

The background of the slide features several stylized, semi-transparent virus particles. These particles are spherical with a light blue base color and are covered with numerous small, colorful protrusions in shades of green, pink, and purple, resembling the surface of a coronavirus. They are scattered across the slide, with some appearing larger and more prominent than others.

CORONA VIRUS ANALYSIS

**Year 2020 & 2021
SQL Analysis**

Surathni Gunawardena



TABLE OF CONTENT

01

INTRODUCTION

02

Problem
Statement

03

Dataset Overview

04

Dataset
Explanation

05

Data Exploration
using SQL



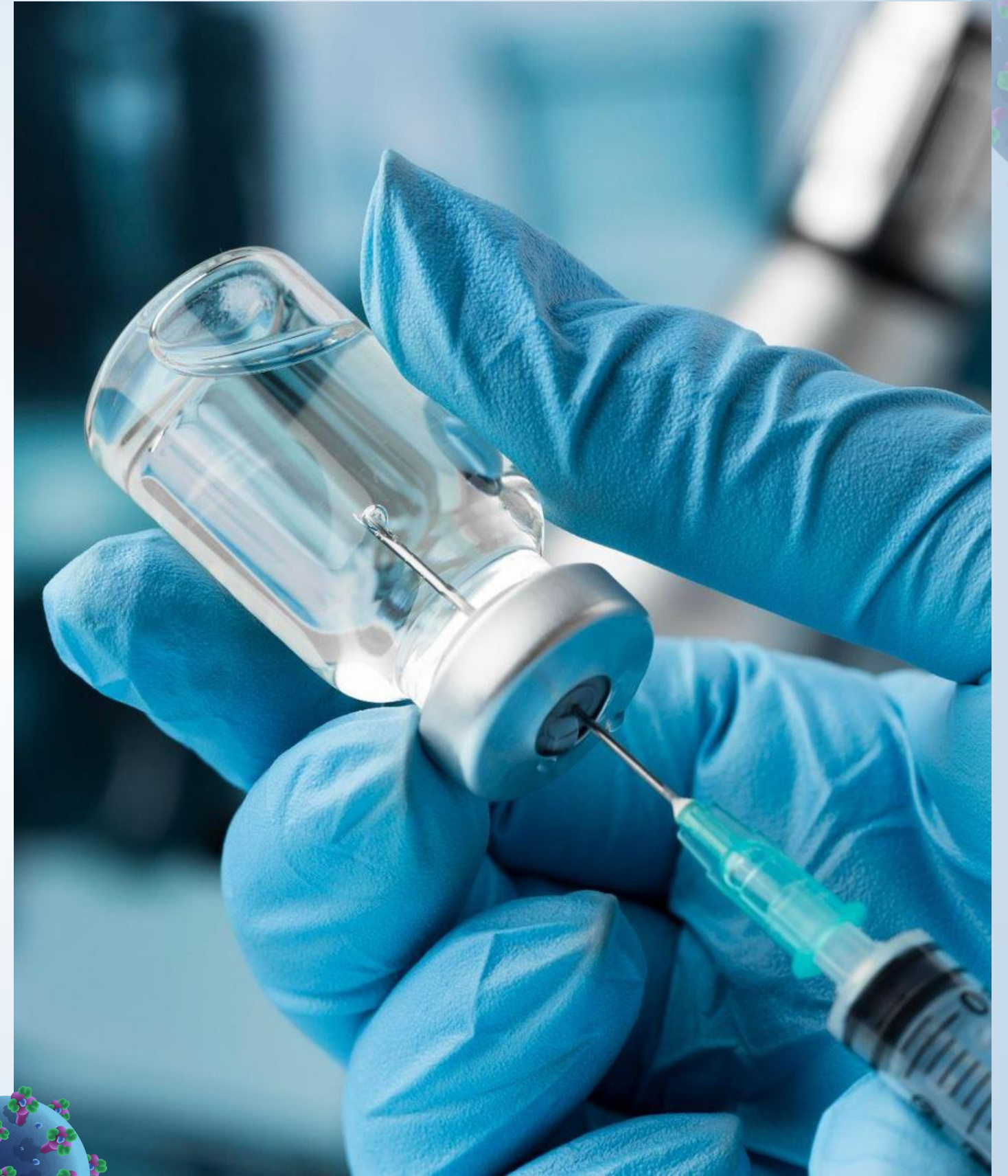
INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by the coronavirus SARS-CoV-2. The first known case was identified in Wuhan, China, in December 2019. The disease quickly spread worldwide, resulting in the COVID-19 pandemic.



PROBLEM STATEMENT

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. This analysis uses SQL to showcase the key insights of its spread globally in year 2020 & 2021.



DATASET OVERVIEW

Results Messages								
	Province	Country_Region	Latitude	Longitude	Date	Confirmed	Deaths	Recovered
1	Afghanistan	Afghanistan	33.93911	67.709953	22-01-2020	0	0	0
2	Afghanistan	Afghanistan	33.93911	67.709953	23-01-2020	0	0	0
3	Afghanistan	Afghanistan	33.93911	67.709953	24-01-2020	0	0	0
4	Afghanistan	Afghanistan	33.93911	67.709953	25-01-2020	0	0	0
5	Afghanistan	Afghanistan	33.93911	67.709953	26-01-2020	0	0	0
6	Afghanistan	Afghanistan	33.93911	67.709953	27-01-2020	0	0	0
7	Afghanistan	Afghanistan	33.93911	67.709953	28-01-2020	0	0	0
8	Afghanistan	Afghanistan	33.93911	67.709953	29-01-2020	0	0	0
9	Afghanistan	Afghanistan	33.93911	67.709953	30-01-2020	0	0	0
10	Afghanistan	Afghanistan	33.93911	67.709953	31-01-2020	0	0	0
11	Afghanistan	Afghanistan	33.93911	67.709953	01-02-2020	0	0	0

Results Messages								
	Province	Country_Region	Latitude	Longitude	Date	Confirmed	Deaths	Recovered
78376	Zimbabwe	Zimbabwe	-19.015...	29.154857	03-06-2021	61	5	19
78377	Zimbabwe	Zimbabwe	-19.015...	29.154857	04-06-2021	52	1	10
78378	Zimbabwe	Zimbabwe	-19.015...	29.154857	05-06-2021	24	0	8
78379	Zimbabwe	Zimbabwe	-19.015...	29.154857	06-06-2021	21	1	30
78380	Zimbabwe	Zimbabwe	-19.015...	29.154857	07-06-2021	49	5	18
78381	Zimbabwe	Zimbabwe	-19.015...	29.154857	08-06-2021	83	6	10
78382	Zimbabwe	Zimbabwe	-19.015...	29.154857	09-06-2021	111	5	161
78383	Zimbabwe	Zimbabwe	-19.015...	29.154857	10-06-2021	64	4	23
78384	Zimbabwe	Zimbabwe	-19.015...	29.154857	11-06-2021	192	3	30
78385	Zimbabwe	Zimbabwe	-19.015...	29.154857	12-06-2021	164	3	22
78386	Zimbabwe	Zimbabwe	-19.015...	29.154857	13-06-2021	107	0	12

DATASET EXPLANATION

- Province: Geographic subdivision within a country/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 CORONA VIRUS data.
- Confirmed: Number of diagnosed CORONA VIRUS cases.
- Deaths: Number of CORONA VIRUS related deaths.
- Recovered: Number of recovered CORONA VIRUS cases

DATA EXPLORATION USING SQL



CODE TO CHECK NULL VALUES

```
SELECT *
FROM [Corona Virus Dataset]
WHERE Province is null or Country_Region is null or
Latitude is null or Longitude is null or Date is null or
Confirmed is null or Deaths is null
or Recovered is null;

SELECT
    SUM(CASE WHEN Province IS NULL THEN 1 ELSE 0 END) AS Province_Null_Count,
    SUM(CASE WHEN Country_Region IS NULL THEN 1 ELSE 0 END) AS Countryregion_Null_Count,
    SUM(CASE WHEN Latitude IS NULL THEN 1 ELSE 0 END) AS Latitude_Null_Count,
    SUM(CASE WHEN Longitude IS NULL THEN 1 ELSE 0 END) AS Longitude_Null_Count,
    SUM(CASE WHEN Date IS NULL THEN 1 ELSE 0 END) AS Date_Null_Count,
    SUM(CASE WHEN Confirmed IS NULL THEN 1 ELSE 0 END) AS Confirmed_Null_Count,
    SUM(CASE WHEN Deaths IS NULL THEN 1 ELSE 0 END) AS Deaths_Null_Count,
    SUM(CASE WHEN Recovered IS NULL THEN 1 ELSE 0 END) AS Recovered_Null_Count
FROM [Corona Virus Dataset];
```

110 %

Results Messages

	Province	Country_Region	Latitude	Longitude	Date	Confirmed	Deaths	Recovered
	Province_Null_Count	Countryregion_Null_Count	Latitude_Null_Count	Longitude_Null_Count	Date_Null_Count	Confirmed_Null_Count	Deaths_Null_Count	Recovered_Null_Count
1	0	0	0	0	0	0	0	0

UPDATING THE NULL VALUE WITH 0

```
--Q2--  
UPDATE [Corona Virus Dataset]  
SET Province = COALESCE(Province, 0),  
    Country_Region = COALESCE(Country_Region, 0),  
    Latitude = COALESCE(Latitude, 0),  
    Longitude = COALESCE(Longitude, 0),  
    Date = COALESCE(Date, 0),  
    Confirmed = COALESCE(Confirmed, 0),  
    Deaths = COALESCE(Deaths, 0),  
    Recovered = COALESCE(Recovered, 0);
```

	Province_Null_Count	Countryregion_Null_Count	Latitude_Null_Count	Longitude_Null_Count	Date_Null_Count	Confirmed_Null_Count	Deaths_Null_Count	Recovered_Null_Count
1	0	0	0	0	0	0	0	0

TOTAL NUMBER OF ROWS

```
--Q3--  
SELECT COUNT(*) AS Total_Rows  
FROM [Corona Virus Dataset];
```

110 %

Results Messages

	Total_Rows
1	78386

**TOTAL
78386**

START DATE & END DATE

```
SQLQuery2.sql - D...54R8FN6\USER (68))*  
  
--Q4--  
  
SELECT  
MIN(TRY_CONVERT(date, Date, 103)) AS start_date,  
MAX(TRY_CONVERT(date, Date, 103)) AS end_date  
FROM [Corona Virus Dataset];
```

110 %

Results Messages

	start_date	end_date
1	2020-01-22	2021-06-13

Start Date

2020/01/22

End Date

2021/06/13

NUMBER OF MONTHS IN THE DATASET

```
--Q5--  
SELECT COUNT(DISTINCT CONCAT(YEAR(TRY_CONVERT(date,Date,105)), '-',  
MONTH(TRY_CONVERT(date,Date,105)))) AS NumberOfMonths  
FROM [Corona Virus Dataset];
```

110 %

Results Messages

	NumberOfMonths
1	18

Total
Months 18

MONTHLY AVERAGE

~vs74D4.sql - DESK...54R8FN6\USER (62))*

```
--Q6--
SELECT
    DATEPART(YEAR, TRY_CONVERT(date, Date,105)) AS Year,
    DATEPART(MONTH, TRY_CONVERT(date, Date,105)) AS Month,
    AVG(CONVERT(float, TRY_CONVERT(int, Confirmed))) AS Average_Confirmed,
    AVG(CONVERT(float, TRY_CONVERT(int, Deaths))) AS Average_Deaths,
    AVG(CONVERT(float, TRY_CONVERT(int, Recovered))) AS Average_Recovered
FROM
    [Corona Virus Dataset]
WHERE
    DATEPART(YEAR, TRY_CONVERT(date, Date,105)) IN (2020, 2021)
GROUP BY
    DATEPART(YEAR, TRY_CONVERT(date, Date,105)),
    DATEPART(MONTH, TRY_CONVERT(date, Date,105))
ORDER BY
    Year, Month;
```

110 %

	Year	Month	Average_Confirmed	Average_Deaths	Average_Recovered
1	2020	1	4.14545454545455	0.123376623376623	0.0928571428571429
2	2020	2	15.2960143304971	0.5935960591133	7.0320197044335
3	2020	3	161.130289065773	8.66066191872643	27.8739002932551
4	2020	4	505.800432900433	41.5222943722944	171.642207792208
5	2020	5	574.849811478844	30.280896522832	318.296397151236
6	2020	6	859.228138528139	29.8175324675325	548.791558441558
7	2020	7	1432.36112274822	35.109551738584	983.05823209049
8	2020	8	1611.84289903645	37.5366568914956	1299.29472140762
9	2020	9	1784.58744588745	34.7772727272727	1438.90670995671

Query executed successfully. | DESKTOP-54R8FN6\SQLEXPRESS ... | D

MINIMUM / MAXIMUM ? VALUES PER YEAR

```
SELECT
    YEAR(TRY_CONVERT(date, Date,105)) AS Year,
    MIN (TRY_CONVERT (int, Confirmed)) AS min_confirmed,
    MIN (TRY_CONVERT (int, Deaths)) AS min_deahs,
    MIN (Recovered) AS min_recovered

FROM
    [Corona Virus Dataset]
GROUP BY
    YEAR(TRY_CONVERT(date, Date,105));
```

	Year	min_confirmed	min_deahs	min_recovered
1	2021	0	0	0
2	2020	0	0	0

```
SELECT
    YEAR(TRY_CONVERT(date, Date,105)) AS Year,
    MAX (TRY_CONVERT (int, Confirmed)) AS max_confirmed,
    MAX (TRY_CONVERT (int, Deaths)) AS max_deahs,
    MAX (Recovered) AS max_recovered

FROM
    [Corona Virus Dataset]
GROUP BY
    YEAR(TRY_CONVERT(date, Date,105));
```

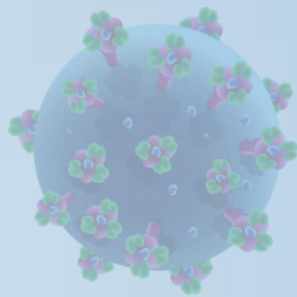
	Year	max_confirmed	max_deahs	max_recovered
1	2021	414188	7374	422436
2	2020	823225	3752	1123456

TOTAL NUMBER OF CASES IN EACH MONTH

```
SELECT
    YEAR(TRY_CONVERT(date, Date,105)) AS Year,
    MONTH(TRY_CONVERT(date, Date,105)) AS Month,
    SUM (TRY_CONVERT (int, Confirmed)) AS Total_confirmed,
    SUM (TRY_CONVERT (int, Deaths)) AS Total_deaths,
    SUM (TRY_CONVERT (INT, Recovered)) AS Total_recovered
FROM
    [Corona Virus Dataset]
GROUP BY
    YEAR(TRY_CONVERT(date, Date,105)),
    MONTH(TRY_CONVERT(date, Date,105))
ORDER BY
    Year, Month;
```

	Year	Month	Total_confirmed	Total_deaths	Total_recovered
1	2020	1	6384	190	143
2	2020	2	68312	2651	31405
3	2020	3	769236	41346	133070
4	2020	4	2336798	191833	792987
5	2020	5	2744333	144561	1519547
6	2020	6	3969634	137757	2535417
7	2020	7	6838092	167613	4693120
8	2020	8	7694938	179200	6202833
9	2020	9	8244794	160671	6647749
10	2020	10	11515841	175484	6782150
11	2020	11	16595938	262247	9172292
12	2020	12	19336799	339996	11924903
13	2021	1	18672205	401893	9164347
14	2021	2	10492664	298239	6719785
15	2021	3	13924790	282620	7888013
16	2021	4	21711021	362387	14205507
17	2021	5	19121083	366549	19131842
18	2021	6	5022282	132657	5544438

VIRUS SPREAD WITH RESPECT TO CONFIRMED CASES

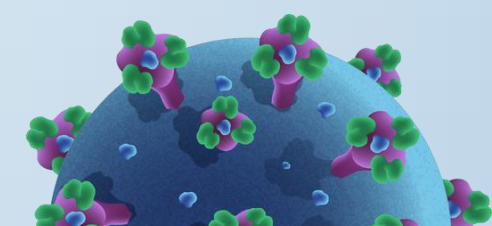


```
SELECT
    COUNT(Confirmed) AS TotalConfirmedCases,
    AVG(TRY_CONVERT(BIGINT, Confirmed)) AS AverageConfirmedCases,
    SUM(POWER(TRY_CONVERT(BIGINT, Confirmed) - AVG_Cases.AverageConfirmedCases, 2))/COUNT(Confirmed) AS VarianceConfirmedCases,
    SQRT(SUM(POWER(TRY_CONVERT(BIGINT, Confirmed) - AVG_Cases.AverageConfirmedCases, 2))/COUNT(Confirmed)) AS StdDevConfirmedCases
FROM
    [Corona Virus Dataset],
    (SELECT AVG(TRY_CONVERT(BIGINT, Confirmed)) AS AverageConfirmedCases FROM [Corona Virus Dataset]) AS AVG_Cases;
```

110 %

Results Messages

	TotalConfirmedCases	AverageConfirmedCases	VarianceConfirmedCases	StdDevConfirmedCases
1	78386	2156	157288925	12541.4881493386



VIRUS SPREAD WITH RESPECT TO RECOVERED CASES

```
SELECT
    COUNT(Recovered) AS TotalRecoveredCases,
    AVG(TRY_CONVERT(BIGINT, Recovered)) AS AverageRecoveredCases,
    SUM(POWER(TRY_CONVERT(BIGINT, Recovered) - AVG_Cases.AverageRecoveredCases, 2))/COUNT(Recovered) AS VarianceRecoveredCases,
    SQRT(SUM(POWER(TRY_CONVERT(BIGINT, Recovered) - AVG_Cases.AverageRecoveredCases, 2))/COUNT(Recovered)) AS StdDevRecoveredCases
FROM
    [Corona Virus Dataset],
    (SELECT AVG(TRY_CONVERT(BIGINT, Recovered)) AS AverageRecoveredCases FROM [Corona Virus Dataset]) AS AVG_Cases;
```

10 %

Results Messages

	TotalRecoveredCases	AverageRecoveredCases	VarianceRecoveredCases	StdDevRecoveredCases
1	78386	1442	107029523	10345.5073824342

COUNTRY HAVING HIGHEST NUMBER OF CONFIRMED CASES

```
SELECT TOP 1
    Country_Region AS Country,
    SUM(CAST(Confirmed AS INT)) AS TotalConfirmedCases

FROM
    [Corona Virus Dataset]
WHERE
    ISNUMERIC(Confirmed)=1
GROUP BY
    Country_Region
ORDER BY
    SUM(CAST(Confirmed AS INT)) DESC;
```

	Country	TotalConfirmedCases
1	US	33461982



USA

COUNTRY HAVING LOWEST NUMBER OF DEATH CASES

```
Country_Region AS Country,  
SUM(CAST(Deaths AS INT)) AS TotalDeathCases  
  
FROM  
[Corona Virus Dataset]  
WHERE  
ISNUMERIC(Deaths)=1  
GROUP BY  
Country_Region  
HAVING  
SUM(CAST(Confirmed AS INT)) = 0  
ORDER BY  
TotalDeathCases ASC;
```

Results	Messages
Country	TotalDeathCases



TOP 5 COUNTRIES HAVING HIGHEST RECOVERED CASES

```
SELECT TOP 5
    Country_Region AS Country,
    sum(cast(Recovered AS INT)) AS TotalRecoveredCases
FROM
    [Corona Virus Dataset]
WHERE
    ISNUMERIC(recovered) =1
GROUP BY
    Country_Region
ORDER BY
    TotalRecoveredCases DESC;
```

110 %

Results Messages

	Country	TotalRecoveredCases
1	India	28089649
2	Brazil	15400169
3	US	6303715
4	Turkey	5202251
5	Russia	4745756

- INDIA
- BRAZIL
- US
- TURKEY
- RUSSIA

THANK
YOU!

