

# **COVID-19 Data Analysis**

An analysis of COVID-19 data

# Introduction

## 1 Purpose of the analysis

The analysis aims to ensure data accuracy and uncover insights from the COVID-19 dataset, helping to understand trends and statistics.

## 2 Brief overview of the dataset

The dataset contains COVID-19 related information such as confirmed cases, deaths, and recoveries, along with other details.



# Checking for Missing Values

- Query: Check for NULL values in the dataset

## The Query

```
SELECT *  
FROM [Corona Virus].[dbo].[Corona Virus  
Dataset]  
--where Province , Country , Region ,Latitude  
,Longitude, Deaths, Recovered ,Confirmed  
,Date is null
```



# Handling Missing Values

- Query: Update NULL values with zeros for all columns

## The Query

```
update [Corona Virus].[dbo].[Corona Virus  
Dataset]  
set Confirmed = coalesce(Confirmed,0)
```



# Total Number of Rows

- Query: Calculate the total number of rows in the dataset

The Query:

```
select count(*) as total_Rows  
from [Corona Virus].[dbo].[Corona Virus  
Dataset]
```

The Output

```
total_Rows  
-----  
28543
```

# New Columns for Calculations

- Query: convert year to Date time
- Query : Convert Deaths, confirmed and Recovered to float

## The Query:

--Made three new columns to combine

```
alter table [Corona Virus].[dbo].[Corona Virus Dataset]
add day_column int,
    month_column int,
    year_column int
```

--insert in each column from the date column

```
update [Corona Virus].[dbo].[Corona Virus Dataset]
```

```
set day_column = CAST(SUBSTRING(Date, 1, 2) AS int),
    month_column = CAST(SUBSTRING(Date,4,2) as int),
    year_column = CAST(SUBSTRING(Date, 7, 4) AS int);
```

--make the converted Date column to the right data type to make calculation

```
ALTER TABLE [Corona Virus].[dbo].[Corona Virus Dataset]
ADD Converted_Date date;
```

--updated and combine the three column

```
UPDATE [Corona Virus].[dbo].[Corona Virus Dataset]
SET Converted_Date= TRY_CONVERT(datetime,
    CONCAT(year_column, '-', month_column, '-', day_column));
```

- make new columns to convert

```
alter table [Corona Virus].[dbo].[Corona Virus Dataset]
```

```
add Converted_Deaths float,
converted_Confirmed float,
Converted_Recovered float;
```

-- update them with new data type to make calculations

```
UPDATE [Corona Virus].[dbo].[Corona Virus Dataset]
SET Converted_Deaths = TRY_CAST(Deaths AS int),
    converted_Confirmed = TRY_CAST(Confirmed AS int),
    Converted_Recovered = TRY_CAST(Recovered AS int);
```

-- Delete any zeros and null values

```
DELETE FROM [Corona Virus].[dbo].[Corona Virus Dataset]
WHERE Converted_Deaths IS NULL OR Converted_Deaths = '0'
    OR converted_Confirmed IS NULL OR converted_Confirmed = '0'
    OR Converted_Recovered IS NULL OR Converted_Recovered = '0';
```

# Start and End Dates

- Queries: Extract start and end dates from the dataset

## The Query:

-

--Start\_date

```
select MIN(Converted_Date) as Start_date  
from [Corona Virus].[dbo].[Corona Virus Dataset]
```

--- End\_date

```
select MAX(Converted_Date) as End_date  
from [Corona Virus].[dbo].[Corona Virus Dataset]
```

## The Output:

Start\_date

-----

2020-01-24

(1 row affected)

End\_date

-----

2021-06-13

# Number of Months Present

- Query: Calculate the number of unique months present in the dataset

The Query:

```
SELECT COUNT(DISTINCT  
MONTH(Converted_Date)) AS  
num_of_months  
FROM [Corona Virus].[dbo].[Corona Virus  
Dataset];
```

The Output :

num_of_months
12



# Monthly Averages for Confirmed, Deaths, and Recovered Cases

- Query: Calculate monthly averages for confirmed, deaths, and recovered cases

## The Query:

```
SELECT
  MONTH(Converted_Date) AS month,
  ROUND(AVG(converted_Confirmed), 2) AS
avg_confirmed,
  ROUND(AVG(Converted_Deaths), 2) AS
avg_deaths,
  ROUND(AVG(Converted_Recovered), 2) AS
avg_recovered
FROM
  [Corona Virus].[dbo].[Corona Virus Dataset]
GROUP BY
  MONTH(Converted_Date)
ORDER BY
  MONTH(Converted_Date);
```

## The Output:

month	avg_confirmed	avg_deaths	avg_recovered
1	4264.66	109.31	4136.88
2	3204	92.97	3244.3
3	4298.07	96.76	3057.52
4	5602.6	129.16	4081.27
5	5240.71	122.16	5470.12
6	3338.45	102.22	3238.9
7	3733.25	92.22	2598.07
8	3756.91	88.93	3162.75
9	3801.94	77.88	3346.48
10	4627.9	75.51	3298.81
11	6932.82	108.08	4434.35
12	6180.77	116.71	5502.98

# Most Frequent Values for Confirmed, Deaths, and Recovered Cases Each Month

- Query: Identify the most frequent values for confirmed, deaths, and recovered cases each month

## The Output :

month	most_freq_confirmed	most_freq_Deaths	most_freq_Recoverd
1	126	2	39
2	1	1	14
3	24	1	1
4	9	1	2
5	1	1	1
6	20	3	5
7	21	1	5
8	145	22	989
9	2	1	1
10	19	1	8
11	66	4	51
12	1024	10	2206

## The Query:

```
WITH MonthFreq AS(
Select MONTH(Converted_Date) as month,
converted_Confirmed as Confirmed,
Converted_Deaths as Deaths,
Converted_Recovered as Recoverd,

ROW_NUMBER() over ( partition by
month(Converted_Date) order by count (*)desc)
as rn

from[Corona Virus].[dbo].[Corona Virus Dataset]

group by
MONTH(Converted_Date),converted_Confirmed ,
Converted_Deaths,Converted_Recovered
)

select month,Confirmed as
most_freq_confirmed
,Deaths as most_freq_Deaths
,Recoverd as most_freq_Recoverd

from MonthFreq

where rn = 1;
```

# Minimum Values for Confirmed, Deaths, and Recovered Cases Per Year

The Output :

year	min_Confirmed	min_Deaths	min_Recovered
2021	1	1	1
2020	1	1	1

- Queries: Find the minimum values for confirmed, deaths, and recovered cases each year

The Query :

```
select  
YEAR(Converted_Date) as year,  
  
min(converted_Confirmed) as  
min_Confirmed,  
  
min(Converted_Deaths) as min_Deaths,  
  
min(Converted_Recovered) as min_Recovered  
from [Corona Virus].[dbo].[Corona Virus Dataset]  
group by YEAR(Converted_Date);
```

# Maximum Values for Confirmed, Deaths, and Recovered Cases Per Year

The Output :

year	max_Confirmed	max_Deaths	max_Recovered
2021	414188	7374	422436
2020	823225	3410	1123456

- Queries: Find the maximum values for confirmed, deaths, and recovered cases each year

The Query :

```
select  
YEAR(Converted_Date) as year,  
max(converted_Confirmed) as  
max_Confirmed,  
  
max(Converted_Deaths) as max_Deaths,  
  
max(Converted_Recovered) as  
max_Recovered  
  
from [Corona Virus].[dbo].[Corona Virus  
Dataset]  
group by YEAR(Converted_Date);
```

# Total Number of Cases of Confirmed, Deaths, and Recovered Each Month

- Query: Calculate the total number of confirmed, deaths, and recovered cases each month

## The Query :

```
select
month(Converted_Date) as month,
sum(converted_Confirmed)asTotal_Confirmed,

sum(Converted_Deaths) as Total_Deaths,

sum(Converted_Recovered) as Total_Recovered

from [Corona Virus].[dbo].[Corona Virus Dataset]

group by month(Converted_Date)
order by month;
```

## The Output :

Month	Total_Confirmed	Total_Deaths	Total_Recovered
1	9326801	239054	9047361
2	6552180	190126	6634597
3	11162089	251274	7940387
4	20415880	470656	14872163
5	19595017	456750	20452795
6	8219255	251665	7974181
7	6581728	162580	4580405
8	7235816	171281	6091450
9	7402368	151638	6515606
10	9394631	153288	6696585
11	14240010	221988	9108146
12	13282469	250802	11825913

# Spread of Coronavirus with Respect to Confirmed Cases

- Query: Calculate total confirmed cases, their average, variance, and standard deviation

## The Output :

total_confirmed	avg_confirmed	variance_confirmed	stdev_confirmed
133408244	4673.94	321129066.39	17920.07

## The Query :

```
SELECT
    ROUND(SUM(converted_Confirmed),2)
AS total_confirmed_cases,

    ROUND( AVG(converted_Confirmed),2) AS
avg_confirmed_case,

    ROUND(VAR(converted_Confirmed),2) AS
variance_confirmed_cases,

    ROUND(STDEV(converted_Confirmed),2)
AS stdev_confirmed_cases

FROM
    [Corona Virus].[dbo].[Corona Virus
Dataset];
```

# Spread of Coronavirus with Respect to Death Cases Per Month

- Query: Calculate total death cases, their average, variance, and standard deviation per month

## The Query :

```
SELECT
    month(Converted_Date) as month,
    ROUND(SUM(Converted_Deaths),2) AS total_Deaths,

    ROUND( AVG(Converted_Deaths),2) AS avg_Deaths,

    ROUND(VAR(Converted_Deaths),2) AS
variance_Deaths,

    ROUND(STDEV(Converted_Deaths),2) AS
stdev_Deaths

FROM  [Corona Virus].[dbo].[Corona Virus Dataset]

group by month(Converted_Date)
order by month;
```

## The Output :

month	total_Deaths	avg_Deaths	variance_Deaths	stdev_Deaths
1	239054	109.31	51141.31	226.14
2	190126	92.97	48649.66	220.57
3	251274	96.76	80416.49	283.58
4	470656	129.16	149729.92	386.95
5	456750	122.16	181723.62	426.29
6	251665	102.22	114905.05	338.98
7	162580	92.22	51210.79	226.3
8	171281	88.93	51434.45	226.79
9	151638	77.88	43599.11	208.8
10	153288	75.51	35587.88	188.65
11	221988	108.08	49555.14	222.61
12	250802	116.71	69805.35	264.21

# Spread of Coronavirus with Respect to Recovered Cases

- Query: Calculate total recovered cases, their average, variance, and standard deviation per month

## The Query :

```
SELECT
    month(Converted_Date) as month,

    ROUND(SUM(Converted_Recovered),2) AS total
'
    ROUND( AVG(Converted_Recovered),2) AS avg
'
    ROUND(VAR(Converted_Recovered),2) AS variance
'
    ROUND(STDEV(Converted_Recovered),2) AS stdev

FROM  [Corona Virus].[dbo].[Corona Virus Dataset]
group by month(Converted_Date)
order by month;
```

## The Output :

month	total	avg	variance	stdev
1	9047361	4136.88	59567024.82	7717.97
2	6634597	3244.3	45515958.73	6746.55
3	7940387	3057.52	59858550.93	7736.83
4	14872163	4081.27	280591735.08	16750.87
5	20452795	5470.12	957446798.69	30942.64
6	7974181	3238.9	198098512.35	14074.75
7	4580405	2598.07	62884301.2	7929.96
8	6091450	3162.75	93637670.32	9676.66
9	6515606	3346.48	128163882.86	11320.95
10	6696585	3298.81	167209487.92	12930.95
11	9108146	4434.35	103282567.4	10162.8
12	11825913	5502.98	709581840.99	26637.98



# Country with the Highest Number of Confirmed Cases

- Query: Identify the country with the highest number of confirmed cases

## The Query :

```
SELECT TOP 5
  [Country Region],
  MAX(converted_Confirmed) AS
highest_confirmed_cases
FROM
  [Corona Virus].[dbo].[Corona Virus Dataset]
GROUP BY
  [Country Region]
ORDER BY
  highest_confirmed_cases DESC;
```

## The Output :

Country Region	highest_confirmed_cases
Turkey	823225
India	414188
US	240089
France	117900
Brazil	100158

# Country with the Lowest Number of Death Cases

- Query: Identify the country with the lowest number of death cases

## The Query :

```
SELECT TOP 5
  [Country Region],
  min(Converted_Deaths) AS
Lowest_Deaths_cases

FROM
  [Corona Virus].[dbo].[Corona Virus Dataset]
GROUP BY

  [Country Region]

ORDER BY
  Lowest_Deaths_cases DESC;
```

## The Output :

Country Region	Lowest_Deaths_cases
Turkey	14
Mexico	8
US	6
Spain	5
Peru	4

# Top 5 Countries with the Highest Number of Recovered Cases

- Query: Identify the top 5 countries with the highest number of recovered cases

## The Query :

```
SELECT TOP 5
  [Country Region],

  max(Converted_Recovered) AS
highest_Recovered_cases

FROM
  [Corona Virus].[dbo].[Corona Virus Dataset]

GROUP BY
  [Country Region]
ORDER BY
  highest_Recovered_cases DESC;
```

## The Output :

Country Region	highest_Recovered_cases
Turkey	1123456
India	422436
Brazil	388340
US	150267
Colombia	89557

# Conclusion

## 1 Summary of key findings and insights

Through the analysis, we discovered important trends such as the average monthly values for confirmed cases, deaths, and recoveries, as well as the most frequent values for each month. Additionally, we identified minimum and maximum values per year

## 2 Recommendations for further analysis or actions

Further analysis should delve into specific regions or demographics to understand COVID-19 case variability, while predictive modeling can inform future public health interventions and policies.

