Capstone Project
World Bank Education Analysis

Team Insight Seekers

Anirudh Upadhyay
Fahad Mehfooz
Pooja Rana
Saifuddin Raja
Shubham Bareja
Varun Nayyar

Steps Involved:

1. Problem Statement
2. Data Summary
3. Data Preparation
4. Null Value Treatment
5. Exploratory Data Analysis (EDA)
6. Conclusions
7. Challenges Faced
8. Future Scope of Work

1. Problem Statement:

We are provided with a dataset named World Bank EdStats that contains 3665 indicators describing education access (primary, secondary and tertiary), GDP, literacy rate, population growth and technical aptitude (PISA, TIMSS, PIRLS).

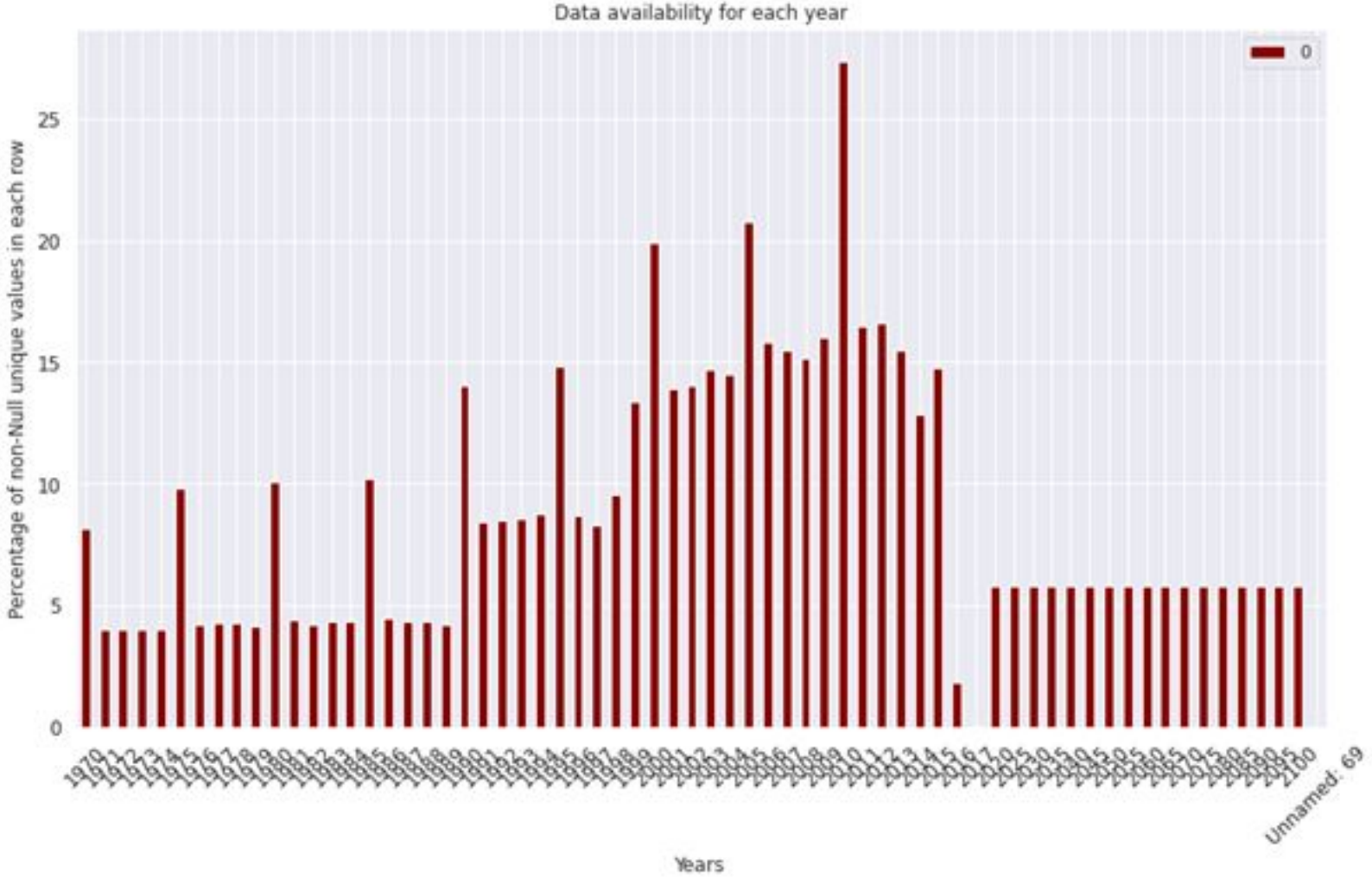
The idea of the project is to analyze the data, mark the variation of indicators across the globe and group the countries according to their similarities and differences.

2. Data Summary:

The Data set is composed of 3 Main data store files ,
which are briefly described below :

1) EdStatsData (886931 rows and 69 columns)

This is our primary file. This dataset is the one which contains the scores for all countries for various indicators over the years starting from 1970. It also contains the projections for multiple indicators for future decades. The dataset has peculiar features. A majority of values are either missing or null. Given below is a graph displaying the distribution of data for every year.


Data availability for each year
0
Percentage of non-Null unique values in each row
0
5
10
15
20
25
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
Unnamed: 69
Years

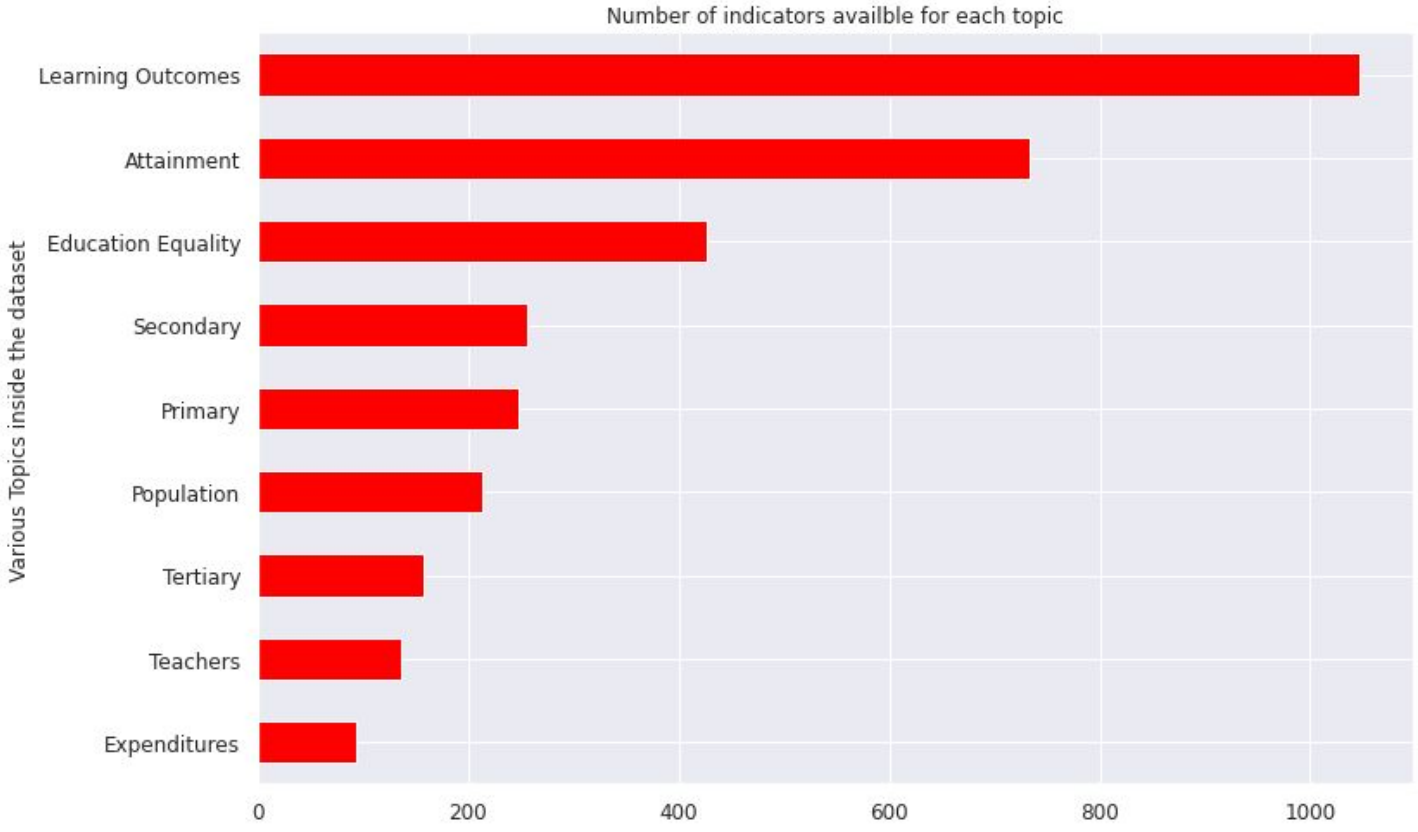
2. EdStatsSeries (3666 rows and 20 columns) –

EdStatsSeries carries all information about the indicators which are used by the World Bank to monitor progress of economies. There are approximately 3700 indicators used to determine which countries have improved in terms of education and other parameters. The dataset contains the description of the indicators, their units of measurements and their sources.

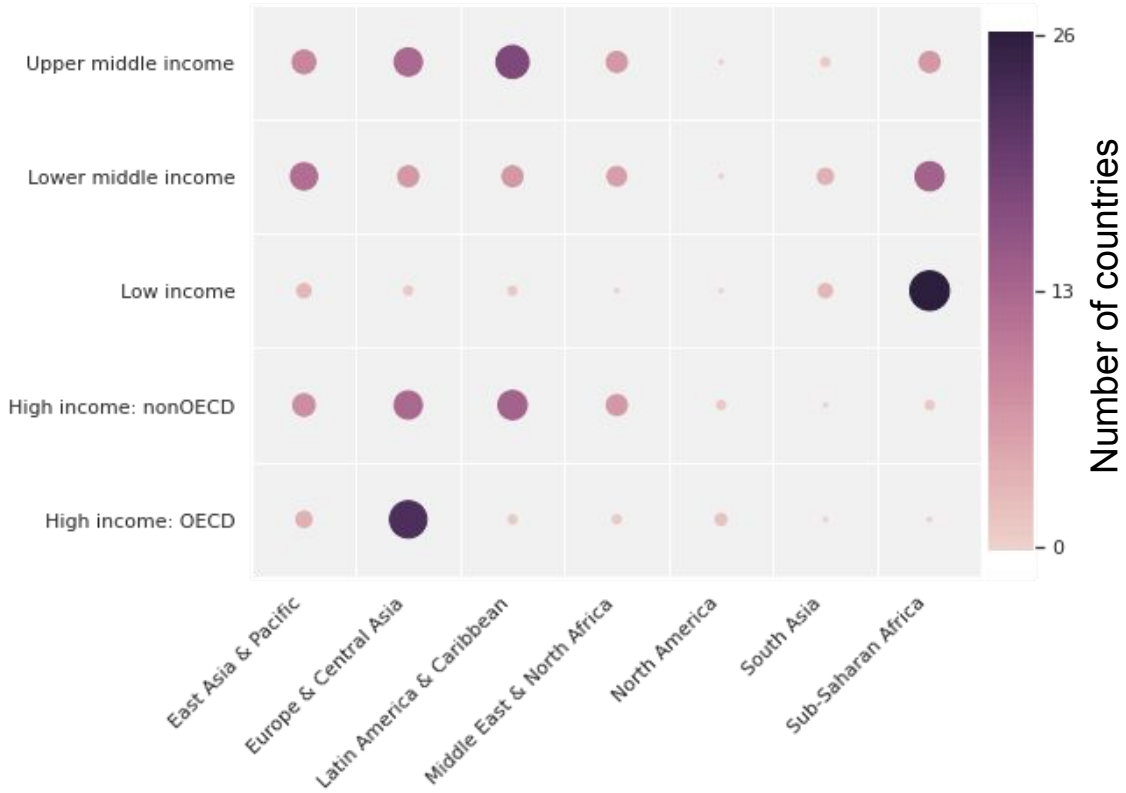
3. EdStatsCountry (242 rows and 31 columns)

EdStats Country carries important information of all the countries that we will encounter in the dataset. Indicators such as latest population census, region and income group of a particular country are provided.

Segmentation of indicators into major groups (Topics)



Comparative analysis of Geographic-Zones and Economic-Groups



3. Data Preparation:

After exploring the dataset, we tried to form a key approach to process the dataset to find the underlying patterns across all countries for all the indicators.

We performed all these operations using: `pandas, matplotlib, seaborn and plotly`

Our main goal was to:-

1. Remove the futile data.
2. Make a befitting dataset.
3. Identify Patterns.
4. Develop Hypothesis.

4. Null Value Treatment:

As we had a large dataset so in order to abridge it, we formulated a way to remove extraneous data.

We removed the future projection data , as it does not hold real information and only contains future predictions so we have discarded data after 2015 till 2100

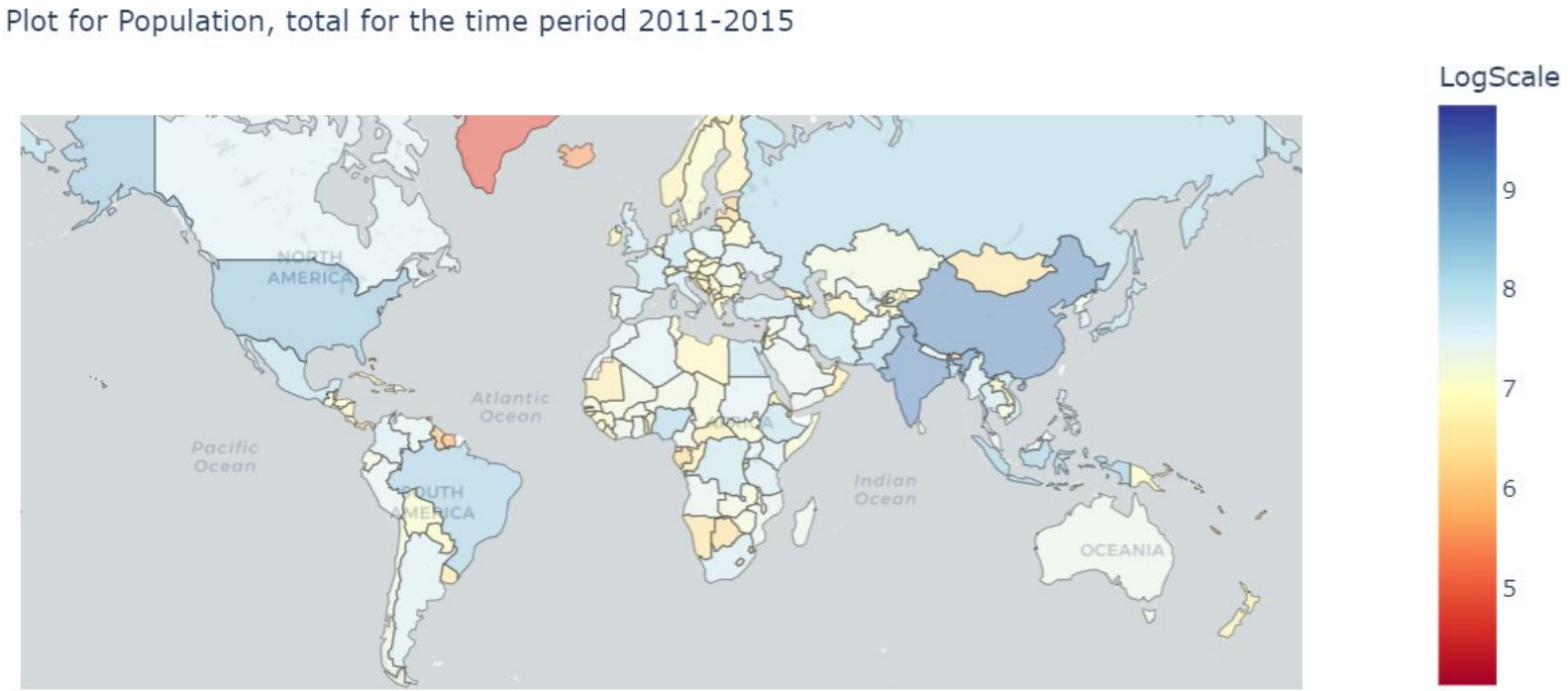
NaN Values:

- We dropped all the rows which have null values in every year column.
- This resulted in a remarkable reduction (almost 60%)
- The final dataset had around 356K records.

General Overview of World GDP

Given above is the plot for GDP (PPP) on the logarithmic scale.
From the Geo-plot it is evident that USA and China have the highest GDP,
Followed by European Countries ,
African countries have the lowest GDP

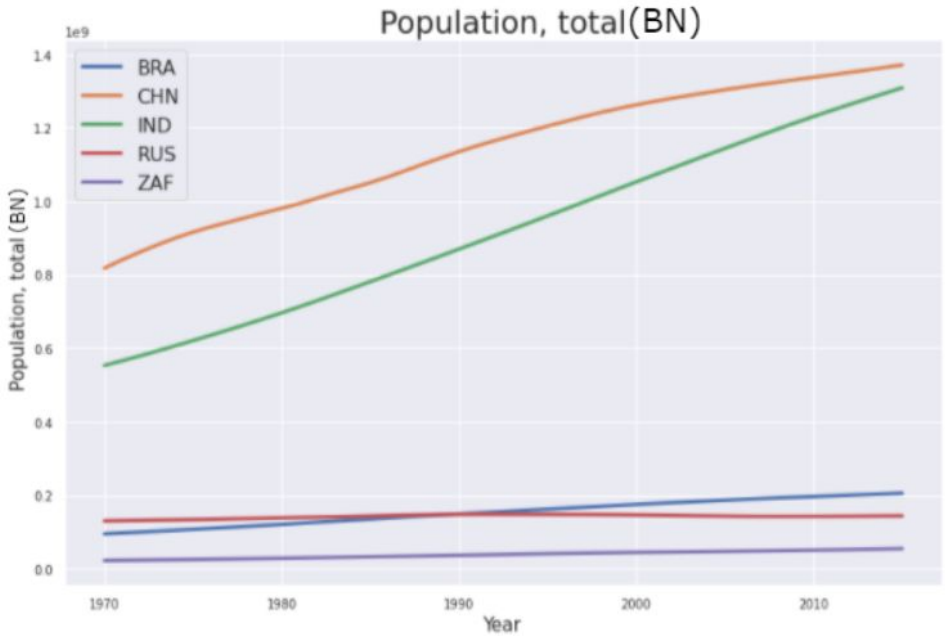
General Overview of World Population



Given above is the plot for total population on the logarithmic scale.
From the Geo-plot it is evident that Asian countries are highly populated and African and European countries have less population

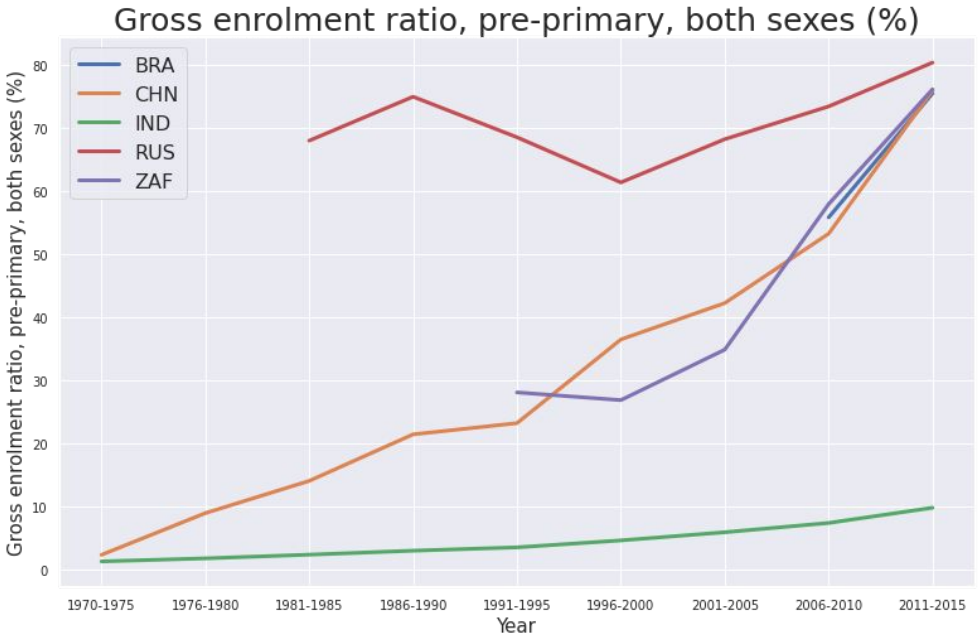
EDA for BRICS Countries

1. Population Analysis



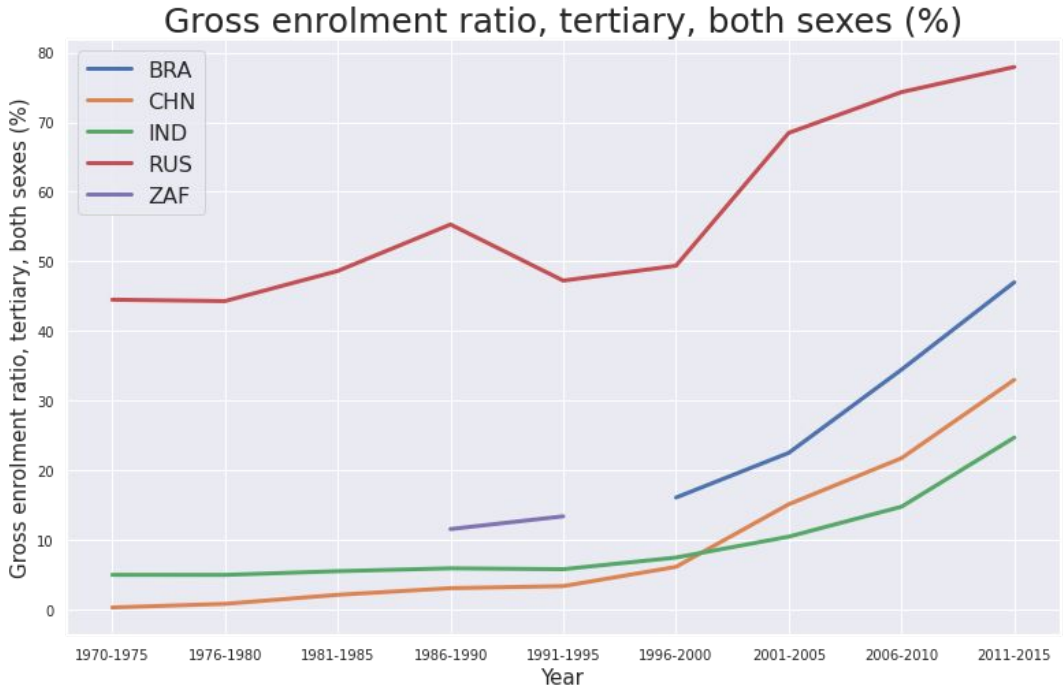
- As depicted by the graph, the population trends for INDIA and CHINA are the similar. Same can be said for the other 3 countries.
- The populations for Russia, SA,
  Brazil stayed relatively stable.

2. PRE-PRIMARY GROSS Enrollment Ratio Analysis



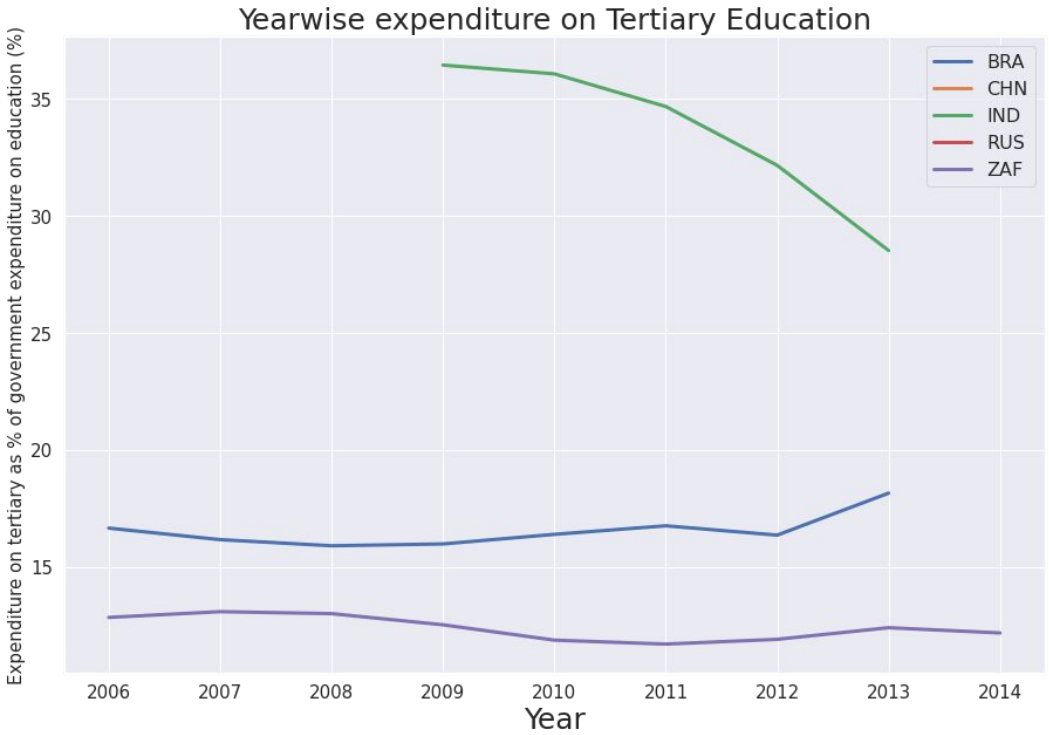
- INDIA performs poorly when it comes to enrollment of children in pre-primary classrooms
- The enrollment % for India is merely 10% towards the end of 2015 whereas it is roughly 80% for other BRICS members

3. TERTIARY GROSS Enrollment Ratio Analysis



- From this graph we can see that Russia has always had a greater percentage of its young population in the universities and colleges(around 80% of youth aged between 16 and 21 in 2015).
- The number of young adults in colleges has been rising for almost every BRICS country throughout the last 4 decades

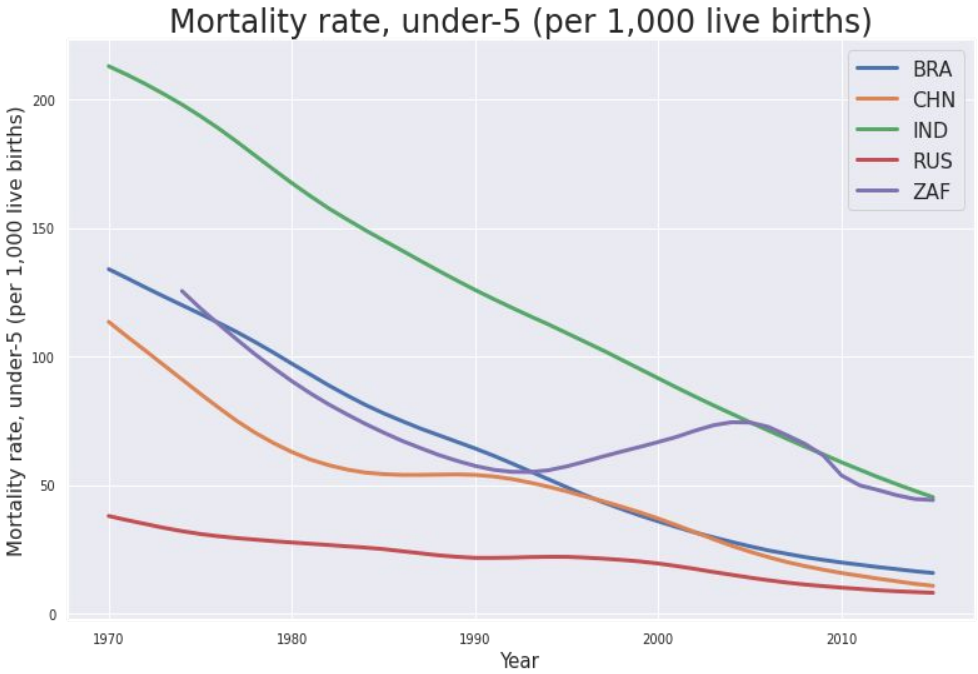
WHY WE SAW WHAT WE SAW?

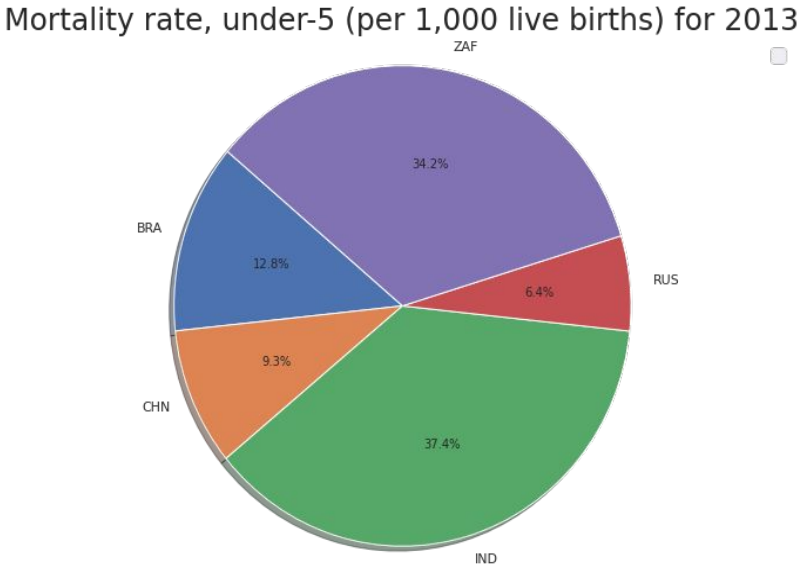


- From this graph we can see that
  the expenditure on tertiary education (%) was highest in India. This could help explain as India is the only country with higher Tertiary Engagement from pre-primary than other countries.

- The opposite is true for Brazil, its pre-primary engagement is highest, but tertiary is far lower, because it has very low expenditure in that area.

4. Mortality Rate Analysis

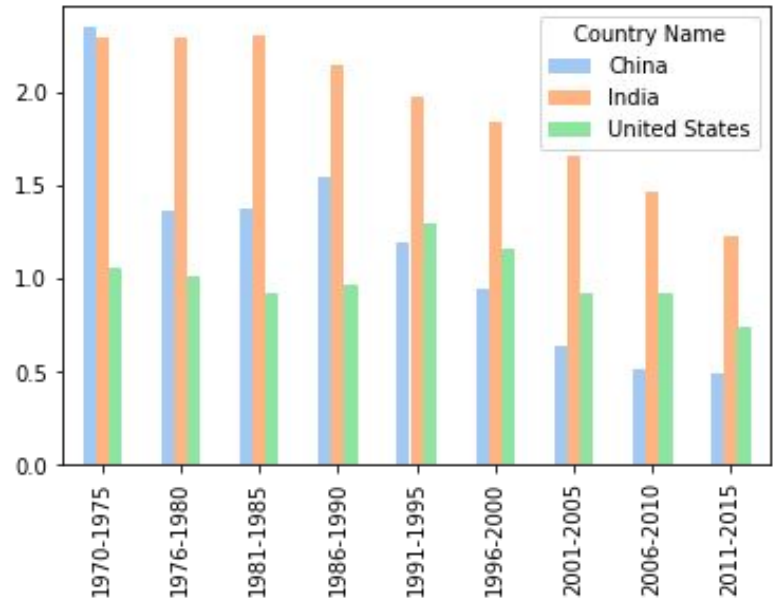




As countries become more educated, the mortality rates drop, and the rate of drop is higher in countries where more focus was laid on tertiary education for previous generations.

EDA for INDIA, CHINA & USA

1. Population Growth Rate Analysis



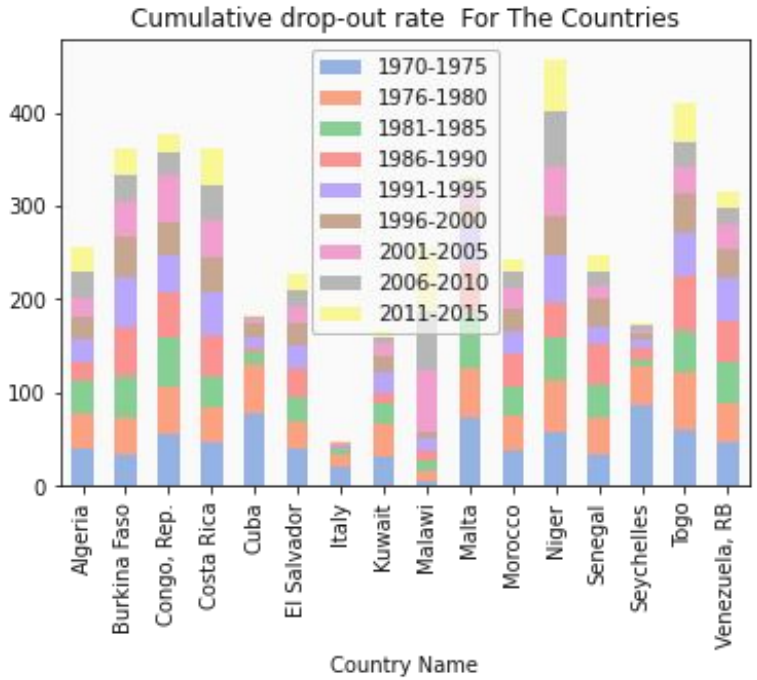
USA:The 1990 to 2000 population increase was the HIGHEST.

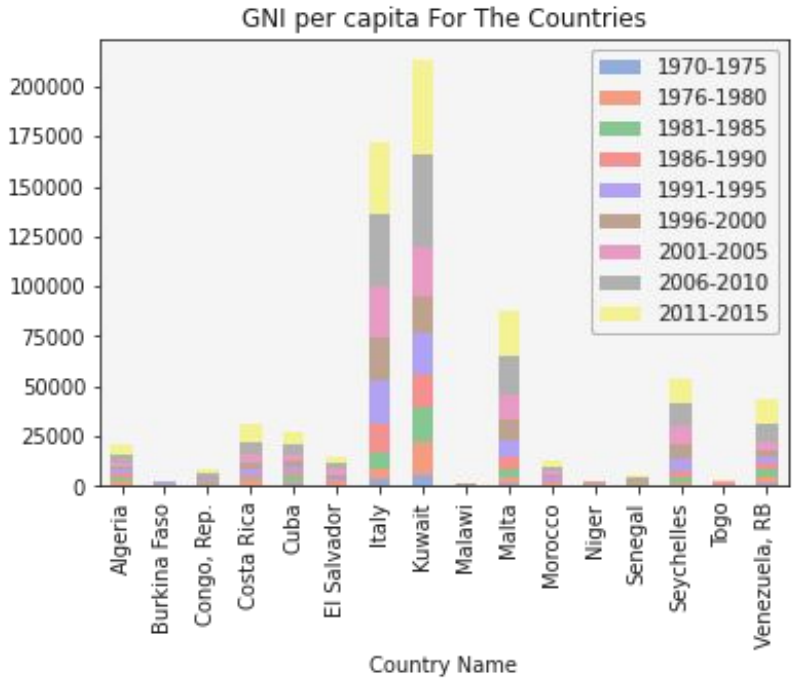
While immigration played an important role in the population surges in all three areas, a large part of the increase also was due to domestic migration and rising birth rates

Plot for Population growth (annual %) for the time period 2011-2015

Given above is the plot for population growth for Asia on the logarithmic scale.
From the Geo-plot it is evident that **India surpassed China** in terms of growth rate in the last decade.

2. GNI v/s DROP-OUT RATE ANALYSIS





- Took the countries where GNI per capita Indicator was available
- The countries having the highest dropout rate affected the GNI per capita for those countries. In case of country Togo, GNI per capita got highly affected

Conclusions

- From the BRICS countries, we can conclude that :
  - The populations of India and China showed growth over the years, with India's growth rate overtaking China's after 2000s.
  - India's pre-primary and tertiary enrollment percentage is pretty low when compared to other BRICS countries, but increasing from 1995.
  - Impact of year-wise expenditure by government on tertiary education shows the increase in tertiary enrollment percentage
  - The mortality rate has dropped for every country, the highest drop being observed in India where more focus was laid on education over the years
- Immigration in the USA from 1990-2000 increased the population growth rate for that 10-year time period
- The countries having the highest dropout rate affected the GNI per capita for those countries

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 compare 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 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.