World Bank Indicators Group Exercise

Group: Dalton Bode, Daniel Brickman, Colin Beveridge, Sawyer Tucker, Chris Nash, Darrell Gerber

Date: 1/13/2022

Financial Sector ETL Report

**Introduction:** For this project, the group analyzed three separate tables from the Financial Sector section of the Indicator page of the World Bank Website. The first (API Inflation) and second (Metadata Inflation) tables come from the “Inflation, consumer prices (annual %)” [(Inflation, consumer prices (annual %), 2] link and the third (API Broad Money Growth) and fourth (Metadata Broad Money Growth) tables come from the “Broad money growth (annual %)” [Broad Money (% of GDP), 1] link. The group came up with several comparisons to perform given the dataset. First, the group wanted to analyze how inflation over time from the first table correlates with the classified income group of the country in the second table. Additionally, the group decided to compare inflation across different regions of the world and compare inflation between the United States and some other countries in comparable income groups. Finally, the group decided to compare net broad money growth from the third table to inflation from the first table. These tables need to be merged with their respective metadata to compare across the selected variables properly. The second table must be merged to the first by country code to identify each country with an income group. Then, the third and fourth tables must be merged on country code as well. In order to answer the fourth question, it is better to simply compare the final graphs of inflation and broad money growth over time, rather than merging the tables and comparing directly.

**Data Sources:** The group found the data on the indicator page of the World Bank Website. (<https://data.worldbank.org/indicator>) It was accessed on January 13th, 2022.

Citations

1. Broad money (% of GDP). Data. (n.d.). Retrieved January 13, 2022, from <https://data.worldbank.org/indicator/FM.LBL.BMNY.GD.ZS?view=chart>
2. Inflation, consumer prices (annual %). Data. (n.d.). Retrieved January 13, 2022, from <https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?view=chart>

**Extraction:** The group obtained the World Bank data by accessing the proper tables on their website, and then downloading CSV files for each.

1. An important step is that when opening the files, do not open them directly out of excel. Go to where they are stored and open from there. A few group members tried to open the file from excel and then navigating to the file, and excel asked the member to parse the file manually, which is difficult to do correctly.
2. The columns that were not strictly necessary were deleted in excel. In the Inflation API table, the group deleted the Indicator Name and Indicator Code columns, because they are the same throughout the table. In the Inflation Metadata table, the SpecialNotes and TableName columns were deleted. For the Broad Money Growth API data, the same columns were deleted as in the Inflation API and in the Broad Money Growth Metadata, the same columns were deleted as in the Inflation Metadata (the tables mirror each other).
3. The group then loaded that data into python through pandas, using the function pandas.read\_csv(‘filelocation’).

When loading the API data files (the big tables) for both categories, the skiprows argument is included, so pandas.read\_csv(‘filelocation’,skiprows=4). This is because the data has a header that changes the data frame format if pandas is not told to ignore it. Ensure that each table is clearly named so that the correct tables can be merged.

**Transformation:**

1. First, the API tables were merged with their respective Metadata tables on Country Code. To be clear, Inflation API was merged with Inflation Metadata and Broad Money API was merged with Broad Money Metadata.
2. After merging the tables, the group noticed an extra column with an “unnamed” label. There is a comma at the end of each row that causes the creation of an extra empty column in the pandas data frame. The group dropped this column.
3. Next, the group dropped all rows for countries that had no yearly data. This was done using the dropna(axis=0, thresh=4). The thresh argument sets a threshold for all rows that have all but four NaN entries. This is done because in every row, there are at least 4 non-zero columns that come from the tables. It is best to inspect the table before selecting the number for thresh. It should be equal to the number of columns that are not years.
4. Next, the group filtered out countries into their own income group data frames to allow for comparison between and within each income group by country. This was done with basic filtering in pandas. Filter the old data frame by the ‘IncomeGroup’ column and create new data frames for each income group. Also, there is an aggregate row for each income group. The group also extracted this row into its appropriate new data frame. Here is the basic code structure1.

APIhighincome = APIall[APIall[‘IncomeGroup’] == ‘High income’]

APIhighincome.append(APIall[APIall['Country Name'] == 'High income'])

**Load:** Finally, the group loaded the tables into csv files named according to the variable they represent. For the high-income group for inflation, the output file name would be InflationMerged\_HighIncome.csv. For lower middle income broad money, the name would be BroadMoneyMerged\_LowerMiddleIncome.csv. These are now all exported to csv files where they can be opened and used for creating visualizations to answer the proposed questions.

**Conclusion:** Overall, to answer the four questions presented in the beginning of the document, these cleaning steps are very important to ensure the data is clean enough to plot and use.