Introduction

Our group chose to look at data concerning global emissions and their possible connections to severe storm events in the United States. Our data was pulled from various sources including the National Oceanic and Atmospheric Administration (NOAA) agency as well as the Global Carbon Project and Global-Warming.org.

Questions

Our goal was to answer the following questions:

- What is the overall trend of emissions in relation to severe weather events over a certain period of time? What types of storms are most affected?
- How does increase or decrease in storms affect overall damage caused by storms to property? To crops?
- What is the distribution of CO2 emissions by continent/region?
- What are the correlation between different fossil fuel sources and other factors with co2 emissions
- Which countries are highest in carbon emissions, has this changed within recent years ? (top 10 countries)
- How does temperature change over time and how does it relate to carbon emissions?
- How has CO2 emissions changed in the top 5 countries over time? As GDP increases or decreases does this impact CO2 emissions?
- What are the most consumed energy types and changes in energy consumption patterns in the United States?
- What is the Energy Intensity Per Capita and by GDP of Top 5 Energy Consumers?

Approach

We followed a general approach of "divide and conquer," delegating tasks based on each member's comfort level working with each technology. This way we could complete multiple tasks in parallel and speed up our production timeline. Additionally, this allowed for us to use each team member's strengths to our best advantage. This approach led to Zach leading the Kafka portion, Armin heading the machine learning section, and Salma and Carol working through the visuals and Dash dashboard. Other tasks were completed when needed by whoever had spare time to complete them.

Data Sources

We used five different datasets for this project including three datasets on emissions, one on temperatures, and one on storm data. For the emissions dataset, two were static CSVs while the other was pulled from an API. Meanwhile, the temperature dataset was also a static CSV and the storm data was collected by combining many CSVs from the National Oceanic and Atmospheric Administration (NOAA) together into one. Each of these sources is cited below.

Andrew, R. & Peters, G. (2022). The Global Carbon Project's Fossil CO2 Emissions Dataset, Version 27. Retrieved January 13, 2023 from https://doi.org/10.5281/zenodo.7215364.

Global-warming.org. (2023) Daily global seasonal cycle and trend value, Version 1. Retrieved January 18, 2023 from https://global-warming.org/api/co2-api.

NOAA. (2020, December). Storm Events Database, Version 3.1. Retrieved January 13, 2023 from https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles.

NOAA. (2022, September). Climate at a Glance: Statewide Average Temperature, Version . Retrieved January 13, 2023 from https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series.

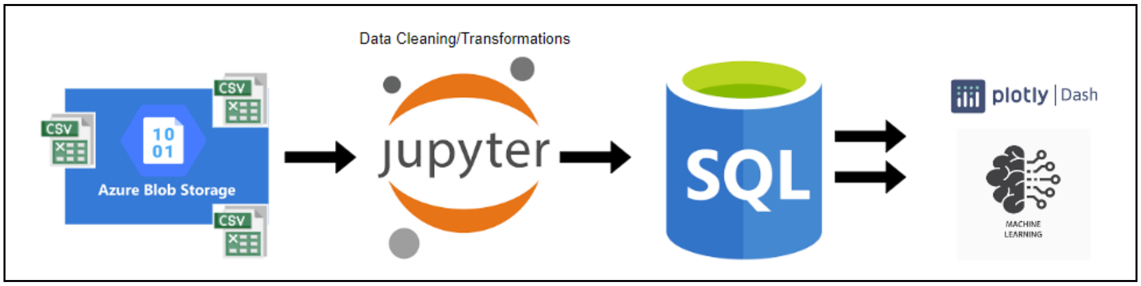
Vanous, B. (2021). Countries CO2 Emissions and More, Version 1. Retrieved January 16, 2023 from https://www.kaggle.com/datasets/lobos/c02-emission-by-countries-growth-and-population.

Discussion

Data Pre-Processing & Cloud ETL

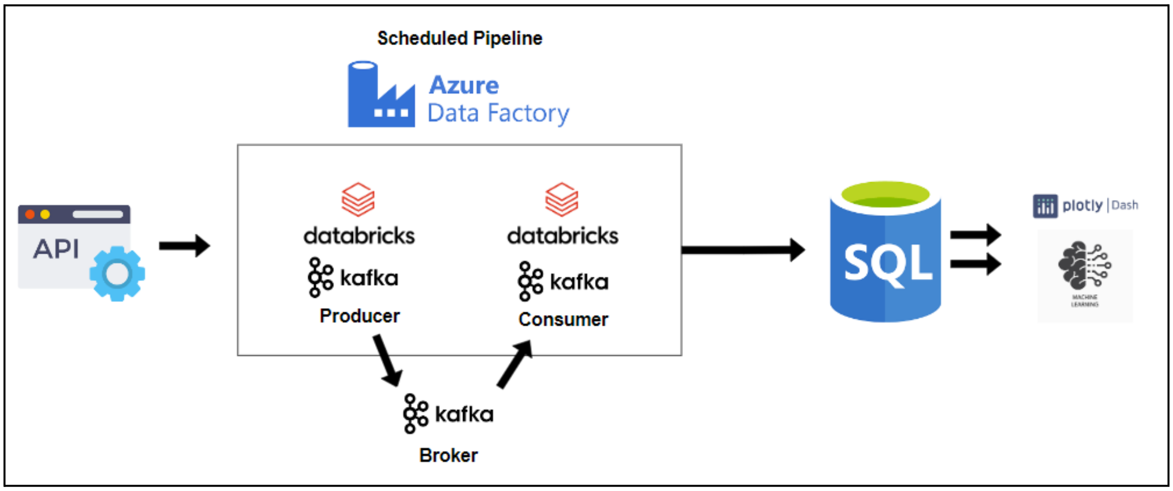
Once we had collected all of our data sources, the first step in the project was to clean the data and transform it into a more usable format. This process was broken down into two portions: one to handle the static data pre-processing and another to handle the API data via cloud ETL.

Each of the static datasets was put through a similar process that required extracting the data from a source, transforming or cleaning the data, then loading the cleaned data elsewhere. For this process we were able to set up a blob storage container within Microsoft Azure to hold the CSVs. By doing this, we were able to access our blob using python code via jupyter notebook and extract the data to perform operations on it. Depending on the dataset, various cleaning actions were performed on the data. These actions could include dropping columns or duplicates, renaming columns, modifying or creating new columns, filling nulls or zeros, etc. Once the dataset was in an acceptable form, we then loaded it to our SQL database to be stored for future use creating visuals or building a machine learning model. A basic pipeline diagram can be seen below for our static data.



A similar pipeline was built to handle the CO2 emissions API data; however, this dataset was not static. Everyday one entry would be added to the dataset, so we would need to run our pipeline at least once a day to maintain all the relevant data. Additionally, other previous entries were sometimes revised, so we also had to be able to update those entries whenever necessary. To handle this situation we turned our attention to Cloud resources, more specifically Kafka, Azure Databricks, and Azure Data Factory.

Using Azure Databricks, we built a Kafka producer that would send individual rows of the dataset as a message to a Kafka broker. For this project we used Confluent Kafka as a broker with the instructors controlling all of admin privileges related to that. Then again using Databricks, we built a Kafka consumer that would be able to receive the messages from the broker. Once the messages were received, they were transformed and formatted correctly, then placed into a dataframe. This data frame was then written to the SQL server, where the data would be housed. For the final step of this process, we used Data Factory to schedule and automatically trigger the pipeline to run a few times a day. By doing this, the data should update automatically every few hours to the most recent data from the API. The final cloud pipeline can be seen in the diagram below.



More information regarding specific steps taken to each dataset, including cloud ETL, can be found on our github within the ETL folder at the following link: Github ETL Folder.

Machine Learning

For the machine learning portion of our project, we trained several algorithms on two different datasets in order to see which variables lended themselves well to predictive analysis and how accurate of a model we could create.

The first dataset we tried was NOAA's Storm Events dataset which contains records on storms and other significant weather phenomena stemming back to the mid 1900s. The data recorded was in the form of CSVs and tracked things like the start and end date of the weather event, the state and counties it affected, how much damage it caused (property, crop, and loss of life), as well as a few other miscellaneous points.

After following the cleaning procedures outlined in the Pre-Processing section, we spent some time analyzing the data to identify possible target variables and training features, eventually settling on training the model to predict storm damage based on time series forecasting. We made a simple function to take in a target variable and split the date-time objects in the testing set into the year, day, etc. and put them in a feature matrix for the model to train on; however, our results were fairly disappointing.

Using XGBoost, the model was only able to muster up a paltry ~0.005 r^2 value, meaning our regression model was not a good fit for the data. We regrouped in order to analyze why that may be, and we noticed that our target variables had a very large proportion of zeroes present, since most storm events would be relatively mild and not cause any sort of meaningful damage. We theorized that this may be making it difficult for our model to accurately predict outcomes, so after some research online to see how others had tackled such problems previously, we decided it may be worth using one of scikit-learn's meta models, specifically the Zero-Inflated Regressor.

From scikit-learn's documents:

> Meta models are models that depend on other estimators that go in and these models will add features to the input model. One way of thinking of a meta model is to consider it to be a way to "decorate" a model.
>
> There are regression datasets that contain an unusually high amount of zeroes as the targets…The classical machine learning algorithms can have a hard time dealing with such datasets. Take linear regression, for example: the chance of outputting an actual zero is diminishing. Sure, you can get regions where you are close to zero, but modeling an output of **exactly zero** is infeasible in general. The same goes for neural networks. What we can do to circumvent these problems is the following: 1. Train a classifier to tell us whether the target is zero, or not. 2. Train a regressor on all samples with a non-zero target.

We tested different combinations of random forest and linear regression classifiers and regressors, but we found that this model's performance did not vary meaningfully from the XGBoost model.

We then performed some feature engineering on the original dataset in an attempt to improve our model's accuracy, since we thought perhaps time series data alone was not strongly correlated with storm damage. Our thought process was that the same storm could cause wildly different amounts of damage in different areas based on socio-economic conditions such as their investment in infrastructure, shelters, and people. We attempted to provide this data to our model by creating new labels to signify the region each storm occurred in (we did this as opposed to just feeding it the raw state data since that would create a bloated feature matrix after one-hot encoding), and then performed one-hot encoding in order to convert each region to an array of 0's and 1's for the model to interpret. We followed this same approach with the column representing what type of storm event occurred (thunder storm vs blizzard etc.), as we thought this would likely be a good predictor as well.

Unfortunately, even after this we could not get either model's performance to improve in any meaningful sense. Our best guess as to why this is, is that our regional labels were still far too broad to provide meaningful information for the model. When we graphed the feature importance for the XGBoost model, it showed that only a few of the regional labels were meaningful, and even then they were still very lowly scored compared to some of the other features like weather types and storm duration. More precise information was available in the form of county labels for each entry, but there were thousands of unique entries and converting them from categorical data to a format that the model could interpret caused memory errors due to how large the dataframe became. Our takeaway here is that there were too many factors involved in the amount of damage a storm event can cause for our simple models to predict accurately, especially with the features we had available.

We then turned our attention to a different dataset, this time looking at global emissions to predict future temperature changes relative to historical precedents. The dataset contained emissions from different sources, such as coal, gas, flaring, etc. which we believed might be a strong predictor of temperature fluctuations.

We tried several models including XGBoost again, an SVM (support vector machine) model, and a linear regression model.

Breaking down our results, the XGB model had a much more promising r^2 value this time, however it was still relatively weak (only 0.353). The graph of the residuals showed a downward trend as well, indicating it was a poor fit for the data. The SVM model performed similarly (r^2 of 0.382), but this time had an upward trend in the residuals, again indicating it was a poor fit. Our linear regression model did not indicate any fit problems when looking at the residuals, but the r^2 was sub 0, indicating very poor predictive performance.

We theorize that our issues stem in large part from the size of the dataset. Since they only tracked emissions by year, there were less than 100 data points present which makes it difficult to train a model well. This issue becomes especially apparent when changing the random seeds for the models. We wrote a function to check the performance of the first thousand random seeds, and we found that we could improve our model's accuracy by up to 200%.

Overall none of these models are consistently good, leading us to believe the data doesn’t lend itself to this sort of predictive analysis. Our target variables are incredibly complex phenomena, which require either more complex models, or more robust features.

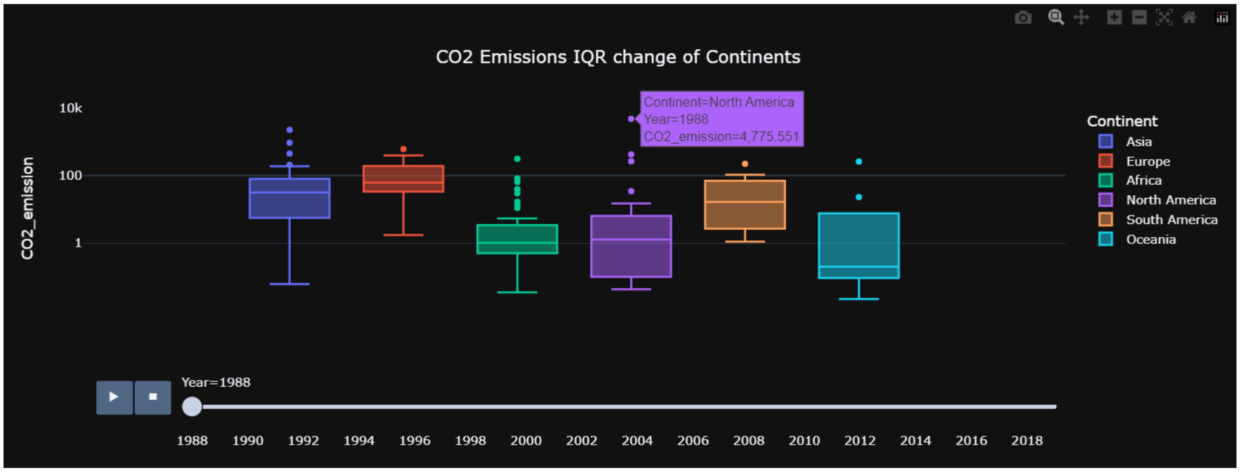
Dash

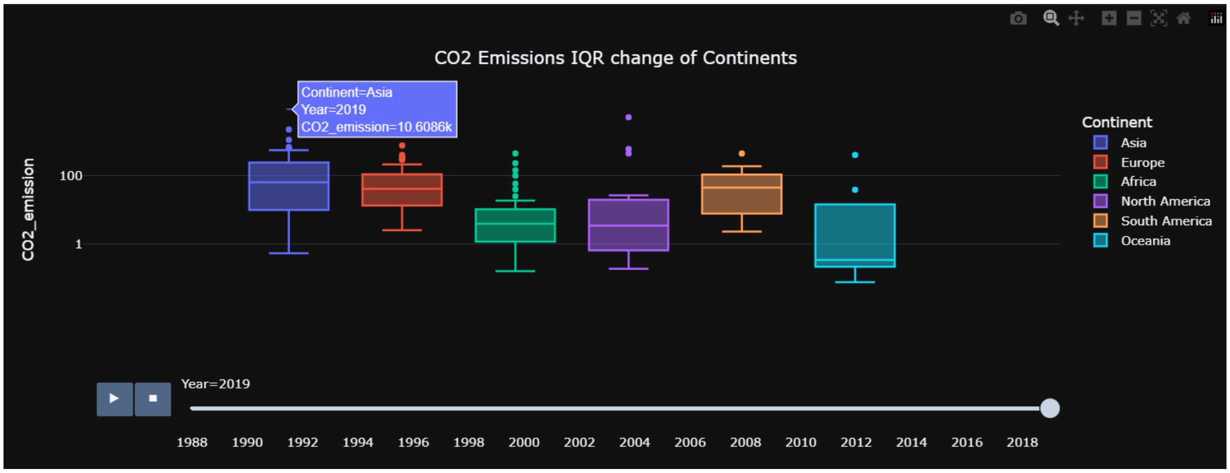
Our group decided to utilize Dash to display our visualizations. We had never worked with Dash before so the first few days were mostly looking up different resources to aid in our task. We watched youtube videos, looked up instructions and starter code from the plotly website, and attended demos for Dash. We also made our napkin drawings to help us set up our visualizations in a way that made the most sense to us and helped answer the questions for our project. Before actually putting our graphs into Dash, we made our visualizations separately in jupyter notebook using libraries such as matplotlib and seaborn in which we were familiar with having used them before. Later, we learned that the library plotly was used in Dash and was a bit more complicated to use the libraries we originally used so we had to change our code to use plotly instead. This took a little while, but once we figured it out we were able to put them on Dash. Finally, once we put our graphs on Dash we used Dash Bootstrap Components, a library for use with plotly Dash, which makes it easier to build more complex and stylish layouts.

Results and Findings

What is the distribution of CO2 emissions by continent/region?

Knowing the distribution of CO2 emissions is useful for businesses to assess their own carbon footprint, comply with emissions regulations, analyze competitors and opportunities, manage their supply chain and make informed investment decisions. This question was answered by grouping countries into their respective continents based on latitude and longitude and using the Countries CO2 Emission dataset. This was done to broaden the scope of the analysis.

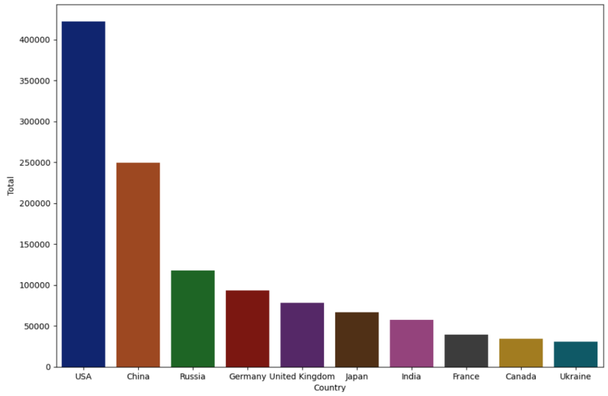


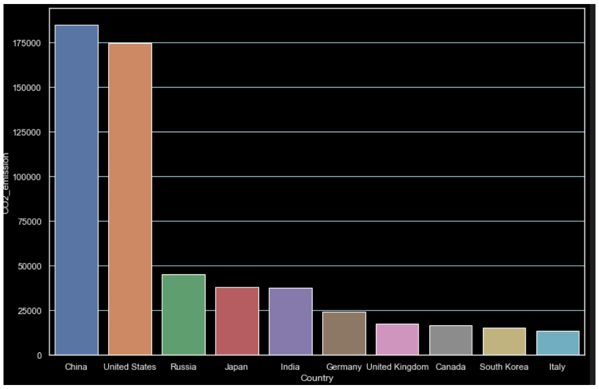


From 1988, Europe had the highest median and interquartile range (IQR) for CO2 emissions, but North America had some countries with emissions that exceeded even those in Europe. However, after 1990, Asia surpassed all other regions with the highest median and IQR of CO2 emissions, and had the highest emission outliers. In order to identify the outliers, we further analyzed the countries with the highest carbon emissions

Which countries are highest in carbon emissions, has this changed within recent years ? (top 10 countries)

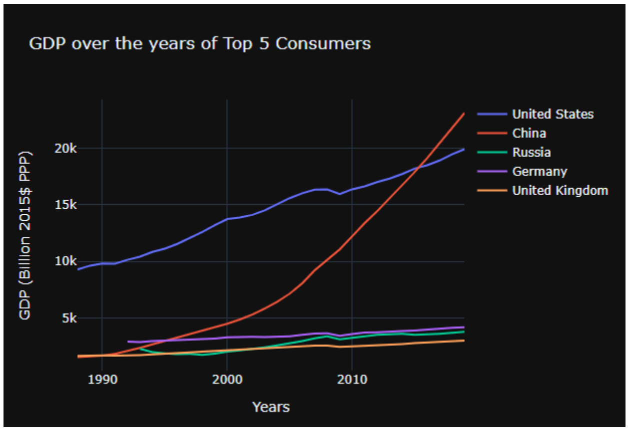
We used the Countries CO2 Emission dataset to help answer this question. We were interested in this question because we have seen trends over time that show the rise and fall of emissions due to the changes in the economy. Furthermore, the United States has been a big contributor to greenhouse gas emissions from things such as burning fossil fuels and we wanted to see if it was the highest in the world or at least in the top 5. The first graph shows the countries with the highest carbon emissions from the early 2000s and the second graph shows the carbon emissions from the last decade. From our hypothesis, we were correct about the United States leading in carbon emissions but that has changed over recent years as China has taken over. We can also see South Korea and Italy have replaced Ukraine and France but the rest of the countries that lead in carbon emissions are still on the top 10 list. Russia still stands in third and Japan has increased in its emissions as well as India.

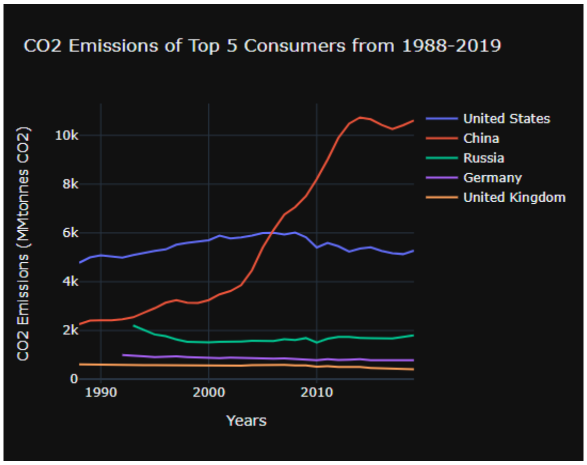




How has CO2 emissions changed in the top 5 countries over time? As GDP increases or decreases does this impact CO2 emissions?

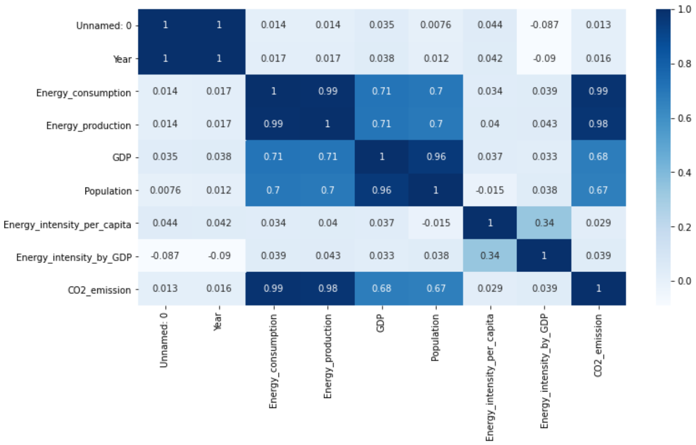
After analyzing the top countries with the highest carbon emission, we wanted to see the trend over the years for the top 5 countries. We also wanted to see if GDP had any effect on carbon emissions and if GDP increased would carbon emissions? As we can see, the United States dominated in the early 90s and 2000s but we see an increase in GDP as the carbon emissions increase as well. There seems to be some correlation between the two. After further research, what severely impacts carbon emissions is whether countries would be able to reduce their energy consumption and production to offset their large increases in GDP. If improvements have not been made or were slow the co2 emissions grew rapidly. Many countries benefit from economic growth by transitioning towards industry, manufacturing, and construction activities that lead to large energy inputs which China has been doing and it is seen in the graphs as it grew rapidly after the early 2000s.

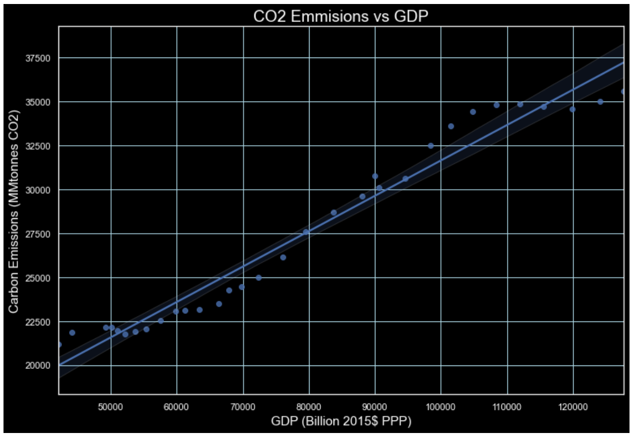


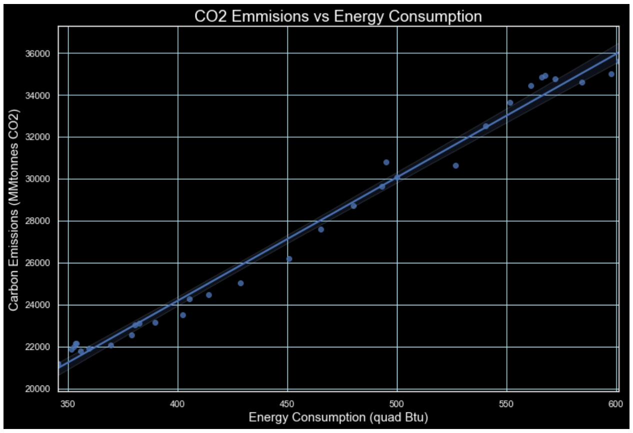


What are the correlations between different fossil fuel sources and other factors with co2 emissions?

To see the correlation between the different fossil fuel sources and co2 emissions we made a heatmap. We specifically looked at the correlation between co2 emissions and population, GDP, energy consumption, and energy production. We hypothesized that the correlation between those variables would be higher and more than .5. When we look at the correlation between co2 and population it is around .67 and GDP is .68 which are closer to 1 meaning it is of higher correlation. The highest correlation we observed were energy consumption and energy production which were .99 for consumption and .98 for production. This is not surprising because energy consumption measures the amount of consumption for specific energy sources and production measures the amount of production for specific energy sources.



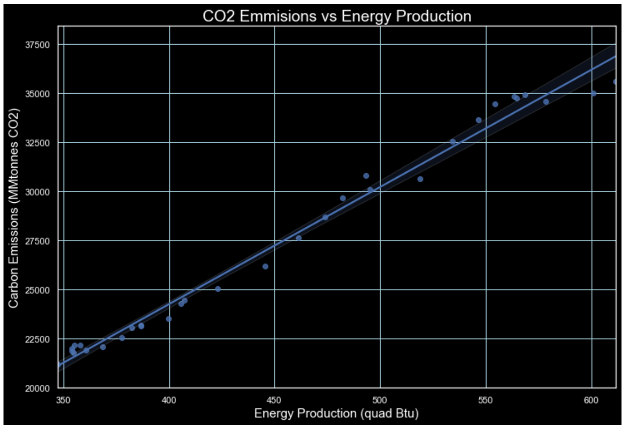


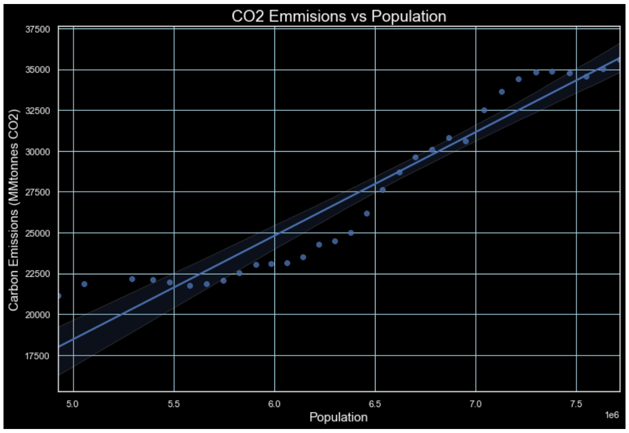


After analyzing the heatmap, we decided to make scatter plots of the gdp, population, energy consumption and production against co2 emissions. As you can see from what our heatmap portrayed, for the first graph of co2 emissions and gdp, as gdp increases so does co2 emissions. The scatter plot shows a positive correlation between the two variables in which the trendline denotes. The next graph shows a positive correlation between co2 and energy consumption. The next two graphs, co2 vs population and co2 vs energy production, both relay the same information in which there is a positive correlation between the variables.

We hypothesized as GDP increases so would co2 emissions. This is mostly because emissions tend to increase when there is more access to money. This is due to the fact that having more money would increase consumption of electricity, heating, transport and other things that require energy. There is a heavy reliance on fossil fuels which increases the co2 emissions. This also gives reason to why there is a positive correlation between consumption, production and co2 emissions.

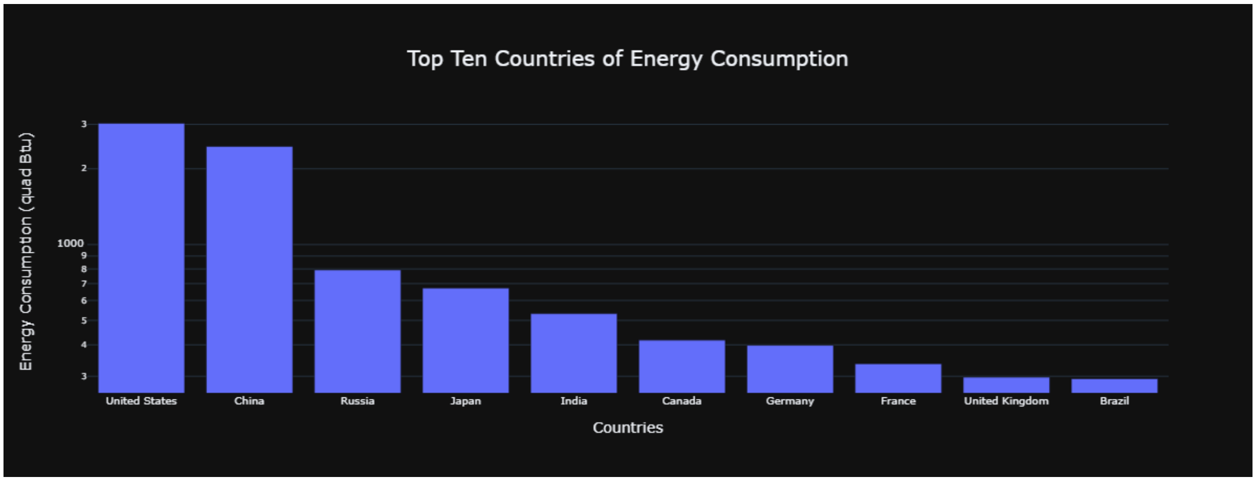
The last graph of co2 emissions and population, the positive correlation makes sense. Human activities are responsible for the increase in greenhouse gasses for things mentioned above like electricity and transportation.



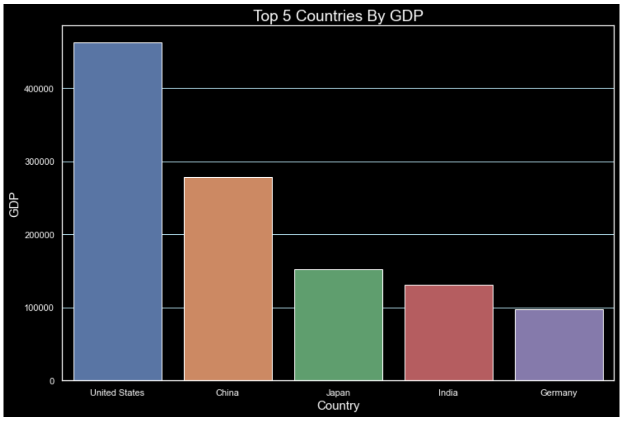


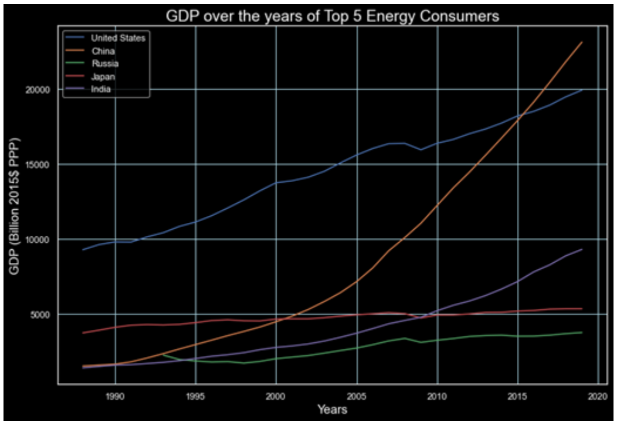
What are the most consumed energy types and changes in energy consumption patterns in the United States?

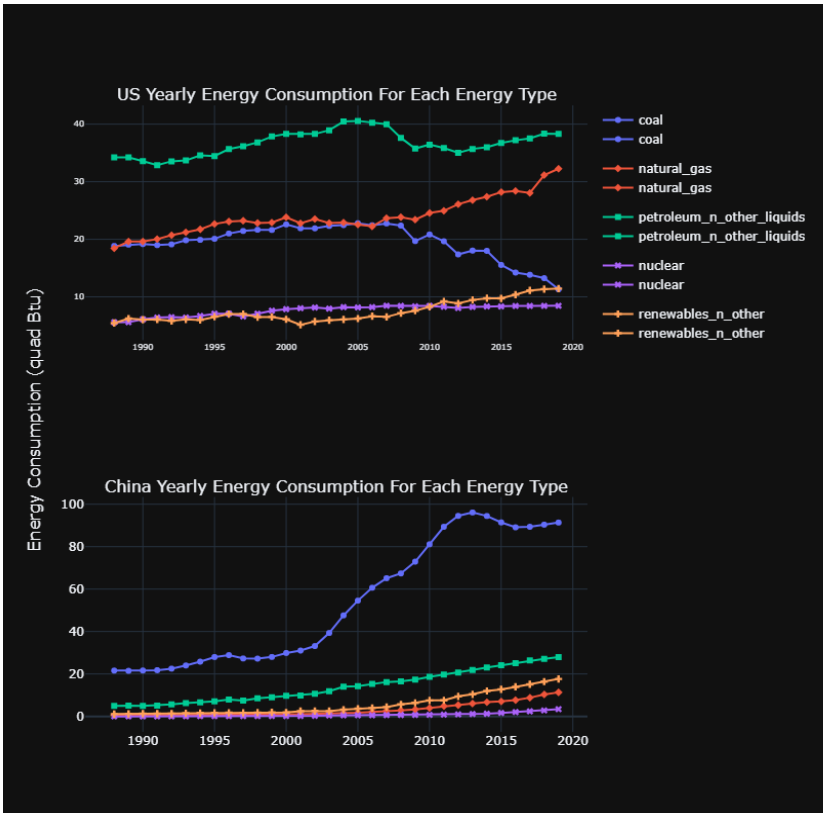
Energy consumption refers to the amount of electrical energy and demand consumed for any billing period. There are many factors to consider when it comes to measuring a country's energy consumption, mainly the difference in consumption between a country, its industry, and its population. Population energy use is often measured in terms of per capita energy intensity, while industrial energy use is often measured in terms of GDP energy intensity, which we will discuss later.



We first evaluated what were the top 10 energy consumers in the world. China and the United States are the largest energy consumers globally due to their status as the two largest economies in the world. From 2000 to 2019, China experienced the quickest increase in GDP among the top energy-consuming nations, with the United States coming in a close second.

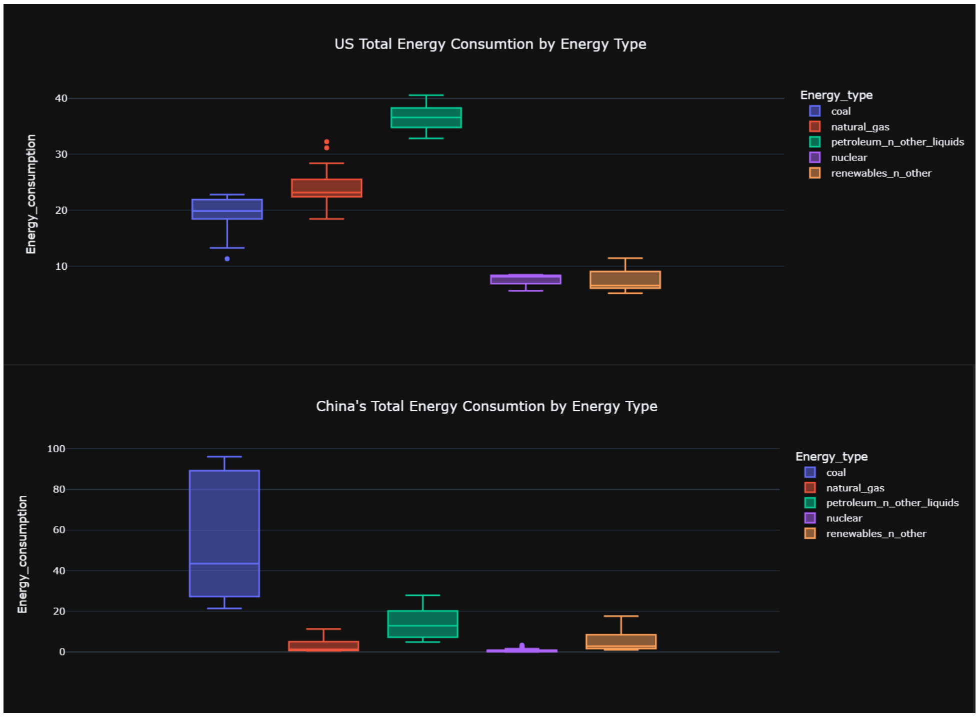

Top 5 Countries By GDP


GDP over the years of Top 5 Energy Consumers

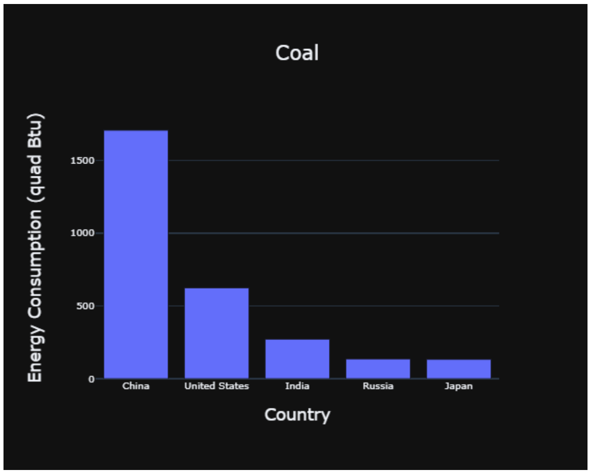

US Yearly Energy Consumption For Each Energy Type
China Yearly Energy Consumption For Each Energy Type

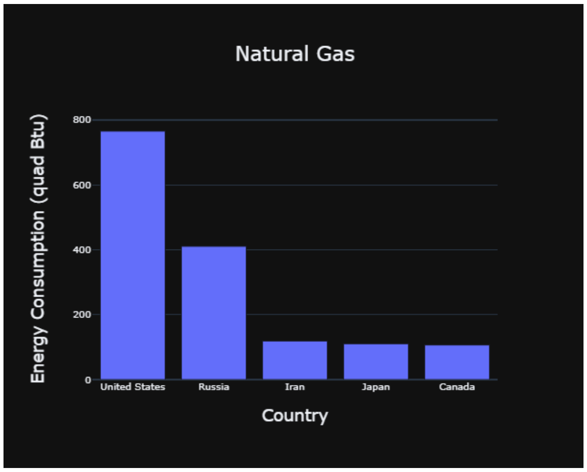
The energy sources in the United States are diverse and can be categorized into primary and secondary, renewable, and nonrenewable, and fossil fuels. The main primary energy sources, found in our dataset, are fossil fuels (petroleum, natural gas, coal), nuclear energy, and renewable sources. The most widely used energy sources in the United States are petroleum/other liquids and natural gas. From 2007 to 2019, the consumption of coal decreased by nearly 50%, and in 2019, it reached the same consumption level as renewables. Nuclear energy experienced growth from 1988 to 1990, then stabilized after reaching a plateau around the year 2000. The US primarily consumes petroleum and natural gas due to abundant reserves, established infrastructure, consumer demand for transportation and heating, government support through policies and incentives, and limited alternatives to meet the demands of the economy.

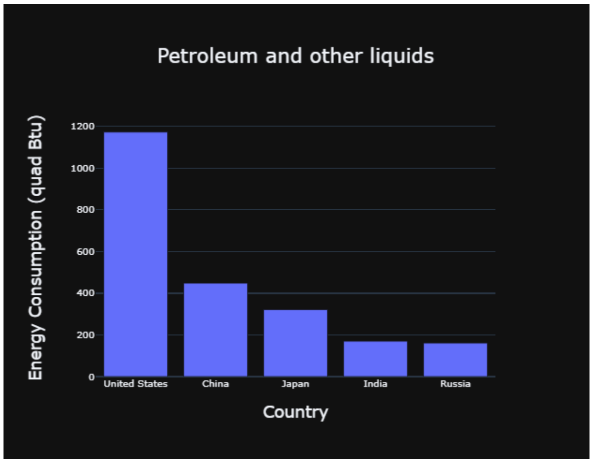
In China, coal is the primary energy source consumed, while consumption of other energy types has slowly increased over time. China experienced rapid economic growth and industrialization starting in the late 20th century, leading to increased energy consumption and heavy reliance on coal-fired power plants. China experienced a 7.3% decrease in coal consumption from 2013 to 2016, demonstrating a disconnection between economic growth and the growth in coal consumption. It is possible that China's coal consumption has already reached its peak.

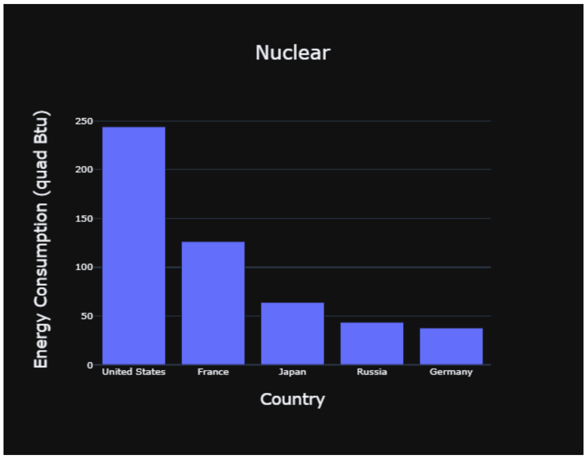


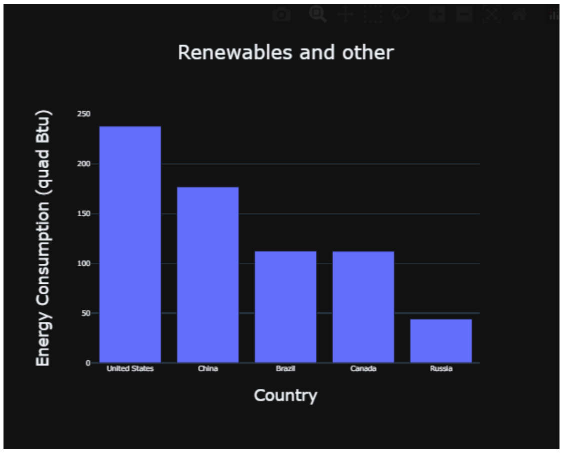
Throughout the entire period, the United States has been the leading user of clean energy which is composed of renewables and nuclear energy. The US leads in total consumption of most energy sources, except coal, where China's consumption is nearly three times that of the US.





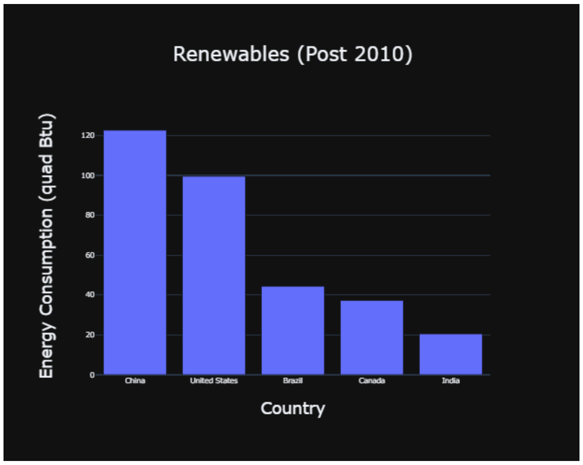


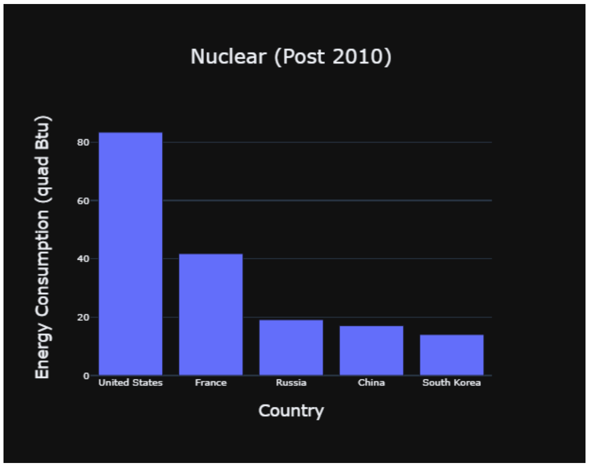




Upon further examination, after 2010, the United States continues to be the major consumer of nuclear energy, however, in regards to renewables and other energy sources, China has boosted its consumption and overtaken the United States to become the primary consumer of renewables. Policy efforts aimed at reducing air pollution have accelerated the decline of coal in China's energy mix. Since

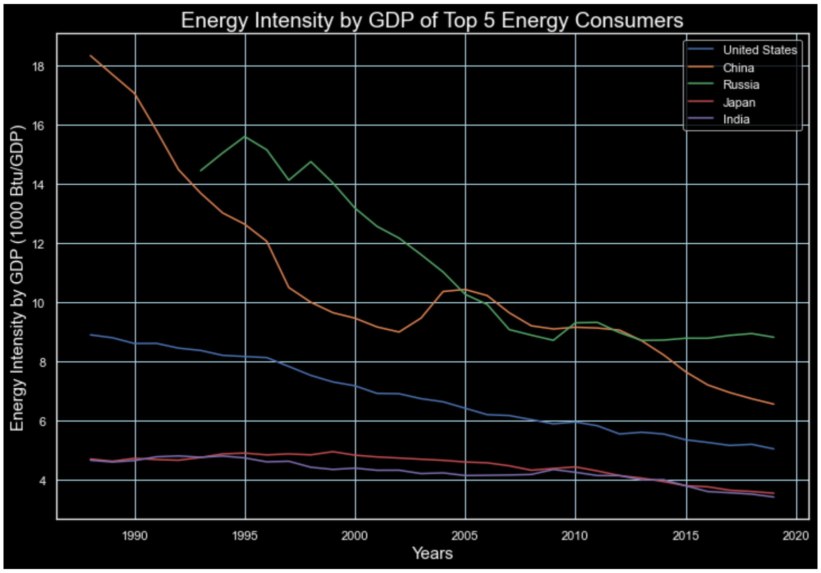
2006, there have been binding targets on air pollution and energy intensity, and reducing coal use has been a key means of compliance. Additionally, advancements in coal plant efficiency and the deployment of clean energy have reduced coal intensity.



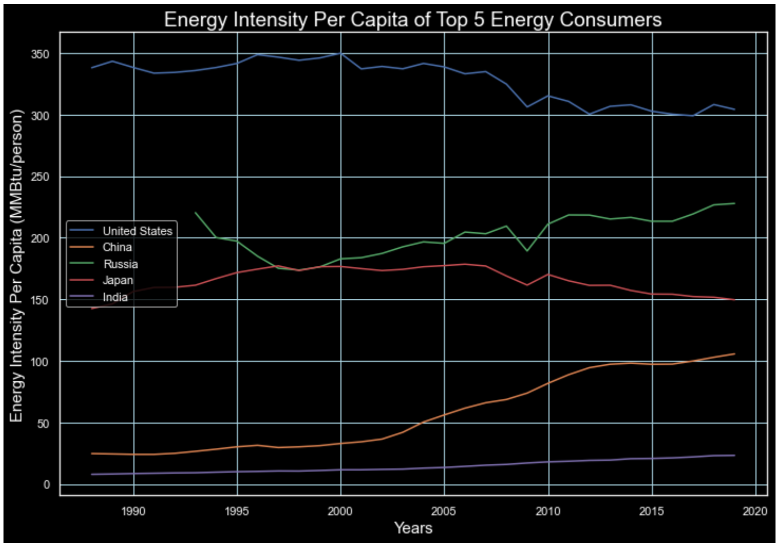


What is the Energy Intensity Per Capita and by GDP of Top 5 Energy Consumers?

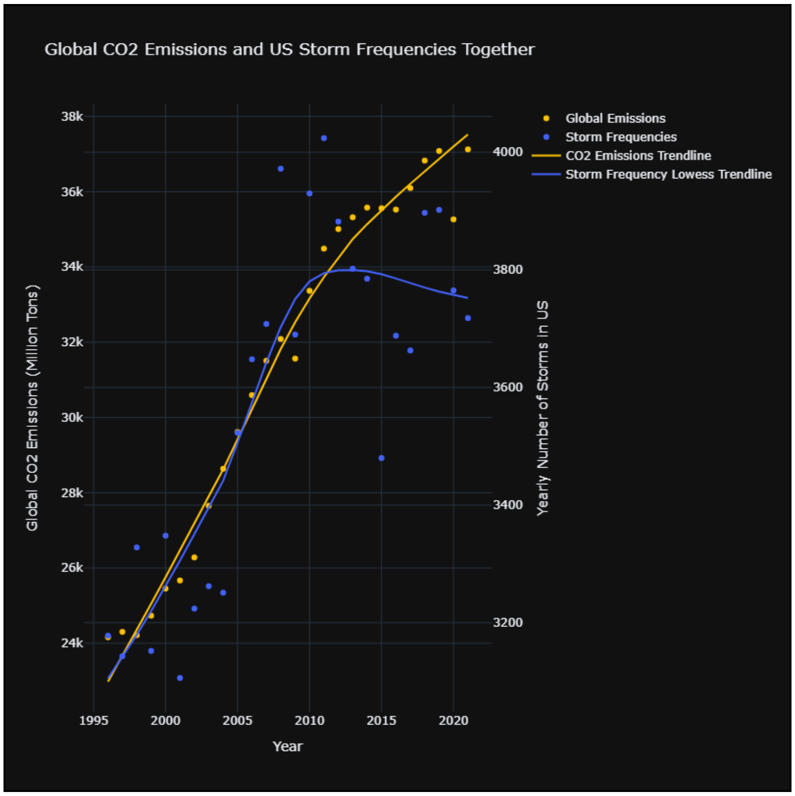
Energy intensity, calculated as the ratio of total energy consumption to real GDP, is a widely used energy indicator and measure of efficiency, with lower values indicating greater efficiency. In this section, it is observed that the top 5 consumers have all reduced Energy Intensity by GDP. Factors such as improvements in processes and equipment, as well as structural and behavioral changes, can influence changes in energy intensity. However, at the economy-wide or end-use sector level, energy efficiency is not a straightforward concept due to the diverse nature of output, including the production of a wide range of goods, the combined transportation of goods and people, and varied housing and climate conditions. Thus, using GDP to calculate an aggregate energy intensity number can obscure rather than shed light on energy use, unless more information on sector details is acquired.



The energy consumption of a population is often gauged by Energy intensity per capita. However, the top 5 consumers seem to have either stagnant or increasing Energy intensity per capita. This could be due to their higher spending capacity on amenities such as technology, appliances, and transportation. Energy intensity per capita can also be influenced by behavioral factors and may not necessarily reflect improvements in energy efficiency. For instance, as people get older, they tend to consume more energy for heating during winter, leading to increased energy intensity even though the heating equipment remains the same. It can be challenging to differentiate between changes in behavior and structural changes, such as demographic shifts like an aging population, which can also drive behavioral changes.

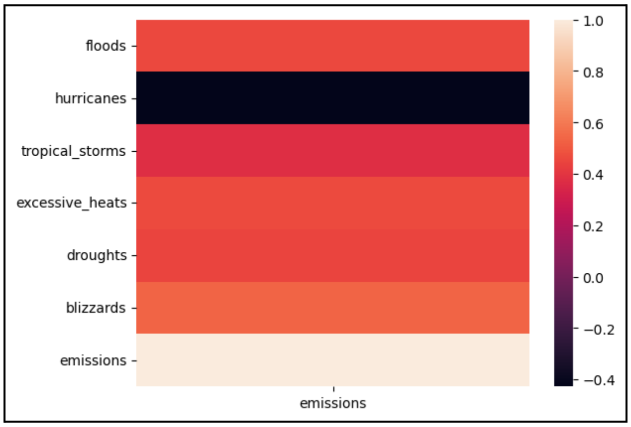


What is the overall trend of emissions in relation to severe weather events over a certain period of time? What types of storms are most affected?

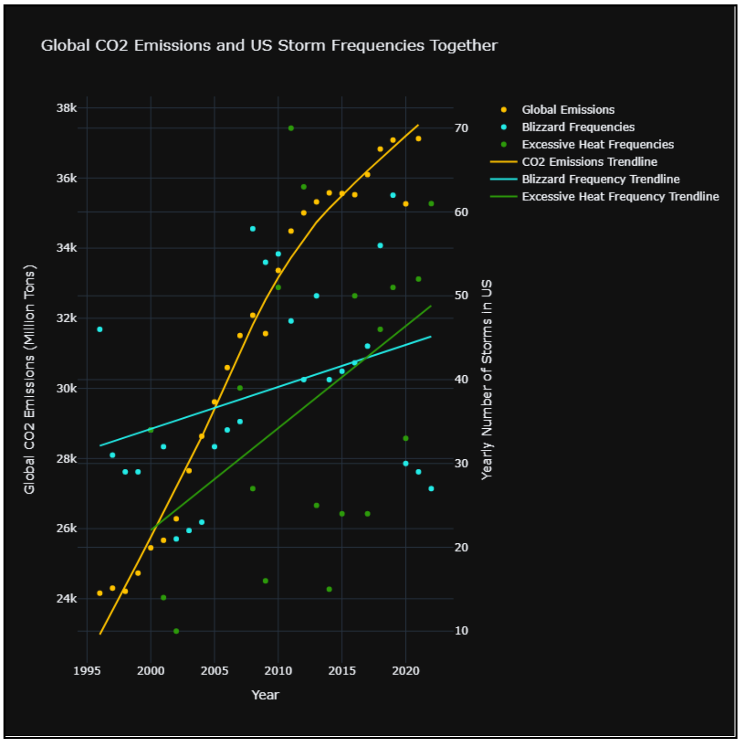


In order to understand the relationship between storm frequency (blue) and global CO2 emissions (yellow), we graphed them together with multiple y-axes. The first y-axis tracks global CO2 emissions by million tons, while the other axis tracks yearly number of storms in the US. By creating both axes, we are able to map a scatterplot of each set of points over one another to better see how they relate or affect each other. Looking at the graph to the left, it is clear that both sets of data trend upward for a majority of the timeframe.

Additionally, they follow very similar upward trends across the first 15 years of the data. To confirm that we saw a strong relationship between the two variables we decided to calculate the correlation coefficient as well. When we did this, we found that the correlation coefficient between the two variables was 0.851. This indicates quite a strong relationship existing between CO2 emissions and storm frequencies. As global CO2 emissions have risen throughout the years, the frequency of storms has increased within the US at a very similar rate.

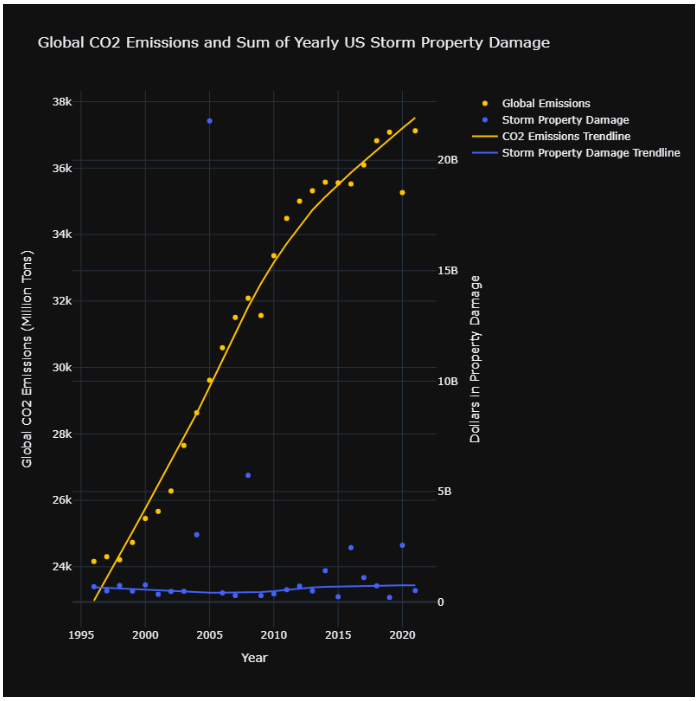


To further analyze the relationship present within this data, we also wanted to break down the storms by type to see which types of storms were more or less affected. Some storm types that we looked at were floods, hurricanes, tropical storms, excessive heats, droughts, and blizzards. To compare the effect of CO2 emissions on these storm types, we graphed a heatmap of their correlation coefficients. In the heatmap to the right, you can see that blizzards and excessive heat appear to have the highest correlations, while hurricanes actually appear to have a negative correlation. It should be noted that while the colors of the heatmap are noticeably different, the actual magnitudes of correlation coefficient are not that different. Overall the magnitude ranges from 0.37 to 0.53, with four of them falling between 0.42 and 0.47. It appears that each individual storm type with the exception of hurricanes, have mild correlations to CO2 emissions.

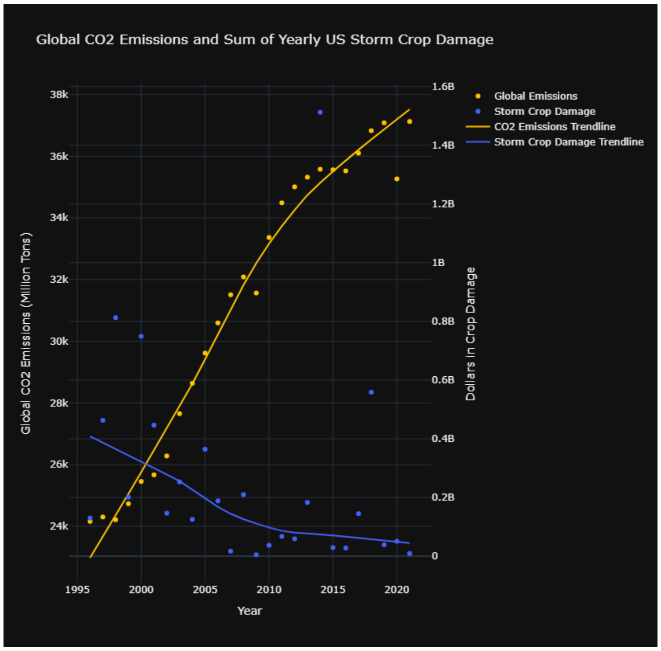


Our final thought for this particular question was to create a similar multi-axes graph as before, but only looking at blizzards and excessive heat. We decided on these two storm types since they appeared to be the most correlated with emissions. By doing this we produced the graph to the right. Overall, this is what we expected to see. Both types of storms have a notable upward trend, but this trend is not nearly as strong as the CO2 emissions trend. As CO2 emissions continue to increase, so do the frequencies of blizzards and excessive heat. This is a very interesting finding because they represent very opposite weather types, but both appear to be happening more often. This graph shows that CO2 emissions can potentially lead to not only more extreme warm weather events, but also cold weather events.

How does increase in storms affect overall damage caused by storms to property? To crops?

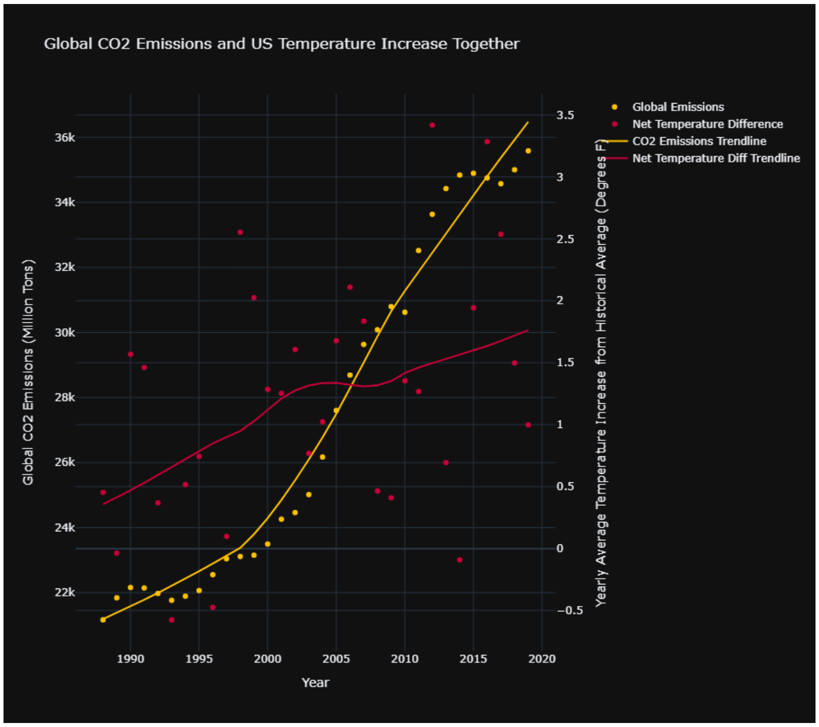


In the previous question we showed an increase in storm frequency throughout the US and how it relates to global CO2 emissions. Within our storm dataset, we also have information containing the dollar amount of damage to both property and crops. So, we are curious whether or not the increase in storm frequency has also brought along an increase in property and crop damage. To assess this we created similar multi-axis plots to showcase both CO2 emissions and the damage variables over time. We began by first looking into property damage, seen to the right. Once graphed, we found that there was little to no correlation between the two. Additionally, we found that the increasing number of storms appear to have almost no effect on the amount of property damage. However, this result could be somewhat skewed by outliers. Particularly, the 2005 datapoint on this graph is nearly 4 times larger than the next closest data point. This year happens to be the year that Hurricane Katrina, the most damaging and deadliest modern US hurricane occurred. Plotting the same data excluding 2005 does result in a similar graph with a mostly flat trendline. For this reason, it appears that there is no relationship between the variables. Additionally, it appears that property damage is more impacted by the severity and intensity of the storm rather than the number of them, which makes sense.



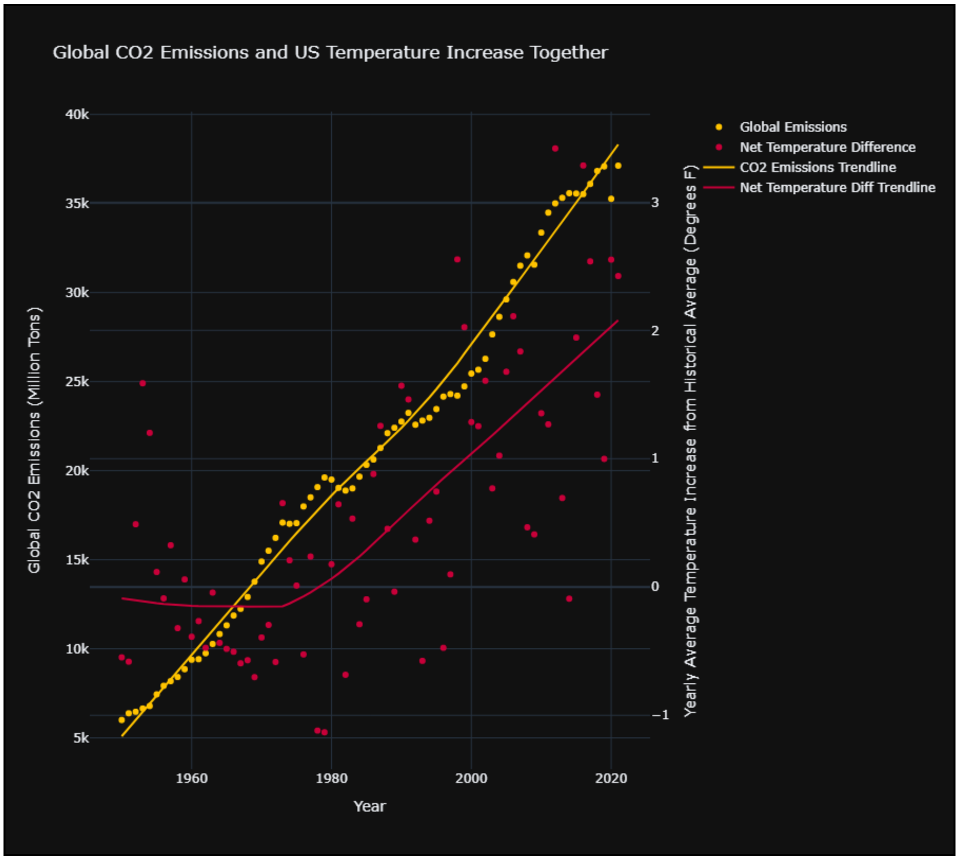
Performing a similar analysis, but instead focused on crops yielded very similar results. Furthermore, crop damage actually had a negatively sloped trendline indicating that even with increasing storm frequency, crops are being damaged less. Again this data is susceptible to outliers as 2014 is notably higher than the other data points this time. Overall, this trend could be for a number of reasons such as better farming techniques, better protection for crops, or even just luck. From a statistical perspective however, it seems there has been little to no effect on the amount of damage done to property or crops even with rising storm frequencies.

How does temperature change over time and how does it relate to carbon emissions?



To develop an understanding of how temperature changes with global CO2 emissions we again used the multi-axis scatterplot. Our dataset on temperature contained a column that calculated the net difference of the average monthly temperature compared to that month's historical average temperature. Using that column and plotting it against time with the CO2 emissions we were able to create the graph to the left. Overall this graph displays an upward trend in temperature differences, but this trend does not seem to follow CO2 emissions that closely. Additionally, looking further into the data we confirm this idea since the correlation coefficient between the two variables was 0.391. One of the problems with this graph was how limited the scope of time was. To counteract this, we decided to use our other emissions dataset, which was slightly less detailed but contained data ranging much further back than 1990.

In doing this, we created the figure seen on the next page. It is a graph with the same setup as the previous one, but the data goes back all the way to 1950. When presented with this level of data, the relationship between rising temperatures and CO2 emissions becomes visibly stronger. This is also backed up by the increase in correlation coefficient to 0.641. Also, when analyzing the graph it appears that after 1973, both trendlines follow an almost identical upward trend. Looking further into this idea, we believe that this drastic switch in the temperature trendline could indicate the time that human emissions surpassed the Earth's natural carbon sink levels. This appears to occur around 17 billion tons of yearly emissions. Further research online shows that we estimated that the Earth absorbed 50% of all emissions in 2012, which would account for approximately 17.5 billion tons of CO2. This would support our hypothesis that global emissions reached a "tipping point" in 1973, where they overtook the Earth's natural ability to absorb CO2. This is what has led to the idea of climate change and has caused many of the issues related to greenhouse gas emissions and rising temperatures today.



Conclusion

We asked many questions over the course of this project, but they can best be summarized by the following: What is the distribution of CO2 emissions by continent/region? What is the correlation between different fossil fuels and other factors with co2 emissions? What countries have the highest CO2 emissions and have they changed over recent years? How has CO2 emissions changed in the top 5 countries and as GDP increases or decreases does that impact CO2 emissions? What are the most consumed energy types and changes in energy consumption patterns in the United States?What is the Energy Intensity Per Capita and by GDP of Top 5 Energy Consumers? What is the trend between storms and emissions over time? What storms are and were most affected by the increasing global emissions? Are storms becoming more damaging over time as measured by property and crop damage? And how is temperature change linked to CO2 emissions over time?

We found strong correlations between CO2 emissions and fossil fuels as well as other factors like GDP, Population, Energy Consumption, and Energy Production. These different variables all contribute to CO2 emissions and we saw trends like when one increases the other would increase as well showing positive correlation. The countries with the highest emissions were countries that industrialized more and used a lot more energy which made sense to see them on the top 10 countries list. We also noticed the correlation between CO2 emissions and GDP so when we noticed how China's GDP increased significantly so did their CO2 emissions and took the place for first from the United States. Thus, we can state there are many factors that contribute to CO2 emissions and emissions are continuously going up.

We noticed intriguing patterns in CO2 emissions and their relation to countries. In the period from 1988 to 2019, Europe had the highest median and interquartile range (IQR) for CO2 emissions, but North America had some countries with higher emissions than Europe. After 1990, however, Asia surpassed all other regions with the highest median and IQR of CO2 emissions, including high emission outliers. The United States and China are the largest global energy consumers due to their status as the largest economies. The top 5 consumers of energy, which were further analyzed, showed a decrease in Energy Intensity by GDP, but either stagnant or increased Energy intensity per capita. This could be due to factors such as improved technology and equipment, as well as consumer behavior. The United States primarily consumes petroleum and natural gas, while China primarily consumes coal. After 2010, the US remained the major consumer of nuclear energy, but China overtook the US to become the primary consumer of renewables.

We found a strong correlation between the frequency of storms and CO2 emissions, as evidenced by our visuals in the previous section on Dash. To reiterate: both were rising at similar rates with an r^2 coefficient of 0.851. As for which types of storms seemed most sensitive to these changes, those appeared to be blizzards and excessive heat waves, which have become both more frequent and more damaging over the last. We can say confidently then that CO2 emissions are correlated with (and potentially a causal factor) more serious weather events, specifically those tied to temperature extremes like the aforementioned blizzards and heat waves.

For the first of our latter questions, our findings aren't as solid; there seems to be either no relationship, or a very weak correlation between storm impacts and emissions, broadly speaking. There is however, an indirect link as the storm intensity is a good predictor of its damage.

Regarding temperature changes and emissions, we found that the strength of the correlation over time depended on how far back we looked. When our dataset was capped at the past 30 years, our r^2 is only 0.391; however, when expanded to include the past 70 years, it jumps to 0.641. This indicates a long-term relationship between temperatures and CO2, and if we went back even further, the relationship may be even stronger.

Next Steps

We found many interesting patterns and connections while analyzing this dataset, but there is still more work to be done, specifically in regards to predictive analysis and identifying future trends and events. We would recommend future groups focus on refining a stronger model using either more robust data, more complex models, or a combination of datasets and models in order to identify how our findings on historical data may impact the future, as the impact of this work and any possible findings cannot be overstated.

The Countries CO2 Emission dataset has two glaring limitations. Firstly, it only covers energy consumption and production up to 2019 and does not account for the impact of events like the COVID-19 pandemic. Secondly, the data is presented on a country level and would benefit from being further broken down by state.