Data Sceience Final Project: Exploring Relationships Between Renewable Energy Consumption And Various Economical And Environmental Indicators Worldwide

Batool Alaidaroos

2023-12-20

Libraries

```
library(tidyverse)        # For data manipulation and visualization
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)            # For reading CSV files
library(caret)            # For data preprocessing
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(RANN)             # For k-nearest neighbors imputation
```

```
## Warning: package 'RANN' was built under R version 4.3.2
```

```r
library(reshape2)         # For data reshaping
```

```
## Warning: package 'reshape2' was built under R version 4.3.2

##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(randomForest)      # For creating a correlation matrix
```

```
## Warning: package 'randomForest' was built under R version 4.3.2

## randomForest 4.7-1.1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(gridExtra)         # For arranging plots in a grid
```

```
## Warning: package 'gridExtra' was built under R version 4.3.2

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:randomForest':
##
##     combine
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(sf)                # For spatial data manipulation and plotting
```

```
## Warning: package 'sf' was built under R version 4.3.2

## Linking to GEOS 3.11.2, GDAL 3.7.2, PROJ 9.3.0; sf_use_s2() is TRUE
```

Loading data

```
#csv files:
renewable_energy_consumption <- read_csv("data/renewable_energy_consumption.csv")
```

```
## Rows: 266 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country_Name, Country Code, Indicator Name, Indicator Code
## dbl (32): 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, ...
## lgl (31): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
co2emissions <- read_csv("data/co2emissions.csv")
```

```
## Rows: 266 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country_Name, Country Code, Indicator Name, Indicator Code
## dbl (31): 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, ...
## lgl (32): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
access_to_electricity <- read_csv("data/access_to_electricity.csv")
```

```
## Rows: 266 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country_Name, Country Code, Indicator Name, Indicator Code
## dbl (32): 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, ...
## lgl (31): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
economic_growth <- read_csv("data/economicGrowth.csv")
```

```
## Rows: 266 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country_Name, Country Code, Indicator Name, Indicator Code
## dbl (61): 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, ...
## lgl  (2): 1960, 2022
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
energy_depletion <- read_csv("data/energy_depletion.csv")
```

```
## Rows: 266 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): Country_Name, Country Code, Indicator Name, Indicator Code
## dbl (52): 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, ...
## lgl (11): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 2022
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
regions <- read_csv("data/regions.csv") # + income level
```

```
## Rows: 265 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Country_Name, Country Code, Region, Income, SpecialNotes
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

First: Data Preprocessing:

```r
# Slicing the data from 2000 to 2021
economic_growth <- select(economic_growth, Country_Name, '2000':'2021' )
renewable_energy_consumption <- select(renewable_energy_consumption, Country_Name, '2000':'2021')
co2emissions <- select(co2emissions, Country_Name, '2000':'2021')
access_to_electricity <- select(access_to_electricity, Country_Name, '2000':'2021')
energy_depletion <- select(energy_depletion, Country_Name, '2000':'2021')
regions <- select(regions, Country_Name, 'Region':'Income')
```

```r
# Pivot the data to long format for better analysis
economic_growth1 <- pivot_longer(economic_growth, cols="2000":"2021",
                                 names_to = "year",
                                 values_to = "eco_growth")

renewable_energy_consumption1 <- pivot_longer(renewable_energy_consumption, cols="2000":"2021",
                                 names_to = "year",
                                 values_to = "renewable_energy_consumption")

co2emissions1 <- pivot_longer(co2emissions, cols="2000":"2021",
                                 names_to = "year",
                                 values_to = "co2emissions")

access_to_electricity1 <- pivot_longer(access_to_electricity, cols="2000":"2021",
                                 names_to = "year",
                                 values_to = "access_to_electricity")
```

```r
energy_depletion1 <- pivot_longer(energy_depletion, cols="2000":"2021",
                                  names_to = "year",
                                  values_to = "energy_depletion")
```

Data Merging:

```r
# Merge datasets on Country_Name and year
merged_data <- merge(economic_growth1, renewable_energy_consumption1, by = c("Country_Name", "year"), al
merged_data <- merge(merged_data, co2emissions1, by = c("Country_Name", "year"), all = TRUE)
merged_data <- merge(merged_data, access_to_electricity1, by = c("Country_Name", "year"), all = TRUE)
merged_data <- merge(merged_data, energy_depletion1, by = c("Country_Name", "year"), all = TRUE)
merged_data <- merge(merged_data, regions, by = c("Country_Name"), all = TRUE)
```

Data Cleaning:

```r
# Remove rows with missing values
merged_data <- merged_data %>% filter(complete.cases(.))
```

Data Imputation:

```r
# Create k-nearest neighbors imputation model
preProcess_missingdata_model <- preProcess(merged_data, method = 'knnImpute')

# Check if any predictors have all missing values
all_missing_columns <- colnames(merged_data)[apply(is.na(merged_data), 2, all)]

# Remove columns with all missing values from the new data point
newdata <- merged_data[, !colnames(merged_data) %in% all_missing_columns]

# Use the imputation model to predict the values of missing data points
if (length(all_missing_columns) > 0) {
  filledData <- predict(preProcess_missingdata_model, newdata = newdata)
  filledData_all_missing <- merged_data[, colnames(merged_data) %in% all_missing_columns]
  filledData <- cbind(filledData, filledData_all_missing)
} else {
  filledData <- newdata
}

# Check if there are any remaining missing values in the imputed data
anyNA(filledData)
```

```
## [1] FALSE
```

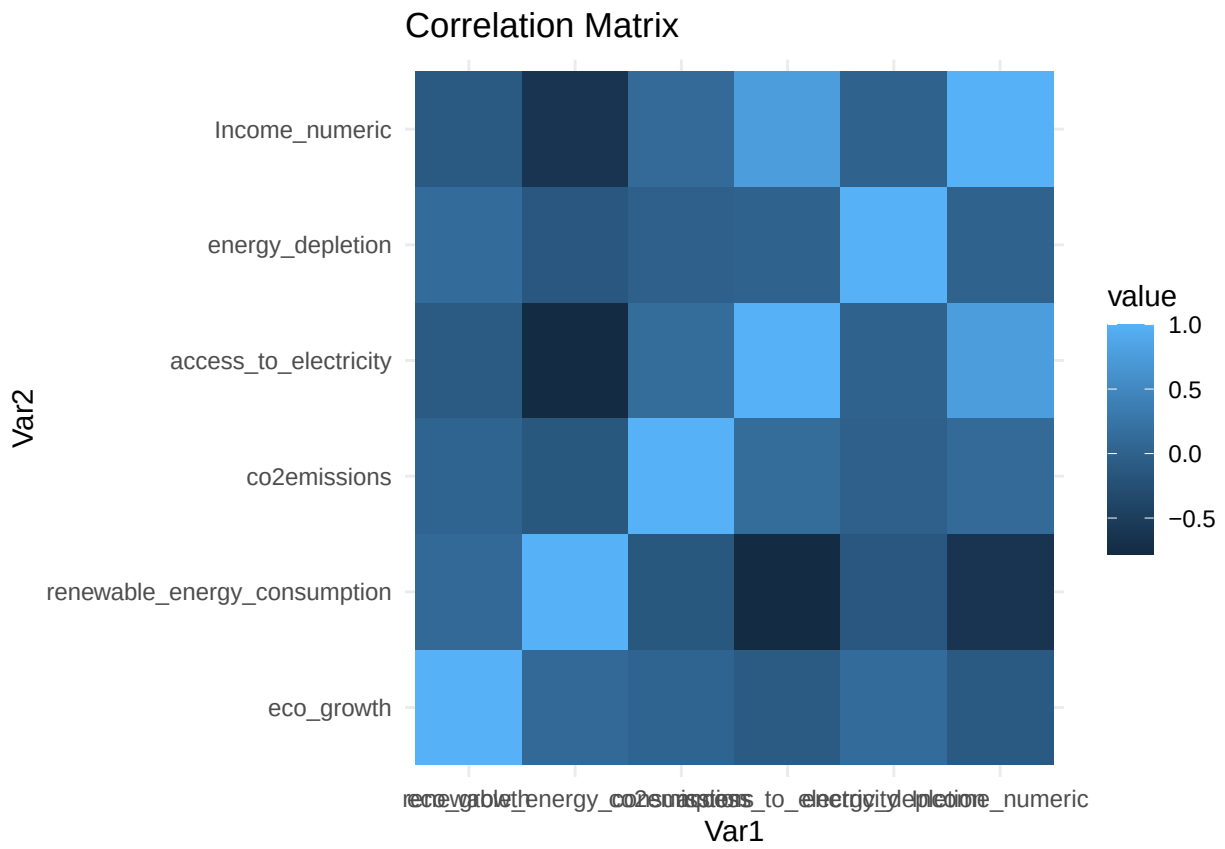
Second: Data Analysis:

Question 1 : What is the relationship between economic growth, renewable energy consumption, CO2 emissions, access to electricity, energy depletion, and income level across countries?

```
# Convert income categories to numeric values
income_numeric <- factor(filledData$Income, levels = c("Low income", "Lower middle income", "Upper middl
filledData$Income_numeric <- as.numeric(income_numeric)

# Correlation matrix including "Income"
correlation_matrix <- cor(filledData[, c("eco_growth", "renewable_energy_consumption", "co2emissions",
print(correlation_matrix)
```

```
##                              eco_growth renewable_energy_consumption
## eco_growth                   1.00000000                   0.1015244
## renewable_energy_consumption  0.10152442                   1.0000000
## co2emissions                  0.03377291                  -0.1392553
## access_to_electricity        -0.10140638                  -0.7801428
## energy_depletion              0.12612740                  -0.1407015
## Income_numeric               -0.11294342                  -0.6394040
##                              co2emissions access_to_electricity
## eco_growth                     0.03377291           -0.10140638
## renewable_energy_consumption  -0.13925534           -0.78014276
## co2emissions                   1.00000000            0.13391019
## access_to_electricity          0.13391019            1.00000000
## energy_depletion              -0.02062747            0.01650088
## Income_numeric                 0.11094259            0.75255233
##                              energy_depletion Income_numeric
## eco_growth                         0.12612740    -0.11294342
## renewable_energy_consumption      -0.14070155    -0.63940401
## co2emissions                      -0.02062747     0.11094259
## access_to_electricity              0.01650088     0.75255233
## energy_depletion                   1.00000000     0.01581723
## Income_numeric                     0.01581723     1.00000000
```

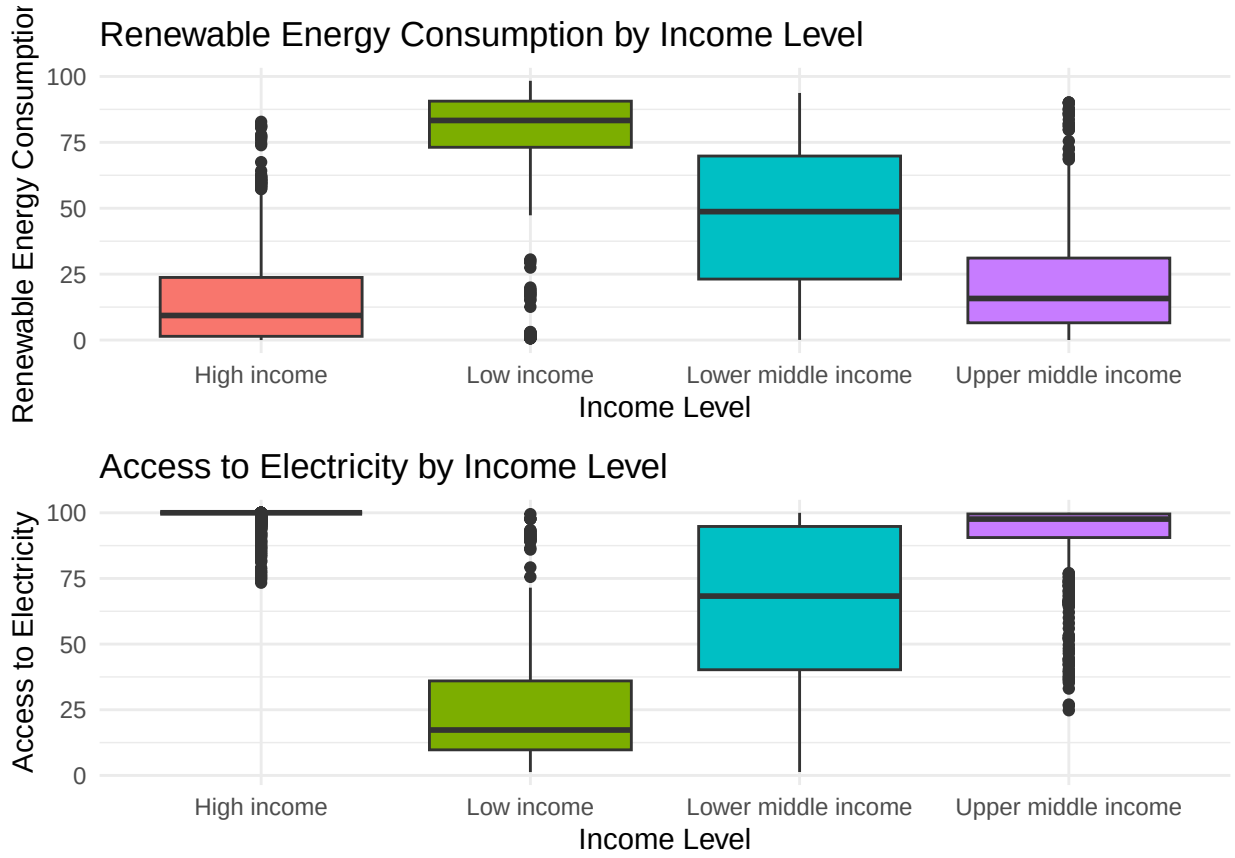
```
# Visualize correlation matrix
correlation_plot <- ggplot(data = melt(correlation_matrix), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  theme_minimal() +
  labs(title = "Correlation Matrix")

correlation_plot
```



Question 2: How does the distribution of renewable energy consumption vary among different income levels, and what insights can be gained by comparing it to the distribution of access to electricity across income levels?

```r
# Box plot for renewable energy consumption by income
plot1 <- ggplot(data = filledData, aes(x = Income, y = renewable_energy_consumption, fill = Income)) +
  geom_boxplot() +
  labs(title = "Renewable Energy Consumption by Income Level",
       x = "Income Level",
       y = "Renewable Energy Consumption",
       fill = "Income Level") +
  theme_minimal() +
  theme(legend.position = "none")

# Box plot for access to electricity by income
plot2 <- ggplot(merged_data, aes(x = Income, y = access_to_electricity, fill = Income)) +
  geom_boxplot() +
  labs(title = "Access to Electricity by Income Level",
       x = "Income Level",
       y = "Access to Electricity",
       fill = "Income Level") +
  theme_minimal() +
  theme(legend.position = "none")
```

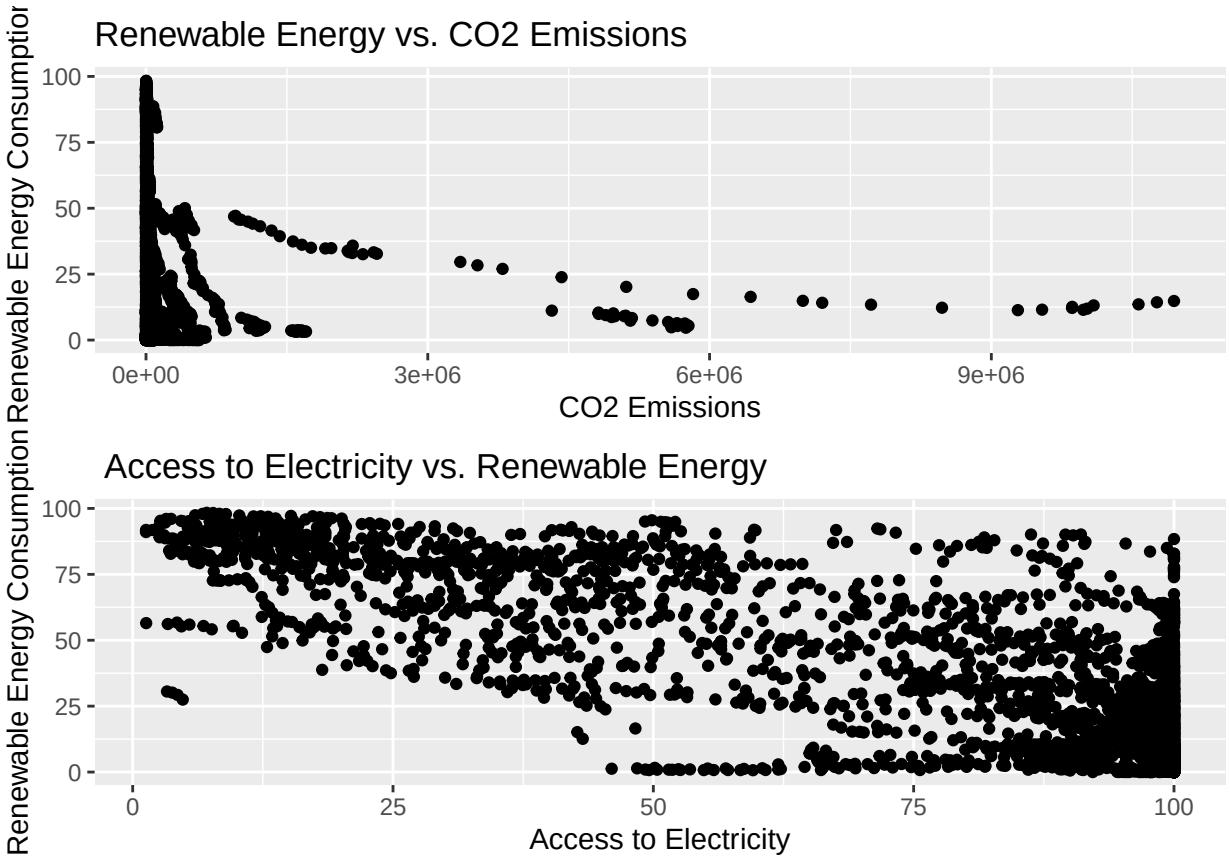
```
# Arrange the plots in a 1x2 grid layout
grid.arrange(plot1, plot2, nrow = 2, ncol = 1)
```



Question 3 : How does renewable energy consumption correlate with various environmental and economic factors?

```
# Scatter plot comparing renewable energy consumption with CO2 emissions
plot1 <- ggplot(merged_data, aes(x = co2emissions, y = renewable_energy_consumption)) +
  geom_point() +
  labs(title = "Renewable Energy vs. CO2 Emissions",
       x = "CO2 Emissions",
       y = "Renewable Energy Consumption")


# Scatter plot comparing renewable energy consumption with access to electricity
plot2 <- ggplot(merged_data, aes(x = access_to_electricity, y = renewable_energy_consumption)) +
  geom_point() +
  labs(title = " Access to Electricity vs. Renewable Energy",
       x = "Access to Electricity",
       y = "Renewable Energy Consumption")
```

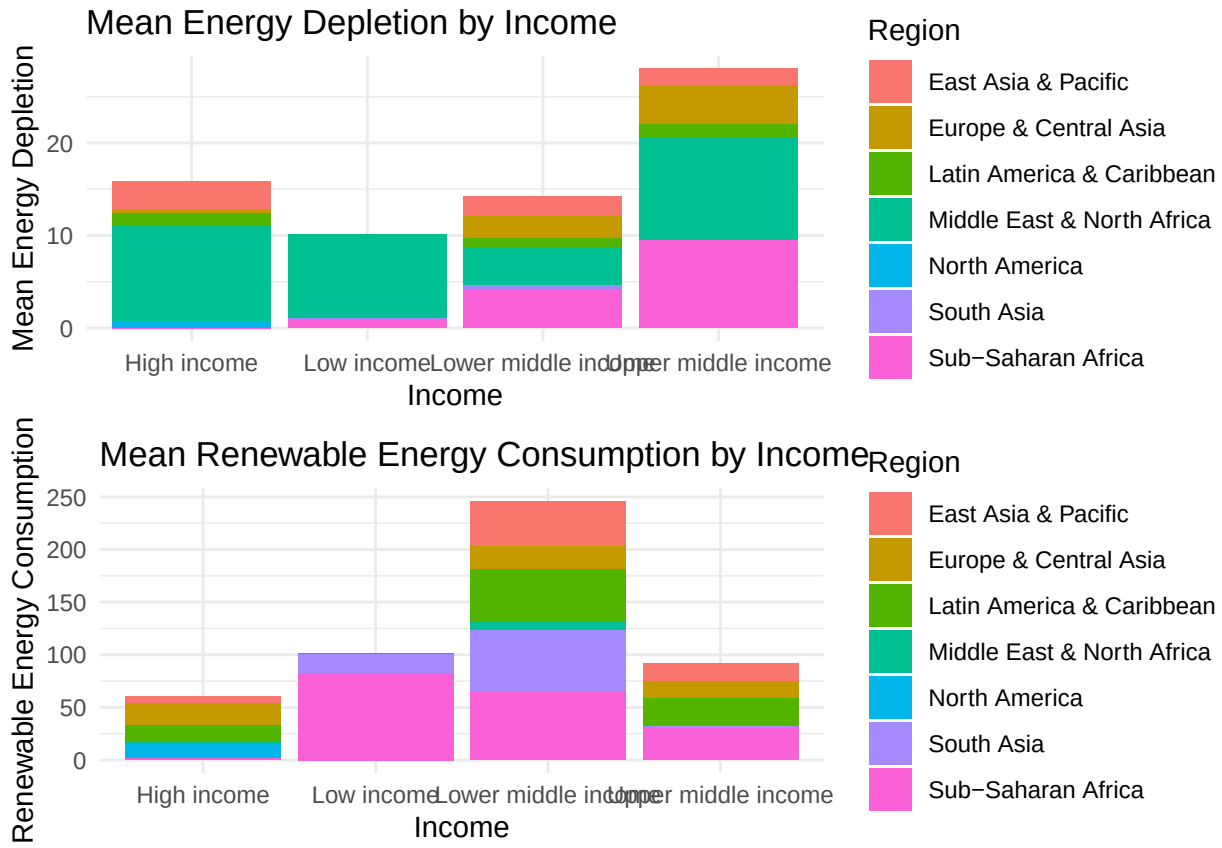
```r
# Arrange the plots in a 2x2 grid layout
grid.arrange(plot1, plot2, nrow = 2, ncol = 1)
```



Question 4: How does the mean energy depletion and mean renewable energy consumption vary across different income levels?

```r
# Bar plot of mean energy depletion by income
plot1 <- ggplot(merged_data, aes(x = Income, y = energy_depletion, fill = Region)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(title = "Mean Energy Depletion by Income",
       x = "Income",
       y = "Mean Energy Depletion") +
  theme_minimal()

# Bar plot of mean renewable energy consumption by income
plot2 <- ggplot(merged_data, aes(x = Income, y = renewable_energy_consumption, fill = Region)) +
  geom_bar(stat = "summary", fun = "mean") +
  labs(title = "Mean Renewable Energy Consumption by Income",
       x = "Income",
       y = "Renewable Energy Consumption") +
  theme_minimal()
```

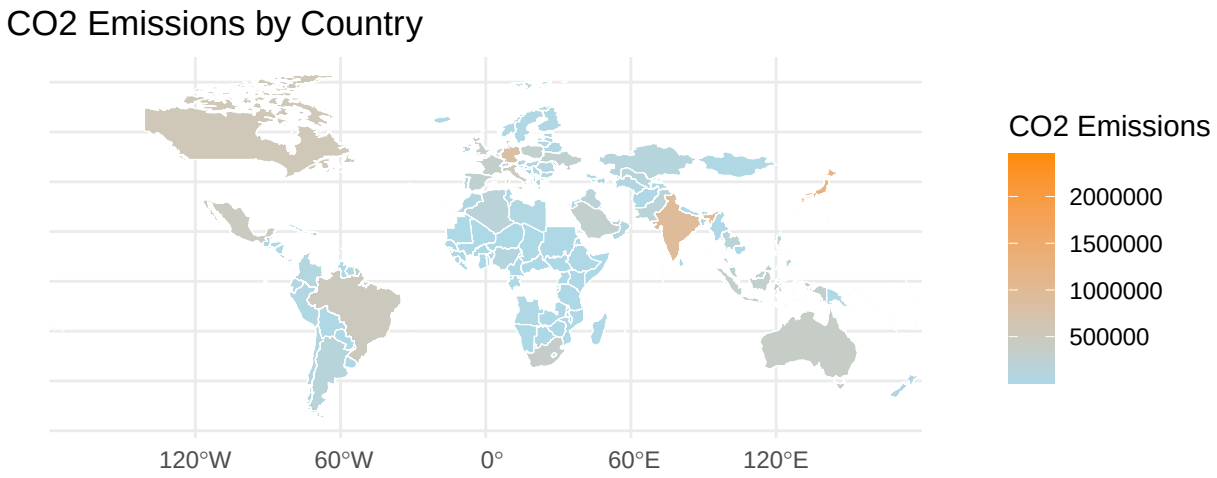
```r
# Arrange the plots in a 1x2 grid layout
grid.arrange(plot1, plot2, nrow = 2, ncol = 1)
```



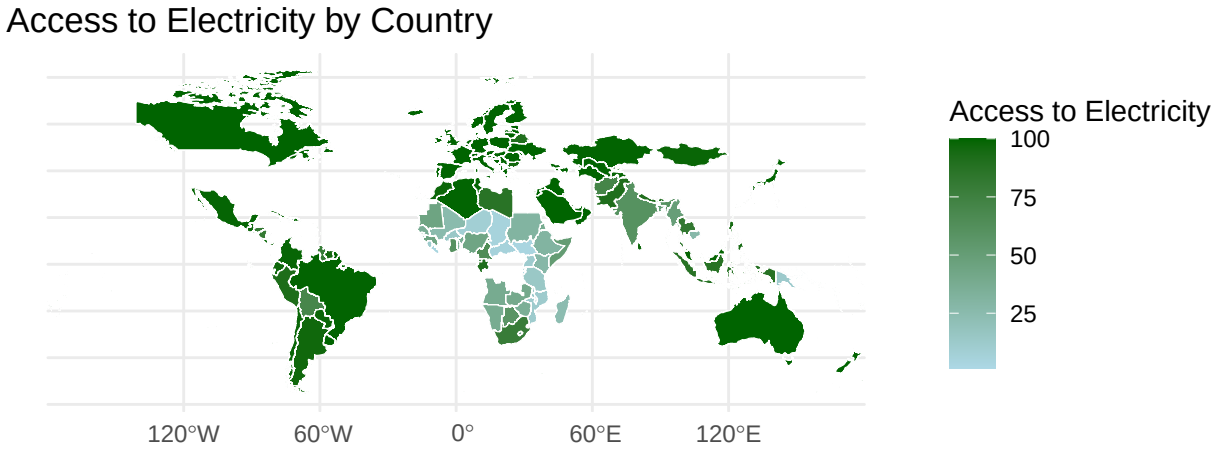
Question 5 : How do environmental and economic indicators, such as CO2 emissions, access to electricity, renewable energy consumption, economic growth, and energy depletion, vary spatially across different countries?

```r
#reading the shape file
world <- st_read("data/WB_countries.shp", quiet = TRUE)

# Merge spatial data with your dataset
merged_sf <- merge(world, merged_data, by.x = "NAME_EN", by.y = "Country_Name")
```

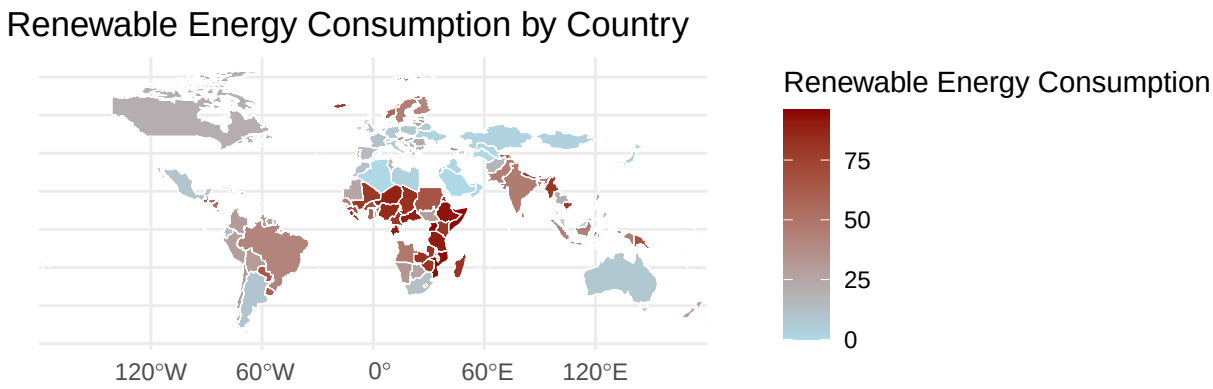
```r
# Plot CO2 emissions on a world map
ggplot() +
  geom_sf(data = merged_sf, aes(fill = co2emissions), color = "white", size = 0.1) +
  scale_fill_gradient(low = "lightblue", high = "darkorange") +
  labs(title = "CO2 Emissions by Country",
       fill = "CO2 Emissions") +
  theme_minimal()
```




```r
# Plot Access to Electricity on a world map
ggplot() +
  geom_sf(data = merged_sf, aes(fill = access_to_electricity), color = "white", size = 0.1) +
  scale_fill_gradient(low = "lightblue", high = "darkgreen") +
  labs(title = "Access to Electricity by Country",
       fill = "Access to Electricity") +
  theme_minimal()
```




```
# Plot Renewable Energy Consumption on a world map
ggplot() +
  geom_sf(data = merged_sf, aes(fill = renewable_energy_consumption), color = "white", size = 0.1) +
  scale_fill_gradient(low = "lightblue", high = "darkred") +
  labs(title = "Renewable Energy Consumption by Country",
       fill = "Renewable Energy Consumption") +
  theme_minimal()
```

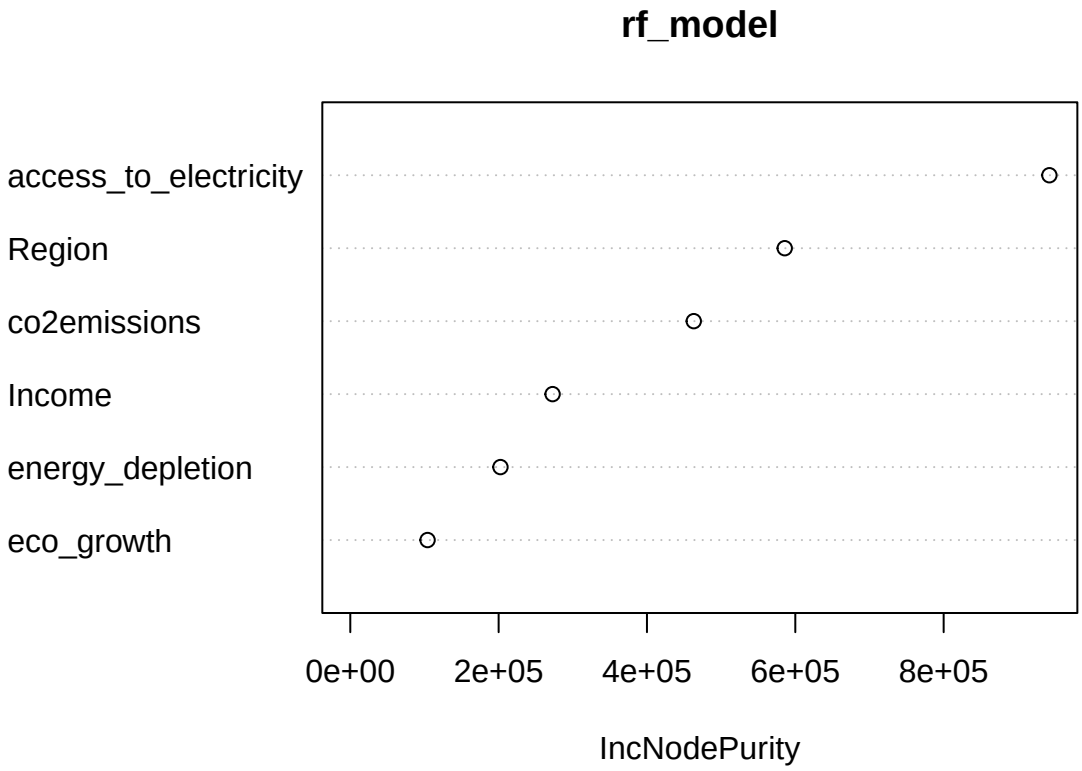


Question 6: Can we predict renewable energy consumption based on other variables? How effective is the machine learning model in predicting renewable energy consumption? and what is the level of agreement between the model's predictions and the actual values, as measured by the confusion matrix?

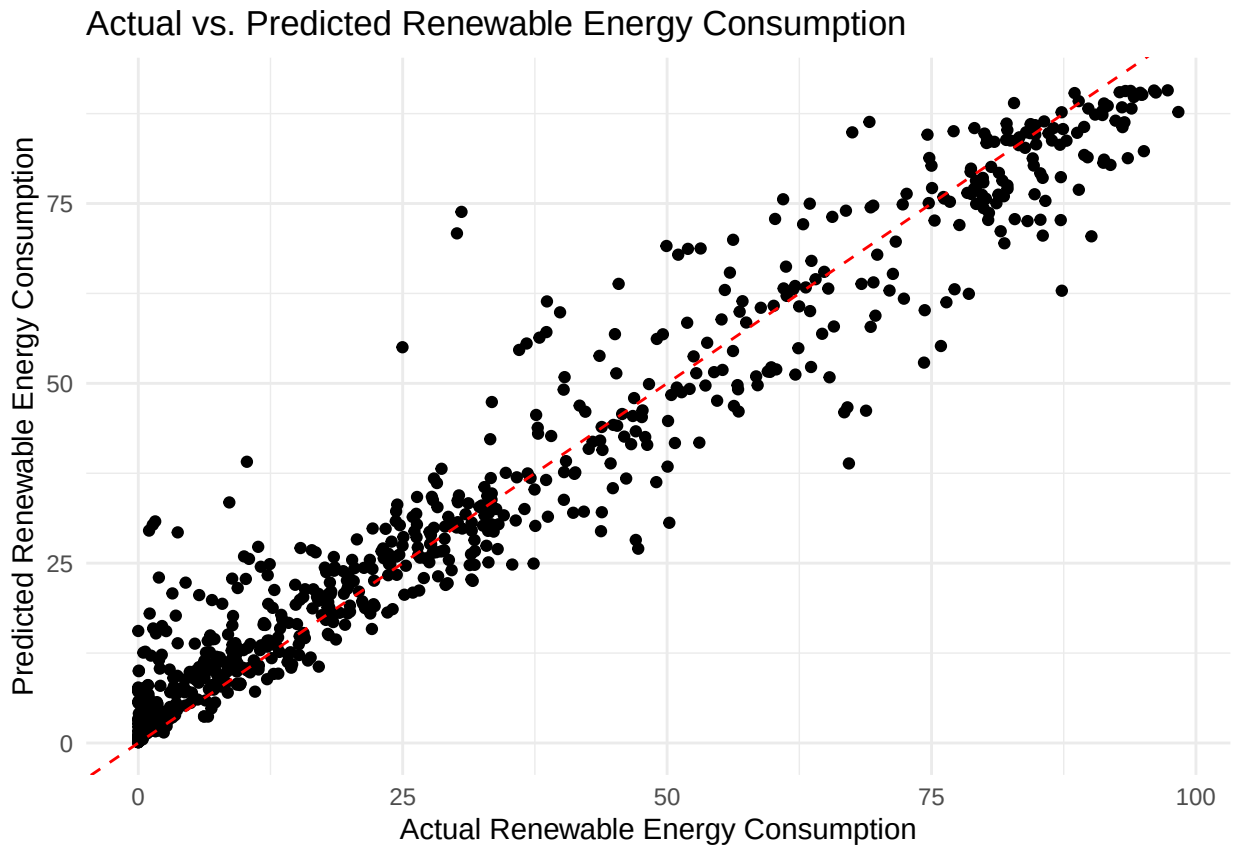
```
# Split the data into training and testing sets to train and evaluate the machine learning model.
set.seed(123)
train_indices <- createDataPartition(filledData$renewable_energy_consumption, p = 0.8, list = FALSE)
train_data <- filledData[train_indices, ]
test_data <- filledData[-train_indices, ]

# Train a Random Forest model
rf_model <- randomForest(renewable_energy_consumption ~ eco_growth + co2emissions + access_to_electrici

# Make predictions on the test set
predictions <- predict(rf_model, newdata = test_data)

# Evaluate model performance
rmse <- sqrt(mean((test_data$renewable_energy_consumption - predictions)^2))
cat("Root Mean Squared Error (RMSE):", rmse, "\n")
```

```
## Root Mean Squared Error (RMSE): 7.509509
```

```r
# Feature importance
importance <- importance(rf_model)
varImpPlot(rf_model)
```




```r
# Include predictions in the test_data dataframe
test_data$predicted_renewable_energy <- predictions

# Visualize actual vs. predicted renewable energy consumption
ggplot(test_data, aes(x = renewable_energy_consumption, y = predicted_renewable_energy)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Actual vs. Predicted Renewable Energy Consumption",
       x = "Actual Renewable Energy Consumption",
       y = "Predicted Renewable Energy Consumption") +
  theme_minimal()
```




```
# Define a threshold for classification
threshold <- 0.5

# Convert renewable energy consumption and predictions to binary classes
actual_class <- ifelse(test_data$renewable_energy_consumption > threshold, 1, 0)
predicted_class <- ifelse(test_data$predicted_renewable_energy > threshold, 1, 0)

# Create a confusion matrix
conf_matrix <- confusionMatrix(factor(predicted_class), factor(actual_class))
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  15   0
##          1  43 675
##
##                Accuracy : 0.9413
##                  95% CI : (0.9218, 0.9572)
##     No Information Rate : 0.9209
##     P-Value [Acc > NIR] : 0.02025
##
##                   Kappa : 0.3912
##
##  Mcnemar's Test P-Value : 1.504e-10
```

```
##
##             Sensitivity : 0.25862
##             Specificity : 1.00000
##          Pos Pred Value : 1.00000
##          Neg Pred Value : 0.94011
##              Prevalence : 0.07913
##          Detection Rate : 0.02046
##    Detection Prevalence : 0.02046
##       Balanced Accuracy : 0.62931
##
##        'Positive' Class : 0
##
```