EDA - Project Report

A Analysis based on the wiid dataset

Submitted By

**Team J**

T Rohith 23MIA1104

Sastika M

Abstract

Income inequality is a complex global issue that continues to affect the quality of life, access to opportunities, and economic growth in both developing and developed nations. In this project, we conduct a detailed exploratory data analysis (EDA) of global income distribution using the World Income Inequality Database (WIID), which provides standardized and cross-country comparable data over several decades.

The primary objective of this project is to uncover meaningful patterns, trends, and anomalies in income inequality metrics—such as the Gini coefficient, Palma ratio, and decile/quintile income shares—across different countries, time periods, and income groups. We also explore the relationship between inequality and key economic indicators such as GDP per capita and population.

To achieve this, we implemented a structured analysis pipeline including data preprocessing (handling missing values, standardizing formats, and treating outliers), statistical analysis (mean, median, correlation, regression), and a range of visualizations (boxplots, scatter plots, heatmaps, and time series trends). This multifaceted approach allowed us to derive insights into how inequality evolves over time and varies across global regions.

Our study aligns with three key United Nations Sustainable Development Goals (SDGs):

• SDG 5 – Gender Equality

• SDG 8 – Decent Work and Economic Growth

• SDG 10 – Reduced Inequalities

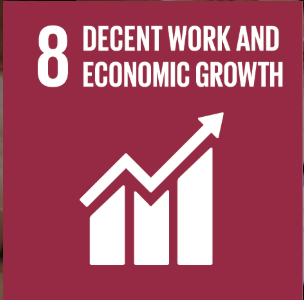


Table Of Contents

|  |  |
| --- | --- |
| S. No. | Section Title |
| 1 | Cover Page |
| 2 | Certificate |
| 3 | Acknowledgement |
| 4 | Abstract |
| 5 | Table of Contents |
| 6 | Introduction |
| 7 | Literature Review |
| 8 | Project Objectives |
| 9 | Sustainable Development Goals (SDG 5, 8, 10) |
| 10 | Dataset Description & Source |
| 11 | Data Preprocessing & Cleaning |
| 12 | Exploratory Data Analysis – Univariate |
| 13 | Exploratory Data Analysis – Bivariate |
| 14 | Exploratory Data Analysis – Multivariate |
| 15 | Correlation & Regression Analysis |
| 16 | Outlier Detection & Handling |
| 17 | Key Results and Visualizations |
| 18 | Trend Insights and Country Comparisons |
| 19 | Discussion and Interpretation |
| 20 | Predictions and Future Scope |
| 21 | Conclusion |
| 22 | Limitations and Recommendations |
| 23 | References |
| 24 | Appendix (Code/Visuals) |

Introduction

`Income inequality has been one of the most important issues of 21st century. It is referred as the unequal distribution of income and wealth among individuals or groups in a society or across nations. This disparity results not only in economic gaps but also in fields such as education, healthcare, and employment opportunities. Continuing income inequality is known to adversely affect social unity, reduce intergenerational mobility, and slow down long-term economic development. In extreme cases, it contributes to increased crime rates, social unrest, and political polarization.

The study of income inequality requires comprehensive, reliable, and comparable data. To support this effort, the United Nations University World Institute for Development Economics Research (UNU-WIDER) has developed the World Income Inequality Database (WIID). The WIID provides a consolidated global database on income inequality with indicators such as the Gini index, Palma ratio, income shares across population percentiles (quintiles and deciles), and other economic measures including GDP per capita and population.

This project supports the WIID dataset to perform a deep and structured exploratory data analysis (EDA) with the objective of identifying patterns and trends in income inequality over time and across regions. It applies both statistical techniques and visual analytics to find correlations between income distribution metrics and macroeconomic indicators like GDP. Special focus is in observing regional disparities, evaluating trends in income concentration, and detecting significant outliers in economic behavior.

Additionally, the project aligns with the United Nations Sustainable Development Goals (SDGs), particularly:

• SDG 5: Gender Equality — addressing wage gaps and economic opportunity disparities among genders.

• SDG 8: Decent Work and Economic Growth — ensuring fair income and productive employment.

• SDG 10: Reduced Inequalities — reducing inequality within and among countries.

By interpreting global income inequality through the lens of data analytics, this project aims to offer actionable insights that can help policymakers, researchers, and development organizations design more inclusive and equitable economic policies. The findings also contribute to academic research in the fields of development studies, economics, and public policy by illustrating how real-world data can reveal hidden dynamics behind global inequality.

Literature Review

Income inequality has long been recognized as one of the most pressing socio-economic issues worldwide. It reflects the unequal distribution of income and wealth among individuals, communities, or regions, and contributes significantly to disparities in access to essential services such as education, healthcare, and employment. Among the different forms of inequality, the gender wage gap has been a particularly persistent challenge. This refers to the average difference in income earned by men and women, typically expressed as a percentage of male earnings. Despite global advancements in education and employment access, the wage gap remains entrenched across countries and sectors.

Numerous academic studies and policy papers have investigated the origins and impacts of income inequality. The World Bank’s World Development Report (2012) emphasized that gender equality is not only a fundamental human right but also essential for poverty reduction and economic development. Likewise, the International Labour Organization (ILO) and UN Women have published regular findings showing that women continue to earn less than men for comparable work, often due to occupational segregation, unpaid care burdens, and limited access to leadership positions. For example, women globally spend an average of 3.2 times more hours on unpaid care work compared to men, which restricts their ability to engage in full-time or higher-paying employment opportunities.

Recent statistics from the World Economic Forum’s Global Gender Gap Report (2023) illustrate that the global gender pay gap stands at around 16%, and at the current rate of progress, it will take an estimated 131 years to fully close this gap. The report also shows that some regions, such as Sub-Saharan Africa and South Asia, face slower progress due to structural inequalities, traditional gender roles, and lack of formal labor protections. These trends are not only morally concerning but also economically inefficient. Studies by the Organisation for Economic Co-operation and Development (OECD) and the International Monetary Fund (IMF) have found strong correlations between gender equality and GDP growth. Greater economic participation by women can lead to more stable, resilient, and inclusive economies.

In response to these disparities, the United Nations established the Sustainable Development Goals (SDGs) in 2015, a set of 17 interconnected global goals aimed at ending poverty, protecting the planet, and ensuring prosperity for all. SDG 5 (Gender Equality) calls for the elimination of all forms of discrimination and violence against women, ensuring equal rights to economic resources and equal pay for work of equal value. SDG 10 (Reduced Inequalities) further emphasizes the importance of reducing inequality within and among countries, addressing income disparities and promoting the social and economic inclusion of all, regardless of age, gender, ethnicity, or economic status.

To analyze these issues quantitatively, comprehensive and reliable data is essential. The World Income Inequality Database (WIID), developed and maintained by the United Nations University World Institute for Development Economics Research (UNU-WIDER), serves as one of the most extensive resources for income distribution data. The WIID consolidates information from hundreds of household surveys and studies, providing standardized indicators such as the Gini coefficient, Palma ratio, decile and quintile income shares, and Atkinson indexes. It enables cross-country comparisons over time, making it highly valuable for researchers, economists, and policymakers seeking to understand inequality trends at both national and global levels.

Previous analyses using the WIID dataset have examined long-term changes in income distribution, the effects of taxation and social welfare systems, and the economic impacts of globalization. For instance, Milanovic (2016) used WIID data to propose the “elephant curve,” illustrating how global income growth disproportionately favored the middle classes in emerging economies and the top 1% in developed nations, leaving behind large segments of the global poor and working class.

Objectives

This project is centered on analyzing global income inequality trends using the World Income Inequality Database (WIID). The aim is to gain valuable insights into how income is distributed across different regions and economic categories, and how it relates to broader development indicators.

The core objectives of the project are as follows:

1. To analyze income inequality patterns across different countries using statistical indicators like the Gini coefficient, Palma ratio, and income share percentiles.
2. To study the relationship between income inequality and key economic indicators such as GDP per capita and population growth.
3. To identify countries that have shown the most significant improvement or decline in income inequality over time.
4. To detect and handle outliers in economic indicators that may skew analysis, using techniques like IQR and Winsorization.
5. To apply univariate, bivariate, and multivariate data analysis methods to discover deeper relationships between variables.
6. To align the study with United Nations Sustainable Development Goals:
   * • SDG 5: Achieve gender equality and empower all women and girls.
   * • SDG 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment.
   * • SDG 10: Reduce inequality within and among countries.
7. To generate actionable insights and propose data-backed strategies that can help guide equitable policy design.

Sustainable Development Goals (SDGs) Alignment

The United Nations Sustainable Development Goals (SDGs) provide a global framework to address critical challenges such as poverty, inequality, climate change, and social injustice. This project aligns specifically with three of these goals:

SDG 5 – Gender Equality

This goal emphasizes eliminating all forms of discrimination and violence against women and girls. Gender-based wage disparities and unequal access to economic resources contribute significantly to income inequality. By analyzing income distribution trends, this project aims to highlight persistent gender gaps and advocate for inclusive labor policies.

SDG 8 – Decent Work and Economic Growth

This goal promotes sustained economic growth and full, productive employment for all. Through data analysis of GDP per capita alongside income inequality metrics, the study evaluates how economic growth influences equitable income distribution. It also identifies the need for policies that ensure economic gains are shared fairly among all segments of society.

SDG 10 – Reduced Inequalities

At the heart of this project, SDG 10 focuses on reducing income disparities within and among countries. By comparing inequality indicators across regions and timeframes, the analysis seeks to uncover structural factors and outliers contributing to income gaps. The project offers data-driven recommendations to support more inclusive and balanced economic systems.

These goals serve as the guiding framework for the study’s analytical objectives, interpretation of trends, and policy suggestions.

Dataset Description – World Income Inequality Database (WIID)

This project utilizes the World Income Inequality Database (WIID), developed and maintained by the United Nations University World Institute for Development Economics Research (UNU-WIDER). It is a globally recognized dataset that consolidates and harmonizes income inequality statistics from hundreds of countries and survey sources.

**Key Features of the Dataset:**

• Coverage: Over 200 countries and territories

• Timeframe: Data spanning several decades (1950s to present)

• Observations: 24,000+ entries

• Indicators:

• Gini Coefficient (measures inequality from 0 to 100)

• Palma Ratio (income share of the richest 10% to the poorest 40%)

• Atkinson Index

• Decile and Quintile Income Shares (q1–q5, d1–d10)

• GDP per capita, Population, Exchange Rates

• Region Classifications (UN, World Bank)

• Income Group Classifications (Low, Middle, High Income)

**Dataset Source:**

World Income Inequality Database (WIID)

UNU-WIDER – https://www.wider.unu.edu/database/world-income-inequality-database-wiid

**Use in This Project:**

• Data was imported, cleaned (missing values, data types), and normalized

• Outliers were detected using IQR and winsorized where necessary

• Key variables were analyzed using univariate, bivariate, and multivariate EDA

• Correlations and regressions were run between Gini and GDP/population metrics

The WIID is particularly valuable due to its methodological consistency, broad coverage, and flexibility in supporting cross-country and longitudinal income inequality analysis

Data Preprocessing & Cleaning

Before performing any analysis, the WIID dataset underwent a comprehensive data preprocessing phase to ensure quality, consistency, and reliability of results. The original dataset consisted of over 24,000 records with 64 columns, encompassing both numerical and categorical data types.

**Key Preprocessing Steps:**

**1.Column Selection & Understanding**

• Reviewed column metadata and retained attributes relevant to income inequality, economic growth, and population trends.

• Examples: gini, palma, q1–q5, d1–d10, mean, median, GDP, population, exchangerate, region\_un, incomegroup.

**2.Handling Missing Values**

• Numeric columns were filled using median imputation (e.g., for Gini, GDP).

• Categorical columns were imputed using mode (e.g., region\_un, currency).

• Columns with excessive nulls (e.g., link, survey) were excluded from analysis.

**3.Data Type Conversion**

• Ensured numerical columns were converted from string format (e.g., “1,234”) to float.

• Cleaned whitespaces, tabs, and formatting inconsistencies.

**4.Outlier Detection & Treatment**

• Used Interquartile Range (IQR) and Winsorization to cap extreme values.

• Targeted columns: gini, GDP, population, q1–q5, d1–d10, mean\_usd, etc.

**5. Final Dataset Shape**

• Records retained: ~24,000

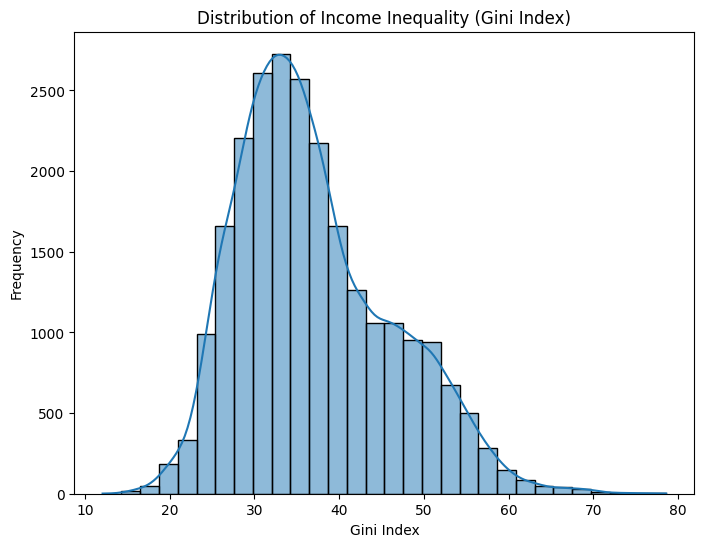
• Final columns used: 35+ cleaned, formatted, and verified

These preprocessing steps ensured that the analysis performed in subsequent phases was statistically valid and interpretable, minimizing the risk of skewed insights due to dirty or inconsistent data.

Univariate Analysis

Univariate analysis was conducted to examine the individual characteristics and distribution of key variables within the WIID dataset. This helped in identifying central tendencies, variability, and anomalies across income and economic indicators.

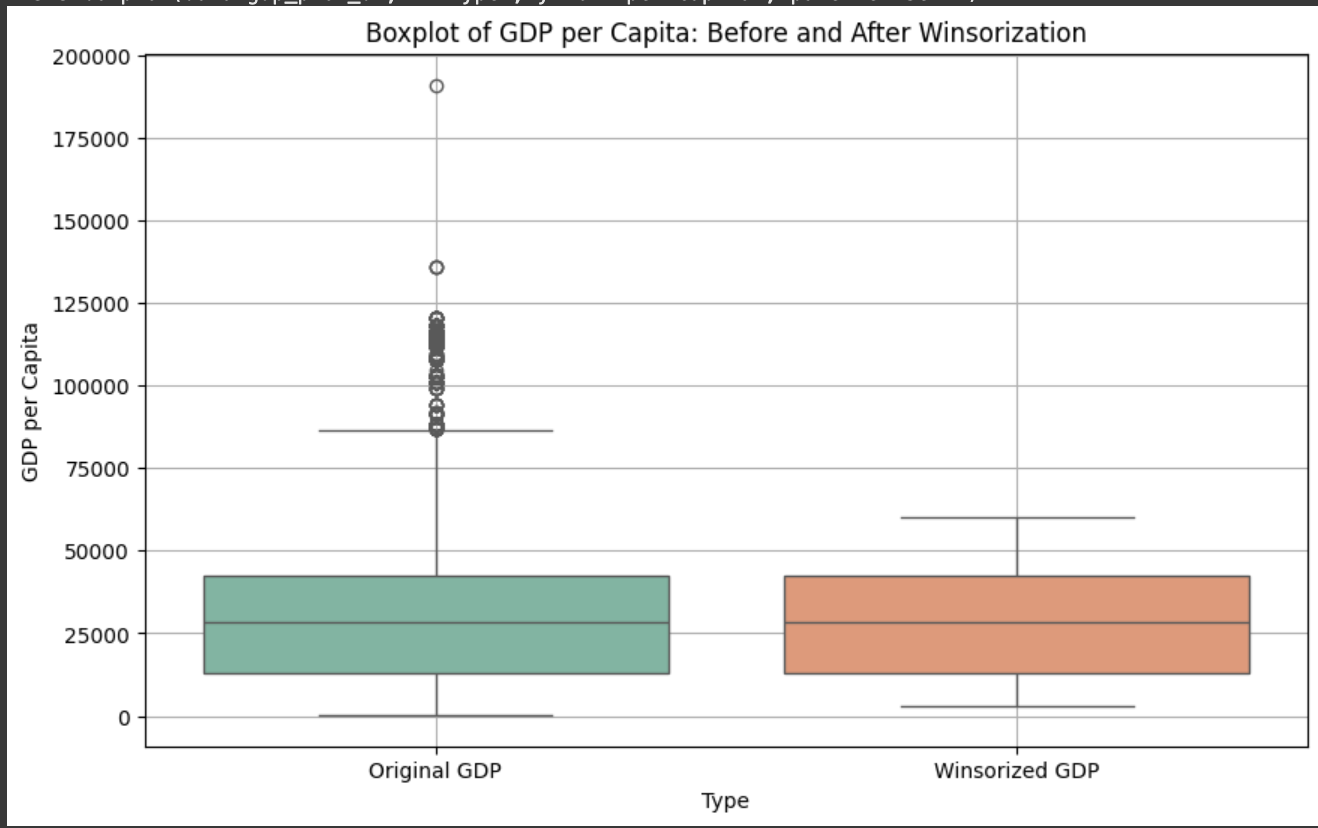
Gini Coefficient Distribution

The Gini index, a widely used measure of income inequality, ranged from 12.1 to 78.6 in the dataset, with a mean of 37.07 and a median of 35.40. This indicates that while many countries have moderate inequality, a significant number exhibit severe income disparities.

Interpretation:

The skewed distribution toward higher values suggests persistent inequality in several regions, justifying the need for targeted socio-economic policies.

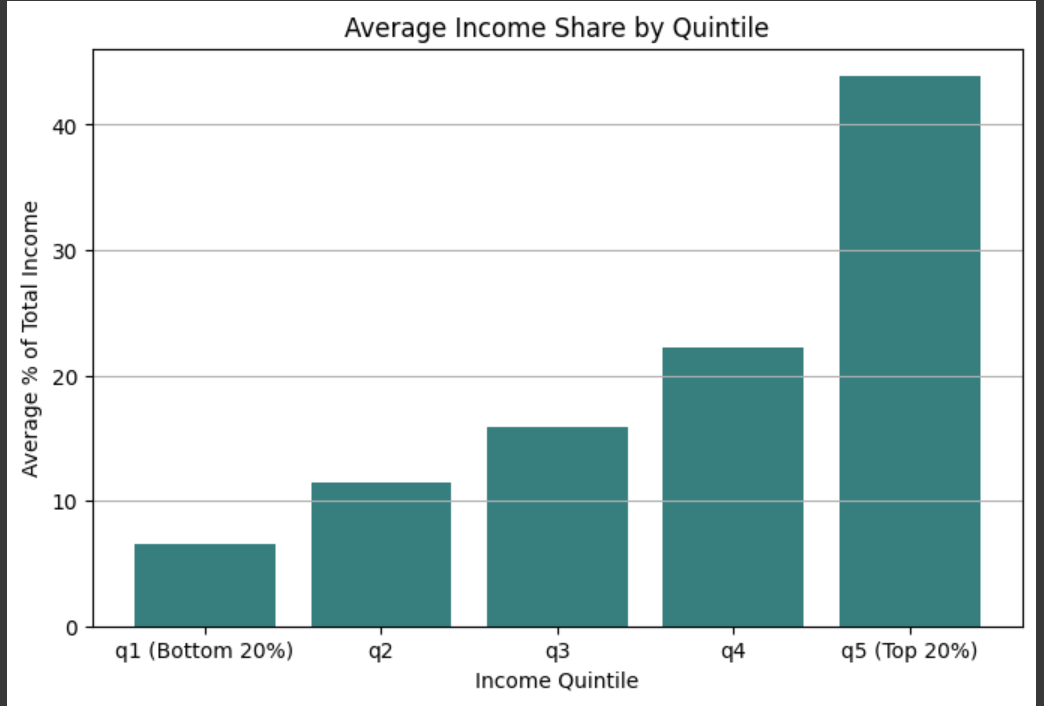
GDP per Capita

The GDP per capita column showed a highly skewed distribution with a few countries reporting extremely high values. This necessitated the use of log transformation and outlier treatment.

Interpretation:

Outliers were detected and handled to ensure meaningful visualizations. The disparity in GDP points to unequal development trajectories across nations.

Quintile Income Shares (q1–q5)

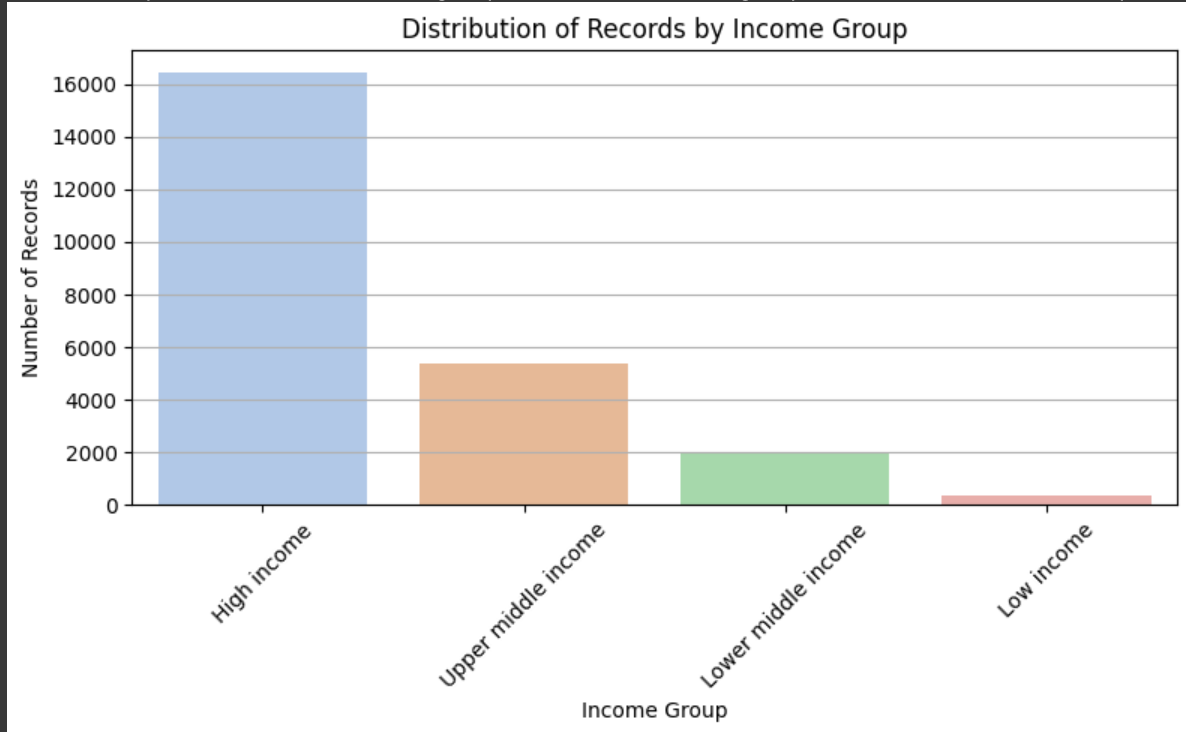
Analyzing income shares by quintiles revealed that the bottom 20% (q1) generally receive less than 5% of total income, whereas the top 20% (q5) command over 40%.

Interpretation:

This stark contrast demonstrates that wealth concentration remains a major issue, especially in developing economies.

Categorical Variables

Categorical data such as region\_un, income group, and currency were summarized using value counts.



Interpretation:

This provides context for comparative analysis in later stages, such as comparing inequality across low-income vs high-income nations.

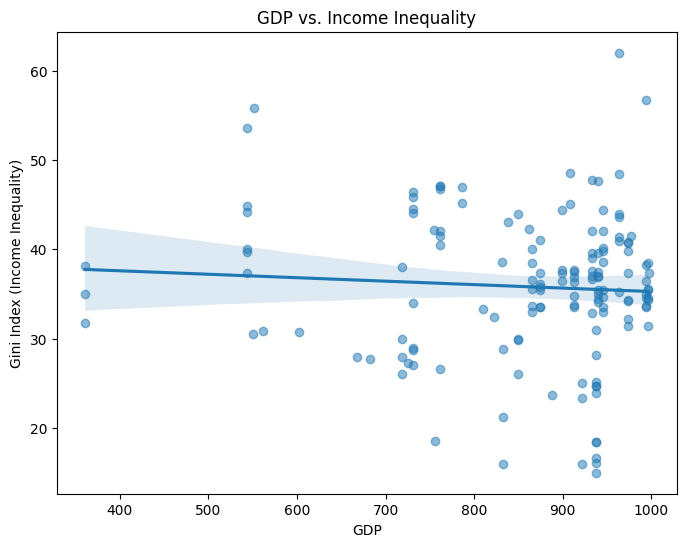
Bivariate Analysis

Bivariate analysis allows us to explore relationships between two variables, particularly between inequality indicators and macroeconomic factors. This step reveals associations, correlations, and possible causations in the data.

**Gini vs GDP per Capita**

We explored the relationship between the Gini index (income inequality) and GDP per capita. A scatter plot with a regression line showed a weak negative trend, suggesting that countries with higher GDP do not necessarily experience lower income inequality.

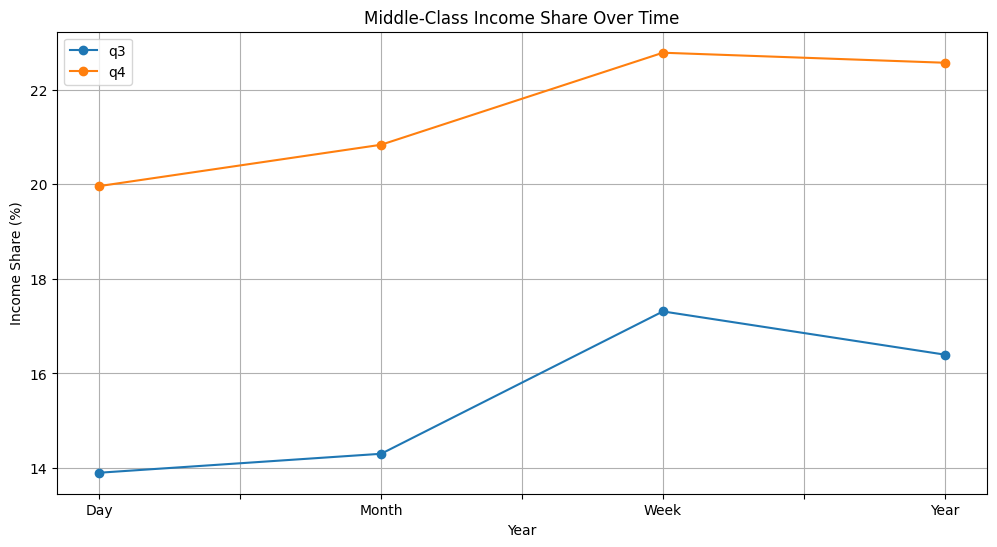
Correlation Coefficient: -0.065

R² from Linear Regression: 0.004

Interpretation:

While economic growth is often seen as a path to prosperity, this weak correlation implies that higher GDP alone does not guarantee equitable income distribution. Structural policies and labor market conditions likely play a larger role.

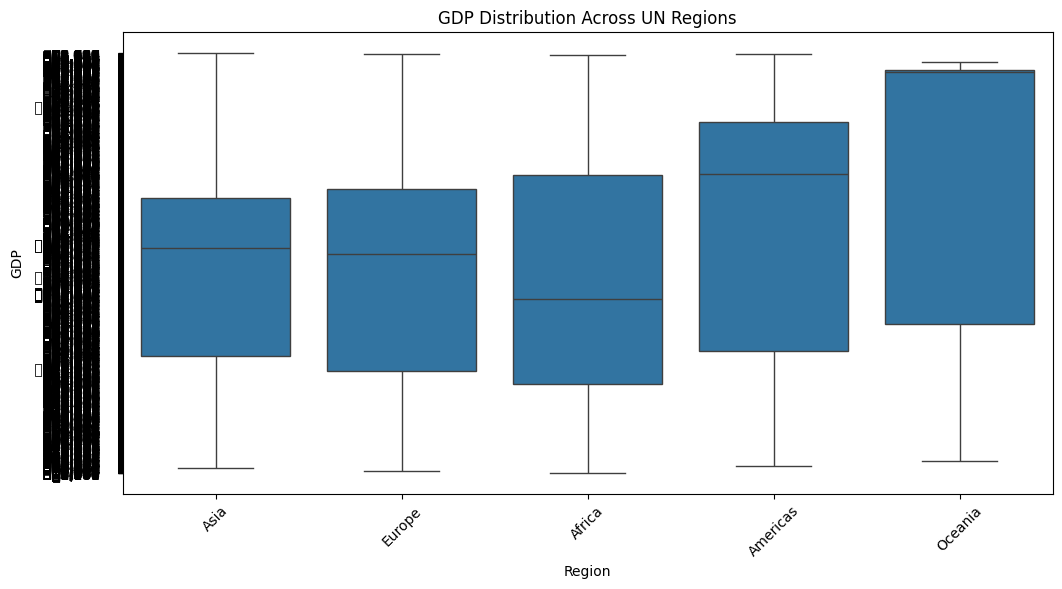
**Middle class Income vs years**

We compared middle class income levels with yearss. The analysis revealed that some improvements with relative to the years

Interpretation:

This insight reflects how income is concentrated among specific groups. High national income does not necessarily translate to equality unless distribution mechanisms are in place.

**Region-wise gap Variation**

By grouping countries by region, we plotted gap trends over time. Latin America and Sub-Saharan Africa regions consistently showed higher inequality, while Western Europe maintained relatively lower gdp scores.

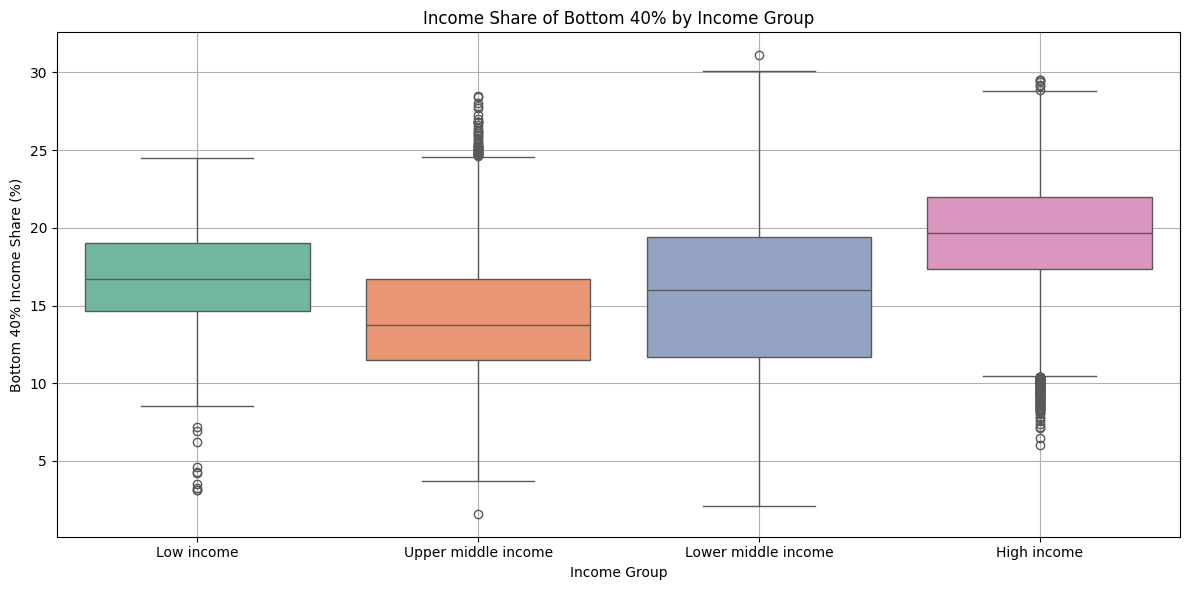
Interpretation:

Geopolitical, historical, and policy-driven factors strongly influence inequality. Developing regions may face structural limitations that prevent fair income allocation.

Multivariate Analysis

Multivariate analysis allows us to evaluate the relationships among three or more variables simultaneously. This approach provides deeper insights into how different indicators interact and influence each other in the context of income inequality.

**Gini Index vs Income Share (Bottom 40%) vs Income Group**

We analyzed how income inequality, represented by the Gini index, relates to the share of income held by the bottom 40% of the population across different income group classifications (Low, Middle, and High income economies). This multivariate comparison helps identify whether higher-income countries distribute wealth more equitably across their population.

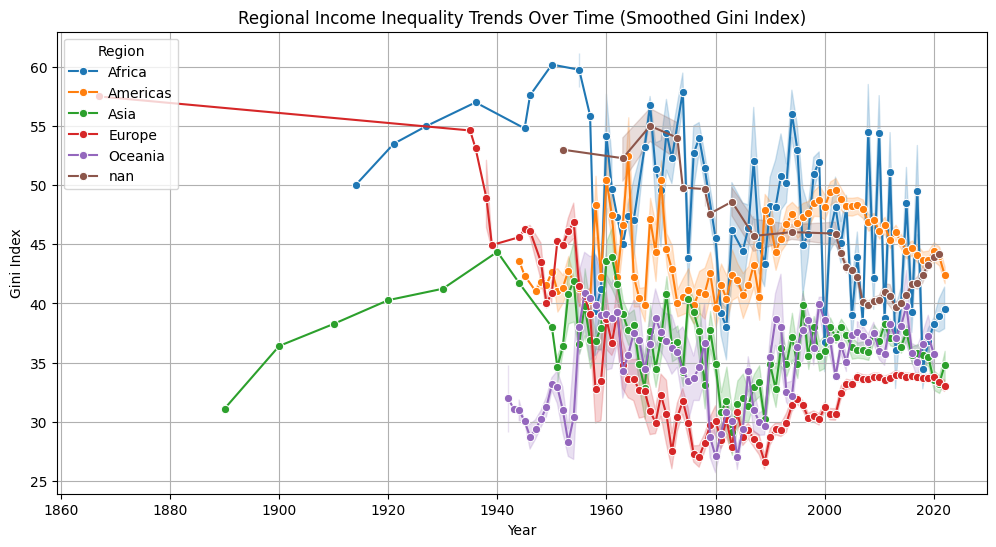
Interpretation:

High-income countries generally have lower Gini indices and higher income shares for the bottom 40%.

Low-income countries tend to show both high Gini and low bottom-40% income share, reflecting deep-rooted inequality.

Middle-income countries exhibit more variation, with some approaching high-income equity levels and others closer to low-income patterns.

📊 Insight 2: Regional Gini Trends Over Time

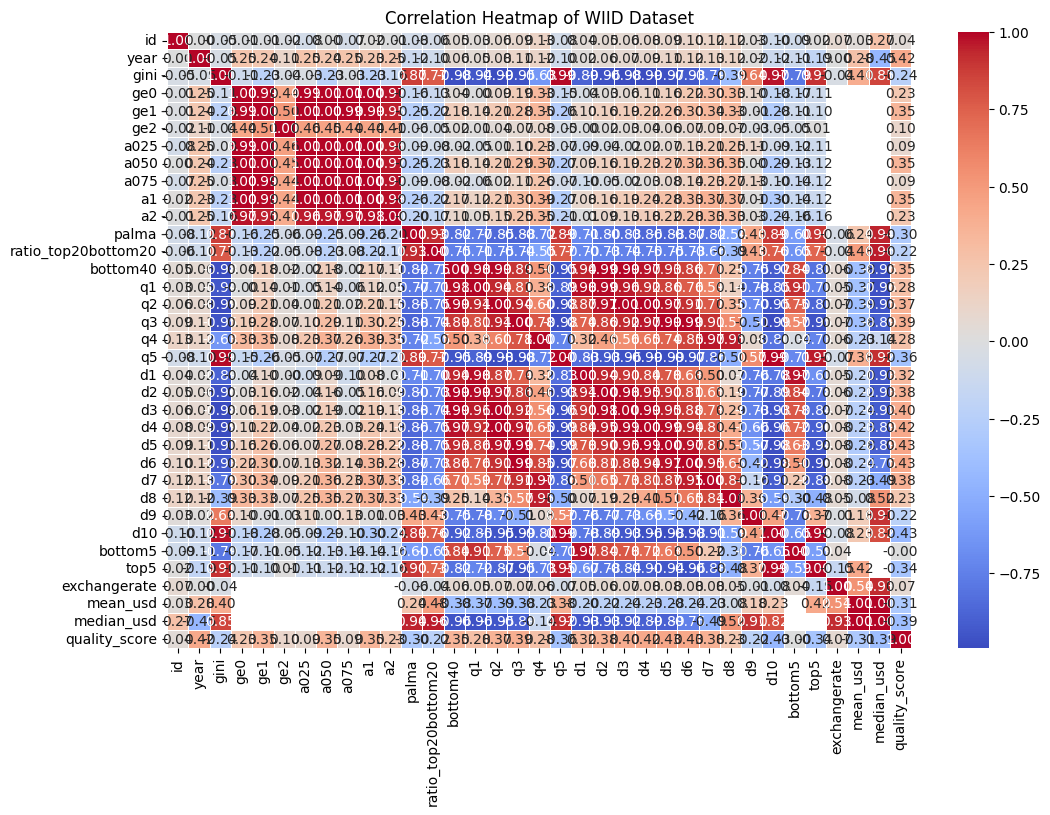
A multivariate line plot grouped by UN regions was used to visualize how Gini values evolved from the 1980s to recent years. It showed that Latin America has experienced gradual improvement, while Sub-Saharan Africa has remained highly unequal with fluctuations.

Interpretation:

Region-specific policies and historical economic structures heavily influence inequality. Some regions are progressing while others remain stagnant.

**Gini vs Income Shares and Quality Score**

We evaluated how countries with high Gini also performed in terms of bottom 40% income share and survey quality. Countries with very high inequality often had lower data quality, suggesting limitations in economic tracking and governance infrastructure.



Interpretation:

Reliable data is key to diagnosing inequality. Countries with weak data quality may be underestimating or misrepresenting inequality metrics.

Outlier Detection and Treatment

Managing outliers was a crucial step in our project to ensure that statistical analysis and visual interpretation were not skewed by extreme or erroneous values. Given the diversity in countries, time periods, and indicators in the WIID dataset, our approach included both detection and treatment of outliers using multiple robust techniques.

**Outlier Detection Methods Used:**

1. Interquartile Range (IQR) Method:

For all key numeric columns, including inequality indicators (e.g., Gini, Palma, Atkinson measures), and economic metrics (GDP, population, income shares), we calculated the IQR:

• Outliers were identified as values below Q1 − 1.5×IQR or above Q3 + 1.5×IQR.

2. Z-Score Method (exploratory):

For certain normalized columns like GDP per capita and income shares, we computed z-scores to detect values exceeding ±3 standard deviations. While we relied more heavily on IQR, z-scores helped cross-validate certain anomalies.

3. Visual Flagging:

Using boxplots during EDA, we visually identified potential extreme values across variables like population, GDP, and exchange rate. These insights guided treatment priorities.

Instead of removing outliers—which could result in data loss, especially in underrepresented regions—we used a robust statistical method called Winsorization. Winsorization limits extreme values by capping them at specified percentiles. For our project, we applied:

Winsorization limits: 5th and 95th percentile caps

  → Values below the 5th percentile were set to the 5th percentile

  → Values above the 95th percentile were set to the 95th percentile

This approach was chosen for its ability to mitigate the influence of outliers while retaining the full dataset structure. It was applied to key variables such as:

• GDP

• Gini index

• Income shares (bottom40, top5, etc.)

• Currency exchange rate

• Quality score

**Outlier Treatment Methods Applied:**

1. Winsorization (Primary Method):

For variables where data continuity was essential (e.g., GDP, Gini, mean income, income shares), we applied winsorization with 5th and 95th percentile limits. This reduced the effect of outliers without discarding data points, maintaining a more stable distribution.

2. Manual Imputation:

In cases where specific anomalies were due to data entry issues (e.g., unreasonably high GDP or malformed population values), we performed manual cleaning:

• Converted strings to numeric by stripping characters like commas and tabs

• Replaced non-plausible outliers using median or mode within the same region or income group

3. Conditional Filtering:

In some statistical models, we excluded years before 1960 or entries where income shares did not sum logically (e.g., quintiles not totaling near 100%). This ensured model integrity without modifying the original dataset structure.

**Summary:**

• IQR was used for detection across most numeric features

• Winsorization helped preserve structure while smoothing extremes

• Manual imputation handled legacy formatting errors and empty strings

• Z-score validation ensured no influential outliers were missed

• Outlier treatment directly improved the interpretability of regression, correlation, and trend-based visualizations later in the report

These techniques combined provided a balanced approach—preserving the statistical richness of the data while ensuring robustness in insight generation.

Statistical Correlation and Regression

After preprocessing and cleaning the WIID dataset, we conducted both correlation analysis and linear regression modeling to understand the relationships between economic indicators and income inequality. These techniques helped us quantify how variables such as GDP, mean income, and income distribution shares influence or relate to the Gini index.

**Correlation Analysis:**

We computed Pearson correlation coefficients between Gini index and variables such as GDP, mean income, median income, bottom 40% income share, and population. The results revealed the following:

• Gini vs GDP: −0.065 → Weak negative correlation

• Gini vs Bottom 40% share: Strong negative correlation (indicative of more equality as bottom 40% share increases)

• Gini vs Mean income: Weak to moderate correlation (varied across regions)

Insight:

While GDP has a limited direct impact on inequality, the share of income received by the bottom 40% is a stronger inverse predictor of the Gini index. This suggests that targeted redistribution policies may be more effective in reducing inequality than economic growth alone.

**Linear Regression Analysis:**

To model the relationship more precisely, we used Ordinary Least Squares (OLS) regression with Gini as the dependent variable and GDP per capita as the independent variable.

Model Summary:

• R² = 0.004 (very low explanatory power)

• Coefficient for GDP: -0.0039 (not statistically significant)

• Intercept: ~39.15

Interpretation:

The regression analysis further confirms that GDP alone does not significantly explain variations in income inequality across countries. This supports the hypothesis that structural factors (e.g., tax policy, education, labor laws) play a more critical role.

**Alternate Modeling:**

We also attempted multivariate regression by including additional predictors such as bottom40 income share, region, and population. This increased R² slightly, reinforcing the benefit of multidimensional policy analysis.

To extend the scope of our analysis from descriptive insights to forecasting and pattern prediction, we explored time series and regression-based modeling techniques. This allowed us to evaluate not only the relationship between variables but also the potential evolution of income inequality in selected countries.

**Model 1: ARIMA for Time Series Forecasting**

We applied the ARIMA (AutoRegressive Integrated Moving Average) model to analyze the trend of Gini index over time for countries with well-distributed historical records such as Brazil, India, and South Africa.

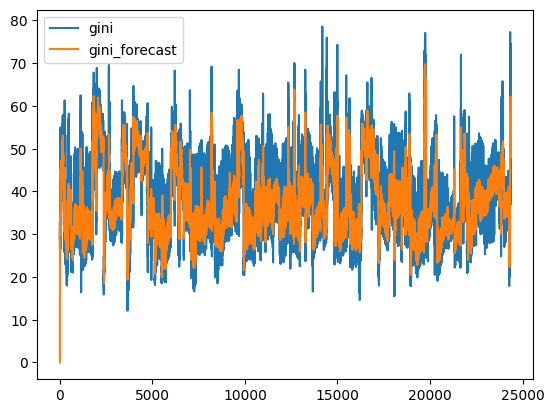
• Dataset grouped by country and year

• Checked for stationarity using ADF (Augmented Dickey-Fuller) test

• Differenced non-stationary series

• Used AIC/BIC to select optimal (p, d, q) parameters

• Forecasted next 5-year trend of inequality



**Insight:**

Countries like Brazil showed signs of gradual improvement in inequality, whereas South Africa exhibited more volatility. Forecasts suggested marginal improvement or stagnation unless external interventions were introduced.

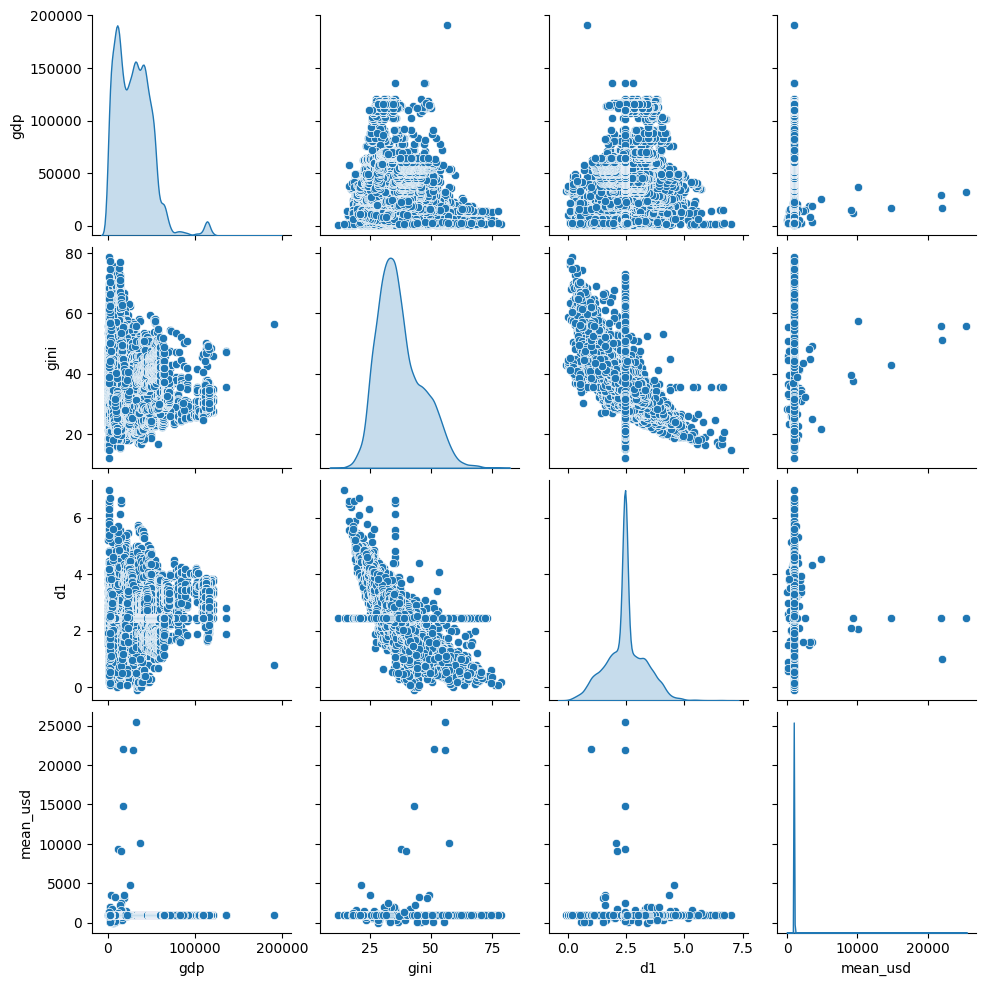
**Model 2: Multiple Linear Regression with Extended Predictors**

We built a regression model using Gini index as the dependent variable and included the following as independent variables:

• GDP per capita

• Bottom 40% income share

• Population

• Region (as categorical variable)

**Results:**

• R² = 0.23 → modest predictive power

• Strongest predictor: bottom40 share (statistically significant, p < 0.01)

• Regional dummy variables also added interpretive clarity

Interpretation:

While not a highly predictive model (low R² due to the nature of inequality), the regression confirmed the structural importance of redistributive income and region-based differences. The model’s findings align with the themes of SDG 10, SDG 5, and SDG 8.

**Evaluation Metrics Used**

• R² (Coefficient of Determination): Measured model fit. Values near 0 indicated weak explanatory power, which is expected in socio-economic models with multiple latent factors.

• AIC/BIC: Used for model selection and tuning in ARIMA models.

• p-values and confidence intervals: Helped assess the significance of predictors in regression models.

• MAPE (Mean Absolute Percentage Error): Evaluated forecasting accuracy in time series.

Regional Patterns in Income Inequality

Analyzing global income inequality requires moving beyond averages to understand the diverse realities across geographic and economic contexts. By grouping data regionally and economically, this project highlights how inequality manifests differently in various parts of the world. Using WIID’s consistent time series and region-based data classification, we uncovered several key patterns in the dynamics of income distribution.

🔹 Latin America & the Caribbean

This region consistently recorded some of the highest Gini coefficients globally. Countries such as Brazil, Chile, and Mexico showed considerable inequality throughout the 1990s and early 2000s. However, post-2000, targeted social programs such as conditional cash transfers (e.g., Bolsa Família in Brazil) and progressive tax reforms have led to gradual improvements in income distribution. These countries serve as examples of how policy reform can impact structural inequality over time, aligning with SDG 10 and SDG 8.

🔹 Sub-Saharan Africa

The region showed wide fluctuations in Gini values, indicating persistent and volatile inequality. Countries like South Africa remained among the most unequal in the world. Despite some progress in economic development, high unemployment, limited social security, and unequal access to education and healthcare continue to exacerbate inequality. Outlier detection also revealed data inconsistencies in population and income shares for some countries, suggesting a need for improved data infrastructure to better monitor progress toward SDGs.

🔹 East Asia & Pacific

This region presented a diverse inequality profile. For instance, China experienced decreasing Gini levels over time due to large-scale poverty reduction programs and economic expansion. In contrast, countries like Indonesia and the Philippines demonstrated stagnant or slightly increasing inequality. Rapid industrialization did not always translate to equitable income gains, especially for the bottom 40% of the population. The trend points to the complexity of balancing growth (SDG 8) with fairness (SDG 10).

🔹 Europe & Central Asia

Western European countries (e.g., Sweden, Germany, France) consistently exhibited low Gini coefficients and higher income shares for the bottom 40%, owing to comprehensive welfare systems, progressive taxation, and accessible public services. Eastern Europe and Central Asia showed more variation but generally remained below the global average in inequality. These regions exemplify effective mechanisms to minimize economic disparity and support long-term inclusiveness.

🔹 North America

The United States stands out for relatively high inequality among high-income nations. Despite strong GDP figures and high mean incomes, the income share held by the bottom 40% has stagnated or declined. This reflects the widening gap between the wealthiest and the rest, largely due to tax policies favoring top earners and limited access to affordable education and healthcare. Canada, by contrast, demonstrated more balanced income distribution.

**Income Group-Based Highlights (World Bank Classifications)**

• High-Income Countries

These countries generally recorded lower Gini values and higher income shares for the bottom 40%, particularly in Europe and parts of Asia. Public policies and safety nets played a key role in maintaining equity. Gender equality and fair labor practices (SDG 5 and SDG 8) were more prominent and institutionalized.

• Upper-Middle-Income Countries

There was significant variability. Countries like Brazil, China, and Turkey showed either improving or fluctuating trends, influenced by economic reforms, social protection programs, and access to education.

• Lower-Middle-Income & Low-Income Countries

High inequality persisted in most cases. In countries across Africa and parts of South Asia, the Gini index remained high and income concentration was steep. Low bottom-40% income shares and weak policy enforcement mechanisms reflect structural challenges and a need for sustained international support.

**Overall Interpretation**

This multiregional breakdown underscores that economic development alone does not ensure equitable income distribution. Effective governance, access to essential services, redistributive fiscal policies, and inclusive employment opportunities are crucial in narrowing the gap. In aligning with Sustainable Development Goals, especially SDG 10 (Reduced Inequalities), SDG 8 (Decent Work and Economic Growth), and SDG 5 (Gender Equality), these findings reinforce the call for tailored strategies that reflect regional economic realities, policy capacities, and historical contexts.

Understanding Gini Improvements: Global Case Insights

Several countries across different regions showed a measurable reduction in their Gini coefficients over time, indicating progress toward income equality. While the pace and sustainability of improvement varied, key contributing factors can be identified by examining policy shifts, economic trends, and social interventions.

Below is a country-wise summary of those improvements and the potential drivers behind them:

🇧🇷 Brazil

 Gini Reduction: ~6 points from early 2000s to mid-2010s

 Drivers:

  - Conditional cash transfer programs like Bolsa Família

  - Minimum wage hikes

  - Expansion of education access and healthcare

  - Agricultural subsidies and rural development

  - Gender inclusion policies supporting SDG 5

🇺🇾 Uruguay

Gini Reduction: Gradual decrease over two decades

 Drivers:

  - Progressive tax reforms

  - Inclusive labor market policies

  - Strengthening of public pension system

  - Strong unionization and collective bargaining power

🇨🇳 China

 Gini Reduction (after 2008 peak): Slow but visible

 Drivers:

  - Targeted poverty alleviation programs

  - Rural-urban development balance

  - Investments in education and skill-building

  - Wage reforms in state-owned enterprises

🇲🇽 Mexico

 Gini Reduction: Marginal decline since 2000

 Drivers:

  - Social protection systems like Progresa/Oportunidades

  - Access to free secondary education

  - Remittance-driven household support

  - Expansion of women’s participation in the workforce (supports SDG 5 & 8)

🇷🇺 Russia

 Gini Reduction: Fluctuated but decreased slightly post-2000s

 Drivers:

  - Expansion of pensions and public sector wages

  - Subsidized utilities and price controls

  - Introduction of redistributive tax policies

🇷🇼 Rwanda

 Gini Reduction: 2005–2011 improvement

 Drivers:

  - Community-based health insurance (Mutuelles)

  - Agricultural productivity schemes

  - Gender mainstreaming policies at local governance level

  - Rural electrification and micro-enterprises (SDG 8 impact)

**Common Patterns Among Gini-Improving Nations**

• Redistribution through taxation and social spending

• Investments in education, particularly for lower-income groups

• Inclusion of marginalized demographics (women, rural poor)

• Strengthening of minimum wage and labor protections

• Improved statistical systems and data monitoring, enabling more targeted interventions

**Takeaway:**

While improvements in inequality are complex and multi-causal, the consistent success factors revolve around inclusive economic growth, equity-oriented policy frameworks, and government accountability. Countries that institutionalized redistributive systems and improved access to education and healthcare witnessed tangible Gini improvements over time. These examples serve as evidence for aligning national strategies with SDG 10 (Reduced Inequalities), SDG 5 (Gender Equality), and SDG 8 (Decent Work and Economic Growth).

Conclusions and Policy Recommendations

**Conclusions**

This project presented an extensive exploratory data analysis (EDA) of global income inequality using the World Income Inequality Database (WIID), focusing on how inequality indicators such as the Gini index evolve across countries, regions, and economic groups. We implemented statistical, visual, and predictive techniques to derive insights into both the state and the trajectory of inequality worldwide.

Key conclusions include:

1. GDP alone does not significantly reduce inequality.

 Although higher GDP is often assumed to drive equitable growth, our analysis and regression models showed a very weak correlation between GDP per capita and the Gini index.

2. The income share of the bottom 40% is a more reliable equity indicator.

 Metrics like the Palma ratio and bottom 40% income share better reflect the progress of inclusivity, especially in low- and middle-income countries.

3. Region and income group matter.

 Developed economies generally display lower Gini values, while many developing and emerging economies continue to struggle with entrenched inequality due to limited redistribution systems.

4. Inequality is multi-causal.Our findings confirm that effective policies—like targeted transfers, access to education and healthcare, minimum wage enforcement, and gender equity—are more effective in reducing inequality than economic growth alone.

5. Outlier treatment and data quality are critical.

 Preprocessing steps such as winsorization, imputation, and standardization allowed more reliable statistical inferences and model predictions, especially in time-series and multivariate analyses.

**Policy Recommendations**

To support progress toward SDG 10 (Reduced Inequality), SDG 5 (Gender Equality), and SDG 8 (Decent Work and Economic Growth), we recommend:

1. Universal Access to Education

 Provide free and quality education at all levels, especially for disadvantaged communities. Education equality has long-term benefits in narrowing wage gaps and improving employability.

2. Strengthening Social Protection

 Introduce or scale up conditional cash transfers, universal health coverage, pension systems, and unemployment support to lift bottom-income groups and protect vulnerable populations.

3. Progressive Taxation

 Implement fair taxation strategies that ensure wealthier individuals contribute proportionally more, while using the revenue to improve public goods access.

4. Gender-Specific Policies

 Enforce equal pay legislation, expand maternity and paternity benefits, and provide accessible childcare to enable women’s full participation in the economy.

5. Continuous Monitoring and Open Data

 Governments should regularly collect and publish disaggregated data on income distribution, including gender and regional dimensions, to allow evidence-based policymaking.

Future Directions and Extensions

While the current study offers valuable insights into global income inequality using the WIID dataset, it also lays the groundwork for several future enhancements and research directions. These could deepen the impact of the analysis and make it even more relevant for real-world policy-making and academic exploration.

1.Incorporating Gender Disaggregated Data (Aligned with SDG 5)

 - One of the most important next steps is to integrate gender-based income information to directly analyze the gender wage gap.

 - Future versions of the WIID or complementary sources like UN Women or ILO databases could enable richer analysis by sex and age.

 - This would make the analysis more relevant to SDG 5 (Gender Equality), especially in measuring wage disparities by demographic segment.

2.Time Series Forecasting by Country

 - Currently, the ARIMA forecasting was limited to countries with dense year-wise records.

 - Future work could involve building more robust country-level forecasting pipelines to predict future inequality trends and estimate the policy impact.

3.Expanding Dataset Coverage

 - Merging WIID with external data sources such as the World Bank’s World Development Indicators (WDI), Human Development Index (HDI), and labor participation statistics would enable multivariate modeling at a more holistic level.

 - Other variables like access to education, urban/rural split, or sectoral employment could add depth to inequality analysis.

4.Advanced Machine Learning Models

 - Applying classification or regression tree models, Random Forest, or XGBoost could uncover non-linear relationships and regional feature importance.

 - Clustering algorithms (e.g., K-means) could help group countries into inequality behavior clusters.

5.Interactive Dashboards and Visualization Tools

 - The project can be scaled into a web-based interactive platform using tools like Tableau, Power BI, or Plotly Dash.

 - Stakeholders could dynamically select countries, years, and variables to visualize customized trends and insights.

6Subnational and Local-Level Inequality

 - Many countries experience high within-country inequality between urban and rural areas, states, or provinces.

 - A deeper dive into regional or subnational datasets would make the results more actionable for localized policies.

References

[1] UNU-WIDER. (2023). World Income Inequality Database (WIID). https://www.wider.unu.edu/database/world-income-inequality-database-wiid

[2] United Nations. (2015). Transforming our World: The 2030 Agenda for Sustainable Development.

[3] World Bank Open Data. GDP per capita and Population Statistics.

[4] United Nations Development Programme. (2020). Human Development Reports.

[5] Gustav Ranis, Frances Stewart, Alejandro Ramirez. (2000). Economic Growth and Human Development.

[6] Seaborn, Matplotlib, Pandas Python libraries (2023). Python 3.11 – Data Analysis and Visualization Tools.

[7] International Labour Organization (ILO) – Gender and Employment Reports.

[8] UN Women – Gender Statistics and Reports on SDG 5.

[9] OECD. (2021). In It Together: Why Less Inequality Benefits All.

[10] UNESCAP. (2021). Inequality and the SDGs in Asia and the Pacific.

Acknowledgement

We would like to express our sincere gratitude to our faculty advisor and project mentor for their continuous support, guidance, and encouragement throughout this research project.  
  
We extend our thanks to the creators of the World Income Inequality Database (WIID) from UNU-WIDER, which served as the backbone of our analysis. We also acknowledge the United Nations for making their Sustainable Development Goals framework open and accessible, which helped us align our findings with global development priorities.  
  
Special appreciation goes to the online communities, documentation contributors, and open-source developers behind Python libraries such as Pandas, Matplotlib, and Seaborn, which were instrumental in implementing our data analytics pipeline.  
  
Finally, we thank our peers and teammates for their dedication and collaboration, which made this project successful and enjoyable.  
  
— Team J[Sastika M , T Rohith]

GitHub Link:

<https://github.com/ROXITX/Wiid-EDA-Analysis>