Exploring Interactions between World Development Indicators

PSTAT 100 Final Project Winter 2023

Jen Rink, Roshan Mehta, Kasturi Sharma, and Jake Jensema

Author contributions

Jen Rink contributed to the Background and Aims section of the report, created the MLR model predicting GDP per capita values, wrote about her methods and did analysis on the results to write about in the discussion section, and edited and compiled the final submission.

Roshan Mehta contributed several charts exploring the relationship between country population and CO2 Emissions over time, in addition to the trend between Forest Area and CO2 Emissions. A discussion of the methods and results was also written.

Kasturi Sharm contributed to the Abstract and Background. As well as creating a graph showing the correlation between government expenditure and GDP per capita.

Jake Jensema contributed charts about the top 5 countries in CO2 emissions and created a SLR model predicting CO2 emissions with a new variable CO22_population.

Abstract

The topic we chose was world development indexes to see whether factors such as GDP or carbon dioxide emissions were correlated. More broadly, we wanted to find out whether these factors could suggest any economic, social, or climate trends. We wanted to see if factors such as education expenditure or population of a country had an affect on its GDP and how carbon dioxide emissions and forest area correlated. To approach this, we created multiple linear regression models and various different graphs to see whether we could find any correlations. With these tools, we were able to find that there are correlation within the several variables such as carbon dioxide emissions and population or education expenditure and GDP per capita.

Introduction

Background

We chose a few indicators from a collection of many to see how they correlate with one another. The indicators we chose were Governemnt Expenditure on Education, CO2 Emissions, Forest Area, and GDP per capita. The government expenditures on education indicator is expressed as a percentage of the GDP. It includes local, regional, and central governments. The CO2 emissions indicator showcases the amount of CO2 in kilotons produced in the area. The forest area indicator showcases the how many square kilometers of forest area are found within the specific country, excluding agricultural areas, urban parks, and gardens. Lastly, the indicator GDP per capita is the gross domestic produce per person. It measures the economic output of a country per person.

The motivation for analyzing data such as this is to find out what sort of effect, if any, does government expenditure on education have on a country's economy and pollution. Does an investment in education mean more people learning about harmful CO2 production and a decrease in CO2 emissions, or does the increase in education lead to more labor and therefore more CO2 emissions because of increase in capital gain? Can we make any statements about the correlation between expenditure on education and the development of a country? As students who have grown up hearing that our generation has expected to find a solution for climate change, we are curious to see how economics and green proposals like forest restoration help combat climate change.

Aims

Analyzing the trends between the variables will help us understand how they interact and influence each other. By utilizing skills learned in this course like building Multiple Linear Regression models, Complex Data Visualization, and identifying outliers we will thoroughly examine the correlations between the variables to come to some conclusions about the affect of a few World Development Indicators on GDP per capita.

By splitting Government Expenditure on Education and Population into three factored levels, we saw that less populated countries implied smaller GDP per capita values (a negative result) and largely populated countries implied larger GDP per capita values. We also saw that Year does not indicate any significant results of GDP per capita (a null result).

Materials and methods

Datasets

The data used in this project comes from 'The World Bank', specifically the World Development Indicators, which can be found here. The WDI is comprised of numerous different topics which were collected by different entities such as the UNESCO Institute for Statistics, or even the countries themselves, but are all compiled to give the most accurate picture of global estimates. The environmental related variables of the data were found using scientific equipment, while others are collected through census surveys, or transactional tracking. The dataset used for this project consists of 5 different variables (Population, GDP, GDP Expenditures for education, CO2 emissions, and Forest Area) from 2012 to 2020, collected from 217 countries/territories around the world. Below is the first four rows of the dataframe.

	Country	Year	GEOC	CO2_emissions	Forest_area	GDP per capita	Population
0	Afghanistan	2012	3.32	8080.00	12084.40	663.14	30466479
1	Albania	2012	2.93	4360.00	7849.17	4247.63	2900401
2	Algeria	2012	7.64	134929.99	19332.00	5610.73	37260563
3	American Samoa	2012	NaN	NaN	173.70	11920.06	53691
4	Andorra	2012	NaN	490.00	160.00	44904.58	71013

The statistical population includes all countries and territories around the globe existing between 2012 and 2020. The sampling frame is all countries and territories reporting economic output and environmental data between this time frame. Overall, the sample and the frame are equal (217) meaning the frame completely overlaps the population. Additionally, the sampling mechanism is a census of the frame. Our population census data also has no scope of inference. Below is a table of variable descriptions from out dataset.

Name	Variable description	Type	Units of measurement
Country Name	Name of each Country	Categorical	N/A
Year	Years range from 2012-2020	Categorical	Year
Government expenditure on education, total (% of GDP)	The total each country spends on education	Numeric	% of GDP
CO2 emissions (kt)	The total CO2 emissions each country exerts	Numeric	Kilotons
Forest area (sq. km)	The total forested area a country has	Numeric	Square Kilometers
GDP per capita	The gross domestic product of a country	Numeric	Per Capita
Population	The total estimated population of a country	Numeric	Individuals

Methods

To begin our analysis on the World Development Indicators, we fit a Multiple Linear Regression (MLR) model with GDP per capita as the response variable and Population, Year, and Government Expenditure on Education (GEOC) as the predictors to see what affect the predictors had on the response. By splitting GEOC and Population into three seperate factor levels (GEOC: 0-5% of GDP, 5-10% of GDP, 10%+ of GDP, Population: 0-3million, 3-300million, 300+million) we were able to see the differences between each group To validate our interpretations of the coefficient estimates from our MLR model, we visualized the relationship between GDP per capita and Government Expenditure on Education and Population with point-and-line plots faceted on Year.

Next, we visualized the total amount of CO2 emissions from the top 5 countries through a set of pie charts which give insight into who the biggest contributors are of CO2 pollution. Then we created a new variable 'CO2_population" which is the ratio between population and CO2 emissions in order to estimate China's CO2 emissions. With this new variable, we ran a Simple Linear Regression (SLR) model with CO2 emissions as the response variable and CO2_population as the predictor.

In the final section of our analysis, several more visualizations were created to answer specific questions about the data. Multiple scatter plots were utilized, as well as fitting a LOESS line and computing a simple linear regression fit. We wanted to investigate the effect population has on CO2 emissions over time and in addition, the trend between CO2 Emissions and Forest Area.

Results

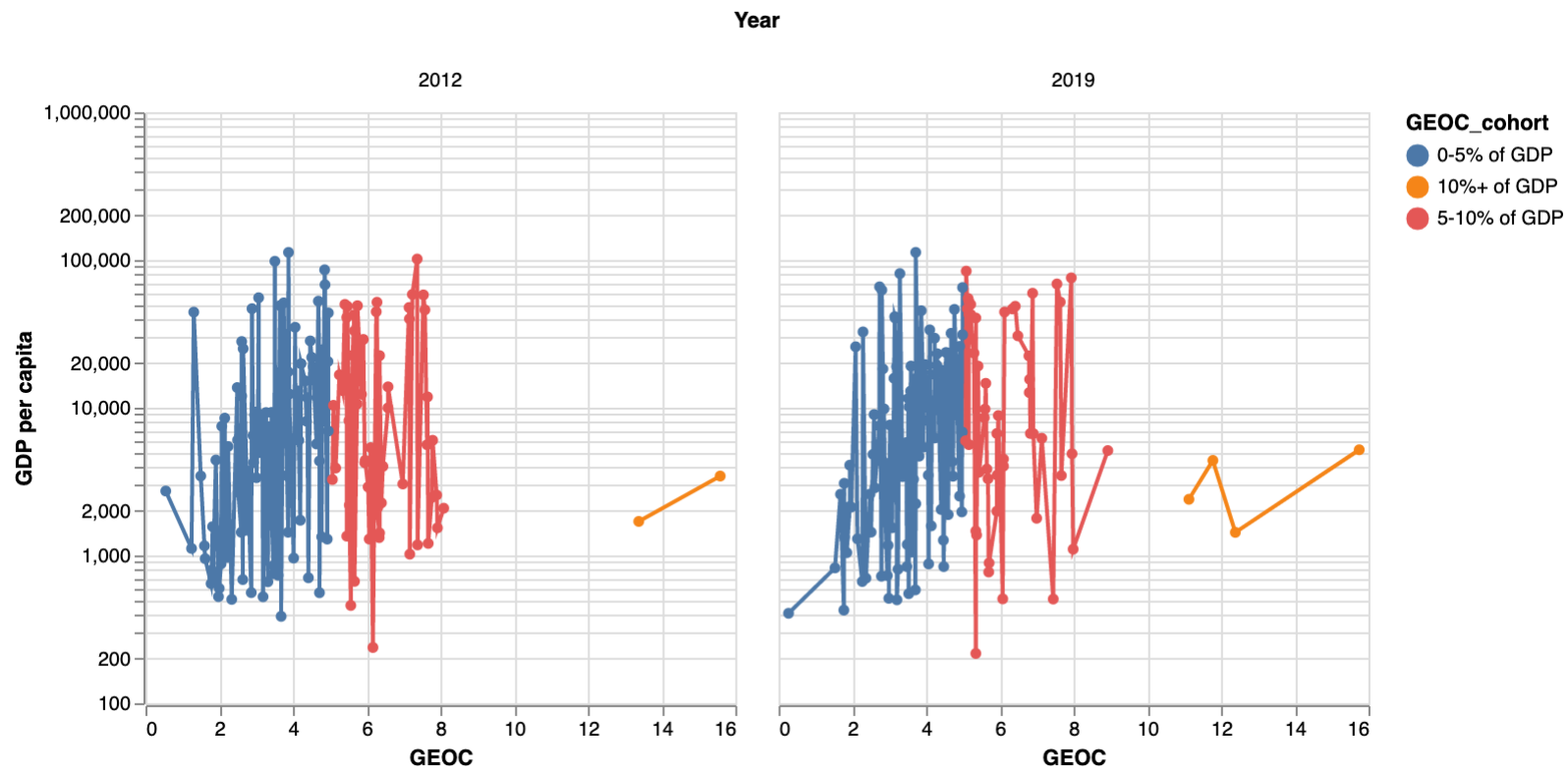
Multiple Linear Regression Model: Predicting GDP per capita

Below is the coefficient table for our MLR model; GDP per capita is log transformed for normalization. The exponentiated estimate column represents the factor by which the estimated GDP per capita (not on the log scale) differs for each variable.

	estimate	standard error	exponentiated estimate
intercept	8.759434	0.134841	6370.503528
Year_2013	0.056013	0.159407	1.057611
Year_2014	0.053842	0.158163	1.055318
Year_2015	-0.049758	0.156314	0.951459
Year_2016	0.003981	0.157032	1.003989
Year_2017	0.062788	0.156565	1.064801
Year_2018	0.104978	0.156972	1.110686
Year_2019	0.137125	0.157707	1.146971
Population_cohort_3million-300million	-0.372828	0.088920	0.688784
Population_cohort_300million+	0.333825	0.305067	1.396299
GEOC_cohort_5-10% of GDP	0.424021	0.084603	1.528094
GEOC_cohort_10%+ of GDP	-0.896429	0.289079	0.408024
error_variance	2.000616	NaN	7.393611

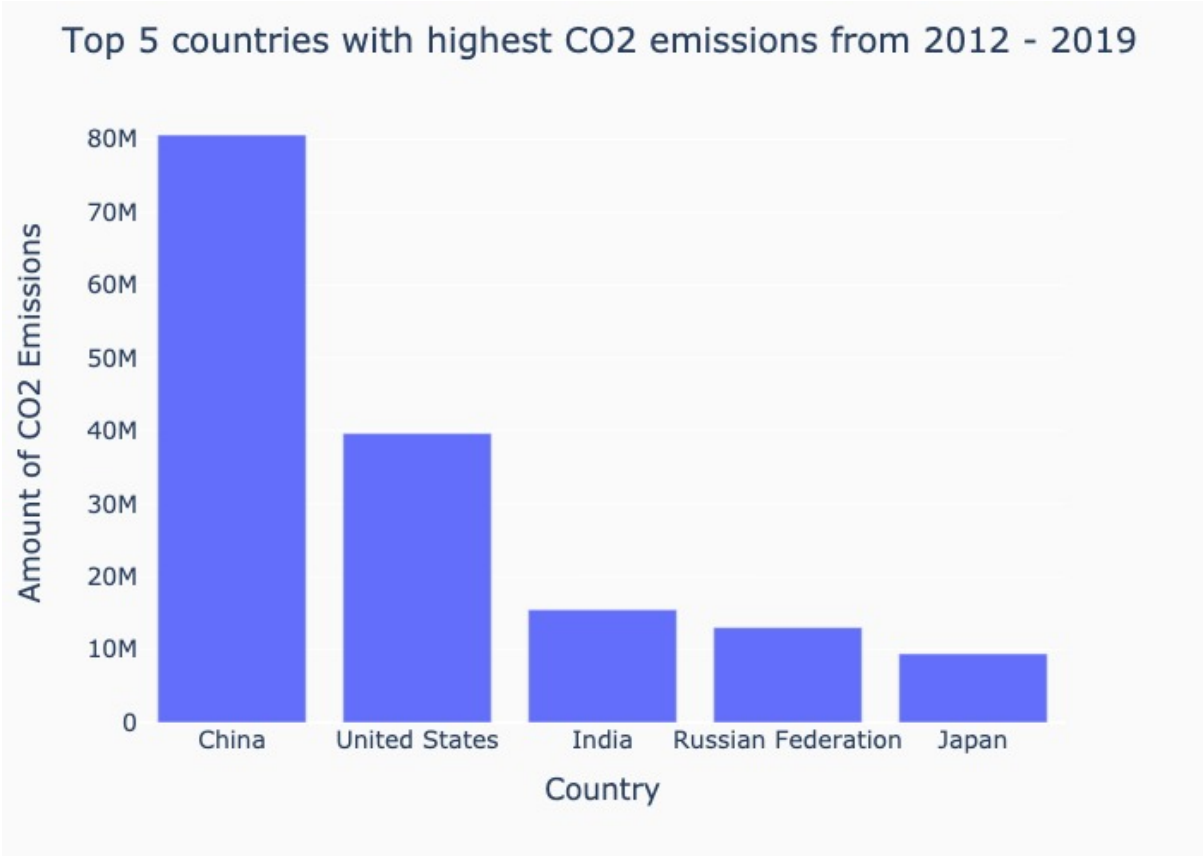
We see that each Year does not seem to indicate difference in GDP per capita values, but that Population and Government Expenditure on Education do indicate differences in GDP per capita.

The point-and-line graph below of GDP per capita vs Government Expenditure on Education corroborates what we deduced from our Multiple Linear Regression Model.

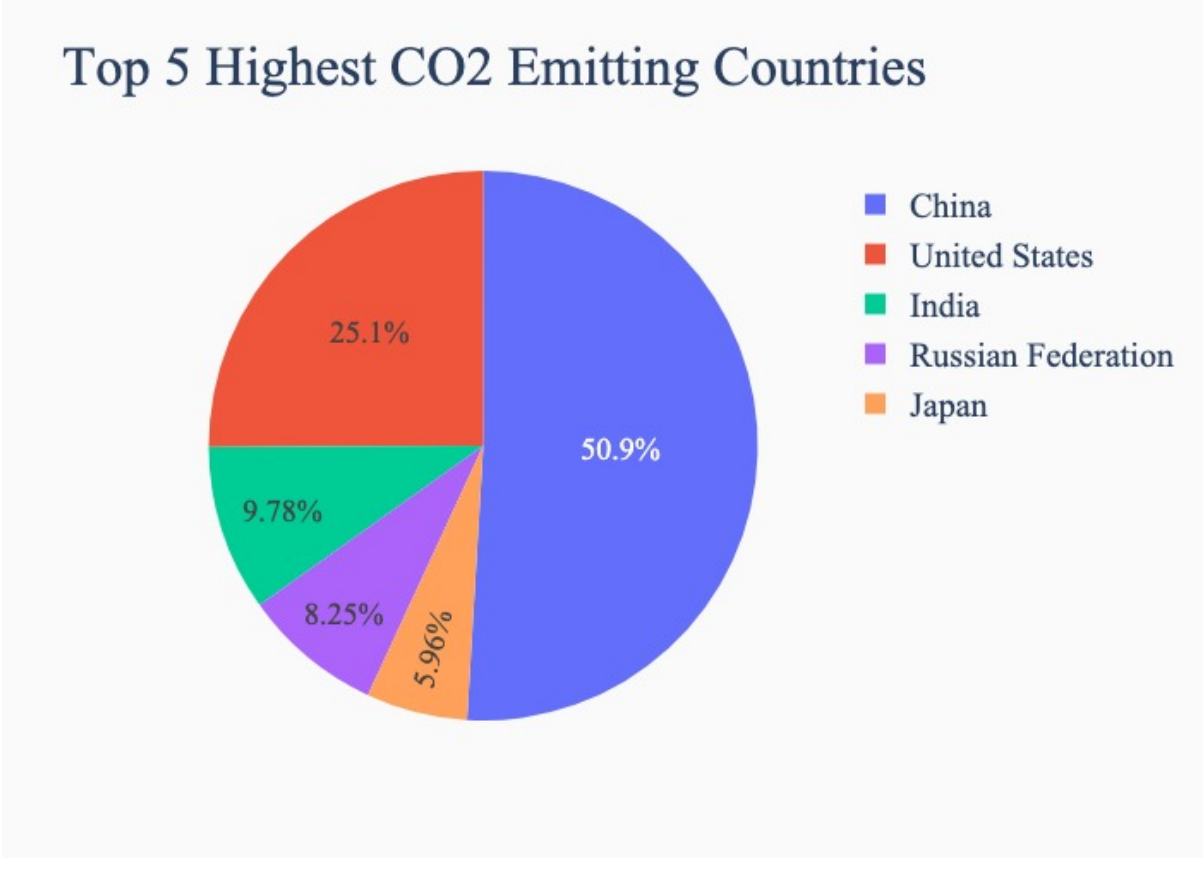


Top CO2 Emitting Countries

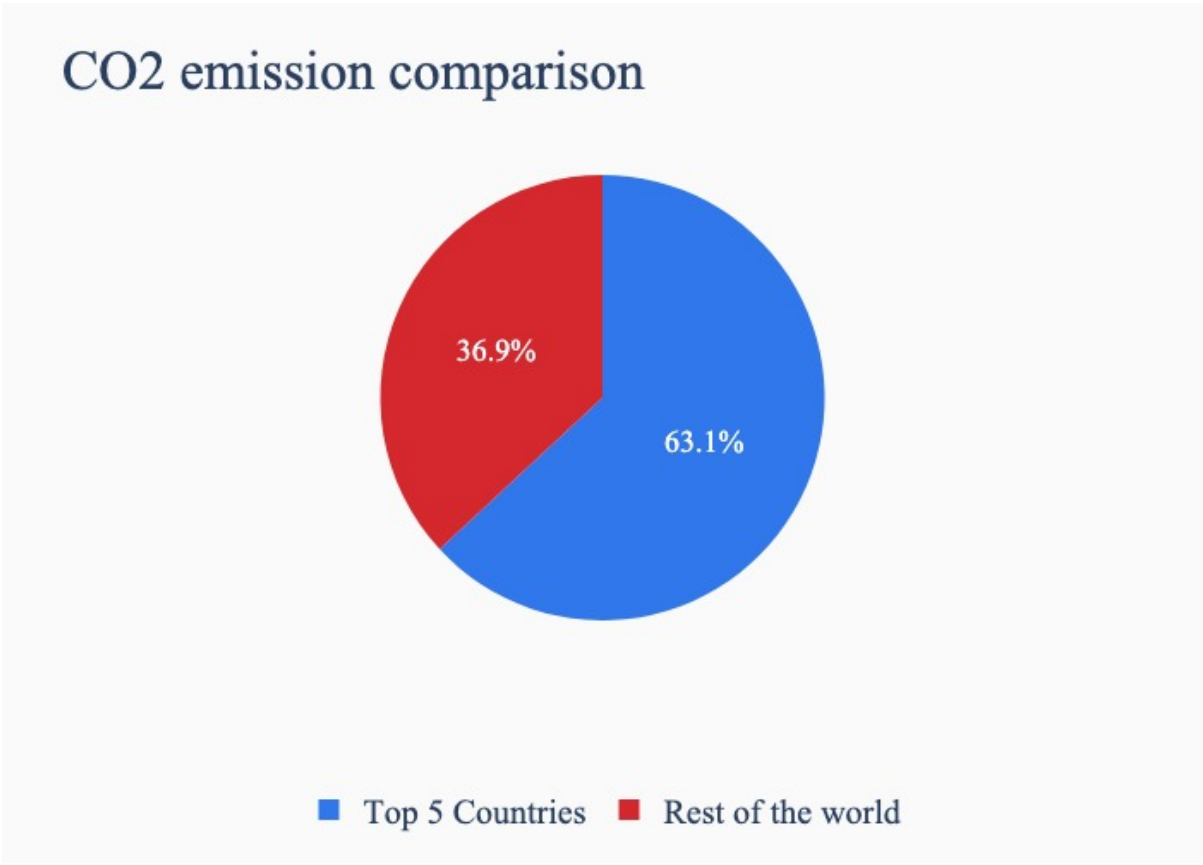
Looking at the countries that emit the most CO2 will give us some insight on who to focus on when building our model.



From this plot we can see that China has accounted for a lot more than the other 4 countries and may even have more emissions than all 4 other countries combined.



We can clearly see now that China does indeed account for more CO2 emissions than the other top 4 countries combined. Now we want to look at how much do these top 5 countries account for in the scope of the whole world.



From this pie chart, we can see that the top 5 countries account for almost two-thirds of the world's total CO2 emissions.

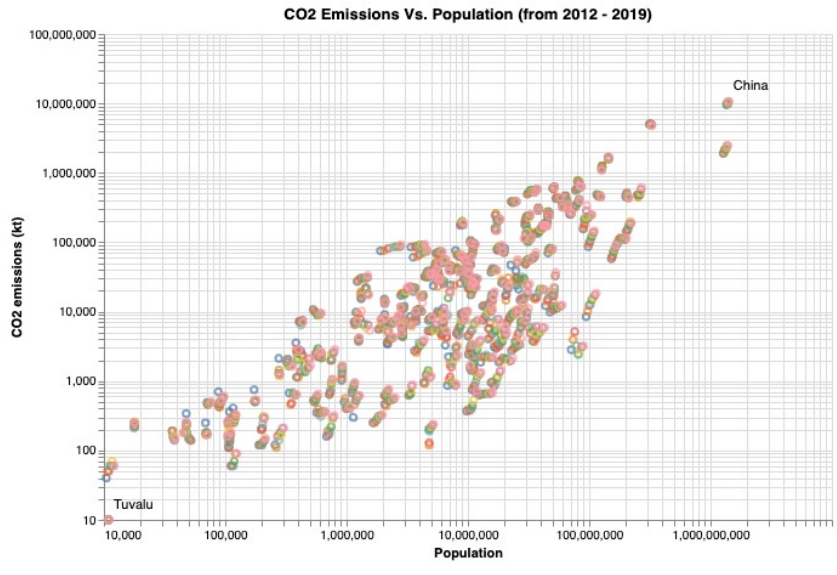
Now we will create a model predicting China's CO2 emissions to see how the country may develop in the future. CO2 emissions will be the response variable and CO2/Population, the ratio between population and CO2 emissions, will be the predictor. Below is the coefficient table we get from our Simple Linear Regression.

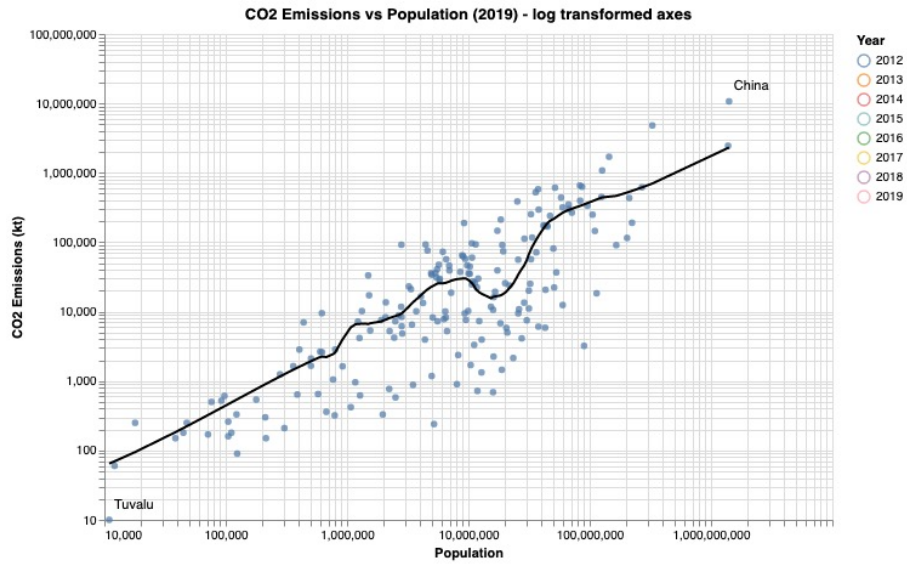
	coefficient estimate	standard error
intercept	-100370.19694648855	271.1329009548612
CO2_Population	-100370.19694648855	271.1329009548612

We also find that our model has a 92% R^2 score, meaning that 92% of the variation in the response variable can be explained by the singular predictor variable CO2_Population that we created. We also see the standard error is about 271 which means the average distance the observed values fall within the regression line is 271. Below we will visualize our SLR model.

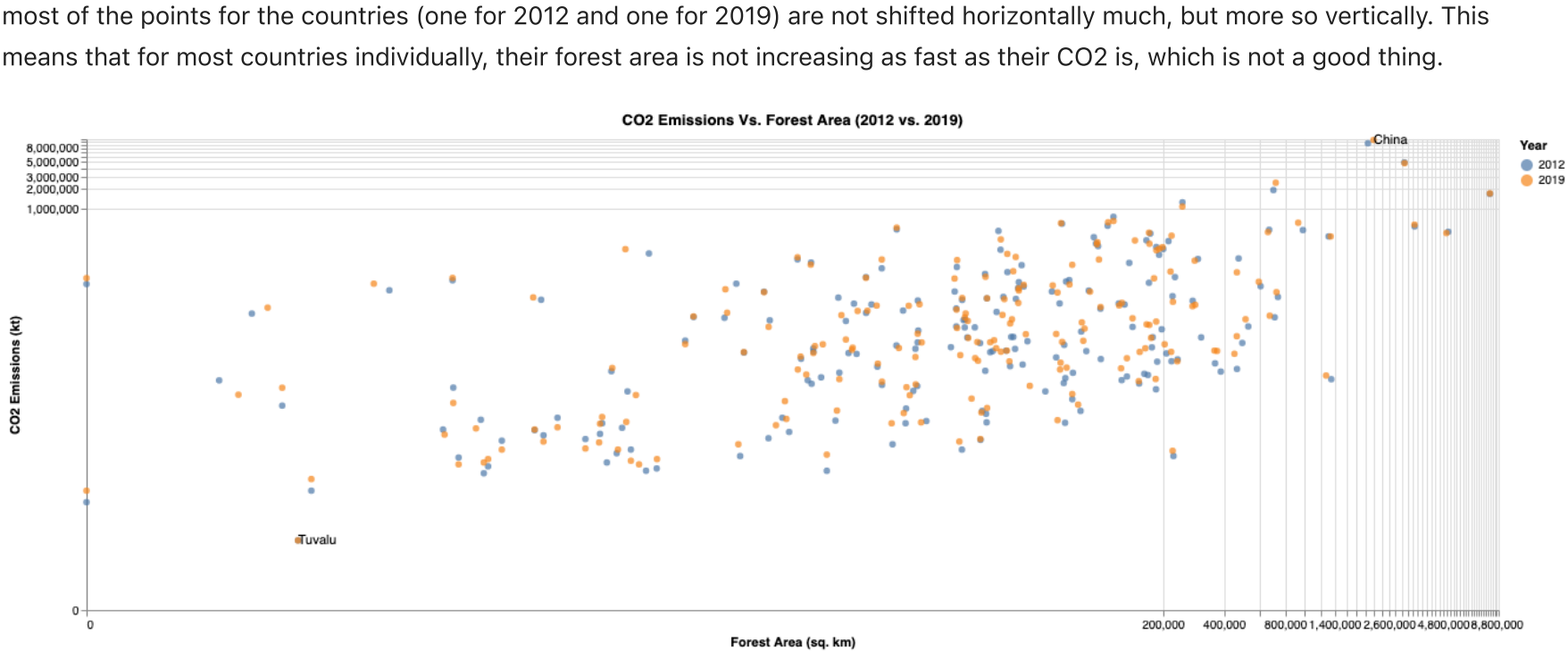
LOESS Line and Visualization: CO2 Emissions, Forest Area, and Population

In the first two graphs that are presented, (population and CO2 over time) we observe a positive, linear trend which indicates that population does indeed lead to a larger amount of carbon dioxide emission. However, over the years presented here, the emissions have been about the same relative to the country's population. Since the emissions have not changed drastically since 2012, the subsequent graph includes only 2019, the most recent year with the most amount of data, along with a LOESS trend line fit to the data points.





In the chart below, we observe something similar to the last two graphs which is quite interesting. There is a positive correlation between CO2 and forest area, meaning when a country has a large area of forests, they also have a large output of CO2. Perhaps forest area as a whole is increasing, but the level of emissions from the biggest countries is just too overpowering. It is important to note that most of the points for the countries (one for 2012 and one for 2019) are not shifted horizontally much, but more so vertically. This means that for most countries individually, their forest area is not increasing as fast as their CO2 is, which is not a good thing.



Discussion

The coefficients for Education Expenditure and Population from our MLR model strongly suggest an influence on GDP per capita. We see that countries that spend within 5-10% of their GDP on Education are predicted to have a slightly higher GDP per capita value, while countries that spend within 10-15% of their GDP on Education are predicted to have a slightly lower GDP per capita value. This may be due to the fact that spending 10-15% of one's GDP per capita value is a fairly large amount and education is an investment that takes years to influence the economy. We also see that countries with larger populations tend to have a larger GDP than smaller countries; this may be because larger populations lead to larger amounts of labor and production. The coefficients' high standard errors may be the product of an imbalance in data from group to group; for example, there are are only 3 countries with a population size over 300 million in our dataset and only 0.02% of our observations indicate a country spent more than 10% of their GDP value on education.

In the next section of our analysis, looking at CO2 emissions, we can see that our new variable CO2_Population gives us a good way for predicting CO2 emission values. This shows that the relationship between the CO2 emissions and how populated a country is has a strong relationship to one another.

The graphs in the final section of our results show an overall increasing level of CO2 emissions in part due to increasing populations, which makes logical sense. Forest area staying around the same level for each country is a bit surprsing however, we would have expected more countries to become more conservative and eco-conscious between 2012 and 2019. This would be something to that we would like to further explore. Perhaps in parts of the world forest area is increasing but it is just not enough to offset the amount of carbon that is being released. To further refine this aim of answering the question, we would go back and take more data from a wide time frame, which would give us a better sense of how the environment has been changing.

In conclusion, we discovered that the most cost-efficient and economically beneficial percentage of GDP to spend on education is between 5-10% of a country's GDP per capita. We also found that larger forest areas do not necessarily indicate lower CO2 emissions probably because factory pollution exponentially increases and forests take decades to grow. This is why it is imperative to limit the amount of harmful pollution industrialized countries produce. Reforestation efforts are important for long term global benefit, but it will not immediately reverse the effects of increased CO2 emissions. The timeframe to act is shortening, and it is imperative we come to an agreement as a world to preserve the health of the Earth.