Beyond GDP: Analyzing Critical Factors Shaping Economic Growth in G20 Countries

Group E3

Xumo Zhu, Malea Van Brocklin, Christine Wu, Ivor Myers, Viveka Dhanda

Department of Mathematics, UC San Diego

Abstract

Understanding the factors driving per capita GDP growth is crucial for promoting sustainable economic development, particularly within the globally influential G20 countries. Our research investigates the determinants that significantly impact economic growth across these major economies, utilizing a comprehensive dataset sourced primarily from Kaggle and compiled from reputable World Bank indicators. By employing robust statistical methods, including multiple linear regression, stepwise regression, and time-series analysis, we examine how variables such as education, healthcare access, trade openness, technological investment, and financial stability contribute to differences in economic performance. Our analysis aims to identify the most impactful indicators influencing economic growth, offering valuable insights into the unique drivers in developed versus emerging economies. The findings of this research are expected to aid policymakers by highlighting strategic areas for targeted investment and policy refinement, facilitating informed decisions that can foster long-term economic prosperity. Furthermore, this study contributes to a deeper, evidence-based understanding of economic dynamics, equipping governments and researchers with practical, data-driven tools to navigate the challenges and opportunities of an increasingly complex global economy.

Keywords: GDP Growth; Economic Development; G20 Countries; Regression Analysis; Time-Series Analysis; Data-Driven Economic Policy

Author roles:
Abstract– Xumo
Problem– Xumo
Previous Analysis– Xumo
Data Source– Malea
Description of Data– Malea
Exploratory Data Analysis– Ivor, Christine, Malea, Xumo
Regression Analysis– Ivor, Christine, Viveka
Time Series Analysis– Ivor
Conclusion–Malea & Christine

1. Problem

Background: Economic growth remains one of the most critical challenges and priorities for nations worldwide, profoundly affecting employment opportunities, individual prosperity, infrastructure development, and overall quality of life. The G20 countries, representing the world's largest and most influential economies, have economic trajectories with significant global implications, prompting continuous exploration by policymakers and economists into factors such as education, healthcare investment, technological advancements, financial policies, and openness to international trade.

Despite extensive research, substantial debate persists regarding the precise drivers of GDP growth. Clarifying these influential factors is crucial for formulating policies that foster sustainable development, especially in our rapidly globalizing world characterized by technological disruption and unexpected global events such as pandemics and climate change. Leveraging statistical analysis and machine learning methods, including regression analysis and time-series forecasting, we can move beyond traditional economic assumptions, delivering empirical, actionable insights into determinants of economic growth.

Research Topic: This project aims to identify and quantify the factors that significantly influence per capita GDP growth in the G20 countries. Utilizing a robust dataset acquired from Kaggle, sourced from World Bank indicators, we will examine variables related to education, healthcare, trade openness, technology investment, and financial stability to determine their impact on GDP growth rates. Our approach includes using multiple linear and stepwise regression analyses to pinpoint key drivers. Additionally, we will capture the dynamics and changes in these relationships over time by incorporating time-series analysis. Ultimately, this comprehensive analytical approach will enable us to build predictive models, enhancing the ability of policymakers and researchers to anticipate future economic trends more accurately.

2. Previous Analysis

Our project proposes an innovative statistical approach to evaluating economic growth determinants among G20 countries. Inspired by existing economic analyses such as those by Jorgenson and Vu on global productivity trends, our methodology diverges by introducing a distinct regression-based statistical framework combined with advanced time-series analysis techniques. Unlike previous research that primarily emphasizes productivity through Total Factor Productivity (TFP), capital, and labor, we incorporate a broader set of socio-economic variables and leverage stepwise regression methods for variable selection to identify influential predictors uniquely tailored to the contemporary G20 context.

Additionally, this project significantly departs from earlier studies by explicitly focusing on nuanced socio-economic indicators including governmental expenditure on education, healthcare infrastructure accessibility, technological advancement initiatives, and measures of international trade openness. While Jorgenson and Vu broadly categorized global economic patterns, our project takes this further by specifically targeting regression-based modeling with stepwise variable selection, combined with robust time-series analyses. This combination will enable us to discern not only static relationships but also dynamic trends over time. The methodological rigor provided by stepwise regression allows us to systematically narrow down influential factors, improving interpretability and policy relevance of our

findings. This paper tries to fill in the respective literature by pointing out which particular, actionable variables are most crucial for sustainable economic growth in G20 countries, enriching the policymakers' toolkit with practical insights and empirical evidence.

3. Data

Data Source: The dataset used for analysis was created by Svetlana Kalacheva and sourced from Kaggle. Titled “G20 Countries Development Indicators”, the dataset was most recently updated in January 2025 and contains information from the World Bank’s collection of development indicators, with data spanning from 2014 to 2023. The original database contains information on all recognized countries from 1960 to 2023.

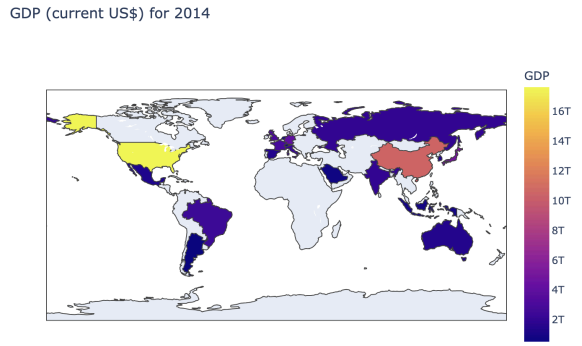
Description of Data: Information regarding the G20 countries is included in the dataset, namely: Argentina, Australia, Brazil, China, France, Germany, India, Indonesia, Italy, Japan, Korea, Mexico, Netherlands, Russian Federation, Saudi Arabia, Spain, Switzerland, Turkiye, United Kingdom, and the United States.

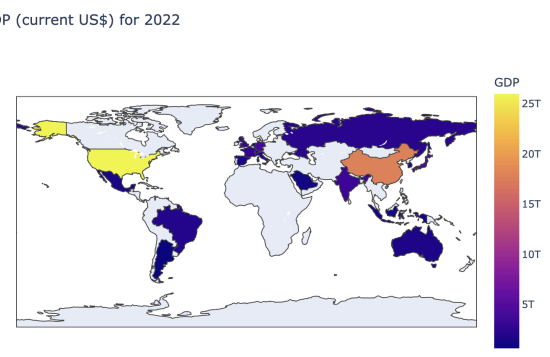
The dataset’s key features are Country Name, Country Code (a three letter code unique to each country), Series Name (the relevant development indicator), Series Code (code representation of the series), and each year ranging from 2014 to 2023.

Our focus development indicator is GDP growth, which is measured in annual percentage change. The dataset contains 52 additional indicators which primarily fall into five different categories: Economic, Health and Population, Environmental, Education, and Social and Infrastructure related indicators. Economic indicators include values related to exports and imports, inflation, merchandise trade, tax revenue, and gross national income (GNI). Health and population indicators include contraceptive prevalence, fertility rates, immunization, life expectancy, mortality rate, population density, and similar metrics. Environmental indicators include electric power consumption, energy use, and protected environmental areas. Education indicators include school enrollment for the primary and secondary levels, completion rates, and the gender parity index (GPI). Social and infrastructure indicators include migration, mobile cellular subscriptions, and military expenditure. These indicators from various sectors allow us to conduct a comprehensive analysis regarding the combined effects of country features upon GDP growth.

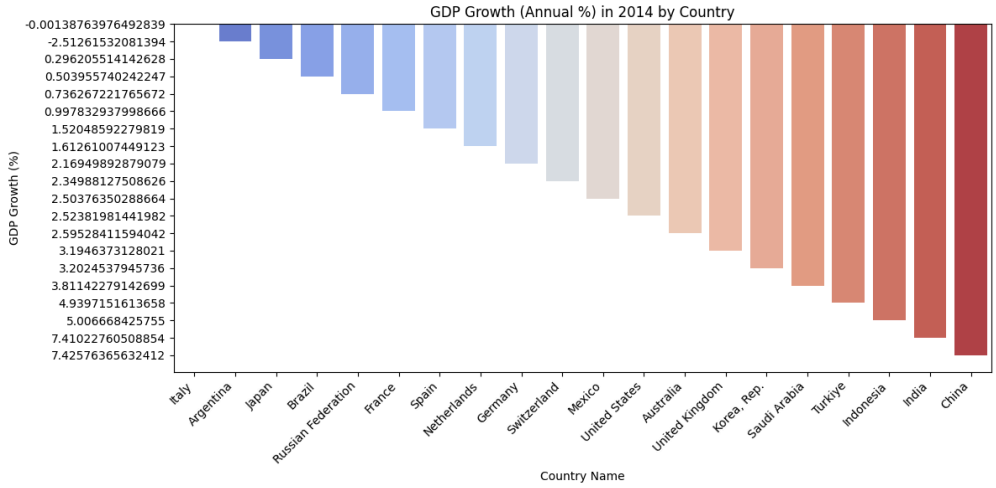
4. Exploratory Data Analysis

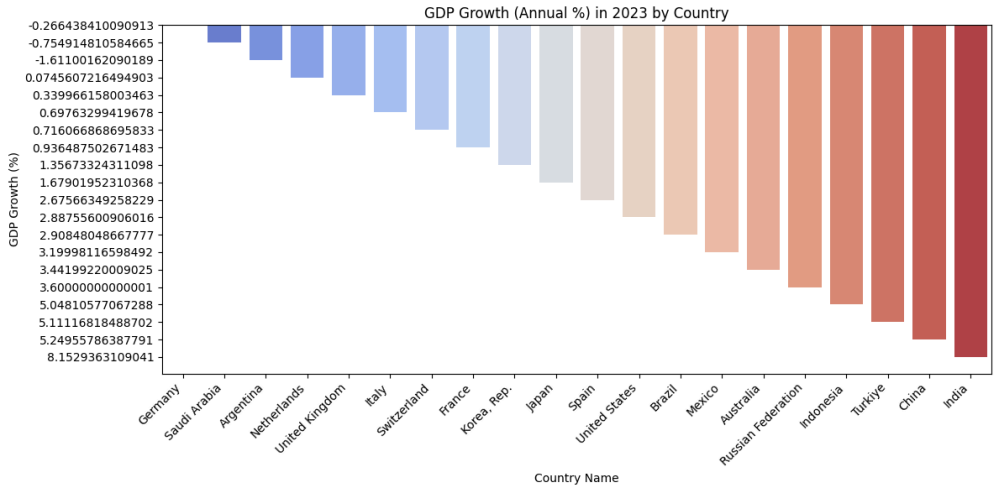
Total GDP For First and Last Date Recorded In Dataset:





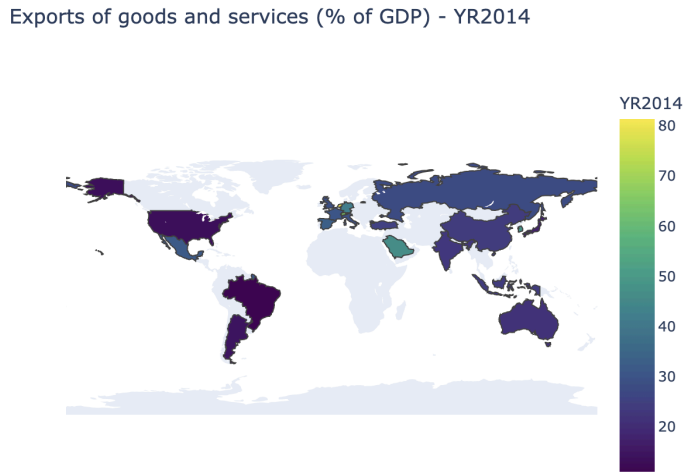
Over the time frame provided by our data, 2014 to 2022–2023 had incomplete GDP reportings– the United States maintained having the largest GDP of any country, and rose by an astounding 47.698144% in the 9-year time frame. However, it was India with the largest overall growth, rising an astounding 64.456228%. Negative growth rates were found in Brazil at -20.525684% and Japan at -13.081159%. The analysis shows that while the US retained absolute economic supremacy, India had the strongest relative growth rate among the examined countries, while Brazil and Japan highlighted some leading indicators associated with negative GDP growth.

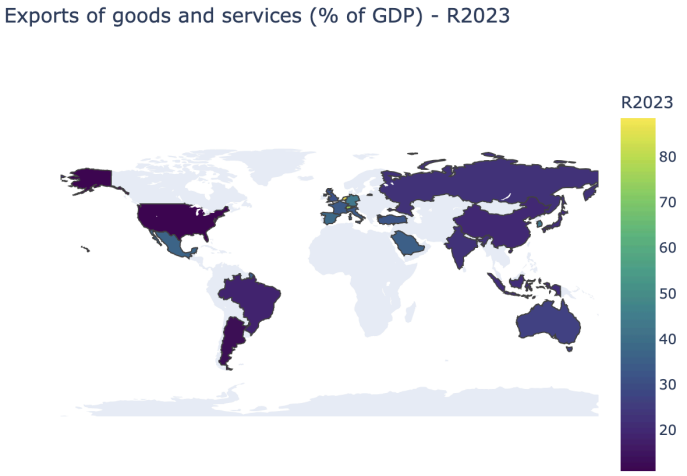




GDP Growth Rates (2014 - 2023): Initially, we were interested in observing the changes to GDP growth percentage by country across the timeframe documented in the dataset, and we found out that doing a bar plot of each country's annual GDP growth percentage for the years 2014 and 2023 would be descriptive. The bar plots show that among the twenty countries under consideration, the highest growth rates for both years were held by India, China, Turkey, and Indonesia, whereas the United States had little variation between the two years and sat close to the median for all countries. Argentina is also an interesting case of negative GDP growth for both years.

Another interesting fact was that Italy, which had negative growth in 2014, now has positive growth in 2023, which somehow establishes a connection between a few development indicators positively influencing GDP growth. A big difference for Russia moved from the bottom quartile in 2014 to the top by 2023. However, it fell by a considerable degree in the UK, which was among the high-performing countries in 2014 and has fallen to the bottom quartile by 2023. This pointed out some key changes that had happened to the countries, which allowed us to move toward finding the pivotal development indicators that influence GDP growth.





Exports of Goods and Services (2014 - 2023): Our analysis aimed to assess the correlation between various economic indicators and GDP, with a specific focus on the exports of goods and services across countries. The annual maps clearly illustrate that Western Europe consistently stands out as one of the most export-dependent regions globally in both examined years. High export dependence is indicated through greater proportions of GDP coming from exports.

In contrast, the United States, China, Brazil, and Argentina show lower export percentages relative to their GDP, suggesting higher self-reliance and a less export-driven economy. And our comparison of annual data reveals shifts in export reliance. Specifically, we observe sustained high export levels in Western Europe, Germany in particular, whereas other regions exhibit a noticeable decline. This decrease in export reliance could reflect escalating trade tensions or evolving economic dynamics that drive nations toward increased domestic economic activities. Such shifts might also indicate changing trade relationships and emerging trade barriers. Overall, the variation in export dependence highlights both regional economic strategies and the broader impact of international relations on trade behaviors.

Handling missing values:

In exploring our data, we eventually identified that the year 2023 had over half of its observations as missing data. Due to this severity, we decided to drop any data in that year since the missing data pattern suggested a systemic issue with data collection or recording during that period, and the amount of missing data was too substantial to allow for any meaningful statistical analysis or imputation.

```
#Count missing data
missing_counts = df_final.isna().sum()
missing_counts = missing_counts[missing_counts > 0]
if not missing_counts.empty:
    print(missing_counts)
```

```
Series Name
Agriculture, forestry, and fishing, value added (% of GDP)                    1
Annual freshwater withdrawals, total (% of internal resources)               20
Births attended by skilled health staff (% of total)                         88
Contraceptive prevalence, any method (% of married women ages 15-49)        154
Domestic credit provided by financial sector (% of GDP)                     127
Electric power consumption (kWh per capita)                                 155
Energy use (kg of oil equivalent per capita)                                147
External debt stocks, total (DOD, current US$)                              117
High-technology exports (% of manufactured exports)                           2
Income share held by lowest 20%                                              48
Industry (including construction), value added (% of GDP)                     1
Mobile cellular subscriptions (per 100 people)                                1
Net barter terms of trade index (2015 = 100)                                 28
Net official development assistance and official aid received (current US$)  117
Poverty headcount ratio at $2.15 a day (2017 PPP) (% of population)          48
Poverty headcount ratio at national poverty lines (% of population)         106
Prevalence of HIV, total (% of population ages 15-49)                        59
Prevalence of underweight, weight for age (% of children under 5)           153
Primary completion rate, total (% of relevant age group)                     67
Revenue, excluding grants (% of GDP)                                         12
School enrollment, primary (% gross)                                          1
School enrollment, primary and secondary (gross), gender parity index (GPI)  46
School enrollment, secondary (% gross)                                       13
Tax revenue (% of GDP)                                                       12
Terrestrial and marine protected areas (% of total territorial area)         40
Time required to start a business (days)                                     60
Total debt service (% of exports of goods, services and primary income)     117
dtype: int64
```

Remaining columns were examined, and since our data is split among 20 different countries and 9 years, it was decided to drop any features that had above 10% missing data, in order to ensure robustness and reliability of our analysis across all countries and years. Additionally, two remaining columns, ‘Revenue, excluding grants (% of GDP)’ and ‘School enrollment, secondary (% gross)’ were dropped since they were missing data across all years for at least one country. Remaining missing data was handled using a forward fill method. We choose a forward fill, as we deemed an annual change of 0% from the year before to be the least significant impact on our models.

5. Data Analysis

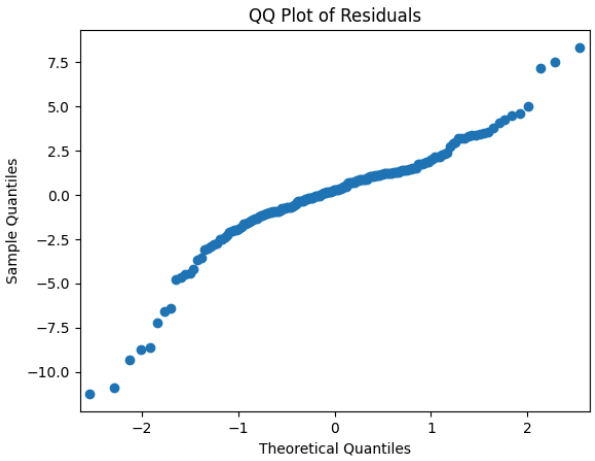
Multiple Linear Regression

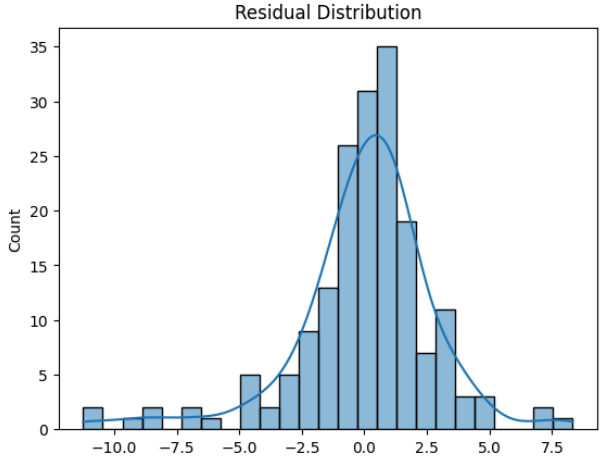
To examine the key determinants of GDP growth in G20 countries, we constructed a Multiple Linear Regression (MLR) model using economic, social, and infrastructural indicators. The dataset contained missing values, which were handled using mean imputation for numerical variables and columns with >30% missing values were dropped to preserve data consistency and important information.

Initial Model: We initially included all independent variables, excluding categorical identifiers such as Country Name, Country Code, and Year to avoid redundancy. Our dependent variable was GDP growth (annual %). We ran an OLS regression on the full dataset to assess initial model performance and variable significance and used evaluation metrics such as R^2, Adjusted R^2, and the F-statistic.

This model yielded R^2 of 0.406, meaning 40.6% of the variation in GDP growth was explained by the selected features. The Adjusted R^2 (0.235) was significantly lower, indicating that some predictors are not significant contributors to the model. The F-statistic (2.373) depicts the model's overall statistical significance, but several individual predictors had high p-values (> 0.05), indicating weak explanatory power.

The Residual Distribution plot showed a slight right skew, suggesting that the residuals were not perfectly normally distributed, potentially leading to biased standard errors. The Q-Q Plot of Residuals showed deviations at both ends, implying non-normal residuals and potential outliers.





Refining Model: From here, we proceeded to Variance Inflation Factor (VIF) analysis to check for multicollinearity. We found some extreme outliers, namely columns suggesting that predictors were highly correlated. Variables with VIF >10 were dropped to improve our model.

We ran a second regression with the modified features. We observed that R^2 decreased to 0.242, and Adjusted R^2 slightly dropped to 0.211, confirming that some removed variables had genuine explanatory power. However, the model now had more reliable coefficient estimates, making interpretation clearer.

At this point, we manually reintroduced GDP, GNI per capita, and Exports of goods and services due to their economic significance and re-ran the model. This version produced an R^2 of 0.353 and an Adjusted R^2 of 0.228.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     GDP growth (annual %)   R-squared:                       0.353
Model:                               OLS   Adj. R-squared:                  0.228
Method:                    Least Squares   F-statistic:                     2.819
Date:                   Mon, 17 Mar 2025   Prob (F-statistic):           2.34e-05
Time:                           05:11:43   Log-Likelihood:                -448.30
No. Observations:                    180   AIC:                             956.6
Df Residuals:                        150   BIC:                             1052.
Df Model:                             29
Covariance Type:               nonrobust
==============================================================================
```

```
==============================================================================
Omnibus:                       47.954   Durbin-Watson:                   2.477
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              128.758
Skew:                          -1.104   Prob(JB):                     1.10e-28
Kurtosis:                       6.506   Cond. No.                     1.21e+16
==============================================================================
```

The Residual plot showed a more centered distribution appeared but heteroscedasticity remained. The QQ plot showed slight improvement but deviation at both ends was also not eliminated completely (normality issue).

Overall, the model provides some insights into GDP growth drivers (eg: Gross Capital Formation (% of GDP) but lacks strong predictive power (Adjusted $R^2 = 0.228$). This is likely due to economic complexities not captured by simple linear regression. The presence of heteroscedasticity and non-normal residuals further limits model reliability. Given these limitations, stepwise regression or alternative modeling techniques may improve predictive accuracy and interpretability.

Stepwise Regression

To investigate the key determinants of GDP growth across countries, we employed stepwise regression techniques (forward and backward selection) using economic, demographic, and infrastructural indicators. The response variable was GDP_growth (annual %). Stepwise regression helps identify the most statistically relevant predictors while avoiding overfitting from including irrelevant variables.

Data Preparation:

The dataset initially contained 180 observations and 35 columns. Missing values were minimal, except for Tax_revenue, which had 81 missing entries (45%) due to being duplicated in the dataset. This redundancy likely arose during data merging. To ensure data integrity, mean imputation was considered for variables with <10% missing data, but ultimately, rows with missing GDP_growth were dropped, resulting in 99 valid observations for analysis.

Statistical Analysis:

In the **Forward Stepwise Regression** approach, we employed the Akaike Information Criterion (AIC) to guide variable selection, aiming to strike a balance between model fit and complexity. This method identified five variables as key predictors of GDP growth: Net_migration, Inflation__GDP_deflator, Population__total, Forest_area, and Tax_revenue. The resulting model explained approximately 14.6% of the variance in GDP growth ($R^2 = 0.146$, Adjusted $R^2 = 0.100$), with an AIC of 512.8, suggesting a modest model fit. Among the predictors, Net_migration ($p = 0.008$), Inflation__GDP_deflator ($p = 0.003$), and Tax_revenue[1] ($p = 0.005$) were statistically significant. However, a severe multicollinearity issue was flagged by condition number (1.52e-28), likely due to the presence of a duplicated Tax_revenue column. Additionally, residual diagnostics including the Omnibus and Jarque-Bera tests indicated non-normality of residuals ($p < 0.001$). These diagnostics highlight limitations in coefficient reliability and suggest that while the model surfaces potentially meaningful variables like inflation and migration, its overall explanatory power remains weak and susceptible to distortion from multicollinearity.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             GDP_growth   R-squared:                       0.146
Model:                            OLS   Adj. R-squared:                  0.100
Method:                 Least Squares   F-statistic:                     3.189
Date:                Mon, 17 Mar 2025   Prob (F-statistic):             0.0106
Time:                        02:58:33   Log-Likelihood:                -250.38
No. Observations:                  99   AIC:                             512.8
Df Residuals:                      93   BIC:                             528.3
Df Model:                           5
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                    0.0026      0.001      2.875      0.005       0.001       0.004
Net_migration             2.364e-06   8.67e-07      2.727      0.008    6.43e-07    4.09e-06
Inflation__GDP_deflator      0.0905      0.030      3.045      0.003       0.031       0.149
Population__total        -8.588e-09   5.49e-09     -1.564      0.121   -1.95e-08    2.32e-09
Forest_area              -2.474e-07   1.62e-07     -1.529      0.130   -5.69e-07    7.39e-08
Tax_revenue[0]              -0.0430      0.053     -0.810      0.420      -0.148       0.062
Tax_revenue[1]               0.0852      0.030      2.875      0.005       0.026       0.144
==============================================================================
Omnibus:                       47.948   Durbin-Watson:                   2.455
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              179.826
Skew:                          -1.588   Prob(JB):                     8.94e-40
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.52e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

For the **Backward Stepwise Regression**, we adopted the Bayesian Information Criterion (BIC), which penalizes model complexity more heavily than AIC, thus promoting parsimony. Surprisingly, the backward selection process retained a large set of 30 variables, including comprehensive economic indicators such as GNI, GNI per capita, Exports, and Imports, as well as demographic variables like Fertility rate, Mortality rate, and Population growth, and categorical variables including Country_Name and Year.

```
Optimal Formula Using Backward Selection: GDP_growth ~ Net_migration + Year +
Life_expectancy_at_birth__total + GNI__PPP + School_enrollment__primary +
Gross_capital_formation + GNI_per_capita__PPP + Tax_revenue + Urban_population_growth
+ Foreign_direct_investment__net_inflows + Population_growth +
Exports_of_goods_and_services + GNI__Atlas_method + Adolescent_fertility_rate +
Country_Name + Inflation__GDP_deflator +
Agriculture__forestry__and_fishing__value_added + Population__total +
Mortality_rate__under_5 + Merchandise_trade + Personal_remittances__received +
Mobile_cellular_subscriptions + GNI_per_capita__Atlas_method + Industry__value_added +
High_technology_exports + Fertility_rate__total + Immunization__measles + Forest_area
+ Mili
```

This model achieved a significantly higher R^2 of 0.527 (Adjusted R^2 = 0.356), explaining over half the variation in GDP growth. The model's AIC was 496.3 and BIC 566.4, both improvements over the forward model. Several predictors emerged as statistically significant, including Gross_capital_formation (p = 0.020), Personal_remittances__received (p = 0.029), Fertility_rate__total (p = 0.025), and Population_density (p = 0.010). Nonetheless, this model also suffered from strong multicollinearity (condition number = 1.66e+16) and non-normal residuals, raising concerns about overfitting and stability of the estimates. Furthermore, the inclusion of Country_Name as a categorical variable introduces interpretation challenges due to dummy coding of individual countries, complicating the analysis of specific national effects.

OLS Regression Results

Dep. Variable:	GDP_growth	R-squared:	0.527
Model:	OLS	Adj. R-squared:	0.356
Method:	Least Squares	F-statistic:	3.084
Date:	Mon, 17 Mar 2025	Prob (F-statistic):	8.79e-05
Time:	02:59:38	Log-Likelihood:	-221.17
No. Observations:	99	AIC:	496.3
Df Residuals:	72	BIC:	566.4
Df Model:	26		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.7169	1.861	-1.460	0.149	-6.428	0.994
Country_Name[T.Australia]	-0.0748	0.038	-1.968	0.053	-0.150	0.001
Country_Name[T.Brazil]	0.0036	0.009	0.420	0.676	-0.014	0.021
Country_Name[T.China]	0.0016	0.001	2.811	0.006	0.000	0.003
Country_Name[T.France]	-0.0002	0.000	-0.454	0.651	-0.001	0.001
Country_Name[T.Germany]	7.679e-08	3.21e-07	0.239	0.812	-5.63e-07	7.16e-07
Country_Name[T.India]	5.605e-09	1.27e-08	0.440	0.661	-1.98e-08	3.1e-08
Country_Name[T.Indonesia]	-2.28e-08	1.47e-08	-1.551	0.125	-5.21e-08	6.5e-09
Country_Name[T.Italy]	-1.187e-09	6.9e-10	-1.721	0.090	-2.56e-09	1.88e-10
Country_Name[T.Japan]	-7.3138	5.900	-1.240	0.219	-19.074	4.447
Country_Name[T.Korea, Rep.]	22.2954	7.824	2.850	0.006	6.699	37.892

...

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 2.57e+16. This might indicate that there are strong multicollinearity or other numerical problems.

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Lastly, we developed a **Custom Model** guided by theoretical relevance rather than statistical selection. This model incorporated 10 features commonly associated with economic growth, including fertility and mortality rates, tax revenue, gross capital formation, military expenditure, population density, GNI (Atlas method), and immunization rates. However, the model performed poorly, yielding an R^2 of 0.110 and an Adjusted R^2 of just 0.008, meaning it explained only 11% of the variance in GDP growth, with virtually no adjustment for model complexity. The AIC of 526.9 also indicated worse fit than both the forward and backward models. None of the predictors were statistically significant (all p-values > 0.1), suggesting that, in isolation, these features did not meaningfully explain variations in GDP growth across countries. These findings imply that while domain knowledge is valuable, data-driven methods may uncover more effective combinations of predictors, though they are not without their own limitations such as overfitting and multicollinearity.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             GDP_growth   R-squared:                       0.110
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     1.084
Date:                Mon, 17 Mar 2025   Prob (F-statistic):              0.383
Time:                        03:00:41   Log-Likelihood:                -252.47
No. Observations:                  99   AIC:                             526.9
Df Residuals:                      88   BIC:                             555.5
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                     -0.0065      0.009     -0.691      0.491      -0.025       0.012
Adolescent_fertility_rate     -0.0225      0.030     -0.759      0.450      -0.082       0.037
Fertility_rate__total          1.9805      1.613      1.228      0.223      -1.225       5.186
Inflation__GDP_deflator        0.0545      0.037      1.478      0.143      -0.019       0.128
Surface_area              -5.516e-08   9.62e-08     -0.574      0.568   -2.46e-07    1.36e-07
Tax_revenue[0]                -0.0059      0.088     -0.067      0.946      -0.181       0.169
Tax_revenue[1]                -0.2147      0.311     -0.691      0.491      -0.832       0.403
Military_expenditure          -0.1816      0.261     -0.697      0.488      -0.700       0.336
Gross_capital_formation        0.1107      0.112      0.984      0.328      -0.113       0.334
Population_density             0.0002      0.004      0.052      0.958      -0.007       0.007
GNI__Atlas_method           6.19e-14   6.87e-14      0.901      0.370   -7.47e-14    1.98e-13
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.06e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

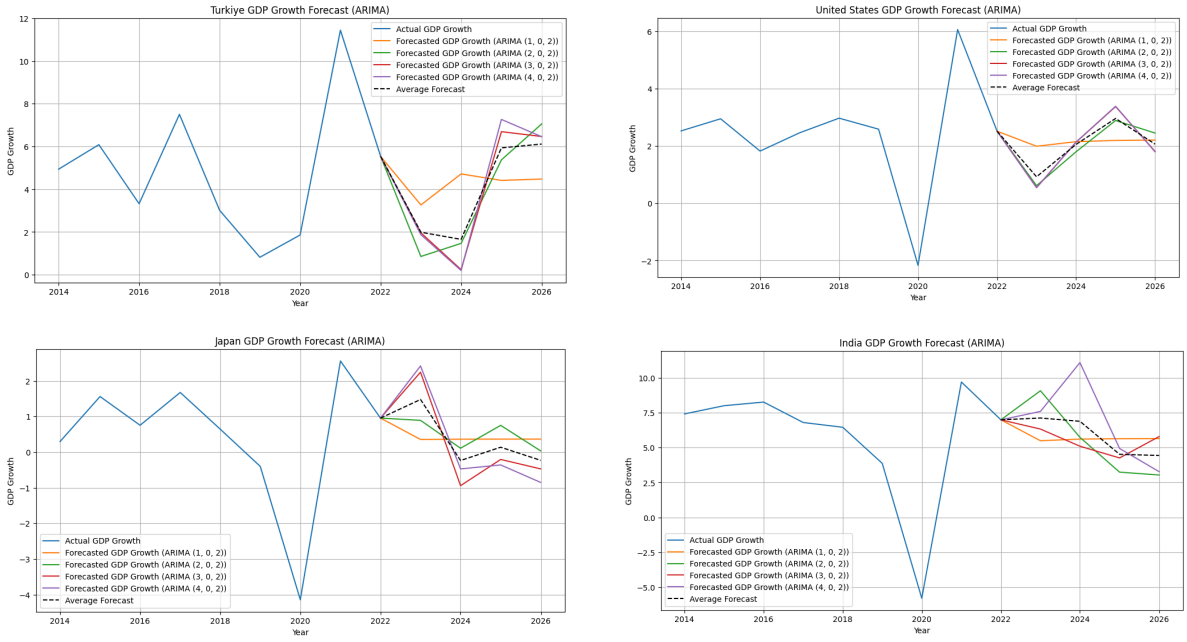
6. Forecasting GDP Growth

A key idea of our project was to forecast GDP growth for future years with a time series analysis using the data provided. We decided that using an ARIMA- autoregressive integrated moving average - model could be a good fit for our data. This is because the statistical model implies that the target variable is

regressed on previous values of the features and its variable can be explained by a linear combination of said features, which we hypothesized to be true. However, unlike with the multiple linear and stepwise regression models, the ARIMA models would be applied to each country individually, to train on and forecast their unique GDP growths.

Data preparation for the ARIMA model involved using our df_cleaned data frame, which implemented a missing data threshold of 10% and applied a forward fill on remaining missing data, as the ARIMA model cannot handle NaN values. Then a for loop was used to create a dictionary of country data frames, that only included data for their specific nation. It was decided to only use the features that the multiple linear regression and stepwise regression deemed statistically significant, so the data frame was further filtered down to those. And since we have a limited number of years of data, but want cross-validation, we set a train/test split of 90%, which equates to training on the first 7 years and testing on the 8th.

We first ran ARIMA models with the 10 statistically significant features and experimented with the autoregressive, differencing, and moving average orders. It was observed that setting the differencing order to 0–implying the time series is stationary–and putting the moving average to 2–representing how many past forecast errors are used to predict the GDP growth–led to models with the lowest relative AIC and BIC values and highest log-likelihood values. It was decided to run each country with 1 to 4 lags–four being the max as that was half of the training observations–as they were in a similar goodness of fit range, and thus the forecasts showed a spread of predictions with varying lag. The average of the forecasts was included as a dashed line as seen in the figures of forecasted GDP growth shown below.



While the plotted forecasting by the ARIMA models seems to display a capturing and application of the variance in each country's dataset, an examination of the coefficients and testing for assumptions implies the opposite.

```
Summary for Turkiye ARIMA (3, 0, 2):
                               SARIMAX Results
==============================================================================
Dep. Variable:             GDP_growth   No. Observations:                    8
Model:                 ARIMA(3, 0, 2)   Log Likelihood                -17.924
Date:                Mon, 17 Mar 2025   AIC                           49.847
Time:                        02:16:58   BIC                           50.403
Sample:                    01-01-2014   HQIC                          46.097
                         - 01-01-2021
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.4739      0.640      6.991      0.000       3.220       5.728
ar.L1         -0.5826      2.094     -0.278      0.781      -4.686       3.521
ar.L2         -0.3317      0.767     -0.432      0.665      -1.835       1.172
ar.L3         -0.7490      0.844     -0.887      0.375      -2.403       0.905
ma.L1         -0.0090    153.985  -5.83e-05      1.000    -301.814     301.796
ma.L2         -0.9910     24.390     -0.041      0.968     -48.794      46.812
sigma2         2.5024     52.693      0.047      0.962    -100.775     105.779
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 0.21
Prob(Q):                              0.94   Prob(JB):                         0.90
Heteroskedasticity (H):              11.30   Skew:                             0.08
Prob(H) (two-sided):                  0.08   Kurtosis:                         2.22
===================================================================================
```

Using Turkey as an example, the constant coefficient delineated that GDP growth does have a baseline trend, but the moving average orders MA(1), and MA(2) terms are completely insignificant with p-values incredibly close to 1. The lag variables of the model, ar.L1-3 have lower p-values but are still statistically insignificant and are more likely just modeling noise. The AIC and BIC are higher implying a poorer fit, but relative to other ARIMA models tested on Turkey data, these values are on the low end. The few positive signs were that Ljung-Box and Jarque-Bera, which are goodness of fit tests, signify that no significant autocorrelation in the residuals and they are not significantly different from a normal distribution. Additionally, the Prob(H) p-value of 0.08 implies a lack of heteroskedasticity, which is an assumption of ARIMA. But ultimately, the coefficients and p-values strongly indicate a poor fit and that the model has severely limited predictive power.

All other countries' models, across the varying autoregressive orders, had similar summary results to those shown above in Turkey, implying that overall our ARIMA model was a poor fit. Tests were performed with all features, not just those identified as statistically significant from the earlier regressions, but saw no significant improvement. The root cause of this is very likely to be the lack of observations in our time series. Our dataset only had reliable data from 2014 to 2022, which is a mere 9 years in the long development of a country. Predictive models such as these often need a much larger sample size to identify statistically significant relations and trends, and then predict on the latest year of data. If we were to create a reliable predictive ARIMA model, the first thing we would need is a dataset that covered a much longer period of time.

7. Conclusion

Our project set out to investigate the critical factors influencing GDP growth across G20 countries using a combination of multiple linear regression, stepwise regression, and time-series forecasting. Through comprehensive statistical analysis of economic, demographic, and infrastructural indicators, we aimed to

identify the most impactful variables contributing to GDP growth and evaluate their potential for predictive modeling to support informed policy decisions.

We developed three multiple linear regression models: one as a baseline and one simplified model which implemented VIF analysis to account for multicollinearity. After determining that the second model had removed critical variables, we reintroduced several to account for complexity. From this we determined that ‘Births attended by skilled staff’, ‘Foreign direct investment’, ‘Gross capital formation’, ‘High tech exports’, ‘Income share held by lowest 20%’, ‘Inflation GDP deflator’, ‘Poverty headcount ratio at national poverty lines’, and ‘Total debt service’ were all statistically significant at a significance level of 0.15. Following this, we performed stepwise regression to further determine which variables to include in our time series analysis. Using forward stepwise regression and the Akaike Information Criterion (AIC), we settled on the following key indicators: ‘Net migration’, ‘Inflation GDP deflator’, ‘Population total’, ‘Forest area’, and ‘Tax Revenue’. Utilizing a combination of the factors discovered through both regression methods, we developed our ARIMA model. The forecasting from these models was able to somewhat capture the information from our test data, but was generally not a great fit.

In relation to our initial project proposal, the final report stayed closely aligned with our original goals. However, we faced several limitations that affected the depth and accuracy of our analysis. One major challenge was the short time span of the dataset, with only nine years of reliable data (2014–2022), which made it difficult for our ARIMA models to predict long-term trends. We also had some data issues, such as missing values, which caused problems like multicollinearity and made our regression models less stable. Although we could handle these issues with imputation and careful feature selection, the small sample size still limited the strength of our results. Also, since economic growth is complex and involves many indicators, using only linear models may not fully capture its patterns. In the future, using more advanced models or larger datasets could help improve accuracy. Despite these challenges, our project provided useful insights and a solid foundation for future research on the factors influencing GDP growth.