CDP Technical Task - Junior Data Analyst

Isabella Morgante

This report documents the approach used to explore data relating to the electricity generation for the top 5 power generating countries by source and over time. The code used to conduct final analyses is displayed below, all historic versions of the code can be found via this projects repository at https://github.com/iz-morgante/CDP_Task.

Inspecting the Data

The dataset was first inspected and assessed for quality and then each analysis question was addressed. To load the data, use the appropriate file path.

```
cdp_data <- read.csv(file = "../Data/Electricity generation by source - top 5.csv", stringsAsFactors = '
```

Examine content and structure of the dataset.

```
head(cdp_data) # View first few rows
```

```
##   Year Country Units  Source Electricity.Generation..GWh.
## 1 1990   China   GWh Biomass                          NA
## 2 1995   China   GWh Biomass                        2897
## 3 2000   China   GWh Biomass                        2421
## 4 2005   China   GWh Biomass                        5200
## 5 2010   China   GWh Biomass                       24839
## 6 2015   China   GWh Biomass                       52743
```

```
str(cdp_data) # View data structure
```

```
## 'data.frame':    420 obs. of  5 variables:
##  $ Year                          : int  1990 1995 2000 2005 2010 2015 2018 1990 1995 2000 ...
##  $ Country                       : Factor w/ 6 levels "China","India",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ Units                         : Factor w/ 1 level "GWh": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Source                        : Factor w/ 10 levels "Biomass","Coal",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Electricity.Generation..GWh.: int  NA 2897 2421 5200 24839 52743 90608 NA NA 1278 ...
```

The output shows that the data contains information about the electricity generation by generation source from 1990-2019. There are 6 *Country* categories representing the top 5 power generating countries as well as global values. There are 10 types of electricity generation listed under the *Sources* column. The electricity generation units are consistent throughout, GWh, as shown bythe *Units* column having only one factor, GWh.

The electricity generation data does contain missing values (NAs). For this exercise, it has been assumed that missing generation values are due to a country not generating electricity by a particular source for that year. Following this assumption, all rows containing missing electricity generation data have been changed to zero for all analyses.

```
names(which(colSums(is.na(cdp_data))>0)) # List columns containing NAs
```

```
## [1] "Electricity.Generation..GWh."
```

```
# Change column name to avoid dots and remove NAs
cdp_data <- rename(cdp_data, ElectricityGeneration_GWh = Electricity.Generation..GWh.) %>% replace(is.na
```

Question 1: Total Electricity Generation

Calculate the total electricity generation per year. How does this change overtime?

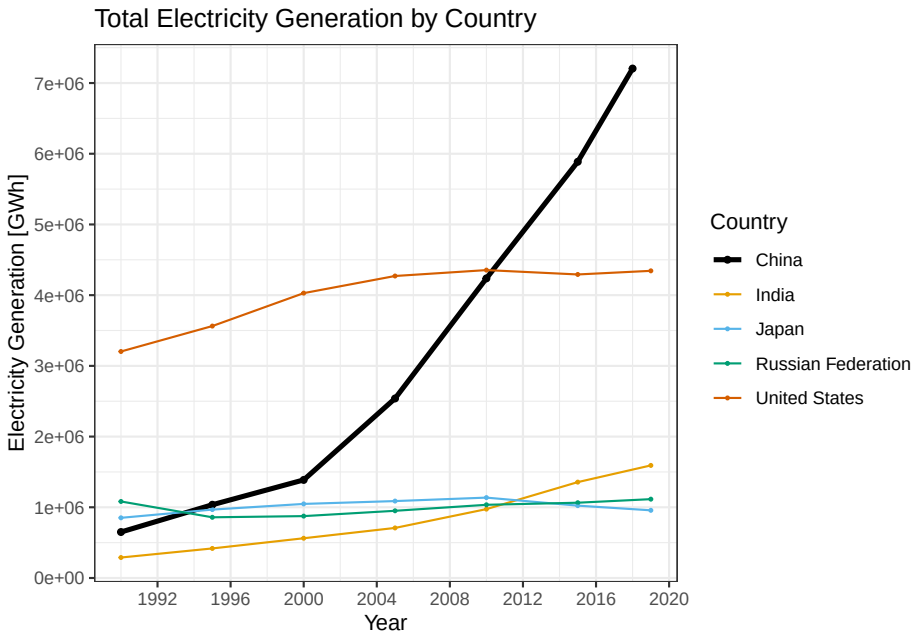
To find the total electricity generated by each country per year, the amount of electricity generation from each source is summed for every year.

```
total_electricity <- cdp_data  %>%
  group_by(Year, Country) %>%
  summarise(Total_ElectricityGeneration_GWh = sum(ElectricityGeneration_GWh))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

The total electricity generation by top 5 power generating countries over 30 years show some interesting tends (see figure below). Japan and the Russian Federation show little change in the total amount of electricity generation over time and the United States' generation appears to have plateaued since 2005. Both India and China have seen an increase in electricity generation.

From the figure, China has shown the most dramatic increase in their electricity generation, with the generation rate increasing. For this reason, I have selected China as my country of choice for further investigation.



Question 2: Fossil Fuel Generation

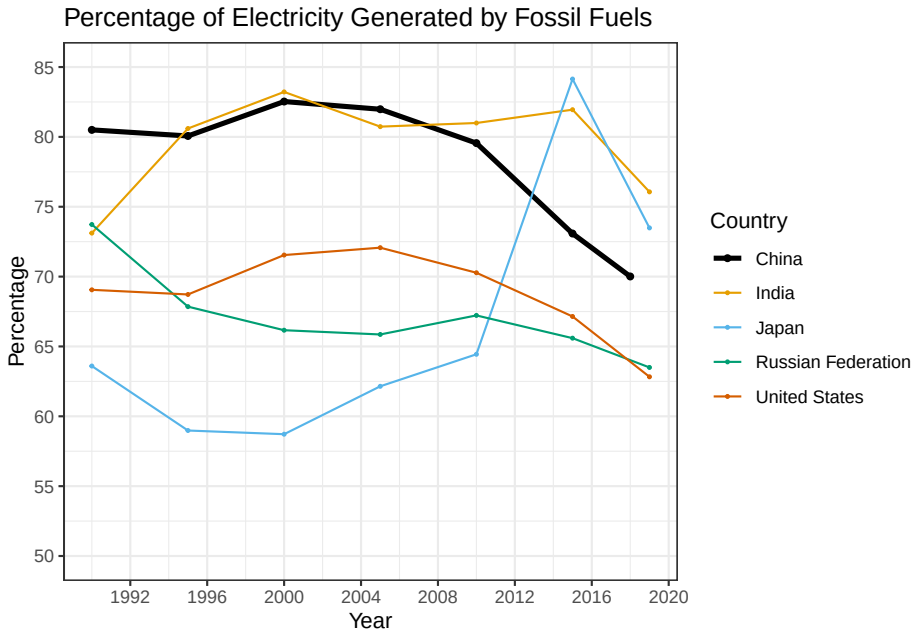
What percentage of electricity is generated from fossil fuel sources per year?

The amount of electricity that is generated by fossil fuels was determined by summing the electricity generated by oil, coal and natural gas for each year. Then the total amount of electricity generation per year for each country was added to the fossil fuel generation dataset. Finally, the percentage of electricity generated by fossil fuels with respect to total electricity generation was found by dividing by fossil fuel generation by total generation. The results may be seen in the figure below.

```
FF_electricity <- filter(cdp_data, Source == "Coal"| Source =="Natural gas" | Source == "Oil") %>% # fi
  group_by(Year, Country) %>%
  summarise(FossilFuel_ElectricityGeneration_GWh = sum(ElectricityGeneration_GWh)) # sum generation val
```

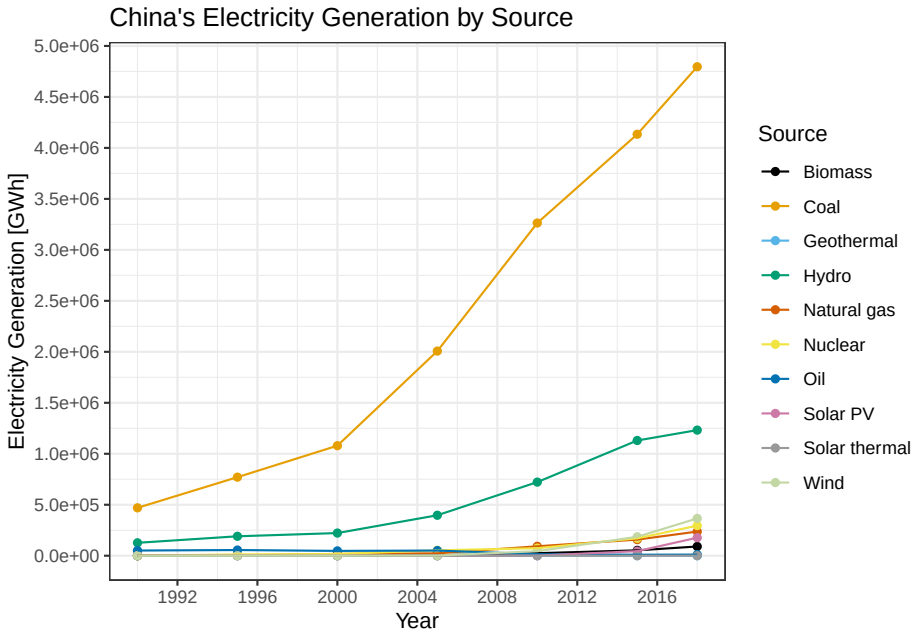
```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
FF_Total <- full_join(FF_electricity, total_electricity, by = c('Year', 'Country')) %>% # join total ge
  mutate(
    Percent_FF = FossilFuel_ElectricityGeneration_GWh / Total_ElectricityGeneration_GWh * 100 # calcula
  )
```



All the countries generate over 50% of their electricity using fossil fuels, with a general decline in their use since 2015. Patterns of fossil fuel usage over time differ between countries. China and the US show a similar trend, with fossil fuel usage peaking in the early 2000s with a consistent decline in usage until present day. Russia's usage has nearly always been declining since 1990, whereas Japan has been increasing usage until peaking in 2015 with the highest recorded proportion of fossil fuel electricity.

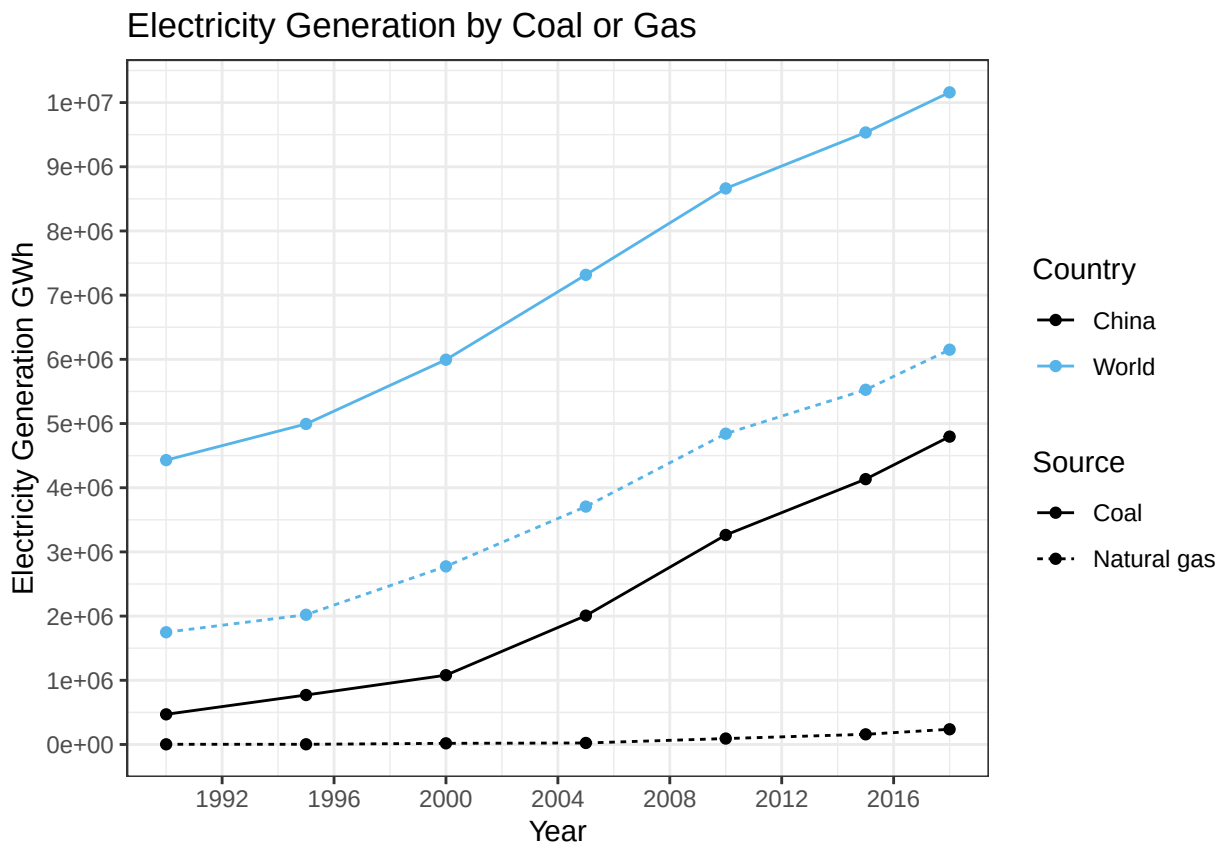
Focusing on China, its electricity generation by fossil fuels has been declining since 2000. However, in comparison to the other countries, its generation of electricity by fossil fuel remains high. A deeper look at China's electricity generation by source (see below figure) reveal that its electricity is predominantly generated with coal, although the use of natural gas appears to be growing.



Question 3: Coal Generation

For the most recent year, what percentage of world coal generation do the top five countries represent? How does this compare to natural gas?

The most recent year recorded is 2019. Generation figures are provided for Japan, US, Russia and India in 2019, however, for the global values and China, the most recent generation figures are from 2018. The graph for electricity generation with coal or natural gas globally and by China can be seen below.

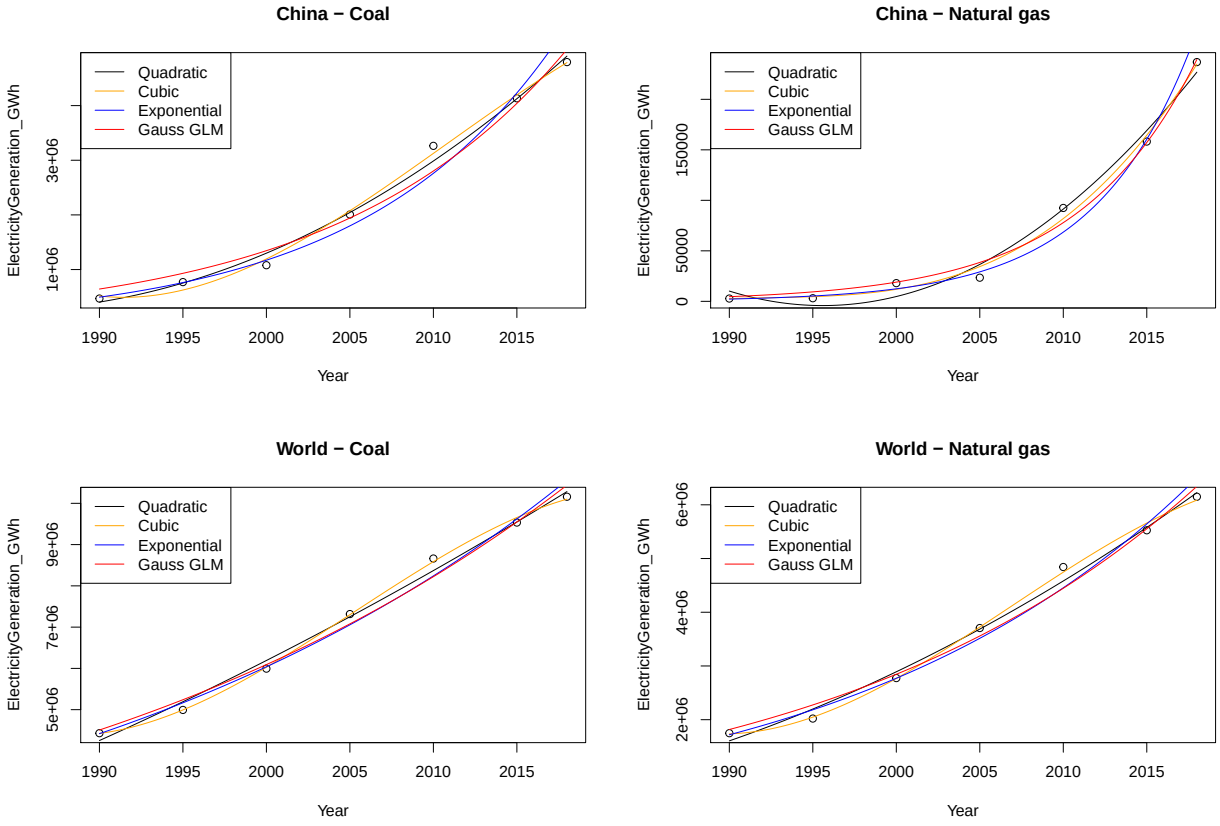


The plot shows a similar trend for each country and each generation source, i.e. that electricity generation for the given sources is increasing at an increasing rate. To form predictive values for electricity generation by coal and gas for the World and China in 2019, models were fitted to the data and the best fitting model was used to predict generation values in 2019.

Four mathematical models were fitted: 2nd and 3rd order polynomials, an allometric curve and a Generalized Linear Model (GLM) with Gaussian distribution and a log link function. Each model was fitted to the data and made predictions for 2019 values. To select the best fitting model, the model with the smallest Akaike information criterion (AIC) was chosen. An example of the fitted models can be seen in the figure below.

```
cg_2019 <- filter(cdp_data, Source == "Coal" | Source=="Natural gas", Year == 2019) # Filter for only 2

for (country in c("China", "World")){
  for (source in c("Coal", "Natural gas")){
    data2Fit <- filter(cdp_data, Source == source, Country == country)

    # Fit models
    QuadFit <- lm(ElectricityGeneration_GWh ~ poly(Year,2), data = data2Fit)
    CubFit <- lm(ElectricityGeneration_GWh ~ poly(Year,3), data = data2Fit)
    ExpFit <- lm(log(ElectricityGeneration_GWh) ~ log(Year), data = data2Fit)
    GLMFit <- glm(ElectricityGeneration_GWh ~ log(Year), data = data2Fit, family = gaussian(link = "log
    #NLS_Fit <- nlsLM(ElectricityGeneration_GWh ~ a * Year^b, data = data2Fit, start = list(a = exp(coe

    # Plot the model fits
    plot(ElectricityGeneration_GWh ~ Year, data = data2Fit)
    curve(predict(QuadFit, newdata = data.frame(Year = x)), col = "black", add = TRUE)
    curve(predict(CubFit, newdata = data.frame(Year = x)), col = "orange", add = TRUE)
```

```
    curve(exp(predict(ExpFit, newdata = data.frame(Year = x))), col = "blue", add = TRUE)
    curve(predict(GLMFit, newdata = data.frame(Year = x), type = "response"), col = "red", add = TRUE)
    legend("topleft", legend = c("Quadratic", "Cubic", "Exponential", "Gauss GLM"),
           col = c("black", "orange", "blue", "red"),
           lty = c(1, 1, 1))
    title(main = paste(country, "-", source, sep=" "))

    # Find AIC for models and their predicted value for 2019
    AIC_values <- data.frame(Model  = c("QuadFit", "CubFit", "ExpFit", "GLMFit"), AIC = c(AIC(QuadFit),

    # Select best fitting model by minimum AIC value
    best_fit <- AIC_values[which.min(AIC_values$AIC), ]

    # Add predicted value to 2019 data
    cg_2019 <- cg_2019 %>% add_row(Year = 2019, Country = country, Units = "GWh", Source = source, Elec
  }
}
```



Finally, to find the percentage of world coal generation the top five countries represent, the coal generation by each country was summed for each year and divided by the global generation for that year. The same method was used to find the values for natural gas.

```
coal_gas_sums_2019 <- filter(cg_2019, Country != "World") %>%
  group_by(Year, Source) %>%
  summarise(ElectricityGeneration_GWh = sum(ElectricityGeneration_GWh))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
percent_coal_2019 <- filter(coal_gas_sums_2019, Source == "Coal") %>% pull(ElectricityGeneration_GWh) /

percent_gas_2019 <- filter(coal_gas_sums_2019, Source == "Natural gas") %>% pull(ElectricityGeneration_
```

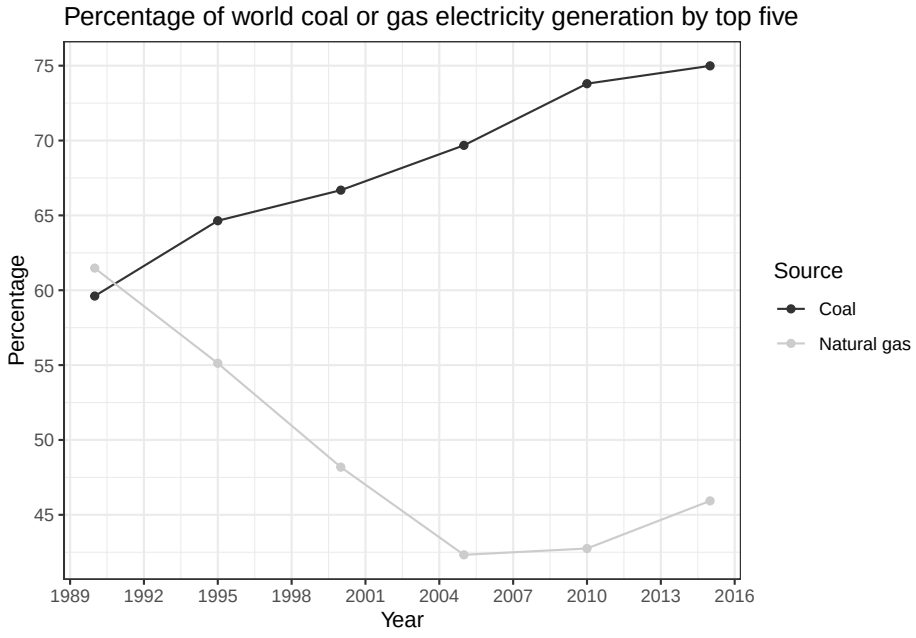
In 2019, the percentage of world coal electricity generation the top five countries represent was 79.29%. In comparison, the percentage gas electricity generation was much lower at 42.26%.

Historically, it appears that percentage of world electricity generation by the top five countries produced by coal is inversely related to production by natural gas (see figure below). Before approx. 1992,a greater proportion of electricity was produced by natural gas, however, this declined as coal production increased over time.

```
# Sum electricity generation by coal and gas for all countries
coal_gas_country <- filter(cdp_data, Source == "Coal" & Country != "World" | Source == "Natural gas" & 
  group_by(Year, Source) %>%
  summarise(ElectricityGeneration_GWh_TopFive = sum(ElectricityGeneration_GWh))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
# Global coal and gas
coal_gas_global <- filter(cdp_data, Source == "Coal" & Country == "World" | Source == "Natural gas" & C
coal_gas_global <- select(coal_gas_global, -c(Country, Units))

# Divide country generation by global value to get proportion and join into single df
coal_gas <- full_join(coal_gas_country, coal_gas_global, by = c('Year','Source')) %>%
  group_by(Year, Source) %>%
  mutate(
    Percent = ElectricityGeneration_GWh_TopFive / ElectricityGeneration_GWh * 100
  )
coal_gas <- rename(coal_gas, ElectricityGeneration_GWh_Global = ElectricityGeneration_GWh)
```



Question 4: Total GHG Lifecycle Emissions

Using data from the previous exercise with the table you have just extracted, calculate the total GHG lifecycle emissions per year for a country of your choice using the provided equation.

The 'Lifecycle greenhouse gas emissions by electricity source' table was extracted from https://en.wikipedia.org/wiki/Emission_intensity using `rvest` package and the appropriate XPath was found by inspecting the source code of the Wikipedia webpage.

```
url <- "https://en.wikipedia.org/wiki/Emission_intensity"
table_XPath <- '//*[@id="mw-content-text"]/div[1]/table[1]'

lifecycle_GHG_emissions <- url %>%
  read_html() %>%
  html_nodes(xpath=table_XPath) %>%
  html_table()
lifecycle_GHG_emissions <-lifecycle_GHG_emissions[[1]]
glimpse(lifecycle_GHG_emissions)
```

```
## Rows: 9
## Columns: 3
## $ Technology                          <chr> "Hydroelectric", "Wind", "Nuclear",~
## $ Description                         <chr> "reservoir", "onshore", "various ge~
## $ `50th percentile  (g CO2-eq/kWhe)` <int> 4, 12, 16, 230, 22, 45, 46, 469, 10~
```

The table contains information about lifecycle greenhouse gas emissions by electricity source. The column names of the lifecycle data were matched to the data provided by CDP and the names of the generation sources were compared between datasets to check for consistency.

```
lifecycle_GHG_emissions <- rename(lifecycle_GHG_emissions, Source = Technology, GHG_EF = `50th percentil

# Check differences between Sorce values
setdiff(lifecycle_GHG_emissions$Source, cdp_data$Source)
```

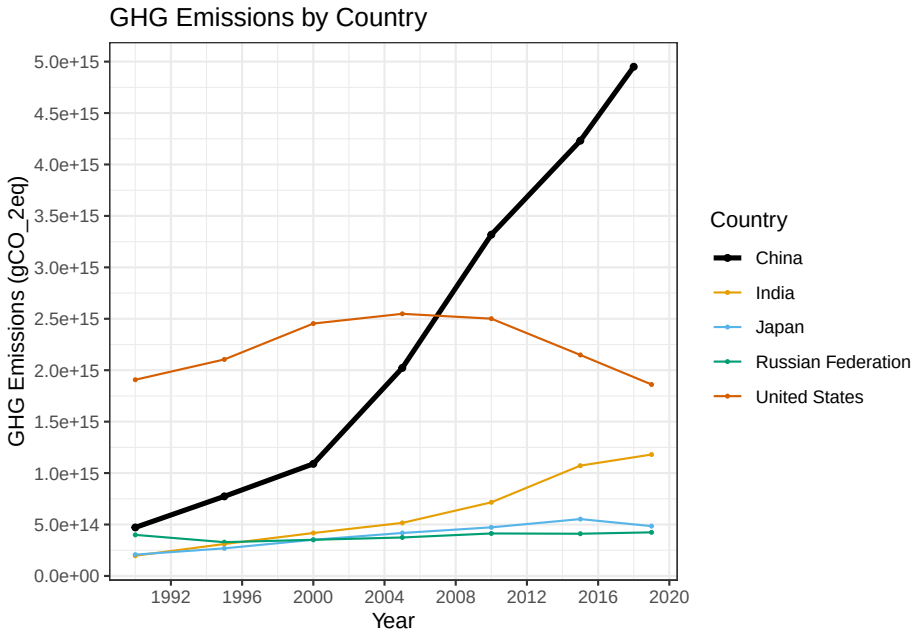
```
## [1] "Hydroelectric"
```

To calculate the total GHG lifecycle emissions per year, the lifecycle emissions data and the CDP electricity data were merged by electricity source. CDP electricity generation was converted from GWh to kWh and the emissions were calculated using:

$$Lifecycle\_GHG\_Emissions(gCO_2eq) = Electricity\_Generation(kWh) * Lifecycle\_GHG\_EF(gCO_2eq/kWh)$$

```
## Add GHG values to cdp data to find emissions
emissions <- full_join(cdp_data, lifecycle_GHG_emissions, by = c('Source')) %>% na.omit() %>% # NAs wil
  mutate(
    ElectricityGeneration_kWh = ElectricityGeneration_GWh * 1e6, #Convert from GWh to kWh
    GHG_Emissions = ElectricityGeneration_kWh * GHG_EF # Calculate emissions
  )

total_emissions <- emissions %>%
  group_by(Year, Country) %>%
  summarise(Total_GHG_Emissions = sum(GHG_Emissions))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```



There is a marked difference between the GHG electricity generation emissions from China in comparison to the other countries, as can be seen from the above figure. China's GHG emissions are over double that of any other country in 2018, and their GHG emissions have been increasing at an increasing rate since 1990. The USA's GHG emissions have been slowly declining and Japan and Russia's emissions have not shown great change over time.