SQL Analysis

About

This project Covid-19 Analysis aims to gain insights into the global and regional impact of the pandemic, examining trends over time, identifying the hotspots, vaccination details, etc., By querying and analyzing relevant data, I aimed to contribute to a better understanding of the virus`s spread and its implications on public health.

The dataset is obtained from the **WHO-COVID-19-global-data.csv**

Purpose

The major aim of this project is to understand how the virus is spreading, see if our actions to control it are working, and figure out where the resources are needed most.

About the Dataset (Schema)

The dataset is obtained from the **WHO-COVID-19-global-data.csv** and from **Kaggle-world-population.** With information about cases, deaths, vaccination details, vaccine types and the populations based on countries.

Covid-19 Global dataset

Column	Description	Datatype
Date_reported	Daily date on which cases are reported	DATE
Country_code	Code for standardized identification	STRING
Country	Specifies the name of country	STRING
WHO_region	WHO region to which the country belongs	STRING
Cases	Number of cases in a specific country on a given date	INTEGER
Deaths	Number of Deaths in a specific country on a given date	INTEGER

Vaccination-data

| Column	Description	Datatype
Country	Code for standardized identification	STRING
ISO3	Specifies the name of country	STRING
WHO_region	WHO region to which the country belongs	STRING
Total_Vaccination	Total count of vaccinations administered	INTERGER
First_Dose	People who receiving the initial vaccine dose	INTEGER
Second_Dose	People who completed the vaccination series with second dose	INTEGER
Booster_Dose	People who receiving additional booster doses	INTEGER

_Vaccination_type_

Column	Description	Datatype
ISO3	Specifies the name of country	STRING
Product_Name	Category of the vaccine product.	STRING
Vaccine_Name	Type of the COVID-19 vaccine.	STRING
Company_Name	Pharmaceutical or biotech company that produced the COVID-19 vaccine.	STRING

Column	Description	Datatype
Country	Code for standardized identification	STRING
Population	Population of the country 2023	INTEGER

ER Diagram
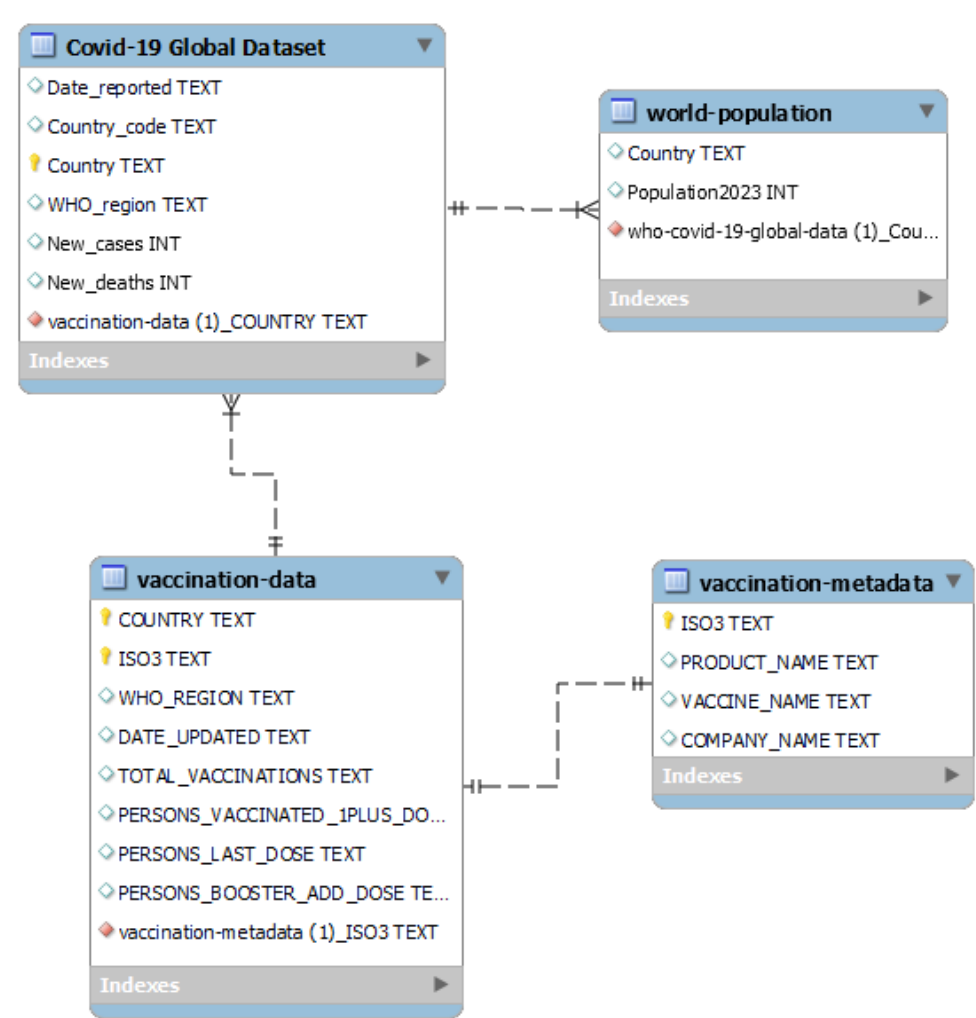
Data Cleaning

Addressing missing data points in the dataset by employing data cleaning techniques. The process involves identifying and handling null values, ensuring data accuracy, and preparing the dataset for further analysis.

Tailoring the dataset by eliminating unnecessary or redundant columns. This enhances data clarity and reduces unnecessary information, streamlining the dataset for analysis.

Analysis Type

- **Global COVID Insight Vault**

  Explore and analyze the global impact of the COVID-19 pandemic.

- **Vaccine Typology Analysis**

  Categorize and evaluate COVID-19 vaccines based on their unique characteristics.

- **Vaccine types Analysis**

  Explore the nuances of COVID-19 vaccine formulations and their impact.

Query Exploration

Global COVID Insight Vault

1. What is the total number of covid-19 cases and deaths globally?

```
Select sum(New_deaths) as Totatal_deaths, sum(New_cases) as Total_cases
from `Covid_data_analysis.Global_data`
```

Query results

	JOB INFORMATION	RESULTS	CHART	PREVIEW

Row	Totatal_deaths ▼	Total_cases ▼
1	6990067	773119173

2. How has the pandemic evolved over time?

```
SELECT EXTRACT(year from date_reported) as Year ,SUM(new_cases) AS total_cases,SUM(new_deaths) A
FROM `Covid_data_analysis.Global_data`
GROUP BY Year
ORDER BY Year;
```

Query results

	JOB INFORMATION	RESULTS	CHART	PREVIEW	JSON

Row	Year ▼	total_cases ▼	total_deaths ▼
1	2020	80316790	1897533
2	2021	200297624	3549358
3	2022	424017189	1248805
4	2023	68487570	294371

3. Which countries have experienced the highest and lowest total covid-19 cases and deaths?

```
WITH CountryTotals AS (
  SELECT Country, SUM(new_cases) AS Total_cases, SUM(new_deaths) AS Total_deaths
  FROM `Covid_data_analysis.Global_data`
  GROUP BY Country
)
SELECT Country, Total_Cases, Total_Deaths
FROM (
  SELECT Country, Total_cases, Total_deaths,
  ROW_NUMBER() OVER (ORDER BY Total_cases DESC) AS RowAsc,
  ROW_NUMBER() OVER (ORDER BY Total_cases ASC) AS RowDesc
  FROM CountryTotals
) AS RANkDATA
WHERE RowDesc <=5 OR RowAsc <= 5;
```

Query results

Row	Country	Total_Cases	Total_Deaths
1	United States of America	103436829	1144877
2	China	99321898	121880
3	India	45008620	533333
4	France	38997490	167985
5	Germany	38437756	174979
6	Tokelau	80	0
7	Holy See	26	0
8	Pitcairn	4	0
9	Democratic People's Republic ...	0	0
10	Turkmenistan	0	0

4. What are the daily trends in new cases and deaths?

```
WITH dailyTrends As(
    Select Date_reported, SUM(new_cases) AS Daily_Total_cases, SUM(new_deaths) AS Daily_Total_deat
    from `Covid_data_analysis.Global_data`
    group by Date_reported
)
select Date_reported, Daily_Total_cases, Daily_Total_deaths
from dailyTrends;
```

Query results

Row	Date_reported	Daily_Total_cases	Daily_Total_deaths
1	2020-01-05	2	3
2	2020-01-12	45	1
3	2020-01-19	90	2
4	2020-01-26	1895	56
5	2020-02-02	12538	310
6	2020-02-09	23059	545
7	2020-02-16	31734	864
8	2020-02-23	9578	692
9	2020-03-01	8272	519
10	2020-03-08	20207	650

5. How does the (Case Fatality Rate) CFR vary between countries?

```
with CFR as(
    Select Country, sum(new_cases) as Total_cases, sum(new_deaths) as Total_Deaths
    from `Covid_data_analysis.Global_data`
    group by Country
)
Select Country, round((Total_deaths/nullif(Total_cases,0)),2)*100 as Fatality_Rate
from CFR
order by Fatality_Rate DESC
```

Query results

Row	Country	Fatality_Rate
1	Yemen	18.0
2	Sudan	8.0
3	Syrian Arab Republic	6.0
4	Egypt	5.0
5	Peru	5.0
6	Somalia	5.0
7	Bosnia and Herzegovina	4.0
8	Liberia	4.0
9	Mexico	4.0
10	Afghanistan	3.0
11	Bulgaria	3.0
12	Algeria	3.0
13	Ecuador	3.0

6. What is the death ratio and total covid - 19 cases based on the population?

```
With DeathRation as (
  Select distinct GD.country, WP.Population2023, sum(GD.New_deaths) as Total_deaths
  from `Covid_data_analysis.Global_data` as GD
  join
  `Covid_data_analysis.World_population` as WP
  using(Country)
  where country is not null
  group by GD.country, WP.population2023
)
Select country,Total_deaths/nullif(population2023,0)*100 as DeathRation
from DeathRation
order by DeathRation DESC
```

Query results

Row	country	DeathRation
1	Peru	0.644999890692...
2	Bulgaria	0.576399988217...
3	Bosnia and Herzegovina	0.510052331985...
4	Hungary	0.481438059895...
5	North Macedonia	0.477542325544...
6	Croatia	0.461231392273...
7	Georgia	0.459514596803...
8	Slovenia	0.454645169660...
9	Montenegro	0.423633446930...
10	Latvia	0.405964121076...
11	San Marino	0.374531835205...
12	Greece	0.368194372899...

Vaccine Typology Analysis

1. How has the administration of first doses evolved over time globally?

```
Select DATE_UPDATED, Sum(PERSONS_VACCINATED_1PLUS_DOSE) as number_of_1st_dose from `Covid_data_a
group by DATE_UPDATED;
```

Row	DATE_UPDATED ▼	number_of_1st_dose
1	2023-11-26	88604440.0
2	2023-11-23	1259722441.0
3	2023-11-22	886838.0
4	2023-11-19	33553850.0
5	2023-11-17	303249030.0
6	2023-11-14	165567890.0
7	2023-11-12	3314463.0
8	2023-11-07	15311319.0
9	2023-11-05	94263266.0
10	2023-10-26	28140956.0

2. Is there a correlation between the rollout of second doses and the decline in new COVID-19 cases?

```
Select v.DATE_UPDATED, sum(g.New_cases) as Totla_cases , sum(v.PERSONS_LAST_DOSE) as number_of_2
left join `Covid_data_analysis.Global_data` as g
on v.DATE_UPDATED = g.Date_reported
group by v.DATE_UPDATED
order by DATE_UPDATED
```

Row	DATE_UPDATED ▼	Total_cases ▼	number_of_2st_dose
1	2022-02-27	11049873	3636858480.0
2	2022-05-22	3782153	5671920.0
3	2022-06-05	3259098	1532575440.0
4	2022-07-24	7173720	51367680.0
5	2022-07-31	7204705	156988560.0
6	2022-08-21	5386339	1093995120.0
7	2022-09-04	4289886	1555484640.0
8	2022-09-11	3269394	12183112320.0

3. Which countries have achieved the highest vaccination coverage for the first dose?

```
with Vaccination_Coverage as (
Select COUNTRY, sum(PERSONS_VACCINATED_1PLUS_DOSE) as Number_of_1st_dose
from
`Covid_data_analysis.Vaccination_Data`
Group by COUNTRY
)
Select * ,
Rank() over(Order by Number_of_1st_dose DESC) as Country_Rank
from Vaccination_Coverage
order by Country_Rank
```

Row	COUNTRY	Number_of_1st_dose	Country_Rank
1	China	1318026807.0	1
2	India	1027419851.0	2
3	United States of America	270227181.0	3
4	Indonesia	204419394.0	4
5	Brazil	184663245.0	5
6	Pakistan	165567890.0	6
7	Bangladesh	151504394.0	7
8	Japan	104737006.0	8
9	Mexico	96727823.0	9
10	Nigeria	93829435.0	10

4. Are there disparities in the distribution of booster doses among different countries?

```
WITH Booster_Dose_Coverage AS (
SELECT COUNTRY,
Sum(PERSONS_BOOSTER_ADD_DOSE) as Number_of_Booster_doses
FROM `Covid_data_analysis.Vaccination_Data`
GROUP BY COUNTRY
)
SELECT*,
    RANK() OVER (ORDER BY Number_of_Booster_doses DESC) AS Country_Booster_Rank
FROM Booster_Dose_Coverage
order by Country_Booster_Rank;
```

Row	COUNTRY	Number_of_Booster_	Country_Booster_Ra
1	China	834060133.0	1
2	India	228816435.0	2
3	United States of America	118438297.0	3
4	Brazil	110565285.0	4
5	Japan	86596573.0	5
6	Indonesia	69356250.0	6
7	Bangladesh	68532189.0	7
8	Viet Nam	57975220.0	8
9	Mexico	57014052.0	9
10	Pakistan	53152375.0	10

Vaccine types Analysis

1. Which companies are the major contributors to the production of COVID-19 vaccines?

```
Select COMPANY_NAME,  count(Vaccine_Name) as Total_vaccines
from `Covid_data_analysis.Vaccine_Type`
group by COMPANY_NAME
order by Total_vaccines DESC
```

Row	COMPANY_NAME	Total_vaccines
1	Pfizer BioNTech	174
2	AstraZeneca	147
3	Moderna	129
4	Janssen Pharmaceuticals	125
5	Beijing Bio-Institute Biological ...	101
6	Gamaleya Research Institute	99
7	Serum Institute of India	92
8	Sinovac	65
9	Bharat Biotech	33
10	Novavax	32

2. How many unique vaccine products are available?

```
Select count(Distinct PRODUCT_NAME) as Unique_Products
from `Covid_data_analysis.Vaccine_Type`
```

Row	Unique_Products
1	35

Conclusion

The comprehensive analysis of the global COVID-19 landscape reveals significant insights. Total cases and deaths globally have been thoroughly examined, shedding light on the magnitude of the pandemic. The evolution of the virus over time demonstrates varying trends, with distinct patterns emerging in different regions. Country-wise comparisons highlight hotspots and areas with both high and low COVID-19 impact. Daily trends provide real-time perspectives, while the Case Fatality Rate offers a measure of severity across nations. Assessing death ratios based on population underscores the varying impact on different countries.

Turning to vaccine typology, the analysis delves into the evolution of first doses administered globally, identifying potential correlations between the rollout of second doses and declines in new cases. Vaccination coverage is scrutinized to recognize countries achieving significant milestones, while an investigation into booster dose distribution reveals potential disparities among nations.

Finally, the analysis of vaccine types elucidates major contributors to vaccine production and emphasizes the diversity of available vaccine products. The overall project serves its primary purpose of contributing to a deeper understanding of the global and regional implications of the pandemic, aiding in informed decision-making and resource allocation for effective public health interventions.