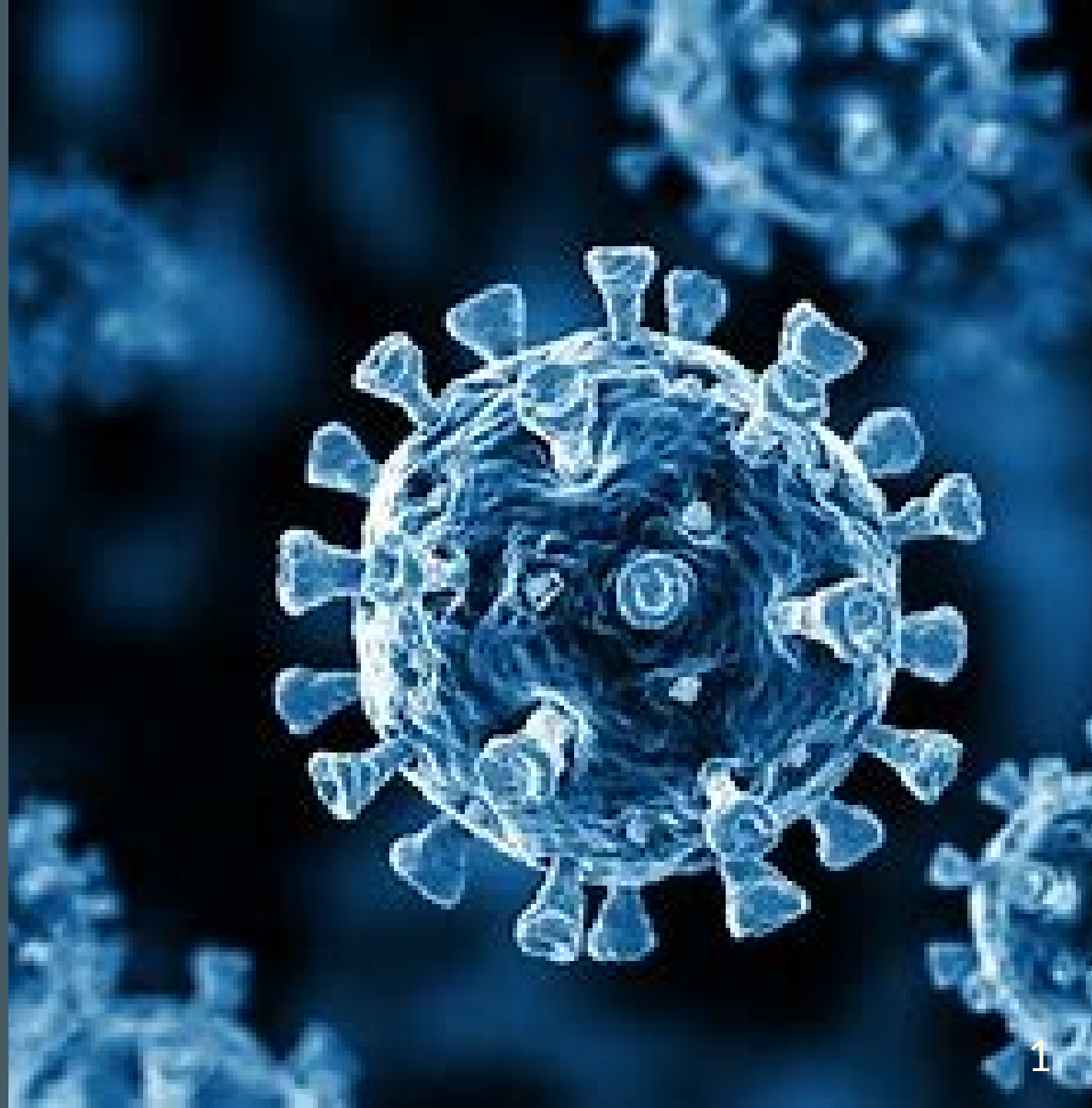
MENTORNESS INTERNSHIP PROJECT

CORONA VIRUS ANALYSIS

ADURAGBEMI OYINLOLA



OVERVIEW:

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.

For this task, I will be analyzing the **CORONA VIRUS** dataset to find answers by writing **SQl (PostgreSQL)** queries, derive meaningful insights and present my findings.

DATASET:

The dataset contains 8 fields:

- Province: Geographic subdivision within a country/region.
- Country/Region: Geographic entity where data is recorded.
- Latitude: North-south position on Earth's surface.
- Longitude: Eat-west position on Earth's surface.
- Date: Recorded date of CORONA VIRUS cases.
- Confirmed: Number of diagnosed CORONA VIRUS cases.
- Deaths: Number of CORONA VIRUS related deaths.
- Recovered: Number of recovered CORONA VIRUS cases.

Q1. Write a code to check NULL values

```
SELECT *
FROM corona_data
WHERE province IS NULL
    OR countryregion 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;
```

0 rows returned

	province character varying	countryregion character varying	latitude double precision	longitude double precision	date date	confirmed integer	deaths integer	recovered integer

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

```
UPDATE corona_data
SET province = 0, countryregion = 0, latitude = 0,
    longitude = 0, confirmed = 0, deaths = 0,
    recovered = 0
WHERE province IS NULL OR countryregion IS NULL OR
    latitude IS NULL OR longitude IS NULL OR
    confirmed IS NULL OR deaths IS NULL OR
    recovered IS NULL;
```

```
0 rows updated
```

Q3. check total number of rows

```
SELECT COUNT(*)
FROM coronavirus;
```

```
1 row returned
```

	count bigint
1	78386

Q4. Check what is start_date and end_date

```
SELECT MIN(date) AS start_date, MAX(date) AS end_date
FROM corona_data;
```

1 row returned

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

Q5. Number of month present in dataset

```
SELECT EXTRACT(YEAR FROM date) AS "Year",
    COUNT(DISTINCT EXTRACT(MONTH FROM date)) AS "monthNumber"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date);
```

2 rows returned

	Year numeric	monthNumber bigint
1	2020	12
2	2021	6

Q6. Find monthly average for confirmed, deaths, recovered

```
SELECT
    EXTRACT(YEAR FROM date) AS "Year", TO_CHAR(date, 'Month') AS "Month",
    ROUND(AVG(confirmed), 2) AS "ConfirmedCases/Month",
    ROUND(AVG(deaths), 2) AS "Deaths/Month",
    ROUND(AVG(recovered), 2) AS "Recovery/Month"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date), TO_CHAR(date, 'Month'),
    EXTRACT(MONTH FROM date)
ORDER BY EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date);
```

18 rows returned

	Year numeric	Month text	ConfirmedCases/Month numeric	Deaths/Month numeric	Recovery/Month numeric
1	2020	January	4.15	0.12	0.09
2	2020	February	15.30	0.59	7.03
3	2020	March	161.13	8.66	27.87
4	2020	April	505.80	41.52	171.64
5	2020	May	574.85	30.28	318.30
6	2020	June	859.23	29.82	548.79
7	2020	July	1432.36	35.11	983.06
8	2020	August	1611.84	37.54	1299.29
9	2020	September	1784.59	34.78	1438.91
10	2020	October	2412.20	36.76	1420.64
11	2020	November	3592.19	56.76	1985.34
12	2020	December	4050.44	71.22	2497.89
13	2021	January	3911.23	84.18	1919.64
14	2021	February	2433.36	69.16	1558.39
15	2021	March	2916.80	59.20	1652.29
16	2021	April	4699.36	78.44	3074.79
17	2021	May	4005.25	76.78	4007.51
18	2021	June	2508.63	66.26	2769.45

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

```
SELECT
    EXTRACT(YEAR FROM date) AS "Year",
    TO_CHAR(date, 'Month') AS "Month",
    MODE() WITHIN GROUP (ORDER BY confirmed) AS "MostFrequentConfirmed",
    MODE() WITHIN GROUP (ORDER BY deaths) AS "MostFrequentDeaths",
    MODE() WITHIN GROUP (ORDER BY recovered) AS "MostFrequentRecovered"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date), TO_CHAR(date, 'Month'),
    EXTRACT(MONTH FROM date)
ORDER BY EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date);
```

18 rows returned

	Year numeric	Month text	MostFrequentConfirmed integer	MostFrequentDeaths integer	MostFrequentRecovered integer
1	2020	January	0	0	0
2	2020	February	0	0	0
3	2020	March	0	0	0
4	2020	April	0	0	0
5	2020	May	0	0	0
6	2020	June	0	0	0
7	2020	July	0	0	0
8	2020	August	0	0	0
9	2020	September	0	0	0
10	2020	October	0	0	0
11	2020	November	0	0	0
12	2020	December	0	0	0
13	2021	January	0	0	0
14	2021	February	0	0	0
15	2021	March	0	0	0
16	2021	April	0	0	0
17	2021	May	0	0	0
18	2021	June	0	0	0

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

```
SELECT EXTRACT(YEAR FROM date) AS "Year", MIN(confirmed) AS "MinConfirmedCases/Year",
    MIN(deaths) AS "MinDeaths/Year", MIN(recovered) AS "MinRecovery/Year"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date);
```

2 rows returned

	Year numeric	MinConfirmedCases/Year integer	MinDeaths/Year integer	MinRecovery/Year integer
1	2021	0	0	0
2	2020	0	0	0

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

```
SELECT EXTRACT(YEAR FROM date) AS "Year", MAX(confirmed) AS "MaxConfirmedCases/Year",
    MAX(deaths) AS "MaxDeaths/Year", MAX(recovered) AS "MaxRecovery/Year"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date);
```

2 rows returned

	Year numeric	MaxConfirmedCases/Year integer	MaxDeaths/Year integer	MaxRecovery/Year integer
1	2021	414188	7374	422436
2	2020	823225	3752	1123456

Q10. The total number of case of confirmed, deaths, recovered each month

```
SELECT
    EXTRACT(YEAR FROM date) AS "Year",
    TO_CHAR(date, 'Month') AS "Month",
    SUM(confirmed) AS "TotalConfirmedCases/Month",
    SUM(deaths) AS "TotalDeaths/Month",
    SUM(recovered) AS "TotalRecovery/Month"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date), TO_CHAR(date, 'Month'),
    EXTRACT(MONTH FROM date)
ORDER BY EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date);
```

18 rows returned

	Year numeric	Month text	TotalConfirmedCases/Month bigint	TotalDeaths/Month bigint	TotalRecovery/Month bigint
1	2020	January	6384	190	143
2	2020	February	68312	2651	31405
3	2020	March	769236	41346	133070
4	2020	April	2336798	191833	792987
5	2020	May	2744333	144561	1519547
6	2020	June	3969634	137757	2535417
7	2020	July	6838092	167613	4693120
8	2020	August	7694938	179200	6202833
9	2020	September	8244794	160671	6647749
10	2020	October	11515841	175484	6782150
11	2020	November	16595938	262247	9172292
12	2020	December	19336799	339996	11924903
13	2021	January	18672205	401893	9164347
14	2021	February	10492664	298239	6719785
15	2021	March	13924790	282620	7888013
16	2021	April	21711021	362387	14205507
17	2021	May	19121083	366549	19131842
18	2021	June	5022282	132657	5544438

Q11. Check how corona virus spread out with respect to confirmed case (Eg.: total confirmed cases, their average, variance & STDEV)

```
SELECT
    ROUND(SUM(confirmed), 2) AS "TotalConfirmedCases",
    ROUND(AVG(confirmed), 2) AS "AvgConfirmedCases",
    ROUND(VARIANCE(confirmed), 2) AS "ConfirmedCasesVar",
    ROUND(STDDEV(confirmed), 2) AS "ConfirmedCasesSpread"
FROM corona_data;
```

1 row returned

	TotalConfirmedCases numeric	AvgConfirmedCases numeric	ConfirmedCasesVar numeric	ConfirmedCasesSpread numeric
1	169065144.00	2156.83	157290931.70	12541.57

Q12. Check how corona virus spread out with respect to death case per month (Eg.: total confirmed cases, their average, variance & STDEV)

```
SELECT
    EXTRACT(YEAR FROM date) AS "Year",
    TO_CHAR(date, 'Month') AS "Month",
    COUNT(deaths) AS "TotalDeaths",
    ROUND(AVG(deaths), 2) AS "AvgDeaths",
    ROUND(VARIANCE(deaths), 2) AS "DeathsVar",
    ROUND(STDDEV(deaths), 2) AS "DeathsSpread"
FROM corona_data
GROUP BY EXTRACT(YEAR FROM date), TO_CHAR(date, 'Month'),
    EXTRACT(MONTH FROM date)
ORDER BY EXTRACT(YEAR FROM date), EXTRACT(MONTH FROM date);
```

18 rows returned

	Year numeric	Month text	TotalDeaths bigint	AvgDeaths numeric	DeathsVar numeric	DeathsSpread numeric
1	2020	January	1540	0.12	4.25	2.06
2	2020	February	4466	0.59	68.34	8.27
3	2020	March	4774	8.66	3901.61	62.46
4	2020	April	4620	41.52	40513.04	201.28
5	2020	May	4774	30.28	20689.25	143.84
6	2020	June	4620	29.82	16933.11	130.13
7	2020	July	4774	35.11	21144.58	145.41
8	2020	August	4774	37.54	23277.87	152.57
9	2020	September	4620	34.78	20107.12	141.80
10	2020	October	4774	36.76	17583.75	132.60
11	2020	November	4620	56.76	27779.81	166.67
12	2020	December	4774	71.22	65359.06	255.65
13	2021	January	4774	84.18	102779.96	320.59
14	2021	February	4312	69.16	68494.76	261.72
15	2021	March	4774	59.20	54397.36	233.23
16	2021	April	4620	78.44	94631.95	307.62
17	2021	May	4774	76.78	131797.08	363.04
18	2021	June	2002	66.26	113020.13	336.18

Q13. Check how corona virus spread out with respect to recovered case (Eg.: total confirmed cases, their average, variance & STDEV)

```
SELECT
    ROUND(SUM(recovered), 2) AS "TotalRecovery",
    ROUND(AVG(recovered), 2) AS "AvgRecovery",
    ROUND(VARIANCE(recovered), 2) AS "RecoveryVAR",
    ROUND(STDDEV(recovered), 2) AS "RecoverySTD"
FROM corona_data;
```

1 row returned

	TotalRecovery numeric	AvgRecovery numeric	RecoveryVAR numeric	RecoverySTD numeric
1	113089548.00	1442.73	107030888.70	10345.57

Q14. Find Country having highest number of the Confirmed case

```
SELECT countryregion, MAX(TotalConfirmed) AS "MaxConfirmed"
FROM (
    SELECT countryregion, SUM(confirmed) AS TotalConfirmed
    FROM corona_data
    GROUP BY countryregion
    ) AS DeathPerCountry
GROUP BY countryregion
ORDER BY "MaxConfirmed" DESC
LIMIT 1;
```

1 row returned

	countryregion character varying	MaxConfirmedCases bigint
1	US	33461982

Q15. Find Country having lowest number of the death case

```
SELECT countryregion, MIN(TotalDeaths) AS "MinDeath"
FROM (
    SELECT countryregion, SUM(deaths) AS TotalDeaths
    FROM corona_data
    GROUP BY countryregion
    ) AS DeathPerCountry
GROUP BY countryregion
ORDER BY "MinDeath"
LIMIT 4;
```

4 rows returned

	countryregion character varying	MinDeath bigint
1	Marshall Islands	0
2	Kiribati	0
3	Dominica	0
4	Samoa	0

Q16. Find top 5 countries having highest recovered case

```
SELECT countryregion, MAX(TotalRecovered) AS "MaxRecovered"
FROM (
    SELECT countryregion, SUM(recovered) AS TotalRecovered
    FROM corona_data
    GROUP BY countryregion
    ) AS DeathPerCountry
GROUP BY countryregion
ORDER BY "MaxRecovered" DESC
LIMIT 5;
```

5 rows returned

	countryregion character varying	MaxRecovered bigint
1	India	28089649
2	Brazil	15400169
3	US	6303715
4	Turkey	5202251
5	Russia	4745756

INFERENCE

- The CORONA VIRUS dataset contains no null values (query 1 and 2).
- There are a total of 78386 datapoints in the dataset (query 3).
- The datapoints in this dataset is collected between the 22nd of January 2020 and 13th of June 2021. This is a period of 18 months (query 4 and 5).
- Queries 6, 7 and 10 presents the average, most frequent and total number of cases values of the *Confirmed*, *Deaths* and *Recovered* on a monthly basis respectively.

- Queries 8 and 9 shows the minimum and maximum values for the *Confirmed*, *Deaths* and *Recovered* on a yearly basis respectively.
- Queries 11, 12 and 13 shows some statistical operations being performed on the *Confirmed*, *Deaths* and *Recovered* fields.
- The *US* has the most number of confirmed COVID cases, with 33461892 recorded cases (query 14).
- *Marshall Islands*, *Kiribati*, *Dominica* and *Samoa*, had 0 record of COVID related death (query 15).
- Query 16 shows that *India* tops the list of countries with most recoveries, followed *Brazil*, *US*, *Turkey*, *Russia*.

THANK YOU, PLEASE GIVE A REVIEW