Global Commodities EDA

Claudeon Susanto

2023-01-09

Introduction to the dataset

The data that I have chosen is **Technology Adoption** and can be downloaded from this [link]. This data provides very useful insights and statistics on each country's development and adoption of technology.

variable	label	iso3c	year	group	category	value
BCG	% children who received a BCG immunization	AFG	1982	Consumption	Vaccines	10.000
ag_harvester	Combine harvesters - threshers in use	AFG	2001	Production	Agriculture	2.000
all_vehicles	Total vehicles (OICA)	AFG	2005	Consumption	Transport	660000.000
aluminum	Aluminum primary production, in metric tons	ALB	1850	Production	Industry	0.000
atm	ATMs	ABW	2011	Consumption	Financial	90.000
bed_acute	Beds for those seeking in-patient acute care	AUS	1960	Non-Tech	Other	67000.000
bed_hosp	Beds in hospitals	AFG	1960	Non-Tech	Hospital (non-drug medical)	1677.093
cabletv	Households that subscribe to cable	AFG	1992	Consumption	Communications	0.000
elec_coal	Electricity from coal (TWH)	ABW	2000	Production	Energy	0.000

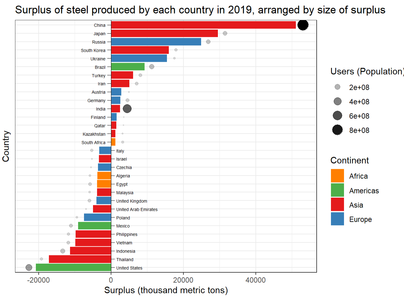
There are 491636 observations and 7 rows in this dataset. The rows are:

1. `variable` : variable name
2. `label` : explanation on what the variable is and what it measures
3. `iso3c` : country code
4. `year`
5. `group` : there are four groups that each variable can be classified as
    - Consumption: technologies that directly increase the consumer's utility
    - Production: technologies involving making goods and services that consumers buy
    - Creation: involves research and development process of technologies
    - Non-tech: not involving technologies
6. `category` : There are 9 categories of the variables, which are Vaccines, Agriculture, Transport, Industry, Financial, Hospital (non-drug medical), Communications, Energy, and Others.
7. `value` : value of the statistics measured

Note that each variable has different periods in which they are available. For example, `railpkm` has observations from year 1834 to 2019. However, `servers` only has observations from the year 2010 onwards.

Furthermore, within each year, not all countries have observations of a variable in a given year. This means that the number of countries observed in the data given a variable differs from year to year. For example, in year 2020, `elec_coal` has 68 different country observations. However, in 1991 it only has 30 different countries.

Plot 1



Insights

This plot depicts the surplus of steel produced by each country in 2019. Surplus can be calculated as follows

$$\text{Surplus} = \texttt{steel\_produced} - \texttt{steel\_demand}$$

A negative surplus means that the country produced less than what it needed, which most likely means the country would have to import steel from other countries. On the other hand, a positive surplus means that the country produced more than it demanded. The surplus could be stored for future use or exported to other countries.

Furthermore, `internetuser` (Number of people with internet access), which is directly correlated to the size of population, is also plotted.

There are some observations that can be gleaned from this plot:

- The absolute magnitude of surplus/deficit tends to be larger in Asian countries (red fill) as compared to that in other continents.
    - Developed East Asian economies (China, Japan, S. Korea) are among countries with the largest surplus of steel production (exporters).
    - Emerging South East Asian economies (Thailand, Indonesia, Vietnam, Philippines, Malaysia) are among countries with the largest deficit of steel production (importers). This is likely because large amounts of steel are needed for construction projects in these developing countries experiencing rapid economic growth, but these countries do not have the required productive capacity and efficient technologies to produce enough steel to sustain their demand.
- Similarly, the magnitude of surplus/deficit tends to be larger in countries where the number of people with internet access (population) is large, such as in China and the US. As the population becomes smaller, the magnitude of surplus/deficit becomes smaller too.
- the US is the country with the largest deficit of steel production, so it is likely to be the largest steel importer.
- China is the country with the largest surplus of steel production, so it is likely to be the biggest exporter of steel. Its surplus is twice than that of Japan's so it is not surprising that the US labelled China as currency manipulator and imposed tariffs on steel.
- It is interesting to note that despite its large population, India has smaller magnitude of surplus compared to smaller countries. This is likely because it is also a developing country.

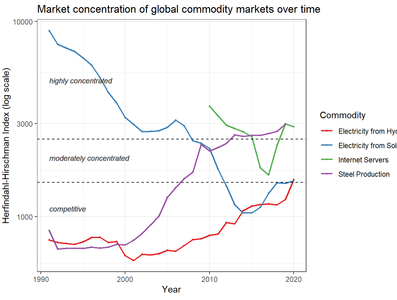
Design choices

`geom_col` was chosen as it can possibly highlight the surplus differences between countries better than line plot or scatter plot. I also removed the y-axis and chose to shift the country labels closer to the bars as doing so makes it easier for the reader to identify which bar belongs to which country.

Regarding the colour palette, "Set1" was chosen from `RColorBrewer` as the difference in hues can help highlight the different continents better. For the `geom_point` , I chose to represent the variable `internetuser` using `alpha` and `size` as doing so can highlight the contrast between population sizes.

This plot is more suitable for a layman as it is easy to compare the sizes of surplus using the height of `geom_col` without any technical knowledge.

Plot 2



Insights

In this plot, we calculated the Herfindahl-Hirschman Index (HHI) for each commodity in the global market to measure market concentration over time. HHI can be defined as

$$\text{HHI} = \sum_{i=1}^{n} s_i^2$$

where s_i is the market share percentage of firm i expressed as a whole number, not a decimal (Source: Investopedia.com). The maximum value of HHI is $100^2 = 10000$ which is only achievable if the market is a pure monopoly.

As HHI increases, market concentration is higher which means there is less competition and can be interpreted as monopoly/oligopoly. Similarly, lower HHI means lower market concentration which leads to higher competition as the market is closer to the perfect competition model. The interpretation of HHI is as follows:

- Highly concentrated: $\text{HHI} \geq 2500$
- Moderately concentrated: $1500 \leq \text{HHI} \leq 2500$
- Competitive: $\text{HHI} \leq 1500$

We would like to measure HHI of global commodities to determine the competitiveness of various markets where each country acts as a firm/supplier. We measure the market share of each country by taking the amount produced divided by global total production. From Plot 2, we can see that market concentrations of various global commodities have significantly changed from over 30 years ago.

- For the market of electricity produced by hydrotechnology, the market still remained competitive in 2019, albeit there has been a slow increase in concentration since the early 2000's. Interestingly, there was a huge jump in market concentration in 2020 so the market is now moderately concentrated.
- For the market of electricity produced by solar technology, the market concentration has drastically decreased from highly concentrated in 1991 to borderline competitive in 2020.
- For the market of servers, the data was only available for 2010 onwards. It experienced a sharp drop in market concentration around the 2015's but market concentration rose up again to highly concentrated in 2020.
- For the market of steel, it experienced a rapid increase in market concentration from 1991 to 2020 (highly concentrated).

As can be seen, market concentration changes differently over time for different types of goods. However, it can be noted that the market concentrations of the different goods are more similar in 2020 than in 1991 where there was a huge gap in market concentrations between Solar Powered Electricity and Steel Production.

Design choices

Line geom is used as it would capture the trend of changing market concentrations better than a scatterplot. Log scale is used for the y-axis as initially the lines were concentrated on the bottom side of the canvas. Regarding the color, "Set1" palette was chosen to differentiate the lines better as other palettes are too light. I also chose to add texts and dashed lines to label the boundaries for different levels of market concentration ratios so that the reader can easily interpret how concentrated a market is without looking at the numbers. Line width and point size were also increased as otherwise they are too thin.

This plot would be appropriate for a technical audience as otherwise it might be difficult to understand what market concentration means and how it can be interpreted.

References

Charles Kenny and George Yang, & Shruti Viswanathan and Michael Pisa. (n.d.). *Technology and development: An exploration of the Data*. Center for Global Development | Ideas to Action. Retrieved November 11, 2022, from https://www.cgdev.org/publication/technology-and-development-exploration-data

Colando, S. (n.d.). *7-19-2022 Tidy Tuesday: Technology Consumption*. RPubs. Retrieved November 11, 2022, from https://rpubs.com/scolando/Tidy-Tuesday-07-19-2022

Rfordatascience. (n.d.). *Tidytuesday/data/2022/2022-07-19 at master · rfordatascience/tidytuesday*. GitHub. Retrieved November 11, 2022, from https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-07-19