Visual Data Science | Topic Selection & Discover

Kiesel Markus | 1228952

November 10, 2021

1 Topic Selection

I decided to use this project to visually explore the education domain and in part answer the follwing question. **How has the level of education changed in different countries?**

2 Discover

First, a search for sutable datasets was performed. The topic of education is of interest to many. Thus, a large number of sutable datasets was discovered from various sources.

Many datasets consider a handful of education indicators for a varying number of countries. The most extensive and global dataset discovered is the one provided by the World Bank [1]. Another comprehensive data source discovered is from the OECD Education Statistics [2].

2.1 The World Bank Education Dataset

The dataset [3] contains 43 092 rows and 65 columns. Data on 162 indicators for 266 countries are provided over the years 1960 to 2020.

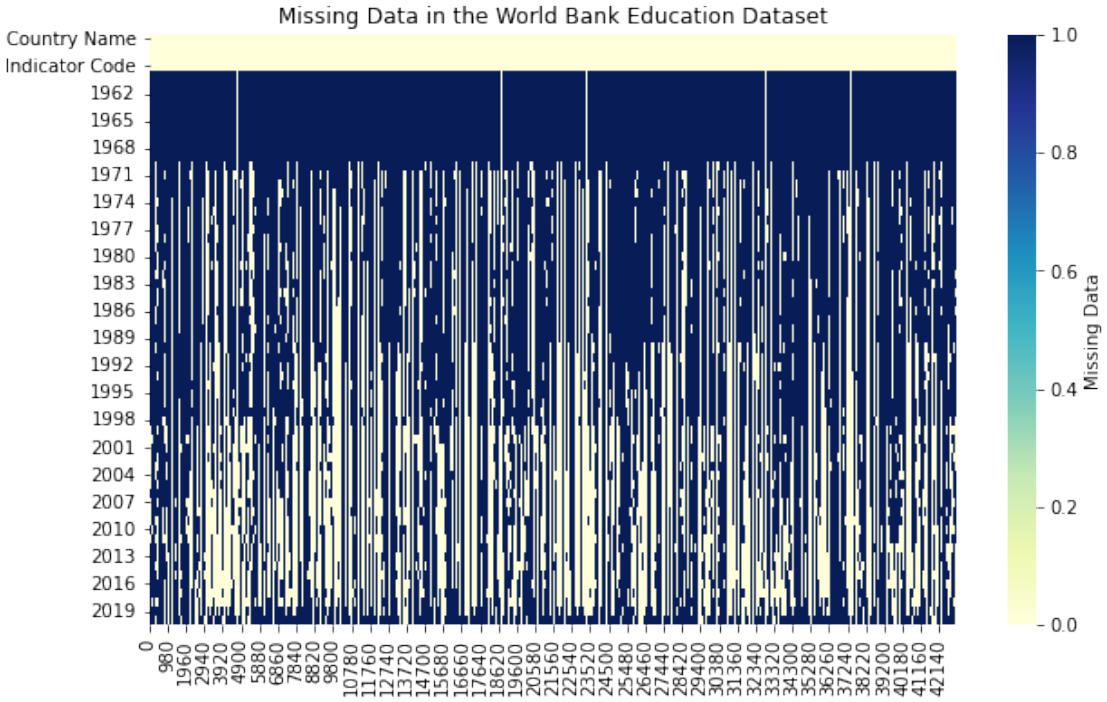
Most of the indicators are provided separately for female and male which allows a distinction by gender. Further, the matadata includes an income group for each country.

To show how detailed the dataset is we show some of the features:

- Population ages 15-64 (% of total population)
- Current education expenditure, primary (% of total expenditure in primary public institutions)
- Government expenditure on education, total (% of GDP)
- Educational attainment, at least Bachelor's or equivalent, population 25+, female (%) (cumulative)
- Trained teachers in lower secondary education (% of total teachers) Educational attainment, at least completed short-cycle tertiary, population 25+, male (%) (cumulative)
- Literacy rate, youth (ages 15-24), gender parity index (GPI)

[1] https://data.worldbank.org/topic/education
[2] https://www.oecd-ilibrary.org/education/data/oecd-education-statistics_edu-data-en
[3] https://api.worldbank.org/v2/en/topic/4?downloadformat=csv

One of the main challenges in this project will be the amount of missing data which we can observe in the following heatmap where the dark spots show the missing data.



2.2 OECD Education Statistics

This data source [4] is not directly one dataset but a database containing various datasets with different education indicators. The datasets usually contain all OECD counties, several indicators, distinguishe between gender and span several years. So this datasats also have a high dimensionality.

3 Whats Next?

In the next step we will explore the data in more detail and based on this decide which indicators to visualize based on which of them are present for most countries and years. The goal ist to visualize the differences in education between different countries and the change over several years. A further focus will be to show the differences between male and female in the education domain. The main dataset used will be the one form the World Bank and The OECD data will be used to give more insight on some indicators for OECD countries.

[4] https://stats.oecd.org/index.aspx