**Capstone Project**

**EDA on World Bank Global Education**

**Arvind Krishna, Keshav Sharma, Jayesh Panchal, Sajjad Ali, Lakshay Arora**

**Data science students**

**Cohort- Boston, Alma Better**

**Introduction**

The **World Bank** Group is the largest financier of **education** in the developing world. They work on **education** programs in 90 countries and are committed to helping them reach SDG4 which calls for access to inclusive and equitable quality education and lifelong learning opportunities for all.

The World Bank Ed Stats All Indicator Query holds over 4,000 internationally comparable indicators that describe education access, progression, completion, literacy, teachers, population, and expenditures. The indicators cover the education cycle from pre-primary to vocational and tertiary education and also holds learning outcome data from international and regional learning assessments (e.g., PISA, TIMSS, PIRLS), equity data from household surveys, and projection/attainment data.

Our analysis on World Bank global education can help us better understand what could be the possible reason that are responsible for change in indicators on education stats for each and every country on the globe.

1. **Problem Statement**

Data provided by the World Bank global Ed Stats are in unformatted manner, corrupted data, and duplicate data and also sometimes it is irrelevant, because it’s a piled-up data coming from various different countries. For doing the analysis on the data the data needs to be in correct format and well organized formed

Various dataset that needs to be cleaned are as follow:

* **EdStatsData**
* **EdStatsCountry**
* **EdStatsCountry-Series**
* **EdStatsSeries**

The main objective of the analysis is to obtain the meaning full information and facts from the given huge datasets as shown above, by cleaning the datasets, doing a proper analysis and visualization and plotting the useful information into different graph and charts so that the trend and relationship between the various indicators on which the analysis is done can be understand easily.

This will also help the World Bank group in taking the future decision based on the past events.

1. **Steps involved**

* **Exploratory Data Analysis**
* **Loading and discovering data**

Now, we need to load our data from the external source, which in this case is uploaded to the drive. Also we would try to spot the nature and properties of the data that we have. The data is divided into 5 different CSV files, each containing the information as follows

* **Importing necessary modules and libraries**

We are importing following libraries for their respective applications:

* Pandas: - Pandas is used to analyse data. It has functions for analysing, cleaning, exploring, and manipulating data.
* Matplotlib: - Matplotlib is a graph plotting library in python that serves as a visualization utility. Most of the Matplotlib utilities lies under the pyplot submodule.
* **Data Cleaning**

Data cleaning is an important step in the data analytics process in which you either remove or update information that is incomplete or improperly formatted.

* **Null values Treatment by different methods**

Dropping and replacing by f fill and b fill

* **Selecting criteria of indicators**
* **Plotting various graph for different indicators**
* **Finding the key facts and relationship between various indicators and countries**
* **concluding**

1. **Data information**

**EdStatsData**: This file contains Education Statistics data (from year 1970 to 2100) of all the countries that are World Bank members and this data is dependent on various indicators that resembles various factors that affect the overall Education growth and development of the country. These indicators can be access, availability, teachers, expenditures, population, literacy, assessments etc.

**EdStatsCountry**: This file contains information from census and surveys across various departments, segregated in a country-wise manner

**EdStatsCountry-Series**: This file contains the Series Code for various indicators, as well as the Data sources from which they have been derived

**EdStatsFootNote**: This file contains year wise indicator names along with their respective description. It tells us the nature of the indicators, if it’s and estimate value or a percentage distribution etc. From context of our analysis this information isn't quite useful, since it’s already evident with the type of data EdStatsData holds.

**EdStatsSeries:**-This file contains regional learning assessments data (e.g. PISA, TIMSS, and PIRLS), equity data from household surveys, and projection/attainment data.

## **4. Selecting Indicators**

The following indicators are taken into consideration for which analysis is to be done:

**Early Childhood Education**

* Enrolment in early childhood education, both sexes (number)
* Percentage of enrolment in pre-primary education in private institutions (%)

**Expenditures**

* Government expenditure on education as % of GDP (%)
* Expenditure on education as % of total government expenditure

**Literacy**

* Adult literacy rate, population 15+ years, both sexes (%)
* Adult illiterate population, 15+ years, both sexes (number)

**Population**

* School age population, pre-primary education, both sexes (number)
* School age population, primary education, both sexes (number)
* School age population, secondary education, both sexes (number)
* School age population, tertiary education, both sexes (number)

**Learning Outcomes**

* PISA: Mean performance on the Reading scale (number)
* PISA: Mean performance on the Mathematics scale (number)

# **Selecting Countries**

List of Countries which are taken into consideration for the analysis.

**India**

**United States**

**United Kingdom**

**Australia**

**Germany**

**South Africa**

**Canada**

**Japan**

**Sri Lanka**

* 1. **Exploring indicators**

### **Early Childhood Education**

Early childhood education (ECE), also called nursery education, refers to the branch of education theory that relates to the teaching of children from birth up to the age of eight. In most cases its referred equivalent to 3rd Grade.

Quality of early childhood education can reflect a significant contribution to the physical, psychomotor, cognitive, social and emotional development of the child, including the acquisition of languages and early age knowledge, which in a long run would reflect towards the country’s literacy. The first eight years of a child's life is a period of tremendous growth and development.

**Observations:**

We are able to observe that the **United States** has some fluctuations in the trend may be due to the subprime crisis in U.S in the year 2007 to 2010 and we can see the significant leap happening in the period 2010 - 2012, followed by a minor drop. Overall, the trend for USA seemed to be consistently showing positive growth. Part of the reason why USA is so developed is because they focus a lot towards early age education of children’s, which contributes significantly to the overall literacy.

For **India**, we observe significant leaps in the periods 2001-2003, 2005-2008 and also at 2010-2013, and we are able to notice that the slopes and degrees of hills are more significant than the slopes and degrees of the valleys, which shows positive signs of growth. India is a developing country and with the above chart, it is evident that sooner or later India would lead the graph in coming years, second only to China.

For **Japan** and **Germany**, we see minor fluctuations throughout, and significant changes in the period 2012-2013. we see a climb for *Germany*, and a drop for *Japan*. Japan has a rigorous process of early education such that students enrol to a school after the age of 8. This clearly affect their rate of early education enrolment rate.

We also see a consistent growth trend for the **United Kingdom**, with no drastic movements.

Also, **South Africa** shows a positive trend for the period 2004-2014, and then has a minor drop.

**Canada** and **Sri Lanka** don’t show any major growth or decline, while **Australia** shows a significant growth in the recent years from 2012 onwards.

*Note:* We haven't included **China** for analysis. Reason being the graph for China was overfitting the charts and overshadowing other countries, which made it hard to analyse trends across other countries. It is clear China is way ahead of other countries. Hence, we selected countries that are homogenous to India.

### **Expenditures**

### It signifies the amount spent by federal government on public educational institutions and other education related institutions. Education plays a very important role for a country to be able to absorb technology and develop the capacity for the realization of development and sustainable growth. Also, Human capital is a productive investment in people that includes knowledge, skills, abilities, ideas, health which often results from expenditures in education.

**Observations:**

**India**: During the period 1997-1999, we see the indicator climb up and remain steady up until 2002, followed by a consistent decline until 2007. Politically, we can observe this as the period towards the end of Vajpayee ministry, and the entry of manmohan ministry.

In the period 2007-2011 we see a consistent increase, post which it remains stable. This is a period wherein the manmohan ministry was prevalent.

**Canada** - We observe a declining trend during the period 1997-2008.and for the period 2008-2011 there is a small climb, post which the indicator remains consistent.

**United Kingdom** - Initially we observe a decline in the period 1997-1999. After which the graph is going up. We see two positive leaps, one at 2001-2002, and the other at 2009-2010, and minor fluctuations for the other periods

**Japan** - We observe the graph to be rather stable. There is a small positive leap we see during the period 2009-2010. It is also noteworthy that the magnitude of the indicator as of 1997, and as of 2015, are relatively similar. The overall graph paints a picture of stability.

**United States** - We see the graph remain consitent for the period 1997-2001, shortly followed by a leap, and significant fluctuations happening throughout the rest of the graph.

**Australia** - we observe a declining trend for the period 1997-2008 post which, there is a significant leap in the period 2008-2010, followed by a drop from 2010-2012. It is also noteworthy that the magnitude of the indicator as of 1997, and as of 2015, are relatively similar. Looks like a considerable baseline.

**Expenditure on education as % of total government expenditure.**

**Observations:**

From the analysis we can conclude that the expenditure on education of India and Sri Lanka decreased in 2014 than that of 1999. Reverse pattern can be seen for South Africa, their expenditure is increased to approx. 18%. United States and United Kingdom are consistently developing well over the years. Also, for countries like Japan and Germany the government expenditure is getting reduced as years go by. whereas of Australia and Canada shows positive signs of growth.

### **Literacy**

Literacy is the process of expanding one's knowledge of reading and writing in order to develop their thinking and learning for the purpose of understanding themselves and the world. Higher literacy rates are associated with high populations, less crime rates, greater economic growth, and higher employment rates. For a person, literacy is a foundational skill required to acquire advanced skills.

**Observations**

We are able to see that the highest literacy rate in the data for India, has an approximate magnitude of 70, and the lowest being approximately 40. The median literacy rate, which has an approximate magnitude of 62, lies towards the higher end for India, which is a positive sign.

This also informs us that the distribution of values for this indicator, is concentrated in the upper region of the graph, which shows that the literacy rate has mostly been above the 50% mark.

**Adult illiterate population, 15+ years, both sexes**

**Observations**:

We see significant inclines during the period 1981-2006, which peaks at 2006. This shows that the highest point of illiterate population for India was recorded at 2006.

Following that, we observe a steep decline in the period 2006-2010, followed by a steady decline from 2010-2011, which shows a positive movement in the overall period through 2006-2011.

There is also a possible relationship which we can observe with regards to the 'Government expenditure on education' indicator we had observed earlier, as we can see the 'Government expenditure on Education' going up during the period 2007-2011, and the adult illiterate population metric declining during the same period.

### **Population**

Population of a country refers to the number of individual livings in that country. These individuals are benefited members of the country and contribute to various aspects of country's Statistics. Population majorly affects the Education and literacy. Rapid growth of population increases the rate of illiteracy and poverty. It also increases the disparity between rich and poor countries in their socio- economic as well as general life. causes imbalance in its use.

**We've selected 4 indicators for this category, they are:**

* School age population, pre-primary education, both sexes (number)
* School age population, primary education, both sexes (number)
* School age population, secondary education, both sexes (number)
* School age population, tertiary education, both sexes (number)

**It is evident that other countries seem insignificant when compared to India in terms of population. Hence, we can narrow down our analysis to the top 4 countries that are most relevant. In this case we are considering the following countries.**

* India
* United States
* Japan
* Germany

**Observations**

For India, we observe a significant leap in the population of official age for pre-primary education, in the period 1997-1998. We are also able to see that across all the graphs, India is on a consistent incline. For the countries Japan, United States and Germany, we see an overall stable trend with minor fluctuations. This shows a potential focus towards stability and consistency in the population.

### **Learning outcomes**

The Learning Outcomes category highlights levels of student learning in reading and mathematics in over 100 countries based on data from international learning assessments like PISA.

We have selected two indicators in this category. Also, we are taking all the available countries for this indicator, merging them as continents and represent on pie charts.

**PISA: Mean performance on the reading scale**

**Observations:**

We see that the values for the top 20 Countries, based on the 'PISA: Mean Performance on the Reading Scale', lie in the range of 495 - 540. It helps us understand the typical range of the highest scores for the indicator.

Also, in the top 20, We have predominantly European countries (11 of the 20 are European Countries), followed by Asian countries (6), Australia and New Zealand from the ANZ region, and Canada from North America.

However, considering just the top 5 countries, we can observe that 4 of them, are Asian, specifically in the East/South-East Asia regions.

So, 4 out of the 6 Asian countries, are in the Top 5, and all Asian Countries except Macao, are in the Top 10

We also see that Finland, which is at #2, is the only European country which is in the Top 5.

**PISA: Mean performance on the mathematics scale.**

**Observations:**

We see that the values for the top 20 Countries, based on the 'PISA: Mean Performance on the mathematics Scale', lie in the range of 500-580.

In the top 20, We have predominantly European countries (10 of the 20 are European Countries), followed by Asian countries (7), Australia and New Zealand from the ANZ region, and Canada from North America.

However, considering just the top 5 countries, we can observe that all 5 of them, are Asian, specifically in the East/South-East Asia regions.

It is noteworthy that 5 of the 7 Asian countries, are in the Top 5, and all Asian Countries except Vietnam, are in the Top 10.

* 1. **Summary:**
* We see that for Enrolment in early Childhood education, the uptrend is significantly positive. It is evident that India is making remarkable and positive growth in this field
* For enrolments In Pre-primary Education in private institutions, we observed that the enrolment during the 2015 has been 2.5X than that of the year 2000.
* The government expenditure on education as % of GDP is rising consistently after the year 2007 and remains stable after 2011, even though there were fluctuations in the past years
* The total expenditure of the government on education (%), for India and Sri Lanka decreased in 2014, then that of 1999, and the expenditure by the United States and the United Kingdom has consistently increased over the years.
* After the year 2006, we see that the illiterate population metric has been decreasing significantly in India, we could be related with the increase in the government expenditures and initiatives towards education.
* Among the Asian countries, China has been holding the lead in PISA scores for Reading and also for Mathematics; Finland leading amongst European countries.

**8. Conclusion:**

Starting with loading the data so far, we have done EDA, null values treatment, and many more exercises.

Finally at last we came to the conclusion on statement i.e.

From the analysis we are able to conclude that India among all is one of the most significantly developing countries. India has shown positive growth over the period of last 20 years (considering from 1995 to 2015). The early childhood enrolment rate of India shows tremendous extension, which signifies that people are more conscious towards the education and also the literacy rate has improved considerably. We can see similar trends for countries like USA. This is clear that USA is currently more prominent and developed in terms of education but the outstanding growth of India over the years shows great potential for the Education domain.

**References-**

* **Education Statistics (EdState) Indicators Data** <https://datatopics.worldbank.org/education>
* **Python Pandas Documentation** <https://pandas.pydata.org/pandas-docs/stable>
* **Python MatPlotLib Documentation** <https://matplotlib.org/stable/index.html>
* **Our EDA Project Documentation**