Final Report ETL Project

By: Michele Bowman, Tania Bukengolts, Prerna Gupta, Laura Koczaja

Overview

The data sets selected by our team measure the child labor rates, secondary education completion rates, and the GDP values for each country, in order to provide a basis for analysis among these three factors.

Extraction

We pulled CSV files from three (3) separate sources for secondary education completion rates by country, child labor rates by country, and gross domestic product (GDP) by country. The sources for our data are:

- Humanitarian Data Exchange
- UNICEF
- World Bank

Transformation

Child Labor Data

The child labor CSV file was imported into a Jupyter notebook file to create a data frame using pandas and numpy. Then, several columns with NaN data were dropped out of the data frame.

	Unnamed: 0	Total	Unnamed: 2	Sex	Unnamed: 4	Unnamed: 5	Unnamed: 6	Place of residence	Unnamed: 8	Unnamed: 9	...
0	NaN	NaN	NaN	Male	NaN	Female	NaN	Urban	NaN	Rural	...
1	Afghanistan	29	NaN	34	NaN	24	NaN	-	NaN	-	...
2	Albania	5	NaN	6	NaN	4	NaN	-	NaN	-	...
3	Algeria	5	NaN	6	NaN	5	NaN	5	NaN	6	...
4	Andorra	-	NaN	-	NaN	-	NaN	-	NaN	-	...

5 rows × 56 columns

The first row was dropped because the column names were loading incorrectly (due to the way the CSV file was organized), and the remaining column names were renamed.

	country	total_percent	male_percent	female_percent
1	Afghanistan	29	34	24
2	Albania	5	6	4
3	Algeria	5	6	5
4	Andorra	-	-	-
5	Angola	23	22	25

Then, the rows that had a dash (–) were replaced with NaN and dropped out of the data frame.

	country	total_percent	male_percent	female_percent
1	Afghanistan	29	34	24
2	Albania	5	6	4
3	Algeria	5	6	5
5	Angola	23	22	25
7	Argentina	4	5	4

Finally, the index was set to the ‘country’ column and the results exported to a CSV file.

Secondary Education Completion Rate

The secondary education completion rate CSV file was also imported into the same Jupyter notebook to create a data frame.

	ISO3 Code	Countries	Total	Male	Female	Urban	Rural	Poorest	Poorer	Middle	Fourth	Richest	Source
0	AFG	Afghanistan	24.0	33.0	15.0	39.0	17.0	12.0	15.0	14.0	23.0	46.0	DHS 2015
1	ALB	Albania	46.0	47.0	44.0	69.0	25.0	16.0	22.0	34.0	61.0	86.0	DHS 2008-09
2	DZA	Algeria	38.0	30.0	47.0	43.0	30.0	22.0	27.0	37.0	41.0	62.0	MICS 2012-13
3	AND	Andorra	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	AGO	Angola	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Eight (8) columns were dropped out of the data frame, then column names were renamed.

	iso3_code	country	total_percent	male_percent	female_percent
0	AFG	Afghanistan	24.0	33.0	15.0
1	ALB	Albania	46.0	47.0	44.0
2	DZA	Algeria	38.0	30.0	47.0
3	AND	Andorra	NaN	NaN	NaN
4	AGO	Angola	NaN	NaN	NaN

Again, the blank rows were dropped out of the data frame, and the index was set to the 'iso3 code' (country code) column.

iso3_code	country	total_percent	male_percent	female_percent
AFG	Afghanistan	24.0	33.0	15.0
ALB	Albania	46.0	47.0	44.0
DZA	Algeria	38.0	30.0	47.0
ARG	Argentina	59.0	53.0	66.0
ARM	Armenia	93.0	88.0	96.0

The results were exported to a CSV file.

GDP Data

The GDP CSV file was imported into the Jupyter notebook to create a data frame. Many columns were dropped out of the data frame because of a reference to years that were not needed for analysis. Column names were again renamed, to maintain consistency across all 3 data frames. Blank rows were dropped out of the data frame as well, and the index was set to the 'iso3 code' (country code) column. The average GDP for each country was also calculated for the years 2008-2017 and loaded into a new column called

‘avg_gdp’. This was not included in the original data. Finally, the results were exported to a CSV file.

iso3_code	country	avg_gdp	2008	2009	2010	2011
ABW	Aruba	2.598939e+09	2.745251e+09	2.498883e+09	2.390503e+09	2.549721e+09
AFG	Afghanistan	1.767182e+10	1.010922e+10	1.243909e+10	1.585657e+10	1.780428e+10
AGO	Angola	1.104351e+11	8.853861e+10	7.030716e+10	8.379950e+10	1.117897e+11
ALB	Albania	1.243410e+10	1.288135e+10	1.204421e+10	1.192695e+10	1.289087e+10
AND	Andorra	3.296477e+09	4.007353e+09	3.660531e+09	3.355695e+09	3.442063e+09

Load

The last step in the process was to load the data from the exported CSV files into a PostgreSQL Database. A relational database was selected over a non-relational database, because our data was organized into table-like data frames and had ‘country’ columns that related to each other, and that could easily be ‘JOINed’ using SQL queries.

SQLAlchemy was used to set up the engine connection so the cleaned child labor, secondary education, and GDP data could be exported into PostgreSQL to query for analysis. The database that we loaded the data into was named ‘ETL Project - Child Labor/Education Rates/GDP’, and the tables were named as the data frames were – ‘child_labor’, ‘education’, and ‘gdp’.

Future Analysis

The data in the tables will help to determine if a correlation exists between child labor, the secondary education completion rate, and GDP for the listed countries. The data can be used to explain which gender by country has completed secondary education and show the GDP relating to that particular country. For instance, a determination can be made as to whether the countries with higher GDPs also have higher percentages of population totals of completed secondary education and lower rates of child labor.