UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA
CDLM DATA SCIENCE

FREEDOM IN THE WORLD: AN EMPIRICAL STUDY OF DIGITAL, GENDER, AND EDUCATIONAL INFLUENCES

PREPARED BY :

ANY DAS (922710)
MAHNOOR FATIMA (924810)
AAROHI MISTRY (925352)
SUMMAN GUL (925663)

DATA SCIENCE LAB – A.Y. 2024/2025
PROFESSORS: M. FATTORE AND S. GERLI

Table of Contents

Introduction

1. Data Description

1.1 Data Source and Overview

This analysis draws on *Freedom in the World*, the flagship annual report published by Freedom House. Recognized globally for assessing political rights and civil liberties, the dataset is grounded in decades of rigorous fieldwork, expert consultations, verified news sources, and NGO reports. Each country and territory is evaluated by independent analysts and undergoes an internal review by Freedom House staff to ensure accuracy and consistency.

Our study uses a panel dataset spanning 2013–2025, comprising 2,723 country-year observations for 209 countries and territories across 13 annual editions. It includes 44 key variables that capture detailed indicators of political rights and civil liberties across six regions: Africa, the Americas, Asia-Pacific, Eurasia, Europe, and the Middle East.

1.2 Key Variables and Indicators

Each observation in the dataset is assessed against a comprehensive framework, which includes the following key indicators:

Primary Ratings: The core of the dataset is represented by Political Rights (PR) and Civil Liberties (CL), two aggregate measures that evaluate the overall level of freedom in a country. Both indicators are scored on a scale from 1, indicating the highest level of freedom, to 7, indicating the lowest, offering a concise overview of a nation's political and civil environment.

Category Scores (A–G): The dataset includes seven disaggregated category scores, labeled A through G, each ranging from 0 to 16, which provide detailed insight into specific facets of freedom. Category A assesses the Electoral Process, B evaluates Political Pluralism and Participation, C measures the Functioning of Government, and D captures Freedom of Expression and Belief. Category E focuses on Associational and Organizational Rights, F on the Rule of Law, and G examines Personal Autonomy and Individual Rights. Together, these scores offer a comprehensive view of the different dimensions that shape a country's overall freedom profile.

Sub-indicators: Each category is composed of discrete questions (e.g., A1, A2, etc.) scored from 0 to 4.

Overall Status: A categorical classification based on the aggregate scores: Free (F), Partly Free (PF) and Not Free (NF).

The dataset also includes fields for additional discretionary questions (Add Q, Add A) used for methodological adjustments.

Table 1:Dataset Coverage Summary

Metric	Value
Total Country-Year Observations	2,723
Unique Countries & Territories	209
Regions	6
Years Covered (Editions)	13 (2013–2025)
Status Categories	3 (F, PF, NF)
Core Variables	44

The distribution of observations across regions and freedom statuses, detailed in Table 2, reveals significant global disparities. Europe accounts for the largest number of observations classified as Free (470), reflecting its overall strong democratic performance. Conversely, Africa has the highest number of observations in the Not Free category (315), highlighting acute challenges to political rights and civil liberties on the

continent. The Americas and Asia show a more mixed distribution, with a substantial number of countries falling into the Partly Free category.

Table 2: Regional Distribution by Status and Region

Status	Africa	Americas	Asia	Eurasia	Europe	Middle East	Total
Free (F)	125	301	222	0	470	13	1,131
Partly Free (PF)	288	125	196	84	81	30	804
Not Free (NF)	315	33	141	139	8	152	788
Total	728	459	559	223	559	195	2,723

2. Descriptive Statistics

The dataset exhibits substantial variation across the key indicators of political rights, civil liberties, and overall freedom. The Total Score, which combines measures of political rights (PR) and civil liberties (CL), averages 57.20 out of 100 with a standard deviation of 30.61, highlighting significant global disparities. Scores range from a minimum of -3 to a maximum of 100. The PR and CL component scores show similar patterns of dispersion, with mean values of 22.54 (std: 13.64) and 34.66 (std: 17.28), respectively. The 1–7 PR and CL ratings have global means around 3.6 and 3.5, indicating that, on average, countries are positioned near the threshold between Partly Free and Not Free status.

The disaggregated category scores (A–G) provide further insight into specific dimensions of governance and freedom. For instance, the Functioning of Government (C) score averages 6.00, while Rule of Law (F) averages 7.78, reflecting the differences in governance quality and legal enforcement across countries. Other categories such as Electoral Process (A) and Political Pluralism (B) also show considerable spread, illustrating diverse challenges across nations.

Table 3: Descriptive Statistics of Core Indicators

Statistic	PR rating	CL rating	PR	CL	A	B	C	D	E	F	G	Total
Count	2,723	2,723	2,723	2,723	2,723	2,723	2,723	2,723	2,723	2,723	2,723	2,723
Mean	3.63	3.49	22.54	34.66	7.22	9.50	6.00	10.41	7.21	7.78	9.25	57.20
Std	2.23	1.95	13.64	17.28	4.49	5.44	3.77	4.74	3.99	4.95	4.19	30.61
Min	1.00	1.00	-3.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-3.00
25%	1.00	2.00	10.00	20.00	3.00	4.00	3.00	7.00	4.00	4.00	6.00	29.50
50%	3.00	3.00	26.00	35.00	9.00	11.00	6.00	11.00	8.00	7.00	9.00	61.00
75%	6.00	5.00	36.00	51.00	12.00	15.00	9.00	15.00	11.00	12.00	13.00	86.50
Max	7.00	7.00	40.00	60.00	12.00	16.00	12.00	16.00	12.00	16.00	16.00	100.00

Overall, these descriptive statistics demonstrate that freedom outcomes are highly heterogeneous across countries and regions. The variation in both aggregate scores and sub-indicators underscores the importance of examining not just overall status but also the nuanced dimensions of political rights, civil liberties, and governance for a comprehensive understanding of global freedom patterns.

Global Freedom Trends

Figure 1 illustrates the evolution of average freedom scores from 2013 to 2025 across six regions. The global trend indicates a gradual decline in political rights and civil liberties over the 13-year period. Europe and the Americas, starting from high average scores, show steady decreases, while Eurasia and the Middle East & North Africa (MENA) remain at lower levels and continue to worsen. Notably, the COVID-19 pandemic around 2020 appears to have accelerated this downward trajectory. Overall, regional disparities are narrowing, not due to improvements in low-scoring regions, but as a result of widespread declines, highlighting a growing global challenge for freedom.

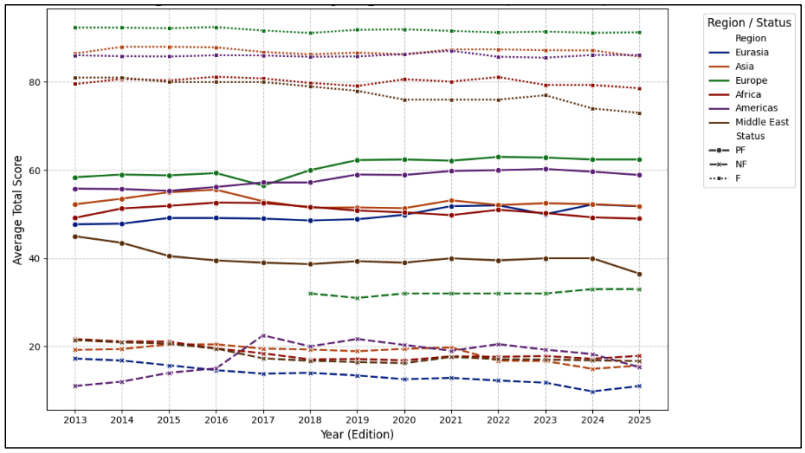


Figure 1: Global Trends of Average Freedom Scores by Region (2013–2025)

Category Score Distribution

Figure 2 presents a box plot of Freedom House category scores (A–G) across regions, showing medians, spreads, and outliers. Europe exhibits consistently high scores with a narrow range, indicating homogeneously strong freedoms. The Americas display a moderate median but a wide spread, reflecting the presence of both Free and Not Free countries. Asia shows substantial variation, with countries ranging from Partly Free to Not Free and a median in the Partly Free category. Africa demonstrates a broad distribution, including both relatively Free nations and highly repressive ones. In contrast, the Middle East & North Africa (MENA) and Eurasia are tightly clustered at low scores, highlighting widespread and consistent repression. This visualization emphasizes regional disparities and internal diversity, revealing both exemplary performers and areas facing severe freedom constraints.

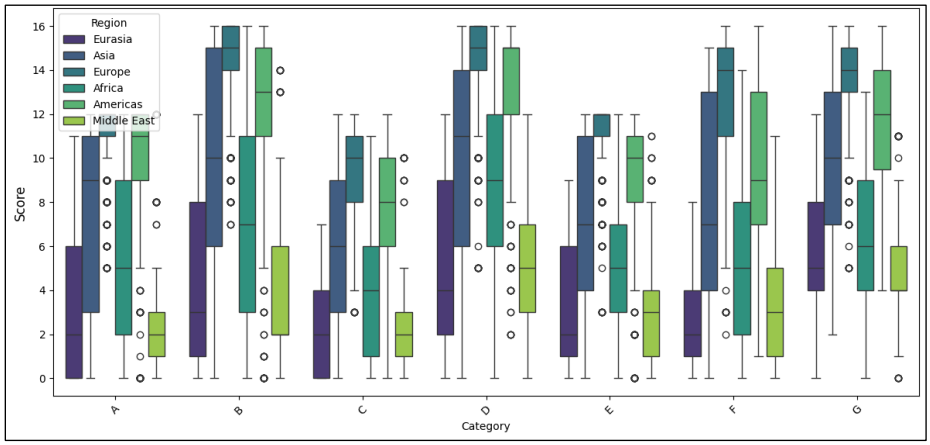


Figure 2: Regional Distribution of Freedom Scores by Category (A–G)

3. Idea and Motivation

The project investigates the relationship between socio-economic variables and measures of freedom using the Freedom in the World dataset. Specifically:

1. What is the impact of internet penetration on freedom of expression across countries?
2. Does female labor force participation correlate with stronger women's rights?
3. Do higher education levels predict greater academic freedom?

By integrating Freedom House indicators (political and civil liberties) with external data sources (World Bank), the study aims to identify patterns that highlight how social, economic, and technological factors influence democratic freedoms.

- **Global Relevance**: Freedom of expression, gender equality, and academic freedom are central to democratic societies. Understanding their drivers provides insight into the conditions that foster open societies.
- **Policy Implications**: Policymakers can benefit from empirical evidence e.g., whether increasing internet access supports free speech, or if expanding women's workforce participation strengthens gender equality.
- **Bridging Datasets**: Combining governance indicators (Freedom House) with socio-economic indicators (World Bank) creates a richer multidimensional analysis beyond political scores alone.
- **Timely Concerns**: With growing digital censorship, gender inequality, and restrictions on academic freedom worldwide, these research questions address some of the most pressing human rights debates of the 21st century.

4. Aim

The aim of this project is to empirically investigate the relationships between freedom indicators and key socio-economic factors. Specifically, the study seeks to:

- Examine how Freedom House indicators, including overall ratings (Political Rights, Civil Liberties, Total Score), category-level scores (A–G), and sub-indicators (A1, A2, …, G4), relate to socio-economic variables such as internet penetration, female labor force participation, and education levels.
- Identify patterns and trends across countries and regions that explain variations in freedoms, highlighting disparities and clusters of similar socio-political dynamics.
- Provide policy-relevant insights on how investments in digital access, education, and gender inclusion can support civil liberties and strengthen democratic governance.
- Offer an evidence-based perspective on the interplay between social and economic development and fundamental freedoms, contributing to informed policy-making and academic discourse.
- Consider temporal and regional contexts, recognizing how global events, technological advancements, and policy shifts influence the observed relationships.

Data Enrichment

To deepen our analysis of Freedom House indicators, we enhanced the raw panel dataset by incorporating three external socio-economic indicators: Internet_Value (internet penetration rate), Female_LFPR (female labor force participation rate), and Tertiary_Enroll_Rate (tertiary education enrollment rate). This enrichment created a clean, consistent, and analysis-ready dataset suitable for robust statistical and regression analyses.

1. Workflow

1.1 Manual Data Cleaning & Country Matching

Principle: Ensure all data entries are consistent and accurate by removing irrelevant rows and harmonizing country names across datasets. This guarantees that analyses are performed on valid, comparable observations.

Metod: Filter irrelevant rows and standardize country names using string matching or reference lists.

Pros: Manual cleaning ensures dataset integrity and prevents errors from mismatched names. When automated approaches are unreliable, this method can maintain high data quality. For large datasets, semi-automated tools or scripts can assist and reduce manual effort while keeping accuracy.

Cons: The process is time-consuming and prone to human errors if not carefully checked. To mitigate these limitations, combining manual checks with automated string matching or fuzzy matching methods can improve efficiency without sacrificing quality.

1.2 Missing Value Imputation

Principle: Fill missing data to allow complete analyses while minimizing distortion of variable distributions.

Formulas:

- Median: $x_i^* = \text{median}(X)$ – robust to outliers

$Where: x_i^* \rightarrow imputed\ value\ for\ missing\ observation$ and

$\text{median}(X) \rightarrow median\ of\ the\ dataset\ X$

- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – suitable for normally distributed variables

$Where: \bar{x} \rightarrow sample\ mean, n \rightarrow number\ of\ observations$ and

$$\sum_{i=1}^{n} x_i \rightarrow sum\ of\ all\ observed\ values$$

Pros: Imputation maintains central tendency and allows analyses without dropping rows. When missingness is low, simple methods like mean or median are effective. Advanced techniques like multiple imputation or regression-based imputation can further improve accuracy when patterns are complex.

Cons: Imputation may reduce natural variability and ignores relationships with other variables. To mitigate this, predictive imputation methods (e.g., regression imputation) can preserve relationships and reduce bias.

1.3 Data Integration / Merging

Principle: Combine multiple datasets to create a richer dataset with additional indicators while keeping alignment across key variables.

Method: Merge datasets using Country + Year keys.

Pros: Produces more comprehensive datasets, allowing richer analysis. Verification scripts or automated checks can help ensure successful merges.

Cons: May reduce sample size if keys are missing or mismatched, and errors can propagate. Automated merge validation and consistency checks help reduce these issues.

1.4 Column Standardization & Categorical Conversion

Principle: Convert variables to appropriate data types to ensure correct analysis and plotting.

Method: Convert numeric, categorical, or factor types as needed.

Pros: Supports accurate aggregation and analysis. Automated type-checking scripts can improve reliability.

Cons: Misclassification can lead to misleading results. Cross-checking with metadata or summary statistics can reduce errors.

2. Enrichment Implementation

2.1 Data Preparation and Enrichment

Objective: The aim was to create a reliable and comprehensive dataset by selecting relevant countries and years, cleaning inconsistent data, standardizing column names, preparing categorical variables, and incorporating key indicators from external sources.

Process & Parameters: We prepared the data using a combination of hands-on cleaning and systematic processing:

Data Cleaning and Selection:

- **Time Frame:** We focused on 2013 to 2023 to align with available data from other sources, ensuring fair comparisons across indicators.
- **Quality Checks:** We manually removed unnecessary rows, text entries, and inconsistent values.
- **Country Matching:** We retained only countries that appeared in both the original dataset and the external sources.
- **Added Indicators:** We brought in three new metrics from outside datasets:
  - Internet penetration rate (Internet_Value)
  - Female labor force participation rate (Female_LFPR)
  - Tertiary education enrollment rate (Tertiary_Enroll_Rate)

Missing Value Handling: We handled missing values according to the data distribution. For skewed data or cases with outliers, we used the median to minimize their effect. For more evenly distributed data, we used the mean to maintain the overall pattern.

Loading the Dataset: We loaded the cleaned and enriched dataset for 2013–2023 using Python's pd.read_excel(), specifying the file path and sheet_name='Freedom'.

Column Name Standardization: We cleaned up the column names to make the dataset easier to work with. Extra spaces at the beginning or end of names were removed to prevent mismatches during merging or grouping. Non-breaking space characters (\xa0) were replaced with regular spaces for consistency. We also standardized the naming by using underscores instead of spaces, which helped avoid syntax errors and made plotting simpler.

Categorical Variable Preparation: We converted key columns such as Country/Territory, C/T, Region, and Status into categorical types. This made it easier to group the data correctly for aggregation and visualizations.

Rationale: Each step was chosen to ensure a clean, consistent, and analysis-ready dataset. The selected time frame maximized data while keeping it consistent. Country matching guaranteed comparable observations. Standardizing column names and converting categorical variables minimized errors in merging, coding, and visualization. Median or mean imputation addressed missing values effectively without bias. Overall, these choices produced a dataset suitable for robust statistical and regression analysis.

Reproducibility: We documented all steps, including cleaning, country selection, column standardization, categorical variable preparation, adding new indicators, handling missing values, and loading the dataset. This ensures that anyone can follow the same process and achieve the same results, providing a solid foundation for the analyses that follow.

2.2 Exploratory Analysis of Enriched Variables

Objective: The goal was to examine the distribution, variation, and patterns of the newly added indicators, while also checking for outliers and differences across regions or political statuses.

Process & Parameters: We performed a thorough exploratory analysis using both statistical summaries and visualizations:

Descriptive Statistical Analysis:

- **Summary Statistics:** Calculated count, mean, median, standard deviation, minimum, maximum, and quartiles for: Internet penetration rate (Internet_Value), Female labor force participation rate (Female_LFPR) and Tertiary education enrollment rate (Tertiary_Enroll_Rate).
- **Grouped Analysis:** Computed aggregated statistics by Region and Status to identify patterns across geographical and political categories.
- **Dataset Verification:** Checked dataset dimensions (rows × columns) to ensure complete coverage and proper structure.

Visual Distribution Analysis:

- **Histograms:** Plotted frequency distributions for each indicator to examine their spread and detect potential outliers.
- **Density Plots:** Created smoothed distribution curves to better understand the shape of each variable, including skewness and multimodality.
- **Comparative Visualization:** Maintained consistent figure sizes, colors, and plotting settings across all visualizations for fair comparison.

Rationale: We used summary statistics to understand the central values and spread of each indicator. Grouping by Region and Status helped reveal differences across geographic and political contexts. Histograms and density plots highlighted features like skewness, multiple peaks, and outliers. Keeping visualization settings consistent made sure that what we saw in the plots reflected the data itself, not the plotting style.

Reproducibility: We carefully documented every step, from computing statistics to creating plots. This means anyone using the same dataset and code can reproduce the same results, providing a solid and reliable foundation for further analysis.

3. Descriptive Analysis of Enriched Data

Descriptive Statistics and Distribution Analysis

The enriched dataset now includes 2,156 observations across 47 variables covering 196 countries, giving us a strong foundation for analysis. There are no missing values for the new indicators, which ensures the data is consistent and reliable. These indicators are Internet_Value (internet penetration rate), Female_LFPR (female labor force participation rate), and Tertiary_Enroll_Rate (tertiary education

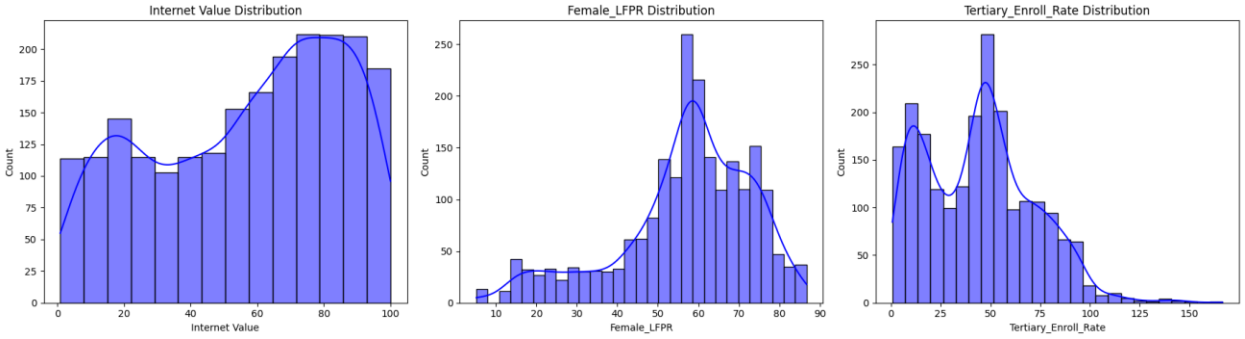
enrollment rate). Table 4 presents their main descriptive statistics, providing a first look at global patterns and variation across countries.

Table 4: Descriptive Statistics for the new indicators

Indicator	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Internet Value	2156	56.94	28.40	0.90	32.68	62.05	81.23	100.00
Female_LFPR	2156	56.83	16.70	5.13	50.06	58.91	68.80	86.72
Tertiary_Enroll_Rate	2156	44.05	27.18	0.66	19.38	45.16	61.12	166.67

Statistical Summary Interpretation

- **Internet Value:** The mean is 56.94 and the median is 62.05. The interquartile range spans from 32.68 to 81.23, showing the central half of countries. The minimum value of 0.90 highlights countries with very limited internet access.
- **Female Labor Force Participation Rate (Female_LFPR):** The mean is 56.83% and the median is 58.91%, with an interquartile range from 50.06% to 68.80%. The overall range and standard deviation (16.70) indicate substantial global variation.
- **Tertiary Education Enrollment Rate (Teritiary_Enroll_Rate):** The mean is 44.05% and the median is 45.16%. With a standard deviation of 27.18% and a maximum of 166.67%, the data shows high dispersion and some extreme outliers.



(a) (b) (c)

Figure 3: Histograms of (a) Internet Value, (b) Female Labor Force Participation, and (c) Tertiary Education Enrollment Rate

Distribution Insights from Histograms

- **Internet Value:** The majority of countries have high internet access, concentrated between roughly 60 and 100. There are fewer countries with low access, forming a long tail toward 0, which clearly highlights the global digital divide.
- **Female Labor Force Participation Rate (Female_LFPR):** The histogram shows a bimodal pattern. One smaller peak appears around 25%, while the main peak is between 50 and 65%. This indicates that countries differ widely in women's workforce participation.
- **Tertiary Enrollment Rate:** Most countries have enrollment rates under 75%, with a few extreme cases going above 100%. This creates a pronounced right-skew, showing that the global average is influenced by a small number of exceptional countries.

Methodologies

1. Descriptive Analysis

Principle: Summarize and characterize the data to gain a clear understanding of its central tendency, variability, and overall distribution before applying more advanced analyses. This step ensures that potential patterns, trends, and anomalies are identified early, guiding subsequent modeling and interpretation.

Theory & Formulas:

- **Mean:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

  Where: $\bar{x} \rightarrow$ Sample mean, $x_i \rightarrow$ Individual observation, $n \rightarrow$ Sample size

- **Standard deviation:** $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

  Where: $s \rightarrow$ Sample standard deviation, $x_i \rightarrow$ Individual observation, $\bar{x} \rightarrow$ Sample mean and $n \rightarrow$ Sample size

- **Median, Kernel Density Estimate**: $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$

  Where: $\hat{f}(x) \rightarrow$ Estimated density at point x, $n \rightarrow$ Sample size, $x_i \rightarrow$ Observed data points, $h \rightarrow$ Bandwidth (smoothing parameter), $K \rightarrow$ Kernel function (e.g., Gaussian)

Pros: Provides a clear summary of central tendency, dispersion, and distribution; identifies outliers and anomalies; simple and interpretable, forming the foundation for subsequent modeling. For additional insight, robust statistics or visual EDA (boxplots, histograms, density plots) can be used.

Cons: Descriptive stats alone cannot reveal causal relationships or complex interactions; may oversimplify data, hiding subtle patterns or nonlinear effects; sensitive to extreme values if robust measures (median, KDE) are not used, which can be mitigated by applying robust measures, non-parametric summaries, or visualization-based exploratory analysis.

2. Correlation Analysis

3.1 Pearson Correlation

Principle: Quantify the strength and direction of a linear relationship between two continuous variables. This measure helps identify whether increases in one variable are associated with increases or decreases in another, providing insight into potential associations before more complex modeling.

Theory & Formulas: $r_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$

Where: $r_{XY} \rightarrow$ Pearson correlation between X and Y

$\mathrm{Cov}(X,Y) \rightarrow$ Covariance between X and Y, $\sigma_X, \sigma_Y \rightarrow$ Standard deviations of X and Y

Pros: Simple, intuitive, and widely used for linear associations. For non-linear relationships or presence of outliers, Spearman or Kendall correlations, or robust measures, can be used to improve insights.

Cons: Only captures linear relationships and is sensitive to outliers. Using rank-based correlations or data transformations can mitigate these issues and better capture non-linear patterns.

2.2 Partial Correlation

Principle: Measure the direct relationship between two variables while removing the effect of one or more additional variables. This allows identification of unique associations that are not confounded by other predictors.

Theory & Formulas: $r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$

Where: $r_{XY \cdot Z} \rightarrow$ Partial correlation between X and Y controlling for Z

$r_{XY} \rightarrow$ Pearson correlation between X and Y

$r_{XZ} \rightarrow$ Pearson correlation between X and Z

$r_{YZ} \rightarrow$ Pearson correlation between Y and Z

Pros: Isolates direct relationships and identifies unique contributions. For non-linear effects, non-parametric or semi-partial correlation methods can provide complementary insights.

Cons: Assumes linear relationships and may miss complex interactions. Transformations or non-parametric alternatives can reduce these limitations.

3. Regression Modeling & Advanced Techniques

3.1 Ordinary Least Squares (OLS)

Principle: Estimate linear relationships between a dependent variable and one or more independent variables by fitting the best linear equation that minimizes the sum of squared errors. This forms the foundation for understanding associations and making predictions.

Theory & Formulas: $y = \beta_0 + \beta_1 x + \epsilon, \quad \hat{\beta} = (X^T X)^{-1} X^T y$

Where: $y \rightarrow$ Dependent variable, $x \rightarrow$ Independent variable, $\beta_0, \beta_1 \rightarrow$ Model coefficients,

$\hat{\beta} \rightarrow$ Estimated coefficient vector

$X \rightarrow$ Design matrix of independent variables and $\epsilon \rightarrow$ Error term

Pros: Simple, interpretable, and widely used for inference, providing clear estimates of relationships between variables. Alternatives like robust regression or regularization methods (Ridge, Lasso) can be applied to mitigate the effects of outliers or multicollinearity.

Cons: Sensitive to outliers and influential points; multicollinearity among predictors can distort coefficient estimates. These issues can be addressed by using robust regression techniques, variance inflation factor (VIF) checks, or feature selection.

3.2 Regression with Interaction Terms

Principle: Examine whether the effect of one independent variable depends on the level of another variable, allowing the model to capture heterogeneous effects across groups.

Theory & Formulas:

$$D3 = \beta_0 + \beta_1 \text{Enroll} + \beta_2 \text{Status} + \beta_3 (\text{Enroll} \times \text{Status}) + \gamma \text{Country} + \delta \text{Edition} + \epsilon$$

Pros: Captures varying effects across different groups or conditions and reveals moderation effects that simple linear models cannot detect. Alternatives like hierarchical or mixed-effects models can provide additional insight when interactions are complex or nested.

Cons: Model complexity increases with multiple interactions, and interpretation becomes more challenging as interactions multiply. Careful visualization, standardization of variables, or using marginal effects plots can help mitigate interpretability issues.

3.3 Mixed-Effects Model (Random Intercepts by Country)

Principle: Account for repeated measurements and group-level variability by including both fixed and random effects, allowing more accurate inference in hierarchical or panel data.

Theory & Formulas:

$$G_{it} = \beta_0 + \beta_1 LFPR_{it} + \beta_2 Status_i + \beta_3 PR\_CL_{it} + \beta_4 Edition_t + u_i + \epsilon_{it}, \quad u_i \sim N(0, \sigma_u^2)$$

Pros: Models both fixed and random effects, capturing hierarchical or nested data structures, and accounts for correlations within groups. Mixed-effects models are especially useful when repeated measurements exist. Alternatives like generalized estimating equations (GEE) or hierarchical Bayesian models can also address group-level correlations.

Cons: More complex and computationally intensive than OLS, and requires careful specification of random effects. Mis-specification can bias results. These issues can be mitigated by model diagnostics, likelihood ratio tests, or comparing models with different random effect structures.

3.4 Clustered Standard Errors

Principle: Adjust standard errors to account for intra-group correlations or heteroskedasticity, improving the robustness of inference in grouped data.

Theory & Formulas: $\widehat{V_{\text{cluster}}} = (X^T X)^{-1} \sum_g X_g^T \widehat{\epsilon_g} \widehat{\epsilon_g^T} X_g (X^T X)^{-1}$

Where: $\widehat{V_{\text{cluster}}} \rightarrow Cluster - robust\ variance - covariance\ matrix$

$X \rightarrow Design\ matrix\ of\ independent\ variables$

$X_g \rightarrow Submatrix\ of\ X\ for\ cluster\ g,\ \widehat{\epsilon_g} \rightarrow Vector\ of\ residuals\ for\ cluster\ g,$

$T \rightarrow Transpose\ operator$ and $(X^T X)^{-1} \rightarrow Inverse\ of\ X^T X$

Pros: Produces robust inference under grouped correlations and is particularly useful in panel or clustered datasets. Alternatives like multi-level modeling or bootstrapped standard errors can also account for intra-group correlations.

Cons: Increases computational complexity and only adjusts standard errors; coefficient estimates remain unchanged. Careful specification of clusters is required, and sensitivity analyses or alternative clustering methods can mitigate potential mis-specification issues.

4. Model Diagnostics & Validation

4.1 Residual Analysis

Principle: Evaluate model assumptions and identify potential issues such as misfit, heteroskedasticity, or multicollinearity. Residual analysis ensures that the model accurately represents the data and guides improvements.

Theory & Formulas:

- Residuals: $e_i = y_i - \hat{y}_i$
- Root Mean Squared Error (RMSE): $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$
- Mean Absolute Error (MAE): $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i|$
- Variance Inflation Factor (VIF): $\text{VIF}_j = \frac{1}{1-R_j^2}$

Where: $e_i \rightarrow Residual\ for\ observation\ i,\ y_i \rightarrow Observedvalue,$

$n \rightarrow Number\ of\ observations,\ \hat{y}_i \rightarrow Predictedvalue,$

$R_j^2 \rightarrow R - squaredfromregressingpredictorjonallotherpredictors$

Pros: Highlights potential problems in model fit, helping detect outliers, influential points, and multicollinearity. Alternative diagnostic tools like leverage plots, Cook's distance, or standardized residual plots can provide additional insights and validate assumptions.

Cons: Interpretation requires experience and statistical knowledge, and residual analysis alone does not provide direct solutions, only diagnostic insights. Sensitivity analyses or model refinements can mitigate detected issues.

4.2 Bootstrap Confidence Intervals

Principle: Assess the uncertainty of model coefficients, particularly in small samples or complex models, by resampling the data and estimating variability.

Theory & Formulas:

Bootstrap Confidence Interval (CI): $95\backslash\%\ \text{CI} = \text{percentiles of } \{\widehat{\beta_1^*}, \widehat{\beta_2^*}, \ldots, \widehat{\beta_B^*}\}$

Where: $\widehat{\beta_b^*} \rightarrow Estimated\ coefficient\ from\ the\ b-th\ bootstrap\ sample$

and $B \rightarrow Total\ number\ of\ bootstrap\ sample\ s$

Pros: Provides robust confidence intervals even for small or non-normal datasets and does not rely on strict parametric assumptions. Alternatives like jackknife resampling or Bayesian credible intervals can be used for additional robustness.

Cons: Computationally intensive, especially for large datasets or many predictors, and interpretation may be complex if the bootstrap distribution is skewed. Increasing the number of resamples or using parallel computing can help mitigate computational challenges.

5. Advanced Machine Learning

5.1 Random Forest Regression

Principle: Predict outcomes using an ensemble of decision trees, capturing non-linear relationships and complex interactions among features. Random Forest improves prediction accuracy and reduces overfitting compared to a single decision tree.

Theory & Formulas: Random Forest Prediction: $\hat{y} = \frac{1}{n}\sum_{i=1}^{n} T_i(X)$

Where: $\hat{y} \rightarrow Predicted\ out\ come,\ n \rightarrow Number\ of\ trees\ in\ the\ forest,$

$T_i(X) \rightarrow Prediction\ from\ the\ i-th\ decision\ tree$

Pros: Random Forest effectively handles complex, non-linear patterns and outliers, while reducing overfitting compared to single decision trees. Alternative ensemble methods such as Gradient Boosting or Extra Trees can offer comparable predictive power and additional flexibility in capturing interactions.

Cons: The model is less interpretable than linear models, making it harder to explain individual predictions. It can also be computationally intensive for large datasets. Techniques like feature importance analysis, partial dependence plots, or dimensionality reduction can help mitigate interpretability and computational challenges.

5.2 Permutation Feature Importance

Principle: Quantify the contribution of each predictor to model performance by measuring the decrease in accuracy when the feature values are randomly shuffled.

Theory & Formulas:

Permutation Feature Importance: $I_j = \text{Original Score} - \text{Shuffled Score}$

Where: $I_j \rightarrow Importance\ of\ feature\ j$

$Original\ Score \rightarrow Model\ performance\ on\ original\ dataset$

$Shuffled\ Score \rightarrow Model\ performance\ after\ shuffling\ feature\ j$

Pros: Permutation feature importance identifies influential features, aiding model interpretation, and is compatible with any supervised learning model. Alternative approaches like SHAP values or LIME can provide more detailed, instance-level explanations and handle correlated features better.

Cons: It can be computationally expensive for large datasets and may be misleading if features are highly correlated. Mitigations include using approximate or grouped feature importance, applying it to decorrelated or orthogonalized features, or combining with other interpretability methods.

5.3 Partial Dependence Plots (PDP)

Principle: Visualize the marginal effect of a feature on the predicted outcome while averaging out the effects of other features. PDPs help interpret complex machine learning models.

Theory & Formulas:

Partial Dependence Plot: $PD(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_s, x_c^{(i)}\right)$

Where: $PD(x_s) \rightarrow Partial\ dependence\ of\ feature\ x_s, n \rightarrow Number\ of\ observations,$

$\hat{f}\left(x_s, x_c^{(i)}\right) \rightarrow Predicted\ outcome\ using\ feature\ x_s\ and\ other\ features\ x_c\ for\ observation\ i$

$x_c^{(i)} \rightarrow Values\ of\ other\ features\ for\ observation\ i$

Pros: PDPs provide intuitive visual interpretation of feature effects and are useful for understanding complex, non-linear models. Alternatives like Individual Conditional Expectation (ICE) plots can show instance-level effects and better reveal heterogeneous patterns.

Cons: PDPs can be misleading if features are correlated and only show average effects, which may not fully capture interactions. Mitigations include using ICE plots, decorrelating features, or combining PDPs with feature interaction analysis.

6. Prediction Computation

Principle: Compute predicted outcomes from the fitted model to evaluate model performance and interpret interactions between predictors. This allows understanding of how independent variables influence the dependent variable, accounting for fixed and random effects.

Theory & Formulas:

Predicted Values: $\widehat{y_i} = X_i\hat{\beta} + u_i$

Where: $\widehat{y_i} \rightarrow Predicted\ value\ for\ observation\ i, \hat{\beta} \rightarrow Estimated\ coefficient\ vector$

$X_i \rightarrow Row\ vector\ of\ independent\ variables\ for\ observation\ i$

$u_i \rightarrow Random/intercept\ term\ for\ observation\ i(if\ applicable)$

Pros: Computing predictions facilitates understanding of the model's behavior, clarifies interactions between variables, and provides a foundation for evaluating accuracy in subsequent analyses. Alternatives like cross-validated predictions or simulation-based approaches can offer more robust insight into predictive performance.

Cons: Interpretation can be challenging for models with multiple predictors or random effects. Predictions alone do not reveal group-specific patterns, so additional analyses or visualizations (e.g., residual plots, predicted vs. observed plots) are needed to fully understand model behavior.

Experiments

1. Assessing the Impact of Internet Penetration on Freedom of Expression

This experiment examines how Internet penetration relates to freedom of expression across countries, using Civil Liberties (CL) ratings as a measure. We performed exploratory data analysis and regression modeling on the enriched dataset from the Data Enrichment section to determine whether higher Internet penetration is linked to greater freedom of expression.

1.1 Data Preparation and Exploration

Objective: Prepare a clean, high-quality dataset suitable for modeling the relationship between Internet penetration and Civil Liberties (CL rating).

Process & Parameters:

- **Dataset Selection:** We used the enriched dataset, which includes Internet Value, Female_LFPR, Tertiary_Enroll_Rate, and CL rating.
- **Column Selection:** CL rating was chosen as the dependent variable, and Internet Value as the main explanatory variable.
- **Data Cleaning:** Both columns were converted to numeric using pd.to_numeric(errors='coerce') to ensure proper use in regression. Rows with missing values were removed using dropna(), leaving a complete dataset.
- **Data Inspection:** We checked the data structure, types, and summary statistics using .info() and .describe() to verify everything was consistent.

Rationale: Converting variables to numeric ensures calculations run correctly. Removing incomplete rows prevents errors and bias, and focusing on the relevant columns keeps the analysis aligned with the research question.

Reproducibility: Using the same dataset and code ensures data preparation is identical across runs. The final sample size can be checked with len(data), guaranteeing consistent result.

1.2 Data Visualization and Relationship Analysis

Objective: Understand the relationship between Internet penetration and CL ratings, identify patterns or anomalies, and guide modeling.

Process & Parameters:

- **Distribution Plots:** Created histograms with kernel density estimates for Internet Value and CL rating. kde=True adds a smooth density curve, and color='skyblue' makes the bars visually distinct.
- **Scatter Plot with Regression Line:** Plotted CL rating against Internet Value with a regression line using sns.regplot(). ci=95 shows the confidence interval, and scatter_kws={'alpha':0.6} reduces overplotting.
- **Correlation Analysis:** Calculated the Pearson correlation coefficient using (data[predictor].corr(data[target])) to quantify the linear association and displayed it in a heatmap using sns.heatmap(). Parameters annot=True, fmt='.2f', cmap='coolwarm', vmin=-1, vmax=1 make the values clear and easy to interpret.

Rationale: Histograms and KDEs help detect skewness, outliers, or multiple peaks. Scatter plots with regression lines reveal potential linear trends, and Pearson correlation quantifies the strength and direction of the relationship, supporting preliminary insights.

Reproducibility: Using the same dataset and consistent plotting parameters ensures that all visualizations and correlations can be reproduced exactly.

1.3 Linear and Nonlinear Modeling

1.3.1 Ordinary Least Squares (OLS) Regression

Objective: Estimate the linear relationship between Internet penetration and Civil Liberties (CL) ratings.

Model Specification: $CL = \beta_0 + \beta_1 \times \text{Internet Penetration} + \epsilon$

Process & Parameters:

- Conducted OLS regression using statsmodels with robust standard errors (cov_type='HC3') to account for heteroskedasticity.
- Applied non-parametric bootstrap sampling (B=2000, random_state=123) to compute a 95% confidence interval for the coefficient of Internet penetration.

Rationale: OLS offers a clear and interpretable estimate of both the size and direction of the association. Using bootstrapping further strengthens the analysis by making the inference more robust and reliable, particularly when the normality assumptions might not hold.

Reproducibility: Reproducibility was ensured by fixing the random seed (123) for bootstrapping, which guarantees identical confidence intervals across runs. Additionally, using the same dataset along with consistent predictor and target selection, and applying the same statsmodels settings (cov_type='HC3'), ensures that the coefficient estimates are fully reproducible.

1.3.2 Random Forest Regression

Objective: Explore potential non-linear relationships between Internet penetration and CL ratings, and validate OLS assumptions.

Process & Parameters:

- **Dataset Split:** Divided the dataset into training (75%) and test (25%) sets using train_test_split with random_state=42 to ensure reproducibility.
- **Model Training:** Fitted RandomForestRegressor with n_estimators=500 and random_state=42.
- **Permutation Feature Importance:** Quantified the contribution of each predictor by measuring the decrease in model R^2 when each feature is randomly permuted.
  - **Model:** The already trained Random Forest model (rf).
  - **Test Set:** Features (X_test) and target (y_test) from the hold-out set.
  - **Number of Repeats (n_repeats):** 40 repetitions per feature to ensure stable estimates.
  - **Random State (random_state):** 42, to guarantee reproducibility of the permutation results.
- **Partial Dependence Plot (PDP):** Computed PDP for Internet penetration to visualize its effect on predicted CL ratings.

Rationale: Random Forest helps uncover non-linear patterns that linear regression might overlook. Partial dependence plots (PDPs) give a visual understanding of how Internet penetration influences predicted civil liberties ratings. Meanwhile, permutation importance quantifies each feature's predictive contribution, and by repeating the permutations, we obtain robust estimates of both feature importance and its variability.

Reproducibility: It was ensured by fixing random_state=42 and using consistent hyperparameters (n_estimators=500) ensures identical train/test splits, R^2 scores, PDPs, and permutation importance results. Consistent preprocessing, including numeric conversion and dropping missing values, guarantees reproducible outcomes.

2. Exploring the Relationship Between Female Labor Force Participation and Women's Rights

In this experiment, we analyzed whether higher female labor force participation (LFPR) is associated with stronger protection and recognition of women's rights across countries. The analysis combined descriptive statistics, correlation assessment, time-series visualization, and mixed-effects regression modeling using the enriched dataset.

2.1 Data Preparation and Exploratory Analysis

Objective: To investigate whether higher female labor force participation is associated with stronger protections for women's rights across countries through comprehensive data preparation, statistical summaries, and visual trend analysis.

Process & Parameters: We conducted an integrated analysis using the following approach:

Dataset Preparation:

- Loaded the Female Labor sheet from *Freedom World.xlsx* using pd.read_excel() to ensure structured access to the relevant data.
- Selected these key variables for analysis:
  - **Dependent variable:** Women's Rights score (G)
  - **Primary predictor:** Female Labor Force Participation Rate (Female_LFPR)
  - **Control variables / covariates:** Political Rights (PR rating), Civil Liberties (CL rating), Edition (Year), Status and Region categorical variables.
- Verified data types and structure using .info() and summary statistics with .describe() to detect potential anomalies or inconsistencies.

Descriptive Statistics:

- Calculated mean, median, standard deviation, minimum, and maximum for Female_LFPR and G using pandas.describe(). Explicitly selected numeric columns for accuracy.
- **Grouped Summaries:** Computed mean Female_LFPR and G by Region and Status to examine potential regional or political differences.

  Parameters: groupby(['Region', 'Status'])[['Female_LFPR','G']].mean().reset_index().
- **Variability & Outlier Check:** Assessed spread and range across groups to detect anomalies or extreme values before modeling.

Time-Series Visualization:

- Filtered data to focus on the 2013-2023 period for consistent temporal analysis.
- Generat Line plots were generated for both Female LFPR and Women's Rights (G) scores using seaborn.lineplot(). Distinct colors were used for each region to differentiate trends, with yearly data points marked (marker='o') for clarity. The linewidth was set to 1 and markersize to 6 for a clean look, and smoothing was disabled (errorbar=None) to reflect actual observed values.
- Adjusted plot aesthetics: grid lines (linestyle='--', alpha=0.5), rotated x-axis ticks for readability, and legend positioned outside the plot for clarity.

Rationale: Focusing on these variables keeps the analysis targeted and avoids unnecessary complexity. Descriptive statistics provide a clear picture of central tendency, spread, and variability, while grouped summaries highlight differences across regions or political freedom categories. Early detection of anomalies ensures reliable data. Line plots reveal longitudinal trends, with markers and color-coding allowing easy comparison across geographic areas.

Reproducibility: Using pandas ensures consistent descriptive statistics and dataset information with .describe() and .info(). Explicitly selecting numeric columns and grouping keys guarantees that all summaries are fully reproducible through groupby(). Matplotlib and seaborn visualizations are reproducible thanks to documented colors, markers, and figure size. Preprocessing steps, including filtering years and coding Region as a categorical variable, ensure that the input data for plotting is consistent. Altogether, these practices make both the analysis results and visual outputs fully reproducible.

2.2 Correlation and Partial Correlation Analysis

Objective: Quantify both raw and adjusted associations between Female Labor Force Participation Rate (Female_LFPR) and Women's Rights (G).

Process & Parameters:

- **Pearson Correlation:** Calculated pairwise correlations among numerical columns: Edition, Female_LFPR, G, PR rating, and CL rating. Missing values were dropped to ensure valid computations. Correlation coefficient and p-value were computed using .corr() or scipy.stats.pearsonr().
- **Partial Correlation:** Computed partial correlations controlling for potential confounders (PR rating, CL rating, and G) to isolate the direct relationship between Female_LFPR and G. Implemented using pingouin.partial_corr() with covar specifying confounders.

Rationale: The Pearson correlation offers a first look at the linear relationship between Female_LFPR and G. To account for potential confounding by political and civil liberties factors, partial correlation was also calculated, isolating the independent effect of Female_LFPR on G and providing a clearer measure of its association.

Reproducibility: The correlation analysis is fully reproducible using the enriched dataset. Explicit column selection, consistent handling of missing values, and standardized specification of confounders ensure identical raw and partial correlation outputs. Using pingouin.pcorr() in Python (or ppcor::pcor() in R) guarantees consistent and reliable partial correlation results.

2.3 Mixed-Effects Regression Modeling

Objective: Formally assess the effect of Female Labor Force Participation Rate (Female_LFPR) on Women's Rights (G) while accounting for repeated measurements and country-level variability.

Model Specification: $G \sim \text{Female\_LFPR} \times \text{Status} + \text{PR\_CL\_Avg} + \text{Edition}$

Process & Parameters:

- **Dependent Variable:** G (Women's Rights score).
- **Fixed Effects:** Female_LFPR, Status, PR_CL_Avg (average of PR and CL ratings), Edition (year).
- **Interaction:** Female_LFPR × Status to examine whether the effect of LFPR differs across political freedom categories.
- **Random Effects:** Random intercepts by Country/Territory to account for unobserved country-level differences.
- **Estimation Method:** Restricted Maximum Likelihood (REML) for unbiased estimation of variance components.
- **Preprocessing Parameters:** PR_CL_Avg calculated as the average of PR and CL ratings: (df_exp2['PR rating'] + df_exp2['CL rating']) / 2

Rationale: Mixed-effects modeling is suitable for hierarchical data, where observations are nested within countries. Including an interaction term allows us to see if the effect of Female_LFPR varies depending on political freedom. REML ensures that variance components are estimated without bias. Alternative

approaches, such as standard OLS or models with random slopes, were considered but not used due to repeated measurements and interpretability concerns.

Reproducibility: The mixed-effects analysis is fully reproducible using statsmodels in Python (smf.mixedlm()) with the same dataset and formula. Explicit preprocessing, including converting categorical variables, casting numeric columns, and calculating PR_CL_Avg, ensures that the model input is consistent. Using the fixed formula and dataset guarantees identical parameter estimates, standard errors, and model summary outputs.

2.4 Model Validation and Interaction Analysis

Objective: Validate the mixed-effects model fit, check model assumptions, and visualize how Female Labor Force Participation (Female_LFPR) affects Women's Rights (G) across political freedom statuses.

Process & Parameters:

1. **Predictions and Residual Analysis:**
   - **Predicted Values and Residuals:** Calculated predicted Women's Rights (G) scores from the fitted mixed-effects model and computed residuals as the difference between observed and predicted values.
   - **Error Metrics:** Evaluated model performance using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to quantify prediction accuracy.
   - **Residual Diagnostics:**
     - Plotted histograms with kernel density estimate (KDE) to inspect the distribution of residuals.
     - Created Q–Q plots to assess normality of residuals.
     - Compared actual vs predicted G scores across observations using line plots.
   - **Multicollinearity Check:** Computed Variance Inflation Factor (VIF) for key predictors (Female_LFPR, PR_CL_Avg, plus intercept) to ensure absence of strong multicollinearity.
2. **Interaction Visualization:**
   - **Sampling LFPR Values:** Generated a sequence of Female_LFPR values spanning the observed range to examine predicted effects continuously.
   - **Holding Covariates Constant:** Fixed PR_CL_Avg at its mean and Edition (year) at its most frequent value to isolate the interaction effect.
   - **Predicted Curves:** Used the fitted mixed-effects model to compute predicted G scores for each Status category (Free, Partially Free, Not Free).
   - **Visualization:** Plotted predicted G scores against Female_LFPR for each Status using distinct colors and consistent line styles for clarity.

Rationale: These steps verify that model assumptions, such as normality and homoscedasticity of residuals, are satisfied. RMSE and MAE provide interpretable measures of prediction accuracy, while VIF ensures that fixed-effect predictors do not exhibit excessive multicollinearity, which could bias coefficients. Interaction plots offer an intuitive understanding of whether and how the effect of Female_LFPR on women's rights varies across political freedom categories, highlighting potential heterogeneity across countries.

Reproducibility: The analysis is fully reproducible using the same mixed-effects model object and enriched dataset. By fixing colors, markers, and LFPR sampling in plotting libraries like matplotlib and seaborn, the visualizations and results remain consistent. Metrics and plots produce identical outputs across runs.

3. Investigating the Role of Higher Education in Academic Freedom

In this experiment, we investigated the relationship between tertiary education levels and academic freedom across countries. Our analysis integrated careful data preparation, exploratory data analysis, correlation and partial correlation assessment, multidimensional visualization, regression modeling, and predictive interpretation. Each step is described in detail below, including the underlying reasoning and guidance for reproducibility.

3.1 Dataset Overview and Exploratory Data Analysis

Objective: To analyze how tertiary education enrollment rates relate to academic freedom across countries, combining data preparation, descriptive statistics, and exploratory visualizations.

Process & Parameters:

- **Dataset Loading:** Loaded the "Education" sheet from Freedom World.xlsx using pd.read_excel().
- **Column Selection:** Kept relevant columns: Country/Territory, Edition (year), Region, Status, PR rating, CL rating, D3 (academic freedom), and Tertiary_Enroll_Rate.
- **Data Inspection:** Verified data types and structure with .info(), and computed descriptive statistics using .describe().

Bar Plots – Tertiary Enrollment by Year and Region:

- Generated bar plots showing average Tertiary Enrollment Rate per year for each region.
- Parameters chosen for clarity and reproducibility: ci=None to display only mean values, palette='bright' for clear region differentiation, legend positioned outside the plot, x-axis labels rotated 45° for readability, and semi-transparent dashed y-axis gridlines to aid comparison of bar heights.

Box Plots – Tertiary Enrollment by Region & Status:

- Created box plots to examine the distribution of Tertiary Enrollment Rates across regions, separated by Freedom Status (Free, Partially Free, Not Free).
- Parameters chosen: The parameters included using `palette='mako'` for clear status-based coloring, rotating x-axis labels for better readability, and adding dashed semi-transparent y-axis gridlines to enhance variability assessment.
- These plots allow evaluation of within-region spread, between-region differences, and the effect of political freedom on enrollment distributions.

Rationale: These steps ensure a clean and well-structured dataset suitable for modeling. Categorical conversion supports proper grouping, interaction terms, and plotting. Bar plots provide a clear view of trends over time and across regions, helping identify periods of rapid change or stagnation. Box plots summarize variability and highlight extreme values or potential outliers without relying solely on numerical summaries. Together, these visualizations help detect structural patterns and inform subsequent correlation and regression modeling.

Reproducibility: Dataset preparation is fully reproducible using the enriched dataset. Consistent column selection, renaming, and data type conversions ensure identical outputs across runs. Visualizations with Seaborn and Matplotlib are reproducible through documented parameters such as colors, figure size, label rotation, and gridlines. Preprocessing steps, including filtering by year and converting Region and Status to categorical types, guarantee that all plots remain consistent and deterministic.

3.2 Correlation Analysis and Multidimensional Visualization

Objective: Quantify both pairwise (raw) and adjusted (partial) associations between tertiary education (Tertiary_Enroll_Rate) and academic freedom indicators (D3), while exploring multi-variable interactions with political ratings (PR and CL).

Process & Parameters:

Raw Correlations:

- Calculated Pearson correlations among all numeric variables: Tertiary_Enroll_Rate, D3, PR rating, CL rating.
- Numeric columns explicitly selected using select_dtypes(include='number') to ensure only continuous variables were included.
- Missing values were automatically dropped to guarantee valid correlation calculations.
- Output: Symmetric correlation matrix showing pairwise linear relationships between numeric variables.

Partial Correlations:

- Computed partial correlations using pingouin.pcorr() to control for confounding effects among variables.
- Pivoted the resulting partial correlation table into a square symmetric matrix for readability and consistency with the raw correlation matrix.
- Parameters: All numeric columns included; each correlation controls for the influence of all other numeric variables to isolate direct associations.

Multidimensional Visualization (Parallel Coordinates):

- **Plot Type:** Parallel coordinates plot using plotly.express.parallel_coordinates.
- **Dimensions:** Tertiary_Enroll_Rate, D3, PR rating, CL rating included to show multi-variable relationships along the same axes.
- **Color Mapping:** PR rating used to color lines with a sequential Viridis scale to highlight variation in political rights.
- **Layout Adjustments:** Layout adjustments included centering the title and increasing the top margin for better visibility, using a white background with larger fonts and proper margins to avoid axis label overlap, and setting the colorbar title to 'PR Rating' with outward ticks for improved clarity.
- **Interactivity:** Plotly ensures dynamic exploration of high-dimensional data, allowing users to hover and filter values along multiple axes.

Rationale: Pearson correlations provide a preliminary measure of linear relationships, while partial correlations isolate the independent effect of tertiary education on academic freedom (D3) by accounting for political ratings, reducing potential confounding bias. Pivoting the partial correlation matrix into a square symmetric format improves interpretability alongside raw correlations. Parallel coordinates plots allow multi-dimensional relationships to be visualized in a single figure, making patterns, clusters, and unusual combinations easier to identify compared to traditional 2D plots.

Reproducibility: Correlation analysis is fully reproducible using the cleaned numeric columns from the dataset. Fixed column selection and consistent handling of missing values ensure identical outputs, and pingouin.pcorr() guarantees consistent partial correlation results. Parallel coordinates visualizations are reproducible using the same columns, color mapping, and layout settings. Using Plotly Express with fixed parameters ensures consistent plots, maintaining identical visual appearance and interpretability across runs.

3.3 Regression Modeling and Interaction Analysis

Objective: Quantify the effect of tertiary education (Tertiary_Enroll_Rate) on academic freedom (D3), adjusting for country- and year-specific differences, and examine whether this effect varies across freedom statuses (Status).

Model Specification: $D3 \sim \text{Tertiary\_Enroll\_Rate} \times C(\text{Status}) + C(\text{Country}) + C(\text{Edition})$

- **Dependent Variable:** D3 (Academic Freedom score).
- **Independent Variables:** Tertiary_Enroll_Rate, Status, Edition.
- **Fixed Effects:** Country and Edition to account for unobserved heterogeneity across countries and years.
- **Interaction Term:** Tertiary_Enroll_Rate × Status to assess differential effects by freedom status.

Process & Parameters:

- Extracted relevant columns: D3, Tertiary_Enroll_Rate, Status, Country, Edition.
- Converted Edition to categorical to model year fixed effects.
- Fitted OLS regression using statsmodels.formula.api.ols.
- Interaction term included explicitly to capture status-dependent effects.
- Clustered standard errors by Country to account for within-country correlation across years.

Rationale: The interaction term allows us to determine whether the effect of tertiary education on academic freedom differs across political freedom categories. Including fixed effects for Country and Edition controls for unobserved heterogeneity at both country and year levels. Clustered standard errors ensure robust inference despite repeated measurements within countries.

Reproducibility: The regression analysis is fully reproducible using the cleaned dataset and statsmodels in Python. A fixed formula, consistent column selection, categorical conversion, and clustered standard error specification ensure identical coefficient estimates and standard errors.

3.4 Interaction Effect Interpretation

Objective: Interpret and visualize how tertiary education (Tertiary_Enroll_Rate) influences academic freedom (D3) differently across freedom statuses (Free, Partially Free, Not Free).

Process & Parameters:

- **Status-Specific Effects:**
  - Baseline: effect for Free countries
  - Interaction terms: Tertiary_Enroll_Rate:C(Status)[T.PF] → additional effect for Partially Free; Tertiary_Enroll_Rate:C(Status)[T.NF] → additional effect for Not Free
  - Computed status-specific effects by summing baseline with respective interaction term.
- **Predicted Academic Freedom Visualization:**
  - Generated 50 evenly spaced Tertiary_Enroll_Rate values for smooth prediction curves.
  - Held Country and Edition constant at their mode values to isolate education and status effects.
  - Predicted D3 values computed for each freedom status using the fitted regression model.
  - Visualization included KDE plots of actual D3 distributions by status (left panel) and line plots of predicted D3 against Tertiary_Enroll_Rate, color-coded by freedom status (right panel).
  - Bright, distinct colors, gridlines, and legends ensure interpretability and clarity.

Rationale: Status-specific coefficients clarify whether the impact of tertiary education is stronger, weaker, or similar depending on a country's freedom status. Combining KDE plots with predicted curves validates the model fit and reveals heterogeneity in academic freedom outcomes across different political contexts.

Reproducibility: Interaction effects and status-specific computations are fully reproducible using the fitted regression model, with consistent column selection, categorical conversions, and interaction terms. Predicted academic freedom visualizations are reproducible using fixed covariates, consistent color mapping, and identical enrollment sampling in Matplotlib or Seaborn.

Results

1. Empirical Findings on Internet Penetration and Freedom of Expression

We analyzed an enriched global dataset of 2,156 country-year observations to study how internet penetration relates to civil liberties protections. Civil liberties ratings (CL rating) are measured on a scale from 1 (most free) to 7 (least free). Using a combination of descriptive statistics, exploratory data analysis, and correlation measures, we consistently found a negative association, highlighting key patterns in how greater digital connectivity generally corresponds to lower civil liberties in certain contexts.

1.1 Descriptive Statistics and Data Distribution

Descriptive statistics for the main variables are presented in Table 5. Civil liberties ratings (CL) had a mean of 3.36 (SD = 1.91) and a median of 3.00, with the middle 50% of countries scoring between 2.00 and 5.00, showing substantial variation. Internet penetration rates (Internet Value) averaged 56.94% (SD = 28.40) with a median of 62.05%, and the interquartile range of 32.68%–81.23% highlights considerable differences in digital access worldwide.

Table 5: Descriptive Statistics of Study Variables

Variable	Count	Mean	Std. Dev.	Minimum	25%	Median	75%	Maximum
CL Rating	2156.00	3.36	1.91	1.00	2.00	3.00	5.00	7.00
Internet Value	2156.00	56.94	28.39	0.90	32.68	62.05	81.23	100.00

The distribution of internet penetration rates (Figure 4, left panel) was approximately normal with slight negative skewness, indicating that while many countries have achieved moderate to high access (cluster around 60–80%), a substantial number remain below 40%. The distribution of civil liberties ratings (Figure 4, right panel) revealed a pronounced bimodal pattern, with concentrations at both the free end (scores 1–2) and the restrictive end (scores 6–7), suggesting polarization in freedom levels and fewer cases in the middle range.

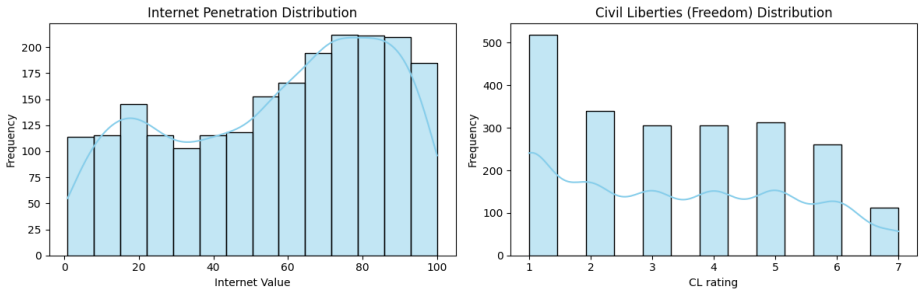


Figure 4: Histograms of Internet Penetration and Civil Liberties Distributions

[Note: Left panel shows the distribution of internet access rates (Internet Value), ranging from 0.9% to 100% penetration. Right panel shows the distribution of civil liberties ratings (CL rating) with bimodal clustering at both ends of the scale: scores 1–2 (most free) and scores 6–7 (least free)]

1.2 Exploratory Relationship Analysis

The relationship between internet penetration and civil liberties was examined through bivariate analysis. Figure 5 presents a scatter plot with a linear regression line illustrating the association between internet penetration rates and civil liberties ratings.

The scatter plot reveals a clear negative relationship between the two variables. The downward slope of the regression line indicates that higher Internet Values systematically correspond to lower CL ratings. While the data points show some dispersion around the trend line, the overall pattern demonstrates a consistent inverse relationship. Notably, observations with Internet Values below 40 predominantly show higher CL ratings (above 4), whereas observations with Internet Values above 60 mainly correspond to CL ratings below 3. This suggests that countries with low internet penetration tend to have more restrictive civil liberties, while higher internet penetration is observed mostly in freer countries.

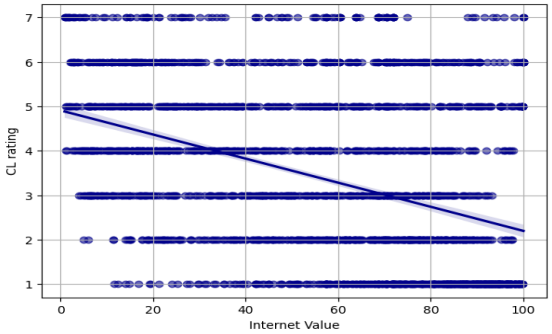


Figure 5: Scatter plot of CL rating vs. Internet Value (with linear fit)

[Note: The scatter plot displays internet penetration (x-axis, 0–100%) and civil liberties ratings (y-axis, 1–7 scale). The blue line represents the linear regression fit with a 95% confidence interval.]

The Pearson correlation analysis confirmed this relationship, yielding a coefficient of r = -0.404, indicating a moderate negative association between Internet Values and CL ratings. In other words, countries with higher internet penetration tend to have lower CL ratings, suggesting that greater internet access is generally associated with freer societies. Although the correlation is statistically significant, the moderate strength indicates that additional factors beyond Internet Value also influence civil liberties levels.

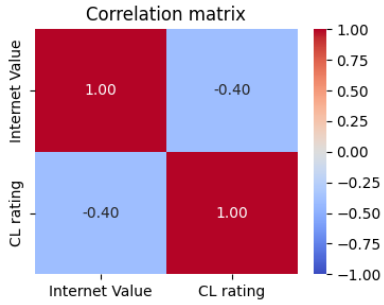


Figure 6: Correlation matrix

[Note: The heatmap illustrates the correlation between CL rating and Internet Value, showing a coefficient of -0.40. The color intensity represents the strength and direction of the correlation, with blue tones indicating negative relationships (Coefficient rounded from -0.404.)]

1.3 OLS Regression & Bootstrap Validation

An ordinary least squares (OLS) regression was conducted to quantify the relationship between internet penetration and civil liberties ratings (CL rating). The results revealed a statistically significant negative relationship, indicating that higher internet penetration is associated with lower civil liberties protections across countries. The regression model was highly significant (F-statistic = 449.3, $p < 0.001$) and explained approximately 16.3% of the variance in civil liberties ratings.

Table 6: OLS Regression Results: Internet Value Predicting CL Rating

Predictor	Coefficient	Robust SE	z-value	p-value	95% CI
Intercept	4.9069	0.075	65.03	<0.001	[4.759, 5.055]
Internet Penetration	-0.0271	0.001	-21.20	<0.001	[-0.030, -0.025]

The analysis yielded several key findings

- **Significant Negative Coefficient**: A one percentage point increase in internet access is associated with a 0.0271 unit decrease in civil liberties ratings on the 7-point scale.
- **Precise Estimation**: The narrow 95% confidence interval [-0.030, -0.025] demonstrates precise estimation of this negative relationship, with the interval excluding zero confirming statistical significance.
- **Substantial Explanatory Power**: The model explains 16.3% of the variance in civil liberties (R^2 = 0.163), indicating internet penetration is an important predictor of freedom levels.
- **Large Effect Size**: The z-value of -21.197 indicates a strong statistical relationship between the variables, far exceeding conventional thresholds for significance.
- **Model Reliability**: The predicted CL rating for zero internet penetration is 4.9069, though this should be interpreted cautiously as an extrapolation beyond observed data.

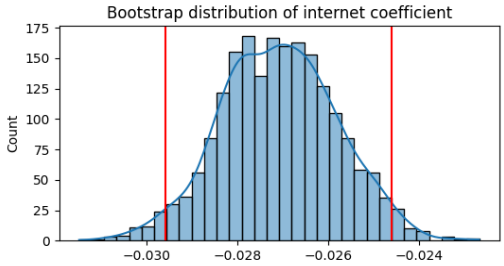


Figure 7: Bootstrap Distribution of Internet Penetration Coefficient

[Note: The figure shows the bootstrap distribution of the internet penetration coefficient based on 2,000 resamples. Red vertical lines indicate the 2.5th and 97.5th percentiles, forming the 95% confidence interval [-0.0296, -0.0246]. The symmetric distribution around the mean estimate (-0.0271) confirms the stability of the negative association.]

Bootstrap validation confirmed the robustness of these results. The nearly identical 95% confidence interval and the absence of positive coefficients in the distribution provide strong evidence that the negative relationship between internet penetration and civil liberties is genuine and not a statistical artifact.

While the relationship is statistically robust, diagnostic tests revealed non-normal residuals and positive autocorrelation (Durbin-Watson = 0.257), suggesting potential model misspecification or omitted variables. These findings indicate that, although internet penetration is strongly associated with civil liberties, additional factors likely influence freedom levels beyond what is captured in this bivariate model.

1.4 Random Forest Model & Partial Dependence

To capture potential non-linear relationships between internet penetration and civil liberties ratings (CL rating), a Random Forest regression was employed. While the OLS model confirmed a general negative association, the Random Forest analysis uncovers more complex patterns that vary across different levels of internet penetration.

Table 7: Random Forest Model Performance Metrics

Dataset	R-squared	Interpretation
Training	0.556	High explanatory power for in-sample data
Testing	0.051	Minimal predictive power for out-of-sample data

The substantial discrepancy between training ($R^2 = 0.556$) and testing ($R^2 = 0.051$) performance indicates that, although internet penetration explains variation in civil liberties within the sample, this relationship does not generalize well to new data. This suggests the presence of complex, context-specific interactions rather than a consistent predictive relationship.

Key Findings:

- **Permutation Importance**: Internet penetration had a high permutation importance (mean = 0.509, std = 0.054), confirming it remains a significant predictor despite generalization challenges.
- **Non-linear Relationship**: Partial dependence analysis reveals a multi-phase relationship not captured by linear models, explaining the Random Forest's superior training performance.

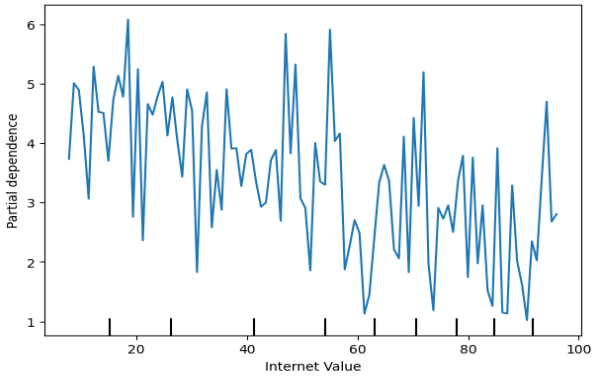


Figure 8: Partial Dependence of Predicted CL Rating on Internet Value

[Note: Partial dependence plot showing the marginal effect of internet penetration on predicted CL ratings. Three phases are observed: stability (0–20% penetration), decline (20–60%), and partial recovery (60–100%)]

The partial dependence analysis reveals three distinct phases in the relationship:

- **Initial Stability Phase (0-20% penetration)**: Civil liberties ratings remain relatively constant, suggesting minimal impact during early stages of internet adoption.
- **Progressive Decline Phase (20-60% penetration)**: Increasing internet access corresponds with a strong negative relationship, reaching the lowest predicted CL ratings around 60% penetration.
- **Partial Recovery Phase (60-100% penetration)**: Beyond 60% penetration, the relationship stabilizes and shows modest improvement, although predictions remain below initial levels.

This non-linear, multi-phase pattern indicates that the relationship between internet penetration and civil liberties evolves through different stages of digital adoption. The contrast between high training performance and low testing performance highlights that, while internet access is an important predictor, the effect is mediated by country-specific factors and does not generalize uniformly across contexts. Nevertheless, the high permutation importance confirms the substantive relevance of internet penetration within broader socio-political settings.

1.5 Synthesis of Findings

Consistent Negative Association: All analyses (correlation, OLS, bootstrap, Random Forest) indicate higher internet penetration is linked to lower civil liberties.

Statistical Robustness: OLS coefficient (-0.027, $p < 0.001$) and bootstrap 95% CI [-0.030, -0.025] confirm the reliability of the effect.

Explanatory Power: OLS explains a moderate portion of variance ($R^2 = 0.163$), while Random Forest highlights context-dependent, non-linear effects.

Variable Importance: Random Forest permutation importance (0.51) underscores the substantive relevance of internet penetration in predicting civil liberties.

Methodological Convergence: Results across multiple approaches consistently support the negative association, though country-specific factors mediate the relationship.

2. Empirical Results: Correlation Between Women's Workforce Participation and Women's Rights

We analyzed female labor force participation (LFPR) and women's rights scores (G) across 2,156 country-year observations. LFPR is measured as the percentage of women aged 15–64 engaged in formal employment, while women's rights scores range from 0 (lowest rights) to 16 (highest rights). Using descriptive statistics, exploratory data analysis, and correlation measures, we examined global patterns and the association between women's economic participation and legal/social rights protections.

2.1 Exploratory Data Analysis

We first examined female labor force participation (LFPR) and women's rights scores to understand global patterns and variation. Descriptive statistics for the main variables are presented in Table 8. LFPR averaged 56.8% (SD = 16.7), ranging from minimal participation (5.1%) to near-universal engagement (86.7%). Women's rights scores averaged 9.5 (SD = 4.1), indicating considerable disparities in gender equality protections across countries.

Table 8: Descriptive Statistics for Female LFPR and Women's Rights Score (G)

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Female_LFPR	2156	56.83	16.70	5.13	50.06	58.91	68.80	86.72
G	2156	9.50	4.12	0.00	6.00	10.00	13.00	16.00

Regional and Political Status Patterns

Analyzing the data by region and political freedom status (Table 9) revealed several notable patterns:

- **Free Countries:** Consistently exhibit the highest women's rights scores, particularly in Europe (G = 14.5) and the Americas (G = 13.0), alongside moderate to high female LFPR (Europe: 68.5%, Americas: 60.7%).
- **Not Free Countries:** Show the lowest women's rights scores, with severe restrictions in the Middle East (G = 4.0) and Africa (G = 4.0), coupled with dramatically reduced female LFPR (Middle East: 27.7%, Europe: 38.1%).
- **Partially Free Countries:** Occupy an intermediate position, with women's rights scores ranging from 6.7 (Middle East) to 9.9 (Europe), and LFPR from 36.8% (Middle East) to 64.2% (Eurasia).

Table 9: Summary of Averages by Region and Status

Region	Status	Avg. Female LFPR (%)	Avg. Women's Rights Score (G)
Africa	Free	51.8	10.8
	Partially Free	59.0	7.2
	Not Free	51.3	4.0
Americas	Free	60.7	13.0
	Partially Free	54.4	9.0
	Not Free	50.7	6.1
Asia	Free	58.9	12.8
	Partially Free	48.9	9.0
	Not Free	63.3	5.7
Europe	Free	68.5	14.5
	Partially Free	55.3	9.9
	Not Free	38.1	5.2
Middle East	Free	70.0	11.0
	Partially Free	36.8	6.7
	Not Free	27.7	4.0

Time-Series Trends (2013–2023)

- European nations maintained consistently high women's rights scores and female participation rates.
- Middle Eastern countries showed improving but still limited female labor force participation despite low women's rights scores.
- Asian and African nations exhibited considerable variability in both metrics, suggesting heterogeneous development patterns.
- The Americas demonstrated stable, moderate-to-high performance in both women's rights and labor participation.

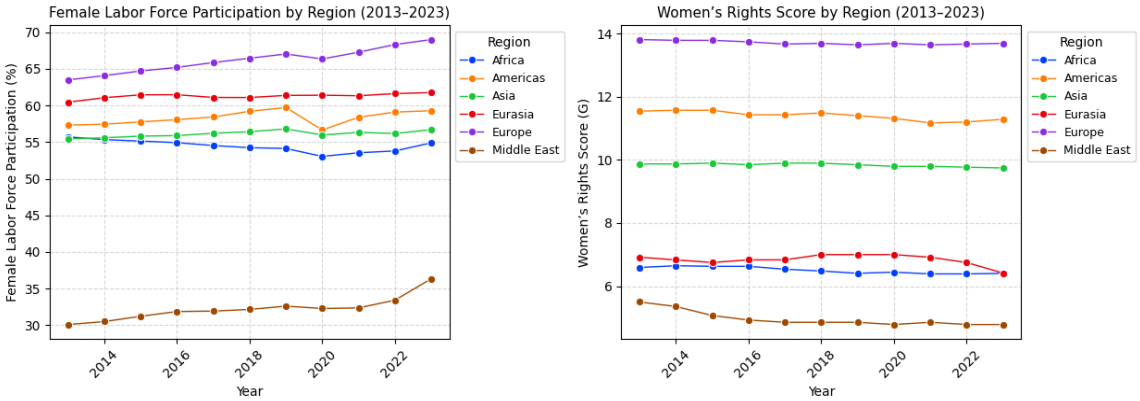


Figure 9: Regional Trends in Female LFPR and Women's Rights (2013-2023)

[Note: Side-by-side comparison of regional trends. Left panel shows female labor force participation rates, with Europe and the Middle East maintaining the highest rates. Right panel shows women's rights scores, with Europe consistently scoring highest and other regions showing varying trajectories. All regions demonstrate stability rather than dramatic changes over the decade.]

These patterns suggest a complex relationship between formal women's rights protections and economic participation, with some regions showing strong alignment, while others demonstrate apparent paradoxes where participation rates either exceed or fall short of what might be expected based on legal rights alone.

2.2 Correlation & Partial Correlation Analysis

We examined the relationships among women's rights (G), female labor force participation (LFPR), political rights (PR rating), and civil liberties (CL rating) using both raw and partial correlations to understand simple associations as well as relationships controlling for other factors.

Key Findings from Raw Correlations:

- Women's rights (G) show a strong positive correlation with female LFPR (r = 0.43), suggesting that stronger legal protections for women are associated with higher economic participation.
- Political rights (PR rating) and civil liberties (CL rating) display strong negative correlations with women's rights (r = -0.89 and r = -0.95, respectively), indicating that higher freedom ratings correspond to better women's rights protections.
- Female LFPR has moderate negative correlations with both political rights (r = -0.34) and civil liberties (r = -0.38).

Table 10: Raw Correlation Matrix

	Edition	PR rating	CL rating	G	Female_LFPR
Edition	**1.00**	0.04	0.03	-0.02	0.03
PR rating	0.04	**1.00**	0.94	-0.89	-0.34
CL rating	0.03	0.94	**1.00**	-0.95	-0.38
G	-0.02	-0.89	-0.95	**1.00**	0.43
Female_LFPR	0.03	-0.34	-0.38	0.43	**1.00**

Partial Correlation Insights:

After controlling for all other variables in the model, several key relationships emerged:

- The positive association between women's rights (G) and female LFPR persists (r = 0.23), though weaker.
- The strong negative relationship between civil liberties and women's rights remains significant (r = -0.68).
- The association between political rights and women's rights becomes negligible (r = -0.03).
- The correlation between civil liberties and female LFPR disappears (r = 0.00).

Table 11: Partial Correlation Matrix (Controlling for All Other Variables)

	Edition	PR rating	CL rating	G	Female_LFPR
Edition	**1.00**	0.04	0.00	0.03	0.03
PR rating	0.04	**1.00**	0.66	-0.03	0.08
CL rating	0.00	0.66	**1.00**	-0.68	0.00
G	0.03	-0.03	-0.68	**1.00**	0.23
Female_LFPR	0.03	0.08	0.00	0.23	**1.00**

The partial correlation analysis shows that the relationship between women's rights and female labor force participation remains meaningful even after accounting for political freedoms, civil liberties, and temporal trends. The correlation decreases from 0.43 to 0.23, suggesting that other factors such as cultural norms, economic structures, and educational opportunities influence this relationship. The strong negative partial correlation between civil liberties and women's rights (r = -0.68) indicates that protections for women's rights are more closely connected to the broader civil liberties framework than to political rights alone.

2.3 Mixed Effects Model

A mixed-effects model was employed to examine the relationship between women's rights scores (G) and female labor force participation, while accounting for country-level random effects and controlling for political context and temporal trends. The model explained substantial variance in women's rights protections and revealed several important patterns.

Key Findings from the Mixed-Effects Model:

- **Female labor force participation** showed a statistically significant positive relationship with women's rights scores ($\beta = 0.017$, $p = 0.003$). This indicates that a one percentage point increase in female labor participation is associated with a 0.017 unit increase in women's rights scores after controlling for other factors.
- **Political and Civil Liberties** (PR_CL_Avg) demonstrated a strong negative relationship with women's rights ($\beta = -0.816$, $p < 0.001$), consistent with the coding where higher freedom ratings correspond to better women's rights protections.
- **Status Effects** revealed that Not Free countries had significantly lower women's rights scores ($\beta = -0.766$, $p = 0.024$) compared to Free countries, while Partially Free countries did not differ significantly from Free countries.
- **Interaction Effects** indicated that the relationship between female labor participation and women's rights was stronger in Not Free countries ($\beta = 0.017$, $p = 0.004$) compared to Free countries.
- **Temporal Trend** showed a small but statistically significant decline in women's rights scores over time ($\beta = -0.009$, $p = 0.013$), equivalent to a 0.09 unit decrease per decade.

Table 12: Mixed-Effects Model Predicting Women's Rights Score (G)

Predictor	Coefficient	Std. Error	z-value	p-value	95% CI
Intercept	28.932	6.990	4.139	<0.001	[15.232, 42.632]
Status [NF]	-0.766	0.338	-2.265	0.024	[-1.429, -0.103]
Status [PF]	-0.218	0.279	-0.783	0.434	[-0.764, 0.328]
Female_LFPR	0.017	0.006	2.924	0.003	[0.005, 0.028]
Female_LFPR × Status [NF]	0.017	0.006	2.849	0.004	[0.005, 0.028]
Female_LFPR × Status [PF]	0.007	0.005	1.437	0.151	[-0.002, 0.016]
PR_CL_Avg	-0.816	0.046	-17.756	<0.001	[-0.906, -0.726]
Edition	-0.009	0.004	-2.490	0.013	—

The model showed good convergence and explained substantial between-country variance (Group Variance = 6.459). Variance Inflation Factors indicated acceptable multicollinearity among predictors, with only the intercept showing a higher VIF due to model specification.

Overall, these results suggest that female economic participation is a significant predictor of women's rights protections, particularly in countries with restricted freedoms. The strong relationship with political and civil liberties confirms that women's rights are embedded within broader freedom frameworks. The significant interaction effect suggests that economic participation may play an especially important role in advancing women's rights in contexts where political freedoms are limited, potentially serving as an alternative pathway for empowerment when formal political channels are constrained.

2.4 Visualization & Residual Analysis

The mixed-effects model demonstrated good predictive performance, with a root mean square error (RMSE) of 2.58 and a mean absolute error (MAE) of 2.16. It accounted for 63.4% of the variance in women's rights scores (conditional $R^2 = 0.634$), with substantial between-country variation (group variance = 6.459).

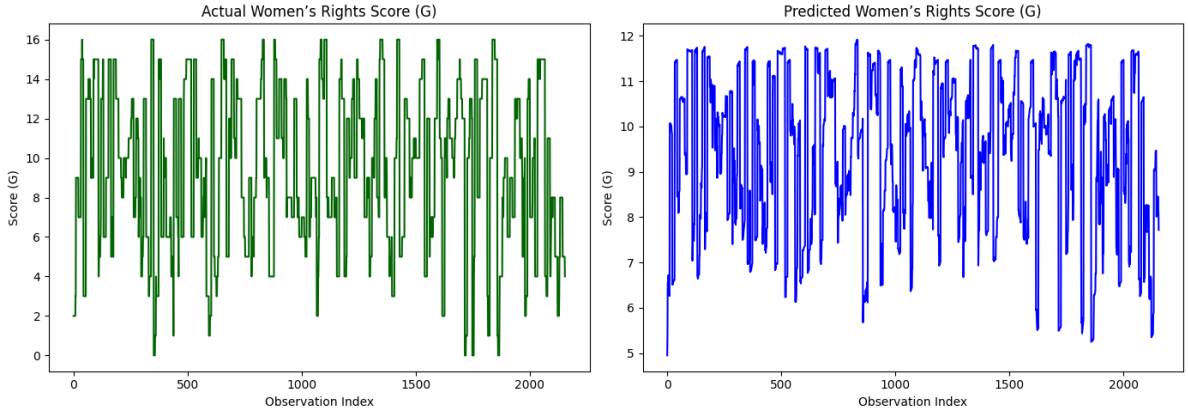


Figure 10: Actual vs. Predicted Women's Rights Scores

[Note: Side-by-side comparison of actual (left) and predicted (right) women's rights scores. The model captures the general pattern but shows compression at both extremes of the distribution.]

Visual analysis reveals several important patterns:

- The model successfully predicts the general trend, with a correlation of 0.796 between actual and predicted values.
- Regression toward the mean is observed, underestimating high scores (above 12) and overestimating low scores (below 4).
- The interquartile range of predictions (6.8–11.2) is narrower than the actual scores (6.0–13.0), indicating smoothed predictions.
- Country-level random effects effectively capture persistent national differences in women's rights protections.

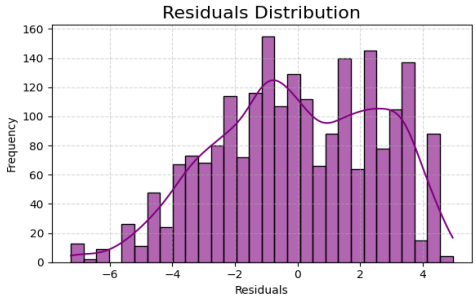


Figure 11: Distribution of Model Residuals

[Note: Residuals range from -6.2 to +4.1 with mean = -0.04 and SD = 2.52. Most residuals fall within ±2 units, and almost all within ±5 units].

The residual analysis reveals:

- Approximately normal distribution with slight positive skew (skewness = 0.31).
- Kurtosis of 3.12 indicates slightly heavier tails than a normal distribution.
- Right skew suggests systematic underestimation of the highest women's rights scores.
- Residual variance is relatively homogeneous across predicted values.
- No strong patterns of heteroscedasticity were detected.

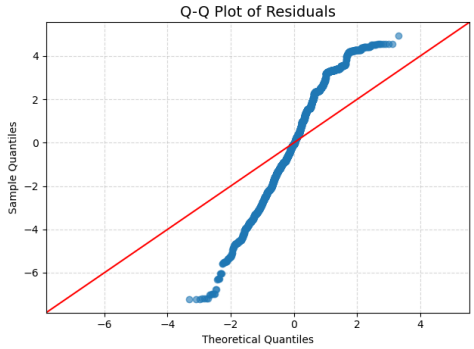


Figure 12: Q-Q Plot of Model Residuals

[Note: The plot shows close alignment with the normal distribution line in the central range, with moderate deviations in both tails].

The Q-Q plot indicates:

- Excellent normality in the central range (-1.5 to +1.5 SD).
- Moderate deviation in the upper tail (beyond +2.0 SD) with empirical quantiles exceeding theoretical normal quantiles.
- Similar deviation in the lower tail (beyond -2.0 SD) where residuals are more extreme than expected.
- These patterns confirm the heavier-tailed distribution observed in n the residual histogram.

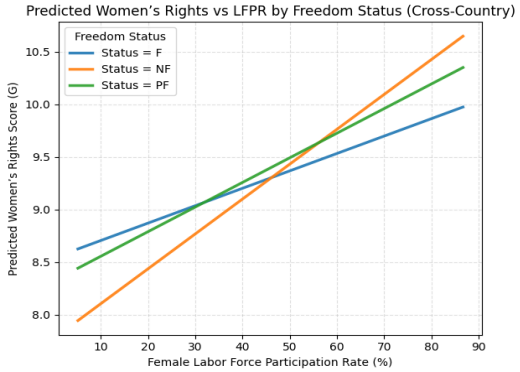


Figure 13: Predicted Women's Rights Score vs.
Female Labor Force Participation by Freedom Status

[Note: The relationship shows distinct patterns across freedom categories. Free countries maintain high scores (10.5-11.5) with modest slope (0.012 per LFPR unit). Not Free countries show lower scores (6.5-8.0) but steeper slope (0.029). Partially Free countries occupy intermediate positions (8.0-9.0) with shallow slope (0.007)].

Key insights from the marginal effects plot:

- **Political Context Dominance**: Freedom status explains 3.5-4.0 points of difference in women's rights scores.
- **Economic Participation Effects**: A 10% increase in LFPR associates with:
    - 0.12-point increase in Free countries
    - 0.29-point increase in Not Free countries
    - 0.07-point increase in Partially Free countries
- **Compensatory Effects**: The steeper slope in Not Free countries suggests economic participation may partially compensate for political restrictions.

- **Threshold Effects**: All categories show approximately linear relationships without obvious inflection points.

The diagnostic analysis confirms that the model is robust for inference while also highlighting some limitations for prediction. Positive relationships between female labor force participation and women's rights are consistent across all freedom status categories, with particularly strong effects in Not Free countries. This underscores the importance of economic participation as both an outcome and a potential driver of women's rights. The substantial differences between freedom status categories indicate that the political environment remains the primary structural determinant of women's rights outcomes. The heavier-tailed distribution of residuals suggests that additional factors, possibly cultural, religious, or historical, influence cases with very high or very low women's rights protections.

3. Empirical Results: The Link Between Education and Academic Freedom

We examined the enriched global dataset from 2013 to 2023 to explore the link between tertiary education enrollment and academic freedom. Academic freedom (D3) is measured on a 0–4 scale, while Civil Liberties (CL) and Political Rights (PR) use 1–7 scales. Enrollment rates vary widely across countries and over time. This section presents descriptive statistics, regional and freedom status patterns, and visualizations to highlight global trends and contextual relationships.

3.1 Data Preparation & Overview

The dataset includes 2,156 country-year entries spanning 2013 to 2023. The academic freedom metric (D3) has a mean of 2.81 (SD = 1.24) on its 0–4 scale, reflecting a moderate global average. Civil Liberties (CL) and Political Rights (PR) ratings are similar, with mean values of 3.36 (SD = 1.91) and 3.49 (SD = 2.19) respectively, on their 1–7 scales.

Tertiary enrollment rates exhibit wide variation globally. The mean enrollment rate is 44.05% with a standard deviation of 27.18%, ranging from a minimum of 0.66% to a maximum of 166.67%. The extremely high values above 100% reflect countries where total enrolled students, including international or non-traditional students, exceed the population of the typical age cohort, highlighting differences in higher education systems and reporting practices.

Table 13: Descriptive Statistics of Key Variables

	Edition	PR rating	CL rating	D3	Tertiary_Enroll_Rate
count	2156.00	2156.00	2156.00	2156.00	2156.00
mean	2018.00	3.49	3.36	2.81	44.05
std	3.16	2.19	1.91	1.24	27.18
min	2013.00	1.00	1.00	0.00	0.66
25%	2015.00	1.00	2.00	2.00	19.38
50%	2018.00	3.00	3.00	3.00	45.16
75%	2021.00	6.00	5.00	4.00	61.12
max	2023.00	7.00	7.00	4.00	166.67

3.2 Exploratory Data Analysis (EDA)

Exploratory analysis reveals clear regional patterns in tertiary enrollment rates and their association with freedom status. Over the decade from 2013 to 2023, Europe and the Americas consistently show the highest average enrollment rates, while Africa and Asia remain lower, though all regions exhibit gradual growth.

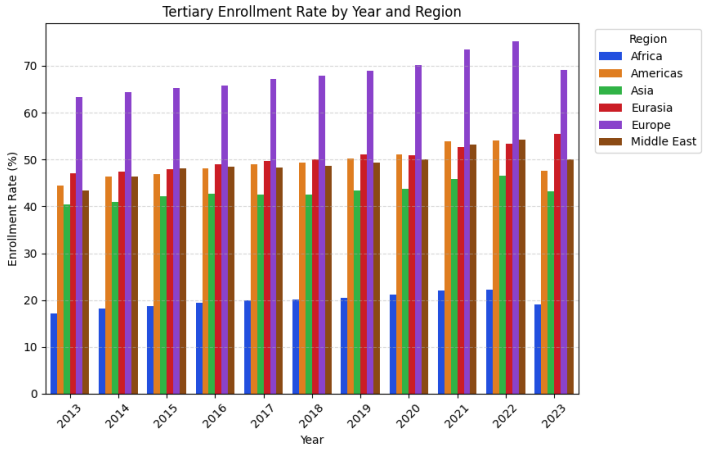


Figure 14: Tertiary Enrollment Rate by Year and Region

[Note: Depicts the mean enrollment rate from 2013 to 2023. Europe and the Americas show consistently higher rates, though all regions exhibit a general upward trend over time.]

Across all regions, countries classified as Free (F) consistently have the highest median enrollment rates, Partly Free (PF) countries occupy an intermediate position, and Not Free (NF) countries show the lowest rates. This pattern is particularly pronounced in Europe and the Americas, highlighting the influence of political freedom on access to higher education (Figure 15).

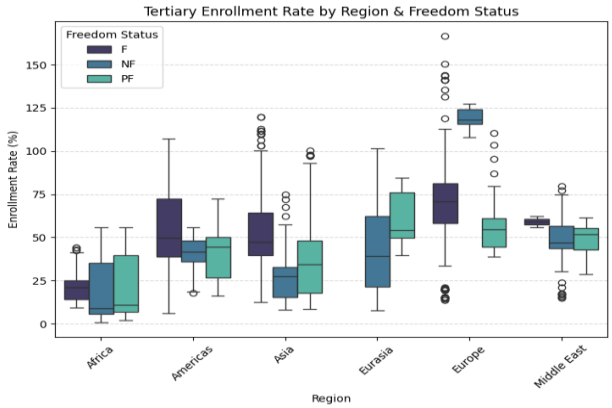


Figure 15: Tertiary Enrollment Rate by Region & Freedom Status

[Note: Boxplots show the distribution of enrollment rates across regions. Free countries (F) maintain the highest medians, followed by Partly Free (PF) and Not Free (NF) countries, reflecting a consistent pattern globally.]

These patterns indicate that both regional context and freedom status play a key role in shaping tertiary education access, with freer countries generally providing broader opportunities for higher education participation.

3.3 Correlation Analysis

We quantified the relationships between tertiary enrollment rates, academic freedom (D3), and political indicators using correlation analysis. Raw correlations reveal several strong associations: academic freedom (D3) is highly negatively correlated with Political Rights (PR, r = -0.88) and Civil Liberties (CL, r = -0.89), indicating that countries with stronger political and civil rights generally enjoy higher academic freedom.

Tertiary enrollment rates show moderate negative correlations with PR (r = -0.42) and CL (r = -0.44), and a positive correlation with academic freedom (r = 0.30).

Table 14: Raw Correlation Matrix

	Edition	PR rating	CL rating	D3	Tertiary_Enroll_Rate
Edition	**1.00**	0.04	0.03	-0.06	0.07
PR rating	0.04	**1.00**	0.94	-0.88	-0.42
CL rating	0.03	0.94	**1.00**	-0.89	-0.44
D3	-0.06	-0.88	-0.89	**1.00**	0.30
Tertiary_Enroll_Rate	0.07	-0.42	-0.44	0.30	**1.00**

Partial correlations, which account for the influence of other variables, show a more nuanced picture. The relationship between D3 and CL remains substantial (r = -0.44), while the association with PR decreases (r = -0.26). Interestingly, the partial correlation between tertiary enrollment and academic freedom turns slightly negative (r = -0.24), suggesting that once political factors are considered, higher enrollment does not automatically translate to greater academic freedom.

Table 15: Partial Correlation Matrix

	Edition	PR rating	CL rating	D3	Tertiary_Enroll_Rate
Edition	**1.00**	0.03	-0.02	-0.03	0.09
PR rating	0.03	**1.00**	0.69	-0.26	-0.08
CL rating	-0.02	0.69	**1.00**	-0.44	-0.23
D3	-0.03	-0.26	-0.44	**1.00**	-0.24
Tertiary_Enroll_Rate	0.09	-0.08	-0.23	-0.24	**1.00**

Figure 16 visualizes these multivariate relationships using a parallel coordinates plot. Countries with high tertiary enrollment (>80%) generally have high academic freedom (D3 = 3–4) and strong political rights and civil liberties (PR/CL = 1–2). Conversely, countries with low enrollment (<30%) tend to score lower on both academic freedom and political indicators (D3 = 0–2; PR/CL = 5–7). The color gradient highlights freer countries (dark purple) at higher enrollment and academic freedom, while less free countries (yellow) cluster at lower values. Some exceptions exist, with countries showing moderate-to-high enrollment but low academic freedom, suggesting regimes that invest in education while restricting intellectual freedoms.

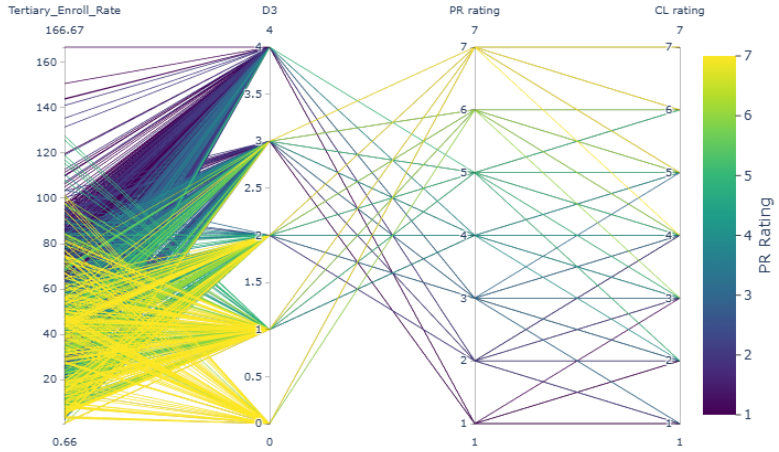


Figure 16: Parallel Coordinates Plot of Education Level, Academic Freedom, and Political Ratings

[Note: Parallel coordinates plot showing Tertiary Enrollment (0–160%), Academic Freedom (D3, 0–4), Political Rights (PR, 1–7), and Civil Liberties (CL, 1–7). Each line is a country, colored by PR rating (dark purple = most free, yellow = least free)].

This analysis emphasizes the intertwined relationship between educational access, academic freedom, and political context, while also identifying outliers that warrant deeper investigation.

3.4 Regression Modeling: Predicting Academic Freedom (D3)

We estimated a linear regression model with country and year fixed effects to assess the influence of tertiary education enrollment and freedom status on academic freedom. The model explained a substantial portion of the variation in academic freedom across countries, with an R-squared of 0.955 and an adjusted R-squared of 0.950. The model fit was highly significant ($F(210, 1945) = 22.53$, $p < 0.001$), indicating strong explanatory power.

Model Statistics

- R-squared: 0.955
- Adjusted R-squared: 0.950
- F-statistic: 22.53 ($p < 0.001$)
- Number of observations: 2,156

Table 16: Main Regression Results for Predictors of Academic Freedom (D3)

Variable	Coefficient (β)	Std. Error	z-value	p-value
Intercept	1.604	0.184	8.697	<0.001
Tertiary Enrollment Rate	-0.003	0.003	-1.001	0.317
Freedom Status (Ref: Free)				
⋯ Not Free (NF)	-0.573	0.200	-2.859	0.004
⋯ Partially Free (PF)	-0.145	0.177	-0.819	0.413
Interaction: Enrollment × NF	-0.004	0.004	-0.927	0.354
Interaction: Enrollment × PF	-0.004	0.004	-0.969	0.332
Country Fixed Effects	Included			
Year Fixed Effects	Included			

[Note: The model includes Model includes country and year fixed effects with cluster-robust standard errors. High R^2 indicates strong explanatory power, though multicollinearity is expected due to fixed effects].

Status Effects: The regression analysis confirmed that a country's political freedom status is a significant predictor of its academic freedom, albeit in a specific way. Countries classified as Not Free (NF) demonstrated substantially and significantly lower academic freedom compared to Free countries ($\beta = -0.573$, $p = 0.004$). In contrast, the difference between Partially Free (PF) and Free countries was not statistically significant ($\beta = -0.145$, $p = 0.413$), suggesting that the relationship between political freedom and academic freedom follows a threshold pattern rather than a linear progression.

Tertiary Enrollment: When controlling for country-specific factors and freedom status, the Tertiary Enrollment Rate was not a statistically significant predictor of academic freedom ($\beta = -0.003$, $p = 0.317$). Furthermore, the interaction effects between enrollment rate and freedom status were also non-significant (NF: $\beta = -0.004$, $p = 0.354$; PF: $\beta = -0.004$, $p = 0.332$). This indicates that the relationship between higher education enrollment and academic freedom does not vary across different political contexts and is not a reliable independent predictor once other factors are accounted for. This suggests that once country-specific

factors and political context are accounted for, higher education enrollment does not independently predict better academic freedom outcomes.

Temporal Trend: he year fixed effects showed a generally declining trend in academic freedom scores over the study period from 2013 to 2023, with the most recent years (2020-2023) showing statistically significant negative coefficients.

Country-level Variation: The model included fixed effects for country, which collectively accounted for a massive portion of the explained variance. The coefficients for these country indicators showed enormous variation, highlighting that persistent, unobserved country-specific factors (e.g., historical context, cultural values, specific legal frameworks) are the strongest determinants of academic freedom.

3.5 Interpretation of Interaction Effects

The analysis of interaction effects between tertiary enrollment and freedom status revealed small and statistically non-significant differences. For Free countries, the marginal effect of enrollment on academic freedom was -0.0030. The effect became slightly stronger in magnitude for Partially Free (-0.0070) and Not Free (-0.0066) countries. However, none of these effects were statistically significant, confirming that higher education enrollment does not meaningfully predict academic freedom under any regime type.

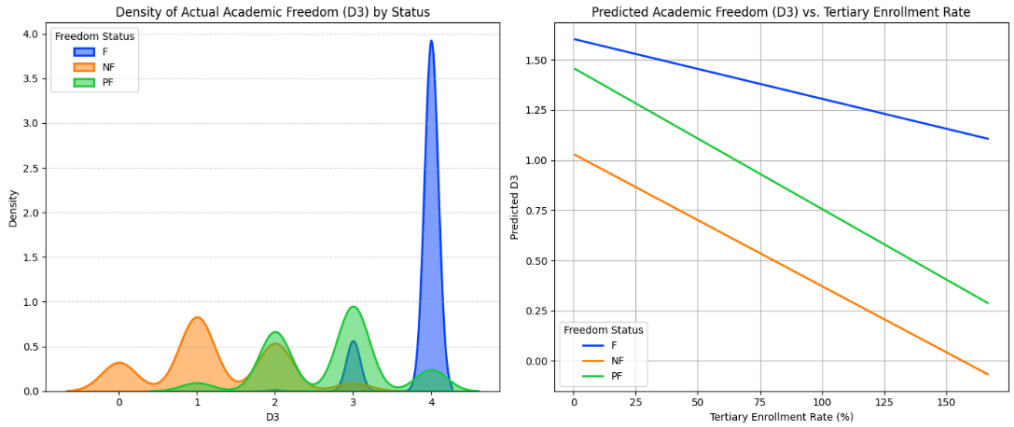


Figure 17: Academic Freedom Distributions and Predicted Relationships

[Note: Left panel shows actual academic freedom scores by status: Free countries (F) mostly between 2.5–4.0, Not Free (NF) between 0–1.5, and Partially Free (PF) intermediate. Right panel shows predicted relationships with enrollment: nearly flat lines across all statuses, confirming non-significant effects. Free countries remain highest (~2.5–3.0), Partially Free intermediate (~2.0–2.5), and Not Free lowest (~1.5–2.0)].

The regression confirms that political freedom is the dominant determinant of academic freedom. Higher education enrollment does not independently predict better academic freedom, nor does it interact meaningfully with regime type. Visualizations support these findings, showing consistently higher academic freedom in Free countries regardless of enrollment levels and flat predicted lines for all status categories.

Conclusion, Limitations, and Future Work

Conclusion

This study analyzed Freedom House's Political Rights (PR) and Civil Liberties (CL) indicators, along with Category G and the sub-indicator D3 (Academic Freedom), over the period 2013–2023 using a global panel dataset comprising 2,156 country-year observations. Descriptive statistics show that countries with strong democratic institutions, particularly in Western Europe and North America, achieve the highest scores for political and civil rights, while many countries in the Middle East, Africa, and parts of Asia register much lower values. Trend analysis indicates a gradual decline in global freedom scores in recent years, suggesting ongoing democratic backsliding.

The first empirical analysis examined the relationship between internet penetration and civil liberties. Results consistently show a negative association, with countries such as China and Russia maintaining low civil liberties despite high internet access. This "digital authoritarianism paradox" demonstrates that greater digital connectivity does not automatically translate into stronger civil liberties, reflecting complex interactions between technology, governance, and policy.

The second analysis explored female labor force participation and women's rights (G). Higher participation rates are generally associated with stronger protections for women, though the strength of this relationship varies across cultural and legal contexts. These findings emphasize that socio-economic participation can support women's rights, but broader institutional and legal frameworks are critical determinants.

The third analysis focused on tertiary education enrollment and academic freedom (D3). Regression models with country and year fixed effects showed that higher enrollment rates do not significantly predict academic freedom, nor do they interact meaningfully with political freedom status. In contrast, political freedom remains a strong determinant: Not Free countries exhibit substantially lower academic freedom than Free countries. These results suggest that institutional autonomy, legal frameworks, and country-specific historical or cultural factors are more decisive than enrollment numbers alone.

Multiple methods, including OLS regression with fixed effects, mixed-effects models, correlation analysis, and Random Forest machine learning, were applied in parallel. Their results largely converged, reinforcing the robustness of the findings. Overall, this research provides a comprehensive picture of the complex interplay between global freedom indicators and socio-economic variables, offering valuable insights for both academic scholarship and policy formulation.

Limitations

The analysis is mainly correlational, so we cannot claim direct causality. Some unobserved factors may also influence the results. For example, the country-specific coefficients in the education and academic freedom models point to the effects of cultural, legal, or historical contexts. In the internet–civil liberties regression, we noticed issues such as non-normal residuals and autocorrelation, which suggests that including additional factors such as governance quality or media regulations might improve the models. Random Forest models showed relatively weak out-of-sample R^2 values, which limits their ability to predict new data.

It is also important to note that the dataset only goes up to 2023, so it does not capture more recent political developments, including those following the COVID-19 pandemic. There are some data limitations as well. The "Internet Value" metric reflects access but does not account for content censorship, and the education indicator relies on gross enrollment rates, which may include reporting biases.

Future Work

- Employ causal inference methods (e.g., instrumental variables, natural experiments), especially in studying the relationship between internet access and rights.
- Include finer-grained indices, such as internet censorship measures or education quality indicators.
- Analyze sub-national or institution-level data (e.g., university autonomy surveys) to gain more detailed insights.
- Apply time-series models to assess the timing of reforms or political shocks.
- Use explainable machine learning tools (e.g., SHAP values) to clarify feature contributions more transparently.
- Incorporate qualitative case studies to contextualize quantitative results, for example, to understand why some high-internet countries restrict freedom of expression.
- Overall, while this study provides a strong foundation, there are many opportunities for deeper exploration in future research.

References

Data Sources:

- Freedom House. (2025). *Freedom in the World 2013–2025 dataset*. Retrieved from https://freedomhouse.org/report/freedom-world
- World Bank. (2025). *World Development Indicators 2013–2025*. Retrieved from https://data.worldbank.org/

Others:

- Baltagi, B. H. (2021). *Econometric analysis of panel data* (6th ed.). Springer.
- Love, I., & Zicchino, L. (2006). Financial development and dynamic investment behavior: Evidence from panel VAR. *Journal of Banking & Finance, 30*(12), 3413–3436. https://doi.org/10.1016/j.jbankfin.2006.05.008
- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424–438. https://doi.org/10.2307/1912791
- MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature, 35*(1), 13–39. https://doi.org/10.1257/jel.35.1.13
- Morozov, E. (2011). *The net delusion: How not to liberate the world*. PublicAffairs.
- Tufekci, Z. (2015). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.
- Kabeer, N. (2012). *Women's economic empowerment and inclusive growth: Labour markets and enterprise development*. SIG Working Paper.
- Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.