Exploring the Evolution of Energy and Emissions

By: John Albright, Clay Grant, Naman Saboo, Christopher Whitney

Defining the Problem:

Our big picture problem is that one of the biggest issues in society today is how climate change has an impact on the future of the world. We want to shed light on what the future may hold and how different aspects of climate change relate to one another. We seek to find answers to the following problems: At what points in time do we see drastic shifts in metrics relating to climate change and what real-world processes were a likely cause? How can we use data we have to make assumptions on the future?

Data Collection:

The "Share of the population with access to clean fuels for cooking" dataset is from the World Health Organization and is processed by Our World In Data. The “Population” dataset was also processed by Our World in Data but came from the United Nations Department of Economic and Social Affairs Population Division. The "Per capita CO2 emissions" dataset describes the total tonnes of CO2 emissions divided by population and comes from the World Bank and Gapminder, who used various data sources to create more clean data. Each population refers to each country's population. The dataset ranges from 1800 to 2022.

In the "Share of the population with access to clean fuels for cooking", there are 6336 samples and 4 features. In the “Per Capita CO2 emissions” data there are 194 samples and 224 features. In the “Population” data, there are 18944 samples and 4 features.

The dataset on the "Share of the population with access to clean fuels for cooking" includes several key features: the country name labeled as "Entity," the country’s 3-digit identification code labeled as "Code," the year of data collection labeled as "Year," and the percentage of the population with access to clean cooking fuels and technologies, labeled as "Residence area type: Total." In the "Per Capita CO2 Emissions" dataset, the features include the country name, labeled as "Country," and the annual CO2 emissions per capita in tonnes, with each year from 1800 to 2021 represented as a separate column. The population dataset includes the country name, labeled as "Name," the country’s identification code, labeled as "Code," the year the data was collected, labeled as "Year," and the estimated population, labeled as "Population - Sex: all - Age: all - Variant: estimates."

Potential sampling issues across these datasets could include missing data for certain years or countries, differences in data collection methods, and variations in the definitions of key terms such as "population" or "access to clean fuels." These inconsistencies may introduce bias or affect the reliability of analyses and predictions, making data cleaning and standardization essential for accurate results.

Data Preparation:

To align our datasets for analysis, we performed two inner merges based on the keys "Country" and "Year," ensuring consistency across all data frames. While this method helped unify the datasets, it may have excluded rows without matching values, effectively dropping data where key information was unavailable. To further standardize the data, we filtered all datasets to include only records within the timeframe of 1990 to 2021.

Additionally, we addressed data type inconsistencies. Some variables, such as CO2 emissions, were initially non-numeric, causing issues with functions requiring numeric inputs. To resolve this, we used the pd.to_numeric function with the errors='coerce' parameter, converting invalid entries into NaN values. This step enabled efficient data cleaning and handling.

To enrich the dataset, we incorporated additional variables, including the percentage of the population with access to clean cooking fuels and total population. This required renaming columns and adjusting data types to facilitate a seamless merge. The final dataset provided a comprehensive view of key metrics for each country and year, forming a solid foundation for our analysis.

Data exploration:

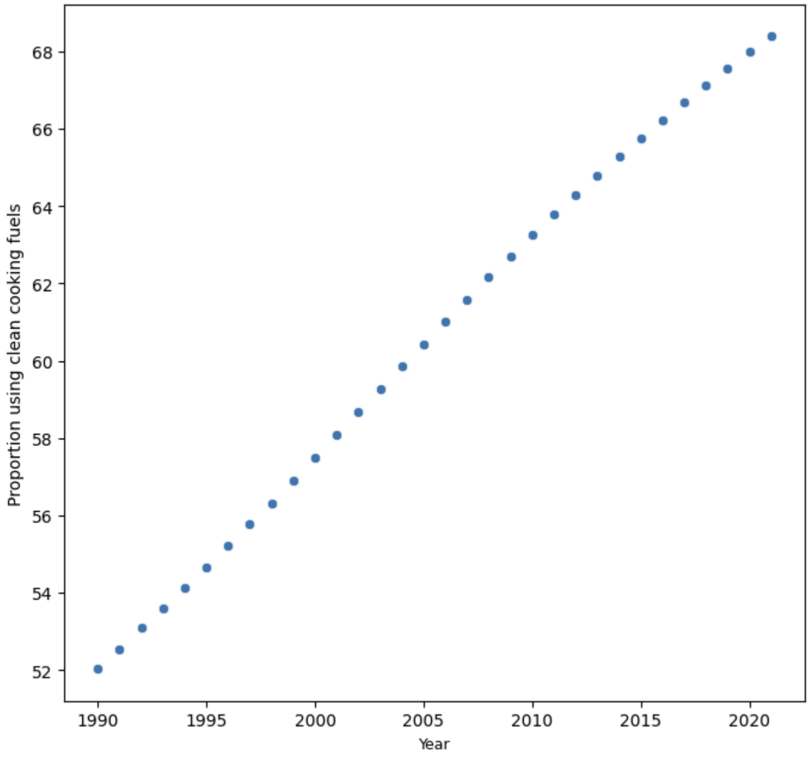

Figure 1

We observe a generally linear increasing trend with this primary reliance increasing each year in our data. We will try to model this.

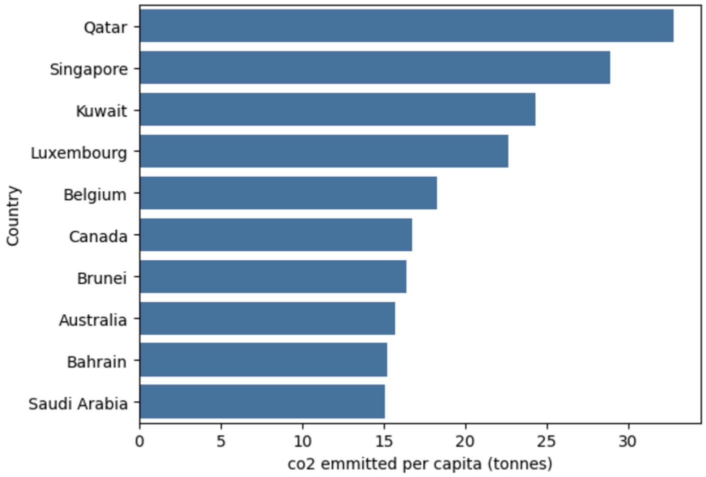

Figure 2

We calculated the top 5 countries emitting the most tonnes of CO2 per capita yearly from 1990 through 2021. We observed that many nations on this list are big in the oil industry with the top 5 of CO2 emissions per capita (tonnes) being Belgium, Kuwait, Luxembourg, Qatar, and Singapore. We then created the following line plot:

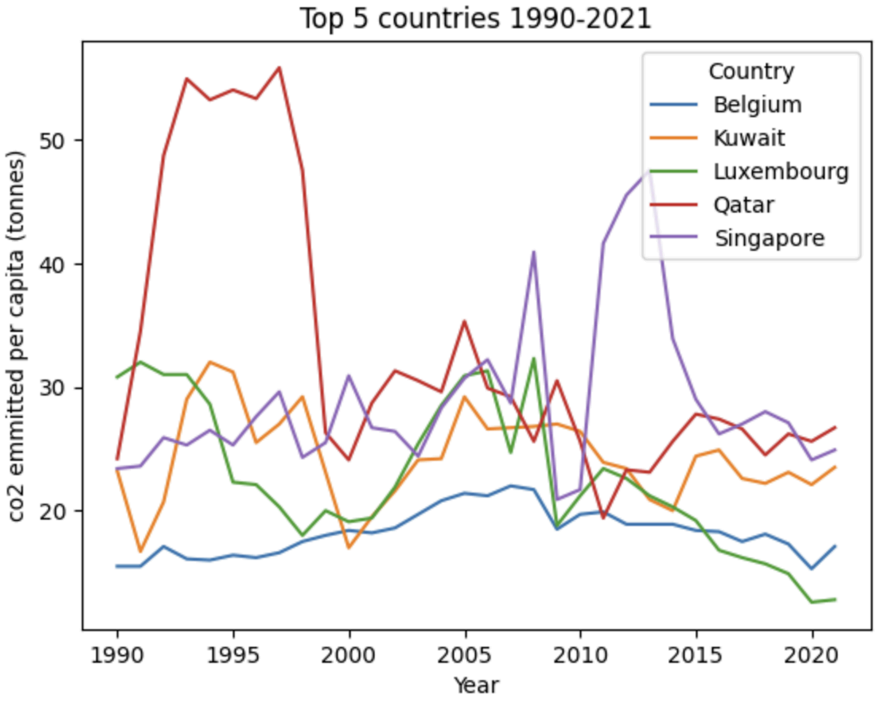

Figure 3

We then created a function that returns the largest absolute change in CO2 emissions for each of the major countries in global emissions (per capita) in adjacent years to determine what significance the peaks have. We derived the following information from the largest absolute differences:

Kuwait's CO2 emissions increased significantly between 1992 and 1993 due to the aftermath of the Gulf War which occurred in 1991. The war led to environmental damage, including the deliberate setting of Kuwait’s oil fields on fire, which burned for months. This caused massive emissions of CO2 as well as the loss of millions of barrels of oil each day. While infrastructure and industrial activities were disrupted during the war, the rebuilding efforts and the return to normal oil production led to a sharp return to normalcy in emissions beginning in early 1993.

From 1998 to 1999, Qatar's CO2 emissions also decreased significantly. This could have been due to a myriad of factors such as an economic downturn in gas and oil production, or changes in the Qatari government. However, while the function illustrates this major change, there are no prominent historical events that directly point to this change. This may indicate a proposal for further research or investigation into the underlying causes, such as potential shifts in energy consumption, production strategies, or other underlying factors.

In 2008-2009, Belgium, Singapore, and Luxembourg saw significant changes in CO2 emissions due to a mix of economic factors, primarily due to the global recession. For example, Belgium experienced a reduction due to the global financial crisis, which decreased industrial output and energy demand leading to lower emissions. Singapore's emissions also dropped, due to fluctuations in global oil prices and a decrease in energy consumption during the global recession of 2008. In Luxembourg, a reduction in CO2 emissions was largely attributed to a decrease in natural gas use as part of reduced energy consumption in response to the downturn.

The second function (parameter_estimates) helps analyze the relationship between two sets of data by calculating several important values. It starts by ensuring the data is in numerical form and then looks at how closely the two sets of data are related by calculating the correlation. It also finds the equation of the best-fit line (the slope and intercept) and uses that line to predict values. Additionally, the function measures how much the actual values differ from the predicted ones using the root means squared error (RMSE), and it calculates the R-squared value to show how well the data fits the line. By providing these statistics, the function gives a clear picture of the relationship between the two variables and how well one can predict the other.

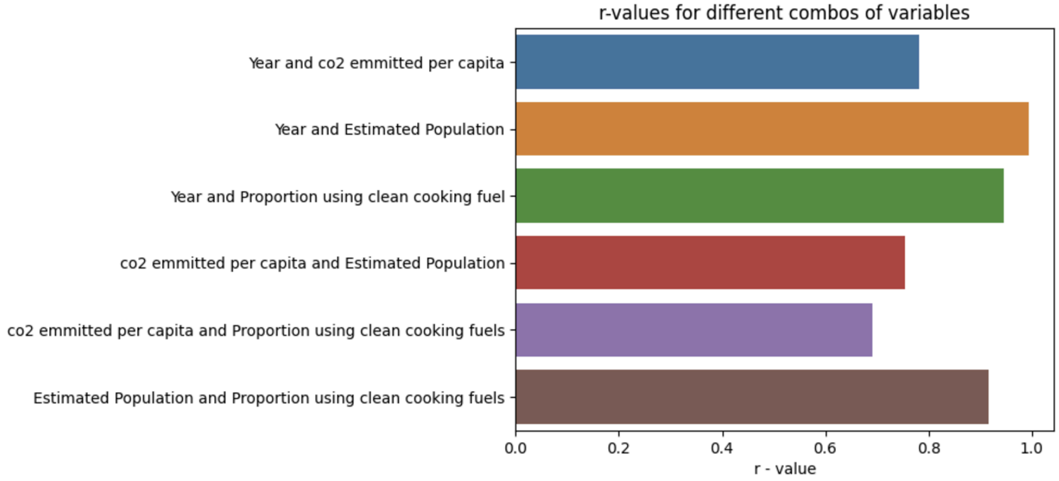

Figure 4

Model Building:

Upon seeing the linear-looking relationship between the proportion of people using clean cooking fuel and year, we decided to use simple linear regression to try to model this relationship efficiently and accurately.The value we got for the slope is 0.548831.

Specifically in our model we only used year as a predictor of the proportion of people using clean cooking fuel. We, however, did have CO2 emissions per capita, Country, and estimated population in which varying models could be built to show relationships.

We created a function (parameter_estimates) function that helps analyze the relationship between two variables and their respective least squares regression lines by calculating several important values. It starts by ensuring the data is in numerical form and then looks at how closely the two sets of data are related by calculating the correlation. The function also finds the equation of the best-fit line (the slope and intercept) and uses that line to predict values. Additionally, the function measures how much the actual values differ from the predicted ones using the root means squared error (RMSE), and it calculates the R-squared value to show how well the data fits the line. By providing these statistics, the function gives a clear picture of the relationship between the two variables and how well one can predict the other.

Model Evaluation:

The question we investigated with this model was “How can we use data we have to make assumptions on the future?” Our model tells us that we have a linear relationship between the attributes of Year and Proportion using clean cooking fuels, which is also supported by Figure 4. We could have run this model across differing variables but instead, we decided to create a function that will do this for us using the mathematical formulas for linear regression.

We trained and tested to compute scores to assess the model. We used 80% of the data in the training set and 20% in the testing set and fit the model on the training set and tested it on the testing set. We ended up with a train score of 0.9989 and a test score of 0.9941 in our first split of data, meaning our model is very effective in predicting the proportion of clean cooking fuels. Running the cells many times led to similar results.

Model Deployment:

Our model demonstrates the potential to predict future CO2 emissions per capita by leveraging correlations between key contributing factors observed in recent years. These predictions can support scientists, policymakers, and stakeholders in making informed decisions to combat climate change more effectively.

While the linear regression model proved highly effective within the scope of our analysis, it has inherent limitations. Linear models are well-suited for short- to medium-term predictions; however, as they do not account for saturation or non-linear trends, they may fail to capture long-term behaviors. For instance, predicting a proportion with a linear model could eventually exceed realistic bounds, prompting the need for alternative approaches, such as logistic regression or other non-linear models. Future work could explore these models to account for potential shifts in trends over time.

Historical data also highlights the impact of unpredictable events—such as economic recessions or geopolitical conflicts—on CO2 emissions. This variability underscores the need for caution when interpreting model results. As George Box famously said, “All models are wrong, but some are useful.” While our model provides valuable insights, it must be applied with an understanding of its assumptions and limitations.

Conclusion:

Through this analysis, we explored the intricate relationships between CO2 emissions, population dynamics, and access to clean cooking fuels over time. By integrating multiple datasets and applying rigorous data preparation and exploration techniques, we uncovered valuable insights into the factors influencing climate change metrics.

Our linear regression model demonstrated strong predictive performance, highlighting the potential to use historical data for future projections. However, we also recognized the inherent limitations of linear models, particularly for long-term predictions where nonlinear behaviors may emerge. This underscores the importance of adaptability in modeling approaches and the need for ongoing research to refine predictive frameworks as new data becomes available.

Ultimately, this project reinforces the critical role of data-driven approaches in understanding and addressing climate change. By identifying trends, investigating anomalies, and leveraging predictive models, we can better equip scientists, and policymakers to make decisions that foster sustainable development and mitigate the impacts of climate change. Moving forward, expanding this work to incorporate additional variables and advanced modeling techniques could further enhance our ability to tackle the challenges posed by climate change.