IDENTIFYING CUSTOMER PROFILES THROUGH K-MEANS CLUSTERING

SHREYA MURALEEDHARAN SWAPNA

INTRODUCTION

This analysis utilized k-means clustering to categorize customers according to demographic characteristics and purchasing patterns. Key features—Age, TotalChildren, and TotalSpent—were carefully engineered to enhance model performance. Marital Status and Education were consolidated to streamline categorical variables for greater interpretability. Data integrity was prioritized by excluding records with missing income information and improbable outliers (specifically, individuals reporting age over 100). The dataset underwent standardization before clustering, ensuring the resulting customer segments would be meaningful and analytically robust.

ANSWER 1) Clustering Customers into Two Groups

K-means clustering was performed with k = 2 to identify distinct customer segments, and results were visualized using fviz_cluster() (*Figure 1.1*). The segmentation incorporated both demographic and purchasing behaviour variables. Variable selection was informed by correlation analysis and domain expertise, corroborated by a heatmap (*Figure 1.2*). The final feature set included *Income, Recency, MntWines, MntMeatProducts, NumWebPurchases, NumCatalogPurchases, and Customer_Since_Days*, thereby capturing aspects of financial capacity and multi-channel shopping patterns. Cluster validity was assessed via a *Silhouette Score of 0.3013*, suggesting moderate separation between groups. Additionally, the clusters accounted for *30.1% of the total variance*, supporting the robustness of the segmentation approach.

The two clusters revealed meaningful behavioural differences:

- *Cluster 1*: Includes *lower-income, less engaged* customers with *longer recency* and *lower spending*. This group may represent price-sensitive or dormant consumers.
- *Cluster 2*: Represents *high-income*, *recently active*, and *high-spending* individuals who show strong engagement, particularly through digital channels.

This segmentation offers actionable strategic value, with Cluster 1 representing lower-engagement customers ideal for re-engagement campaigns, promotions, or loyalty-building efforts, while Cluster 2 consists of high-value customers suited for retention, premium positioning, and upselling. The PCA cluster plot (*Figure 1.3*) visually confirms the clear separation between these segments, allowing marketers to quickly identify natural groupings and derive insights without extensive data analysis.

ANSWER 2) Choosing two variables and visualizing the clustering

The two variables selected for clustering visualization are *MntWines* (amount spent on wine) and *NumWebPurchases*(number of purchases made through the company's website). These variables were chosen because they reflect key customer behaviours—*spending power and digital engagement.* Their inclusion helps distinguish between high-value digital shoppers and low-engagement, low-spending customers.

To visualize the clusters, a scatter plot was created using these two features with k-means clustering (k = 2). Each point represents a customer, coloured by cluster, and labelled by customer ID *(Figure 2.1)*.The silhouette score for this two-variable clustering was calculated to be approximately **0.30**, indicating moderate separation between the clusters.

- *Statistical Observation*: Customers in *Cluster 2* tend to have higher spending on wines and a greater number of web purchases, while *Cluster 1* comprises customers with lower wine expenditure and fewer online purchases.
- *Intuitive Interpretation*: *Cluster 2* customers frequently shop online and invest more in luxury products like wine, making them ideal candidates for personalized digital campaigns, loyalty rewards, or premium subscription offerings. In contrast, *Cluster 1* appears to consist of low-engagement, low-value customers who may benefit from re-engagement efforts, targeted discounts, or cross-channel outreach to boost conversion.

This segmentation allows the marketing team to adopt differentiated strategies: focus on retention and upselling for Cluster 2, and activation or win-back campaigns for Cluster 1.

ANSWER 3) Describe a Buyer Persona for Each Cluster

To translate the clustering results into actionable marketing strategies, two buyer personas were developed, each representing one of the customer segments identified through the k-means clustering analysis.

Cluster 1 is represented by *Olivia Sebastian (Figure 3.1)*, a customer with lower income, longer recency, and limited engagement in online and catalogue-based shopping. She tends to make occasional purchases, often spending less on luxury items like wine and meat products. Olivia represents a segment that is less engaged and may be price-sensitive, making her an ideal target for discount offers, re-engagement campaigns, and onboarding promotions that encourage digital interaction.

Cluster 2 aligns with *Ethan Jones (Figure 3.2)*, a high-income, high-spending customer who is digitally savvy and frequently purchases through web and catalogue channels. He spends more on wine and meat products, indicating a preference for premium goods, and has a relatively recent purchase history. Ethan exemplifies a loyal, high-value customer who responds well to personalized recommendations, exclusive loyalty benefits, and targeted upselling strategies.

These personas were developed using the most influential demographic and behavioural variables from the clustering process—Income, Recency, product-specific spending patterns, digital channel engagement, and tenure with the company. Each persona captures the statistical essence of its respective cluster and offers a practical lens through which marketers can deploy segment-specific strategies for activation, retention, or growth.

ANSWER 4) Evaluating Alternative Cluster Solutions- Comparing 3, 4, and 5 Segments

To explore whether further segmentation improves clustering quality, k-means was applied with **k = 3, 4, and 5**, and results were visualized using PCA-style cluster plots.

- The *3-cluster solution (Figure 4.1)* presents well-separated groups with balanced sizes and minimal overlap. Each cluster has a clear center and spread, making it easy to interpret and suitable for identifying broad differences in customer demographics and purchasing patterns.
- The *4-cluster model (Figure 4.2)* offers greater granularity by uncovering subgroups within existing clusters. However, it introduces moderate overlap—especially in the central region—making interpretation more complex. While it may support more tailored strategies, this comes at the cost of clarity.
- The *5-cluster solution (Figure 4.3)* captures greater diversity and reveals micro-segments, potentially reflecting niche behaviours or hybrid personas. Yet, the increased overlap—particularly among clusters 3, 4, and 5—reduces separation and challenges meaningful differentiation.

In summary, increasing k from 3 to 5 improves segmentation depth but decreases interpretability. The *3-cluster solution* offers the most practical balance of simplicity, separation, and actionable insights, making it the strongest foundation for persona development and targeted marketing strategies.

ANSWER 5) Determining the Optimal Number of Clusters

The *Elbow Method (Figure 5.1)* was used to determine the ideal number of customer segments by evaluating within-cluster variation across different *k* values. A clear elbow is observed at $k = 2$, where the rate of WSS reduction sharply decreases, indicating that two clusters capture the most meaningful structure in the data. This aligns with previous analysis, where the 2-cluster solution revealed distinct groups with different engagement levels. The silhouette score of *0.3013* further supports the validity of this segmentation, suggesting moderate but interpretable separation.

Although higher cluster counts (k = 3–5) slightly increase explained variation, they introduce significant overlap and reduce clarity. Thus, $k = 2$ strikes the best balance between simplicity, separation, and business relevance.

CONCLUSION

K-means clustering with three segments offers the most actionable balance of interpretability and business insight for targeted marketing.

REFERENCES

• OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/
• Grammarly. (2025). Grammarly. Grammarly.com. https://app.grammarly.com/
• Kaggle Dataset: https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/input
• HubSpot Buyer Persona: https://blog.hubspot.com/marketing/buyer-persona-research
• Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. STHDA. https://www.sthda.com/english/

APPENDIX

This appendix contains supporting figures referenced throughout the report, including detailed cluster statistics, buyer persona summaries, and visual outputs from the k-means clustering analysis.

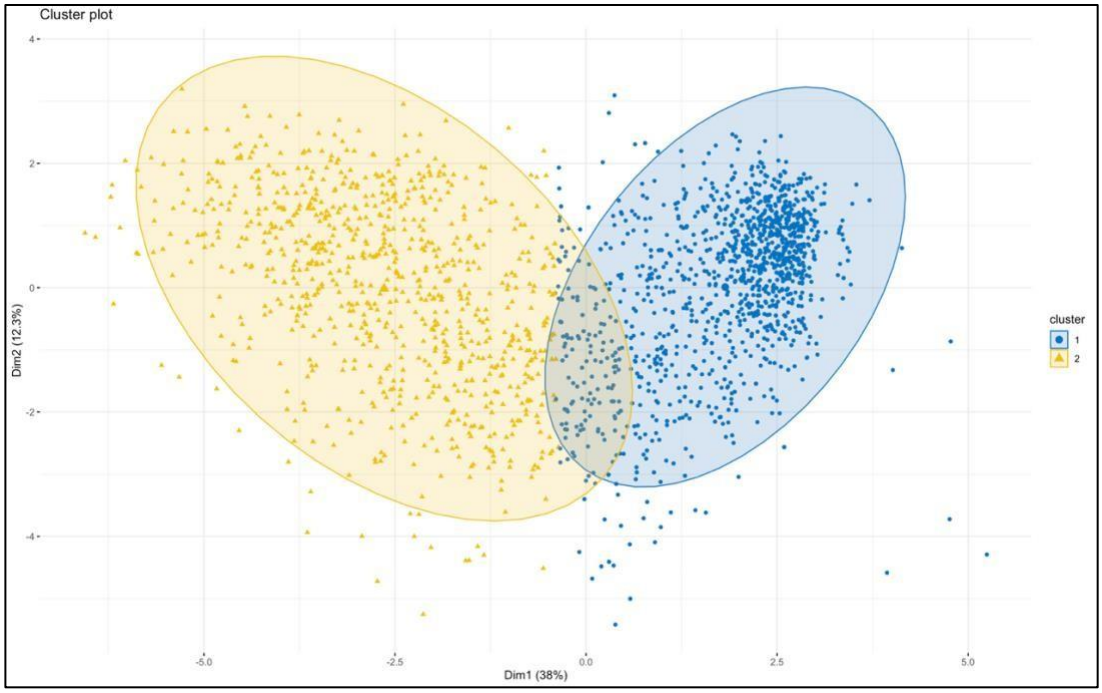

Figure 1.1 : Customer Segmentation Using K-Means Clustering (k = 2)

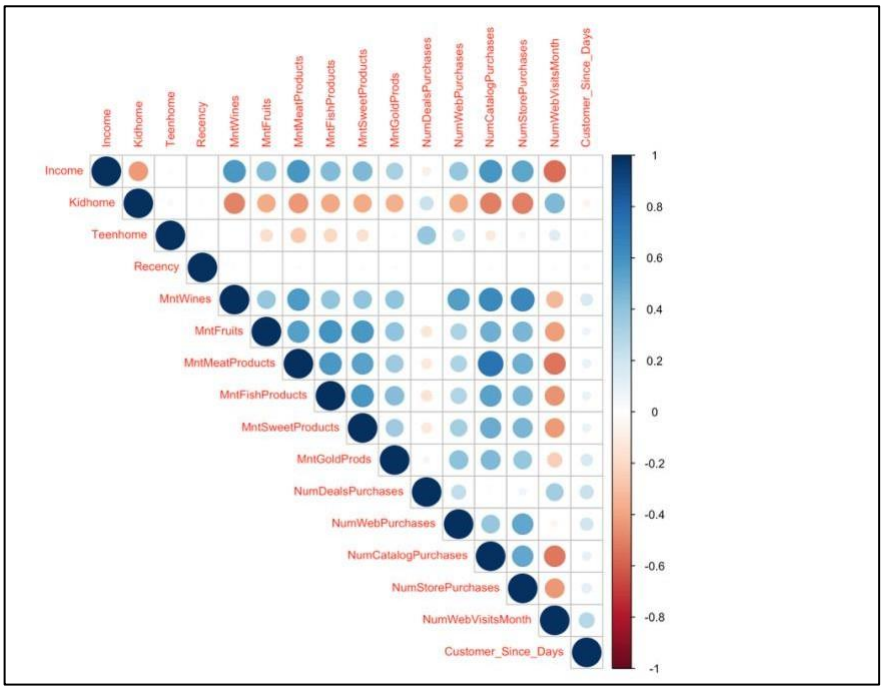

Figure 1.2: Heatmap showing Correlation between Key Customer Variables

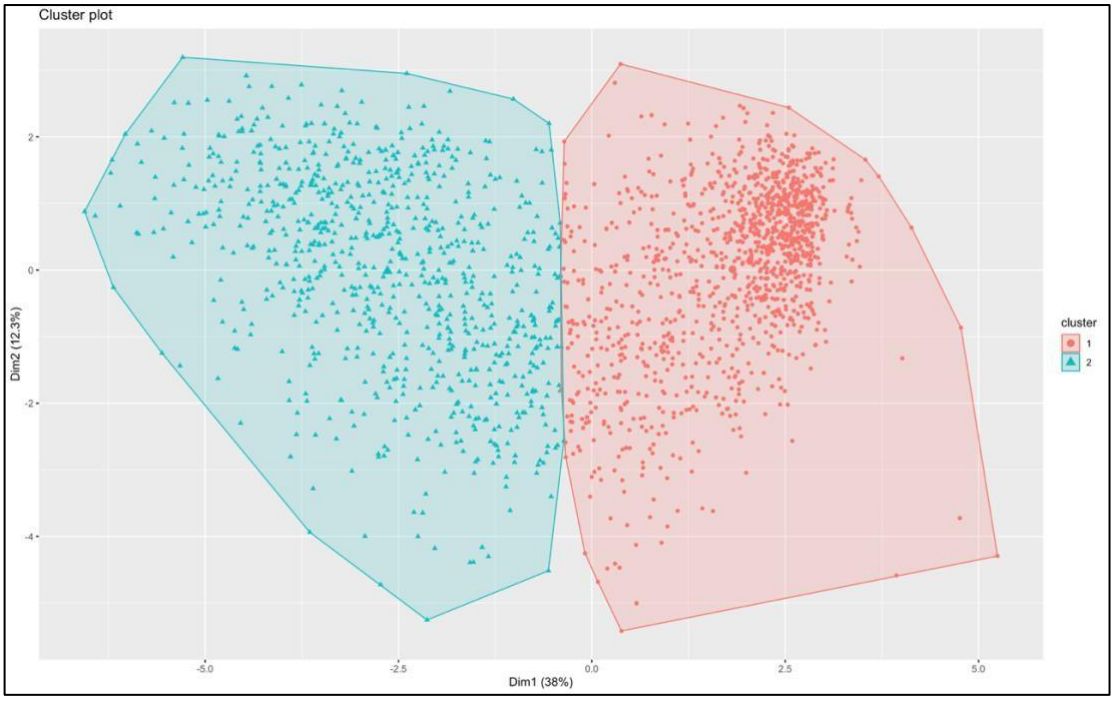

Figure 1.3: PCA-Based Cluster Visualization (k = 2)

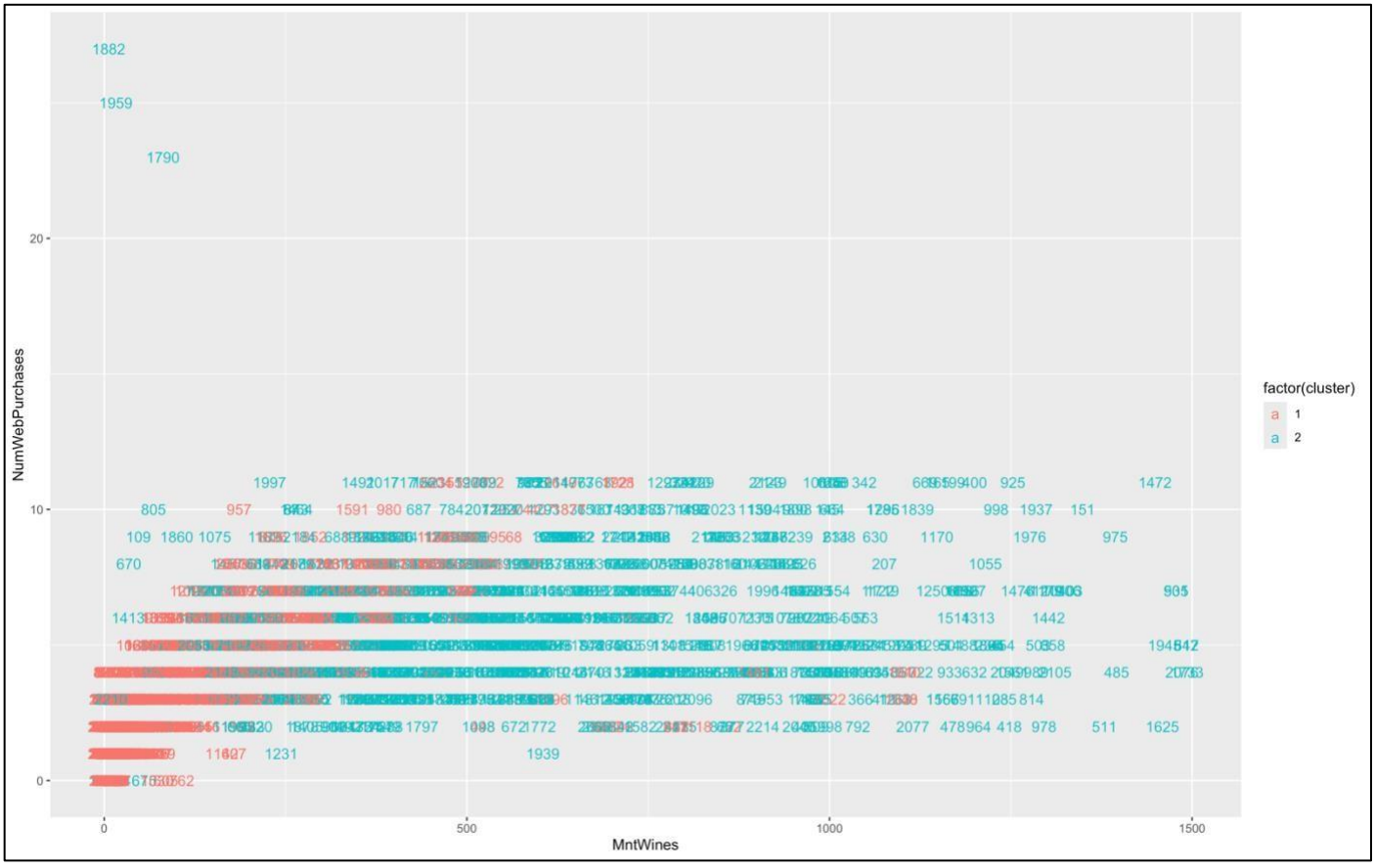

Figure 2.1 : Cluster Plot Using MntWines and NumWebPurchasest (k = 2)

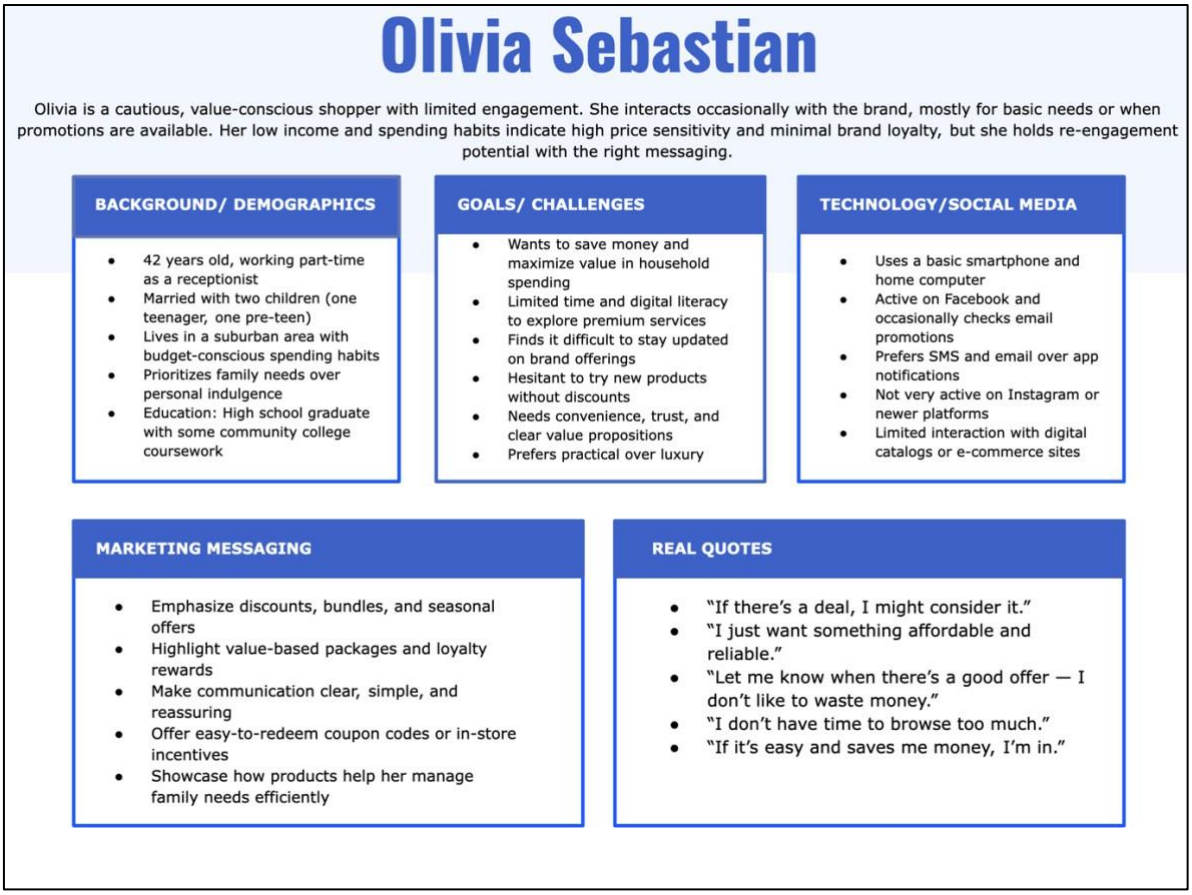

Figure 3.1: Buyer Persona – Value-Hunting Olivia (Cluster 1)

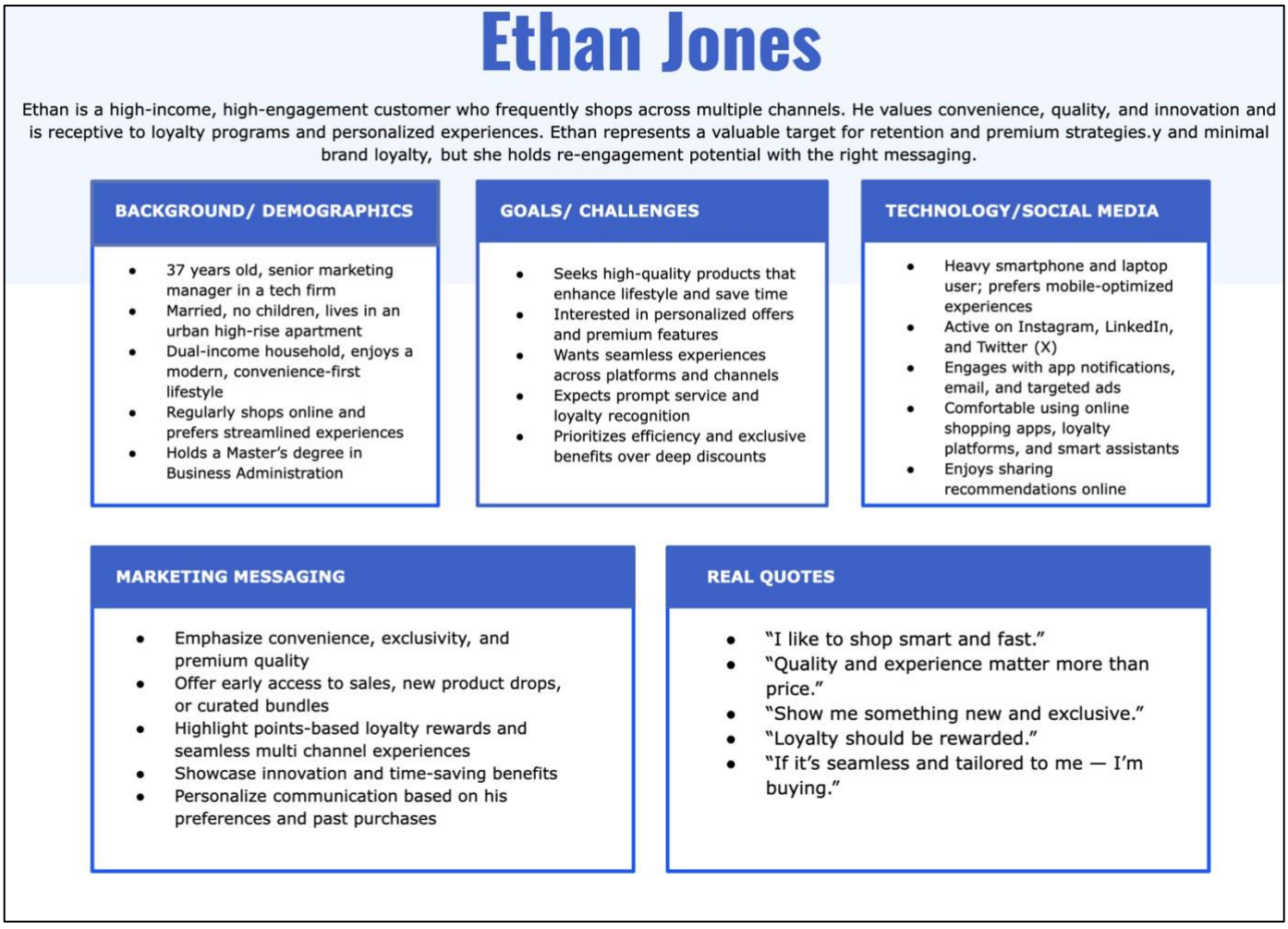

Figure 3.2: Buyer Persona – Engaged Ethan (Cluster 2)

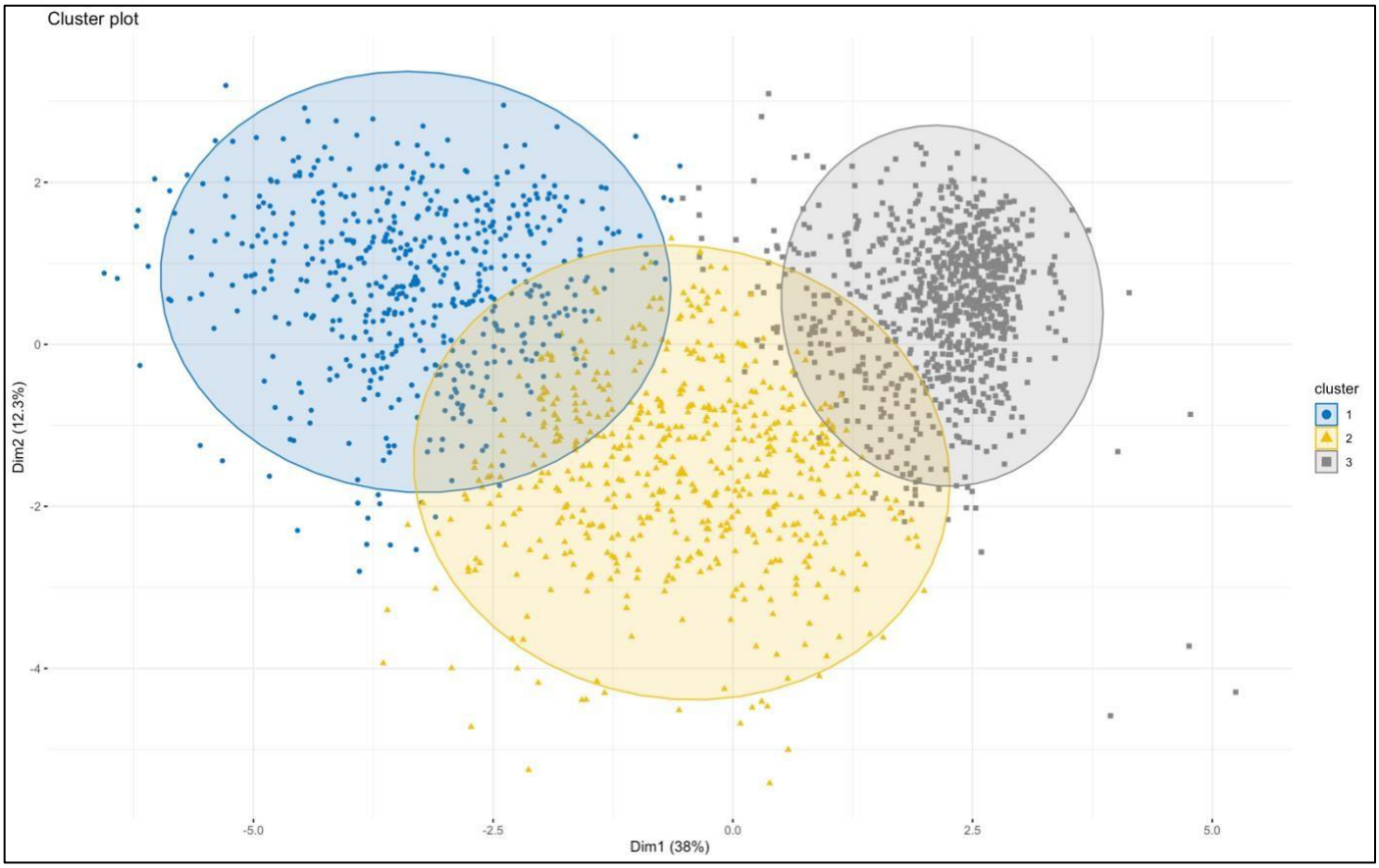

Figure 4.1 : Customer Segmentation Using K-Means Clustering (k = 3)

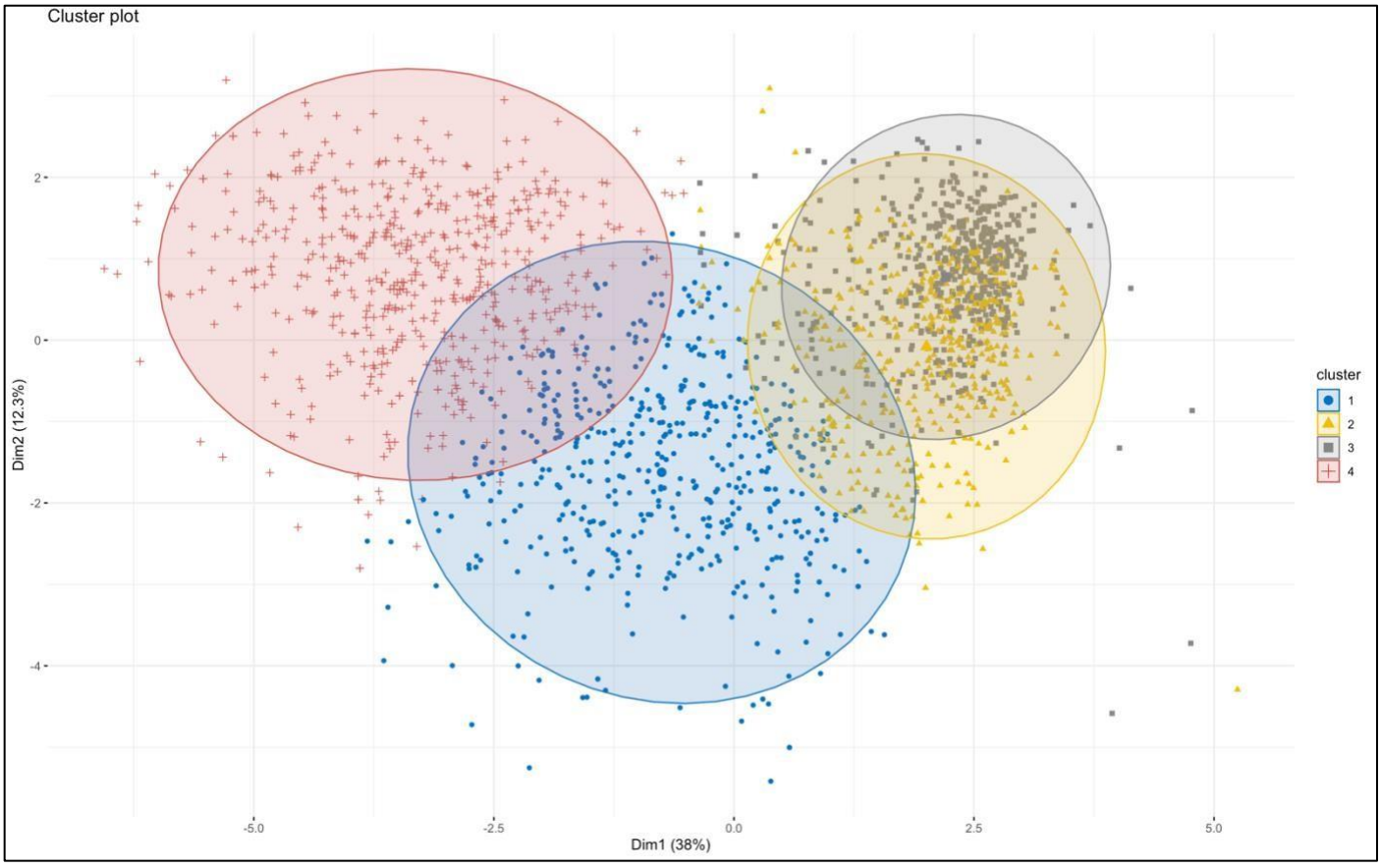

Figure 4.2 : Customer Segmentation Using K-Means Clustering (k = 4)

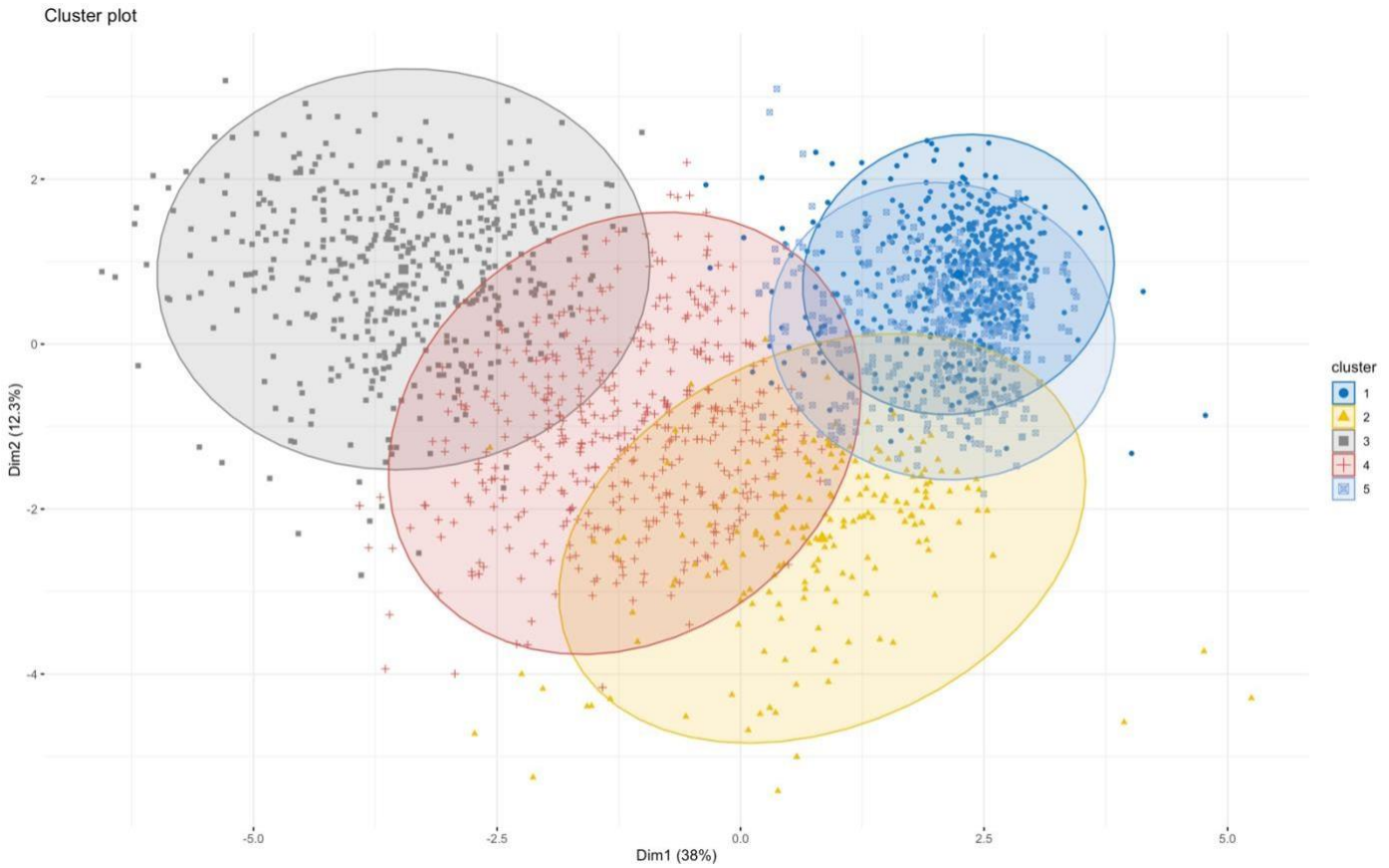

Figure 4.3 : Customer Segmentation Using K-Means Clustering (k = 5)

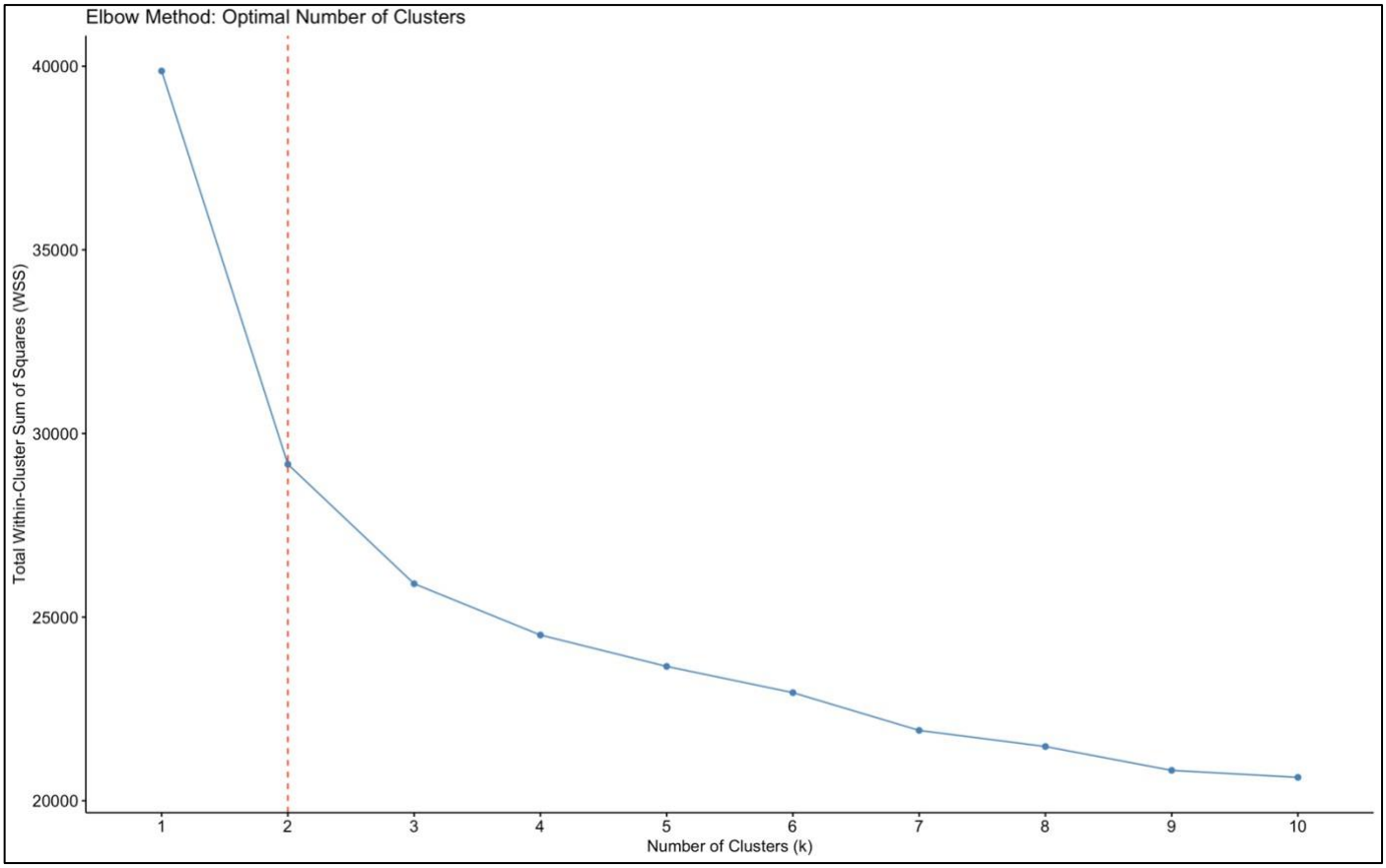

Figure 5.1***: Elbow Method to Determine Optimal Number of Clusters***