Eddieb Sadat
Professor Zadbood
EM622-WS Decision Making via Data Analysis
12/18/22

Final Report

Introduction:

The energy industry spans globally and is the backbone of many societies and the economies. Nearly every product and service on the planet relies on the energy industry in some capacity, not to mention the people who rely on energy to power their homes and transportation. A small disturbance in energy could result in massive domino effects towards society and the economy that could cost billions of dollars. Thus, it has been a priority for many researchers and nations to maintain records and analysis of energy data, including production quantity, usage quantity, energy type, and location. One organization, the World Resources Institute, publishes this data in their magazine, which was compiled to form a large dataset of over 30,000 power plants spanning the globe[1]. This report will analyze the data, formulate questions, and attempt to answer the questions using helpful visuals.

Business Understanding:

This project was completed to satisfy the Final Project assignment for the EM-622 course; The requirements are to find an interesting dataset, analyze to formulate questions, and generate useful and high-quality visuals to answer the questions. This project serves as the opportunity to showcase and utilize the cumulative knowledge gained from EM-622 and serve as practice for future work with data.

The first step of this project was to select an interesting dataset. In a previous report, I analyzed the stock prices of three large Oil and Gas companies because I believed their stock prices could be a decent indicator for global economic status. Continuing with this theme, I selected a dataset that includes information about Power Plants across the world[1], which includes the location of each power plant, the type of fuel used, capacity, and owners. Some interesting questions can be answered with this data:

- Where are most power plants located?
- Which countries have the most energy capacity?
- Which owners own the most power plants?
- What is the most common fuel source to produce energy? Per Country?

Beyond fulfilling the EM-622 assignment, completing this project successfully demonstrates skills that will be useful in the energy industry and beyond. For organizations who

are looking to understand energy usage and needs across the world, simple and accurate visuals are crucial for decision making. For example, UNEP (United Nations Environment Programme) may want to know precisely how, where, and the quantity of energy produced in major regions to assist them in making a more informed decision about potential new policy implementation. Thus, it is imperative to transform the raw-data into a useful tool that can foster positive change in the world.

Data Understanding:

The dataset was downloaded from the repository[1] and subsequently imported into R. To better prepare for producing accurate and useful visuals it is important to first understand the structure and makeup of the data.

```
'data.frame':   29910 obs. of  9 variables:
 $ Country        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan"
 $ Powerplant.Name: chr  "Kajaki Hydroelectric Power Plant Afghanistan" "Mahipar
lu Dam Hydroelectric Power Plant Afghanistan" "Nangarhar (Darunta) Hydroelectric
 $ gppd_idnr      : chr  "GEODB0040538" "GEODB0040541" "GEODB0040534" "GEODB0040
 $ Capacity..MW.  : num  33 66 100 11.6 42 ...
 $ Latitude       : num  32.3 34.6 34.6 34.5 34.6 ...
 $ Longitude      : num  65.1 69.5 69.7 70.4 69.1 ...
 $ Primary.Fuel   : chr  "Hydro" "Hydro" "Hydro" "Hydro" ...
 $ Owner          : chr  "" "" "" "" ...
 $ Source         : chr  "GEODB" "GEODB" "GEODB" "GEODB" ...
```

Fig. 1 - Structure of the Data

Figure 1 shows the structure of the dataset as observed in R, consisting of 9 columns and 29,910 rows. In combination with the information provided by the author of the dataset[1] and the figure above, a breakdown of each variable is listed:

1. Country (chr): Name of the country where the power plant is located.
2. Powerplant.Name (chr): Name of the power plant.
3. Gppq_idnr (chr): ID of the power plant.
4. Capacity..MW. (num): Maximum energy production in Megawatts.
5. Latitude (num): Location in latitude.
6. Longitude (num): Location in longitude.
7. Primary.Fuel (chr): Type of fuel used to generate energy.
8. Owner (chr): Organization that owns the power plant.
9. Source (chr): Source data was retrieved from.

Country, Latitude, and Longitude will be useful in showcasing the location of the power plants. Capacity..MW will be useful in showcasing the maximum energy capacity of each power plant. Owner will be useful in identifying the most common owners. Finally, Source will be useful in showcasing the fuel types used in each power plant for energy production.

Data Preparation:

Next, to produce accurate visuals, the integrity of the data must be confirmed. The first step is to identify whether there are any missing values. Normally, missing values would be represented by 'NA', but in this dataset missing values are represented by a blank character.

```
colSums(mydata == '')
      Country Powerplant.Name       gppd_idnr    Capacity..MW.        Latitude       Longitude    Primary.Fuel
            0               0               0               0               0               0               0
        Owner          Source
        10379              15
```

Fig. 2 - Missing values per Column

Figure 2 shows the number of missing values in each column as observed in R. Owner has the most concerning number of missing values, at nearly 35% of the data missing (10,379 missing out of 29,910), followed by Source, which only has 15 missing values. Recalling figure 1, both Owner and Source are character-type variables. In other words, it is not possible to fill in missing values through quantitative means, such as replacing them with the column mean or range. The best solution would be to generate a predictive model(s) to determine the most likely value for those missing values, which is out of the scope for this assignment. Furthermore, it would not be a good idea to simply remove all rows with missing values because it is important to have documentation of as many power plants as possible (assuming this was a report for environmental analysis or such). To compromise, it was decided that no rows would be removed, however, when generating stats or visuals for Owners, it would be noted that the values may not be indicative of reality since 35% of the data is missing. Since source is not useful for this project, it will be ignored.

It is important to never permanently modify a dataset just in case original data must be referenced. Thus, throughout the modeling process, whenever additional data preparation or manipulation is done, the changes are stored in new variables to keep the original untampered.

Modeling:

One of the interesting questions was to determine where most power plants are located. To answer this question, the Country variable will be utilized. Because Country is a character-type variable, it is categorical. Additionally, each row in the dataset indicates one distinct power plant. Thus, counting the number of power plants for each country results in a quantifiable variable that can be graphed. This results in 164 unique countries with power plant values [Fig. 3]

```
data.frame':   164 obs. of  2 variables:
$ Country: chr  "United States of America" "China" "United Kingdom" "Brazil" ...
$ n      : int  8686 3041 2536 2340 2017 1154 982 861 614 505 ...
```

Fig 3. - Structure of Dataframe with Power Plan quantity per Country

A barplot is a good visual to compare the quantities of power plants per country. However, since 164 variables on one graph would be illegible, the bar graph was generated to compare the top ten countries in terms of quantity of power plants [Fig. 4].

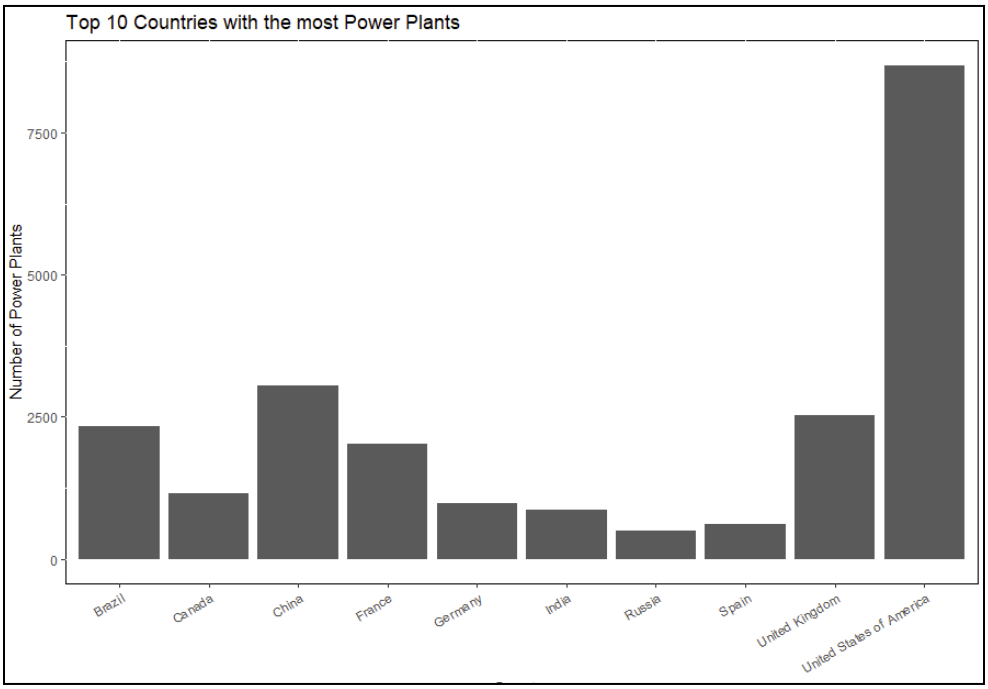

Fig. 4 - Top 10 Countries by Quantity of Power Plants

To compare all the countries, a geographical heatmap is a great visual, as it will color each country according to a gradient based on the quantity of power plants [Fig. 5].

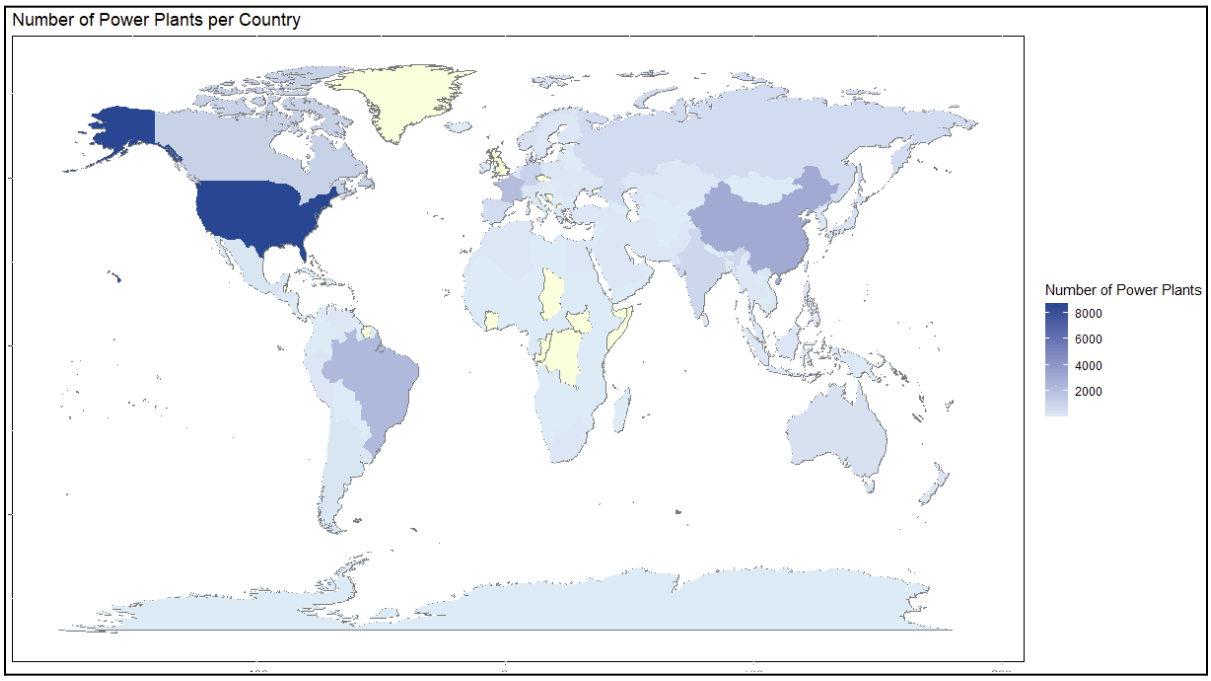

Fig. 5 - Geographical Heatmap of Power Plants per Country

Originally, the maps library was used to attempt creating this map (as demonstrated by prior examples in the EM-622 curriculum). However, no matter the attempt, there was always a discrepancy between the gradient and the countries. Instead, a different method was used utilizing both the maps and ggplot libraries[3]. Furthermore, many of the country names from my dataset did not match the maps names. To resolve this, the names of missing countries were changed manually[3]. Finally, since not all countries were available in my dataset, an initial plot was generated to indicate missing countries with a light yellow color[3].

Next is to determine which countries have the most energy production capacities. To do so, the sum of energy capacity per country is generated using the aggregate function. As mentioned previously, to avoid over-cluttering of data, the top ten countries in terms of summed capacity will be saved for visualization[4] [Fig. 6].

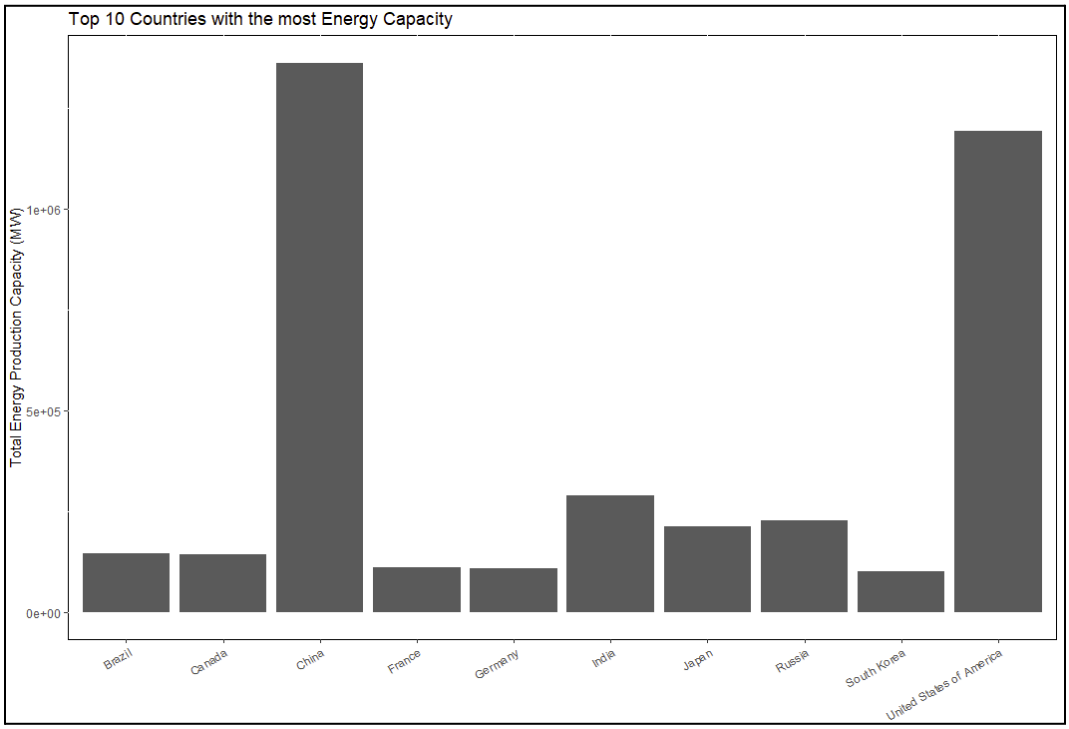

Fig. 6 - Top 10 Countries by Total Energy Production Capacity

Once again, to better visualize all countries, a geographical heatmap is produced [Fig. 7]. The same steps as the previous map were repeated, where a gradient is formed to compare energy capacity per country (darker means more energy capacity and lighter means less). Countries in light yellow were countries with no available data.

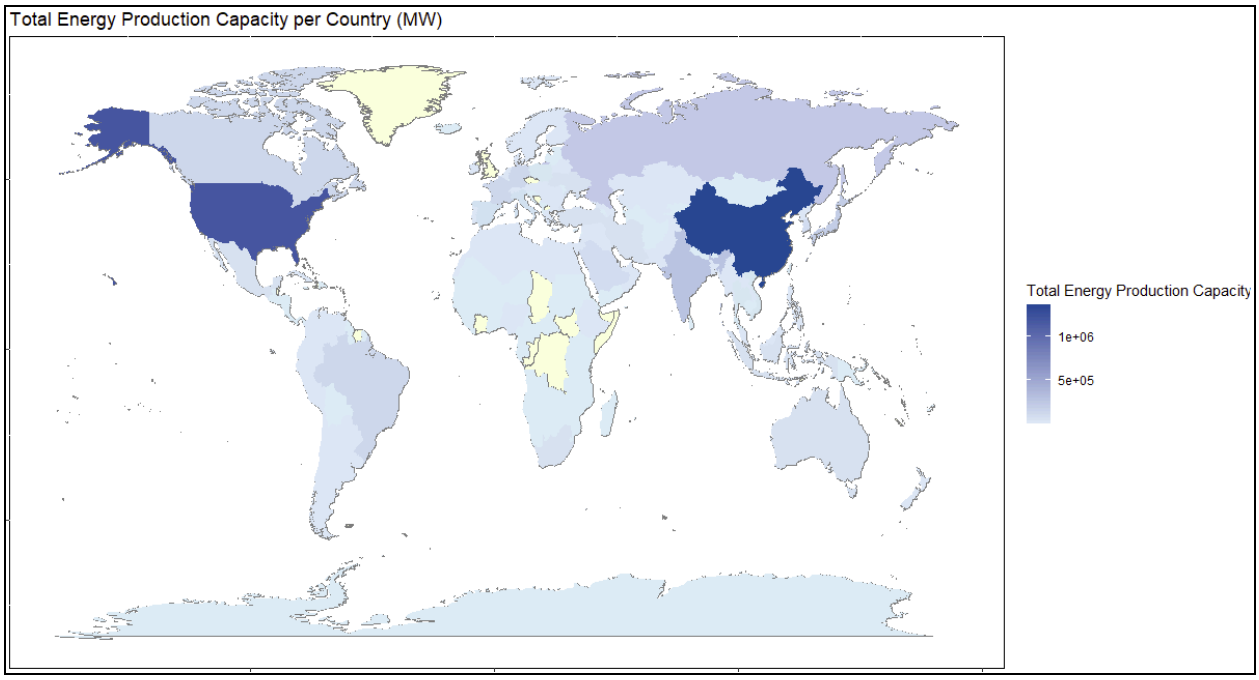

Fig 7. - Geographical Heatmap of Total Energy Production Capacity per Country

To determine who the most prominent power plant owners are, the same technique was utilized when determining the number of power plants per country. To prevent crashing or unwanted columns, the blank rows are omitted [Fig. 8].

```
> colSums(mydata1 == '')
       Country Powerplant.Name       gppd_idnr   Capacity..MW.        Latitude       Longitude    Primary.Fuel
             0              0               0               0               0               0               0
         Owner          Source
             0               0
> str(mydata1)
data.frame':   19531 obs. of  9 variables:
```

Fig. 8 - Sum of Blanks per Column and Structure

After summing up the occurrences per unique Owner, it was determined that there were 9,565 different owners [Fig. 9].

```
> str(ownermax)
data.frame':   9565 obs. of  2 variables
```

Fig. 9 - Structure of Dataframe with Number of Power Plants per Owner

To produce a more legible bar graph, only the top ten owners by number of power plants was graphed [Fig. 10].

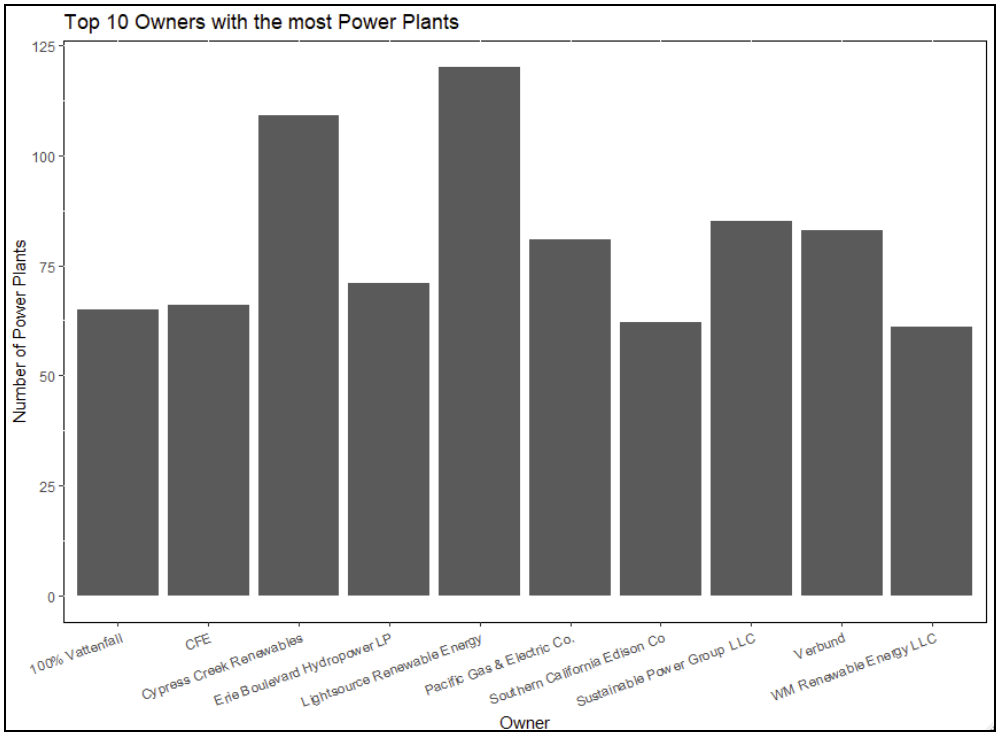

Fig. 10 - Top 10 Owners by Number of Power Plants

Finally, to determine the most common fuel sources, the same technique is utilized again to count the number of occurrences per unique fuel source. A bar graph is produced with the results [Fig. 11].

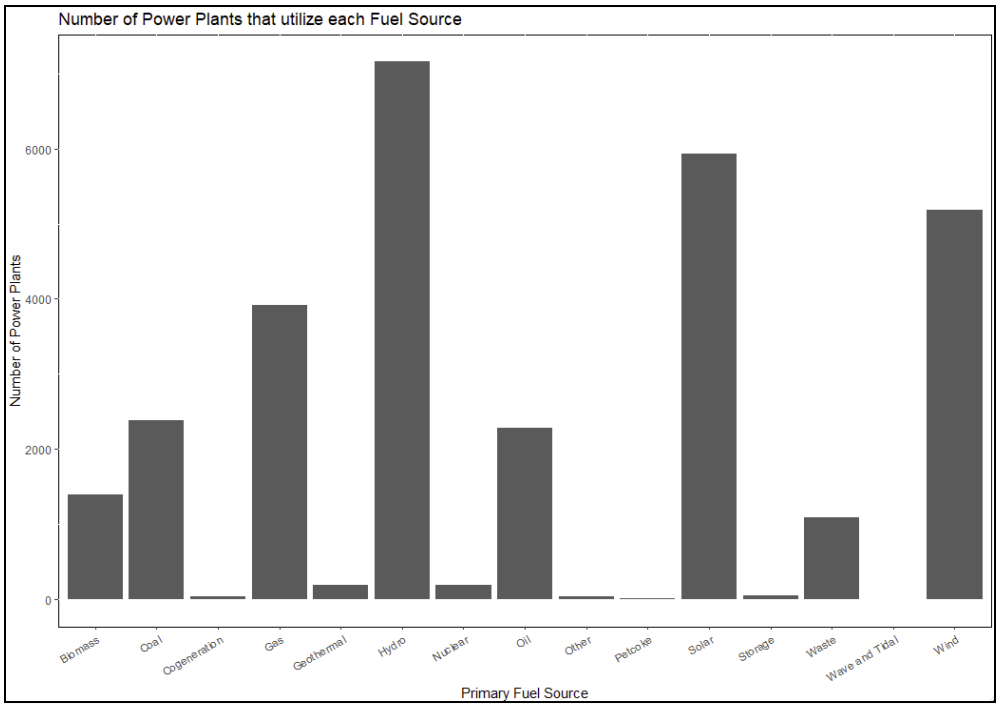

Fig. 11 - Number of Power Plants that utilize each Fuel Source

To better visualize all of the important information gathered so far, an index-type treemap is generated. The data is indexed by country and fuel source, and the size is determined by the maximum energy production capacity [Fig. 12].

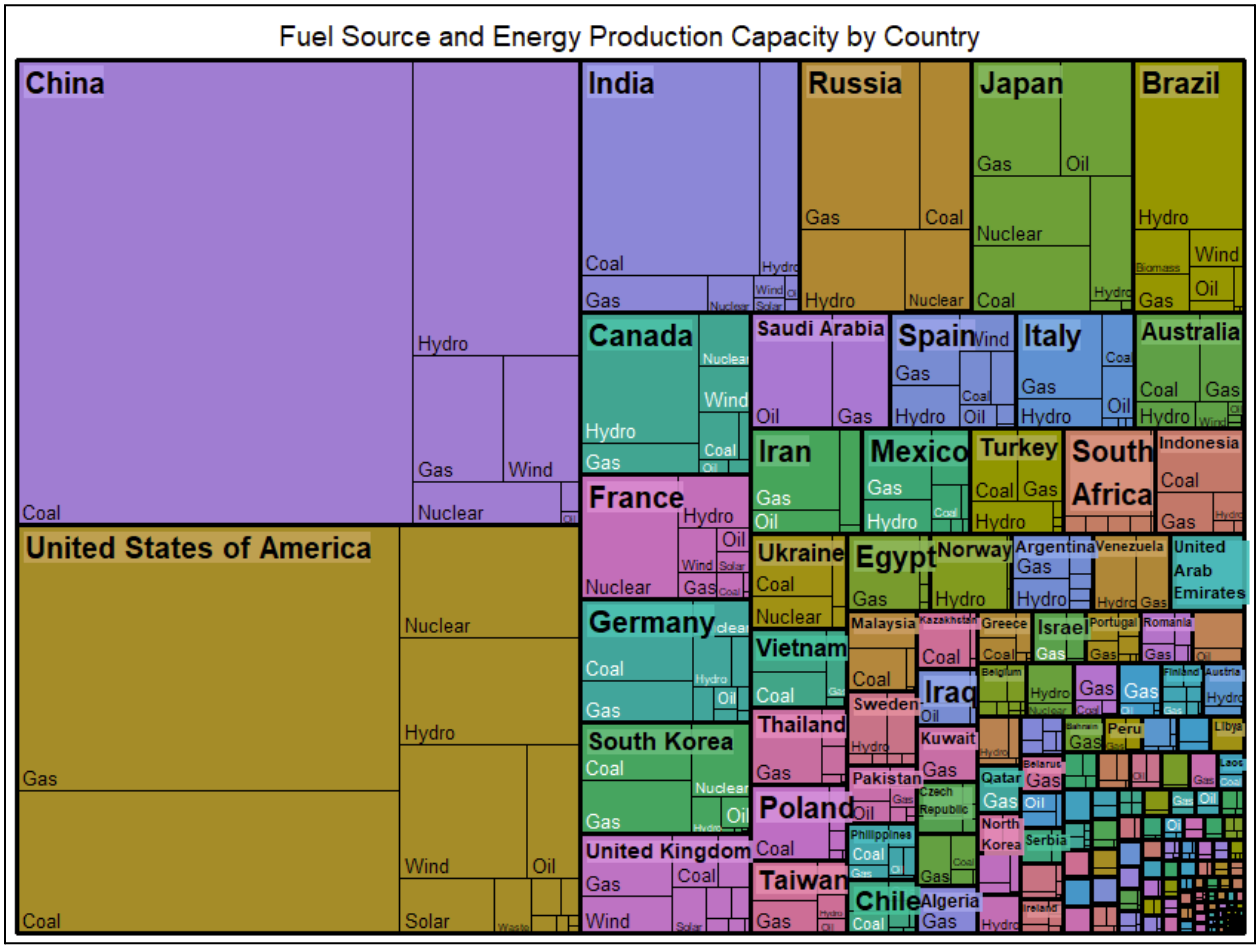

Fig. 12 - Treemap depicting Fuel Source and Max. Energy Capacity per Country

Evaluation:

To be a little informal here - wowzers that is a lot of figures! To effectively evaluate, the evaluations of each question and visual will be done in the same order as presented above.

From Figure 4, it is clear that the United States has the most number of power plants, nearly beating out the second closest country by three times. Notably, the ten countries are either developed and/or carry very large populations. Thus, it makes sense for these countries to have the most power plants, as they will likely need more power to fulfill their societal and economical needs. This sentiment is further supported by Figure 5, where all developed or developing countries indicate respectable numbers of power plants. On the other hand, most of continental Africa and large portions of South-West Asia show a severe lack of power plants as

compared to the rest of the world. This makes sense as most of these countries are underdeveloped.

From Figure 6, it is clear that China and the United States have by far the largest maximum energy production capacity, with each country nearly beating the next closest by five-to-ten times. This is particularly shocking considering that China has about a third the number of power plants, and still somehow manage to have a larger capacity than the United States. This is an interesting relationship that should garner more research to identify the reasons. One potential theory is that the United States has stricter environmental laws, and thus, are limited to certain energy production by the EPA. Other than this anomaly, the rest of the countries seem to follow a proportional relationship between number of power plants and maximum energy production, as proven by how similar Figures 5 and 7 look to each other.

From Figure 10, it seems that all owners generally own a similar number of power plants each, which could make sense since they are incredibly expensive to build, purchase, and maintain, which could set a limit to how many one organization can own. Furthermore, power plants are regional, e.g. in the United States utility companies usually dominate small parts of the country. The limit of expansion could also serve as a reason why this distribution seems to be more normal. Additionally, according to Figure 10, two of the top ten owners are utility companies located in California (Pacific Gas & Electric Co. and Southern California Edison). This serves as a testament to how populous and active California is, as they require many power plants to supply California demand.

From Figure 11, most power plants utilize Hydro (water) as their fuel source, followed by Solar (sun), Wind, and then Gas. Personally, this is quite shocking because I would not have thought that a mass majority of energy production is the result of renewable energy. It makes sense given the stress of environmental health nowadays, but these fuels are typically not favored by smaller countries as they are more expensive to set up or rebuild from a pre-existing power plant. Furthermore, it was a shock to see how little nuclear fuels are used. There have been many news stories regarding the dangers and harm of nuclear power plants, and it made it seem as though a far larger number of energy production utilizes nuclear energy. However, it is likely for these same reasons (particularly in producing nuclear waste) that nuclear fuel is far from being the most utilized. Overall, these were shocking, yet pleasing, discoveries about fuel sources.

Finally, Figure 12 showcases all the important data into a single visual. Looking at the sizes of each square, it is clear that China and the United States have the largest maximum energy production capacities. Furthermore, it is clear to see the different fuel sources used per country. For example, China favors coal, the United States favors gas, and India favors coal. Many of the developed countries utilize efficient fuel sources, and surprisingly, even many developing or underdeveloped countries utilize more efficient fuels like gas or hydro. There is a very clear correlation to fuel source by location, as many coastal countries utilize hydro, and many countries in the Middle-East utilize oil and gas which makes sense given the resources at each regions' hand.

Deployment:

To truly be able to use the results from this report for important decision-making, the integrity dataset must be verified to have been created appropriately, e.g. aggregating data from reliable sources, collecting the most data possible for all countries, etc. Since I do not possess those capabilities, the results of this report can't be guaranteed to be usable for important purposes. However, assuming the dataset is viable for use, the methodologies for organizing, preparing, and visualizing the data are sure to be usable for many industry applications.

Through the organized material and digestible visuals, this report can provide crucial information about countries' source for and supply of energy. It can also provide insight into the capabilities of each country in terms of energy production. The uses for this information are wide, and can be deployed to satisfy many stakeholders in the energy and environmental markets.

Sources:
[1] https://www.kaggle.com/datasets/jaytilala/global-power-plant?resource=download
[2] https://rstudio-pubs-static.s3.amazonaws.com/4305_8df3611f69fa48c2ba6bbca9a8367895.html
[3] https://stackoverflow.com/questions/61838700/query-on-how-to-make-world-heat-map-using-ggplot-in-r
[4] https://www.statmethods.net/management/sorting.html