COVER PAGE

Title: Final Project Report - Data Science 2020

Subtitle: The Role of Education in Poverty Alleviation: A Cross Country Analysis

Date: December 17, 2020

Presenter: Junyung Kim, Danice Pe Benito, Joseph, Mattew, Brandon

Prepared for the fourth assignment of Data Science 2020

Class number: CSCI-4350/CSCI-6350

**Table of Contents**

[Table of Contents ii](#_Toc58537523)

[1. Data-Oriented Workflow Construction 1](#_Toc58537524)

[2. Workflow Description 1](#_Toc58537525)

[3. Data Cleansing 2](#_Toc58537526)

[4. Data Analytics 3](#_Toc58537527)

[5. Data Stewardship 4](#_Toc58537528)

[6. Reference 5](#_Toc58537529)

# **Data-Oriented Workflow Construction**

The project title that Team 3 is focusing on is poverty eradication, one of the Sustainable Development Goal supported by the United Nations (UN). Even though there are many definitions of poverty [1], the World Bank (WB) has its guideline to define the poverty level. It explains: Extreme poverty rates have fallen by more than half since 1990. While this is a remarkable achievement, one-in-five people in developing regions still live on less than $1.90 a day [2].

Most of the time, policymakers follow the goal of higher per capita income or reduce income inequality to achieve the objective of poverty alleviation [3]. However, the level of education is an essential factor for poverty risk. As better education level may positively affect the income per capita, it can be claimed that estimates of education could be of much use for anti-poverty policy. The prime intention behind the study is to validate the hypothesis below:

1. Improvement in education and health can give aid to reduce the poverty level.

Figure 1 shows the data-driven workflow of the project. First, raw data were imported from an external source. Since WB has classified all the world countries on the basis of their incomes in four groups, namely low income; lower middle income, upper middle income and high income countries, countries are sampled from each group so that one can find out how income level gives affects to poverty level. In addition, interpolation and extrapolation was adopted to generate missing values. After the interpolation/extrapolation step, datasets are integrated based on the subject and then it clustering analysis has done to identify time dependent patterns among poverty level and other factors. Regression analysis has been done to identify most relevant factors for poverty eradication. More detailed description of each stage of the workflow will be discussed in Section 2.

|  |
| --- |
|  |
| **Figure 1. Data-oriented workflow.** |

# **Workflow Description**

**2.1. Raw Data**

Table 1 summarizes the source of the dataset used in this project.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **Unit** | **Source** | **URL** |
| Government expenditure on education, total | (% of GDP) | WB | <https://data.worldbank.org/indicator/SE.XPD.TOTL.GB.ZS> |
| Literacy rate, adult total | % of people ages 15 and above | WB | <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS> |
| School Enrollment, tertiary | % gross | WB | <https://data.worldbank.org/indicator/SE.TER.ENRR> |
| School Enrollment, secondary | % gross | WB | <https://data.worldbank.org/indicator/SE.SEC.ENRR> |
| Poverty head count | % of population | UNdata | <https://data.un.org/Data.aspx?d=WDI&f=Indicator_Code%3ASI.POV.NAHC> |
| GDP per capita average annual growth rate | % growth | UNdata | <http://data.un.org/Search.aspx?q=gdp+growth> |

**Table 1. Data description and data source URL**

Since different data sources, data techniques and tools for data collection are used, it is important to have data set with proper meta data base. WB allows database users to download dataset with the corresponding meta data. For data and information preservation, those meta data are well archived in Github repository during the project. Table 2 shows exemplary metadata for *Literacy rate, adult total*.

|  |  |
| --- | --- |
| **Metadata** | **Description** |
| INDICATOR\_NAME | Literacy rate, adult total (% of people ages 15 and above) |
| SOURCE\_NOTE | Adult literacy rate is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life. |
| SOURCE\_ORGANIZATION | UNESCO Institute for Statistics (http://uis.unesco.org/). Data as of September 2020. |

**Table 2. Metadata of *Current health expenditure, per capita***

# **Data Cleansing**

For this Data Science Project, once the team found the datasets needed for the project, the team first had to clean the data. The reason why the team has to clean the data is to make sure that when the team does the data analysis, the data analysis can be understood clearly and can create a clear and understandable visualization of the data using data analytic methods such as linear regression or clustering. Most of the datasets came from a resource called the World Bank (https://data.worldbank.org/) where the team would download the available Excel file. In most of the datasets, the team observed that there were many missing values in the datasets. The team then used two approaches to clean the data for proper data analysis. The two approaches are interpolation and extrapolation: For the interpolation, if there was a missing value in a cell and if the cell was in between two known values, the team replaced missing values (i.e., “NA”) with linearly interpolated value. For the extrapolation, if there are a consistent number of cells with missing values, the team would pick either the first or last value depending which is needed. For example, in a row of 20 columns, if the first 10 columns have missing values while the last 10 has filled in values, the team would insert the value from column 11 in columns 1 to 10. The same would occur if the last 10 columns were missing values where the team would choose the value from column 10 to fill into columns 11 to 20. The final method that the team used for data cleansing is the use of integrating metadata (i.e., country region, country income group) into the main data sheet. The team used this method so that there are more variables for analysis. For example, if there is metadata that shows a country’s income level and geographic location, the team would take those two columns and place them into the main data sheet to allow for better and clearer data analysis.

# **Data Analytics**

**4.1. //**

As described in Section 1, WB has classified all the world countries on the basis of their incomes in four groups. For data preparation purpose, seven countries are sampled from each group.

**4.2. //**

For data interpolation, average annual growth was calculated by using two nearest edge values. For extrapolation, overall average annual growth was calculated by using all available values.

# **Data Stewardship**

Data stewardship is the management and oversight of an organization's data assets to help provide users with high-quality data that is easily accessible in a consistent manner [5]. In terms of data life cycle, one can consider three main elements: data acquisition, data curation and data preservation.

**3.1. Data Acquisition**

Raw data was acquired from the data repository mentioned in Section 2.1 and the basic data processing (i.e., refilling missing data) has done. The processed data are saved in open database repository (Github) and shared with other team members.

**3.2. Data Curation**

The data saved in Github are controlled and managed by the data steward, who is responsible for data-oriented workflow to ensure fitness of data elements - both the data content and metadata. Anyone who wanted to upload new dataset into the Github repository needs to communicate with the data steward, so that the data steward can check whether the dataset and meta data are well organized and completed. In addition, anyone who used the dataset and completed analysis needs to write an analysis report, which includes proper provenance, computer source code, and metadata description, and the data steward sees whether the analysis report was written in compliance with the provenance format.

**3.3. Data Preservation**

Since our team is planning to publish conference paper, this dataset and analysis result may need to be revisited. For future purpose, every content including the dataset, metadata, and provenance will be stored in the Github repository. After the coursework, the Github repository will be managed by the data steward.

# **Reference**

1. Hagenaars, Aldi, and Klaas De Vos. "The definition and measurement of poverty." Journal of human resources (1988): 211-221.
2. Ferreira, Francisco, et al. “The International Poverty Line Has Just Been Raised to $1.90 a Day, but Global Poverty Is Basically Unchanged. How Is That Even Possible?” World Bank Blogs, 4 Oct. 2015, blogs.worldbank.org/developmenttalk/international-poverty-line-has-just-been-raised-190-day-global-poverty-basically-unchanged-how-even.
3. Janjua, Pervez Zamurrad, and Usman Ahmed Kamal. "The Role of Education and Health in Poverty Alleviation a Cross Country Analysis." Journal of Economics, Management and Trade (2014): 896-924.
4. McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." Journal of Open Source Software 2.11 (2017): 205.
5. Rosenbaum, Sara. "Data governance and stewardship: designing data stewardship entities and advancing data access." Health services research 45.5p2 (2010): 1442-1455.