Noor Ahmed, Camila Garcia, Ananya Agarwal, Sonakshi Jain, Camila Berthin

BA222

Professor Leder-Luis

November 17, 2023

Project 2: Analysis of Socio-Economic & Demographic Insights

Our dataset consisted of variables that provided information about multiple socio - economic and demographic factors that aimed to conceptualize the global position of various countries and their internal conditions regarding their economy, healthcare, etc. The information we analyzed was taken from a dataset published under the name 'Global Socio-Economic & Demographic Insights, Popular Indicator of World Development (1973-2022)' and is available on Kaggle. Overall, the aim of the data was to uncover factors that play a significant role in the present conditions as well as the development of selected countries. Furthermore, we were inclined to choose this dataset as most of our group members have a strong interest in economic analysis (specifically macroeconomics) and come from diverse backgrounds, so we wanted to analyze a dataset that combined these two interests. Moreover, in the context of this data set, our group was interested in finding the extent of influence GDP has on variables such as population, fertility rate, foreign investments, and mortality rate, as our intuition suggests that high GDP could indicate a higher amount in certain areas.

The data intends to extensively map the above-mentioned variables over multiple decades and consists of information from 202 countries. It looks at both numerical and categorical variables to achieve this surmountable aim of drawing 'intricate patterns of fertility rates to the

economic pulses echoed in agricultural GDP contributions'[1]. Some categorical variables presented in the dataset are country, region, and subregion. Some of the key numerical variables presented include population total, population density, and GDP. It is important to note that the data points are supplemented by interpolation, meaning any missing data were replaced with substituted values. The authors did their best to ensure the integrity of the time series and maintain the authenticity of the trends, though they acknowledged that added values should be considered as approximations.

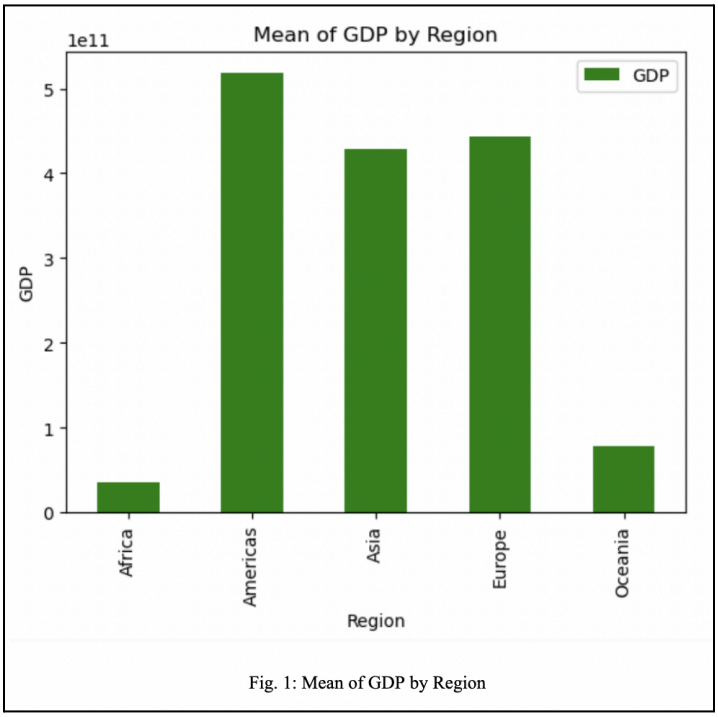

Fig. 1: Mean of GDP by Region

Since the primary focus of our analysis was GDP, we opted to analyze and check the average GDP for each region. We wanted to further analyze whether regions being developed or developing countries affected their GDP. The x-axis of the graph represents the different regions which are Africa, Americas, Asia, Europe, and Oceania. The y-axis represents the mean GDP. The graph is not symmetric and not uniform; rather, it is unimodal and skewed right. This graph provides a snapshot of the economic disparity between different regions. According to the analysis, Americas has the highest total average GDP making it the region with the most economic output and Africa has the lowest GDP indicating it has the

[1] "Global Socio-Economic & Demographic Insights Dataset," Kaggle, last updated November 6, 2023, https://www.kaggle.com/datasets/samybaladram/databank-world-development-indicators?select=world_development_data_imputed.csv.

least economic output among these regions. The GDP of Europe and Asia are relatively close, with Europe slightly higher. Oceania's GDP falls between that of Asia and Africa. Through this analysis, we can assume that developed regions like Americas and Europe have a higher GDP. Whereas, a higher number of developing countries in regions like Asia and Africa have a lower GDP. It's important to note that many factors can influence these numbers, including population, life expectancy, investment, consumer spending, net exports, inflation, etc.

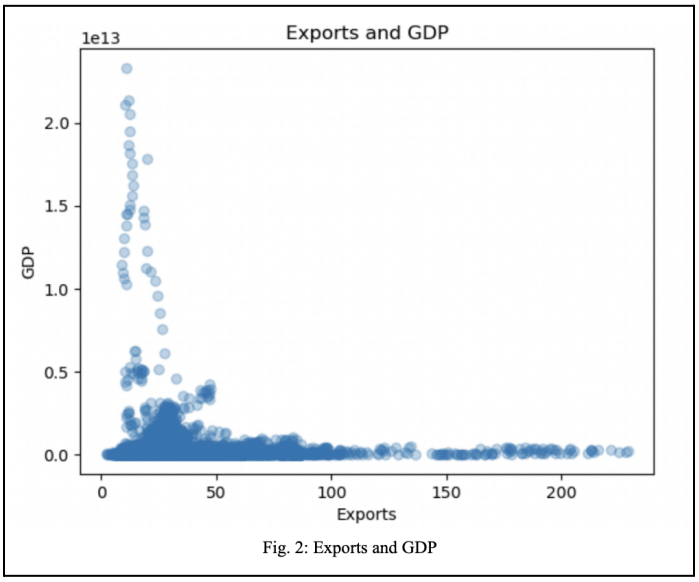

Fig. 2: Exports and GDP

We then ran a scatter plot, in order to further understand how exports have played a role within gross domestic product. Upon further analysis, the calculated correlation came to -0.127 which presents a weak correlation. Within Figure 2, we found that there is no direct linear correlation between both variables. This suggests that a higher number of exports did not contribute to a higher GDP. Moreover, this could be caused due to a number of factors [such as: the value of exports, trade,inflation and specific economic situations]. The instances in which the GDP had a higher value, indicate countries that feature a vast inclusion of economic activities, not limited to just exports. Looking towards the outliers within the scatterplot, these indicate smaller economies that heavily depend upon exports or countries that have developed certain goods that are exported. Overall, the analysis suggests that economic activities contributing to higher GDP extend beyond just the volume of exports.

With this scatterplot, we aimed to understand the relationship between Foreign Direct Investment and Gross National Income. Figure 3 does appear to indicate a positive strong correlation between both variables, also supported by the calculated correlation of 0.683. Based on a separate regression we ran, it revealed that for every additional dollar made by Foreign Direct Investment (FDI), Gross National Income (GNI) increases by 0.017 cents. This is predictive as an increase in GNI can be seen as increased trust in FDI in the country but it can also be attributed to factors like political changes, specific microeconomic agenda of investing enterprises like market entry, or a new government policy.

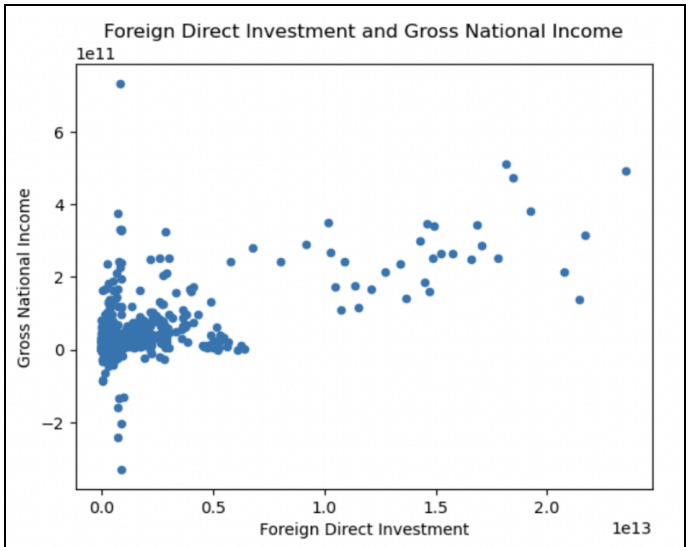

Fig. 3: Foreign Direct Investment and Gross National Income

Fig. 4: Mortality Rate Versus Life Expectancy

For this regression analysis, we aimed to see whether life expectancy and mortality rate have a relevant relation between them. Based on the graph in Figure 4, we can predict that countries with higher life expectancy at birth (in total years) will have a lower mortality rate (under 5 years since birth). Statistically, as a year increases in life expectancy in a country, the mortality rate (per 1000 births) of children under 5 years decreases by -3.9586. Thus, the

regression predicts that countries with an increase in life expectancy correspond with a decrease in mortality rate. Additionally, the regression performed was fit and strong for these two variables as demonstrated by the r-squared value of 0.777. We can attribute this to possibly a greater advancement in healthcare facilities but we cannot assume this to be casual as other variables like geographical location, population, and GDP can affect these results.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.3719	0.039	61.381	0.000	2.296	2.448
Americas	-1.3680	0.058	-23.539	0.000	-1.482	-1.254
Asia	-0.6048	0.056	-10.848	0.000	-0.714	-0.495
Europe	-2.1990	0.058	-37.838	0.000	-2.313	-2.085
Oceania	-1.3720	0.080	-17.097	0.000	-1.529	-1.215

Fig. 5: Urban Population Growth and Region

Figure 5 represents our final regression and looks into understanding if there is any relationship between Urban Population Growth rate within different regions. We decided to use urban instead of the overall population growth as cities/towns (urban centers) may have more opportunities for work and a dense amount of residents. All the coefficients for regions represented in the regression can be interpreted relative to the omitted region (base level), Africa. Compared to Africa's Urban Population Growth (UPG), the region of the Americas has a lower UPG rate of -2.26. We focus primarily on the Americas as it is the region with the highest GDP, based on our analysis. All other regions also resulted in negative coefficients, showcasing a significant decrease in UPG rate compared to the region of Africa. Moreover, the p-values are less than 0.05 and 0 is not in each confidence interval. Therefore, by looking at the different coefficients for UPG in these regions, we can predictably say that there are differences in each UPG between the regions. Considering Africa had the lowest GDP, the region is seeing continuous growth in its urban population, and no other region was shown to meet their rate of growth.

Following our two regression analyses, our group has come to the conclusion that there are no causal relationships between variables. There are too many factors that could affect GDP which is why one variable cannot solely reflect a causal relationship. Throughout our analysis, we finalized that the GDP is shown to have been impacted by multiple global socio-economic factors. With our focus on life expectancy, mortality rate, and urban population growth it can be said that the large overarching denominating factor had been centered around both the region and population. Areas of focus that the dataset should have included, to shed more light and give more insight into, could have perhaps been noneconomic factors. Because of the number of variables for this dataset, there was an overwhelming amount of data left to our discretion. With the countries variable, we could have narrowed our analysis in a country within each region to uncover which countries were contributing to their region's high or low GDP. This should further be examined to have a more comprehensive outlook on what factors like migration rate or mortality rate look for a country with a strong or weak economy. Nevertheless, based on an overall dataset correlation we did, it appears that some variables did not have a strong relationship with each other. This could have been due to the database from which the data was collected, along with the authors' decision to generate values for missing data. Therefore, if the dataset had not been manipulated perhaps there could have been stronger correlations which would have made it easier to utilize effective regressions.

In terms of each team member's participation, we all came together and put in effort to complete the project. Both team members Ananya and Camila G retained their focus on the regressions and their written analysis of the variables being analyzed. Team members Sonakshi, Noor, and Camila B appointed their main focus to graphs concerning the overall dataset and their analyses as well as verbalizing the coding into the report.