*-- Project Status: [Completed]*

Project Summary :

Capstone project 1 - World Bank Global Education Analysis

**About the Dataset :**

We are provided with a dataset named World Bank EdStats that contains 4000 indicators describing education access(primary, vocational and tertiary), progression, literacy, teacher, population etc.

The World Bank is an international financial institution which provides monetary assistance to governments of low and middle countries for socio-economic development. Each year, the World Bank collects various statistics through international statistical communities and globally coordinated programs to monitor the growth and progress of various economies.

**Problem Statement :**

We were provided with 5 datasets and asked to find insights and trends for different groups of countries and come up with a rough estimate as to which countries are alike and which are dissimilar.

**Approach taken :**

The idea of the project is to analyze the data and mark the variation of indicators across the globe, and also group the countries according to their similarities and differences.

**Insights from exploring the Data :**

By analysing and plotting the required graphs, we find that most of the data is available in the year range 2000 to 2010 and after the year 2020, most of the data are missing and only projections are observed. All government expenditures across various countries in the education sector among the income groups from the year 1998 onwards have become stable and in the years 1983 and 1986 government mainly focused on the lower middle income group.

Finally, due to the emergence of the internet around the year 1995 shows a stark growth of literacy rate and the highest growth has been observed in the High income OECD(The Organisation for Economic Co-operation and Development) group.

**Challenges faced:**

It was very challenging to completely understand the data and to comprehend the relevance of each CSV file As the percentage of missing data was huge, it took a lot of effort to decide on the final data to keep for analysis Filtering out the best indicators from 3700 indicators to keep for analysis deciding on the set of countries to work based on economy and geography.

**Future scope of work:**

Working out on Top European powers and comparing their positions based on different indicators considering the amount of indicators in the data, if we dig deep enough, various micro trends can be unearthed, which we were not able to extensively cover during this short duration. This dataset can also be used to measure compensation of teachers, if we are to advise the education ministry on management of funds. Learning Assessment Indicators for Mathematics and Science can be used to predict populations that tend to have a knack for technology.

Data Visualization Methods Used :

* Time Series Line graphs
* Scatter plot

Python Libraries used :

**For Graphing :**

* Matplotlib
* Seaborn

**Data wrangling :**

* Numpy
* Pandas

The Structure of the main IPYNB notebook :

* Initial Library imports
* Exploring File : EdStatsCountry-Series.csv
  + Contains indicators and data sources for certain countries
* Exploring File : EdStatsCountry.csv
  + Contains list of all countries that are present in the data (total of 241 countries) - along with other features specific to the country like - Region, Income Group specific to the country etc.
* Exploring File : EdStatsSeries.csv
  + Contains list of all indicators and the definition of each indicator
  + Reading data and exploring which columns are necessary
* Exploring File : EdStatsData.csv
  + contains each country, with list of indicators (3665 unique indicators); contains measurement value for each indicator from years 1970 to 2017; from 2020 to 2100 - contains projections
  + Reading data and exploring which columns are necessary
* Exploring File : EdStatsFootNote.csv
  + Contains the estimations and uncertainty bounds for each year - looks like some years are missing
* Utility Function 1 (income\_group\_plot) is used to plot the graphs of various income group
* Utility Function 2 (Country\_indicator) is used to plot the graphs of developed and developing country trends based on country indicator in various age , sex

Project done by :

* Name : **DEBARTHA ROY CHOWDHURY**
* Github link: [*deba-roy/World\_Bank\_Global\_Education\_Analysis: Exploratory Data Analysis of Educational data by World bank across 240+ countries around the World (github.com)*](https://github.com/deba-roy/World_Bank_Global_Education_Analysis)
* Google Drive link: *https://drive.google.com/drive/folders/1xXf0Xkz0z\_Rxo2DFgWecA8Oa3lEnZnmo?usp=sharing*