CORONA VIRUS ANALYSIS

COVID-19 DATASET ANALYSIS

USING SQL BY

PREM NIMJE

CONTENTS

INTRODUCTION

- The CORONA VIRUS pandemic has had a significant impact on public health and has created an urgent
- need for data-driven insights to understand the spread of the virus. As a data analyst, you have been
- tasked with analyzing a CORONA VIRUS dataset to derive meaningful insights and present your findings.

DATASET VIEW

Description of each column in dataset:

Province: Geographic subdivision within a country/region.

Country/Region: Geographic entity where data is recorded.

Date: Recorded date of CORONA VIRUS data.

Longitude: East-west position on Earth's surface.

Latitude: North-south position on Earth's surface.

Confirmed: Number of diagnosed CORONA VIRUS cases.

Deaths: Number of CORONA VIRUS related deaths.

Recovered: Number of recovered CORONA VIRUS cases.

DATA ANALYSIS OF THE DATASET

- Loading the Dataset
- Cleaning Dataset
- Analyzing the questions and answers

Loading the dataset

- SELECT * FROM corona_data.`corona virus dataset`;

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Province	Country/Region	Latitude	Longitude	Date	Confirmed	Deaths	Recovered
Afghanistan	Afghanistan	33.93911	67.709953	22-01-2020	0	0	0
Afghanistan	Afghanistan	33.93911	67.709953	23-01-2020	0	0	0
Afghanistan	Afghanistan	33.93911	67.709953	24-01-2020	0	0	0
Afghanistan	Afghanistan	33.93911	67.709953	25-01-2020	0	0	0
Afghanistan	Afghanistan	33.93911	67.709953	26-01-2020	0	0	0

irus dataset 1

Cleaning the Dataset

- Checking for missing value/Null Value

```
-- To avoid any errors, check missing value / null value
-- Q1. Write a code to check NULL values
USE corona_data;
SELECT COUNT(*) AS null_value
FROM corona_data.`corona virus dataset`
WHERE 'ALL COLUMNS' IS NULL;
-- Anwer: We have 0 null values, which means we do not have any missing values..
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	null_value
▶	0

Result 3

Output

IF NULL VALUES ARE PRESENT,UPDATE THEM WITH ZEROS FOR ALL COLUMNS

- The output for NULL values being present is zero, meaning there are no NULL values in any of the columns.

```
-- Q2. If NULL values are present, update them with zeros for all columns.
UPDATE corona_data.`corona virus dataset`
SET
Province = COALESCE(Province, 0),
`Country/Region` = COALESCE(`Country/Region`, 0),
Latitude = COALESCE(`Latitude`, 0),
Longtitude = COALESCE(`Longtitude`, 0),
Date = COALESCE(`Date`, 0),
Confirmed = COALESCE(Confirmed, 0),
Deaths = COALESCE(Deaths, 0),
Recovered = COALESCE(Recovered, 0);
```

Analyzing the questions and answers

```sql
-- Q3. check total number of rows
USE corona_data;
SELECT COUNT(*) AS total_number_of_rows
FROM corona_data.`corona virus dataset`;
-- Anwer:The total number of rows is 78386
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_number_of_rows
78386

Result Grid

Form Editor

Field Types

Check what is start_date and end_date

```sql
USE corona_data;
SELECT
    MIN(STR_TO_DATE(Date, '%d-%m-%Y')) AS start_date,
    MAX(STR_TO_DATE(Date, '%d-%m-%Y')) AS end_date
FROM
    corona_data.`corona virus dataset`;
-- Anwer: The start_date is '2020-01-22' and the end_date is '2021-06-13'
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

start_date	end_date
2020-01-22	2021-06-13

Result Grid
Form Editor
Field Types

Number of month present in dataset

```sql
USE corona_data;
SELECT COUNT(DISTINCT MONTH(STR_TO_DATE(Date, '%d-%m-%Y'))) AS num_of_months
FROM corona_data.`corona virus dataset`;
-- Anwer: Number of month present is 12
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	num_of_months
▶	12

Result Grid

Form Editor

Field Types

Find monthly average for confirmed, deaths, recovered

```sql
USE corona_data;
SELECT
      MONTH(STR_TO_DATE(Date, '%d-%m-%Y')) AS month,
      AVG(Confirmed) AS avg_confirmed,
      AVG(Deaths) AS avg_deaths,
      AVG(Recovered) AS avg_recovered
FROM corona_data.`corona virus dataset`
GROUP BY MONTH(STR_TO_DATE(Date, '%d-%m-%Y'))
ORDER BY MONTH;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

month	avg_confirmed	avg_deaths	avg_recovered
1	2958.2814	63.6812	1451.4555
2	1203.1187	34.2777	769.1034
3	1538.9638	33.9302	840.0799
4	2602.5778	59.9805	1623.2136
5	2290.0519	53.5306	2162.9021
6	1357.8852	40.8357	1220.1533
7	1432.3611	35.1096	983.0582
8	1611.8429	37.5367	1299.2947
9	1784.5874	34.7773	1438.9067
10	2412.1996	36.7583	1420.6431
11	3592.1944	56.7634	1985.3446
12	4050.4397	71.2183	2497.8850

Result Grid
Form Editor
Field Types

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

```sql
USE corona_data;
SELECT
    MONTH(STR_TO_DATE(Date, '%d-%m-%Y')) AS month,
    MAX(Confirmed) AS most_frequent_confirmed,
    MAX(Deaths) AS most_frequent_deaths,
    MAX(Recovered) AS most_frequent_recovered
FROM (
    SELECT
        Date,
        Confirmed,
        Deaths,
        Recovered,
        COUNT(*) AS frequency
    FROM corona_data.`corona virus dataset`
    GROUP BY Date, Confirmed, Deaths, Recovered
) AS subquery
GROUP BY MONTH(STR_TO_DATE(Date, '%d-%m-%Y'))
ORDER BY MONTH;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

month	most_frequent_confirmed	most_frequent_deaths	most_frequent_recovered
1	300462	4475	87090
2	134975	3907	98389
3	100158	3869	102138
4	401993	4249	299988
5	414188	4529	422436
6	134154	7374	231456
7	75866	1595	140050
8	85687	1505	95881
9	97894	1703	101468
10	99264	3351	388340
11	207933	2259	139292
12	823225	3752	1123456

Result Grid
Form Editor
Field Types

Find minimum values for confirmed, deaths, recovered per year

```sql
USE corona_data;
SELECT
    YEAR(STR_TO_DATE(Date, '%d-%m-%Y')) AS year,
    MIN(Confirmed) AS min_confirmed,
    MIN(Deaths) AS min_deaths,
    MIN(Recovered) AS min_recovered
FROM
 `corona virus dataset`
GROUP BY
    YEAR(STR_TO_DATE(Date, '%d-%m-%Y'))
ORDER BY YEAR;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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

Result Grid

Form Editor

Find maximum values of confirmed, deaths, recovered per year

```sql
USE corona_data;
SELECT
    YEAR(STR_TO_DATE(Date, '%d-%m-%Y')) AS year,
    MAX(Confirmed) AS max_confirmed,
    MAX(Deaths) AS max_deaths,
    MAX(Recovered) AS max_recovered
FROM
    `corona virus dataset`
GROUP BY
    YEAR(STR_TO_DATE(Date, '%d-%m-%Y'))
ORDER BY
    YEAR;
-- Anwer: :Run the code to see the maximum values for confirmed, deaths, and recovered cases per year.
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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

Result Grid

Form

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

```sql
USE corona_data;
SELECT
    MONTH(STR_TO_DATE(Date, '%d-%m-%Y')) AS month,
    SUM(Confirmed) AS total_confirmed_cases,
    SUM(Deaths) AS total_death_cases,
    SUM(Recovered) AS total_recovered_cases
FROM
    `corona virus dataset`
GROUP BY
    MONTH(STR_TO_DATE(Date, '%d-%m-%Y'))
ORDER BY
    MONTH;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

month	total_confirmed_cases	total_death_cases	total_recovered_cases
1	18678589	402083	9164490
2	10560976	300890	6751190
3	14694026	323966	8021083
4	24047819	554220	14998494
5	21865416	511110	20651389
6	8991916	270414	8079855
7	6838092	167613	4693120
8	7694938	179200	6202833
9	8244794	160671	6647749
10	11515841	175484	6782150
11	16595938	262247	9172292

Result Grid
Form Editor
Field Types

Check how corona virus spread out with respect to confirmed case

```sql
-- 11. Check how corona virus spread out with respect to confirmed case
-- (Eg.: total confirmed cases, their average, variance & STDEV )
USE corona_data;
SELECT
    SUM(Confirmed) AS total_confirmed_cases,
    AVG(Confirmed) AS average_confirmed_cases,
    VARIANCE(Confirmed) AS variance_confirmed_cases,
    STDDEV(Confirmed) AS stddev_confirmed_cases
FROM
    `corona virus dataset`;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_confirmed_cases	average_confirmed_cases	variance_confirmed_cases	stddev_confirmed_cases
169065144	2156.8283	157288925.07796532	12541.488152446875

Result Grid

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

```sql
-- Q12. Check how corona virus spread out with respect to death case per month
-- (Eg.: total confirmed cases, their average, variance & STDEV )
USE corona_data;
SELECT
    MONTH(STR_TO_DATE(Date, '%d-%m-%Y')) AS month,
    SUM(Deaths) AS total_death_cases,
    AVG(Deaths) AS average_death_cases,
    VARIANCE(Deaths) AS variance_death_cases,
    STDDEV(Deaths) AS stddev_death_cases
FROM
    `corona virus dataset`
GROUP BY
    MONTH(STR_TO_DATE(Date, '%d-%m-%Y'))
ORDER BY
    MONTH;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

month	total_death_cases	average_death_cases	variance_death_cases	stddev_death_cases
1	402083	63.6812	78999.5307609659	281.0685517110833
2	300890	34.2777	34848.64785490521	186.67792546229245
3	323966	33.9302	29781.93292256146	172.57442719754704
4	554220	59.9805	67898.57559453539	260.5735512183372
5	511110	53.5306	76767.73838185583	277.06991605343194
6	270414	40.8357	46243.20314719306	215.04232873365433
7	167613	35.1096	21140.154944373826	145.39654378414167
8	179200	37.5367	23272.99645685882	152.55489653517785
9	160671	34.7773	20102.7692237308	141.78423475030925
10	175484	36.7583	17580.07101972725	132.589860169348
11	262247	56.7634	27773.793596962234	166.6547136955995

Result Grid
Form Editor
Field Types

Check how corona virus spread out with respect to recovered case-- (Eg.: total confirmed cases, their average, variance & STDEV)USE corona_data;

```sql
USE corona_data;
SELECT
    SUM(Recovered) AS total_recovered_cases,
    AVG(Recovered) AS average_recovered_cases,
    VARIANCE(Recovered) AS variance_recovered_cases,
    STDDEV(Recovered) AS stddev_recovered_cases
FROM
    `corona virus dataset`;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

total_recovered_cases	average_recovered_cases	variance_recovered_cases	stddev_recovered_cases
113089548	1442.7264	107029523.26229636	10345.507395110999

Result Grid

Find Country having highest number of the Confirmed case

```
USE corona_data;
SELECT
    `Country/Region` AS Country,
    MAX(Confirmed) AS Highest_Confirmed_Cases
FROM
    `corona virus dataset`
GROUP BY
    `Country/Region`
ORDER BY
    Highest_Confirmed_Cases DESC
LIMIT 1;
-- Anwer:The country with highest confirmed cases is 'Turkey'
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Country	Highest_Confirmed_Cases
Turkey	823225

Result Grid

Find Country having lowest number of the death case

```
USE corona_data;
SELECT
    `Country/Region`,
    MIN(Deaths) AS lowest_death_cases
FROM
    `corona virus dataset`
GROUP BY
    `Country/Region`
ORDER BY
    lowest_death_cases ASC
LIMIT 1;
-- Anwer: The country with the lowest number of death case is 'Afghanistan'
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Country/Region	lowest_death_cases
Afghanistan	0

Result Grid

Find top 5 countries having highest recovered case

```sql
USE corona_data;
SELECT
    `Country/Region`,
    SUM(Recovered) AS highest_recovered_cases
FROM
    `corona virus dataset`
GROUP BY
    `Country/Region`
ORDER BY
    highest_recovered_cases DESC
LIMIT 5;
-- Anwer: The countries with the highest reported cases are India, Brazil, the United States, Turkey, and Russia.
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Country/Region	highest_recovered_cases
India	28089649
Brazil	15400169
US	6303715
Turkey	5202251
Russia	4745756

Result Grid

Form Editor

INSIGHT

- Highest and Lowest Averages:

The average number of recovered cases is highest in the 12th month and lowest in the second month.

- **Highest and Lowest Frequencies:**
  - The 6th month sees the highest frequency of deaths, while the 8th month has the lowest.
  - The 12th month has the highest frequency of recoveries, with the 1st month having the lowest.
- **Highest Total Cases:**
  - The 4th month records the highest confirmed cases, while the 7th month has the lowest.
  - The 4th month also records the highest number of deaths, with the 9th month having the lowest.
  - The 5th month sees the highest total recovered cases, with the 7th month having the lowest.
- **Year-wise Observations:**
  - 2020 records the highest values for confirmed and recovered cases, while 2021 records the highest for deaths.
- **Overall Statistics:**
  - The total spread of COVID-19 is substantial, with over 169 million confirmed cases.
  - The average confirmed cases variance and standard deviation indicate significant variability in spread.
- **Country-specific Observations:**
  - Turkey has the highest confirmed cases, while Afghanistan has the lowest number of deaths.
  - India, Brazil, the United States, Turkey, and Russia are the countries with the highest reported cases.

CONCLUSION

The insights drawn from the data suggest a dynamic and varied pattern of COVID-19 spread across different time periods and regions. While some months and years have seen peaks in confirmed cases, deaths, and recoveries, others have experienced relatively lower numbers. The data also highlights the disparities among countries in terms of the severity of the pandemic. Understanding these patterns is crucial for formulating effective public health policies and interventions to mitigate the impact of the virus. Further analysis could delve into factors influencing these variations, such as government responses, healthcare infrastructure, and population demographics.