UNIVERSITY OF CAPE TOWN

STA5004W

ADVANCED ANALYTICS MINOR DISSERTATION

Personal Finance:

A Statistical Analysis of the Habits and Behaviours of the South African Consumer

---

Author:
Ehsaan Rajak

Student Number:
RJKEHS001

Supervisors:
Prof. Francesca Little
Dr. Allan Clark

Acknowledgements

I would like to express my deepest gratitude to my parents for their unwavering support, boundless encouragement, and relentless determination throughout my academic journey. Their belief in me has been a constant source of motivation, guiding me through the challenges and triumphs of this dissertation. My mother, Khatija Rajak, has been the greatest source of inspiration throughout my life. She has always shown me that it's possible to keep improving yourself and its never too late to keep learning and pushing yourself to be better. My father, Afzal Rajak, has been the unwavering rock of our family and I know that none of what we have achieved would be possible without his sacrifices.

I extend my heartfelt thanks to my supervisors, Prof. Francesca Little and Dr. Allan Clark, for their invaluable guidance, unwavering patience, and insightful feedback. Their expertise, encouragement, and mentorship have been instrumental in shaping this research and my growth as a scholar. Without their ability to simply continue the conversation whenever I had not made progress for an extended period of time, this dissertation would not be where it is right now.

A special mention goes to the Department of Statistical Sciences at the University of Cape Town, which has been my academic home for the past few years. The faculty and staff have created an environment conducive to learning and exploration, providing me with resources and support that have been integral to the completion of this dissertation. It truly has been a second home in the toughest of times.

I am also grateful to 22seven who provided the data for this dissertation and allowed me the freedom to explore it in the best way I saw fit for this dissertation. The feedback and discussion I had with the 22seven team was an invaluable part of the refinement of this dissertation.

Thank you all for your contributions, guidance, and unwavering support throughout this academic pursuit.

Abstract

Understanding consumer spending behavior and the efficacy of budget setting is crucial in managing personal finances. This dissertation employs clustering methodologies and statistical models to explore the intricate dynamics of financial habits and their relationship with budget establishment using longitudinal spending data from the 22seven platform. The initial chapters delve into the analysis of consumer spending behavior through various clustering techniques, unveiling fundamental drivers of spending patterns. While highlighting the role of spending control in distinguishing consumer clusters, the study emphasizes its correlation with wealth creation potential.

Subsequently, the investigation into the impact of budget setting on expenditure habits reveals compelling evidence. Individuals who set a budget exhibited a significant reduction in spending, indicating an average decrease of approximately 38% compared to non-budget setters. However, limitations in observational data analysis, including incomplete financial account linkage and potential sample bias, caution against drawing absolute conclusions. This dissertation underscores the complexity of consumer financial decision-making, calling for continued exploration and refinement of methodologies to better grasp the nuanced interplay between budget setting and expenditure patterns. While providing valuable insights, this study serves as a stepping stone for future research, encouraging a deeper understanding of effective spending control, the balance between consumption and savings, and the broader efficacy of budgeting in managing overspend.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the realm of personal finance advice, the maxim is simple: spend less than you earn and save the rest. While a relatively simple concept, it is one that many people, particularly in South Africa, struggle with. Bengtsson (2012) found that in South Africa "marginal propensity to save is close to zero or even negative" implying that South Africans are not particularly good at saving. When provided with extra income, they choose to spend rather than save. More recently Mutual (2023) reported that 34% of working South Africans had taken out a personal loan in 2023 and 54% had to dip into savings in order to make ends meet, further highlighting the dire situation of the South African consumer.

Such spending behaviors have attracted substantial attention within the realm of behavioral economics and consumer finance. Theories such as the Theory of Planned Behavior (TPB) and Behavioral Economics have offered frameworks to comprehend the complexities of consumer spending habits. The TPB posits that individual intentions are determined by attitudes toward the behavior, subjective norms, and perceived behavioral control. In the context of spending, this theory suggests that an individual's attitude towards saving versus spending, social influences, and their perceived ability to control their spending significantly impact their financial choices (Ajzen, 1991).

Furthermore, Behavioral Economics, spearheaded by researchers like Kahneman and Tversky, explores how psychological factors influence economic decisions. Prospect Theory, a cornerstone of Behavioral Economics, highlights that individuals often make financial choices based on the potential for gains or losses rather than the final outcome. This theory illuminates why individuals may be inclined to spend more when experiencing windfalls or bonuses, as the emotional response to potential gains might overshadow the rationality of saving for the future (Kahneman and Tversky, 2013).

Additionally, concepts such as hyperbolic discounting have garnered attention in understanding impulsive spending behaviors. Hyperbolic discounting refers to the tendency for individuals to prefer immediate rewards over larger delayed rewards, even if the latter would be more beneficial in the long term. This phenomenon plays a significant role in consumer spending habits, as individuals may prioritize immediate gratification through spending, disregarding the long-term consequences of financial decisions (Ainslie, 1975).

The interplay of these theoretical frameworks emphasizes that consumer spending habits are not solely determined by rational economic calculations but are significantly influenced by psychological, social, and cognitive factors. Understanding these theories provides a foundation for comprehending the intricacies behind the spending patterns observed, particularly within the South African context where challenges in savings and increased borrowing rates prevail, contributing to the broader discourse on consumer financial behavior.

The goal of this dissertation is to shine a quantitative lens on the understanding of consumer spending behaviour, particularly in the South African context. To do so, a unique dataset was obtained from the personal financial management app, 22seven. With the aim of understanding their users better, 22seven agreed to embark on a collaborative journey with the author of this dissertation. As part of the agreement, 22seven identified a subset of their data which they were willing to share.[1] In exchange for the statistical analysis and findings in the data, the author of this dissertation was permitted to use said findings for the purposes of this dissertation. After collaborating with stakeholders at the business, two main research questions were identified which the author aims to answer here:

1. **Do spending patterns over time reveal distinct groups?**
2. **Does setting a budget work in reducing spend?**

1.1 Aims and Objectives

In addition to the core research questions, the directives provided by 22seven allowed for a flexible exploration of these questions and the definition of subsidiary objectives, as well as an exploration of the methodology for tackling these queries.

Research Question 1: Do spending patterns over time reveal distinct groups?

To explore this inquiry, a clustering approach has been embraced, as elaborated in Chapter 4. Various methods for clustering longitudinal data have been applied and the resulting user groups have been compared with the goal of identifying distinct clusters that can be easily characterized.

[1] More details on the structure of these data will be shared in Chapter 3.

Research Question 2: Does setting a budget work in reducing spend?

The primary focus centers on addressing this pivotal inquiry, as detailed in Chapter 5. This question is further divided into two related sub-questions:

1. Is a user's spend significantly reduced after the act of setting a budget on the app?
2. Compared to users who do not set a budget, how much do users that do set a budget save?

To address both of these sub-questions, a mixed modeling framework has been employed which shall be detailed in Chapter 5.

Chapter 2

Literature Review

In the pursuit of understanding the diverse landscape of consumer behavior, clustering methodologies have emerged as pivotal tools to uncover patterns and segment consumers based on their spending habits, preferences, and financial behaviors. These methodologies encompass a range of techniques aimed at grouping individuals with similar characteristics or behaviors, offering valuable insights into the heterogeneity within consumer populations.

One of the most commonly used methods for clustering subjects is the k-Means algorithm (Gupta and Chandra, 2019). It has emerged as a particularly popular approach due to its relative simplicity and speed of execution. It has been applied in a variety of fields for applications such as Customer Segmentation (Ezenkwu et al., 2015), Medical Image Segmentation (Ng et al., 2006) and helping to provide evidence for economic theories (Wielechowski et al., 2021). Due to its success in a wide variety of fields, it is often used as a first step in many clustering applications. However, owing to its non-parametric nature, there exists a competing field of clustering based in statistical theory and often grouped together under the term Latent Class Models (Magidson and Vermunt, 2002).

Unlike K-Means, Latent Class Models (LCMs) assume that the observed data is a result of an underlying categorical variable that divides the population into distinct classes. These classes or clusters are characterized by unique patterns of responses or behaviors, allowing for a deeper understanding of the heterogeneity within a population (Clogg, 1995). While a version of k-Means exists to deal with longitudinal data, (Genolini and Falissard, 2010), the advantage of LCMs is that the statistical theory allows for explicit modeling of the time component in the data. These advantages (increased interpretability due to theory and flexible modeling of time effects) have made LCMs increasingly popular in medical applications (Liu et al., 2021; Formann and Kohlmann, 1996).

The applicability of these methods and the parallels to medical data are clearly seen. For each user of the app, rather than measuring some medical quantity of interest, consumer financial metrics are measurable. However, due to the lack of availability of these data, it appears as though there exists a gap in the literature. The author of this dissertation aims to fill this gap by performing clustering using a variety of approaches and comparing the results. The result should allow for evaluating the effectiveness of these methods in understanding heterogeneity among consumer spending habits and gaining greater insight into factors influencing spending habits.

Furthermore, one of the tools used to help consumers reign in spending is that of the budget for which there also exists some theory. Budgeting is a fundamental tool in personal finance management, aiming to allocate income towards specific expenses, savings, and investments within a predefined framework. Behavioral economics and psychological theories contribute to understanding how budgeting influences spending behaviors.

The Mental Accounting Theory proposed by Richard Thaler explores how individuals mentally separate their finances into different categories or "accounts," assigning different purposes to each. This theory suggests that people tend to spend differently based on the mental categorization of money, leading to diverse spending behaviors. For instance, individuals might be more inclined to overspend using a credit card (often considered as "not real money" at the point of purchase) rather than using cash, which has a more immediate impact on their mental account for available funds (Thaler, 1985).

Moreover, research has shown that the effectiveness of budgeting systems is closely tied to behavioral aspects. The Implementation Intention Theory suggests that forming specific plans or intentions can significantly impact goal attainment. When individuals set precise and actionable goals within their budget—such as allocating a certain percentage of income towards savings or restricting spending on non-essential items—it increases the likelihood of adhering to the budget (Gollwitzer, 1999).

Behavioral economics also sheds light on the concept of nudges in budgeting. Nudges are subtle interventions that steer individuals towards making better decisions without limiting their freedom of choice. In the context of personal finance, nudges can be implemented through various means, such as reminders, visual cues, or default settings, to encourage individuals to stick to their budgets (Thaler and Sunstein, 2009).

However, despite the theoretical foundations and strategies suggested by behavioral economics and psychological theories, the implementation and success of budgeting practices vary among individuals. Factors like self-control, impulsivity, socioeconomic status, and cultural influences also play pivotal roles in determining the effectiveness of budgeting tools. Cultural norms, for instance, might influence the perceived importance of saving or spending within a community, impacting individual budgeting habits.

One of the factors influencing budgeting compliance is the ease with which expenses are recorded. The theory of consumer budgeting as proposed by Heath and Soll (1996) posits that under this

scenario, budget compliance will be relatively high. This is further corroborated by both lab and field studies (Heath and Soll, 1996; Stilley et al., 2010).

However, research on planning fallacies has shown that individuals tend to be overoptimistic when planning (Buehler et al., 2010). This implies that spending will be higher than budget, because budgets are a plan for future spending (Novemsky and Kahneman, 2005; Ülkümen et al., 2008). Similarly, investigations into predicting consumer budgets have revealed that individuals tend not to factor in unforeseen circumstances when estimating their expenses (Sussman and Alter, 2012). Consequently, it is likely that spending will surpass the budgeted amount due to plans being based on forecasts, which fail to encompass all potential expenses. These discoveries propose the notion that consumers will likely exceed their planned budget significantly (Kahneman and Tversky, 2013).

Theories concerning mental accounting and consumer budgeting provide several grounds to suggest that budgets can impact a consumer's spending behavior to some extent. For instance, establishing a budget could serve as a self-regulation mechanism, enabling consumers to limit their expenditure by imposing strict guidelines such as "do not exceed X amount on Y". Additionally, budget setting might constrain spending as earmarked money for specific purposes could pose psychological barriers against utilizing it otherwise. Moreover, the practice of monitoring expenses could levy a psychological toll on deviating from the budget, creating further motivation to curb spending (Thaler, 1985; Heath and Soll, 1996).

To the best of the author's knowledge, these budgeting theories have not been extensively tested with large real-world data. A recent entrant, (Lukas and Howard, 2023), into the literature did have access to similar data as this dissertation. Lukas and Howard (2023) found that setting a budget did have a positive effect in reducing spend and that the effect persists over time. However, this was done in a first-world country (Great Britain) in which the economic situation is quite different from South Africa. To further bolster the literature, this dissertation performs an investigation into the effectiveness of budgeting in the South African market.

Chapter 3

Data Overview

22seven is a popular personal finance app operating primarily in South Africa. The goal of the platform is to enable South Africans to take better control of their finances and help them make better money decisions. It does so by aggregating data from multiple financial service providers across a user's accounts and by using this data, provides budgeting features as well as personal insights on users' spending habits. A simplified overview of how the platform works as extracted from their website (https://www.22seven.com) is shown in Figure 3.1.

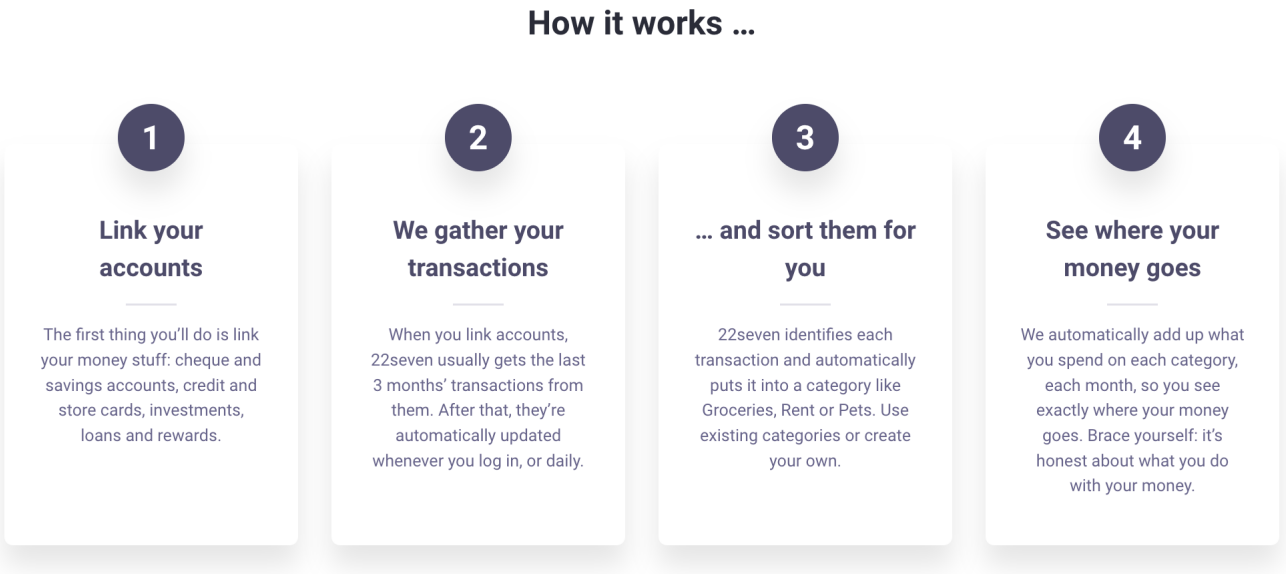
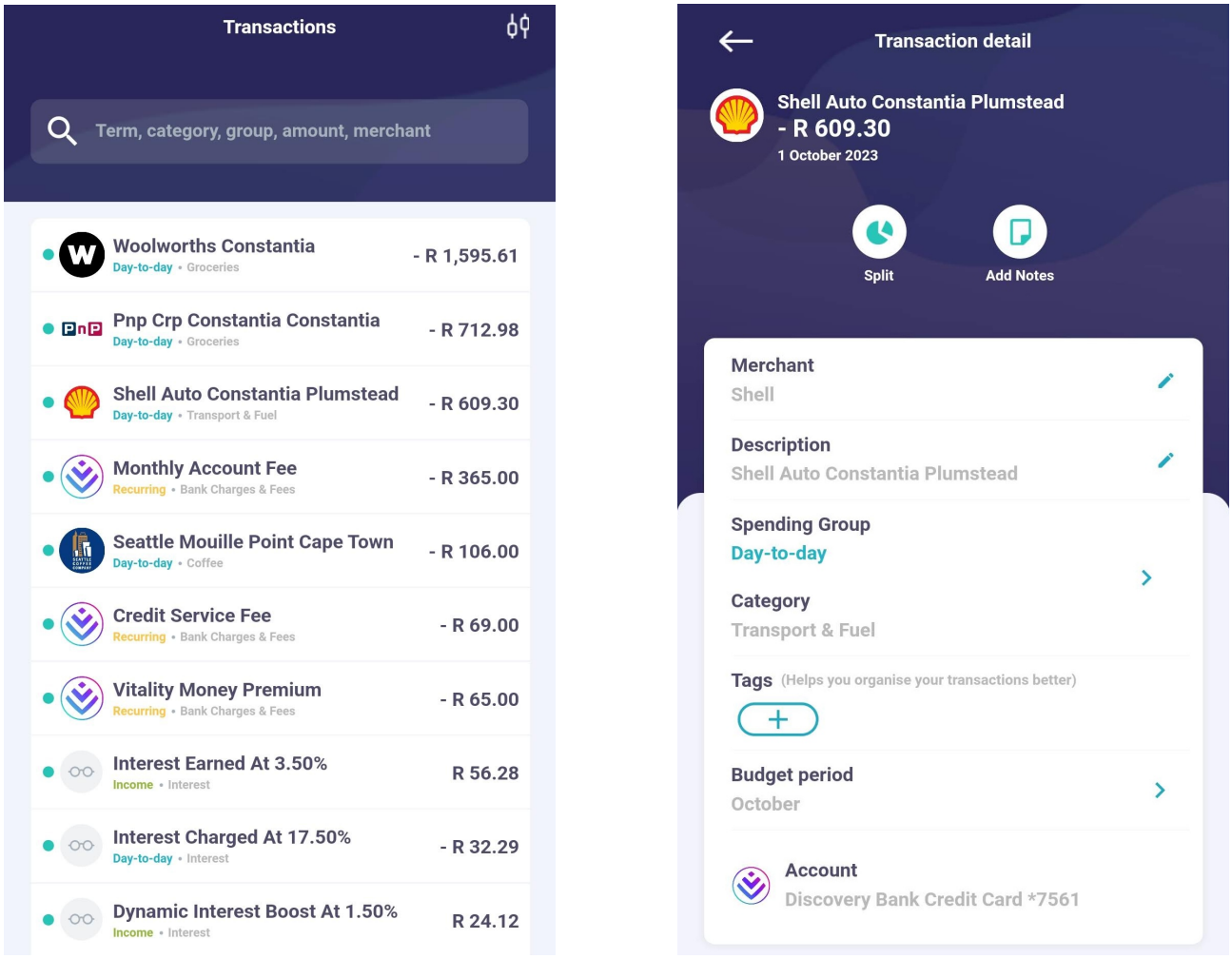


Figure 3.1: 22seven How it works 22seven (2023)

With over 700 000 registered users and having been recently acknowledged as one of the top 200 fintechs globally (Browne, 2023), 22seven is undoubtedly a market leader in this space. With access to such interesting data, it is easy to see that there is no dearth of options when it comes to exploring these data from a statistical standpoint.

One of the key features of the 22seven platform is the automatic categorisation of a user's financial transactions into categories. This is illustrated in Figure 3.2.



(a) All transactions screen

(b) Individual transaction screen

Figure 3.2: Illustration of 22seven categorisation feature

Figure 3.2a displays a list of transactions that a user can see and can be retrieved from their linked accounts. Figure 3.2b displays the details for a single transaction that are assigned by 22seven. Before continuing, it is important to **note that the illustrations shown here are not the exact data shared for this dissertation. They are included to illustrate the data generating mechanism before specifying the exact data shared in the following subsections.**

For every transaction, the following information can be seen:

- Description
- Amount

- Category
- Date

In addition to this, it is known to which customer the transaction belongs as there exists a unique user identifier (UUID) which is also attached to the transaction. This UUID is a randomly generated sequence of letters and numbers and thus does not contain any personally identifiable information of the user. It does, however, link to a set of demographic information which are voluntarily provided by the user. This can include age, gender and ethnicity which are self-reported by a user.

The other main feature mentioned previously is that of budgeting. This feature is illustrated in Figure 3.3. For each category of spend, there is a budgeted amount (on the right of the /) and an actual spend amount (on the left of the /). This data is aggregated by month from the transaction data shown before. For example, the user has gone over budget in the Eating Out & Takeaways and Groceries categories by spending R1 302 and R5 272 while budgeting R779 and R2 209 respectively. A category the user has remained within budget on is Housekeeping in which they budgeted R400 and have only spent R350.

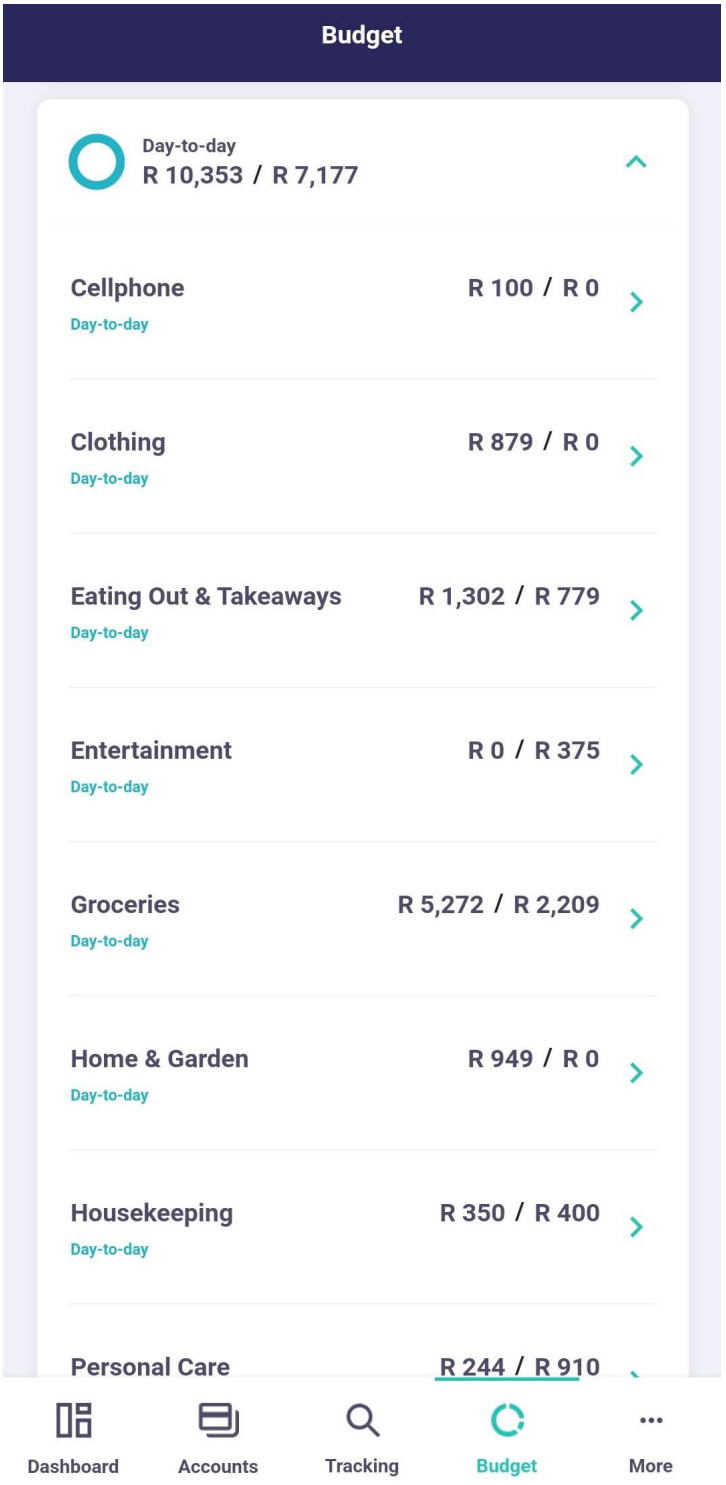


Figure 3.3: Illustration of 22seven budgeting feature

Specifics of data shared: Do spending patterns over time reveal distinct groups?

To answer this question, a profile of spend was created for users active in the first 6 months of 2022. This profile was created with the aim to represent what a user typically spends in the days following the receipt of their salary. To illustrate this, an example for how this was created is shown in Table 3.1.

	Day 0	Day 1	Day 2	Day 3
Actual Spend	2500	0	1000	200
Cumulative Spend	2500	2500	3500	3700
Cumulative Proportion of Salary Spent	10%	10%	14%	15%

Table 3.1: Illustrative example for the creation of spend profiles for a user earning a salary of 25 000

Assume the example user earned a salary of R25 000. On the day they received their salary, they spent R2 500 which represents 10% of their salary. By day 3, they have spent a total of R3 700 which is approximately 15% of their total salary. This cumulative proportion of salary spent is hereafter referred to as a customer's spend profile. For each customer, their spend profiles are calculated. The primary data therefore have the following fields:

- UUID
- Day
- Cumulative Proportion of Salary Spent

After a visual examination of the data, it was determined that the raw spend profiles contained extreme outliers which made it dificult to model the data well. To overcome this, a log transform with base 10 is applied to the spend profiles. These are the data used to perform the clustering.

In addition to this, a second set of data was shared. These included the UUID of the user to be able to link the two sets of data. The second set of data contained the estimate of monthly income, a user's age as well as their proportion spend in the three main spending groups within the app:

- Day to Day: This encompasses the variable expenses that an ordinary person incurs and includes items such as Groceries, Fuel etc.
- Recurring: These are analogous to fixed costs for an ordinary person and would include items such as Debit Orders for Medical Aid, Insurance, etc.

- Invest-Save-Repay: These are amounts used to pay down debt and/or put into savings and investment vehicles to grow money

Specifics of data shared: Does setting a budget work?

In order to answer this question, it was determined that the budget data as illustrated previously in Figure 3.3 was required. The data shared thus contained the following information:

- UUID
- month
- category
- actual spend
- budgeted spend

In terms of the scope of the data, a sample of customers who were active from 2018-2022 were identified. This cohort comprised 1 317 unique customers. In addition, only budgets in the following three categories were considered: Groceries, Eating Out & Takeaways and Transport & Fuel. While 22seven does track more than 50 categories of spend, these 3 are the most populous which is unsurprising as they are the "necessities" of living. Therefore, these data are also more likely to be complete and accurate.

In addition to this, a second set of data with the user's demographic data was shared. These included age, ethnicity and gender as well as a UUID to be able to link these demographics to the budgeting data. It is noted that the sample of users was selected such that all demographic information for all users was available.

Chapter 4

Do spending patterns over time reveal distinct groups?

4.1 Introduction

Understanding consumer behavior in the realm of financial transactions is a complex yet crucial endeavor for various industries and researchers alike. The intricacies of spending habits, financial patterns, and the underlying factors driving these behaviors have sparked a growing interest in employing sophisticated analytical techniques to uncover insights.

This chapter delves into a comprehensive exploration of clustering analysis applied to longitudinal spending data. The primary objective is to dissect and categorize consumers based on their spending profiles over time. The dataset used in this study comprises detailed transactional information, allowing for a nuanced examination of spending dynamics, patterns, and trends exhibited by different groups of consumers.

The exploration commences with an overview of multiple clustering techniques, each offering unique perspectives and methodologies to uncover hidden structures within the spending data. Through the implementation of diverse methodologies such as Group Based Trajectory Modeling (GBTM), k-means for Longitudinal Data, Growth Mixture Modeling (GMM), Longitudinal Latent Profile Analysis (LLPA), and Feature Based Clustering (FBC), the analysis aims to capture and understand the varying spending trajectories exhibited by distinct clusters of consumers.

Additionally, the chapter incorporates a comparative analysis of the effectiveness of these clustering methodologies. By evaluating the degree of agreement and disagreement between different clustering techniques, insights emerge regarding the consistency and robustness of identified consumer clusters across varied analytical approaches.

Moreover, the examination extends beyond clustering techniques. It encompasses a covariate analysis that investigates how demographic factors, such as age, income, and proportion of spending across different expense categories, relate to the identified consumer clusters. This facet of the analysis seeks to unearth connections between consumer spending behaviors and demographic

attributes, shedding light on the interplay between financial habits and socio-economic characteristics.

Ultimately, the overarching goal of this chapter is to provide a comprehensive understanding of consumer spending behavior through a multifaceted analytical lens. By dissecting spending patterns, identifying consumer clusters, and examining the influence of demographic factors, the study attempts to offer valuable insights into the intricacies of financial behaviors, paving the way for informed decision-making and future research directions in consumer finance analytics.

4.2 Methods

This dissertation focuses on the examination of various methods employed for clustering longitudinal data. The research investigates three distinct categories of clustering methods: **cross-sectional clustering**, **mixture modeling**, and **feature-based clustering**. The chapter begins with a brief overview of each of these methods before delving into a more detailed examination later in this section.

Cross-sectional analysis involves the observation of multiple features of interest $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ at a single time point for a set of n individuals. However, in the case of longitudinal data, there is an encounter with a set of time points $(t_1, \ldots, t_{T_i})$ representing the instances at which a measured variable is observed, along with a set of observations $(y_{i1}, \ldots, y_{1T_i})$ for each individual. Notably, the observation times may not align across all individuals, resulting in varying sets of measurements. Consequently, applying cross-sectional analysis tools directly to cluster longitudinal data is typically impractical. Fortunately, in the research, data is recorded at regular intervals, enabling its representation as a set of features. This representation facilitates the application of ordinary cross-sectional clustering methods to explore the clustering of longitudinal data. In the study, the purely non-parametric **k-means** method is compared with **Latent Profile Analysis**, a statistical theory-based approach.

In contrast to cross-sectional analysis, which disregards the temporal aspect, the **mixture modeling** approach incorporates temporal information by modeling the profiles over time. This approach aims to create latent profiles that describe different clusters of individuals. The present study explores two methods within this category: **Group-Based Trajectory Modeling** and **Growth Mixture Modeling**. Both methods involve fitting linear models to the trajectories and utilizing them to assign cluster memberships.

In the **feature-based clustering** approach, the time series aspect of the data is once again disregarded. Instead, the longitudinal profiles are transformed into a set of features before performing clustering. The features may include statistics such as mean value, minimum, maximum, and others derived from the trajectory data. Subsequently, standard cross-sectional clustering is conducted on these derived features rather than the original data. In the study, a **k-means Clustering** approach is employed on the derived features.

4.2.1 k-means

k-means clustering is a widely used unsupervised machine learning algorithm that aims to partition a given dataset into k distinct clusters based on the similarity of data points. It is an iterative algorithm that minimizes the within-cluster sum of squares (WCSS) by iteratively assigning data points to the nearest cluster centroid and updating the centroids until convergence of the algorithm is achieved.

The k-means clustering algorithm is initialised as described below:

1. Select the number of clusters, denoted by k, that the algorithm should aim to identify.
2. Randomly initialize k cluster centroids, denoted by $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$, as the starting points for the algorithm. Each centroid is a vector in the same feature space as the input data.

The main steps of the k-means algorithm involve iteratively assigning data points to the nearest cluster centroid and updating the centroids until convergence is achieved. This process can be described as follows:

1. **Assign Data Points to Nearest Centroids:** For each data point $\boldsymbol{X}_i$ in the dataset, calculate its distance to each centroid $\boldsymbol{\mu}_j$ using a distance metric, typically the Euclidean distance.

$$d(\boldsymbol{X}_i, \boldsymbol{\mu}_j) = ||\boldsymbol{X}_i - \boldsymbol{\mu}_j||^2$$

   where $||\cdot||$ denotes the Euclidean norm. The data point $\boldsymbol{X}_i$ is then assigned to the nearest centroid by selecting the centroid that minimizes the distance:

$$c_i = \arg\min_j d(\boldsymbol{X}_i, \boldsymbol{\mu}_j)$$

   where c_i represents the cluster assignment of data point $\boldsymbol{X}_i$.

2. **Update Centroids:** After assigning all data points to the nearest centroids, the next step is to update the centroids based on the newly formed clusters. The centroid of each cluster is recalculated as the mean of all data points assigned to that cluster:

$$\boldsymbol{\mu}_j = \frac{1}{\sum_{i=1}^{n} \mathbb{1}(c_i = j)} \sum_{i=1}^{n} \boldsymbol{X}_i \cdot \mathbb{1}(c_i = j)$$

   where $\mathbb{1}(c_i = j)$ is an indicator function that equals 1 if $c_i = j$ and 0 otherwise.

3. Repeat steps 1 and 2 until convergence is achieved. Convergence is typically determined when there is no or minimal change in the cluster assignments or centroids. Various stopping criteria, such as a maximum number of iterations or a predefined threshold below which any change in the WCSS is deemed insignificant, can be used.

The final output of the k-means clustering algorithm is a set of k cluster centroids and the assignments of data points to these clusters (James et al., 2013).

4.2.2 Latent Profile Analysis

Latent profile analysis (LPA) is a statistical approach used to uncover hidden clusters or profiles within a population. It assumes that individuals can be assigned to one of several unknown clusters based on their characteristics. LPA is often referred to as Gaussian Mixture Modeling because each cluster is characterized by a Multivariate Normal Distribution. The clusters are defined by their mean vector, denoted as $\boldsymbol{\mu}_g$, and their variance-covariance matrix, denoted as $\boldsymbol{\Sigma}_g$, where $g = 1, \ldots, G$. Here, G represents the specified number of total clusters.

Rather than assigning individuals to a single cluster, LPA assigns them probabilistically to all clusters. This means that each individual is associated with a probability of belonging to each cluster, instead of being exclusively assigned to one. The probabilities of cluster membership are derived from the observed variables and the estimated parameters of the specified statistical model.

For an individual i at time t_j, given the cluster membership g, the observed variable $y_{i,j}$ is modeled as follows:

$$y_{i,j} = \mu_{g,j} + \epsilon_{g,i,j} \tag{1}$$

where $\mu_{g,j}$ represents the cluster-specific mean at time t_j , $\epsilon_{g,i,j} \sim N(0, \sigma_{g,j})$, and $\sigma_{g,j}$ is the cluster-specific standard deviation at time t_j.

To calculate the probability density of the observations for individual i, we need to consider all G clusters. This is achieved by marginalizing over all possible cluster assignments, resulting in the following equation:

$$f(\mathbf{y}_i) = \sum_{g=1}^{G} \pi_g \prod_{j=1}^{n} \phi(y_{i,j} | \mu_{g,j}, \sigma_{g,j}) \tag{2}$$

In the above equation, $\phi(\cdot)$ represents the probability density function of the normal distribution with mean $\mu_{g,j}$ and standard deviation $\sigma_{g,j}$. The term π_g denotes the cluster proportion for cluster g, where $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$. The cluster proportion represents the probability of an individual belonging to a specific cluster.

To estimate the latent profile model, two essential quantities need to be computed: a) the parameters of the clusters, denoted as $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_G)$, and b) the cluster membership matrix $\boldsymbol{Z}$, where $z_{i,g}$ represents the probability of individual i belonging to cluster g. These quantities are typically estimated using the Maximum-Likelihood estimation method with the EM-algorithm.

The EM-algorithm consists of two iterative steps: the E-step and the M-step. In the E-step, the cluster assignment probabilities ($\boldsymbol{Z}$) are estimated based on the current parameters $\boldsymbol{\theta}$ and the observed variables $\boldsymbol{y}$. This step calculates the probability of each individual belonging to each cluster, considering the estimated parameters. In the M-step, the parameters $\boldsymbol{\theta}$ are updated using the current cluster assignments $\boldsymbol{Z}$. This step maximizes the likelihood of the observed data by adjusting the parameters based on the current assignments. The iterations between the E-step and M-step are repeated until there is minimal change in the log-likelihood, indicating convergence of the model estimation (Teuling et al., 2021; Oberski, 2016).

4.2.3 Group Based Trajectory Modelling

Group Based Trajectory Modelling (GBTM) is a statistical approach that extends the concepts of Latent Profile Analysis (LPA) by explicitly incorporating the time component into the modeling process. While LPA assumes independence of observations at each time point, GBTM recognizes the interdependencies among observations over time, allowing for a more accurate representation

of complex longitudinal data.

GBTM involves the specification of a set of G latent time series profiles or clusters, similar to LPA. However, instead of characterizing each profile using a Multivariate Normal distribution as in LPA, GBTM utilizes a set of linear regression coefficients, denoted as $\boldsymbol{\beta}_g$, to describe the trajectory of each cluster profile. These coefficients capture the relationship between the observed variables and the time variable within each cluster.

At a given time point $t_{i,j}$, where i represents the individual and j denotes the specific time point, GBTM considers a design vector $\mathbf{x}_{i,j}$ associated with that time point. The trajectories of the observed variables, given membership to a particular cluster g, are modeled as:

$$y_{i,j} = \mathbf{x}_{i,j}^T \boldsymbol{\beta}_g + \epsilon_{g,i,j} \tag{3}$$

In the above equation, $\boldsymbol{\beta}_g$ represents the cluster-specific regression coefficients, which describe how the observed variables change over time within each cluster. The term $\epsilon_{g,i,j}$ denotes the error term associated with the gth cluster at time point $t_{i,j}$, assumed to follow an independent and identically distributed (iid) Normal distribution with mean zero and standard deviation σ_g. This error term captures the variability in the observed trajectories around the cluster-specific regression line.

The marginal mean of the observed variable $y_{i,j}$ can be computed as the sum of the products of the cluster probabilities π_g and the corresponding design vector and regression coefficients:

$$E[y_{i,j}] = \sum_{g=1}^{G} \pi_g \mathbf{x}_{i,j}^T \boldsymbol{\beta}_g \tag{4}$$

The design vector $\mathbf{x}_{i,j}$ can take different forms depending on the desired trajectory shapes. In its simplest form, it can be represented as $(1, t_{i,j})$, resulting in linear trajectories. However, to capture more complex and flexible trajectories, polynomial or spline-based design vectors are often employed. These design vectors allow for smooth and nonlinear patterns in the trajectories, enabling a more accurate representation of the underlying dynamics of the data.

The parameters of the GBTM model, including the cluster probabilities π_g, regression coefficients $\boldsymbol{\beta}_g$, and error variances σ_g, are estimated through Maximum Likelihood Estimation (MLE). The estimation process typically employs the Expectation-Maximization (EM) algorithm, which itera-

tively optimizes the likelihood function by alternately updating the latent class assignments and estimating the model parameters until convergence of an algorithm is achieved.

GBTM provides a powerful framework for identifying latent clusters with distinct trajectory patterns in longitudinal data. By explicitly considering the time component and allowing for flexible trajectory shapes, GBTM enables a more nuanced understanding of how individuals or groups evolve over time, shedding light on the underlying processes and heterogeneity within the data (Teuling et al., 2021; Oberski, 2016).

4.2.4 Growth Mixture Modelling

Growth Mixture Modelling (GMM) is an extension of Group Based Trajectory Modelling (GBTM) that allows for the inclusion of individual-specific random effects in the model formulation. By incorporating these random effects, GMM provides a more flexible and nuanced approach to capturing individual differences in the growth trajectories within each latent cluster.

Similar to GBTM, GMM specifies a set of G latent clusters or profiles, with each cluster characterized by its own set of regression coefficients $\boldsymbol{\beta}_g$. However, in GMM, the model formulation incorporates individual-specific random effects, denoted as $\mathbf{u}_{g,i}$, which capture the unique characteristics of each individual within a particular cluster.

The observed variable $y_{i,j}$ at time point $t_{i,j}$ is modeled as a combination of the fixed effects (described by $\mathbf{x}_{i,j}^T \boldsymbol{\beta}_g$), the individual-specific random effects (represented by $\mathbf{z}_{i,j}^T \mathbf{u}_{g,i}$), and the residual error term ($\epsilon_{g,i,j}$):

$$\begin{aligned} y_{i,j} &= \mathbf{x}_{i,j}^T \boldsymbol{\beta}_g + \mathbf{z}_{i,j}^T \mathbf{u}_{g,i} + \epsilon_{g,i,j} \\ \mathbf{u}_{g,i} &\sim N(0, \boldsymbol{\Sigma}_g) \\ \epsilon_{g,i,j} &\sim N(0, \sigma_g^2) \end{aligned} \tag{5}$$

Here, $\mathbf{z}_{i,j}$ represents the design matrix associated with the individual-specific random effects, allowing for the incorporation of various individual-level predictors or time-varying covariates. The random effects $\mathbf{u}_{g,i}$ are assumed to follow a Multivariate Normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_g$, capturing the individual-specific deviations from the overall cluster trajectory. The residual errors $\epsilon_{g,i,j}$ are assumed to be normally distributed with mean zero and variance $\sigma_{\epsilon,g}^2$, representing the unexplained variability within each cluster.

The marginal mean of the observed variable $y_{i,j}$ given the individual-specific random effects $\mathbf{u}_{g,i}$ is computed as:

$$E[y_{i,j}|\mathbf{u}_{g,i}] = \sum_{g=1}^{G} \pi_g \left(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_g + \mathbf{z}_{i,j}^T \mathbf{u}_{g,i}\right) \tag{6}$$

This equation represents the expected value of $y_{i,j}$ conditional on the individual-specific random effects, weighted by the cluster probabilities π_g.

The parameters of the GMM model, including the cluster probabilities π_g, regression coefficients $\boldsymbol{\beta}_g$, individual-specific random effects covariance matrix $\boldsymbol{\Sigma}_g$, and residual variances σ_g^2, are estimated through Maximum Likelihood Estimation (MLE) using the Expectation-Maximization (EM) algorithm. The EM algorithm iteratively maximizes the likelihood function by updating the latent class assignments, estimating the model parameters, and repeating these steps until convergence is achieved.

By incorporating individual-specific random effects, GMM allows for the identification of distinct subgroups with unique growth trajectories within each latent cluster. This modeling approach provides a more comprehensive understanding of individual differences and captures the heterogeneity in the data beyond what can be explained by the fixed effects alone (Teuling et al., 2021; Oberski, 2016).

4.2.5 Adjusted Rand Index

In order to compare the results of the different clustering methods, a performance measure is usually required and for this dissertation, the Adjusted Rand Index (ARI) is chosen to be used. The ARI is a statistical measure used to evaluate the similarity between two different clusterings of data points. It quantifies the level of agreement or similarity between the clusters produced by different algorithms or methods.

Calculation of ARI:

The ARI computes a similarity score by measuring the proportion of agreements between two sets of clusters while considering the expected random agreements. It operates by comparing all pairs of samples and counting pairs that are assigned in the same or different clusters in both the predicted and true clusterings.

The formula for ARI involves:

- **a**: The number of pairs of data points that are in the same cluster in both the predicted and true clusterings.
- **b**: The number of pairs of data points that are in different clusters in both the predicted and true clusterings.
- **c**: The number of pairs of data points that are in the same cluster in the predicted clustering but in different clusters in the true clustering.
- The number of pairs of data points that are in different clusters in the predicted clustering but in the same cluster in the true clustering.

The ARI formula can be expressed as:

$$ARI = \frac{\text{Coefficient} - \text{Expected}}{\text{Maximum} - \text{Expected}}$$

Here, the *Coefficient* term is calculated as $\frac{a+b}{\binom{n}{2}}$, where n is the total number of data points. The *Expected* term represents the expected value of agreement between two random clusterings, and the *Maximum* term denotes the maximum possible value of the agreement index.

Interpretation of ARI Values:

- An ARI value close to 1 indicates strong agreement or similarity between the clusterings.
- A value around 0 suggests no agreement between the clusterings, which could be due to random chance.
- A negative value implies that the observed agreement is less than what is expected by chance.

Usage in Comparative Analysis:

In this study, the ARI was utilized to assess the agreement between different clustering methods. Pairwise ARIs were computed between various clustering algorithms used in the analysis, such as Feature Based Clustering, Growth Based Trajectory Modelling (GBTM), Growth Mixture Modelling (GMM), Longitudinal Latent Profile Analysis (LLPA), and k-means for Longitudinal Data. This approach allowed for the quantification of the degree of similarity between the resulting clusters and determination of the level of consistency across different methods in clustering similar user profiles (Santos and Embrechts, 2009).

4.3 Results

The analysis begins by plotting the trajectories of all individuals in Figure 4.1. From this, it is not immediately evident that there are natural clusters in the data; therefore, an exploratory analysis is conducted to identify any possible longitudinal clusters.

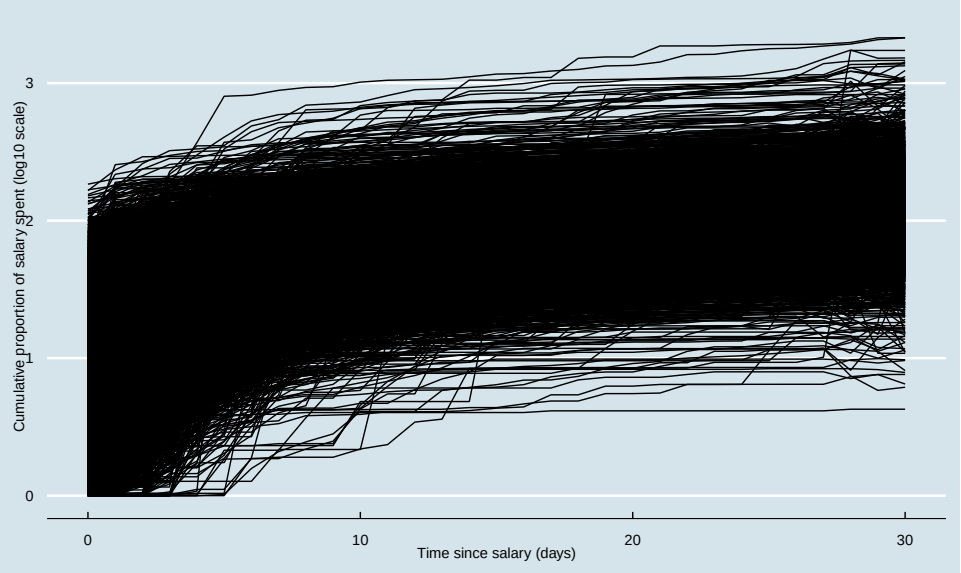


Figure 4.1: Trajectories of individuals

4.3.1 k-means

The analysis commences with the k-means approach, investigating clusters of sizes 2 to 8. Due to computational limitations, fitting larger number of clusters was not investigated. For each cluster size, the algorithm is run 10 times with random different starting positions, and the best result is chosen as the cluster assignment. Here, 'best' is defined as the assignment that minimises the within-cluster sum of squares.

Before delving into choosing the optimal number of clusters, the validity of the clusters formed using the k-means analysis is first investigated. This is done in two ways: firstly, by a statistical test comparing the cluster sizes formed under each of the cluster sizes. If there were no heterogeneity in the data, it would be expected that the profiles would be distributed fairly uniformly and thus any attempt to cluster would simply partition the data in a fairly even manner. In such a case, the number of observations in each cluster would be almost equal between clusters. The presence of heterogeneity is checked by means of a chi-squared goodness of fit test, comparing the cluster

sizes as illustrated in Table 4.1 to the case in which the observations are equally distributed between the clusters. The p-value for each test was found to be infinitesimally small, indicating that the clusters formed might indeed explain some heterogeneity in the data.

Number of Clusters	Cluster Label							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
k=2	7911	6482						
k=3	6516	4743	3134					
k=4	5620	4111	2999	1663				
k=5	3805	3221	3145	2687	1535			
k=6	3217	3182	3085	2631	1507	771		
k=7	2983	2738	2653	2598	2147	734	541	
k=8	2390	2305	2256	2068	1962	1929	810	673

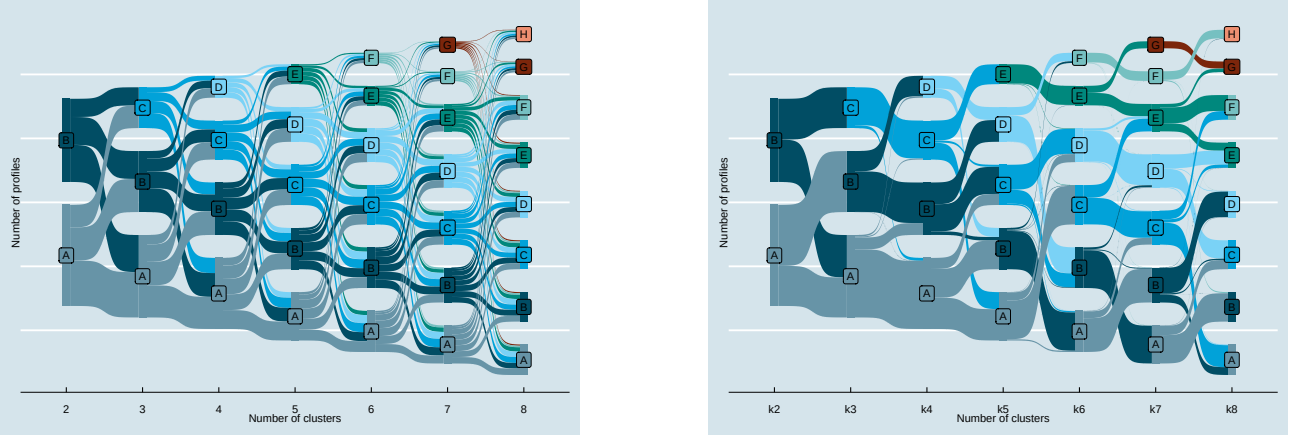
Table 4.1: Cluster sizes for k-means clustering of longitudinal profiles of spend. The cluster labels A to H are labels assigned to distinguish the different clusters and have no inherent meaning across the differing number of clusters.

The second way in which the cluster validity was examined was by means of a visual test to identify if observations which cluster together at a lower number of clusters continue to cluster together as the number of clusters is increased. To illustrate the point, consider the case when at each cluster size, observations are randomly assigned to one of the clusters based on the distribution observed in Table 4.1. i.e., when there are 2 clusters, 7911 observations are randomly assigned to cluster A and 6482 observations to cluster B; when there are 3 clusters, 6516 are assigned to cluster A, 4743 to cluster B, and 3134 to cluster C, and so on.

In such a case, the Goodness of Fit test conducted above would still be passed, but this means that the clusters are fairly meaningless. It is expected that if two trajectories are very similar, they should fall in the same cluster even as the number of clusters is increased. Naturally, there will be some splintering as the k-means algorithm guarantees that k clusters will always be returned, and it is up to the researcher to determine at which value of k, it is no longer informative to splinter the observations.

The point is further emphasised in Figure 4.2a in which the precise randomisation described above is conducted, illustrating the flow of individuals between clusters. Incrementing the number of

clusters simply results in observations being assigned based on the sample weight of the previous clustering. This is probably most clearly seen moving between k=3 and k=4. Across all 4 clusters, it can be seen quite easily that the largest contribution comes from the A cluster in the previous increment, followed by B and then C. The proportion contribution is also quite similar between the 4 clusters.



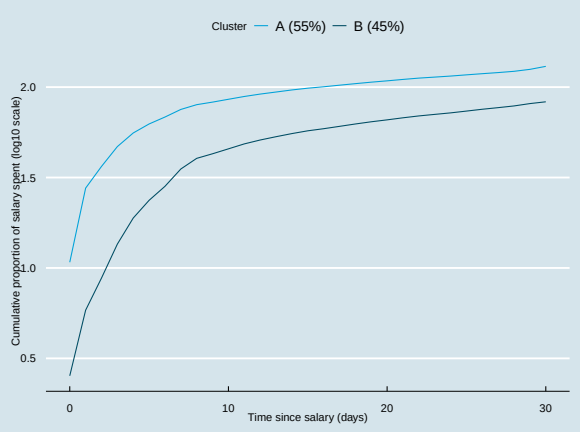
(a) Random assignment

(b) Actual clusters formed from k-means

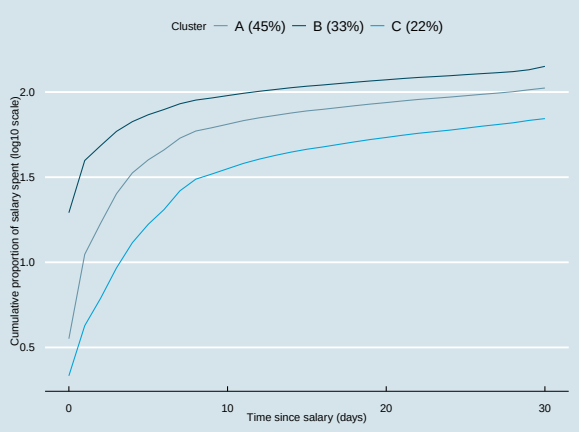
Figure 4.2: Flow of trajectories in the same cluster as the number of clusters is incremented

Figure 4.2a can be contrasted with Figure 4.2b which illustrates the actual flow of trajectories clustered with k-means as the number of clusters investigated is incremented. For instance, when moving from k=2 to k=3, something quite different is observed. At k=3, cluster C consists entirely of observations which were in cluster B at k=2 and cluster B at k=3 consists entirely of observations which were initially in cluster A at k=2. This indicates that similar profiles are indeed moving together, lending credence to the idea of validity of the clusters.

What is happening between k=2 and k=3 can be attempted to be explained by looking at the mean cluster profiles in Figure 4.3. In the simplest sense, Cluster A_2 can simply be described as spending more than cluster B_2. On closer inspection, something interesting shows up. Since a log_{10} scale of the data is used, any values above 2 can be considered as spending more than 100% of salary, classifying the two cluster centroids as overspenders vs underspenders. Moving to 3 clusters, these two cluster centroids are still seen, but there is an additional centroid which is spending pretty close to 100% i.e., just about breaking even. Thus, consistency in the clusters is observed, and more information about the heterogeneity in the data is gleaned as more clusters are added.



(a) k=2



(b) k=3

Figure 4.3: Mean trajectory profiles for k-means clustering

The question that then follows is how many clusters should be fitted to the data? Naturally, as the number of clusters is increased, the better the fit to the data will be, but there needs to be a tradeoff between interpretability and fitting to the data. To aid this decision, Figure 4.4 is produced, in which the Bayesian Information Criterion (BIC), Weighted Root Mean Square Error (WRMSE), and Weighted Mean Absolute Error (WMAE) for each of the cluster sizes from 2 to 8 are computed. The weighted measures are weighted by the probability of assignment to a cluster. Since k-means assigns observations to a cluster deterministically, these weighted measures are equivalent to the unweighted measures.

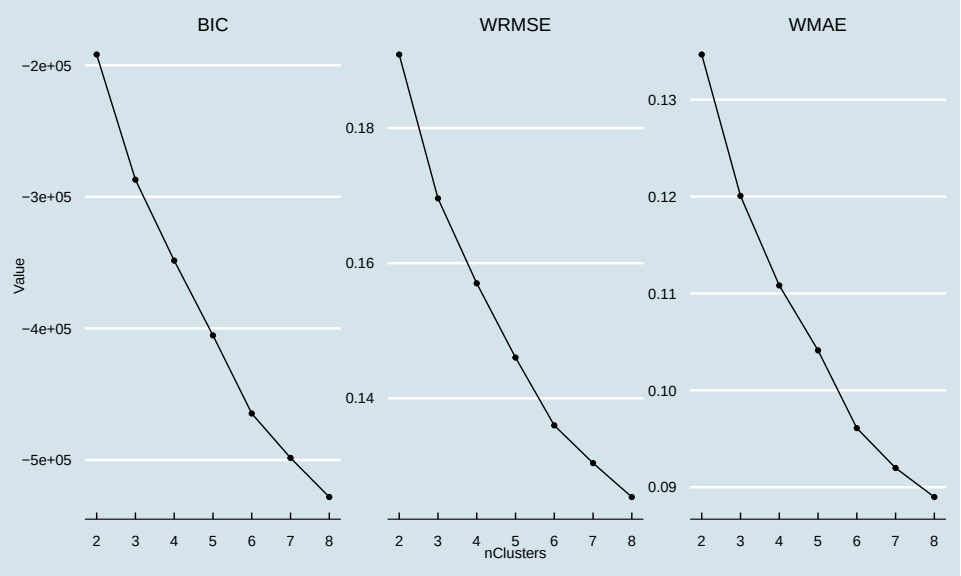
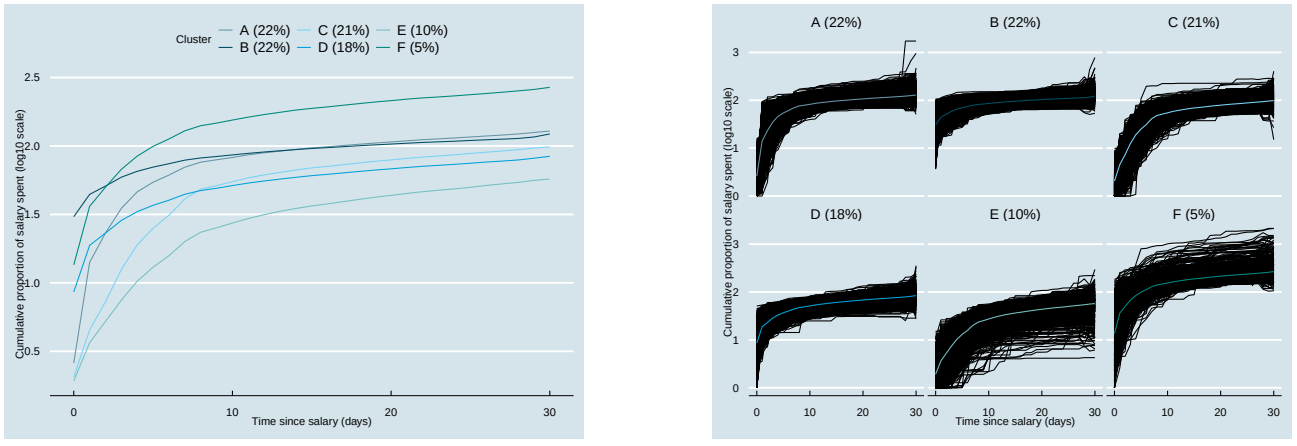


Figure 4.4: Fit metrics for k-means clustering

The typical elbow test is not clear in this plot, but it can be seen that post k=6, the marginal gain in fit does not seem to be as useful as before k=6. For this reason, k=6 is selected as the "best" choice and the resulting clusters are produced in Figure 4.5.



(a) Mean profiles

(b) Mean profiles with the trajectories assigned

Figure 4.5: Final chosen k-means clustering

From the resulting cluster profiles, it is straightforward to see that cluster F consists of individuals who tend to spend significantly above what their income implies they are able to spend. Clusters D and E end the month at similar places, but cluster D tends to start the month by spending more.

Clusters A and B exhibit a similar pattern in ending at a similar place but with B starting much higher. Cluster C has a similar shape to cluster E, but the initial rate of spend seems higher, resulting in slightly more overspending. A more detailed characterisation of clusters will be provided in Section 4.4.

4.3.2 Latent Profile Analysis

As with the k-means analysis, the investigation includes various numbers of clusters or, in this case, latent profiles. The analysis includes a range for the number of latent profiles between 2 and 8. Following the investigation of the validity of the clusters formed, several fit metrics are produced to aid in the decision of the "best" number of latent profiles. In Table 4.2, the number of profiles assigned to each latent profile for each of the sizes specified is presented. A Chi-Squared Goodness of Fit test was performed, revealing that at k=8, the result was not significantly different from a uniform distribution, but at each of the smaller cluster sizes, the resulting p-value was again very close to zero.

Number of Clusters	Cluster Label							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
k=2	6060	8333						
k=3	5674	3817	4902					
k=4	4526	3321	2479	4067				
k=5	3286	1486	2913	2605	4103			
k=6	3605	1839	1791	3471	1960	1727		
k=7	2182	2625	2309	2805	1357	1121	1994	
k=8	2061	1771	2339	1799	2444	2106	1023	850

Table 4.2: Cluster sizes for Latent Profile Analysis of longitudinal profiles of spend

In Figure 4.6, the visual test is created to check for the validity of clusters as the number of latent profiles increases. It is noted that at lower values, the clustering appears consistent, but as the number of latent profiles shifts from 6 to 7 or 7 to 8, there is significantly more mixing compared to the k-means analysis, which may indicate that a lower number of latent profiles is sufficient to describe the heterogeneity in the trajectories.

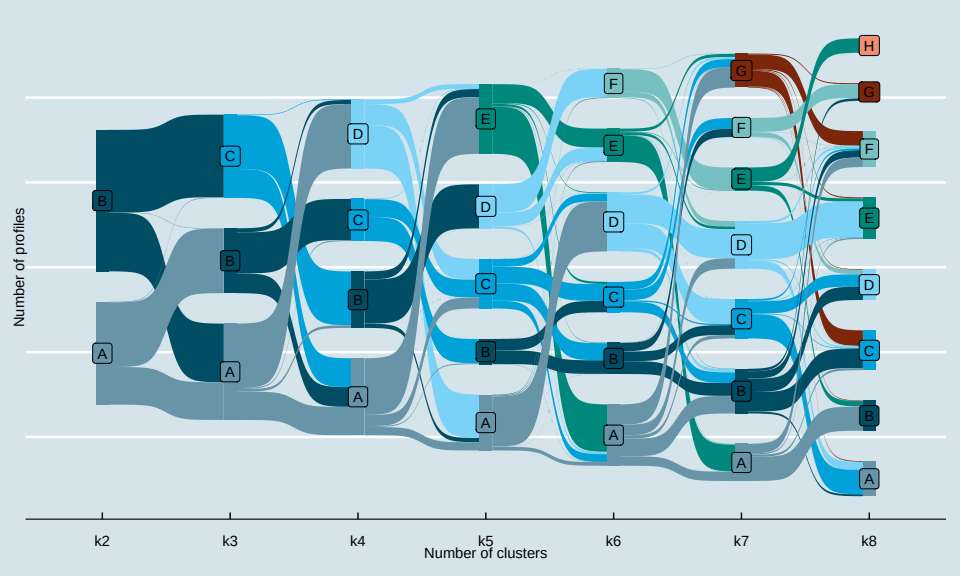


Figure 4.6: Flow of trajectories in the same cluster as we increment the number of latent profiles for Latent Profile Analysis

In order to set the final number of latent profiles, fit metrics are produced in Figure 4.7. From this, the hypothesis that fewer latent profiles might be sufficient is reaffirmed as the WMAE and WRMSE are flat between having 7 and 8 profiles, indicating that a better fit to the data is not necessarily gained. While this suggests that 7 latent profiles might be ideal, the change in the fit metrics is fairly small between 6 and 7 latent profiles. For consistency with other analyses, the number of latent profiles is set at 6, and the resulting clusters are presented in Figure 4.8.

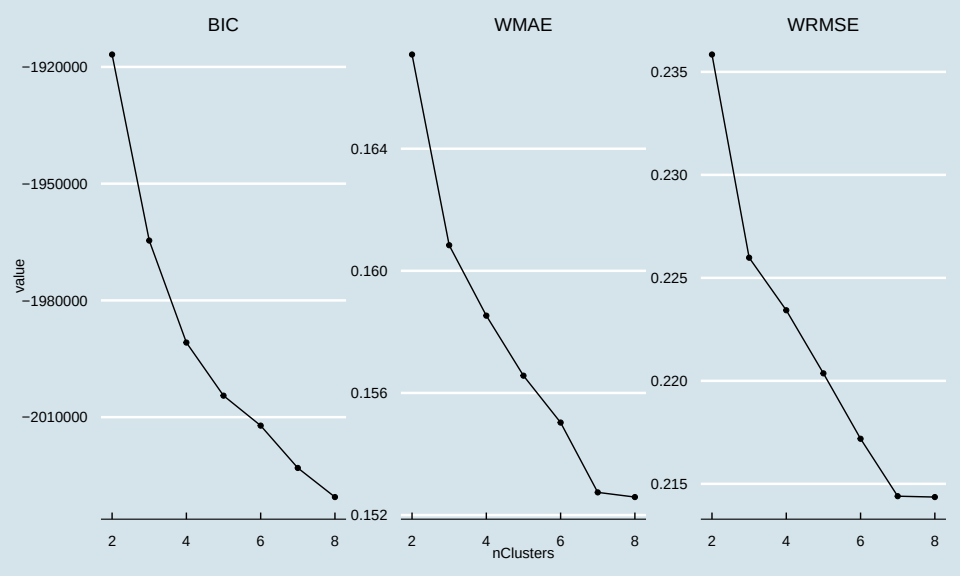
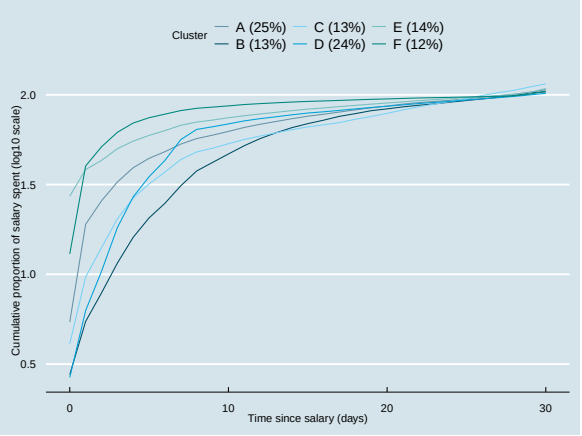
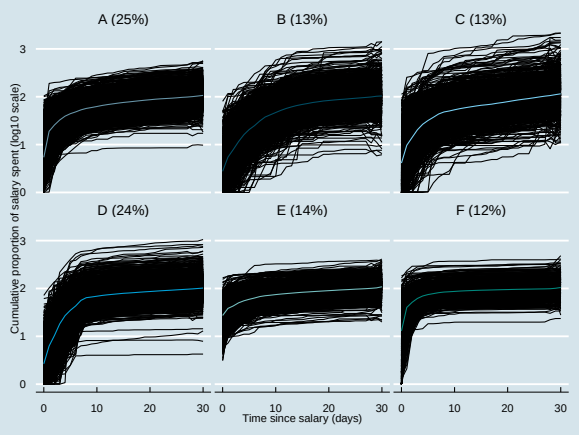


Figure 4.7: Fit metrics for Latent Profile Analysis



(a) Mean profiles



(b) Mean profiles with the trajectories assigned

Figure 4.8: Final chosen Latent Profile Analysis clustering

It is interesting to note that the Latent Profile cluster means all end the month at a similar point, i.e., roughly 100% of salary spent. This contrasts with the k-means analysis in which the distinguishing factor appeared to be the end-of-month position. Rather, the LPA analysis indicates that a distinguishing factor is how much is spent on the day of receiving a salary.

4.3.3 Group Based Trajectory Modelling

As mentioned in Section 4.2.3, GBTM implements a regression model to model the trajectories. In order to do so, it is necessary to specify the design vector which for this dissertation, is chosen to be a B-spline basis of degree 3 with three interior knots at $t = 7, 14$ & 21. These knots are chosen to coincide with the weeks of spend that comprise a month. Spending patterns are thus allowed to be flexible between weeks. i.e. the relationship of spend is allowed to be different in week 1 after receiving salary and week 4 when most people have already done the majority of their spending. The fitted regression equation for a profile has form:

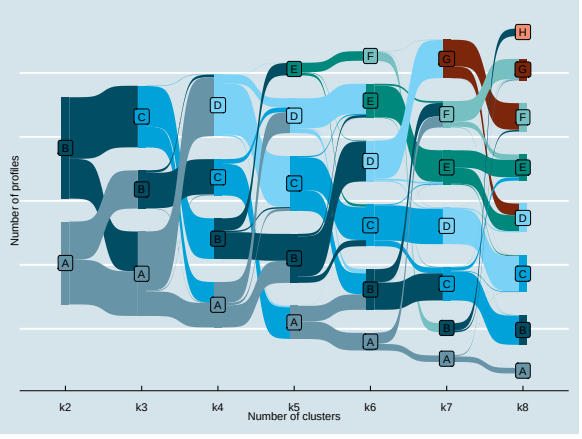
$$y_i = \beta_0^g + \beta_1^g t_i + \beta_2^g t_i^2 + \beta_3^g t_i^3 + \sum_{j=1}^{K} \gamma_j^g B_{j,3}(t_i) + \epsilon_i \tag{7}$$

Here, y_i represents the cumulative spend for user$_i$, t_i is the time variable, $\beta_0, \beta_1, \beta_2, \beta_3$ are regression coefficients, γ_j are spline coefficients, and $B_{j,3}(t_i)$ are the B-spline basis functions of degree 3 with knots at $t = 7, 14$, and 21 for group g.

As done previously, the model is fitted with 2 to 8 latent profiles and validity checks are produced in Figure 4.9. At all cluster sizes, a p-value for the Chi-Squared Goodness of fit test close to 0 is obtained and visual inspection of Figure 4.9b does not raise any issue of note.

	A	B	C	D	E	F	G	H
k=2	6474	7919						
k=3	6595	2985	4813					
k=4	3545	3178	2833	4837				
k=5	2587	3826	4285	2717	978			
k=6	1380	3169	3267	3186	2647	744		
k=7	508	703	2602	2855	2679	2046	3000	
k=8	439	2325	2873	2243	2044	2205	1669	595

(a) Cluster sizes for Group Based Trajectory Modelling



(b) Trajectories in the same cluster as the number of latent profiles for GBTM is incremented

Figure 4.9: GBTM Validity Checks

The investigation then moves on to determining the optimal number of latent profiles for which fit metrics are produced in Figure 4.10. It is not immediately obvious the optimal number of clusters to be chosen but a small kink at k=6 in the BIC and WRMSE plots is noticed, resulting in the selection of this as the optimal number of latent profiles.

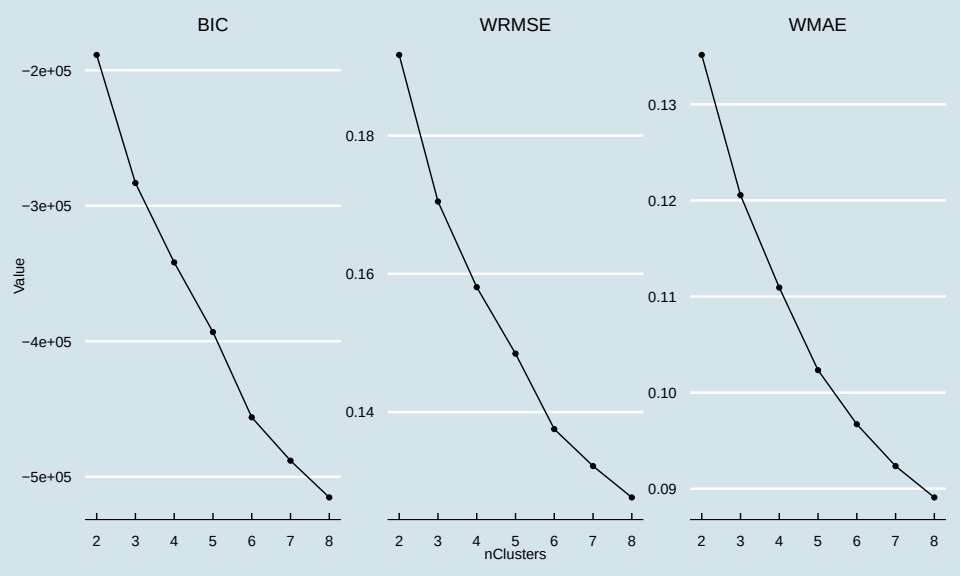
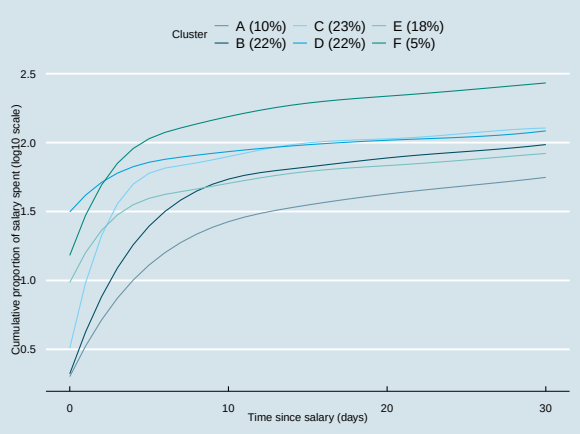
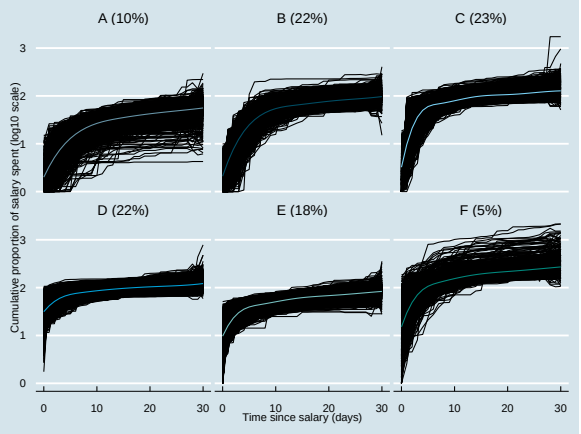


Figure 4.10: Fit metrics for Group Based Trajectory Modelling

In Figure 4.11, the resulting model profiles along with the profiles which would be assigned to the respective profiles are produced. It is interesting to note that these profiles are visually very similar to those found by k-means clustering but naturally smoother as they are defined by a smooth mean function.



(a) Mean profiles



(b) Mean profiles with the trajectories assigned

Figure 4.11: Final chosen Group Based Trajectory Modelling clustering

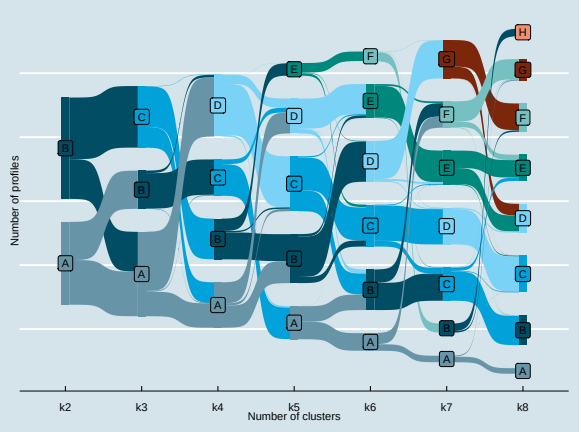
4.3.4 Growth Mixture Modelling

GMM can be considered an extension of GBTM, and the fitted profiles follow the same form as Equation 4.3.3 with the addition of the random effects as described in Section 4.2.4.

The resulting validity checks for fitting the model with varying number of latent profiles are produced in Figure 4.12. At all cluster sizes, a p-value for the Chi-Squared Goodness of fit test close to 0 is obtained. Visual inspection of Figure 4.12b does not raise any concerns, but it is noticed that GMM seems to produce some very consistent clusters compared to previous methods.

	A	B	C	D	E	F	G	H
k=2	6447	7946						
k=3	6211	4333	3849					
k=4	4674	3178	3863	2678				
k=5	4706	1917	2422	3643	1705			
k=6	3861	2406	2827	2604	1645	1050		
k=7	4217	2048	1705	2280	1076	1940	1127	
k=8	3043	1951	2401	1645	1026	1905	1553	869

(a) Cluster sizes for Growth Mixture Modelling



(b) Trajectories in the same cluster as the number of latent profiles for GMM is incremented

Figure 4.12: GMM Validity Checks

The investigation then moves on to determining the optimal number of latent profiles for which fit metrics are produced in Figure 4.2.3. It is not immediately obvious the optimal number of clusters to be chosen, but a small kink at k=6 in the WMAE plot is noticed, resulting in the selection of this as the optimal number of latent profiles.

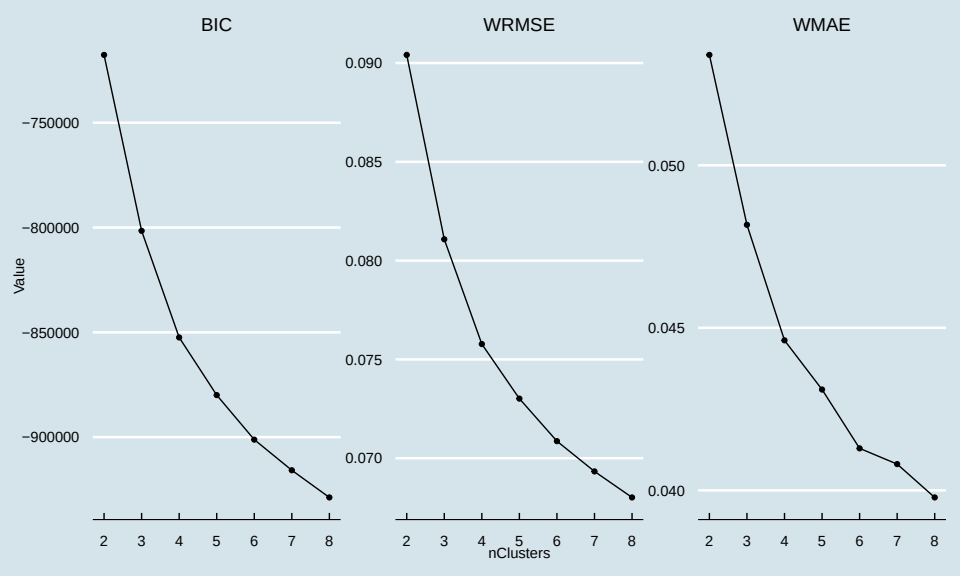
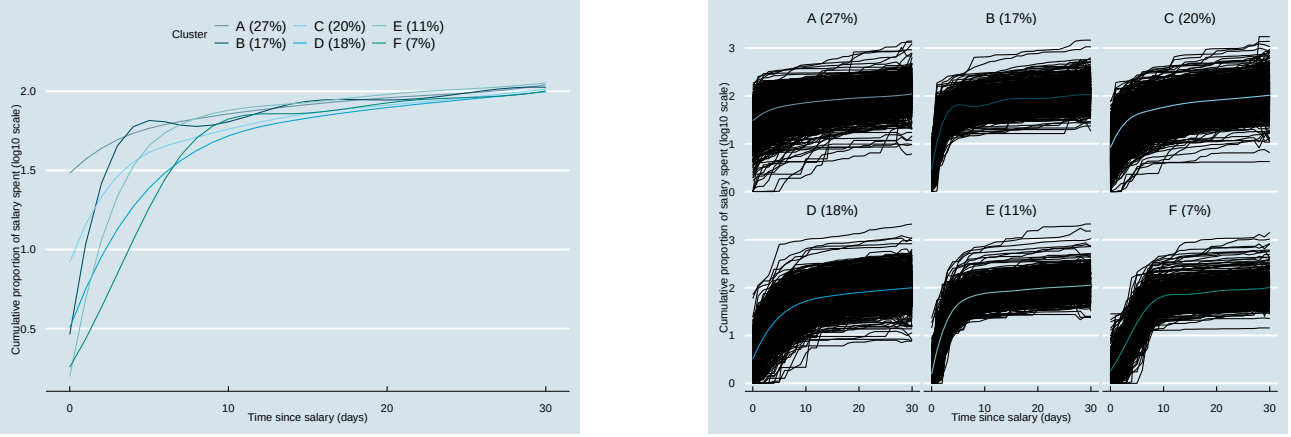


Figure 4.13: Fit metrics for Growth Mixture Modelling

In Figure 4.14, the resulting model profiles are produced along with the profiles which would be assigned to the respective profiles. It is interesting to note that these profiles are visually very similar to those found by the Latent Profile Analysis but are naturally smoother as they are defined by a smooth function. While GMM allows for a better fit to the data, it appears that this could potentially come at the cost of fitting to the noise in the data, with cluster B in particular having an implausible shape. The definition of the longitudinal data is that it is cumulative, so the expectation is that the profiles should be monotonic, which the profile for cluster B does not follow.



(a) Mean profiles

(b) Mean profiles with the trajectories assigned

Figure 4.14: Final chosen Growth Mixture Modelling clustering

4.3.5 Feature Based Clustering

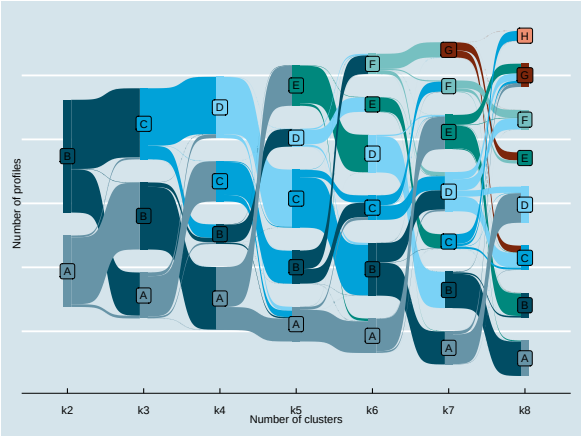
After performing the clustering in the above sections, it was noted that the distinguishing features seem to be around the rate of spend at specific time points. This was a more anecdotal observation and came from visual examinations of the resulting cluster profiles from the different methods. In order to examine the feature-based clustering, a judgement call based on a combination of the aforementioned results and domain knowledge was used to create a set of features for each spend profile. The set of features is described below:

- t_0: Value of spend profile at time 0
- t_7: Value of spend profile at time 7
- t_{30}: Value of spend profile at time 30
- $t_7 - t_0$: Value of spend profile at time 7 minus the value of spend profile at time 0
- $t_{30} - t_0$: Value of spend profile at time 30 minus the value of spend profile at time 0

Rather than having a set of 30 time points for each profile, the dimensionality of the profile has now effectively been reduced to 5 while believing that the important information regarding the shape of the profiles is still preserved. Following this, a standard k-means clustering approach was performed on the derived features in order to produce the clustering of users. The validity checks and fit metrics are presented in Figures 4.15 from which k=6 is chosen as the number of clusters to fit.

	A	B	C	D	E	F	G	H
k=2	5581	8812						
k=3	3592	5247	5554					
k=4	4921	1530	3152	4790				
k=5	2710	2590	4534	1400	3159			
k=6	2686	4119	1915	2934	1156	1583		
k=7	2607	2836	1139	3060	2670	895	1186	
k=8	2766	1933	1886	2821	873	1527	1814	773

(a) Cluster sizes for Feature Based Clustering



(b) Trajectories in the same cluster as the number of latent profiles is incremented for Feature Based Clustering

Figure 4.15: Feature Based Clustering Validity Checks

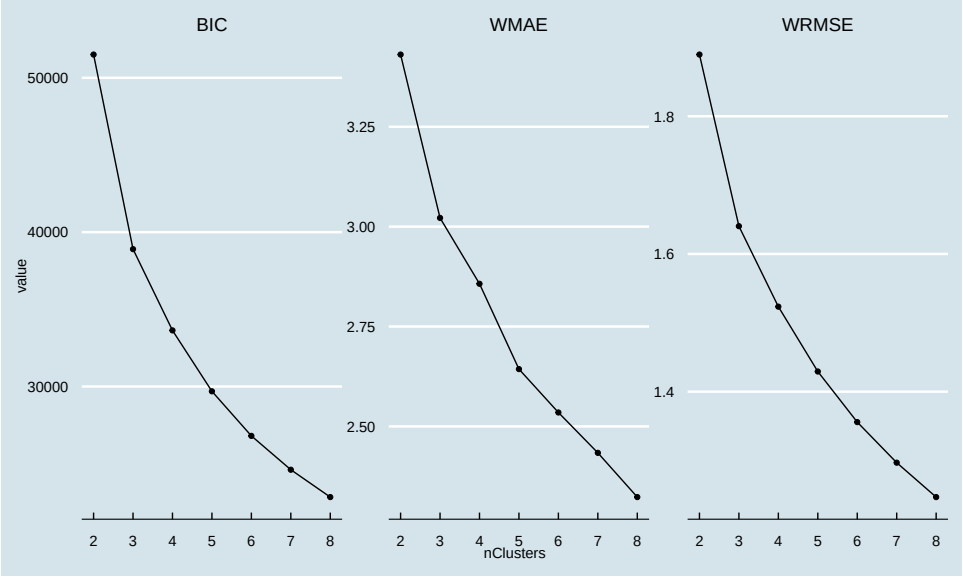
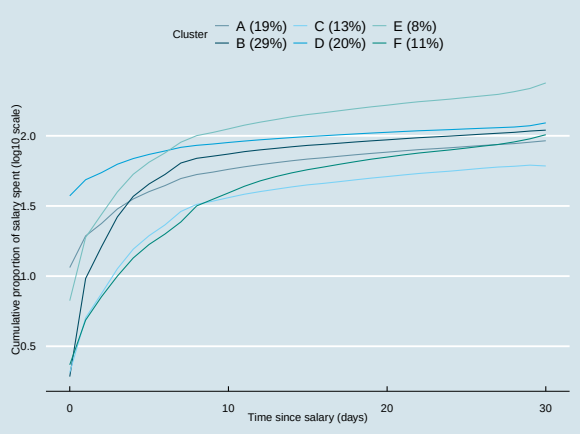
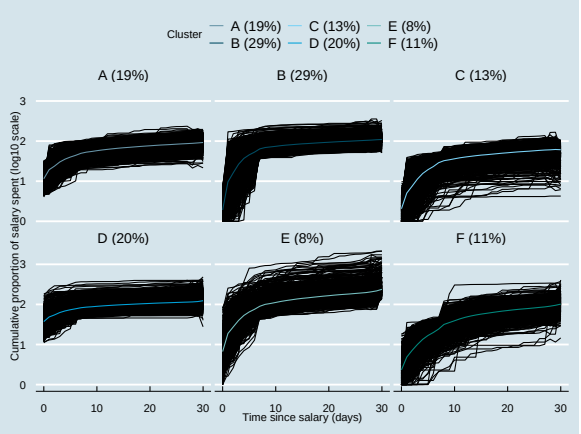


Figure 4.16: Fit metrics for Feature Based Clustering

The resulting mean profiles depicted in Figure 4.17a appear to follow a similar pattern as the KML and GBTM cluster means with an apparent distinguishing factor being the level of overspend at the end of the month. Cluster C appear to be those who underspend while Cluster A are the overspenders. The other clusters tend to end the month spending around as much as they earn with distinguishing factors being how much money is spent at the beginning of the month.



(a) Mean profiles



(b) Mean profiles with the trajectories assigned

Figure 4.17: Final chosen Feature Based Clustering

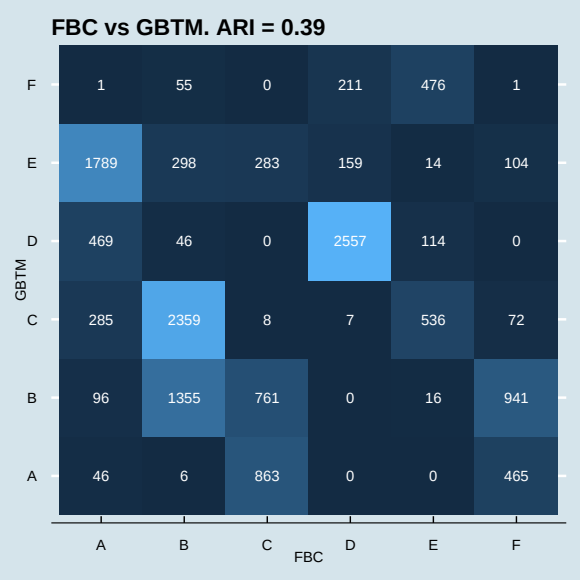
4.3.6 Comparative Analysis of Clustering Techniques

In order to evaluate the effectiveness of the clustering, it is important not just to look at the clustering results in isolation but also to examine the level of agreement between the different clustering methods. In the ideal case, the same users would be clustered together regardless of the method chosen, validating the reliability of any inferences made. That is, if the groups formed are indeed structurally significant, a consistent grouping of users should emerge regardless of the algorithm used.

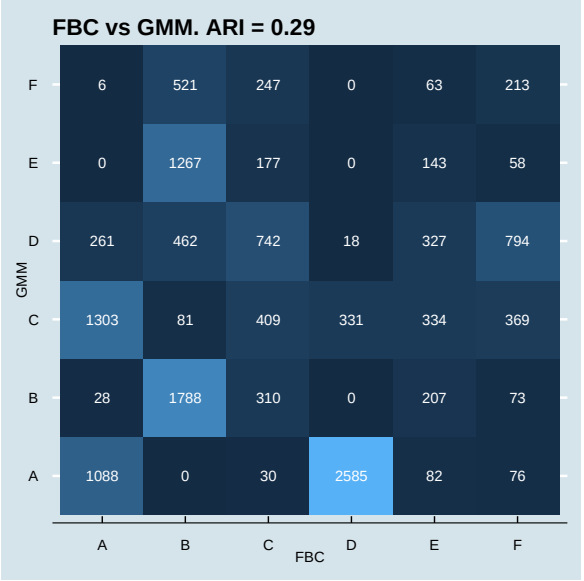
In Figure 4.18, this comparison is illustrated visually by performing pairwise cross tabulations between the different assignments for the different methods. Additionally, the Adjusted Rand Index score is included for each pairwise comparison, providing a quantitative measure of similarity between two clusterings that adjusts for random agreement. A value close to 0 indicates only random agreement, whereas a value of 1 would indicate that the two methods cluster the same users together.

It is noteworthy that the level of agreement does not appear to be very high between many of the clustering methods. In fact, apart from the pair of GBTM and KML, all other ARI values $\leq$ 0.4, indicating weak association. However, it is interesting to observe that the agreement between GBTM and KML stands at 0.92, which is a very high level of agreement. This implies that these two methods are capturing the same underlying dynamics in the spend profiles. It is interesting to note that KML does not explicitly model the time component of the profiles, whereas GBTM does.

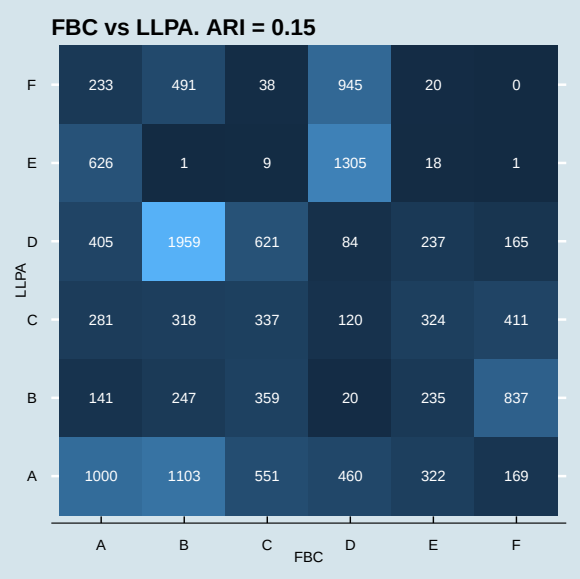
In other words, both the non-parametric KML and parametric GBTM tend to capture the same dynamics.



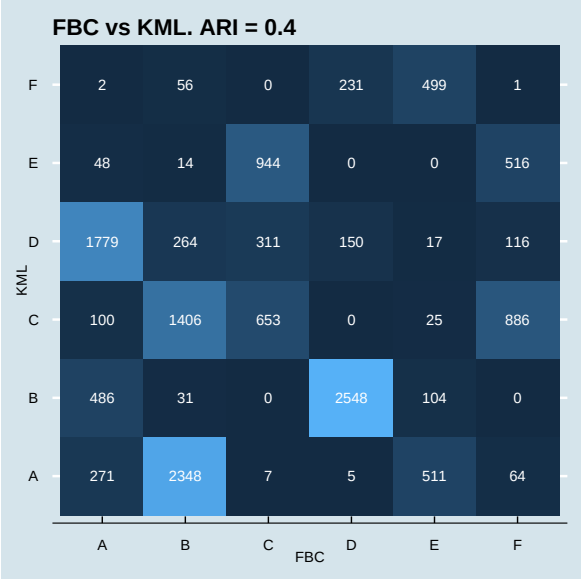
(a) Feature Based Clustering compared to Group Based Trajectory Modelling. (Adjusted Rand Index = 0.39)



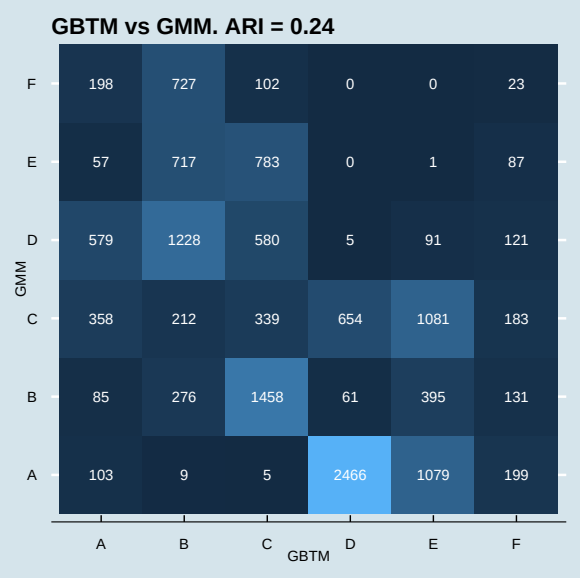
(b) Feature Based Clustering compared to Growth Mixture Modelling (Adjusted Rand Index = 0.29)



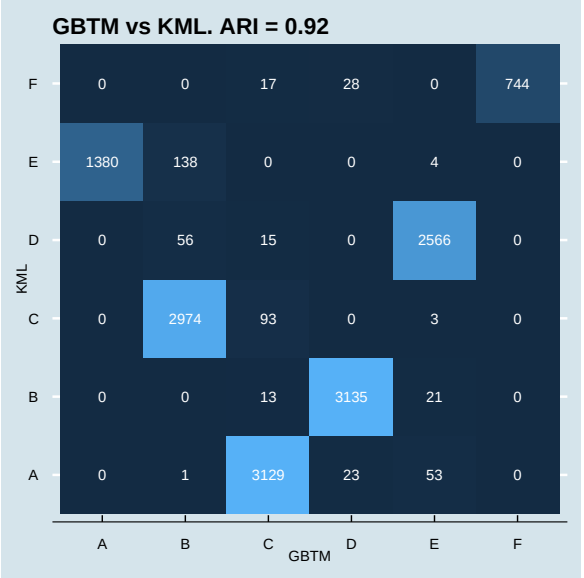
(c) Feature Based Clustering compared to Longitudinal Latent Profile Analysis. (Adjusted Rand Index = 0.15)



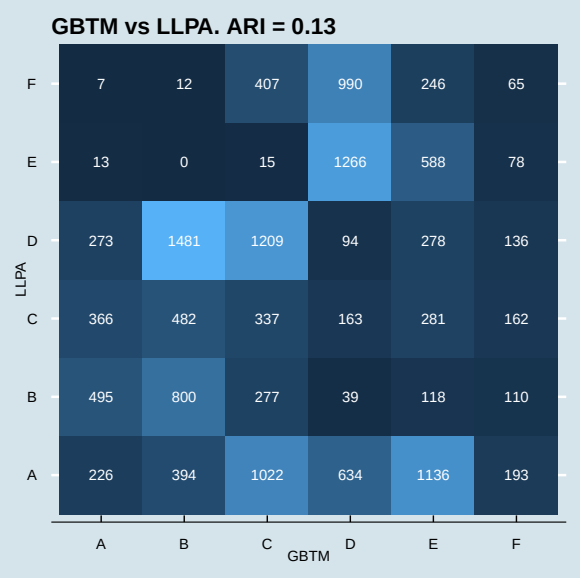
(d) Feature Based Clustering compared to k-means for Longitudinal Data. (Adjusted Rand Index = 0.4)



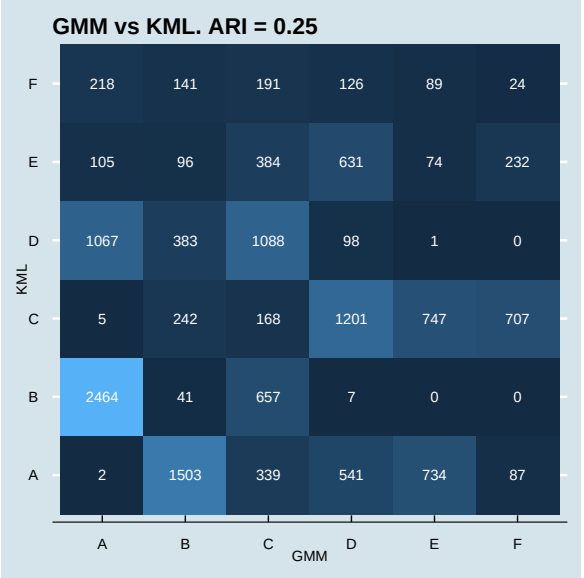
(e) Group Based Trajectory Modelling compared to Growth Mixture Modelling. (Adjusted Rand Index = 0.24)



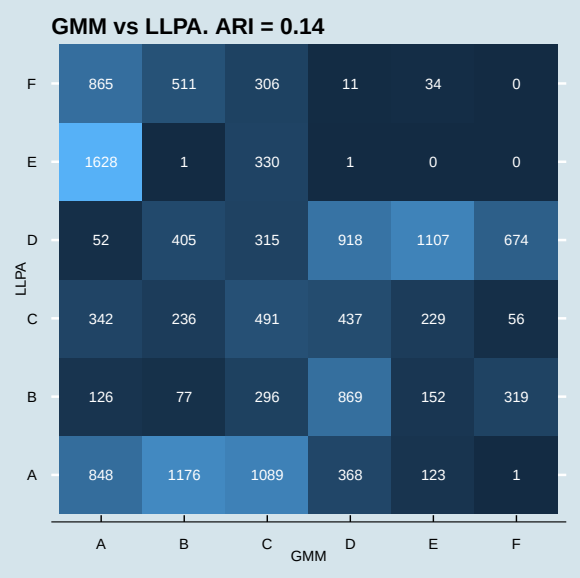
(f) Group Based Trajectory Modelling compared to k-means for Longitudinal Data. (Adjusted Rand Index = 0.92)



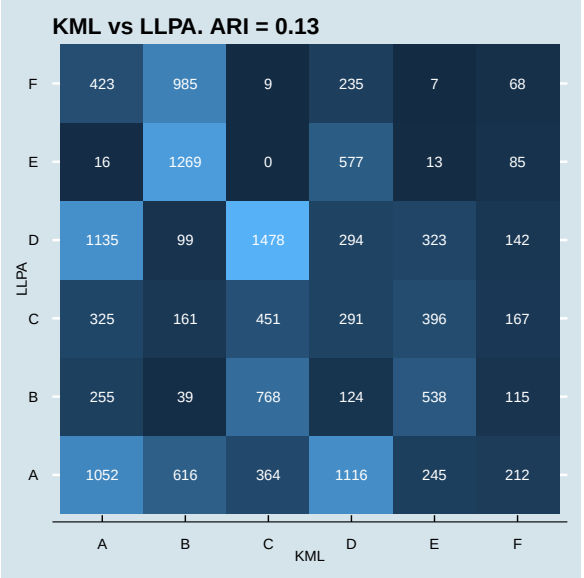
(g) Group Based Trajectory Modelling compared to Longitudinal Latent Profile Analysis. (Adjusted Rand Index = 0.13)



(h) Growth Mixture Modelling compared to k-means for Longitudinal Data. (Adjusted Rand Index = 0.25)



(i) Growth Mixture Modelling compared to Longitudinal Latent Profile Analysis. (Adjusted Rand Index = 0.14)

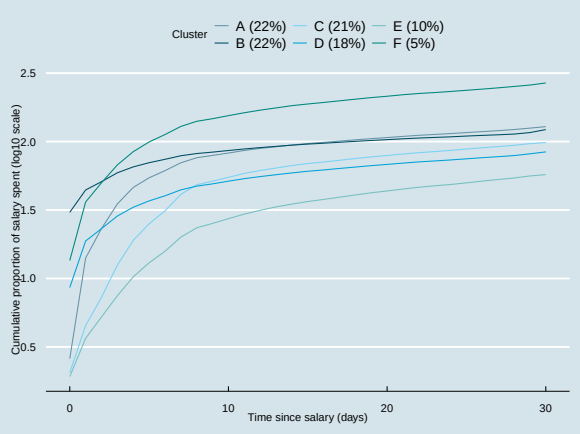


(j) k-means for Longitudinal Data compared to Longitudinal Latent Profile Analysis. (Adjusted Rand Index = 0.13)

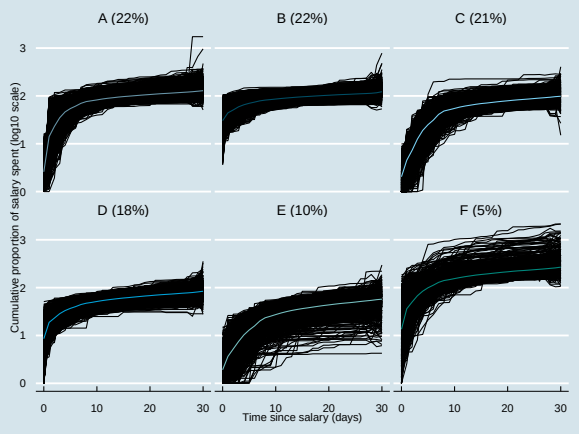
Figure 4.18: Comparison of user assignments to clusters based on different clustering methods

4.4 Covariate Analysis

To delve deeper into the factors influencing clustering outcomes, an exploration of additional customer features—age, income, and expenditure distribution across distinct spending groups—was conducted. This covariate analysis aimed to discern how these demographic and spending-related attributes correlate with the clusters identified by the k-means for Longitudinal Data method. This analysis was primarily performed visually through Figure 4.20, and for ease of reference, the cluster assignments are reproduced in Figure 4.19. Due to the high level of agreement between the KML and GBTM methods, as well as the relatively low level of agreement between the other methods, this examination considers a single clustering as sufficient for this investigation.

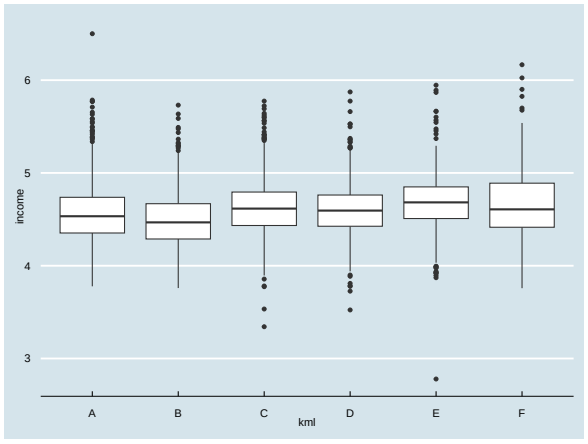


(a) Mean profiles

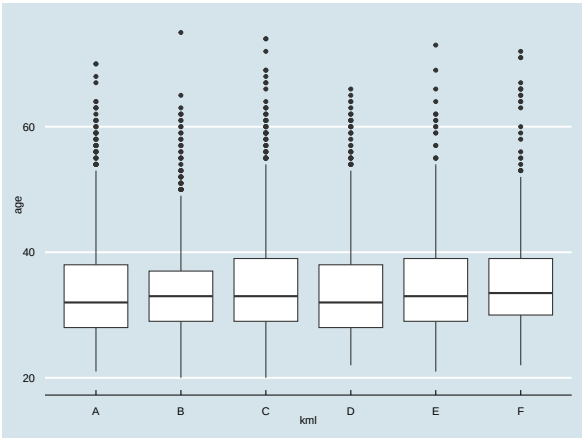


(b) Mean profiles with the trajectories assigned

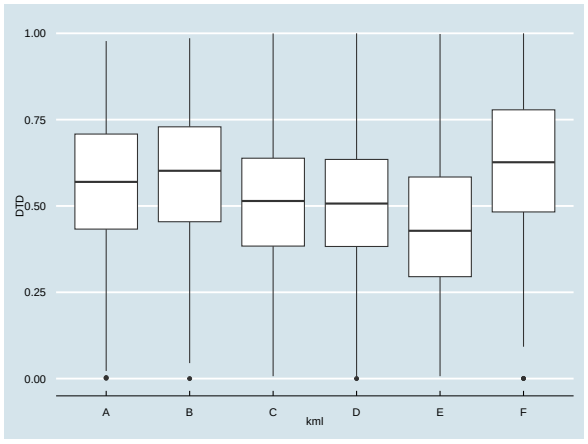
Figure 4.19: Final chosen k-means clustering



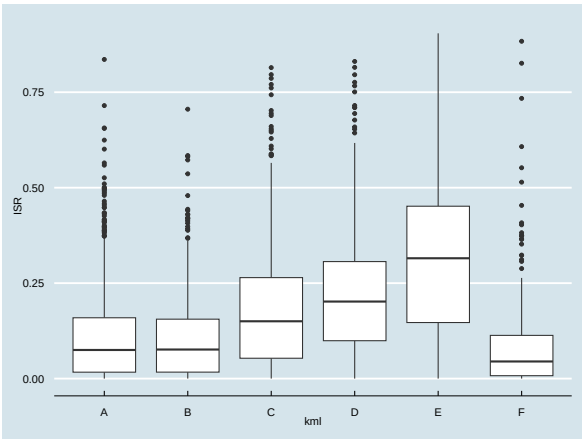
(a) Income (log10 scale) vs KML Cluster Assignments



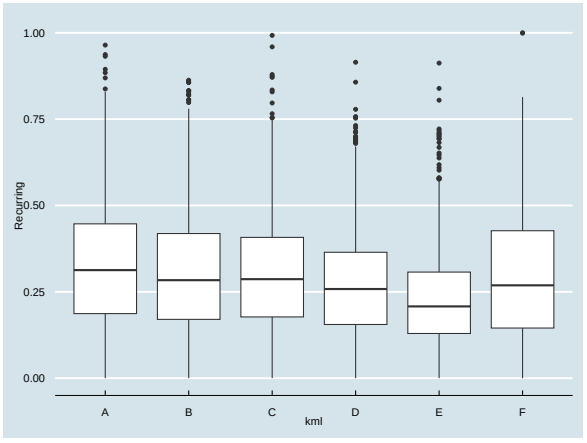
(b) Age vs KML Cluster Assignments



(c) Proportion of spend on Day to Day expenses vs KML Cluster Assignments



(d) Proportion of spend on Invest-Save-Repay expenses vs KML Cluster Assignments



(e) Proportion of spend on Recurring expenses vs KML Cluster Assignments

Figure 4.20: Covariate Analysis of the KML clustering

Demographic Attributes: The examination of age and income distributions across clusters didn't exhibit distinct trends strongly associated with the identified consumer clusters. These attributes seemed less influential in defining spending behavior among the clustered groups.

Spending Composition: The three primary spending groups—Day to Day expenses, Recurring expenses, and Invest-Save-Repay expenses—offered noteworthy insights into the spending dynamics across clusters.

Day to Day Expenses: A discernible pattern emerged showcasing a progressive decrease in the proportion of expenditure on Day to Day expenses from clusters A to E. Conversely, Cluster F displayed a relative increase in this spending category. This trend underscores the association between controlled spending habits and reduced allocation towards variable expenses.

Recurring Expenses: Similar to the Day to Day expenses, a declining trend was observed in Recurring expenses as the analysis moved from clusters A to E, coupled with a relative rise in Cluster F. This trend aligns with the overarching theme of controlled spending behavior among certain clusters.

Invest-Save-Repay (ISR) Expenses: In contrast, the proportion of spending allocated to Invest-Save-Repay expenses demonstrated an inverse relationship with the other spending categories. As the expenditure on Day to Day and Recurring expenses decreased across clusters A to E, Cluster F exhibited a relatively lower proportion of spend in the ISR category. This pattern underscores the capacity to channel surplus funds towards investment or debt repayment among clusters with controlled spending.

The covariate analysis unveiled a compelling narrative emphasizing the pivotal role of spending control in defining consumer clusters. While demographic attributes like age and income didn't exhibit strong associations with spending patterns, the distribution of expenses across distinct categories highlighted a clear correlation between controlled spending behavior and the capacity to allocate surplus funds towards productive financial avenues.

4.5 Conclusion

The analysis of consumer spending behavior through clustering methodologies and covariate examination has unveiled compelling insights into the multifaceted nature of financial habits and their interplay with demographic attributes. Despite encountering varied results across different

clustering techniques, this study discovered intriguing consistencies that shed light on the fundamental drivers of consumer spending patterns.

The utilization of diverse clustering methods highlighted the complexity inherent in longitudinal spending data. While the agreement between methodologies such as k-means for Longitudinal Data and Group Based Trajectory Modeling (GBTM) hinted at common underlying spending dynamics among consumer clusters, the divergence among other methods underscored the intricate and context-dependent nature of consumer behavior analysis.

A pivotal revelation from this study was the significant role played by spending control in distinguishing consumer clusters. Notably, clusters exhibited distinct degrees of overspending by the end of each month, indicating that the ability to manage both fixed and variable expenses significantly influenced spending trajectories. The covariate analysis reinforced this observation, showcasing a correlation between controlled spending, increased investment in productive assets, and the potential for wealth creation.

However, this study merely scratches the surface of a larger, more intricate landscape of consumer financial behavior. The nuances of what constitutes effective spending control, the balance between present consumption and future savings, and the efficacy of budgeting as a tool for managing overspending remain open questions ripe for further exploration.

In the next chapter, further exploration into the efficacy of budgeting in controlling overspending will be conducted.

Chapter 5

Does setting a budget work in reducing spend?

5.1 Introduction

In this chapter, the aim is to answer the titular question by means of analyzing the budget data as previously described. The chapter begins with a simple exploratory analysis of budget vs actual spend data before moving into a longitudinal analysis. The aim of the longitudinal analysis is to compare spend data before the act of setting a budget to spend data after the act of setting a budget. The main question of interest is whether the act of setting a budget resulted in a decrease in spend. This was evaluated by means of both a non-parametric test as well as within the mixed modeling framework. Finally, a comparison is made between budgeters and non-budgeters to evaluate if there is a difference in spending between the two groups. This was again done within the mixed modeling framework, and propensity score matching was conducted to account for possible biases.

5.2 Methods

5.2.1 Linear Mixed Effect Models (LMMs)

Linear Mixed-Effect Models (LMMs) represent a powerful statistical framework for analyzing data with hierarchical or nested structures, where observations are not independent and may exhibit both fixed and random effects. LMMs extend the traditional linear regression model by incorporating random effects, allowing for the modeling of correlated data while accounting for within-group variability.

Consider a dataset with n observations on a response variable Y, which can be expressed as:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \varepsilon_i, \tag{8}$$

where:

- Y_i is the response for the ith observation,
- $\mathbf{X}_i$ is the ith row of the fixed-effects design matrix,
- $\boldsymbol{\beta}$ is the vector of fixed-effects coefficients,
- $\mathbf{Z}_i$ is the ith row of the random-effects design matrix,
- $\mathbf{u}_i$ is the vector of random effects for the ith observation,
- ε_i is the random error term for the ith observation.

The random effects $\mathbf{u}_i$ are assumed to follow a multivariate normal distribution:

$$\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{9}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the random effects. The error terms ε_i are assumed to be independently and identically distributed with mean zero and constant variance σ^2 and are independent of $\mathbf{u}_i$.

Estimation of the fixed-effects coefficients $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$ can be performed using maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML). This typically involves numerical optimization techniques, such as the Newton-Raphson algorithm. The likelihood function for the LMM is expressed as:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{y}) = \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{10}$$

where $f(Y_i \mid \boldsymbol{\beta}, \boldsymbol{\Sigma})$ is the conditional density of Y_i given the model parameters.

Interpreting the results of LMMs involves examining the estimated fixed-effects coefficients $\hat{\boldsymbol{\beta}}$, which represent the average effects of the fixed predictors, as well as the covariance structure captured by $\hat{\boldsymbol{\Sigma}}$, which characterizes the correlation among random effects.

LMMs offer several advantages, including the ability to model nested or clustered data, account for within-group variability, and handle unbalanced designs. They are particularly valuable in longitudinal studies, repeated measures experiments, and any context where data exhibit complex dependencies (Pinheiro and Bates, 2006).

5.2.2 Generalised Linear Mixed Effect Models (GLMMs)

Generalized Linear Mixed-Effect Models (GLMMs) represent an extension of Linear Mixed-Effect Models (LMMs) that accommodate non-normal response variables through the use of link functions. GLMMs combine the flexibility of generalized linear models (GLMs) with the ability to account for correlated and hierarchical data structures, making them a valuable tool in statistical analysis.

Consider a dataset with n observations on a response variable Y, which is not necessarily normally distributed. A GLMM formulates the relationship between the response variable and the predictors as follows:

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i, \tag{11}$$

where:

- $g(\cdot)$ is the link function,
- μ_i is the expected value of Y_i given the model,
- Y_i is the response for the ith observation,
- $\mathbf{X}_i$ is the ith row of the fixed-effects design matrix,
- $\boldsymbol{\beta}$ is the vector of fixed-effects coefficients,
- $\mathbf{Z}_i$ is the ith row of the random-effects design matrix,
- $\mathbf{u}_i$ is the vector of random effects for the ith observation.

The random effects $\mathbf{u}_i$ follow a multivariate normal distribution, as in LMMs:

$$\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{12}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the random effects.

Estimation of the fixed-effects coefficients $\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$ in GLMMs can be challenging due to the nonlinear nature of the link function and the presence of random effects. Various methods, including maximum likelihood estimation (MLE), restricted maximum likelihood estimation (REML), and Bayesian approaches, are employed for parameter estimation.

The likelihood function for GLMMs depends on the specific distributional assumption for the response variable. For instance, in the case of a binary response (e.g., logistic regression), the likelihood function would be based on the binomial distribution. The estimation procedure typically involves numerical optimization techniques to maximize the likelihood.

Interpreting the results of GLMMs involves examining the estimated fixed-effects coefficients $\hat{\boldsymbol{\beta}}$, which represent the effects of the predictors on the transformed scale, and the covariance structure captured by $\hat{\boldsymbol{\Sigma}}$, which characterizes the correlation among random effects.

GLMMs are particularly useful in situations where the response variable is categorical, count-based, or follows a non-normal distribution. They can be applied to a wide range of data types, including binary outcomes, Poisson counts, and more (Stroup, 2012).

5.2.3 Zero-Inflated Generalized Linear Mixed Effect Models (ZIGLMMs)

Zero-Inflated Generalized Linear Mixed-Effect Models (ZIGLMMs) are an extension of GLMMs that address the challenges posed by datasets with excessive zero values. In many real-world scenarios, especially in fields like epidemiology, ecology, and social sciences, count data are collected, and a substantial proportion of these counts are zero. ZIGLMMs are designed to account for this excessive zero-inflation in the data, making them a powerful tool for analyzing such datasets.

The excess zeros in the data can arise from two distinct processes:

1. **Structural Zero:** These are zeros that occur due to the nature of the data, and they would be zeros regardless of the predictor variables. For example, in a study on the number of accidents in a city, a rural area with no traffic lights will always have zero accidents, irrespective of any other factors.

2. **Sampling Zero:** These are zeros that result from the stochastic nature of data collection. For example, in a survey about the number of doctor visits in a year, a person who had no doctor visits during the survey period due to random chance.

ZIGLMMs handle these excess zeros by employing a two-part model. The first part is a logistic regression model that estimates the probability of a zero count, distinguishing between structural and sampling zeros. The second part is a count-based generalized linear mixed-effects model that predicts non-zero counts for the observations.

Mathematically, the model can be expressed as follows:

For the structural zero part:

$$\text{logit}(\pi_i) = \mathbf{X}_i\boldsymbol{\gamma} + \mathbf{Z}_i\mathbf{v}_i, \tag{13}$$

where:

- π_i is the probability of observing a zero count for the i-th observation.
- $\boldsymbol{\gamma}$ is a vector of fixed effects coefficients specific to the structural zero part.
- $\mathbf{v}_i$ is the vector of random effects for the structural zero part.

For the count-based part, you continue to use the GLMM formulation:

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i, \tag{14}$$

where the components are as defined in the GLMM section above.

The overall prediction for the count for the i-th observation is then given by:

$$Y_i = \begin{cases} 0, & \text{with probability } \pi_i \\ \text{Non-zero count}, & \text{with probability } (1-\pi_i) \end{cases} \tag{15}$$

Estimation in ZIGLMMs involves fitting both the logistic and count-based parts simultaneously. The logistic part helps identify and model the excess zeros, while the count-based part models the non-zero counts.

ZIGLMMs are especially useful in fields where count data is common, and there is a clear need to distinguish between structural and sampling zeros. They provide a more accurate and nuanced understanding of the data, which can lead to better insights and predictions. However, fitting ZIGLMMs can be computationally intensive, and model selection and interpretation require careful consideration (Zuur et al., 2009).

5.2.4 Propensity Score Matching

Propensity Score Matching (PSM) is a statistical technique commonly employed in observational studies to reduce bias when estimating treatment effects. PSM aims to balance the covariate dis-

tributions between treatment and control groups, making them more comparable and allowing for a more valid assessment of causal effects (Caliendo and Kopeinig, 2008).

Consider a study where we want to estimate the causal effect of a binary treatment variable, denoted as T, on an outcome variable Y, in the presence of several covariates, denoted as $\mathbf{X}$. The propensity score, denoted as $e(\mathbf{X})$, is defined as the probability of receiving the treatment given the covariates:

$$e(\mathbf{X}) = \Pr(T = 1 \mid \mathbf{X}). \tag{16}$$

The fundamental assumption underlying PSM is the conditional independence assumption, which states that, conditional on the propensity score, treatment assignment is independent of the outcome variable. This assumption is crucial for the validity of PSM.

The PSM procedure typically involves the following steps:

1. **Estimation of Propensity Scores**: Fit a model (usually logistic regression) to estimate the propensity scores $e(\mathbf{X})$ for each observation. The covariates $\mathbf{X}$ are used as predictors in this model.

2. **Matching**: Match treated and control units based on their estimated propensity scores. Several matching algorithms can be used, such as nearest neighbor matching, kernel matching, or propensity score caliper matching.

3. **Assessment of Balance**: After matching, assess the balance of covariates between the treated and control groups using standardized mean differences or other balance diagnostics. The goal is to achieve covariate balance to reduce bias.

4. **Estimation of Treatment Effect**: Estimate the treatment effect by comparing the outcomes of the matched treated and control groups. Common estimators include the difference in means or regression models on the matched dataset.

5. **Inference**: Perform statistical tests or construct confidence intervals to assess the statistical significance of the estimated treatment effect.

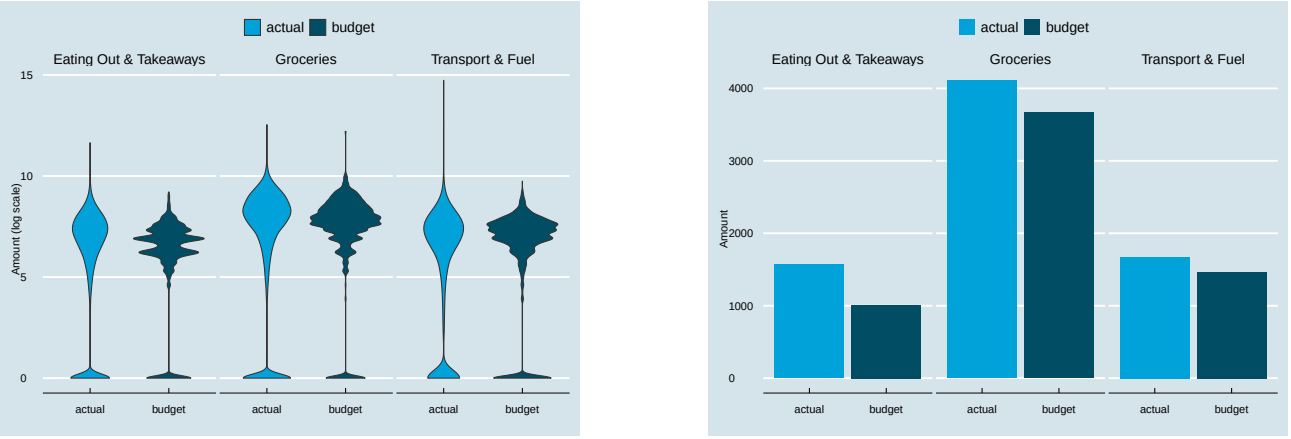
6. **Sensitivity Analysis**: Conduct sensitivity analyses to assess the robustness of the results to potential violations of the conditional independence assumption or different matching methods.

The estimated treatment effect obtained through PSM represents the average causal effect of the treatment on the outcome among the subpopulation of units that are comparable in terms of their propensity scores. It is essential to interpret the results within the context of the matched sample and consider the limitations of the method.

PSM is a valuable tool for making causal inferences in observational studies, especially when randomized controlled trials are not feasible. However, it relies on strong assumptions, such as the conditional independence assumption, which may be challenging to validate in practice. Sensitivity analyses and careful consideration of potential sources of bias are critical aspects of interpreting PSM results (Guo and Fraser, 2014).

5.3 Exploratory Analysis

To gain insights into the correlation between budgeted and actual expenditures in real-world scenarios, an initial examination of the raw data is conducted, as presented in Figure 5.1. The graphical representation in Figure 5.1b illustrates the average values of budgeted and actual expenditures across three distinct categories. Notably, the analysis reveals that, on average, individuals tend to surpass their budgeted expenditure.



(a) Distribution of spend across categories

(b) Mean spend across categories

Figure 5.1: Actual vs Budgeted Spend

Further insights are gleaned from Figure 5.1a, wherein the variability in actual expenditures appears significantly greater than that of budgeted expenditures, as evidenced by the elongated tails in the distribution. Moreover, budgeted expenditures exhibit a discernible "spiky" pattern, imply-

ing that individuals tend to allocate budget amounts in round figures, in contrast to the relatively smoother distribution observed in actual spending patterns. It is also worth noting that the distribution of expenditures, when analyzed on a logarithmic scale, displays a degree of symmetry and approximates a near-normal distribution.

However, a peculiar cluster of observations around zero warrants brief scrutiny. It is imperative to acknowledge that the 22seven dataset is observational in nature. Consequently, data is not reported but rather recorded as transactions are observed within a user's linked accounts. In instances where users do not have any recorded observations in a specific category for a given month, a spending value of R 0 is attributed to that category for that particular month. This practice introduces a minor complexity, as it effectively injects zeros into the dataset instead of treating them as missing data. The distinction between true zeros (indicating that a customer genuinely did not spend any money in a given category) and missing data (suggesting that a customer spent in that category but the transaction data was not recorded) cannot be definitively ascertained due to the inherent limitations of the data. Given this ambiguity, it is determined that, for the purposes of analysis moving forward, these zeros will be considered as true zeros.

The hypothesis that individuals commonly exceed their budgeted expenditures can be empirically investigated through the application of a Linear Mixed Effect Model, formulated as follows:

$$\begin{aligned} Y_{ijt} = \beta_0 &+ \beta_1 \cdot \text{budget}_{\text{Indicator}} + \beta_{2.1} \cdot \text{category}_1 \\ &+ \beta_{2.2} \cdot \text{category}_2 + \beta_3 \cdot \text{budget}_{\text{Indicator}} \cdot \text{category} \\ &+ c_i + m_t + \epsilon_{ijt} \\ &\text{with } c_i \overset{\text{iid}}{\sim} N(0, \sigma_c^2), \text{ and } \epsilon_{ijt} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2) \end{aligned} \tag{17}$$

In this formulation, Y_{ijt} represents the monetary amount for individual i in category j during month t. The binary covariate $\text{budget}_{\text{Indicator}}$ takes the value 1 for actual expenditures and 0 for budgeted amounts. Similarly, $category_1$ and $category_2$ are binary covariates representing different expenditure categories. The terms c_i and m_t denote individual-specific deviations from the overall mean for individual i and month t, respectively. [2]

Of particular interest is the parameter β_1, which provides insights into the magnitude of the disparity between actual and budgeted expenditures and the likelihood of its statistical significance. The

[2] The full regression results are available in Table 6.1 within the Appendix

fitted model yields an estimated value of $\hat{\beta}_1 = -561.24$, accompanied by a standard error of 74.28. Employing Satterthwaite's method, as implemented in the lmerTest package in R, yields a minuscule p-value ($< 1 \times 10^{-13}$) (Kuznetsova et al., 2017). This outcome strongly suggests the presence of a genuine effect, indicating that, on average, individuals tend to exceed their budgeted expenses. It is important to note, however, that this conclusion should be interpreted with caution due to the non-normally distributed nature of the data, potentially violating some of the assumptions underlying the model. Nevertheless, this finding offers valuable insights and warrants consideration, although further investigation may be required to fully assess the magnitude of this effect. For the present analysis, abstaining from pursuing additional exploration is recommended, as the evidence appears sufficiently robust for the intended purposes.

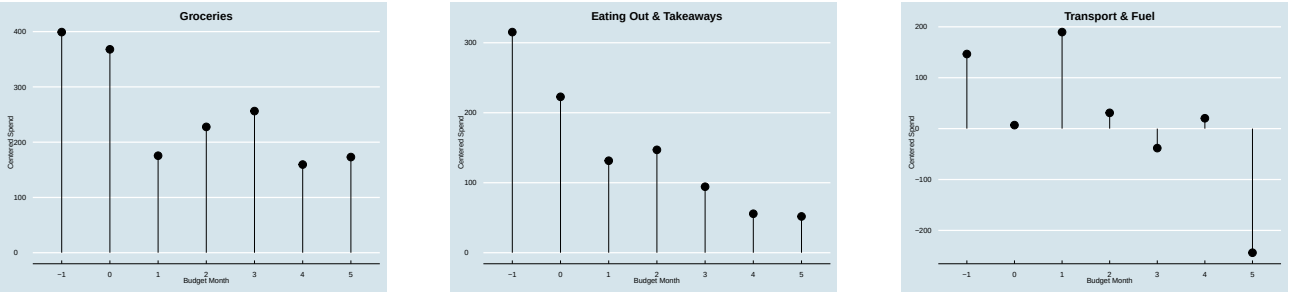
5.4 Longitudinal Analysis

As highlighted in the preceding section, it is apparent that individuals frequently exceed their planned budgets. This raises the pertinent question of whether the act of setting a budget is, in essence, ineffective. While this may seem counterintuitive, it necessitates a more thorough investigation. The hypothesis posits that, despite individuals often surpassing their budgetary constraints, the mere act of establishing a budget can still contribute to expenditure reduction.

To empirically examine this hypothesis, the analysis focuses on users' spending behaviors during two distinct time periods: the month immediately prior to setting their initial budget and the five months following the budget's initiation. However, it is important to acknowledge the inherent challenge posed by the fact that users set their budgets at various points in time. For instance, User A may have initiated their first budget in June 2020, while User B might have done so as late as June 2022. Given the fluctuating prices and economic conditions over this two-year period, a straightforward comparison of raw expenditure values would be inappropriate. To mitigate the impact of price variability, all expenditures are normalized by aligning them with their respective calendar months. This approach minimizes the influence of price fluctuations, providing a more reliable measure of expenditure trends over time. From this point, the normalized expenditure is referred to as centered spend.

Figure 5.2 displays the average expenditures by customers during what is termed the "Budget Month." This designation corresponds to a timeframe relative to the initial budget setting. Specifically, Month 0 denotes the month in which a budget was first established, Month -1 represents the month immediately prior to budget initiation, and Month 1 signifies the month immediately fol-

lowing the creation of a budget. Notably, the analysis reveals compelling evidence suggesting that setting a budget indeed contributes to expenditure reduction, a trend observed across all three categories investigated.



(a) Groceries (b) Eating Out & Takeaways (c) Transport & Fuel

Figure 5.2: Actual vs Budgeted Spend

The assertion can be more rigorously tested by means of the Wilcoxon Signed Rank Sum test. The choice to employ this non-parametric alternative to a paired t-test is made as the data exhibit patterns that deviate from the normality assumption. A one-sided test for a significant decrease in spend in the month prior to setting a budget compared to the month post setting a budget as well as five months following the setting of a budget is conducted. Formally, the hypotheses are as laid out below, and the same two tests are performed for each of the different categories. Results are summarized in Table 5.1.

Test 1:

$$\begin{aligned} H_0 &: \text{The median difference between a user's spending before setting a budget} \\ &\quad \text{and 1 month after setting a budget is equal to 0} \\ H_A &: \text{The median difference between a user's spending before setting a budget} \\ &\quad \text{and 1 month after setting a budget is greater than 0} \end{aligned} \tag{18}$$

Test 2:

$$\begin{aligned} H_0 &: \text{The median difference between a user's spending before setting a budget} \\ &\quad \text{and 5 months after setting a budget is equal to 0} \\ H_A &: \text{The median difference between a user's spending before setting a budget} \\ &\quad \text{and 5 months after setting a budget is greater than 0} \end{aligned} \tag{19}$$

	pre vs post	pre vs 5 months post
Groceries	0.0002	0.0002
Eating Out & Takeaways	0.0017	<1e-4
Transport & Fuel	0.0341	<1e-4

Table 5.1: p-values for the Wilcoxon Signed Rank Sum tests specified in Equations 18 and 19

The results of the analysis seem to indicate that setting a budget has an almost immediate impact in reducing one's spend in the Groceries and Eating Out & Takeaways categories but not in the Transport & Fuel category. A proposed explanation for this is the fact that in the South African market, consumers are price takers when it comes to transport costs as fuel prices are dictated by the national government and are standard across the options that consumers have to choose from. Consumers therefore have very little choice and are less in control of their own spend in this category.

A further piece of evidence on the effectiveness of setting a budget on reducing one's spend is provided by means of a broken stick regression. This is done naively by placing a knot at budget month 1 and fitting a linear mixed effects model with the dependent variable as the centered spend. The fixed effect is simply the budget month, and the slope is allowed to vary after budget month 1. A random intercept term is included for both the customer as well as the category. The described linear mixed effects model can be represented as follows:

Let Y_{ij} be the centered spend for the i-th customer in the j-th category and budget month, where $i = 1,2,\ldots,n$ (number of customers) and $j = 1,2,\ldots,m$ (number of categories).

The model can be expressed as:

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{BudgetMonth}_{ij} + \beta_2 \cdot \text{BudgetMonth}_{ij} \cdot \text{AfterMonth1}_{ij} + u_{0i} + u_{1j} + \varepsilon_{ij} \tag{20}$$

where β_0 is the overall intercept, β_1 is the fixed effect for the budget month, β_2 is the interaction term for the budget month and an indicator variable AfterMonth1_{ij} that takes the value 1 if the budget month is greater than 1, and 0 otherwise. u_{0i} represents the random intercept for the i-th customer, which accounts for individual customer-specific variability, u_{1j} represents the random

intercept for the j-th category, which accounts for category-specific variability, and ε_{ij} is the random error term. Results of fitting this model are shown in Table 5.2.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	96.146	83.784	34.095	1.148	0.259
BudgetMonth	-121.469	61.834	13911.342	-1.964	0.050
I(BudgetMonth * AfterMonth1)	96.519	68.137	13908.819	1.417	0.157

Table 5.2: Estimated coefficients for the broken stick regression as described in Equation 20

The outcomes derived from the broken stick regression analysis provide compelling evidence that establishing a budget has a significant and immediate impact on reducing expenditures. This assertion is supported by the negative sign of the estimated coefficient $\hat{\beta}_1$ along with the notably low p-value associated with it. While the positive sign of the $\hat{\beta}_2$ coefficient suggests a subsequent upward trend in spending, it is important to note that this effect is not particularly robust, as evidenced by its relatively high p-value. Consequently, it can be confidently asserted that the act of budget setting leads to a short-term reduction in spending, and remarkably, this effect persists over an extended period, demonstrating its persistence over time.

5.5 Causal Analysis

In this section, a direct investigation into the central question posed by this chapter is undertaken: "Does setting a budget work in reducing spend?". However, before delving into the analysis, it is crucial to acknowledge and reflect upon the inherent observational nature of the dataset. The decision to establish a budget is entirely driven by the individual, introducing the possibility of confounding variables that may influence this choice and thereby introduce bias into the analysis.

In an ideal research scenario, there would be the luxury of randomly assigning consumers to two groups: a treatment group (comprising those who set a budget) and a control group (consisting of individuals who do not set a budget). In such a setup, effective randomization would facilitate a straightforward comparison of outcomes between the two groups. Regrettably, the experiment does not align with this ideal, as individuals voluntarily choose whether or not to set a budget. To mitigate the impact of potential confounders and enhance the validity of the findings, Propensity Score Matching is employed. This technique allows for effective control of the influence of confounding variables, thereby bolstering the reliability of the research outcomes.

After completing the Propensity Score Matching procedure to mitigate confounding influences, the subsequent step in the analysis entails the application of a Generalized Linear Mixed Model (GLMM). This statistical framework becomes particularly advantageous in this study due to the non-normal distribution of the data, a characteristic that aligns with the reality of expenditure patterns. GLMMs are well-suited for such scenarios where data do not adhere to a normal distribution and exhibit correlation among observations from the same individuals over time. By harnessing the power of GLMMs, the relationship between setting a budget and its impact on expenditure is effectively modeled, accounting for the complex dependencies inherent in longitudinal data. This approach enhances the precision of the estimations regarding the causal effect of budget setting, enabling more robust and accurate conclusions while accommodating the unique data distribution characteristics.

5.5.1 Propensity Score Matching

To execute Propensity Score Matching, users are initially categorized into two distinct groups: those who have opted to establish a budget and those who have not. Subsequently, a probit Generalized Linear Model (GLM) is employed to compute propensity scores by aligning users based on key attributes such as age, ethnicity, gender, and income. The outcomes of this matching process are presented in Figure 5.3, affording the opportunity to assess the efficacy of the matching strategy. Pre matching, it is noticeable that the data are likely biased, particularly with respect to age and ethnicity; however, the use of PSM matching reduces the bias of these confounders significantly. Notably, a substantial reduction in the overall distance score is evident, and when scrutinizing the matched variables, an Absolute Standardized Mean Difference of less than 0.05 is observed across the board. This outcome signifies the effectiveness of the matching approach in reducing the biases inherent in the observational study, enhancing the robustness of the analysis.

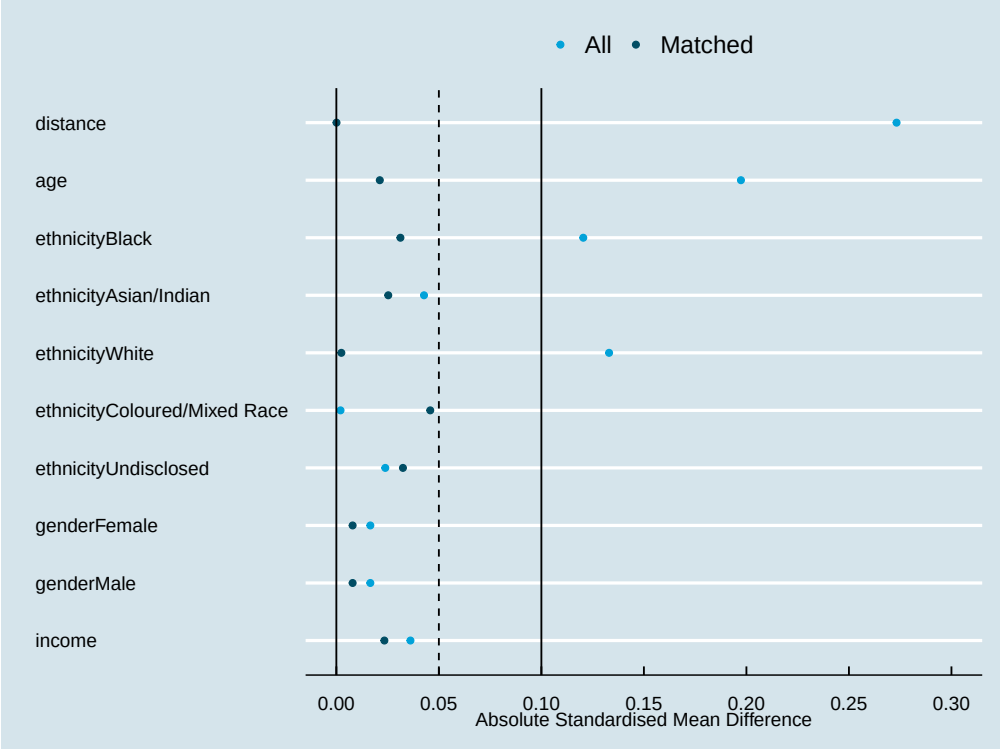


Figure 5.3: Results of performing propensity score matching

5.5.2 Generalised Linear Mixed Effect Model

The propensity scores obtained above are then used as weights to weight the observations for the GLMM fitted. As noted in Figure 5.1, a cluster of zeros is observed, indicating zero inflation in the data. The model specified must therefore account for this while also addressing the non-normality in the data. A two-part model is required to properly account for this. The implementation of these two parts, the zero-inflated part and the GLMM with a Gamma link function part, is described in more detail below.

1. Zero-Inflated Part:

Let Y_{ijkl} represent the count of the spend for customer i in year j, month k, and category t. The zero-inflated part models the probability of observing zero spend ($Y_{ijkl} = 0$) as a function of a Bernoulli process with probability π_{ijkl}.

The logistic regression part for the zero-inflated model is as follows:

$$\begin{aligned}\text{logit}(\pi_{ijkl}) = \beta_0 &+ \beta_1 \text{income}_i + \beta_2 \text{age}_i + \beta_3 \text{gender}_i \\ &+ \beta_4 \text{budgeted\_amount}_{ijkl} + \beta_5 \text{set\_budget}_{ijkl} \\ &+ u_i + v_{jkl} \end{aligned} \tag{21}$$

Where:

- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the fixed-effect coefficients.
- $\text{income}_i, \text{age}_i, \text{gender}_i, \text{budgeted\_amount}_{ijkl}, \text{set\_budget}_{ijkl}$ are the covariates for income, age, gender, budgeted amount, and budget tracking for customer i in year j, month k, and category l. Note that both income and budgeted_amount are recorded on a log scale.
- u_i represents the random effect of customers, assumed to follow a normal distribution with mean zero and some variance.
- v_{jkl} represents the random effect of the nested year/month/category effect, assumed to follow a normal distribution with mean zero and some variance.

The fitted zero-inflated part of the model is given in Table 6.2 within the appendix.

2. Gamma GLMM Part:

Assuming that non-zero spend ($Y_{ijkl} > 0$) follows a Gamma distribution, the conditional distribution of non-zero spend given that it's non-zero is modeled using the Gamma GLMM. The conditional mean of spend is modeled as:

$$\begin{aligned}E(Y_{ijkl} | Y_{ijkl} > 0) = \mu_{ijkl} = \exp(&\alpha_0 + \alpha_1 \text{income}_i + \alpha_2 \text{age}_i + \alpha_3 \text{gender}_i \\ &+ \alpha_4 \text{budgeted\_amount}_{ijkl} + \alpha_5 \text{set\_budget}_{ijkl} \\ &+ \gamma_i + \delta_{jkl}) \end{aligned} \tag{22}$$

Where:

- $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$ are the fixed-effect coefficients for the Gamma part.
- $\text{income}_i, \text{age}_i, \text{gender}_i, \text{budgeted\_amount}_{ijkl}, \text{set\_budget}_{ijkl}$ are the covariates for income, age, gender, budgeted amount, and budget tracking for customer i in year j, month k, and category l. Note that both income and budgeted_amount are recorded on a log scale.

- γ_{ij} represents the random effect of customers, assumed to follow a normal distribution with mean zero and some variance.
- δ_{ijk} represents the random effect of the nested year/month/category effect, assumed to follow a normal distribution with mean zero and some variance.

The fitted conditional Gamma part of the model is given in Table 5.3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.204	0.077	80.465	0.000
income	0.072	0.001	56.146	0.000
age	0.015	0.001	10.245	0.000
gender_{male}	0.137	0.032	4.303	0.000
budgeted_amount	0.077	0.002	41.439	0.000
set_budget	-0.483	0.014	-33.303	0.000

Table 5.3: Estimated conditional gamma part of the GLMM as specified in Equation 22

The key coefficient of interest in the table above is the one related to the "set_budget" variable. It's worth noting that the estimated coefficient is -0.483, which implies that when people set a budget, their spending tends to decrease by a factor of approximately $exp(-0.483) = 0.62$. This reduction corresponds to a substantial 38% decrease in average spending for those who have a budget compared to those who don't.

When examining the other factors, it can be interpreted in a similar way. Both income and age are positively associated with spending, meaning that people with higher incomes and older individuals tend to spend more. Additionally, males appear to spend more than females.

However, it should be noted that there are some data limitations:

- **Incomplete Financial Account Linkage**: It can't be certain if users have linked all their financial accounts, which means that some of their spending data may be missing from the analysis.
- **Sample Selection Bias**: There's a possibility of bias in the sample because the data are limited to users of 22seven which may not be broadly representative of people as a whole.
- **Unaccounted Confounders**: There's no data on certain app usage behaviors like login fre-

quency or budget tracking activity. Without this information, it's impossible to control for their potential impact on spending.

Even though every effort was made to minimize bias and control for confounding factors using the available data, it's still possible that there are hidden variables and influences that weren't measured. Therefore, any conclusions drawn in this chapter should be taken with a degree of caution, and further research is needed to explore these associations in more depth.

5.6 Conclusion

In this chapter, an in-depth exploration was undertaken to understand the intricate relationship between budget setting and individual spending behaviors using data derived from the 22seven platform. Through an array of statistical models and methodologies, the goal was to discern whether the act of setting a budget correlates with observable changes in expenditure patterns.

The investigation revealed compelling evidence suggesting that establishing a budget significantly impacts expenditure habits. The analysis, particularly the Generalized Linear Mixed-Effect Model (GLMM), underscored a noteworthy finding: individuals who set a budget exhibited a substantial reduction in spending, approximately a 38% decrease on average compared to those who did not establish a budget.

However, it's imperative to acknowledge the complexities inherent in observational data analysis. Despite meticulous efforts to mitigate biases and confounding factors through propensity score matching and rigorous statistical modeling, certain limitations persist. Incomplete financial account linkage and potential sample selection bias within the dataset might influence the robustness of the conclusions.

Moreover, while the findings suggest a strong association between budget setting and reduced spending, causality cannot be definitively established. Unmeasured variables and unaccounted behaviors within the app, such as login frequency or budget tracking activity, might exert an influence that the analysis did not capture.

Therefore, while the results of this analysis shed light on the potential efficacy of budget setting in curbing expenditures, caution should be exercised in drawing absolute conclusions. Further research, encompassing a more comprehensive dataset and accounting for additional behavioral nuances, is essential to corroborate and deepen the understanding of the relationship between budgeting and spending habits.

This chapter serves as a significant step toward unraveling the dynamics of budget setting and its impact on individual financial behaviors. However, it also highlights the complexity of human financial decision-making, calling for continued exploration and refinement of methodologies to better grasp the nuanced interplay between budget setting and expenditure patterns.

Chapter 6

Discussion and Conclusions

The comprehensive analysis conducted across the chapters of this dissertation provides a multifaceted understanding of consumer spending behavior and the impact of budget setting on expenditure patterns. Through the use of diverse clustering methodologies and statistical models, this research has revealed significant insights while also highlighting the intricacies and complexities inherent in studying financial habits.

The exploration of consumer spending behavior through clustering methodologies illuminated the multifaceted nature of financial habits and their interplay with demographic attributes. While various clustering techniques showcased divergent results, certain consistencies emerged, indicating underlying drivers of consumer spending patterns. Notably, the study identified the pivotal role of spending control in distinguishing consumer clusters, emphasizing its influence on managing both fixed and variable expenses and its correlation with potential wealth creation.

Subsequently, the investigation into the relationship between budget setting and individual spending behaviors demonstrated a substantial impact of establishing a budget on expenditure habits. The analysis, particularly through the Generalized Linear Mixed-Effect Model (GLMM), indicated a considerable reduction in spending among individuals who set a budget compared to those who did not, marking a notable average decrease of approximately 38

However, it is essential to acknowledge the complexities and limitations encountered in observational data analysis. Despite rigorous efforts to mitigate biases and confounding factors, inherent limitations such as incomplete financial account linkage and potential sample selection bias within the dataset persist. The inability to establish causality and account for unmeasured variables and behaviors within the app underscores the need for caution in drawing absolute conclusions from these findings.

While the results suggest a strong association between budget setting and reduced spending, further research is imperative to validate and expand upon these findings. A more comprehensive dataset, accounting for additional behavioral nuances, and refining methodologies will be crucial in deepening our understanding of the intricate relationship between budgeting and expenditure patterns.

In conclusion, this dissertation represents a significant step towards unraveling the dynamics of consumer financial behavior and the impact of budget setting. The complexity of human financial decision-making calls for continued exploration and refinement of methodologies to gain a more nuanced understanding of the interplay between budget setting and individual expenditure habits. This study paves the way for future research to delve deeper into the nuances of effective spending control, the balance between present consumption and future savings, and the broader efficacy of budgeting in managing overspend.

Therefore, while the findings presented herein offer valuable insights, they serve as a starting point for further inquiry and refinement in comprehending the intricate landscape of consumer financial behavior.

References

22seven (2023). 22seven home page. https://www.22seven.com. [Accessed 03-01-2023].

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological bulletin*, 82(4):463.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Bengtsson, N. (2012). The marginal propensity to earn and consume out of unearned income: Evidence using an unusually large cash grant reform. *The Scandinavian Journal of Economics*, 114(4):1393–1413.

Browne, R. (2023). From banking giants to lending up-and-comers — here are the world's top 200 fintech companies — cnbc.com. https://www.cnbc.com/2023/08/02/here-are-the-worlds-top-200-fintechs-cnbc-and-statista.html. [Accessed 16-10-2023].

Buehler, R., Griffin, D., and Peetz, J. (2010). The planning fallacy: Cognitive, motivational, and social origins. In *Advances in experimental social psychology*, volume 43, pages 1–62. Elsevier.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72.

Clogg, C. C. (1995). Latent class models. In *Handbook of statistical modeling for the social and behavioral sciences*, pages 311–359. Springer.

Ezenkwu, C. P., 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 (IJARAI)*, 4(10):40–44.

Formann, A. K. and Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical methods in medical research*, 5(2):179–211.

Genolini, C. and Falissard, B. (2010). Kml: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American psychologist*, 54(7):493.

Guo, S. and Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications*, volume 11. SAGE publications.

Gupta, M. K. and Chandra, P. (2019). A comparative study of clustering algorithms. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 801–805. IEEE.

Heath, C. and Soll, J. B. (1996). Mental budgeting and consumer decisions. *Journal of consumer research*, 23(1):40–52.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112, chapter 12.4.1, pages 515–519. Springer.

Kahneman, D. and Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Liu, Z., Liu, R., Zhang, Y., Zhang, R., Liang, L., Wang, Y., Wei, Y., Zhu, R., and Wang, F. (2021). Latent class analysis of depression and anxiety among medical students during covid-19 epidemic. *BMC psychiatry*, 21:1–10.

Lukas, M. F. and Howard, R. C. 2023). The influence of budgets on consumer spending. *Journal of Consumer Research*, 49(5):697–720.

Magidson, J. and Vermunt, J. (2002). Latent class models for clustering: A comparison with k-means. *Canadian journal of marketing research*, 20(1):36–43.

Mutual, O. (2023). Old Mutual Savings & Investment Monitor 2023. https://www.oldmutual.co.za/savingsmonitor/. [Accessed 25-11-2023].

Ng, H., Ong, S., Foong, K., Goh, P.-S., and Nowinski, W. (2006). Medical image segmentation using k-means clustering and improved watershed algorithm. In *2006 IEEE southwest symposium on image analysis and interpretation*, pages 61–65. IEEE.

Novemsky, N. and Kahneman, D. (2005). The boundaries of loss aversion. *Journal of Marketing research*, 42(2):119–128.

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. *Modern statistical methods for HCI*, pages 275–287.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184. Springer.

Stilley, K. M., Inman, J. J., and Wakefield, K. L. (2010). Planning to make unplanned purchases? the role of in-store slack in budget deviation. *Journal of consumer research*, 37(2):264–278.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Sussman, A. B. and Alter, A. L. (2012). The exception is the rule: Underestimating and overspending on exceptional expenses. *Journal of Consumer Research*, 39(4):800–814.

Teuling, N. D., Pauws, S., and Heuvel, E. v. d. (2021). Clustering of longitudinal data: A tutorial on a variety of approaches. *arXiv preprint arXiv:2111.05469*.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science*, 4(3):199–214.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Ülkümen, G., Thomas, M., and Morwitz, V. G. (2008). Will i spend more in 12 months or a year? the effect of ease of estimation and confidence on budget estimates. *Journal of Consumer Research*, 35(2):245–256.

Wielechowski, M., Cherevyk, D., Czech, K., Kotyza, P., Grzęda, Ł., and Smutka, L. (2021). Interdependence between human capital determinants and economic development: K-means regional clustering approach for czechia and poland. *Entrepreneurial Business and Economics Review*, 9(4):173–194.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., Smith, G. M., Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data. *Mixed effects models and extensions in ecology with R*, pages 261–293.

Appendix

Additional Results for Chapter 5

Tables

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1353.734	186.044	6.190	7.276	0.000
$\text{budget}_{indicator}$	-561.242	74.278	131941.365	-7.556	0.000
$\text{category}_{Groceries}$	2561.307	75.135	132688.662	34.090	0.000
$\text{category}_{Transport\&Fuel}$	143.579	77.720	132709.952	1.847	0.065
$\text{budget}_{indicator}$ * $\text{category}_{Groceries}$	119.180	104.490	131941.365	1.141	0.254
$\text{budget}_{indicator}$ * $\text{category}_{Transport\&Fuel}$	352.341	107.887	131941.365	3.266	0.001

Table 6.1: Estimated Linear Mixed Effects Model as specified in Equation 17

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.235	0.235	9.527	0.000
income	-0.495	0.003	-176.547	0.000
age	0.016	0.004	3.859	0.000
gender_{Male}	-0.215	0.088	-2.435	0.015
budgeted_amount	-0.240	0.006	-41.597	0.000
set_budget	1.411	0.043	33.171	0.000

Table 6.2: Estimated Zero-inflated part of the GLMM as specified in Equation 21