IDS 702 Team Project

Part I: Exploratory Data Analysis

Yellow Team - Suzy Anil, Sukhpreet Sahota, Xianchi Zhang, Yuanjing Zhu

2022-10-17

## Data Overview

Upon deciding on the World Inequality Database as our source (<https://wid.world/data/>), our group quickly realized not only its vastness but the variation in data points that are collected to answer key economic and social inequality questions. These key characteristics of the database are important to understand as the database is open-access, compiling valid data from national databases, surveys, fiscal data, and wealth rankings. This combination of data imports is key to ensuring the validity of the available data.

For our project, the key resultant that our group wants to analyze is a nation’s carbon footprint (The Total National CO2 Footprint). A nation’s total carbon footprint is equal to the combination of CO2 footprint and the footprint of other greenhouse gases. This leads to 2 distinct research questions:

1. How do income brackets affect a nation’s carbon footprint? (top 10%, middle 40%, bottom 50%)
2. How have changes in a nation’s average wealth affected their carbon footprint?

During the duration of our study, our main objective is to understand how the different economic status groups affect a nation’s carbon footprint. Our assumptions or hypotheses leading to these questions is that there are no CO2 changes between any of the income brackets and changes within a country’s average wealth over time have not had an impact on it’s respective carbon footprint.

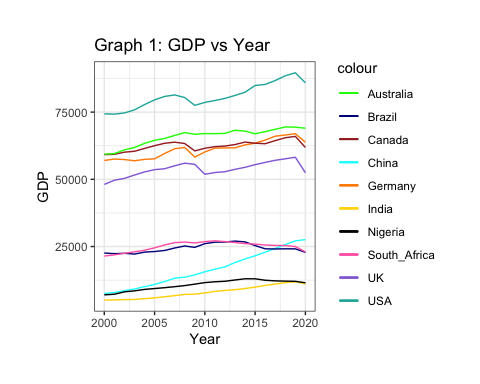
While the dataset/database is vast, we have narrowed to the following key variables that will help analyze demographic statistics of a country:

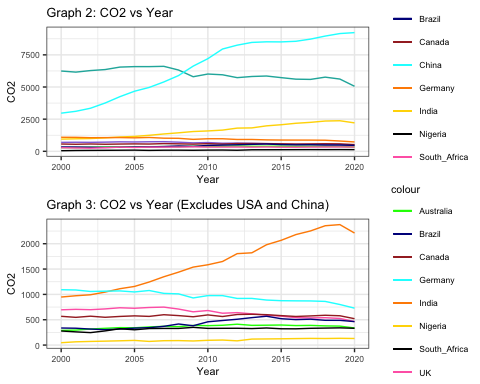
1. The National Income for the Respective Nation - *which according to the World Inequality Database is defined as “National income aims to measure the total income available to the residents of a given country. It is equal to the gross domestic product (the total value of goods and services produced on the territory of a given country during a given year), minus fixed capital used in production processes (e.g. replacement of obsolete machines or maintenance of roads) plus the net foreign income earned by residents in the rest of the world. National income has many limitations. However it is the only income concept that has an internationally agreed definition (established by the United Nations System of National Accounts, see SNA 2008). So we use it as our reference concept (with tax havens correction).”*
2. The respective nation’s Gross Domestic Product
3. The Income Inequality within a nation (as determined by the following income brackets: Top 10%, Middle 40%, and Bottom 50%)
4. The respective nation’s total population
5. The respective nation’s market-value national wealth - *which according to the World Inequality Database is defined as “Net national wealth is the total value of assets (cash, housing, bonds, equities, etc.) owned by the national economy, minus its debts. The national economy - in the national accounts sense - includes all domestic sectors, i.e. all entities that are resident of a given country (in the sense of their economic activity), whether they belong to the private sector, the corporate sector, the governement sector.”*
6. The respective nation’s wealth-to-income ratio (which assesses the net national wealth to the net national income)
7. Years, from the beginning of the century to 2020 (2000 - 2020)
8. Countries

While the world inequality database maintains data for around 300 countries/regions throughout the world, we have narrowed down that total selection/population to a sample size of 10 countries: The United States, China, India, Germany, the United Kingdom, Canada, Australia, Brazil, Nigeria, and South Africa, to effectively analyze and assess these data questions over the 20 year selected period.

*It is important to note: to help standardize the findings for all countries, the US dollar was the currency selected for the appropriate variables*

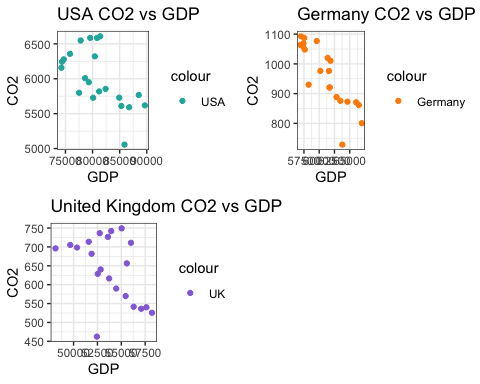
## Primary Relationship of Interest

 From Graph 1, we have visualized the trend of average income per adult GDP over the years 2000-2020. We see general positive linear trends across all countries with USA, Australia and Canada with the top three highest GDP.

 In Graph 2, we visualized the carbon emissions released by each country over the years. The results of the graph show that the USA and China are major contributors of carbon emissions, the other countries are clustered close together towards the bottom of the graph. In order to see the trend among these countries (non-major CO2 contributors), we chose to plot Graph 3 while omitting the two. Most of the countries show a constant trend and some show a slight decrease post-2010, this could be due to reduced coal use and the transition to renewable energy use. The only country that has a contradictory trend is India, which increased in carbon emissions significantly from 2000-2020. India’s main source of energy generation is coal-fired and the growth of renewable slowed over the years which explains the spike after 2012. USA, China and India are known as the three major carbon emitters and our data shows parallel results.



Correlation Matrix



From the correlation matrix, we see that all the countries have a generally strong positive correlation between carbon emissions and average wealth. However, USA, Germany and UK have negative correlations. The three graphs above model the relationship between carbon emissions and GDP for those countries and we see the same relationship as we do with wealth. There is not a very strong compared to other countries (see Appendix) but there is an overall negative trend. Based on the similar behavior between GDP and wealth with carbon emissions, we will dive deeper into income brackets and average wealth in our model to infer how significant this relationship is.

## Other Characteristics

As mentioned within our data overview, the World Inequality Database is an extensive. For example, the database is comprised of data points within 7 distinct categories:

1. Average and Total Income - This category provides a breakdown of the national and personal income and savings
2. Average and Total Wealth - This category provides a breakdown of the national and personal wealth by assessing assets, liabilities, market valuation of companies, etc.
3. Income Inequality - The income inequality measures the national and personal income/savings between different ranges and can be customized based on the percent range (i.e. top 10% (from 90-100%), etc.)
4. Wealth Inequality - Similar to income, the wealth inequality measures the total value of assets, liabilities, etc. over different ranges and can be customized based on the percent range (i.e. top 10% (from 90-100%), etc.)
5. Carbon Macro and Average - This category assesses carbon emissions and provides breakdowns by national, household, imports, and other territorial groups.
6. Carbon Inequality - The carbon inequality measures the national carbon footprint between different ranges and can be customized based on the percent range (i.e. top 10% (from 90-100%), etc.)
7. Other data variables available on World Inequality Database - A category that maintains other demographic and financial data points, from market exchange rates and taxes to population size and employed population

From these broad categories, we subset to our variables to answer our two research questions. This consisted of analyzing aggregated views of a nation’s income, wealth, carbon footprint, and population size. To see changes over time, we have selected a 20 year timeline (from 2000 - 2020), which, in our opinion, should be a long enough time frame to demonstrate any social or economic changes. Lastly, in selecting our sample size of 10 countries, we wanted to assess the impacts/research questions geographically across the world. For this reason, we selected up to 2 of the largest countries in terms of size, population, income, and other demographic means per continent in order to standardize and have a better understanding throughout the globe.

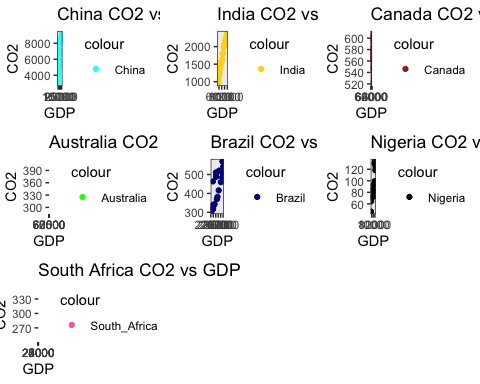
## Potential Challenges

While the World Inequality Database is a great resource and maintains great data, it also has many limitations and presents many challenges. Three of the key constraints we ran into were:

1. The inability to analyze data beyond and/or within our selected 20 year time frame. This was seen from both a country aspect as well as variable aspect. Our decided 20 year window was optimal for most countries and most variables. For example, during our exploratory phase, we found that even for several countries (22 countries to be exact), they did not have basic national income and GDP information. It was important for us to remove these countries that didn’t have data within our 20 year window as to not skew/impact our understanding. It can be assumed that for smaller countries, some key data points may not be available due to limited resources. In addition, we found that as the World Inequality Database has become more widely used for research purposes, data collection of interesting data points varies and is only available and can only go as far back based on the sources/when the variable was created. While we may prefer to garner a more significant understanding/trend, we are unable to based on this limitation.
2. The standardization of the data across countries due to multiple factors: population size, currency, etc. While we have tried to control for these factors by including population as a variable and normalizing all financial data to US dollars, this is an important limitation to outline now and remember for future model creation, etc.
3. The resistance to include more variables was due to many variables being dependent on the primary independent variables resulting in collinearity in our model. For example, wealth income is made up of assets and debt so we were unable to include those in our analysis, rather we chose to include the aggregate feature for our model to be more well-rounded.

## Appendix

### CO2 vs GDP graphs (cont.)



### Descriptive Statistics

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | year | National\_income\_USA | GDP\_USA | CO2\_USA |
|  | Min. :2000 | Min. :64323 | Min. :74252 | Min. :5057 |
|  | 1st Qu.:2005 | 1st Qu.:66668 | 1st Qu.:77806 | 1st Qu.:5729 |
|  | Median :2010 | Median :69570 | Median :80419 | Median :5950 |
|  | Mean :2010 | Mean :69369 | Mean :80921 | Mean :6010 |
|  | 3rd Qu.:2015 | 3rd Qu.:72530 | 3rd Qu.:84920 | 3rd Qu.:6321 |
|  | Max. :2020 | Max. :76075 | Max. :89639 | Max. :6608 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_China | GDP\_China | CO2\_China |
|  | Min. : 6259 | Min. : 7533 | Min. :2966 |
|  | 1st Qu.: 9138 | 1st Qu.:10893 | 1st Qu.:4672 |
|  | Median :13275 | Median :15618 | Median :7202 |
|  | Mean :13957 | Mean :16456 | Mean :6595 |
|  | 3rd Qu.:18483 | 3rd Qu.:21571 | 3rd Qu.:8514 |
|  | Max. :23087 | Max. :27578 | Max. :9229 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_India | GDP\_India | CO2\_India |
|  | Min. : 4552 | Min. : 5052 | Min. : 947.9 |
|  | 1st Qu.: 5283 | 1st Qu.: 5913 | 1st Qu.:1157.1 |
|  | Median : 6855 | Median : 7737 | Median :1584.6 |
|  | Mean : 7119 | Mean : 8046 | Mean :1622.9 |
|  | 3rd Qu.: 8647 | 3rd Qu.: 9915 | 3rd Qu.:2067.3 |
|  | Max. :10503 | Max. :11895 | Max. :2377.1 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_Germany | GDP\_Germany | CO2\_Germany |
|  | Min. :48168 | Min. :56905 | Min. : 728.6 |
|  | 1st Qu.:49847 | 1st Qu.:57632 | 1st Qu.: 875.8 |
|  | Median :53318 | Median :61601 | Median : 976.1 |
|  | Mean :52663 | Mean :61162 | Mean : 959.5 |
|  | 3rd Qu.:54640 | 3rd Qu.:63462 | 3rd Qu.:1057.7 |
|  | Max. :57716 | Max. :66973 | Max. :1092.7 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_United\_Kingdom | GDP\_United\_Kingdom | CO2\_United\_Kingdom |
|  | Min. :43076 | Min. :48108 | Min. :462.5 |
|  | 1st Qu.:46503 | 1st Qu.:52440 | 1st Qu.:569.8 |
|  | Median :47626 | Median :53726 | Median :656.5 |
|  | Mean :47556 | Mean :53771 | Mean :641.3 |
|  | 3rd Qu.:49268 | 3rd Qu.:55579 | 3rd Qu.:710.9 |
|  | Max. :50901 | Max. :58218 | Max. :749.2 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_Canada | GDP\_Canada | CO2\_Canada |
|  | Min. :50203 | Min. :59203 | Min. :520.7 |
|  | 1st Qu.:52204 | 1st Qu.:61487 | 1st Qu.:564.5 |
|  | Median :53779 | Median :62511 | Median :575.4 |
|  | Mean :53544 | Mean :62450 | Mean :574.3 |
|  | 3rd Qu.:55098 | 3rd Qu.:63435 | 3rd Qu.:588.4 |
|  | Max. :56411 | Max. :65977 | Max. :608.3 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_Australia | GDP\_Australia | CO2\_Australia |
|  | Min. :49197 | Min. :59361 | Min. :290.0 |
|  | 1st Qu.:53336 | 1st Qu.:64499 | 1st Qu.:341.3 |
|  | Median :54886 | Median :66983 | Median :374.4 |
|  | Mean :54137 | Mean :65894 | Mean :363.2 |
|  | 3rd Qu.:55740 | 3rd Qu.:67929 | 3rd Qu.:388.6 |
|  | Max. :56479 | Max. :69495 | Max. :412.0 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_Brazil | GDP\_Brazil | CO2\_Brazil |
|  | Min. :18859 | Min. :22191 | Min. :304.2 |
|  | 1st Qu.:19574 | 1st Qu.:22924 | 1st Qu.:336.8 |
|  | Median :20612 | Median :24150 | Median :459.5 |
|  | Mean :20803 | Mean :24353 | Mean :428.2 |
|  | 3rd Qu.:21701 | 3rd Qu.:25356 | 3rd Qu.:502.9 |
|  | Max. :23411 | Max. :27024 | Max. :569.2 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_Nigeria | GDP\_Nigeria | CO2\_Nigeria |
|  | Min. : 6475 | Min. : 7025 | Min. : 46.76 |
|  | 1st Qu.: 8752 | 1st Qu.: 9374 | 1st Qu.: 81.08 |
|  | Median :10450 | Median :11509 | Median : 92.18 |
|  | Mean : 9880 | Mean :10718 | Mean : 98.12 |
|  | 3rd Qu.:11329 | 3rd Qu.:12127 | 3rd Qu.:124.26 |
|  | Max. :12032 | Max. :13003 | Max. :133.63 |

Exploratory Analysis of National Income, GDP, and CO2 for Countries and Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | National\_income\_South\_Africa | GDP\_South\_Africa | CO2\_South\_Africa |
|  | Min. :18132 | Min. :21449 | Min. :245.9 |
|  | 1st Qu.:20314 | 1st Qu.:23617 | 1st Qu.:320.9 |
|  | Median :21278 | Median :25577 | Median :328.3 |
|  | Mean :21061 | Mean :25032 | Mean :317.1 |
|  | 3rd Qu.:22179 | 3rd Qu.:26459 | 3rd Qu.:331.5 |
|  | Max. :22939 | Max. :27121 | Max. :349.8 |