



Capstone Project

EDA

On

World Bank Global Education
Analysis

From Beginning to End.

Data Cleaning

Fixing incorrect, corrupted, incorrectly formatted, duplicate and incomplete data within a dataset.

Conclusion

Summary of Facts and Observation recorded from Analysis.

Introduction

A brief note about the project.



Analysis and Visualization

Analysing various indicator.



References



Introduction

Exploratory Data Analysis (EDA) is the approach of analyzing data, gathering and summarizing the important characteristics of the information, and using simple visualizations that make it easier to understand.

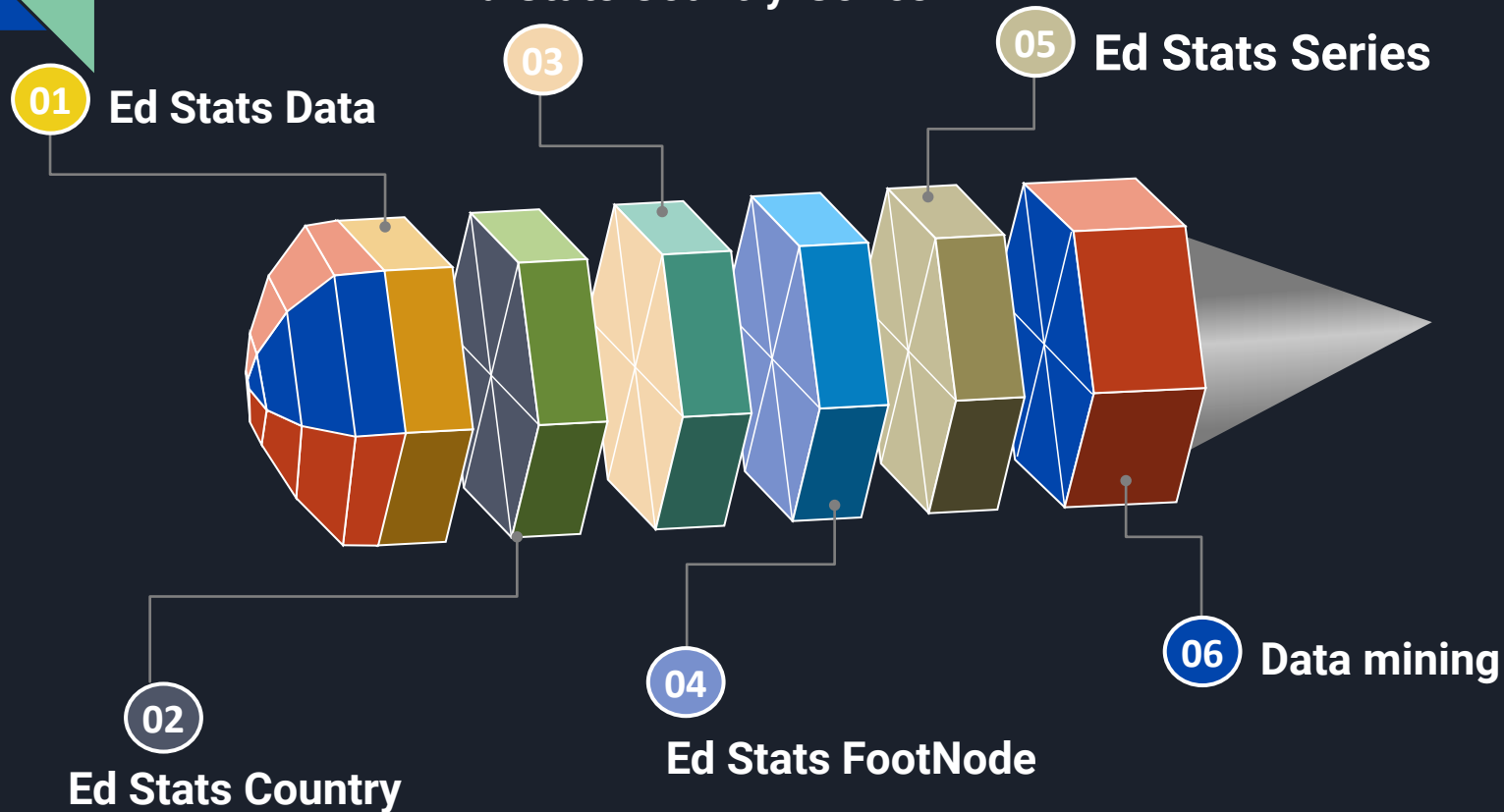


World Bank: Global Education

The world bank group is the largest financier of the education in the developing world. World bank edstats provides the data related to education for every country.



Loading and Discovering data



Data cleaning



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



In this Data Cleaning we had dropped the part of the data which was unnecessary in doing analysis

**IF YOUR DATA COLLECTION IS WRONG,
ANY CONCLUSION IS WRONG!**

Garbage Data In



Analysis Pipeline

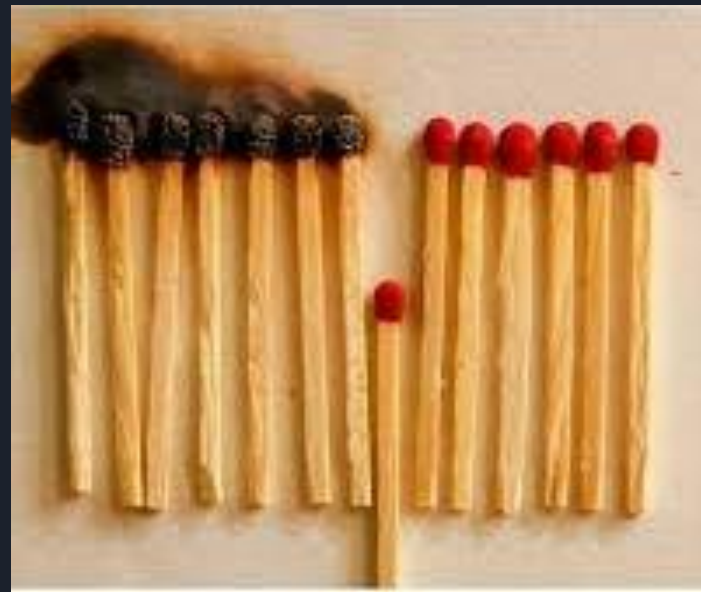


Garbage Data Out



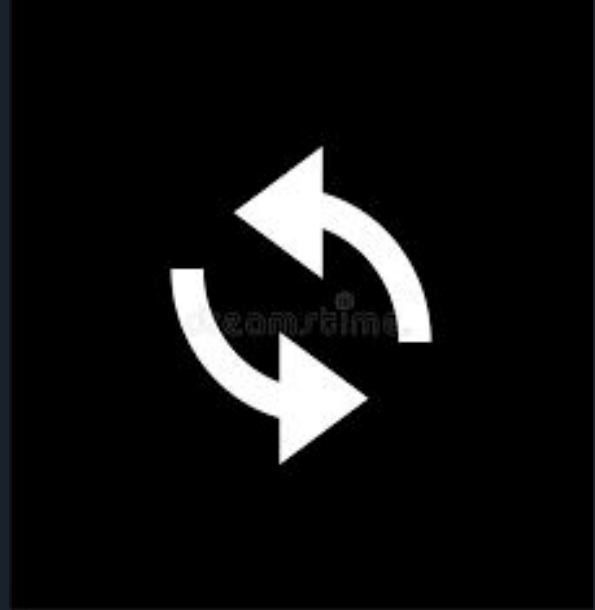
Dropping as needed, is good!


The data columns with all 'NaNs' are dropped, such that the data remains consistent.
We then try to treat and clean the rest of the data as appropriate.



Replacing the NaN values

We had use two approaches, using 'ffill' and 'bfill' to perform upward and downward fill for 'NaN' values and also column wise mean , and row wise mean, as best applicable.





Extracting meaningful chunk out of file for the particular indicator
by dropping unnecessary columns and return a column wise
matrix of country's data over the years


Enrolment in early childhood education, both sexes (number)										
	Year	Australia	Canada	Germany	India	Japan	South Africa	Sri Lanka	United Kingdom	United States
1	1999	NaN	NaN	2332585.0	NaN	NaN	NaN	NaN	NaN	7183122.0
2	2000	NaN	NaN	2297821.0	NaN	3425697.0	NaN	NaN	NaN	7110066.0
3	2001	272996.0	NaN	2398104.0	2617403.0	3463487.0	NaN	NaN	NaN	7538720.0
4	2002	263541.0	NaN	2352829.0	4623168.0	3492456.0	518985.0	NaN	NaN	7667685.0
5	2003	211627.0	NaN	2316687.0	5507559.0	3555478.0	NaN	NaN	NaN	NaN
6	2004	214059.0	NaN	2238270.0	5050006.0	3586547.0	NaN	NaN	NaN	7435568.0
7	2005	221672.0	NaN	2232306.0	4761485.0	3615999.0	685749.0	NaN	NaN	7361682.0
8	2006	212998.0	NaN	2443550.0	5264053.0	3630184.0	NaN	NaN	989596.0	7342289.0
9	2007	212402.0	486874.0	2420124.0	5366340.0	3614874.0	NaN	NaN	1004677.0	7512518.0
10	2008	215526.0	NaN	2410081.0	6576483.0	3606830.0	NaN	NaN	1108148.0	7191333.0

Result of Dropping and Replacing

	Year	Australia	Canada	Germany	India	Japan	South Africa	Sri Lanka	United Kingdom	United States
1	1970	250086.0	889508.0	3876767.0	31490248.0	5095959.0	652143.0	340204.0	1966175.0	11633696.0
2	1971	251115.0	865708.0	3842603.0	31836122.0	5188082.0	672199.0	344428.0	1949416.0	11403609.0
3	1972	251054.0	831302.0	3786292.0	32494288.0	5267009.0	694273.0	349662.0	1906443.0	11035671.0
4	1973	250261.0	790105.0	3697424.0	33118470.0	5336434.0	716971.0	354854.0	1845099.0	10648013.0
5	1974	249254.0	749323.0	3575594.0	33700136.0	5409003.0	738413.0	358413.0	1775736.0	10291870.0
6	1975	248444.0	718258.0	3413043.0	34240160.0	5542572.0	756518.0	358572.0	1729657.0	10095371.0
7	1976	250874.0	707713.0	3236478.0	34743112.0	5645246.0	777599.0	358283.0	1670245.0	9929981.0
8	1977	254612.0	702801.0	2964479.0	35570288.0	5817110.0	797782.0	356292.0	1598068.0	9894422.0
9	1978	258208.0	706416.0	2671381.0	36402984.0	5981775.0	816670.0	354115.0	1523781.0	9901996.0
10	1979	259626.0	714253.0	2407797.0	37193140.0	6087079.0	833902.0	354968.0	1454956.0	9922180.0

Analysis and visualization





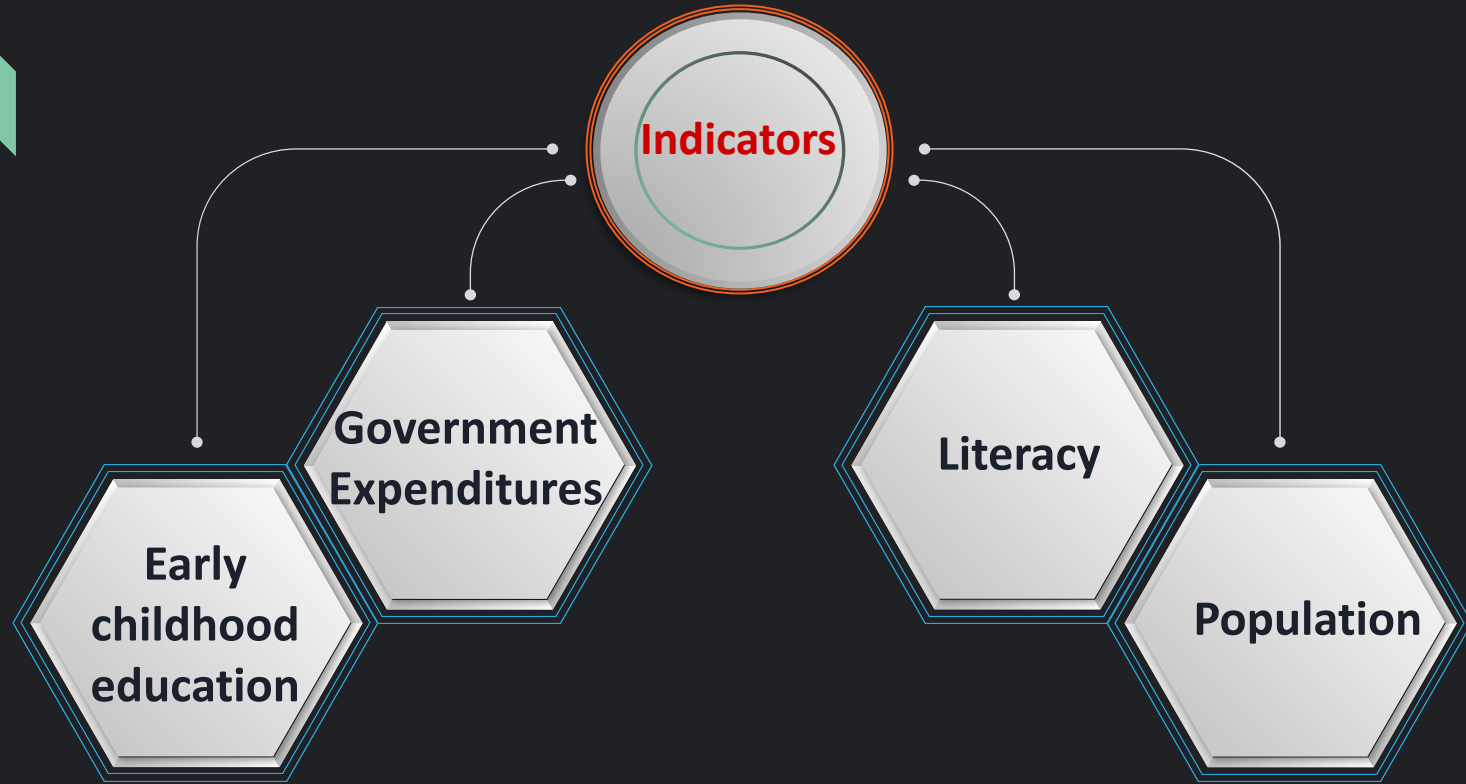
Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data.



We then use 'Data visualization', to create charts, graphs, or other forms of visualization, which makes information easier to analyse and interpret.



Types of Indicators



Indicators in Depth

Early Childhood Education

- Enrolment in early childhood education, both sexes.
- Percentage of enrolment in pre-primary education in private institutions.

Government 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.

Population

- School age population, pre-primary education, both sexes Education Equality.
- School age population, primary education, both sexes.
- School age population, secondary education, both sexes.
- School age population, tertiary education, both sexes.



The Countries in focus



Canada

United Kingdom

Germany

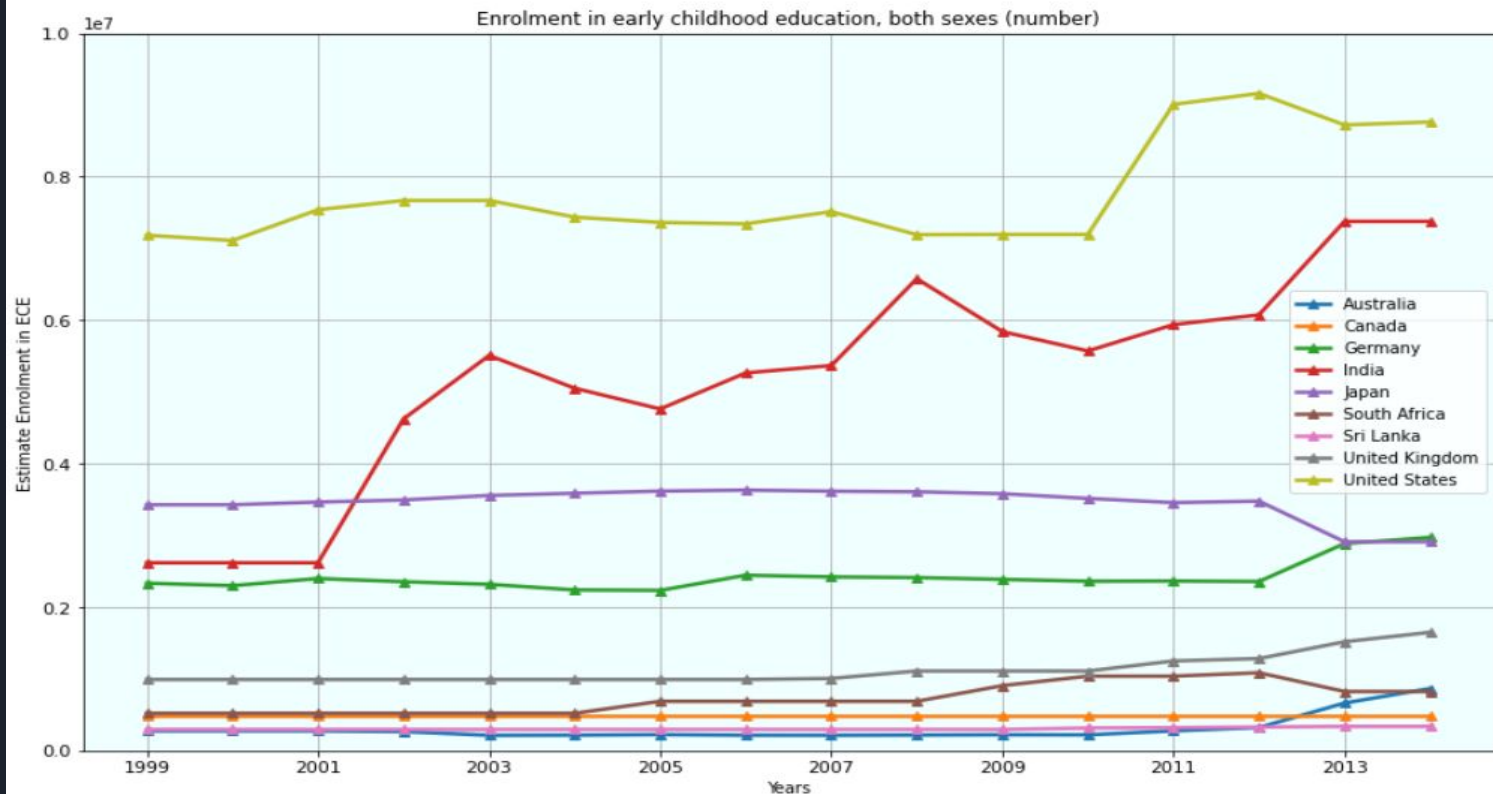
India

Japan

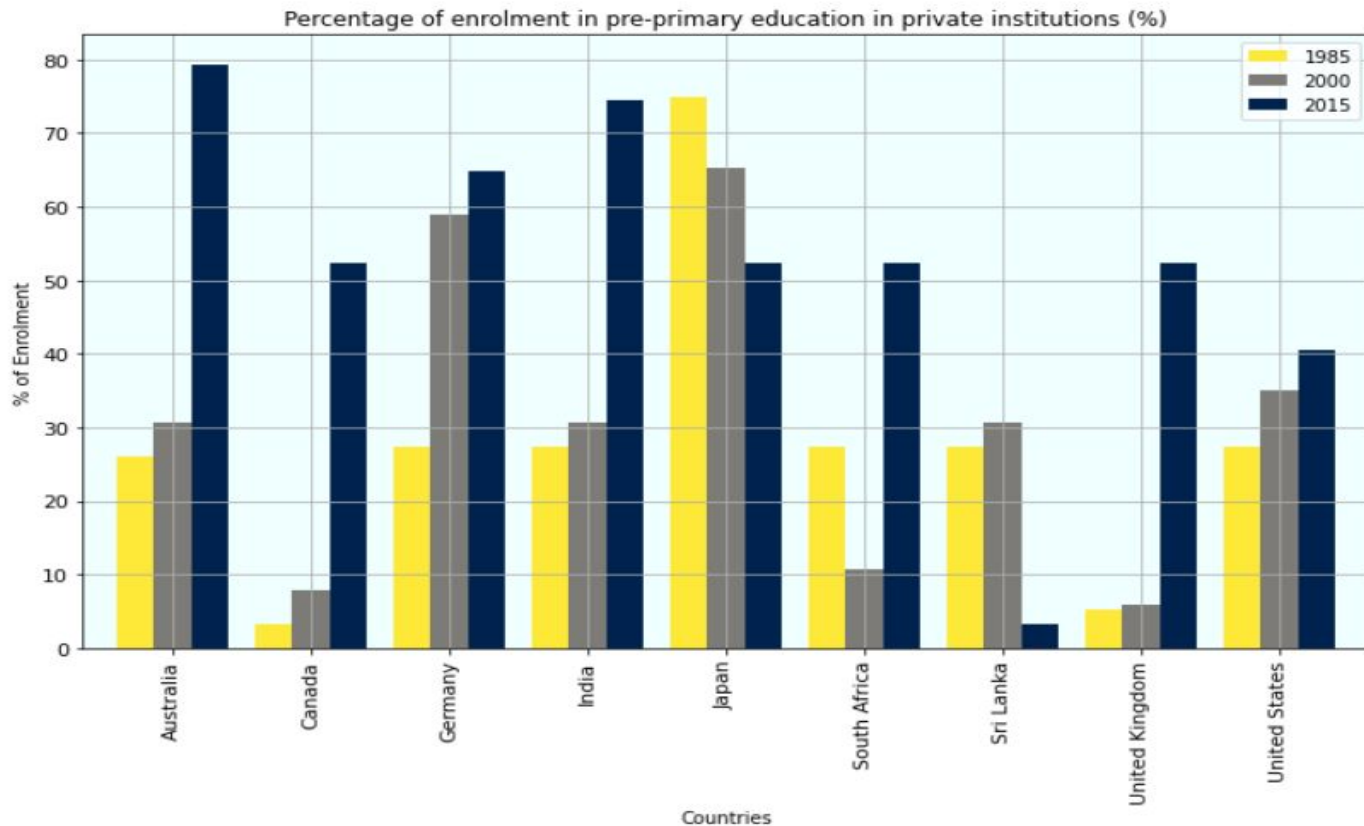
United States

South Africa

Australia

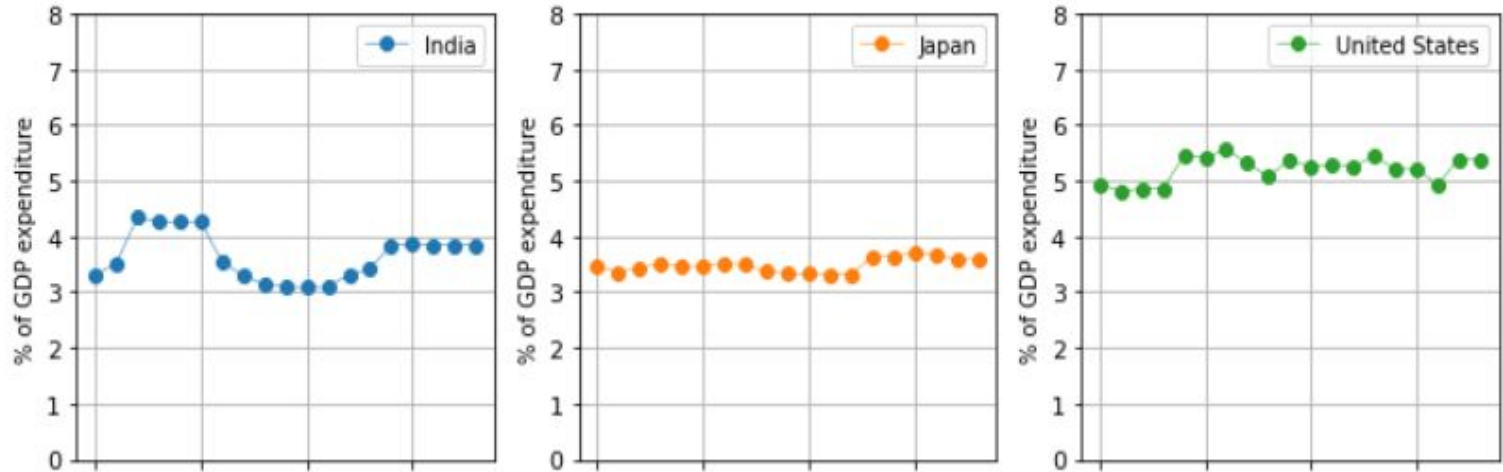


For India, we observe significant leaps in the periods 2001-2003, 2005-2008 and also at 2010-2013, and we are able to notice that the slopes and degrees of hills are more significant than the slopes and degrees of the valleys, which shows positive signs of growth. India shows a positive and promising uptrend.



As you can see from the graph that the Japan and Sri Lanka are only the countries where enrollment in pre primary education in private institutions decrease during the time interval of 1985-2015. As for the rest of the countries the enrollments for pre-primary in private institutions increases.

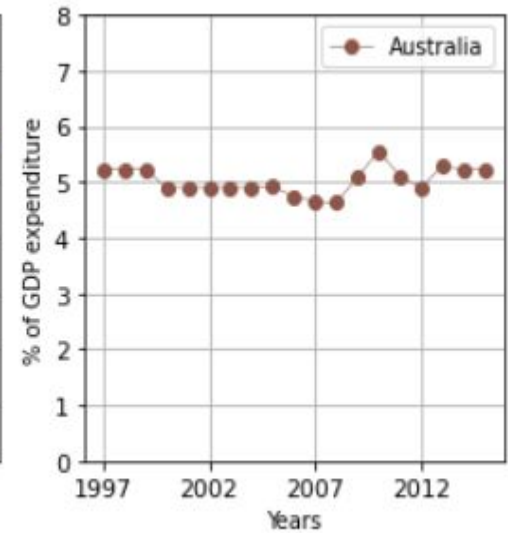
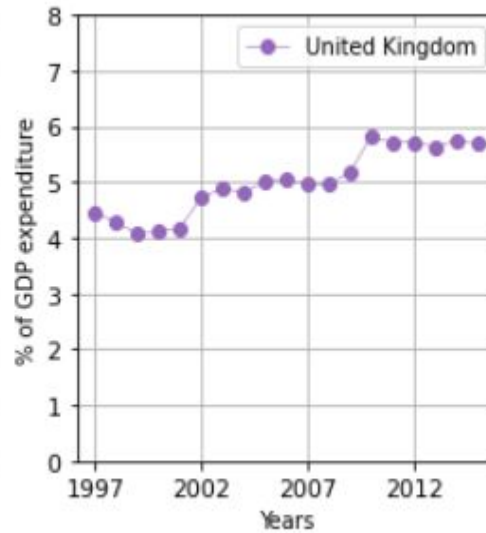
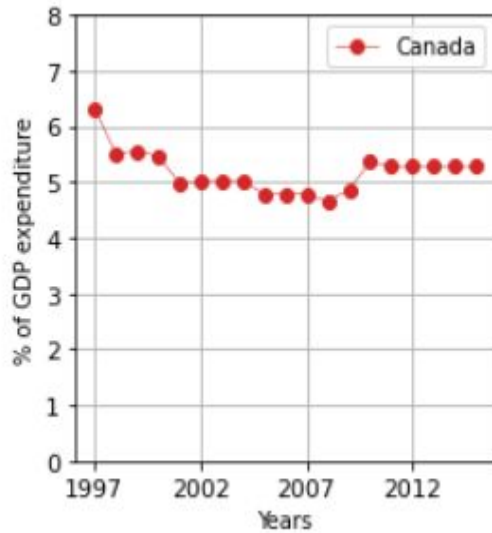
Government expenditure on education as % of GDP (%)



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. In the period 2007-2011 we see a consistent increase, post which it remains stable.

Japan: The graph seems to be more 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.

United States: We observe the graph remains consistent during the period 1997-2001, shortly followed by a leap, and significant fluctuations happened throughout the rest of the graph.

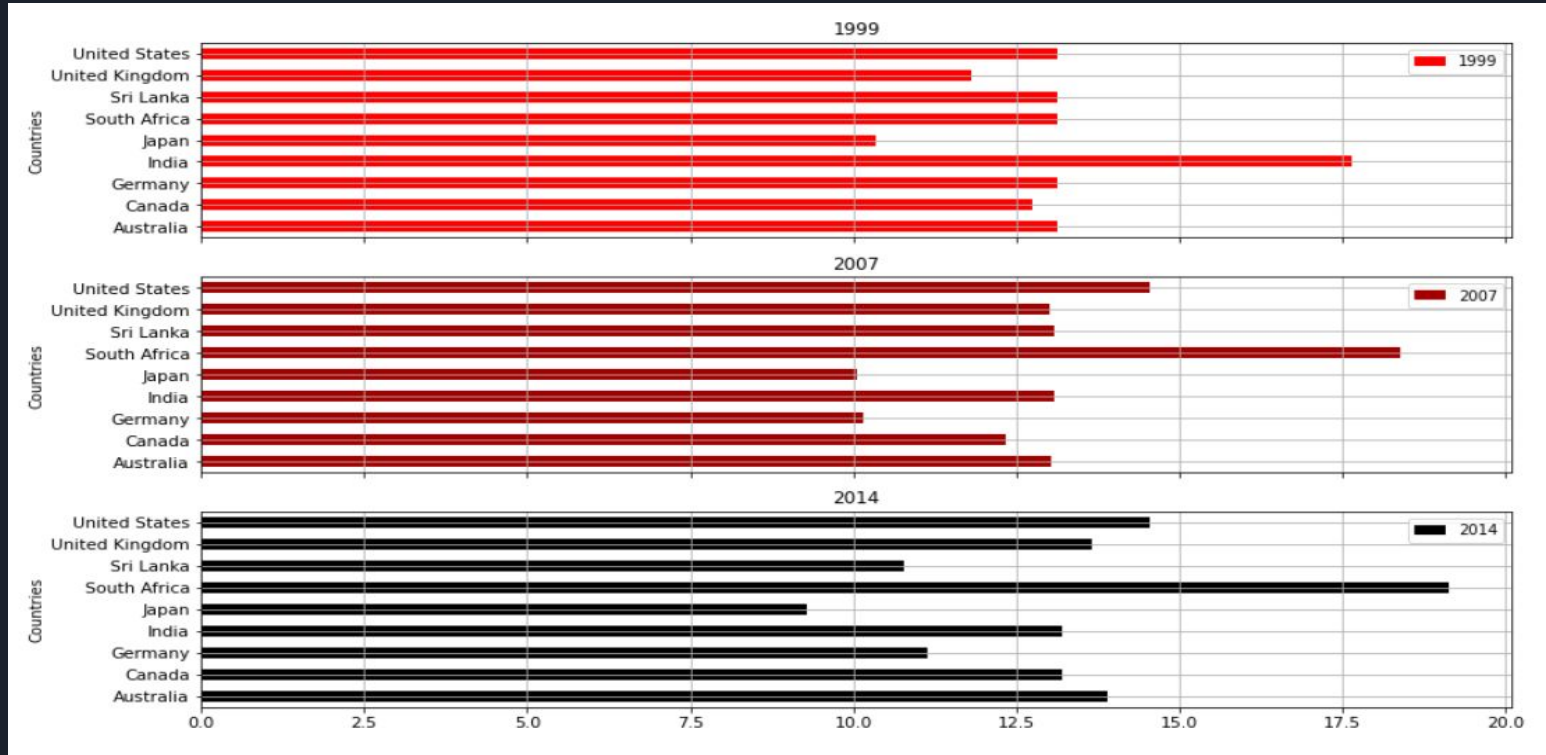


Canada : we observe a declining trend during the period 1997-2008 and for the period 2008-2011 there is a small climb, post which the graph remains stable.

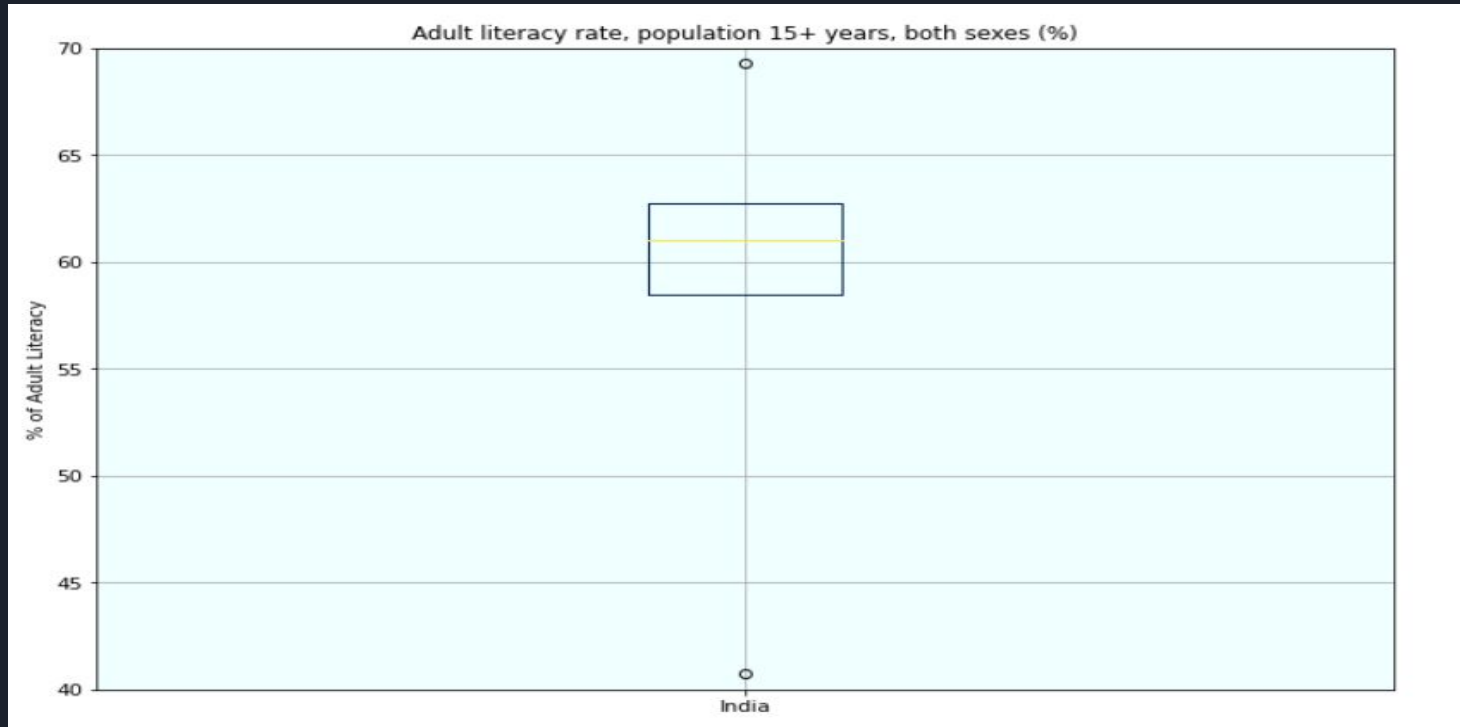
United Kingdom : Initially we observe a decline in the period 1997-1999. After which the graph is going up. We see two positive leaps, one at 2001-2002, and the other at 2009-2010, and minor fluctuations for the the other periods.

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 (%)

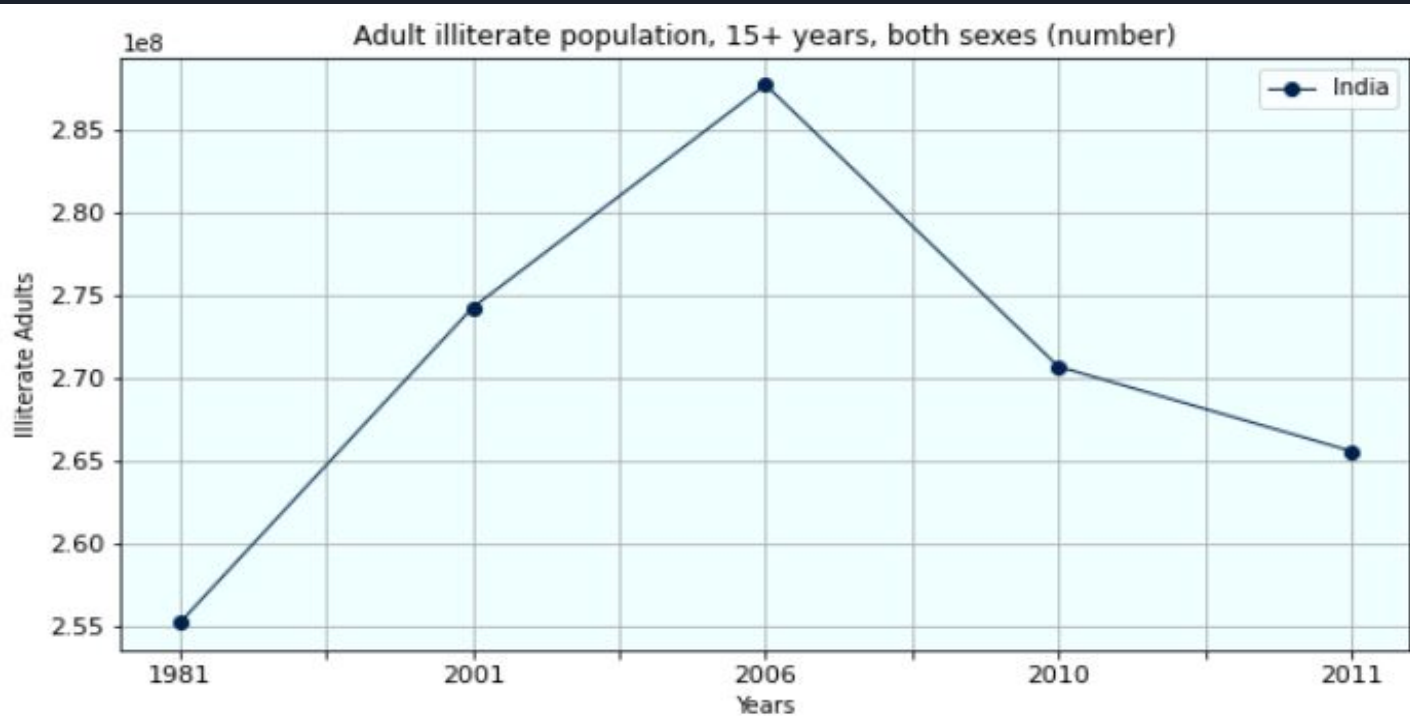


From the analysis we can conclude that the expenditure on education of India and Sri Lanka decreased in 2014 than that of 1999. Reverse pattern can be seen for South Africa, their expenditure is increased to approx 18%. United States and United Kingdom are consistently developing well over the years.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



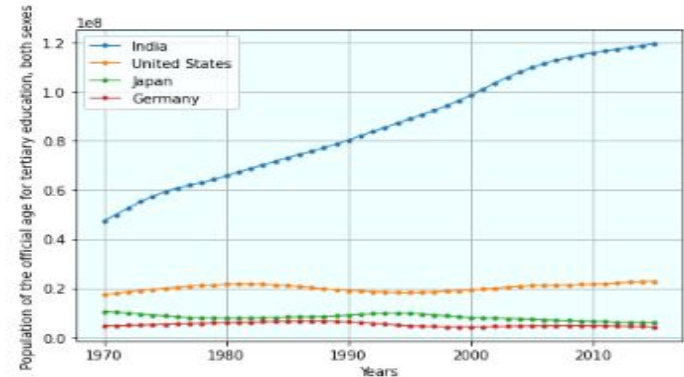
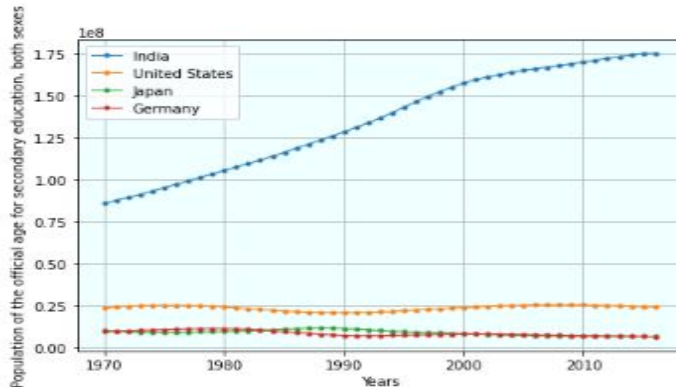
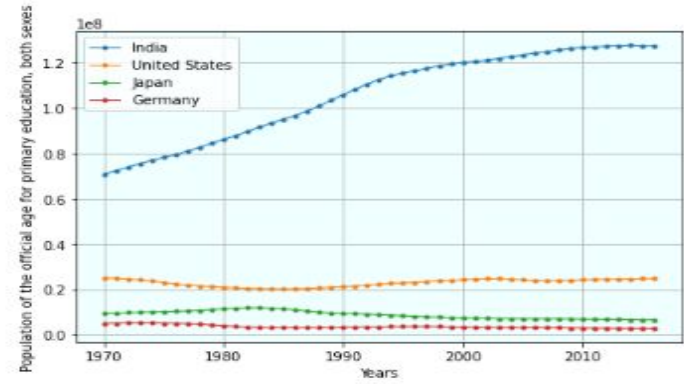
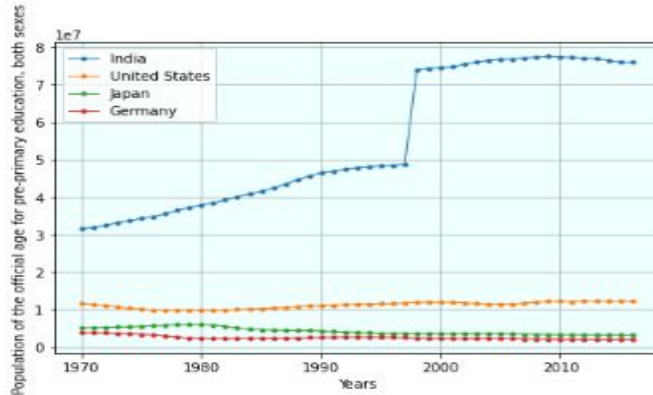
We are able to see that the highest literacy rate for India, has an approx magnitude of 70, and the lowest being approx 40. The median literacy rate, which has approx magnitude of 62, lies towards the higher end for India, which represents a positive growth.

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



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: we also observed a steep decline in the period 2006-2010, followed by a steady decline from 2010-2011, which shows a positive sign in the overall period through 2006-2011.

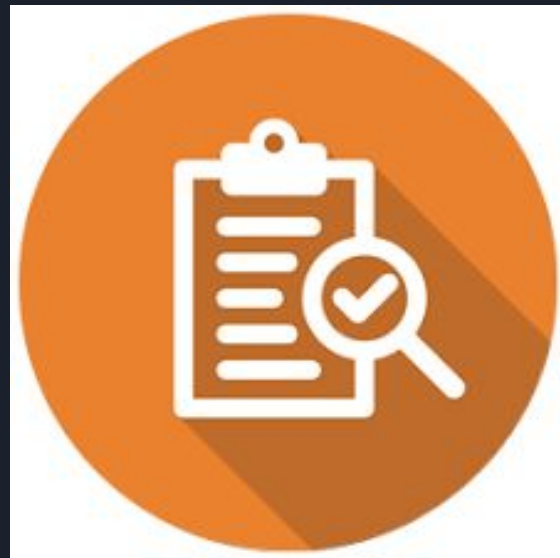
Based on the observation earlier, we see that as 'Government expenditure on Education' is going up during 2007-2011, the adult illiterate population metric is declining during the same period.

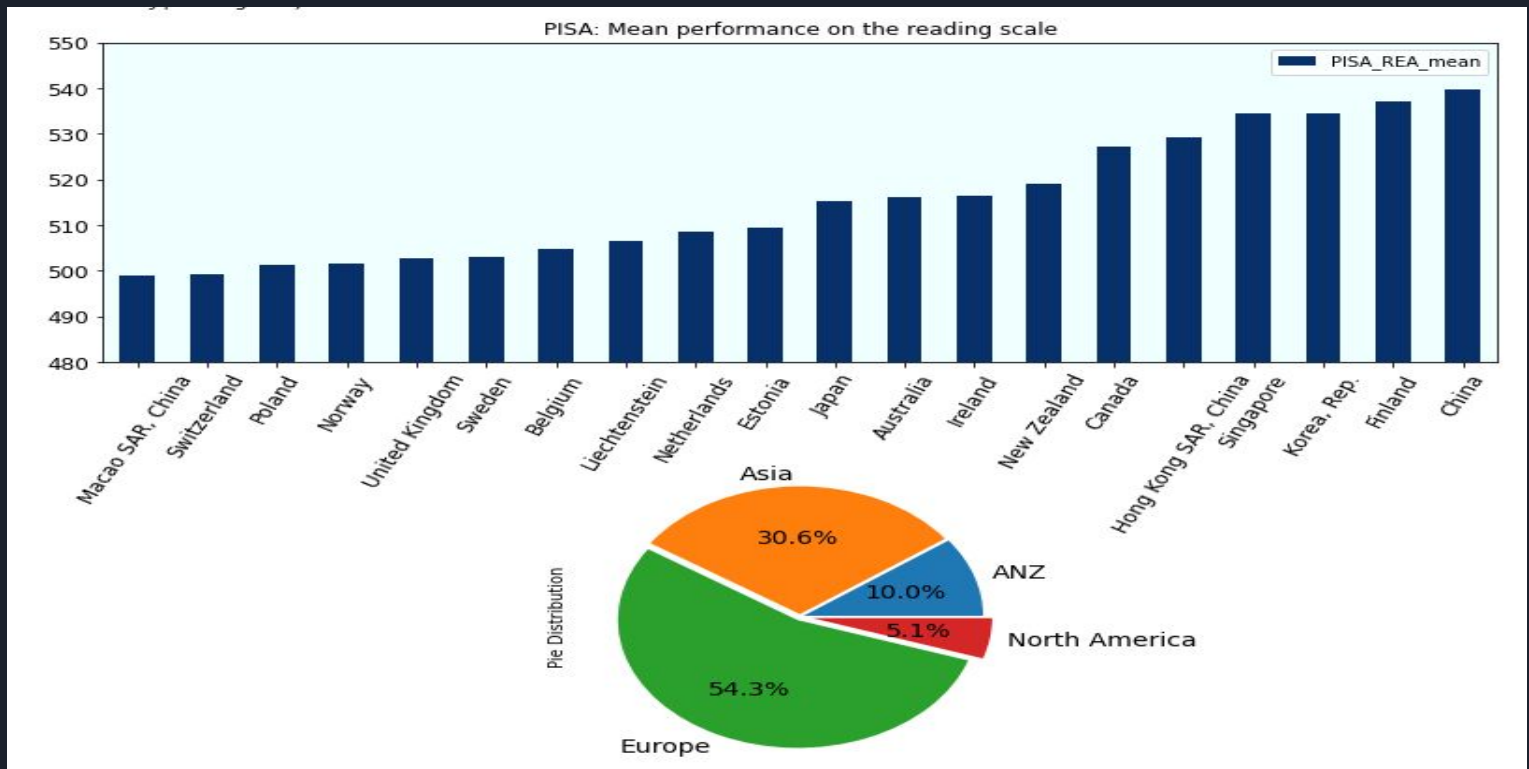


For india, We observe a significant leap in the population of official age for pre-primary education, in the period 1997-1998. We are also able to see that across all the graphs, india is on a consistent incline. For the countries Japan, United States and Germany, we see an overall stable trend with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

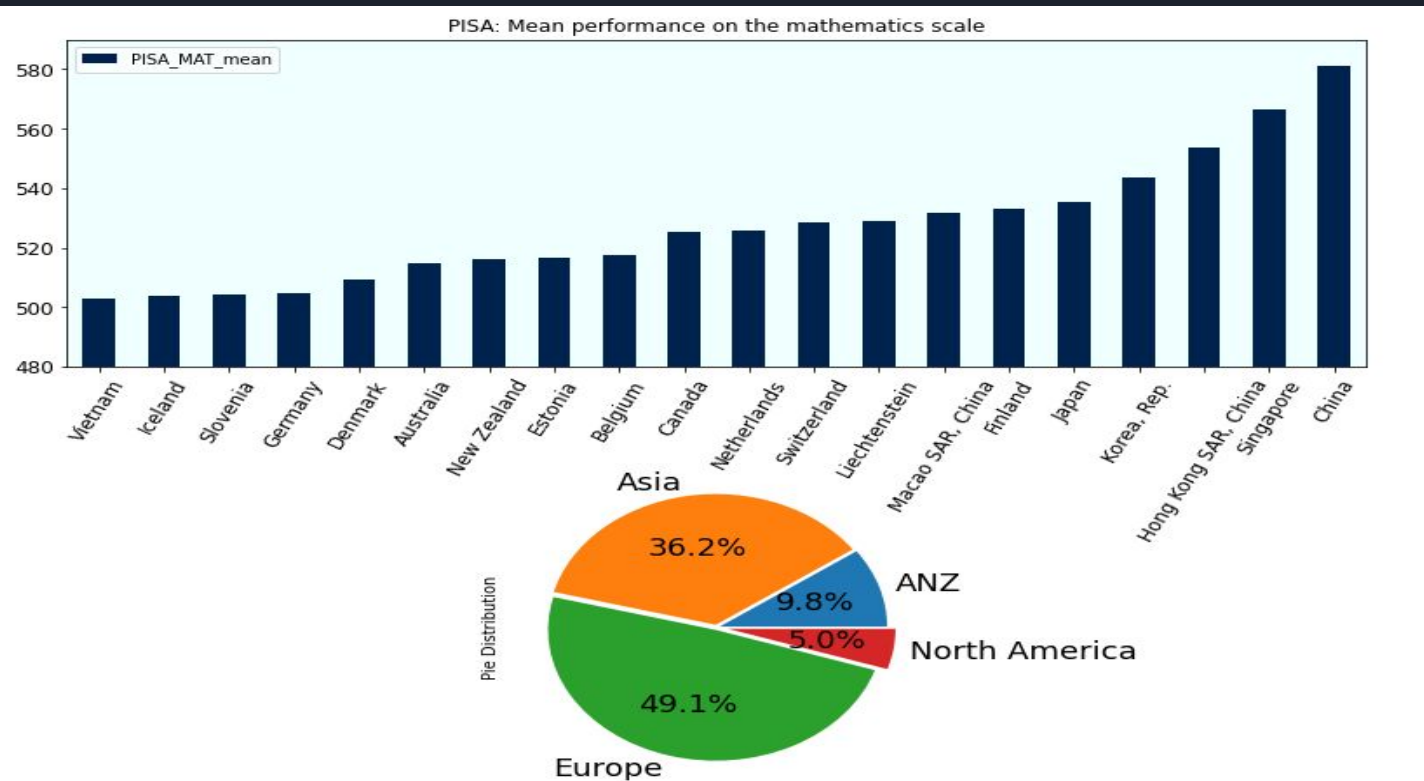
Learning outcomes

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.





For the top 20 countries we have predominantly European countries (11 out of 20 are European Countries), followed by 6 Asian countries, 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 second is the only European country which is in the Top 5 category.

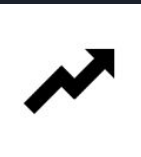


For the top 20 countries we have predominantly European countries (10 out of 20 are European Countries), followed by 7 Asian countries, Australia and New Zealand from the ANZ region, and Canada from North America. However, considering just the top 5 countries, we can observe that all 5 of them, are Asian, specifically in the East/South-East Asia regions. It is noteworthy that 5 out of 7 Asian countries, are in the Top 5, and all Asian Countries except Vietnam, are in the Top 10.

Conclusion



- 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 enrollment during the 2015 has been 2.5X than that of the year 2000.
- The govt 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, than 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.





References

- **Education Statistics (EdStats) 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**

Thank You