Capstone Project

World Bank Global Education Analysis

By
Tejas Thakur and Aditya Singh

Contents

Problem Statement

- Analysis of Correlation between different indicators.

- Selection of Indicators to perform further analysis.
- Analysis of performance of different regions based on indicators (GDP, Literacy, PISA, Enrollment ratio etc).

- Analysis of countries based on different indicators (GNI, Literacy, PISA, Enrollment Ratio etc).

- Finding valuable insights from the above analyses.

Summary Of Data

Main Dataset Name - World Bank Education Statistics All Indicator Query containing 4000 internationally comparable indicators that describe education access, progression, completion, literacy, teachers etc. from the years - (1970 - present) and projections till 2100.

The main dataset has five sub datasets. Below contained value of rows and columns are for two custom datasets used in the analysis.

Shape -

Dataset Name	Rows	Columns
Main_df	886390	52
Edstats_country_df	241	32

Important Columns - Country_Name, Country_Code, Indicator_Name, Indicator_Code, Region, Income_Group, Years(1970-2020).

Missing Data

Preliminary Data Inspection

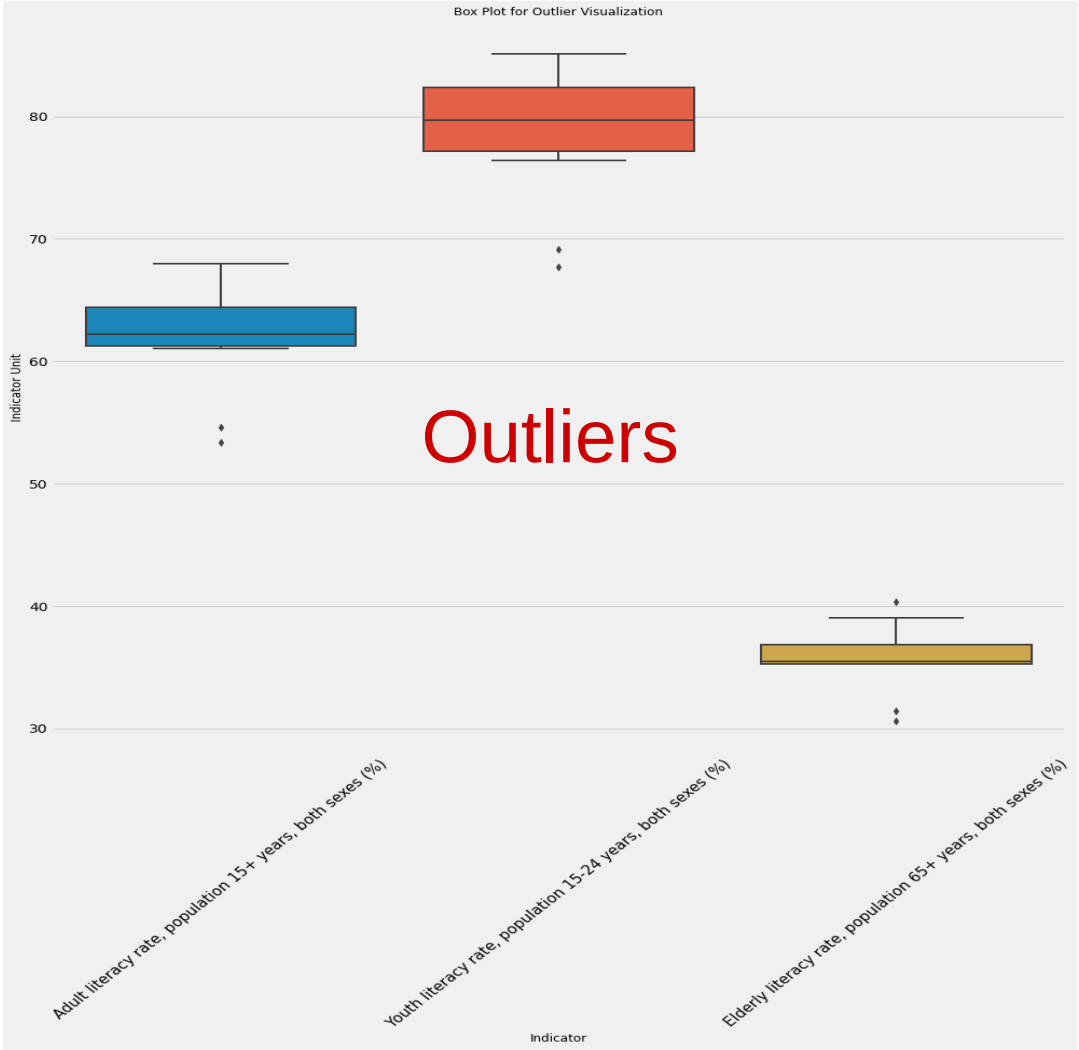
head(),tail(),info(),describe(),
shape,columns

Drop Most Irrelevant Rows

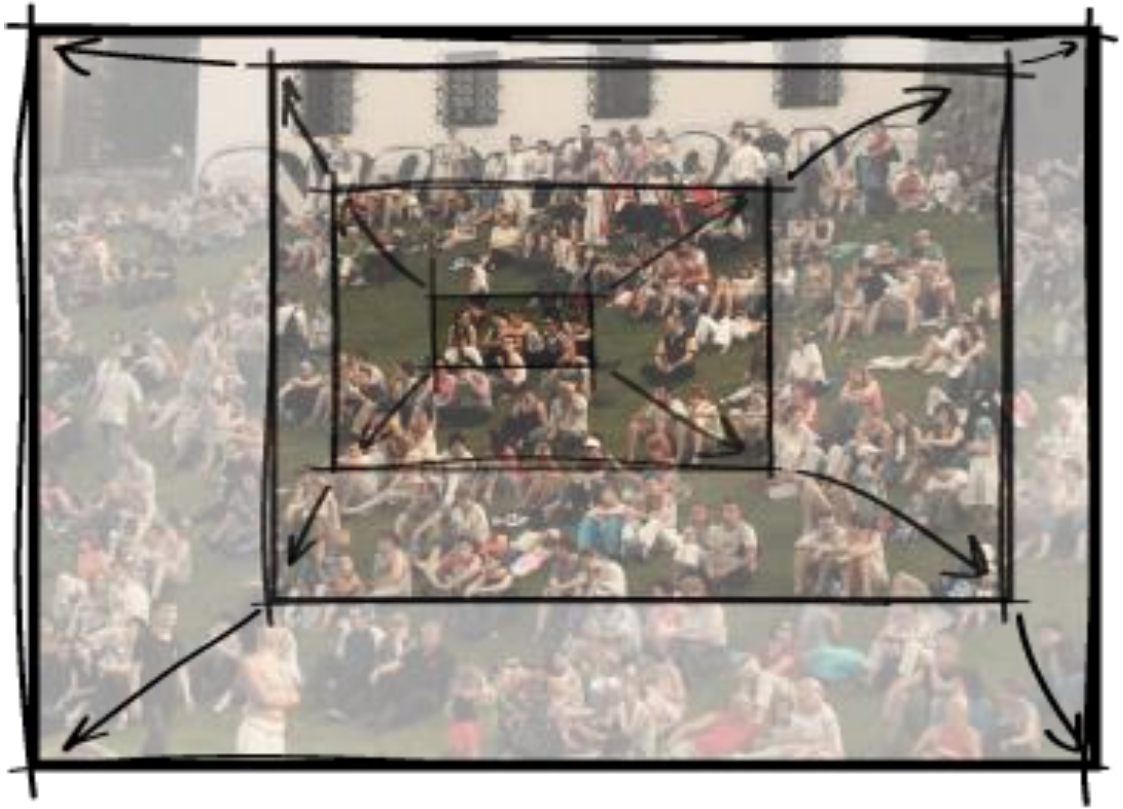
Values recorded for less than 5 years

Fill as many NaNs possible

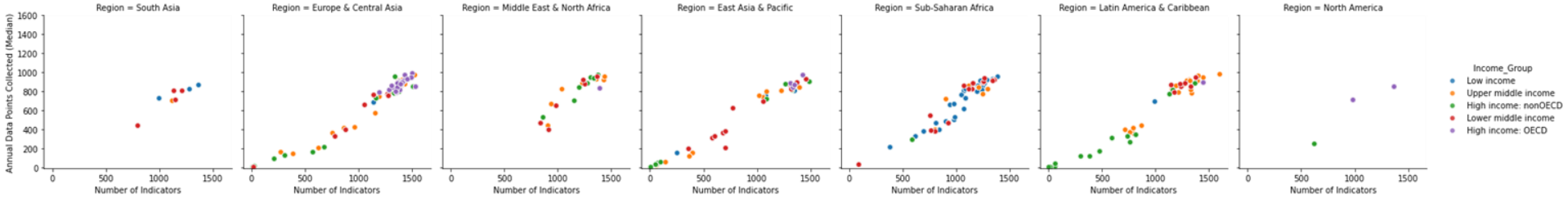
Interpolation

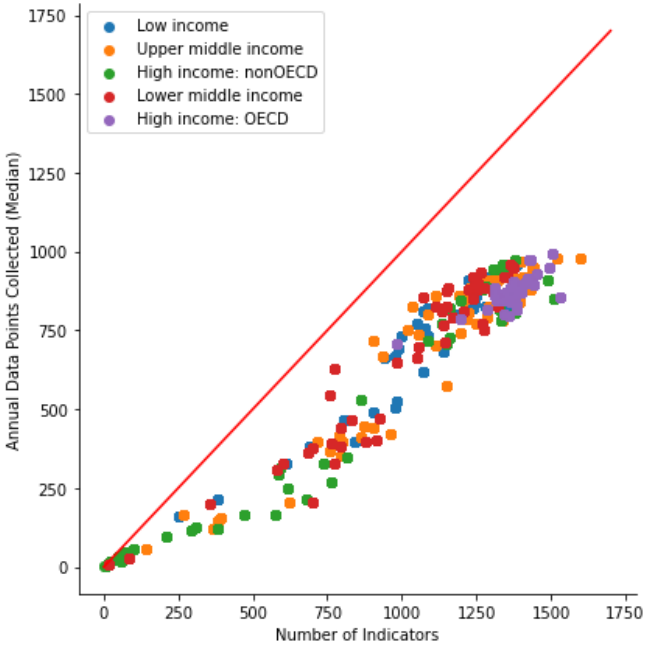

Box Plot for Outlier Visualization
Outliers
Indicator Unit
80
70
60
50
40
30
Adult literacy rate, population 15+ years, both sexes (%)
Youth literacy rate, population 15-24 years, both sexes (%)
Elderly literacy rate, population 65+ years, both sexes (%)
Indicator

Exploration Methodology



Extent of Data Collection

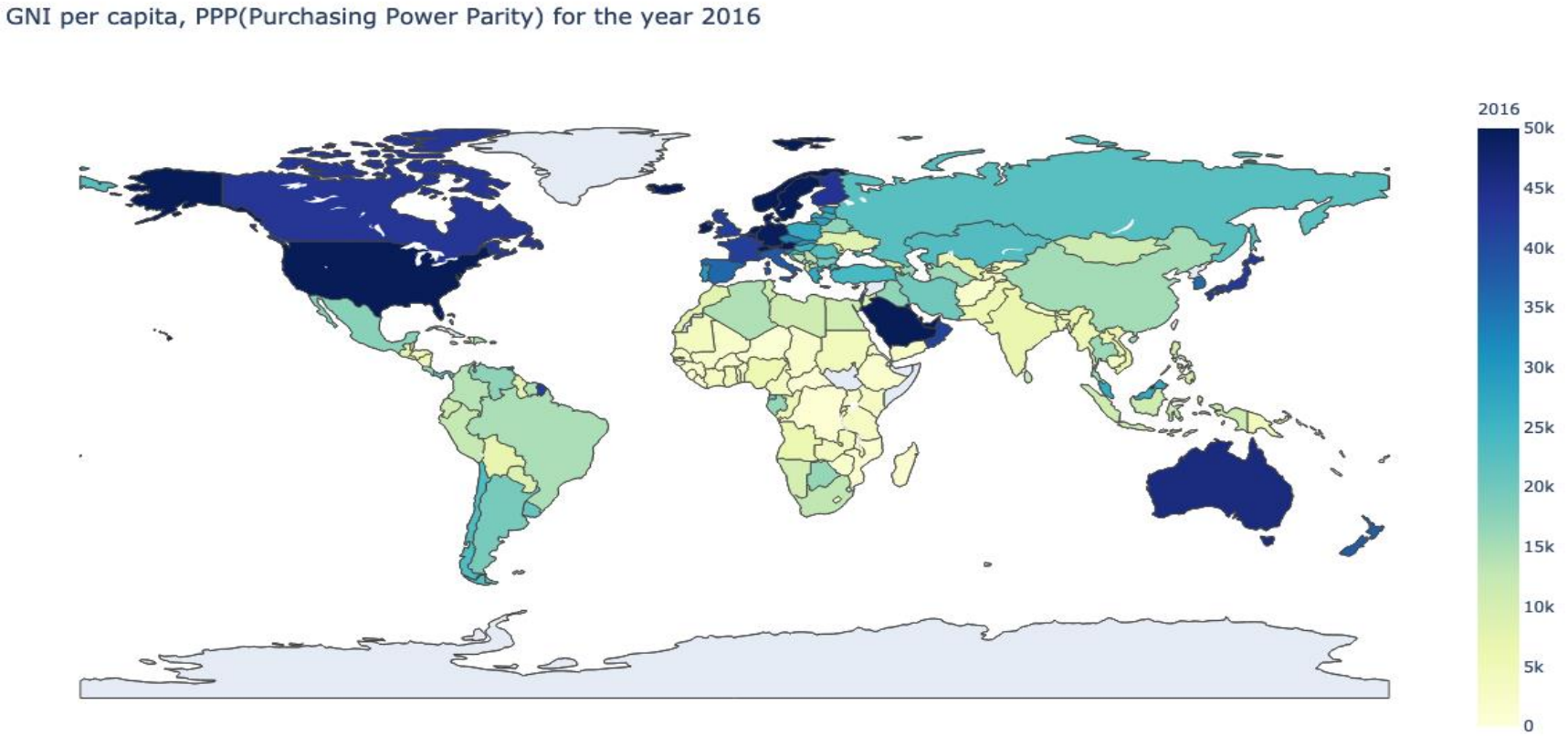




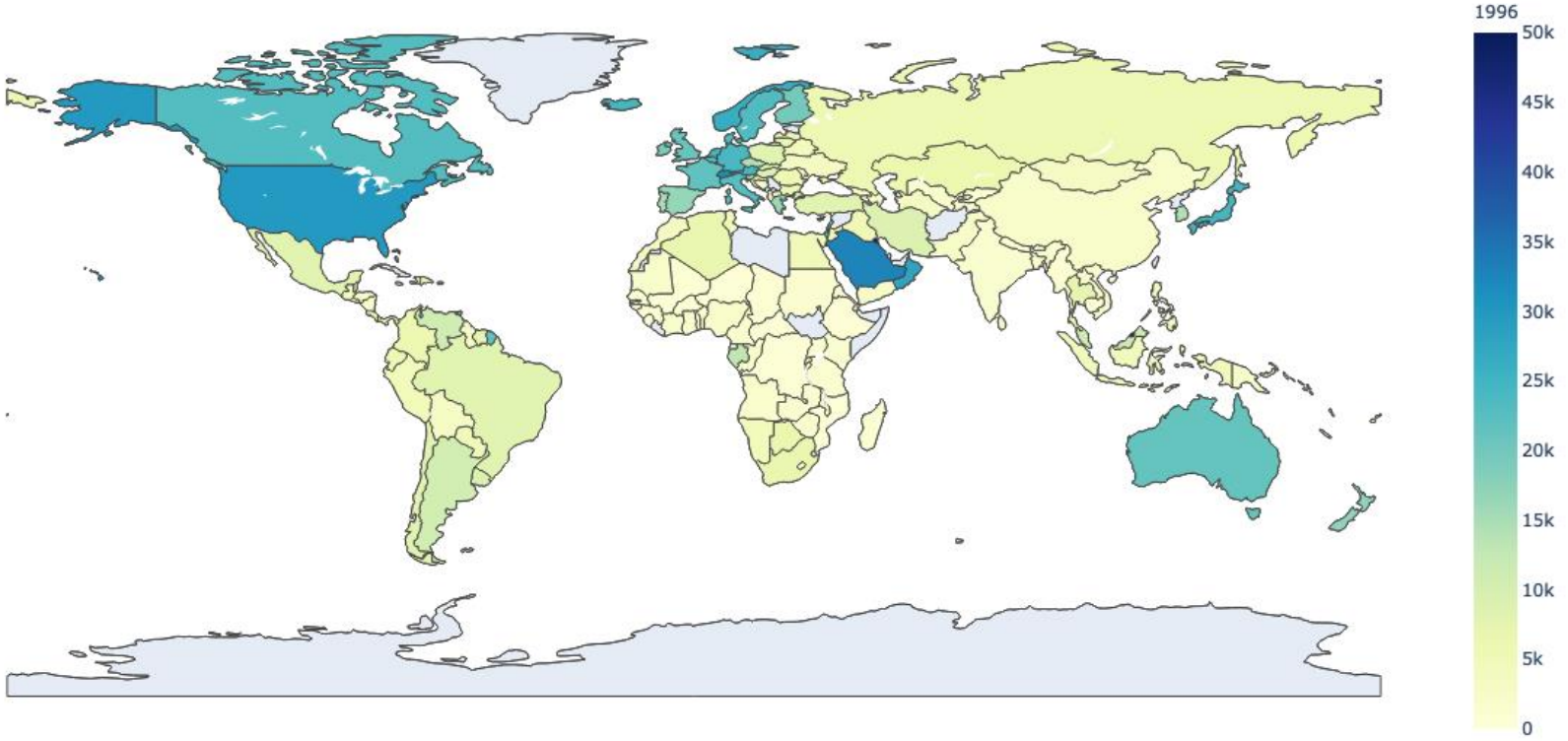
Comparison based on Indicators

- Various Indicators taken for comparison based on Regions and Countries are listed below-
  - GNI per capita, PPP
  - Unemployment Rate
  - Gross Enrollment Ratio
  - PISA
  - PIAAC
  - LITERACY RATE
  - BARRO-LEE

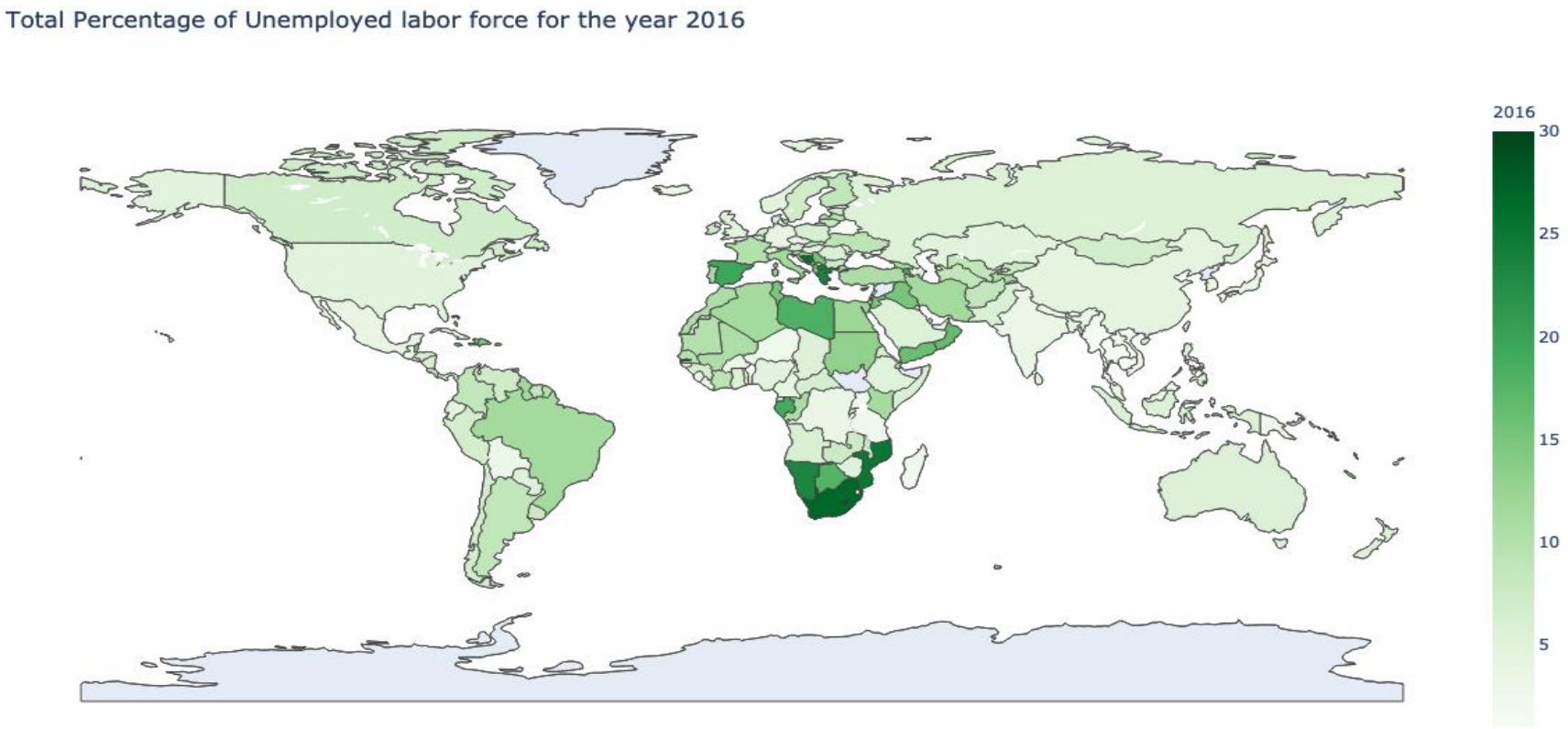
GNI per capita, PPP



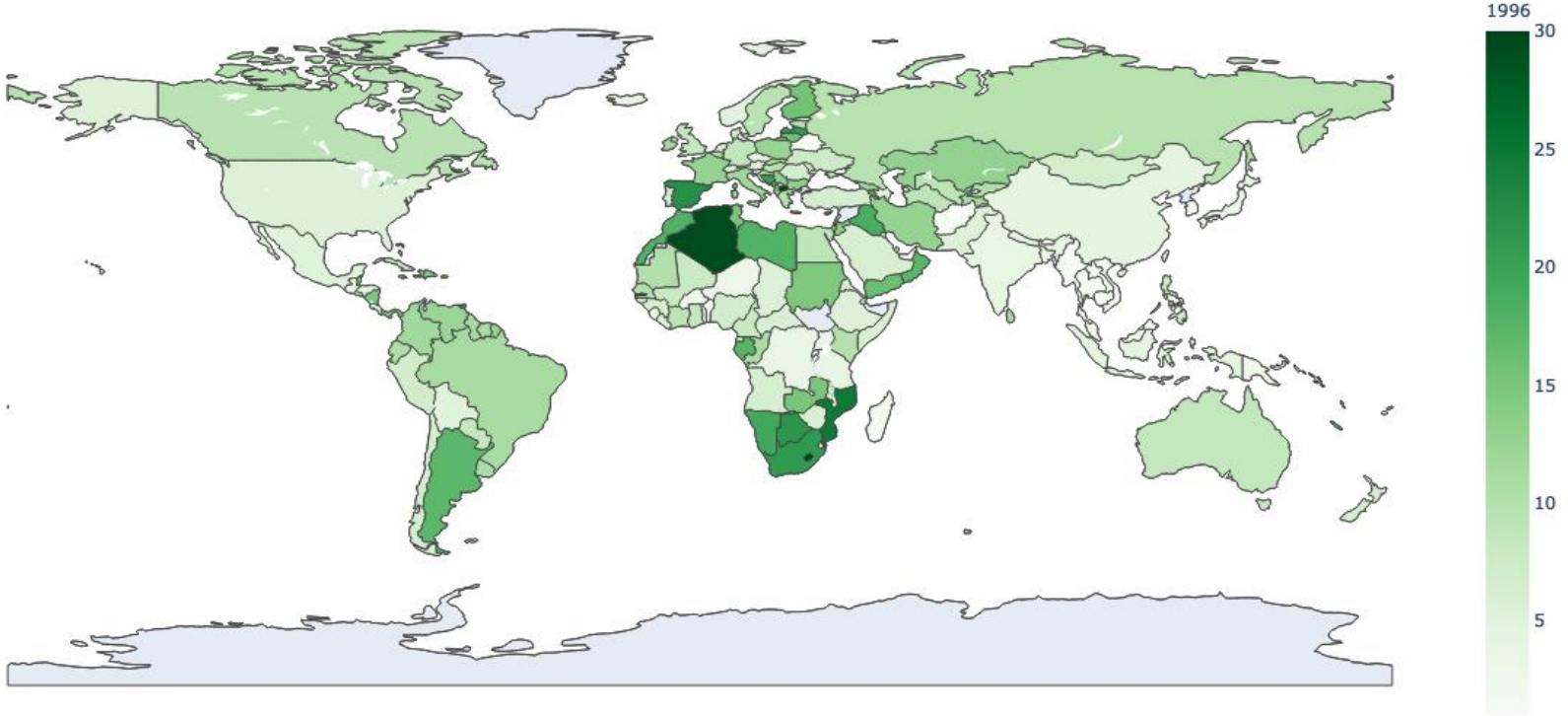
GNI per capita, PPP(Purchasing Power Parity) for the year 1996



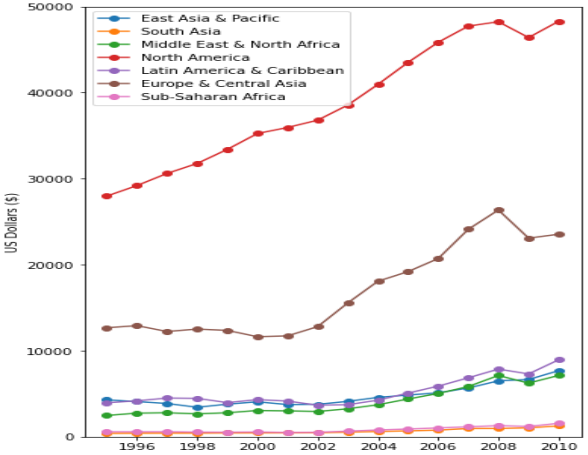
Percentage of Total Unemployed Labor force

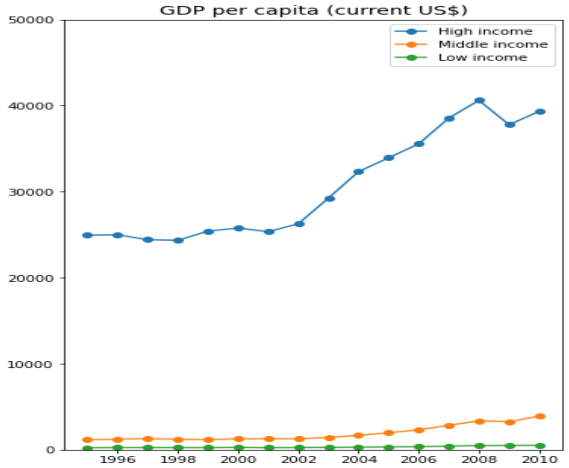


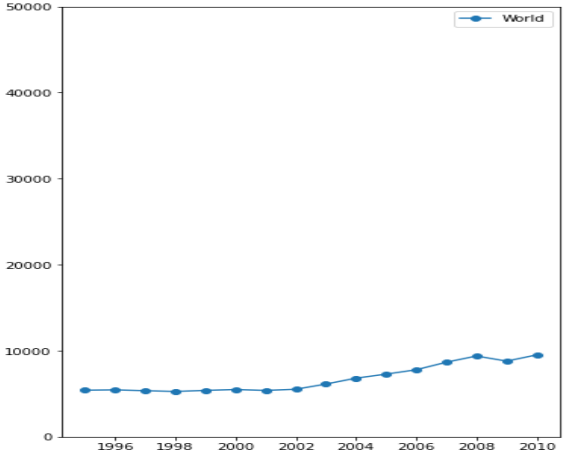
Total Percentage of Unemployed labor force for the year 1996

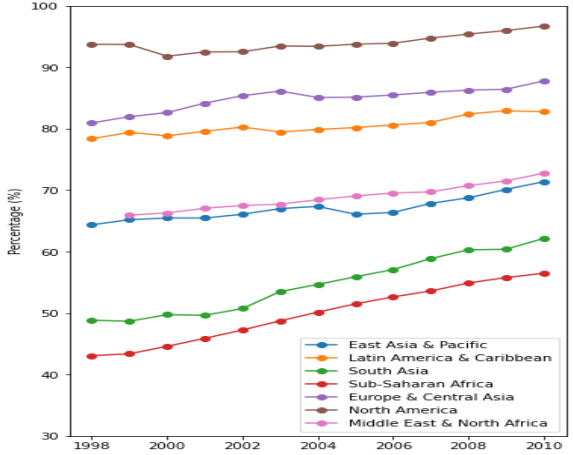


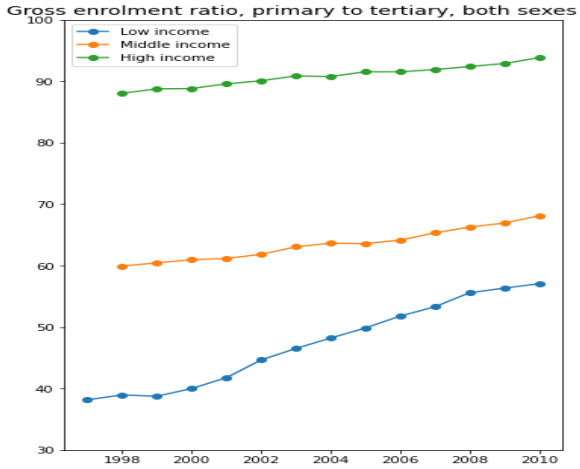
GDP vs Enrollment (1995-2010)

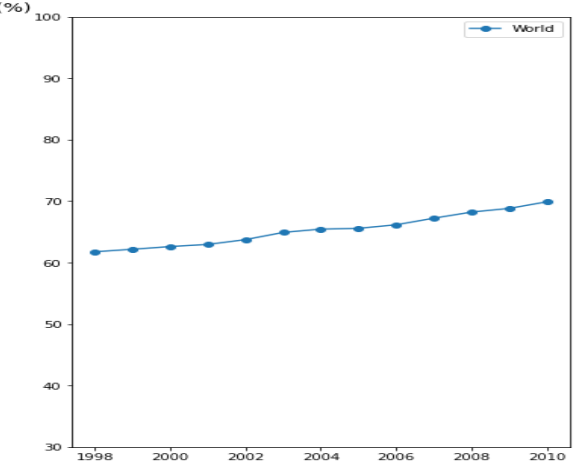




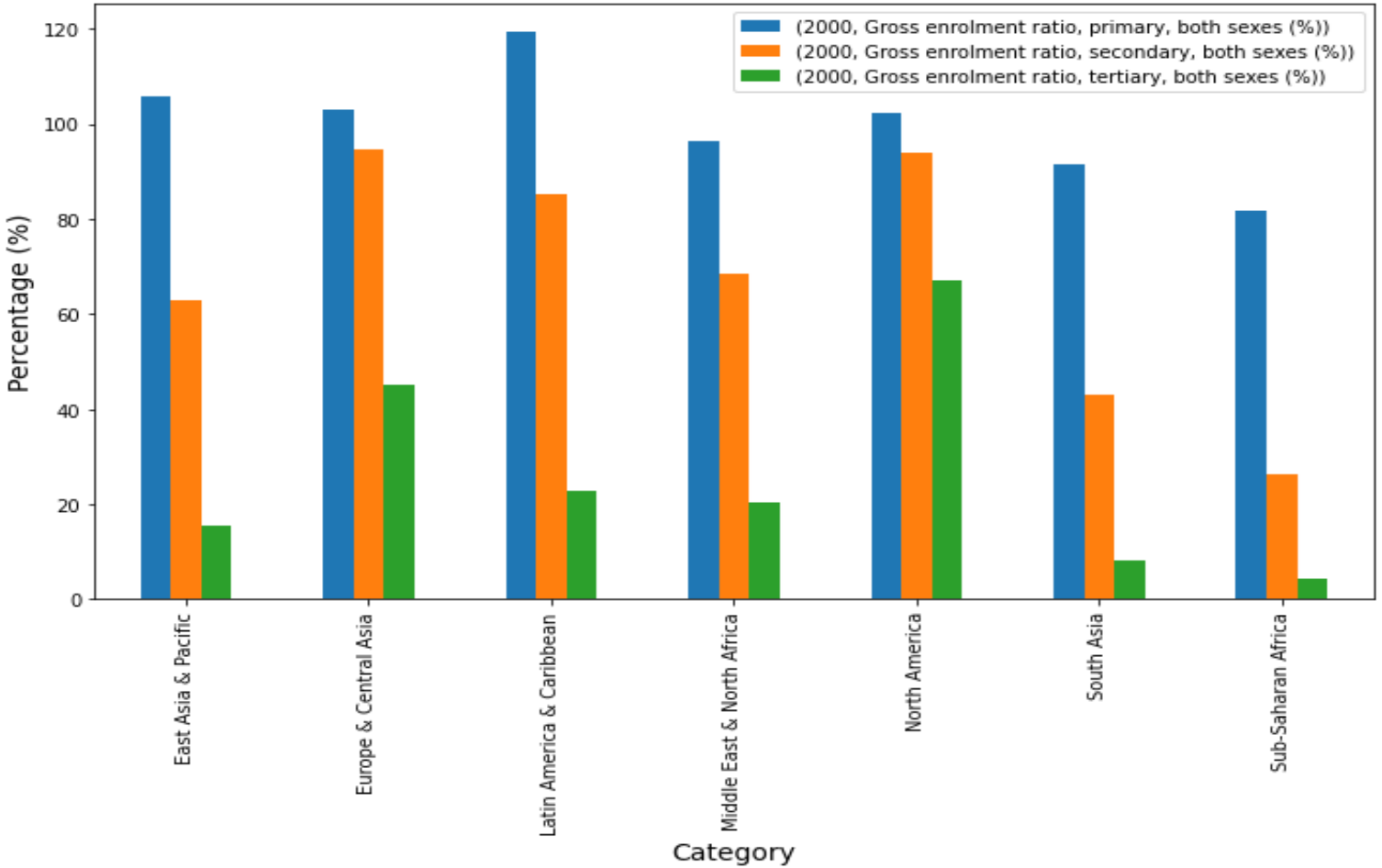




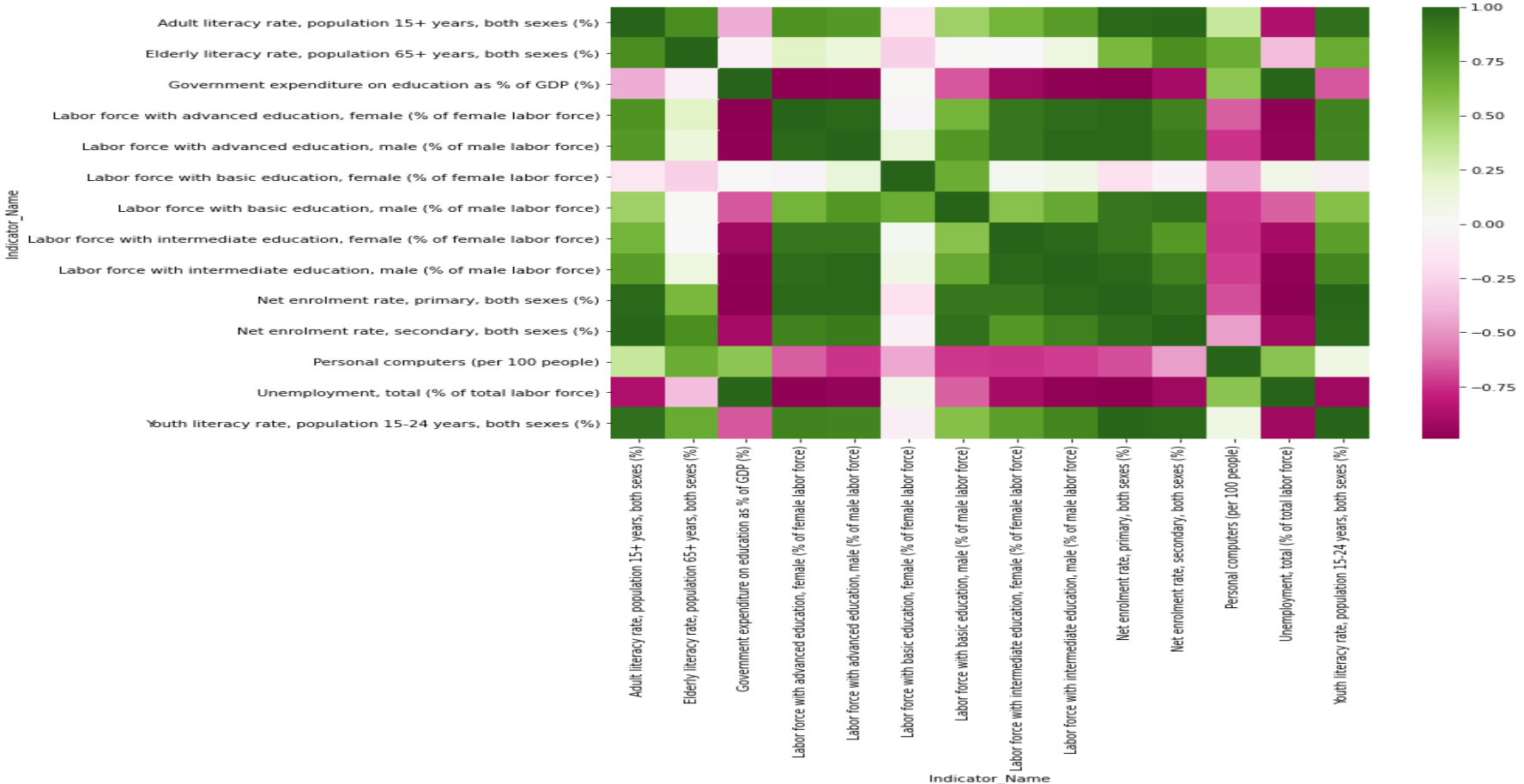


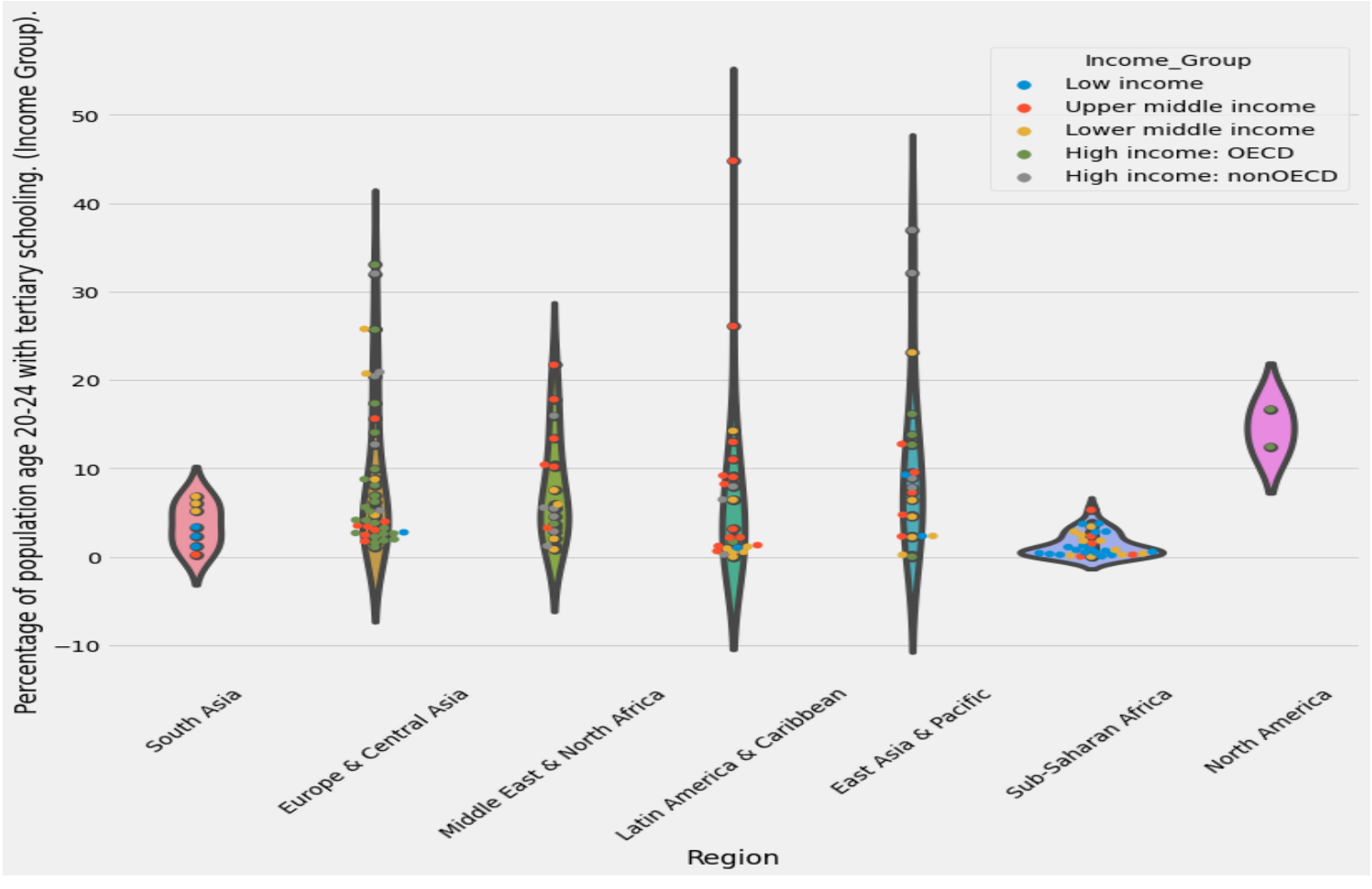


Gross Enrollment Ratio(Region based)

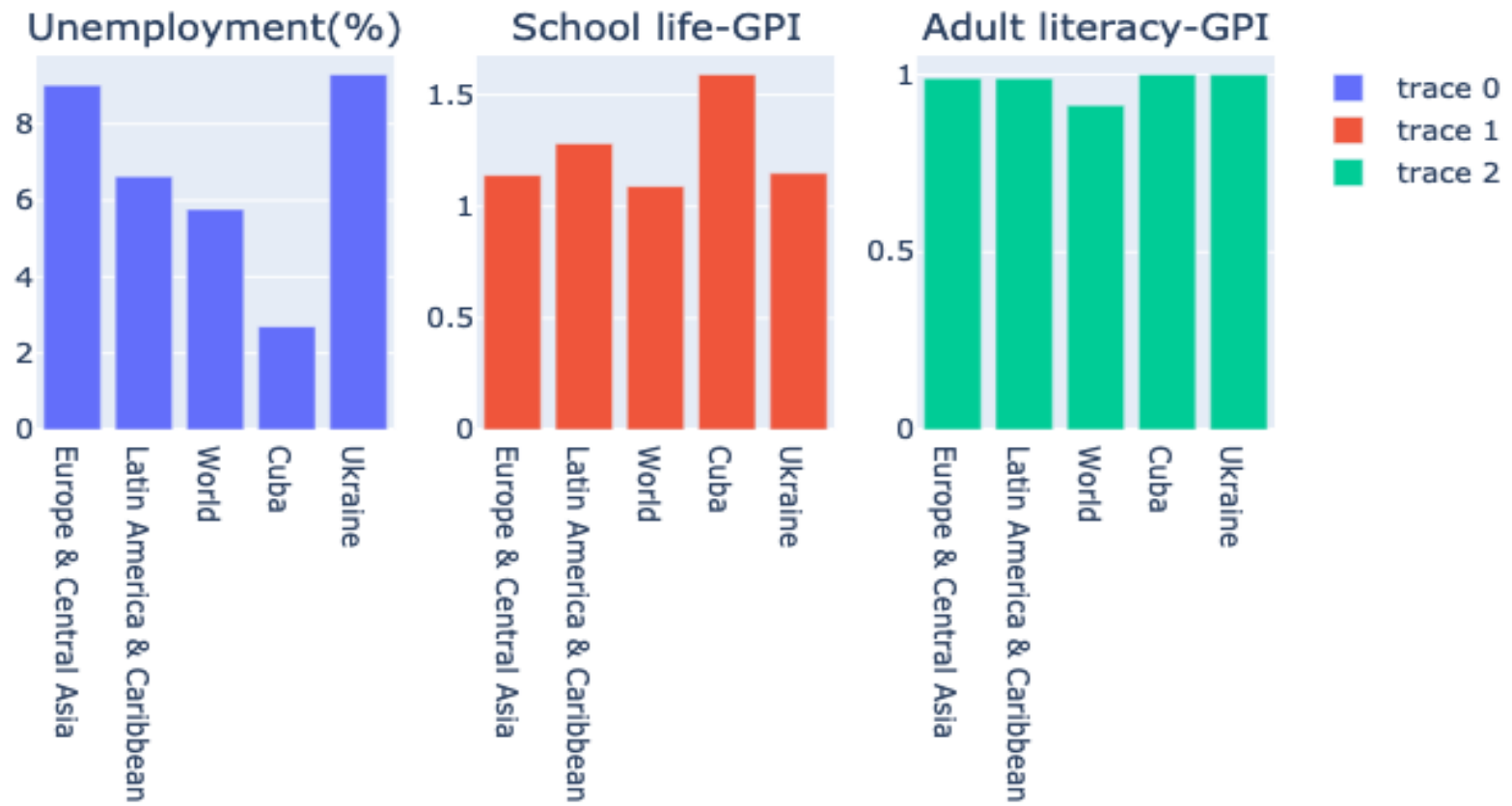


Correlation Between Indicators

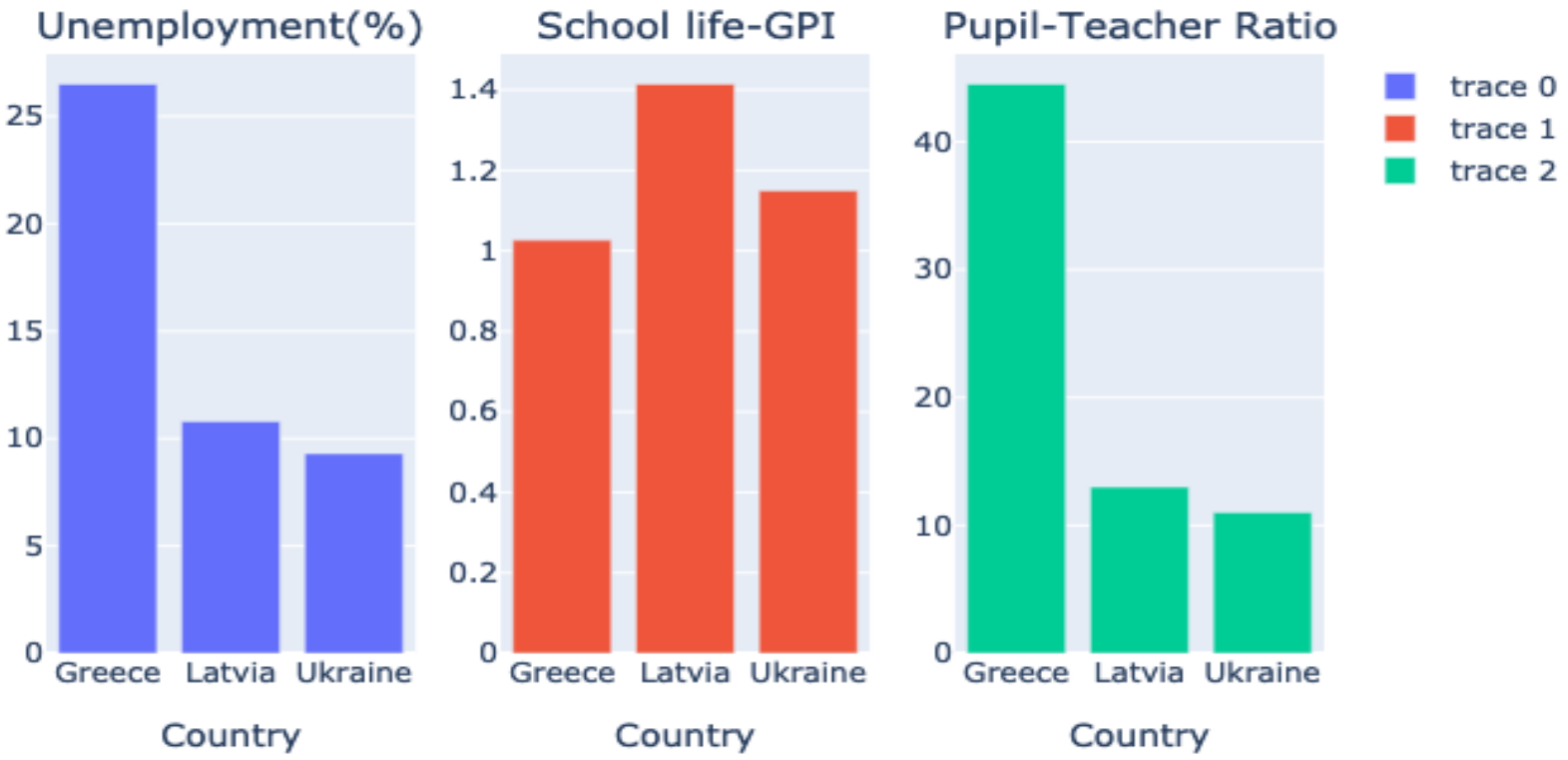



Comparison of countries based on Region
Income_Group
Low income
Upper middle income
Lower middle income
High income: OECD
High income: nonOECD
Percentage of population age 20-24 with tertiary schooling. (Income Group).
50
40
30
20
10
0
−10
South Asia
Europe & Central Asia
Middle East & North Africa
Latin America & Caribbean
East Asia & Pacific
Sub-Saharan Africa
North America
Region

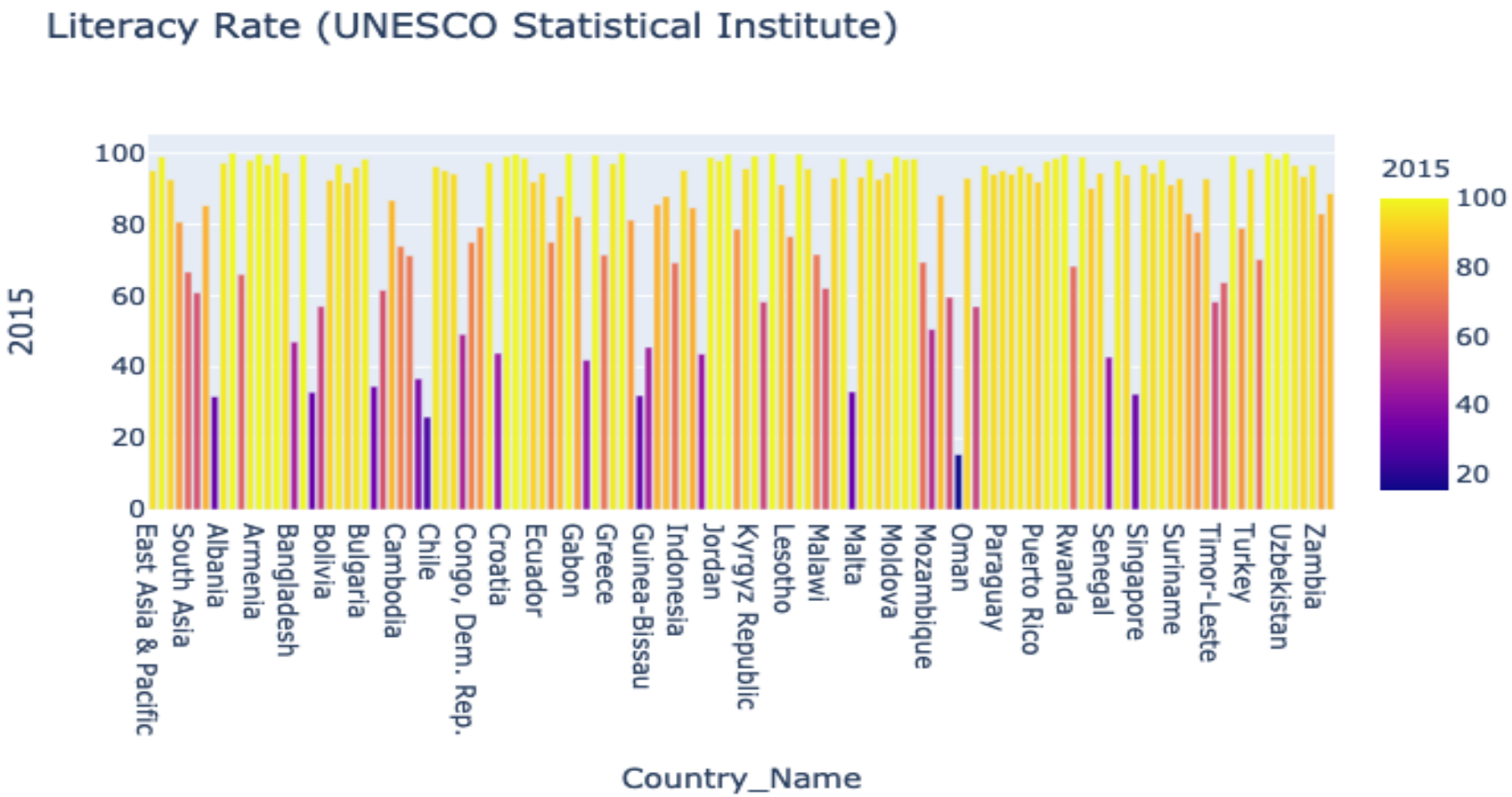
Percentage of population with Tertiary Schooling(20-24).



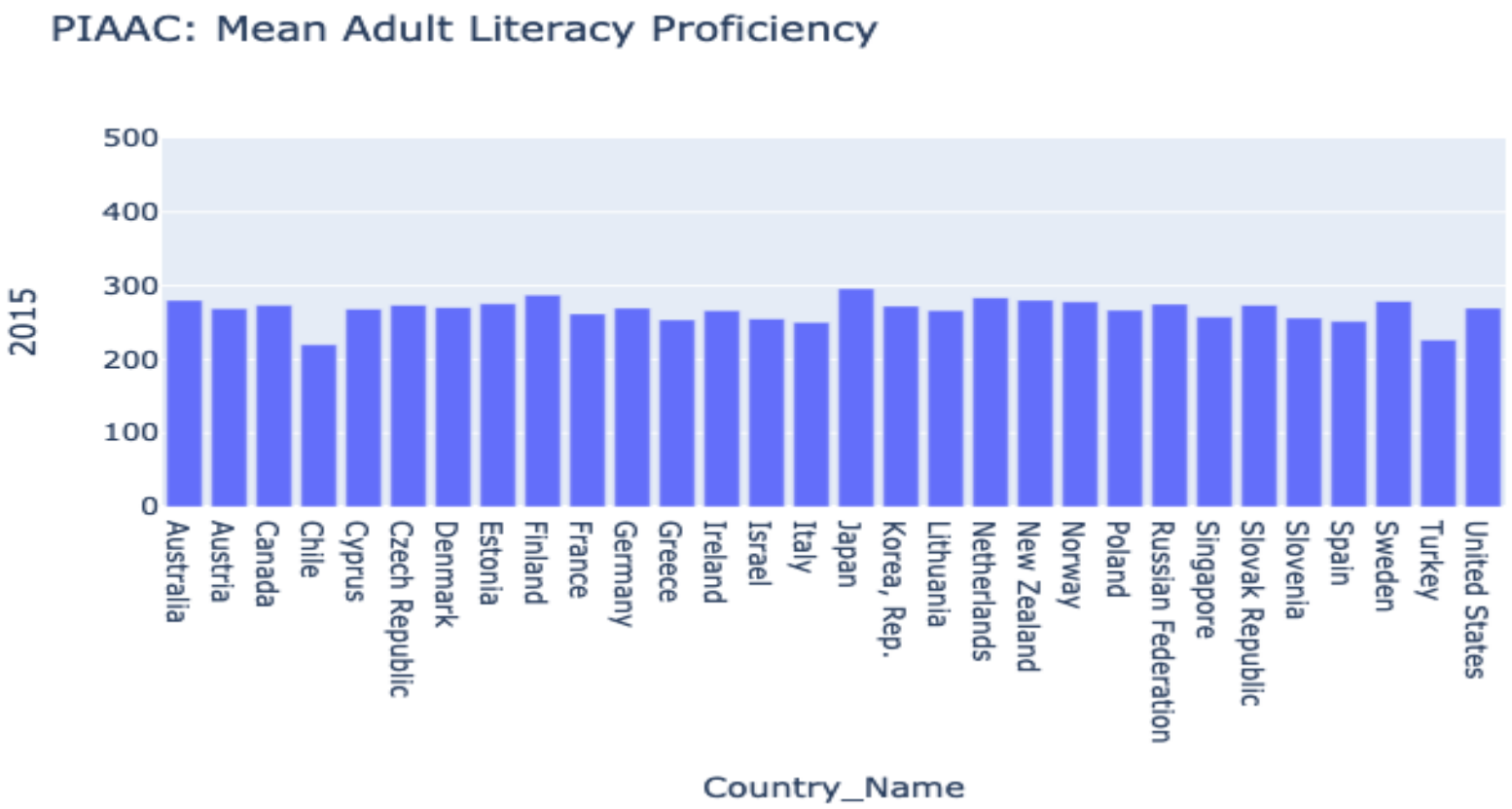
Percentage of population with Tertiary Schooling(20-24).



UNESCO - Literacy Rate



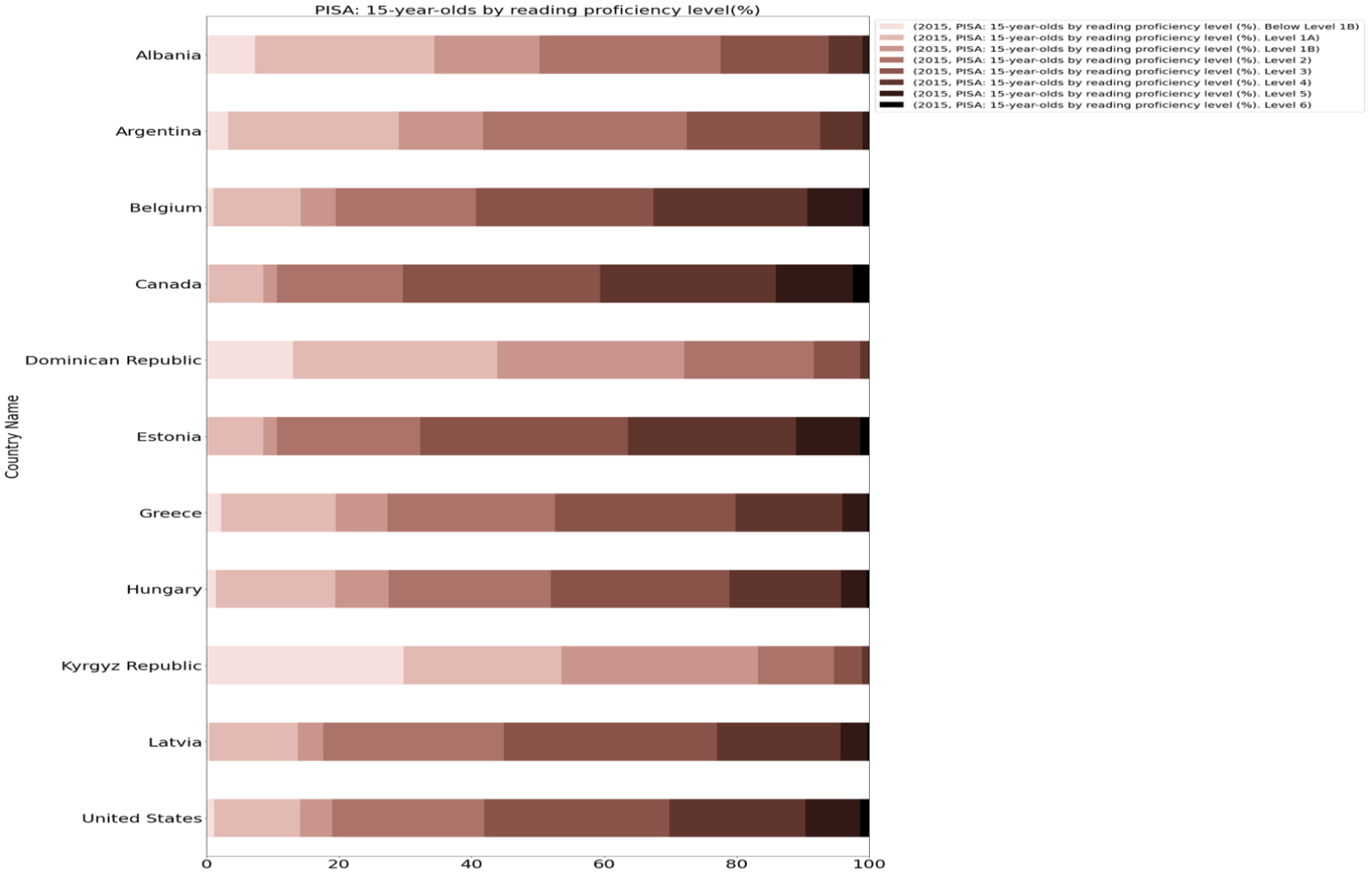
PIAAC - Literacy Scores



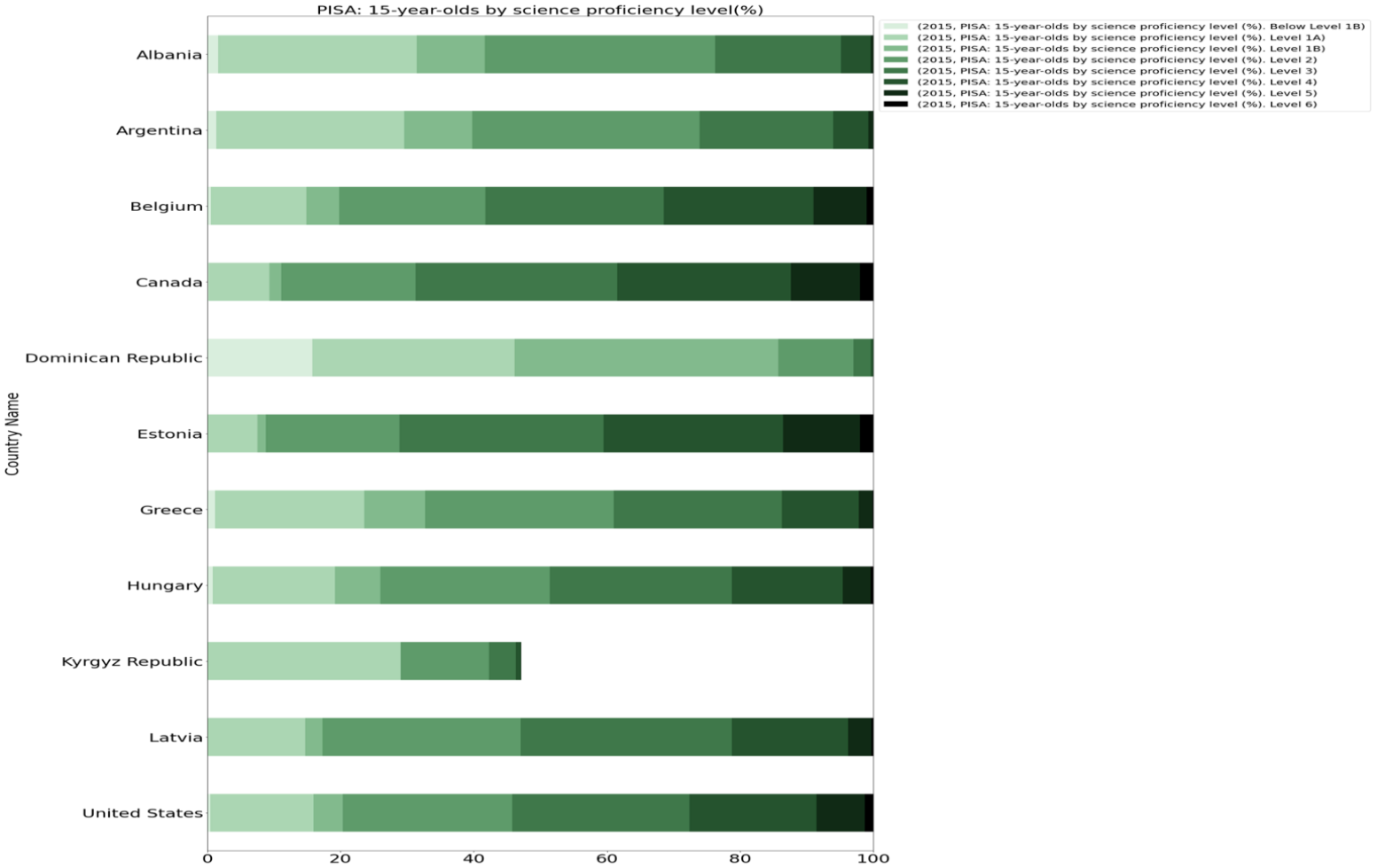
PISA

- It has three indicators namely proficiency for Reading, Science and Mathematics.
- The below slides contain comparisons based on these three indicators in the following order:
  - PISA Reading Proficiency
  - PISA Science Proficiency
  - PISA Mathematics Proficiency

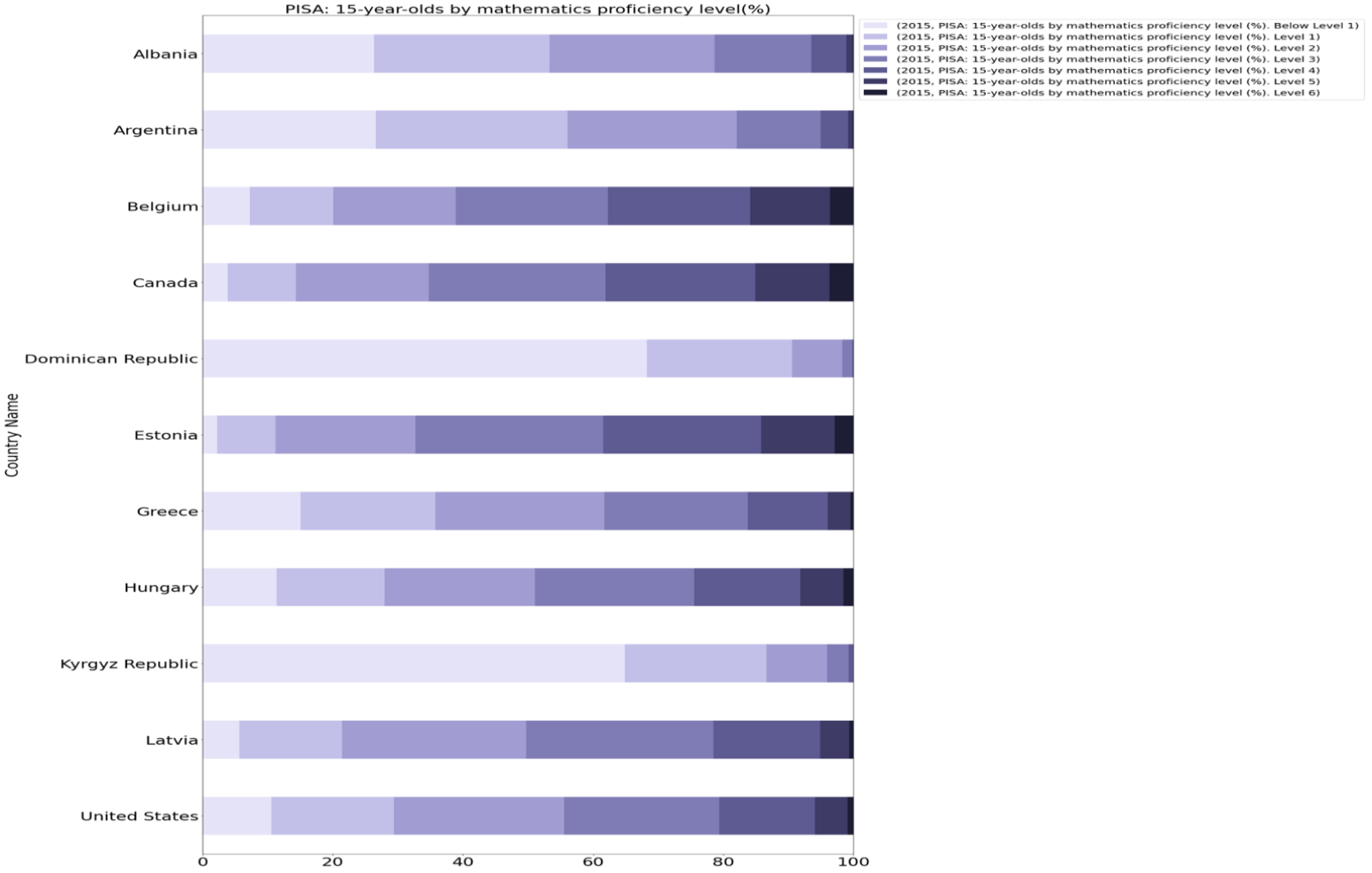
PISA - Reading Proficiency



PISA - Science Proficiency



PISA - Mathematics Proficiency



Challenges

- Inconsistency in accepted indicators.
- NaNs and missing values.
- Possible important indicators missed because of methods used to clean data.
- Thousands of indicators with particularly unique measurement methodology.
- Manual keyword search was performed to find critical indicators because of which many indicators are yet to be analyzed.
- Fair comparison between countries was challenging.
- Plotting choropleth maps using Geopandas was tricky because of mismatch between countries of Geopandas dataframe and pandas dataframe.

Conclusion

- Few countries, which are now in the high-income category, have had significantly large economic growth in the last couple of decades as compared to the middle and low income countries.
- There is large inequality between countries in different income groups in term of economic strength (GDP per capita) and enrolment in educational institutions.
- As the level of educational difficulty and expense increases the enrolment naturally decreases across the world.
- Strong correlations were found between:
  - Government expenditure on education, unemployment rate, and number of personal computers per 100 people
  - A slightly negative relationship was found between government expenditure and youth literacy rate which goes against expectation

- From the Barro-Lee indicator, we can see that a number of lower middle income and upper middle income countries outperform the high income OECD and non-OECD countries based on percentage of tertiary schooling especially in Europe, Middle East and Latin America.

- Ukraine shows high rate of unemployment which is quite contrast having high percent of tertiary education.

- Ukraine performs significantly well when compared to high income countries like Greece and Latvia, having the lowest pupil-teacher ratio shows the reason why Ukraine has a higher tertiary education percentage.

- High income and developed economies shows higher rate of proficiency for reading, science and mathematics for 15 year olds, most importantly across all levels as opposed to lesser developed economies.

- Significant increase of GNI, PPP all over the world barring some countries from Africa and Asia over the span of 20 years.

- Significant decrease of Unemployment rate in Algeria over the span of 20 years.
- Marginal decrease of Unemployment percentage all over the world from 1996-2016.