COVID-19 DATA ANALSIS

RUCHI BHAVSAR

ABSTRACT:

COVID-19, a flu that started with affecting a few countries has now spread worldwide and was declared a pandemic by WHO (World Heath Organization) on March 2020 and since has been identified as a threat to the entire world population. Various studies are being conducted by researchers and scientists of all countries to find a cure for the pandemic. This work represents a dataset which is focused on COVID-19 at a global level and also provides analysis and spread of the disease. Multiple factors play a role for the infection to spread, therefore, extensive research is being carried to better fight the disease. The dataset used in this project was provided by Johns Hopkins University Center of System Science and Engineering. They aggregate the data from WHO, national and regional public health and have made it available on a public GitHub.

Since the data we receive is never in the desired format, I have processed the data, performed data cleaning and, aggregation along with visualization.

INTROUCTION:

The unprecedented speed of Covid-19's transmission is the main concern among everyone. The origin of this virus has been traced back to a wet market in Wuhan, China. From there, it has reached every corner of the world in no time. Doctors and researchers are learning new things about SARS-CoV2 everyday, which is making it difficult to prevent. Since, they cannot predict what might happen in the near future, projects like these come into picture.

After the data cleaning and processing the dataset given, the 2 datasets are merged and aggregated for better extraction of data. When merging the 2 datasets, we create a DataFrame to represent the data in a table format. Other factors, such as climate and population are also taken into consideration when plotting the graph.

DATA ACQUISITION:

Johns Hopkins University Center of System Science and Engineering(JHU CSE) provided the data to the public. They receive the data from various sources such as the WHO, regional and public health institution. The data is available to everyone and is updated daily for better prediction of models. The dataset had following columns:

1. Date - Daily covid cases for each country
2. Country/Region - All the countries with covid cases
3. Province/State - All the provinces with covid cases
4. Confirmed - confirmed covid cases
5. Recovered - People who recovered from covid
6. Deaths - Number of people who died due to covid

This was accompanied by another two datasets, ‘worldpopulation.json’ and ‘climate.json’. These two datasets were helpful to check how climate plays a role in affecting the spread of Covid-19.
The columns in ‘world-population.json’:

1. Rank - Rank of the country based population size
2. Country - Name of the country
3. Population - Total population of the country
4. World - Percentage of the population in terms of total world population

The columns in ‘climate.json’:

1. high - highest recorded temperature every month
2. low - lowest recorded temperature per month
3. dryDays - Number of days without rainfall
4. snowDays - Number of days of snowfall
5. rainfall - Number of days of rainfall

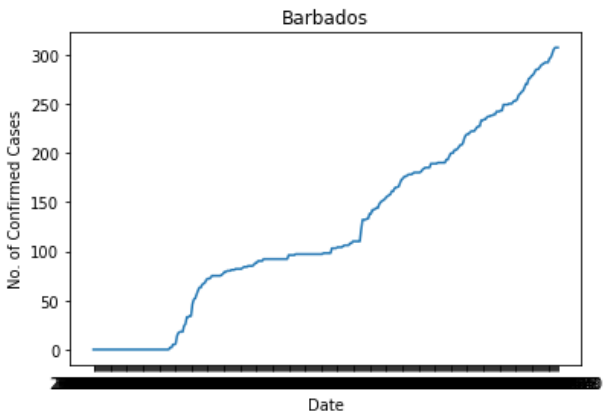
Data Processing and Cleaning:

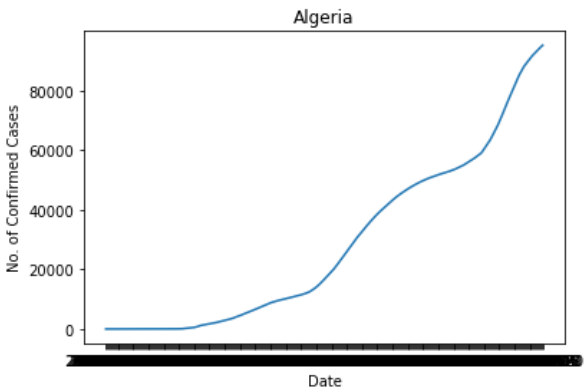
For data processing and cleaning, unwanted columns were dropped.
Eg:
df.drop("Province/State", axis=1, inplace=True)

DATA ANALYSIS:

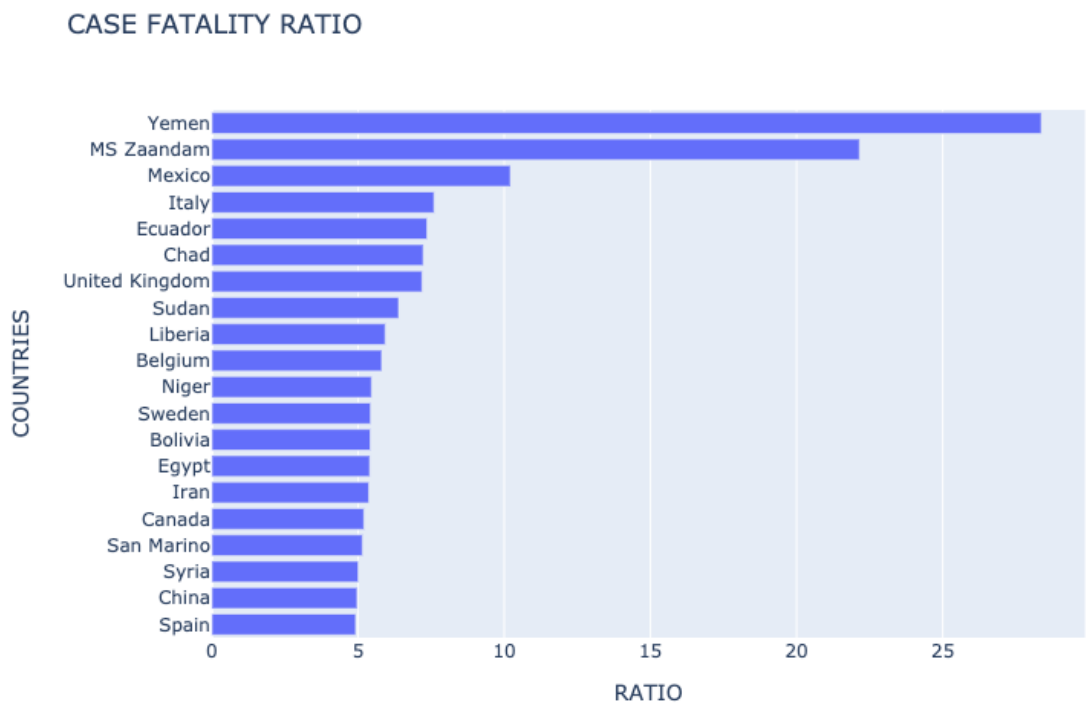
To plot the graphs between the data sets, matplotlib was used. We first plotted the number of confirmed cases over the time. The graphs that were plotted in this project were:

1. **Number of confirmed cases over time for each country**





2. **A bar plot that shows number of deaths over 100 confirmed cases**

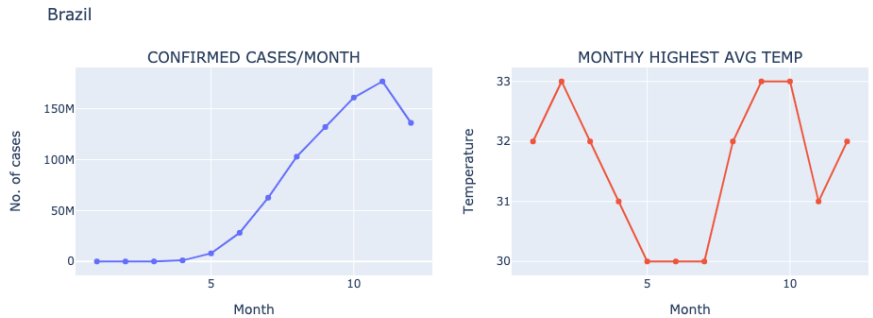


3. **Ratio for top 10 countries with highest number of confirmed cases per capita**

TOP 10 COUNTRIES WITH HIGHEST CONFIRMED CASES PER CAPITA

	country	ratio
0	Qatar	10.67
1	Andorra	8.69
2	Bahrain	8.40
3	San Marino	6.89
4	Holy See	5.16
5	Luxembourg	5.01
6	Panama	5.00
7	Chile	4.69
8	Kuwait	4.68
9	Israel	4.46

4. **Graph for monthly number of confirmed cases with the average monthly temperature**



2020

5. Research question related to COVID-19:

Even though we have considered the climatic changes in a country that might affect the growth of the virus, we haven't taken into consideration how the living conditions in the country, economic stability and overall health of the citizens. For countries that are underdeveloped or developing, the resources to fight COVID-19 are not readily available and therefore can lead to mass spreading of the virus in public areas. How will these factors play a role in the spread of SARS-CoV-2?

CONCLUSION:

The biggest uncertainty for the future of COVID-19 is how weather will affect it's transmission dynamics. Some preliminary laboratory trials suggest that the virus maybe particularly sensitive to weather and with the rising temperatures could reduce transmission rates. It is still uncertain whether the winter weather will increase the risks, especially in underdeveloped countries.

Projections suggest that, COVID-19 will decrease during the summer, rebound by spring and rise again next winter. However, this remains uncertain and the probability of the cases doubling weekly is >20% in summer. To conclude, it cannot be predicted when the risk for COVID-19 will end but with the methods learnt in this project we can try to flatten the curve as much as possible.