Analysis of Societal Factors with Potential Relationships to Changes in GDP Over Time

Final Project Group 42: Ben Wu, Bradley Tian, Emily Bidle, Jack Wang, Tim Gao, Tracey Ley

Introduction

The GDP, or Gross Domestic Product, is a common benchmark utilized by media and research institutions alike for measuring the performance and outputs of a nation's economy. A higher GDP indicates more robust and plentiful growth of a country's economy. Due to GDP's significance in both analyzing the past and predicting the future, this investigation attempts to identify factors that could influence or be utilized to estimate GDP changes. More specifically, we aim to answer the following question:

What societal factors - both economic and non-economic - are proportionally related to changes in GDP over time?

Description of Data

The primary dataset utilized in this investigation contains comprehensive data on overall global development as of 2018, many of which can be used to identify factors that may contribute or relate to changes in GDP across nations worldwide. Within the dataset, some key variables stood out among the others. These factors include compulsory education, labor force size, annual population growth, age dependency, public access to electricity, measles immunization status, and life expectancy. We initially used these data to search for potential correlations between such factors and the GDP utilizing regression analysis. Subsequently, we established several hypotheses on several factors with relatively high correlations and conducted hypothesis testing to verify our hypotheses. The scope of data we employed spans between the year 2017 and 2018.

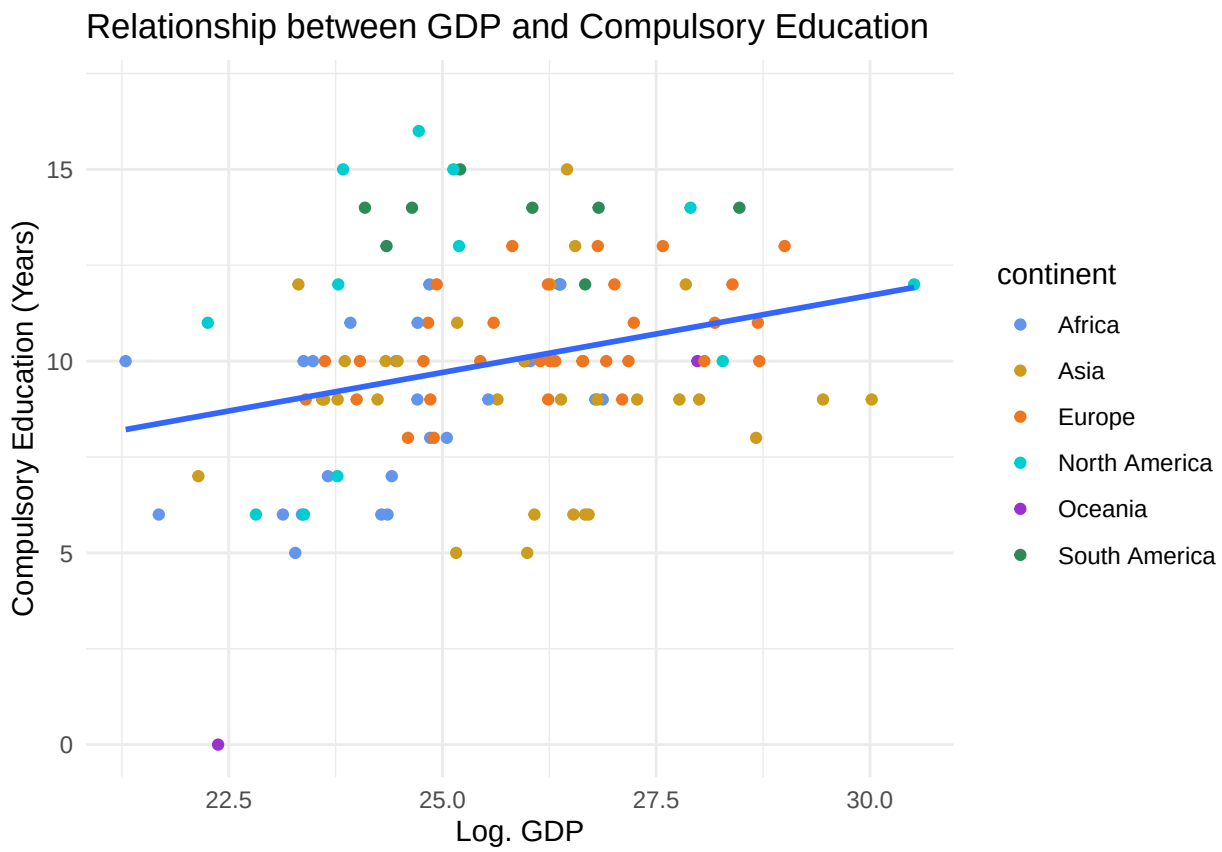
Due to natural volatility and randomness of data collected from large populations, we have determined that a correlation coefficient above 0.65 would be considered a fairly strong correlation between the two variables. A coefficient between 0.65 and 0.3 would be considered moderate, while below 0.3 would be considered weak or insignificant.

Although we are looking at the effect of societal factors on GDP, insinuating GDP as the response variable traditionally on the y-axis, we settled on log of GDP as the variable along the x-axis instead to better compare and showcase the changes in exploratory variables between each independent graph.

Upon initial testing, we have discovered that raw GDPs create too many outliers and incomparable values that severely compromise the accuracy of regression analysis. After conducting contextual research, we have decided to use the log of GDPs to better portray countries' GDPs in relation to each other. This scale change allows us to compare more countries with disparate raw GDPs without losing proportional accuracy.

Exploratory Data Analysis

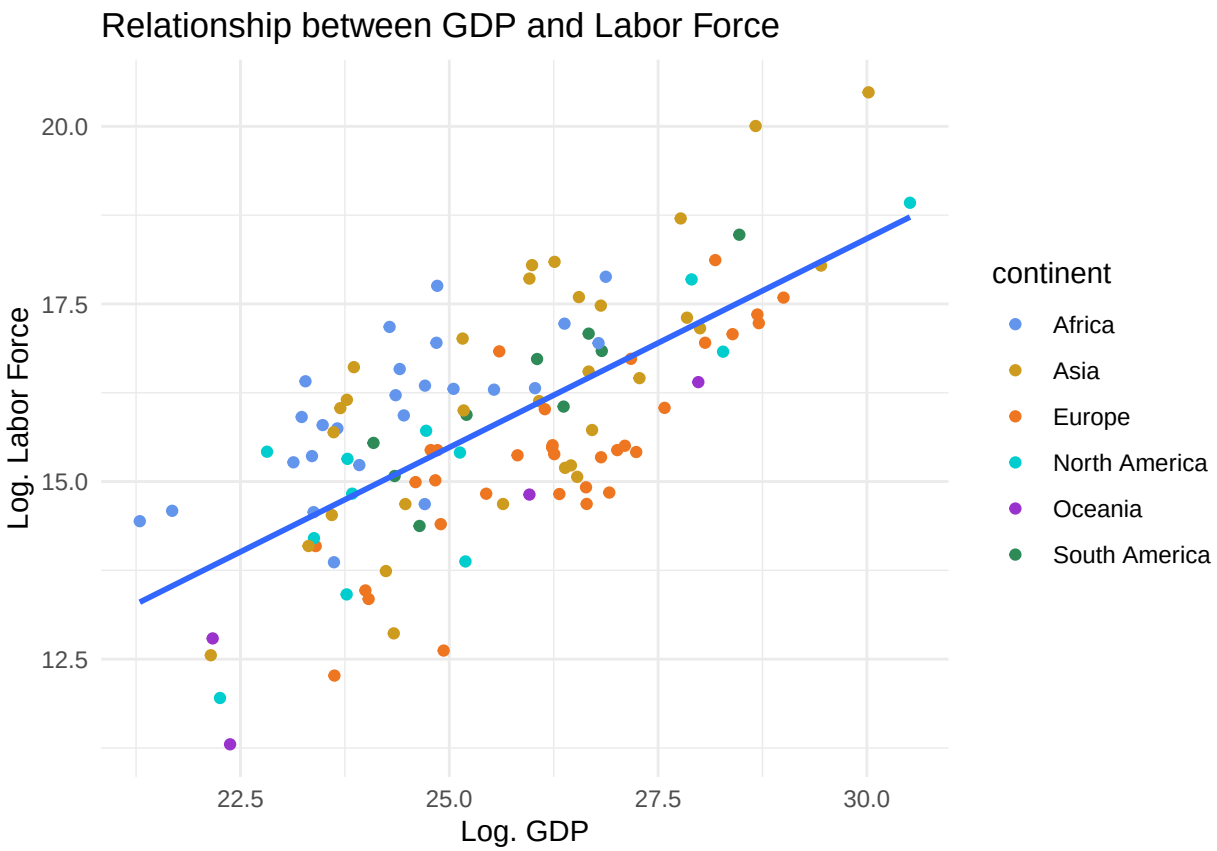
Figure 1 by Bradley Tian




```
##
##  Pearson's product-moment correlation
##
## data:  log_GDP and education
## t = 3.0275, df = 108, p-value = 0.003084
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0975660 0.4437022
## sample estimates:
##       cor
## 0.2796978
```

To analyze the relationship between nations' GDP and their compulsory education levels, we constructed a scatter plot with GDP on the x-axis and Compulsory Education, in Years, on the y-axis. The linear model regressing compulsory education on GDP reveals a weak positive correlation between the two variables. More precisely, the correlation coefficient is approximately 0.28. Furthermore, upon evaluating each continent separately, we found that no particular continent shows homoscedastic residuals nor a reliable correlation stronger than the 0.28 overall. Therefore, it is likely that GDP has no direct impact on levels of compulsory education of nations worldwide.

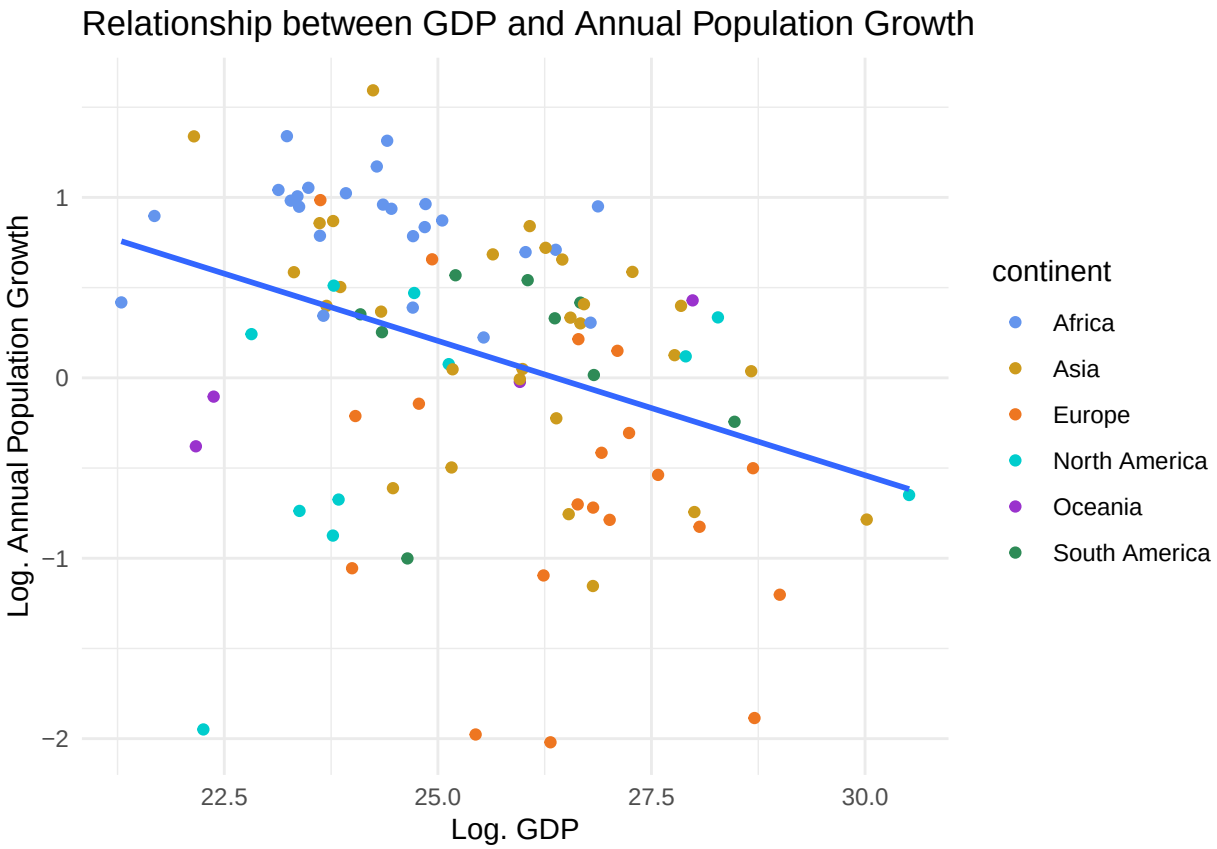
Figure 2 by Bradley Tian




```
##
##  Pearson's product-moment correlation
##
## data:  log_GDP and log_labor
## t = 10.127, df = 112, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5813661 0.7765367
## sample estimates:
##       cor
## 0.6913595
```

The scatter plot shown above visualizes the relationship between GDP and labor size of nations worldwide. The x-axis represents GDP, and the y-axis represents the size of the labor force. Regressing labor size on GDP, we have calculated a correlation coefficient of 0.69, which indicates a relatively strong, positive correlation between the two variables. We have also observed that the distribution of data is relatively elliptical, albeit with few outliers towards the lower end. According to the correlation coefficient and the homoscedasticity of the scatter plot, there is a clear association between GDP and Labor Size of nations worldwide.

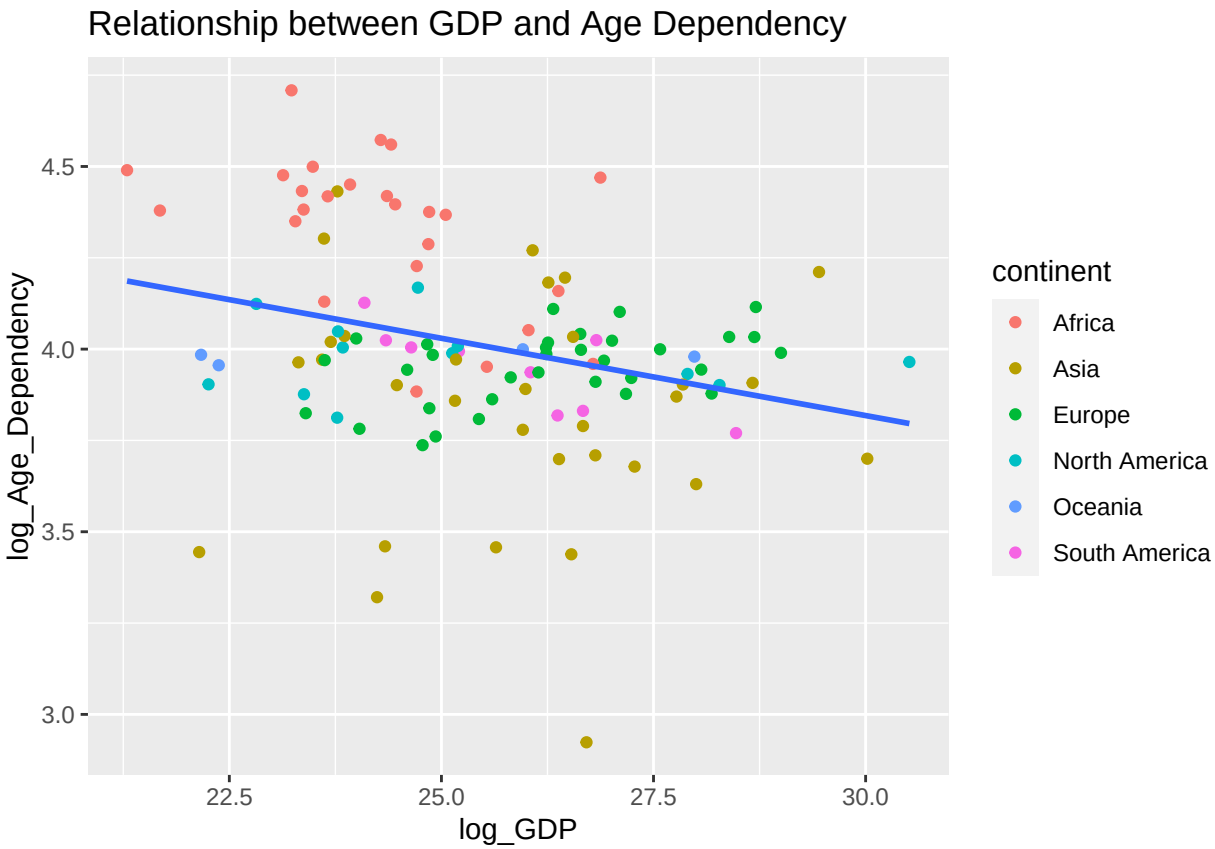
Figure 3 by Emily Bidle




```
##
##  Pearson's product-moment correlation
##
## data:  catalog$GDP and catalog$pop_ann_growth
## t = -1.8259, df = 112, p-value = 0.07053
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.34320223  0.01434431
## sample estimates:
##       cor
## -0.170019
```

We plotted a scatter plot with GDP on the x-axis and annual population growth on the y-axis to explore the relationship between the two variables. There appears to be a weak negative association between GDP and annual population growth with a correlation coefficient of -0.17 when all countries are compared on the same graph. However, after evaluating the continents separately, it can be noted that countries in Africa and Asia (excluding significant outliers), showcase a trend where countries with a lower GDP have a higher annual population growth rate. It can be hypothesized that high population growth in low-income countries may slow their development and contribute to a lower GDP.

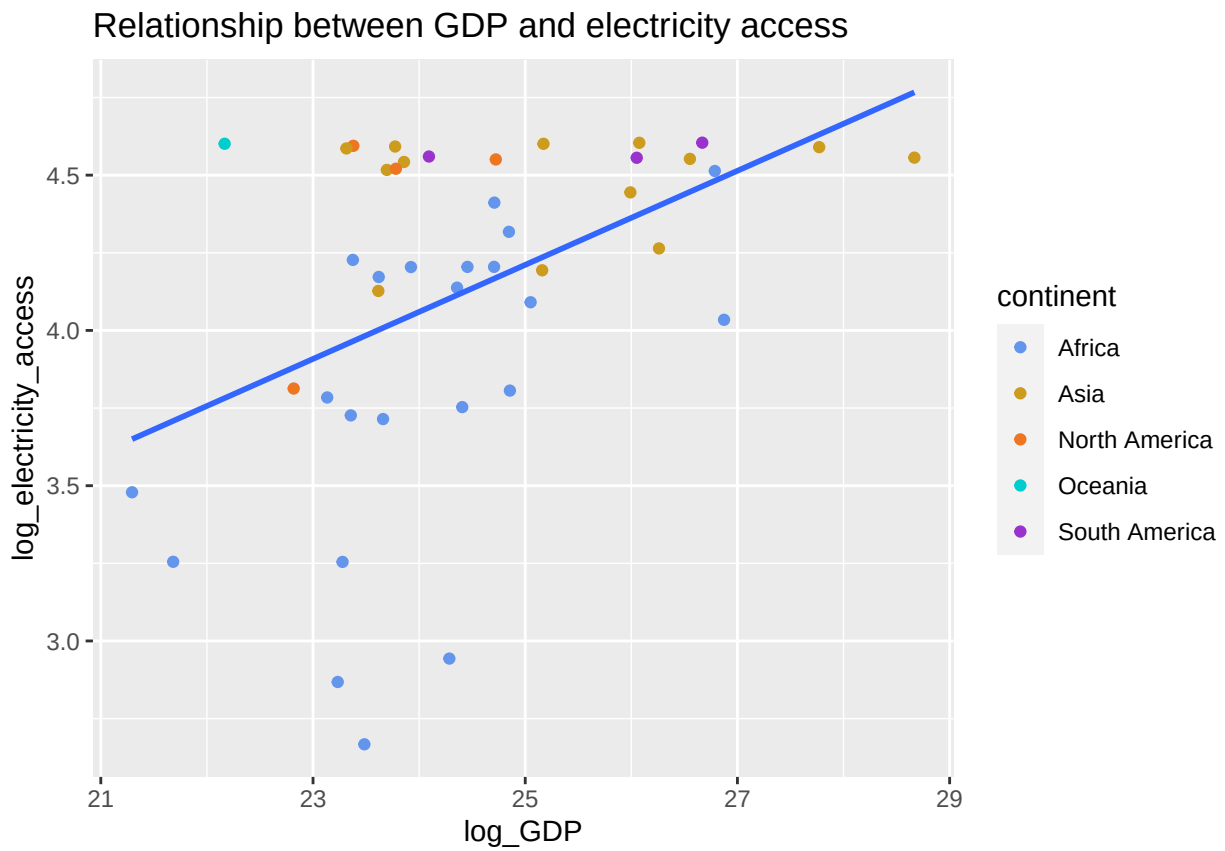
Figure 4 by Tracey Ley




```
## [1] -0.2919155
```

Mathematically, there is a weak negative correlation of -0.29 between GDP and age dependency, but, graphically, there appears to be little to no correlation. This difference could be attributed to several extreme outliers, making the association less reliable. Notably, most countries besides those from the continents of Africa and Asia seem to have stabilized at around 40 dependents per 100 working-age individuals. A reasonable balance between the two is expected in order to have a stable economy. Africa and Asia have a negative and positive slope, respectively. Changes in child birth rate as GDP increases is a possible explanatory hypothesis for these variables. For example, much of Asia's age dependency proportions are very low, and this can be attributed to their declining birth rates, however, no definitive conclusion of causality can be made.

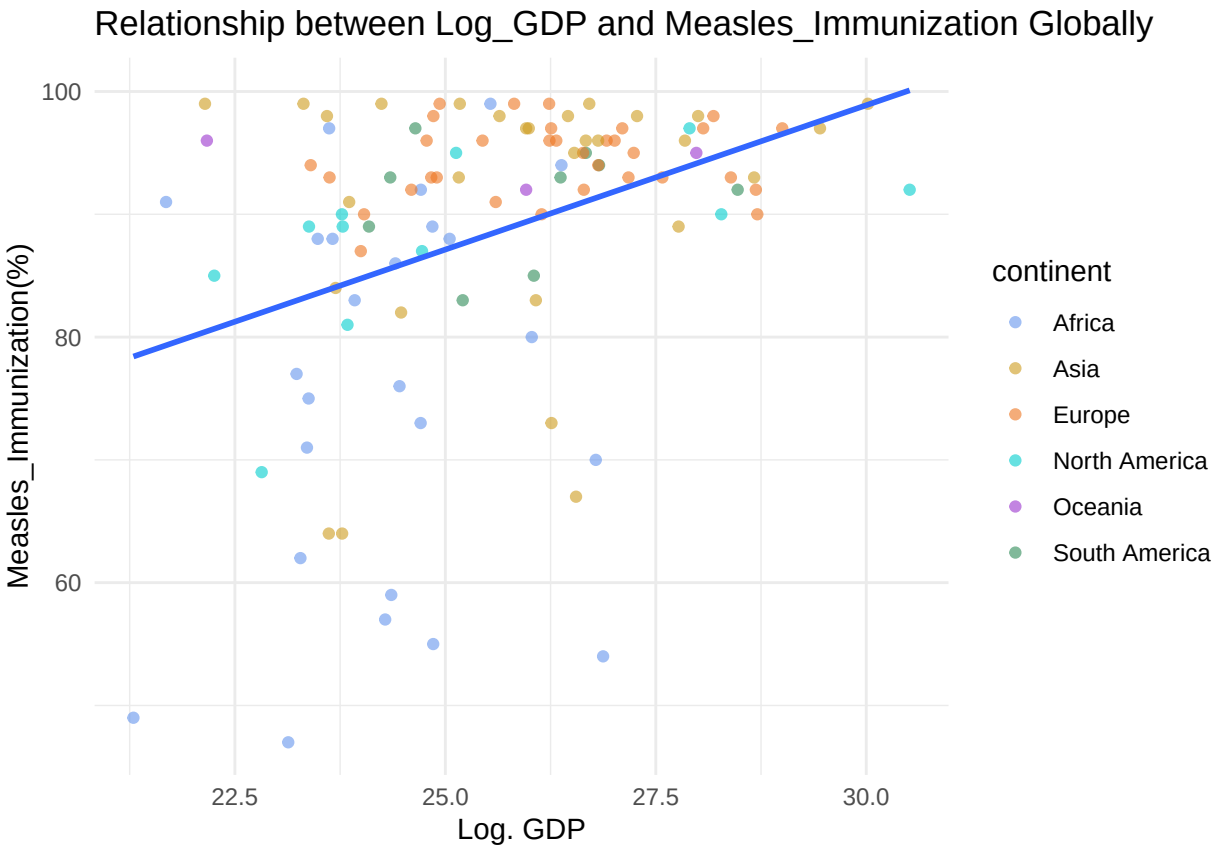
Figure 5 by Tim Gao




```
##
##  Pearson's product-moment correlation
##
## data:  log_GDP and log_electricity_access
## t = 3.2531, df = 41, p-value = 0.002288
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1766355 0.6630869
## sample estimates:
##       cor
## 0.4529498
```

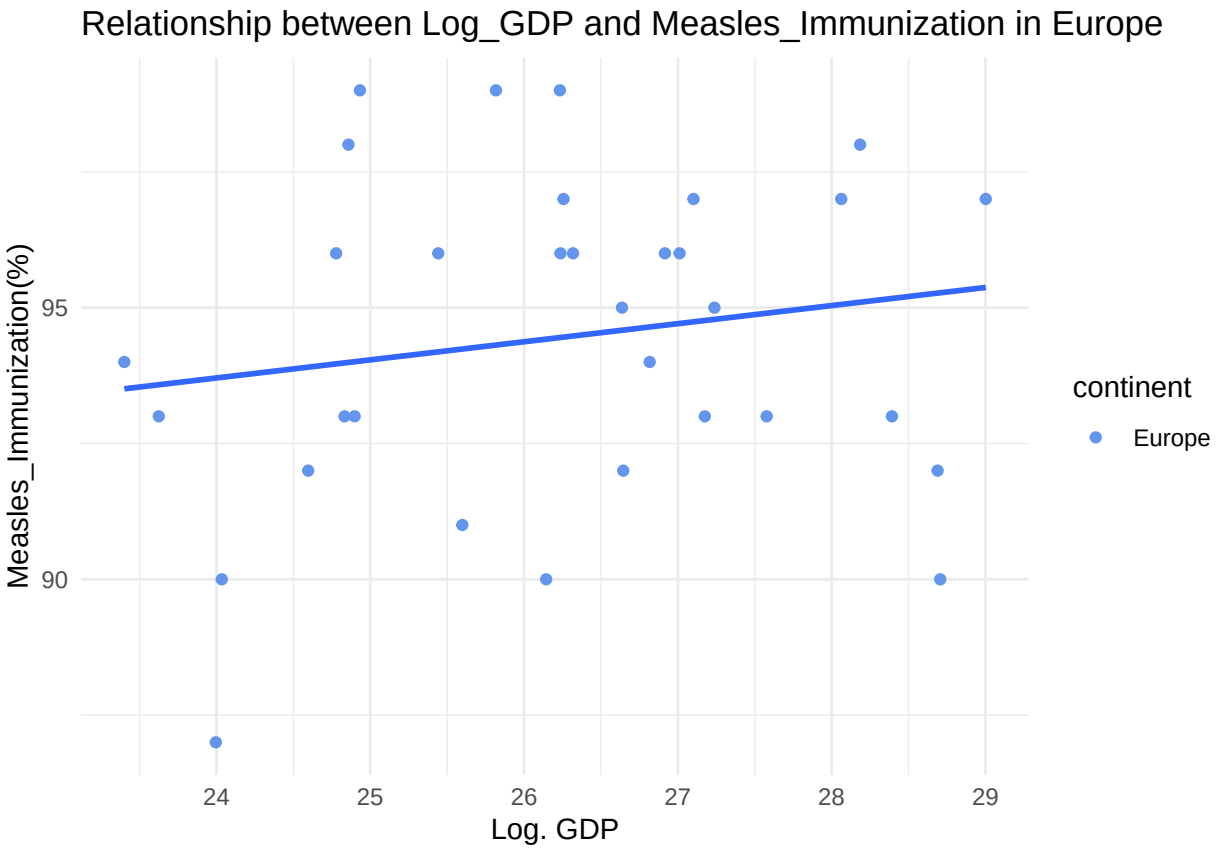
This graph visualizes the correlation between a nation's GDP and the percentage of the population that has access to electricity. The correlation coefficient was calculated to be about .453. However, it is worth noting that this is after the outliers, being the nations with 100% access to electricity, are removed. Nevertheless, even with the outliers added back in, the correlation coefficient is still almost unchanged at about 0.433. Therefore, regardless of accounting for outliers or not, there seems to be little correlation between the two factors. Even when accounting for individual continents or observing the shape that the scatter plots form, it is difficult to find signs of relation.

Figure 6 by Ben Wu



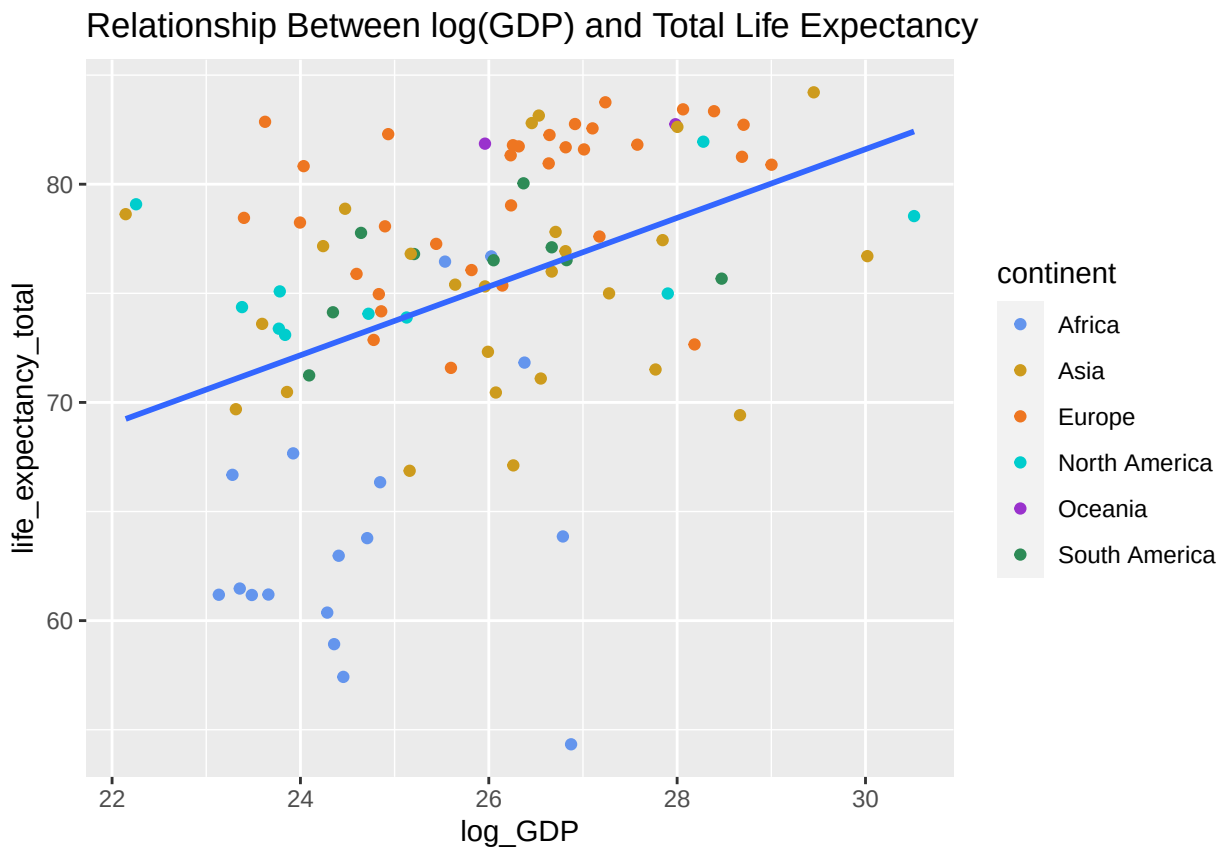
To better understand the correlation between a country's GDP and its citizens' wellbeing, the Measles_Immunization percentage is taken into consideration. By its name, Measles_Immunization would reflect a country's level of Medicare, people's acceptance to vaccination, and their education level, etc. However, as shown by the graph above, the correlation between Measles_Immunization and Log.GDP is rather weak, with a positive correlation coefficient being 0.3697, less than the standard of 0.5. This may probably be because of the difference between irregular emphasis development on economic growth, rather than those education, medical services that are more closely related to citizens' wellbeing. Given such a scenario and the sparsely spread out scatterplot, it pushes us to closely look at the correlation condition for Europe, a continent that is generally well-developed.

Figure 7 by Ben Wu



As shown from the scatter plot and its linear regression above, its correlation coefficient is even weaker than GDP versus Measles_Immunization rate globally, being only 0.1715. This GDP versus Measles_Immunization percentage in Europe, plus 0.3697 correlation glocally, alludes to the fact that there is no direct correlation between countries' Log.GDP and their citizens' measles immunization percentage. There are too many confounding variables, like the accessibility of local medical services, cultural acceptance to vaccinations, citizens' general education level, etc. And measles immunization percentage is not closely correlated with GDP.

Figure 8 by Jack Wang




```
##
##  Pearson's product-moment correlation
##
## data:  log_GDP and banknew$life_expectancy_total
## t = 4.3999, df = 95, p-value = 2.831e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2309474 0.5645588
## sample estimates:
##       cor
## 0.4114412
```

To understand the relationship between the GDP of a country and its total life expectancy, a scatter plot was constructed by using the World Bank data relating the logarithm of GDP to the total life expectancy, and a regression line was also plotted on the graph. The correlation coefficient is r=0.411 which indicates that there is a weak correlation overall when all countries are plotted in the graph. Since there is no strong correlation, another hypothesis is that different continents might have different environments which is affecting the total life expectancy, and by eliminating the effects of continents the correlation might be different. For all intents and purposes, further research is required to identify meaningful relationships.

Data Analysis

Section 1: GDP & Labor Participation Rates by Bradley Tian

As shown in Figure 2, there exists a relatively strong correlation between a nation's GDP and the size of its labor force. It was thus hypothesized that, on average, higher GDPs are related to better labor force engagement. To test the validity of this conjecture and examine further relationships between GDP and labor force, we collected additional data from DataBank, the World Bank's central database, on GDP and labor participation rates (labor force size / total population) of nations worldwide for the year 2017. Utilizing the given dataset, we calculated labor participation rates for the year 2018 via the same methodology.

Average Labor Participation Rate for 2017:

```
## [1] 67.65723
```

Average Labor Participation Rate for 2018:

```
## [1] 46.40418
```

To effectively gauge the relationship between labor participation rates and GDP, we consolidated the two groups of data into labor-GDP ratios, which are calculated by dividing labor participation ratios with logged GDP.

Average Labor-GDP Ratio for 2017:

```
## [1] 2.747909
```

Average Labor-GDP Ratio for 2018:

```
## [1] 1.829137
```

In context of our predictions, we hereby established a null hypothesis stating that changes in labor participation rate are directly in proportion with changes in GDP; that is, any difference in labor-GDP ratios across different years is simply due to chance.

R2018 = R2017

In contrast, we established an alternative hypothesis stating that changes in labor participation rates are not directly related to changes in GDP; that is, differences between labor-GDP ratios across different years are real and incites the need for further research.

R2018 != R2017

With two ratio samples from 2017 and 2018, we conducted a two-sample hypothesis test comparing the difference between averages of ratios of the two years. As we seek to identify differences - both positive and negative - between the two averages, we designed this test to be two-tailed.

We began by calculating the respective standard error for each sample. Since the population standard deviation is unknown, we bootstrapped the population SD with standard deviations of the samples.

Standard Error of the 2017 Sample:

```
## [1] 0.03167095
```

Standard Error of the 2018 Sample:

```
## [1] 0.03169461
```

With the two sample standard errors, we then calculated the standard error of the difference:

```
## [1] 0.04480623
```

We then calculated the z-score as such:

(observed difference between averages of 2017 and 2018 ratios – expected difference, which is 0) / (standard error of difference)

Resulting in the value below:

```
## [1] -20.50546
```

The P-Value is then calculated as pnorm(-z) + 1-pnorm(z):

```
## [1] 9.621409e-94
```

As shown above, the p-value is extremely small, strongly suggesting that the null hypothesis should be rejected. Therefore, we conclude that labor participation rates and GDP are not directly related, and that additional variables related to the labor force, such as employment rates and supply chain statuses, should be investigated to better identify factors related to changes in GDP of nations worldwide.

Section 2: GDP & Measle Immunization Rates by Tracey Ley

Essentially, we strove to test the change in the measles immunization rate in children aged 12-23 months between 2 years. Vaccination contributes to healthy individuals and a healthy population contributes to economic growth. From the world bank's database resource online, we were able to secure the relevant immunization rates each year as well as the GDP in US dollars to compare with our data in the 2018 dataset. For each world bank dataset, we narrowed the results to only produce the column corresponding to 2017, then found the ratios (calculated via the same methodology in section 1) between immunization and log of GDP to set up our 2 sample hypothesis testing comparing the immunization rate changes conditional on GDP.

Average Immunization-GDP Ratio for 2017:

```
## [1] 3.477288
```

Average Immunization-GDP Ratio for 2018:

```
## [1] 1.829137
```

The null hypothesis becomes that there is no difference between the ratios of measles immunization and gdp from each year and that any observed difference is simply due to chance. Conversely, the alternative hypothesis states that the observed difference between ratios is real.

We then calculated the SE of both ratios by first calculating individually then pooling the samples by square rooting the sum of each standard error squared. From the difference between the average ratios divided by the SE, we could find both the z score and p value.

Standard Error of 2017 Sample:

```
## [1] 0.03894705
```

Standard Error of 2018 Sample:

```
## [1] 0.03169461
```

Standard Error of Differences:

```
## [1] 0.05021376
```

Z-Score:

```
## [1] -32.82271
```

P-Value:

```
## [1] 2
```

Since the p-value is approximately 0, which is extremely small and less than our stated alpha of 0.05, we reject our null hypothesis. Therefore, a significant difference beyond chance exists in the progression of immunization rates and GDP, suggesting that although changes in measles immunization rates may be moderately correlated to changes in GDP, we cannot declare immunization rates as an influencing/related factor for GDP change across nations worldwide.

Conclusion

Firstly, our exploration of the world bank dataset and variables of interest, in conjunction with GDP, suggested measles immunization and labor participation rates as variables to further investigate through hypothesis testing. In both of our 2-sample hypothesis tests, we created ratios of the respective variable over GDP between the year 2017 and 2018 in order to see whether observed differences are significant or due to mere chance. After sampling, we discovered that the p-values were extremely small, leading us to reject our null hypotheses. Since we are comparing the difference of ratios between GDP and labor participation rates as well as GDP and measles immunization rate. By rejecting the null hypothesis, we understood that the ratios are inconsistent, which suggests that it is highly unlikely for either of the variables to be directly related to GDP over time. Further research on other lurking variables, such as education and population growth rates, is needed to better identify likely factors that contribute to or indirectly influence GDP growth.

Due to the wild variability in the data resulting from such a large population sample, it was difficult to interpret much of our results with many of our exploratory variables having little to no correlation with GDP. Additionally, the variables of interest were not related to each other, further complicating the difficulty for any solid conclusions. If we were to conduct the tests again, we could subdivide the dataset differently, possibly by looking at variables only within a single continent or by looking at variables from one country over time and their relation/correlation with each other. Expansion upon such analysis would yield more accurate results that help us better understand the composition of GDP and its interrelationships with other components of societies across the world.

References

"GDP (Current US$)." DataBank, The World Bank, 2021, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

"Immunization, Measles (% of Children Ages 12-23 Months)." DataBank, The World Bank, 2021, https://data.worldbank.org/indicator/SH.IMM.MEAS?end=2018&start=2008.

"Labor Force Participation Rate, Total (% of Total Population Ages 15-64) (Modeled ILO Estimate)." DataBank, The World Bank, 2021, https://data.worldbank.org/indicator/SL.TLF.ACTI.ZS.