Capstone - World Sustainability Data Science Project

Ammar A. Thuraya

6/5/2022

Introduction

In this project, we will use the latest global Sustainability dataset. The 'global sustainability' dataset tracks the performance of 173 countries against a range of sustainability metrics over 19 years from 2000 to 2018. The dataset was generated from several merged data sources with real-world data. There are more than 50 fields and many records with missing data in different records in the original dataset.

This project will measure the evolution and progress towards sustainability across different countries from the year 2000 to 2018. Through the data analysis in this study, we will select a few variables that could affect the progress, namely the following: (1) % of renewable energy consumption out of total energy consumption (2) % of access to electricity out of the total population (3) Annual production-based CO2 emissions in millions of tons (4) the time effect (year) (5) the countries with data points in the dataset

Then, we will do some cleaning on the original dataset, including omitting columns that are not addressed in this study. We will also clean the long field names and replace them with shorter, more meaningful names. Then, we will remove NAs from the dataset and create subsets that are useful for different parts of this project.

Once all work on the datasets is done, we will create a training dataset with +80% of the data records, and the remaining data will be used for the validation and test sets.

Afterward, we will build different models to predict the level of CO2 emissions contributed by different countries in the world, which directly impacts Earth's sustainability. We will start with simpler models that measure different variables' impact like energy, electricity, and time (years.) Then, we will build more advanced models like regularization and neural networks to find more optimum Root Mean Square Error (RMSE) to measure the quality of the implemented models. The less (smaller) the RMSE value, the better the model predicts the target variable results.

Creating the World Sustainability (wsus) and the validation datasets

Top rows of wsus subset with new selected column names:

```
##   country countryID year CO2em energy electricity
## 1   Aruba       ABW 2000 2.378 0.1753     91.6604
## 2   Aruba       ABW 2001 2.407 0.1805    100.0000
## 3   Aruba       ABW 2002 2.437 0.1814    100.0000
## 4   Aruba       ABW 2003 2.561 0.1846    100.0000
## 5   Aruba       ABW 2004 2.616 0.1871    100.0000
## 6   Aruba       ABW 2005 2.719 0.1866    100.0000
```

Top Renewable Energy Countries in year 2000:

```
##            country countryID year CO2em  energy electricity
## 1 Congo, Dem. Rep.       COD 2000 0.893 97.9403    6.700000
## 2         Ethiopia       ETH 2000 3.464 95.5545   12.700000
## 3           Uganda       UGA 2000 1.361 95.0125    7.318620
## 4         Tanzania       TZA 2000 2.571 93.7255    9.056112
## 5       Mozambique       MOZ 2000 1.323 93.6202    6.089136
```

Top production-based CO2 emission countries in year 2000:

```
##              country countryID year    CO2em  energy electricity
## 1      United States       USA 2000 5998.070  5.4297   100.00000
## 2              China       CHN 2000 3349.295 29.6030    96.93087
## 3 Russian Federation       RUS 2000 1471.052  3.4964   100.00000
## 4              Japan       JPN 2000 1264.844  3.7761   100.00000
## 5              India       IND 2000  978.427 51.5537    59.34105
```

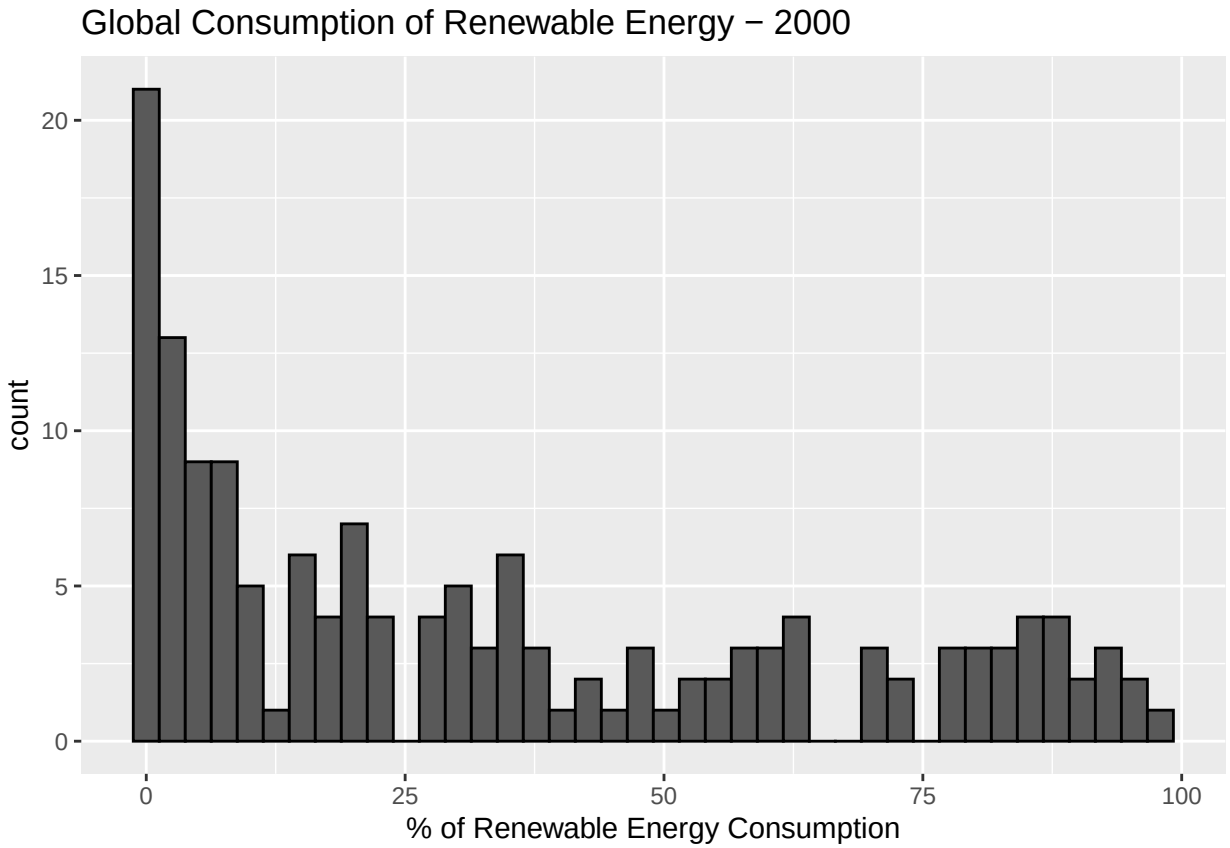
Top Renewable Energy Countries in year 2018:

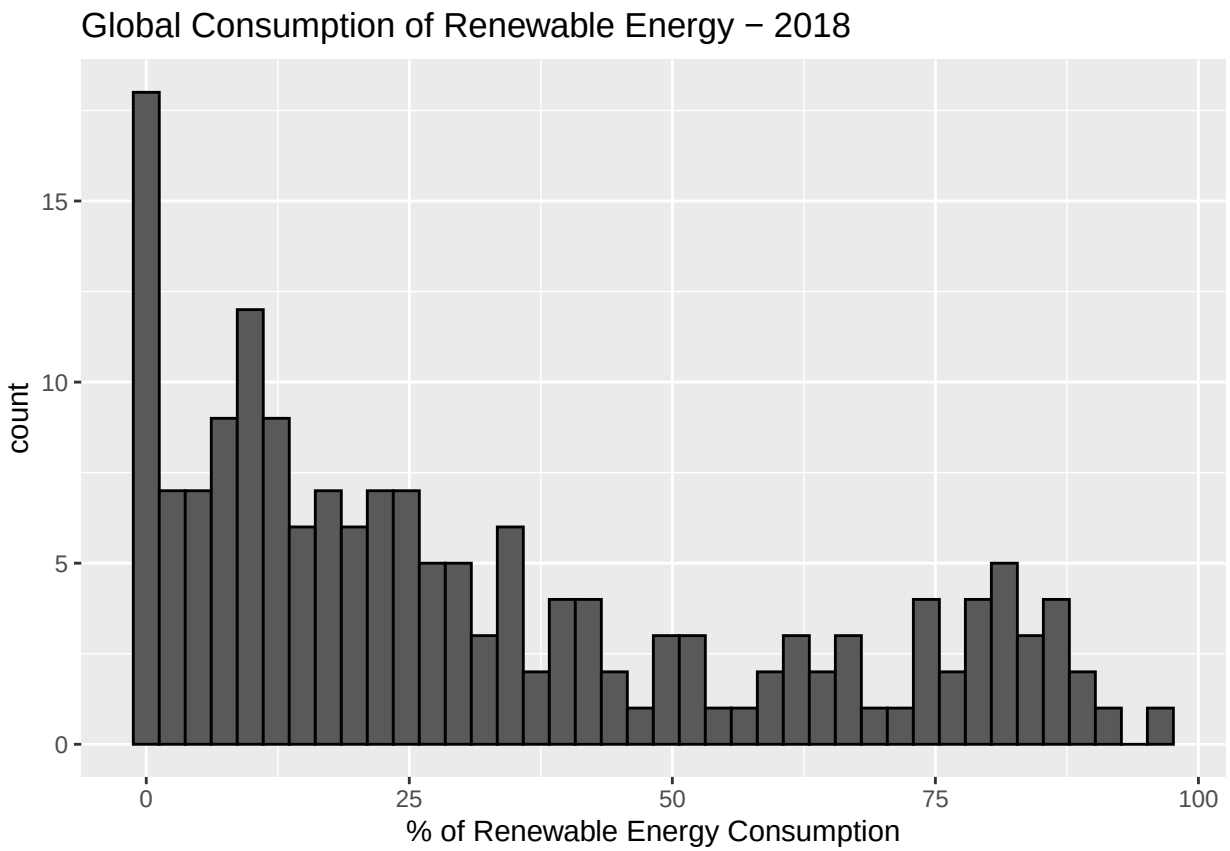
```
##            country countryID year  CO2em  energy electricity
## 1 Congo, Dem. Rep.       COD 2018  2.231 96.3837    18.75225
## 2           Uganda       UGA 2018  5.385 90.3339    42.70000
## 3         Ethiopia       ETH 2018 16.185 89.9231    45.05340
## 4            Gabon       GAB 2018  4.803 89.8871    89.84184
## 5          Liberia       LBR 2018  1.274 87.2133    24.81808
```

Top production-based CO2 emission countries in year 2018:

```
##              country countryID year    CO2em  energy electricity
## 1              China       CHN 2018 9956.569 13.1238    100.0000
## 2      United States       USA 2018 5424.882 10.1072    100.0000
## 3              India       IND 2018 2591.324 31.6892     95.1933
## 4 Russian Federation       RUS 2018 1691.360  3.1805    100.0000
## 5              Japan       JPN 2018 1135.688  7.3877    100.0000
```

plots


Global Consumption of Renewable Energy − 2000
count
20
15
10
5
0
0
25
50
75
100
% of Renewable Energy Consumption



Top production-based CO2 emission Countries in 2000 (+5M tons):

```
## # A tibble: 1 x 2
##   country       countryID
##   <chr>         <chr>
## 1 United States USA
```

Top production-based CO2 emission Countries in 2018 (+5M tons):

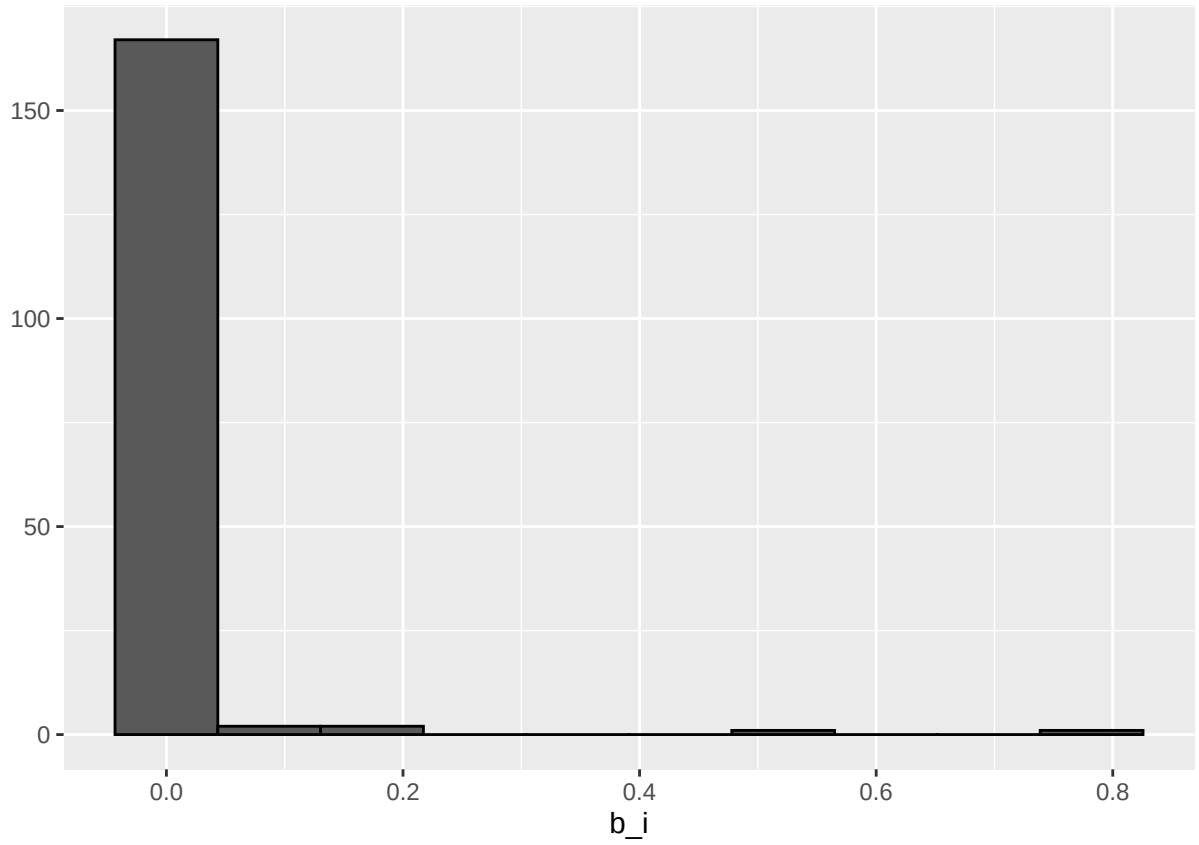
```
## # A tibble: 2 x 2
##   country       countryID
##   <chr>         <chr>
## 1 China         CHN
## 2 United States USA
```

Creating the Prediction Models

Model 1 - with Countries Global Impact:

```
CO2em_avgs <- train_set %>% group_by(countryID) %>%
  summarize(b_i = mean(CO2em - mu))
```

```
pred_CO2em_M1 <- test_set %>%
  left_join(CO2em_avgs, by='countryID') %>%
  mutate(pred = mu + b_i) %>%
  pull(pred)
```



Measuring the Quality of the Model:

When we use the fitted model to make predictions on new observations, we can use several different metrics to measure the quality of the model, including R-squared, RMSE, and MAE. Following is a brief description of each: 1. R-squared (Multiple R-squared): measures the strength of the linear relationship between the predictor and the response. The higher the multiple R-squared, the better the predictor variables can predict the response variable. 2. RMSE (Root Mean Squared Error): measures the average prediction error made by the model in predicting the value for a new observation. This is the average distance between the true value of an observation and the value predicted by the model. Lower values for RMSE indicate a better model fit. 3. MAE (Mean Absolute Error): measures the average absolute difference between the true value of an observation and the value predicted by the model. This metric is generally less sensitive to outliers than RMSE, and lower values for MAE indicate a better model fit.

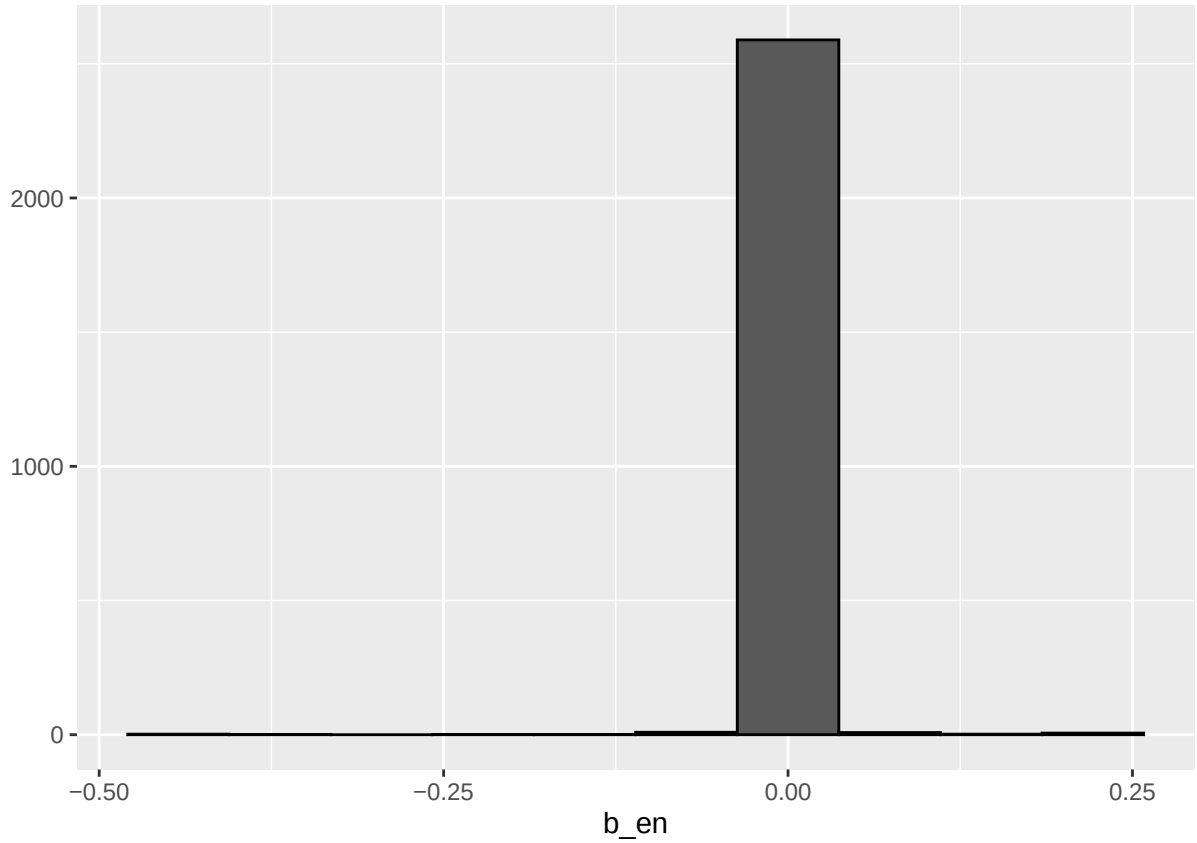
Model 1 Quality Results:

```
##   R_squared        RMSE         MAE
## 0.929682993 0.029757706 0.007538095
```

Model 2 - with Renewable Energy Effect

```
energy_avgs <- train_set %>%
  left_join(CO2em_avgs, by='countryID') %>%
  group_by(energy) %>%
  summarize(b_en = mean(CO2em - mu - b_i))

pred_CO2em_M2 <- test_set %>%
  left_join(energy_avgs, by='energy') %>%
  mutate(pred = mu + b_en) %>%
  pull(pred)
```



Model 2 Quality Results:

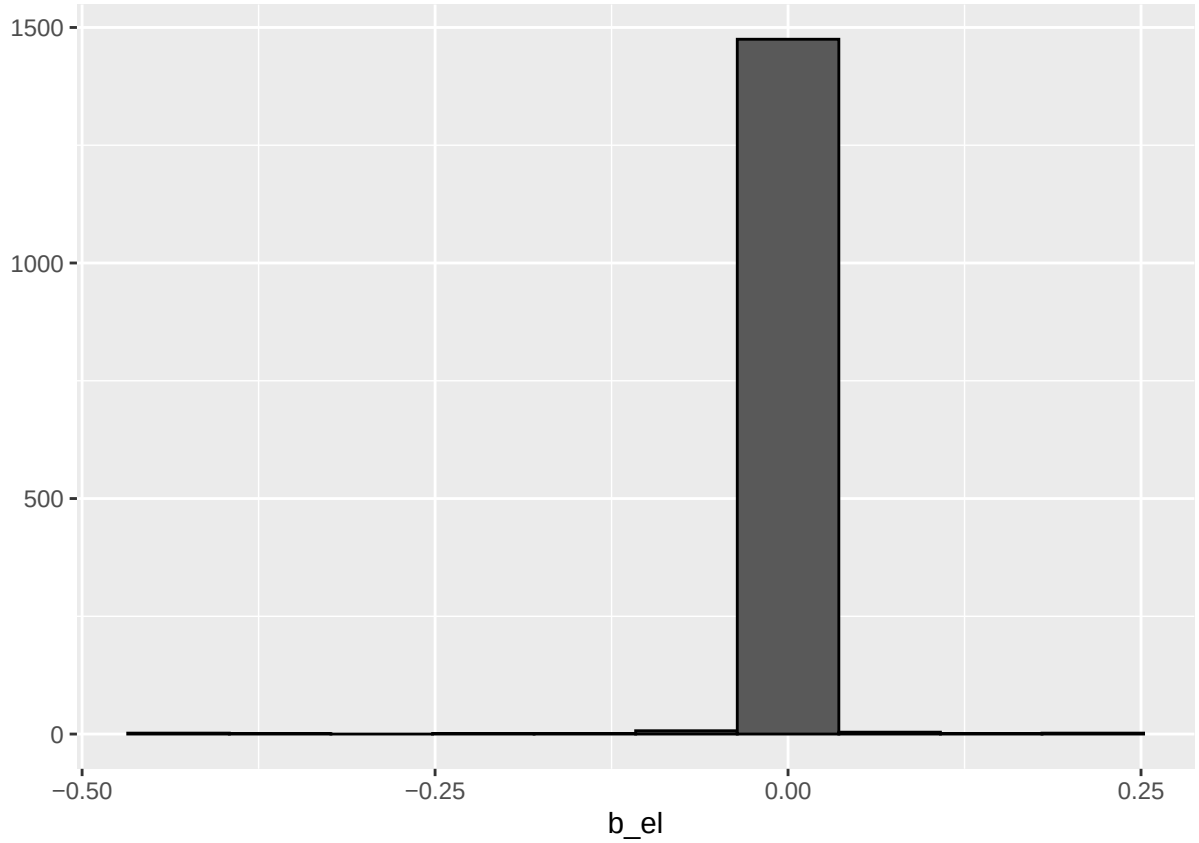
```
##  R_squared       RMSE        MAE
## 0.05764174 0.01400107 0.01290223
```

Model 3 - with Access to Electricity effect

```
electric_avgs <- train_set %>%
  left_join(CO2em_avgs, by='countryID') %>%
  group_by(electricity) %>%
```

```
  summarize(b_el = mean(CO2em - mu - b_i))

pred_CO2em_M3 <- test_set %>%
  left_join(electric_avgs, by='electricity') %>%
  mutate(pred = mu + b_el) %>%
  pull(pred)
```



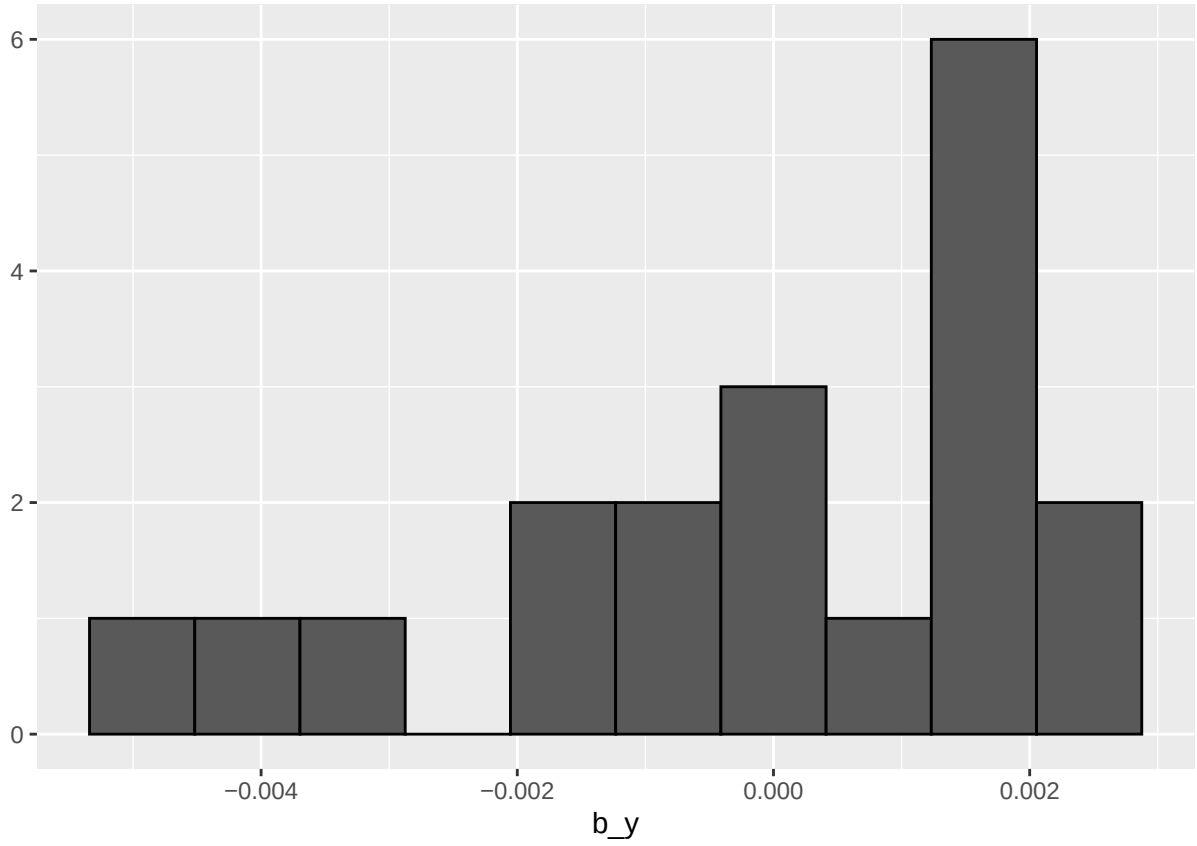
Model 3 Quality Results:

```
##    R_squared         RMSE          MAE
## 0.0005767342 0.1140557582 0.0380331187
```

Model 4 - with Time (year) effect

```
year_avgs <- train_set %>%
  left_join(CO2em_avgs, by='countryID') %>%
  group_by(year) %>%
  summarize(b_y = mean(CO2em - mu - b_i))

pred_CO2em_M4 <- test_set %>%
  left_join(year_avgs, by='year') %>%
  mutate(pred = mu + b_y) %>%
  pull(pred)
```



Model 4 Quality Results:

```
##    R_squared         RMSE          MAE
## 8.046762e-05 9.630319e-02 3.085448e-02
```

Model 5 - using Regularization method

Our regularization model will use the country, energy, electricity, and time (year) effects. Then, we will tune the overall impact of the regularization term by multiplying its value by a scalar known as lambda (also called the regularization rate.) We will then use 10-folds cross-validation to select the lambda values.

```
lambdas <- 10^seq(0, 100, 0.25)

regularize_RMSEs <- sapply(lambdas, function(l){
  b_i <- train_set %>%
    group_by(countryID) %>%
    summarize(b_i = sum(CO2em - mu)/(n()+l))

  b_en <- train_set %>%
    left_join(b_i, by="countryID") %>%
    group_by(energy) %>%
    summarize(b_en = sum(CO2em - b_i - mu)/(n()+l))

    b_el <- train_set %>%
```

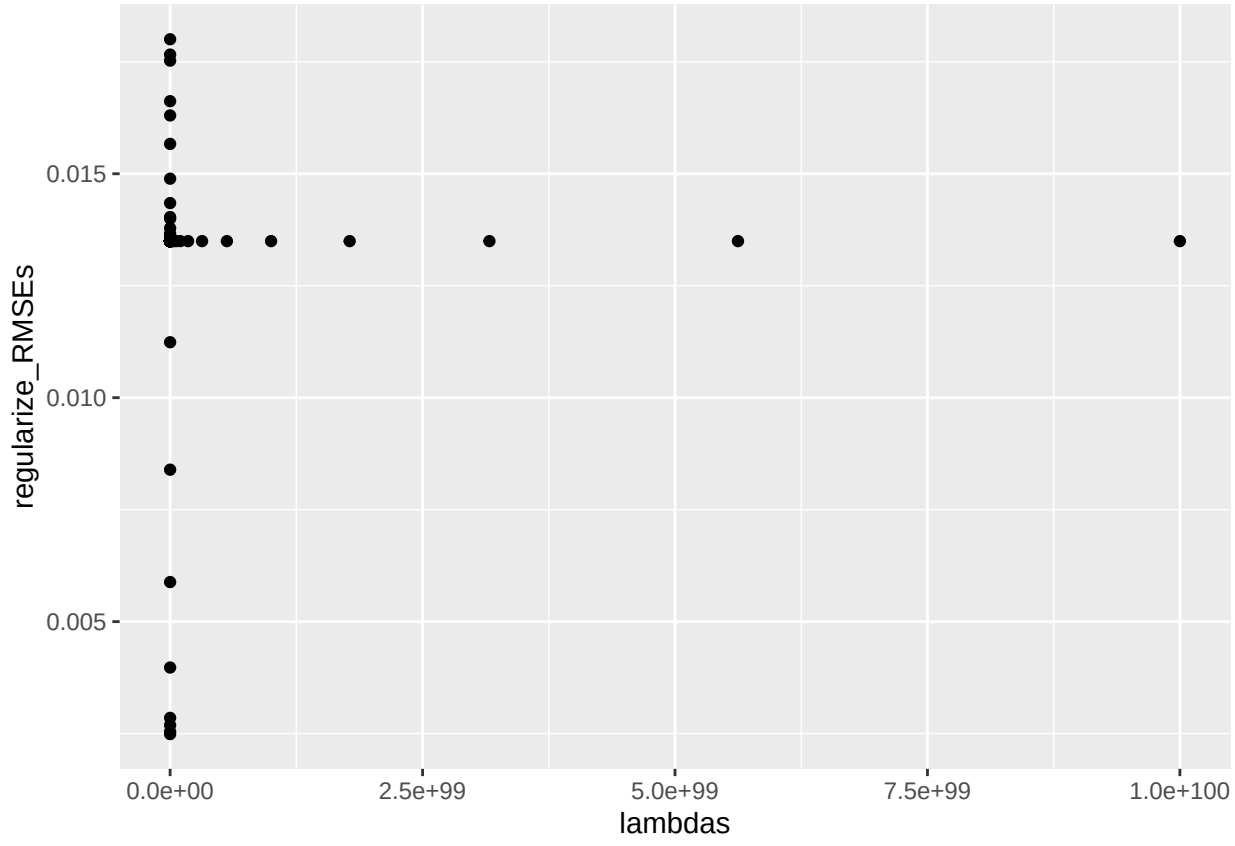
```
    left_join(b_i, by="countryID") %>%
    group_by(electricity) %>%
    summarize(b_el = sum(CO2em - b_i - mu)/(n()+l))

    b_y <- train_set %>%
    left_join(b_i, by="countryID") %>%
    group_by(year) %>%
    summarize(b_y = sum(CO2em - b_i - mu)/(n()+l))

  pred_CO2em <- test_set %>%
    left_join(b_i, by = "countryID") %>%
    left_join(b_y, by="year") %>%
    left_join(b_en, by = "energy") %>%
    left_join(b_el, by = "electricity") %>%
    mutate(pred = mu + b_i + b_y + b_en + b_el) %>% .$pred

  return(RMSE = RMSE(pred_CO2em, test_set$CO2em, na.rm=TRUE))

})
```



Minimum Lambda:

```
## [1] 3.162278
```

Model 5 RMSE result:

	M1_RMSE	M2_RMSE	M3_RMSE	M4_RMSE	M5_RMSE	M6_RMSE
RMSE	0.03	0.014	0.11	0.096	0.0025	0.095

```
## [1] 0.002488802
```

Model 6 - using Neural Network Model

A neural network can be imagined as a system consisting of many highly interconnected nodes, called 'neurons,' organized in layers that process information using dynamic state responses to external inputs. A neural network consists of three layers: Input Layer (that takes inputs based on existing data,) Hidden Layer (that optimizes the weights of the input variables in order to improve the predictive power of the model,) and Output Layer (output of predictions based on the data from the input and hidden layers.)

```
nn <- neuralnet(CO2em ~ energy + electricity, data=train_maxmindf, hidden=c(2,1), linear.output=FALSE, t

nn$result.matrix
```

```
##                                [,1]
## error                   8.22893484
## reached.threshold       0.00896715
## steps                  79.00000000
## Intercept.to.1layhid1  -0.80817855
## energy.to.1layhid1      3.26498600
## electricity.to.1layhid1 -2.78171959
## Intercept.to.1layhid2  -0.41007641
## energy.to.1layhid2      7.16991880
## electricity.to.1layhid2 -3.05519746
## Intercept.to.2layhid1  -0.49077126
## 1layhid1.to.2layhid1    0.23214678
## 1layhid2.to.2layhid1    3.81339402
## Intercept.to.CO2em     -1.96218182
## 2layhid1.to.CO2em      -3.48159698
```

Model 6 Quality Results:

```
##  R_squared       RMSE        MAE
## 0.03036075 0.09489150 0.02902651
```

Final RMSE results for all Models

Results:

As you can see from the above, we have implemented six different models to measure the global CO2 emission impact and how different countries can impact the global CO2 emissions by considering the time (progress in years), usage of renewable energy %, and access to electricity % effects. Each model generated a different RMSE result, and the Regularization Model generated the most optimized RMSE. In the Regularization model, we have used 10-folds cross validation to select the lambda value, and we were able to get the best (minimum) lambda (3.162) and corresponding RMSE (0.0025) outputs of this model.

It is important to note that we have done a lot of data cleansing before applying the different machine models to the training dataset. Then, we applied the validation (or test data-subset) to validate the outputs and yield the RMSE values for each model accordingly.

It is also important to note that we scaled the effect variables to have similar scales from 0 to 1 to avoid considerable fluctuation (some columns had percentage values, and some had considerable data variation like the CO2 emissions column in the original set.) That helped get normalized and more consistent views on the data before applying and testing the different models.

Conclusion and future considerations

We have used different prediction models to measure the CO2 emissions global impact with different variables that could affect sustainability, like usage of renewable energy, access to electricity, time (year), and the different countries that are driving the impact.

A few other variables could have affected sustainability that was not measured, namely: usage of renewable electricity out of total electricity usage and Natural resources rents, which is the difference between the price of a commodity and the average cost of producing it. The Total natural resources rents are the sum of oil rents, natural gas rents, coal rents, mineral rents, and forest rents. There are many other variables in the dataset that could be helpful to measure and consider in building future prediction models.

An updated World sustainability dataset should help provide many of the missing values in the current dataset (NAs,) and add the latest years that were not included in this dataset (e.g., from 2018 to 2021.)