***Analysis of data to study the effects of climate change based on Energy Use, CO2 Intensity, and Agricultural Land Area.***

**Student Name: Harshini Netha**
**Student ID : 22033688**
**Course : Applied Data Science 1**

**Abstract**

By examining changes in energy consumption, CO2 intensity, agricultural land area, using the World Bank data, this analysis aimed to determine the impact of climate change. We will concentrate on data from 1990 to 2020, which spans a 30-year period and covers different nations. With the appropriate visualizations, analysis will be done to compare the relationship between various features covering nations and other comparisons.

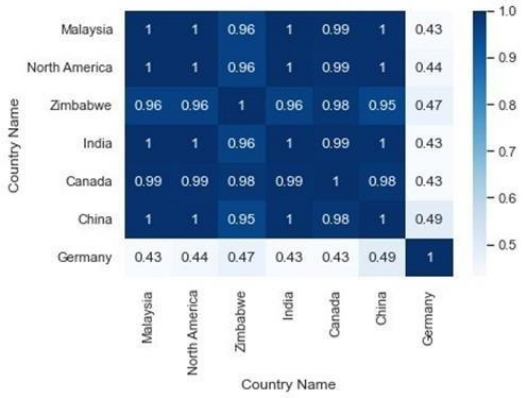
**GitHub Link**: https://github.com/HarshiniNetha/ADS-Rework-Assignment

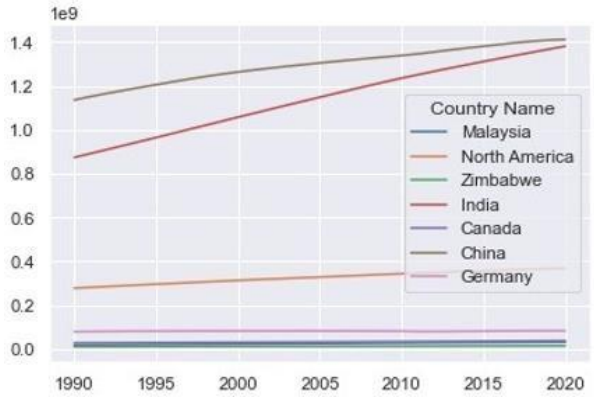
## *Report*

This report describes the findings of the analysis of the data which is taken from the climate change World Bank repository.

The data which has been taken from the repository contains a CSV file which has 20216rows and 65 columns. The columns are: Country Name, Country Code, Indicator Name, Indicator Code and 61 columns containing years 1960 to 2020. There are 266 country names and country codes for the corresponding countries. There are 76 indicator names in total for each country. The country names include countries such as Aruba, Malaysia, North America, Guyana, Zimbabwe, India, China, Oman, and Honduras. Indicator names include urban population, total population, forest land, arable land, renewable electricity output, total greenhouse gas emissions.

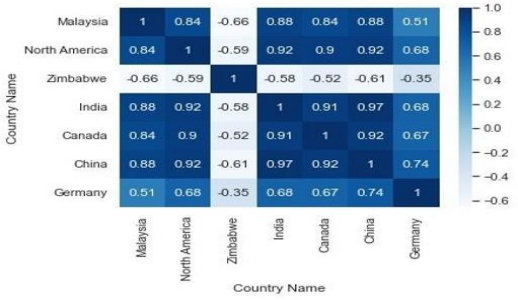
In the next part, the various analysis results are described. We first started with the indicator name "Population, total". Also we removed columns which are not needed: country code, indicator name, indicator code and the columns for the years 1960-1989. We are analyzing data from 1990-2020. For each of the indicator names whose analysis is done, we take into account 7 countries: Malaysia, North America, Zimbabwe, India, Canada, China and Germany. Since the data frame has the years as columns, the data frame was transposed to change the years as rows and the country names as columns. After that the **datarame.describe()** function was used which returns the statistical results such as count, mean, standard deviation, min, max, 25%, 50% and 75%. These results are returned for each county. From the results we found that the highest mean value for total population isfrom China and the lowest from Zimbabwe. The standard deviation is lowest for Germanyand highest for India. After that we calculatedthe correlation among all the countries and plotted the correlation matrix. It was found that Germany has the least correlation with all other countries. It can also be seen over the years, both China and India has a gradual increase in the total population compared to the other countries.

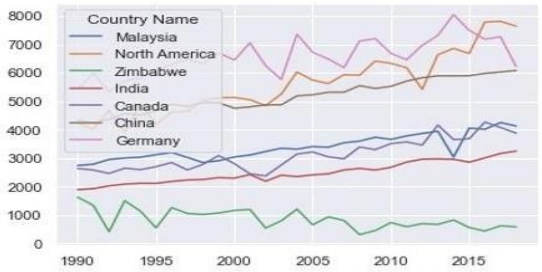



For the next indicator name "CO2 emissions from liquid fuel consumption (kt)", the highestmean is for North America and the lowest is for Zimbabwe. Even though India has a bigger population than North America, the overall CO2 emissions is lower. The plot for the countries over all the years shows how much CO2 emissions are generated from North America with India in the 3rd place.

For the indicator "cereal yield (kg per hectare)", we find that the highest mean is Germany followed by China. The lowest mean is Zimbabwe. The highest standard deviation is North America with the lowest from Zimbabwe.

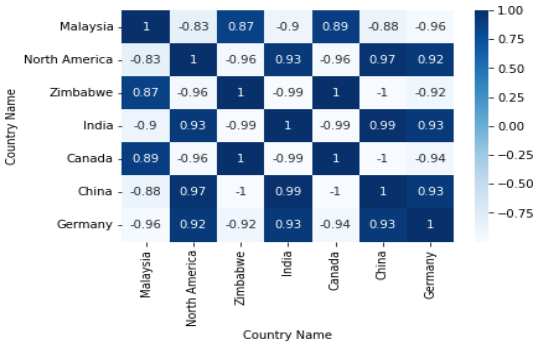


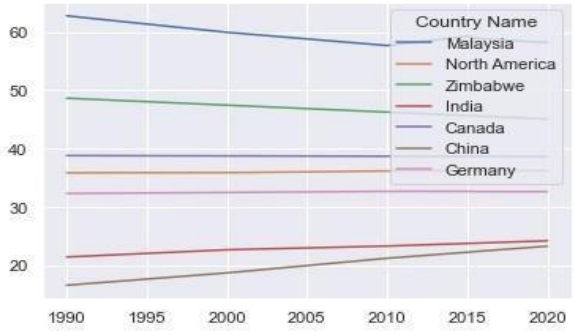




In the case of "Arable land (% of land area)", the plot for all countries over the years shows that India has the highest percentage of arable land followed by Germany. Malaysia has the lowest percentage of arable land.

The plot for the indicator "forest area (% of land area)" for all the countries over the years shows that Malaysia has the highest percentage of forest area with Zimbabwe being the second highest. However, there is a decrease in the values over the years whereas both India and China are having an increase in forest area.


The next part of the report deals with the correlation between three different indicators for the 7 countries. The first one is the correlation between arable land and forest land of each country. From the results, it can be seen that apart from Malaysia, Canada and Germany, the rest of the countries has a negative correlation. This shows that increase in one variable leads to decrease in the other variable. This shows that in those countries there is deforestation going on with the highest value in the case of India. Increasing arable land leads to decrease in the forest area.




The next is the correlation between total population and cereal yield for all countries. It can be seen that Zimbabwe has a negative correlation. This shows that there is shortage of cereal yield for the population. The rest of the countries have a positive correlation value suggesting that if total population increases, there will be increase in proper cereal yield.

The last correlation is between total population and CO2 emissions from liquid fuel. It can be seen that all the countries has a positive correlation value. This means that increase in population will also mean increase in CO2 emissions.