Statistical Models and Stochastic Methods

Biofuel analysis

Jan Izquierdo

Sergi Ocaña

Index

1. Context and Objective

In our project, the main goal is to investigate the relationship between global energy production and consumption. We are using two sizable datasets that cover the years 2000–2010. One dataset depicts the patterns of biofuel energy use in all countries during this period, while the other records the amount of biofuels produced, both expressed in thousands of barrels.

1.1 Process to achieve the objective

Our main goal is to create predictive models by using the knowledge extracted from the information related to biofuel production and consumption, to project how the most key nations will develop in terms of biofuels over the next several years. Our aim is to develop a prediction model by utilizing the past trends and patterns. Once our model has been applied in a few critical countries, we will compare the results with actual numbers and conduct a thorough analysis to assess how close our predictions are to the empirical results.

2. Datasets

2.1 Brief analysis of the datasets

For this project, we have procured two interrelated datasets encompassing global data for all countries also clustered by regions within the years 2000 to 2010. One dataset delineates biofuel production, while the other encapsulates biofuel consumption. The project's predictor variable will be historical data on the production and consumption of biofuels. These data, which cover prior years and show the development of production and consumption patterns, will be the main source of information used to project future output and usage of biofuels.

2.2 Distribution of the data

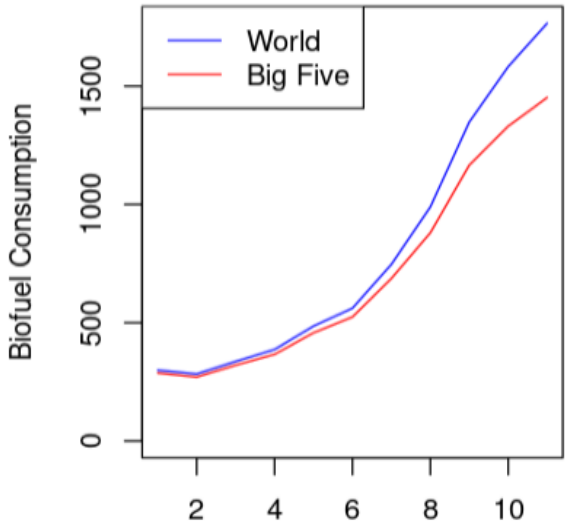
Our initial step is to start by looking at the data in the files to see what kind of distribution it has. It is clear from observation that some countries have numerical values that are noticeably higher than those of other countries. As a result, we

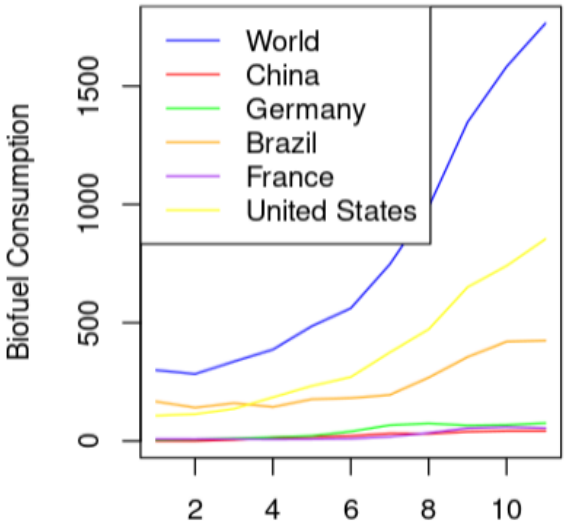
separate the top five nations and carry out an extensive analysis throughout the dataset. After making this comparison, we examine these nations percentage contributions over each year, and the results show that a very small number of countries have a major impact on the total results.

```
 [1] 95.91394 95.41884 95.30989 94.66503 94.15503 93.37546 92.01507 88.94228 86.48789 84.10223
[11] 82.21336
```

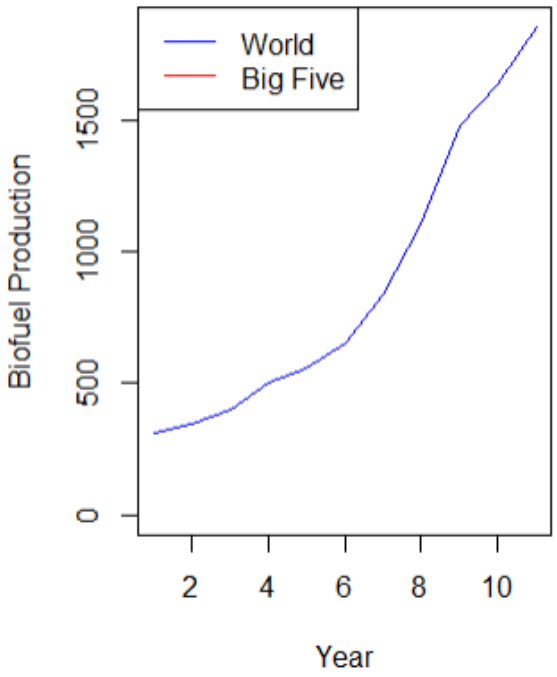
We further enhance the analytical depth by generating line plots representations to show the significance of these nations. In our case the five countries are China, France, Brazil, Germany and the United States.

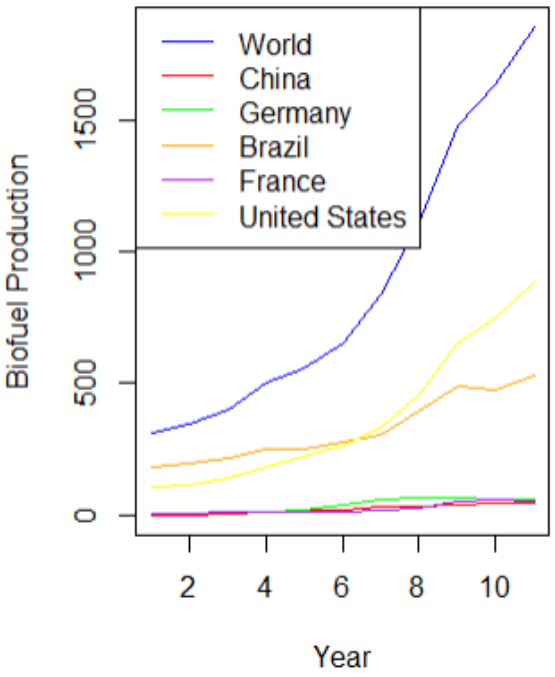
Consumption





Production



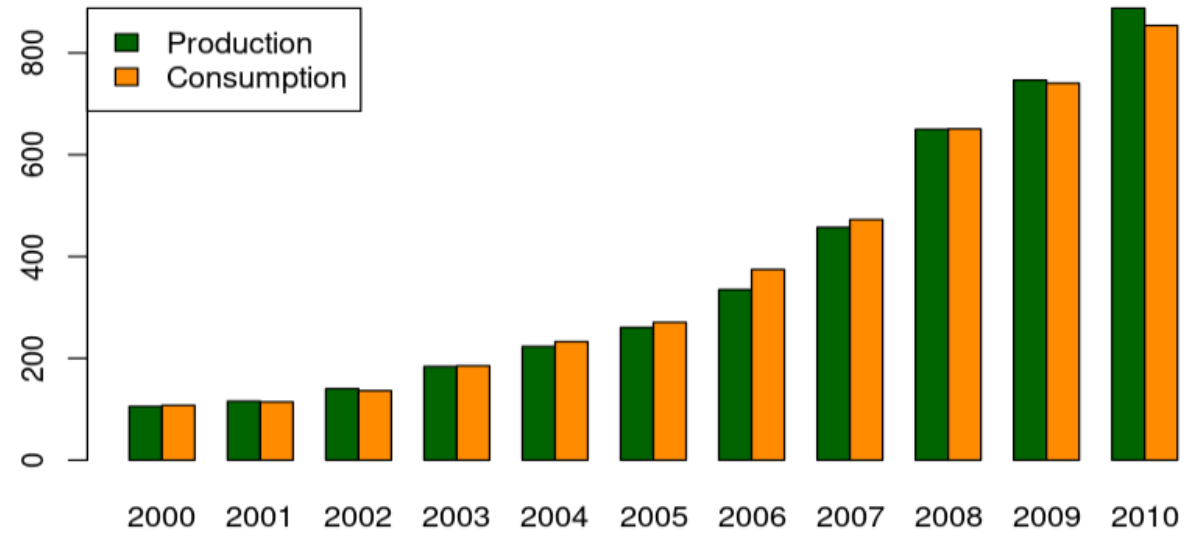


Upon reviewing the results, it becomes apparent that the data follows a skewed distribution pattern.

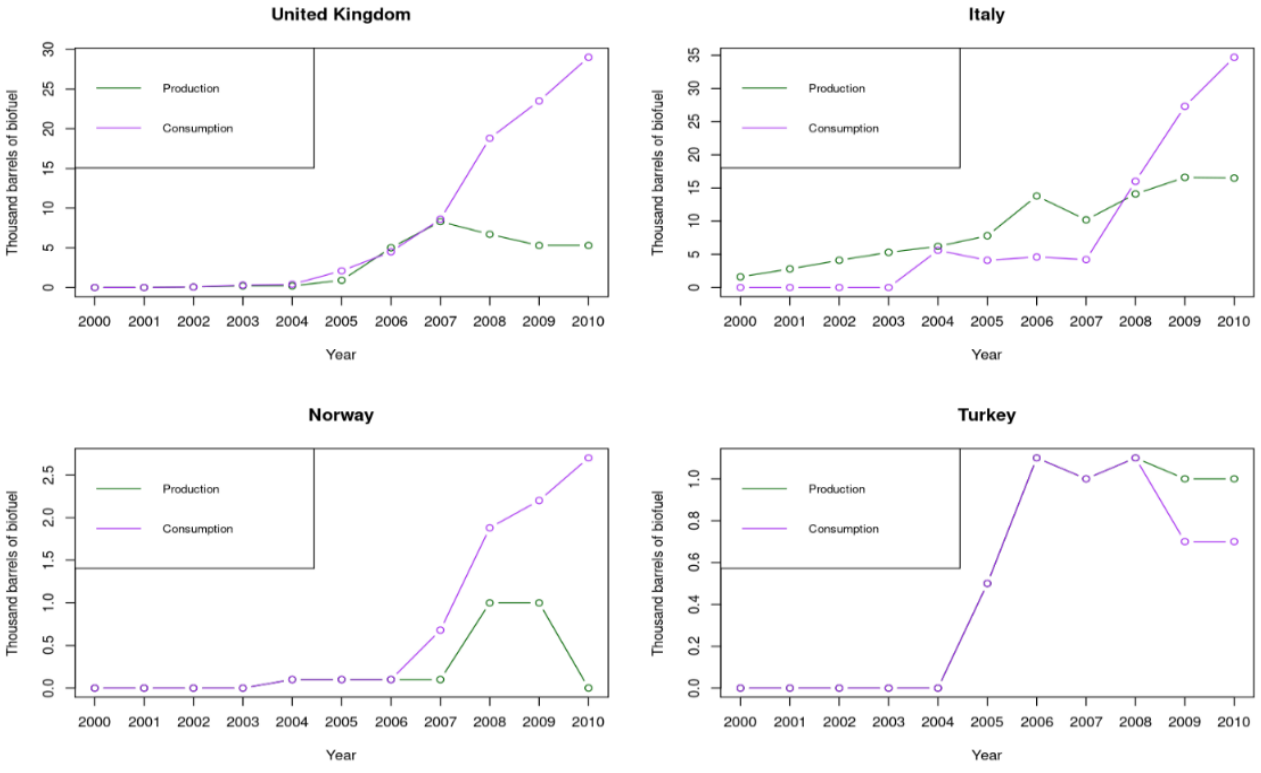
2.3 Basic plots

As we have more than 200 observations our analysis will focus on some of the most significant countries, before creating our model we are going to do some graphics in order to better understand the information we have in our hands.

First we see the trend over the past years of the biggest contributor of the table, the United States.



Not all countries were following a positive correlation as these following cases show.



For that reason we decided to make a correlation matrix for each of the selected countries so we could see the relationship between production and consumption.

```
$`United States`
            Production Consumption
Production   1.0000000   0.9982607
Consumption  0.9982607   1.0000000

$`United Kingdom`
            Production Consumption
Production   1.0000000   0.7154708
Consumption  0.7154708   1.0000000

$Italy
            Production Consumption
Production   1.0000000   0.8374989
Consumption  0.8374989   1.0000000

$Norway
            Production Consumption
Production   1.0000000   0.6179933
Consumption  0.6179933   1.0000000

$Turkey
            Production Consumption
Production    1.000000    0.974178
Consumption   0.974178    1.000000
```

In conclusion, the United States and Turkey display the strongest relationships between biofuel production and consumption, while Norway shows a relatively weaker association.

3. Linear regression

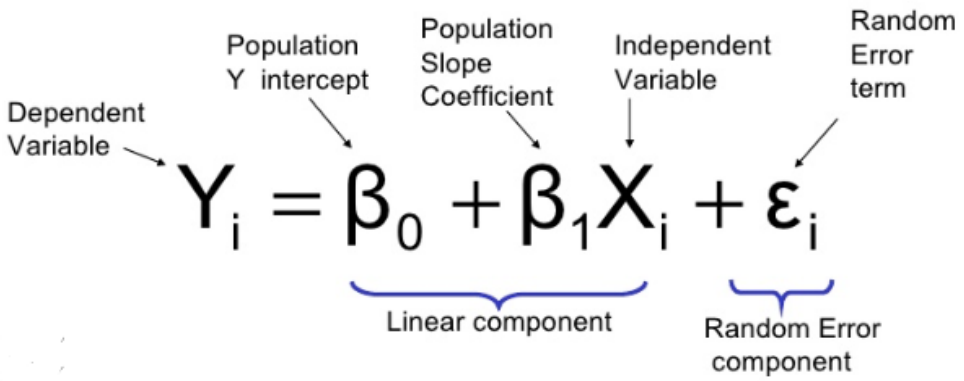
3.1 Models

In the pursuit of understanding the relationship between two variables, an introductory approach entails the utilization of a linear model. This model, which we have used in previous assignments of this course, aims to capture and describe the link between variables.

Central to the theoretical foundation of linear regression is the portrayal of 1 as a constant representing the effect exerted by the predictor on the response, while 0 serves as another constant term. This conceptual framework denotes that for every incremental unit change in x, the response undergoes an alteration by ᵞ1 units. Within this framework, εi symbolizes the associated error intrinsic to the model, while Xi signifies the predictor variable being assessed.

This model's essence lies in fitting a line that effectively summarizes the association observed amidst the variables. The line serves as a representation of the core relationship between the response and predictor. This model serves as a method to

summarize the general trend among a bunch of data points, acknowledging that the variables aren't tightly interconnected. It aims to describe how two things might move together but doesn't insist that one always dictates the other. Instead, it's about capturing a sense of their relationship, acknowledging there's a link without expecting it to be a precise, step-by-step connection. It's a way to simplify and understand how these variables generally behave in relation to each other.



3.2 Prediction with the model

To predict what the biofuel consumption and production would be in a given year we created a function that uses a linear regression model, this function uses the consumption or production data from the database for every year to create the model that will be used to predict the data from a certain country in the year that you prefer, to do this the function uses the given data from all years from the selected country and creates the model from that.

3.2.1 Prediction based on year

Prediction for a year

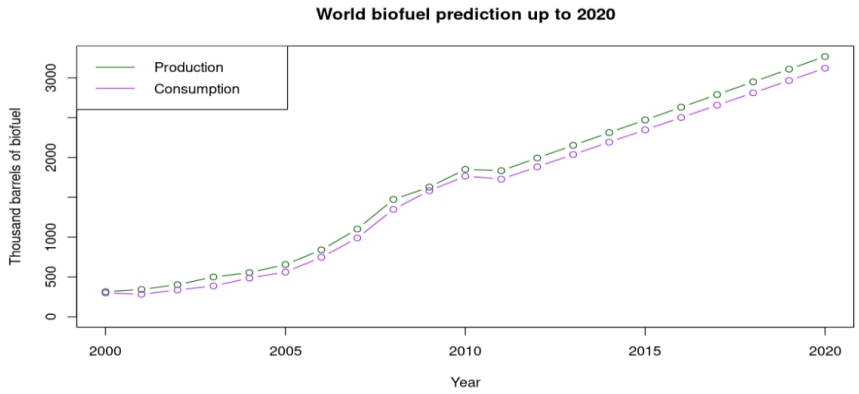
The function performs 2 linear models, one for production and one for consumption, the linear model uses the data of the production database or the consumption database for a given region to project the data for production and consumption of a certain year. Then the function analyzes the r squared function of the model to know how much of the variance in the data is explained by the model.

```
For World in 2015
Production
       fit     lwr      upr
1 2477.929 1941.52 3014.338
Consumption
       fit      lwr      upr
1 2350.405 1758.835 2941.975

R-squared for Production Model: 0.9173619
R-squared for Consumption Model: 0.8957997
```

In this example we predicted the data for the world in 2015, our model estimates our production to be 2478 thousand barrels of biofuel, the maximum would be 3014 and the minimum 1942. For consumption our estimate would be 2350, the maximum 2942 and our minimum 1759. The variance of the data explained by this model according to r squared is a 92% in the case of production and a 90% in the case of consumption.

Prediction up to a year

This linear model enables a methodical procedure in the predictive study of biofuel production and consumption. The db_p and db_c datasets serve as a reference for historical production and consumption data for a given nation and are extracted using the clean_pred function to start the prediction process. Then the function uses the clean_plot function to visualize a projection of the production and consumption values of the past and future years up to the year you use as input, in this case 2020.



R-Squared

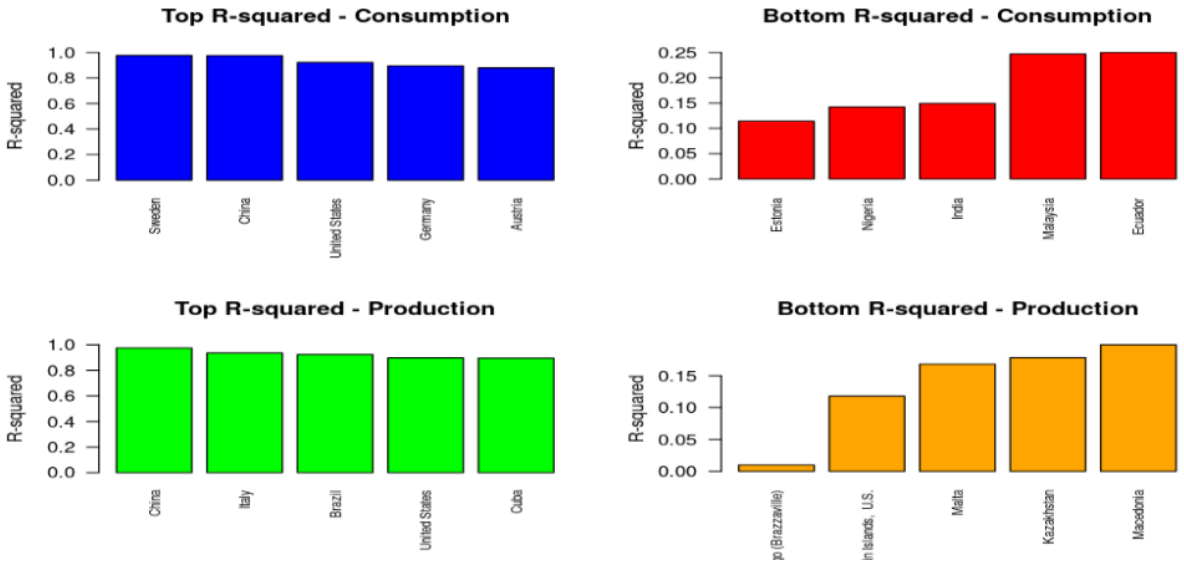
After doing our model we have conducted an R-squared test in order to clarify how much of the variability of the data can be explained by the model. For that we have

generated a function that iterates for each column (countries) except the last one which is world and the regions column in order to not duplicate data and lead to confusion in the output function. Although we have columns with NA's we just ignore them and continue. When we already have all the R-squared values we simply do a mean to see how well the data fits in our model.

Results	Datasets
[1] 0.7778141 [1] 0.8605092	Consumption Production

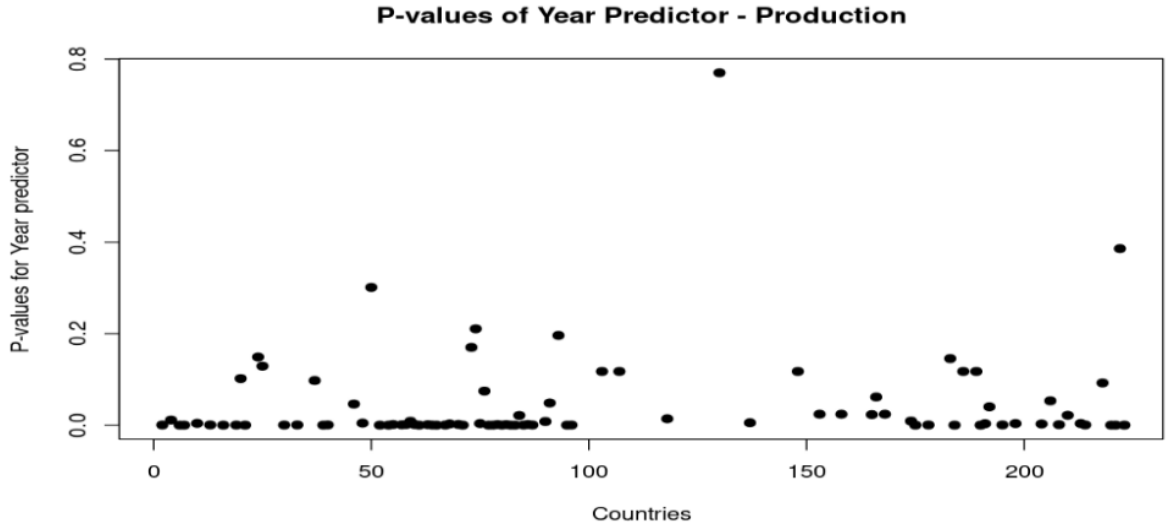
As we can see from the obtained results the linear model fits the data better in the production dataset.

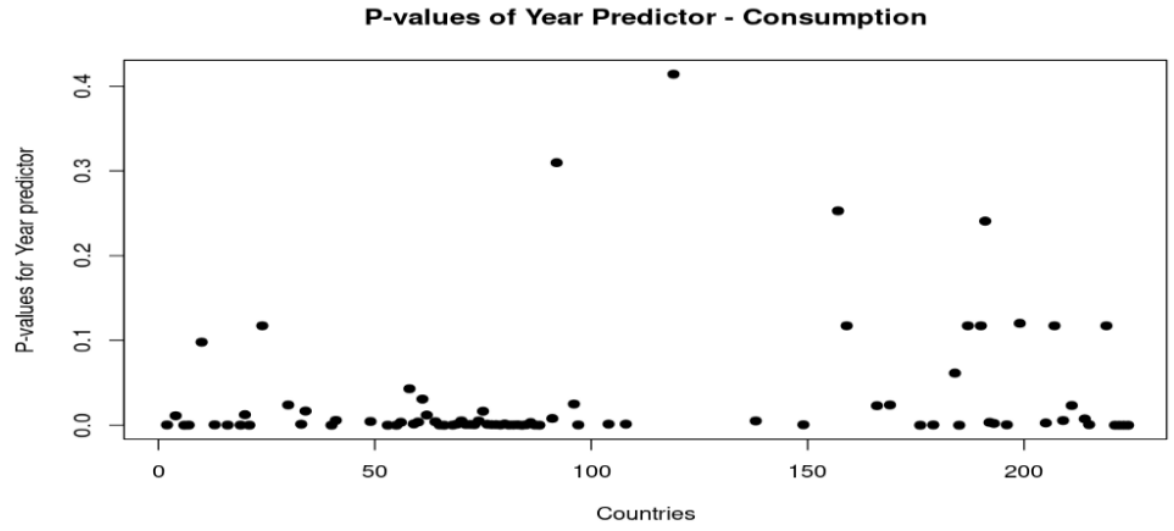
Reusing the function to calculate the mean of the R-Squared we also can get which countries have a better fit in the model and which have a worse fit.



P-value

In order to assure us that all the calculations in our models were correct we create a function that gives us the p-value for each country and plotted the results of each country to get a better understanding.





The predictor variable year has consistently low p-values across all the countries, indicating its statistical importance. This implies a strong correlation between year and the research outcome which are the future production and consumption values for each nation.

3.2.2 Prediction based on production or consumption

This function performs 2 linear models, one for production and one for consumption. Using the production or consumption data to create a data frame where each year corresponds with a value and using the model to project this relation of year and value the function can find a year where the requested value matches consumption and a year where it matches production.

```
Production 2000 for World can be found in the year 2011
Consumption 2000 for World can be found in the year 2012
```

In this case the model has predicted that in the year 2011 the global production will be of 2000 thousands of barrels of biofuel in the world, while 2000 thousands of barrels will be consumed in 2012.

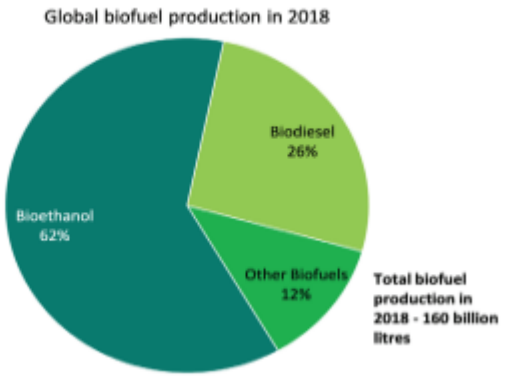
```
Production 50 for Spain can be found in the year 2021
Consumption 50 for Spain can be found in the year 2016
```

This can highlight the differences between the production and consumption rate of a country, for example the production of Spain will be 50 in the year 2021, while the consumption will be 50 in 2016. This lets us see the difference between Spain's production and consumption rate. We can observe that the consumption rate is much higher, as it will reach the chosen value quicker.

4. Conclusion

4.1 Comparison with real data

Production



In 2018, 160 billion litres of biofuels were produced globally. Bioethanol is the largest biofuel globally with a share of 62% followed by FAME biodiesel at 26%. Rest of the biofuels including HVO (Hydrogenated vegetable oil), renewable diesel, cellulosic ethanol etc. had a share of 12%. Americas dominate the biofuel production globally. North and South America together produce 75% of all biofuels globally with Europe having a share of 14%. In 2018, 59.3 billion m3 of biogas was produced globally with an equivalent energy content of 1.36 EJ. During 2000 – 2018, the sector experienced an annual growth rate of 9%.

To follow the process of this conclusion the data you need to know is

- The standard barrel is equivalent to 160 liters.
- The global production of biofuel in 2018 was 160 billion liters.

According to our prediction in 2018 the world production is around 2949.3 thousand barrels per day. We have to transform our real world data from billions of liters to thousands of barrels per day to be able to compare it with our real world results.

$$160\ billion\ liters\ a\ year\ \frac{10^9\ liters}{1\ billion\ liters}\frac{1\ barrel}{160\ liters}\frac{1\ thousand\ barrels}{1000\ barrels} = 10^6\ thousands\ of\ barrels\ a\ year$$

$$10^6 thousands\ of\ barrels\ a\ year\ \frac{1year}{365\ days} = 2739.7\ thousands\ of\ barrels\ per\ day$$

If we compare the predicted output for 2018 with our real world data we can see that our prediction model is accurate, as the data predicted by our function was of 2949.3 thousands of liters per day

```
Production
       fit     lwr      upr
1 2949.309 2337.56 3561.057
```

We can check the percentage of error in this prediction

$$\frac{2949.3-2739.7}{2739.7} * 100 = 7.65\%$$

Our prediction had a 7.65% margin of error in this case

```
Production 2740 for World can be found in the year 2016
95% Confidence Interval for Production prediction: 2013 to 2018
```

To check if our prediction model from production or consumption data works, we introduce the data that we know belongs to 2018 to the function, analyzing the result we can see that our model is wrong by 2 years as according to the function our data belongs to 2016. The function also provides us with the confidence interval for this prediction, we can see that our data does fit inside this interval.

To reassess our evaluations we do the same tests again for South, Central and North America, we know that the real world data from the americas is a 75% of the

world total, so $2739.7 * 0.75 = 2054.8$ thousands of barrels per day produced in the Americas. We will use this data to check the error of our function in order to create a column for all the data for the production values of the Americas, then we run the function for 2055 thousands of barrels per day for this new column, we get that our production is from the year 2015, so we can see that our predictions aren't entirely correct and we also get a confidence interval for our prediction.
The errors in this model prediction might be because of unpredictable events such as wars, pandemics, economic recessions, or others.

```
Production 2055 for North & Central & South America can be found in the year 2015
95% Confidence Interval for Production prediction: 2012 to 2017
```

```
Production 2005 for North America & Central & South America can be found in the year 2014
Consumption 2005 for North America & Central & South America can be found in the year 2016
```

In the case of Europe using the same formula we get that as per real world data the production of barrels will be approximately 385 thousands per day. The result from our prediction is that this quantity will arrive in 2015 with a margin of 2 years, nevertheless the prediction is wrong by a margin of 3 years in this case, as our real world data is from 2018.
We have observed that the errors in this model prediction might be because of unpredictable events such as wars, pandemics, economic recessions, or others.

```
Production 385 for Europe can be found in the year 2015
95% Confidence Interval for Production prediction: 2013 to 2017
```

Consumption

In our analysis of the consumption model we have found an error. After doing the calculations with the same conversion factors we find that for Germany in 2018 the consumption was around 72 thousand barrels per day. Once done this if we put the data in our model the result year is 2008. The reason why this happens, we think, is because if we plot the data for consumption in Germany we can see that 72 has already been surpassed in 2008, so the model looks into the data and doesn't make any prediction.

```
Consumption 72 for Germany can be found in the year 2008
```

```
[1]  4.9  6.8 10.8 15.7 21.9 40.7 67.0 74.0 66.0 68.0 75.5
```

In the next case, Brazil, the country always presents a positive trend in the consumption values so, the model works correctly. We found data for the year 2018 and as we did the calculations to convert to barrels, we got a value of 632. With the country fitted we also get wrong with our prediction but this time we are inside the confidence interval.

```
Production 632 for Brazil can be found in the year 2013
Consumption 632 for Brazil can be found in the year 2016
```

5. Bibliography

[1] Biofuel consumption production . data.world. (2016, November 11). https://data.world/doe/biofuel-consumption-production

[2] Brazil: Biofuels Annual (2023) USDA Foreign Agricultural Service. Available at: https://fas.usda.gov/data/brazil-biofuels-annual-8 (Accessed: 23 November 2023).

[3] Bioenergy in Germany facts and figures 2020 - FNR. Available at: https://www.fnr.de/fileadmin/allgemein/pdf/broschueren/broschuere_basisdaten_bioenergie_2020_engl_web.pdf (Accessed: 01 December 2023).

[4] Global Bioenergy Statistics 2020 - WBA - . Available at:

https://www.worldbioenergy.org/uploads/201210%20WBA%20GBS%202020.pdf (2023, December 03).

6. Appendix

Analysis of biofuel production and consumption

Jan Izquierdo & Sergi Ocaña

2023-11-26

Importing the data

```
data_p<-read.csv("./data/prod0010thousandbarrelsperday.csv")
data_c<-read.csv("./data/cons0010thousandbarrelsperday.csv")
```

Fixing the data

```
db_p <- data.frame(t(data_p[-1]))
colnames(db_p) <- data_p[, 1]
db_c<-data.frame(t(data_c[-1]))
colnames(db_c) <- data_c[, 1]

# Function to remove non-numeric characters at the beginning of each row
remove_X_from_row_names <- function(df) {
  rownames(df) <- gsub("^X", "", rownames(df))
  return(df)
}

# Remove 'X' characters from row names of db_p and db_c
db_p <- remove_X_from_row_names(db_p)
db_c <- remove_X_from_row_names(db_c)
#sapply(db_p, class)
```

Transforming the data, from character to numeric

```
db_p[] <- lapply(db_p, function(x) {
  suppressWarnings(as.numeric(as.character(x)))
})

db_c[] <- lapply(db_c, function(x) {
  suppressWarnings(as.numeric(as.character(x)))
})

#sapply(db_p, class)
#sapply(db_c, class)
```

Distribution across all data of consumption and production

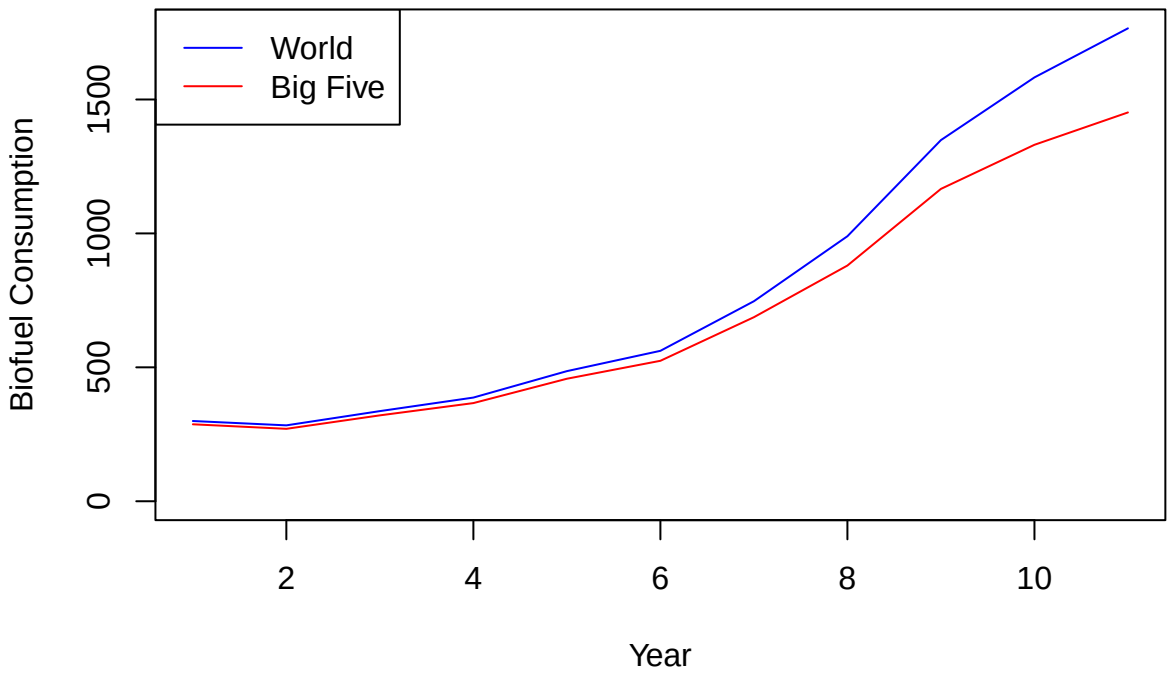
Selection of countries of who we will perform analysis later

```
#For consumption
sums_c <- colSums(db_c, na.rm = TRUE)

db_c$World <- rowSums(db_c[, c("Asia & Oceania", "Europe", "North America", "Central & South America",
              "Africa")], na.rm = TRUE)

db_c$sum_of_columns <- rowSums(db_c[, c("China", "Germany", "Brazil", "France", "United States")],
              na.rm = TRUE)

db_c$percentage <- (db_c$sum_of_columns / db_c$World) * 100

#For production
sums_p <- colSums(db_p, na.rm = TRUE)


db_p$World <- rowSums(db_p[, c("Asia & Oceania", "Europe", "North America", "Central & South America",
              "Africa")], na.rm = TRUE)

db_p$sum_of_columns <- rowSums(db_c[, c("China", "Germany", "Brazil", "France", "United States")],
              na.rm = TRUE)

db_p$percentage <- (db_p$sum_of_columns / db_p$World) * 100
```

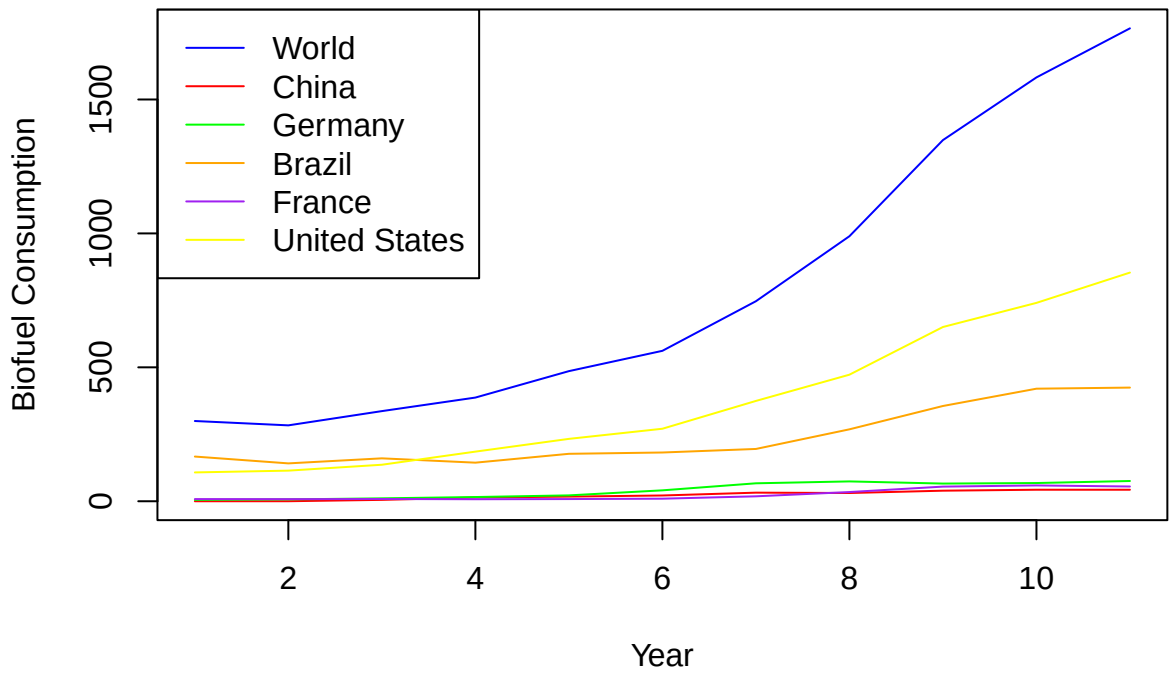
Plot the selected countries' consumption against the World

For consumption

```
# Plot 1: Line plot
plot(db_c$World, type = "l", col = "blue", xlab = "Year", ylab = "Biofuel Consumption",
     ylim = c(0, max(db_c$World, na.rm = TRUE)))
lines(db_c$sum_of_columns, type = "l", col = "red")
legend("topleft", legend = c("World", "Big Five"), col = c("blue", "red"), lty = 1)
```



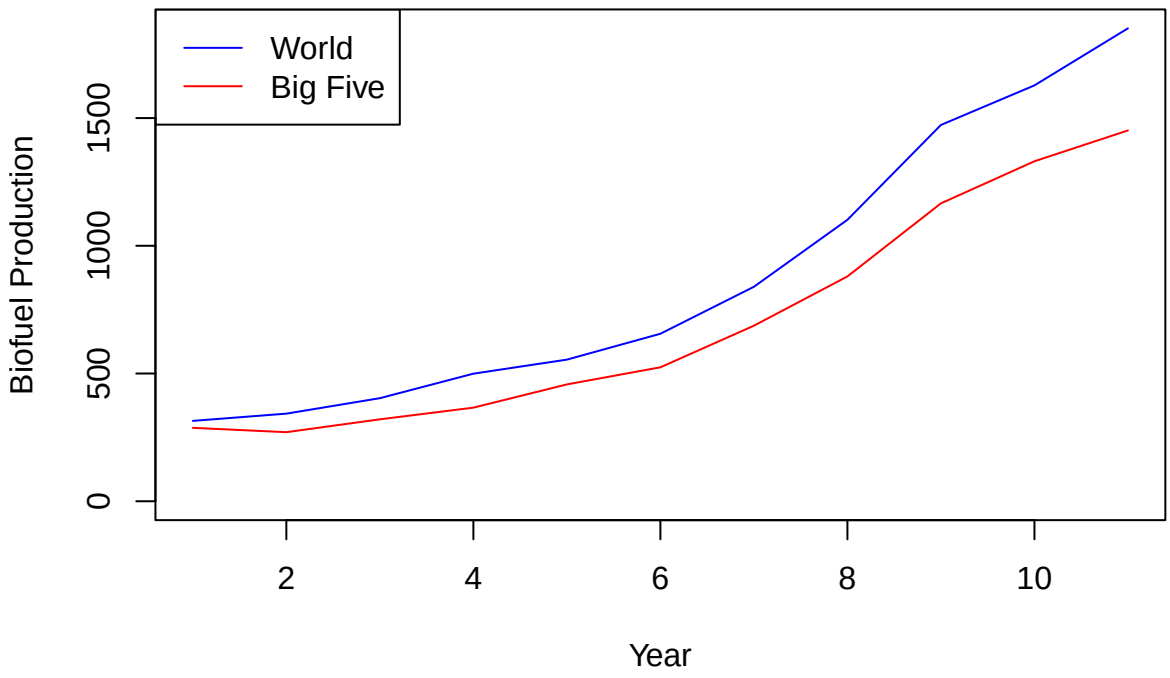

```r
# Plotting a line plot for selected columns
plot(db_c$World, type = "l", col = "blue", xlab = "Year", ylab = "Biofuel Consumption",
     ylim = c(0, max(db_c$World, na.rm = TRUE)))

lines(db_c$China, type = "l", col = "red")
lines(db_c$Germany, type = "l", col = "green")
lines(db_c$Brazil, type = "l", col = "orange")
lines(db_c$France, type = "l", col = "purple")
lines(db_c$`United States`, type = "l", col = "yellow")

legend("topleft", legend = c("World", "China", "Germany", "Brazil", "France", "United States"),
       col = c("blue", "red", "green", "orange", "purple", "yellow"), lty = 1)
```



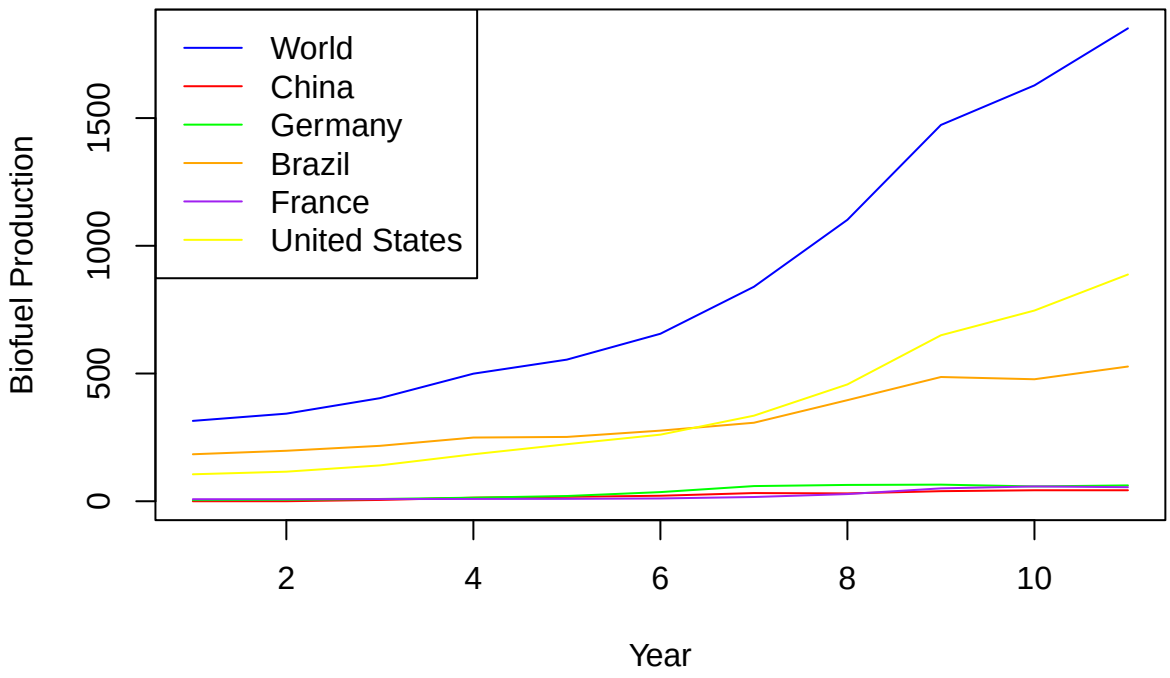

```
## The dominance of just a few countries in contributing to a significant portion of
## the world's biofuel consumption indicates a highly skewed distribution. This skewed distribution
## suggests a high level of inequality or concentration, where a small number of countries,
## holding a substantial percentage of the total consumption
```

For production

```
# Plotting for db_p
# Plot 1: Line plot
plot(db_p$World, type = "l", col = "blue", xlab = "Year", ylab = "Biofuel Production",
     ylim = c(0, max(db_p$World, na.rm = TRUE)))
lines(db_p$sum_of_columns, type = "l", col = "red")
legend("topleft", legend = c("World", "Big Five"), col = c("blue", "red"), lty = 1)
```

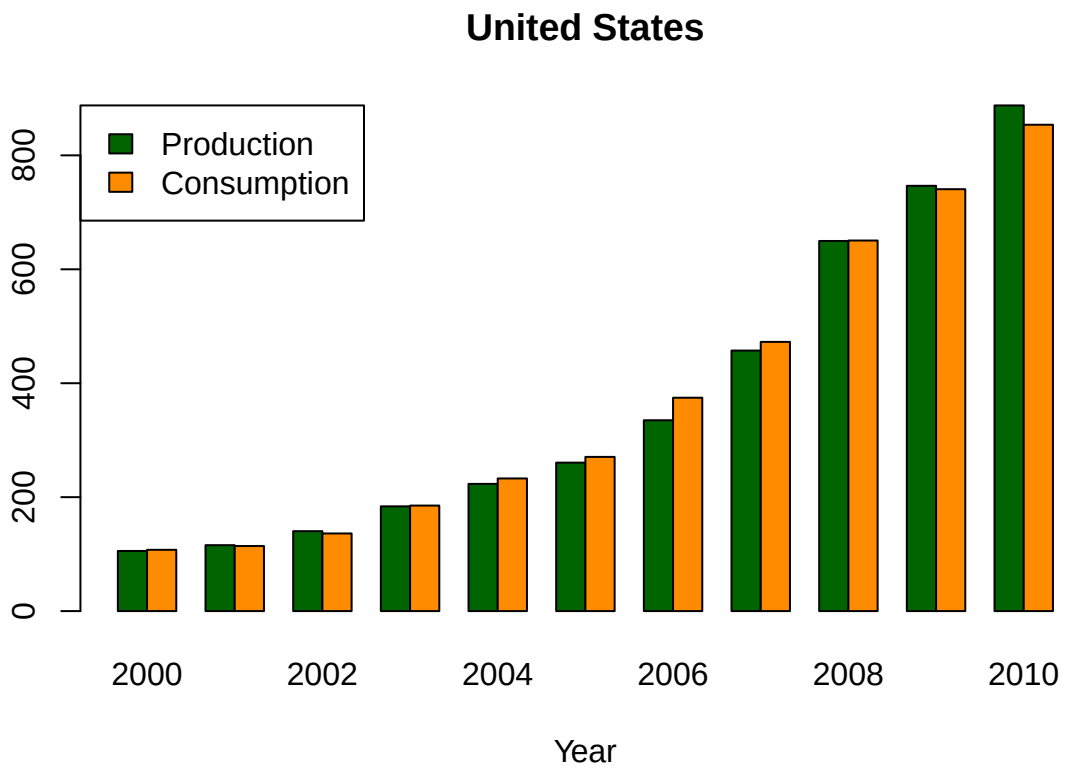



```r
# Plotting a line plot for selected columns
plot(db_p$World, type = "l", col = "blue", xlab = "Year", ylab = "Biofuel Production",
     ylim = c(0, max(db_p$World, na.rm = TRUE)))

lines(db_p$China, type = "l", col = "red")
lines(db_p$Germany, type = "l", col = "green")
lines(db_p$Brazil, type = "l", col = "orange")
lines(db_p$France, type = "l", col = "purple")
lines(db_p$`United States`, type = "l", col = "yellow")

legend("topleft", legend = c("World", "China", "Germany", "Brazil", "France", "United States"),
       col = c("blue", "red", "green", "orange", "purple", "yellow"), lty = 1)
```

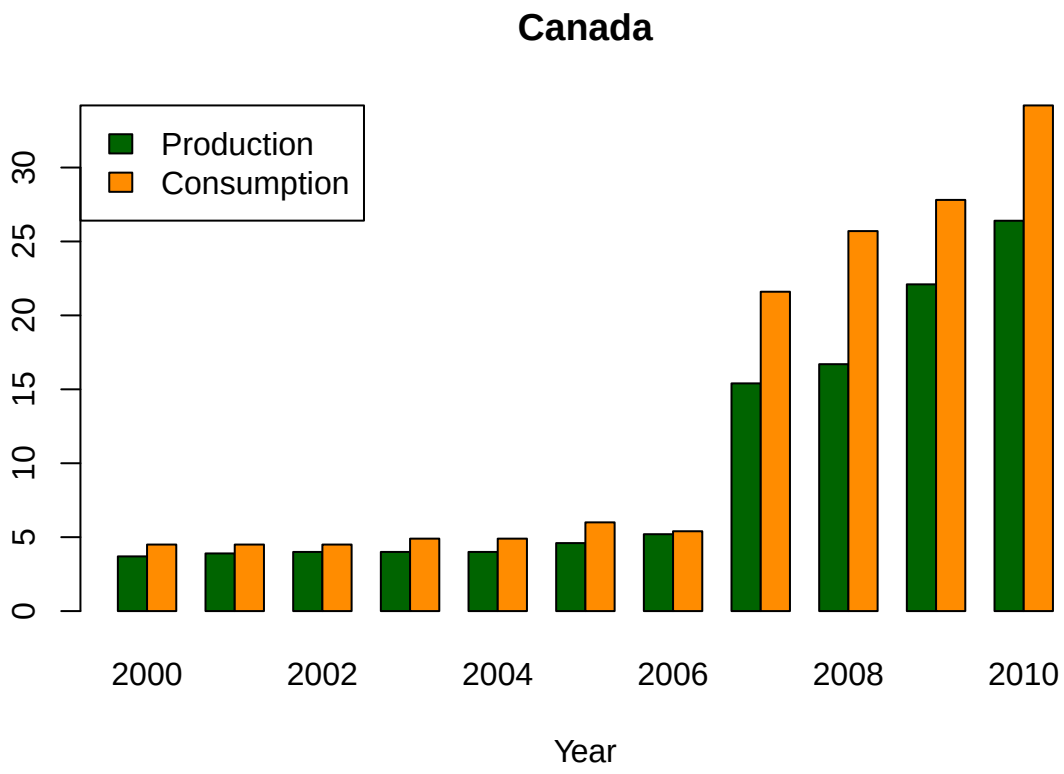


Data representation of United States of America and Canada

```
combined_freq<-t(cbind(db_p$`United States`, db_c$`United States`))
barplot(combined_freq, beside=T, col=c("darkgreen","darkorange"),
        legend.text = c("Production", "Consumption"),args.legend = list(x = "topleft"),
        xlab="Year", names.arg=rownames(db_p), main="United States")
```

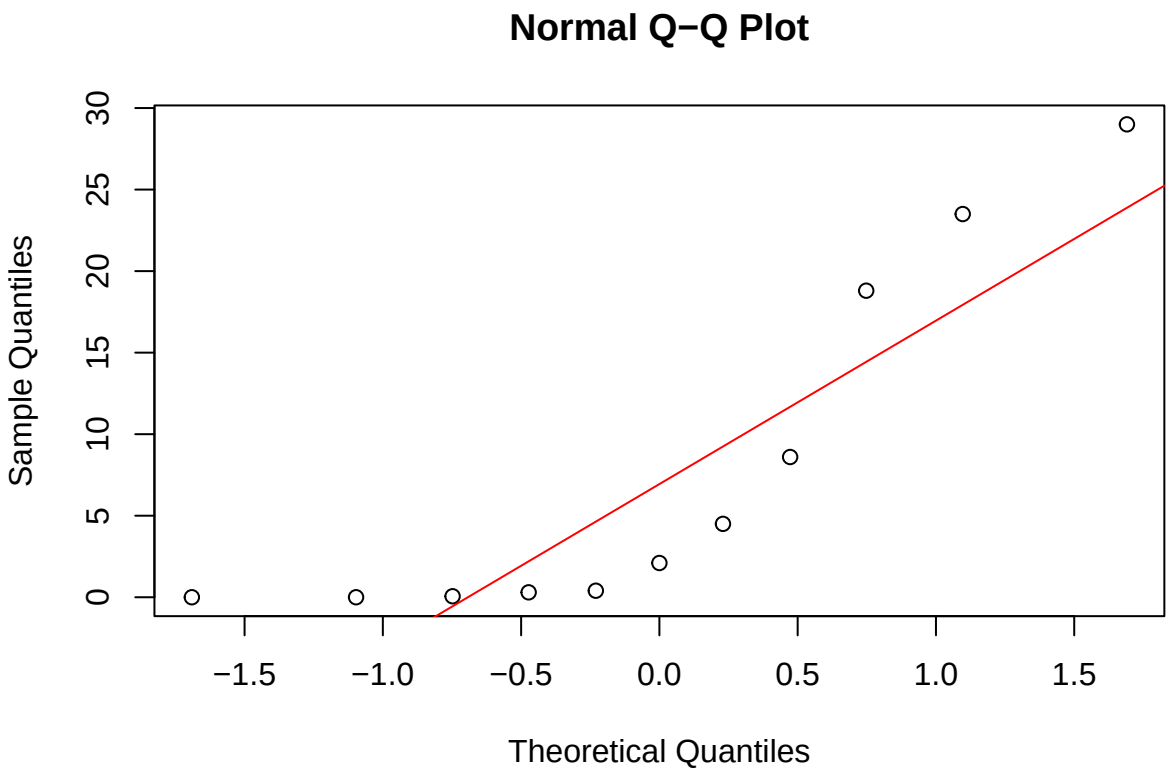



```
combined_freq<-t(cbind(db_p$Canada, db_c$Canada))
barplot(combined_freq, beside=T, col=c("darkgreen","darkorange"),
        legend.text = c("Production", "Consumption"),args.legend = list(x = "topleft"),
        xlab="Year", names.arg=rownames(db_p), main="Canada")
```



Does it follow normal distribution??

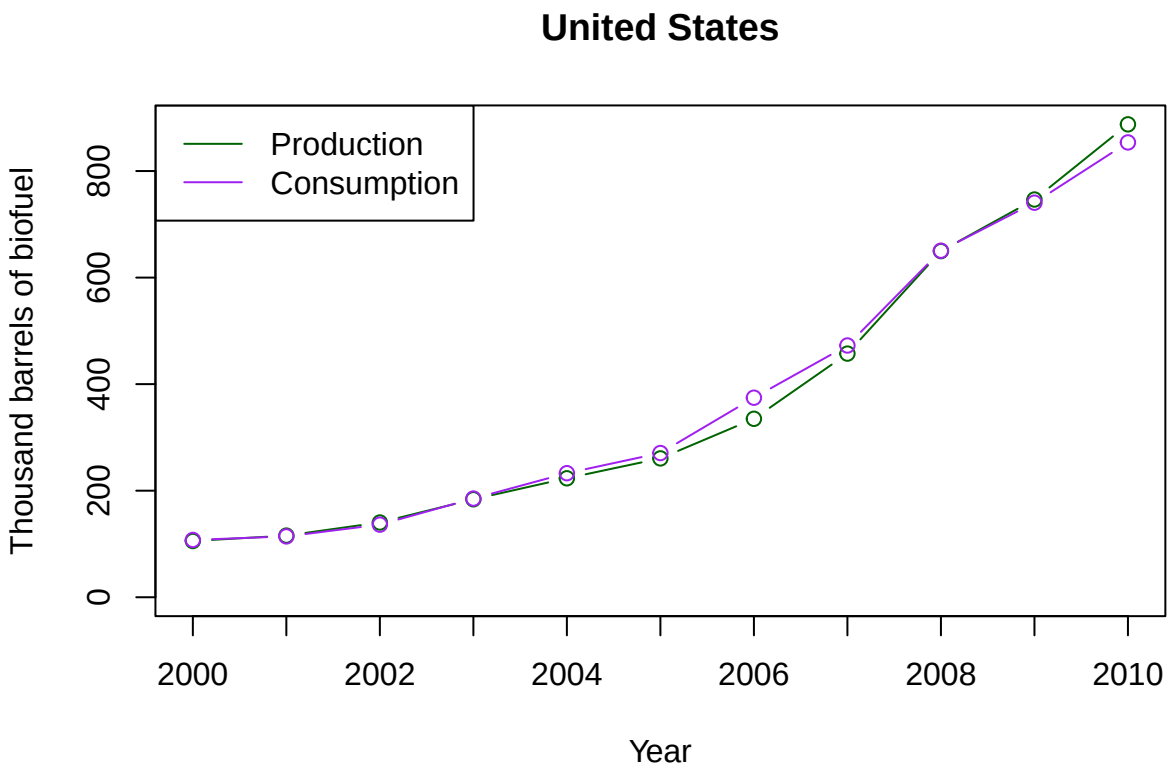
```
# Extract 'United States' consumption data
us_consumption <- as.numeric(db_c[, 'United Kingdom'])

# Q-Q plot for 'United States' consumption
qqnorm(us_consumption)
qqline(us_consumption, col = "red")  # Add reference line for normal distribution
```



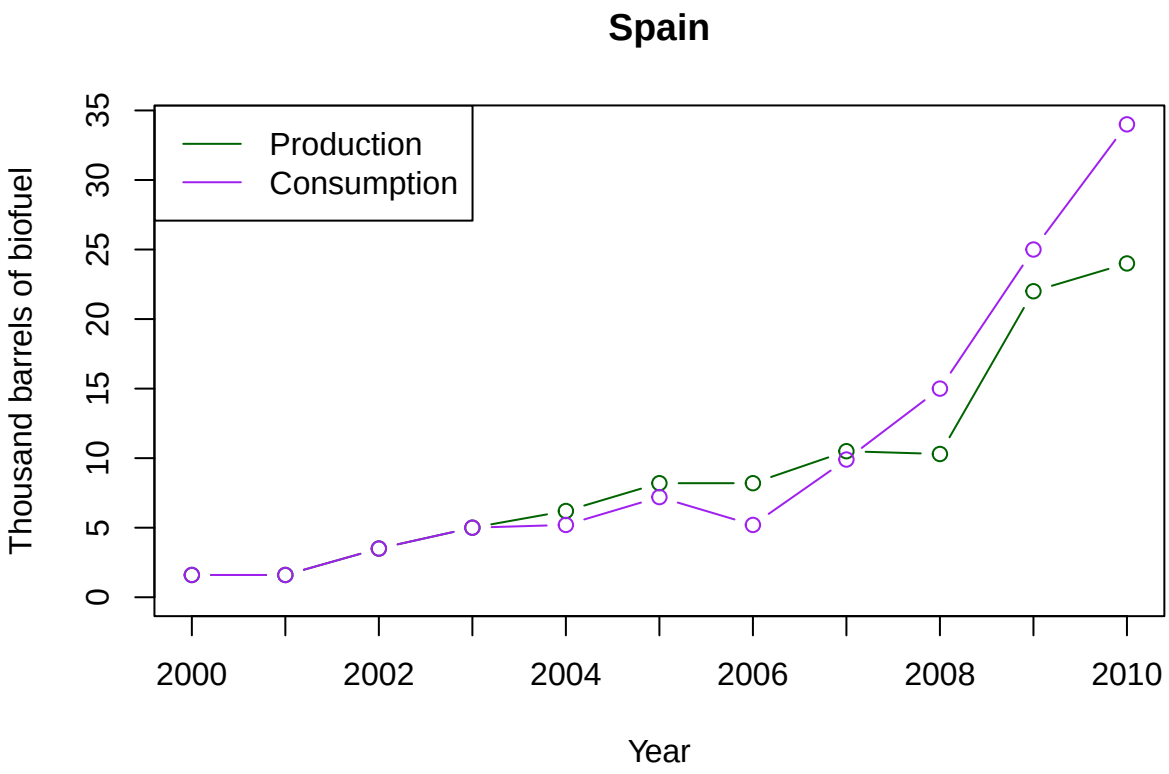

```
## We can observe that our data follows a normal distribution
```

Plotting the data from a country

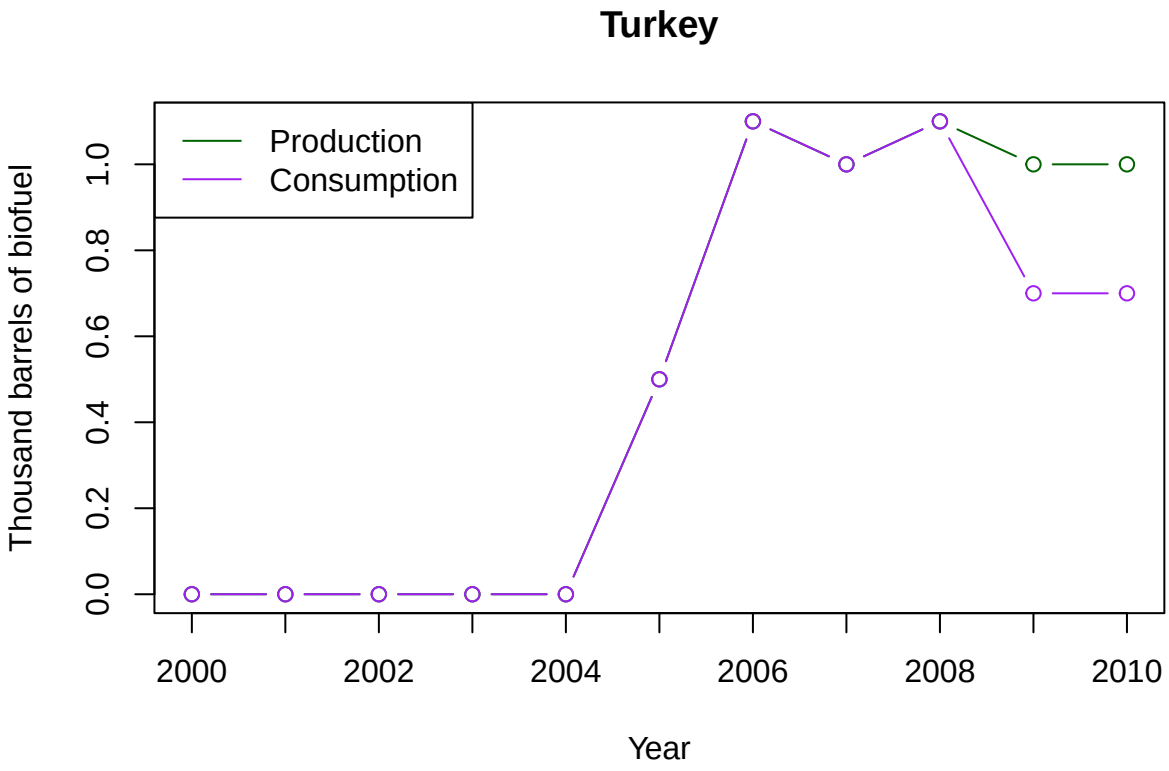
```
plot_country <- function(country_name) {
  max_value <- max(max(db_p[[country_name]], na.rm = TRUE), max(db_c[[country_name]],
         na.rm = TRUE), na.rm = TRUE)
  plot(db_p[[country_name]], col="darkgreen", type="b", ylab="Thousand barrels of biofuel",
         xlab="Year", xaxt="n", ylim = c(0, max_value))
  axis(1, at = 1:length(rownames(db_p)), labels = rownames(db_p))
  legend("topleft", legend = c("Production", "Consumption"), col = c('darkgreen', 'purple'),
         lty = 1)
  lines(db_c[[country_name]], col="purple", type="b")
  title(main = country_name)
}

# Usage example
plot_country("United States")
```




```
plot_country("Spain")
```




```
plot_country("Turkey")
```



##Correlations

```r
# Function to calculate correlation matrix for specified countries
calculate_correlation_matrix <- function(data, countries) {
  correlation_matrices <- list()

  for (country in countries) {
    country_data <- data.frame(
      Year = as.numeric(rownames(db_c)),
      Production = db_p[[country]],
      Consumption = db_c[[country]]
    )

    correlation_matrix <- cor(country_data[, c("Production", "Consumption")])

    # Store correlation matrices along with country names
    correlation_matrices[[country]] <- correlation_matrix
  }

  return(correlation_matrices)
}

# Specified countries
specified_countries <- c("United States", "United Kingdom", "Italy", "Norway", "Turkey")

# Generating correlation matrices for specified countries
correlation_matrices_specified <- calculate_correlation_matrix(db_c, specified_countries)
```

```
correlation_matrices_specified
```

```
## $`United States`
##             Production Consumption
## Production   1.0000000   0.9982607
## Consumption  0.9982607   1.0000000
##
## $`United Kingdom`
##             Production Consumption
## Production   1.0000000   0.7154708
## Consumption  0.7154708   1.0000000
##
## $Italy
##             Production Consumption
## Production   1.0000000   0.8374989
## Consumption  0.8374989   1.0000000
##
## $Norway
##             Production Consumption
## Production   1.0000000   0.6179933
## Consumption  0.6179933   1.0000000
##
## $Turkey
##             Production Consumption
## Production    1.000000    0.974178
## Consumption   0.974178    1.000000
```

Prediction of production and consumption for a country for any year

```
db_p$Year <- as.numeric(rownames(db_p))
db_c$Year <- as.numeric(rownames(db_c))

pred <- function(country, year) {
  # Production
  p_country <- db_p[[country]]
  p_model <- lm(p_country ~ Year, data = db_p)
  resp <- predict(p_model, newdata = data.frame(Year = year), interval = "prediction")

  # Consumption
  c_country <- db_c[[country]]
  c_model <- lm(c_country ~ Year, data = db_c)
  resc <- predict(c_model, newdata = data.frame(Year = year), interval = "prediction")

  # Calculate R-squared for production and consumption models
  p_r_squared <- summary(p_model)$r.squared
  c_r_squared <- summary(c_model)$r.squared

  cat("Production\n"); print(resp)
  cat("Consumption\n"); print(resc)

  cat("\nR-squared for Production Model:", p_r_squared, "\n")
```

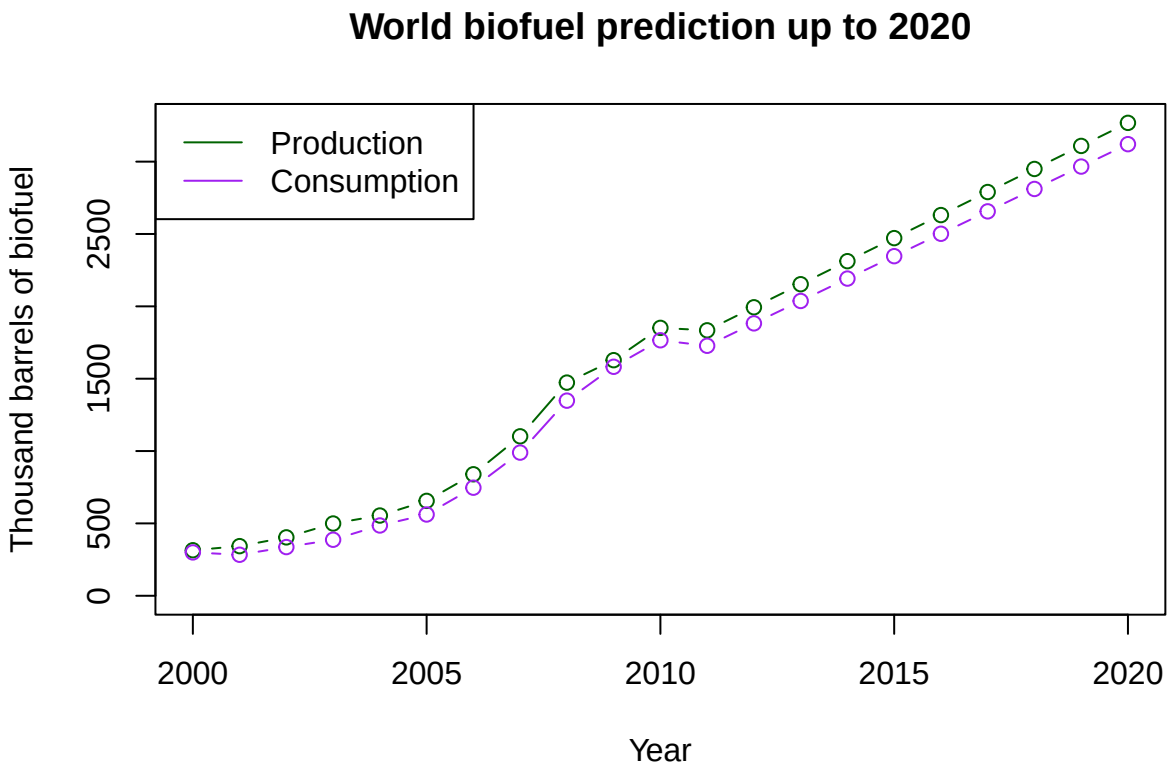
```
  cat("R-squared for Consumption Model:", c_r_squared, "\n")

}

pred("Europe", 2019)
```

```
## Production
##        fit      lwr      upr
## 1 464.6996 375.2637 554.1356
## Consumption
##        fit      lwr      upr
## 1 589.6636 440.0804 739.2468
##
## R-squared for Production Model: 0.9358675
## R-squared for Consumption Model: 0.8985361
```

Plot of the prediction for a country up to a year

```
clean_pred<-function(country, year){
 #production
 p_country<-db_p[[country]]
 p_model <- lm(p_country ~ Year, data = db_p)
 resp<- predict(p_model, newdata = data.frame(Year=year))
 #consumption
 c_country<-db_c[[country]]
 c_model <- lm(c_country ~ Year, data = db_c)
 resc<- predict(c_model, newdata = data.frame(Year=year))
 return(c(resp[[1]], resc[[1]]))
}

clean_plot<-function(country, p_country, c_country, total_years){
  max_value <- max(max(p_country, na.rm = TRUE), max(c_country, na.rm = TRUE), na.rm = TRUE)
  plot(total_years,p_country, col="darkgreen", type="b", ylab="Thousand barrels of biofuel",
       xlab="Year", ylim = c(0, max_value))
  legend("topleft", legend = c("Production", "Consumption"), col = c('darkgreen', 'purple'),
       lty = 1)
  lines(total_years,c_country, col="purple", type="b")
  name<-paste(country, "biofuel prediction up to", total_years[length(total_years)], sep=" ")
  title(main = name)
}

predict_plot<-function(country, year){
  c=2011
  total_years=(rownames(db_p))
  p_country<-db_p[[country]]
  c_country<-db_c[[country]]
  while (c<=year){
    #add years to total year counter
    total_years<-(c(total_years, c))
    #calculate production for current year
    p_country<-c(p_country, clean_pred(country, c)[1])
```

```
    #calculate consumption for current year
    c_country<-c(c_country, clean_pred(country, c)[2])
    c=c+1
  }
  clean_plot(country, p_country, c_country, total_years)
}
predict_plot("World", 2020)
```



#R-Squared

```
# Fit the linear regression model for consumption
c_country <- db_c[["United States"]]
c_model <- lm(c_country ~ Year, data = db_c)

# Calculate R-squared for consumption model
r_squared <- summary(c_model)$r.squared

# Displaying the R-squared value
cat("R-squared value for the consumption model:", r_squared, "\n")
```

```
## R-squared value for the consumption model: 0.9216657
```

```
cat("With this R-squared value, approximately", round(r_squared * 100, 2), "% of the variability
in biofuel consumption of the country can be explained with the linear regression model.\n")
```

```
## With this R-squared value, approximately 92.17 % of the variability
## in biofuel consumption of the country can be explained with the linear regression model.
```

```
calculate_mean_r_squared_consumption <- function(db_c) {
  selected_countries <- c('World', 'Africa', 'Europe', 'North America', 'Asia & Oceania',
                      'Central & South America', 'sum_of_columns')
  r_squared_values_c <- data.frame(Country = character(), R_squared_Consumption = numeric(),
                      stringsAsFactors = FALSE)

  for (country in selected_countries) {
    c_country <- db_c[[country]]

    # Check for missing values (NA)
    if (any(is.na(c_country))) {
      cat("Warning: Missing values (NA) found in", country, "\n")
      next
    }

    c_model <- lm(c_country ~ Year, data = db_c)
    c_r_squared <- summary(c_model)$r.squared

    r_squared_values_c <- rbind(r_squared_values_c, list(Country = country,
                       R_squared_Consumption = c_r_squared))
  }

  mean_r_squared_c <- mean(r_squared_values_c$R_squared_Consumption, na.rm = TRUE)

  return(mean_r_squared_c)
}


mean_r_squared_consumption <- calculate_mean_r_squared_consumption(db_c)
print(mean_r_squared_consumption)
```

```
## [1] 0.7778141
```

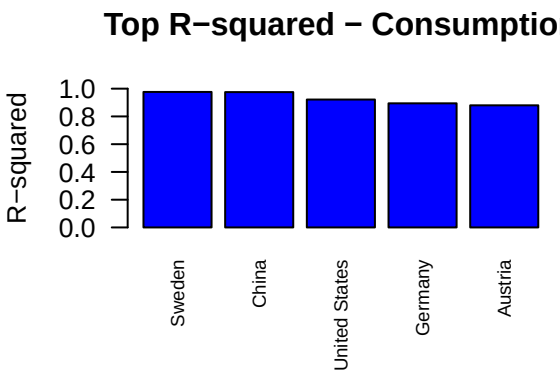
```
## An R squared value of approximately 0.7778141 suggests that around 77.78141 % of the variability
## observed in the consumption data across the countries can be explained by the linear relationship
## with time according to the model.
```

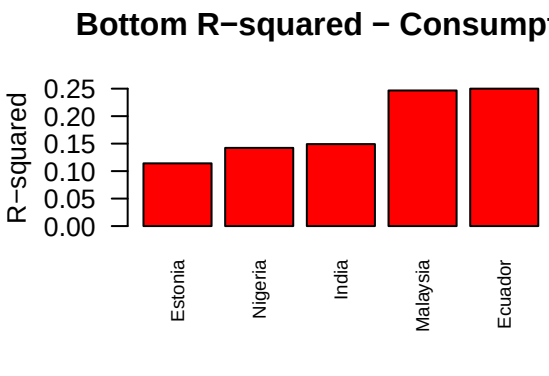
#Top R-Squared values

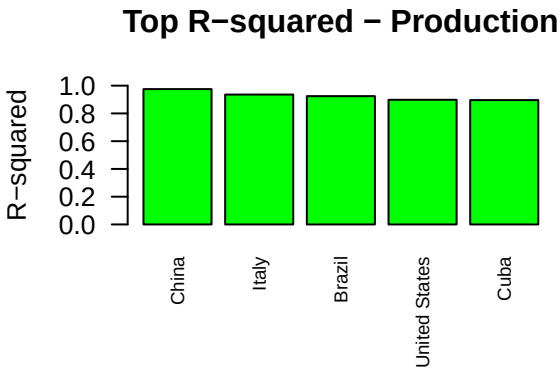
```
get_top_bottom_r_squared <- function(db_c, db_p) {
  get_top_r_squared <- function(db) {
    columns_to_exclude <- c("World", "Europe", "Africa", "Asia & Oceania", "North America",
                            "Central & South America", "sum_of_columns", "percentage")
    countries <- setdiff(names(db)[-ncol(db)], columns_to_exclude)
    r_squared_values <- data.frame(Country = character(), R_squared = numeric(),
                            stringsAsFactors = FALSE)

    for (country in countries) {
      c_country <- db[[country]]
```

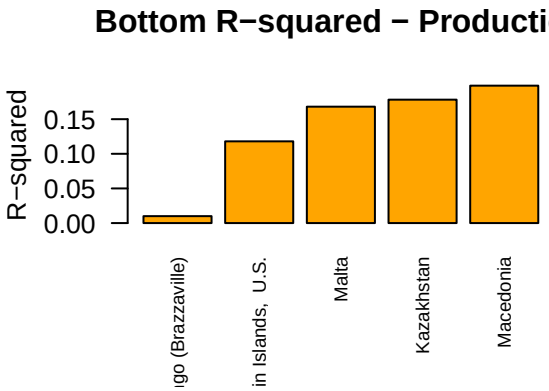
```
      suppressWarnings({
        if (any(is.na(c_country))) {
          next
        }
      })

      c_model <- lm(c_country ~ Year, data = db)
      c_r_squared <- summary(c_model)$r.squared

      r_squared_values <- rbind(r_squared_values, list(Country = country, R_squared = c_r_squared))
    }

    top_r_squared <- r_squared_values[order(r_squared_values$R_squared, decreasing = TRUE), ]
    top_r_squared <- head(top_r_squared, 5)

    bottom_r_squared <- r_squared_values[order(r_squared_values$R_squared), ]
    bottom_r_squared <- head(bottom_r_squared, 5)

    return(list(top = top_r_squared, bottom = bottom_r_squared))
  }


  db_c_top_bottom <- get_top_r_squared(db_c)
  db_p_top_bottom <- get_top_r_squared(db_p)


  par(mfrow = c(2, 2))

  # Plot for db_c top R-squared
  barplot(db_c_top_bottom$top$R_squared, names.arg = db_c_top_bottom$top$Country,
          main = "Top R-squared - Consumption", ylab = "R-squared", col = "blue", las = 2,
          ylim = c(0, 1), cex.names = 0.7)

  # Plot for db_c bottom R-squared
  barplot(db_c_top_bottom$bottom$R_squared, names.arg = db_c_top_bottom$bottom$Country,
          main = "Bottom R-squared - Consumption", ylab = "R-squared", col = "red", las = 2,
          cex.names = 0.7)

  # Plot for db_p top R-squared
  barplot(db_p_top_bottom$top$R_squared, names.arg = db_p_top_bottom$top$Country,
          main = "Top R-squared - Production", ylab = "R-squared", col = "green", las = 2,
          ylim = c(0, 1), cex.names = 0.7)

  # Plot for db_p bottom R-squared
  barplot(db_p_top_bottom$bottom$R_squared, names.arg = db_p_top_bottom$bottom$Country,
          main = "Bottom R-squared - Production", ylab = "R-squared", col = "orange", las = 2,
          cex.names = 0.7)
}

get_top_bottom_r_squared(db_c, db_p)
```

```
## Warning in summary.lm(c_model): essentially perfect fit: summary may be
## unreliable
```



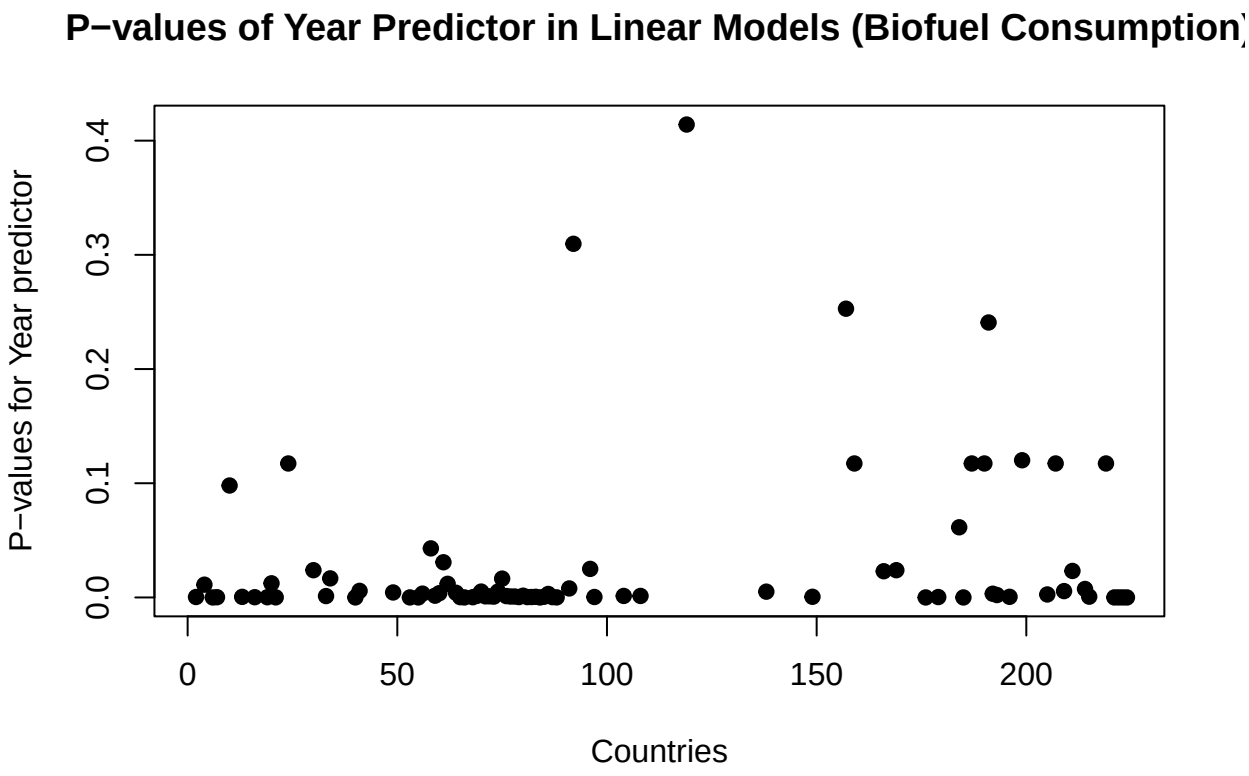







```r
create_lm_and_plot_p_values <- function(data) {
  p_values <- numeric()

  # Suppress warnings while performing calculations
  suppressWarnings({
    for (col in names(data)[-1]) {
      country_data <- data[[col]]

      # Check for missing values (NA)
      if (any(is.na(country_data))) {
        next
      }


      lm_country <- lm(country_data ~ Year, data = data)

      # Extract the p-value for 'Year' predictor in each linear model and store it
      p_value <- summary(lm_country)$coefficients["Year", "Pr(>|t|)"]
      p_values <- c(p_values, p_value)
    }
  })

  # Plotting the p-values
  plot(p_values, pch = 19, xlab = "Countries", ylab = "P-values for Year predictor",
       main = "P-values of Year Predictor in Linear Models (Biofuel Consumption)")
}
```

```
create_lm_and_plot_p_values(db_c)
```




```
## Just to ensure that the model is correct
```

From what year can this data be from?

```
year_pred <- function(country, input) {
  value <- data.frame(Production = input)

  # Production
  p_country_df <- data.frame(Year = db_p[["Year"]], Production = db_p[[country]])
  p_model <- lm(Year ~ Production, data = p_country_df)
  p_pred <- predict(p_model, newdata = value, interval = "confidence")

  # Consumption
  c_country_df <- data.frame(Year = db_c[["Year"]], Production = db_c[[country]])
  c_model <- lm(Year ~ Production, data = c_country_df)
  c_pred <- predict(c_model, newdata = value, interval = "confidence")

  cat("\nProduction", input, "for", country, "can be found in the year",
      round(p_pred[1], digits = 0))
  cat("\n95% Confidence Interval for Production prediction:", round(p_pred[2], digits = 0), "to",
```

```
      round(p_pred[3], digits = 0))

  cat("\nConsumption", input, "for", country, "can be found in the year", round(c_pred[1],
      digits = 0))
  cat("\n95% Confidence Interval for Consumption prediction:", round(c_pred[2], digits = 0), "to",
      round(c_pred[3], digits = 0))
}

year_pred("World", 2740)
```

```
##
## Production 2740 for World can be found in the year 2016
## 95% Confidence Interval for Production prediction: 2013 to 2018
## Consumption 2740 for World can be found in the year 2016
## 95% Confidence Interval for Consumption prediction: 2013 to 2019
```

```
db_p$`North & Central & South America`<-rowSums(db_p[, c("North America","Central & South America")],
                      na.rm = TRUE)
db_c$`North & Central & South America`<-rowSums(db_c[, c("North America","Central & South America")],
                      na.rm = TRUE)

year_pred("North & Central & South America", 2055)
```

```
##
## Production 2055 for North & Central & South America can be found in the year 2015
## 95% Confidence Interval for Production prediction: 2012 to 2017
## Consumption 2055 for North & Central & South America can be found in the year 2016
## 95% Confidence Interval for Consumption prediction: 2013 to 2019
```

```
year_pred("Europe", 385)
```

```
##
## Production 385 for Europe can be found in the year 2015
## 95% Confidence Interval for Production prediction: 2013 to 2017
## Consumption 385 for Europe can be found in the year 2012
## 95% Confidence Interval for Consumption prediction: 2010 to 2014
```

```
year_to_year_pred <- function(country, input) {
  # Summing production values for North America and Central & South America
  p_sum <- db_p[["North America"]] + db_p[["Central & South America"]]
  c_sum <- db_c[["North America"]] + db_c[["Central & South America"]]

  value <- data.frame(Production = input)

  # Production prediction
  p_country_df <- data.frame(Year = db_p[["Year"]], Production = p_sum)
  p_model <- lm(Year ~ Production, data = p_country_df)
  p_pred <- predict(p_model, newdata = value)

  # Consumption prediction
  c_country_df <- data.frame(Year = db_c[["Year"]], Production = c_sum)
```

```r
  c_model <- lm(Year ~ Production, data = c_country_df)
  c_pred <- predict(c_model, newdata = value)

  cat("\nProduction", input, "for", country, "can be found in the year",
      round(p_pred, digits = 0))
  cat("\nConsumption", input, "for", country, "can be found in the year",
      round(c_pred, digits = 0))
}

year_to_year_pred("North America & Central & South America", 2005)
```

```
##
## Production 2005 for North America & Central & South America can be found in the year 2014
## Consumption 2005 for North America & Central & South America can be found in the year 2016
```