PSTAT 100 - Final Project

Jacob Hornby & Sofia Ward

2025-03-17

Table of contents

Introduction

Greenhouse gas (GHG) emissions are a critical driver of climate change, with countries around the world grappling with strategies to mitigate their environmental impact while balancing economic growth and increasing population demands.

Understanding the relationship between national **GHG emissions** and socioeconomic factors, such as **Gross Domestic Product (GDP)** and **population size**, is crucial for designing effective climate policies.

GDP serves as a key indicator of a country's economic activity, reflecting industrial output, consumption, and energy use, all of which contribute to emissions. Higher GDP levels often correlate with increased energy consumption, but they may also enable investment in cleaner technologies. Similarly, population size plays a significant role in shaping emissions trends, as larger populations demand more energy, transportation, and industrial output. However, variations in policy, energy efficiency, and economic structure can lead to differing levels of emissions intensity among countries with similar socioeconomic profiles.

Question of Interest

What are the top driving factors and key contributing countries responsible for greenhouse gas (GHG) emissions?

Historically, developed nations have undergone industrial revolutions that fueled economic expansion but also contributed significantly to environmental degradation. This raises important questions: Are developed nations still the primary contributors to emissions, or have emerging economies become dominant players? Are there hidden relationships between socioeconomics and emissions that go beyond simple economic growth?

By exploring these interactions, we seek to understand how economic development and demographic factors influence environmental sustainability.

Data Description

Country GHG Historical Emissions (1990 - 2019)

This data set includes 195 **worldwide country emissions** of all greenhouse gases across sectors; namely energy, industrial processes, agriculture, waste, and land use: change and forestry.

- **30 Years of Data**
- **Observational Units**: Annual GHG Emissions ($MtCO_{2e}$) per Country or Region
- Due to the different capacity and reporting requirements, **not all countries have a complete inventory**.

Variable Name	Variable Type	Description
`Country`	Character	Recorded Country Name
`Data source`	Character	Emissions Data Source (`Climate Watch`)
`Sector`	Character	Emission Sector Classification (`Total including LUCF`, LUCF: Land Use Change and Forestry)
`Gas`	Character	Measured Type of Gas (`All GHG`)
`Unit`	Character	Emission Measurement Units ($MtCO_{2e}$)
`2019, ..., 1990`	Double	Recorded Emissions for Given Year

Sources and Links:

- Climate Watch: Open Data
  - GHG Emissions
  - Methodology

GDP Per Capita / Population (1960 - 2019)

Annual GDP per capita and total population for 264 **countries** at approximately five-year intervals from **1960 to 2019**.

- **60 Years of Data**
- **Observational Units**: Annual GDP per **Country**.
- ***Limitations and Exceptions***:
  - "Current population estimates for developing countries lack reliable recent census data. Population estimates are from demographic modeling and so are susceptible to biases and errors from shortcomings in both the model and the data. Because future trends cannot be known with certainty, population projections have a wide range of uncertainty."
  - "GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. **It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.**"

Variable Name	Data Set	Variable Type	Unit of Measure	Description
`Country Name`	Both	Character	N/A	Recorded Country Name
`Country Code`	Both	Character	ISO 3166-1 alpha-3 Code	Standardized Country Code
`Indicator Name`	GDP per Capita	Character	N/A	Data Value Units (`GDP per capita (current US$)`)
`Indicator Code`	GDP per Capita	Character	N/A	Corresponding Indicative Code (`NY.GDP.PCAP.CD`)
`1960, ..., 2019`	GDP per Capita	Double	US Dollars	Weighted GDP Average of Country
`1960, ..., 2019`	Population	Double	N/A	Total Country Population

Sources and Links:

- World Bank: Open Data
  - Population
  - GDP Per Capita

Data Cleaning

Tidying / Missing Values

To ensure proper analysis and readability, the **data sets** should be **tidied** by:

- **Reshaping them into a grouped column format** (Moving years from columns to rows).
- **Ensuring years are stored as numeric variables**.
- **Handling missing values** appropriately.

GHG Historical Emissions needed format reshaping, with years being one column for easy access.

Given the substantial amount of missing values in the GDP per Capita and Population data sets, we aim to maintain data quality while minimizing information loss. To achieve this, we used mean imputation, filling in missing entries with the average GDP per capita specific to each country. The Population data set needed similar formatting and no missing values in order to combine data frames.

Merging

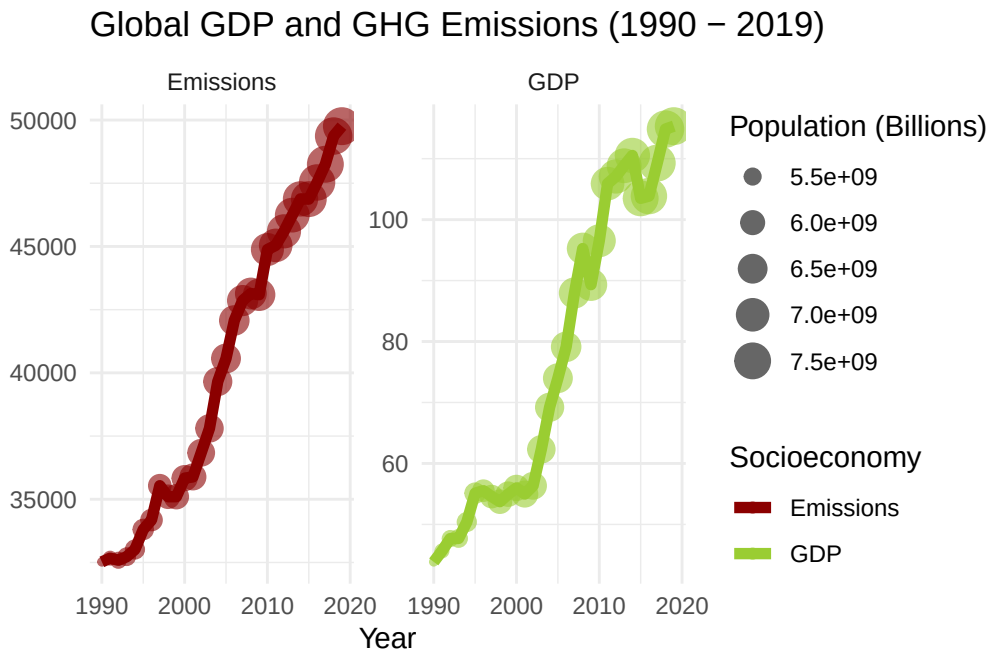
All three data sets were combined into one informative data set for analysis, void of missing values. The head and tail of the prepared data set appears as follows:

Country	Year	Emissions	Country Code	GDP	Population
World	1999	35101.90	WLD	54.96173	6034491778
World	1998	35099.21	WLD	53.72709	5954005533
World	1997	35537.18	WLD	54.57085	5872254371
World	1996	34179.33	WLD	55.53384	5789623830
World	1995	33805.61	WLD	55.12439	5706689093

Country	Year	Emissions	Country Code	GDP	Population
Fiji	2004	0.07	FJI	34.34932	817860
Fiji	2003	-0.38	FJI	29.37892	816076
Fiji	2002	-0.50	FJI	23.60258	815257
Fiji	2001	-0.33	FJI	21.39626	813925
Fiji	2000	-0.59	FJI	21.76569	811006

Exploratory Data Analysis

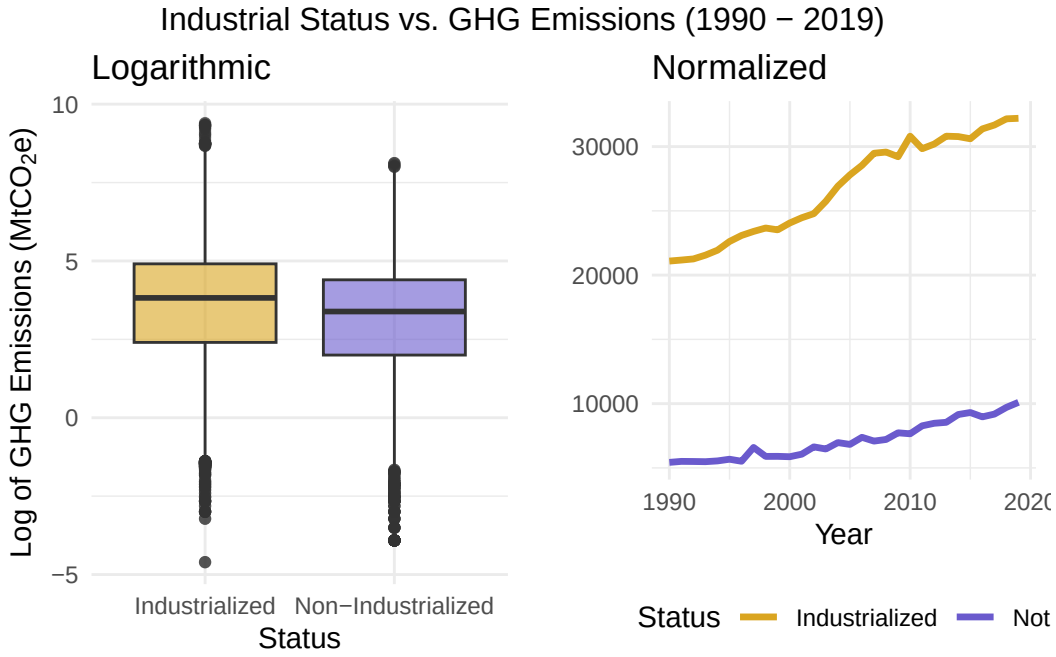
Before we dive into individual contributions, let's take a look at global relationships between 1990 and 2019.



The figure above demonstrates side-by-side plots of global GHG emissions and GDP per capita, reflecting a similar exponential increase from 2000 to 2010.

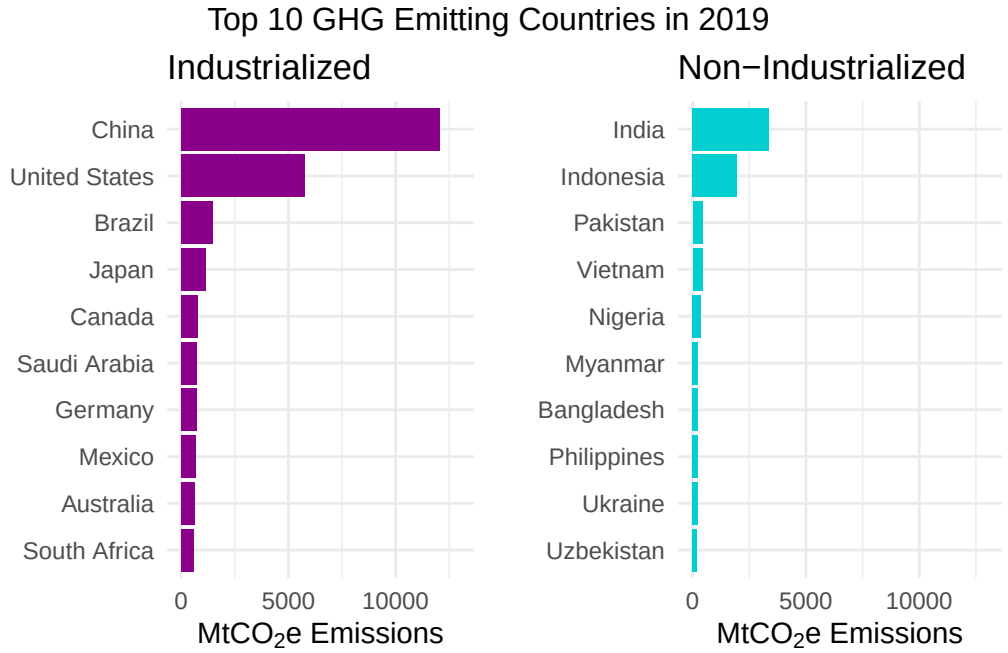
From 1995 to 2005, a plateau in GDP proceeds an apparent spike in GHG levels, possibly due to many growing countries' heavy coal usage to power their industrial expansion. The World Trade Organization, established in 1995, aids with global trade, industrial production and economic growth, especially in developing countries. With trade networks established and populations growing by the 2000's, GHG byproducts continue to steadily rise.

Let's take a closer look at the industrialized countries, specifically the early, late and ongoing nations. The classifier we will be using to determine a country's industrialization status is based on whether their normalized GDP per capita is greater than or equal to 45:



These plots demonstrate the aggregated annual GHGs from selected industrialized and non-industrialized countries from 1990 to 2019. Industrialized countries contribute more GHG emissions on average than non-industrialized countries. This could be due to an industrialized country's economy being more capital-intensive and resource-heavy than those in developing nations.

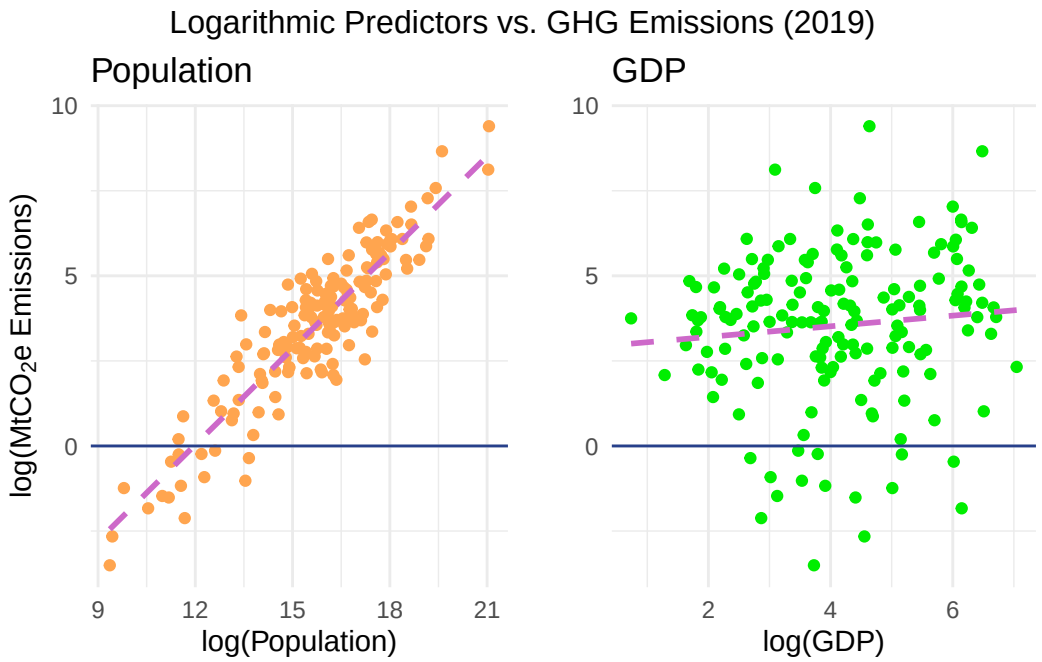
Within these industrialized and non-industrialized countries, which ones have emitted the largest amount of greenhouse gases? Looking at the data from the most recent year of 2019, we obtain the following top ten each:



Based on the above bar plots, we can see that China has by and far the largest amount of GHG emissions for any country, with approximately 1.205541×10^4; the United States' emission value of 5771 isn't even a close second. For non-industrialized countries, India's 3363.6 and Indonesia's 1959.71 top their pack pretty steadily; both are larger than the remaining top eight industrialized country emissions, which in turn are higher than the other top eight non-industrialized country emissions.

It's important to note the classification of "Industrialized" vs. "Non-Industrialized" countries: the largest normalized GDP per year as being higher than 45. As such, India and Indonesia are classified as "Non-Industrialized", when in reality they are technically considered to be newly-industrialized countries; such is also the case for Brazil (2019 GHG Emissions = 1451.63), Mexico (2019 GHG Emissions = 670.84), and South Africa (2019 GHG Emissions = 562.19) among others.

The previous plots demonstrate that the normalized GDP per capita appears to have a generally positive relationship with a country's GHG emissions per year, but what about when we throw the country's population into the mix? Once again, we will be analyzing with the year 2019, wherein most countries would have had their highest population yet:



Logarithmic transformations were used to ensure relationships are detailed in a concise manner void of outliers.

The left plot demonstrates the relationship between the logarithmic population and GHG emissions in 2019. After transformation, values cluster into a straight, positive form (represented with the magenta-colored, dotted regression line), indicating that the variables have a positive relationship with one another: where there's more people, there's generally more emissions. The intercept of the regression starts at a logarithmic $MtCO_{2e}$ value of less than 0 (the x-axis is represented by the blue-colored line), indicating that countries with low populations have reduced (or perhaps even little to no) GHG emissions than those with larger populations.

By contrast, the right plot demonstrates the relationship between the logarithmic normalized GDP per capita and GHG emissions in 2019. Compared to the population, transformed values are more scattered and inconsistent in relation; however, the regression line showcases a positive relationship between our variables, just like our initial observations when comparing industrialized countries. On average, the greater GDP a country has, the more greenhouse gases the country ends up emitting. Although the regression's intercept is higher than with the population, the coefficient and slope are significantly weaker, meaning that the population has a larger influence on the emission of greenhouse gases.

Given everything we have analyzed up to this point, we can finally determine the relationship between each country, its maximum gross domestic product per capita, and population along with the total number of emissions from 1990 to 2019.

Our top ten and bottom five countries with the largest amount of greenhouse gas emissions are as follows:

Country	Emissions	Population	Max GDP	Industrialized?
China	212226.96	1262645000	103.16630	Yes
United States	177781.18	282162411	653.97518	Yes
India	61248.93	1056575549	21.99599	No
Brazil	51449.94	174790340	133.45612	Yes
Indonesia	43353.95	211513823	42.35569	No
Japan	35546.13	126843000	487.03477	Yes
Germany	27149.61	82211508	480.59993	Yes
Canada	24141.24	30685730	527.78390	Yes
United Kingdom	18607.79	58892514	506.66827	Yes
Australia	17897.52	19153000	682.50107	Yes

Country	Emissions	Population	Max GDP	Industrialized?
Nauru	2.42	10337	98.62389	Yes
Kiribati	2.16	84396	18.88505	No
Tuvalu	0.63	9394	41.59030	No
Fiji	-12.94	811006	64.17488	Yes
Bhutan	-118.05	591021	34.16176	No

Even when all years are put into play, it's clear that the country's population plays a stronger part in greenhouse gas emissions than the maximum gross domestic product per capita; the top six all have populations greater than $100,000,000$ people, while the bottom five all have populations less than $1,000,000$ people! By comparison, lower ranking countries such as Fiji and Nauru have larger gross domestic products per capita than India or Indonesia, and otherwise, Brazil and China are the only two countries in the top ten with a maximum gross domestic product per capita value of less than 200.

Summary

This project, utilizing several Exploratory Data Analysis techniques, aimed to analyze the relationship between greenhouse gas emissions and socioeconomic issues among several countries. Considering greenhouse gases are a leading contributor to climate change, with elements such as a steadily increasing world population and industrialized economic activity not helping the case, the natural first step to combating these issues is to recognize the significance of these relationships to make plans going forward.

Based on our exploratory data analysis, the socioeconomic factor with the strongest relationship with national greenhouse gas emissions is the country's population. Many of the globe's most populous countries, including China and India with over $1,000,000,000$ each by 2019, have exhibited significantly larger greenhouse gas emissions over the course of a 30 period, with China topping the charts at over $200,000$ emissions of $MtCO_{2e}$ and the United States falling close behind at over $175,000$ emissions. In contrast, countries with significantly lower populations, such as many Oceanic countries of less than $1,000,000$ people each, have emitted little to no $MtCO_{2e}$ over the course of the 1990's, 2000's and 2010's.

However, the gross domestic product per capita of a country did not render as insignificant. Many industrialized countries, defined as having a normalized GDP of over 45, maintained consistently stronger emission averages than the majority of non-industrialized countries, once more a generally positive relationship. However, certain countries classified as "Non-Industrialized" due to their low GDPs, particularly India and Indonesia, fared stronger greenhouse gas emissions than even most other "Industrialized" countries, most likely due to their larger population and status as a newly-industrialized country.

Despite the project achieving its goals of determining the primary causes of greenhouse gas emissions and which countries exhibit the most of them, it is not without its limitations. The most significant of them would have to be the classification of country industrialization, a factor in which requirements have no agreed consensus. The choice to use 45 as the threshold for a normalized GDP value came from the data itself than anything else, leading to a number of countries classified as either "Industrialized" or even "Non-Industrialized" to be either incorrect or really be newly-industrialized countries instead. On top of that, many correlations and conclusions are made based on plots and raw data on more recent years: stronger usage of machine learning models and larger data could contribute to a more accurate portrayal of predictors with appropriate statistical significance and interpretability.

While it can be difficult to cease climate change due to a number of factors out of our control, the correlation between population and greenhouse gas emissions can indeed be addressed. However, with the very few amount of places to explore left on Earth, one may need to consider… would it be smarter to advise space travel at this point, reduce the amount of pregnancies in larger communities, or simply distribute the population more equally? And would any of those methods technically be applicable considering the general hesitance or lack of substantial wealth from the general population?

References

- **Climate Watch**: https://www.climatewatchdata.org/
- **World Bank Group: Open Data**: https://data.worldbank.org/
- **World Resources Institute**: https://www.wri.org/
- **World Trade Organization**: https://www.wto.org/