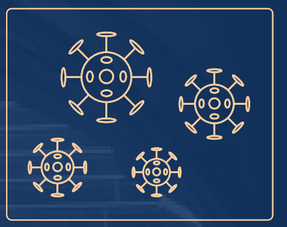
Corona Virus Analysis

By- Roopam Barua

Table of Contents

About the Project



Coronavirus disease 2019 (COVID-19) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first identified in Wuhan, China, in late 2019. It quickly spread globally, leading to a pandemic declared by the World Health Organization (WHO) in March 2020.

COVID-19 primarily spreads through respiratory droplets when an infected person coughs, sneezes, or talks. It can also spread by touching surfaces contaminated with the virus and then touching the face.

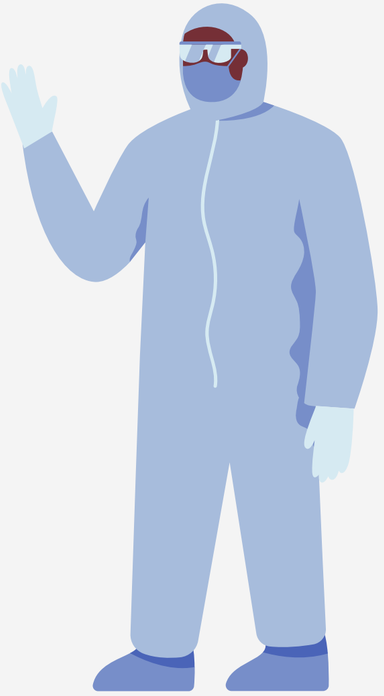
Symptoms can range from mild to severe and include fever, cough, fatigue, shortness of breath, loss of taste or smell, and in severe cases, pneumonia and acute respiratory distress syndrome (ARDS).

The pandemic has had profound global consequences, affecting health systems, economies, and daily life worldwide. Lockdowns, travel restrictions, and social distancing measures have been implemented to curb the spread of the virus. The Objective of this analysis is to discover the global impact of the virus in order to derive meaningful insights which will help aid decision making to control and manage the spread of the virus.

Data Set Description

Data set is of c.s.v file type, imported it to MYSQL using table data import wizard for analysis

- **Provinces**
- **Country/Region**
- **Latitude**
- **Longitude**
- **Date**
- **Confirmed**
- **Deaths**
- **Recovered**



Exploratory Data Analysis

To avoid any errors, check missing value / null value

Q1. Write a code to check NULL values

```
14 •  select * from corona_virus_dataset where
15    Provinces is Null or
16    Country is Null or
17    Latitude is Null or
18    Longitude is Null or
19    Date is Null or
20    Deaths is Null or
21    Recovered  IS null;
```

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

Provinces	Country	Latitude	Longitude	Date	confirmed	Deaths	Recovered

The table is showing blank because there is no null value

Q2. If NULL values are present, update them with zeros for all columns.

No Null value is present in the data set

Q3. check total number of rows

```
24      #Count the total no of rows in the dataset
25 •    select count(*) from corona_virus_dataset;
```

Result Grid | Filter Rows: | Export: | Wrap

count(*)
78386

Q4. Check what is the start_date and end_date

```
27      # select the Start date and End date ( Date value was in string, converted it into Date value)#
28 •    update corona_virus_dataset set Date=str_to_date(Date, "%d-%m-%Y");
29 •    select Min(Date) as Start_Date, Max(Date) as End_Date from corona_virus_dataset;
```

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

	Start_Date	End_Date
▶	2020-01-22	2021-06-13

First I had to convert the date type from string to date value as during table creation the date type was varchar

Q5. Number of month present in dataset

```
31      # select the total no of months#
32 •    set @Start_date = "2020-01-22";
33 •    set @End_date = "2021-06-13";
34 •    select timestampdiff(month, @Start_date, @End_date) as Number_of_months;
35
```

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

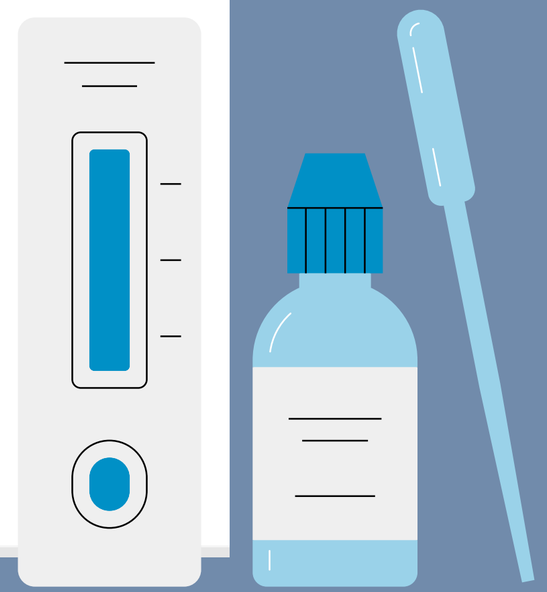
	Number_of_months
▶	16

Q6. Find monthly average for confirmed, deaths, recovered

```
36      #Find monthly average for confirmed, deaths, recovered
37 •    select year(Date) as YEARS , month(Date) as MONTHS, Floor(Avg(confirmed)) as AVG_CONFIRM,
38      Floor(Avg(Deaths)) as AVG_DEATH, Floor(Avg(Recovered)) as AVG_RECOVER
39      from corona_virus_dataset group by year(Date), month(Date);
```

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

YEARS	MONTHS	AVG_CONFIRM	AVG_DEATH	AVG_RECOVER
2020	1	4	0	0
2020	2	15	0	7
2020	3	161	8	27
2020	4	505	41	171
2020	5	574	30	318
2020	6	859	29	548
2020	7	1432	35	983
2020	8	1611	37	1299
2020	9	1784	34	1438
2020	10	2412	36	1420
2020	11	3592	56	1985
2020	12	4050	71	2497
2021	1	3911	84	1919



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

```
41      ## Find most frequent value for confirmed, deaths, recovered each month
42 •    SELECT month(date) as month, confirmed, Deaths, recovered, COUNT(*) AS frequency
43      FROM corona_virus_dataset
44      GROUP BY confirmed, Deaths, recovered, month
45      ORDER BY frequency desc;
```

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

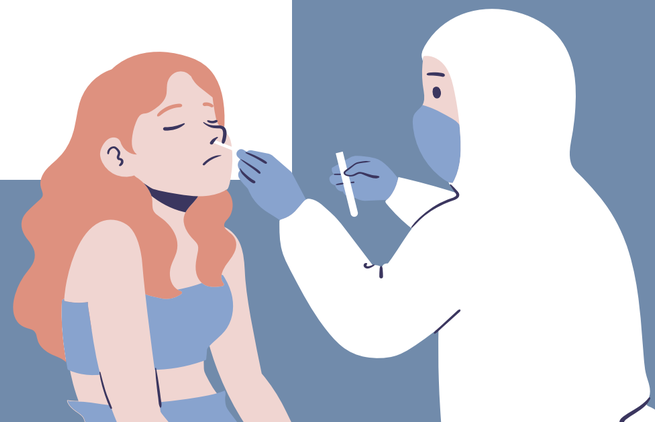
month	confirmed	Deaths	recovered	frequency
2	0	0	0	4781
3	0	0	0	3138
1	0	0	0	2416
5	0	0	0	2395
4	0	0	0	2073
6	0	0	0	1759
7	0	0	0	1245
9	0	0	0	1135
8	0	0	0	1116
10	0	0	0	1112
12	0	0	0	1087

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

```
47      ## Find minimum values for confirmed, deaths, recovered per year
48 •    select year(Date), min(confirmed), min(Deaths), min(recovered)
49      from corona_virus_dataset group by year(Date);
```

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

year(Date)	min(confirmed)	min(Deaths)	min(recovered)
2020	0	0	0
2021	0	0	0



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

```
51      ## Find maximum values of confirmed, deaths, recovered per year
52 •    select year(Date), max(confirmed), max(Deaths), max(recovered)
53      from corona_virus_dataset group by year(Date);
```

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

year(Date)	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

```
55    ## The total number of case of confirmed, deaths, recovered each month
56 •  select year(Date) as Year, month(Date) as Months, sum(confirmed) as Total_confirmed,
57    sum(Deaths) as Total_Deaths, sum(recovered) as Total_recovered
58    from corona_virus_dataset group by year(Date), month(Date);
```

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

Year	Months	Total_confirmed	Total_Deaths	Total_recovered
2020	1	6384	190	143
2020	2	68312	2651	31405
2020	3	769236	41346	133070
2020	4	2336798	191833	792987
2020	5	2744333	144561	1519547
2020	6	3969634	137757	2535417
2020	7	6838092	167613	4693120
2020	8	7694938	179200	6202833
2020	9	8244794	160671	6647749
2020	10	11515841	175484	6782150

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

```
62      ##  Check how corona virus spread out with respect to confirmed case
63 •    select sum(confirmed) as Total_confirmed, avg(confirmed), variance(confirmed),
64       stddev(confirmed) from corona_virus_dataset;
```

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

Total_confirmed	avg(confirmed)	variance(confirmed)	stddev(confirmed)
169065144	2156.8283	157288925.07796532	12541.488152446875

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

```
66      ##  Check how corona virus spread out with respect to death case per month
67      --      (Eg.: total confirmed cases, their average, variance & STDEV )
68 •    select year(Date) as Year, month(Date) as Months, round(sum(Deaths),2)
69      as Total_Deaths, round(avg(Deaths),2) as Avg_Deaths, round(variance(Deaths),2)
70      as Variance_Deaths, round(stddev(Deaths),2) as Stddev_Deaths
71      from corona_virus_dataset group by year(Date), month(Date);
```

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

Year	Months	Total_Deaths	Avg_Deaths	Variance_Deaths	Stddev_Deaths
2020	1	190	0.12	4.25	2.06
2020	2	2651	0.59	68.32	8.27
2020	3	41346	8.66	3900.79	62.46
2020	4	191833	41.52	40504.27	201.26
2020	5	144561	30.28	20684.91	143.82
2020	6	137757	29.82	16929.45	130.11
2020	7	167613	35.11	21140.15	145.4
2020	8	179200	37.54	23273	152.55
2020	9	160671	34.78	20102.77	141.78
2020	10	175484	36.76	17580.07	132.59

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

```
73      ## Check how corona virus spread out with respect to recovered case
74      --      (Eg.: total confirmed cases, their average, variance & STDEV )
75  •   select sum(recovered) as Total_recovered, round(avg(recovered),2)
76      as Avg_recovered, round(variance(recovered),2) as variance_recovered,
77      round(stddev(recovered),2) as stddev_recovered
78      from corona_virus_dataset;
```

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

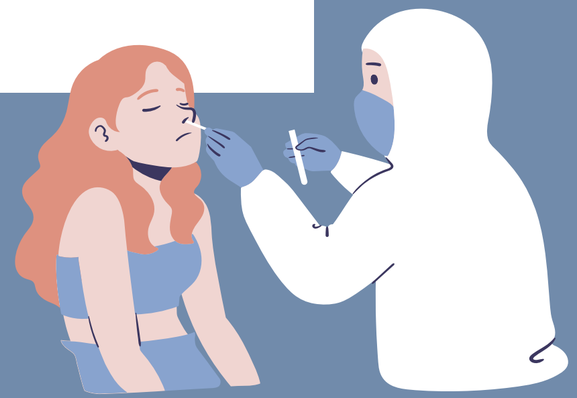
Total_recovered	Avg_recovered	variance_recovered	stddev_recovered
113089548	1442.73	107029523.26	10345.51

Q14. Find Country having highest number of the Confirmed case

```
80      ## Find Country having highest number of the Confirmed case
81  •   select country, sum(confirmed) as total_confirmed from
82      corona_virus_dataset group by Country order by total_confirmed desc limit 1;
```

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

country	total_confirmed
US	33461982



Q15. Find Country having lowest number of the death case

```
84      ## Find Country having lowest number of the death case
85  •   select country, sum(Deaths) as Total_Deaths
86      from corona_virus_dataset group by Country
87      order by Total_Deaths asc limit 4;
```

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

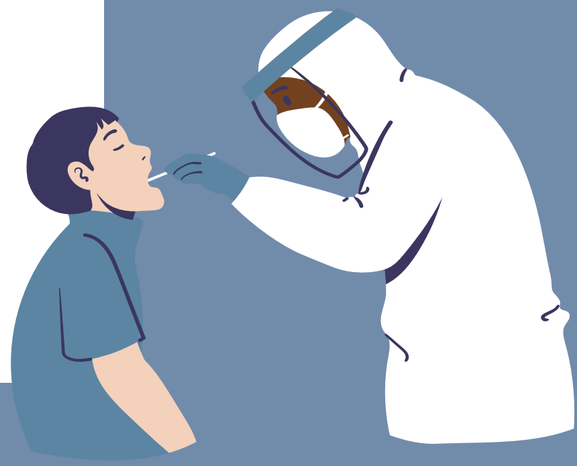
country	Total_Deaths
Dominica	0
Marshall Islands	0
Kiribati	0
Samoa	0

Q16. Find top 5 countries having highest recovered case

```
89      ## Find top 5 countries having highest recovered case
90  •   select country, sum(recovered) as Total_recovered
91      from corona_virus_dataset group by country
92      order by Total_recovered desc limit 5;
```

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

country	Total_recovered
India	28089649
Brazil	15400169
US	6303715
Turkey	5202251
Russia	4745756



Insights

- The result from the analysis indicates a high variability in the number of confirmed cases and recovered cases, which proves that the spread of the virus is uneven, with some areas or time periods experiencing higher or lower cases compared to others.

- January 2020 recorded the least total death cases and least total confirmed cases of 190 and 6,384, respectively. The low variability indicates significant consistency and stability. This means the death cases and confirmed cases were relatively similar in different regions, proving a more uniform spread of the virus in terms of fatalities during the month. However, the regions witnessed a significant increase and decrease in the spread of the virus from February 2020 to May 2021.

- Marshall Islands, Samoa, Dominica, and Kiribati had the lowest count of death cases with a total of 0 respectively.

Insights

• The United States has the highest number of confirmed cases, with 33.5 million people infected with the virus. India recorded at most 29.6 million confirmed cases, and Brazil had a total of 17.4 million confirmed cases.

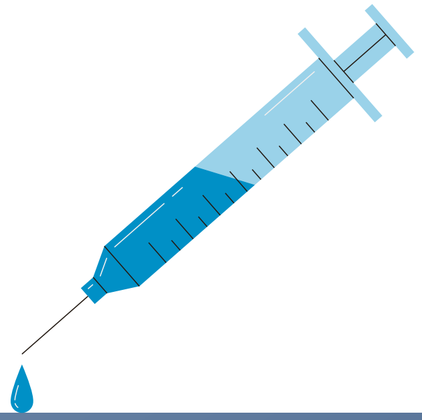
• The top five countries with the highest recovered cases are India, Brazil, United States, Turkey, and Russia.

• The top three highest average confirmed cases occurred in April 2021 with a total of 4,699, December 2020 with a total of 4,050, and May 2021 with a total of 4,005.

• The top three highest average recovered cases occurred in May 2021 with a total of 4,007, April 2021 with a total of 3,074, and June 2021 with a total of 2,769.

• The year 2020 had the maximum number of confirmed cases and recovered cases, whereas 2021 recorded more death cases.

• The highest number of deaths were reported in January 2021.



Recommendation

• Identify and focus on high-risk areas using regulations such as travel restrictions, and quarantine policies.

• Incorporate the use of predictive modelling to forecast potential outbreaks and take adequate measures to prevent them.

• Distribute resources such as medical equipment, healthcare workers, and hospital beds to areas with greater number of cases in order to control and manage the spread of the virus effectively.