Data Management Project - Parts 1 & 2

A Research Paper
Presented to
The MDM4UG Class
The Academy for Gifted Children - PACE

In Partial Fulfilment
of the Requirements for the Course
Mathematics of Data Management
By Jonathan Chiang
January 2019

Contents

Chapter 1

The Problem and Its Scope

1.1 Introduction

Gross domestic product (GDP) per capita is a measure of a country's total economic output divided by its population, i.e. how productive the average member of society is. GDP is important because it gives a basic overview of how well a country's economy is doing. Several factors affect a country's GDP per capita. Statistical analysis can identify which of these factors are the most significant. The scope of this study was limited to identifying the significance of factors, and not going beyond to suggest what ways to change them in a positive way or suggest why they are correlated with GDP. Many of these factors are complex and in fact interact with each other, but it still should hold true that the factor that accounts for the most variance in the data will be the most significant. The factors studied are as follows: IQ, Democracy Index, age of population, education spending, and murder rate per capita.

Definition of factors as used in the research: **gross domestic product** (GDP) is a monetary measure of the market value of all the goods and services produced by a country in a year. An **intelligence quotient** (IQ) is a score derived from several standardized tests designed to assess human intelligence. The **Democracy Index** is an index published yearly by the Economist Intelligence Unit that gives a value to the state of democracy in a country. The value ranges from zero to ten, with zero being a full authoritarian regime with little personal influence over government, while a ten is an ideal democracy. **Age** is a measure of the length of a human's existence

in years. **Education spending** is defined as a country's total government education expenditure divided by its total GDP, expressed in terms of a percentile, e.g. Denmark spends 8.7% of its total GDP on education. **Murder rate per capita** is a measure of how many murders are committed in a country in a year divided by its total population. The reason crime rate was not used is because "crime" is a very vague term that is hard to create a metric upon. Murder rate is regardless a good indicator of a country's social and political stability.

The Pearson Correlation Coefficient (also known as the r-value) is a measure of the linear correlation between two variables X and Y. A value of $+1$ means a perfect positive correlation, a value of -1 means a perfect negative correlation, and a value of zero means no correlation. For a sample $\{(x_1, y_1), ..., (x_n, y_n)\}$:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{1.1}$$

where $\bar{x}$ and $\bar{y}$ are the respective sample means. The coefficient of determination (also known as the r^2-value) is the proportion of the variance in the dependent variable Y that is predictable from the independent variable X.

1.2 Problem

1.2.1 Problem

The factors studied are as follows: IQ, Democracy Index, age of population, education spending, and murder rate per capita. After analysis of data, a single factor was presented as the most significant in determining a country's GDP per capita. Additionally, the other factors were ranked according to how relevant they are in determining GDP per capita (how strongly correlated they are with GDP per capita). In doing so, this also provided a specific value for how significant each factor is (namely, the Pearson correlation coefficient).

1.2.2 Hypothesis

IQ is the most significant factor in determining a country's economy, followed by democracy index, educational spending, age, and murder rate per capita in that order.

1.3 Research Methodology

Materials needed: Google Sheets, data sets for each factor; for each factor:

1. Data was gathered from a reliable source that provided the value of a given metric for a comprehensive list of countries. The data was ported to Google Sheets. The first column contained the name of the country, and the second column contained the value of each factor's metric. The country list was sorted in alphabetical order.

2. Data was imported from a constant dataset of a country's GDP (The World Bank Group, 2018), using data from the relevant year. For example, if the IQ data was measured in 2002, the 2002 GDP dataset would be used. The data was matched appropriately in a third column with each country.

3. Using the graphing tools available in Sheets, a graph was created with the factor plotted on the x-axis and the GDP per capita plotted on the y-axis.

4. The r^2-value was calculated using the the tools provided.

Chapter 2

Presentation, Analysis, and Interpretation of Data

This chapter presents the data, analysis, and interpretation of different factors that may affect GDP per capita, by country. Data sets from different sources were used, resulting in varying reliability of results. Not all data was available for every country in every data set.

2.1 Net Migration per Capita

When net migration per capita by country is compared to GDP per capita by country, the raw, unfiltered data suggests that $r^2 = 0.228$. While this is a decent correlation, there are a few outliers that can be filtered out. The five right-most countries in the x-axis (net migration per capita) (Singapore, Equatorial Guinea, Qatar, Oman, and Bahrain in that order from left to right) all have relatively small populations, which are bound to be more sensitive to migration. If we remove the five outliers, the r^2 value becomes much more realistic and rises to a whopping 0.43. Refer to Figure 2.2. Furthermore, using a polynomial model, an r^2 value of 0.565 is achieved, meaning that net migration can be reasonably expected to account for approximately 56.6% of the variation for GDP.

Even visually, a very strong correlation can be seen with the naked eye. There is a positive correlation. What this trend means is that countries with low/negative net migration usually have a low GDP per capita, while countries with high/positive net migration usually have a high GDP per

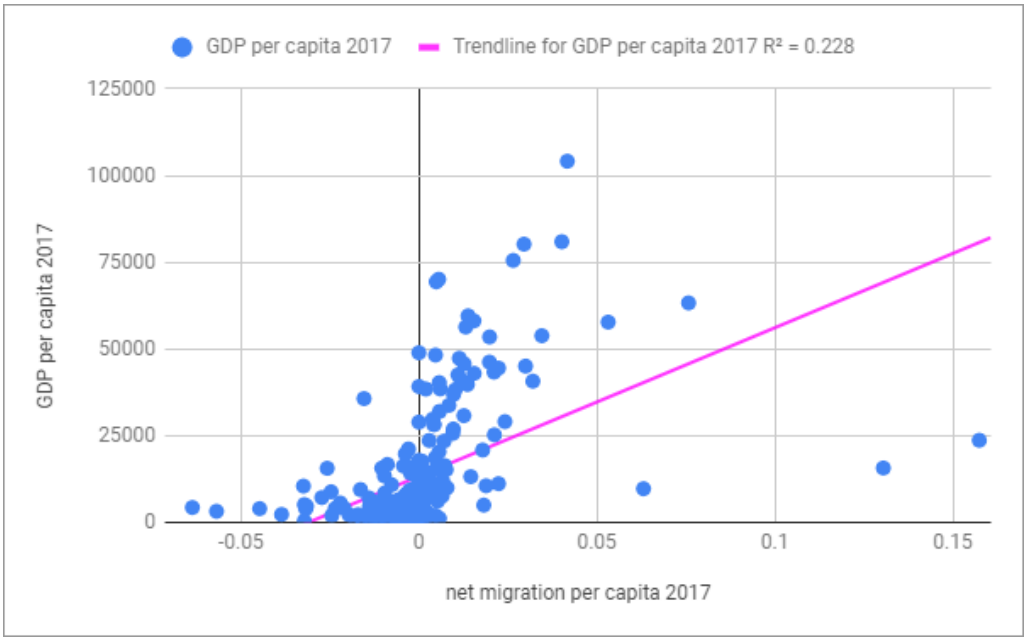

Figure 2.1: Net migration per capita vs. GDP per capita (2017), source [2]

capita. Why this occurs is beyond the scope of this research; it is possible that a high GDP would attract more immigrants to a country, while it is also possible that a high emigration rate would reduce the skilled labour force and by extension the GDP per capita of a country. The rightmost data point is Luxembourg with a net migration per capita of 0.042, a total population of 600 000, and a GDP per capita of US$104 103.

2.2 Government Expenditure on Education as a Percentage of GDP

The method used to obtain this data was different and more complicated than the typical method outlined previously. Due to the nature of the data set and unavailability of data, the most recent data measurement was used for each country. This was accomplished through the use of the function

=LOOKUP(9.99E+307, E6:BK6),

which is a simple way to obtain the most recent data point in a row/column,

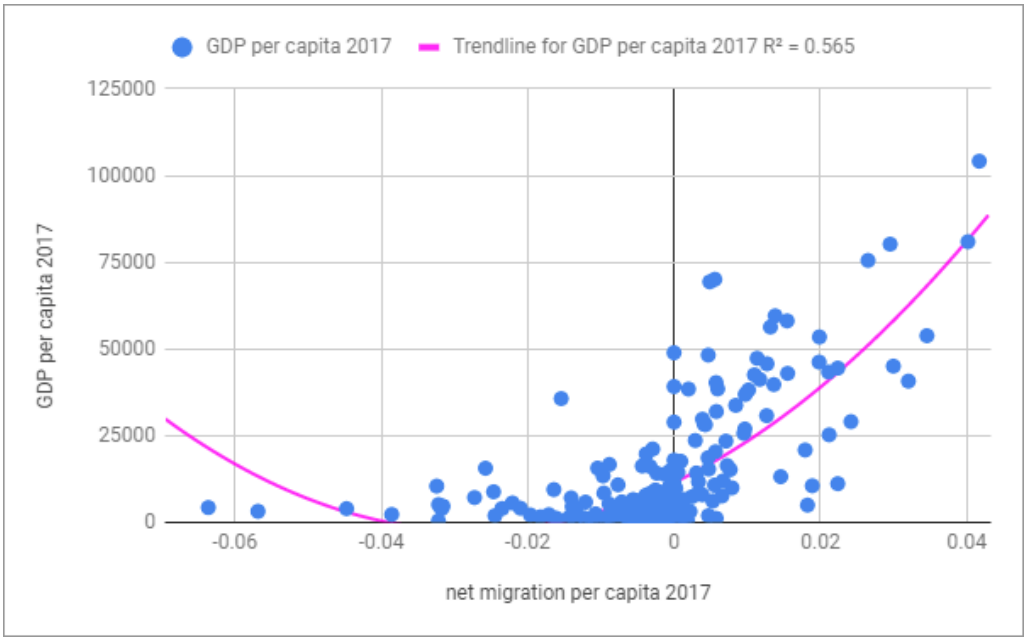

Figure 2.2: Net migration per capita vs. GDP per capita (2017), adjusted for outliers, source [2]

based on the input in the second input area. Unfortunately the data must only be compared to the 2018 GDP of each country, but it still holds that the correlation should be present (if any) albeit not as accurate. Removing Nauru as an outlier (the only data point was from 2002 showing that they spent 79.1% of their government expenditure on education, which is unrealistic and likely inaccurate), we get Figure 2.3.

Ultimately, even considering the outliers, the r^2 value is very small and can be considered statistically insignificant. No matter what kind of model (linear, polynomial, exponential logarithmic, moving average, etc.) the r^2 is not meaningfully significant. This shows that government expenditure on education is probably not a good indicator for GDP per capita, and accounts for little variance within the data.

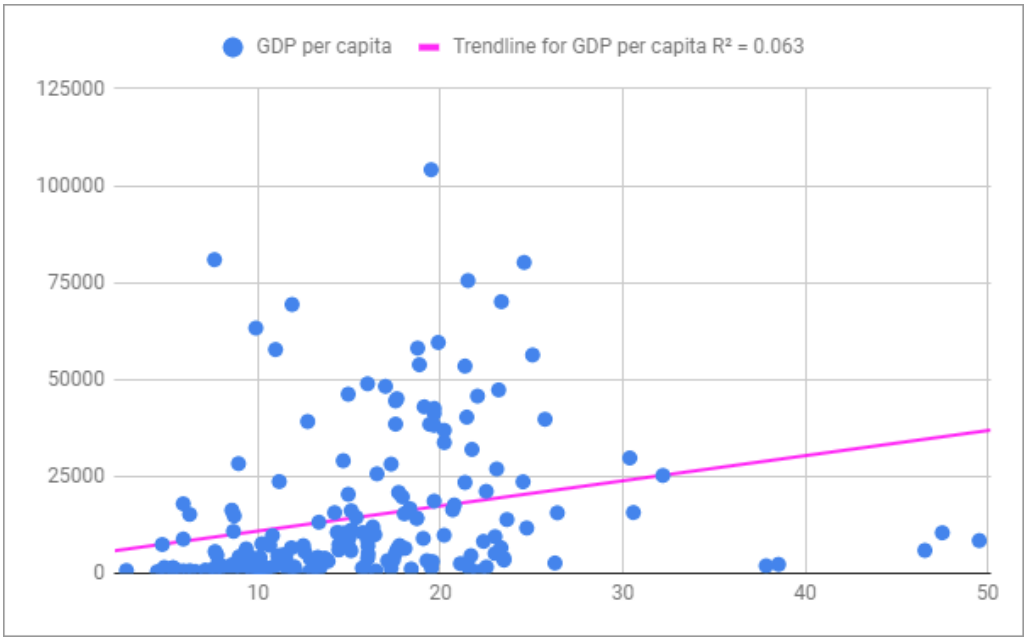

Figure 2.3: % Government expenditure on education (various dates) vs. GDP per capita (2017), excepting Nauru (2002), source [2]

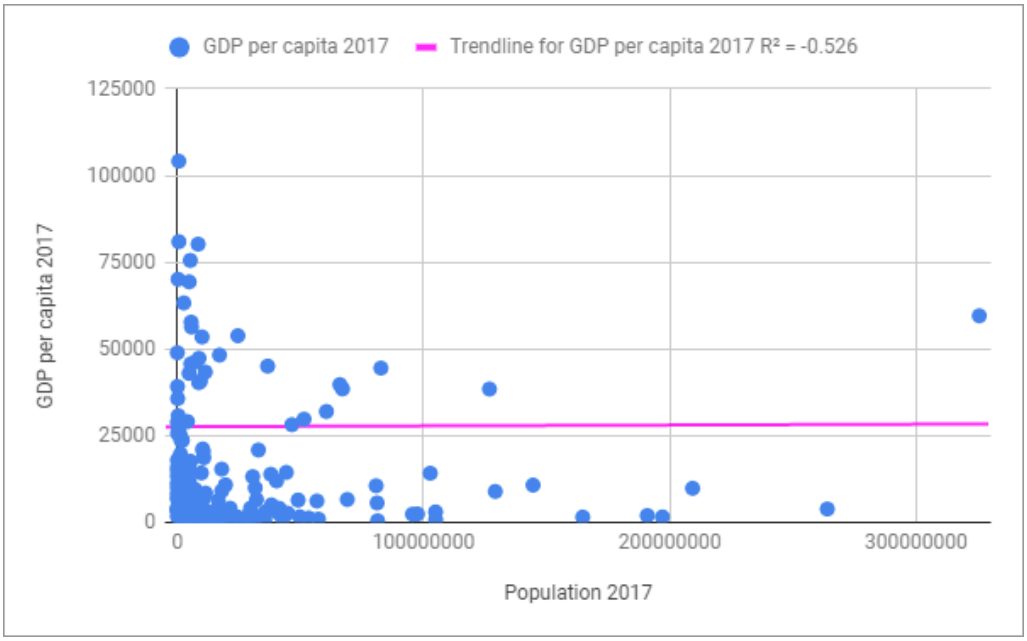

Figure 2.4: Population vs. GDP per capita (2017), source [2]

2.3 Population

Population is already inherently not a great factor to use when determining GDP. The goal of this paper is to provide countries with factors that they can effect change on to positively change GDP per capita. However, population is not something very easy for a government to regulate. Regardless, population should still be considered for the sake of noticing trends as with any other factor. This leads to Figure 2.4.

As clear from the graph, using a polynomial regression provides an r^2 value of -0.526. This means that the chosen model is a very poor representation of the data. Every model apart from polynomial regression yields poorer results. The reason for this is probably due to the high amount of countries with low population and/or low GDP per capita.

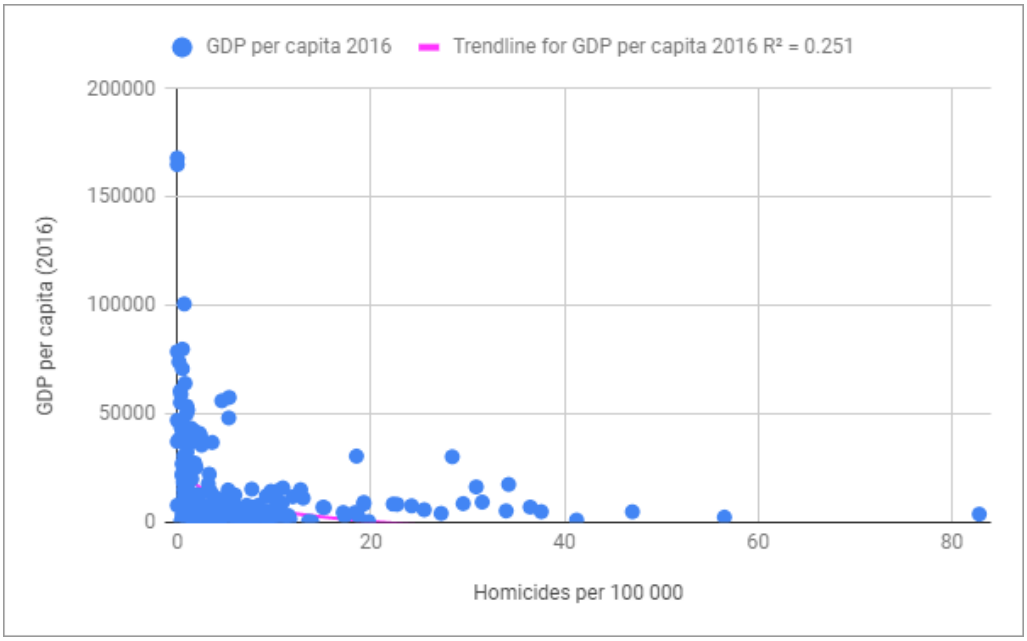

Figure 2.5: Homicides per 100 000 (various years, $\leq$ 2016) vs. GDP per capita (2016), adjusted for outliers, source [2]

2.4 Homicides per 100 000

Again, the data for this factor was not as readily available for each year. As such, in accordance with the method outlined in Section 2.2, the most recent data was used for each country, meaning the data is from various years. This introduces some inaccuracy but the trend should still be apparent, if any. Similar to population, in Figure 2.5, there is a big cluster of data points near the origin (0, 0) relative to the large outliers. However, contrasting to Section 2.2, a mild trend can still be found if using a logarithmic regression, with $r^2 = 0.251$. In addition to this, cleaning up the extreme outliers which are El Salvador at 82.8 homicides per 100 000 in 2016, Monaco at US$168 010 GDP per capita in 2016, and Liechtenstein at US$164 993 GDP per capita in 2016, using exponential regression yields an r^2 of -0.253. This does not actually change the r^2 value by much. The trend means that a low homicide rate is correlated with a high GDP, and a high homicide rate is correlated with a high GDP.

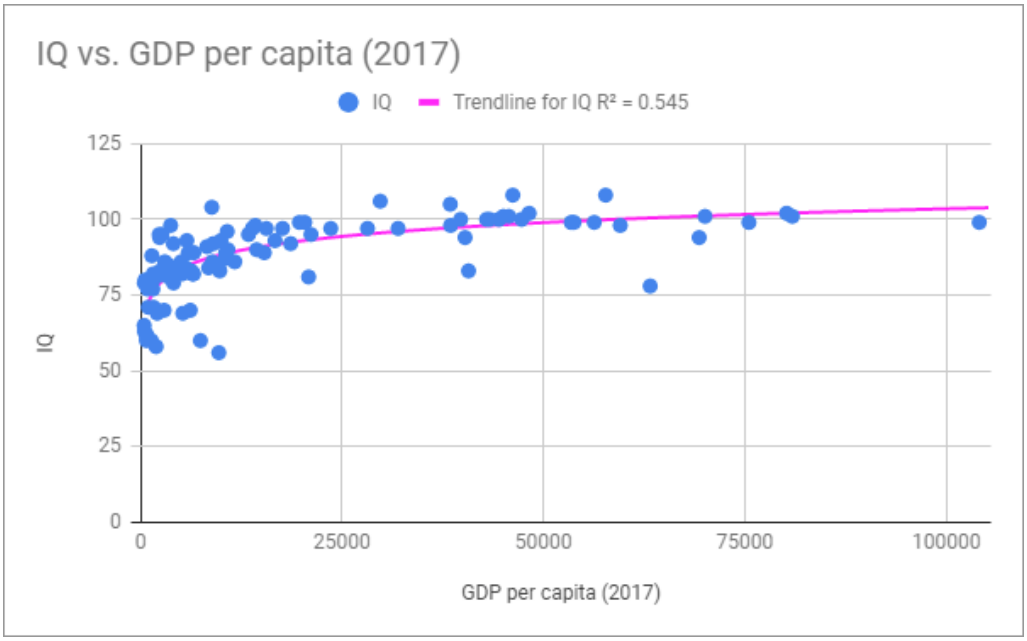

Figure 2.6: Intelligence Quotient (IQ) vs. GDP per capita (2017). IQ data taken from [1] and GDP data taken from [2]

2.5 Intelligence Quotient

The intelligence quotient is not as reliable as the other factors, and its measurement can be affected by many things. Additionally, the data retrieved from certain countries may be more or less reliable to a certain degree. One criticism of IQ is that "... IQ was developed by West Europeans for West Europeans according to West European standards. It is still debatable whether this procedure can be applied to people(s) with entirely different social structures, cultures, values and ways of thinking." [1]

From the r^2 value of 0.545, we see that there is a very strong correlation between intelligence and GDP per capita. This result was obtained through logarithmic regression.

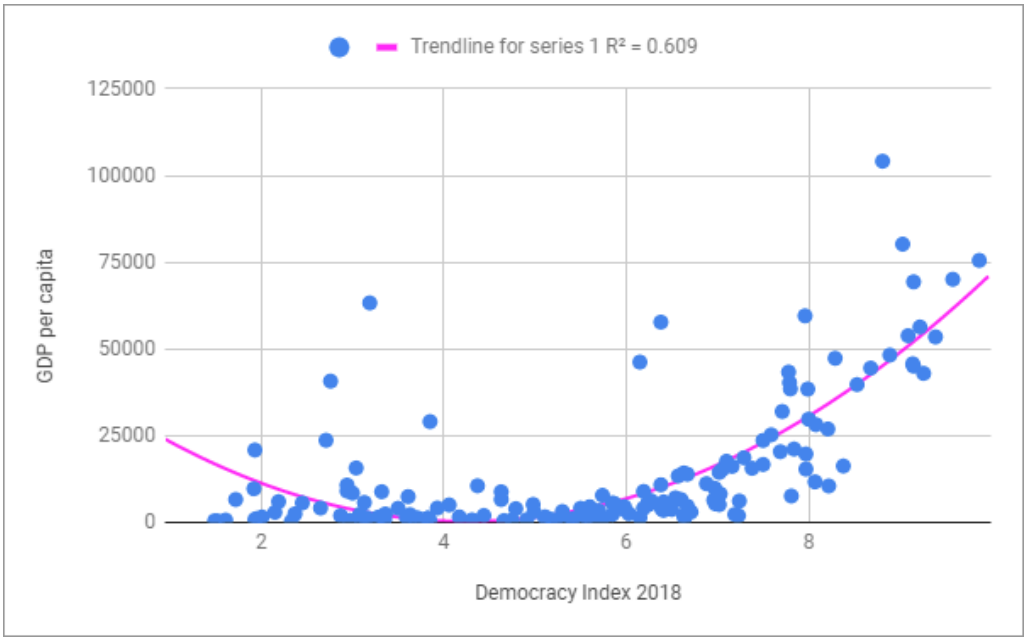

Figure 2.7: Democracy Index (DI) (2018) vs. GDP per capita (2017). DI data taken from [3] and GDP data taken from [2]

2.6 Democracy Index

The Democracy Index from 2018 published by The Economist was compared against the GDP per capita for each country. Using a polynomial regression, an r^2 value of a whopping 0.609 was obtained. Refer to Figure 2.6. A very clear pattern emerges, with more democratic countries earning more per person. The main exception on the left side is Qatar with a GDP of US$63 249 per capita, with a democracy score of only 3.19. The main explanation for this is its vast oil reserves, meaning that it can have a large economy even with a mostly authoritarian government restricting personal freedoms and rights.

2.7 Comparing Factors

Ranking	Factor	r^2-value
1	Democracy Index	0.609
2	Net Migration per Capita	0.565
3	IQ	0.545
4	Homicide Rate	0.251
5	% GDP Gov't Spending on Education	0.063
6	Population	-0.526

Chapter 3

Summary of Findings, Conclusions, and Recommendations

3.1 Summary

The research aimed to discover the significance of the correlation between certain factors and GDP per capita. It also aimed to identify the factor with the highest correlation with GDP per capita. The factors studied were as follows: IQ, Democracy Index, age of population,education spending, and homicide rate per capita. The experiment used a spreadsheet program (Google Sheets) in order to communicate the findings visually and also obtain the r^2, which was used as the basis for determining correlation. The information was gathered from similar, reliable sources. The information was mainly sourced from World Bank, only using other sources where information regarding certain factors were not available (such as Democracy Index and IQ). Some of the problems with the factors were identified, such as the measurement of IQ. For some of the factors, outliers significantly affected the graph and correlation value, so they were removed and a logical explanation for the appearance of the outliers was identified.

3.2 Findings

The study which aimed to determine the factor most highly correlated with GDP per capita found that Democracy Index was the most significant factor in determining GDP per capita. This factor was closely followed by Net Migration per Capita and IQ. The top three factors had an r^2-value of greater than 0.5, which means that the correlation was significant. One factor, Homicide Rate per 1000, had a lesser correlation with GDP per Capita, yielding an r^2-value of 0.251. The final factors studied, Government Spending on Education (as % of GDP) and Population had very small or negative correlations with GDP per capita. The explanation for the negative r^2-value is that Google Sheets uses a version of r^2 that can yield negative results. How this is interpreted is that the given line of best fit is not well representative of the data. All curves tested (linear, polynomial, etc.) yielded negative r^2 values, of which the highest value (least negative) was selected. As can be shown from the results, the summary of the results goes as such:

1. GDP per capita has a significant correlation with some measurable factors.

2. Democracy Index, Net Migration per Capita, and IQ are some of the significant factors in determining a country's GDP per capita.

3. The data sets and method of obtaining results was reliable.

3.3 Conclusions

The explanations for why some factors affect GDP per capita is beyond the scope of this paper. The goal of the paper was to identify how closely each factor correlates with GDP per capita. Since GDP per capita is a measurement of the global economy and thus a by-product of human behaviour and decision-making, the causal relationships between the factors and GDP per capita cannot be determined from this paper alone. For example, this paper can not answer the question of whether or not a factor determines GDP per capita or vice versa. In the real world, economists have suggested that many of these relationships are two-way. Some reasonable hypotheses can be suggested, such as a country's net migration being determined by their GDP per capita (citizens from poor countries want to seek better economies, and

citizens from richer countries have an incentive to stay), which is consistent with the findings. However, the verification of these hypotheses cannot be determined here. Some reasonable explanations for outliers can be formed, such as oil-rich economies being very wealthy per capita yet not needing a highly-educated and intelligent population.

3.4 Recommendations

As suggested in Section 3.3 Conclusions, further research by economists and behavioural psychologists is necessary to determine the nature of the relationships between certain factors and how they affect GDP. However, this paper identifies certain factors that should be studied further (Democracy Index, IQ, Net Migration per Capita), along with a method to identify other factors that could be significant, yet were not studied in this paper. Not only should further research study look at the reasons for why some factors are significant, further research should also identify why some factors are not significant. The original hypotheses supposed that population and government spending on education would have been more significant correlated with GDP per capita, yet the results suggested otherwise. Once the reasons for these relationships are found, government policy can be enacted in order to potentially increase GDP per capita and thus provide benefits for citizens of countries.

Bibliography

[1] Lars Eglitis. Iq by country, 2010.

[2] Multiple Sources. The world bank group, 2019.

[3] The Economist Intelligence Unit. Democracy index 2018: Me too?, 2018.