Economic Power and its effects on CO2 and Pollution around the globe

Github link- https://github.com/dhand33p/USCDSCI510

Project Description: For this final project, I decided to create a CSV after web scraping three websites. Each website pertained to different data, specifically about the growth of domestic product per capita(how much a person makes), carbon dioxide emissions, and pollution levels for each respective country. After collecting all this data, I joined the three datasets and created one CSV regarding each country's data.

Motivation: I was interested in learning the effects of high economic status in different countries and how they are affected by pollution and other factors. One might concur that more developed countries have more CO2 emissions, so I analyzed this statement. By understanding how countries are affected around the globe by these CO2 emissions as a result of economic power(especially understating how pollution levels of high CO2 Emission countries vary).

Data Sources:

1. CO2 Emissions by Country - Worldometer
2. World's Most Polluted Countries in 2021 - PM2.5 Ranking | IQAir
3. GDP per Capita - Worldometer

For each data source, I used the requests and beautiful soup library to extract the data from the websites. Specifically, by analyzing the HTML code of each website and finding the table element in each one(although it was challenging to find a Table element for one of the websites) and collected the data. Once the data had been collected, I added each country's information to a list; however, the data needed to be cleaned. I cleaned the data using regex expressions and extracted the relevant information from the list. For the first website, I only extracted CO2 emissions from 2016 and world percentage share to find leaders in CO2 emissions. For the second website, I selected the pollution level from 2021(a higher level means worse pollution). For the third, I selected the GDP per Capita(Purchasing power parity), adjusted for inflation.

Afterward, I created three different dictionaries. The first website would have a dictionary pertaining to Country and CO2, the second would have Country and Pollution, and the third would have country and GDP. Once the dictionary was created, I converted them to a pandas data frame, which resulted in having a total of three different data frames (shown below).

	Country	CO2_emis_2016
0	China	29.18
1	United States	14.02
2	India	7.09
3	Russia	4.65
4	Japan	3.47

	Country	Poll_2021
0	Chad	75.9
1	Pakistan	66.8
2	Tajikistan	59.4
3	India	58.1
4	Oman	53.9

	Country	GDP_Per_Cap
0	Qatar	128647.0
1	Macao	115367.0
2	Luxembourg	107641.0
3	Singapore	94105.0
4	Brunei	79003.0

Finally, to generate the CSV file(shown below), I joined all three data frames, which resulted in about 97 different rows; before the join, each data frame had roughly 115 rows(data lost due to missing information and join issues).

	Country	CO2_emis_2016	Poll_2021	GDP_Per_Cap
0	China	29.18	32.6	16842.0
1	United States	14.02	10.3	59928.0
2	India	7.09	58.1	7166.0
3	Russia	4.65	12.3	25763.0
4	Japan	3.47	9.1	42067.0

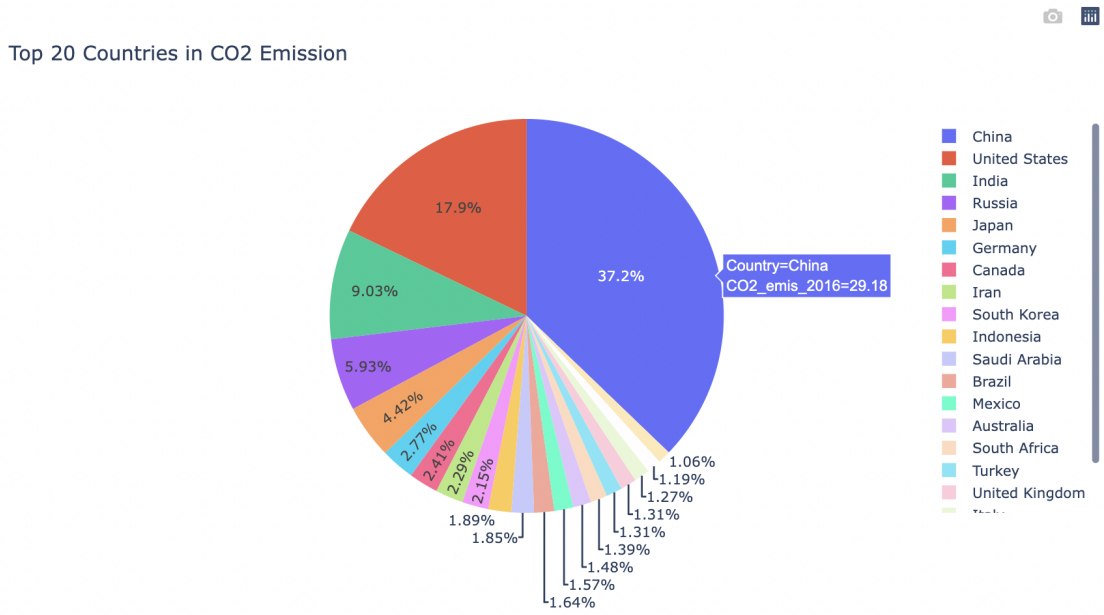
Data Analysis: Once I had collected my data and created the CSV file, I checked any correlation and relationships between the columns. Specifically, this would be the first step to making my visualizations; here, I found some interesting insights. There was little correlation between pollution and CO2 emissions (using the pandas' correlation matrix function). Similarly to the previous correlation, CO2 and GDP per capita provided little correlation, which was surprising as one would expect a country with higher GDP to emit more CO2 emissions. Using the same function, I found the most correlation between GDP and pollution levels, which provided the reasoning for higher economic status in those countries, allowing cars and other machinery to emit more polluting factors.

	CO2_emis_2016	Poll_2021
CO2_emis_2016	1.000000	0.053336
Poll_2021	0.053336	1.000000

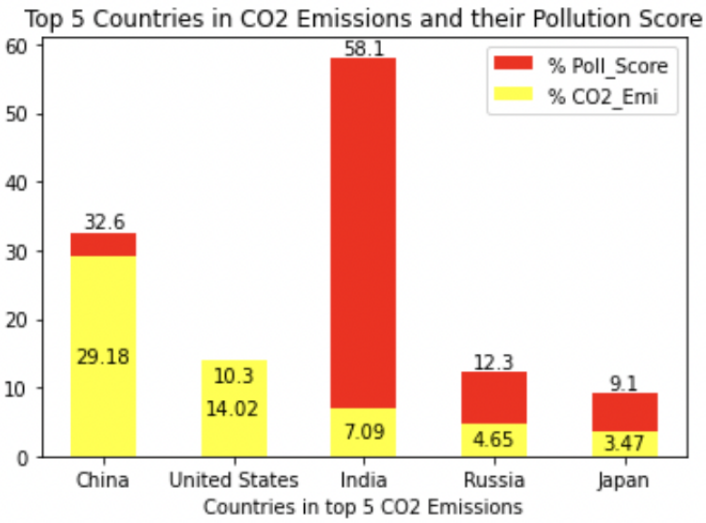
	GDP_Per_Cap	Poll_2021
GDP_Per_Cap	1.000000	-0.322456
Poll_2021	-0.322456	1.000000

	CO2_emis_2016	GDP_Per_Cap
CO2_emis_2016	1.000000	0.040108
GDP_Per_Cap	0.040108	1.000000

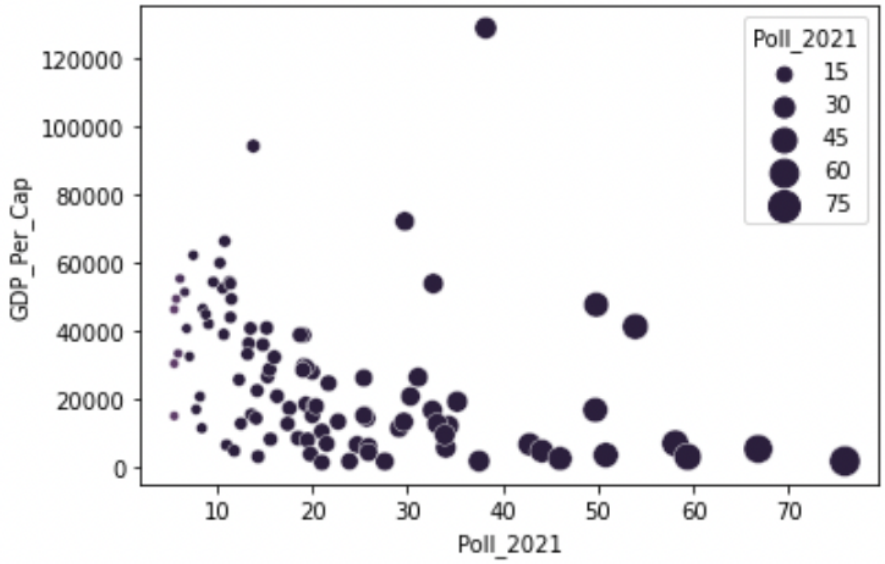
The pie chart showed some interesting findings; the top two counties in CO2 emission make up more than half of the pie chart. Here we see the impact of developed countries around the globe.



To test how the global leaders in CO2 emissions are affected(if at all) by their CO2 emissions regarding pollution. Besides India, the remaining four are significantly impacted by their CO2 pollution. Through these findings, the rest of the countries are facing more extremities in global warming due to their emissions(location near the sea, more extreme weather conditions, etc.).



This scatterplot looks at the relationship between REAL GDP PER CAPITA between all countries, explicitly comparing their pollution levels. The scatter plot represents the correlation between GDP and pollution; here, we can see how higher GDP tends to have lower pollution levels.



Findings: Through this analysis, some insightful trends were found. By understanding the global leaders of CO2 emissions, those countries are not significantly affected by their CO2 emissions concerning pollution. Also, the Pearson coefficient showed a slight -0.322 between GDP and pollution level. This does support the fact that a country's pollution level is affected by its economic status(countries with higher GDP have lower pollution levels). This person coefficient and scatter plot allow us to provide evidence that the global leaders in CO2 are affecting other developing countries concerning pollution besides themselves.