**ORIE 4741** **Final Report** **Fall 2021**

# Contents



# SUMMARY


In this project, we seek to help understand the variables that help promote economic growth within a nation. As our data is multidimensional and includes variables over many different countries, groups of countries, and times, we were able to multiple analyses from different perspectives. This includes looking at a single time frame and using sparse regression to determine which variables are contribute positively to GDP growth, and looking at a specific group of countries (grouped by development) over time to perform a time series analysis. These methods and models allow us to investigate the budgeting decisions a nation might make while also analyzing global economic patterns. As the data is multidimensional, data processing was an important step in this project, as shown in section 3. In section 4, we perform the spare regression for a *single* time frame to investigate important features. In section 5 and 6, we look at the data across time with both automated machine learning and an ARIMA model for time series analysis, respectively. Then we show the value of our project with a case study in section 7. Lastly, we discuss the limitations of our models and analyses while also discussing the many opportunities for future work in section 8.

***Methods used from class includes sparse regression (section 4), automated machine learning for hyper parameter tuning (section 5), an the ARIMA model for time series analysis (section 6). There are multiple plots throughout the paper for the visualization of our models, including in the appendix. Discussions of results, confidence, and fairness are also included in section 8***

# 1 Introduction

The economic status of a nation can have enormous effects on the quality of life for its citizens. When a nation has less resources that it can allocate to its citizens, the citizens cannot thrive and end up living in worse environments; in turn, a nation without comfortable and educated citizens cannot grow and lift itself from poverty. Hence, a complex issue for underdeveloped nations is determining how to allocate resources and how to make budgeting decisions, as they are under tight constraints and these decisions can create feedback effects with regards to their economic status. Likewise, this issue is still prominent for developed nations because it would be advantageous to make decisions that secures a nation's power and stabilizes the lives of its citizens.

There are many factors that may be used to decide how a country should budget and allocate resources. The choices a nation must make may include, for example: whether or not to export certain products, how much of a product to produce and export, how much money to spend on education, health care, and other social services, how much to spend on infrastructure, etc. The factors that go into these decisions are both specific to the nation and dependent on the dynamics of the global economy, which depends on the place in time. Thus, the decisions a nation must make are truly complex and these best decision is not usually chosen with quantitative or empirical methods. Instead, policy makers, diplomats, legal officials, and other governmental figures might gather to decide the best course of action. However, using machine learning would be extremely advantageous, as there are more objective techniques that can help governmental officials determine how to help their country grow.

The question that we seek to answer is "What steps should a country take to develop?" Development has traditionally been looked at through an economic perspective, although there are many other social factors experts study to determine development such as the average lifespan of a person and happiness or well being. We will primarily look at development through the lens of economic development with metrics like GDP and GDP percent growth. **The goal of this project is to analyze factors that may contribute to a countries GDP and find opportunities for growth.** We will look at factors such as trading patterns, infrastructure, education, and other economic and social variables to indicate what variables hold the most potential for growth when the variables are changed. This project is important to understanding the dynamics of world development and the results can help a country understand how to better utilize its resources to optimize growth.

**We are going to answer our question with the results of a few different analyses: robust, sparse regression for finding variables important to a nation's economy, time series forecasting for model and actual use of our results, and lastly, a case study to implement our results and show its effectiveness.**

# 2 Dataset

Data is taken from the World Bank World Development Indicators:

https://databank.worldbank.org/source/world-development-indicators

The dataset is multidimensional and includes statistics for every country on the topics of Economic Policy & Debt, Education, Environment, Financial Sector, Gender, Health, Infrastructure, Poverty, Private Sector & Trade, Public Sector, Social Protection & Labor, and Social Health. Examples of some of the features are "Population ages 15-64, female (% of female population)," "GDP growth (annual %)," "Age dependency ratio (% of working-age population)," and "Merchandise exports (current US$)." There are also standard world development related statistics like total debt and GINI index. In total, there are 1443 of these features for each country. These features also exist for groups of countries like continents, regions, and groups based on development (low vs middle vs high income) and they exist for different years from 1960 to 2020. None of our features are ordinal and all are real continuous values, as warranted by the World Bank Indicators format.

# 3 Data Processing

Our data is multi-dimensional: the dataset includes 1443 features for each country for each time from 1960 to 2020. There are multiple ways we could analyze the dataset to answer our question, and thus, there are multiple ways we could

clean our data and use feature engineering techniques. Because we want to understand what features are most important without regards to any country, we will not look at country-specific or region-specific models (until Section 7, which includes a case study for a specific country). Instead, when conducting the time series analysis, we will group countries based on current development, as defined by the World Bank which defines the development of countries by gross national income (GNI) per capita [2]. Instead of manually grouping each country into different levels of development and creating a definition of development unique to this project, we will use groupings provided by the World Bank. As there are multiple features for each country, the data and features also exist for regions, continents, and levels of development. Continents include groupings like "Latin America & the Caribbean," while regions are smaller geographical locations, like "Africa Eastern and Southern" or "Middle East and North Africa."

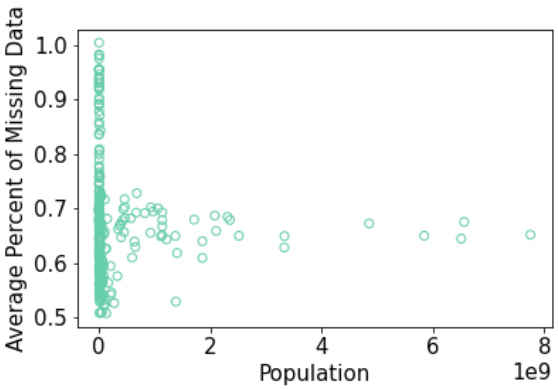
### 3.1 Cleaning & Feature Engineering

Almost all of the features have missing data for at least one country at some time. During a first attempt, we used only features that did not have missing data for the first $n$ years, where $n$ was a threshold that we could use to select for the number of features. This proved to be an impractical method for selecting features as a large portion of features have missing data (recall the any one feature holds data for each country at each time; thus, it is unlikely that this statistic was available for each of the more than 200 countries for each year). We decided to select for nonmissingness without tolerances: each feature is missing $x$ number of years for some country and we get the average percent of years that it is missing for each country. Then we loop through the features and their respective percentage and select for the features with lowest percent of missing data.

One might wonder whether or not this technique loses features that are important for distinguishing between countries; some features might have missingness that tells us something about the nation or time with missing data. We decided to see which countries had the most missing data. The following list is a partial list of those countries:

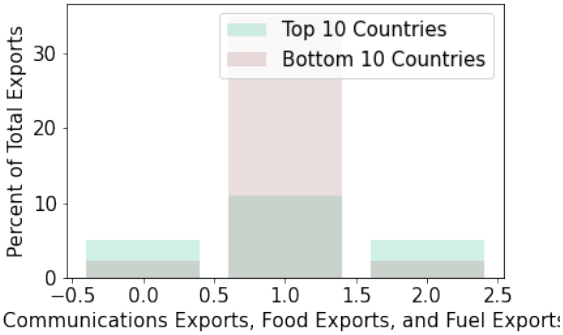
$$
\begin{aligned}
[0.9544583]&\,American Samoa\\
[0.93709864]&\,Andorra\\
[0.75776161]&\,Antiguaand Barbuda\\
[0.83512359]&\,Aruba\\
[0.88856549]&\,Bermuda\\
[0.9533495]&\,British Virgin Islands\\
&\vdots
\end{aligned}
$$

It seems as though these countries do have something in common: their small population. We then compared population to missingness and got the following plot:




Population versus the average percent of missing data across time.

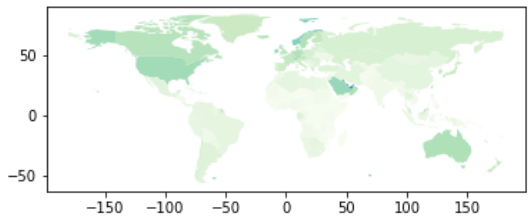
From the plot, we see that countries with small populations have a much larger range of missing data and that other countries have a steady amount of missing data. This confirms our beliefs: countries with very small populations have high levels of missing data, while other countries have about the same amount of missingness. We see that the missingness does hold information, but not information that is very relevant to our project.




Here, as part of a preliminary analysis, we looked at the top 10 and bottom 10 countries in terms of GDP and choose some of the trade statistics to compare, as GDP is a measure calculated by exports and imports. Exports in this

dataset were recorded as a percentage of total exports, so we did not have to worry about bias between the top and bottom 10. The plot below compares three of the exports for the bottom and top 10 countries and shows that countries with low GDPS tend to have food exports be a large percentage of their merchandise exports, whereas high GDP countries, while still having food exports be a leading percentage, have a smaller difference between percentages. These are the type of results that we want to show and prove in our analyses below.

Lastly, when looking through our data, we created the map below.




We already see that wealth might have a lopsided distribution. We want to further investigate what helps create wealth next.

# 4 Simple Models for Analysis

For analytical reasons, we conducted a simple least squares regression some of the variables as part of our preliminary analysis so get a sense of what variables, subject to our bias when choosing these variables, might have a sizeable impact on the wealth of a nation. Although not as informative as a machine learning model, as we will rarely find new nations to test this model on, this serves as a useful tool for analyzing which features hold the most weight when solving for GDP. Thus, sparse regression should return valuable information on which features (trade decisions, social factors, etc.) are most important and correlated with growth.

*Note: plots for this analysis are included in the appendix, if the reader wishes to view them.*

The matrix that we are using for this analysis is for a certain year. By "predicting" the GDP, we can interpret the coefficients to understand which of the statistics have a positive/negative contribution to the GDP. We found that communication export had a particularly large coefficient, implying that for the specific year that the matrix was taken from, increased trade in communications as a percent of total trade would increase GDP more than other trades, in general. Similarly, we found that "Rural population (% of total population)" was the second highest coefficient, implying it is an important factor. However, this is more likely to just be correlated with a high GDP and is not something a country can directly change to help their economy. The importance of such analyses of the results are discussed in Section 7 and Section 8, where we implement our methods once again to show how the results should be interpreted and discuss the limitations and use of the models.

# 5 Automated Machine Learning

Performing a well posed analysis where we consider our features across time but without looking at any one country will help us make broader, generalized statements about trade and other factors that contribute to growth.

## 5.1 Naive Automated Machine Learning

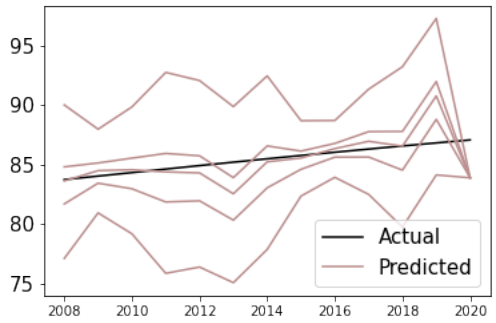
A naive strategy for using automated machine learning would be to blindly enter our data into an automated machine learning tool and analyze the results. We ran Miscrosoft's "Fast and Lightweight Automated Machine Learning," or flaml, on our data to test this strategy. We split our dataset generically with the first half being the training set and second fourth years being the validation set. The model that was returned by flaml was a decision trees model. However, we know that this is unsound for our application, as each of our rows represent a different time. Although the relation of the predictions to each of the other predictions is dependent on the rows, the predictions themselves will also change over time.
That is, getting new rows or new data (which would be moving forward in time) would not be predicted well with decision trees because the boundaries of the leafs should change over time.

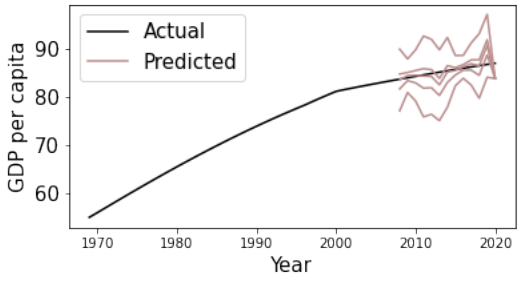
## 5.2 Automated Machine Learning: Hyperparameter Tuning

A better strategy might be to use automated machine learning for a specific model and tune the hyperparameters. We used automachine learning to tune the hyperparameters of a robust regression model. Regresson was used in order to capture continuous variables and robust

is necessary because of there are many outliers in with regards to wealth, as a few countries have economies much greater than most of other countries. The results of the given model for each of the low, middle, and high income group of countries are given below.

| Model | Test Error |
|---|---|
| Model 1: Low | 7.93 |
| Model 2: Middle | 18.75 |
| Model 3: High | 58.70 |




**The results above are shown for multiple predictions that are the simulations of different decisions a country might make. Each predicted line is the GDP of a country if one of the variables is changed by some random amount, which is analogous to a country making a decision to increase or decrease a certain variable at the time the prediction starts. As we can see from the plots, there are certain decisions that are predicted to raise the GDP more than others. This is one of the primary uses of the models and creating simulations like these can help answer our initial questions about the variables a nation can control to create wealth.**

# 6 Time Series Analysis

Performing a time series analysis will help us understand which what steps a country should take for growth by helping us understand the current dynamics of an economy. We let the first half of the dataset be the training set, while the second half was split in half to be the validation and test set. **We ran an autoregressive integrated moving average model to forecast the change in GDP**. *Recall from class some details about time series forecasting* from class:

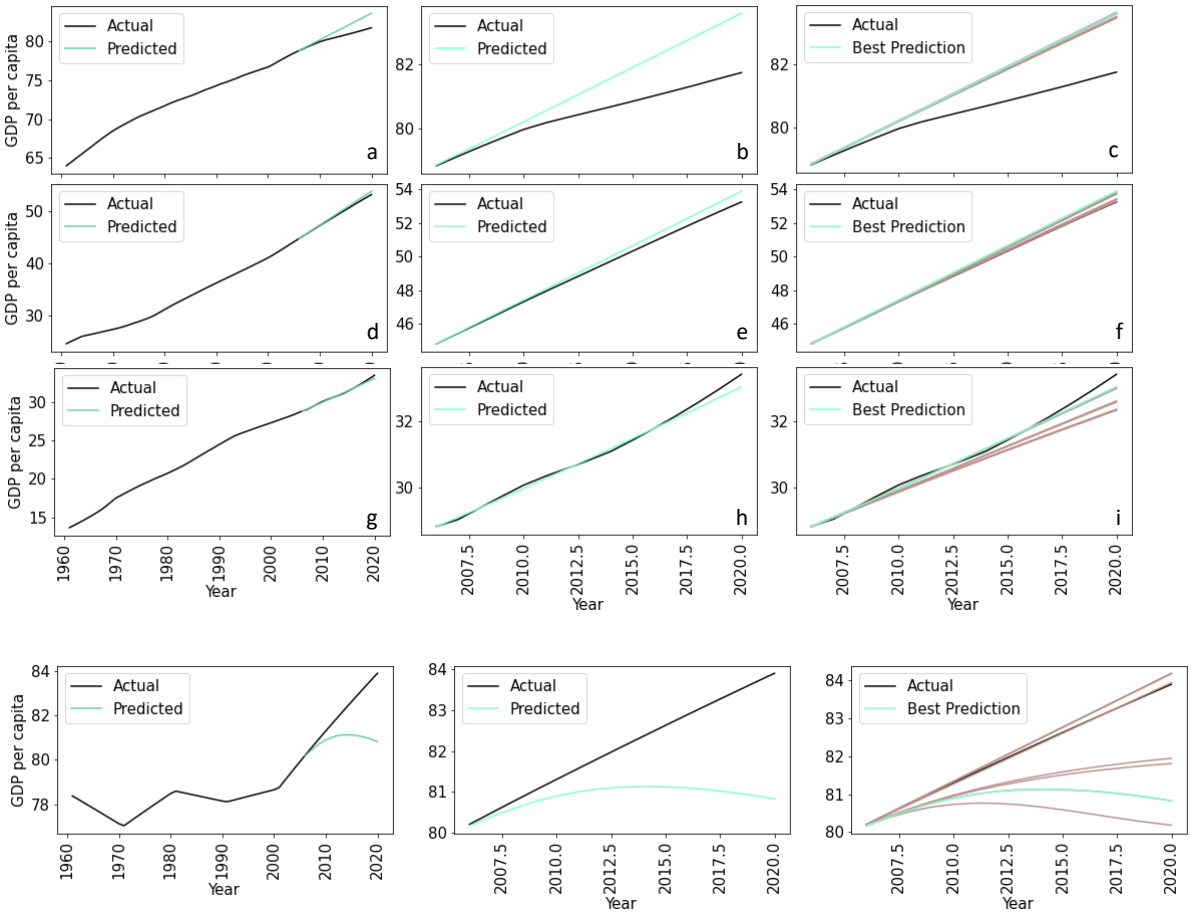
The auto regressive model finds a weight $w$ that places its coefficients on the lagged outcomes and each prediction is based off the lagged outcomes and the previous prediction. At each prediction step, we are only predicting the value at the next time. The auto-regressive moving average (ARMA) model writes the prediction, still a single value for a single time, is a linear combination of the $p$ most recent observations and the $q$ past residuals. This is under the condition that the time series is stationary, or independent of the time. This problem can be solved with least squares linear regression where we are minimizing the sum of the squared residuals for each time where the residuals.

If the time series is not stationary, as in this project, we might have an ARIMA(p,d,q) process. As before, $p$ here is the number of past observations we look at, $q$ is the number of past residuals we look at, and $d$ is a parameter for the backshift operator. A time series is ARIMA(p,d,q) if $(1-B)^d a_t$ is ARMA(p,q) where $B$ is the backshift operator which shifts the value back one [1].

We get stationarity if we have time series like $a_t = a_{t-1} + r_t$ where $a_t - a_{t-1}$ is stationary because $a_t - a_{t-1}$ is random. For our project, we have $a_t = a_{t-1} + g_t$ where $g_t$ represents that added GDP that is a function of the growth rate (so not the growth rate itself, but the added GDP as a result of the growth rate).

In this project, we can make an exception, because over short periods of time, the growth rate doesn't change much. This is especially true because we are looking at the rates for high, middle, and low income countries, which have averaged statistics and are less likely to have large jumps or dips. Thus, we can treat $g_t$ as a constant when predicting over short distances. So we are able to use ARIMA and the assumptions are (approximately) true.

To measure error, we will look at the mean squared error, or mean squared residual here, where the residual is the difference between the predicted and actual. At each step, there is a




new prediction, so we will take the square root of the sum of the squared residuals.

Because we need that p, d, and q, are integers, we can do a grid search when cross validating, as opposed to other searches like randomized search as an automated machine learning technique. Thus, we **performed a grid search to find the best hyperparameters when cross validating**.

The time series results for each of the groups are given below.

| Model | Test Error |
|---|---|
| Model 1: Low | 7.93 |
| Model 2: Middle | 18.75 |
| Model 3: High | 58.70 |

By looking at the table and the plot, we see that the performed the best on the low income countries. However, this is not a definite or fixed statement that is true for all time, as it is possible the low income countries evolved in an easy-to-predict pattern for the last few years, but this can always change. Looking at the plots on the previous page, we see that the models overpredicted for the high income countries and underpredicted for the low income countries for that past decade or so. There was an unexpected amount of growth the low income countries, while high income countries tapered off in wealth on average for the last decade.

These errors can probably be mitigated updating the model each year, which is more realistic, as nations actually do get updated data each year or even more frequently. We follow this thought process in the next section where we show how a country might actually use this information for development and growth.

# 7 How Can the Results Be Used? A Case Study

Here we will look at a specific nation and show how the analyses above can be used for its in-

tended purpose: to find ways a country can make better decisions for economic growth. In this section, the model was retrained for country-specific data. The last three plots above show the results of the second model (time series analysis) for Germany. Germany was chosen because of its unique trajectory and interesting changes over time. As we can see here, the model does poorly due to a lack of constant updates (**this will be revisited in section 8** where the practical use of the model is discussed). We can also see that most of the predictions underpredict, as did the best as returned from the validation set. When the sparse regression model, which does not use data that changes over time, was ran, we got that the highest weights were also for communication exports and lowest were for 'Secure Internet servers (per 1 million people)'. We realize when actually making budgeting decisions, it is necessary to observe all of the highest and lowest features and *use knowledge related to the economy of the specific country or group of countries to make educated decisions* instead of only using the qualitative results, which do not distinguish between variables that are easy to manipulate versus those that are not.

# 8 Discussion & Future Work

The contributions of this project categorized as identifying drivers of growth for groups of nations and modeling the change in growth for groups of nations. The most valuable contribution would be that it could be used for future analysis for a specific country. As in the section 7, which showed the results for the nation Germany, a specific nation can be chosen and the methods used can be repeated for economic analysis. Then a nation might be able to find that changing certain variables can be profitable, while also modeling how these changes will effect its future. In addition, just as specific nations can be inputted to be analyzed, the feature being predicted can also be switched out. Other variables that might be useful at any one time, like fertility rate, life span, etc., can be easily swapped out using these methods. The models can be used to predict and model economic development for the future.

## 8.1 Flaws & Difficulties

One of the largest difficulty in analyzing this dataset was in cleaning the data and selecting features. The dataset is multidimensional, as was described above, which caused greater complexity. While this was also beneficial, as the dataset could be analyzed from multiple perspectives, it also introduces more room for error. Also, a large difficulty was in the missingness of the data. With such a large number of variables and features, there was more data to keep track of and much of it was missing. The data was missing due to the nature of the data: it is difficult to collect certain statistics like number of people with access to internet per hundred thousand. Statistics like these can generally be backfilled, however, as is usually the case with time series analyses.

One possible cause for error is in the choice of features. Features were chosen based on missingness and the missingness was not relevantly informative for this project, but it is possible that some of the features that had large amounts of missing data could have been more helpful in our analysis.

Lastly, the models are analyses are limited in that the features that might seem important as determined by the modeled may not be features that a nation directly has quick control of. For example, the model may return "total area" as a feature that can be changed for more growth. A country cannot (easily) gain land. This can be corrected by selecting features that normalize such statistics, such as "total rural area as percent of total area," because a country can more easily control the amount of rural vs urban land as opposed to total land. This was not an issue in the models above as we selected normalized variables, but this may be an issue in the future when more features are taken into account.

## 8.2 Fairness and Weapons of Math Destruction (WMD)

Although we are do not have data on specific individuals and are not creating conclusions for these individuals, we have considered the consequences of our model on the fairness and equality between nations. We believe that because our model can only be used for the economic growth and modeling of a nation, there is less opportunity for the model itself to create inequality. However, we realize that the analyses may only be accessed to nations or people that have more resources, which may further increase inequality. From another perspective, the predictions of these models will not have negative consequences as they will not be used to discriminate or make any decisions that are meant to harm others, but

only to create wealth within a nation.

## 8.3 Future Work

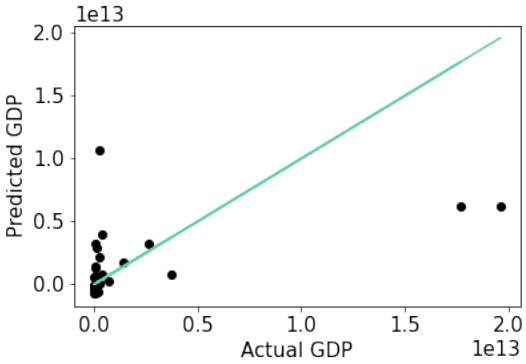
There are many opportunities for improvement which can be done in the future. First, a better method for selecting features could be used, or even better, most or all of the features could be used with proper data imputation methods. Unsupervised learning, in fact, is well suited for our dataset and question and would be a good option for the future. Clustering and matrix completion could both be informative for answering our question.

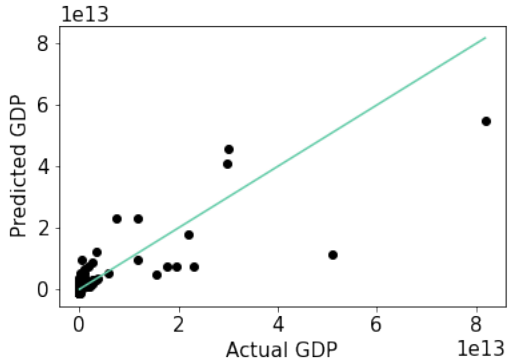
Due to using data that has one dimension being time, we could also view this problem from the perspective of dynamics. Methods such as the "sparse identification of nonlinear dynamics," or SINDy, perform sparse regression on time series data thought to come from dynamical systems. Viewing the economy as a dynamical systems could provide new information, as this method also takes into account the variety of features that may exist in the system.

Lastly, if these models are actually used, they would need to be constantly updated. The features that help a country grow depends not only on the current development of a nation, but on the current time period. As seen in the time series plots above, it is crucial to constantly update the model and make only short term predictions instead of long term predictions, as there is a nonstationary property to the data. This might mean the model would be applied as a sliding window. It would be logical to constantly update the model even if the data was not nonstationary, but **it is also important to consider the limitations in how long into the future the time series model can predict. The initial regression model, which did not depend on time and only found variables that contributed to the GDP with sparseness**, does not necessitate a sliding window or constant re-updates, although it is still more useful to constantly improve and update the model.

# 9 Appendix

Plots as mentioned in section 4:




# References

[1] Madeleine Udell (2021), *Feature Engineering*, ORIE 4741 Lecture

[2] World Bank Data Help Desk (2022) *World Bank Country and Lending Groups*, The World Bank Group