25847333_Project_1_Report

Project 1: Global Development Indicators Report

Introduction

Global Development Indicators (GDI's) are statistical measures that assist countries in measuring their development and improvements across different aspects over time. GDI's are often measured and recorded by international companies and organizations such as the World Bank. The data that we will be looking at throughout this report is retrieved from the World Bank Development Indicators service. Throughout this report, the following questions will help us analyse the different Global Development Indicators:

- What is the distribution of government expenditure on education?
- What are the trends in electric power consumption over the past three years?
- How does internet usage vary across GDP levels for the top three economies?
- How has the estimated level of corruption control changed over the years?
- What is the relationship between population growth and urban population?

Data Cleaning and Preparation

According to Wickham, Çetinkaya-Rundel and Grolemund, there are 3 rules to ensure that a data set is tidy. The first rule being that each variable has its own column, and each column has its own variable. The second rule is that each observation is a row, and each row is an observation. Lastly, each value has a cell, and each cell has a single value (74).

The Global Development Data Tidying Process

- When taking a quick glance at the provided dataset, we can clearly see that the first step to tidying this dataset would be to make use of the pivot function. Pivot_longer would joins columns together in order to make it neater. In our dataset we can combine the different years into one column called the "Year" column. I have also created a new column called Values where I have stored all the values in.
- However, we can clearly see that the Value column contains many missing values that have been replaced by 2 full stops (..). This is not useful to our analysis and therefore needs to be removed. However, in order to remove the full stops we first need to replace them with Na to convert them to explicit missing values. I have made use of mutate and the na_if function to convert the 2 full stops(..) to NA values.
- Thirdly, in order to remove the values that are NA, I have filtered my dataset to only include the values that are NOT NA using the(!(is.na)).
- When looking at our new 'Year' column that we have made using the pivot longer function we can see that it contains the 4-digit year followed by square brackets with numbers inside of them. However, having both makes the dataset look untidy. I have used the mutate function along with the str_extract function in order to extract the part of the string that contains 4 digits which represent the year. This gets rid of all other brackets and letters included in the column.
- In the next code chunk, I have used the mutate function to convert both the 'Year' and the 'Value' column to numeric. This allows us to perform any mathematical equations on the dataset. In addition to this, the values in the Value column are in scientific notation. In order to remove the scientific notation, I have used the format function to ensure that the scientific notation is set to FALSE.
- In the next step I have renamed the column names to remove the space that appears in the existing column names. This just makes it more convenient when working with the dataset as it avoids having to use back ticks (") every time we call or use the column names.
- Lastly, I had to convert the Value column to numeric again. After changing it from a character to numeric I arranged the year column in ascending order.

Transforming the Data Process

The Distribution of Government Expenditure on Education

- In order to answer this question, we first need to create a database which focuses only on the series_name that we are interested in. For this specific question we want to focus on expenditure on Education. In order to do this, I have filtered the new dataset to include series_names that are included in the vector. This includes primary, secondary and tertiary education.

- Secondly, I used the mutate function to calculate the mean for each of the education categories. I used na.rm to ignore any NA values that may mess up my analysis. I then used the summarise function to only include the series_name and the new average for each category.

Electric Power Consumption Process

- Firstly, I created a new dataset that includes only the data about electric power consumption. This was done by filtering the series_name. I also filtered between the years 2000 and 2020. I wanted to conduct my analysis over a period of 20 years.
- In order to compare the countries and who has the highest electric power consumption I need to calculate the average for every country. This is done by using summarize function and calculating the mean for the value column. I have also gotten rid of all NA values using na.rm function. After completing these steps, I ungrouped the group by function from the second line of code.
- Lastly, I want to select the top 5 countries with the highest average consumption. I arranged the column from biggest to smallest and used the slice_head function to select the top 5 countries with the highest average.

Internet Usage Process vs GDP

- The first step was to create a new dataset for this section. I filtered the data to only include information about the Individuals using the Internet. I then selected which columns I wanted to appear in my dataset. Finally, I renamed the value column to be more descriptive.
- In the second code chunk, we did the exact same steps to retrieve the necessary data about GDP. I used the filter function to select on data about GDP and then renamed the column from value to GDP.
- In order to work with both datasets, I had to join them together using the inner join function. I joined them together by country name and the year. This creates a new dataset that includes both of the above datasets that are crucial for our analysis.
- After joining both datasets, I had to find the average (mean) of GDP for each country. In order to ensure no outliers such as NA values incorrectly skewed my analysis, I had to ensure that all NA values were ignored. This was done using the na.rm function. After

working out the mean GDP I arranged the values in descending order to be able to select the top 3 countries with the highest GDP.

- In the next code chunk, I have filtered the dataset to include the rows for the top 3 countries with the highest GDP. The %in% function checks to see if the country name is in the combined dataset.
- Lastly, I have created a basic mathematical function to convert internet usage to a percentage. I did this by multiplying by 100.

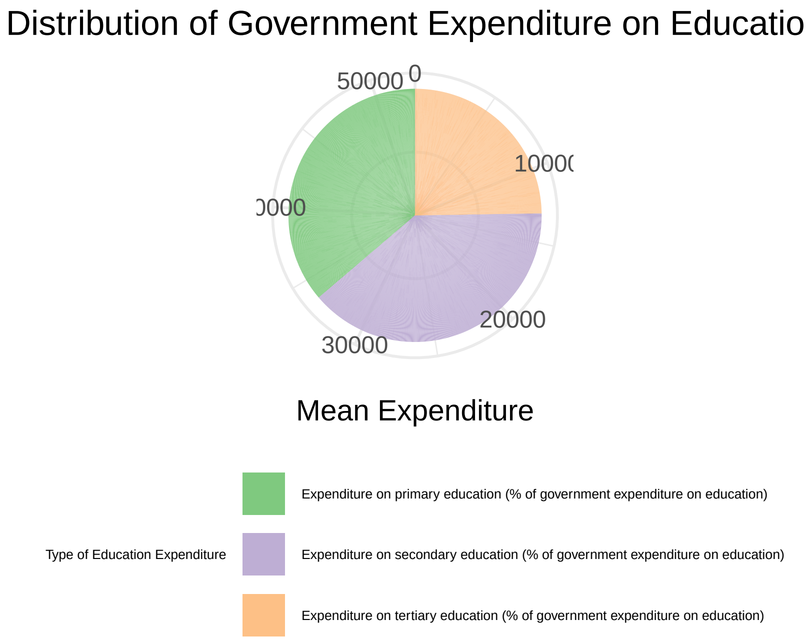
Corruption Control Process

- The first step was to create a new dataset that was filtered to only show the series_name that were called 'Control of Corruption'. I also filtered the dataset too only include rows that were over the last 20 years. This includes from 2002 to 2022. The rename function allows me to rename the value column.
- To get rid of all NA values I filtered the data where the rows were NOT equal to NA.
- I then filtered the data to only show the values that were equal to or greater than 0 to get rid of any outliers that would affect my data analysis.

Population Growth vs Urban Population Process

- The first step was to create a new dataset that filtered series_name to population growth. I only wanted the columns that ere the country's name, year and value. I then renamed the value column to population growth to provide more clarity about the column.
- I then had to do the exact same steps for the next dataset that I created. I filtered where series_name was equal to Urban Population. I also renamed the value column to a more descriptive/ informative name.
- In order to use both datasets I needed to join them together using an inner join. I joined them by the country name and year columns.
- Eventually I had to work out the average for population growth. I did this using the mean function. After calculating the averages, I arranged them in descending order to be able to select the top 3 countries with the highest population growth average.
- Lastly, I filtered the dataset to include the top 3 countries with the highest population growth average.

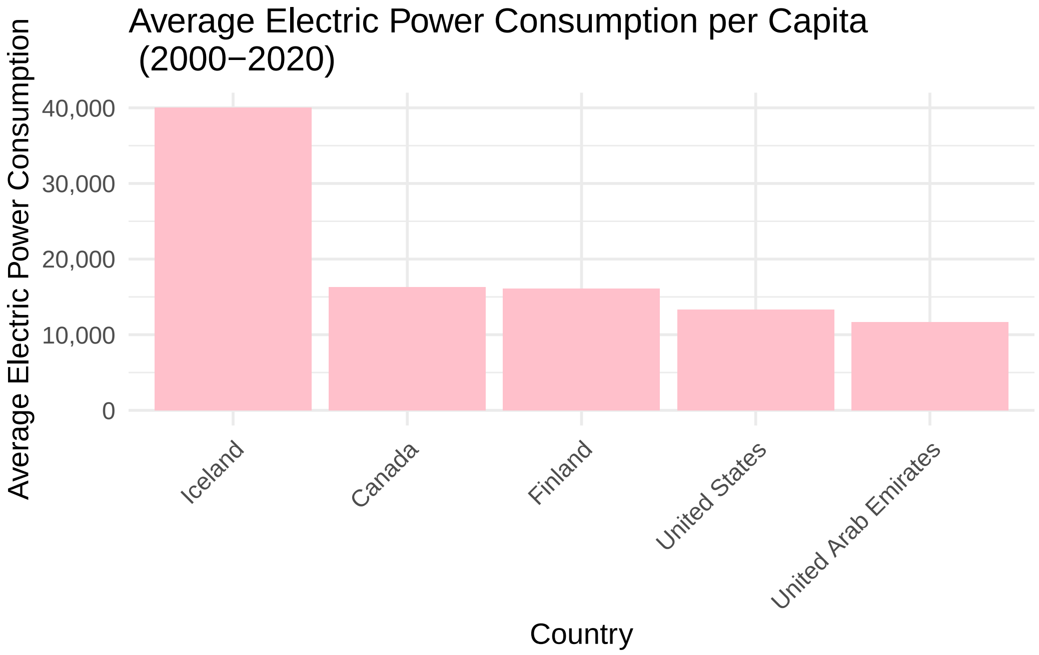
Distribution of Government Expenditure on Education



This visualisation is a bar chart that displays the distribution of government expenditure on the different types of education. We can clearly see that the government spends the most amount of money on the secondary education sector. The second most funded sector is the primary sector. This could indicate a specific hierarchy between the education categories. It could indicate that secondary education is deemed the most important and primary education is deemed the second most important.

Our statistical summaries seem to be a good indication of how the government distributes its funding between the 3 education categories. We can see that the average expenditure for secondary education is at 36,13%, while the average expenditure for the primary sector is slightly lower at 31,70& and the average expenditure for tertiary is only at 20,21%.

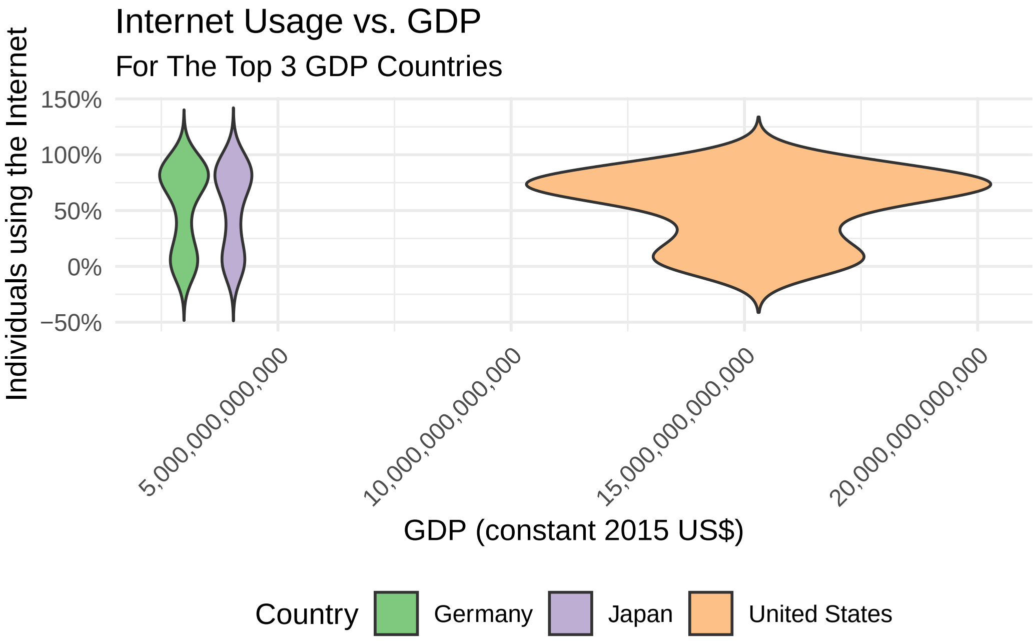
Electric Power Consumption Per Capita



The above graph is a bar graph showing the top 5 countries and their average electric power consumption. The graph shows data ranging from 2000 to 2022. We can see that there is a trend for the average estimates of electricity consumption. More developed and industrialised countries typically consume more electricity on average because of factors like population density, and colder/warmer climates that require significant heating or cooling.

When looking at the statistical summaries we can see that the average amount of electricity consumed per person per year is about 5134,841 kWh. The standard deviation is almost 7,430 kWh, this suggests that there is a lot of variation in the amount of power consumed in the countries.

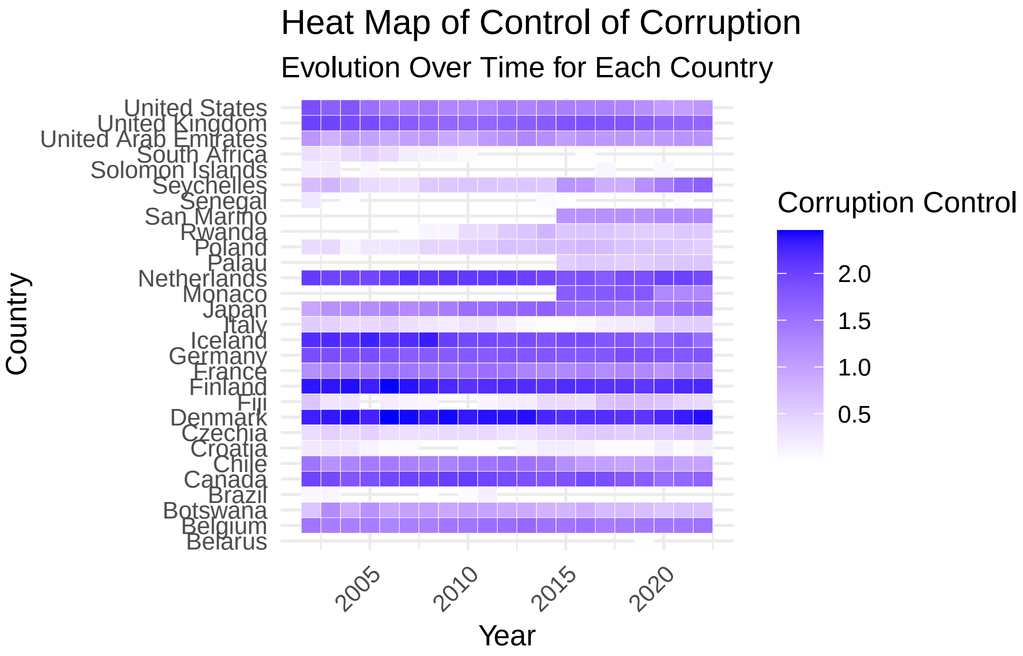
Internet Usage Across GDP Levels for the Top 3 Economies



The above graph is a violin plot that shows the relationship between the number of individuals who use the internet and the GDP of a country. We can see that United States of America has the highest GDP. They also have the widest section on the violin plot indicating that the higher the GDP the more individuals use the internet This suggests that as countries become wealthier, they are likely to have more individuals having access to resources like the internet.

When looking at the statistical summaries for this data we can see that the median internet usage for each country is typically high. This could suggest that the internet usage for each country is increasing as the years go on.

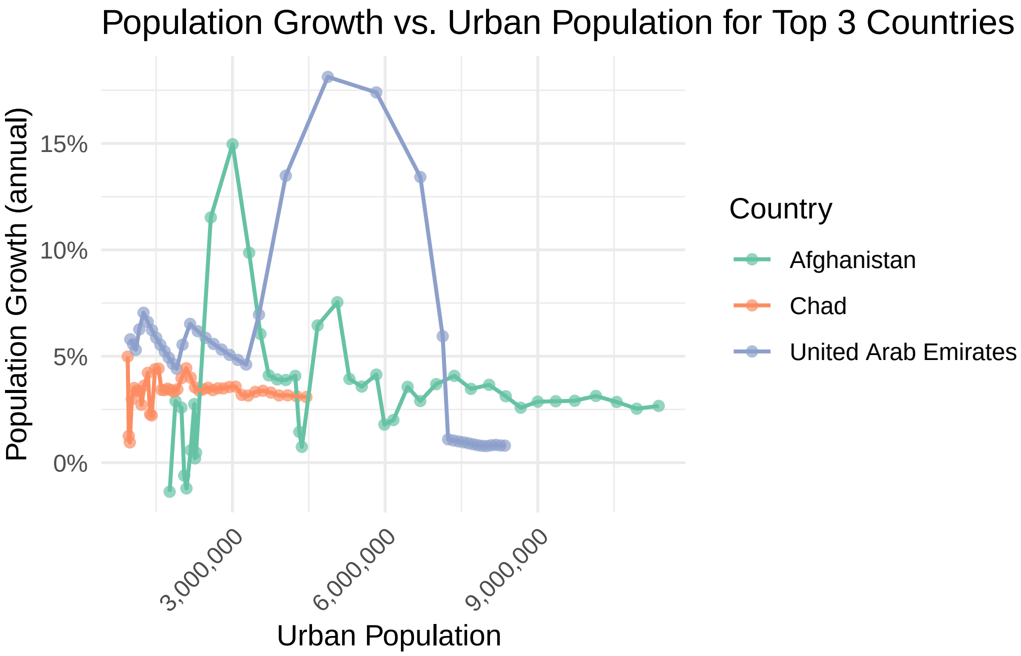
Estimated Level of Corruption Control



This visualization is a heat map that represents each country's control of corruption over time. The darker the shades on the map indicate the higher control of corruption. By analysing this graph, we can clearly see that countries such as Finland and Denmark have high control over corruption while countries like south Africa and Brazil have little to no control of corruption.

Our statistical summaries give us a range of values which highlight the variety of levels of governance effectiveness across these countries from 2000 to 2020.

Population Growth vs. Urban Population



The above visualization is a line graph that represents the relationship between population growth and urban population for the top 3 countries by average population growth. The graph shows that when urban population is low population growth is generally higher than normal.

Narrative

The main objective of my analysis is to conduct an exploratory analysis of global development indicators. This was done by examining relationships between, government expenditure and education, electric power consumption, internet usage vs GDP levels, corruption control over time, and the relationship between population growth and urban population.

Over the years, there seems to be an ongoing trend with government expenditure when it comes to the education sectors. The governments spend most money on the primary sector indicating that these years are some of the most important years of a person's life. Electric power consumption seems to be much higher for those countries that are more developed than others. Trends show that those countries who have extreme weather conditions tend to use more electricity. For example, the use of heaters and air conditioners. There is a strong relationship between the internet usage and the country's GDP. The number of individuals who use the internet is greatly impacted by the country's wealth and access to resources. Corruption control tends to be much lower in countries that have a low GDP. Lastly, there seems to be an inverse relationship between urban population size and population growth. This analysis shows that growth tends to be higher in less urbanized regions.

Key Takeaways:

- Secondary education receives the highest government funding.
- More developed and industrialized countries tend to use more electrical power.
- Wealthier countries tend to have higher internet usage.
- Poorer countries have less control of corruption.
- There is an inverse relationship between urban population and population growth

Implications

- Secondary education receiving more funding than the other sectors imply that it is prioritized over other education sectors due to its importance in developing knowledge and skills.
- The consequence is that countries with lower incomes could have a harder time building solid government structures and maintaining accountability, which could lead to a weaker control over corruption.

Limitations

- A limitation of my study could be interpretational bias. How I have interpreted the data, and analysis could be completely different to how someone else could interpret my findings. This could potentially lead to conclusions; however, I have included strong summaries about my findings in an attempt to avoid this.
- Another limitation for my analysis were the number of missing values and outliers in my data. However, this was easy to handle as i just removed them to ensure my data analysis was not incorrectly interpreted.