Analysis

Distinguishing payment user groups by clustering transaction data

February, 2024

Jelmer Reijerink

DeNederlandscheBank
EUROSYSTEEM

Distinguishing Payment User Groups by Clustering Transaction Data

Author: Jelmer Reijerink. With thanks to colleagues at DNB, and in particular Marie-Claire Broekhoff, Carin van der Cruijsen and Ellen van der Woerd for the useful exchange of views. All remaining errors are mine.

De Nederlandsche Bank n.v.
P.O. Box 98 1000 AB
Amsterdam
Internet: www.dnb.nl
Email: info@dnb.nl

Summary

- Using payment diary surveys from Dutch consumers, we identify five distinct payment user groups, each exhibiting different demographic characteristics and payment behaviours. We describe these groups as: Family-Centric Middle-Income Consumers, Senior Traditional Banking Users, High-Income Urban Consumers, Financially Challenged Consumers and Young & Low Value Purchase Consumers.
- In order to be able to identify these groups of individuals, who share similar socio-demographic characteristics and payment behaviour patterns, we perform a cluster analysis - a popular technique in machine learning.
- To see if we can generalise our findings, we use euro area data to perform a similar analysis outside of the Netherlands. This analysis reveals the following five groups: Small Household Seniors, Low-Income City Residents, Rural Families, Affluent Online Consumers, and Urban Families. This indicates that, despite significant variations in demographic characteristics and payment behaviour among individuals across countries, we can still identify common consumer groups.
- Understanding different user groups provides researchers and policymakers with valuable insights into the diverse factors influencing payment choices, enabling the adaptation of strategies that meet the specific needs and preferences of each segment.

1 Introduction

Consumers are increasingly moving away from using cash as a payment method at the point-of-sale (POS). In the entire euro area, cash accounted for 59% of the value spent at POS locations in 2022, marking a 12 percentage point decline from 2016 (European Central Bank, 2022). In the Netherlands, where only 15% of all euros spent at POS locations in 2022 was paid in cash (De Nederlandsche Bank & Dutch Payments Association, 2023), the use of cash is at an all-time low. Despite the broad adoption of newer payment methods like payment cards and mobile phones, certain consumers continue to use cash as a means of payment, each to varying degrees.

Existing literature on payment behaviour has extensively explored various factors influencing how consumers make transactions. First of all, transaction details, such as purchase price and transaction costs, significantly impact payment choices. Studies reveal that cash usage is lower for more expensive purchases, and consumers tend to avoid payment methods with high transaction costs (Jonker, 2007; Klee, 2008; European Central Bank, 2022; De Nederlandsche Bank & Dutch Payments Association, 2023). Payment behaviour is also closely related to standard personal characteristics, with cash usage showing strong correlations with demographics across multiple

countries. For instance, lower-income, older, and less-educated consumers tend to use more cash for their transactions (Bagnall et al., 2016; Arango-Arango et al., 2018; van der Cruijsen and Plooij, 2018). Additionally, first-generation migrants from cash-oriented countries are more likely to use cash in the Netherlands (Kosse and Jansen, 2013). Furthermore, the characteristics of the region in which individuals reside play a role, as the share of debit card transactions tends to be higher in urbanised regions (van der Cruijsen and Knoben, 2021). Cash is of particular importance for vulnerable groups, such as those with low digital literacy, certain physical disabilities, mild intellectual disabilities, and financial difficulties (van der Cruijsen and Reijerink, 2023). The collective findings from these studies shed light on the complex nature of payment behaviour, underscoring the relevance of contextual factors in shaping individuals' payment preferences and highlighting the diverse influences on payment choices.

Prior studies have primarily focused on analysing various individual factors influencing payment behaviour using standard regression models. However, these studies have not addressed the identification of different groups of consumers, thereby failing to provide a comprehensive overview of the diverse user types within the payments domain. Previous research that aimed to segment various consumer groups (e.g., Dutch Payments Association, 2021) primarily focused on examining consumers' values and attitudes regarding payments rather than emphasising other demographic factors. Understanding different user types can provide valuable insights into which groups rely more on cash and are more vulnerable to changes in cash availability or acceptance. Moreover, this knowledge contributes to a comprehensive understanding of payment behaviour, which can be helpful when making decisions regarding the future role of cash in the payment landscape. (Dutch Payments Association, 2021)

To address this crucial gap in the literature and to gain a deeper understanding of different consumer groups and their characteristics, we leverage machine learning techniques that allow us to define distinct user groups. Machine learning methods have found wide application in the payment domain, particularly in fraud detection, identifying suspicious payment transactions and credit scoring (Doerr et al., 2021). Central banks often use these techniques to enhance data quality, arrive at richer contextual insights, and gather more comprehensive information (Bank for International Settlements, 2022). Despite the limited presence of machine learning applications in payment diary data analysis, these techniques can be useful in understanding consumer payment choice, as highlighted by Shy (2020). Particularly, the combination of regression analysis and machine learning techniques could strengthen the reliability of algorithms describing consumer payment choice, as they complement each other effectively.

Utilising payment diary data collected in the Netherlands, this study conducts a cluster analysis based on transaction details and individuals' personal characteristics to unveil comprehensive patterns. A cluster analysis is a popular machine learning technique that is useful for unsupervised data analysis and pattern recognition. This means that this technique can be applied to datasets where labels or categories are unknown prior to the analysis, such as ours. Using cluster validation statistics, we find

that the optimal number of clusters is five. We employ a k-means cluster analysis with $k = 5$ to reveal these five clusters, each characterised by specific demographic and payment-related features. Cluster 1 (Family-Centric Middle-Income Consumers, 32% of respondents) comprises individuals with average age, income, and education levels, exhibiting a high usage of mobile banking apps and low reliance on cash. Cluster 2 (Senior Traditional Banking Users, 29% of respondents) consists of older individuals with slightly lower incomes, preferring traditional internet banking and with a high use of cash (28% of total purchase value). Cluster 3 (High-Income Urban Consumers, 20% of respondents) represents high-income, highly educated individuals, who extensively use both internet and mobile banking, with the least reliance on cash. Cluster 4 (Financially Challenged Consumers, 12% of respondents) includes individuals facing financial difficulties, with lower incomes and a relatively high usage of cash. Lastly, cluster 5 (Young & Low Value Purchase Consumers, 8% of respondents) consists of younger individuals with an average age of 18, predominantly using mobile banking apps, making infrequent and inexpensive transactions, and using cash relatively often. In summary, by performing a k-means cluster analysis we identify five statistically distinct groups within the set of respondents, characterised by their payment behaviour and demographic traits.

Additionally, we complement the analysis of Dutch survey data by running a similar cluster analysis on data from other euro area countries. This extended analysis indicates that we can also define distinct payment user groups when considering a broader European context. We define the following clusters: Small Household Seniors, Low-Income City Residents, Rural Families, Affluent Online Consumers, and Urban Families. From the euro area data it appears that countries with higher cash usage comprise a larger portion of individuals categorised as Low-Income City Residents or Urban Families, and a smaller portion as Affluent Online Consumers. This understanding of the clusters provides researchers and policymakers with valuable insights into the diverse factors influencing payment choices, enabling the adaptation of strategies that meet the specific needs and preferences of each segment.

Our study is structured as follows. Section 2 describes the Dutch data that we use in this study, and provides an explanation of the data preprocessing and cleaning steps. Section 3 outlines the steps we take to perform the cluster analysis by providing explanations for variable selection and deciding on the clustering technique and parameters. Section 4 offers the results on the Dutch dataset, and we rerun a similar analysis on the euro area data in Section 5. We conclude in Section 6.

2 Data

We use payment survey data collected from Dutch consumers by the Ipsos research agency on behalf of the De Nederlandsche Bank (DNB) and the Dutch Payments Association (DPA). The main goal of this Survey on Consumers' Payments (SCP) is to gain insight into payment behaviour in the Netherlands at the POS. We refer to Jonker et al. (2018) for a comprehensive overview of the survey details. Each day, roughly 65 Dutch consumers aged 12 years and over registered all the

payments they made during this day. For each transaction, respondents indicated the amount paid, the payment instrument that was used for the payment, and the industry the payment was made in. Additionally, respondents answered questions about payment preferences and personal characteristics. All data is weighted and validated on a yearly basis[1]. As some of the variables we are interested in were not part of the SCP prior to June 2020, the data used in this analysis spans the period from June 2020 until December 2022. As we aim to find specific groups for users of cash at the POS, we are interested in respondents that actually made a purchase on the reporting day. This is why we exclude respondents that did not make a purchase.

Next, we select a specific subsample of the dataset. Respondents of the SCP can participate in the survey at most once per quarter. Potentially, respondents could have participated in the SCP up to 10 times within the observed timeframe. Even though the majority of the respondents (around 61%) only participated once, we ensured that only one observation per respondent was randomly selected. Limiting the selection to a single reporting day per respondent ensures that each participant's contribution is represented equally and avoids potential bias resulting from multiple responses from the same individual. By selecting these observation randomly, we ensure that they are distributed evenly across time. After this selection, our dataset contains 26,123 observations. The variables of interest for this study are presented in Table A.1 and Table A.2 in Appendix A.

We perform a number of data preprocessing and cleaning steps to handle missing values, address inconsistencies and improve the quality of our data, as this usually enables machine learning algorithms to make more accurate predictions and reliable decisions. All data preprocessing and cleaning steps are explained in detail in Appendix B, and result in a dataset of 25,862 observations. We present an overview of the summary statistics of the resulting dataset in Table C.1 in Appendix C to provide insight into the characteristics of the dataset and to get a general feel for the data. As mentioned before, the respondents are a good representation of the Dutch population, so the summary statistics are in line with what we would expect. For example, the average number of purchases is 2.13 and the average purchase value is €26.75.

3 Cluster analysis

Clustering, or grouping data, is a machine learning technique used in various fields to uncover hidden patterns and structures within datasets. The primary goal of clustering is to identify similarities and dissimilarities among data points and organise them into distinct clusters based on their shared characteristics. By doing so, clustering helps us gain valuable insights, such as segmenting customers into different user groups.

[1]The Ipsos research agency ensures that the survey is filled in by a sample that is representative for the Dutch population. The total value of debit card transactions at the POS is corrected based on actual debit card transactions reported to the DPA.

3.1 Dimension reduction

Before we can perform the cluster analysis on our dataset, we need to decide on the selection of variables in our analysis. We choose to incorporate all the demographic and payment-related variables that were presented in Appendix C. By including this diverse set of variables, we can gain valuable insights into the cash usage patterns of different groups and how these patterns are influenced by demographic characteristics. This approach enables us to identify distinct user segments based on spending behaviour, payment preferences, and demographics, providing a more comprehensive understanding of payment user groups and their different determinants.

In our analysis, we encounter some challenges related to the large number of variables we have used and the diverse types of data they represent. Traditional cluster analysis methods rely on distance measures. However, these methods struggle when dealing with such a wide range of variables (Aggarwal et al., 2001). Additionally, most clustering algorithms are designed for numerical data and may not handle mixed data types well, like the combination of categorical and continuous variables we have used. Furthermore, these algorithms are sensitive to the scale of the data, and our variables are not scaled. One key reason for not scaling the variables is that scaling can result in a loss of valuable information. This is why we employ a technique called dimensionality reduction to overcome the aforementioned challenges. The idea behind this method is to simplify our data while preserving its important characteristics. Even though dimensionality reduction can also introduce its own level of information loss, this transformation does allow us to reveal hidden patterns and structures within the data. By using the extracted factors instead of the original variables, we can achieve more informative representations of the data, leading to improved clustering outcomes and potentially avoiding overfitting. We choose to adopt a Factor Analysis of Mixed Data (FAMD) as our dimensionality reduction method. Full details of this analysis and the steps that were executed can be found in Appendix D.

3.2 Clustering technique and parameters

In order to cluster the observations, we first need to select a clustering technique. In this analysis we work with unlabelled data - data without any defined categories or groups prior to the clustering analysis. Therefore, we have to use an unsupervised learning technique if we want to cluster the data. In this study we use the k-means algorithm, an algorithm that is widely used due to its simplicity and efficiency. This algorithm involves several steps. First, a number of k centres is chosen. Each k centre represents the centre of a cluster (a centroid). Next, data points are iteratively assigned to one of these centroids, based on the similarity to the features that are provided. The features we use are the five dimensions that have been calculated using the dimension reduction method described in Section 3.1. Data point similarity is calculated using Euclidean distances. This is a popular distance measure that is suitable for datasets where variables have numerical values; it simply calculates the straight-line distance between two points. After assigning

all data points to a centroid, the algorithm then recalculates the centroids and reassigns data points until no further reassignments occur. The k-means algorithm offers several advantages, including its straightforward implementation and scalability to large datasets. The algorithm also outperforms other unsupervised methods such as hierarchical clustering in terms of computational speed.

Nonetheless, there are certain limitations to the k-means algorithm that we need to address. First, the presence of outliers significantly impacts the performance of k-means; however, we have mitigated this concern by employing various data preprocessing steps (Appendix B). Furthermore, this method may not be ideal for datasets with a high number of variables, but we have already addressed this issue through the prior execution of FAMD (Section 3.1). Another challenge lies in selecting an appropriate value for k, the number of clusters, as this decision can be subjective. To help us in identifying an optimal number of clusters, we make use of validation statistics. We present these validation statistics and their meaning in Appendix E. Both of the chosen validation statistics suggest that choosing five clusters offers the best balance between granularity and meaningfulness in segmenting the data, indicating an appropriate k-value of five for the k-means clustering analysis. This is why we proceed with the clustering analysis utilising five clusters.

4 Results Dutch data

In this section we present the results of the k-means clustering analysis by first comparing the sizes of the clusters and looking at the cash usage per cluster. Next, we calculate the mean values of all variables per cluster in order to see how the clusters differ from one another. Finally, we perform an exploratory analysis of the results in order to gain a more intuitive understanding of the distinct clusters.

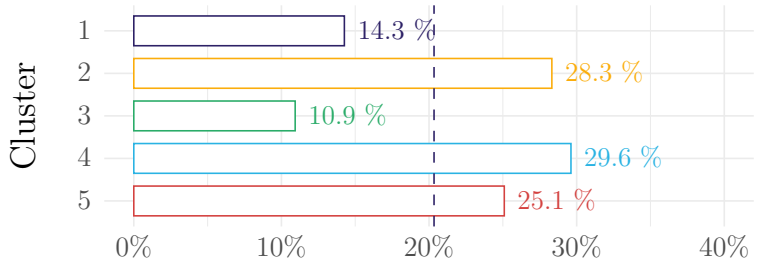
4.1 Cluster sizes and cash usage per cluster

Clustering results from a k-means cluster analysis using $k = 5$ are presented in Table 1, providing information on the size of each cluster. Notably, the first three clusters represent the majority, making up for 80% of all observations. Although the last two clusters are relatively smaller in size, their identification suggests that they possess enough distinctive characteristics from the other clusters.

As explained (Section 2), we calculate the average share of the purchase value that respondents paid in cash. To help us identify five distinct payment user groups among the respondents, we also calculate this per cluster separately (presented in Figure 1). The average of the share of cash usage across all respondents is 20%. However, cash usage varies significantly across clusters. Cluster 2 (28%), cluster 4 (30%) and cluster 5 (25%) have higher-than-average cash usage, while cluster 1 (14%) and cluster 3 (11%) have relatively low cash usage.

Table 1 Size of the clusters (NL data)

	Cluster	Number of observations	% of observations
□	1	8,202	31.71
○	2	7,455	28.83
△	3	5,101	19.72
⊠	4	3,102	11.99
⊕	5	2,002	7.74



Average share of purchase value paid in cash
The dotted line represents the average across all clusters

Figure 1 Respondents' average share of their purchase value paid in cash, per cluster

4.2 Variable means per cluster

To look more closely into the differences per cluster, we calculate the mean values of all variables for each cluster. These mean values are presented in Table E.1 in Appendix E. The table provides insights into the specific features that define each cluster. We also investigate the robustness of the reported findings when subjected to modification in the models. This is essential to enhance the credibility and reliability of our findings. The results of these additional robustness checks are presented in Appendix F.[2]

Next, we calculate each variable's deviation from its mean value for all clusters separately. Subsequently, by plotting these deviations we facilitate a more comprehensive presentation of the distinctions among the individual clusters. This allows us to get a more intuitive understanding of the results. We present these outcomes in Figure 2. The figure highlights that most clusters have significant deviations for certain variables. For instance, we find that respondents in cluster 5 exhibit an average age (18 years) that is almost 65% lower than the overall average age (51 years). Or, along similar lines, respondents within cluster 4 report an average level of financial difficulty that is almost 50% higher than the average of the complete dataset, thereby emphasising the significant differences in this particular cluster in terms of financial difficulty.

[2]Additionally, we find that the distribution of the weekdays is nearly uniform across clusters, indicating no bias towards specific purchase days within any particular cluster.

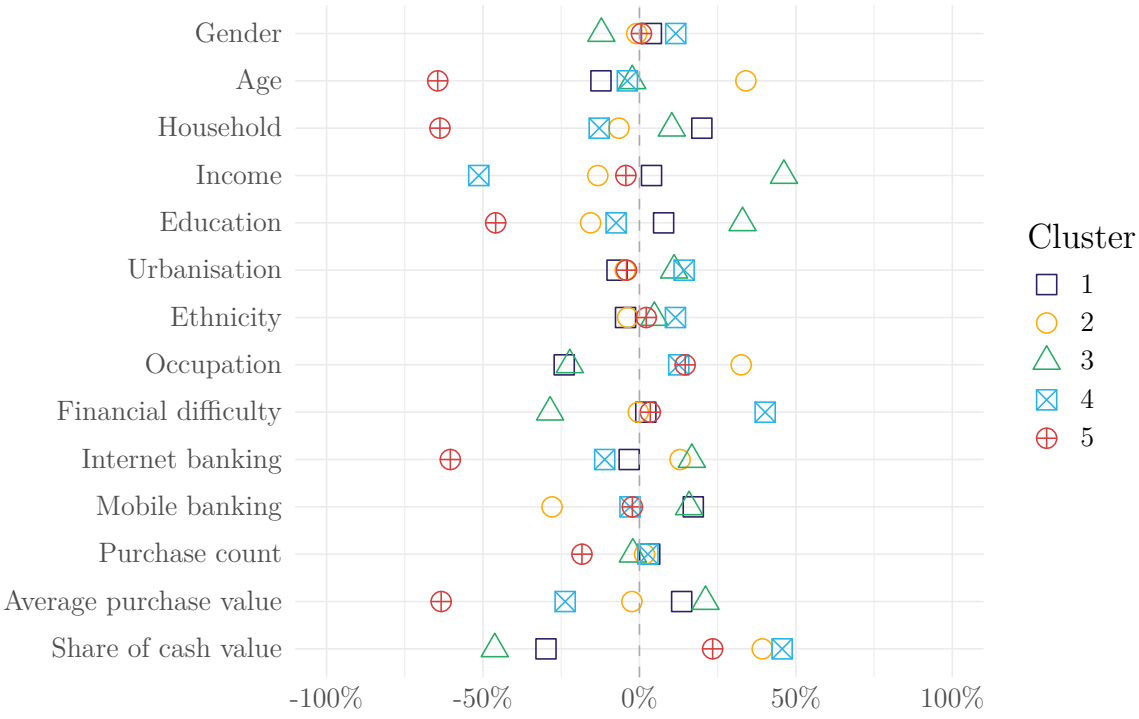


Figure 2 Variables' deviation from the mean value, per cluster

Following the visualisation with the distinct features per cluster, we perform a profiling analysis to interpret the clusters. By comparing the profiles of the different clusters, we can identify key variables that differentiate them. These differentiating variables provide insights into the specific demographics and behaviours that define each cluster. Based on the variable differences, we aim to describe each cluster in terms of their demographic and payment-related characteristics. We provide the following descriptions:

□ **1 - Family-Centric Middle-Income Consumers (32% of respondents)**

This group contains individuals with average age, income and education levels. Compared to the other clusters, this group more frequently includes households with children, and has a higher proportion of employed individuals. Additionally, this group demonstrates a high usage of mobile banking apps compared to others. Their average purchase value is higher than the overall average and substantially higher than most of the other clusters. Remarkably, they rely much less on cash for transactions, with only 14% of their total purchase value paid in cash.

○ **2 - Senior Traditional Banking Users (29% of respondents)**

This group stands out for its notably older age distribution compared to other clusters. Individuals in this group have slightly lower incomes and education levels. Additionally, this group has a relatively low proportion of employed individuals compared to the other clusters. These individuals tend to use traditional internet banking more frequently than mobile banking apps. 28% of the total purchase value of these individuals is paid in cash, which is relatively high.

△ **3 - High-Income Urban Consumers (20% of respondents)**

This cluster represents the group with the highest income and education levels among all clusters. Individuals in this cluster have the most stable employment status and make the highest average purchase value. Both internet banking and mobile banking usage are substantially higher in this group. Moreover, they are more likely to reside in urban areas. This cluster has the lowest share of cash usage, accounting for only 11% of their total purchase value.

☒ **4 - Financially Challenged Consumers (12% of respondents)**

This group comprises individuals who have relatively low incomes and face financial difficulties, making it challenging for them to make ends meet. They have slightly below-average education levels. This group is less likely to have kids and tends to reside in urban areas. The average purchase values are quite low for this cluster. Cash usage is high compared to the other clusters and the average, with 30% of their purchase value conducted in cash.

⊕ **5 - Young & Low Value Purchase Consumers (8% of respondents)**

Individuals in this group are much younger, with an average age of 18, and many of these individuals live with their parents. Their education levels are relatively low, likely due to ongoing studies. They primarily rely on mobile banking apps, possibly reflecting a generational mobile-first trend. The frequency of their purchases is lower than the overall average, and the transactions they make are relatively inexpensive. 25% of their total purchase value is paid in cash, which is relatively high.

5 Results of cluster analysis using euro area data

Rather than solely focusing on finding distinct groups of payment users in the Netherlands, we are also interested in exploring data from other euro area countries to determine if we can generalise our findings more broadly. In order to do so, we use data that is collected for the Study on the Payment Attitudes of Consumers in the Euro area (SPACE) provided by the European Central Bank (2022). This study looks at the behaviour and preferences of consumers relating to cash, card and other available payment methods. The data was collected directly from around 50,000 euro area consumers but does not include observations from Germany and the Netherlands, because in those countries data was collected using own, country-specific, surveys. The questionnaire that was used by the ECB had a high degree of similarity compared to the questionnaire that was used in the Netherlands, which means that the most important demographic and payment variables can also be used to perform a cluster analysis using the euro area data, and these findings can be compared to the situation in the Netherlands.

We first clean our dataset using an approach similar to the process we applied for the data from the Netherlands, as described in Appendix B. Next, we employ a dimension reduction (Factor Analysis of Mixed Data), similar to the analysis explained in Appendix D.[3] The resulting dataset contains 32,754 observations. Rather than a priori finding the optimal number of clusters through various validation statistics, we aim to identify five distinct clusters immediately.[4] This approach is chosen to assess alignment with clusters discovered in the Dutch data.

We present the outcomes of this cluster analysis below. We start by looking at cluster sizes and cash usage per cluster, just as we did with the Dutch dataset. Then, we calculate the average values per cluster for the variables utilised in our analysis. We finalise by comparing the results per country.

Table 2 Size of the clusters (euro area data)

	Cluster	Number of observations	% of observations
+	1	8,356	25.51
×	2	6,557	20.02
◇	3	6,402	19.55
▽	4	6,381	19.48
✳	5	5,058	15.44

[3] The appropriate number of dimensions that we use to reduce the euro area data is five, similar to the Dutch data.

[4] If we were to determine an ideal number of clusters beforehand, the cluster validation statistics (outlined in Appendix E for Dutch data) would not conclusively identify the best number of clusters.

5.1 Cluster sizes and cash usage per cluster

Cluster sizes are presented in Table 2. It appears that the first four clusters are more or less equal in terms of size, and the fifth cluster is somewhat smaller. Despite being the smallest cluster, the fifth cluster still accounts for approximately one-sixth of all respondents.

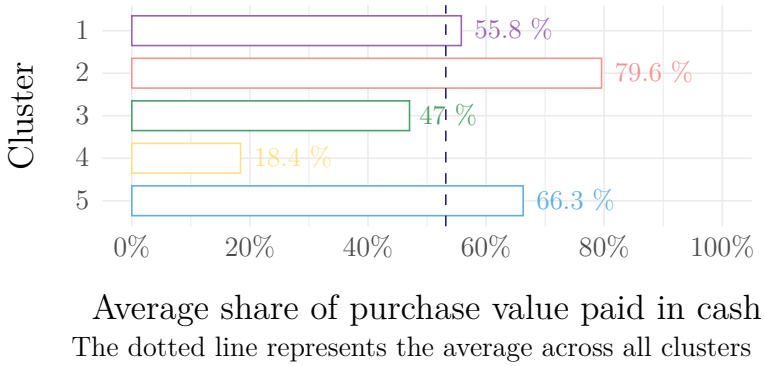


Figure 3 Respondents' average share of their purchase value paid in cash, per cluster

Next, we find that cash usage varies across the different clusters, as presented in Figure 3. On average, individuals spent 53% of their total purchase value using cash. Cash usage is highest in the second and fifth cluster, with respectively 80% and 66% of the purchase value paid in cash. Cash usage is lowest in the fourth cluster, with only 18% of the value paid in cash.

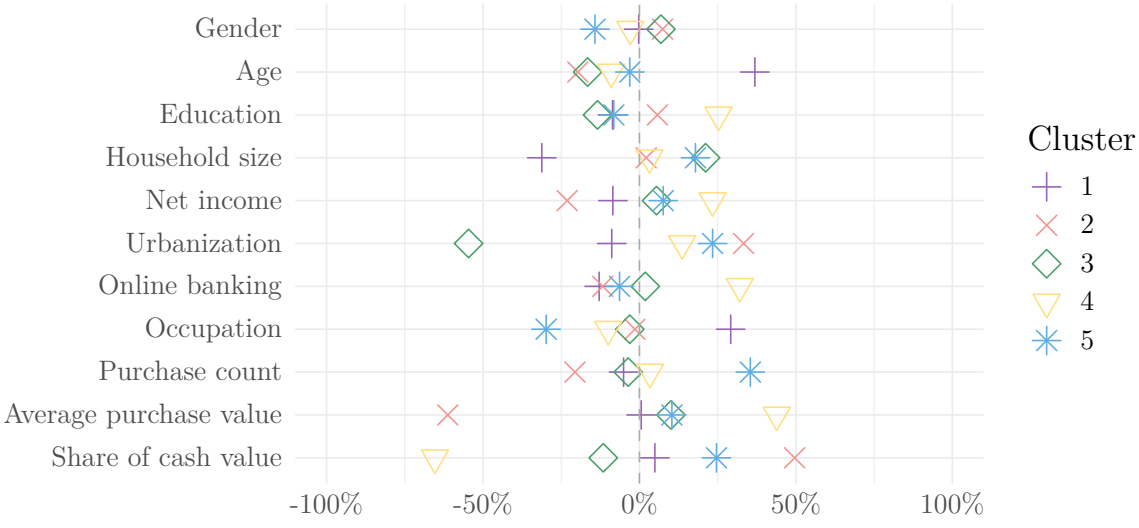


Figure 4 Variables' deviation from the mean value, per cluster

5.2 Variable means per cluster

We also want to have a better understanding of the differences between the clusters for the euro area data, which is why we calculated the means per cluster for all variables used (see Table G.1 in Appendix G). Again, for interpretation purposes, we plotted the deviations from the mean values for each variable, per cluster. The results of these calculations are presented in Figure 4. Based on this figure, we provide a profiling analysis of the five clusters:

+ 1 - Small Household Seniors (26% of respondents)

This cluster is characterised by the oldest individuals (average age of 66.4) with moderate education levels and income. These households are typically small in size, and they are more often unemployed. Their cash usage is slightly above average, with 56% of their total purchase value paid in cash.

× 2 - Low-Income City Residents (23% of respondents)

This group represents individuals who are most likely to live in an urban area. They have the lowest average income among all clusters, which also explains their low average purchase size. The cash usage of this cluster is highest of all clusters, accounting for 80% of their total purchase value.

◇ 3 - Rural Families (21% of respondents)

This cluster includes individuals that are the least likely to live in an urban area. Their education levels are slightly lower than average, and their household size is largest of all clusters. This cluster shows an average share of cash usage, with 47% of the total purchase value.

▽ 4 - Affluent Online Consumers (19% of respondents)

Individuals in this group have a much higher education level than average. Their income levels are high, and they are more likely to live in an urban area. They use online banking services more often than individuals in the other clusters, and the purchases these individuals make are the most expensive of all clusters. The average share of cash usage is low, with only 18% of the total purchase value paid in cash.

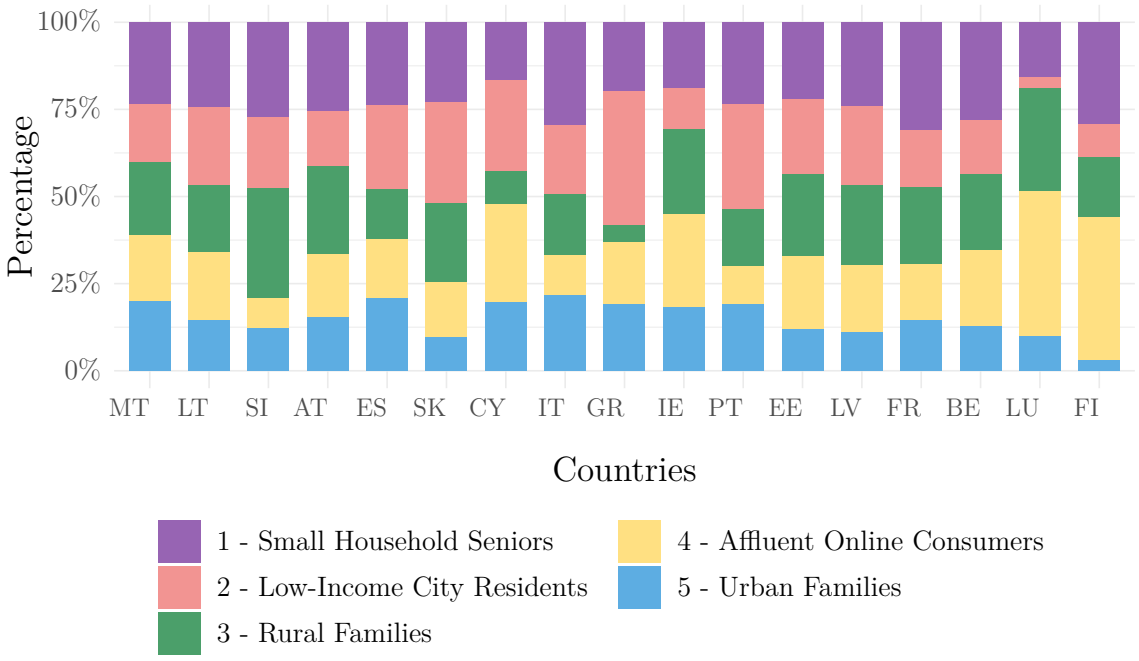
✳ 5 - Urban Families (11% of respondents)

This group comprises individuals who have a large household size, and who more often reside in urban areas. While income levels are average, the amount of purchases is highest among all clusters. These individuals still rely on cash for a significant share of their total purchase value (66%).

5.3 Results per country

Now that we have defined and interpreted the different clusters within the euro area dataset, we can compare the relative sizes of the clusters per country. We present this visual comparison in Figure 5. The countries in the figure are arranged based on their usage of cash in POS transactions. Countries on the left-hand side of the figure (such as Malta, Lithuania) represent a high usage of cash at the POS, whereas countries on the right-hand side (such as Luxembourg, Finland) exhibit low shares of cash in POS transactions.

Overall, we find that cluster 1 is the largest for most countries. Also, there does not seem to be a trend between a country's cash usage and the relative size of cluster 1. This indicates that there is a substantial group of Small Household Seniors in each country, regardless of a country's cash intensity. Similarly, we do not find a relation between countries' cash usage and the relative size of cluster 3 - the Rural Families. Hence, the size of cluster 1 and 3 (within which cash usage is average) appears to be unrelated to the cash intensity of a country. Nonetheless, there are some noteworthy differences between the cash-intense countries and the countries where cash is used much less. We find that the relative size of cluster 4 - the Affluent Online Consumers - is larger in countries where cash usage is relatively low, such as Luxembourg and Finland. In contrast, clusters 2 and 5 - the Low-Income City Residents and the Urban Families - appear to be smaller in those countries. Thus, countries with higher cash usage comprise a larger portion of individuals categorised as Low-Income City Residents or Urban Families, and a smaller portion as Affluent Online Consumers.



Note: data from Germany and the Netherlands was not collected in the ECB survey, and could therefore not be shown in this plot.

Figure 5 Cluster division per euro area country

6 Conclusion

The primary objective of this study was to identify and characterise potential groups of payment users within payment diary survey data from the Netherlands. We identify five distinct payment user groups in our dataset, which we describe as: Family-Centric Middle-Income Consumers, Senior Traditional Banking Users, High-Income Urban Consumers, Financially Challenged Consumers and Young & Low Value Purchase Consumers.

We partition these user groups by employing an unsupervised learning approach using a k-means clustering analysis, because our dataset is unlabelled. This method partitions the dataset into k clusters based on similarities between data points. By calculating the means for all variables per cluster, we highlight each cluster's defining characteristics. This relatively straightforward concept helps us to reveal multiple consumer groups in the dataset, extending beyond previous studies that only examine the relationship between individual variables and cash usage. Finally, the groups are described using a cluster profiling analysis.

Next, we extend the analysis to data from other euro area countries. Employing a comparable methodology, an examination of cluster profiling similarly reveals the identification of five coherent clusters. We describe the five groups as: Small Household Seniors, Low-Income City Residents, Rural Families, Affluent Online Consumers, and Urban Families. Furthermore, we find that the relative share of low-income city residents and urban families is larger in countries that are cash-intense, and that the relative share of affluent online consumers is smaller in those countries.

We find multiple similarities between the identified groups in the Dutch data and the euro area data. First of all, both datasets contain a relatively large group with older consumers that are more likely to live alone (the Senior Traditional Banking users and the Small Household Seniors). Secondly, both datasets contain people that can be categorised as urban and low income consumers (the Financially Challenged Consumers and the Low-Income City Residents). Lastly, we show that higher educated digital individuals are present in both datasets (High-Income Urban Consumers and the Affluent Online Consumers). These findings indicate that, even though demographics and payment behaviour of individuals can differ significantly between countries, there are still similarities between different consumer groups.

Identifying payment user groups allows us to form a comprehensive and more holistic view of payment behaviour, which can help inform decisions regarding the future of cash. This analysis not only reveals the diverse payment behaviours but also highlights the link between socio-economic factors and payment preferences. It sets the stage for policy measures aiming to improve financial inclusion and broaden access to payment systems across all population segments. By better understanding these connections, policymakers can take tailored actions, from supporting digital inclusion programmes for less digitally literate groups to fostering innovative payment solutions that suit older consumers' needs.

References

Aggarwal, C., Hinneburg, A., and Keim, D. (2001). *On the surprising behavior of distance metrics in high dimensional space*. Springer, Berlin, Heidelberg.

Arango-Arango, C. A., Bouhdaoui, Y., Bounie, D., Eschelbach, M., and Hernandez, L. (2018). Cash remains top-of-wallet! International evidence from payment diaries. *Economic Modelling*, 69:38–48.

Bagnall, J., Bounie, D., Huynh, K. P., Kosse, A., Schmidt, T., Schuh, S., et al. (2016). Consumer cash usage: A cross-country comparison with payment diary survey data. *International Journal of Central Banking*, 12(4):1–61.

Bank for International Settlements (2022). Machine learning in central banking. *IFC Bulletin*, 57. Retrieved from https://www.bis.org/ifc/publ/ifcb57.pdf.

De Nederlandsche Bank & Dutch Payments Association (2023). Point-of-sale payments in 2022. Retrieved from https://www.dnb.nl/media/m00mskgh/points-of-sale.pdf.

Doerr, B. S., Gambacorta, L., and Serena, J. M. (2021). How do central banks use big data and machine learning? *SUERF Policy Brief*, 37:1–6.

Dutch Payments Association (2021). Infographic betaalsegmenten consumenten. Retrieved from https://www.betaalvereniging.nl/wp-content/uploads/Infographic-betaalsegmenten-update-2021.pdf.

European Central Bank (2022). Study on the Payment Attitudes of Consumers in the Euro Area (SPACE II). Retrieved from https://www.ecb.europa.eu/stats/ecb_surveys/space/shared/pdf/ecb.spacereport202212~783ffdf46e.en.pdf.

Jonker, N. (2007). Payment instruments as perceived by consumers - results from a household survey. *De Economist*, 155:271–303.

Jonker, N., Hernandez, L., de Vree, R., and Zwaan, P. (2018). From cash to cards: How debit card payments overtook cash in the netherlands. *DNB Occasional Studies*, 16(1).

Klee, E. (2008). How people pay: Evidence from grocery store data. *Journal of Monetary Economics*, 55(3):526–541.

Kosse, A. and Jansen, D.-J. (2013). Choosing how to pay: the influence of foreign backgrounds. *Journal of Banking & Finance*, 37(3):989–998.

Shy, O. (2020). Alternative methods for studying consumer payment choice. *FRB Atlanta Working Paper*, 2020-8.

Statistics Netherlands (2022). Internettoegang en internetactiviteiten; persoonskenmerken. Retrieved from https://www.cbs.nl/nl-nl/cijfers/detail/84888NED.

van der Cruijsen, C. and Knoben, J. (2021). Ctrl+c ctrl+pay: Do people mirror electronic payment behavior of their peers? *Journal of Financial Services Research*, 59(1-2):69–96.

van der Cruijsen, C. and Plooij, M. (2018). Drivers of payment patterns at the point of sale: Stable or not? *Contemporary Economic Policy*, 36(2):363–380.

van der Cruijsen, C. and Reijerink, J. (2023). Uncovering the digital payment divide: Understanding the importance of cash for groups at risk. *DNB Working Paper No. 781*.

Appendix A Variable description Dutch dataset

Table A.1 Variable overview - demographic variables

Variable	Description	Type
Gender	Respondent's gender	1 = male 2 = female
Age	Respondent's age	Discrete
Household	Respondent's household composition	1 = living with parents 2 = student house 3 = single without kids 4 = together without kids 5 = single with kids 6 = together with kids
Income	Gross yearly income in euro	1 = less than 14,300 2 = 14,300 - 23,400 3 = 23,400 - 38,800 4 = 38,800 - 51,300 5 = 51,300 - 65,000 6 = 65,000 - 77,500 7 = 77,500 - 103,800 8 = more than 103,800
Education	Highest completed level of education	1 = primary education 2 = lower secondary education 3 = upper secondary education 4 = post-secondary education 5 = higher vocational education 6 = university
Urbanisation	Degree of urbanisation	1 = not urbanised 2 = hardly urbanised 3 = moderately urbanised 4 = strongly urbanised 5 = very strongly urbanised
Ethnicity	Respondent and parents both born in the Netherlands	1 = yes (native) 2 = no (immigrant)
Occupation	Respondent's occupation type	1 = in service of government 2 = employed 3 = self-employed or freelance 4 = no paid job
Financial difficulty	Difficulty making ends meet	1 = very easy 2 = easy 3 = not hard and not easy 4 = hard 5 = very hard

Notes: This table shows an overview and description of all the demographic variables that were used for the cluster analysis. Descriptive statistics of these variables can be found in Table C.1.

Table A.2 Variable overview (continued) - payment variables

Variable	Description	Type
Internet banking	Do you use internet banking via your computer?	0 = no 1 = yes
Mobile banking	Do you use a banking app on your smartphone?	0 = no 1 = yes
Purchase count	Number of purchases an individual made	Discrete
Average purchase value	Average value per purchase in euro	Discrete
Share of cash value	Share of the total purchase value paid with cash	Discrete

Notes: This table shows an overview and description of all the payment variables that were used for the cluster analysis. Descriptive statistics of these variables can be found in Table C.1.

Appendix B Data preprocessing steps

The goal of this study is to group individuals who share similar characteristics. This can be problematic if values are missing, because the computation of measures of similarity are difficult to define for these values. Similarly, outliers can negatively impact the results of our analysis. This section describes the steps that we take to deal with missing values and outliers.

As the percentage of missing values[5] for most variables is generally below 1%, the decision to omit these observations from the analysis is justified. For the variable that captures individuals' level of urbanisation, however, there are relatively many missing values (around 10%). These values are missing because the level of urbanisation is assigned automatically based on an individual's postal code. As not all individuals chose to report their postal code, these levels of urbanisation could not be assigned. There are several ways to deal with these missing values, each with its own advantages and disadvantages. Despite the possibility that certain individuals may be less inclined to disclose their location (i.e. due to privacy concerns), we observe that the degree of urbanisation does not seem to correlate with the other variables in the dataset. This supports the selection of a random imputation method to substitute these missing values. Using this method, we randomly select values from the distribution of the non-missing data to replace the missing values. Advantages to this are that it ensures that the data distribution is preserved, it is easy to implement, it is computationally efficient and has the benefit of only imputing values that are already observed in the dataset.

For the variable that captures an individual's income group, we also find that the number of respondents that indicated they did not want to answer the question is relatively high (26%). Excluding these observations without further consideration would not be advisable, because those who choose not to disclose their income often fall within specific income brackets, such as extremely high or low incomes. Their absence in the dataset could skew the results, a phenomenon known as non-response bias. While income is likely associated with other dataset variables, we choose not to employ a multivariate imputation method that depends on these variables, because we intend to use the preprocessed and cleaned dataset for a subsequent cluster analysis.

[5]These include responses recoded as missing such as "I don't know" or "I don't want to answer"

Multivariate imputation methods can introduce data leakage or dependencies between variables, potentially undermining the quality of results. This is why we choose to also employ random imputation for the missing income group values, drawing from the distribution of the non-missing data. We extend the analysis by performing a robustness check where we use mean imputation, another imputation method that is described in Appendix F, for handling the missing income observations.

Finally, we observe that the standard deviation of the variable representing the average purchase value is relatively high. Additionally, the variable appears not to be normally distributed, implying the presence of several extreme positive values. While these high positive values remain feasible within the context of purchase value, not correcting for the skewed distribution potentially means that these few observations can badly impact the result of our analysis. To mitigate this concern, we employ a logarithmic transformation on the variable. This transformation reshapes the distribution, resulting in a more symmetric bell curve resembling a normal distribution.

Appendix C Summary statistics

In this appendix we present the summary statistics (Table C.1). Demographic variables from the summary statistics are in line with what we would expect, as the dataset is a good representation of the Dutch population. We find that the average number of purchases (2.13) and the average purchase value (€26.75) are both in line with findings reported in other studies (i.e. DNB and DPA, 2023). The share of the total purchase value that respondents paid in cash is calculated for each respondent individually. Taking the average of this number across all respondents reveals a share of cash value of 20.34%. If we look at the breakdown of the total value of payments from other studies that use same dataset (DNB and DPA, 2023), we find that 15% is paid in cash. This different finding can be explained by the fact that cash is used less for higher value purchases, resulting in a lower share of cash when aggregating across respondents. Finally, it appears that on average digital banking products such as internet banking (77.02%) and mobile banking (79.57%) have high rates of adoption in the Netherlands. Similar findings were presented by Statistics Netherlands (2022).

Table C.1 Summary statistics

Demographic variables	
Gender	45.39% male
Age (s.d.)	50.74 (17.82)
Household	together without kids: 38.83% together with kids: 26.24% single without kids: 21.98%
Income	23,400 - 38,800: 21.69% 38,800 - 51,300: 20.12% 51,300 - 65,000: 13.69%
Education	higher vocational education: 28.96% post-secondary education: 22.42% lower secondary education: 21.69%
Urbanisation	strongly urbanised: 34.07% very strongly urbanised: 20.33% hardly urbanised: 19.78%
Ethnicity	yes (native): 85.26% no (immigrant): 14.74%
Occupation	employed: 48.00% no paid job: 39.61% in government service: 6.66%
Financial difficulty	easy: 41.77% not hard and not easy: 36.02% very easy: 14.46%
Payment variables	
Internet banking	77.02% yes
Mobile banking	79.57% yes
Purchase count	min: 1, mean (s.d.): 2.13 (1.6), max: 18
Average purchase value	min: 0.01 euro, mean (s.d.): 26.75 euro (39.61), max: 1287.33 euro
Share of cash value	min: 0%, mean (s.d.): 20.34% (37.64), max: 100%

Note: This table shows the summary statistics of the dataset that is used for the cluster analysis. For categorical variables exceeding three categories, it presents the three highest answer shares. Data from the SCP is used. The number of observations is 25,862.

Appendix D Dimension reduction

In this study, we adopt Factor Analysis of Mixed Data (FAMD) as our chosen dimensionality reduction method. Unlike ordinary principal component analysis, FAMD is better suited to datasets with mixed data types, as it efficiently handles numerical and categorical information. The goal of FAMD is to find the most relevant components that can effectively describe our data, regardless of its type. This means we can handle both numerical and categorical information, which is essential for our analysis.

Determining the appropriate number of dimensions for FAMD is an important step. Although this selection always involves some level of subjectivity, there are guidelines to support this process. In this specific scenario we find that almost 25% of the variance is explained by the first five dimensions. Also, adding more than five dimensions does not significantly improve this level of variation anymore. These arguments provide good grounds to choose to reduce the amount of dimensions to five. Detailed results of all dimensions and their contribution to the original variables are presented in Table D.1.

Table D.1 Contribution of the variables to the dimensions

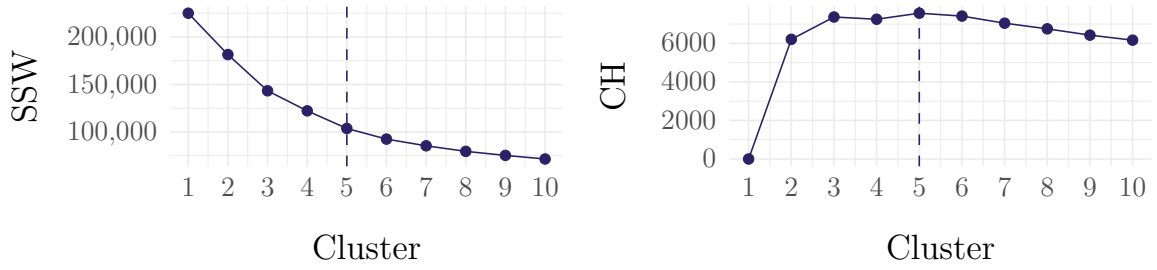
	1	2	3	4	5
Gender	1%	1%	6%	1%	0%
Age	6%	28%	0%	0%	0%
Household	12%	29%	32%	11%	37%
Income	19%	3%	13%	20%	13%
Education	14%	14%	15%	22%	11%
Urbanisation	0%	0%	1%	16%	1%
Ethnicity	0%	0%	1%	7%	0%
Occupation	21%	2%	6%	8%	9%
Financial difficulty	8%	2%	16%	14%	1%
Internet banking	0%	13%	0%	1%	2%
Mobile banking	10%	1%	3%	0%	1%
Purchase count	0%	1%	1%	0%	5%
Log average purchase value	2%	7%	4%	1%	5%
Share of cash value	6%	0%	1%	0%	14%

Note: This table shows the contribution of all variables to the dimensions calculated in the FAMD. The individual values can be interpreted as the weight of these variables in the individual dimensions.

Appendix E Cluster analysis

The validation statistics used for selecting the optimal number of clusters are presented in Figure E.1. The first panel in this figure displays the sum of squares within-cluster (SSW), which quantifies the distance between each centroid and its corresponding data points. Our objective is to identify the "elbow" point on the SSW curve, where further increasing the number of clusters no longer significantly reduces the SSW. This point represents a balance between minimising within-cluster distances and preventing overfitting, ensuring a meaningful clustering solution.

According to this assessment, the optimal number of clusters appears to be five. The second panel of the figure illustrates the Calinsky-Harabasz index (CH) for various numbers of clusters. The CH index assesses how well data points in a cluster are similar to each other compared to other clusters, with higher values indicating denser and better-separated clusters. Notably, this validation statistic also suggests that the ideal number of clusters is five, as the CH value reaches its peak at this value. Both metrics indicate that an appropriate k-value for the k-means clustering analysis is five.



(a) Within-cluster sum of squares (SSW) (b) Calinsky-Harabasz index (CH)

Figure E.1 Cluster validation statistics for various number of clusters

Table E.1 Variable means per cluster

Variable	1 Family-Centric Middle-Income Families	2 Senior Traditional Banking Users	3 High-Income Urban Consumers	4 Financially Challenged Consumers	5 Young & Low Value Purchase Consumers
Gender	1.60 (0.49)	1.53 (0.50)	1.36 (0.48)	1.73 (0.45)	1.56 (0.50)
Age	44.5 (11.4)	68.0 (9.24)	49.6 (12.9)	48.8 (16.0)	18.0 (8.61)
Household	4.92 (1.29)	3.83 (0.59)	4.52 (1.18)	3.57 (1.12)	1.49 (1.41)
Income	4.39 (1.42)	3.67 (1.37)	6.18 (1.60)	2.06 (1.22)	4.04 (2.01)
Education	4.12 (1.00)	3.23 (1.37)	5.09 (1.08)	3.54 (1.36)	2.07 (1.35)
Urbanisation	3.14 (1.20)	3.23 (1.23)	3.75 (1.20)	3.86 (1.18)	3.24 (1.24)
Ethnicity	1.10 (0.30)	1.10 (0.31)	1.20 (0.40)	1.28 (0.45)	1.17 (0.38)
Occupation	2.11 (0.57)	3.69 (0.75)	2.16 (0.89)	3.13 (0.99)	3.19 (0.99)
Financial difficulty	2.43 (0.65)	2.38 (0.72)	1.70 (0.70)	3.34 (0.90)	2.47 (0.85)
Internet banking	0.75 (0.44)	0.87 (0.34)	0.90 (0.30)	0.68 (0.46)	0.30 (0.46)
Mobile banking	0.93 (0.25)	0.57 (0.49)	0.92 (0.27)	0.77 (0.42)	0.78 (0.42)
Purchase count	2.20 (1.66)	2.17 (1.56)	2.09 (1.50)	2.19 (1.76)	1.74 (1.36)
Average purchase value	30.4 (38.6)	26.1 (37.3)	32.4 (53.0)	20.4 (27.3)	9.80 (14.6)
Share of cash value	14.3 (32.2)	28.3 (42.3)	10.9 (28.5)	29.6 (42.7)	25.1 (41.5)

Note: The values in the table represent the means of all variables for each cluster after performing k-means cluster analysis on the dataset following dimension reduction. The standard deviations are shown in parentheses.

Appendix F Robustness checks

This section extends the main analysis by examining whether the reported results are robust to changes in the specified models. First, we check for stability of the clustering solution by randomly selecting and reordering a different subsample of the dataset. If the clusters are consistent across different subsamples, this suggests that the clustering algorithm is robust and not highly sensitive to the specific composition of the data. Second, as mentioned in Appendix B, around 26% of the observations had a missing value for the income group variable. Therefore, introducing a robustness check to assess the impact of our chosen imputation method is a crucial step in ensuring the reliability of our cluster analysis results. This is why we rerun the analysis using a mean imputation method for the income group variable, rather than a random imputation method. Mean imputation is a technique used to replace missing values in the dataset by substituting them with the mean value of the available data for that variable. Last, to verify whether our conclusions remain intact after isolating the effects of the COVID-19 pandemic, we rerun the cluster analysis on a subset of the data that only contains survey data that was collected after March 2022.

The robustness analyses provide several key insights[6]. Firstly, they confirm that the relative cluster sizes closely align with the baseline results, thus providing strong support for the existence of five clusters. Additionally, we find that the mean values of the variables within each cluster are similar to the initial results. While there may be slight variations in the variable means compared to the baseline results, these findings continue to underscore the reliability of the initial interpretations.

Appendix G Extended cluster analysis euro area

Table G.1 Variable means per cluster using data from euro area countries

Variable	1 Small Household Seniors	2 Low-Income City Residents	3 Rural Families	4 Affluent Online Consumers	5 Urban Families
Gender	1.51 (0.50)	1.63 (0.48)	1.62 (0.49)	1.47 (0.50)	1.30 (0.46)
Age	66.4 (8.73)	38.9 (13.5)	40.5 (12.7)	44.1 (13.0)	47.0 (13.5)
Education	2.02 (0.72)	2.33 (0.61)	1.91 (0.56)	2.76 (0.46)	2.03 (0.85)
Household size	1.84 (0.70)	2.73 (1.18)	3.24 (1.13)	2.76 (1.17)	3.15 (1.10)
Net income	2.98 (1.07)	2.50 (1.08)	3.43 (0.83)	4.02 (1.08)	3.50 (0.89)
Urbanisation	0.60 (0.49)	0.88 (0.32)	0.30 (0.46)	0.75 (0.43)	0.82 (0.39)
Online banking	0.60 (0.49)	0.60 (0.49)	0.70 (0.46)	0.90 (0.29)	0.64 (0.48)
Occupation	2.92 (0.34)	2.22 (0.54)	2.19 (0.46)	2.03 (0.42)	1.58 (0.71)
Purchase count	2.20 (1.28)	1.84 (0.98)	2.24 (1.29)	2.40 (1.40)	3.14 (1.82)
Average purchase value	33.1 (54.9)	12.7 (16.7)	36.2 (56.9)	47.3 (99.6)	36.3 (60.0)
Share of cash value	0.56 (0.45)	0.80 (0.37)	0.47 (0.45)	0.18 (0.33)	0.66 (0.40)

Note: The values in the table represent the means of all variables for each cluster after performing k-means cluster analysis on the dataset following dimension reduction. The standard deviations are shown in parentheses. Data from the ECB Study on the payment attitudes of consumers in the euro area (SPACE) 2022 was used (ECB, 2022).

[6]The outputs of the robustness analyses are available upon request.