A Comparative Study of Dimensionality Reduction on GDP Per Capita (1990–2023)

Presented to

Dr. Peter Gao
Department of Mathematics
San Jose State University

In Partial Fulfillment

Of the Requirements for the Class
Math 250

By

Olivia Hartnett
Priyanka Goel
Veda Sahithi Bandi

May 2025

Contents

1 Introduction

Understanding economic development across countries requires analyzing complex, high-dimensional data that spans multiple years and regions. Dimensionality reduction techniques offer a way to simplify such datasets by projecting them into lower dimensions, while preserving the essential structures and relationships. These methods are particularly useful for discovering hidden patterns, comparative analysis, and clear visualization.

In this study, we apply three dimensionality reduction methods—Principal Component Analysis (PCA), Factor Analysis, and t-distributed Stochastic Neighbor Embedding (t-SNE)—to explore global economic development trends. PCA and Factor Analysis are linear techniques that aim to summarize the data using components or latent factors, while t-SNE is a nonlinear approach designed for uncovering local structures in high-dimensional data. Each method offers a different lens to understand how countries' economic trajectories compare over time.

Our goal is to evaluate how different dimensionality reduction techniques group countries with similar GDP trends and to interpret the resulting low-dimensional representations. We aim to identify common patterns of economic growth, differences across regions, and whether nonlinear methods reveal structures that linear methods miss. By comparing the results across techniques, we highlight the trade-offs between interpretability and the ability to capture complex patterns in the data.

2 Methods

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that projects high-dimensional data onto a smaller set of orthogonal directions, called principal components, which capture the maximum variance in the data. These components are uncorrelated and ordered by how much variation they explain. PCA is useful for compressing data, reducing redundancy, and visualizing patterns in lower dimensions, especially when variables are correlated, allowing the main structure to be summarized by just a few components.

In practice, PCA is often computed using singular value decomposition (SVD). The right singular vectors define the new axes, and the projections onto these directions are called principal component scores. Each component's importance is determined by the square of the corresponding singular value. When variables have different units or variances, scaling the data is important to ensure that no variable dominates the analysis.

PCA assumes that the most meaningful structure in the data is captured by the directions with the most variance. It is best suited for data where important trends are linear and when random noise affects all directions equally. It also requires the data to be centered and, in most cases, scaled. While PCA is computationally efficient and reveals global patterns, it may miss nonlinear relationships and is sensitive to outliers. One common mistake is overinterpreting the components. Since they are chosen purely to maximize variance, they don't necessarily correspond to meaningful real-world concepts.

In this study, PCA helps us understand global economic trends by identifying common patterns in GDP per capita trajectories and projecting country-level data into a lower-dimensional space.

2.2 Factor Analysis

Factor Analysis (FA) is a statistical method used to model the correlations among the observed variables in terms of a smaller number of unobserved variables called latent factors. FA is often used when we believe that what we see in the data is a reflection of some deeper structure. Unlike PCA, which is a descriptive technique, factor analysis is a *data-generating model*. This means it assumes that the data we observe was produced by a specific probabilistic process involving the latent factors. Mathematically, we model each observed data point x as:

$$x = Wz + \mu + \Psi$$

where W is a matrix of factor loadings (how much each factor contributes to each variable), z is a vector of latent factors, μ is the mean of the data, and Ψ is a covariance matrix and assumed to be independent across variables and represent variations not explained by common factors. A key assumption of the model is that the factors z follow a standard normal distribution and may or may not be uncorrelated.

FA also assumes that each variable's total variance can be split into two parts: the part explained by the common factors called *communality* and the part that is unique to that variable called *uniqueness*. This is useful when we want to reduce redundancy and understand which variables are driven by the same underlying sources.

However, FA comes with its own challenges like overinterpretability. Because there are many valid ways to represent the same factor structure, the model is "unidentifiable" without additional constraints. Rotations like *varimax* are often applied to make the loadings more interpretable. FA also assumes the factors are linearly related to the observed variables.

In this project, we use factor analysis to see if a small number of common trends can explain the variation in GDP per capita across countries and years. This allows us to study economic development from the perspective of shared latent structures.

2.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique used primarily to visualize high-dimensional data. It works by preserving the local structure of the data: points that are close together in high dimensions remain close in the lower dimension. Unlike PCA and factor analysis, the axes in t-SNE do not have a direct interpretation. The goal is not to explain variance or correlation, but to explore patterns and uncover potential clusters.

The main assumption behind t-SNE is that the data lie on a lower-dimensional manifold and that local similarity patterns in high dimensions should be reflected in the reduced space. The algorithm models similarity between points as probabilities and tries to preserve those similarities by minimizing a divergence between high- and low-dimensional probability distributions.

t-SNE is powerful for revealing clusters and structure in complex, nonlinear datasets. It is especially useful when traditional linear techniques fail to separate meaningful patterns. However, t-SNE does not preserve global structure, is sensitive to hyperparameter choices, and can create misleading patterns if the assumptions are not met. It also performs best with large datasets and may give unstable results on smaller ones. It also has issues with overinterpretability, so results should be interpreted with caution.

In our project, we applied t-SNE to GDP per capita data to explore economic similarities among countries. Despite our dataset being smaller than typically recommended, the algorithm produced a meaningful embedding that revealed groupings aligned with economic regions.

3 Data

Gross Domestic Product (GDP) per capita is a key indicator of economic performance and living standards, reflecting the average income per person in a country. It enables comparisons of prosperity and development across nations. When examined over time, GDP per capita reveals not only a country's current wealth but also its economic trajectory, whether it is growing, stagnant, or in decline.

We use GDP per capita data obtained from *Our World in Data*, which compiles global development indicators from various sources, primarily the World Bank. The dataset covers nearly 200 countries and territories from 1990 to 2023, with GDP per capita expressed in 2021 international dollars—an adjusted unit that accounts for inflation and purchasing power parity to enable cross-country comparisons of income and living standards.

The original dataset is in long format, with each row representing a combination of country, year, and GDP per capita value. To prepare the data for analysis, we pivoted it to wide format so that each row corresponds to a country and each column to a specific year (from 1990 to 2023, the period we focus on in this study). This format allowed us to treat each country's GDP trajectory as a high-dimensional vector suitable for dimensionality reduction techniques. All further preprocessing and transformations are discussed in the Exploratory Data Analysis section.

3.1 Exploratory Data Analysis

We began by addressing missing values in the GDP per capita dataset. Thirteen countries had gaps in their GDP series between 1990 and 2023. To preserve the structure and continuity of each country's economic trajectory, we calculated the average annual rate of change in GDP per capita for each country and used this to forward-fill and backward-fill missing entries. This imputation method helped maintain realistic growth patterns without distorting trends.

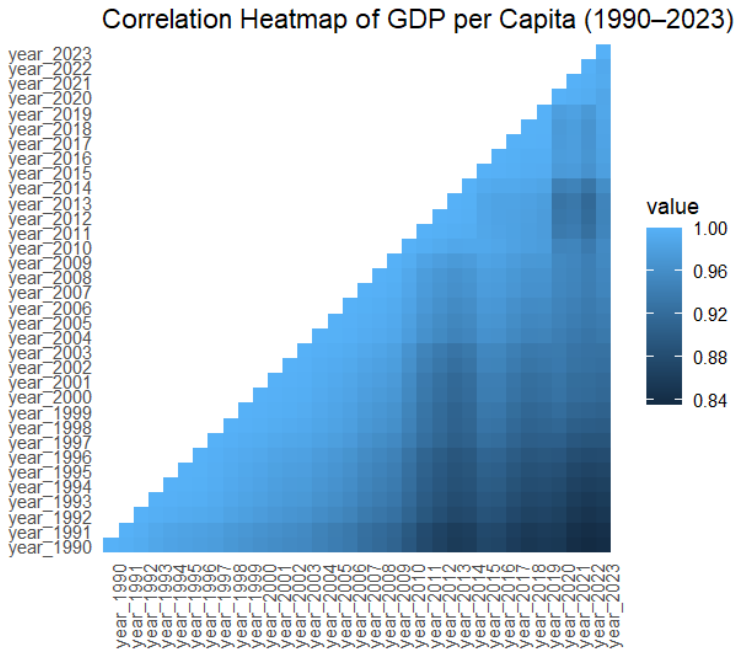

Figure 1: Correlation heatmap of GDP per capita (1990–2023).

After preprocessing, we examined the structure of the dataset. Figure 1 shows a correlation heatmap of GDP per capita values from 1990 to 2023. The heatmap highlights strong positive correlations across years, particularly in recent decades. This indicates that countries' GDP trajectories tend to evolve smoothly over time. Such temporal coherence supports the use of dimensionality reduction techniques, as it suggests that the data can be effectively summarized in a lower-dimensional space.

We then assessed the distribution of GDP per capita across countries. This confirmed that the GDP per capita data is heavily right-skewed, with a long tail driven by a small group of extremely wealthy countries, such as Luxembourg, Qatar, and Bermuda. This skewness can distort results in methods like PCA and FA, where high-variance features may dominate the first few components or factors, making them reflect scale rather than structure.

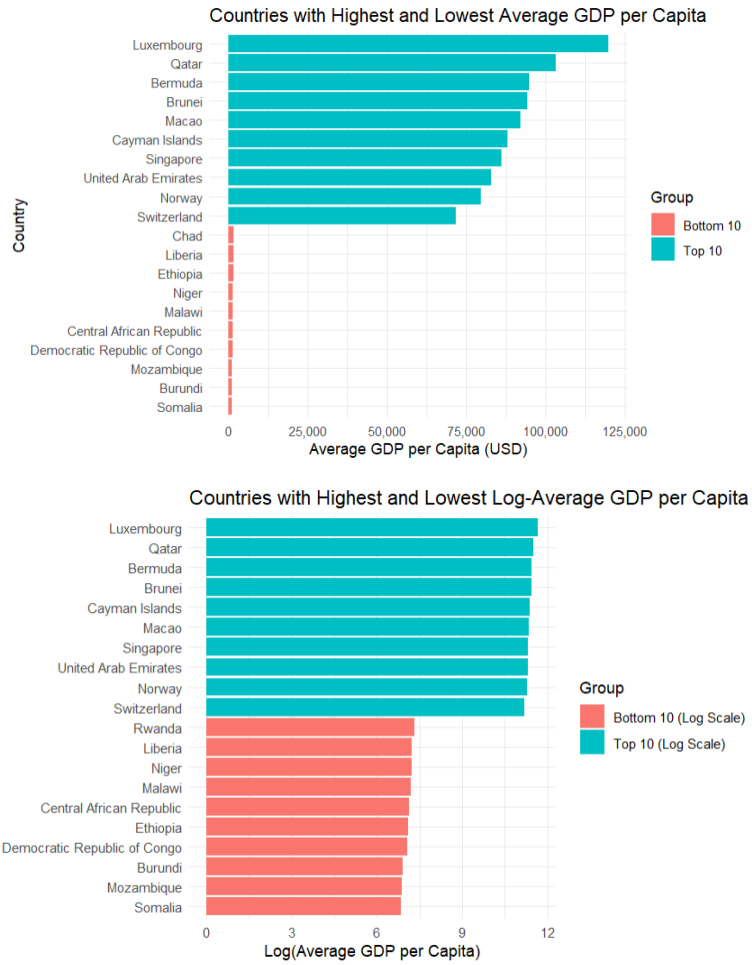

Figure 2: Top and bottom 10 countries by average GDP per capita (1990–2023)

To address this, we applied a logarithmic transformation to the GDP per capita values. As shown in Figure 2, the transformation compresses differences in magnitude while preserving rank order, reducing the influence of extreme outliers and stabilizing variance across countries. More importantly, the log transformation enhances visibility among mid- and low-income countries, whose variation was previously overshadowed. This adjustment makes the data more suitable for PCA and FA by aligning with their assumptions about scale and variance. Without transformation, dimensionality reduction would likely overemphasize the richest countries,

underemphasize the poorest countries, and obscure meaningful global patterns.

Together, these preprocessing steps—log transformation and standardization—ensure that the GDP per capita data is well-prepared for dimensionality reduction. They allow PCA, factor analysis, and t-SNE to uncover meaningful structure in global economic trajectories without being skewed by extreme values or scale differences.

4 Application and Results

Since the GDP per capita values across years are highly correlated, dimensionality reduction techniques like PCA and FA are particularly well-suited to uncover underlying patterns in the data by summarizing shared trends over time. Although both methods reduce dimensionality, they differ in purpose: PCA helps summarize overall economic levels, and FA aims to capture shared economic trends across countries. To complement these linear methods, we also apply t-SNE, a nonlinear technique designed for visualization, which can capture complex, local patterns in the data that PCA and FA might miss. Together, these methods provide a multi-perspective view of global economic trajectories.

4.1 Principal Component Analysis

By reducing the dataset to a few uncorrelated principal components, PCA allows us to summarize each country's economic trajectory and visualize similarities and differences between countries in a low-dimensional space. We applied PCA to the log-transformed and standardized GDP per capita data from 1990 to 2023.

The scree plot (Figure 3) shows that the first principal component (PC1) explains approximately 97.1% of the total variance, while PC2 accounts for another 2.0%. Together, PC1 and PC2 capture over 99% of the total variance in the dataset. This rapid drop-off in explained variance suggests that nearly all meaningful variation across countries can be visualized in just two dimensions, justifying our use of a 2D PCA plot.

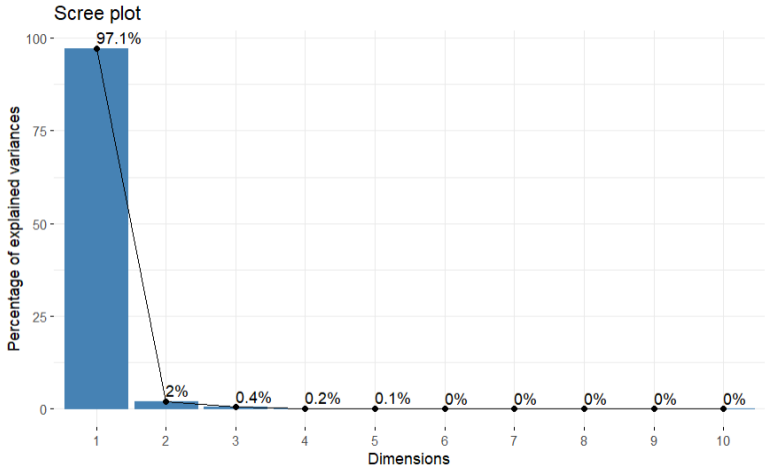

Figure 3: Scree plot showing the variance explained by each principal component

The loadings associated with PC1 are positive and nearly uniform across all years, suggesting that PC1 represents the overall level of GDP per capita over the time period. Countries with high PC1 scores tend to have consistently high income levels across the years, while those

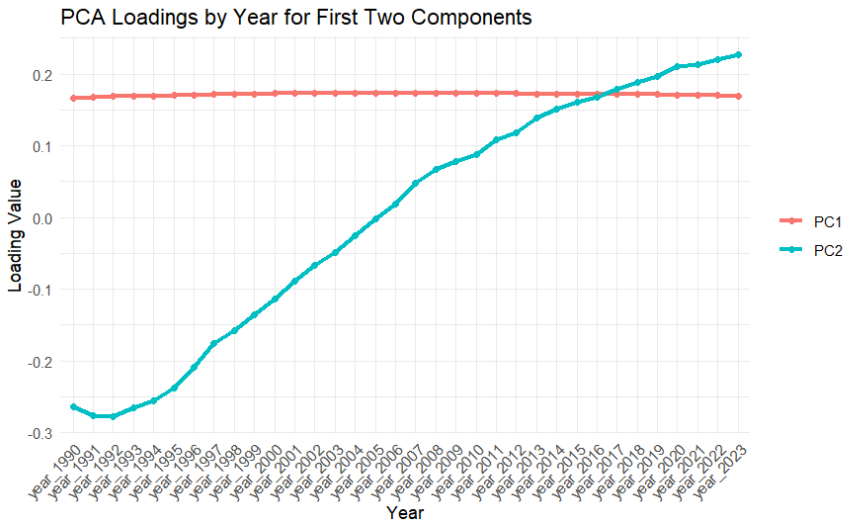

Figure 4: Loadings for PC1 and PC2 across years

with low scores tend to have persistently low income. Thus, PC1 reflects long-term economic standing.

PC2, by contrast, distinguishes countries based on their trajectory of change. The loadings for PC2 are negative for earlier years and positive for more recent years, indicating that this component captures whether a country's GDP has increased or decreased over time. Countries that started with low GDP and grew significantly have higher PC2 scores, while those with stagnant or declining GDP have lower or even negative PC2 values. In this sense, PC2 helps identify growth patterns, separate from overall wealth.

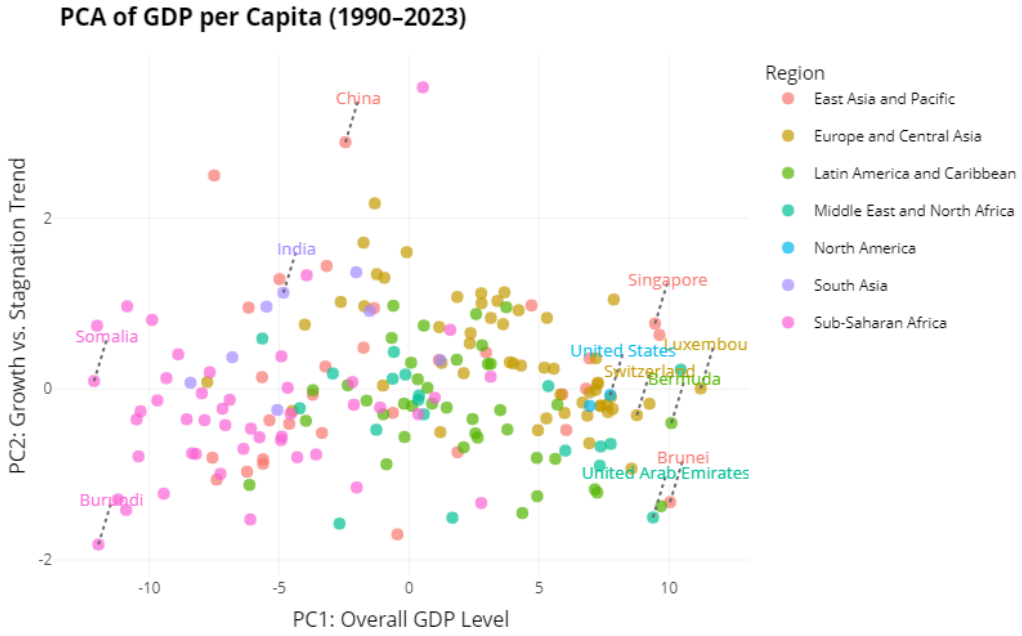

Figure 5: PCA plot of countries colored by region

The PCA plot highlights meaningful regional patterns and distinct country-level economic

trajectories. Sub-Saharan African countries like Somalia and Burundi appear far to the left, reflecting low overall GDP levels. Somalia's slightly elevated PC2 score suggests modest recent growth, while Burundi remains low on both components, indicating persistent economic stagnation. India, representing South Asia, lies lower on PC1 but higher on PC2, reflecting its consistent upward growth from a lower base. China, in the East Asia and Pacific region, shows a mid-level PC1 score and one of the highest PC2 values in the dataset, capturing its dramatic rise in income over the past two decades. Singapore also stands out with both high economic standing and strong growth, as reflected in its high PC1 and PC2 scores.

High-income countries such as Bermuda, Switzerland, Luxembourg, and the United States are located far to the right on PC1, confirming their consistently high GDP per capita. However, their near-zero or negative PC2 values suggest that their income levels have remained relatively stable over time. This includes wealthy countries like the United Arab Emirates and Brunei, which appear low on PC2, indicating limited recent growth despite their high GDP levels. Latin America and the Caribbean countries have mid-level incomes and generally flat or inconsistent growth trajectories, potentially reflecting economic volatility or stalled development. These patterns confirm that PCA effectively captures both income levels and growth dynamics, offering a powerful lens for comparing regional and national economic progress.

4.2 Factor Analysis

To explore latent structures underlying the economic development of countries, we applied factor analysis to the log-transformed and standardized GDP per capita dataset covering the years 1990 to 2023. Before model fitting, we verified that the data was suitable for FA. Bartlett's test of sphericity was statistically significant, indicating strong correlations among variables, and the Kaiser-Meyer-Olkin (KMO) index was 0.97, confirming excellent sampling adequacy.

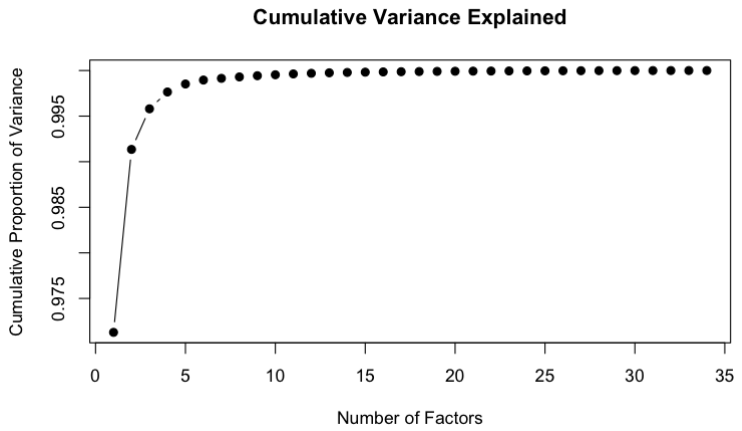

Figure 6: Scree plot showing the variance explained by each latent variable.

A two-factor model was extracted using the `factanal()` function in R. Although the scree plot suggested that a single dominant factor could explain most of the variance, retaining two factors allowed us to capture meaningful temporal dynamics. The initial unrotated results resembled PCA but lacked interpretability. To address this, we applied Varimax rotation, which enhanced the clarity of the factor structure. The rotated factor loadings revealed a clear pattern: Factor 1 loaded heavily on earlier years (primarily 1990–2005), while Factor 2 had

increasing loadings on recent years, especially from 2008 onward. This allowed us to interpret Factor 1 as representing historical economic strength and Factor 2 as capturing modern GDP influence.

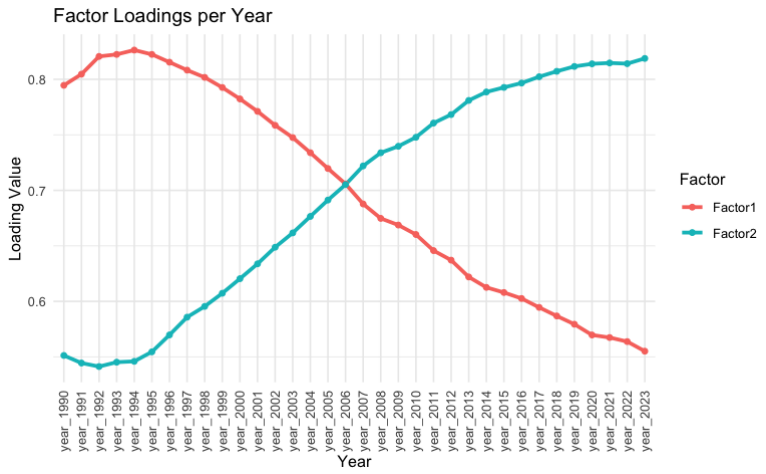

Figure 7: Rotated factor loadings from 1990 to 2023.

The estimated factor scores provided latent coordinates for each country in the two-factor space. Countries such as Luxembourg and Switzerland scored highly on both factors, indicating sustained economic strength across decades. In contrast, countries like Brunei and the United Arab Emirates had high scores on Factor 1 but low on Factor 2, suggesting strong early economic influence that did not persist. Conversely, emerging economies such as China and Singapore showed low scores on Factor 1 but high on Factor 2, reflecting rapid growth in the last few decades. Countries like Somalia and Burundi scored low on both factors, indicating consistently low GDP levels over time.

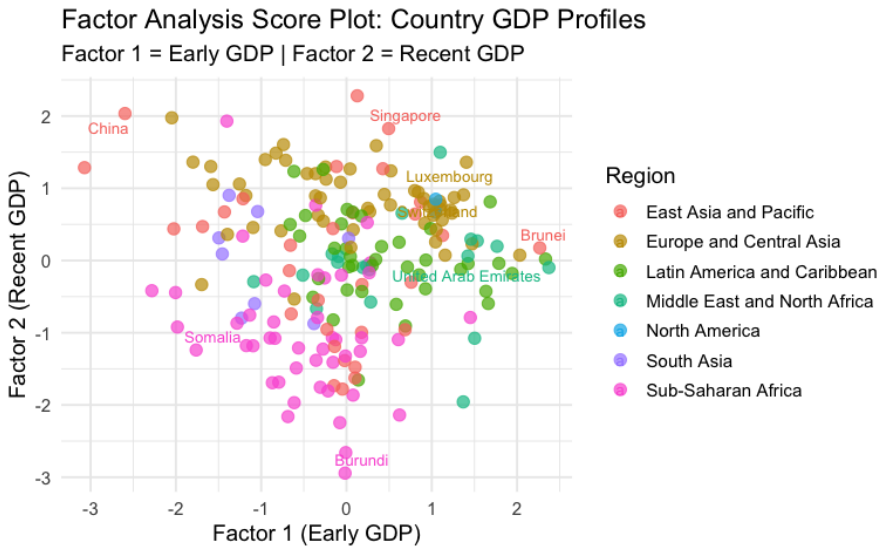

Figure 8: Factor Analysis score plot colored by world region. Countries are positioned by early (factor 1) and recent (factor 2) GDP influence, with selected nations labeled for interpretability.

4.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

To further explore the structure of GDP per capita data across countries, we applied t-Distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear dimensionality reduction technique that preserves local similarities among high-dimensional observations. We used the same log-transformed and standardized dataset employed in PCA and Factor Analysis to ensure a consistent basis for comparison. The primary tuning parameter in t-SNE is the *perplexity*, the effective number of neighbors each point has in high-dimensional space. Perplexity controls the balance between local and global structure in the resulting embedding. We ran t-SNE with perplexity values ranging from 5 to 50, to examine how the layout of countries evolved across different scales.

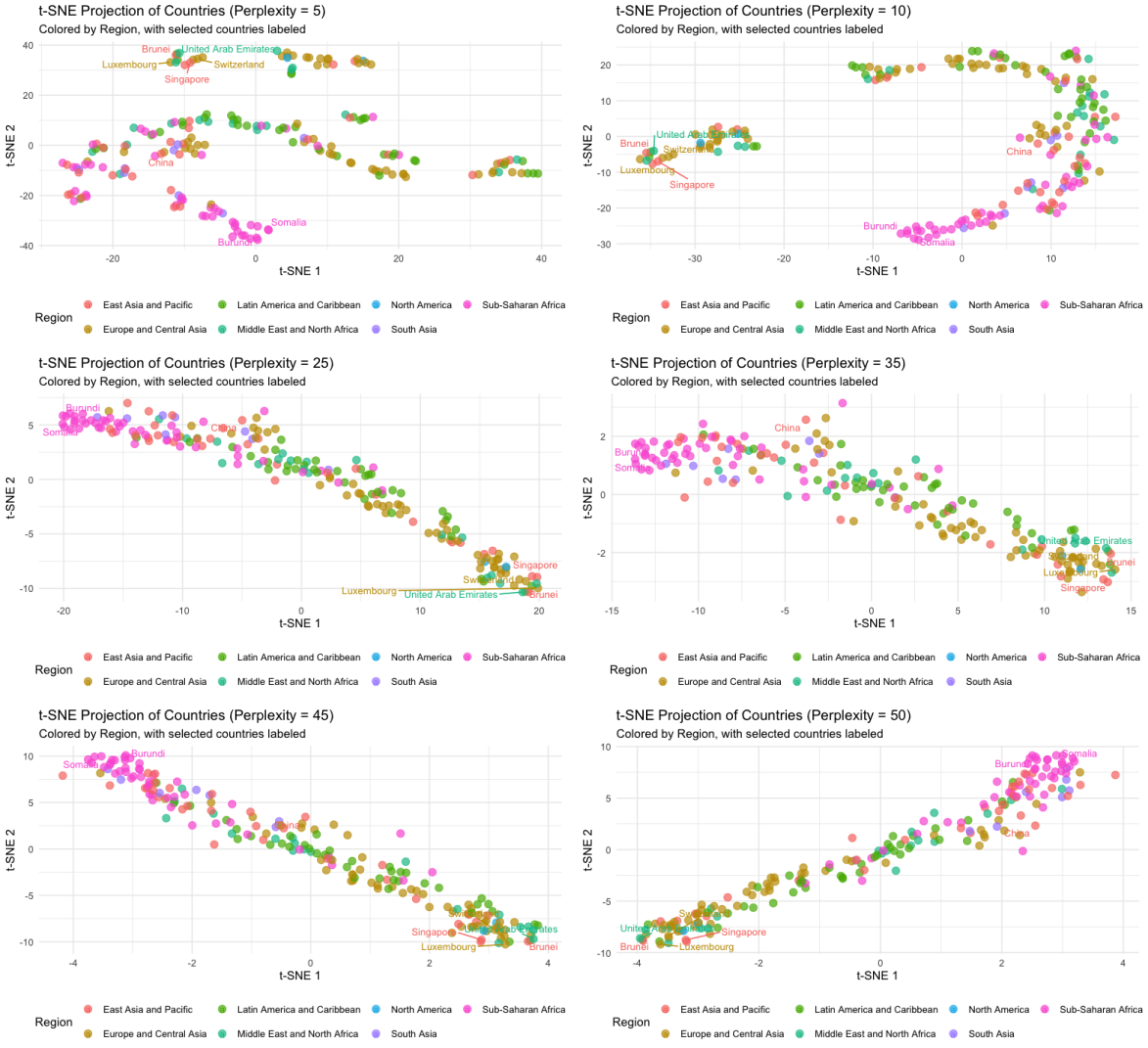

Figure 9: t-SNE projections of countries based on GDP per capita trajectories at increasing perplexity values (5, 10, 25, 35, 45, 50).

At low perplexity values (5 and 10), the plots emphasize local neighborhood structure. Countries with very similar GDP patterns—such as Somalia and Burundi in Sub-Saharan Africa, or Switzerland, Luxembourg, and Singapore among high-income nations—form compact clusters. These clusters are isolated from one another, highlighting fine-grained similarity within regions

but providing limited insight into global structure. China appears somewhat apart from both low and high-income countries, reflecting its trajectory of rapid growth from a lower base.

As perplexity increases (from 20 to 50), the t-SNE layout expands, gradually revealing broader economic gradients. The distinction between low-income countries (e.g., Somalia, Burundi) and high-income countries (e.g., Luxembourg, Switzerland) becomes clearer and more structured. The projection at perplexity 40 and 50 shows a quasi-linear gradient from bottom-left to top-right (or right-to-left in flipped views), where countries are aligned along a global income spectrum. While regional clusters are still visible, countries begin to space out more smoothly along economic lines.

Importantly, we observe that the axes flip and rotate across different perplexity values. For instance, the same cluster of Sub-Saharan African countries may appear in the top-right in one frame and bottom-left in another. This is a known and expected behavior in t-SNE since the method does not preserve axis orientation across runs. These transformations do not affect the underlying relationships between countries; their relative positions to one another remain consistent even when the coordinate system changes.

Regional coherence remains strong throughout: Sub-Saharan Africa consistently appears as a distinct low-income group; Europe and Central Asia form dense clusters among high-income countries; and South Asia, along with East Asia & Pacific, exhibits upward transitions depending on each country's growth trajectory. The Middle East and North Africa, as well as Latin America and the Caribbean, often appear interspersed within middle-income ranges, reflecting mixed growth patterns and economic diversity within those regions.

Overall, t-SNE offers a flexible, nonlinear view of the global economic landscape. It captures both local neighborhood structures and broader global trends, depending on the perplexity setting. However, the method's lack of axis interpretability and sensitivity to initialization and hyperparameters warrant careful consideration in comparative analyses.

5 Discussion and Comparative Analysis

This study applied three dimensionality reduction techniques—Principal Component Analysis (PCA), Factor Analysis (FA), and t-Distributed Stochastic Neighbor Embedding (t-SNE)—to explore patterns in global GDP per capita trajectories from 1990 to 2023. Each method offered distinct insights into the structure of the data and revealed complementary perspectives on global economic development.

PCA efficiently summarized global economic variation, with its first component (PC1) alone capturing over 97% of the total variance. PC1 effectively ranked countries by their average income level across the years. The second component, accounting for a smaller portion of variance ($\sim$2%), added important interpretive value by capturing long-term growth trends. It helped distinguish countries with rising GDP trajectories from those with stable or early economic influence. Together, PC1 and PC2 offered a clear view of both economic scale and growth dynamics. However, interpretation beyond the first two components was limited.

Factor Analysis, especially after Varimax rotation, extracted a more interpretable two-factor structure. Factor 1 loaded on early years and represented historical economic influence, while Factor 2 reflected more recent economic strength. This allowed us to assess not just the level of development, but also the timing of economic emergence across countries. For instance, China and Singapore exhibited high scores on recent factors, while Brunei and the UAE showed stronger influence earlier in the period. FA thus provided richer temporal interpretation compared to PCA.

t-SNE contributed a nonlinear and visually intuitive view of the data. At low perplexities, it revealed tight clusters of economically similar countries, preserving fine-grained local relationships. As perplexity increased, it unveiled global gradients that separated low- and high-income nations along a continuous spectrum. Countries like Somalia and Burundi consistently appeared at one end, while Luxembourg and Switzerland appeared at the other. Although t-SNE lacks interpretable axes and is sensitive to hyperparameter tuning, it proved effective at exposing nuanced and nonlinear structures that PCA and FA might miss.

Each method comes with trade-offs. PCA is computationally efficient and mathematically transparent but assumes linearity and emphasizes global variance. FA offers strong interpretability through its modeling of latent structures but is sensitive to factor extraction and rotation choices. t-SNE excels at uncovering local patterns and nonlinear relationships but lacks stability and global interpretability, requiring careful control of perplexity and initialization.

Despite these differences, all three methods consistently grouped countries with similar economic behavior and highlighted broad patterns of global inequality, sustained performance, and emerging growth. For example, Sub-Saharan African nations clustered consistently at the low end of all projections, while countries like China, India, and Vietnam showed upward movement consistent with rapid development.

There are several opportunities for extending this analysis. Incorporating other socioeconomic indicators—such as education, life expectancy, or CO_2 emissions—could reveal deeper latent structures and improve interpretability. Future work might also examine regional or income-specific patterns through localized dimensionality reduction, or apply time-series modeling to directly capture growth dynamics. These extensions could offer a more detailed understanding of how economic trajectories evolve across space and time.

References

[1] Our World in Data. *GDP per Capita (2025)*. Retrieved from https://ourworldindata.org/grapher/gdp-per-capita-worldbank?tab=table

[2] Van der Maaten, L. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[3] MATH 250 Course Website. Slides 10, 13 and 17.

[4] Strang, G. (2019). *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press.

[5] Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

[6] Heiss, A. *Data Visualization for Fall 2023*. Retrieved from https://datavizf23.classes.andrewheiss.com/

[7] R Core Team. `Rtsne` function. Retrieved from https://www.rdocumentation.org/packages/Rtsne/versions/0.17/topics/Rtsne

[8] R Core Team. `factanal` function. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/factanal

[9] R Core Team. `prcomp` function. Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp