



OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

CLUSTER ANALYSIS

MARKET RESEARCH SURVEY

Taylor Tews
MSc Consumer Psychology Student,
University of Galway

PS6183 – Advanced Research Methods

April 9th, 2025



Executive Summary

A cluster analysis was conducted to segment consumers based on shopping behaviors, aiming to market grocery products more effectively.

Cluster Analysis

The data was cleaned and prepared for use in the analysis. The variables chosen to cluster were variables indicating the places consumers purchase goods at. A K-Means clustering strategy was used to determine there were three clusters present in the data. Each cluster has been assigned a distinct name based on key traits. These clusters are below.

Cluster 1: Web Wanderers

The Web Wanderers cluster views the website without purchasing items. Furthermore, this cluster is the youngest with a mean age of 53.8 and has the lowest mean income of €36,023.

Cluster 2: Offline Originals

Consumers belonging to the Offline Originals prefer to shop in-store or via the catalog and not online. This cluster has the highest mean income of €72,719, which could indicate a longing for a prestigious shopping experience while at the store.

Cluster 3: Balanced Buyers

The Balanced Buyers shopping behavior relies on the shopping at all different places including the catalog, in-store, and online. This cluster has the oldest mean age of 58.1.

Marketing Recommendations

Based on the key traits from each cluster, the following marketing actions were recommended in-line with consumer psychology literature

- Web Wanderers
 - Increase number of product reviews on website to leverage social proof which will increase perceived value of products for customers.
- Offline Originals
 - Appeal to the customers sense of touch and desire for prestigious products by incorporating smart technology into packaging.
- Balanced Buyers
 - Use an Omnichannel marketing strategy to mirror a similar marketing message across all consumer touchpoints during the shopping experience.

Table of Contents

	Page
Data Description & Cleaning	03
K-Means Clustering	04
Cluster Shopping Behavior	05
Cluster Demographics	08
Marketing Recommendations	10
References	13
Appendix	14

Data Description

The data is from customer records from a grocery firm's database. The entire database includes 2240 data points with 29 attributes. The dataset used to analyze the clusters in this analysis contains 1500 unique customer records. Appendix A is the data dictionary used to explain the variables included in this dataset. The variables are grouped by attributes aligned with the 4Ps of a marketing mix. The categories are person, price, place, and promotion. Data from these categories will help to understand each consumer and their shopping behaviors during analysis.

Data Cleaning



Handle for Missing Values

Before analysis, the dataset was cleaned for ease and precision of results. First, the data was checked for missing values in the dataset. Missing values can affect the quality and reliability of the results (Enders, 2004). The only missing data in the dataset was from the income variable, which had 19 missing values. To handle the missing values in the data, the MICE package was implemented. The MICE package imputes in data using a predictive model. The package used predictive mean matching to fill in the missing values of the income variable for the best match. Filling the income values using this method will help maintain quality and reliability in the analysis. No more missing values were present in the dataset after this step was completed.



Adding and Filtering Variables

The next step in data preparation ensuring proper variables for the cluster analysis. An Age variable was created from the current variable Year_Birth for easier comprehension of shoppers' age as a demographic variable. A summary of the dataset revealed Age to have a max value of 132. Next, any Age data points above 100 were filtered out and swapped with NA. This will exclude any age values that do not make logical sense and may have been errors in the data. Removing the errors in the data will clean the data to make the final analysis more accurate. Similarly, YOLO and Absurd in the variable Marital_Status have been filtered out of the column for having low numbers and not applying to the column.



Selecting Variables

A correlation test of the behavioral variables in the data set was executed to select variables for the cluster analysis. The correlation test mirrored the groupings in the data dictionary of the Products and Place category to be correlated the most. A cluster analysis test can be conducted using the variables in the products category as well as the Place category. For use in the cluster analysis, the variables in the Place category were chosen to understand how the company can best group individuals by where they shop to sufficiently target practical marketing information. Consumers are beginning to shop by online methods. Understanding where consumers shop can be critical to enhancing customer experience.

Furthermore, a second dataset was created to choose only the variables needed for the cluster analysis. The variables: ID, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, and NumWebVisitsMonth are the Place variables used in the cluster analysis.

Data Cleaning Continued

DATA NORMALIZATION

The final step in data cleaning was to normalize the data. This normalization step avoids any bias in the cluster analysis results due to scaling values since data clustering relies on distance measures. To scale the data, another data frame was created with standardization applied.

DATA OUTLIERS

Outliers were shown in the data after viewing the data from scaling. Data points further than 3.7 standard deviations away from the mean were removed to ensure accuracy of the data. 3.7 standard deviations was chosen as the threshold from observing the data to remove outliers without removing too many valuable data points. Next, multivariate outliers were removed from the dataset for the same reasoning. Finally with removal of all univariate and multivariate outliers, a final dataset was ready for use in cluster analysis.

K-Means Clustering

The use of both the Elbow method and the Silhouette method were compared in the selection of K for the cluster analysis. Each method indicated using K = 3 would be best suited for the dataset in the K-mean cluster analysis. This information is displayed by the graphs in Appendix C and Appendix D.

K-Means cluster analysis was then preformed using K = 3 and used the dataset prepared. From the analysis clusters were determined by distance metrics between the characters. Cluster 1 having 711 characters, Cluster 2 having 389 characters and cluster 3 having 345 characters. A table summary of the cluster characteristics is located below and will be discussed further in the report. The visualization to the right illustrates the grouping of these characters into clusters.

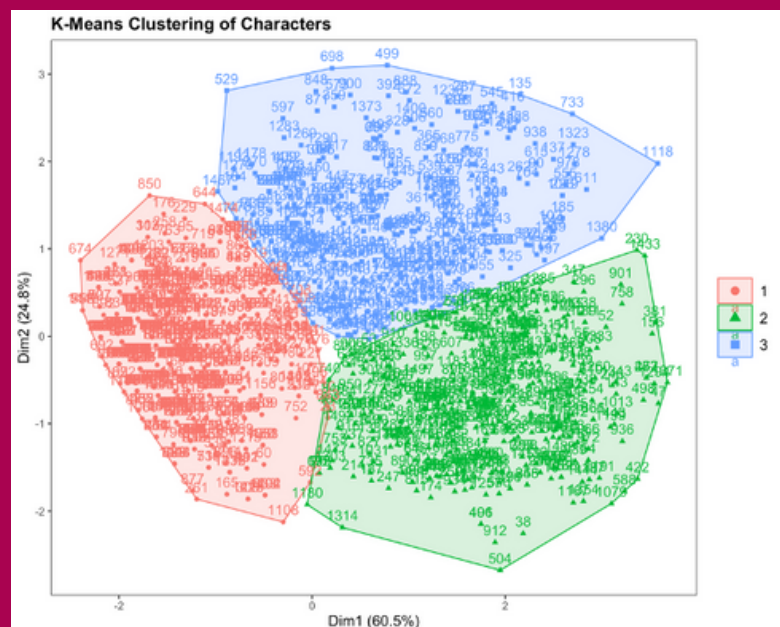


Figure 1: Clustering of Characters Visualization

Cluster Number	# of Web Purchases	# of Catalog Purchases	# of Store Purchases	# of Web Visits per Month	Character Count
1	-0.72	-0.69	-0.76	0.45	709
2	0.19	0.97	0.85	-1.24	388
3	1.20	0.09	0.57	0.43	345

Figure 2: Table Summary of Cluster Analysis

Cluster Shopping Behavior

CLUSTER 1: WEB WANDERERS

This cluster has significant negative behaviors in each category of the cluster analysis resembling a “shop but not buy” purchase behavior. The cluster analysis assigns scores of the following for this cluster:

◆ Number of Website Purchases Score: -0.72

A **high negative** number would indicate this cluster does not make many purchases from the website.

◆ Number of Catalog Purchases Score: -0.69

A **high negative** number would indicate this cluster does not make many purchases from the catalog.

◆ Number of Store Purchases Score: -0.76

A **high negative** number would indicate this cluster does not make many purchases from the store.

◆ Number of Website Visit Per Month Score: 0.45

This number is **low**, yet significant in the positive direction. This indicates the cluster may frequently visit the website.

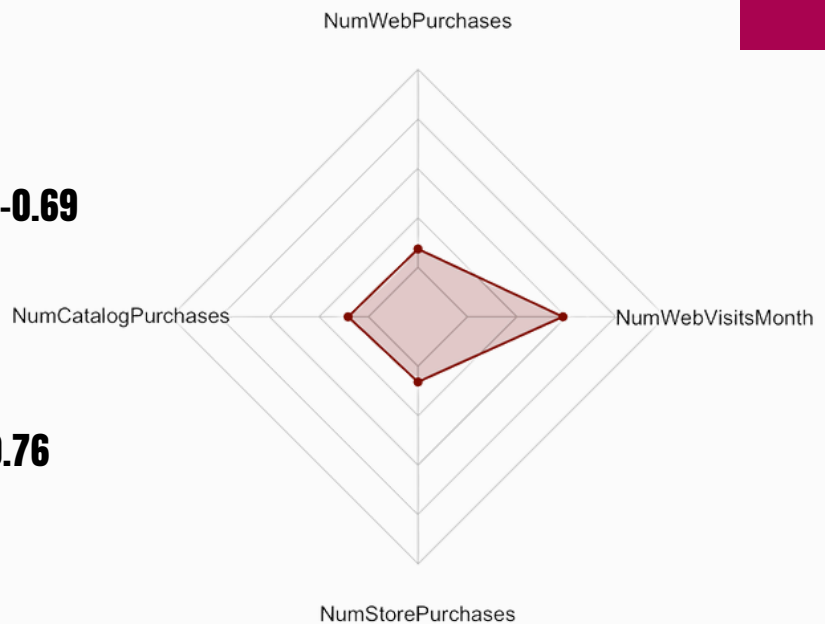


Figure 3: Radar Graph – Web Wanderers

KEY FINDINGS: WEB WANDERERS

Key takeaways from the results indicate cluster 1 to have low engagement into making purchases. Even though they have low engagement on making purchases, this cluster is active on the website. This skew in behavior patterns can be noticed on the radar graph above with the long stretch moving towards the number of website visits per month.

The purchase behavior of this cluster has been deemed with the name of an Web Wanderers. The Web Wanderers cluster is known for their stalking by viewing the website, but not making purchases on the website, store, or catalog place of sales.

Cluster Shopping Behavior

CLUSTER 2: OFFLINE ORIGINALS

This cluster has significant positive cluster analysis scores in each category of the cluster analysis that resembles physical shopping purchase behavior. The cluster analysis assigns scores of the following for this cluster:

◆ Number of Website Purchases Score: 0.19

This score is **low** indicating the possibility of online purchases from this consumer, not a driving force of behavior.

◆ Number of Catalog Purchases Score: 0.97

A high number would indicate this cluster **does** make many purchases from the catalog.

◆ Number of Store Purchases Score: 0.85

A high score in this category means this cluster **does** make in-store purchases.

◆ Number of Website Visit Per Month Score: -1.24

A **high negative** score in this category means this cluster does not visit the website.

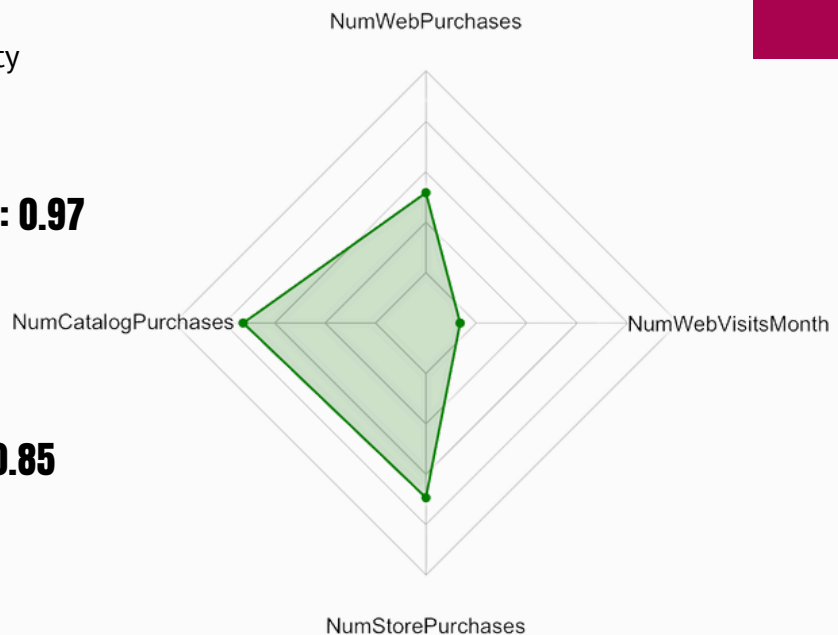


Figure 4: Radar Graph – Offline Originals

KEY FINDINGS: OFFLINE ORIGINALS

The analysis of this cluster is driven by the high scores in both Catalog and Store purchases and the strong negative score in the number of website visits per month. This indicates behavior of traditional shopping, formulating the name of a Offline Originals. Traditional shopping is the shopping behaviors that came before the digital era of online shopping. The Offline Originals are going to shop using the company's catalog and in-store purchases. The graph above illustrates the divide from this cluster's emphasis away from online modes of purchasing behaviors.

Cluster Shopping Behavior

CLUSTER 3: BALANCED BUYERS

This cluster has balanced purchased behaviors in each category of the cluster analysis. The cluster analysis assigns scores of the following for this cluster:

★ Number of Website Purchases Score: 1.20

This score is **high** indicating online shopping is a typical behavior for individuals in this cluster.

★ Number of Catalog Purchases Score: 0.09

This number is close to 0 meaning this cluster is neither for/or against shopping from the catalog, but it will not be a significantly strong purchase behavior either way.

★ Number of Store Purchases Score: 0.57

A significant number to indicate this cluster makes store purchases.

★ Number of Website Visit Per Month Score: 0.43

Similarly, to the number of stores purchases this customer also has the behavior of visiting the website for the company.

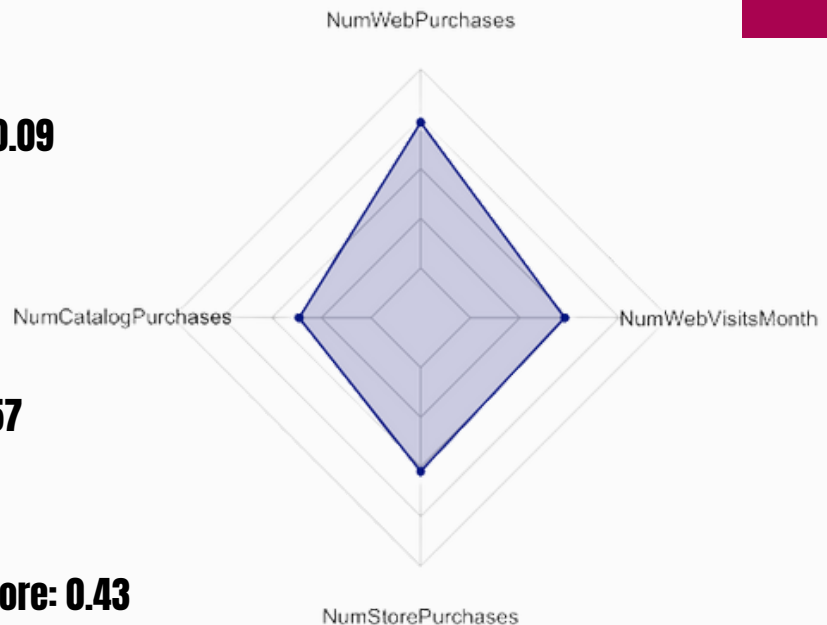


Figure 5: Radar Graph – Balanced Buyers

KEY FINDINGS: BALANCED BUYERS

This cluster is well balanced in many places to shop. The shoppers view the online website for the company. Their behaviors align with making purchases via store and website purchases. Due to the nature of no behaviors in this cluster leaning away from purchase behaviors, this group is named the Balanced Buyers. The balance of purchase behaviors is noted on the radar graph above with emphasis into web purchasing.

Cluster Demographics

The demographic make-up of each cluster is an integral part to the location of shopping behaviors described above. In each cluster there are similarities and differences but ensuring marketing actions are targeted towards the clusters demographics and habitual shopping behaviors will elicit the most success. This section of the report will analyze and connect demographics of the clusters with the shopping behaviors. Below is the average demographic information for each. All clusters top percentage of Education and marital status matched, therefore the percentage of the top two for both of those categories were added up to get a percentage.

CLUSTER 1: WEB WANDERERS

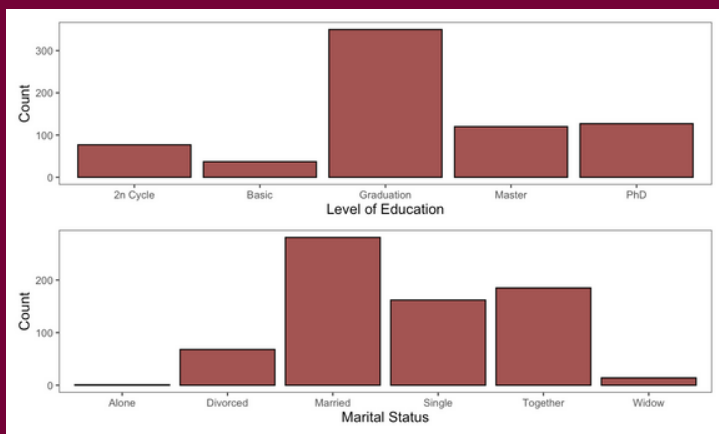


Figure 6: Education and Marital Status - Web Wanderers

Mean Age: 53.8

Mean Income: €36,023

Education: 32.99% (Graduated and PhD)

Marital Status: 32.30% (Married and Together)

CLUSTER 3: OFFLINE ORIGINALS

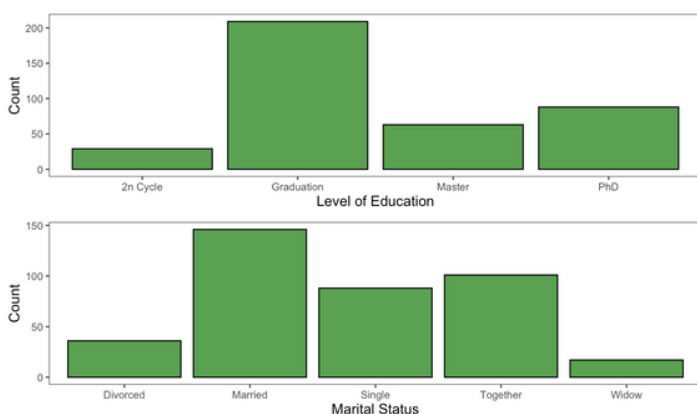


Figure 7: Education and Marital Status - Offline Originals

Mean Age: 57.8

Mean Income: €72,719

Education: 20.59% (Graduated and PhD)

Marital Status: 17.10% (Married and Together)

CLUSTER 3: BALANCED BUYERS

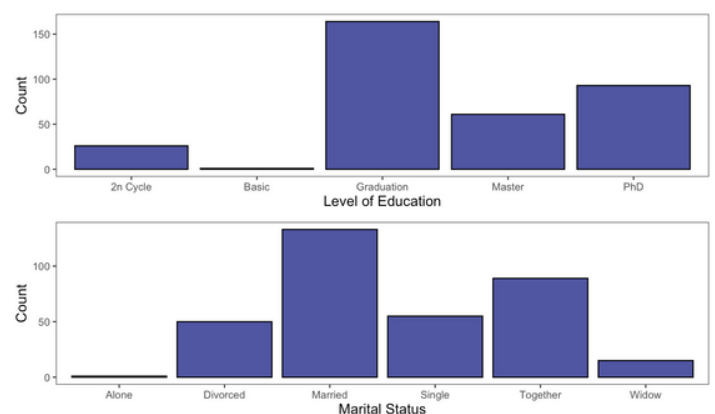


Figure 8: Education and Marital Status - Balanced Buyers

Mean Age: 58.1

Mean Income: €58,248

Education: 17.74% (Graduated and PhD)

Marital Status: 15.39% (Married and Together)

Cluster Demographics:

Key Similarities and Differences

Across all clusters, the mean age remains in the 50s. The average age of the clusters in the Generation X category is between 1961 and 1980 birth years (Gurău, 2012). The population can range between reaching a possible peak in careers and nearing retirement. Though the ages are close to one another, an ANOVA test revealed there is a significant difference in ages across clusters ($p\text{-value} = 1.48\text{e-}10$). A Tukey analysis revealed the most notable difference from this category's clusters' is the Web Wanderers (Cluster 1) are the youngest (53.8) versus the other two clusters' ages, 57.8 (Offline Originals) and 58.1 (Balanced Buyers). This population tends to have a separation between digitally proficient individuals and individuals who stray away from technology (Lissitsa & Kol, 2016). This age could explain the Web Wanderers' amount of website visits having a slightly younger mean age than the other clusters.

The section on average income has more differences than similarities. An ANOVA test revealed significant differences across clusters for incomes ($p\text{-value} = <2\text{e-}16$). Similarly, a Tukey test confirmed that the differences between all groups were significant. The Web Wanderers have the lowest average income of €36,023. This low income level may relate to how this cluster has the purchasing behavior of shopping online but rarely purchasing any products. The balanced buyers have the second-highest mean income of €58,248. This income level allows for this cluster to have flexibility in their purchases. This cluster commonly makes store and website purchases with reliance on viewing the websites. An assumption is that this cluster may be viewing the company website to compare prices, as they have enough money to buy products but not enough to disregard pricing. Finally, the Offline Original cluster has an average income of €72,719. This cluster may be less focused on the price of products with a high-income level and enjoy the physical shopping in-store and catalog.

The education and marital status of all three clusters have the highest percentage of the cluster in Graduated or PhD and Married and Together as their one and two highest percentage. Therefore, I have collected the percentages of individuals in the clusters for comparison between groups. The Web Wanderers have the highest percentage of educated individuals and the highest percentage of individuals with a partner. This demographic may be busy daily, only allowing time to surf the web to shop. An interesting point is that Offline Originals have a smaller percentage of educated individuals and smaller couples when related to Web Wanderers. Similarly, the Balanced Buyers record even smaller percentages in both categories.

Cluster	Mean Age	Mean Income	Education (%)	Marital Status (%)
Web Wanderers	53.8	€36,023	32.99	32.30
Offline Originals	57.8	€72,719	20.59	17.10
Balanced Buyers	58.1	€58,248	17.74	15.39

Figure 9: Summary of Demographic by Cluster

Marketing Recommendations

Clustering based upon location shopping behavior occurs allowed the cluster analysis to reveal the varied shopping behavior of individuals in similar demographics. Each of the clusters provides the company with different opportunity to accurately reach each cluster of consumers. The marketing recommendations below are tailored towards each cluster based upon shopping behavior and demographic key takeaways supported with consumer psychology literature.

Web Wanderers

A couple different marketing techniques should be used to target the Web Wanderers cluster. The techniques rely on the clusters' key characteristics of being potentially price-sensitive due to low income, high website engagement, and hesitancy to make purchases.

One consumer psychology framework that can be leveraged to encourage purchase behavior for price-sensitive individuals is increasing methods that emphasize social proof. The Web Wanderers are viewing the website without making a purchase. This behavior may insinuate a high-perceived risk in purchasing the product. One method could include increasing online reviews available on the website for products. Increasing access to online reviews can nudge consumers towards believing increased perceived value of the products based on the theory of social proof (Cialdini & Goldstein, 2004). Specifically, in online grocery shopping, online reviews are critical for the evaluation of how consumers view the products and experience. To make the reviews more impactful, having the company promote customer to leave detailed reviews rather than numerical rates will furthermore support the persuasion of this marketing strategy (Wang et al., 2024).



Marketing Recommendations

Continued

Offline Originals

The Offline Originals have a high average income and prefer to buy products in person or from the catalog. From the key characteristics making this group, touch appears to be an integral part of the shopping behavior. Companies like Coca-Cola have found success in changing the shape of soda bottles to appeal to consumers' sensory desires (Spence & Gallace, 2011). Specifically, even nondiagnostic haptic cues have been found to have an impact on product judgments (Krishna & Morrin, 2008). Furthermore, packaging in the food industry is constantly adapting and changing for different regulations. The industry has developed different packaging systems to combat the changes and attract new customers.

For example, an intelligent packaging system is a food packaging system that incorporates monitors into the food packaging to let the customer know real time data about the purchase (Yan et al., 2022). Incorporating an intelligent packaging system into the product profile may help the Offline Originals understand the prestige and high tech of the product offering.



Balanced Buyers

A key strength of the Balanced Buyer Shopper is this shopper uses many different channels to shop and receive information from. An Omnichannel marketing approach needs to be the main priority to market to this consumer. Omnichannel marketing refers to integrating the customer experiences with touchpoints on multiple different channels (Nguyen et al., 2022). This approach will help the consumer view consistency in the various places they shop (i.e. online, in-store, and catalog).

Marketers need to be cognizant of cognitive overload for the consumers in this category. Excessive information can cause the consumer to have a reduction in decision quality (Meyer, 1998). The information the Balanced Buyers are receiving entering across many marketing channels. If the message is consistent using the omnichannel marketing approach the consumers will have less decision fatigue.



References

- Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, 55(1), 591–621. <https://doi.org/10.1146/annurev.psych.55.090902.142015>
- Comparing Generation X and Millennial consumers. *Journal of Consumer Marketing*, 29(2), 103–113. <https://doi.org/10.1108/07363761211206357>
- Enders, C. K. (2004). The Impact of Missing Data on Sample Reliability Estimates: Implications for Reliability Reporting Practices. *Educational and Psychological Measurement*, 64(3), 419–436. <https://doi.org/10.1177/0013164403261050>
- Gurău, C. (2012). A life-stage analysis of consumer loyalty profile: Comparing Generation X and Millennial consumers. *Journal of Consumer Marketing*, 29(2), 103–113. <https://doi.org/10.1108/07363761211206357>
- Krishna, A., & Morrin, M. (2008). Does Touch Affect Taste? The Perceptual Transfer of Product Container Haptic Cues. *Journal of Consumer Research*, 34(6), 807–818. <https://doi.org/10.1086/523286>
- Lissitsa, S., & Kol, O. (2016). Generation X vs. Generation Y – A decade of online shopping. *Journal of Retailing and Consumer Services*, 31, 304–312. <https://doi.org/10.1016/j.jretconser.2016.04.015>
- Meyer, J. (1998). Information overload in marketing management. *Marketing Intelligence & Planning*, 16(3), 200–209. <https://doi.org/10.1108/02634509810217318>
- Nguyen, A., McClelland, R., Hoang Thuan, N., & Hoang, T. G. (2022). Omnichannel marketing: Structured review, synthesis, and future directions. *The International Review of Retail, Distribution and Consumer Research*, 32(3), 221–265. <https://doi.org/10.1080/09593969.2022.2034125>
- Reliability Estimates: Implications for Reliability Reporting Practices. *Educational and Psychological Measurement*, 64(3), 419–436. <https://doi.org/10.1177/0013164403261050>
- Spence, C., & Gallace, A. (2011). Multisensory design: Reaching out to touch the consumer. *Psychology & Marketing*, 28(3), 267–308. <https://doi.org/10.1002/mar.20392>
- Wang, Y., Xu, X., & Zhu, Q. (2024). Investigating the Signaling Mechanism of Consumer Reviews Toward Online Grocery Shopping. *IEEE Transactions on Engineering Management*, 71, 10727–10739. <https://doi.org/10.1109/TEM.2024.3403938>
- Yan, M. R., Hsieh, S., & Ricacho, N. (2022). Innovative Food Packaging, Food Quality and Safety, and Consumer Perspectives. *Processes*, 10(4), 747. <https://doi.org/10.3390/pr10040747>

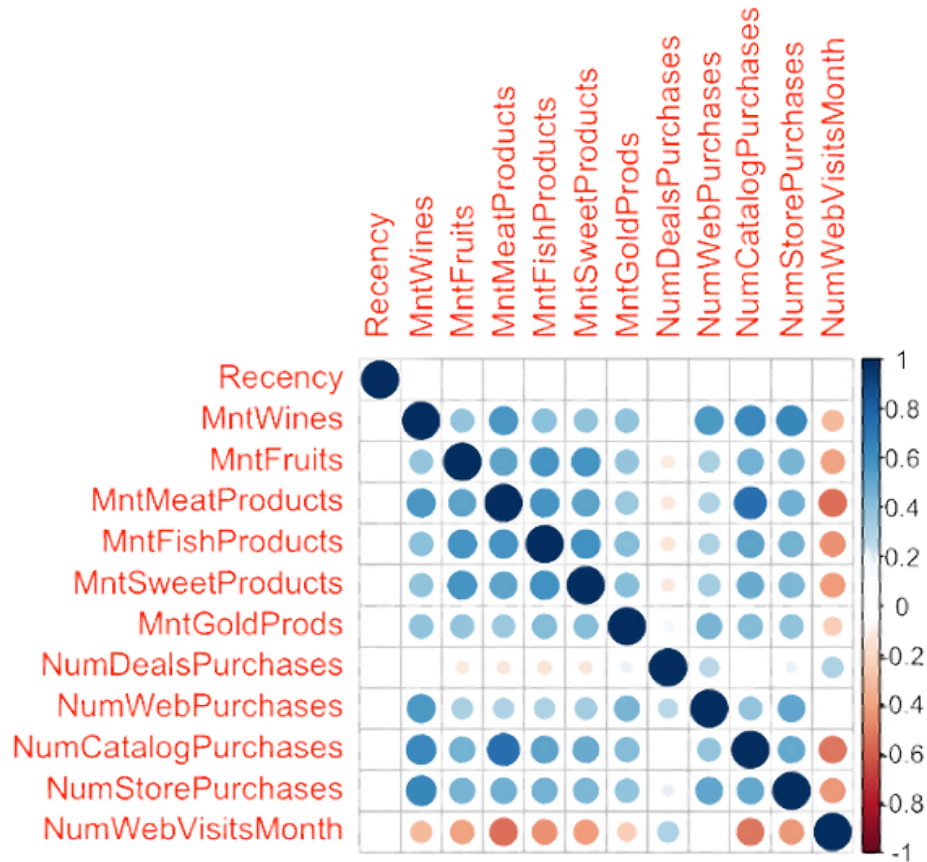
Appendix

Appendix A: Data Dictionary

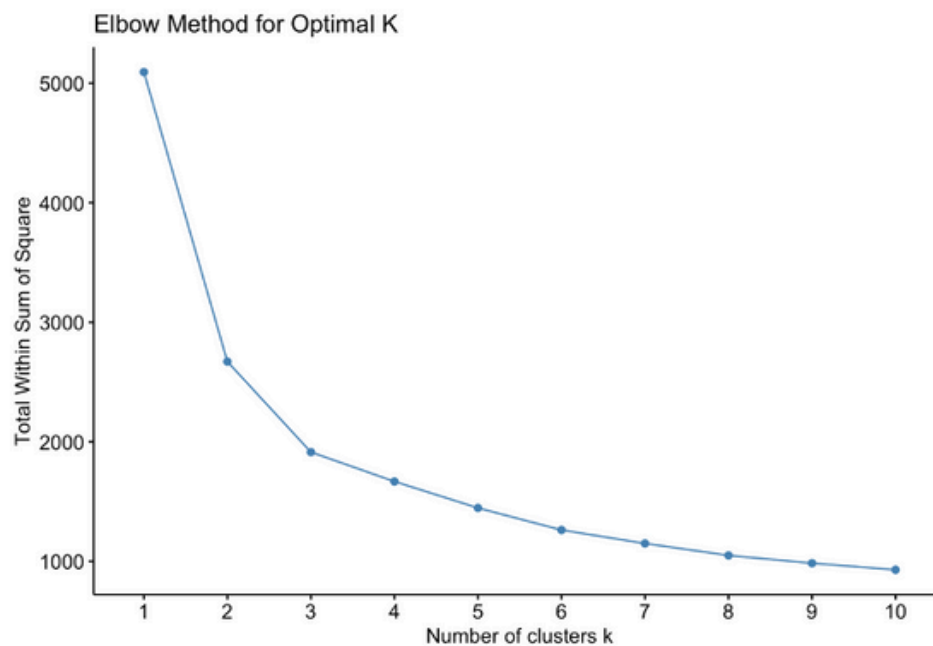
CATEGORY	VARIABLE	DESCRIPTION
PEOPLE	ID	Customer's unique identifier
	Year_Birth	Customer's birth year
	Education	Customer's education level
	Marital_Status	Customer's marital status
	Income	Customer's yearly household income
	Kidhome	Number of children in customer's household
	Teenhome	Number of teenagers in customer's household
	Dt_Customer	Date of customer's enrollment with the company
	Recency	Number of days since customer's last purchase
	Complain	1 if the customer complained in the last 2 years, 0 otherwise
PRODUCTS	MntWines	Amount spent on wine in last 2 years
	MntFruits	Amount spent on fruits in last 2 years
	MntMeatProducts	Amount spent on meat in last 2 years
	MntFishProducts	Amount spent on fish in last 2 years
	MntSweetProducts	Amount spent on sweets in last 2 years
	MntGoldProds	Amount spent on gold in last 2 years
PROMOTION	NumDealsPurchases	Number of purchases made with a discount
	AcceptedCmp1	1 if accepted the offer in the 1st campaign, 0 otherwise
	AcceptedCmp2	1 if accepted the offer in the 2nd campaign, 0 otherwise
	AcceptedCmp3	1 if accepted the offer in the 3rd campaign, 0 otherwise
	AcceptedCmp4	1 if accepted the offer in the 4th campaign, 0 otherwise
	AcceptedCmp5	1 if accepted the offer in the 5th campaign, 0 otherwise
	Response	1 if accepted the offer in the last campaign, 0 otherwise
PLACE	NumWebPurchases	Number of purchases made through the company's website
	NumCatalogPurchases	Number of purchases made using a catalogue
	NumStorePurchases	Number of purchases made directly in stores
	NumWebVisitsMonth	Number of visits to company's website in the last month

Appendix

Appendix B: Correlation Test

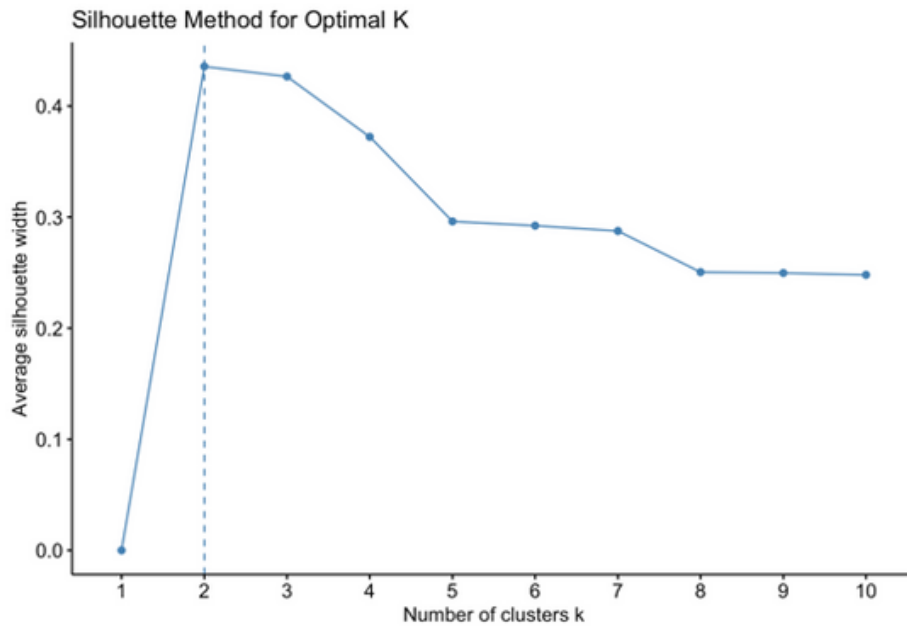


Appendix C: Elbow Method Test for Choosing K



Appendix

Appendix D: Silhouette Method Test for Choosing K



Appendix E: Summary of Cluster Characteristics Table

Cluster	NumWeb Purchases	NumCatalog Purchases	NumStore Purchases	NumWebVisits Month
1	-0.72	-0.69	-0.76	0.45
2	0.19	0.97	0.85	-1.24
3	1.20	0.09	0.57	0.43

Appendix F: Summary of Education by Cluster

Cluster #	Education	Count	Percentage
1	Graduation	350	24.2
1	PhD	127	8.79
1	Master	120	8.30
1	2n Cycle	77	5.33
1	Basic	37	2.56
2	Graduation	209	14.5
2	PhD	88	6.09
2	Master	63	4.36
2	2n Cycle	29	2.01
3	Graduation	164	11.3
3	PhD	93	6.44
3	Master	61	4.22
3	2n Cycle	26	1.80
3	Basic	1	0.0692

Appendix

Appendix G: Summary of Marital Status by Cluster

Cluster #	Marital Status	Count	Percentage
1	Married	281	19.5
1	Together	185	12.8
1	Single	162	11.2
1	Divorced	68	4.72
1	Widow	14	0.971
1	Alone	1	0.0693
2	Married	146	10.1
2	Together	101	7.00
2	Single	88	6.10
2	Divorced	36	2.50
2	Widow	17	1.18
3	Married	133	9.22
3	Together	89	6.17
3	Single	55	3.81
3	Divorced	50	3.47
3	Widow	15	1.04