Factors in a Countries GDP:

Introduction

Gross Domestic Product (GDP) is the measure of the value for all goods and services produced in a specific time period across a country. GDP is used to help determine or measure a country's wealth, and a higher GDP is correlated with a better economy. Growing GDP is important for a country's success and prosperity.

In this project, we utilized data on countries all over the world and examined which factors affected GDP the most to identify areas needing improvement to directly grow their GDP.

Data

For this project we used two primary data sources: Countries of the World: A Simple Example, and a Kaggle dataset called countries_of_the_world.csv.

2.1 Countries of the World: A Simple Example
We used this website to obtain information about the size and population of each country. This page has information on 250 countries, including the country name, capital, population and area (in km). To scrape the data into an R file for us to use, we used the xml2 R package and found the xml text for the four data features listed above. We then had to remove extra space and "\n" text from the country feature using gsub(). The process produced a data frame with 250 entries and four features. The scraping code is included in the R script "Wrangling_Project.R".

2.2 countries_of_the_world.csv
This Kaggle dataset is a collection of 227 data entries, including statistics on countries' economies and the trends in their population changes. We removed features from the dataset that were not relevant to the analysis we are conducting or that were scraped previously from dataset *2.1*. We also found and replaced commas with periods before loading the data into R because the dataset read all periods found in numbers as commas. We read this dataset into R and used gsub() to remove extra space from the country column. The reading and data processing code is included in the R script "Wrangling_Project.R".

https://www.scrapethissite.com/pages/simple/

2.3 wrangling_countries.csv
The datasets both shared country names, so we used vertical integration to merge our two datasets together based on country. Dataset *2.1* had more countries than dataset *2.2*, leading us to have a lot of missing values. To deal with this, we completely omitted any rows that had missing values. While this did trim our data down a lot, we thought this was the best way to handle the missing data points because of how much countries differ, making imputation of missing fields not an option. We thought that imputing missing values would falsify our results

more than removing rows with missing values would. We wrote this merged dataset into a new CSV file called wrangling_countries.csv. The data integration and cleaning code is included in the R script "Wranging_Project.R".

https://www.kaggle.com/datasets/fernandol/countries-of-the-world

Table 1: Data dictionary

Column	Type	Description
Country	Text	Unique country name
Capital	Text	Unique capital name
Population	Numeric	Country population
Area	Numeric	Country area (in km)
Region	Text	Region countries are located in
Net Migration	Numeric	Number of immigrants coming in each year
Infant Mortality	Numeric	Mortality rate of infants (per 1000 births)
GDP	Numeric	GDP ($ per capita)
Literacy	Numeric	% of population that is literate
Phones	Numeric	Number of phones per 1000 people
Crops	Numeric	% of land dedicated to farming
Climate	Numeric	Climate rating
Birthrate	Numeric	Rate of birth
Deathrate	Numeric	Rate of death
Agriculture	Numeric	Agriculture emphasis
Industry	Numeric	Industry emphasis
Service	Numeric	Service emphasis

Analysis

The goal of this project was to examine what factors affect the GDP of each individual country the most to give them direction on how to improve their GDP.

3.1 GDP correlation

To gain knowledge on which factors would affect GDP the most, we first started by finding the correlation between GDP and each feature in our dataset to ensure that we are exploring the features that are relevant. Table 2 displays the correlation coefficients we came up with.

Table 2: Correlation Coefficients

Correlation with GDP	
Data Column:	Correlation:
Population	-0.080
Area	-0.031
Migration	**0.308**
Infant	-0.563
Literacy	**0.525**
Phones	**0.894**
Crops	-0.169
Climate	-0.099
Birthrate	-0.538
Deathrate	-0.201
Agriculture	-0.504
Industry	0.052
Service	**0.325**

From finding the correlation between each feature and the GDP, we were able to determine that the **yearly migration**, the **percentage of the population that is literate**, the **number of phones per 1000 people** and the **service rating** had the highest positive correlation with GDP. We then decided to investigate these four features individually to find out what kind of impact they have.

3.2 GDP and Migration

Using a ggplot2 scatterplot to visualize GDP per capita to the net migration of each country. The scatter plot (figure 1) suggests that there is a weak positive correlation between GDPs per capita and net migration, meaning that the higher the net migration the higher the GDP per capita. Something that should also be noted is that there is a large number of countries with a net migration of 0 or near 0, which could be caused by strong policies against immigration within a country. The positive correlation here makes sense logically since countries with prosperous economies will be a more likely migration location than countries without.

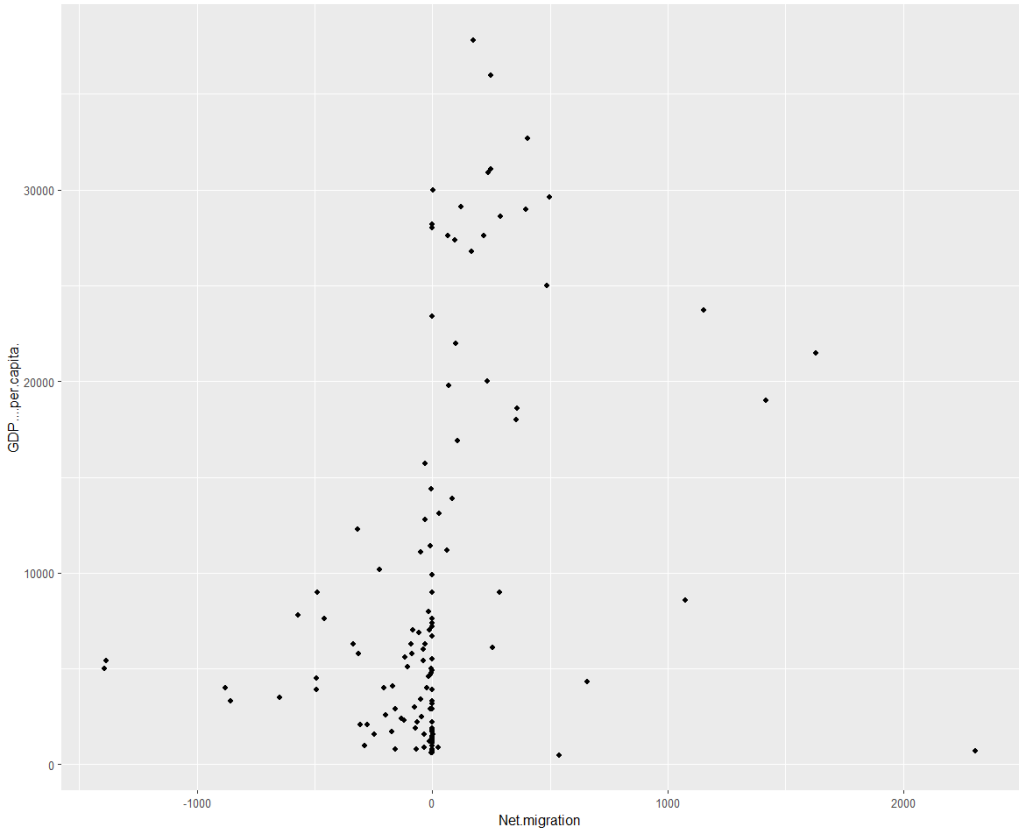

Figure 1: GDP per Capita vs Net Migration

3.3 GDP and Service

Using the data, we were able to compare GDP per capita and service to determine how impactful the service industry is on GDP. Our findings show that there is a positive correlation between GDP and service in a country. It appears that when a country provides a small to middling number of services, GDP remains about the same. However, when a country is the top third in services provided, they see a spike in GDP compared to other countries. This is most likely because the more services a country can provide, the better their economy tends to be.

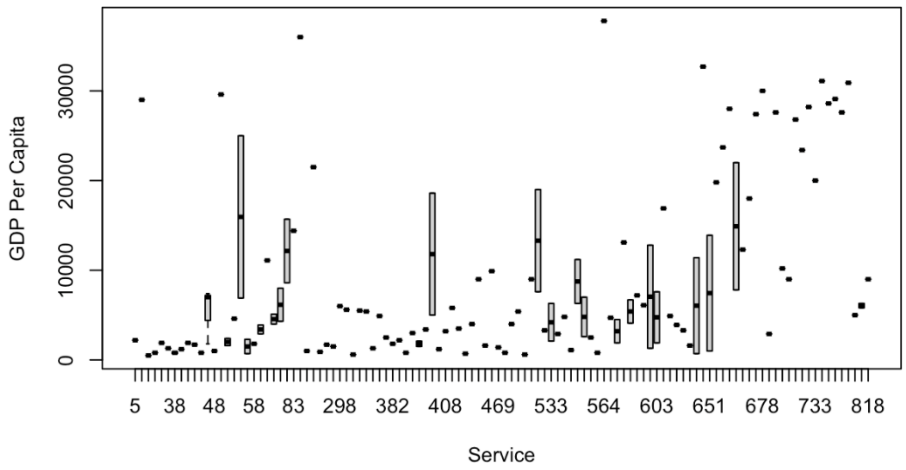

Figure 2: GDP per Capita vs Service

3.4 GDP and Literacy

Continuing with ggplot2 scatterplots, it can be seen that higher literacy rates for a country create much higher GDP outputs. There is a very strong, positive correlation between the two features at the rate of 0.525. In theory, as a country increases its literacy rate, it should see a subsequent increase in GDP as well. However, GDPs vary extensively at the highest rates. A country with little to no literacy, even up to 50%, is sure to have a low GDP to match. Once the 50% threshold, GDP begins to vary, but really takes off after the 75% mark. Countries with 100% literacy rate may also have low GDP but tend to have much higher GDPs to match.

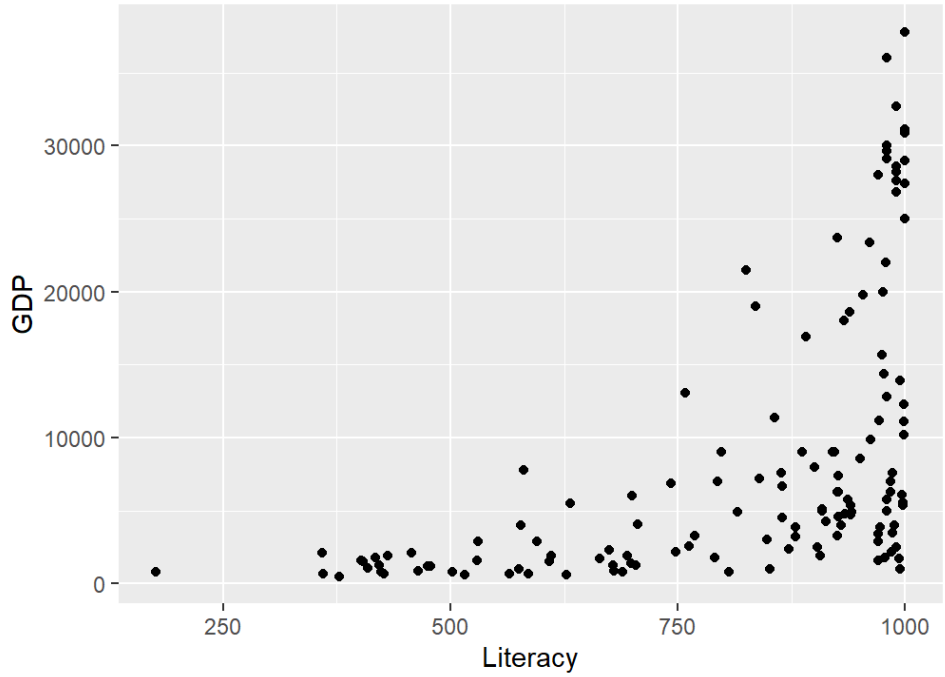

Figure 3: GDP per Capita vs Literacy

It is important to note that literacy may have a strong correlation with other features such as phones, industry, etc. that may help to improve literacy rate in unison, generating GDP growth.

3.5 GDP and Phones

We know that the GDP per capita of a country and the number of phones per 1,000 people have a very high, positive, and linear correlation of 0.894. Figure 4 shows the linear relationship between the two and indicates that as the number of phones per 1,000 people increases, GDP does as well.

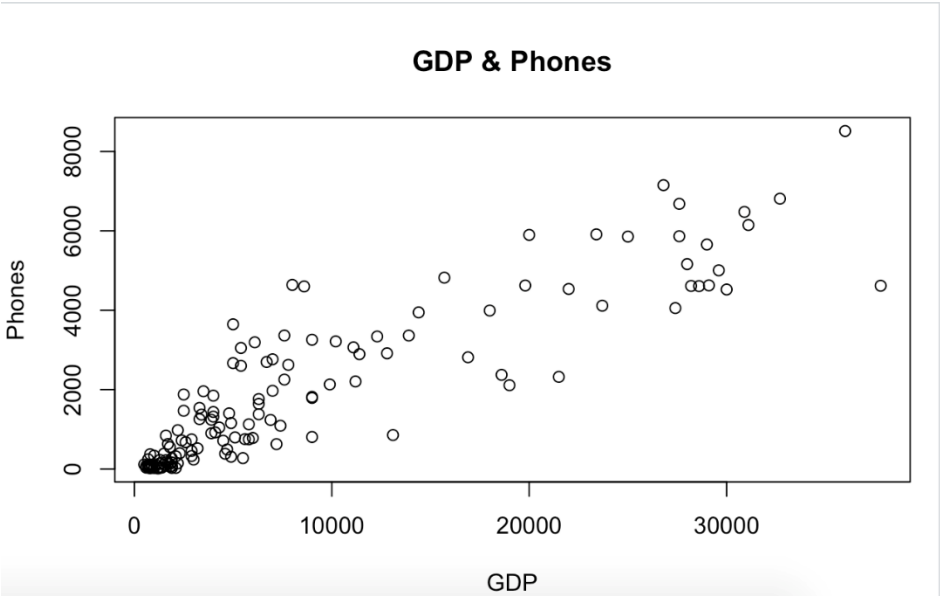

Figure 4: GDP per Capita vs Phones per 1,000 people

Figure 4 also shows concentrated data around 0 for both features and becomes sporadic as the values increase. To determine if the data is normally distributed, we conducted a Shapiro-Wilks normality test on both GDP and phones, shown below in Table 3.

Table 3: Shapiro-Wilks normality test

Data:	**W:**	**P-Value:**
GDP Per Capita	0.77863	4.724e-13
Phones Per 1000 People	0.85504	3.135e-10

The tests yielded a p-value of 4.724e-13 for GDP per capita and a p-value of 3.135e-10 for phones per 1,000 people. Both p-values were below our alpha of 0.05, leading us to reject the null hypothesis and accept the alternative hypothesis that our data is not normally distributed.

All code for analysis and visualization is included in the R script “Wrangling_Project2.R”.

Conclusion

This project incorporates descriptive data on countries scraped from a Countries of the World website, as well as statistical data from a CSV file downloaded from Kaggle to investigate what factors impact a country's GDP the most. We found the correlation between each statistical data feature and GDP, and found that the four features with the highest, positive correlation were the yearly migration, the percentage of the population that is literate, the number of phones per 1,000 people and the service rating. We further analyzed these for features to discover more about their relationship with GDP.

Using scatter plots and summary statistics, our initial assumptions that the yearly migration, the percentage of the population that is literate, the number of phones per 1,000 people and the service rating impact a countries GDP in a positive way was confirmed. This shows us that advancements in technology, education and population are major factors in a country's GDP per capita and that countries should focus on these areas to improve their GDP.

We had several limitations in our project that should be acknowledged. First, the number of countries in our scraped data was different than the number of countries from our Kaggle dataset, causing us to lose a lot of data points. Second, we focused our analysis on the four variables with the highest positive correlation, but there were three data features that had a negative correlation over –0.5 (infant mortality, birthrate and agriculture rating) that we did not look further into. These variables could have given us insight into the reasoning behind countries with low GDP per capita. Finally, our analysis was introductory and only went into four different features, and there are lots of other factors that play into a country's GDP that could be explored.