**COVID-19 Database Project**

Explore COVID-19 data using SQL!

**Project Overview**  
In this project, I performed queries on a COVID-19 database. This database contains statistical data on COVID-19, including the number of infections and deaths worldwide, the number of tests conducted, vaccines administered daily, and general information about regions and their healthcare quality.

**Objectives**:

* Utilize fundamental SQL functionalities
* Answer questions using correct SQL queries
* Discover interesting insights using analytical functions
* Apply data-cleaning knowledge
* Study up-to-date COVID-19 data

**My Role**  
I have just joined a team of COVID-19 researchers. My task is to analyze historical data and healthcare data across different regions.

**Database**  
The COVID-19 data is provided by **Our World in Data**.

**Part 0**

**Data Filtering**

1. Check if any of the tables have duplicate rows

My query:

SELECT \*

FROM (SELECT \*,

ROW\_NUMBER() OVER (PARTITION BY iso\_code, date ORDER BY date ASC) as row\_num

FROM da-nfactorial.covid19.cases

ORDER BY iso\_code, date ASC) as Table1

WHERE row\_num>1

1. iso\_code should contain only 3 letters. Are there any countries with iso\_code that contain different amounts of letters? Kosovo – OWID\_KOS

My query:

SELECT t.iso\_code,

validation,

r.location as Country

FROM (SELECT

iso\_code,

CASE

WHEN REGEXP\_CONTAINS(iso\_code,'^[A-Z]{3}$') THEN 'Valid'

ELSE 'Invalid'

END AS validation

FROM da-nfactorial.covid19.regions

) as t

JOIN da-nfactorial.covid19.regions r

ON t.iso\_code = r.iso\_code

WHERE validation = 'Invalid'

1. We need to identify if there are any Islands included in this database. Look for countries containing “Islands” in the name.

My query:

SELECT

location as Country,

iso\_code

FROM da-nfactorial.covid19.regions

WHERE location like '%Islands%'

1. Identify countries that have their name input in ‘’

My query:

SELECT

location as Country,

iso\_code

FROM da-nfactorial.covid19.regions

WHERE REGEXP\_CONTAINS(location, '\\(.\*\\)')

1. Change all the NULL values to 0.

My query:

SELECT iso\_code,date,

COALESCE(total\_cases,0) as total\_cases,

COALESCE(new\_cases,0) as new\_cases,

COALESCE(total\_deaths,0) as total\_deaths,

COALESCE(new\_deaths,0) as new\_deaths,

FROM da-nfactorial.covid19.cases

WHERE new\_cases IS NULL

**Part 1**

**Question 1: Which country had the highest deathrate from COVID-19**

My query:

WITH Table1 as (SELECT Country, iso\_code, date, prob,

RANK() OVER (ORDER BY prob DESC) as Ranking

FROM (SELECT

r.location as Country,

c.iso\_code,

date,

(c.total\_deaths\*100 / c.total\_cases) as Prob

FROM da-nfactorial.covid19.cases c

JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

GROUP BY 1,2,3,4)

ORDER BY Prob DESC)

SELECT \*

FROM Table1

WHERE Ranking =1

ORDER BY Date ASC

**Question 2: What is the percentage of COVID-19 infected population of every country. What is the deathrate of COVID-19 in every country?**

My query:

With Table1 as (

SELECT Country, t1.iso\_code, all\_cases,all\_deaths,

all\_cases\*100 / d.population as InfectRate,

all\_deaths\*100/d.population as DeathRate

FROM (SELECT

r.location as Country,

r.iso\_code,

SUM(c.new\_cases) as all\_cases,

SUM(c.new\_deaths) as all\_deaths

FROM da-nfactorial.covid19.cases c

JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

GROUP BY 1,2

ORDER BY 1 ASC) as t1

LEFT JOIN da-nfactorial.covid19.demography d

ON d.iso\_code = t1.iso\_code)

SELECT \*

FROM Table1

ORDER BY InfectRate DESC

**Question 3: What is the COVID-19 infected population rate in the world? What is the COVID-19 death rate in the world?**

My query:

SELECT all\_cases,all\_deaths,world\_population,

all\_cases\*100/world\_population as InfectRate,

all\_deaths\*100/world\_population as DeathRate

FROM(SELECT

SUM(c.new\_cases) as all\_cases,

SUM(c.new\_deaths) as all\_deaths,

SUM(distinct(d.population)) as world\_population

FROM da-nfactorial.covid19.cases c

LEFT JOIN da-nfactorial.covid19.demography d

ON c.iso\_code = d.iso\_code)

**Question 4: Which countries did well in terms of COVID-19 treatment?**

In this project, a country is considered to have managed treatment well if the latest recorded number of ICU patients is lower than the number observed initially.

My query:

WITH Table1 as (SELECT distinct(Country),

FIRST\_VALUE(date) OVER(PARTITION BY iso\_code ORDER BY date ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) as FirstDate,

FIRST\_VALUE(icu\_patients) OVER(PARTITION BY iso\_code ORDER BY date ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) as FirstDayPatients,

LAST\_VALUE(date) OVER(PARTITION BY iso\_code ORDER BY date ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) as LastDate,

LAST\_VALUE(icu\_patients) OVER(PARTITION BY iso\_code ORDER BY date ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) as LastDatePatients,

FROM (SELECT distinct(r.location) as Country,c.iso\_code,h.date as date, icu\_patients

FROM da-nfactorial.covid19.cases c

LEFT JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

JOIN da-nfactorial.covid19.hospital h

ON h.iso\_code = c.iso\_code))

SELECT Country,FirstDate,FirstDayPatients,LastDate,LastDatePatients,

FirstDayPatients - LastDatePatients

FROM Table1

WHERE FirstDayPatients - LastDatePatients > 0

ORDER BY FirstDayPatients - LastDatePatients DESC

**Question 5: How did the UK do in terms of COVID-19 treatment?**

Provide data on new confirmed cases and deaths, the number of new tests, new vaccine doses, and the number of patients admitted to hospitals and intensive care units for the first time, by month. This can be done for all three years or for a representative year.

My query:

WITH hospital as (SELECT \*,

icu\_patients - LAG(icu\_patients) OVER (order by date) as new\_icu\_patients,

hosp\_patients - LAG(hosp\_patients) OVER (order by date) as new\_hosp\_patients

FROM (SELECT iso\_code,

date,

IFNULL(icu\_patients, 0) as icu\_patients,

IFNULL(cast(hosp\_patients as float64), 0) as hosp\_patients

FROM da-nfactorial.covid19.hospital

) as t

WHERE iso\_code="GBR" order by date),

Table1 as (SELECT

DATE\_TRUNC(a.date, month) as month,

a.iso\_code,

new\_cases,

new\_deaths,

new\_tests,

new\_vaccinations,

new\_icu\_patients,

new\_hosp\_patients

FROM

da-nfactorial.covid19.cases a

left join hospital b

on a.iso\_code=b.iso\_code and a.date=b.date

left JOIN da-nfactorial.covid19.tests c

ON a.iso\_code=c.iso\_code and a.date=c.date

left JOIN da-nfactorial.covid19.vaccinations d

ON a.iso\_code=d.iso\_code and a.date=d.date

WHERE

a.iso\_code = "GBR"

and a.date between '2021-01-01' and '2021-12-31'

)

Select month,

SUM(new\_cases) as new\_cases,

SUM(new\_deaths) as new\_deaths,

SUM(new\_tests) as new\_tests,

SUM(new\_vaccinations) as new\_vaccinations,

SUM(new\_icu\_patients) as new\_icu\_patients,

SUM(new\_hosp\_patients) as new\_hosp\_patients

FROM Table1

GROUP BY month

ORDER BY month

**Question 6: How did the number of new confirmed cases change on a daily basis within countries?**

**To answer this question, use the relative change formula.**  
Relative change = (new cases - new cases on the previous day) / new cases on the previous day \* 100

**Provide the following information**:

* Country names
* Observation date
* New confirmed cases
* New cases on the previous day
* Relative change

**Additionally, add a column named Trend that contains the following information**:

* ‘Increase’ if the relative change is positive
* ‘Decrease’ if the relative change is negative
* ‘No change’ if there is no change

My query:

WITH Table1 as (SELECT Country,date,new\_cases,prev\_new\_cases,

CASE

WHEN prev\_new\_cases != 0 THEN (new\_cases - prev\_new\_cases) \* 100 / prev\_new\_cases

ELSE NULL

END AS relat\_diff

FROM (SELECT r.location as Country, c.date as date, c.new\_cases,

lag(new\_cases) OVER(PARTITION BY r.location ORDER BY c.date) as prev\_new\_cases

FROM da-nfactorial.covid19.cases c

LEFT JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

ORDER BY 1 ASC))

SELECT \*,

CASE

WHEN new\_cases > prev\_new\_cases THEN "Increase"

WHEN new\_cases < prev\_new\_cases THEN "Decrease"

WHEN new\_cases = prev\_new\_cases THEN "No change"

ELSE NULL

END AS Trend

FROM Table1

**Question 7: Which countries recorded the highest number of confirmed cases between March 20 and March 30, 2020?**

We want the country with the highest number of confirmed cases on a given day to be ranked 1, the second highest to be ranked 2, and so on. You should identify the top-ranked country for each day from March 20 to March 30.

My query:

SELECT \*

FROM (SELECT r.location as Country, c.date as date, c.new\_cases,

RANK() OVER(PARTITION BY date ORDER BY new\_cases DESC) as Ranking

FROM da-nfactorial.covid19.cases c

LEFT JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

WHERE date BETWEEN '2020-03-20' AND '2020-03-30'

ORDER BY 2 ASC)

WHERE Ranking = 1

ORDER BY date ASC

**Question 8: What are the top 25 countries in terms of COVID-19 Death Rate?**

My query:

WITH Table1 as (SELECT \*,

RANK() OVER(ORDER BY death\_rate DESC) as Ranking

FROM(SELECT r.location as Country, c.date as date, c.new\_deaths, d.population,

(c.new\_deaths/d.population)\*100 as death\_rate

FROM da-nfactorial.covid19.cases c

LEFT JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

LEFT JOIN da-nfactorial.covid19.demography d

ON c.iso\_code = d.iso\_code

ORDER BY 5 DESC))

SELECT \*

FROM Table1

WHERE Ranking BETWEEN 1 and 25

ORDER BY Ranking ASC

**Вопрос 9:What should Kazakhstan expect in the next 5 days?**

I decided to use the pandemic growth factor between two days as a measure of prediction. It is calculated by dividing the number of confirmed cases on a given day by the number of confirmed cases on the previous day.

Pandemic growth factor for day N = (number of cases on day N) / (number of cases on day N-1)

For a more accurate result, take the average of the pandemic growth factor over the last 10 days.

Number of cases after N days = (number of cases today) \* (pandemic growth factor) ^ N

My query:

WITH Table1 as (SELECT \*,

Coalesce(CASE

WHEN prev\_cases !=0 THEN new\_cases / prev\_cases

ELSE 0

END,0) as Gfact

FROM(SELECT r.location as Country, c.date as date, c.new\_cases,

lag(new\_cases) OVER(ORDER BY c.date ASC) as prev\_cases

FROM da-nfactorial.covid19.cases c

LEFT JOIN da-nfactorial.covid19.regions r

ON c.iso\_code = r.iso\_code

WHERE r.location = 'Kazakhstan'

AND date <= '2021-07-30'))

SELECT \*,

new\_cases \* (POWER(GfactAVG,5)) as forecast

FROM(SELECT \*,

AVG(Gfact) OVER(ORDER BY Date ROWS BETWEEN 9 preceding AND CURRENT ROW) as GfactAVG

FROM Table1

ORDER by date DESC)

**Part 2**

**Define four questions about COVID-19 that you want to answer based on data analysis.**

* Write SQL queries to obtain the data needed to successfully answer your questions.
* Visualize the obtained data (using histograms or other charts) that answer your question.
* Explain the answer in 1-2 sentences.
* The queries should include analytical functions.

**Question 1: On which day was the highest number of confirmed COVID-19 cases reported worldwide?**

My query and explanation with graph:

SELECT \*

FROM(SELECT date, new\_cases,

rank() OVER(ORDER BY new\_cases DESC) as ranking

FROM(SELECT date,

SUM(new\_cases) as new\_cases

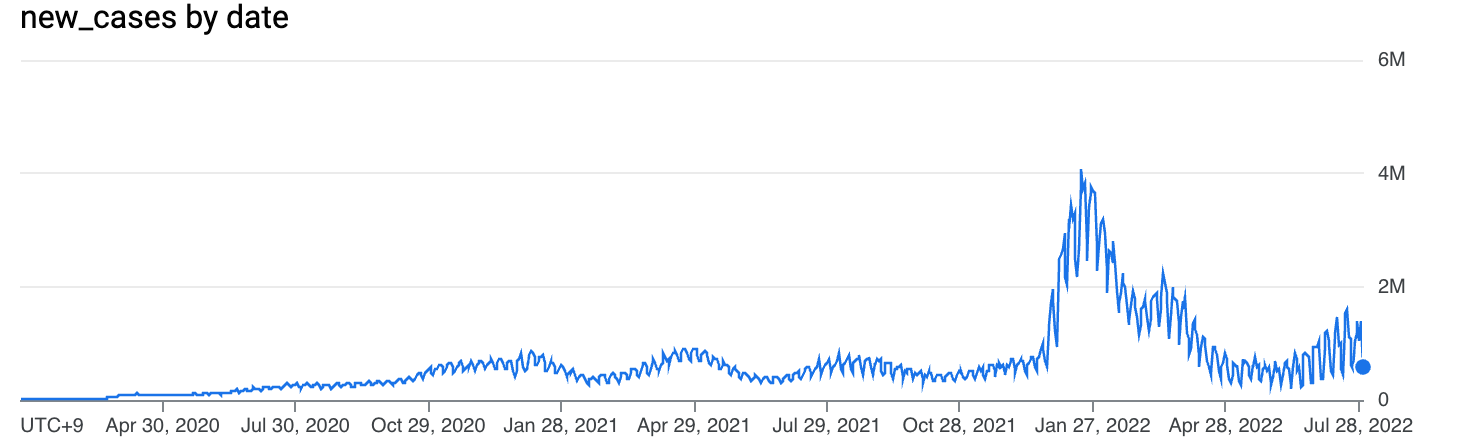
FROM da-nfactorial.covid19.cases

GROUP BY 1))

ORDER BY ranking ASC

Для графика:

ORDER BY date ASC вместо ranking ASC



As the chart shows, the majority of confirmed cases were recorded in early 2022 rather than in 2020. This is likely due to two reasons:

-By that time, COVID-19 was perceived as a less dangerous and less lethal threat due to vaccination.

-Easing or complete removal of lockdown measures worldwide.

**Question 2: At what point did countries have the highest percentage of fully vaccinated population?**

My query with graph and explanation:

SELECT \*

FROM(SELECT \*,

RANK() OVER(PARTITION BY iso\_code ORDER BY ratio DESC) as ratio\_ranking

FROM(SELECT \*, fully\_vaccinated / population as ratio

FROM(SELECT v.iso\_code,v.date,v.people\_fully\_vaccinated as fully\_vaccinated, d.population

FROM da-nfactorial.covid19.vaccinations v

JOIN da-nfactorial.covid19.demography d

ON v.iso\_code = d.iso\_code

GROUP BY 1,2,3,4

ORDER BY 1 ASC)))

WHERE ratio\_ranking = 1

ORDER BY 1 ASC



This scatter plot demonstrates that the vast majority of countries had the highest percentage of vaccinated population relative to the total population closer to the end of the period. This result is explained by the fact that the vaccine became available only after the virus emerged, and the vaccination process itself takes a considerable amount of time. However, it is worth noting that there are countries where the maximum percentage of vaccinated population was less than 10%.

**Question 3: In which country was the highest percentage of the population vaccinated in a single day?**

My query with graph and explanation:

WITH Table1 as (SELECT \*,

RANK() OVER(PARTITION BY iso\_code ORDER BY vac\_rate DESC) as ranking

FROM(SELECT v.iso\_code,v.date,v.total\_vaccinations as people\_vaccinated, d.population, v.total\_vaccinations\*100 / d.population as vac\_rate

FROM da-nfactorial.covid19.vaccinations v

JOIN da-nfactorial.covid19.demography d

ON v.iso\_code = d.iso\_code

GROUP BY 1,2,3,4

ORDER BY 1 ASC)

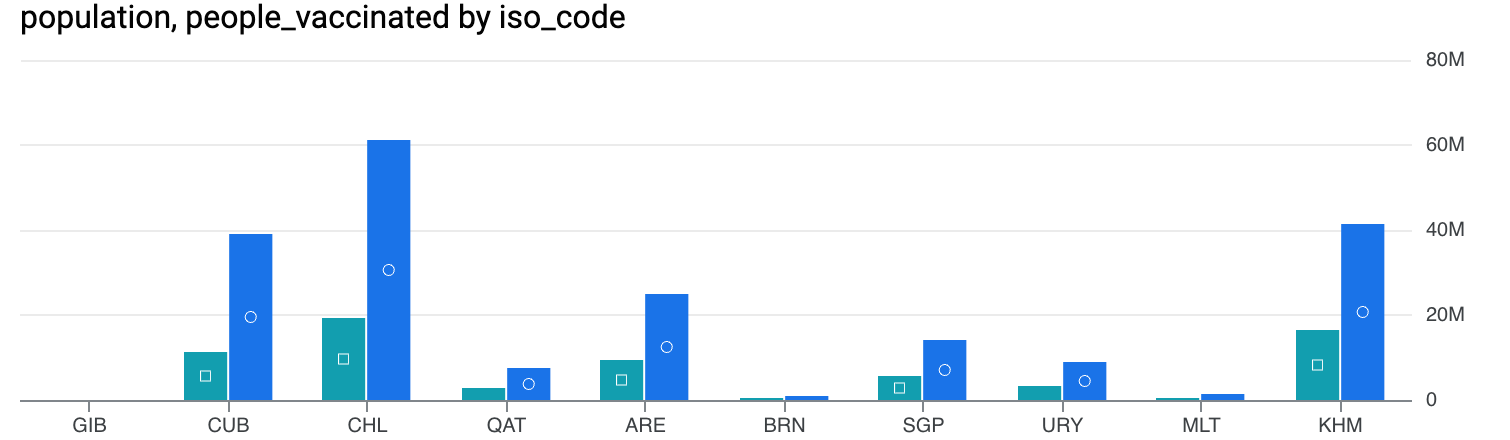
ORDER BY iso\_code ASC ,ranking ASC)

SELECT \*

FROM Table1

WHERE ranking = 1

ORDER BY table1.iso\_code ASC



This bar chart shows the ratio of the country's population to the number of people who received the vaccine on the most productive vaccination day in each country. To our surprise, in all cases, the number of vaccinated people significantly exceeded the country's population. This may be due to the inclusion of foreign visitors to the country for vaccination purposes.

**Question 4: In which countries was the mortality rate higher than the global average?**

My query with graph and explanation:

WITH Table1 as (SELECT \*,total\_deaths \*100 / population as death\_rate

FROM(SELECT c.iso\_code, sum(c.new\_deaths) as total\_deaths,d.population

FROM da-nfactorial.covid19.cases c

JOIN da-nfactorial.covid19.demography d

ON c.iso\_code = d.iso\_code

GROUP BY 1,3)),

Table2 as (SELECT AVG(death\_rate) as avg\_death\_rate

FROM Table1)

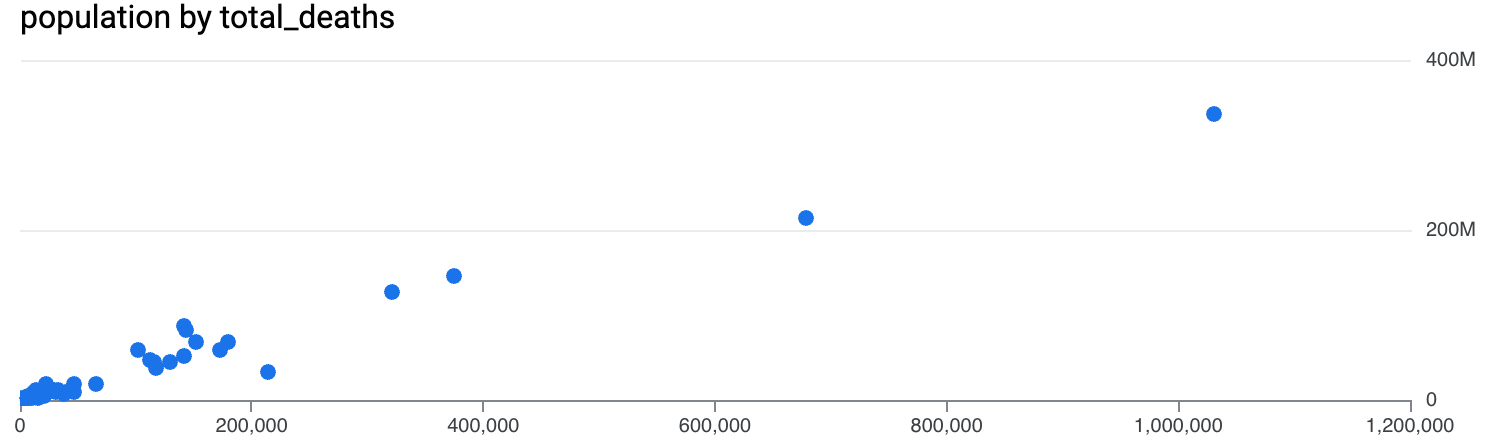
SELECT \*

FROM Table1

CROSS JOIN Table2

WHERE Table1.death\_rate > Table2.avg\_death\_rate

ORDER BY Table1.death\_rate DESC



This chart shows the ratio of the country's population to the number of COVID-19 deaths in the context of countries with a COVID-19 mortality rate exceeding the global average. The X-axis represents the total population of the country, while the Y-axis shows the number of COVID-19 deaths. The chart does not reveal any cases of exceptionally high mortality levels.