

Project Description:

I developed a project that involves data cleaning, processing, and visualization using Python, leveraging libraries such as Pandas, Matplotlib, and Seaborn. The dataset, sourced from Kaggle, contains COVID-19 case and death statistics from 2020 to 2024. The purpose of this project is to enhance my Python programming skills, particularly in data manipulation and visualization, which will better prepare me for future roles in data analysis, data science, and related fields.

Purpose:

This project was aimed at improving my proficiency in Python, focusing on data handling and visual storytelling through visualization tools. By applying these skills to a real-world dataset, I aim to solidify my foundation in Python, positioning myself for future roles that require data analysis, processing, and insightful data-driven decisions.

Case Study:

Title: Understanding COVID-19 Trends Through Data Visualization (2020-2024)

Objective: To analyse and visualize the progression of COVID-19 cases and deaths from 2020 to 2024 using Python, with the goal of identifying key trends, peaks, and anomalies in the data. This case study will focus on developing skills in data cleaning, processing, and visualization, and explore how these insights can assist in public health decisions.

Background: COVID-19 has had a profound impact globally, and tracking its spread over time has been crucial for governments and health organizations. This study utilizes real-world COVID-19 data from Kaggle, spanning four years, to create a detailed overview of the pandemic's progression. The analysis focuses on cumulative case growth, daily new cases, and death rates.

Data Overview:

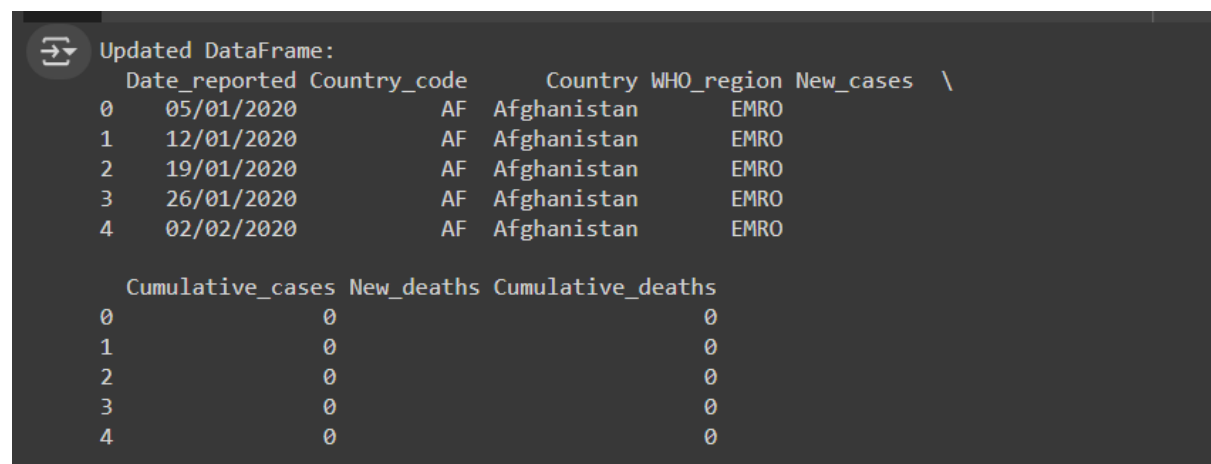
- **Dataset:** Kaggle COVID-19 cases and deaths from 2020 to 2024.
- **Fields:** Date reported, country, cumulative cases, cumulative deaths, new cases, and new deaths.
- **Data Cleaning:** Handled missing values, formatted date fields, and filtered for key insights.
- **Data Processing:** Aggregated data by month/year, performed groupings based on countries and time periods, and calculated critical metrics such as case fatality rate.

Data Cleaning

The Excel dataset on COVID-19, which contains over 50,000 rows, had all its columns, including "Date reported," "Country code," "Country," "WHO region," "New cases," "Cumulative cases," "New deaths," and "Cumulative deaths," merged into a single cell. This made it necessary to split the columns into separate fields for analysis. Additionally, the dataset contained multiple entries for the same country across different dates, which needed to be aggregated and counted as one data point to get a clearer overall picture.

Handling such a large volume of data and performing these transformations manually in Excel would have been extremely difficult, if not impossible, due to the sheer size and complexity. To address this, I utilized the Pandas library in Python, which allowed me to efficiently split the columns, as they were separated by semicolons (";"). This automated approach enabled me to clean and structure the data quickly, paving the way for further analysis and visualization while avoiding the pitfalls of manual processing.

After cleaning the dataset, I saved the updated file and re-imported it using Pandas' `pd.read_csv` prompt so I could continue with the data visualization phase. Fortunately, the dataset did not contain any missing or faulty values, meaning the data cleaning process was completed at this point, and I was ready to focus on analysing and visualizing the data.



	Date_reported	Country_code	Country	WHO_region	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
0	05/01/2020	AF	Afghanistan	EMRO	0	0	0	0
1	12/01/2020	AF	Afghanistan	EMRO	0	0	0	0
2	19/01/2020	AF	Afghanistan	EMRO	0	0	0	0
3	26/01/2020	AF	Afghanistan	EMRO	0	0	0	0
4	02/02/2020	AF	Afghanistan	EMRO	0	0	0	0

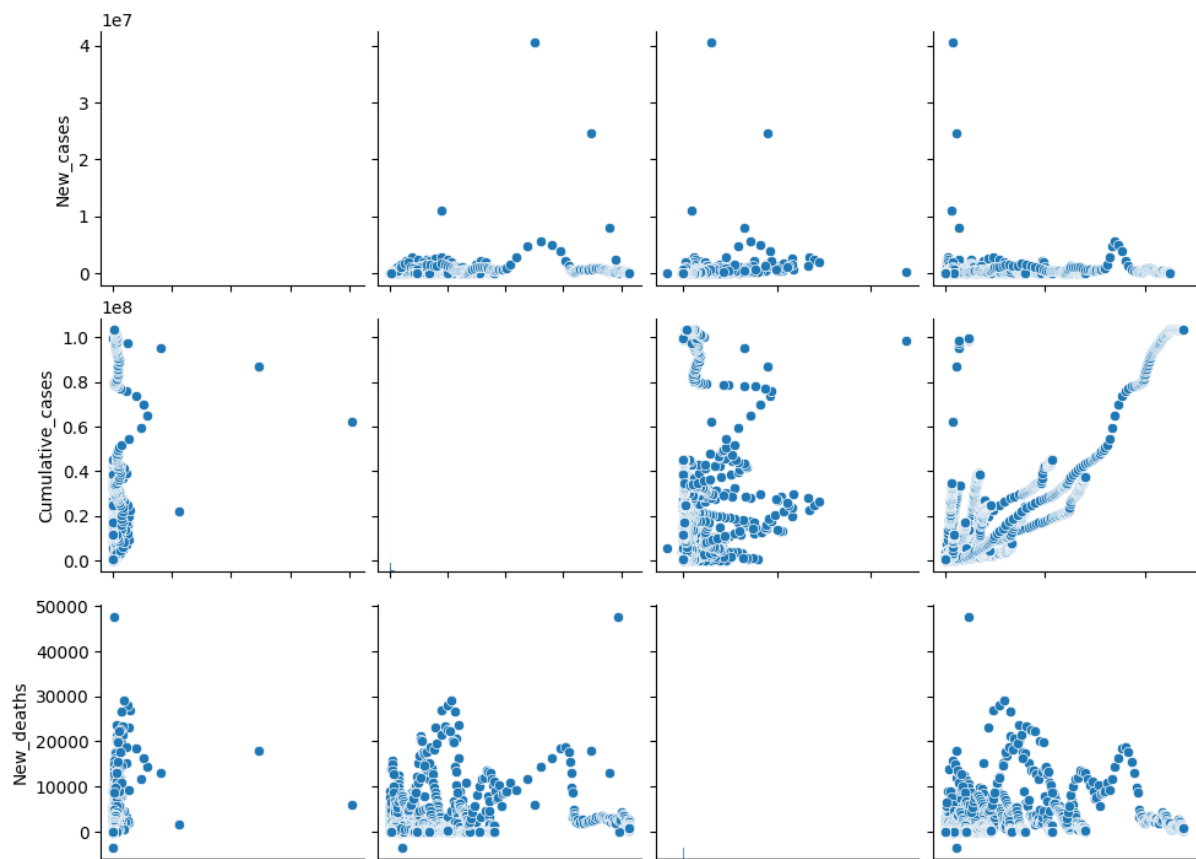
```
[12] df.to_csv('path_to_cleaned_dataset.csv', index=False)
```

STEP 2: Scatterplot Preparation

After completing the data cleaning process, we proceeded to convert four key columns—'new cases', 'Cumulative cases', 'New deaths', and 'Cumulative deaths'—from their initial data types (most likely strings or mixed types) into numeric formats.

Why: This step was essential because these columns represent numerical data, which is crucial for performing calculations and creating visualizations like correlation matrices and histograms. If left as strings, operations such as addition, correlation analysis, or plotting would not function properly.

Method: The `pd.to_numeric()` function was used to convert the values in each column to a numeric type. This method ensures that any non-numeric entries, such as strings or missing data, are either converted correctly or appropriately handled to avoid errors during analysis.



These are the visualizations generated after executing the code in the step 2

Inferences from these pair plots

1. New Cases vs. Cumulative Cases:

- There seems to be a somewhat positive trend between 'new cases' and 'Cumulative cases,' indicating that as new cases rise, cumulative cases

also increase. However, the data points are spread out, showing varying rates of new cases across different time periods or countries.

2. New Cases vs. New Deaths:

- There is no clear linear trend between 'new cases' and 'New deaths.' This suggests that higher new case counts do not directly correlate with higher death counts on a day-to-day basis, possibly due to varying healthcare responses and other factors.

3. Cumulative Cases vs. Cumulative Deaths:

- A clearer positive trend is observed between 'Cumulative cases' and 'Cumulative deaths.' This makes sense as more cumulative cases over time would lead to an increase in cumulative deaths, though the slope may vary due to recovery rates and other health interventions.

4. New Deaths vs. Cumulative Deaths:

- There is also no strong direct relationship between 'New deaths' and 'Cumulative deaths,' which may suggest that cumulative deaths grow steadily regardless of day-to-day fluctuations in new deaths.

Overall, the visualizations suggest that while there are relationships between cumulative cases and cumulative deaths, day-to-day new cases and deaths may fluctuate due to several factors, making it harder to discern clear trends at that level.

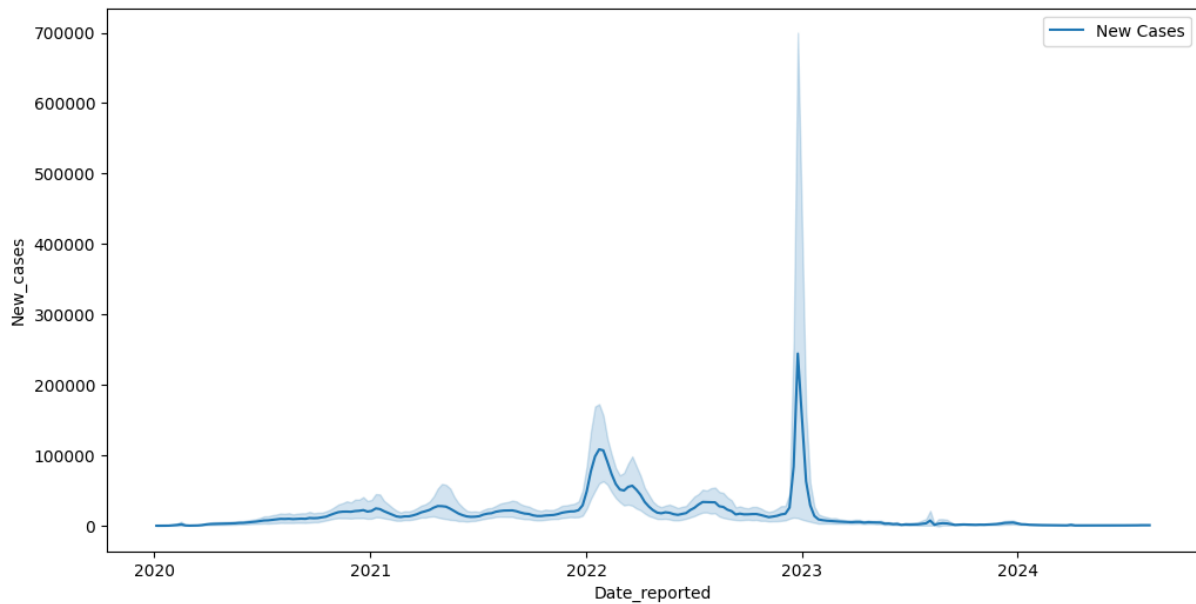
Step 3: Data Visualization using matplotlib & Seaborn

The scatterplots generated in the previous step lacked clarity and did not adequately represent the data. Therefore, it became necessary to explore more suitable methods for accurately visualizing the information.

Thus, we utilized Python's Matplotlib and Seaborn libraries to generate more detailed graphical representations from the dataset

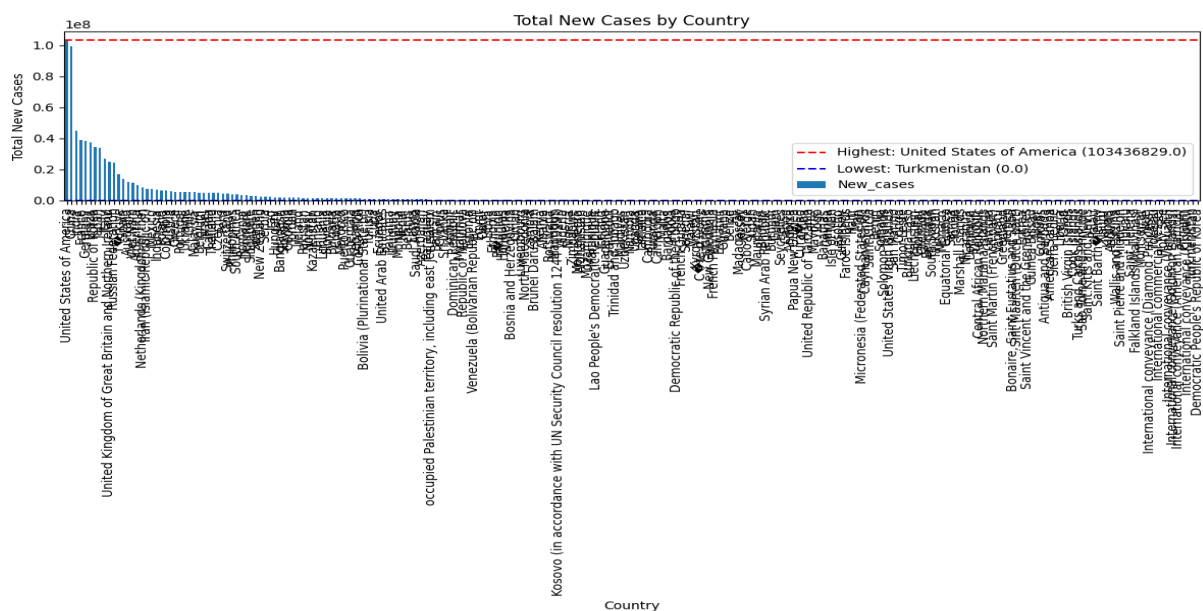
Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Seaborn is built on top of Matplotlib and provides a high-level interface for drawing attractive and informative statistical graphics.

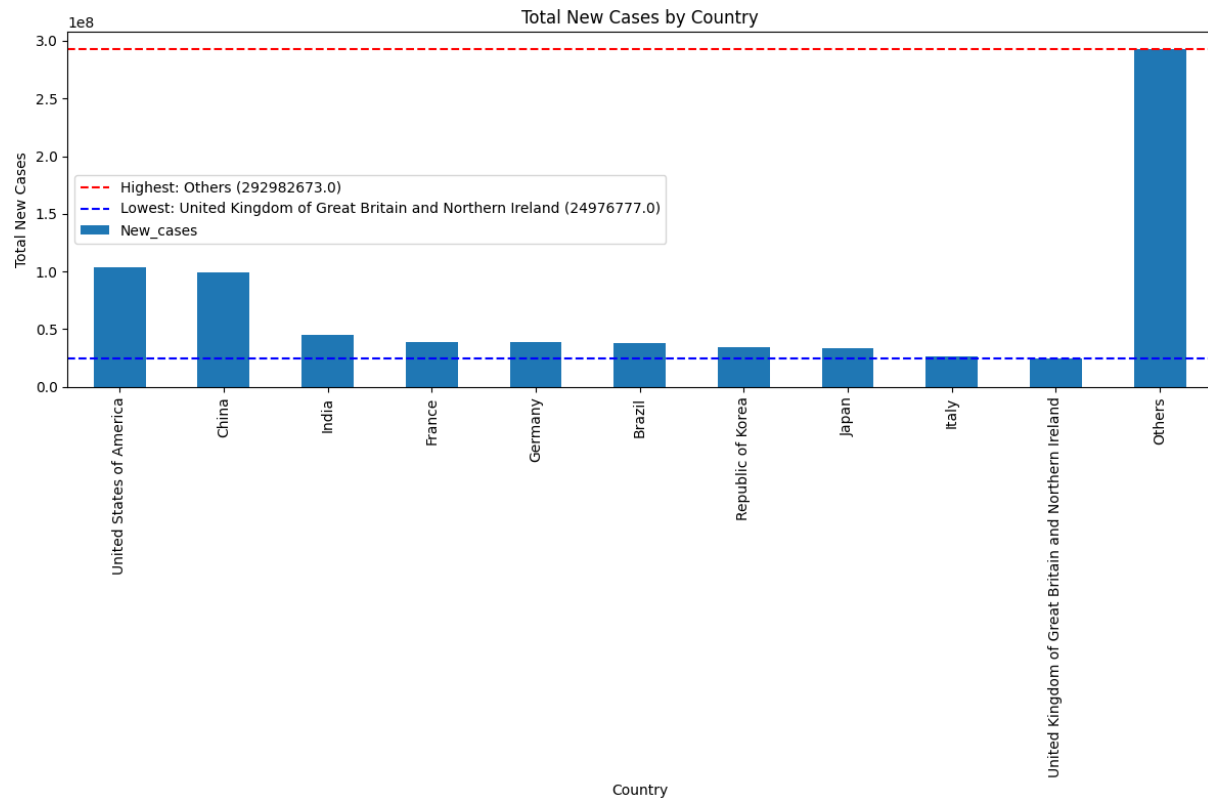


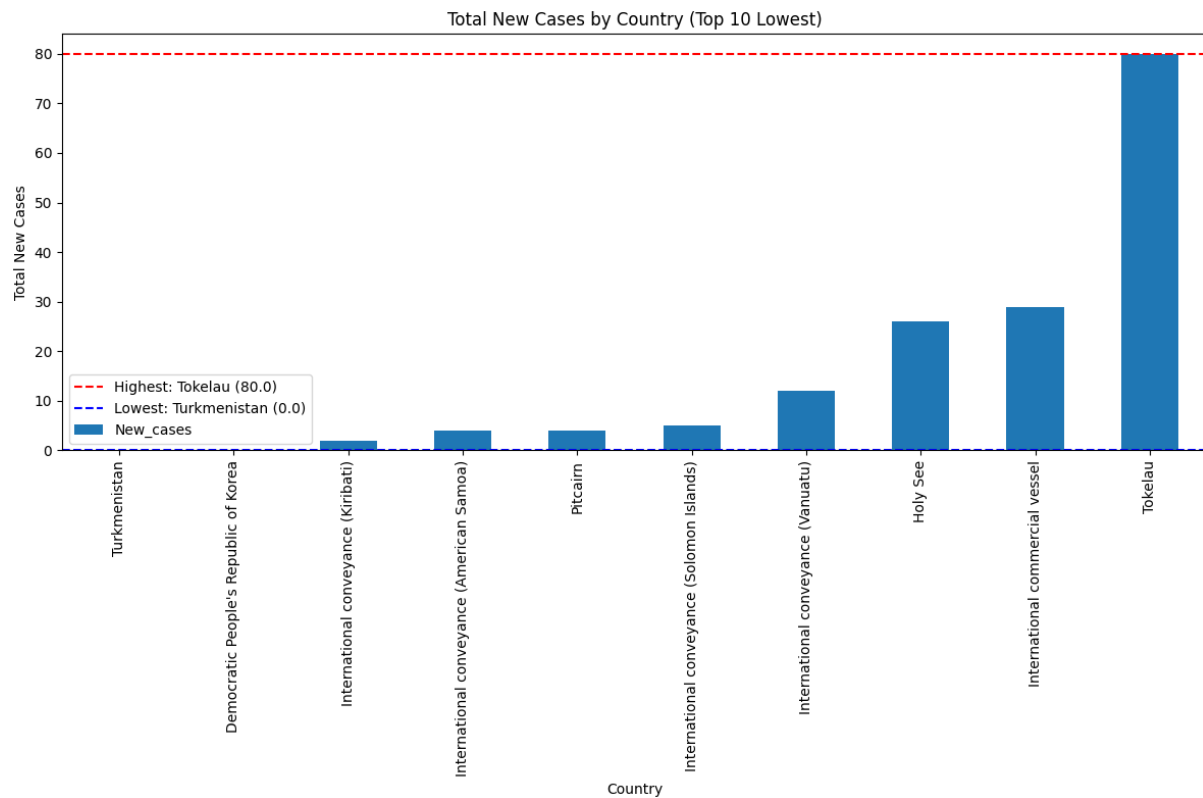
- The graph shows the trend of new COVID-19 cases reported between 2020 and 2024.
- The number of new cases remains relatively low throughout 2020 and 2021, with small peaks visible in the data.
- A sharp and significant peak occurs in early 2023, indicating a major spike in new cases. This spike quickly declines but marks the highest point in the dataset.
- Following the 2023 peak, the number of new cases drops sharply and remains at relatively low levels through the rest of the period.

Step 4: More Data Visualizations



To visualize the 'Aggregate new cases by country' data, we continued to employ Matplotlib. However, due to the dataset's extensive size and numerous countries, the resulting graph became overly cluttered. To enhance clarity and visual appeal, we divided the data into top 10 and bottom 10 countries, creating separate graphs. This approach ensured that key insights from the dataset remained accessible to readers.

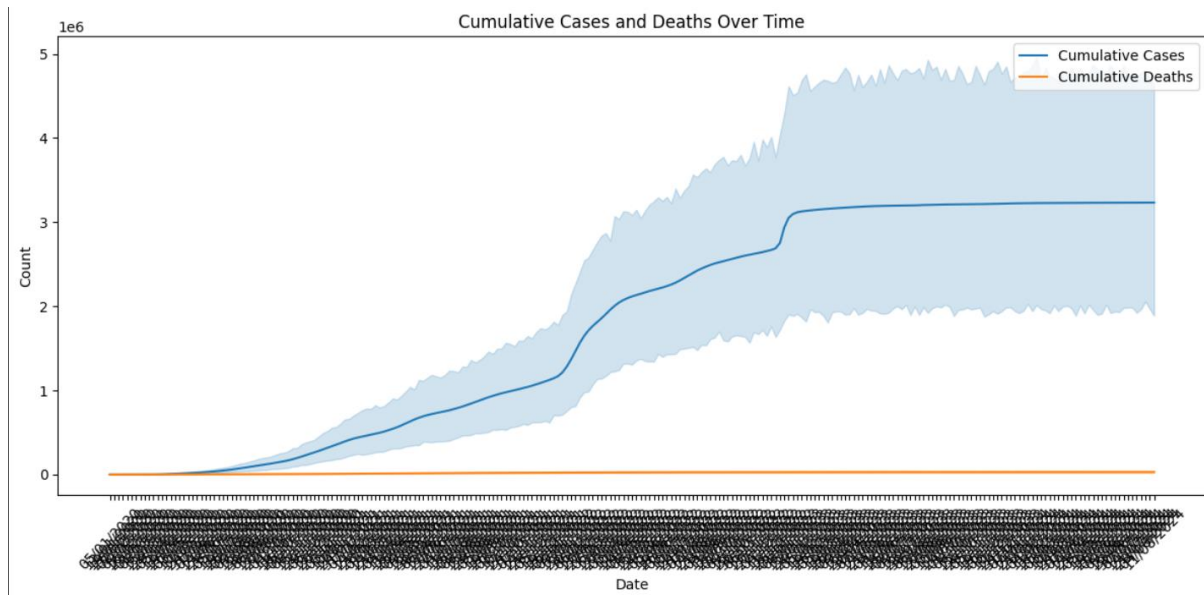




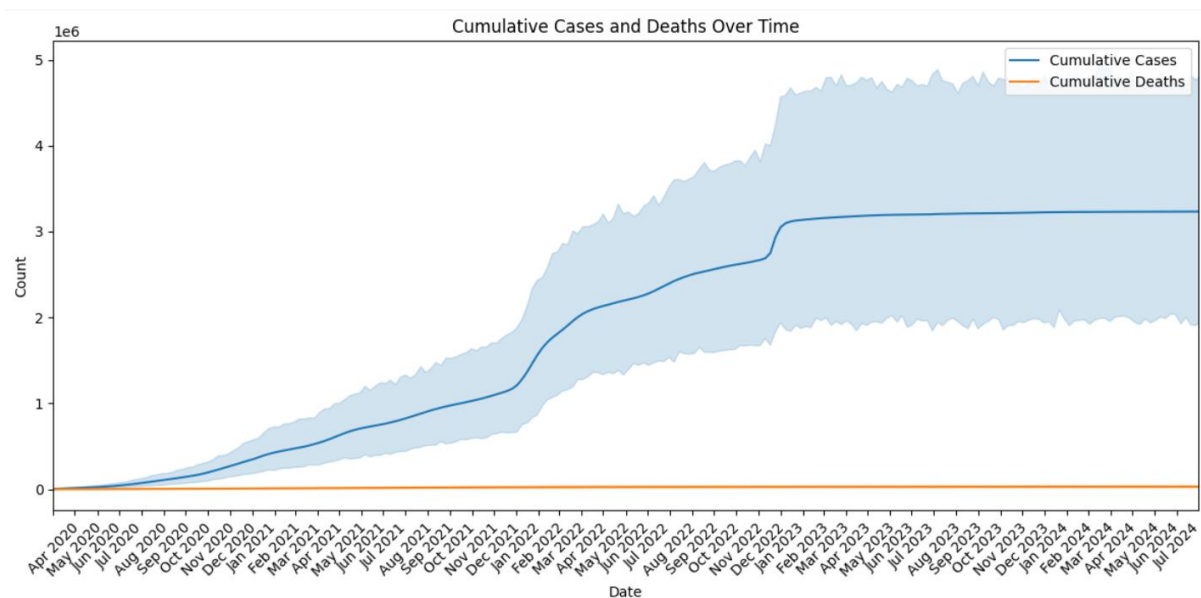
Key improvements:

- **Conciseness:** The sentence is more concise and direct.
- **Clarity:** The reason for the initial graph's issue and the solution are more clearly stated.
- **Focus:** The emphasis is on the goal of making the data more visually appealing and accessible.

Upon constructing a timeline of cumulative cases and deaths over time, a comprehensive graph was initially produced. Nevertheless, due to the extensive nature of the dataset, the x-axis labels became excessively congested.



To enhance clarity, the x-axis labels were subsequently divided by month

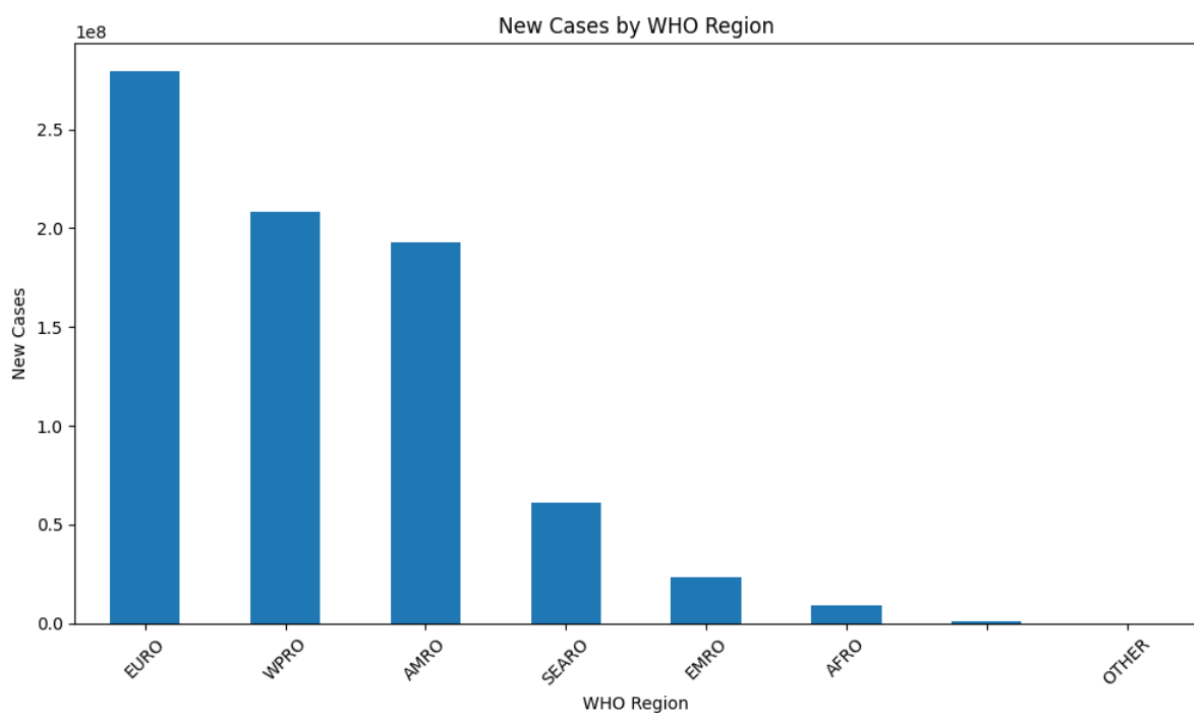


We calculated the total number of new cases across different WHO regions and generated this graph. Since the visualization was clear and not cluttered, no further modifications were necessary

Explanation of WHO Regions:

1. **EURO (European Region):** Covers Europe, including countries like the UK, Germany, Italy, and many more. This region appears to be the hardest hit by COVID-19.

2. **WPRO (Western Pacific Region):** Includes countries in East Asia, Oceania, and the Pacific Islands, such as China, Japan, South Korea, Australia, and New Zealand.
3. **AMRO (Region of the Americas):** Represents North and South America, including countries like the USA, Canada, Brazil, and Argentina.
4. **SEARO (South-East Asia Region):** Comprises countries such as India, Bangladesh, and Indonesia, representing South Asia's COVID-19 cases.
5. **EMRO (Eastern Mediterranean Region):** Consists of the Middle Eastern countries and North Africa, such as Saudi Arabia, Egypt, and Iran.
6. **AFRO (African Region):** Covers Sub-Saharan Africa, including countries like Nigeria, South Africa, and Kenya.
7. **OTHER:** Likely refers to regions not falling directly under the jurisdiction of WHO regions, possibly small territories or global organizations.



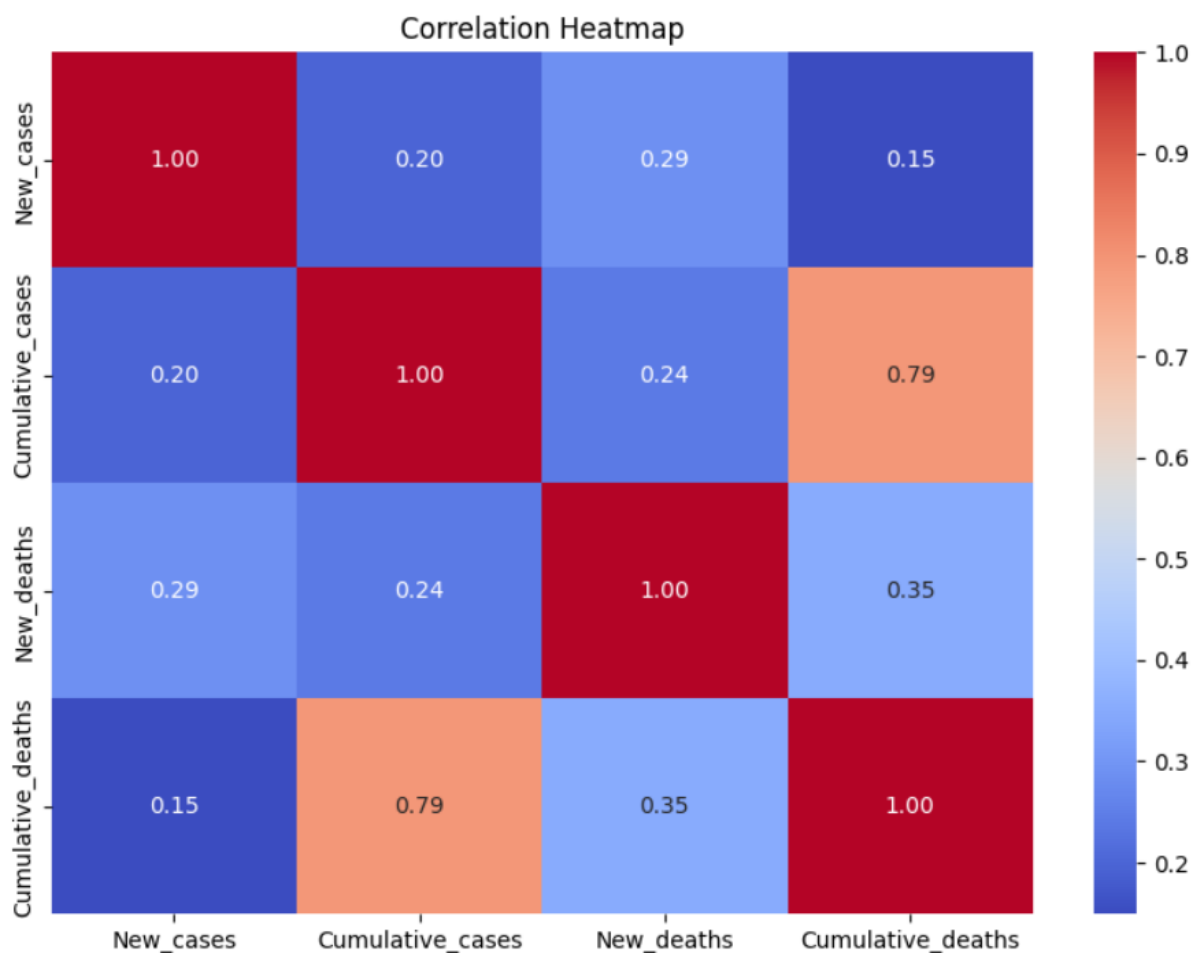
Inferences from the Bar Chart:

- **THE EURO** (European Region) has the highest number of new cases, surpassing 2.5 billion, indicating that Europe was the most affected WHO region in terms of new COVID-19 cases.
- **THE WPRO** (Western Pacific Region) and **AMRO** (Region of the Americas) follow, both with significant numbers of new cases.

- **SEARO** (South-East Asia Region) has fewer new cases compared to the top three, but still a notable amount.
- **EMRO** (Eastern Mediterranean Region) and **AFRO** (African Region) report considerably fewer new cases compared to the other regions.
- The **OTHER** category has a negligible number of cases, suggesting it represents a smaller or less impacted group.

The chart visually represents the disparity in COVID-19 impact across different geographical regions, highlighting Europe, the Americas, and the Western Pacific as the most affected areas.

Lastly, we decided to plot a correlation heatmap between 4 metrics ('new cases', 'Cumulative cases', 'New deaths', 'Cumulative deaths')



Inference from the Correlation Heatmap

Understanding the Relationships

This correlation heatmap provides insights into the relationships between four key metrics related to COVID-19: new cases, cumulative cases, new deaths, and cumulative deaths. The colour intensity and values within each cell represent the strength and direction of the correlation between the corresponding variables.

Key Findings:

1. **Strong Positive Correlation Between New Cases and Cumulative Cases:** As expected, the correlation between new cases and cumulative cases is very strong and positive (1.00). This indicates that a higher number of new cases directly leads to an increase in the cumulative total of cases over time.
2. **Moderate Positive Correlation Between New Cases and New Deaths:** A moderate positive correlation (0.29) exists between new cases and new deaths. This suggests that while there is a relationship between increased cases and deaths, other factors like age, health conditions, and healthcare access may also influence mortality rates.
3. **Strong Positive Correlation Between Cumulative Cases and Cumulative Deaths:** Similar to the relationship between new cases and cumulative cases, a strong positive correlation (0.79) is observed between cumulative cases and cumulative deaths. This indicates that a higher total number of cases is associated with a higher total number of deaths.
4. **Moderate Positive Correlation Between New Deaths and Cumulative Deaths:** A moderate positive correlation (0.35) exists between new deaths and cumulative deaths. While there is a relationship, it's not as strong as the other correlations, suggesting that factors like improved treatment methods or changes in population demographics might influence the relationship between daily deaths and the cumulative death count.

Overall, the heatmap confirms the expected relationships between these COVID-19 metrics. It highlights the strong association between new cases and cumulative cases, as well as the relationship between cases and deaths. However, it also suggests that other factors beyond the simple correlation between cases and deaths may influence mortality rates.

STEP 5: Showing Key Statistical Measures

Summary Statistics:

	New_cases	Cumulative_cases	New_deaths	Cumulative_deaths
count	3.902800e+04	5.784000e+04	25001.000000	5.784000e+04
mean	1.988104e+04	1.792594e+06	282.323947	2.001023e+04
std	2.707823e+05	7.797691e+06	1214.392195	8.186440e+04
min	-6.507900e+04	0.000000e+00	-3432.000000	0.000000e+00
25%	4.300000e+01	4.162750e+03	4.000000	2.800000e+01
50%	3.930000e+02	4.588300e+04	20.000000	5.650000e+02
75%	3.968000e+03	5.220872e+05	105.000000	6.974500e+03
max	4.047548e+07	1.034368e+08	47687.000000	1.194158e+06

The Following is basic statistical measures such as mean, count standard deviation etc, and here are the inferences from the same.

1. New cases:

- **Mean:** 19,881.0419,881.0419,881.04
- **Standard Deviation (std):** 270,782.3270,782.3270,782.3
- **Min:** -65,079-65,079-65,079
- **25th Percentile (25%):** 434343
- **50th Percentile (50%) / Median:** 393393393
- **75th Percentile (75%):** 3,9683,9683,968
- **Max:** 40,475,48040,475,48040,475,480

2. Cumulative cases:

- **Mean:** 1,792,5941,792,5941,792,594
- **Standard Deviation (std):** 7,797,6917,797,6917,797,691
- **Min:** 000
- **25th Percentile (25%):** 4,162.754,162.754,162.75
- **50th Percentile (50%) / Median:** 45,88345,88345,883
- **75th Percentile (75%):** 522,087.2522,087.2522,087.2
- **Max:** 103,436,800103,436,800103,436,800

3. New deaths:

- **Mean:** 282.32282.32282.32
- **Standard Deviation (std):** 1,214.391,214.391,214.39

- **Min:** -3,432-3,432-3,432
- **25th Percentile (25%):** 444
- **50th Percentile (50%) / Median:** 202020
- **75th Percentile (75%):** 105105105
- **Max:** 47,68747,68747,687

4. Cumulative deaths:

- **Mean:** 20,010.2320,010.2320,010.23
- **Standard Deviation (std):** 81,864.481,864.481,864.4
- **Min:** 000
- **25th Percentile (25%):** 282828
- **50th Percentile (50%) / Median:** 565565565
- **75th Percentile (75%):** 6,974.56,974.56,974.5
- **Max:** 1,194,1581,194,1581,194,158

Inferences and Insights

1. Distribution of New Cases:

- **Mean vs. Median:** The mean of new cases is significantly higher than the median, suggesting that there are some extremely high values that are skewing the average upwards.
- **Wide Range:** The range from minimum to maximum is very wide (-65,079-65,079 to 40,475,48040,475,48040,475,480), indicating a high variability in the number of new cases reported.
- **Presence of Negative Values:** The minimum value being negative suggests possible data errors or inconsistencies that should be investigated.

2. Cumulative Cases:

- **Mean vs. Median:** The mean is also higher than the median, similar to new cases, indicating skewness due to very high cumulative case counts.
- **Wide Range:** The range is quite large (000 to 103,436,800103,436,800103,436,800), showing substantial variability in cumulative cases reported.

- **High Variability:** The high standard deviation relative to the mean indicates that some locations or time periods have extremely high cumulative case counts compared to others.

3. New Deaths:

- **Mean vs. Median:** The mean of new deaths is much higher than the median, indicating a few extreme values.
- **Presence of Negative Values:** The minimum value being negative suggests possible data issues or errors.
- **Wide Range:** The range from -3,432 to 47,687 is significant, reflecting substantial variability in reported new deaths.

4. Cumulative Deaths:

- **Mean vs. Median:** Like new deaths, the mean is much higher than the median, indicating a skewed distribution due to extreme values.
- **Wide Range:** The range is from 000 to 1,194,158, showing a large variability in the cumulative number of deaths.

General Observations

- **Skewed Data:** For both new cases and New deaths, the presence of extreme values skews the mean significantly away from the median. This skewness should be considered when interpreting central tendencies.
- **Data Quality:** Negative values in new cases and new deaths suggest potential data issues or reporting errors that need to be addressed.
- **High Variability:** The large standard deviations in all columns reflect high variability in the data, which might be due to differences in reporting practices, regional outbreaks, or other factors.

Recommendations

1. **Investigate Negative Values:** Check and clean data for any erroneous negative values in new cases and New deaths.
2. **Consider Data Transformation:** Applying transformations (e.g., logarithms) to handle skewed data might provide better insights, especially for highly skewed distributions.
3. **Regional Analysis:** Conduct further analysis to understand regional differences in case and death counts, which might help in better understanding the data's variability.

And lastly the Highest & Lowest values of key determiners of the dataset are as follows

```
Highest Values:
New_cases          40475477.0
Cumulative_cases   103436829.0
New_deaths         47687.0
Cumulative_deaths  1194158.0
dtype: float64
Lowest Values:
New_cases          -65079.0
Cumulative_cases    0.0
New_deaths         -3432.0
Cumulative_deaths   0.0
dtype: float64
```

CONCLUSION

With this, I conclude the Python project using a sample dataset on COVID-19. Throughout this journey, I've gained a deeper understanding of key Python libraries like Pandas, Matplotlib, and Seaborn, which have been instrumental in data cleaning, processing, and visualization. This project has provided me with hands-on experience in handling large datasets, drawing meaningful insights from raw data, and transforming it into visually appealing charts and graphs.

By working with real-world data, I encountered challenges such as managing inconsistencies, missing values, and the complexities of aggregating information across various dimensions, all of which have strengthened my skills in data handling. I learned the importance of data preprocessing and how proper data visualization can lead to more effective decision-making and interpretation.

The dataset, sourced from Kaggle, offered a detailed view of COVID-19 cases and deaths across multiple countries, although its authenticity has not been verified by official sources. Despite this, it has served as an excellent resource for learning and honing my data analysis skills. All the code I used during this project will be provided in my GitHub repository, and the dataset will be uploaded there as well for anyone interested in exploring or replicating the project.

This experience has been an invaluable learning opportunity that will certainly benefit me in future roles requiring data-driven solutions, analytics, or Python programming.

Bibliography & Notes

GitHub Repository: <https://github.com/0904aniruddh>

Kaggle: <https://www.kaggle.com/datasets/abdoomoh/daily-covid-19-data-2020-2024?resource=download>