The Direction of Power Plant and Energy Generation

Professor Qiwei Li

Date:
November 23, 2021

Group 7 Team Members:
Haeun Kim
Akarsh Sherman

Table of Contents

1. Introduction

The purpose of this project was to use statistical modelling in R to analyze a large dataset and find interesting trends and findings about that data. We wanted to find trends in global power plant data by studying a combination of datasets from Our World in Data and World Resources Institute which look at power plants across the world spanning from early 1965 to August 2020, and continue to be updated.

Our primary dataset[1] titled “global_power_plant_dataset_last” includes 34,936 instances of 36 variables including country, latitude, longitude, primary fuel source and capacity(Mw) of individual power plants around the world from 1965 to 2020.

The second dataset[2] titled “Primary-energy-consumption-from-fossil fuels-nuclear-renewables" includes 4285 instances of 6 variables including country(entity), year, fossil fuel rate, renewable energy rate and nuclear energy rate power plants around the world from 1965 to 2019. This dataset combined various renewable energy sources, the percentage of the changing trend of fossil fuels, renewable energy, and nuclear energy can be seen at a glance.

The third dataset[3] titled “share-elec-produc-by-source” includes 6190 instances of 11 variables including country(entity), year, and percentage of electricity production by fuel source(coal, gas, hydro, solar, wind, oil, nuclear). This dataset outlines how much of total electricity production in a country in a certain year was contributed by power plants using a certain fuel source.

The fourth dataset[4] titled “share-energy-consum-by-source” includes 4284 instances of 11 variables. This dataset uses the same variables as the above dataset but tracks percentage of electricity consumption in a country in a certain year versus electricity production. This dataset is also missing a few instances of data for some countries, hence, the lower total instances.

From the beginning of recorded history, human beings have depended on fossil fuels as our primary energy source. However, the scarcity of fuel sources and their detrimental effects to the environment due to emissions (Fossil fuel usage is directly related to carbon dioxide and other chemical substances emission) mean that finding alternative cleaner forms of energy are more important than ever. As a result, renewable energy options are on the rise as more renewable energy plants are built around the world in order to replace fossil fuel.

Over the course of the project, we began by comparing levels of production between renewable sources of energy like solar, wind and hydroelectricity and more conventional forms of energy like fossil fuels. We also wanted to compare the contribution of each energy source to total energy production. The results from these basic questions were used to guide deeper exploration regarding greenhouse gas emissions, power plant capacity and efficiency and overall trends in renewable power plant geolocation.

2. Data Cleaning

The first thing we did was changing the name of the file of the dataset (for example, 'global_power_plant_dataset_last' to 'global', and 'share-elec-produc-by-source' to 'production', ect) because original files' names were too complicated. The data we used clearly existed and the column names were clear so we could easily find and figure it out. And when I looked through the data, there was almost no data loss and all the numbers were well organized to the decimal point. So, datasets required rather minimal cleaning.

We found the missing instances of data in dataset[4] were from before the year 1985. So, in order to use the data, we decided to remove all instances of data from before 1985 in all our datasets. We decided to do this by creating a new dataframe with instances after 1985 for each fuel source.

For instance,

```
# Making the data frame for coal from 1985 to 2020
value <- coal.world
coal.world.prod <- data.frame(value)
coal.world.prod$year <- 1985:2020
coal.world.prod$name[1:36] <- "coal"
```

We also repeated that process for all the other fuel sources and then combined the resulting data frames using *rbind()* to get the *new.production* data frame.

3. Questions and Finding

In this section we will discuss the questions our group came up with and analyze our results. We will also discuss possible reasons for certain trends and what can be learnt from those findings. The questions we will discuss are aimed to help us to understand the trends in the world energy production and where we see it going in the near future.

3.1 About Energy

3.1.1 How has the energy production ratio of the world changed since 1985?

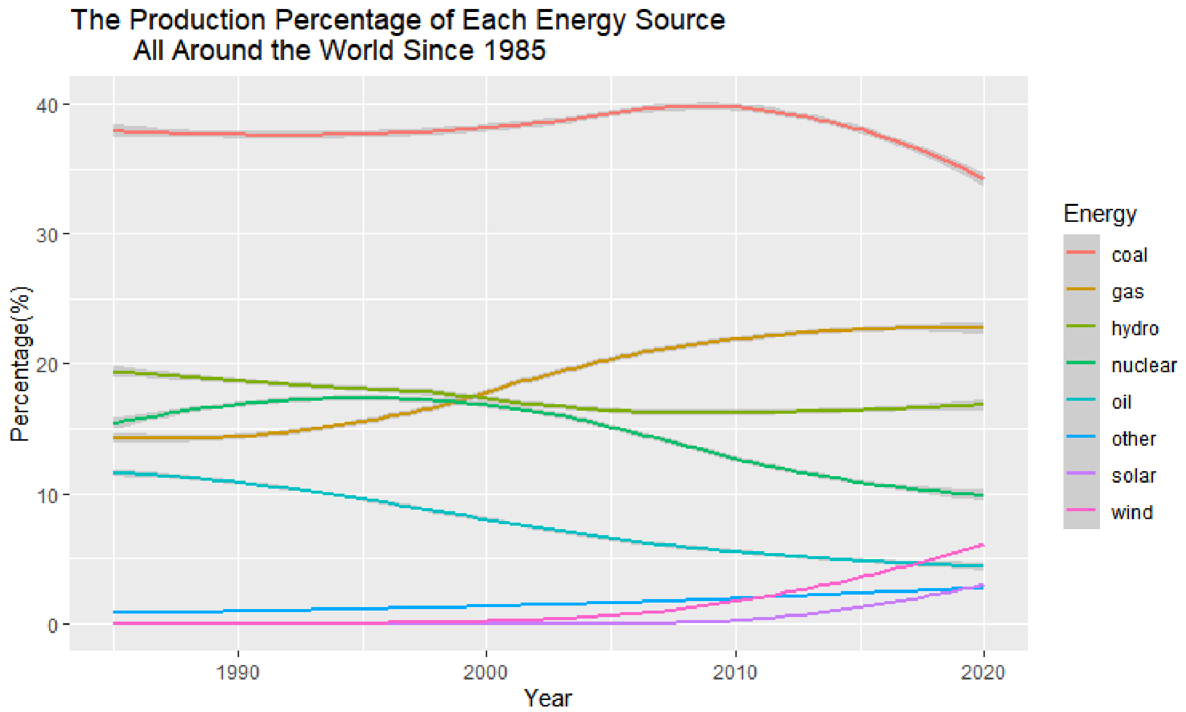

Figure 1. Trend line for the production percentage of each energy source over the world from 1985 to 2020.

First, we wanted to look at a broad trend line to see how energy production had changed according to our data. Since we had some missing data for some countries before 1985, we decided to exclude that data while looking at the trendline and instead focus on data after 1985. We noticed that approaching 2020, all renewable forms of energy had positive trends, with solar and wind energy showing the greatest percent increase in the last 10 years. Natural Gas also experiences a steady increase up until 2015 where it starts to level out. The use of coal experiences an upward trend approaching 2010, but declines into a steep downward trend approaching 2020. We can figure out that there is a shift from coal to natural gas. (When we searched for it, the world tends to change from coal to natural gas because natural gas influences a big reduction in greenhouse gas emissions, but also much less fine particulate matter pollution.) Nuclear energy has also been on a downward trend since 2000. The above trends suggest that the use of green energies is rising. These observations serve more to lay a foundation that can be used to find trends in energy production by major energy producers.

3.1.2. How does the production of each energy source in the five countries with the largest energy production change?

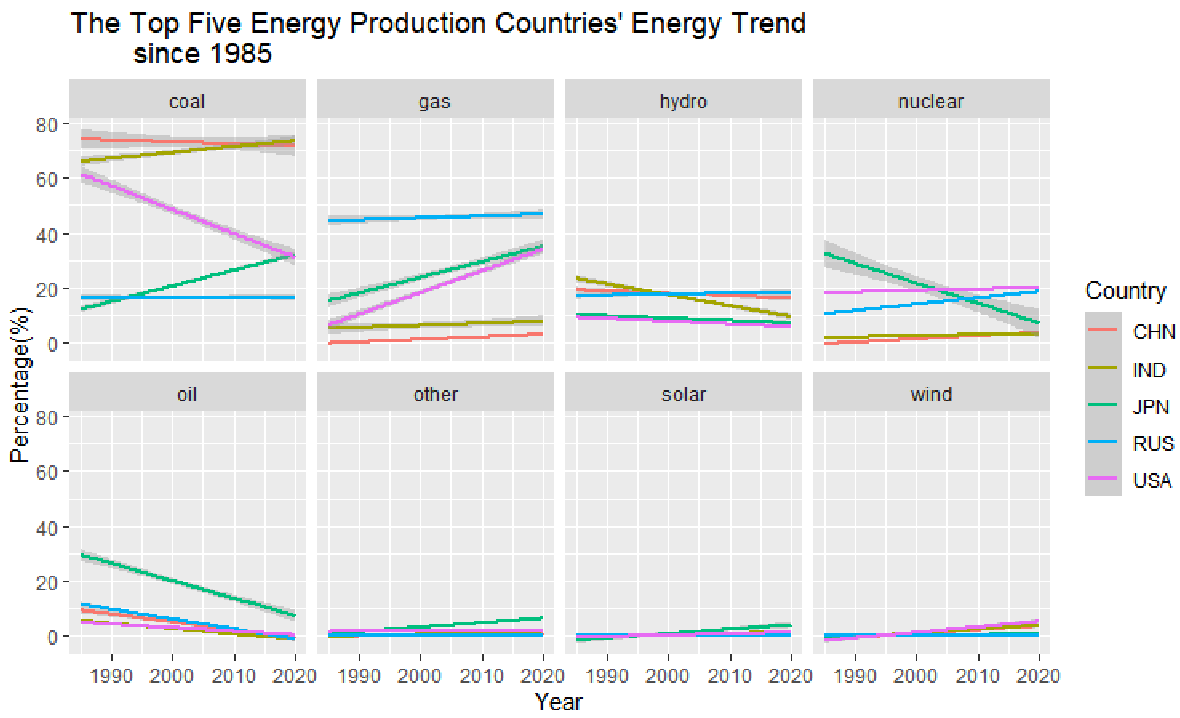

Figure 2. Trend line of each energy source production percentage of top 5 energy production countries from 1985 to 2020.

We were then curious as to how the largest energy producers have changed their energy production since 1985. We chose to compare the five countries with the largest energy production as they have high demands that need to be met with cheap energy production. Unsurprisingly, we found that these major countries still use coal and natural gas as a large percent of their total energy production. We noticed that hydro, wind, solar and other(photovoltaic, geothermal and other smaller green fuels) only accounted for a small part of their energy production. Oil production is decreasing for all countries. We assumed that this occurred with the increase of renewable energy and gas energy because our dataset figures are indicated in percentages. We noticed that while the USA and China have reduced their coal use, India and Japan are increasing their coal use. Gas production is on the rise for all countries. India, Japan, USA and China are using less hydro energy, while Russia is slowly increasing their production of hydro energy. Except for Japan, all countries have increased their use of nuclear

energy. It can be seen that Japan's nuclear power production has decreased a lot since the Fukushima nuclear power plant explosion. However, there is a slight upward trend amongst the use of these green fuels which suggests that these technologies are getting better and may be used more effectively in the future.
If we do more additional work, we want to focus on changes in actual production. Because our data is expressed as a percentage, we want to compensate for the disadvantage that no matter what specific energy increases, it decreases if it increases relatively less compared to other energies.

3.1.3. Are there any differences between energy consumption and production? If so, why does it happen?

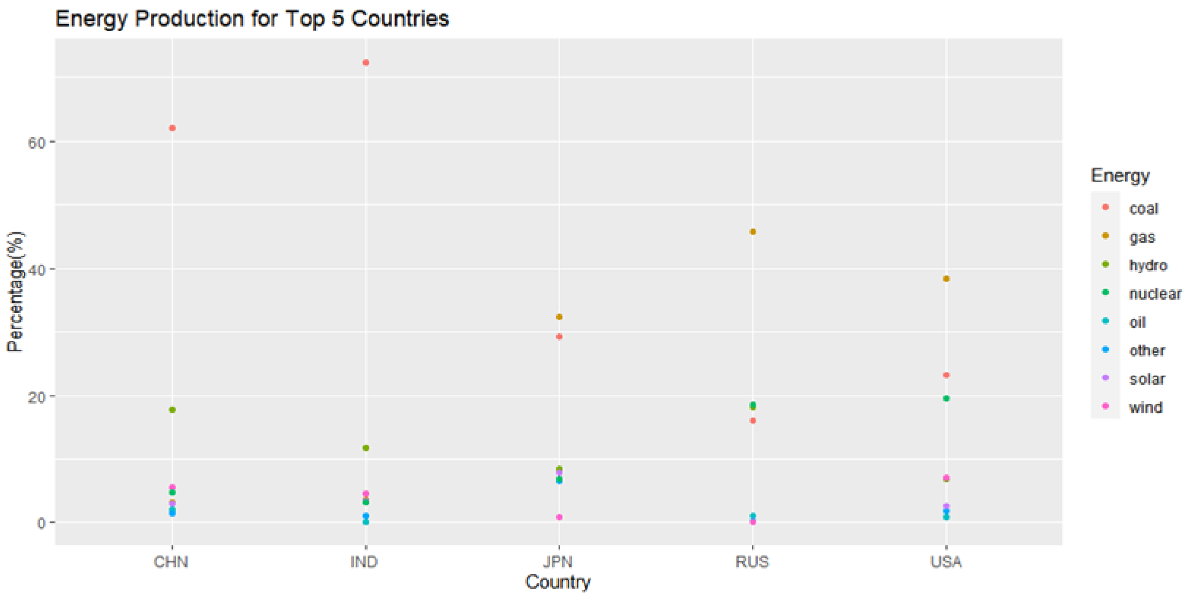

Figure 3. Point plot of each energy source production for the top 5 energy production percentage of 2019.

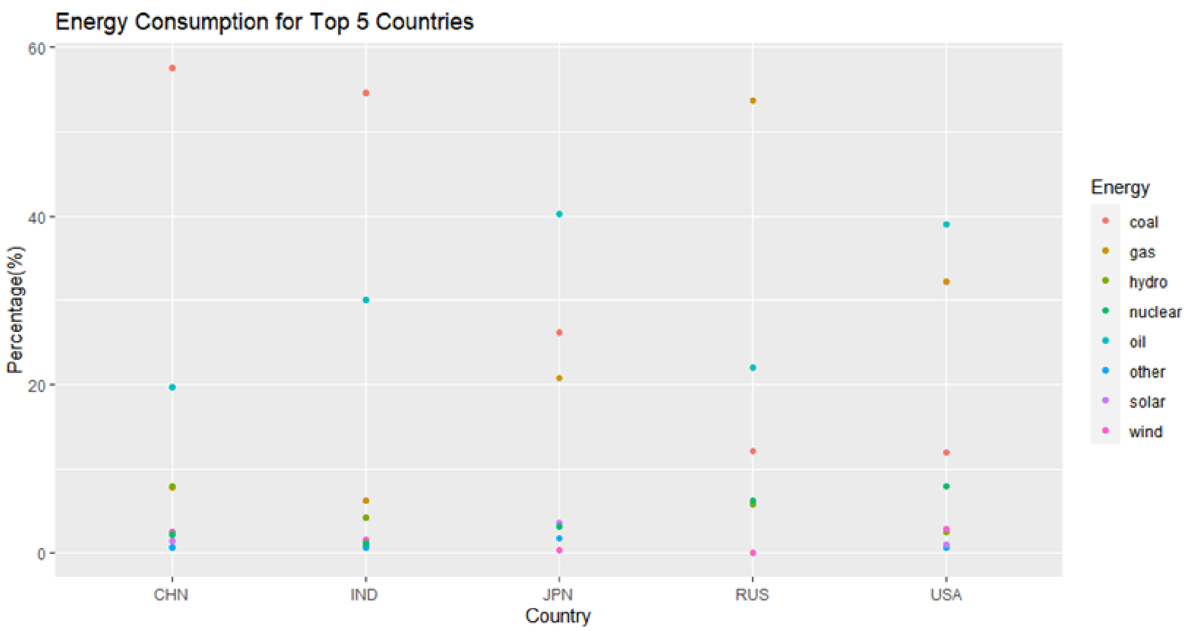

Figure 4. Point plot of each energy source consumption for the top 5 energy production percentage of 2019.

We also wanted to look at the difference between energy production and energy consumption amongst the top energy producers in the world. We picked to compare in the year 2019 as that is the latest year where we have all the data points for the whole year. We found that coal was both the highest produced and consumed in both China and India. They also used hydro electricity to produce between 15-20% of their total electricity. However hydro electricity only constituted around between 5-10% of their energy consumption. Oil accounted for around 40% of total energy production for the USA, Japan and Russia. An additional investigation found that the difference between energy production and consumption occurs because energy loss occurs in the process of transporting and storing energy, and the created energy is stored or exported to other countries. This discrepancy in levels suggests that there is a difference in expected needs and actual requirements. Some of this discrepancy also suggests that there are no viable storage/distribution methods for renewable energies.

3.2 About Renewable Energy

3.2.1. Are there environmentally friendly countries that use more renewable and nuclear energy than fossil fuel?

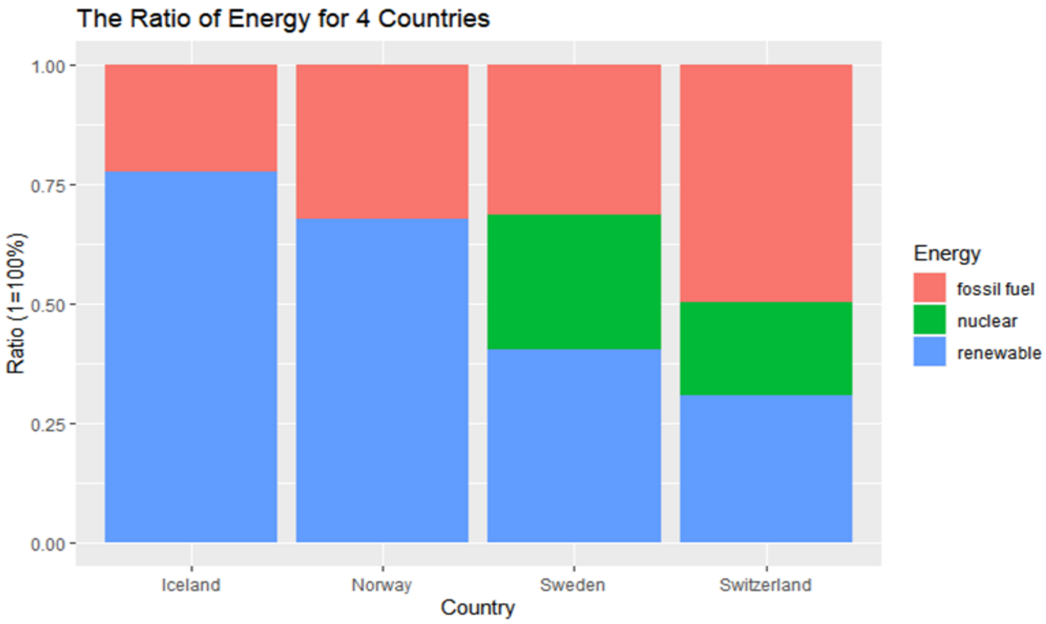

Figure 5. Bar plot for the ratio of energy for 4 eco-friendly countries in 2018.

We also wanted to look at whether it was possible for countries to be dependent on green energies. We determined that if a country was using green energies, i.e. renewable energies or nuclear energy to produce 50% or more of their total energy production, they were dependent on green energies. After manipulating the data, we found that only four countries fit that criteria, namely Iceland, Norway, Sweden and Switzerland. We found that Iceland and Norway used only renewable energies to produce close to 75% of their energy while Sweden and Switzerland used a combination of green and nuclear energy. We came to the conclusion that it is possible for countries to be dependent on green energies but it appears that it is a lot easier for smaller countries with less energy requirement to transition to depending on renewable sources. This could mean that renewable sources of energy are not yet efficient enough to support larger demands and may require more innovation before they are viable to larger countries.

3.2.1.1 Will these eco-friendly countries have low levels of greenhouse gases, or carbon dioxide?

We obtained above that there are four countries which produce more renewable energy and nuclear energy than fossil fuel. We assumed that those countries have less carbon dioxide compared to other countries because we thought that the use of fossil fuel is directly related to generating carbon dioxide. So, we found another dataset about carbon dioxide emission figures all around the world, and made this plot.

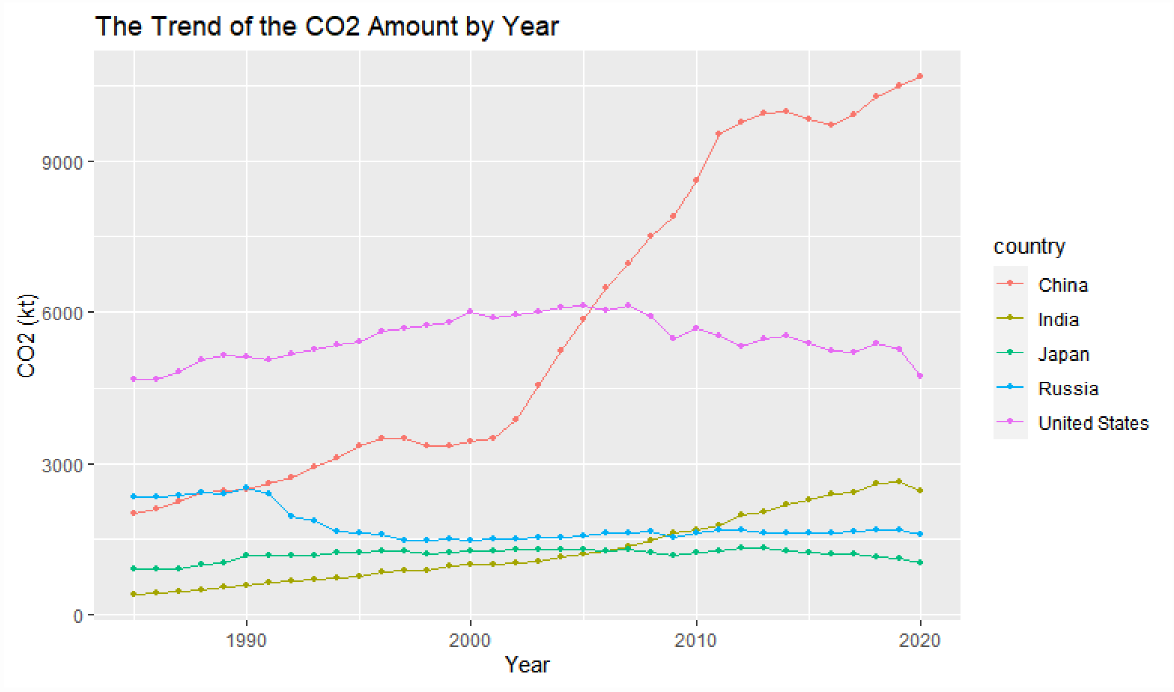

Figure 6. Trend line of CO2 amount for top 5 energy production countries from 1985 to 2020.

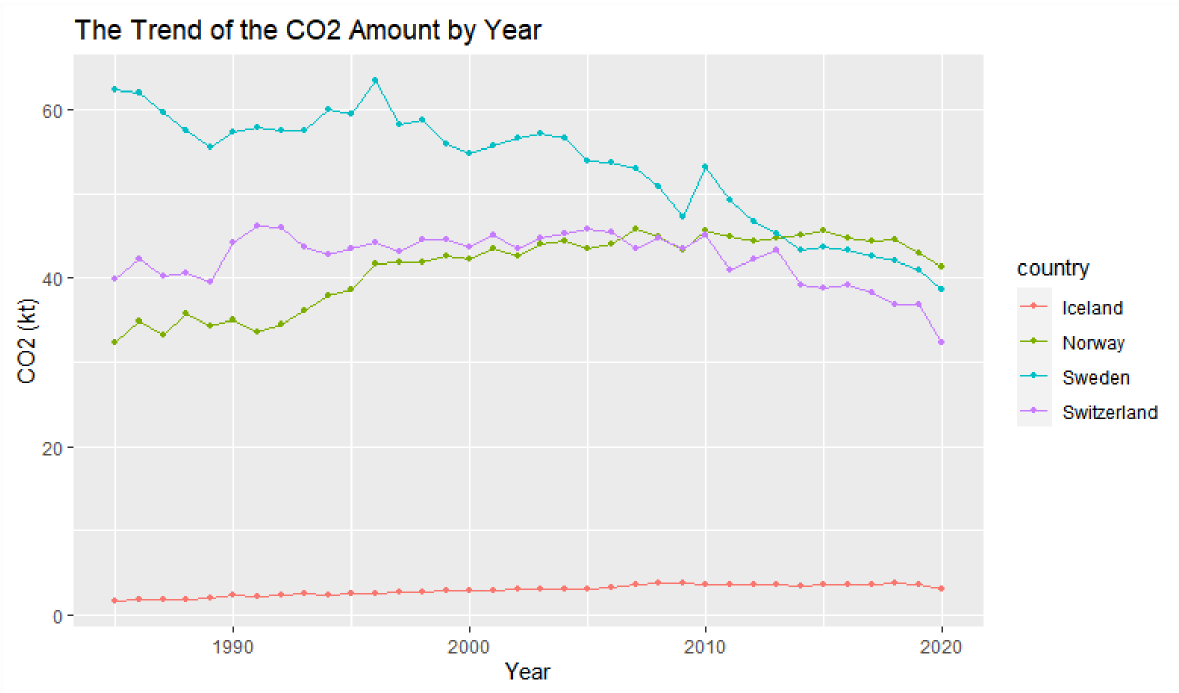

Figure 7. Trend line of CO2 amount for 4 eco-friendly countries from 1985 to 2020.

The graph above shows the amount of change in carbon dioxide by year in the top five countries that produce the most energy in the world, and below is the amount of change in carbon dioxide by year in the environmentally friendly countries we figure out from above. We can see that the amount of carbon dioxide is much less in environmentally friendly countries than in the top five energy production countries. When we focus the carbon dioxide value on 2020, the carbon dioxide emission of Japan, has the lowest carbon dioxide among top five energy production countries, is assumed more than 780kt, while the carbon dioxide emission of Sweden, has the highest carbon dioxide among eco-friendly countries, is assumed more than 40kt. And when comparing Sweden's CO2 emission to China's CO2 emission, China's CO2 emission is more than 230 times of Sweden's CO2 emission. We could figure out that there is a significant difference between the two groups regarding carbon dioxide emission.

However, we couldn't define that this significance occurred because eco-friendly countries generate energy by using renewable energy and nuclear energy more than fossil fuel. When we researched the cause of CO2, we figured out that fossil fuel is a significant cause of CO2 emission, but CO2 also depends on agriculture, livestock industry, population, and other things. According to the InsideClimate News, it says that "Emission from livestock accounts for about 14.5 percent of total greenhouse gas emissions, globally". We concluded that the eco-friendly countries have less CO2, but carbon dioxide can vary and change according to the characteristics of the country. In the future, we want to focus on the carbon dioxide emission's reason and percentage for each country. If so, we can find out the real values of CO2 emission from fossil fuel, and its value might be correct and much more accurate than what it is now.

3.3 About Energy Plant

3.3.1. Is the amount of energy generated proportion to the number of energy plants?

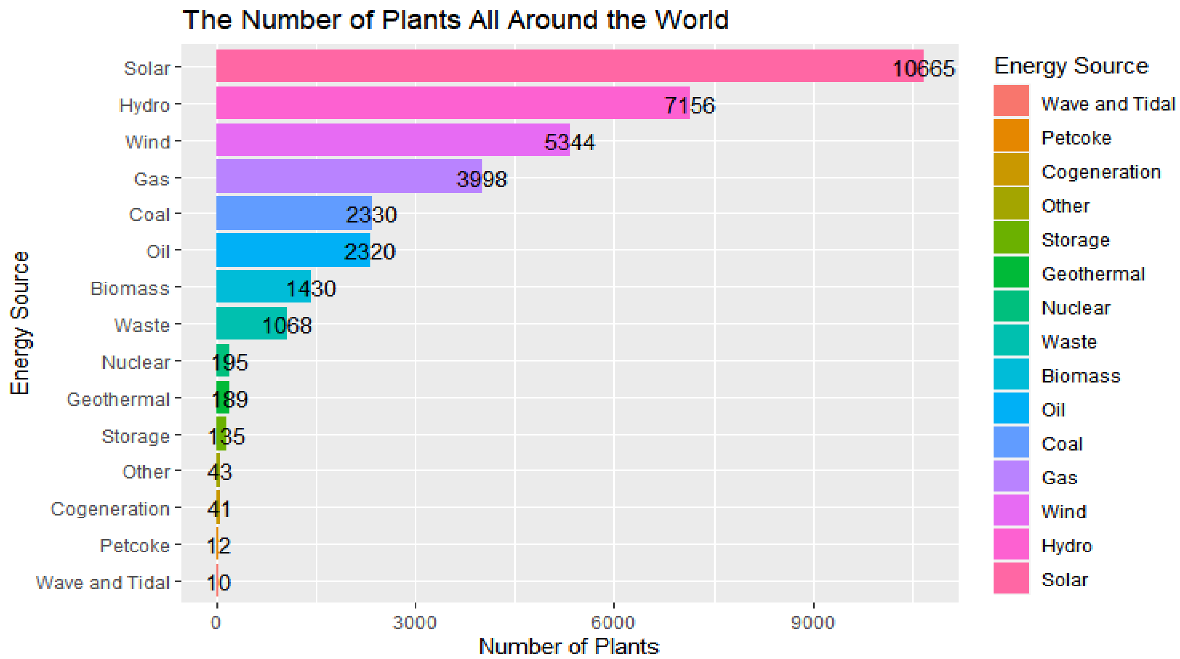

Figure 8. Horizontal Bar plot for the number of plants for each energy source over the world.

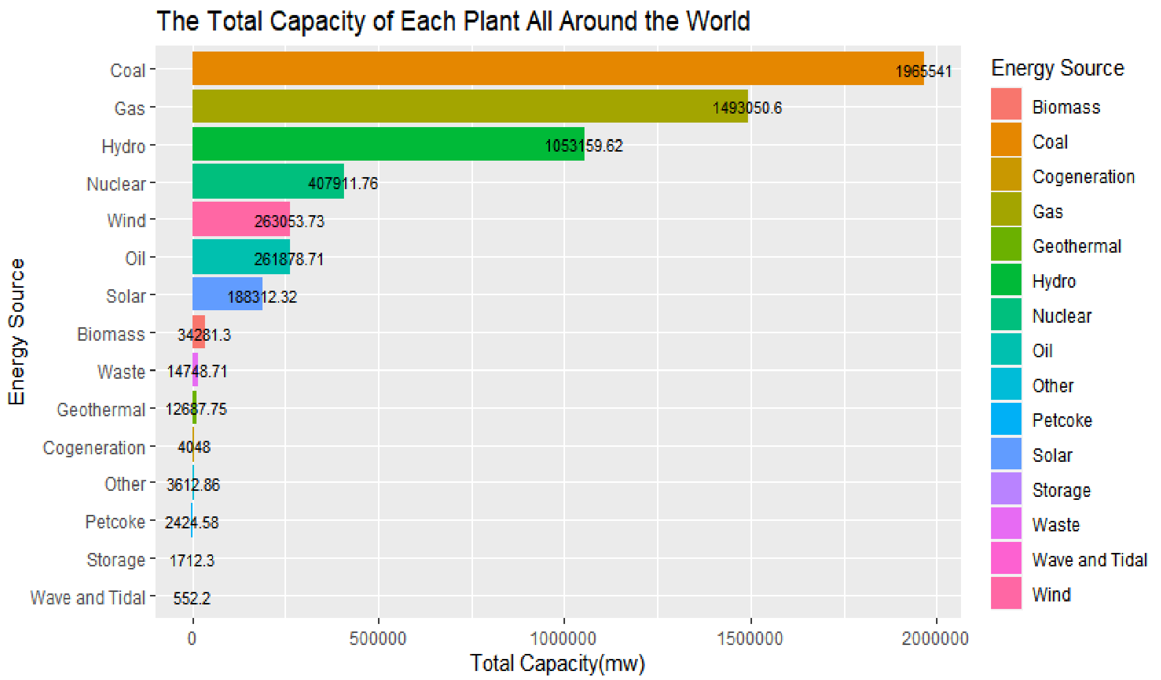

Figure 9. Horizontal Bar plot for the total energy capacity for each energy source over the world.

At this point, we were curious about the relationship between the number of energy plants and the total capacity of each energy plant. The result was quite surprising to us. We can see that fossil fuels such as coal, oil, and gas generate much more energy than renewable energy. First of all, the three energy sources with the largest number of power plants are renewable energy, but if you look at the total capacity of power plants around the world, you can see that fossil fuels are the first and second, and renewable energy is the third. Solar energy accounts for the largest number in the world, but its total capacity is too small. On the other hand, there were 2,330 coal plants worldwide, but the capacity to produce was overwhelmingly large. And we can also see that coal produces the largest capacity among fossil fuels compared to the number. Gas has more power plants than coal, but its total capacity is less than coal, and in the case of oil, the number of power plants is similar to coal, but the amount of energy that can be generated is too small compared to coal. Hydro energy seems to produce a similar amount of energy compared to the number, and wind energy has the similar total capacity as oil, but the number of power plants is more than twice that of oil power plants. One interesting plant is nuclear energy. It has only 195 power plants all around the world, but it's total capacity is much bigger than solar, oil, and wind.

We expected that the number of energy plants will have a significant relation with the total capacity of the energy since renewable energy plants are easier and common to find, however, according to the graph, it derives that the amount of energy is not generated proportionally to the number of plants. We observed that the renewable energy plants that are most common to find are not producing enough energy compared to the fuel energy and we can predict that increasing the number of renewable energy plants will not highly affect the total energy capacity of overall plants around the world.

3.3.2. Which energy plant can produce the most energy?

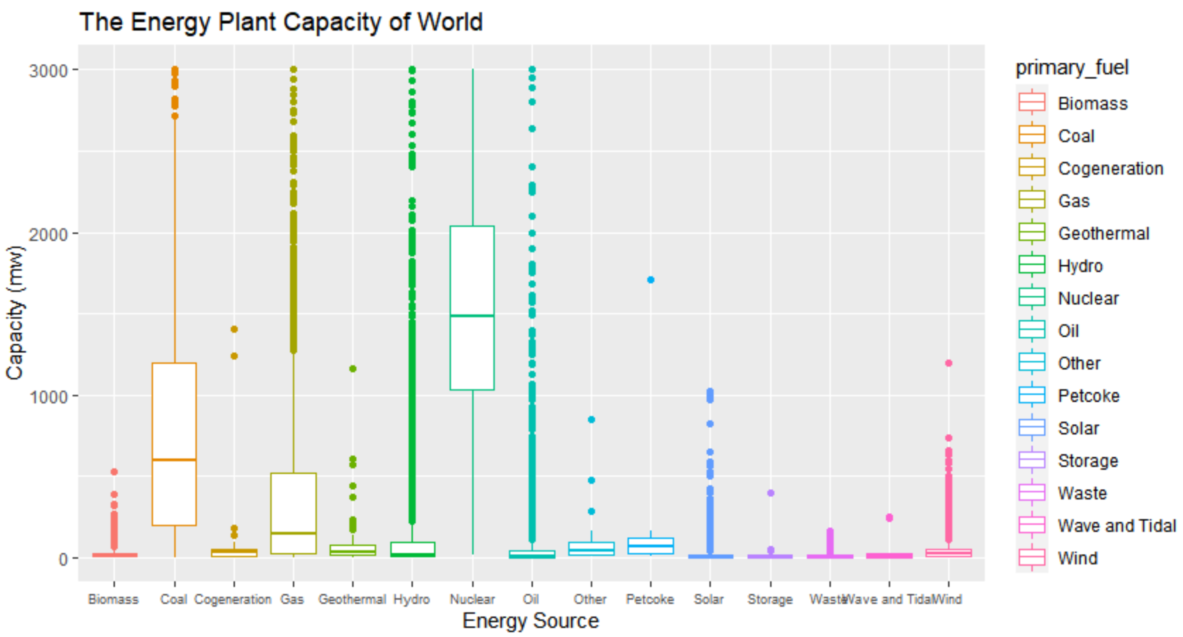

Figure 10. Boxplot of the energy plant capacity for each energy source over the world.

We can see above the total number of power plants and the total capacity of each energy source, and after finding out that the relationship between the two is not very relevant, we created this graph to find out the distribution of capacity that each energy plant can produce.

We can figure out that nuclear power plants have the largest capacity all through other energy sources. Following energy sources are Coal, and Gas. When we compare it with the above question graph, we can easily understand why the nuclear power plant has a low number of power plants, but the generating capacity accounts for quite a large scale. When we look over renewable energy, the energy producing capacity is too low. We could find out why solar energy and wind energy have the largest number of power plants and the smallest total capacity here.

3.3.2.1 Is it possible for renewable energy and other energy to replace fossil fuel and slow down global warming?

We made the conclusion to the question “Is it possible for renewable energy and other energy to replace fossil fuel?” by using the findings we found through question 3.3.1 and 3.3.2. We concluded it’ll become possible someday in the far future, but renewable energy will not be able to keep up with global warming, of which carbon dioxide is the main cause. The reason we made this conclusion is that all around the world should make the same number of plants already established from now on to cover the fossil fuels. Many experts now announce that the renewable energy plants will produce more energy than fossil fuels by 2030 and that renewable energy will replace fossil fuels by 2050, but global temperature is estimated to increase by 1.5 degrees by 2030 with current temperature increase. However, we are open to infinite possibilities for the answer to this question. Once we're not experts, we can't easily judge what many experts predict with the dataset we have, and we always open the possibility that our answers could change infinitely as new energy sources continue to develop and many countries are planning and creating large-scale renewable energy projects to generate bigger and more energy. But the conclusion we made here is what we predicted with the data and graphs we have now. If possible, we would like to make up for this by comparing the growth of renewable energy and decline of fossil fuels each year with energy-related data created and provided by each country from next year.

3.3.3. How do renewable energy plants in the States spread?
- 3.3.3.1 Does it affect the geometrical characteristics?

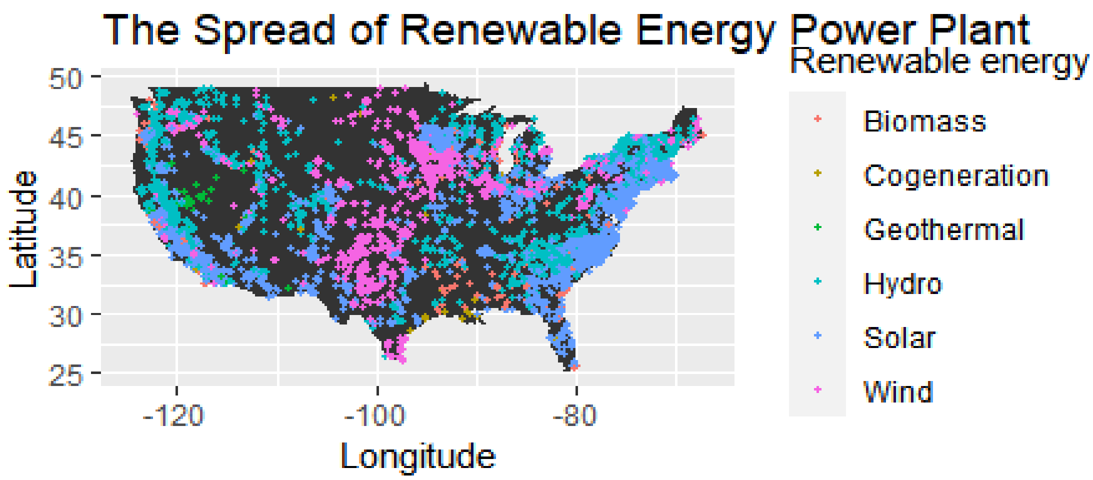

Figure 11. Map of the United States to show the spread of renewable energy power plants.

We gained the number of plants and the capacity of each energy plant all over the world. So, at this question, we actually wanted to figure out the geographical features of energy plants around the world, but we could no longer do it because we didn't know the geographic features of the world well and had difficulty identifying the geographic features one by one. So, we decided to identify the geographical features of energy plants by limiting them to the United States where we live and where we can know the geographical features.

Looking at the distribution of power plants in the United States, more energy plants are distributed in the east than in the west, centering in the central United States. We can figure it out, but it was hard to recognize any other details and other energy plants because this plot looks so messy. So, we separated it by energy plant.

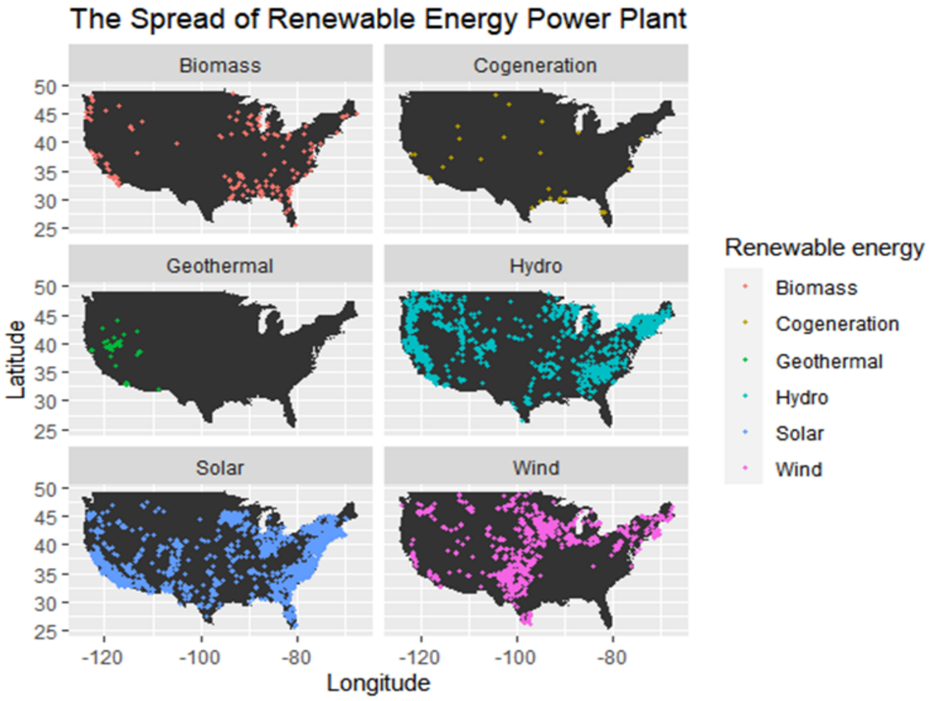

Figure 12. Map of the United States to show the spread of each renewable energy power plant.

We thought that each energy plant might be influenced by geometrical characteristics when we separated them by energy source. In the case of wind plants, they are concentrated in Mid America. Looking at the wind plant, it is concentrated in the central United States. The establishment condition for the wind plant is that the average annual wind speed should be 9 mph, round hills; open plates and water; and mountain gaps that funnel and intensity wind.

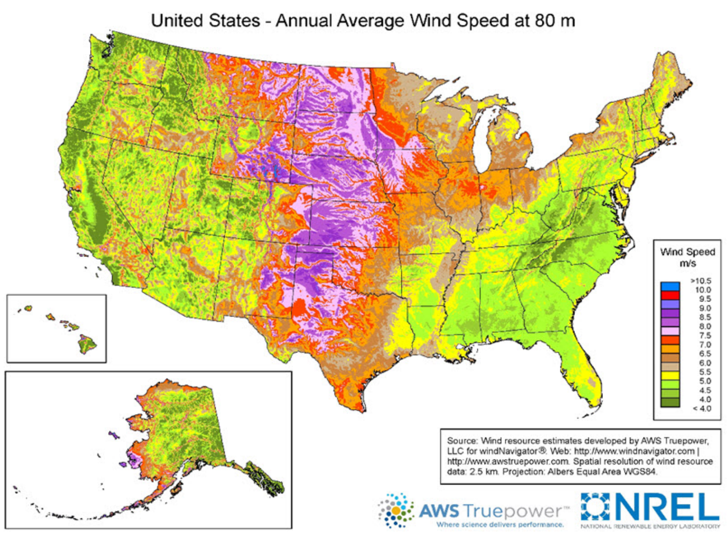

Figure 13. Map of the United States to show annual average wind speed to compare location of wind power plants [8]

The graph above is a map graph related to wind speed and is obtained from energy information administration, which is shown by wind speed. It can be seen that wind power plants are geographically affected by the fact that the graph of the region with high wind speed and the distribution of wind power plants we obtained above are similar. In the case of geothermal, California is in an active geothermal area called the Ring of Fire. So it's the best place to build geothermal power plants, and geothermal power plants are gathered in California. Hydroelectric power plants must be considered to be geographically affected. This is because it must be built in an area with water or in an area with dams. Looking at the distribution of hydroelectric power plants, it is concentrated and distributed along the coast, while there are few in the central region, so hydroelectric power generation also seems to be subject to geographic factors. However, in the case of solar plants, we found out that solar plants and hydro plants are concentrated in East and West America. Since it is installed in areas with high sunlight, it should be considered that it is more affected by the weather than by geographical influence. In the case of cogeneration plants, they do not receive geographic factors when we searched that cogeneration is the use of a heat engine or power station to generate electricity and useful heat at the same time, and biomass also does not receive geographic factors. We can see that biomass is concentrated in Florida. This is because there are more raw materials for biomass in Florida than in other states.

And we also learned that wind, hydro, and geothermal power plants are geographically affected, but other power plants are more affected by weather, surrounding facilities, and the presence or absence of raw materials than by geography.

We thought Texas' weather was hitting hard in four seasons, so we predicted that there would be a lot of solar power plants in Texas, but contrary to our expectations, Texas had fewer solar power plants than other states. So we searched on the Internet to get the answer which states produce the most solar energy, and found that Texas was the second-largest solar energy production in the United States after California. So we wanted to find out and study more later whether it was because the capacity of Texas' energy plants was larger than that of other states, or whether we got them wrong, or whether our dataset does not include all the power plants in the USA.

4. Conclusion

We analyzed the data regarding the energy production for different sources, mainly renewable energy, nuclear energy, and fossil fuel to compare the production amount and carbon dioxide, and how the renewable sources play a role as an alternative energy source for fossil fuel. For the production amount, we compared the top 5 energy production countries with the 8 different energy sources to see the trend of the energy source, and we observed that the fossil fuels such as coal and gas have been constantly used since 1985 while renewable energy sources such as wind and solar are kept increasing but they still have low percentage comparing to others. We also compared the top 4 eco-friendly countries that use a high percentage of renewable energy sources using the trend of the carbon dioxide amount, and it clearly showed that higher percentages of the renewable energy sources produce lower carbon dioxide amounts. From these results, we wanted to know the relationship between the number of energy plants and total energy capacity and we found out that increasing the number of energy plants does not affect the energy capacity due to higher energy density for fossil fuel while renewable energy plants have a low total capacity of energy even it has the greatest number of plants. Based on the analysis, we expected that renewable energy sources are beneficial for the eco-friendly and it can play a role as fossil fuel as alternative energy but there might be some limitations due to the environmental factors and plant capacities due to the lower energy density.

We found many interesting observations through the data analysis. However, it does not provide enough evidence to confirm that renewable energy sources can be used instead of fossil fuel, so we could analyze the energy density of each energy source to compare it with total capacity additionally in the future to support the results of the analysis. In addition, we can extend the analysis to the relationship between the environmental factors and renewable energy sources since many countries have different environmental factors such as land, location, and weather so it can be beneficial to explore the limitation of energy capacity by comparing how renewable energy spreads around each continent.

5. Reference

[1] Gustin, G. (2020, December 7). *As beef comes under fire for climate impacts, the industry fights back*. Inside Climate News. Retrieved November 21, 2021, from https://insideclimatenews.org/news/21102019/climate-change-meat-beef-dairy-methane-emissions-california/.

[2] Arbabyazd, L. (2021, August 18). *Renewable vs nuclear energy generation (1965-)*. Kaggle. Retrieved November 23, 2021, from https://www.kaggle.com/donjoeml/energy-consumption-and-generation-in-the-globe.

[3] *Renewables could displace fossil fuels to power the world by 2050, report claims*. Energy Central. (2021, April 24). Retrieved November 23, 2021, from https://energycentral.com/news/renewables-could-displace-fossil-fuels-power-world-2050-report-claims.

[4] Wikimedia Foundation. (2021, November 23). *Cogeneration*. Wikipedia. Retrieved November 24, 2021, from https://en.wikipedia.org/wiki/Cogeneration.

[5]Production and consumption of Energy. (n.d.). Retrieved November 24, 2021, from http://www.stats.gov.cn/english/statisticaldata/yearlydata/yb2004-e/html/7ie.htm.

[6] *Can renewable energy sources replace fossil fuels?* Prysmian Group. (n.d.). Retrieved November 24, 2021, from https://www.prysmiangroup.com/en/insight/sustainability/can-renewable-energy-sources-replace-fossil-fuels.

[7] Ritchie, H., & Roser, M. (2020, May 11). *CO_2 and greenhouse gas emissions*. Our World in Data. Retrieved November 24, 2021, from https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.

[8] *U.S. Energy Information Administration - EIA - independent statistics and analysis*. Where wind power is harnessed - U.S. Energy Information Administration (EIA). (n.d.). Retrieved November 24, 2021, from https://www.eia.gov/energyexplained/wind/where-wind-power-is-harnessed.php.

6. Code

```
# Bringing out the library
library(ggplot2)
library(maps)
library(ggmap)

# recall all the dataset
global <- read.csv("globalpowerplant.csv")
product <- read.csv("production.source.csv")
consume <- read.csv("consumer.source.csv")
nuclear <- read.csv("nuclear.csv")
co2 <- read.csv("co2.data.csv")

########## Plot 1 ##########
# Plot1 for world energy production trend

# Bring out the data index of world
world <- which(product$Entity == "World")

# Bring out the data index for each energy resource
coal_world <- product$Coal....electricity.[world]
gas_world <- product$Gas....electricity.[world]
hydro_world <- product$Hydro....electricity.[world]
solar_world <- product$Solar....electricity.[world]
wind_world <- product$Wind....electricity.[world]
oil_world <- product$Oil....electricity.[world]
other_world <- product$Other.renewables....electricity.[world]
nuclear_world <- product$Nuclear....electricity.[world]

# Making the data frame for coal from 1985 to 2020
value <- coal_world
coal_world_prod <- data.frame(value)
coal_world_prod$year <- 1985:2020
coal_world_prod$name[1:36] <- "coal"

# Making the data frame for nuclear from 1985 to 2020
value <- nuclear_world
nuclear_world_prod <- data.frame(value)
nuclear_world_prod$year <- 1985:2020
nuclear_world_prod$name[1:36] <- "nuclear"

# Making the data frame for gas from 1985 to 2020
value <- gas_world
gas_world_prod <- data.frame(value)
gas_world_prod$year <- 1985:2020
gas_world_prod$name[1:36] <- "gas"

# Making the data frame for hydro from 1985 to 2020
value <- hydro_world
hydro_world_prod <- data.frame(value)
```

```
hydro_world_prod$year <- 1985:2020
hydro_world_prod$name[1:36] <- "hydro"

# Making the data frame for solar from 1985 to 2020
value <- solar_world
solar_world_prod <- data.frame(value)
solar_world_prod$year <- 1985:2020
solar_world_prod$name[1:36] <- "solar"

# Making the data frame for wind from 1985 to 2020
value <- wind_world
wind_world_prod <- data.frame(value)
wind_world_prod$year <- 1985:2020
wind_world_prod$name[1:36] <- "wind"

# Making the data frame for oil from 1985 to 2020
value <- oil_world
oil_world_prod <- data.frame(value)
oil_world_prod$year <- 1985:2020
oil_world_prod$name[1:36] <- "oil"

# Making the data frame for other resources from 1985 to 2020
value <- other_world
other_world_prod <- data.frame(value)
other_world_prod$year <- 1985:2020
other_world_prod$name[1:36] <- "other"

# row-bind all the dataframe made above
new_production <- rbind(coal_world_prod, gas_world_prod, hydro_world_prod,
                        oil_world_prod, other_world_prod, solar_world_prod,
                        wind_world_prod, nuclear_world_prod)

# Plot (1)
# The production percentage of each energy source of world since 1985
ggplot(data = new_production) +
  geom_smooth(mapping = aes(x = year, y = value, color = name),
              method = 'loess', formula = 'y ~ x') +
  labs( x = "Year", y= "Percentage(%)", color = "Energy",
        title = "The Production Percentage of Each Energy Source
        All Around the World Since 1985")

########## Plot 2 ##########
# Plot 2 for the production trend of each energy source for top five countries
# where produce the most energy all around the world since 1985

# Bring out the data index of each contries (top five countries)
prod_5 <- which(product$Code == "USA" & product$Year >= 1985 &
                               product$Year <= 2020 |
                               product$Code == "CHN" & product$Year >= 1985 &
                               product$Year <= 2020 |
                               product$Code == "IND" & product$Year >= 1985 &
```

```
                                product$Year <= 2020 |
                                product$Code == "RUS" & product$Year >= 1985 &
                                product$Year <= 2020 |
                                product$Code == "JPN" & product$Year >= 1985 &
                                product$Year <= 2020)

# Make the data frame for country name
df_prod_5 <- data.frame(product$Code[prod_5])

# Bring out the data for each energy resource from index we figured above
coal_prod_5 <- product$Coal....electricity.[prod_5]
gas_prod_5 <- product$Gas....electricity.[prod_5]
hydro_prod_5 <- product$Hydro....electricity.[prod_5]
solar_prod_5 <- product$Solar....electricity.[prod_5]
wind_prod_5 <- product$Wind....electricity.[prod_5]
oil_prod_5 <- product$Oil....electricity.[prod_5]
nuclear_prod_5 <- product$Nuclear....electricity.[prod_5]
other_prod_5 <- product$Other.renewables....electricity.[prod_5]


# Making the data frame for coal
coal_prod_5 <- data.frame(coal_prod_5)
coal_prod_5$year <- rep(1985:2020)
coal_prod_5$name[1:180] <- "coal"
names(coal_prod_5) <- c("value", "year", "name")

# Making the data frame for gas
gas_prod_5 <- data.frame(gas_prod_5)
gas_prod_5$year <- rep(1985:2020)
gas_prod_5$name[1:180] <- "gas"
names(gas_prod_5) <- c("value", "year", "name")

# Making the data frame for hydro
hydro_prod_5 <- data.frame(hydro_prod_5)
hydro_prod_5$year <- rep(1985:2020)
hydro_prod_5$name[1:180] <- "hydro"
names(hydro_prod_5) <- c("value", "year", "name")

# Making the data frame for solar
solar_prod_5 <- data.frame(solar_prod_5)
solar_prod_5$year <- rep(1985:2020)
solar_prod_5$name[1:180] <- "solar"
names(solar_prod_5) <- c("value", "year", "name")

# Making the data frame for nuclear
nuclr_prod_5 <- data.frame(nuclear_prod_5)
nuclr_prod_5$year <- rep(1985:2020)
nuclr_prod_5$name[1:180] <- "nuclear"
names(nuclr_prod_5) <- c("value", "year", "name")

# Making the data frame for wind
```

```
wind_prod_5 <- data.frame(wind_prod_5)
wind_prod_5$year <- rep(1985:2020)
wind_prod_5$name[1:180] <- "wind"
names(wind_prod_5) <- c("value", "year", "name")

# Making the data frame for oil
oil_prod_5 <- data.frame(oil_prod_5)
oil_prod_5$year <- rep(1985:2020)
oil_prod_5$name[1:180] <- "oil"
names(oil_prod_5) <- c("value", "year", "name")

# Making the data frame for other resources
other_prod_5 <- data.frame(other_prod_5)
other_prod_5$year <- rep(1985:2020)
other_prod_5$name[1:180] <- "other"
names(other_prod_5) <- c("value", "year", "name")

# bind all the dataframe made above and make new dataframe we'll use
new_df_prod_5 <- rbind(coal_prod_5, gas_prod_5, hydro_prod_5,
                         oil_prod_5, other_prod_5, solar_prod_5,
                         wind_prod_5, nuclr_prod_5)
new_prod_5 <- cbind(new_df_prod_5, df_prod_5)

# Rename the column's name
names(new_prod_5) <- c("value", "year", "name", "country")

# Plot (2)
# The production percentage of each energy source for the top five countries
# the most production energy all aroudn the world
ggplot(data = new_prod_5) +
  geom_smooth(mapping = aes(x = year, y = value, color = country),
              method = 'lm', formula = 'y ~ x') +
  facet_wrap(~name, nrow = 2) +
  labs( x = "Year", y= "Percentage(%)", color = "Country",
        title = "The Top Five Energy Production Countries' Energy Trend
        since 1985")


########## Plot 3 ##########
# By using the data frame we figured out above, we compare the differences
# between production and consumption of energy

# bring out the index which is suitable for conditions
# this index for consumer energy rate for top 5 countries
cons_5 <- which(consume$Code == "USA" &consume$Year == 2019|
                  consume$Code == "CHN"  &consume$Year == 2019|
                    consume$Code == "IND" &consume$Year == 2019|
                  consume$Code == "RUS" &consume$Year == 2019|
                    consume$Code == "JPN" &consume$Year == 2019)

# this index for production energy rate for top 5 countries
```

```
prod_5 <- which(product$Code == "USA"&product$Year == 2019 |
                  product$Code == "CHN"&product$Year == 2019|
                 product$Code == "IND"&product$Year == 2019 |
                  product$Code == "RUS"&product$Year == 2019|
                 product$Code == "JPN"&product$Year == 2019)

# bring out the energy rate by using the index from dataset
prod_value <- product$Code[prod_5]
cons_value <- consume$Code[cons_5]

# make the new data frame with the data we figured out above
df_prod_5 <- data.frame(prod_value)
df_cons_5 <- data.frame(cons_value)

# Bring out the data index for each energy resource for production
coal_prod_5 <- product$Coal....electricity.[prod_5]
gas_prod_5 <- product$Gas....electricity[prod_5]
hydro_prod_5 <- product$Hydro....electricity.[prod_5]
solar_prod_5 <- product$Solar....electricity.[prod_5]
wind_prod_5 <- product$Wind....electricity.[prod_5]
oil_prod_5 <- product$Oil....electricity.[prod_5]
nuclear_prod_5 <- product$Nuclear....electricity.[prod_5]
other_prod_5 <- product$Other.renewables....electricity.[prod_5]

# Bring out the data index for each energy resource for consumption
coal_cons_5 <- consume$Coal....sub.energy.[cons_5]
gas_cons_5 <- consume$Gas....sub.energy.[cons_5]
hydro_cons_5 <- consume$Hydro....sub.energy.[cons_5]
solar_cons_5 <- consume$Solar....sub.energy.[cons_5]
wind_cons_5 <- consume$Wind....sub.energy.[cons_5]
oil_cons_5 <- consume$Oil....sub.energy.[cons_5]
nuclear_cons_5 <- consume$Nuclear....sub.energy.[cons_5]
other_cons_5 <- consume$Other.renewables....sub.energy.[cons_5]

# Production
# Making the data frame for coal
value <- coal_prod_5
coal_prod_5 <- data.frame(value)
coal_prod_5$name <- "coal"
# Combining the new data frame with the the data frame for country name
coal_prod_5 <- cbind(coal_prod_5, df_prod_5)

# Making the data frame for gas
value <- gas_prod_5
gas_prod_5 <- data.frame(value)
gas_prod_5$name <- "gas"
gas_prod_5 <- cbind(gas_prod_5, df_prod_5)

# Making the data frame for hydro
value <- hydro_prod_5
hydro_prod_5 <- data.frame(value)
```

```
hydro_prod_5$name <- "hydro"
hydro_prod_5 <- cbind(hydro_prod_5, df_prod_5)

# Making the data frame for solar
value <- solar_prod_5
solar_prod_5 <- data.frame(value)
solar_prod_5$name <- "solar"
solar_prod_5 <- cbind(solar_prod_5, df_prod_5)

# Making the data frame for solar
value <- nuclear_prod_5
nuclear_prod_5 <- data.frame(value)
nuclear_prod_5$name <- "nuclear"
nuclear_prod_5 <- cbind(nuclear_prod_5, df_prod_5)

# Making the data frame for wind
value <- wind_prod_5
wind_prod_5 <- data.frame(value)
wind_prod_5$name <- "wind"
wind_prod_5 <- cbind(wind_prod_5, df_prod_5)

# Making the data frame for oil
value <- oil_prod_5
oil_prod_5 <- data.frame(value)
oil_prod_5$name <- "oil"
oil_prod_5 <- cbind(oil_prod_5, df_prod_5)

# Making the data frame for other resources
value <- other_prod_5
other_prod_5 <- data.frame(value)
other_prod_5$name <- "other"
other_prod_5 <- cbind(other_prod_5, df_prod_5)

# Consumer
# Making the data frame for coal
value <- coal_cons_5
coal_cons_5 <- data.frame(value)
coal_cons_5$name <- "coal"
coal_cons_5 <- cbind(coal_cons_5, df_cons_5)

# Making the data frame for gas
value <- gas_cons_5
gas_cons_5 <- data.frame(value)
gas_cons_5$name <- "gas"
gas_cons_5 <- cbind(gas_cons_5, df_cons_5)

# Making the data frame for hydro
value <- hydro_cons_5
hydro_cons_5 <- data.frame(value)
hydro_cons_5$name <- "hydro"
hydro_cons_5 <- cbind(hydro_cons_5, df_cons_5)
```

```
# Making the data frame for solar
value <- solar_cons_5
solar_cons_5 <- data.frame(value)
solar_cons_5$name <- "solar"
solar_cons_5 <- cbind(solar_cons_5, df_cons_5)

# Making the data frame for solar
value <- nuclear_cons_5
nuclear_cons_5 <- data.frame(value)
nuclear_cons_5$name <- "nuclear"
nuclear_cons_5 <- cbind(nuclear_cons_5, df_cons_5)

# Making the data frame for wind
value <- wind_cons_5
wind_cons_5 <- data.frame(value)
wind_cons_5$name <- "wind"
wind_cons_5 <- cbind(wind_cons_5, df_cons_5)

# Making the data frame for oil
value <- oil_cons_5
oil_cons_5 <- data.frame(value)
oil_cons_5$name <- "oil"
oil_cons_5 <- cbind(oil_cons_5, df_cons_5)

# Making the data frame for other resources
value <- other_cons_5
other_cons_5 <- data.frame(value)
other_cons_5$name <- "other"
other_cons_5 <- cbind(other_cons_5, df_cons_5)

# bind all the dataframe made above for production
new_prod_5 <- rbind(coal_prod_5, gas_prod_5, hydro_prod_5, oil_prod_5,
                    other_prod_5, solar_prod_5, wind_prod_5, nuclear_prod_5)

# make the new column to indicate "production"
new_prod_5$names <- rep("production")

# bind all the dataframe made above for consumption
new_cons_5 <- rbind(coal_cons_5, gas_cons_5, hydro_cons_5, oil_cons_5,
                    other_cons_5, solar_cons_5, wind_cons_5, nuclear_cons_5)

# make the new column to indicate "consumption"
new_cons_5$names <- rep("consumer")

# Make the point plot for production
ggplot(data = new_prod_5) +
  geom_point (mapping = aes(x = prod_value, y = value, color = name)) +
  labs(x = "Country", y = "Percentage(%)",
       title = "Energy Production for Top 5 Countries", color = "Energy")
```

```
# Make the point plot for consumption
ggplot(data = new_cons_5) +
  geom_point(mapping = aes(x = cons_value, y = value, color = name))+
  labs(x = "Country", y = "Percentage(%)",
       title = "Energy Consumption for Top 5 Countries", color = "Energy")

########## Plot 4 ##########
# Bring out the index where fossil fuel production rate is less than 50 in 2018
below_50 <- which(nuclear$Fossil.fuels....sub.energy. <= 50 &
                    nuclear$Year == 2018)

# Find out the countries name and make the data frame
country <- nuclear$Entity[below_50]
country <- data.frame(country)

# Find out the fossil fuel and make the data frame
fossil_50 <- nuclear$Fossil.fuels....sub.energy.[below_50]
fossil_50 <- data.frame(fossil_50)
# Name it that it is Fossil Fuel
fossil_50$name <- rep("Fossil Fuel")
# Change the column name
names(fossil_50) <- c("value", "name")
# Make the new column for country name
fossil_50 <- cbind(fossil_50, country)

# Find out the renewable and make the data frame
renew_50 <- nuclear$Renewables....sub.energy.[below_50]
renew_50 <- data.frame(renew_50)
renew_50$name <- rep("renewable")
names(renew_50) <- c("value", "name")
# Make the new column for country name
renew_50 <- cbind(renew_50, country)

# Find out the nuclear and make the data frame
nuclear_50 <- nuclear$Nuclear....sub.energy.[below_50]
nuclear_50 <- data.frame(nuclear_50)
nuclear_50$name <- rep("Nuclear")
names(nuclear_50) <- c("value", "name")
# Make the new column for country name
nuclear_50 <- cbind(nuclear_50, country)

# Make the data frame (final version)
new_new <- rbind(fossil_50, renew_50, nuclear_50)

# Make the bar plot which indicates energy production ratio
ggplot()+
  geom_bar(data = new_new, mapping = aes(x = country, y = value,
                                         fill = as.factor(name)),
           position = "fill" , stat="identity") +
  labs( x = "Country", y = "Ratio (1=100%)", fill = "Energy",
        title = "The Ratio of Energy for 4 Countries")
```

```
########## Plot 5 ##########
# This plot is made in order to look over the trend of the CO2 amount by year
# We're going to compare the amount of
# CO2 eco-friendly countries to top five countries

# bring out the index of eco-friendly countries between 1985 and 2020
environ_ctr <- which(co2$year >= 1985 & co2$country == "Iceland" |
                       co2$year >= 1985 & co2$country == "Norway" |
                       co2$year >= 1985 & co2$country == "Sweden" |
                       co2$year >= 1985 & co2$country == "Switzerland")

# bring out the index of top five energy production countries
# between 1985 and 2020
top_5_ctr <- which( co2$year >= 1985 & co2$country == "United States" |
                      co2$year >= 1985 & co2$country == "China" |
                      co2$year >= 1985 & co2$country == "India" |
                      co2$year >= 1985 & co2$country == "Japan" |
                      co2$year >= 1985 & co2$country == "Russia")

# Bring out the amount of CO2 by using the eco-friendly index
tb_environ <- co2$co2[environ_ctr]

# Bring out country by using the eco-friendly index
tb_environ_cty <- co2$country[environ_ctr]

# Bring out year by using the eco-friendly index
tb_environ_yr <- co2$year[environ_ctr]

# column bind with the amount of CO2, country, and year
new_tb_environ <- cbind(tb_environ, tb_environ_cty, tb_environ_yr)

# make the data frame
df_environ <- data.frame(new_tb_environ)

# rename the column name
names(df_environ) <- c("co2", "country", "year")

# Bring out the amount of CO2 by using the top 5 production index
tb_top_5 <- co2$co2[top_5_ctr]

# Bring out country by using the top 5 production index
tb_top_cty <- co2$country[top_5_ctr]

# Bring out year by using the top 5 production index
tb_top_yr <- co2$year[top_5_ctr]

# column bind with the amount of CO2, country, and year
new_tb_top <- cbind(tb_top_5, tb_top_cty, tb_top_yr)

# make the data frame
```

```
df_top <- data.frame(new_tb_top)

# rename the column name
names(df_top) <- c("co2", "country", "year")

# Make the line plot included point plot for eco-friendly countries
ggplot(data = df_environ) +
  geom_line(mapping = aes(x = as.numeric(year), y = as.numeric(co2),
                          color = country)) +
  geom_point(mapping = aes(x = as.numeric(year), y = as.numeric(co2),
                           color = country), cex = 0.9) +
  labs(x = "Year", y = "CO2 (kt)",
       title = "The Trend of the CO2 Amount by Year")

# Make the line plot included point plot for top five countries
ggplot(data = df_top) +
  geom_line(mapping = aes(x = as.numeric(year), y = as.numeric(co2),
                          color = country)) +
  geom_point(mapping = aes(x = as.numeric(year), y = as.numeric(co2),
                           color = country), cex = 0.9) +
  labs(x = "Year", y = "CO2 (kt)",
       title = "The Trend of the CO2 Amount by Year")

########## Plot 6 (1) ##########
# Plot 6 (1) for the number of power plants all around the world

# by using table we can figure out the number of each plants
plant_number <- table(global$primary_fuel)

# In order to make the plot decreasing sort
plant_number <- sort(plant_number, decreasing = FALSE)

# Making the data frame for it
plant_number <- data.frame(plant_number)

# Making the plots by using ggplot, and it indicates the figures
ggplot(plant_number) +
  geom_bar(mapping = aes(x = Freq, y = Var1, fill = Var1), stat = "identity") +
  labs(x = "Number of Plants", y = "Energy Source", fill = "Energy Source",
       title = "The Number of Plants All Around the World") +
  geom_text(aes(x = Freq, y = Var1, label = Freq))

########## Plot 6 (2) ##########
# Plot 6 (2) for the total energy capacity which can generate
# from each energy plant all around the world

# figure out the sum capacity for each energy plant
sum_solar <- sum(global$capacity_mw[which(global$primary_fuel == "Solar")])
sum_hydro <- sum(global$capacity_mw[which(global$primary_fuel == "Hydro")])
sum_wind <- sum(global$capacity_mw[which(global$primary_fuel == "Wind")])
sum_coal <- sum(global$capacity_mw[which(global$primary_fuel == "Coal")])
```

```
sum_oil <- sum(global$capacity_mw[which(global$primary_fuel == "Oil")])
sum_gas <- sum(global$capacity_mw[which(global$primary_fuel == "Gas")])
sum_biomass <- sum(global$capacity_mw[which(global$primary_fuel == "Biomass")])
sum_waste <- sum(global$capacity_mw[which(global$primary_fuel == "Waste")])
sum_nuclear <- sum(global$capacity_mw[which(global$primary_fuel == "Nuclear")])
sum_geother <- sum(global$capacity_mw[which(global$primary_fuel ==
                                              "Geothermal")])
sum_storage <- sum(global$capacity_mw[which(global$primary_fuel == "Storage")])
sum_other <- sum(global$capacity_mw[which(global$primary_fuel == "Other")])
sum_cogen <- sum(global$capacity_mw[which(global$primary_fuel ==
                                            "Cogeneration")])
sum_petcoke <- sum(global$capacity_mw[which(global$primary_fuel == "Petcoke")])
sum_wnt <- sum(global$capacity_mw[which(global$primary_fuel ==
                                          "Wave and Tidal")])


# make the dataframe for the value of capacity
df_sum <- data.frame(c(sum_solar, sum_hydro, sum_wind, sum_coal, sum_oil,
                       sum_gas, sum_biomass, sum_waste, sum_nuclear,
                       sum_geother, sum_storage, sum_other, sum_cogen,
                       sum_petcoke, sum_wnt))

# round up to the second decimal place.
df_sum <- round(df_all, 2)

# make the dataframe for 'energy source name' by the order of the sum dataframe
df_name <- data.frame(c("Solar", "Hydro", "Wind", "Coal", "Oil", "Gas",
                        "Biomass", "Waste", "Nuclear", "Geothermal",
                        "Storage", "Other", "Cogeneration", "Petcoke",
                        "Wave and Tidal"))

# make the new dataframe by using cbind
new_sum <- cbind(df_sum, df_name)

# change the column name
names(new_sum) <- c("value", "energy")

# make the barplot for total capacity of each plant all around the world
ggplot(new_sum) +
  geom_bar(mapping = aes(x = value, y = reorder(energy, value), fill = energy),
           stat = "identity") +
  labs(x = "Total Capacity(mw)", y = "Energy Source", fill = "Energy Source",
       title = "The Total Capacity of Each Plant All Around the World") +
  geom_text(aes(x = value, y = energy, label = value), cex = 3)


########## Plot 7 ##########
# Which energy plant produces the most energy?

# barplot for energy plant produces the energy
ggplot(global) +
```

```
  geom_boxplot(mapping = aes(x = primary_fuel, y = capacity_mw,
                             color = primary_fuel)) +
  labs(x = "Energy Source", y = "Capacity (mw)",
       title = "The Energy Plant Capacity of World") +
  scale_y_continuous(limits = c(1, 2000)) +
  theme(axis.text.x = element_text(size= 7))

########## Plot 8 ##########
# This plot is in order to look the spread of energy plants in the States

# Bringing out USA
us_map <- map_data("state")

# Making the dataframe for USA
production_usa <- global[which(global$country == "USA"), ]

# Making the dataframe for longitude and latitude (map plot)
production_usa <- production_usa[which(production_usa$longitude <= -65 &
                                         production_usa$longitude >= -130), ]
production_usa <-  production_usa[which(production_usa$latitude <=50 &
                                          production_usa$latitude >= 23), ]

# Making the dataframe for renewable energy (scatter plot)
production_usa <- production_usa[which(production_usa$primary_fuel == "Biomass"|
                                         production_usa$primary_fuel ==
"Cogeneration"|
                                         production_usa$primary_fuel ==
"Geothermal"|
                                         production_usa$primary_fuel == "Hydro"|
                                         production_usa$primary_fuel == "Wind"|
                                         production_usa$primary_fuel ==
"Solar"), ]


# Making the map plot for the spread of renewable energy power plant.
ggplot(us_map) +
  geom_polygon(mapping = aes(x = long, y = lat, group = group)) +
  coord_quickmap() +
  geom_point(data = production_usa,
             mapping = aes(x = longitude, y = latitude, color = primary_fuel),
             cex = 0.9) +
  labs(x = "Longitude", y = "Latitude", color = "Renewable energy",
       title = "The Spread of Renewable Energy Power Plant")

# Making the map plot for the spread of renewable energy power plant
# divided by the energy source
ggplot(us_map) +
  geom_polygon(mapping = aes(x = long, y = lat, group = group)) +
  coord_quickmap() +
  geom_point(data = production_usa,
             mapping = aes(x = longitude, y = latitude, color = primary_fuel),
```

```
            cex = 0.7) +
  facet_wrap(~primary_fuel, nrow = 3) +
  labs(x = "Longitude", y = "Latitude", color = "Renewable energy",
       title = "The Spread of Renewable Energy Power Plant")
```