# Predicting Carbon Dioxide Emissions: A Comparison Across Agri-Food Systems and GDP Per Capita

Natalie Assaad, Julia Pfeiffer, Larissa Cybyk, Eva Butler, Ethan Banerjee, and Owen Shaffer

DS 3001: Foundations of Machine Learning

Professor Terence Johnson

December 14, 2024

# Abstract

Carbon dioxide (CO2) is the main greenhouse gas emitted by human activities such as the combustion of fossil fuels for transportation, electricity, and energy (Environmental Protection Agency). As the world's population continues to increase each year, these emissions will only increase and further damage the environment. We seek to build a model predicting CO2 emission levels over time by region and assess various economic and agricultural trends related to CO2 emissions. Our data is a combination of two datasets from Kaggle. The first dataset contains data on CO2 emission levels from a variety of sources in the agrifood industry, such as pesticide manufacturing, forest fires, and food processing (Bello, 2023). This dataset also includes data for over 200 regions from 1990-2020. Our second dataset contains GDP per capita data for various regions between 1960-2020 (Tas, 2022). By combining these datasets for our analysis, we will be able to predict total emission levels over time per region, assessing the accuracy of our model by comparing it to the original data, as well as study the relationship between emissions and GDP per capita and analyze which agrifood variable has the highest correlation with predicted emissions.

We produced a linear regression model to predict total emission values for each region over time, which demonstrated a reasonable level of accuracy. The training MSE (19.71) and RMSE (4.44) indicated moderate prediction error on the training set, while the test MSE (28.79) and RMSE (4.98) were slightly higher, pointing to minor overfitting but acceptable generalization to unseen data. The residuals were largely centered around zero, though some outliers and deviations may have impacted the model's predictive performance. Regions with the highest prediction error included India, China, and Indonesia – nations that tend to have higher total emissions overall. From our predictions, we found that low median GDP was associated

with extreme levels of emissions (both high and low), an observation reinforced by the Environmental Kuznets Curve (Ansari, 2023). However, this data included a significant number of outliers among the lowest emitters, suggesting that there are numerous low-emitting regions with high GDP. Additionally, urban population was most correlated with total predicted CO2 emissions, with higher levels of emissions suggesting a higher median urban population and a larger spread of values. Overall, a more comprehensive dataset with less missing values could improve the accuracy of this model. Nonetheless, this predictive model could be beneficial in highlighting areas for improvement in CO2 emissions as well as suggesting trends between economic and environmental factors. These insights can act as the foundation for future research and models, as well as guide strategy development for reducing CO2 emissions and curating a more sustainable agrifood industry.

## Introduction

Agriculture-related carbon dioxide (CO2) emissions account for approximately 60% of global emissions (Bello, 2023) and are expected to rise alongside population growth. Our project aims to explore this critical issue by analyzing emissions data from an Agri-Food dataset along with economic data from the World Bank, allowing us to model and predict total emission levels over time by region. These datasets offer considerable historical depth, with records dating back to at least 1990, and provide detailed data that will enhance our analyses. For instance, the Agri-Food dataset includes nearly every major source of agricultural CO2 emissions while the data from the World Bank provides a critical economic component to our analyses. By combining these datasets, our model will help us understand how emissions evolve over time and how they correlate with GDP per capita and agrifood factors.

This approach could reveal valuable insights into the relationship between the agrifood industry and economic development. An important implication of this analysis is that it will allow us to evaluate potential disparities between developed and developing nations, as the latter are often criticized for contributing disproportionately to global emissions. Many developing nations are becoming leading manufacturers due to lower costs of production, which often incentives cheaper, carbon-intensive production (Rooper, 2024). According to the Climate Leadership Council, developing countries may have contributed to nearly 95% of global emissions increases in the last ten years – a trend that is expected to continue over time. While developing countries are currently the largest emitters, they also suffer the most severe consequences of climate change, with income losses five times greater than those in developed nations. Nonetheless, developed countries have contributed the most historical CO2 emissions since 1850 (Beynon & Wickstead, 2024).

The combination of our two datasets will allow us to further study these economic and environmental disparities between developed and developing nations. By producing and analyzing our predictive model, we hope to guide discussion on reducing emissions while supporting sustainable economic growth and agrifood practices. Since income losses from climate change are five times greater in developing countries, our model can act as the foundation for producing strategies to address these economic vulnerabilities, especially in regions where emissions and agrifood factors are closely tied to population growth or resource limitations. We may also identify whether higher GDP areas, which often have the resources to adopt greener technologies, are truly using these advantages or whether their economic growth could be correlated with greater emissions outsourcing. Moreover, by identifying which agrifood factors are most correlated to total emissions in specific regions, our model could act as a tool to produce policies for minimizing carbon-intensive practices and supporting greener interventions in regions capable of these investments. Ultimately, this project not only provides a clearer picture of the global emissions landscape but also offers a basis for designing strategies that balance combating climate change and addressing food security for a growing population.

After producing our linear regression model, we discovered various important findings. Our model was largely accurate, revealing trends in predicted emissions that were all comparable to our original data. Countries such as China, Brazil, Indonesia, and the United States were noted as some of the highest emitters over time, with China displaying an increase in emissions over time. Additionally, our model showed the highest level of prediction error for countries that tended to be top emitters (e.g., India, China, Indonesia), perhaps due to their increased variation in emission levels over time. In our comparison of GDP and total predicted emissions, we found that both high and low emitters had a very low median GDP, with a significant number of outliers in low emitting areas. This suggests that there are a large number of low emitting countries with high GDP. Finally, urban population level was the most highly correlated factor with emission level, with the top emitters exhibiting higher median urban population levels and a wider spread of population values. This increased spread of values suggests that other factors may be contributing to higher levels of total emissions – an area for potential future research.

## Data

Our project combines two datasets from Kaggle. The first, sourced from the Food and Agriculture Organization (FAO) and the Intergovernmental Panel on Climate Change (IPCC), details CO2 emissions from the agrifood industry in 236 regions from 1990 to 2020. This dataset includes emissions data for various sources such as: savanna fires, forest fires, crop residues, rice cultivation, drained organic soils, pesticides manufacturing, food product transportation, net forest conversion, household food consumption, food retail, on-farm electricity, food packaging, agrifood systems waste disposal, food processing, manufacturing fertilizers, Industrial Processes and Product Use (IPPU), manure applied to soil, manure left on pastures, fires on organic soils, fires in humid tropical forests, and on-farm energy use. Other demographic variables in this dataset are rural, urban, and total (male/female) populations. Additionally, total emissions, land covered by forests, and average temperature is recorded for each region. The second dataset, provided by the World Bank, tracks GDP per capita for 266 regions between 1960 and 2020.

In combining these two datasets, we faced some difficulties. While we joined our data together based on the variables for geographical area, the two datasets had variation in this naming structure. For example, the agrifood dataset contains data from both “China” and “China, mainland” while the GDP dataset does not. The GDP dataset lists certain geographical areas that are not in the agrifood dataset, such as “Fragile and conflict affected situations” or “Least developed countries: UN classification.” Additionally, the agrifood dataset contains data from 1990-2020 while the GDP dataset begins in 1960. These variations required more effort in the data cleaning process when combining the datasets, and we eventually had to leave out certain geographical areas from the GDP dataset. Although the GDP dataset contains more geographical areas, it also contains a large amount of missing data. Because of this, we may not be able to find significant relationships between certain variables or make accurate predictions for geographical areas or years with too much missing data. To resolve this issue, if a region was missing more than five years of data, we dropped it from the set. Otherwise, we calculated the mean GDP for that geographical area and imputed it into the dataset. We also dropped unnecessary columns which helped to limit the amount of missing values.

After merging these two datasets and conducting an exploratory analysis, observations in our study became a region’s GDP per capita and its data on various agrifood measures for one specific year (ranging from 1990 to 2020). We conducted some exploratory data analysis, providing valuable insights and trends from the combined data. Figure 1 contains a line plot displaying total CO2 emissions from 1990 to 2020 for higher emitting countries including Australia, Brazil, Canada, China, Democratic Republic of the Congo, Germany, India, Indonesia, Japan, Mexico, Myanmar, Thailand, United States, and the Russian Federation. While China clearly became the highest CO2 emitter over time, other countries showed notable trends. For example, Indonesia experienced a significant spike in CO2 emissions around 1997, temporarily surpassing other nations in total emissions during that period. In contrast, Brazil, which had the highest emissions in 1990, saw a steep decline in emissions around 2010. Additionally, we observed a noticeable upward trend in total emissions over time across regions, highlighting the importance of studying these trends and pinpointing areas for improvement (Figure 2).

**Figure 1**

*Total Emissions in High Emitting Countries*

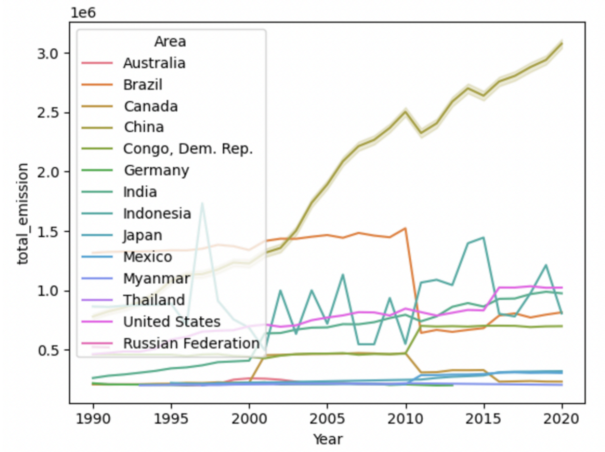
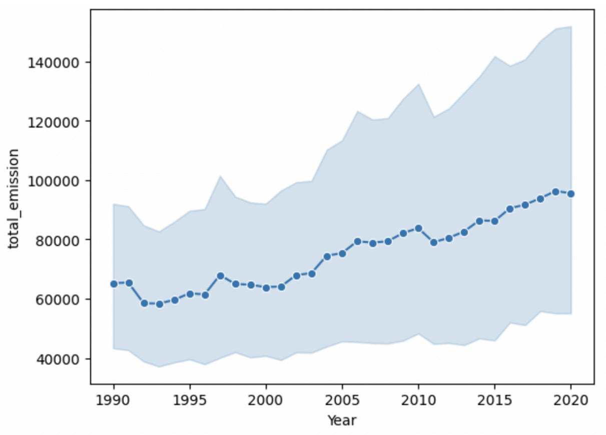


**Figure 2**

*Total Emissions Between 1990-2020*



**Methods**

After combining our data and exploring the variables, we continued to produce our prediction model. Since our goal is to predict emissions levels over time per region, we will be conducting a linear regression that will allow us to compare the predicted emission levels to the true values from the original dataset. For our analysis, we will construct a pipeline to ensure consistent data manipulation throughout the process. This pipeline will also enable us to scale numeric features, fill missing values, and encode categorical features. The full pipeline consisted of a numeric pipeline and a categorical encoder. The numeric pipeline used an imputer to fill missing values with the most frequent value in that column and then scaled all values using Scikit-learn's standard scaler. To encode the categorical features, we used the OneHotEncoder class. Once the data was split into training and test sets and transformed with the pipeline, we trained a linear regression model on the test set.

Throughout this process, one of the main weaknesses we found was the quality of our data. Although we were able to find datasets with a large amount of entries, there were a lot of missing values. We were able to clean the data by opting to exclude certain variables with a significant amount of missing values. We also had to remove countries or years entirely if they did not have crucial values like the total emissions or GDP. This may reflect in the success of our model.

We evaluated our model using regression metrics such as RMSE to measure the error of our predictions on the test data. The success of our model relies on low error and the ability to accurately predict the total emissions of a region based on variables such as GDP, total population, and amount of forestland. To assess our model's accuracy, we created a scatter plot comparing the model's predicted and actual values as well as a histogram reflecting the distribution of residuals. To effectively communicate our results we used a combination of visualizations, performance metrics, and tables. We created graphs that visualize the correlation between features and the predicted total emissions, allowing us to show what values were the most impactful and the accuracy of our model. We then created a graph of the predicted emissions over time for the countries with the emissions, which is a helpful contrast to the similar graph that we created with the actual values during data exploration. Finally, we produced a variety of boxplots that compared the agrifood variables with our total predicted emissions quantiles.

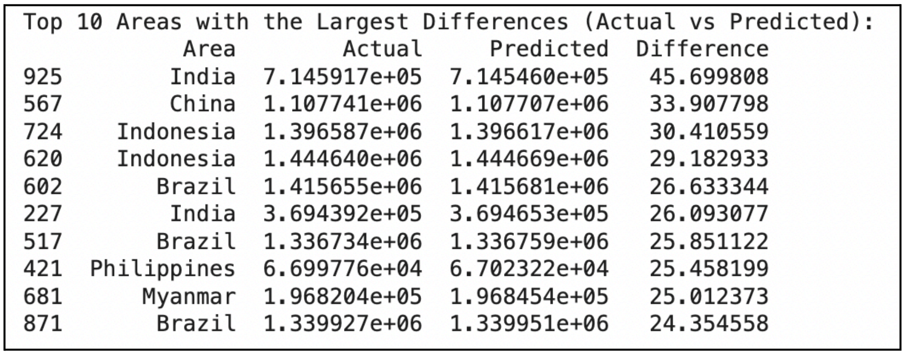
**Results**

Overall, our linear regression model performed with notable accuracy. The training MSE (19.71) and RMSE (4.44) indicated a moderate level of error in predictions on the training set. Whereas the test MSE (28.79) and RMSE (4.98) were slightly higher, suggesting minor overfitting but an acceptable generalization to new data. The model's residuals were mostly centered around zero, with some outliers/deviations that could have influenced the model's predictive accuracy.

Figure 3 represents countries that had the highest level of difference between actual and predicted values from our model. Countries such as China, India, and Indonesia tended to have higher levels of emissions, which could have allowed for more variation in accurately predicting their total emission values. This makes sense from an environmental perspective, as China and India are often discussed as two of the nations with the most weight in defining the environmental future of the planet. Both countries have intensifying urban populations, and since this was a factor with high correlation to total emissions, the continuation or discontinuation of this trend could be creating higher variability in our model's projections.

**Figure 3**

*Countries with the Largest Prediction Error*

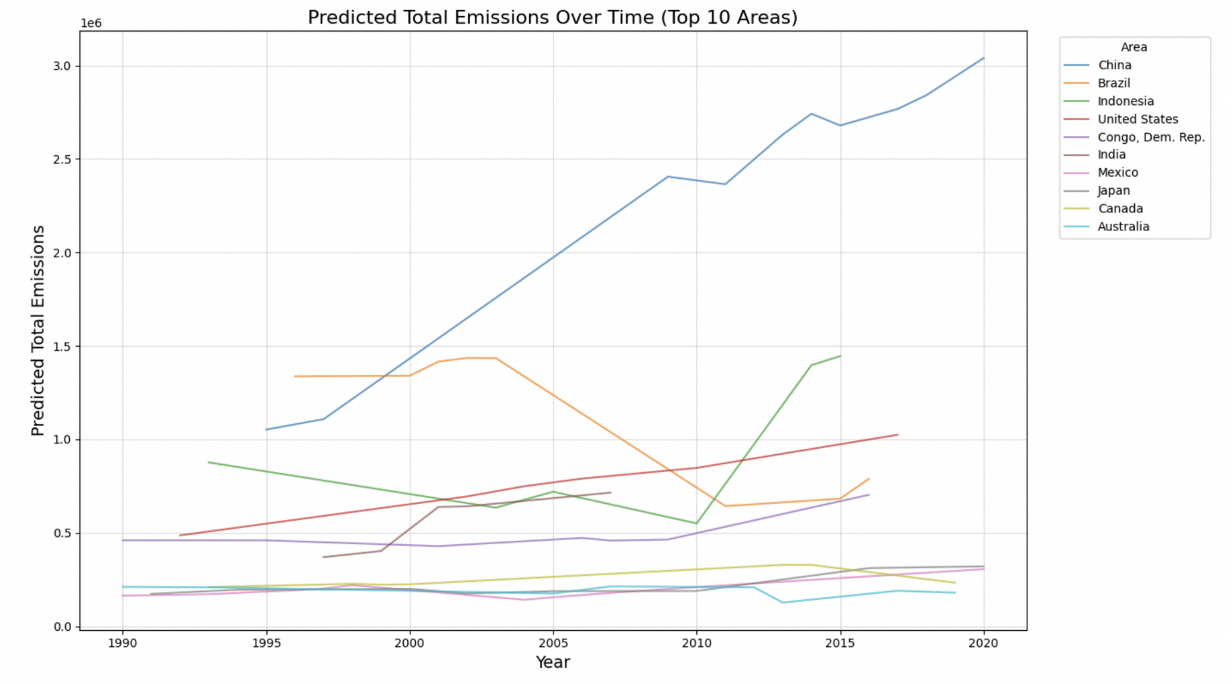


```
Top 10 Areas with the Largest Differences (Actual vs Predicted):
            Area        Actual     Predicted  Difference
925        India  7.145917e+05  7.145460e+05   45.699808
567        China  1.107741e+06  1.107707e+06   33.907798
724    Indonesia  1.396587e+06  1.396617e+06   30.410559
620    Indonesia  1.444640e+06  1.444669e+06   29.182933
602       Brazil  1.415655e+06  1.415681e+06   26.633344
227        India  3.694392e+05  3.694653e+05   26.093077
517       Brazil  1.336734e+06  1.336759e+06   25.851122
421  Philippines  6.699776e+04  6.702322e+04   25.458199
681      Myanmar  1.968204e+05  1.968454e+05   25.012373
871       Brazil  1.339927e+06  1.339951e+06   24.354558
```

Our linear regression produced the results in Figure 4 for the countries with the highest levels of predicted emissions over time. Due to some missing data in the dataset, not all years are accounted for in the prediction for each country (this could be improved with future studies incorporating more comprehensive datasets). Overall, the model showed similar results to the actual data (Figure 1) – with China being the country with the largest level of emissions that are increasing over time. It also reflected similar trends to the real data for Brazil and Indonesia - with Brazil seeing a significant drop in emission levels around 2010 while Indonesia saw an increase.

**Figure 4**

*Total CO2 Emissions in Highest Emitting Countries (1990-2020)*



The distribution of GDP values over different emission quantiles is reflected in Figure 5. The first two quantiles represent the lowest emissions, both with a relatively low median GDP, and a large amount of outliers. This suggests that there are numerous low-emission countries with a very high GDP. The last emission quantile (representing the highest emission levels) also has a comparatively low median GDP, with far fewer outliers. Overall, these results show a lower range of GDP values for countries with the top emissions.

Although the lowest and highest emission quantiles have similarly low median GDP's, one important observation is the lack of outliers in the highest emission quantile. This suggests that it could be easier to lower emissions as a high GDP nation that has greater access to influential resources and technology. In support of this idea, the Environmental Kuznets Curve (Figure 6) theorizes that environmental degradation initially increases with economic growth but eventually decreases as societies develop and prioritize sustainability among other factors (Ansari, 2023). However, at the turning point of the Kuznets curve, choice is introduced. For example, as a country sees greater economic opportunity, will it adopt more sustainable practices, such as transitioning to predominantly vegetarian diets to reduce emissions from red meat production? Similarly, will advancements in women's education lead to less births per capita and further contribute to environmental improvement? These considerations highlight the need to study other factors in addition to GDP when assessing environmental impact.

**Figure 5**

*Predicted CO2 Emissions by GDP per Capita*

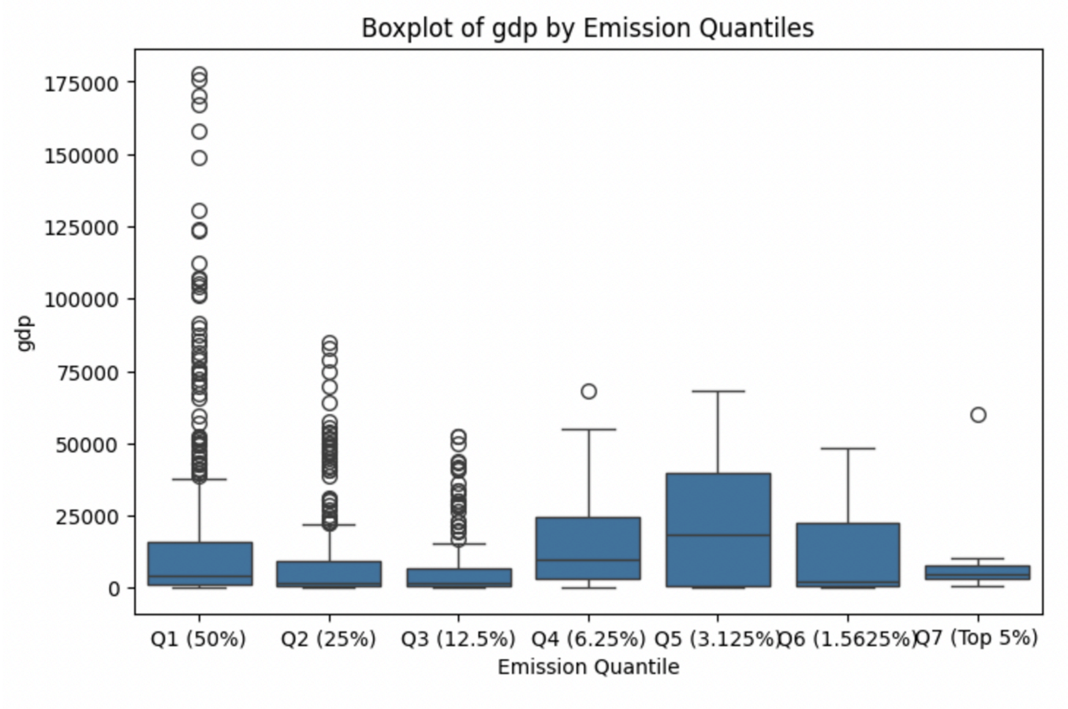
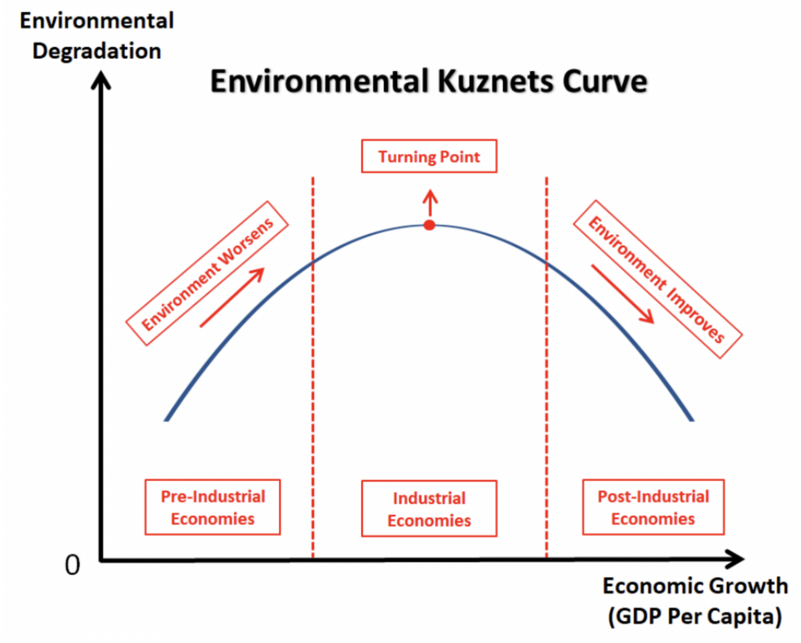


**Figure 6**

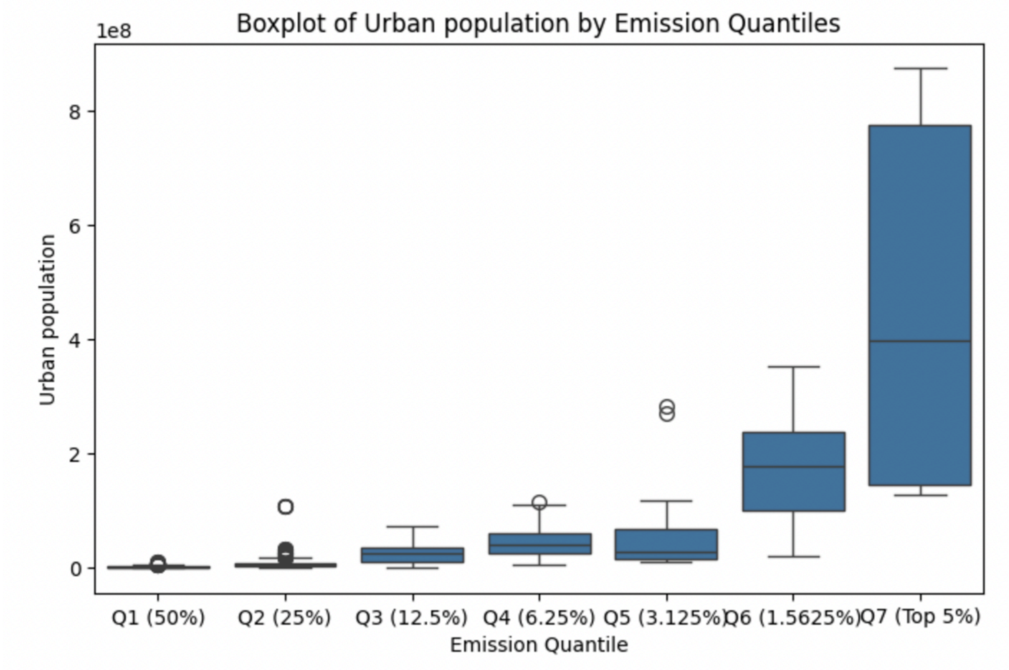
*Environmental Kuznets Curve*


Environmental Degradation
Environmental Kuznets Curve
Turning Point
Environment Worsens
Environment Improves
Pre-Industrial Economies
Industrial Economies
Post-Industrial Economies
0
Economic Growth (GDP Per Capita)

In our model, urban population level was one of the most highly correlated variables with total emission levels (Figure 7). The lowest emitters (Q1 to Q5) have extremely low median urban population levels and a low spread of population values. From quantiles 6 to 7, the median urban population significantly increases and the spread of values expands. Evidently, the top 5% of emitters (Q7) have the largest median urban population compared to other quantiles. This trend is unsurprising as areas with higher urban populations require more infrastructure, transportation, housing, and overall energy to support a larger amount of people. Nonetheless, the wider spread of values in Q5-Q7 could suggest that other factors (such as industrialization level or primary use of nonrenewable energy sources) may also play a role in higher emission levels.

**Figure 7**

*Predicted CO2 Emissions by Urban Population*



## Conclusion

Our project aimed to investigate how CO2 emission levels have evolved over time, how GDP per capita is related to emission levels, and what variables are the most correlated with total predicted emissions in order to advance our understanding of this global concern. To achieve these objectives, we combined two large datasets to parse out the most highly correlated variables with CO2 emissions as well as study the relation between emission levels and GDP per capita. In merging our two datasets and conducting an exploratory data analysis, there were a few issues that may have impacted our results. The GDP dataset and the agrifood dataset contained differences in regions which required us to remove areas that did not match between datasets. Additionally, while the cleaned data was quite comprehensive with a variety of variables and a large time span, some of the columns contained a notable amount of missing data. Specifically, countries with more than five years of missing data were dropped from the merged dataset, which slightly changed the nature of our study – rather than investigating emissions on a truly global scale, our study focused on countries that had adequate amounts of data. While our boxplot displaying GDP by emission quantiles had a low median GDP in the highest emission group with few outliers (indicating that countries with lower GDP are responsible for the top emissions), this may not accurately represent the trend on a global scale as we removed various countries that had too much missing data. Instead, these results showcase this trend with select countries that had enough data to be included in our study. Future research investigating this global relationship between GDP and CO2 emissions should address these gaps in our data and attempt to include more geographic regions in their analysis. Furthermore, we must acknowledge our missing data when analyzing other variables related to emission levels. Our most highly correlated variable with total emission levels was urban population, with our boxplot showing the highest emitters having the largest urban populations. Our results may reflect the removal of countries or regions that did not have enough data to be included in the analysis, potentially excluding areas that may have shown some outliers or any opposing data.

Our linear regression model, while providing valuable insights, had a few limitations that future research could address. The most critical limitation of our model was the substantial amount of missing data in both the GDP and agrifood emissions datasets. Excluding countries with five or more years of missing data may have skewed our final results and created bias within the data. The model demonstrated particular challenges in correctly predicting emissions for high emitting countries like China and India. These nations showed the highest prediction errors, suggesting that their emission patterns might be more complex than our linear regression model could effectively capture. While a linear regression model may be helpful for some regions, other regions might require a more advanced non-linear analysis to capture their trends in carbon emission and economic data. Our approach was further limited by the constrained number of variables we could effectively include in our analysis. The complex interactions between emissions and economic factors are too nuanced and need a more sophisticated model to capture the relationships. These limitations are not just shortcomings but opportunities for further research. Future studies could focus on developing more comprehensive and complete datasets, exploring non-linear modeling techniques, and incorporating a broader range of variables beyond GDP per capita and urban population. By acknowledging these constraints, we provide a strong foundation for a more nuanced investigation into global CO2 emissions, emphasizing the urgent need for continued research that can reveal the relationship between economic development and environmental impact.

With the rise of CO2 emissions worldwide, more studies must be conducted to fully understand and address this global matter (U.S. Global Change Research Program, 2022). Our study provides an ample starting point for future research to fully investigate this issue. Future research studying CO2 emissions over time in various regions can utilize this model as a starting point to create a model that can generalize better or predict future environmental/economic trends. For far too long, data availability has limited environmental monitoring and modeling. This is especially the case for developing countries which may not have the resources to devote to climate tracking. While many developed countries choose to lay blame on these developing countries for lack of data reporting or environmental standards, it is important to analyze models like ours to understand how these countries could develop more sustainably. There are many emergent technologies that are making global environmental data more equitable across country boundaries. These come in many forms, but particularly useful to a study like this would be developments in Landsat and LiDAR remote sensing techniques (Wasser, 2022). Data collection like this is done remotely through satellites, and while some data gathering inequality still exists, it vastly increases the ability for continental-global scale data collection. These satellite systems use ground truthing methods to report land use changes and other environmental factors, which can help us better to understand environmental changes occurring in relation to CO2 emissions. This is especially important for developing countries which may not have advanced data gathering or reporting capabilities.

Understanding CO2 emissions is also critical for anticipating the broader economic, environmental, and societal impacts of climate change. By studying emission data and its relationship with variables like GDP, urbanization, and industrial activity, researchers can identify opportunities to develop more sustainable growth strategies. While our study revealed meaningful correlations and trends, it also highlights the complexities and challenges in analyzing CO2 emissions. Addressing the gaps in data and incorporating more sophisticated methods would be critical for future research. As the global community works towards lowering emissions to mitigate climate change, continued research efforts like this can play a vital role in informing policies, fostering international cooperation, and securing a sustainable future for the next generations.

# References

Ansari, S. (2023, June 19). The Kuznets Curve. Economics Online. https://www.economicsonline.co.uk/definitions/the-kuznets-curve.html/

Bello, A. L. (2023, July 17). Agri-Food CO2 Emission Dataset - Forecasting ML. Kaggle. https://www.kaggle.com/datasets/alessandrolobello/agri-food-co2-emission-dataset-forecasting-ml

Beynon, J., & Wickstead, E. (2024, October 17). Climate and development in three charts: An update. Center for Global Development. https://www.cgdev.org/blog/climate-and-development-three-charts-update

Environmental Protection Agency. (n.d.). Overview of Greenhouse Gases. EPA. https://www.epa.gov/ghgemissions/overview-greenhouse-gases#carbon-dioxide

Rooper, H. (2024, June 20). Emissions growth in the developing world . Climate Leadership Council. https://clcouncil.org/blog/emissions-growth-in-the-developing-world/#:~:text=As%20emissions%20decrease%20in%20the,this%20trend%20will%20likely%20continue.

Tas, O. C. (2022, March 19). World GDP(GDP, GDP per capita, and annual growths). Kaggle. https://www.kaggle.com/datasets/zgrcemta/world-gdpgdp-gdp-per-capita-and-annual-growths

U.S. Global Change Research Program. (2022). Atmospheric Carbon Dioxide | GlobalChange.gov. Www.globalchange.gov. https://www.globalchange.gov/indicators/atmospheric-carbon-dioxide

Wasser, L. (2022, August 31). *The Basics of LiDAR - Light Detection and Ranging - Remote Sensing | NSF NEON | Open Data to Understand our Ecosystems*.

Www.neonscience.org.

https://www.neonscience.org/resources/learning-hub/tutorials/lidar-basics