# DETERMINANTS OF ECONOMIC GROWTH OF A NATION

**ASHISH U. SOLLAPUR**

**RYERSON UNIVERSITY**

**TORONTO**

**STUDENT ID 501140340**

**DATA ANALYTICS, BIG DATA AND PREDICTIVE ANALYTICS**

**DR. DERYA KICI**

**04 APRIL 2022**

# Contents

# Abstract

Development of any economy is based on a broad range of indices such as agricultural income, gross domestic product (GDP) growth, employment rate, income equality index etc. An economy's growth is indicated by such factors which can be used as guidance tools for the respective economy to decide on future course of action. The data points can help understand where the resources need to be optimally allocated for better growth. The data may also present facts where there has been a misuse of available resources thus inviting urgent corrective action.

One of the several reliable data metrics is published by The World Bank featuring a set of various monthly indicators in several broad groups such as climate change, education, health etc. The published data includes a set of data points since the year 1960 for all the countries in the world. However, since the data from 1960 will be very exhaustive for my study, I plan to analyse the data starting from 2001 up to 2019. Some of the metrics have missing data as several aspects were started to be measured only until a few decades ago. In this project, I aim to analyse the datasets published by The World Bank under the following broad groups: (1) Economy & Growth (2) Education (3) Financial Sector (4) Health

By analysing a set of attributes under each of the above broad groups, I plan to carry out exploratory data analysis to understand correlation and causality between the variables. I will also carry out time-series analysis to see how various attributes have changed over years and their interlinkage. This will help me understand if one parameter has a positive or negative impact with some other parameter; or it may be completely independent of other attributes in the dataset.

The dataset contains missing values and probable outliers in certain attributes. After cleaning the dataset, I will first attempt to draw patterns from the data. These patterns would be valid for most of the economies.

In the second part of this project, I plan to short-list 3 economies and carry out regression analysis, K-nearest neighbours classification and K-means clustering on the data points of these economies. This will provide a deeper understanding about the data for such economies. I will later interpret the findings and recommend areas where more focus is required to achieve faster and balanced overall growth.

My research questions are as follows:

1. How strongly correlated are the factors within education, health and financial sector to the GDP of an economy?
2. Which are the top 3 attributes that affect the GDP growth either positively or negatively?
3. What measures should be implemented to increase GDP in the countries selected for study?

As I draw conclusions from the data points, I plan to use several data points from different datasets which are correlated to each other. The data is published in the form of various datasets interlinked with each other. Thus, assessing data metrics from these groups will enable us draw insights about the overall health of an economy, its direction, areas of improvement, strengths, and weaknesses etc.

I will carry out the study using R-programming to analyse data and present my findings. The related files are uploaded on the Github account which can be located on the below link:

https://github.com/ashishsoll/Project

## Literature Review

Economic growth means overall development of a region in terms of quality of living, sustainability, environmental friendliness, average income of individuals, inflation, employment opportunities, healthy living, higher education opportunities among other aspects. Each of these stated parameters directly impacts the quality of living of any individual. For example, higher inflation leads to depreciation of the local currency eventually reducing the purchasing power. People of a region have higher productivity if they are healthy and eventually contribute more towards the overall growth of the region. Ample employment opportunities mean a robust workforce ready to deliver higher output in times of requirements. All these positive factors have a multiplier effect on the dependant industries which leads to total economic growth and development. While each of these parameters are important for the growth, one should also take into consideration the environmental impact. Every development should intend to have minimal or no environmental impact, albeit it can leave a positive development at times.

This literature review aims to understand the research already done in the field of economic growth and the parameters directly affecting it. There are papers published for individual aspects such as the impact of education, health, inflation etc. on the growth of economies within the region. Such studies assess the impact of only certain parameters on the GDP growth of localised economies. Hongyi and Huang (2009) studied how health and education affect the economic growth in China. Whereas Aziz and Azmi (2017) studied the factors affecting the GDP growth in Malaysia. There are empirical studies on the economic growth of African countries and its factors such as one carried out by Eggoh, Houeninvo, and Sossou (2015).

Aziz and Azmi (2017) studied the various parameters affecting the GDP growth in Malaysia. The authors tried to find the impact that inflation, foreign direct investment (FDI) and women's participation in the workforce on the GDP growth in Malaysia. The authors studied more than 30 years of data starting from 1982. Some of the methods they used to analyse were Ordinary Least Squares (OLS) method, Regression analysis to carry their in-depth analysis. The authors wanted to try and find out the reasons and factors affecting the growth of the GDP. They concluded that in their study, inflation does not contribute significantly to the GDP growth. However, the FDI has a huge contribution in deciding the GDP growth of an economy. Furthermore, the paper states that there is a positive impact of the female labour participation on the GDP growth. The paper's findings are consistent with other relevant studies carried out by the authors earlier. This study is limited to the geographical region of Malaysia.

Eggoh, Houeninvo, and Sossou (2015) studied the impact of human capital and the impact of health on the growth of an economy, especially in the African context. The authors mention that until their study was carried out, there were opposing results regarding the impact of health on the growth of the GDP. The research used two main methods of assessing the impact of health on GDP: the OLS method and the Generalized Method of Moment (GMM) and assessed the data from 1996 to 2010 of the majority of African countries. Indicators such as inflation rate, GDP per capita and share of exports and imports as a percentage of the economy's GDP were used to understand the changes. The study concluded by stating that more the spending on health, lower was its impact on the GDP growth. The authors cited inefficiencies such as leakages, red tape and corruption as the main reasons for their conclusion. They mentioned that the cossuption levels should be brought down and health spending increased to have overall positive impact on the GDP growth.

Karavaeva (2021) studied the relationship between sustainable development the economic growth in the European Region. The author states that while levels of education were significant deciding factors to the overall economic growth, there was not enough evidence in stating that healthcare support contributed to overall economic growth. The author studied 4 of the 17 interdependent factors contributing to a sustainable future for everyone. These 4 factors were Poverty, Health, Education, and Inequality levels. The author analysed the interdependence of these parameters by employing statistical tests such as correlation, carrying out the Durbin Watson autocorrelation test among other significant analyses. The findings were that higher income inequality promotes higher income growth in the region. Poverty alleviation and health of an individual do not contribute significantly to the economic growth of European Union (EU) member countries.

Deme and Mahmoud (2020) studied the effect of quantity and quality of education on per capita real GDP growth in the African region, specifically for the low- and middle-income countries. They carried out regression analysis for 34 African nations where the data consisted of various education parameters. The Graduate Record Examination (GRE) quantitative and verbal test scores were used to assess the outcome of the quantity and quality of education levels. The studies found that there is statistical correlation between growth in the primary school enrolments and per-capita economic growth. The findings were similar for secondary level education as well. The quality of the education was assessed by the quantum of government spending as a percentage of GDP growth on the education levels. The results supported the theory that higher economic growth is achieved by growth in the quality of education. However, the studies found that due to factors such as corruption and pilferage, there is loss in the quality of education being offered at the ground level. The authors concluded that there is strong association quantity of education and economic growth and that the policy makers should target school enrollment as a tool to achieve economic growth.

Another paper published by Chowdhury (2003) analysed the world bank data to understand the income distribution of the world. This paper focused on the income distribution analysis in terms of inequalities and disparities. The author discussed the growth-inequality parameter and studied the relationship between GDP growth rate and the income inequality by way of regression analysis and Theil's entropy index. The paper concluded two trends concerning the income equality: (1) The inter-regional component of the inequality is stronger than the global inequality. This inequality increases as time passes. (2) The inequality between two regions increases whereas it decreases within a region over time.

Cooray (2009) published a paper on The Financial Sector and Economic Growth where the author evaluated and assessed the financial sector size, activity, efficiency, and their interaction with the economic growth of 35 low and medium sized economies. The paper states that for the economic growth, the financial sector size, activity and efficiency are important parameters. If the resources are directed for productive uses of the financial sector of the 35 economies under study, it can induce higher and faster growth. To increase the efficiency, the overhead costs should be decreased, and bank concentration should be increased. The skill levels of the population of each of these economies can be increased to promote economic growth. This research paper's findings are consistent with the results of Beck and Levine (1999) which states that there is a positive correlation between economic growth and the development of the financial sector.

Marquez-Ramos (2019) studied the impact of education and literacy on the growth of an economy. The author studied the data from Spain's perspective. The study mentioned that the most common assumption when studying this relationship is linearity between the education level and growth of an economy. However, the study assumed that an economy may respond differently depending on the literacy level of its population and thus went on to assess the relationship dynamically. For education levels, the admission numbers for secondary and college levels were considered. The author applied the Ng-Perron unit root test and regression analysis to carry out the research. The paper's concluding remarks mention that there are inconsistencies in the patterns observed. The author states that education plays a vital role in the economic growth of Spain. However, there were irregularities observed in the patterns and thus, education plays a dual role.

Grant (2019) ventured to study the effect of primary, secondary and tertiary levels of education on the economic activity of a country. The study did not focus on any specific geographical region and hence can be considered to be a global research. The data from a sample set of countries was studied and this sample data was divided among countries with varying levels of income. The sample countries were divided into three groups – low income, lower middle income, and upper middle income. The data for the study was sourced from the World Bank's publicly available datasets. Instead of applying any statistical methods to assess the impact, the study was carried out empirically. Thus, only observational analyses were made. The study mentions that if the primary education attendance levels are doubled, it will have a positive impact of 4% on food availability. More than 170 million people can come out of poverty if all the children have at least the primary education. The paper mentions that for secondary education, the countries should strive to achieve gender equality as women who complete secondary education have higher chances of earning more than their male counterparts. Tertiary and voluntary education leads to higher aptitude for the students and it enables them to think independently. The paper mentions that the low income countries should focus more on enrolling students and spend more on the education infrastructure while the middle income countries need to stop pilferage of funds so that they are put to better use for infrastructure development. Overall, the higher levels of education lead to higher economic activity eventually leading to higher GDP growth.

While there have been studies concentrating on the factors affecting economic growth in certain specific regions, there are very few research available for assessing the reasons for economic growth and finding its determinants on a global level. I plan to study three determinants of economic growth – Education, financial sector and health and their impact on the growth of the economy. These determinants were selected based on empirical analyses of existing research available. In this research, I aim to find out the if these parameters affect the economic growth of a nation and to what extent.

## Data Description

World Bank conducts ongoing research of various parameters to depict the economic well being of a nation. The data is available from the year 1960 to 2020. However, for the purpose of this study, the data is selected only from 2001 to 2019 as there is incomplete data for the year 2020 and the data for the years prior to 2001 is pre-dated for the purpose of this study. As the research focuses on understanding the impact of the financial sector, health conditions and literacy levels in a country and their impact on the GDP growth of the country, only the relevant set of data published by the World Bank is used.

This study utilises four different data sets for the analyses. Each of the datasets has several parameters as detailed below:

## Dataset Summary

| Dataset | Attributes | Regions | Start Year | End Year |
|---|---|---|---|---|
| Financial Sector | 77 | 267 | 2001 | 2019 |
| Health | 256 | 267 | 2001 | 2019 |
| Education | 163 | 267 | 2001 | 2019 |
| Economy and Growth | 267 | 255 | 2001 | 2019 |

As the available data is multi-tiered – For each attribute, data is available for several years for multiple countries, the description of the important attributes relevant to my study from each of the four data sets is provided below:

### Financial Sector:

| Feature Name | Description | Type of Variable | Basic Statistics |
|---|---|---|---|
| Country Name | Name of the country | Nominal | Examples:<br>Canada, Australia, Italy |
| Indicator Name | Types of various indicators monitored by World Bank | Nominal | Examples:<br>Bank capital to assets ratio (%),<br>Automated teller machines (ATMs) (per 100,000 adults) |
| 2001-2019 | Year for which the data is available | Categorical | 2001:<br>Min: $-3.312\text{x}10^{11}$ |

| | | | |
|---|---|---|---|
| | | | 1st Quartile: 4<br>Median: 24<br>Mean: 1.068 x $10^{12}$<br>3rd Quartile: 1.09 x $10^{7}$<br>Max: 1.19 x $10^{15}$ |

Bank capital to assets ratio (%) (Continuous variable) – This ratio signifies the strength of the banking sector in the country with respect to its quantum of lending.

Automated teller machines (ATMs) (per 100,000 adults) (Discrete variable) - Automated teller machines are computerized telecommunications devices that provide clients of a financial institution with access to financial transactions in a public place.

Inflation, consumer prices (annual %) (Continuous variable) - Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals.

### Health:

Life expectancy at birth, total (years) (Continuous variable) - Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

People using at least basic sanitation services (% of population) (Continuous variable) - The percentage of people using at least basic sanitation services, that is, improved sanitation facilities that are not shared with other households.

## Education:

| **Feature Name** | **Description** | **Type of Variable** | **Basic Statistics** |
|---|---|---|---|
| Country Name | Name of the country | Nominal | Examples:<br>Canada, Australia, Italy |
| Indicator Name | Types of various indicators monitored by World Bank | Nominal | Examples:<br>Current education expenditure, total (% of total expenditure in public institutions),<br>Literacy rate, adult total (% of people ages 15 and above) |
| 2001-2019 | Year for which the data is available | Categorical | 2019:<br>Min: 0<br>1st Quartile: 8<br>Median: 46<br>Mean: 4.698 x $10^6$<br>3rd Quartile: 92<br>Max: 3.391 x $10^9$ |

Current education expenditure, total (% of total expenditure in public institutions) - Current expenditure is expressed as a percentage of direct expenditure in public educational institutions (instructional and non-instructional) of the specified level of education.

Educational attainment, at least Bachelor's or equivalent, population 25+, total (%) (cumulative) - The percentage of population ages 25 and over that attained or completed Bachelor's or equivalent.

Educational attainment, at least completed primary, population 25+ years, total (%) (cumulative) - The percentage of population ages 25 and over that attained or completed primary education.

Literacy rate, adult total (% of people ages 15 and above) - Adult literacy rate is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life.

### Economy & Growth:

| **Feature Name** | **Description** | **Type of Variable** | **Basic Statistics** |
|---|---|---|---|
| Country Name | Name of the country | Nominal | Examples:<br>Canada, Australia, Italy |
| Indicator Name | Types of various indicators monitored by World Bank | Nominal | Examples:<br>GDP (constant 2015 US$),<br>Exports of goods and services (constant 2015 US$) |
| 2001-2019 | Year for which the data is available | Categorical | 2015:<br>Min: -8.531 x $10^{14}$<br>1st Quartile: 10<br>Median: 1.224 x $10^{7}$ |

| | | | |
|---|---|---|---|
| | | | Mean: 1.14 x $10^{13}$<br>3rd Quartile: 2.32 x $10^{10}$<br>Max: 1.17 x $10^{16}$ |

GDP (constant 2015 US$) (Continuous Variable) - GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. Data are in constant 2015 prices, expressed in U.S. dollars.

Exports of goods and services (constant 2015 US$) (Continuous Variable) - Exports of goods and services represent the value of all goods and other market services provided to the rest of the world.

Imports of goods and services (constant 2015 US$) (Continuous Variable) - Imports of goods and services represent the value of all goods and other market services received from the rest of the world.

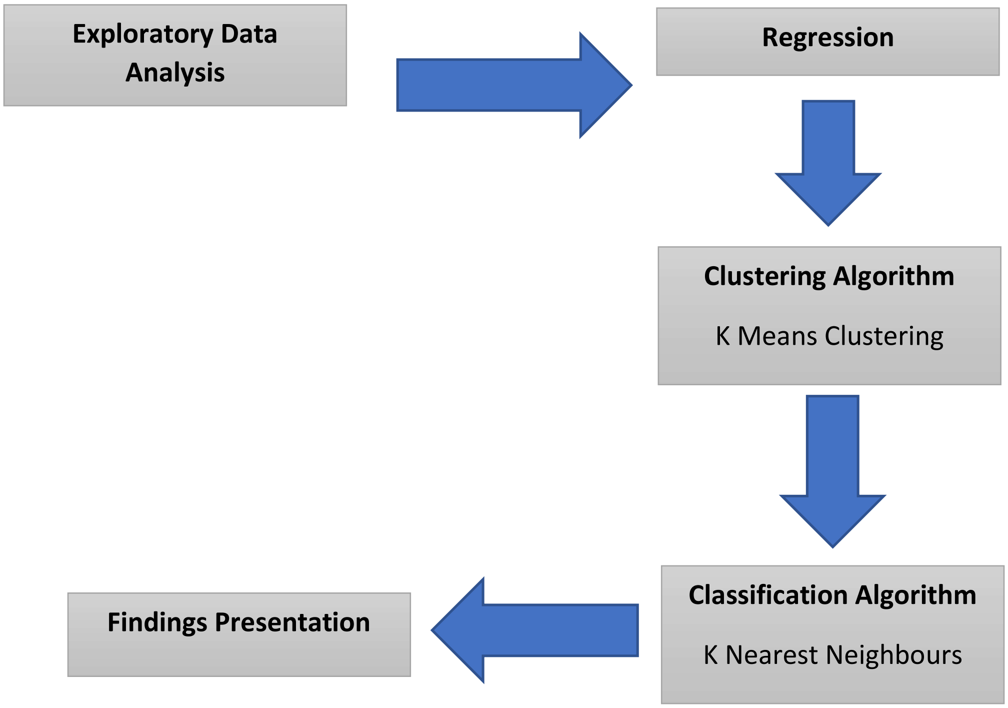
As the number of variables available in each of the datasets is more than 200, the study restricts the number to only a few variables relevant for this research. These variables are tabulated below:

| | |
|---|---|
| Health | 1. Life expectancy at birth, total (years)<br>2. People using at least basic sanitation services (% of population)<br>3. Physicians (per 1,000 people) |
| Education | 1. Literacy rate, adult total (% of people ages 15 and above)<br>2. Current education expenditure, total (% of total expenditure in public institutions) |
| Financial sector | 1. Net foreign assets (current LCU) |

| | |
|---|---|
| | 2. Account ownership at a financial institution or with a mobile-money-service provider (% of population ages 15+) |
| Economic Growth | 1. GDP (constant 2015 US$) (Continuous Variable) |

I have chosen only one variable of GDP in the Economic Growth dataset as I am assessing the impact of health, education and financial sector improvements on the GDP growth of countries. Thus, the GDP will be the dependant variable and others will be independent variables.

## Methodology and Tools to be used:



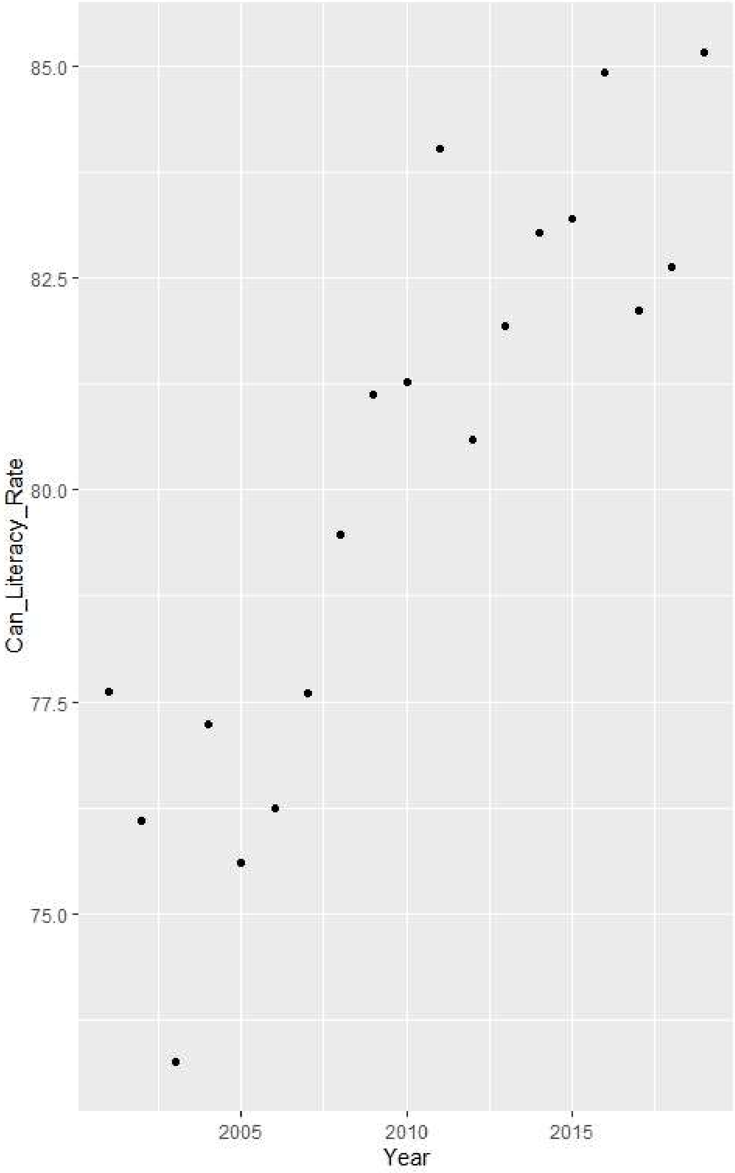
I plan to start with exploratory data analysis on the 4 datasets. This will help understand the basic statistical measures of central tendencies. This study will analyse the correlation between various attributes and their interlinkages using three different approaches – Regression, Clustering analysis and Classification analysis.

After exploring the dataset, I will carry out regression analysis and implement clustering algorithm and classification algorithm on the data points of these economies.

These analyses will yield details regarding the correlation of the data points and their impact on the GDP growth attribute. Thus, the dependent variable will be the GDP growth rate which I will analyse using the above-mentioned methods.

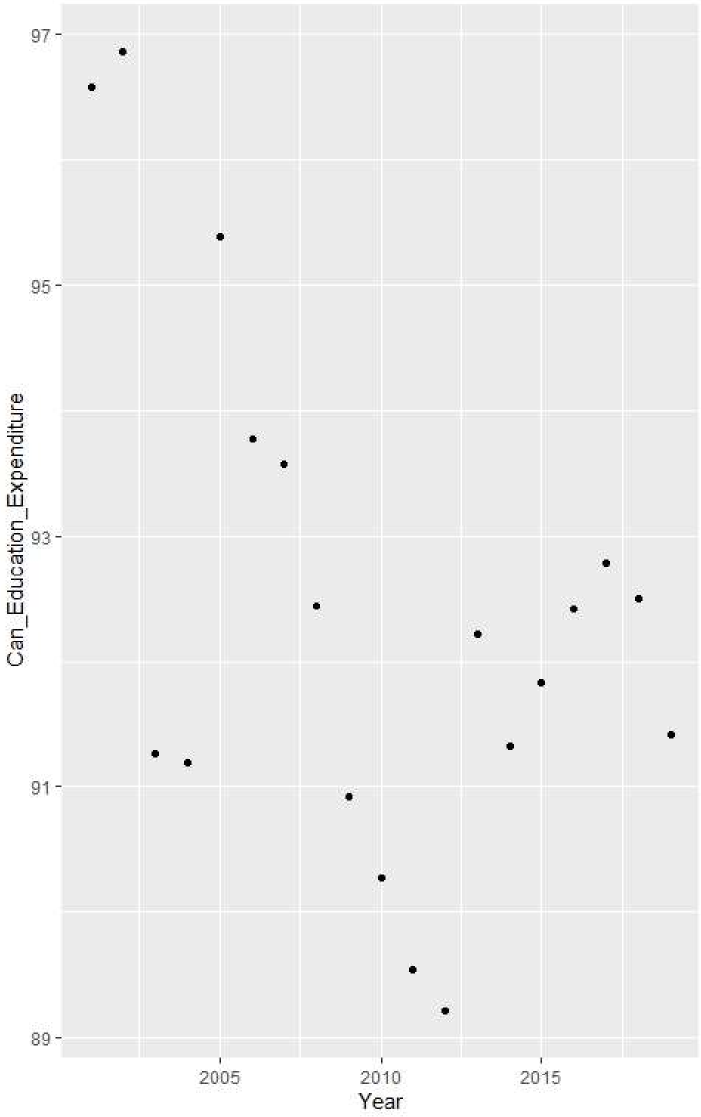
Finally, I will present the findings of the research including recommendations to increase the GDP growth rate of the countries.

## Descriptive Statistics:



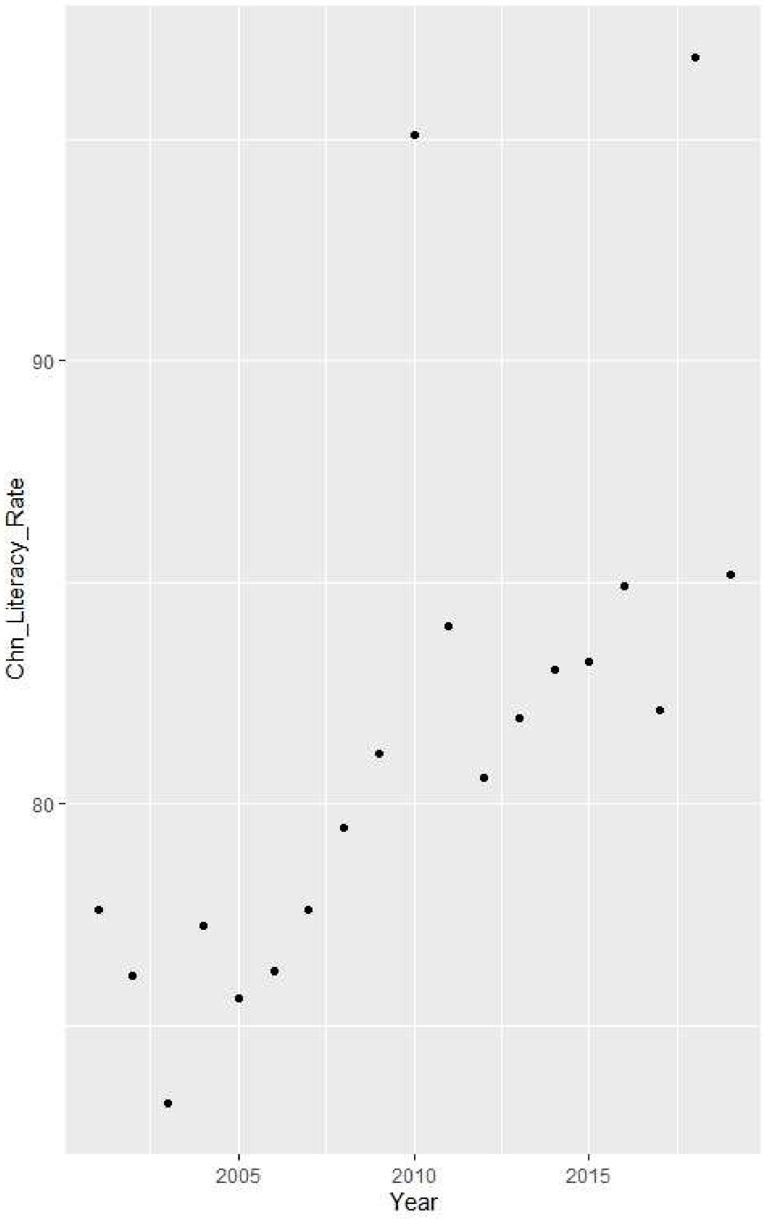
Canada Literacy Rate vs Year

This image shows the plot of data points between the year and the literacy rate in Canada.



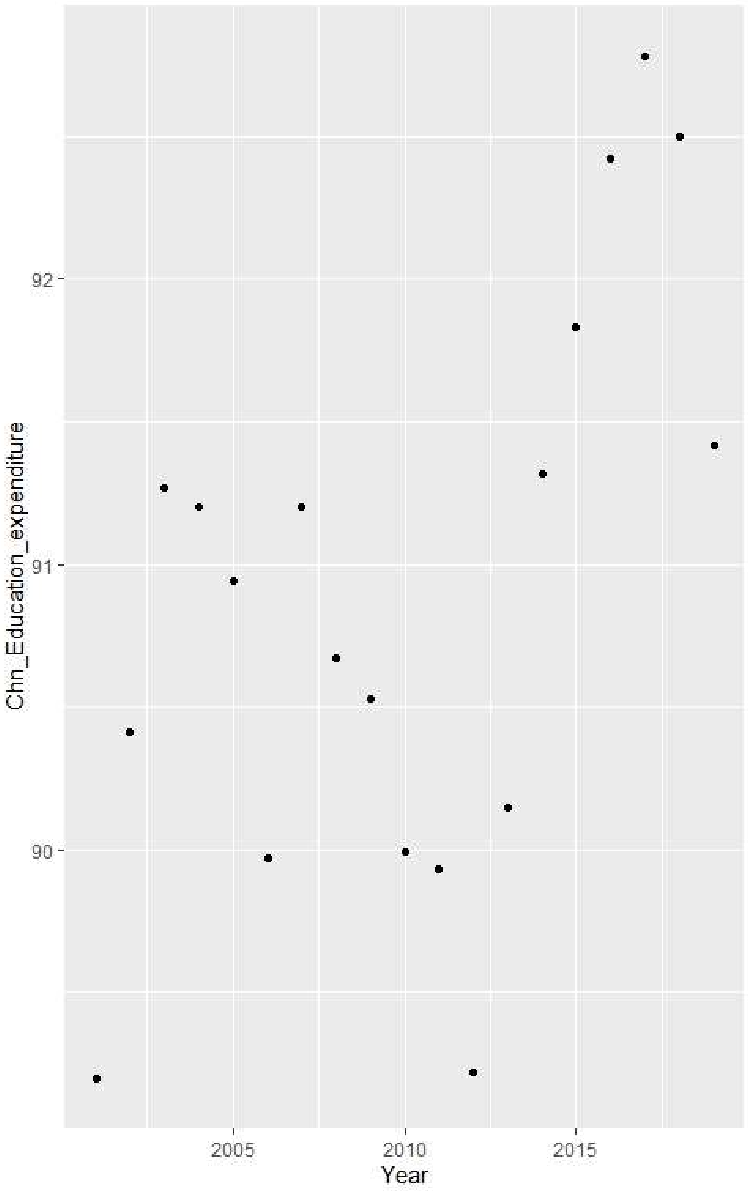
Canada Education Exposure vs Year

This image shows the plot of data points between the year and the education exposure in Canada.



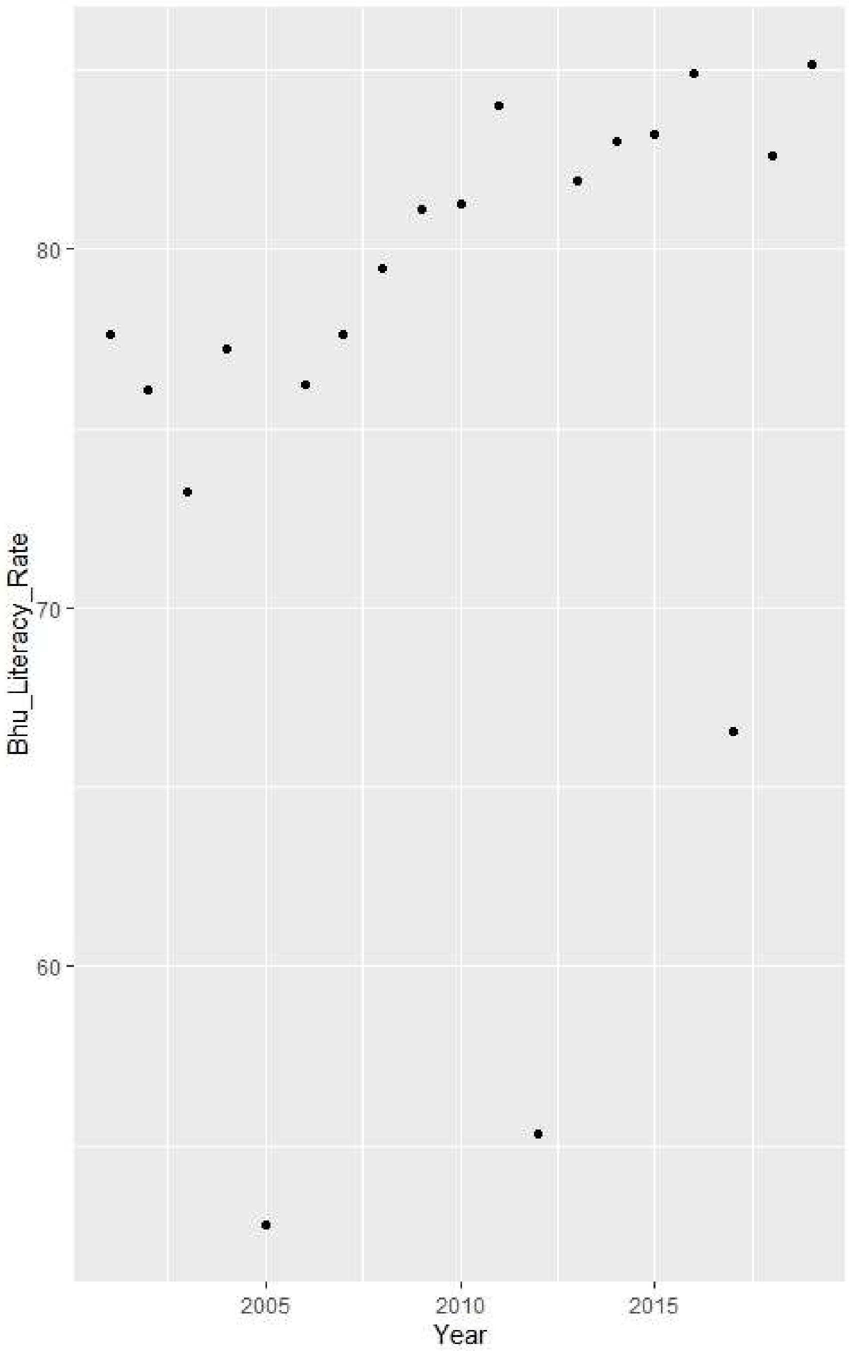
<u>China Literacy Rate vs Year</u>

This image shows the plot of data points between the year and the literacy rate in China.



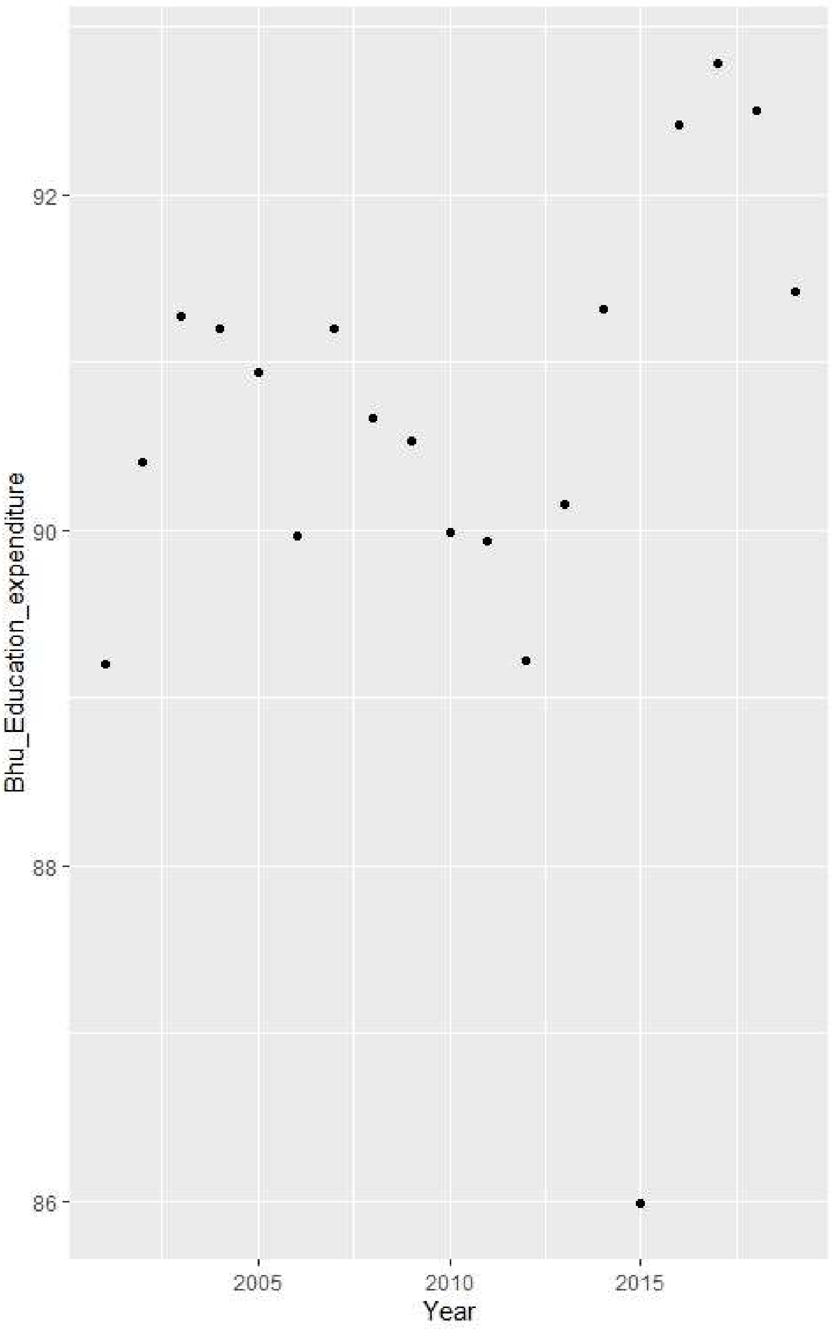
China Education Expenditure vs Year

This image shows the plot of data points between the year and the education exposure in China.



Bhutan Literacy Rate vs Year

This image shows the plot of data points between the year and the literacy rate in Bhutan.



Bhutan Education Expenditure vs Year

This image shows the plot of data points between the year and the education exposure in Bhutan.

The plots show that the data is widely distributed. The null values can be replaced with mean values.

## Regression

Regression is a form of analysis where an equation assesses the relationship between a set of independent variables and a single dependent variable. This method tries to fit a straight line among all the different data points. As the straight line cannot pass through all the data points, the model has an in-built error. However, it is one of the simplest forms of analysis to assess the dependency between variables.

The dependent variables are coded as follows in the model:

| | |
|---|---|
| Y_Canada_Education1: | Literacy Rate in Canada |
| Y_Canada_Education2: | Current Education Expenditure in Canada |
| Y_Canada_Health1: | Life expectancy at Birth in Canada |
| Y_Canada_Health2: | People using basic sanitation facilities in Canada |
| Y_Canada_Health3: | Physicians per 1000 people in Canada |
| Y_Canada_Financial_Sector1: | Bank Capital to Assets Ratio in Canada |
| Y_Canada_Financial_Sector2: | Percentage Inflation in Canada |

| | |
|---|---|
| Y_Bhutan_Education1: | Literacy Rate in Bhutan |
| Y_ Bhutan _Education2: | Current Education Expenditure in Bhutan |
| Y_ Bhutan _Health1: | Life expectancy at Birth in Bhutan |
| Y_Bhutan_Health2: | People using basic sanitation facilities in Bhutan |
| Y_Bhutan_Health3: | Physicians per 1000 people in Bhutan |
| Y_Bhutan_Financial_Sector1: | Bank Capital to Assets Ratio in Bhutan |
| Y_Bhutan_Financial_Sector2: | Percentage Inflation in Bhutan |

| | |
|---|---|
| Y_China_Education1: | Literacy Rate in China |
| Y_China_Education2: | Current Education Expenditure in China |
| Y_China_Health1: | Life expectancy at Birth in China |
| Y_China_Health2: | People using basic sanitation facilities in China |
| Y_China_Health3: | Physicians per 1000 people in China |
| Y_China_Financial_Sector1: | Bank Capital to Assets Ratio in China |
| Y_China_Financial_Sector2: | Percentage Inflation in China |

The regression method was carried out for 3 different datasets – one each for Canada, Bhutan and China. Summary of the regression equation is as follows:

GDP_Canada = 107.402717 – (1.015220 * Y_Canada_Health2) – (0.025809 * Y_Canada_Financial_Sector1)

GDP_Bhutan = (0.0684506 * Y_Bhutan_Health2) – (0.0915461 * Y_Bhutan_Health1)

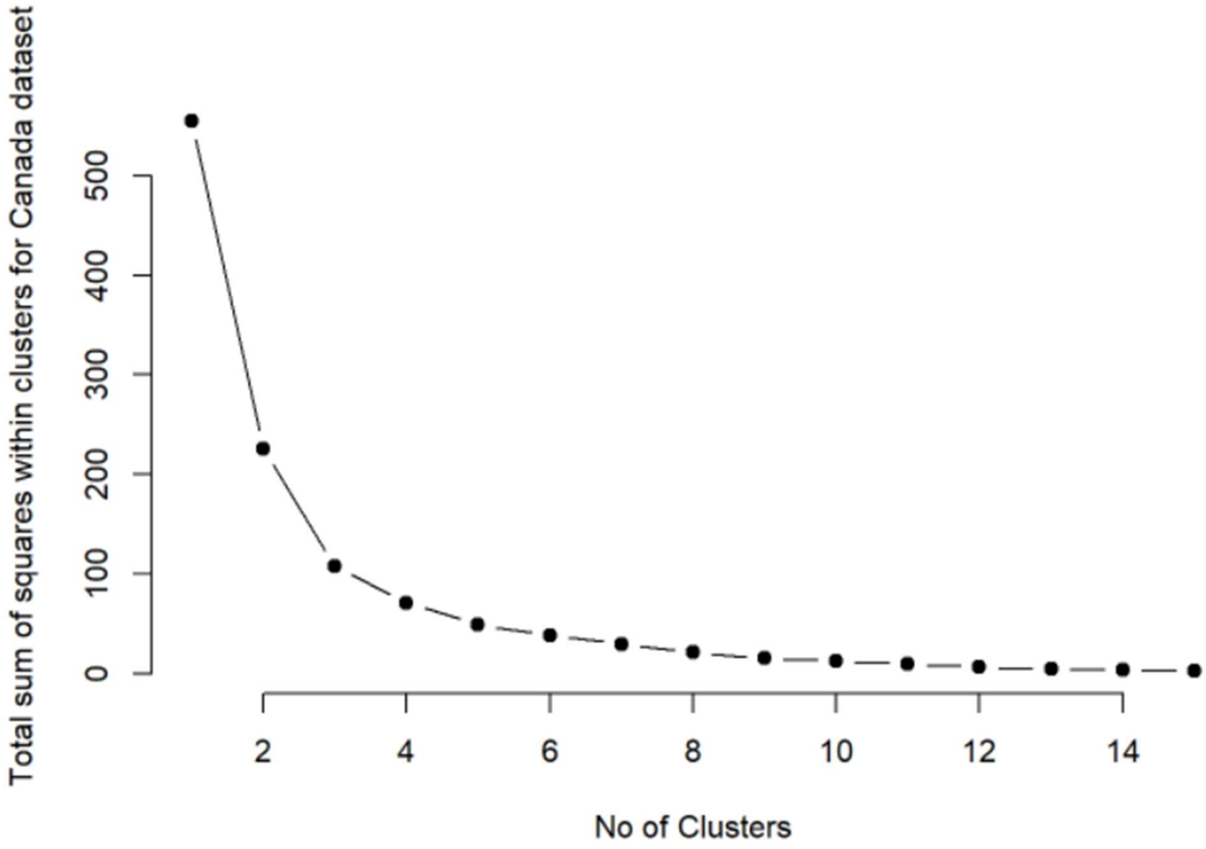
GDP_China = 0.01660 * Y_China_Financial_Sector1

The above equations consider the dependent factors at a confidence level of 95%

| **Country** | **Adjusted R-Squared** | **p-value** |
|---|---|---|
| Canada | 0.9838 | $3.881*10^{-10}$ |
| Bhutan | 0.9939 | $1.843*10^{-12}$ |
| China | 0.9948 | $7.25*10^{-13}$ |

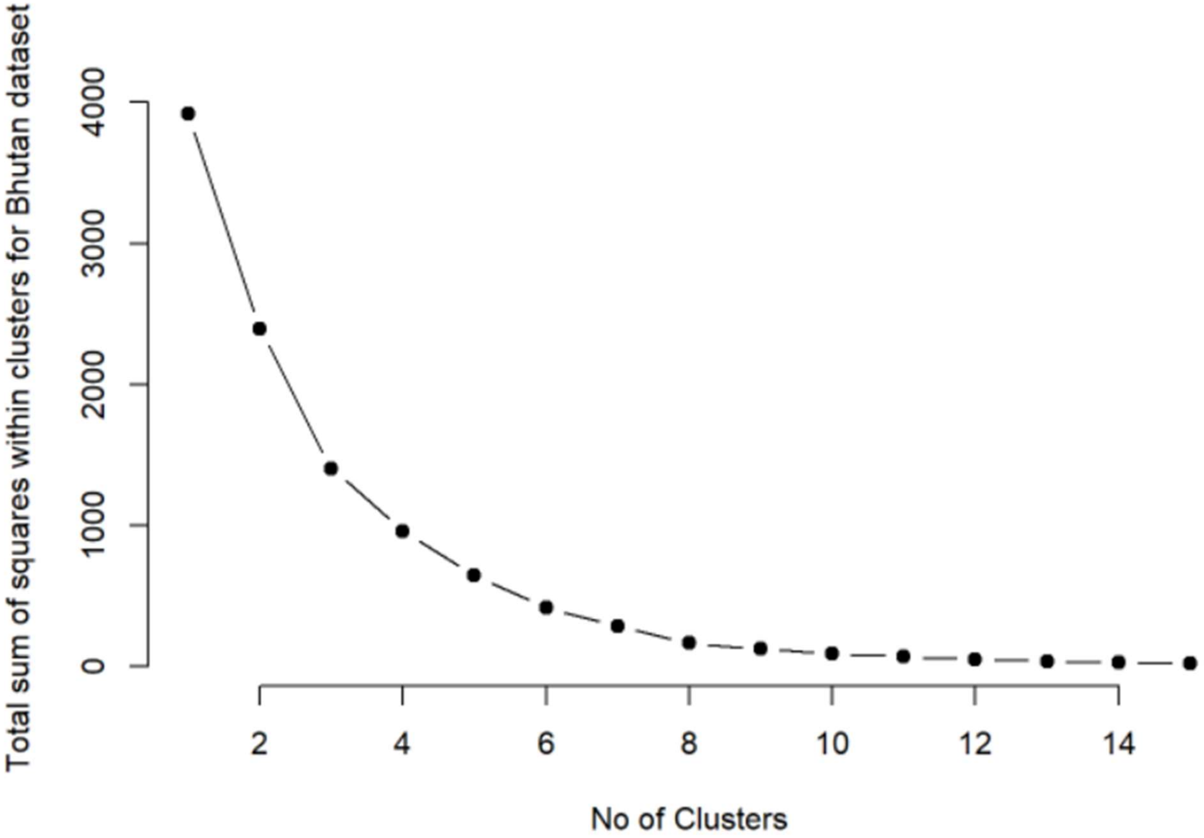
## K Means Clustering

K Means clustering is an unsupervised algorithm used for clustering analysis. This algorithm tries to group similar data points together which helps in understanding which specific groups contribute in a similar manner towards the change in the dependent variable. The ‘k’ in the model name specifies the total number of clusters that the data points are to be divided into. The ‘k’ value can also be the value where the total within cluster sum of squares is the least. This can be done using the elbow method where in the chart of the clusters vs within cluster sum of squares is plotted. However, the ‘k’ value can also be approximated as the integral value of square root of the total number of variables in the dataset.



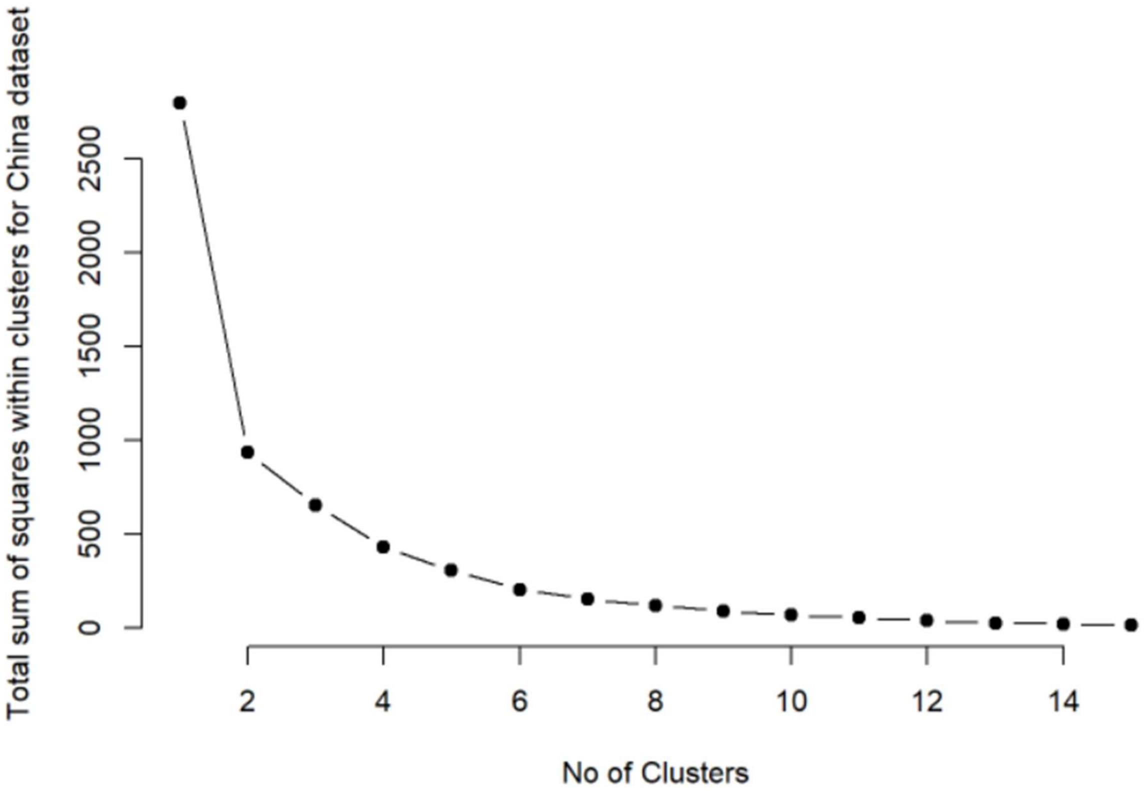
The figure depicts the plot between the total number of clusters and the total sum of square within the clusters for Canada dataset.

Note that the when the number of clusters increases beyond 4, the total sum of squares within the clusters does not decrease considerably. Hence, choose 4 clusters.



The figure depicts the plot between the total number of clusters and the total sum of square within the clusters for Bhutan dataset.

Note that the when the number of clusters increases beyond 6, the total sum of squares within the clusters does not decrease considerably. Hence, choose 6 clusters.



The figure depicts the plot between the total number of clusters and the total sum of square within the clusters for China dataset.

Note that the when the number of clusters increases beyond 6, the total sum of squares within the clusters does not decrease considerably. Hence, choose 6 clusters.

| Dataset | Total Sum of Squares | Total Clusters | Within Sum of Squares | Total Within Sum of Squares |
|---|---|---|---|---|
| Canada | 554.31 | 4 | 13.56, 3.23, 8.81, 45.14 | 70.76 |
| Bhutan | 3916.10 | 6 | 205.18, 25.85, 120.09, 0, 0, 63.66 | 414.80 |
| China | 2794.96 | 6 | 57.43, 26.58, 0, 67.89, 0, 52.0 | 203.92 |

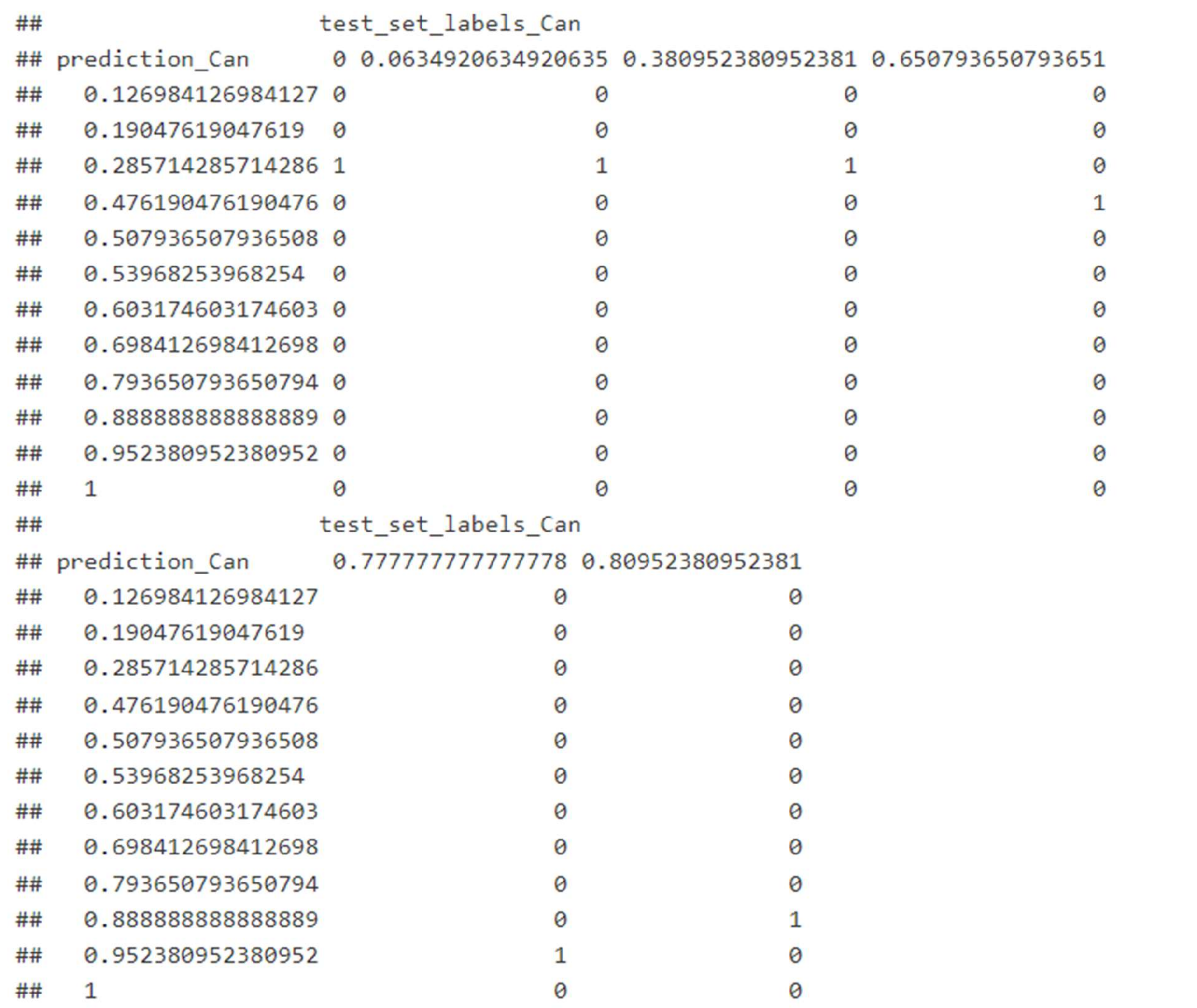
## KNN Classification

K-Nearest Neighbours (KNN) Classification is a supervised algorithm which is used for classification problems. Supervised algorithms require a training data set used to fit the model and a separate test data set on which the model is tested.

For the purposes of this study, the train-test split is done on the 70-30 scale. The total dataset is split into two parts – one containing 70% of the total data points used for training the KNN classification model and the other 30% data points used for testing the model. The model classifies 'k' nearest neighbours to group similar data points together. The value of k should be the integral value of the square root of total sample size.

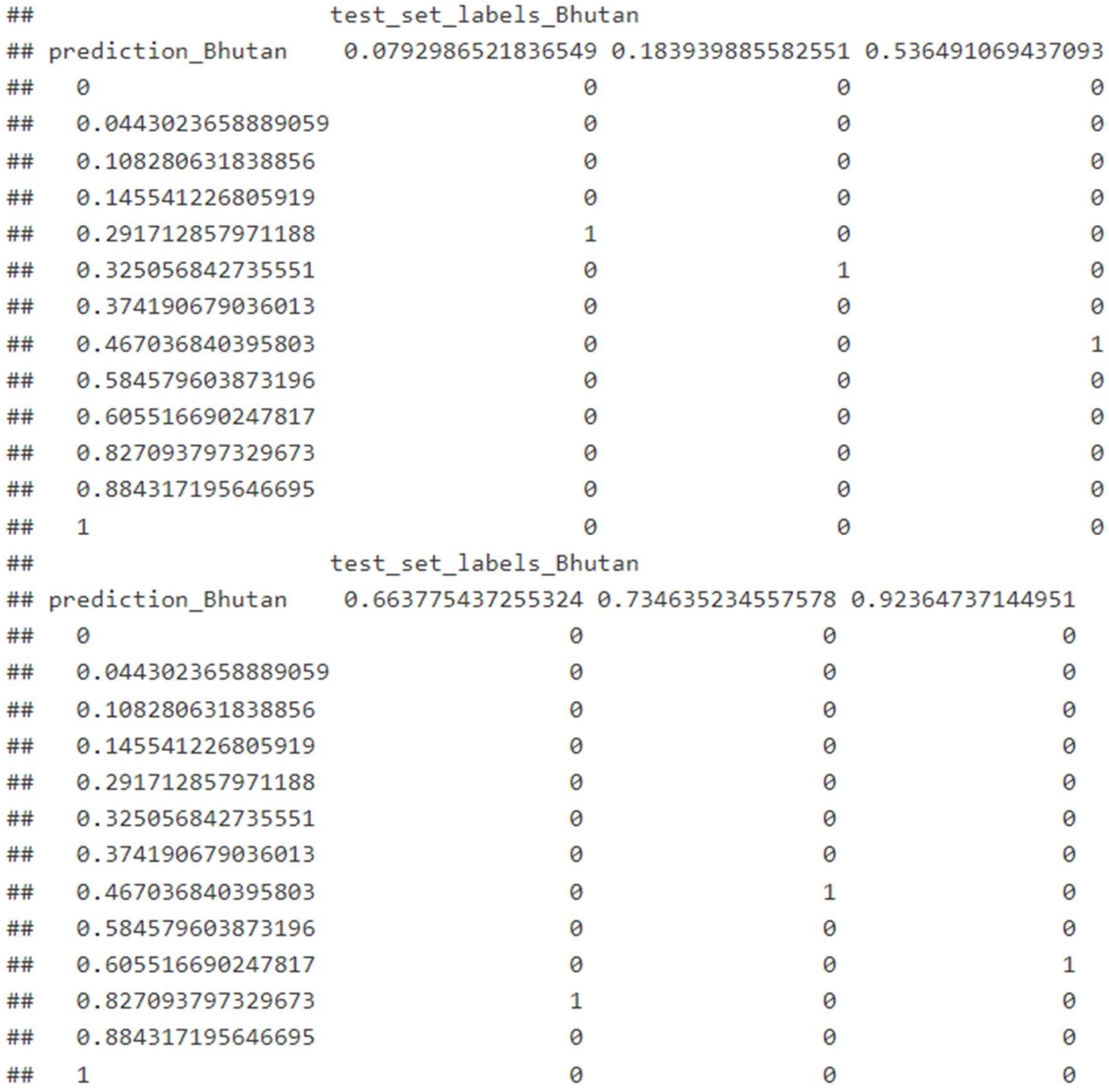
The KNN algorithm does not perform well with outliers and hence should be avoided if the data has too many outliers.

## Prediction for Canada dataset



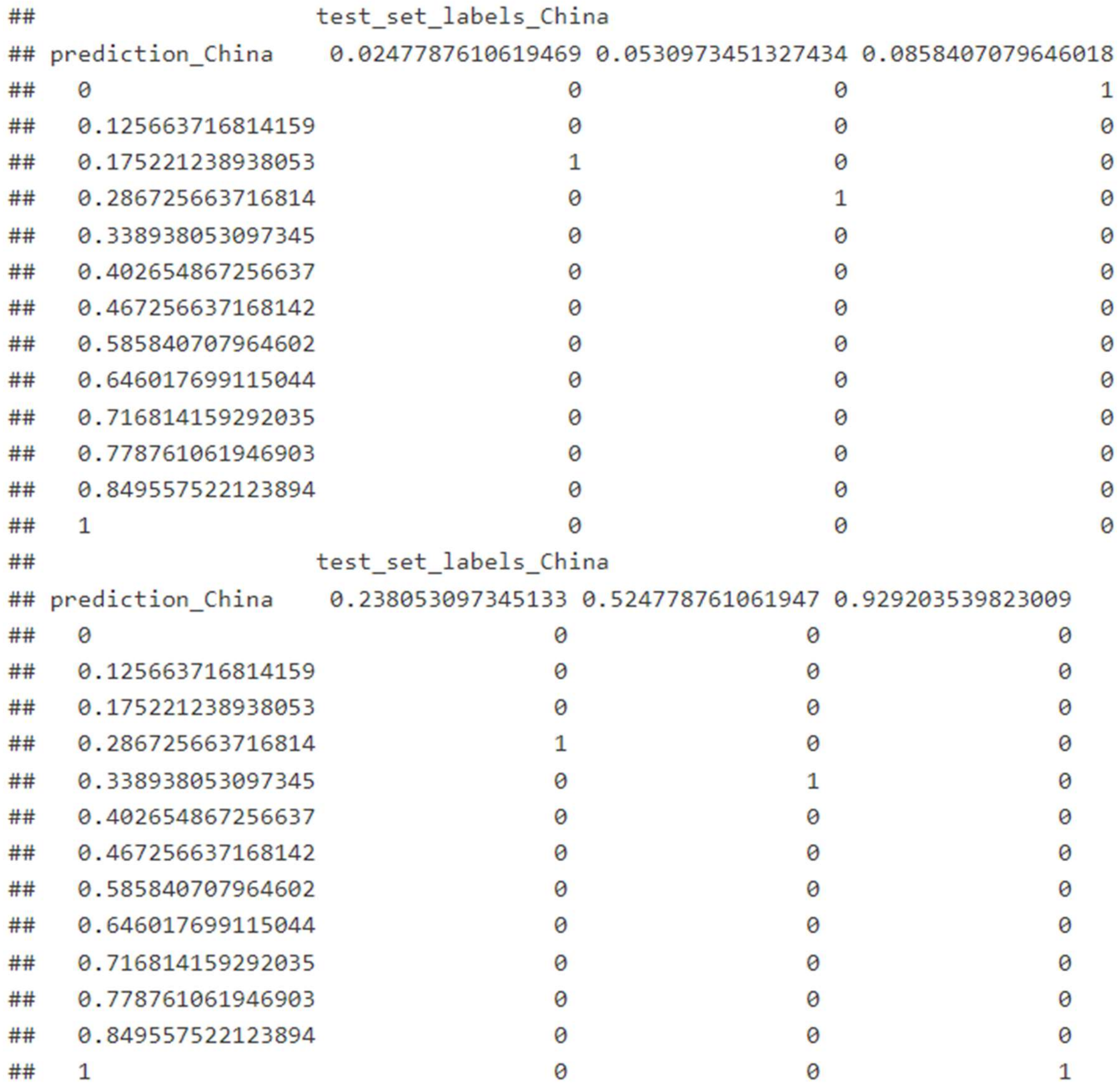
```
##                     test_set_labels_Can
## prediction_Can      0 0.0634920634920635 0.380952380952381 0.650793650793651
##   0.126984126984127 0                  0                 0                 0
##   0.19047619047619  0                  0                 0                 0
##   0.285714285714286 1                  1                 1                 0
##   0.476190476190476 0                  0                 0                 1
##   0.507936507936508 0                  0                 0                 0
##   0.53968253968254  0                  0                 0                 0
##   0.603174603174603 0                  0                 0                 0
##   0.698412698412698 0                  0                 0                 0
##   0.793650793650794 0                  0                 0                 0
##   0.888888888888889 0                  0                 0                 0
##   0.952380952380952 0                  0                 0                 0
##   1                 0                  0                 0                 0
##                     test_set_labels_Can
## prediction_Can      0.777777777777778 0.80952380952381
##   0.126984126984127                 0                0
##   0.19047619047619                  0                0
##   0.285714285714286                 0                0
##   0.476190476190476                 0                0
##   0.507936507936508                 0                0
##   0.53968253968254                  0                0
##   0.603174603174603                 0                0
##   0.698412698412698                 0                0
##   0.793650793650794                 0                0
##   0.888888888888889                 0                1
##   0.952380952380952                 1                0
##   1                                 0                0
```

## Prediction for Bhutan dataset



```
##                       test_set_labels_Bhutan
## prediction_Bhutan    0.0792986521836549 0.183939885582551 0.536491069437093
##   0                                   0                 0                 0
##   0.0443023658889059                  0                 0                 0
##   0.108280631838856                   0                 0                 0
##   0.145541226805919                   0                 0                 0
##   0.291712857971188                   1                 0                 0
##   0.325056842735551                   0                 1                 0
##   0.374190679036013                   0                 0                 0
##   0.467036840395803                   0                 0                 1
##   0.584579603873196                   0                 0                 0
##   0.605516690247817                   0                 0                 0
##   0.827093797329673                   0                 0                 0
##   0.884317195646695                   0                 0                 0
##   1                                   0                 0                 0
##                       test_set_labels_Bhutan
## prediction_Bhutan    0.663775437255324 0.734635234557578 0.92364737144951
##   0                                  0                 0                0
##   0.0443023658889059                 0                 0                0
##   0.108280631838856                  0                 0                0
##   0.145541226805919                  0                 0                0
##   0.291712857971188                  0                 0                0
##   0.325056842735551                  0                 0                0
##   0.374190679036013                  0                 0                0
##   0.467036840395803                  0                 1                0
##   0.584579603873196                  0                 0                0
##   0.605516690247817                  0                 0                1
##   0.827093797329673                  1                 0                0
##   0.884317195646695                  0                 0                0
##   1                                  0                 0                0
```

## Prediction for China Dataset



```
##                     test_set_labels_China
## prediction_China     0.0247787610619469 0.0530973451327434 0.0858407079646018
##   0                                   0                  0                  1
##   0.125663716814159                   0                  0                  0
##   0.175221238938053                   1                  0                  0
##   0.286725663716814                   0                  1                  0
##   0.338938053097345                   0                  0                  0
##   0.402654867256637                   0                  0                  0
##   0.467256637168142                   0                  0                  0
##   0.585840707964602                   0                  0                  0
##   0.646017699115044                   0                  0                  0
##   0.716814159292035                   0                  0                  0
##   0.778761061946903                   0                  0                  0
##   0.849557522123894                   0                  0                  0
##   1                                   0                  0                  0
##                     test_set_labels_China
## prediction_China     0.238053097345133 0.524778761061947 0.929203539823009
##   0                                  0                 0                 0
##   0.125663716814159                  0                 0                 0
##   0.175221238938053                  0                 0                 0
##   0.286725663716814                  1                 0                 0
##   0.338938053097345                  0                 1                 0
##   0.402654867256637                  0                 0                 0
##   0.467256637168142                  0                 0                 0
##   0.585840707964602                  0                 0                 0
##   0.646017699115044                  0                 0                 0
##   0.716814159292035                  0                 0                 0
##   0.778761061946903                  0                 0                 0
##   0.849557522123894                  0                 0                 0
##   1                                  0                 0                 1
```

## Summary

This study has analysed various factors of three datasets – Health, Education and Financial Sector and their impact on the growth of the GDP of three economies.

The regression analysis provided insights regarding which factors majorly affect the GDP growth. The K Means clustering analysis provided a set of correlated parameters which affect the GDP of the countries. The KNN classification algorithm also informed the different parameters which influence the GDP in a similar manner.

This study approached the analysis using three different approaches. While one of these three methods may be more precise and useful than the others, this study provided a good comparison of different methods of analysis. The unsupervised learning technique - K means clustering - has proved to be a better measure among the three methods used in this study.

The original set of questions will be attempted to be answered:

1. How strongly correlated are the factors within education, health and financial sector to the GDP of an economy?

A. Out of the several factors considered for this study, it can be said that the Financial_Sector1, Health_1 and Health_2 are the strongly correlated attributes foreach of the economies.

2. Which are the top 3 attributes that affect the GDP growth either positively or negatively?

A. The top 3 attributes contributing to the GDP growth are financial capital to assets ratio, life expectancy at birth and basic sanitation facilities.

3. What measures should be implemented to increase GDP in the countries selected for study?

A. As per this study, to increase the GDP of Canada, people with basic sanitation facilities should be lowered. Also, the total lending by the banks on existing capital should be increased.

To increase the GDP of Bhutan, there should be an increase in the basic sanitation facilities provided to people, while life expectancy can be reduced.

To increase the GDP of China, the total funding by the banks on existing capital should be increased.

While this study recommends the above aspects for the overall GDP growth of the selected 3 countries, there are other parameters outside the scope of this study that directly or indirectly influence the GDP growth. Thus, the above answers to the questions should be considered in line with other aspects.

Clustering analysis can be used to have even further understanding regarding the interplay of various attributes. Techniques such as DBScan analysis can help understand the distribution of the data and the density connectivity and inform better insights.

## References:

Aziz, R. N. A. R., & Azmi, A. (2017). Factors affecting gross domestic product (GDP) growth in Malaysia. *International Journal of Real Estate Studies*, 11(4), 61-67. https://www.utm.my/intrest/files/2017/09/07-FACTOR-AFFECTING-GROSS-DOMESTIC-PRODUCT-GDP-GROWTH-IN-MALAYSIA1.pdf

Beck, T., & Levine, R. (1999). *A new database on financial development and structure* (Vol. 2146). World Bank Publications.

Chowdhury K. (2003). Empirics for World Income Distribution: What Does the World Bank Data Reveal? *The Journal of Developing Areas*, *36*(2): 59–83. http://www.jstor.org/stable/4192920

Cooray A. (2009). The Financial Sector and Economic Growth. *Economic Record* https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2009.00584.x

Deme, M., & Mahmoud, A. M. A. (2020). Effect of Quantity and Quality of Education on Per Capita Real-GDP Growth: Evidence from Low- and Middle-Income African Countries. Applied Economics, 52:57, 6248-6264. https://doi.org/10.1080/00036846.2020.1789058

Eggoh, J., Houeninvo, H., & Sossou, G. A. (2015). Education, health and economic growth in African countries. Journal of Economic Development, 40(1), 93. https://www.researchgate.net/profile/Hilaire-Houeninvo/publication/282646931_EDUCATION_HEALTH_AND_ECONOMIC_GROWTH_IN_AFRICAN_COUNTRIES/links/5615219208aec62244119a7c/EDUCATION-HEALTH-AND-ECONOMIC-GROWTH-IN-AFRICAN-COUNTRIES.pdf

Grant Catherine (2017). The contribution of education to economic growth. *Knowledge, evidence and learning for development* K4D_HDR_The_Contribution_of_Education_to_Economic_Growth_Final.pdf (publishing.service.gov.uk)

Hongyi, L. I., & Huang, L. (2009). Health, education, and economic growth in China: Empirical findings and implications. *China Economic Review*, 20(3), 374-387. https://doi.org/10.1016/j.chieco.2008.05.001

Indicators | Data. (2020). The World Bank. https://data.worldbank.org/indicator?tab=featured

Karavaeva, K. (2021). The Relationship between Sustainable Development and GDP Growth in EU Countries. *Charles University* http://hdl.handle.net/20.500.11956/126546

Marquez-Ramos, Laura. (2019) Education and Economic Growth: An Empirical Analysis of Nonlinearities. *Applied Economic Analysis* https://www.emerald.com/insight/content/doi/10.1108/AEA-06-2019-0005/full/html