Name: Tanvi Shekhar Sawant

COVID Data Analysis

1. Dataset used: You can download the dataset from following link: https://raw.githubusercontent.com/datasets/covid-19/main/data/countries-aggregated.csv
2. Load and pre-process the dataset
    a. Load the dataset from a given URL.
    b. Check for null values and display the data types of each column.
    c. Convert the 'Date' column to datetime format.
    d. Display the first few rows of the dataset to understand its structure.

3. **Global Trends Visualization:**
    a. Aggregate the data by date and calculate the global totals for confirmed cases, deaths, and recoveries.
    b. Plot the trends of these global metrics over time using line plots.
    c. Insights: There is a gradual increase in the number of confirmed cases from 2020-08 onwards.
    d. The number of deaths suddenly increased from April 2020 globally. However, they were less than the number of confirmed cases worldwide.
    e. The recoveries saw a small dip towards the start of 2021, indicating the lockdown might have been revoked. However it again increased, but decreased steeply after August 2021. This anomaly arises due to no data being recorded after that month.
4. **Countries Comparison**
    a. Define a function to compare the trends of confirmed cases and deaths across multiple specified countries.
    b. Example usage of the function is provided to compare India and the US.
    c. India and US being most affected by the COVID, show similarities in the number of confirmed cases and deaths.
    d. However, India is still seen to have less affected compared to the US.
    e. **India also showed more better recovery** compared to US. This might indicate that better strategic measures were taken to control the outbreak.
5. **Comparative Analysis**
    a. Define a function to perform comparative analysis between top 10 most affected countries.
    b. The top 10 most affected countries were: **US, India, Brazil, France, UK, Russia, Turkey, Germany, Italy, and Spain.**
    c. **US again tops being the most affected countries** with highest confirmed cases and deaths compared to other 9 countries.
    d. **India and Brazil showed better recoveries** in spite of being 2nd and 3rd ranked for confirmed cases.
    e. Analyse the growth rate of cases and deaths using moving averages.
    f. The confirmed cases increased drastically after April 2020, indicating the start of the pandemic. The deaths started increasing around August start week.
6. **Correlation Matrix**
    a. Calculate the correlation matrix between confirmed cases, deaths, and recoveries.

   b. Visualize the correlation matrix using a heatmap.
   c. The **deaths and the confirmed cases show correlation of 0.96** indicating that the confirmed cases are more likely to lead to deaths. The **correlation between confirmed cases and recoveries is in negative**, indicating they are very less likely to be recoveries after confirmed case.
7. What is the trend of confirmed cases, deaths, and recoveries globally over time?
   a. Aggregate the data by date to calculate the global totals for confirmed cases, deaths, and recoveries.
   b. Use line plots to display the trends of these global metrics over time.
   c. From August 2020 onward, there is a gradual increase in the number of confirmed cases.
   d. A noticeable increase in the number of deaths occurred globally from April 2020, though deaths remained lower than the number of confirmed cases worldwide.
   e. Recoveries showed a slight dip at the beginning of 2021, suggesting that lockdowns might have been lifted. Recoveries then increased again, but experienced a sharp decline after August 2021, likely due to a lack of recorded data beyond that month.
8. Which country has the highest number of confirmed cases, deaths, and recoveries as of the latest date in the dataset?
   a. Identify the country with the highest number of confirmed cases, deaths, and recoveries as of the latest date in the dataset.
   b. Plot a bar chart to visualize these highest metrics.
   c. **US has highest number of confirmed cases and deaths**, and **India has highest recoveries**. (Mera Bharat Mahan 😌)
9. How do the trends of confirmed cases differ between the top 5 most affected countries?
   a. Identify the top 5 countries with the highest number of confirmed cases.
   b. Plot a line chart to visualize these highest metrics.
   c. US has high confirmed cases, again!
   d. European countries like **Brazil, France, and Germany have the comparatively very less confirmed cases** than US.
10. What is the growth rate of confirmed cases, deaths, and recoveries globally?
   a. Calculate the daily growth rate for confirmed cases, deaths, and recoveries globally.
   b. Plot the growth rates over time.
11. What are the cumulative confirmed cases, deaths, and recoveries for the top 10 most affected countries?
   a. The top 10 most affected countries show a stark contrast in the number of confirmed COVID-19 cases, with the United States, India, and Brazil leading significantly.
   b. The visualization for cumulative deaths highlights a similar pattern, with the United States, Brazil, and India reporting the highest number of deaths.
   c. The United States leads in reported deaths, followed by Brazil and India, reflecting their higher confirmed case counts and possibly varying healthcare capacities and pandemic response strategies.

   d. While the United States and Brazil consistently appear in the top ranks for both confirmed cases and deaths, other countries like Russia, France, and the United Kingdom also show significant impacts, albeit with varying degrees of severity.
   e. The data underscores the global disparity in pandemic impact, influenced by factors such as population density, healthcare infrastructure, testing capacity, and public health measures.
12. How does the recovery rate compare across different countries?
   a. First define a function to find the recovery rate for the top 10 most affected countries.
   b. The plot illustrates the recovery rates for each of the top 10 most affected countries over time.
   c. Countries like the United States, Brazil, and India show varying recovery rates throughout the pandemic period.
   d. **Recovery rates fluctuate over time**, influenced by factors such as healthcare capacity, treatment protocols, and public health interventions.
   e. **India, Peru and Brazil have shown significant better recoveries.**
   f. Next categorise all countries to continents.
   g. Use plotly to plot interactive graph. Creating bar chart for every continent.
   h. In Asia, Brunei had most recovery rate. Followed by China (😳).
13. Correlation Matrix
   a. High correlation between confirmed cases and death.
   b. Least correlation between confirmed cases and recovered.
14. Impact of Lockdown Measures
   a. Define a lockdown date as there is no exact dataset that could be found. So we are assuming the date based on known facts from Google. (not quite reliable)
   b. Categorize the data into pre-lockdown and post-lockdown periods.
   c. Plot the impact of lockdown measures on the confirmed cases for multiple countries. (Top 5 affected countries)
   d. **India has stagnant growth during second lockdown.**
   e. **European countries also show stagnant growth till start of 2022.** After that there is a steep increase.
15. How does the number of tests conducted relate to the number of confirmed cases in various countries?
   a. The number of tests each day divided by the number of confirmed cases each day. The series is smoothed by averaging daily figures over a rolling 7-day window.
   b. Not all countries report testing data on a daily basis. To generate this series we assume that testing changed equally on a daily basis over any periods in which no data was reported. Source Official data collated by Our World in Data – Last updated 10 April 2024 – processed by Our World in Data
   c. Plot the relationship between tests conducted and confirmed cases for various countries using an interactive plot.
   d. Using dropdown, check the relationship for each selected country.
16. What is the distribution of confirmed cases, deaths, and recoveries by continent
   a. Map countries to their respective regions.
   b. Plot the distribution of confirmed cases and deaths by region.
   c. North America: Highest median confirmed cases and deaths with significant variability, indicated by many outliers.

   d. Europe: High median confirmed cases and deaths with wide interquartile ranges and numerous outliers.
   e. Asia: Broad range of confirmed cases and deaths with many outliers, showing significant variability.
   f. Africa, South America, and Oceania: Lower median numbers of confirmed cases and deaths, with fewer but notable outliers.
17. Countries with highest confirmed cases on the last date recorded
   a. **US not just topping the GDP at this point.**
   b. **India having 527151 deaths** on last recorded date.
18. Mortality Rate
   a. Calculating the mortality rate using the formula
   ```
   df['Mortality_Rate'] = (df['Deaths'] / df['Confirmed']) * 100
   ```
   b. The highest mortality rate was seen on April 30 and May first week.
19. Active cases worldwide
   a. Make a filtered_data dataframe to find the recovered data till 2021-08-05, as the data is not recorded after this date.
   b. Based on this dataframe calculate the active cases using this formula filtered_df['Active'] = filtered_df['Confirmed'] - filtered_df['Deaths'] - filtered_df['Recovered']
   c. The graph shows gradual increase over time with a small steep increase around December 20, 2021.

Concluding thoughts:

a. US was most affected by this pandemic. It could be due to the high health insurance rates, poor strategies to control the pandemic, even after the second wave. High poverty could also be a contributing factor.
b. India was also most affected, given the population density of the country. However, around the second wave, proper measures were seem to be taken to control the outbreak. Timely vaccinations, strict lockdowns, better infrastructure etc.
c. Brazil, Australia and Dominico also showed greater recovery rates compared to other countries. Strategies like quarantine, limiting the foreigners, better access to healthcare etc could be the major reasons.
d. Overall, COVID affected many people and transformed the way we lived our lives for the pandemic years.
   The only good thing came out of it was WFH and online classes :)