0.1675299	-0.34570667
-0.64362695	-0.7749754	-0.55608583	-0.537322423	-0.5565287	-0.54106403
1.11129597	0.6574949	**1.61035784**	**1.892126709**	**1.0551122**	**2.02365253**
-0.07309533	0.6107232	-0.06874885	-0.003956948	0.3411473	0.03504016
1.52090798	**1.1920956**	0.85068479	0.444464621	0.3862427	0.3165232

Finally, from the cluster centers of each of the spending categories, it's evident that customers often bundle meat and wine together while purchasing, and they buy the other four categories in a separate bundle.

Part D: T-Test for Validating Spending Differences in Campaigns and Offers

T-test is a popular statistical analysis tool often used to analyze marketing campaign effectiveness; for instance, Barajas *et al.* (2012) used it for analyzing targeted display advertisements and Darsareh *et al.* (2019) used to study a social marketing campaign.

Now that the core customer segments have been figured out and patterns have been found in customers' buying behavior, it is essential to analyze the effectiveness of campaigns to understand whether they are making any difference in customer spending.

Smart Fresh used six campaigns to increase customer spending. Using PCA, we found three groups of offers: Group 1 (campaigns 3& 4), Group 2 (latest campaign & campaign 1), and Group 3 (campaigns 2 & 5). We ran three T-tests, selecting one campaign for Group 1, Group 2, and Group 3, and the results of two are mentioned as the other one had similar results.

Hypothesis 1: Campaign 3 (Wellness Products)

Null Hypothesis (H_0):

The mean spending of customers who accepted the wellness product offer is less than or equal to the mean spending of customers who declined the offer.

Alternative Hypothesis (H_1):

The mean spending of customers who accepted the wellness product offer is greater than the mean spending of customers who declined the offer.

Result:

With a 95% confidence level, t = 8.3774, df = 180.86, p-value = 1.466e-14, the mean spending for customers who accepted the offer was 75.89, and the mean spending for customers who declined the offer was 34.63. With a large t-statistic and p-value smaller than 0.05, we can reject the null hypothesis (Kim, 2015).

Hypothesis 2: Latest Campaign (Meat)

Null Hypothesis (H_0):

The mean spending on meat by customers who accepted the offer is less than or equal to that of customers who declined the offer.

Alternative Hypothesis (H_1):

The mean spending on meat by customers who accepted the offer is greater than the mean spending on meat by customers who declined the offer.

Result:

With t = 9.0833, df = 392.97, p-value < 2.2e-16, and a 95% confidence level, the null hypothesis is rejected since the p-value is much lower than 0.05. Besides, the mean spending for customers who accepted offers was 293.76, and for those who didn't, it was 144.64. As a result, we can confidently say that customers who accepted the offer spends significantly over those who didn't accept the offer.

Insights:

From the T-tests, we can prove that the campaigns have impacts on customer spending and thus should be used strategically to increase the customer base in multiple categories for SmartFresh Retail.

Final Recommendations:

Repositioning SmartFresh Retail Brand:

Throughout the analysis, it has been highlighted that SmartFresh Retail is perceived as a retail chain focusing on wine and meat rather than an overall retail focusing on all product categories equally. Compared to the current retail chains, the point of parity (PoP) can be with premium high-retail brands like Waitrose and John Lewis. The point of difference (POD) for SmartFresh retail is clearly on wine and meat.

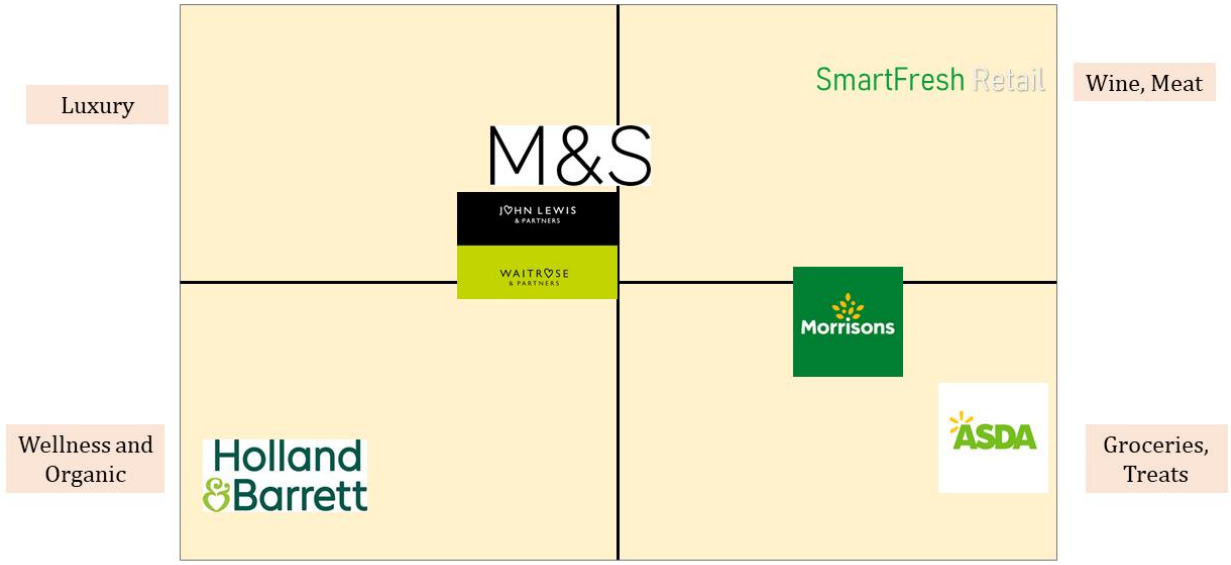

Figure 17 Current Customer Brand Perceptions for Retail Chains

However, if they want to change the consumer perception, they can opt for a rebranding or a brand repositioning strategy, which is often used as a business strategy to change the focus on the consumer segments(Miller, Merrilees and Yakimova, 2014).

A brand-building pyramid model can be utilized to enhance the brand repositioning of SmartFresh Retail (Keller, 2013). For repositioning the brand, they must focus on how they want to be perceived in the customer's mind. If they want to focus on organic, treats, and wellness, they can promote them through bundle offers in the Regular I and Regular II categories. From the clustering and PCA analysis, it has been seen that customers tend to have similar purchase patterns for this category. Next, they can offer special loyalty cards for Loyal I and Loyal II categories in luxury items since they already have a high purchase basket. As they already specialize in wine and meat, they can offer a cheaper option for price-sensitive irregular customers. Thus, they can pull them into the business.

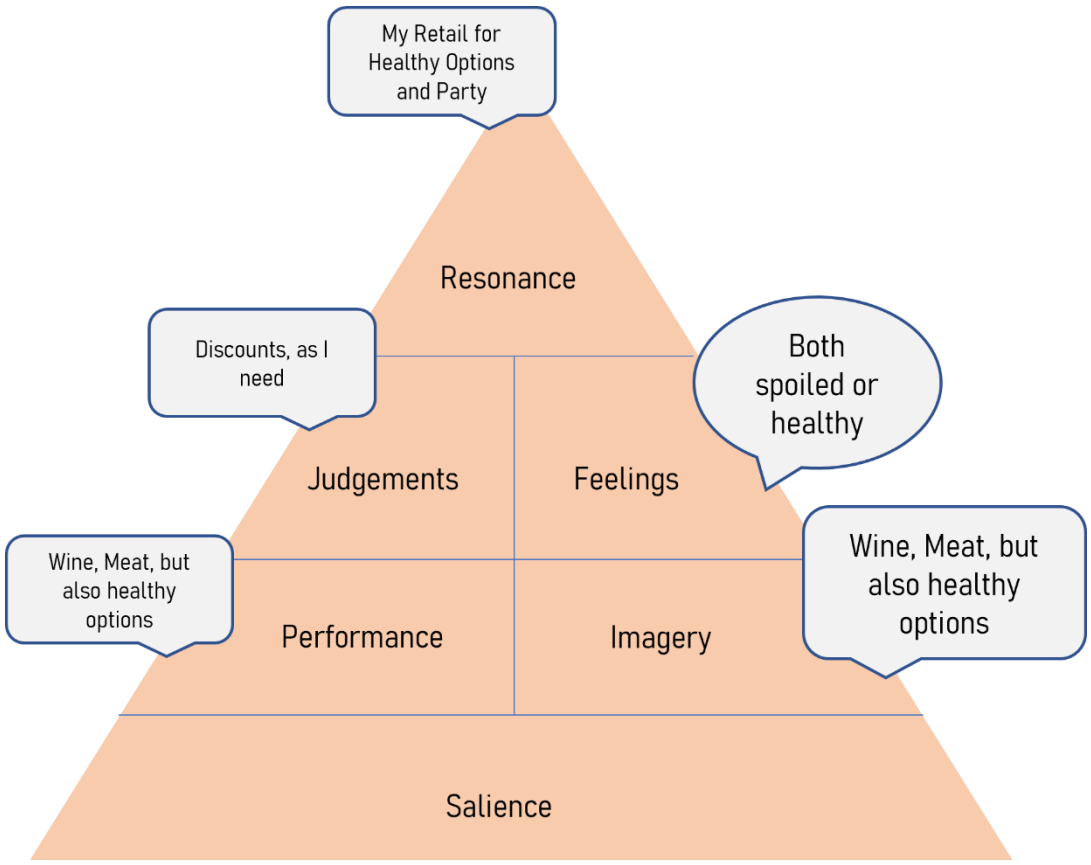

Figure 18 Brand Building Pyramid (Keller, 2013)

From T-tests, it's seen that offers have a high correlation with higher spending, so they can develop strategies to constantly provide customers more offers and create a new perception for overall brand repositioning.

References

Abdullah, A., Doucouliagos, H. and Manning, E. (2015) 'Does Education Reduce Income Inequality? A Meta-Regression Analysis', *Journal of Economic Surveys*, 29(2), pp. 301–316. Available at: https://doi.org/10.1111/joes.12056.

Barajas, J. *et al.* (2012) 'Marketing campaign evaluation in targeted display advertising', in *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*. New York, NY, USA: Association for Computing Machinery (ADKDD '12), pp. 1–7. Available at: https://doi.org/10.1145/2351356.2351361.

Brosdahl, D.J.C. and Carpenter, J.M. (2011) 'Shopping orientations of US males: A generational cohort comparison', *Journal of Retailing and Consumer Services*, 18(6), pp. 548–554. Available at: https://doi.org/10.1016/j.jretconser.2011.07.005.

Darsareh, F. *et al.* (2019) 'B Butterfly Campaign: A social marketing campaign to promote normal childbirth among first-time pregnant women', *Women and Birth*, 32(2), pp. e166–e172. Available at: https://doi.org/10.1016/j.wombi.2018.06.007.

Keller, K.L. (2013) *Strategic Brand Management: Building, Measuring, and Managing Brand Equity*. Pearson.

Kim, T.K. (2015) 'T test as a parametric statistic', *Korean Journal of Anesthesiology*, 68(6), pp. 540–546. Available at: https://doi.org/10.4097/kjae.2015.68.6.540.

Liu, H. (2021) 'Big data precision marketing and consumer behavior analysis based on fuzzy clustering and PCA model', *Journal of Intelligent & Fuzzy Systems*, 40(4), pp. 6529–6539. Available at: https://doi.org/10.3233/JIFS-189491.

Liu, H. and Yu, L. (2005) 'Toward integrating feature selection algorithms for classification and clustering', *IEEE Transactions on Knowledge and Data Engineering*, 17(4), pp. 491–502. Available at: https://doi.org/10.1109/TKDE.2005.66.

Miller, D., Merrilees, B. and Yakimova, R. (2014) 'Corporate Rebranding: An Integrative Review of Major Enablers and Barriers to the Rebranding Process', *International Journal of Management Reviews*, 16(3), pp. 265–289. Available at: https://doi.org/10.1111/ijmr.12020.

Appendix

Appendix

PART A Codes:

Figure 1:

```
age_data <- data.frame(
  Date_of_Birth = c("1940 to 1950", "1950 to 1960", "1960 to 1970", "1970 to 1980", "1980 to 1990", "1990 and more"),
  Customer_no = c(107, 460, 506, 740, 363, 61),
  Percentage = c(4.78, 20.54, 22.59, 33.04, 16.21, 2.72),
  Age = c("75 to 85", "65 to 75", "55 to 65", "45 to 55", "35 to 45", "25 to 35")
)
ggplot(age_data, aes(x = Date_of_Birth, y = Customer_no, fill = Age)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Distribution of Customers by Year and Age",
       x = "Date of Birth",
       y = "Customer Numbers",
       fill = "Age Group") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```

Figure 2:

```
library(ggplot2)
data <- data.frame(
  Category = c("Wine", "Meat", "Luxury Goods", "Wellness Products", "Treats", "Organic Foods"),
  Revenue = c(680816, 373968, 98609, 84057, 60621, 58917)
)
ggplot(data, aes(x = Category, y = Revenue, fill = Category)) +
  geom_bar(stat = "identity") +
```

```
 labs(title = "Revenue Earned by Category",
      x = "Spending Category",
      y = "Revenue") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
scale_y_continuous(labels = scales::comma)
```

EDA:

```
smart.df <-read.csv("E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 1\\EDA File.csv")
str(smart.df)
spending_columns <- smart.df[, c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat", "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")]

summary(spending_columns)


par(mfrow = c(2, 1))

for (col in c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat", "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods"))
{
 # Histogram with density curve
 hist(smart.df[[col]], breaks = 30, prob = TRUE, main = paste("Histogram with Density Curve for", col), xlab = col, col = "lightblue")
 lines(density(smart.df[[col]], na.rm = TRUE), col = "red", lwd = 2)
}

# List of spending columns
spending_columns <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat", "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")
```

```
# Loop through each spending column
for (col in spending_columns)
 {
 # Create a horizontal boxplot
 boxplot(smart.df[[col]],
        main = paste("Boxplot for", col),
        xlab = "Amount Spent",
        ylab = col,
        col = "lightgreen",
        horizontal = TRUE)  # Horizontal boxplot


 # List of spending columns
 spending_columns <- c("Spend_Wine", "Spend_OrganicFood", "Spend_Meat", "Spend_WellnessProducts", "Spend_Treats", "Spend_LuxuryGoods")


 # Loop through each spending column
 for (col in spending_columns) {
  # Create a horizontal boxplot
  boxplot(smart.df[[col]],
         main = paste("Boxplot for", col),
         xlab = "Amount Spent",
         ylab = col,
         col = "lightgreen",
         horizontal = TRUE)  # Horizontal boxplot
 }


summary(spending_columns)


install.packages("moments")
library(moments)
```

```
summary(spending_columns)
smart.df <-read.csv("E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 1\\EDA File.csv")
str(smart.df)
summary(smart.df$Age)

hist(smart.df$Age,
    main = "Histogram for Age of Customers",
    xlab = "Age",
    ylab = "Number of Customers",
    breaks = 15,
    col = "lightblue",
    freq = FALSE)
lines(density(smart.df$Age, na.rm = TRUE),
     col = "darkred",
     lwd = 2)

boxplot(smart.df$Age,
       xlab = "Age",
       main = "Boxplot for Age of Customers",
       horizontal = TRUE,  # Horizontal boxplot
       col = "lightblue")

options(scipen = 999)
summary(smart.df$Annual_Income)

hist(smart.df$Annual_Income,
    main = "Histogram for Annual Income of Customers",
    xlab = "Annual Income",
    ylab = "Relative Frequency",
```

```
    breaks = 25,  # Number of bins
    col = "lightblue",
    freq = FALSE)  # freq=FALSE for relative frequency (density)

# Add a density curve
lines(density(smart.df$Annual_Income, na.rm = TRUE),
     col = "darkred",
     lwd = 2)
boxplot(smart.df$Annual_Income,
       xlab = "Annual Income",
       main = "Boxplot for Annual Income of Customers",
       horizontal = TRUE,  # Horizontal boxplot
       col = "lightblue")
```

Part B:

Spending PCA

```
# Step 1: Load a Demo Dataset
spend <- read.csv("E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 1\\PCA Spending.csv")
head(spend)
summary(spend)
str(spend)

# Step 2: Standardize the Data
spend.sc <- spend
spend.sc[1:6] <- data.frame(scale(spend[1:6]))
summary(spend.sc)

library(corrplot)
corrplot(cor(spend.sc[1:6]), method = 'ellipse', order="hclust")
```

```
# This parameter orders the correlation matrix using hierarchical clustering.

# Hierarchical clustering groups similar variables together, making the plot easier to interpret by placing highly correlated variables next to each other.


# Step 3: Perform PCA

spend.pc <- prcomp(spend.sc[1:6], scale=TRUE)

summary(spend.pc)


# Step 4: Visualize PCA Results

# Scree Plot

screeplot(spend.pc, type = "lines", main = "Scree Plot")

plot(spend.pc, type="l")


# Biplot

biplot(spend.pc)


# Step 5: Cumulative Variance Explained

cumsum(summary(spend.pc)$importance[2,]) # The cumsum function computes the cumulative sum of the elements in the vector.

summary(spend.pc)$importance[3,]


# Step 6: Create a New Dataset with Selected Components (First 3 Components)

pca_scores <- spend.pc$x[, 1:3]

head(pca_scores)


# Step 7: Principal Component Loadings

spend.pc$rotation



### Performing rotation
```

```
# Install the necessary packages
install.packages("psych")  # For Varimax rotation
library(psych)

# Apply Varimax rotation using the 'principal' function from the 'psych' package
pca_rotated <- principal(spend.sc[, 1:6], nfactors = 3, rotate = "varimax")

# View the rotated factor matrix
pca_rotated$loadings
```

Offer PCA:

```
# Step 1: Load a Demo Dataset
offers <- read.csv("E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 1\\PCA Offers.csv")
head(offers)
summary(offers)
str(offers)

# Step 2: Standardize the Data
offers.sc <- offers
offers.sc[1:6] <- data.frame(scale(offers[1:6]))
summary(offers.sc)

library(corrplot)
corrplot(cor(offers.sc[1:6]), method = 'ellipse', order="hclust")
# This parameter orders the correlation matrix using hierarchical clustering.
# Hierarchical clustering groups similar variables together, making the plot easier to interpret by placing highly correlated variables next to each other.

# Step 3: Perform PCA
offers.pc <- prcomp(offers.sc[1:6], scale=TRUE)
summary(offers.pc)
```

```
# Step 4: Visualize PCA Results

# Scree Plot

screeplot(offers.pc, type = "lines", main = "Scree Plot")

plot(offers.pc, type="l")


# Biplot

biplot(offers.pc)


# Step 5: Cumulative Variance Explained

cumsum(summary(offers.pc)$importance[2,]) # The cumsum function computes the cumulative sum of the elements in the vector.

summary(offers.pc)$importance[3,]


# Step 6: Create a New Dataset with Selected Components (First 3 Components)

pca_scores <- offers.pc$x[, 1:3]

head(pca_scores)

# Step 7: Principal Component Loadings

offers.pc$rotation



# Apply Varimax rotation using the 'principal' function from the 'psych' package

pca_rotated <- principal(offers.sc[, 1:6], nfactors = 3, rotate = "varimax")


# View the rotated factor matrix

pca_rotated$loadings
```

Part C: Clustering

```
## Step 1: Explore the data ----
```

```
smart <- read.csv("E:/UoB/Semester 2/MABS/Assessments/Assignment 1/Clustering.csv") ##Note: set YOUR path

str(smart)

summary(smart)


## Step 2: Training a model on the data ----

interests <- smart[1:14]

interests_z <- as.data.frame(lapply(interests, scale))

summary(interests_z)


k.max <- 10

#data <- scaled_data

wss <- sapply(1:k.max,

        function(k){kmeans(interests_z, k, nstart=50, iter.max = 15 )$tot.withinss})

wss

plot(1:k.max, wss,

   type="b", pch = 19, frame = FALSE,

   xlab="Number of clusters K",

   ylab="Total within-clusters sum of squares")



set.seed(2345)

smart_clusters <- kmeans(interests_z, 5)


## Step 3: Check the results ----

smart_clusters


# look at the size of the clusters

smart_clusters$size
```

```
# look at the cluster centers
smart_clusters$centers

# look at other parameters
smart_clusters$cluster

# Check the results ----
# look at the size of the clusters
smart_clusters
smart_clusters$tot.withinss
smart_clusters$betweenss
smart_clusters$size
smart_clusters$centers

## Step 4: Visualize clusters ----
library (cluster)
clusplot(interests_z, smart_clusters$cluster, color = TRUE, shade = TRUE, labels = 4, lines = 0, main =
"K-means Cluster Plot")

# Add cluster labels to the original data
smart$Cluster <- smart_clusters$cluster

# Save the dataset with cluster labels
write.csv(smart, "clustered_customers.csv", row.names = FALSE)
```

Part D: T-test

```
# Install and load necessary libraries
install.packages("readxl")
library(readxl)
install.packages("car")
library(car)
```

```
# Read the Excel file
group_1 <- read_excel("E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 1\\T Test\\Group 1.xlsx")

# Calculate descriptive statistics for Wellness_Accept and Wellness_Decline
wellness_stats <- data.frame(
  Metric = c("Mean", "Standard Deviation", "Sample Size", "Min", "Max"),
  Wellness_Accept = c(mean(group_1$Wellness_Accept, na.rm = TRUE),
                      sd(group_1$Wellness_Accept, na.rm = TRUE),
                      length(na.omit(group_1$Wellness_Accept)),
                      min(group_1$Wellness_Accept, na.rm = TRUE),
                      max(group_1$Wellness_Accept, na.rm = TRUE)),

  Wellness_Decline = c(mean(group_1$Wellness_Decline, na.rm = TRUE),
                       sd(group_1$Wellness_Decline, na.rm = TRUE),
                       length(na.omit(group_1$Wellness_Decline)),
                       min(group_1$Wellness_Decline, na.rm = TRUE),
                       max(group_1$Wellness_Decline, na.rm = TRUE))
)

# Display descriptive statistics
print(wellness_stats)

# Combine the data into a single data frame for the Levene's Test and T-test
data_levene <- data.frame(
  Score = c(group_1$Wellness_Accept, group_1$Wellness_Decline),
  Group = rep(c("Wellness_Accept", "Wellness_Decline"),
              times = c(length(group_1$Wellness_Accept),
                        length(group_1$Wellness_Decline)))
```

```
)

# Levene's Test for equality of variances (Check homogeneity of variance)
levene_test <- leveneTest(Score ~ Group, data = data_levene)
print(levene_test)

# Independent Samples T-Test
# The data argument should be the combined data frame with a grouping variable (Group)
t_test <- t.test(Score ~ Group, data = data_levene, var.equal = FALSE)

# Print the T-test results
print("Independent Samples T-Test :")
print(t_test)

# Load your data
group_2 <- read_excel("E:\\UoB\\Semester 2\\MABS\\Assessments\\Assignment 1\\T Test\\Group 2.xlsx")
head(group_2)

# Calculate descriptive statistics for Spend_Meat_Accept and Spend_Meat_Decline
meat_stats <- data.frame(
  Metric = c("Mean", "Standard Deviation", "Sample Size", "Min", "Max"),
  Spend_Meat_Accept = c(mean(group_2$Spend_Meat_Accept, na.rm = TRUE),
                        sd(group_2$Spend_Meat_Accept, na.rm = TRUE),
                        length(na.omit(group_2$Spend_Meat_Accept)),
                        min(group_2$Spend_Meat_Accept, na.rm = TRUE),
                        max(group_2$Spend_Meat_Accept, na.rm = TRUE)),

  Spend_Meat_Decline = c(mean(group_2$Spend_Meat_Decline, na.rm = TRUE),
                         sd(group_2$Spend_Meat_Decline, na.rm = TRUE),
```

```
                length(na.omit(group_2$Spend_Meat_Decline)),
                min(group_2$Spend_Meat_Decline, na.rm = TRUE),
                max(group_2$Spend_Meat_Decline, na.rm = TRUE))
)

# Display the results
print(meat_stats)

# Combine data into a single frame with a grouping column for Levene's Test
data_levene <- data.frame(
  Score = c(group_2$Spend_Meat_Accept, group_2$Spend_Meat_Decline),
  Group = rep(c("Spend_Meat_Accept", "Spend_Meat_Decline"),
          times = c(length(group_2$Spend_Meat_Accept),
                  length(group_2$Spend_Meat_Decline)))
)

# Levene's Test for variance equality (to check if the variances are equal between groups)
levene_test <- leveneTest(Score ~ Group, data = data_levene)
print("Levene's Test:")
print(levene_test)

# Perform Independent Samples T-Test
t_test <- t.test(group_2$Spend_Meat_Accept, group_2$Spend_Meat_Decline, var.equal = FALSE)
print("Independent Samples T-Test:")
print(t_test)
```