E-Learning Statistics Analysis

Azadeh MOJARAD POHIER

OpenClassroom

Mentor: Rim ROMDHANE

June 2021

Contents

Introduction:

This analysis is done for EdTech start-up, which offers online training content for high school and university level audiences.

The data is taken from The World Bank EdStats website:

https://datacatalog.worldbank.org/dataset/education-statistics

The World Bank EdStats Query holds around 2,500 internationally comparable education indicators for access, progression, completion, literacy, teachers, population, and expenditures. The indicators cover the education cycle from pre-primary to tertiary education. The query also holds learning outcome data from international learning assessments (PISA, TIMSS, etc.), equity data from household surveys, and projection data to 2050.

Objectives:

1. Which countries have a strong potential of customers for our services?
2. For each of these countries, how will this customer potential evolve?
3. In which countries should the company operate as a priority?

Data discovery:

To check the data, all the five Csv files (EdStatsData, EdStatsCountry , EdStatsFootNote, EdStatsCountry-Series, EdStatsSeries) are downloaded. For more information I checked PIB.csv (API_NY.GDP.PCAP.CD_DS2_en_csv_v2_2445354). To read all the files, we need, python libraries, Pandas. After checking all the files, we select “EdStatsData & EdStatsCountry” for this project.

Important sets of data for this project:

- Internet
- Education
- Population
- Economy

Now we can start first part of our data cleaning.

Data Cleaning:

For data cleaning part, since the dataset is huge, it is necessary to filter the dataset for several times.

Using groupby() function to see the number of indicators in Data Frame. The data includes 3665 indicators, name of countries and years from 1970 to 2100.

This project has 2 parts:

1) Analyze data for previous years. (1970-2020)
2) Analyze data for future. (2025-2100)

Part1:

After checking all indicators on the website, we select the most important one by following:

Internet

→ Internet users (per 100 people)(IT.NET.USER.P2)

Population

→ Population, ages 15-24, total(SP.POP.1524.TO.UN)
→ School age population, tertiary education, both sexes (number) (SP.TER.TOTL.IN)
→ School age population, secondary education, both sexes (number)(SP.SEC.TOTL.IN)

Education

→ Enrolment in tertiary education, all programs, both sexes (number) (SE.TER.ENRL)
→ Enrolment in secondary education, both sexes (number)(SE.SEC.ENRL)

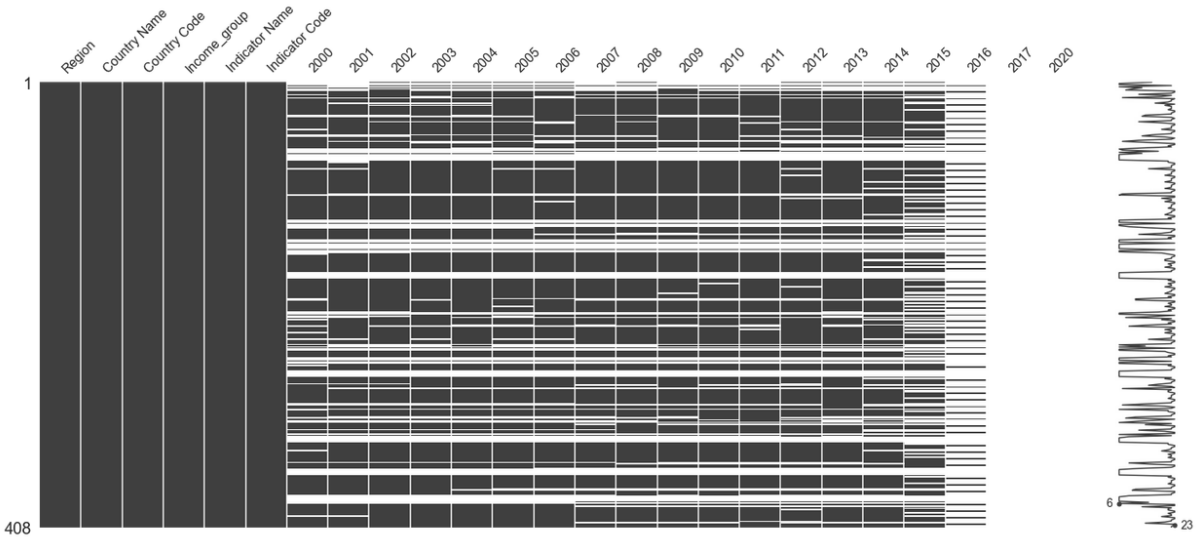
Expenditures on education

→ Expenditure on education as % of total government expenditure (%) (SE.XPD.TOTL.GB.ZS)
→ GDP per capita based on purchasing power parity (PPP) , (current international dollar) (NY.GDP.PCAP.PP.CD)

Now creating new DataFrame with the name of countries, selected indicators, the countries with high income and year from 2000 to 2020.

Data Quality:

After creating new Data frame, need to check ***null values*** by using msno faunction.



We must to Drop years, '2016', '2017' & '2020' because of missing values more than 50%.

For other years using fillna() function to fill all the null values with their median because the distribution is skewed.

Pivot table:

Using pivot table to summarize the data between features "Country Name" and "Indicator Code" in 2015.

By checking pivot table description, it shows us that there exists seven countries with incorrect data (huge difference between mean and median in 'IT.NET.USER.P2'). We can remove them since they are so small countries.

Since our company is looking for young student so the next part is to find the countries with high population in age 15-24, more than 1,000,000.

To merge:

By using Merge() function we can add two data frame 'pivot_table' & 'data_keep', to have a Data Frame with all years & selected indicators.

Duplicated:

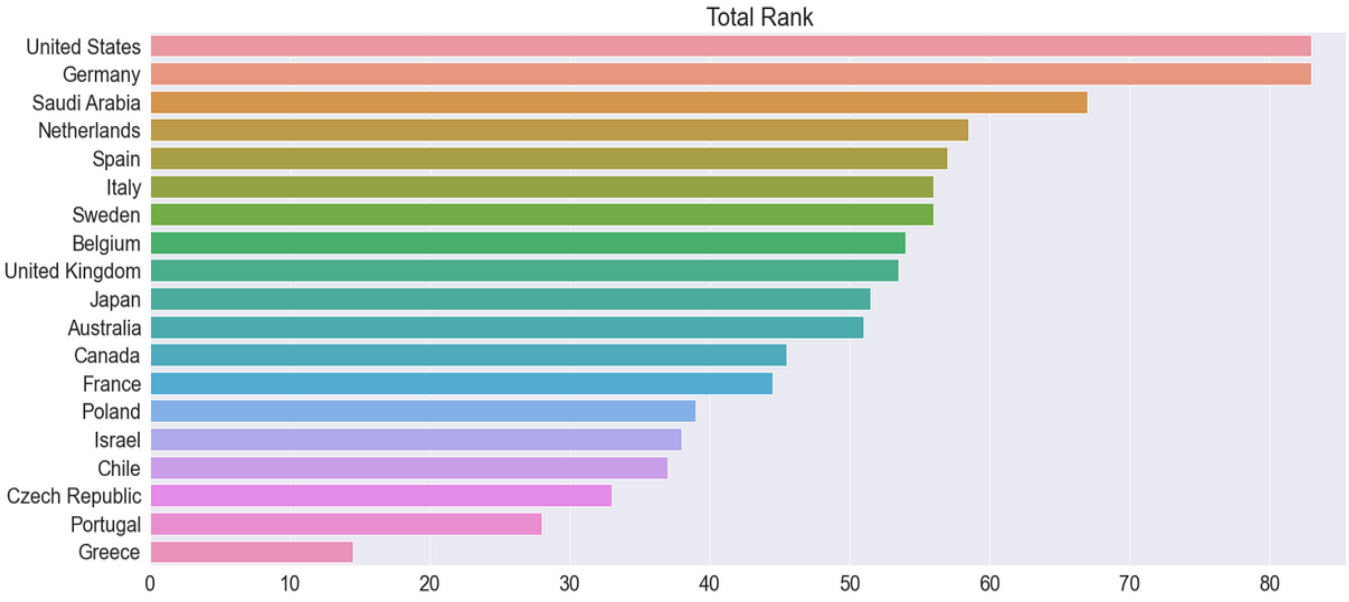
After merging, we need to check duplicated data between 'Country Name' & 'IT.NET.USER.P2' by using duplicated() function. There are 133 duplicated observation, so we need the drop_duplicates() function to drop all of them.

Finally, the dataset is clean now so we can start our analysis.

Data Analysis

After cleaning our dataset it remain 19 countries for our analyzing.

The next step is ranking the rest of countries by highest population age 15-24, internet users, enrolment in tertiary school and GDP per capita by using rank() function.

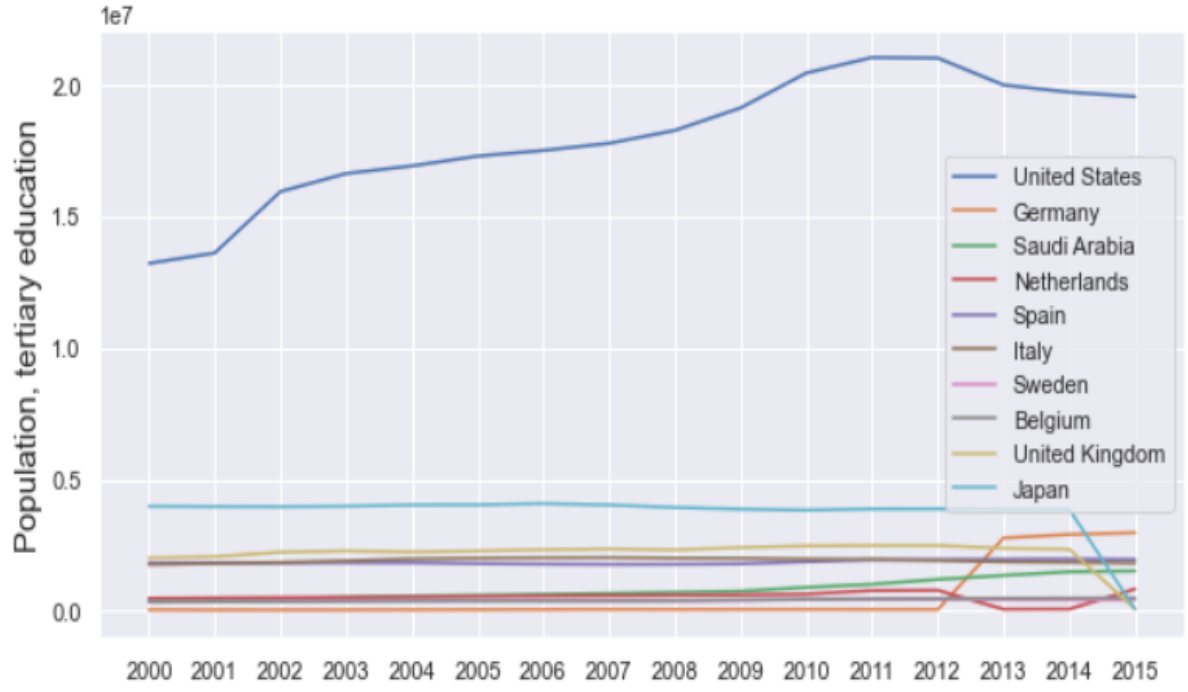


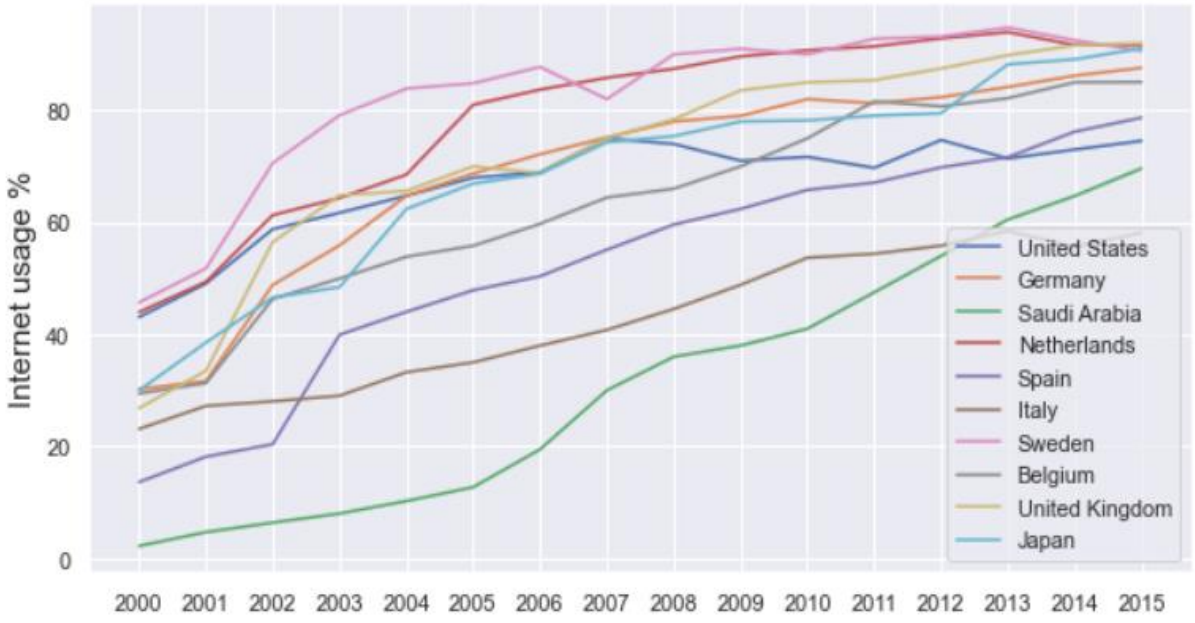
Here we have all the ranked countries, to make a comparison, we keep just top 10 of them:

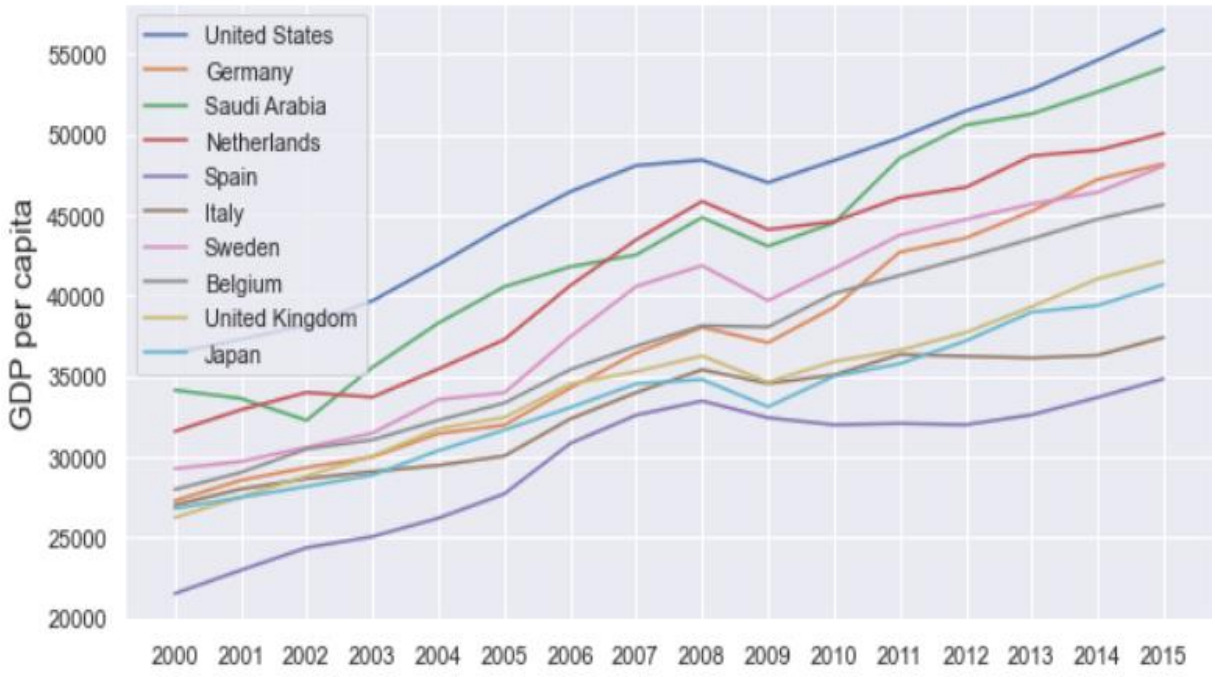
'United States', 'Germany', 'Saudi Arabia', 'Netherlands', 'Spain', 'Italy', 'Sweden', 'Belgium', 'United Kingdom', 'Japan'.

Visualization:

For visualization, by importing Seaborn we can use sns.lineplot() function to show the history of our dataset from 2000 to 2020 for selected countries.





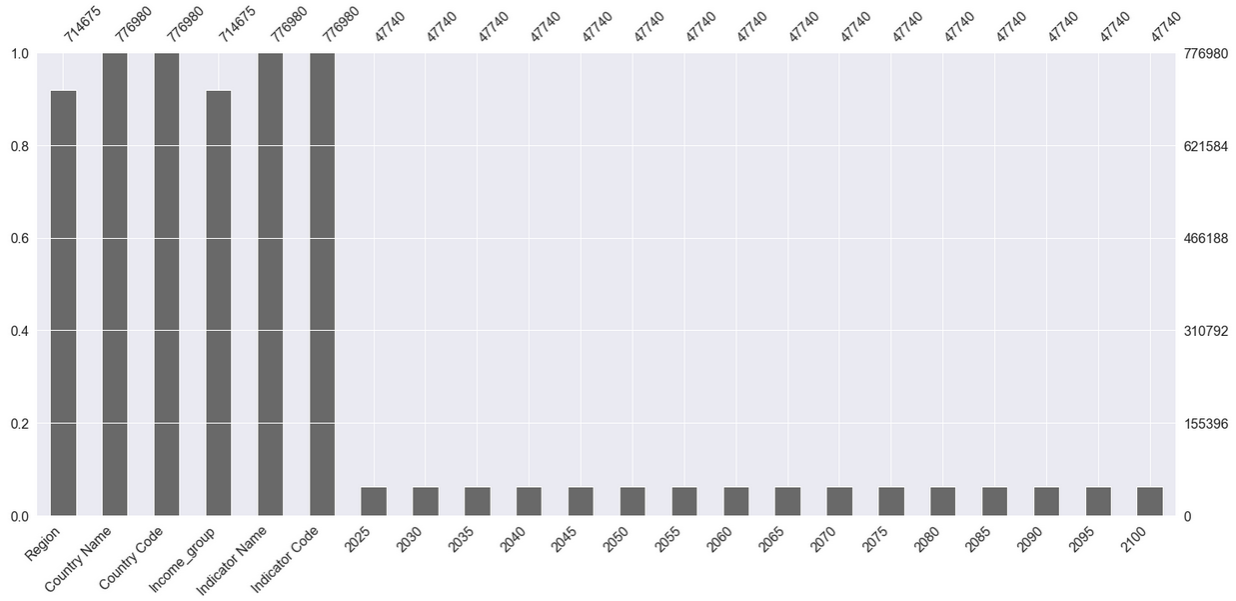


By comparing these 4 graphs, it is obvious that USA is the best countries for this project, with highest young population.

Part 2:

Projection

For projection we are working on the years from 2025 to 2100.

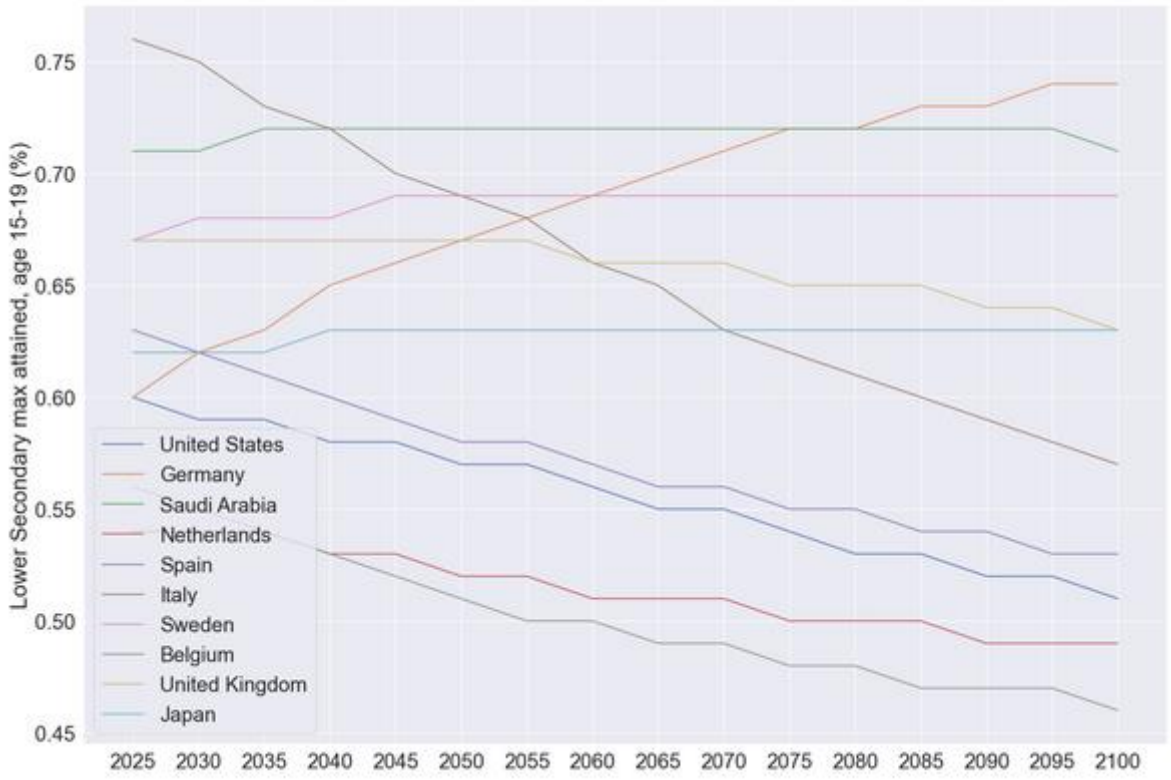


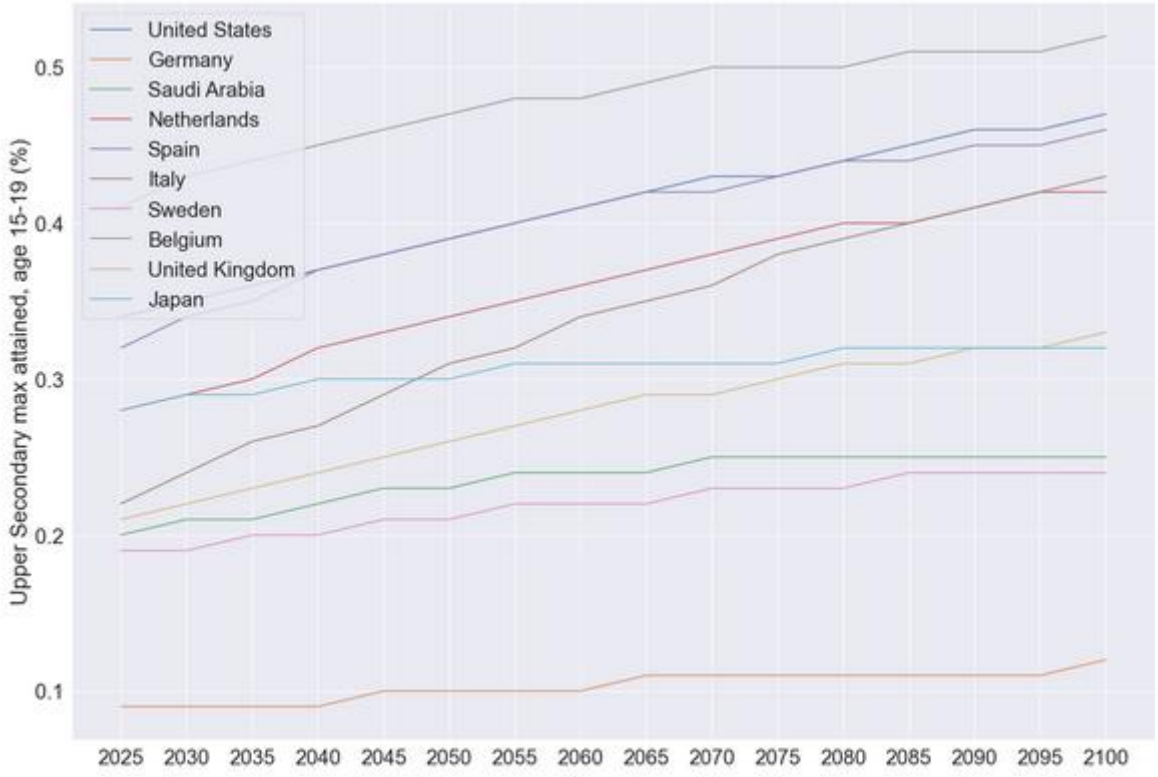
Since there are many missing values in our dataset, we just make a sample from our data. Therefore, by using notnull() function we keep just all non-null variables. Then try to find interesting indicators for this part from the website, following by:

- Wittgenstein Projection: Percentage of the population age 15-19 by highest level of educational attainment. Upper Secondary. Total(PRJ.ATT.1519.3.MF)

- Wittgenstein Projection: Percentage of the population age 15-19 by highest level of educational attainment. Lower Secondary. Total(PRJ.ATT.1519.2.MF)

Visualization:





For this part the Upper Secondary population is going up for all the selected countries but Lower Secondary population is going down for some of these countries gradually except of Germany.