ORIE 3120: Final Paper

Amer Islam, Lingxuan Shi, Yoo Jin Bae

Introduction

Migration from country to country across the world has become easier than ever before for the common person as a result of developments in technology, political relationships, and travel. This phenomenon has supported the growth of not only voluntary immigration, but refugeeism as well. Disease, war, wrongful persecution, and natural disasters are still prevalent in many places across the globe and continue to drive people displace people both internationally and domestically. This exploration will investigate the economic and demographic factors that may bear an influence on refugeeism and internal displacement. This project is built off a Refugee Migration dataset from the UNHCR that breaks down populations of refugees and displaced peoples by nations of origin, nations of asylum, and demographic distinctions. The overall analysis aims to determine what factors could lead a nation or state to produce more displaced people, and what factors could lead a nation or state to offer asylum to more displaced people.

In order to produce a dataset that would allow us to evaluate the influence of different determinants on movement of displaced peoples, the pandas module was used in a Jupyter notebook running Python to combine the UNHCR dataset with data from the World Bank that describes different nations' performances in different economic indexes such as GDP per capita, annual GDP growth rate, and annual inflation rate. An area of concern in selecting variables of interest was that the countries that tended to produce more displaced peoples characteristically tended not to consistently collect and share a wide range of economic and demographic indexes. For this reason, variables that did not omit countries of interest were carefully included in the final dataset. The data was filtered to only include that which was collected in 2019 given that it was the most recent year for which the bulk of the data was available for countries of interest. Furthermore, given the spread of COVID-19 pandemic in early 2020, the little data available for this last year has likely been influenced by several confounding factors such as travel bans and heightened death tolls. This project will explore the influences of various determinants of international and domestic displacement by addressing several more specific questions.

1. Which indicators are strongest linear predictors for the number of refugees/internally displaced people produced for each particular country?
2. What is the influence of the total number of refugees hosted by a country on the total number of refugees produced by that particular country?
3. What is the best combination of variables for producing a model of the number of refugees/internally displaced people produced for each particular country?
4. Does a logistic model do a better job of predicting the total number of refugees produced by a country than a linear model?
5. To Are the number of male and female refugees produced by a country statistically distinct from one another?

Which indicators are strongest predictors for the number of refugees/internally displaced people to anticipate and prepare for shifts in the number of refugees produced in different nations, it would be helpful to know which factors best predict this metric. Visualization 1A for the most part seems to indicate that the higher the Gross Domestic Product (GDP) per capita, the more refugees scaled for

population there are per country, with darker circles tending to be larger, unusually with the exception of the two largest circles, representing Somalia and Afghanistan; these outlier values may be anomalous as a result of the influence of another variable. Visualization 1B seems to indicate less of a clear trend or pattern between Inflation Rate and our variable of interest in comparison.

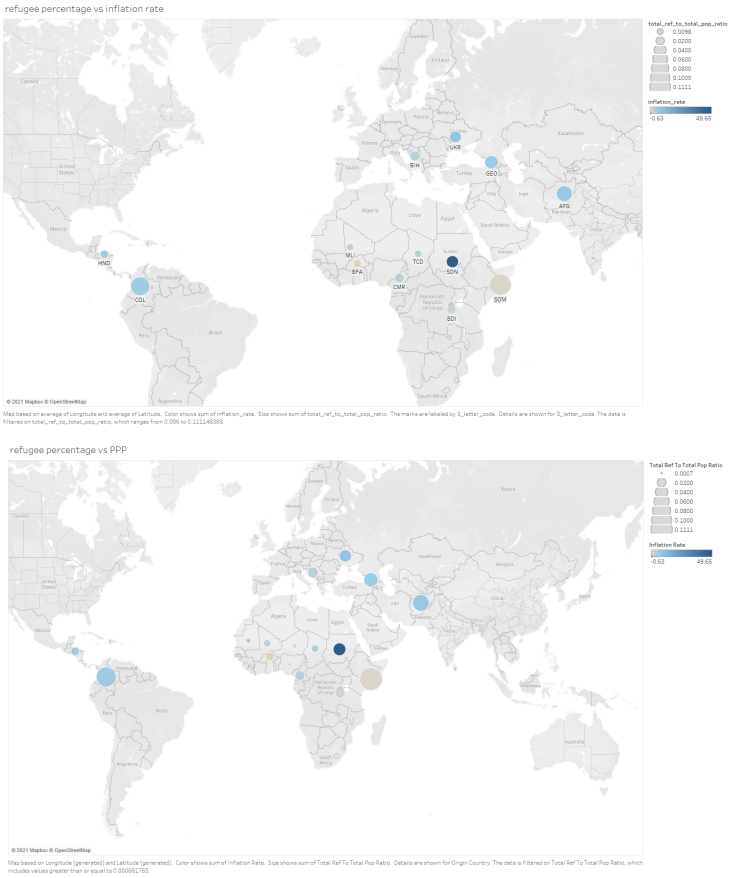


Figure 1A: GDP per Capita, Inflation Rate vs Refugees scaled to Population by Country

As part of addressing this question, this exploration used linear regression to fit several models for economic indicators so as to predict the total number of refugees produced by a particular nation. Each of these models unfortunately produced low R-squared scores and high P-values when assessed at a 95% confidence level. The best performing, albeit weak, model for the economic indicators was the inflation rate, which had an R-squared value of 0.027, and a P-value of 0.098 (Figure 1C).

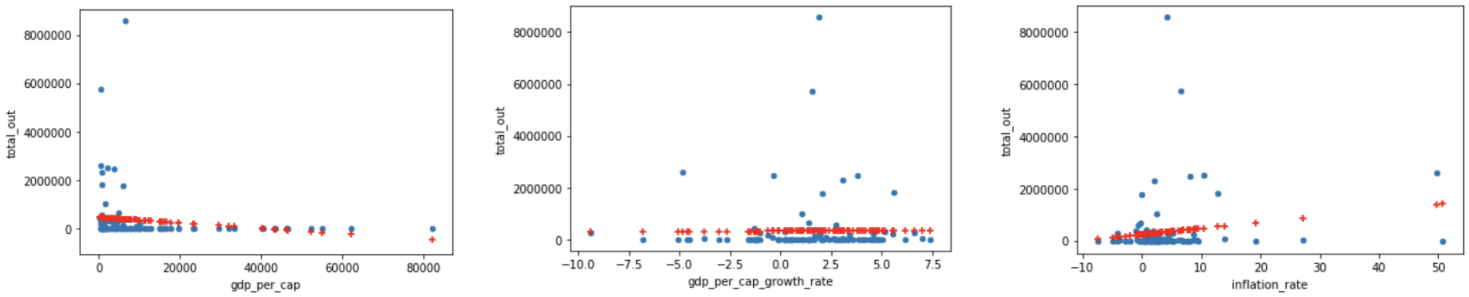


Figure 1B: Linear Reg. Plots for GDP per Capita, GDP Growth Rate, Inflation Rate vs Total Refugees Produced

OLS Regression Results

Dep. Variable:	total_out	**R-squared:**	0.024
Model:	OLS	**Adj. R-squared:**	0.014
Method:	Least Squares	**F-statistic:**	2.505
Date:	Fri, 21 May 2021	**Prob (F-statistic):**	0.117
Time:	13:41:33	**Log-Likelihood:**	-1594.2
No. Observations:	104	**AIC:**	3192.
Df Residuals:	102	**BIC:**	3198.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.58e+05	1.34e+05	3.418	0.001	1.92e+05	7.24e+05
gdp_per_cap	-11.2978	7.138	-1.583	0.117	-25.456	2.861

OLS Regression Results

Dep. Variable:	total_out	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.010
Method:	Least Squares	**F-statistic:**	0.01089
Date:	Fri, 21 May 2021	**Prob (F-statistic):**	0.917
Time:	13:42:33	**Log-Likelihood:**	-1595.5
No. Observations:	104	**AIC:**	3195.
Df Residuals:	102	**BIC:**	3200.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.289e+05	1.22e+05	2.701	0.008	8.74e+04	5.7e+05
gdp_per_cap_growth_rate	3967.5131	3.8e+04	0.104	0.917	-7.14e+04	7.94e+04

OLS Regression Results

Dep. Variable:	total_out	**R-squared:**	0.027
Model:	OLS	**Adj. R-squared:**	0.017
Method:	Least Squares	**F-statistic:**	2.793
Date:	Fri, 21 May 2021	**Prob (F-statistic):**	0.0978
Time:	13:43:10	**Log-Likelihood:**	-1594.1
No. Observations:	104	**AIC:**	3192.
Df Residuals:	102	**BIC:**	3197.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.385e+05	1.23e+05	1.942	0.055	-5146.974	4.82e+05
inflation_rate	2.311e+04	1.38e+04	1.671	0.098	-4318.970	5.05e+04

Figure 1C: Reg. Results for GDP per Capita, GDP Growth Rate, Inflation Rate vs Total Refugees Produced

What is the influence of the total number of refugees hosted by a country on the total number of refugees produced by that particular country?

During data preprocessing, it became apparent that several of the nations that produced high numbers of refugees also hosted a high number of refugees. This was not what may be typically expected given that if conditions in a country were such that people were fleeing it to seek asylum elsewhere, it would not traditionally make sense for that country to be a destination for others seeking asylum.

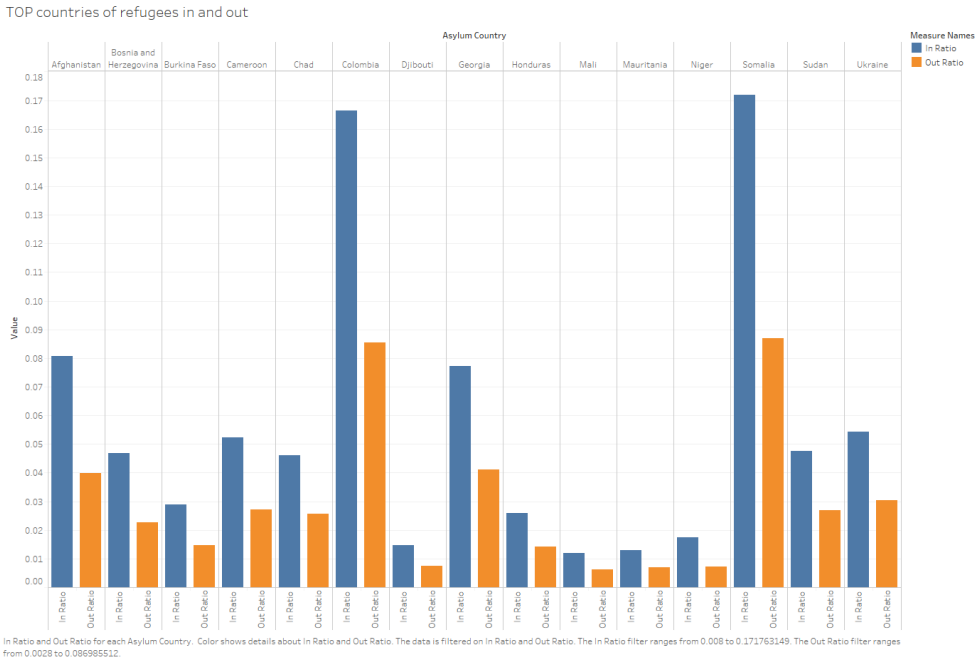
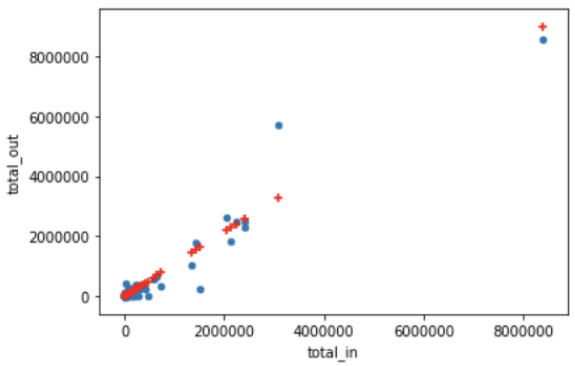


Figure 2A: Refugees leaving the Country to seek Asylum vs Refugees receiving Asylum in a Particular Country

As depicted in the diagram above (Figure 2A), there appears to be some correlation between the number of refugees produced by a country and the number of asylum seekers being hosted by that same country. That being said, strictly based on this visualization, this pattern may simply be a consequence of these nations simply having a high population size, predisposing them to both produce more refugees given the high overall population, but potentially to also have the manpower or capital to host more people.

Given the economic predictors were not producing strong models, the exploration opted to use refugee statistics themselves to produce a model predicting total refugees produced. Fascinatingly enough, the model that used the total number of refugees that a country hosts to predict the total number of refugees produced by a country produced an extremely strong model (Figure 2C), not only in comparison to the economic indicators, but overall.



OLS Regression Results

Dep. Variable:	total_out	**R-squared:**	0.924
Model:	OLS	**Adj. R-squared:**	0.923
Method:	Least Squares	**F-statistic:**	1235.
Date:	Fri, 21 May 2021	**Prob (F-statistic):**	8.46e-59
Time:	13:43:47	**Log-Likelihood:**	-1461.7
No. Observations:	104	**AIC:**	2927.
Df Residuals:	102	**BIC:**	2933.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1.876e+04	3.2e+04	-0.586	0.559	-8.23e+04	4.48e+04
total_in	1.0766	0.031	35.136	0.000	1.016	1.137

Figure 2B: Linear Reg. Plot Total Refugees Hosted vs Total Refugees Produced **Figure 2C: Reg. Results for Total Refugees Hosted vs Total Refugees Produced**

This model had an R-squared value of 0.924, indicating that the number of refugees hosted in a country explained roughly 92.4% of the variability of the data for the number of refugees produced by a country. The coefficient for total refugees hosted was 1.0766. Furthermore the model produced a P-value of 0.000, indicating that the relationship between these two variables is statistically significant given it is lower than the $\alpha = 0.05$. Contextually, this relationship is most likely the result of internally displaced peoples (IDPs), who are refugees that relocated within the nation they originated from. For each nation, an abundance of their refugees likely originated within the country itself and could not manage to cross the border so as to find refuge in a foriegn state. Another contextual explanation could be that those nations that produce refugees may border several other countries that are also subject to war, famine, or other forces that drive refugeeism. As a consequence, it is possible that many of these refugees shifted from the dangerous regions of their nations of origin to the safe regions of their host nations; the danger and threat would not be uniform across nations as a whole, but instead vary within. This would explain why these countries whose adverse living conditions have pushed people out have also been destinations of refuge for those individuals leaving other countries.

What is the best combination of variables for producing a model of the number of refugees/internally displaced people produced for each particular country?

This investigation, especially in the context of linear regression modelling, has looked at the significance and strength of individual variables on the number of refugees produced by particular countries. Given that there are a wide range of potential influences on refugeeism across countries, it would be likely useful to construct a multivariate model to predict creation and movement of displaced people in the future. To extend the exploration in this context, this investigation elected to use the AIC method to evaluate different groups of variables so as to find the strongest possible model. AIC, or Akaike information criterion runs multiple distinct combinations of variables and drops those that have larger P-values until it eventually returns a model with only the most statistically significant variables.

Given that the significance of model selection rests on there being a range of variables to test, for this segment of the exploration, In the investigation's code, the model was fit using a train and test split, so as to use the training data to select and fit the model, and the testing data to predict our total refugees produced, evaluating the strength split as a different, wider set of variables was used to test for the most statistically significant model. The model continued to drop non-significant variables until only those whose P-values indicated strength of model were left.

Dep. Variable:	diff	**R-squared (uncentered):**	0.982
Model:	OLS	**Adj. R-squared (uncentered):**	0.972
Method:	Least Squares	**F-statistic:**	97.02
Date:	Sat, 22 May 2021	**Prob (F-statistic):**	1.50e-07
Time:	19:09:28	**Log-Likelihood:**	-109.97
No. Observations:	14	**AIC:**	229.9
Df Residuals:	9	**BIC:**	233.1
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Year	-62.2142	12.225	-5.089	0.001	-89.870	-34.558
ppp	58.9072	12.317	4.783	0.001	31.045	86.769
gdp	-2.349e-06	4.51e-07	-5.204	0.001	-3.37e-06	-1.33e-06
life	2536.9863	467.329	5.429	0.000	1479.814	3594.159
secondary	-179.6622	27.748	-6.475	0.000	-242.433	-116.892

Figure 3A: AIC Modelling for a range of Variables

Our final, most statistically significant model included the following variables: Year, Purchasing Power Parity (PPP), Gross Domestic Product (GDP), and Life Expectancy. This model had an AIC score of 229.9 and statistically significant P-values falling below the α value of 0.05. The AIC model summary output is expressed below.

Total Refugees Produced = -62.2142(Year) + 58.9072(Purchasing Power Parity) - 2.349e-06(Gross Domestic Product) + 2356.9863(Life Expectancy)

Does a logistic model do a better job of predicting the total number of refugees produced by a country?

As an alternative approach to linear regression, logistic regression was also applied to the dataset. Logistic regression is used to assess whether or not there is a notable relationship between a dependent variable and various independent variables, going along perfectly with . This approach was especially appropriate for this project, since there is one dependent variable, total_out, and multiple economic indices being examined.

The model was built using Python. Prior to coding, multiple packages needed in building a logistic model on python were imported: numpy, pandas, and statsmodels.api. After the needed packages were imported, the same dataset that we used in linear regression, aggregate.csv, was imported. Then, because there were 5 null values identified inside economic index columns, data cleaning was done by

removing rows containing them. Data manipulation was also done on the total_out column, since in order to run logistic regression, the dependent variable needs to be categorised. First, the average total_out among all the countries were identified, then the total_out values for each country were categorised into whether or not it was above/equal to or below the total average. Finally, the logistic regression model was built with the sorted total_out values as the dependent variable, and the economic indices, gdp_per_capita, gdp_per_capita_growth_rate, and inflation_rate, as the independent variables.

```
Optimization terminated successfully.
         Current function value: 0.320165
         Iterations 9
```

Logit Regression Results

Dep. Variable:	totalOut	**No. Observations:**	123
Model:	Logit	**Df Residuals:**	119
Method:	MLE	**Df Model:**	3
Date:	Sat, 22 May 2021	**Pseudo R-squ.:**	0.2030
Time:	14:35:47	**Log-Likelihood:**	-39.380
converged:	True	**LL-Null:**	-49.410
Covariance Type:	nonrobust	**LLR p-value:**	0.0001651

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-0.7328	0.477	-1.536	0.124	-1.667	0.202
gdp_per_cap	-0.0002	9.61e-05	-2.496	0.013	-0.000	-5.15e-05
gdp_per_cap_growth_rate	0.0286	0.093	0.306	0.760	-0.155	0.212
inflation_rate	0.0268	0.032	0.850	0.395	-0.035	0.089

Figure 4: Logistic regression model for GDP per Capita, GDP Growth Rate, Inflation Rate vs Total Refugees Produced

The logistic regression model shows that the relationship between the number of total refugees produced per country and the GDP per capita is the only relationship with a p-value that is below 0.05. Therefore, through the model generated, it can be concluded from this model that the GDP per capita is the only economic index that has a statistically significant relationship with the total number of refugees produced per country.

Are the number of male and female refugees produced by a country statistically distinct from one another?

In the context of contemporary culture, the phrase “Women and Children first,” is frequently heard in times of distress or danger, especially in popular media. This investigation opted to explore whether or not these so-called societal norms gave rise to legitimate differences in the gender-based breakdown of refugee populations. The original dataset already included a demographic breakdown of refugees produced by Gender, which allowed the investigation to briefly review the data at the beginning of the exploration. A cursory reading of the original dataset suggested some mild variation between male and female refugee populations, but it was not obvious as to whether or not these differences were statistically significant.

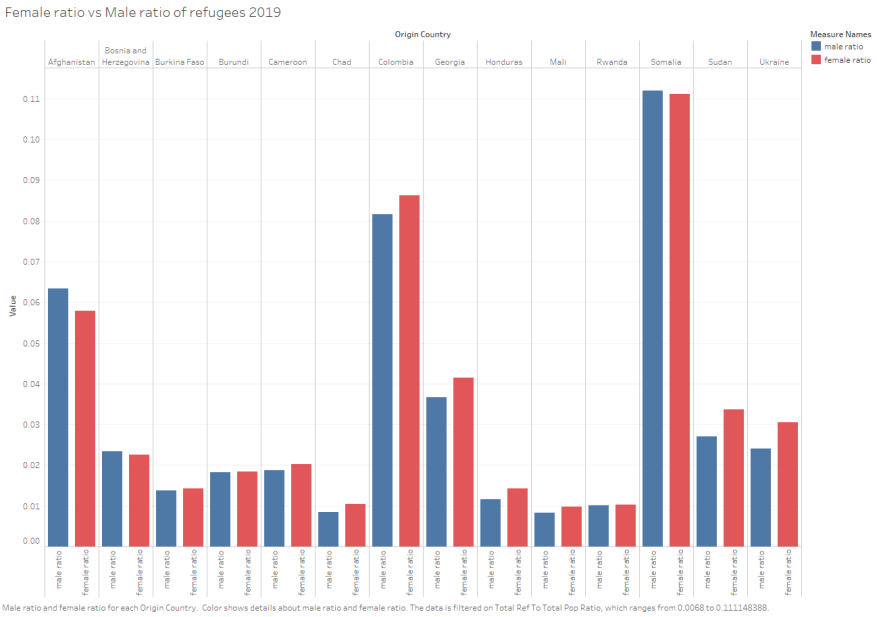


Figure 5A: Male Refugees vs Female Refugees

The diagram above (Figure 4A) does depict a general variation in the data for the ratio of male refugees to female refugees, but one demographic group was not consistently above the other. In order to further extend this exploration, this study opted to use a two sided T-test to determine whether or not there was a statistically significant difference between these refugee demographic groups. T-testing is a statistical method used to determine whether or not two entities have statistically significant differences between one another or not using their mean values.

```
Ttest_indResult(statistic=-1.1186849245951211, pvalue=0.2706838602090723)
```

Figure 5B: T-test output for Male Refugees vs Female Refugees

The T-test comparing number of male refugees to female refugees produced a P-value of 0.2706838602090723, which fell above the 95% confidence level α of 0.05, indicating that contrary to the previously mentioned societal norms, there is not a statistically significant difference between male and female refugee populations.

Conclusion and Review

From the initial stages of this investigation, the exploration faced obstacles in terms of the nature of the data it was handling. The variables appended to the base dataset were limited on the basis that the nations that produce the highest numbers of refugees did not frequently or accurately collect economic or demographic data through means or national survey or census. Omitting those nations that did not have data for variables of interest would severely restrict the investigation, and ignore some of the most significant producers of refugees on the planet. Thus, exploration has been forced to rely on several more limited national metrics for developing and testing models. It is possible that this is the reason that many of the models for linear and logistic regression did not demonstrate a strong or statistically significant relationship. That being said, it was very telling that the number of Refugees hosted in a country was a

strong predictor of the number of Refugees produced. This heavily implies that the United Nations High Commissioner for Refugees and similar organizations would benefit most from using historical and current refugee metrics themselves to anticipate and predict future changes in refugee populations. That being said, the AIC analysis revealed that a multivariate approach to using economic indicators could produce a strong model for total refugees produced. Based on this observation, the results of this exploration would support the UN collecting this data itself. It is unfortunate that that these developing nations which are subject to the adversity that drives the growth of refugee populations characteristically do not have much data available to evaluate; these nations are stuck in situations where those countries that could benefit most from data-driven predictions and policies in regard to refugees are those who have the least data available themselves. In the future, governments and NGOs should make it a priority to collect this data by any means possible given its immense value for developing policies and strategies to tackle the issues of hosting refugees and facilitating their movement.

Bibliography

Core Dataset

- UNHCR Dataset containing data for Refugees and Displaced Persons categorized by Country of Origin, Country of Asylum, and Demographic Breakdown
    - https://www.unhcr.org/r<efugee-statistics/download/?url=E1ZxP4

Merged Datasets

- Kaggle Dataset for Latitude and Longitude by Country
    - https://www.kaggle.com/eidanch/counties-geographic-coordinates
- World Bank Dataset for GDP per Capita
    - https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2019&name_desc=false&start=1997
- World Bank Dataset for GDP per Capita Growth Rate
    - https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG?end=2019&name_desc=false&start=1999
- World Bank Dataset for Inflation Rate
    - https://data.worldbank.org/indicator/NY.GDP.DEFL.KD.ZG?end=2019&name_desc=false&start=2019
- World Bank Dataset for Population
    - https://data.worldbank.org/indicator/SP.POP.TOTL?name_desc=false
- World Bank Dataset for PPP
    - https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?view=chart
- World Bank Dataset for Life Expectancy
    - https://data.worldbank.org/indicator/SP.DYN.LE00.IN?view=chart
- School enrollment, primary (% gross)
    - https://data.worldbank.org/indicator/SE.PRM.ENRR?most_recent_year_desc=false