PPOL564 - Project Proposal

Maryam Khalid Shah - ms4684

10/31/2020

Problem Statement: *The effect of change in the state of democracy on a country's education budget (between 2008 and 2017).*

For the purpose of this analysis, I use the Economic Intelligence Unit's (EIU) definition of democracy and their scoring, ranking and categorization of countries according to their state of democracy[1]. I define education budget as the percentage of government expenditure on education for that year. I have limited my analysis to 10 years (from 2008 to 2017) because of two main reasons. Firstly, 10 years would allow me to see more significant changes as opposed to analyzing only 2 or 3 years of data. Secondly, Democracy Index data is only available for every year starting from 2008. Moreover, government expenditure on education data is missing for 2018-2019 for a lot of countries so I am excluding those 2 years. I also limit my analysis to only UN member states.

Data sources: I intend to use the following data sources:

- The Economic Intelligence Unit's Democracy Index Reports 2008-2017.
- 'Our World in Data' dataset[2] on the percentage of government expenditures on education by country. This dataset uses data from the UNESCO Institute for Statistics published by the World Bank.
- United Nations Member States List[3]

Plan to obtain data:

The Economic Intelligence Unit's Report for each year is available as a pdf for registered users and includes a table of the Democracy Index and Category scores for 167 countries for that year. After registering as a user, I will download 10 pdf files and will scrape table data from these pdf files in Python.

I will scrape the table on the website 'Our World in Data' to obtain data from 2008 to 2017, and will scrape the UN website to obtain a list of UN member states.

Methods I aim to employ: To obtain my data, I aim to employ web scraping. For this, I will use the package 'requests'[4] for downloading the website, BeautifulSoup[5] from 'bs4' to parse the website as well as 'PyPDF2' to scrape pdfs.

I will perform column-wise manipulations to add year as a separate column for each country, and then merge my 10 Democracy Index datasets on the relevant variable such that I have a tidy data set with country names as the unit of observation and year and the other variables as columns. I will then use the country-converter package to standardize country names for this dataset as well as the dataset I obtain from 'Our World in Data' before merging the two.

Once I have my merged dataset, I will create a data frame of UN member countries (which I will parse from the UN website). I will then perform a 'left' merge (with my UN data frame on the left and my merged dataset on the right) on the variable 'country' to obtain my final dataset.

I will also convert the data types as needed e.g. Democracy Index Score to float, years to integers, countries to strings etc. and explore and clean the dataset. While exploring, I will use the package 'missingno' to visually assess missing values.

To calculate change in the Democracy Index Score from one year to the next, I would have to create extra columns e.g. the new column ‘change_dem_index’ would calculate the difference in the current and previous democracy index score and store the value in the row of the current democracy index.

I plan to create graphs using plotnine, seaborn[6], and matplotlib.pyplot[7]. Initially, I will use these graphs to understand and explore my data e.g. by plotting types of regime on the x-axis and the percentage of observations in each category using a bar plot. I will also create numerical summary statistic tables to understand my data better.

In the later stages of my project, I will use my data to graph line plots to visually assess the effect of change in the state of democracy on countries’ expenditure on education (as a % of GDP). I will also facet my data to view the trend across the types of regimes. I also plan to visualize how each component used to calculate the index changed over the years and its effect on the Democracy Index.

Furthermore, I plan to use machine (statistical) learning components from Week 13 to derive insights from my data.

What “success” means with respect to my project: For my project, I define success in two stages (the second stage depends on the first). The first stage would be to obtain data and create a clean dataset (as mentioned above). Since this would involve some scraping from pdfs, it may not be that straightforward as scraping directly from the web. Therefore, success at Stage 1 of my project would mean a comprehensive, clean dataset that is ready for exploration and analysis.

Stage two success means deriving interesting insights from my data − specifically finding a statistically significant effect of change in the state of democracy on a country’s education budget.

Works Cited

1. EIU Democracy Index 2019. The Economist Intelligence Unit. *https://www.eiu.com/topic/democracy-index*. Retrieved 30-10-2020.
2. Our World in Data Share of Education in Government Expenditure (% of government expenditure) (World Bank (2019)). UNESCO Institute for Statistics. *https://ourworldindata.org/grapher/share-of-education-in-government-expenditure?tab=table&time=2008..latest&country=*$NPL_{ZAF}PER_{USA}ITA_{CAF}AFG_{DZA}ATG$ Retrieved on 30-10-2020.
3. United Nations Member States. United Nations. *https://www.un.org/en/member-states/*. Retrieved on 30-10-2020.
4. Chandra, R. V., & Varanasi, B. S. (2015). Python requests essentials. Packt Publishing Ltd.
5. Richardson, L. (2007). Beautiful soup documentation. April.
6. Waskom, M. et al., 2017. mwaskom/seaborn: v0.8.1 (September 2017), Zenodo. Available at: https://doi.org/10.5281/zenodo.883859
7. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.