A Data-Driven Analysis and Forecasting of Gender Inequality Trends

Contents

1. Introduction

Gender inequality worldwide consistently affects sectors such as education, employment opportunities, political representation and healthcare access. The UNDP developed the Gender Inequality Index (GII) to track differences in women's reproductive health status, empowerment levels, and participation in the labour market. It's crucial to understand the factors responsible for gender inequality evolution to create effective policy interventions and build more equitable societies.

The report analyzes gender inequality patterns from 1990 to 2022 using statistical modeling techniques. We employ multiple regression analysis to quantify the association between GII and key socio-economic variables, examining their relative contributions to gender inequality.

This analysis examines the following explanatory variables:

- Adolescent Birth Rate (births per 1,000 women ages 15-19)
- Labour Force Participation Rate, Female (% of the population ages 15 and older)
- Labour Force Participation Rate, Male (% of the population ages 15 and older)
- Maternal Mortality Ratio (deaths per 100,000 live births)
- Share of Seats in Parliament, Female (% held by women)
- Share of Seats in Parliament, Male (% held by men)
- Population with at Least Some Secondary Education, Female (% ages 25 and older)
- Population with at Least Some Secondary Education, Male (% ages 25 and older)

All models were designed using data from 1990 to 2017, and validated against observed values from 2018 to 2022. After identifying key drivers of gender inequality through regression analysis, we employ ARIMA time-series modeling to project GII trends from 2023 to 2027. This approach allows us to assess the trajectory of gender inequality and anticipate future developments based on past trends.

Accordingly, the objectives of this study are:

1. Identify key socio-economic factors influencing GII through multiple linear regression.
2. Quantify the relative impact of these variables on gender inequality.
3. Use ARIMA modelling to predict future GII trends after 2022.
4. Evaluate the accuracy of regression and time-series models in analyzing gender inequality.
5. Serve as a resource for policymakers and governments seeking to develop effective strategies to reduce gender inequality.

2. Data Analysis

2.1 Data Source

This study relied on United Nations Development Programme (UNDP) data. This dataset contains information about 195 countries spanning from 1990 to 2022 and features the Gender Inequality Index (GII) along with yearly country rankings and previously discussed socio-economic variables.

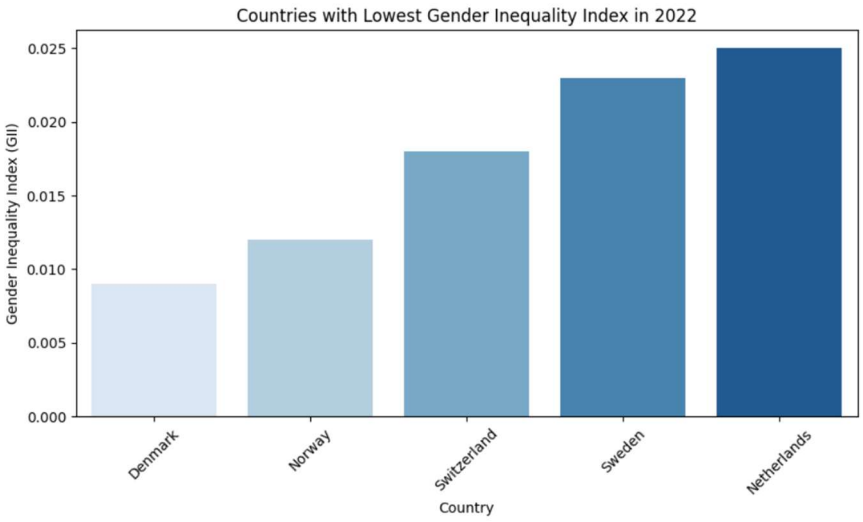
2.2 Data Cleaning & Preparation

Specific preprocessing steps were implemented to maintain the consistency and reliability of the data. The analysis proceeded without imputation or exclusion because the key indicators contained no missing values. The dataset's dimension and note columns were excluded since they contained no values. We reorganized the data using a pivot table transformation, establishing country and year as indexes and indicators as columns. This method made the organisation of values clear for every independent variable.

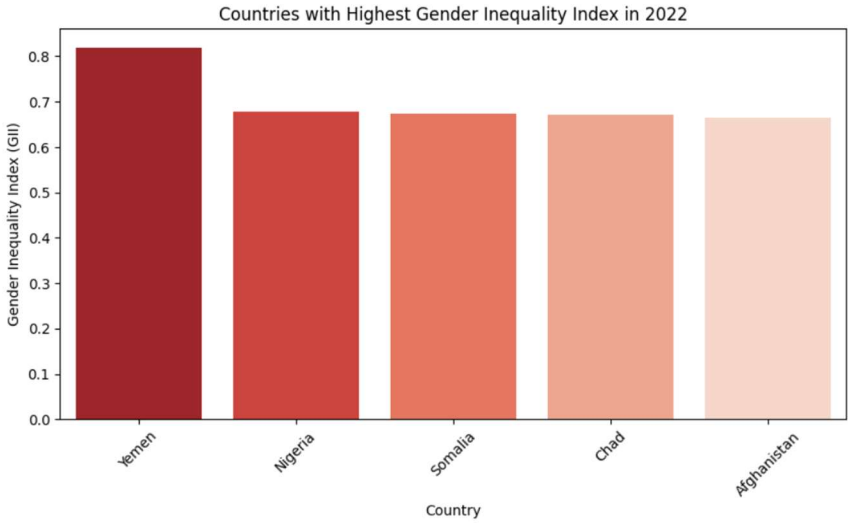
2.3 Exploratory Data Analysis (EDA)

Our exploratory data analysis (EDA) of global gender inequality patterns utilized the latest Gender Inequality Index (GII) data from 2022 alongside historical data from 1990 to 2022. This section investigates countries with the highest and lowest levels of gender inequality alongside those showing the quickest and slowest advancement over time.

2.3.1 Top and Bottom 5 Countries by GII in 2022



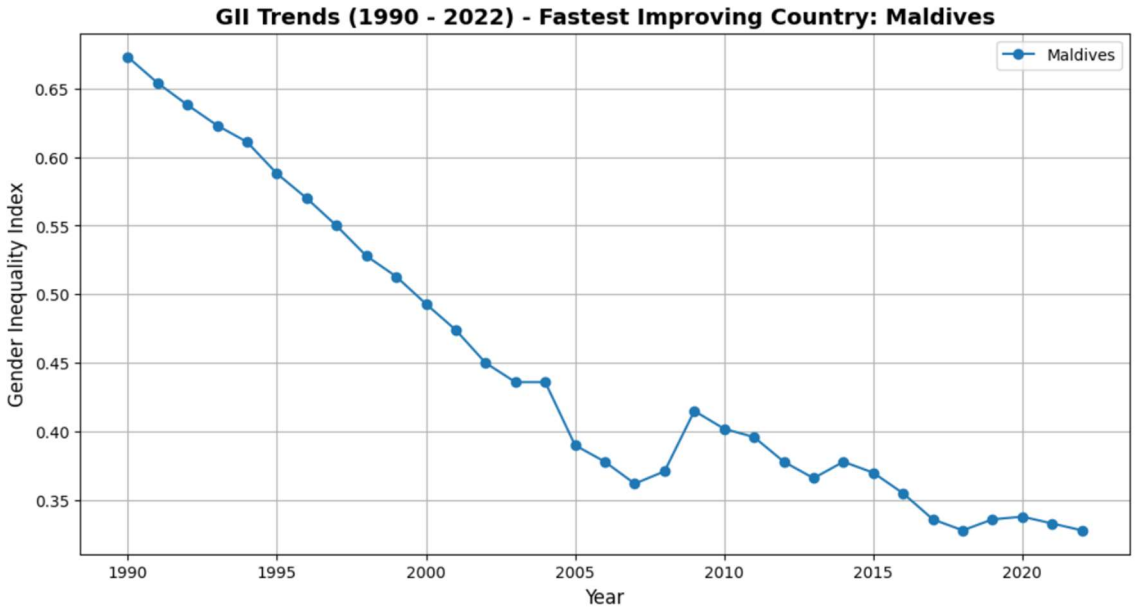
The nations of Denmark, Norway, Switzerland, Sweden, and the Netherlands earned top rankings as the five most gender-equal countries in 2022. The low Gender Inequality Index scores demonstrate minor differences between men and women across education, workforce inclusion, maternal health, and political representation.



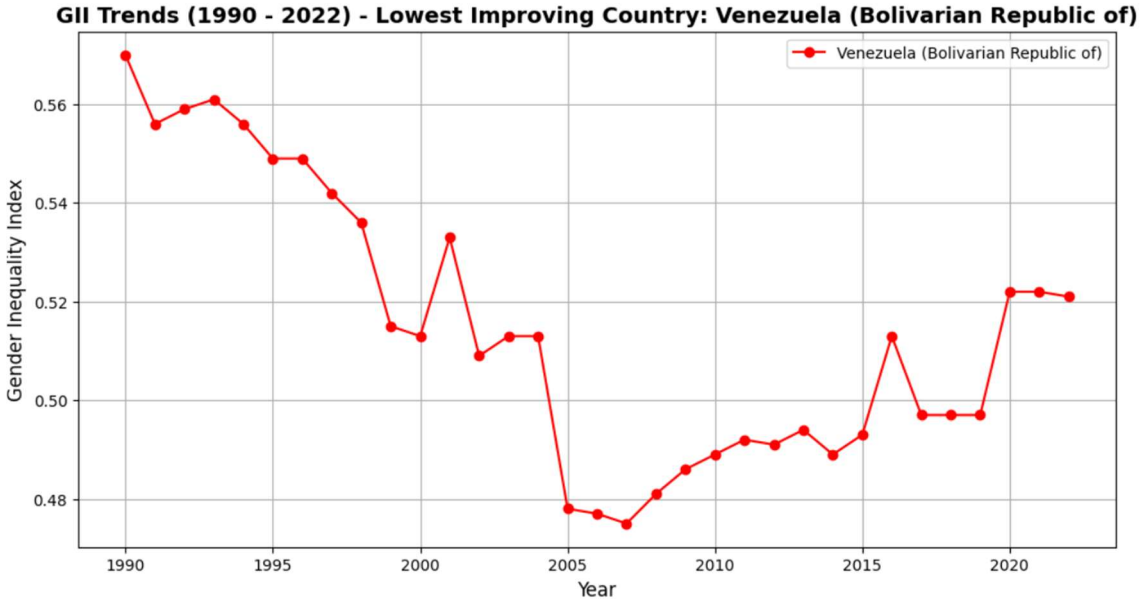
Yemen, Nigeria, Somalia, Chad, and Afghanistan occupy the bottom positions regarding gender equality, as indicated by their highest Gender Inequality Index values. The nations exhibit significant gender differences, which include poor educational opportunities for women alongside limited economic roles for females, lower female political involvement and higher maternal mortality rates.

2.3.2 GII Trends Over Time: Fastest vs. Slowest Improving Countries

To examine historical trends in gender inequality, GII data from 1990 to 2022 was analyzed to determine the country that has shown the most significant reduction in GII and the one with the least improvement.



The Maldives stands out for its significant GII reduction, demonstrating a major improvement in gender equality throughout the last thirty years. Between 1990 and 2022, the country's GII score declined from over 0.65 to below 0.30 due to progress in education, workforce involvement, and maternal health improvements. There were slight deviations throughout 2005 and 2010, yet the general path reflects a steady decline.



In contrast, Venezuela has experienced minimal advancement in GII while facing fluctuating results and periods of stagnation. The trend decreased slightly during the early 2000s but remained stable since 2010 when recent years presented minimal declines. The data indicates a slowdown in gender equality advancement that appears to correlate with more considerable political, economic, and social challenges.

The exploratory analysis delivers insights about worldwide gender inequality patterns throughout history and establishes the groundwork for future research into the factors behind these differences.

3. Methodology

3.1 Regression Analysis

3.1.1 Assessing linearity in relationships

An evaluation of the linear relationship between the Gender Inequality Index (GII) and all chosen independent variables was needed before the completion of regression modelling. The assumption of a linear relationship between predictors and the response variable is a core part of multiple linear regression, making checking this assumption essential for evaluating the model's suitability.

The relationship between each independent variable and the Gender Inequality Index (GII) was analyzed using scatter plots. Examining these plots allows one to visually assess trends and patterns and identify deviations from linear behaviour. The scatter plots reveal that different independent variables display multiple levels of linearity with one another.

Strong negative correlation: The proportion of people who have completed at least some secondary education forms a strong negative relationship with the Gender Inequality Index (GII). Nations that achieve higher educational levels demonstrate a linear decrease in gender inequality.

Strong positive correlation: The adolescent birth rate and the maternal mortality rate display a strong positive correlation with GII while showing that higher values for these indicators lead to greater gender inequality. The effect of maternal mortality decreases at higher levels, which indicates a possible non-linear relationship.

Weaker relationship: The relationship between GII and other variables demonstrates unclear patterns or non-linear characteristics. The correlation between female labour force participation and GII is moderately negative, although its distribution suggests a potential non-linear effect due to observed curvature. The participation rate for men in the labour force fails to show any clear pattern. The proportion of parliamentary seats held by female members shows an inverse relationship to GII values, while male-dominated representation increases gender inequality levels.

The scatter plot analysis (shown in the Appendix) reveals that certain variables fit linear modelling while others need non-linear modelling. The subsequent sections will evaluate multiple linear regression (MLR) and polynomial regression to assess their effectiveness in capturing these observed relationships.

3.1.2 Multiple Linear Regression (MLR) with Ordinary Least Squares (OLS)

The Multiple Linear Regression (MLR) model using Ordinary Least Squares (OLS) was applied to examine the relationship between the Gender Inequality Index (GII) and the independent variables. The model achieved an R^2 score of 0.855, indicating that the independent variables can explain 85.5% of the variation in GII. The Mean Absolute Error (MAE) of 0.0577 and Mean Squared Error (MSE) of 0.0055 suggest that the variation in GII can be closely explained by the independent variables.

The model's effectiveness becomes more apparent in the actual vs. predicted values scatter plot since the data points align closely with the expected trend.

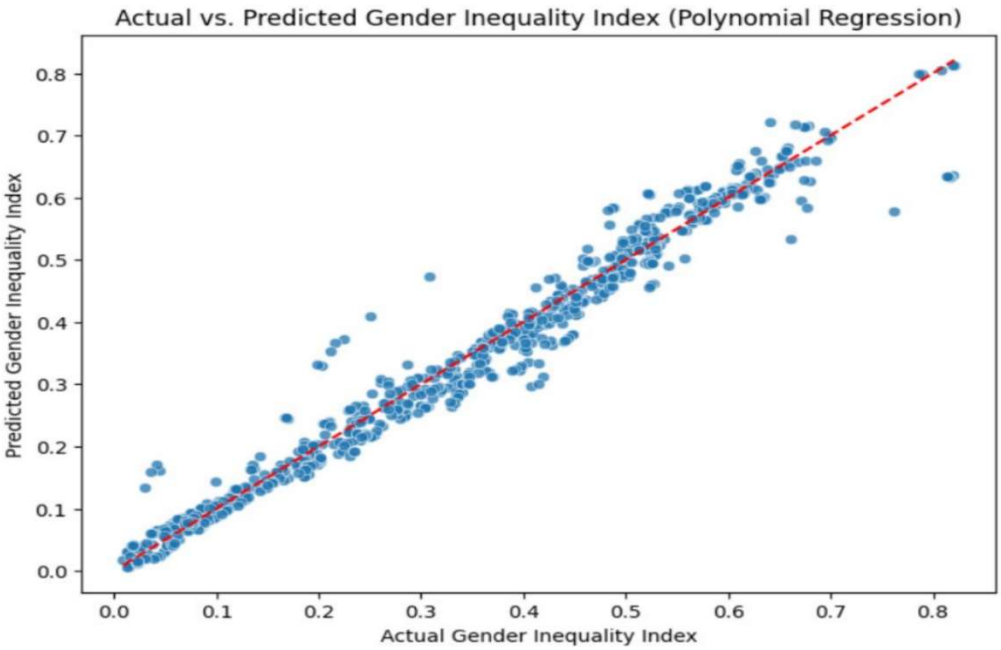
Analysis from regression results shows how each independent variable affects GII values. The study reveals that both the adolescent birth rate and maternal mortality ratio show strong positive coefficients, which validate their link to elevated levels of gender inequality. Negative coefficients for female labour force participation and female secondary education levels confirm their importance as factors diminishing GII. Data shows male labour force participation and parliamentary representation rates correlate positively with gender inequality, but male secondary education levels show little significance.

The regression output (shown in the Appendix) identifies potential multicollinearity issues despite the model's strong performance, which can affect coefficient interpretation.

3.1.3 Polynomial Regression with Ridge Regularization

A Polynomial Regression model with Ridge Regularization was applied to handle the possible non-linear relationship between the Gender Inequality Index (GII) and its predictors. To ensure model stability and flexibility, ridge regularization was employed to prevent overfitting by setting the regularization parameter (alpha) at 0.01.

The model reached an R^2 score of 0.9701, which shows that it accounts for about 97% of variations in GII. The predictive performance improved significantly, as demonstrated by the Mean Absolute Error (MAE) reaching 0.0216 and the Mean Squared Error (MSE) achieving 0.0011.



The scatter plot comparing actual vs. predicted GII values suggests a strong alignment, further validating the effectiveness of this approach.

Although polynomial regression improves the model's ability to account for non-linear effects, it also increases model complexity, making direct interpretation of coefficients less intuitive.

3.1.4 Comparison of Regression Model Performance

The polynomial regression model performed better by attaining higher R^2 scores and lower error values, demonstrating its better data fitting capability. The comparison between actual and predicted values in scatter plots shows how the polynomial regression model approaches the ideal trend line more closely than the other model.

The polynomial regression demonstrates enhanced predictive power but creates greater complexity, obscuring the effect interpretation of individual variables. The MLR model offers better clarity on how predictors influence gender inequality but has the potential to oversimplify some connections between variables.

3.2 Time Series Analysis with ARIMA

3.2.1 Justification for ARIMA Selection

We chose the AutoRegressive Integrated Moving Average (ARIMA) model to predict time series because it effectively captures time dependencies and long-term trends in organized historical data. The Gender Inequality Index (GII) develops through time because of policy changes and socio-economic and demographic shifts, making ARIMA an appropriate tool for modelling and forecasting future patterns.

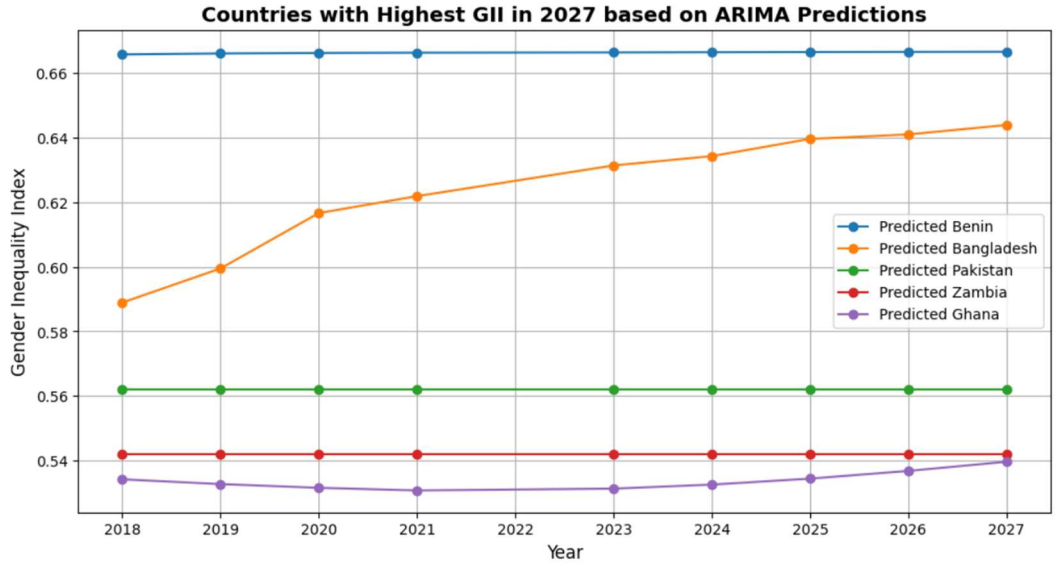
3.2.2 Model Training and Evaluation

The data was split into three segments: The data analysis was segmented into three distinct periods comprising a training set from 1990 to 2017 followed by a validation set from 2018 to 2022 and ending with a forecasting set from 2023 to 2027. Model fitting was conducted using the training period while the validation period helped measure forecasting accuracy and extended the model for future predictions.

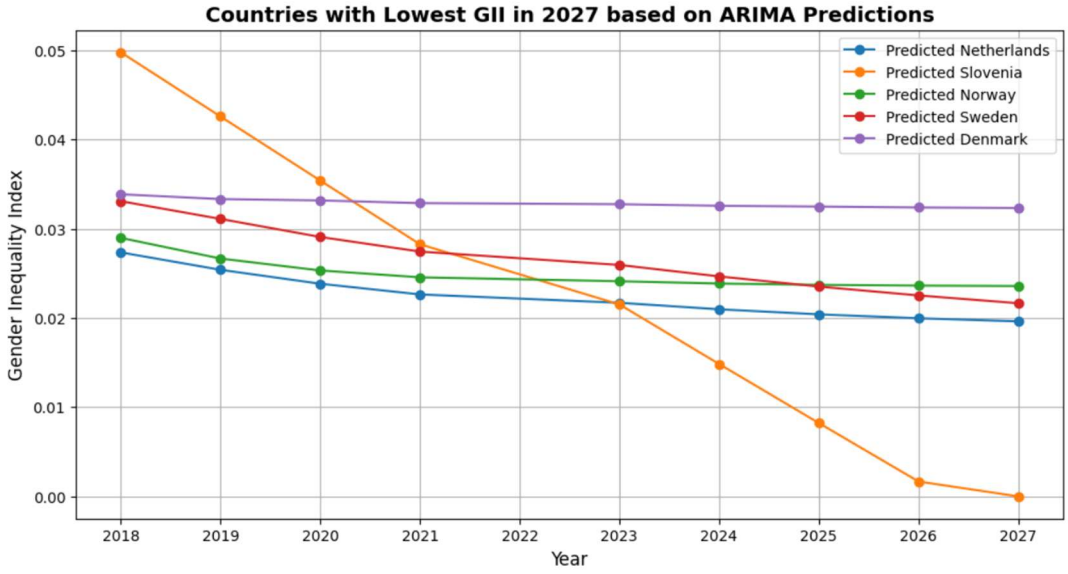
The optimal ARIMA setup was identified by testing various combinations of autoregressive (p), differencing (d), and moving average (q) parameters. A separate model was developed for every country's GII time series to account for unique trend variations.

3.2.3 Model Performance Evaluation

The study showed excellent predictive accuracy with MAE at 0.0178 and MSE at 0.0006 demonstrating very close agreement between forecasted and observed values. The R^2 score of 0.9794 demonstrates that the model accounts for nearly all variations within GII trends throughout time. The Mean Absolute Percentage Error (MAPE) stood at 14.02% which reflects moderate forecast uncertainty alongside an overall prediction accuracy of 85.98%.



The five countries projected to have the highest GII values in **2027** are Benin, Bangladesh, Pakistan, Zambia, and Ghana. Bangladesh shows a steady increase in gender inequality over time, while others, such as Benin, maintain consistently high GII values. The upward trajectory suggests that without significant policy interventions, gender disparities in these countries may continue to widen.



Conversely, the five countries expected to have the lowest GII values by 2027 are the Netherlands, Slovenia, Norway, Sweden, and Denmark. These nations demonstrate a declining trend in GII, signifying continuous progress toward gender equality. Notably, Slovenia exhibits the steepest decline, suggesting significant improvements in gender parity.

4. Challenges & Limitations

Our study of the Gender Inequality Index (GII) uncovered numerous methodological, computational and interpretative challenges that affected our results' practicality. The dataset provided extensive information yet lacked detailed data about cultural and policy effects on gender disparities, which restricted the model from generalizing historical trends.

Multiple linear regression (MLR) served as a strong baseline model but faced limitations due to multicollinearity, compromising the clarity of individual coefficient interpretations. We tackled this by building the polynomial regression, which addressed non-linear patterns. The model's enhanced performance introduced a risk of overfitting, which could impair its generalizability across varying economic scenarios.

Regarding interpretation issues, the models show statistical connections yet fall short of determining cause-and-effect relationships, implying that although the results provide helpful trend examination, they need supplementary qualitative information for effective policy formation.

All in all, our approach works effectively to showcase significant trends and delivers data-based insights which support strategic gender equity initiatives yet require contextual understanding to maintain balanced and practical decision-making.

5. Social Implications

Our analysis suggests that strategic interventions can address the following gender equality issues:

Relations between Reproduction and Health Outcomes - In areas with high GII scores, we observe significantly higher Adolescent Birth Rates and Maternal Mortality Ratios. Limited access to quality reproductive healthcare and comprehensive sex education contributes to these adverse health outcomes. Policymakers should implement targeted public health campaigns and allocate additional resources to enhance maternal and adolescent health services. These actions could reduce health disparities and help lower GII scores.

Educational & Economic Opportunities - Our findings demonstrate that increased Labour Force Participation Rates among females and a higher percentage of populations with at least some secondary education (for both females and males) correlate with lower GII scores. Expanding access to quality education and fostering equitable, child-friendly workplaces are essential to narrowing income gaps and boosting women's economic empowerment. These efforts drive social progress and improve overall community well-being.

Political Representation - A higher share of seats in parliament that women hold is linked to improved gender equality. When women actively participate in political decision-making, they shape policies that address gender-specific challenges. Enhancing female representation in legislative bodies promotes more inclusive and democratic governance, ultimately supporting sustained reductions in the GII.

Regional Variations - Our data analysis and the ARIMA forecasting reveals significant regional differences in progress toward gender equality. Western Europe and North America appear poised for rapid improvement, while Sub-Saharan Africa and South Asia progress more slowly. These disparities underscore the need for interventions tailored to each region's unique socio-economic and cultural context.

These insights underscore the need for coordinated policy efforts that integrate healthcare, education, economic empowerment, and political inclusion to drive meaningful and sustained reductions in gender inequality.

6. Conclusion

This study analyzed gender inequality trends across 195 countries from 1990 to 2022, identifying key socio-economic factors through multiple regression and forecasting future trends using ARIMA modeling. Our findings indicate that higher adolescent birth rates and maternal mortality ratios contribute to increased gender inequality, while greater female labor force participation and higher education levels are associated with lower GII scores. The ARIMA model, validated with an R^2 score of 0.9794 and 85.98% prediction accuracy, projects persistent inequality in countries like Benin and Bangladesh, while nations such as Slovenia and Norway are expected to continue improving gender equity. These results emphasize the need for targeted policy interventions to address regional disparities.

Overall, this paper underscores the importance of leveraging statistical models to understand and anticipate gender inequality trends. The insights derived from our analysis provide valuable evidence for policymakers, educators, and advocacy groups aiming to implement data-driven strategies for gender equity. Future research should integrate alternative modeling techniques, such as machine learning approaches, and incorporate additional socio-political variables to enhance the efficency of gender inequality predictions.

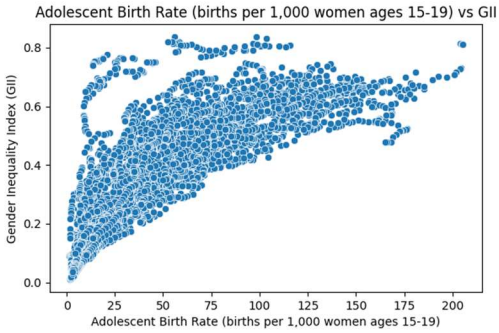
References

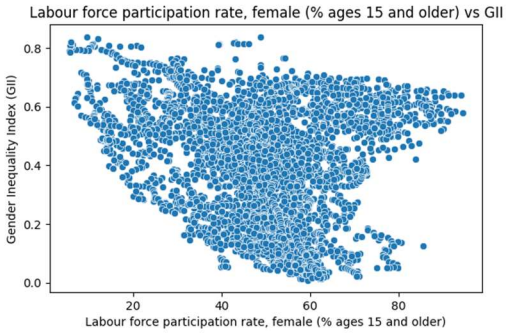
OpenAI. (2023). *ChatGPT* (Feb 20 Version) [Large language model]. https://chat.openai.com

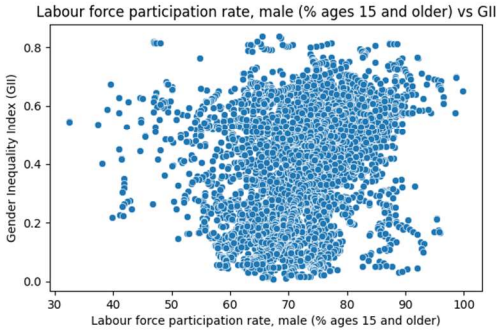
United Nations Development Programme. *Human Development Report data center: Documentation and downloads*. Retrieved [February 10, 2025], from https://hdr.undp.org/data-center/documentation-and-downloads

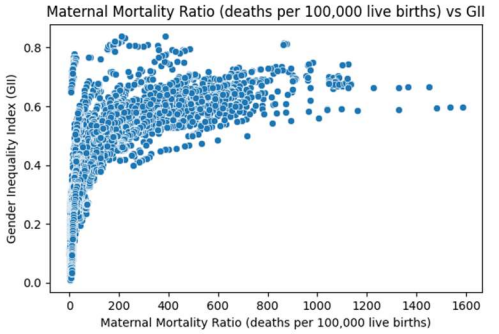
Appendix

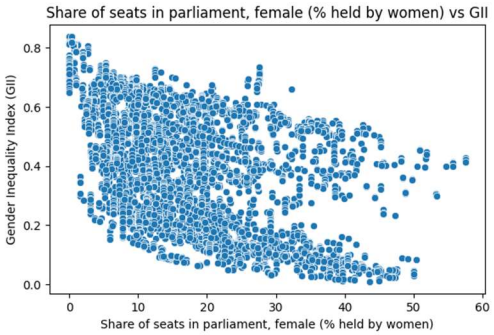
Appendix1

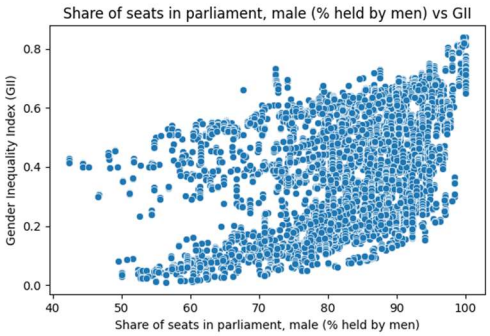


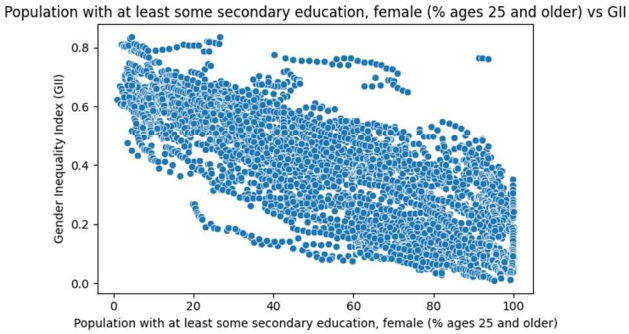


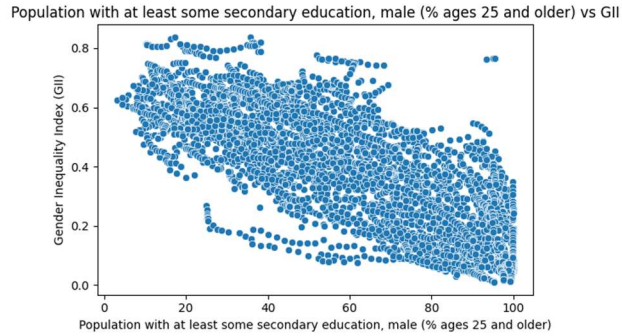












Appendix2

	coef
const	2.007e-05
Adolescent Birth Rate (births per 1,000 women ages 15-19)	0.0019
Labour force participation rate, female (% ages 15 and older)	-0.0027
Labour force participation rate, male (% ages 15 and older)	0.0029
Maternal Mortality Ratio (deaths per 100,000 live births)	0.0002
Share of seats in parliament, female (% held by women)	-0.0013
Share of seats in parliament, male (% held by men)	0.0033
Population with at least some secondary education, female (% ages 25 and older)	-0.0011
Population with at least some secondary education, male (% ages 25 and older)	-0.0003