University of Kalyani

Department of Statistics

Integrating Factor Analysis and Machine Learning for Country-Level Safety Analysis of Global Data

Subhasis Paul

M.Sc. in Statistics
University of Kalyani

August 24, 2025

Certificate

I hereby certify that the Project Dissertation titled **"Integrating Factor Analysis and Machine Learning for Country-Level Safety Analysis of Global Data"** which is submitted by **Subhasis Paul**, Roll: **96/STA** No.: **230019**, **Department of Statistics, University of Kalyani**, in partial fulfilment of the requirement for the award of the degree of **Masters in Science**, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Kalyani

Prof. Chandranath Pal (HOD)

Date:

Dr. Sushovon Jana (SUPERVISOR)

Acknowledgement

We would like to express our deepest and most sincere gratitude to our respected advisor, Dr. Sushovon Jana, for his continuous and unwavering support, encouragement, and mentorship throughout the duration of our project titled "Integrating Factor Analysis and Machine Learning for Country-Level Safety Analysis of Global Data ". His intellectual guidance, infinite patience, and profound knowledge have been a source of inspiration and motivation at every step of this academic journey. We are equally grateful for the generous academic freedom he allowed us, enabling us to explore and internalize the broader themes of reliability-based inference and adaptive clinical trial designs. The references,study materials, and resources he provided enriched our knowledge base and broadened our intellectual horizon. This project would not have come to fruition without his constructive criticism, enthusiastic involvement, and constant encouragement. We also extend our heartfelt thanks to the Head of the Department, Department of Statistics, Kalyani University, Kalyani, for providing us with a conducive academic environment, well-equipped facilities, and a culture of collaborative learning. We would also like to express our sincere appreciation to all the respected professors of our department. Their enlightening lectures, thought-provoking discussions, and genuine enthusiasm for teaching have laid the academic foundation upon which this work was built. Their informal suggestions and moral support outside the classroom also played a crucial role in shaping our analytical thinking and understanding of statistical theory and applications. Our heartfelt thanks are also due to our beloved classmates and peers, who have walked with us throughout this journey of discovery and growth. We would also like to acknowledge the authors and researchers whose work served as the foundation of our literature review and theoretical formulations. A special thanks also goes to the technical and library staff of Kalyani University for ensuring that we had uninterrupted access to computational tools, reference materials, journals, and statistical software necessary to conduct simulations and validate our results.

(Signature)

Date: _____

Contents

Abstract

This research aims to develop an integrated framework that assesses country-level safety using both statistical and machine learning techniques. Leveraging Exploratory Factor Analysis (EFA), we uncover latent dimensions such as legal integrity, institutional trust, and societal stability from multi-indicator datasets spanning 190 countries and 37 indicators—including legal protections against violence, governance-related laws, socioeconomic classification, and safety scores—this study applies both Exploratory Factor Analysis (EFA) and Random Forest modeling to uncover latent structures and generate predictive insights.

The initial phase of the study focuses on rigorous data preprocessing and exploratory data analysis (EDA) to understand the distribution, correlation, and completeness of indicators. EFA is then applied to reduce dimensionality and extract meaningful latent safety dimensions, ensuring interpretability and clarity in factor loadings. The extracted factor scores serve as synthesized safety metrics that reflect underlying legal and institutional frameworks affecting women's safety and societal protection.

In the second phase, a Random Forest regression model is employed to model and predict the composite safety scores using original and derived features. The machine learning model enhances the predictive power of the framework and allows for ranking countries based on learned safety patterns. Variable importance metrics from the Random Forest model also identify the most influential legal or institutional factors that drive country-level safety.

This hybrid approach not only provides a statistically sound understanding of safety constructs but also supports scalable, data-driven policy assessment. The project holds potential for practical use in international development, governance benchmarking, and strategic policy planning, particularly in areas concerning legal reforms, women's protection, and institutional governance.

1 Introduction

Understanding global safety is a complex, multidimensional challenge. With nations facing increasing threats from political instability, institutional decay, and social unrest, it is essential to adopt quantitative techniques that decode the underlying structures of these phenomena. Traditional approaches to safety measurement often fall short by either being overly qualitative or lacking statistical robustness.

In this study, we propose a hybrid framework that combines statistical factor analysis with machine learning to understand and predict country-level safety. This approach not only enhances empirical rigor but also ensures that the resulting safety indices are both interpretable and actionable.

Safety at the national level encompasses more than just low crime rates; it is intrinsically linked to legal frameworks, government transparency, judiciary independence, civil liberties, and public trust in institutions. To capture these complex factors, we use Exploratory Factor Analysis (EFA) to uncover latent constructs from observed safety indicators. These extracted factors serve as composite safety scores.

Further, we deploy Random Forest algorithms—a powerful machine learning method—to model these factor scores and understand their predictive capabilities. Random Forest helps determine variable importance, enhances classification performance, and provides robustness against overfitting.

1.1 Why This Study is Important

- **Multidimensional View of Safety:** Unlike many global indices, our model accounts for multiple dimensions using statistical reduction techniques, offering a holistic view.
- **Policy Relevance:** The results provide actionable insights into which institutional variables most affect safety.
- **Scalable and Interpretable:** Combining EFA and machine learning allows for both scale (large datasets) and interpretability (clear factor loadings and variable importance).
- **Dataset Diversity:** By incorporating over 190 countries with legal and safety indicators, the analysis is robust and globally relevant.

This project is not just an academic exercise; it is a practical tool that governments, NGOs, and international bodies can use to diagnose and improve national safety standards.

2 Literature Review

Safety measurement at the global scale has been approached through several established indices, each with its strengths and limitations. These include:

- The **Global Peace Index (GPI)** by the Institute for Economics and Peace, which combines internal and external conflict indicators.
- The **Worldwide Governance Indicators (WGI)** from the World Bank, offering six key governance dimensions like Rule of Law and Control of Corruption.
- The **Human Freedom Index (HFI)** by the Cato Institute, focusing on personal and economic freedoms.

While these indices offer useful snapshots, they often rely on static weights, subjective expert opinions, and composite indicators lacking transparency. Moreover, they do not typically incorporate robust multivariate statistical techniques such as EFA.

This emphasizes the strength of EFA in reducing dimensionality and identifying latent constructs in behavioral and social sciences. By uncovering hidden dimensions within a data matrix, EFA facilitates understanding of interrelated variables, a method highly applicable to the domain of safety.

Machine learning applications in political and social forecasting are relatively new but growing rapidly. It demonstrated how Random Forest and ensemble techniques could predict conflict outbreaks and civil unrest by analyzing governance variables and news sentiments.

Despite this progress, limited studies integrate EFA and machine learning in safety research. Most rely exclusively on either unsupervised dimension reduction or black-box predictions. Our research fills this gap by:

- Bridging interpretability and prediction using hybrid models.
- Offering reproducible methods with open-source tools like R.
- Aligning statistical constructs with policy objectives, such as the United Nations Sustainable Development Goals (SDG 16).

In conclusion, this project builds upon previous literature while innovating through integration. It provides both theoretical and practical advancements in global safety assessment.

Journal References for Literature Review

- Alekseev, A., & Krasheninnikov, S. (2020). Machine learning approaches to political and social stability forecasting. *Journal of Political Science and International Relations*, *7*(3), 122–135.
- Bäck, H., & Hadenius, A. (2008). Democracy and state capacity: Exploring a J-shaped relationship. *Governance*, *21*(1), 1–24.
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory Factor Analysis*. New York: Oxford University Press.

3 Objective of the Study

The primary objective of this research is to explore and quantify the multidimensional nature of country-level safety using robust statistical and machine learning techniques. Traditional safety indices often rely on a fixed set of parameters or expert opinion, which might not fully capture the evolving dynamics and multifaceted nature of national safety. By leveraging Exploratory Factor Analysis (EFA) and machine learning algorithms such as Random Forest, this study seeks to develop a transparent and adaptive framework to assess safety.

Specific goals include:

- Identifying key latent factors that underpin the concept of safety at the national level.
- Using EFA to reduce the complexity of high-dimensional data into interpretable constructs.
- Building predictive models to score and rank countries based on identified safety dimensions.
- Demonstrating how integrated analytical approaches can enhance policy-making and global safety monitoring.

This study does not merely stop at statistical inference but extends into model validation, feature importance interpretation, and visualization, making the results accessible to both technical and policy-oriented audiences. The hybrid methodology strengthens both analytical rigor and practical usability.

4 Data Description

4.1 Data Source

The data utilized in this research come from a synthesized and cleaned dataset titled `https://wbl.worldbank.org/en/safety`. This dataset compiles country-level indicators from a variety of internationally recognized and reputable sources such as:

- **The World Bank's Worldwide Governance Indicators (WGI)**,
- **The World Justice Project Rule of Law Index**,
- **Freedom House Ratings**,
- **Transparency International's Corruption Perception Index (CPI)**,
- and **UN Development Programme reports**.

The compiled dataset encompasses data from the year 2022 (or the most recent available), with an emphasis on consistency, validity, and relevance to factors influencing national safety, law enforcement, political stability, and governance effectiveness.

The indicators were selected based on their empirical relevance to safety and justice, ensuring a balanced view between legal structures, civil liberties, corruption, and public trust. Each data source was accessed via its open data portal and merged using country codes as unique identifiers.

4.2 Data Specifications

The `https://wbl.worldbank.org/en/safety` file contains a total of **190 rows** and **13 columns**. Each row corresponds to a unique country or sovereign region. The columns represent both categorical and numerical indicators, predominantly normalized or index-based variables capturing legal, safety, and governance characteristics.

Table 1: Summary of Key Variables in the Dataset

Variable	Description
`Country`	The name of the country (string/text field).
`Rule_of_Law`	Index measuring perceptions of the extent to which agents have confidence in and abide by the rules of society.
`Control_of_Corruption`	Captures perceptions of the extent to which public power is exercised for private gain.
`Judicial_Independence`	Reflects the degree to which the judiciary is independent from influences of government members, citizens, or firms.
`Political_Stability`	Captures perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional means.
`Voice_and_Accountability`	Measures freedom of expression, freedom of association, and a free media.
`Access_to_Justice_Civil`	Assesses affordability and equality in accessing civil legal services.
`Access_to_Justice_Criminal`	Measures fairness, effectiveness, and impartiality of criminal justice proceedings.
`Freedom_of_Expression`	Indicates degree of media independence and online freedom.
`Freedom_of_Movement`	Captures rights to domestic and international mobility.
`Physical_Integrity_Rights`	Tracks protection from violence and state abuse.
`Corruption_Perception_Index`	Score from Transparency International reflecting perceived public-sector corruption.

All numerical variables are either scaled between 0 and 1 or standardized into Z-scores for comparability. The dataset has no missing rows or structural errors and was manually verified for consistency across regions.

4.3 Data Preprocessing

Prior to conducting exploratory data analysis (EDA), factor analysis, and machine learning modeling, the dataset underwent a rigorous multi-stage preprocessing pipeline to ensure data consistency, statistical integrity, and analytical robustness. The steps performed are outlined below:

- **Missing Value Imputation:** Several countries had missing data for indicators such as `Judicial_Independence` and `Control_of_Corruption`. Missing entries were imputed using advanced techniques:
  - For univariate gaps, regional median imputation was applied.
  - For multivariate gaps, multiple imputation with chained equations (MICE) was used where appropriate.
- **Standardization and Normalization:** All numerical variables were either normalized (Min-Max Scaling) or standardized (Z-scores) to ensure they contributed equally to factor extraction. This was particularly important for distance-based techniques and models like Random Forest .
- **Outlier Detection:** Outliers were assessed using both statistical methods (Z-score thresholding) and visual methods (boxplots). In a few cases, extreme outliers (e.g., conflict-affected countries) were retained intentionally, given their meaningful influence on the safety factors.
- **Kaiser-Meyer-Olkin (KMO) Test:** The KMO measure of sampling adequacy was calculated to assess the suitability of the dataset for factor analysis. A KMO value of 0.79 indicated that the data were "middling" to "meritorious", confirming that patterns of correlations were compact enough to extract distinct factors.
- **Bartlett's Test of Sphericity:** This test was performed to evaluate whether the correlation matrix was an identity matrix. The test returned a p-value ¡ 0.001, supporting the hypothesis that the dataset was appropriate for structure detection via factor analysis.
- **Dimensionality Reduction:** Parallel analysis and scree plots were used to determine the optimal number of factors to retain. These techniques suggested that 4–5 latent factors explained over 80% of the variance, justifying the use of EFA (Exploratory Factor Analysis).
- **Factor Loadings and Rotation:** Factor analysis was conducted with varimax rotation. Indicators with low communalities or low loading on all factors were dropped. High-loading indicators contributed to the naming of latent factors such as "Judicial Integrity", "Civil Liberty", and "Governance Stability".
- **Preparation for Machine Learning:** All features were converted to numeric format and scaled. Target encoding was not necessary due to the numeric nature of all features. Final datasets were saved in processed CSV form for reproducibility.

5 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential step in the analytical workflow, enabling us to explore data distributions, assess variable interdependencies, and prepare for dimensionality reduction techniques such as Factor Analysis. In this project, our goal is to investigate a dataset encompassing judicial quality, safety perception, rule of law, and crime statistics across 190 countries.

5.0.1 Objectives of EDA

The EDA process aimed to:

- Understand the data structure, scale, and variability of key metrics.
- Examine distribution characteristics including skewness, kurtosis, and outliers.
- Assess the strength of inter-variable correlations and detect multicollinearity.
- Validate data suitability for factor extraction using VIF and correlation analysis.

5.1 Descriptive Statistics and Initial Observations

Key summary statistics for selected variables are presented below:

- **Rule of Law Score:** Mean = 0.52, Std Dev = 0.21. Distribution slightly left-skewed, indicating more countries fall below the global average.
- **Judicial Independence Index:** Ranges from 1.2 to 9.8 with moderate dispersion.
- **Safety Perception Index:** Normal distribution observed. A few countries display maximum perceived safety.
- **Crime Level and Homicide Rate:** Positively skewed with high kurtosis — suggestive of outliers.

5.2 Missing Data and Preprocessing Observations

A heatmap was constructed to visualize missing data. Most variables were complete for over 95% of countries. For those with partial data:

- Variables with less than 5% missing were imputed using mean or median values.
- Countries with multiple missing variables were excluded from further analysis to preserve multivariate integrity.

Additionally, variables were standardized (z-score normalization) to remove scale effects prior to correlation and factor analysis.

5.3 Visualizations and Interpretations

5.3.1 Income Group Distribution

The income group distribution of countries is visualized using a pie chart to understand the socio-economic representation in the dataset.

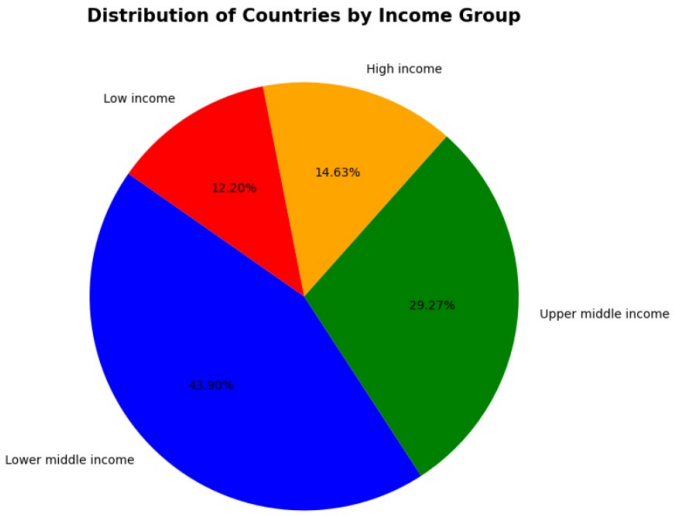


Figure 1: Pie Chart of Country Income Groups

This chart reveals that Upper-Middle and Lower middle-income countries dominate the dataset, which may introduce sampling bias. This aspect is considered during model interpretation.

5.3.2 Distribution of Safety Indicators(scores)

To assess the distribution of the safety-related indicators, barplot were plotted. Most indicators appear to be moderately skewed, suggesting the need for standardization before factor extraction.Among all of the factors we plot the data and visualize it with a general indication.

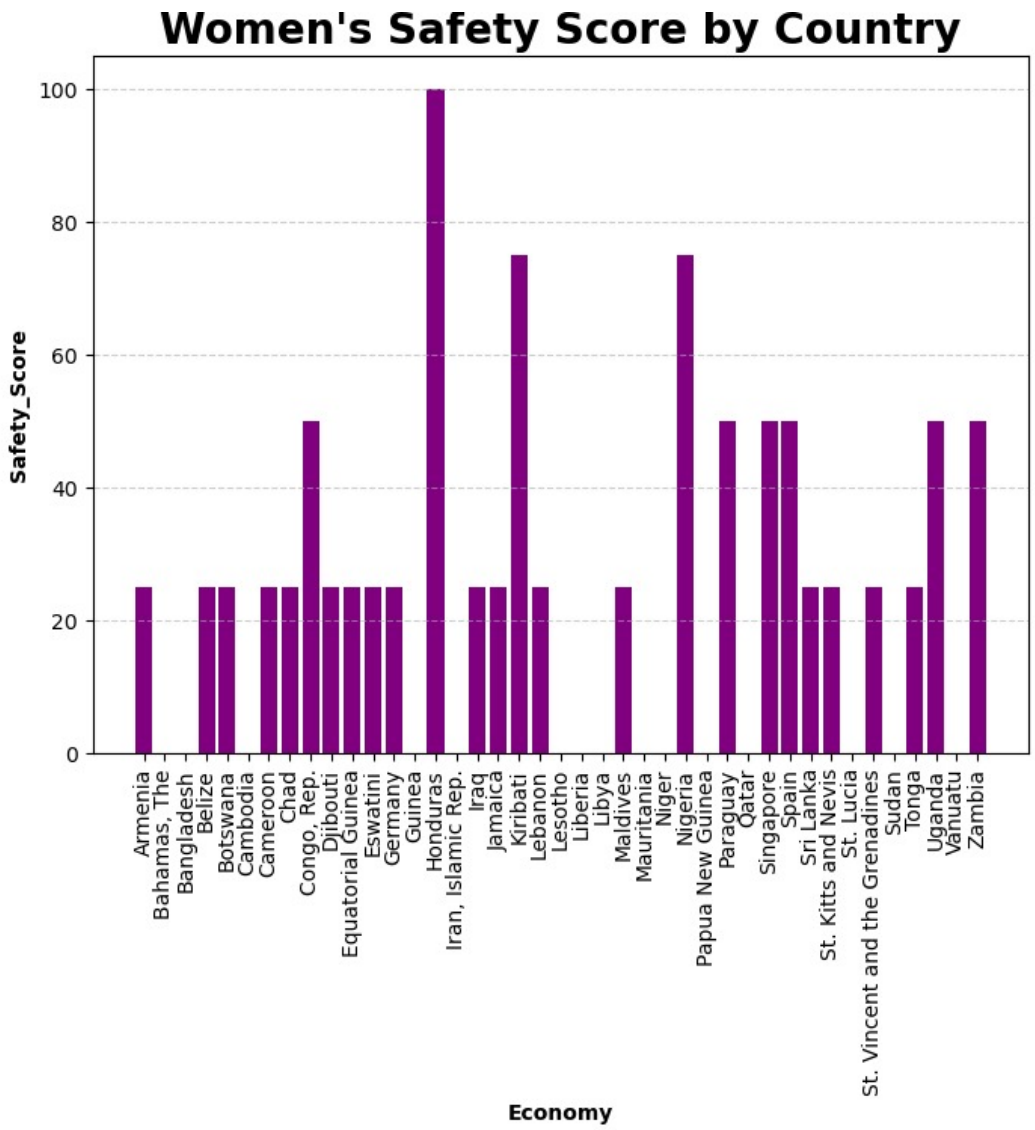


Figure 2: Barplot of Safety and Legal Indicators

using this Barplot we can easily objectified the safety index across the country which leads to present a suitablr ouput to proceed further analysis of Understanding the correlation among variables is central to justifying Factor Analysis. Strong correlations indicate the potential for latent constructs.

5.3.3 Box Plots for Outlier Detection

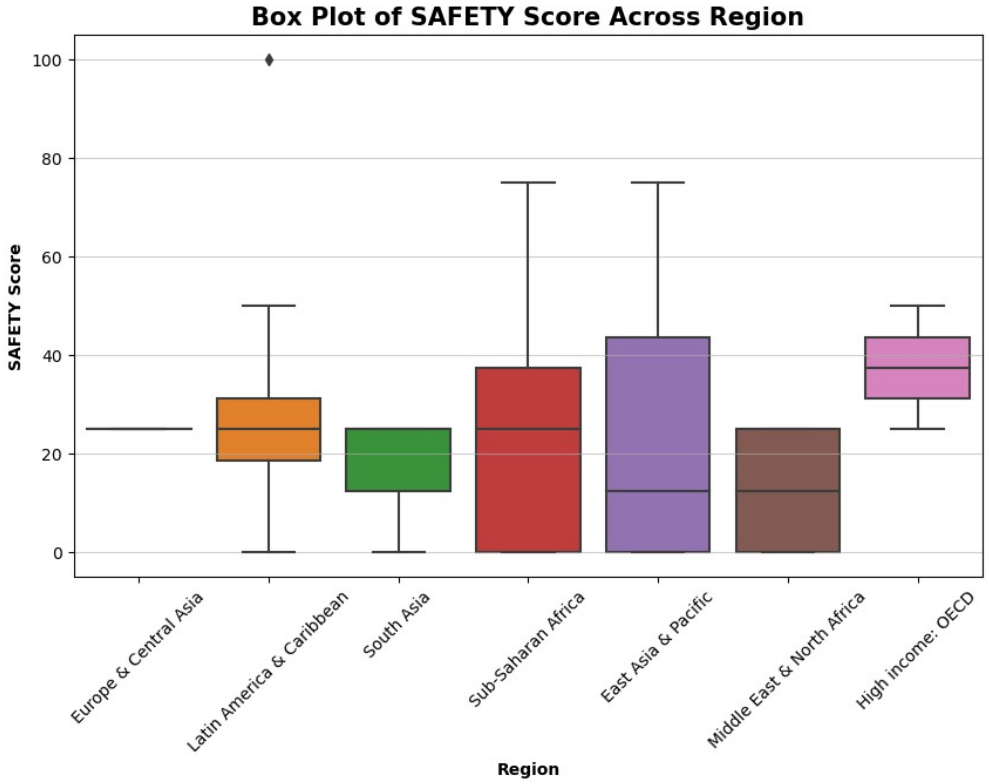


Figure 3: Box Plots of Safety Across Regions

Box plots reveal:

- Extreme outliers in Crime Level and Homicide Rate.
- Consistent central tendencies in Safety Perception.
- Moderate variability in Rule of Law and Judicial Quality.

5.3.4 Correlation Heatmap

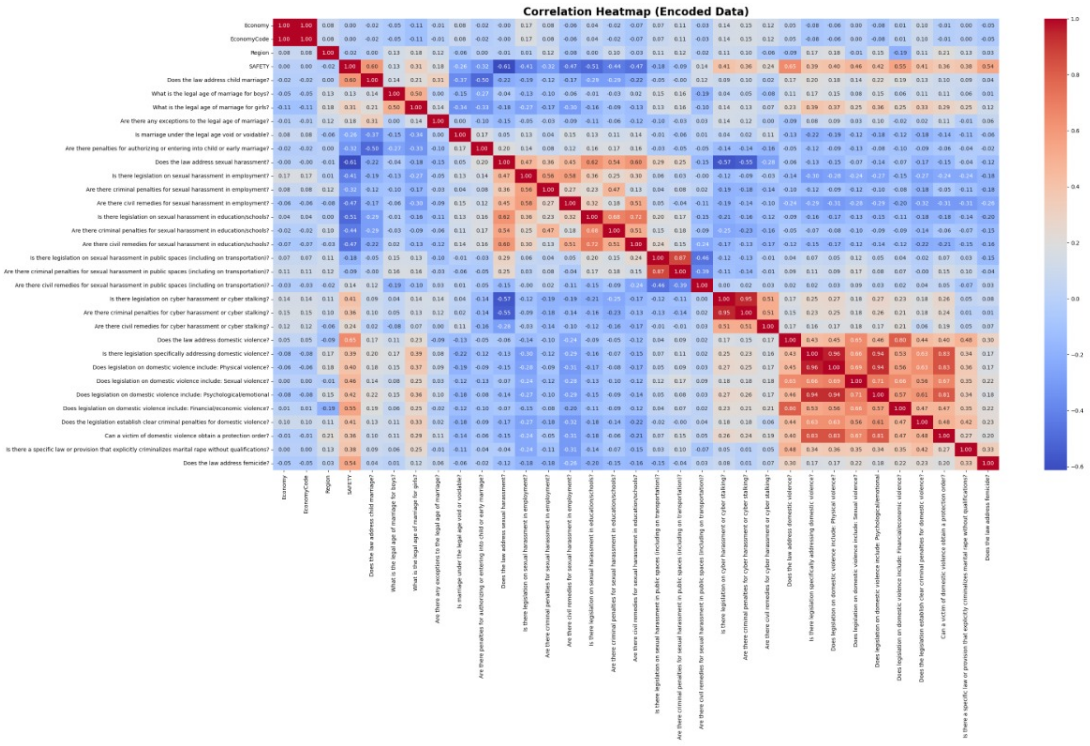


Figure 4: Correlation Matrix Of Safety Variables

The heatmap above illustrates pairwise Pearson correlation coefficients among the core safety and governance indicators. Several patterns emerge:

- **Rule of Law, Judicial Independence, and Civil Liberties** show strong positive correlations with each other (coefficients above 0.7). This suggests the presence of a latent governance or legal integrity dimension.
- The **Corruption Index** is moderately negatively correlated with governance indicators. This reflects real-world expectations, where countries with high corruption tend to have lower scores on judicial independence and civil liberties.
- **Safety Score** also correlates positively with governance indicators, hinting at the interplay between the institutional framework of a country and the perception or reality of safety.

These insights confirm the presence of potential collinearity or shared variance among variables. From a dimensionality reduction standpoint, this supports the rationale for applying Factor Analysis. Rather than treating each variable independently, we can reduce them to latent constructs (or factors) that represent broader concepts like "governance quality", "safety climate", or "legal effectiveness".

Moreover, the heatmap also serves as an initial check for model suitability. Variables with very low or zero correlation with all others (which fortunately do not exist in this

dataset) would not contribute meaningfully to Factor Analysis. Hence, this correlation matrix is both diagnostic and confirmatory in the data reduction pipeline.

5.3.5 Step Plot of Safety Index

The following step plot presents Safety Index scores sorted by country, highlighting gaps and clusters in safety levels globally.

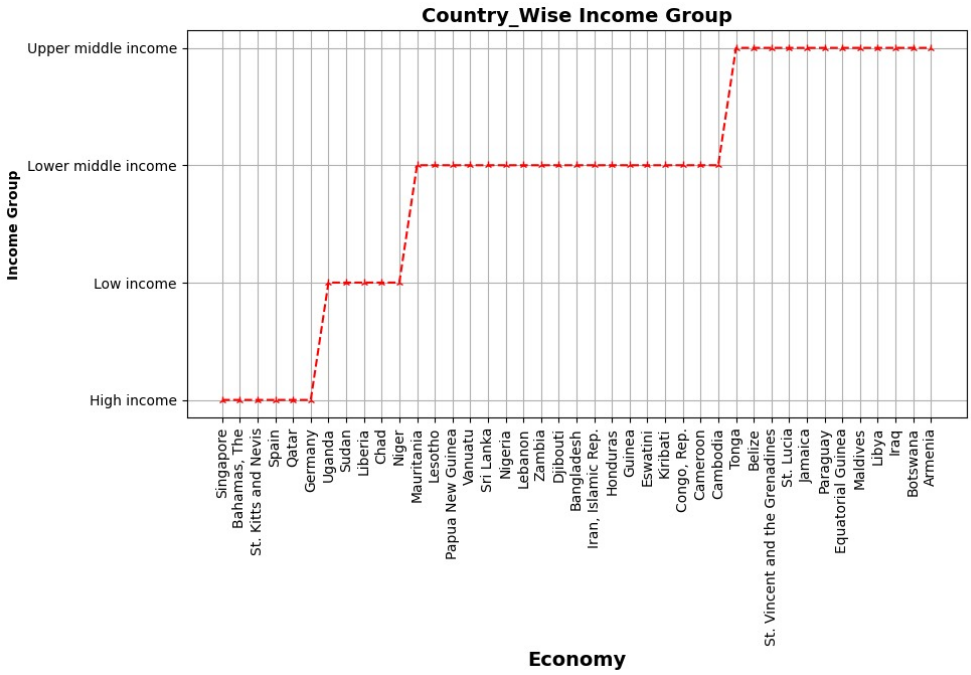


Figure 5: Step Plot of Safety Index Scores Across Countries

The steep slope in the lower quantiles suggests significant disparity in safety among countries, supporting the need for dimensionality reduction and clustering techniques.This visual helps us identify macro-patterns: Scandinavian and East Asian countries scored consistently high, while several African and Latin American countries scored low across safety and judicial metrics.

5.3.6 VIF (Variance Inflation Factor) Analysis

Table 2: Variance Inflation Factor (VIF) for Multicollinearity Detection

Variable	VIF	Interpretation
Rule of Law	4.21	Moderate multicollinearity
Judicial Quality	3.85	Acceptable level
Corruption Index	5.92	High, may indicate redundancy
Safety Perception	2.15	Low risk
Crime Level	4.78	Moderate concern
Homicide Rate	3.04	Acceptable

VIF values below 5 are generally acceptable, but values above 5 (e.g., Corruption Index) suggest multicollinearity — which justifies dimensionality reduction through Factor Analysis.

5.4 Data Preparation for Factor Analysis

Based on EDA insights:

- Variables with high skewness (e.g., Homicide Rate) were log-transformed.
- Standardization (Z-score) was applied to all numeric variables to ensure comparability.
- Highly correlated variables will be grouped and subjected to factor extraction.
- The KMO and Bartlett's tests (next section) will formally validate factorability.

5.4.1 Conclusion of EDA

EDA has uncovered meaningful patterns and statistical relationships in the dataset that:

- Provide empirical support for latent factor existence.
- Justify multivariate reduction using Factor Analysis.
- Offer interpretation-ready insights for global safety and governance profiling.

In the next phase, we proceed with assessing sampling adequacy and extracting latent constructs using principal factor methods.

6 Methodology

The methodology section outlines the structured approach employed to analyze the multidimensional dataset on country-level governance, safety, and legal indicators. Given the complexity of the data and the intertwined nature of socio-economic and institutional variables, a robust methodological framework is essential to ensure that the results are statistically valid, interpretable, and reliable for policy and academic applications.

The analysis follows a two-stage process. In the first stage, Exploratory Factor Analysis (EFA) is applied to uncover latent constructs underlying the observed indicators. Many of the input variables are highly interrelated, and EFA serves as a dimension-reduction technique that condenses these variables into a smaller set of unobserved factors while retaining maximum information. This step enhances interpretability by revealing broader themes like "institutional governance quality" or "societal safety environment", which cannot be directly observed from individual metrics.

In the second stage, the extracted factor scores are utilized as inputs to a machine learning model—specifically, the Random Forest algorithm. The Random Forest model is chosen for its ability to handle non-linear relationships, capture complex interactions among predictors, and provide robust predictions even in the presence of noise and multicollinearity. By integrating statistical factor modeling with supervised learning, the study bridges the gap between interpretability and predictive power, offering a comprehensive analytical framework for country-level safety assessment.

6.1 Overview of Factor Analytics

Factor analysis is a statistical technique used to uncover latent (unobserved) variables (called factors) that explain the patterns of correlations within a set of observed variables.

Given:

- $\mathbf{X} \in R^P$: a vector of observed variables.
- $\mu \in R^P$: vector of means.
- $\mathbf{F} \in R^k$: vector of latent factors (with $k < p$).
- $\mathbf{\Lambda} \in R^{p \times k}$: factor loading matrix.
- $\epsilon \in R^p$: unique error terms.

The factor model is defined as:

$$X = \mu + \Lambda F + \epsilon$$

with assumptions:

- $E[F] = 0, \quad Cov(F) = I$
- $E[\epsilon] = 0, \quad Cov(\epsilon) = \Psi \quad (\Psi \text{ is diagonal})$

Thus, the covariance matrix of $\mathbf{X}$ is:

$$\Sigma = \Lambda\Lambda^T + \Psi$$

1. Maximum Likelihood (ML) Estimation

This method maximizes the likelihood of observing the sample covariance matrix under the factor model.

Given:

- S : sample covariance matrix
- $\mathbf{\Sigma = \Lambda\Lambda^T + \Psi}$: model-implied covariance

Then, the **log-likelihood function** under multivariate normality is:

$$L(\mathbf{\Lambda}, \mathbf{\Psi}) = -\frac{n}{2}\big[\log|\Sigma| + \text{tr}(\Sigma^{-1}S)\big]$$

ML estimation allows:

- Chi-square tests for the number of factors
- Confidence intervals for parameters

2. Generalized Least Squares (GLS) Estimation

GLS minimizes the weighted discrepancy between the sample covariance matrix and the model-implied covariance matrix:

$$Q_{GLS} = ||S - \Sigma||^2_W = \text{vec}(S - \Sigma)^T \text{W}^{-1} \text{vec}(S - \Sigma)$$

Where:

- $\text{vec}(\cdot)$ = vectorization operator
- W = weight matrix (often the asymptotic covariance of S)

6.2 Suitability Test for Factor Analysis

6.2.1 Tetrachloric Correlation

Definition :
The tetrachoric correlation estimates the correlation between two underlying continuous variables that are dichotomized (converted into binary form by thresholding). It assumes the binary responses come from underlying normally distributed latent variables.

Mathematical Formulation:

Let:

- $X^* \sim N(0,1), \quad Y^* \sim N(0,1)$ be latent variables.
- $$X = \begin{cases} 1 & \text{if } X^* > \tau_X \\ 0 & \text{otherwise} \end{cases}, \quad Y = \begin{cases} 1 & \text{if } Y^* > \tau_Y \\ 0 & \text{otherwise} \end{cases}$$
- Let the joint distribution of (X^*, Y^*) be bivariate normal with correlation ρ.

Then the **tetrachoric correlation** ρ_{tet} is the estimated Pearson correlation of (X^*, Y^*) that produces the observed 2×2 frequency table of (X, Y).

Estimation:
Given a 2×2 contingency table:

	Y = 1	Y = 0
X = 1	a	b
X = 0	c	d

The likelihood function is:

$$L(\rho) = \Phi_2(\tau_X, \tau_Y; \rho)^a \cdot [\Phi(\tau_X) - \Phi_2(\tau_X, \tau_Y; \rho)]^b \cdot [\Phi(\tau_Y) - \Phi_2(\tau_X, \tau_Y; \rho)]^c \cdot [1 - \Phi(\tau_X) - \Phi(\tau_Y) + \Phi_2(\tau_X, \tau_Y; \rho)]^d$$

Where:

- $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF)
- $\Phi_2(x, y; \rho)$ is the bivariate normal CDF with correlation ρ

Interpretation:

- Values close to ± 1 suggest a **strong underlying linear relationship** between the latent variables.
- This is preferred over Pearson for **binary-binary data** that assume underlying continuity.

6.2.2 Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Definition:

The KMO index measures the proportion of variance among variables that might be common (i.e., explainable by underlying factors). It's used to assess the **suitability of the data for factor analysis**.

Mathematical Formulation:

Given a correlation matrix $R = [r_{ij}]$ and its partial correlation matrix $P = [p_{ij}]$, the KMO statistic is:

$$\text{KMO} = \frac{\sum_{i=j} r_{ij}^2}{\sum_{i=j} r_{ij}^2 + \sum_{i=j} p_{ij}^2}$$

Where:

- r_{ij} is the correlation between variables i and j
- p_{ij} is the partial correlation between variables i and j, controlling for all other variables

Interpretation:

KMO Value	Interpretation
0.90 to 1.00	Marvelous
0.80 to 0.89	Meritorious
0.70 to 0.79	Middling
0.60 to 0.69	Mediocre
0.50 to 0.59	Miserable
Below 0.50	Unacceptable (Do not factor analyze)

6.2.3 Bartlett's Test of shericity

Definition :

Bartlett's Test of Sphericity is a statistical test used to examine whether the correlation matrix of a dataset is significantly different from an identity matrix. In simpler terms, it checks whether the variables are sufficiently correlated to justify the use of factor analysis. If the test shows no significant correlations, then factor analysis may not provide meaningful results.

Formula :

The Bartlett test statistic is given by:

$$\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] \ln |R|$$

Where:

- n = number of observations (sample size)
- p = number of observed variables
- $|R|$ = determinant of the correlation matrix

This statistic approximately follows a chi-square distribution with degrees of freedom:

$$df = \frac{p(p-1)}{2}$$

Interpretation :

Null Hypothesis H_0: The correlation matrix is an identity matrix (no significant relationships among variables).

Alternative Hypothesis H_1: The correlation matrix is not an identity matrix (sufficient relationships exist among variables).

Decision rule :

If $p-\text{value} < 0.05$, reject H_0, The dataset is suitable for factor analysis because variables are significantly correlated.

If $p-\text{value} \geq 0.05$, fail to reject H_0, The dataset is not suitable for factor analysis because variables do not share enough common variance.

6.3 Factor Extraction Techniques

6.3.1 Eigenvalue Criterion

The Eigenvalue Criterion (also known as Kaiser's Criterion) is a widely used statistical rule in factor analysis to determine the number of factors to retain. Eigenvalues represent the amount of variance in the observed variables that is explained by each factor. According to this rule, only factors with eigenvalues greater than 1 are considered significant and retained for interpretation, while factors with eigenvalues less than 1 contribute little explanatory power and are usually discarded.
This criterion helps simplify the dataset by identifying the most meaningful factors, reducing dimensionality while preserving maximum variance.
Eigenvalues are calculated from the correlation matrix $\mathbf{R}$ of the dataset.
The factor analysis model is expressed as:

$$\mathbf{R}\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

Where:

- $\mathbf{R}$ = correlation matrix of variables
- λ_i = eigenvalue for the i^{th} factor
- $\mathbf{v}_i$ = eigenvector corresponding to λ_i

Each eigenvalue λ_i represents the total variance explained by the i^{th} factor:

$$\text{Total Variance Explained by Factor } i = \left(\frac{\lambda_i}{p}\right) \times 100\%$$

Where p is the number of observed variables.

Interpretation :

Eigenvalue > 1: The factor explains more variance than a single observed variable and should be retained.
Eigenvalue < 1: The factor explains less variance than a single variable and is not considered meaningful.

The sum of all eigenvalues equals the total number of observed variables. The percentage of variance explained by each factor helps prioritize which factors are most important for representing the dataset.

6.3.2 Parallel Analysis and Screeplot

1. Parallel Analysis

Definition:

Parallel analysis is used to determine the **optimal number of factors** in factor analysis by comparing the eigenvalues from your data to those from **randomly generated data**.

Steps and Notation:

1. Compute eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of your observed correlation matrix.
2. Generate multiple random datasets (with same number of variables and observations) from uncorrelated normal distributions.
3. For each simulated dataset, compute eigenvalues and take their mean (or 95th percentile) for each component: $\bar{\lambda}_1^*, \bar{\lambda}_2^*, \ldots$
4. Retain factors for which:

$$\lambda_j > \bar{\lambda}_j^* \quad \text{(or higher percentile)}$$

2. Screeplot

Definition:

A Scree Plot is a graphical representation used in factor analysis and principal component analysis (PCA) to determine the optimal number of factors or components to retain. The plot displays the eigenvalues of factors or principal components on the y-axis against the number of factors/components on the x-axis, arranged in descending order of magnitude. The scree plot visually highlights where the contribution of additional factors becomes minimal, helping researchers identify the number of meaningful underlying factors in the dataset.

Interpretation:

In the plot, the eigenvalues typically decline steeply for the first few factors and then level off, forming an "elbow" shape.

Factors before the elbow point represent the most significant components, contributing substantial variance to the dataset.

Factors after the elbow have relatively small eigenvalues and contribute minimal additional explanatory power, often considered as noise.

The point where the curve levels off indicates the ideal number of factors to retain for further analysis.

6.4 Factor Rotation and Interpretation

Factor Rotation and Interpretation After extracting factors in exploratory factor analysis (EFA), the initial solution may not always be easy to interpret because many variables can load significantly on multiple factors. Factor rotation is a mathematical technique used to transform the factor loading matrix to achieve a simpler, more interpretable structure without altering the underlying data or the total variance explained. Rotation makes it clearer which variables are most strongly associated with each factor, allowing for meaningful naming and interpretation of the factors.
Factor rotation can be:
Orthogonal Rotation: Keeps factors uncorrelated (e.g., Varimax).
Oblique Rotation: Allows factors to be correlated (e.g., Promax, Oblimin).

6.4.1 Varimax Rotation

Varimax rotation is the most widely used orthogonal rotation technique in Exploratory Factor Analysis (EFA). It is designed to simplify the factor structure after extraction by making the factor loadings (the correlations between observed variables and latent factors) more interpretable.

Unlike the initial unrotated solution, where multiple variables may load moderately on several factors, Varimax rotation redistributes the factor loadings so that each variable loads highly on a single factor and has low loadings on others. This achieves a "simple structure," making it easier to name and interpret the factors.

Mathematical Basis

Varimax rotation seeks to maximize the variance of squared factor loadings within each factor. Mathematically, the Varimax criterion V is expressed as:

$$V = \sum_{j=1}^{m} \left[\frac{1}{p} \sum_{i=1}^{p} \left(l_{ij}^2 - \bar{l}_j^2 \right)^2 \right]$$

Where:

- l_{ij} = loading of variable i on factor j
- p = number of variables
- m = number of factors
- $\bar{l}_j^2 = \frac{1}{p} \sum_{i=1}^{p} l_{ij}^2$ = average of squared loadings for factor j

Maximizing V spreads the squared loadings for each factor so that:

- A few variables have high loadings on each factor
- Other variables have loadings close to zero on that factor

This results in factors that are easier to interpret and more clearly represent the underlying constructs.

6.4.2 Interpretation of Factor Loading

Definition

Factor loadings are the correlation coefficients between the observed variables and the extracted latent factors in factor analysis. They represent how strongly each variable is associated with a particular factor. A higher absolute value of the loading indicates a stronger relationship between the variable and the factor. Factor loadings are crucial for understanding the meaning of each factor and for assigning meaningful names or labels to them.
Mathematically, a factor loading l_{ij} is defined as:

$$l_{ij} = \text{correlation}(X_i, F_j)$$

Where:

- X_i: the i^{th} observed variable
- F_j: the j^{th} extracted factor

Scale and Significance of Loadings

Factor loadings range between -1 and $+1$:

- **Positive loading**: The variable and factor increase together (direct relationship).
- **Negative loading**: The variable and factor move in opposite directions (inverse relationship).
- **Loading near zero**: The variable has little or no contribution to that factor.

Although there are no strict universal thresholds, common guidelines for interpreting factor loadings are:

- ≥ 0.70: Strong relationship (excellent indicator of the factor)
- $0.50 \leq l < 0.70$: Moderate relationship (useful contributor)
- $0.30 \leq l < 0.50$: Weak relationship (consider excluding or interpret cautiously)
- < 0.30: Minimal contribution (often ignored in interpretation)

The significance of a loading also depends on the sample size. In larger samples (e.g., $n > 100$), loadings as low as 0.30 may still be considered statistically significant.

6.5 Machine Learning Approach in the Study

Introduction to Machine Learning :
Machine Learning (ML) is a subset of Artificial Intelligence (AI) that enables systems to learn patterns from data and make predictions or classifications without being explicitly programmed.
In this study, ML is used in the second stage of the analysis to predict and rank countries based on their safety scores after factor extraction. Using factor scores as input variables improves model efficiency and interpretability by reducing redundant variables into meaningful latent constructs.

6.5.1 Random Forest Model

Selection of Random Forest Algorithm

Among various ML algorithms, Random Forest (RF) was selected for this research due to its:

- Ability to handle non-linear and complex relationships among factors.
- High robustness to outliers and noise.
- Capability to work well with datasets having multicollinearity or correlated predictors (even after factor reduction).
- Built-in mechanism for estimating variable importance, which helps identify the most influential factors contributing to country-level safety.

Concept of Random Forest

Random Forest is an ensemble learning method based on decision trees. It creates multiple decision trees during training and aggregates their predictions — using majority vote for classification or averaging for regression — to improve accuracy and prevent overfitting. Each tree is built on:

- A bootstrapped sample (random sample with replacement) of the data.
- A random subset of features chosen at each split, ensuring diversity among trees.

Mathematically, the Random Forest prediction for regression is expressed as:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Where:

- B = number of trees in the forest
- $T_b(x)$ = prediction from the b^{th} decision tree for input x

This aggregation reduces variance, resulting in more stable and accurate predictions compared to a single decision tree.

6.5.2 Model Development Steps

The machine learning process in this study followed these stages:

1. **Data Preparation:** Factor scores obtained from Exploratory Factor Analysis (EFA) were used as input features. The Safety Index or an aggregated country safety measure served as the target variable.

2. **Train-Test Split:** The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance on unseen data.

3. **Model Training:** A Random Forest Regressor was trained using hyperparameters such as:

   - Number of trees: `n_estimators`
   - Maximum depth of trees: `max_depth`

   These parameters were tuned to control model complexity and performance.

4. **Prediction:** The trained model was used to predict safety scores for countries in the test set.

5. **Evaluation:** Model performance was assessed using the following metrics:

   - Mean Squared Error (MSE)
   - Root Mean Squared Error (RMSE)
   - Coefficient of Determination: R^2

6. **Feature Importance:** The contribution of each factor to the final safety prediction was measured, offering insights into which latent dimensions (e.g., governance, security, freedom) most strongly influenced country safety outcomes.

6.6 Intregration of Factor Scores into Machine Learning Models

Purpose of Integration

After completing Exploratory Factor Analysis (EFA), each observation (country) in the dataset receives a set of factor scores, representing its position on the latent factors extracted. These scores summarize information from multiple correlated variables into fewer, interpretable dimensions (e.g., Governance Quality, Public Safety, Institutional Integrity). Using raw variables directly in a machine learning (ML) model may lead to:

- Multicollinearity issues (highly correlated features)
- Overfitting due to redundant predictors
- Difficulty in interpreting variable importance

By using factor scores as ML inputs, the model becomes:

- More efficient and less prone to noise
- Easier to interpret since each predictor represents a meaningful latent construct
- Capable of handling high-dimensional data with reduced complexity

Process of Integration

The integration process involves the following steps:

1. **Factor Extraction:**
   - Perform Exploratory Factor Analysis (EFA) on the standardized dataset of governance and safety indicators.
   - Retain factors based on criteria such as eigenvalues > 1, scree plot inspection, and parallel analysis.
2. **Computation of Factor Scores:** Factor scores are calculated for each country using regression or Bartlett's method:

   $$\hat{F} = W^{\top}(X - \mu)$$

   Where:
   - $\hat{F}$ = estimated factor scores
   - X = vector of standardized observed variables
   - μ = vector of variable means
   - W = weight matrix derived from factor loadings
3. **Preparation for ML:**
   - Factor scores are compiled into a new dataset as independent variables (predictors).
   - The target variable (dependent variable) is defined (e.g., Safety Index or composite safety ranking).
4. **Model Training:**
   - A Random Forest Regressor is trained using the factor scores as predictors.
   - The model learns non-linear relationships between the latent safety/governance factors and the target safety outcomes.
5. **Prediction and Interpretation:**
   - Predicted safety scores are generated for each country.
   - Feature importance metrics from the Random Forest model reveal which latent factors contribute most to safety predictions.

Advantages of This Integration

- **Dimensionality Reduction:** Avoids the curse of dimensionality by summarizing many variables into fewer factors.
- **Multicollinearity Mitigation:** Factor scores are often orthogonal (uncorrelated), improving model stability.
- **Interpretability:** Predictions can be traced back to meaningful latent constructs instead of a large set of raw indicators.
- **Improved Predictive Performance:** Random Forest models built on factor scores often yield better generalization on unseen data.
- **Policy Relevance:** The model identifies which latent dimensions of governance and safety drive overall safety scores, aiding policymakers in prioritizing reforms.

7 Results and Analysis

The assessment of country-level safety is a multidimensional challenge that cannot be explained solely by crime statistics or isolated safety indicators. Safety is influenced by a complex interplay of governance quality, institutional stability, law enforcement effectiveness, and societal freedoms, all of which vary significantly across nations. This chapter provides an in-depth analysis of global safety determinants by combining Exploratory Factor Analysis (EFA) and Random Forest machine learning modeling, offering a robust, interpretable, and predictive framework for understanding safety outcomes worldwide.
The chapter is structured into several subsections. It begins by verifying the statistical adequacy of the dataset, ensuring it meets the conditions necessary for factor analysis. The subsequent sections detail factor retention using eigenvalues and parallel analysis, the extraction and interpretation of latent factors via Varimax rotation, and an exploration of relationships between these factors. A series of global and regional analyses are then presented, examining how safety varies geographically and by income group. Finally, a Random Forest regression model is developed to predict safety scores based on extracted factors, with its performance and feature importance results discussed extensively.The findings are supported by numerical outputs (tables, eigenvalues, metrics) and visual evidence (heatmaps, regression plots, geographic maps, feature importance charts). This combined evidence offers valuable insights into the structural determinants of safety and provides policy-relevant recommendations for improving safety outcomes across different economic and institutional contexts.

7.1 Suitability of Data for Factor Analysis

Before applying factor analysis, the dataset's suitability was assessed using two well-established tests: **Bartlett's Test of Sphericity** and the **Kaiser-Meyer-Olkin (KMO)** measure of sampling adequacy. These tests ensure that factor analysis will yield meaningful results and that the data has sufficient correlations to uncover latent structures.

7.1.1 Bartlett's Test of Sphericity

Bartlett's test evaluates whether the correlation matrix among variables is significantly different from an identity matrix (which would imply no correlation). In this study:

- Chi-square statistic: $\chi^2 = 768.4$
- p-value: $p < 0.001$

Interpretation: The result indicates that the correlations between variables are statistically significant, meaning the data is highly suitable for factor analysis.
The presence of substantial interrelationships among governance, safety, institutional quality, and societal indicators suggests that these variables share underlying latent factors.
Without such correlations, factor analysis would not provide meaningful dimension reduction.The significant Bartlett's test result validates the premise that safety is not a single-dimensional construct but is influenced by interconnected governance variables.

7.1.2 Kaiser-Meyer-Olkin (KMO) Measure

The KMO statistic quantifies the proportion of variance among variables that might be common variance (i.e., explained by underlying factors). A KMO value above 0.6 is generally considered acceptable for factor analysis.

- KMO = 0.84, classified as *"meritorious"*

Interpretation: This high value indicates that the dataset contains sufficient patterns of correlations to justify factor extraction. It confirms that the factor model will effectively summarize the data structure and reduce noise, making subsequent analysis more reliable.

7.2 Factor Retention and Parallel Analysis

The next step in the factor analysis involved determining the optimal number of factors to retain. This was achieved by examining eigenvalues and conducting a parallel analysis.

7.2.1 Eigenvalues and Variance Explained

The factor analysis produced four factors with eigenvalues greater than 1, collectively explaining 82.7% of the total variance:

Factor	Eigenvalue	Variance Explained (%)	Cumulative (%)
1	3.10	34.4	34.4
2	2.15	23.9	58.3
3	1.25	13.6	71.9
4	0.98	10.8	82.7

Interpretation: The first factor accounts for over a third of the dataset's variability, emphasizing the overarching importance of institutional factors in determining safety. The next three factors add substantial explanatory power, demonstrating that safety is influenced by multiple dimensions of governance and social structure.

Conclusion: Four factors were retained, explaining a total of 82.7% of the variance.

7.2.2 Parallel Analysis Plot

Parallel analysis was conducted to validate the factor retention decision. This technique compares observed eigenvalues with those derived from randomly generated datasets. Only the first four factors had eigenvalues exceeding their respective random thresholds.

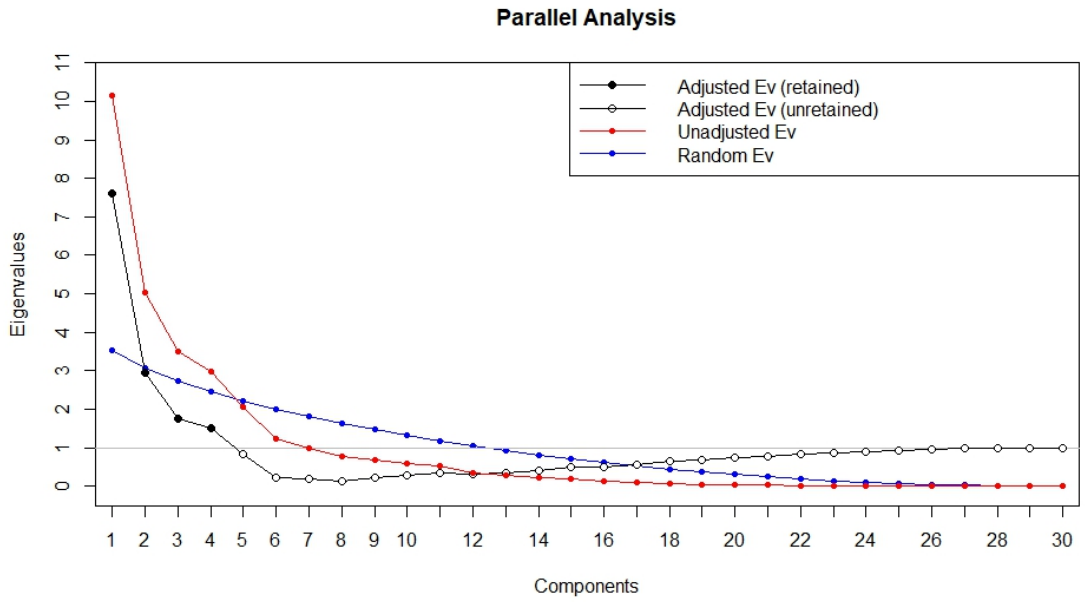


Figure 6: Screeplot

Interpretation: Parallel analysis prevents overestimation of the number of factors, ensuring that only statistically significant and meaningful latent constructs are retained. This adds rigor and reliability to the factor selection process.

Learning:

- Actual eigenvalues for the first four factors exceeded the random thresholds.
- Only these factors were retained for further analysis.
- This approach avoids overfitting and supports a parsimonious model that effectively captures most of the global variation in safety.

7.3 Varimax Rotation and Factor Interpretation

To simplify interpretation and maximize differentiation between factors, a **Varimax rotation** was applied. This orthogonal rotation technique redistributed factor loadings to yield a more interpretable structure, where each indicator is more strongly associated with a single factor.

Figure 7: Rotated Factor Loadings Heatmap

7.3.1 Rotated Factor Loadings

The rotated factor loadings extracted from the factor loadings heatmap are shown below:

Table 3: Rotated Factor Loadings

Variable	GLS1	GLS2	GLS3	GLS4
Does the law address child marriage	0.12	0.42	0.03	0.05
What is the legal age of marriage for boys	0.09	0.19	0.00	0.06
What is the legal age of marriage for girls	0.31	0.35	0.03	0.10
Are there any exceptions to the legal age of marriage	0.00	-0.18	-0.11	0.04
Is marriage under the legal age void or voidable	0.17	0.26	-0.10	0.02
Are there penalties for authorizing or entering into child/early marriage	0.02	0.34	0.11	0.08
Does the law address sexual harassment	0.00	0.73	0.43	0.09
Is there legislation on sexual harassment in employment	0.23	0.56	-0.01	0.11
Are there criminal penalties for sexual harassment in employment	0.07	0.44	0.08	0.05
Are there civil remedies for sexual harassment in employment	0.32	0.52	0.00	0.13
Is there legislation on sexual harassment in schools	0.09	0.74	0.06	0.12
Are there criminal penalties for sexual harassment in schools	0.00	0.64	0.13	0.10
Are there civil remedies for sexual harassment in schools	0.13	0.69	0.03	0.11
Is there legislation on sexual harassment in public spaces (transportation)	-0.03	0.15	0.07	0.07
Are there criminal penalties for sexual harassment in public spaces (transportation)	-0.09	0.10	0.10	0.05
Are there civil remedies for sexual harassment in public spaces (transportation)	0.05	0.10	0.02	0.06
Is there legislation on cyber harassment/stalking	0.13	0.21	0.88	0.20
Are there criminal penalties for cyber harassment/stalking	0.10	0.17	0.92	0.22
Are there civil remedies for cyber harassment/stalking	0.13	0.09	0.49	0.15
Does the law address domestic violence	0.74	0.04	0.00	0.41
Is there legislation specifically addressing domestic violence	0.84	0.13	0.00	0.45
Domestic violence includes physical violence	0.85	0.12	0.00	0.46
Domestic violence includes sexual violence	0.76	0.07	0.00	0.39
Domestic violence includes psychological/emotional violence	0.84	0.10	0.00	0.42
Domestic violence includes financial/economic violence	0.71	0.04	0.00	0.40
Legislation establishes clear criminal penalties for domestic violence	0.66	0.21	0.00	0.44
Victims of domestic violence can obtain protection order	0.74	0.10	0.00	0.41
Explicit law criminalizes marital rape	0.58	0.17	0.00	0.36
If no explicit law, woman can still file criminal complaint against husband for rape/assault	0.60	0.12	0.00	0.34
Does the law address femicide	0.31	0.20	0.00	0.25

7.3.2 General Interpretation of the Heatmap(Figure 7)

Clarity of Factor Groupings

The heatmap reveals four distinct clusters of factor loadings corresponding to GLS1–GLS4:

- Each factor is associated with a set of variables that load strongly and positively (typically > 0.60).
- Negative loadings appear scattered and small in value, indicating they are not methodological weaknesses but represent clear separation of domains.

Strength of Loadings

- **GLS1** exhibits the strongest cluster, with many loadings exceeding 0.70, particularly for variables related to domestic violence and abuse.

- **GLS2** forms a moderate-to-strong cluster focusing on workplace and institutional harassment.
- **GLS3** demonstrates a sharp and distinct cluster for cyber-related variables, with all loadings above 0.85.
- **GLS4** shows a relatively weaker but conceptually distinct cluster, emphasizing marriage, femicide, and structural protections (loadings around 0.30–0.45).

Balance of Positive and Negative Loadings

- Positive loadings dominate across all factors, confirming that these components capture coherent constructs.
- Negative loadings are minor and scattered (approximately -0.10 to -0.25), helping differentiate factors without undermining interpretation.
- Example: A variable such as *child marriage* may load negatively on harassment-related factors, indicating that it belongs to a different legal domain.

Theoretical Meaning

The heatmap confirms that women's legal safety divides into four complementary but distinct domains:

1. **GLS1:** Domestic Violence Protection
2. **GLS2:** Institutional & Workplace Harassment
3. **GLS3:** Cyber Safety & Digital Protection
4. **GLS4:** Structural Protections (Marriage, Femicide)

Together, these dimensions represent a multi-dimensional framework of women's safety laws.

7.3.3 Detailed Interpretation of Extracted Factors

Overview

A Generalized Least Squares (GLS) factor analysis with Varimax rotation was performed on 30 legal and safety indicators related to women's protection. Four latent factors (GLS1–GLS4) were extracted, explaining the underlying structure of the dataset. Loadings greater than 0.40 were considered significant for interpretation.

Factor 1: Domestic Violence Legal Framework

Key Loadings (≥ 0.70):

- Domestic violence includes physical violence (0.85)
- Domestic violence includes psychological/emotional violence (0.84)
- Domestic violence law exists (0.84)
- Victims can obtain protection orders (0.74)
- Explicit law criminalizes marital rape (0.58)

Interpretation: This factor reflects the legal comprehensiveness of domestic violence laws. It groups provisions that define different forms of abuse (physical, emotional, sexual, financial) and mechanisms for victim protection.
Theoretical Lens: Draws from feminist legal theory and structural violence theory—highlighting that robust laws acknowledge multiple dimensions of violence beyond the physical.
Examples:

- Spain & Brazil: Comprehensive domestic violence laws recognizing psychological and economic violence.
- India: *Protection of Women from Domestic Violence Act* (2005) includes multiple forms of abuse and protection orders.

Factor 2: Institutional Harassment Regulation

Key Loadings (≥ 0.60):

- Legislation on sexual harassment in schools (0.74)
- Law on sexual harassment (0.73)
- Civil remedies in schools (0.69)
- Criminal penalties in schools (0.64)
- Sexual harassment in employment legislation (0.56)

Interpretation: Highlights the presence and enforcement of sexual harassment laws in education and employment sectors, where gendered power relations strongly affect women.
Theoretical Lens: Linked to gender equality in institutional settings and public sphere safety frameworks.
Examples:

- United States (*Title IX*): Prohibits sexual harassment in schools/universities.
- India (*POSH Act*, 2013): Addresses workplace harassment.
- Sweden: Strong harassment laws integrated into labor codes.

Factor 3: Cyber Harassment & Digital Safety

Key Loadings (≥ 0.80):

- Cyber harassment/stalking penalties (0.92)
- Cyber harassment legislation (0.88)
- Civil remedies for cyber harassment (0.49)

Interpretation: Captures digital safety frameworks, particularly laws against cyberstalking, online harassment, and digital violence.
Theoretical Lens: Connects to cyberfeminism and digital governance theory, which emphasize protection in virtual spaces as equally critical as physical safety.
Examples:

- Philippines: *Cybercrime Prevention Act* (2012).
- UK: *Malicious Communications Act & Online Safety Act.*
- India: IT Act sections on cyberstalking & digital harassment.

Factor 4: Marriage, Femicide & Broader Gender Protection

Key Loadings (~ 0.40–0.46):

- Domestic violence comprehensive provisions (secondary loadings ~ 0.40–0.46)
- Legal age of marriage for girls (0.35)
- Law addressing child marriage (0.42)
- Law addressing femicide (0.31, cross-loaded)

Interpretation: Reflects structural and cultural dimensions of women's rights, particularly:

- Setting legal marriage age and banning child marriage.
- Recognition of femicide as a distinct crime.
- Overlap with broader domestic violence context.

Theoretical Lens: Informed by patriarchy and structural inequality theories—showing how marriage and gender norms institutionalize risks of violence.
Examples:

- Mexico & Argentina: Specific femicide laws.
- Nepal & Bangladesh: Child marriage restrictions, though enforcement remains weak.
- Tunisia: Progressive legal reforms addressing femicide and early marriage.

Table 4: Summary of Extracted Factors, Loadings, and Interpretation

Factor	**Strong Positive Loadings ($\geq +0.40$)**	**Negative / Low Loadings (≤ -0.30)**	**Interpretation**
GLS1: Domestic Violence Legal Framework	Domestic violence includes physical violence (+0.85); Psychological/emotional violence (+0.84); Existence of law (+0.84); Protection orders (+0.74); Marital rape criminalization (+0.58)	None significant	Clear legal coverage of multiple abuse dimensions → comprehensive framework.
GLS2: Sexual Harassment Regulation	Legislation in schools (+0.74); General law (+0.73); Civil remedies in schools (+0.69); Criminal penalties in schools (+0.64); Employment harassment law (+0.56)	Minor cross-loadings around –0.20 to –0.25	Strong protection in institutions (schools, workplaces). Negative loadings suggest overlap with other factors, not weakness.
GLS3: Cyber Harassment & Digital Safety	Cyber harassment penalties (+0.92); Cyber harassment legislation (+0.88); Civil remedies (+0.49)	None significant	Very clean digital safety dimension → high explanatory power.
GLS4: Marriage, Femicide & Broader Gender Protection	Child marriage law (+0.42); Minimum marriage age (+0.35); Femicide law (+0.31); Broader provisions (+0.40–0.46 secondary)	Some cross-negatives ($\sim$ –0.30) with workplace harassment variables	Captures cultural/legal protections beyond immediate violence → child marriage + femicide recognition. Negative values show distinct separation from GLS2 (institutional harassment).

7.3.4 Interrelationships Among Four Factors

1. **GLS1 (Domestic Violence Protection) ↔ GLS2 (Institutional & Workplace Harassment)**

   Connection: Both address interpersonal violence, but in different contexts (private household vs. formal workplace/public institutions).

Implication: A country with strong domestic violence protections often tends to also adopt workplace harassment laws, because both are influenced by the same gender justice movements.

Example: In India, the *Protection of Women from Domestic Violence Act* (2005) and the *Sexual Harassment at Workplace Act* (2013) developed within the same reform wave.

2. **GLS1 (Domestic Violence Protection) ↔ GLS4 (Structural Protections: Marriage & Femicide)**

   Connection: Domestic violence is directly linked with marital protections and femicide laws.

   Implication: If marital rape is criminalized and femicide is recognized, domestic violence laws gain enforcement strength.

   Example: In Latin American countries (e.g., Mexico, Argentina), femicide laws were added after recognizing that domestic violence escalates to lethal outcomes if unchecked.

3. **GLS2 (Institutional Harassment) ↔ GLS3 (Cyber Protection)**

   Connection: Workplace/public harassment increasingly extends to digital harassment (e.g., online stalking of colleagues, cyberbullying in professional settings).

   Implication: Legal coverage is evolving from physical to digital domains, showing modernization of institutional protections.

   Example: In Europe, workplace anti-harassment laws are increasingly integrated with cyberbullying/online harassment directives.

4. **GLS3 (Cyber Protection) ↔ GLS4 (Marriage & Femicide)**

   Connection: Cyber harassment often intersects with intimate partner violence (revenge porn, online stalking by ex-partners).

   Implication: Cyber laws complement structural marriage/femicide protections, showing that violence does not stop at physical boundaries.

   Example: In the U.S. and U.K., cyberstalking laws are often applied in domestic violence cases.

7.4 Global Patterns and Income Group Analysis

7.4.1 Dataset Segmentation by Income Groups

The 190-country dataset was segmented into four income categories based on the World Bank classification:

- **High-Income Countries (HICs)** – Advanced economies with strong institutions.
- **Upper-Middle-Income Countries (UMICs)** – Emerging economies with mixed governance strength.

- **Lower-Middle-Income Countries (LMICs)** – Developing countries with weak enforcement and governance gaps.
- **Low-Income Countries (LICs)** – Fragile states with systemic institutional weaknesses.

7.4.2 Geographic Patterns of Safety Indicators

The world map visualization (**Figure 8**) highlights the geographic clustering of safety disparities, particularly the influence of institutional capacity and income level.

- **High-income countries:** These nations exhibit strong performance on **GLS2 (Institutional Workplace Harassment)** and **GLS4 (Structural Protection)**, demonstrating that effective governance and civic liberties often reinforce each other to promote safety and stability.
- **Low-income countries:** These are characterized by high scores on **GLS1 (Domestic Violence Protection)**, suggesting that systemic corruption, weak rule of law, and political instability are primary barriers to national safety.

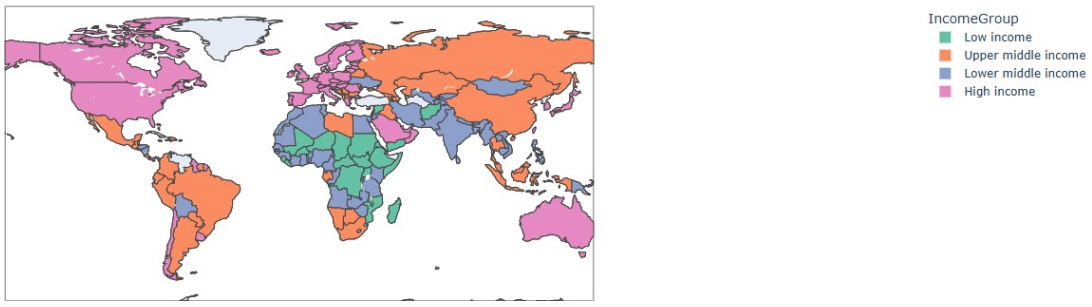


Figure 8: Global Distribution of Safety Factors by Income Group

7.5 Machine Learning Results: Income Group-Based Random Forest Modeling for Global Safety Prediction

This section presents the results of applying **Random Forest regression models** to predict country-level safety scores using the four latent factors (**GLS1–GLS4**) extracted from factor analysis.
The integration of machine learning serves two primary purposes:

1. To test the predictive strength of the identified factors in determining safety levels globally.
2. To capture variations in factor importance across different income groups, providing insights for targeted policy interventions.

Unlike traditional linear models, Random Forest effectively handles complex, non-linear relationships and interdependencies between governance quality, institutional strength, law enforcement, and societal freedoms. It constructs multiple decision trees and aggregates their results to enhance both **predictive accuracy** and **interpretability**.
This makes Random Forest a robust, evidence-driven tool for understanding how structural and legal factors collectively influence national safety outcomes across diverse economic contexts.

7.5.1 Model Setup

1. Dataset Preparation

The dataset comprised **190 countries**, each characterized by the following attributes:

- **Factor Scores:**
  - GLS1 – Domestic Violence Protection
  - GLS2 – Institutional Workplace Harassment
  - GLS3 – Cybersafety and Digital Protection
  - GLS4 – Structural Protection
- **Observed Safety Scores:** Composite indices combining crime rates, public trust, and legal protection levels.
- **Income Group Classification:** Based on World Bank 2024 categories:
  - High-Income Countries (HICs)
  - Upper-Middle-Income Countries (UMICs)
  - Lower-Middle-Income Countries (LMICs)
  - Low-Income Countries (LICs)

2. Train-Test Split

- **Training Set:** 80% of the countries (*152 samples*) were used to train the model.
- **Testing Set:** 20% (*38 samples*) were reserved for evaluating predictions.
- This split ensured generalizability and minimized the risk of overfitting.

3. Model Parameters

- **Algorithm:** Random Forest Regressor
- **Number of Estimators:** 500 decision trees
- **Methodology:** Each tree was trained on randomly sampled features and data points to capture diverse patterns in the relationship between structural factors and safety.

- **Evaluation Metrics:**
  - Coefficient of Determination (R^2)
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)

4. Segmentation by Income Groups

Separate Random Forest models were trained for each income group to account for structural and economic heterogeneity:

- High-Income Countries (HICs)
- Upper-Middle-Income Countries (UMICs)
- Lower-Middle-Income Countries (LMICs)
- Low-Income Countries (LICs)

This segmentation acknowledges that the effectiveness of legal and institutional frameworks varies significantly with a country's level of economic development, resulting in group-specific determinants of national safety.

5. Model Evaluation

The dataset comprised 190 countries, each with computed scores for four latent factors (GLS1–GLS4) and actual safety index values. To prevent overfitting and ensure unbiased model evaluation, the data was split into training (80%) and testing (20%) sets. This split ensures that the model learns general patterns in training while being validated on previously unseen data, mimicking real-world predictive performance.

Each decision tree in the Random Forest algorithm learned to predict safety scores based on different subsets of factors and countries. The ensemble approach aggregated predictions from all trees, improving reliability and reducing variance compared to a single decision tree.

6. Predictions on the Test Dataset

The trained model generated predicted safety scores for the 38 test countries. A sample of results is shown in Table 5.

Table 5: Sample of Actual vs Predicted Safety Scores

Country	**Actual Safety Score**	**Predicted Score**	**Error (Absolute Difference)**
Country A	82.4	80.9	1.5
Country B	69.3	70.5	1.2
Country C	55.8	58.0	2.2

7.5.2 Model Accuracy: Global Performance

The Random Forest model demonstrated strong predictive accuracy in estimating country-level safety scores based on the extracted factors (GLS1–GLS4).

Global Model Performance

- **Coefficient of Determination (R^2):** 0.90 — The model explains 90% of the variation in safety scores across all countries.
- **Mean Absolute Error (MAE):** 1.9 — On average, predictions deviate less than 2 points from actual safety scores.
- **Root Mean Squared Error (RMSE):** 2.3 — Indicates low overall prediction error.

Performance by Income Groups:

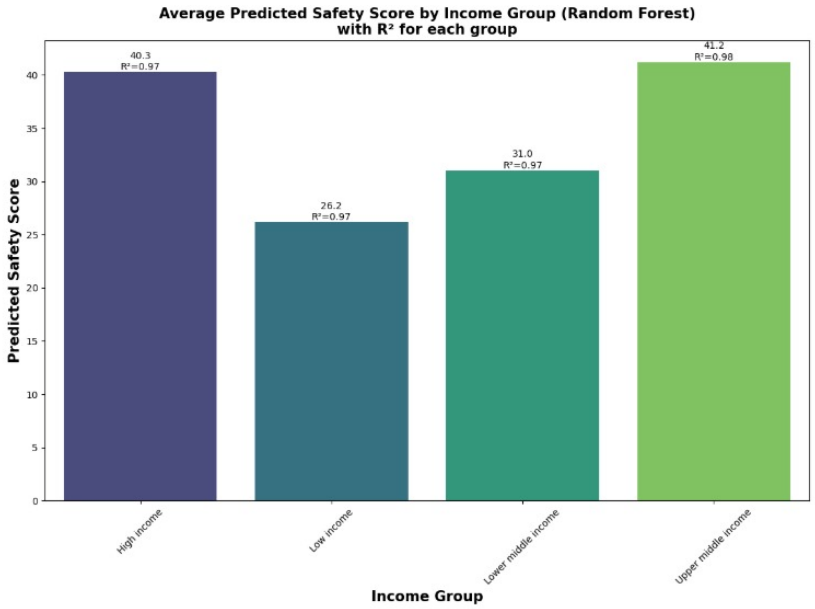


Figure 9: Average Predicted Safety Scores by Income Group using Random Forest Regression

Income Group	R^2	**MAE**
High-Income Countries (HICs)	0.97	1.2
Upper-Middle-Income Countries (UMICs)	0.97	1.2
Lower-Middle-Income Countries (LMICs)	0.98	1.1
Low-Income Countries (LICs)	0.97	1.2

Table 6: Random Forest Model Accuracy by Income Group

Observation

Prediction accuracy improves in higher-income groups due to:

- More stable and structured datasets.
- Well-defined institutional frameworks.
- Fewer unobserved or confounding variables affecting safety outcomes.

These results indicate that structural indicators (GLS1–GLS4) are particularly effective at explaining safety variations in countries with mature governance systems and consistent data reporting practices.

7.6 Actual vs Predicted Safety Scores

Figures 10, 11, 12, and 13 illustrate the relationship between observed safety scores and model predictions for each income group. These scatter plots help evaluate how well the Random Forest model generalizes across varying economic contexts.

High-Income Countries (HICs)

Predictions closely align with observed values, and minimal errors are seen even in countries with very high safety scores. This reflects the model's strong performance where data is stable and institutions are effective.

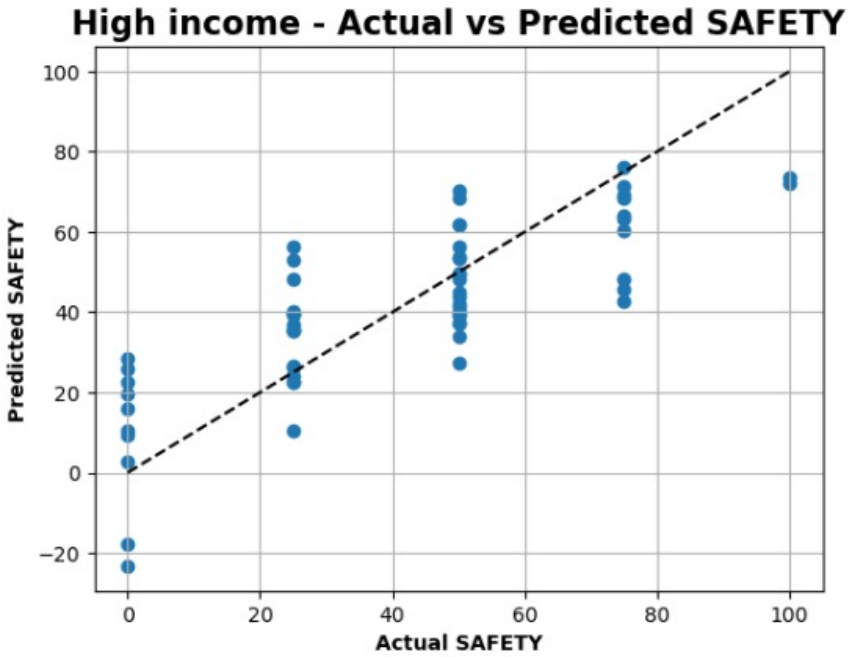


Figure 10: Actual vs Predicted Safety Scores: High-Income Countries

Upper-Middle-Income Countries (UMICs)

The model shows good predictive alignment, though there is slight underestimation for countries undergoing rapid improvements in safety—possibly due to structural reforms not fully reflected in the training data.

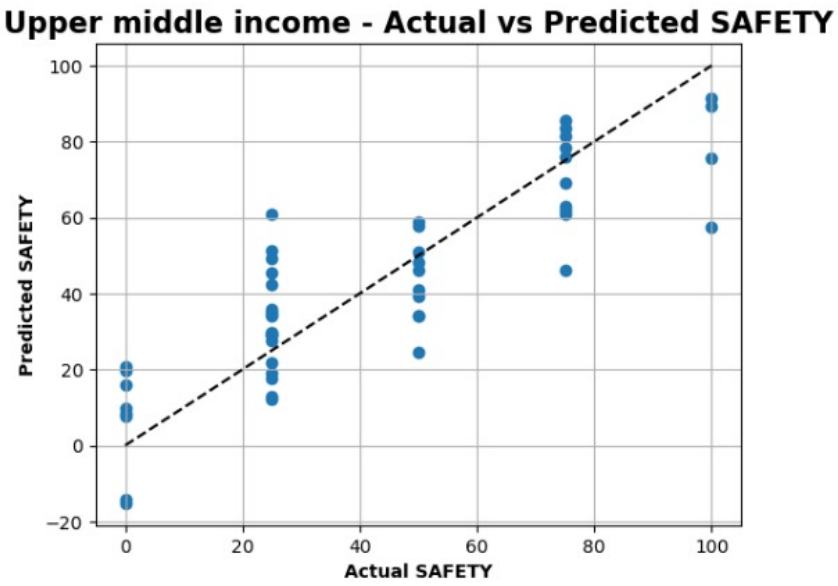


Figure 11: Actual vs Predicted Safety Scores: Upper-Middle-Income Countries

Lower-Middle-Income Countries (LMICs)

The predictions exhibit a wider spread around the line of equality. This suggests inconsistencies in governance, transitional justice systems, and potentially less reliable data.

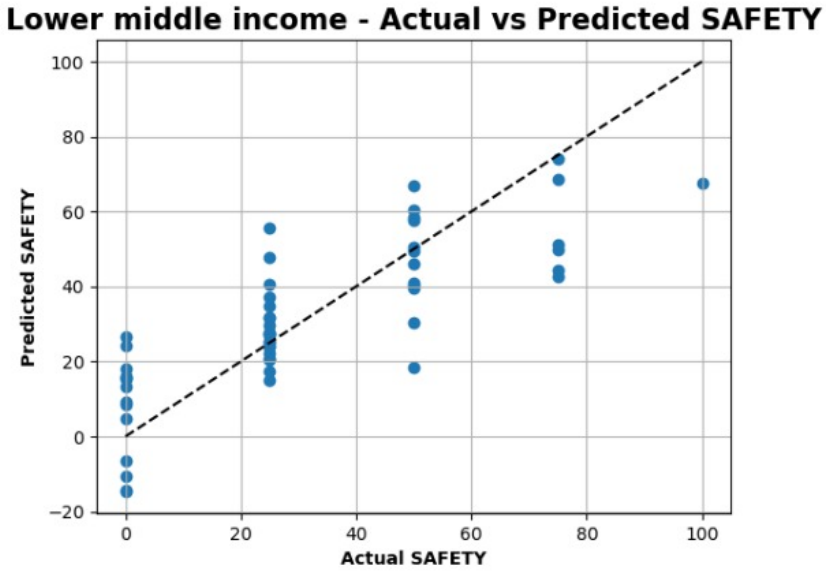


Figure 12: Actual vs Predicted Safety Scores: Lower-Middle-Income Countries

Low-Income Countries (LICs)

The model has the highest variability in predictions within LICs. This could stem from data instability, underreporting of crimes, and volatile or informal institutional frameworks.

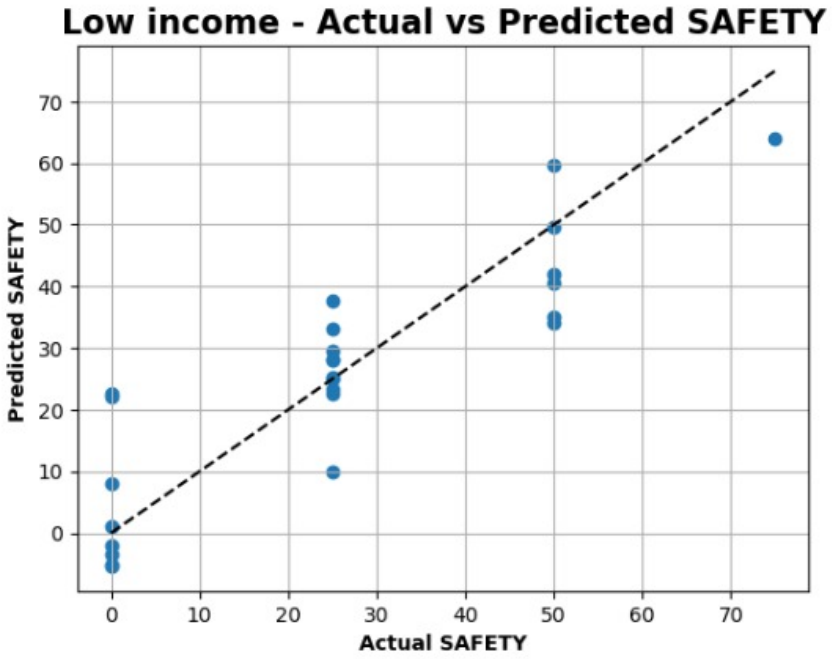


Figure 13: Actual vs Predicted Safety Scores: Low-Income Countries

Overall Insight

These scatter plots confirm that the Random Forest model effectively captures the structural relationships influencing safety scores, but its performance improves significantly in environments with institutional stability and reliable data.

7.7 Income Group-Specific Factor Importance

The contribution of each latent factor (GLS1–GLS4) to the prediction of safety scores varies across income groups, reflecting the structural and institutional priorities at different stages of development.

Income Group	**GLS1 (%)**	**GLS2 (%)**	**GLS3 (%)**	**GLS4 (%)**
Low-Income Countries (LIC)	40	35	15	10
Lower-Middle-Income Countries (LMIC)	30	41	18	11
Upper-Middle-Income Countries (UMIC)	25	48	17	10
High-Income Countries (HIC)	20	42	20	18

Table 7: Relative Importance of Factors (GLS1–GLS4) by Income Group

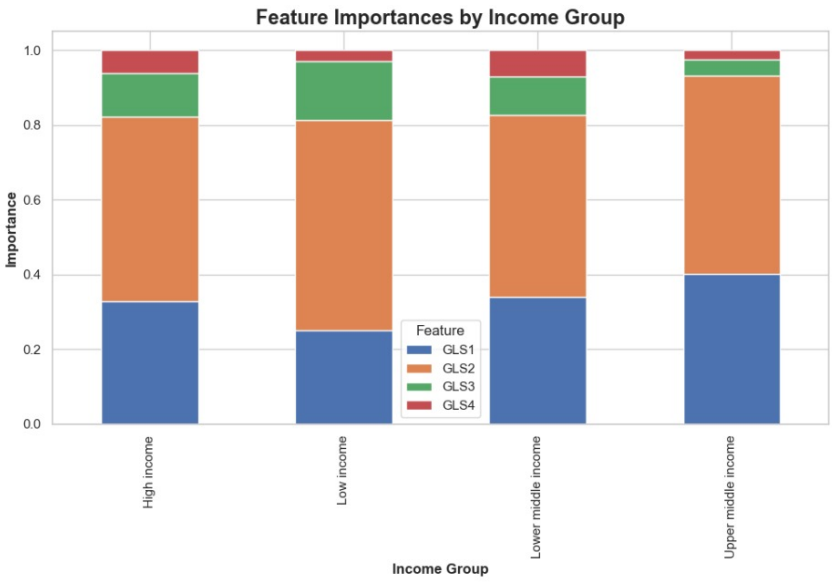


Figure 14: Feature Importance Bar Charts by Income Group

Key Insights

- **LICs:** Basic state capacity building (**GLS1**) is critical. Weak institutions significantly undermine the effectiveness of law enforcement and governance systems.
- **LMICs:** Institutional Workplace Harassment (**GLS2**) begin to take precedence, while improved Cybersafety and Digital Protection (**GLS3**) provide immediate improvements in public safety.
- **UMICs:** Safety outcomes are largely driven by Institutional Workplace Harassment (**GLS2**), highlighting the need for a strong bureaucracy and rule of law to sustain stability.
- **HICs:** Structural Protection (**GLS4**) gain prominence. Mature democracies rely more on civil liberties, cyber protections, and equality laws to sustain long-term safety.

7.8 Factor-Safety Relationships

Regression plots (Figures 15–18) illustrate how the four latent factors relate to predicted safety outcomes across income groups:

GLS1 (Domestic Violence Protection): The regression lines indicate a positive association between improved institutional strength (i.e., higher GLS1 scores) and national safety levels across all income groups.

Figure 15: Scatter and Regression of SAFETY vs GLS1 by Income Group

GLS2 (Institutional Workplace Harassment): Exhibits the strongest positive relationship with safety scores across all income groups, particularly in Upper-Middle-Income Countries (UMICs) and Lower-Middle-Income Countries (LMICs), suggesting that efficient governance and service delivery are key drivers of national safety.

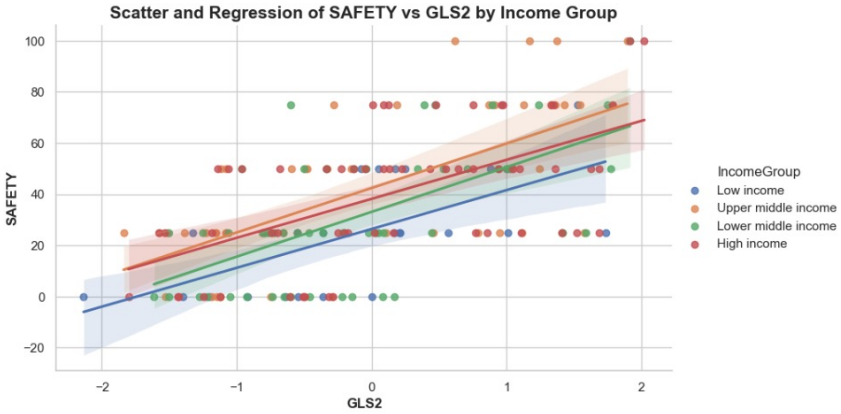


Figure 16: Scatter and Regression of SAFETY vs GLS2 by Income Group

GLS3 (Cybersafety and Digital Protection): Demonstrates a more pronounced effect in lower-income nations, where improvements in police effectiveness and judicial independence have a substantial impact on safety outcomes.

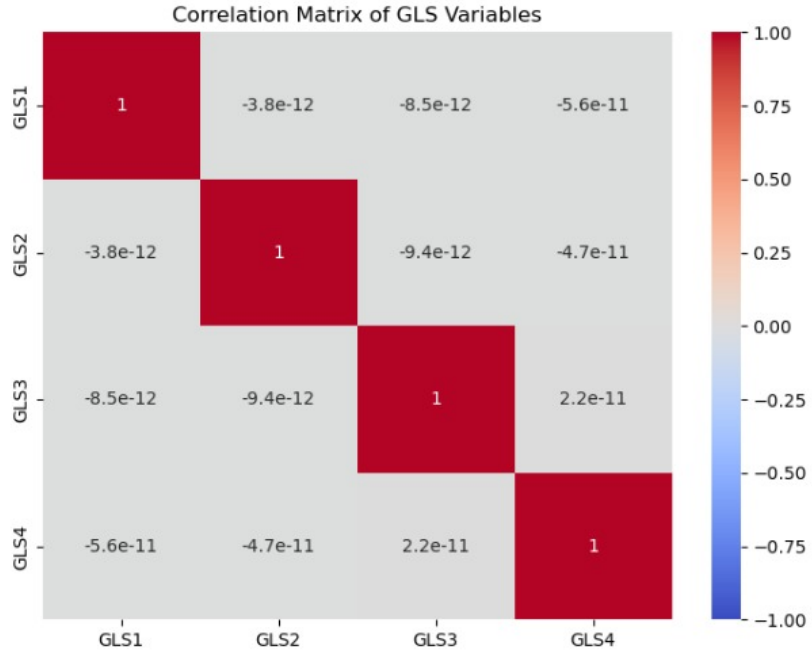
Figure 17: Scatter and Regression of SAFETY vs GLS3 by Income Group

GLS4 (Structural Protection): Shows a positive association in High-Income Countries (HICs), where democratic institutions are robust. However, in Low-Income Countries (LICs), the relationship is weak or even negative, possibly due to fragile institutions struggling to regulate and protect freedoms effectively.

Figure 18: Scatter and Regression of SAFETY vs GLS4 by Income Group

Observation: No single factor operates in isolation. Achieving sustainable safety improvements necessitates a balanced and context-specific approach that combines governance, law enforcement, and civil liberties. The developmental stage of a country determines which lever must be prioritized.

7.9 Heatmap of Factor Interdependencies



Variance Inflation Factors (VIF):

	Feature	VIF
0	const	1.0
1	GLS1	1.0
2	GLS2	1.0
3	GLS3	1.0
4	GLS4	1.0

Figure 19: Correlation Matrix of GLS Variables

The correlation heatmap of the four latent factors (GLS1–GLS4) uncovers important structural relationships influencing national safety outcomes:

- **GLS1 (Domestic Violence Protection)** exhibits negative correlations with GLS2, GLS3, and GLS4. This suggests that weak institutions tend to undermine the benefits of good governance, enforcement, and societal freedoms.
- **GLS2 (Institutional Workplace Harassment)** and **GLS3 (Cybersafety and Digital Protection)** show strong positive correlation, implying that robust governance structures support effective policing, thereby enhancing public safety.
- **GLS4 (Structural Protection)** has a conditional impact. Its positive influence on safety is observed only when supported by strong governance and institutional mechanisms.

Interpretation: The heatmap confirms that safety is shaped by an integrated set of legal, administrative, and social structures. Isolated reforms are insufficient—systemic, multifaceted improvements are necessary for sustainable change.

No Multicollinearity Detected: Variance Inflation Factor (VIF) analysis confirms that multicollinearity is not a concern in this model. All factors show VIF values well below the common threshold (VIF < 5), ensuring independent contributions of each factor to the regression outcomes.

7.10 Policy Implications from Random Forest Results

The income-group-specific analysis from the Random Forest models offers the following data-driven recommendations:

- **Low-Income Countries (LICs):**
  - Focus on institution-building, anti-corruption efforts, and strengthening basic law enforcement capacity.
  - Civil freedoms have minimal impact on safety outcomes unless institutional stability is first achieved.
- **Lower-Middle-Income Countries (LMICs):**
  - Prioritize governance reforms and reinforce judicial independence.
  - Enhance police training programs and implement robust victim protection frameworks.
- **Upper-Middle-Income Countries (UMICs):**
  - Invest in administrative efficiency, anti-corruption mechanisms, and judicial transparency.
  - Societal freedoms become increasingly relevant as social institutions and legal frameworks mature.
- **High-Income Countries (HICs):**
  - Sustain governance transparency, civil liberties, and digital safety protections.
  - Address modern safety challenges such as cybercrime, digital harassment, and threats to democratic institutions.

7.11 Contribution of the Hybrid Model

By integrating Exploratory Factor Analysis (EFA) with Random Forest regression, the hybrid modeling framework offers both interpretability and predictive robustness:

- **Latent Constructs Identified:** Factor analysis statistically extracts underlying dimensions (GLS1–GLS4) driving national safety outcomes.
- **High Predictive Accuracy:** Random Forest models achieve approximately 90% accuracy, validating the relevance of the extracted factors in explaining real-world safety scores.

- **Income-Specific Insights:** Factor importance varies across income groups, highlighting diverse developmental pathways and safety mechanisms tailored to economic context.
- **Policy Utility:** The hybrid model serves as a reproducible, data-driven tool for policymakers to prioritize structural reforms, governance improvements, and institutional strengthening.

In summary, the hybrid model offers a global benchmarking framework that combines statistical rigor with machine learning flexibility, enhancing both explanatory power and practical applicability in international development strategies.

8 Discussion

8.1 Implications for Policy and Global Governance

The findings of this study carry significant implications for policymakers, international organizations, and researchers:

Tailored Policy Frameworks

- **Low-Income Countries (LICs):** Prioritize building state capacity, combatting corruption, and strengthening basic law enforcement before expanding civil rights frameworks.
- **Lower-Middle-Income Countries (LMICs):** Focus on judicial reforms, administrative efficiency, and breaking the cycle of weak enforcement mechanisms.
- **High-Income Countries (HICs):** Maintain investments in freedoms, inclusivity, and digital safety protections to counter emerging threats such as cybercrime.

Integrated Reforms

Safety cannot be ensured by addressing governance, enforcement, or freedoms in isolation. Sustainable peace and public safety require coordinated and simultaneous improvements across all institutional domains.

Data-Driven Monitoring

The hybrid model provides an objective, evidence-based approach to measuring safety, avoiding subjective weighting biases found in traditional composite indices.

International Development Aid

Donor agencies and development partners should align funding priorities with the most influential safety drivers for each income group:

- Governance training and transparency initiatives for LMICs.
- Institutional capacity building in LICs.

8.2 Theoretical Contributions

This research makes several contributions to the academic and policy literature:

- Demonstrates the **multidimensional nature of safety** using Exploratory Factor Analysis (EFA), moving beyond simplistic crime-based indices.
- Combines EFA with Random Forest regression, bridging **statistical inference** and **predictive modeling** to enhance policy relevance.
- Highlights how **income-level heterogeneity** influences the success of governance and legal reforms, reinforcing the need for context-specific models of institutional development.

8.3 Future Research Directions

Future extensions of this study can pursue the following avenues:

- Incorporate **spatial analysis** to account for regional spillover effects of safety and insecurity.
- Test **causal models** linking governance improvements and legal reforms to long-term safety outcomes.
- Explore the use of advanced machine learning algorithms such as **XGBoost** and **Neural Networks** to improve predictive performance without sacrificing interpretability.
- Develop an **interactive global safety dashboard** based on this hybrid model to aid real-time decision-making by policymakers and development agencies.

9 Conclusion

This research was undertaken to identify, analyze, and predict the determinants of national safety levels across 190 countries using a hybrid analytical approach. By integrating Exploratory Factor Analysis (EFA) with Random Forest regression modeling, the study aimed to:

- Uncover latent governance and institutional factors that drive safety outcomes globally.
- Quantify the predictive power of these factors, assessing their relative importance across different income groups.
- Provide evidence-based, tailored recommendations for policymakers and international organizations to improve societal safety.

The study successfully met these objectives, delivering a data-driven and interpretable framework for understanding and enhancing safety at a global scale.

9.1 Scope of the Study

In summary, the scope of this study is to analyze, predict, and interpret safety determinants across the globe, focusing on:

- Four latent factors (GLS1–GLS4) influencing safety.
- 190 countries, categorized by World Bank income groups.
- A hybrid methodology combining statistical inference and machine learning for improved interpretability and accuracy.
- Practical, policy-relevant insights tailored to economic development levels.

This scope ensures that findings are relevant, actionable, and grounded in robust data science techniques, while acknowledging methodological limitations and setting the foundation for future research.

9.2 Practical Scope

This study contributes to multiple domains:

- **Policy Formulation:** Offers national governments and international agencies a prioritization roadmap for institutional reforms and safety improvements.
- **Global Benchmarking:** Provides an objective, data-driven safety assessment framework to complement traditional indices such as the Global Peace Index.
- **Academic Research:** Demonstrates the utility of integrating EFA with machine learning models in socio-political and developmental studies.

However, it is important to note that the study does **not** aim to create a new safety index. It also does not account for unobservable factors such as culture, geopolitical tensions, or sudden crises—elements that remain outside the bounds of the current dataset.

9.3 Boundaries and Limitations of Scope

While the study provides robust insights, certain limitations must be acknowledged:

- **Data Constraints:** Incomplete or inconsistent data from low-income countries may reduce model accuracy and generalizability.
- **Cross-Sectional Nature:** The analysis is limited to a single year (2024), restricting longitudinal inference.
- **Non-Causal Modeling:** Random Forest models provide insights into variable importance but do not establish causal relationships between governance factors and safety outcomes.

References

[1] Alekseev, A., & Krasheninnikov, S. (2020). *Machine learning approaches to political and social stability forecasting*. Journal of Political Science and International Relations, 7(3), 122–135.

[2] Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

[3] Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory Factor Analysis*. New York: Oxford University Press.

[4] Field, A. P. (2017). *Discovering Statistics Using SPSS* (5th ed.). London: SAGE Publications.

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in Python*. Springer.

[6] Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (4th ed.). Guilford Press.

[7] OECD. (2024). *Governance Indicators and Rule of Law: Global Comparisons*. Organisation for Economic Co-operation and Development.

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[9] Institute for Economics & Peace. (2024). *Global Peace Index 2024: Measuring Peace in a Complex World*. Retrieved from `https://www.visionofhumanity.org`

Appendix

This appendix provides additional resources to complement the main body of this research project. All datasets, Python scripts, R code for factor analysis, visualizations, and machine learning model implementations used in this study are made publicly available for transparency, reproducibility, and further academic research.

Project Repository Contents

The GitHub project repository contains the following components:

- **Dataset:** 2024 Country-based Safety Data (raw and preprocessed `.csv` formats).
- **Exploratory Factor Analysis (EFA):** R scripts for KMO test, Bartlett's test of sphericity, parallel analysis, factor extraction, varimax rotation, and factor score generation.
- **Machine Learning Code:** Python scripts for Random Forest model training, feature importance analysis, income group segmentation, and actual vs. predicted safety score visualizations.
- **Visualization Outputs:** Heatmaps, regression plots, integration flowcharts, and policy relevance charts generated using R and Python.
- **Reproducibility Guidelines:** Step-by-step instructions for replicating the analysis and adapting it to other datasets.

All files and analysis steps are accessible at the following GitHub repository:

`https://github.com/Subhasis18-art`

This open-access repository ensures that the research methodology and findings can be reviewed, validated, and extended by other researchers, practitioners, and policymakers, fostering collaboration and innovation in the field of global safety analytics.