Analysis of biofuel production and consumption

Jan Izquierdo & Sergi Ocaña

2023-11-26

Importing the data

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

Fixing the data

```r
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

```r
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

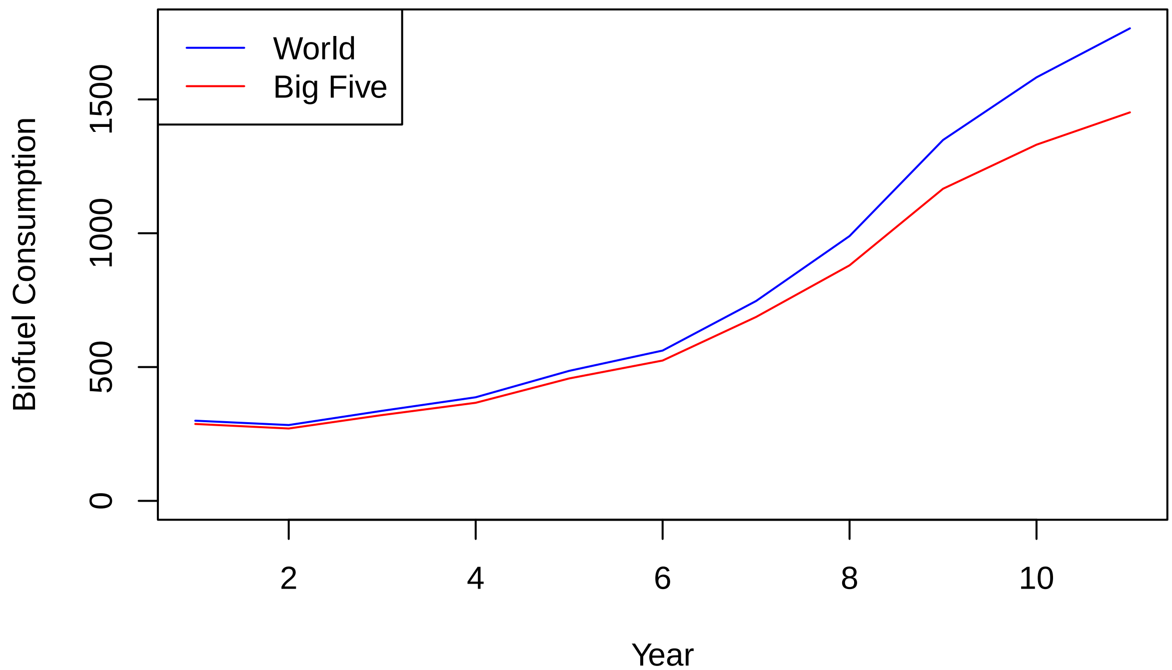
Selection of countries of who we will perform analysis later

```
#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
```

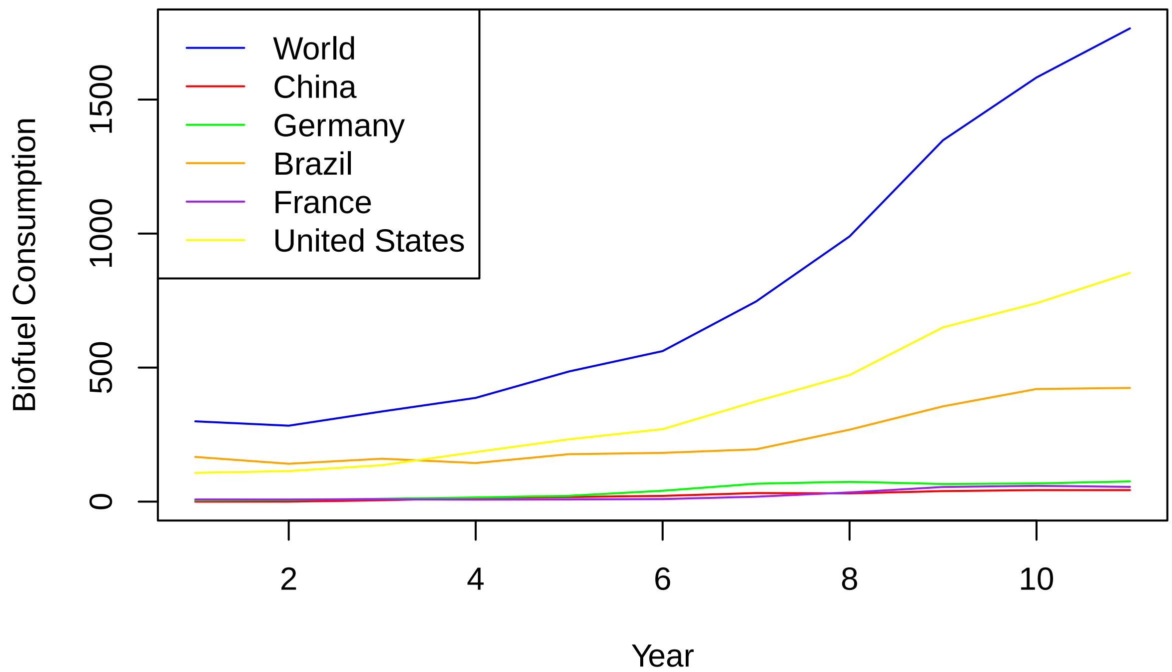
Plot the selected countries' consumption against the World

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)
```



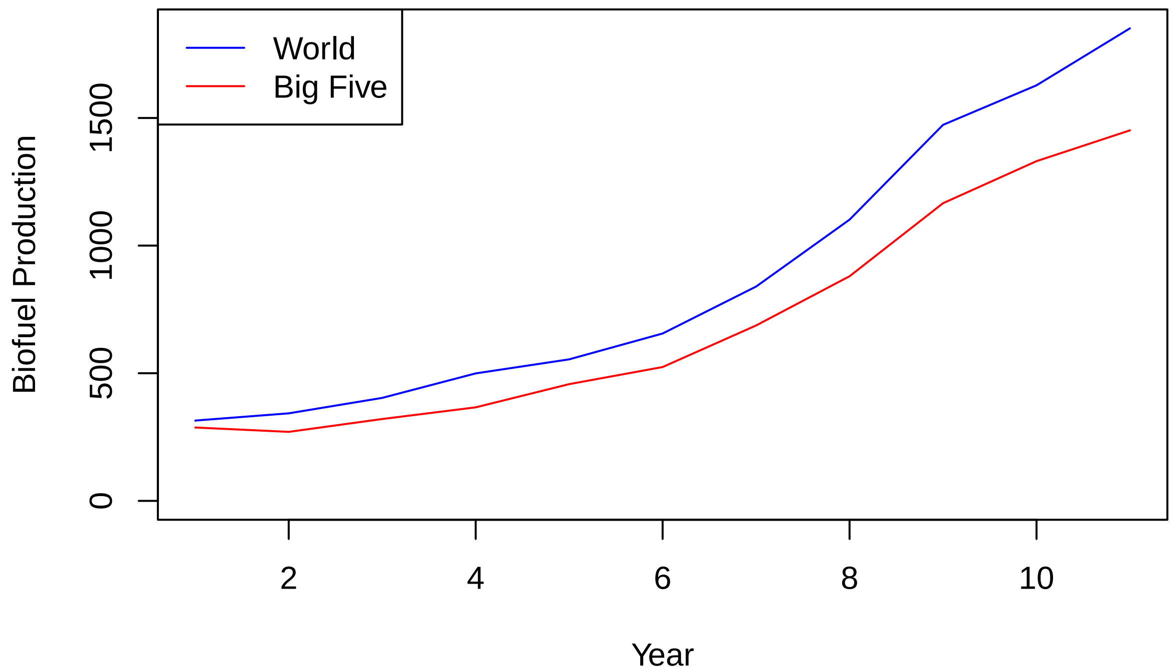

```r
# Plotting a line plot for selected columns
plot(db_c$World, type = "l", col = "blue", xlab = "Year", ylab = "Biofuel Consumption",
     ylim = c(0, max(db_c$World, na.rm = TRUE)))

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)
```



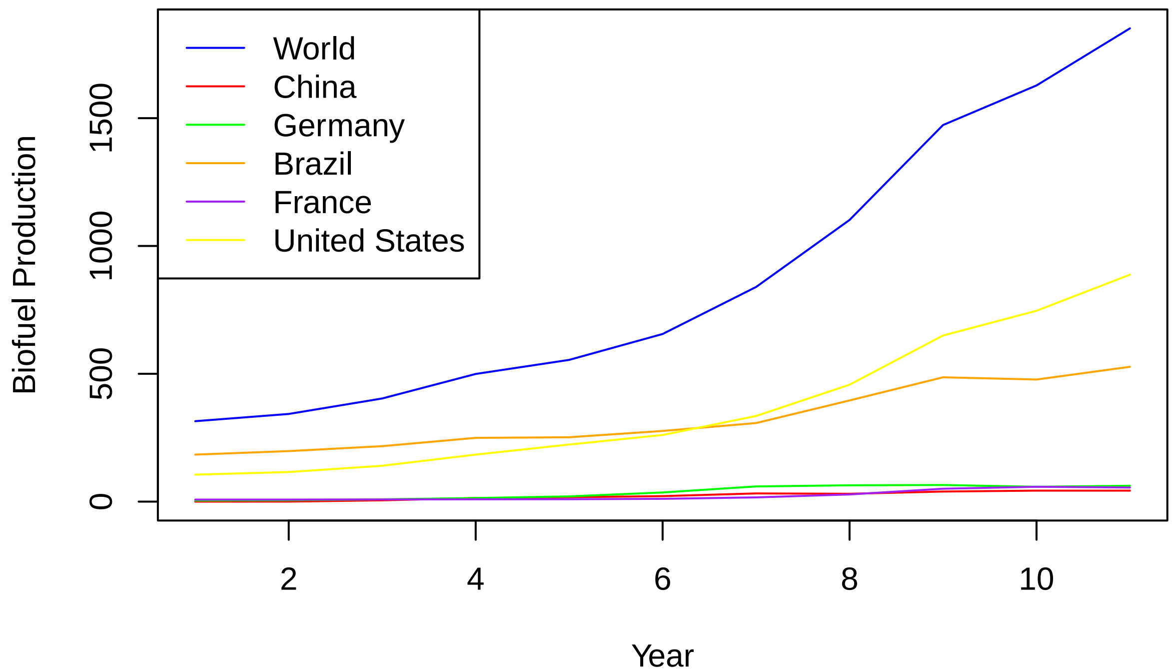

```
## The dominance of just a few countries in contributing to a significant portion of
## the world's biofuel consumption indicates a highly skewed distribution. This skewed distribution
## suggests a high level of inequality or concentration, where a small number of countries,
## holding a substantial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

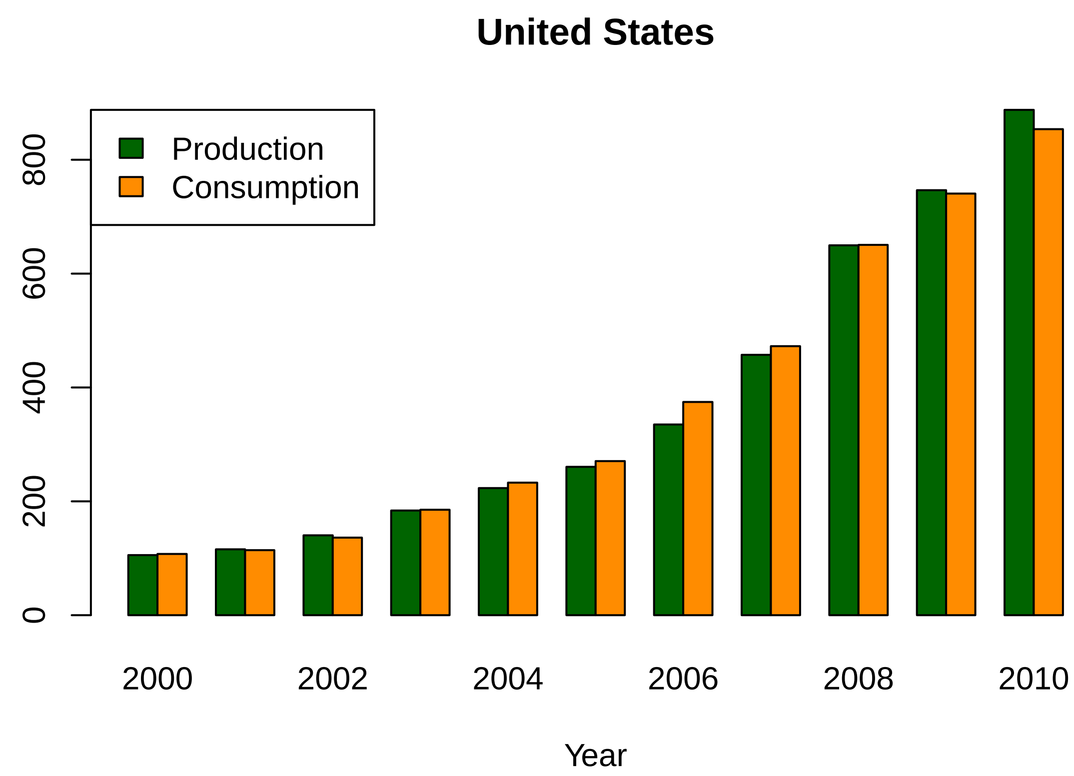



```r
# Plotting a line plot for selected columns
plot(db_p$World, type = "l", col = "blue", xlab = "Year", ylab = "Biofuel Production",
     ylim = c(0, max(db_p$World, na.rm = TRUE)))

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)
```

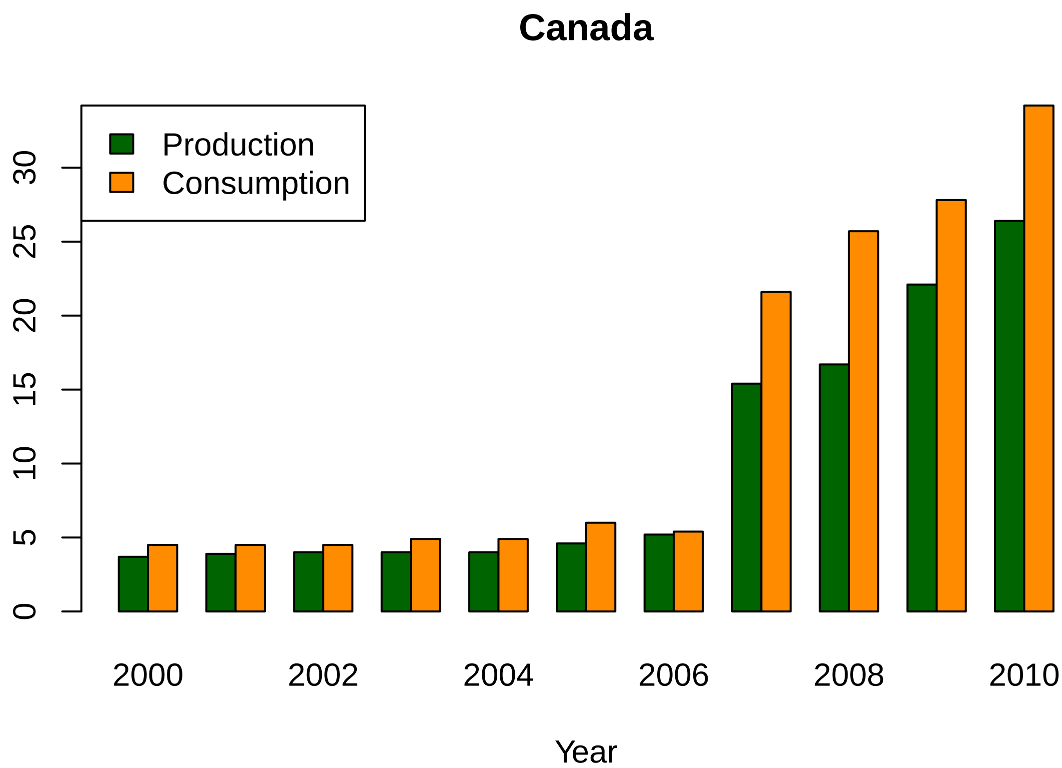


Data representation of United States of America and Canada

```
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")
```

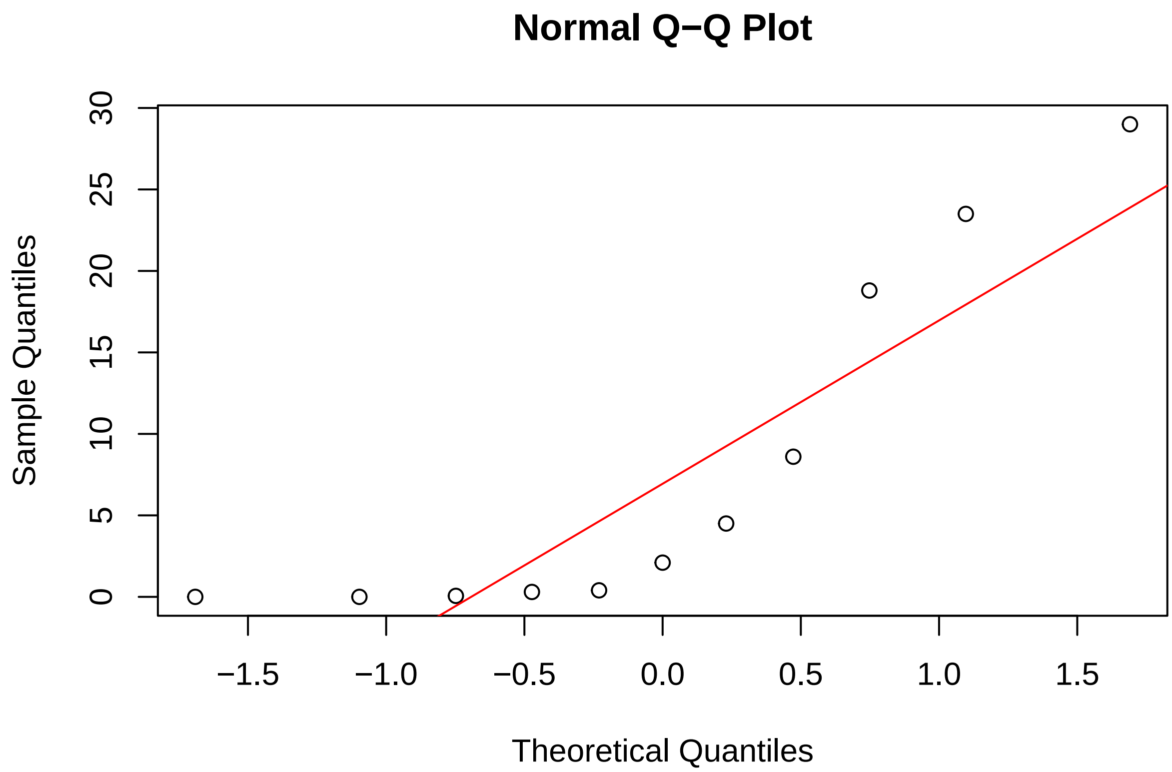



```
combined_freq<-t(cbind(db_p$Canada, db_c$Canada))
barplot(combined_freq, beside=T, col=c("darkgreen","darkorange"),
        legend.text = c("Production", "Consumption"),args.legend = list(x = "topleft"),
        xlab="Year", names.arg=rownames(db_p), main="Canada")
```



Does it follow normal distribution??

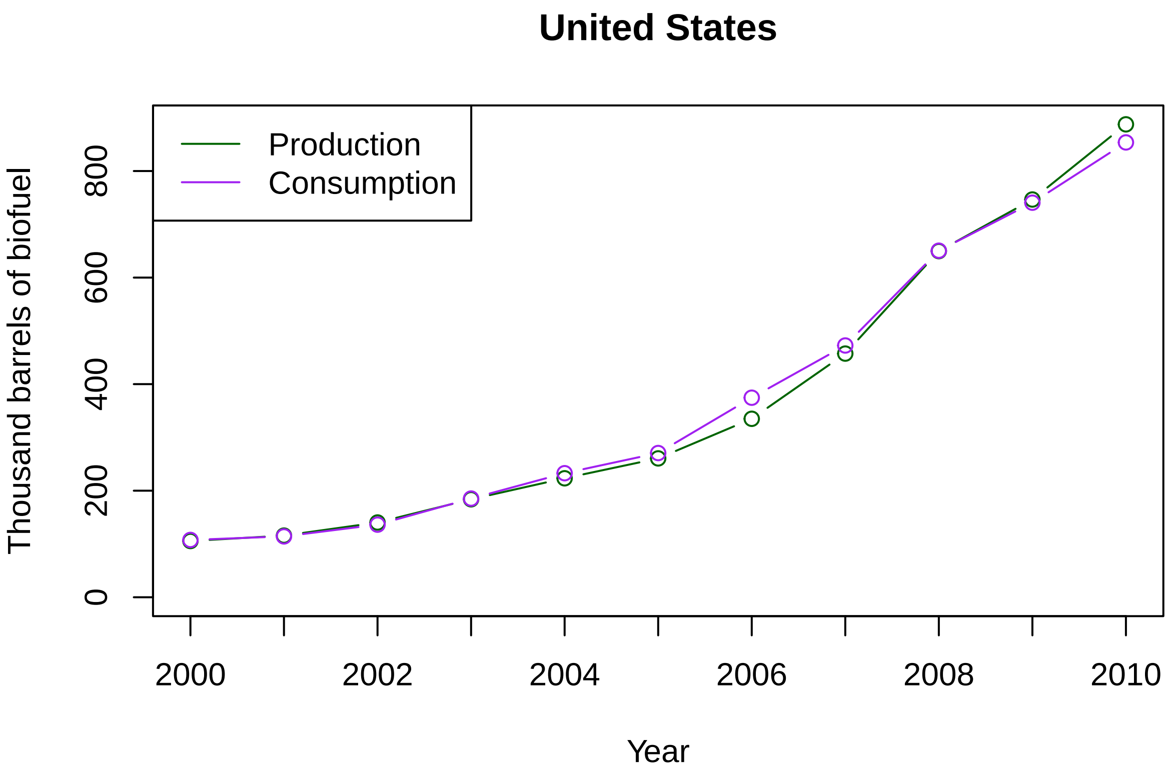
```
# Extract 'United States' consumption data
us_consumption <- as.numeric(db_c[, 'United Kingdom'])

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



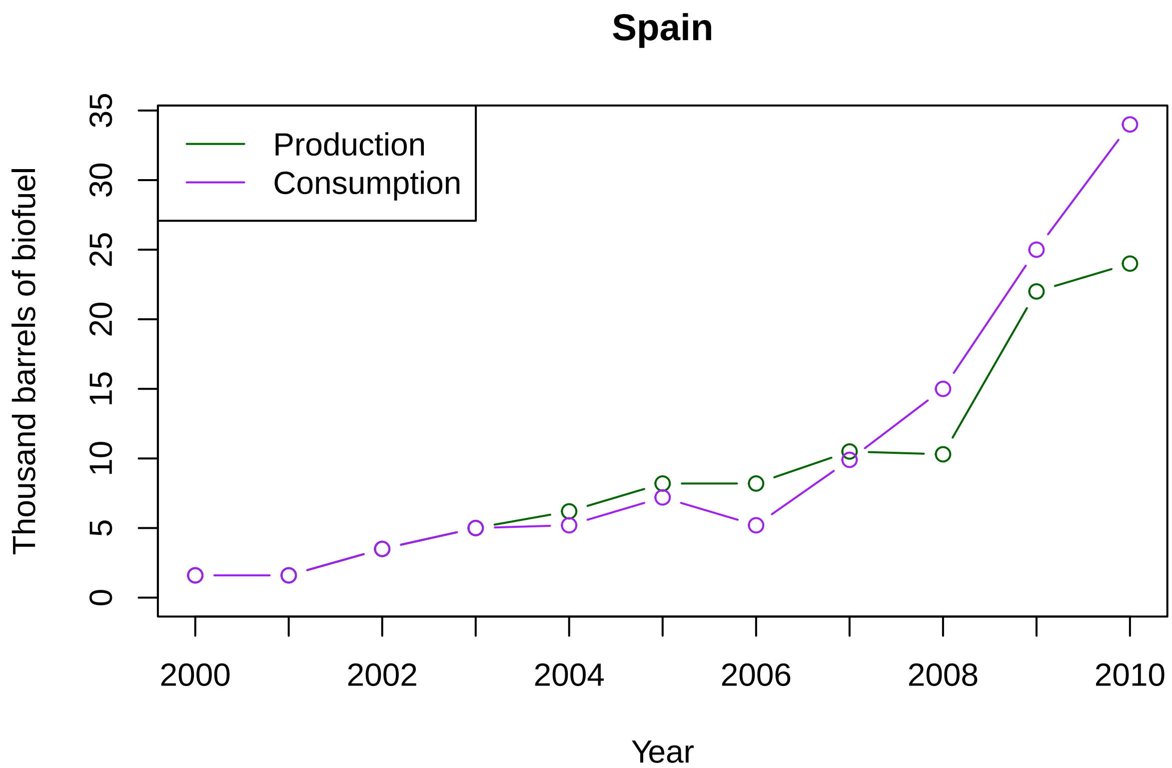

```
## We can observe that our data follows a normal distribution
```

Plotting the data from a country

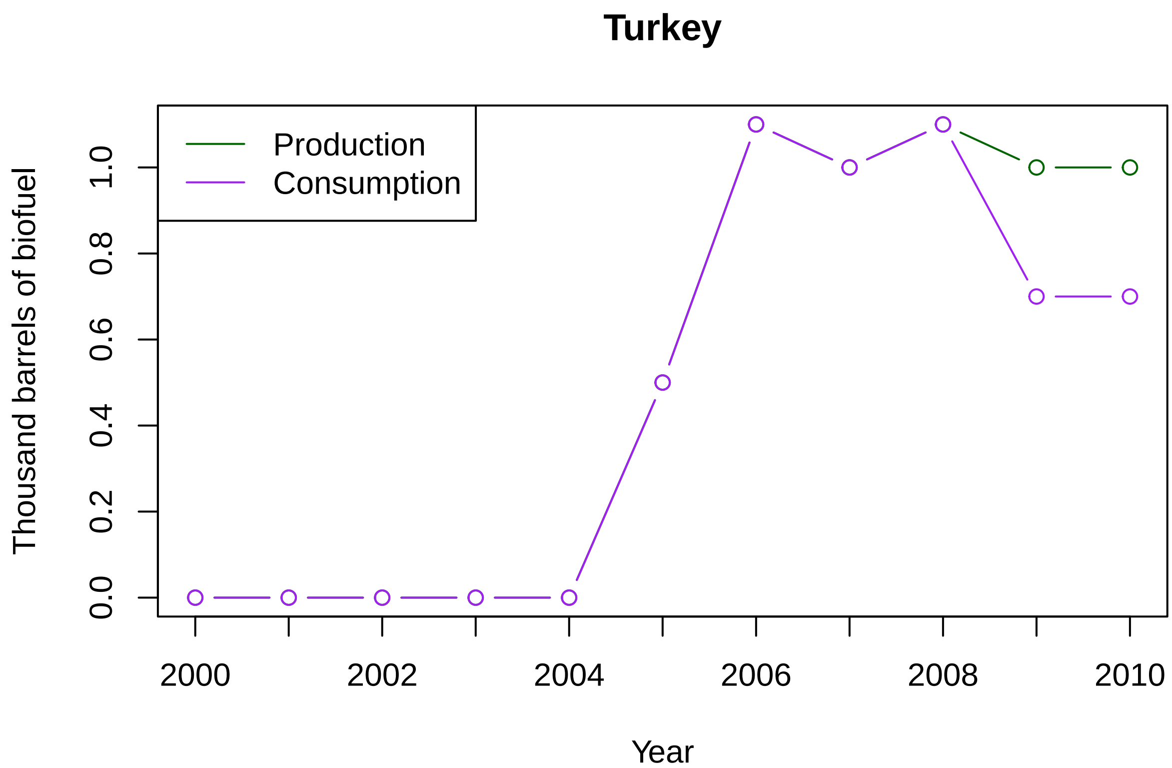
```r
plot_country <- function(country_name) {
  max_value <- max(max(db_p[[country_name]], na.rm = TRUE), max(db_c[[country_name]],
         na.rm = TRUE), na.rm = TRUE)
  plot(db_p[[country_name]], col="darkgreen", type="b", ylab="Thousand barrels of biofuel",
         xlab="Year", xaxt="n", ylim = 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")
```

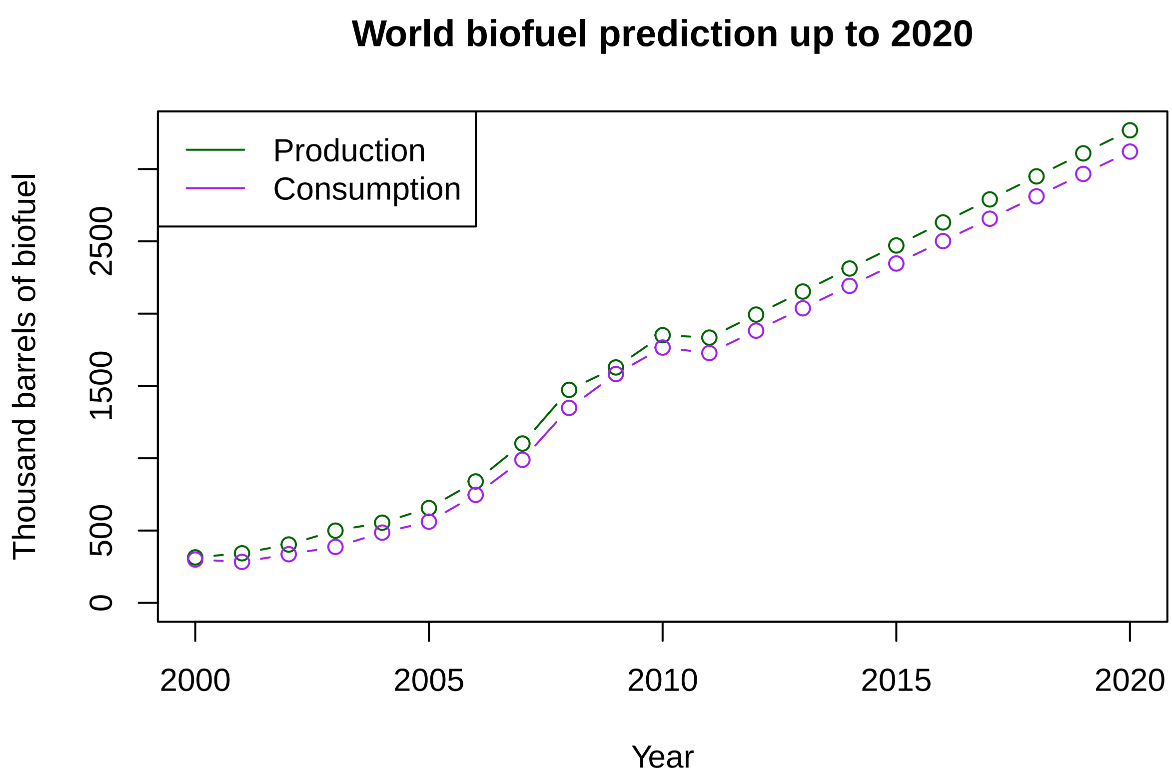
```r
  cat("R-squared for Consumption Model:", c_r_squared, "\n")

}

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

```r
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])
```

```r
    #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

```r
# 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
```

```r
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
```

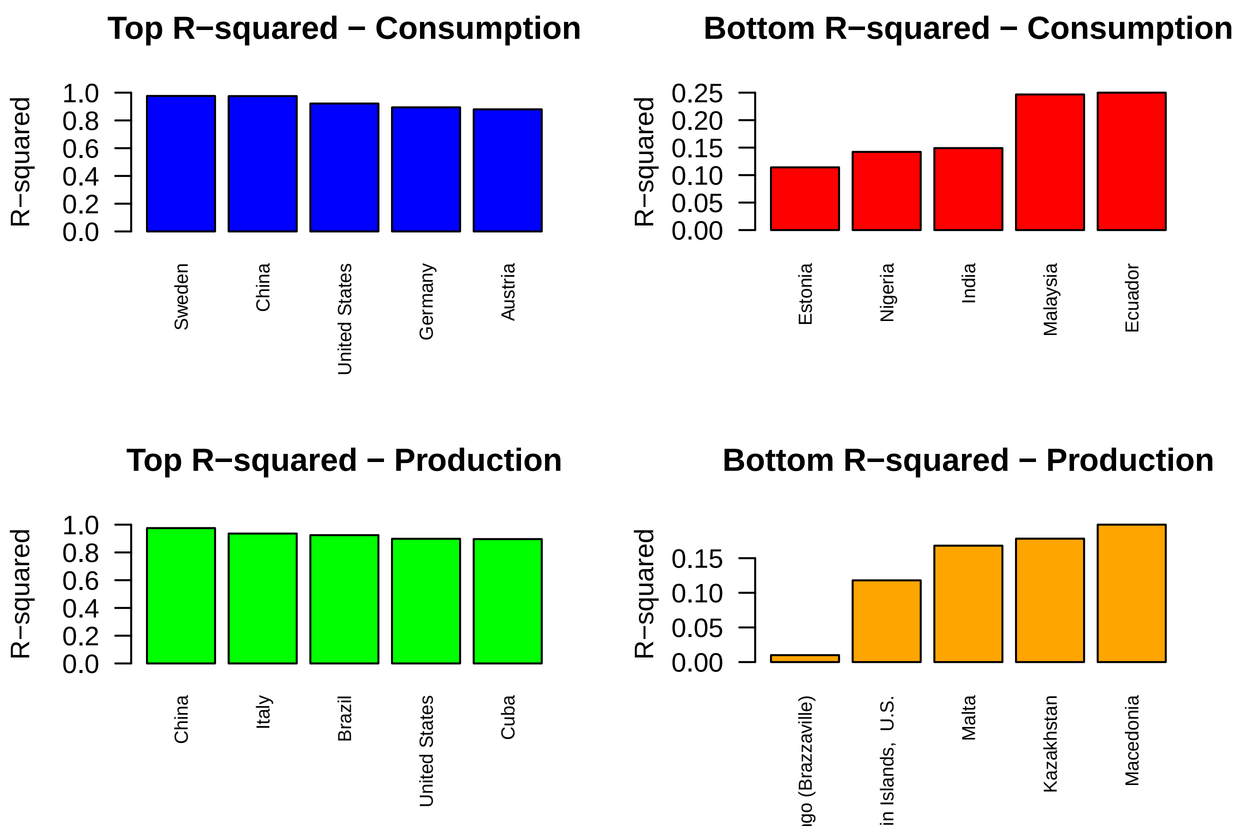
```
## An R squared value of approximately 0.7778141 suggests that around 77.78141 % of the variability
## observed in the consumption data across the countries can be explained by the linear relationship
## with time according to the model.
```

#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]]
```

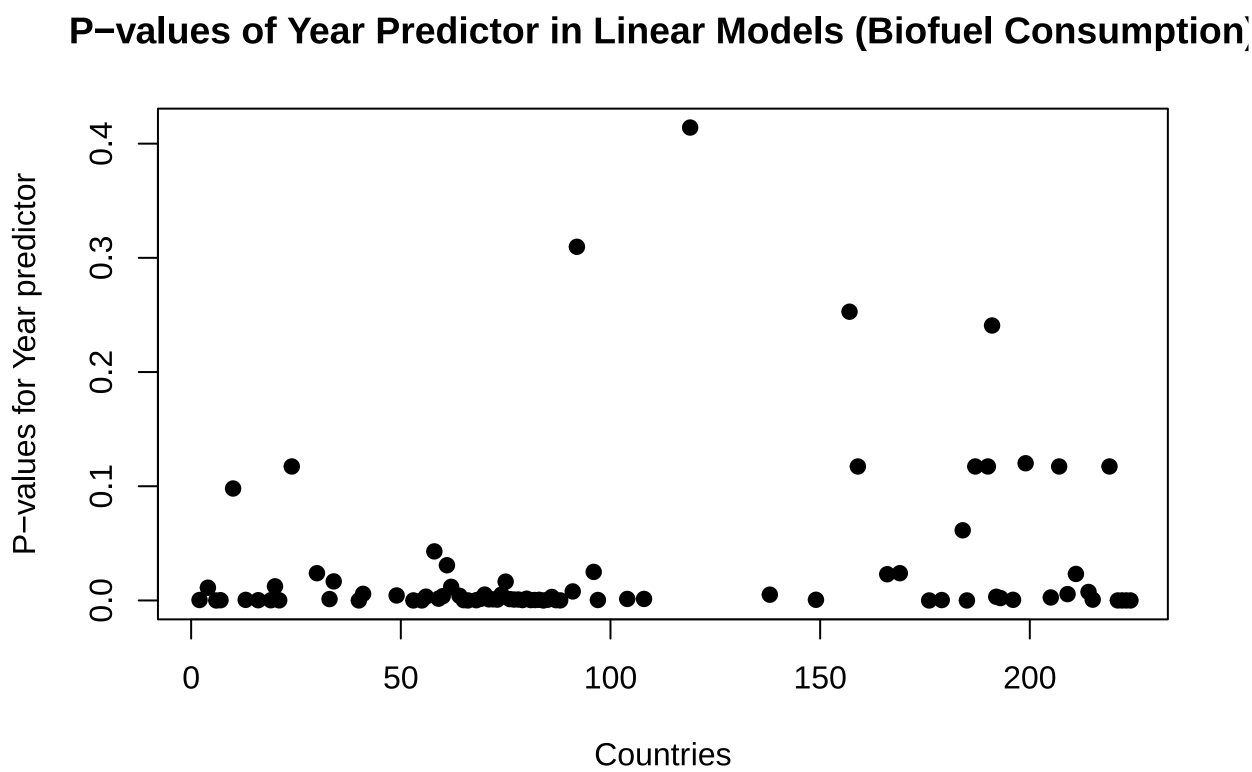
```r
      suppressWarnings({
        if (any(is.na(c_country))) {
          next
        }
      })

      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
```




```
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
```