Customer Segmentation and Marketing Analytics: A Data-Driven Approach for SmartFresh Retail

# **Executive Summary**

**SmartFresh Retail is an omnichannel lifestyle retailer focusing on premium wines, organic foods, and luxury goods.** This report examines the company’s efforts to refine its customer segmentation, aiming to strengthen personalized marketing, retention, and revenue**.** By analyzing data from 2,240 customers using exploratory data analysis, t-tests, principal component analysis, and K-means clustering, it became evident that preserving naturally skewed distributions is critical in a luxury context. High-value outliers, often the core drivers of profit, remain distinctly visible when not compressed by transformations.

The t-tests reveal that customers who accept promotional offers spend substantially more on premium categories such as wine and luxury goods, highlighting the effectiveness of aspirational campaigns. Principal component analysis extracts core spending dimensions, enabling K-means to segment customers into three distinct clusters: (1) elite or affluent consumers, (2) moderate-income shoppers, and (3) budget-conscious buyers. Each group exhibits unique purchasing patterns and promotional responsiveness, underscoring the need for targeted strategies. Exclusive loyalty benefits may best serve affluent patrons, while carefully timed upselling can encourage moderate-income customers to try premium products. Overall, these findings empower SmartFresh Retail to tailor its marketing approach and remain competitive in the high-end retail arena.

Table of Contents

[Executive Summary 3](#_Toc193090731)

[Introduction 5](#_Toc193090732)

[Analysis and Interpretation 6](#_Toc193090733)

[Data Pre-Processing 6](#_Toc193090734)

[Exploratory Data Analysis (EDA) 6](#_Toc193090735)

[T-Test Analysis 8](#_Toc193090736)

[Factor Analysis (PCA) 9](#_Toc193090737)

[K-Means Clustering 11](#_Toc193090738)

[Conclusion 13](#_Toc193090739)

[References 15](#_Toc193090740)

[Appendix 16](#_Toc193090741)

**List of Figures**

[Figure 1: Boxplot of key Variables (Source: Author) 7](#_Toc192983793)

[Figure 2: Histogram of Key Variables (Source: Author) 8](#_Toc192983794)

[Figure 3: Scree Plot: PCA (Source: Author) 10](#_Toc192983795)

[Figure 4: Elbow Graph for Optimal K (Source: Author) 11](#_Toc192983796)

[Figure 5: K-Means Clustering Results (Source: Author) 12](#_Toc192983797)

**List of Tables**

[Table 1: Descriptive Statistics for key variables (Source: Author) 6](#_Toc192983819)

[Table 2: T-Test Results (Source: Author) 9](#_Toc192983820)

[Table 3: K- Means Clusters Result (Source: Author) 12](#_Toc192983821)

# **Introduction**

Retail businesses today face a rapidly evolving landscape, where shifting consumer expectations and intense competition make personalized marketing more important than ever (Arora, et al., 2021). This challenge is especially evident in **luxury retail**, where a small group of high-value customers can account for a disproportionately large share of total revenue (Boston Consulting Group, 2024). Unlike mass-market contexts, where spending is often more evenly distributed luxury consumers display varied motivations, such as exclusivity, status, or occasional indulgence (Kapferer & Bastien, 2012). Understanding these nuances is essential for maximizing marketing effectiveness and sustaining profitability in a premium environment.

SmartFresh Retail, a multi-channel retailer offering premium wines, organic foods, and luxury goods, operates within this dynamic setting. Its customers range from casual shoppers to affluent connoisseurs, each with distinct behaviours and motivations. To improve marketing efficiency, SmartFresh plans to refine its customer segmentation by analysing demographic attributes, purchase histories, and promotional engagement data. Research suggests that well-targeted personalization can raise revenues by 10–30% (Arora, et al., 2021), but such strategies require not just detailed data but also a clear grasp of consumer behaviour, especially in high-end markets where aspiration and exclusivity strongly influence spending (Fionda-Douglas & Moore, 2009).

Luxury data often reveals a **skewed distribution**, with a small segment of wealthy customers dominating total sales (Boston Consulting Group, 2024). While standard practice might suggest data transformation of such skewed data to achieve more normal distributions, doing so risks obscuring the premium spenders who are crucial for profitability and targeted engagement (Kapferer & Bastien, 2012). Preserving raw distributions ensures these top-tier customers remain clearly identifiable, aligning more closely with the strategic needs of a luxury-oriented retailer.

This report addresses several core questions: Who are the main luxury spenders? Which demographic and behavioural factors drive promotional acceptance? How can SmartFresh pinpoint segments offering the greatest potential for upselling and cross-selling? To explore these issues, the report examines data pre-processing choices, especially handling missing values and outliers, applies t-tests, principal component analysis (PCA), and K-means clustering to derive insights, and critically evaluates the decision to retain skewed data. Ultimately, the analysis translates these findings into actionable marketing strategies, enabling SmartFresh to deepen customer relationships, enhance campaign targeting, and maintain a competitive advantage in the evolving luxury market.

# Analysis and Interpretation

## Data Pre-Processing

Data pre-processing served to address missing entries and extreme outliers without losing significant insights about top spenders. A small fraction (<5%) of rows contained missing data (for example, Annual\_Income or Kidhome). Where the gaps were minimal and random, entire rows were omitted to avoid bias; for key variables (like Kidhome), median imputation was used to maintain sample size (Hair, et al., 2018). Extremely high values—such as incomes exceeding £600,000 or wine expenditures above £1,000—were kept intact. In a luxury retail context, these outliers can represent the most influential customers rather than statistical anomalies. Finally, while many monetary variables showed a strong right-skew, no data transformation was performed. Compressing outliers might mask the genuine gap between moderate and elite spenders, which is central to identifying high-value segments in a premium environment.

## Exploratory Data Analysis (EDA)

EDA is critical for revealing distributional characteristics, detecting skew, and identifying the presence of a high-value minority—an expected pattern in the luxury retail segment (Arora, et al., 2021).

Table 1 presents the descriptive statistics for Annual\_Income and the six spending categories: Spend\_Wine, Spend\_OrganicFood, Spend\_Meat, Spend\_WellnessProducts, Spend\_Treats, and Spend\_LuxuryGoods. The table includes the mean, median, standard deviation (SD), and the minimum and maximum values.

Table 1: Descriptive Statistics for key variables (Source: Author)

A table with numbers and a number of data

AI-generated content may be incorrect.

The mean frequently exceeds the median, signifying right-skew. For instance, while the median annual income stands at £50,000, a maximum of £666,666 indicates a small group of affluent individuals. Similarly, the gap between typical spending (e.g., a median of £174 for wine) and an upper limit of £1,493 underlines the presence of outliers who purchase premium goods far above the norm. Each category has a wide range, with many customers spending little or zero and a few spending significantly more. This pattern is characteristic of luxury retail, where a small elite drives a large portion of sales (Kapferer & Bastien, 2012).

**Boxplots of Annual Income and Spending Variables**

Figure 1: Boxplots of Income & Spending depict the median, interquartile range, and any outliers. Each category has a majority of values clustered toward the lower end, with a few outliers stretching significantly higher.

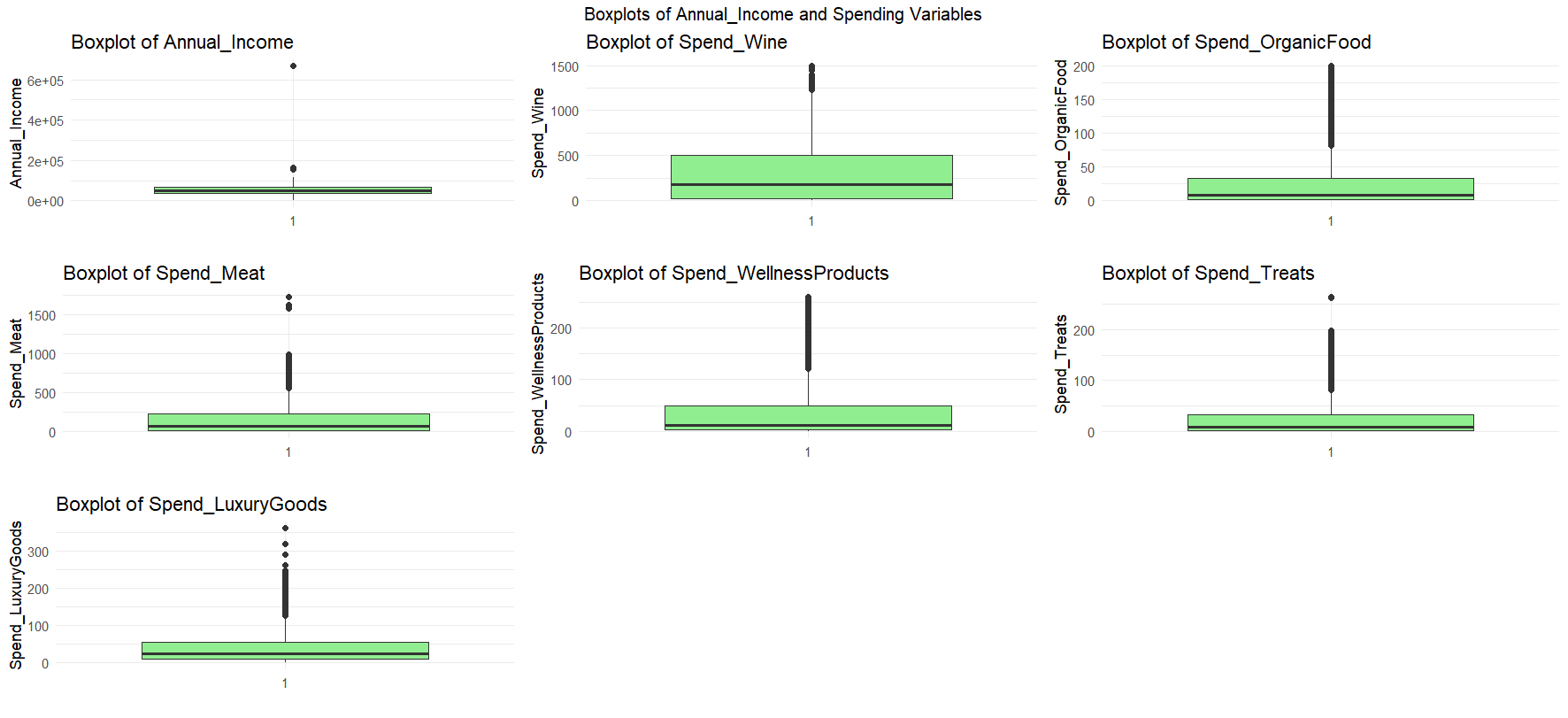


Figure 1: Boxplot of key Variables (Source: Author)

These boxplots reveal that most customers spend at relatively modest levels, while a few high spenders stand out well above the median. In a luxury retail setting, these outliers aren’t anomalies but genuine elite customers who play a crucial role in driving overall revenue (Kapferer & Bastien, 2012). The plots show a clear pattern—most values are concentrated at the lower end, while a few high-spending outliers stretch the distribution. This strong right-skew is common in luxury markets, where a small group of affluent customers drives a significant share of the revenue.

**Histograms of Annual Income and Spending Variables**

Figure 2: Histograms of Income & Spending reveal a long tail for each variable, indicating that most observations cluster at the lower end, with a small number of high spenders or earners.

A graph of a person

AI-generated content may be incorrect.

Figure 2: Histogram of Key Variables (Source: Author)

Most customers display relatively modest income or spending, consistent with the long tail pattern, whereas a small subset of affluent or high-spend individuals extends the right tail and contributes disproportionately to revenue. Recognizing this distribution is crucial for targeting top-value customers in luxury retail (Boston Consulting Group, 2024).

## 

## T-Test Analysis

A t-test compares the means of two groups to see if observed differences are statistically significant (Fionda-Douglas & Moore, 2009). Here, the objective was to determine whether customers who **accepted promotional offers** spend differently from those who did not.

**Table 2. T-Test Results for Spending by Offer Acceptance** compares mean spending on wine, organic food, meat, wellness products, treats, and luxury goods for two groups: (1) customers who accepted an offer, and (2) those who did not.

Table 2: T-Test Results (Source: Author)

A table with numbers and a few black text

AI-generated content may be incorrect.

**Interpretation**:

* **Wine & Luxury Goods**: Offer acceptors allocate significantly more (p < 0.05), implying these premium categories resonate strongly with promotion-responsive customers.
* **Essentials** (OrganicFood, Meat, Wellness, Treats): No significant difference emerges, suggesting discounts or promotions on staple items may not specifically drive higher spend among those who accept offers.

**Marketing Implications**:  
These findings reinforce the idea that **aspirational campaigns**, featuring indulgent or luxury products are more likely to appeal to higher-value patrons. The results provide a strong indication that high-value spenders respond to premium-oriented promotions.

## Factor Analysis (PCA)

Principal Component Analysis (PCA) condenses correlated variables into fewer dimensions, aiding segmentation by clarifying underlying behavioural or demographic factors (Hair, et al., 2018). Here, PCA was applied to nine variables—Annual\_Income, Kidhome, Teenhome, and six spending categories—to identify core axes of variation without discarding potentially important demographic elements.

Figure 2: Scree Plot shows that the **first principal component (PC1)** explains approximately **47.1%** of the total variance, the **second (PC2)** adds around **13.3%,** and the **third (PC3)** contributes a further **8.7%.** Beyond this point, each subsequent component explains progressively less variance, indicating diminishing returns in explanatory power (Bruin, 2011). Retaining the first three PCs thus captures nearly 70% of the variance, striking a practical balance between interpretability and completeness.

A graph with numbers and a line

AI-generated content may be incorrect.

Figure 3: Scree Plot: PCA (Source: Author)

**Interpretation of Scree Plot**:  
PC1 often reflects an overall affluence and premium spending dimension, PC2 highlights differences in family composition and possibly selective indulgence, while PC3 offers more nuanced distinctions for e.g., smaller contrasts between categories or moderate spenders.

From a marketing perspective, focusing on these three components enables a more streamlined segmentation strategy. PC1 distinguishes genuinely high-spend, high-income customers from those with lower incomes or more children; PC2 can separate out those who selectively purchase luxury items versus essential goods; and PC3 may capture subtle behavioural patterns. This clarity informs promotional campaigns ensuring that aspirational items target the high-affluence segment, while moderate or family-oriented customers receive tailored offers. Crucially, no data transformation was performed, preserving the real gap between elite spenders and the rest, which aligns with the reality that a small minority of high-value patrons drives a large portion of revenue in premium markets (Sehgal, 2024).

## K-Means Clustering

Clustering techniques, such as **K-means**, segment customers based on similarity, allowing businesses to tailor marketing strategies. The **Elbow Method** helps determine the optimal number of clusters by minimizing intra-cluster variance while maintaining managerial feasibility (Savic, et al., 2019).

**Elbow Method and Optimal K**

The figure 4: Elbow Method plots the total within-cluster sum of squares (WSS) against various k values. A distinct bend at **k=3** suggests three clusters strike a balance between meaningful separation and managerial feasibility.

A graph of a number of clusters

AI-generated content may be incorrect.

Figure 4: Elbow Graph for Optimal K (Source: Author)

While k=2 might yield slightly higher silhouette scores, it merges moderate and high spenders into a broad segment. K=3 ensures a nuanced distinction that better reflects real consumer differences.

**K-Means Clusters Table and Visualization**

**Table 3. depicts K-Means Clusters (k=3), Z-Score Means** presents the final cluster centres for variables including Annual\_Income, Kidhome, Teenhome, and each spending category.

Table 3: K- Means Clusters Result (Source: Author)

A table with numbers and a number of objects

AI-generated content may be incorrect.

**K-Means Clustering Results**

The **Figure 5**: K-Means Clustering Results depicts how customers group in the PCA space.

A graph showing different colored circles

AI-generated content may be incorrect.

Figure 5: K-Means Clustering Results (Source: Author)

**Cluster Characteristics:**

1. **Cluster 1: Elite/Affluent**

* Positive z-scores for income, wine, luxury, etc.
* Represents the high-value segment vital for premium promotions.

1. **Cluster 2: Moderate-Income/Occasional Indulgers**

* Moderate overall spend, occasional interest in premium categories.
* Amenable to “stepping up” with curated deals or seasonal offers.

1. **Cluster 3: Budget-Oriented**

* Negative z-scores on premium spend, lower annual income.
* Likely responsive to loyalty points or value-based promotions.

**Interpretation of Clusters**:  
Had the dataset been log-transformed, the distinction between Cluster 1 and Cluster 2 might be diluted. In a luxury context, preserving raw data ensures the “elite” segment remains distinct, facilitating tailored strategies for each group.

**Marketing Implications**:

* **Cluster 1**: Prioritize exclusive events, personal shopper services, and premium (VIP) loyalty schemes.
* **Cluster 2**: Offer occasional indulgences, themed bundles, or mid-level premium experiences.
* **Cluster 3**: Maintain engagement via essentials-based offers, loyalty incentives, and smaller “treat” promotions without diluting the brand’s upscale identity.

# Conclusion

This study integrated data pre-processing, exploratory data analysis, t-tests, principal component analysis, and K-means clustering to gain a nuanced understanding of SmartFresh Retail’s customer base within a luxury setting. A deliberate choice was made to retain outliers so that the genuine gap between moderate and elite spenders remained visible. This decision reflects the reality of the luxury retail sector, where a relatively small cohort of high-value customers often generates a disproportionate share of overall revenue (Balchandani, et al., 2025).

The exploratory phase confirmed a strong right-skew, with many customers spending modestly and a small number investing significantly in premium goods. These distributional patterns highlight how typical statistical norms (e.g., normality assumptions) may not apply in luxury retail (Roux, et al., 2017). By recognizing the “long tail,” SmartFresh can focus on top-tier spenders who are most receptive to exclusive promotions. T-tests reinforced this perspective, showing that customers who accepted promotional offers were notably more likely to spend on wine and luxury items, suggesting that indulgent or aspirational campaigns resonate strongly with high-spend patrons.

Principal component analysis distilled multiple spending variables into a few key dimensions, distinguishing overall spending intensity from preferences for aspirational versus practical items. This step ensured that correlated categories (e.g., wine and luxury) did not overshadow each other in the subsequent clustering. K-means clustering then revealed three distinct segments: elite, moderate-income, and budget-oriented, each offering different opportunities for targeted marketing. The elite segment, with its high disposable income and affinity for premium goods, stands to benefit most from VIP experiences, private previews, and high-end loyalty schemes. In contrast, moderate-income customers might respond to carefully curated promotions that encourage them to sample more exclusive products without alienating them through excessive premium pricing. Finally, budget-oriented families are likely to remain engaged through loyalty points, essentials-based bundles, and occasional “treat yourself” offers that align with their cost-conscious behavior.

Reflecting on these outcomes, it is evident that a data-driven segmentation approach, tailored to the unique demands of the luxury sector, enables more effective and personalized marketing strategies. By continually refining its segmentation model through periodic data updates, campaign tracking, and deeper behavioral analysis, SmartFresh can not only optimize retention but also capture additional revenue from an evolving consumer base. The emphasis on preserving outliers and focusing on premium behaviors underscores a practical alignment with real-world spending patterns, thereby equipping SmartFresh to maintain a competitive edge in the high-end retail market.

# References

Arora, N. et al., 2021. *The value of getting personalization right—or wrong—is multiplying.* [Online]   
Available at: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying

Balchandani, A. et al., 2025. *The State of Luxury: How to navigate a slowdown.* [Online]   
Available at: https://www.mckinsey.com/industries/retail/our-insights/state-of-luxury

Boston Consulting Group, 2024. *True-Luxury Global Consumer Insight.* [Online]   
Available at: https://web-assets.bcg.com/ab/c9/75e5c157466091957025a8c858ab/true-luxury-global-consumer-insight-2024-summary-english.pdf

Bruin, J., 2011. *UCLA: Statistical Consulting Group.* [Online]   
Available at: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/

Fionda-Douglas, A. & Moore, C. M., 2009. The anatomy of the luxury fashion brand. *The Journal of Brand Management,* p. 347–363.

Hair, J. F., W. C. B., Babin, B. J. & Anderson, R. E., 2018. *Multivariate Data Analysis.* s.l.:CENGAGE.

Kapferer, J.-N. & Bastien, V., 2012. *The Luxury Strategy: Break the Rules of Marketing to Build Luxury Brands.* London: Kogan Page.

Roux, E., Tafani, E. & Vigneron, F., 2017. Values Associated with Luxury Brand Consumption and the Role of Gender. *Journal of Business Research,* pp. 102-113.

Savic, A., Bjelobaba, G., Janicijevic, S. & Stefanovic, H., 2019. An Application of PCA Based K-Means Clustering for Customer Segmentation in One Luxury Goods Company. *University for Business and Technology in Kosovo .*

Sehgal, P., 2024. *Harnessing the Power of Strategic Segmentation in the Luxury Market.* [Online]   
Available at: https://www.pallavisehgal.com/strategy-marketing/harnessing-the-power-of-strategic-segmentation-in-the-luxury-market  
[Accessed 2025].

# Appendix

## Complete R-Code for SmartFresh Retail Analysis

* Loading Libraries

## A screen shot of a computer program AI-generated content may be incorrect.

* Data Importing & Cleaning

## A computer screen shot of a computer code AI-generated content may be incorrect.

## Question 1: Descriptive Analysis

## A screenshot of a computer program AI-generated content may be incorrect.

## Question 2: t-Test

## A computer screen shot of a program AI-generated content may be incorrect.

* **Question 3: Factor Analysis (PCA)**

A screenshot of a computer program

AI-generated content may be incorrect.

* **Question 4: Clustering (K-means)**

**A computer screen shot of a program

AI-generated content may be incorrect.**