# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

**Group 10 - 4CSD**

**Baylon, Karyle Zhienelle V.**

**Francisco, Leann Joy Y.**

**Magtanong, Ralph Daven M.**

---

**Use Python to analyze relationships and interactions between multiple variables in a dataset. You will explore correlations, identify underlying factors, reduce dimensionality, and classify groups based on their characteristics.**

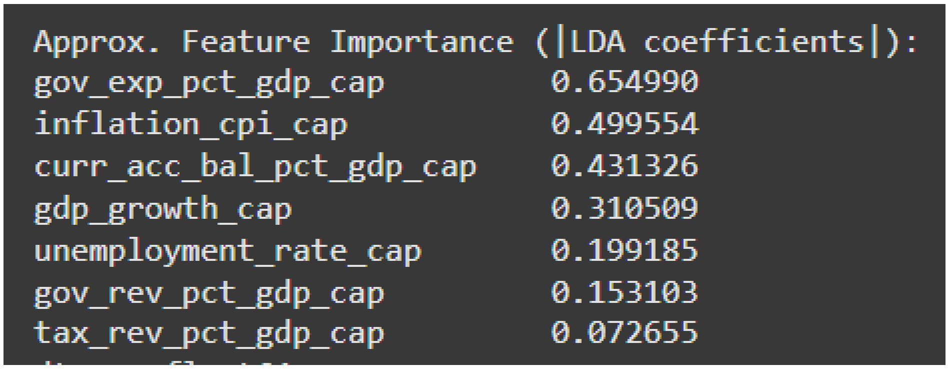
**QUESTIONS:**

1. **Which combination of features best separates high-income vs. low-income countries?**

PRELIMINARY TESTS:

**[Bartlett] chi2=6791.66, dof=21, p-value=0**

**[KMO] overall=0.672**



```
Approx. Feature Importance (|LDA coefficients|):
gov_exp_pct_gdp_cap          0.654990
inflation_cpi_cap            0.499554
curr_acc_bal_pct_gdp_cap     0.431326
gdp_growth_cap               0.310509
unemployment_rate_cap        0.199185
gov_rev_pct_gdp_cap          0.153103
tax_rev_pct_gdp_cap          0.072655
```

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

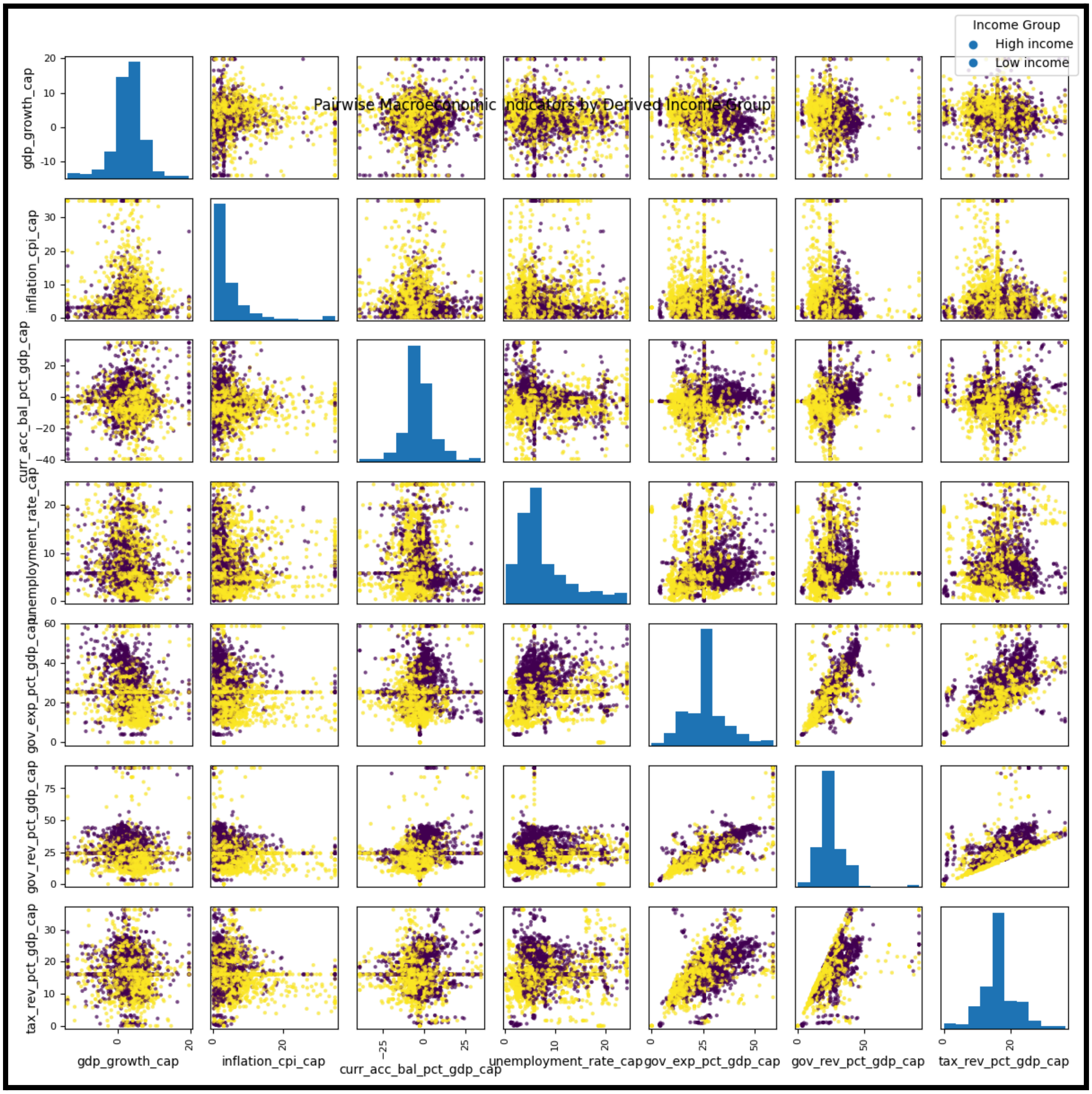


***Chart type:*** DA Scatter Plot

***Encoding:***

- **x-axis and y-axis:** Values of macroeconomic indicators
- **Color:** Defines the two categorical groups

- **Marks:** Circle points that represent each country observation
- **Histogram:** Distribution of each variable individually

  **JUSTIFICATION**:

  A scatter plot matrix allows simultaneous comparison of pairwise relationships among multiple variables while highlighting group differences across all feature

- Main takeaway in one sentence
  - Based on the scatter plot matrix, the combination of gov_exp_pct_gdp_cap (government expenditure) and tax_rev_pct_gdp_cap (tax revenue), both as a percentage of GDP, provides the clearest separation between high- and low-income countries.

- One design decision and its benefit
  - Using **color encoding for IncomeGroup** makes it easy to visually detect where clusters of high vs. low income diverge across the feature pairs, improving interpretability without requiring separate charts.

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

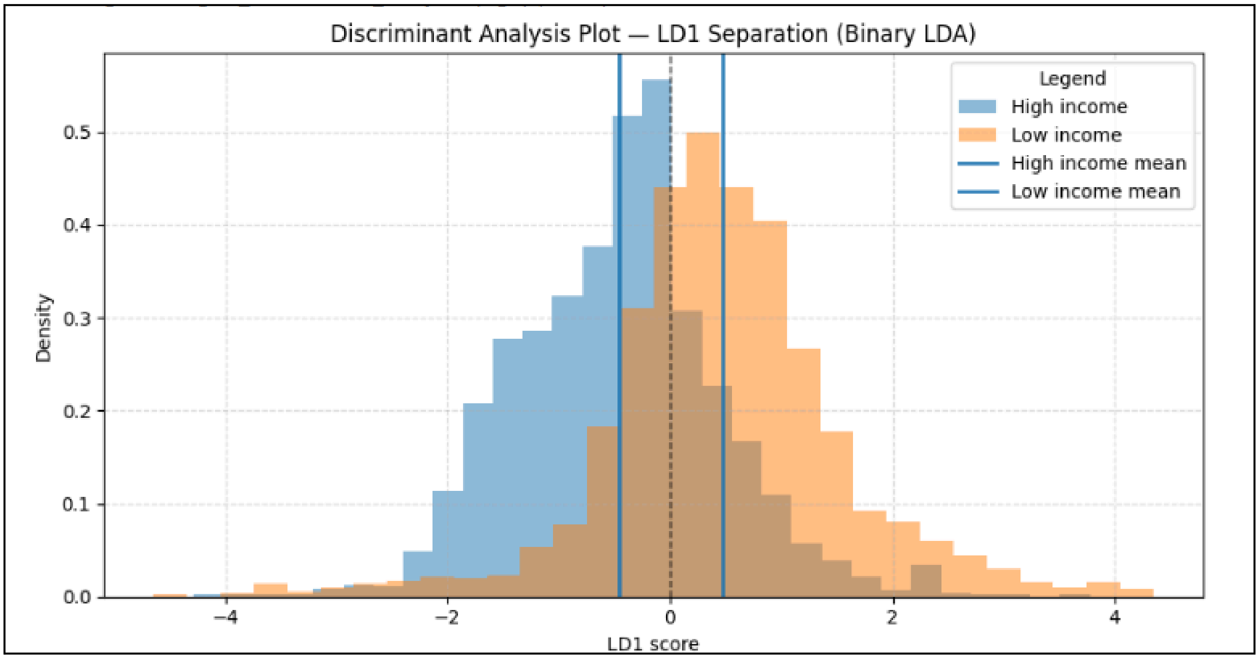


***Chart type:*** Histogram with density overlay (Discriminant Analysis Plot along LD1).

***Encoding:***

- **x-axis (x-position):** Linear Discrimination 1 score
- **y-axis (y-position):** Density
- **Color (fill):** Income group category
- **Vertical lines:** Class means (average LD1 score for each group).

  **JUSTIFICATION**:

  The plot directly shows how the single discriminant function separates the two income groups by comparing their score distributions, highlighting both overlap and distinct group means.

- Main takeaway in one sentence

  - The discriminant analysis plot demonstrates that the features effectively separate high-income and low-income countries, as evidenced by their distinct mean scores on the first linear discriminant axis, although their score distributions still overlap noticeably.

- One design decision and its benefit
  - Including **vertical lines for class means** helps the viewer quickly identify the central tendency of each group, improving interpretability beyond the raw histograms.

2. **Can we reduce the number of macroeconomic indicators into a smaller set of key dimensions that explain most of the variance across countries?**

PRELIMINARY TESTS:

**[Bartlett] chi2=11392.47, dof=45, p-value=0**

**[KMO] overall=0.624**

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

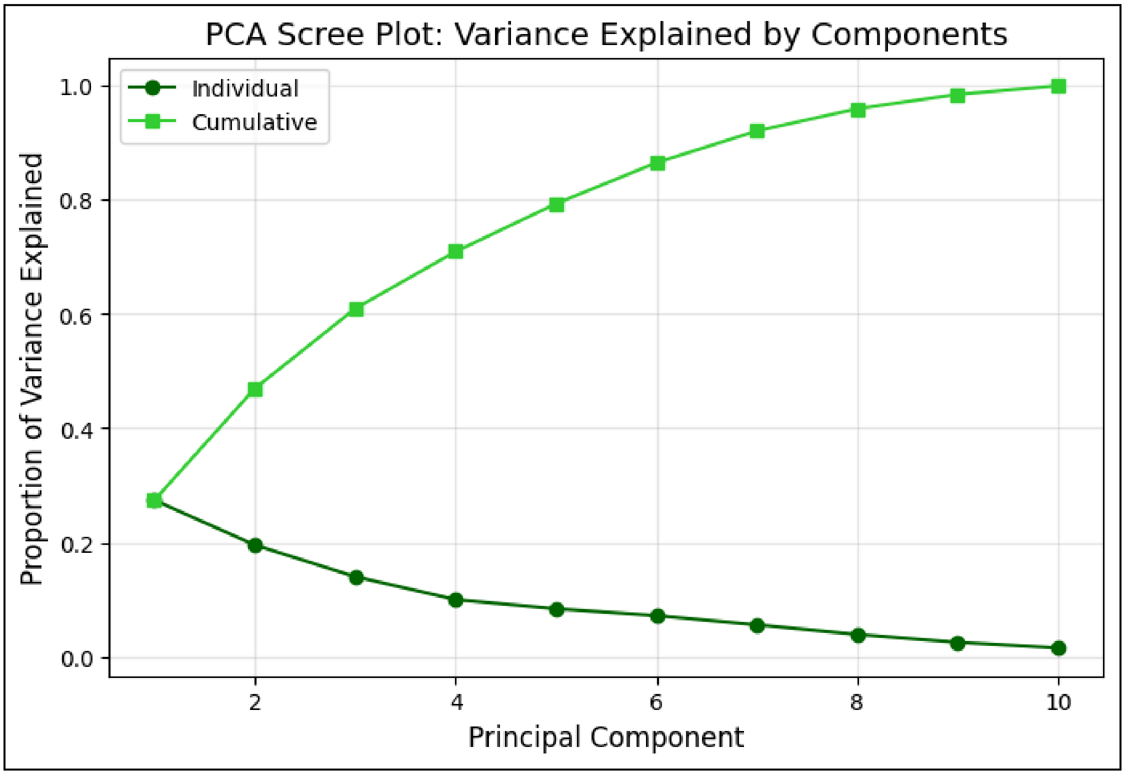


***Chart type***: *PCA Scree Plot*

***Encoding***:

- **X-axis:** Principal components (PC1, PC2, ... PC10)
- **Y-axis:** Proportion of variance explained (quantitative)
- **Lines:** Two lines: one for individual variance explained per component, one for cumulative variance explained.

**JUSTIFICATION**: The most effective visual tool for determining the quantity of elements to keep in PCA is a **scree plot**. It makes it simple to determine the "elbow point" at which extra components add minimal value by displaying both the cumulative coverage and the individual contributions of each principal component. By using two lines (individual and cumulative), we can see how much each component contributes as well as how they add up to account for the majority of the variance.

- Main takeaway in one sentence
    - More than 70% of the variance can already be explained by the first three principal components, indicating that the macroeconomic indicators can be effectively reduced down to a small number of dimensions.

- One design decision and its benefit
    - We used two lines (one for individual and one for cumulative variance explained) in distinct shades of green, which improves clarity by letting viewers see both perspectives of variance contribution without switching plots.

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

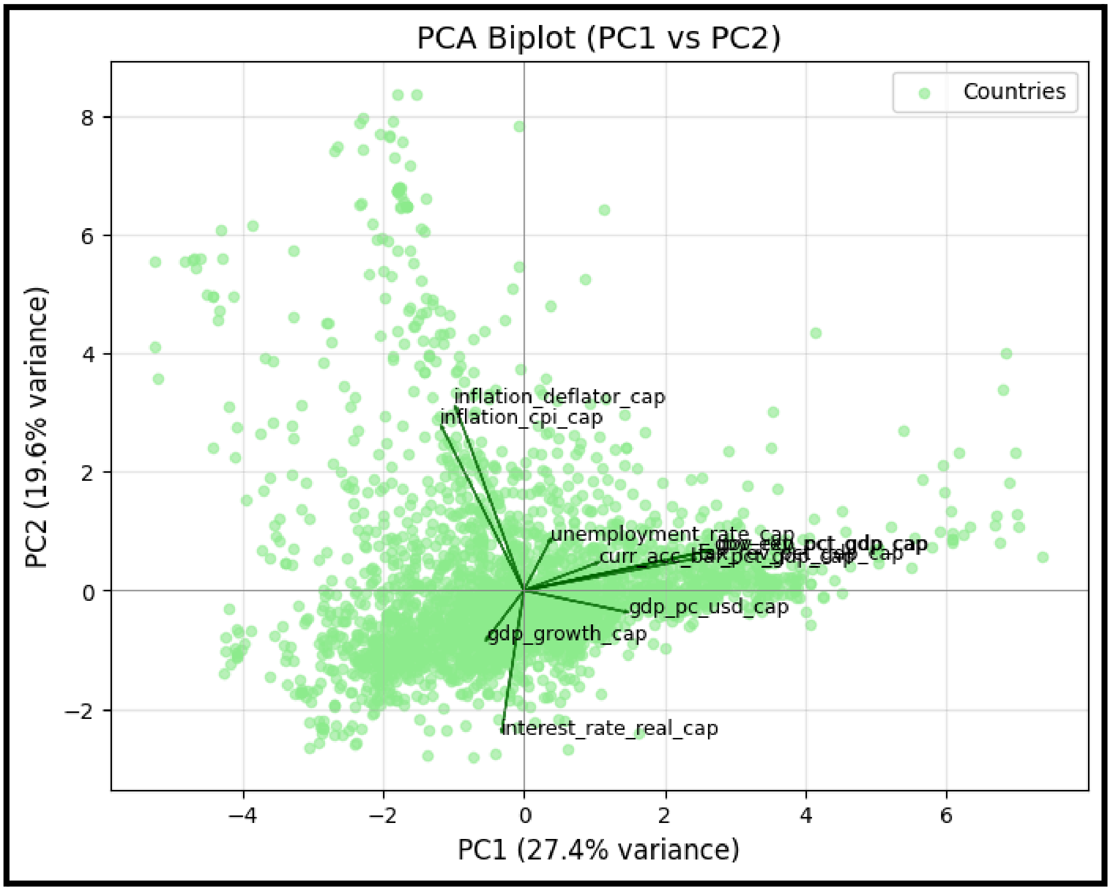


***Chart type:*** *PCA Biplot (PC1 vs PC2)*

***Encoding:***

- **X-axis:** PC1 scores (27.4% variance explained)
- **Y-axis:** PC2 scores (19.6% variance explained)
- **Points**: Each observation (country-year)
- **Arrows:** Variable loadings showing contribution of each macroeconomic indicator to PC1 and PC2.

**JUSTIFICATION**: Since a **biplot** illustrates the relationship between variables and observations (countries) and principal components, it is perfect for interpreting PCA results. The arrows show the indicators causing the separation, while the scatter shows clustering patterns among nations (e.g., inflation variables pointing together suggest they load similarly). Because of this dual perspective, the abstract idea of PCA is easier to understand and can be linked back to the source data.

- Main takeaway in one sentence
  - The first two components show clear patterns in which inflation-related variables cluster together, whereas growth and fiscal indicators load differently, implying that macroeconomic variance is driven by a few unique dimensions.

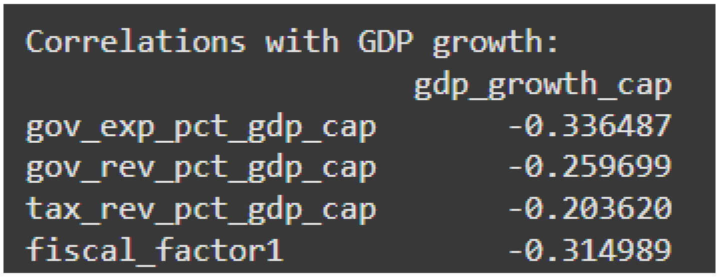
- One design decision and its benefit
  - We included percentage labels of variance explained (e.g., “PC1: 27.4%”) on the axes, which helps viewers immediately understand how much information is retained in each component, making the interpretation of dimensionality reduction clearer.

3. **How strongly are fiscal variables (government expenditure, revenue, and tax revenue as % of GDP) correlated with GDP growth?**

PRELIMINARY TESTS:

**[Bartlett] chi2=394.31, dof=3, p-value=0**

**[KMO] overall=0.691**



```
Correlations with GDP growth:
                    gdp_growth_cap
gov_exp_pct_gdp_cap      -0.336487
gov_rev_pct_gdp_cap      -0.259699
tax_rev_pct_gdp_cap      -0.203620
fiscal_factor1           -0.314989
```

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

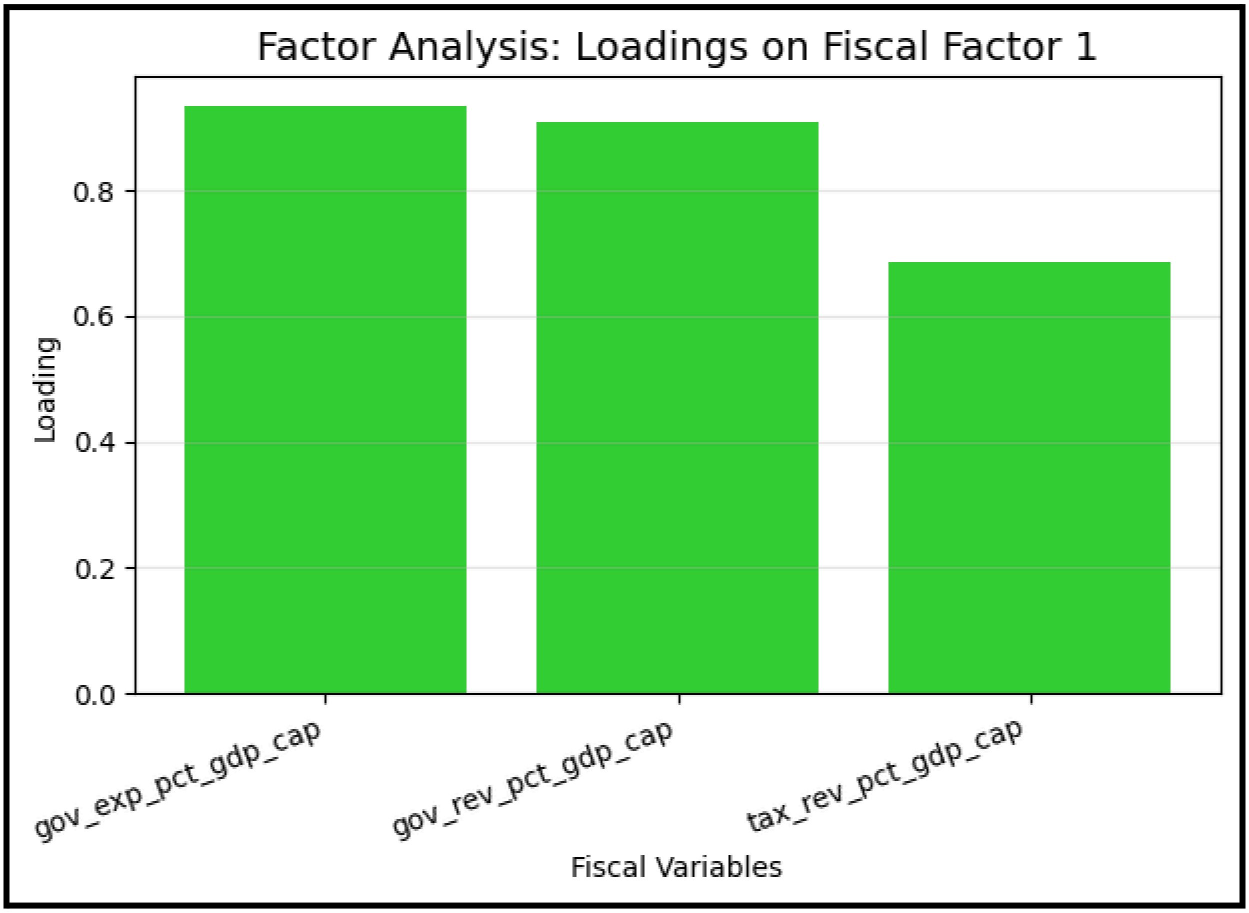


***Chart type***: *Factor Loadings Bar Chart*

***Encoding***:

- **X-axis:** Fiscal variables
- **Y-axis:** Factor loadings on the first fiscal factor

**JUSTIFICATION**: The most effective way to show how much each fiscal variable contributes to the underlying factor is a **bar chart of factor loadings**. It makes it simple to compare the relative strength of government expenditure, revenue, and tax revenue in forming the latent “fiscal stance” dimension. By displaying standardized loadings as bars, we can clearly see which variables dominate the factor and which play a smaller role, without visual clutter.

- Main takeaway in one sentence
    - The fiscal variables of government expenditure, government revenue, and tax revenue as a percentage of GDP show weak negative correlations with GDP growth, suggesting that higher fiscal intensity is weakly associated with lower GDP growth.

- One design decision and its benefit
    - We chose to standardize the fiscal variables before applying Factor Analysis so that each was measured on the same scale. This ensures that differences in units or magnitudes do not bias the factor loadings. By doing so, we can confidently interpret the loadings as genuine relative contributions rather than artifacts of scaling.

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

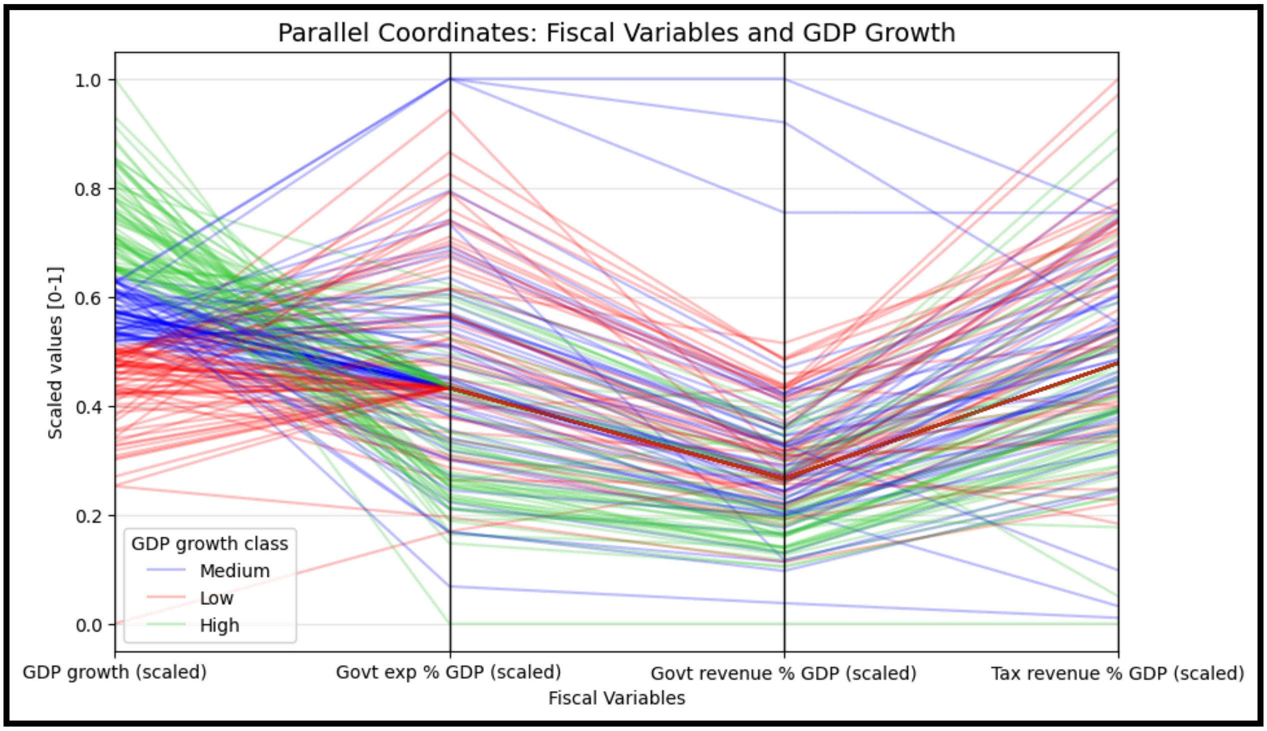


***Chart type:*** *Parallel Coordinates Plot*

***Encoding:***

- **X-axis:** Fiscal Variables
- **Y-axis:** Normalized values (0–1) for each variable
- **Color**: GDP growth class (Low, Medium, High)

**JUSTIFICATION**: A **parallel coordinates plot** is the most effective tool for comparing multiple fiscal variables alongside GDP growth across countries in a single view. It allows us to visualize entire fiscal profiles at once and detect whether growth classes (Low, Medium, High) show systematic differences in expenditure, revenue, and tax revenue patterns. By using colors to represent GDP growth classes, we can highlight trends across dimensions while making it easy to spot overlaps and divergences between groups.

# MULTIVARIATE DATA ANALYSIS WITH PYTHON - LAB ACT 5

*DATA ANALYSIS AND VISUALIZATION (CS ELEC 3C)*

- Main takeaway in one sentence
  - Countries with higher GDP growth tend to cluster at the lower end of government expenditure and revenue, showing a negative association between fiscal intensity and growth.

- One design decision and its benefit
  - I normalized all variables to a [0–1] range before plotting so they would be visually comparable on the same axis. Without scaling, GDP growth percentages and fiscal shares could not be meaningfully compared on one chart. This design choice helped me clearly see the relative patterns and avoid misleading impressions caused by differences in units.