Global Carbon Emissions

Pragna Goura
1001721805

1. Introduction

This project aims to analyze the total output of carbon emissions by each country since the 18th century. Global pollution is a hot topic in our world and with this dataset, we can analyze what fossil fuel and what country contributes the most to the carbon pollution in our air today.

Challenges of using this dataset:

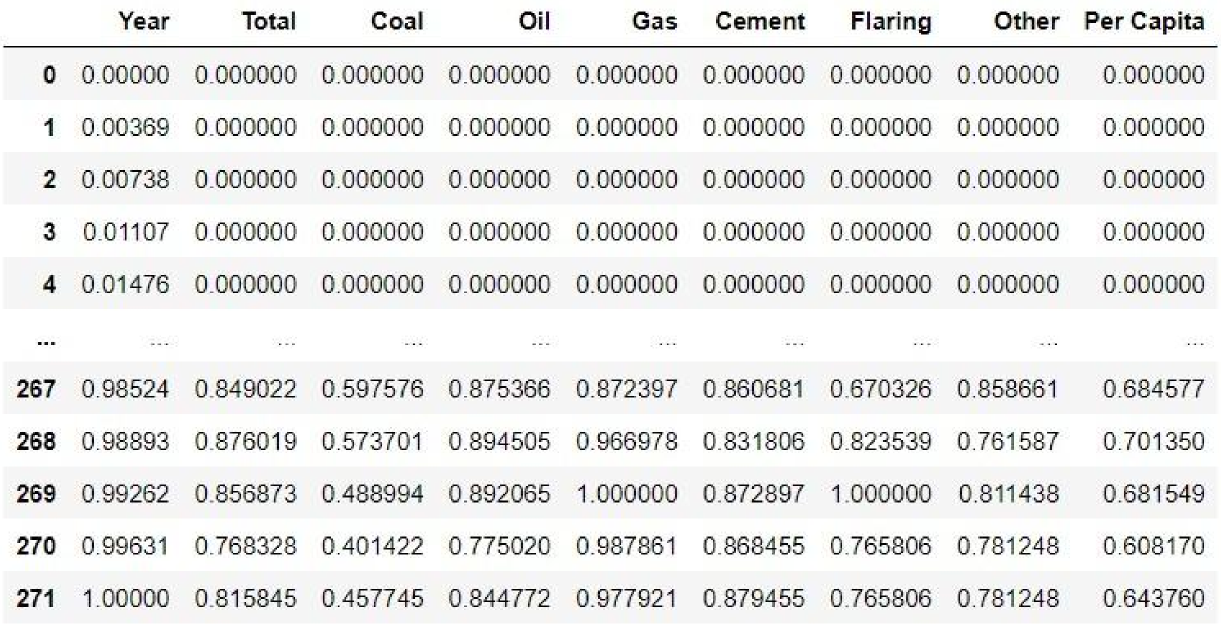
a) Over time, as technology developed, so did the increase of by-product that is CO2. We see a general trend in the emissions rate but there are some parts of history where the use of a specific fossil fuel was more or less. We have some outliers in this dataset, but they are kept due to the nature of time and how the global economy and political acts have prevented some fossil fuels and helped increase others.

2. The Dataset

The dataset came with 11 columns and around 63000 rows. Every 272 rows are the number of years that document the increase of fossil fuels in every country. The fossil fuels in this set are coal, oil, gas, cement, flaring, and others.

3. Data Preprocessing

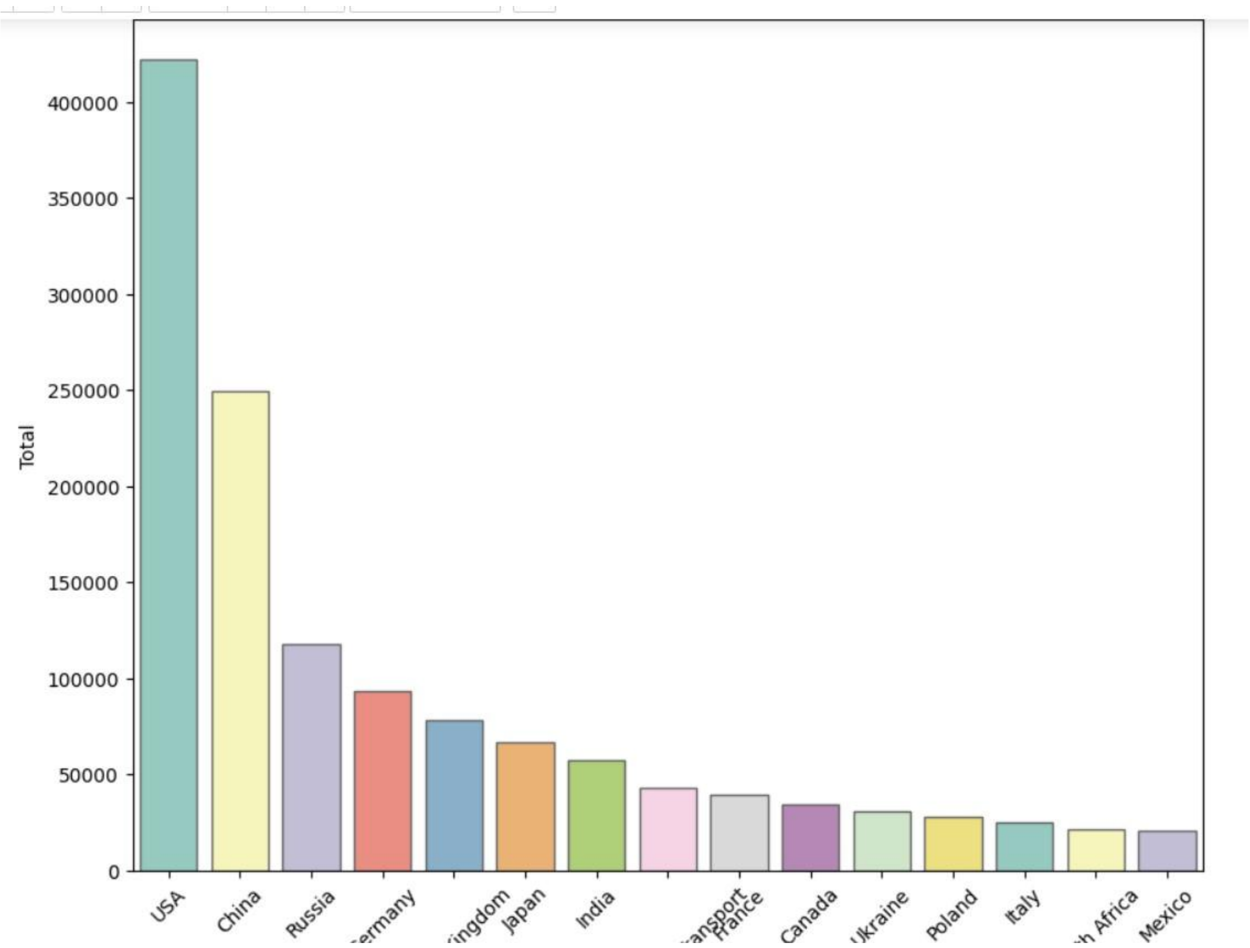
The data came pretty organized and with some NaN values. When looking at where these NaN values were placed in the dataset, I could infer that the reason for no values is that there were simply no carbon emissions produced. The dataset takes us back to the 1750's almost a century before the industrial revolution. Even if the technology was created, it was not in use the way it was in the 1850s. Once that was done, I made sure to encode and normalize my data. There were no classification variables in my dataset for the machine to encode, so I only focused on normalizing my data frame from 0-1. I did this for both the entire dataset and that dataset that only holds the US statistics.



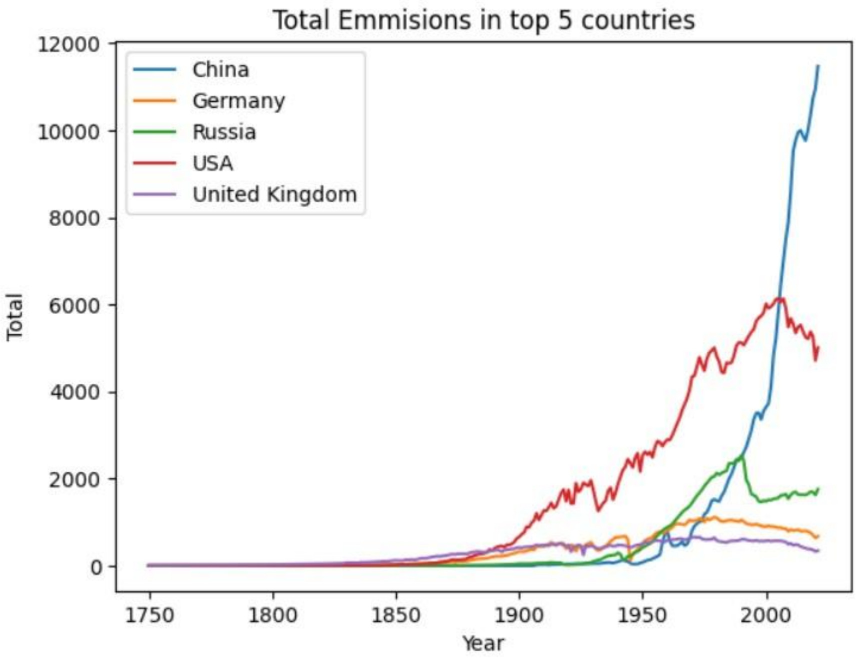
	Year	Total	Coal	Oil	Gas	Cement	Flaring	Other	Per Capita
0	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	0.00369	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.00738	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.01107	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.01476	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...
267	0.98524	0.849022	0.597576	0.875366	0.872397	0.860681	0.670326	0.858661	0.684577
268	0.98893	0.876019	0.573701	0.894505	0.966978	0.831806	0.823539	0.761587	0.701350
269	0.99262	0.856873	0.488994	0.892065	1.000000	0.872897	1.000000	0.811438	0.681549
270	0.99631	0.768328	0.401422	0.775020	0.987861	0.868455	0.765806	0.781248	0.608170
271	1.00000	0.815845	0.457745	0.844772	0.977921	0.879455	0.765806	0.781248	0.643760

4. EDA

The first thing I wanted to analyze was what country was the highest contributor to the carbon emissions and after graphing, we see that the US outweighs the other countries, about a ⅓ more than China throughout modern world history.

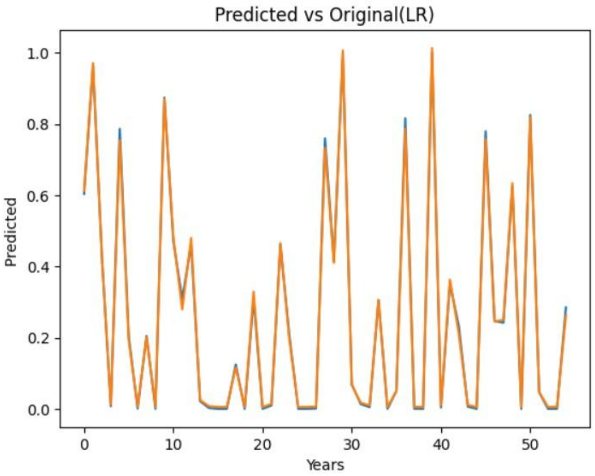


But then I decided to see what the spread of carbon emission looked like over time, so I plotted a line graph of the top five countries.

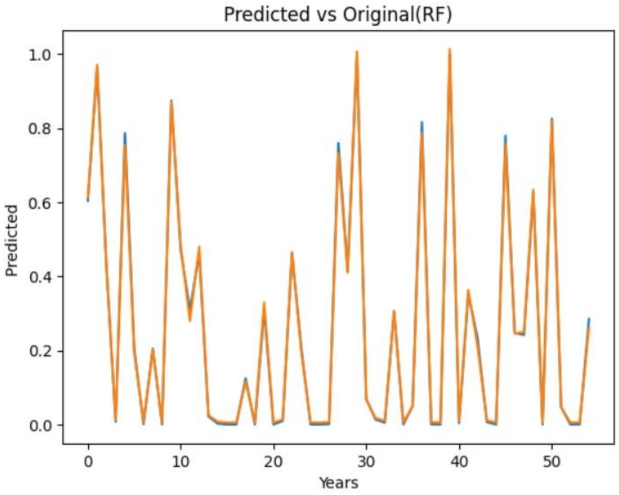


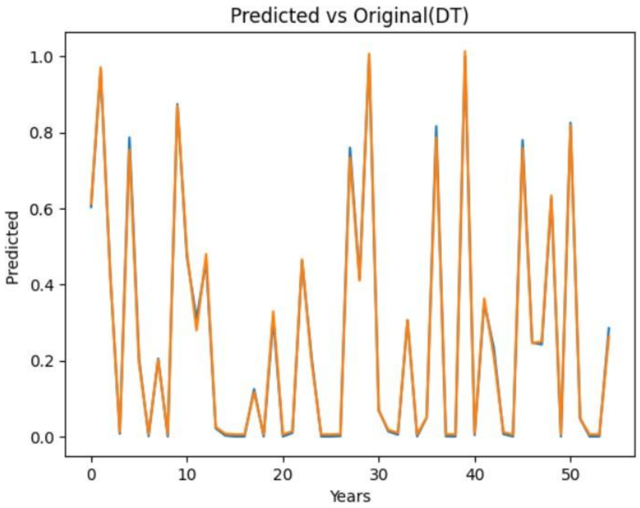
As we see here, in the current day, China is the country that contributes to most of the carbon emissions at the present time and the US has added to the pollution in a steady way over time.

5. The Models



I first decided to go with linear regression as it was one of the simpler regressive models. After my first attempt at a linear regression model, the r2 score was a perfect score. Although this is something that is wanted, it is not realistic. I try importing Ridge which is another linear regression model that prevents overfitting. For this, the r2 score was 0.99 which isn't too much of a difference.





For the other two models, the r2 scores were also closer to 1.0, the Random Forest was 0.998 and the Decision Tree was 0.996 but for these two models, the RSME is closer to zero as well.