THE CURIOSITY CUP 2022

A Global SAS® Student Competition

A Review of Impact of Greenhouse Gas Emissions

Group Name - SASPICIOUS

Abstract

There has been solid evidence about climate change and its effect. To make significant changes, however, we need to pay attention to countries and sectors that matter. Our team sourced 2 climate data from Our World In Data and Worldbank and conducted exploratory data analysis such as time series analysis, clustering, forecasting to find out different clusters of countries. There is also another time series forecasting on the sector to find out which sector should we pay more attention to.

Introduction

The United Nations Climate Change Conference in Glasgow (COP26) which was held from November 1 to November 12, 2021 brought together 120 international leaders and nearly 40,000 registered participants. In the conference, countries reaffirmed their commitment to the Paris Agreement objective of keeping global average temperature rise below 2°C above pre-industrial levels, with efforts to keep it below 1.5°C. The representatives expressed their deepest concern and emphasized that human activities have caused roughly 1.1°C of warming to date, and that repercussions are already being felt. Most countries also agreed that this issue has to be highly valued and committed to solve environmental problems and climate warming problems.

With this as a context, the team is inspired to conduct research to understand the impact of the greenhouse effect to the environment. greenhouse effects are caused by the presence of water vapor, carbon dioxide, methane, and certain other gasses in the air. For the scope of this research, the team is focusing on carbon dioxide and the team is conducting a few analyses to explore the effect of carbon dioxide towards the greenhouse effect and the environment. The team is using two main datasets from Worldbank and Our World in Data. The details of the datasets will be discussed in the next section. For this research, the team will focus on clustering the countries based on various aspects and identifying the countries and sectors that should be paying attention. The tools that are used by the team include SAS studio on demand, Python and Tableau.

Data

There are two datasets used for this research. The primary dataset comes from The World Bank. This dataset contains environmental data as well as other greenhouse gasses emissions data from 1960 to 2018. This dataset has a total of 266 regions (including continent and other organizations that countries joined) and 76 dimensions (attributes).

Another dataset is Ghg Emissions by Sector which is maintained by Our World in Data. This dataset is available from the Our World in Data website. In this dataset, it contains co2 and greenhouse gas emissions data for each sector of the countries from 1990 to 2016. This dataset will be mainly used for studying sector related CO2 emissions analysis.

Problem

1. What are the countries that play a big role in climate change across time?
2. What countries contribute to high CO2 emissions?
3. What are the key factors that affect CO2 emissions?
4. What are the sectors that caused the top countries in CO2 emissions ranking emitted high CO2 emissions?

Data Cleaning/Validation

The missing values issue is severe in both source datasets. As shown in figure1, half of the columns of our worldbank dataset contain more than 50% missing values over the total observations. The team has decided to drop the columns that contain more than 50% missing data.

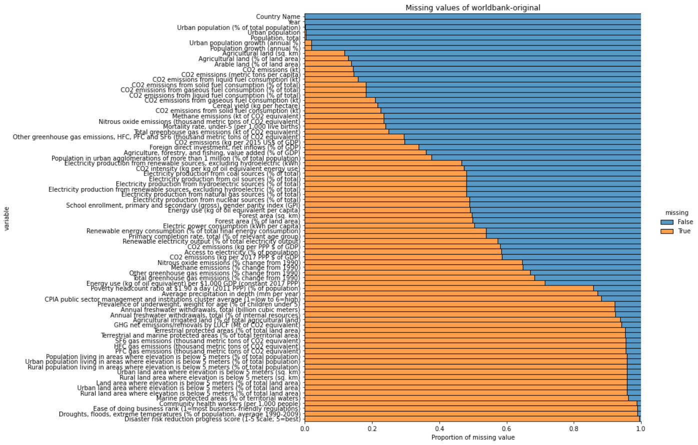
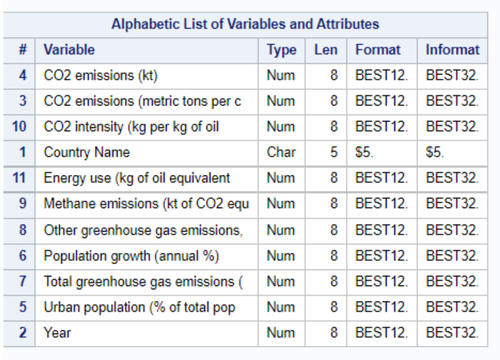


Figure 1: Barchart of Missing Values Percentage in Worldbank Dataset

(The orange color represents the missing value percentage per each column.)

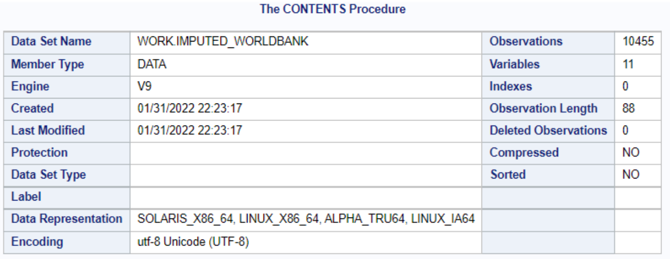
Following this, the team then uses Boruta feature selection algorithm to further reduce the dimension of the dataset and choose the most reasonable columns for future data analysis. After the team chose the most relevant attributes, the team used multivariate imputation to fill in the missing values with a machine learning algorithm. The team has also done the validation to make sure the values the imputation fill in is reasonable. The content and the list of variables and attributes of the imputed dataset is shown in figure 2

Figure 2: Overview of the imputed world bank dataset



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
4	CO2 emissions (kt)	Num	8	BEST12.	BEST32.
3	CO2 emissions (metric tons per c	Num	8	BEST12.	BEST32.
10	CO2 intensity (kg per kg of oil	Num	8	BEST12.	BEST32.
1	Country Name	Char	5	$5.	$5.
11	Energy use (kg of oil equivalent	Num	8	BEST12.	BEST32.
9	Methane emissions (kt of CO2 equ	Num	8	BEST12.	BEST32.
8	Other greenhouse gas emissions,	Num	8	BEST12.	BEST32.
6	Population growth (annual %)	Num	8	BEST12.	BEST32.
7	Total greenhouse gas emissions (	Num	8	BEST12.	BEST32.
5	Urban population (% of total pop	Num	8	BEST12.	BEST32.
2	Year	Num	8	BEST12.	BEST32.



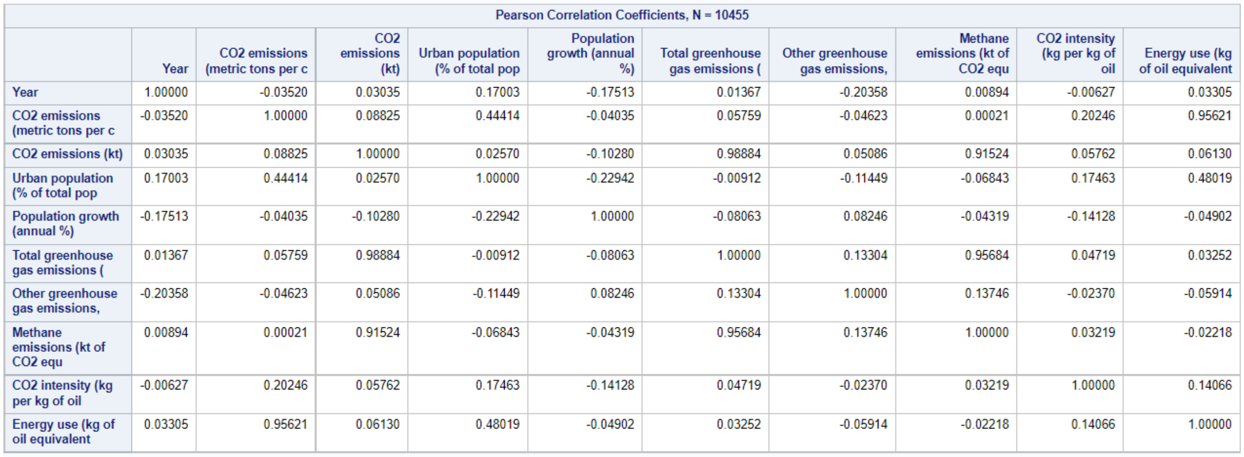
The CONTENTS Procedure

Data Set Name	WORK.IMPUTED_WORLDBANK	Observations	10455
Member Type	DATA	Variables	11
Engine	V9	Indexes	0
Created	01/31/2022 22:23:17	Observation Length	88
Last Modified	01/31/2022 22:23:17	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Analysis

The SAS tool that was used for this analysis task is SAS online studio from SAS® OnDemand for Academics.

Correlation Matrix



Pearson Correlation Coefficients, N = 10455

	Year	CO2 emissions (metric tons per c	CO2 emissions (kt)	Urban population (% of total pop	Population growth (annual %)	Total greenhouse gas emissions (	Other greenhouse gas emissions,	Methane emissions (kt of CO2 equ	CO2 intensity (kg per kg of oil	Energy use (kg of oil equivalent
Year	1.00000	-0.03520	0.03035	0.17003	-0.17513	0.01367	-0.20358	0.00894	-0.00627	0.03305
CO2 emissions (metric tons per c	-0.03520	1.00000	0.08825	0.44414	-0.04035	0.05759	-0.04623	0.00021	0.20246	0.95621
CO2 emissions (kt)	0.03035	0.08825	1.00000	0.02570	-0.10280	0.98884	0.05086	0.91524	0.05762	0.06130
Urban population (% of total pop	0.17003	0.44414	0.02570	1.00000	-0.22942	-0.00912	-0.11449	-0.06843	0.17463	0.48019
Population growth (annual %)	-0.17513	-0.04035	-0.10280	-0.22942	1.00000	-0.08063	0.08246	-0.04319	-0.14128	-0.04902
Total greenhouse gas emissions (	0.01367	0.05759	0.98884	-0.00912	-0.08063	1.00000	0.13304	0.95684	0.04719	0.03252
Other greenhouse gas emissions,	-0.20358	-0.04623	0.05086	-0.11449	0.08246	0.13304	1.00000	0.13746	-0.02370	-0.05914
Methane emissions (kt of CO2 equ	0.00894	0.00021	0.91524	-0.06843	-0.04319	0.95684	0.13746	1.00000	0.03219	-0.02218
CO2 intensity (kg per kg of oil	-0.00627	0.20246	0.05762	0.17463	-0.14128	0.04719	-0.02370	0.03219	1.00000	0.14066
Energy use (kg of oil equivalent	0.03305	0.95621	0.06130	0.48019	-0.04902	0.03252	-0.05914	-0.02218	0.14066	1.00000

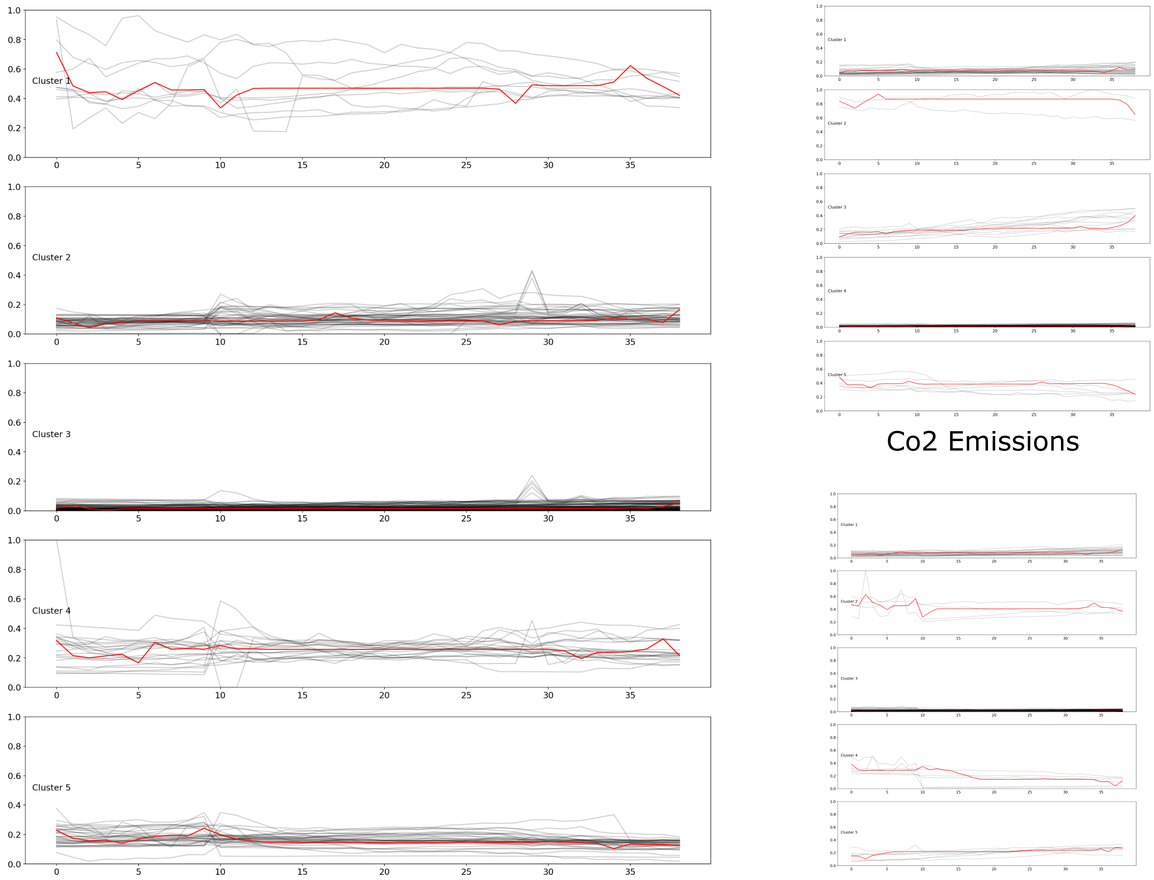
Figure 3: Correlation Matrix of the imputed world bank dataset

The Pearson Correlation Coefficient is used to understand the relationship between two variables that are measured on the same interval or ratio scale in our imputed dataset. The Pearson Correlation Coefficient ranges from -1 to 1, and this measures the strength of the association between two continuous variables. It was found that Total greenhouse gas emissions and Methane emissions (kt of CO2 equivalent) have a strong positive correlation towards CO2 emissions (kt) with a coefficient value of 0.98884 and 0.91524. For CO2 emissions (metric tons per capita), there is only one attribute, Energy use (kg of oil equivalent per capita) showing high correlation with a coefficient value of 0.95621. Based on this analysis, we could conclude that CO2 emission of a country may be influenced by different sectors and issues. In order to identify which sectors and countries may pay attention to, clustering and further sector analysis will be conducted.

Visualization

Clustering

Different countries require different solutions for climate issues. Clustering countries based on their historical results could better design tailored solutions for them. Algorithm used for this part is Dynamic Time Warping Distance Metric for Time Series Clustering. The k chosen is 5 based on previous research (Alcántara et al., 2006). Additionally, 6 more countries are removed before feeding into this clustering because their data can be stand alone from the rest of the cluster. They are China, the United States, India, and Aruba, Qatar, and the Russian Federation.



Co2 Emissions

Co2 Emissions Per Capita

Greenhouse Gasses

Figure 4: Time Series Clustering Results

Based on figure 4, co2 emissions per capita cluster 1 and cluster 2 have the most significant differences. The top cluster contains a total of 9 countries which includes the United Arab Emirates, Australia, Bahrain, Brunei Darussalam, Canada, Kuwait, Luxembourg, Saudi Arabia, Trinidad and Tobago. And the bottom cluster contains 43 countries. It is still not clear what the common ground is amongst all the countries in the feature yet. This part will leave for future investigation.

Sector Analysis and Forecasting

There are many factors that cause the rise of greenhouse gasses. In order to understand which sectors are responsible for the rapid growth of greenhouse gasses, the team carried out a visualization analysis of the greenhouse gasses emitted by different sectors around the world, and found that the top three sectors of greenhouse gas emissions are electricity & heat, transport and manufacturing/construction energy. The visualization is available under the appendix.

Among all the countries around the world, the United States has been the country with highest cumulative CO2 emissions since the year 1750 and now is still the country with the second highest annual CO2 emission according to a report from the Union of Concerned scientists, it is also one of the most "stand out" countries from our previous clustering analysis. This indicates that the United States plays a significant role in climate change and the carbon reduction policies adopted by the United States will affect the global greenhouse effect to a certain degree. The result also tally with the world sector trend in terms of greenhouse gas emissions. The team further forecasts the sectors and the result is shown in figure 5. The team hereby conclude that the United States and worldwide should revise their sector, especially energy to reduce the greenhouse gasses emissions.

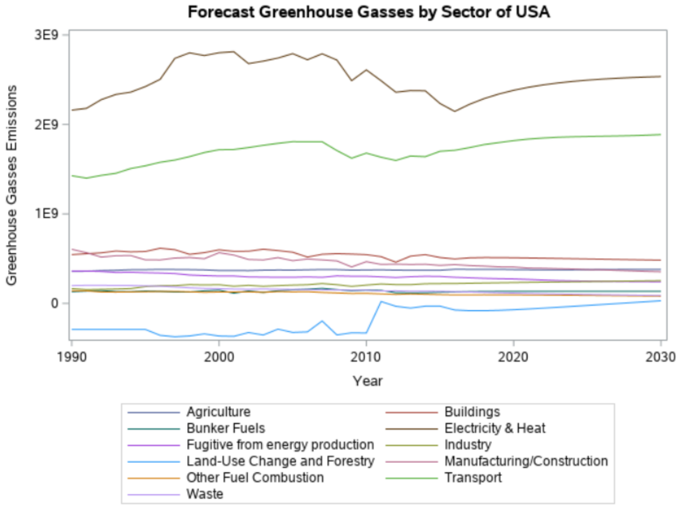


Figure 5: Forecast Greenhouse Gasses by Sector of USA

Suggestions for Future Studies

As this research is only focused on understanding the CO2 situation in various countries, there is no detailed research on the impact of CO2 on other fields of countries or the impact of other fields in countries on climate warming. In the future the team will consider diving deeper by mining more data, connecting them to CO2, trying to find out the impact of greenhouse gasses on the world, understanding its trends and the effectiveness of current carbon reduction policies.

Conclusion

The team highlighted that the methane emissions and carbon dioxide emissions are key factors to the rise of greenhouse gas emissions. The team also found that energy use per capita also highly correlated with the carbon dioxide emissions per capita which indicated that the energy system is the main contributor to the greenhouse gas emissions. In the following analysis, the team carried out a few clustering to cluster the countries based on carbon dioxide emissions and the team figured out the countries can be grouped into five clusters according to the pattern of rise of carbon dioxide in their countries. In order to know which sectors require urgent attention to reduce the greenhouse gas emissions, the team also carried out an analysis on the carbon dioxide by sector. The United States has been selected as the country to further investigate the sector wise impact on greenhouse gas emissions. The study revealed that the United States has almost a similar pattern as the world trend, their top sectors of highest greenhouse gas emissions are the same, which are electricity & heat, transport manufacturing energy and this also proves the hypothesis the team made earlier that the energy system is the main cause of the greenhouse effect.

References

Article in conference proceedings Alcántara, V., Duarte, R., & Obis Artal, M. 2006. “Regional decomposition of CO2 emissions in the world: a cluster analysis.” *Working papers (Universitat Autònoma de Barcelona. Departament d'Economia Aplicada)* Available at https://ddd.uab.cat/pub/estudis/2006/hdl_2072_2097/wpdea0306.pdf

APPENDIX

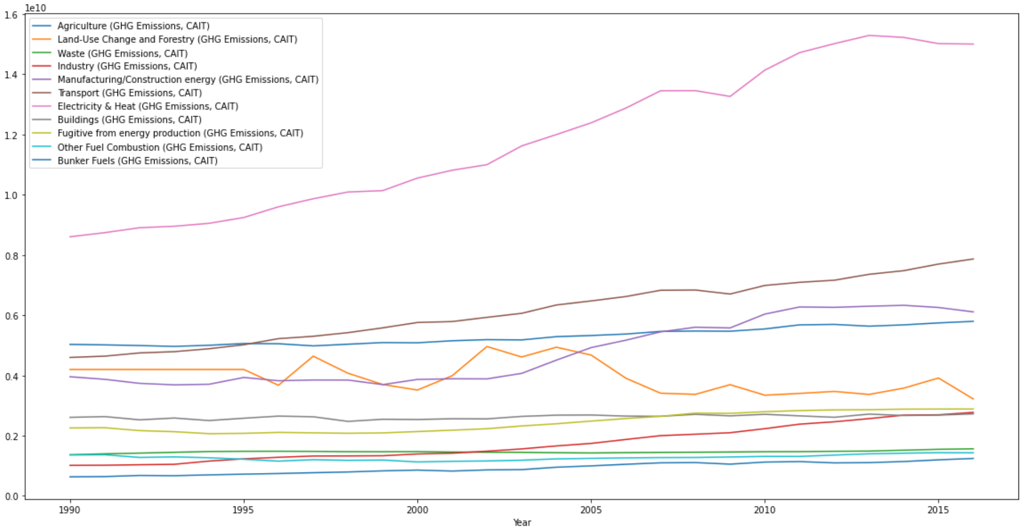


Figure: Global Sector Trend

Total CO2 Emission (Metric Tons per Capita) Around the Globe 1980-2020

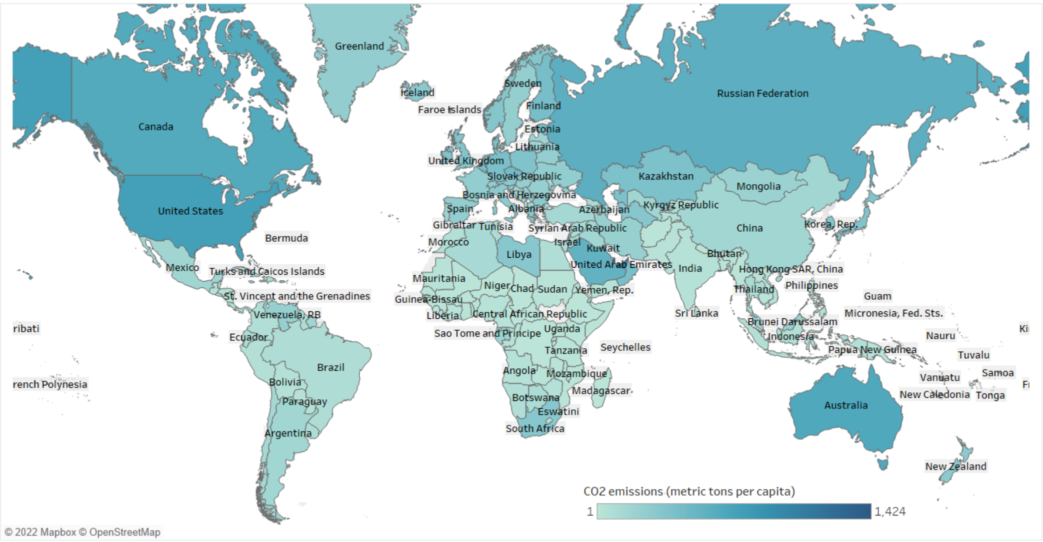


Figure: Map Visualization of Total CO2 Emission (Metric Tons per Capita) Around the Globe 1980-2020

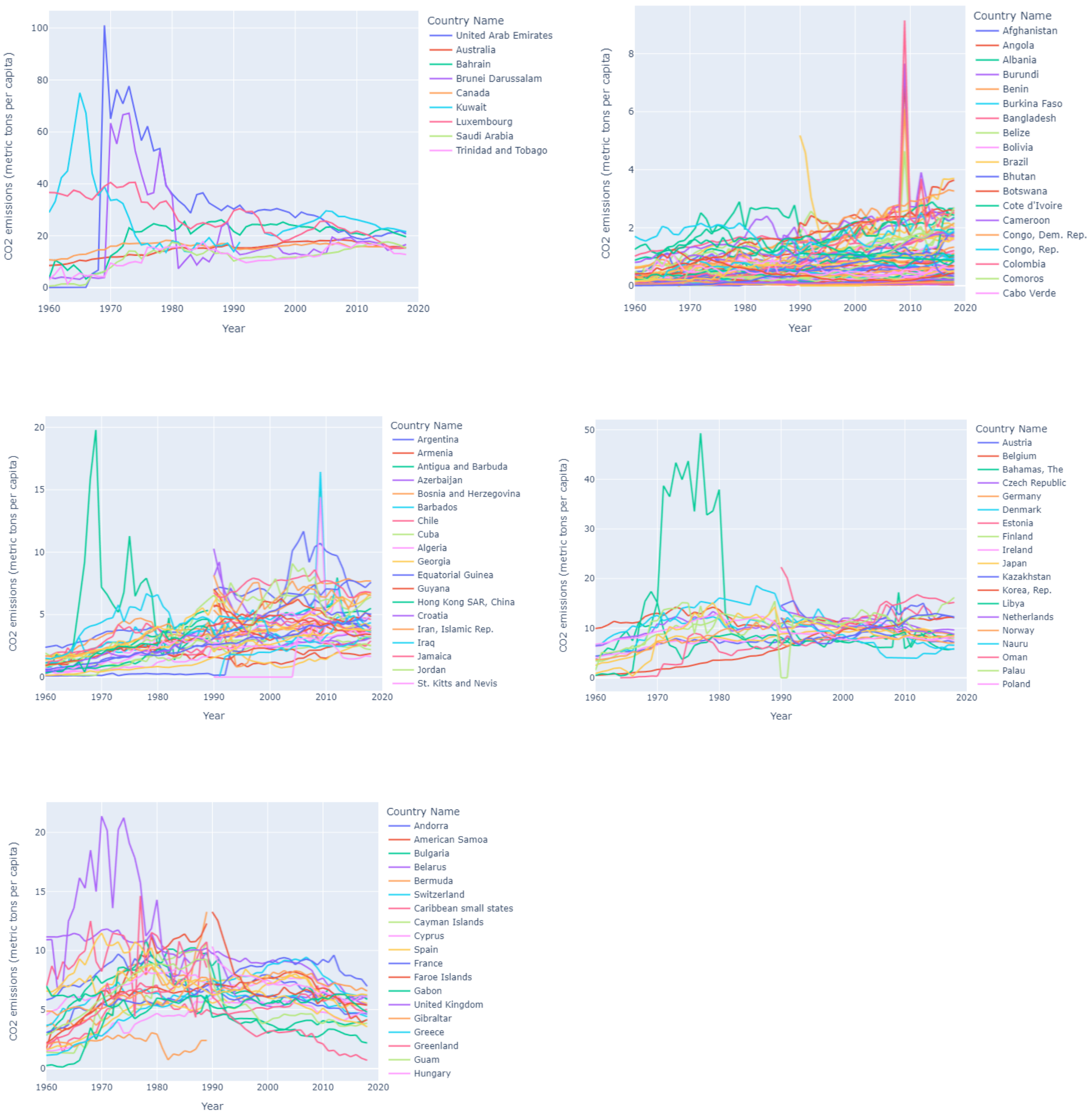


Figure: Time Series CO2 Emission Per Capita Clustering