A Clustering-Based Analysis of Synthetic Socioeconomic Profiles and Employment Status

Rahma Atef, 2231172262

Abstract

This study explores latent socioeconomic structure in employment-related financial behavior using unsupervised learning. Traditional analyses often treat employment status as a binary variable, overlooking heterogeneity within populations.To address this limitation, we apply K-Means clustering to a synthetic dataset of demographic and financial attributes, with Principal Component Analysis (PCA) used for visualization. The results identify three distinct clusters (k = 3): moderate-income employed individuals, high-income employed individuals, and a lower-income cluster. Students appear across all clusters, indicating diverse financial profiles within this group. These findings show that employment status alone does not fully capture financial well-being and that clustering can reveal meaningful population subgroups. Because the analysis is based on synthetic data, the results are exploratory and motivate validation on real-world datasets.

I. Introduction

Unemployment remains a pressing global issue,affecting over 186 million people worldwide [1] and creating serious financial and social challenges. While prior research often focuses on predicting employment status using supervised models based on individual characteristics such as education, income, and savings [2], these approaches simplify employment into a binary outcome and overlook the diversity of financial and work behaviors within populations.

A key gap in the literature is the lack of methods that reveal natural socioeconomic groupings. Many individuals, including students, self-employed workers, and gig workers, do not fit neatly into "employed" or "unemployed" categories, and traditional models fail to capture this heterogeneity. Understanding these nuanced patterns is essential for designing targeted interventions and effective support mechanisms.

This study aims to uncover latent socioeconomic clusters by applying unsupervised learning to a synthetic dataset of demographic and financial attributes. We hypothesize that individuals will form distinct clusters that differ in financial stability and employment composition.In particular, we expect lower income, saving and expenses clusters to exhibit higher unemployment rates, while financially stable and more educated clusters will show higher employment rates.

Our contributions are as follows:

- Identify distinct socioeconomic clusters that reveal diversity in financial behavior and employment outcomes beyond traditional binary employment labels.
- Analyze how demographic and financial factors, including income, savings, education, and financial stress, relate to employment outcomes within each cluster.
- Reveals subgroups that traditional models miss, potentially improving policy targeting
- Provide insights that may inform researchers, educators, and financial planners when designing interventions tailored to different population groups.

II. Related Work

Prior research on socioeconomic patterns and employment has applied unsupervised clustering to reveal structure in labor market participation, financial behavior, and demographic characteristics. These studies can be broadly categorized into three themes: individual-level clustering, regional or macro-level clustering, and clustering with dimensionality reduction.

A. Individual-Level Socioeconomic Clustering

Several studies have applied clustering to individual-level socioeconomic data to examine social stratification and economic vulnerability [10], [6]. These approaches typically use features such as income, household conditions, education, and demographic background to identify disadvantaged or at-risk populations. A key advance of this work is the demonstration that individuals with similar employment labels or income levels can exhibit substantial heterogeneity in financial stability and living conditions. However, previous studies treated employment status as a contextual rather than an analytical outcome. Differences between employed, unemployed, self-employed, or student groups were rarely examined systematically, limiting information on how labor market outcomes relate to financial behavior.

B. Regional and Macro-Level Clustering

Another line of research has focused on analyzing or grouping regions or countries based on labor market indicators, unemployment rates, or economic structure [4], [5]. These studies successfully revealed geographic disparities and structural patterns in employment, providing guidance for region-specific policy and macroeconomic planning. Although informative on a macro scale, such studies abstract away from individual-level behavior, offering limited understanding of how personal demographic and financial characteristics shape employment outcomes.

C. Clustering with Dimensionality Reduction

To improve interpretability in high-dimensional datasets, several studies combined clustering with dimensionality reduction methods, particularly Principal Component Analysis (PCA) [7]. These approaches identified dominant dimensions—such as income security, educational attainment, or household financial pressure—that drive cluster formation. The main contribution of this research lies in improving visualization and reducing redundancy among correlated features. However, these studies often emphasized methodology over practical labor market insights, with limited attention to linking clusters to specific employment outcomes.

Despite these advances, existing research exhibits several key limitations. Most studies focus on demographic or economic variables alone, rarely combining both, which restricts the ability to capture complex interactions that influence employment outcomes. In addition, few studies explicitly relate the discovered clusters to employment categories, leaving the relationship between financial behavior and employment types underexplored. Many analyses are cross-sectional, preventing examination of temporal dynamics or longitudinal patterns in financial stability and employment. Although dimensionality reduction techniques such as PCA improve visualization, previous work often provides limited actionable information for researchers, policymakers, or financial planners. As a result, existing approaches are limited in their ability to uncover latent financial subgroups within employment categories without imposing predefined labels, motivating the need for an unsupervised, individual-level clustering framework.

Our study addresses these gaps by clustering a combined set of demographic and financial attributes at the individual level and explicitly analyzing how employment categories are distributed across clusters post hoc. In contrast to much of the prior literature, this study places employment outcomes at the center of the analysis rather than treating them as contextual information. This approach bridges methodological clustering research with practical labor market analysis, enabling a more nuanced understanding of how demographic and financial characteristics jointly shape employment and financial behaviors.

III. Methods and Experiments

A. Overview

This study investigates unemployment patterns by applying unsupervised clustering to a synthetic individual level socioeconomic dataset. The goal is to identify natural groupings of individuals based on demographic and financial features and examine how employment outcomes vary across these clusters. The workflow includes data preprocessing, feature transformation, clustering using K-Means, cluster validation, dimensionality reduction via Principal Component Analysis (PCA), and post-hoc analysis of cluster characteristics.

B. Design and Setup

Analyses were conducted in Python 3.12 usingpandas (2.2.0), numpy (1.26.4), scikit-learn (1.5.0) for clustering, scaling, PCA, and silhouette analysis, and matplotlib (3.8.3) for visualization. The workflow ensures reproducibility, with scripts for preprocessing, clustering, PCA, and visualization stored in a local repository and shareable upon request.

TABLE I
Synthetic Dataset Attributes

Attribute	Type	Description
Age	Numeric	Age in years
Education Level	Categorical	Highest education achieved
Region	Categorical	Geographic location
Income	Numeric	Monthly income in USD
Expenses	Numeric	Monthly expenses in USD
Savings	Numeric	Total savings in USD
Credit Score	Numeric	Creditworthiness rating
Employment Status	Categorical	Employed / Unemployed / Self-employed / Student

C. Materials and Data

The synthetic dataset was obtained from Kaggle [9] and contains 32,424 synthetic individual records. Each record includes:

- **Demographic:** age, education level, geographic region
- **Financial:** income, expenses, savings, credit score
- **Employment:** employment status (employed, unemployed, self-employed, student)

Synthetic data allows controlled exploration of latent socioeconomic patterns while avoiding privacy concerns, providing a reproducible environment to analyze the relationship between demographic/financial features and employment outcomes.

D. Procedure

1) **Exploratory Data Analysis (EDA) and Cleaning:** Feature distributions, missing values, and inconsistencies were examined. No missing values were detected. Nonsense or corrupted entries were removed, and minor errors (e.g., single incorrect values) were corrected using mean or median imputation. Irrelevant or redundant columns were also removed.
2) **Outlier Detection and Removal:** Outliers in numeric features (income, expenses, savings, age, credit score) were identified using the Interquartile Range (IQR) method with a threshold of 1.5×IQR. Only 12 rows of outliers were removed to reduce noise while minimally affecting the dataset.
3) **Encoding and Scaling:**Categorical variables (education level, region) were one-hot encoded for clustering; employment status was encoded only for post-hoc analysis and excluded from the clustering feature set.
4) **Clustering:** K-Means clustering was implemented using `sklearn.cluster.KMeans` with parameters: `n_clusters` tested from 2 to 10, `n_init`=10, and `random_state`=42 for reproducibility. Only demographic and financial features were used, excluding employment status. The optimal number of clusters ($k = 3$) was determined via silhouette analysis. Employment status was analyzed post-hoc to evaluate its distribution across clusters.
5) **Dimensionality Reduction and Visualization:** PCA reduced the high-dimensional dataset to two components for visualization, with `n_components`=2 for visualization. This configuration captured 34.7% of total variance facilitating 2D inspection of cluster separation despite the limited variance explained.
6) **Cluster Analysis:** Each cluster was examined for:
   - Distribution of demographic and financial attributes
   - Distribution of employment outcomes
   - Feature correlations and internal patterns using heatmaps and correlation matrices

E. Analysis Methods

Clusters were validated using silhouette scores to ensure well-separated and interpretable groupings. The optimal number of clusters ($k = 3$) was determined using silhouette score and elbow methods. Post-hoc analyses examined the relationships between cluster membership and employment categories, financial stability, and demographic factors through descriptive statistics means and standard deviations for continuous variables, frequency distributions for categorical variables. PCA visualizations with two components facilitated identification of patterns within and across clusters. Feature relationships were analyzed using Pearson correlation matrices and visualized with heatmaps.

F. Reproducibility and Ethics

No real human subjects or personally identifiable information were involved; therefore, IRB approval was not required. All analyses were conducted in Python 3.12 with specified libraries, and the use of synthetic data ensures full reproducibility and avoids privacy concerns. Scripts can be shared for replication.

This section presents the quantitative outcomes of the K-Means clustering analysis applied to the synthetic socioeconomic dataset. The aim was to identify natural groupings of individuals based on demographic and financial features.

IV. Results

This section presents the outcomes of the K-Means clustering analysis applied to the synthetic socioeconomic dataset. Our goal was to identify natural groupings of individuals based on demographic and financial attributes. We first validate the number of clusters, then summarize cluster sizes and financial characteristics, examine employment distributions, visualize cluster separation, and analyze correlations among features.

A. Cluster Validation

The optimal number of clusters was determined using both the Elbow and Silhouette methods. Both metrics indicated three clusters ($k = 3$) as most appropriate (Figures 1 and 2).

Takeaway: Both validation metrics consistently support three distinct clusters.

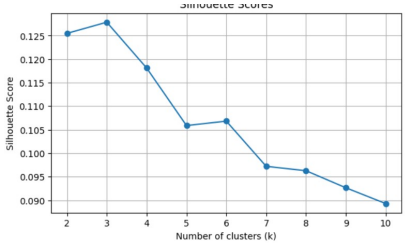

Fig. 1. Silhouette analysis indicating optimal cluster separation at $k = 3$.

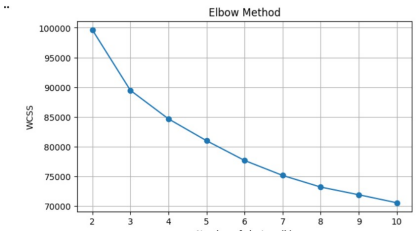

Fig. 2. Elbow method showing $k = 3$ as optimal number of clusters.

TABLE II
CLUSTER SUMMARY STATISTICS (FINANCIAL FEATURES)

Cluster	Size	Income (USD/month)	Expenses (USD/month)	Savings (USD total)
0	4,631	3,279 ± 1,012	1,910 ± 805	114,220 ± 52,310
1	2,534	291 ± 120	2,035 ± 890	219,502 ± 75,600
2	3,304	5,115 ± 1,560	3,127 ± 1,220	419,632 ± 120,450

B. Cluster Sizes and Financial Summary

Table II summarizes cluster sizes and mean financial attributes. Cluster 2 has the highest mean income ($5,115 ± $1,560) and savings ($419,632 ± $120,450), Cluster 1 shows the lowest mean income ($291 ± $120), and Cluster 0 is intermediate. **Takeaway:** Cluster 2 exhibits the highest financial prosperity, Cluster 1 shows low income but unexpectedly high savings, and Cluster 0 is intermediate.

C. Employment Distribution

Table III presents counts and percentages of employment categories per cluster.

TABLE III
CLUSTER COUNTS BY EMPLOYMENT STATUS

Cluster	Employed	Student	Self-employed	Unemployed	Total
0	3,476 (75%)	1,155 (25%)	0 (0%)	0 (0%)	4,631
1	0 (0%)	384 (15%)	2,020 (80%)	130 (5%)	2,534
2	2,286 (69%)	1,014 (31%)	4 (0%)	0 (0%)	3,304

Takeaway: Employment distribution varies across clusters. Cluster 1 is predominantly self-employed, while students appear in all clusters.

D. Cluster Visualization

Figure 3 shows a two-dimensional PCA projection of the clustered data, and Figure 4 presents standardized feature values per cluster.

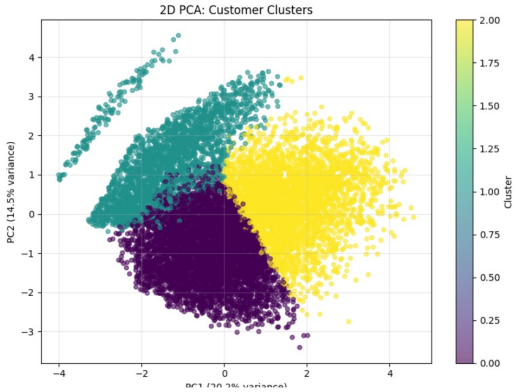

Fig. 3. PCA projection of clustered data ($k = 3$). Variance explained: PC1 = 14.5%, PC2 = 20.2%.

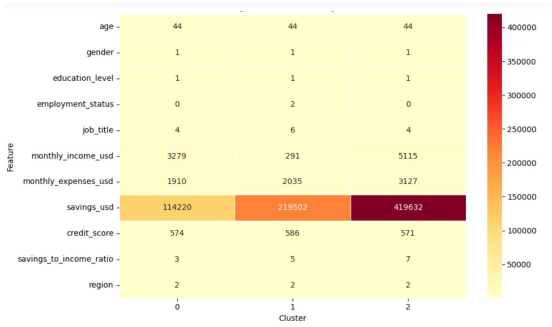

Fig. 4. Heatmap of standardized financial and demographic features per cluster.

Takeaway: PCA demonstrates mostly separated clusters with minor overlap. The heatmap highlights variation in financial attributes, whereas demographic features are more uniform.

E. Correlation Analysis

Pearson correlation coefficients between features are shown in Figure 5 and summarized below:

- Income vs. Expenses: $r = 0.55$
- Savings vs. Expenses: $r = 0.46$
- Age vs. Income: $r = 0.08$
- Credit Score vs. Income: $r = 0.12$

Fig. 5. Correlation matrix of financial and demographic features.

Takeaway: Financial variables are moderately correlated, while demographic features show weak correlations with financial attributes.

F. Secondary Observations

- Students constitute 25%, 15%, and 31% of Clusters 0, 1, and 2, respectively.

- The high mean savings in Cluster 1 results from the synthetic data generation, surprisingly includes a small number of extreme high-saving records.

V. Discussion

The clustering analysis revealed three distinct socioeconomic groups, each with characteristic financial behaviors and employment profiles. Cluster 0 consisted mainly of employed individuals with moderate income, expenses, and savings, reflecting stable and predictable financial patterns. Cluster 2 was composed primarily of high-income employed individuals with elevated expenses and the strongest savings-to-income ratios, consistent with expected financial stability. In contrast, Cluster 1 included self-employed, unemployed individuals, and students with low monthly income but unexpectedly high savings, highlighting irregular financial patterns.The sharp separation of employment categories across clusters reflects the structure of the synthetic dataset and should not be interpreted as representative of real-world labor markets.

These patterns suggest that employment type strongly influences financial stability, but other factors, such as irregular income from self-employment or accumulated savings, can lead to unexpected outcomes within low-income groups. The distribution of students across all clusters further emphasizes that demographic labels alone do not predict financial behavior; students exhibit diverse financial situations depending on part-time work, family support, and regional economic conditions. This heterogeneity underscores the importance of considering both employment type and contextual factors when assessing financial well-being.

Our findings align with previous research demonstrating the link between employment type and financial stability [3,5], but they extend these insights by uncovering heterogeneity within clusters. For example, prior studies typically report average characteristics, whereas our approach reveals subgroups, such as students with varying income and savings profiles, and self-employed individuals whose irregular income inflates cluster-level savings. This highlights the added value of unsupervised clustering combined with post-hoc analysis for understanding complex socioeconomic behaviors.

The implications for policy and practice are significant. Interventions such as financial aid, career guidance, or employment support programs should account for heterogeneity within demographic groups rather than applying uniform solutions. Tailoring support to the specific needs of students, self-employed workers, or low-income households can enhance program effectiveness and ensure resources are directed where they are most needed.

Several limitations should be acknowledged. First, the dataset is synthetically generated, which, while allowing controlled experimentation, may not capture all nuances of real-world financial and employment behavior. Second, the cross-sectional nature of the analysis prevents assessment of temporal dynamics, such as income volatility or transitions between employment types. Third, factors like secondary income sources, government support, or detailed regional economic indicators were not incorporated, which may influence cluster composition and observed patterns. Finally, although regional context was

conceptually considered, it was not explicitly modeled, limiting the precision of interpretations regarding geographic effects on financial behavior.

Overall, this study demonstrates that unsupervised clustering can uncover meaningful subgroups within heterogeneous populations and provides a nuanced understanding of the interplay between employment, financial behavior, and demographic characteristics. The results highlight both predictable patterns among stable employed groups and unexpected patterns among irregular-income populations, emphasizing the value of detailed cluster-based analysis for informing policy and further research.

VI. Conclusion

This study demonstrates that unsupervised clustering of financial and demographic data can uncover meaningful subgroups with distinct employment and financial patterns. Key insights reveal that stable employed individuals exhibit predictable financial behavior, whereas irregular-income groups, including self-employed workers and students, display diverse outcomes and unexpected savings patterns, highlighting heterogeneity often overlooked by conventional analyses. Understanding these differences provides a more nuanced perspective on socioeconomic behavior, emphasizing that demographic labels alone are insufficient to predict financial stability or employment outcomes.

The practical implications are significant. Policymakers, educators, and financial planners can leverage these findings to design targeted interventions, such as customized financial aid, career guidance, and employment support programs, which account for subgroup specific needs rather than applying uniform strategies. This approach can improve resource allocation, program effectiveness, and support for populations with irregular or variable income.

Future work should extend this analysis by incorporating real-world data, considering additional socioeconomic factors such as secondary income sources, government support, or regional economic indicators, and exploring alternative clustering or dimensionality reduction techniques. Such efforts would allow tracking temporal changes in financial stability and employment transitions, thereby providing a deeper understanding of socioeconomic dynamics over time.

Overall, this work advances understanding of the interplay between employment, financial behavior, and demographic factors, demonstrating the value of unsupervised clustering for revealing hidden heterogeneity within populations and informing more precise, data-driven interventions in labor market and financial planning.

References

[1] International Labour Organization (ILO), World Employment and Social Outlook: Trends 2024. Geneva: ILO, 2024.

[2] Z. Othman, S. W. Shan, I. Yusoff, and C. P. Kee, "Classification techniques for predicting graduate employability," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 4-2, pp. 1712–1720, Sep. 2018, doi:10.18517/ijaseit.8.4-2.6832.

[3] R. Finnie and D. Gray, "Labour-force participation of older displaced workers," *Canadian Public Policy*, vol. 38, no. 4, pp. 1–25, 2012.

[4] P. McCann, "The regional and urban policy of the European Union," *Regional Studies*, vol. 49, no. 6, pp. 978–998, 2015.

[5] V. Monastiriotis, "Regional growth and unemployment disparities," *European Urban and Regional Studies*, vol. 21, no. 2, pp. 123–145, 2014.

[6] F. Booysen, M. Van Der Berg, R. Burger, M. Von Maltitz, and H. Du Rand, "Using an asset index to assess trends in poverty in seven Sub-Saharan African countries," *World Develop.*, vol. 34, no. 6, pp. 1108–1130, 2006.

[7] S. Vyas and L. Kumaranayake, "Constructing socio-economic status indices: how to use principal components analysis," *Health Policy Plan.*, vol. 21, no. 6, pp. 459–468, 2006.

[8] B. Hashemian, E. Massaro, I. Bojic, J. M. Arias, S. Sobolevsky, and C. Ratti, "Socioeconomic characterization of regions through the lens of individual financial transactions," *PLOS ONE*, vol. 12, no. 6, p. e0187031, 2017.

[9] Miadul, "Personal Finance ML Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/miadul/personal-finance-ml-dataset. [Accessed: Dec. 24, 2023].

[10] S. Sujarwoto and G. Tampubolon, "Spatial clustering of diet, physical activity, and childhood obesity and its socio-economic determinants: evidence from a nationwide survey in Indonesia," *PLoS ONE*, vol. 11, no. 3, p. e0150489, Mar. 2016.