Christopher Mulya | Natalie Coutinho | Pranav Kumar Mahesh

Exploring the Interplay of Macroeconomic and Social Variables on Unemployment in Indonesia

Introduction:

The purpose of this report is to present a detailed analysis of socio-economic indicators pertaining to Indonesia, spanning the years 1980-2022. Through rigorous examination, the report aims to unveil significant trends, patterns, and interrelationships within the dataset. The insights derived from this analysis serve to inform policy formulation, decision-making processes, and future research endeavors, contributing to the holistic understanding of Indonesia's development trajectory.

Dataset Overview:

The dataset at hand comprises 43 observations spanning the years 1980 to 2022 and encompasses 14 variables, each representing a unique socio-economic indicator related to Indonesia. The variables include crucial aspects such as compensation of employees, education expenditure, population growth, net migration, GDP growth, lending interest rates, inflation rates, life expectancy, consumption expenditure, foreign direct investment, tertiary school enrollment, tax revenue, short-term debt, and unemployment rates. These indicators offer a comprehensive snapshot of Indonesia's economic and social landscape over the study period.

```
Variable types and non-null counts:
<class 'pandas.core.frame.DataFrame'>
Index: 43 entries, 1980 to 2022
Data columns (total 14 columns):
 #   Column                                                          Non-Null Count  Dtype
---  ------                                                          --------------  -----
 0   Compensation of employees (% of expense)                        37 non-null     float64
 1   Adjusted savings: education expenditure (current US$)           42 non-null     float64
 2   Population growth (annual %)                                    43 non-null     float64
 3   Net migration                                                   43 non-null     float64
 4   GDP growth (annual %)                                           43 non-null     float64
 5   Lending interest rate (%)                                       37 non-null     float64
 6   Inflation, consumer prices (annual %)                           43 non-null     float64
 7   Life expectancy at birth, total (years)                         42 non-null     float64
 8   Households and NPISHs final consumption expenditure (% of GDP)  43 non-null     float64
 9   Foreign direct investment, net inflows (% of GDP)               43 non-null     float64
 10  School enrollment, tertiary (% gross)                           43 non-null     float64
 11  Tax revenue (% of GDP)                                          38 non-null     float64
 12  Short-term debt (% of exports of goods, services and primary income)  42 non-null     float64
 13  Unemployment, total (% of total labor force) (national estimate)      38 non-null     float64
dtypes: float64(14)
memory usage: 5.0+ KB
None
```

Figure 1. Variable Information

The above figure provides a more detailed breakdown of the dataset's structure, showcasing the data types and non-null counts for each variable. It becomes apparent that some variables have missing values in certain years, potentially requiring further attention during the analysis.

Summary Statistics:

The summary statistics (Figure 2) offer a nuanced view of the dataset's central tendencies and variabilities. For instance, exploring "Compensation of employees (% of expense)," we observe a mean value of approximately 18.78%, indicating the average share of expenses allocated to employee compensation. The standard deviation of 5.57% suggests a moderate dispersion around the mean. The minimum and maximum values (7.43% and 27.78%, respectively) provide insights into the range of variation over the years.

```
Summary Statistics:
```

Indicator Name	count	mean	std	min	25%	50%	75%	max
Compensation of employees (% of expense)	37.0	1.878173e+01	5.568715e+00	7.426203e+00	1.470848e+01	1.637655e+01	2.399085e+01	2.777834e+01
Adjusted savings: education expenditure (current US$)	42.0	1.065173e+10	1.280355e+10	5.921390e+08	8.011616e+08	2.503211e+09	2.330928e+10	3.771320e+10
Population growth (annual %)	43.0	1.497622e+00	4.441259e-01	6.365551e-01	1.253628e+00	1.410099e+00	1.760234e+00	2.379330e+00
Net migration	43.0	-2.584514e+04	4.797391e+04	-1.084920e+05	-5.202400e+04	-2.880400e+04	7.319000e+03	4.737500e+04
GDP growth (annual %)	43.0	4.956022e+00	3.468557e+00	-1.312673e+01	4.740124e+00	5.308595e+00	6.420715e+00	1.000000e+01
Lending interest rate (%)	37.0	1.674676e+01	5.616760e+00	8.520000e+00	1.240333e+01	1.597917e+01	2.082500e+01	3.215417e+01
Inflation, consumer prices (annual %)	43.0	8.871831e+00	8.672454e+00	1.560130e+00	4.929370e+00	6.757317e+00	9.949279e+00	5.845104e+01
Life expectancy at birth, total (years)	42.0	6.579200e+01	3.377954e+00	5.875400e+01	6.327000e+01	6.632700e+01	6.863275e+01	7.051800e+01
Households and NPISHs final consumption expenditure (% of GDP)	43.0	6.136298e+01	4.316667e+00	5.303797e+01	5.787863e+01	6.157706e+01	6.351623e+01	7.394380e+01
Foreign direct investment, net inflows (% of GDP)	43.0	1.072251e+00	1.229591e+00	-2.757440e+00	3.414733e-01	1.268301e+00	2.022334e+00	2.916115e+00
School enrollment, tertiary (% gross)	43.0	1.814500e+01	1.184930e+01	3.389360e+00	8.518350e+00	1.480152e+01	2.751289e+01	4.263317e+01
Tax revenue (% of GDP)	38.0	1.410358e+01	3.621455e+00	8.310105e+00	1.108287e+01	1.429561e+01	1.623984e+01	2.194643e+01
Short-term debt (% of exports of goods, services and primary income)	42.0	2.781233e+01	1.128223e+01	1.109974e+01	1.988145e+01	2.431931e+01	3.446384e+01	5.557260e+01
Unemployment, total (% of total labor force) (national estimate)	38.0	4.637553e+00	1.762756e+00	1.660000e+00	3.003000e+00	4.376500e+00	6.081000e+00	8.060000e+00

Figure 2. Summary Statistics

Similarly, for the "Adjusted savings: education expenditure (current US$)" indicator, the mean value of approximately 10 billion US dollars signifies the average annual adjusted savings in education expenditure. The standard deviation of 12.80 billion US dollars implies a considerable variability in annual savings. These summary statistics provide a foundation for understanding the magnitude and variability of each indicator.

By systematically examining these statistics for all 14 variables, we can develop a nuanced understanding of the dataset's characteristics. This detailed exploration lays the groundwork for subsequent statistical analyses and informs the selection of appropriate methods to answer scientific questions related to Indonesia's socio-economic trends.

Null Values Investigation:

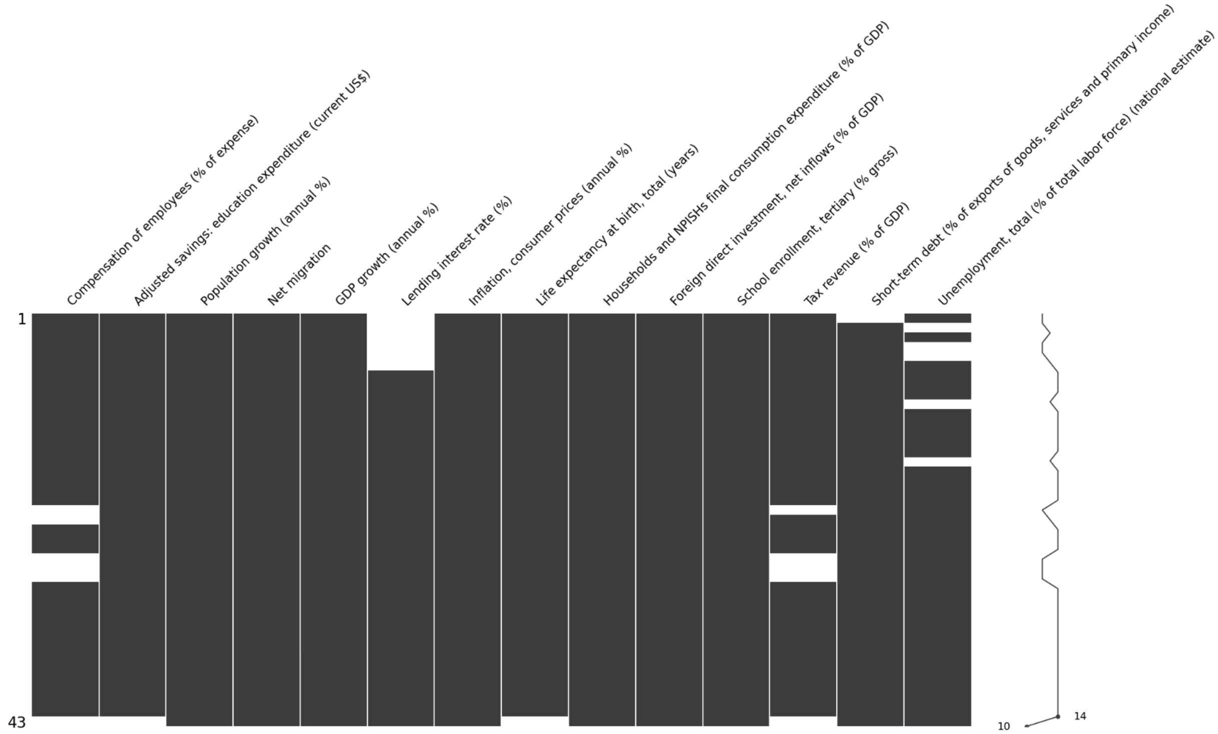


Figure 3. Missing Value Insight

The matrix, shown in Figure 3, depicts the missing values contained in our dataset. Notably, key variables such as compensation of employees, lending interest rate, life expectancy, tax revenue, and unemployment rate display notable instances of missing information.

To address these missing values, we imputed them using a straightforward imputation method, employing forward fill (and backward fill for NA values that start from 1980, since there were no previous values to forward fill from). This approach aimed to fill the gaps in our dataset, ensuring a more comprehensive and reliable foundation for subsequent analyses. The selection of forward fill was based on the structured nature of our data, organized by year, where variables tend to exhibit trends over time. By adopting the forward fill (and backward fill) imputation technique, we aimed to maintain the integrity of the data while mitigating the impact of missing values on the overall analytical outcomes.

Univariate analysis (distribution and spread for every continuous attribute):

Compensation of employees (% of expense)

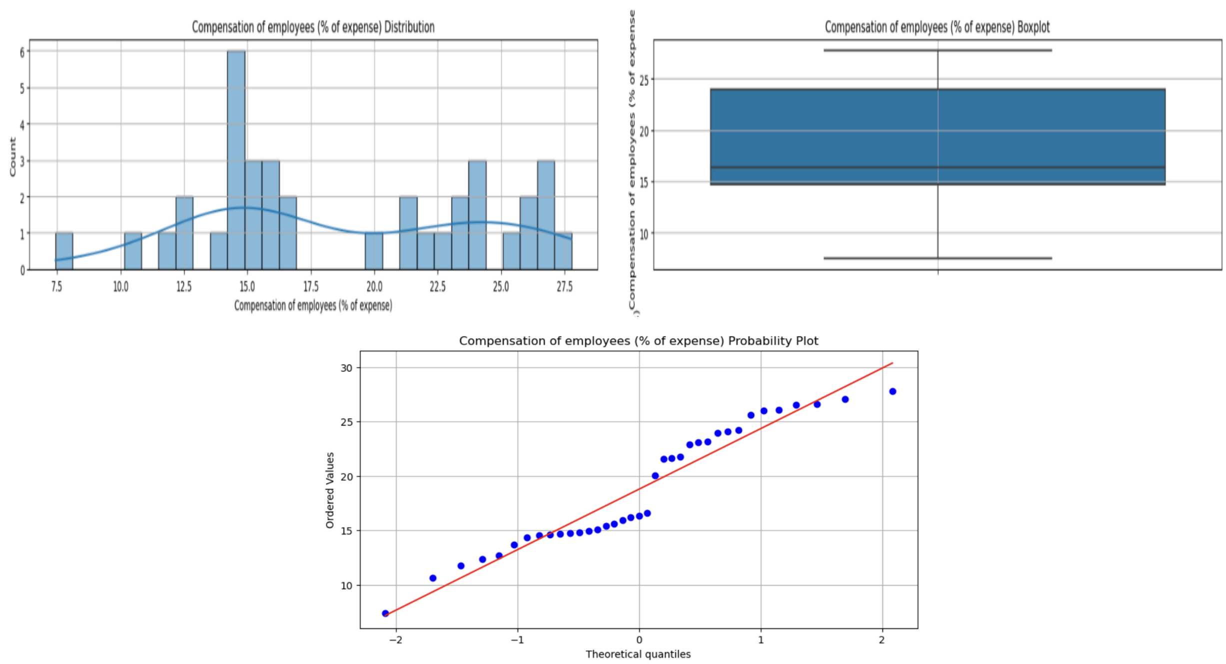
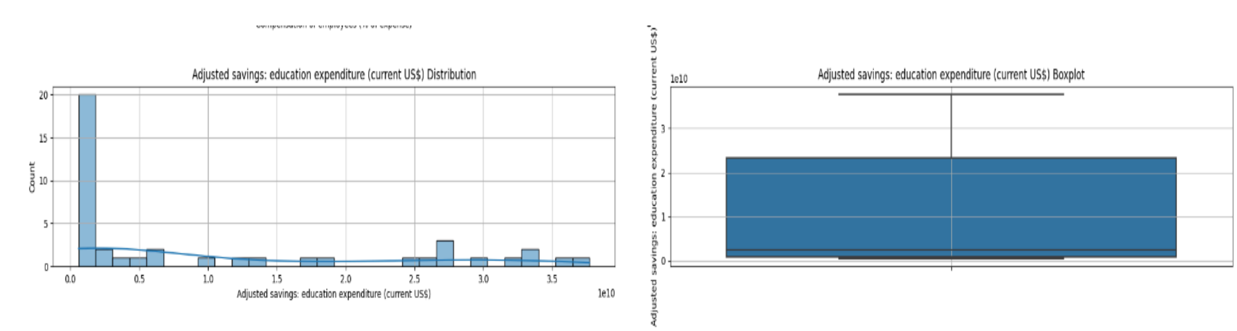


Figure 4. Distributional Plot, Boxplot and Q-Q Plot of Employee Compensation

Examining the histogram and box plot, it clearly exhibits a positive skewness with a computed value of 0.0780, falling within the range of -0.5 to +0.5. This observation suggests a marginal departure from perfect symmetry, indicating a distribution that is approximately symmetric. The q-q plot further reinforces the data's deviation from a normal distribution. Notably, there are no discernible outliers within the dataset, contributing to its overall stability and consistency.

Adjusted savings: education expenditure (current US$)



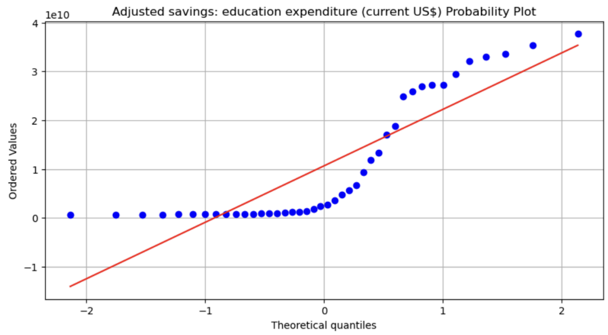


Figure 5. Distributional Plot, Boxplot and Q-Q Plot of Education Expenditure

Examining the histogram and box plot, it clearly exhibits a positive skewness, characterized by a skewness value of 0.9171, surpassing the typical range of +0.5 to +1. This deviation from the symmetric distribution suggests a pronounced rightward tail, indicating a substantial concentration of values towards the lower end. While the assertion of a normal distribution does not hold true based on the q-q plot, it is pertinent to note that the prevailing skewness suggests a more likely right-skewed orientation. Notably, an absence of outliers is discernible within the dataset, contributing to the coherence of the observed distribution.

Population growth (annual %)

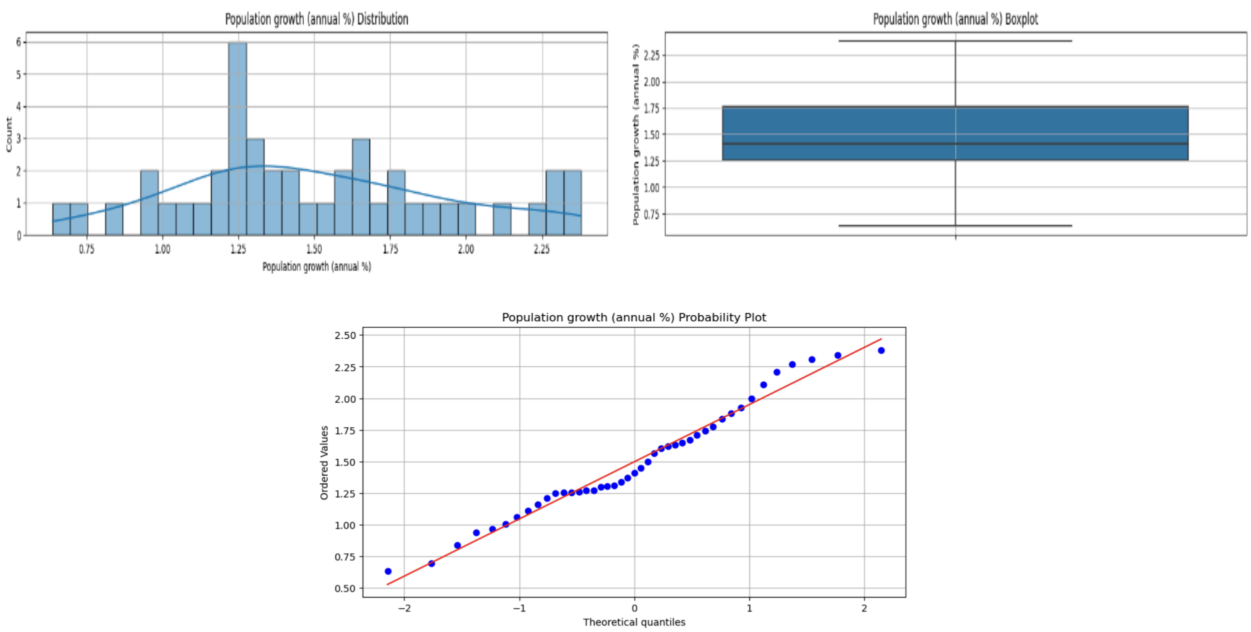


Figure 6. Distributional Plot, Boxplot and Q-Q Plot of Population Growth

Examining the histogram and box plot, it clearly exhibits a positive skewness, as indicated by a skewness value of 0.2911, falling within the range of -0.5 to +0.5. This suggests a moderate rightward deviation from symmetry. While as seen in the q-q plot the data appears to align with a normal distribution, it is good to note that the right skewness does not appear to be overly pronounced, and the absence of outliers in the dataset reinforces the reliability of this observation.

Net migration

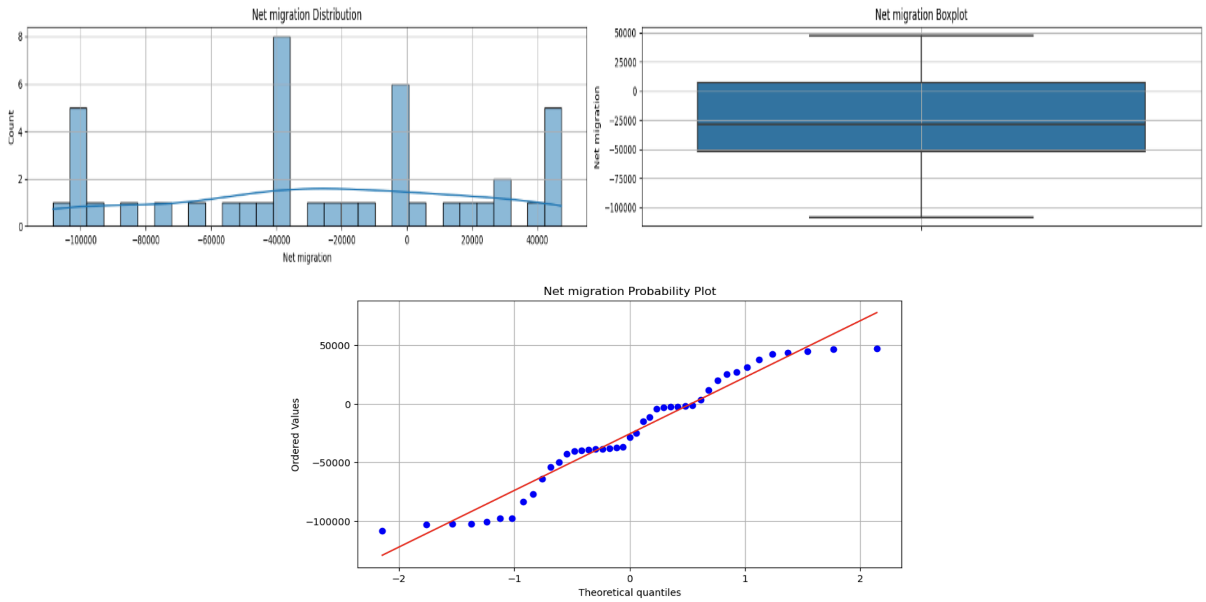


Figure 7. Distributional Plot, Boxplot and Q-Q Plot of Net Migration

Examining the histogram and box plot, it clearly exhibits the data is negatively skewed with a value of -0.1817, falling within the range of -0.5 to +0.5. This indicates a moderate leftward skewness, implying that the majority of observations tend to cluster towards the higher end of the distribution. However, it is crucial to note that based on the q-q plot the data deviates from a normal distribution. The absence of outliers in the dataset is noteworthy, as it suggests a relatively uniform distribution without extreme values.

GDP growth (annual %)

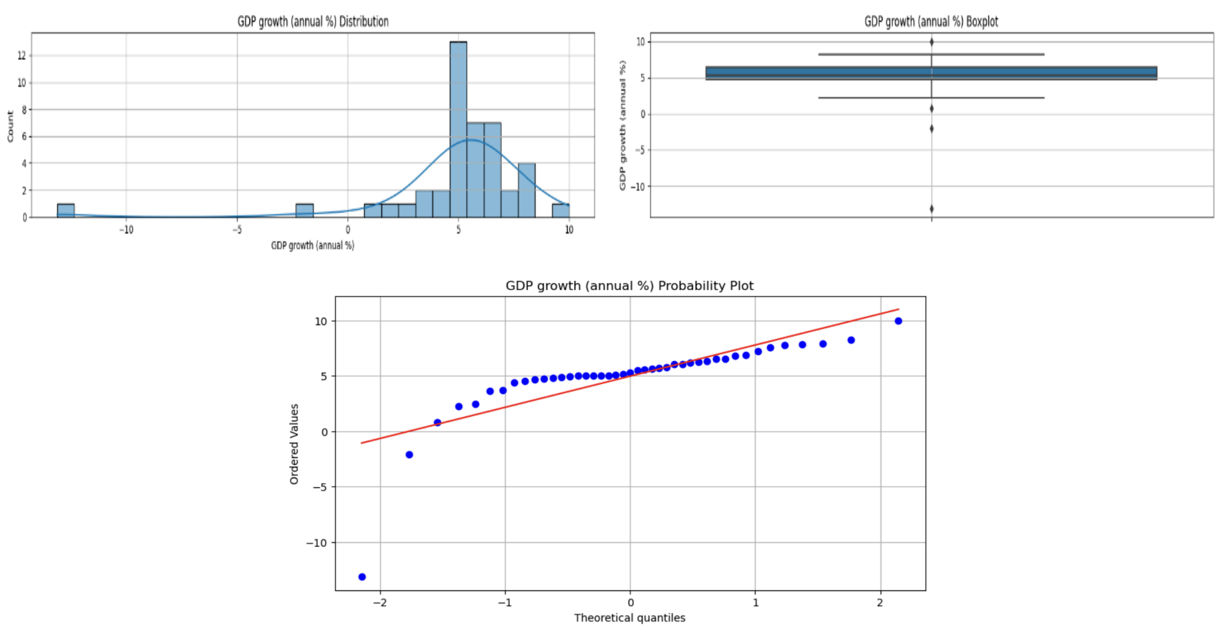


Figure 8. Distributional Plot, Boxplot and Q-Q Plot of GDP Growth

The data exhibits a notable negative skewness as we see while referring to the histogram and boxplot, quantified by a skewness value of -3.6367. This value, surpassing the conventional range of -1 to +1, underscores the pronounced asymmetry looking at the q-q plot as it does not lie on the fitted line of the distribution with a discernible leftward skewed. This observation denotes a substantial departure from the characteristics of a normal distribution. Surprisingly, there is evidence of outliers in the data when we look at the boxplot.

Foreign direct investment, net inflows (% of GDP)

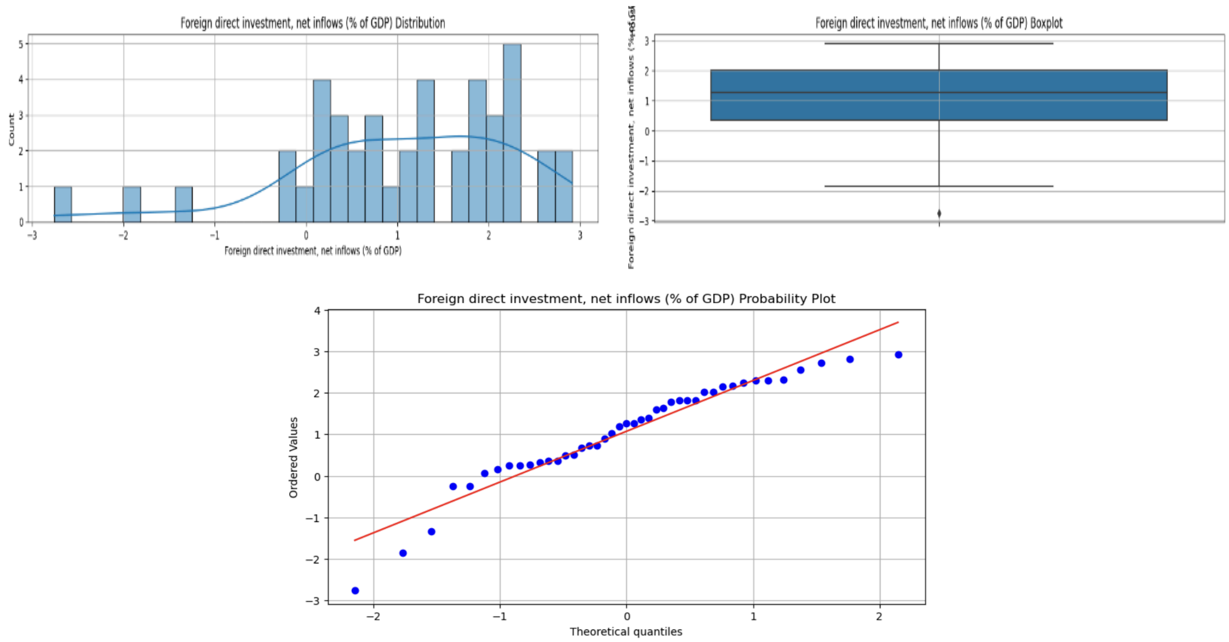


Figure 9. Distributional Plot, Boxplot and Q-Q Plot of Foreign Direct Investment

Analyzing the histogram and boxplot exhibits a notable negative skewness, as evidenced by a skewness coefficient of -0.9687, falling within the range of -1 to -0.5. This observation indicates a pronounced leftward tail in the distribution, signifying an asymmetry of information. It is pertinent to highlight that the dataset does not conform to a normal distribution, thereby deviating from the expected symmetrical pattern with the absence of outliers.

School enrollment, tertiary (% gross)

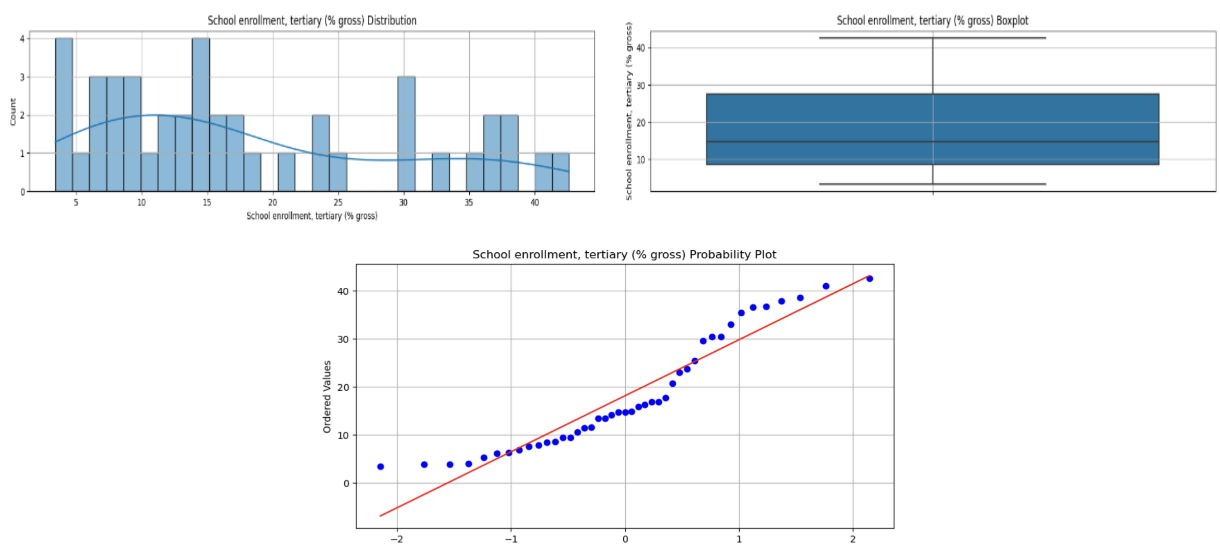


Figure 10. Distributional Plot, Boxplot and Q-Q Plot of Tertiary School Enrollment

On carefully analyzing the data the histogram shows a positive skewness with a value of 0.6718, falling within the range of +0.5 to +1. This observation signifies a departure from a symmetrical distribution in the q-q plot, leaning towards a right-skewed pattern. Notably, the data does not adhere to the characteristics of a normal distribution. Furthermore, upon scrutiny, no outliers are discernible within the data as shown in the boxplot.

Unemployment, total (% of total labor force) (national estimate)

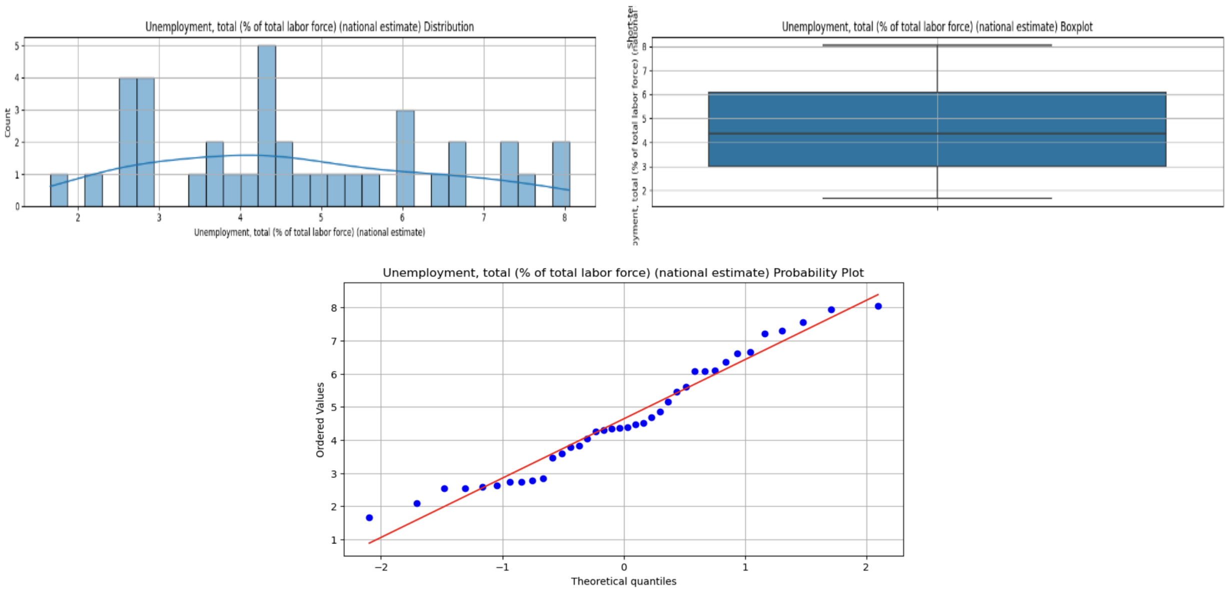


Figure 11. Distributional Plot, Boxplot and Q-Q Plot of Unemployment Rate

The histogram and box plot both reveal a positive skewness, as indicated by a skewness value of 0.3160. This falls within the range of -0.5 to +0.5, suggesting a moderate rightward deviation from a perfectly symmetrical distribution. Despite this moderate skewness, the q-q plot affirms the data's conformity to a normal distribution, thereby displaying a tendency towards a right-skewed pattern. Furthermore, a thorough examination of the dataset reveals an absence of apparent outliers.

Exploration of correlation among variables:

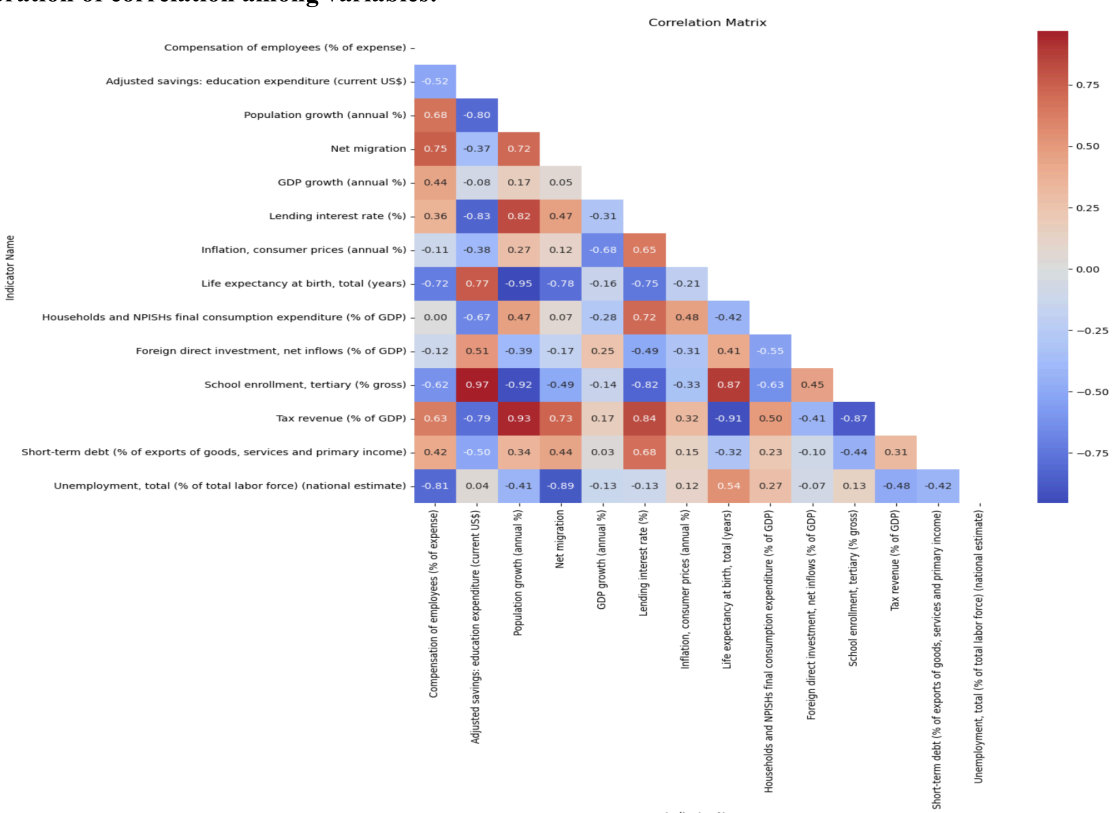


Figure 12. Correlations Among Socio-economic Indicators in Indonesia (1980-2022)

The correlation matrix (Figure 12) sheds light on the interrelationships between the selected socio-economic indicators in Indonesia from 1980 to 2022. Correlation coefficients range from -1 to 1, where -1 indicates a perfect negative correlation, 1 signifies a perfect positive correlation, and 0 denotes no correlation.

Bivariate analysis (relationship between different variables, correlations):

The Bivariate Analysis section explores the relationships between various socio-economic indicators and the unemployment rate in Indonesia.

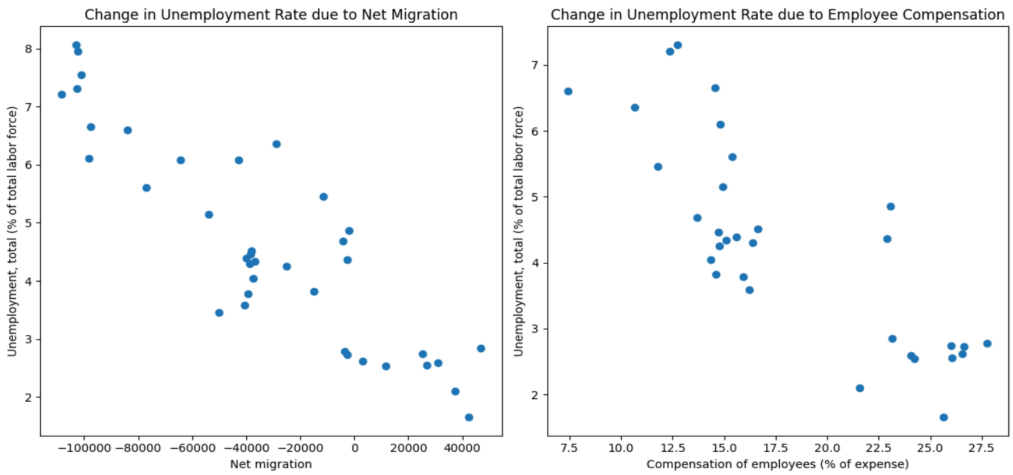


Figure 13. Bivariate plots between net migration, compensation of employees, and unemployment rate in Indonesia

Figure 13 presents two bivariate plots, each shedding light on distinct relationships. The first plot illustrates the relationship between Net Migration and the Unemployment Rate, revealing a robust negative correlation of -0.89 (refer to Figure 12 for the correlation matrix). This indicates that as net migration increases, the unemployment rate tends to decrease, suggesting a potential connection between migration patterns and employment trends. In parallel, the second plot delves into the correlation between Compensation of Employees and the Unemployment Rate, uncovering a significant negative correlation of -0.81 (see Figure 12). This implies that higher compensation allocated to employees corresponds to lower unemployment rates. These findings provide valuable insights into the dynamics between migration, labor costs, and unemployment, underscoring the interconnected nature of socio-economic factors.

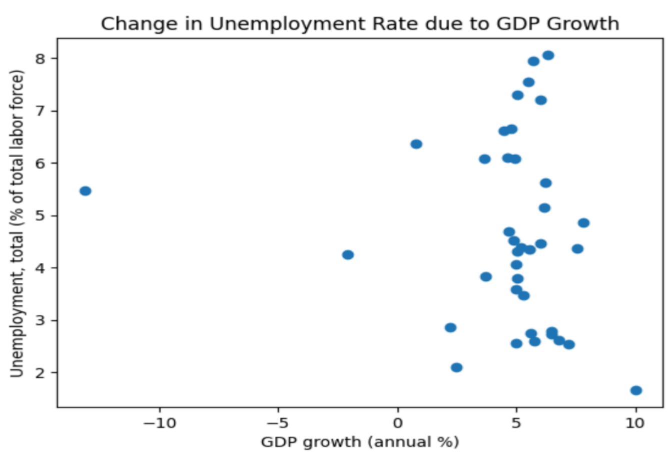


Figure 14. Bivariate plot between GDP Growth and Unemployment Rate in Indonesia

Figure 14 displays a bivariate plot illustrating the relationship between GDP growth and the unemployment rate in Indonesia. The plot reveals a very weak negative correlation, with a value of -0.13 (see Figure 12), between the two variables. Most data points are dispersed vertically, suggesting that changes in GDP growth have limited predictive power over fluctuations in the unemployment rate. This weak correlation implies that other factors, not solely captured by GDP growth, significantly contribute to the dynamics of unemployment in the Indonesian context. In-depth analysis and consideration of these additional factors are essential for a comprehensive understanding of the unemployment landscape in the country.

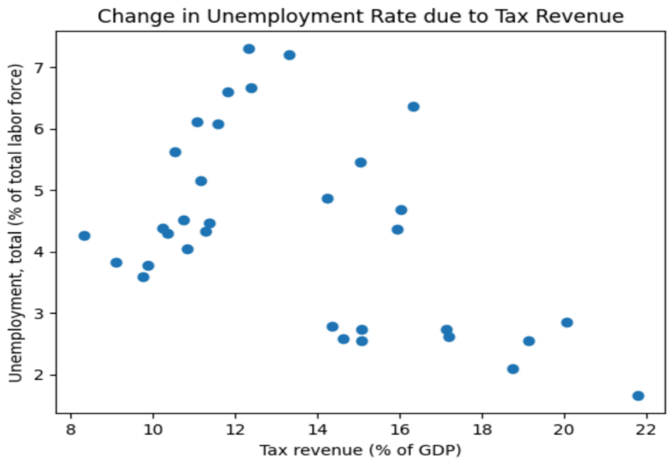


Figure 15. Bivariate plot between Tax Revenue and Unemployment Rate in Indonesia

Figure 15 illustrates a bivariate plot examining the relationship between tax revenue and the unemployment rate in Indonesia. The plot uncovers a moderate negative correlation, with a value of -0.48 (see Figure 12 for values) between these two variables. The observed correlation suggests that as tax revenue fluctuates, there is a tendency for the unemployment rate to exhibit corresponding variations. This finding implies a potential connection between fiscal policies impacting tax revenue and labor market conditions.

Scientific Questions:

Impact of Labor Costs on Unemployment

- Question: What is the impact of variations in compensation of employees (% of expense) on the unemployment rate in Indonesia, and how can this relationship inform strategies for mitigating unemployment risks?
- Rationale: Exploring the correlation between compensation of employees and unemployment rates helps unravel potential links between labor costs and workforce stability, offering practical implications for economic policies and business strategies.
- Statistical Analysis Technique: Simple Linear Regression Analysis. This technique will help quantify the impact of variations in compensation of employees (% of expense) on the unemployment rate. This technique allows for modeling the relationship between the dependent variable (unemployment rate) and independent variable (labor costs). By analyzing the coefficients and significance levels, we can identify the strength and significance of the relationship, providing insights for formulating strategies to mitigate unemployment risks.

Policy Effectiveness in Mitigating Unemployment Risks

- Question: What policy measures are most effective in reducing unemployment rates? Are there differences in the effectiveness of monetary policy, fiscal policy, and labor market reforms in tackling unemployment?
- Rationale: Understanding the comparative effectiveness of policies can guide policymakers in formulating targeted and efficient strategies to address unemployment challenges.
- Statistical Analysis Technique: Multiple Regression Analysis. This technique quantifies the effectiveness variables included in the analysis, specifically focusing on monetary policy (lending interest rates, short term debt, GDP growth, and inflation rates), fiscal policy (tax revenue, educational expenditure, and foreign direct investments), and labor market reforms (net migration, compensation of employees, school enrollment, life expectancy, population growth) impacting unemployment. This approach allows for understanding the unique contribution of each policy variable and their combined effects, guiding policymakers in formulating targeted and efficient strategies to address unemployment challenges.

Temporal Patterns and Seasonality

- Question: Is there evidence of autocorrelation in unemployment rates, potentially linked to specific economic cycles or time periods?
- Rationale: Understanding the presence of autocorrelation in unemployment rates provides insights into the persistence of trends over time and the potential influence of past unemployment levels on current rates. This knowledge is crucial for policymakers and analysts seeking to develop accurate forecasts and strategies to address unemployment challenges.
- Statistical Analysis Technique: Time Series Analysis. This method is employed to investigate autocorrelation in unemployment rates over time. Specifically, the autocorrelation function (ACF) and Autoregressive Integrated Moving Average (ARIMA) models are utilized to assess the dependency of current unemployment rates on past values. By examining historical data, this approach enables the identification of patterns and trends in unemployment dynamics, providing valuable insights for policymaking and economic analysis.

Results and Analysis:

Impact of Compensation of Employees on Unemployment Rate

Our investigation aimed to discern the impact of variations in compensation of employees (% of expense) on the unemployment rate in Indonesia, offering insights into strategies for mitigating unemployment risks. Employing a simple linear regression analysis, we found compelling evidence of a significant negative relationship between these two variables. The regression model yielded an R-squared value of 0.649, indicating that approximately 64.9% of the variance in the unemployment rate can be explained by changes in the compensation of employees. The regression coefficient for the compensation of employees was estimated to be -0.2529 ($p < 0.001$), suggesting that for every one-unit increase in the percentage of expenses allocated to employee compensation, the unemployment rate decreases by approximately 0.2529 percentage points. This negative coefficient aligns with economic intuition, implying that higher labor costs may deter employers from hiring, consequently reducing the unemployment rate. Our findings are consistent with theoretical expectations and empirical evidence in labor economics, supporting the notion that labor market dynamics are influenced by the cost of employment.

The regression equation derived from the simple linear regression (SLR) analysis is:

Unemployment rate = 8.9575 - 0.2529 * Compensation of Employees (% of Expense)

This equation models the relationship between the unemployment rate and the percentage of expenses allocated to employee compensation.

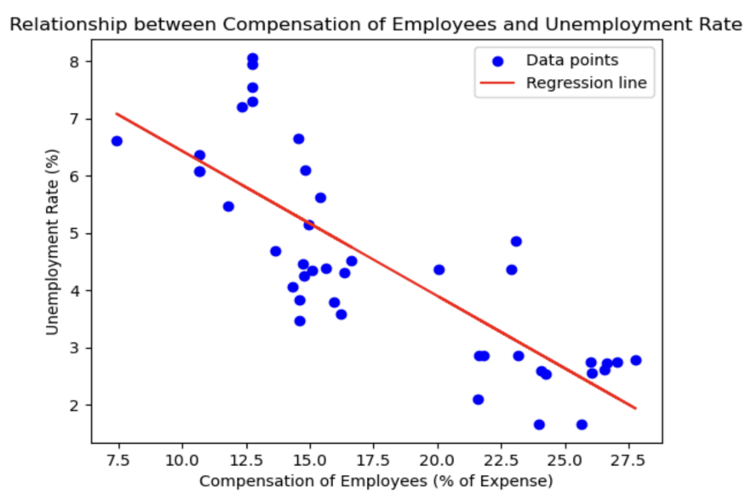


Figure 16. Scatter Plot of Unemployment Rate against Compensation of Employees (% of Expense)

Assumptions underlying the regression analysis were thoroughly examined to ensure the validity of our results. Diagnostic tests, including the Omnibus, Jarque-Bera, and Durbin-Watson tests, were conducted to assess key assumptions such as linearity, independence of errors, homoscedasticity, and normality of residuals. The Omnibus test evaluates the skewness and kurtosis of the residuals, yielding a non-significant p-value of 0.263, indicating that the skewness and kurtosis of the residuals are within acceptable limits, supporting the assumption of normality. Similarly, the Jarque-Bera test, assessing the normality of residuals based on skewness and kurtosis, also yielded a non-significant p-value of 0.295, further confirming the normality assumption.

Moreover, the Durbin-Watson test was conducted to assess the independence of errors, revealing a result of 0.508, suggesting the presence of positive autocorrelation in the residuals. However, despite this violation of the independence of errors assumption, our analysis proceeded as the other diagnostic tests did not reveal significant violations. Visually observing the scatter plot (Figure 16), we noted a negative relationship between the compensation of employees and the unemployment rate, with the regression line indicating a clear downward trend. This plot serves as a robust visual representation, reinforcing the statistical inference drawn from the regression analysis. Together, these diagnostic tests and visual representations support the validity of our regression analysis and the reliability of the observed relationship between labor costs and unemployment rates.

Impact of different policy measures in reducing unemployment

This investigation delves into the effectiveness of various policy measures in mitigating unemployment in Indonesia, focusing on monetary policies, fiscal policies, and labor reforms. Comparing the impact of these strategies can guide policy makers in formulating targeted and efficient strategies to address unemployment challenges.

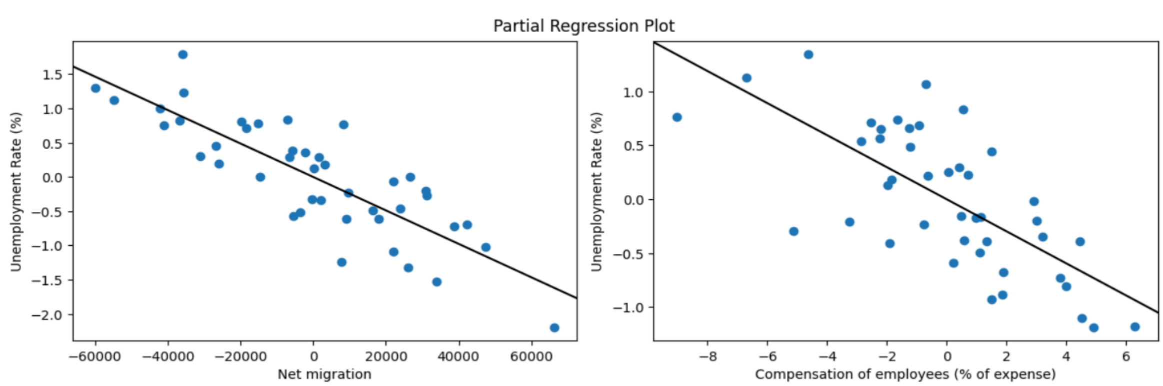
By conducting multiple regression and employing stepwise selection to identify significant factors, the study highlights strong evidence indicating that net migration, compensation of employees, education expenditure, and foreign direct investment are particularly influential in reducing the unemployment rate, accounting for 93.1% of its variance as indicated by an R-squared value of 0.931.

Notably, net migration and compensation of employees represent labor market reforms, while education expenditure and foreign direct investment signify fiscal policies. Interestingly, factors related to monetary policy show no significant impact on unemployment.

The regression equation derived from the multiple regression analysis:

Unemployment rate = 6.8938 - 0.1485 * Compensation of Employees + 0.1659 * Foreign Direct Investment - 0.00002 * Net Migration - 5.4e-11 * Education Expenditure

Diagnostic tests were conducted to provide information about the overall goodness of fit and certain assumptions about the multiple regression model. The Durbin-Watson test yielded a value of 1.823, suggesting a slight positive correlation in the residuals. While values close to 2 indicate no correlation, deviations from this suggest autocorrelation. Despite this minor correlation, the model appears to adequately capture most of the data’s information. The Jarque-Bera test produced a value of 0.586 which is relatively high, indicating that the residuals may be normally distributed, supported by a skewness value close to zero, signifying approximately symmetrical distribution.



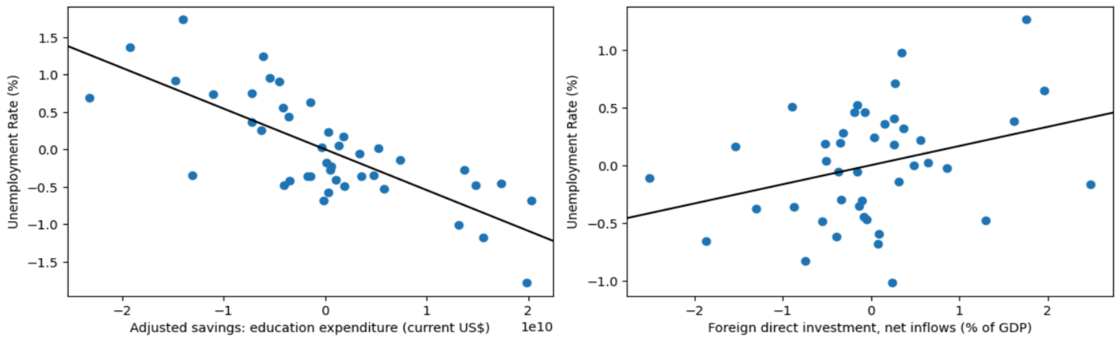


Figure 17. Partial Regression Plot on Unemployment Rate

Based on the above scatter plots (Figure 17), it's evident that net migration, compensation of employees, and education expenditure exhibit a negative relationship with the unemployment rate. This implies that an increase in skilled workers migrating to Indonesia, higher compensation for employees relative to expenses, and greater investment in education are associated with lower unemployment rates.

Conversely, foreign direct investment shows a positive relationship with the unemployment rate. While higher levels of foreign direct investment may stimulate economic activity, they can also lead to job displacement or competition for local labor, consequently contributing to higher unemployment rates.

Time Series Analysis

Utilizing the Augmented Dickey Fuller test (ADF Test) to assess the stationarity of the time series data on unemployment rates in Indonesia. Our results indicated that we have non-significant p-value i.e., 0.58 exceeds 0.05, indicating non-stationarity of data, we plan to apply differencing techniques to transform the data into a stationary form. Here, we have taken alpha=0.05.

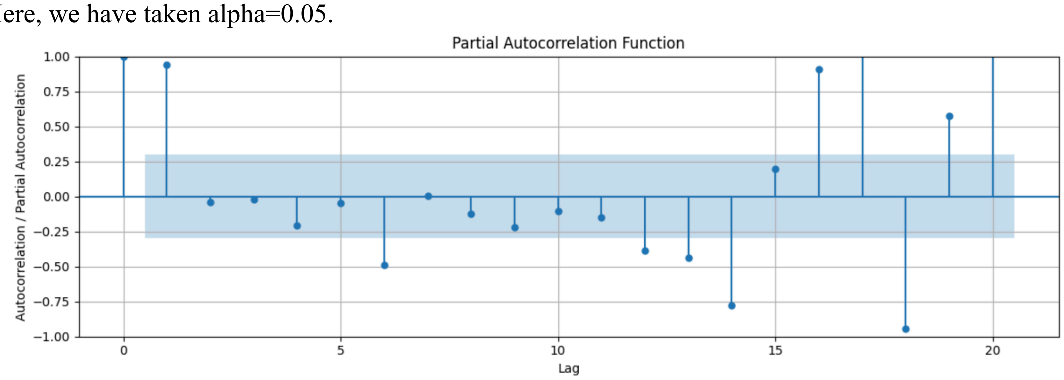


Figure 18. Partial Autocorrelation Plot

Auto-Regressive parameter (p):

- The Auto-Regressive parameter in an ARIMA model is denoted as 'p'. It is determined by examining the Partial Autocorrelation Function (PACF) plot (Figure 18).
- This rapid decay indicates a more direct relationship between each observation and its predecessors in the time series. This rapid decay indicates a more direct relationship between each observation and its predecessors in the time series.
- 'p' is defined by the significant lag before which the PACF plot cuts off to 1. This indicates the lag beyond which the partial autocorrelation becomes statistically insignificant.
- Notably, the PACF decays at a faster rate than the Autocorrelation Function (ACF), which makes it particularly useful in identifying the order of the Auto-Regressive (AR) component in the ARIMA model.

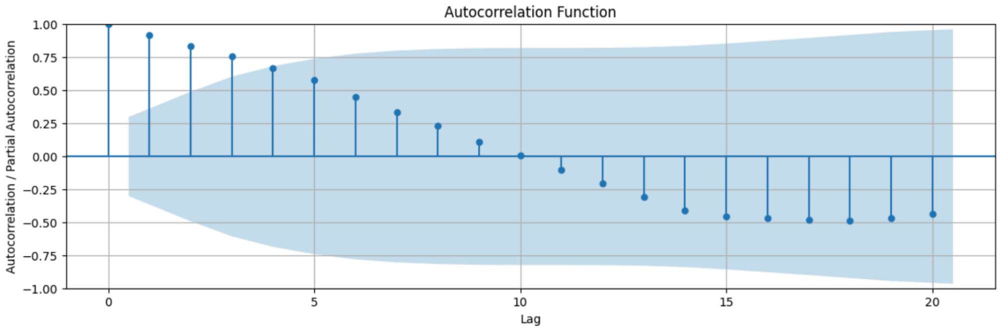


Figure 19. Autocorrelation Plot

Moving-Average parameter (q):

- The Moving-Average parameter in an ARIMA model is denoted as ‘q’. It is determined by examining the Autocorrelation Function (ACF) plot (Figure 19).
- ‘q’ is defined by the significant lag before which the ACF plot cuts off to 1. This indicates the lag beyond which the autocorrelation becomes statistically insignificant.
- The ACF plot steadily decreases without any large jumps; it indicates autocorrelation.

The observed pattern of the ACF plot, characterized by a steady decrease without any significant spikes or jumps, holds critical implications for our analysis of the unemployment rate in Indonesia. Autocorrelation, as depicted by this trend, suggests a strong relationship between consecutive observations in the dataset. This finding is particularly relevant in the context of analyzing unemployment dynamics over time, where understanding the sequential nature of data is essential for identifying underlying trends and patterns.

In the context of Indonesia’s unemployment rate, the consistent decline in autocorrelation indicates that the influence of past observations on future ones gradually diminishes over time. This insight underscores the evolving nature of unemployment trends and the importance of considering historical data when assessing the current state of employment in the country. By recognizing these temporal dependencies, policymakers and analysts can better anticipate shifts in unemployment rates and formulate targeted interventions to address labor market challenges.

The absence of large jumps or spikes in the ACF plot, on the other hand, suggests a relative stability in the autocorrelation structure of the unemployment rate data. This stability implies a certain degree of predictability or regularity in the underlying process governing unemployment dynamics, offering valuable insights for policymakers tasked with implementing long-term strategies to promote job creation and economic stability in Indonesia.

Furthermore, our analysis employed an Autoregressive Integrated Moving Average model to examine the trends of unemployment rates in Indonesia over time. The model specification chosen was ARIMA(1, 1, 1), indicating the presence of a first-order autoregressive term, a first-order differencing term, and a first-order moving average term. Assumptions underlying the ARIMA were thoroughly examined to ensure the validity of our results.

The estimated coefficient for the autoregressive term (AR) is 0.8242. This suggests a positive relationship between the current unemployment rate and its lagged value, indicating persistence in unemployment levels over time. The coefficient is statistically significant at the 0.05 level, implying that past unemployment rates have a significant impact on the current rate.

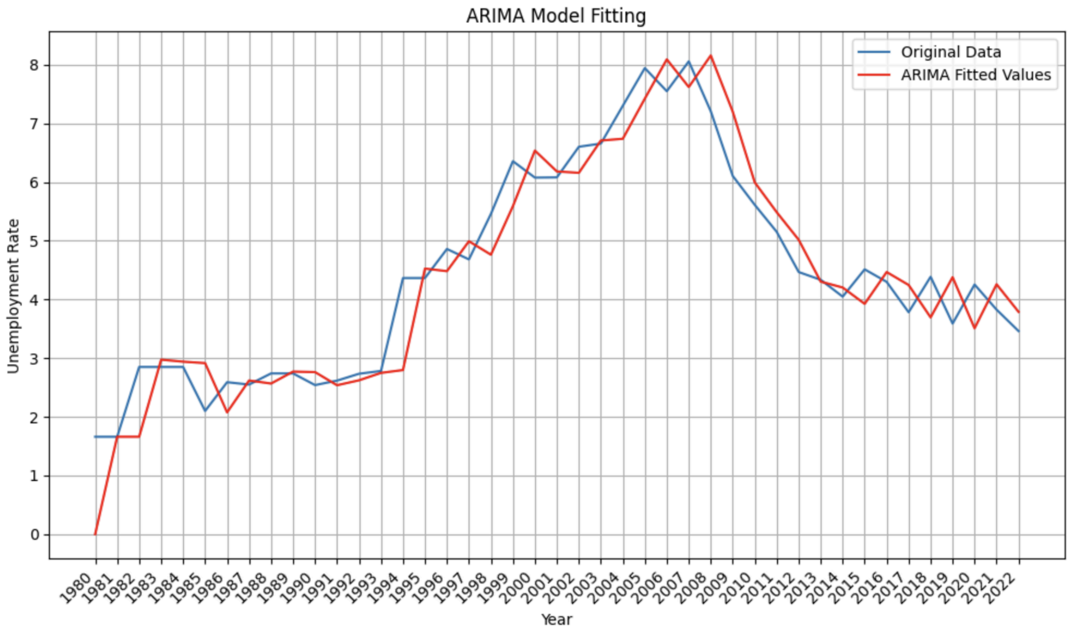


Figure 20. ARIMA Model - Observed Vs Fitted Values

The estimated coefficient for the moving average term (MA) is -0.7299. This negative coefficient suggests that past forecast errors tend to have an opposite effect on the current unemployment rate. However, the coefficient is not statistically significant at the 0.05 level. This means that we cannot reliably estimate the impact of past forecast errors on the current unemployment rate.

The estimated residual variance is 0.3090, indicating the variability in the unemployment rate that is not explained by the model. Despite this, the value is statistically significant at the 0.05 level, suggesting that the model captures a notable portion of the variability in the unemployment rate. Now this represents the magnitude of the random fluctuations or "random steps" that are seen as substantial fluctuations in the unemployment rate that are not explained by the model's autoregressive and moving average components.

Diagnostic tests such as the Ljung-Box test for autocorrelation and the Jarque-Bera test for normality are conducted to assess the adequacy of the model. The Ljung-Box test statistic (Q) of 0.74 and Prob (Q) 0.39 along with the Jarque-Bera test statistic of 1.36 and Prob (JB) 0.51 suggest that the residuals are not significantly autocorrelated and follow a normal distribution, respectively. The Heteroskedasticity (H) statistic is reported as 0.88, with a corresponding probability value (Prob(H)) of 0.82 exceeding the significance level of 0.05, stating that there is no significant evidence of heteroskedasticity in the residuals. As mentioned above the data seems slightly right-skewed with a positive skewness value of 0.43. And finally, the kurtosis value (reported as 3.18) is slightly higher than the typical excess kurtosis value of 3 for a normal distribution (heavier on the tails). Additionally, the plot (Figure 20) displays the observed values of the unemployment rate alongside the corresponding predictions generated by the ARIMA model. We see that the model is well-fitted as the predicted values closely follow the observed values. It indicates that the ARIMA model accurately captures the underlying patterns and dynamics in the data.

Conclusion:

In summary, our comprehensive analysis of socio-economic indicators in Indonesia revealed significant insights into the dynamics of unemployment and its relationship with various factors. Specifically, we identified a notable negative relationship between compensation of employees and the unemployment rate, indicating higher labor costs may lead to reduced unemployment rates. However, it is important to acknowledge the violation of the independence of errors assumption detected by the Durbin-Watson test and the correlational nature of the observed relationship, which warrants further exploration and investigation to better understand the complexities underlying unemployment trends in Indonesia.

Additionally, our findings highlighted the effectiveness of labor market reforms, such as those focusing on net migration and compensation, as well as fiscal policies like education expenditure and foreign direct investment, in mitigating unemployment. These measures not only enhance the skill base of the workforce but also ensure fair compensation, thus fostering employment opportunities. Although these models capture a significant portion of unemployment variance, they overlook crucial factors like technological advancements and broader government policies, which could also influence unemployment rates significantly.

Moreover, our time series analysis underscored the persistence of unemployment patterns over time accounting for significant autocorrelation dynamics. It’s worth noting that initially, the data was found to be non-stationary according to the Augmented Dickey-Fuller (ADF) test. Therefore, we applied first-order differencing to achieve stationarity before proceeding with the analysis.

While our data passed all the diagnostic tests required to run regression analysis and time series analysis there were few exceptions that reinforced the validity of our results. One notable challenge we encountered was missing data. Addressing missing data is crucial for ensuring the accuracy and reliability of our analysis. While we employed techniques such as backward and forward fill to handle missing values, the presence of incomplete data sets can introduce biases and limitations in our findings.

Recommendations:

1. **Focus on Labor Market Reforms:** The study highlights the effectiveness of labor market reforms, like incentivizing skilled migration and ensuring fair compensation, in reducing unemployment. Prioritizing policies to attract skilled workers and improve local employee welfare can significantly impact unemployment rates in Indonesia.
2. **Increase Investment in Education:** Increasing investment in education is a long-term strategy to reduce unemployment, involving initiatives to enhance access to quality education, vocational training, and aligning curriculum with labor market demands.
3. **Careful Monitoring of Foreign Direct Investment:** While foreign direct investment (FDI) can boost economic growth, the study indicates a potential positive link with unemployment due to job displacement or labor competition. Policy makers must closely monitor FDI's impact on the labor market and enact measures to optimize its benefits while mitigating negative effects on employment.
4. **Reassess Monetary Policies:** Since factors related to monetary policies showed no significant impact on unemployment in the analysis, policy makers may need to reassess the effectiveness of existing monetary policies in addressing unemployment challenges. This could involve exploring alternative strategies or adjusting existing policies to better support job creation and economic growth.

Appendix:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

file_path = 'indonesia_data.csv'

# Read the CSV file with a tab as the delimiter
df = pd.read_csv(file_path, sep=',', skiprows=3)

# Specify columns to drop
columns_to_drop = ["Country Name", "Country Code", "Indicator Code", "Unnamed: 67"]

# Drop the specified columns
df = df.drop(columns=columns_to_drop, errors='ignore')

df.head()
```

```
# Specify a list of indicator names you want to select
indicator_names_to_select = ['Compensation of employees (% of expense)',
                             'Adjusted savings: education expenditure (current
US$)',
                             'Population growth (annual %)',
                             'Net migration',
                             'GDP growth (annual %)',
                             'Lending interest rate (%)',
                             'Inflation, consumer prices (annual %)',
                             'Life expectancy at birth, total (years)',
                             'Households and NPISHs final consumption expenditure
(% of GDP)',
                             'Foreign direct investment, net inflows (% of GDP)',
                             'School enrollment, tertiary (% gross)',
                             'Tax revenue (% of GDP)',
                             'Short-term debt (% of exports of goods, services and
primary income)',
                             'Unemployment, total (% of total labor force)
(national estimate)'
                            ]

# Select rows with the specified indicator names
updated_df = df[df['Indicator
Name'].isin(indicator_names_to_select)].set_index('Indicator Name')

# Transpose the df to have the predictors as column names
transposed_df = updated_df.transpose()

#Rearrange columns
transposed_df = transposed_df[['Compensation of employees (% of expense)',
```

```
                                'Adjusted savings: education expenditure (current
US$)',
                              'Population growth (annual %)',
                              'Net migration',
                              'GDP growth (annual %)',
                              'Lending interest rate (%)',
                              'Inflation, consumer prices (annual %)',
                              'Life expectancy at birth, total (years)',
                              'Households and NPISHs final consumption expenditure
(% of GDP)',
                              'Foreign direct investment, net inflows (% of GDP)',
                              'School enrollment, tertiary (% gross)',
                              'Tax revenue (% of GDP)',
                              'Short-term debt (% of exports of goods, services and
primary income)',
                              'Unemployment, total (% of total labor force)
(national estimate)']]

# Filter based on years
final_df = transposed_df.loc['1980':'2022']
final_df
```

```
#Number of NaN values in each column
final_df.isna().sum()

import missingno as msno

# Visualize missing values using a matrix
msno.matrix(final_df)
plt.show()

# Display basic information about the dataset
print("Number of variables (columns):", final_df.shape[1])
print("Number of observations (rows):", final_df.shape[0])

# Display variable types and non-null counts
print("\nVariable types and non-null counts:")
print(final_df.info())
```

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=[24, 48])

# List of columns you want to visualize
columns_to_visualize = ['Compensation of employees (% of expense)',
                         'Adjusted savings: education expenditure (current US$)',
                         'Population growth (annual %)',
```

```
                         'Net migration',
                         'GDP growth (annual %)',
                         'Lending interest rate (%)',
                         'Inflation, consumer prices (annual %)',
                         'Life expectancy at birth, total (years)',
                         'Households and NPISHs final consumption expenditure (% of
GDP)',
                         'Foreign direct investment, net inflows (% of GDP)',
                         'School enrollment, tertiary (% gross)',
                         'Tax revenue (% of GDP)',
                         'Short-term debt (% of exports of goods, services and
primary income)',
                         'Unemployment, total (% of total labor force) (national
estimate)'
                         ]

# Iterate through the columns
for i, column in enumerate(columns_to_visualize, start=1):
    # Distributional plot
    plt.subplot(len(columns_to_visualize), 2, i * 2 - 1)
    sns.histplot(final_df[column], kde=True, bins=30)
    plt.title(f"{column} Distribution")
    plt.grid(True)
    print(f'skew value of {column} is', final_df[column].skew())

    # Boxplot
    plt.subplot(len(columns_to_visualize), 2, i * 2)
    sns.boxplot(y=final_df[column])
    plt.title(f"{column} Boxplot")
    plt.grid(True)

# Adjust layout for better spacing
plt.tight_layout()

# Add spacing between subplots
plt.subplots_adjust(hspace=0.5)

# Show the plot
plt.show()
```

```
import scipy.stats as stats
from scipy.stats import shapiro

for column in columns_to_visualize:
    print(column)

    # Handling NA values by dropping them
    data = final_df[column].dropna()
```

```
    stat, p = shapiro(data)
    print(f'Statistics={stat:.3f}, p={p:.3f}')

    # Shapiro-Wilk test
    alpha = 0.05
    if p > alpha:
        print('Sample looks Gaussian (fail to reject H0)')
    else:
        print('Sample does not look Gaussian (reject H0)')

    # Probability plot
    plt.figure(figsize=[10, 5])

    stats.probplot(data, plot=plt, dist='norm', fit=True)
    plt.title(f"{column} Probability Plot")
    plt.grid(True)

    plt.show()
```

```
# Display summary statistics
print("\nSummary Statistics:")
final_df.describe().T

# Custom function to annotate heatmap while handling masked values
def annotate_heatmap(ax, fmt=".2f"):
    for i in range(correlation_matrix.shape[0]):
        for j in range(correlation_matrix.shape[1]):
            if not mask[i, j]:
                ax.text(j + 0.5, i + 0.5, format(correlation_matrix.values[i, j],
fmt),
                        ha="center", va="center", color="black", fontsize=10)

# Visualize correlation matrix
correlation_matrix = final_df.corr()

# Visualize half of the correlation matrix
mask = np.triu(np.ones_like(correlation_matrix, dtype=bool))

plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", mask=mask)
plt.title("Correlation Matrix")

# Call custom function to annotate heatmap
annotate_heatmap(plt.gca())

plt.show()
```

```
# old dataset for comparison
final_df

# Fill missing values using forward fill followed by backward fill
final_df_filled = final_df.fillna(method='ffill').fillna(method='bfill')

# Display the DataFrame after filling missing values
final_df_filled
```

```
import statsmodels.api as sm

# Extracting independent and dependent variables
X = final_df_filled['Compensation of employees (% of expense)']
y = final_df_filled['Unemployment, total (% of total labor force) (national
estimate)']

# Adding a constant term to the independent variable
X = sm.add_constant(X)

# Fitting the linear regression model
model = sm.OLS(y, X).fit()

# Printing the model summary
model.summary()
```

```
# Stepwise selection
def stepwise_selection(X, y, initial_list=[], threshold_in=0.05, threshold_out =
0.05, verbose=True):
    included = list(initial_list)
    while True:
        changed=False
        # forward step
        excluded = list(set(X.columns)-set(included))
        new_pval = pd.Series(index=excluded, dtype=float)
        for new_column in excluded:
            model = sm.OLS(y,
sm.add_constant(pd.DataFrame(X[included+[new_column]]))).fit()
            new_pval[new_column] = model.pvalues[new_column]
        best_pval = new_pval.min()
        if best_pval < threshold_in:
            best_feature = new_pval.idxmin()
            included.append(best_feature)
            changed=True
            if verbose:
                print('Add  {:30} with p-value {:.6}'.format(best_feature,
best_pval))

        # backward step
```

```
        model = sm.OLS(y, sm.add_constant(pd.DataFrame(X[included]))).fit()
        # use all coefs except intercept
        pvalues = model.pvalues.iloc[1:]
        worst_pval = pvalues.max()
        if worst_pval > threshold_out:
            changed=True
            worst_feature = pvalues.idxmax()
            included.remove(worst_feature)
            if verbose:
                print('Drop {:30} with p-value {:.6}'.format(worst_feature,
worst_pval))
        if not changed:
            break
    return included

import numpy as np
import statsmodels.api as sm

# Define predictors (independent variables) and the target variable
X = final_df_filled[['Compensation of employees (% of expense)',
          'Adjusted savings: education expenditure (current US$)',
          'Population growth (annual %)',
          'Net migration',
          'GDP growth (annual %)',
          'Lending interest rate (%)',
          'Inflation, consumer prices (annual %)',
          'Life expectancy at birth, total (years)',
          'Households and NPISHs final consumption expenditure (% of GDP)',
          'Foreign direct investment, net inflows (% of GDP)',
          'School enrollment, tertiary (% gross)',
          'Tax revenue (% of GDP)',
          'Short-term debt (% of exports of goods, services and primary income)']]

y = final_df_filled['Unemployment, total (% of total labor force) (national
estimate)']
result = stepwise_selection(X, y)

# Fit the final model
final_model = sm.OLS(y, sm.add_constant(X[result])).fit()
final_model.summary()
```

```
import matplotlib.pyplot as plt
import statsmodels.api as sm

# Visualize the time series data
plt.figure(figsize=(10, 6))
plt.plot(final_df_filled.index, final_df_filled['Unemployment, total (% of total
labor force) (national estimate)'], marker='o', linestyle='-')
```

```
plt.title('Yearly Unemployment Rate')
plt.xlabel('Year')
plt.ylabel('Unemployment Rate')

# Rotate x-axis labels diagonally
plt.xticks(rotation=45, ha='right')

plt.grid(True)
plt.subplots_adjust(bottom=0.2)
plt.tight_layout()
plt.show()

# Augmented Dickey-Fuller test for stationarity
result = sm.tsa.stattools.adfuller(final_df_filled['Unemployment, total (% of total
labor force) (national estimate)'])
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])

import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Create a figure with subplots
fig, axes = plt.subplots(2, 1, figsize=(12, 8))

# Plot ACF on the first subplot
plot_acf(final_df_filled['Unemployment, total (% of total labor force) (national
estimate)'], lags=20, ax=axes[0])
axes[0].set_title('Autocorrelation Function')

# Plot PACF on the second subplot
plot_pacf(final_df_filled['Unemployment, total (% of total labor force) (national
estimate)'], lags=20, ax=axes[1])
axes[1].set_title('Partial Autocorrelation Function')

# Add labels and grid
for ax in axes:
    ax.set_xlabel('Lag')
    ax.set_ylabel('Autocorrelation / Partial Autocorrelation')
    ax.grid(True)

plt.tight_layout()
plt.show()
```

```
from statsmodels.tsa.ar_model import AutoReg

# Fit an Autoregressive (AR) model
ar_model = AutoReg(final_df_filled['Unemployment, total (% of total labor force)
(national estimate)'], lags=2)
```

```
ar_model_fit = ar_model.fit()

# Print model summary
print(ar_model_fit.summary())

import matplotlib.pyplot as plt

# Obtain the fitted values
fitted_values = ar_model_fit.fittedvalues

# Obtain the observed values
observed_values = final_df_filled['Unemployment, total (% of total labor force)
(national estimate)']

# Plot observed vs fitted values
plt.figure(figsize=(10, 6))
plt.plot(observed_values, label='Original Data')
plt.plot(fitted_values, color='red', label='AR Fitted Values')
plt.xlabel('Time')
plt.ylabel('Unemployment (% of total labor force)')
plt.title('AR Model Fitting')

# Rotate x-axis labels diagonally
plt.xticks(rotation=45, ha='right')
plt.legend()
plt.grid(True)
plt.subplots_adjust(bottom=0.2)
plt.tight_layout()
plt.show()

# Calculate the moving average with a window size of your choice
window_size = 3
moving_avg = final_df_filled['Unemployment, total (% of total labor force)
(national estimate)'].rolling(window=window_size).mean()

# Plot original data and moving average
plt.figure(figsize=(10, 6))
plt.plot(final_df_filled.index, final_df_filled['Unemployment, total (% of total
labor force) (national estimate)'], label='Original Data')
plt.plot(final_df_filled.index, moving_avg, color='red', label='Moving Average
(Window Size = {})'.format(window_size))
plt.title('Moving Average Model Fitting')
plt.xlabel('Year')
plt.ylabel('Unemployment Rate')

# Rotate x-axis labels diagonally
plt.xticks(rotation=45, ha='right')
plt.legend()
plt.grid(True)
```

```
plt.subplots_adjust(bottom=0.2)
plt.tight_layout()
plt.show()
```

```
from statsmodels.tsa.arima.model import ARIMA

# Define the non-stationary time series data
y = final_df_filled['Unemployment, total (% of total labor force) (national
estimate)']

# Define the ARIMA model (p, d, q)
p = 1  # Autoregressive (AR) order
d = 1  # Integration (differencing) order
q = 1  # Moving Average (MA) order

# Fit the ARIMA model
model = ARIMA(y, order=(p, d, q))
arima_result = model.fit()

# Print the summary of the ARIMA model
print(arima_result.summary())

# Plot the original data and the ARIMA model's fitted values
plt.figure(figsize=(10, 6))
plt.plot(y.index, y, label='Original Data')
plt.plot(y.index, arima_result.fittedvalues, color='red', label='ARIMA Fitted
Values')
plt.title('ARIMA Model Fitting')
plt.xlabel('Year')
plt.ylabel('Unemployment Rate')

# Rotate x-axis labels diagonally
plt.xticks(rotation=45, ha='right')
plt.legend()
plt.grid(True)
plt.subplots_adjust(bottom=0.2)
plt.tight_layout()
plt.show()
```