Knowledge Discovery for Social Good: A Detailed Analysis of Economic Growth Using Heterogeneous Datasets

1st Anish Sethi
Information Technology and Web Science
Rensselaer Polytechnic Institute
Troy, USA
sethia@rpi.edu

2nd Thilanka Munasinghe
Information Technology and Web Science
Rensselaer Polytechnic Institute
Troy, USA
munast@rpi.edu

Abstract—**Knowledge discovery for social good is a challenging task, especially when combining data from heterogeneous data sources. In this paper, we present our work on analyzing the economic growth of 28 countries based on various datasets containing education level, employment, government expenditure, and gross value added by various sectors. While performing the analysis, we encountered several challenges due to missing data and the nature of the multiple non-uniform indicators used. We utilized knowledge representation mechanisms to overcome these challenges and utilized several clustering methods, such as agglomerative clustering, mean-shift clustering, and birch clustering in our analysis. We also utilized dimensionality reduction techniques to lower the dimensions and find the two most correlated principal components that would be based on clustering on countries' growth in a specific year. Unsurprisingly, we observed that education is a significant positive indicator of direct and indirect economic growth since an increase in education quality leads to better employment and eventually increases a country's Gross Domestic Product (GDP). However, during this analysis, it became clear the dearth of data engineering mechanisms that cater to diverse data sources available in sustainable development goals. The work outlined in this paper could act as a guide for someone wishing to perform a similar analysis using such heterogeneous data to advance state of the art in knowledge discovery for social good.**

Index Terms—**Education, Growth, Economy, Unemployment, Gender Parity, Agglomerative Hierarchical Clustering, Birch Clustering, Mean Shift Clustering, KNN Clustering, Principle Component Analysis, Sustainable Development Goals, Social Good, Knowledge Discovery, Data Engineering**

I. Introduction

Educating the youth of a nation is an essential factor in terms of economic growth. Consistent economic growth is vital to a nation's progress; it helps reduce unemployment and brings political stability; it promotes socio-economic development. It is exceptionally critical in developing countries where there are ample opportunities to grow. A well-educated society helps individuals, families, and the country as a whole to grow continuously. This paper investigates the relationship between education growth and the effects of different economic attributes that influence development. Based on our initial analysis of the effects of primary, secondary, and tertiary education on a nation's GDP, it is strongly felt that both education and economic factors are good indicators of a country's growth. Previous preliminary analysis using a linear model showed an accurate GDP prediction based on Education and growth based on sectors. We also found out that GDP is highly correlated to tertiary education and value-added by most countries' service sectors.

Our analysis focuses on the effects of enrollment in education at different levels and individual aspects of a nation's economy like value-added by sectors, sector-wise employment of people, and type of national expenditure such as import and export. Unlike existing research and analysis [1] [2], our research focuses on the relationship between different genders in terms of growth instead of general health and well being.

According to research done by Dominic J. Brewer and Patrick J. McEwan in the Economics of Education (2010) [3], the quality of education is measured based on cognitive skills. Societies with higher cognitive skills play an essential role when assessing economic growth. Also, it means that economic growth depends on workers' skills, and in developing countries, high enrollment in schools does not mean a high literacy rate. This makes it worthy of analyzing whether school enrollment by itself is an appropriate indicator of a country's economic growth. According to Eric Hanushek and Ludger Woessmann [4], this is one of the crucial features that developing countries look after when assessing growth.

We analyzed data collected from the United Nations and World Bank about Education, Growth, and Gross Value added by sectors. It is generally the case that education and GDP are always positively correlated because more people find employment with an increase in education level. Since employers are ready to provide highly specialized jobs with high pay, the spending or consumption increases, and effectively it improves the GDP.

In the 2009 report on Gender equality, economic growth, and employment [5], Åsa Löfström states that without gender parity, sustainable development and growth could not be achievable. It was observed that GDP per capita is affected strongly by gender equality, and more women in the government sector could promote the economic growth of a nation. Hence we feel it is imperative that we include the effects of gender disparity in school enrollment and employment when analyzing the growth of a nation.

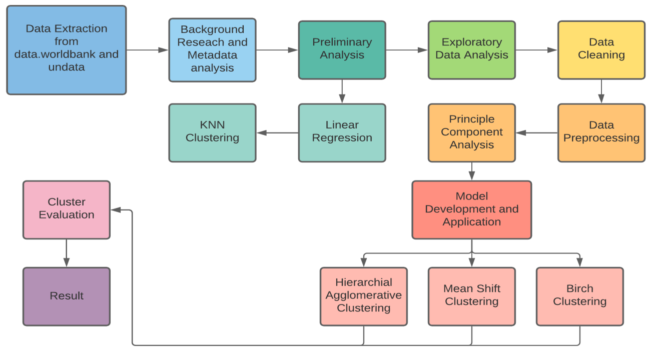

Data Extraction from data.worldbank and undata
Background Reseach and Metadata analysis
Preliminary Analysis
Exploratory Data Analysis
Data Cleaning
KNN Clustering
Linear Regression
Principle Component Analysis
Data Preprocessing
Cluster Evaluation
Model Development and Application
Result
Hierarchial Agglomerative Clustering
Mean Shift Clustering
Birch Clustering

Fig. 1. Workflow Diagram

The workflow diagram (Fig.1) explains the multiple processes undertaken for this analysis and the steps performed to obtain the result. The following were the results of the Preliminary Analysis - For countries such as India, China, USA, and Greece, there was a significant increase in the GDP predicted using Linear regression with an increase in tertiary education, Syria on the other hand, correlated with the industrial sector and hence did not show a high predicted GDP. Clustering of the data was based on the predicted GDP into countries with low, average, and high growth based on the education levels of the country was done using KNN clustering [6] algorithm with an accuracy of 81.7%.

II. Data Collection

The data was compiled using information from 187 countries by the United Nations and the world bank. It spans a total of 49 years beginning from 1970 to 2019 and consists of data about employment to population ratios, enrollment in schools by gender as a percentage of population [7], employment in various sectors by gender, government consumption, GDP, GDP per capita, Literacy rates, Gross national income and GDP by type of expenditure and gross value added (in US dollars).

III. Data pre-processing and Exploratory Data Analysis

Out of the 187 countries that we selected, 28 countries of interest, namely - Bangladesh, Belgium, Bolivia, Brazil, China, Colombia, Denmark, Germany, Ghana, Greece, Iceland, India, Iran, Islamic Republic of Iraq, Japan, Lebanon, Libya, New Zealand, Norway, Pakistan, Singapore, Sri Lanka, Sweden, Switzerland, Syrian Arab Republic, United Arab Emirates, United Kingdom, and the United States. The data was merged into one data set on the basis of country and year, and attributes chosen for analysis were - Primary, Secondary and Tertiary Enrollment for males and females as a percentage of the population, GDP per capita (in US Dollars), Value added by economic activity and type of expenditure by the GDP. While calculating the value added by economic activity, gross value added by each economic activity was converted to a percentage of total gross value added.

A correlation plot (Fig.2) of the data-set shows that the economic indicators are highly correlated amongst themselves and average correlation with enrollment by gender and levels. It is also noticeable that the value added by the agriculture sector correlates negatively with education, whereas other activities (ISIC-JP) correlate highly. This is in line with our previous analysis, where employment in the agriculture sector did not promote GDP growth highly, which was dependent on tertiary education.

A. *Government Consumption and GDP Per Capita*

Initial Analysis of the government spending data shows us that for most developing countries such as Brazil, Bolivia, Colombia, India, government spending has increased sharply in the decades after 1990 due to various factors such as lax economic policies regarding trade and introduction of ICT industries which contributed to the economic boom [8]. The only exception is Bangladesh (Fig.3), which we saw a small spike right before the 1990s. Bolivian spending and GDP per capita increased considerably with their curbing of the hyperinflation through the later 20th Century and a gradual increase in GDP per capita was visible in 2005 from $1034 to $3548 in 2018, whereas government spending in Brazil increased exponentially after the 1990s and went from 62.4 Billion Brazilian Real in 1994 to 655 Billion Brazilian Real in 2009 (Fig.4). Colombia, on the other hand, experienced an economic boom in 1990 due to its liberal economic policies and went from $120 billion in 1990 to $750 billion in 2010 shown in (Fig.5) [9].

Developed countries like Sweden, Denmark, Germany, USA, Switzerland, Iceland, Belgium, and Singapore generally saw a constant increase in the past 50 years (Fig.6). We saw a sudden drop in spending expenditure in some European countries such as Germany, Belgium and Greece after the after those countries began to adopt Euro currency in 2000, as shown in (Fig.7) for Belgium.

For almost all countries, an increase in government spending resulted in a subsequent increase in GDP per capita in the coming years, as shown in Germany (Fig.8). This is because in times of slow economic growth, government spending often leads to increased economic activity, and a rise in tax would eventually lead to a higher GDP [10].

B. *Employment by Sector*

A higher number of females are employed in the agriculture and in the industry sectors, with agriculture being the biggest employer of about 40% females and 50% males in developing countries such as Bolivia (shown in Fig.9), Colombia, Ghana, Sri Lanka, India, and Pakistan whereas a meager 4% of the population (Male and Female) are employed in the agriculture sector by developed countries like Germany (shown in Fig.10), Belgium, US, United Kingdom, Singapore, and Denmark.

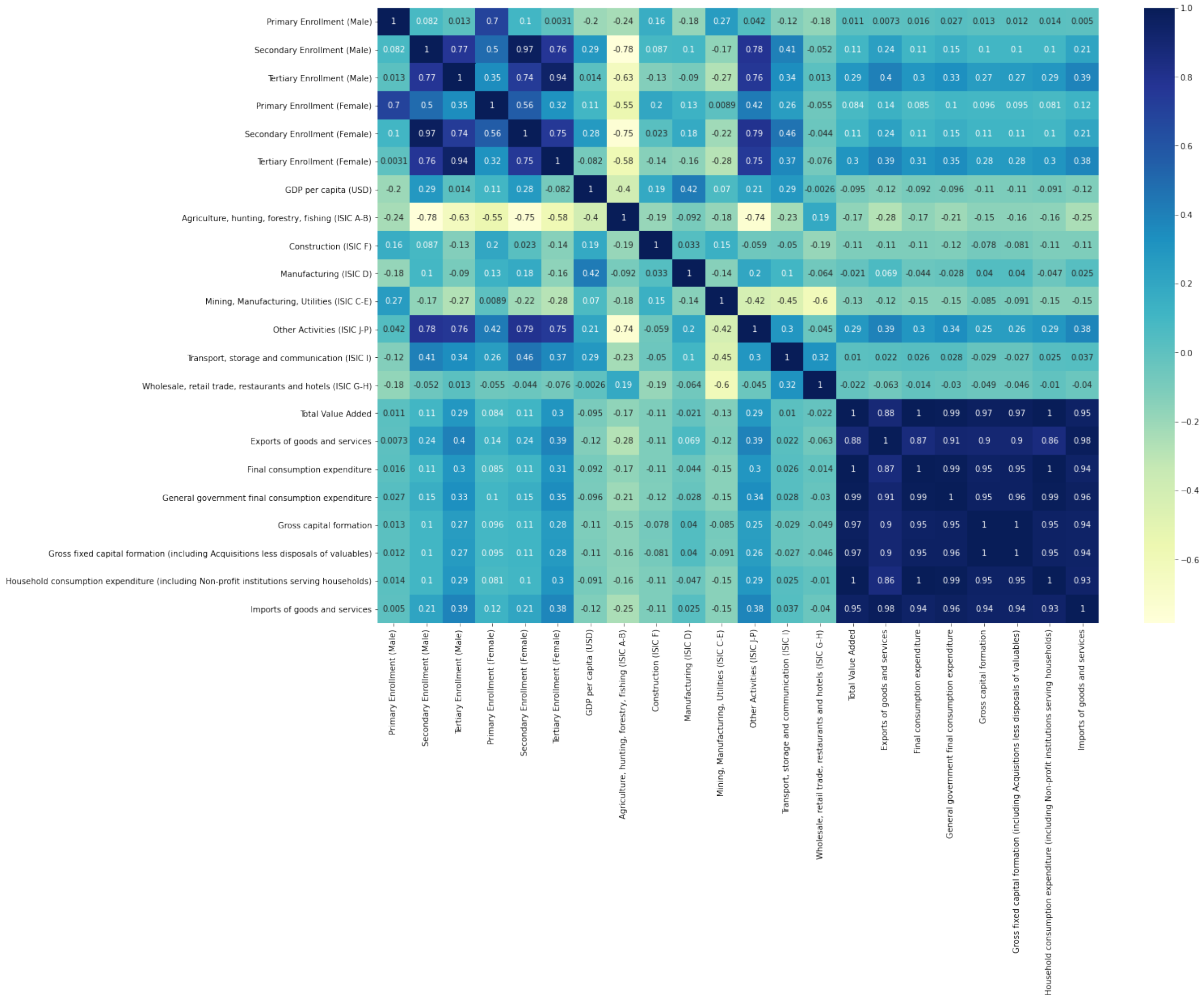


Fig. 2. Correlation between attributes

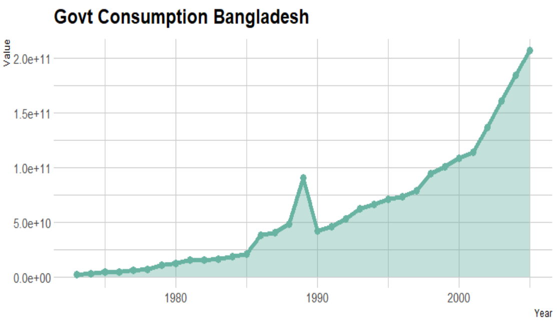


Fig. 3. Government consumption for Bangladesh (in BDT)

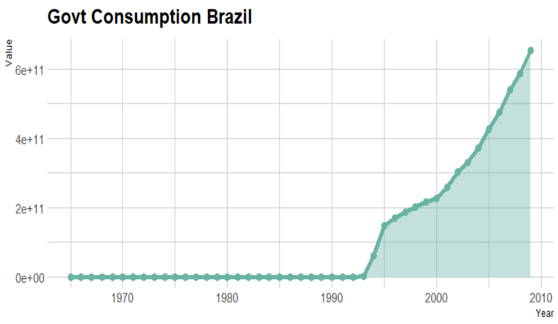


Fig. 4. Government consumption for Brazil (in BRL)

In those countries, employment in the agriculture sector has dropped drastically in the past decade.

Industrial sector employs less than 15% of the women in developed countries (shown in Fig.11) and close to 25% in developing countries (shown in Fig.12). During this time period we also notice a slight drop in employment in the industry sector and a huge increase (0-40%) in the agriculture sector in the developing nations (Fig.9 and Fig.12). Male employment in the industrial sector is similar in all countries with 30% employed. Bangladesh is the only country to have

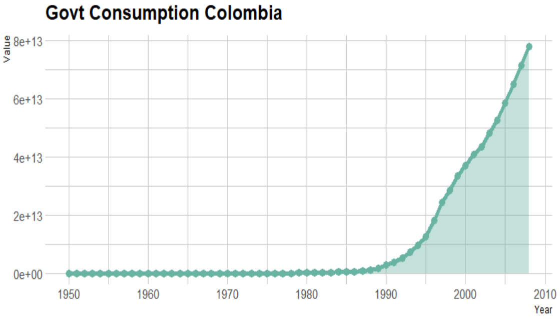


Fig. 5. Government consumption for Colombia (in COP)

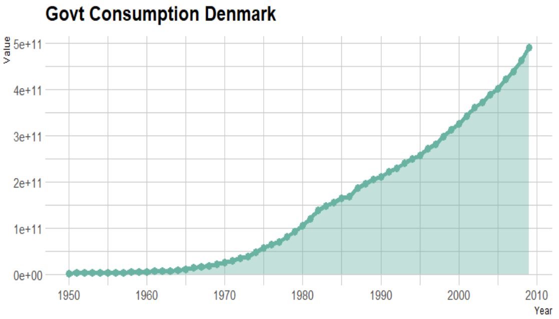


Fig. 6. Government consumption for Denmark (in DKK)

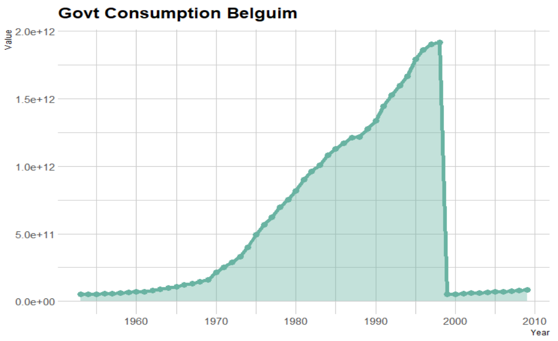


Fig. 7. Government consumption drops in Belgium after adopting EUR

more women employed in the industry sector and to have an increase in percentage employment as well (Fig.13).

Service sector employment accounts for about 60% of males in developed (Fig.14) and 40% developing countries (Fig.15) and is constantly rising in accordance to the drops in the agriculture and industry sector employment.

The service sector is the largest employer of women in Brazil, with more than 70% of women working in the service sector over the past 30 years. A trend is seen with the service sector, hiring about 80% women in all countries, but this trend continues to increase constantly in only developed countries.

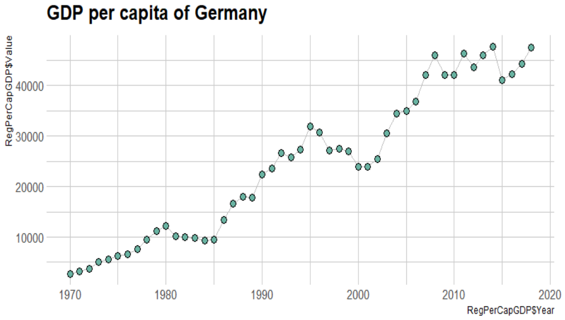


Fig. 8. GDP per capita of Germany over time (in USD)

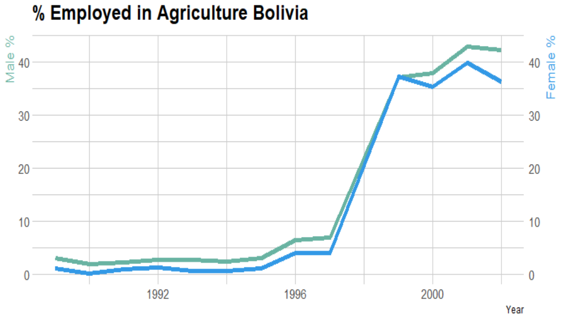


Fig. 9. Employment in Agriculture Bolivia

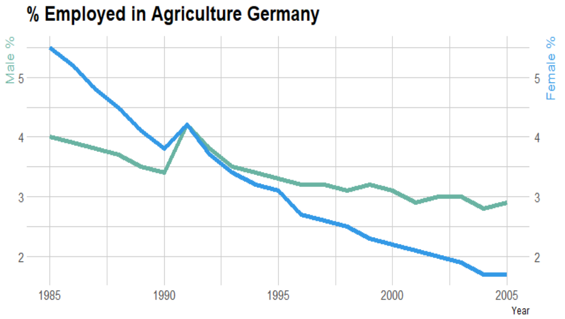


Fig. 10. Employment in Agriculture Germany

In Colombia and Bolivia, employment drops around the year 2000 for males and females (in the case of Bolivia), and the difference is seen in the slight increase in employment in the agriculture and industry sectors. Bolivia also employed most women in the service sectors at 90% and 60% males, which later drops to 60% and 40%, respectively (Fig.16). Service sector employment accounts for about 60% of males in developed (Fig.14) and 40% developing countries (Fig.15) and is constantly rising in accordance to the drops in the agriculture and industry sector employment.

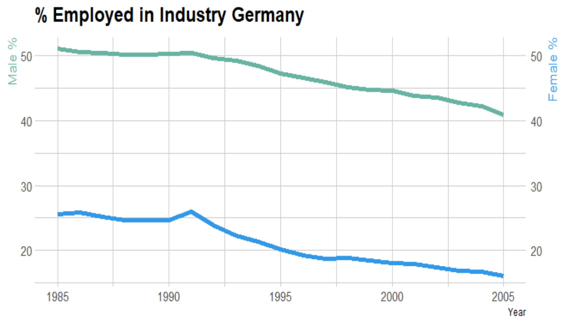


Fig. 11. Employment in Industry Germany

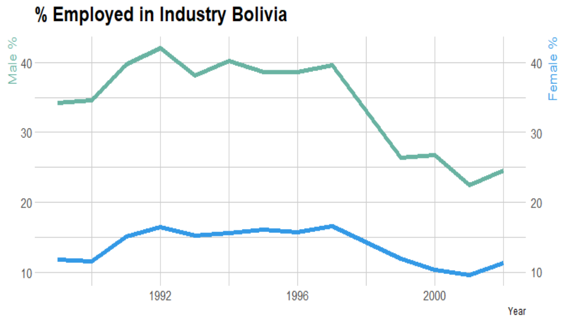


Fig. 12. Employment in Industry Bolivia

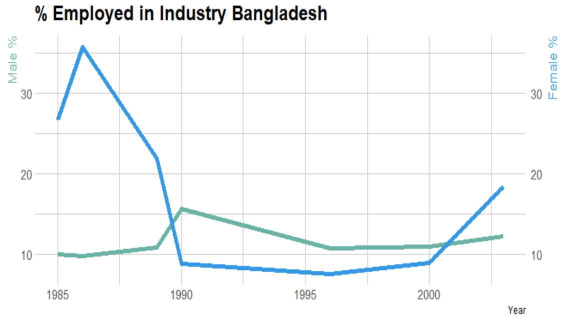


Fig. 13. Employment in Industry Bangladesh

C. Enrollment in schools by gender

Primary school enrollment in most countries reflects the increase in employment and correlates greatly with increases in GDP. Notable exceptions include Bangladesh and Brazil (Fig.17) with a sudden drop in enrolment in primary schools in the early 2010s. Apart from that, a constant enrollment is seen in most developed countries such as Belgium and Denmark (Fig.18). Most nations have a slightly higher male enrollment rate in primary schools than females with the one exception of India, where it is quite high until 2010 and then falls lower than female enrollment (Fig.19).

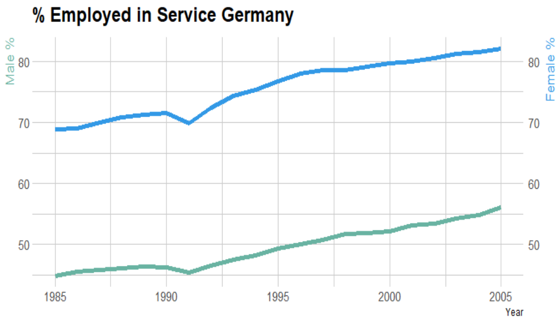


Fig. 14. Employment in Service Germany

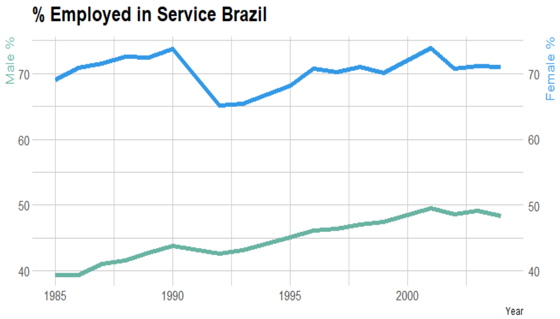


Fig. 15. Employment in Service Brazil

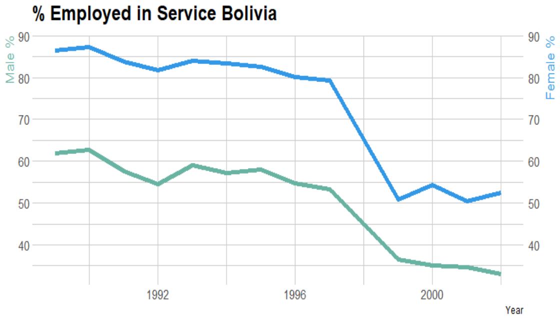


Fig. 16. Employment in Service Bolivia

Secondary School enrollment is more in males and increases gradually with time for both genders in developing countries such as Colombia, India, Sri Lanka, Bolivia (Fig.20). Whereas in developed countries such as Germany, Belgium, and Denmark (Fig.21), it is almost constant for the past 50 years. One notable exception is secondary school enrollment in China (Fig.22), where it drops suddenly in the mid-1980s to resume a constant increase after that.

Tertiary school enrollment is quite low for all countries when compared with Primary and Secondary school enrollment (Fig.23). It is generally higher in females than males

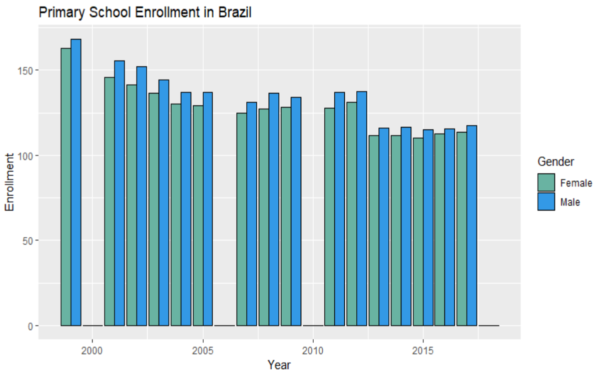


Fig. 17. Primary School Enrollment in Brazil

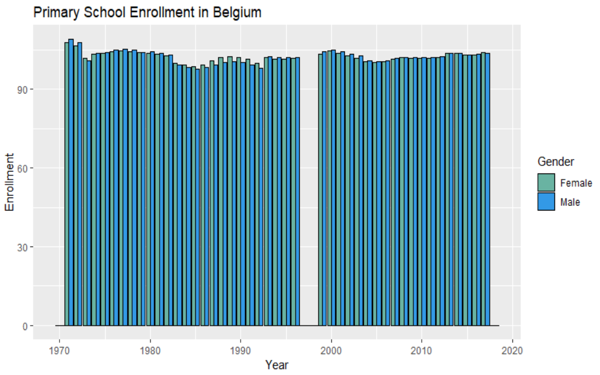


Fig. 18. Primary School Enrollment in Belgium

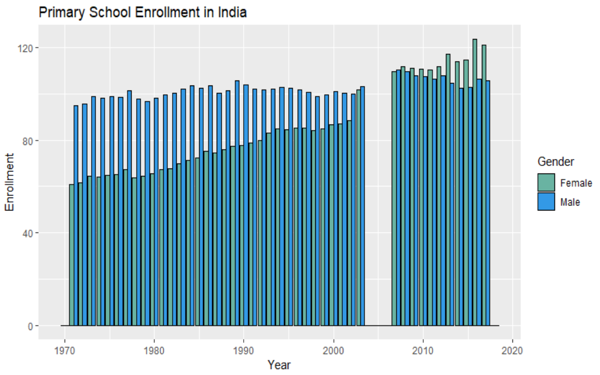


Fig. 19. Primary School Enrollment in India

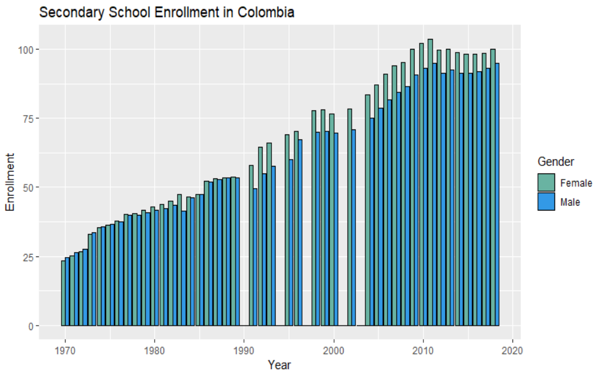


Fig. 20. Secondary School Enrollment in Colombia

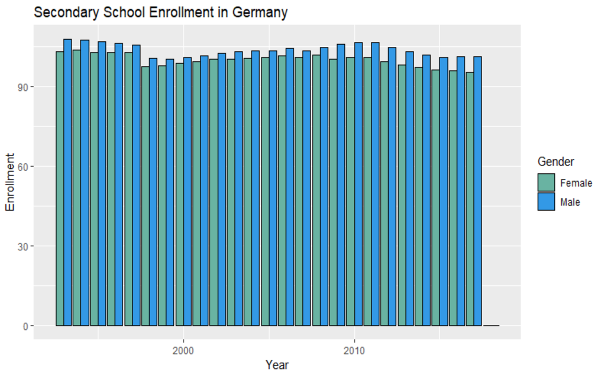


Fig. 21. Secondary School Enrollment in Germany

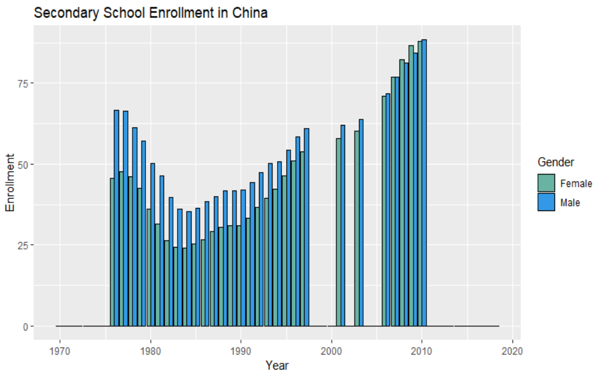


Fig. 22. Secondary School Enrollment in China

and is constantly increasing with a high correlation with the GDP (Fig.24). A notable exception is India, where enrollment is quite higher in males than females (Fig.25).

Model Development and Application

The data was centered for PCA using the mean value of attributes, and inner covariance was calculated to measure the correlated between the attributes [11]. Principle Component Analysis (PCA) was performed on the covariance matrix to find the most correlated attributes that are needed to cover 99% of the variance. The data was then normalized, and the dimensionality was reduced using the fit_transform function. This analysis outlined that we needed two principal components [PC1 and PC2] for the optimal coverage (alpha ¿ 99%).

A dendrogram was then visualized using the centroid variance minimization algorithm on the principal components to find the optimal number of clusters for Hierarchical Agglomerative Clustering [12] [13].

D. Birch Clustering

Birch Clustering uses a clustering feature tree to evaluate the data. [14] [13]. In this analysis, 1372 samples were used, and

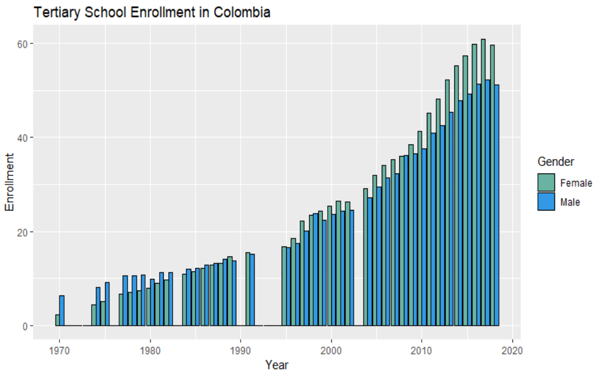


Fig. 23. Tertiary School Enrollment in Colombia

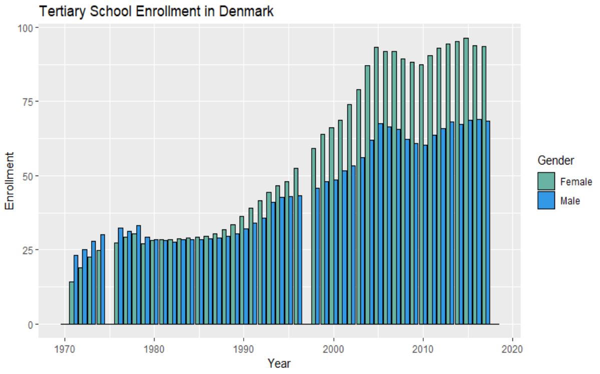


Fig. 24. Tertiary School Enrollment in Denmark

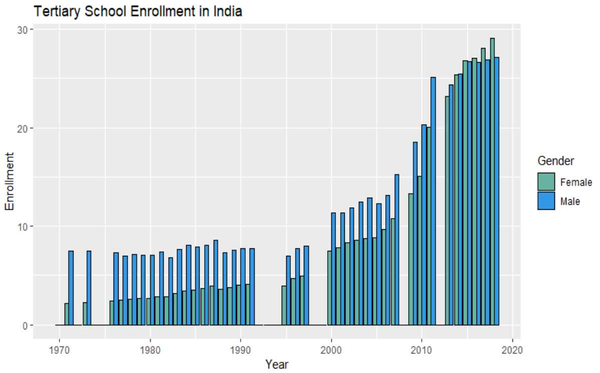


Fig. 25. Tertiary School Enrollment in India

the branching factor was set at 50 with 5 clusters to a threshold of 0.15. A higher threshold would cause the clusters to diverge, and an optimal solution is never reached; on the other hand, we noticed a lower threshold would cause over fitting [15]. When we performed the Birch clustering, the clusters are formed with the labels 0 - 2 where 0 represents a medium growth cluster, 1 represents a high growth cluster and 2 represents the countries with the lowest economic growth. These points are then plotted on a scatter plot where the purple cluster is high growth, green is medium growth and red is low growth (Fig.26).

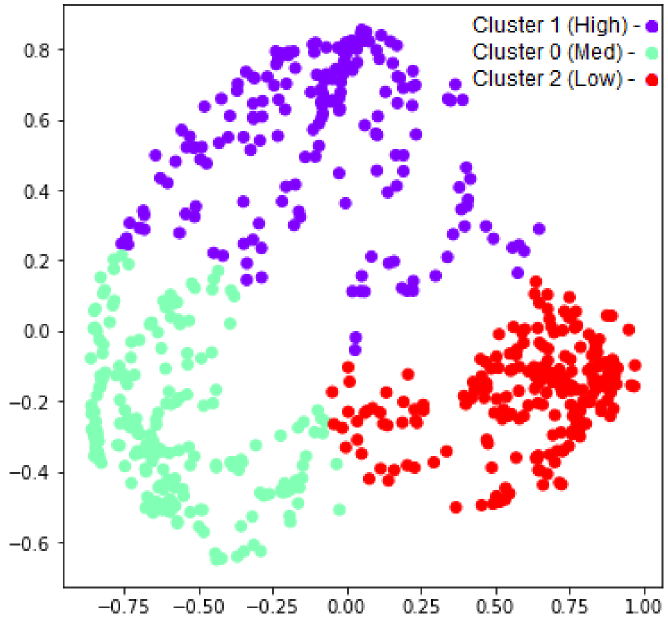


Fig. 26. Plot of the Birch Clusters

E. Hierarchical Agglomerative Clustering

Agglomerative clustering is a type of clustering which uses a bottom-up approach to cluster data points by treating each point as a single cluster [12] [13].

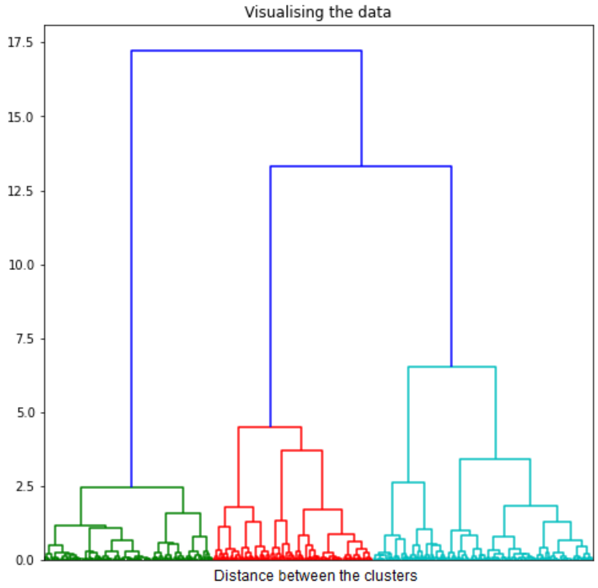


Fig. 27. Dendogram of the linkage function

In order to find the optimal number of clusters required, a linkage function was implemented, which takes the distance between pairs of data points and groups them on the basis of similarity. For this analysis, the Ward's minimum variance [16] [13] method was implemented where it minimizes the cluster variance and, at each step, merges the closest pair of

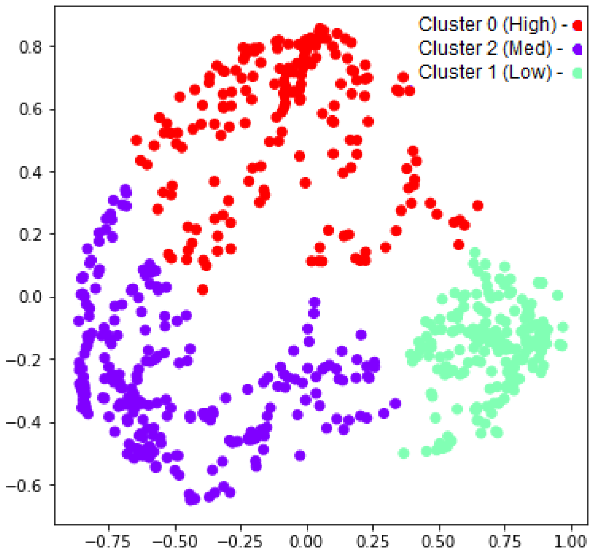


Fig. 28. Plot of the Agglomerative Clusters

clusters based on the distance between them [17]. As visible in the plot (Fig.27), the linkage function can then be plot as a dendrogram where it outlines the number of clusters needed on the basis of the branches from the root. In our analysis, the optimal number of clusters were 3, and the model was fit to the two principal components of the data set. This model resulted in 3 clusters on the basis of growth with labels from 0 to 2 where similar to Birch Clustering, and they were clustered into high, medium, and low growth nations (Fig.28). Here label 0 is used for high growth (red) clusters, 1 for medium growth (purple), and 2 for low growth (green) clusters.

F. *Mean-shift Clustering*

Mean shift clustering is a centroid based algorithm that updates the centroids to be the means in a given cluster [14] [13]. Duplicates points in a cluster were then removed after the optimal clusters were processed. For the estimate bandwidth parameter, the quantile is set as 0.2 since that gives the most accurate clusters. A value of less than 0.2 causes over-fitting and the formation of 9 clusters whereas any higher than 0.2 causes under-fitting and 1-2 clusters. Labels selects are from 0 to 5 where 0 signifies countries with low growth where as countries with high growth are represented by 1. As shown in the scatter plot (Fig.29) High growth clusters are in green, Medium High in orange, Medium in blue, Medium low growth in red and low growth nations in purple. Mean shift clustering at optimal tuning gives us 5 clusters which seem similar to the clusters formed using agglomerative clustering.

Results

The algorithms successfully clustered all 28 countries on the basis of their growth, depending on education, employment, and value-added by economic sectors.

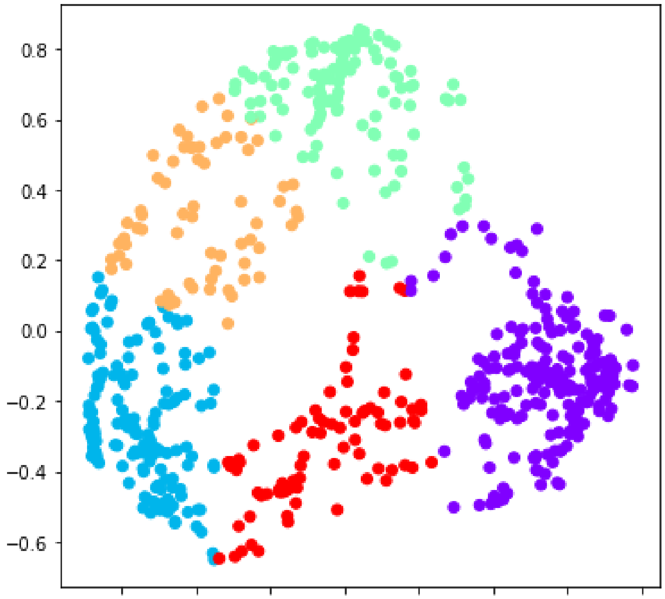
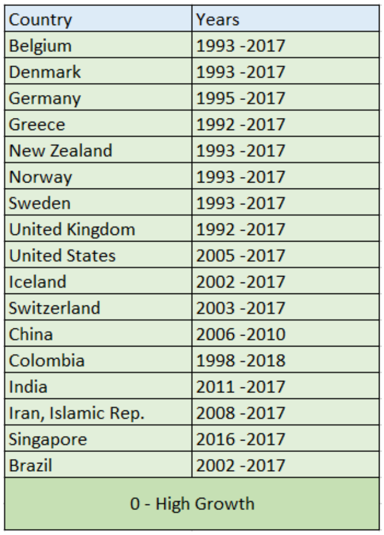


Fig. 29. Plot of the Mean Shift Clusters

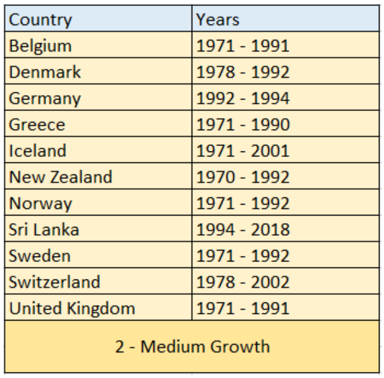


Country	Years
Belgium	1993 -2017
Denmark	1993 -2017
Germany	1995 -2017
Greece	1992 -2017
New Zealand	1993 -2017
Norway	1993 -2017
Sweden	1993 -2017
United Kingdom	1992 -2017
United States	2005 -2017
Iceland	2002 -2017
Switzerland	2003 -2017
China	2006 -2010
Colombia	1998 -2018
India	2011 -2017
Iran, Islamic Rep.	2008 -2017
Singapore	2016 -2017
Brazil	2002 -2017
0 - High Growth	

Fig. 30. High growth clusters (Agglomerative Clustering)

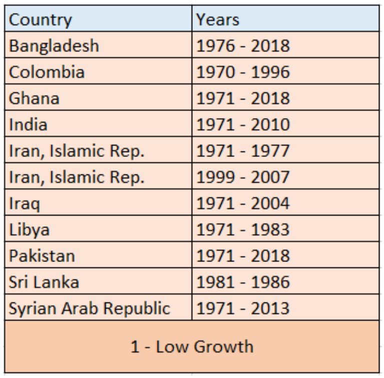
G. *Agglomerative Clustering*

Of the 3 clusters of High, Medium, and Low growth formed by Agglomerative clustering, cluster 0 signified the countries with High growth and consisted of developed nations such as Belgium, Denmark, Germany, Greece, Iceland, New Zealand, Norway, Sweden, Switzerland, United Kingdoms and the US for the years 1990s to 2019 and a few developing countries such as India, Columbia, China, Singapore, Brazil from the mid-2000s to present (Fig.30). The same developed countries had medium growth from 1970s to early 1990s and were labelled as cluster 2 for those years. This group also consisted of Sri Lanka for the years 1994 to present (Fig.31).



Country	Years
Belgium	1971 - 1991
Denmark	1978 - 1992
Germany	1992 - 1994
Greece	1971 - 1990
Iceland	1971 - 2001
New Zealand	1970 - 1992
Norway	1971 - 1992
Sri Lanka	1994 - 2018
Sweden	1971 - 1992
Switzerland	1978 - 2002
United Kingdom	1971 - 1991
2 - Medium Growth	

Fig. 31. Medium growth clusters (Agglomerative Clustering)



Country	Years
Bangladesh	1976 - 2018
Colombia	1970 - 1996
Ghana	1971 - 2018
India	1971 - 2010
Iran, Islamic Rep.	1971 - 1977
Iran, Islamic Rep.	1999 - 2007
Iraq	1971 - 2004
Libya	1971 - 1983
Pakistan	1971 - 2018
Sri Lanka	1981 - 1986
Syrian Arab Republic	1971 - 2013
1 - Low Growth	

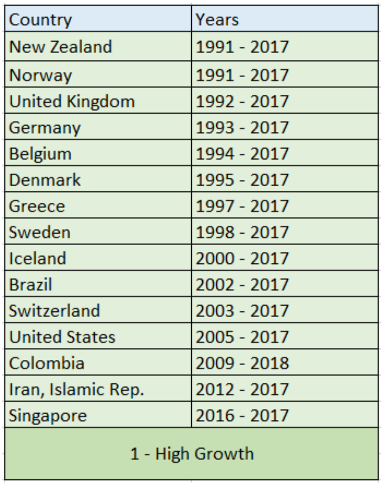
Fig. 32. Low growth clusters (Agglomerative Clustering)

Most developing countries where labelled as cluster 1, which represented low growth nations. This included countries such as Bangladesh, Ghana, Pakistan, Syria, for the years 1970 to 2019. It is also interesting so see that the countries such as India and China that were labeled as High or Medium growth for the years after 2000s, were labeled in this group for the years 1970 to mid-1990s (Fig.32).

H. *Birch Clustering*

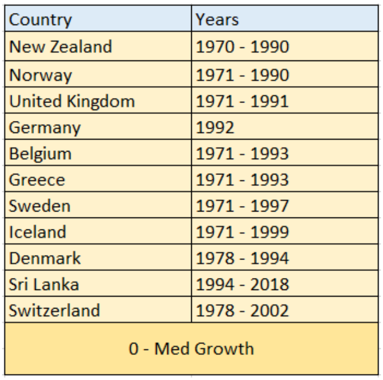
Birch Clustering [13] [15] method similarly divided the countries into 3 clusters of High, Medium, and Low growth based on the attributes. Cluster 1 represented High growth and consisted of Developed Countries from the 1990s to the present and some developing countries such as Brazil, Colombia, and Iran for the past ten years. (Fig.33).

Medium growth was represented in cluster 0, and it was similar to agglomerative clustering. This cluster consisted of developed nations from the 1970s to 1990s that were also clustered as High Growth for the past 30 years (1990s to present) (Fig.34).



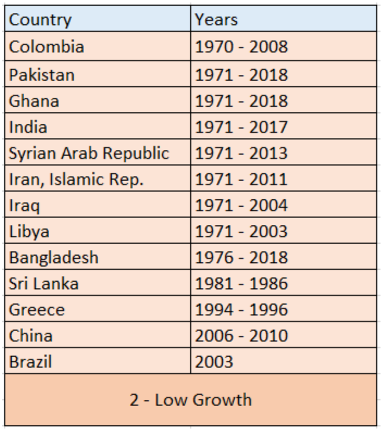
Country	Years
New Zealand	1991 - 2017
Norway	1991 - 2017
United Kingdom	1992 - 2017
Germany	1993 - 2017
Belgium	1994 - 2017
Denmark	1995 - 2017
Greece	1997 - 2017
Sweden	1998 - 2017
Iceland	2000 - 2017
Brazil	2002 - 2017
Switzerland	2003 - 2017
United States	2005 - 2017
Colombia	2009 - 2018
Iran, Islamic Rep.	2012 - 2017
Singapore	2016 - 2017
1 - High Growth	

Fig. 33. High growth clusters (Birch Clustering)



Country	Years
New Zealand	1970 - 1990
Norway	1971 - 1990
United Kingdom	1971 - 1991
Germany	1992
Belgium	1971 - 1993
Greece	1971 - 1993
Sweden	1971 - 1997
Iceland	1971 - 1999
Denmark	1978 - 1994
Sri Lanka	1994 - 2018
Switzerland	1978 - 2002
0 - Med Growth	

Fig. 34. Medium growth clusters (Birch Clustering)



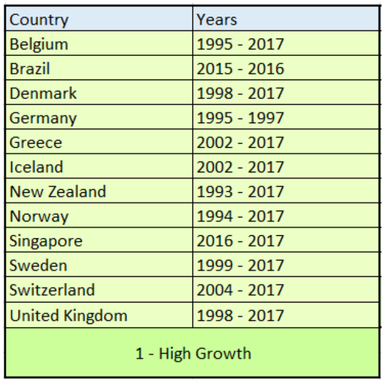
Country	Years
Colombia	1970 - 2008
Pakistan	1971 - 2018
Ghana	1971 - 2018
India	1971 - 2017
Syrian Arab Republic	1971 - 2013
Iran, Islamic Rep.	1971 - 2011
Iraq	1971 - 2004
Libya	1971 - 2003
Bangladesh	1976 - 2018
Sri Lanka	1981 - 1986
Greece	1994 - 1996
China	2006 - 2010
Brazil	2003
2 - Low Growth	

Fig. 35. Low growth clusters (Birch Clustering)

Cluster 2 (Fig.35) here represented countries with low growth and consisted of developing countries from 1970 to the present. A notable exception here was Greece that moved from medium growth to low growth in 1994 and jumped to high growth in 1997. China and India on the other hand remained in low growth even after the year 2005 unlike the agglomerative algorithm where they moved to the high growth cluster.

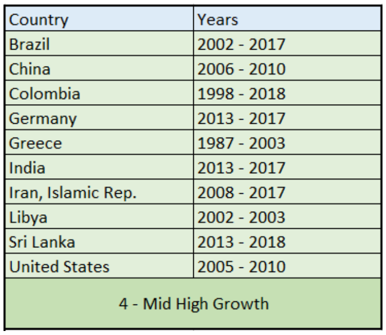
I. *Mean-shift Clustering*

The Mean Shift algorithm found 5 clusters optimal, namely ”High growth”, ”Medium-High growth”, ”Medium growth”, ”Medium-Low growth” and ”Low growth”. These are very similar to the 3 clusters formed by Agglomerative and Birch algorithms where most developed countries for the years 1990s to present are in the ”High growth” cluster (Fig.36) with the exception of Iran which in this case moved down to the ”Medium-High growth” cluster.



Country	Years
Belgium	1995 - 2017
Brazil	2015 - 2016
Denmark	1998 - 2017
Germany	1995 - 1997
Greece	2002 - 2017
Iceland	2002 - 2017
New Zealand	1993 - 2017
Norway	1994 - 2017
Singapore	2016 - 2017
Sweden	1999 - 2017
Switzerland	2004 - 2017
United Kingdom	1998 - 2017
1 - High Growth	

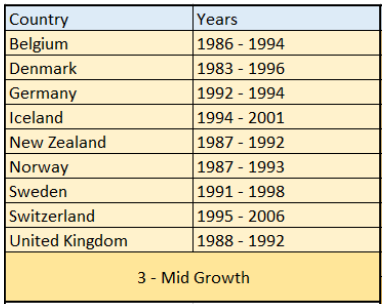
Fig. 36. High growth clusters (MeanShift Clustering)



Country	Years
Brazil	2002 - 2017
China	2006 - 2010
Colombia	1998 - 2018
Germany	2013 - 2017
Greece	1987 - 2003
India	2013 - 2017
Iran, Islamic Rep.	2008 - 2017
Libya	2002 - 2003
Sri Lanka	2013 - 2018
United States	2005 - 2010
4 - Mid High Growth	

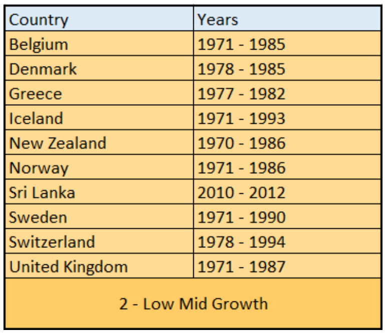
Fig. 37. Medium - High growth clusters (MeanShift Clustering)

The ”Medium-High growth” cluster (Fig.37) consists of developing nations from the range of 2000s to present due to the economic boom from the 90s. These were present in the ”High growth” or the ”Medium growth” clusters of the other two algorithms. The ”Medium growth” cluster (Fig.38) consists of developed nations in the 1990s which are experiencing steady growth in the tome period and the ”Medium-Low growth” cluster consists of developed nations before the 1990s (Fig.39).



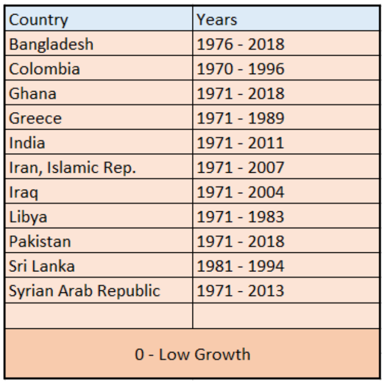
Country	Years
Belgium	1986 - 1994
Denmark	1983 - 1996
Germany	1992 - 1994
Iceland	1994 - 2001
New Zealand	1987 - 1992
Norway	1987 - 1993
Sweden	1991 - 1998
Switzerland	1995 - 2006
United Kingdom	1988 - 1992
3 - Mid Growth	

Fig. 38. Medium growth clusters (MeanShift Clustering)



Country	Years
Belgium	1971 - 1985
Denmark	1978 - 1985
Greece	1977 - 1982
Iceland	1971 - 1993
New Zealand	1970 - 1986
Norway	1971 - 1986
Sri Lanka	2010 - 2012
Sweden	1971 - 1990
Switzerland	1978 - 1994
United Kingdom	1971 - 1987
2 - Low Mid Growth	

Fig. 39. Medium - Low growth clusters (MeanShift Clustering)



Country	Years
Bangladesh	1976 - 2018
Colombia	1970 - 1996
Ghana	1971 - 2018
Greece	1971 - 1989
India	1971 - 2011
Iran, Islamic Rep.	1971 - 2007
Iraq	1971 - 2004
Libya	1971 - 1983
Pakistan	1971 - 2018
Sri Lanka	1981 - 1994
Syrian Arab Republic	1971 - 2013
0 - Low Growth	

Fig. 40. Low growth clusters (MeanShift Clustering)

The ”Low growth” cluster for mean shift algorithm consists of developing nations from the 1970s to the 1990s (Fig.40) with one exception of Greece for the years 1981 to 1989 due to a lower rate of tax caused by deficiencies in the tax structure [18].

Conclusion

The UN and the World Bank data were available in the form of individual values for 187 countries; data for 28 countries was extracted and was append together to form the data set for the analysis. The country name and year column were then removed from the testing data along with the GDP values. Inner covariance of the data set was calculated after centering the data and was used to perform principal component analysis to find the least number of dimensions needed to cover 99% of the spread. The two dimensions’ eigenvalues were projected on the data set to obtain the principle data set or a reduced basis for the clustering algorithms.

The Agglomerative Cluster’s linkage function was plot as a dendrogram to verify the optimal number of clusters. The algorithm was then utilized on the principal values of the data set for 3 clusters. Similarly, mean shift and birch clustering algorithms were implemented at optimal tuning for 5 and 3 clusters, respectively. From the analysis of the different clusters formed by the three algorithms, it is evident that developed nations such as the United States, United Kingdoms, Sweden, Denmark, Germany, Belgium, Iceland, Singapore, New Zealand, and Belgium are considered to be high growth for the last three decades along with some developing nations like India, China, Brazil, and Iran. They experience a boost in growth during the past decade. Even by current standards, these developed nations were still considered to be experiencing a medium growth in the 1990s when most other countries were lagging. By comparison, most developed countries in the 1990s were growing at a rate equal to many developing countries today, and they were both clustered in medium growth clusters. These developing nations were considered low growth before the 1990s and moved on to medium growth after the high growth phase in the 90s. A few countries have directly gone from the low growth cluster before the 1990s to the high growth cluster by today, and those countries are India, China, and Iran.

Our analysis outlines the importance of enrollment in schools at various levels and shows us how gender equality goes hand in hand with the growth of a nation. As seen in the exploratory data analysis, many of the developed nations clustered as high growth clusters had equal enrollment in all levels of education based on gender and had high employment of females in the service sector. We can also see employment by sectors where agriculture is negatively correlated with education and value-added by other sectors correlate positively. Our analysis work aligns on investigating multiple UN Sustainable Development Goals; Goal 5 (Gender Equality), Goal 4 (Quality Education), Goal 8: Decent Work, and Economic Growth [19] [20] [21].

We encourage the UN, World Bank, and other organizations to introduce a consistent and comprehensive data collection and data management plans to their member countries. Introducing and requesting those countries to follow common standards during the data collection can reduce the incompatibilities in data that affect future analysis work and get the maximum benefit from the collected data without any compromise.

IV. Future Work and Code Repository Access

We intend to keep this project work as an open-source project and contribute to the UN's Sustainable Development Goals efforts. In the future, we are planning to implement an interactive data visualization dashboard based on these results and upcoming analysis work. The data sets along with the complete code repository can be found on https://github.com/anishsethi96/ICDE-Economic-Growth

Acknowledgment

We would like to express our gratitude to Information Technology and Web Science (ITWS) program at Rensselaer Polytechnic Institute (RPI) for their generous support during this project. We appreciate the effort done by the United Nations and World Bank for keeping these data publicly available. We sincerely express our appreciation for the government agencies and respective authorities of those countries who collected the data and shared them with the UN and other organizations for the public use.

References

[1] W. M. Cole, "Wealth and health revisited: Economic growth and well-being in developing countries, 1970 to 2015," *Social Science Research*, vol. 77, p. 45–67, 2019.

[2] A. Bhargava, D. T. Jamison, L. J. Lau, and C. J. Murray, "Modeling the effects of health on economic growth," *Journal of Health Economics*, vol. 20, no. 3, p. 423–440, 2001.

[3] D. J. Brewer and P. J. McEwan, *Economics of Education.* Elsevier Publications, 2010.

[4] E. A. Hanushek and L. Woessmann, "The role of cognitive skills in economic development," *Journal of Economic Literature*, vol. 46, no. 3, p. 607–668, 2008.

[5] Å. Löfström, "Gender equality, economic growth and employment," *Swedish Ministry of Integration and Gender Equality*, 2009.

[6] T. Kodinariya and P. Makwana, "Review on determining of cluster in k-means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 90–95, 01 2013.

[7] "How can gross school enrollment ratios be over 100 percent?" [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/114955-how-can-gross-school-enrollment-ratios-be-over-100

[8] C. Avgerou, "The link between ict and economic growth in the discourse of development," in *Organizational information systems in the context of globalization*. Springer, 2003, pp. 373–386.

[9] S. D. Muñoz G.M., *Boom and Bust in Colombia 1990–2013*. Palgrave Macmillan, New York, 2016.

[10] "How does government spending affect the economic growth," Oct 2020. [Online]. Available: https://www.elearnmarkets.com/blog/government-spending-affect-the-economy/

[11] A. Dey, "Machine learning algorithms: a review," *International Journal of Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1174–1179, 2016.

[12] "Agglomerative hierarchical clustering." [Online]. Available: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/

[13] M. J. Zaki and W. Meira, *Data mining and machine learning: fundamental concepts and algorithms*. Cambridge University Press, 2020.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An efficient data clustering method for very large databases," p. 103–114, 1996. [Online]. Available: https://doi.org/10.1145/233269.233324

[16] R. K. Blashfield, "Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods." *Psychological Bulletin*, vol. 83, no. 3, p. 377, 1976.

[17] T. Kurita, "An efficient agglomerative clustering algorithm using a heap," *Pattern Recognition*, vol. 24, no. 3, pp. 205 – 209, 1991. [Online]. Available: http://www.sciencedirect.com/science/article/pii/003132039190062A

[18] M. Johnston, "Understanding the downfall of greece's economy," Aug 2020. [Online]. Available: https://www.investopedia.com/articles/investing/070115/understanding-downfall-greeces-economy.asp

[19] "Goal 4: Quality education." [Online]. Available: https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-4-quality-education.html

[20] "Goal 5: Gender equality." [Online]. Available: https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-5-gender-equality.html

[21] "Goal 8: Decent work and economic growth." [Online]. Available: https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-8-decent-work-and-economic-growth.html

[22] "Employment to population ratio, 15 , total (%) (modeled ilo estimate)." [Online]. Available: https://data.worldbank.org/indicator/SL.EMP.TOTL.SP.ZS

[23] "Literacy rate, adult total (% of people ages 15 and above)." [Online]. Available: https://data.worldbank.org/indicator/SE.ADT.LITR.ZS

[24] "School enrollment, primary, female (% gross)." [Online]. Available: https://data.worldbank.org/indicator/SE.PRM.ENRR.FE

[25] "School enrollment, primary, male (% gross)." [Online]. Available: https://data.worldbank.org/indicator/SE.PRM.ENRR.MA

[26] "School enrollment, secondary, female (% gross)." [Online]. Available: https://data.worldbank.org/indicator/SE.SEC.ENRR.FE

[27] "School enrollment, secondary, male (% gross)." [Online]. Available: https://data.worldbank.org/indicator/SE.SEC.ENRR.MA

[28] "School enrollment, tertiary, female (% gross)." [Online]. Available: https://data.worldbank.org/indicator/SE.TER.ENRR.FE

[29] "School enrollment, tertiary, male (% gross)." [Online]. Available: https://data.worldbank.org/indicator/SE.TER.ENRR.MA

[30] "Undata gdp by type of expenditure at current prices - us dollars," http://data.un.org/Data.aspx?d=SNAAMA&f=grID:101;currID:USD;pcFlag:0.

[31] "Undata gross value added by kind of economic activity at current prices - us dollars." [Online]. Available: http://data.un.org/Data.aspx?d=SNAAMA&f=grID:201;currID:USD;pcFlag:0

[32] "Undata per capita gdp at current prices - us dollars." [Online]. Available: https://data.un.org/Data.aspx?q=gdp&d=SNAAMA&f=grID:101;currID:USD;pcFlag:1