Nottingham University
Business School
UK | CHINA | MALAYSIA

Exploratory Research: Introducing Consumer Stability as an Informing Customer Segmentation Feature

By: Jonathan James Lankfer

Supervisor: Madeleine Ellis

Student ID: 14321456

Course: Business Analytics MSc

Year of Publication: 2022

A Dissertation presented in part consideration for the degree of *Business Analytics MSc, University of Nottingham*. The work is the sole responsibility of the candidate.

Abstract

Static simplicity has encapsulated the academic and industry application of customer segmentation ever since its introduction to the wider consumer behaviour discussion in the 1950s. Evolving traditional analytical methods to address their coarseness, this research provides justification for the introduction of consumer stability to extend and conceivably displace lifetime and single-visit basket analysis which saturates the existing segmentation conversation. Principal component analysis is applied to demonstrate consumer stability's importance, represented in this research by temporally lagged RFM measures within an FMCG scenario. A k-Means clustering model is implemented to extend this exploratory research, providing an example of how latent behaviours could provide an opportunity for marketers to enhance their output by evaluating temporal consumption pathways.

Key words: Principal Component Analysis, k-Means Cluster modelling, Customer Segmentation, Consumer stability, RFM

Acknowledgements

First, I would like to thank my supervisor Madeleine Ellis for her insightful comments, recommendations, and friendly conversation provided on this dissertation. This support, alongside guidance from the Business Analytics MSc teaching staff and N/LAB team, has contributed not only insight to this dissertation, but a deeper appreciation and awareness of the wider data analytics field. I would also like to express my sincere gratitude to the Business School for awarding me the Nottingham University Business School Alumni Scholarship which has allowed me the opportunity to undertake this MSc course. I am also incredibly thankful to UoN Sport for their continued funding and lifestyle support which has allowed me to pursue my sporting endeavours alongside my academic studies. A particular mention to my lifestyle advisor Freddie Fairbairn, who across the past four years has provided constant support and advice for all university life had to throw at me.

Table of Contents

List of Tables

List of Figures

1. Introduction

Customer segmentation attempts to detail heterogenous sub-groups from a wider homogeneous consumer market using statistical or analytical means, differentiated by shared behavioural characteristics or similarities. Exceeding an intuitive logic, an evidence-based method to understanding customer behaviour justifies and guides tailored business activity to each heterogenous group, principally through adaptive marketing delivery. Business analytics is providing an advanced sophistication to segmentation, discerned via segmentation accuracy and wider feature inclusion. Due to the dominance of segmentation that exists within consumer behaviour and business analytics literature, provoked by an industry acceptance of consumer importance, many traditional methods of segmentation have been comprehensively explored. Despite this acceptance of the complexity of consumers, there is a lack of research regarding a customer's stability to reflect how their behaviour differs over time.

This research aims to introduce consumer stability as a customer segmentation feature and explore its significance as an informing feature when performing customer segmentation. This introductory chapter will discuss the background and context behind the study, followed by the research purpose, methodology approach, and finally, the research outline.

1.1 Background

Customer segmentation has been a long-standing marketing practice and is universally accepted as a key directive within industry to gaining competitive advantage. The benefit of segmentation is encapsulated

under the assumption tailored marketing activity outperforms a blanket mass-marketing approach. The rise of the web and current internet era underpinning consumption behaviour has meant when tracked across time, consumers are in an unrivalled position of power against the corporations they are purchasing from. Accessibility to alternatives, substitute competition, and ease of price comparison are just some of the many forces enabled in today's internet era. Therefore, to retain competitive advantage within this customer-orientated market, companies need to further improve the effectiveness of their targeted behaviour, signalling the necessity for higher-performing customer segmentation methods.

Current customer segmentation practice relies primarily on the use of cluster modelling which is an unsupervised machine learning approach to form groups of customers based upon mutually shared features. Unsupervised learning refers to the application of analytics to data not already labelled; in the case of customer segmentation this is observed through an initially unsegmented customer base. Therefore, cluster modelling when applied to customer segmentation seeks to assign unlabelled customers to a sub-group in which they exhibit shared behavioural characteristics.

Historically, segmentation has utilised intuition or logic-driven rules which rely on some combination of demographic, geographic, psychographic, or behavioural indicators. Whilst still discussed in current consumer behaviour literature, computational developments of the 20th and 21st century has meant the capabilities for sophisticated analytics have progressively increased, driving the application of machine learning to customer segmentation forwards. Advancements comprise of increasingly complex data sets providing rich feature inclusion, meta-parameter tuning to enhance segmentation accuracy, and amplified capabilities to identify

non-intuitive underlying patterns that make up consumer segments. Therefore, whilst the bridge between consumer behaviour and machine learning is currently robust, the continual expanding scope of machine learning coupled with unexplored consumer behaviour considerations leaves customer segmentation primed for further address.

Epitomising the unperfected-nature of customer segmentation practice, Weinstein (1993, cited in Bock and Uncles 2002:215-216) refers to a US healthcare company who commissioned 18 segmentation studies over a 5-year period without adopting one. Highlighting the lack of effective segmentation delivery existing within industry, a disparity emerges in which current practice does not mirror the saturated customer segmentation literature depicting a well-established and optimised academic segmentation approach.

1.2 Research Purpose

Reaffirming the breadth of coverage customer segmentation has received across consumer behaviour and business analytics literature in a theoretical and applied context, the application of cluster modelling is well established as the appropriate method to achieving customer segmentation. Coverage given to clustering within customer segmentation predominantly involves the use of customer lifetime measures, such as Recency-Frequency-Monetary (RFM), or single-visit analysis using basket data as indicators of consumer behaviour. In spite of this coverage, both methods have the potential to deliver sub-optimal segmentation results. In the former, cognitive processes cannot be aggregated across a consumer's lifetime to mirror any given time frame. Denounced because generalised aggregation ignores formal temporal influxes such as seasonality as well as informal forces such as lifestyle changes. In parallel, the latter neglects consumption trends across time by assuming

behaviour can be duplicated based on single visit behaviour. Therefore, the unexplored nature of analysing a consumer's stability poses a forward-facing problem towards the application of customer segmentation within a business context, reflected in an unoptimized segmentation model and potential for sub-optimal business decisions.

Given the lack of academic consideration to changing customer behaviour, this research study aims to introduce consumer stability within customer segmentation literature. Underlying this aim, two research objectives emerge. Firstly, to utilise consumer stability as a feature when performing cluster modelling in application to customer segmentation. This objective seeks to discover if consumer stability is an informing feature when set against traditional customer lifetime methods to segment customers using cluster modelling? The second research objective is to explore consumer stability paths that are observed within consumer behaviour. This objective will seek to answer if common or dominant pathways exist, and are these optimal to business performance? The significance of the former could expose a new vein of further temporal business analytics research into consumer stability and possesses implications for cluster modelling accuracy when applied to customer segmentation. The significance of the latter could provide optimal consumer pathways, which if identified, can be employed by marketers within industry to enhance decision making, such as the application of evidence-based customer journey mapping. Therefore, this exploratory research holds valuable implications for both academia and industry, contributing a novel insight to the vast body of literature surrounding customer segmentation as well as providing a real-world value to marketing action within business practice.

1.3 Methodology Approach

This study presents a revised approach to traditional customer segmentation which establishes scope for temporal consumption behaviours to be expressed by customers, represented by lagged RFM measures drawn from a large FMCG transactional data source. Conceptually following an automatic latent temporal topic modelling study implementing Non-negative tensor factorisation (Smith et al. 2016), this exploratory research will apply Principal Component Analysis (PCA) to determine the effectiveness of using stability-based measures. PCA acts as a dimensionality reduction technique based on variance, applied in this context to evaluate feature importance. PCA will be supplemented by correlation analysis to further analyse the engineered consumer stability features. Subsequently, to discover possible stability trends, common pathways within the sub-groups will be evaluated using a k-Means clustering model to identify optimal consumer stability trends that are considered preferred, given contextual performance goals within the FMCG sector.

1.4 Research outline

This dissertation is organised as follows. It will begin in Chapter 1: Introduction, by introducing the research study with reference to the background of customer segmentation, purpose of the research, and the methodology taken.
This introductory background is supplemented in Chapter 2: Literature Review, which reviews the relevant literature to the topic, inclusive of background to customer segmentation, business analytics applications to the task, summary of k-Means cluster modelling techniques, and additional feature engineering methodology centred around PCA.

Chapter 3: Methodology and Research Design follows, which describes and justifies the methodology taken to introduce consumer stability within a customer segmentation approach. This will include an outline of the research design, data exploration completed, methodology to generate and evaluate temporal stability features via feature engineering and dimensionality reduction, and lastly a description of the k-Means modelling approach used.

In Chapter 4: Findings and Results, analytical results of the PCA and k-Means model are harnessed to demonstrate the suitability of consumer stability as an informing segmentation feature.

Mobilising the research findings and results, Chapter 5: Discussion and Limitations provides academic and industry relevant consequences of the research. This will revisit research objectives and answer associated questions that shaped the design and motivation of this research.

Lastly, Chapter 6: Conclusion wraps up with concluding remarks summarising the most significant outcomes of the completed research.

1.5 Summary and Research Gap

In summary, a research gap emerges amongst customer segmentation literature in which the consumer has been continually treated as static, violating the assumption of the consumer as a cognitively complex decision-maker. This exploratory research addresses this gap, introducing consumer stability as an informing feature when performing customer segmentation via cluster modelling. Smith et al. (2016) is positioned as an outlying piece of academic research amongst customer segmentation literature, providing a novel approach to mitigate the static coarseness traditional segmentation. This acts as a steppingstone by establishing scope for exploration into temporal representation of a non-static consumer, held under the hypothesis that consumer stability is an informing customer segmentation feature.

2. Literature Review

2.1 Background to Customer Segmentation

Convention has agreed that to succeed, a business needs to satisfy the customer; "if they [companies] take care of their customers, market share and profits will follow" (Kotler 2005:5). Unlike its agreed upon importance, a sufficient consumer-centric understanding and approach is yet to be defined. By nature, consumer behaviour is not an exact art. Driven by cognitive human instinct, decision-making is fluid and exceptionally unpredictable. Long before the depiction of the consumer being king (Kelly 1973), business practices have been fascinated by the concept of the consumer. Smith's seminal work announcing market segmentation recommends looking at a homogeneous market as fragmented, possessing heterogenous attributes requiring modified approaches to product preferences and segments' needs (Smith 1956). The notion of a set homogenous market has since been comprehensively discarded (Beane and Ennis 1987; Vyncke 2002; Engel et al. 1972). Accepting this heterogeneous market, a priority for businesses is understanding what fragments their audience, the principal motivator for customer segmentation.

The continued desire for complex consumer understanding has retained customer segmentation as a focal point for business attention to gain competitive advantage. Competitive advantage is observed through an increased understanding of the consumer and their needs (Engel et al. 1972; Hunt and Arnett 2004). Firstly, accurate understanding can allow marketing activity to be tailored to gain increased trust (Weinstein 2006). Secondly, appropriate adjustments can be made to shifting market demands and enable distinct offerings to consumers (Engel et al. 1972). Ultimately, the benefit of customer segmentation is achieving increased

profitability because a segmented marketing approach yields increased sales compared to blanket mass-marketing approaches (Cross 1999; Wind 1978). Defined as the 'heavy half' of the heterogenous market, a company can disproportionately increase performance by proactively targeting the segments that account for up to 80% of total sales (Twedt 1964:71-72). Weinstein completed a national segmentation study with 203 U.S marketing executives in the technology industry; concluding "Segmenters (concentrated and differentiated marketers) were found to be significantly more effective target marketers than nonsegmenters (undifferentiated marketers)" (Weinstein 1993:1-2), verifying the significance of Twedt's heavy half.

20th century awareness of its benefit meant customer segmentation occurred well in advance of refined statistical testing or current machine learning methods that can be realistically executed on developed computers and software (Jenson 1996; Pearl 1988). Underpinning early attempts were intuitive factors derived from qualitative interpretation and logic-driven consumer-understanding. Kotler (1997) refined this traditional segmentation practice, raising four primary variables utilised to perform a divisive segmentation of the market: demographic, geographic, psychographic, and behavioural. Other variables had also been mentioned including firmographics, cognitive reasoning, purchasing approaches, situational factors, lifestyle, and personality (e.g., Aaker 1995; Bonoma and Shapiro 1983; Dickson 1993).

Whilst many of these traditional consumer factors are well-established in current practice, they have faced recurring criticism for being unrepresentative of consumer behaviour motivators (Ahlm et al. 2007). Geographic and demographic dividers, whilst interesting to know and supportive in qualitative descriptions, are not proficient to foreseeing behaviour as they are unable to capture motives or drivers of consumers'

behaviour (e.g., Lancioni and Oliva 1995; Tynan and Drayton 1987; Schultz 2002). This supports the complexity problem of segmentation driven by cognitive behaviour. A consumer could share demographic characteristics such as gender, age, or occupation but inevitably hold opposing values, motivations, and beliefs (Morgan et al. 2003). In contrast, whilst reflective of cognitive moments, behavioural and psychographic variables are explorative in its research process. Therefore, much of the rich information gleaned lacks validity by being a product of qualitative primary research (Lesser and Hughes 1986; Yankelovich and Meer 2006). Criticism of traditional segmentation methods is reinforced when conceptualised in practice. Weinstein (1993, cited in Bock and Uncles 2002:215-216) anecdotally refers to a US healthcare company who commissioned 18 segmentation studies over a 5-year period without adopting one.

The daunting breadth of segmentation possibilities and lack of shared agreement over segmenting variables left traditional unrefined methods ineffective and unoptimized for implementation in practice. Seeking to concentrate the diluted spread of segmentation methods, Bock and Uncles (2002) produced a refined list of 5 taxonomies of generic differences between consumers. This list includes preferences for product benefits, consumer interaction effects, choice barriers, bargaining power, and profitability. For this research study, recognising these differentiating taxonomies for consumer heterogeneity can support the identification of relevant segmentation variables, the methods used to form segments, and the appraisal and optimization of existing segmentations. This challenges pre-conceived notions of intuitive segmenting methods amongst traditional literature and re-thinks possible ways to consider the consumer both in analytical and applied descriptive means.

2.2 Business Analytics Application to Customer Segmentation

Business analytics bridges the gap from explorative to informative information gain by providing business solutions through quantifiable means. Defined, "business analytics is a set of techniques and processes that can be used to analyze data to improve business performance through fact-based decision-making" (Lui et al. 2018:840). Big data has transformed the business environment and harnessing this evolving knowledge source is vital to understanding the phenomena it represents (ibid). Referring to the retail sector this study has targeted, but in recognition of its transferability to any sector or industry, the data volume, variety, and velocity collected through transactional data alone is vast (Chang et al. 2014). Consequently, big data's attributes exposes traditional manual methods of segmentation as redundant when considering the richness of information available and demand for complex solutions to match the intricacies of the consumer's cognitive decisions. Thus, leveraging the requirement for functional business analytics application.

Pertinent research studies within recent analytical customer segmentation literature can be classified into two overlapping categories, differentiated by their approach to segmentation. Firstly, to focus upon customer's purchase history and the consumer's personal characteristics; referred to in this study as customer centric. Secondly, to focus upon the products within a customer's basket during a single visit to segment consumers given category linkages; referred to as basket centric.

Exploring the former, customer centric segmentation requires quantifying observed behavioural variables, predominantly derived from traditional manual identifiers. Due to the admitted volume, variety, and velocity of big data, the scope of applied analytics to calculating behaviour is broad.

RFM is regularly utilised as an effective tool to guiding behavioural consumer analysis; recognised by its commonplace amongst research studies performed in the last decade (e.g., Dogan et al. 2018; Hu and Yeh 2014; Zalaghi and Varzi 2014; Chen et al. 2012). The emphasis on RFM as a combined approach to evaluating consumer behaviour revisits Bock and Uncles proposed profitability taxonomy (Bock and Uncles 2002). The heavy half of customers can be identified by RFM, viewing profitability as a multifaceted entity past the generic measure of total spend. RFM as a customer centric segmentation can be complimented via lifetime behavioural measures, such as lifetime interactions and orders (Cui et al. 2006), optimised customer lifetime value (CLV) (Venkatesan and Kumar 2004), hidden and actual economic expenditure (Bhattacharyya 1999), and statistical transaction patterns such as average quantity or spend per visit. Other non-behavioural customer variables have also been drawn upon such as the application of psychographic variables, informed by a customer's intention to purchase and lifestyle habits (e.g., Hong and Kim 2012; Liao et al. 2011). Returning to the criticism of traditional methods' explorative data collection methods, proposed psychographic methods lacks sufficient replicability due to the reliance on survey data against the scale of data presented within modern business practice.

The varied discussion amongst literature surrounding the suitable application of RFM and additional supplementary variables demonstrates the continued evolution of customer centric segmentation and ongoing strides to understand the complex consumer. To reduce the complexity effects brought upon by customer centric segmentation, basket analysis draws alternative conclusions without immediate consideration for who the customer is, instead observing what they purchase. Basket centric segmentation, also known as association rule mining, finds similarities between products bought together in a single visit and which are most frequently bought together to suggest conclusions about what common

types of behaviour a selected consumer displays (e.g., Agrawal et al. 1993; Chen et al. 2005). Commonly observed behaviours can support evaluations of consumer's predicted lifestyle, purchasing patterns, and cross-category associations which are becoming increasingly targeted methods of segmentation in recent literature (e.g., Miguéis et al. 2012; Han et al. 2014; Park et al. 2014). Basket analysis is appropriately mentioned within this study due to the relevance within the retail and fast-moving consumer goods (FMCG) sector. Broad product and category offerings have the potential to reveal underlying, manually unobservable patterns which may not support logically derived behaviours. This can harness additional contextualised data types such as loyalty program data (Banerjee 2018), category types (Cil 2012), and geographic and time series data (Tang et al. 2008). Basket centric analysis produces conceptual insights into consumption patterns per visit, reflecting conditional behaviours which can be lost across customer purchasing history. Recent attempts have been made to combine these two orientations to customer segmentation (Griva et al. 2018) but hold undeveloped repercussions within academic literature and business practice due to its novelty and author's request for industry assessment.

Smith et al. (2016) provides a novel approach to mitigate the static coarseness traditional segmentation which builds upon latent temporal topic modelling approach by implementing a Non-negative tensor factorisation (NTF). This draws upon conceptually similar motives to this study, casting dimensionality reduction to represent latent behaviours within transactional data. Grounds for further exploratory research into the field of non-static consumption patterns, observed via latent temporal trends to reflect purchasing trends, is established by Smith et al. (2016). Despite its novel value, due to its narrow scope concerning temporal purchasing patterns within topic modelling, further research into other

consumer stability measures is required to firmly introduce consumer stability into the customer segmentation discussion.

2.3 Dimensionality Reduction via PCA and Correlation Analysis

To satisfy the first objective of this study, utilising consumer stability as a feature when performing customer segmentation, PCA and correlation analysis have been selected as respective modes of dimensionality reduction to evaluate this selection's degree of effectiveness. Dimensionality reduction seeks to project data in fewer dimensions, serving to support data visualisation, model performance by tackling the curse of dimensionality, and finally feature engineering and importance evaluations (Singh 2020). Highlighting the latter due to its relevance within this study, Jolliffe and Cadima (2016) review dimensionality reduction comprehensively; presenting PCA as a function that provides the best variation-preserving two-dimensional plot of the data. Simplified, PCA identifies features that account for the largest amount of total variation within a data set, signalling effective information gain. PCA achieves this by finding the eigenvectors of the covariance matrix between all the features, with the first principal component capturing the highest variance, second principal component capturing the second highest variance, and so on. PCA has been used consistently amongst literature when performing feature importance (e.g., Liu et al. 2019; Boutsidis et al. 2008), reflecting its reliability as a feature importance evaluation technique.

To supplement the feature importance evaluation conducted by PCA, Ibrahim et al. (2021) draw upon correlation analysis to extend their feature selection choices alongside PCA. Correlation analysis considers the strength of association among the input features, seeking to identify and remove features with high correlation. Correlation is considered

detrimental because a high correlation could contribute to a skewed modelling output by over-emphasising a statistical relationship if more than one input feature display the same relationship, referred to as dependency. Therefore, as well as dimensionality reduction removing redundant features, correlation analysis is a requirement to avoid engineering duplicated, and thus dependant input features which would skew feature performance within PCA.

Jolliffe and Cadima (2016) raise the importance of standardisation within their PCA review due to the reliance of the criterion, variance, being on a distance unit of measurement. This presents an undesirable effect as a feature with a dominating unit of measurement will obscure a feature with a negatable unit of measurement. Obscuring certain features skews distance measures to features with the largest measurement scale, rather than degree of influence in informing the model. Standardisation is the industry standard for counteracting this skew, performing a uniformed change of scale to all input features. A standardised version of observation, x_{ij}, undertakes centring by subtracting the mean of variable j, written as, $\overline{x_j}$, and then is divided by the standard deviation, s_j, of the n observations of variable j (Jolliffe and Cadima 2016:6):

Equation 1. Standardisation scaler.

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}$$

Therefore, standardisation provides a suitable pre-processing operation to ensure PCA and correlation analysis are invariant to changes in unit measurement amongst input features; a requirement amongst data sets where disparity amongst features' unit of measurement is conceivable, like transaction data where order, sales, and domain data is all involved (Proagrica n.d.).

2.4 Cluster Modelling and k-Means Algorithm

Cluster modelling is presented as the primary data mining method conducted by recent academic studies to perform customer segmentation. Figuratively, clustering discovers inherent structure within data, divided by a given algorithm's determinant of what seems to make sense (Patel and Mehta 2011). Observations are assigned membership by exhibiting similar characteristics to others in their group, more so than characteristics of other groups (Francis 2012, cited in Gnararaj et al. 2014:60). Analytical clustering emerged from statistical methods, discussed in detail by Berger and Magliozzi (1992), providing the foundation for probabilistic statistical techniques such as Gaussian Mixture modelling (Allenby et al. 1999). Four heavily featured descriptive clustering models exist within customer segmentation literature: k-Means, hierarchical, density-based, and model-based analysis. Selection in practice is contextually conditional, "there is no clustering technique that is universally applicable in uncovering the variety of structures present in multidimensional data sets" (Jain et al. 1999:268). Whilst acknowledging the breadth of cluster modelling techniques, in recognition of the feature importance focus held by this study, this review will narrow the scope of review to k-Means only due to its selection as part of the research's methodology.

K-Means clustering is considered the industry standard for customer segmentation, following a fast and flexible partitioning method (Kanugo et al. 2002). Summarised, the k-Means algorithm, given n unlabelled observations existing in d-dimensional space, R^d, and an integer k, partitions n such that each observation has a minimised Euclidean distance from its associated k cluster (Method 1).

Method 1. Generic k-Means clustering algorithm.

1. *Decide on the number of clusters, k.*
2. *Initialize the k cluster centroids.*
3. *Assign the n data points to the nearest clusters.*
4. *Update the centroid of each cluster using the data points therein.*
5. *Repeat steps 3 and 4 until the changes in positions of centroids are zero.*

Source: *Ezenkwu et al. 2015:41*

Euclidean distance is relied upon as the only distance measure available to minimise across iterations across k-Means clustering literature, discussed broadly within studies (e.g., Chui et al. 2009; Ezenkwu et al. 2015; Huerta-Munoz et al. 2017). A special case (n =2) of the Minkowski metric, *Euclidean distance* $d_2(x_i, x_j)$ for point x_i is defined as:

Equation 2. Euclidean distance.

$$d_2(x_i, x_j) = \left(\sum_{k=1}^{d} \left(x_{i,k} - x_{j,k}\right)^2\right)^{1/2}$$

squaring the distance between centroid, x_j, and observation, x_i, to heavily penalise errors. Silhouette analysis can be used to reduce the ambiguity of selecting a k value by measuring how close each point it to neighbouring clusters. The silhouette score $s(x_i)$ for point x_i is defined as:

Equation 3. Silhouette score.

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i),\ a(x_i)\}}$$

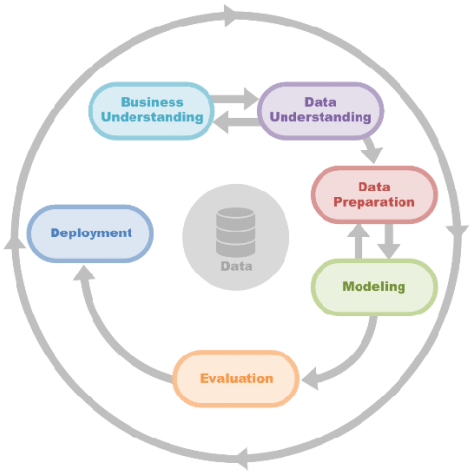
where x_i is an element in cluster π_k, $a(x_i)$ is the average distance of x_i to all other elements in the cluster π_k, and $b(x_i)$ is the minimum average distance of x_i to all other elements in cluster π_l where $l \neq k$. To inform the

value of k, the average silhouette score is taken across every observation and the highest predominantly suggests the most appropriate k value, reinforced by contextual domain knowledge (Shutaywi and Kachouie 2021).

3. Methodology and Research Design

The following research design is proposed as a suitable methodology to introduce consumer stability within the existing customer segmentation literature discussion. The business analytics process underlying this exploratory research observes the CRISP-DM theoretical framework, provided by Shearer (2000). The CRISP-DM framework invites a completeness to analytical tasks by suggesting a cyclical process, ensuring comprehensive fulfilment of the data mining life cycle (Figure 1):

Figure 1. CRISP-DM (data-mining framework)



Source: Shearer 2000, cited in Tounsi et al. 2020

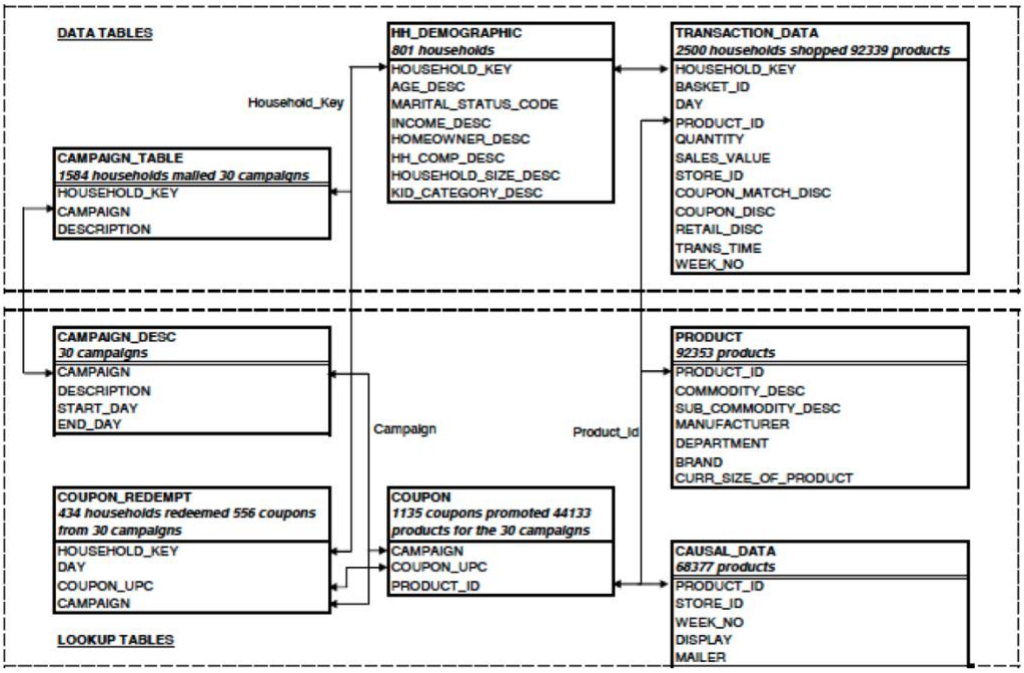
3.1 Business and Data Understanding

3.1.1 Data Collection

Dunnhumby provide the entirety of the data used within this study, specifically 'The Complete Journey' data set (Dunnhumby 2014a). This data was accessed online via Dunnhumby's available source files which offer rich real-world data sets across an array of subject contexts. Reasoning for selecting The Complete Journey source file was due to its

FMCG transactional data spine, offering comprehensive coverage of consumption behaviour within a retail context. FMCG was selected over other potential industries, such as subscription-based service data or slow-moving consumer goods (SMCG), because of its relevance to consumer stability. Relevance is justified because consumer stability has a higher likelihood of being observed within FMCG due to the fast-paced and fluid nature of consumption present and therefore offers opportunity to capture rich and observable behavioural trends, above that of other industries. The Complete Journey data set possesses 8 inter-connected tables (Figure 2).

Figure 2. 'The Complete Journey' table details.



Source: Dunnhumby 2014b:3

Extensive detail of the entire source files' data tables and contained individual features can be found via 'The Complete Journey User Guide' (Dunnhumby 2014b). This studies' scope is narrowed to the transactional data (TRANSACTION_DATA) table. A breakdown of the transaction table's features are displayed in Figure 3. This selection recognises the table's value both in terms of holding temporal information (DAY, TRANS_TIME,

and WEEK_NO), required when considering consumer stability, as well as offer relevant consumption measures such as sales totals (SALES_VALUE) and quantity (QUANTITY). Scope will be widened to consider demographic (HH_DEMOGRAPHIC Table) and product information (PRODUCT Table) to supplement initial data exploration and cluster summaries in the latter evaluation segment of this study.

Figure 3. Transaction data table description.

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
BASKET_ID	Uniquely identifies a purchase occasion
DAY	Day when transaction occurred
PRODUCT_ID	Uniquely identifies each product
QUANTITY	Number of the products purchased during the trip
SALES_VALUE	Amount of dollars retailer receives from sale
STORE_ID	Identifies unique stores
COUPON_MATCH_DISC	Discount applied due to retailer's match of manufacturer coupon
COUPON_DISC	Discount applied due to manufacturer coupon
RETAIL_DISC	Discount applied due to retailer's loyalty card program
TRANS_TIME	Time of day when the transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 - 102

Source: Dunnhumby 2014b:4

Two key attributes of Dunnhumby's data source files are ethical credibility and scale. Firstly, Dunnhumby has completed their own anonymisation of the source file, including the removal of personally identifiable information. Therefore, privacy and confidentiality is enforced providing fundamental integrity to this study. Secondly, the expansive scale of the source file offers significant opportunity for complex and well-informed analysis and evaluations to be derived; the source file holds 2,595,732 individual transactions across a 710-day period. This avoids the frequent detrimental effect of overfitting due to a small data set and allows for a greater coverage of actual behaviours displayed due to a larger sample size. An essential assumption in conjunction with the latter attribute is the sample data being representative of the population, justified within the data exploration segment of the methodology.

3.1.2 Data Exploration

Data exploration allows for deeper underlying understanding of the available data set and provides a fundamental base for navigation and analysis. In broad overview, The Complete Journey source file includes 2,595,732 individual transactions across a 710-day period from 2,500 unique households. As an FMCG retailor, the product department variety is vast, ranging from groceries to travel and leisure to toys. To satisfy the fore-mentioned assumption of the data set being representative of a 'normal' population, demographic information is evaluated against its coverage of all category types of demographic information (Table 1). This breadth of demographic inclusion, coupled with the substantial scale of the source file and multi-year period, satisfies the assumption that the data is representative of a 'normal' population. This assumption enables the replicability of the study methodology to inspire further academic research and ensures the applicability of conclusions to a broader audience within industry.

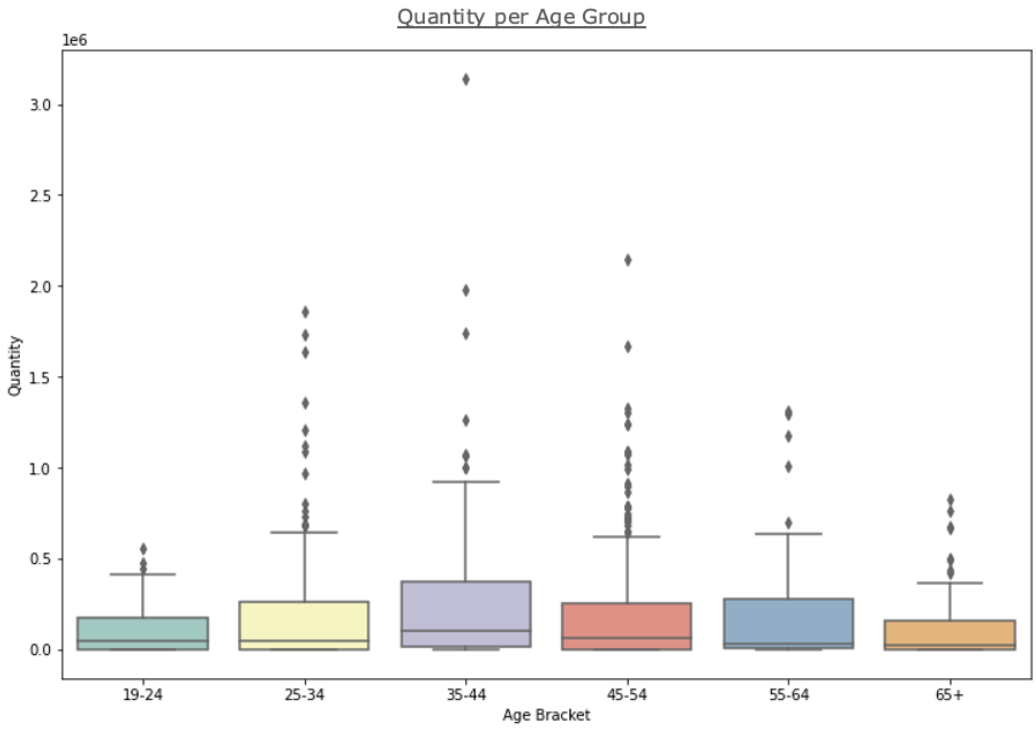
Table 1. Unique demographic feature values.

Demographic Feature.	**Unique values.**
Age	'19-24','25-34','35-44','45-54','55-64','65+'
Salary	'Under 15K','15-24K','25-34K','50-74K','75-99K','100-124K','125-149K','150-174K','175-199K','200-249K','250K+'
Homeowner Status	'Homeowner','Unknown','Renter','Probable Renter', 'Probable Owner'
Household Composition	'Single Female','Single Male','1 Adult Kids','2 Adults No Kids','2 Adults Kids'
Household size	'1','2','3','4','5+','None/Unknown'

Having attributed the source file to possess representative depth, further exploration of the source file provides insight into the nature of consumer behaviour involved with the profiled retailor. Firstly, a varied purchase

quantity emerges across age groups. Visualised in Figure 4, age bracket 35-44 display high quantity purchasing, whereas the younger and older groups both appear to purchase less.

Figure 4. Quantity by age.



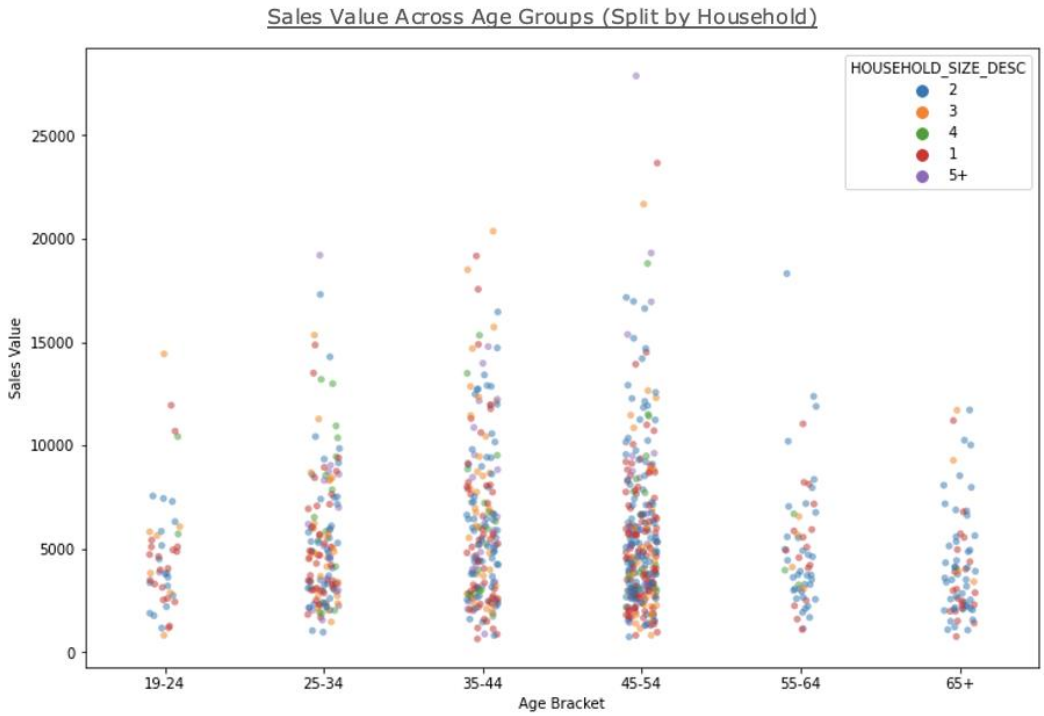
Evaluating quantities likely correlation to spend value, substituting quantity with sales value displays a similar result when used to compare consumption by age. Again, the 35-44 age bracket holds the highest purchasing involvement, reflected by holding the highest average sales value of £6,402.82. Spend descends adjacently to the 35-44 bracket as age increases/decreases respectively (Table 2).

Table 2. Spending per age bracket.

Age bracket.	Sales Value.
19-24	£4,704.32
25-34	£5,478.42
35-44	£6,402.82
45-54	£5,755.09
55-64	£5,054.73
65+	£4,241.03

Exploring consumption relative to demographic factors further, household size does not immediately correlate to a particular spending behaviour (Figure 5). Inferred from a lack of observed pattern across the two categorical features in relation to sales value, no household size prevails as an identifiably high or low spender. What this visualisation does demonstrate is a large degree of variance across spend, often rooted by extreme outlying data.

Figure 5. Spending by age - categorised by household size.



Contextual considerations intuitively suggest large quantity and sales values are unexpected within FMCG transactional data set, due to the commonality of regular but small purchasing within the industry. Statistically describing quantity to explore this high variance, it can be observed a household's average quantity purchased sits within an interquartile range of 1.31 to 115.58 units (Table 3). Greatly exceeding the interquartile range, maximum average purchase per household is 6393.96 units (Table 3). We can therefore perceive the 'average' consumer to purchase within a rounded IQR quantity range of 1 to 116 units, recognising an abnormality of purchasing behaviour outside this range.

Table 3. Quantity per household.

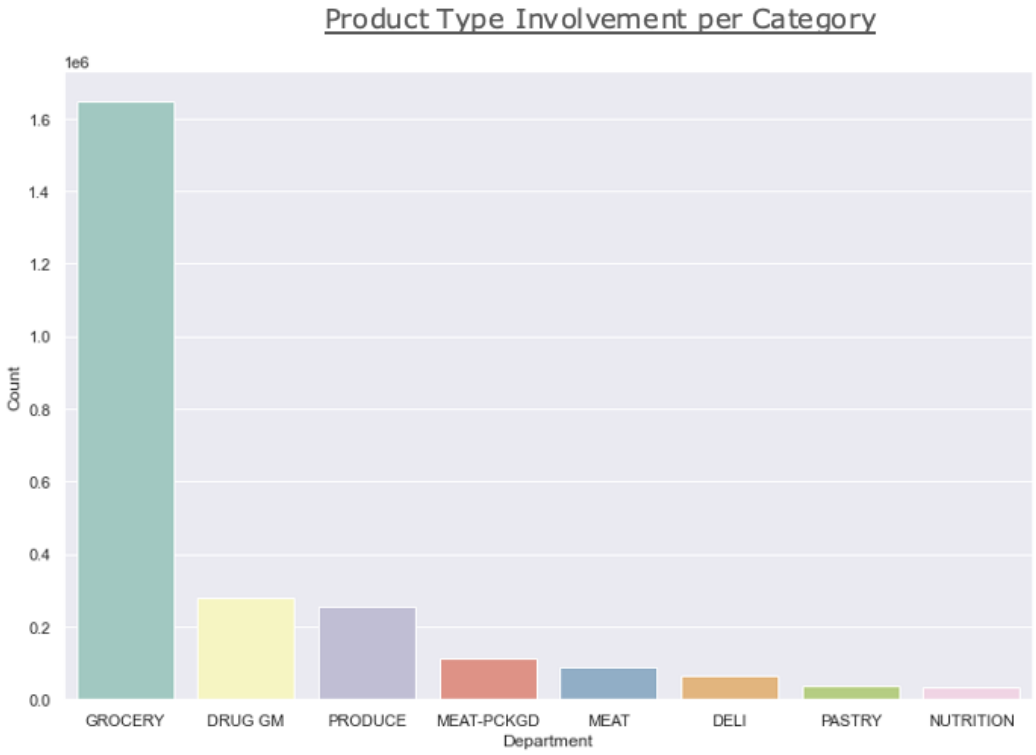
	Quantity per household.
Count	2,500
Mean	92.08
Std	242.58
Min	0.88
25%	1.31
50%	14.95
75%	115.58
Max	6,393.96

Exploring any imperfections of the data which may disturb the application of data analytics processes, the data source does not possess missing data but does possess negative values. Identifying either characteristic could hold implications for data preparation by requiring pre-processing adjustments. Fictitiously, missing data may be handled via imputation or deletion techniques; well discussed by Kang (2013). Nonetheless, due to its lack of existence within this data missing data will not be referenced within data preparation stages. Negative values often exist within raw transactional data where refunds, discounts, and returns are being issued. Dissecting the presence of negative values through domain awareness, our transaction data includes applied discounts (RETAIL_DISC,

COUPON_DISC, and COUPON_MATCH_DISC). Due to their appropriateness within the data table and potential for engineering features, such as original retail price and marketing campaigns' effectiveness, the existence of negative values does not present an immediate issue to this study.

Lastly, by exploring the most popular products being purchased it can be observed grocery shopping dominates the category involvement of customers (Figure 6). Awareness of this heavily weighted preference to grocery shopping will support qualitative and statistical conclusions within the results of the cluster modelling. This is because segments displaying behaviour against this preference could provoke unique insights about differing product involvement.

Figure 6. Involvement per product department.



3.2 Data Preparation

Having explored the data available, enabling a provisional business and data understanding, data preparation is required to ensure the raw data is clean, noise-free, and consistent. These characteristics are principally challenged by outlying, missing, or skewed data. Preparation within this methodology is split between data cleansing and feature engineering. The former cleans the raw data being used by handling any imperfections in the data. The latter establishes the premise for this research by generating features to reflect consumer stability.

3.2.1 Data Cleansing

As identified within section 3.2.2, whilst no missing or negative values within the data required address, outlying values were recognised significantly outside the IQR of quantity purchases per household. Statistically, a mean quantity of 92.08 units was obtained despite this being in the far upper quartile of average household quantity (Table 3). Outlying data can negatively influence data analytics tasks by disguising abnormal results and evaluations. Further, machine learning algorithms are sensitive to range and distribution attribute values. Metaphorically, data outliers are considered noise which spoil and mislead the algorithmic training process by interfering with information gain from informative values. Addressing this, a refined transaction list is generated which excludes any transaction above the quantity of 208 units, removing 23,131 transactions. This value is elected by adding one standard deviation to the 75% value and treating this as a ceiling measure for normal consumption. In doing so, a statistical summary of consumption when excluding outlying quantity values demonstrates a household's averaged 1.32 units per transaction (Table 4), accounting for an average of 1342.69 units per household across the 710-day period (Table 5).

Table 4. Average quantity per household per transaction.

	Average quantity per household per transaction.
Count	2,500
Mean	1.32
Std	0.22
Min	0.88
25%	1.19
50%	1.27
75%	1.39
Max	4.46

Table 5. Average quantity per household across period.

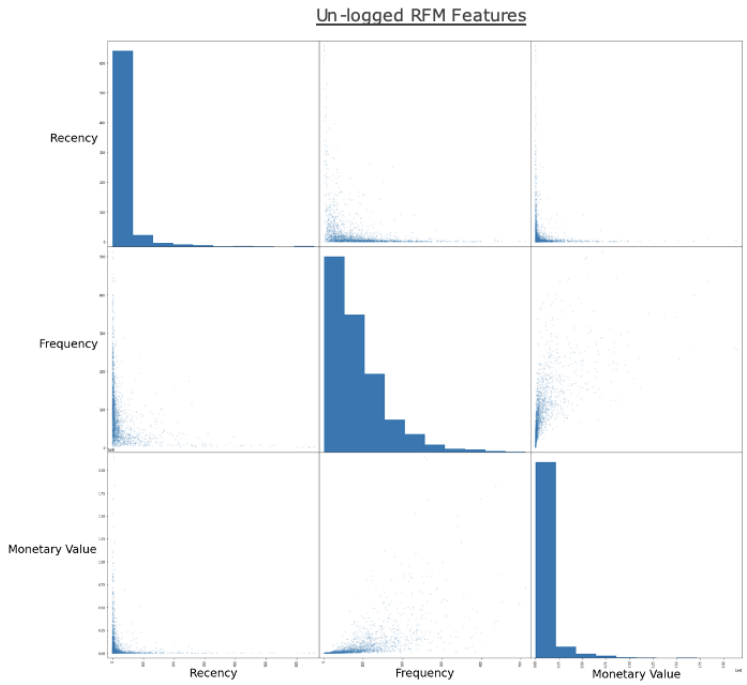
	Average quantity per household.
Count	2,500
Mean	1,342.69
Std	1,298.56
Min	5.0
25%	414.75
50%	956.0
75%	1,861.5
Max	11,216.0

Having cleansed the data set of outlying values, data preparation also includes a scaling transformation to remove the effects of skewed, non-linear relationships so that featured data approximately conforms to normality. Reiterating the significance Jolliffe and Cadima (2016) places on standardisation as a scaling transformation, a dominating unit of measurement skews distance measures to features with the largest measurement scale, rather than degree of influence in informing the model. Whilst any analytical process is vulnerable to the effects of skew due to their learning sensitivity, this proposed research is especially vulnerable due to the reliance of k-Means and PCA on distance measures, further discussed within section 3.4. Due to this reliance, it is appropriate

to note scaling's significance to ensure integrity and optimisation of this methodology by preventing the damaging repercussions of skew.

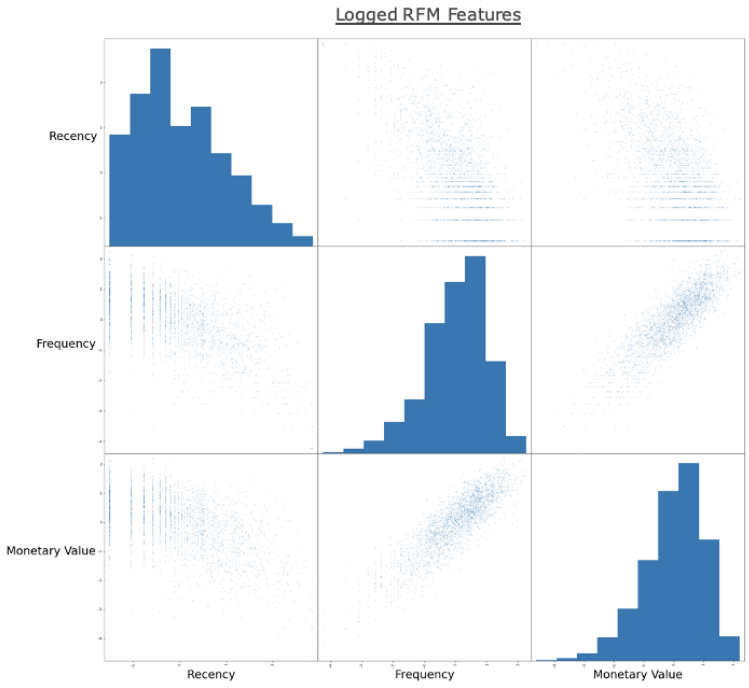
Standardisation is the industry standard for counteracting this skew, performing a uniformed change of scale to all input features, often complimented by a log transformation. This pairing is motivated by the fact standardisation normalises a data set's distribution to a centre where $\mu = 0$ and $\sigma^2 = 1$. However, the distribution shape remains unchanged. By contrast, a log transformation removes skew within the data itself. Whilst not a compulsory data preparation step, when exploring the data through a dummy static RFM analysis we can observe a heavily right-tailed skew within the data (Figure 7). A static RFM analysis is justified as a representative dummy procedure due to the same underlying feature inclusion and engineering methodology proposed within section 3.3.2. Therefore, demand to perform a log transformation to remove this skew before performing standardisation is considered necessary.

Figure 7. Un-logged RFM features.



Applying a log transformation to a data point (x) returns an associated value (y) where $y = \log_e x$. This transformation is statistically reversible, $x = e^y$. Displayed in figure 8, the log transformation pulls the data into a perceived normal distribution, desired due to the reduction in skew.

Figure 8. Logged RFM features.



Following a log transformation, standardisation scales raw data by converting points into a confined range, performing a uniformed change of scale to all input features (Equation 1). The effect of standardisation on the previously logged static RFM measure can be viewed in Table 6 and 7, illustrating the first 5 households static RFM measures before and after standardisation is applied. A standardisation function was defined which calculated the mean and standardisation from the feature the lagged counterpart relates to. This is because temporal data is encoded within lagged variables by both their absolute values and their relative values. Therefore, by standardizing them collectively ensures the information encoded in their relative values is preserved, rather than standardising each column independently. In conjunction, standardisation and log

transformations provide a suitable data preparation operation to ensure analysis is invariant to changes in unit measurement amongst input features, a key attribute in the reliability and accuracy of analysis results and research conclusions.

Table 6. Static RFM features pre-standardisation.

Household Key.	Recency.	Frequency.	Monetary Value.
1	1.60943	4.36844	15.97276
2	3.76120	3.80666	14.30404
3	2.07944	3.80666	15.24807
4	4.43082	3.40119	13.03559
5	2.07944	3.49650	12.15935

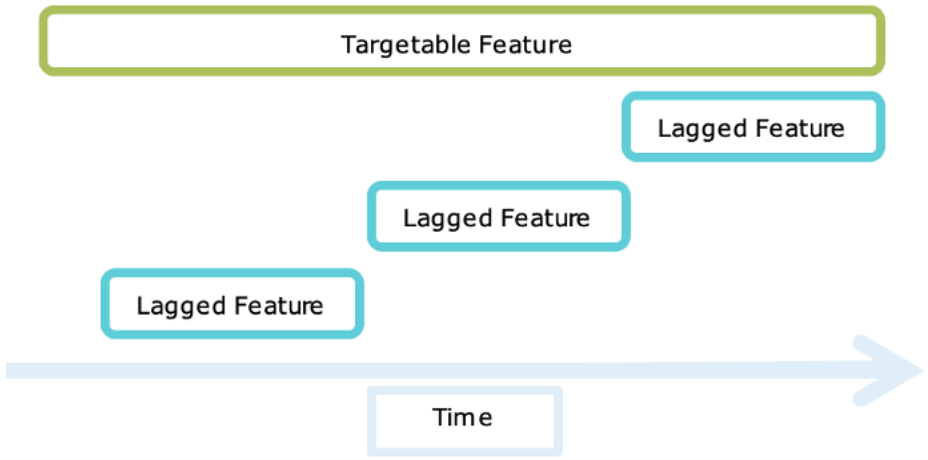
Table 7. Static RFM features post-standardisation.

Household Key.	Recency.	Frequency.	Monetary Value.
1	-0.42928	0.30755	0.79016
2	1.02307	-0.27846	0.05509
3	-0.11205	0.27846	0.47094
4	1.47505	-0.70068	-0.50365
5	-0.11205	-0.60141	-0.88964

3.2.2 Feature Engineering

To introduce consumer stability as a customer segmentation feature, lagged RFM features are generated which seek to uncover a customer's latent temporal behaviour across weekly windows. Lagged features are essentially a past tense snapshot of a targetable feature (Figure 9). A 7-day weekly window was selected as an appropriate tumbling window size due to the fast-paced nature of FMCG customer behaviour and recognition of the high volume of transactions occurring within each weekly period. Whilst narrow, an even smaller daily window size was rejected because of FMCG's consumption being susceptibility to inconsistent behaviour and so daily trends may vary too much, disguising any observable conclusions.

Figure 9. Lagged feature illustration.



Lagged RFM features are generated following accepted practice of RFM feature generation, with the additional condition of occurring within the relevant window space. Briefly defining each lagged feature:

- Monetary value is defined as total financial spend by a household within the given window.
- Recency is defined as presence of purchase by a household within the given window.
- Frequency is defined as the number of purchases by a household within the given window.

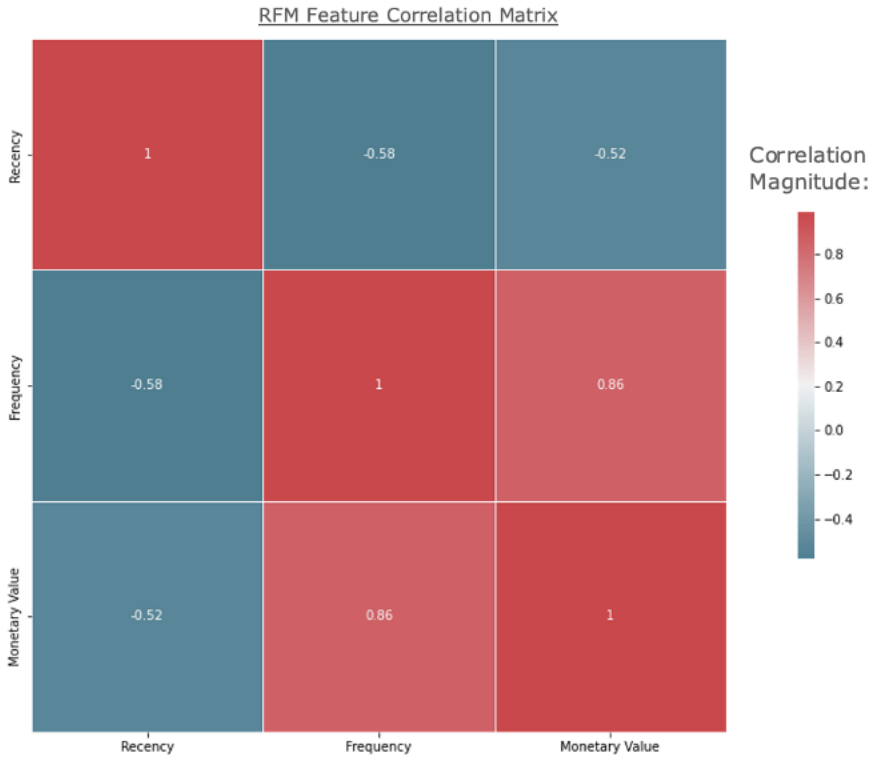
The respective static RFM feature is also included within the engineered table to support the evaluation of consumer stability as an informing customer segmentation feature. This is appropriate as a point of comparison, utilised within the subsequent PCA.

3.3 Principal Component Analysis

Having applied the fore-mentioned log and standardisation transformations to the engineered lagged feature inclusive data set, correlation analysis is implemented to evaluate the inter-dependency

amongst the generated features, defined as multicollinearity. Correlation is a statistical measure to expresses the strength of relationship between two features. Correlation analysis is an appropriate step prior to PCA occurring due to the procedural risk of multicollinearity because admitting dependant features within a data set during analysis will overemphasise certain relationships whilst shadowing others. Figure 10 displays a correlation matrix for our dummy static RFM data to demonstrate its functionality, utilising an annotated heatmap to sufficiently visualise magnitude and direction of relationships.

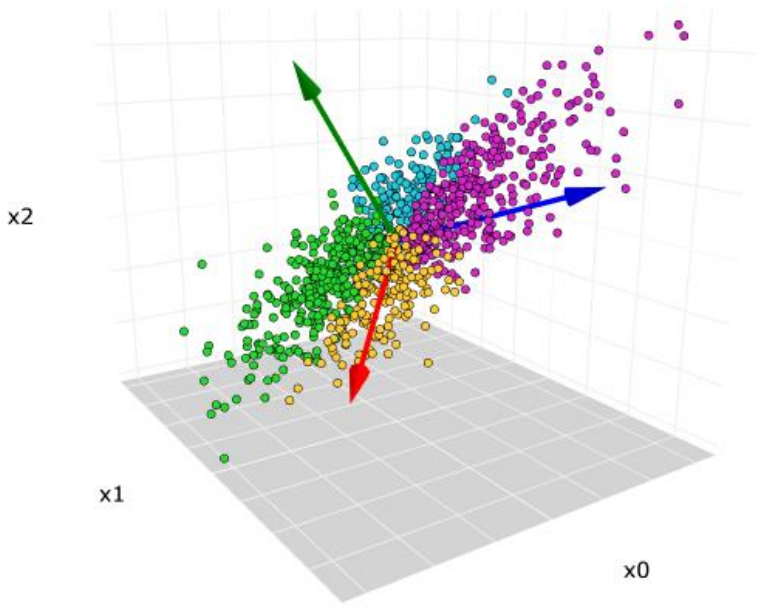
Figure 10. Static RFM correlation matrix.



After ensuring appropriate feature inclusion measures are taken following correlation analysis, PCA can be applied to evaluate the feature importance of our engineered variables. PCA provides the best variation-preserving two-dimensional plot of the data, achieved by creating new linear combinations of original axes (Figure 11). PCA is advantageous because this two-dimensional plot identifies features that account for the

largest amount of total variation within a data set, signalling effective information gain. To derive a measure of feature importance, an incremental measure of captured variance will determine the number of principal components, or axes, to carry forwards. PCA tends to disregard dimensions after 70% of variance is captured. Having captured sufficient variance, PCA provides a statistical list of which variables contribute most to these components, an accepted indication of feature importance.

Figure 11. Principal component analysis visualisation.



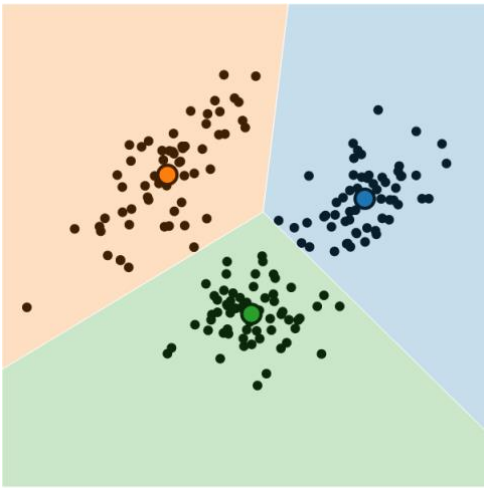
Source: Cheng (2022)

3.4 k-Means Cluster Model

A cluster modelling approach is applied to satisfy the second objective of discovering common consumer behaviour pathways. Due to the possible deterrent of diluting this research's objectives and prevailing popularity of k-Means as a clustering model for customer segmentation, justified in section 2.4, a k-Means approach is the sole clustering model chosen to satisfy this second objective. K-Means is an unsupervised learning

approach which represents sub-groups of values by centroids (Figure 12). The motivation for clustering is due to the conclusions and inferences that can be made from centroid positioning within the axes. This opportunity is enabled when centroid values are reverse transformed after being logged and standardised to exist within the original axes space.

Figure 12. k-Means centroid visualisation.

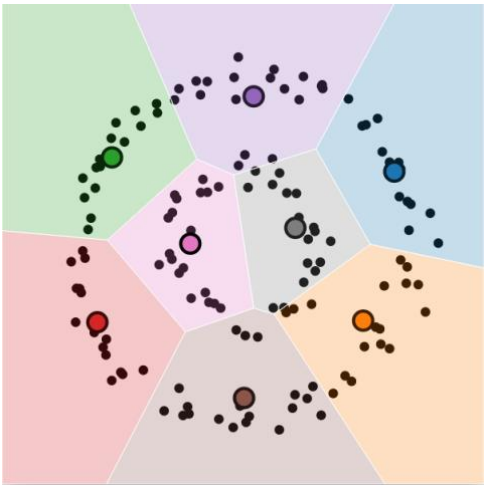


Source: Broyelle (2021)

Deconstructing the generic k-Means methodology for our purposes (Method 1), to first decide the number of clusters a silhouette score is calculated for a range of k values (Equation 3). A silhouette score is utilised as a metric to measure the effectiveness of a clustering model by calculating the average distance of values associated within a given centroid to neighbouring clusters. This technique combines both the notion of cohesion within clusters and separation between others. Average distance refers to the Euclidean distance metric (Equation 2). To inform the value of k, the average silhouette score is taken across every observation and the highest predominantly suggests the most appropriate k value, reinforced by contextual domain knowledge and brief requirement caveats. An implicit assumption using k-Means is clusters are isotropic, meaning clusters are uniformed in size and roughly spherical. Vulnerability is therefore recognised when the distribution of data cannot

form convex partitions, seen in figure 13. Violation of this assumption can be identified by illustrating clusters in a scatter plot, a step included in the customer segmentation analysis of section 4.2.

Figure 13. Non-convex concentric ring.



Source: Broyelle (2021)

As stated, this exploratory research is designed to introduce consumer stability as an informing customer segmentation feature. Therefore, meta-parameter tuning is side-lined despite recognition of its importance within analytics. To still offer sufficient coverage to the construction of the k-Means model used, it is important to note this model utilised a mean aggregator and no initialisation occurred. Common reasoning for applying both are to mitigate outliers within the data. Justification for neglecting outliers is placed because of prior data preparation steps to remove outlying data and create normalising tendencies through a log transformation and standardisation. These steps sufficiently protect the model's design against the sensitive nature of distance-based models to noisy data.

3.5 Evaluation

3.5.1 Feature Importance Evaluation

Evaluating the result of PCA will provide the basis for satisfying the first research objective question, seeking to discover if consumer stability is an informing feature when set against traditional methods such as customer lifetime or single-visit basket analysis. It will achieve this by observing attained principal components and identifying which features are most informative within these components. An explained variance metric will endorse the number of principal components considered sufficient by evaluating the extent original features contribute to components generated. To guide feature inclusion of PCA and further conclusions, the magnitude and direction of correlations are statistically evaluated. A key indicator of success in this section of the evaluation is recognising if engineered lagged features are present and informative within the identified principal components.

3.5.2 Customer Segmentation Analysis

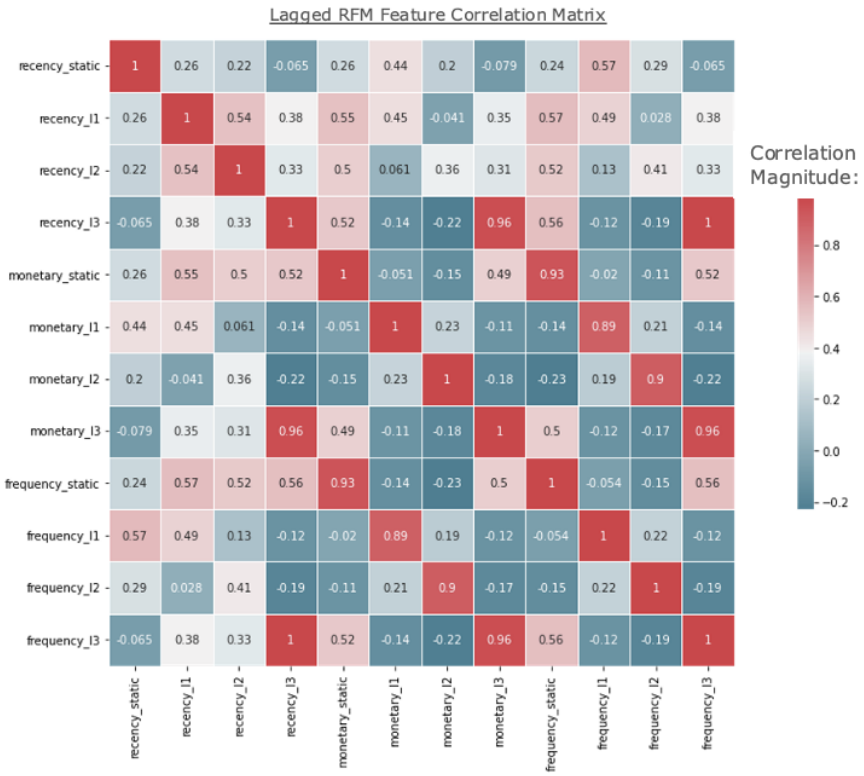
Evaluating the results of the cluster modelling provides opportunity to discover non-static consumer behaviour pathways from shared patterns in consumption. This will be achieved quantifiably by statistically analysing centroid trends produced from the k-means model. Attention within this step will tend towards considering how lagged features interact with each other within formed segments.

4. Findings and Results

4.1 Feature Importance Evaluation

Figure 14 displays the correlation matrix for our engineered variables, utilising an annotated heatmap to sufficiently visualise magnitude and direction of relationships. Immediately, recognising correlation visually with a dark red shade and high magnitude figure, frequency and monetary variables can be observed to display clear correlation. Inferred by each relative feature all possessing a correlation magnitude above a threshold of 0.89. The statistical significance of this magnitude is it exhibits a highly dominant correlation, in effect exerting the same information within the data. As discussed, multicollinearity provides high risk as this correlation will overemphasise this relationship, negatively influencing PCA and clustering results.

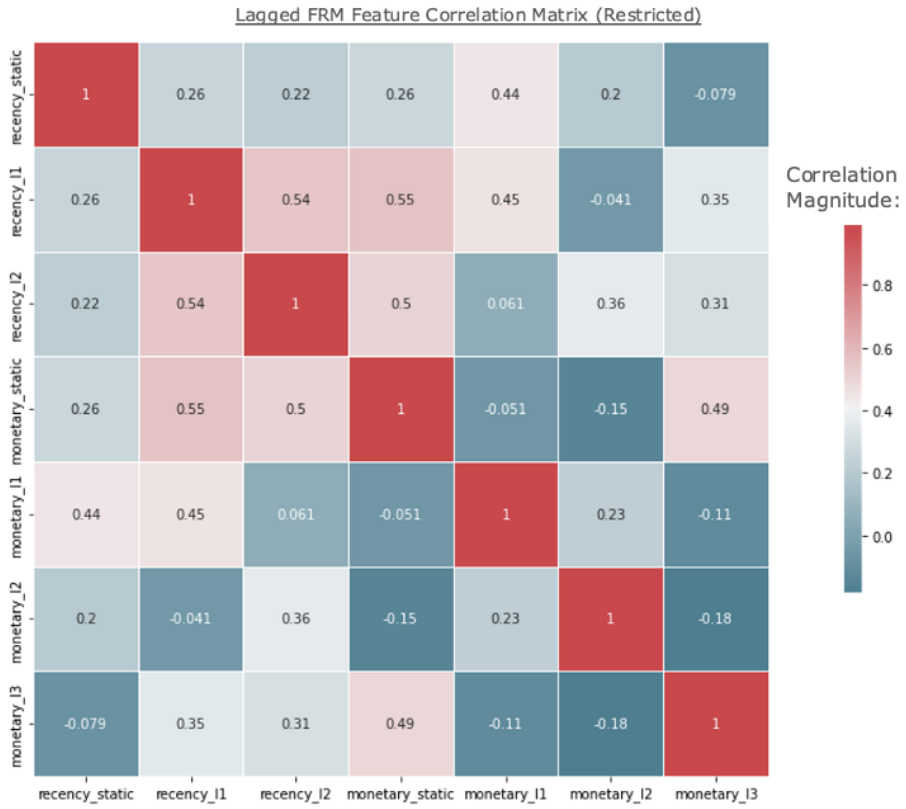
Figure 14. Un-Restricted lagged feature correlation matrix.



When dissecting this first identified relationship between relative frequency and monetary features, it is perhaps logical that frequencies involvement in the respective monetary calculation will lead to a likelihood of correlation. Whilst not identifiable within a static RFM calculation due to the consistent use of both metrics separately within academia, when taken in much smaller window sizes it is possible this inter-dependency becomes more apparent than a singular lifetime measure. Based on this first identified relationship, the corrective measure was taken to remove all frequency variables. Without an informed understanding of this correlation and the variables mathematical relationship, the decision could seem drastic. Nevertheless, appreciating the risk multicollinearity poses, this is deemed an appropriate decision.

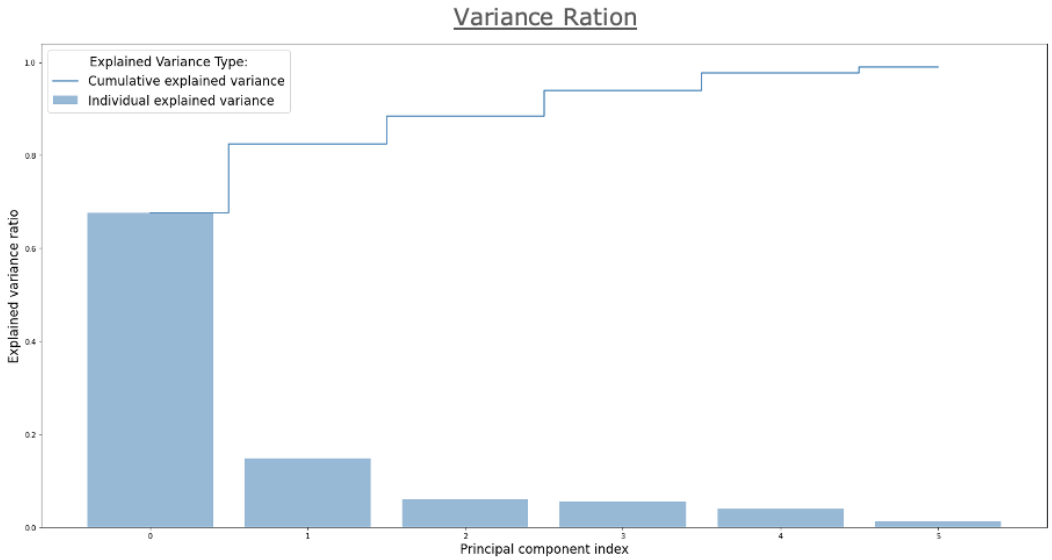
A second correlated relationship within the feature space is the third lagged version of the relative recency, frequency, and monetary feature. This poses an issue because despite dropping the third lagged frequency feature, a significant correlation still exists between the third recency and monetary feature. To avoid admitting a correlation of magnitude of 0.96, the decision to drop the third lagged recency feature was also taken, observable in figure 15.

Figure 15. Restricted lagged feature correlation matrix.



Having appropriately dropped variables displaying multicollinearity, PCA is implemented as a tool to evaluate feature importance of our generated lagged RFM measures. PCA achieves this by producing principal components which reduce the dimensional space features exist on. Significance of a feature's presence within attained components indicates reliance of those features for information gain and therefore a signal of importance. An explained variance metric endorses the number of principal components considered by producing a ratio of each component's contribution to overall variance. The ratio of explained variance is calculated by dividing the eigenvalue by total eigenvalues. Industry standard advises a cumulative explained ratio of 70% is considered enough principal components to be representative of a data set's complete variance. Applying this to our data, it can be observed 2 principal components are required; recognising a cumulative variance of 0.7 is achieved after the second component step (figure 16).

Figure 16. Explained variance ratio.



Accepting 2 components sufficiently encapsulate variance, evaluating the extent original features contribute to these two components grant an indication of which are most informative. The importance of each feature is reflected by the magnitude of the corresponding values in the eigenvectors; a higher magnitude indicating a higher importance. Seen in figure 17, the third lagged monetary feature demonstrates an overwhelming contribution to the first principal component; holding a 0.97 weighting (Table 8). Figuratively, this weighting suggests the value of a household's monetary spend in the third lagged window is a key influencer in the position of that household on the principal axes. In contrast, a broader spread of features contribute to the second principal component's variance with the first recency feature providing the greatest weight. This suggests an incorporation of multiple features contribute to the determinant of the second component without singular reliance on one affecting a values position on the axes.

Figure 17. Feature component contribution.

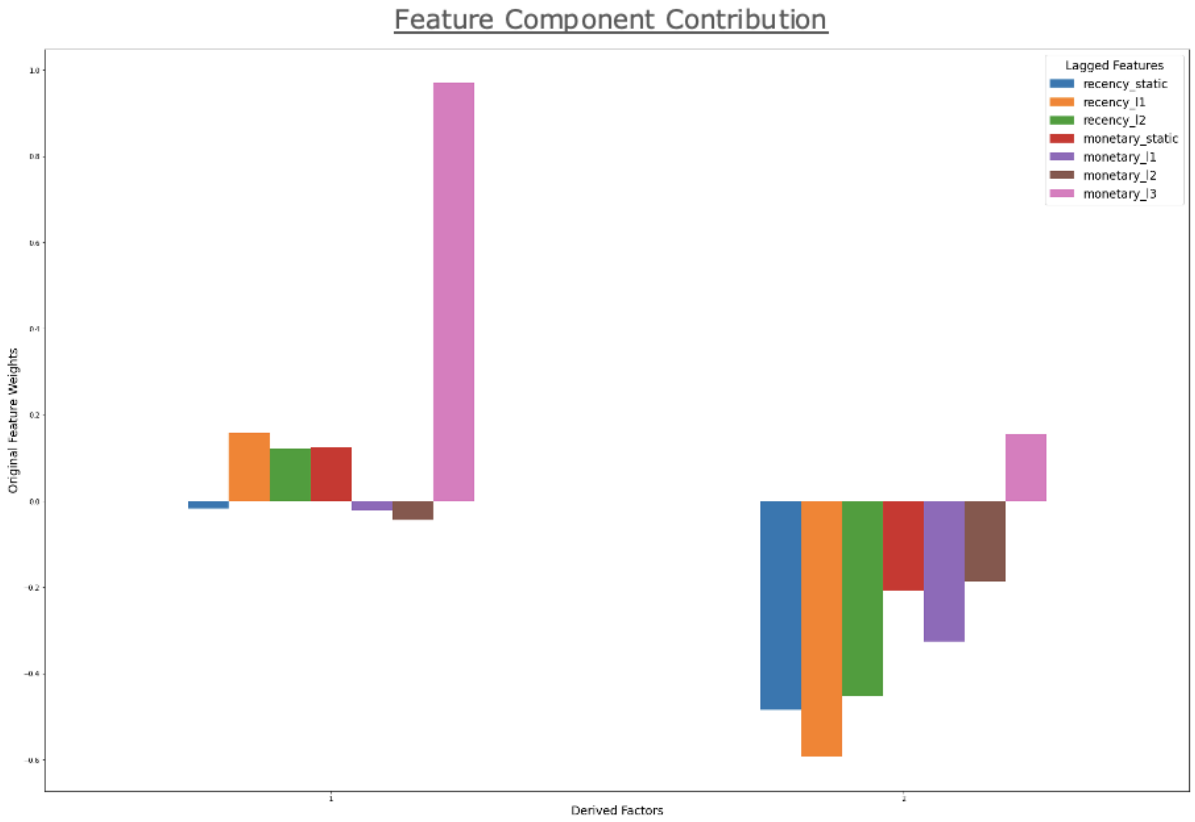


Table 8. Principal component contribution per feature.

Feature.	Principal Component 1.	Principal Component 2.
'recency_static'	0.02	0.48
'recency_l1'	0.16	0.59
'recency_l2'	0.12	0.45
'monetary_static'	0.12	0.21
'monetary_l1'	0.02	0.33
'monetary_l2'	0.04	0.19
'monetary_l3'	0.97	0.16

4.2 Cluster Modelling

Firstly, a k value of 2 is selected as the most appropriate number of clusters the points should be segmented upon. Evaluating the silhouette scores collected when passing a range of k values (Table 9) alongside insights from feature contribution to principal components (Figure 17), it can be observed this selection is intuitively justified because of the third lagged monetary feature's influencing dominance over the first principal

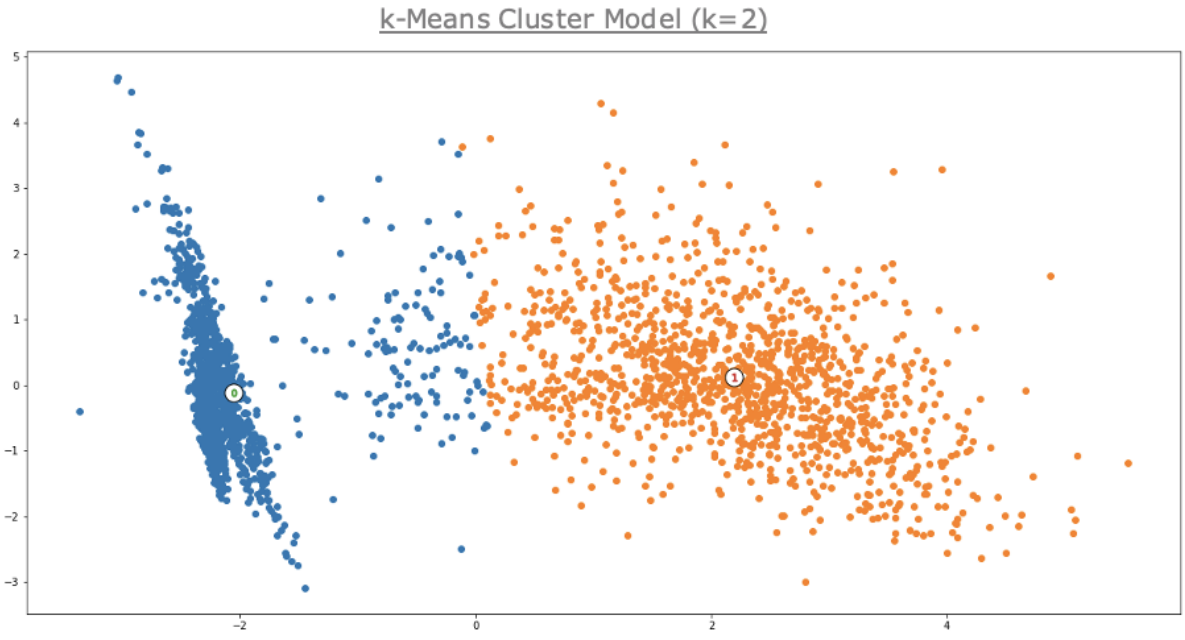
component. Therefore, based on a k value of 2, it is highly likely a household's monetary value in the third lagged window will be divisive in segmenting households.

Table 9. Silhouette score - cluster range 2-8.

Cluster Size.	Silhouette Score.
2	0.64062
3	0.52378
4	0.46081
5	0.45047
7	0.42480
8	0.41762

Implementing a k-Means cluster model with a k value of 2, a clustered model of the data in 2 dimensions can be viewed in Figure 18. Immediately, a geometrically non-spherical segment shape can be observed. This anisotropic cluster shape violates of the assumption k-Means conducts convex partitions on isotropic data. Results and findings are still presented from this model to satisfy the second objective of this dissertation, withholding underlying repercussions of this violation until further dissected in section 5.2, Limitations.

Figure 18. k-Means cluster model.



Quantitatively evaluating this cluster model, the following segment archetypes can be derived which display the statistical summary of the cluster centre of each segment (Table 10). Archetypes have been inverse transformed to return values relevant to original axes dimensions, accounting for the log and standardisation transformations.

Table 10. Segment archetypes.

Feature.	Segment 0.	Segment 1.
recency_static	0.826156	0.687834
recency_l1	0.895779	1.537505
recency_l2	0.927585	1.410354
monetary_static	1.914671	3.107979
monetary_l1	0.686598	0.579604
monetary_l2	0.729664	0.580236
monetary_l3	1.182871	75.726218

The first statistically significant takeaway when evaluating cluster centre archetypes is a higher recency and recency figure in the first and second lagged window displayed by Segment 1 in comparison to Segment 0. This contradicts the traditional RFM feature, recency, which suggests Segment 0 have consumed more recently. Exposing this static coarseness by using temporal windows, Segment 1 households consistently consumed in proximity to weekly windows more recently. Secondly, Segment 1 displays an extreme monetary centre value of 75.726 in the third lagged window which drastically surpasses any relative monetary value across both segments. This can be classified as irregular when compared to relative values on their respective axes. Irregularity is reinforced when considering household's mean average quantity and sales value within each segment respectively because a lack of differentiation between segment's consumption highlights the unexpectedness of Segment 1's third lagged monetary centre value (Table 11).

Table 11. Cluster sales and quantity averages.

Segment.	Quantity.	Sales Value.
0	1.246118,	2.922842
1	1.264026	2.717483

Evaluating the pathway Segment 0 follows in reference to changes in relative feature groups across windows, it can be conceived Segment 0 are a suboptimal customer group. A decrease from the second to first recency to recency value can be observed, indicating Segment 0 are purchasing less in proximity to the window than the previous. Further, the same segment's cluster centre is positioned such that monetary spend decreases between each window. Both temporal patterns are indicative of a downturn in consumption behaviour, considered suboptimal relative to business performance due to FMCG priorities often transfixed on maximising spend and transaction quantity. Evaluating Segment 1 under the same conditions, it can be conceived Segment 1 is an optimal customer group. Identified primarily due to the cluster's increasing recency values, these household's sit significantly higher in terms of volume of purchases completed in recency to the window, with Segment 1 having a recency value in the first window of 1.537505 compared to Segment 0's respective value being just 0.895779. However, monetary spend has little to no change moving from 0.580236 to 0.579604. This demonstrates a lack of spend increase aligned to increased recency across the same period. Discussion of both pathways and their implications for marketing activity are explored in the subsequent section 6, Discussion and limitations.

5. Discussion and Limitations

As a prerequisite to the ensuing section, the discussion and limitations for this research have been combined to offer a succinct and interconnected rationale for academic discussion and practical recommendations. In doing so, limitations associated to particular points of discussion can be mentioned coherently, encapsulating their weighted significance in reference to one another.

Invoking the findings and results of the PCA and cluster modelling, this section will return to the objectives of this exploratory research to provide grounds for discussion in academia and industry. Firstly, determining if consumer stability is an informative customer segmentation feature when performing cluster modelling, it can be observed via PCA results that consumer stability is an informing feature. Reached because both principal components relying on the temporally representative features to contribute to their variance coverage (Figure 17). Reliance can be recognised because each principal component's leading feature contributor is a lagged feature. Therefore, consumer stability can be judged to hold informative value when segmenting customers, due to its competitiveness against static RFM features as well as the dominance of the recency feature in the third lagged window within the first principal component.

Considering the implications of this evaluation within academia, findings reinforce the latent temporal behavioural work of Smith et al. (2016) by substantiating the introduction of consumer stability within the customer segmentation academic discussion. This is because temporally representative lagged RFM features exceeded the information gain that relative lifetime RFM features provided. Further, finding sufficient information gain via consumer stability justifies the same author's

criticism of the static coarseness of traditional segmentation methods. Lifetime RFM can be considered a suboptimal feature in informing segmentation under the recognised lesser contribution static RFM features provided to principal components in direct comparison to relative lagged features. Profiling this criticism indicates the required shift towards the further research and application of consumer stability as a customer segmentation feature within academia. For this exploratory research to achieve completeness in satisfying the first objective of establishing consumer stability as an informing customer segmentation feature set against traditional methods, a similar study would be advised against single-visit basket analysis methods. Raised within section 2.2, Business Analytics Application to Customer Segmentation, basket centric segmentation finds similarities between products bought together in a single visit and which are most frequently bought together to suggest conclusions about what common types of behaviour a selected consumer displays. Fulfilling this, consumer stability could be cemented into the analytical customer segmentation discussion having effectively challenged the two most referenced traditional segmenting methods.

Alongside academic implications, the practical value of satisfying the first objective is stressed when considering the optimised consumer-orientated marketing decisions it can facilitate. Endorsing consumer stability, temporal behaviour can be employed to guide business activity as a substitute of, or additional rationale for, lifetime customer metrics. Overall, this research allows retailors, primarily FMCG, to view customer's behaviour from an alternative point of view, surpassing lifetime or single-visit snapshots to instead consider behaviour as fluctuating and non-static. Dissected in direct reference to this research's findings, to view consumption in shortened periodic windows in proximity to present allows for richer information gain reflective of actual behaviours, mitigating the aforementioned criticism of traditional customer and basket centric

approaches. Specifically, demonstrating the feature importance of lagged RFM measures could allow for refined rolling approach to segmentation in which marketing activity prioritises consumer stability as the persuading indicator of a customer's behaviour. For instance, a marketer could recognise the significance of the first lagged feature of each relative measure as a key indicator of consumer stability. For instance, a household in Segment 1 with an increased recency value in the first window in comparison to the second could be considered highly desirable because their consumption habits in reference to their stability are optimal, purchasing more recently than previously. Intuitively this presents a logical reasoning, suggesting a customer purchasing more than the previous week is likely to be targetable for further consumption. However, due to the resounding acceptance of static RFM measures currently in place this offers statistical evidence for this reasoning to break with traditional practices in industry.

Considering the larger landscape in which customer segmentation exists, its purpose for marketers is two-fold, enforcing optimal consumption and rectifying suboptimal consumption. Alongside direct marketing activity instanced previously, motivated by targeting optimal consumers reflective of their stability, satisfaction of the first objective also suggests wider indirect implications for advertisers to discover bridging products that induce consumption to supplement management of suboptimal pathways. Identifying consumer stability allows the shifting mentality of marketers to recognise customers as temporally manoeuvrable and therefore requiring attention, modified not just on preferences or habits but also in consideration of their stability state.

Although this proposed research methodology presents positive academic and practical implications within the broad customer segmentation field, a key limitation underpinning the presented findings and discussion is

exploratory naivety. 3 rolling windows was conceived as an appropriate number to introduce lagged RFM measures whilst ensuring computational efficiency and interpretability. However, the overriding influence of the third lagged monetary feature as a divisive axes disorientated the attained PCA and cluster model's results. Firstly, when considering this effect of a small window sample on PCA, correlation analysis removed a large number of engineered features. Whilst it was appropriately justified within this study, a wider array of lagged features may have revealed varying population correlations and justified a decision to keep certain dropped features, such as all relative frequency features. By reducing the dimensions in which PCA occurred, results were constrained to a small number of axes and therefore each component was susceptible to a dominating feature, in this case monetary in the third lagged window. Had a wider array been presented, variance coverage may have been apportioned allowing for more features to contribute to the principal components, representative of a generalised population behaviour.

Whilst the proposed methodology has not yet been thoroughly optimised or evaluated against other clustering methods, it does support the extraction of temporal-orientated clusters that reflect differing consumer stability behaviours. Satisfying the second research objective, observing if common or dominant pathways exist, and are these optimal to business performance, cluster centre archetypes presented by a k-Means clustering model offer two contrasting stability pathways. Further academic research will support these proposed pathways, in particular reference to the cross-examination of how pathways fare when a larger volume of rolling windows are constructed. With transparent discern, Segment 1's third lagged monetary value is drastically above the relative feature's norm. Therefore, continued investigation and exploration of how this value sits when a greater number of lagged features per relative measure is of particular interest for further academic research. Accepting this irregularity, we can still consider this exploratory research as a second

step, following Smith et al. (2016), towards the introduction of consumer stability within customer segmentation literature. Moreover, alongside the observable benefit of deriving latent temporal behaviours, this research can be considered uniquely valuable due to the deployment of stability measures against static lifetime RFM in a shared environment. Therefore, recognising the contribution of engineered stability measures within PCA and high involvement to cluster creation strongly indicates the opportunity for valuable information gain as an exciting avenue amongst saturated, and potentially suboptimal traditional discussion.

A second major limitation of this methodology was the reliance on a k-Means clustering model. Apparent within figure 18, the data displays an anisotropic cluster shape. This violates k-Means' assumption that data being partitioned is isotropic, misplacing some points due to the assumption of globular clusters. This could lead to suboptimal cluster centres being derived due to noise stretching clusters into an inappropriately formed spherical shape. To combat this drawback, evaluation of multiple models would be an essential next step in the continuation of introducing consumer stability as a customer segmentation feature. Particular attention would be placed on ensuring inclusion of model's adept to manage classes that form non-spherical shapes such as a density-based model e.g., DBSCAN. Recognising this drawback limits the breadth of coverage this research can definitively provide to the satisfaction of the second research objective. Whilst this research has suggested stability pathways that exist, and their optimality to business performance, to this end, these findings lack technical integrity due to the suboptimal management k-Means possesses when handling anisotropic classes.

6. Conclusion

By means of significance, the most remarkable outcome of this exploratory research is the reliance of derived principal components on introduced consumer stability features. Gained via evaluation of PCA results, these findings provide evidence to support the proposed hypothesis of this dissertation, introducing consumer stability as an informing customer segmentation feature. This conclusion is justified by recognising the dominating presence of the third lagged monetary value in the first principal component, and outweighing contribution of all lagged features within the second (Figure 17). Consequently, the most significant implication of this research is the potential inferiority of traditional segmentation methods to temporally reflective techniques. Evolving customer and basket centric approaches, in this case through temporally modifying RFM analysis, consideration of latent temporal consumption and a customer's ability to fluidly adapt their own purchasing position has been demonstrated as an area of targetable focus for consumer behaviour academia and marketing practice.

7. References

Aaker DA (1995) *Strategic Market Management*, Wiley, New York

Agrawal R, Imieliński T, and Swami A (1993) 'Mining association rules between sets of items in large databases', *ACM SIGMOD International conference on management of data*, Vol. 22(2):207-216

Ahlm S, Holmstrom M, and Stenman V (2007) *Traditional market segmentation – an evaluating approach*, School of Economics and management, Lund University, Available at: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=1339435&fileOId=2434865, accessed 27th June 2022

Allenby GM, Leone RP, and Jen L (1996) 'A Dynamic Model of Purchase Timing with Application to Direct Marketing', *Journal of the American Statistical Association*, Vol. 94(446):365-374

Banerjee K (2018) 'Enhancing Customer Loyalty using Market Basket Analysis', *Aspire Systems*, Available at: https://blog.aspiresys.com/digital/big-data-analytics/enhancing-customer-loyalty-using-market-basket-analysis/, accessed 1st July 2022

Beane TP and Ennis DM (1987) 'Market segmentation: A review', *European Journal of Marketing*, Vol. 21(5):20-42

Berger P and Magliozzi T (1992) 'The effect of sample size and proportion of buyers in the sample on the performance of list segmentation equations generated by regression analysis', *Journal of Direct Marketing*, Vol. 6(1):13–22

Bhattacharyya DK (1999) 'On the Economic Rationale of Estimating the Hidden Economy', *The Economic Journal*, Vol. 109(456):348-359

Bock T and Uncles M (2002) 'A taxonomy of differences between consumers for market segmentation', *International Journal of Research in Marketing*, Vol. 19(3):215-224

Bonoma TV and Shapiro BP (1983) *Segmenting the industrial market*, Lexington Books, Lexington

Boutsidis C, Mahoney MW, and Drineas P (2008) 'Unsupervised Feature Selection for Principal Component Analysis', *KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas, Nevada, 24th-27th August:61-69

Broyelle A (2021) 'Deep (Deep Deep) Dive into K-Means Clustering', *Github*, Available at: https://antoinebrl.github.io/blog/kmeans/, accessed 15/08/2022

Chang RM, Kauffman RJ, and Kwon Y (2014) 'Understanding the paradigm shift to computational social science in the presence of big data', *Decision Support Systems*, Vol. 63(1):67-80

Chen YL, Tang K, Shen RJ, and Hu YH (2005) 'Market basket analysis in a multiple store environment', *Decision Support Systems*, Vol. 40(2):339-354

Chen YS, Cheng CH, Lai CJ, Hsu CY, and Syu HJ (2012) 'Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment', *Computers in Biology and Medicine*, Vol. 42(2):213-221

Cheng C (2022) 'Principal Component Analysis (PCA) Explained Visually with Zero Math', *Towards Data Science*, Available at: https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d, accessed 13th August 2022

Chui CY, Chen YF, Kuo IT, and Ku HC (2009) 'An intelligent market segmentation system using k-means and particle swarm optimization', *Expert Systems with Applications*, Vol. 36(3):4558-4565

Cil I (2012) 'Consumption universes based supermarket layout through association rule mining and multidimensional scaling', *Expert Systems with Applications*, Vol. 39(10):8611-8625

Cross L (1999) 'Segmentation: When Less Is More', Graphic Arts Monthly, Vol. 71(2):96-106

Cui G, Wong M, and Lui HK (2006) 'Machine learning for direct marketing response models: Bayesian networks with evolutionary programming', *Management Science*, Vol. 52(4):597-612

Dickson PR (1993) *Marketing management*, The Dryden Press, Orlando

Dogan O, Ayin E, and Bulut ZA (2018) 'Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry', *International Journal of Contemporary Economics and Administrative Sciences*, Vol. 8(1):1-19

Dunnhumby (2014a) 'The Complete Journey', Dunnhumby Source Files, Available at: https://www.dunnhumby.com/source-files/, accessed 1st July 2022

Dunnhumby (2014b) 'The Complete Journey User Guide', Dunnhumby Source files, Available at: https://www.dunnhumby.com/source-files/, accessed 1st July 2022

Engel JF, Fiorillo HF, and Cayley MA (1972) *Market Segmentation Concepts and Applications*, Holt, Rinehart, and Winston, New York

Ezenkwu CP, Ozuomba S, and Kalu C (2015) 'Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services', *International Journal of Advanced Research in Artificial Intelligence*, Vol. 4(10):40-44

Gnararaj TN, Kumar KR, and Monica N (2014) 'Survey on mining clusters using new k-mean algorithm from structured and unstructured data', *International Journal of Advances in Computer Science and Technology*, Vol. 3(2):60-65

Griva A, Bardaki C, Pramatari K, and Papakiriakopoulos D (2018) 'Retail Business Analytics: Customer Visit Segmentation using Market Basket Data', *Expert Systems with Applications*, Vol. 100(1):1-16

Han S, Ye Y, Fu X, and Chen Z (2014) 'Category role aided market segmentation approach to convenience store chain category management', *Decision Support Systems*, Vol. 57(1):296-308

Hu YH and Yeh TW (2014) 'Discovering valuable frequent patterns based on RFM analysis without customer identification information', *Knowledge-Based Systems*, Vol. 61(3):76-88

Huerto-Munoz DL, Rios-Mercado RZ, and Ruiz R (2017) 'An iterated greedy heuristic for a market segmentation problem with multiple attributes', *European Journal of Operational Research*, Vol. 261(1):75-87

Hunt SD and Arnett DB (2004) 'Market Segmentation Strategy, Competitive Advantage, and Public Policy: Grounding Segmentation Strategy in Resource-Advantage Theory', *Australasian Marketing Journal*, Vol. 12(1):7-25

Ibrahim S, Nazir S, and Velastin SA (2021) 'Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis', *Journal of Imaging,* Vol. 7(225):1-16

Jensen FV (1996) *An Introduction to Bayesian Networks*, Springer, New York

Jolliffe IT and Cadima J (2016) 'Principal Component Analysis: A Review and Recent Developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* Vol. 374(2065):1-16

Kang H (2013) 'The prevention and handling of the missing data', *Korean Journal of Anaesthesiology,* Vol. 64(5): 402-406

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, and Wu AY (2002) 'An Efficient k-Means Clustering Algorithm: Analysis and

Implementation', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24(7):881-892

Kelly WT (1973) New Consumerism: Selected Readings, Grid, Inc., Columbus, OH:87

Kotler P (1997) *Marketing Management: Analysis, Planning, Implementation, and Control*, (9th ed.), Prentice Hall International, Upper Saddle River

—— , Wong V, Saunders J, and Armstrong G (2005) *Principles of Marketing*, (4th ed.), Pearson, Edinburgh Gate:5

Lancioni R and Oliva TA (1995) 'Penetrating purchaser personality', *Marketing Management*, Vol. 3(4):22-29

Lesser JA and Hughes MA (1986) 'The generalizability of Psychographic Market Segments across Geographic Locations', *Journal of Marketing*, Vol. 50(1):18-27

Liao S, Chen Y, and Hsieh H (2011) 'Mining customer knowledge for direct selling and marketing', *Expert Systems with Applications*, Vol. 38(5):6059-6069

Lui Y, Han H, and DeBello JE (2018) 'The Challenges of Business Analytics: Successes and Failures', *51st Hawaii International Conference on System Sciences Proceedings*, Manoa, 3rd January:840

Lui Y, Singleton A, and Arribas-Bel D (2019) 'A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification', *Geo-spatial Information Science*, Vol. 22(4):251-264

Miguéis VL, Camanho AS, Falcão E, and Cunha J (2012) 'Customer data mining for lifestyle segmentation', *Expert Systems with Applications*, Vol. 39(10):9359-9366

Morgan CM, Levy DJ, and Fortin M (2003) 'Psychographic Segmentation', Communication World, Vol. 20(1):22-26

Park CH, Park YH, and Schweidel DA (2014) 'A multi-category customer base analysis', *International Journal of Research in Marketing*, Vol. 31(3):266-279

Patel VR and Mehta RG (2011) 'Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm', *IJCSI International Journal of Computer Science Issues*, Vol. 5(2):331-336

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA

Proagrica (n.d.) 'Benefits of data standardization in the agricultural supply chain', *Proagrica: News and Insights,* Available at: https://proagrica.com/news/benefits-of-data-standardization-in-the-agricultural-supply-chain/, accessed 24th July 2022

Schultz DE (2002) 'Behavior change; do your segments?', *Marketing News*, 22nd July:5-6

Smith WR (1956) 'Product differentiation and market segmentation as alternative marketing strategies', *Journal of Marketing*, Vol. 21(1):3-8

Shearer C (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing*, Vol. 5(4):13–22

Shutaywi M and Kachouie NN (2021) 'Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering', *Entropy*, Vol. 23(6):1-17

Singh A (2020) 'Importance of Dimensionality Reduction', *Analytics Vidhya*, Date Printed: 14/06/2020, Available at: https://medium.com/analytics-vidhya/importance-of-dimensionality-reduction-d6a4c7289b92, accessed 24th July 2022

Smith G, Goulding J, and Smith A (2016) *Automatic Temporal Retail Segmentation From big Data*, Business School, University of Nottingham, Available at:

http://www.cs.nott.ac.uk/~pszgss/AMA2016preprint.pdf, accessed 19th July 2022

Hong T and Kim E (2012) 'Segmenting customers in online stores based on factors that affect the customer's intention to purchase', *Expert Systems with Applications,* Vol. 39(2):2127-2131

Tang K, Chen YL, and Hu HW (2008) 'Context-based market basket analysis in a multiple-store environment', *Decision Support Systems*, Vol. 45(1):150-163

Tounsi Y, Anoun H, and Hassouni L (2020) 'CSMAS: Improving Multi-Agent Credit Scoring System by Integrating Big Data and the new generation of Gradient Boosting Algorithms', *The 3rd International Conference on Networking, Information Systems & Security,* Association for Computer Machinery, New York, accessed 2nd August 2022

Twedt DW (1964) 'How Important to Marketing Strategy is the Heavy User?', *The Journal of Marketing*, Vol. 28(1):71-72

Tynan AC and Drayton J (1987) 'Market segmentation, *Journal of Marketing Management*, Vol. 2(3):301-335

Venkatesan R and Kumar VA (2004) 'Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy', *Journal of Marketing*, Vol. 68(4):106-125

Vyncke P (2002) 'Lifestyle Segmentation: From Attitudes, Interests and Opinions, to Values, Aesthetic Styles, Life Visions and Media Preferences', *European Journal of Communication*, Vol. 17(4):445-463

Weinstein A (1993) 'Market Selection in Technology-Based Industry: Insights from Executives', *American Marketing Association Winter Educators' Conference Proceedings*, Newport Beach, CA, 20th-23rd February:1-2

Weinstein A (2006) 'A strategic framework for defining and segmenting markets', *Journal of Strategic Marketing*, Vol. 12(2):115-127

Wind Y (1978) 'Issues and Advances in Segmentation Research', *Journal of Marketing Research*, Vol. 15(3):317-337

Yankelovich D and Meer D (2006) 'Rediscover market segmentation', *Harvard Business Review*, Vol. 84(2):122-134

Zalaghi Z and Varzi Y (2014) 'Measuring customer loyalty using an extended RFM and clustering technique', *Management Science Letters*, Vol. 4(5):905-912