AN IN-DEPTH ANALYSIS INTO THE REAL CONTRIBUTERS TO GREENHOUSE GAS EMISSIONS

By Tristan Nicholls

STAT6020 Predictive Analytics

Table of Contents

Abstract

Climate change and more so human induced climate change has been a highly discussed issue for some time since the world has industrialized. Humans have become aware of the potential effect they may have on planet earth and the impact this could have on future generations. Therefore, a global effort is being made to reduce human's effect on the environment, led by events such as the United Nations Climate Change Conference, also known as the COP26 and the Glasgow Conference.

This report investigates this issue of climate change. More specifically, it investigates the issue surrounding greenhouse gas emissions and some of the main contributors. The goal of this report is not to point fingers, but deduce relationships between key gas emission and population measurements in an effort to conclude where the biggest impact can be made in improving our atmosphere.

World Bank Data, collated by the World Bank Group, under the category of climate change was extracted and analyzed in this investigation. It comprised of 77 key variables which may or may not have an effect on climate change, 18 of which were used for the purpose of this analysis. Population, renewable energy use and gas emission variables were used in an effort to deduce key relationships. Firstly, exploratory data analysis was conducted to determine the shape of the data and the variable of interest, total greenhouse gas emissions. Secondly, once the data has been pre-processed and readied for analysis, certain selected machine learning models were employed to help influence better decisions from a governing standpoint. Lastly, these models were evaluated to help guide the report towards some fascinating discoveries and conclusions.

Introduction

While there are a number of ways in which humans effect the environment, increasing greenhouse gas emissions have been at the forefront of discussion. Most concern has been placed around global warming and as a consequence rising sea levels along with increased natural disasters.

Figure 1 - Global Change in Temperature

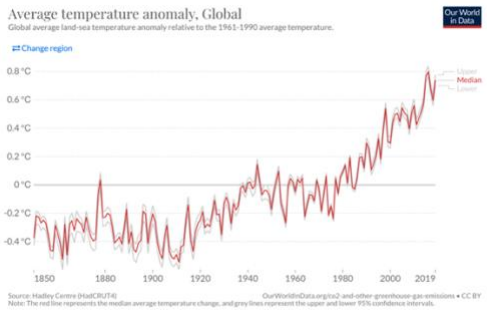


Figure 1 shows the increase in global temperature which has sparked such a deep discussion into the health of this planet. From around 1980 till 2019, planet earth experienced an increase in temperature by approximately 0.7 degrees Celsius.

Figure 2 - Change in Atmospheric CO2 Concentration

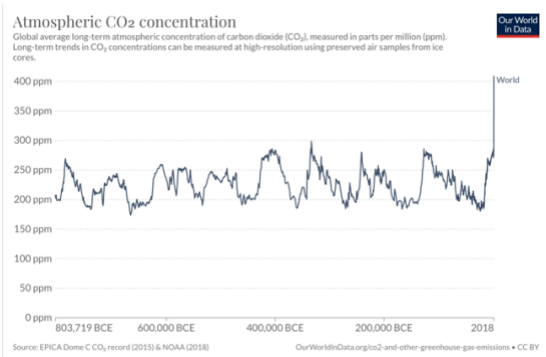


Figure 2 showcases the sudden surge in atmospheric CO2 concentration which has caused concern amongst humans. According to this graph, planet earth has never experienced this level of atmospheric CO2 concentration which has motivated quick action from most countries to mitigate.

Figure 3 - Trend and Share of CO2 Emissions by Country

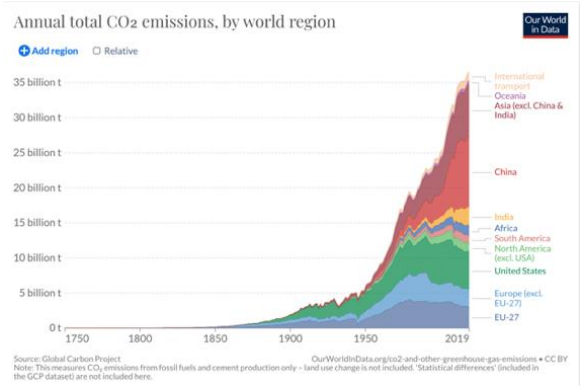


Figure 3 shows annual total CO2 emissions by world region. This figure highlights that Chin and the USA make up a large proportion of CO2 emissions in comparison to the rest of the world. This graph also illustrates that some locations are starting to experience a decrease in their CO2 emissions whilst others are still experiencing a sharp incline.

Figure 4 - Predictions Based on Actions Taken Now

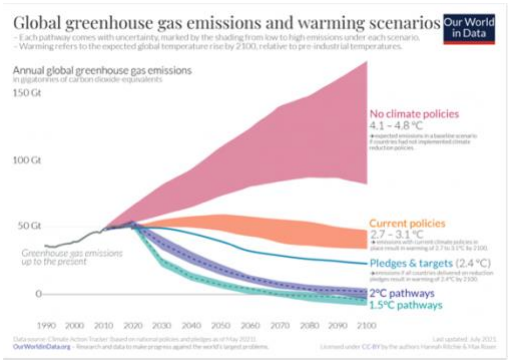


Figure 4 illustrates some scenarios based off what actions the human race decide to take. This illustration suggests that under current climate policies, planet earth will endure a temperature increase of between 2.7 and 3.1 degrees Celsius by 2100.

Figure 5 - Actions Starting to be Taken By Many Countries

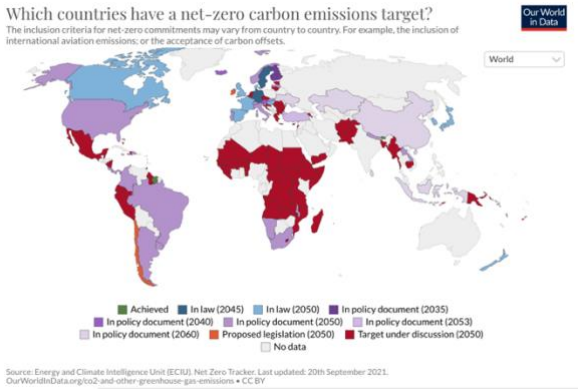


Figure 5 summarizes which countries are pushing towards a net-zero carbon emissions target and by what year. As seen, around half the planet has committed or started talking about accomplishing this goal while other have yet to do so.

It is these figures that have brought around concern amongst humans as to what the future of this planet looks like. In light of the upcoming Glasgow Conference on climate change, the remainder of this report dives deep into the data collected by the World Bank to determine the legitimacy of these claims, and determine where world leaders must focus their attention to yield the most significant improvements within their climate change objectives.

Data

The data used for this analysis was collected and organized by the World Bank Group. The World Bank Group is a global partnership comprising of five institutions who are working for sustainable solutions that reduce poverty and build shared prosperity in developing countries. (World Bank, 2021). The data on climate change is part of a larger group of World Development Indicators all collected by The World Bank and is considered to be some of the most current and accurate global development data available.

It is important to note that the data collected is at the time that measurement was last taken and ranges between the years 2001 – 2021. For example, the CO2 emissions recorded for Ukraine could have been a measurement taken anywhere between the years 2001 and 2021. This does raise some problems as a large difference in gas emissions can be expected between these two different time stamps. Additionally, comparing two measurements from two completely different time stamps can lead to some inaccuracies in the results.

For this analysis, 18 different variables were used out of the 77 in the original dataset. These variables mostly apply to gas emissions as well as population statistics. The summary statistics of these 18 variables can be seen in appendix 1 and appendix 2.

Analysis

Before any analysis could proceed, the dataset was organized more appropriately. Firstly, only the necessary variables of interest were extracted from the 77 in the original dataset. Secondly, the names of variables were hard to understand and therefore changed to be more meaningful. The code and output for these manipulations have been omitted here but have been included in appendix 1 and appendix 2 for reference.

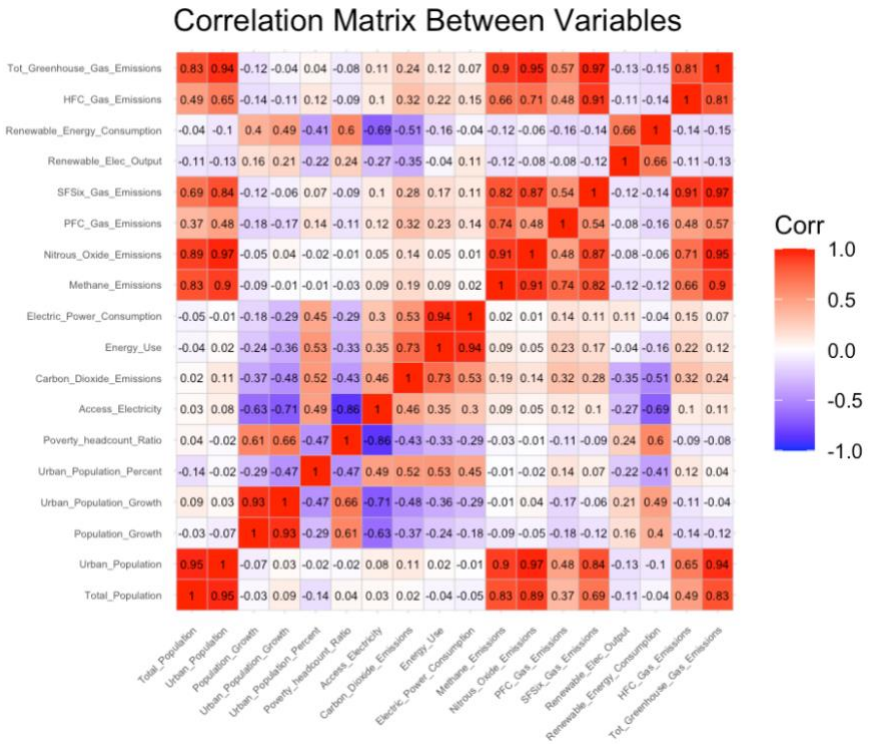
Figure 6 - Exploratory Analysis: Correlation Between Variables

```
wbcc4 <- na.omit(wbcc3)

library(ggcorrplot)

## Loading required package: ggplot2

corrwbcc4 <- cor(na.omit(wbcc4))
ggcorrplot(corrwbcc4, lab = TRUE, lab_size = 1.8, tl.cex = 4.4, title = "Correlation Matri
x Between Variables")
```



In figure 6, the construction and presentation of a correlation matrix is displayed. It shows the correlation amongst all variables with a lighter shade of colour leaning towards no correlation, blue leaning towards a strong negative correlation and red leaning towards a strong positive correlation. Naturally, each variable will have a perfect, positive relationship with itself. However, some other relatively strong positive relationships exist between Total Population and Methane Emissions along with Nitrous Oxide Emissions and Total Greenhouse Gas Emissions. Moreover, Total Greenhouse Gas Emissions appears to have a strong positive relationship with both Methane Emissions, SF6 Gas Emissions and Nitrous Oxide Emissions. This is understandable as all three contribute to Total Greenhouse Gas Emissions and so an increase in one would most definitively cause an increase in the other. There a few other

interesting correlations seen in this figure. Electric Power Consumption appears to have a strong positive relationship with Energy use. Urban Population appears to have a strong positive relationship with Methane Emissions, Nitrous Oxide Emissions and Total Greenhouse Gas Emissions but this can be expected as it shows a Pearson correlation coefficient of 0.95 with Total Population which also has a strong positive relationship with all three variables. There are not nearly as many strong negative relationships within this dataset. However, Access to Electricity appears to have a strong but negative relationship with the Poverty Headcount Ratio along with a much weaker but also negative relationship with Urban Population Growth. It is important to note that correlation by no means indicates causation and that no conclusions can be drawn from the matrix. However, this correlation matrix indicates that this analysis may encounter an issue of multicollinearity, especially in the case of regression modelling, which may need to be mitigated via the use of regularization methods.

```
wbcc4_hist <- hist(wbcc4$Tot_Greenhouse_Gas_Emissions, xlab = "Total Greenhouse Gas Emissi
ons", main = "Total Greenhouse Gas Emissions")
```

Figure 7 - Histogram Showing Distribution of Total Greenhouse Gas Emissions

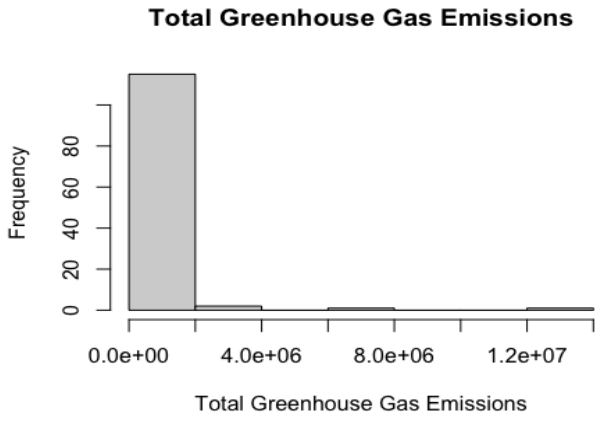


Figure 8 - Summary Statistics of Total Greenhouse Gas Emissions

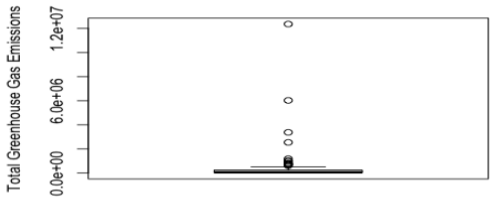
```
summary(wbcc4$Tot_Greenhouse_Gas_Emissions)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     2040    22475    62600   366342   249375 12355240
```

Figure 7 shows a histogram for Total Greenhouse Gas Emissions for all observations in the dataset, excluding those with NA values, as this will be the outcome variable of interest. As can be seen, this data is heavily skewed to the right with a large variation in this particular variable. Figure 8 shows the numbers behind this skew where it can be seen that the median is significantly smaller than the mean. This suggests that outliers may be present pulling the mean significantly higher than the median. This level of skew can have a negative impact on the standard error of many implemented machine learning models.

```
wbcc4_boxplot <- boxplot(wbcc4$Tot_Greenhouse_Gas_Emissions, ylab = "Total Greenhouse Gas
Emissions", data = wbcc4)
```

Figure 9 - Boxplot Showcasing All Observations (Excluding NA's)



The spread in the data of this particular variable can be visually seen by the boxplot in figure 9. Again a large amount of skew is observed. Additionally, a number of outliers can be seen which explains the large mean in comparison to the median. These large outliers essentially pull the mean higher whilst the median remains the same regardless.

```
wbcc5 <- subset(wbcc4, wbcc4$Tot_Greenhouse_Gas_Emissions > 22475 & wbcc4$Tot_Greenhouse_G
as_Emissions < 249375)

wbcc5_hist <- hist(wbcc5$Tot_Greenhouse_Gas_Emissions, xlab = "Total Greenhouse Gas Emissi
ons", main = "Total Greenhouse Gas Emissions (Middle 50%)")
```

Figure 10 - Total Greenhouse Gas Emissions (Middle 50% of Observations)

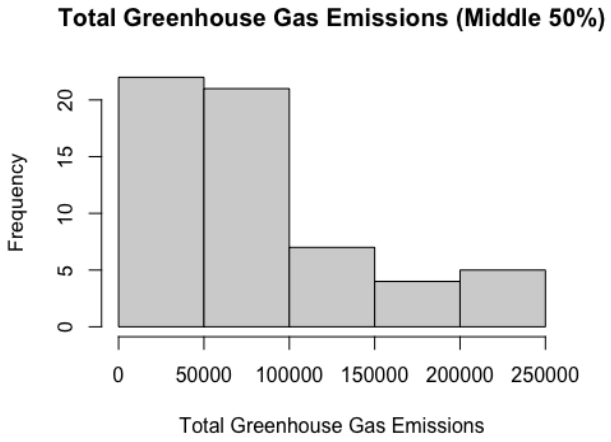


Figure 10 shows the distribution of data for the variable Total Greenhouse Gas Emissions for a subset of the original dataset. The middle 50% (interquartile range) of values for this particular variable were used to create a subset which is seen in figure 10. Consequently, much less skew in the data is observed. However, there is still some right skew which could still lead to some inaccuracies in the results of certain models namely, the residual errors of a regression model.

Figure 11 - Summary Statistics of Total Greenhouse Gas Emissions (Middle 50% of the Data)

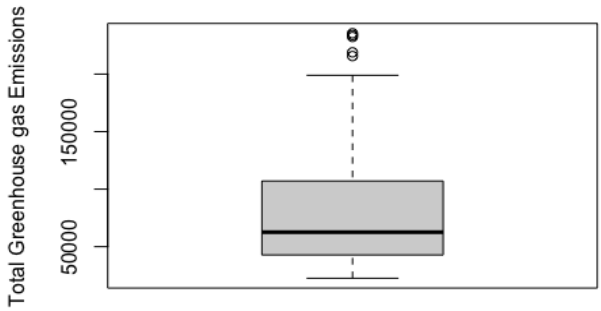
```
summary(wbcc5$Tot_Greenhouse_Gas_Emissions)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22550   42800   62600   84576  107065  235510
```

Figure 11 shows the new summary statistics for the subset middle 50% of the data. As seen, the mean and median are much closer together where a gap of only 21976 occurs, appose to the original where a difference of 12292640 occurred in Total Greenhouse Gas Emissions.

```
wbcc5_boxplot <- boxplot(wbcc5$Tot_Greenhouse_Gas_Emissions, ylab = "Total Greenhouse gas
Emissions", data = wbcc5)
```

Figure 12 - Boxplot for Total Greenhouse Gas Emissions (Middle 50% of Data)



According to the boxplot in figure 12, there is much less skew for the middle 50% of the original data. However, some outliers still exist and can therefore still expect some pull upwards in the mean and residual error of most models implemented. It is important to acknowledge that the sample size has decreased from 217 to 119 after NA's were removed and further to 59 when outliers and extreme values were removed in an effort to reduce the skew of the data. Observations are never removed without a valid reason to do so. In this scenario, some models could not be evaluated properly given such a large amount of skew and therefore justifies the exclusion of these more extreme values at the cost of sample size.

Multiple Linear Regression and Model Regularization/Variable Selection

Figure 13 - Multiple Regression Model Analysis

```
wbcc5_lm <- lm(wbcc5$Tot_Greenhouse_Gas_Emissions ~ ., data = wbcc5)
summary(wbcc5_lm)

##
## Call:
## lm(formula = wbcc5$Tot_Greenhouse_Gas_Emissions ~ ., data = wbcc5)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41750  -9851   2252  10421  71205
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -3.314e+04  4.529e+04  -0.732 0.468535
## Total_Population           -5.406e-04  4.135e-04  -1.307 0.198413
## Urban_Population            3.093e-03  8.184e-04   3.780 0.000501 ***
## Population_Growth           1.574e+04  1.188e+04   1.325 0.192547
## Urban_Population_Growth    -1.009e+04  9.516e+03  -1.060 0.295169
```

```
## Urban_Population_Percent     -3.483e+02  2.808e+02  -1.240 0.221876
## Poverty_headcount_Ratio      -6.635e+01  3.266e+02  -0.203 0.840015
## Access_Electricity            4.689e+02  3.565e+02   1.315 0.195691
## Carbon_Dioxide_Emissions      5.655e+03  3.316e+03   1.705 0.095696 .
## Energy_Use                    2.320e+01  9.291e+00   2.497 0.016640 *
## Electric_Power_Consumption   -9.561e+00  3.209e+00  -2.980 0.004831 **
## Methane_Emissions             1.326e+00  2.219e-01   5.973 4.73e-07 ***
## Nitrous_Oxide_Emissions       5.753e-01  4.518e-01   1.273 0.210020
## PFC_Gas_Emissions             6.755e+01  3.095e+01   2.183 0.034825 *
## SFSix_Gas_Emissions           5.071e+01  2.678e+01   1.894 0.065316 .
## Renewable_Elec_Output        -4.129e+01  1.746e+02  -0.237 0.814206
## Renewable_Energy_Consumption -4.901e+01  3.347e+02  -0.146 0.884284
## HFC_Gas_Emissions             1.632e-01  3.034e+00   0.054 0.957352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21430 on 41 degrees of freedom
## Multiple R-squared:  0.9079, Adjusted R-squared:  0.8698
## F-statistic: 23.79 on 17 and 41 DF,  p-value: 5.714e-16
```

Figure 13 shows the input and output of a multiple linear regression model on the dataset which included only the middle 50% of values for Total Greenhouse Gas Emissions. Linear regression models are sensitive to outliers and extreme values, therefore, to get an accurate estimate of which variables are statistically significant along with their coefficients, it is justifiable that this subset of the original data be used. According to the regression model labelled, "wbcc5_lm," there are a number of statistically significant variables which have an impact on Total Greenhouse Gas Emissions. Firstly, The Urban Population of a particular country appears to be statistically significant to the 0.1% level. The standard error of this estimated coefficient computed to be 0.0008184 which creates a 68% confidence interval of [0.0022746, 0.0039114] for the coefficient of Urban Population on Total Greenhouse Gas Emissions as 68% of observations lie within one standard deviation of the mean. The estimated coefficient indicates that an increase in the Urban Population by 1 person would yield an increase in Total Greenhouse Gas Emissions by 0.003093 Kt of CO2 equivalent in a given country.

Methane Emissions also appeared to be statistically significant to the 0.1% level in relation to Total Greenhouse Gas Emissions. The standard error of 0.2219 allowed a 68% confidence interval of [1.1041, 1.5479] to be created for this particular coefficient. The estimated coefficient of 1.326 for Methane Emissions indicates that 1 Kt of CO2 equivalent increase in Methane Emissions yields approximately a 1.326 Kt of CO2 equivalent increase in Total Greenhouse Gas Emissions.

Electric Power Consumption (EPC) produced a statistically significant result to the 1% level. This particular variable produced an estimated coefficient of -9.561 with a standard error of 3.209. Thus, a 68% confidence interval for the coefficient of EPC equates to [-12.77, -6.352]. Ultimately, an increase in EPC by 1 kWh per capita yields a decrease in TGGE by approximately 9.561 Kt of CO2 equivalent.

In addition, PFC Gas Emissions (PFCGE) had a statistically significant relationship with TGGE to the 5% level. Producing an estimated coefficient of 67.55 with a standard error 30.95 allows a 68% confidence interval of [36.6, 98.5] to be created. Essentially, an increase in PFCGE by 1 Kt of CO2 Equivalent yields and increase in TGGE by approximately 67.55 Kt of CO2 equivalent.

Moreover, Energy Use (EU) appeared to have a positive relationship with TGGE, also to the 5% level. With a computed estimated coefficient of 23.20 and a standard error of 9.291, a 68% confidence interval can be determined as, [13.909, 32.491]. These results indicate that an increase in EU by 1 kg of oil equivalent per capita would yield an estimated increase in TGGE by 23.20 Kt of CO2 equivalent.

Both SF6 Gas Emissions (SF6G) and Carbon Dioxide Gas Emission (CO2) proved to be statistically significant to the 10% level with coefficient estimates of 50.71 and 565.5 respectively. According to their standard errors, these estimates generated 68% confidence intervals of [23.93, 77.49] for SF6G and [233.9, 897.1] for CO2. Therefore, an

increase in 1 Kt of CO2 equivalent for SF6G and CO2 would yield an increase in TGGE by 50.71 and 565.5 Kt of CO2 equivalent respectively.

The regression model as whole produced an adjusted R squared value of 0.8698 indicating that 86.98% of the variation was explained by the linear model. However, the residual standard error of 21430 appears relatively high. Given values for each variable in the model and given their coefficients. Given the residual standard error, the predicted value for TGGE will be no further than 21430 Kt of CO2 equivalent from the actual amount 68% of the time. This could be attributed to the right skew in the data even after reducing the observations to only include the interquartile range for TGGE. With the data skewed right, it is expected that a linear regression model would yield a higher residual error than expected for both the coefficients and the model as a whole. Also, it is important to acknowledge that this is a very high dimensional dataset after removal of many observations. There are 18 variables with only 59 observations while a commonly recommend ratio is 10 observations to each variable which would equate to 180 observations in this scenario. Therefore, some regularization or variable selection methods would need to be performed.

```
plot(wbcc5_lm, which = 1)
```

Figure 14 - Checking for Linearity and Constant Variance

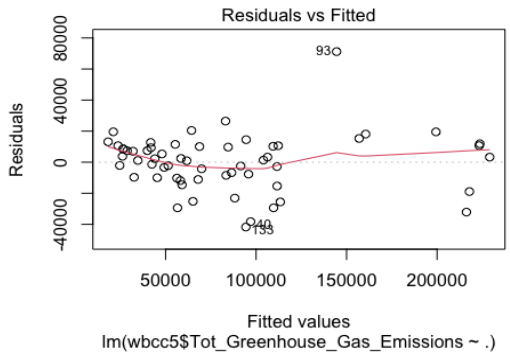


Figure 14 displays a residuals vs fitted plot to check the assumptions of linearity and constant variance. As seen, the red line approximately follows the dotted line with only some minor deviation. This minor deviation could be attributed to observation 40, 133 and 93, all of which are deemed outliers. The assumption of constant variance looks to be satisfied as observations appear to deviate from the dotted line by a consistent amount as the fitted values increases. Observations labelled as outliers will not be removed from this set of data as the sample size is already relatively small in comparison to the number of variables.

```
library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.0-2

wbcc5_matrix <- model.matrix(wbcc5$Tot_Greenhouse_Gas_Emissions ~ ., data = wbcc5)
wbcc5_st <- scale(wbcc5)
X <- wbcc5_matrix[,-1]
Y <- wbcc5$Tot_Greenhouse_Gas_Emissions
wbcc5_lasso = glmnet(X, Y, alpha = 1)
plot(wbcc5_lasso, label = TRUE)
```

Figure 15 - Lasso Regression Showcasing Shrinkage of Variables Towards Zero

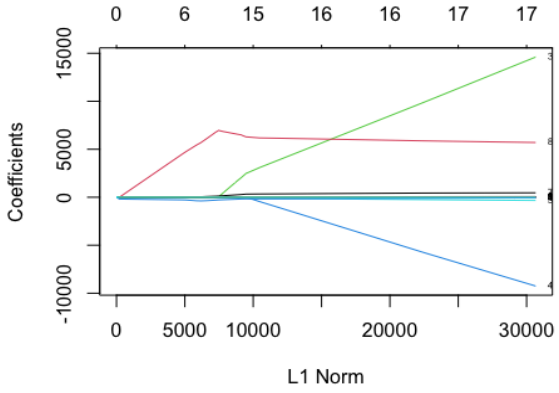
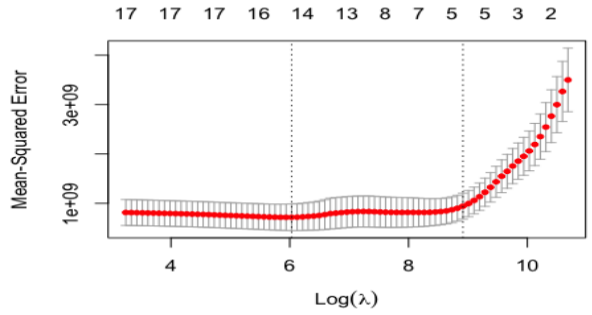


Figure 15 displays a visual representation of the Lasso Regression model. A Lasso regression model is used to shrink variables towards zero while only leaving some. Therefore, it works as variable shrinkage and variable selection. The graph in Figure 15 reads right to left and shows each variable as it shrinks towards zero as the L1 Norm decreases towards zero.

```
cv_lasso = cv.glmnet(X, Y, alpha = 1)
plot(cv_lasso)
```

Figure 16 - CV Lasso Regression Plot to Show Number of Variables Recommended



The plot shown in figure 16 summarizes the recommended log(lambda) value along with the recommended number of variables after the Lasso Regression model has been implemented. According to this graph, 5 variables are recommended to achieve the lowest mean-squared error whilst limiting the number of variables used in the model. After five variables have been used, the mean square error does not decrease enough to justify increasing the dimensionality.

Figure 17 - Coefficients Recommended by the Lasso

```
coef(cv_lasso)
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                                          1
## (Intercept)                   3.175382e+04
## Total_Population              .
## Urban_Population              1.489306e-03
## Population_Growth             .
## Urban_Population_Growth       .
## Urban_Population_Percent      .
## Poverty_headcount_Ratio       .
## Access_Electricity            .
## Carbon_Dioxide_Emissions      3.152445e+03
## Energy_Use                    .
## Electric_Power_Consumption    .
## Methane_Emissions             1.117019e+00
## Nitrous_Oxide_Emissions       .
## PFC_Gas_Emissions             .
## SFSix_Gas_Emissions           3.829367e+01
## Renewable_Elec_Output         .
## Renewable_Energy_Consumption -2.459379e+02
## HFC_Gas_Emissions             .
```

Figure 17 shows a summary of which variable coefficients have shrunk to zero and which have shrunk towards zero but still remain in the model. As seen, there is a slight difference in this result compared to the original linear regression model. Energy Use and PFC Gas Emissions have been dropped from the model whilst Renewable Energy Consumption has being added.

Figure 18 - Multiple Linear Regression Model with Selected Fewer Variables

```
wbcc5_lm_lasso <- lm(wbcc5$Tot_Greenhouse_Gas_Emissions ~ wbcc5$Urban_Population + wbcc5$C
arbon_Dioxide_Emissions + wbcc5$Methane_Emissions + wbcc5$SFSix_Gas_Emissions + wbcc5$Rene
wable_Energy_Consumption, data = wbcc5)
summary(wbcc5_lm_lasso)
```

```
##
## Call:
## lm(formula = wbcc5$Tot_Greenhouse_Gas_Emissions ~ wbcc5$Urban_Population +
##     wbcc5$Carbon_Dioxide_Emissions + wbcc5$Methane_Emissions +
##     wbcc5$SFSix_Gas_Emissions + wbcc5$Renewable_Energy_Consumption,
##     data = wbcc5)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51225 -12299   -569   8143  89034
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        4.650e+03  1.181e+04   0.394   0.6954
## wbcc5$Urban_Population             1.970e-03  3.740e-04   5.267 2.58e-06 ***
## wbcc5$Carbon_Dioxide_Emissions     7.344e+03  1.673e+03   4.389 5.46e-05 ***
## wbcc5$Methane_Emissions            1.448e+00  1.720e-01   8.418 2.42e-11 ***
## wbcc5$SFSix_Gas_Emissions          5.170e+01  2.648e+01   1.953   0.0562 .
```

```
## wbcc5$Renewable_Energy_Consumption -3.252e+02  1.445e+02  -2.251   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24030 on 53 degrees of freedom
## Multiple R-squared:  0.8504, Adjusted R-squared:  0.8363
## F-statistic: 60.25 on 5 and 53 DF,  p-value: < 2.2e-16
```

Figure 18 shows the multiple linear regression model along with its output for those variables selected by the Lasso regression model previously. The most noticeable difference is the change in significance of these variables. CO2 has experienced an increase in statistical significance whilst renewable energy consumption has achieved statistical significance to the 5% level. The other three variables remained the same. However, there is a noticeable change in the coefficients. CO2 appears to have a more impactful effect on TGGE in this particular model compared to Urban Population which decreased. Methane Emissions and FS6G both increased ever so slightly with regards to their effect on the response variable.

Figure 19 - Creating a New Subset with Top 25% of Values for Total Greenhouse Gas Emissions

```
wbcc6 <- subset(wbcc4, wbcc4$Tot_Greenhouse_Gas_Emissions > 249375)
summary(wbcc6$Tot_Greenhouse_Gas_Emissions)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   263240   368370   477300  1273897   822732 12355240
```

Figure 19 shows the manipulation of the original dataset to create another subset of those observations in the top 25% for Total Greenhouse gas Emissions. Looking at the summary statistics, there still appears to be a very large skew towards the right as the mean is again much higher than the median. Ultimately, this casts doubt over the validity of using these observations for any further analysis.

```
wbcc6_boxplot <- boxplot(wbcc6$Tot_Greenhouse_Gas_Emissions, ylab = "Total Greenhouse Gas
Emissions", data = wbcc6)
```

Figure 20 - Boxplot for Top 25% Observations for Total Greenhouse Gas Emissions

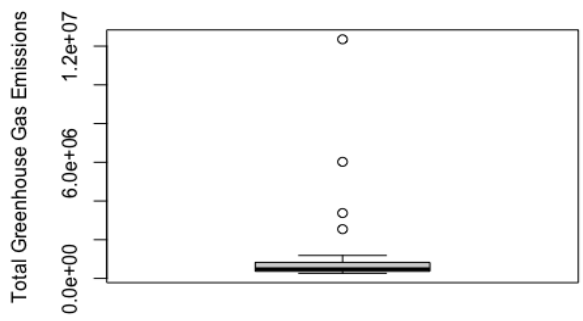


Figure 20 displays a boxplot for those observations within the highest 25% of Total Greenhouse Gas Emissions. Clearly, this data is still largely skewed to the right with a few outliers. Therefore, there is no use in attempting to perform a regression analysis as such high and extreme values will yield obscure residual errors. In addition, this particular subset only has 30 observations which is not ideal in a regression model with 18 variables.

Regression Tree Analysis

```
library(tree)
wbcc5_tree <- tree(formula = wbcc5$Tot_Greenhouse_Gas_Emissions ~ ., data = wbcc5)
plot(wbcc5_tree)
text(wbcc5_tree, cex = 0.5)
```

Figure 21 - Regression Tree for Total Greenhouse Gas Emissions (Middle 50% of Observations)

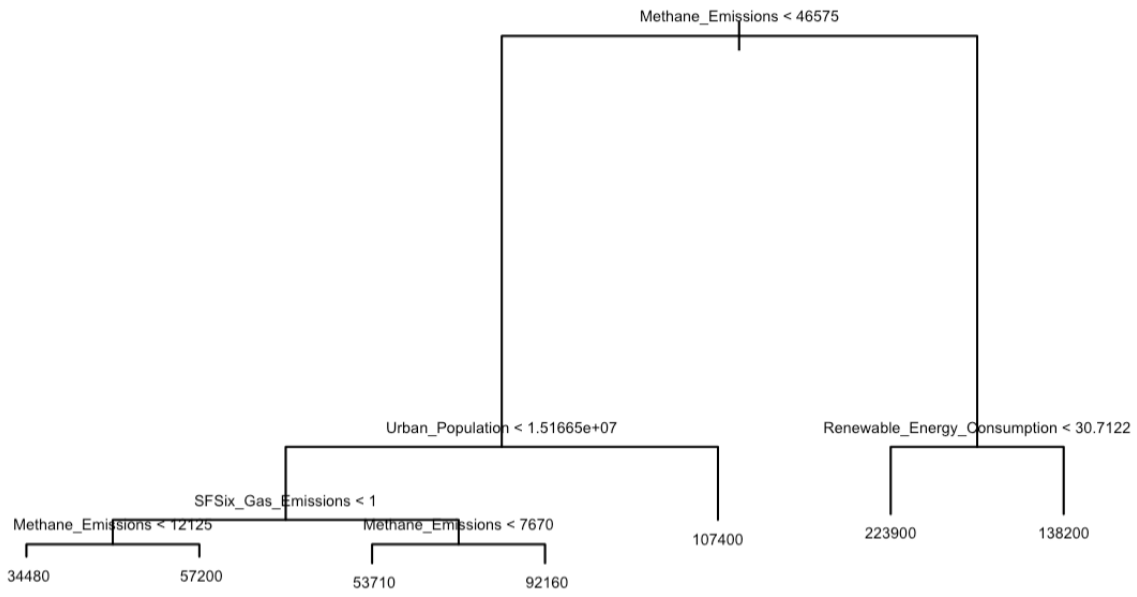


Figure 21 displays a regression tree diagram with Total Greenhouse Gas Emissions set as the response variable and for the subset containing the middle 50% of observations. The regression tree consists of four different components; the root node, internal nodes, terminal nodes and branches that connect these three. The variables selected at each node are those that minimize the total residual sum of squares in the resulting child nodes. The terminal nodes or leaf nodes are the average of all observations that fall into that particular node. Therefore, looking at the regression tree one can categorize a particular observation into one of the seven categories of Total Greenhouse Gas Emissions. In the creation of this regression tree, four different variables were used. Methane Gas Emissions, Urban Population, Renewable Energy Consumption and SF6 gas Emissions were used to create the nodes, all of which were considered important and significant in the Lasso regression model.

Therefore, to predict the estimated TGGE of a new observation, it is trivial to follow the rules in the regression tree. More importantly, for a new observations TGGE to be predicted it only requires a value for these four variables. This highlights an important component of regression and classifications trees in that they perform variable selection embedded within the algorithm.

Figure 22 - Decision Rules for Created Regression Tree

```
wbcc5_tree

## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 59 2.046e+11  84580
##    2) Methane_Emissions < 46575 49 6.764e+10  64880
##      4) Urban_Population < 1.51665e+07 40 2.319e+10  55320
##        8) SFSix_Gas_Emissions < 1 26 7.672e+09  45840
##         16) Methane_Emissions < 12125 13 8.124e+08  34480 *
##         17) Methane_Emissions > 12125 13 3.504e+09  57200 *
##        9) SFSix_Gas_Emissions > 1 14 8.843e+09  72930
##         18) Methane_Emissions < 7670 7 7.204e+08  53710 *
##         19) Methane_Emissions > 7670 7 2.949e+09  92160 *
##      5) Urban_Population > 1.51665e+07 9 2.453e+10 107400 *
##    3) Methane_Emissions > 46575 10 2.483e+10 181100
##      6) Renewable_Energy_Consumption < 30.7122 5 9.528e+08 223900 *
##      7) Renewable_Energy_Consumption > 30.7122 5 5.502e+09 138200 *
```

Figure 22 lays out the decision rules at each node starting from the root to the leaves. The staggering of the numbers indicates how high or low a particular node is on the tree. This also gives an indication as to how important each of these variables are to the Total Greenhouse Gas Emissions. Therefore, Methane Emissions appears to be very important to determining how much TGGE a country will emit. The Urban Population of a particular country also appears to be of high importance. Therefore, a country with Methane emissions less than 46575 Kt of CO2 equivalent but an Urban Population of more than 15166500 is estimated to produce 107400 Kt of CO2 equivalent in Total Greenhouse Gas Emissions.

Random Forest

Figure 23 - Random Forest Algorithm

```
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

wbcc5_RF <- randomForest(wbcc5$Tot_Greenhouse_Gas_Emissions ~ ., data = wbcc5)
wbcc5_RF

##
## Call:
##  randomForest(formula = wbcc5$Tot_Greenhouse_Gas_Emissions ~ .,      data = wbcc5)
##                Type of random forest: regression
```

```
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 1154835384
##                     % Var explained: 66.69
```

Figure 23 shows the input and output of a random forest algorithm. Essentially, 500 regression trees were created and the average was taken. The model shows that 66.69% of the variation in the data was explained by the random forest.

```
varImpPlot(wbcc5_RF, main = "Importance of Variables")
```

Figure 24 - Variable Importance

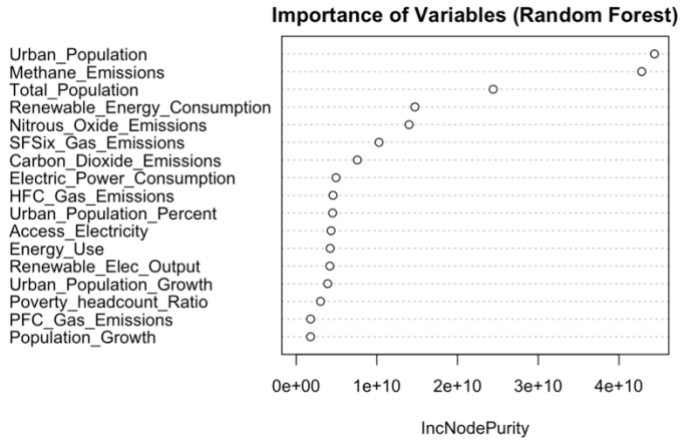


Figure 24 displays a simple graph which shows the importance of each variable as predicted by the random forest algorithm. The importance of each variable has been computed by Node Purity. Again, Urban Population and Methane Emissions appear to be the most important variables in this model. However, they are followed by Total Population as an important predictor for Total Greenhouse Gas Emissions. Renewable Energy Consumption and Nitrous Oxide Emissions are the last two "important" variables before the remainder appear relatively close, as per the random forest algorithm.

Classification Tree Analysis

Figure 25 - Creating a New Categorical Variable Based on Total Greenhouse Gas Emissions

```
quantile(wbcc4$Tot_Greenhouse_Gas_Emissions, prob = seq(0, 1, length = 11), type = 5)

##       0%      10%      20%      30%      40%      50%      60%      70%
##     2040    10336    18043    34488    46050    62600    93766   195996
##      80%      90%     100%
##   328536   654080 12355240

summary(wbcc4$Greenhouse_Emissions_Cat)

##  00-10%  10-20%  20-30%  30-40%  40-50%  50-60%  60-70%  70-80%  80-90% 90-100%
##      12      12      12      12      11      12      12      12      12      12
```

Ultimately, it appears that for most countries, reduction of gas emissions has become a competition. With this, the full dataset (excluding NA's values) has been split into groups categorized by percentiles of Total Greenhouse Gas Emissions being emitted. The summary displays how many observations were placed in each category, with the total number of observations equal to 59. The code used to create this categorical variable has been included in appendix 4.

```
wbcc7 <- wbcc4[, -18]
wbcc7_tree_class <- tree(wbcc7$Greenhouse_Emissions_Cat ~., data = wbcc7)
plot(wbcc7_tree_class)
text(wbcc7_tree_class, cex=0.35)
```

Figure 26 - Classification Tree with % Group As the Response

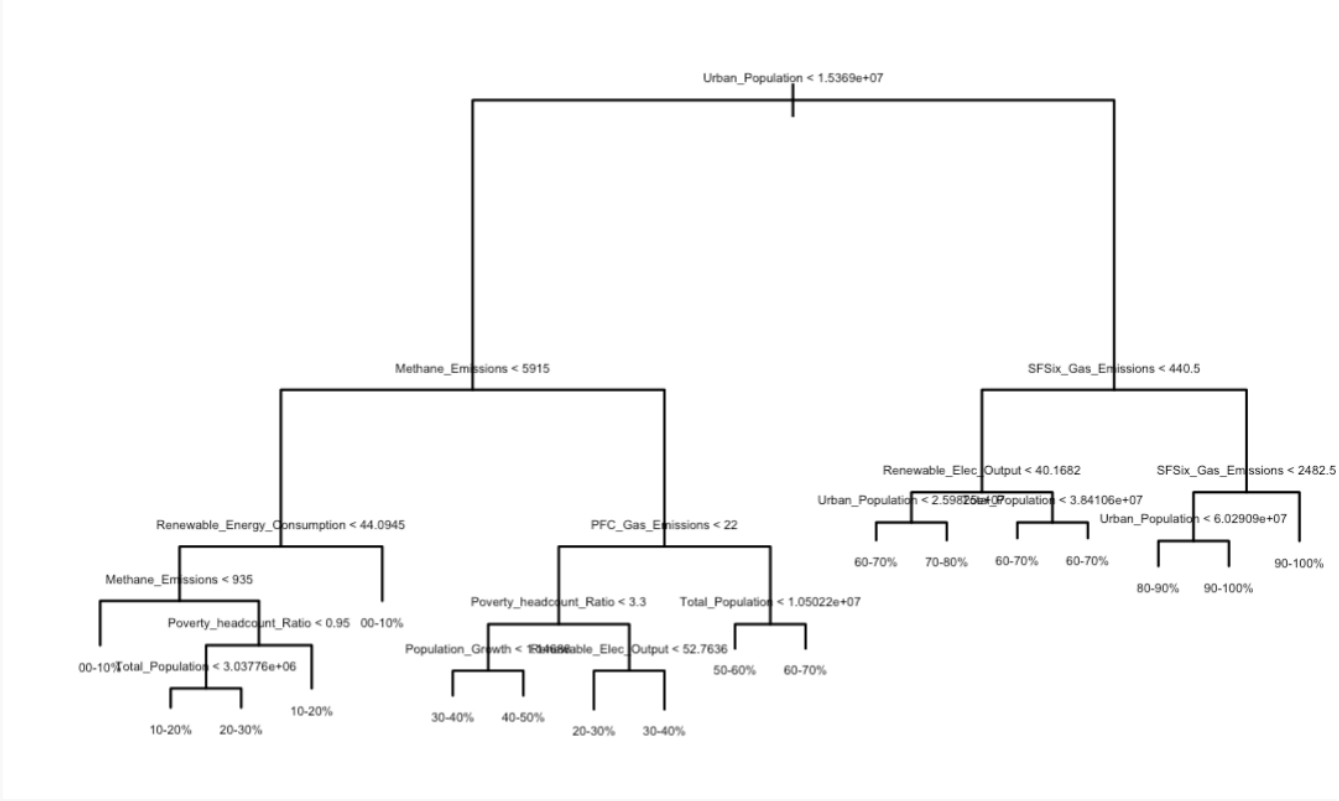


Figure 26 displays a classification tree with the percent grouping for Total Greenhouse Gas Emissions as the response variable in the leaf nodes. Unlike the previously created regression tree, this classification tree is assigning observations to a particular class and includes all observations from the dataset (excluding those with NA's). Nine variables were used in the construction of this tree which yielded 18 terminal nodes. Overall, the misclassification rate of this particular tree came to 26.05% .(Appendix 5). To reduce this misclassification rate, the tree created would likely need to be pruned. Such pruning would lead to a higher classification accuracy but increased bias.

Naïve Bayes

Figure 27 - Naive Bayes Model

```
library(naivebayes)

## naivebayes 0.9.7 loaded

##
## Attaching package: 'naivebayes'

## The following object is masked from 'package:data.table':
##
##     tables

no_obs <- dim(wbcc4)[1]
accuracy <- rep(0,10)
for(i in 1:10){
wbcc4_test_index <- sample(no_obs, size = as.integer(no_obs*0.10), replace = FALSE)
wbcc4_training_index <- -wbcc4_test_index
wbcc4_NaiveBayes <- naive_bayes(Greenhouse_Emissions_Cat ~., data = wbcc4[wbcc4_training_i
ndex,])
wbcc4_pred_class <- predict(wbcc4_NaiveBayes, newdata = wbcc4[wbcc4_test_index, -19], type
= "class")
cont_table <- table(wbcc4_pred_class, wbcc4$Greenhouse_Emissions_Cat[wbcc4_test_index])
accuracy[i] <- sum(diag(cont_table))/sum(cont_table)
}
accuracy

##  [1] 0.2727273 0.4545455 0.5454545 0.2727273 0.2727273 0.6363636 0.6363636
##  [8] 0.1818182 0.4545455 0.6363636

mean(accuracy)

## [1] 0.4363636
```

Figure 27 shows the implementation of the naïve bayes function on the full dataset with only the NA's removed. The rationale behind implementing this model is to determine the accuracy to which an observation can be assigned into one of the 10 categories. The dataset is split into 10% for testing and the remaining 90% for training. To do this, a loop was run 10 times and the accuracy of each iteration was stored inside a vector, "accuracy." The mean of these accuracies came to 43.64% with the highest recorded being 63.64% and lowest 18.19%. This is clearly a very large variation and an accuracy of 43.64% is not sufficient to use in practice.

Conclusion

There are a number of conclusions that can be drawn from this in depth analysis of Greenhouse Gas Emissions. It is inevitable that certain gas emissions would prove to have a statistically significant impact on the amount of greenhouse gas emitted from a particular country given the latter is partly derived by the former. However, it is not as obvious that population, energy/electrical use or renewable energy use would prove to also have a significant impact.. After careful comparison of various statistical learning methods, this investigation found Urban Population, Methane Gas Emissions, SF6 Gas Emissions and Carbon Dioxide Gas Emissions all to aid in increasing the total greenhouse gases emitted into our earths atmosphere. The use of renewable energy proved to help lower the Total Greenhouse Gases emitted by enough to make it significant. Given this, in preparation for the up coming Glasgow Conference on climate change, world leaders should be focusing their attention on reducing the amount of Methane, SF6 and Carbon Dioxide the country emits whilst increasing the use of renewable energy sources. Carbon Dioxide and SF6 yield a much larger increase in Total Greenhouse Gas Emissions per unit and therefore with an emphasis on these two gases, countries will get a better "bang for their buck."

Contrary, these conclusions are made lightly. The strongest conclusion from this analysis to be made is that there is a greater need for more accurate, timely and precise measurement of these variables by all countries, especially on an issue as serious as climate change. A dataset that started with 217 observations could only be analyzed using a small portion of 59 observations. Consequently, the error associated with many predictions increases drastically. If each country were to measure and record these variables, an analysis such as this could draw much stronger and more confident conclusions. However, methods could have been used that work with high dimensional datasets more effectively such as cluster or principle component analysis. Lastly, the data used for this analysis were measurements taken a various points in time over 20 years. Therefore, further projects could take a dive into one specific country and compare the change in these measurements over time.

References

Ritchie, H, Roser, M. (2020). *CO2 and Greenhouse Gas Emissions.* Retrieved 24th October 2021 from https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,

Group, World Bank. (2021). *World Development Indicators.* Retrieved October 19, 2021 from https://databank.worldbank.org/source/world-development-indicators

Appendices

Appendix 1 - Initial Summary Statistics

```
setwd("~/Documents/UniversityofNewcastle/STAT6020 - 2021/Assignments")
wbcc <- read.csv("wbcc_bc.csv", header = TRUE)

wbcc2 <- wbcc[, c(2,76, 78, 75, 77, 79, 74, 17, 37, 29, 28, 47, 49, 51, 52, 23, 26, 46, 44
)]
summary(wbcc2)
```

```
##    country           SP.POP.TOTL         SP.URB.TOTL         SP.POP.GROW
##  Length:217         Min.   :1.083e+04   Min.   :     5498   Min.   :-1.7206
##  Class :character   1st Qu.:7.866e+05   1st Qu.:   458630   1st Qu.: 0.3236
##  Mode  :character   Median :6.625e+06   Median :  3899416   Median : 1.0576
##                     Mean   :3.562e+07   Mean   : 20154875   Mean   : 1.1438
##                     3rd Qu.:2.569e+07   3rd Qu.: 11327426   3rd Qu.: 1.9202
##                     Max.   :1.402e+09   Max.   :861289359   Max.   : 4.1242
##                                         NA's   :2
##   SP.URB.GROW      SP.URB.TOTL.IN.ZS  SI.POV.DDAY    EG.ELC.ACCS.ZS
##  Min.   :-1.5938   Min.   : 13.35    Min.   : 0.00   Min.   :  6.721
##  1st Qu.: 0.6897   1st Qu.: 42.79    1st Qu.: 0.30   1st Qu.: 84.762
##  Median : 1.5985   Median : 62.38    Median : 1.70   Median :100.000
##  Mean   : 1.7766   Mean   : 61.39    Mean   :13.78   Mean   : 86.470
##  3rd Qu.: 2.8528   3rd Qu.: 80.98    3rd Qu.:19.00   3rd Qu.:100.000
##  Max.   : 5.6748   Max.   :100.00    Max.   :78.80   Max.   :100.000
##  NA's   :2         NA's   :2         NA's   :55      NA's   :1
##  EN.ATM.CO2E.PC     EG.USE.PCAP.KG.OE   EG.USE.ELEC.KH.PC  EN.ATM.METH.KT.CE
##  Min.   : 0.02617   Min.   :    9.563   Min.   :   39.06   Min.   :      0
##  1st Qu.: 0.75731   1st Qu.:  569.577   1st Qu.:  912.60   1st Qu.:   2390
##  Median : 2.54060   Median : 1233.319   Median : 2604.41   Median :   9200
##  Mean   : 4.18922   Mean   : 2244.358   Mean   : 4245.86   Mean   :  42798
##  3rd Qu.: 5.91693   3rd Qu.: 2712.758   3rd Qu.: 5539.15   3rd Qu.:  33710
##  Max.   :32.41564   Max.   :17922.704   Max.   :53832.48   Max.   :1238630
##  NA's   :26         NA's   :45          NA's   :75         NA's   :26
##  EN.ATM.NOXE.KT.CE EN.ATM.PFCG.KT.CE EN.ATM.SF6G.KT.CE EG.ELC.RNEW.ZS
##  Min.   :     0    Min.   :    0.0   Min.   :    0     Min.   :  0.000
##  1st Qu.:   765    1st Qu.:    0.0   1st Qu.:    0     1st Qu.:  1.601
##  Median :  3430    Median :    0.0   Median :    0     Median : 16.176
##  Mean   : 15625    Mean   :  511.5   Mean   : 1186     Mean   : 30.466
##  3rd Qu.: 12180    3rd Qu.:  134.0   3rd Qu.:  366     3rd Qu.: 52.803
##  Max.   :538790    Max.   :20578.0   Max.   :57054     Max.   :100.000
##  NA's   :26        NA's   :80        NA's   :80
##  EG.FEC.RNEW.ZS   EN.ATM.HFCG.KT.CE EN.ATM.GHGT.KT.CE
##  Min.   : 0.000   Min.   :     0    Min.   :      30
##  1st Qu.: 5.859   1st Qu.:     0    1st Qu.:    9280
##  Median :20.864   Median :   105    Median :   39060
##  Mean   :28.839   Mean   :  5663    Mean   :  240177
##  3rd Qu.:43.972   3rd Qu.:  1203    3rd Qu.:  107065
```

```
##  Max.   :96.384   Max.   :300896    Max.   :12355240
##  NA's   :4        NA's   :80        NA's   :26
```

Appendix 2 - Variables Names Changed

```
library(data.table)
setnames(wbcc2, "SP.POP.TOTL", "Total_Population")
setnames(wbcc2, "SP.URB.TOTL", "Urban_Population")
setnames(wbcc2, "SP.POP.GROW", "Population_Growth")
setnames(wbcc2, "SP.URB.GROW", "Urban_Population_Growth")
setnames(wbcc2, "SP.URB.TOTL.IN.ZS", "Urban_Population_Percent")
setnames(wbcc2, "EG.ELC.ACCS.ZS", "Access_Electricity")
setnames(wbcc2, "EN.ATM.CO2E.PC", "Carbon_Dioxide_Emissions")
setnames(wbcc2, "EG.USE.PCAP.KG.OE", "Energy_Use")
setnames(wbcc2, "EG.USE.ELEC.KH.PC", "Electric_Power_Consumption")
setnames(wbcc2, "EN.ATM.GHGT.KT.CE", "Tot_Greenhouse_Gas_Emissions")
setnames(wbcc2, "EN.ATM.METH.KT.CE", "Methane_Emissions")
setnames(wbcc2, "EN.ATM.NOXE.KT.CE", "Nitrous_Oxide_Emissions")
setnames(wbcc2, "EN.ATM.PFCG.KT.CE", "PFC_Gas_Emissions")
setnames(wbcc2, "EN.ATM.SF6G.KT.CE", "SFSix_Gas_Emissions")
setnames(wbcc2, "EN.ATM.HFCG.KT.CE", "HFC_Gas_Emissions")
setnames(wbcc2, "EG.ELC.RNEW.ZS", "Renewable_Elec_Output")
setnames(wbcc2, "EG.FEC.RNEW.ZS", "Renewable_Energy_Consumption")
setnames(wbcc2, "SI.POV.DDAY", "Poverty_headcount_Ratio")
summary(wbcc2)
```

```
##    country          Total_Population    Urban_Population    Population_Growth
##  Length:217         Min.   :1.083e+04   Min.   :     5498   Min.   :-1.7206
##  Class :charactzer   1st Qu.:7.866e+05   1st Qu.:   458630   1st Qu.: 0.3236
##  Mode  :character   Median :6.625e+06   Median :  3899416   Median : 1.0576
##                     Mean   :3.562e+07   Mean   : 20154875   Mean   : 1.1438
##                     3rd Qu.:2.569e+07   3rd Qu.: 11327426   3rd Qu.: 1.9202
##                     Max.   :1.402e+09   Max.   :861289359   Max.   : 4.1242
##                                         NA's   :2
##  Urban_Population_Growth Urban_Population_Percent Poverty_headcount_Ratio
##  Min.   :-1.5938         Min.   : 13.35           Min.   : 0.00
##  1st Qu.: 0.6897         1st Qu.: 42.79           1st Qu.: 0.30
##  Median : 1.5985         Median : 62.38           Median : 1.70
##  Mean   : 1.7766         Mean   : 61.39           Mean   :13.78
##  3rd Qu.: 2.8528         3rd Qu.: 80.98           3rd Qu.:19.00
##  Max.   : 5.6748         Max.   :100.00           Max.   :78.80
##  NA's   :2               NA's   :2                NA's   :55
##  Access_Electricity Carbon_Dioxide_Emissions   Energy_Use
##  Min.   :  6.721    Min.   : 0.02617         Min.   :    9.563
##  1st Qu.: 84.762    1st Qu.: 0.75731         1st Qu.:  569.577
##  Median :100.000    Median : 2.54060         Median : 1233.319
##  Mean   : 86.470    Mean   : 4.18922         Mean   : 2244.358
##  3rd Qu.:100.000    3rd Qu.: 5.91693         3rd Qu.: 2712.758
##  Max.   :100.000    Max.   :32.41564         Max.   :17922.704
##  NA's   :1          NA's   :26               NA's   :45
##  Electric_Power_Consumption Methane_Emissions Nitrous_Oxide_Emissions
##  Min.   :   39.06           Min.   :      0   Min.   :     0
##  1st Qu.:  912.60           1st Qu.:   2390   1st Qu.:   765
##  Median : 2604.41           Median :   9200   Median :  3430
```

```
##  Mean   : 4245.86           Mean   :  42798   Mean   : 15625
##  3rd Qu.: 5539.15           3rd Qu.:  33710   3rd Qu.: 12180
##  Max.   :53832.48           Max.   :1238630   Max.   :538790
##  NA's   :75                 NA's   :26        NA's   :26
##  PFC_Gas_Emissions SFSix_Gas_Emissions Renewable_Elec_Output
##  Min.   :    0.0   Min.   :    0       Min.   :  0.000
##  1st Qu.:    0.0   1st Qu.:    0       1st Qu.:  1.601
##  Median :    0.0   Median :    0       Median : 16.176
##  Mean   :  511.5   Mean   : 1186       Mean   : 30.466
##  3rd Qu.:  134.0   3rd Qu.:  366       3rd Qu.: 52.803
##  Max.   :20578.0   Max.   :57054       Max.   :100.000
##  NA's   :80        NA's   :80
##  Renewable_Energy_Consumption HFC_Gas_Emissions Tot_Greenhouse_Gas_Emissions
##  Min.   : 0.000               Min.   :     0    Min.   :      30
##  1st Qu.: 5.859               1st Qu.:     0    1st Qu.:    9280
##  Median :20.864               Median :   105    Median :   39060
##  Mean   :28.839               Mean   :  5663    Mean   :  240177
##  3rd Qu.:43.972               3rd Qu.:  1203    3rd Qu.:  107065
##  Max.   :96.384               Max.   :300896    Max.   :12355240
##  NA's   :4                    NA's   :80        NA's   :26
```

Appendix 3 - Removed Country from Data for Purpose of Analysis

```
wbcc3 <- wbcc2[, -1]
summary(wbcc3)

##  Total_Population    Urban_Population    Population_Growth
##  Min.   :1.083e+04   Min.   :     5498   Min.   :-1.7206
##  1st Qu.:7.866e+05   1st Qu.:   458630   1st Qu.: 0.3236
##  Median :6.625e+06   Median :  3899416   Median : 1.0576
##  Mean   :3.562e+07   Mean   : 20154875   Mean   : 1.1438
##  3rd Qu.:2.569e+07   3rd Qu.: 11327426   3rd Qu.: 1.9202
##  Max.   :1.402e+09   Max.   :861289359   Max.   : 4.1242
##                      NA's   :2
##  Urban_Population_Growth Urban_Population_Percent Poverty_headcount_Ratio
##  Min.   :-1.5938         Min.   : 13.35           Min.   : 0.00
##  1st Qu.: 0.6897         1st Qu.: 42.79           1st Qu.: 0.30
##  Median : 1.5985         Median : 62.38           Median : 1.70
##  Mean   : 1.7766         Mean   : 61.39           Mean   :13.78
##  3rd Qu.: 2.8528         3rd Qu.: 80.98           3rd Qu.:19.00
##  Max.   : 5.6748         Max.   :100.00           Max.   :78.80
##  NA's   :2               NA's   :2                NA's   :55
##  Access_Electricity Carbon_Dioxide_Emissions   Energy_Use
##  Min.   :  6.721    Min.   : 0.02617         Min.   :    9.563
##  1st Qu.: 84.762    1st Qu.: 0.75731         1st Qu.:  569.577
##  Median :100.000    Median : 2.54060         Median : 1233.319
##  Mean   : 86.470    Mean   : 4.18922         Mean   : 2244.358
##  3rd Qu.:100.000    3rd Qu.: 5.91693         3rd Qu.: 2712.758
##  Max.   :100.000    Max.   :32.41564         Max.   :17922.704
##  NA's   :1          NA's   :26               NA's   :45
```

```
##  Electric_Power_Consumption Methane_Emissions Nitrous_Oxide_Emissions
##  Min.   :   39.06           Min.   :      0   Min.   :     0
##  1st Qu.:  912.60           1st Qu.:   2390   1st Qu.:   765
##  Median : 2604.41           Median :   9200   Median :  3430
##  Mean   : 4245.86           Mean   :  42798   Mean   : 15625
##  3rd Qu.: 5539.15           3rd Qu.:  33710   3rd Qu.: 12180
##  Max.   :53832.48           Max.   :1238630   Max.   :538790
##  NA's   :75                 NA's   :26        NA's   :26
##  PFC_Gas_Emissions SFSix_Gas_Emissions Renewable_Elec_Output
##  Min.   :    0.0   Min.   :    0       Min.   :  0.000
##  1st Qu.:    0.0   1st Qu.:    0       1st Qu.:  1.601
##  Median :    0.0   Median :    0       Median : 16.176
##  Mean   :  511.5   Mean   : 1186       Mean   : 30.466
##  3rd Qu.:  134.0   3rd Qu.:  366       3rd Qu.: 52.803
##  Max.   :20578.0   Max.   :57054       Max.   :100.000
##  NA's   :80        NA's   :80
##  Renewable_Energy_Consumption HFC_Gas_Emissions Tot_Greenhouse_Gas_Emissions
##  Min.   : 0.000               Min.   :     0    Min.   :      30
##  1st Qu.: 5.859               1st Qu.:     0    1st Qu.:    9280
##  Median :20.864               Median :   105    Median :   39060
##  Mean   :28.839               Mean   :  5663    Mean   :  240177
##  3rd Qu.:43.972               3rd Qu.:  1203    3rd Qu.:  107065
##  Max.   :96.384               Max.   :300896    Max.   :12355240
##  NA's   :4                    NA's   :80        NA's   :26
```

Appendix 4 - Creating a Categorical Variable

```
wbcc4$Greenhouse_Emissions_Cat <- as.factor(
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 2040, "0%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 10336, "00-10%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 18043, "10-20%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 34488, "20-30%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 46050, "30-40%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 62600, "40-50%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 93766, "50-60%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 195996, "60-70%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 328536, "70-80%",
  ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 654080, "80-90%",
ifelse(wbcc4$Tot_Greenhouse_Gas_Emissions < 12355241, "90-100%", "Other")))))))))))
```

Appendix 5 - Classification Tree Summary

```
summary(wbcc7_tree_class)

##
## Classification tree:
## tree(formula = wbcc7$Greenhouse_Emissions_Cat ~ ., data = wbcc7)
## Variables actually used in tree construction:
## [1] "Urban_Population"             "Methane_Emissions"
## [3] "Renewable_Energy_Consumption" "Poverty_headcount_Ratio"
## [5] "Total_Population"             "PFC_Gas_Emissions"
```

```
## [7] "Population_Growth"            "Renewable_Elec_Output"
## [9] "SFSix_Gas_Emissions"
## Number of terminal nodes:  18
## Residual mean deviance:  1.297 = 131 / 101
## Misclassification error rate: 0.2605 = 31 / 119
```

Appendix 6 - Attempted LDA With Very Low Accuracy

```
library(MASS)
set.seed(0)
no_obs <- dim(wbcc4)[1]
lda.accuracy <- rep(0,10)
for(i in 1:10){
wbcc4_lda_test_index <- sample(no_obs, size = as.integer(no_obs*0.2), replace = FALSE)
wbcc4_lda_training_index <- -wbcc4_lda_test_index
wbcc4_lda.fit <- lda(Greenhouse_Emissions_Cat ~ ., data = wbcc4, subset = wbcc4_lda_traini
ng_index)
wbcc4_lda.pred <- predict(wbcc4_lda.fit, wbcc4[wbcc4_lda_test_index, -19])
wbcc4_lda.Pred_class <- wbcc4_lda.pred$class
Lda_Cont_tab <- table(wbcc4_lda.Pred_class, wbcc4$Greenhouse_Emissions_Cat[wbcc4_lda_test_
index])
lda.accuracy[i] <- sum(diag(Lda_Cont_tab))/sum(Lda_Cont_tab)
}
lda.accuracy

##  [1] 0.17391304 0.17391304 0.30434783 0.04347826 0.21739130 0.13043478
##  [7] 0.04347826 0.04347826 0.21739130 0.21739130

mean(lda.accuracy)

## [1] 0.1565217
```