**World Bank Global Education Analysis**

Project by: DEBARTHA ROY CHOWDHURY

**Abstract:**

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. 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.

**Introduction:**

The World Bank EdStats (Education Statistics) portal is a comprehensive data and analysis source for key topics in education such as access, completion, learning, expenditures, policy, and equity. Data sources include administrative country data from international learning assessments (PISA, TIMSS, PIRLS, PIAAC, and EGRA) and three regional learning assessments (SACMEQ, PASEC, LLECE); World Bank databases. It includes World Bank Education Projects Database classified by activities, components, and sub-sectors of all World Bank Education Projects since 1970.

**Problem Statement:**

The World Bank EdStats 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.

Explore and analyze the data to identify variation of indicators across the globe, which countries are more alike and different. Feel free to add more extensive analyses and details.

**Methodology:**

At first I have imported 5 datasets and checked null values of each. Then read the datasets one by one to drop insignificant columns. Thereafter trying to visualise which columns or dataset will be important to analyse the data. Necessary libraries like numPy, pandas, matplotlib, seaborn, random (to get an arbitrary integer value within range) have been imported for plotting different graphs according to the data. Lastly, based on the graphs, adequate analysis have been done to find meaningful insights.

**Dataset analysation:**

From the data source provided we get 5 Datasets namely EdStatsCountry-Series.csv, EdStatsCountry.csv, EdStatsData.csv , EdStatsFootNote.csv and EdStatsSeries.csv.

*The details of the datasets are as follows:-*

* (EdStatsCountry-Series.csv) -It contains indicators and data sources for certain countries.
* (EdStatsCountry.csv) -It contains list of all countries that are present in the data (Which is a total of 241 countries) - along with other features specific to the country like region, income group specific to the country etc.
* (EdStatsData.csv) – It contains each country, with list of indicators (Contains about 3665 unique indicators); contains measurement value for each indicator from years 1970 to 2017; from 2020 to 2100 – which contains projections only.
* (EdStatsFootNote.csv) – It contains the estimations and uncertainty bounds for each year and it seems to look like some years are missing.
* (EdStatsSeries.csv) – It contains list of all indicators and the definition of each indicator.

**Null value handling and Data cleaning:**

In EdStatsCountry-Series.csv we have unnamed: 3 which have 0 not null values so we remove it.

In EdStatsCountry.csv we have unnamed: 31 which have 0 not null values so we remove it.

In EdStatsData.csv we have unnamed: 69 which have 0 not null values so we remove it.

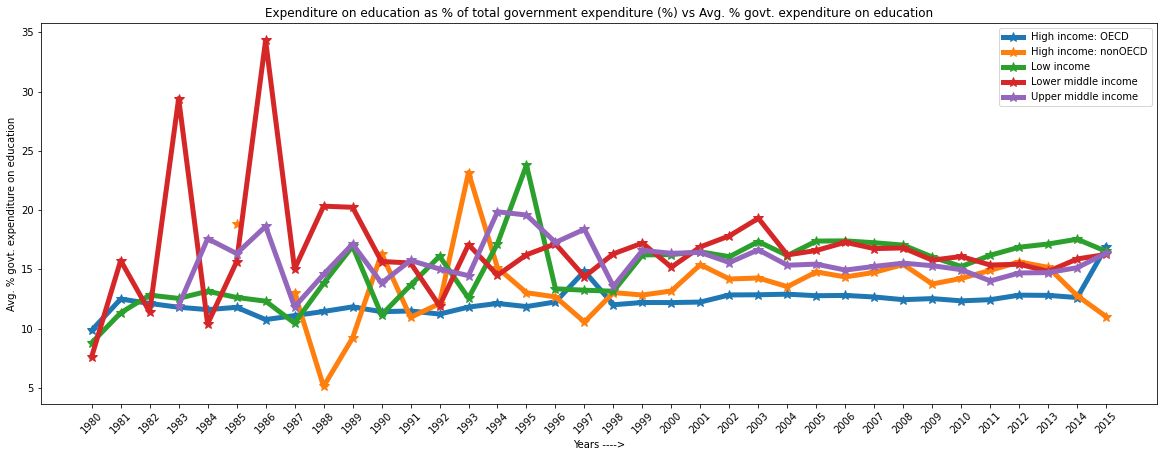
In EdStatsFootNote.csv we have unnamed: 4 which have 0 not null values so we remove it.

In EdStatsSeries.csv we have Unnamed: 20, License Type, Related indicators, Other web links, Notes from original source, Unit of measure which have 0 not null values so we remove it.

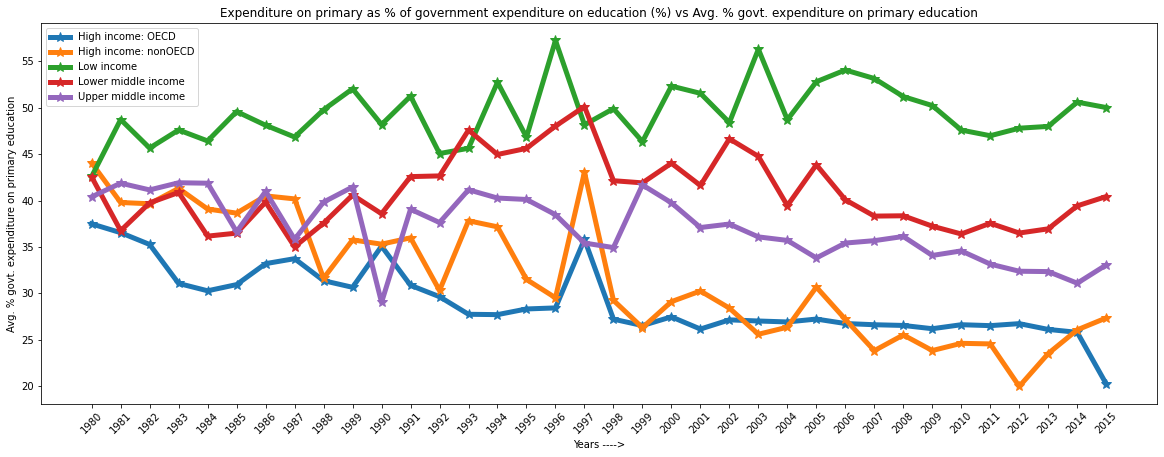
**Visualization:**

We want to visualize the statistical data present in our Dataset and find insights to it.

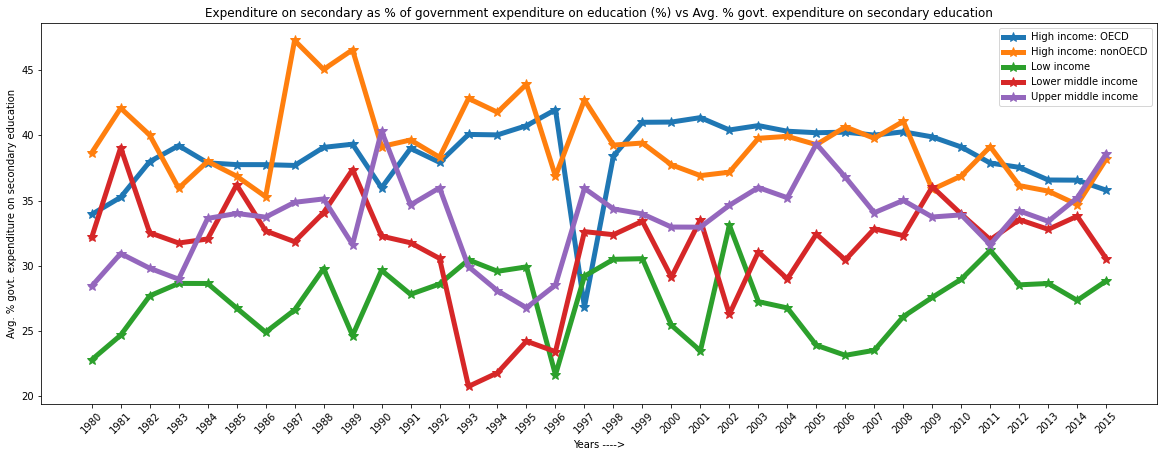
**Observation 1:**

****

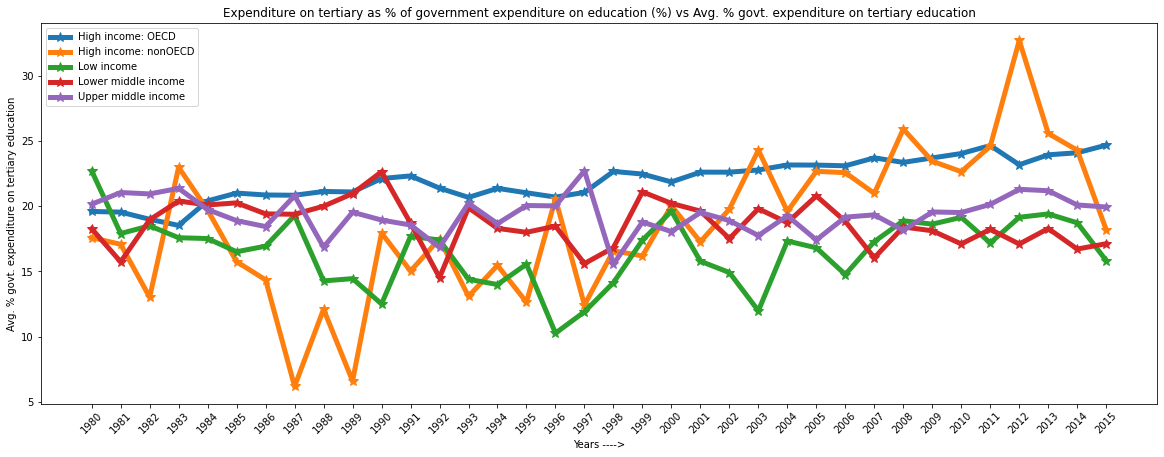
**Observation 2:**

****

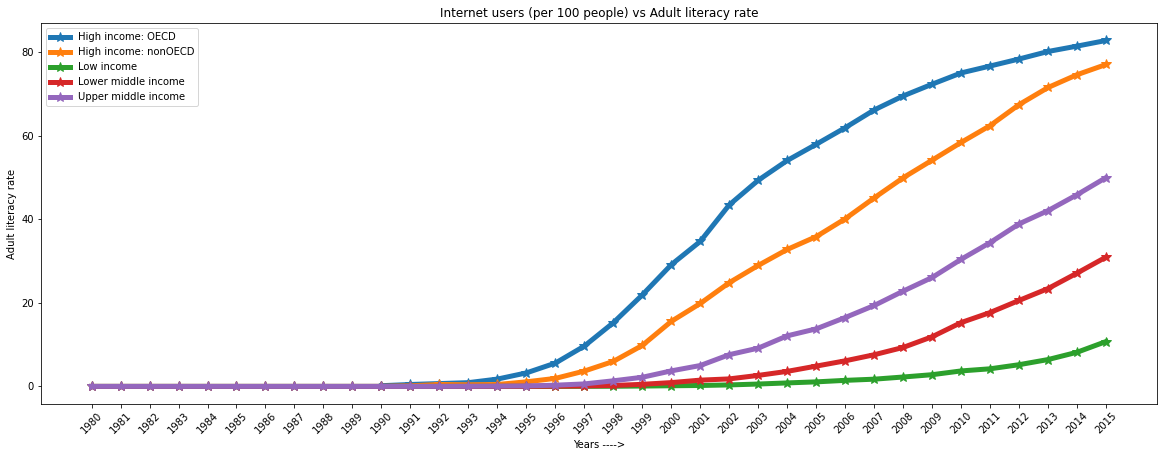
**Observation 3:**

****

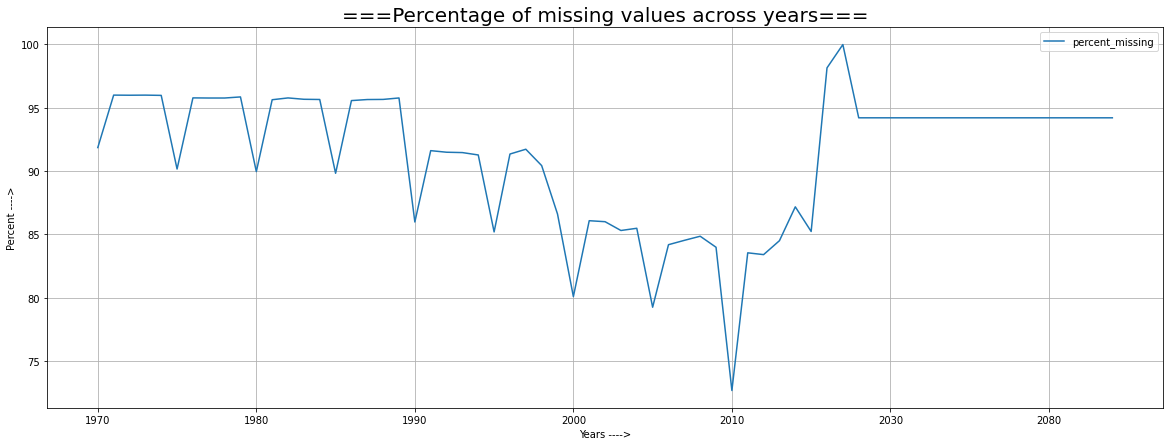
**Observation 4:**

****

**Observation 5:**

****

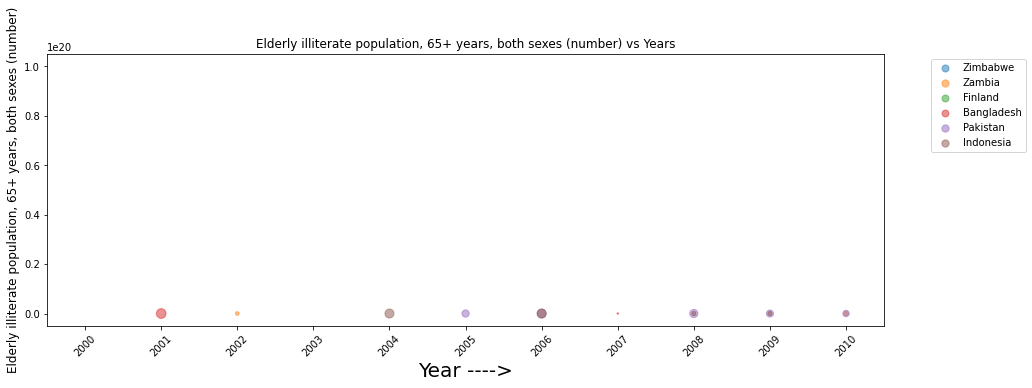
**Observation 6:**

****

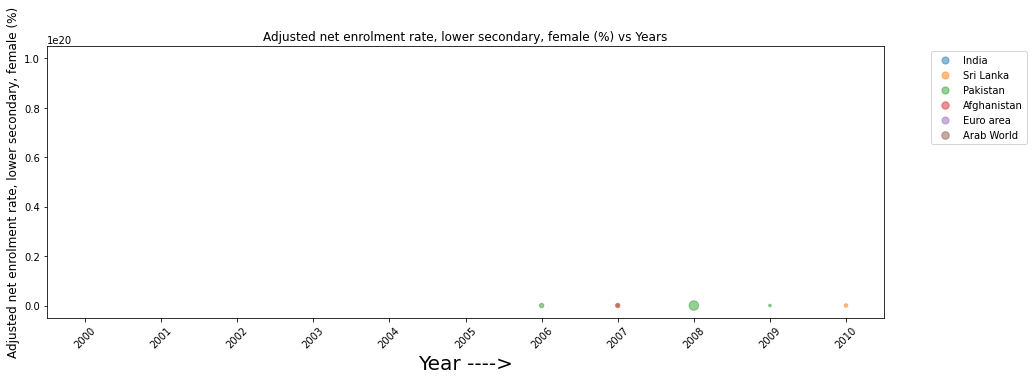
**Observation 7:**

****

**Observation 8:**

****

**Observation 9:**

****

**Future prospect and Conclusion:**

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 3665 indicators to keep for analysis deciding on the set of countries to work based on economy and geography.

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. We can introspect through datasets how student to teacher ratio depends on the socio-economical background of the country.

**Acknowledgement:**

It is my foremost duty to express my deep regards and gratitude to my peers, under whose encouragement and timely support, I was able to undertake this project. I am also grateful to the articles, projects and the tutorial videos provided by the concerned authors, without whose precious guidance the project would not have been successfully done.