THE JOURNAL OF INDONESIA SUSTAINABLE DEVELOPMENT PLANNING

VOL. 7 NO. 1 – APRIL 2026

E-ISSN: 2722-0842 | P-ISSN: 2721-8309

Available online at
http://journal.pusbindiklatren.bappenas.go.id/

Research Paper

Segmentation of Adult Respondents' Well-being Profiles Based on Daily Stress, Social Networks, Personal Resources, and Lifestyle Using Clustering Method

Ajang Sopandi[1], Siti Ummi Masruroh[2*], Neneng Tati Sumiati[3] Cindy Rahayu[4], Rona Nisa Sofia Amriza[5], and Doni Febrian[6]

[1]Faculty of Engineering, University of Muhammadiyan Tangerang, Indonesia
[2,6]Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta, Indonesia
[3]Faculty of Psychology, Syarif Hidayatullah State Islamic University Jakarta, Indonesia
[4]School of Computer Science Bina Nusantara University Jakarta, Indonesia
[5]Information System of Telkom University Bandung, Indonesia

***Correspondence author**: ummi.masruroh@uinjkt.ac.id*

Abstract

Existing literature on the topic of wellbeing mostly utilized scales and methods that are abstract and variable-centered yet assume homogeneity within the population being studied. This research utilizes a person-centered approach to classify the sample of 15,977 adults from a large-scale online survey about their wellbeing according to variables related to their stress, social networks, personal resources, and lifestyle. Factor analysis of mixed data (FAMD) is performed to reduce 22 variables of mixed types to 14 principal components that account for 77.51% of the variance in the data. Using these components, eight segments of well-being are classified by K-Means clustering and validated using Silhouette analysis. These segments range from those with low levels of stress, high levels of meditation, and clear goals for their lives to those with high levels of stress, no sense of accomplishment in their careers, and few social connections outside of work. Interestingly, another variable that was revealed as significantly different within each of the stress levels groups was the notion of whether or not the individual feels like they have enough money to cover their needs. Finally, the methods used in this research can be replicated to evaluate the wellbeing of the general population and to inform the creation of interventions to improve the lives of those with certain types of wellbeing profiles.

Keywords: Wellbeing Segmentation; Clustering; Factor Analysis of Mixed Data; K-Means; Person-centered

ARTICLE INFO
Received: May 23, 2025
Received in revised form: November 14, 2025
Accepted: April 30, 2026

doi: 10.46456/jisdep.v7i1.1046

This is an open access article under the CC BY-SA license
©Sopandi *et al* (2026)

THE JOURNAL OF INDONESIA SUSTAINABLE DEVELOPMENT PLANNING
Published by Centre for Planners' Development, Education, and Training (Pusbindiklatren), Ministry of National Development Planning/National Development Planning Agency (Bappenas), Republic of Indonesia

Address: Jalan Proklamasi 70, Central Jakarta, Indonesia 10320
Phone: +62 21 31928280/31928285
Fax: +62 21 31928281
E-mail: journal.pusbindiklatren@bappenas.go.id

Supported by Indonesian Development Planners Association (PPPI)

Please cite this article in APA Style as:
Sopandi, A., Masruroh, S. U., Sumiati, N.T., Rahayu, C., Amriza, N. S., & Febarian, D. (2026). Segmentation of Adult Respondents' Well-being Profiles Based on Daily Stress, Social Networks, Personal Resources, and Lifestyle Using Clustering Method. *The Journal of Indonesia Sustainable Development Planning, Vol 7(1),* 140-153. https://doi.org/10.46456/jisdep.v7i1.1046

1. Introduction

The concept of well-being encompasses multiple meanings, ranging from the health of an individual's physical body to their mental and social health. In order to appropriately evaluate the degree of progress that is made by a society or its individuals, it is necessary to consider each of these different facets of well-being (UNESCO, 2022). To better comprehend the well-being of a person in the present day, it is necessary to incorporate metrics that evaluate the daily lives of those individuals. Therefore, a study that considers each of these different variables, such as those related to an individual's physical lifestyle, social relationships, personal resources, and daily stressors, will help to gain an understanding of the biological, psychological, and social well-being of those individuals. Consequently, an understanding of how each of these different components of an individual's health and well-being interacts with each other could be utilized to create holistic profiles of each individual.

While there is an understanding of each of these different variables that contribute to the well-being of an individual, there is a problem in the methods that are traditionally utilized for analyzing it. For instance, many research studies utilize methods that evaluate the independent effects of each of those variables, without considering the holistic and often non-linear effects that those variables have upon the well-being of those individuals (M. Wu et al., 2024). Furthermore, most of the methods that are utilized in these studies are based upon the assumption of population homogeneity, which ignores any subpopulations within those groups, whose well-being is affected differently as a result of habits, stressors, lifestyle or environment. Thus, a shift in the methods for evaluating the health of individuals should also shift towards a person-centered approach that considers various behavioral and environmental variables (Andretta & McKay, 2020).

The necessity of utilizing such a person-centered analytical paradigm is supported by several recent publications in the existing literature. For instance, Leite et al. published a study utilizing data from 35,936 individuals from Europe and performed cluster analysis on the data to determine the distribution of individuals according to their values within Schwartz's human value priorities. As a result, the authors identified four main groups within the population, each with a different focus on social or growth values. The authors found that individuals had higher levels of well-being and greater social life integration within within groups with focus on growth values (Leite et al., 2021). These findings support the concept that an individual's well-being is dependent upon their own formed system of values and motivations. Additional support for this paradigm was provided by Lazić et al. in a study with 945 undergraduate students as sample. The authors utilized latent profile analysis to investigate the internal structure of the students' well-being. Consequently, they identified four profiles within the students according to their levels of well-being (Low, mixed, moderately low, and high), and later validated those segments by correlating them with emotional distress and emotional regulation strategies they used. Students with high levels of well-being used predominantly adaptive strategies for emotional regulation, while those with low and mixed levels of well-being employed primarily maladaptive and even destructive strategies for emotional regulation (Lazić et al., 2021).

Beyond applying such analyses to the general population, studies can also be performed on specific societal demographics. For instance, Pasanen et al. studied individuals who live alone in Finland. The authors used latent class analysis to assess the health profiles of 884 adults who live alone in the country. They classified the population according to their physical, social, and mental health and revealed four different health profiles (Languishing, Managing, Healthy, and Flourishing). However, they also found that individuals within certain health profiles have robust mental and social health yet poor physical health (Pasanen et al., 2021). The importance of incorporating physical health considerations into the formation of health profiles is further supported by Hennessey et al. in their study of 366 adolescents. Using latent profile analysis, they identified three profiles according to their levels of well-being. For one, participants showed a higher level of well-being when linked to a higher volume of physical activities, displaying stronger capability of emotional literacy, and a greater sense of belongings in a school environment. These findings suggest the need to incorporate physical activity considerations within any profile analysis of well-being (Hennessey et al., 2024).

Despite the theoretical insights produced by these studies, a synthesis of the current literature indicates a few notable gaps in the field. The majority of studies that have been published in relation to well-being have utilized scales that are relatively abstract and general in their description of the well-being of the individuals being surveyed. Thus, existing literature has a lack of effort to bridge the gap

between psychological well-being and daily habits. There has also been a lack of research investigating how individuals can be segmented according to both their daily behaviors as well as some of their biological and socioeconomic factors.

Finally, the majority of current studies that aim to segment populations according to their well-being utilize methods like Latent Profile Analysis or Latent Class Analysis. Both of these methods rely upon a number of strict assumptions of the data, particularly the assumption of local conditional independence of the individuals' responses to the survey questions. Because these assumptions are typically violated by real-world data sets, a variety of issues can emerge from utilizing these methods. By contrast, K-means clustering provide a deterministic and non-parametric alternative to these techniques that resolves the issue of calculating distances between individuals. However, K Means Clustering tends to be invalid and flawed when the data contains mixed data types (Ufeli et al., 2025). One of the solutions to this problem is by introducing Factor Analysis of Mixed Data (FAMD) as an advanced dimensionality reduction technique to synthesize the heterogeneous mix of continuous and categorical survey variables. FAMD mathematically integrates Principal Component Analysis for numerical data and Multiple Correspondence Analysis for categorical data, creating a harmonized, unified analytical framework without sacrificing the underlying structural relationships of the original variables (Singh et al., 2024). The ultimate objective is to apply the highly efficient K-means clustering algorithm directly to these optimized, continuous principal coordinates to partition the adult population into robust segments.

These gaps in the current literature indicates the importance of the current research endeavor. By segmenting individuals according to a variety of daily habits and factors related to their well-being, this research will provide an actionable taxonomy of the different forms of well-being among adults. Furthermore, each of these segmented groups will provide an understanding of the behaviors that typically define the well-being of individuals in each group, which may help to reveal some non-obvious behavior configurations that are associated with benefits like reduced stress or improved socioeconomic outcomes. Additionally, by establishing an algorithm based upon methods like FAMD and K-means clustering, researchers will gain an efficient method of analyzing well-being surveys that incorporate various and mixed data types. Finally, by bridging the gap between deterministic machine learning algorithms and the complexity of human psychological states and traits, this study may provide insights into the structure of human psychology that can be used to foster well-being among society's individuals.

2. Methods

To discover the meaningful segments of individuals experiencing well-being within the survey's adult respondents, the Knowledge Discovery in Databases (KDD) process will be employed. The KDD process is a framework that helps to extract meaningful patterns from data. The framework consists of five primary stages: Data Selection, Data Cleaning and Preprocessing, Data Transformation, Data Mining, and Interpretation (Palacios et al., 2021). Each of these stages will be performed to account for the multidimensional data within the survey.

Data Selection

The dataset consists of the responses from the website www.Authentic-Happiness.com. The data from this study were collected to evaluate the various ways that individuals choose to structure their lifestyles and their behaviors to maximize their satisfaction with the lives that they live.

The dataset consists of 22 different variables that describe the lives of adult individuals. These variables include measures of physical and lifestyle factors (such as steps taken per day, intake of fruits and vegetables, hours of sleep per day, and body mass index), social network factors (such as the size of their social circle and the number of social interactions per day), personal factors (such as having a vision for their life, number of hours spent in a state of "flow" in the past seven days, and the number of minutes spent meditating per day), and stress-related factors (such as the stress level experienced daily, the number of vacation days lost due to stress, and the number of days it takes for an individual to feel like they have sufficient income to cover their financial needs). These variables consist of both continuous and categorical variables.

Data Cleaning

Data cleaning is essential for ensuring the accuracy and reliability of the data set by reducing the errors and handling unexpected values. A few measure that can be applied in this step are:

1. Missing values can be replaced with the mean or the most probable value in the data set.
2. Noisy data can be smoothed using techniques such as binning, regression, and clustering.
3. Duplicates in the data can be removed.

Data cleaning is crucial for enhancing the quality of the data and enhancing the effectiveness of data mining (W.-T. Wu et al., 2021).

Data Transformation

Given the high number of variables (22) involved in this project, there is a potential risk of encountering the "curse of dimensionality." To resolve this issue, Factor Analysis of Mixed Data (FAMD) is utilized. FAMD is an advanced pre-processing technique that generalizes the concepts of PCA to accommodate both numerical and categorical variables. More specifically, the principles of PCA are applied to the continuous variables to extract the variance within the group, while the principles of MCA are applied to the categorical variables to extract the inertia within that group (Delimiro Visbal-Cadavid, 2020). Each of these variables is weighted to ensure that neither variable type dominates the other. The FAMD procedure involves:

1. Standardizing numerical variables to a mean of 0 and variance of 1.
2. Applying One-hot encoding to categorical variables and scaling them based on their frequency.
3. Calculating principal components that maximize the explained inertia (variance) across both data types.

Reducing the 22 variables to principal components will help to address the curse of dimensionality and aid in the interpretability of the results of the algorithm. Furthermore, it may allow for components to be created that represent certain concepts within the data set, such as a component that represents Holistic Health that incorporates variables like FRUITS_VEGGIES, DAILY_STEPS, and BMI_RANGE, or a component that represents social interactions that includes variables like CORE_CIRCLE and SUPPORTING_OTHERS.

Data Mining

With the data successfully transformed into a continuous space, the core data mining step is executed using the K-means clustering algorithm. The K-means algorithm partitions the respondents into distinct, non-overlapping well-being profiles by grouping data points around central centroids. The algorithm optimizes these clusters by iteratively minimizing the Within-Cluster Sum of Squares (WCSS), mathematically represented as:

$$WCSS = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where K represents the total number of clusters, C_k represents the set of data points assigned to cluster K, μ is the centroid of cluster K, and $||x_i-\mu_k||$ denotes the Euclidean distance between a specific data point x_i and its respective centroid (Chong, 2021). To determine the optimal number of clusters (K), two popular graph-based evaluation techniques such as the Silhouette Score and the Elbow Method will be employed to assess the cohesion and separation of the resulting segments (Saputra et al., 2020).

Evaluation and Interpretation

The final stage of the KDD process involves utilizing the discovered patterns to extract knowledge. The clusters are evaluated based upon the values of their centroids across the original 22 survey variables. Through analyzing the different levels of each variables within each cluster, it is possible to formulate holistic and descriptive profiles of the well-being of adults within each cluster. Thus, insights can be extracted regarding the wellbeing (or lack thereof) of adult humans according to these discovered clusters.

3. Results and Discussions

Data Selection

The dataset consists of 15,977 survey responses with 24 attributes. Most of the columns are of numerical types, with five of them being categorical values.

Table 1. Dataset Columns

Column name	Data type	Description
TIMESTAMP	Numerical	Date when survey was completed
FRUITS_VEGGIES	Numerical	HOW MANY FRUITS OR VEGETABLES DO YOU EAT EVERYDAY? (counts)
DAILY_STRESS	Categorical	HOW MUCH STRESS DO YOU TYPICALLY EXPERIENCE EVERYDAY? (Stress Level 1-5)
PLACES_VISITED	Numerical	HOW MANY NEW PLACES DO YOU VISIT? (counts)
CORE_CIRCLE	Numerical	HOW MANY PEOPLE ARE VERY CLOSE TO YOU? (counts)
SUPPORTING_OTHERS	Numerical	HOW MANY PEOPLE DO YOU HELP ACHIEVE A BETTER LIFE? (counts)
SOCIAL_NETWORK	Numerical	HOW MANY PEOPLE DO YOU INTERACT WITH DURING A TYPICAL DAY? (counts)
ACHIEVEMENT	Numerical	HOW MANY REMARKABLE ACHIEVEMENTS ARE YOU PROUD OF? (counts)
DONATION	Numerical	HOW MANY TIMES DO YOU DONATE YOUR TIME OR MONEY TO GOOD CAUSES? (counts)
BMI_RANGE	Numerical	WHAT IS YOUR BODY MASS INDEX (BMI) RANGE? (BMI Range, with 1 is below the index of 25, and 2 is larger than 25)
TODO_COMPLETED	Numerical	HOW WELL DO YOU COMPLETE YOUR WEEKLY TO-DO LISTS? (counts of completed weekly tasks)
FLOW	Numerical	HOW MANY HOURS DO YOU EXPERIENCE "FLOW"? (hours) IN A TYPICAL DAY
DAILY_STEPS	Numerical	HOW MANY STEPS (IN THOUSANDS) DO YOU TYPICALLY WALK EVERYDAY? (in thousand steps per week)
LIVE_VISION	Numerical	FOR HOW MANY YEARS AHEAD IS YOUR LIFE VISION VERY CLEAR FOR? (years)
SLEEP_HOURS	Numerical	ABOUT HOW LONG DO YOU TYPICALLY SLEEP? (hours)
LOST_VACATION	Numerical	HOW MANY DAYS OF VACATION DO YOU TYPICALLY LOSE EVERY YEAR ?(counts)
DAILY_SHOUTING	Numerical	HOW OFTEN DO YOU SHOUT OR SULK AT SOMEBODY? (counts per months)
SUFFICIENT_INCOME	Categorical	HOW SUFFICIENT IS YOUR INCOME TO COVER BASIC LIFE EXPENSES? (yes or no question)
PERSONAL_AWARDS	Numerical	HOW MANY RECOGNITIONS HAVE YOU RECEIVED IN YOUR LIFE? (counts)
TIME_FOR_PASSION	Numerical	HOW MANY HOURS DO YOU SPEND EVERYDAY DOING WHAT YOU ARE PASSIONATE ABOUT? (counts)
WEEKLY_MEDITATION	Numerical	HOW MANY TIMES DO YOU HAVE THE OPPORTUNITY TO SELF-REFLECT IN A TYPICAL WEEK (counts)
AGE	Categorical	AGE GROUPS (eg. 21 to 35, 36 to 50, etc)
GENDER	Categorical	MALE OR FEMALE DATA
WORK_LIFE_BALANCE_SCORE	Numerical	SCORE CALCULATED BY AH.COM ALGORITHM

The key features and summaries of the available data are then summarized to count the mean and range of the data.

Table 2. Dataset Summary

Column name	Summary
TIMESTAMP	2015-2021
FRUITS_VEGGIES	Range: 0-5, Mean: 2.92, Std Deviation: 1.44
DAILY_STRESS	Range: 0-5, Mean: 2.79, Std Deviation: 1.37
PLACES_VISITED	Range: 0-10, Mean: 5.23, Std Deviation: 3.31
CORE_CIRCLE	Range: 0-10, Mean: 5.51, Std Deviation: 2.84
SUPPORTING_OTHERS	Range: 0-10, Mean: 5.62, Std Deviation: 3.24
SOCIAL_NETWORK	Range: 0-10, Mean: 6.47, Std Deviation: 3.09

Column name	Summary
ACHIEVEMENT	Range: 0-10, Mean: 4, Std Deviation: 2.76
DONATION	Range: 0-10, Mean: 2.72, Std Deviation: 1.85
BMI_RANGE	Yes: 58.9%, No: 41.1%
TODO_COMPLETED	Range: 0-10, Mean: 5.75, Std Deviation: 2.62
FLOW	Range: 0-10, Mean: 3.19, Std Deviation: 2.36
DAILY_STEPS	Range: 0-10, Mean: 5.7, Std Deviation: 2.89
LIVE_VISION	Range: 0-10, Mean: 3.75, Std Deviation: 3.23
SLEEP_HOURS	Range: 3-10, Mean: 7.04, Std Deviation: 1.2
LOST_VACATION	Range: 0-10, Mean: 2.9, Std Deviation: 3.69
DAILY_SHOUTING	Range: 0-10, Mean: 2.93, Std Deviation: 2.68
SUFFICIENT_INCOME	Yes: 72.9%, No: 27.1%
PERSONAL_AWARDS	Range: 0-10, Mean: 5.71, Std Deviation: 3.09
TIME_FOR_PASSION	Range: 0-10, Mean: 3.33, Std Deviation: 2.73
WEEKLY_MEDITATION	Range: 0-10, Mean: 6.23, Std Deviation: 3.02
AGE	21-35: 38.2%, 36-50: 29.1%, 51 or More: 21.2%, Less than 20: 11.4%
GENDER	Female: 61.7%, Male: 38.3%
WORK_LIFE_BALANCE_SCORE	Range: 480-820, Mean: 667, Std Deviation: 45

Timestamp and WORK_LIFE_BALANCE_SCORE was removed due to being irrelevant, and missing a value most of the time

Data Cleaning

During the initial analysis of the dataset, it has been found that there were none missing values besides the timestamp column. However, some of the timestamp data were populating the FRUIT_VEGIES column of the dataset. Said data was removed.

Data Transformation

The dataset was first fitted using a full 22-component FAMD to examine the variance explained by each component. The resulting eigenvalue decomposition is summarized in Table 3.

Table 3. Eigenvalues Summary

Component	Eigenvalue	% of variance	% of variance(Cumulative)
0	4.755	10.69%	10.69%
1	3.34	7.50%	18.19%
2	2.861	6.43%	24.62%
3	2.647	5.95%	30.57%
4	2.585	5.81%	36.38%
5	2.519	5.66%	42.04%
6	2.503	5.63%	47.66%
7	2.472	5.56%	53.22%
8	2.377	5.34%	58.56%
9	2.267	5.09%	63.65%
10	2.071	4.65%	68.31%
11	1.871	4.21%	72.51%
12	1.174	2.64%	75.15%
13	1.047	2.35%	77.51%
14	0.963	2.16%	79.67%
15	0.898	2.02%	81.69%
16	0.875	1.97%	83.65%
17	0.821	1.85%	85.50%
18	0.778	1.75%	87.25%
19	0.738	1.66%	88.91%
20	0.716	1.61%	90.52%
21	0.704	1.58%	92.10%

The first component alone accounted for 10.69% of the total variance of the dataset, the highest percentage amongst all the components. The first six components accounted for 42.04% of the variance

of the dataset, while the 21 components within the optimized FAMD model accounted for 92.05% of the variance of the dataset.

General rule of thumb that is often applied to factor analysis is that any components with an eigenvalue less than 1 should be discarded (The Kaiser Rule) (Grippo et al., 2021). An eigenvalue of 1 indicates that a component is explaining the same amount of information as all of the original variables in the dataset combined. Thus, any component that has an eigenvalue less than 1 is explaining less information than the raw variables alone, indicating that such a component is somewhat useless.

Considering both the eigenvalues and the percentage of variance of the components, it is determined that an optimized FAMD model should have 14 components. This model accounts for 77.51% of the variance of the dataset.

Data Mining

Two complementary methods were employed to identify the optimal number of clusters for K-Means: the Elbow Method and Silhouette Analysis.

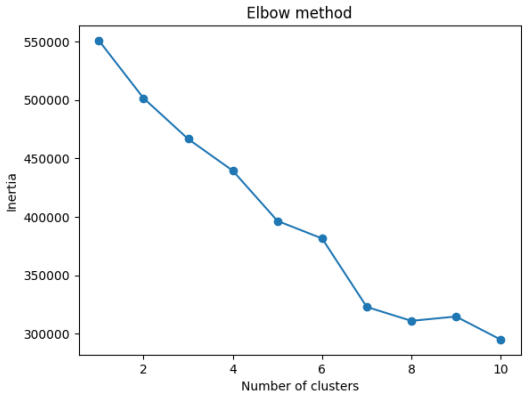

Figure 1. Elbow Method Result

The Elbow Method plots within-cluster inertia (sum of squared distances between observations and their cluster centroid) for K value 1 to 10 (number of clusters). The inertia declines steeply as the number of clusters increases from k = 1 to k = 4, but more gradually beyond k = 8.

Silhouette analysis was conducted across k = 2 to k = 10 on the 14-component FAMD-reduced data. The Silhouette Scores for each configuration are reported in Table 4.

Table 4. Silhouette Scores

Number of Clusters (k)	Silhouette Score
2	0.088
3	0.110
4	0.143
5	0.147
6	0.152
7	0.162
8	0.179
9	0.167
10	0.166

For each value of k considered, Silhouette Scores increased from k = 2 to k = 8 (where the highest Silhouette Score was obtained), and then slightly decreased at k = 10. Thus, Silhouette Analysis suggests that k = 8 is the most appropriate number of clusters.

The final K-Means algorithm with k = 8 was performed on the 14-component FAMD output. The cluster labels were mapped to the original dataset for profiling, as displayed in Table 5.

Table 5. Cluster Distribution

Cluster Label	Count	% of Total	Dominant Daily Stress Level
0	1,449	9.1%	Stress Level 4
1	3,203	20.1%	Stress Level 3
2	2,478	15.5%	Stress Level 1
3	1,195	7.5%	Stress Level 3
4	2,052	12.8%	Stress Level 5
5	3,407	21.3%	Stress Level 2
6	1,511	9.5%	Stress Level 4
7	676	4.2%	Stress Level 0

Initial observation suggests that Daily stress level (DAILY_STRESS) emerged as a near-perfect discriminating variable for these clusters. Where each cluster contained a single particular value of DAILY_STRESS. Thus, FAMD has preserved this variable within the reduced data, and K-Means has learned to form clusters according to these stress levels.

Interpretation

To interpret the behavioral character of each cluster, a Lift Index was calculated for all continuous wellbeing variables. The index normalizes each cluster's mean against the population mean (Lift = 100 corresponds to average; >100 above average; <100 below average). Table 6 presents selected high-signal features.

Table 6. Lift Index for Selected Continuous Variables

Feature	Cl.0	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6	Cl.7
DAILY_SHOUTING	116	102	63	108	150	84	124	50
LIVE_VISION	103	106	114	86	80	107	75	122
LOST_VACATION	121	94	67	111	147	86	123	61
TIME_FOR_PASSION	104	100	115	94	79	106	75	133
WEEKLY_MEDITATION	93	100	114	98	85	105	85	126
DONATION	121	104	104	91	96	102	73	103
SUPPORTING_OTHERS	119	105	102	93	97	101	76	102
FLOW	103	104	111	95	84	105	76	120

Several patterns emerge from these clusters. Cluster 7 has the most extreme profile of the clusters, having the lowest values of daily shouting (lift of 50), high time for passion (133), weekly meditation (126), and live vision (122). Despite the small size of this cluster (4.2% of the total sample), Cluster 7 represents the most psychologically healthy segment of the population with its low stress levels (all with a stress level of 0), high levels of aspiration, and self-regulation in meeting those aspirations.

Cluster 4 exhibits the opposite of the characteristics of Cluster 7. It has the highest values of daily shouting (lift of 150), lost vacation (147), and the lowest values of both live vision (80) and time for passion

(79). All individuals in this cluster have stress levels of 5, indicating that it is the most distressed and burnt-out segment of this population.

Cluster 6 is also a concerning segment of the population. It has relatively high daily shouting (124), high lost vacation (123), but low levels of the variables that indicate willingness to engage with others and pursue interesting activities and goals: donation (73), supporting others (76), live vision (75), and flow experiences (76). Most individuals in this cluster are between the ages of 21 and 35 (66.8% of the members of this cluster). They are similar to the Cluster 4, but instead of extreme anger, they seem defined by a general lack of purpose and activity.

Cluster 2 has high levels of weekly meditation (114), live vision (114), and low levels of daily shouting (63) and lost vacation time (67). Most individuals in this cluster have low stress levels (all with a stress level of 1), as well as high levels of time for passion (115). Thus, this cluster is another with high levels of self-care and low stress, similar to Cluster 7.

Cluster 0 has above-average values of the variables measuring willingness to engage in prosocial behaviour: supporting others (119), donation (121), and daily shouting (116). Additionally, this cluster has above-average levels of lost vacation time (121). However, stress levels for this cluster are relatively high; all individuals have stress levels of 4. Overall, this section potentially reflects the emotionally volatile, but high activity of prosocial behavior segment of the cluster.

Overall, clusters 1 (36.5%), 3 (5.6%), and 5 (4.9%) lie in the middle of the lift variable. Each of these clusters has most of the variables near the average for the population. These three clusters are the largest group with average characteristics (collectively 48.8% of the total sample) and differ primarily in their stress levels (clusters 1, 3, and 5 have stress levels of 3, 3, and 2, respectively).

The cross-tabulation of the five categorical variables according to cluster further reveals degrees of differentiation within the data according to these variables.

Table 7. Stress Level Distribution By Clusters

Cluster	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
0	0%	0%	0%	0%	100%	0%
1	0%	0%	0%	100%	0%	0%
2	0%	100%	0%	0%	0%	0%
3	0%	0%	0%	100%	0%	0%
4	0%	0%	0%	0%	0%	100%
5	0%	0%	100%	0%	0%	0%
6	0%	0%	0%	0%	100%	0%
7	100%	0%	0%	0%	0%	0%

The variable of daily stress level is what definitively separates the individuals within each cluster. Each cluster has an even 100% association with one level of stress. Cluster 4 has 100% of its members with a stress level of 5, and Cluster 7 has 100% of its members with a stress level of 0. Thus, the contribution of this variable to the Component scores of the FAMD analysis is strong; it is the only variable to possess a strong contribution to components 5–9, with the highest contribution (0.393) to component 6. Thus, it is this variable that primarily indicates the separation of clusters according to the FAMD analysis.

Table 8. Sufficient Income Distribution By Cluster

Cluster	Insufficient (1)	Sufficient (2)
0	13.8%	86.2%
1	0.0%	100.0%

Cluster	Insufficient (1)	Sufficient (2)
2	18.8%	81.2%
3	100.0%	0.0%
4	41.3%	58.7%
5	22.7%	77.3%
6	47.1%	52.9%
7	19.7%	80.3%

The SUFFICIENT_INCOME category exhibits considerable variation within the clusters. Cluster 3 contains 100% of its members with insufficient income, while cluster 1 contains 100% of individuals with sufficient income levels. Both of these clusters, however, have the same stress level of 3. Thus, this variable acts as a second distinguishing factor within the FAMD analysis.

Table 9. Age Distribution By Cluster

Cluster	< 20	21–35	36–50	51+
0	1.9%	8.3%	55.1%	34.7%
1	10.1%	36.6%	31.2%	22.1%
2	11.9%	36.2%	25.2%	26.7%
3	13.6%	48.7%	25.6%	12.1%
4	11.9%	37.3%	31.6%	19.2%
5	10.3%	39.5%	29.6%	20.6%
6	20.3%	66.8%	7.9%	5.1%
7	16.1%	32.0%	22.3%	29.6%

The age of individuals within each cluster exhibits some variation. Cluster 6, which contains high levels of stress, low levels of live vision, and low levels of social interactions online, contains 66.8% of its members between the ages of 21 and 35, and 20.3% of those within the cluster are under the age of under 20. Cluster 0 contains the majority of individuals within the age group between 36 and 50 years of age (55.1% of the individuals within the cluster fall within that age group), as well as 34.7% of those within the cluster who are 51 years of age or beyond. Cluster 7, those with high levels of meditation, contain the most even distribution of individuals of each age group, with 29.6% of those within the cluster being 51 or older.

Table 10. Gender Distribution By Cluster

Cluster	Female	Male
0	66.9%	33.1%
1	65.4%	34.6%
2	52.4%	47.6%
3	64.4%	35.6%
4	71.1%	28.9%
5	60.2%	39.8%
6	60.6%	39.4%
7	44.7%	55.3%

Each of the clusters is predominantly composed of female individuals, as the total population is comprised of 61.7% females. The exceptions to this include Cluster 7, which is 44.7% female and 55.3% male (the gender distribution for this cluster is even, and this cluster is the only one to contain a majority of male individuals), and Cluster 4 (those with the maximum level of daily stress) who are 71.1% female. Cluster 2, those with low levels of stress and high levels of mindfulness, is the most balanced in relation to its gender distribution, with 52.4% of the individuals within the cluster being female, and the remaining 47.6% being male.

Table 11. BMI Range Distribution By Cluster

Cluster	BMI Range 1	BMI Range 2
0	39.6%	60.4%
1	58.7%	41.3%
2	65.5%	34.5%
3	55.6%	44.4%
4	51.8%	48.2%
5	61.9%	38.1%
6	71.1%	28.9%
7	62.9%	37.1%

The range of body mass index (BMI) of individuals within each cluster exhibits some variation. Cluster 0 contains 60.4% of its members with a BMI that falls within range 2 (the higher range of body mass levels), compared to the population average value of 41.1%. Cluster 6, again, those with high levels of stress and low levels of live vision online, has the lowest percentage of its members within BMI range 2, with only 28.9% of the members of this cluster having a body mass within that range. Furthermore, the low-stress clusters (Clusters 2 and 7) contain the majority of their members within BMI range 1 (the healthier range of body masses), at 65.5% for Cluster 2 and 62.9% for Cluster 7. This is in line with the healthier behavioral profiles of these segments.

Discussion

This paper demonstrates the application of FAMD as a dimensionality reduction technique for mixed wellbeing data, followed by K-Means clustering to identify distinct individual segments.

The Silhouette Score measurements for the various values of k between 2 and 8 all showed an improvement over the preceding k. The maximum score of 0.179 for k = 8 was lower than the scores obtained for other research such as (Delimiro Visbal-Cadavid, 2020), but expected to be lower due to the nature of the data used. Specifically, the high dimensionality and overlapping values of the wellbeing variables. However, the monotonic increase in Silhouette Score with k indicates that the clusters with k = 8 are capturing meaningful structure within the dataset, rather than overfitting to noise.

The FAMD model indicated that the most important dimensions for wellbeing with respect to the clustered data were DAILY_STRESS, income sufficiency, and aspirational clarity. Furthermore, each cluster exactly corresponds to one of the DAILY_STRESS levels, indicating that the importance of this categorical variable was recognised by the model. Additionally, Clusters 1 and 3 were entirely separated along the dimension of SUFFICIENT_INCOME, indicating that such nuances within the categorical data are also preserved by the model—an expected outcome given that FAMD was applied to ensure the harmonisation of both numerical and categorical data.

An analysis of the resulting 8 clusters with the lift index and cross-tabulation of the categorical variables indicated differences in the characteristics of each segment of the population. For instance, Clusters 7 and 2 contained individuals with low levels of stress and who followed wellbeing and meditation practices. Cluster 4 contained individuals with high levels of stress and low levels of wellbeing, indicating a potential for adverse effects on their mental and physical health. Cluster 6 contained individuals with high levels of stress, who did not like to participate in social activities, and who were of a younger age

group compared to the other clusters. Cluster 6 reflects a different risk pattern in comparison to Cluster 4 in that it is more prone to early career disengagement rather than chronic burnout

Each of these segments suggests ways in which individuals might be benefited by the implementation of wellbeing programmes. For instance, individuals within Cluster 4 and Cluster 6 would benefit from stress management program and exposure to workplace boundary-setting initiatives. On the other side, the average segments of the cluster (1,3, and 5) could benefit more from a preventive measure rather than full intervention. Additionally, the differences between individuals according to their income within Clusters 1 and 3 indicate that economic security is another variable that should be incorporated into the creation of wellbeing programmes for individuals.

The framework presented here would be suitable for deployment within systems that monitor the wellbeing of large populations. Furthermore, the scalability of the algorithm for new data points indicates that this model can be applied to each new wave of surveys to determine the wellbeing of those individuals. Future work would involve determining the stability of these eight segments across different waves of the surveys.

Conclusion

This study aims to address three main gaps in the literature regarding wellbeing. Each of these gaps relates to either the lack of models that consider behavioral variables for segmentation, methods that can handle mixed data types, or the actionability of current wellbeing profile studies. Thus, each of the study's objectives is addressed by the FAMD–K-Means algorithm, illustrating how this study contributes to the existing literature on wellbeing.

One of the main contributions of this study is the use of Factor Analysis of Mixed Data (FAMD) to address the issue of using mixed types of variables within wellbeing clustering studies. Many current segmentation methods have used Latent Profile or Latent Class analyses, methods that assume certain types of independence within the data collected from individuals regarding their wellbeing. Furthermore, these analyses tend to struggle with the inclusion of categorical variables in segmentation studies. By applying FAMD to the wellbeing data, these issues are avoided; FAMD can mathematically account for both continuous and categorical variables while preserving the distinctions among categorical variables that reflect wellbeing. Thus, this method offers an advancement over the methods used in previous segmentation studies of wellbeing.

In addition to presenting a method that can effectively analyze wellbeing data, this study also contributes to knowledge of wellbeing through the segmentation results themselves. Specifically, segmentation of the wellbeing data allows for the division of adults into eight different wellbeing segments, each of which demonstrates different characteristics in relation to daily life and wellbeing. For instance, Cluster 3 contains individuals who are financially insecure yet experience similar levels of stress to those who are financially secure and classified as Cluster 1. Thus, the existence of this cluster indicates that financial security is associated with wellbeing independently of stress levels, a finding that is hard to discover through the analysis of either variable alone.

Finally, within each of these segments, individuals of different ages have been found to exhibit different characteristics. For instance, Cluster 6 contains a higher percentage of adults under the age of 35 than other clusters with different risk factor, and warranting a different intervention in comparison to the older population of Cluster 4 with the chronic burnout profile. Thus, this study indicates that segmentation of individuals according to their wellbeing can reveal differences across age groups, findings that are difficult to analyze with methods that are not specifically centered on individuals' wellbeing.

Given the above methodology, it is possible to employ this approach within wellbeing monitoring systems for large-scale organizations or for public health in general. due to the low computational cost of both the FAMD and the K-Means models, it is possible to apply this methodology to a series of wellbeing surveys over time to monitor changes between these different segments of the population.

Limitations

Several limitations of this study exist. First, the current model is static in its nature, and adapting the study to consider streaming or longitudinal data would enhance its utility in monitoring wellbeing over time. Second, while the model can determine the contribution of each variable to the formation of clusters through lift analysis, automated tools like SHAP (SHapley Additive exPlanations) could be incorporated to determine the contribution of each feature to the membership of each observation in the formed clusters. Finally, while the Silhouette Scores for each k-value were relatively low, other clustering algorithms exist that might be better suited to find clusters within the data.

Acknowledgments

Give credit to funding bodies and departments that have been of help during the project, for instance, by supporting it financially.

References

Andretta, J. R., & McKay, M. T. (2020). Self-efficacy and well-being in adolescents: A comparative study using variable and person-centered analyses. *Children and Youth Services Review*, *118*, 105374. https://doi.org/10.1016/j.childyouth.2020.105374

Chong, B. (2021). K-means clustering algorithm: A brief review. *Academic Journal of Computing & Information Science*, *4*(5), 37–40. https://doi.org/10.25236/AJCIS.2021.040506

Delimiro Visbal-Cadavid, A. M.-M. (2020). Use of Factorial Analysis of Mixed Data (FAMD) and Hierarchical Cluster Analysis on Principal Component (HCPC) for Multivariate Analysis of Academic Performance of Industrial Engineering Programs. *Journal of Southwest Jiaotong University*, *55*(5). https://jsju.org/index.php/journal/article/view/745

Grippo, C., Jagmohan, P., Helbich, T. H., Kapetas, P., Clauser, P., & Baltzer, P. A. T. (2021). Correct determination of the enhancement curve is critical to ensure accurate diagnosis using the Kaiser score as a clinical decision rule for breast MRI. *European Journal of Radiology*, *138*, 109630. https://doi.org/10.1016/j.ejrad.2021.109630

Hennessey, A., MacQuarrie, S., & Petersen, K. J. (2024). Exploring physical, subjective and psychological wellbeing profile membership in adolescents: A latent profile analysis. *BMC Psychology*, *12*(1), 720. https://doi.org/10.1186/s40359-024-02196-5

Lazić, M., Jovanović, V., Gavrilov-Jerković, V., & Boyda, D. (2021). A person-centered evaluation of subjective well-being using a latent profile analysis: Associations with negative life events, distress, and emotion regulation strategies. *Stress and Health*, *37*(5), 962–972. https://doi.org/10.1002/smi.3056

Leite, Â., Ramires, A., Vidal, D. G., Sousa, H. F. P. E., Dinis, M. A. P., & Fidalgo, A. (2021). Hierarchical Cluster Analysis of Human Value Priorities and Associations with Subjective Well-Being, Subjective General Health, Social Life, and Depression across Europe. *Social Sciences*, *10*(2), 74. https://doi.org/10.3390/socsci10020074

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. *Entropy*, *23*(4). https://doi.org/10.3390/e23040485

Pasanen, T. P., Tamminen, N., Martelin, T., Mankinen, K., & Solin, P. (2021). Profiles of subjective health among people living alone: A latent class analysis. *BMC Public Health*, *21*(1), 1335. https://doi.org/10.1186/s12889-021-11396-2

Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). *Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method*. 341–346. https://doi.org/10.2991/aisr.k.200424.051

Singh, A., Pandey, U. K., Singh, A., Bhatt, H., Mishra, S., & Mishra, V. K. (2024). Improvising Performance of Machine Learning Algorithm using FAMD on Coronary Artery Disease. *2024 International Conference on Control, Computing, Communication and Materials (ICCCCM)*, 569–574. https://doi.org/10.1109/ICCCCM61016.2024.11039854

Ufeli, C. P., Sattar, M. U., Hasan, R., & Mahmood, S. (2025). Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms. *Information*, *16*(6). https://doi.org/10.3390/info16060441

UNESCO. (2022). *UNESCO strategy on education for health and well-being*. UNESCO. https://doi.org/10.54675/MSST2323

Wu, M., Yang, C., Zhang, Y., Umeda, M., Liao, J., & Mawditt, C. (2024). Longitudinal patterns and sociodemographic profiles of health-related behaviour clustering among middle-aged and older adults in China and Japan. *Ageing & Society*, *44*(11), 2464–2483. https://doi.org/10.1017/S0144686X2200143X

Wu, W.-T., Li, Y.-J., Feng, A.-Z., Li, L., Huang, T., Xu, A.-D., & Lyu, J. (2021). Data mining in clinical big data: The frequently used databases, steps, and methodological models. *Military Medical Research*, *8*(1), 44. https://doi.org/10.1186/s40779-021-00338-z