COVID-19 Case Study

Overview: *The COVID-19 pandemic, caused by the SARS-CoV-2 virus, emerged in late 2019 and rapidly spread globally, leading to significant health, economic, and social impacts. This unprecedented health crisis highlighted the crucial role of data analysis in managing such pandemics. By meticulously tracking and analyzing data on confirmed cases, recoveries, and deaths, policymakers and health professionals can make informed decisions to control the spread of the virus and allocate resources effectively.*

Dataset Details:

Dataset *(Open link in a new tab and download the dataset.)*

This case study utilizes three key datasets, each providing daily updates on different aspects of the pandemic for various countries and regions:

- ***Confirmed Cases Dataset:*** *Contains the cumulative number of confirmed COVID-19 cases per day for each country and region. The data spans from January 22, 2020, to May 29, 2021, with over 276 geographic entries.*
- ***Deaths Dataset:*** *Records the cumulative number of deaths attributed to COVID-19, structured similarly to the confirmed cases dataset. This provides crucial information for assessing the lethality and outbreak severity in different areas.*
- ***Recovered Cases Dataset:*** *Includes data on the cumulative number of individuals who have recovered from COVID-19, which is vital for understanding the disease's progression and the effectiveness of treatment protocols.*

Each dataset includes columns for Province/State, Country/Region, geographic coordinates (Lat, Long), and a series of dates representing daily cumulative totals.

Objective of the Case Study:

The primary objectives of this case study are:

- ***Practical Application of Python:*** *To demonstrate and enhance practical skills in using Python for data analysis, focusing on data manipulation, cleaning, and visualization.*
- ***Insightful Data Analysis:*** *To provide analytical insights into the dynamics of COVID-19's spread, recovery, and mortality rates across different regions and over time.*
- ***Skill Development:*** *To equip students with the knowledge to handle real-world data using Python libraries such as Pandas for data manipulation, Matplotlib for data visualization, and Numpy for numerical data processing.*

Analysis Questions:

1. ***Data Loading:*** *How do you load the COVID-19 datasets for confirmed cases, deaths, and recoveries into Python using Pandas?*
2. ***Data Exploration***
    1. *After loading the datasets, what is the structure of each dataset in terms of rows, columns, and data types?*
    2. *Generate plots of confirmed cases over time for top countries.*
    3. *Generate plots of confirmed cases over time for China.*

3. ***Handling Missing Data:*** *Identify these missing values and replace them using a suitable imputation method such as forward filling for time-series data.*

4. ***Data Cleaning and Preparation:*** *Replace blank values in province column with "All Provinces"*

5. ***Independent Dataset Analysis***

    1. *Analyze the peak number of daily new cases in Germany, France, and Italy. Which country experienced the highest single-day surge, and when did it occur?*

    2. *Compare the recovery rates (recoveries/confirmed cases) between Canada and Australia as of December 31, 2020. Which country showed better management of the pandemic according to this metric?*

    3. *What is the distribution of death rates (deaths/confirmed cases) among provinces in Canada? Identify the province with the highest and lowest death rate as of the latest data point.*

6. ***Data Transformation***

    1. *Transform the 'deaths' dataset from wide format (where each column represents a date) to long format where each row represents a single date and columns are now country names, ensuring that the date column is in datetime format. How would this transformation be executed?*

    2. *What is the total number of deaths reported per country up to the current date?*

    3. *What are the top 5 countries with the highest average daily deaths?*

    4. *How have the total deaths evolved over time in the United States?*

7. ***Data Merging***

    1. *How would you merge the transformed datasets of confirmed cases, deaths, and recoveries on the 'Country/Region' and 'Date' columns to create a comprehensive view of the pandemic's impact?*

    2. *Analyze the monthly sum of confirmed cases, deaths, and recoveries for countries to understand the progression of the pandemic. [From the merged dataset]*

    3. *Redo the analysis in last question for United States, Italy and Brazil.*

8. ***Combined Data Analysis***

    1. *For the combined dataset, identify the three countries with the highest average death rates (deaths/confirmed cases) throughout 2020. What might this indicate about the pandemic's impact in these countries?*

    2. *Using the merged dataset, compare the total number of recoveries to the total number of deaths in South Africa. What can this tell us about the outcomes of COVID-19 cases in the country?*

    3. *Analyze the ratio of recoveries to confirmed cases for the United States on a monthly basis from March 2020 to May 2021. Which month experienced the highest recovery ratio, and what could be the potential reasons?*