ETL Report

Group 5:
Jack Lynn
Stanley Perez
Tavianne Kemp
Sam Wainright

Process Date: 13 July 2022

Introduction

Set the stage. Introduce the problem that you are trying to solve. Identify sources of data. Describe why the data needs to be transformed.

About WorldBank: A partnership of 5 financial institutions that provide loans and grants, policy advice, and research to low- and middle- income countries (The World Bank, 2022a).

Data.worldbank.org supplies a data catalogue organized by "indicator". Upon selecting a data set, general information and a broad overview visualization that gives an initial look at the trend over a period of time in line, bar, or map form.

Group 5 focused on "Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)" under the "Social Development" indicator. Before beginning the ETL process, the following questions are considered for the exploration:

Our questions:

1. How do the indicators differ in countries, especially in those that are labeled as low income?
2. What is the labor force participation rate per region?
3. What is the labor force participation rate per income group?
4. In regions that show children under the age of 14 working and attending school, what is their literacy rate upon entering adolescence?

Data Sources

The data can be found on the World Bank Open Data page and selecting "Browse by Indicator". The data "Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)" data is located under the "Social Development" indicator (The World Bank, 2022b). This was accessed July 2022 and at the time of extraction, was last updated on June 2022.

Extraction

The data is extracted from data.worldbank.org. The source page (The World Bank, 2022b) supplies the data in csv, xml, or excel format under the download section. Download the data as a csv file. The files provided upon downloading are, “Data”, “Metadata – Country”, and “Metadata – Indicators”. Keep all three csv files saved. Transformation process will occur with the “Data” csv file.

Transformation

Load in both the “data” csv and the “metadata-country” csv:

1. Inner merge the metadata onto the data via the country code to get the regions.
2. Remove columns that are not essential: indicator code, table name and country code.
    - Special notes column is also removed. “Special notes” supplies geopolitical and/or socio-economic comments about the country, which may be helpful in forming conclusions, but not necessary for initial data exploration.
3. The resulting table will be left with: “country name”, “indicatory name”, “region”, various years and “income group”.
4. Years span from 1960 to 2022. These are aggregated into a single “years” column.
    - The “year” column displays the singular year value per indicator
    - Hard copy (.copy()) is necessary to make sure that we aren’t referring back to the original data frame. If it was not included, it would be writing over itself
    - This singular column allows us to look at data for a specific year.
5. Handle nulls:
    - Apply N/A to nulled regions using .fillna() and setting it equal to N/A
    - The nulled regions are smaller developments/communities like “late-demographic dividend”, “sub-saharan Africa”, “heavily indebted poor countries”.
6. Rename columns to omit spaces

Load

1. Load transformed data as a csv table

Conclusion

References

The World Bank. (2022a, May 18). What We Do. Retrieved July 13, 2022, from data.worldbank.org website: https://www.worldbank.org/en/what-we-do

The World Bank. (2022b, June 30). Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate) | Data. Retrieved July 13, 2022, from data.worldbank.org website: https://data.worldbank.org/topic/social-development?view=chart