01.020 DESIGN THINKING PROJECT III

A Food Security Index (FSI) Prediction Model

Prepared for:

Zeyu Sun
Lecturer, 10.022 Modelling Uncertainty
Singapore University of Technology and Design

and

Mei Xia Lin
Lecturer, 10.022 Modelling Uncertainty
Singapore University of Technology and Design

Prepared by:

Group 5
Nadine Verelia Moiras (1008481)
Valerie Lim Li Ting (1008199)
Jing Yang Loke(1008156)
Beatriz Acosta Ong (1008239)
Mariano Guillermo Vicente Perdices (1008005)

TERM 3 2024

Introduction: Food security is defined as when all people, at all times, have physical and economic access to sufficient safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life. In order to achieve that, we can use the **Food Security Index (FSI)** which is essential in measuring and comparing the food security status of countries. With economic, health, and agricultural factors being important aspects of the FSI, it is crucial to understand the relationship of these factors in regard to the FSI so that more effective policies and interventions can be built.

Problem Statement: How might we build a robust model to predict and better understand the factors influencing the Food Security Index (FSI) using economic, health, and agricultural indicators?

Exploratory Analysis & Key Hypothesis:

1) **Economic Stability:** Refers to a situation where all the essential economic resources of a country are available to its citizens, and no economic swings interrupt their daily lives. Stable economies ensure affordability and access to food. Key variables include **GDP, trade openness, food inflation rate, and expenditure on agriculture**.

2) **Agricultural Productivity:** Refers to the ratio of agricultural inputs to outputs - the higher the agricultural output, the greater the agricultural productivity of a farm. Effective agricultural systems ensure food abundance, diversity, and resilience against shortages. Key variables include **agriculture, fishing & forestry GDP, amount of agricultural land, and crop yield or production levels**.

3) **Health Indicators:** Refers to the health status and determinants of health of the population when subjected to inadequate nutrition and food access. Health indicators help assess how effectively food is utilized to meet the nutritional needs of the population. Key variables include **percentage of children stunted, the prevalence of undernourishment, and food insecurity prevalence**.

4) **Social Indicators**: Refers to assessing the overall well-being and quality of life of a population. Key variables like **HDI (Human Development Index)** combine health, education and income levels to evaluate the broader social conditions that affect food security.

Data Collection: Countries chosen based on the availability of data, while still ensuring a diverse representation from all the continents and different levels of income: for **Models 1 and 2**, our **X variables like GDP, trade openness, food inflation rate, and expenditure on agriculture**, as well as agricultural-related variables like **percentage of urbanized population and others** were sourced from World Bank and FAO (Food and Agriculture Organization). While **our Y variable (Food Security Index - FSI)** for all our models and **X variables of our Model 3 and 4 (HDI, percentage of children stunted, prevalence of undernourishment, and others)** were sourced from Economist Impact. The **year 2021** was chosen as it provides a comprehensive snapshot of global food security trends, with most of the datasets being readily available for this period and each source was chosen based on their reliability and relevance to the study's focus areas.

Data Cleaning: To address missing data, we used alternative countries with similar economic, health, or agricultural profiles. This preserved variability and ensured the analysis remained representative without reducing the sample size or introducing bias.

Model Development: Once the data was cleaned, we proceeded to understand the relationships between the variables and to test various potential models.

1) **Model 1 - Economic Indicators** (Focused on access and conditions influencing a country's food security)

- **GDP (Gross Domestic Product):** Reflects the overall economic output of a country, influencing both food availability (through domestic production) and food access (through purchasing power).
- **Trade Openness:** Measures the level of integration into international markets, influencing access to food imports and the ability to respond to food supply shocks.
- **Food Inflation Rate:** A key indicator of food price volatility, which can affect the affordability of food for the population.
- **Expenditure on Agriculture:** Represents government investment in agriculture, which can boost food production capacity and improve food availability.
- **Percentage of Urbanized Population:** Urbanization can affect food access, as urban areas often have better infrastructure and more diverse food options, but also more inequality in access

From the pairplot **(see Appendix A)**, we conclude that only urbanization has a sufficient correlation with FSI (R^2 value of 0.6). We can also see a linear relationship from the plot.

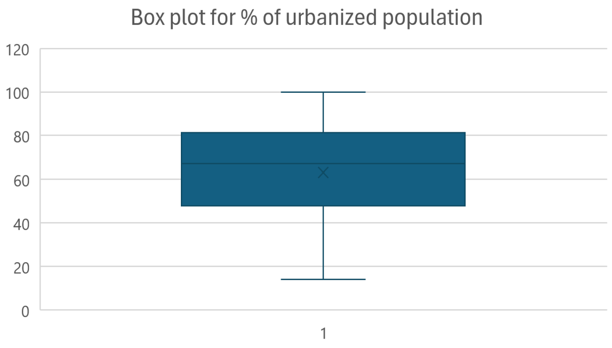


Fig 1: Box plot of predictor(s) of model 1

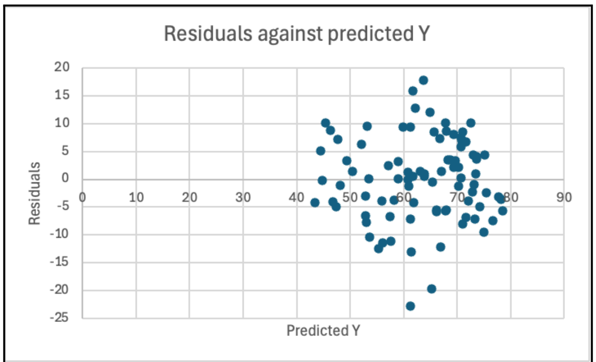


Fig 1.1: Residual plots of Model 1

The box plot of the variable suggests no data outliers. We then proceeded to create the linear regression model, summarized in **Appendix B**. The summary shows that the adjusted R^2 value is 0.594, which is relatively acceptable but can be better. Other than that, the p-value is $9.80 * 10^{-19}$, which is less than 0.05, indicating that the model is statistically significant. The standard error is 7.56, which can be improved. Next, we check for the rest of the linear regression assumptions.

From the scatter plot of residuals against predicted y, we do not see a pattern in the data points, indicating that the assumption of homoscedasticity and normality of residuals are satisfied. For both plots, there are 2 points that seem to deviate from the rest: Angola (Urbanization = 67.46, FSI = 45.5) and Haiti (Urbanization = 57.964, FSI= 38.5). Both are considered underdeveloped nations and we suspect there are unpredictable factors that might have caused the large residual values.

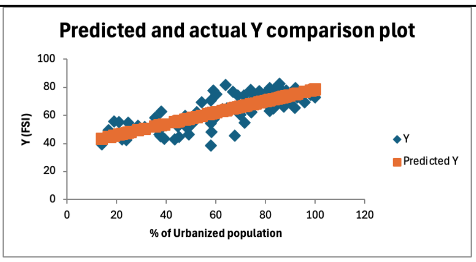


Fig 1.2 Predicted and actual Y comparison of model 1

Result: The model suggests percentage of urbanized population has a positive relationship with FSI, possibly due to infrastructure development that comes with urbanization; FSI = (37.7 + 0.408) * percentage of urbanized population and adjusted R^2 is 0.594

2) <u>**Model 2 - Agriculture and Economic Indicators**</u> (Building on urbanization, the model incorporated additional economic and agricultural variables to capture their impact on food availability and economic access.)

- **GDP Growth:** Measures the rate at which a country's economy is growing, directly influencing food security by increasing the capacity to import or produce food.
- **Consumer Price Index (CPI):** Indicates inflation, particularly the cost of goods, including food. Higher CPI often correlates with higher food prices, which can decrease food access for vulnerable populations.
- **Agriculture, Fishing & Forestry GDP (%):** This variable is a direct measure of the importance of the agricultural sector in a country's economy. Countries with high agricultural GDP tend to have higher food availability and a stronger agricultural infrastructure.
- **Amount of Agricultural Land:** Reflects the availability of land for food production, a key determinant of agricultural output.

From the pairplot **(see Appendix C)**, we conclude that only the percentage of GDP from agriculture, fishing, forestry has a sufficient correlation with FSI (R^2 value of 0.6).

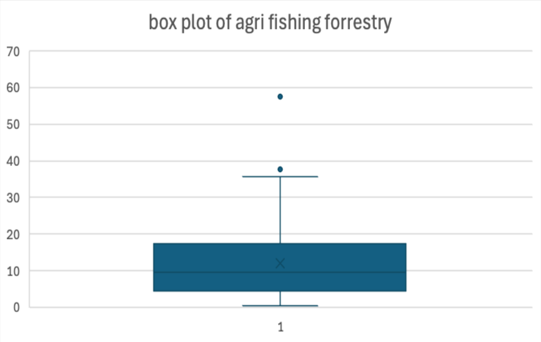


Figure 2: Box plot of predictor(s) for model 2

The box plot suggests 2 outlier values, but considering that these values are not illogical, we decided to keep them. We suspect a slight non-linear relationship due to its concave shape. After testing different trendlines and their R^2 values, we deduced that a polynomial trendline with a R^2 value of 0.7022 is best suited for the model and better than the linear model with a value of 0.6.

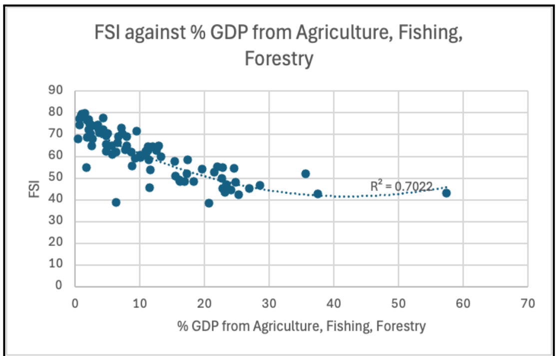


Fig 2.1: Scatter plot of Model 2

We proceed to add another column containing the squared values of the X predictor variable, making the model based on that and the original values to predict FSI. The summary shows that the adjusted R^2 value is higher (0.694) with p-values lower than 0.05.

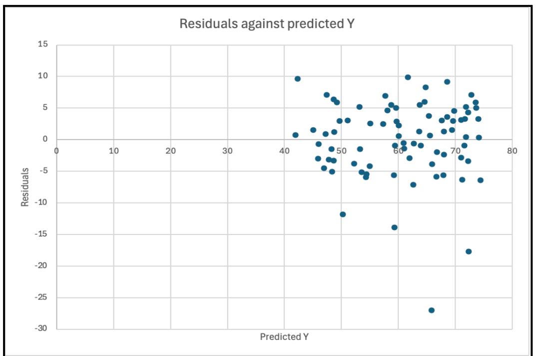


Fig 2.2: Residual Plots of Model 2

From the scatter plot of residuals against predicted y, we do not see a pattern in the data points, indicating that the assumption of homoscedasticity and normality of residuals are satisfied.

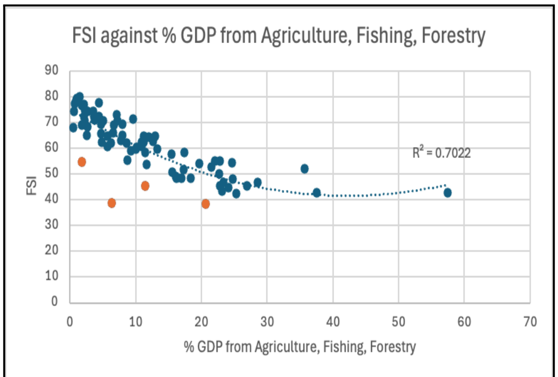


Fig 2.3: Scatter Plot of Model 2

Still, there is room for improvement as the model outliers indicate the presence of factors not taken into account by this model. The points with relatively big residual values are highlighted in the plot above.

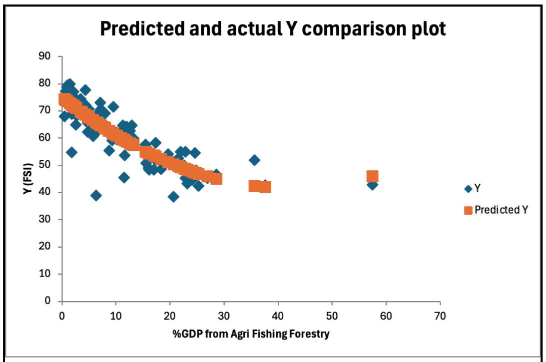


Fig 2.4: Predicted and actual Y comparison of model 2

Result: The model reflects a non-linear relationship between the percentage of GDP from agriculture, fishing, and forestry and FSI, best captured by a polynomial trendline with an adjusted R^2 of 0.694. We deduce that countries with a higher percentage of GDP attributed to agricultural sectors tend to be underdeveloped, and may lack the infrastructure and economic strength to support food security. Furthermore, over-reliance on the agricultural sectors means these countries are susceptible to climate change and natural factors, further reducing the stability of the economy.

3) **Model 3 & 4: Health, Agricultural, and Social Indicators**

In the third iteration, we expanded the analysis to include agricultural, economic, health, and social indicators, identifying key correlations through pairplot analysis which includes: **(1) percentage of children stunted, (2) prevalence of moderate to severe food insecurity, (3) percentage of children underweight, (4) human development index (HDI), and (5) percentage of GDP from agriculture, fishing, forestry.** However, HDI seems to be correlated with all the other factors. Additionally, percentage of children underweight and percentage of children stunted seem to correlate with each other (R^2 value of 0.71). As such, we chose to **use HDI as the sole predictor of one of the models**. We used the remaining variables as the predictor for another model, choosing percentage of children stunted over percentage of children underweight as it seems to be more correlated with FSI.

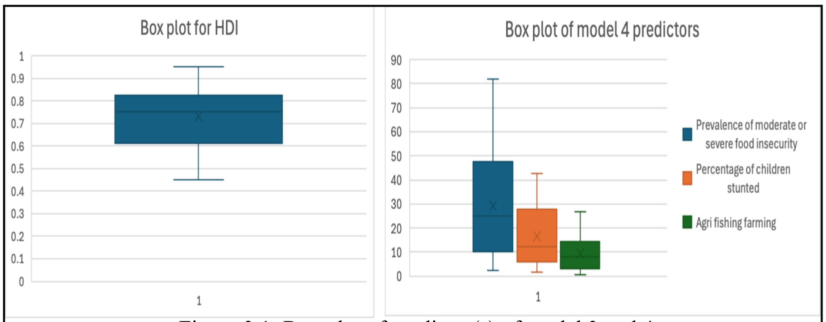


Figure 3.1: Box plot of predictor(s) of model 3 and 4

Model 3 focused solely on the **Human Development Index (HDI)**, a composite measure that includes health, education, and income, to test whether overall human development is a reliable predictor of food security. While **HDI** showed some correlation with FSI, it did not fully capture the granular health and agricultural factors influencing food security.

- **HDI:** A composite index that measures a country's average achievements in three basic dimensions of human development:
  - **Health:** Life expectancy at birth, indicating the average number of years a newborn is expected to live.
  - **Education:** Combined measure of mean years of schooling (for adults aged 25 and older) and expected years of schooling (for children entering school).
  - **Income:** Gross National Income (GNI) per capita, adjusted for purchasing power parity (PPP).

HDI and FSI show a linear relationship from the plot, so no transformation is required. The result of the linear regression model is summarised in **Appendix F**. The summary shows a satisfactory adjusted R^2 value, 0.843, which is the highest so far. Furthermore, p-values from the regression model are lower than 0.05 and the standard error is relatively small (4.26).

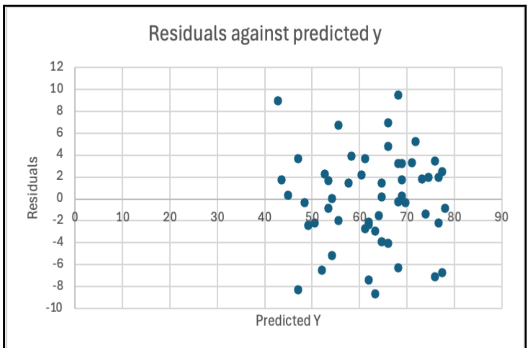


Figure 3.2: Residual Plot of Model 3

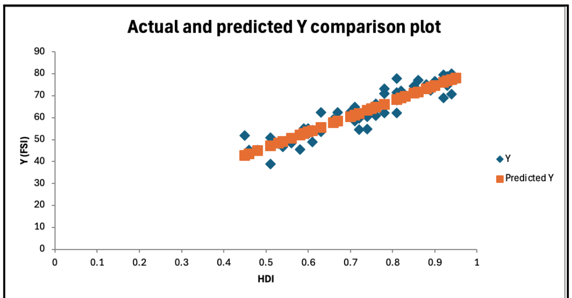


Figure 3.3 Predicted and actual Y comparison of model 3

Plotting the residuals against predicted values shows no pattern and appears to be randomly scattered. From this, we conclude that both the assumptions of normality of residuals and

Results: The model suggests HDI has a strong positive relationship with FSI; FSI = 11.122757 + 70.528329(HDI) and adjusted R^2 is 0.843.

constant variance (homoscedasticity) are satisfied.

- **Model 4** combines **Agriculture, Fishing & Forestry GDP**, **% of children stunted**, and **prevalence of moderate/severe food insecurity**.
  - **Agriculture, Fishing, and Forestry GDP (% of total GDP):** The share of a country's total Gross Domestic Product (GDP) generated by the agriculture, fishing, and forestry sectors.
  - **Percentage of Children Stunted:** The proportion of children under five whose height-for-age is below the World Health Organization (WHO) standards, indicating chronic malnutrition.
  - **Prevalence of Moderate or Severe Food Insecurity:** The percentage of the population experiencing moderate or severe challenges in accessing food, as measured by the Food Insecurity Experience Scale (FIES).

Since model 4 consists of multiple predictors, we first perform normalization on each predictor variable in order to prevent certain variables from dominating the model. From the plots, all predictors appear to have a linear relationship with FSI. It seems the quadratic relationship between percentage of GDP from agriculture, fishing, and forestry with FSI is less apparent due to the slightly different selection of countries based on the availability of data.

To confirm this, we created a model that takes into account the **quadratic relationship (Appendix G)** and observed that the p-value of the percentage of GDP from agriculture, fishing, and forestry squared is 0.234. Since it is greater than 0.05, the predictor is not statistically significant and can be removed. With the **removed predictor (Appendix H)**, the adjusted R^2 value is 0.780, which is good but not better than model 3. The standard error is relatively small (5.06) and all p-values show that the model is statistically significant. Plotting the residuals, we obtain:

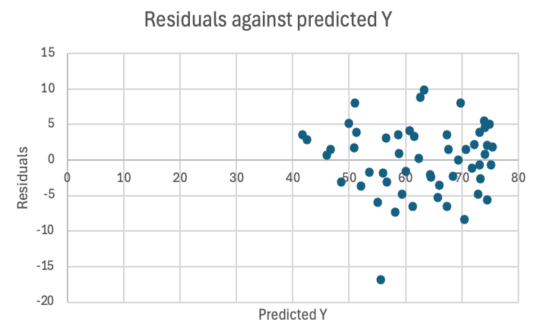


Figure 4: Residual plots of Model 4

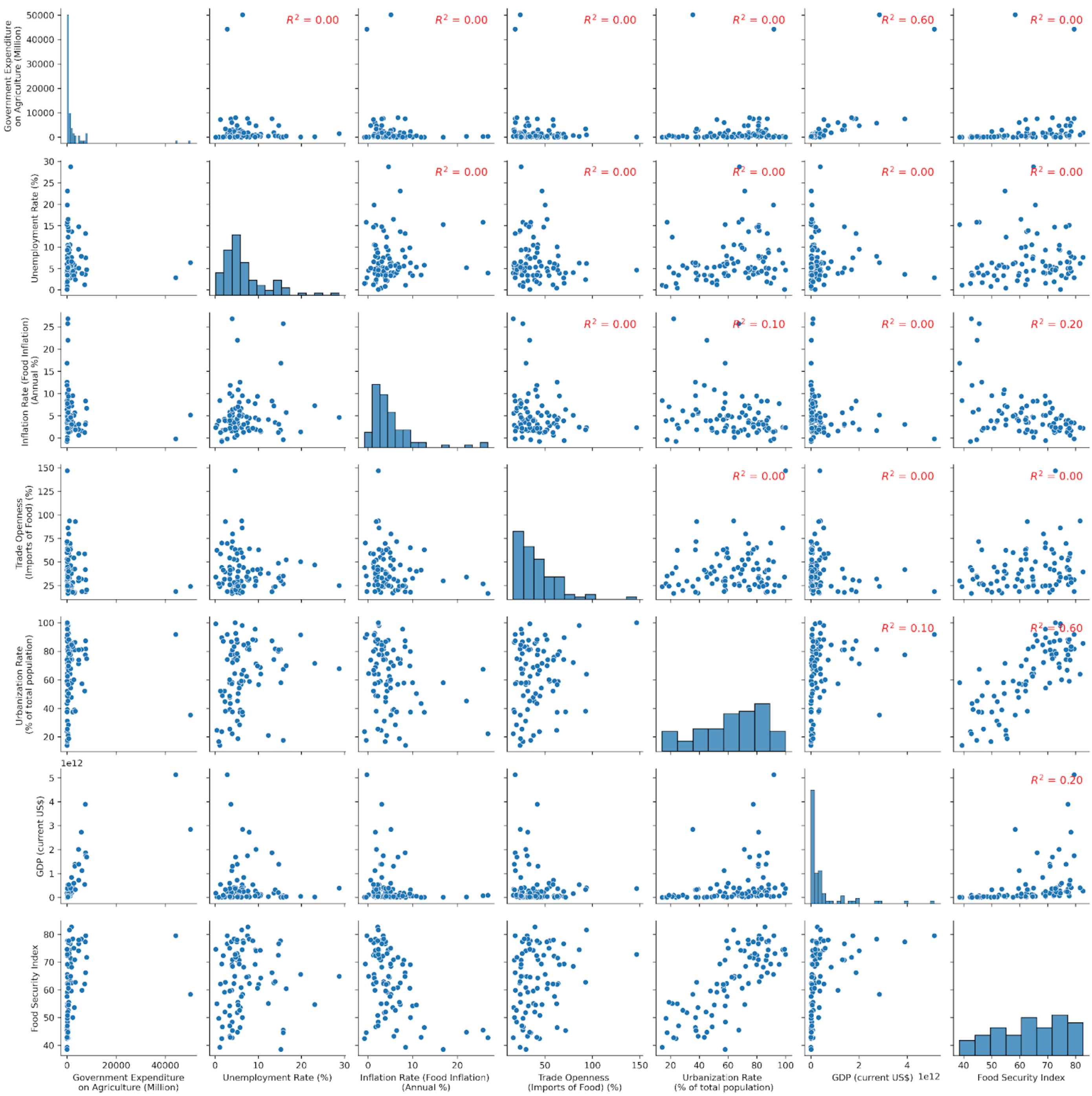
The residual plots show that there is one clear value with an unusually big residual. Upon further investigation, this value is Sudan. With a relatively low prevalence of food insecurity (49.4) and % of GDP from agriculture, fishing, forestry (6.36), Sudan has the lowest FSI of all the countries in the dataset (38.8). We then found out that in 2021, Sudan experienced a military coup, which we suspect is one of the main reasons behind such a low FSI. Apart from that, the homoscedasticity and normality assumption seem to be satisfied.

Result: The model suggests negative relationships between all predictors with FSI; FSI = 62.741 - 2.909 (prevalence of moderate or severe food insecurity) - 4.203 (percentage of children stunted) - 3.544 (agri fishing forestry) and Adjusted R^2 of 0.780.

Conclusion: We present 3 final models for 3 separate datasets, namely model 1, 2, and 3. For the third dataset, we chose model 3 as it is the one with better metrics. However, model 4 has more practical value as it gives a more detailed insight, allowing policymakers to target particular issues surrounding food security.

Models	Adjusted R^2	Standard Error	Formulas
Model 1	0.594022458	7.563021884	FSI = (37.7 + 0.408) * percentage of urbanized population
Model 2	0.694401378	6.13551459	FSI= 75.163 - 0.593 (Agri, fishing, forestry GDP) + 0.0188 $(Agri,\ fishing,\ forestry\ GDP)^2$
Model 3	0.842546619	4.25915624	FSI = 11.122 + 70.528 (HDI)
Model 4	0.779538034	5.039182922	FSI = 62.741 - 2.909 (prevalence of moderate or severe food insecurity) - 4.203 (percentage of children stunted) - 3.544 (agri fishing forestry)

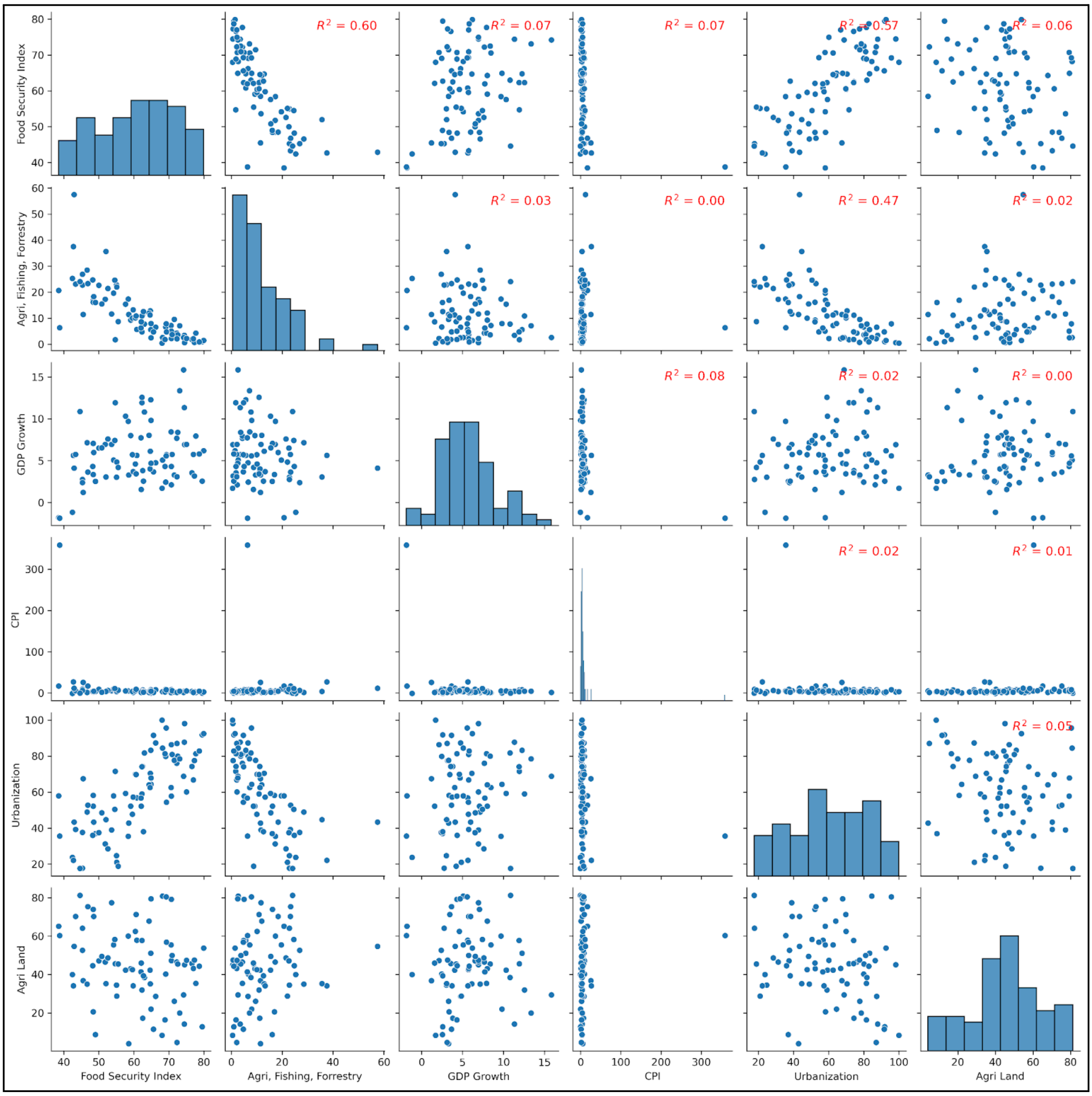
Appendix:



Appendix A: Pairplot for first dataset exploration

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.773749874							
R Square	0.598688867							
Adjusted R Square	0.594022458							
Standard Error	7.563021884							
Observations	88							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	7338.531107	7338.531107	128.2975684	9.80E-19			
Residual	86	4919.139802	57.19930002					
Total	87	12257.67091						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	37.69654526	2.409724883	15.64350583	6.9788E-27	32.90617074	42.48691979	32.90617074	42.48691979
% of Urbanized population	0.407627623	0.035987725	11.32685165	9.80E-19	0.336086394	0.479168853	0.336086394	0.479168853

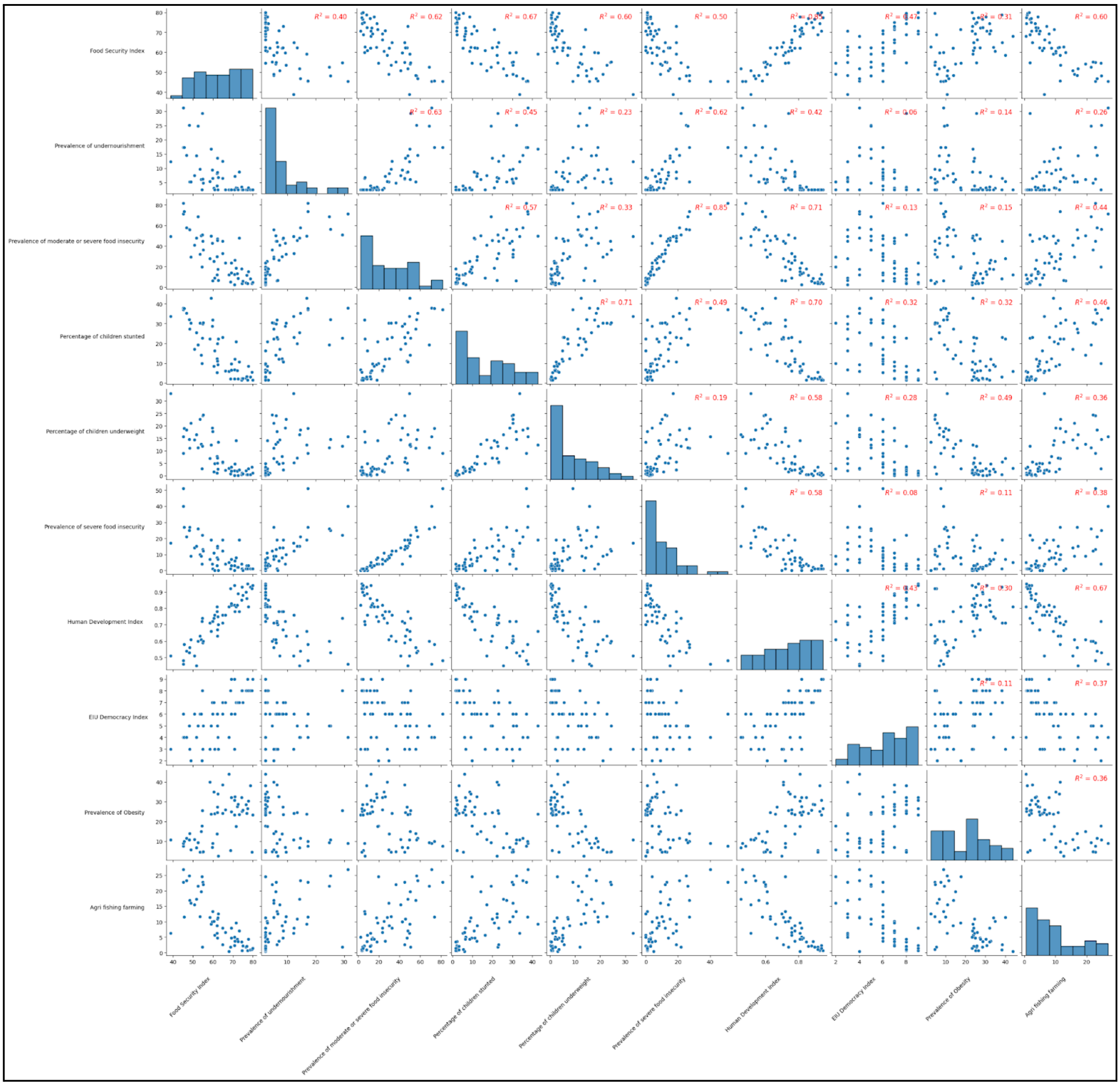
Appendix B: Model 1 linear regression



Appendix C: Pairplot for second dataset exploration

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.837995967							
R Square	0.70223724							
Adjusted R Square	0.694401378							
Standard Error	6.134551459							
Observations	79							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	6745.166829	3372.583415	89.61837654	1.01642E-20			
Residual	76	2860.086842	37.6327216					
Total	78	9605.253671						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	75.16327193	1.365004657	55.0644802	5.2658E-63	72.44462954	77.88191432	72.44462954	77.88191432
Agri, fishing, forrestry GDP	-1.593006913	0.162198783	-9.821324706	3.62161E-15	-1.916053786	-1.26996004	-1.916053786	-1.26996004
Agri, fishing, forrestry GDP squared	0.018871552	0.003707399	5.090239824	2.51775E-06	0.011487626	0.026255477	0.011487626	0.026255477

Appendix D: Model 2 linear regression



Appendix E: Pairplot for third dataset exploration

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.919551286							
R Square	0.845574568							
Adjusted R Square	0.842546619							
Standard Error	4.259156254							
Observations	53							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	5065.827667	5065.827667	279.2564837	2.47341E-22			
Residual	51	925.1610119	18.140412					
Total	52	5990.988679						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	11.12275668	3.143829532	3.53796431	0.000869566	4.811255535	17.43425783	4.811255535	17.43425783
X Variable 1	70.52832936	4.220481141	16.71096896	2.47341E-22	62.05535992	79.0012988	62.05535992	79.0012988

Appendix F: Model 3 linear regression

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.892466608								
R Square	0.796496647								
Adjusted R Square	0.779538034								
Standard Error	5.039812922								
Observations	53								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	4	4771.802393	1192.950598	46.96708729	5.11985E-16				
Residual	48	1219.186286	25.39971429						
Total	52	5990.988679							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	62.74150943	0.692271545	90.63135683	2.51167E-55	61.3496042	64.13341466	61.3496042	64.13341466
Prevalence of moderate or severe food insecurity	-3.053900355	1.125487993	-2.713401098	0.009220649	-5.316845632	-0.790955077	-5.316845632	-0.790955077
Percentage of children stunted	-3.753181945	1.203872972	-3.117589672	0.003078233	-6.173730787	-1.332633103	-6.173730787	-1.332633103
Agri fishing forestry	-6.936628619	2.984899992	-2.323906542	0.024410914	-12.93817229	-0.935084948	-12.93817229	-0.935084948
agri fishing forestry squared	3.303513985	2.738339335	1.206393212	0.233578372	-2.20228626	8.80931423	-2.20228626	8.80931423

Appendix G: Model 4 Regression (Quadratic)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.889002993							
R Square	0.790326321							
Adjusted R Square	0.777489157							
Standard Error	5.063177672							
Observations	53							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	4734.836041	1578.27868	61.56549209	1.20939E-16			
Residual	49	1256.152639	25.63576813					
Total	52	5990.988679						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	62.74150943	0.69548094	90.21312562	3.92279E-56	61.34388816	64.13913071	61.34388816	64.13913071
Prevalence of moderate or severe food insecurity	-2.909012865	1.124249795	-2.587514695	0.012684318	-5.168277412	-0.649748317	-5.168277412	-0.649748317
Percentage of children stunted	-4.203772656	1.149773407	-3.656174888	0.000624163	-6.514328823	-1.89321649	-6.514328823	-1.89321649
Agri fishing forestry	-3.544718342	1.006875339	-3.520513618	0.000941578	-5.56811009	-1.521326595	-5.56811009	-1.521326595

Appendix H: Model 4 Regression (Linear)

References

Corteva Agriscience (2022) 'Global Food Security Index (GFSI).' https://impact.economist.com/sustainability/project/food-security-index/download-the-index.

Food and Agriculture Organization of the United Nations 'FAOSTAT.' https://www.fao.org/faostat/en/#data/IG (Accessed: December 5, 2024).

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2012. Introduction to linear regression analysis. 5th ed. Hoboken, NJ: Wiley. Available at: https://statanaly.com/wp-content/uploads/2023/05/IntroductiontoLinearRegressionAnalysisbyDouglasC.MontgomeryElizabethA.PeckG_.GeoffreyViningz-lib.org_.pdf [Accessed 5 Dec. 2024].

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/AG.LND.AGRI.ZS.(Accessed: December 5, 2024)

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG.(Accessed: December 5, 2024)

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/NE.IMP.GNFS.ZS?end=2023&start=2000.(Accessed: December 5, 2024)

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?view=chart.(Accessed: December 5, 2024)

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD (Accessed: December 5, 2024).

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG.(Accessed: December 5, 2024)

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?view=chart (Accessed: December 5, 2024).

World Bank Group 'World Bank Open Data,' World Bank Open Data. https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS.(Accessed: December 5, 2024)