Journal of Financial and Management Sciences, 2 (3) 242 –262, (2026)

DOI: https://doi.org/10.70970/fxyq7d50

Behavioural Segmentation of Credit-Constrained Households Using Unsupervised Machine Learning

Jean Paul Manirafasha, Eric Nizeyimana, David Hagumyuwumva

Received: May 22, 2026

Revised: July 8, 2026
Accepted: August 13, 2026
Published online: September 10, 2026

Abstract

Credit-constrained households are frequently treated as a homogeneous group in financial analysis, despite substantial heterogeneity in their financial behavior. This simplification limits the effectiveness of financial inclusion policies by obscuring important behavioral differences across households. This study addresses this limitation by developing an interpretable unsupervised machine learning framework to uncover latent behavioral segments among credit-constrained households using FinScope microdata from the Rwanda National Institute of Statistics (NISR). A purposive filtering approach is applied to isolate credit-constrained households, followed by trimmed variance feature selection to retain the most informative financial variables while mitigating the influence of extreme outliers. The selected variables are standardized and analyzed using K-Means clustering. Model evaluation results indicate that a three-cluster solution provides the optimal segmentation structure, as supported by the Elbow Method and a Silhouette Score of approximately 0.68, reflecting satisfactory cluster separation and internal cohesion. The resulting clusters reveal distinct behavioral profiles, including low-income debt-burdened households, high-income high-leverage households, and moderate-income wealth-accumulating households. These findings demonstrate that credit constraint is not a uniform condition but a multidimensional phenomenon shaped by diverse financial structures, highlighting deviations from traditional assumptions of homogeneous household behavior. In conclusion, the study provides a scalable and interpretable analytical framework that enhances behavioral segmentation in household finance and supports more targeted, evidence-based financial inclusion policies and institutional decision-making.

Keywords: Financial Inclusion, Credit-Constrained Households, K-Means Clustering, Behavioral Segmentation, Unsupervised Learning

jeanpaulfasha@gmail.com, eric.nizeyimana@auca.ac.rw, david.hagumuwumva@auca.ac.rw
Department of Big Data Analytics, Adventist University of Central Africa (AUCA), Kigali, Rwanda

1. Introduction

Household financial behavior is inherently heterogeneous, yet credit-constrained households are frequently treated as a homogeneous group in financial analysis and policy design. This simplification obscures critical differences in income stability, asset ownership, and debt exposure, thereby limiting the effectiveness of financial inclusion strategies. As financial systems become increasingly data-rich, the ability to accurately segment households based on underlying financial behavior has become a strategic priority for regulators and financial institutions seeking to design targeted and evidence-based interventions [1, 2].

However, the growing complexity of household financial data characterized by high dimensionality, non-linear relationships, and extreme wealth outliers poses significant challenges for traditional segmentation and econometric approaches. These methods typically rely on predefined outcome variables and are highly sensitive to multicollinearity and distributional distortions, leading to information loss and reduced interpretability [3, 4]. As a result, existing approaches are often insufficient for capturing latent heterogeneity within credit-constrained populations, particularly in emerging economy contexts where financial structures are more complex and less formalized.

Despite these advances, limited research has applied unsupervised machine learning techniques specifically to identify behavioral heterogeneity within credit-constrained households. This represents a critical gap, as such methods are well-suited for uncovering latent structures in high-dimensional financial data without requiring predefined labels.

In response, this study aims to develop an interpretable unsupervised machine learning framework to identify latent behavioral segments among credit-constrained households using FinScope data from Rwanda. The study applies a robust feature selection approach alongside clustering techniques to group households based on financial similarity rather than predefined classifications.

This study contributes to the literature in three ways. First, it provides empirical evidence that credit-constrained households are behaviorally heterogeneous. Second, it introduces a trimmed variance feature selection approach to improve clustering robustness in the presence of extreme financial values. Third, it offers policy-relevant insights by linking behavioral segments to targeted financial inclusion strategies.

2. Literature Review

Understanding household financial behaviour and credit access requires integrating perspectives from economics, finance, and data science. Financial inclusion is a multidimensional construct that extends beyond access to formal financial services to include the effective use of credit, savings, and risk management instruments for consumption smoothing and wealth accumulation [1, 2]. Accordingly, financial exclusion is not merely a lack of access, but a manifestation of constrained financial behaviour.

Traditional approaches to analysing household financial behaviour have relied heavily on demographic segmentation, classifying individuals based on observable characteristics such as income, age, or geographic location [3]. While operationally convenient, these approaches implicitly assume behavioural homogeneity within demographic groups and fail to capture latent differences in financial decision-making, particularly under conditions of credit constraint.

From a theoretical perspective, the Life-Cycle Hypothesis (LCH) and Permanent Income Hypothesis (PIH) provide foundational models of household financial behaviour [10, 11]. Both frameworks assume that individuals can smooth consumption over time through access to credit markets. However, this assumption is frequently violated in practice. In the presence of binding credit constraints, households are unable to borrow against future income, leading to deviations from optimal consumption behaviour.

Empirical evidence shows that such constraints result in precautionary savings, reduced consumption, and increased reliance on informal financial mechanisms [7].

These deviations underscore a critical limitation of classical economic theory: the assumption of frictionless financial markets. In reality, credit-constrained households exhibit heterogeneous behavioural responses depending on income stability, asset ownership, and access to financial services. This heterogeneity cannot be adequately captured using traditional demographic or theory-driven models.

In response, financial segmentation has evolved toward behaviourally informed approaches that aim to group individuals based on financial actions rather than static characteristics [3]. This shift reflects a broader recognition that financial vulnerability is driven by complex, multidimensional interactions that are not directly observable through conventional variables.

Empirical research has increasingly leveraged large-scale financial datasets; however, most studies do not explicitly isolate credit-constrained households as a distinct analytical category. For instance, Williams [13] provides descriptive insights into household debt dynamics but does not examine behavioural segmentation within constrained populations. Similarly, supervised machine learning approaches, such as those developed by White [5], focus on predicting outcomes like default risk, rather than identifying latent behavioural structures among financially vulnerable but non-defaulting households.

Recent studies have explored unsupervised learning techniques for financial segmentation. Chen et al. [12] apply K-Means clustering to transaction data, while Nguyen [9] uses principal component analysis to examine wealth structures.

However, these studies primarily focus on consumption patterns or dimensionality reduction and do not explicitly address behavioural heterogeneity within credit-constrained populations. Zhao et al. [8] further highlight the existence of a "missing middle," comprising households that are neither fully excluded nor fully integrated into formal credit markets, reinforcing the need for more nuanced analytical frameworks.

In the African context, Hall [6] demonstrates that behavioural financial patterns exhibit cross-market similarities, suggesting the applicability of data-driven segmentation methods. Nevertheless, existing studies continue to treat credit-constrained households as a relatively homogeneous group, limiting the identification of nuanced behavioural differences.

Taken together, the literature reveals two key gaps. First, there is limited application of unsupervised learning methods to explicitly analyse behavioural heterogeneity within credit-constrained populations. Second, existing approaches do not adequately address the challenges posed by high-dimensional financial data with extreme value distributions.

To address these gaps, this study conceptualises credit constraint as a continuum of behavioural manifestations rather than a binary condition. Under this framework, unsupervised machine learning techniques are employed to uncover latent behavioural segments, providing a more granular and empirically grounded understanding of household financial exclusion.

3. Data and Methodology

3.1 Data

This study uses secondary household-level data from the FinScope 2024 survey conducted by the Rwanda National Institute of Statistics. The dataset provides nationally representative information on household socio-economic conditions, including income, assets, liabilities, expenditure patterns, and access to formal and informal financial services.

The original dataset consists of 13,994 households and approximately 25 variables. For the purpose of this study, the analytical focus is restricted to credit-constrained households, defined as those that satisfy at least one of the following conditions: (i) reported denial of credit, (ii) discouraged from applying due to anticipated rejection, or (iii) limited access to formal credit services.

Applying these criteria yields a working sample of 4,458 households, which constitutes the effective population for the analysis. The household is treated as the unit of analysis, reflecting its role as the primary financial decision-making entity.

A subset of variables capturing key dimensions of financial behavior is selected, including income, assets, debt, and selected socio-demographic indicators. Missing values are handled using imputed variables provided within the dataset to preserve consistency. Continuous variables are subsequently standardized to ensure comparability across different measurement scales.

The resulting dataset forms a structured and analysis-ready input for the unsupervised learning pipeline.

4. Methodology

This study adopts an exploratory quantitative research design grounded in unsupervised machine learning to identify latent behavioral segments among credit-constrained households. The choice of this design is motivated by the absence of predefined outcome variables and the need to uncover intrinsic patterns directly from high-dimensional financial data. Unlike supervised approaches, which rely on labeled outcomes, the objective here is to reveal underlying structures in household financial behavior without prior assumptions. The analysis is therefore situated within a data-driven paradigm commonly used in computational social science and data mining.

The empirical analysis is based on secondary data obtained from the FinScope 2024 survey conducted by the Rwanda National Institute of Statistics. The dataset provides nationally representative information on household socio-economic and financial conditions, including income, assets, debt, expenditure patterns, and access to formal and informal financial services. The original dataset consists of 13,994 households and approximately 25 variables. For the purpose of this study, the focus is restricted to credit-constrained households, defined as those that have experienced at least one of the following conditions: denial of credit, discouragement from applying due to perceived rejection, or limited access to formal financial services. Applying these criteria yields a working sample of 4,458 households, which forms the analytical population. The household is treated as the unit of analysis, reflecting its role as the primary economic decision-making entity.

The methodological approach is implemented as a structured and sequential data transformation pipeline in which each stage produces an output that directly informs the subsequent step, ensuring coherence, transparency, and reproducibility throughout the analysis. The process begins with data preparation, where the dataset is filtered to isolate the target population of credit-constrained households, thereby reducing the analytical space from 13,994 to 4,458 observations. Following this population filtering, relevant variables capturing key dimensions of household financial behavior—such as income, assets, debt, and selected socio-demographic indicators—are extracted. Missing values are handled using imputed variables provided within the dataset to preserve consistency and minimize information loss.

To ensure comparability across variables measured on different scales, all continuous variables are standardized using z-score normalization, allowing each feature to contribute proportionally to subsequent analysis. The transformation is defined as:

$$Z_{ij} = (X_{ij} - \mu_j) / \sigma_j$$

where X_{ij} represents the value of variable j for household i, μ_j is the mean of variable j, and σ_j is the corresponding standard deviation. This normalization step produces a scale-invariant dataset, which is essential for subsequent variance-based feature selection and distance-based clustering.

Following standardization, feature selection is conducted to reduce dimensionality and improve the signal-to-noise ratio in the data. Given the presence of extreme values typical of financial datasets, a trimmed variance approach is applied. This involves excluding observations from both tails of each variable distribution prior to computing variance, thereby ensuring that variability reflects typical household behavior rather than outliers. The variance of each feature is computed as:

$$s^2 = (1 / (n - 1)) \Sigma_{i=1}^{n} (x_i - \bar{x})^2$$

where x_i represents the trimmed observations and $\bar{x}$ is the trimmed mean. Features exhibiting higher trimmed variance are retained, as they capture greater heterogeneity in financial behavior, while low-variance variables are excluded. This process reduces the dataset from approximately 25 variables to a smaller set of informative features, ultimately yielding a core subset of variables that serve as the input for clustering.

The refined and standardized dataset is then used to perform clustering using the K-Means algorithm. The objective of K-Means is to partition the households into K groups such that within-cluster similarity is maximized and between-cluster similarity is minimized. This is operationalized by minimizing the within-cluster sum of squared distances between observations and their respective cluster centroids:

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

where C_k denotes the set of observations in cluster k and μ_k represents the centroid of that cluster. The algorithm is initialized using the K-Means++ method to improve convergence and reduce sensitivity to initial centroid selection. Multiple random initializations are performed to ensure stability of the clustering solution. At this stage, households are grouped based on similarity in their financial profiles, generating preliminary cluster structures that reflect latent behavioral patterns.

To determine the optimal number of clusters, internal validation techniques are employed. Specifically, the elbow method is used to examine the rate of decrease in within-cluster variance as the number of clusters increases, while the Silhouette score is used to assess the degree of separation between clusters. These complementary metrics ensure that the selected clustering solution achieves a balance between compactness and separation, thereby enhancing both statistical validity and interpretability.

Once the optimal clustering structure is established, the results are further examined through visualization and interpretation. Dimensionality reduction techniques are used to project the high-dimensional data into a lower-dimensional space for visualization purposes, facilitating the identification of cluster separation patterns. The resulting clusters are then profiled based on their financial characteristics, allowing for economic interpretation of each segment. This step transforms the clustering output into meaningful insights that can inform financial inclusion strategies and policy design.

To ensure robustness, the entire analytical process is designed to be reproducible and internally consistent. Validity is supported through the use of theoretically grounded variables and established clustering evaluation metrics, while reliability is enhanced by the use of standardized data, consistent preprocessing procedures, and repeated model runs to confirm stability of results. Ethical considerations are addressed through the use of anonymized secondary data and adherence to principles of responsible data handling and non-discriminatory interpretation.

Overall, the methodology integrates data filtering, standardization, feature selection, clustering, validation, and interpretation into a coherent analytical framework in which each stage builds directly upon the previous one. This structured pipeline ensures that the analytical process remains logically connected, methodologically rigorous, and fully reproducible, thereby enabling the robust identification of latent behavioral segments among credit-constrained households.

5. Results

Following the data preparation and filtering procedures, the final analytical sample consists of 4,458 credit-constrained households described by a reduced set of high-informative financial variables. Initial exploratory analysis reveals substantial dispersion in key indicators such as income, assets, and debt, with pronounced right-skewness and the presence of extreme values. This confirms that household financial data are highly heterogeneous and justifies the application of robust feature selection techniques to avoid distortion from outliers.

To address this, a trimmed variance approach was applied to all candidate variables. The results show clear differentiation in variability across financial indicators. Asset-related variables exhibit the highest trimmed variance, with total assets (1.76×10^{12}), houses (7.33×10^{11}), and net worth (4.77×10^{11}) dominating the distribution, followed by debt (4.41×10^{11}) and income (6.99×10^{10}). In contrast, variables such as rent (5.85×10^{8}) and equity (1.57×10^{9}) display comparatively low variability, indicating limited discriminatory power for segmentation. Based on this criterion, five variables—income, debt, net worth, houses, and total assets—were retained as the core features for clustering, as they capture the most significant variation in household financial behavior.

5.1 Trimmed Variance Feature Selection

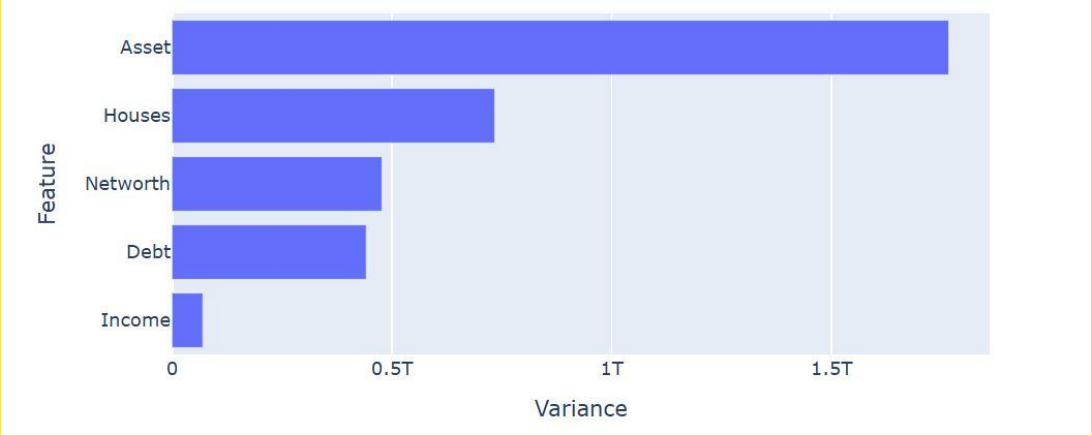

Figure 1*: Distribution of selected high-variance variables after trimming*

The optimal number of clusters was determined using both the elbow method and Silhouette analysis. The inertia curve shows a pronounced decline between K = 1 and K = 3, after which the marginal reduction becomes negligible, indicating an elbow point at K = 3. Complementary to this, the Silhouette score reaches approximately 0.85 at K = 2 and 0.68 at K = 3. While K = 2 provides slightly higher cohesion, K = 3 offers a more informative segmentation structure with acceptable separation and improved interpretability. Based on this trade-off, K = 3 was selected as the optimal clustering solution.

5.2 Determination of the Optimal Number of Clusters

K-Means Model: Inertia vs number of clusters

Figure 2*: Inertia as a function of the number of clusters (Elbow Method)*

Silhouette analysis was conducted to assess the quality of cluster separation. As shown in Figure 2, the Silhouette Score reaches a high value of approximately 0.85 at K = 2 and 0.68 at K = 3, indicating strong intra-cluster cohesion and clear inter-cluster separation.

5.3 K-Means Model: Silhouette Score vs Number of Clusters

***Figure 3**: Silhouette Score across different numbers of clusters*

Based on both the Elbow Method and Silhouette analysis, K = 3 is selected as the optimal number of clusters. This choice balances model simplicity, interpretability, and clustering performance, ensuring that the resulting segments are both statistically meaningful and economically interpretable.

5.4 Cluster Naming and Financial Interpretation

The K-Means algorithm was then applied using the selected features and optimal cluster size. The resulting clustering structure demonstrates clear separation in the feature space, as confirmed by dimensionality reduction visualization. The clusters exhibit minimal overlap and well-defined boundaries, indicating that the algorithm successfully identified distinct behavioral groups within the dataset. Stability checks using multiple initializations confirm the robustness of the clustering solution.

5.5 Mean Household Finances by Cluster

***Figure 4**: Mean financial indicators across clusters*

Figure 4 illustrates clear differences in financial structure across clusters. Cluster 1 exhibits the highest levels of income, assets, and debt, indicating a high-capacity but highly leveraged financial profile. In contrast, Cluster 0 shows consistently low values across all financial indicators, reflecting severe financial constraints and limited capacity for financial accumulation. Cluster 2 occupies an intermediate position, indicating a transitional group experiencing gradual wealth accumulation and relatively balanced financial conditions.

5.6 PCA Representation of Clusters

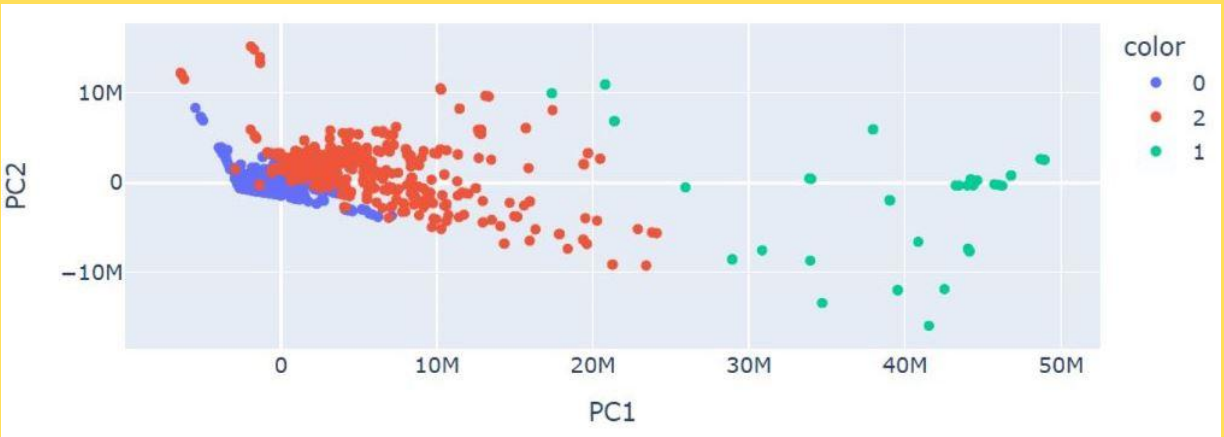

***Figure 5**: PCA projection of households showing cluster separation*

Figure 5 presents a three-dimensional PCA projection of the clustered data. The visual ization shows clear separation between the three clusters, with limited overlap, indicating that the clustering algorithm successfully identified distinct groups within the dataset.

Comparative analysis across clusters highlights substantial structural differences in financial behavior. The contrast between low-income constrained households and asset-rich but leveraged households demonstrates that credit constraint does not arise from a single underlying condition. Instead, it reflects multiple financial realities, including income insufficiency, leverage pressure, and transitional financial growth.

Overall, the empirical results demonstrate that credit-constrained households are not homogeneous but instead form distinct behavioral segments with fundamentally different financial structures. The combination of robust feature selection, statistically validated clustering, and clear financial differentiation provides strong empirical support for the segmentation framework and establishes a solid foundation for subsequent interpretation and policy analysis.

6 Discussion and conclusion

The empirical results provide clear and internally consistent evidence that credit-constrained households exhibit substantial heterogeneity in their financial structures and behavioral patterns. Using a final analytical sample of 4,458 households, the clustering model identifies three statistically distinct segments (K = 3), a choice supported by the Elbow Method and a Silhouette score of 0.68, indicating strong cluster cohesion and separation. These results demonstrate that the identified clusters are not only computational artifacts but reflect structurally meaningful differences in household financial behavior.

A closer examination of the cluster profiles reveals economically interpretable patterns. Cluster 0, representing the most financially constrained group, is characterized by low average income (Rwf 460,800), moderate debt (Rwf 330,800), and very low net worth (Rwf 222,700). This cluster exhibits minimal asset ownership and limited financial buffers, indicating severe liquidity constraints and high vulnerability to economic shocks. In this segment, credit constraint is primarily driven by insufficient financial capacity and restricted access to productive financial opportunities.

In contrast, Cluster 1 represents a high-capacity but highly leveraged segment. Households in this group report high average income (Rwf 2.89 million), substantial debt levels (Rwf 6.88 million), and high net worth (Rwf 21.86 million). Despite their strong asset base, the disproportionately high debt burden indicates elevated financial risk and potential exposure to repayment stress. This finding highlights a critical nuance: credit constraint does not necessarily imply low wealth but may instead reflect structural inefficiencies in credit allocation or risk management within financially active households.

Cluster 2 captures an intermediate and transitional group, with moderate income (Rwf 1.16 million), moderate debt (Rwf 2.79 million), and net worth of Rwf 3.27 million. This segment reflects households undergoing gradual wealth accumulation, balancing income generation with manageable levels of financial exposure. The presence of this cluster suggests that credit constraint evolves along a continuum, rather than existing as a fixed condition, with households potentially transitioning across financial states over time.

These empirical findings extend theoretical expectations from the Life-Cycle Hypothesis and the Permanent Income Hypothesis. While these models assume access to credit markets that facilitates consumption smoothing, the results demonstrate that credit constraints disrupt this mechanism, leading to observable deviations such as low-income households exhibiting constrained consumption, asset-rich households maintaining high leverage, and transitional households balancing competing financial pressures. This indicates that real-world financial behavior under constraint is shaped by both capacity limitations and structural market frictions.

Methodologically, the study demonstrates the effectiveness of combining trimmed variance feature selection with K-Means clustering in high-dimensional financial datasets. The feature selection process reduces the dataset from approximately 25 variables to 15 variables, and ultimately to 5 core financial indicators (income, debt, net worth, houses, assets), ensuring that clustering is driven by the most informative dimensions of financial behavior. This contributes to both the stability and interpretability of the model, as evidenced by the clear statistical separation between clusters.

From a policy and institutional perspective, the results provide strong justification for moving beyond uniform financial inclusion strategies. The quantitative differences across clusters indicate that a one-size-fits-all approach is inefficient. Instead, Cluster 0 requires income support and financial protection mechanisms, Cluster 1 requires debt regulation and risk management interventions, and Cluster 2 benefits from asset-building and financial growth programs. This segmentation-based framework allows for more precise targeting of interventions and more efficient allocation of financial resources.

In conclusion, the study establishes that credit-constrained households are structurally heterogeneous, with three distinct financial profiles supported by strong statistical evidence. By integrating robust feature selection, clustering, and validation within a coherent analytical pipeline, the research provides both a methodological contribution and actionable empirical insights. The findings underscore the importance of data-driven, segment-specific approaches in financial inclusion policy and demonstrate the value of unsupervised machine learning in uncovering latent behavioral structures within complex socio-economic data.

References

[1] Beck, T., Demirgüç-Kunt, A., & Levine, R. (2007). Finance, inequality and the poor. Journal of Economic Growth, 12(1), 27–49.

[2] Allen, F., Demirgüç-Kunt, A., Klapper, L., & Peria, M. S. M. (2016). Financial inclusion: Global evidence. World Bank Economic Review, 30(1), 2–30.

[3] Kumar, A. (2021). Clustering algorithms in banking: A review. International Journal of Bank Marketing, 39(2), 230–255.

[4] Deaton, A. (1991). Saving and liquidity constraints. Econometrica, 59(5), 1221–1248.

[5] White, B. (2019). Demographic shifts and household debt: Evidence from the SCF. Journal of Applied Econometrics, 34(6), 900–920.

[6] Hall, A. (2025). AI-driven policymaking in Africa. African Development Review, 37(1), 10–25.

[7] Thompson, S. (2020). Behavioral manifestations of credit constraints. Journal of Consumer Research, 47(5), 789–805.

[8] Zhao, X., et al. (2021). Exploring the "missing middle" in credit markets. Journal of Financial Services Research, 60(2), 195–220.

[9] Nguyen, T. (2023). Dimensionality reduction in household wealth surveys. Computational Economics, 41(4), 567–589.

[10] Modigliani, F., & Brumberg, R. (1954). Utility analysis and the consumption function. In Post-Keynesian Economics, 3(2), 388–436.

[11] Friedman, M. (1957). A Theory of the Consumption Function. Princeton University Press.

[12] Chen, L., et al. (2022). Unsupervised learning for credit risk: A K-Means approach. FinTech Review, 8(3), 101–115.

[13] Williams, R. (2019). The 2019 Survey of Consumer Finances: A methodological overview. Federal Reserve Bulletin, 105(3), 1–35.