The Influence of Economic Indicator and Human Development on Carbon Dioxide Emission

*CISC 5450 Final Project

Sheng Yun
Graduate School of Arts and Sciences
Fordham University
New York, USA
syun13@fordham.edu

Jingyan Xu
Graduate School of Arts and Sciences
Fordham University
New York, USA
jxu246@fordham.edu

***Abstract*—The objective of this research is to analyze the effect of Gross Domestic Product output and the Human Development Index against CO2 emissions in many countries across the globe over the years. In this study, we used a variety of methods, including time series and statistical modeling. We also constructed graphs that show the relationship between the variables. The results of this study found that GDP output effect was positive and strongly correlated to CO2 while HDI positive and insignificant effect against CO2 over the years.**

***Keywords*—CO_2 Emission, GDP, HDI, Time series, correlations**

I. Introduction

There are multiple factors that affect the environmental protection efforts, such as economic growth, technology, culture, and industrial policies. For instance, CO2 emission is one of the major factors that lead to global warming and environmental pollution, but it is tied to energy consumption and production. When we are too fixated on economic growth, we inevitably ignore environmental management, which in the end will do more harm than good to society as well. Goods and energy production, which are major components in the economy, requires a lot of energy consumption. Burning fossil fuels remain the single largest contributor to the greenhouse gas [1] but they are needed everywhere in our economy because heating households and producing electricity requires burning fossil fuel. The higher the demand in heat and electricity, the more fossil fuels required to produce energies.

On the other hand, economic growth and the quality of the environment also significantly impact the level of human development. It is true that economies with higher output have higher human developments, the growth also pollutes the environment. Environmental issues do not exist only in developed or least developed countries; it is a world problem. The decline in environmental quality leads to natural disasters like mass forest fires and an increase in the number of major hurricanes, and homes and lives were lost. Based on the background, we can formulate the two problems we want to investigate in this research:

1. How does the GDP affect CO2 emissions?
2. What is the effect of HDI on CO2 emissions?

II. Dataset sources and details

In our research, three datasets are used and all of them are retrieved from Our World in Data. The three datasets are the CO2 emission dataset, the national GDP dataset, and Human Development Index dataset. Each dataset has different number of entities or countries, and time range of data, therefore making join operations of datasets necessary.

Furthermore, in order to categorize countries in our dataset into different regions, we utilized the dataset from the World Bank [2] as it categorizes countries into different regions. For example, in the metadata file, the United States is categorized as a North American nation, and Mexico is categorized as a Latin American nation. We joined the four datasets together to clean all the missing values and to make sure we have all the information available.

A. National GDP

The national GDP dataset [3] has 4 attributes: Entity (country name), Code (a unique three-letter code representing the country), Year, and GDP. We count that there are 183 countries included in the GDP dataset, and the time ranges from 1950 to 2019. Not all countries have the same relative frequency occurrence in the dataset, which means there exist null values for multiple countries. The GDP column measures the economic outputs across different countries throughout history by using import, export, and prices of final products. It is adjusted for inflation and converted into the same currency (US Dollars) by using ICP PPP. The latest update of the dataset dates back to 2021, and it was published by Robert Feenstra, Robert Inklaar, and Marcel P. Timmer, three economics academic scholars [4].

It is important to mention that GDP is by far one of the best measurement to tell the well-being of an economy. There are a few ways to view GDP, one is that it's the total income of all the people in the economy, another way is the total expenditure on the output of goods and services of the economy. The underlying idea is that it measures people's incomes which are what people cares about. Income and expenditure can largely be seen as the same thing because income must equal expenditure. The fundamental concept states that there are

always a buyer and a seller in a trasaction and that every dollar of expenditure by a buyer will become a dollar of income to the seller. GDP is a good indicator of the health of the economy because it shows whether the output of goods and services can better satisfy the demands of entities in an economy, like households or the government.

B. CO_2 Emission

The CO_2 emission dataset [5] has 4 attributes: Entity, Code, Year, and annual CO_2 emissions. In this dataset, the time spans until 2021. Similarly, there are many countries that don't record their CO_2 level every year so there are missing values. The dataset measures annual production-based emissions of CO_2 across time and countries in tonnes. It was published by Global Carbon Project in 2022 [6].

There are many types of greenhouse gases and here we only used the CO_2 emission dataset to represent the environmental impact because we deemed CO_2 emission should be a good representation of greenhouse gas emissions. Greenhouse gas emissions strengthen greenhouse effect which leads to climate change, and CO_2 emissions account for overwhelming majority of total greenhouse gas emissions per year, for about 72% [7].

C. Human Development Index (HDI)

The HDI dataset [8] has 4 attributes: Entity, Code, Year, and HDI. There are 190 countries included in this dataset, and the time ranges from 1990 to 2021. HDI is a composite measure of a country's performance in three basic human development concepts, life expectancy at birth which measures whether citizens can have a long and healthy life, average and expected years of schooling which measures a country's ability to provide access to knowledge, GNI per capita which measures the standard of living in a country. HDI is computed with the geometric mean of normalized indices of these three measures. The United Nations Development Programme first published this computed data in 1990 [9].

III. Relative and Previous Work done by Others on the Same Dataset

Carbon intensity is a key measurement to determine the environmental-economic balance. It is a measurement of the quantity of CO2 emitted per unit of GDP. A high carbon intensity implies that there is a high level of CO2 emissions relative to the size of the economy. Studies have shown that carbon intensity had reached its peak back in 1951, but it also notes a significant abnormality, which is that during 1950-1980, the carbon intensity in China nearly tripled and maintained twice the global average ever since, while the rest of the world has seen a gradual downward trend since the peak was reached. The study carefully examined China's industrial policies during the Great Leap Forward, the time period of political instability called the Cultural Revolution, and later the Economic Reform. The study found that despite the temporary drop in GDP, the carbon intensity remained high because the CO2 emission level remained the same, and later the decline in carbon intensity in China was largely due to new technologies that improved energy efficiency.

Another study [11] [12] shows that the decoupling of economic growth and energy consumption exists in many countries. In many rich countries like Germany, Denmark, the UK, or Sweden, CO2 emission levels have largely maintained the same or even declined while the GDP outputs have been increasing. Countries like the US also begin to see similar trends that decoupling has started. One reason why this phenomenon occurs is that these countries are replacing fossil fuels with low-carbon energy. The decline in the cost of clean energy technologies has made the effort to address the greenhouse gas emission problem faster.

Education outcomes and GDP per capita are two of the three pillars used when computing HDI. There are studies that have tried to map out a relationship between education outcomes and GDP per capita. A study [13] shows that richer countries tend to produce better education outcomes, but differences are large even among countries with similar income per capita. Another study shows that while income is an important factor that affects expenditure on education and education outcomes when the income level is above a certain threshold, the relationship between education outcome and income level becomes nonexistent.

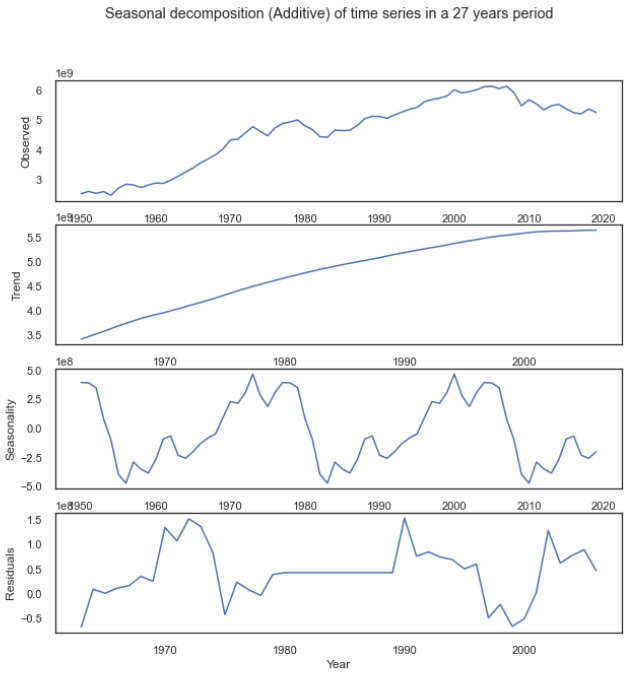


Fig. 1: Time Series analysis for US CO_2 Emission from 1950 to 2019

IV. Results and Interpretation

This report will focus on the USA greenhouse gas emission data and the relationship with the economic performance, human development index, and a possible yearly seasonal

trend. Related work has demonstrated a relatively strong relationship between CO_2 emission and GDP growth over the year since 1950 for an individual country or geographical region. However, there is a gap for a study investigating the time series for CO_2 emission and the correlations between the emission, economic performance, and human development level since World War II in the United States. It will be worth investigating whether the economic fluctuation is caused by CO_2 emission or the other way around; furthermore, whether the CO_2 emission drop will indicate the change in the human development index. More specifically, we are focusing on whether CO_2 emission drops will indicate an economic recession, whether this will cause a human development index change, and if the USA has a relatively better performance in recovering from the recession and performing the emission reduction plans.

In this study, as we mentioned in the previous introduction, we will demonstrate the data since 1950. We will provide information on economic performance, the human development index, and Greenhouse gases emission. Necessary Graphs and plots will be demonstrated, and partial code will be shown in the **Appendix**.

A. Time Series for USA's CO_2 Emission (1950 - 2019)

The first question we are interested in is the trend of the US CO_2 emission after 1950. To demonstrate this, we plot the time series for the emission of CO_2 in US with the Matplotlib and the Streamlit tools. We first used Matplotlib to plot the time series; however, Matplotlib only supports limited access to the data. We can see the increasing trend of the 's CO_2 over the years, but it is challenging to spot the subtle changes between the years. Thus, we used the second approach, the embedded tool from the streamlit library(**Appendix I**). It provides real-time zoom-in and zoom-out, making it easy to see the value of any specific year.

Though we can see a growing trend and a possible similar periodic pattern in 1953-1981 and 1981-2008, we cannot see a more particular periodic pattern in the plot. To better illustrate the periodic patterns we spot, we plot the trend, seasonality, and residuals for a frequency of 27 years of seasonality decomposition(Fig. 1). Since we realize that the seasonality we think that may exist did not increase as the year increase, so we are performing the additive time series. Then, similarly, we see a pattern in the seasonality plot shown in 14 years(**Appendix I**). These show that there is seasonality in the data, but the pattern shown in the data is in a 27 years span or 14 years span. The necessity of performing time series forecasting needs more extensive investigation at this stage since there may be more relationship inside of this data. For the convenience of further investigation, we will also prepare a time series plot for the GDP in the USA on the same scale(**Appendix I**). However, the GDP data seems to not follow the same seasonality.

B. Cumulative Plot & Rate of Change

After having the CO_2 time-series plot, we can claim that the CO_2 emission of the USA is an overall increase over the years,

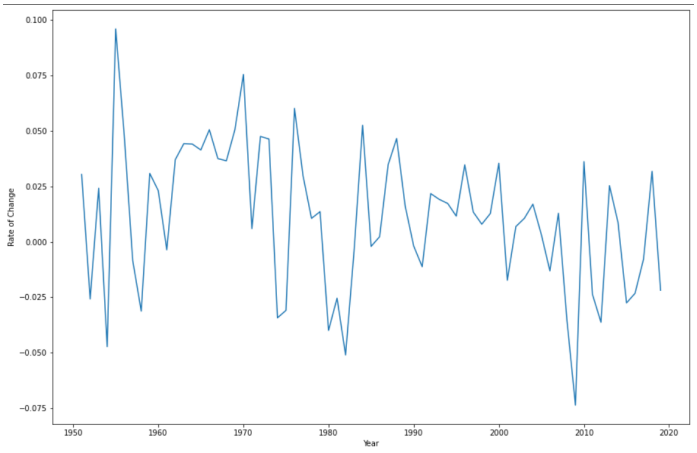


Fig. 2: Rate of Change of US CO_2 Emission from 1950 to 2019

but how much the greenhouse gas emission has accumulated over the year still needs to be discovered. Thus, we plotted the cumulative plot of the CO_2 from 1950 to 2019(**Appendix I**) and the rate of CO_2 change over the years (Fig. 2). With the plots we mentioned above, we can see the cumulative trend of CO_2 in the US and the rate of CO_2 emission change in the US. We do not see a particular pattern in the rate of CO_2 emission change plot, but we can still spot some of the spikes and crashes in the plot. For example, we can easily spot the only significant drop that indicates a 7% drop in the 2009 CO_2 emission, which may correspond to the financial crisis[16] or the 2009 Copenhagen climate conference[17](less likely since there is no clear plan result). Also, we can spot a significant spike in 1955 that has a gross rate of almost 10%. However, whether these dramatic changes were triggered by economic spikes or human development is questionable.

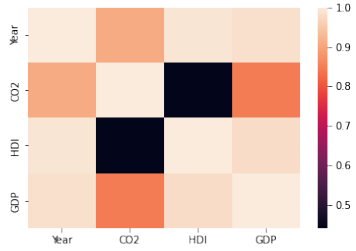
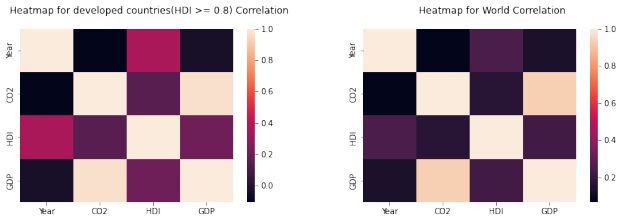


Fig. 3: Heatmap for USA Data Correlations

C. Features Correlations

Then we want to know the relationship between the GDP, CO_2 emission, and HDI. We plot a heat map of the US data (Fig. 3). For reference, we also plot the heat map for the World data. After analyzing the heatmap, we can claim a relatively large relationship between CO_2 emission and GDP, but the relationship between CO_2 emission and the HDI is ignorable. The most substantial relationship we can see is between HDI



(a) 1a (b) 1b

Fig. 4: Benchmarks for Correlations

and GDP but this is because the calculations for HDI strongly involve the GDP per capita. This gave us confidence to claim that there is some relationship between the CO_2 emission and the economic development in the US but we still need more evidence and context to claim the relationship.

One fact we need to validate is whether the relationship between CO_2 and GDP is a universal relationship or a characteristic of the US. We plot the overall correlation heatmap for the world data (Fig. 4(a)) and (HDI ≥ 0.8) developed countries data (Fig. 4(b)) to investigate this. From the plots, we found similar levels of correlations. We also plot the HDI data in all the countries in America within a bar chart(**Appendix I**), but we cannot see an apparent correlation. Thus, we will further investigate the relationship between CO_2 emission and the GDP in USA data since it is universal that these two features are correlated.

D. CO_2 Emission in Regions

One possibility is that countries in the same region may have a similar way of developing, which may conclude a similar amount of emission and rate of change overall. We first plot the Barchart of neighboring countries(North America: Bermuda, Canada, USA) for averages of CO_2 emission on the same scale (Fig. 5). Then we realize that the USA has a way more substantial CO_2 emission compared to the other two countries in North America, according to the dataset. Thus, we decide to align the rates of the changes(**Appendix I**) to see if there are any similar patterns in these plots. However, we can only spot that Canada seems to have similar spikes and drops with the USA as we mentioned previously, but the subtle changes in between are not showing a high-level similarity. Bermuda, on the other hand, has not had enough data to demonstrate similarity.

To better see if the amount of USA emission is substantial in America (Also if they share any pattern similarities), we also plot the graph for all the countries in America(**Appendix I**). We can conclude that the USA has a way more prominent CO_2 emission than any other country in America since it is well-known that the USA is the most powerful country and the most economically powerful country in the world. We would like to see if the GDP of the USA holds a similar statement. So we plot the graph of the GDP for all American countries(**Appendix I**), and we can see the overall plot is showing a high-level similarity. This gave us more confidence that CO_2 emission may be relevant to economic growth.

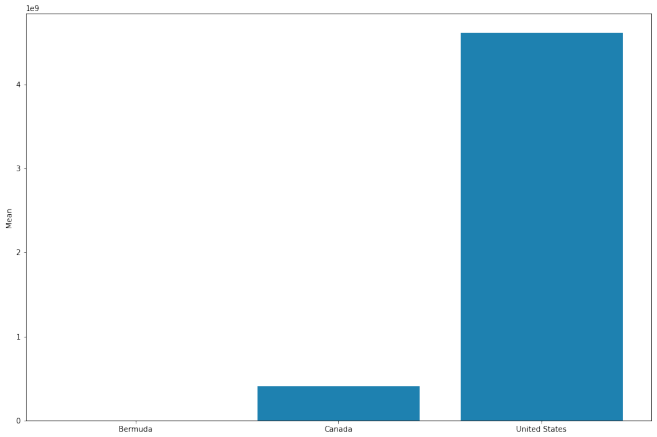


Fig. 5: Barchart of the CO_2 Emission in North America

E. Regionality

Although the correlation heatmaps in Subsection C show no significant relationship between HDI and CO_2 emission, we still want to check if there is any regionality in the relationship between HDI and CO_2 emission. We plot the scatter plots of HDI of all countries versus their CO_2 emission in the most recent available year in the dataset(2017) (Fig. 6). The regionality has a structure that is too complex to interpret(even with the outliers(like China and USA) removed from the plot(**Appendix I**)). We also plot a similar plot with the CO_2 emission data against the GDP data(**Appendix I**)

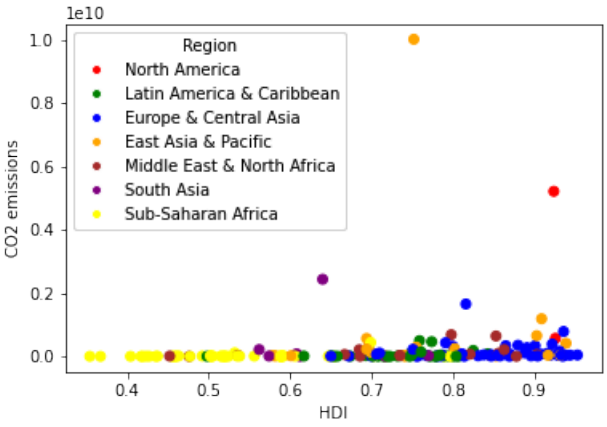


Fig. 6: Scatter Plots of HDI of all countries Versus their CO_2 Emission in the Most Recent available year

So far we have created and analyzed multiple plots on our data, and we have the following hypothesis

- Time series analysis plot for the USA CO_2 emission, and the corresponding 'Trend', 'Seasonality', and 'Residual Plots'. We found out that there is a strong possibility that the data for emission can follow a 27-year, or more precisely, a 14-year seasonality.
- The heatmap for analyzing the correlations in the USA data implies that there exists a strong correlation between CO_2 emission and GDP data.
- Other analysis (**Appendix**) may not bring us more relationship information, but they directed us to focus only on the CO_2 data.

F. Data Distribution and Statistical Characteristics

Before we go further, we would like to know better about our data, and we would like to know how the data distributed and what they are the extreme values. We plot the histogram to understand better the cardinality of the CO_2 emission data in the USA, and we gather the quantile and descriptive statistics to understand the data better.

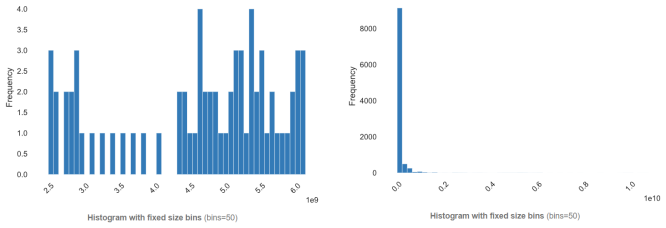
In the histogram, we can see the distribution of the USA CO_2 emission(Fig. 7(a)) is very different from the whole world data(Fig. 7(b)). This is because that some countries(such as China) have an extremely big value in the emission while some of the countries' data are very small. In the distribution of the USA, we cannot have any claims from the graph since the data are outlined similarly. So we further investigate quantile statistics and descriptive statistics(Also in **Appendix I, II**).

- We have the Quantile statistics in our USA CO_2 emission data:
  The minimum value in the data is 2,489,462,300; 5-th percentile is 2,615,554,435; Q1: 3,739,325,650; median is 4,857,628,850; Q3 value is 5,467,644,125; 95-th percentile falls in to 6,038,797,630; and the maximum value of the data is 6,137,603,600; The range of the whole data is 3,648,141,300; and the Interquartile range (IQR) is 1,728,318,475.
- Then we have the Descriptive statistics
  - Standard deviation: 1,130,548,933
  - Coefficient of variation (CV): 0.2452098968
  - Kurtosis: -0.9183033697
  - Mean: 4,610,535,496
  - Median Absolute Deviation (MAD): 680,121,150
  - Skewness: -0.5934579331
  - Sum: $3.227374847 \times 10^{11}$
  - Variance: $1.27814089 \times 10^{18}$
  - Monotonicity: Not monotonic

The above information will be useful in the following regression analysis.

V. Discussion and Pre-Modeling

In this section, we will discuss the possible directions the data modeling can go. Analysis of the characteristics of the data will be provided, and a basic model evaluation will be demonstrated. However, modeling, evaluations of time series, and prediction are beyond the scope of this study.



(a) USA Histogram (b) World Histogram

Fig. 7: CO_2 emission Histogram

A. Regression Analysis

In this section, we will utilize the automated machine learning web app(used PyCaret and Streamlit Libraries) created by one of the authors to perform the automated regression analysis on the CO_2 emission data in the USA. From the previous sections, we know that there is a strong relationship between greenhouse gas emissions and economic performance. Now we will do a regression analysis to quantify the relationship and find the best modeling method.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	84,023,409.8304	12,665,537,704,713,604.0000	101,435,704.0334	0.9851	0.0228	0.0188	0.0170
catboost	CatBoost Regressor	103,641,328.5012	17,612,568,652,923,588.0000	121,173,432.3259	0.9795	0.0258	0.0222	0.1630
gbr	Gradient Boosting Regressor	111,518,897.4447	18,890,956,748,618,876.0000	128,023,743.9187	0.9776	0.0276	0.0241	0.0080
rf	Random Forest Regressor	102,138,657.6256	18,238,519,000,370,156.0000	118,528,936.7128	0.9770	0.0257	0.0222	0.0190
knn	K Neighbors Regressor	142,848,877.2000	32,770,437,239,590,092.0000	164,575,166.8000	0.9624	0.0365	0.0321	0.0070
xgboost	Extreme Gradient Boosting	127,106,657.2000	25,561,101,302,444,852.0000	149,305,877.6000	0.9610	0.0321	0.0274	0.0190
dt	Decision Tree Regressor	136,303,646.7200	27,489,917,327,032,320.0000	158,181,197.4099	0.9578	0.0348	0.0297	0.0060
ada	AdaBoost Regressor	136,219,246.1496	28,399,456,601,392,104.0000	161,154,390.6113	0.9548	0.0344	0.0292	0.0100
lar	Least Angle Regression	340,487,281.0957	155,680,664,044,857,792.0000	385,933,955.8423	0.7550	0.0791	0.0715	0.0060
llar	Lasso Least Angle Regression	340,487,268.3809	155,680,663,873,940,032.0000	385,933,953.9429	0.7550	0.0791	0.0715	0.0050
lasso	Lasso Regression	340,487,281.0446	155,680,664,043,310,400.0000	385,933,955.8328	0.7550	0.0791	0.0715	0.1680
lr	Linear Regression	340,487,281.0957	155,680,664,044,859,616.0000	385,933,955.8423	0.7550	0.0791	0.0715	0.4940
ridge	Ridge Regression	340,326,059.5256	155,672,068,021,668,768.0000	385,900,870.2562	0.7549	0.0791	0.0714	0.0060
en	Elastic Net	337,164,028.4312	155,903,374,272,071,232.0000	385,818,120.7849	0.7534	0.0791	0.0706	0.0060
omp	Orthogonal Matching Pursuit	436,583,528.6502	248,197,314,292,452,544.0000	490,918,974.7277	0.5916	0.1117	0.1000	0.0060
br	Bayesian Ridge	554,352,716.3483	385,203,307,977,783,872.0000	614,567,570.5669	0.3990	0.1468	0.1334	0.0060
dummy	Dummy Regressor	882,872,172.8000	1,206,749,292,063,516,160.0000	1,063,875,856.0000	-0.2438	0.2557	0.2291	0.0050

Fig. 8: Barchart of the CO_2 emission in North America

The app mentioned above will automatically return the scoreboard of the performance of the state-of-the-art regression algorithms performed on the data. Since we are focusing on the relationship between two features, so we are dropping the irrelevant columns in the dataset. After performing the regression models on those two features, surprisingly, we found out there is not enough evidence to claim that these two features have a conventional regression relationship(**Appendix I** and **Appendix II**). We further plot the scatter plot for GDP VS. CO_2 emission, but the graph shows a bit of linearity, but it seems very periodic. We realize that we might need to model it in a higher dimension, or there are more features to form the regression. Higher dimensional regression modeling is beyond our study here. Furthermore, there is a possibility that certain trends got interrupted by uncertain events and certain policies like emission reduction plans. We do not hold enough information on how much impact and when it will

impact the data, so the analysis may need more relative data to support it.

However, we found out the regression model performs well when the year data is involved, so we decide to do the regression along with the year data. From the R2 value in the chart, we can see there is a strong regression relationship(Fig. 8) between the CO_2 emission, GDP and the Year(Time)($R^2 > 0.7$). Extensive plots analysis will be shown in the **Appendix**. Yet, the relationship we found in the correlation matrix in the previous section is still not resolved, and this apparently exceeds the scope of this work. We will claim that there is a strong correlation bet CO_2 emission and GDP, but the relationship may need more information to interpret.

B. Time Series Analysis

We will now perform the time series forecast on the data. Continuing from the previous discussion on the time series we recall the time series plot from Section III A. We found an apparent seasonal pattern for a 27-year period and a 14-year period. After a careful experiment of all possible year frequencies, we found out that 27 is the period of years the time series will have the most apparent trend and least produced residuals given that the relationship of the additive time series is as follows.

$$Y = Trend + Seasonality + Residuals$$

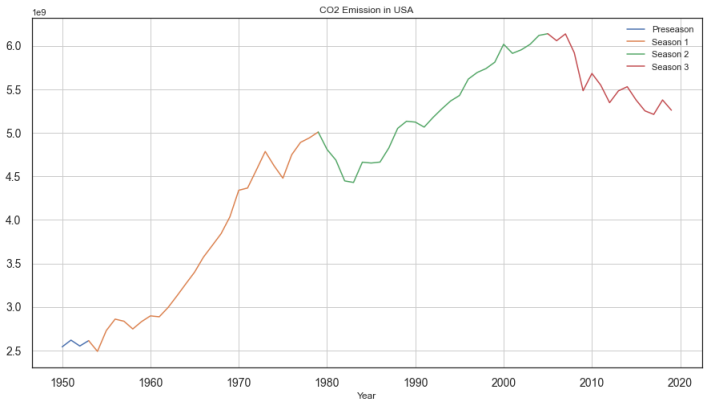


Fig. 9: Apparent Pattern in the time series

From a long-sighted perspective, we can claim that the USA CO_2 emission has a long-run increase followed by a sudden crash in a 27-year period, specifically, from 1953 to 1980, and from 1981 to 2008. More subtle seasonality is not as apparent as this trend. We can use this seasonality to forecast the future data of the CO_2 emission in the USA, but this work will not be in the scope of this work. To better demonstrate this seasonality, we provide a time series with different seasons labeled in the plot(Fig. 9).

We noticed that this seasonality is not ideally fit in the original data(though it is the best frequency we can get), and there might be a little bit shift in the recent trend. We think it might be possible that this seasonal trend is getting shorter in the time span, and will likely be more frequent(The frequency is decreasing). Thus, we might need further investigation on how much it will change and if this will be dominant. Also, we noticed that there was a drop after 2009, as previously mentioned, and the emission is not showing a positive relationship with the GDP. We think it is possible that the US found a way to grow economically without increasing emission, but this claim needs more information and future data.

VI. Conclusions

In this study, we can conclude that there exists a strong and positive correlation between GDP and CO2 emissions while the correlation between HDI and CO2 emissions is positive but ignorable. It actually confirms our understanding that economic growth largely drives CO2 emissions because increases in spending on goods and transportation require energy to function.

We also mentioned the seasonality effect in the positive association between GDP and CO2. The study about historical carbon intensity in China in Section 3 provides an important direction for any future research about the two variables. Ongoing economical recessions and industrial policies affect output and thereby influence CO2 emissions as well. We should factor in political and economical contexts in the research to try to explain this phenomenon as it can help us better understand the relationship.

We should also consider other datasets that are possibly more representative, for example, instead of national GDP, we can use the Gini coefficient or GDP per capita, as they can be more ideal to indicate the wellness of a country's economy because there are certain countries that have high GDP but low GDP per capita due to high population. We can also consider the Historical Index of Human Development (HIHD) [14] instead of HDI because one reason why CO2 and HDI show an ignorable relationship is that HDI tends to remain on the same level over the years. The HIHD changes more quickly because the indices are derived non-linearly. [15] One other thing we should consider is to further split the dataset into least developed, developing, and developed countries. We can see from the study about education outcomes vs. GDP per capita in Section 3 that rich countries tend to have better education outcome, but when the income per capita passes a certain level, the relationship between the two variables become ignorable. We may want to examine if similar logic can be applied to CO2 vs. GDP study because rich countries, albeit with higher economic output and demand, also have better technologies to combat climate change, while poorer countries don't have the luxury to pursue renewable energy technologies so that they are willing to sacrifice environmental management for better economic growth.

Acknowledgement

All the available plots and results screenshots during this study will be provided in the **Appendix I**, and the partial code for generating the plots will be demonstrated in a format of a python notebook in **Appendix II**. We would like to thank

the open-source library: Streamlit and PyCaret for providing easy access to multiple resources.

REFERENCES

[1] Sources of Greenhouse Gas Emissions. (2022, August 5). US Environmental Protection Agency. https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions

[2] The World Bank. (2022). GDP (current USD) [Dataset]. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

[3] National GDP. (2022, November 28). Our World in Data. https://ourworldindata.org/grapher/national-gdp

[4] Feenstra, R. C., Inklaar, R. and Timmer, M.P. (2015), "The Next Generation of the Penn World Table". American Economic Review, 105(10), 3150-3182

[5] Annual CO2 emissions. (2022, November 11). Our World in Data. https://ourworldindata.org/grapher/annual-co2-emissions-per-country

[6] Pierre Friedlingstein, Michael O'Sullivan, Matthew W. Jones, Robbie M. Andrew, Luke Gregor, Judith Hauck, Corinne Le Quéré, Ingrid T. Luijkx, Are Olsen, Glen P. Peters, Wouter Peters, Julia Pongratz, Clemens Schwingshackl, Stephen Sitch, Josep G. Canadell, Philippe Ciais, Robert B. Jackson, Simone R. Alin, Ramdane Alkama, Almut Arneth, Vivek K. Arora, Nicholas R. Bates, Meike Becker, Nicolas Bellouin, Henry C. Bittig, Laurent Bopp, Frédéric Chevallier, Louise P. Chini, Margot Cronin, Wiley Evans, Stefanie Falk, Richard A. Feely, Thomas Gasser, Marion Gehlen, Thanos Gkritzalis, Lucas Gloege, Giacomo Grassi, Nicolas Gruber, Özgür Gürses, Ian Harris, Matthew Hefner, Richard A. Houghton, George C. Hurtt, Yosuke Iida, Tatiana Ilyina, Atul K. Jain, Annika Jersild, Koji Kadono, Etsushi Kato, Daniel Kennedy, Kees Klein Goldewijk, Jürgen Knauer, Jan Ivar Korsbakken, Peter Landschützer, Nathalie Lefèvre, Keith Lindsay, Junjie Liu, Zhu Liu, Gregg Marland, Nicolas Mayot, Matthew J. McGrath, Nicolas Metzl, Natalie M. Monacci, David R. Munro, Shin-Ichiro Nakaoka, Yosuke Niwa, Kevin O'Brien, Tsuneo Ono, Paul I. Palmer, Naiqing Pan, Denis Pierrot, Katie Pocock, Benjamin Poulter, Laure Resplandy, Eddy Robertson, Christian Rödenbeck, Carmen Rodriguez, Thais M. Rosan, Jörg Schwinger, Roland Séférian, Jamie D. Shutler, Ingunn Skjelvan, Tobias Steinhoff, Qing Sun, Adrienne J. Sutton, Colm Sweeney, Shintaro Takao, Toste Tanhua, Pieter P. Tans, Xiangjun Tian, Hanqin Tian, Bronte Tilbrook, Hiroyuki Tsujino, Francesco Tubiello, Guido R. van der Werf, Anthony P. Walker, Rik Wanninkhof, Chris Whitehead, Anna Willstrand Wranne, Rebecca Wright,Wenping Yuan, Chao Yue, Xu Yue, Sönke Zaehle, Jiye Zeng, and Bo Zheng (2022), Earth System Science Data, 14, 4811–4900, 2022, DOI: 10.5194/essd-14-4811-2022.

[7] Olivier, J. G. J., Peters, J. A. H. W. (2020). Trends in global CO2 and total greenhouse gas emissions: 2020 Report (PBL publication number: 4331). PBL Netherlands Environmental Assessment Agency. https://www.pbl.nl/en/publications/trends-in-global-co2-and-total-greenhouse-gas-emissions-2020-report

[8] Human Development Index. (2022, November 29). Our World in Data. https://ourworldindata.org/grapher/human-development-index

[9] United Nation Development Programme. (2022). Human Development Index (HDI) [Dataset]. https://hdr.undp.org/data-center

[10] Ritchie, H. (2017, May 11). Carbon intensity in China's recent history – Politics matters a lot in achieving both prosperity and sustainability. Our World in Data. https://ourworldindata.org/chinese-turbulence-how-periods-of-political-reform-affect-the-carbon-intensity-of-economies

[11] Ritchie, H. (2021a, November 30). A number of countries have decoupled economic growth from energy use, even if we take offshored production into account. Our World in Data. https://ourworldindata.org/energy-gdp-decoupling

[12] Ritchie, H. (2021b, December 1). Many countries have decoupled economic growth from CO2 emissions, even if we take offshored production into account. Our World in Data. https://ourworldindata.org/co2-gdp-decoupling

[13] Roser, M., Ortiz-Ospina, E. (2016) - "Global Education". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/global-education'

[14] Historical Index of Human Development. (2018). Our World in Data. https://ourworldindata.org/grapher/human-development-index-escosura

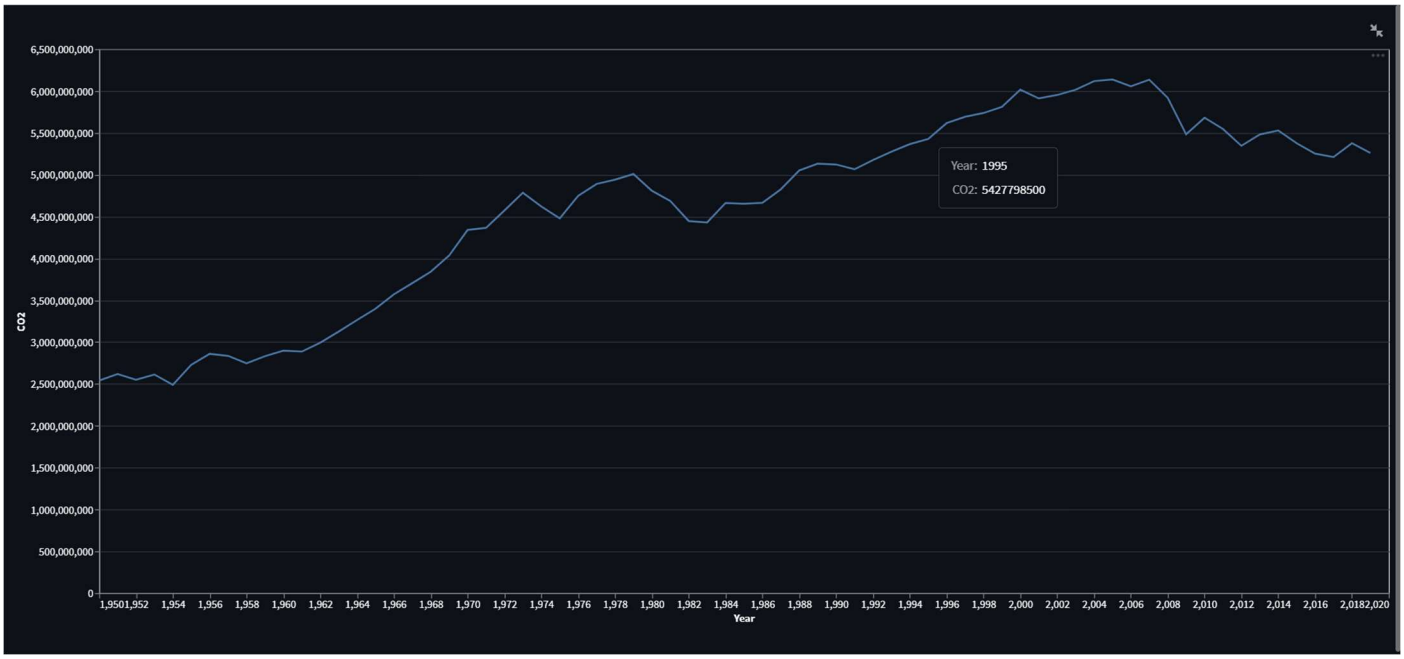
[15] Human Development Index vs. Historical Index of Human Development. Our World in Data. https://ourworldindata.org/grapher/hdi-vs-hihd

[16] Kotz, D. M. (2009). The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism. Review of radical political economics, 41(3), 305-317.

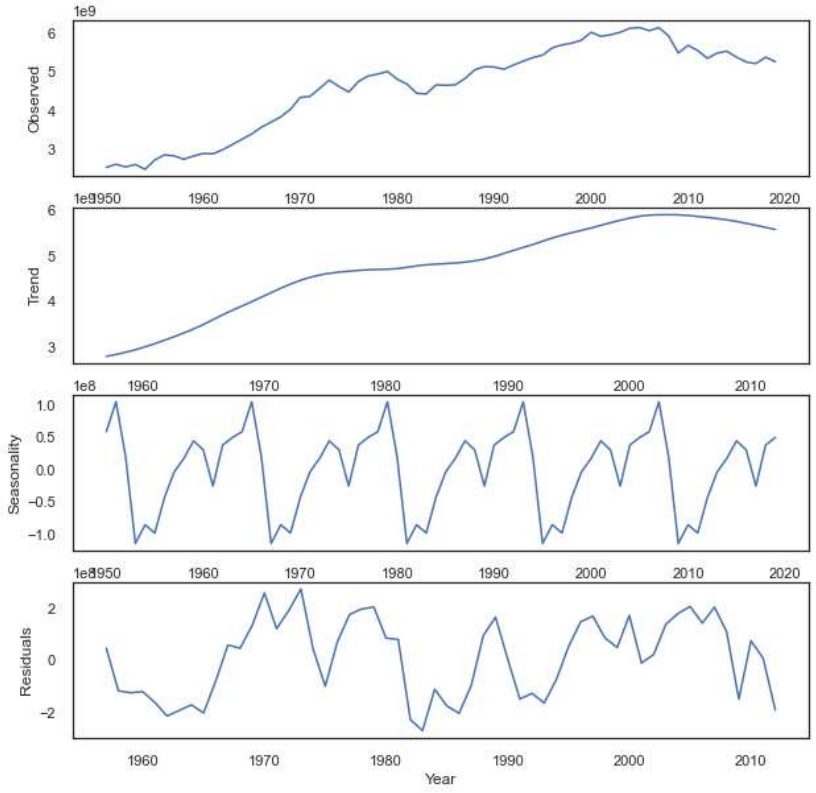
[17] Dimitrov, R. S. (2010). Inside UN climate change negotiations: The Copenhagen conference. Review of policy research, 27(6), 795-821.

[18] Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. American journal of epidemiology, 156(3), 193-203.

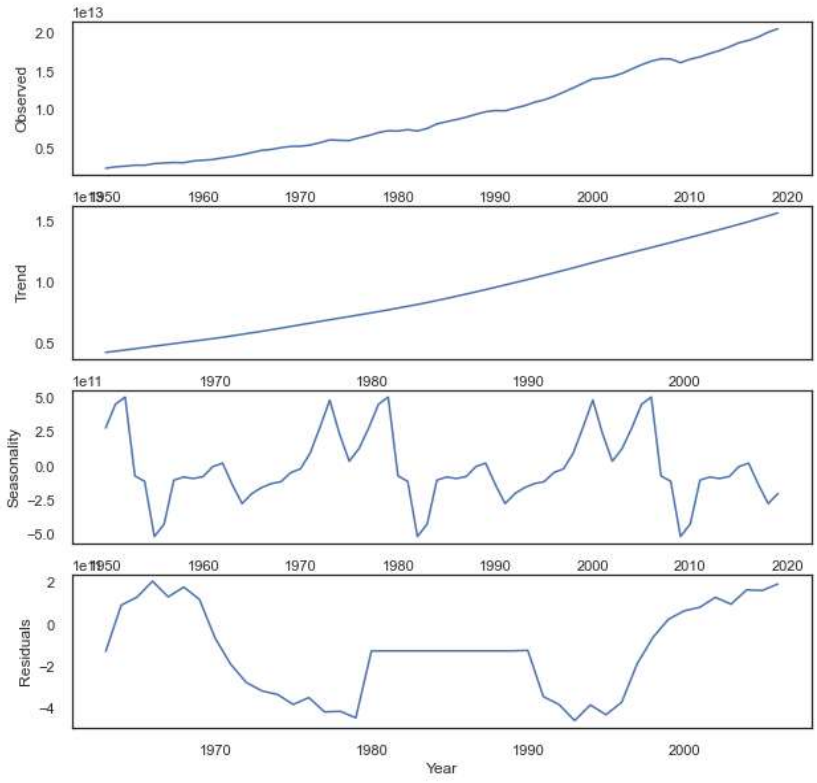
APPENDIX I

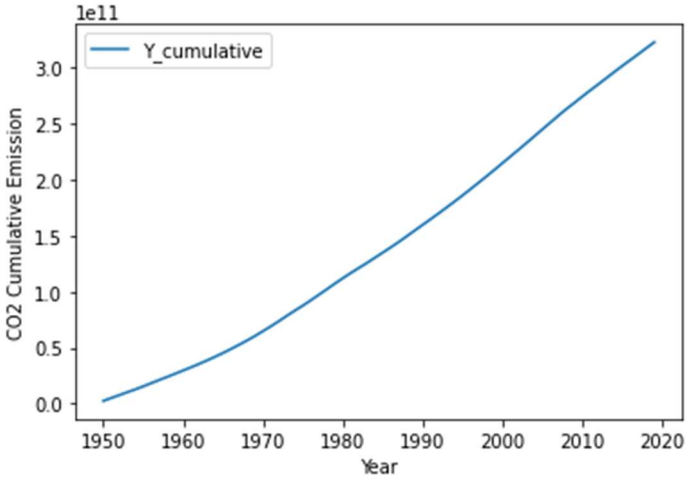


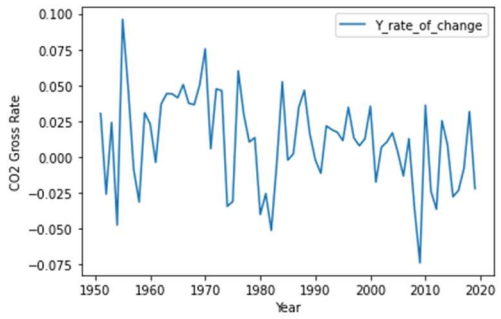
Seasonal decomposition (Additive) of time series in a 17 years period



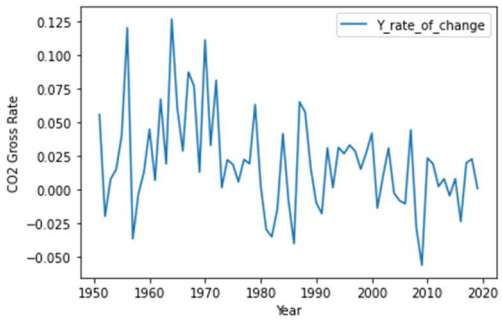
GDP Seasonal decomposition (Additive) of time series in a 27 years period



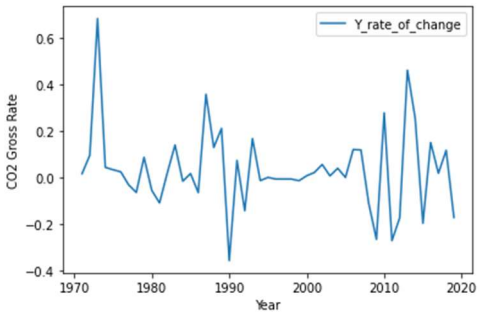




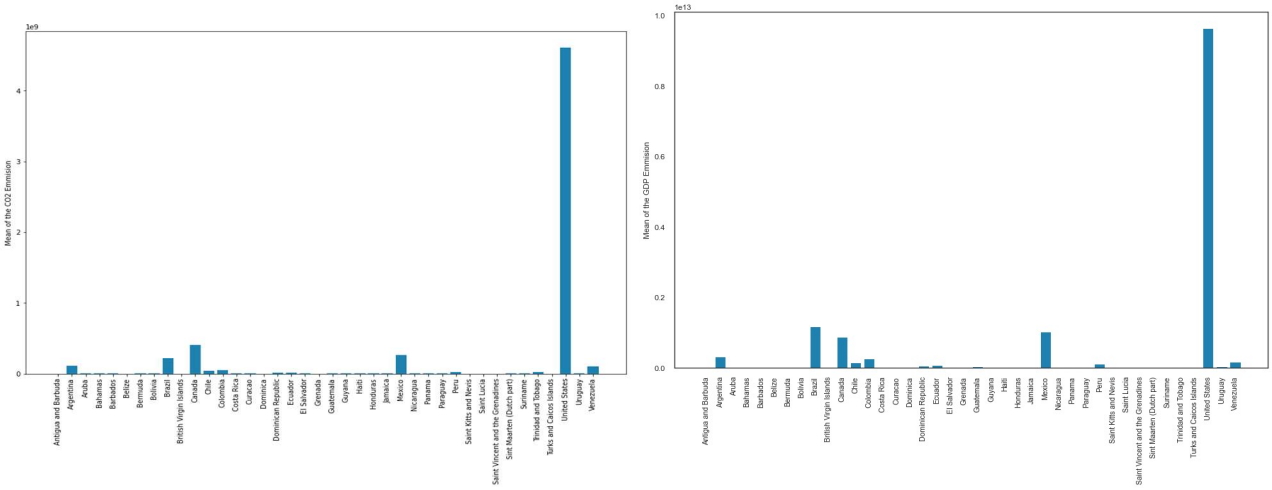
USA

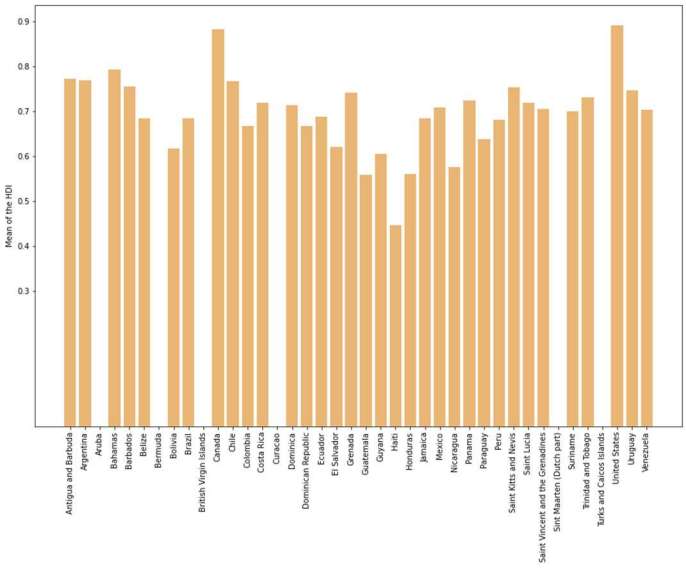


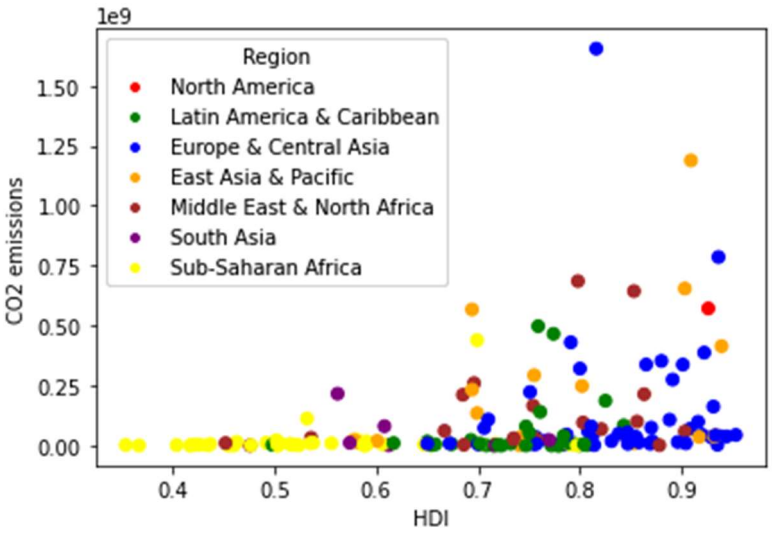
CANADA



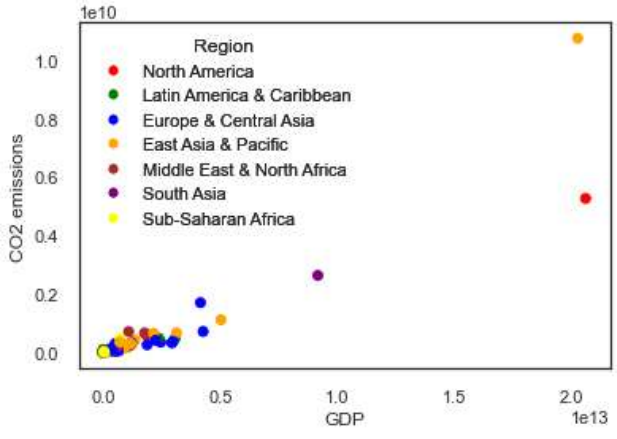
BERMUDA



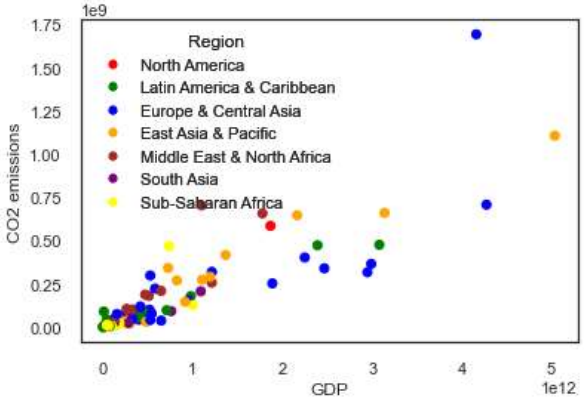




OUTLIERS REMOVED



GDP VS CO2 (2019)



GDP VS CO2 (2019 OUTLIERS REMOVED)

	Model	MAE	MSE	RMSE	↓ R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	123,680,720.1280	31,489,466,444,909,656.0000	157,681,666.8988	0.9420	0.0317	0.0264	0.0160
ada	AdaBoost Regressor	161,638,620.2585	42,677,160,745,712,752.0000	199,622,391.7762	0.9258	0.0416	0.0346	0.0080
rf	Random Forest Regressor	158,421,555.0976	44,375,673,585,578,584.0000	201,335,228.4734	0.9181	0.0399	0.0329	0.0200
knn	K Neighbors Regressor	163,941,664.0000	41,216,158,929,518,592.0000	194,538,268.8000	0.9062	0.0402	0.0349	0.0070
catboost	CatBoost Regressor	177,488,029.6272	53,298,451,745,470,424.0000	218,854,548.4683	0.9006	0.0446	0.0378	0.0740
gbr	Gradient Boosting Regressor	188,142,169.1299	58,984,712,877,078,376.0000	230,929,924.8623	0.8957	0.0475	0.0402	0.0070
dt	Decision Tree Regressor	191,021,575.6800	60,671,111,559,940,672.0000	233,287,358.7796	0.8926	0.0480	0.0408	0.0070
xgboost	Extreme Gradient Boosting	190,980,116.8000	60,658,442,240,943,720.0000	233,267,148.8000	0.8926	0.0480	0.0408	0.0070
br	Bayesian Ridge	549,730,662.4000	384,686,947,535,788,416.0000	600,458,857.6000	-0.0033	0.1487	0.1355	0.0050
en	Elastic Net	549,714,716.8000	384,688,358,003,048,448.0000	600,445,932.8000	-0.0035	0.1487	0.1355	0.0050
lar	Least Angle Regression	549,714,752.0000	384,688,473,108,171,968.0000	600,446,006.4000	-0.0035	0.1487	0.1355	0.0060
llar	Lasso Least Angle Regression	549,714,752.0000	384,688,473,108,171,968.0000	600,446,006.4000	-0.0035	0.1487	0.1355	0.0060
lr	Linear Regression	549,714,745.6000	384,688,390,644,799,872.0000	600,445,958.4000	-0.0035	0.1487	0.1355	0.0100
lasso	Lasso Regression	549,714,716.8000	384,688,358,003,048,448.0000	600,445,932.8000	-0.0035	0.1487	0.1355	0.0070
ridge	Ridge Regression	549,714,684.8000	384,688,368,310,969,984.0000	600,445,932.8000	-0.0035	0.1487	0.1355	0.0050
omp	Orthogonal Matching Pursuit	549,714,752.0000	384,688,473,108,171,968.0000	600,446,006.4000	-0.0035	0.1487	0.1355	0.0050
dummy	Dummy Regressor	1,023,553,910.4000	1,473,974,927,931,683,584.0000	1,176,911,660.8000	-0.8506	0.2890	0.2707	0.0050
huber	Huber Regressor	1,428,403,438.6940	2,407,892,711,031,334,400.0000	1,518,167,411.1350	-4.8478	0.4994	0.3496	0.0070

CO2 VS GDP REGRESSION ANALYSIS

APPENDIX II

graphs_usa

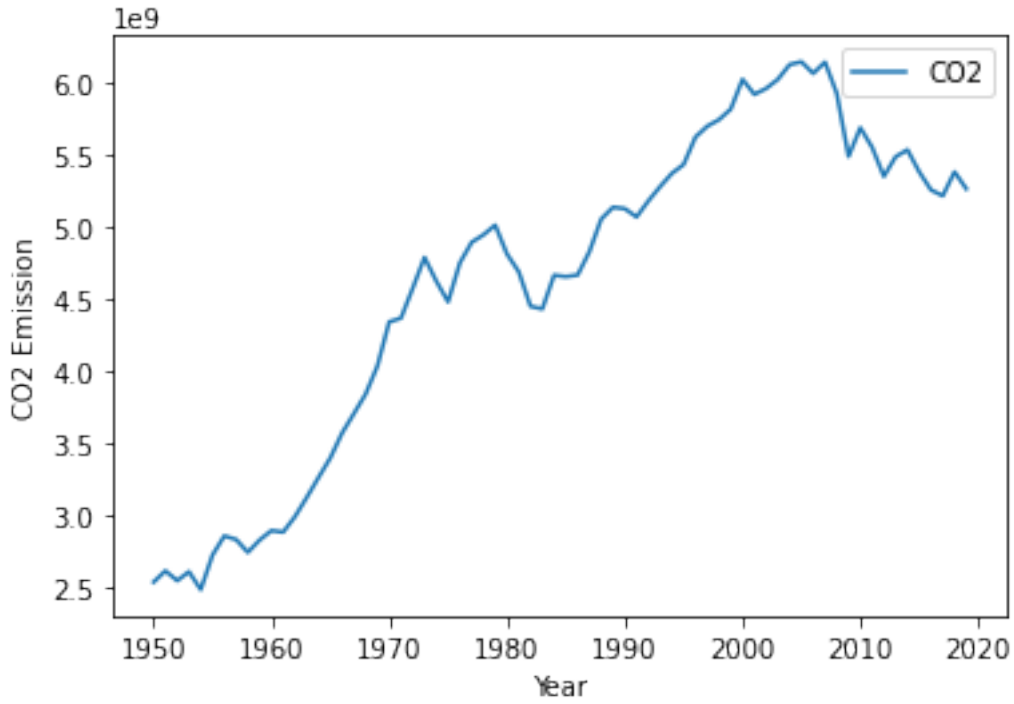
December 17, 2022

[249]:
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
import plotly.express as px
from scipy.signal import savgol_filter
```

[416]:
```
# Load data from a CSV file
co2 = pd.read_excel('CO2.xlsx')
hdi = pd.read_excel('HDI.xlsx')
gdp = pd.read_excel('national-gdp.xlsx')

# Data merging and preprocessing
df_merged = pd.merge(co2, hdi, on=['Code', 'Year'], how='left')
df_gdp_merge = pd.merge(df_merged, gdp, on=['Code', 'Year'], how='left').
 ↪drop(["Entity_y", "Entity"],axis=(1))

df = df_gdp_merge[df_gdp_merge['Region'].notna()]
df = df.reset_index(drop=True)
```

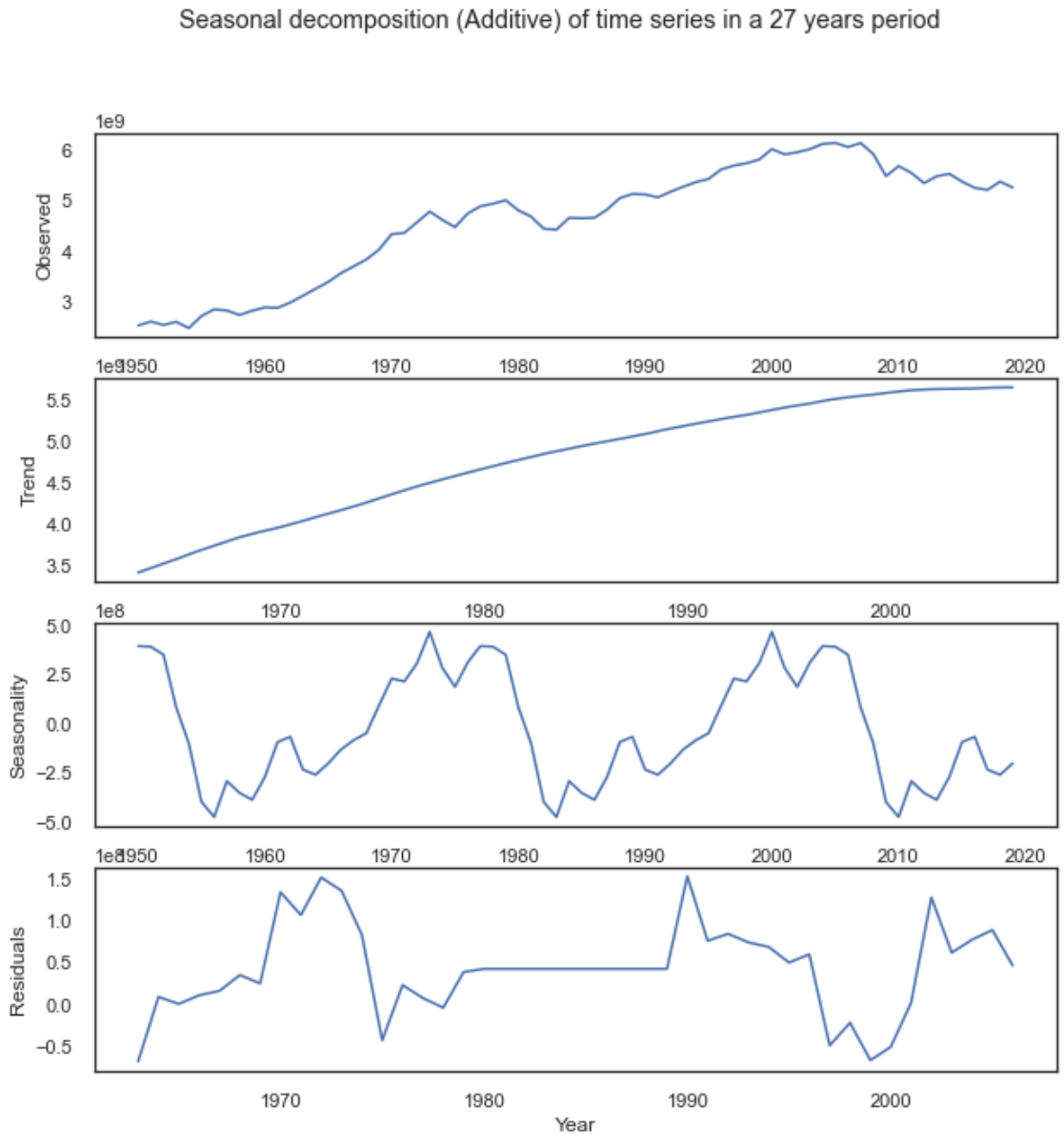
[ ]:
```
# Select the X country as United States:
X_Country = df[df["Code"] == "USA"]
```

[ ]:
```
# Time Series on USA:
X_Country.plot(x="Year", y="CO2", ylabel= 'CO2 Emission')
plt.show()
```



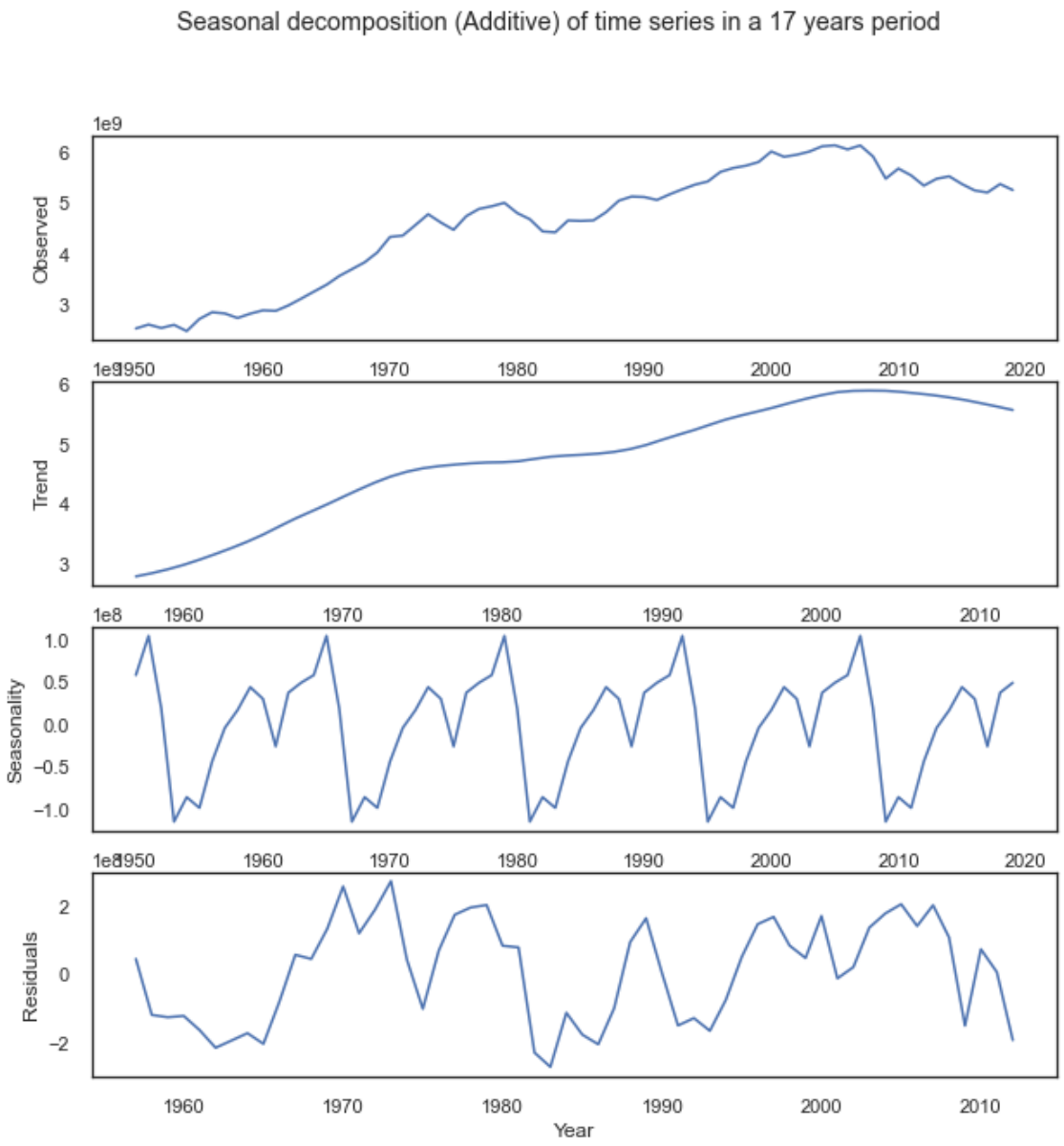
[392]:
```
from statsmodels.tsa.seasonal import seasonal_decompose

# Load the data
# Decompose the time series into trend, seasonality, and residuals
X_Country_CO2 = X_Country[['CO2', 'Year']]
X_Country_CO2.set_index('Year', inplace=True)
# Convert the data to a Pandas series with a PeriodIndex
result = seasonal_decompose(X_Country_CO2['CO2'], model='additive', period= 27)

# Create the plot
fig, axs = plt.subplots(4, 1, figsize=(10, 10))

result.observed.plot(ax=axs[0])
result.trend.plot(ax=axs[1])
result.seasonal.plot(ax=axs[2])
result.resid.plot(ax=axs[3])

# Add labels and title
axs[0].set_ylabel('Observed')
axs[1].set_ylabel('Trend')
axs[2].set_ylabel('Seasonality')
axs[3].set_ylabel('Residuals')
plt.suptitle('Seasonal decomposition (Additive) of time series in a 27 years␣
 ↪period')
```

```
# Save the plot or display it
plt.savefig('seasonal_decomposition.png')
plt.show()# 27 (1.5~-0.5)
```



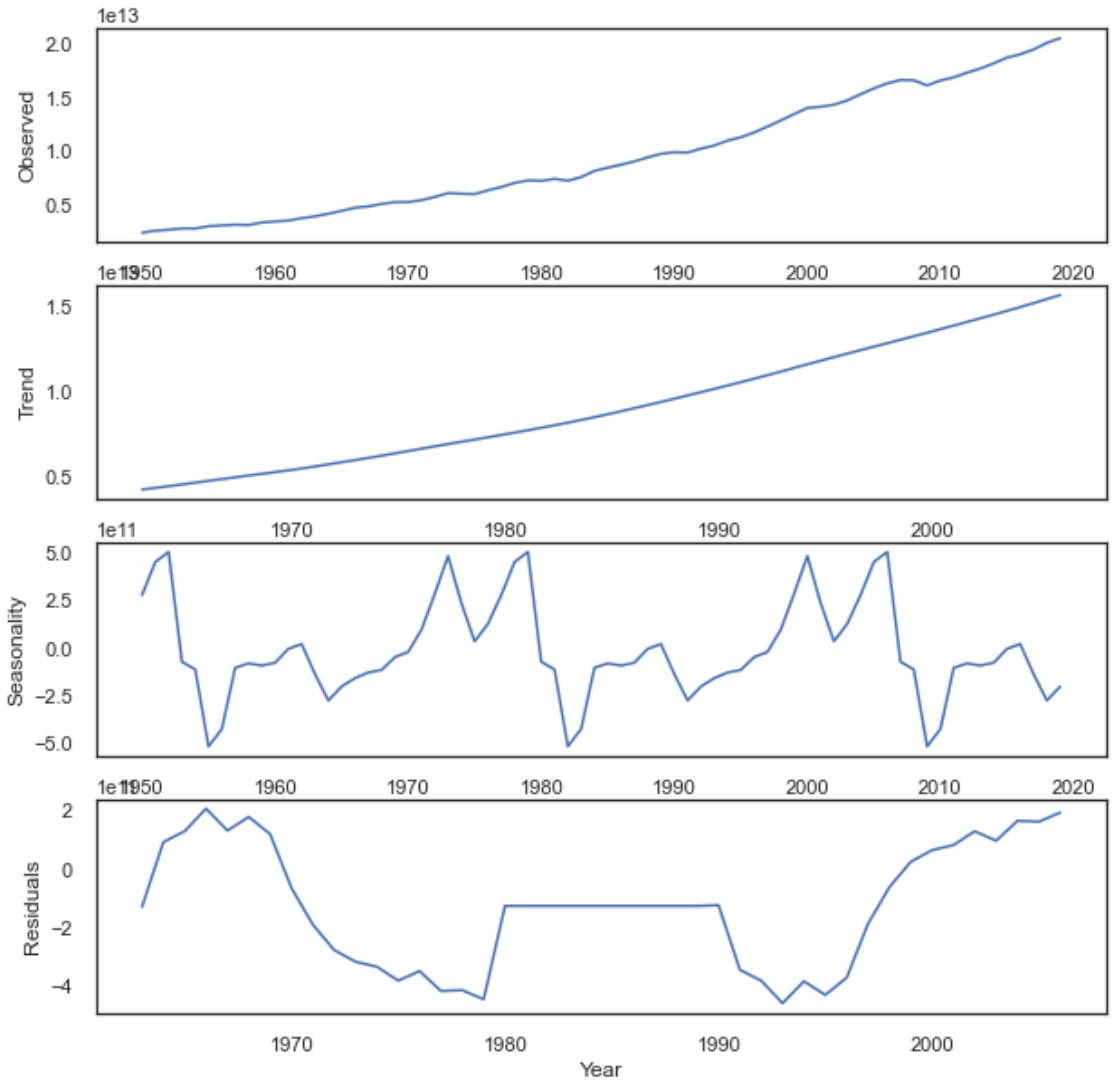
[408]:
```
# Convert the data to a Pandas series with a PeriodIndex
result = seasonal_decompose(X_Country_CO2['CO2'], model='additive', period= 14)

# Create the plot
fig, axs = plt.subplots(4, 1, figsize=(10, 10))
```

```
result.observed.plot(ax=axs[0])
result.trend.plot(ax=axs[1])
result.seasonal.plot(ax=axs[2])
result.resid.plot(ax=axs[3])

# Add labels and title
axs[0].set_ylabel('Observed')
axs[1].set_ylabel('Trend')
axs[2].set_ylabel('Seasonality')
axs[3].set_ylabel('Residuals')
plt.suptitle('Seasonal decomposition (Additive) of time series in a 17 years␣
 ↪period')
# Save the plot or display it
plt.savefig('seasonal_decomposition_16.png')
plt.show()
```



[412]:
```
X_Country_CO2 = X_Country[['GDP (output, multiple price benchmarks)', 'Year']]
X_Country_CO2.set_index('Year', inplace=True)
# Convert the data to a Pandas series with a PeriodIndex
result = seasonal_decompose(X_Country_CO2['GDP (output, multiple price␣
 ↪benchmarks)'], model='additive', period= 27)

# Create the plot
fig, axs = plt.subplots(4, 1, figsize=(10, 10))

result.observed.plot(ax=axs[0])
```

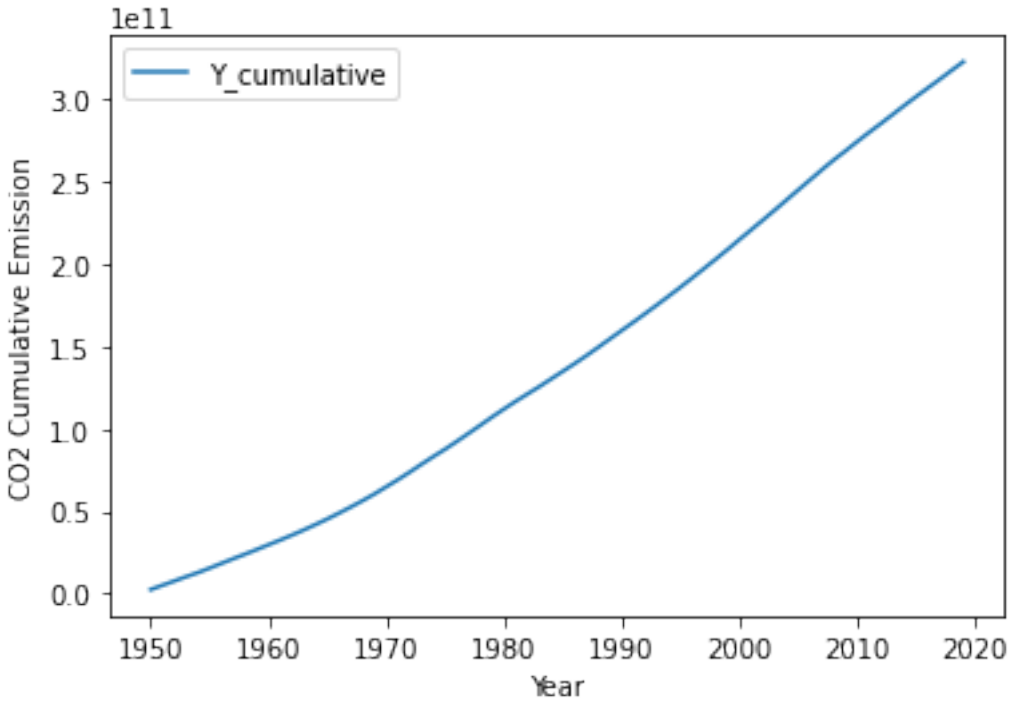
```
result.trend.plot(ax=axs[1])
result.seasonal.plot(ax=axs[2])
result.resid.plot(ax=axs[3])

# Add labels and title
axs[0].set_ylabel('Observed')
axs[1].set_ylabel('Trend')
axs[2].set_ylabel('Seasonality')
axs[3].set_ylabel('Residuals')
plt.suptitle('GDP Seasonal decomposition (Additive) of time series in a 27␣
 ↪years period')
# Save the plot or display it
plt.savefig('seasonal_decom_16.png')
plt.show()
```



[ ]:
```
# Cumulative plot of Y for X
X_Country["Y_cumulative"] = X_Country["CO2"].cumsum()
X_Country.plot(x="Year", y="Y_cumulative", ylabel="CO2 Cumulative Emission")
plt.show()
```

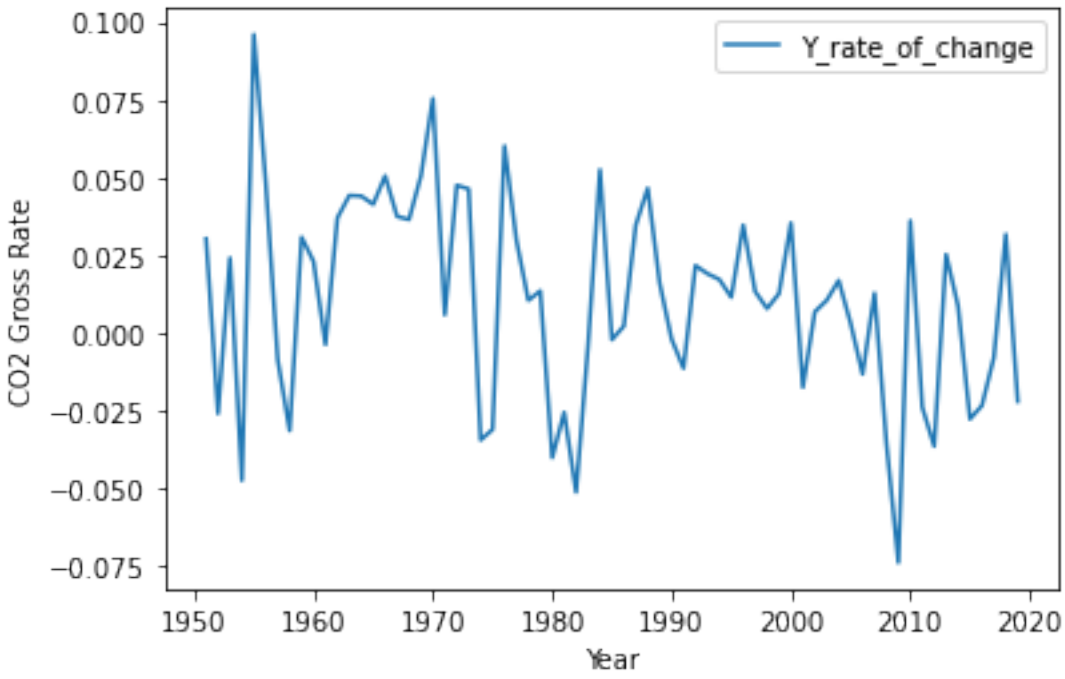
```
C:\Users\s24Yu\AppData\Local\Temp\ipykernel_30796\1601110225.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  X_Country["Y_cumulative"] = X_Country["CO2"].cumsum()
```



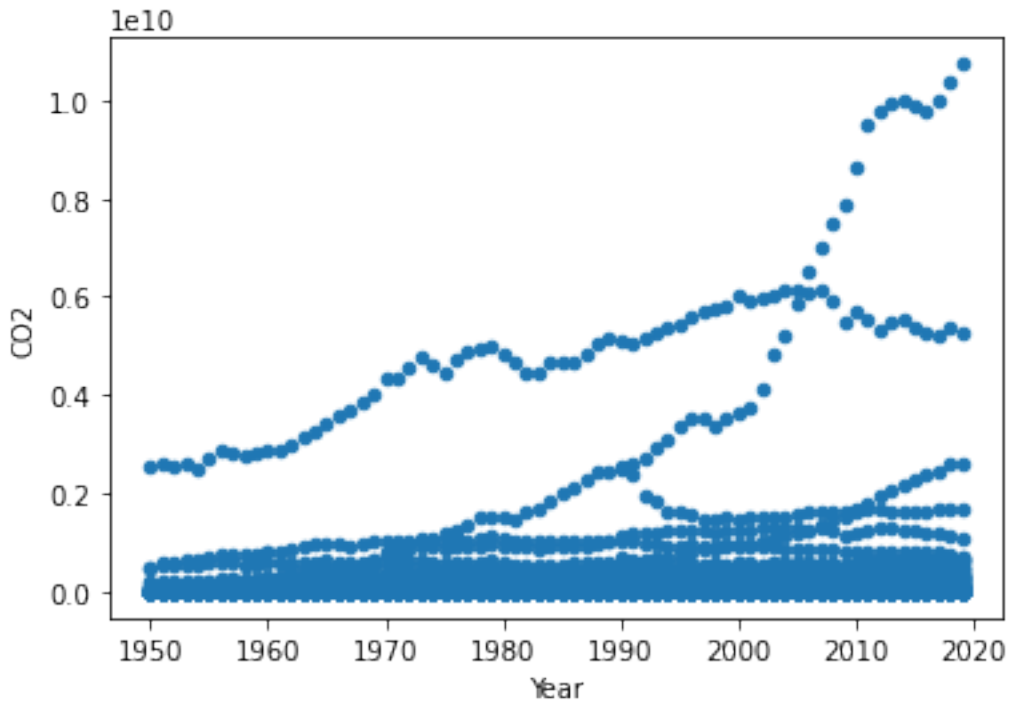
[ ]:
```
X_Country["Y_rate_of_change"] = X_Country["CO2"].pct_change()
X_Country.plot(x="Year", y="Y_rate_of_change", ylabel="CO2 Gross Rate")
plt.show()
```

```
C:\Users\s24Yu\AppData\Local\Temp\ipykernel_30796\527911188.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  X_Country["Y_rate_of_change"] = X_Country["CO2"].pct_change()
```



[ ]:
```
df.plot(x="Year", y="CO2", kind="scatter")
plt.show()
```

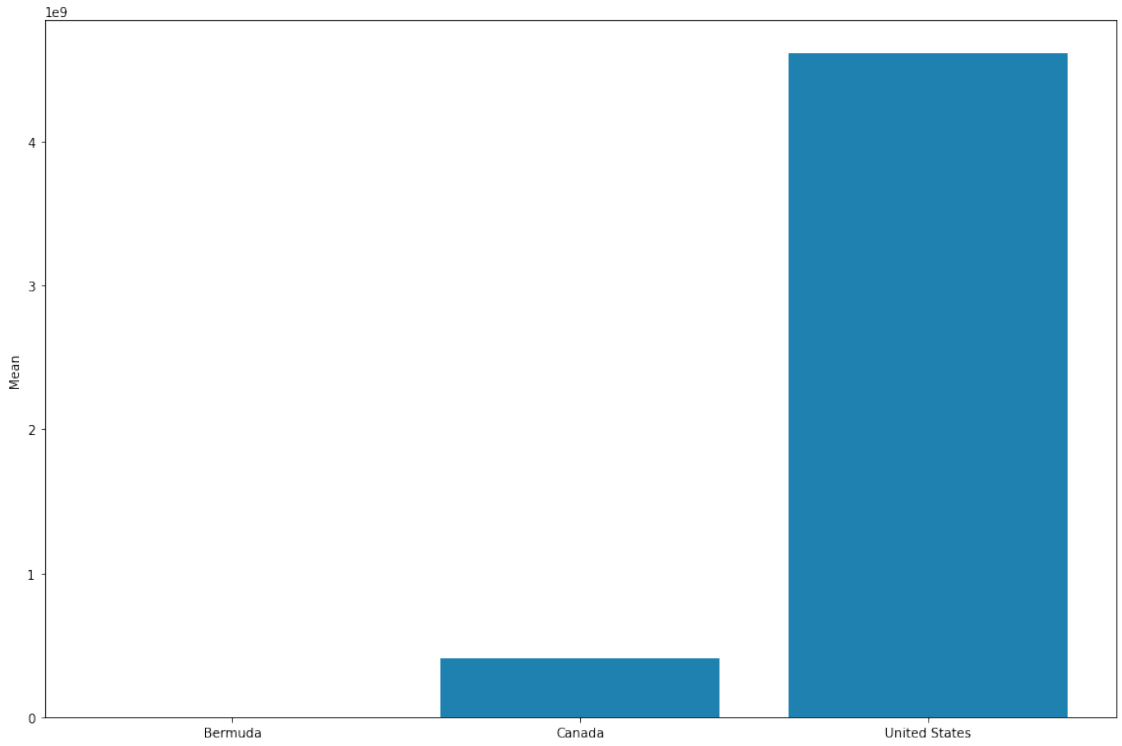
```
c:\Users\s24Yu\anaconda3\lib\site-
packages\pandas\plotting\_matplotlib\core.py:1114: UserWarning: No data for
colormapping provided via 'c'. Parameters 'cmap' will be ignored
  scatter = ax.scatter(
```



[ ]:
```
NA = df[df['Region'] == 'North America']
NA = NA.reset_index(drop=True)
display(NA)
NA_1 = NA.groupby('Entity_x').mean().reset_index()
plt.figure(figsize=(15,10))
plt.bar(range(len(NA_1)), NA_1['CO2'], color="#1e81b0")
plt.xticks(range(len(NA_1)), NA_1['Entity_x'])
plt.ylabel('Mean')
plt.savefig('hst_na.png')
```

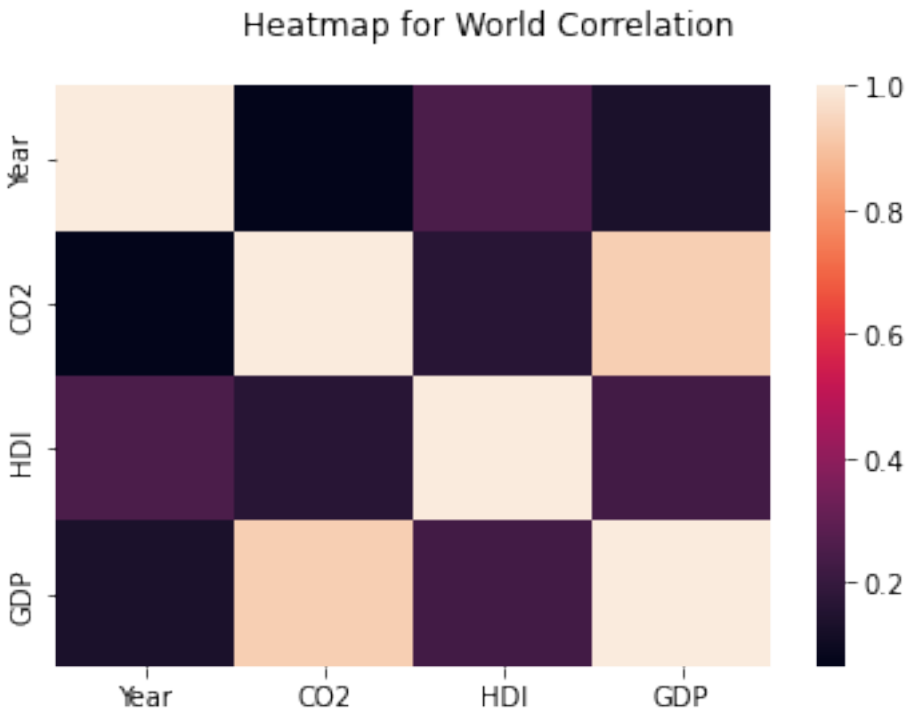
```
          Entity_x Code  Year           CO2    HDI         Region  \
0          Bermuda  BMU  1970  2.271680e+05    NaN  North America
1          Bermuda  BMU  1971  2.308320e+05    NaN  North America
2          Bermuda  BMU  1972  2.528160e+05    NaN  North America
3          Bermuda  BMU  1973  4.250240e+05    NaN  North America
4          Bermuda  BMU  1974  4.433440e+05    NaN  North America
..             ...  ...   ...           ...    ...            ...
185  United States  USA  2015  5.376578e+09  0.920  North America
186  United States  USA  2016  5.251758e+09  0.922  North America
187  United States  USA  2017  5.210957e+09  0.924  North America
188  United States  USA  2018  5.376657e+09    NaN  North America
189  United States  USA  2019  5.259144e+09    NaN  North America

     GDP (output, multiple price benchmarks)
```

```
0                                1.081255e+10
1                                2.050390e+10
2                                1.638498e+10
3                                1.607617e+10
4                                1.421993e+10
..                                        ...
185                              1.878540e+13
186                              1.909520e+13
187                              1.954300e+13
188                              2.015530e+13
189                              2.059580e+13

[190 rows x 7 columns]
```

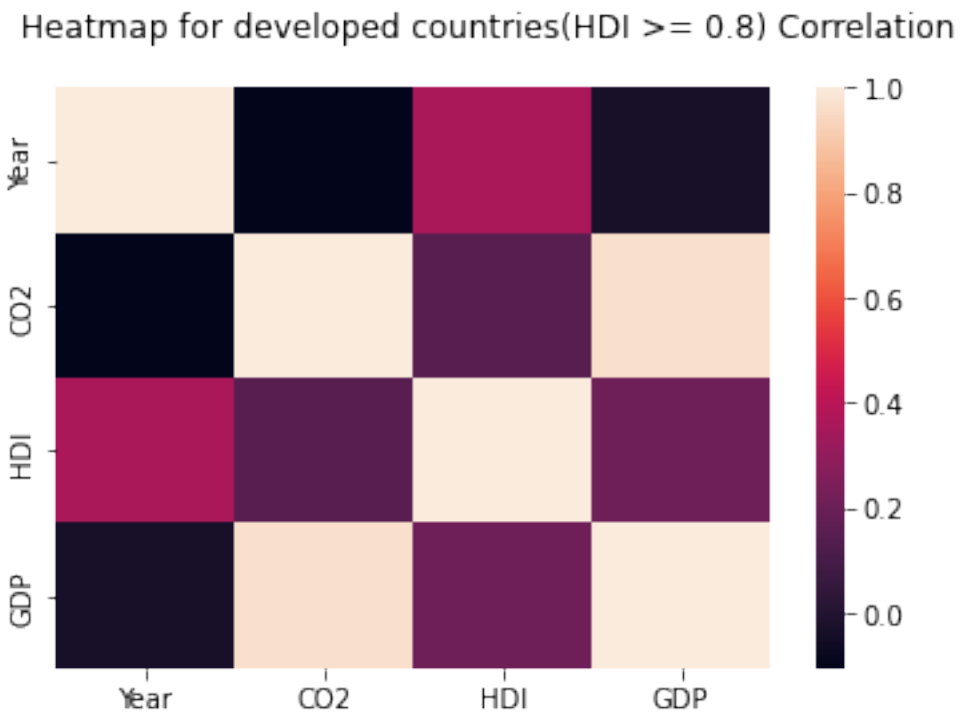


[ ]:
```
df['GDP'] = df['GDP (output, multiple price benchmarks)']
df = df.drop('GDP (output, multiple price benchmarks)', axis = (1))
```

[ ]:
```
sns.heatmap(df.corr())
plt.suptitle('Heatmap for World Correlation')
plt.savefig('hm_wrld.png')
plt.show()
```



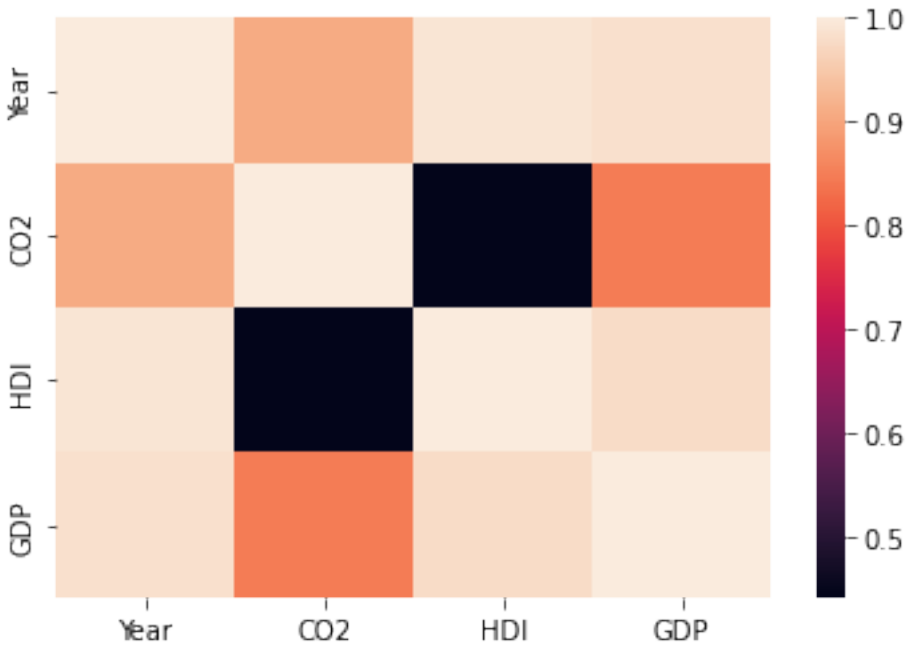
[ ]:
```
DEV = df[df['HDI'] >= 0.8]
display
DEV = DEV.reset_index(drop=True)
sns.heatmap(DEV.corr())
plt.suptitle('Heatmap for developed countries(HDI >= 0.8) Correlation')
plt.savefig('hm_dev.png')
plt.show()
```



[ ]:
```
X_Country['GDP'] = X_Country['GDP (output, multiple price benchmarks)']
X_Country = X_Country.drop(['GDP (output, multiple price␣
↪benchmarks)','Y_rate_of_change'], axis = (1))
```

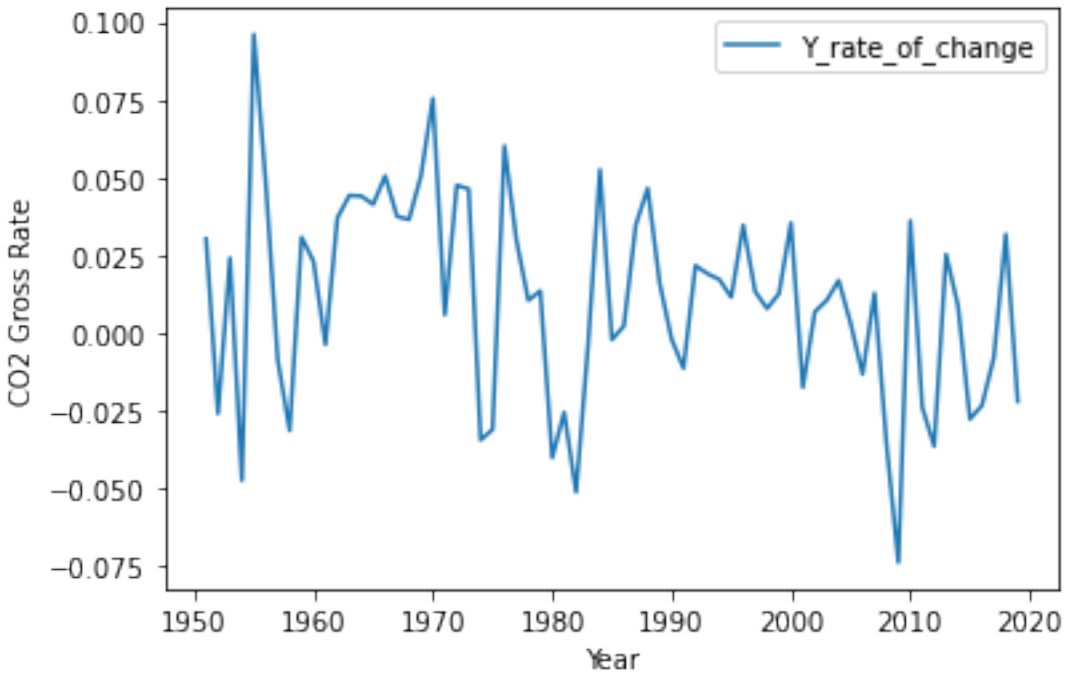
```
C:\Users\s24Yu\AppData\Local\Temp\ipykernel_30796\1650941638.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  X_Country['GDP'] = X_Country['GDP (output, multiple price benchmarks)']
```

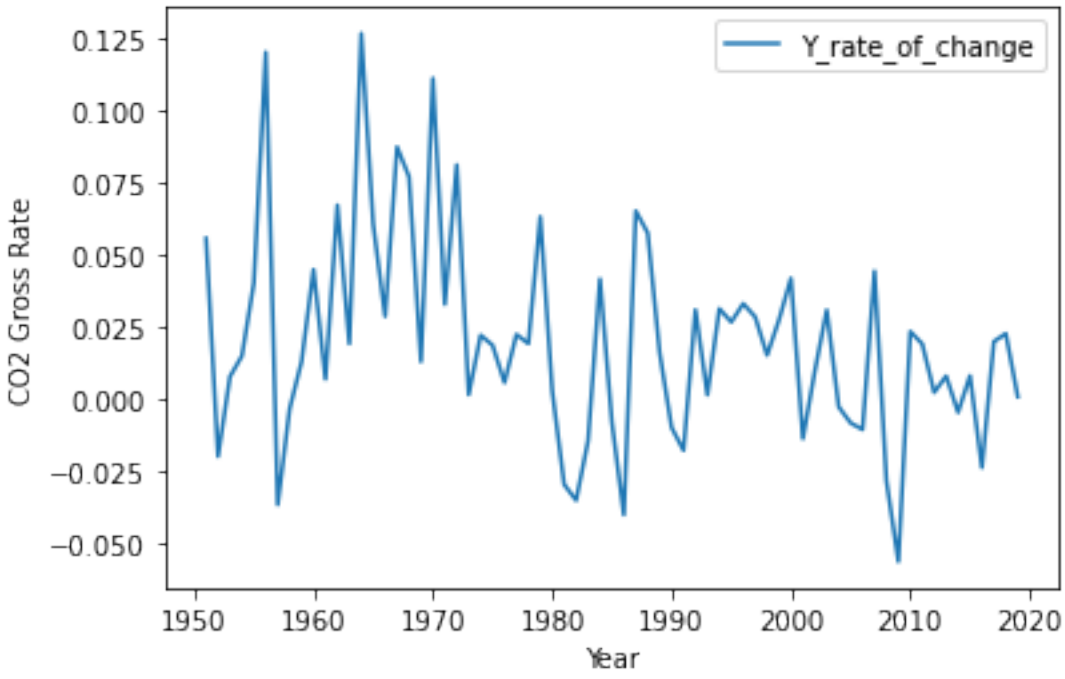
[ ]:
```
sns.heatmap(X_Country.corr())
plt.savefig('hm_usa.png')
plt.show()
```

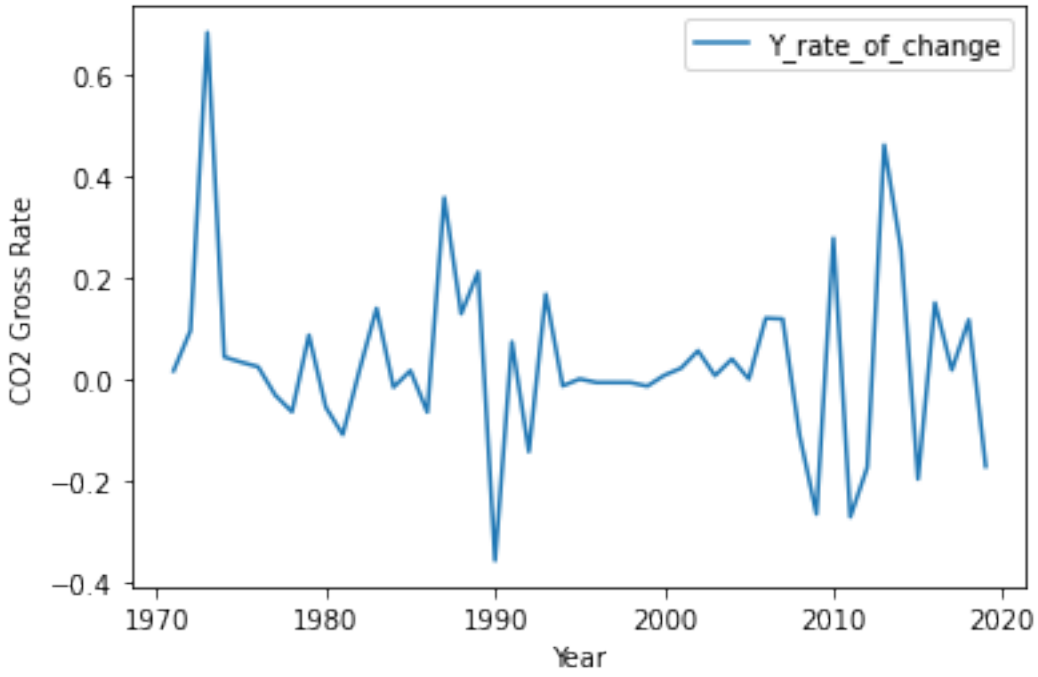


[ ]:
```
X_Country.plot(x="Year", y="Y_rate_of_change", ylabel="CO2 Gross Rate")
can = df[df["Code"] == "CAN"]
ber = df[df["Code"] == "BMU"]
can["Y_rate_of_change"] = can["CO2"].pct_change()
ber["Y_rate_of_change"] = ber["CO2"].pct_change()
can.plot(x="Year", y="Y_rate_of_change", ylabel="CO2 Gross Rate")
ber.plot(x="Year", y="Y_rate_of_change", ylabel="CO2 Gross Rate")

plt.show()
```

```
C:\Users\s24Yu\AppData\Local\Temp\ipykernel_30796\1406352151.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  can["Y_rate_of_change"] = can["CO2"].pct_change()
C:\Users\s24Yu\AppData\Local\Temp\ipykernel_30796\1406352151.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  ber["Y_rate_of_change"] = ber["CO2"].pct_change()
```


Y_rate_of_change
CO2 Gross Rate
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
1950
1960
1970
1980
1990
2000
2010
2020
Year


Y_rate_of_change
CO2 Gross Rate
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
1950
1960
1970
1980
1990
2000
2010
2020
Year

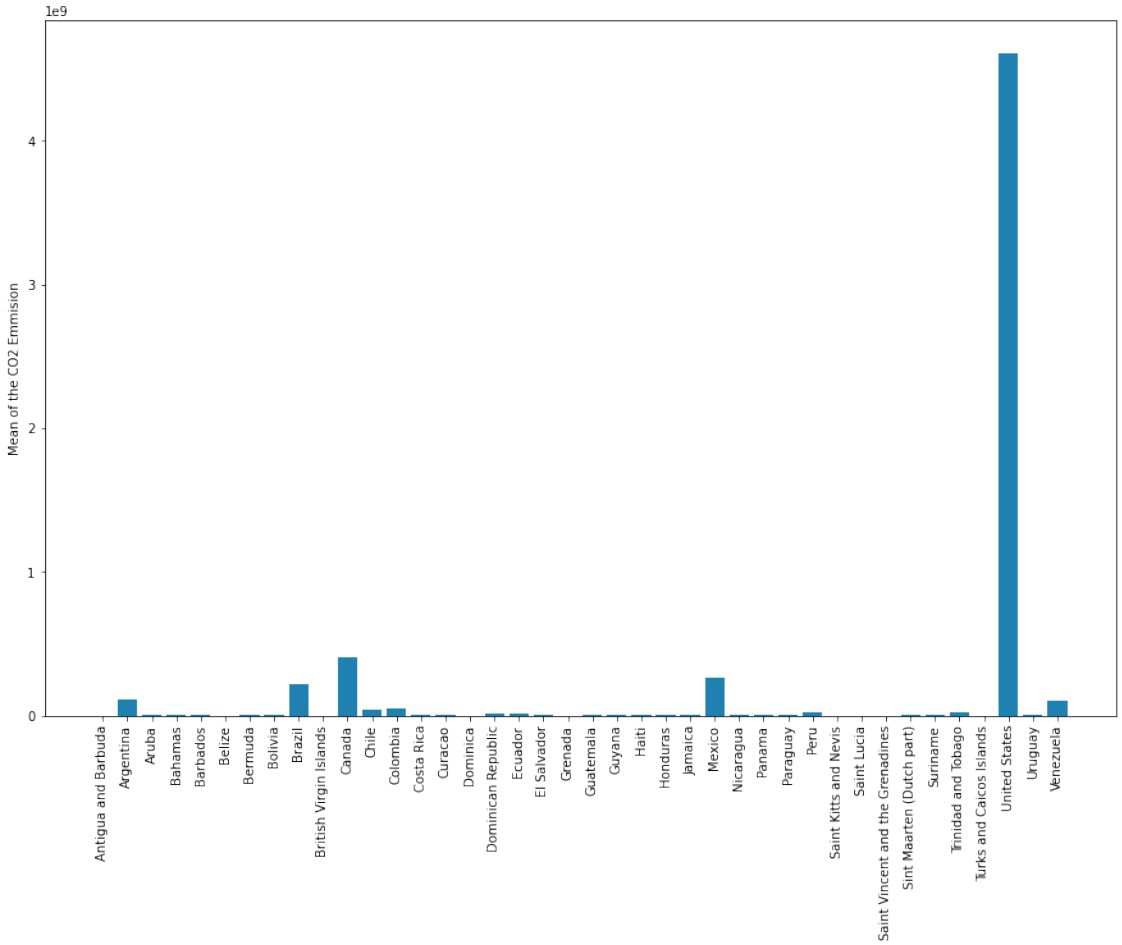


[ ]:
```
AM = df[df['Region'] == 'Latin America & Caribbean']
NA = df[df['Region'] == 'North America']
AM = pd.concat([AM, NA], axis = 0)
AM = AM.reset_index(drop=True)
display(AM)
AM_1 = AM.groupby('Entity_x').mean().reset_index()
plt.figure(figsize=(15,10))
plt.bar(range(len(AM_1)), AM_1['CO2'], color='#1e81b0')
# plt.bar(range(len(AM_1)), AM_1['CO2'], color='#1e81b0')
plt.xticks(range(len(AM_1)), AM_1['Entity_x'], rotation='vertical')
plt.ylabel('Mean of the CO2 Emmision')
```

	Entity_x	Code	Year	CO2	HDI	\
0	Antigua and Barbuda	ATG	1970	4.616640e+05	NaN	
1	Antigua and Barbuda	ATG	1971	4.250240e+05	NaN	
2	Antigua and Barbuda	ATG	1972	3.737280e+05	NaN	
3	Antigua and Barbuda	ATG	1973	3.297600e+05	NaN	
4	Antigua and Barbuda	ATG	1974	4.286880e+05	NaN	
...	...	...	...	...	...	
2348	United States	USA	2015	5.376578e+09	0.920	
2349	United States	USA	2016	5.251758e+09	0.922	
2350	United States	USA	2017	5.210957e+09	0.924	
2351	United States	USA	2018	5.376657e+09	NaN	
2352	United States	USA	2019	5.259144e+09	NaN	

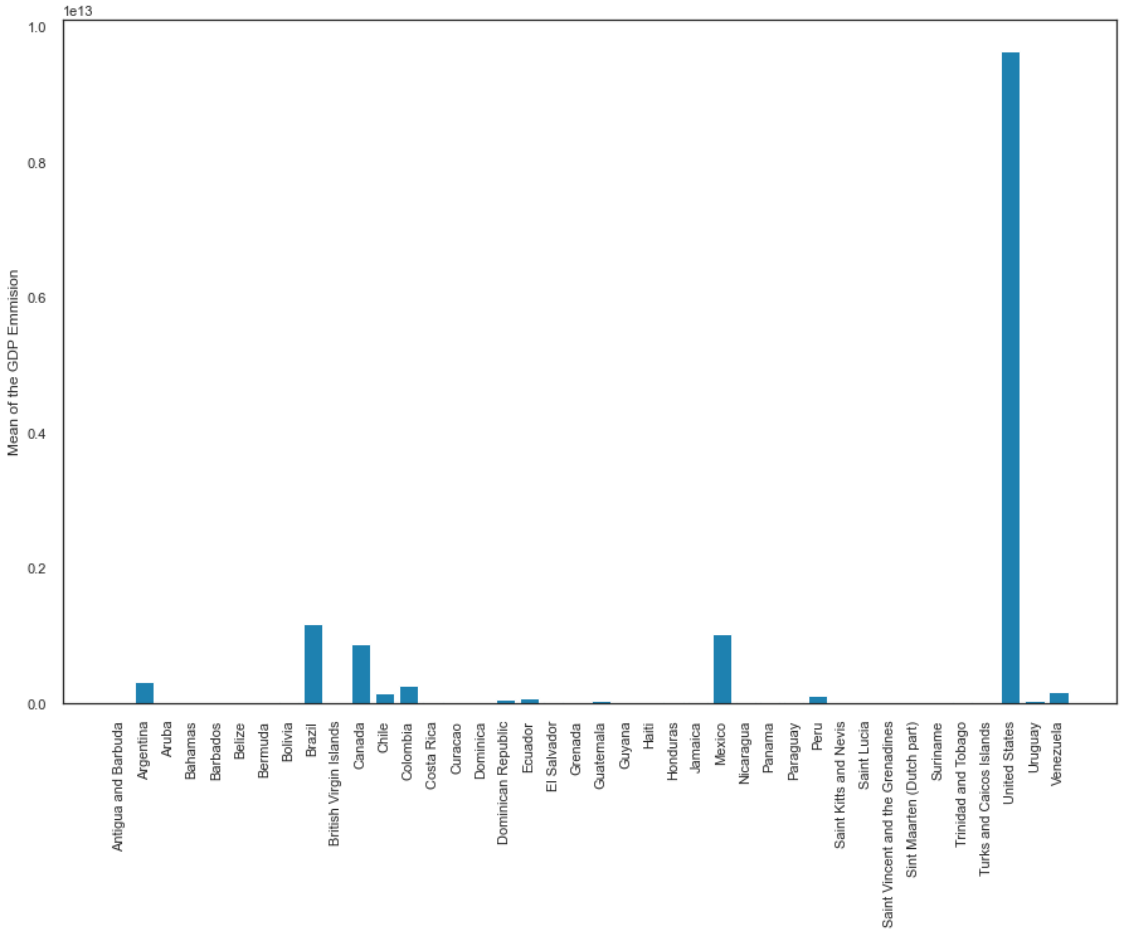
```
                        Region  GDP (output, multiple price benchmarks)
0     Latin America & Caribbean                              4.001085e+08
1     Latin America & Caribbean                              4.306256e+08
2     Latin America & Caribbean                              4.613090e+08
3     Latin America & Caribbean                              4.962153e+08
4     Latin America & Caribbean                              4.848474e+08
...                         ...                                       ...
2348              North America                              1.878540e+13
2349              North America                              1.909520e+13
2350              North America                              1.954300e+13
2351              North America                              2.015530e+13
2352              North America                              2.059580e+13

[2353 rows x 7 columns]
```

[ ]: Text(0, 0.5, 'Mean of the CO2 Emmision')



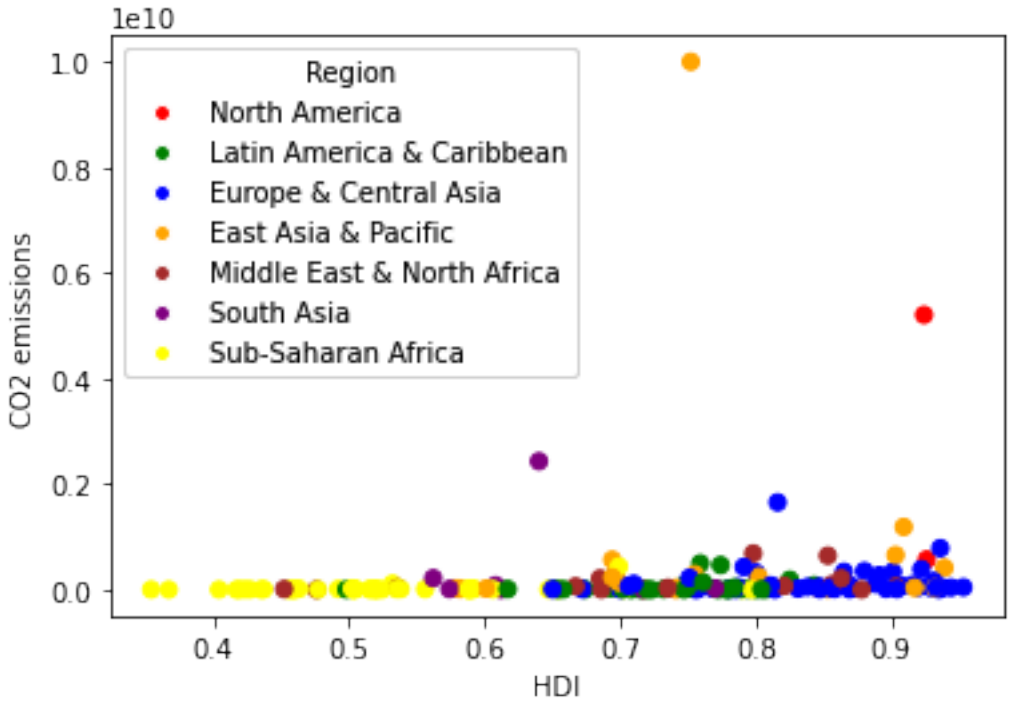
[413]:
```
plt.figure(figsize=(15,10))
plt.bar(range(len(AM_1)), AM_1['GDP (output, multiple price benchmarks)'],␣
↪color='#1e81b0')
# plt.bar(range(len(AM_1)), AM_1['CO2'], color='#1e81b0')
plt.xticks(range(len(AM_1)), AM_1['Entity_x'], rotation='vertical')
plt.ylabel('Mean of the GDP Emmision')
```

[413]:
```
Text(0, 0.5, 'Mean of the GDP Emmision')
```



[ ]:
```
fig, ax = plt.subplots()
world_2017 = df[df['Year']==2017]
# world = df.drop('World', axis=1)
# create a mapping from region to color
region_colors = {
    'North America': 'red',
    'Latin America & Caribbean': 'green',
    'Europe & Central Asia': 'blue',
    'East Asia & Pacific': 'orange',
```

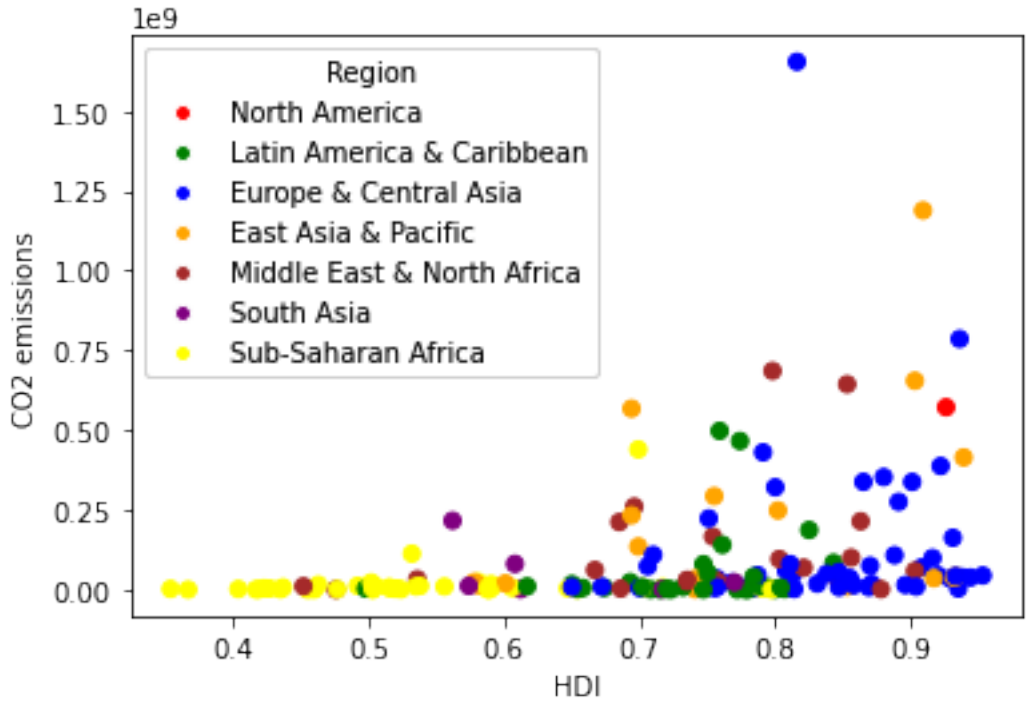
```
    'Middle East & North Africa': 'brown',
    'South Asia': 'purple',
    'Sub-Saharan Africa': 'yellow',
}
# create a list of colors based on the region of each data point
colors = [region_colors[region] for region in world_2017.Region]


scatter = ax.scatter(world_2017.HDI, world_2017.CO2, c=colors)

# add legend
# legend1 = ax.legend(*scatter.legend_elements(), title="Region")

legend_elements = [
    plt.Line2D([0], [0], marker='o', color='w', label=region,␣
 ↪markerfacecolor=color)
    for region, color in region_colors.items()
]
legend1 = ax.legend(handles=legend_elements, title="Region")
ax.add_artist(legend1)


# add axis labels
ax.set_xlabel('HDI')
ax.set_ylabel('CO2 emissions')
plt.savefig('sct_17.png')
plt.show()
```



[ ]:
```
fig, ax = plt.subplots()

world_2017 = df[df['Year']==2017]
world_2017 = world_2017[world_2017['CO2'] <= 2e9]
world_2017 = world_2017[world_2017['HDI'] != 0]
world_2017_nona = world_2017[world_2017['HDI'].notna()]
world_2017_nona = world_2017_nona[world_2017_nona['CO2'] != 0]
world_2017_nona = world_2017_nona[world_2017_nona['CO2'].notna()]
# create a list of colors based on the region of each data point

# create a list of colors based on the region of each data point
colors = [region_colors[region] for region in world_2017_nona.Region]
display(world_2017_nona)

scatter = ax.scatter(world_2017_nona.HDI, world_2017_nona.CO2, c=colors)
display(scatter)
# add legend
#legend1 = ax.legend(*scatter.legend_elements(), title="Region")

legend_elements = [
    plt.Line2D([0], [0], marker='o', color='w', label=region,␣
 ↪markerfacecolor=color)
    for region, color in region_colors.items()
]
```

```
legend1 = ax.legend(handles=legend_elements, title="Region")
ax.add_artist(legend1)


# add axis labels
ax.set_xlabel('HDI')
ax.set_ylabel('CO2 emissions')
plt.savefig('sct_nona.png')
plt.show()
```

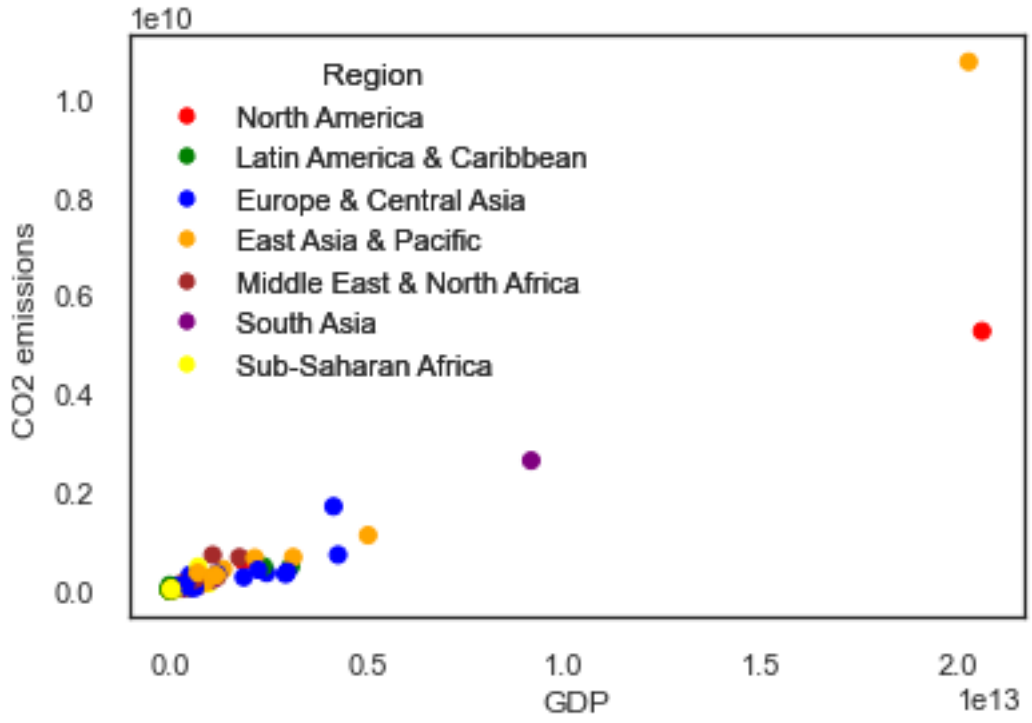
```
                 Entity_x Code  Year          CO2    HDI  \
47                Albania  ALB  2017    5564149.0  0.785
107               Algeria  DZA  2017  166411380.0  0.754
157                Angola  AGO  2017   24304090.0  0.581
237    Antigua and Barbuda  ATG  2017     490976.0  0.780
307             Argentina  ARG  2017  186898880.0  0.825
...                   ...  ...   ...          ...    ...
9987            Venezuela  VEN  2017  140384980.0  0.761
10037             Vietnam  VNM  2017  233040750.0  0.694
10068               Yemen  YEM  2017   10588620.0  0.452
10133              Zambia  ZMB  2017    6842875.0  0.588
10199            Zimbabwe  ZWE  2017    9596071.0  0.535

                         Region  GDP (output, multiple price benchmarks)
47          Europe & Central Asia                              3.497495e+10
107    Middle East & North Africa                              4.721880e+11
157             Sub-Saharan Africa                             2.301510e+11
237     Latin America & Caribbean                              1.367487e+09
307     Latin America & Caribbean                              1.022510e+12
...                           ...                                       ...
9987    Latin America & Caribbean                              1.276834e+10
10037         East Asia & Pacific                              6.472250e+11
10068  Middle East & North Africa                              3.978057e+10
10133           Sub-Saharan Africa                             5.333115e+10
10199           Sub-Saharan Africa                             4.431674e+10

[169 rows x 7 columns]


<matplotlib.collections.PathCollection at 0x20cca8808b0>
```



[414]:
```
fig, ax = plt.subplots()

world_2019 = df[df['Year']==2019]
# world_2017 = world_2017[world_2017['CO2'] <= 2e9]
world_2019 = world_2019[world_2019['CO2'] != 0]
world_2019_nona = world_2019[world_2019['CO2'].notna()]
world_2019_nona = world_2019_nona[world_2019_nona['CO2'] != 0]
world_2019_nona = world_2019_nona[world_2019_nona['CO2'].notna()]
# create a list of colors based on the region of each data point

# create a list of colors based on the region of each data point
colors = [region_colors[region] for region in world_2019_nona.Region]
display(world_2019_nona)

scatter = ax.scatter(world_2019_nona['GDP (output, multiple price␣
 ↪benchmarks)'], world_2019_nona.CO2, c=colors)
display(scatter)
# add legend
#legend1 = ax.legend(*scatter.legend_elements(), title="Region")

legend_elements = [
    plt.Line2D([0], [0], marker='o', color='w', label=region,␣
 ↪markerfacecolor=color)
    for region, color in region_colors.items()
```

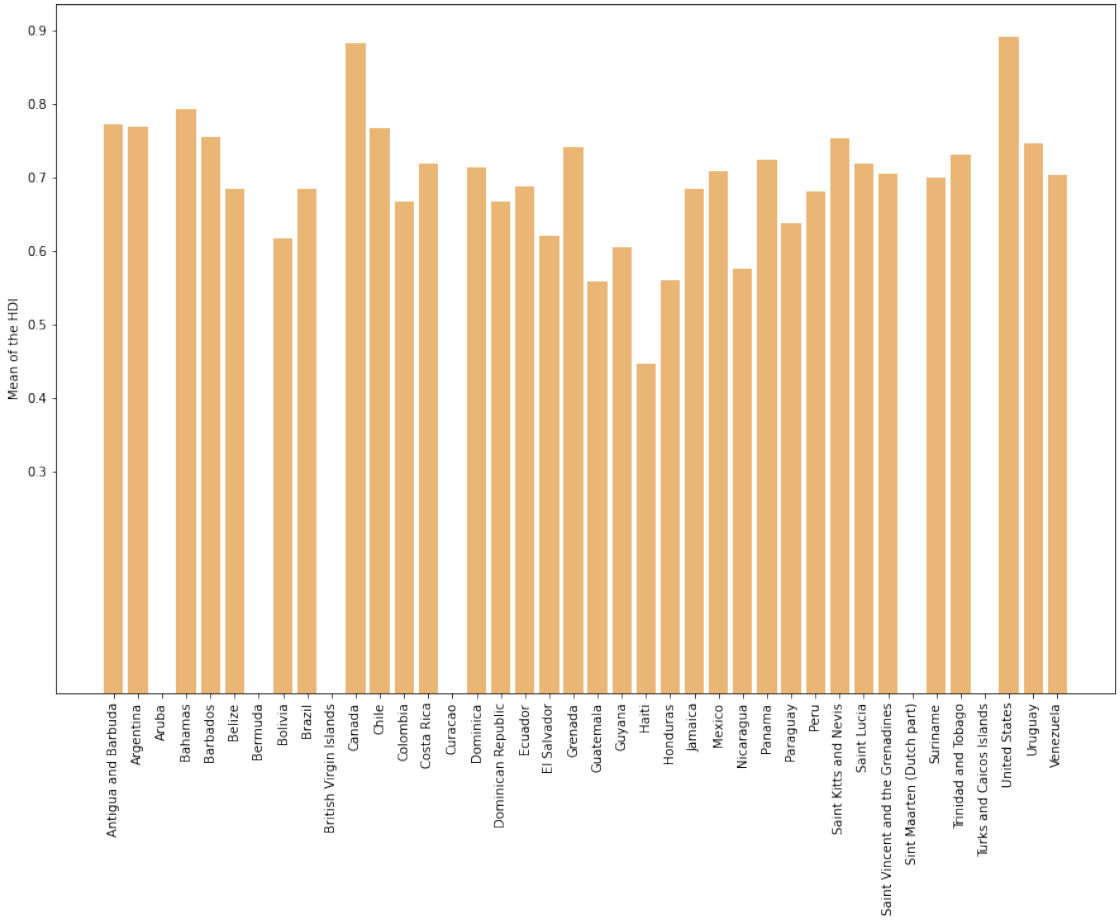
```
]
legend1 = ax.legend(handles=legend_elements, title="Region")
ax.add_artist(legend1)


# add axis labels
ax.set_xlabel('GDP')
ax.set_ylabel('CO2 emissions')
plt.savefig('sct_nona.png')
plt.show()
```

```
                 Entity_x Code  Year          CO2  HDI  \
49                Albania  ALB  2019    4947485.0  NaN
109               Algeria  DZA  2019  179504830.0  NaN
159                Angola  AGO  2019   21818020.0  NaN
189              Anguilla  AIA  2019     153888.0  NaN
239    Antigua and Barbuda  ATG  2019     498304.0  NaN
...                    ...  ...   ...          ...  ...
9989             Venezuela  VEN  2019   89111144.0  NaN
10039              Vietnam  VNM  2019  341004930.0  NaN
10070                Yemen  YEM  2019   12683843.0  NaN
10135               Zambia  ZMB  2019    7747163.0  NaN
10201             Zimbabwe  ZWE  2019   11114607.0  NaN

                         Region  GDP (output, multiple price benchmarks)
49          Europe & Central Asia                              3.610304e+10
109    Middle East & North Africa                              5.074880e+11
159             Sub-Saharan Africa                             2.278560e+11
189             Sub-Saharan Africa                             2.256805e+08
239     Latin America & Caribbean                              1.603854e+09
...                           ...                                       ...
9989    Latin America & Caribbean                              7.160107e+09
10039         East Asia & Pacific                              7.241230e+11
10070  Middle East & North Africa                              5.182806e+10
10135           Sub-Saharan Africa                             5.678371e+10
10201           Sub-Saharan Africa                             4.082657e+10

[182 rows x 7 columns]


<matplotlib.collections.PathCollection at 0x20ce89898e0>
```



[ ]:
```
plt.figure(figsize=(15,10))
plt.bar(range(len(AM_1)), AM_1['HDI'], color='#eab676')
plt.xticks(range(len(AM_1)), AM_1['Entity_x'], rotation='vertical')
plt.yticks(np.arange(0.3, 1, 0.1))
plt.ylabel('Mean of the HDI')
```

[ ]:
```
Text(0, 0.5, 'Mean of the HDI')
```



[ ]:
```
from pandas_profiling import ProfileReport

USA = X_Country.drop(['Entity_x','Code','Region'], axis = (1))
profile_df = USA.profile_report()
profile_df
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[ ]:

[406]:
```
USA = X_Country.drop(['Entity_x','Code','Region','Year'], axis = (1))
profile_df = USA.profile_report()
profile_df
#USA data profiling without year:
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[406]:

[ ]:
```
wrld_pfl = df.drop(['Entity_x','Code','Region'], axis = (1))
profile_df = wrld_pfl.profile_report()
profile_df
#World data profiling:
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]

<IPython.core.display.HTML object>
```

[ ]:

[315]:
```
X_Country = df[df["Code"] == "USA"]
X_Country.to_csv("usa_data", index = False)
```

[403]:
```
from pycaret.regression import *
usa_data = X_Country.drop(['Entity_x', 'Region', 'HDI', 'Code'], axis=1)
```

[273]:
```
setup(data=usa_data, target='CO2')
exp_reg = compare_models()
#consider a linear regresson model
lr = create_model('lr')
```

```
<pandas.io.formats.style.Styler at 0x20ce3327a90>
```

```
<IPython.core.display.HTML object>

<pandas.io.formats.style.Styler at 0x20cdff99400>

Processing:   0%|          | 0/85 [00:00<?, ?it/s]

<IPython.core.display.HTML object>

<pandas.io.formats.style.Styler at 0x20ce1ee9fa0>

Processing:   0%|          | 0/4 [00:00<?, ?it/s]
```

[ ]:
```
lr = create_model('lr')
```

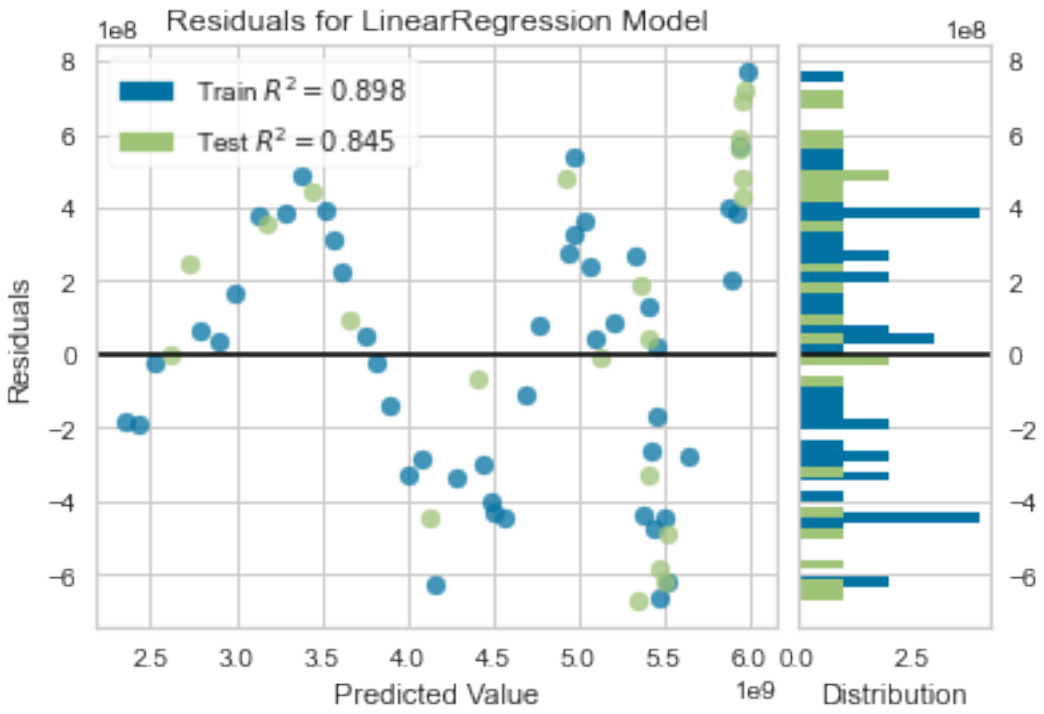
```
<IPython.core.display.HTML object>

<pandas.io.formats.style.Styler at 0x20cca8d0880>

Processing:   0%|          | 0/4 [00:00<?, ?it/s]
```

[ ]:
```
lr_tuned = tune_model(lr)
```

```
<IPython.core.display.HTML object>

<pandas.io.formats.style.Styler at 0x20cd25f3790>

Processing:   0%|          | 0/7 [00:00<?, ?it/s]

Fitting 10 folds for each of 4 candidates, totalling 40 fits
Original model was better than the tuned model, hence it will be returned. NOTE:
The display metrics are for the tuned model (not the original one).
```

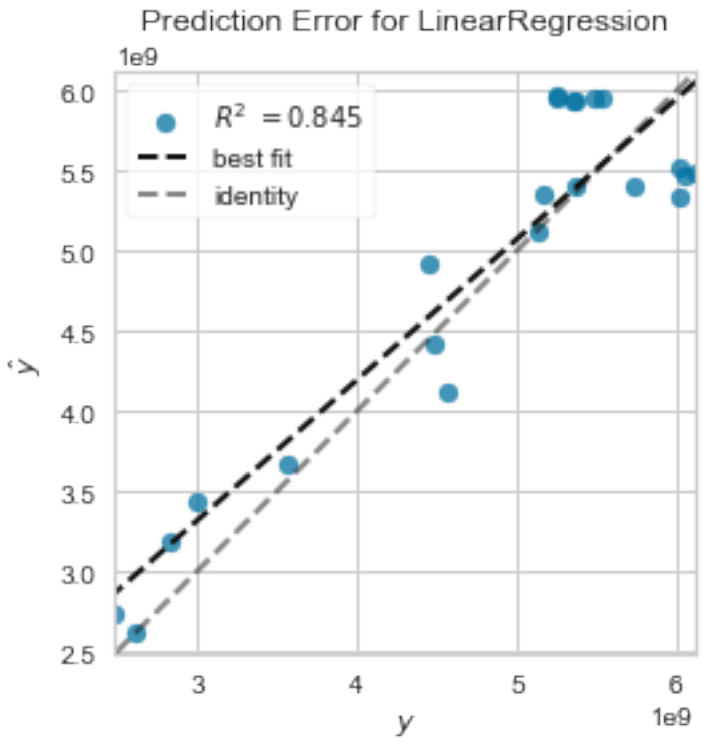
[ ]:
```
#Below is the snipet of generating the corresponding graphs
plot_model(lr_tuned)
```

```
<IPython.core.display.HTML object>
```



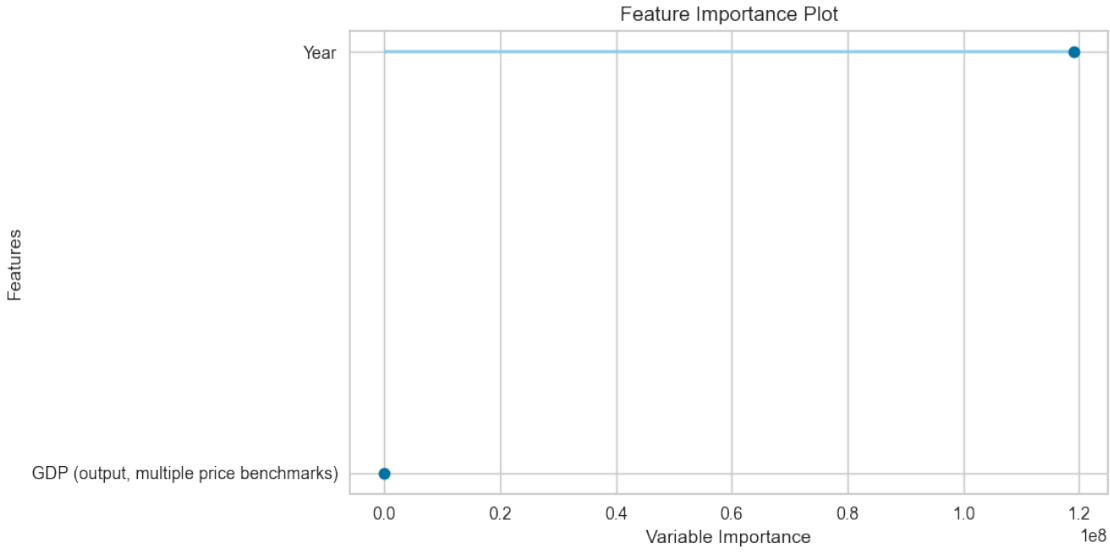
[ ]:
```
plot_model(lr_tuned, plot = 'error')
```

```
<IPython.core.display.HTML object>
```



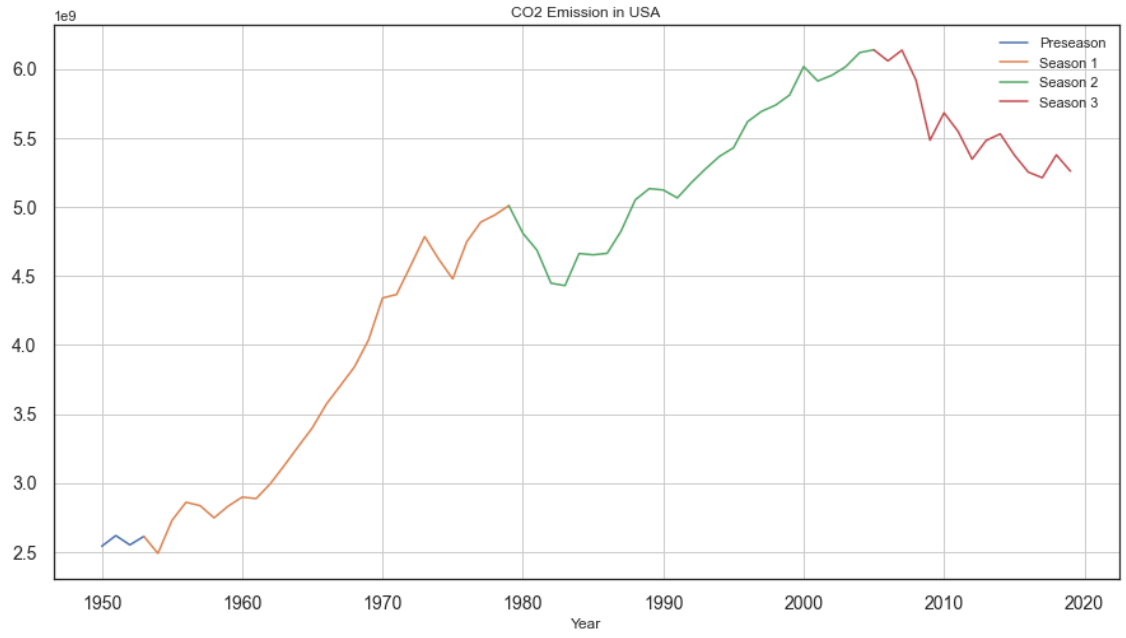
[ ]:
```
plot_model(lr_tuned, plot = 'feature')
#The Year feature is the most important feature to do this regression analysis
```

```
<IPython.core.display.HTML object>
```



[ ]:
```
final_lr = finalize_model(lr_tuned)
```

[405]:
```
preseason.CO2.plot(figsize=(15,8), title= 'CO2 Emission in USA', fontsize=14,␣
↪label='Preseason')
season1.CO2.plot(figsize=(15,8), title= 'CO2 Emission in USA', fontsize=14,␣
↪label='Season 1')
season2.CO2.plot(figsize=(15,8), title= 'CO2 Emission in USA', fontsize=14,␣
↪label='Season 2')
season3.CO2.plot(figsize=(15,8), title= 'CO2 Emission in USA', fontsize=14,␣
↪label='Season 3')

plt.legend()
plt.grid()
plt.show()
plt.savefig('time series.png')
```



<Figure size 432x288 with 0 Axes>

[ ]:

Overview

Dataset statistics

Number of variables	5
Number of observations	70
Missing cells	40
Missing cells (%)	11.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.9 KiB
Average record size in memory	41.8 B

Variable types

Numeric	5

Alerts

`df_index` is highly correlated with `Year` and 3 other fields (Year, CO2, HDI, GDP (output, multiple price benchmarks))	**High correlation**
`Year` is highly correlated with `df_index` and 3 other fields (df_index, CO2, HDI, GDP (output, multiple price benchmarks))	**High correlation**
`CO2` is highly correlated with `df_index` and 3 other fields (df_index, Year, HDI, GDP (output, multiple price benchmarks))	**High correlation**
`HDI` is highly correlated with `df_index` and 3 other fields (df_index, Year, CO2, GDP (output, multiple price benchmarks))	**High correlation**
`GDP (output, multiple price benchmarks)` is highly correlated with `df_index` and 3 other fields (df_index, Year, CO2, HDI)	**High correlation**
`HDI` has 40 (57.1%) missing values	**Missing**
`df_index` is uniformly distributed	**Uniform**
`Year` is uniformly distributed	**Uniform**
`df_index` has unique values	**Unique**
`Year` has unique values	**Unique**
`CO2` has unique values	**Unique**
`GDP (output, multiple price benchmarks)` has unique values	**Unique**

Reproduction

Analysis started	2022-12-18 05:10:53.926382
Analysis finished	2022-12-18 05:11:02.472882
Duration	8.55 seconds
Software version	pandas-profiling v3.4.0 (https://github.com/pandas-profiling/pandas-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22Pandas%20Profiling%20Report%22%2C%20%22dataset%22%3A%20%7B%22description%22%3A%20%22%22%2C%20%22cr

Variables

Select Columns

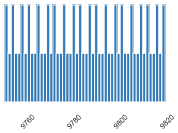
df_index

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION (This variable has a high correlation with 4 fields: Year, CO2, HDI, GDP (output, multiple price benchmarks))
UNIFORM
UNIQUE

Distinct	70
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	9784.5

Minimum	9750
Maximum	9819
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	688.0 B



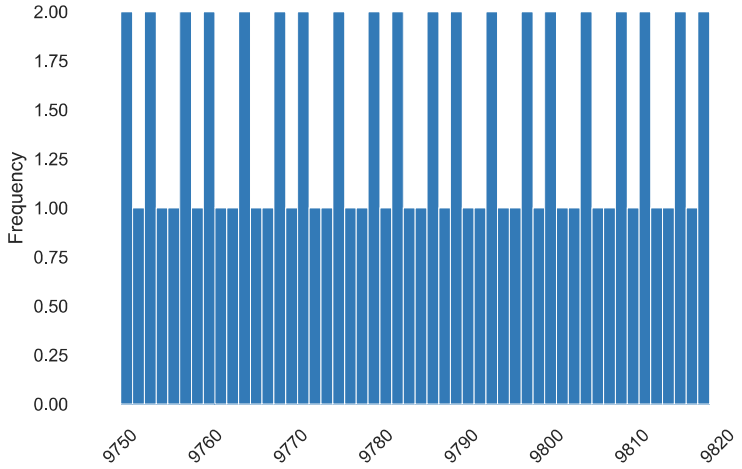
Quantile statistics

Minimum	9750
5-th percentile	9753.45
Q1	9767.25
median	9784.5
Q3	9801.75
95-th percentile	9815.55
Maximum	9819
Range	69

Descriptive statistics

Standard deviation	20.35108515
Coefficient of variation (CV)	0.002079931028
Kurtosis	-1.2
Mean	9784.5
Median Absolute Deviation (MAD)	17.5
Skewness	0
Sum	684915
Variance	414.1666667

Interquartile range (IQR)	34.5	**Monotonicity**	Strictly increasing



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
9750	1	1.4%
9794	1	1.4%
9800	1	1.4%
9799	1	1.4%
9798	1	1.4%
9797	1	1.4%
9796	1	1.4%
9795	1	1.4%
9793	1	1.4%
9802	1	1.4%
Other values (60)	60	85.7%

Value	Count	Frequency (%)
9750	1	1.4%

Value	Count	Frequency (%)
9751	1	1.4%
9752	1	1.4%
9753	1	1.4%
9754	1	1.4%
9755	1	1.4%
9756	1	1.4%
9757	1	1.4%
9758	1	1.4%
9759	1	1.4%

Value	Count	Frequency (%)
9819	1	1.4%
9818	1	1.4%
9817	1	1.4%
9816	1	1.4%
9815	1	1.4%
9814	1	1.4%
9813	1	1.4%
9812	1	1.4%
9811	1	1.4%
9810	1	1.4%

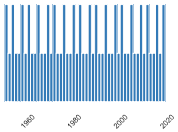
Year

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION (This variable has a high correlation with 4 fields: df_index, CO2, HDI, GDP (output, multiple price benchmarks))
UNIFORM
UNIQUE

Distinct	70
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1984.5

Minimum	1950
Maximum	2019
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	688.0 B

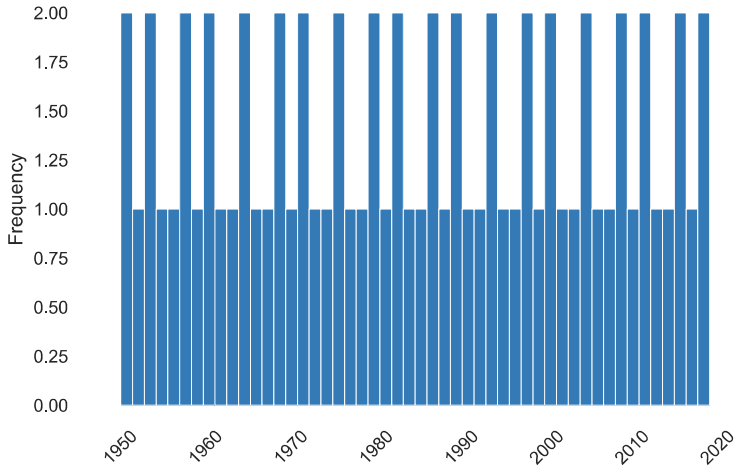


Quantile statistics

Minimum	1950
5-th percentile	1953.45
Q1	1967.25
median	1984.5
Q3	2001.75
95-th percentile	2015.55
Maximum	2019
Range	69
Interquartile range (IQR)	34.5

Descriptive statistics

Standard deviation	20.35108515
Coefficient of variation (CV)	0.01025501897
Kurtosis	-1.2
Mean	1984.5
Median Absolute Deviation (MAD)	17.5
Skewness	0
Sum	138915
Variance	414.1666667
Monotonicity	Strictly increasing



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
1950	1	1.4%
1994	1	1.4%
2000	1	1.4%
1999	1	1.4%
1998	1	1.4%
1997	1	1.4%
1996	1	1.4%
1995	1	1.4%
1993	1	1.4%
2002	1	1.4%
Other values (60)	60	85.7%

Value	Count	Frequency (%)
1950	1	1.4%
1951	1	1.4%
1952	1	1.4%

Value	Count	Frequency (%)
1953	1	1.4%
1954	1	1.4%
1955	1	1.4%
1956	1	1.4%
1957	1	1.4%
1958	1	1.4%
1959	1	1.4%

Value	Count	Frequency (%)
2019	1	1.4%
2018	1	1.4%
2017	1	1.4%
2016	1	1.4%
2015	1	1.4%
2014	1	1.4%
2013	1	1.4%
2012	1	1.4%
2011	1	1.4%
2010	1	1.4%

CO2

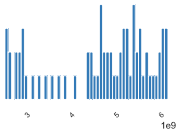
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION (This variable has a high correlation with 4 fields: df_index, Year, HDI, GDP (output, multiple price benchmarks))

UNIQUE

Distinct	70
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	4610535496

Minimum	2489462300
Maximum	6137603600
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	688.0 B

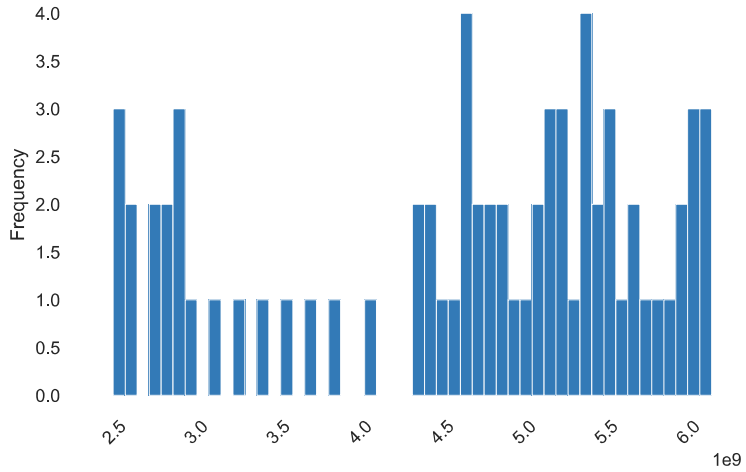


Quantile statistics

Minimum	2489462300
5-th percentile	2615554435
Q1	3739325650
median	4857628850
Q3	5467644125
95-th percentile	6038797630
Maximum	6137603600
Range	3648141300
Interquartile range (IQR)	1728318475

Descriptive statistics

Standard deviation	1130548933
Coefficient of variation (CV)	0.2452098968
Kurtosis	-0.9183033697
Mean	4610535496
Median Absolute Deviation (MAD)	680121150
Skewness	-0.5934579331
Sum	$3.227374847 \times 10^{11}$
Variance	$1.27814089 \times 10^{18}$
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
2541485300	1	1.4%
5365579000	1	1.4%
6016350700	1	1.4%
5810331600	1	1.4%
5737129500	1	1.4%
5691864600	1	1.4%
5616430600	1	1.4%
5427798500	1	1.4%
5274363000	1	1.4%
5952699000	1	1.4%
Other values (60)	60	85.7%

Value	Count	Frequency (%)
2489462300	1	1.4%
2541485300	1	1.4%
2551219500	1	1.4%

Value	Count	Frequency (%)
2612971300	1	1.4%
2618711600	1	1.4%
2728512000	1	1.4%
2747107600	1	1.4%
2831923200	1	1.4%
2835772200	1	1.4%
2859995000	1	1.4%

Value	Count	Frequency (%)
6137603600	1	1.4%
6135287300	1	1.4%
6117963000	1	1.4%
6057163300	1	1.4%
6016350700	1	1.4%
6015804400	1	1.4%
5952699000	1	1.4%
5918868500	1	1.4%
5911988000	1	1.4%
5810331600	1	1.4%

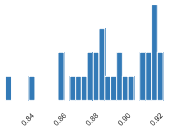
HDI

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
(This variable has a high correlation with 4 fields: df_index, Year, CO2, GDP (output, multiple price benchmarks))
MISSING

Distinct	27
Distinct (%)	90.0%
Missing	40
Missing (%)	57.1%
Infinite	0
Infinite (%)	0.0%
Mean	0.8909666667

Minimum	0.826
Maximum	0.924
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	688.0 B

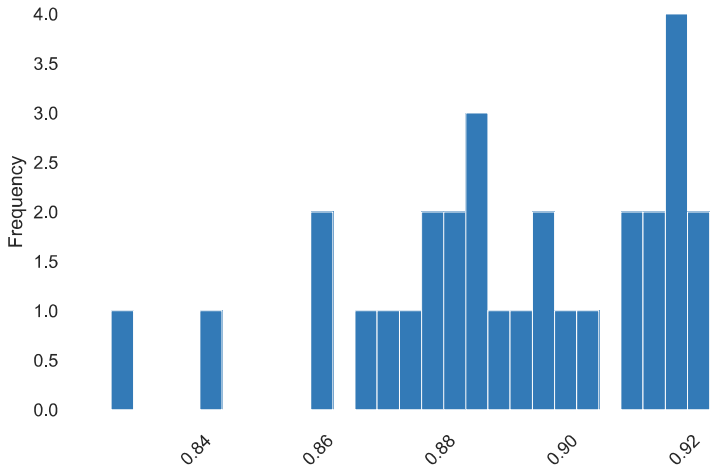


Quantile statistics

Minimum	0.826
5-th percentile	0.84955
Q1	0.8775
median	0.891
Q3	0.913
95-th percentile	0.9211
Maximum	0.924
Range	0.098
Interquartile range (IQR)	0.0355

Descriptive statistics

Standard deviation	0.02468069654
Coefficient of variation (CV)	0.02770103244
Kurtosis	0.29191704
Mean	0.8909666667
Median Absolute Deviation (MAD)	0.019
Skewness	-0.7228985394
Sum	26.729
Variance	0.0006091367816
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=27)

Value	Count	Frequency (%)
0.918	2	2.9%
0.91	2	2.9%
0.885	2	2.9%
0.893	1	1.4%
0.922	1	1.4%
0.92	1	1.4%
0.916	1	1.4%
0.917	1	1.4%
0.914	1	1.4%
0.905	1	1.4%
Other values (17)	17	24.3%
(Missing)	40	57.1%

Value	Count	Frequency (%)
0.826	1	1.4%
0.841	1	1.4%

Value	Count	Frequency (%)
0.86	1	1.4%
0.861	1	1.4%
0.867	1	1.4%
0.871	1	1.4%
0.875	1	1.4%
0.877	1	1.4%
0.879	1	1.4%
0.881	1	1.4%

Value	Count	Frequency (%)
0.924	1	1.4%
0.922	1	1.4%
0.92	1	1.4%
0.918	2	2.9%
0.917	1	1.4%
0.916	1	1.4%
0.914	1	1.4%
0.91	2	2.9%
0.905	1	1.4%
0.901	1	1.4%

GDP (output, multiple price benchmarks)

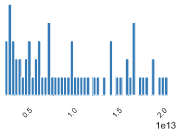
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION (This variable has a high correlation with 4 fields: df_index, Year, CO2, HDI)

UNIQUE

Distinct	70
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	$9.626140571 \times 10^{12}$

Minimum	2.47563×10^{12}
Maximum	2.05958×10^{13}
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	688.0 B

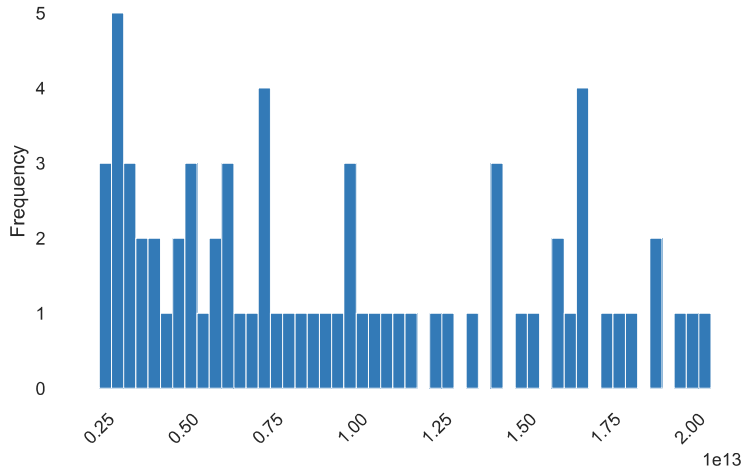


Quantile statistics

Minimum	2.47563×10^{12}
5-th percentile	2.8735045×10^{12}
Q1	4.98286×10^{12}
median	8.38458×10^{12}
Q3	1.4349025×10^{13}
95-th percentile	1.895579×10^{13}
Maximum	2.05958×10^{13}
Range	1.812017×10^{13}
Interquartile range (IQR)	9.366165×10^{12}

Descriptive statistics

Standard deviation	$5.470355334 \times 10^{12}$
Coefficient of variation (CV)	0.5682812643
Kurtosis	-1.12221439
Mean	$9.626140571 \times 10^{12}$
Median Absolute Deviation (MAD)	4.4621×10^{12}
Skewness	0.4418840404
Sum	6.7382984×10^{14}
Variance	$2.992478748 \times 10^{25}$
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

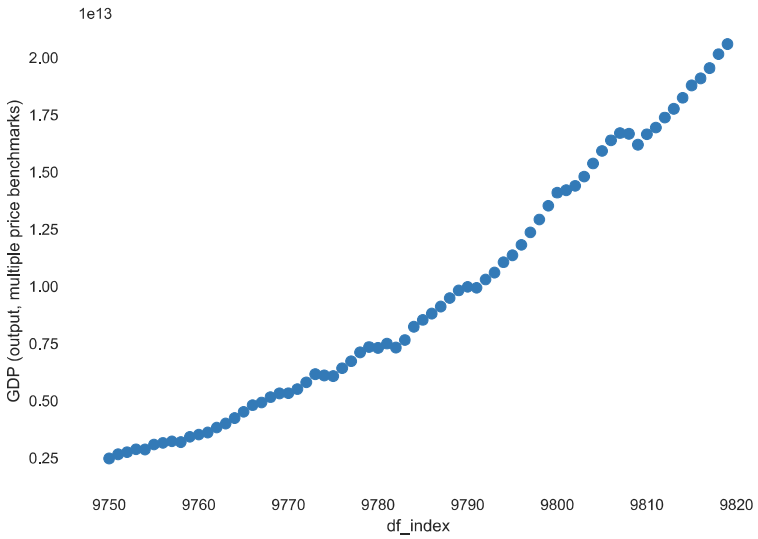
Value	Count	Frequency (%)
2.47563×10^{12}	1	1.4%
1.10546×10^{13}	1	1.4%
1.4096×10^{13}	1	1.4%
1.35268×10^{13}	1	1.4%
1.29243×10^{13}	1	1.4%
1.23603×10^{13}	1	1.4%
1.18158×10^{13}	1	1.4%
1.13617×10^{13}	1	1.4%
1.06025×10^{13}	1	1.4%
1.43969×10^{13}	1	1.4%
Other values (60)	60	85.7%

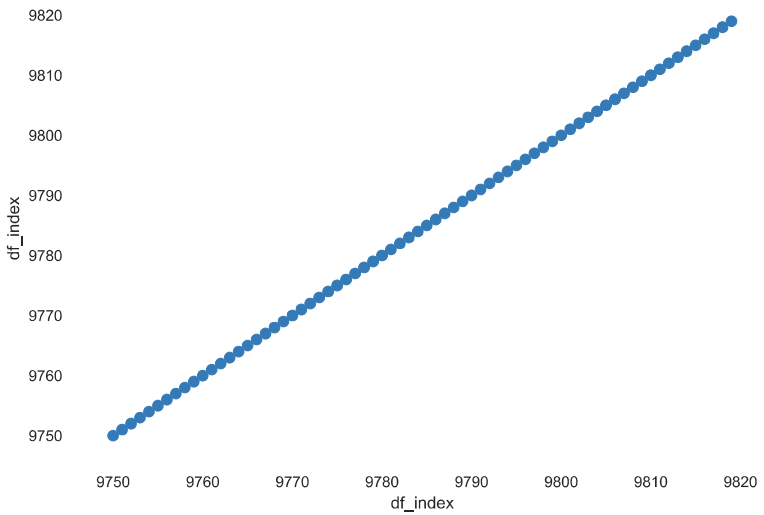
Value	Count	Frequency (%)
2.47563×10^{12}	1	1.4%
2.66082×10^{12}	1	1.4%
2.75193×10^{12}	1	1.4%

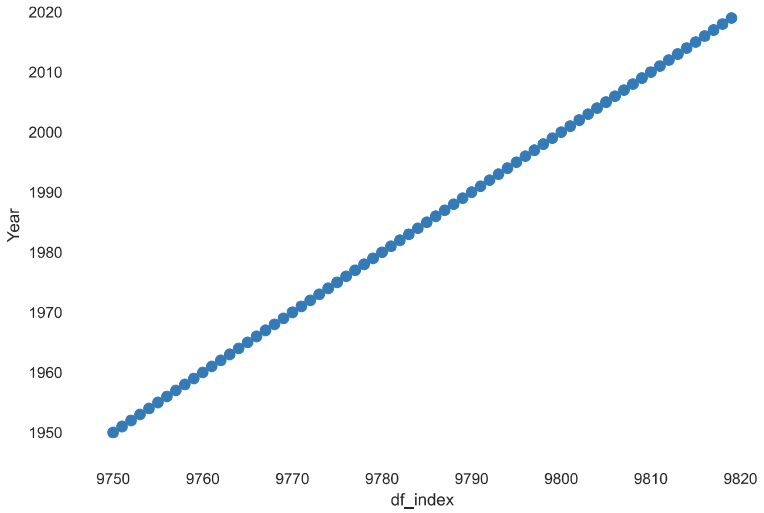
Value	Count	Frequency (%)
2.86954×10^{12}	1	1.4%
2.87835×10^{12}	1	1.4%
3.08656×10^{12}	1	1.4%
3.15759×10^{12}	1	1.4%
3.19375×10^{12}	1	1.4%
3.2264×10^{12}	1	1.4%
3.42602×10^{12}	1	1.4%

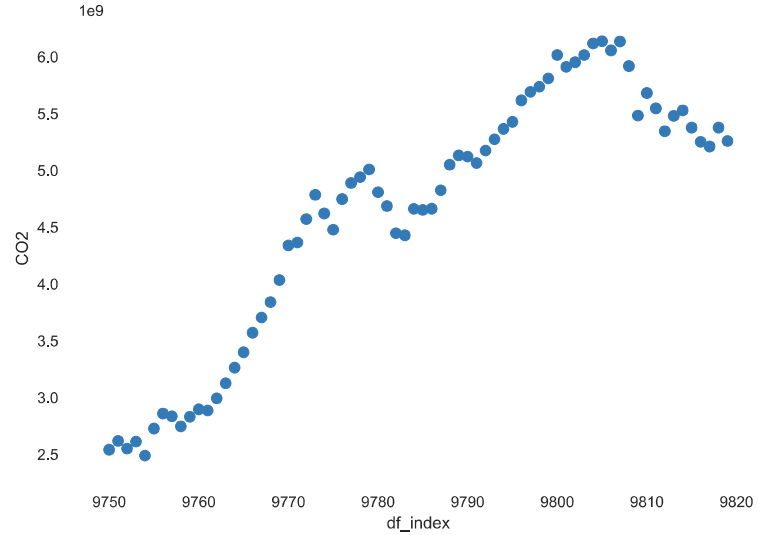
Value	Count	Frequency (%)
2.05958×10^{13}	1	1.4%
2.01553×10^{13}	1	1.4%
1.9543×10^{13}	1	1.4%
1.90952×10^{13}	1	1.4%
1.87854×10^{13}	1	1.4%
1.82442×10^{13}	1	1.4%
1.7764×10^{13}	1	1.4%
1.7383×10^{13}	1	1.4%
1.69439×10^{13}	1	1.4%
1.67024×10^{13}	1	1.4%

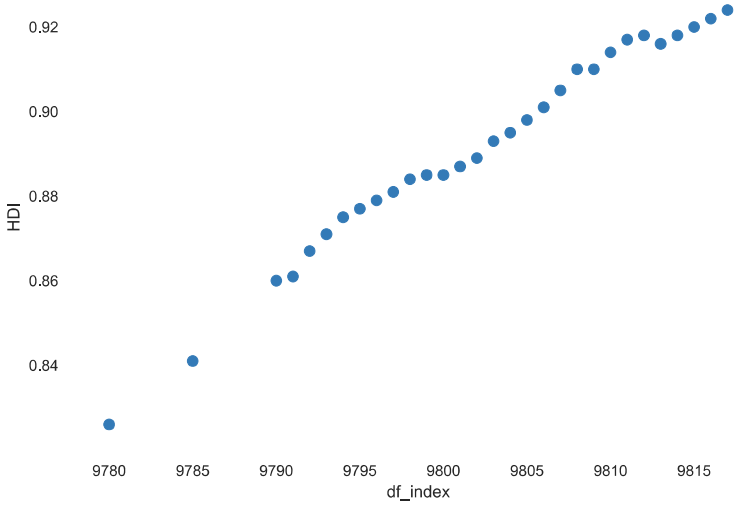
Interactions

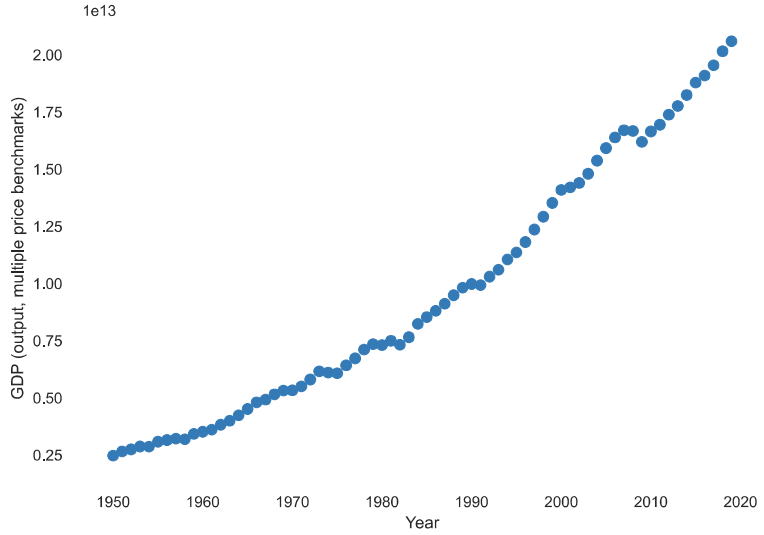


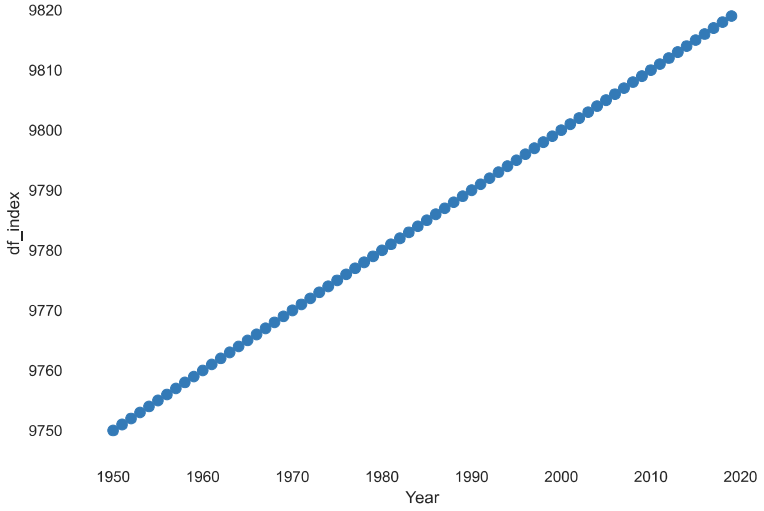

9820
9810
9800
9790
9780
9770
9760
9750
df_index
9750
9760
9770
9780
9790
9800
9810
9820
df_index

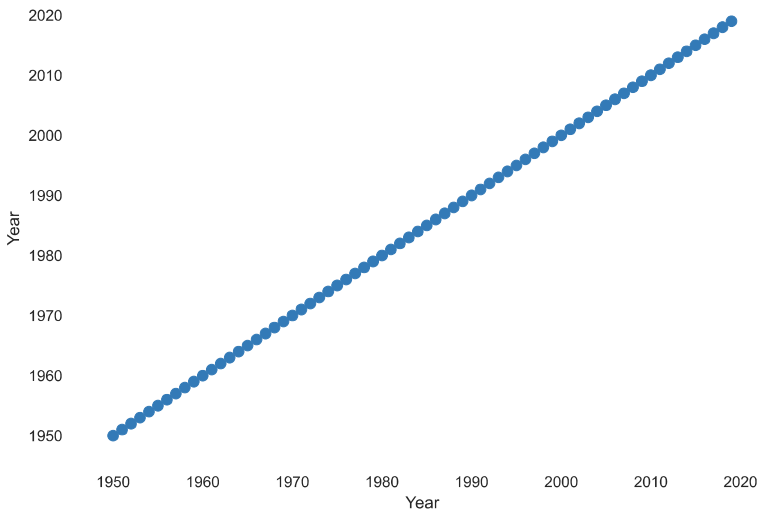

2020
2010
2000
1990
1980
1970
1960
1950
Year
9750
9760
9770
9780
9790
9800
9810
9820
df_index

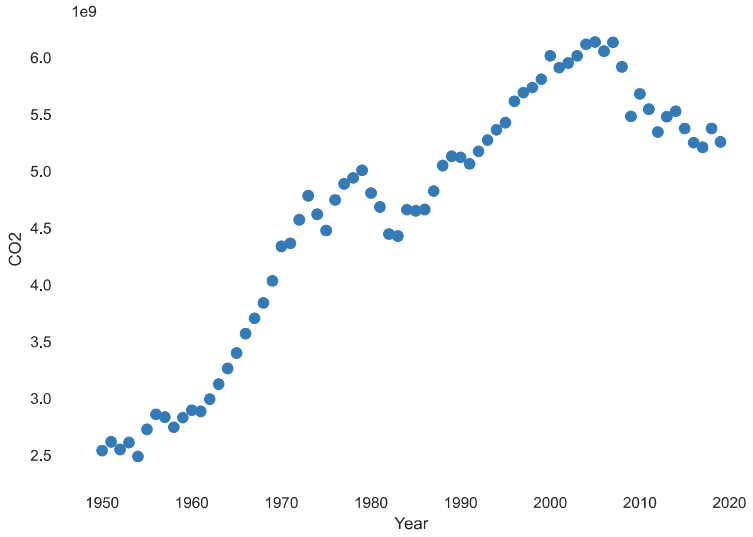

1e9
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
CO2
9750
9760
9770
9780
9790
9800
9810
9820
df_index

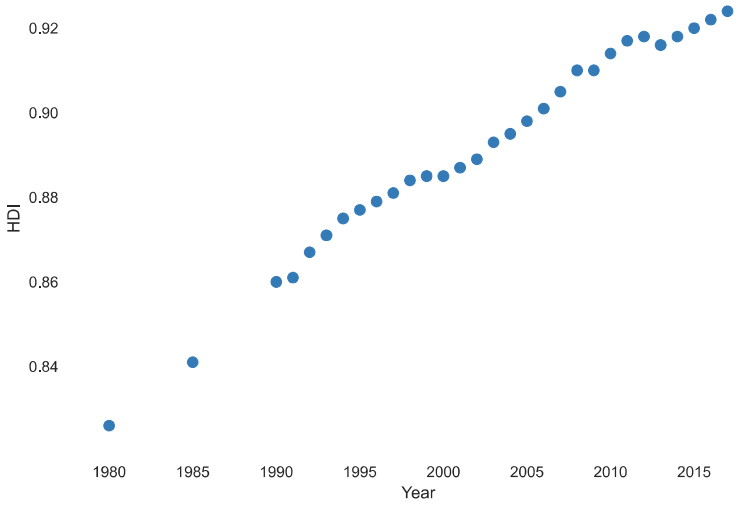

0.92
0.90
0.88
0.86
0.84
HDI
9780
9785
9790
9795
9800
9805
9810
9815
df_index

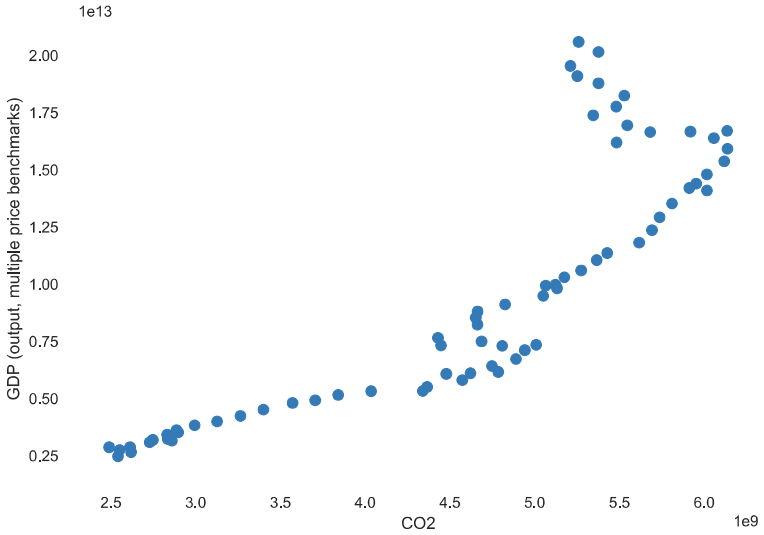

1e13
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
GDP (output, multiple price benchmarks)
1950
1960
1970
1980
1990
2000
2010
2020
Year

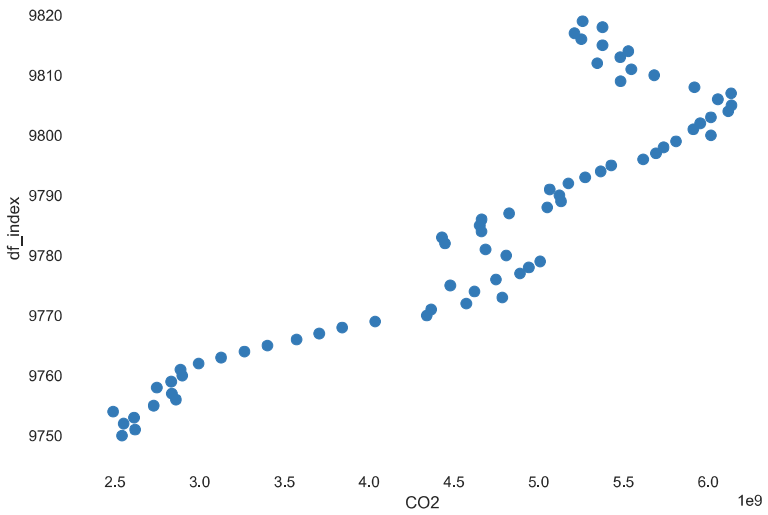

9820
9810
9800
9790
9780
9770
9760
9750
df_index
1950
1960
1970
1980
1990
2000
2010
2020
Year

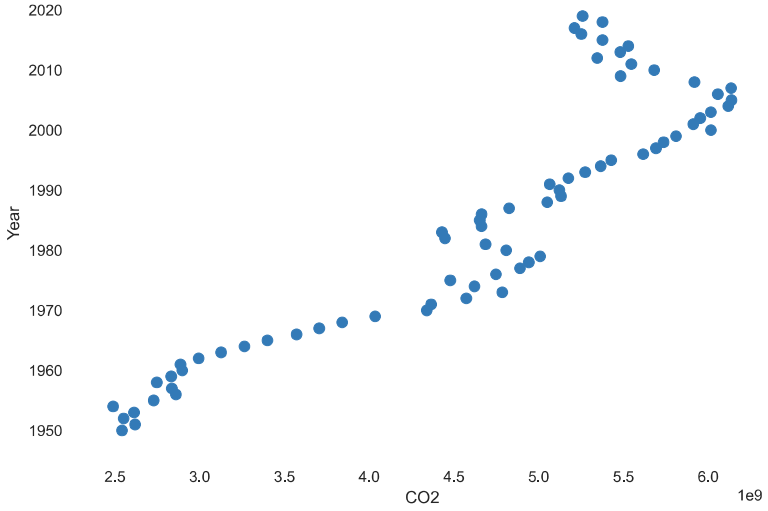

2020
2010
2000
1990
1980
1970
1960
1950
Year
1950
1960
1970
1980
1990
2000
2010
2020
Year

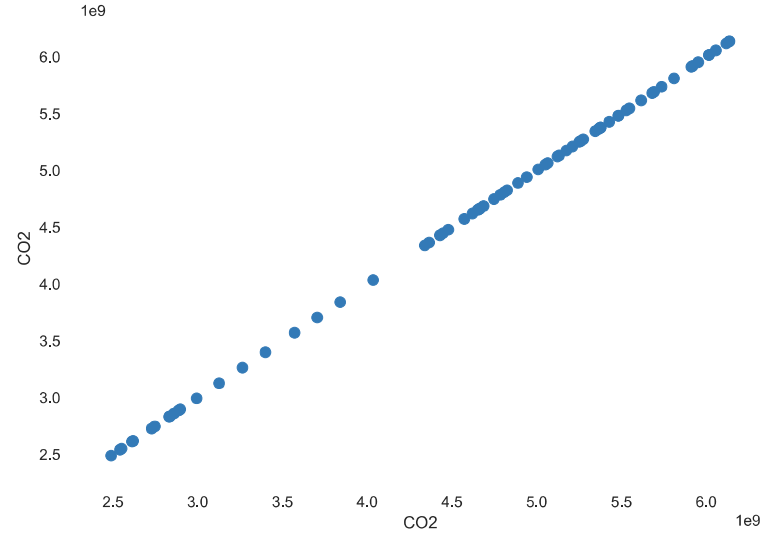

1e9
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
CO2
1950
1960
1970
1980
1990
2000
2010
2020
Year

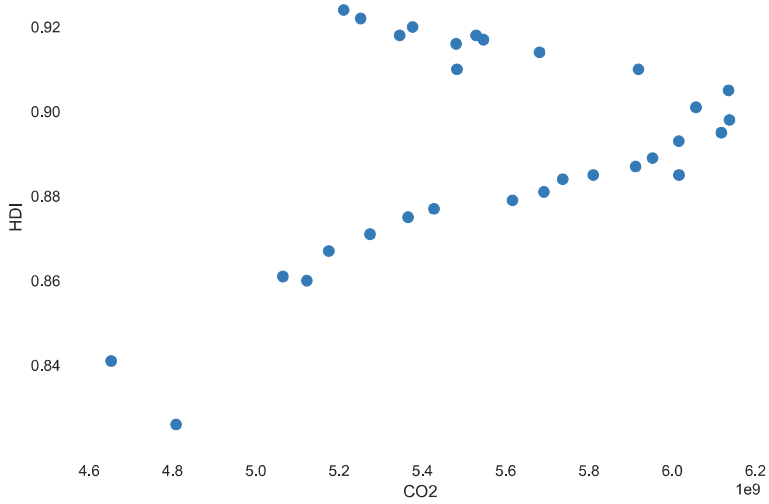

0.92
0.90
0.88
0.86
0.84
HDI
1980
1985
1990
1995
2000
2005
2010
2015
Year

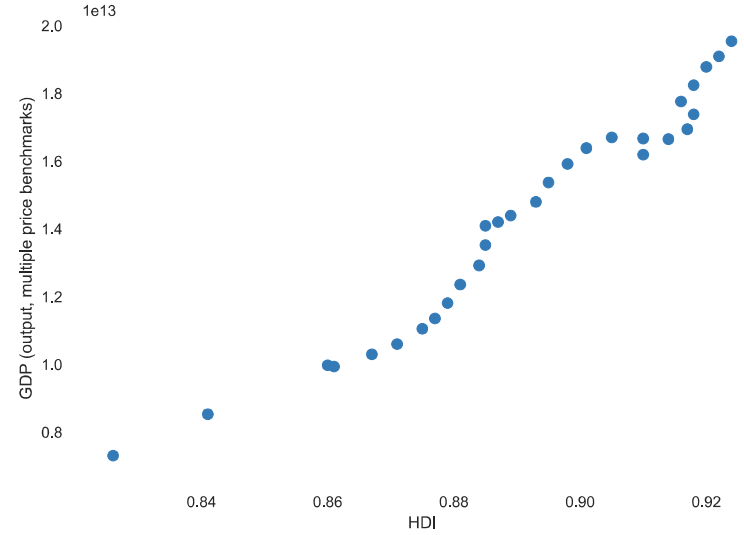

1e13
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
GDP (output, multiple price benchmarks)
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e9
CO2

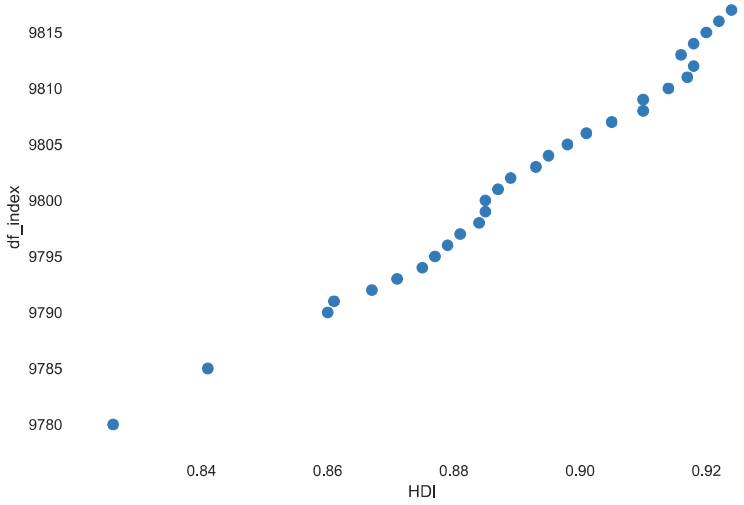

9820
9810
9800
9790
9780
9770
9760
9750
df_index
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e9
CO2

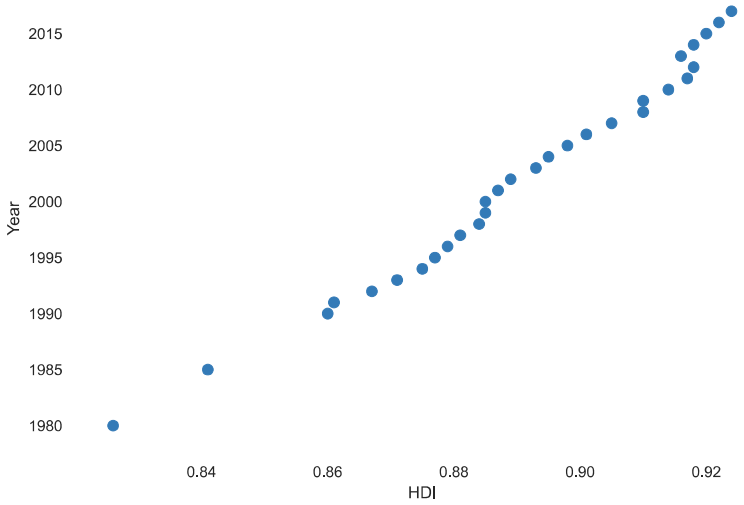

2020
2010
2000
1990
1980
1970
1960
1950
Year
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e9
CO2

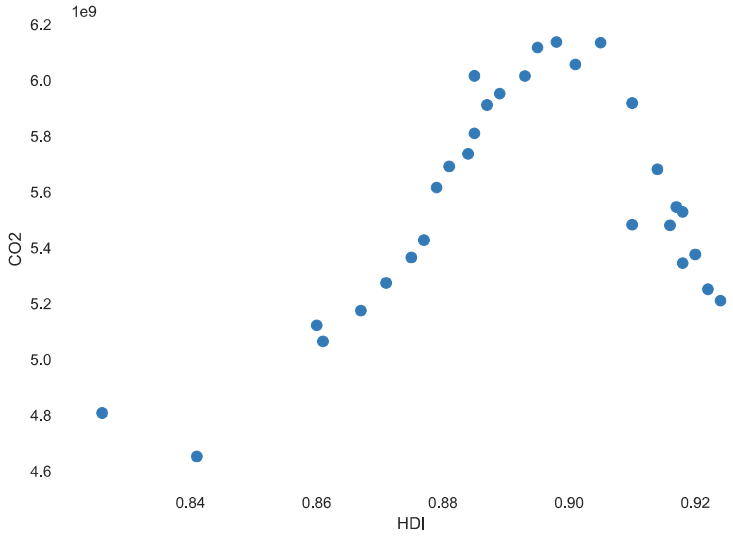

1e9
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
CO2
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e9
CO2

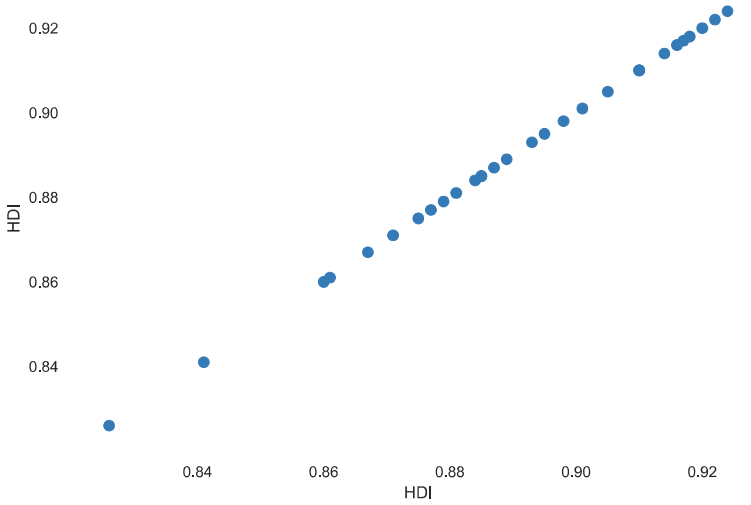

0.92
0.90
0.88
0.86
0.84
HDI
4.6
4.8
5.0
5.2
5.4
5.6
5.8
6.0
6.2
1e9
CO2

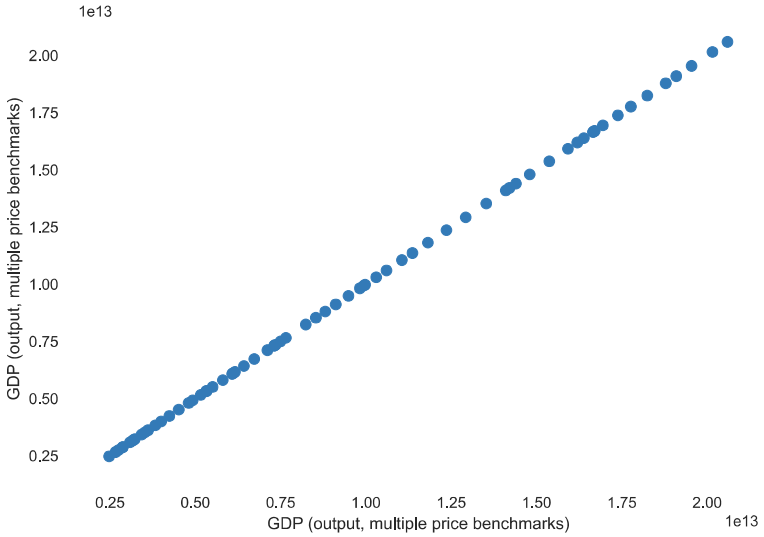

1e13
2.0
1.8
1.6
1.4
1.2
1.0
0.8
GDP (output, multiple price benchmarks)
0.84
0.86
0.88
0.90
0.92
HDI

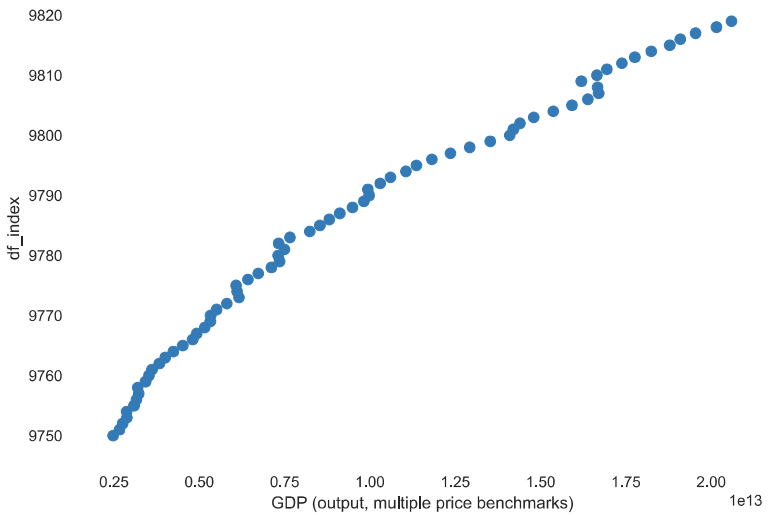

9815
9810
9805
9800
9795
9790
9785
9780
df_index
0.84
0.86
0.88
0.90
0.92
HDI

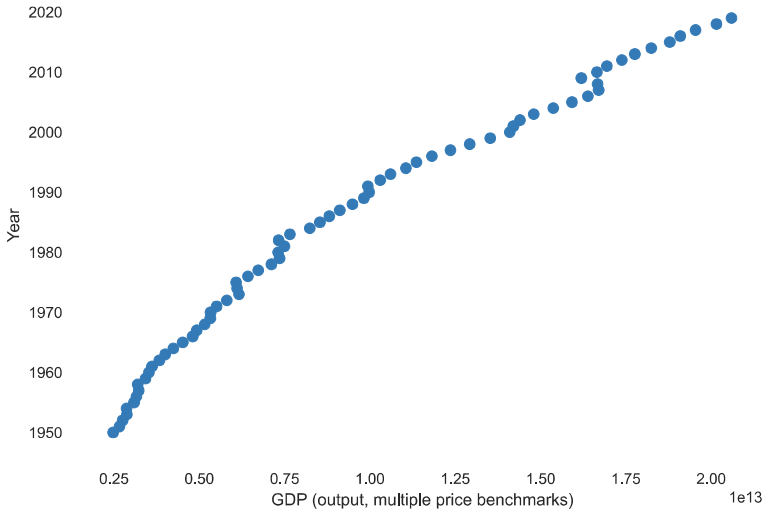

2015
2010
2005
2000
1995
1990
1985
1980
Year
0.84
0.86
0.88
0.90
0.92
HDI

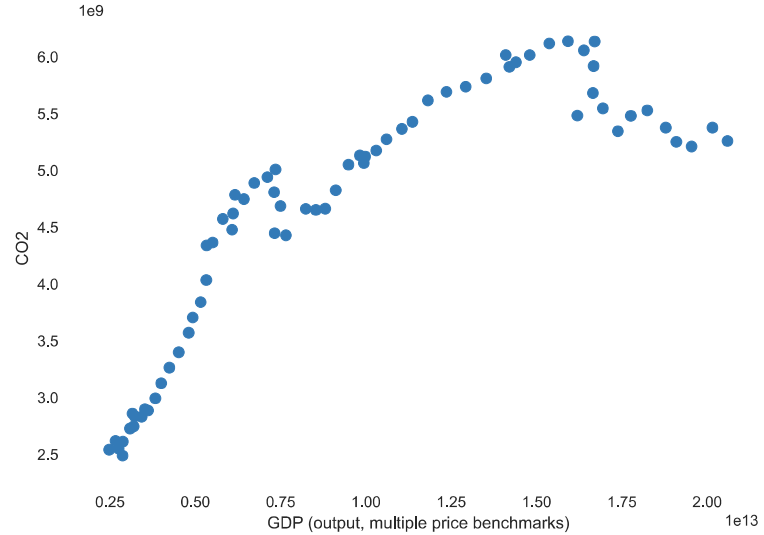

1e9
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
CO2
0.84
0.86
0.88
0.90
0.92
HDI

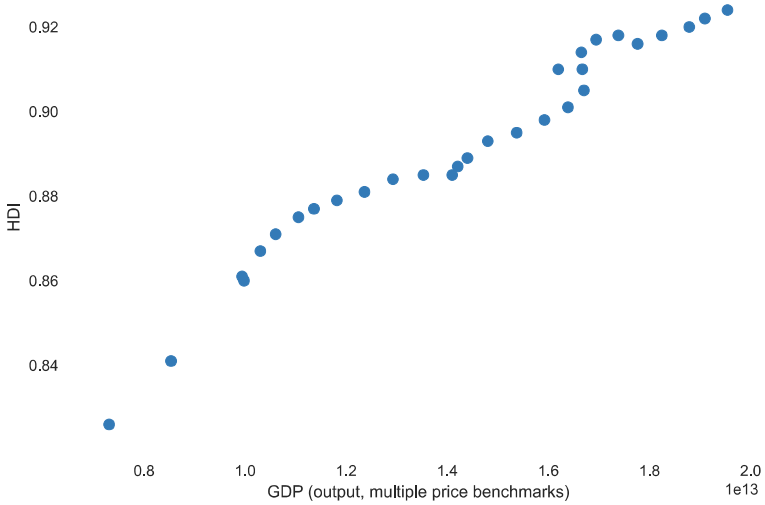

0.92
0.90
0.88
0.86
0.84
HDI
0.84
0.86
0.88
0.90
0.92
HDI


1e13
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
GDP (output, multiple price benchmarks)
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e13
GDP (output, multiple price benchmarks)

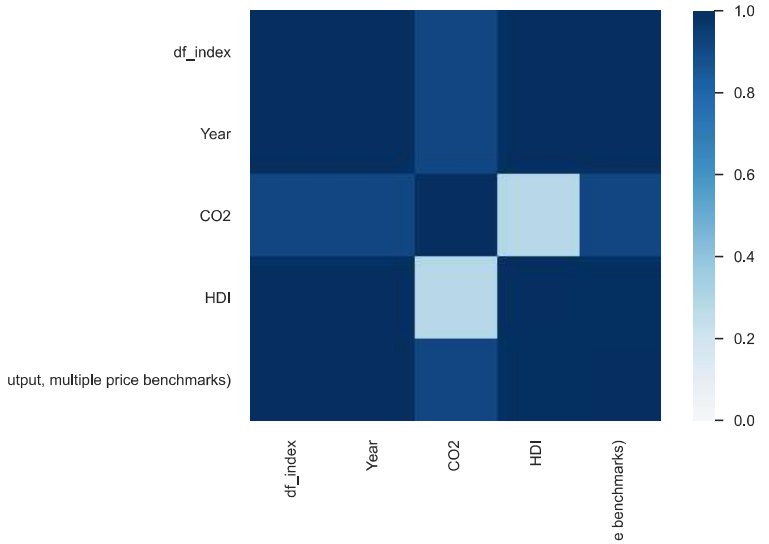

9820
9810
9800
9790
9780
9770
9760
9750
df_index
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e13
GDP (output, multiple price benchmarks)

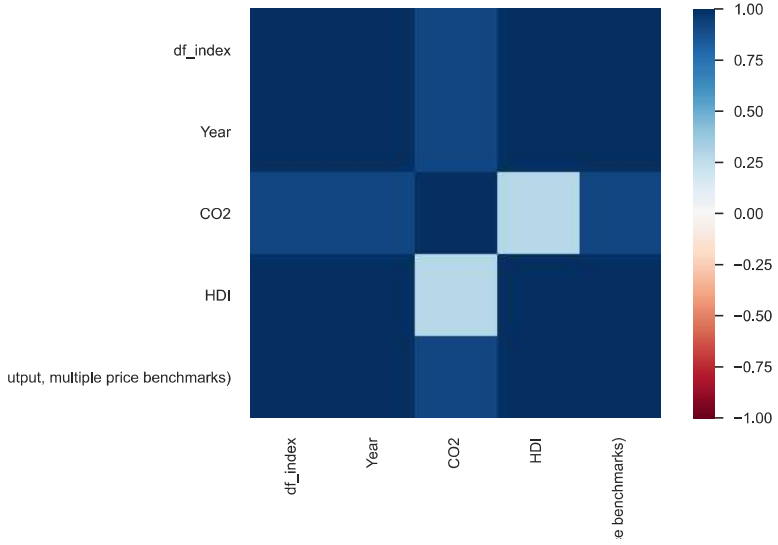

2020
2010
2000
1990
1980
1970
1960
1950
Year
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e13
GDP (output, multiple price benchmarks)

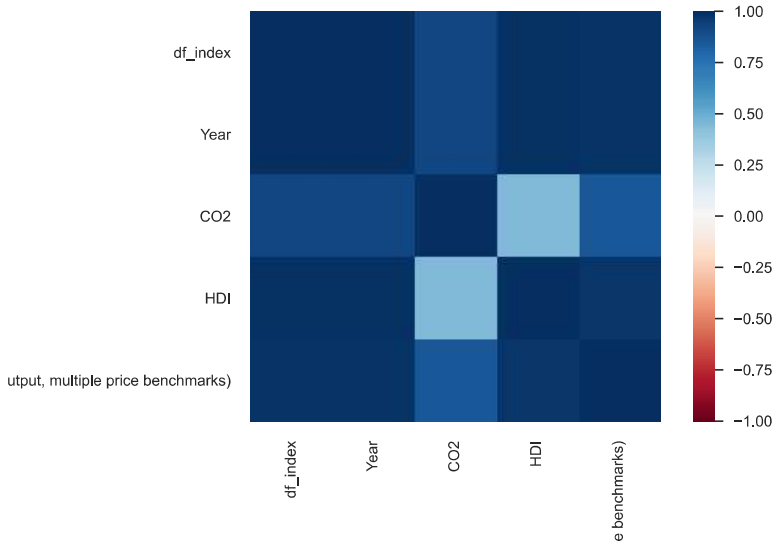


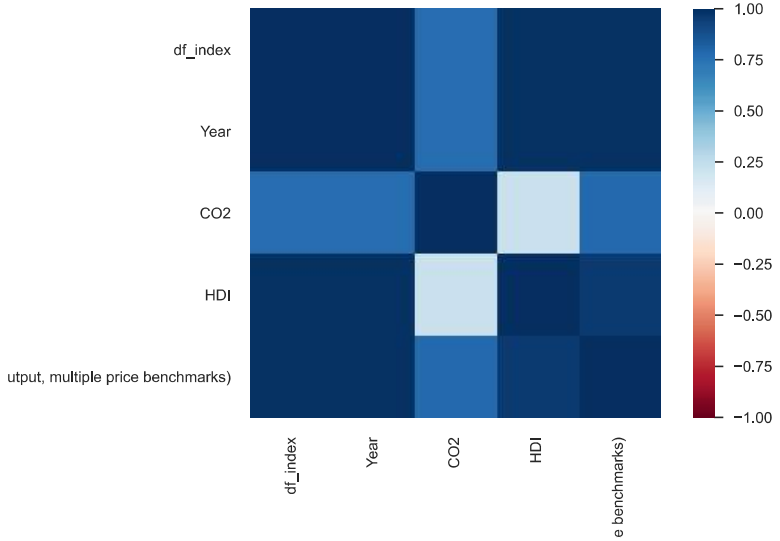


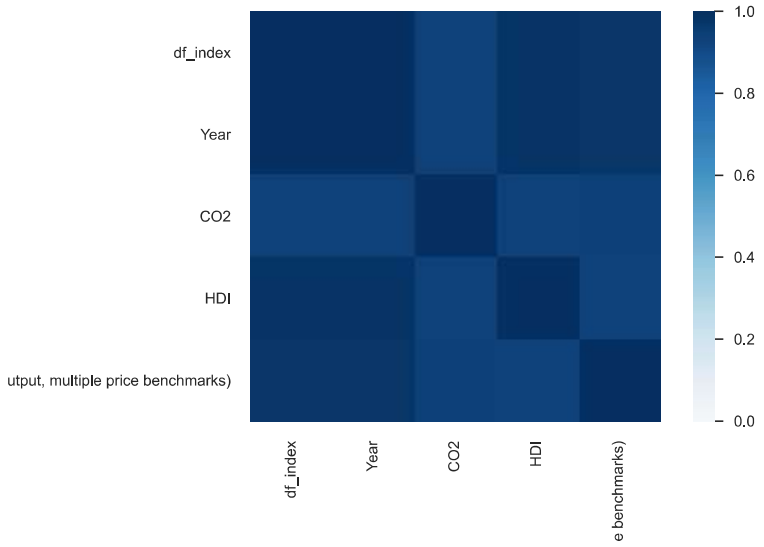
Correlations


df_index
Year
CO2
HDI
utput, multiple price benchmarks)
1.0
0.8
0.6
0.4
0.2
0.0
df_index
Year
CO2
HDI
e benchmarks)

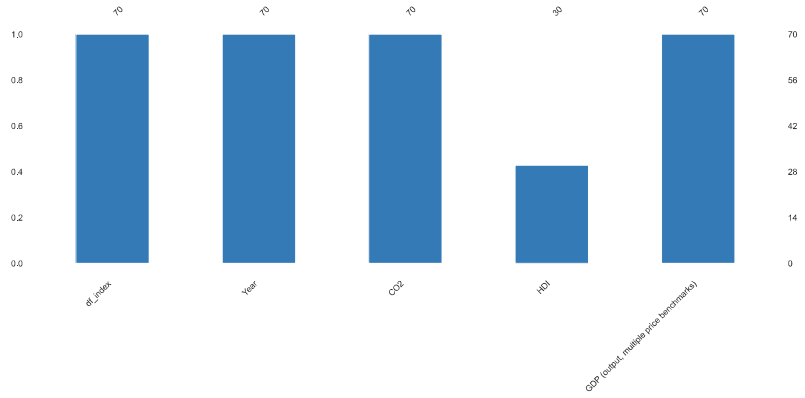

df_index
Year
CO2
HDI
utput, multiple price benchmarks)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
df_index
Year
CO2
HDI
e benchmarks)


df_index
Year
CO2
HDI
utput, multiple price benchmarks)
df_index
Year
CO2
HDI
e benchmarks)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

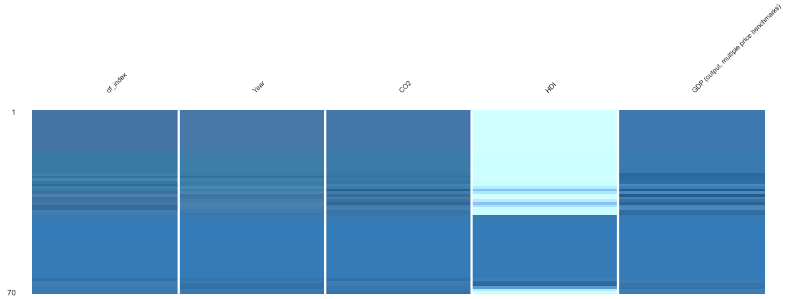

df_index
Year
CO2
HDI
utput, multiple price benchmarks)
df_index
Year
CO2
HDI
e benchmarks)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



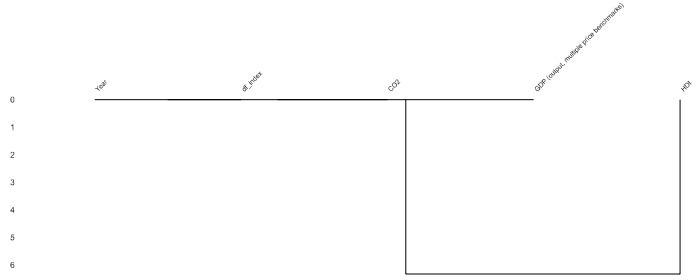
Missing values



A simple visualization of nullity by column.



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.



The dendrogram allows you to more fully correlate variable completion, revealing trends deeper than the pairwise ones visible in the correlation heatmap.

Sample

First rows

	df_index	Year	CO2	HDI	GDP (output, multiple price benchmarks)
0	9750	1950	2.541485e+09	NaN	2.475630e+12
1	9751	1951	2.618712e+09	NaN	2.660820e+12
2	9752	1952	2.551220e+09	NaN	2.751930e+12
3	9753	1953	2.612971e+09	NaN	2.878350e+12
4	9754	1954	2.489462e+09	NaN	2.869540e+12
5	9755	1955	2.728512e+09	NaN	3.086560e+12
6	9756	1956	2.859995e+09	NaN	3.157590e+12
7	9757	1957	2.835772e+09	NaN	3.226400e+12
8	9758	1958	2.747108e+09	NaN	3.193750e+12
9	9759	1959	2.831923e+09	NaN	3.426020e+12

Last rows

	df_index	Year	CO2	HDI	GDP (output, multiple price benchmarks)
60	9810	2010	5.681392e+09	0.914	1.665170e+13
61	9811	2011	5.546629e+09	0.917	1.694390e+13
62	9812	2012	5.345454e+09	0.918	1.738300e+13
63	9813	2013	5.480926e+09	0.916	1.776400e+13
64	9814	2014	5.528871e+09	0.918	1.824420e+13
65	9815	2015	5.376578e+09	0.920	1.878540e+13
66	9816	2016	5.251758e+09	0.922	1.909520e+13
67	9817	2017	5.210957e+09	0.924	1.954300e+13
68	9818	2018	5.376657e+09	NaN	2.015530e+13
69	9819	2019	5.259144e+09	NaN	2.059580e+13