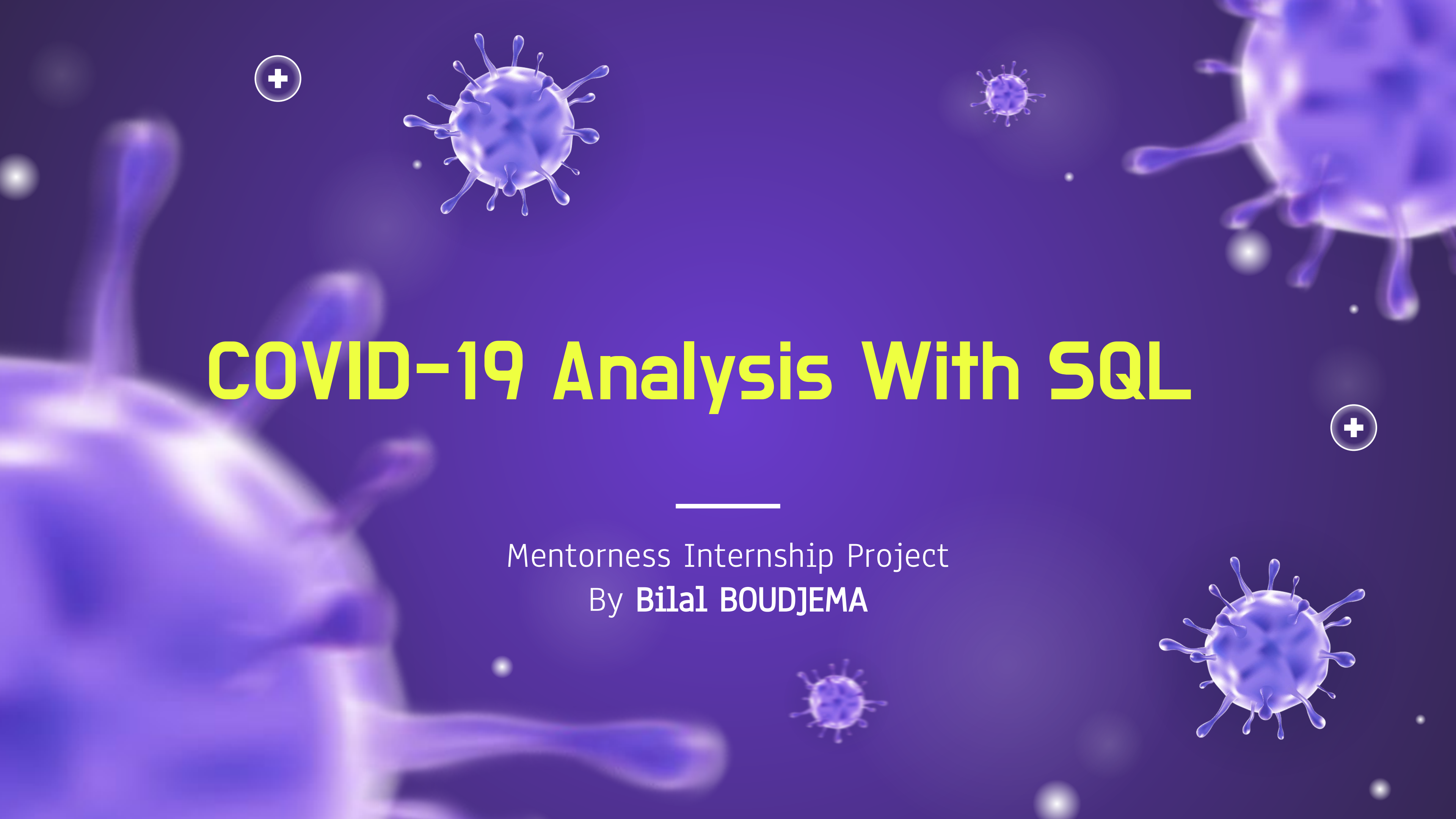


COVID-19 Analysis With SQL

Mentorness Internship Project

By **Bilal BOUDJEMA**

A PICTURE IS WORTH A THOUSAND WORDS

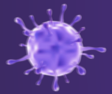
TABLE OF CONTENTS

01

Project Overview

02

Dataset Description

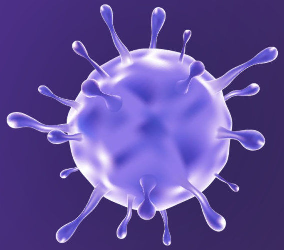


03

Analysis (Queries)

04

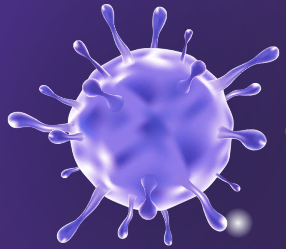
Summary



Project Overview

The COVID-19 pandemic has profoundly affected public health, highlighting the pressing necessity for data-driven analysis to comprehend its transmission patterns.

As a data analyst, my assignment involves delving into a COVID-19 dataset to extract valuable insights and deliver your conclusions.



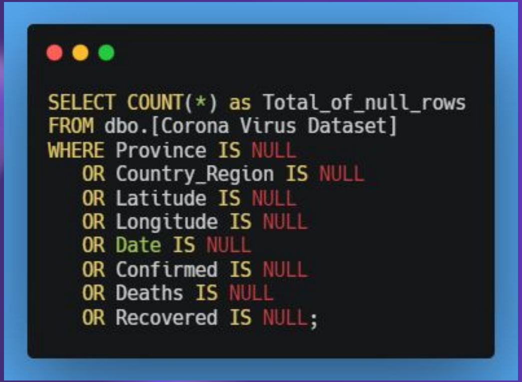
Dataset Description

Column Descriptions in the Dataset:

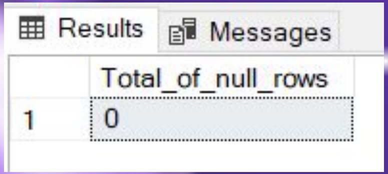
- **Province**: A geographic division within a country or region.
- **Country/Region:** The geographical entity where the data is documented.
- **Latitude:** The north-south position on the Earth's surface.
- **Longitude:** The east-west position on the Earth's surface.
- **Date:** The recorded date of the COVID-19 data.
- **Confirmed:** The count of diagnosed COVID-19 cases.
- **Deaths:** The tally of COVID-19 related fatalities.
- **Recovered:** The number of individuals who have recuperated from COVID-19.

Data Exploration and Analysis

- Checking for NULL Values

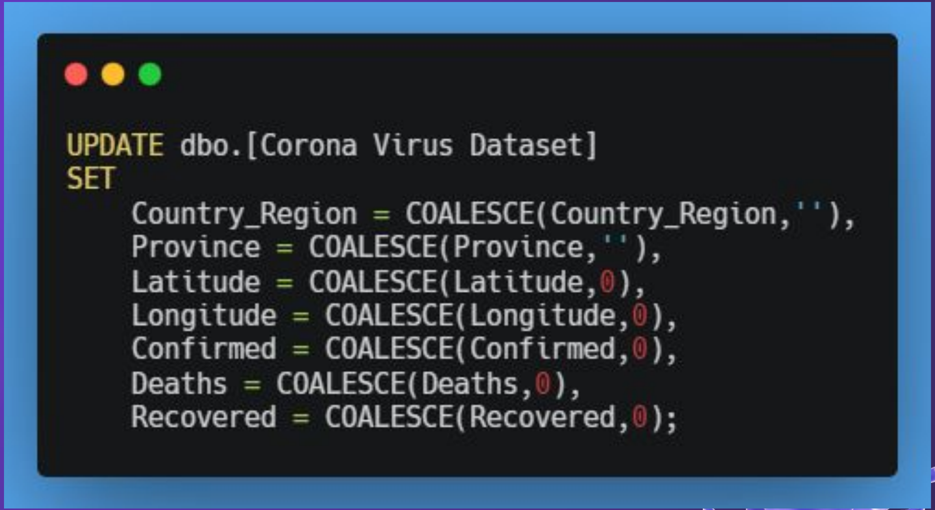


```sql
SELECT COUNT(*) as Total_of_null_rows
FROM dbo.[Corona Virus Dataset]
WHERE Province IS NULL
   OR Country_Region IS NULL
   OR Latitude IS NULL
   OR Longitude IS NULL
   OR Date IS NULL
   OR Confirmed IS NULL
   OR Deaths IS NULL
   OR Recovered IS NULL;
```



	Total_of_null_rows
1	0

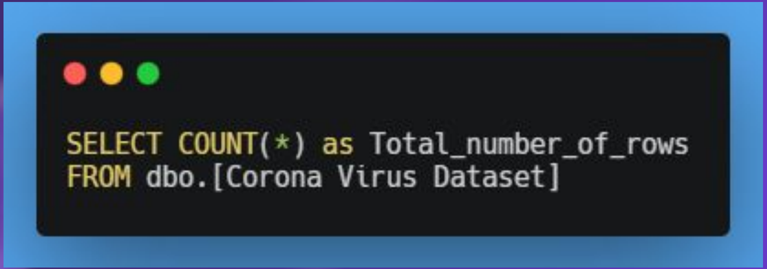
- Checking for NULL Values



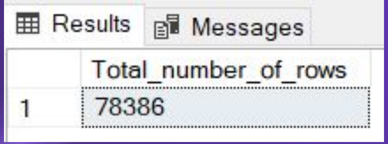
```sql
UPDATE dbo.[Corona Virus Dataset]
SET
    Country_Region = COALESCE(Country_Region,''),
    Province = COALESCE(Province,''),
    Latitude = COALESCE(Latitude,0),
    Longitude = COALESCE(Longitude,0),
    Confirmed = COALESCE(Confirmed,0),
    Deaths = COALESCE(Deaths,0),
    Recovered = COALESCE(Recovered,0);
```

Data Exploration and Analysis

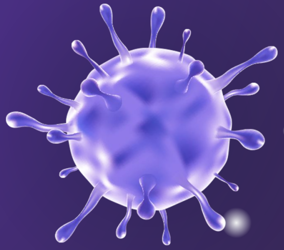
- Check the total number of rows



```sql
SELECT COUNT(*) as Total_number_of_rows
FROM dbo.[Corona Virus Dataset]
```

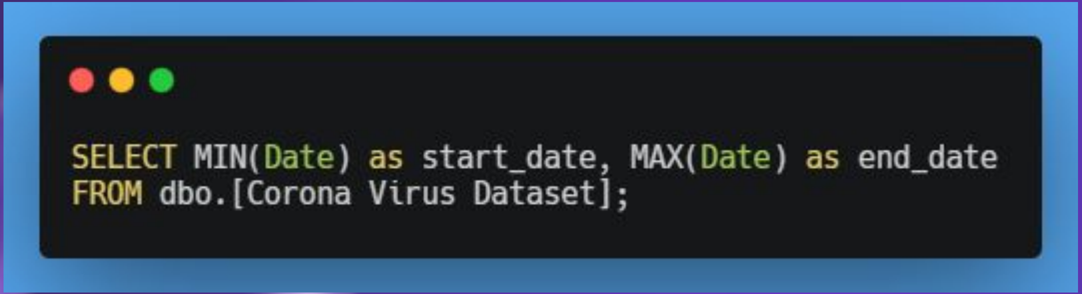


	Total_number_of_rows
1	78386

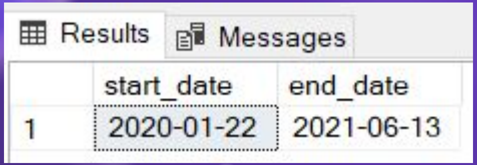


Data Exploration and Analysis

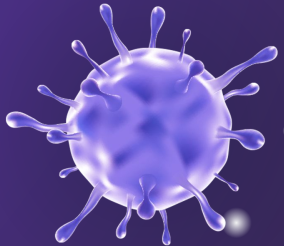
- Check what is the start date and end date



```sql
SELECT MIN(Date) as start_date, MAX(Date) as end_date
FROM dbo.[Corona Virus Dataset];
```

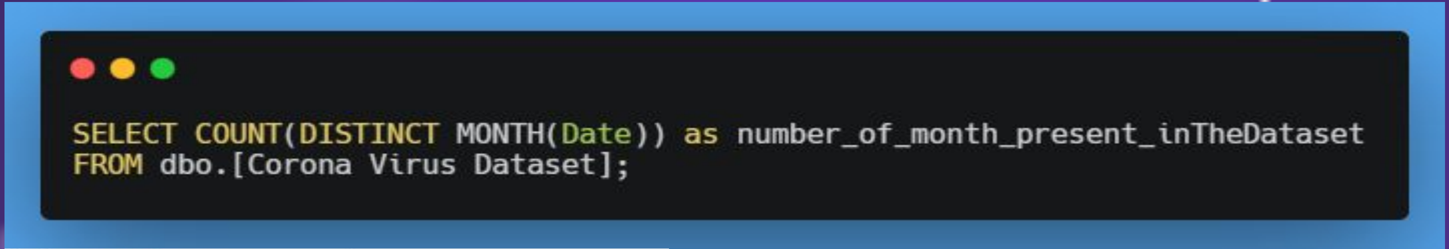


	start_date	end_date
1	2020-01-22	2021-06-13

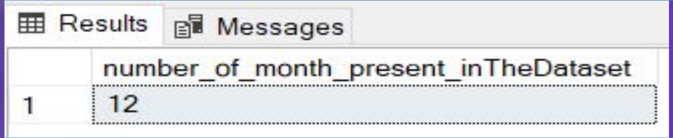


Data Exploration and Analysis

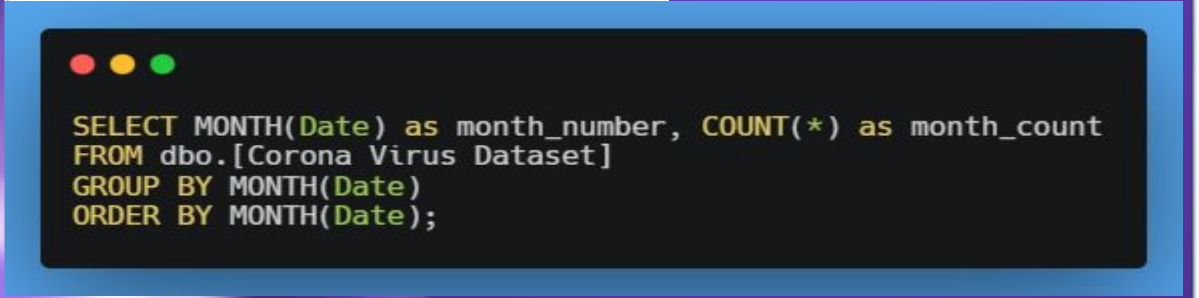
- Check number of months in the dataset



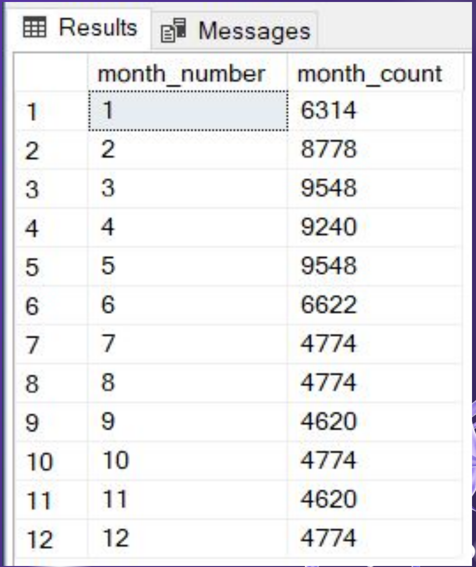
```
SELECT COUNT(DISTINCT MONTH(Date)) as number_of_month_present_inTheDataset
FROM dbo.[Corona Virus Dataset];
```



	number_of_month_present_inTheDataset
1	12



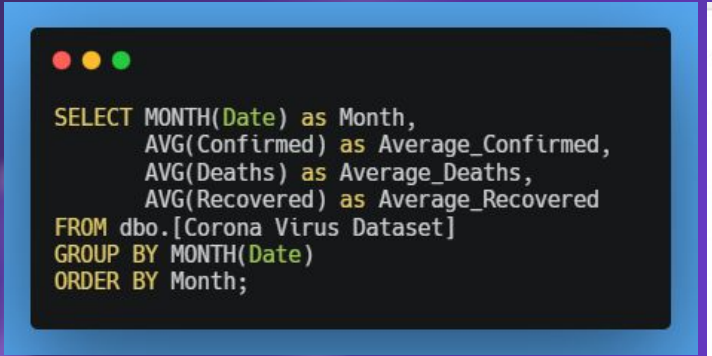
```
SELECT MONTH(Date) as month_number, COUNT(*) as month_count
FROM dbo.[Corona Virus Dataset]
GROUP BY MONTH(Date)
ORDER BY MONTH(Date);
```



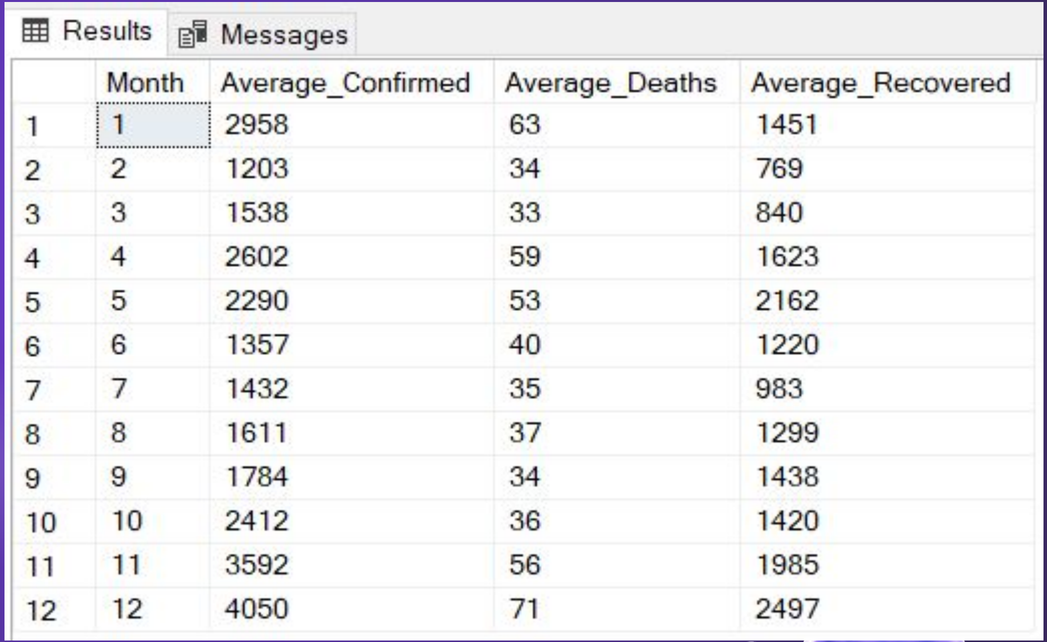
	month_number	month_count
1	1	6314
2	2	8778
3	3	9548
4	4	9240
5	5	9548
6	6	6622
7	7	4774
8	8	4774
9	9	4620
10	10	4774
11	11	4620
12	12	4774

Data Exploration and Analysis

- Find the monthly average for confirmed, deaths, recovered



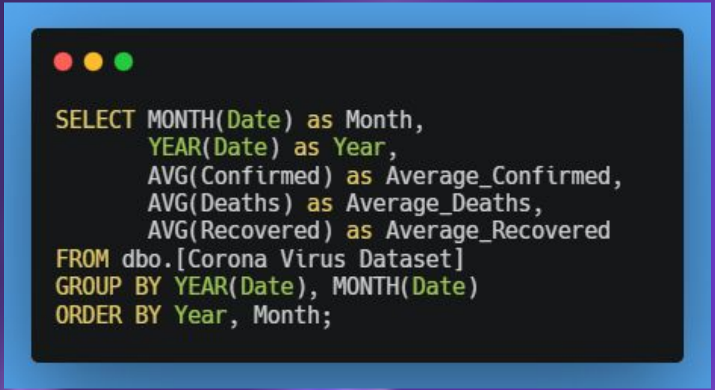
```
SELECT MONTH(Date) as Month,
       AVG(Confirmed) as Average_Confirmed,
       AVG(Deaths) as Average_Deaths,
       AVG(Recovered) as Average_Recovered
FROM dbo.[Corona Virus Dataset]
GROUP BY MONTH(Date)
ORDER BY Month;
```



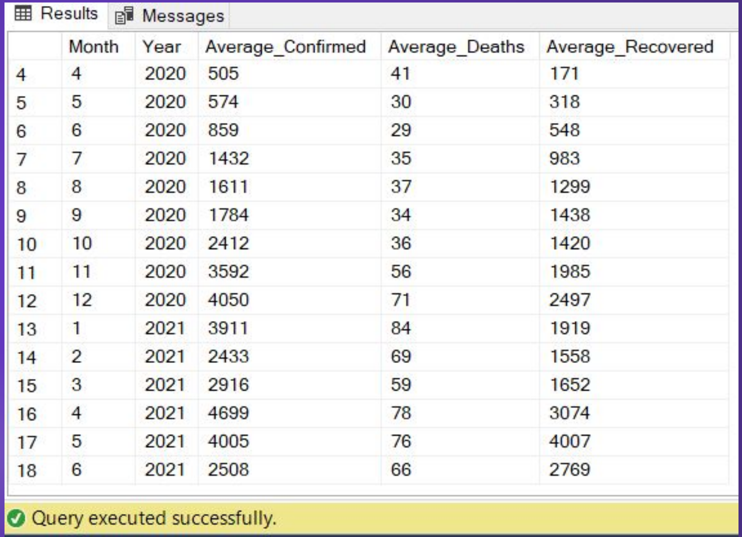
	Month	Average_Confirmed	Average_Deaths	Average_Recovered
1	1	2958	63	1451
2	2	1203	34	769
3	3	1538	33	840
4	4	2602	59	1623
5	5	2290	53	2162
6	6	1357	40	1220
7	7	1432	35	983
8	8	1611	37	1299
9	9	1784	34	1438
10	10	2412	36	1420
11	11	3592	56	1985
12	12	4050	71	2497

Data Exploration and Analysis

- Find the monthly average for confirmed, deaths, recovered

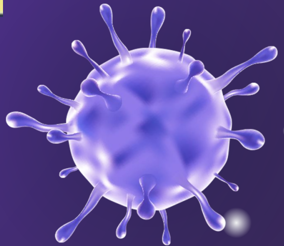


```
SELECT MONTH(Date) as Month,
       YEAR(Date) as Year,
       AVG(Confirmed) as Average_Confirmed,
       AVG(Deaths) as Average_Deaths,
       AVG(Recovered) as Average_Recovered
FROM dbo.[Corona Virus Dataset]
GROUP BY YEAR(Date), MONTH(Date)
ORDER BY Year, Month;
```



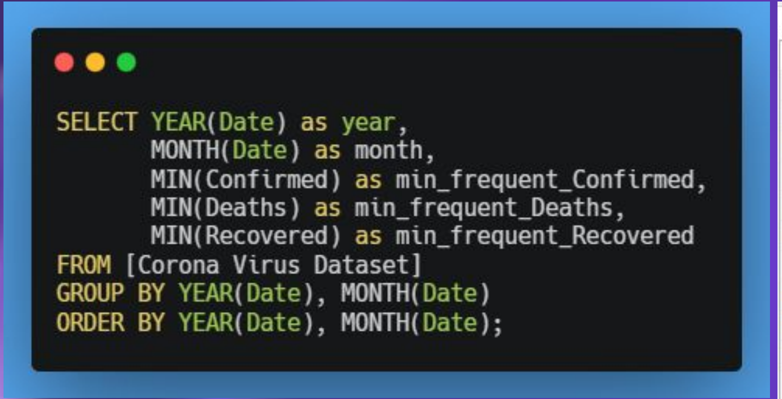
	Month	Year	Average_Confirmed	Average_Deaths	Average_Recovered
4	4	2020	505	41	171
5	5	2020	574	30	318
6	6	2020	859	29	548
7	7	2020	1432	35	983
8	8	2020	1611	37	1299
9	9	2020	1784	34	1438
10	10	2020	2412	36	1420
11	11	2020	3592	56	1985
12	12	2020	4050	71	2497
13	1	2021	3911	84	1919
14	2	2021	2433	69	1558
15	3	2021	2916	59	1652
16	4	2021	4699	78	3074
17	5	2021	4005	76	4007
18	6	2021	2508	66	2769

Query executed successfully.

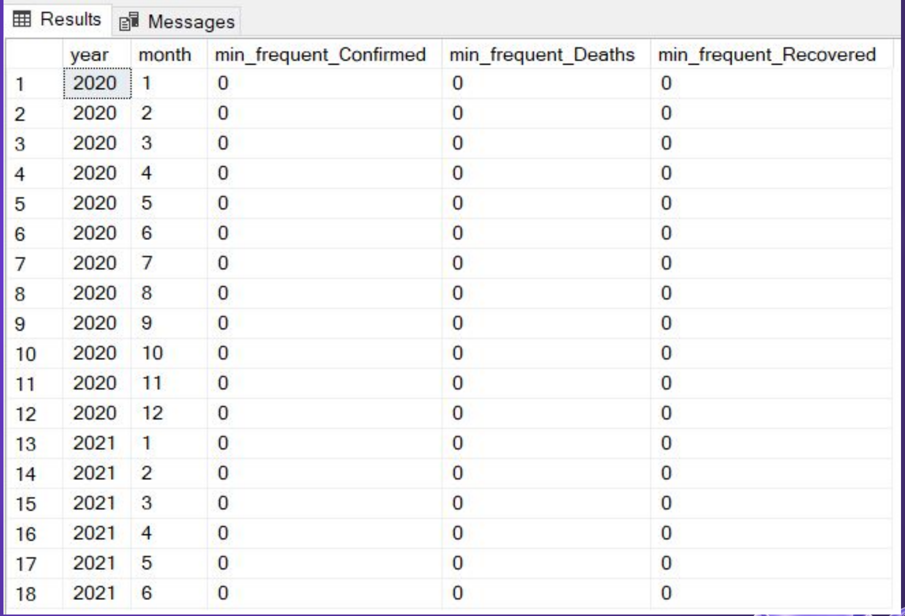


Data Exploration and Analysis

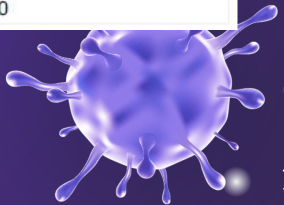
- Find minimum values for confirmed, deaths, recovered per year



```
SELECT YEAR(Date) as year,
       MONTH(Date) as month,
       MIN(Confirmed) as min_frequent_Confirmed,
       MIN(Deaths) as min_frequent_Deaths,
       MIN(Recovered) as min_frequent_Recovered
FROM [Corona Virus Dataset]
GROUP BY YEAR(Date), MONTH(Date)
ORDER BY YEAR(Date), MONTH(Date);
```

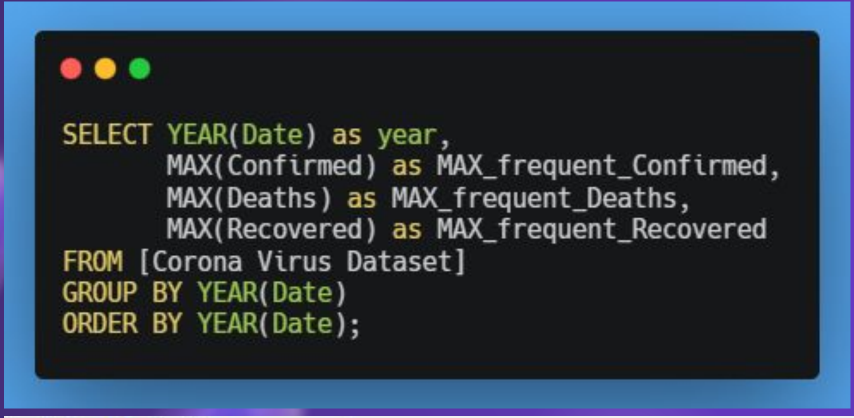


	year	month	min_frequent_Confirmed	min_frequent_Deaths	min_frequent_Recovered
1	2020	1	0	0	0
2	2020	2	0	0	0
3	2020	3	0	0	0
4	2020	4	0	0	0
5	2020	5	0	0	0
6	2020	6	0	0	0
7	2020	7	0	0	0
8	2020	8	0	0	0
9	2020	9	0	0	0
10	2020	10	0	0	0
11	2020	11	0	0	0
12	2020	12	0	0	0
13	2021	1	0	0	0
14	2021	2	0	0	0
15	2021	3	0	0	0
16	2021	4	0	0	0
17	2021	5	0	0	0
18	2021	6	0	0	0

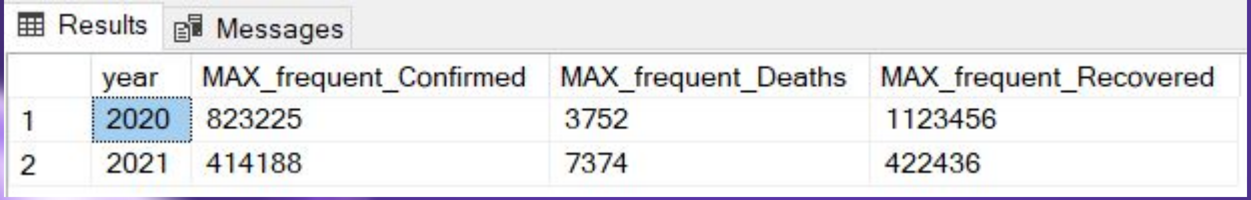


Data Exploration and Analysis

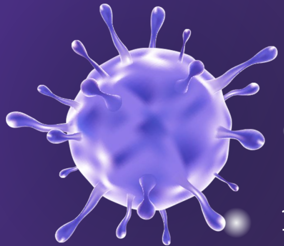
- Find maximum values of confirmed, deaths, recovered per year



```
SELECT YEAR(Date) as year,
       MAX(Confirmed) as MAX_frequent_Confirmed,
       MAX(Deaths) as MAX_frequent_Deaths,
       MAX(Recovered) as MAX_frequent_Recovered
FROM [Corona Virus Dataset]
GROUP BY YEAR(Date)
ORDER BY YEAR(Date);
```

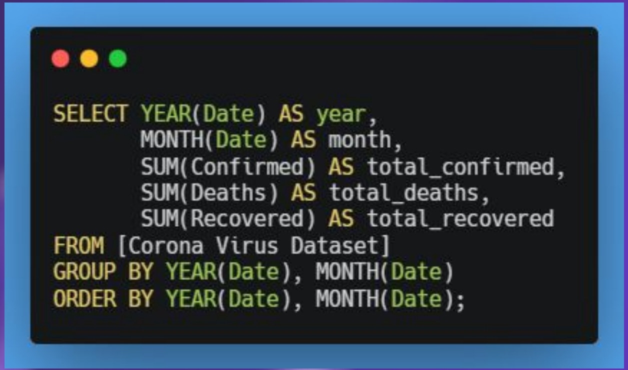


	year	MAX_frequent_Confirmed	MAX_frequent_Deaths	MAX_frequent_Recovered
1	2020	823225	3752	1123456
2	2021	414188	7374	422436

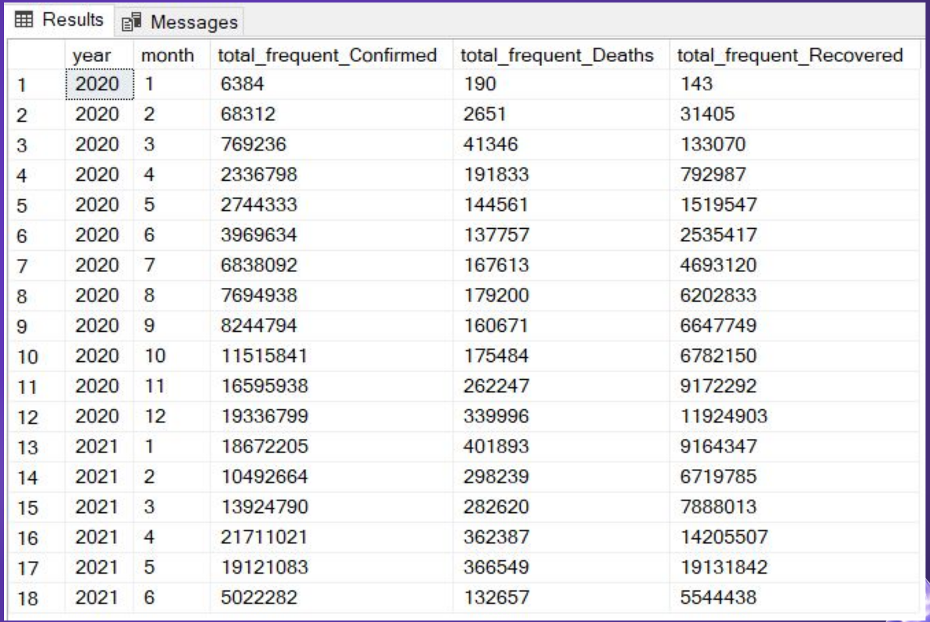


Data Exploration and Analysis

- The total number of case of confirmed, deaths, recovered each month



```
SELECT YEAR(Date) AS year,
       MONTH(Date) AS month,
       SUM(Confirmed) AS total_confirmed,
       SUM(Deaths) AS total_deaths,
       SUM(Recovered) AS total_recovered
FROM [Corona Virus Dataset]
GROUP BY YEAR(Date), MONTH(Date)
ORDER BY YEAR(Date), MONTH(Date);
```



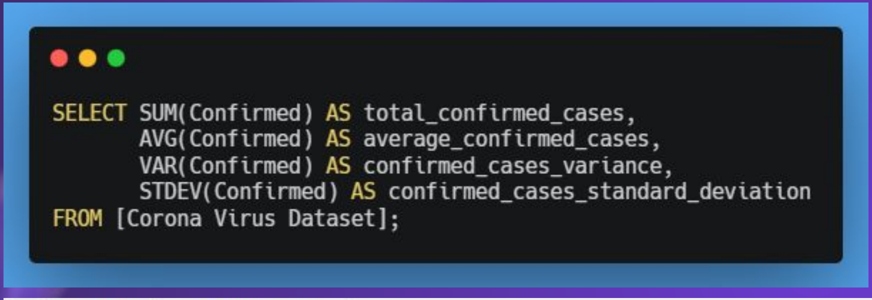
	year	month	total_frequent_Confirmed	total_frequent_Deaths	total_frequent_Recovered
1	2020	1	6384	190	143
2	2020	2	68312	2651	31405
3	2020	3	769236	41346	133070
4	2020	4	2336798	191833	792987
5	2020	5	2744333	144561	1519547
6	2020	6	3969634	137757	2535417
7	2020	7	6838092	167613	4693120
8	2020	8	7694938	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

169,065,144

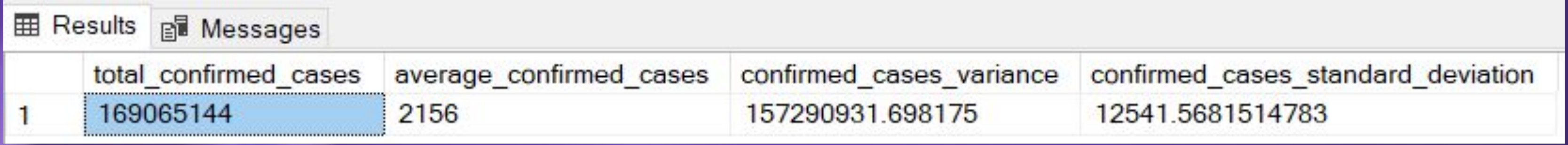
Total Confirmed cases

Data Exploration and Analysis

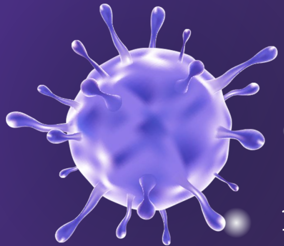
- Check how coronavirus spread out with respect to confirmed case



```sql
SELECT SUM(Confirmed) AS total_confirmed_cases,
       AVG(Confirmed) AS average_confirmed_cases,
       VAR(Confirmed) AS confirmed_cases_variance,
       STDEV(Confirmed) AS confirmed_cases_standard_deviation
FROM [Corona Virus Dataset];
```

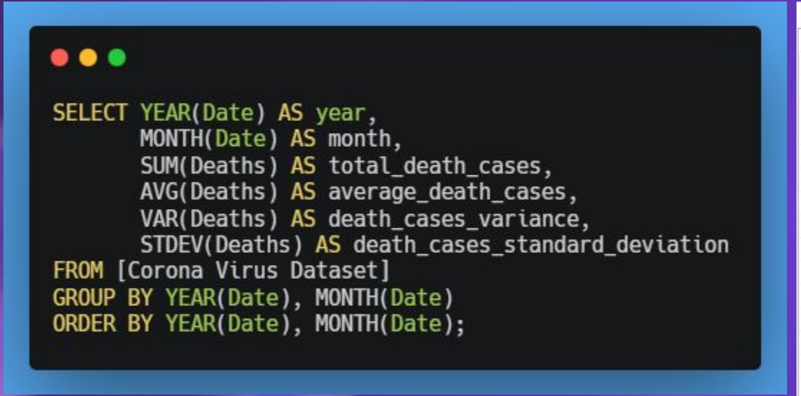


	total_confirmed_cases	average_confirmed_cases	confirmed_cases_variance	confirmed_cases_standard_deviation
1	169065144	2156	157290931.698175	12541.5681514783

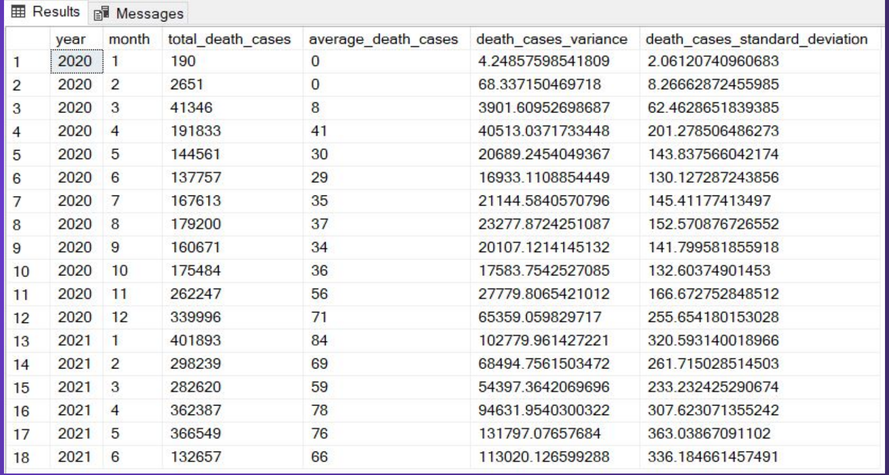


Data Exploration and Analysis

- Check how coronavirus spread out with respect to death case per month



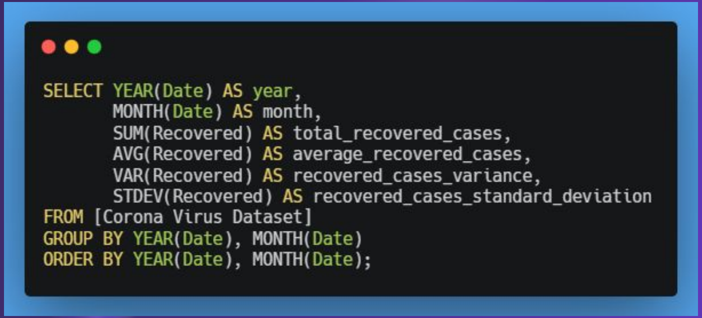
```
SELECT YEAR(Date) AS year,
       MONTH(Date) AS month,
       SUM(Deaths) AS total_death_cases,
       AVG(Deaths) AS average_death_cases,
       VAR(Deaths) AS death_cases_variance,
       STDEV(Deaths) AS death_cases_standard_deviation
FROM [Corona Virus Dataset]
GROUP BY YEAR(Date), MONTH(Date)
ORDER BY YEAR(Date), MONTH(Date);
```



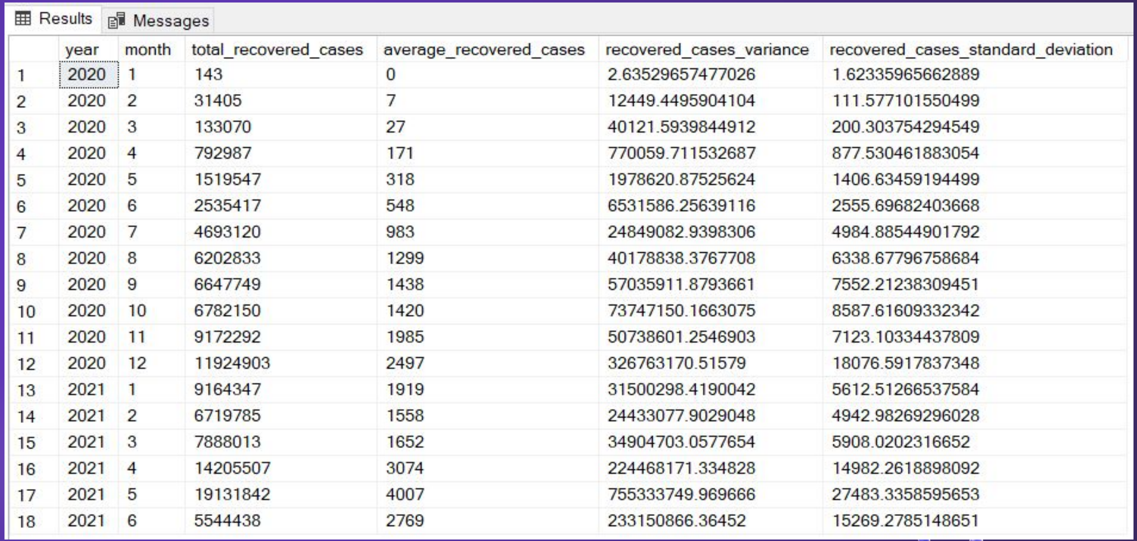
	year	month	total_death_cases	average_death_cases	death_cases_variance	death_cases_standard_deviation
1	2020	1	190	0	4.24857598541809	2.06120740960683
2	2020	2	2651	0	68.337150469718	8.26662872455985
3	2020	3	41346	8	3901.60952698687	62.4628651839385
4	2020	4	191833	41	40513.0371733448	201.278506486273
5	2020	5	144561	30	20689.2454049367	143.837566042174
6	2020	6	137757	29	16933.1108854449	130.127287243856
7	2020	7	167613	35	21144.5840570796	145.41177413497
8	2020	8	179200	37	23277.8724251087	152.570876726552
9	2020	9	160671	34	20107.1214145132	141.799581855918
10	2020	10	175484	36	17583.7542527085	132.60374901453
11	2020	11	262247	56	27779.8065421012	166.672752848512
12	2020	12	339996	71	65359.059829717	255.654180153028
13	2021	1	401893	84	102779.961427221	320.593140018966
14	2021	2	298239	69	68494.7561503472	261.715028514503
15	2021	3	282620	59	54397.3642069696	233.232425290674
16	2021	4	362387	78	94631.9540300322	307.623071355242
17	2021	5	366549	76	131797.07657684	363.03867091102
18	2021	6	132657	66	113020.126599288	336.184661457491

Data Exploration and Analysis

- Check how coronavirus spread out with respect to recovered case



```
SELECT YEAR(Date) AS year,
       MONTH(Date) AS month,
       SUM(Recovered) AS total_recovered_cases,
       AVG(Recovered) AS average_recovered_cases,
       VAR(Recovered) AS recovered_cases_variance,
       STDEV(Recovered) AS recovered_cases_standard_deviation
FROM [Corona Virus Dataset]
GROUP BY YEAR(Date), MONTH(Date)
ORDER BY YEAR(Date), MONTH(Date);
```



	year	month	total_recovered_cases	average_recovered_cases	recovered_cases_variance	recovered_cases_standard_deviation
1	2020	1	143	0	2.63529657477026	1.62335965662889
2	2020	2	31405	7	12449.4495904104	111.577101550499
3	2020	3	133070	27	40121.5939844912	200.303754294549
4	2020	4	792987	171	770059.711532687	877.530461883054
5	2020	5	1519547	318	1978620.87525624	1406.63459194499
6	2020	6	2535417	548	6531586.25639116	2555.69682403668
7	2020	7	4693120	983	24849082.9398306	4984.88544901792
8	2020	8	6202833	1299	40178838.3767708	6338.67796758684
9	2020	9	6647749	1438	57035911.8793661	7552.21238309451
10	2020	10	6782150	1420	73747150.1663075	8587.61609332342
11	2020	11	9172292	1985	50738601.2546903	7123.10334437809
12	2020	12	11924903	2497	326763170.51579	18076.5917837348
13	2021	1	9164347	1919	31500298.4190042	5612.51266537584
14	2021	2	6719785	1558	24433077.9029048	4942.98269296028
15	2021	3	7888013	1652	34904703.0577654	5908.0202316652
16	2021	4	14205507	3074	224468171.334828	14982.2618898092
17	2021	5	19131842	4007	755333749.969666	27483.3358595653
18	2021	6	5544438	2769	233150866.36452	15269.2785148651

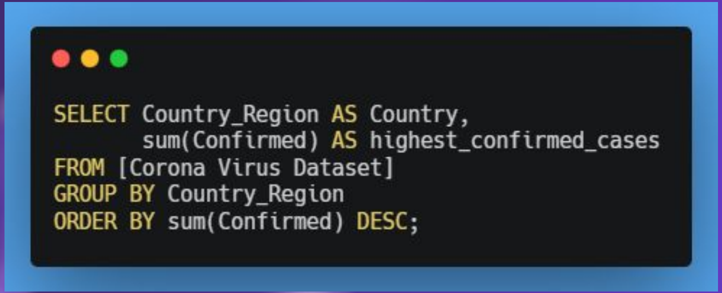
FINDINGS

Country having highest number of the Confirmed case

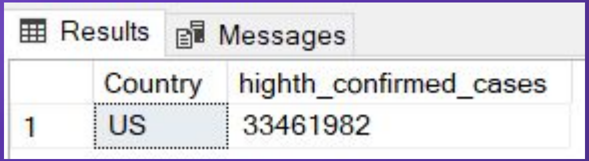
334,619,82 Confirmed cases in U.S

Data Exploration and Analysis

- Find Country having highest number of the Confirmed case



```
SELECT Country_Region AS Country,
       sum(Confirmed) AS highest_confirmed_cases
FROM [Corona Virus Dataset]
GROUP BY Country_Region
ORDER BY sum(Confirmed) DESC;
```



	Country	highth_confirmed_cases
1	US	33461982

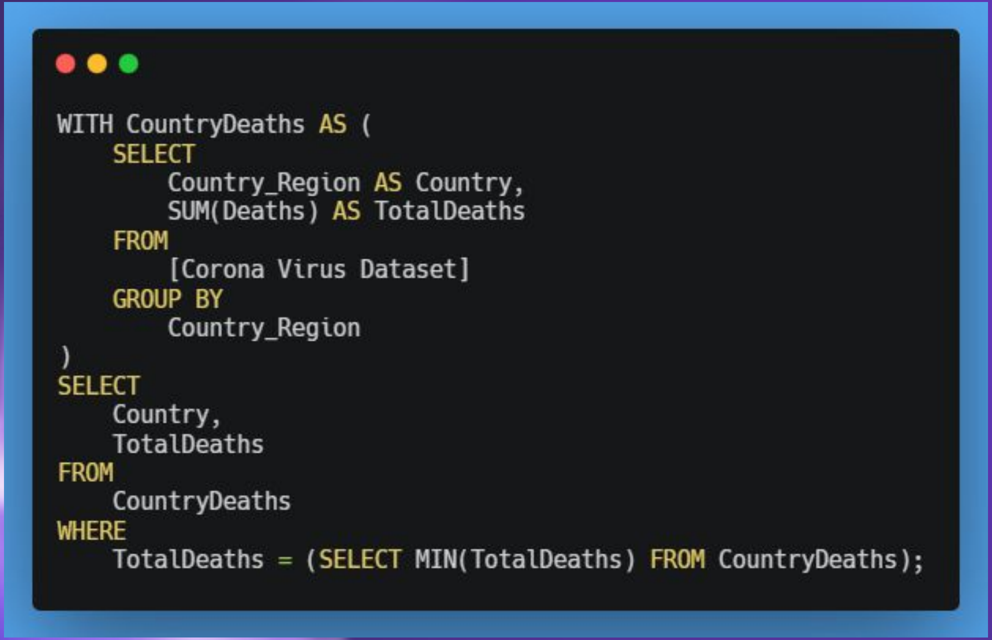
FINDINGS

Country having lowest number of the death case

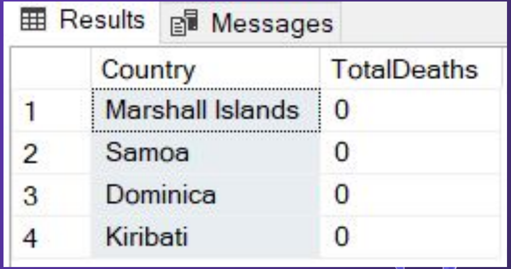
- Marshall Islands
- Samoa
- Dominica
- Kiribati

Data Exploration and Analysis

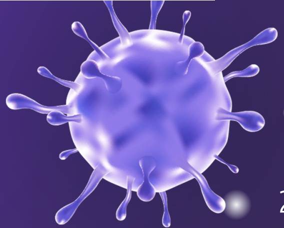
- Country having lowest number of the death case



```
WITH CountryDeaths AS (
    SELECT
        Country_Region AS Country,
        SUM(Deaths) AS TotalDeaths
    FROM
        [Corona Virus Dataset]
    GROUP BY
        Country_Region
)
SELECT
    Country,
    TotalDeaths
FROM
    CountryDeaths
WHERE
    TotalDeaths = (SELECT MIN(TotalDeaths) FROM CountryDeaths);
```



	Country	TotalDeaths
1	Marshall Islands	0
2	Samoa	0
3	Dominica	0
4	Kiribati	0



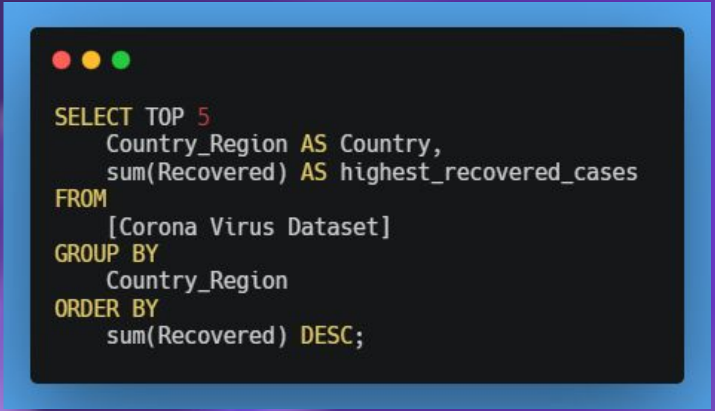
FINDINGS

Top 5 countries having highest recovered case

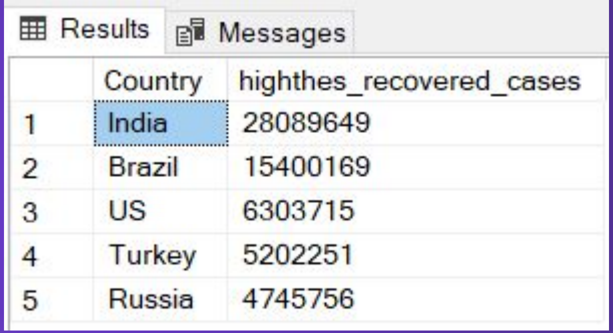
- India
- Brazil
- US
- Turkey
- Russia

Data Exploration and Analysis

- Top 5 countries having highest recovered case



```sql
SELECT TOP 5
    Country_Region AS Country,
    sum(Recovered) AS highest_recovered_cases
FROM
    [Corona Virus Dataset]
GROUP BY
    Country_Region
ORDER BY
    sum(Recovered) DESC;
```



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

Insights

Insights

- COVID-19 Pandemic Duration: January 22, 2020, to June 13, 2021.
- India Leads in Recovered Cases then Brazil.
- Lowest Death Counts: Samoa, Dominica, Kiribati, and the Marshall Islands.
- Highest Confirmed COVID-19 Cases was in United States.
- Peak Recovered Cases: April 2021.
- Peak Death Rate: January 2021.

Summary

Summary

Data Gathering:

- Collection of data from multiple sources including hospitals, health departments, and research institutes.
- Gathering information on confirmed cases, deaths, recoveries, and demographic details.

Data Cleaning:

- Removing inconsistencies, errors, and missing values from the collected data.
- Ensuring the accuracy of the information for analysis.

Exploratory Analysis:

- Utilizing SQL queries to uncover patterns and trends within the data.
- Investigating factors such as age and gender to understand their influence on outcomes.

Aggregation:

- Summarizing key metrics such as total cases, deaths, and recovery rates.
- Aggregating data for different countries, regions, and time periods to compare the virus impact and track its progression.

Thank You