Global Economic Indicators Between 1987-2021 For the Top 10 Highest GDP Countries in 2022

Phase 1 - Report

Syed Hassan - shassa25@illinois.edu
Gordon Ochi - gochi2@illinois.edu
Noah Rogers - noahrr2@illinois.edu

1. **Identify a dataset**
    a. *25 years worth of global economic indicators data for the top 10 countries by GDP in 2022 (United States, China, Japan, Germany, United Kingdom, France, India, Italy, Brazil, Canada)*
    https://databank.worldbank.org/source/global-economic-monitor-(gem)#

2. **Develop a target**
    a. **U_0 - Zero data cleaning case:** Plotting macroeconomic variables in the United States from 1991 onwards, such as the Gross Domestic Product (GDP) or Consumer Price Index (CPI). Likely applicable to other countries as well, with different time ranges depending on when a specific country began collecting data consistently.

    b. **U_1 - Main use case:** On a general level this would be an in-depth exploratory analysis of the dataset, relating countries and various statistics, perhaps lending itself to predictive analysis for future data. Specific tasks could be computing new statistics for each country, such as the import/export ratio to determine a trade surplus or deficit, or global macroeconomic indicators relating statistics for various countries. Specific examples of queries are given in Section 2.b.i.
        Queries:
        1. Import/export ratio for a given country.
        2. Correlation of GDP to unemployment rate.
        3. Correlation of Retail Sales Volume to CPI.
        4. Correlation of the Stock Markets to CPI.
        5. Correlation of Retail Sales Volume to Total Reserves.
        6. Stock market total value in USD in the United States versus other countries.
        7. Import/export relations among various countries (e.g., does China have increased exports when the United States has increased imports).
        8. Industrial production versus GDP (how has a country's industrial production changed versus GDP).
        9. Comparison of seasonally adjusted CPI to not seasonally adjusted CPI (how much of a difference does seasonally adjusting CPI make?).
        10. Core CPI versus CPI (is the Fed's claim that gas and food are volatile and Core CPI is a more stable/reliable indicator of inflation valid?).
        11. Removing potentially redundant columns. Also may need to do some background research and determine if certain statistics aid our use case, or are unnecessary information (e.g., total reserves).
        12. Clean up null data instances as much as feasible (some data reported monthly, others quarterly). Split the dataset into a combination of quarterly and yearly statistics, and monthly.

13. United States unemployment rate as a leading/trailing indicator for worldwide unemployment rate increase/decrease?
14. Correlation of industrial production to exports/imports.
15. How do industrial production and exports from China in the years 2019 to the end of 2020 correlate with United States imports, retail sales, and CPI? In other words, do supply chain disruptions as evidenced by macroeconomic variables correlate with increased inflation?
16. Checking for outliers. We will want to be sure that values for a particular statistic in the dataset fall within a reasonable range, and potentially replace any obvious outliers.

c. **U_2 - Data cleaning not sufficient:** A use case where this particular dataset would never be sufficient would be for the purposes of cross-correlating with stock market data to predict stock market price changes. Price movement in the stock market is incredibly complex, often resulting from numerous factors beyond simple monthly or quarterly macroeconomic data. While this dataset might aid stock market equity price predictions, it would likely be insufficient by itself to contribute significantly in any algorithmic or statistical models.

3. **Describe the dataset**
   a. **Dataset Narrative:** The original dataset is broken up into groupings of rows by country and series that report back financial indicators by year from January 1987 till July of 2020. The financial indicators are split into yearly metrics (formatted as “${YEAR} [${YEAR}]”), monthly metrics per year (formatted as “${YEAR}M01 [${YEAR}M01]” following through M01-M12), and quarterly metrics per year (formatted as “${YEAR}Q1 [${YEAR}Q1]” following through Q1-Q4). A diagram of the dataset is shown below.

Global Economic Indicators 1987 - 2021	
Column Name	Column Type
Country	VARCHAR
Country Code	VARCHAR
Series	VARCHAR
Series Code	VARCHAR
Yearly Financial Indicators (From 1987...2020)	FLOAT
Monthly Financial Indicators Per Year (From 1987...2020 Month 7)	FLOAT
Quarterly Financial Indicators Per Year (From 1987...2020 Month 7)	FLOAT

b. **Column Metadata:**

The “Country” column refers to a Nation in the world.
The “Country Code” column represents a Nation’s corresponding abbreviated symbol (such as “USA” for the United States, or "JPN” for Japan).
The “Series” column refers to what kind of data is being represented in the row (GDP, CPI Price, etc…).
The “Series Code" column is utilized as an kind of abbreviation for these types of represented data such as “CPI Price,not seas.adj,,,” being represented by the series code “CPTOTNSXN”.
Finally, each financial indicator split into Yearly, Monthly per Year, and Quarterly per Year data segments, is of the data type that suits the series in that row (i.e. if the series uses financial metrics in terms of USD, then the financial indicator for these columns match in USD).

c. **Spatial Extent:** Spatially, this data covers the top 10 countries by GDP: The United States, China, Japan, Germany, United Kingdom, France, India, Italy, Brazil, Canada.
d. **Temporal Extent:** This dataset covers data from January 1987 to July 2020, however earlier dates tend to have a higher likelihood of null data.

4. **List obvious data quality problems**
   a. The data from many yearly reports are missing for rows of interest in the dataset especially in the years between 1987 - 1995. See screenshot below.

Series	1988 (1988)	1989 (1989)	1990 (1990)	1991 (1991)	1992 (1992)
Exports Merchandise, Customs, constant US$, millions, seas. adj.	..	..	..	531179.12..	563806.51..
				..	121953.88..
					..
Exports Merchandise, Customs, current US$, millions, not seas. adj.	..	..	..	131.22242..	137.56222..
				401.23	508.72
				31623	35793
				71970	84940
				103939	107863
				169602	178027
				269407.55..	305860.66..
				315224	339650
				402876	430196
				421730	448164.5
Exports Merchandise, Customs, current US$, millions, seas. adj.	..	..	..	402.76699..	509.40003..
				31699.905..	35746.435..
				71651.925..	84266.045..
				103950.85..	107804.01..
				128548	134811
				269931.32..	306322.34..
				315747	340013
				403246.42..	430373.69..
				421922.9	448163.5
				..	136015.00..

b. Some data are provided quarterly, others only monthly. See example shown below.

93M07 [	1993M08 [	1993M09 [	1993M10 [	1993M11 [	1993M12 [	1993Q1 [1	1993Q2 [1	1993Q3 [1	1993Q4 [1	1994 [199	1994M01 [	1994M02 [	1994M03 [
9.06391	69.33653	69.47284	69.74546	69.92721	69.92721	..	..	..	..	71.11614	70.10896	70.42701	70.79051
	..	..	..	..	..	..	..	..	..	..	..	..	..
6.44045	66.57839	66.67035	66.94622	67.13014	67.26808	..	..	..	..	68.15319	67.26808	67.452	67.63592
846975	2.840909	2.763997	2.752294	2.744546	2.810963	..	..	..	..	2.595605	2.45098	2.515723	2.651779
9.02866	69.25528	69.3006	69.52722	69.75384	69.93514	..	..	..	..	70.93227	70.02579	70.16176	70.38838
66.6399	66.8245	66.96295	67.23984	67.28599	67.28599	..	..	..	..	68.40512	67.47059	67.70134	67.93209
	..	..	..	..	..	1682400	1702200	1720500	1753400	7287200	..	..	..
	..	..	..	..	..	1682400	1702200	1720500	1753400	7287200	..	..	..
3.70599	34.35137	34.30625	34.62637	35.09327	35.75609	..	..	..	..	36.2304	36.94277	37.29855	36.45209
3.70599	34.35137	34.30625	34.62637	35.09327	35.75609	..	..	..	..	36.2304	36.94277	37.29855	36.45209
6363.38	77685.29	78416.45	76862.12	76249.59	75524.47	..	..	..	..	76607.91	76494.11	78337.77	79601.57
6.9	6.8	6.7	6.8	6.6	6.5	..	..	..	..	6.1	6.6	6.6	6.5
77E+11	1.77E+11	1.78E+11	1.79E+11	1.80E+11	1.81E+11	..	..	..	..	2.24E+12	1.81E+11	1.81E+11	1.83E+11
74E+11	1.80E+11	1.81E+11	1.81E+11	1.80E+11	1.79E+11	..	..	..	..	2.25E+12	1.81E+11	1.82E+11	1.85E+11
750596	0.749801	0.750596	0.753773	0.749801	0.742653	0.749801	0.75642	0.750331	0.748742	0.764297	0.742653	0.74583	0.746624
741204	0.738077	0.737295	0.735731	0.733385	0.729476	0.740683	0.746416	0.738858	0.732864	0.752411	0.733385	0.737295	0.738077
3032.52	47882.27	49487.15	49580.86	50576.82	50319.56	140228.6	144265.1	145401.9	150477.2	662702.6	49542.65	51291.61	52943.37
48316.8	48610.1	50526.8	53889.3	51433.5	48719.2	134619	144541.5	147453.7	154042	663251.8	46514.1	46653.7	54662.7
4371.27	64830.63	67315.6	71492.76	68596.16	65601.58	179503	191100.8	196517.5	205690.5	866834.1	62632.35	62552.72	73213.13
4803.37	64874.4	67119.9	67389.92	68963.49	68980.4	189304	193280.1	196797.7	205333.8	880205.7	67553.36	69567.3	71731.54
	..	..	..	..	..	2351987	2365644	2376913	2409288	9886789	..	..	..
799783	0.800875	0.801464	0.801554	0.802554	0.804311	0.79607	0.797726	0.800707	0.802806	0.816298	0.808298	0.807735	0.809447
	..	..	..	..	..	2462800	2477100	2488900	2522800	10352600	..	..	..
37422	37833	38413	40006	40049	41420	113707	116241	113668	121475	512627	39843	38293	42487
803645	0.803645	0.802817	0.802817	0.803645	0.806131	0.800055	0.803093	0.803369	0.804198	0.819732	0.81193	0.81193	0.813587
35464.2	36876.3	37955.5	41147.6	40293.8	41412.2	114026.8	117913.5	110296	122853.6	512625.1	37560.5	37125.6	46138.6
5790.19	47239.61	47928.51	49910.56	49901.92	51497.49	142835.5	145715.5	141958.3	151310	627744.1	49292.45	47407.85	52488.89
4129.16	45886.28	47277.9	51254.03	50138.78	51371.56	142521.1	146823.1	137293.4	152764.4	625109.1	46260.74	45725.1	56710.07
1	1	1	1	1	1	..	..	..	..	1	1	1	1
1	1	1	1	1	1	..	..	..	..	1	1	1	1
3.13955	63.02232	63.43466	63.44766	63.90893	64.48361	..	..	..	..	66.45259	64.00322	64.95686	66.07921

c. Some statistical data seems repetitive, e.g. there are six different types of import data. While the labels indicate that these are computed slightly differently, it is likely that most are a bit redundant.

d. Some columns also do not seem useful. For example, the Series Code column does not seem useful for our purposes. Country Code is also redundant.

e. The default arrangement with the time series along the top axis and the statistics of interest listed repetitively in the third column for each country is perhaps not ideal - see the sample below from OpenRefine. These axes will need to be

swapped such that the statistics are listed at the top, and the index iterates by time step and country.

All		Country	Country Code	Series	Series Code	1987 [1987]	1987M01 [1987M	1987M02 [1987M
1.		United States	USA	Core CPI,not seas.adj.,,,	CORENS	53.70627845	52.61579564	52.75210599
2.		United States	USA	CPI Price, % y-o-y, median weighted, seas. adj.,	CPTOTSAXMZGY	..	..	..
3.		United States	USA	CPI Price, seas. adj.,,,	CPTOTSAXN	52.24043012	51.2212168	51.405135
4.		United States	USA	CPI Price, % y-o-y, not seas. adj.,,	CPTOTSAXNZGY	3.578211654	1.364877161	1.914311759
5.		United States	USA	Core CPI,seas.adj.,,,	CORESA	53.56179719	52.53067311	52.66664552
6.		United States	USA	CPI Price,not seas.adj.,,,	CPTOTNSXN	52.43738665	51.3182609	51.50285896

5. **Devise an initial plan**

S_1 - Description of dataset and matching use case:

Go through the dataset and describe the various statistics, and whether they match our use case U_1. Check our list of combined statistics from Section 2.b.i and confirm that they fit the use case.

S_2 - Profiling of dataset and identification of quality problems:

The data quality problems explained in Section 4 will be an issue and need to be addressed for U_1, and other instances of data quality issues have all been addressed under some of the queries in U_1 to ensure the dataset will be good enough.

S_3 - Data cleaning process and tools:

Step 1 - Columns and dataset reshuffling

The first step in this process will be splitting the dataset into a quarterly and yearly statistics dataset, as well as a monthly dataset. After this the time series row that is currently the top row in the dataset after the Country-Series Code columns will need to be transposed with the Series column, such that the individual statistics become the primary columns. Then each step in the index will be rearranged such that the 10 country's data for that time step are listed adjacent to one another vertically. This initial task can likely be completed efficiently in Python with the Pandas package.

Step 2.1 - Quarterly and Yearly Dataset Cleaning

After the initial reshuffling of the dataset outlined above, likely the only needed tool for completing the remaining cleaning would be OpenRefine, perhaps with a bit of the regex extension within OpenRefine. We will first investigate and mediate the fundamental data problems listed in Section 4, and will then implement the relevant queries listed in Section 2.b.i.

Step 2.2 - Monthly Dataset Cleaning

The separate monthly dataset obtained from the splitting in Step 1 will be processed in a nearly identical process. OpenRefine will be used to fix

fundamental data problems listed in Section 4, and relevant queries from Section 2.b.i will be implemented.

Step 3 - Combined Statistics Cleaning/Processing

Once the two cleaned datasets are obtained from Steps 2.1 and 2.2, it will be possible to create a third sub-dataset that examines the queries from Section 2.b.i having to do with statistical relations among the various countries. Datalog or SQLite will be considered for this task, since it will involve combining information from separate datasets.

Step 4 - Workflow Diagram and Provenance

We will use the workflow histories from OpenRefine, as well as documented changes with Datalog/SQLite, to generate a workflow diagram in YesWorkFlow.

S_4 - Checking the new dataset:

We'll confirm that the obvious data quality problems that were explained in Section 4, will be addressed and fixed, and then work towards gathering the data in the outline provided in U_1 for (2). We will ensure that the results for the queries and cleaning for U_1 will make sense and be helpful towards answering our questions, and at that point when we are satisfied the dataset will be D'.

S_5 - Documentation of types and amounts of changes to dataset:

We're going to make a YesWorkflow diagram based on the history of our cleaning in the OpenRefine tool, and also keep a record of any and all changes that we make through the usage of python, regex, and Datalog/SQLite tools.

Tentative Assignment of Tasks:

Task 1 - Assigned to Gordon Ochi. Task must be completed by 07/10.

Task 2.1 - The quarterly and yearly cleaning will be done in group working sessions with all three team members between 07/10 and 07/17, ensuring a consistent understanding of the cleaning process among team members. The timing of completion of this task will align with the end of week 9.

Task 2.2 - Assign to Noah Rogers. Will follow the same process defined by the group in Task 2.1, but for the monthly data. Task must be completed by the end of week 10 (07/24).

Task 3 - Assign to Syed Hassan. Task must be completed by the end of week 11 (07/31)

Task 4 - This will be completed in a final group working session the weekend of (07/30). Task must be completed by the end of week 11 (07/31). Thisfinal group working session will also serve as a means of wrapping up the project and reaching a consensus on the final form for submission.