Project Report

2024-04-07

Project 1 - Project Report

Project 1 for the course COMP4010 - Data Visualization

Group member

Name	Student ID	E-mail	Title
Hoang Trung Thanh	V202100516	21thanh.ht@vinuni.edu.vn	Project Manager
Tran Tue Nhi	V202000079	20nhi.tt@vinuni.edu.vn	Data Analyst
Nguyen Minh Tuan	V202000254	20tuan.nm@vinuni.edu.vn	Data Analyst

Introduction

The dataset that we choose is the dataset taken from Our World in Data (OWID). The dataset contains all the important informations about the energy and electric generation/consumption of a country, as well as some important metrics such as GDP, total CO2 emissions, etc. Each row in the table corresponding to a country in a specific year, with the content of the row containing all the information above.

This dataset is important, as it tells the overall picture about the past and current situation of energy and electricity in many countries in the world, which can give us insights about the effect of it on environmental problems such as air pollution, global warming, etc. It can also tells us the development of a country, and the sustainability of such development (is renewable energy in development ? what is the emission per capita ? and so on). One of the big problem with this dataset is the abudance of N/A entries. There are a lot of data missing, and on a variety of columns. As such, it is our job to clean the data and extract all the important available data, as well as fill in the missing data from outside if necessary.

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)
```

```r
library(showtext)
```

```
## Loading required package: sysfonts
```

```
## Loading required package: showtextdb
```

```r
font_add_google("Open Sans", "Open Sans")
showtext_auto()
```

```r
df <- read.csv("owid-energy.csv")
```

Question 1

- **Question**: What are the patterns and comparisons in electricity generation among the top 5 countries by energy consumption/electricity generation/GDP and globally in 2023? How do these patterns differ between developed and developing countries or across continents?

We intended to answer the question for year 2023, as 2023 is thought to be the most recent year which have available data for patterns of electricity generation for each country. In this question, we primary needs the data about the share of each type of electricity generation in a country, which is encoded as "x_share_elec" in the table. We want to answer the above question because it can give us insights about what is the different between practices in the top countries in GDP, as well as the difference between low-income countries and high-income countries with regard to this problem.

However, due to the lack of data in 2023, we can only answer the question for the most recent year which have available data, which is 2021. As such, we will choose the data in 2021 in this question.

In the first aspect of the question, we will plot the constitution of energy generation of 5 countries with the highest GDP worldwide. However, due to our lack of information about GDP in the dataset, we have to get 5 countries in the Internet. From our research, 5 countries with the highest GDP are China, the US, India, Japan and Germany.

```r
df_2021 = df[df$year == 2021, ]

df_2021_top_5 <- df_2021[df_2021$country == "China" | df_2021$country == "India" | df_2021$country == "
```

From that, we will obtain the important metrics for each country, which are the share of each type of electricity generation.

```r
df_2021_top_5_important_metric <- df_2021_top_5[,c("country", "population", "gdp", "biofuel_share_elec"
```

```r
df_2021_top_5_important_metric
```

```
##              country population gdp biofuel_share_elec coal_share_elec
## 4349           China 1425893504  NA              2.003          62.932
## 7731         Germany   83408560  NA              8.057          28.253
## 9287           India 1407563904  NA              2.070          74.173
## 10185          Japan  124612528  NA              3.851          32.510
## 20674 United States  336997632  NA              1.307          21.624
##       fossil_share_elec gas_share_elec hydro_share_elec low_carbon_share_elec
## 4349             66.289          3.213           15.323                33.711
## 7731             48.487         16.354            3.377                51.513
## 9287             78.053          3.745            9.356                21.947
## 10185            71.002         35.119            8.256                28.998
## 20674            60.509         38.037            5.936                39.491
##       nuclear_share_elec oil_share_elec other_renewables_share_elec
## 4349               4.803          0.144                       2.003
## 7731              11.873          3.880                       8.098
## 9287               2.563          0.135                       2.070
## 10185              6.387          3.373                       4.166
## 20674             18.742          0.848                       1.746
##       renewables_share_elec solar_share_elec wind_share_elec
## 4349                 28.908            3.854           7.727
## 7731                 39.640            8.474          19.691
## 9287                 19.384            3.986           3.973
## 10185                22.611            9.254           0.935
## 20674                20.750            3.960           9.108
```

However, entries such as "renewables_share_elec" can be overlapping with other types of eletricity generation, so we hypothesize that the electricity generation types which are the most common are: Biofuel, Coal, Gas, Hydro, Nuclear, Oil, Solar and Wind. We test if our assumption is right by add them together.

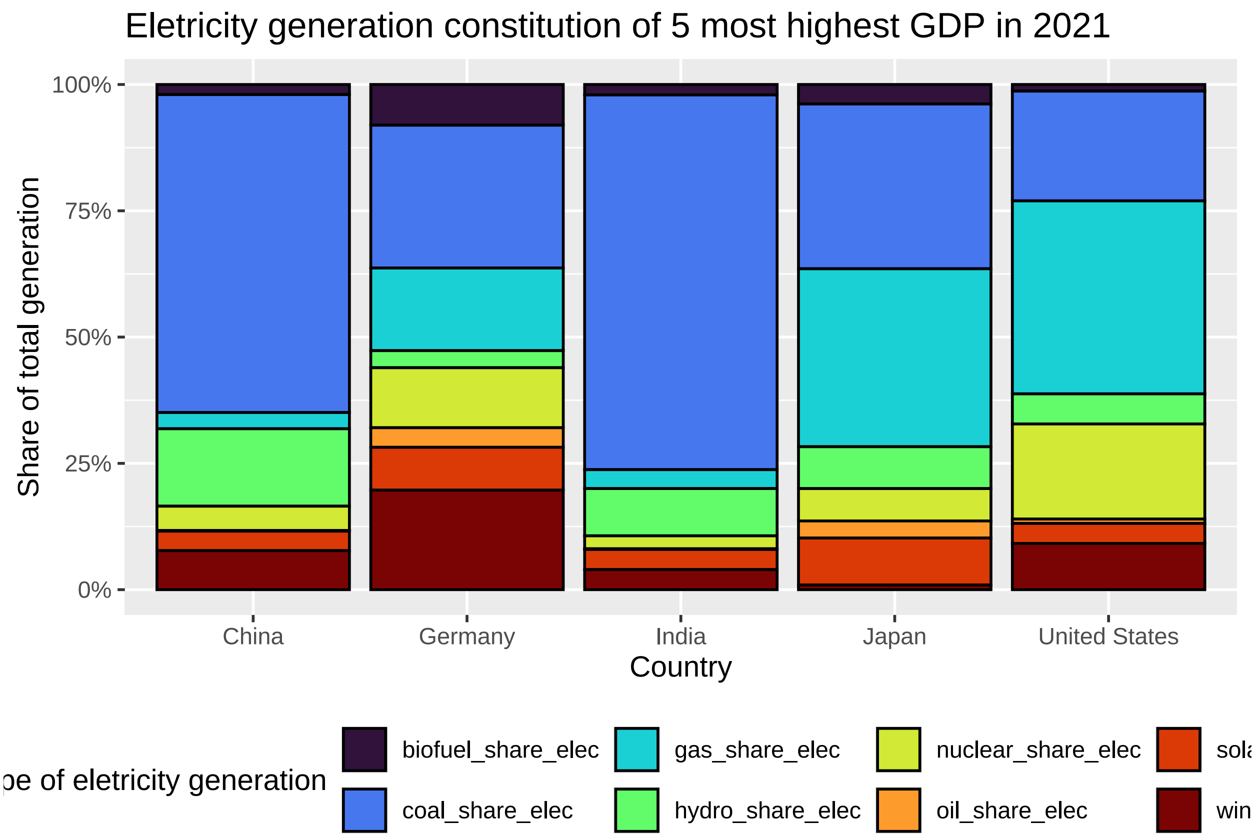
```
df_2021_top_5_important_metric$total = df_2021_top_5_important_metric$biofuel_share_elec + df_2021_top_5
df_2021_top_5_important_metric$gas_share_elec +
df_2021_top_5_important_metric$hydro_share_elec +
df_2021_top_5_important_metric$nuclear_share_elec +
df_2021_top_5_important_metric$oil_share_elec +
df_2021_top_5_important_metric$solar_share_elec +
df_2021_top_5_important_metric$wind_share_elec
```

We see that, the results are approximately 100%, so other kinds of electricity generation is there, but not too much.

```
df_2021_top_5_important_metric <- pivot_longer(df_2021_top_5_important_metric, cols = c("biofuel_share_e
```

In this question, to illustrate the constitution of 5 highest GDP countries, we will use the percent stacked bar chart.

```
ggplot(df_2021_top_5_important_metric, aes(fill=eletric_share_type, y=percentage, x=country)) +
  geom_bar(position="fill", stat="identity", color = "black") +
  scale_fill_viridis_d(option = "turbo") +
  theme(legend.position = "bottom") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Eletricity generation constitution of 5 most highest GDP in 2021", x = "Country", y = "S
```



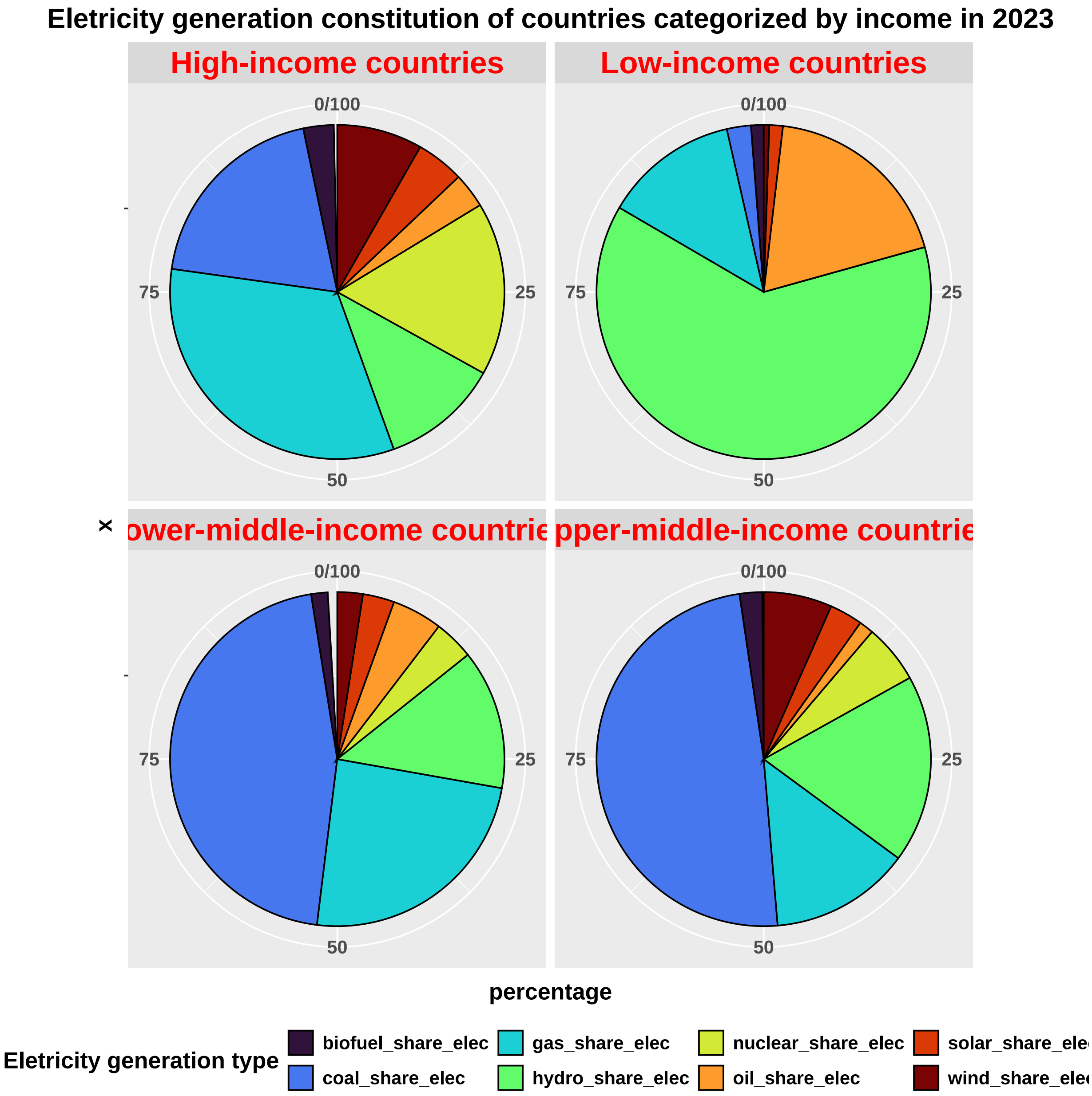
We can see that, there is a wide difference between the constitution of electricity generation in 5 countries above, for example, China and India tends to use more coal as the means of generating electricity, whereas Japan and the US favor gas as the main resource for electricity generation. Overall, we cannot tell a general trend from just the graph above. We need more information.

We suppose that the electrical generation constitution of a country correlates with the development stage of a country. We try to plot 4 pie charts corresponding to: High-income countries, Low-income countries, Lower-middle-income countries, Higher-middle-income countries to see if there is some correlation between 5 countries above and the pie charts generated.

```
df_2021_categorize <- df_2021[df_2021$country == "High-income countries" | df_2021$country == "Low-incom
```

```
df_2021_categorize_important_metric <- pivot_longer(df_2021_categorize, cols = c("biofuel_share_elec",

ggplot(df_2021_categorize_important_metric, aes(x = "", y=percentage, fill=eletric_share_type)) +geom_ba
  scale_fill_viridis_d(option = "turbo") +
  facet_wrap(.~ country,  nrow = 2) +
  theme(legend.position = "bottom", plot.title = element_text(hjust = .5), text=element_text(face="bold
  coord_polar("y", start=0) +
  labs(title = "Eletricity generation constitution of countries categorized by income in 2023", fill =
```



Discussion

- We can see that there is little correlation between 5 top countries regarding to GDP. Each have its own energy constitution distinct from each other.
- However, there is a significant correlation between the electricity generation of those countries with the electricity generation constitution separated by income. China and India have its electricity majorly produced by using coal primary and hydroelectric secondary, which directly relates to country which have lower-middle income or upper-middle-income, while Germany, Japan and the US tends to have a more varied constitution, which mostly includes coal, wind, gas and nuclear electricity generation.
- From that results, we can postulate that the electricity generation constitution mostly depends on the development stage of that country, in which one of the criteria is the GDP per capita. More research is needed, although.

Question 2

- **Question**: How has Vietnam's energy generation evolved over the past decade, particularly in terms of renewable versus non-renewable sources? How does Vietnam's energy consumption and efficiency compare to regional and global averages?

To answer this question, we intend to utilize an animated Treemap that aims to succinctly visualize the evolution of Vietnam's energy generation over the past decade, highlighting the shift from non-renewable to renewable sources alongside global averages. By representing four key data points—Vietnam's fossil and renewable energy, and the global averages for each—this visualization provides a clear comparison of Vietnam's energy transition and its standing on the global stage. This efficient and engaging approach offers insights into Vietnam's progress and challenges in adopting sustainable energy, facilitating a better understanding of its energy dynamics in relation to worldwide trends.

In the dataset, we already have the fossil electricity and renewable electricity data in Vietnam. However, we don't have the global averages of the aforementioned data. Therefore, we have to write a Python script to calculate those data. Moreover, as this data contains so much noise and null data, we have to preprocess the data using a Python script, which is already committed to our GitHub repo.

Importantly, instead of solving this question using R, I have discussed it with Professor Dung and got his approval to use other programming languages to tackle this problem. This question is visualized using Javascript, D3.js, and a cloud service provider called Observable. The link to our Observable workspace is here: https://observablehq.com/d/72ba4dd430929194

Data Preprocessing:

To successfully produce an animated Treemap, we have to preprocess the data first. Here are the steps that we employ to clean our data:

- Import CSV file
- Calculate average values of global renewable and fossil electricity for each year
- Remove all the data fields, except the Vietnam renewable electricity, Vietnam fossil electricity, global renewable & fossil electricity
- Filter the year from 1985 to 2021.
- Transpose columns to rows
- Put all values in a quote
- Return data in TSV format

Data Labeling:

An important aspect of Treemap is that we have to label the regions for the data that is visualized. Each different region will have a different color on our chart. Since the regions in Q2 are not many, we decided to manually label them.

Analysis:

Observable notebook: https://observablehq.com/d/72ba4dd430929194

The source code is in the notebook above. In this section, I only go through the implementation result.

The result is a treemap that can represent 4 different data: - Fossil electricity in Vietnam - Renewable electricity in Vietnam - Global average fossil electricity - Global average renewable electricity

There is a button and a slider on the top left of the visualization, which can be used to animate the Treemap. Once that button is hit, the Treemap will animate and grow bigger / smaller to show the changes in energy data through the years

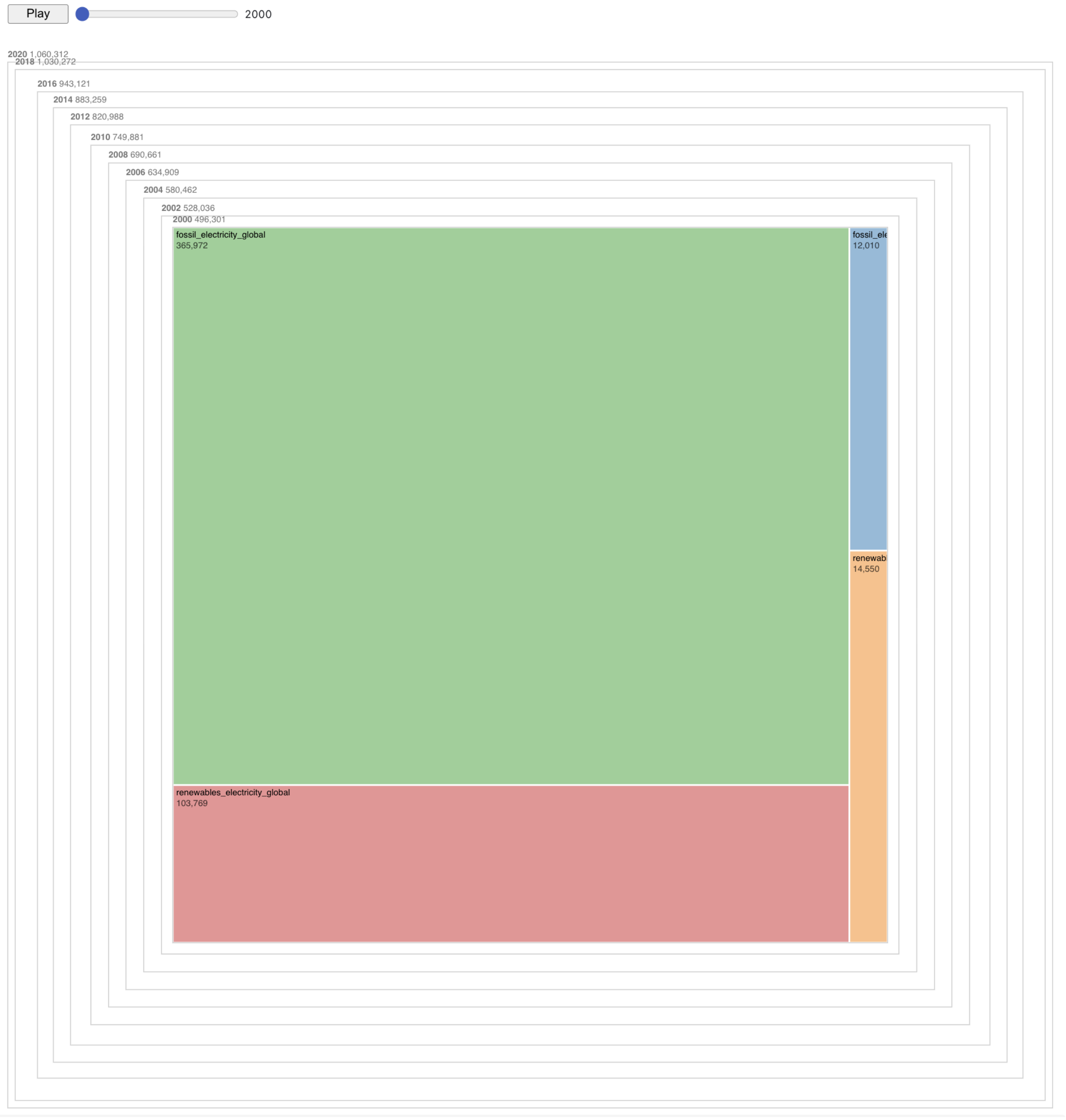


Figure 1: Treemap in 2000

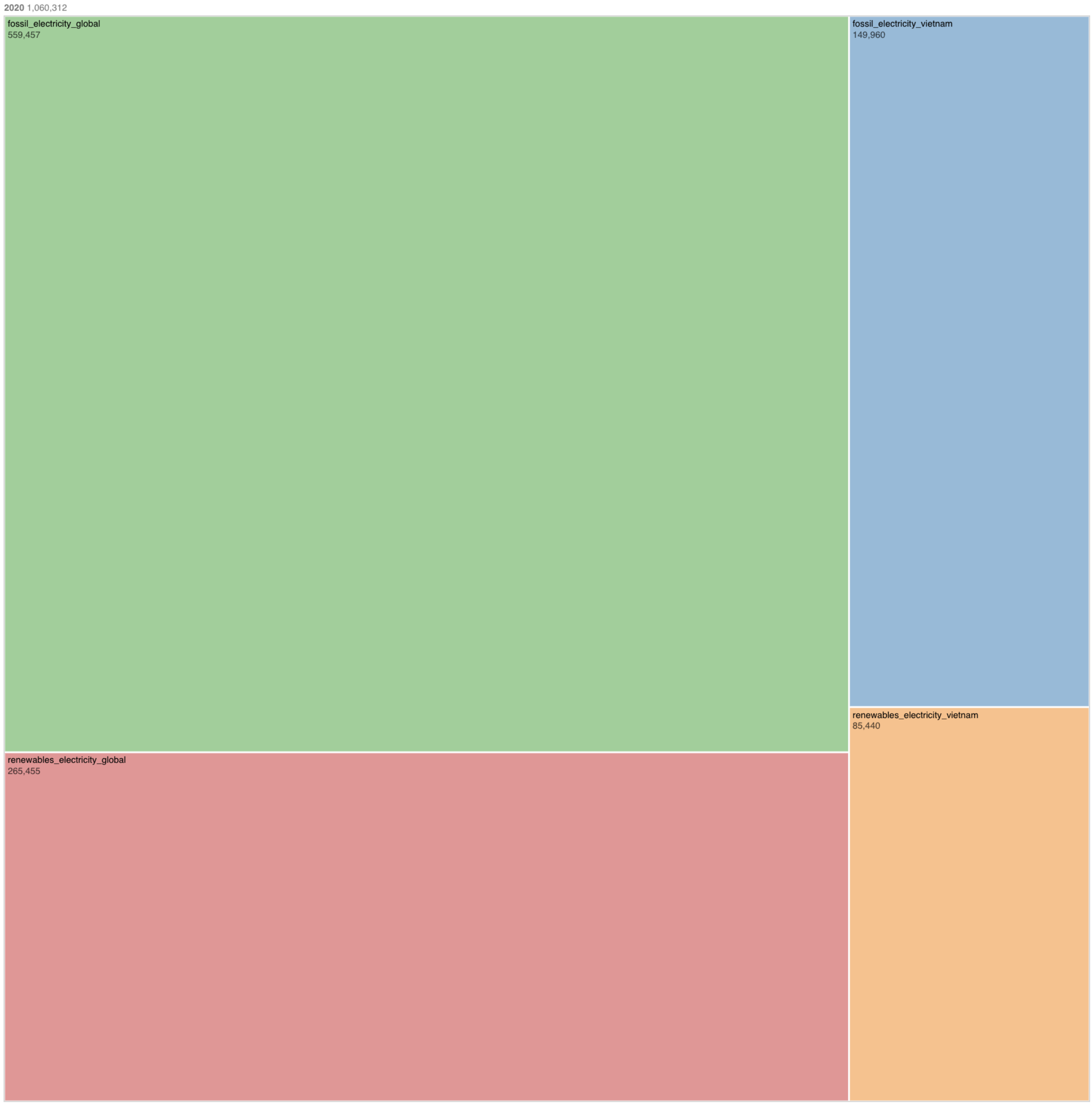


Figure 2: Treemap in 2000

Discussion:

- Surprisingly, the amount of electricity generated from renewable energy in Vietnam in 2000 was greater than the amount of fossil energy. I didn't know that until I worked on this question. In my opinion, this is because of the large amount of electricity generated from hydro power dams, which is considered renewable.
- Over the course of 20 years, the proportion of fossil electricity in Vietnam gradually increased, which made the portion of renewable electricity decreased. I believe that this was because of the newly built thermal power plant in Vietnam. This contrasts with the global trend, which remains stable from 2000 to 2020.
- One of the technical issues with the chart is that it can not represent correctly the name of the data region in the year 2000 because the chart is too small. This problem is solved when the timeframe is changed to the year 2020, which makes the chart bigger. This is a native issue of Treemap, so it can be solved by our team.

Conclusion:

The analysis highlights distinct electricity generation patterns, showing that the energy mix in the top five countries by consumption is more closely tied to their development stage rather than GDP size. Developed countries like Germany, Japan, and the USA have diversified energy sources including renewables, whereas developing countries such as China and India rely heavily on coal and hydroelectric power. Vietnam's shift from a renewable-heavy mix to more fossil fuels over the past decade contrasts with global trends towards renewables. This suggests a development-induced energy transition, with further research needed on its socio-economic and environmental impacts. Enhanced data visualization techniques could also provide clearer insights into these global and regional energy dynamics.