Reducing Greenhouse Gas Emissions
Where Are Efforts Best Focused?

Adam Englund

1 Abstract

Greenhouse gas emissions are considered the key contributor to global warming[1]. Our goal is to see if the relative influence of the key indicators of greenhouse gas emissions can be measured and compared to get a sense of where effort should be prioritized. All of the data is from the World Bank Group, under the topic of climate change. The methodological approach is to create a linear regression model from which the relative importance of different indicators of greenhouse gas emissions can be determined. The key takeaway from this analysis is that oil based energy may dwarf other concerns in regard to reducing greenhouse gas emissions, however more granular data is required to form any definitive conclusions.

2 Introduction

We hear a lot about reducing emissions to tackle the problem of climate change. There is reporting on emissions targets, the Paris Climate Accords, carbon trading systems, over-population etc. It's reported that burning fossil fuels is something that needs to end, of beef and dairy being disastrous to the environment, and the challenges of renewable energy. There is agreement (for the most part) that all of these challenges are important in addressing climate change. However, whenever we are faced with finding a solution to a multi-faceted problem, a good first step is to priotize the various challenges involved. Can we do this with the problem of greenhouse gas emissions? Are we able to able to provide relative measures of the importance of the key contributors to greenhouse gas emissions?

3 Data

The data provided has been extracted from global development data made available by the World Bank Group, it is comprised of indicators that were categorized under the topic of climate change. An initial dataset was provided based on this data and then a subset of variables were selected that were considered relevant, plus some new variables were created based on the existing data (details below).

Since the goal is to try to determine the relative affects different variables have on greenhouse gas emissions, the variable 'EN.ATM.GHGT.KT.CE' (total greenhouse gas emissions as kt of CO2 equivalent) was initially chosen as the response variable. However, it was decided that a per capita measure of emissions is a fairer way to compare nations. So a new variable was created for this called 'EN.ATM.GHGT.KT.PC' (while trying to use the same naming convention) calculated by dividing 'EN.ATM.GHGT.KT.CE' by the total population ('SP.POP.TOTL').

To aid in some visualization, some other variables were also created:

- By classifying the total greenhouse gas emissions per capita as kt of CO2 equivalent ('EN.ATM.GHGT.KT.PC') as either 'Low', 'Medium, 'High' or 'Very High', by using the inter-quartile range of 'EN.ATM.GHGT.KT.PC'. This new variable was called 'EN.ATM.GHGT.CLS.PC'.

- By classifying the total production of electricity by fossil fuels as either 'Low', 'Medium, 'High' or 'Very High', by summing % electricity produced by coal, natural gas and petroleum and using the inter-quartile range. This new variable was called 'EG.ELC.FOSS.CLS'.

The dataset required further cleaning as there were a number of missing values that affected the majority of the dataset, there were many indicators that were reflecting the same information, and there were also indicators that were clearly not relevant to the planned analysis. To handle removal of the indicators for the previously stated reasons, the indicators were sorted according to which had the largest number of missing values (please refer to Appendix C). This prioritized the order in which the indicators were considered to be surplus to requirements. It was considered appropriate to use a mixture of energy usage indicators, natural resource usage indicators, and some economic indicators as the predictors for greenhouse gas emissions, and the reasons for keeping or removing a variable were based on this logic as well as whether a variable was duplicating information.

It was also decided to convert the electricity production indicators from percentages to kilowatts per hour. This was done by multiplying 'EG.USE.ELEC.KH.PC' (electric power consumption kWh per capita) by the percentage indicators for different energy sources (i.e. 'EG.ELC.COAL.ZS', 'EG.ELC.NUCL.ZS' etc.). This was then taken a step further to break down the predictor 'EG.USE.PCAP.KG.OE' (energy use kg of oil equivalent per capita) into energy use from electricity and other forms of energy usage. This was done by first converting 'EG.USE.PCAP.KG.OE' from units in kg of oil equivalent to units in kWh (at a ratio of 1:11.63) to match the units used by 'EG.USE.ELEC.KH.PC'. 'EG.USE.ELEC.KH.PC' was then subtracted from the converted value of 'EG.USE.PCAP.KG.OE' (in kWh) to create a new predictor called 'EG.USE.OTHR.KH.PC' (energy use from source other than electricity kWh per capita).

```
wbcc_cln$EG.USE.OTHR.KH.PC <- wbcc_cln$EG.USE.PCAP.KG.OE*11.63 - wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.COAL.KH.PC <- wbcc_cln$EG.ELC.COAL.ZS * wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.HYRO.KH.PC <- wbcc_cln$EG.ELC.HYRO.ZS * wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.NGAS.KH.PC <- wbcc_cln$EG.ELC.NGAS.ZS * wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.NUCL.KH.PC <- wbcc_cln$EG.ELC.NUCL.ZS * wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.PETR.KH.PC <- wbcc_cln$EG.ELC.PETR.ZS * wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.RNWX.KH.PC <- wbcc_cln$EG.ELC.RNWX.ZS * wbcc_cln$EG.USE.ELEC.KH.PC
wbcc_cln$EG.ELC.RNEW.KH.PC <- wbcc_cln$EG.ELC.RNWX.ZS + wbcc_cln$EG.ELC.HYRO.KH.PC
wbcc_cln$EG.ELC.FOSS.KH.PC <- wbcc_cln$EG.ELC.COAL.KH.PC + wbcc_cln$EG.ELC.NGAS.KH.PC + wbcc_cln$EG.ELC.PETR.KH.PC
```

At the end of this process there were 16 indicators and 130 countries left from the original dataset, please see Appendix D for the full list of indicators with their descriptions.

4 Methods

Some exploratory analysis was initially conducted, first using a simple scatter plot to view the most obvious relationship, and then using principal component analysis and k-means clustering to try to determine whether patterns or relationships exist in the higher dimensional data we have. This revealed some evidence of some interesting relationships within the data. From there, linear regression was performed with feature selection to determine a model that would allow us to compare the relative impact of the most important predictors on the response. More detailed information for each of these steps is outlined in the rest of this section.

A relationship that seemed like it would have a strong correlation is between the CO2 emissions per capita ('EN.ATM.CO2E.PC') and energy use per capita ('EG.USE.PCAP.KG.OE'). The scatter plot in Figure 1 shows this relationship based on the cleaned dataset, it also contains information about the level of fossil fuel use in electricity production:

There appears to be a strong positive linear relationship, however we can immediately spot some interesting outliers. Qatar, Iceland, and Trinidad and Tobago are all conspicuously anomalous, and there seems to be some clear separation between nations that have over 10 metric tons of CO2 emissions per capita. Most of the countries in this group appear to be oil producing nations e.g. U.S., Russia, Canada, some middle eastern countries etc. One of the nations that immediately pops out is Canada, when we look into the data Canada is producing 63% of it's electricity from renewable energy sources, and the very high CO2 emissions seem

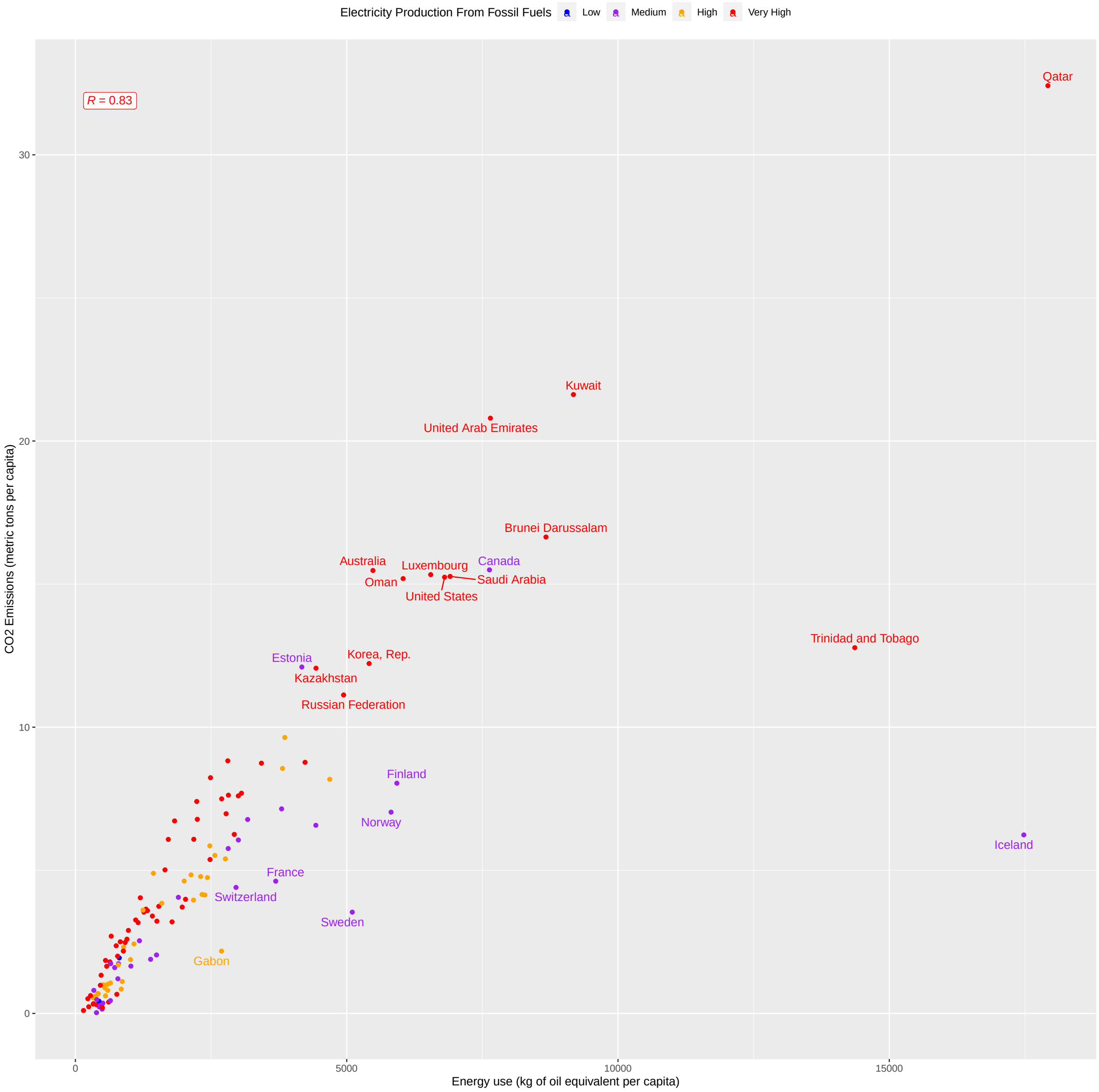

Figure 1: Energy Us vs CO2 Emissions

to be accounted for by transportation and Canada's petroleum production (and in particular it's use of tar sands for this purpose[2]).

```
##    country EG.ELC.RNEW.ZS
## 33  Canada       63.01145
```

This highlights something interesting missing from the indicators we have available i.e. we have indicators for CO2 emissions from gaseous, liquid and solid fuel consumption, but there is no indicator for CO2 emissions from gas flaring (and potentially other high CO2 emitting activities of petroleum production). The same can be said of energy use kg of oil equivalent per capita ('EG.USE.PCAP.KG.OE'). An attempt has been made to separate out the electricity usage component from this variable (as detailed below), but we are unable to extract precise granular information about the fuel sources.

Given the above, we should expect that energy use per capita ('EG.USE.PCAP.KG.OE') will also have a strong relationship with the total greenhouse gas emissions per capita ('EN.ATM.GHGT.KT.PC'). From this, plus the interest we have in exploring the relationship of different types of electricity production, it was decided that breaking down the energy use per capita to account for electricity consumption would be a good idea. The new predictors created for this were outlined in the data section above.

4.1 Data Exploration Using PCA and Clustering

Some initial data exploration was performed using principal component analysis (PCA) and k-means clustering to see if we could find anything interesting in the data. All of the predictors as well as the response were used in both PCA and k-means. The silhouette width criterion (see Figure 2) was used to compare the performance of the clustering, this suggested that 2-means clustering might be best.

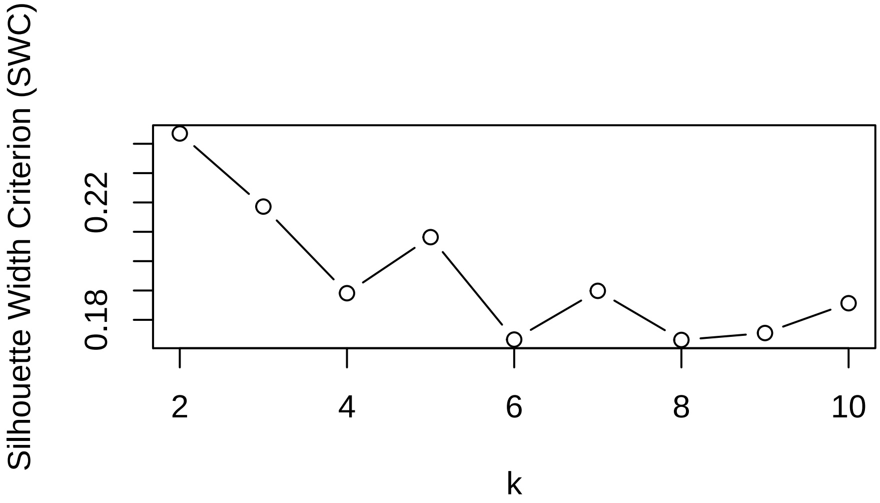

Figure 2: Comparing SWC for different values of k in k-means clustering

The PCA output using the first two principal components is shown in Figure 3, the top plot has coloured the data points according to the k-means clustering output, and in the bottom plot they are coloured by the response variable 'EN.ATM.GHGT.KT.PC' (total greenhouse gas emissions per capita kt of CO2 equivalent). Here we see evidence that the first two principal components are doing a pretty good job of explaining the response variable as we can see a clear pattern emerging. It is also evident when comparing the two plots that the k-means clustering has been effective at identifying very high emitters from the rest of the dataset in the 2 clusters.

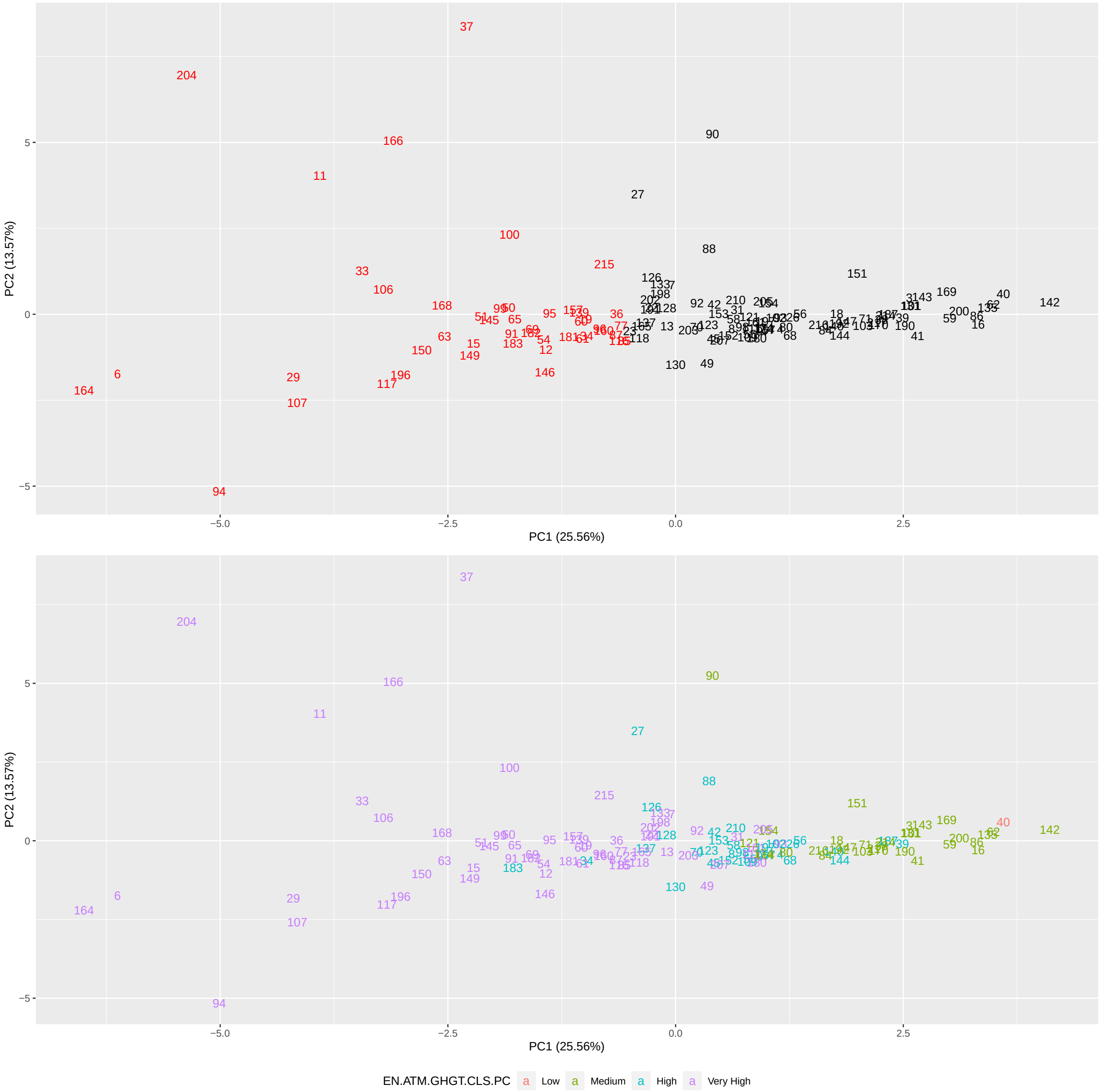

Figure 3: PCA Output Categorized by 2-means Clustering (top). And Low, Medium, High & Very High Levels of Total Greenhouse Gas Emissions Per Capita kt of CO2 Equivalent (bottom).

4.2 Linear Regression and Feature Selection

Given the results above, there is evidence that there is a subset of predictors that could work well in a predictive model, however the predictors in their original form don't satisfy the assumptions for linear regression. So the response and some of the predictors have been transformed to produce a model that does satisfy the assumptions for linear regression. A significant regression equation was found (F(16, 113) = 41.42, $p < 0.001$), with an R^2 of 0.85. However, as can be seen in Figure 4, even after these transformations have been included in the model there are still issues with normality and homoscedasticity, although the linearity assumption seems to be satisfied as the residuals are not too far away from 0.

```
##
## Call:
## lm(formula = log(EN.ATM.GHGT.KT.PC) ~ log(EG.USE.OTHR.KH.PC) +
##     log(EG.ELC.COAL.KH.PC + 1) + log(EG.ELC.NGAS.KH.PC + 1) +
##     log(EG.ELC.NUCL.KH.PC + 1) + log(EG.ELC.HYRO.KH.PC + 1) +
##     log(EG.ELC.RNWX.KH.PC + 1) + log(EG.ELC.PETR.KH.PC + 1) +
##     EG.ELC.ACCS.ZS + log(AG.LND.AGRI.K2 + 1) + log(AG.LND.FRST.K2 +
##     1) + AG.YLD.CREL.KG + NV.AGR.TOTL.ZS + I(1/ER.H2O.FWTL.K3) +
##     ER.LND.PTLD.ZS + IC.BUS.EASE.XQ + BX.KLT.DINV.WD.GD.ZS, data = wbcc_cln)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01654 -0.14627  0.01402  0.16968  0.94267
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.291e+01  6.935e-01 -18.615  < 2e-16 ***
## log(EG.USE.OTHR.KH.PC)      6.774e-01  5.605e-02  12.085  < 2e-16 ***
## log(EG.ELC.COAL.KH.PC + 1)  1.415e-02  8.031e-03   1.762 0.080704 .
## log(EG.ELC.NGAS.KH.PC + 1) -7.795e-03  7.885e-03  -0.989 0.324940
## log(EG.ELC.NUCL.KH.PC + 1) -1.640e-02  8.095e-03  -2.025 0.045189 *
## log(EG.ELC.HYRO.KH.PC + 1) -2.909e-02  1.206e-02  -2.413 0.017444 *
## log(EG.ELC.RNWX.KH.PC + 1) -2.830e-02  1.141e-02  -2.479 0.014648 *
## log(EG.ELC.PETR.KH.PC + 1)  1.565e-02  1.277e-02   1.226 0.222689
## EG.ELC.ACCS.ZS              9.584e-03  2.424e-03   3.953 0.000135 ***
## log(AG.LND.AGRI.K2 + 1)     2.792e-02  2.675e-02   1.044 0.298914
## log(AG.LND.FRST.K2 + 1)     3.049e-02  2.312e-02   1.319 0.189915
## AG.YLD.CREL.KG              1.750e-05  1.225e-05   1.429 0.155902
## NV.AGR.TOTL.ZS             -2.825e-03  6.309e-03  -0.448 0.655187
## I(1/ER.H2O.FWTL.K3)         8.758e-03  1.102e-02   0.795 0.428501
## ER.LND.PTLD.ZS              6.374e-04  3.153e-03   0.202 0.840174
## IC.BUS.EASE.XQ             -1.232e-03  1.053e-03  -1.169 0.244716
## BX.KLT.DINV.WD.GD.ZS       -8.899e-04  3.096e-03  -0.287 0.774274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3485 on 113 degrees of freedom
## Multiple R-squared:  0.8543, Adjusted R-squared:  0.8337
## F-statistic: 41.42 on 16 and 113 DF,  p-value: < 2.2e-16
```

To try to fix the model assumptions, nine outliers were identified and removed from the dataset to check whether this improved the situation. Given that some of the data could be from very different time periods (since the data was collected over a 20 year period and only the latest values were included), this could explain some of these outliers and it would be preferable to remove them. But even if they are legitimate data points, it was considered safe to remove them as the goal is to find some type of gauge for the relative affect of the predictors.

The new residuals vs fitted and normal Q-Q plot in Figure 5 shows that removing the outliers means that the model now meets the normality and homoscedasticity assumptions for linear regression. R^2 also improved to 0.91.

The next step was to see if the dimensions could be reduced and find the most important predictors. LASSO was considered a good approach for this as it provides feature selection, the ability to perform cross validation, and provides some indication of the order of importance of the predictors.

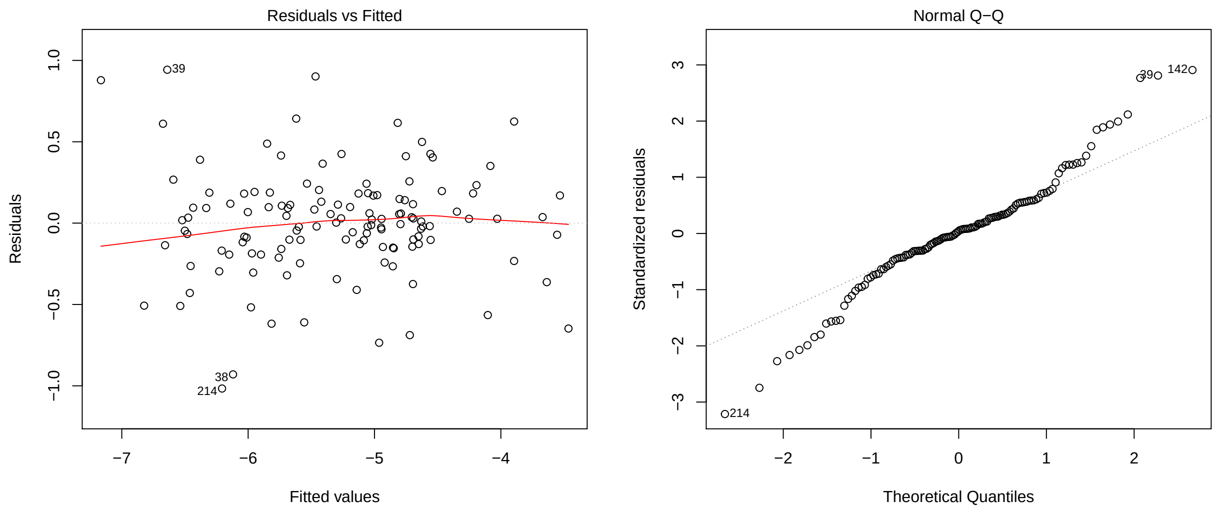

Figure 4: Checking for Linearity, Homoscedasticity & Normality Using Residuals vs Fitted and Q-Q Plots in Linear Regression Model with Transformed Predictors

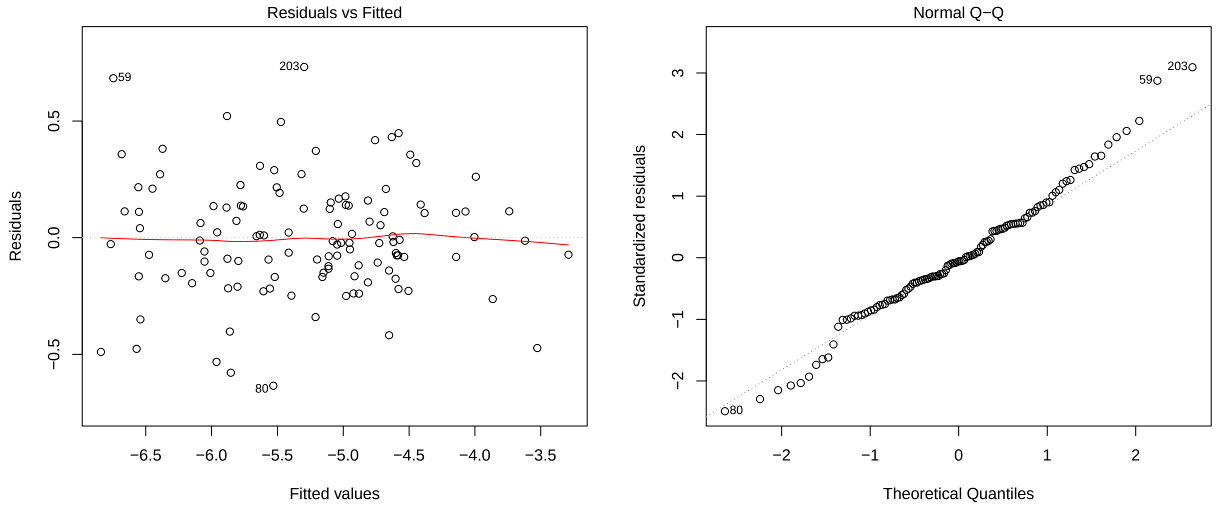

Figure 5: Residuals vs Fitted and Q-Q Plots in Linear Regression Model with Transformed Predictors, and After Outliers Have Been Removed

Since the cross validation uses a random sampling process (10-fold cross validation was used in this case), it was run multiple times (this was done manually and hasn't been included in this report). The plot of the mean-squared error (MSE) (averaged over the 10 folds) vs the natural logarithm of the penalty term, was used to help with the feature selection. The results varied in these plots from between as many as 12 variables at the minimum MSE, to as few as 3 variables at one standard error from the minimum MSE (see Figure 6 for an example of this plot).

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                -12.967879452
## (Intercept)                  .
## log(EG.USE.OTHR.KH.PC)       0.720182075
## log(EG.ELC.COAL.KH.PC + 1)   .
## log(EG.ELC.NGAS.KH.PC + 1)   .
## log(EG.ELC.NUCL.KH.PC + 1)  -0.006699607
## log(EG.ELC.HYRO.KH.PC + 1)  -0.007393383
## log(EG.ELC.RNWX.KH.PC + 1)  -0.006105832
## log(EG.ELC.PETR.KH.PC + 1)   0.003086215
## EG.ELC.ACCS.ZS               0.007704162
## log(AG.LND.AGRI.K2 + 1)      0.019679456
## log(AG.LND.FRST.K2 + 1)      .
## AG.YLD.CREL.KG               .
## NV.AGR.TOTL.ZS              -0.004362957
## I(1/ER.H2O.FWTL.K3)          .
## ER.LND.PTLD.ZS               .
## IC.BUS.EASE.XQ               .
## BX.KLT.DINV.WD.GD.ZS         .
```

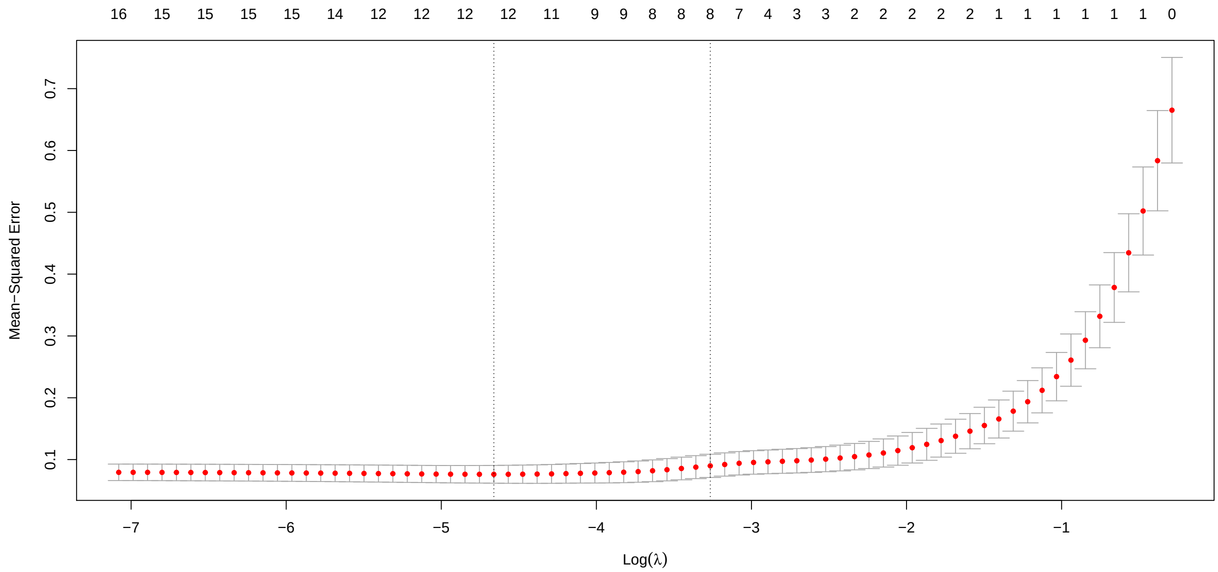

Figure 6: Comparing values of log(λ) against MSE in LASSO for Feature Selection

After reviewing the results from running the cross validation multiple times, the mid-point between the minimum MSE and one standard error away was considered a good point at which to do the feature selection (the was done in R by retrieving the coefficients and supplying the lambda value by subtracting the lambda value at 1 SE from the lambda value at the minimum MSE). At this mid-point, there were usually 8 predictors left (the output in the report may differ due to the randomness of the cross validation process). These 8 predictors were most commonly:

log(EG.USE.OTHR.KH.PC), log(EG.ELC.NUCL.KH.PC + 1), log(EG.ELC.HYRO.KH.PC + 1), log(EG.ELC.RNWX.KH.PC + 1), log(EG.ELC.PETR.KH.PC + 1), log(AG.LND.AGRI.K2 + 1), EG.ELC.ACCS.ZS, NV.AGR.TOTL.ZS

These 8 predictors will be used in the linear regression model for the analysis provided in the next section of the report. However, it was decided to also add ‘log(EG.ELC.COAL.KH.PC + 1)’ and ‘log(EG.ELC.NGAS.KH.PC + 1)’, even though they don’t have much affect on the overall model, we are interested in comparing them to the other electricity usage variables.

5 Results and Discussion

The very low p-value for the F-statistic suggests that this model is statistically significant in predicting the response. Given the number of predictors, we can’t read too much into the value of R^2 of 0.9, however it suggests that a large part of the variance is explained by this model. Our residuals vs fitted and normal Q-Q plots in Figure 7 look good in regard to the model satisfying the assumptions for linear regression of linearity, homoscedasticity and normality. (See Appendix B for the linear regression equation.)

```
## 
## Call:
## lm(formula = log(EN.ATM.GHGT.KT.PC) ~ log(EG.USE.OTHR.KH.PC) + 
##     log(EG.ELC.COAL.KH.PC + 1) + log(EG.ELC.NGAS.KH.PC + 1) + 
##     log(EG.ELC.NUCL.KH.PC + 1) + log(EG.ELC.HYRO.KH.PC + 1) + 
##     log(EG.ELC.RNWX.KH.PC + 1) + log(EG.ELC.PETR.KH.PC + 1) + 
##     EG.ELC.ACCS.ZS + NV.AGR.TOTL.ZS + log(AG.LND.AGRI.K2 + 1), 
##     data = wbcc_cln)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.64471 -0.16228 -0.02188  0.13997  0.69527 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                -13.878375   0.509091 -27.261  < 2e-16 ***
## log(EG.USE.OTHR.KH.PC)       0.774371   0.045482  17.026  < 2e-16 ***
## log(EG.ELC.COAL.KH.PC + 1)   0.006678   0.005789   1.154 0.251177    
## log(EG.ELC.NGAS.KH.PC + 1)  -0.010670   0.006094  -1.751 0.082739 .  
## log(EG.ELC.NUCL.KH.PC + 1)  -0.017653   0.005997  -2.944 0.003959 ** 
## log(EG.ELC.HYRO.KH.PC + 1)  -0.013690   0.008142  -1.681 0.095545 .  
## log(EG.ELC.RNWX.KH.PC + 1)  -0.021994   0.007346  -2.994 0.003402 ** 
## log(EG.ELC.PETR.KH.PC + 1)   0.007112   0.008610   0.826 0.410597    
## EG.ELC.ACCS.ZS               0.010647   0.001683   6.326 5.56e-09 ***
## NV.AGR.TOTL.ZS              -0.008114   0.004961  -1.635 0.104813    
## log(AG.LND.AGRI.K2 + 1)      0.052285   0.014842   3.523 0.000623 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2575 on 110 degrees of freedom
## Multiple R-squared:  0.909,  Adjusted R-squared:  0.9007 
## F-statistic: 109.9 on 10 and 110 DF,  p-value: < 2.2e-16
```

The table below summarizes the percentage changes in total greenhouse gas emissions per capita as kt of CO2 equivalent, for every 1% increase in the indicator. (The conversions used to interpret the log transformed dependent and independent variables are provided in Appendix A.)

Indicator	% Change in Emissions
Access to electricity (% of population)	1.07%
Energy use other than electricity consumption (kWh per capita)	0.77%
Agricultural land (sq. km)	0.05%
Electricity production from oil sources (kWh per capita)	0.01%
Electricity production from coal sources (kWh per capita)	0.01%
Electricity production from natural gas sources (kWh per capita)	-0.01%
Electricity production from hydroelectric sources (kWh per capita)	-0.01%

Indicator	% Change in Emissions
Electricity production from nuclear sources(% of total)	-0.02%
Electricity production from renewable sources, excl hydroelectric (kWh)	-0.02%
Agriculture, forestry, fishing, and hunting, value added (% of GDP)	-0.81%

Fortunately, seven of our ten predictors are in the same unit of measure kWh, and immediately puts into perspective relative affects of energy consumption as electricity versus energy consumption in forms that are not electricity. For example, a 1% increase in electricity consumption from coal production per capita results in a 0.01% increase in greenhouse gas emissions per capita, whereas a 1% increase in energy consumption excluding electricity per capita results in a 0.77% increase in greenhouse gas emissions, or 77 times more than coal based electricity consumption.

To put the indicator for agricultural land (km^2) into perspective, we'll use an example using data for Australia. According to the dataset, Australia has a total of 3588950 km^2 of agricultural land, and according to the Australian Bureau of Statistics (ABS), Australia has an average farm size[3] of 43.31 km^2. So a 1% increase in agricultural land (km^2) in Australia, is the equivalent of an extra 82867 average sized farms, resulting in an increase of 0.05% greenhouse gas emissions per capita (kt of CO2 equivalent).

For agriculture, forestry, fishing, and hunting, value added (% of GDP), every 1% of GDP increase in value added results in a reduction of 0.81% in greenhouse gas emissions.

And finally, for every increase of 1% of the population that has access to electricity, greenhouse gas emissions per capita (kt of CO2 equivalent) increase by 1.07%.

When analyzing these results, there is one variable that sticks out above all others: energy use other than electricity consumption (kWh per capita). This variable was derived from the original indicator for energy use (kg of oil equivalent per capita), where "total energy use refers to the use of primary energy before transformation to other end-use fuels (such as electricity and refined petroleum products". Since electricity usage has been removed to create the new variable, it's safe to assume that the majority of this value is due to energy derived from oil. Unfortunately we don't have precise figures about the make up of the original energy use per capita variable, so we can't draw any definitive conclusions, however, assuming oil makes up the majority of "energy use other than electricity consumption (kWh per capita)", it dwarfs all other concerns in regard to greenhouse gas emissions, as improvements in other areas will provide fractional benefits compared to addressing the oil problem.

6 Conclusions

This attempt to model the relative affects of the different contributors to greenhouse gas emissions has revealed that there is an elephant in the room in regard to reducing greenhouse gas emissions. In this report we named it: "energy use other than electricity consumption (kWh per capita)". We don't have any further information about the granular components of this variable, however, if we are to assume that it is comprised mainly of oil based energy, then addressing oil based energy appears to be the number one priority, and it could be argued that improvements in other contributors to greenhouse gas emissions are negligible in comparison.

However, if we were to identify a 2nd priority, there is evidence to suggest that access to electricity is a significant factor in the amount of greenhouse gas emissions, suggesting that renewable energy in developing nations (given that this variable is most likely to rise in developing nations, and that existing infrastructure may not be as widespread), may be a better focus for investment of renewable energy solutions as a priority over developed nations, and also has the benefit of potentially creating new economic opportunities.

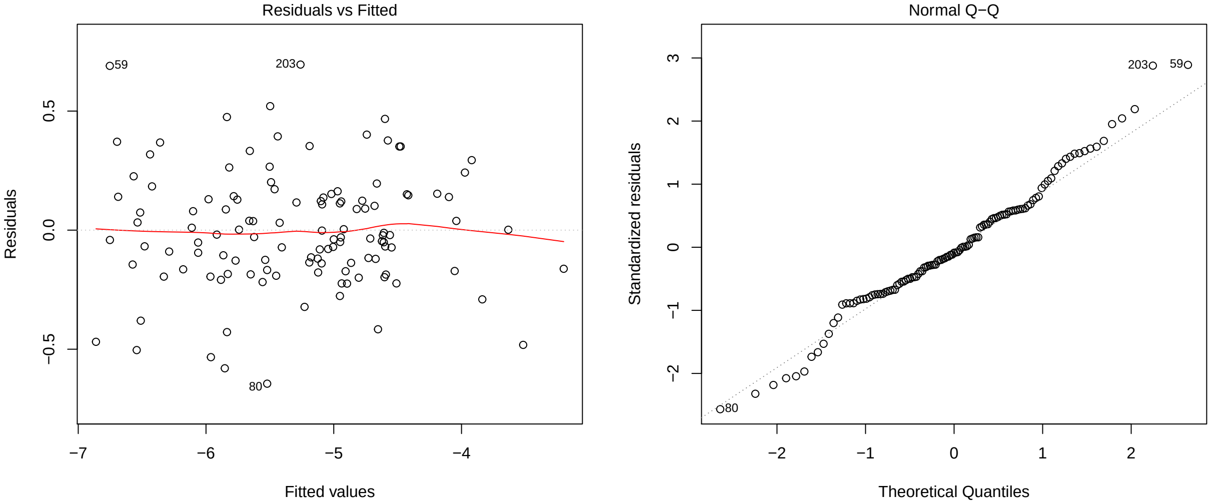

Figure 7: Residuals vs Fitted and Q-Q Plots in Linear Regression Model with Transformed Predictors, Outliers Removed & Feature Selection Performed

7 References

1. Josie Garthwaite. Stanford study finds stark differences in the carbon-intensity of global oil fields.
2. Special Report: Global Warming of 1.5ºC.
3. 2015-16 Agricultural Census, Australian Buruea of Statistics.
4. Clay Ford. Interpreting Log Transformations in a Linear Model.

---

8 Appendix A

From Interpreting Log Transformations in a Linear Model[4]. . .

Both dependent/response variable and independent/predictor variable(s) are log-transformed: Interpret the coefficient as the percent increase in the dependent variable for every 1% increase in the independent variable.

Only the dependent/response variable is log-transformed: Exponentiate the coefficient, subtract one from this number, and multiply by 100. This gives the percent increase (or decrease) in the response for every one-unit increase in the independent variable.

The dependent/response variable 'EN.ATM.GHGT.KT.PC' was log transformed.

So for 'EG.ELC.ACCS.ZS', a 1% increase (since the variable is already a percentage and wasn't transformed) is equal to: $(e^{0.010647} - 1) \times 100 =$ 1.0703881% increase in 'EN.ATM.GHGT.KT.PC'.

For 'NV.AGR.TOTL.ZS', a 1% increase (since the variable is already a percentage and wasn't transformed) is equal to: $(e^{-0.008114} - 1) \times 100 =$ -0.808117% increase in 'EN.ATM.GHGT.KT.PC'.

All other independent/predictor variables were log transformed, so the coefficients for these predictors can be interpreted as the percentage increase in 'EN.ATM.GHGT.KT.PC' for every 1% increase in the predictor.

---

9 Appendix B

The equation for the linear regresson model used in the analysis:

$$\begin{aligned}
log(EN.ATM.GHGT.KT.PC) = 0.77(log(EG.USE.OTHR.KH.PC)) + 0.007(log(EG.ELC.COAL.KH.PC + 1)) \\
-0.01(log(EG.ELC.NGAS.KH.PC + 1)) - 0.02(log(EG.ELC.NUCL.KH.PC + 1)) \\
-0.01(log(EG.ELC.HYRO.KH.PC + 1)) - 0.02(log(EG.ELC.RNWX.KH.PC + 1)) \\
+0.007(log(EG.ELC.PETR.KH.PC + 1)) - 0.008(NV.AGR.TOTL.ZS) \\
+0.01(EG.ELC.ACCS.ZS) + 0.05(log(AG.LND.AGRI.K2 + 1))
\end{aligned}$$

---

10 Appendix C

10.1 Part I Count of missing values by indicator variable

```r
wbcc <- read.csv("wbcc_bc.csv", stringsAsFactors=T)
# get a count of the number of missing values in each column
wbcc_na_count <- data.frame(sapply(wbcc, function(y) length(which(is.na(y)))))
```

```r
# get the variable names
variables <- rownames(wbcc_na_count)
# get the counts
na_count <- wbcc_na_count$sapply.wbcc..function.y..length.which.is.na.y....
# create a data frame using the variable names and counts from above
wbcc_na_count <- data.frame(variables, na_count)
# sort them by count of missing value (desc)
wbcc_na_count_sorted <- wbcc_na_count[order(wbcc_na_count$na_count, decreasing=T),]
# show the variables with counts > 100
head(wbcc_na_count_sorted, 20)
```

```
##            variables na_count
## 67    IS.ROD.PAVE.ZS      166
## 72    SH.MED.CMHW.P3      157
## 54 EN.CLC.GHGR.MT.CE      155
## 53    EN.CLC.DRSK.XQ      134
## 66    IQ.CPA.PUBS.XQ      130
## 59 EN.URB.MCTY.TL.ZS       96
## 13 AG.LND.IRIG.AG.ZS       93
## 46 EN.ATM.HFCG.KT.CE       80
## 51 EN.ATM.PFCG.KT.CE       80
## 52 EN.ATM.SF6G.KT.CE       80
## 18    EG.ELC.COAL.ZS       76
## 19    EG.ELC.HYRO.ZS       76
## 20    EG.ELC.NGAS.ZS       76
## 21    EG.ELC.NUCL.ZS       76
## 22    EG.ELC.PETR.ZS       76
## 24    EG.ELC.RNWX.KH       76
## 25    EG.ELC.RNWX.ZS       76
## 28 EG.USE.ELEC.KH.PC       75
## 73    SH.STA.MALN.ZS       67
## 74       SI.POV.DDAY       55
```

---

11 Appendix D

The below table contains the predictors used in the analysis:

Indicator	Description
AG.LND.AGRI.K2	Agricultural land (sq. km)
AG.LND.FRST.K2	Forest area (sq. km)
AG.YLD.CREL.KG	Cereal yield (kg per hectare)
ER.H2O.FWTL.K3	Annual freshwater withdrawals, total (billion cubic meters)
ER.PTD.TOTL.ZS	Terrestrial and marine protected areas (% of total territorial area)
EG.ELC.ACCS.ZS	Access to electricity (% of population)
EG.ELC.COAL.KH.PC	Electricity production from coal sources (kWh per capita)
EG.ELC.HYRO.KH.PC	Electricity production from hydroelectric sources (kWh per capita)
EG.ELC.NGAS.KH.PC	Electricity production from natural gas sources (kWh per capita)
EG.ELC.NUCL.KH.PC	Electricity production from nuclear sources (kWh per capita)
EG.ELC.PETR.KH.PC	Electricity production from oil sources (kWh per capita)
EG.ELC.RNWX.KH.PC	Electricity production from renewable sources, excluding hydroelectric (kWh)
EG.USE.OTHR.KH.PC	Energy use other than electricity consumption (kWh per capita)
NV.AGR.TOTL.ZS	Agriculture, forestry, fishing, and hunting, value added (% of GDP)
IC.BUS.EASE.XQ	Ease of doing business index (1=most business-friendly regulations)
BX.KLT.DINV.WD.GD.ZS	Foreign direct investment, net inflows (% of GDP)

Below are the countries used in the analysis (after outliers have been removed):

```
##  [1] Angola                 Albania                United Arab Emirates
##  [4] Argentina              Armenia                Australia
##  [7] Austria                Azerbaijan             Belgium
## [10] Benin                  Bangladesh             Bulgaria
## [13] Bosnia and Herzegovina Belarus                Bolivia
## [16] Brazil                 Botswana               Canada
## [19] Switzerland            Chile                  China
## [22] Congo, Dem. Rep.       Congo, Rep.            Colombia
## [25] Costa Rica             Cyprus                 Czech Republic
## [28] Germany                Denmark                Dominican Republic
## [31] Algeria                Ecuador                Egypt, Arab Rep.
## [34] Eritrea                Spain                  Estonia
## [37] Ethiopia               Finland                France
## [40] United Kingdom         Georgia                Ghana
## [43] Greece                 Guatemala              Honduras
## [46] Croatia                Haiti                  Hungary
## [49] Indonesia              India                  Ireland
## [52] Iran, Islamic Rep.     Iraq                   Israel
## [55] Italy                  Jamaica                Jordan
## [58] Japan                  Kazakhstan             Kenya
## [61] Kyrgyz Republic        Cambodia               Korea, Rep.
## [64] Kuwait                 Lebanon                Libya
## [67] Sri Lanka              Lithuania              Luxembourg
## [70] Latvia                 Morocco                Moldova
## [73] Mexico                 North Macedonia        Malta
## [76] Myanmar                Mongolia               Mozambique
## [79] Mauritius              Malaysia               Namibia
## [82] Nigeria                Nicaragua              Netherlands
##  [85] Norway               Nepal                New Zealand
##  [88] Oman                 Pakistan             Panama
##  [91] Peru                 Philippines          Poland
##  [94] Portugal             Paraguay             Qatar
##  [97] Romania              Russian Federation   Saudi Arabia
## [100] Sudan                Senegal              El Salvador
## [103] Suriname             Slovak Republic      Slovenia
## [106] Syrian Arab Republic Togo                 Thailand
## [109] Tajikistan           Tunisia              Turkey
## [112] Tanzania             Ukraine              Uruguay
## [115] United States        Uzbekistan           Venezuela, RB
## [118] Vietnam              South Africa         Zambia
## [121] Zimbabwe
## 217 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Zimbabwe
```