**Capstone Project**

**EDA on World Bank Global Education**

**Neel Naik , Piyush Nirwan , Prasad Khadatkar , Saransh Jain , Shivank Shukla**

**Data Science Students**

**Cohort- Azaadi, 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 hold 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 reasons that are responsible for change in indicators on education stats for each and every country on the globe.

**1.Problem Statement**

It is always important to understand what even seems to be the problem or the question. Making an assumption or not understanding fully the problem will lead to wrong conclusions and will result in wrong actions. Identifying the problem is naturally also one of the hardest tasks. Like A. Einstein stated:

**'If I had an hour to solve a problem I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions.'**

* Government Expenditure on Education and its impact on the reach of education to everyone.
* Impact of economic class or economic background on access to education.
* With time, the government has become more careful with allocating funds for education

**2. 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 resemble 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 an estimated value or a percentage distribution etc. From the 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

**3. Selecting Indicators**

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

**Economy:-**

* GDP at market prices (current US$) NY.GDP.MKTP.CD
* GDP per capita (current US$) NY.GDP.PCAP.CD

**Early Childhood Education**

* Out-of-school children of primary school age, both sexes (number) SE.PRM.UNER
* Enrolment in pre-primary education, female (number) SE.PRE.ENRL.FE
* Enrolment in secondary education, female (number) SE.SEC.ENRL.FE
* Enrolment in tertiary education, all programmes, female (number) SE.TER.ENRL.FE
* Pupil-teacher ratio in primary education (headcount basis) SE.PRM.ENRL.TC.ZS

**Expenditures:-**

* Government expenditure on education as % of GDP (%) SE.XPD.TOTL.GD.ZS
* Expenditure on education as % of total government expenditure (%) SE.XPD.TOTL.GB.ZS

**Literacy among working population:-**

* Labor force, total SL.TLF.TOTL.IN

**Population:-**

* Population growth (annual %) SP.POP.GROW

**Learning Outcomes:-**

* PISA: Mean performance on the Reading scale (number) LO.PISA.REA
* PISA: Mean performance on the Mathematics scale (number) LO.PISA.MAT

**4. Process**

When we start using the data, it might be a combination from different sources or it might not be of the highest quality. A process known as data cleaning is the fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. What we aim to achieve is clean data. And to tell the truth, that is a science on its own. There are plenty of tools, theories, and methods to use.

**● 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 analyze data. It has functions for analyzing, 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 lie 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 all the null columns from 2026-2095

Filtering out all the indicators having null values

* **Creating the DataFrame**

On the basis of income categories and region , two Dataframes were constructed.

**5. Analysis**

**5.1 Analysis on income group**

* From our previous analysis we came to know that Higher income countries constitute only about 1/4 in numbers and still controls 60% of world GDP.
* Despite making ⅓ in numbers, lower income countries have only about 0.5% of Global GDP share.
* This parity shows on multiple indicators such as population growth rate, Female enrolment in Tertiary Education, Teacher to Pupil Ratio, Mortality Rate.
* On one hand the overall education level of high income countries has always been good, But over the years, especially over the last two decades the female enrolment in tertiary education and labor force has increased exponentially.
* On the other hand, lower income countries are still struggling for any significant improvement in the education sector.
* Middle income countries have shown consistent improvement in overall education, Be it Teacher to pupil ratio, percentage of population getting access to primary education, Female participation in all levels of education.

**5.2 Analysis by Region**

* From our analysis on the global GDP with respect to income groups on the world mapwe can clearly see that high income countries are concentrated in Europe and North america.
* Most of the low income Countries are in Africa with few exceptions in the Middle East(Afghanistan).
* Middle income countries are spread throughout the world with high concentration in Asia, with representation in Europe, Latin America, Africa.
* From all this analysis we are stating the obvious, Europe and North America being the center of education in the world due their world renowned institutes attracting talent from all across the Globe.
* Whereas sub saharan Africa is constantly struggling to feed their population. With not enough resources to allocate towards education these countries have to rely on organizations such as World Bank and its sister organizations for basic necessities like food and education.
* This trend also shows in the mortality rate being high in africa and low in europe and north america, and lack of education and awareness contributes towards high population growth in africa and relatively low population growth rate in europe and north america.
* With some analysts stating that this century will be Asia’s century.Most of the emerging economies are from Asia, these emerging economies countries are spending a large percentage of their GDP towards educating their children and reeducating their labor force.

**5.3 Analysis on Big Economies of the world**

* The analysis on the big economies of the world over the last three decades shows the clear trend of high government expenditure on education.
* The analysis of labor force on these big economies shows a well educated(trained) labor force. India has the second largest trained labor force.
* Even countries like Japan with a small aging population have a relatively large amount of educated labor force.
* PISA reading scores and PISA mathematics scores show a clear picture of how the spending on education has been fruitful for these countries.
* With the trend of technology integration in social life and a possibility of Metaverse, it is expected to see a further increase in education spending for these countries.

**6. Suggestions**

* **Awareness programmes** should be incorporated in order to have a sustainable population growth so that proper resources and facilities can be provided to everyone.
* There should be **more spending on education** as a percentage of GDP by the respective governments.
* As we have analyzed that the **skilled and educated workforce** can contribute better in the development of the nation, we must be focused towards education of children and re-education of the labor force.
* In order to increase the quality of education we should follow the norms suggested by the world bank for **Teacher to Pupil ratio**.
* **Better data collection** helps in **better data analysis** and providing solutions to ongoing problems and upcoming problems, so we should try to collect the proper data.

**References-**

* **Education Statistics (EdState) IndicatorsData** [**Education Statistics | Home (worldbank.org)**](https://datatopics.worldbank.org/education/)
* **Python Pandas Documentation** [**https://pandas.pydata.org/pandas-docs/stable**](https://pandas.pydata.org/pandas-docs/stable)
* **Python MatPlotLib Documentation** [**https://matplotlib.org/stable/index.html**](https://matplotlib.org/stable/index.html)
* **Python Seaborn Documentation** [**seaborn: statistical data visualization — seaborn 0.12.0 documentation (pydata.org)**](https://seaborn.pydata.org/)
* **Python plotly Documentation** [**https://plotly.com/python-api-reference/**](https://plotly.com/python-api-reference/)
* **Our EDA Project Documentation**