Databases

Final Project

Samuel Blanco Ponce

June 2023

Objectives

The objectives may be divided into two parts.

Part I: study correlation between different economic variables. In particular the Pearson correlation coefficient will be computed for:
- Any pair of the top 5 stocks in the USA.
- Per capita GDP PPP of any pair of the countries CHN, ESP, GBR, SWE, USA.
- Any pair of the following economic factors in the US economy: per capita GDP PPP, poverty rate, unemployment rate, and

Part II: study the effect of COVID-19 on multiple economic factors, these being:

- Gross Domestic Product (current US$)
- Gross Domestic Product per capita (current US$)
- Gross Domestic Product based on Purchasing Power Parity (current international US$)
- Gross Domestic Product based on Purchasing Power Parity per capita (current internal US$)
- Gini Index
- Poverty headcount ratio at $2.15 a day (based on 2017 PPP)
- Total percentage of the labor force unemployment (according to ILO estimates)
- Total CO_2 equivalent* emissions (in kilotons)
- Per capita CO_2 equivalent emissions (in metric tons)
- Price of current five largest stocks

* 'equivalent' since other emissions are considered as well.

The effect will be measured by computing the average growth of any particular of the aforementioned values during the period 2010-2019 (pre COVID) and compare it for the average growth in the post 2020 (post COVID).

For any pair of the above we will compute the Pearson Correlation Coefficient.

Data

For each of the five countries in the study we have chosen the top five stocks in terms of national market capitalization as measured by https://companiesmarketcap.com/ and such that monthly stock price data is available during the period 2010-2021 as measured by https://finance.yahoo.com/.

For example, the top 8 stocks by market cap in Spain are IDEXY, IBE.MC, SAN, BBVA, AMS.MC, CABK.MC, CLNX.MC and NTGY.MC (see *Figure 1*). However, only IBE.MC, SAN, BBVA, CABK.MC and NTGY.MC comply with the aforementioned data criteria (available data during the 2010-2021 period), and thus it is these five stocks that are included in the study. The complete list of stocks is given in *Figure 2*.

Figure 1. Top 8 stocks in Spain (June, 2023). Screenshot taken from https://companiesmarketcap.com/.

Country	China (People's Republic of) (CHN)	Spain (ESP)	United Kingdom (The) (GBR)	Sweden (SWE)	United States of America (The) (USA)
Top 5 Stocks	China Mobile (0941.HK)	Banco Bilbao Vizcaya Argentaria (BBVA)	AstraZeneca (AZN)	Assa Abloy (ASAZF)	Apple (APPL)
	ICBC (1398.HK)	CaixaBank (CABK.MC)	HSBC (HSBC)	Atlas Copco (ATCO-B.ST)	Amazon (AMZN)
	Kweichow Moutai (600519.SS)	Iberdrola (IBE.MC)	Linde (LIN)	Hexagon (HXGBF)	Alphabet (Google) (GOOG)
	PetroChina (601857.SS)	Naturgy (NTGY.MC)	SHELL (SHEL)	Investor AB (INVE-B.ST)	Microsoft (MSFT)
	China Construction Company (601939.SS)	Santander (SAN)	Unilever (UL)	Volvo Group (VOLVF)	NVIDIA (NVDA)

Figure 2. Top five stocks by market cap in China, Spain, United Kingdom, Sweden, and the United States of America (June, 2023).

For any particular stock (e.g. SAN) a `.csv` file containing the monthly stock price data for the period 2010-2022 is downloaded from Yahoo Finance. The directory `/DataStocks` contains all 25 such `.csv` files, which were later combined into a single `.csv` file named `stockprices` so as to more easily import it.

The prices were imported into the table `stockprices` with attributes `country`, `stock`, and `price`. For the latter the Open price (i.e. the price of the stock upon opening of the corresponding trading day) was used.

The remaining economic factors were downloaded from the [World Bank's Databank](#) and may be found in the file `Data.csv` inside the directory `/Data`. All such factors are measured yearly and data was included for the period 2010-2020. The table below indicates which table (in MySQL) stored which data:

Table	Measure
`gdp`	Gross Domestic Product (current US$)
`gdppc`	Gross Domestic Product per capita (current US$)
`gdpppp`	Gross Domestic Product based on Purchasing Power Parity (current international US$)
`gdppppppc`	Gross Domestic Product based on Purchasing Power Parity per capita (current internal US$)
`giniindex`	Gini Index
`poverty`	Poverty headcount ratio at $2.15 a day (based on 2017 PPP)
`unemployment`	Total percentage of the labor force unemployment (according to ILO estimates)
`co2`	Total CO2 equivalent emissions (in kilotons)
`co2pc`	Per capita CO2 equivalent emissions (in metric tons)

Figure 3. MySQL tables created from the data.

Data Analysis

Part I: How do economic variables correlate?

For every pair (x, y) of the variables above we compute the Pearson Correlation Coefficient $r(x, y)$, defined as

$$r(x, y) := \frac{\overline{(xy)} - (\overline{x})(\overline{y})}{\left(\overline{x^2} - \overline{x}^2\right)^{1/2} \left(\overline{y^2} - \overline{y}^2\right)^{1/2}}$$

We will use such coefficients to compute the correlation between different pairs of economic factors.

CHN	1				
ESP	0.89356	1			
GBR	0.95806	0.96924	1		
SWE	0.99062	0.93688	0.97822	1	
USA	0.9836	0.9508	0.99181	0.99212	1
GDP PPP PC	CHN	ESP	GBR	SWE	USA

Figure 4. Pearson correlation coefficient between per capita Gross Domestic Product based on Purchasing Power Parity between different countries.

The high coefficients in the table above indicate that economic growth is heavily correlated between all countries in the study i.e. whenever one of the five economies in question is 'doing well' (has a high GDP PPP per capita) the other four are likely to also be doing well, and vice versa.

AAPL	1				
AMZN	0.93305	1			
GOOG	0.94417	0.95816	1		
MSFT	0.9441	0.9765	0.94961	1	
NVDA	0.87717	0.93227	0.9014	0.88716	1
Stock prices	AAPL	AMZN	GOOG	MSFT	NVDA

Figure 5. Pearson correlation coefficient between current top 5 stocks in the USA based on market capitalization.

Similarly, stock prices between the top stocks in the US change similar to one another.

We will now study the correlation between the following four factors in the US economy:
- GDP PPP per capita
- Poverty rate
- Unemployment rate
- CO_2 emissions per capita

If we partition the four factors of the table above into two groups depending on whether they tend to indicate economic growth (positive) or not (negative), then we may expect that positive-positive and negative-negative correlation to be high, while positive-negative correlation to be low.

Positive	Negative
GDP PPP per capita	Poverty rate
CO2 emissions per capita	Unemployment rate

Figure 6. Example of expected correlation results (for each row, each element in the first column should correlate negatively to the element in the right column).

However, when computing the correlation coefficients we find that there are three correlations that do not follow this logic, these being:
1. GDP PPP per capita with CO_2 emissions per capita (Pearson coefficient of -0.89886)
2. Unemployment rate with CO_2 emissions per capita (Pearson coefficient of 0.9251)
3. Poverty rate with Unemployment rate (Pearson coefficient of -0.30929)

The first and second correlations may be explained by disregarding CO_2 emissions as a pointer to economic growth, maybe due to western countries attempting to adopt a larger portion of their power from renewable sources. As for the third correlation, we have no reasonable explanation.

GDP PPP per capita	1			
Poverty rate	-0.30929	1		
Unemployment rate	-0.76904	-0.3431	1	
CO2 emissions per capita	-0.89886	-0.21652	0.9251	1
	GDP PPP per capita	Poverty rate	Unemployment rate	CO2 emissions per capita

Figure 7. Pearson correlation coefficient between different economic factors in the US economy.

An important result is that -due to its negative correlation with poverty and unemployment- GDP PPP per capita can be used as a low-resolution indicator of other relevant economic factors.

Part II: How did COVID affect the economy?

Such effect is computed by first creating two new tables in MySQL: `precovidgrowth` and `postcovidgrowth` which will, as their names suggest, store the average growth of the different economic indicators. To create such tables the `.sql` files `PreCOVID` and `PostCOVID` may be run.

Once the tables are created we compare their data to study the effect of COVID in the economy.

Country	Stock	Pre COVID avg growth	Post COVID avg growth	Difference in growth
CHN	0941.HK	-0.0517162	-0.243	-0.19128
CHN	1398.HK	0.19213	-0.761982	-0.95411
CHN	600519.SS	2.22522	1.73645	-0.48877
CHN	601857.SS	-0.588597	-0.00439059	0.58421
CHN	601939.SS	0.362455	-0.483447	-0.8459
ESP	BBVA	-0.58131	1.42913	2.01044
ESP	CABK.MC	0.24623	1.4136	1.16737
ESP	IBE.MC	0.464626	0.824014	0.35939
ESP	NTGY.MC	0.577756	0.457253	-0.1205
ESP	SAN	-0.837196	0.203516	1.04071
GBR	AZN	0.780194	1.16622	0.38602
GBR	HSBC	-0.19801	-0.0328522	0.16516
GBR	LIN	0.876681	1.7306	0.85392
GBR	SHEL	0.144558	0.536324	0.39177
GBR	UL	0.636902	-0.260894	-0.8978
SWE	ASAZF	2.49013	0.31782	-2.17231
SWE	ATCO-B.ST	1.47124	1.36385	-0.10739
SWE	HXGBF	1.5388	1.6219	0.0831
SWE	INVE-B.ST	1.25264	1.35306	0.10042
SWE	VOLVF	0.908009	1.08041	0.1724
USA	AAPL	2.12883	2.69829	0.56946
USA	AMZN	2.51584	0.680178	-1.83566
USA	GOOG	1.43428	1.59134	0.15706
USA	MSFT	1.54551	1.65888	0.11337
USA	NVDA	2.80164	4.32729	1.52565

Figure 8. Average growth of stocks in both pre-COVID and post-COVID periods.

The average change in growth is 3.0458 (computed from *Figure 8*), meaning that the top stocks generally did better during the one year period *after* COVID.

Country	Economic measure	Pre COVID avg growth	Post COVID avg growth	Difference in growth
CHN	GDP	10.1162	2.8553	-7.26094
USA	GDP	3.98151	-1.49901	-5.480523
GBR	GDP	1.68252	-5.33587	-7.01839
SWE	GDP	1.11073	2.46772	1.35699
ESP	GDP	0.0252818	-8.41682	-8.4421
CHN	GDP pc	9.49217	2.61077	-6.88141
USA	GDP pc	3.29539	-2.44127	-5.73665
GBR	GDP pc	0.975805	-5.68114	-6.65695
SWE	GDP pc	0.0882055	1.72985	1.64164
ESP	GDP pc	-0.111592	-8.86311	-8.75152
CHN	GDP PPP	6.76932	3.22591	-3.54341
USA	GDP PPP	3.2908	-3.17555	-6.46636
GBR	GDP PPP	3.45965	-5.8446	-9.30425
SWE	GDP PPP	3.27929	-1.48048	-4.75977
ESP	GDP PPP	3.49755	-11.4857	-14.98327
CHN	GDP PPP pc	6.76932	3.22591	-3.54341
USA	GDP PPP pc	3.2908	-3.17555	-6.46636
GBR	GDP PPP pc	3.45965	-5.8446	-9.30425
SWE	GDP PPP pc	3.27929	-1.48048	-4.75977
ESP	GDP PPP pc	3.49755	-11.4857	-14.98327
CHN	Gini Index	-1.4639	-3.14136	-1.67746
USA	Gini Index	0.414587	-4.33735	-4.75193
GBR	Gini Index	-0.285938	-0.609758	-0.32382

SWE	Gini Index	0.657257	-1.36519	-2.02244
ESP	Gini Index	-0.276619	1.74928	2.0259
CHN	Poverty Rate	5.66984	400	394.33016
USA	Poverty Rate	-7.33333	-80	-72.66667
GBR	Poverty Rate	15	0	-15
SWE	Poverty Rate	17.7619	33.3333	15.57142
ESP	Poverty Rate	5.90548	12.5	6.59451
CHN	Unemployment Rate	-38.1446	400	438.14462
USA	Unemployment Rate	0.740741	-80	-80.74074
GBR	Unemployment Rate	16.6667	0	-16.66667
SWE	Unemployment Rate	16.0317	33.3333	17.30158
ESP	Unemployment Rate	5.17276	12.5	7.32723

Figure 9. Average growth of economic measures in both pre-COVID and post-COVID periods.

As can be appreciated in *Figure 9*, COVID had a very different effect depending on the country in question (e.g. poverty increased only in China, Spain, and Sweden in the post COVID period).

Conclusion

We have seen that per capita GDP PPP is a good economic indicator due to its negative correlation to poverty and unemployment. As such, it shows that the economies of the different countries are closely affiliated. Our study on stock prices demonstrates that these are also interconnected.

COVID-19 had a very different economic impact regarding the country, probably due to the different measures that such countries adopted in response to the pandemic.

What remains to be understood is the negative correlation between poverty and unemployment.