Analyzing the Global Impact of COVID-19 using SQL

Presented by Anshu Kumari for Mentorness
Under Batch MIP-DA-09

Overview

The COVID-19 pandemic has significantly impacted global health and economies, creating a need for data-driven insights. This project uses SQL to analyze a comprehensive COVID-19 dataset, aiming to understand the virus's spread, impact, and recovery patterns. By examining confirmed cases, deaths, and recoveries across different regions and time periods, we provide detailed insights into the pandemic's progression.

Our analysis involves cleaning the data and executing SQL queries to answer key questions about the virus's impact, including monthly trends, regional disparities, and country-specific analyses.

Key Objectives

01

Data Exploration

03

Data Analysis

02

SQL Proficiency

04

Insight Generation

Importing Data in MYSQL

01

Create Database

```
create database covid;
```

02

Load Data from CSV

- Open the database schema.
- Right-click and select "Table Data Import Wizard".
- Choose the location of the CSV file.

03

Check the data type

- Check and adjust data types as needed.
- Import the data into the database.

DATASET

- **Province:** Geographic subdivision within a country or region.
- **Country/Region:** Geographic entity where data is recorded.
- **Latitude:** North-south position on Earth's surface.
- **Longitude:** East-west position on Earth's surface.
- **Date:** Recorded date of COVID-19 data.
- **Confirmed:** Number of diagnosed COVID-19 cases.
- **Deaths:** Number of COVID-19 related deaths.
- **Recovered:** Number of recovered COVID-19 cases.

KEY QUESTIONS ADDRESSED IN THE ANALYSIS

1. Checking for NULL Values
2. Handling NULL Values
3. Counting Total Rows
4. Date Range
5. Number of Months in Dataset
6. Monthly Averages
7. Most Frequent Values
8. Minimum Values per Year
9. Maximum Values per Year
10. Total Cases Each Month
11. Spread Analysis - Confirmed Cases
12. Spread Analysis - Death Cases
13. Spread Analysis - Recovered Cases
14. Country with Highest Confirmed Cases
15. Country with Lowest Death Cases
16. Top 5 Countries by Recovered Cases

Q1. Write a code to check NULL values

```
SELECT
    SUM(CASE WHEN Province IS NULL THEN 1 ELSE 0 END) AS column1_nulls,
    SUM(CASE WHEN Country IS NULL THEN 1 ELSE 0 END) AS column2_nulls,
    SUM(CASE WHEN Latitude IS NULL THEN 1 ELSE 0 END) AS column3_nulls,
    SUM(CASE WHEN Longitude IS NULL THEN 1 ELSE 0 END) AS column4_nulls,
    SUM(CASE WHEN Date IS NULL THEN 1 ELSE 0 END) AS column5_nulls,
    SUM(CASE WHEN Confirmed IS NULL THEN 1 ELSE 0 END) AS column6_nulls,
    SUM(CASE WHEN Deaths IS NULL THEN 1 ELSE 0 END) AS column7_nulls,
    SUM(CASE WHEN Recovered IS NULL THEN 1 ELSE 0 END) AS column8_nulls
FROM corona_virus;
```

column1_nulls	column2_nulls	column3_nulls	column4_nulls	column5_nulls	column6_nulls	column7_nulls	column8_nulls
0	0	0	0	0	0	0	0

Q3. Check total number of rows

```
SELECT
    COUNT(*) AS total_no_of_rows
FROM
    corona_virus;
```

total_no_of_rows
78386

Q4. Check what is start_date and end_date

```
SELECT
    MIN(date) AS starting_date, MAX(date) AS ending_date
FROM
    corona_virus;
```

starting_date	ending_date
2020-01-22	2021-06-13

Q5. Number of month present in dataset

```
SELECT
   COUNT(DISTINCT date_format(date,'%y,%m')) AS no_of_months
FROM
   corona_virus;
```

no_of_months
18

Q6. Find monthly average for confirmed, deaths, recovered

```
SELECT
    date_format(date,'%m,%y') AS months,
   ROUND(AVG(Confirmed), 2) AS avg_confirmed,
   ROUND(AVG(Deaths), 2) AS avg_deaths,
   ROUND(AVG(Recovered), 2) AS avg_recovered
FROM
   corona_virus
GROUP BY months;
```

Q7. Find most frequent value for confirmed, deaths, recovered each month

```
SELECT month, 'Confirmed' AS category, value, freq
FROM (SELECT date_format(date,'%m,%y') AS months, Confirmed AS value,
COUNT(Confirmed) AS freq FROM corona_virus GROUP BY MONTH(date) , Confirmed
ORDER BY month , freq DESC LIMIT 1) AS confirmed_data
UNION ALL SELECT month, 'Deaths' AS category, value, freq
FROM (SELECT MONTH(date) AS month, Deaths AS value, COUNT(Deaths) AS freq
FROM corona_virus GROUP BY MONTH(date) , Deaths
ORDER BY month , freq DESC LIMIT 1) AS deaths_data
UNION ALL SELECT month, 'Recovered' AS category, value, freq
FROM (SELECT MONTH(date) AS month, Recovered AS value,
COUNT(Recovered) AS freq FROM corona_virus
 GROUP BY months , Recovered
ORDER BY month , freq DESC LIMIT 1) AS recovered_data;
```

month	category	value	freq
01,20	Confirmed	0	1373
01,20	Deaths	0	1530
01,20	Recovered	0	1511

Q8. Find minimum values for confirmed, deaths, recovered per year

```
SELECT year(date) as year,
 MIN(Confirmed) AS min_confirmed,
 MIN(Deaths) AS min_deaths,
 MIN(Recovered) AS min_recovered
FROM corona_virus
group by year;
```

year	min_confirmed	min_deaths	min_recovered
2020	0	0	0
2021	0	0	0

Q9. Find maximum values of confirmed, deaths, recovered per year

```
SELECT year(date) as year,
    MAX(Confirmed) AS max_confirmed,
    MAX(Deaths) AS max_deaths,
    MAX(Recovered) AS max_recovered
FROM corona_virus
group by year;
```

year	max_confirmed	max_deaths	max_recovered
2020	823225	3752	1123456
2021	414188	7374	422436

Q10. The total number of case of confirmed, deaths, recovered each month

```
SELECT  date_format(date,'%m,%y')as months,
    SUM(Confirmed) AS total_confirmed,
    SUM(Deaths) AS total_deaths,
    SUM(Recovered) AS total_recovered
FROM corona_virus group by months;
```

months	total_confirmed	total_deaths	total_recovered
01,20	6384	190	143
02,20	68312	2651	31405
03,20	769236	41346	133070
04,20	2336798	191833	792987
05,20	2744333	144561	1519547
06,20	3969634	137757	2535417
07,20	6838092	167613	4693120
08,20	7694938	179200	6202833
09,20	8244794	160671	6647749
10,20	11515841	175484	6782150
11,20	16595938	262247	9172292
12,20	19336799	339996	11924903
01,21	18672205	401893	9164347
02,21	10492664	298239	6719785
03,21	13924790	282620	7888013
04,21	21711021	362387	14205507
05,21	19121083	366549	19131842
06,21	5022282	132657	5544438

Q11. Check how corona virus spread out with respect to confirmed case

```
SELECT
    round(SUM(Confirmed),2) AS total_confirmed_cases,
    round(AVG(Confirmed),2) AS average_confirmed_cases,
    round(VARIANCE(Confirmed),2) AS variance_confirmed_cases,
    round(STDDEV(Confirmed),2) AS std_dev_confirmed_cases
FROM
    corona_virus;
```

total_confirmed_cases	average_confirmed_cases	variance_confirmed_cases	std_dev_confirmed_cases
169065144	2156.83	157288925.08	12541.49

Q12. Check how corona virus spread out with respect to death case per month

```
SELECT
    date_format(date,'%m,%y')as months,
    SUM(Deaths) AS total_Death_cases,
    round(AVG(Deaths),0) AS
average_Death_cases,
    round(VARIANCE(Deaths),0) AS
variance_Death_cases,
    round(STDDEV(Deaths),0) AS
std_dev_Death_cases
FROM
    corona_virus
GROUP BY months;
```

months	total_Death_cases	average_Death_cases	variance_Death_cases	std_dev_Death_cases
01,20	190	0	4	2
02,20	2651	1	68	8
03,20	41346	9	3901	62
04,20	191833	42	40504	201
05,20	144561	30	20685	144
06,20	137757	30	16929	130
07,20	167613	35	21140	145
08,20	179200	38	23273	153
09,20	160671	35	20103	142
10,20	175484	37	17580	133
11,20	262247	57	27774	167
12,20	339996	71	65345	256
01,21	401893	84	102758	321
02,21	298239	69	68479	262
03,21	282620	59	54386	233
04,21	362387	78	94611	308
05,21	366549	77	131769	363
06,21	132657	66	112964	336

Q13. Check how corona virus spread out with respect to recovered case

```
SELECT
    SUM(Recovered) AS total_Recovered_cases,
    round(AVG(Recovered),2) AS average_Recovered_cases,
    round(VARIANCE(Recovered),2) AS variance_Recovered_cases,
    round(STDDEV(Recovered),2) AS std_dev_Recovered_cases
FROM
    corona_virus;
```

total_Recovered_cases	average_Recovered_cases	variance_Recovered_cases	std_dev_Recovered_cases
113089548	1442.73	107029523.26	10345.51

Q14. Find Country having highest number of the Confirmed case

```
SELECT
    country, SUM(Confirmed)
FROM
    corona_virus
GROUP BY country
ORDER BY SUM(Confirmed) DESC
LIMIT 1;
```

country	SUM(Confirmed)
US	33461982

Q15. Find Country having lowest number of the death case

```
select country from(
SELECT country, SUM(Deaths) AS TotalDeaths,
      RANK() OVER (ORDER BY SUM(Deaths) ASC) AS DeathRank
FROM corona_virus
GROUP BY country
ORDER BY TotalDeaths ASC)as a where Deathrank=1;
```

country
Dominica
Kiribati
Marshall Islands
Samoa

Q16. Find top 5 countries having highest recovered case

```
SELECT
    country, SUM(Recovered)
FROM
    corona_virus
GROUP BY country
ORDER BY SUM(Recovered) DESC
LIMIT 5;
```

country	SUM(Recovered)
India	28089649
Brazil	15400169
US	6303715
Turkey	5202251
Russia	4745756

INSIGHTS

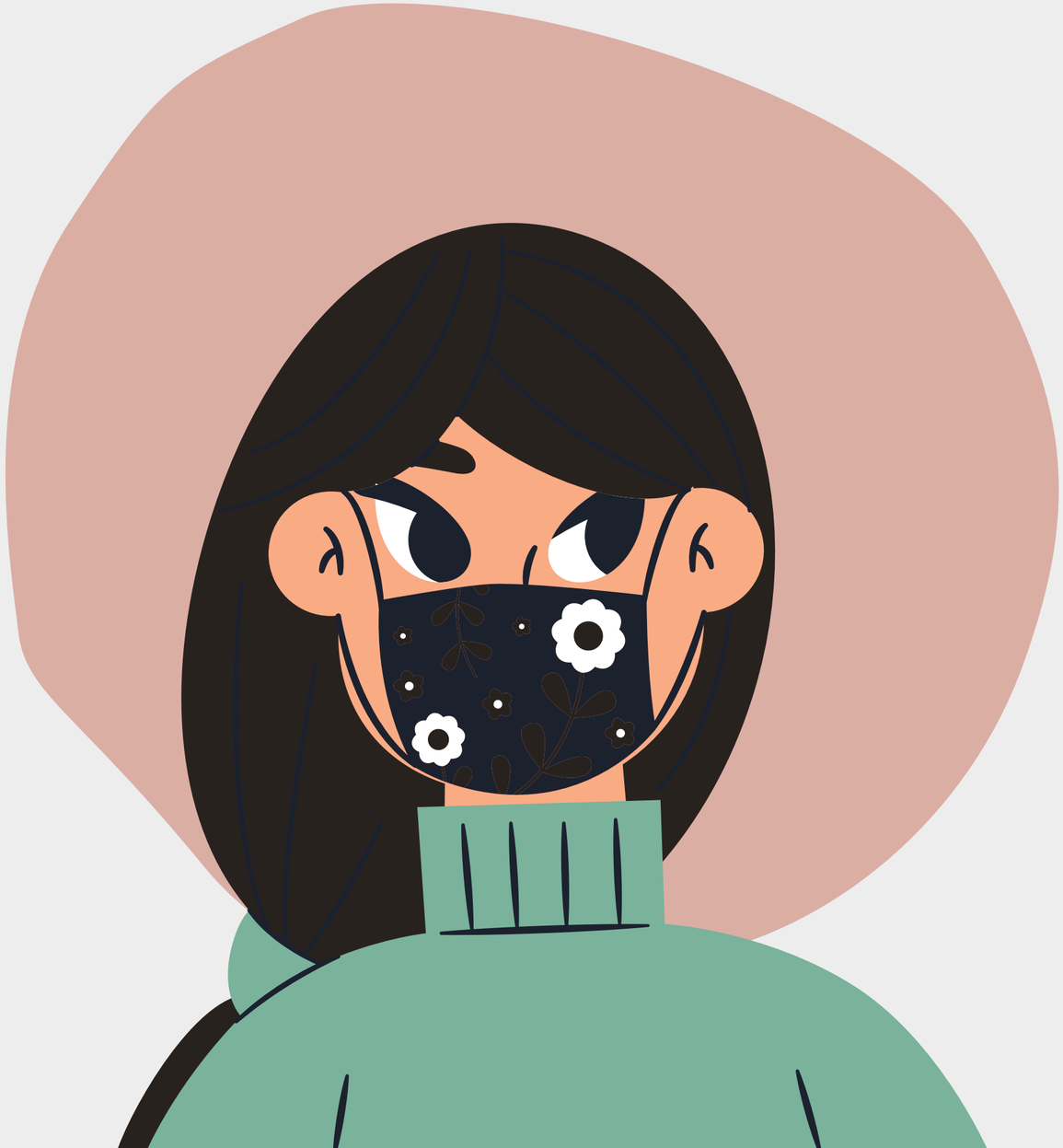
- The dataset covers the period from January 22, 2020, to June 13, 2021, encompassing 18 months of data on confirmed cases, deaths, and recoveries.

- From January 2020 to June 2021, average monthly confirmed cases, deaths, and recoveries rose significantly, peaking between November 2020 and April 2021, before declining in mid-2021.

- A total of 169,065,144 confirmed cases were recorded globally, with the US having the highest number of confirmed cases at 33,461,982.

- The dataset shows a global total of 3,522,465 deaths and 113,089,548 recoveries, reflecting significant impacts and recovery efforts worldwide.

- Countries like Dominica, Kiribati, Marshall Islands, and Samoa reported the lowest number of deaths, indicating effective containment or low exposure, whereas India, Brazil, the US, Turkey, and Russia reported the highest number of recoveries.



Thank You!

We appreciate your time and interest in our COVID-19 data analysis.

If you have any questions or need further information, please feel free to ask.

@Anshu Kumari