

CORONA VIRUS ANALYSIS



**Data-Driven Insights into the Spread and Impact of
CORONA VIRUS**

Author: Buntty Patil

Date: 06-06-2024

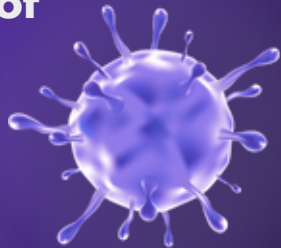


TABLE OF CONTENTS



01

Project Overview

02

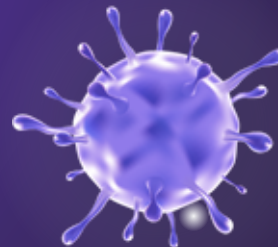
Dataset

03

Analysis Report

04

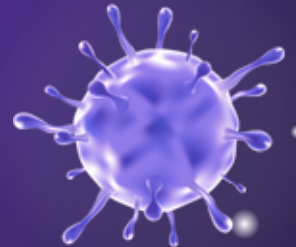
Conclusion





Note

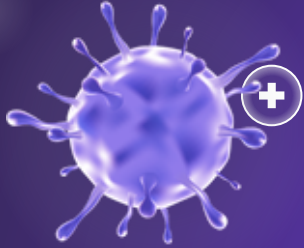
This project is associated with 'Mentorless' while working as a data analyst intern. All analysis questions have been assigned by the 'Mentorless' team, and the same has been completed using MySQL.





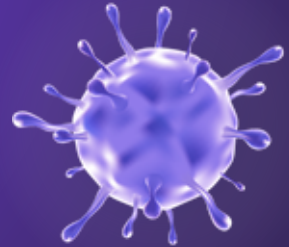
01

Project Overview



Project Overview

The CORONA VIRUS pandemic has significantly impacted 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.



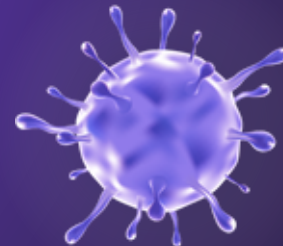
02

Dataset

DATASET



VARIABLE	DESCRIPTION
province	Geographic subdivision within a country/region.
country_or_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 CORONA VIRUS data.
confirmed	Number of diagnosed CORONA VIRUS cases.
deaths	Number of CORONA VIRUS related deaths.
recovered	Number of recovered CORONA VIRUS cases.



03

Analysis Report

Q-1 Write a code to check the null values



```
# Checking null values
SELECT
    SUM(CASE WHEN province IS NULL THEN 1 ELSE 0 END) AS province_null_counts,
    SUM(CASE WHEN country_or_region IS NULL THEN 1 ELSE 0 END) AS country_or_region_null_counts,
    SUM(CASE WHEN latitude IS NULL THEN 1 ELSE 0 END) AS latitude_null_counts,
    SUM(CASE WHEN longitude IS NULL THEN 1 ELSE 0 END) AS longitude_null_counts,
    SUM(CASE WHEN date IS NULL THEN 1 ELSE 0 END) AS date_null_counts,
    SUM(CASE WHEN confirmed IS NULL THEN 1 ELSE 0 END) AS confirmed_null_counts,
    SUM(CASE WHEN deaths IS NULL THEN 1 ELSE 0 END) AS deaths_null_counts,
    SUM(CASE WHEN recovered IS NULL THEN 1 ELSE 0 END) AS recovered_null_counts
FROM corona_virus_dataset;
```

Result Grid							
Filter Rows: <input type="text"/>							
Export:							
Wrap Cell Content:							
province_null_counts	country_or_region_null_counts	latitude_null_counts	longitude_null_counts	date_null_counts	confirmed_null_counts	deaths_null_counts	recovered_null_counts
0	0	0	0	0	0	0	0

The dataset did not contain any null values.



Q-2 Check total number of rows



```
# Check total number of rows  
SELECT COUNT(*) AS num_rows  
FROM corona_virus_dataset;
```

Result Grid	
	num_rows
▶	78386

A total of 78386 rows are present in the dataset.



Q-3 Check what the start date and end date is



```
# Check what is start_date and end_date
SELECT
    MIN(date) AS start_date,
    MAX(date) AS end_date
FROM corona_virus_dataset;
```

Result Grid			Filter Rows:	
	start_date	end_date		
▶	2020-01-22	2021-06-13		

Start date: January 22, 2020

End date: June 13, 2021



Q-4 Number of months present in the dataset

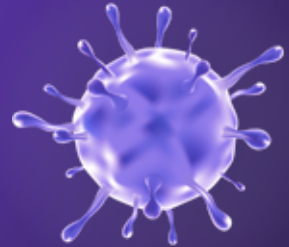


```
# Number of month present in dataset
SELECT
    COUNT(DISTINCT DATE_FORMAT(date, '%Y-%m')) AS num_months
FROM corona_virus_dataset;
```

Result Grid	
	num_months
▶	18






A total of 18 months present in the dataset.

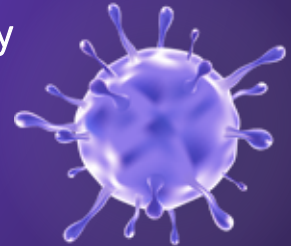


Q-5 Find the monthly average for confirmed, deaths, and recovered cases.

```
# 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_dataset
GROUP BY months;
```

Result Grid		 Filter Rows:	<input type="text"/>	Export: 
	months	avg_confirmed	avg_deaths	avg_recovered
▶	01-2020	4.15	0.12	0.09
	02-2020	15.30	0.59	7.03
	03-2020	161.13	8.66	27.87
	04-2020	505.80	41.52	171.64
	05-2020	574.85	30.28	318.30
	06-2020	859.23	29.82	548.79
	07-2020	1432.36	35.11	983.06
	08-2020	1611.84	37.54	1299.29
	09-2020	1784.59	34.78	1438.91
	10-2020	2412.20	36.76	1420.64
	11-2020	3592.19	56.76	1985.34
	12-2020	4050.44	71.22	2497.89
	01-2021	3911.23	84.18	1919.64
	02-2021	2433.36	69.16	1558.39
	03-2021	2916.80	59.20	1652.29
	04-2021	4699.36	78.44	3074.79
	05-2021	4005.25	76.78	4007.51
	06-2021	2508.63	66.26	2769.45

- Based on the monthly average for confirmed, deaths, and recovered cases, it is evident that there was a significant peak in average cases during the first wave, from April 2020 to January 2021. There was a gradual increase during this period.
- In February and March 2021, there was a slight decrease in the monthly average cases.
- However, in April 2021, the monthly average cases increased again and reached a peak, which lasted until June 2021.



Q-6 Find the minimum values for confirmed, deaths, and recovered cases per year.

```
# Minimum values of confirmed, deaths, recovered per year
SELECT
    DATE_FORMAT(date, '%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
ORDER BY year;
```

Result Grid				
	year	min_confirmed	min_deaths	min_recovered
▶	2020	0	0	0
	2021	0	0	0

- The minimum number of confirmed, deaths, and recovered cases was 0 in both 2020 and 2021.
- In 2020, the minimum value could be 0 during the initial stages of the pandemic when the virus had not yet spread nationwide.
- In 2021, when the number of cases was at its peak, it was not possible for the minimum values in 2021 to be 0. This could happen if the Null values were replaced by 0 in the dataset.

Q-7 Find the maximum values for confirmed, deaths, and recovered cases per year.



```
# Maximum values of confirmed, deaths, recovered per year
SELECT
    extract(YEAR FROM date) AS year,
    MAX(confirmed) AS max_confirmed,
    MAX(deaths) AS max_deaths,
    MAX(recovered) AS max_recovered
FROM corona_virus_dataset
GROUP BY year
ORDER BY year;
```

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

- The maximum number of confirmed cases in 2020 is comparatively higher than in 2021.
- In 2020, the reported number of deaths is lower than in 2021.
- The number of recovered cases is lower in 2021, possibly due to a decrease in confirmed cases for the same period.

Q-8 Find the total number of confirmed, deaths, and recovered cases per month.

```
# Total number of case of confirmed, deaths, recovered each
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_dataset
GROUP BY months;
```

Result Grid		 Filter Rows:	<input type="text"/>	Export:
	months	total_confirmed	total_deaths	total_recovered
▶	01-2020	6384	190	143
	02-2020	68312	2651	31405
	03-2020	769236	41346	133070
	04-2020	2336798	191833	792987
	05-2020	2744333	144561	1519547
	06-2020	3969634	137757	2535417
	07-2020	6838092	167613	4693120
	08-2020	7694938	179200	6202833
	09-2020	8244794	160671	6647749
	10-2020	11515841	175484	6782150
	11-2020	16595938	262247	9172292
	12-2020	19336799	339996	11924903
	01-2021	18672205	401893	9164347
	02-2021	10492664	298239	6719785
	03-2021	13924790	282620	7888013
	04-2021	21711021	362387	14205507
	05-2021	19121083	366549	19131842
	06-2021	5022282	132657	5544438

- The total number of confirmed and recovered cases peaked in December 2020 during the first wave of COVID-19.
- In April 2021, the total confirmed and recovered cases reported were the highest across both waves.

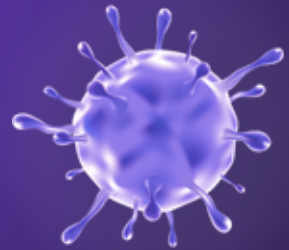


Q-9 Check how corona virus spread out with respect to confirmed cases.

```
# Check how corona virus spread out with respect to confirmed case
SELECT
    SUM(confirmed) AS total_confirmed,
    ROUND(AVG(confirmed), 2) AS avg_confirmed,
    ROUND(VARIANCE(confirmed), 2) AS var_confirmed,
    ROUND(STDDEV(confirmed), 2) AS stddev_confirmed
FROM corona_virus_dataset;
```

Result Grid				
Filter Rows: <input type="text"/>				
Export: <input type="button" value=""/>				
Wrap				
	total_confirmed	avg_confirmed	var_confirmed	stddev_confirmed
▶	169065144	2156.83	157288925.08	12541.49

- The global impact of the virus is evident with a total of 169,065,144 confirmed cases.
- On average, each record in the dataset reports 2,157 confirmed cases, indicating high transmission rates of the virus.
- The high variance (157,288,925.08) and standard deviation (12,541.49) indicate significant variability in the spread of the virus in different areas and at different times.
- A high standard deviation indicates varying numbers of confirmed cases, highlighting localized outbreaks or surges in specific regions or periods.

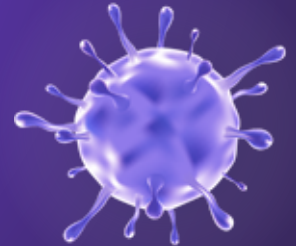


Q-10 Check how corona virus spread out with respect to deaths per month.

```
# Check how corona virus spread out with respect to deaths
SELECT
    DATE_FORMAT(date, '%m-%Y') AS months,
    SUM(deaths) AS total_deaths,
    ROUND(AVG(deaths), 0) AS avg_deaths,
    ROUND(VARIANCE(deaths), 0) AS var_deaths,
    ROUND(STDDEV(deaths), 0) AS stddev_deaths
FROM corona_virus_dataset
GROUP BY months;
```



	months	total_deaths	avg_deaths	var_deaths	stddev_deaths
▶	01-2020	190	0	4	2
	02-2020	2651	1	68	8
	03-2020	41346	9	3901	62
	04-2020	191833	42	40504	201
	05-2020	144561	30	20685	144
	06-2020	137757	30	16929	130
	07-2020	167613	35	21140	145
	08-2020	179200	38	23273	153
	09-2020	160671	35	20103	142
	10-2020	175484	37	17580	133
	11-2020	262247	57	27774	167
	12-2020	339996	71	65345	256
	01-2021	401893	84	102758	321
	02-2021	298239	69	68479	262
	03-2021	282620	59	54386	233
	04-2021	362387	78	94611	308
	05-2021	366549	77	131769	363
	06-2021	132657	66	112964	336

- Significant rise in deaths from 190 in January 2020 to 41,346 in March 2020.
- Highest number of deaths recorded at 339,996 in December 2020
- Average death peaks in December 2020 with an average of 71 deaths.
- Variance and standard deviation peak in December 2020 indicating higher variability.

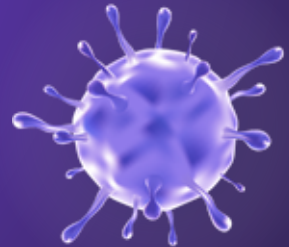


Q-11 Check how the corona virus spread out with respect to recovered cases.

```
# Check how corona virus spread out with respect to recovered cases
SELECT
    SUM(recovered) AS total_recovered,
    ROUND(AVG(recovered), 2) AS avg_recovered,
    ROUND(VARIANCE(recovered), 2) AS var_recovered,
    ROUND(STDDEV(recovered), 2) AS stddev_recovered
FROM corona_virus_dataset;
```

Result Grid		 Filter Rows:	Export:  Wrap C	
	total_recovered	avg_recovered	var_recovered	stddev_recovered
▶	113089548	1442.73	107029523.26	10345.51

- The total number of recovered cases from COVID-19 is 113,089,548, indicating significant progress in treatment and recovery efforts.
- The average number of recovered cases is 1,442.73, providing insight into the recovery rate during the analyzed period.
- The variance in recovered cases is 10,702,952.26. A high variance indicates a wide spread in the number of recovered cases, which can be attributed to fluctuations over time, different waves of infections, and varying recovery rates.
- The standard deviation is 10,345.51, indicating significant fluctuations in the number of recovered cases due to varying impacts of the pandemic over time.



Q-12 Find the country having the highest number of confirmed cases.

```
# Country having highest number of the Confirmed case
SELECT
    country_or_region,
    SUM(confirmed) AS total_confirmed
FROM corona_virus_dataset
GROUP BY country_or_region
ORDER BY total_confirmed DESC
LIMIT 1;
```

Result Grid			Filter Rows:
	country_or_region	total_confirmed	
▶	US	33461982	

- The US (United States) is the country with the highest number of confirmed cases.

Q-13 Find the country having the lowest number of deaths

```
# Find Country having lowest number of the death case
SELECT
    country_or_region,
    SUM(deaths) AS total_deaths
FROM corona_virus_dataset
GROUP BY country_or_region
ORDER BY total_deaths ASC
LIMIT 1;
```

Result Grid			Filter Rows:
	country_or_region	total_deaths	
▶	Dominica	0	

- Dominica is the country with the lowest number of deaths.

Q-14 Find the top 5 countries having the highest number of recovered cases.

```
# Find top 5 countries having highest recovered case
SELECT
    country_or_region,
    SUM(recovered) AS total_recovered
FROM corona_virus_dataset
GROUP BY country_or_region
ORDER BY total_recovered DESC
LIMIT 5;
```

Result Grid		
Filter Rows:		
	country_or_region	total_recovered
▶	India	28089649
	Brazil	15400169
	US	6303715
	Turkey	5202251
	Russia	4745756

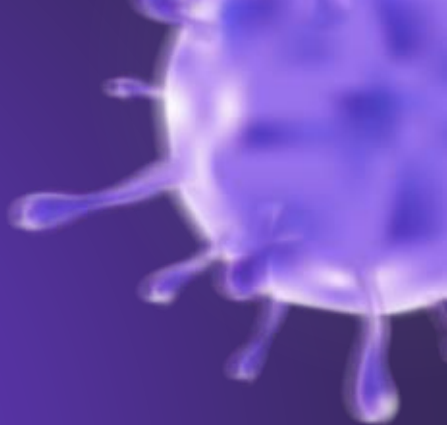
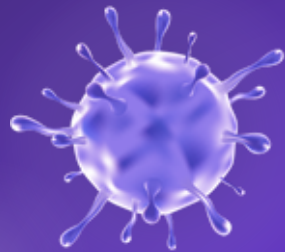
Top 5 countries by total recovered cases:

1. INDIA
2. BRAZIL
3. US
4. TURKEY
5. RUSIA



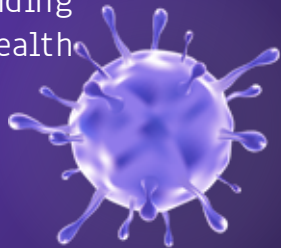
04

Conclusion



CONCLUSION

In summary, our analysis of the coronavirus dataset reveals the extensive impact of COVID-19, with 169 million confirmed cases highlighting its global reach. The high average of 2,156 confirmed cases per record and significant variability underscores the uneven spread and frequent surges. These insights stress the importance of continuous monitoring, targeted interventions, and efficient resource allocation to manage the pandemic effectively. Understanding these patterns helps us to better prepare for and respond to future public health challenges, ensuring more resilient and adaptable healthcare systems.





THANKS!

Does anyone have any questions?

Mail me at:
buntypatil1305@gmail.com



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**

