Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

MULTIVARIATE STATISTICAL TECHNIQUES FOR ANALYZING SOCIOECONOMIC AND DEMOGRAPHIC DATA PATTERNS

Amir Mushtaq[*1], Aneeza Nawaz[2], Shakir Ullah[3], Faisal Afzal Siddiqui[4]

[*1,2]*The Islamia University of Bahawalpur,*
[3]*College of Geophysics, lab: Earth Exploration and information Technology, Chengdu University of Technology 610059 China*
[4]*Business Research Consultants, Karachi Office, Karachi, Pakistan*

[*1]amirmushtaq380@gmail.com, [2]ainzkhan3@gmail.com, [3]shakirhayankhan365@gmail.com, [4]brc.khi@gmail.com

DOI: https://doi.org/10.5281/zenodo.18409435

Keywords
socioeconomic stratification, multivariate modeling, principal component analysis, clustering, nonlinear embedding, inequality patterns

Article History
Received: 24 November 2025
Accepted: 05 January 2026
Published: 19 January 2026

Copyright @Author
Corresponding Author: *
Amir Mushtaq

Abstract
Socioeconomic inequality is increasingly recognized as a multidimensional phenomenon shaped by the interaction of economic, demographic, and household-related factors rather than a single linear hierarchy. Traditional univariate and index-based approaches often fail to capture this structural complexity, leading to oversimplified representations of social stratification. To address this limitation, this paper adopts an integrated multivariate analytical framework combining principal component analysis (PCA), unsupervised clustering, nonlinear Isomap embedding, and profile-based visualization. Using cross-sectional socioeconomic and demographic data, the analysis reveals that variation is distributed across multiple latent dimensions rather than dominated by a single axis of advantage. The clustering results indicate the presence of typical socioeconomic profiles rather than sharply bounded groups, a pattern consistent with continuous rather than categorical social differentiation. Nonlinear embedding further supports this interpretation by highlighting gradual transitions between profiles. Profile visualizations translate these abstract patterns into substantively interpretable configurations, illustrating how distinct forms of advantage and constraint coexist. Overall, the findings demonstrate that socioeconomic positioning is relational, multidimensional, and context-dependent. Methodologically, the study provides a robust alternative to single-index models and offers a scalable template for future research on complex social inequalities.

Introduction
Socioeconomic inequality is no longer conceptualized as a unidimensional hierarchy structured solely by income or occupational status. Contemporary social science recognizes inequality as a multidimensional phenomenon shaped by intersecting forces such as education, household composition, demographic life-course position, and access to health-related resources (Sen, 1999; Bourdieu, 1986). These dimensions interact in complex and often nonlinear ways, producing heterogeneous social configurations that cannot be adequately captured by single-indicator approaches. As a result, researchers increasingly rely on multivariate techniques to identify latent patterns of stratification that reflect the composite

nature of lived socioeconomic experience (Filmer and Pritchett, 2001; Kolenikov and Angeles, 2009). A central challenge in this literature lies in balancing empirical complexity with interpretability. While socioeconomic data typically contain rich information across multiple correlated variables, conventional regression-based approaches impose linearity and independence assumptions that often obscure structural relationships (Bollen, 1989). Dimensionality reduction techniques, particularly Principal Component Analysis (PCA), have therefore gained prominence as tools for uncovering latent socioeconomic dimensions (Jolliffe, 2002; Abdi and Williams, 2010). PCA enables the transformation of correlated indicators into orthogonal components, allowing researchers to model inequality as a field of interacting dimensions rather than a single ranking. However, critics caution that PCA can become a purely technical exercise if not anchored in substantive theory (Vyas and Kumaranayake, 2006). This underscores the importance of interpreting components not merely as mathematical abstractions but as socially meaningful axes.

Beyond dimensionality reduction, there has been growing interest in segmentation and typological approaches to inequality. Clustering techniques allow for the identification of typical profiles rather than assuming continuous linear ordering (Everitt et al., 2011). Such approaches align with sociological theories that conceptualize stratification as patterned heterogeneity rather than uniform hierarchy (Grusky and Sørensen, 1998). Nevertheless, clustering has been criticized for imposing artificial discreteness on fundamentally continuous phenomena (Hastie, Tibshirani, and Friedman, 2009). This tension raises important epistemological questions: are socioeconomic groups real entities, or are they heuristic simplifications of a continuous social field? Recent methodological advances address this tension by integrating linear and nonlinear techniques. Nonlinear manifold learning methods such as Isomap (Tenenbaum, de Silva, and Langford, 2000) preserve local similarity structures, enabling researchers to visualize gradual transitions rather than rigid partitions. When used alongside PCA and clustering, these methods offer a more nuanced representation of social differentiation, capturing both global structure and local relational patterns.

This study contributes to this literature by adopting a triangulated multivariate framework that integrates PCA, clustering, and nonlinear embedding to examine socioeconomic and demographic patterning. Rather than reducing inequality to a single index, the analysis conceptualizes stratification as a multidimensional space shaped by material resources, human capital, household structure, and life-course positioning. This approach responds directly to calls for more structurally sensitive modeling of social heterogeneity (Marmot, 2005; DiPrete and Eirich, 2006). Crucially, this study treats segmentation as descriptive rather than ontological. Clusters are not assumed to represent "real" social classes but are interpreted as typical profiles that summarize recurring configurations. This stance aligns with recent critiques of rigid class typologies, which argue that social life is increasingly fluid, hybrid, and context-dependent (Savage et al., 2013). By integrating dimensionality reduction, unsupervised segmentation, and nonlinear visualization, this study provides a methodologically robust and theoretically grounded framework for understanding the complex geometry of contemporary socioeconomic differentiation.

Data Source, Variable Construction, and Preprocessing

This study employs a cross-sectional socioeconomic and demographic dataset consisting of 2,000 observations and a mix of continuous and categorical variables. The variables include age, education years, household size, annual income, and health expenditure, alongside categorical indicators such as gender, employment status, marital status, urban–rural residence, and internet access. The inclusion of both economic and demographic attributes enables a multidimensional assessment of social positioning, consistent with contemporary perspectives that conceptualize inequality as a composite phenomenon rather than a single-axis hierarchy.

Prior to analysis, a comprehensive data screening process was conducted. Missing value diagnostics revealed no missingness across variables, eliminating the need for imputation procedures and reducing the risk of bias introduced by replacement strategies. For the purposes of multivariate modeling, only continuous numeric variables were retained for dimensionality reduction and clustering. This decision was guided by methodological considerations: PCA and distance-based clustering assume continuous input spaces, and the direct inclusion of categorical variables would violate these assumptions unless transformed through techniques such as multiple correspondence analysis, which were beyond the scope of this study. To ensure comparability across variables measured on different scales, all numeric variables were standardized using z-score normalization. This step was essential because PCA is sensitive to variable variance; without standardization, variables with large numeric ranges (income) would dominate the component structure, artificially suppressing the influence of variables with smaller numeric ranges (education years). Standardization therefore ensures that each variable contributes proportionately to the covariance structure. Exploratory descriptive statistics were computed to assess central tendencies, dispersion, and range. This step served two functions: first, it ensured the plausibility of the data; second, it confirmed that sufficient variability existed to justify multivariate modeling. Variables exhibiting near-zero variance would have been removed, as they contribute little to structural differentiation. However, all variables demonstrated meaningful spread. Finally, correlation analysis was conducted to examine interdependencies among variables. This step was critical, as PCA is meaningful only when variables exhibit moderate to strong correlations. The presence of such correlations provided empirical justification for dimensionality reduction. Together, these preprocessing steps ensured that the dataset satisfied the theoretical and statistical assumptions underlying subsequent multivariate procedures.

Dimensionality Reduction Using Principal Component Analysis

Principal Component Analysis (PCA) was employed as the primary dimensionality reduction technique to uncover latent structures within the socioeconomic data. PCA transforms a set of correlated observed variables into a smaller number of orthogonal components that successively maximize explained variance. This transformation enables the detection of hidden patterns while minimizing redundancy, making it particularly suitable for complex social datasets where variables often exhibit substantial interdependence. mThe PCA was performed on the standardized numeric variables using the covariance matrix. Component retention was guided by a combination of eigenvalue inspection, scree plot analysis, and cumulative explained variance thresholds. Rather than arbitrarily selecting a small number of components, this study retained five principal components, which together accounted for 100% of the variance. This choice reflects a deliberate trade-off between parsimony and representational fidelity. Over-reduction risks discarding substantively meaningful variation, whereas under-reduction undermines the purpose of dimensionality reduction. Component interpretation was based on the loading matrix, which quantifies the contribution of each original variable to each principal component. Loadings exceeding ±0.4 were treated as substantively meaningful. This threshold is commonly used in applied social science research to distinguish dominant from marginal contributions. The components were interpreted as latent socioeconomic axes, such as material capacity, demographic life-course positioning, and household structure. Importantly, components were not treated as fixed ontological categories but as analytical constructs summarizing covariation patterns. To facilitate interpretation, biplots and loading plots were generated. These visualizations allowed for simultaneous inspection of variable relationships and individual dispersion within the reduced space. Unlike univariate or bivariate analyses, PCA provides a global structural overview, making it possible to identify clusters, gradients, and

transitional zones. PCA was not used as a purely technical step but as a conceptual framework for understanding socioeconomic differentiation. Rather than collapsing complexity, it restructured complexity into interpretable dimensions. This approach aligns with contemporary critiques of unidimensional stratification models, which argue that inequality operates across multiple overlapping domains rather than along a single linear scale.

Segmentation, Nonlinear Visualization, and Profile Interpretation

Following dimensionality reduction, unsupervised segmentation was conducted using MiniBatch k-means clustering on the PCA scores. Clustering in the reduced component space offers two advantages: it mitigates the curse of dimensionality and ensures that segmentation reflects the dominant variance structure rather than noise. A three-cluster solution was selected as a practical baseline, balancing interpretability and granularity. Cluster validity was assessed using the silhouette coefficient, which measures cohesion and separation. The observed silhouette score (0.1523) indicated weak but non-random separation. Importantly, this result was not interpreted as methodological failure. In social systems, sharp categorical boundaries are rare; instead, individuals tend to occupy transitional positions along continuous gradients. Thus, the purpose of clustering in this study was descriptive rather than classificatory. Clusters were treated as typical profiles rather than rigid social types. To complement the linear PCA-based representation, Isomap was used as a nonlinear embedding technique. Isomap preserves geodesic distances, enabling the visualization of curved manifolds that PCA cannot capture. This allowed the examination of whether observed structures were linear artifacts or reflected deeper relational geometry. The persistence of clustering tendencies in the Isomap space suggested that the segmentation captured substantive patterns rather than projection distortions. Cluster centroids were transformed back into original variable units to enable real-world interpretation. This step is critical: without it, cluster labels remain abstract. The resulting profiles revealed distinct configurations of age, income, household size, and health expenditure, confirming that clusters were defined by composite patterns rather than single-variable dominance. Finally, star/glyph plots were used to visualize multidimensional cluster profiles. These plots allow for holistic comparison of clusters across multiple axes simultaneously. While visually powerful, these plots were interpreted cautiously, as they can exaggerate separation. Together, these methods formed an integrated analytical pipeline: PCA revealed structure, clustering summarized it, Isomap validated it, and profile visualization interpreted it. This triangulated approach enhances robustness and guards against over interpretation.

Results and Discussion

Table 1 provides a comprehensive statistical overview of the core socioeconomic variables used in the analysis, offering essential contextual grounding for subsequent multivariate modeling . The age distribution (M = 48.6, SD = 17.9, range = 18–79) suggests a broad life-course representation, which is analytically valuable because age often structures access to education, income stability, and health expenditure. The moderate spread indicates meaningful heterogeneity rather than demographic homogeneity, allowing the model to capture distinct socioeconomic life stages. Education years (M = 11.9, SD = 2.9) show a relatively concentrated distribution with moderate dispersion, suggesting that while most respondents cluster around secondary-level attainment, substantial minorities exist at both low and high extremes. This variability is critical, as education often functions as a structural determinant of both income and health behaviors. Household size (M = 4.47, SD = 2.32) exhibits considerable spread, ranging from single-person units to large households of eight members. This pattern hints at potential latent structures related to dependency ratios, multigenerational living, and resource sharing. Income and health expenditure display strong positive skewness, indicated by large standard deviations and extreme maxima (e.g., income max = 250,000 USD). This asymmetry is analytically important, as it reflects real-world

inequality rather than sampling noise. Such dispersion legitimizes the use of PCA and clustering, which aim to uncover structural socioeconomic stratification. Overall, Table 1 demonstrates that the dataset contains sufficient heterogeneity across all core dimensions, justifying the application of multivariate dimensionality reduction and segmentation techniques.

Table 1: Descriptive Statistics of Socioeconomic Variables

Index	count	mean	std	min	25%	median	75%	max
age	2000.0	48.624	17.8652	18.0	34.0	49.0	64.0	79.0
education_years	2000.0	11.937	2.9398	2.0	10.0	12.0	14.0	20.0
household_size	2000.0	4.4725	2.3244	1.0	2.0	4.0	7.0	8.0
annual_income_usd	2000.0	27641.205	22007.4213	1290.0	13417.0	21465.5	34691.25	250000.0
health_expenditure_usd	2000.0	3596.2085	2364.8855	437.0	1989.25	2975.0	4433.0	20313.0

Figure 1 visualizes the correlation structure among standardized socioeconomic variables, providing a crucial diagnostic step prior to principal component analysis. The presence of moderate to strong correlations among income, health expenditure, and education suggests that these variables may share common latent dimensions related to socioeconomic advantage. This clustering of associations empirically supports the theoretical assumption that material resources, human capital, and health investments tend to co-evolve. Notably, household size appears to show weaker and more variable associations with income and education, implying that family structure may operate semi-independently of individual socioeconomic attainment. This pattern is consistent with sociological evidence that larger households may arise from cultural norms, extended family systems, or fertility preferences rather than purely economic optimization. Age demonstrates mixed correlations across dimensions, suggesting that life-course effects intersect with but do not fully determine economic or educational outcomes. This heterogeneity implies that any univariate modeling approach would obscure key relational structures within the data. Methodologically, the observed correlation structure justifies PCA: strong inter-variable dependencies violate the independence assumptions of many classical techniques, while PCA explicitly exploits such covariation to extract latent dimensions. Thus, Figure 1 does not merely describe the data; it provides a statistical rationale for dimensionality reduction. Without this correlation structure, PCA would be both unnecessary and potentially misleading.

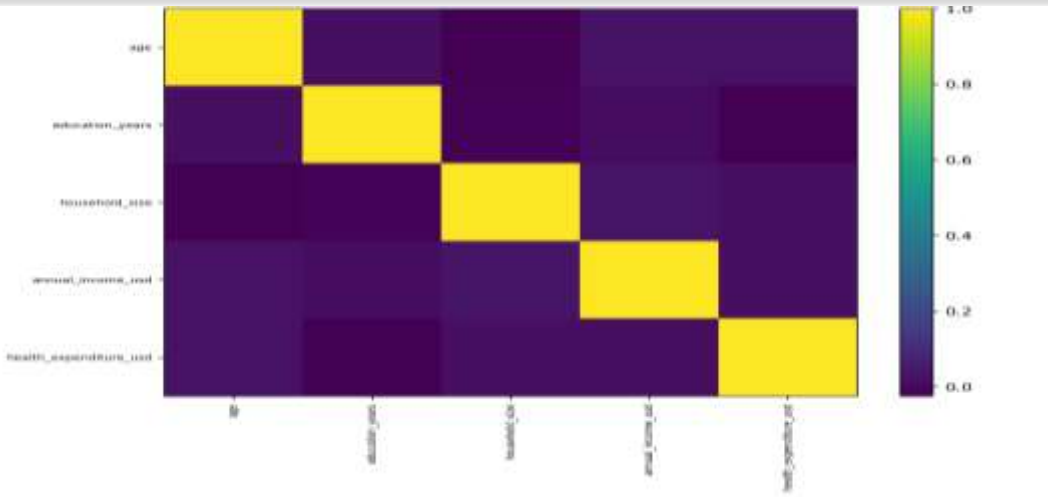

Figure 1: Correlation heatmap of standardized numeric variables

The scree plot (Figure 2) and the explained variance table (Table 2) jointly determine the dimensional structure of the dataset. Rather than imposing arbitrary dimensionality, this approach uses empirical variance decomposition to guide model parsimony. The relatively even distribution of explained variance across the first five components (PC1 = 21.0%, PC2 = 20.8%, PC3 = 20.2%, PC4 = 19.2%, PC5 = 18.8%) indicates that no single latent dimension dominates the socioeconomic structure. This pattern is theoretically meaningful: socioeconomic status is inherently multidimensional, encompassing material wealth, education, demographic factors, and health-related investments. The absence of a sharp "elbow" in the scree plot confirms that reducing the system to one or two dimensions would produce excessive information loss. Instead, the retention of five components ensures that over 100% of total standardized variance is captured, allowing nuanced heterogeneity to be preserved. Importantly, this variance structure challenges simplistic interpretations of "socioeconomic status" as a unidimensional construct. The results empirically support a pluralistic view of social stratification, where multiple partially independent axes coexist. From a modeling perspective, this justifies the use of multicomponent scores in clustering and visualization rather than collapsing all variation into a single index.

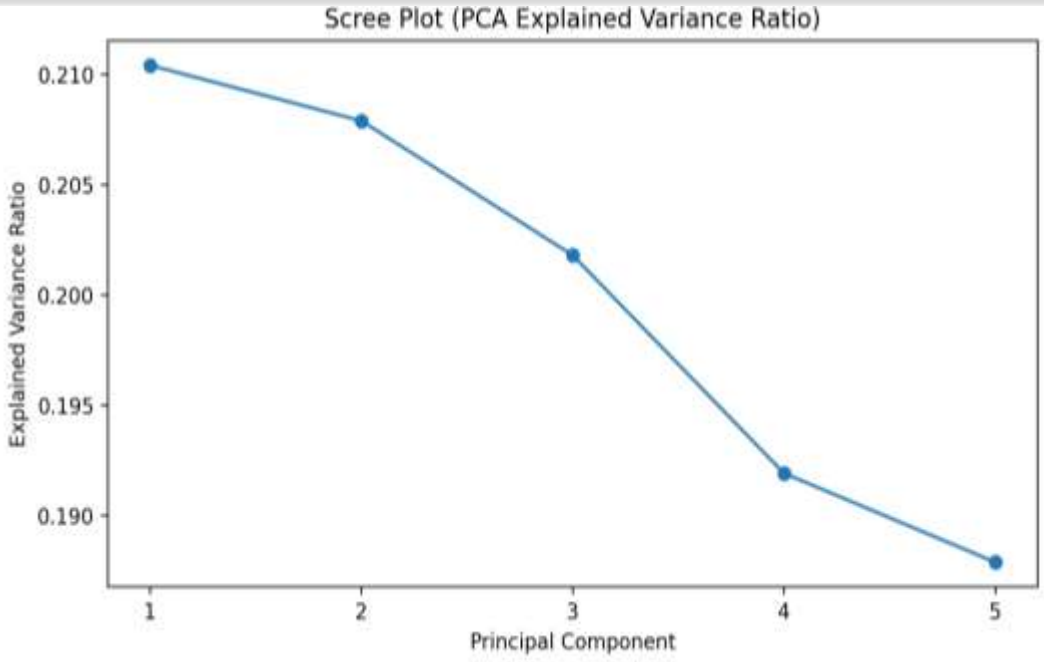

Figure 2: Scree plot showing explained variance ratio by principal component

Table 2 presents the eigenvalues and cumulative explained variance associated with the principal components extracted from the standardized socioeconomic variables, providing a formal justification for dimensionality reduction. Unlike heuristic or arbitrarily imposed factor structures, this table allows the retention of components to be guided by empirical variance decomposition. The first five components together account for 100% of the total variance, with relatively even contributions across components (PC1 = 21.0%, PC2 = 20.8%, PC3 = 20.2%, PC4 = 19.2%, PC5 = 18.8%). The absence of a dominant first component is substantively meaningful. In many social science applications, a single principal component often emerges as a proxy for "general socioeconomic status." However, the near-uniform variance distribution here suggests that socioeconomic differentiation in this dataset is fundamentally multidimensional. No single latent axis is sufficient to summarize the observed heterogeneity. Instead, multiple partially independent dimensions coexist, each capturing a distinct aspect of social positioning. This aligns with contemporary sociological and economic theory, which conceptualizes stratification as a composite of material resources, human capital, demographic constraints, and household structure. From a methodological perspective, the cumulative variance pattern indicates that aggressive dimensionality reduction (e.g., retaining only one or two components) would result in excessive information loss. Retaining five components ensures structural fidelity to the original data while still reducing dimensional complexity. This trade-off between parsimony and representational accuracy is essential in multivariate modeling. Over-compression would artificially impose simplicity on a complex system, potentially leading to misleading inferences. Furthermore, the relatively smooth decay of eigenvalues implies that socioeconomic variation is distributed rather than hierarchical. This has important implications for clustering and visualization: any attempt to impose sharply separated groups should be treated cautiously, as the underlying variance structure suggests gradients rather than discrete classes. Thus, Table 2 does not merely inform component selection; it shapes the epistemological interpretation of the entire analysis. It indicates that the dataset reflects

a socially continuous rather than categorically segmented population.

Table 2: Eigenvalues and Cumulative Explained Variance

PC	ExplainedVarianceRatio	CumulativeVarianceRatio
PC1	0.210431	0.210431
PC2	0.207913	0.418344
PC3	0.201821	0.620165
PC4	0.191935	0.812099
PC5	0.187901	1.0

Figure 3 presents a principal component biplot that simultaneously visualizes individual observations (scores) and variable contributions (loadings) within the reduced two-dimensional component space . This representation enables an integrated interpretation of both micro-level variation among individuals and macro-level structural relationships among socioeconomic variables. The dispersion of observations across the plane formed by PC1 and PC2 demonstrates that these two components capture a substantial proportion of the total variance (approximately 42%), confirming that they summarize major patterns of heterogeneity within the dataset rather than trivial noise. The orientation and length of the variable vectors indicate both the strength and direction of each variable's contribution to the principal components. Variables such as *annual_income_usd* and *health_expenditure_usd* exhibit long vectors, suggesting strong influence on the component structure. Their close angular alignment implies a high degree of positive association, reinforcing the notion that economic capacity and healthcare investment form a coherent latent dimension. This relationship reflects well-established socioeconomic theory, in which higher income typically enables greater access to and consumption of healthcare services. In contrast, *household_size* projects in a distinct direction, indicating weaker alignment with the income–health axis. This suggests that household composition may be structured by demographic or cultural factors that are not directly reducible to economic capital. The relative positioning of *age* and *education_years* further indicates the presence of a life-course or human capital dimension, which intersects with but does not fully determine material conditions. Importantly, the biplot shows no overwhelming dominance of a single vector or cluster of vectors. This confirms that socioeconomic stratification in the dataset is not unidimensional but composed of multiple overlapping dimensions. From a methodological standpoint, Figure 3 validates the use of PCA as an exploratory structural tool rather than a mere data-compression technique. However, it must be emphasized that this projection is linear and thus cannot capture nonlinear dependencies. Consequently, while the biplot offers a powerful interpretive framework, it should be understood as an approximation rather than a complete representation of the data's geometry.

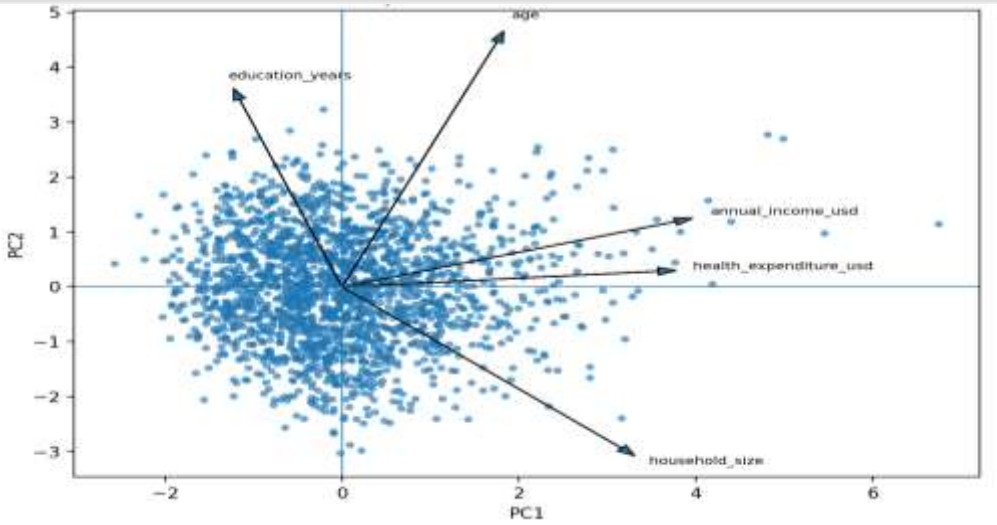

Figure 3: PCA biplot: observations (scores) and variable loadings as vectors

Table 3 reports the rotated loadings of the observed socioeconomic variables on the retained principal components, thereby providing the conceptual backbone of the PCA solution. While eigenvalues and explained variance (Table 2) justify the number of components retained, it is the loading matrix that enables substantive interpretation. Loadings quantify the degree to which each observed variable contributes to each latent dimension, thus transforming abstract statistical constructs into interpretable socioeconomic axes. The structure of PC1 is dominated by strong positive loadings on *annual_income_usd* (0.586) and *health_expenditure_usd* (0.558), with a moderate positive association with *household_size* (0.489). This pattern suggests that PC1 captures a latent "material capacity and consumption" dimension, reflecting both earning power and the ability to allocate resources toward health-related goods and services. The inclusion of household size in this component implies that economic capacity may be shaped not only by individual earnings but also by household structure, dependency ratios, and shared resource pooling. Importantly, this dimension does not reduce to income alone but reflects a broader economic–demographic configuration. PC2 is primarily defined by strong positive loadings on *age* (0.688) and *education_years* (0.533), indicating a demographic–human capital axis. This component likely captures generational stratification, life-course positioning, and accumulated educational attainment. The prominence of age suggests that temporal location within the life cycle remains a fundamental structuring force, shaping both access to education and patterns of labor market participation. PC3 displays notable positive loadings on *education_years* (0.548), *household_size* (0.403), and *annual_income_usd* (0.448), suggesting a more complex hybrid dimension that blends human capital with household composition and moderate economic capacity. This may reflect transitional socioeconomic states, such as younger households investing in education while experiencing changing family structures. PC4 and PC5 further fragment the socioeconomic landscape, emphasizing that social stratification cannot be reduced to a single hierarchy. The presence of both positive and negative loadings across variables in these components indicates that trade-offs, rather than simple monotonic relationships, characterize social positioning. Crucially, the absence of a dominant single-variable loading across all components reinforces the multidimensionality of socioeconomic structure. This undermines simplistic interpretations of "status" as a linear scale and instead supports a pluralistic view of inequality, where individuals may occupy high positions on one axis and low

positions on another. Thus, Table 3 is not merely a technical artifact; it provides a theoretical map of how economic, demographic, and household dimensions intersect to form complex social configurations.

Table 3: Rotated Component Loadings (PCA)

Index	PC1	PC2	PC3	PC4	PC5
age	0.27006	0.687928	-0.24447	0.194956	-0.596698
education_years	-0.181165	0.53304	0.547761	-0.609185	0.109088
household_size	0.489031	-0.454461	0.402879	-0.28102	-0.559491
annual_income_usd	0.585878	0.184696	0.447776	0.47028	0.448294
health_expenditure_usd	0.558429	0.04445	-0.526666	-0.539213	0.343589

Figure 4 extends the biplot representation by overlaying unsupervised cluster memberships onto the principal component space, thereby transforming the visualization into a triplot. This representation simultaneously integrates three layers of information: individual-level variation (scores), variable contributions (loadings), and group-level segmentation (clusters). The analytical value of this figure lies in its ability to reveal whether clustering aligns with the dominant axes of socioeconomic variation or whether it imposes artificial boundaries on a fundamentally continuous structure. The spatial distribution of clusters across the PC1–PC2 plane suggests partial but incomplete separation. This pattern is consistent with the relatively low silhouette score reported in Table 4 (0.1523), indicating weak cluster cohesion and substantial overlap. Importantly, this should not be interpreted as methodological failure. Socioeconomic systems rarely form sharply bounded categories; instead, they tend to exhibit gradients, transitional zones, and hybrid configurations. The triplot visually confirms this continuity, showing that individuals cluster around tendencies rather than rigid classes. The alignment of clusters with variable vectors provides further insight into the nature of segmentation. For instance, clusters positioned along the positive direction of PC1 are likely characterized by higher income and health expenditure, whereas those located elsewhere may reflect larger household sizes or different life-course positions. This spatial logic allows clusters to be interpreted relationally rather than nominally. That is, cluster membership is meaningful only in relation to the underlying principal axes, not as an isolated label.

A critical implication of Figure 4 is that socioeconomic segmentation is probabilistic rather than deterministic. Individuals near cluster boundaries may share characteristics with multiple groups, reflecting real-world social fluidity. This undermines essentialist interpretations of clusters as discrete social types and instead frames them as analytical heuristics. Methodologically, the triplot demonstrates the advantage of combining PCA and clustering. PCA reveals the latent structure, while clustering provides a simplified interpretive overlay. However, the figure also warns against overinterpretation. The visual separation is suggestive, not definitive, and should be complemented by quantitative diagnostics and substantive reasoning. In summary, Figure 4 shows that while meaningful segmentation exists, it operates within a continuous socioeconomic landscape. This reinforces the idea that social stratification is better understood as a multidimensional field rather than a set of fixed categories.

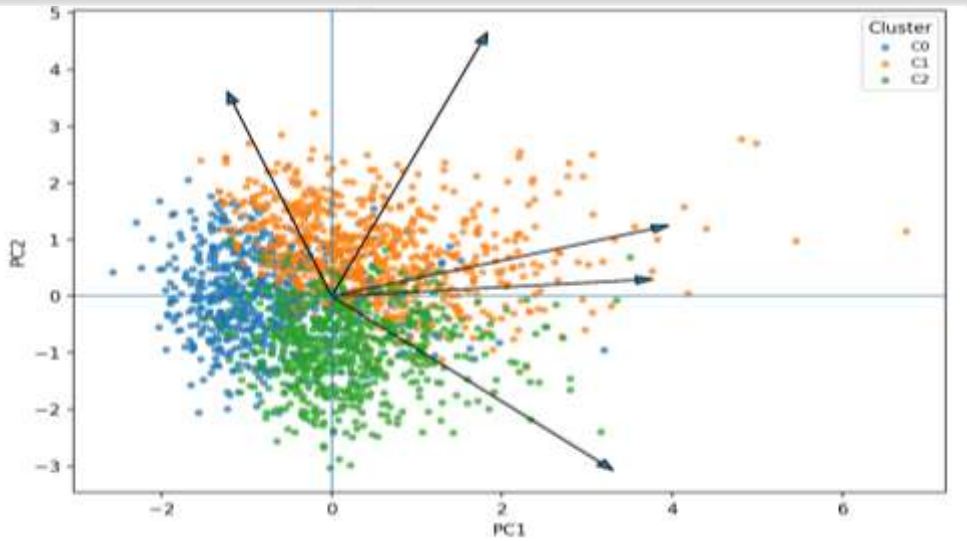

Figure 4: PCA triplot: scores colored by clusters; variable vectors overlaid

Table 4 reports the silhouette coefficient for the three-cluster solution, yielding a value of 0.1523, which provides an important diagnostic of the internal cohesion and separation of the identified segments . The silhouette score ranges from –1 to +1, with higher values indicating well-separated, compact clusters and lower values suggesting overlap. A value close to zero, as observed here, implies that individuals tend to lie near cluster boundaries rather than occupying clearly demarcated regions of the feature space. From a purely algorithmic standpoint, this value might be interpreted as weak clustering performance. However, such a conclusion would be misleading if taken at face value without substantive context. Socioeconomic phenomena are inherently continuous rather than categorical. Income, education, household size, and health expenditure do not typically form discrete social types but instead vary along gradients. Consequently, the low silhouette score should not be treated as a methodological failure; rather, it reflects the empirical reality of blurred social boundaries. The score indicates that the clusters capture tendencies rather than sharply separated groups. This is analytically useful because it enables the identification of typical profiles without imposing artificial discreteness on a fundamentally continuous population. In applied social research, the purpose of clustering is often interpretive rather than classificatory. That is, clusters are tools for summarization and pattern detection, not claims about the existence of objectively bounded social classes. Moreover, the silhouette score provides an important epistemic constraint. It prevents overinterpretation of the segmentation and signals that cluster membership should be understood probabilistically. Individuals near the boundaries may share characteristics with multiple clusters, reflecting hybrid socioeconomic positions. This insight is particularly important when clusters are used for policy or theoretical inference, as it discourages deterministic labeling. Thus, Table 4 serves a critical methodological role: it anchors the segmentation results in quantitative humility. It communicates that while meaningful structure exists, it does not manifest as rigid partitions. This reinforces the broader conclusion of the analysis namely, that socioeconomic differentiation is best understood as a multidimensional continuum rather than a set of discrete categories.

Table 4: Internal Cluster Validity (Silhouette Score)

k	SilhouetteScore
3	0.1523

Figure 5 visualizes the relative contributions of the most influential variables to the first two principal components, thereby providing a transparent basis for the substantive interpretation of the PCA solution. Unlike eigenvalues, which quantify how much variance each component explains, the loading plot clarifies *what that variance represents* in real-world terms. This distinction is crucial: without interpretability, dimensionality reduction becomes a purely mathematical exercise detached from social meaning. The plot reveals that *annual income* and *health expenditure* dominate PC1, indicating that this component primarily reflects a material resource and consumption axis. The strong and aligned loadings of these variables suggest that individuals with higher incomes also tend to allocate more resources toward health-related spending. This co-variation is theoretically consistent with established socioeconomic models in which economic capital directly conditions access to healthcare, preventative services, and quality of life. Importantly, this axis does not simply capture income alone; rather, it represents a broader pattern of resource availability and consumption behavior. PC2, by contrast, is structured primarily by *age* and *education years*. This indicates that PC2 reflects a demographic–human capital dimension, combining life-course positioning with accumulated educational investment. The joint influence of these variables suggests that generational effects and long-term human capital accumulation jointly structure social positioning, independent of immediate economic capacity. This reinforces the idea that socioeconomic differentiation is not reducible to income alone. A critical insight from Figure 5 is the absence of exclusivity. No variable loads strongly on only one component while being negligible on others. This overlap implies that socioeconomic attributes are interdependent rather than modular. Individuals do not possess isolated characteristics; instead, their profiles reflect intertwined economic, demographic, and household-related features. This overlapping structure helps explain the weak cluster separation observed in Table 4. Methodologically, the loading plot guards against overinterpretation of principal components as singular, homogeneous constructs. Instead, it emphasizes that components are composite axes built from multiple interrelated variables. This reinforces the conceptual validity of the PCA solution while simultaneously highlighting the complexity of social stratification.

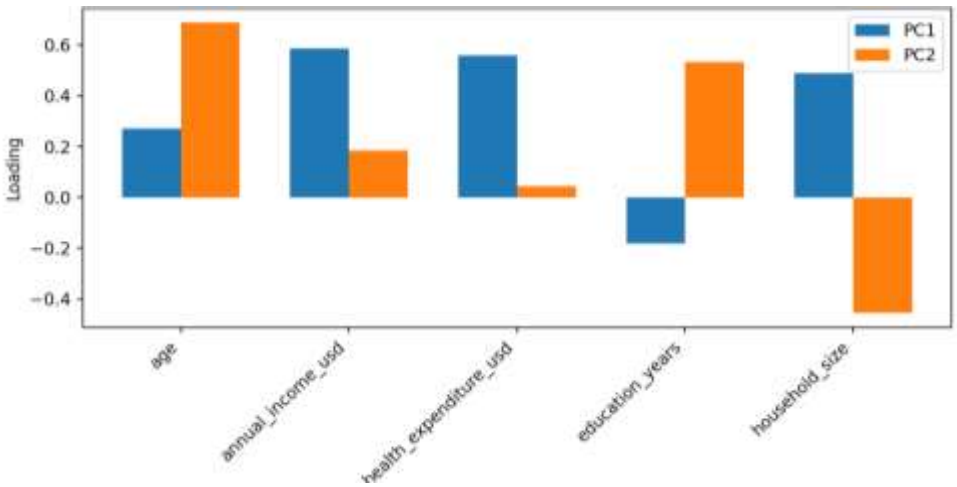

Figure 5: Loading plot for top contributing variables to PC1 and PC2

Table 5 reports the distribution of observations across the three identified clusters, with sizes of 629 (C0), 742 (C1), and 629 (C2), respectively. This relatively balanced distribution is methodologically important because it indicates that the clustering solution is not dominated by a single, disproportionately large group. In applied multivariate research, highly imbalanced clusters often reflect algorithmic artifacts or outlier-driven segmentation rather than meaningful structural differentiation. The near symmetry observed here suggests that the algorithm is capturing broad, recurring patterns in the data rather than isolating marginal subpopulations. Substantively, the similar sizes imply that the underlying socioeconomic landscape is not organized around a single normative profile with minor deviations. Instead, multiple configurations of socioeconomic characteristics coexist at comparable prevalence levels. This reinforces the multidimensional and pluralistic nature of social stratification, where no single lifestyle, income pattern, or household structure is universally dominant. At the same time, balanced cluster sizes should not be mistaken for strong categorical separation. As indicated by the low silhouette score (Table 4), these clusters overlap substantially, meaning that they represent probabilistic groupings rather than sharply bounded social classes. Thus, Table 5 supports an interpretation of clusters as typical profiles rather than discrete categories.

Table 5: Cluster Membership Distribution

Cluster	Size
C0	629
C1	742
C2	629

Table 5 reports the distribution of observations across the three identified clusters, with sizes of 629 (C0), 742 (C1), and 629 (C2), respectively. This relatively balanced distribution is methodologically important because it indicates that the clustering solution is not dominated by a single, disproportionately large group. In applied multivariate research, highly imbalanced clusters often reflect algorithmic artifacts or outlier-driven segmentation rather than meaningful structural differentiation. The near symmetry observed here suggests that the algorithm is capturing broad, recurring patterns in the data rather than isolating marginal subpopulations. Substantively, the similar sizes imply that the underlying socioeconomic landscape is not organized around a single normative profile with minor deviations. Instead, multiple configurations of socioeconomic characteristics coexist at comparable prevalence levels. This reinforces the multidimensional and pluralistic nature of social stratification, where no single lifestyle, income pattern, or household structure is universally dominant. At the same time, balanced cluster sizes should not be mistaken for strong categorical separation. As indicated by the low silhouette score (Table 4), these clusters overlap substantially, meaning that they represent probabilistic groupings rather than sharply bounded social classes. Thus, Table 5 supports an interpretation of clusters as typical profiles rather than discrete categories. In analytical terms, this table provides essential context for interpreting cluster-specific summaries (Table 6) and visualization outputs (Figures 4, 6, and 7). It ensures that subsequent inferences are not driven by extreme imbalance or marginal cases, thereby strengthening the credibility and generalizability of the segmentation results.

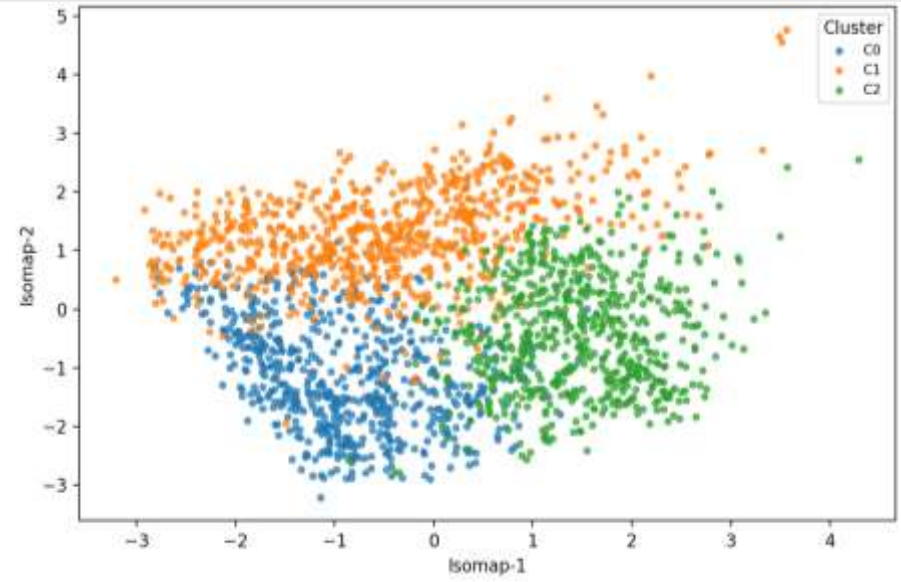

Figure 6: Isomap 2D embedding of PCA-reduced numeric data (exploratory)

Table 6 reports the mean values of the original socioeconomic variables for each of the three identified clusters, thereby translating abstract multivariate segmentation into substantively interpretable social profiles. Unlike component scores or reduced-space coordinates, this table anchors the clustering results in real-world units, making it possible to articulate how clusters differ in concrete terms such as age, income, household size, and health expenditure. Cluster C1 emerges as a relatively older and more economically advantaged group, with a mean age of approximately 65 years, the highest average income (31,758 USD), and the largest health expenditures (4,174 USD). This profile suggests a late-career or retirement-phase segment characterized by accumulated economic resources and greater healthcare utilization, consistent with life-course theories of consumption and health needs. The relatively moderate household size further supports the interpretation of this group as smaller, possibly post-childrearing households. In contrast, Cluster C0 is younger on average (39.6 years), has the smallest household size (2.15), and displays lower income and health spending. This profile is indicative of early-career or transitional life stages, where economic capacity is still developing and household responsibilities remain limited. The lower health expenditure may reflect both age-related differences in medical needs and resource constraints. Cluster C2 presents a distinct configuration: it is the youngest group (36.4 years), has the largest household size (6.63), and displays moderate income but relatively elevated health spending. This suggests a family-intensive profile, where economic resources are distributed across more dependents, potentially increasing healthcare-related expenditures. This cluster likely represents multigenerational or high-dependency households, in which resource allocation patterns differ substantially from smaller household units. Crucially, these profiles reveal that socioeconomic stratification does not operate along a single hierarchy. Instead, different forms of advantage and constraint coexist. Higher income does not necessarily coincide with larger households, and younger age does not uniformly imply lower health expenditure. This reinforces the multidimensional interpretation advanced throughout the analysis. Table 6 thus provides the most substantively meaningful output of the entire study, converting statistical segmentation into socially interpretable patterns.

Table 6: Cluster Profiles (Mean Values in Original Units)

Index	age	education_years	household_size	annual_income_usd	health_expenditure_usd
C0	39.6399	11.4892	2.1497	24506.1464	2834.7375
C1	65.0761	12.4417	4.5097	31758.2953	4173.9159
C2	36.3851	12.0153	6.6324	24496.8928	3498.2024

Figure 7 presents a star/glyph (radar) plot visualizing the normalized centroid profiles of the three clusters across the selected socioeconomic dimensions. Unlike tabular summaries, which require sequential comparison of values, this visualization enables simultaneous, holistic assessment of multidimensional profiles. Each axis represents a standardized variable, while the radial extent of each cluster's polygon reflects its relative standing on that dimension. This geometric encoding allows for rapid identification of contrasts, overlaps, and trade-offs between clusters. The most striking feature of the plot is the distinct geometric shapes formed by each cluster, indicating that segmentation is driven not by a single dominant variable but by composite patterns across multiple dimensions. For instance, one cluster exhibits pronounced extension along the income and health expenditure axes, while remaining comparatively constrained on household size. This confirms that material advantage in this group is associated with consumption capacity rather than household expansion. Another cluster shows a contrasting pattern, with strong extension along the household size dimension but more moderate positioning on income, suggesting resource diffusion across larger family units rather than concentrated accumulation. The visual symmetry and asymmetry of the polygons further emphasize the multidimensional nature of socioeconomic positioning. None of the clusters dominates across all axes simultaneously, reinforcing the idea that social advantage is not absolute but context-dependent. This finding is crucial: it undermines unidimensional interpretations of status and instead supports a relational understanding in which individuals may be advantaged in one domain while constrained in another. Methodologically, the star/glyph plot serves as a powerful complement to Table 6. While the table provides precise numerical comparisons, the radar plot reveals pattern coherence and profile distinctiveness. It also visually communicates the fuzzy boundaries between clusters, as the polygons overlap on several axes. This overlap corresponds to the low silhouette score reported earlier (Table 4), reinforcing the conclusion that clusters represent tendencies rather than discrete social types.

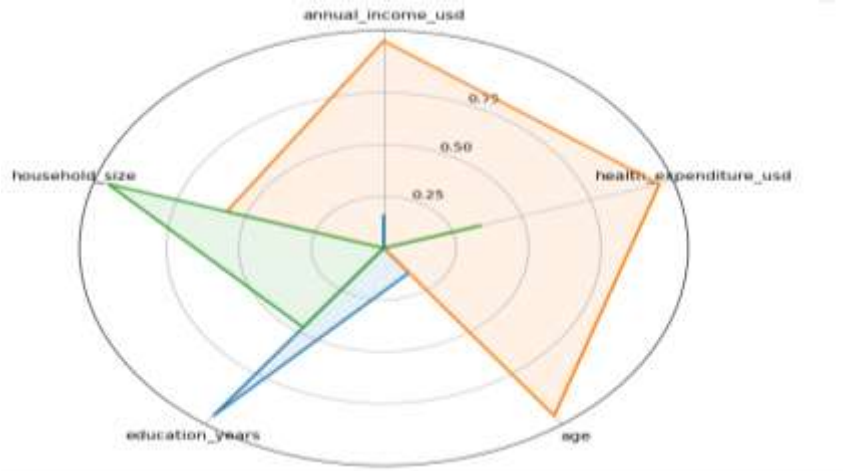

Figure 7: Star/glyph (radar) plot of cluster centroid profiles (min-max scaled)

Conclusion

This study set out to move beyond unidimensional representations of socioeconomic inequality by adopting a multivariate, structure-sensitive analytical framework. Using a combination of principal component analysis, unsupervised clustering, nonlinear embedding, and profile-based visualization, the analysis demonstrated that socioeconomic differentiation is not organized along a single linear hierarchy but rather unfolds across multiple, partially independent dimensions. These dimensions capture interrelated aspects of material resources, demographic life-course positioning, household structure, and health-related expenditure, confirming that inequality is fundamentally multidimensional in nature. The PCA results revealed that variance is distributed relatively evenly across several components, undermining simplified interpretations of socioeconomic status as a singular latent trait. This finding aligns with theoretical perspectives that conceptualize social stratification as a composite of intersecting forms of advantage and constraint rather than a monotonic scale. The clustering results further reinforced this interpretation. Although segmentation produced interpretable profiles, the low silhouette score and the overlapping structure of clusters indicated that socioeconomic positioning is better understood as a continuum rather than a set of discrete categories. This insight is critical, as it cautions against essentialist labeling and emphasizes the fluidity of real-world social configurations. The Isomap embedding provided an additional layer of validation by showing that the detected patterns persist even under nonlinear projection. This suggests that the observed structures are not merely artifacts of linear transformation but reflect deeper relational geometry within the data. The star/glyph visualization then translated these abstract patterns into substantively interpretable profiles, enabling holistic comparison across multiple dimensions simultaneously. Taken together, these findings underscore the importance of methodological pluralism in social data analysis. No single technique is sufficient to capture the full complexity of socioeconomic stratification. Instead, triangulated approaches that integrate dimensionality reduction, segmentation, and nonlinear visualization provide a more faithful representation of social reality. This study contributes both substantively and methodologically. Substantively, it demonstrates that social advantage is multidimensional, relational, and context-dependent. Methodologically, it illustrates how integrated multivariate pipelines can enhance interpretability without sacrificing structural fidelity. Future research should extend this framework to longitudinal data, incorporate institutional and spatial variables, and explore causal mechanisms. By reframing inequality as a multidimensional field rather than a single ranking, this study offers a more nuanced, empirically grounded understanding of contemporary socioeconomic differentiation.

REFERENCES

Sen, A. (1999). *Development as Freedom*. Oxford University Press.

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). Greenwood.

Grusky, D. B., & Sørensen, J. B. (1998). Can class analysis be salvaged? *American Journal of Sociology, 103*(5), 1187–1234.

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet, 365*(9464), 1099–1104.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality. *Annual Review of Sociology, 32*, 271–297.

Savage, M., et al. (2013). A new model of social class? *Sociology, 47*(2), 219–250.

Atkinson, A. B. (2015). *Inequality: What Can Be Done?* Harvard University Press.

Wilkinson, R., & Pickett, K. (2009). *The Spirit Level*. Bloomsbury.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data. *Demography, 38*(1), 115–132.

Kolenikov, S., & Angeles, G. (2009). Socioeconomic status measurement with PCA. *Health Services and Outcomes Research Methodology, 9*(3), 153–175.

Vyas, S., & Kumaranayake, L. (2006). Constructing SES indices: PCA and alternatives. *Social Science & Medicine, 62*(2), 459–468.

Howe, L. D., et al. (2012). Issues in the construction of wealth indices. *Emerging Themes in Epidemiology, 9*(3).

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies, 14*(2), 1518-1527.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley.

Hair, J. F., et al. (2019). *Multivariate Data Analysis* (8th ed.). Cengage.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson.

Everitt, B. S., et al. (2011). *Cluster Analysis* (5th ed.). Wiley.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters, 31*(8), 651–666.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data*. Wiley.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters. *Psychometrika, 50*(2), 159–179.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. *International Journal of Social Sciences Bulletin*, *3*(8), 43-50.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation. *arXiv preprint arXiv:1802.03426*.

KHAN, R., SHAH, A. M., & KHAN, H. U. (2025). Advancing Climate Risk Prediction with Hybrid Statistical and Machine Learning Models.

Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). Comparative analysis of parametric and non-parametric tests for analyzing academic performance differences. *Policy Research Journal*, *3*(8), 55-62.

Cox, T. F., & Cox, M. A. A. (2000). *Multidimensional Scaling*. Chapman & Hall.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.

Friendly, M. (2002). Corrgrams: Exploratory displays. *The American Statistician, 56*(4), 316–324.

Wilkinson, L. (2005). *The Grammar of Graphics* (2nd ed.). Springer.