STAT3622 Report- An in-depth analysis of global air and plastic pollution

by Iqra Abbasi

The motivation and objective

Pollution, both in gaseous and plastic forms, has been a well-known and perhaps one of the most dire global challenges of modern society. Interestingly, pollution emission and waste mismanagement can be traced to countries over a span of multiple years to better understand the highest sources of emitters over a period of time.

The aim of this project is thus to track different geographic regions across the globe and study their emission behaviors over time. This involves first using datasets related to plastic production to investigate plastic waste trends, and then to look at Greenhouse Gas (GHG) emissions without land use, land use change, and forestry (LULUCF) to infer air pollution levels by different countries, and how these measurements impact their rank in the World Air Quality Index.

The Datasets Used

The data used in this project comes from publicly-available, well-trusted datasets. Because this project consists of both air pollution and plastic pollution analysis, datasets relating to both air and plastic production were acquired, the details of which are listed below:

1. Global plastic production and mismanaged plastic waste per person by country from 1990 to 2015 (Source: Our World in Data)

2. Global pollutant emissions including GHG emissions by country from 1990 to 2014 (Source: UN Data)

3. Air Pollution Rank by country (Source: World Air Quality Index)

A preview of of the datasets is shown below. Plastic waste and mismanaged plastic waste are taken to be in per million tonnes and gaseous emissions are measured in kilotonnes. Additionally, rank for each country is taken from the World Air Quality index that classifies countries into one of four groups: Good, Moderate, Unhealthy (for sensitive groups) and Unhealthy.

```
## [1] "Snippet of unprocessed data for mismanaged plastic waste"
```

Country	waste_2010	waste_2019	wastepercapita_2010	wastepercapita_2019
Albania	29705	69833	0.032	24.239153
Algeria	520555	764578	0.086	17.758995
Angola	62528	236946	0.045	7.445279
Anguilla	52	0	0.010	0.000000

Country	waste_2010	waste_2019	wastepercapita_2010	wastepercapita_2019
Antigua and Barbuda	1253	627	0.051	6.463917
Argentina	157777	465808	0.026	10.401912

```
## [1] "Snippet of total global mismanaged plastic waste"
```

Entity	Code	Year	Mismanaged.waste....global.total.
Albania	ALB	2010	0.0933
Algeria	DZA	2010	1.6347
Angola	AGO	2010	0.1964
Anguilla	AIA	2010	0.0002
Antigua and Barbuda	ATG	2010	0.0039
Argentina	ARG	2010	0.4955

```
## [1] "Snippet of unprocessed data for per-capica plastic waste for each country"
```

Entity	Code	Year	Per.capita.plastic.waste..kg.person.day.
Albania	ALB	2010	0.069
Algeria	DZA	2010	0.144
Angola	AGO	2010	0.062
Anguilla	AIA	2010	0.252
Antigua and Barbuda	ATG	2010	0.660
Argentina	ARG	2010	0.183

```
## [1] "Snippet of unprocessed greenhouse gas emissions for eaach country from 1990 to 2014"
```

country	year	CO2	GHGs	HFCs	CH4	N2O	SF6
Australia	2014	393126.9	522397.1	10787.350	98076.11	20084.54	129.6054
Australia	2013	396913.9	526882.7	10034.128	99857.20	19756.45	128.9446
Australia	2012	406462.8	537377.6	9353.066	100796.84	20342.38	127.5522
Australia	2011	403705.5	534089.8	8837.851	101085.54	20034.58	125.0019
Australia	2010	406201.0	533917.4	8166.067	99447.73	19698.30	121.0317
Australia	2009	408448.5	537889.9	7468.944	101886.83	19607.63	119.4511

Preprocessing and preparing data for visualising

After loading in the data, several preprocessing steps were applied to facilitate various types of visualisations, including:

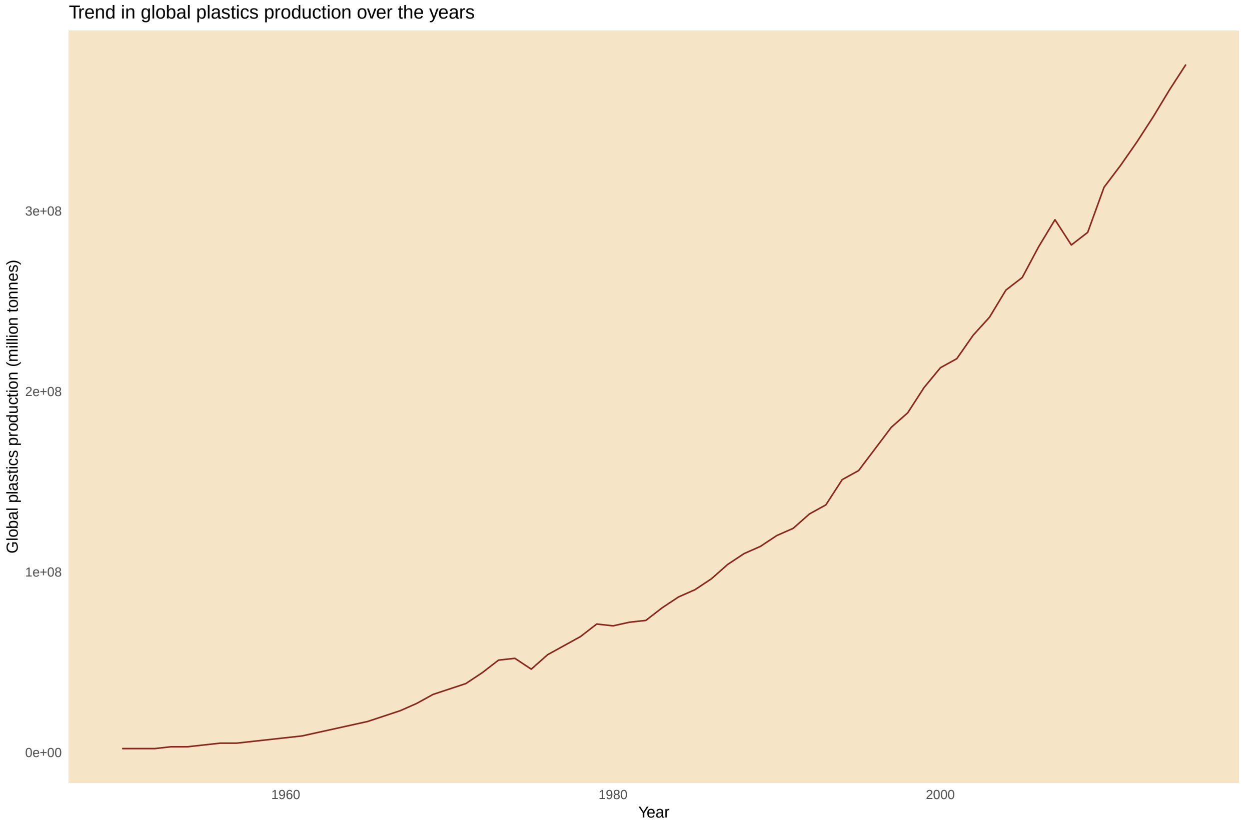
1. Imputing missing values (taking mean if missing value is in time-series format).

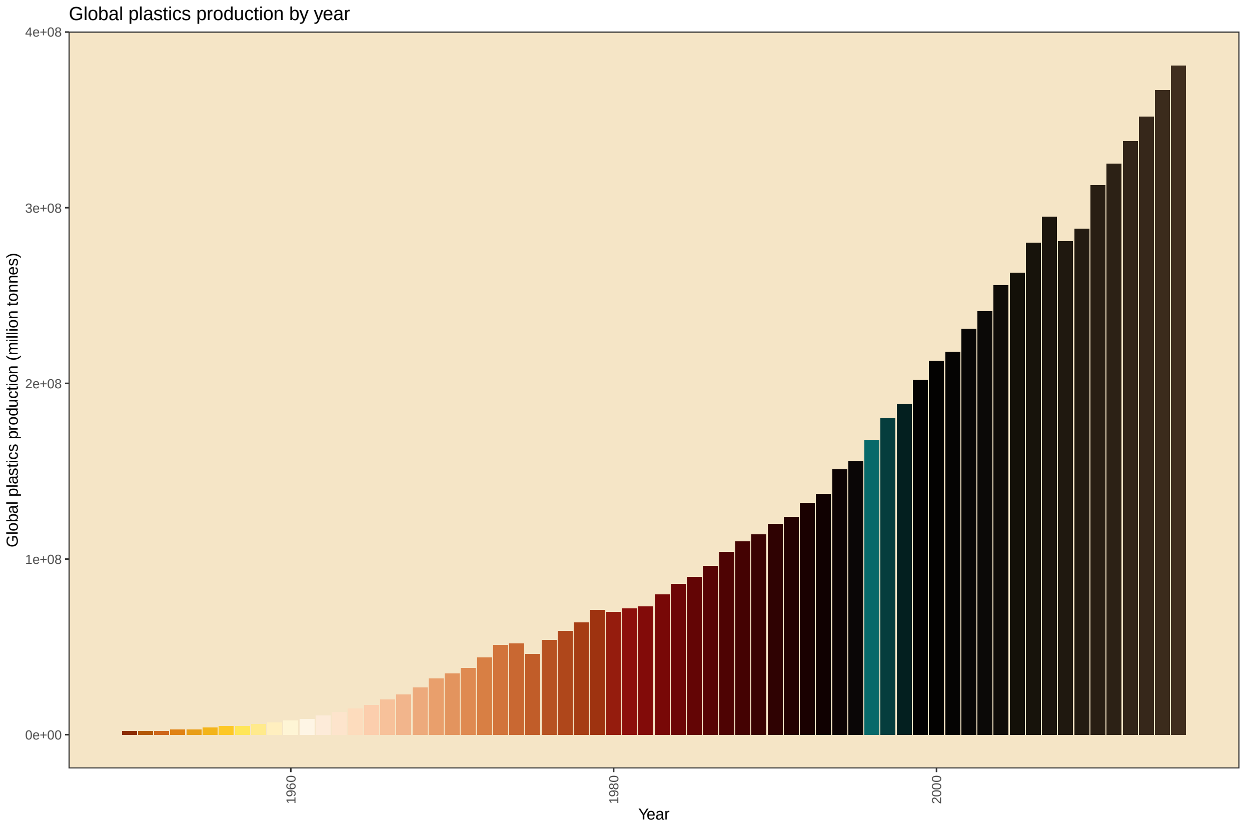
2. Converting datasets to ‘long’ format for plotting trends of explanatory variables (such as each row representing a year and its row values containing the emission values of the countries in that year).

3. Joining world rank index with the GHG emissions dataset for classification models.

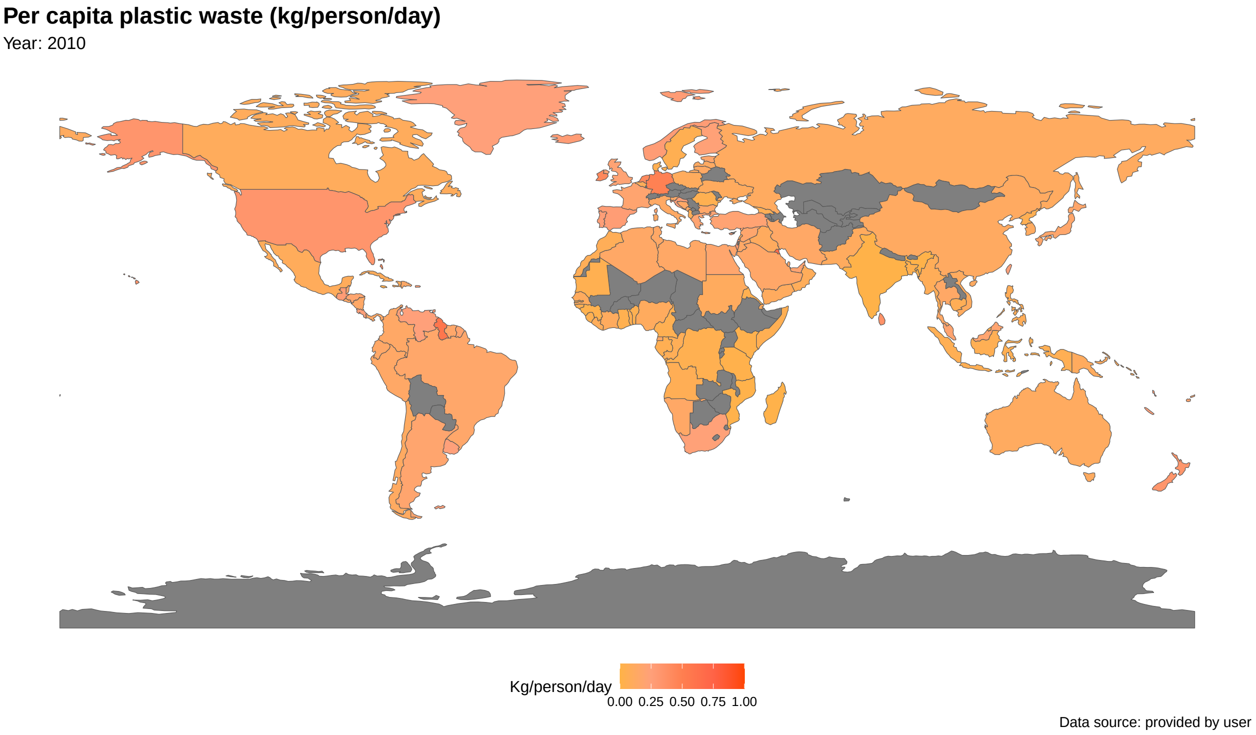
Analyzing Global mismanaged waste and global plastic production:

The following line and bar plot shows the trend in global plastic waste production over the years. The plots show a clear growing trend overall which has seen an accelerated increase especially after the late 90s.




Global plastics production by year
4e+08
3e+08
2e+08
1e+08
0e+00
Global plastics production (million tonnes)
1960
1980
2000
Year

Another way to better understand the trend of plastic pollution is the per capita waste generated by a person for each day for every country. The darker the shade, the higher the proportion of plastic waste generated by the country. The plot below indicates that countries such as the United States and Germany have a higher per capita plastic waste generation per day. There is no data available for countries that are greyed out, and many countries in Africa and Asia fall in the 0-0.25 kg threshold of plastic waste production per day.

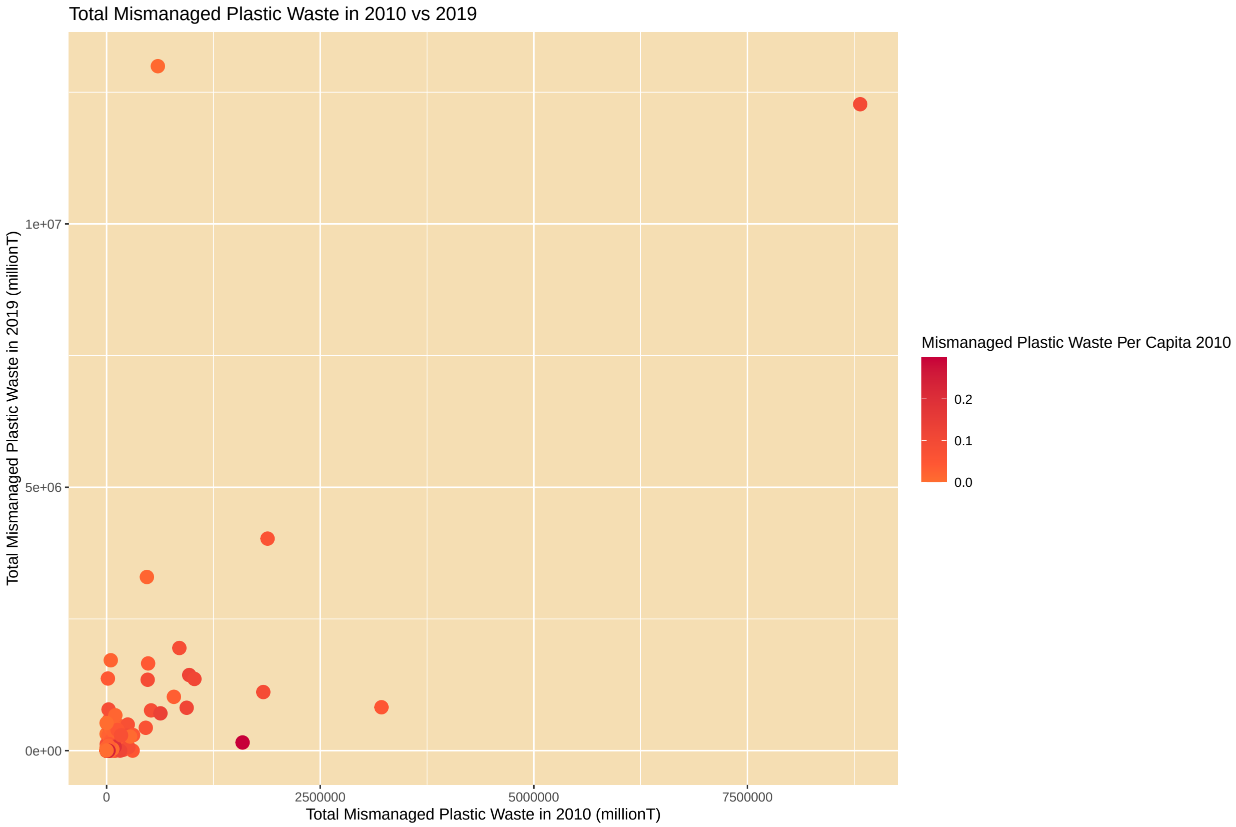


Per capita plastic waste (kg/person/day)

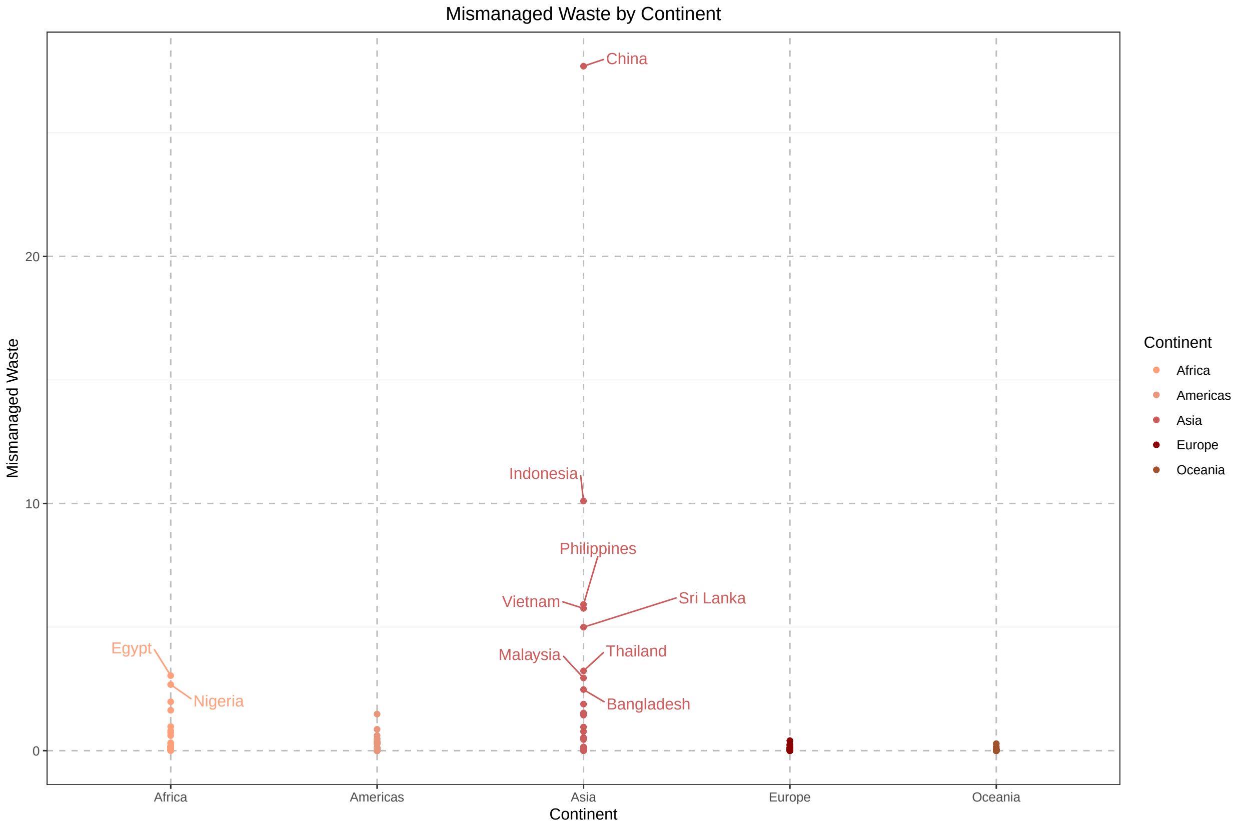
Year: 2010

Data source: provided by user

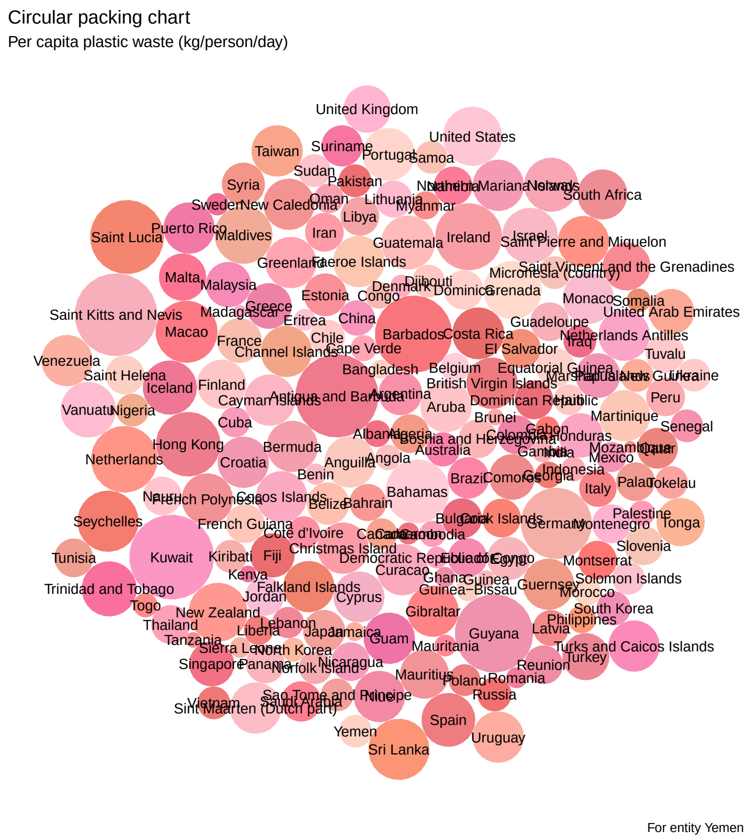
It is also useful to compare how mismanaged plastic waste has changed across an extended period of time. Here, mismanaged waste refers to waste that ended up in landfills or in the ocean. The plot below shows data points for each country, and a clear positive relationship is seen between the amount of mismanaged plastic waste in 2010 and that in 2019. Another observation is that there are only a few countries with a net plastic waste higher than 2500000 million tonnes, and that most are clustered around the origin.



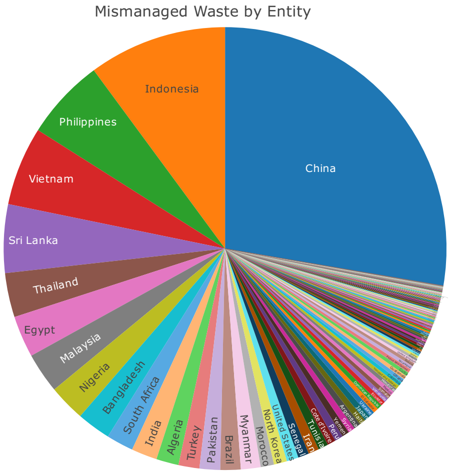
To analyze how countries' mismanaged global waste production varies by continent, it is useful to utilize R's 'countrycode' library to attach the continent of each country to the dataframe in a separate column called 'Continent'. By grouping the data points by continent and making the following plots, we see that Asia is ahead of other continents in terms of countries with high-producing mismanaged waste, and China is far ahead of other countries in terms of the mismanaged waste it produces.



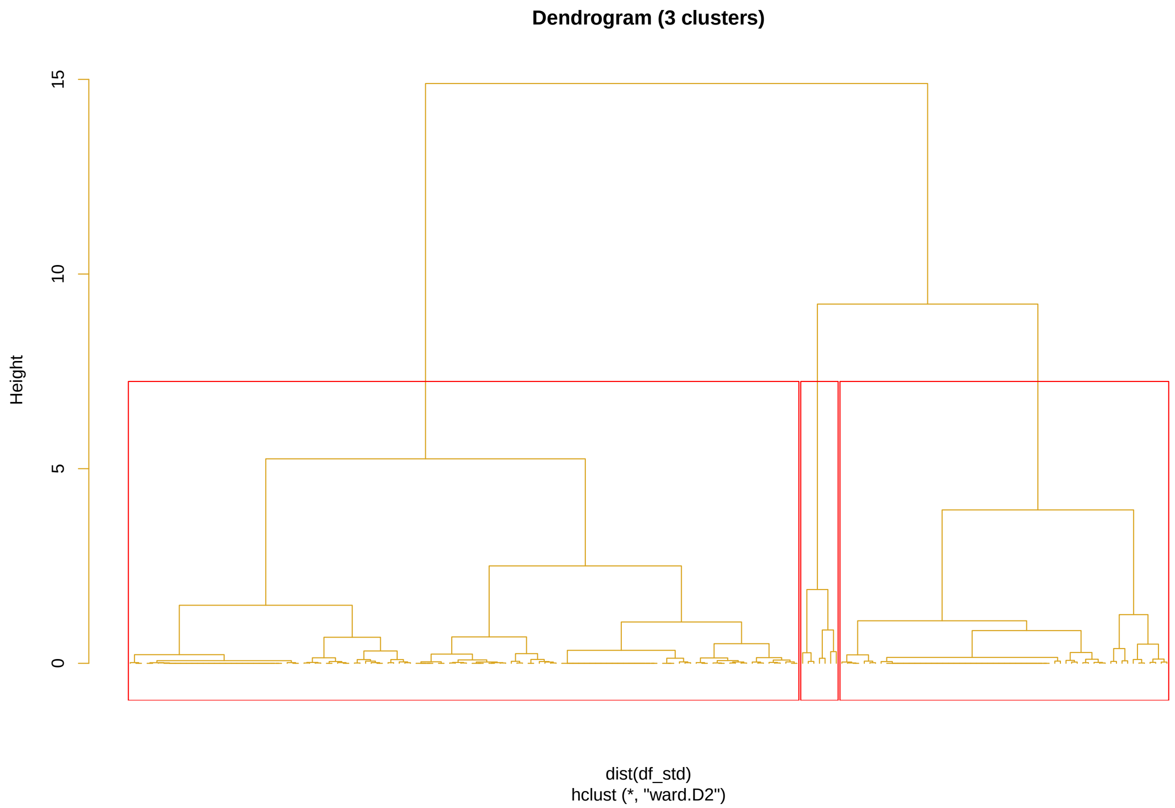
It is worth noting that plastic waste per capita is not the same measurement as mismanaged waste. Plastic waste per capita can be an indicator of waste generated, but that waste can be recycled or repurposed. Mismanaged waste, on the other hand, is waste that goes directly to landfills or the ocean. The circular-packing graph and pie chart below capture these differences - China takes the lead in mismanaged global waste according to the pie chart, and the circular packing chart shows that countries such as Kuwait and Germany produce the most plastic waste per capita -they do not necessarily have high values of mismanaged waste, suggesting that they have adequate measurements in place to repurpose the plastic waste produced.

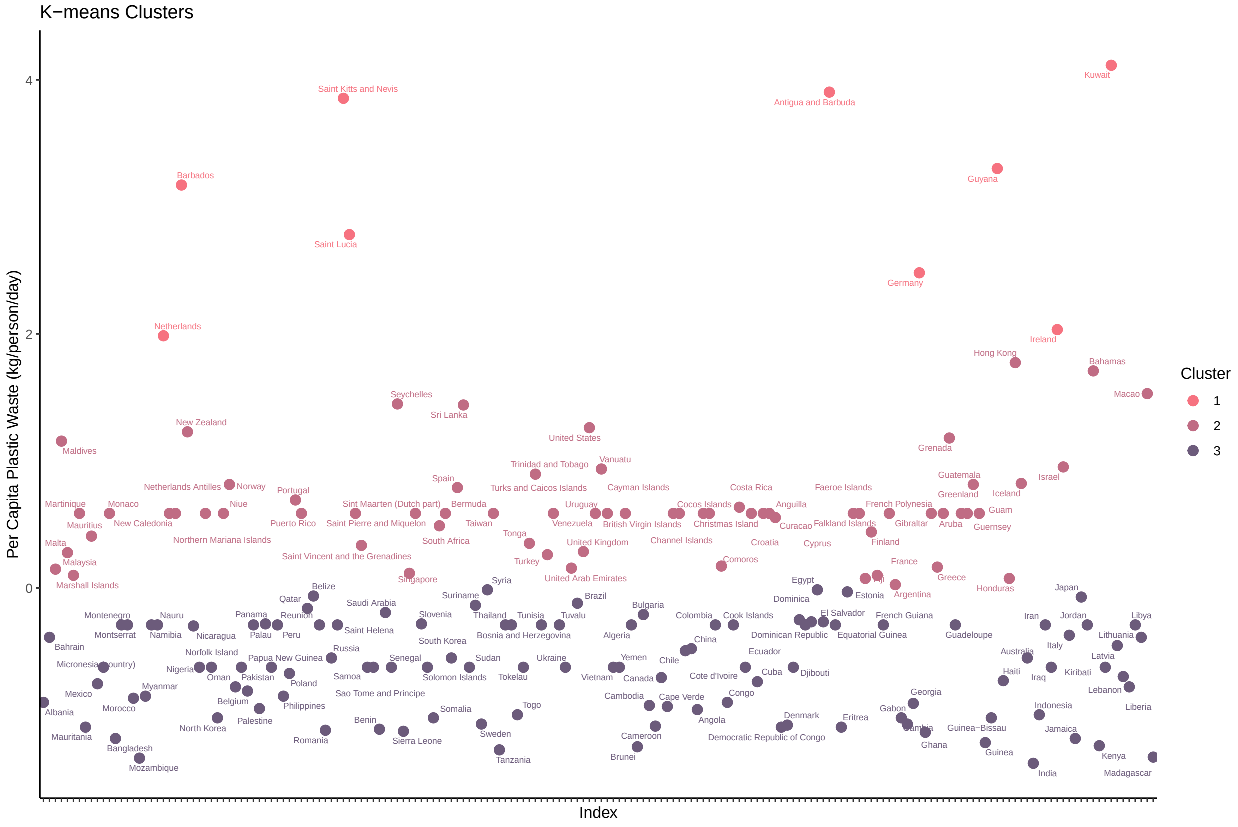

Circular packing chart

Per capita plastic waste (kg/person/day)

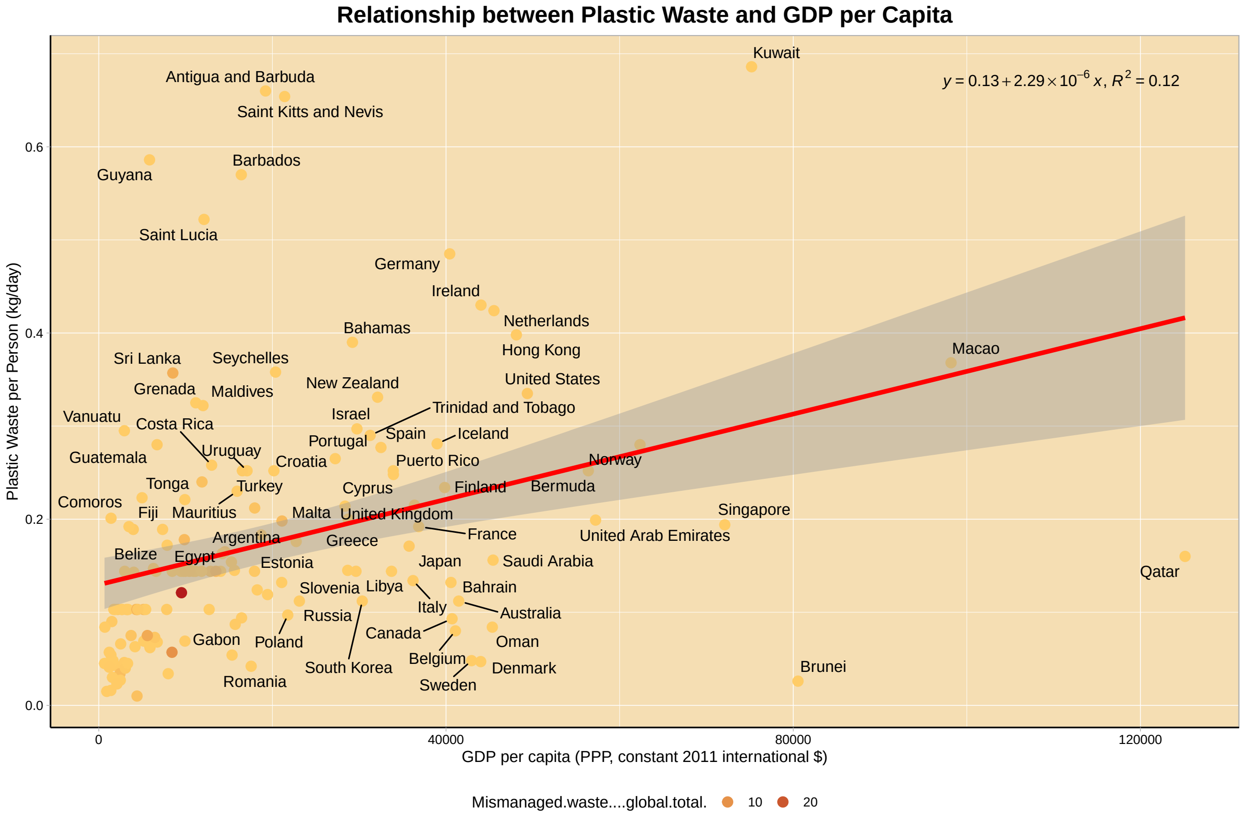


To better study the similarity of countries according to the cluster they belong to, hierarchical clustering is used. Because a value of the optimal number of clusters is not known, a dendrogram is first created using Ward's method. Results from the dendrogram show three distinct clusters. We therefore go ahead and assign three clusters to our dataframe using the K-means method, and get the following plot as a result. The data is divided into three distinct clusters - countries with high plastic waste per capita can now be distinguished from countries with medium and low plastic waste per capita.


Dendrogram (3 clusters)
Height
15
10
5
0
dist(df_std)
hclust (*, "ward.D2")

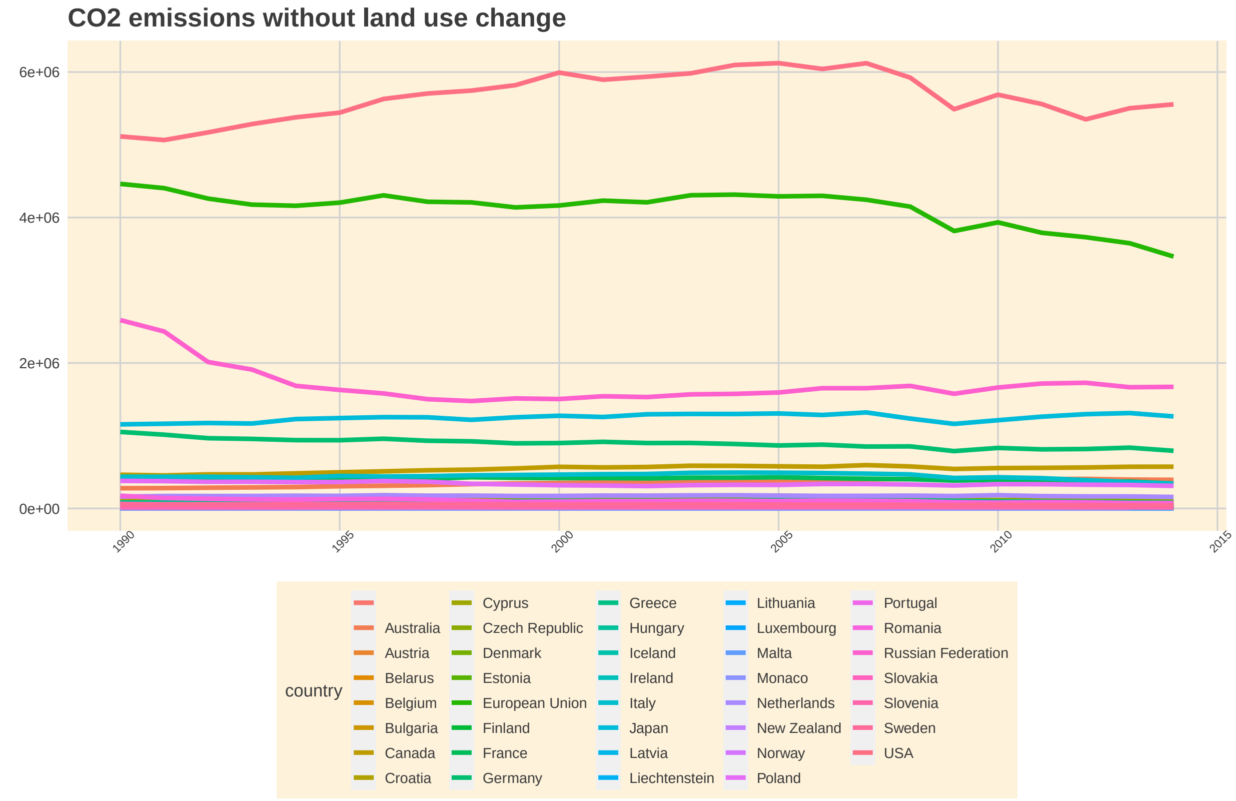

K−means Clusters
Per Capita Plastic Waste (kg/person/day)
4
2
0
Index
Cluster
● 1
● 2
● 3
Saint Kitts and Nevis
Antigua and Barbuda
Kuwait
Barbados
Guyana
Saint Lucia
Germany
Netherlands
Ireland
Hong Kong
Bahamas
Macao
Seychelles
Sri Lanka
New Zealand
United States
Maldives
Grenada
Trinidad and Tobago
Vanuatu
Guatemala
Israel
Spain
Netherlands Antilles
Norway
Portugal
Turks and Caicos Islands
Cayman Islands
Costa Rica
Faeroe Islands
Greenland
Iceland
Martinique
Monaco
Niue
Sint Maarten (Dutch part)
Bermuda
Uruguay
Cocos Islands
Anguilla
French Polynesia
Guam
Mauritius
New Caledonia
Puerto Rico
Saint Pierre and Miquelon
Taiwan
Venezuela
British Virgin Islands
Christmas Island
Curacao
Falkland Islands
Gibraltar
Aruba
Guernsey
Malta
Northern Mariana Islands
South Africa
Tonga
United Kingdom
Channel Islands
Croatia
Cyprus
Finland
Malaysia
Saint Vincent and the Grenadines
Turkey
Comoros
France
Greece
Marshall Islands
Singapore
United Arab Emirates
Fiji
Honduras
Japan
Belize
Syria
Egypt
Estonia
Argentina
Qatar
Suriname
Brazil
Dominica
Saudi Arabia
Bulgaria
El Salvador
Montenegro
Nauru
Panama
Reunion
Slovenia
Thailand
Tunisia
Tuvalu
Colombia
Cook Islands
French Guiana
Iran
Jordan
Libya
Montserrat
Namibia
Nicaragua
Palau
Peru
Saint Helena
South Korea
Bosnia and Herzegovina
Algeria
Dominican Republic
Equatorial Guinea
Guadeloupe
Lithuania
Bahrain
China
Italy
Norfolk Island
Russia
Ecuador
Australia
Micronesia (country)
Papua New Guinea
Senegal
Sudan
Ukraine
Yemen
Chile
Latvia
Nigeria
Oman
Pakistan
Samoa
Solomon Islands
Tokelau
Vietnam
Canada
Cote d'Ivoire
Cuba
Djibouti
Haiti
Iraq
Kiribati
Poland
Myanmar
Lebanon
Mexico
Sao Tome and Principe
Cambodia
Cape Verde
Congo
Georgia
Belgium
Morocco
Philippines
Somalia
Togo
Angola
Gabon
Indonesia
Liberia
Albania
Palestine
Benin
Sweden
Denmark
Eritrea
Gambia
Guinea−Bissau
Mauritania
North Korea
Romania
Sierra Leone
Cameroon
Democratic Republic of Congo
Jamaica
Bangladesh
Tanzania
Brunei
Ghana
Guinea
Kenya
Mozambique
India
Madagascar

One study that can be done from the data sources we have is to evaluate if a country's GDP corresponds with the plastic waste per capita produces. By plotting GDP against plastic waste per captia, and plotting a line of best fit using linear regression, we see that there is a clear positive relationship between GDP per capita and plastic waste generated per capita. We also notice that the confidence interval of the best fit line is also relatively wide, and that data points are not tightly bound towards the generated line, showing that there are variations from the coefficients that the linear regression model predicted. The poor linear relationship is also justified by the R squared value of just 0.12.

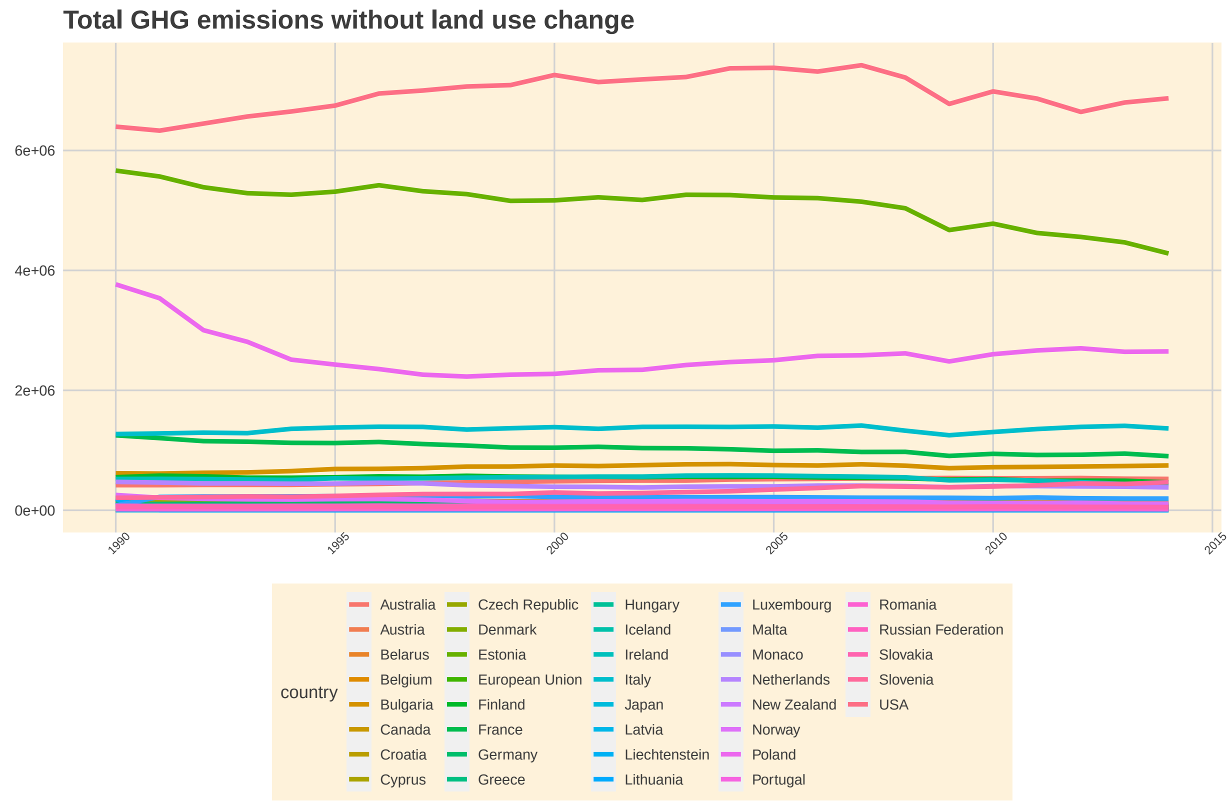


Exploring Green House Emissions across countries

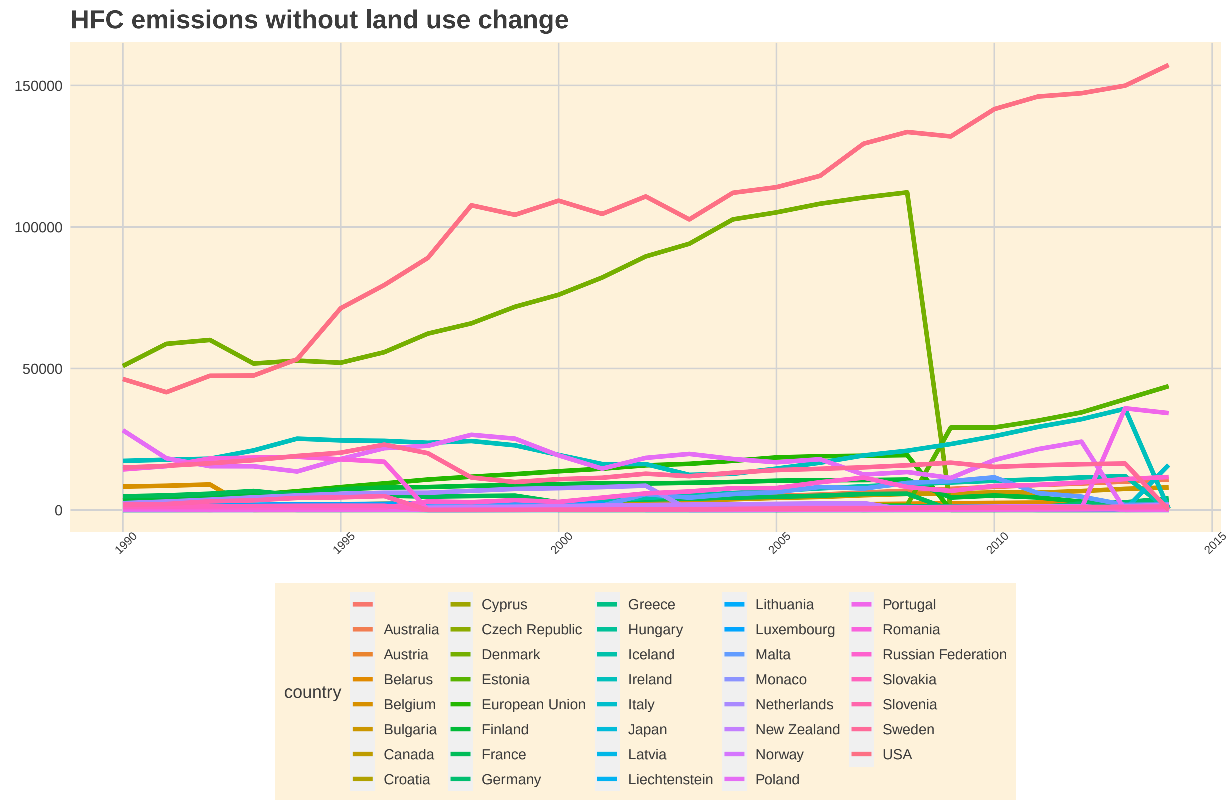
To now explore the emission of green house gases and other gaseous pollutants across countries, we create line plots for each pollutant, for each country. The following line plot produces the trend of CO2 emission without land use change (this dataset looks at gaseous emissions without those contributing to Land use, land-use change, and forestry, or LULUCF, as human-induced land use is necessary for economic and social development and cannot be avoided). Plotting CO2 emissions without land use change shows that there has generally been a steadily slow decline, and the net emissions by European Union countries signified by the green line has seen a dramatic drop in CO2 emissions, while emissions from USA saw a gradual drop but shows a gradual recent increase.


CO2 emissions without land use change
6e+06
4e+06
2e+06
0e+00
1990
1995
2000
2005
2010
2015
country
Australia
Austria
Belarus
Belgium
Bulgaria
Canada
Croatia
Cyprus
Czech Republic
Denmark
Estonia
European Union
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Japan
Latvia
Liechtenstein
Lithuania
Luxembourg
Malta
Monaco
Netherlands
New Zealand
Norway
Poland
Portugal
Romania
Russian Federation
Slovakia
Slovenia
Sweden
USA

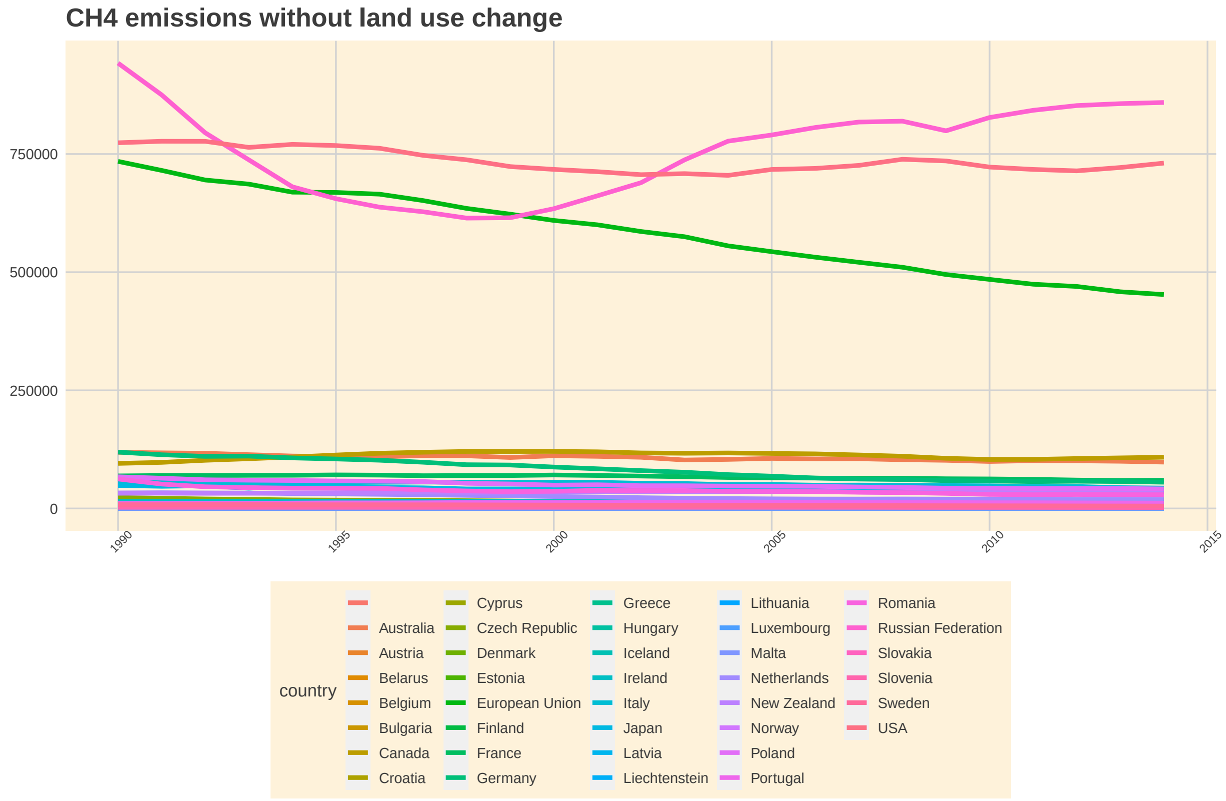
Next, plotting total Greenhouse Gas (GHG) emissions without land use change for each country shows that there has also been a general decline. The European Union in particular can be seen as having a dramatic decline from its previously high numbers here. The USA on the other hand continues having steady numbers since 2010.



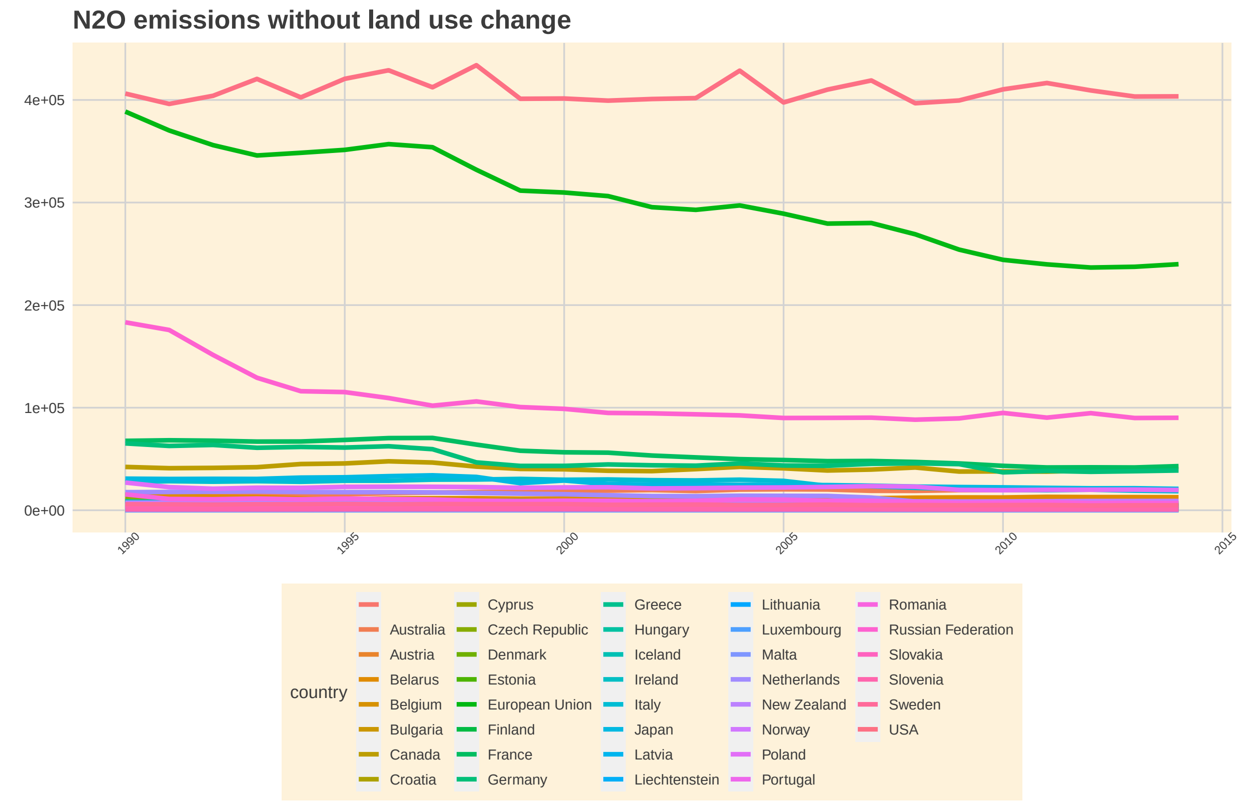
HFC emissions is also a type of greenhouse gas that can trap heat, and plotting its emissions over the years shows that it started fluctuating in the late 90s, and Denmark surprisingly saw a dramatic increase then sharp decline for HFC emissions in the past 30 years. But perhaps the most noteworthy is the emissions by USA; it has seen steady growth since the 90s and surpasses all other countries by a huge margin.



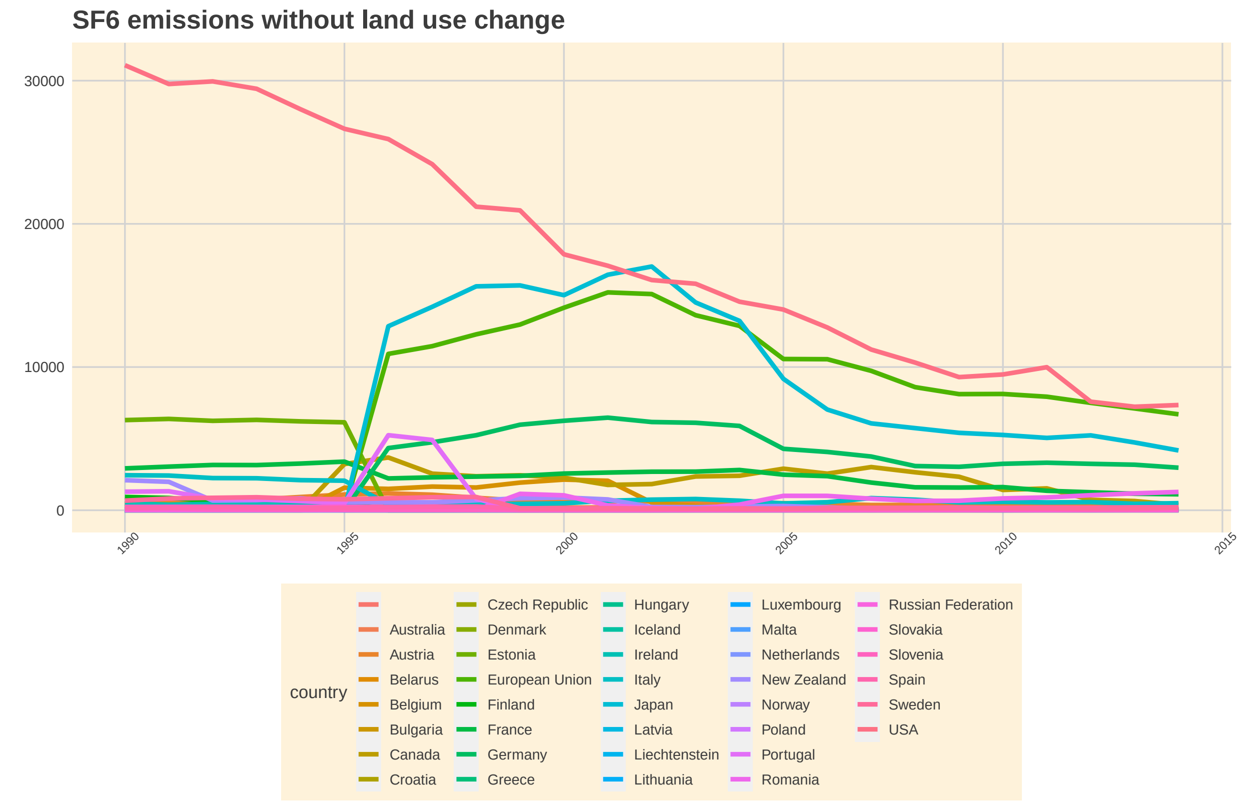
By tracking CH4 emissions, on the other hand, we see that there are three leading drivers for CH4 emissions: the Russian Federation, the USA, and the European Union. The Russian Federation saw a moderate decline but has been climbing up in terms of its CH4 emissions, and the European Union is still steadily seeing declining numbers for CH4. The USA is also far ahead in terms of CH4 emissions, though there are signs of slow decrease over the past few decades.



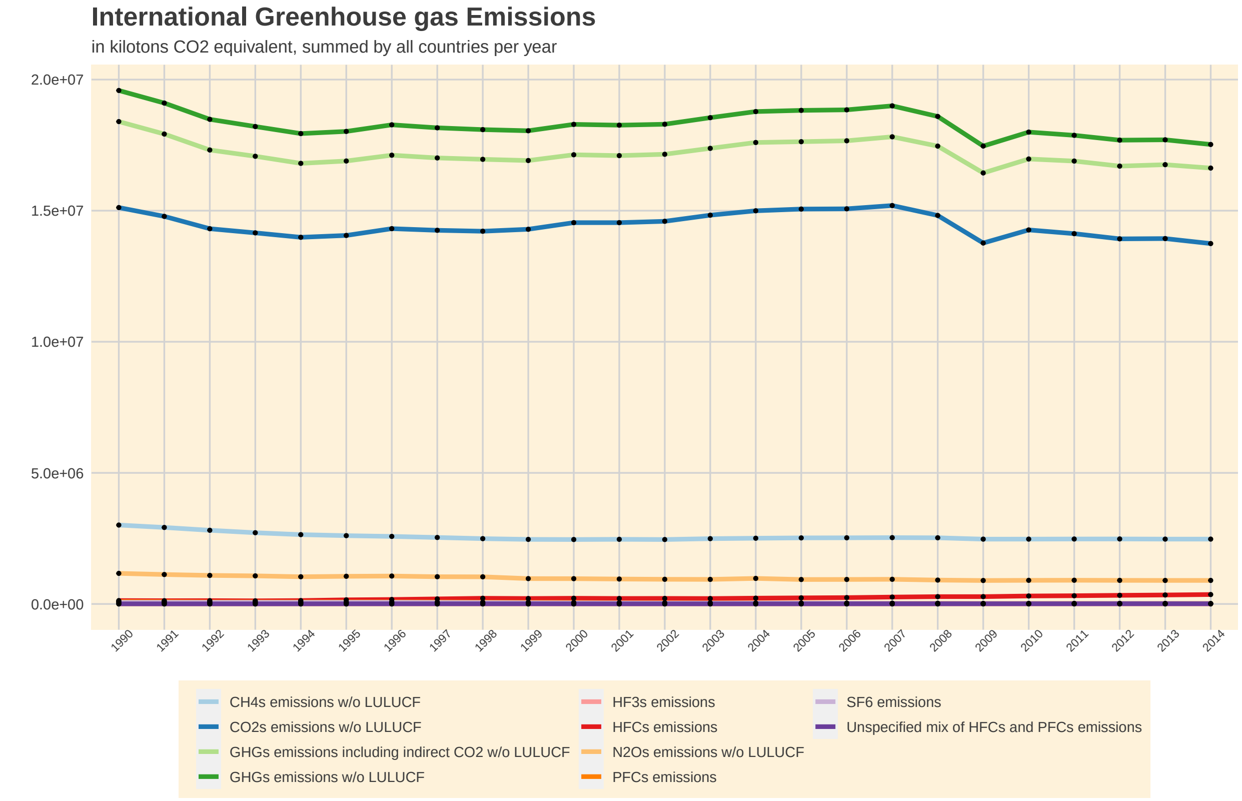
For N2O emissions, we again see that the European Union , the USA, and the Russian Federation take the lead against other countries. Russia and the European Union have been seeing a steady decline in N2O emissions since the late 80s. The N2O emissions produced by US have seen recent fluctuations, but no significant decrease. Other countries have also been experiencing a decline in N2O emissions.



Finally, by looking at SF6 trends over the years, we see that most countries saw a sharp increase in SF6 emissions at around late 1990s, followed by a rapid recent decline. The US had the highest rate of SF6 emissions, after which it saw a steep decline in the early 90s. Japan, the European Union, and Germany in particular saw steep peaks in SF6 emissions at the turn of the century, followed by a slow decline.

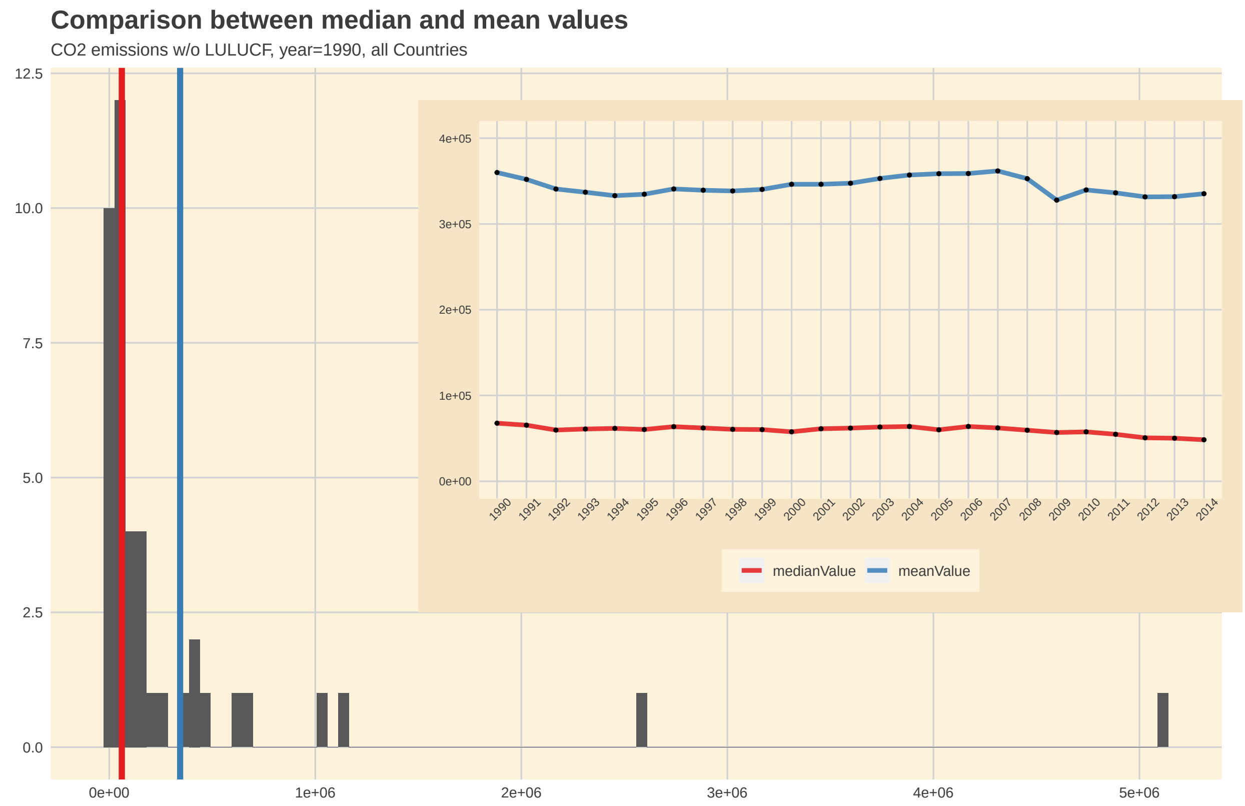


After examining the distribution of emissions by country across several years, we look at the combined greenhouse emission trend globally. This involves summing up all the emission values over the years to visualize the overall trend from 1990 to 2014. We therefore see not only the individual emissions, but all of them combined to form the total GHG emissions. From the generated line plot, we see that CO2 emissions without LULUCF are the biggest contributors of net Greenhouse Gases (the lines in green), and CH4 is the second biggest driver, though it is far behind CO2.

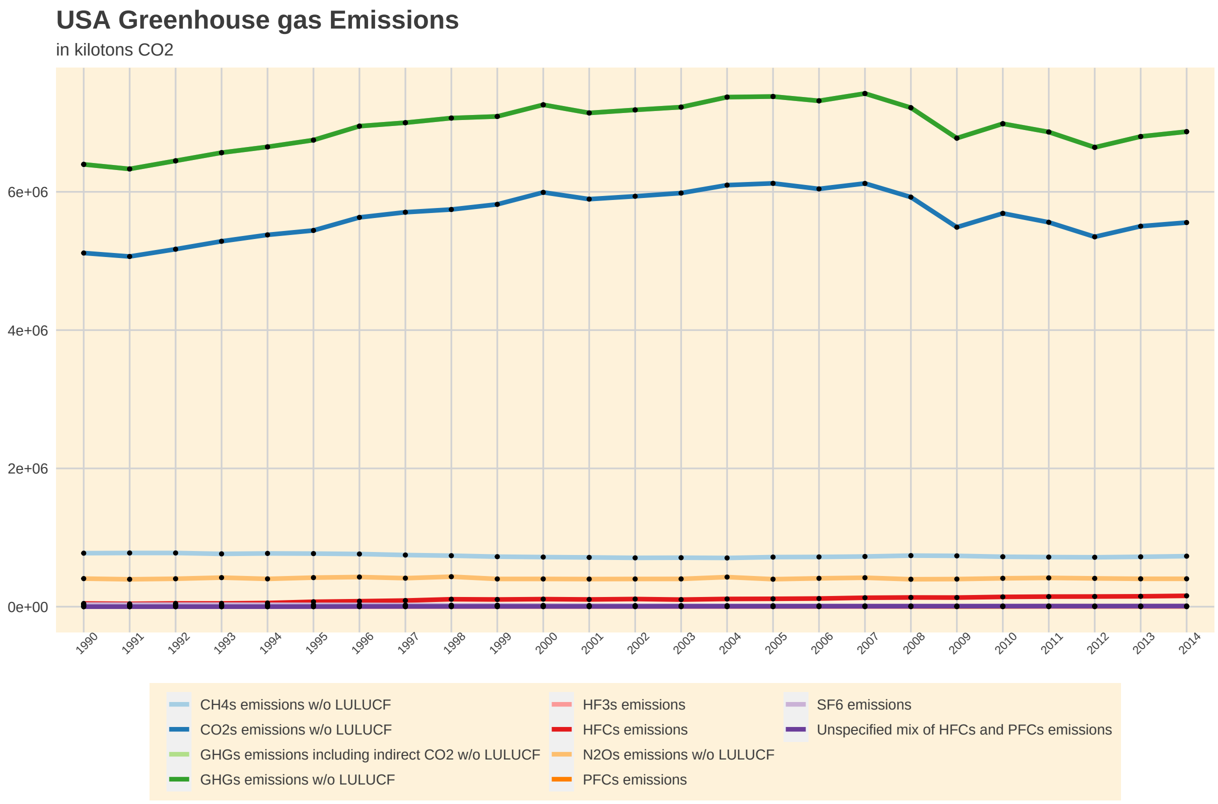

International Greenhouse gas Emissions

in kilotons CO2 equivalent, summed by all countries per year

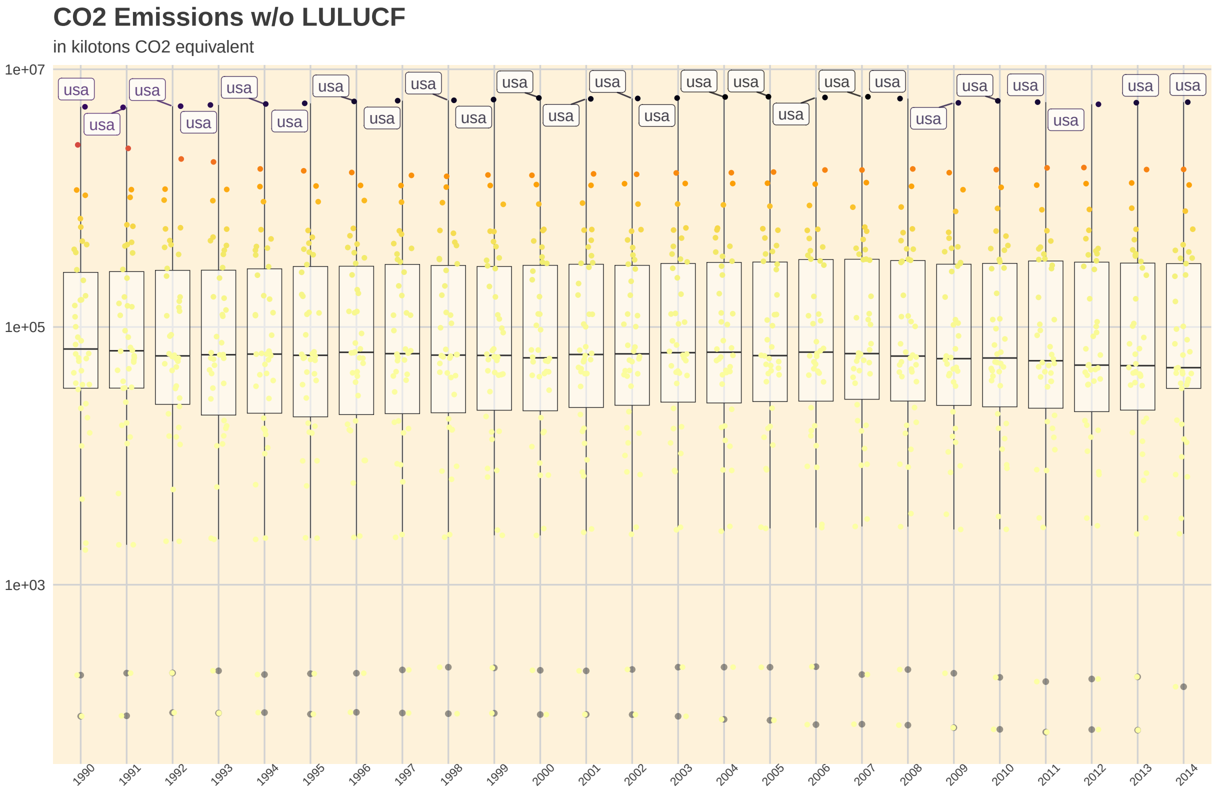
We next visualize the distribution of CO2 emissions and compare the median and mean values of this distribution. From the line plot, the median value seems to have been more steady over the years than the mean value. We also see that the median is greater than the mean in this distribution, and this significant difference between the two shows the effect that outliers can have on the mean of the data.



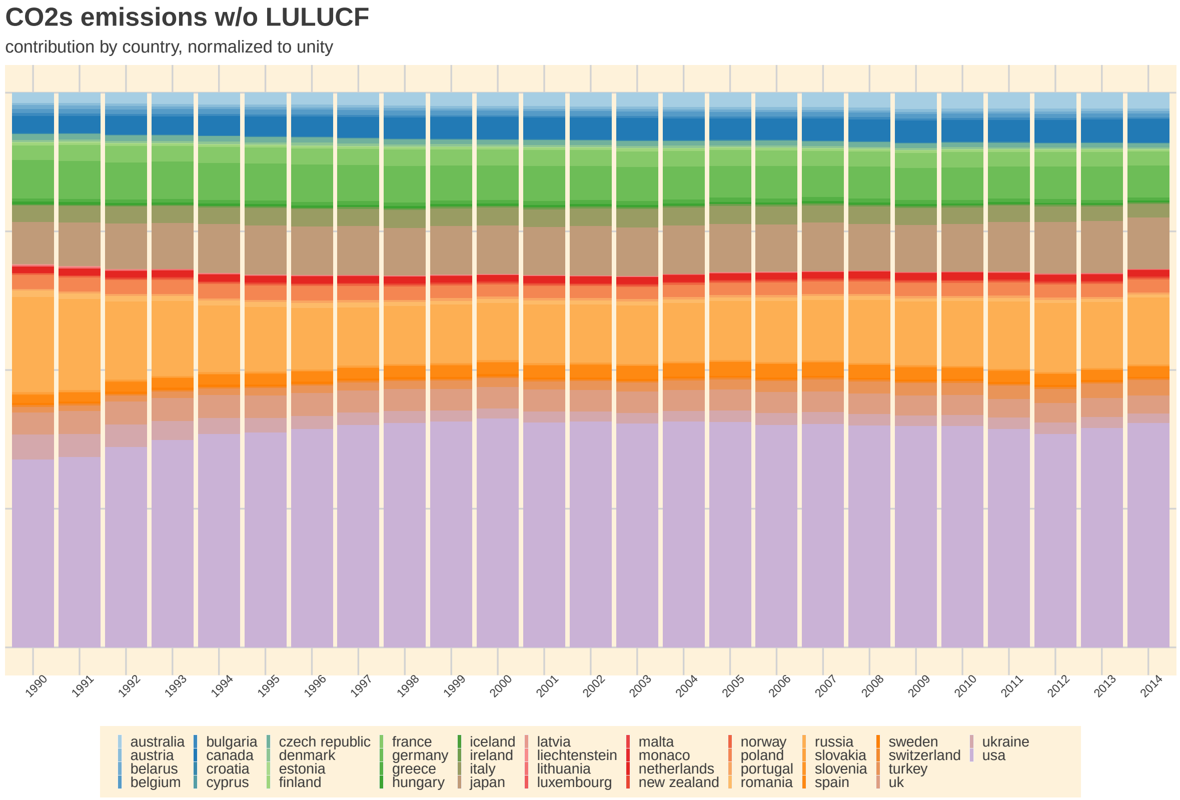
To look at emissions of a particular case, we look at emissions produced by USA. We see that GHG emissions were steadily increasing until 2008-2009, after which it slowly declined but has picked up in the past decade. Since the line trend for CO2 is similar to that of all GHG emissions, it is reasonable to assume that the increase in GHG emissions was heavily driven by the increase in CO2 emissions for this case.



To further investigate CO2 emissions, we plot CO2 emissions as boxplots for each year, and see that USA has consistently been in the upper end of the box plot, showing that it has heavily been contributing towards global CO2 emissions.



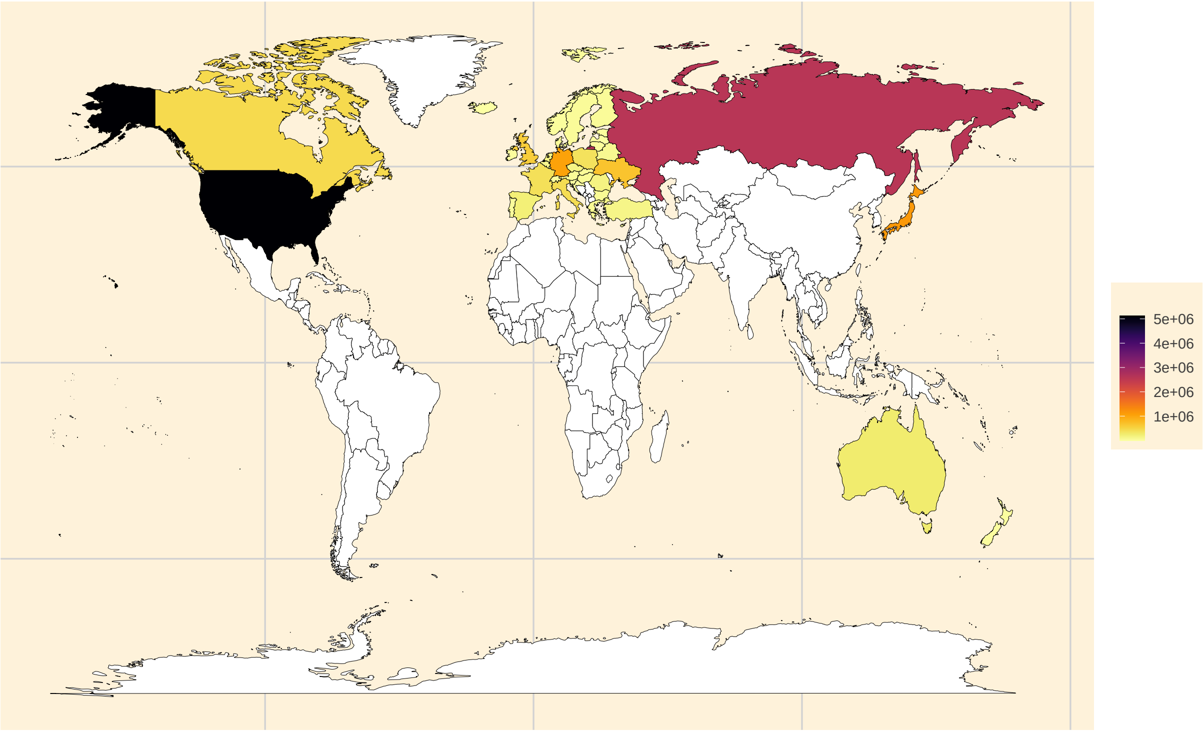
We can also create a stacked histogram by country and year to see which countries are the top three contributors of CO2. As we can see, the top three contributors are USA (by a huge margin), Russia, and Japan.



By using a world map to compare CO2 emissions, we also see how CO2 levels have changed across the years in a manner that is easy to visualise at a global level. We see that primarily, Russia's CO2 emissions have declined in 2014 since 1990, and some other countries saw slighter declines.

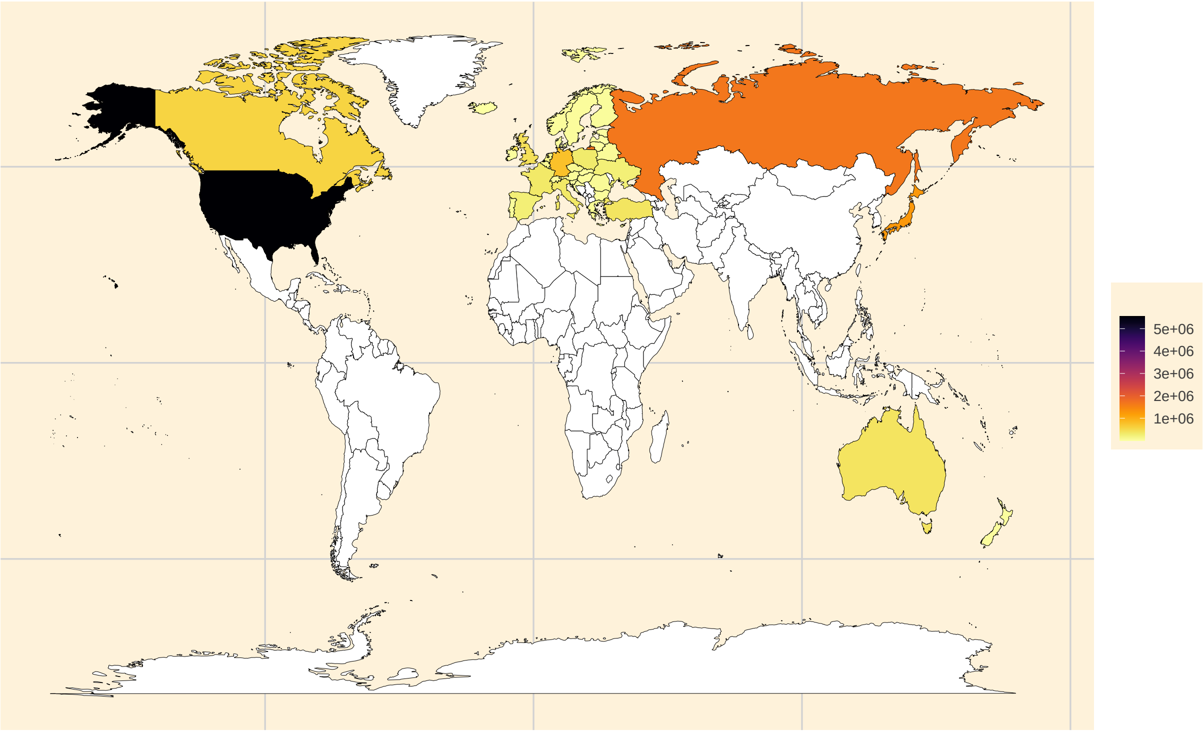
CO2s emissions w/o LULUCF

in kiloton, year 1990



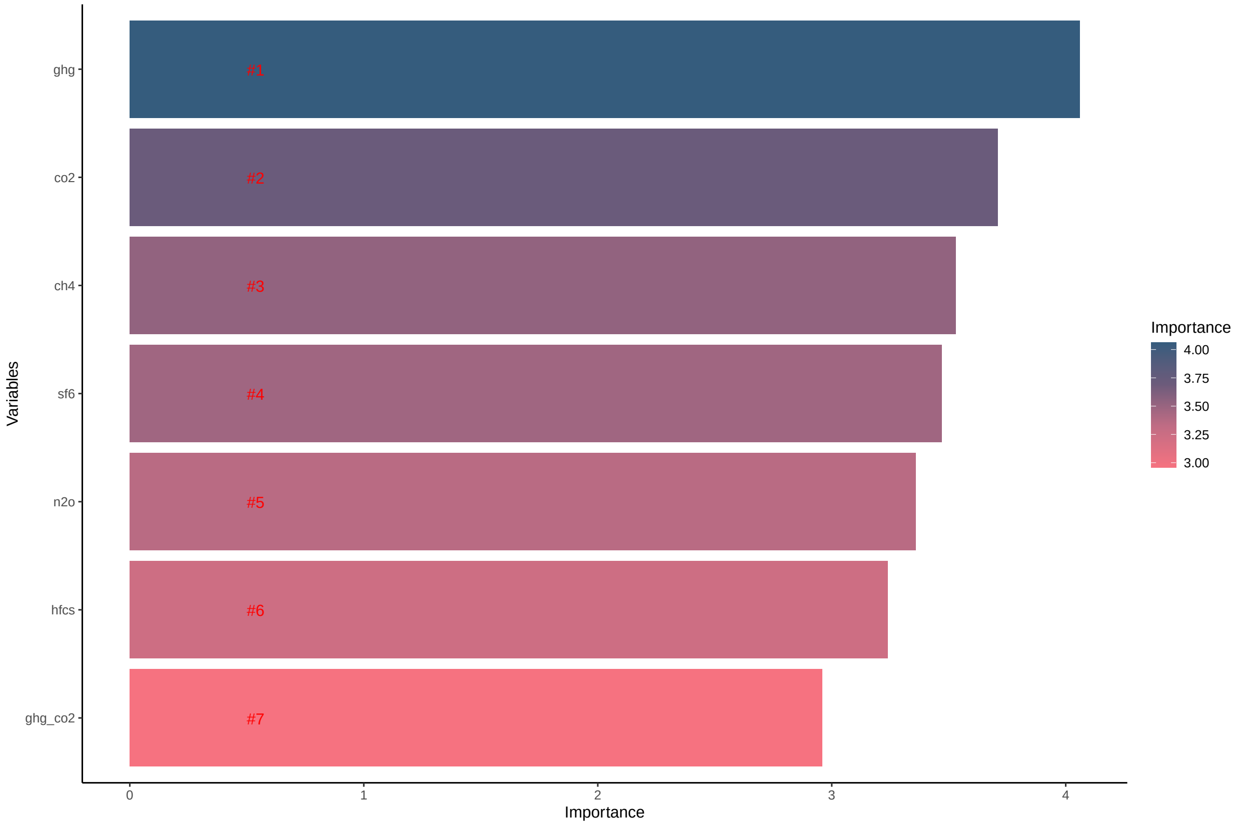
CO2s emissions w/o LULUCF

in kiloton, year 2014



Classification

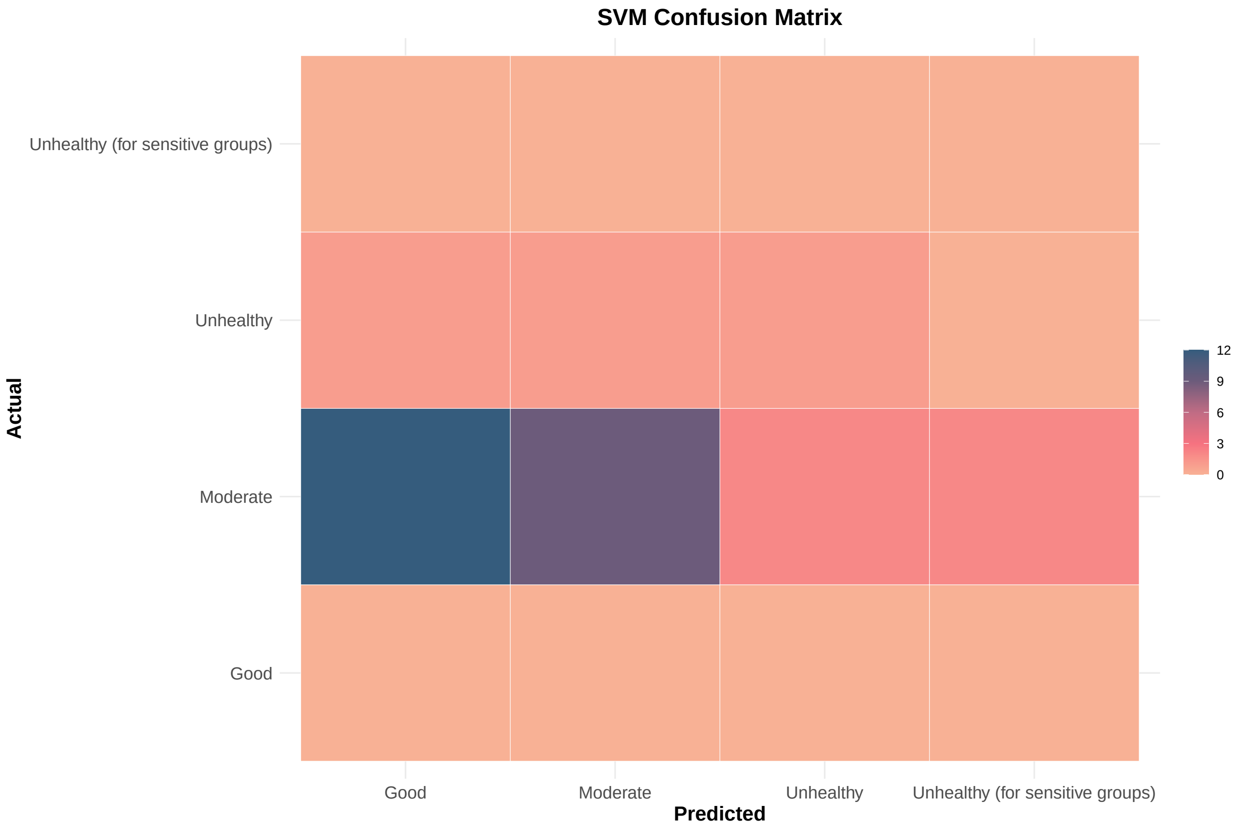
By using data from the World Air Quality Index, which classifies countries as ‘Good’, ‘Moderate’, ‘Unhealthy (for sensitive groups)’, and ‘Unhealthy’ based on the quantity and types of pollutants emitted, this section attempts to classify the countries present in our dataset into one of the four groups. The first choice was random forest, and after performing random forest classification on the merged dataset, we get an accuracy of 100 percent, signifying that the model may not be picking up sufficient details to produce a more realistic accuracy rate. After running feature importance on the random forest model, we see that CO2 and CH4 bore roughly the same importance, with other emissions being much less important for the model.



```
## [1] "Random Forest Accuracy: 100 %"
```

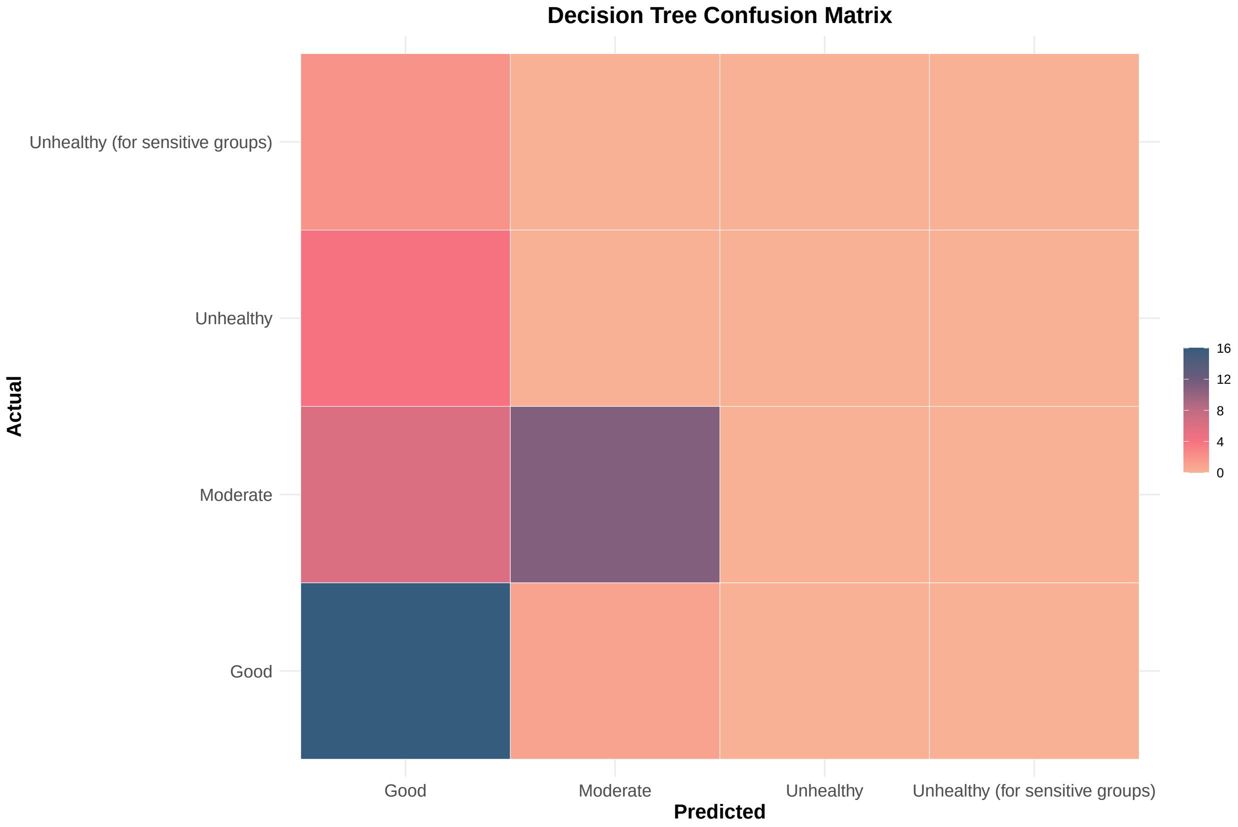
After Random Forest, SVM was tested to see if the countries can be correctly categorized into their correct groups. The kernel chosen was ‘polynomial’, and the results from the confusion matrix with the accuracy of 37.5% shows that the model was moderately effective in predicting the ranks: a lot of its predictions were ‘Moderate’ and while it was good at correctly predicting the ‘Moderate’ rank, it also frequently predicted ‘Moderate’ as ‘Good’.

```
## [1] "SVM Accuracy: 35.71 %"
```



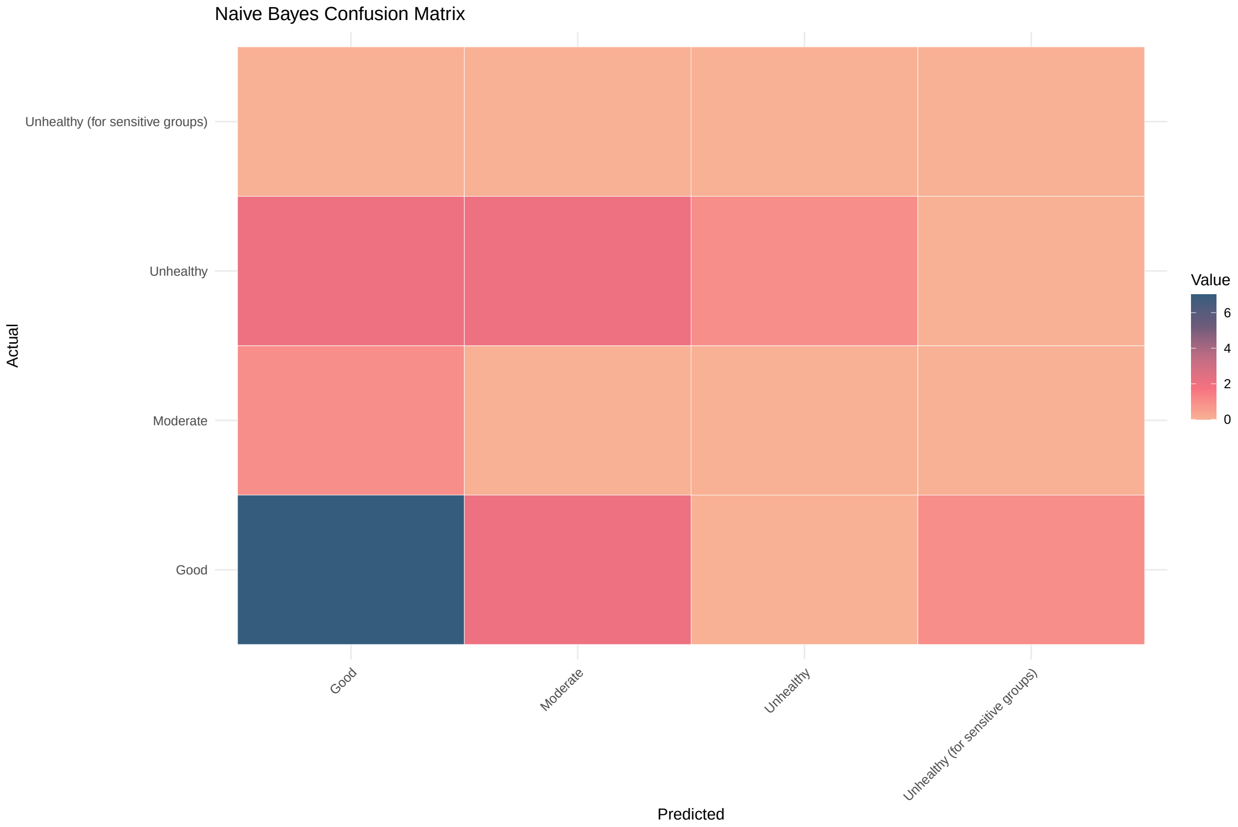
Decision trees were also tested for this classification problem, and they delivered a significantly better accuracy of 67.5%. Moreover, the confusion matrix indicates that it was mostly able to classify 'Moderate' and 'Good' ranks, but attributed the 'Good' rank to both the 'Unhealthy' ranks, making it weaker in terms of ranking countries that are in the 'Unhealthy' category.

```
## [1] "Decision Tree Accuracy: 67.5%"
```

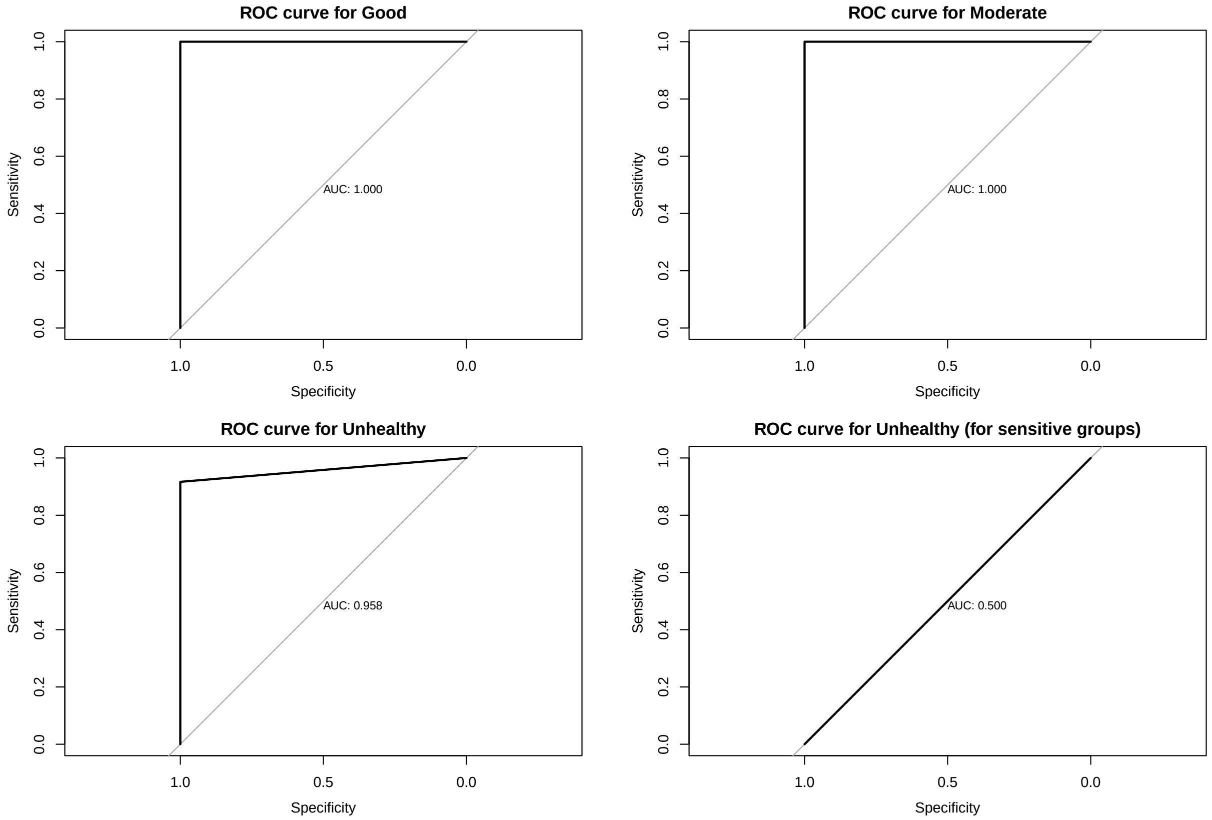


Next, we look at the Naive Bayes classifier to see if it performs better. The Naive Bayes assumes independence among all the factors and is a relatively simpler model better for classifying smaller datasets such as ours. Results from the confusion matrix and the accuracy rate of 50% shows that it was moderately effective in classifying the countries correctly. It performed well for the 'Good' category but performed poorly for the 'Unhealthy' category. A constant poor performance for the 'Unhealthy' category is because there were only a few training samples of this category in the dataset, making the models weaker towards classifying it correctly.

```
## [1] "Naive Bayes Accuracy: 50 %"
```



Another way of classifying these ranks is by considering a 'One-vs-All' approach when classifying; specifically, we train separate models as binary classifiers to detect one of the four categories. We select decision trees for its relatively better performance for this approach, and the ROC AUC curves generated below show that a one-vs-all approach performs much better at classifying countries according to their ranks. If one wanted to correctly classify a country in the 'Unhealthy (for sensitive groups)' category, the sample will be passed to all four models, and a 'True' value will be generated only by the 'Unhealthy (for sensitive groups)' model.



From the above analyses and computations, we can make the final inferences: 1. A 'one-vs-all' model using decision trees is more reliable for classification of countries' ranks rather than using models for multi-class classification. 2. a country's GDP bears a positive relationship with per capita plastic waste production, but the positive relation is not tightly bound across all countries. 3. A country's per capita plastic waste is not the same indicator of its mismanaged waste. 4. The general trend of greenhouse gases is slightly declining, but Carbon Dioxide remains the biggest contributor to GreenHouse Gas emissions, and the USA continues to emit the most amount of CO2 by a large margin.