Kovács *et al. J Big Data* (2021) 8:141
https://doi.org/10.1186/s40537-021-00529-4

Journal of Big Data

RESEARCH Open Access

Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis

Tibor Kovács[1], Andrea Ko[1] and Asefeh Asemi[2*]

*Correspondence: asemi.asefeh@uni-corvinus.hu
[2] The Doctoral School of Economics, Business, and Informatics, Corvinus University of Budapest, Fővám tér 8, 1093 Budapest, Hungary
Full list of author information is available at the end of the article

Abstract

Identifying investment patterns as part of customer segmentation is one of the most important tasks in retail banking. Clustering customers effectively is an important element of improving marketing policy and strategic planning. There are several methods for identifying similar groups of customers and describing their characteristics to offer them appropriate products. However, using machine learning methods is rare, and the application is limited for certain types of data. The aim of this study is to investigate the benefits of using a two-stage clustering method using neural-network-based Kohonen self-organizing maps followed by hierarchical clustering for identifying the investment patterns of potential retail banking customers. The unique benefit of this method is the ability to use both categorical and numerical variables at the same time. This research examined 1,542 responses received for an online investment survey, focusing on the questions that are related to the respondents' investment preferences and their current financial assets. The research utilizes descriptive statistics and multiple correspondence analysis (MCA) to understand the variables and Kohonen self-organizing maps (SOMs), in combination with hierarchical clustering, to identify customer groups and describe the characteristics of these clusters. The analysis was able to identify clusters of potential customers with similar preferences and gained insights into their investment patterns related to their investment portfolio and investment behavior, including their savings profile, attitude to risk-taking, and preferences for investment advice. These findings were supported by additional insights through the application of multiple correspondence analysis (MCA) describing patterns of financial instruments and portfolios. The main contribution of the research is the combined application of the machine learning methods Kohonen SOM, hierarchical clustering, and MCA for investment pattern analysis in the retail banking business.

Keywords: Investment Patterns, Factors Affecting Investment, Customer Clustering, Retail Banking Business, Kohonen Self-Organizing Maps, Multiple Correspondence Analysis

Introduction

Investors follow different investment patterns depending on the circumstances they find themselves in. These factors may be environmental or personal. Environmental factors could be related to the economic conditions of the community and the resources

available to or the constraints bound by investors for investment. Personal circumstances could be related to the amount of savings, the personality traits of the individuals, or the amount of knowledge they have about investing. Investment patterns can be observed in various financial assets. These include the purchase of securities, monetary or paper (financial) assets, cryptocurrencies, and relatively liquid real assets such as gold, real estate, or art collections. In fact, an "investment pattern is the investment in different investment avenues with vital consideration of risk and return" [1]. Strong organizational relationships with customers are required to run a successful business [2]. Studying customer behavior in the management of investment affairs is in many respects almost a new issue [3]. Many records are created every day through several agencies about customers in different parts of the world. Structured, semi structured, and unstructured data are being created at a rapid pace from heterogeneous sources such as reviews, ratings, feedback, trading details, investment data, etc., leading to Big Data. This information generated about customers may hold many frequent patterns that can be filtered and analyzed to provide suggestions about the products, items, or offerings that customers might be interested in. The internet also offers a vast amount of information about potential customers through a variety of investment websites. They attract customers by offering products and services online, tailored to different customer groups, and many of them also provide free consulting services. Online searches also result in a group of products and services with features that the customer can see and check before buying them [4]. In financial and investment companies, the customer relationship management system is a repository of customer information and experience, holding all customer profiles. With this information, it is possible to customize the products or service offerings for each unique customer based on their personal needs before they decide to invest. To better manage customer investments, companies need to categorize, cluster, and segment them based on their individual circumstances. "One of the most effective tools to understand consumers' motivation and behavior is segmentation" [5]. Customer segmentation therefore could help to optimize marketing policy and strategic planning to maximize profitability. How the customer uses a product and evaluates it is also important because it affects reuse behavior. Barczak, Ellen, and Pilling [6] reflected on this by examining the consumption behavior of bank customers after purchase. These are the reasons why researchers are trying to help investment companies and their clients make the best investment decisions based on their characteristics. On the other hand, one of the problems investment companies are facing is the lack of sufficient information about customers. Therefore, experts try to discover hidden patterns in customer data. In this way, they can provide suitable investment options for customers. There are several ways for investment companies to use all available opportunities, strengthening competition between companies and discovering hidden knowledge about customers [7, 8]. The latest research in artificial intelligence (AI), especially machine learning methods and neural networks, provides new opportunities for analyzing customers' behavior [9–11]. Future AI trends in business include forecasting customer behavior [12], predicting customers' responses to direct marketing [13], analyzing customer churn, and predicting market evolution [13, 14].

This study aims to explore potential customers' investment patterns to provide them with appropriate products in retail banking. We applied a novel, two-stage clustering

approach: combining Kohonen self-organizing maps [15, 16] with hierarchical clustering as an unsupervised machine learning method for exploring the respondents' investment characteristics. This approach enables us to use both categorical and numerical variables that are common in survey instruments. In addition, we used multiple correspondence analysis (MCA) to gain insights into the patterns potential customers use to think about financial instruments and portfolios. This study investigates customers from multiple dimensions that may influence their investment patterns: the respondents' current investment portfolio, the magnitude of their available savings, how they are making ends meet, and their views on appropriate investment products and risk profiles. This research utilizes the responses of customers (as potential investors) to an online investment questionnaire published by a leading Hungarian financial portal. The paper first discusses the theoretical background and reviews the published scientific works. We performed two types of analysis: an analysis of the subjects of the published articles in relation to the factors influencing the investment and a semantic analysis in relation to the basic concepts related to the research problem. We used the Voyant tool [17] for term analysis and the Yewno tool [18] to find the semantic relation between basic concepts in the research knowledge map. This is followed by introducing the investment questionnaire, the data collected, and the methodology applied. The data analysis and findings sections describe the data dimensions that are relevant from an investment point of view and discuss the customer clusters identified by applying the two-stage clustering method. The paper concludes by summarizing the results and discussing the applicability of the method.

Literature review

Edmondson and Mcmanus [19] stated that theory is developed as an outcome of a study, new ideas that contest conventional wisdom, challenge prior assumptions, integrate prior streams of research to produce a new model, or refine the understanding of a phenomenon. A review of our scientific literature shows that related research studies focus on different aspects of this topic. Here, in addition to summarizing the key concepts of the research topic, we reviewed the past literature related to this study. The following is a general analysis of the subjects and concepts of published articles related to this research. Scientific databases and web analysis tools were used to perform this analysis.

Retail banking is the direct provision of banking services to individuals. This area of banking includes a variety of services, such as cash cards, credit cards, debit cards, current and savings accounts, mortgages, and personal loans. These services are provided to customers through various service channels, such as branch chains, ATMs, internet banking, and telephone banking. Retail banking can also be defined as receiving deposits from people and lending them to individuals, firms, and companies. By this definition, banks act as direct financial intermediaries. Customer experience is influenced by the contribution of both the customers and the company. All the events encountered by customers before and after a transaction are part of the customer experience. Keiningham et al. [20] believe that innovation in the business model (BMI) is crucial for a company's ability to achieve long-term growth and sustainability. Creating innovations to improve the value of products or services or delivering appropriate offers helps customers to use products and services more effectively. What customers encounter is personal

and may involve sensory, emotional, rational, and physical aspects to create a memorable experience. In retail banking, both investors' experiences and the qualities of investment funds play an important role in the success of services and products offered by the company. Maklan and Klaus [21] presented the customer experience quality (EXQ) scale to measure the quality of the customer experience. They concluded that key features of the customer experience play an important role in assessing service quality or customer satisfaction in the market. Klaus [22] then explained the conceptualization and implementation of customer experience (CX) quality on the EXQ scale. Kuppelwieser and Klaus [23] developed this scale and systematically examined the psychometric properties of EXQ. They studied the nature of the relationships between these dimensions, as well as between their dimensions and cases. The results of their research showed that customers evaluate and understand experience as a general evaluation and do not differentiate between the meanings of different stages or dimensions of experience. Wewege and Thomsett [24] have addressed this issue with the publication of the third edition of their book titled *The Digital Banking Revolution*. They stated how fintech companies are transforming the retail banking industry through disruptive financial innovation. Innovation in customer segmentation is one of the important issues in the retail banking industry.

Marco et al. [25] show that the use of cognitive analytics management is a valid tool to describe new technology implementations for businesses. They found that a self-organizing map better classifies the customer base of a retailer by paring two machine learning algorithms. Fatima and Sharma [26] identified certain biases affecting investor decision-making and segmented investors accordingly. They used factor analysis, and the findings revealed that eight extracted factors affect investment decisions: tend to fall into imitator, stereotypical, independent individualist, risk-tolerant, efficient planner, confident, passive, and competent confirmer. Jääskeläinen [27] analyzed customer data from a local retail bank using machine learning to detect the attributes of investors. They used different clustering algorithms. He found that a customer invests if he or she has an investor profile, higher account balance, job, and marketing permission. Some authors have examined the impact of various factors that influence customer decisions [9, 28–31]. Some studies have been conducted on customer clustering and segmentation based on customer behavioral perspectives, customer behavioral factors, demographic factors, and environmental objects [32–38]. Goncarovs [39] described a five-step customer segmentation method consisting of gathering quantitative information, creating specific microsegments, sorting microsegments, and creating final customer segments. Artificial intelligence and machine learning play an important role in identifying investment patterns for banking innovations, with the addition of risk capital and other emerging technologies also considered by scientists [40]. Boone and Roehm [41] examined the use of artificial neural networks (ANNs) as an alternative means for the segmentation of retail databases. They concluded that ANNs are useful for retailers in market segmentation because they offer more homogeneous segmentation solutions than the mixed model and K-means clustering algorithms and are less sensitive to initial start-up conditions. Ying Li and Feng Lin [42] used data mining methods to segment clients in the securities industry from the perspective of customer value and customer behavior. They believe that the clustering algorithm could be used as a customer segmentation method

commonly used in data mining. Li, Wu, and Lin [43] proposed a two-stage clustering algorithm based on a self-organizing feature map, which uses the self-organizing feature to cluster the raw data initially. Then, the behavioral and value features of the segmented groups of customers were filtered out using a data mining tool. They concluded that, in terms of behavioral features, security customers are segmented into general type, important type, and silent type. Bigné et al. [44] examined neural networks, specifically SOM, as an alternative to traditional statistical segmentation methods and identified segments in the mature market. The results show the superiority of nonhierarchical clustering and SOM over hierarchical clustering and show their complementary nature. Mak, Ho, and Ting [45] presented a financial data mining model for extracting customer behavior. They aimed to increase the availability of decision support data and hence, increase customer satisfaction. Their simulation experiments showed that the proposed method can improve the turnover of a financial company and deepen the understanding of investment behavior. Saluja and Shaikh [46] decoded the investment pattern of large payers, such as foreign institutional investors (FIIs) and domestic institutional investors (DIIs), using the decision tree method of machine learning techniques. Chen, Ho, and Liu [47] analyzed whether personality creates significant differences in financial performance. They used investor personalities rather than sentiment, which is difficult to predict because of noise. They applied statistics tests and machine learning algorithms to achieve their goal. Albert, Merunka, & Valette-Florence [48] and Lamprinopoulou & Tregear [49] used the MCA criterion to review and analyze the literature, distinguish key factors in motivating the network to share knowledge, accelerate innovation, reduce transaction costs, improve reputation, and create new market opportunities.

Clustering data with both categorical and numerical variables is a challenge, as many clustering algorithms either expect categorical or numerical data [50]. K-means clustering, which is a widely used method, expects only numerical data, while others [51, 52] use only categorical variables. One way to overcome this problem is to cluster first the numerical variables and then combine the results with the categorical variables and apply a clustering algorithm that is designed for categorical variables [53]. Another approach for two-stage clustering is the combination of self-organizing maps with K-means clustering, [54], which is proposed for market segmentation. Artificial neural networks, such as self-organizing maps, have demonstrated their learning capabilities; they can handle a large amount of data and can handle mixed variables of categorical and numerical data. The result of the self-organizing map has well-defined distances for the nodes, wherein hierarchical clustering could be easily applied to aggregate them to the optimal number of clusters. We applied this method to our dataset.

To obtain a general understanding of the research conducted in relation to the "factors affecting investment", we analyzed all English language articles published in academic journals between 2010 and 2020. We searched "factors affecting investment" in the "keyword" field by the "SuperSearch"[1] tool available at the library of the Corvinus University of Budapest. We created a document from all the subjects of these articles and then analyzed this document using the Voyant web-based tool [17]. The

[1] http://www.lib.uni-corvinus.hu/eng.

Table 1 Factors affecting investment from articles published in academic journals in 2010–2020

No	Factor affecting investment	Count
1	Financial (e.g., Saving & Capital)	89
2	Business Decision-Making	65
3	Environmental	60
4	Economic	35
5	Market	27
6	Foreign Direct Investment	24
7	Management	23
8	Risk	22
9	Development	17
10	Investors	14

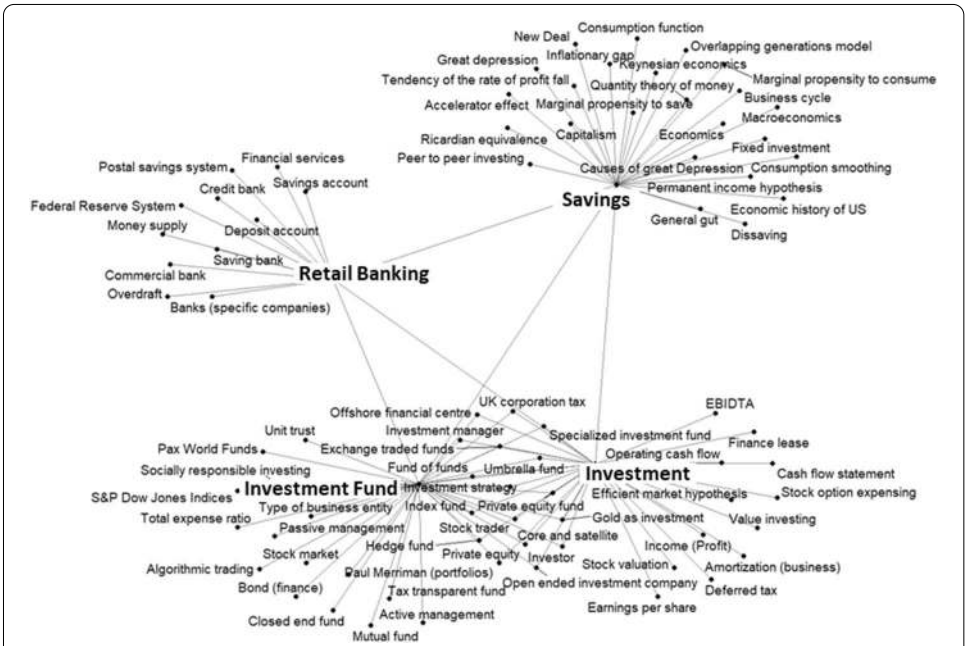

Fig. 1 Knowledge map of the four concepts: investment, investment funds, saving, and retail banking

document was classified into 10 groups based on the factors' relative frequencies. Table 1 shows the factor frequencies in the entire corpus. The factor is the subject in the corpus, while the count is the frequency of the factor in the corpus.

To further analyze the relationship between the basic concepts related to our research, we prepared a knowledge map. For this reason, we used the Yewno Semantic Search Tool [18] to draw a scientific map of the research concepts. The semantic relationship has four main concepts: investment, investment funds, savings, and retail banking, depicted in Fig. 1. There are overlapping concepts between investment and investment funds, investors, stock traders, investment companies, and several investment instruments, such as private equity funds or exchange-traded funds. The various types of funds and investment portfolios are the most important concepts for

investment funds, while performance metrics (e.g., earnings per share, stock valuation) are the most important concepts for investment. Retail banking as a concept refers to several types of banks (a reference to specific companies has been removed from the semantic map), while savings refers to various economic theories around consumption and investment. Our model was formulated with reference to these concepts: the types of investment funds the person finds appropriate are analyzed in relation to the person's saving and consumption patterns, approach to performance and risk, and the types of assets their family holds to care for the future.

In summary, the research background shows that many methods have been used to classify and cluster customers and analyze these clusters; however, the use of AI, especially neural networks, is still rare. There are several research studies on customer experience and investment patterns examining customer experiences from different perspectives. They examined this relationship from different aspects, including the impact of demographic factors. Having both categorical and numerical data as a source for clustering presents challenges. The practical use of Kohonen self-organizing maps to overcome these challenges is rare, and practical application is not well documented for analyzing investment patterns. We propose here a novel approach using Kohonen self-organizing maps in combination with hierarchical clustering and describe the practical application of this method. Our research could help companies develop better investment proposals based on customers' investment patterns using this approach.

Research questions and methodology

This section introduces the methodological approach of the study: the research process, the data collection methods, and the data analysis techniques. It describes how the components are interconnected and form a logical sequence and will evaluate and justify the reasons for the proposed methodological options.

Our main goal is to demonstrate the effectiveness of the two-stage clustering method using Kohonen self-organizing maps (SOMs) and hierarchical clustering by exploring investment patterns in potential retail banking customers. We explored this through answering three research questions:

- Could we identify investment patterns using this clustering method?
- Could we describe the important investment factors?
- Could we recommend appropriate investment products for potential investors?

Prior to selecting the variables for building the machine learning model, we analyzed the variables related to investment and financial awareness. Three variables were excluded from the model:

- Current investment portfolio consists of…"
- "Influencing factors of investment decisions"
- "Which (fintech) product have you heard of and have you used?"

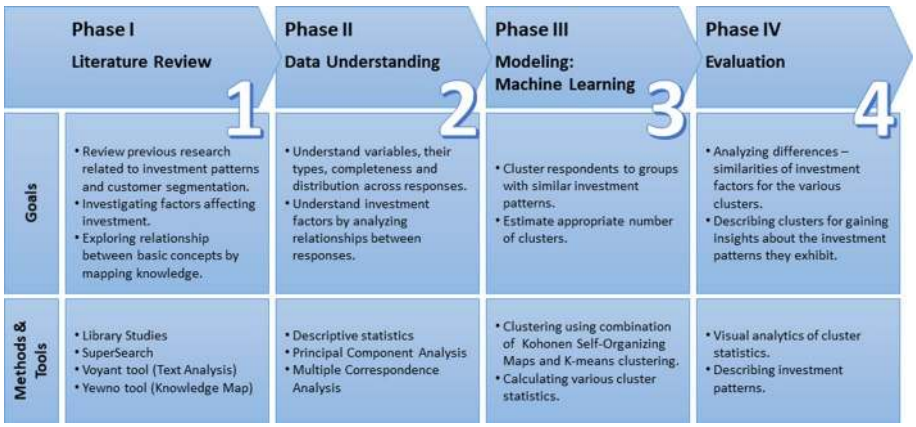

Fig. 2 - Summary of research phases and processes

Having the following five variables selected to build the machine learning model to identify clusters of potential customers with similar preferences and investment patterns:

- "Which product is appropriate for you based on your opinion?"
- "What are the important factors for your long-term savings?"
- "How much savings do you have?"
- "How do you make ends meet?"
- "What Investment your family has to provide financial stability?"

Research process and data analysis techniques

Our study is experimental and exploratory research. The data collection was both quantitative and qualitative; therefore, a mixed method was used in this study. Details of each research phase are shown in Fig. 2. We used descriptive statistics, principal component analysis, and multiple correspondence analysis to understand the variables that are relevant to investment. It is followed by an unsupervised machine learning method that combines Kohonen SOMs and hierarchical clustering to identify clusters of potential customers with similar investment patterns in the retail banking context. To implement the models, the calculations were carried out in the R environment [55]. Principal component analysis, multiple correspondence analysis, and hierarchical clustering were performed using the 'factoextra' package [56], while Kohonen self-organizing maps were created using the 'kohonen' package [57, 58]. The research was conducted in four phases. In phase one, we present a review of the literature and research background. In phase two, the data were explored using descriptive statistics, principal component analysis, and multiple correspondence analysis to understand the relevant investment factors. In phase three, unsupervised machine learning was used to identify clusters of potential investors. In the last phase, the characteristics of these clusters were analyzed further to describe similarities and differences. The confidence levels of our results are limited by the number of samples and variables available in our research. Therefore, the results should be recognized as qualitative rather than quantitative.

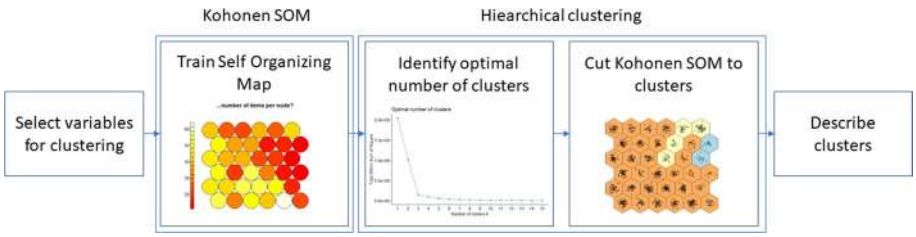

Fig. 3 Process of the two-step clustering

The identification of customer clusters was begun by applying the Kohonen SOM method, as the dataset describing customer behavior included both categorical (e.g., types of financial assets they hold) and numerical data (e.g., amount of discretionary savings they hold). The output of the Kohonen SOM is a map with a predefined number of nodes (in our case 36), holding differing numbers of customers. This initial step helps overcome the problem of having mixed data of categorical and numerical variables and creates a clustered representation of customers, where the similarity between the nodes (distances from neighbors) can be measured as Euclidean distance. The second step of our method is the hierarchical clustering of the nodes using node distances and the subsequent estimation of the ideal number of final clusters using well-known methods, such as the silhouette method [59], elbow method or gap statistics [60]. The combination of these two methods is proposed because it overcomes the problem of having mixed categorical and numerical variable data, and the Kohonen method can handle large datasets efficiently. Other methods, such as K-means or hierarchical clustering, were discarded due to the nature of our dataset. The process of our two-step clustering process is illustrated on Fig. 3.

Data collection, data quality

This study uses data that were collected through an online investment questionnaire published by a leading Hungarian financial portal called "Portfolio". The questionnaire is accessible in a web-based format at https://www.portfolio.hu/befektetesi-kerdoiv/?page=1 in the Hungarian language. "Portfolio" is an online financial portal in Hungary with a user visit count over 15 million per month as of December 2020, ranking as the 28th most visited side in Hungary [61]. "Portfolio" has a distinct emphasis on business, financial, and economic news. In addition to its online media platforms, the enterprise also offers a trading platform and presents a personal analysis of financial markets. The company also has activities in the field of commercial enterprises and organizes annual professional fora in the fields of agriculture, insurance, lending, asset management, corporate finance, capital markets, the car sector, monetary IT, and real estate [62]. Our data consist of 1542 responses to the online questionnaire received through the portfolio.hu website in 2019. The investment questionnaire was designed in partnership with Corvinus University of Budapest and the Dorsum company, one of the region's leading providers of innovative investment software. It is the result of joint research with the purpose of determining how conscious their readers

Table 2 The structure of the Dorsum questionnaire

Section	Subject	No. of Questions	Question Type
1	Affinity to information technology	6	Single/Multiple choice
2	Financial awareness, investment position and risk appetite	5 4 1	Multiple choice Y/N Numeric/fill-in-the-blank
3	Spending habits and savings	6	Multiple choice
4	Personality and decision-making traits	10 10	5-point Likert scale Single/Multiple choice
5	Financial Planning for the Future	2	Open-ended
6	Banking service and satisfaction (Number of banks currently used, customer type: regular, premium, private banking client)	1 3 1 19 3	Numeric/fill-in-the-blank Single/Multiple choice Ranking 5-point Likert scale Y/N
7	Demographic Data (Age, gender, residence, education, job)	1 6	Numeric/fill-in-the-blank Single/Multiple choice
Total		76 Questions	

are about their finances. This questionnaire is published on portfolio.hu in Hungarian, translated to English for the analysis and the explanations of findings and insights.

The web-based investment questionnaire has 74 variables (questions), grouped into seven main sections exploring the respondent's financial awareness and affinity to information technology and novel financial services, as well as collecting data about their demographics (Table 2). The questionnaire starts with asking about the respondent's affinity to information technology and their use of social media and online, followed by questions about their financial awareness, investment portfolios, investment approaches, risk profiles, and spending and savings habits. The next section of the questionnaire is about the respondent's personality profile, followed by a single open-ended question about short- and long-term financial planning. The next, substantial section is about the respondent's relationship with banks and their understanding and experience with novel banking (fintech) products. The questionnaire concludes with questions about the respondents' demographics. There are categorical variables (questions), either single-choice or multiple-choice; numerical variables measured on a 5-point Likert scale; and unstructured, textual data for the open-ended questions.

Data quality was good, and missing data were below 1%, except for some demographic data. In a significant portion, 45% of respondents did not provide information about their age; therefore, we could not use this in our models. However, for those who provided this information, 30–34 years was the most frequent age bracket. Interestingly, 7.8% of respondents stated that their age was 15–19 years old. Similarly, a poor response rate was received for the highest levels of education: 77% did not provide this information. There were better responses for other demographic data: most respondents were from the capital city (50%) or from larger cities and towns (20% + 17%). Only 8% were from rural areas. The most frequent (42%) current occupation is a "graduate employee", meaning that they are employed by a company, that their occupation requires a graduate degree, and that they do not have management duties.

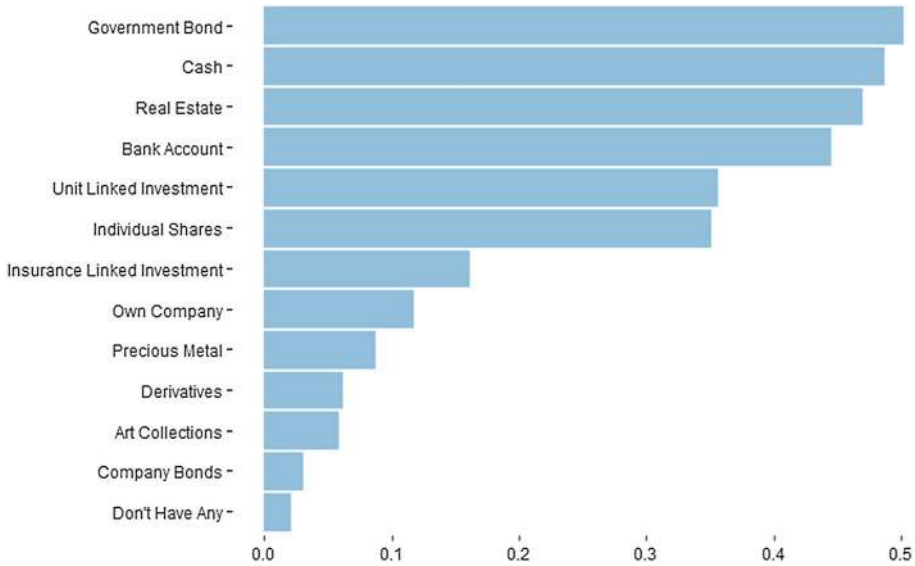

Fig. 4 Current investment portfolio of respondents

Results and discussion

This section provides an overview of the results of descriptive statistics and multiple correspondence analysis. We discuss, among other influencing factors of investment decisions, familiarity with novel financial products, important factors for long-term savings, risk appetite, appropriate investment products based on respondents' opinions, and means to provide financial stability. The confidence levels of the results described below are limited by the limited number of samples and therefore should be recognized as qualitative results.

Understanding variables: descriptive statistics and multiple correspondence analysis (MCA)

The following section introduces the data using descriptive statistics and MCA.

Current investment portfolio

Thirteen investment product options were presented to the respondents to select from, indicating what assets they currently hold (multiple choice option). On average, they reported holding 3 different products (with a range of 1 to 9). Government bonds were the most popular choice, followed by cash, real estate and a bank account in popularity (Fig. 4). The multiple correspondence analysis did not reveal any obvious dimensions of preferences among the variables.

Influencing factors of investment decisions

Respondents were asked to rank their preferences of where they seek advice for investment decisions, with choices of relying on their own opinions; through family, friends,

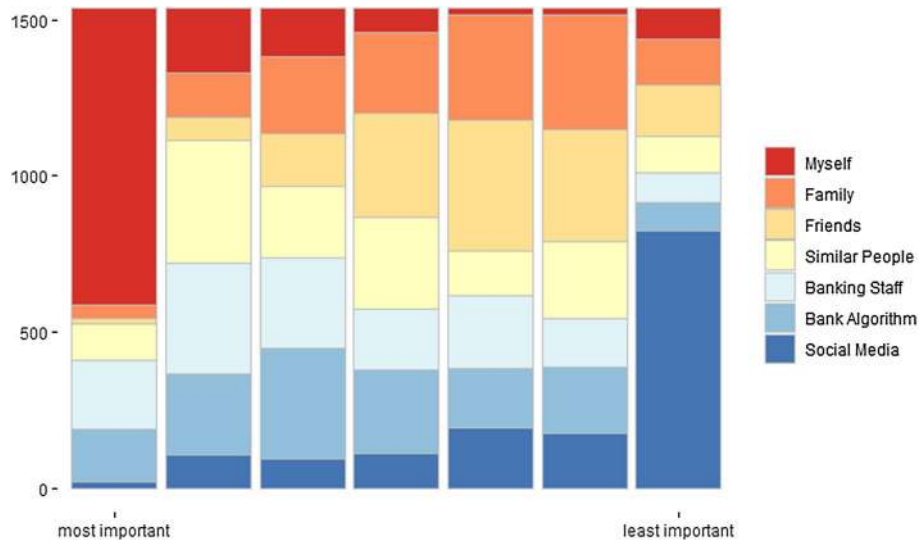

Fig. 5 Influencing factors and information sources of investment decisions

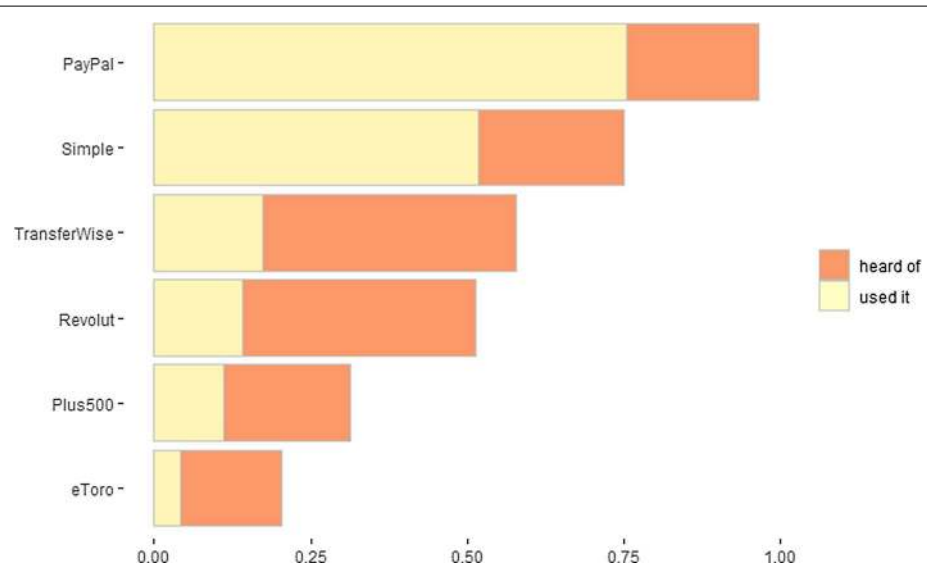

Fig. 6 How familiar respondents are with novel financial (fintech) products

or banking staff, or bank algorithms and social media recommendations. Most of the respondents (62%) chose to rely on their own opinions and ranked social media (54%) as the least important factor (Fig. 5).

Familiarity with novel financial (fintech) products

Respondents were asked to indicate what novel financial (fintech) products they know about or have used. PayPal was the most well-known product; 96% of the respondents knew it, and 75% used it. It was followed by Simple, a local fintech product used for online and mobile payments (Fig. 6).

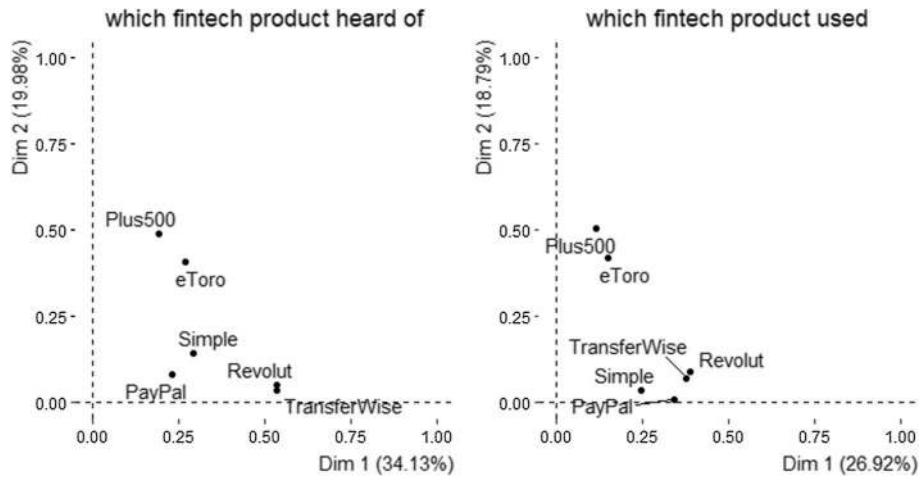

Fig. 7 MCA analysis of knowledge of and experience with fintech products

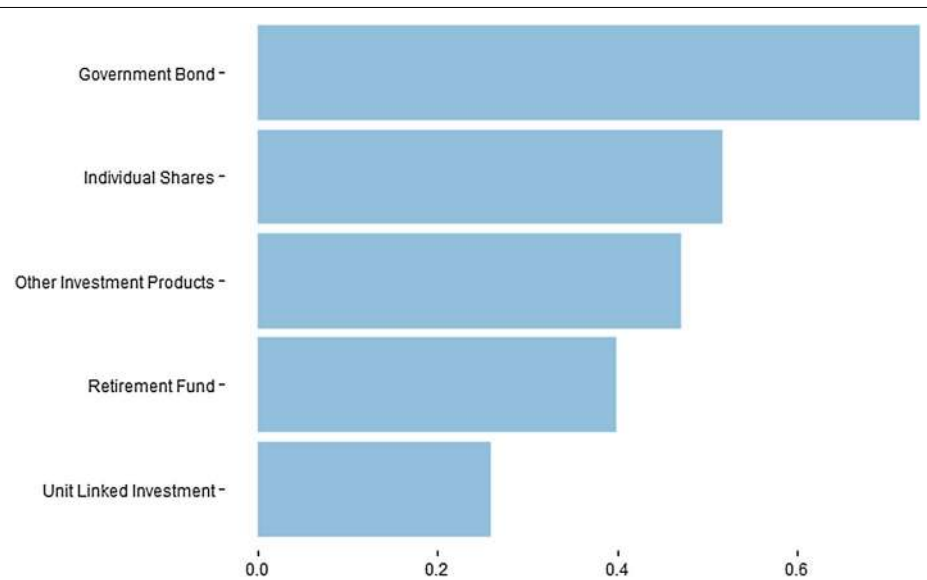

Fig. 8 Which investment product respondents think is appropriate for them

The MCA analysis revealed that Plus 500 and eToro are rather different products from the others based on the respondents' knowledge and prior experience. This is in line with the nature of the products, the former being online trading and the others being predominantly online payment platforms (Fig. 7).

Appropriate investment products based on respondents' opinion

Respondents were asked to choose multiple investment products from a list that they think is appropriate for them. On average, they chose 2 out of the 5 options. The most popular was government bonds (74%), followed by individual shares (Fig. 8). The MCA analysis revealed that respondents thought that there were two groups of products:

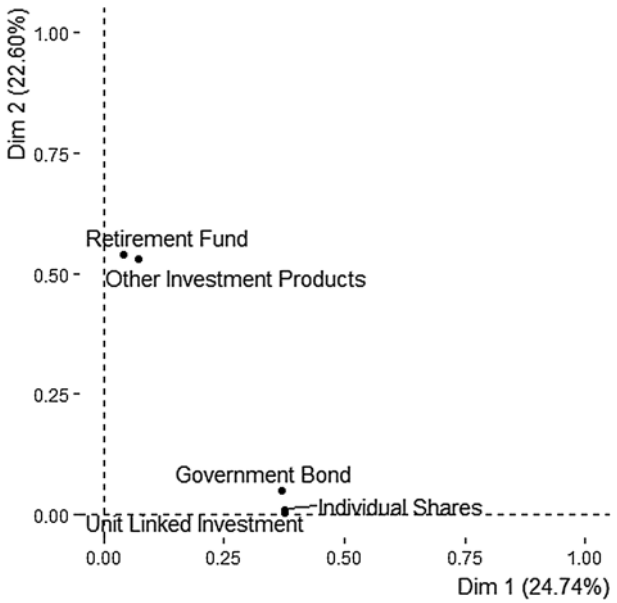

Fig. 9 MCA analysis of question "which product is appropriate for you"

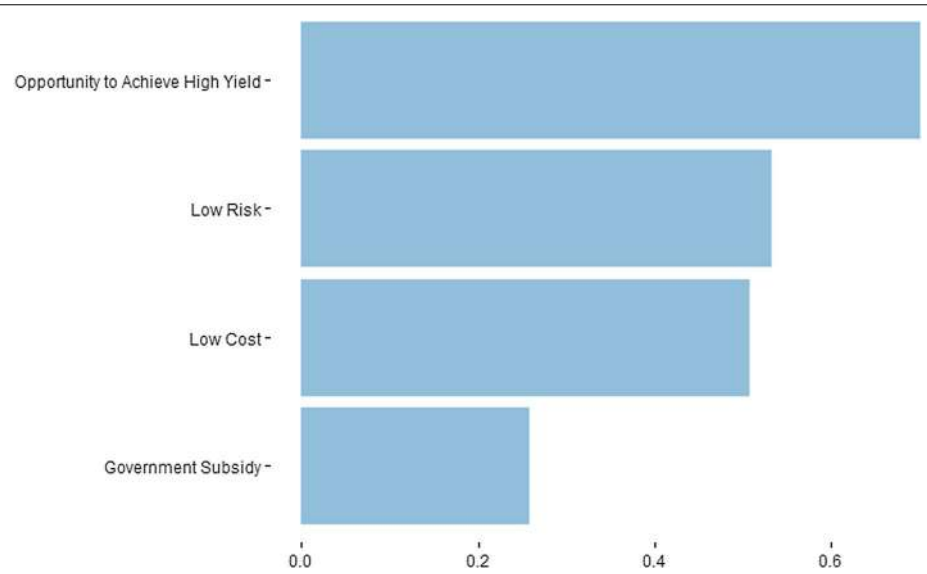

Fig. 10 Risk- and yield-related factors for investment decisions

retirement funds and other investment products, were one choice, while government bonds, individual shares, and unit-linked investment were the other (Fig. 9).

Important factors for your long-term savings and risk appetite

Respondents were asked to choose the factors that are important for long-term investment (multiple choice). Opportunity to achieve high yield was chosen as the most important factor (almost 70%), followed by low risk (Fig. 10). Interestingly, 29% of respondents chose low risk and high yield simultaneously, even though these options

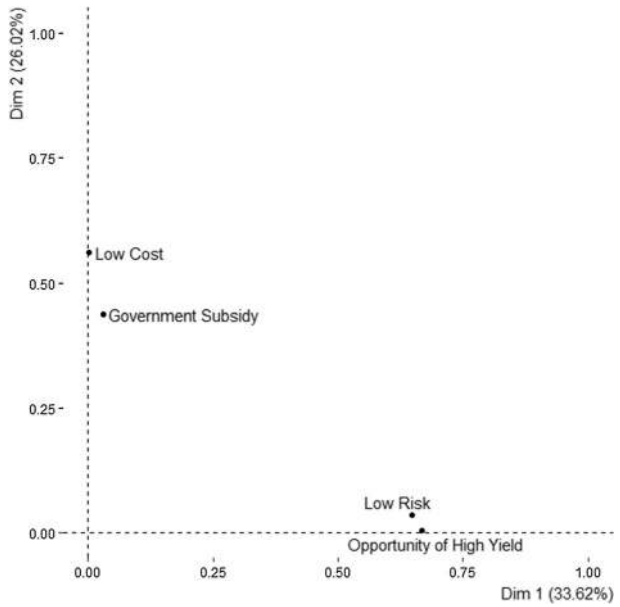

Fig. 11 MCA analysis of the risk- and yield-related factors for investment decisions

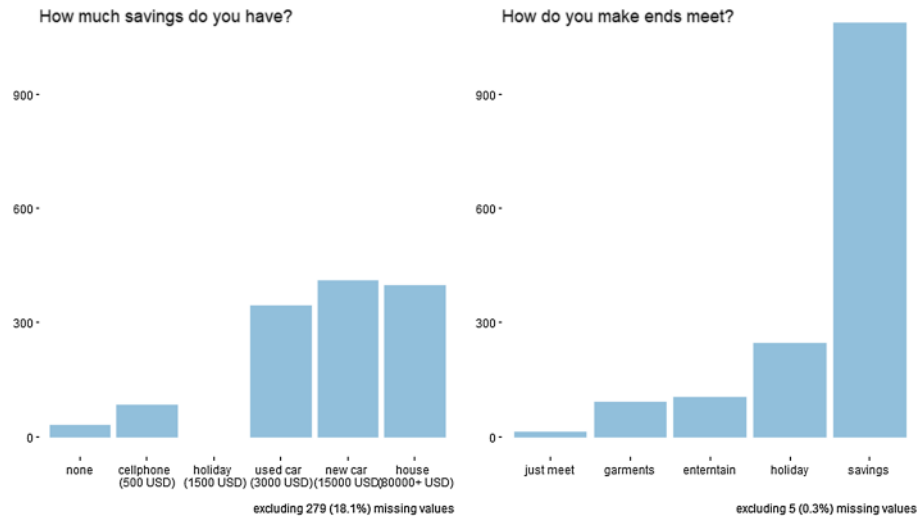

Fig. 12 Value and the ability to increase discretionary savings

are mutually exclusive. One explanation of this choice could be that they are seeking a balanced portfolio with multiple products and different risk profiles. The MCA analysis revealed that respondents look at investment cost and risk profile as two different factors: low cost and government subsidy (providing guaranteed income) are one set of factors, while low risk and high yield are the other (Fig. 11).

Savings and making ends meet

Respondents were asked to indicate how much savings they have and how they make ends meet. Most respondents have significant funds and could make discretionary

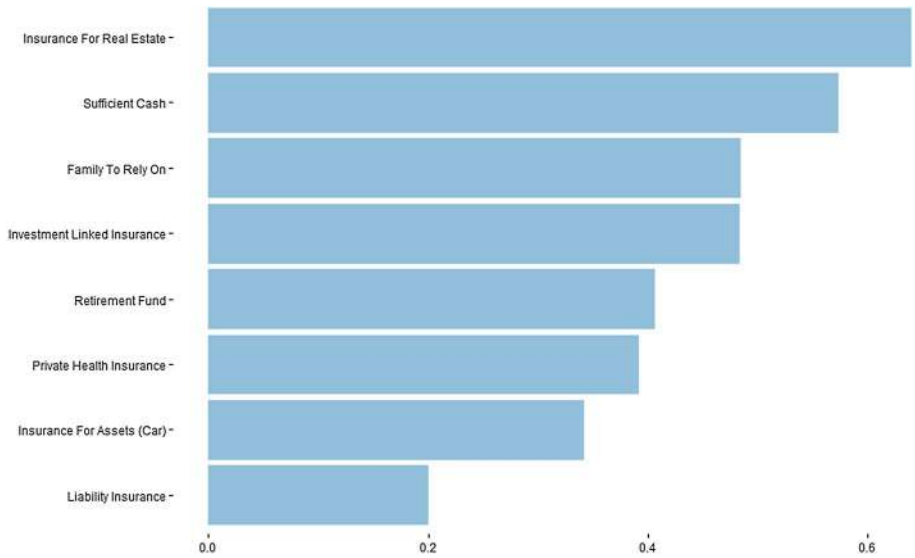

Fig. 13 Means to provide financial stability

Fig. 14 MCA analysis of means providing financial stability

savings at month's end, indicating that they could be targets for investment products (Fig. 12).

Means to provide financial stability

The next question was about different means to provide financial stability (multiple choice). Most respondents (64%) indicated having insurance for the real estate they own, followed by having sufficient cash (Fig. 13). The MCA analysis revealed two dimensions of the variables: the first being financial instruments (insurance products and retirement funds) and the second being more traditional support structures: family and cash (Fig. 14).

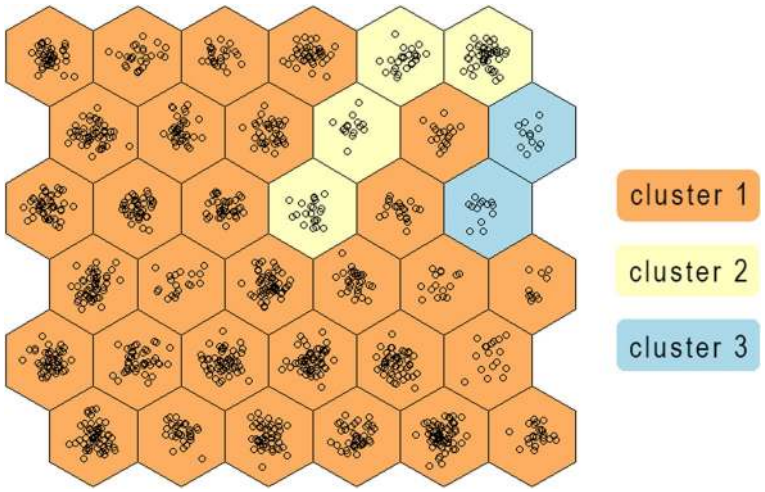

Fig. 15 Kohonen map with clusters of respondents

Clustering customers using Kohonen SOMs

Customer responses to the five selected questions were analyzed by building a 6×6 Kohonen self-organizing map with hexagonal topology. The neural network was trained using 100,000 iterations, followed by hierarchical clustering of the nodes. Using a neural network enables the use of both categorical and numerical variables, using Euclidean distance for the numerical variables and Hamming distance [63] for the categorical variables. As the output of the Kohonen map includes the distance matrix of the nodes, it enables hierarchical clustering. The optimal number of clusters was determined using the average silhouette approach [59, 64], suggesting three clusters of customers (Fig. 15). The validity of the method was checked by training the Kohonen SOM repeatedly and checking the differences, while the optimal number of clusters was checked by using the gap statistics [60] and the elbow method, suggesting similar values to the optimal numbers of clusters. The hierarchical clustered nodes of the Kohonen SOM were therefore cut into 3 main clusters, as suggested by the silhouette approach. There was a large cluster containing most respondents (89%) and two smaller clusters (9% and 2%).

The Kohonen map also makes it possible to analyze the clusters by individual variables used for creating the self-organizing map. In this way, we could gain insights about the groups to use when selecting products, offerings, or ways of communication with the clusters of customers. The following section describes the pertinent features of the clusters.

Insights gained from analyzing the clusters

Figure 15 shows the box-and-whisker plots for customer discretionary savings and funds they are left with after making ends met. Boxes represent the interquartile range, lines the minimum and maximum values without outliers, whiles dots the outliers. Respondents from Cluster 1 have significant funds and make ends meet with large discretionary savings. Cluster 2 respondents have little savings; however, they could still make ends meet left with sufficient funds; therefore, they could also be

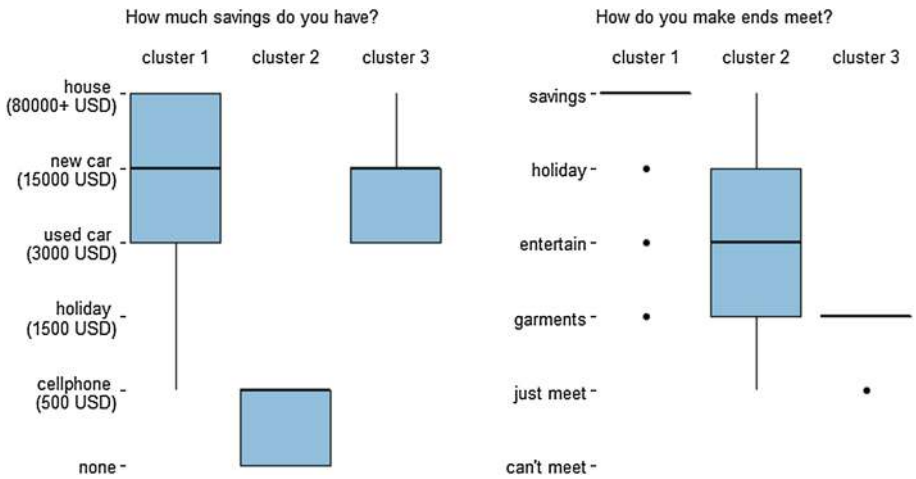

Fig. 16 Current funds and savings potential of clusters

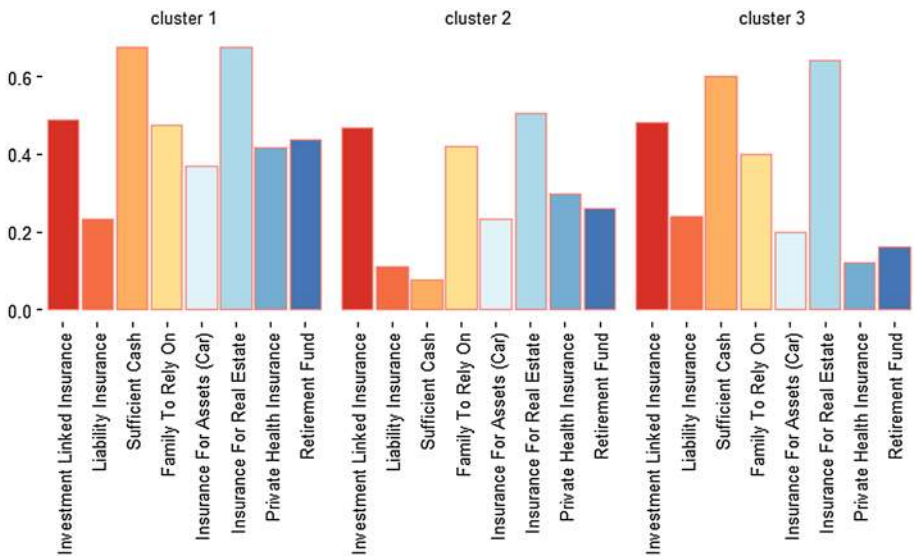

Fig. 17 Differences in how clusters think about financial stability

targets for investment products. Cluster 3 respondents, the smallest group, have significant savings but make ends meet with the smallest amount (Fig. 16).

When thinking about financial stability for the future, the difference between the clusters could be described: while Cluster 1 thinks about all instruments as almost equally important, Cluster 2 underplays the importance of cash and liability insurance, and Cluster 3 deemphasizes private health insurance and retirement funds (Fig. 17).

The approach toward investment risk and potential yield is rather similar to Clusters 1 and 2. Cluster 3 is different from these two clusters in that they underrate the

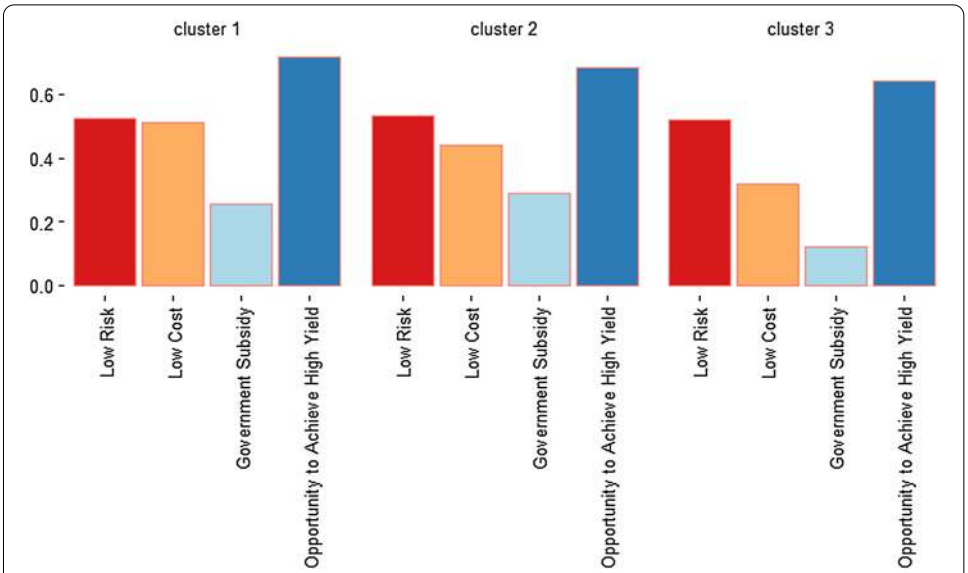

Fig. 18 Risk appetite and cost consciousness of the three clusters

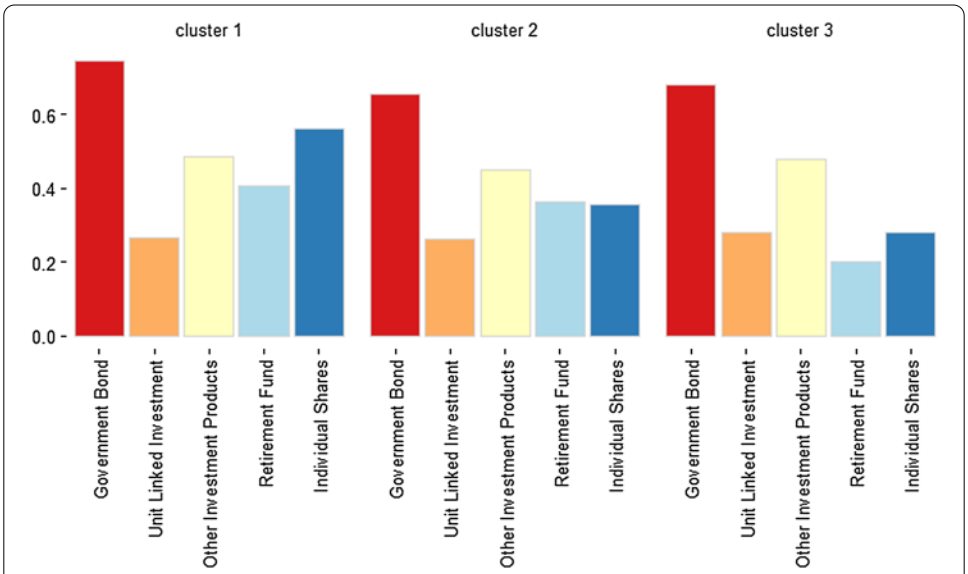

Fig. 19 How clusters think about appropriate investment products

value of government subsidies received on certain investment products and seem to be the least cost conscious (Fig. 18).

Finally, our two-stage clustering method enables us to describe what investment products the three clusters think are appropriate for them. All three clusters think about government bonds as an important element of their portfolio. Cluster 1 would overweight individual shares in their portfolio compared to the other two groups of potential customers, in line with their balanced approach to investment risks and potential yields. The difference between Clusters 2 and 3 is mainly how they think

about retirement funds, Cluster 3 underplaying its importance in line with their thinking about providing financial stability (Fig. 19).

Summary of insights about the clusters

In summary, the three clusters could be identified using the two-stage clustering method:

- The largest group of the most affluent customers, having significant savings and making ends meet easily, left with sufficient discretionary savings (Cluster 1), and thinking about financial stability using a multitude of means would be looking at a balanced portfolio of both government bonds with low risk and individual shares of potentially higher yield.
- The second group reported having little savings but still making ends meet with sizeable savings potential (Cluster 2). This group of potential customers may overlook the importance of insurance products when thinking about financial stability in the future and would be looking at an investment portfolio that is overweight in government bonds.
- The smallest group had significant savings but reported making ends meet with less disposable income than the other two groups (Cluster 3). This group underrates the importance of retirement funds and private health insurance when thinking about financial stability and wants to have the least number of individual shares and retirement funds in their portfolio.

In addition to these insights, multiple correspondence analysis highlighted that retirement funds and other investment products are perceived to be rather different from government bonds, individual shares, and unit-linked investment groups; therefore, their offering might require different communication strategies. Furthermore, risk and yield are perceived to be rather different from the cost of investment (including subsidies, which may be used to offset investment costs), which could also be used when designing communication protocols.

Conclusion

The main goal of this research was to demonstrate the effectiveness of the two-stage clustering method to explore and identify investment patterns in potential retail banking customers. The results confirmed that the method is effective in identifying distinct groups of customers, describing their investment patterns and investment factors. The unique feature of this research compared with the previous ones is the use of a new AI-related approach to targeting "potential customers". This was supported by the application of text analysis and a knowledge map in the literature review to reveal the characteristics of the research field. As a result of the literature review, we found that the ten important factors related to investments are financial, business decision-making, environmental, economic, market, foreign direct investment, management, risk, development, and investors. Similar to the literature review, our experimental results obtained from the analysis of respondents' data showed that the "risk" factor is an important factor for potential customers. The analysis of the semantic

relationship of basic concepts resulted in a knowledge map. The four main concepts of the knowledge map were investment, investment funds, savings, and retail banking (Fig. 1). The main contribution of this research is that MCA and Kohonen SOM have been combined for the clustering of potential customers. Other researchers applied Kohonen SOM and MCA as well, but they targeted different problems than ours. Elsäßer and Wirtz [65] examined the success factors of branding in a business-to-business setting and analyzed their performance impact on customer satisfaction and brand loyalty. Lamprinopoulou and Tregear [49] investigated the structure and content of network relations among SME clusters and explored the link to marketing performance. Albert et al. [48] used MCA for those respondents who expressed their love for a particular brand. He applied MCA to estimate the coordinates in a multi-dimensional space of the words that express the feeling of love. It is noteworthy that customer information was not examined here, but the data of ordinary respondents were examined, and based on this, future customers were predicted. We could identify three clusters of respondents, which were described by their current investment patterns and by the most important investment factors for the long term. Lai et al. [30] studied seven major effective factors on the decision-making underlying the R&D investment process along with treating R&D investment behavior. Hwang et al. [9] estimated the probability of customers' return using a machine learning approach on the received feedback comments and satisfaction ratings regarding the previous usage of the service. Higuchi and Maehara [38] used a factor-cluster analysis to cluster customers. Our investigation also resulted in appropriate investment patterns (funds and products) for potential investors. The analysis of the online investment questionnaire highlighted many important insights about the respondents as potential customers that could help to build better relationships with them and to offer more appropriate products to them. In our study, the experiences of respondents as potential customers were analyzed. Klaus and Maklan [21] provided the scale EXQ to measure customer experience quality. After one year, Klaus presented an updated customer experience quality (EXQ) scale that challenged the conceptualization and operationalization of customer experience. Kuppelwieser and Klaus [23] systematically explored the scale's psychometric properties and found that the EXQ scale comprises two or more dimensions rather than one. They explored the nature of the relationships between these dimensions and increased the understanding of the role in which customers experience quality in different research settings. In our study, analyzing the responses to individual questions showed that government bonds seem to be the most popular assets among the respondents when thinking about investment choices. However, they may also think of a portfolio of assets, wherein selected individual shares, other investment products, and retirement funds would play a role. Government bonds and individual shares or unit-linked investment would be one set of choices, while retirement funds or other investment products would be the second set of considerations when thinking about appropriate investment portfolios. Respondents think about the opportunity for high yield as the most important factor for investment, followed by the low risk and/or low cost associated with this opportunity. Most of them think about more than one factor. Low-risk and high-yield opportunities are understood as potentially mutually exclusive factors. Low cost and government-provided

subsidies are considered related factors. Regarding the respondents' current portfolio of investments, most of them claim they have 3 or more types of assets. Government bonds and individual shares are the most popular ones, followed by retirement funds and unit-linked investment products. When thinking about the measures to provide financial stability, insurance for the real estate they may own, sufficient cash savings, and family members they could rely on are considered the most important ones in this order.

Information about demographics was provided with many gaps and missing data, limiting its usability. However, those who provided information about themselves represented most age groups (with significant numbers from the 30- to 34-year-old group). It is also interesting that a significant portion of the respondents were employed as a "knowledge worker" ("graduate employee").

Kohonen SOM-based clustering using multiple questions of "Assets the family has to provide financial stability", "Savings the person has", "How to make ends meet", "Appropriate investment products", and "Important investment factors" resulted in three distinct groups of potential customers. When thinking about appropriate investment products, government bonds are always part of their preferred portfolio, with different mixes of other products varying by clusters. When thinking about risks and yield opportunities as factors of investment, the opportunity for high yield is always identified as an important one. The three clusters can be described: The first group of respondents or potential customers, which is the largest group, had significant savings (Cluster 1). They easily meet their needs. This group of customers maintains their financial stability by using various tools. To that end, they seek to have a balanced set of low-risk, high-yield individual government bonds. According to the second group, they had little savings but again they had the potential to meet their needs by saving (Cluster 2). This group of potential customers is likely to overlook the importance of insurance products and think of an overweight government bond investment portfolio. According to the smallest group, they had significant savings (Cluster 3). This group probably considers the importance of private pension and health insurance funds and is interested in having the least number of stocks and pension funds.

The unique features of this research are the mixed methodological approach in the analysis of investment patterns. The limitation of the study is that data were collected in Hungary and the respondents were financially aware; they were interested in financial issues. Future research includes the application of the combination of Kohonen SOM and MCA for investment pattern analysis. Designing a recommendation system based on the results can be a research project for the future.

Acknowledgements
This work was supported by the National Research, Development, and Innovation Fund of Hungary [grant number: 1.3.1-VKE-2018-00007 and GINOP- 2.2.1-18-2018-00010].

Authors' contributions
Authors have a contributions equal share in the study. All authors read and approved the final manuscript.

Funding
No Funding.

Availability of data and materials
Data was collected in partnership with the Corvinus University of Budapest, the Dorsum company and Portfolio in the 1.3.1-VKE-2018–00,007 project. According to the Consortia agreement and the Head of the project's consent, project data can be used for additional research and publications by the authors.

Declarations

Ethics approval and consent to participate
This article does not contain any studies with human or animal participants performed by any of the authors.

Consent of participation
Not applicable.

Competing interests
Work at the Corvinus University of Budapest helped design and develop the survey in conjunction with commercial companies (Dorsum and Portfolio).

Author details
[1]Corvinus University of Budapest, Fővám tér 8, 1093 Budapest, Hungary. [2]The Doctoral School of Economics, Business, and Informatics, Corvinus University of Budapest, Fővám tér 8, 1093 Budapest, Hungary.

Received: 23 July 2021 Accepted: 16 October 2021
Published online: 02 November 2021

References

1. Sabhaya RJ. An analysis of investment pattern of people during the period of 2018–19 in surat city. Int J Psychosoc Rehabil. 2020;24(6):1236–46.
2. Woodcock N, Green A, Starkey M. Social CRM as a business strategy. J Database Mark Cust Strategy Manag. 2011. https://doi.org/10.1057/dbm.2011.7.
3. Tejeda-Lorente Á, Bernabé-Moreno J, Herce-Zelaya J, Porcel C, Herrera-Viedma E. A risk-aware fuzzy linguistic knowledge-based recommender system for hedge funds. Proc CompSci. 2019. https://doi.org/10.1016/j.procs.2019.12.068.
4. Scheinbaum A, editor. Online consumer behavior: theory and research in social media, advertising, and e-tail. New York: Routledge/Taylor & Francis Group; 2012.
5. Arli D. Investigating consumer ethics: a segmentation study. JCM. 2017. https://doi.org/10.1108/JCM-08-2016-1908.
6. Barczak G, Ellen PS, Pilling BK. Developing typologies of consumer motives for use of technologically based banking services. J Bus Res. 1997. https://doi.org/10.1016/S0148-2963(96)00032-X.
7. Athanassopoulos AD. Customer satisfaction cues to support market segmentation and explain switching behavior. J Bus Res. 2000. https://doi.org/10.1016/S0148-2963(98)00060-5.
8. Persson A, Ryals L. Making customer relationship decisions: analytics v rules of thumb. J Bus Res. 2014. https://doi.org/10.1016/j.jbusres.2014.02.019.
9. Hwang S, Kim J, Park E, Kwon SJ. Who will be your next customer: a machine learning approach to customer return visits in airline services. J Bus Res. 2020. https://doi.org/10.1016/j.jbusres.2020.08.025.
10. Loureiro SMC, Guerreiro J, Tussyadiah I. Artificial intelligence in business: state of the art and future research agenda. J Bus Res. 2020. https://doi.org/10.1016/j.jbusres.2020.11.001.
11. Calderon-Monge E, Pastor-Sanz I, Sendra Garcia FJ. Analysis of sustainable consumer behavior as a business opportunity. J Bus Res. 2020. https://doi.org/10.1016/j.jbusres.2020.07.039.
12. V. P. Semenov, V. v. Chernokulsky, and N. v. Razmochaeva, Research of artificial intelligence in the retail management problems. 2017. doi: https://doi.org/10.1109/CTSYS.2017.8109560.
13. Soltani-Fesaghandis G, Pooya A. Design of an artificial intelligence system for predicting success of new product development and selecting proper market-product strategy in the food industry. Int Food Agribusiness Manag Rev. 2018;21:847–64.
14. Burez J, van den Poel D. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009. https://doi.org/10.1016/j.eswa.2008.05.027.
15. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982. https://doi.org/10.1007/BF00337288.
16. Kohonen T. The self-organizing map. Proc IEEE. 1990;78(9):1464–80. https://doi.org/10.1109/5.58325.
17. S. Sinclair and G. Rockwell, "Voyant Tools," 2016. http://voyant-tools.org/. Accessed 05 Jan 2021.
18. "Yewno," 2020. https://discover.yewno.com/. Accessed 05 Jan 2021.
19. A. C. Edmondson and S. E. Mcmanus, Methodological fit in management field research, 2007. [Online]. https://www.jstor.org/stable/20159361
20. Keiningham T, et al. Customer experience driven business model innovation. J Bus Res. 2020. https://doi.org/10.1016/j.jbusres.2019.08.003.
21. Maklan S, Klaus P. Customer Experience: Are We Measuring the Right Things? Int J Market Res. 2011. https://doi.org/10.2501/IJMR-53-6-771-792.
22. Klaus P. Customer experience: the origins and importance for your business. In: Klaus P, editor. Measuring customer experience. London: Palgrave Macmillan UK; 2015. p. 1–21 (**10.1057/9781137375469_1**).
23. Kuppelwieser VG, Klaus P. Measuring customer experience quality: the EXQ scale revisited. J Bus Res. 2020. https://doi.org/10.1016/j.jbusres.2020.01.042.
24. L. Wewege and M. C. Thomsett, The digital banking revolution: how fintech companies are transforming the retail banking industry through disruptive financial innovation. Walter de Gruyter GmbH & Co KG, 2019. https://doi.org/10.1515/9781547401598
25. de Marco M, Fantozzi P, Fornaro C, Laura L, Miloso A. Cognitive analytics management of the customer lifetime value: an artificial neural network approach. JEIM. 2021. https://doi.org/10.1108/JEIM-01-2020-0029.

26. Fatima A, Sharma JK. Segmenting Investors on their Biases Manifested in Investment Decision-Making by Individual Investors. SJOM. 2021. https://doi.org/10.33215/sjom.v4i4.663.
27. J. Jääskeläinen, "Segmentation of investor customers using machine learning in banking," Lappeenranta, 2021 [Online]. http://urn.fi/URN:NBN:fi-fe2021051730210. Accessed 09 Sep 2021.
28. Mishra KC, Metilda MJ. A study on the impact of investment experience, gender, and level of education on overconfidence and self-attribution bias. IIMB Manag Rev. 2015. https://doi.org/10.1016/j.iimb.2015.09.001.
29. Aren S, Aydemir SD. The factors influencing given investment choices of individuals. Proc Soc Behav Sci. 2015. https://doi.org/10.1016/j.sbspro.2015.11.351.
30. Lai Y-L, Lin F-J, Lin Y-H. Factors affecting firm's R&D investment decisions. J Bus Res. 2015. https://doi.org/10.1016/j.jbusres.2014.11.038.
31. Kuhnen CM, Miu AC. Socioeconomic status and learning from financial information. J Financ Econ. 2017. https://doi.org/10.1016/j.jfineco.2017.03.002.
32. Plath DA, Stevenson TH. Financial services consumption behavior across Hispanic American consumers. J Bus Res. 2005. https://doi.org/10.1016/j.jbusres.2004.03.003.
33. Shim G, Lee S, Kim Y. How investor behavioral factors influence investment satisfaction, trust in investment company, and reinvestment intention. J Bus Res. 2008. https://doi.org/10.1016/j.jbusres.2006.05.008.
34. T. Zhang, X. Huang, J. Tang, and X. Luo, Case study on cluster analysis of the telecom customers based on consumers' behavior. 2011. doi: https://doi.org/10.1109/ICIEEM.2011.6035407.
35. Oprescu AE. The strategic marketing planning—general framework for customer segmentation. Ann Spiru Haret Univ Econ Ser. 2014;5(1):51–9.
36. R. Ait Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, Combining RFM model and clustering techniques for customer value analysis of a company selling online. 2015. doi: https://doi.org/10.1109/AICCSA.2015.7507238.
37. Dhawan D, Mehta SK. Saving and investment pattern: assessment and prospects. ACRN J Finance Risk Perspect. 2019. https://doi.org/10.35944/jofrp.2019.8.1.008.
38. Higuchi A, Maehara R. A factor-cluster analysis profile of consumers. J Bus Res. 2021. https://doi.org/10.1016/j.jbusres.2020.09.030.
39. P. Goncarovs, Using data analytics for customers segmentation: experimental study at a financial institution. 2018. doi: https://doi.org/10.1109/ITMS.2018.8552951.
40. Santos RS, Qin L. Risk capital and emerging technologies: innovation and investment patterns based on artificial intelligence patent data analysis. JRFM. 2019. https://doi.org/10.3390/jrfm12040189.
41. Boone DS, Roehm M. Retail segmentation using artificial neural networks. Int J Mark Res. 2002. https://doi.org/10.1016/S0167-8116(02)00080-0.
42. Ying Li and Feng Lin, Customer segmentation analysis based on SOM clustering. 2008. doi: https://doi.org/10.1109/SOLI.2008.4686353.
43. Y. Li, Y. Wu, and F. Lin, Research on Customer Segmentation Based on a Two-Stage SOM Clustering Algorithm. 2009. doi: https://doi.org/10.1109/ICMSS.2009.5302076.
44. Bigné E, Aldas-Manzano J, Küster I, Vila N. Mature market segmentation: a comparison of artificial neural networks and traditional methods. Neural Comput App. 2010. https://doi.org/10.1007/s00521-008-0226-y.
45. Mak MKY, Ho GTS, Ting SL. A financial data mining model for extracting customer behavior. Int J Eng Bus Manag. 2011. https://doi.org/10.5772/50937.
46. Saluja MS, Shaikh Y. Decoding investment pattern of fiis and diis in indian stock market using decision tree. IJACR. 2017;8:3.
47. Chen T-H, Ho R-J, Liu Y-W. Investor personality predicts investment performance? A statistics and machine learning model investigation. Comput Hum Behav. 2019. https://doi.org/10.1016/j.chb.2018.09.027.
48. Albert N, Merunka D, Valette-Florence P. When consumers love their brands: exploring the concept and its dimensions. J Bus Res. 2008. https://doi.org/10.1016/j.jbusres.2007.09.014.
49. Lamprinopoulou C, Tregear A. Inter-firm relations in SME clusters and the link to marketing performance. J Bus Ind Mark. 2011. https://doi.org/10.1108/08858621111156412.
50. Lam D, Wei M, Wunsch D. Clustering data of mixed categorical and numerical type with unsupervised feature learning. IEEE Access. 2015;3:1605–16. https://doi.org/10.1109/ACCESS.2015.2477216.
51. Guha S, Rastogi R, Shim K. Rock: a robust clustering algorithm for categorical attributes. Inf Syst. 2000;25(5):345–66. https://doi.org/10.1016/S0306-4379(00)00022-3.
52. V. Ganti, J. Gehrke, and R. Ramakrishnan, CACTUS---clustering categorical data using summaries. 1999. doi: https://doi.org/10.1145/312129.312201.
53. He Z, Xu X, Deng S. Clustering mixed numeric and categorical data: a cluster ensemble approach; 2005. arXiv:cs/0509011.
54. Kuo RJ, Ho LM, Hu CM. Integration of self-organizing feature map and K-means algorithm for market segmentation. Comput Oper Res. 2002;29(11):1475–93. https://doi.org/10.1016/S0305-0548(01)00043-0.
55. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Core Team; 2020. https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing. Accessed 19 Dec 2020.
56. A. Kassambara and F. Mundt, "factoextra: Extract and Visualize the Results of Multivariate Data Analyses," Apr. 01, 2020. http://www.sthda.com/english/rpkgs/factoextra. Accessed 19 Dec 2020.
57. Wehrens R, Kruisselbrink J. Flexible self-organizing maps in kohonen 3.0. J Stat Softw. 2018;87(7):1–18. https://doi.org/10.18637/jss.v087.i07.
58. Wehrens R, Buydens LMC. Self- and super-organizing maps in R: The kohonen package. J Stat Softw. 2007;21(5):1–19. https://doi.org/10.18637/jss.v021.i05.
59. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987. https://doi.org/10.1016/0377-0427(87)90125-7.
60. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Series B Stat Methodol. 2001. https://doi.org/10.1111/1467-9868.00293.
61. "portfolio.hu traffic overview." https://www.similarweb.com/website/portfolio.hu/. Accessed 31 Jan 2021.

62. "portfolio.hu conferences." https://www.portfolio.hu/en/events. Accessed 31 Jan 2021.
63. Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J. 1950. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.
64. Kaufman L, Rousseeuw PJ. Partitioning around medoids (program pam). Find Grp Data. 1990;344:68–125.
65. Elsäßer M, Wirtz BW. Rational and emotional factors of customer satisfaction and brand loyalty in a business-to-business setting. JBIM. 2017. https://doi.org/10.1108/JBIM-05-2015-0101.

Publisher's Note