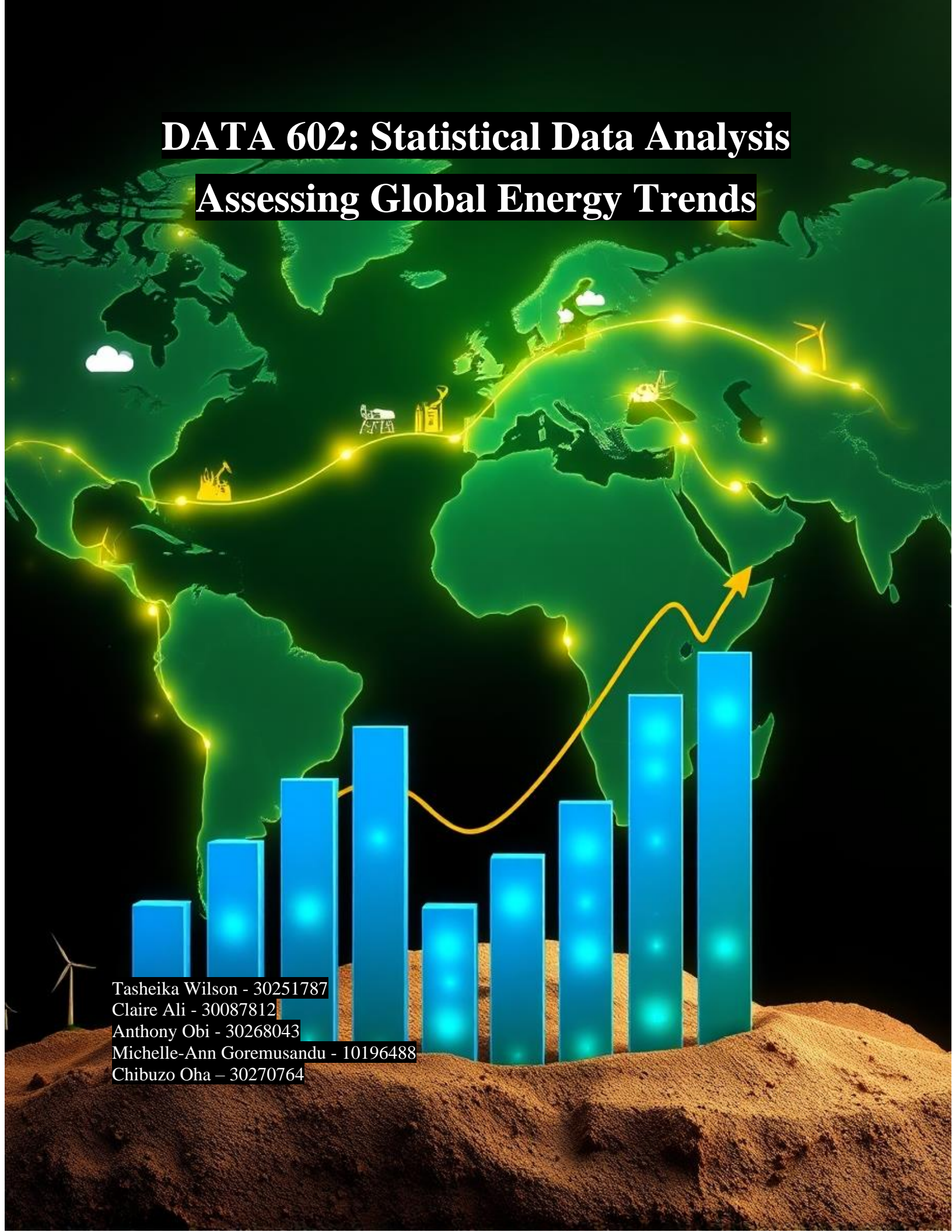


# DATA 602: Statistical Data Analysis

## Assessing Global Energy Trends



Tasheika Wilson - 30251787  
Claire Ali - 30087812  
Anthony Obi - 30268043  
Michelle-Ann Goremusandu - 10196488  
Chibuzo Oha - 30270764

# Table of Contents

<b>Introduction</b> .....	3
<b>Dataset</b> .....	3
<b>Guiding Questions</b> .....	4
<b>Question 1</b> .....	4
<b>Question 2</b> .....	4
<b>PART 1: Exploring Different Forms of Energy Globally</b> .....	5
Total Energy Production and Consumption by Country .....	5
Natural Gas.....	7
Coal and Lignite .....	10
Shares of Renewables in Electricity Production .....	13
<b>PART 2: Hypothesis Testing - Bootstrap Method</b> .....	16
Analyzing the Trend in the Renewable Energy Share of Electricity Production: A Comparison Between 1990 and 2020 .....	16
<b>Has the share of renewable energy in electricity production significantly changed         from 1990 to 2020?</b> .....	16
<b>PART 3 : Simple Linear Regression</b> .....	20
Investigating the relationship between oil products domestic consumption and O2 emissions in Canada .....	20
<b>Does oil products domestic consumption (Mt) affect the level of CO2 emissions in         Canada?</b> .....	20
Simple Linear Regression Model .....	22
Estimating the model .....	23
Hypothesis Testing.....	25
Assessing the fit of the model .....	26
Assessing Model Assumptions .....	27
Conclusion .....	29
Recommendation.....	29
Reference .....	30

## Introduction

Energy is fundamental to modern civilization, driving every aspect of our daily lives, from powering industries and transportation to homes and digital infrastructure (International Energy Agency, 2020). The importance of energy cannot be overstated as it plays a key role in technological enhancements, economic growth, and societal development.

As energy is a key part of our lives, this project aims to assess the global energy trends and their environmental impacts. Assessing the trends will give us keen insights into how different forms of energy production and consumption are changing across the globe. Energy production and consumption also come with environmental impacts. This analysis will help us investigate the relationship between energy usage and CO2 emissions, allowing us to assess how shifts toward renewable energy impact the global carbon footprint. By examining these trends and relationships, we aim to generate insights that will help inform our own daily choices and foster more sustainable energy consumption practices. Understanding these dynamics is crucial for making more informed decisions on energy usage, climate change, and global sustainability initiatives.

## Dataset

The dataset used in this project was obtained from Kaggle. The data within the dataset is sourced from Enerdata, a company specializing in global energy intelligence and licensed under World Bank Dataset Terms of Use. The dataset provided global energy statistics from various countries over 30 years, from 1990 to 2020. The data is structured in a tabular format consisting of 21 columns and 1365 unique rows. Within the dataset, there are various energy and emission-related metrics. The key metrics can be divided into the following categories:

- Emissions and Energy Intensity
- Energy Production and Consumption
- Fossil Fuels Production and Consumption
- Renewable Energy Shares in Electricity Production

## Guiding Questions

The guiding questions being explored aim to primarily explore different forms of energy production and how they have changed over time, by country or by region. They also aim to assess if renewable energy's contribution to production has changed. Finally, the relationship between CO2 emissions and Oil product consumption in Canada will be investigated.

Further details of the questions are:

### Question 1

Has the share of renewable energy in electricity production significantly changed from 1990 to 2020?

- Null Hypothesis: The proportion of share of renewable energy in electricity production in 2020 is less than or equal to the proportion in 1990.
- Alternative Hypothesis: The proportion of the shares of renewable energy in electricity production in 2020 is greater than the proportion in 1990.

### Question 2

Does oil products domestic consumption (Mt) affect the level of CO2 emissions in Canada?

- Null Hypothesis: There is no significant difference in the relationship between oil products domestic consumption (Mt) and CO2 emissions from fuel combustion (MtCO2)
- Alternative Hypothesis: There is a positive relationship between oil products domestic consumption (Mt) and CO2 emissions from fuel combustion (MtCO2)

# PART 1: Exploring Different Forms of Energy Globally

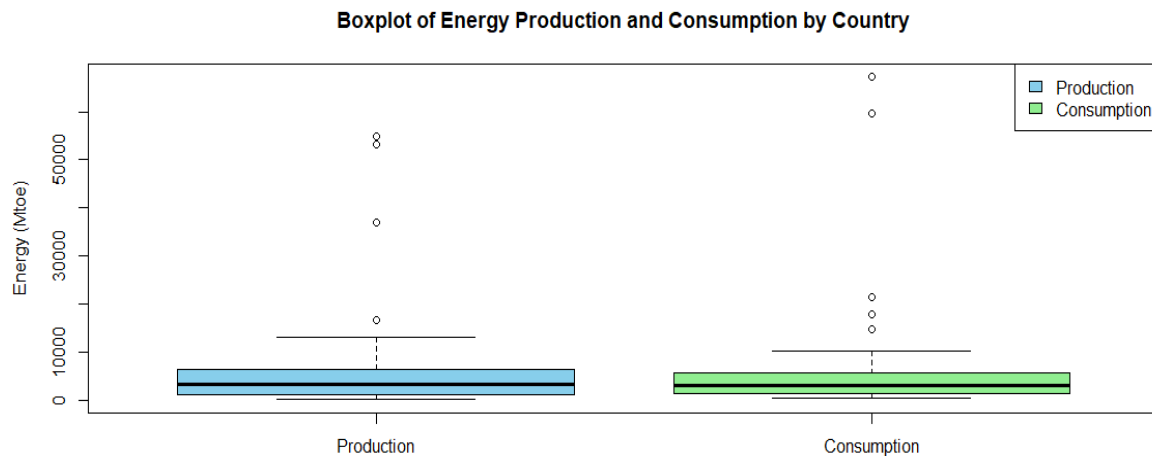
## Total Energy Production and Consumption by Country

To examine the relationship between energy consumption and production, a country-based analysis was conducted over 30 years of data. To facilitate this exploration, a graph of Total Energy and Total Consumption was plotted. The box plot shows the spread of the total energy and consumption for the entire dataset. The production boxplot has several outliers, with one extreme outlier at over 50,000 Mtoe. The consumption boxplot also has outliers, but they are less extreme.

```
My_data_grp_country = My_data %>% group_by(country) %>%
  reframe(
    Total.energy.production..Mtoe. =
sum(Total.energy.production..Mtoe.),
    Total.energy.consumption..Mtoe. =
sum(Total.energy.consumption..Mtoe.))

boxplot(My_data_grp_country$Total.energy.production..Mtoe.,
  My_data_grp_country$Total.energy.consumption..Mtoe.,
  main = 'Boxplot of Energy Production and Consumption by Country',
  ylab = 'Energy (Mtoe)',
  names = c("Production", "Consumption"),
  col = c('skyblue', 'lightgreen'),
  border = 'black',
  horizontal = FALSE)

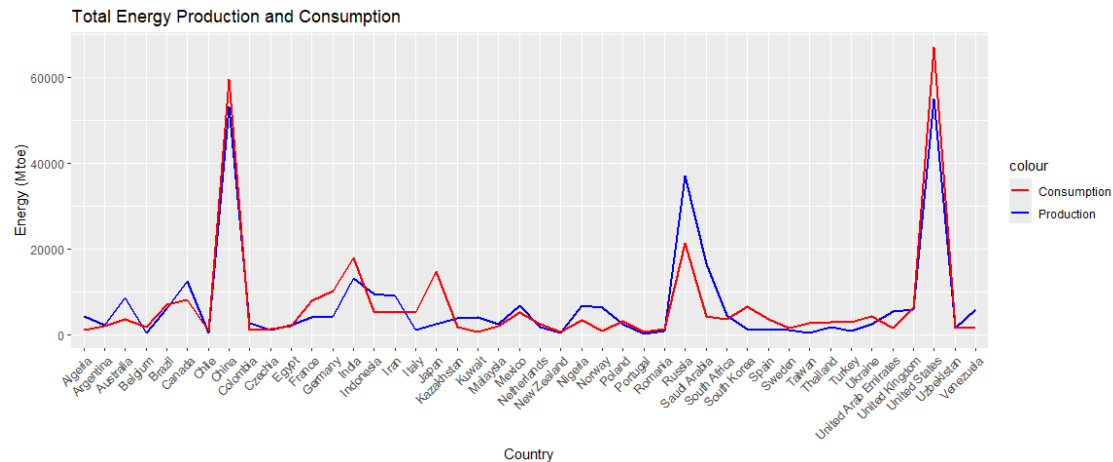
# Add a Legend
legend("topright",
  legend = c("Production", "Consumption"),
  fill = c("skyblue", "lightgreen"),
  border = "black")
```



The line graph reveals critical insights about energy production and consumption across different countries. Countries where production lines are above or close to consumption lines are generally self-sufficient or export energy. Countries where the consumption line is higher rely on energy imports. A country producing more than it consumes might focus on export strategies or energy market dominance, while countries consuming more than they produce may need to focus on energy efficiency or import strategies.

```
# Line plot for energy production and consumption by country
ggplot(My_data_grp_country, aes(x = country)) +
  geom_line(aes(y = Total.energy.production..Mtoe., color = "Production",
group = 1),
            linewidth = 1) +
  geom_line(aes(y = Total.energy.consumption..Mtoe., color = "Consumption",
group = 1),
            linewidth = 1) +
  labs(title = "Total Energy Production and Consumption",
       x = "Country",
       y = "Energy (Mtoe)") +
  scale_color_manual(values = c("Production" = "blue", "Consumption" =
"red")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```





## Natural Gas

Natural gas is a fossil fuel energy source primarily composed of methane, with smaller amounts of other hydrocarbons and non-hydrocarbon gases. The bargraphs below show that generally, regions that are high producers of natural gas are consequently high domestic consumers of natural gas and vice versa.

```
data$`Natural gas production (bcm)` <- as.numeric(data$`Natural gas
production (bcm)`)
```

## Warning: NAs introduced by coercion

```
data$`Natural gas production (bcm)` <- ifelse(is.na(data$`Natural gas
production (bcm)`),
                                             0, data$`Natural gas production
(bcm)`)
```

```
data$`Natural gas production (bcm)` <- round(data$`Natural gas production
(bcm)` , 2)
data$`Natural gas domestic consumption (bcm)` <-
  as.numeric(as.character(data$`Natural gas domestic consumption (bcm)`))
```

```
data$`Natural gas domestic consumption (bcm)` <-
  round(data$`Natural gas domestic consumption (bcm)` ,
        2)
```

```
plot1 <- ggplot(data, aes(x=Region, y=`Natural gas production (bcm)` ,
fill=Region)) +
  geom_bar(stat="identity") +
  labs(title="Total Natural Gas Production \n by Region (1990-2020)",
       x="Region", y="Production (bcm)") +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
        plot.title=element_text(size=10,face = "bold", hjust = 0.5))
```

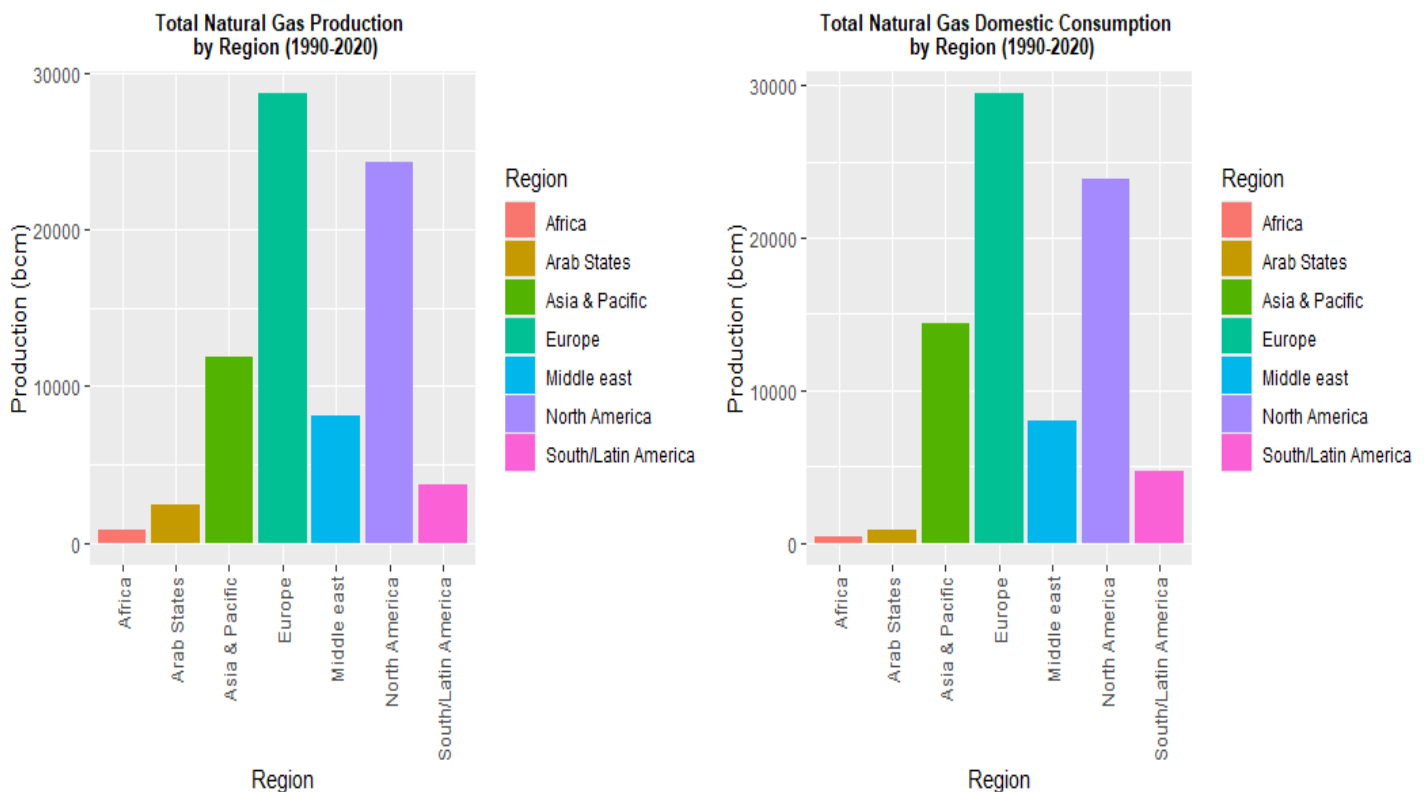
```
plot2 <- ggplot(data, aes(x=Region, y=`Natural gas domestic consumption
(bcm)` ,
```

```

    fill=Region)) + geom_bar(stat="identity") +
  labs(title="Total Natural Gas Domestic Consumption \n by Region (1990-
2020)",
    x="Region",
    y="Production (bcm)" ) +
  theme(axis.text.x = element_text(angle=90, vjust=0.5, hjust=1),
    plot.title=element_text(size=10,face = "bold", hjust = 0.5))

grid.arrange(plot1, plot2, ncol=2)

```



Looking further at the spread of the data, we see that the average natural gas production is higher in Europe and North America, making them the top producers. Upon further analysis, we see that Russia is the top European producer (600+bcm) and the United States is the top North American Producer (500+bcm). As of 2021, the United States significantly outproduces Russia in natural gas. The gap between them is larger than shown in the graph.

```

data_countries <- data %>%
  filter(country %in% c("Russia", "United States", "China", "Iran",
"Canada"))

ggplot(data_countries, aes(x = country, y = `Natural gas production (bcm)`),

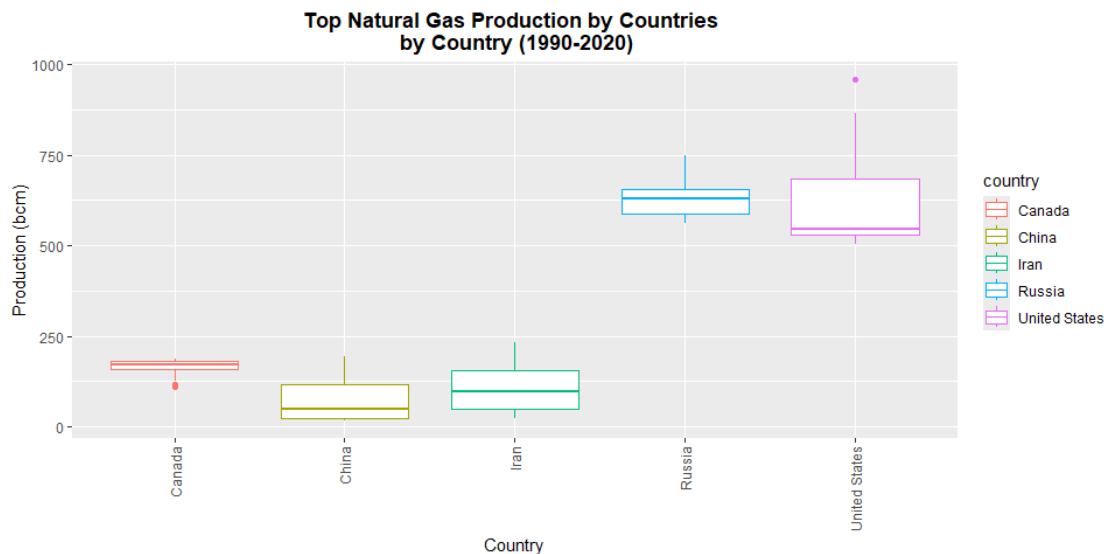
```



```

    color = country, group = country)) +
geom_boxplot() +
  labs(title = "Top Natural Gas Production by Countries \n by Country
(1990-2020)",
    x = "Country", y = "Production (bcm)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
    plot.title = element_text(face = "bold", hjust = 0.5))

```



From the above plot, we see that these top producing countries have been on a steady incline since approximately 2016. A country of interest was Iran; although not the top global producer, it has the highest proportion of natural gas energy to total energy consumption among the top natural gas producers. While this may indicate a shift in energy usage for the countries, the reasons are more complex. Some of the higher natural gas consumption can be attributed to (*Why Iran Consumes Five Times More Gas Than Turkey?*, 2023):

1. A decline in the northern hemisphere's temperatures, causing increased household consumption of natural gas for heating
2. Gas leakage in transmission and distribution (approximately 7bcm/yr)

```

data <- data %>%
  mutate(
    NatGasToTotalEnergy =
      `Natural gas domestic consumption (bcm)` / `Total energy consumption
(Mtoe)`
  )

data_countries <- data %>%
  filter(country %in% c("Russia", "United States", "Canada", "China",
    "Iran"))

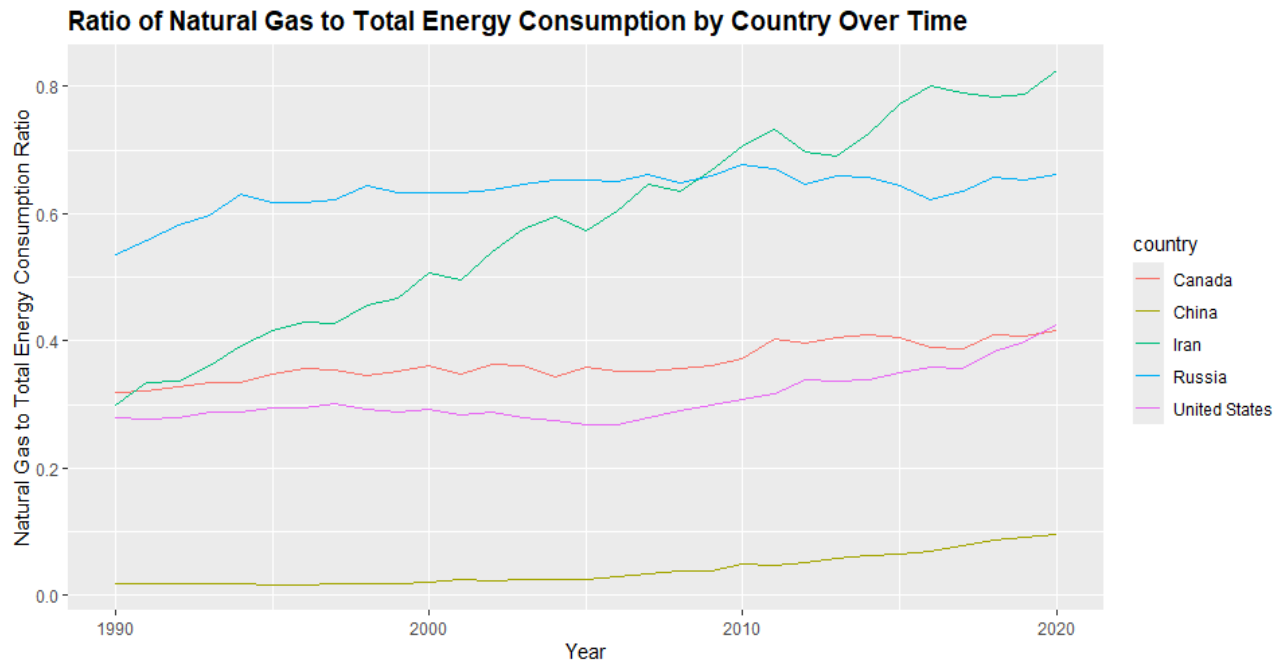
ggplot(data_countries, aes(x = Year, y = NatGasToTotalEnergy,

```

```

    color = country, group = country)) + geom_line() +
labs(title = "Ratio of Natural Gas to Total Energy Consumption by Country
Over Time",
      x = "Year",
      y = "Natural Gas to Total Energy Consumption Ratio") +
theme(
  plot.title = element_text(face = "bold", size = 14))

```



## Coal and Lignite

Total Coal consumption witnessed a steady increase from 1990, peaking in 2014. Since 2014, the demand for coal has been on the decline, partly due to a shift to more sustainable and renewable energy sources.

```

Coal_Lignite
=Energy_data[,c("country", "Year", "Coal.and.lignite.domestic.consumption..Mt."
)]
Coal_Lignite$Coal.and.lignite.domestic.consumption..Mt. <-
  as.numeric(gsub("[^0-9.]", "",
Coal_Lignite$Coal.and.lignite.domestic.consumption..Mt.))

## Warning: NAs introduced by coercion

#x=mean(Coal_Lignite$Coal.and.lignite.domestic.consumption..Mt., na.rm =
TRUE)
grouped_data_year <- Coal_Lignite %>%
  group_by(Year) %>%
  summarise(
    sum_value = sum(Coal.and.lignite.domestic.consumption..Mt., na.rm =
TRUE),

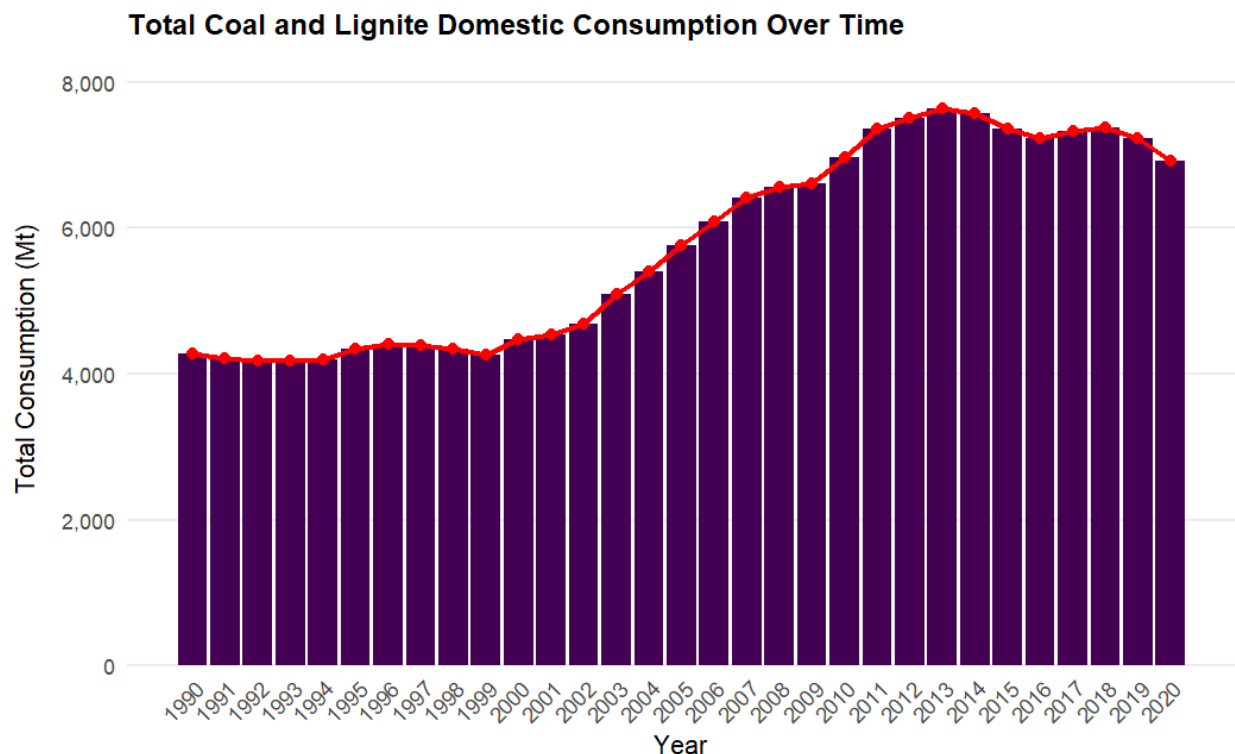
```

```

.groups = 'drop')

ggplot(data = grouped_data_year, aes(x = Year, y = sum_value)) +
  geom_col(fill = viridis(1)) +
  geom_line(color = "red", linewidth = 1, group = 1) +
  geom_point(color = "red", size = 2) +
  labs(title = "Total Coal and Lignite Domestic Consumption Over Time",
       x = "Year",
       y = "Total Consumption (Mt)") +
  scale_y_continuous(labels = comma_format(),
                    expand = expansion(mult = c(0, 0.1))) +
  scale_x_continuous(breaks = unique(grouped_data_year$Year)) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    axis.title.y = element_text(margin = margin(r = 10)),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank()
  )

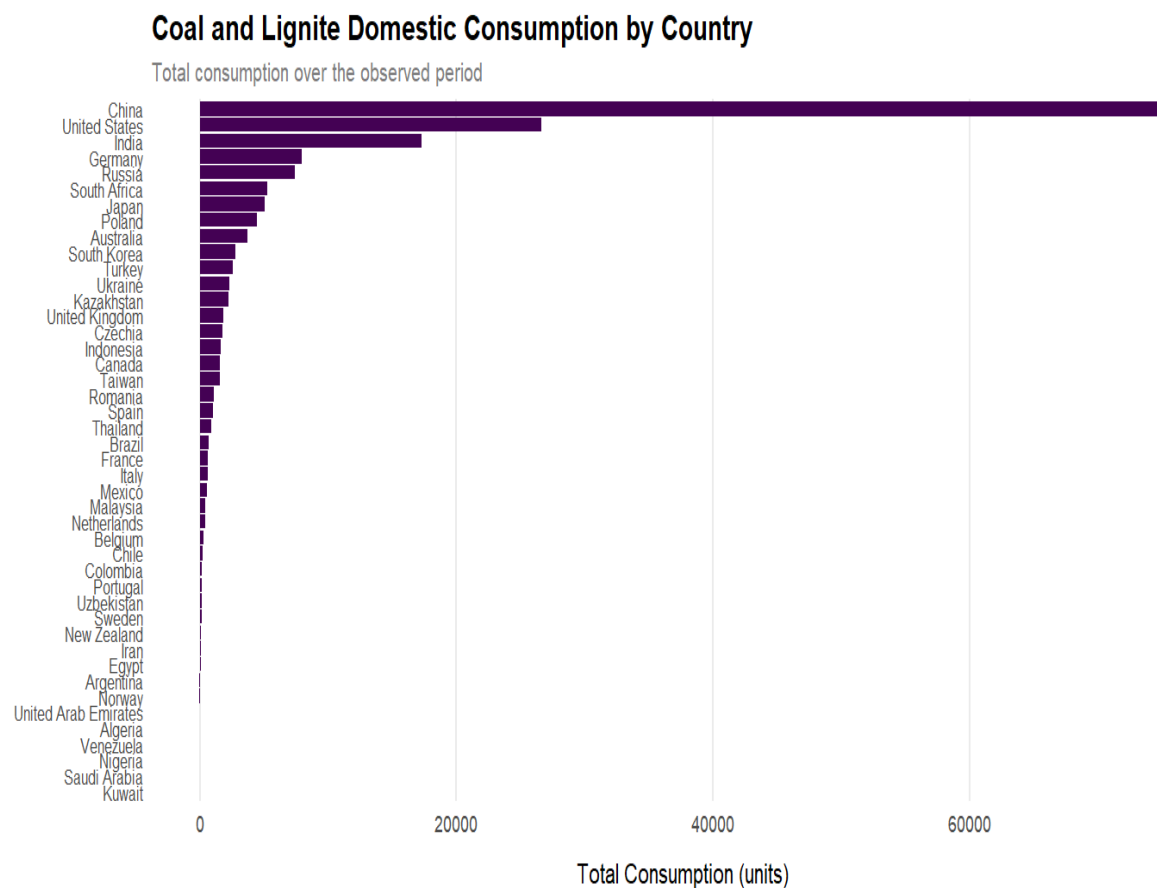
```



China is the biggest consumer of coal followed by the United States and India. Most of the oil-producing countries, including Saudi Arabia, Nigeria, and Venezuela, have little or zero consumption of coal.

```
grouped_data <- Coal_Lignite %>%
  group_by(country) %>%
  summarise(
    sum_value = sum(Coal.and.lignite.domestic.consumption..Mt., na.rm =
TRUE),
    .groups = 'drop') %>%
  arrange(desc(sum_value))

ggplot(data = grouped_data, aes(x = reorder(country, sum_value), y =
sum_value)) +
  geom_bar(stat = "identity", fill = viridis(1)) +
  labs(title = "Coal and Lignite Domestic Consumption by Country",
    subtitle = "Total consumption over the observed period",
    x = "",
    y = "Total Consumption (units)") +
  coord_flip() +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    axis.text.y = element_text(size = 8),
    axis.title.x = element_text(margin = margin(t = 10)),
    panel.grid.major.y = element_blank(),
    panel.grid.minor.x = element_blank())
```



# Shares of Renewables in Electricity Production

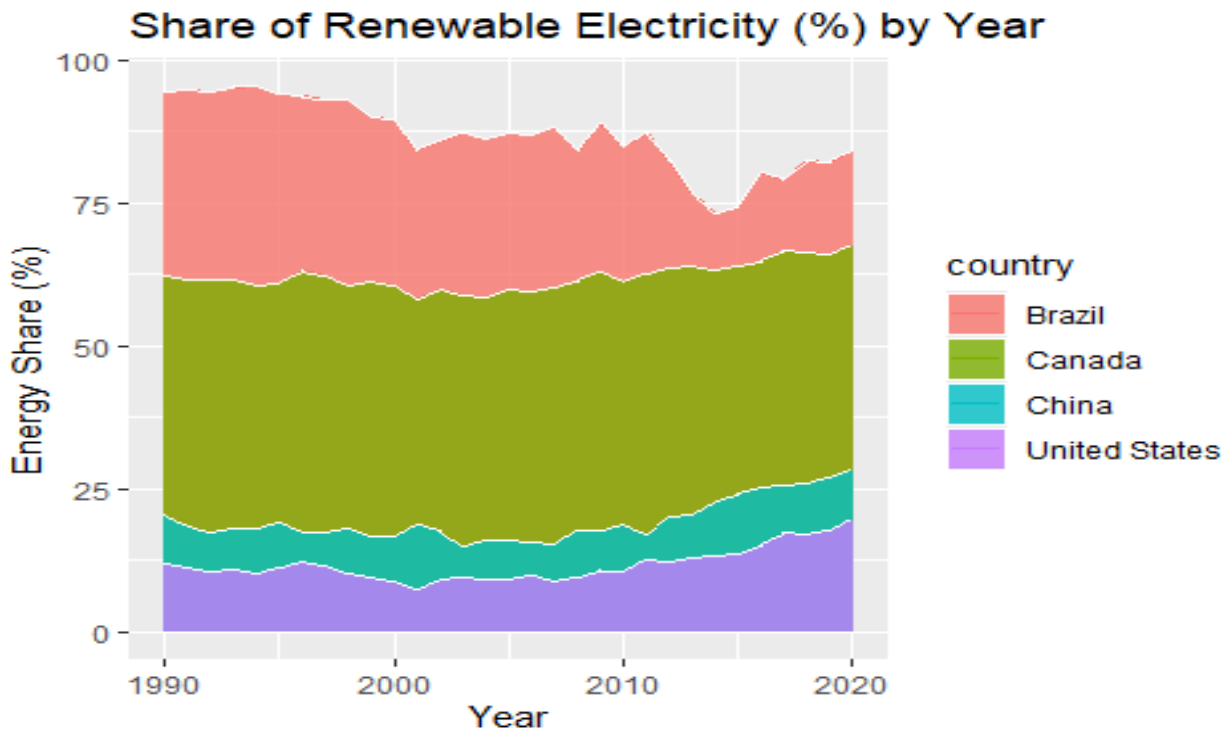
Shares of renewables in energy production have been steadily increasing globally over the past few decades. This growth is primarily attributed to ambitious renewable energy policies and falling production costs for solar and wind power technologies. Based on this dataset, we will explore how it has changed.

To evaluate the progression of the country's shares of renewable electricity production, background research was conducted to determine the current highest shares. Globally, China leads, producing approximately 31% of renewable electricity. This is nearly three times that of the next greatest shares by country, which is the United States with 11%. The United States is followed by Brazil, with 6.4%, and Canada, with 5.4% (IRENASTAT Online Data Query Tool, n.d.).

The following exploration seeks to assess these countries and determine what share of overall electricity produced in each country is renewable. The shares for the countries were plotted against each other.

```
m <- dplyr::select(energy_df, country, Year,
Share.of.renewables.in.electricity.production...)
v <- m %>% filter(grepl('Brazil|China|Canada|United.States', country))

ggplot(data = v, aes(x=Year,
y=Share.of.renewables.in.electricity.production..., fill=country)) +
geom_line(aes(colour=country)) + geom_area(alpha=0.8,
colour="white", aes(y=Share.of.renewables.in.electricity.production...), position = 'identity') + labs(title = "Share of Renewable Electricity (%) by
Year",
x = "Year",
y = "Energy Share (%)")
```



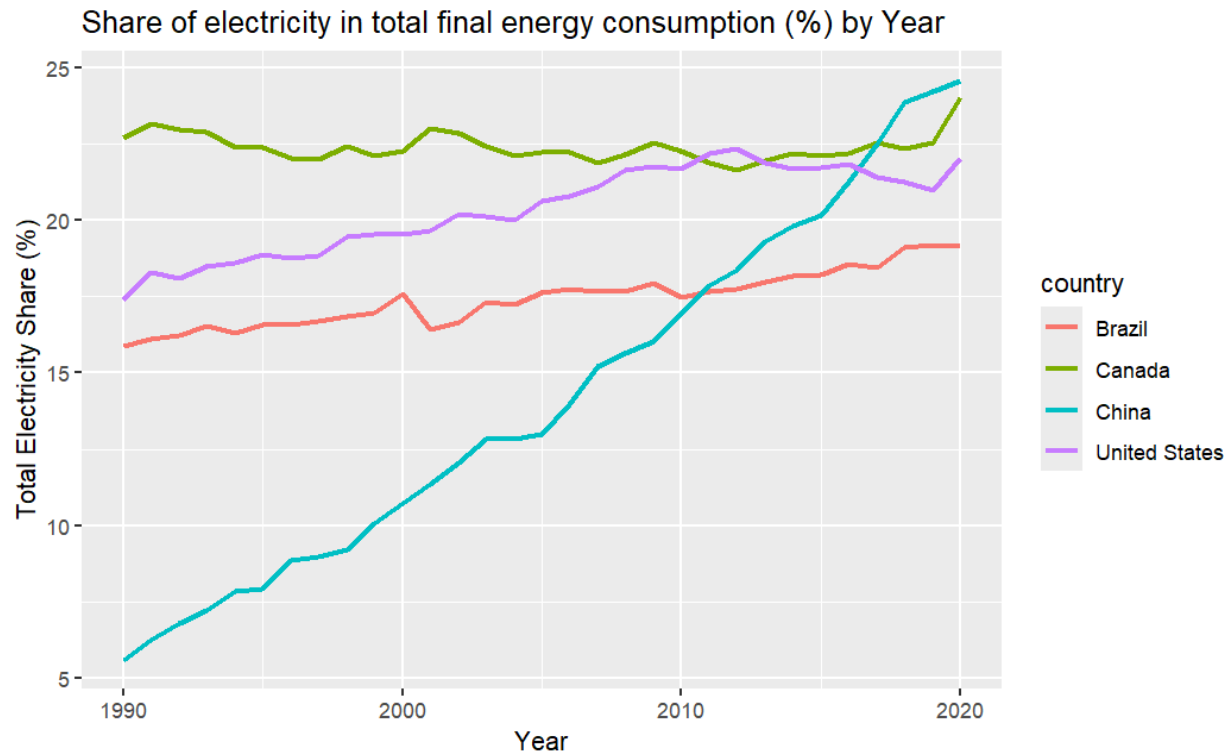
Brazil yielded the highest overall national proportion. While Canada is the fourth global producer of electricity, Canada yielded energy shares that were only marginally lower than that of Brazil. In comparison, while China and the United States are quite large global producers, most of the total electricity generated by each country is non-renewable.

The data was then examined to determine each country's share of electricity for their total final energy consumption.

```
n <- dplyr::select(energy_df, country, Year,
Share.of.electricity.in.total.final.energy.consumption....)
t <- n %>% filter(grepl('Brazil|China|Canada|United.States', country))

ggplot(data = t, aes(x=Year,
y=Share.of.electricity.in.total.final.energy.consumption...., fill=country))
+ geom_line(aes(colour=country), lwd=1) + labs(title = "Share of electricity in
total final energy consumption (%) by Year",
x = "Year",
y = "Total Electricity Share (%)")
```





The United States and Canada yield relatively large, fairly stable trend lines for each proportion. Brazil is similarly stable, yielding slightly lower overall proportions than the previous two. Notably, China had very little electricity shares in 1990; however, the resultant trend yields a strong, positive correlation, that eventually exceeds the prior three countries. While China may be the global leader in renewable electricity production, proportionally, they lack sustainable energy consumption practices, which may contribute to its extremely high levels of air pollution. China uses primarily coal to generate electricity; while efforts have been made to generate alternate energy sources, a drastic change will be needed to reduce environmental impact (Huang, 2024). If this steadily increasing trend continues, human welfare and ecological well-being may be devastated.

## PART 2: Hypothesis Testing - Bootstrap Method

### Analyzing the Trend in the Renewable Energy Share of Electricity

#### Production: A Comparison Between 1990 and 2020

#### **Has the share of renewable energy in electricity production significantly changed from 1990 to 2020?**

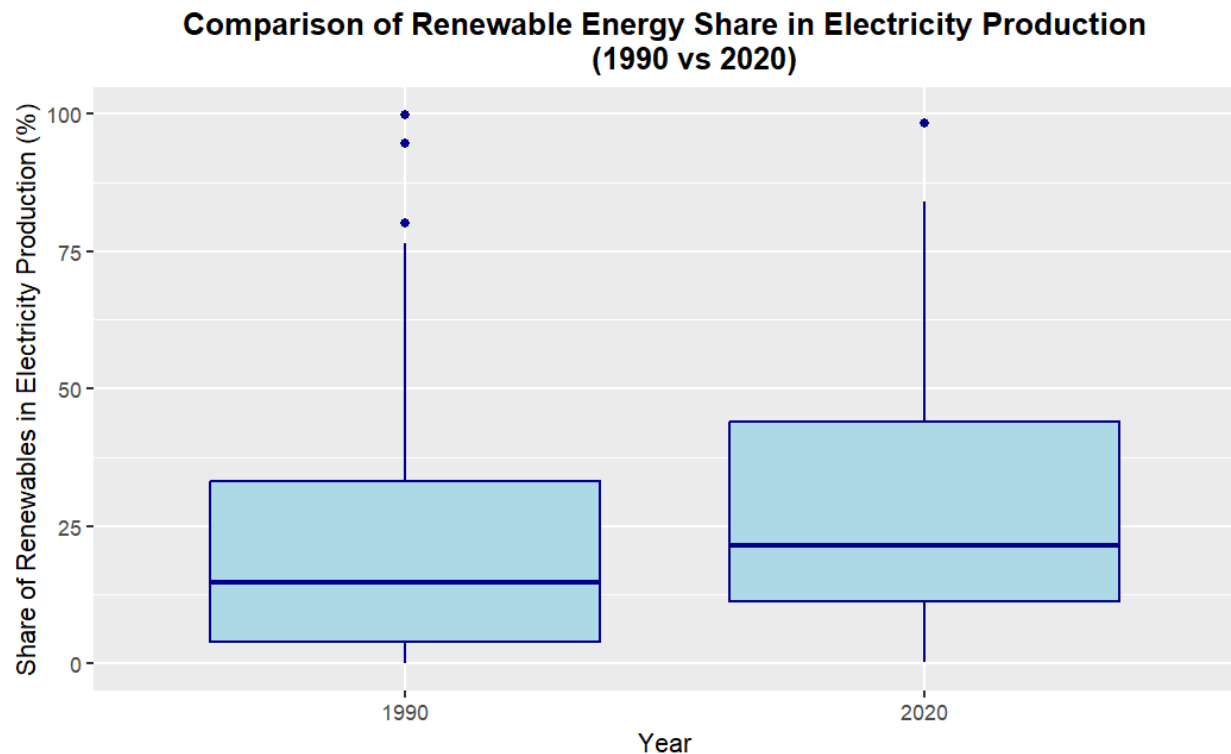
Null Hypothesis: The proportion of share of renewable energy in electricity production in 2020 is less than or equal to the proportion in 1990.

Alternative Hypothesis: The proportion of the shares of renewable energy in electricity production in 2020 is greater than the proportion in 1990.

This visualization allows us to quickly assess whether there's been a significant change in the use of renewables for electricity production globally from 1990 to 2020, and how the distribution of this share has changed. From the plot, there is a change in the median line indicating that there has been an increase from 1990 to 2020.

```
Data_1990 <- Energy_data %>% filter(Year ==1990)
Data_2020 <- Energy_data %>% filter(Year ==2020)
# Combine the data for 1990 and 2020
combined_data <- Energy_data %>%
  filter(Year %in% c(1990, 2020)) %>%
  mutate(Year = as.factor(Year))

ggplot(combined_data, aes(x = Year, y =
Share.of.renewables.in.electricity.production...)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(title = "Comparison of Renewable Energy Share in Electricity
Production
(1990 vs 2020)",
x = "Year",
y = "Share of Renewables in Electricity Production (%)")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5,))
```



To investigate whether the shares have increased a hypothesis test will be conducted. However, to determine where a parametric or non-parametric approach is used, the Shapiro-Wilk method will be used to test the normality of our data. The Shapiro-Wilk test is a hypothesis test that is applied to a sample with a null hypothesis stating that the sample has been generated from a normal distribution. It is an appropriate test to check the normality of the variable within the dataset, as the size of the variables is less than 5000. The dataset being used is less than 5000 so this test is appropriate.

```
Data_1990 <- Energy_data %>% filter(Year ==1990)
Data_2020 <- Energy_data %>% filter(Year ==2020)
D2020<-Data_2020$Share.of.renewables.in.electricity.production....
D1990<-Data_1990$Share.of.electricity.in.total.final.energy.consumption....

shapiro.test(D2020)

##
##  Shapiro-Wilk normality test
##
## data:  D2020
## W = 0.89326, p-value = 0.0006798

shapiro.test(D1990)

##
##  Shapiro-Wilk normality test
```

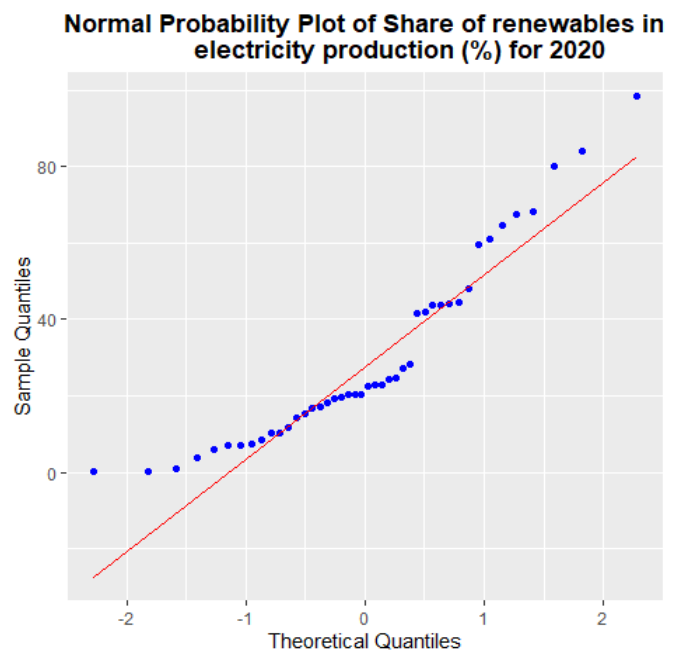
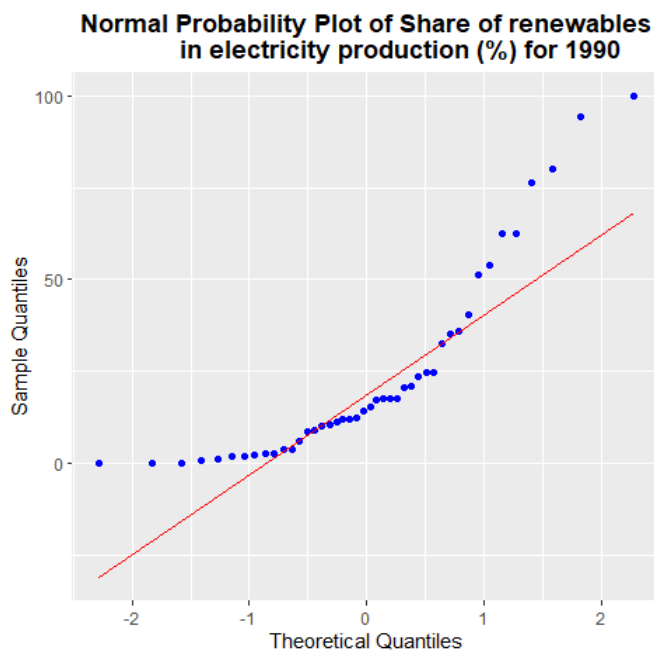
```
##
## data: D1990
## W = 0.85432, p-value = 5.637e-05
```

A plot can also be used to check the normality of the variable.

```
y=
ggplot(Data_2020,aes(sample=Share.of.renewables.in.electricity.production....
)) +
  stat_qq(color = "blue") +
  stat_qq_line(color = "red") +
  ggtitle("Normal Probability Plot of Share of renewables in
          electricity production (%) for 2020") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles")+theme(plot.title = element_text(face = "bold",
hjust = 0.5,))

x=ggplot(Data_1990,aes(sample=Share.of.renewables.in.electricity.production..
..)) +
  stat_qq(color = "blue") +
  stat_qq_line(color = "red") +
  ggtitle("Normal Probability Plot of Share of renewables
          in electricity production (%) for 1990") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles")+theme(plot.title = element_text(face = "bold",
hjust = 0.5,))

plot_grid(x,y,align="h",ncol=2,nrow=1)
```



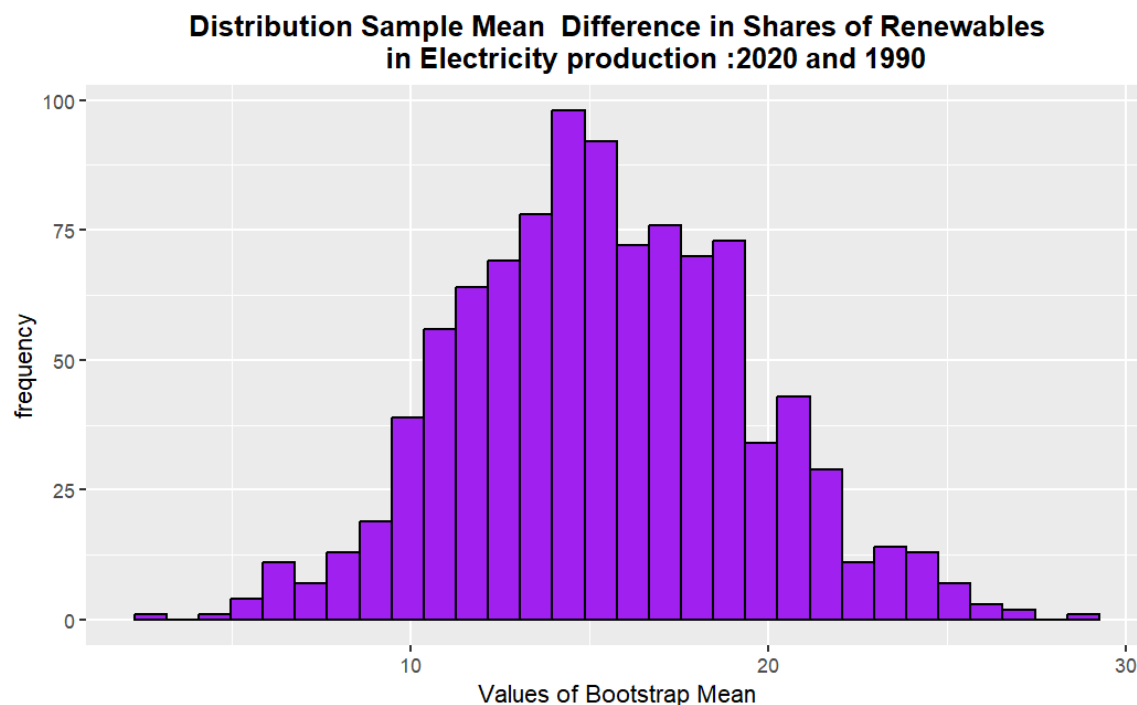
From the results from the Shapiro-Wilks, the P-value is very small, lower than the significance level of 0.05. Therefore the null hypothesis is rejected. It concludes that the distribution of the variable does not follow a normal distribution. This can also be seen in the Normality plots. Based on this conclusion, a non-parametric approach needs to be taken to investigate the difference in the shares of renewables in electricity production across the two years.

Since the investigation looks into whether the share of renewables in electricity production is greater in 2020 compared to 1990, the bootstrap approach will be used. Specifically, testing the difference between two population

```
set.seed(123)

Data_1990 <- Energy_data %>% filter(Year ==1990)
Data_2020 <- Energy_data %>% filter(Year ==2020)
D2020<-Data_2020$Share.of.renewables.in.electricity.production....
D1990<-Data_1990$Share.of.electricity.in.total.final.energy.consumption....

shares_2020<- do(1000)*mean(resample(D2020))
shares_1990<- do(1000)*mean(resample(D1990))
diff<-shares_2020-shares_1990
ggplot(data=diff,aes(x=mean))+geom_histogram(color="black",fill="purple",bins
=30)+
  xlab("Values of Bootstrap Mean")+ ylab("frequency")+
  ggtitle("Distribution Sample Mean Difference in Shares of Renewables
in Electricity production :2020 and 1990")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5,))
```



The Confidence interval for our population difference needs to be calculated to see whether the null hypothesis is true. Using the quantile function:

```
quantile(diff$mean,c(0.025,0.975))
```

```
##      2.5%      97.5%  
## 7.813046 23.959014
```

The confidence interval is given as [7.814,23.959]. Both ends of the interval are positive, which strongly suggests that there was a real increase in the share of renewables from 1990 to 2020. Therefore, the null hypothesis is rejected. The interval suggests that the increase in the share of renewables was at least 7.814% and could be as high as 23.959%. Since the interval doesn't include zero, we can conclude that the difference is statistically significant at the 95% confidence level. The wide interval might indicate substantial variability in renewable adoption rates among different countries.

This increase may suggest that policies and technologies promoting renewable energy between 1990 and 2020 have had a measurable impact.

## PART 3 : Simple Linear Regression

### Investigating the relationship between oil products domestic consumption and O2 emissions in Canada

#### Does oil products domestic consumption (Mt) affect the level of CO2 emissions in Canada?

First, let's look at the trends of oil products consumption and emission for Canada over the 30 years of the dataset.

```
x =ggplot(CA_data, aes(x = Year, y =  
CO2.emissions.from.fuel.combustion..MtCO2.)) +  
  geom_line(color = "#1E88E5", size = 1.2) +  
  geom_point(color = "#FFC107", size = 3) +  
  geom_smooth(method = "loess", se = FALSE, color = "#D81B60", linetype =  
"dashed", size = 1) +  
  labs(  
    title = "CO2 Emissions from Fuel Combustion in Canada (1990-2020)",  
    subtitle = "Annual emissions measured in million tonnes of CO2",  
    x = "Year",  
    y = "CO2 Emissions (MtCO2)"  
  )+  
  theme_minimal() +  
  theme(
```



```

    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5, color = "grey50"),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = 10),
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(color = "grey90"))

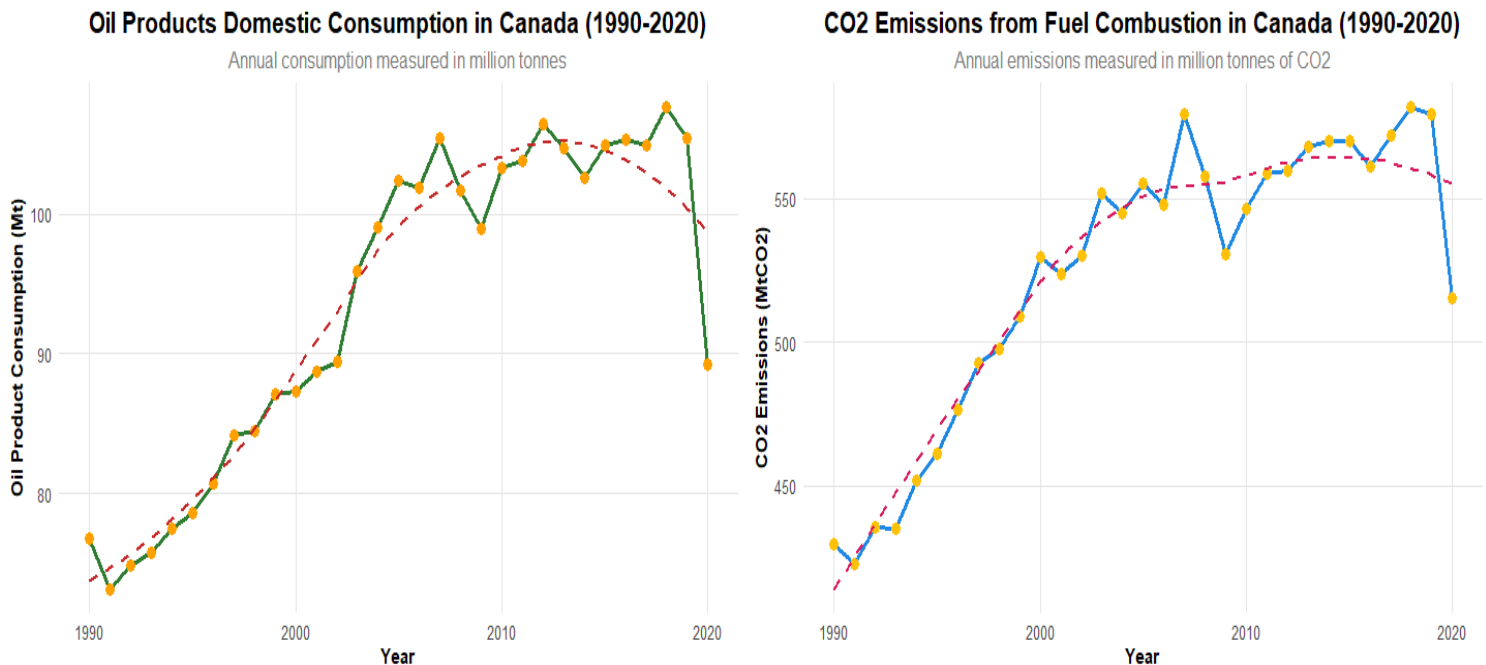
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

y= ggplot(CA_data, aes(x = Year, y = Oil.products.domestic.consumption..Mt.))
+
  geom_line(color = "#2E7D32", size = 1.2) +
  geom_point(color = "#FFA000", size = 3) +
  geom_smooth(method = "loess", se = FALSE, color = "#C62828", linetype =
"dashed", size = 1) +
  labs(
    title = "Oil Products Domestic Consumption in Canada (1990-2020)",
    subtitle = "Annual consumption measured in million tonnes",
    x = "Year",
    y = "Oil Product Consumption (Mt)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5, color = "grey50"),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = 10),
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(color = "grey90")
  )

plot_grid(y,x,align="h",ncol=2,nrow=1)

```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Based on both plots, we see an increasing trend to a point, then a big decrease from 2019 to 2020. This may be attributed to the Covid-19 pandemic.

```
favstats(~CO2.emissions.from.fuel.combustion..MtCO2.,data=Energy_data)

##      min      Q1   median      Q3      max      mean      sd      n
missing
## 7.597759 99.9472 221.0363 423.8332 9716.772 550.3752 1213.058 1364
0

favstats(~Oil.products.domestic.consumption..Mt.,data=Energy_data)

##   min    Q1  median    Q3   max   mean    sd   n missing
## 1.949 12.809 27.75218 77.50086 888.491 68.84885 128.3397 1364      0
```

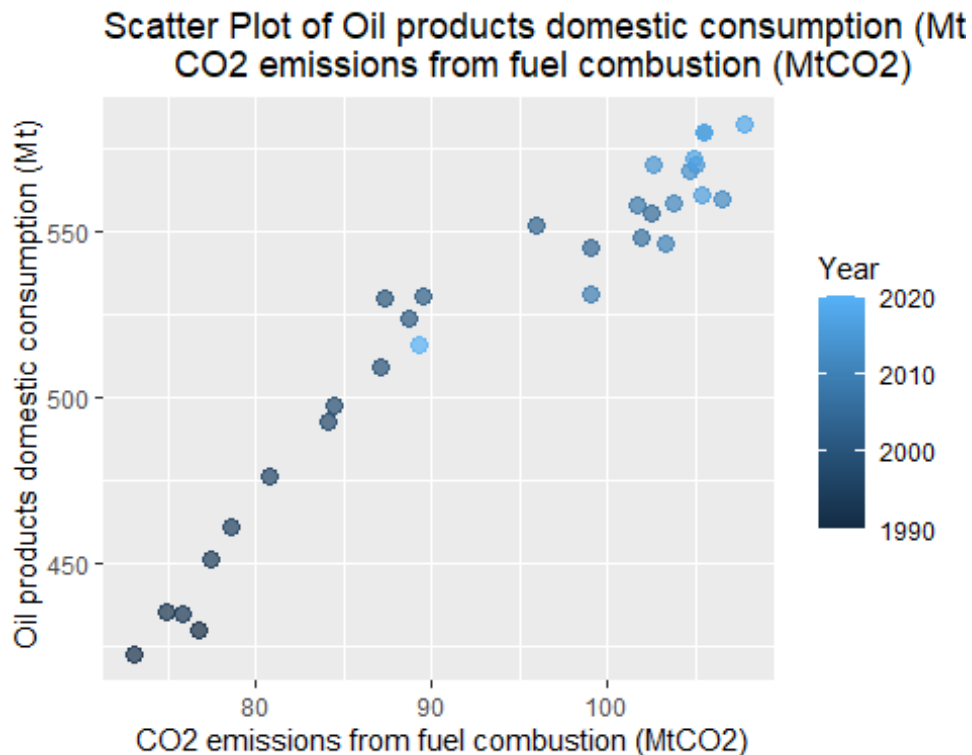
## Simple Linear Regression Model

The aim is to investigate the linear association between oil products' domestic consumption (Mt) and CO2 emissions from fuel combustion (MtCO2). In this analysis our variables are:

- Dependent - CO2 emissions from fuel combustion (MtCO2)
- Independent - Oil products domestic consumption (Mt)

A scatter plot was constructed to visualize the relationship.

```
ggplot(CA_data, aes(x=Oil.products.domestic.consumption..Mt.,
                    y=CO2.emissions.from.fuel.combustion..MtCO2.))+
  geom_point(aes(color = Year), size = 3, alpha = 0.7)+
  labs(title = "Scatter Plot of Oil products domestic consumption (Mt)
CO2 emissions from fuel combustion (MtCO2) ",
       x=" CO2 emissions from fuel combustion (MtCO2)",
       y=" Oil products domestic consumption (Mt)")
```



From the visualization above, we can see that there is a positive relationship between the two variables, indicating that as the oil product consumption increases, CO2 Emissions will also be greater.

### Estimating the model

```
lm(CO2.emissions.from.fuel.combustion..MtCO2.~Oil.products.domestic.consumpti
on..Mt.,data=CA_data)

##
## Call:
## lm(formula = CO2.emissions.from.fuel.combustion..MtCO2. ~
Oil.products.domestic.consumption..Mt.,
##     data = CA_data)
##
```

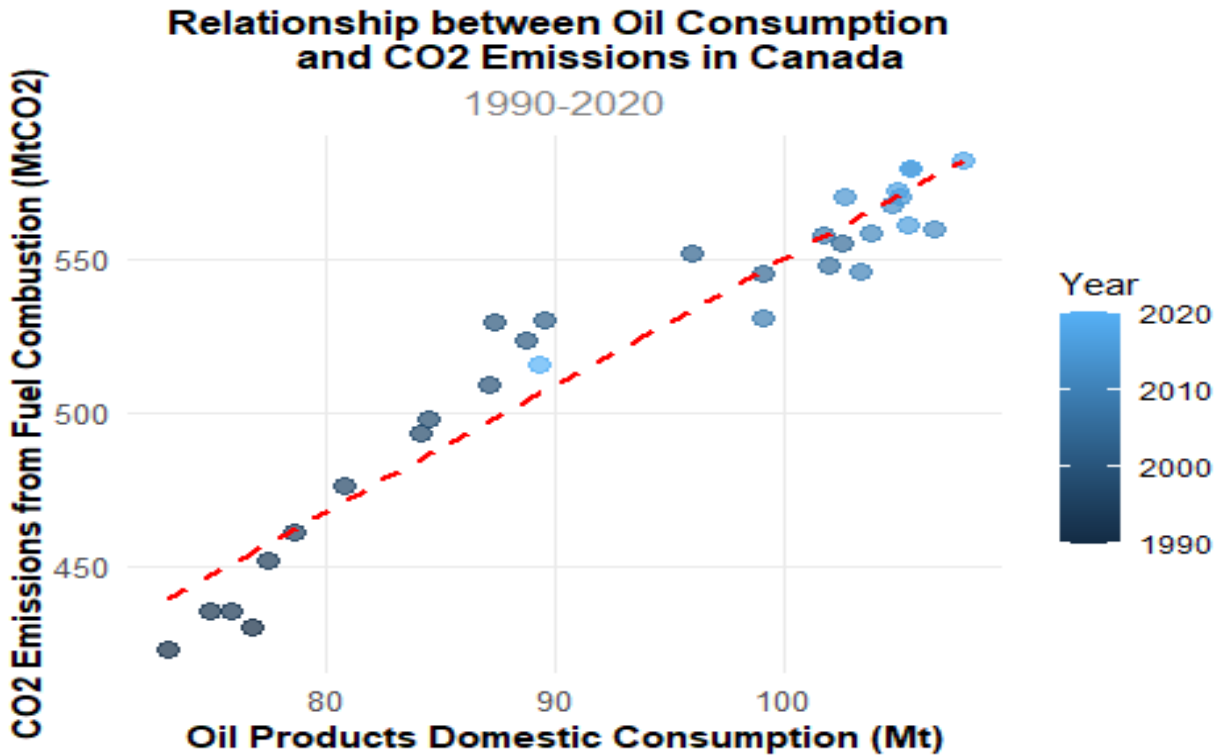
```
## Coefficients:
##                               (Intercept)
Oil.products.domestic.consumption..Mt.
##                               137.185
4.132
```

Based on the results from the linear regression model, we can construct the equation as CO2 emissions from fuel **combustion = 137.19 + 4.13 × Oil products domestic consumption**. So we can say that:

- 137.19 represents the estimated CO2 emissions when oil product consumption is zero.
- 4.13 indicates that for every unit increase in oil products domestic consumption, CO2 emissions are estimated to increase by 4.13 units (MtCO2).

To visualize the regression:

```
ggplot(CA_data, aes(x = Oil.products.domestic.consumption..Mt.,
                    y = CO2.emissions.from.fuel.combustion..MtCO2.)) +
  geom_point(aes(color = Year), size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = FALSE, linetype = "dashed")
+
  labs(title = "Relationship between Oil Consumption
    and CO2 Emissions in Canada",
    subtitle = "1990-2020",
    x = "Oil Products Domestic Consumption (Mt)",
    y = "CO2 Emissions from Fuel Combustion (MtCO2)",
    color = "Year") +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 12, hjust = 0.5),
    plot.subtitle = element_text(size = 12, hjust = 0.5, color = "grey50"),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = 10),
    legend.position = "right",
    panel.grid.minor = element_blank()
  )
## `geom_smooth()` using formula = 'y ~ x'
```



Looking at the strength of the relationship between the two variables :

```
cor(CA_data$Oil.products.domestic.consumption..Mt.,
    CA_data$CO2.emissions.from.fuel.combustion..MtCO2.)
## [1] 0.963729
```

The model suggests a positive linear relationship between oil consumption and CO2 emissions from fuel combustion. For every million-tonne increase in oil product consumption, CO2 emissions are estimated to increase by 4.13 million tonnes. Even with no oil consumption, there's an estimated baseline emission of 137.19 million tonnes of CO2, likely from other fuel sources.

## Hypothesis Testing

To further investigate the relationship, a hypothesis test can be used to potentially determine a positive association

Null Hypothesis: There is no significant difference in the relationship between oil products domestic consumption (Mt) and CO2 emissions from fuel combustion (MtCO2)

Alternative Hypothesis: There is a positive association between oil products domestic consumption (Mt) and CO2 emissions from fuel combustion (MtCO2)

We are investigating

```

model<-
lm(CO2.emissions.from.fuel.combustion..MtCO2.~Oil.products.domestic.consumption..Mt.,
    data=CA_data)
summary(model)

##
## Call:
## lm(formula = CO2.emissions.from.fuel.combustion..MtCO2. ~
Oil.products.domestic.consumption..Mt.,
##     data = CA_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.4974 -10.5213  -0.7743   8.4620  31.6002
##
## Coefficients:
##                                Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                   137.1845     20.0451   6.844 1.62e-
07 ***
## Oil.products.domestic.consumption..Mt.    4.1321      0.2125  19.446 < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.55 on 29 degrees of freedom
## Multiple R-squared:  0.9288, Adjusted R-squared:  0.9263
## F-statistic: 378.2 on 1 and 29 DF,  p-value: < 2.2e-16

```

To test the hypothesis that there is a positive association, the following formula is used:

```

1-pt(q=19.446,df=29)

## [1] 0

```

The P value for this test is 0, which is less than the 5% level of significance thus we can reject the null hypothesis in favor of the alternative. From this, we can conclude that there is a positive association between the two variables.

### Assessing the fit of the model

- The Multiple R-squared of 0.9288 indicates that about 92.88% of the variance in CO2 emissions is explained by oil product consumption.
- The Adjusted R-squared (0.9263) is very close to the Multiple R-squared, suggesting that the model isn't overfitted.
- The F-statistic (378.2) with a p-value of 0 indicates that the model is statistically significant.

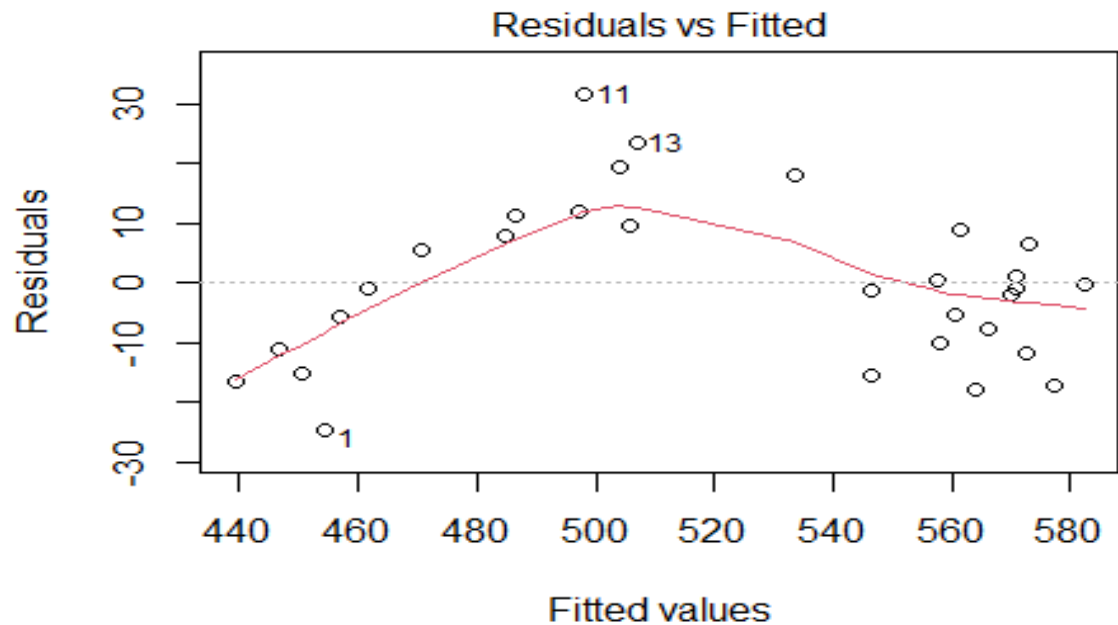


## Assessing Model Assumptions

Looking at the residual plots from the model will help to check if the model assumptions are met.

The first plot being looked at is the Residuals vs Fitted plot. For this plot, no patterns should be seen. From the plot below we can say that there is some curve happening in the plot. Because some patterns arise in the data, it suggests a violation of the assumption of homoscedasticity. The curve value indicates that the relationship between the independent variable and the dependent variable might not be fully captured.

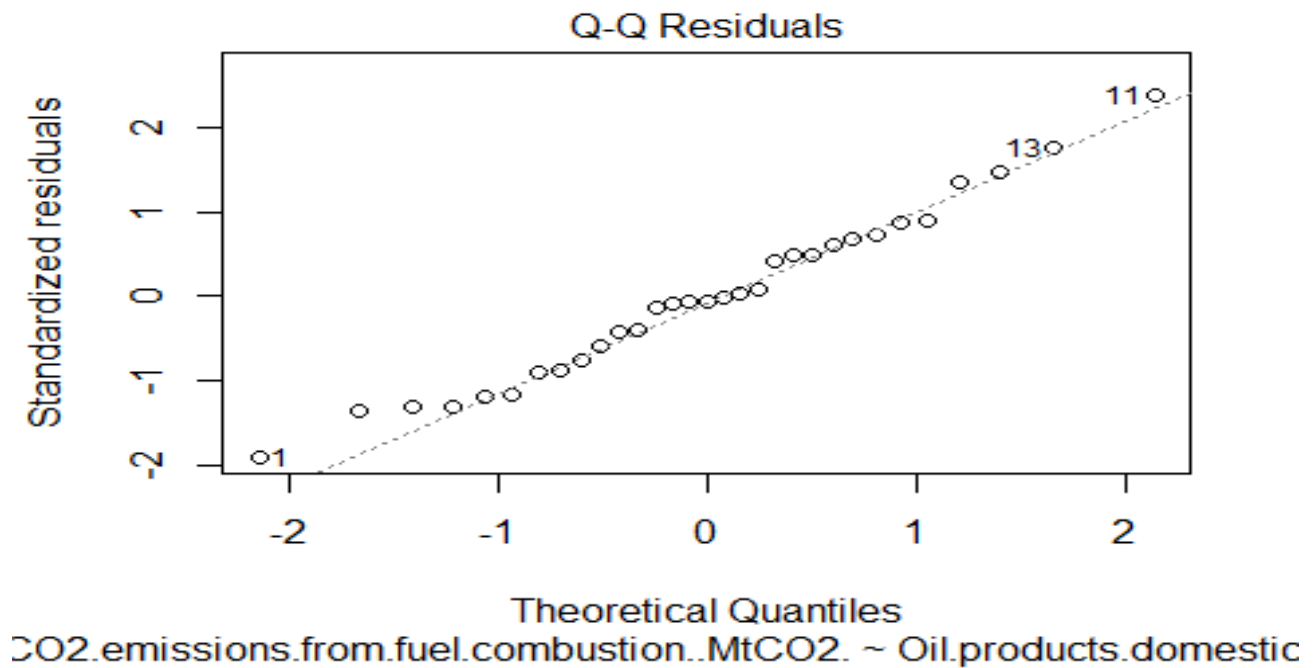
```
# Residuals vs Fitted  
rf=plot(model, which = 1)
```



CO2.emissions.from.fuel.combustion..MtCO2. ~ Oil.products.domestic

The Q-Q plot checks for normality within the data. This plot shows the points following the diagonal line closely. However, there are some deviations at the tail, especially at both extremes. These points may represent outliers or non-normality in the extreme value of the residuals

```
# Normal Q-Q  
qq=plot(model, which = 2)
```



Overall, based on the residuals, we can conclude that while there is a very strong correlation between the two variables, the model does show that some nonlinearity may exist.

## Conclusion

From the EDA analysis, we found that countries and regions tend to consume energy that has been produced domestically. The shares of electricity and natural gas consumption for total energy consumption remain relatively unchanged from 1990 to 2020 for most top-producing countries. China and Iran were both notable outliers; both showed a steep increase in shares. Coal production has been on a downward trend since 2013, although China still relies heavily on coal for energy production.

From our investigation into whether the shares of renewables in electricity production have increased, the null hypothesis was rejected. This gives us strong evidence for our alternative hypothesis, suggesting that there has been an increase in shares of renewables in electricity production. This could indicate a trend toward more renewable sources of energy.

From assessing whether there was a linear relationship between the relationship between oil consumption and CO2 emissions from fuel combustion, it was concluded that there is a strong positive association between the two variables.

## Recommendation

For our overall project, there are a few recommendations to improve the overall accuracy of the results:

- 1) For our coal consumption analysis, a comparison on a per capita basis would have yielded a stronger investigation, due to the discrepancies in population across different countries.
- 2) When investigating the validity of the model, it was found that the residual vs fitted plot had some pattern in the plot, putting the validity of the model in question. To improve the model, initial transformations could be applied to the selected variables. Some common transformations that could be used are:
  - a. Log transformation
  - b. Square root or power transformation
  - c. Polynomial regression

## Reference

- IRENASTAT Online Data Query Tool . (n.d.). Pxweb.irena.org; International Renewable Energy Agency.  
[https://pxweb.irena.org/pxweb/en/IRENASTAT/IRENASTAT\\_\\_Power%20Capacity%20and%20Generation/RESHARE\\_2024\\_H2.px/Huang, Y. \(2024, April 24\).](https://pxweb.irena.org/pxweb/en/IRENASTAT/IRENASTAT__Power%20Capacity%20and%20Generation/RESHARE_2024_H2.px/Huang, Y. (2024, April 24).)
- Council on Foreign Relations. (n.d.). China's battle against Air Pollution: An update. Council on Foreign Relations. <https://www.cfr.org/blog/chinas-battle-against-air-pollution-update>
- Select table. PxWeb. (n.d.). <https://pxweb.irena.org/pxweb/en/IRENASTAT>
- Huang, Y. (2024, April 24). China's Battle Against Air Pollution: An Update. Council on Foreign Relations. <https://www.cfr.org/blog/chinas-battle-against-air-pollution-update>
- Why Iran Consumes Five Times More Gas Than Turkey? (2024, February 1). Iran International. <https://www.iranintl.com/en/202312091166>
- Mikhail-Mks. (2020). World energy data 1990 - 2020. Kaggle.com.  
<https://www.kaggle.com/datasets/shub218/energy-data-1990-2020/data>