Detecting Socioeconomic Structures and Well-Being Dynamics: An Exploratory Data Analysis within the Digital Humanities Framework

by Constantinos Panayi
Postgraduate Student, MSc Digital Humanities
National and Kapodistrian University of Athens, University of Cyprus, ATHENA Research Centre

***Abstract*—This study presents a structured Exploratory Data Analysis (EDA) and multivariate modelling workflow applied to a socioeconomic and lifestyle dataset comprising 30 observations and 10 numerical variables. Within a Digital Humanities (DH) framework, the analysis integrates descriptive statistics, outlier treatment, correlation analysis, linear regression, Principal Component Analysis (PCA), and K-means clustering in order to uncover latent structural patterns. Results reveal a dominant socioeconomic gradient, a distinct well-being dimension inversely related to age, and strong predictive relationships between income and expenditure as well as age and health status. Dimensionality reduction preserves over 90% of total variance, while clustering confirms income-based stratification and highlights internal heterogeneity within the middle-income group. The study demonstrates how combined statistical and unsupervised learning techniques can illuminate complex behavioural and economic structures even within small-scale datasets. The present project was conducted as the Second Assignment for the *Data Analytics* course of the MSc Digital Humanities programme, jointly offered by the *National and Kapodistrian University of Athens*, the *University of Cyprus*, and the *ATHENA Research Centre*.**

***Index Terms*—Socioeconomic Patterns, Well-Being Analysis, Exploratory Data Analysis, Digital Humanities, Supervised and Unsupervised Learning, Linear Regression, Principal Component Analysis, K-means Clustering**

I. Introduction

Quantitative data analysis enables the systematic investigation of socioeconomic and behavioural structures through integrated descriptive and multivariate techniques. Rather than examining variables in isolation, the present study develops a coherent analytical pipeline that progresses from univariate exploration to regression modelling and unsupervised learning. The objective is to identify dominant structural dimensions, quantify predictive relationships, and detect latent group formations embedded within the dataset.

A. Aims

The study is guided by the following objectives:

- To examine the distributional properties and dispersion patterns of key variables.
- To detect and manage statistical outliers in a methodologically robust manner.
- To quantify linear associations through correlation and regression modelling.
- To reduce dimensional complexity via Principal Component Analysis (PCA).
- To identify latent clusters using K-means clustering and assess their socioeconomic interpretation.

II. Data Management

A. Dataset Overview

The dataset comprises 30 individual-level observations and 10 numerical variables capturing complementary demographic, socioeconomic, and well-being dimensions, as summarised in Table I. Data were provided by the instructors of the *Data Analytics* course of the MSc Digital Humanities programme.

More specifically, the dataset includes:

- **Demographic variable**: age, representing the biological life stage of each individual.
- **Human capital variable**: education_years, reflecting accumulated formal educational attainment.
- **Economic activity variables**: income, weekly_work_hours, monthly_expenditure, and savings, jointly describing labour participation, earning capacity, consumption behaviour, and financial accumulation.
- **Well-being variables**: health_index and satisfaction_score, capturing objective physical condition and subjective life evaluation respectively.
- **Mobility variable**: commute_time_minutes, indicating daily travel burden and potential lifestyle constraints.
- **Administrative variable**: id, serving exclusively as a unique identifier and excluded from statistical modelling.

The coexistence of economic, demographic, behavioural, and subjective indicators enables a multidimensional analytical perspective. In contrast to datasets focused exclusively on financial metrics, the present structure allows for the examination of how material conditions intersect with well-being outcomes and life-course characteristics.

TABLE I: Dataset Variables and Descriptions

Variable Name	Description
id	Unique numerical identifier for each observation.
age	Age of the individual (years).
income	Annual income (monetary units).
education_years	Total years of formal education completed.
weekly_work_hours	Average number of hours worked per week.
health_index	Composite physical health indicator (0–1 scale).
satisfaction_score	Subjective life satisfaction score (1–10 scale).
monthly_expenditure	Average monthly consumption expenditure.
savings	Accumulated financial savings.
commute_time_minutes	Average daily commute time (minutes).

All variables are continuous (integer or float), ensuring compatibility with parametric statistical methods, regression modelling, and dimensionality reduction techniques. The balanced inclusion of structural (income, work hours), behavioural (expenditure, savings), and evaluative (health, satisfaction) dimensions renders the dataset particularly suitable for exploring latent socioeconomic gradients and identifying potential clustering patterns across individuals.

B. Data Quality and Pre-Processing

Before proceeding to modelling and multivariate exploration, an initial diagnostic inspection of the dataset is conducted in order to establish a structured understanding of its internal coherence, statistical properties, and potential anomalies. As outlined above, the dataset contains 30 observations and 10 numerical variables, all stored in consistent integer or float format. Moreover, no missing values or duplicate records are detected across any of the variables, thereby ensuring data completeness, consistency, and reliability for subsequent analyses.

As a first step, descriptive statistics are computed for all numerical variables. These measures provide an initial overview of central tendency and dispersion, allowing a preliminary identification of potential irregularities.

The results indicate substantial variability across socioeconomic and lifestyle features. Age ranges from early adulthood to late middle age, suggesting a heterogeneous life-course structure. Income and savings exhibit wide dispersion, reflecting differentiated financial positions. Education years display a notably broad range, signalling the possible presence of extreme values. Weekly work hours cluster around full-time employment levels, while health index and satisfaction score show relatively limited dispersion, suggesting moderate homogeneity in well-being indicators. Monthly expenditure and commute time present moderate variability, indicating differentiated consumption and mobility patterns.

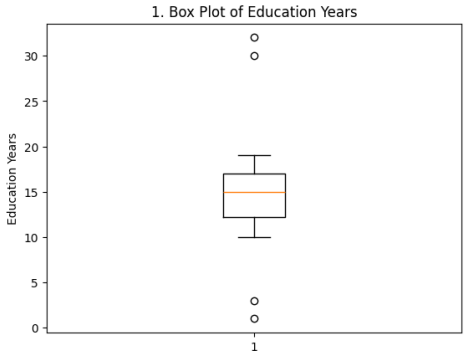

Fig. 1: Box plot of education_years

Fig. 2: Box plot of income

To enhance interpretability and visually assess dispersion, three key variables are represented through box plots; *education_years*, *income*, and *weekly_work_hours*, as illustrated in Figures 1, 2, and 3, respectively. Among the available techniques, the Interquartile Range (IQR) method is selected due to its robustness, as it relies on quartiles and is therefore less sensitive to extreme observations, making it particularly suitable for skewed distributions.

The box plot for *education_years* (Figure 1) reveals a relatively concentrated distribution around typical schooling lengths, with the IQR spanning a moderate central range and a median close to the sample's central educational attainment. Nevertheless, several extreme observations are clearly visible at both tails of the distribution. At the lower end, very small values deviate sharply from the main cluster, while exceptionally high values represent substantial departures at the upper tail. Despite the overall limited dispersion, these extreme observations introduce marked heterogeneity. Given the relatively small sample size, such deviations must be interpreted cautiously, as trends may appear more exaggerated during EDA.

In contrast, the box plot for *income* (Figure 2) displays substantial variability but no statistically significant outliers. Most observations fall within a well-defined IQR, and the distribution appears balanced, indicating stable

Fig. 3: Box plot of weekly_work_hours

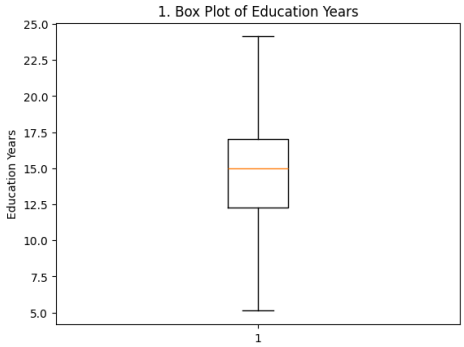

Fig. 4: Box plot of education_years after IQR capping

socioeconomic differentiation without extreme anomalies.

Similarly, the box plot for *weekly_work_hours* (Figure 3) shows moderate spread around standard full-time employment levels. The distribution is relatively symmetric and reveals no extreme outliers, though slight asymmetry suggests variation in workload intensity across individuals.

C. Data Cleansing

Following visual inspection, the Interquartile Range (IQR) method is formally applied to all numerical variables (excluding the identifier *id*) in order to identify statistical outliers. According to this method, observations falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are classified as outliers.

The analysis confirms the presence of four outlier observations, all belonging exclusively to the variable *education_years*. These correspond to extreme low and high values that clearly fall outside the acceptable statistical boundaries, as already observed in the initial box plot.

In addressing these outliers, multiple methodological options are considered. Removing the affected observations would result in a loss of approximately 13% of the dataset, which is statistically significant given the small sample size. Alternatively, a logarithmic transformation could reduce skewness; however, this would alter interpretability and is inappropriate for a discrete and substantively meaningful variable such as years of education.

Consequently, a winsorisation (capping) strategy is adopted as a balanced compromise. The extreme values are replaced with the corresponding lower and upper IQR bounds, thereby stabilising dispersion while preserving all observations and maintaining interpretative clarity. The updated distribution of *education_years* after capping is presented in Figure 4.

As shown in Figure 4, the distribution now appears considerably more regular and symmetric. The IQR captures the central 50% of observations within a compact and coherent range, while no extreme values extend beyond the whiskers. The median remains stable, confirming that the central tendency of the variable has not been distorted.

It is important to acknowledge the methodological trade-off inherent in winsorisation. While the procedure enhances statistical stability and prevents extreme values from disproportionately influencing regression and multivariate analysis, it simultaneously reduces visibility of genuinely exceptional cases. This balance is particularly crucial in small samples, where each observation exerts comparatively greater influence on overall results.

Overall, the dataset following preprocessing exhibits improved statistical stability, preserved completeness, and enhanced suitability for subsequent correlation, regression, and clustering procedures.

III. Methodology

A. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) constitutes the foundational methodological framework of the study and is employed to systematically examine the structure, distribution, and variability of the dataset prior to any modelling procedures. EDA enables the identification of dominant patterns, dispersion properties, and potential relationships among variables through the combined use of descriptive statistics and visual analytics.

Visualisation techniques are selected according to variable type and analytical purpose:

- **Numerical distributions**: Histograms and box plots are used to assess central tendency, dispersion, skewness, and the presence of outliers in variables such as income, savings, age, and weekly work hours.
- **Bivariate relationships**: Scatter plots are employed to visually inspect potential linear associations between selected predictor–target pairs, including income and expenditure, age and health index, and work hours and savings.
- **Correlation structure**: A Pearson correlation matrix and heatmap are used to examine the strength and direction of linear relationships across all numerical variables.

Through this process, EDA provides an informed basis for subsequent regression modelling and multivariate analysis.

B. Regression Analysis and Modelling

To quantify the extent to which specific socioeconomic variables explain variation in selected outcomes, both simple and multiple linear regression models are constructed. The modelling approach is exploratory in nature, focusing on interpretative insight rather than predictive optimisation.

1) Simple Linear Regression: Three simple regression models are estimated:

1) **Income as predictor of monthly expenditure**, examining whether earning capacity explains consumption behaviour.
2) **Age as predictor of health index**, assessing the relationship between life-course progression and physical well-being.
3) **Weekly work hours as predictor of savings**, investigating whether labour intensity contributes to financial accumulation.

Model performance is evaluated using coefficient estimates, R^2 values, and visual inspection of regression lines and residual dispersion.

2) Multiple Linear Regression: Multiple regression models are employed to estimate the combined contribution of several predictors on selected target variables. This approach allows the estimation of each predictor's unique effect while controlling for the influence of others.

Three configurations are examined:

- **Satisfaction score as target**, predicted by income, savings, expenditure, weekly work hours, commute time, education years, and health index.
- **Income as target**, predicted by savings, monthly expenditure, and weekly work hours.
- **Age as target**, predicted by health index and satisfaction score.

Goodness-of-fit measures and coefficient interpretation are used to assess explanatory strength and relative predictor importance.

C. Dimensionality Reduction: Principal Component Analysis (PCA)

Given the multidimensional structure of the dataset, Principal Component Analysis (PCA) is applied to reduce dimensional complexity while preserving as much variance as possible. Prior to PCA, all numerical variables (excluding the identifier *id*) are standardised to ensure scale comparability. Three principal components are extracted, capturing the dominant latent dimensions of the dataset. The PCA projection is examined through:

- 2D visualisation (PC1–PC2 plane) to identify broad structural gradients.
- 3D visualisation (PC1–PC2–PC3) to explore deeper latent separations and potential outliers.

This procedure enables the identification of underlying socioeconomic gradients and behavioural dimensions that may not be visible when examining variables individually.

D. Unsupervised Learning: K-Means Clustering

To detect natural groupings within the dataset without predefined labels, K-means clustering is applied to the standardised numerical variables. The number of clusters is set to $k = 4$, balancing interpretability and structural granularity. The clustering process iteratively assigns observations to centroids in order to minimise within-cluster variance.

Cluster interpretation is conducted within the PCA space, allowing visual examination of how groups align with latent socioeconomic dimensions. This approach facilitates the detection of both dominant stratification patterns and internally heterogeneous subgroups. Overall, the methodological framework integrates EDA, regression, dimensionality reduction, and clustering techniques in order to provide a coherent and multi-layered understanding of the dataset's latent structure.

IV. Findings and Analysis

A. Univariate Analysis

Following data cleaning and preprocessing, the first stage of the substantive analysis focuses on the independent examination of each numerical variable. Univariate analysis enables the identification of distributional shape, central tendency, dispersion, and potential asymmetries prior to the exploration of inter-variable relationships. This step is essential for establishing a statistically informed baseline and for interpreting subsequent regression and multivariate findings within the broader context of the dataset's internal variability.

In this stage, measures of Central Tendency (mean, median, and mode) and Dispersion (range, variance, standard deviation, and IQR) are computed for each numerical variable. These statistical measures are then visualised through histograms, which provide a clearer understanding of distributional shape, including symmetry, skewness, and variability. The combined histogram representation is presented in Figure 5.

Based on the histograms and the corresponding summary statistics, the following observations can be made:

1) **age**: The age distribution exhibits slight right skewness, with a greater concentration of individuals in younger age groups. The mean and median are relatively close, indicating mild asymmetry. Dispersion is moderate, and the IQR shows that most values fall within a compact mid-range. Overall, age does not follow a normal distribution but remains fairly balanced.
2) **income**: Income displays significant right skewness, characterised by a concentration of middle-income individuals and some substantially higher values that elevate the mean. The large variance and standard deviation reflect important socioeconomic heterogeneity within the sample.

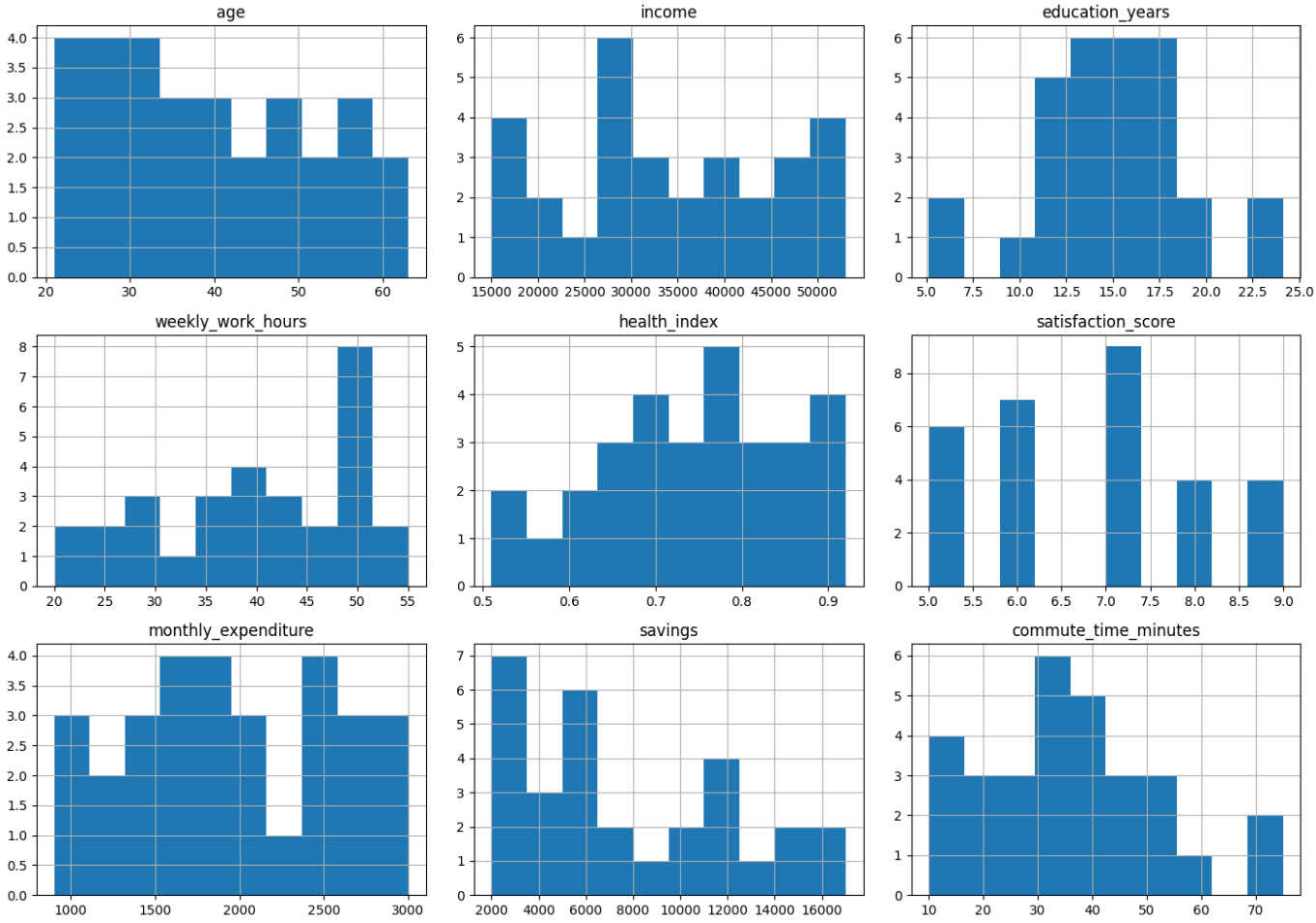

Fig. 5: Histograms of numerical variables (Univariate Distributions)

3) **education_years**: After IQR-based capping, the distribution becomes more regular and symmetric. The mean and median are nearly identical, while the IQR indicates a narrow central range, suggesting relatively homogeneous educational attainment across the dataset.
4) **weekly_work_hours**: Working hours cluster around standard full-time levels. The distribution shows slight left skewness due to individuals working longer hours. Variability is moderate and the IQR captures a compact central range.
5) **health_index**: The health index displays limited dispersion with slight left skewness, as many values fall in the higher range. Mean and median are close, indicating relative stability in health outcomes.
6) **satisfaction_score**: Satisfaction scores are concentrated within a relatively narrow range, mostly between moderate and high values. Variability is low, suggesting generally positive subjective well-being within the sample.
7) **monthly_expenditure**: Expenditure exhibits moderate right skewness, with most values clustered in the middle range. Dispersion remains contained, indicating differentiated but not extreme spending behaviour.
8) **savings**: Savings show strong right skewness, with many low-savings observations and a small number of substantially higher values. This results in high variance and a heavy-tailed distribution.
9) **commute_time_minutes**: Commute time is moderately right-skewed, with most individuals commuting within a typical mid-range interval. A small subset of longer commute times increases dispersion.

Overall, the univariate analysis reveals a dataset characterised by moderate dispersion across most variables, pronounced right skewness in financial indicators (income and savings), and relatively compact distributions in well-being measures. These findings establish the distributional foundation necessary for interpreting subsequent correlation, regression, and clustering analyses, particularly with respect to socioeconomic heterogeneity and life-course effects.

B. Bivariate Analysis

Following the univariate exploration of individual variables, the analysis proceeds to the examination of pairwise relationships. Bivariate analysis aims to identify the strength, direction, and structural significance of linear associations between numerical variables, thereby providing a bridge between descriptive statistics and multivariate modelling.

1) Correlation Analysis: Pearson correlation coefficients (r) are computed to quantify the degree of linear association between all pairs of numerical variables. The resulting correlation matrix is visualised through a heatmap, presented in Figure 6, where deeper red tones indicate stronger positive correlations and deeper blue tones denote stronger negative relationships.

The heatmap reveals two dominant structural clusters within the dataset:

1) **Economic Cluster**: income, monthly_expenditure, savings, and weekly_work_hours exhibit strong positive intercorrelations, indicating coherent financial behaviour patterns. Higher income is associated with greater expenditure and savings, while increased work hours tend to align with stronger economic activity.
2) **Well-being Cluster**: health_index and satisfaction_score show a strong positive association, suggesting that better health correlates with higher subjective well-being.

Age emerges as a central structural variable, displaying positive associations with income and savings, and strong negative correlations with health_index and satisfaction_score. This pattern reflects the dual role of age as both an economic progression factor and a potential driver of declining well-being.

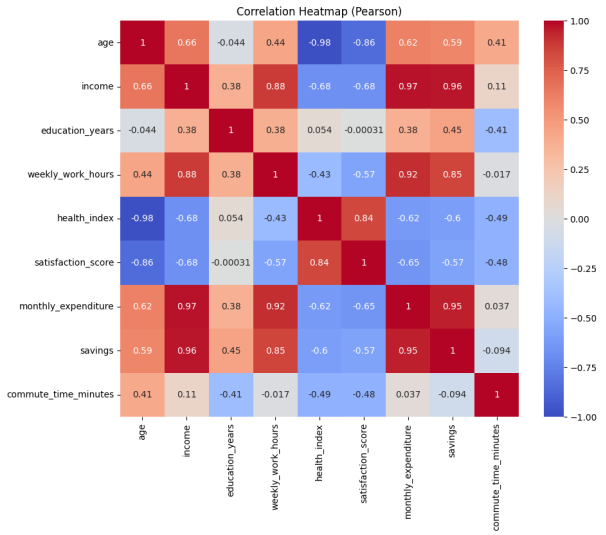

Fig. 6: Pearson Correlation Heatmap

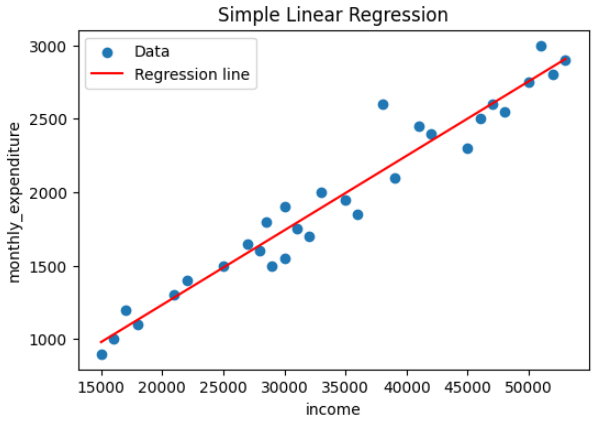

Fig. 7: Simple Linear Regression: Income → Monthly Expenditure

Education_years and commute_time_minutes appear comparatively less integrated within the broader correlation structure, exhibiting weaker and more selective associations.

Overall, the correlation analysis uncovers a clear structural pattern characterised by an economic gradient and an inverse well-being dimension structured around age.

2) Simple Linear Regression: To further quantify key bivariate relationships, simple linear regression models are estimated for selected predictor–target pairs. Each model captures the direction and strength of the relationship through a fitted regression line and corresponding R^2 value.

a) Income as predictor of monthly expenditure: The regression model between income and monthly expenditure, illustrated in Figure 7, reveals a strong positive linear relationship. The upward-sloping regression line closely aligns with the observed data points, indicating that higher income levels systematically correspond to higher consumption expenditure. The high R^2 value confirms strong predictive capacity.

b) Age as predictor of health index: The regression between age and health_index (Figure 8) demonstrates a strong negative linear relationship. The downward-sloping regression line indicates that health status declines steadily with increasing age. The concentration of points around the line suggests limited unexplained variability and strong explanatory power.

c) Weekly work hours as predictor of savings: The regression model between weekly_work_hours and savings (Figure 9) shows a positive linear trend, indicating that greater labour intensity is generally associated with increased financial accumulation. Although dispersion around the regression line is higher compared to the previous models, the overall association remains statistically meaningful.

In summary, the bivariate regression results confirm the structural patterns identified in the correlation matrix. In-

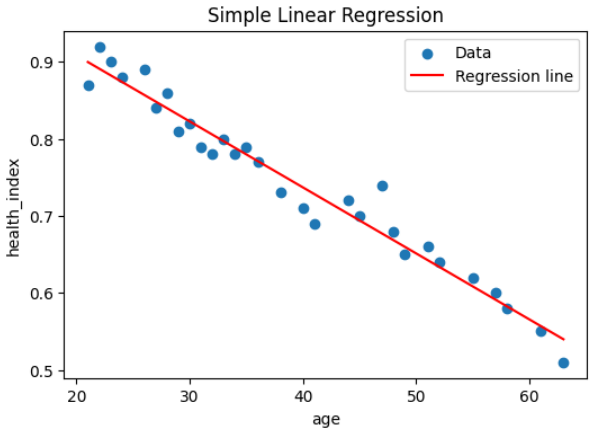

Fig. 8: Simple Linear Regression: Age → Health Index

Fig. 9: Simple Linear Regression: Weekly Work Hours → Savings

come strongly predicts expenditure, age robustly predicts health decline, and work intensity moderately predicts savings behaviour. These findings provide a quantitative foundation for the subsequent multivariate and dimensionality reduction analyses.

C. Multivariate Analysis

Building upon the bivariate findings, the analysis proceeds to a multivariate framework in order to examine how multiple variables interact simultaneously and to uncover latent structural dimensions within the dataset. Multivariate analysis enables a deeper and more integrated understanding of socioeconomic and well-being patterns, moving beyond pairwise associations toward system-level structure.

1) Multiple Linear Regression: Multiple linear regression models are estimated to evaluate the joint contribution of several predictors on selected target variables. This approach allows for the estimation of each predictor's unique effect while controlling for the influence of the others.

Three configurations are examined:

a) Satisfaction Score as Target: The first model predicts *satisfaction_score* using income, savings, monthly_expenditure, weekly_work_hours, commute_time_minutes, education_years, and health_index. Among all predictors, *health_index* emerges as the dominant positive determinant of life satisfaction, while *weekly_work_hours* exerts a moderate negative effect. Economic variables show comparatively limited independent contribution once health and workload are controlled for, suggesting that subjective well-being is structured primarily around physical condition and labour intensity rather than purely financial metrics.

b) Income as Target: The second model predicts *income* using savings, monthly_expenditure, and weekly_work_hours. Monthly expenditure and savings display strong positive coefficients, confirming the coherence of the economic cluster identified earlier. Weekly work hours show a negative coefficient, potentially reflecting differences in hourly wage structures or occupational composition.

c) Age as Target: The third model estimates *age* as a function of health_index and satisfaction_score. The results illustrate a strong negative association between age and both predictors, with health emerging as the most influential factor. This result reinforces the life-course gradient previously identified, linking increasing age with declining health and moderately lower life satisfaction.

Overall, the multiple regression models confirm the presence of two dominant structural dimensions: an internally coherent economic system and a well-being axis structured around age and health.

2) Principal Component Analysis (PCA): To further investigate latent structural dimensions, Principal Component Analysis (PCA) is applied to the standardised numerical variables (excluding the identifier *id*). PCA reduces dimensional complexity while preserving as much variance as possible.

Three principal components are extracted, jointly explaining over 90% of total variance.

a) PC1 – Socioeconomic Gradient: The first principal component captures the dominant economic dimension of the dataset. As illustrated in the 2D projection (Figure 10), observations are distributed along a clear horizontal gradient corresponding to income-related variables such as savings and expenditure.

b) PC2 and PC3 – Secondary Social Dimensions: The second and third components capture additional variability linked to education, health, and behavioural characteristics. The 3D projection (Figure 11) reveals clearer group separation and highlights potential outliers that are not fully visible in the 2D representation.

Overall, PCA confirms that the dataset is structured primarily around a dominant socioeconomic gradient, supplemented by secondary behavioural and well-being dimensions.

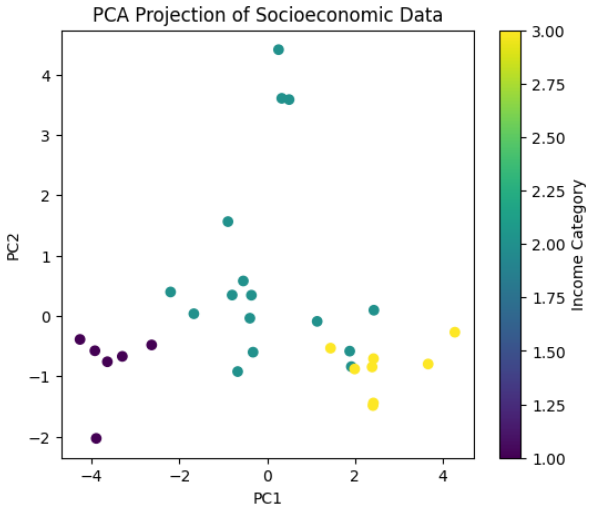

Fig. 10: PCA Projection (PC1–PC2 Plane)

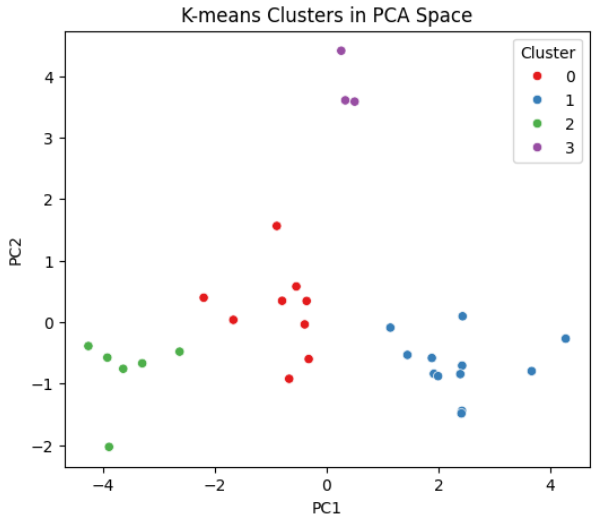

Fig. 12: K-Means Clustering in PCA Space (2D)

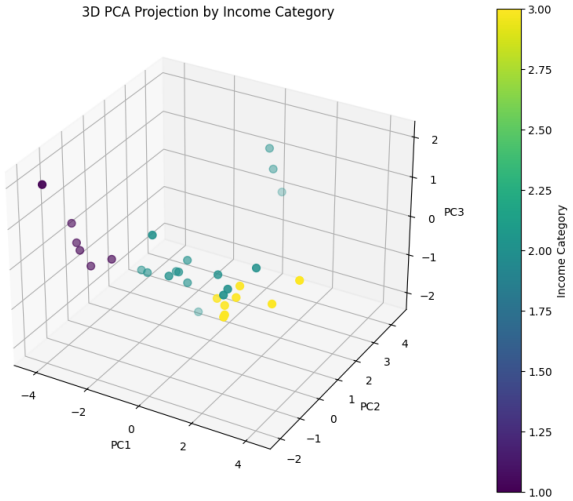

Fig. 11: 3D PCA Projection (PC1–PC2–PC3)

3) K-Means Clustering: To detect natural groupings without predefined labels, K-means clustering is applied to the standardised data with $k = 4$ clusters.

a) 2D Cluster Structure: The clustering structure projected onto the PCA plane (Figure 12) reveals strong alignment with the socioeconomic gradient identified in PC1. Low- and high-income observations form cohesive and internally homogeneous clusters, while middle-income cases display greater dispersion.

b) 3D Cluster Interpretation: The 3D PCA cluster visualisation provides a more refined representation of inter-cluster distances. It highlights the relative homogeneity of the extreme income groups and the internal heterogeneity of the middle-income category, which splits across multiple structural positions.

Overall, taken together, the multivariate analysis confirms that the dataset is organised around a dominant economic dimension with a secondary well-being axis structured around age and health. Multiple regression quantifies these relationships, PCA condenses them into latent components, and clustering reveals coherent socioeconomic groupings while exposing internal heterogeneity within the middle-income category.

V. Conclusion

The present study developed a coherent and multi-layered analytical framework in order to examine the structural properties of a multidimensional socioeconomic and lifestyle dataset. By integrating Exploratory Data Analysis, regression modelling, Principal Component Analysis, and K-means clustering, the analysis moved progressively from descriptive inspection to latent structural interpretation.

The findings reveal the presence of a dominant socioeconomic gradient organising the dataset, primarily structured around income, expenditure, savings, and work intensity. This economic dimension demonstrates strong internal coherence and high explanatory capacity in both bivariate and multivariate regression models. At the same time, a distinct well-being axis emerges, characterised by the strong inverse relationship between age and health status, and the central role of health in shaping life satisfaction.

Multiple linear regression confirms that financial indicators strongly predict income dynamics, whereas subjective well-being depends more substantially on health conditions and workload intensity than on purely monetary variables. Principal Component Analysis further condenses these relationships into a small number of latent components, preserving over 90% of total variance while clearly exposing the underlying socioeconomic gradient. K-means clustering reinforces this structure, revealing

cohesive low- and high-income clusters and highlighting notable heterogeneity within the middle-income group.

Importantly, the study demonstrates that even within a relatively small dataset, integrated statistical and unsupervised learning techniques can uncover meaningful structural patterns. The combination of descriptive statistics, modelling, dimensionality reduction, and clustering provides a comprehensive perspective on how demographic, economic, and well-being variables interact as a system rather than as isolated indicators.

While the limited sample size constrains generalisability, the analytical framework applied here illustrates the methodological value of structured, stepwise quantitative analysis. Future research could expand the dataset, incorporate longitudinal observations, or introduce additional behavioural and contextual variables in order to further refine the understanding of socioeconomic stratification and well-being dynamics.

Overall, the findings confirm the existence of a coherent socioeconomic structure shaped by financial capacity, labour engagement, and life-course effects, while simultaneously highlighting the central importance of health in determining subjective well-being outcomes.

Acknowledgements

The author would like to thank the instructors of the *Data Analytics* course for providing the dataset and methodological guidance. The author also gratefully acknowledges the *Union of Greek Shipowners* and the *Georgios and Andrianí Lordos Foundation* for the financial support provided through the award of scholarships that enabled the completion of his postgraduate studies.