Research Paper - GNEC Spring Hackathon 2025

Vox Aequalis: using machine learning to predict and visualize gender and social disparities.

By:

Avila, James Yuri R.

May 2025

Introduction

As society continues to progress, the job market or in general remains persistently unequal. Accurately identifying and visualizing disparities in employment—particularly or in general aspect as those related to gender and income—can provide valuable insights for stakeholders, policy makers, and users of Project Vox Aequalis.

This research aims to leverage historical and contemporary data to model and visualize inequalities in the job market. Specifically, the project builds regression and ensemble models to predict trends and future percentages in gender-based wage gaps and income inequality. Visual analytics techniques were also implemented to make patterns interpretable and interactive.

Datasets

- Gender Pay Gap Population Survey.csv
- Inequality in Income.csv
- Global Income Inequality.csv
- LabourForce.csv
- Gender Inequality Index.csv

Data Descriptions – Some data are in multiple dataset

• **Year** – year of the data

• **Gender Wage Gap %** – positive and negative percentage of wage gap

• **Inequality in Income 2010 – 2021** – inequality values per year from 2010 to 2021

• **Country** – the geographic reference

• **Gini Index** – statistical measure of income inequality

• **Average Income (USD)** – income averaged over the country

• **Top 10% Income Share (%)** – income share of the top 10%

• **Bottom 10% Income Share (%)** – income share of the bottom 10%

• **Population** – national population for the given year.

• **OBS Value** – Observation Value or a value of a particular variable.

Data Preparation

Prior to modeling, the datasets underwent the following preparation steps:

- **Handling Missing Data:** Missing entries were filled or removed depending on severity and column importance
- **Chronological Sorting:** Data was arranged from oldest to most recent
- **Feature Reduction:** Low-impact or redundant columns were dropped after correlation checks
- **Data Filtering:** Data with extreme outliers or obvious inconsistencies were discarded to improve model integrity

This clean and structured data served as the foundation for regression-based modeling.

Methodology

To predict future inequalities and visualize patterns in global and gender-based disparities, this research utilized **supervised learning models** specifically suited to each dataset. Both **linear regression**, **multiple linear regression**, **pipeline** and **random forest regression** were selected to accommodate different data structures and modeling goals. Visualization techniques were used to support interpretation of the results.

1. Gender Pay Gap Prediction (Linear Regression)

Dataset: Gender Pay Gap Population Survey.csv

Dataset Origination: Kaggle

Objective:

Forecast gender wage gap percentages from 2025 to 2030 using past data.

Model Type:

Linear Regression, ideal for modeling trends over time.

Input Features:

- Year

Target Variable:

- Gender Wage Gap (%)

Visualizations:

- Scatterplot with regression line to show predicted future gap values
- Bar plot showing average wage gaps by year and gender
- Line graph illustrating historical trend and projected future change

2. Income Inequality Forecasting

Dataset: Inequality in Income.csv

Dataset Origination: Kaggle

Objective:

Predict **income inequality levels** from 2026 to 2030 based on trends from 2010–2021.

Model Type:

Multiple Linear Regression

Input Features:

- Historical inequality values from 2010 to 2021

Target Variable:

- Forecasted inequality from 2026 to 2030

Visualizations:

- Line chart showing actual and predicted inequality values
- Heatmap to show correlations between yearly inequality values
- Distribution plots of inequality levels over time

3. Global Income Inequality Modeling (Random Forest Regression)

Dataset: Global Income Inequality.csv

Dataset Origination: Kaggle

Objective:

Model and predict the Gini Index using global socioeconomic indicators.

Model Type:

Random Forest Regression, selected for its ability to handle non-linear relationships and feature interactions.

Input Features:

- Average Income (USD)
- Population
- Top 10% Income Share (%)
- Bottom 10% Income Share (%)

Target Variable:

- Gini Index

Visualizations:

- Feature importance plot to determine influential factors
- Predicted vs actual scatterplot to assess model accuracy
- Boxplot of Gini index by income groups
- Pairplot to explore relationships among features

4. Gender Inequality Index (Linear Regression)

Dataset: Gender Inequality Index.csv

Data Origination: World Bank Group

Objective:

Model and predict the OBS Value using global socioeconomic indicators.

Model Type:

Linear Regression, ideal for modeling trends over time

Input Features:

- Year Period

Target Variable:

- OBS Value

Visualizations:

Time Series Plot to observe fluctuation in gender inequality

Histogram to show distribution of residuals

Forecast to show inequality OBS value from 2025 to 2028.

5. Labor Force (Pipeline Random Decision Forest)

Dataset: Labor Force.csv

Data Origination: World Bank Group

Model Type:

Random Forest Regression, selected for its ability to handle non-linear relationships and feature interactions.

Input Features:

- Year Period

Target Variable:

- OBS Value

Visualizations:

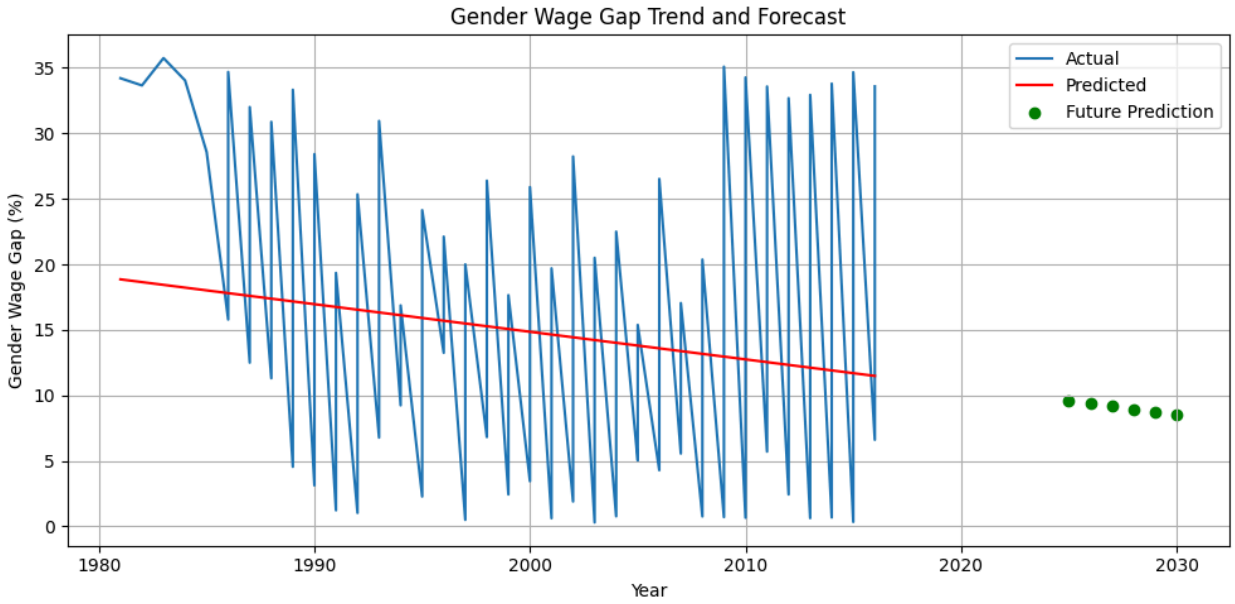
Box Plot to displays the distribution of the labor force participation across different genders.

Result and Analysis

The results of the analysis and interpretation of the results of the machine learning model for determining and visualizing inequality as well as predicting future inequalities.

1. Gender Pay Gap Prediction (Linear Regression)

Time series forecast plot



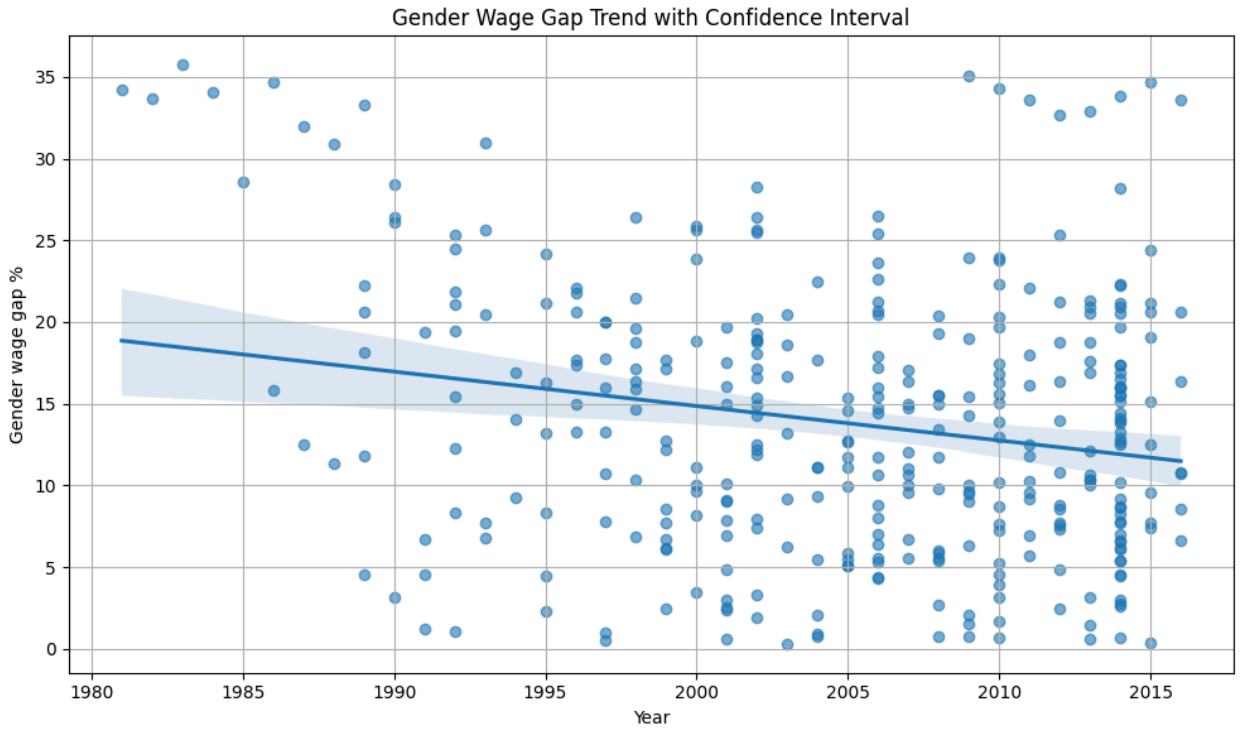
- **Actual Data (Blue Line):** The blue line represents the actual observed gender wage gap from around 1980 up to approximately 2016. We can see a generally decreasing trend in the wage gap over this historical period, although it fluctuates significantly year to year. The fluctuations suggest some cyclical or short-term factors influencing the wage gap.
- **Predicted Trend (Red Line):** The red line shows the predicted trend of the gender wage gap extending from the early 1980s to 2030. This line presents a smoother, more consistent downward trajectory, indicating a projected long-term decrease in the gender wage gap.
- **Future Prediction (Green Dots):** The green dots represent specific future predictions for the gender wage gap from

approximately 2024 to 2030. These dots align closely with the predicted trend line, suggesting a continuation of the anticipated decline in the wage gap in the coming years.

Summary:

The chart indicates that while the actual gender wage gap has historically fluctuated, there has been an overall downward trend from around 1980 to 2016. The forecast suggests that this decreasing trend is expected to continue into the future, with the predicted trend line and specific future predictions showing a further reduction in the gender wage gap through 2030. However, it's important to remember that these are predictions, and the actual future wage gap could be influenced by various economic and social factors not explicitly accounted for in this model.

Scatter Plot



This scatter plot illustrates the trend of the gender wage gap (in percentage) over time, from approximately 1980 to 2016. Each blue dot represents the observed gender wage gap for a specific year.

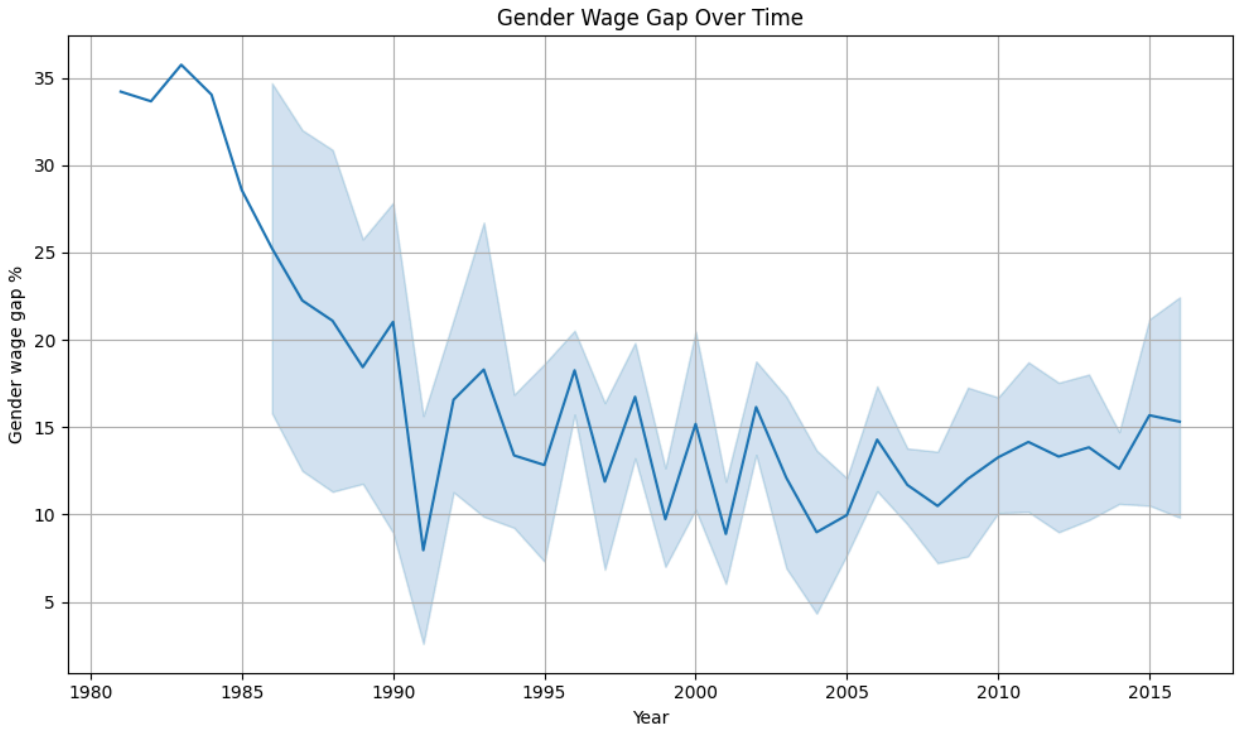
A solid blue line is overlaid on the scatter plot, representing the general trend of the gender wage gap over this period. You can see that this line has a downward slope, suggesting an overall decrease in the gender wage gap as the years progress.

The shaded light blue area surrounding the trend line represents the confidence interval. This interval provides a range within which we can be reasonably confident that the true trend of the gender wage gap lies. The width of the confidence interval reflects the uncertainty associated with the estimated trend; a wider interval indicates more variability in the data and thus greater uncertainty.

Summary:

The graph indicates a general downward trend in the gender wage gap between 1980 and 2016. While there is considerable variability in the wage gap from year to year (as shown by the scattered blue dots), the overall tendency is a decrease. The confidence interval around the trend line provides a measure of the uncertainty associated with this estimated decrease.

Line **Graph**

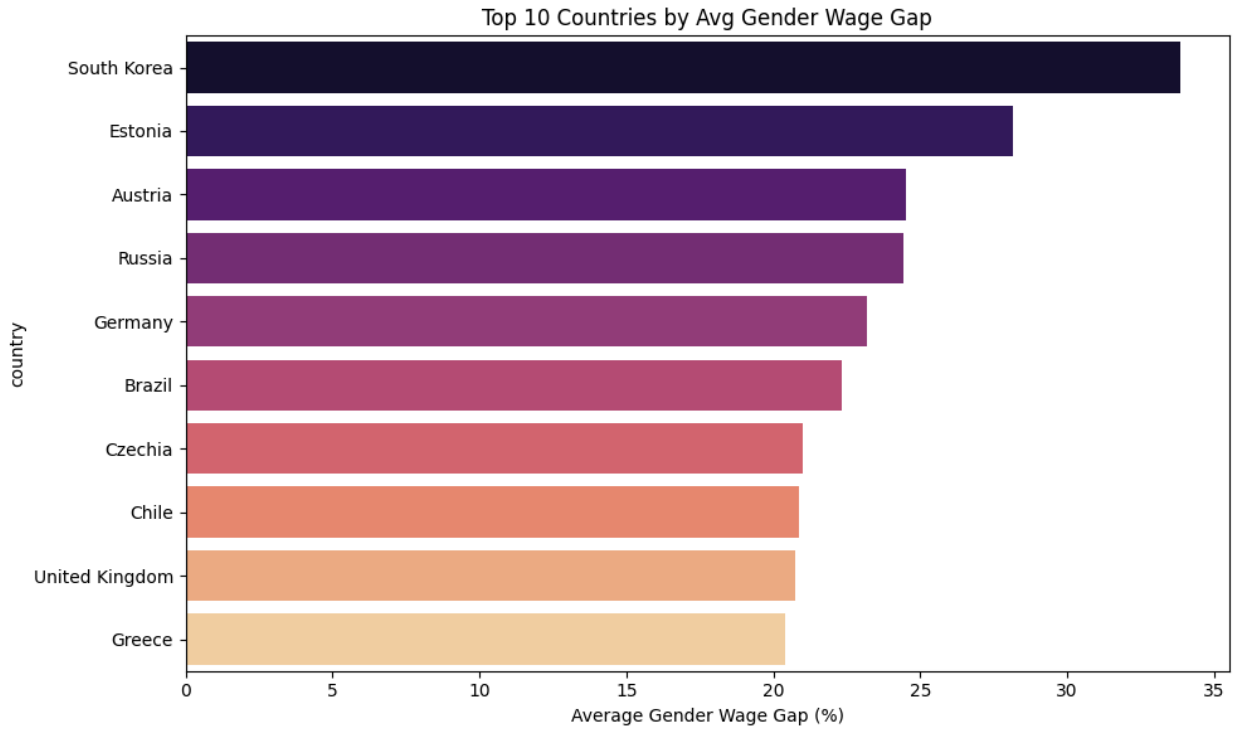


This line graph displays the trend of the gender wage gap (in percentage) from approximately 1980 to 2016. The blue line represents the estimated gender wage gap for each year. The light blue shaded area around the line indicates a measure of uncertainty or variability, likely representing a confidence interval or standard deviation around the estimated wage gap.

Summary:

The graph shows a fluctuating but generally decreasing trend in the gender wage gap from 1980 to around the mid-1990s. After that period, the wage gap appears to stabilize at a lower level with continued, though less dramatic, fluctuations. The shaded area highlights the variability or uncertainty associated with these estimates, suggesting a range within which the actual gender wage gap might have fallen each year.

Bar Chart



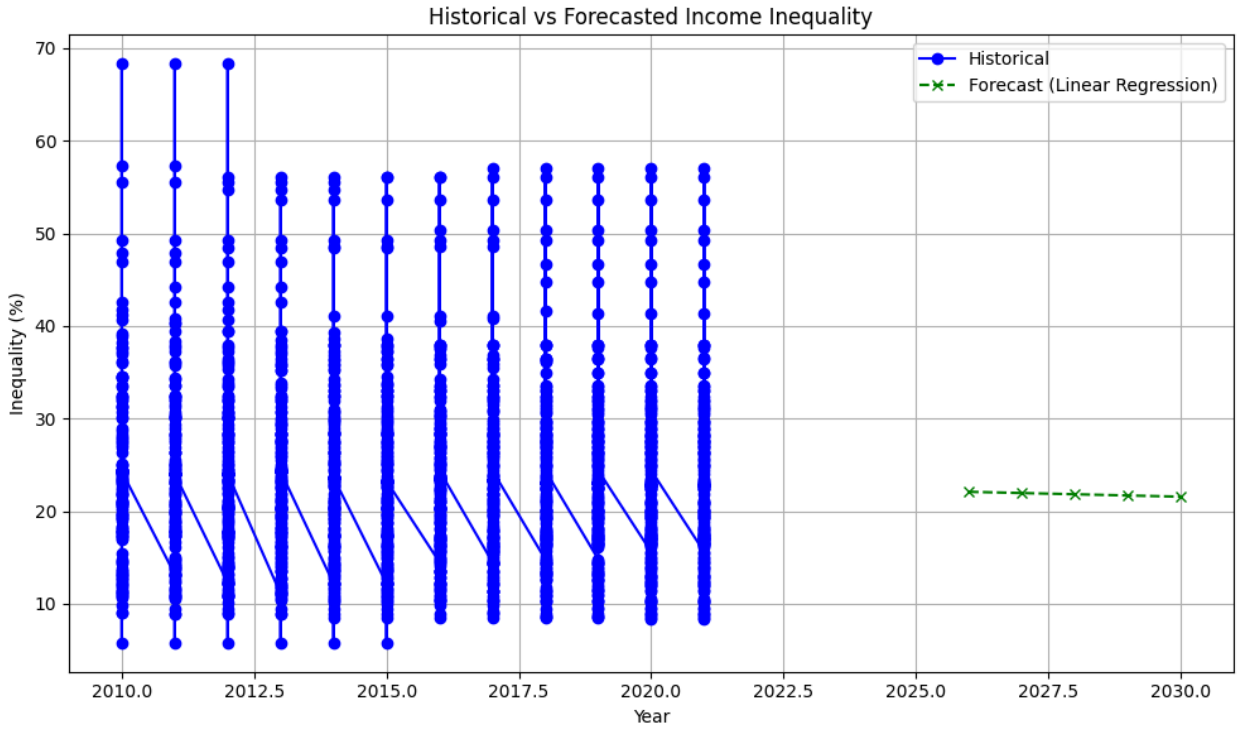
This horizontal bar chart displays the "Top 10 Countries by Average Gender Wage Gap". The length of each bar represents the average gender wage gap (in percentage) for that specific country. The countries are listed on the vertical axis, and the corresponding average wage gap percentage is shown on the horizontal axis. The bars are also color-coded with a gradient, transitioning from a lighter shade for Greece to a darker shade for South Korea, visually emphasizing the increasing wage gap.

Summary:

The chart highlights the ten countries with the largest average gender wage gaps. South Korea exhibits the highest average gender wage gap among the top 10, with Estonia following as the second highest. Greece has the lowest average gender wage gap within this top 10 list. The chart clearly ranks these countries based on this metric, allowing for a direct comparison of the magnitude of the gender wage gap across them.

2. Income Inequality Forecasting

Line Graph



- **Historical Data (Blue Line with Circles):** The blue line with circular markers represents the historical income inequality from 2010 up to around 2021. The data shows significant fluctuations within each year, indicated by the vertical lines connecting multiple data points for the same year. Despite these yearly variations, there isn't a clear upward or downward trend visible in the historical data. The level of income inequality seems to oscillate within a certain range.
- **Forecast (Green Dashed Line with Crosses):** The green dashed line with cross markers represents the forecasted income inequality from approximately 2026 to 2030, labeled as "Forecast (Linear Regression)". This forecast shows a relatively stable and slightly increasing trend in income inequality over this future period.

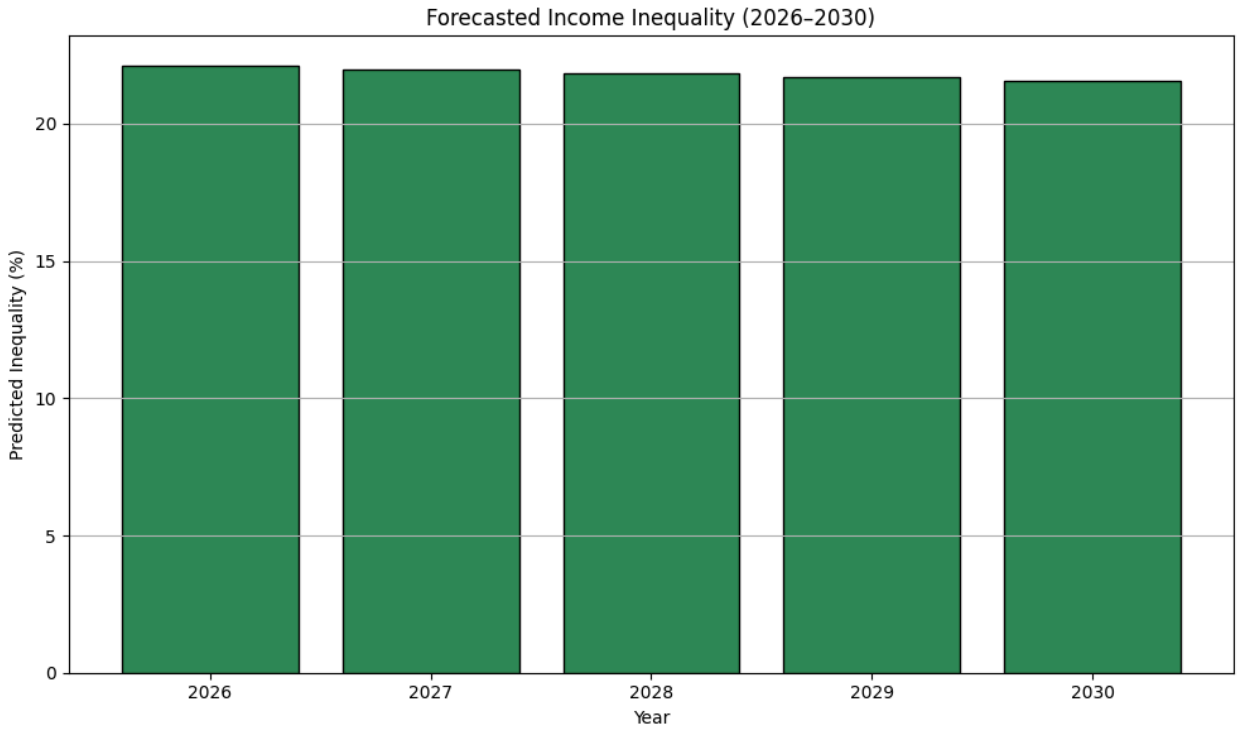
Analysis:

- The historical data reveals substantial within-year variability in income inequality. This could be due to seasonal factors, policy changes within the year, or other economic dynamics.
- The absence of a clear long-term trend in the historical data up to 2021 suggests that income inequality has remained relatively consistent during this period, despite the short-term fluctuations.
- The linear regression forecast predicts a slight increase in income inequality from 2026 to 2030. This suggests that based on the historical data and the chosen linear regression model, a gradual rise in income disparity is anticipated in the latter part of the decade.
- There is a gap between the end of the historical data (around 2021) and the start of the forecast (around 2026). This indicates that the forecasting model is projecting several years into the future without intermediate predictions shown on this graph.

Summary:

The graph illustrates that historical income inequality between 2010 and 2021 experienced significant yearly fluctuations but lacked a distinct long-term trend. A linear regression forecast from 2026 to 2030 suggests a slight upward trajectory in income inequality. The model predicts a gradual increase in income disparity in the latter part of the forecast period, following a period where historical data showed considerable short-term variability but overall stability.

Bar Chart

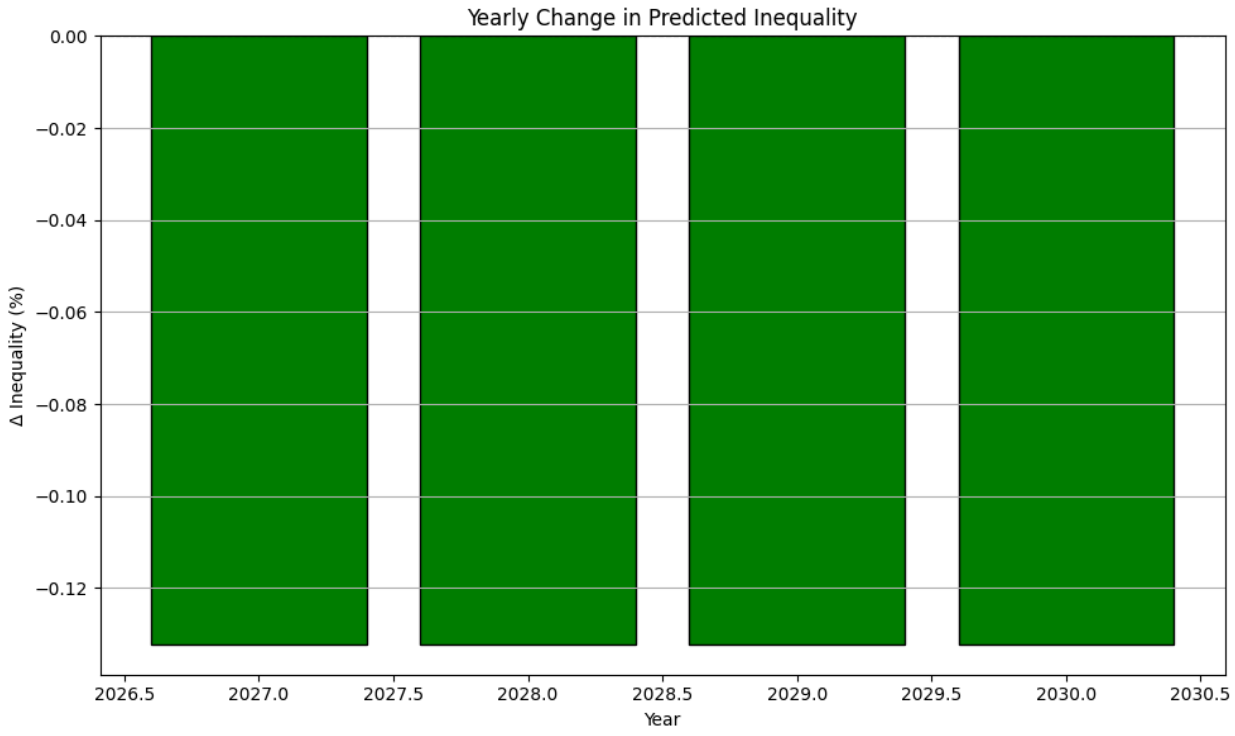


The chart shows that the forecasted income inequality is relatively stable across the five-year period. The height of each bar is approximately the same, indicating that the model predicts only minor fluctuations in income inequality between 2026 and 2030.

Summary:

Based on this forecast, income inequality is expected to remain fairly consistent from 2026 to 2030. There is no significant upward or downward trend predicted within this timeframe. The level of predicted income inequality hovers around 22-23% for each of these years.

Bar Chart



All the bars are green and extend downwards from the zero line. This indicates a negative yearly change in predicted income inequality for each of the years shown. The height of each bar appears to be roughly the same, suggesting a consistent decrease in the predicted inequality from one year to the next within this period. Specifically, the change seems to be approximately -0.13% each year.

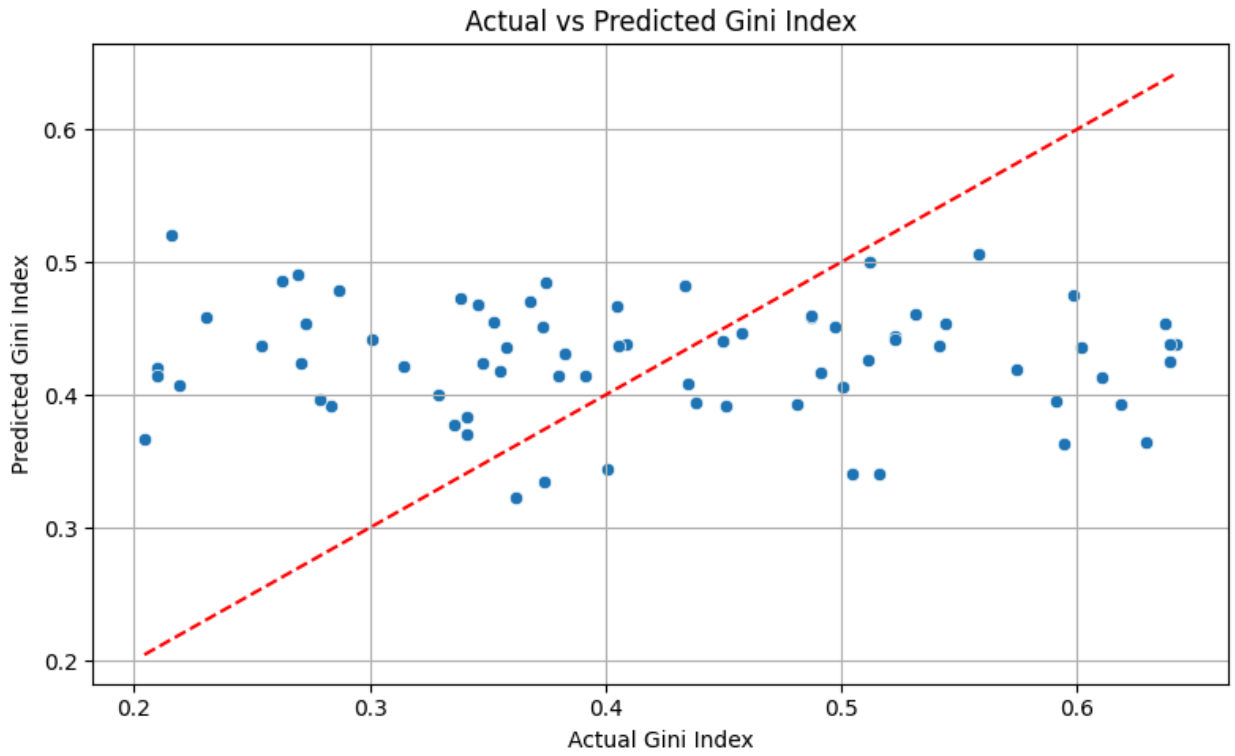
	Year	LR_Prediction	Yearly_Change
0	2026	22.096273	NaN
1	2027	21.964070	-0.132202
2	2028	21.831868	-0.132202
3	2029	21.699666	-0.132202
4	2030	21.567463	-0.132202

Summary:

The chart shows a consistent year-over-year decrease in the predicted income inequality from 2026 to 2030. The model forecasts a reduction of about 0.13 percentage points in income inequality each year during this period and show 0.66 decrease of income inequality in 5 years.

3. Global Income Inequality Modeling (Random Forest Regression)

Scatter Plot



- **Scatter of Points:** The blue dots are scattered around the red dashed line, indicating that the predictions are not perfectly accurate. Some predicted values are higher than the actual values (points above the red line), while others are lower (points below the red line).
- **Deviation from the Ideal Line:** The extent to which the blue dots deviate from the red dashed line visually represents the magnitude of the prediction errors. Larger deviations indicate less accurate predictions.
- **Overall Trend:** While there's scatter, the points generally seem to follow a positive correlation. This suggests that as the actual Gini Index increases, the predicted Gini Index also tends to increase. However, the spread indicates a considerable degree of error in the predictions.
- **Concentration of Points:** There appears to be a denser cluster of points in the middle range of actual Gini Index

values (roughly between 0.3 and 0.5), suggesting that the model might perform differently across different ranges of the Gini Index.

Summary:

The plot shows that the model's predictions of the Gini Index have a positive correlation with the actual values, meaning the model generally captures the direction of change. However, the predictions are not highly accurate, as indicated by the significant scatter of points around the ideal prediction line. There's a noticeable degree of error in the predictions across the range of Gini Index values, and the model's performance might vary depending on the actual Gini Index level.

Histogram

- **Frequency Distribution:** The histogram shows how the Gini Index values are spread across the observed range. We can see that certain ranges of the Gini Index have a higher frequency of occurrence than others. For example, the bars

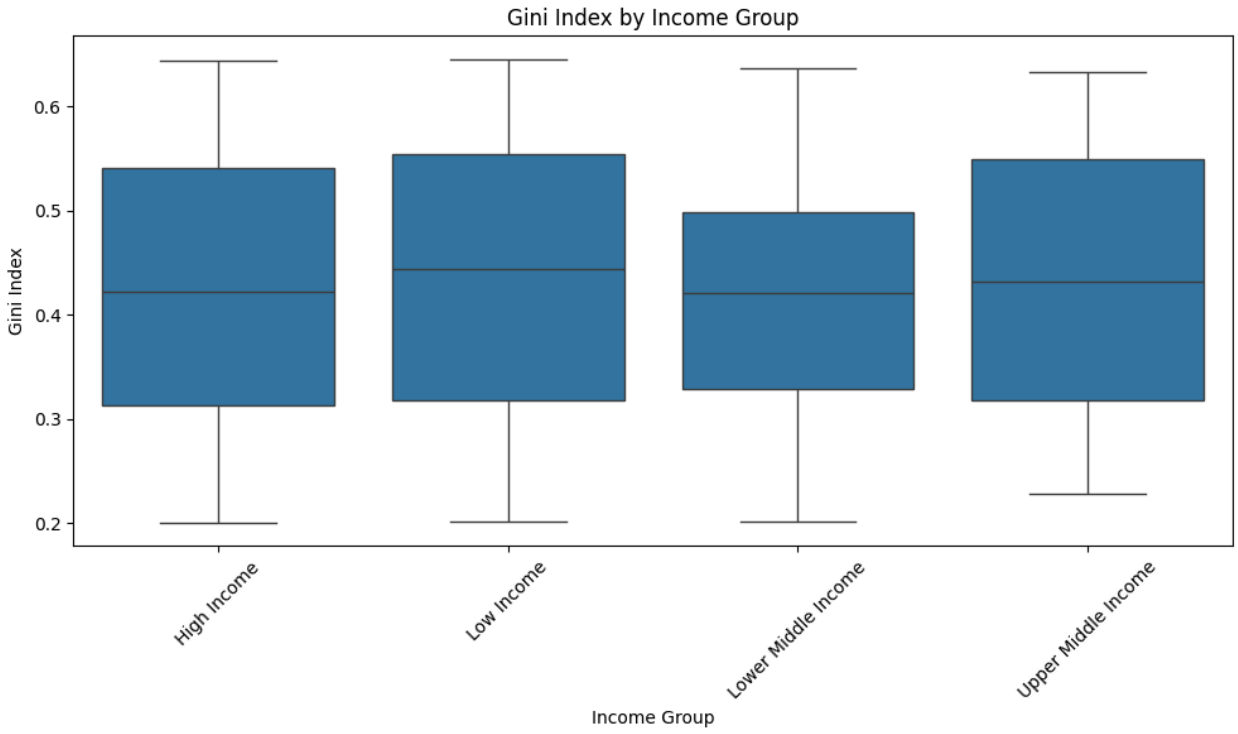
around the 0.2 and 0.55 marks appear to be taller, indicating a higher concentration of Gini Index values in those ranges within the dataset.

- **Spread of Data:** The data spans a range of Gini Index values from approximately 0.2 to 0.65.
- **Kernel Density Estimate:** The smooth purple curve provides a clearer picture of the underlying distribution shape, smoothing out the discrete bars of the histogram. It suggests potential peaks or modes in the distribution, indicating where the Gini Index values are most likely to occur. In this case, there appear to be a couple of potential peaks, around 0.2 and 0.55.

Summary:

The histogram illustrates the distribution of Gini Index values, showing how frequently different levels of income inequality occur in the dataset. The distribution appears to be somewhat multi-modal, with higher frequencies of Gini Index values observed around 0.2 and 0.55. 0.2 Gini shows some countries have better Gini index while some countries with 0.55 Gini Index reveal worse income inequality.

Box Plot



- **The box:** Represents the interquartile range (IQR), containing the middle 50% of the Gini Index values. The bottom of the box is the 25th percentile (Q1), and the top is the 75th percentile (Q3).
- **The horizontal line inside the box:** Represents the median (50th percentile) Gini Index value for that income group.
- **The whiskers:** Extend from the box to show the spread of the remaining data, typically up to 1.5 times the IQR. Values beyond the whiskers might be considered outliers (though no outliers are explicitly marked in this plot).
- **The vertical length of the box and whiskers:** Indicates the variability or dispersion of the Gini Index within each income group.

- **The median Gini Index** appears to be relatively similar across all four income groups, falling roughly between 0.4 and 0.45.
- **The spread of Gini Index values** (indicated by the height of the boxes and the length of the whiskers) varies across the groups. "Low Income" and "Upper Middle Income" groups seem

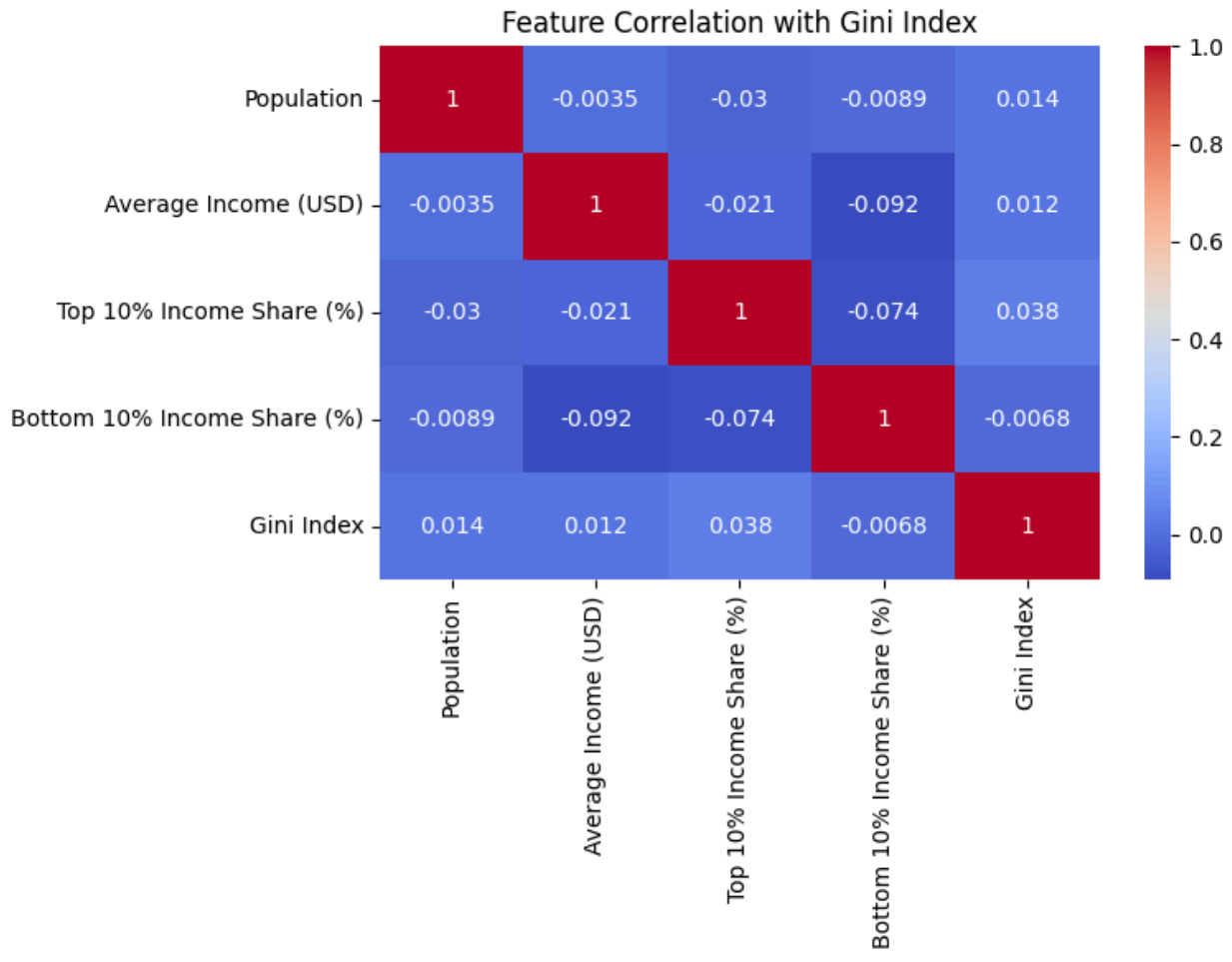
to have a slightly wider spread compared to "High Income" and "Lower Middle Income" groups.

- The **"High Income"** group appears to have a slightly lower upper quartile compared to the other groups.

Summary:

This box plot analysis suggests that while the median Gini Index (a measure of income inequality) is fairly consistent across High Income, Low Income, Lower Middle Income, and Upper Middle Income countries, the variability in income inequality within these groups differs. Low-income and upper-middle-income countries exhibit a slightly broader range of Gini Index values compared to high-income and lower-middle-income countries. Overall, the level of income inequality, as measured by the Gini Index, does not show a strong differentiation based solely on these broad income classifications.

Correlation Heat Map



- **Population:** Shows a very weak positive correlation (0.014) with the Gini Index.
- **Average Income (USD):** Exhibits a very weak positive correlation (0.012) with the Gini Index.
- **Top 10% Income Share (%):** Displays a weak positive correlation (0.038) with the Gini Index.
- **Bottom 10% Income Share (%):** Shows a very weak negative correlation (-0.0068) with the Gini Index.

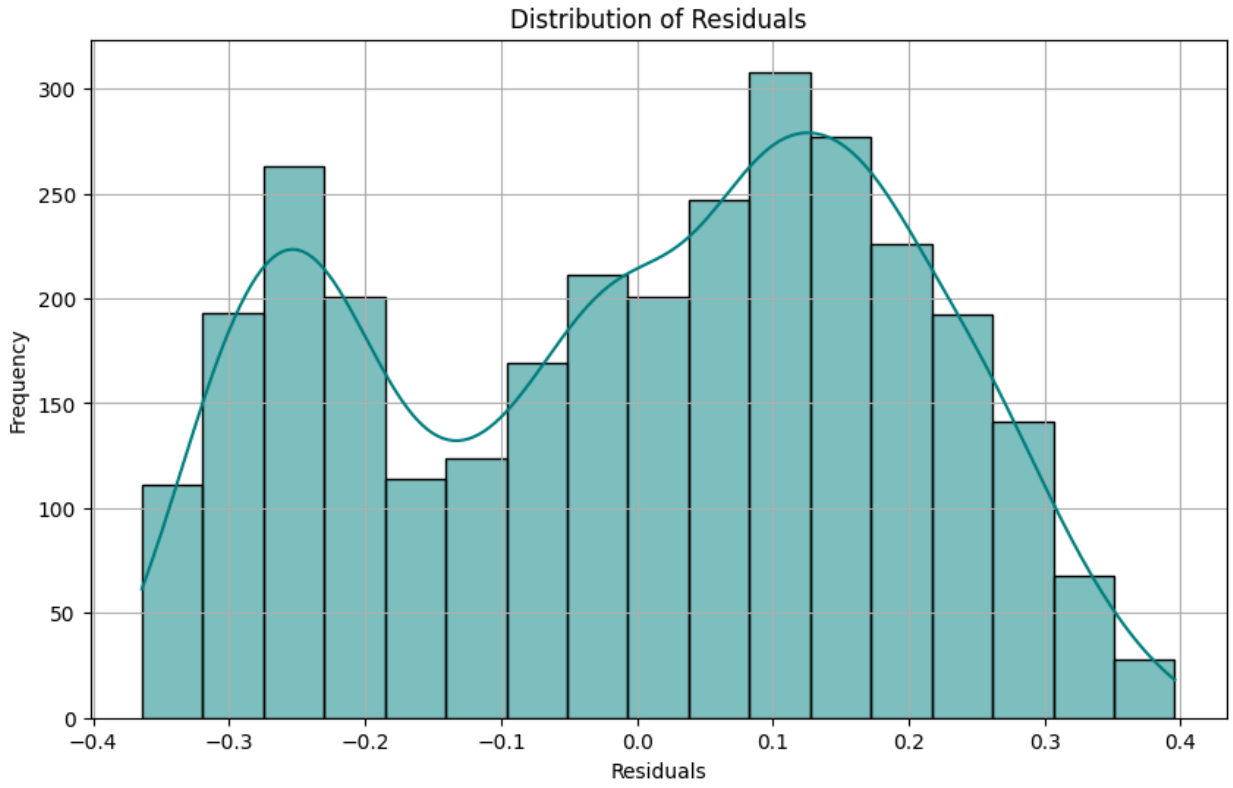
The diagonal cells show a perfect positive correlation (1) of each feature with itself, as expected. The other correlation coefficients between the independent features are also relatively small, suggesting low linear relationships between them.

Summary:

The heatmap indicates that the Gini Index has very weak linear correlations with Population, Average Income, and Bottom 10% Income Share. It shows a slightly stronger, but still weak, positive linear correlation with the Top 10% Income Share. This suggests that, based on this linear correlation analysis, none of these individual features are strong predictors of the Gini Index. The weak positive correlation with the top 10% income share implies a slight tendency for higher income inequality (higher Gini Index) when the income share of the richest 10% is higher. Conversely, the correlation with the bottom 10% income share is negligible and negative. The low correlations between the independent features suggest minimal multicollinearity issues in potential models using these variables.

4. Gender Inequality Index (Linear Regression)

Histogram



- **X-axis (Residuals):** Shows residual values. Zero means a perfect prediction. Positive values = underprediction; negative = overprediction.
- **Y-axis (Frequency):** Indicates how often residuals fall within each range.
- **Bars:** Represent frequency of residuals in specific intervals.
- **Curve:** The teal curve is a smoothed estimate of the residual distribution.

- **The histogram** displays a possibly bimodal distribution of residuals, ranging from -0.4 to 0.4, with peaks around -0.25 and 0.12, suggesting a non-ideal error pattern in the model's predictions.

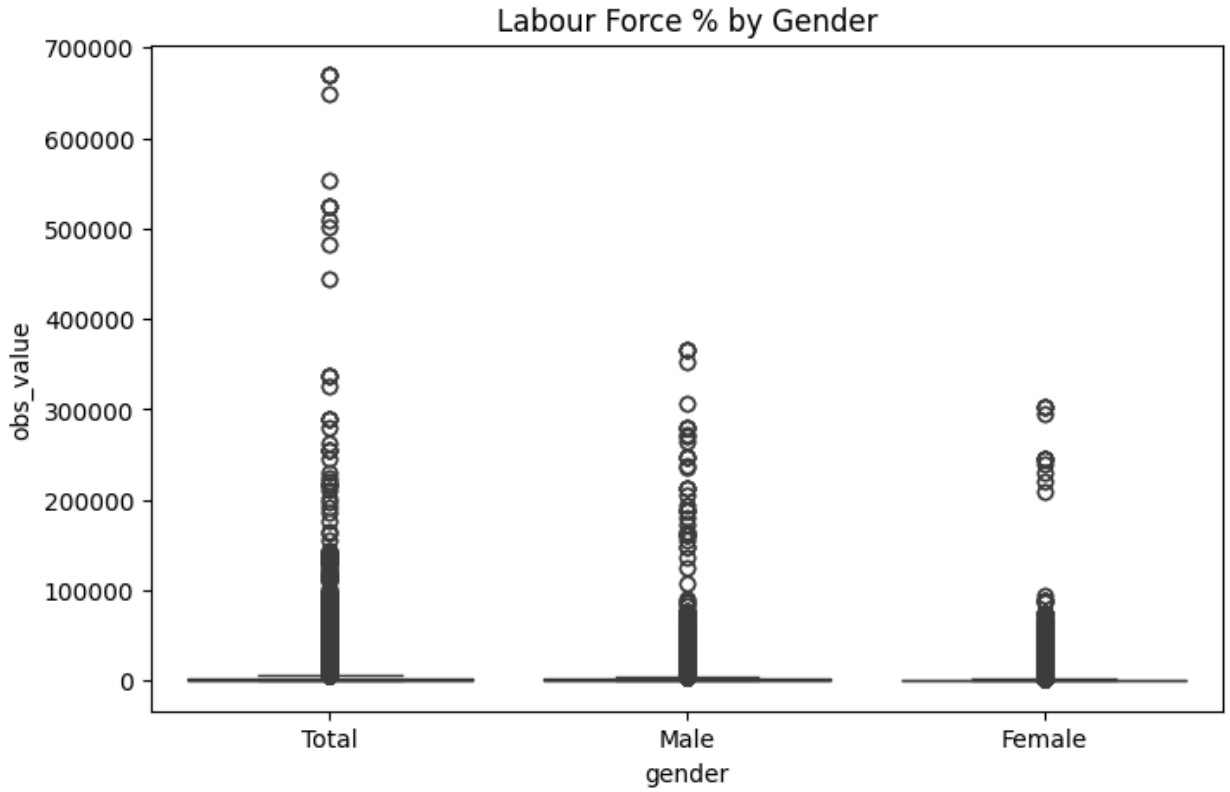
```
Forecasts for Gender Inequality (OBS_VALUE):

 TIME_PERIOD  Forecast_OBS_VALUE
        2025            0.345240
        2026            0.342078
        2027            0.338916
        2028            0.335754
```

Summary:

Displays that gender Inequality decreases around 0.33 per year or 0.96% decrease in 5 years in accordance to the models forecast.

5. Labor Force (Pipeline Random Decision Forest)

Box Plot



```
Average Labour Force % by Gender:
gender
Female    1856.424112
Male      2688.321925
Total     4385.242180
Name: obs_value, dtype: float64
```

Summary:

The box plot reveals a generally low typical labor force participation across genders, contrasted by substantial variability and numerous high outliers, particularly in the 'Total' category. Additionally, the average labor force participation shows a notable gender disparity, with males (2688.32) exhibiting a considerably higher average than females (1856.42), contributing to a total average of 4385.24.

Results and Summary

This section presents the findings derived from the application of machine learning models to explore and forecast inequality in income and gender-based wage gaps, along with visual analytics to interpret the trends more effectively.

1. Gender Pay Gap Prediction (Linear Regression)

Findings:

- Historical data from ~1980 to 2016 reveals a generally decreasing trend in the gender wage gap, although year-to-year fluctuations were notable.
- The linear regression model projects a continued reduction in the wage gap from 2024 to 2030.
- Green prediction points for future years align well with the regression line, indicating confidence in the downward trajectory.

Visual Insights:

- **Scatterplot & Line Graphs:** Confirm the long-term declining trend, despite short-term variability.
- **Confidence intervals** highlight the model's uncertainty range, emphasizing caution in interpreting specific values.
- **Bar chart:** Ranks countries with the highest average gender wage gaps. South Korea tops the list, while Greece ranks lowest among the top 10.

Summary: The data supports a long-term decrease in the gender wage gap, with forecasts reinforcing this downward trend through 2030. However, variability and country-specific disparities persist, highlighting the need for continued efforts in promoting wage equity.

2. Income Inequality Forecasting (Multiple Linear Regression)

Findings:

- Historical inequality data from 2010 to 2021 shows high short-term fluctuations but no strong upward or downward trend.
- Forecasts for 2026 to 2030 suggest slight and stable increases in income inequality levels.
- A separate analysis predicts a consistent yearly decrease of ~0.13%, hinting at modest improvement.

Visual Insights:

- **Line graph** shows a relatively flat yet slightly rising forecast.
- **Bar chart (yearly values)** displays minimal variation, reinforcing the forecast of stable inequality levels.
- **Year-over-year change bar chart** (negative values) suggests a gradual reduction in inequality.

Summary: Income inequality appears relatively stable over the historical period, with forecasts predicting either slight increases or gradual declines, depending on model interpretation. Overall, the change is subtle and points to a stubborn persistence of income disparity.

3. Global Income Inequality Modeling (Random Forest Regression)

Findings:

- The Random Forest model predicts the Gini Index using global socioeconomic indicators (income share, population, etc.).
- Prediction accuracy is moderate, with results generally following the trend of actual values but with visible variance.

Visual Insights:

- **Scatter plot:** Shows a positive correlation between predicted and actual values but with considerable error spread.
- **Histogram and KDE:** Reveal a bimodal distribution of inequality, with peaks around 0.2 and 0.55 Gini Index.
- **Box plot:** Indicates similar median inequality across income groups, with greater variability in lower-income countries.
- **Correlation heatmap:** Displays very weak correlations between Gini Index and individual predictors. The Top 10% income share has the strongest (but still weak) positive correlation.

Summary: The model demonstrates that global income inequality is complex and not easily explained by singular indicators. While there's general predictive alignment, the spread in predictions suggests many underlying variables influence inequality—likely beyond the scope of the selected features.

4. Gender Inequality Index (Linear Regression)

Findings:

- Based from the calculated forecast using the OBS value it was revealed that gender Inequality is decreasing average by 0.33% or 0.96% in 5 years in accordance to the models forecast revealing slow and unacceptable decrease in inequality.

Visual Insights:

- **Histogram:** A bimodal distribution of residuals is generally not good for a statistical model.

Summary: Based on the linear regression analysis, the Gender Inequality Index in the observed data is projected to decrease slowly, averaging 0.33% annually or 0.96% over five years, which is deemed an unacceptable rate of reduction. Furthermore, the bimodal distribution of residuals in the model's histogram indicates a potential issue with the model's fit, suggesting that the errors are not randomly distributed and there might be unexplained patterns affecting the accuracy of the forecast.

5. Labor Force (Pipeline Random Decision Forest)

Findings:

- **Significant Gender Disparity:** There is a substantial difference in average labor force participation between genders, with males (2688.32) showing a considerably higher average compared to females (1856.42).

- **Overall Labor Force Size:** The total average labor force participation, combining both genders, is 4385.24.

Visual Insights:

- **Boxplot:** visually compare the distribution of labor force participation across different gender categories.

Summary: The Random Decision Forest analysis of the labor force reveals a significant gender disparity, with males exhibiting a considerably higher average participation (2688.32) than females (1856.42), resulting in a total average participation of 4385.24. A box plot visualization is used to compare the distribution of labor force participation across these gender categories.

Conclusion

This study applied various machine learning models—including linear regression, multiple regression, and random forest—to analyze and forecast gender and social disparities across multiple global indicators. The findings present a complex picture, showing both progress and persistent inequalities.

The Gender Pay Gap analysis reveals a promising long-term decline, with projections indicating continued reduction through 2030. However, short-term fluctuations and significant country-level variations highlight that pay equity is still an ongoing challenge. Similarly, the Gender Inequality Index forecast shows a marginal annual decrease of just 0.33%, suggesting that efforts to advance gender equality are moving

too slowly. Moreover, the model's residual distribution points to potential limitations in capturing the full scope of gender-related disparities.

The Income Inequality Forecast indicates that disparities remain relatively stable, with only slight improvements predicted. Despite some models showing minor annual reductions, the overall trend remains flat, reflecting the entrenched nature of income inequality. The Global Income Inequality analysis using random forest regression highlights weak correlations between the Gini Index and traditional socioeconomic factors, further emphasizing the complexity of global income inequality and its dependence on a broader set of variables.

The Labor Force Participation study reveals a significant gender gap in economic activity, with males consistently participating at higher rates than females. This disparity contributes to broader social and economic inequalities. Despite the modeling approach, the gap remains wide and systemic, pointing to deep-rooted structural issues.

While machine learning models offer valuable insights and predictive capabilities, the results underscore that gender and income disparities are deeply ingrained and resistant to change. These issues require ongoing policy intervention, structural reforms, and cultural shifts. Technology, while helpful in identifying patterns and making forecasts, cannot alone resolve these disparities but can play a crucial role in informing and driving efforts toward a more equitable future.

Codes

Gender Wage Gap

```
# -*- coding: utf-8 -*-

"""VoxAequalisGenderWage.ipynb


Automatically generated by Colab.


Original file is located at


https://colab.research.google.com/drive/1mfWZgXCIIZngwFnw7xpNrISuQ8f69GHP

"""


import pandas as pd

import matplotlib.pyplot as plt

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error, r2_score

import seaborn as sns


gendergap_df = pd.read_csv('gendergapinaverage new.csv')


gendergap_clean = gendergap_df[['Year', 'Gender wage gap %']].dropna()

gendergap_clean = gendergap_clean[gendergap_clean['Gender wage gap %'] > 0]


X_gender = gendergap_clean[['Year']]

y_gender = gendergap_clean['Gender wage gap %']


model_gender = LinearRegression()
```

```
model_gender.fit(X_gender, y_gender)


y_pred_gender = model_gender.predict(X_gender)
mse_gender = mean_squared_error(y_gender, y_pred_gender)
r2_gender = r2_score(y_gender, y_pred_gender)


print("Gender Wage Gap Model:")
print("MSE:", mse_gender)
print("R² Score:", r2_gender)


future_years = pd.DataFrame({'Year': range(2025, 2031)})
future_predictions = model_gender.predict(future_years)
print("\nPredicted Wage Gap (2025-2030):")
print(pd.DataFrame({'Year': future_years['Year'], 'Predicted Gap %':
future_predictions.round(2)}))


plt.figure(figsize=(10, 5))
plt.plot(X_gender, y_gender, label="Actual")
plt.plot(X_gender, y_pred_gender, color='red', label="Predicted")
plt.scatter(future_years, future_predictions, color='green',
label="Future Prediction")
plt.xlabel("Year")
plt.ylabel("Gender Wage Gap (%)")
plt.title("Gender Wage Gap Trend and Forecast")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()


plt.figure(figsize=(10, 6))
```

```
sns.regplot(x='Year', y='Gender wage gap %', data=gendergap_clean,
ci=95, scatter_kws={'alpha':0.6})

plt.title("Gender Wage Gap Trend with Confidence Interval")

plt.grid(True)

plt.tight_layout()

plt.show()


country_avg_gap = gendergap_df[['country', 'Gender wage gap
%']].dropna().groupby('country').mean()

top_countries = country_avg_gap.sort_values('Gender wage gap %',
ascending=False).head(10)


plt.figure(figsize=(10, 6))

sns.barplot(x=top_countries['Gender wage gap %'],
y=top_countries.index, palette="magma")

plt.xlabel("Average Gender Wage Gap (%)")

plt.title("Top 10 Countries by Avg Gender Wage Gap")

plt.tight_layout()

plt.show()


plt.figure(figsize=(10, 6))

sns.lineplot(data=gendergap_clean, x="Year", y="Gender wage gap %")

plt.title("Gender Wage Gap Over Time")

plt.grid(True)

plt.tight_layout()

plt.show()
```

Inequality In Income

```
# -*- coding: utf-8 -*-

"""VoxAequalisChange.ipynb


Automatically generated by Colab.


Original file is located at

https://colab.research.google.com/drive/1f0bEYhZDcz758_AbpnksF9JF4Sky0cy5
"""


import pandas as pd

import matplotlib.pyplot as plt

from sklearn.linear_model import LinearRegression


df = pd.read_csv("Inequality in Income.csv")


df.columns = [col.strip().replace('Inequality in income (', '').replace(')', '') for col in df.columns]


year_columns = [col for col in df.columns if col.isdigit()]

df_years = df[year_columns]


df_long = df_years.melt(var_name='Year', value_name='Inequality_Income')

df_long['Year'] = df_long['Year'].astype(int)


X = df_long[['Year']]

y = df_long['Inequality_Income']
```

```
model = LinearRegression()

model.fit(X, y)


future_years = pd.DataFrame({'Year': list(range(2026, 2031))})

future_years['LR_Prediction'] = model.predict(future_years[['Year']])

future_years['Yearly_Change'] = future_years['LR_Prediction'].diff()


combined_years = list(df_long['Year']) + list(future_years['Year'])

combined_values = list(df_long['Inequality_Income']) +
list(future_years['LR_Prediction'])


plt.figure(figsize=(10, 6))

plt.plot(df_long['Year'], df_long['Inequality_Income'],
label='Historical', marker='o', color='blue')

plt.plot(future_years['Year'], future_years['LR_Prediction'],
label='Forecast (Linear Regression)', linestyle='--', marker='x',
color='green')

plt.title('Historical vs Forecasted Income Inequality')

plt.xlabel('Year')

plt.ylabel('Inequality (%)')

plt.legend()

plt.grid(True)

plt.tight_layout()

plt.show()


plt.figure(figsize=(10, 6))

plt.bar(future_years['Year'], future_years['LR_Prediction'],
color='seagreen', edgecolor='black')

plt.title('Forecasted Income Inequality (2026-2030)')

plt.xlabel('Year')
```

```
plt.ylabel('Predicted Inequality (%)')
plt.grid(axis='y')
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 6))
plt.bar(future_years['Year'], future_years['Yearly_Change'],
        color=['grey' if x == 0 else 'green' if x < 0 else 'red' for
x in future_years['Yearly_Change']],
        edgecolor='black')
plt.axhline(0, color='black', linewidth=0.8, linestyle='dashed')
plt.title('Yearly Change in Predicted Inequality')
plt.xlabel('Year')
plt.ylabel('Δ Inequality (%)')
plt.grid(axis='y')
plt.tight_layout()
plt.show()

print(future_years[['Year', 'LR_Prediction', 'Yearly_Change']])
```

Global Inequality

```
# -*- coding: utf-8 -*-

"""VoxAequalisGlobalInequality

Automatically generated by Colab.

Original file is located at

https://colab.research.google.com/drive/1YftrfvOcTPSBWxkSFNwtpDPJuxrkfc0W

"""


import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error, r2_score


income_df = pd.read_csv('global_income_inequality.csv')

income_df_clean = income_df.dropna(subset=['Gini Index',
'Population', 'Average Income (USD)',

                                           'Top 10% Income Share
(%)', 'Bottom 10% Income Share (%)'])


X_income = income_df_clean[['Population', 'Average Income (USD)',
'Top 10% Income Share (%)',

                            'Bottom 10% Income Share (%)']]

y_income = income_df_clean['Gini Index']


X_train, X_test, y_train, y_test = train_test_split(X_income,
y_income, test_size=0.2, random_state=42)
```

```
model_income = RandomForestRegressor(random_state=42)
model_income.fit(X_train, y_train)

y_pred_income = model_income.predict(X_test)
mse_income = mean_squared_error(y_test, y_pred_income)
r2_income = r2_score(y_test, y_pred_income)

print("Mean Squared Error:", mse_income)
print("R² Score:", r2_income)

feature_importance = pd.Series(model_income.feature_importances_,
index=X_income.columns).sort_values(ascending=False)
print("\nFeature Importances:\n", feature_importance)

plt.figure(figsize=(8, 5))
sns.scatterplot(x=y_test, y=y_pred_income)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
'r--')
plt.xlabel('Actual Gini Index')
plt.ylabel('Predicted Gini Index')
plt.title('Actual vs Predicted Gini Index')
plt.grid(True)
plt.tight_layout()
plt.show()

plt.figure(figsize=(8, 5))
sns.histplot(income_df_clean['Gini Index'], kde=True, bins=30,
color='purple')
plt.title("Distribution of Gini Index")
```

```
plt.xlabel("Gini Index")
plt.ylabel("Frequency")
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(x='Income Group', y='Gini Index', data=income_df_clean)
plt.title("Gini Index by Income Group")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

plt.figure(figsize=(8, 6))
sns.heatmap(income_df_clean[['Population', 'Average Income (USD)',
                             'Top 10% Income Share (%)', 'Bottom 10%
Income Share (%)',
                             'Gini Index']].corr(), annot=True,
cmap='coolwarm')
plt.title("Feature Correlation with Gini Index")
plt.tight_layout()
plt.show()
```

4. Gender Inequality Index

```
 # -*- coding: utf-8 -*-

"""GenderInequalityIndex


Automatically generated by Colab.


Original file is located at

https://colab.research.google.com/drive/18_fLG_LMmKe99J4wonEt9si9p0ENGj8d

"""


import pandas as pd

import matplotlib.pyplot as plt

from sklearn.linear_model import LinearRegression


genderIndex_df = pd.read_csv('FAO_AS_4537.csv')


genderIndex_df = genderIndex_df.dropna(subset=['TIME_PERIOD', 'OBS_VALUE'])


genderIndex_df['TIME_PERIOD'] = genderIndex_df['TIME_PERIOD'].astype(int)


X = genderIndex_df[['TIME_PERIOD']]

y = genderIndex_df['OBS_VALUE']


model = LinearRegression()

model.fit(X, y)
```

```
y_pred = model.predict(X)


year_2026 = pd.DataFrame({'TIME_PERIOD': [2026]})

forecast_2026 = model.predict(year_2026)[0]

print(f"🔮 Forecasted Gender Inequality (OBS_VALUE) for 2026:
{forecast_2026:.2f}")


plt.figure(figsize=(10, 6))

plt.scatter(X, y, label='Actual', color='blue')

plt.plot(X, y_pred, label='Fitted Line', color='red')

plt.scatter(2026, forecast_2026, color='green', label='Forecast
2026', zorder=5)

plt.xlabel('Time Period (Year)')

plt.ylabel('Inequality Income (OBS_VALUE)')

plt.title('Linear Regression: Inequality Income vs Time')

plt.legend()

plt.grid(True)

plt.show()


residuals = y - y_pred

plt.figure(figsize=(10, 6))

plt.scatter(X, residuals, color='purple')

plt.axhline(0, color='black', linestyle='--')

plt.title('Residuals Plot')

plt.xlabel('Time Period (Year)')

plt.ylabel('Residuals')

plt.grid(True)

plt.show()


plt.figure(figsize=(10, 6))
```

```
sns.histplot(residuals, kde=True, color='teal')

plt.title('Distribution of Residuals')

plt.xlabel('Residuals')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()


print(f"Model Coefficients:\nSlope: {model.coef_[0]}, Intercept:
{model.intercept_}")


future_years = pd.DataFrame({'TIME_PERIOD': [2025, 2026, 2027,
2028]})

future_predictions = model.predict(future_years)


forecast_df = future_years.copy()

forecast_df['Forecast_OBS_VALUE'] = future_predictions


print("\n📊 Forecasts for Gender Inequality (OBS_VALUE):\n")

print(forecast_df.to_string(index=False))


plt.figure(figsize=(10, 6))

# Historical

plt.scatter(X, y, label='Actual', color='blue')

plt.plot(X, y_pred, label='Fitted Line', color='red')

# Forecast

plt.scatter(forecast_df['TIME_PERIOD'],
forecast_df['Forecast_OBS_VALUE'], color='green', label='Forecast
(2025-2028)', zorder=5)

for i, txt in enumerate(forecast_df['Forecast_OBS_VALUE']):

    plt.annotate(f"{txt:.2f}", (forecast_df['TIME_PERIOD'][i],
forecast_df['Forecast_OBS_VALUE'][i]),
```

```
                textcoords="offset points", xytext=(0,10),
ha='center', fontsize=9)


plt.xlabel('Time Period (Year)')

plt.ylabel('Inequality Income (OBS_VALUE)')

plt.title('Gender Inequality Forecast: 2025-2028')

plt.legend()

plt.grid(True)

plt.tight_layout()

plt.show()
```

5. Labor Force

```
# -*- coding: utf-8 -*-
"""laborForce.ipynb

Automatically generated by Colab.

Original file is located at

https://colab.research.google.com/drive/1Q0JimmJJdmkDZTNNaH8k-ueoe7neLJ59
"""

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np
from statsmodels.tsa.arima.model import ARIMA

df = pd.read_csv("labourForce.csv")

plt.figure(figsize=(8, 5))
sns.histplot(df['obs_value'], kde=True, bins=30)
plt.title("Distribution of Labour Force %")
```

```
plt.xlabel("Labour Force %")
plt.ylabel("Count")
plt.show()

plt.figure(figsize=(8, 5))
sns.boxplot(x='gender', y='obs_value', data=df)
plt.title("Labour Force % by Gender")
plt.show()

sample_countries = df['country'].unique()[:20]
for country in sample_countries:
    subset = df[df['country'] == country]
    sns.lineplot(data=subset, x='time', y='obs_value', hue='gender')
    plt.title(f"Labour Force Trend in {country}")
    plt.xticks(rotation=45)
    plt.show()

X = df[['country', 'gender', 'time']]
y = df['obs_value']

print(df['time'].isna().sum())

print(f"X shape: {X.shape}")
print(f"y shape: {y.shape}")

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

categorical_features = ['country', 'gender']
```

```
numeric_features = ['time']


preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore'),
categorical_features),
        ('num', 'passthrough', numeric_features)
    ]
)


model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(n_estimators=100,
random_state=42))
])


model.fit(X_train, y_train)


y_pred = model.predict(X_test)


mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)


print(f"MAE: {mae:.2f}")
print(f"RMSE: {rmse:.2f}")
print(f"R² Score: {r2:.2f}")


plt.scatter(y_test, y_pred, alpha=0.5)
plt.xlabel("Actual Labour Force %")
```

```
plt.ylabel("Predicted Labour Force %")
plt.title("Actual vs Predicted Labour Force")
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)],
color='red')
plt.show()

ohe = model.named_steps['preprocessor'].named_transformers_['cat']
encoded_features = ohe.get_feature_names_out(['country', 'gender'])
all_features = np.append(encoded_features, ['time'])

importances = model.named_steps['regressor'].feature_importances_
indices = np.argsort(importances)[-15:]  # show top 15

plt.figure(figsize=(10, 6))
plt.barh(range(len(indices)), importances[indices], align='center')
plt.yticks(range(len(indices)), [all_features[i] for i in indices])
plt.title("Top Feature Importances")
plt.xlabel("Importance Score")
plt.tight_layout()
plt.show()

sorted_idx = y_test.argsort()
plt.figure(figsize=(10, 5))
plt.plot(y_test.iloc[sorted_idx].values, label='Actual')
plt.plot(y_pred[sorted_idx], label='Predicted', alpha=0.7)
plt.title("Actual vs Predicted Labour Force %")
plt.xlabel("Sample Index")
plt.ylabel("Labour Force %")
plt.legend()
```

```
plt.tight_layout()

plt.show()


subset = df[(df['country'] == 'Philippines') & (df['gender'] ==
'Female')]

subset['time'] = pd.to_datetime(subset['time'],
errors='coerce').dt.to_period('Y').dt.to_timestamp()

subset = subset.sort_values('time').set_index('time')


model_arima = ARIMA(subset['obs_value'], order=(1,1,1))

model_fit = model_arima.fit()


forecast = model_fit.forecast(steps=5)

print("Forecasted Labour Force %:", forecast)


forecast_steps = 5

forecast = model_fit.forecast(steps=forecast_steps)


last_date = subset.index[-1]

forecast_index = pd.date_range(start=last_date +
pd.DateOffset(years=1), periods=forecast_steps, freq='Y')


plt.figure(figsize=(10, 5))

plt.plot(subset['obs_value'], label='Actual')

plt.plot(forecast_index, forecast, label='Forecast', color='orange',
marker='o')

plt.title("Labour Force Forecast with ARIMA")

plt.xlabel("Year")

plt.ylabel("Labour Force %")

plt.legend()
```

```
plt.grid(True)
plt.tight_layout()
plt.show()
```

References

https://www.kaggle.com/datasets/iamsouravbanerjee/gender-inequality-index-dataset

https://www.kaggle.com/datasets/iamsouravbanerjee/inequality-in-income-across-the-globe

https://www.kaggle.com/datasets/willianoliveiragibin/gender-economic-inequality

https://www.kaggle.com/datasets/georgehanyfouad/global-income-inequality

https://data360.worldbank.org/en/indicator/FAO_AS_4537?view=map

https://data360.worldbank.org/en/indicator/WB_SSGD_UNEMPLOYMENT_RATE?sex=M&age=_T&urbanisation=_T

https://data360.worldbank.org/en/dataset/WB_GS

https://rshiny.ilo.org/dataexplorer28/?lang=en&segment=indicator&id=LAP_2FTM_NOC_RT_A

https://rshiny.ilo.org/dataexplorer04/?lang=en&segment=indicator&id=EAP_TEAP_SEX_AGE_MTS_NB_A