**COVID-19 Database project**

**Исследуйте данные о COVID-19 с помощью SQL!**

**Обзор проекта**

В этом проекте вы будете выполнять запросы для базы данных COVID-19. Эта база данных содержит статистические данные о COVID-19 такие как количество заражении и смертей во всем мире, количество тестов и полученных вакцин в день и общую информацию о регионах и их качестве медицины.

Ваши задачи:

* использовать основные функционалы SQL
* отвечать на вопросы с помощью правильных запросов
* найти интересные инсайты в данных используя аналитические функции
* применить знания по очистке данных
* изучить актуальные данные о COVID-19

**Ваша роль**

Вы только что попали в команду исследователей по COVID-19. Ваша задача - проанализировать исторические данные и данные о состоянии медицины в регионах. Вы собираете информацию для отчета о любых закономерностях, которые, возможно, стоит изучить подробнее.

**База данных**

Актуальные данные по COVID-19 предоставлены [Our World in Data](https://ourworldindata.org/coronavirus).

Чтобы помочь вам с дальнейшими запросами, ниже представлена информация о параметрах в таблицах. Все таблицы между собой связаны двумя столбцами: **iso\_code и date**.

*Cases*

| **Параметр** | **Описание** |
| --- | --- |
| iso\_code | Трехбуквенный код страны |
| date | Дата наблюдения |
| total\_cases | Общее количество подтвержденных случаев COVID-19 |
| new\_cases | Новые подтвержденные случаи COVID-19 |
| total\_deaths | Общее количество смертей, связанных с COVID-19 |
| new\_deaths | Новые смерти, связанные с COVID-19 |

*Demography*

| **Параметр** | **Описание** |
| --- | --- |
| population | Численность населения |
| population\_density | Число жителей, приходящееся на 1 км² территории |
| median\_age | Медианный возраст населения |
| aged\_65\_older | Доля населения в возрасте 65 лет и старше |
| gdp\_per\_capita | Валовой внутренний продукт по паритету покупательной способности |
| extreme\_poverty | Доля населения, живущего в крайней бедности |
| cardiovasc\_death\_rate | Смертность от сердечно-сосудистых заболеваний в 2017 году (годовое число умерших на 100 000 населения) |
| diabetes\_prevalence | Распространенность диабета (% населения в возрасте от 20 до 79 лет) в 2017 году |
| female\_smokers | Доля женщин, которые курят |
| male\_smokers | Доля мужчин, которые курят |
| handwashing\_facilities | Доля населения, имеющего базовые средства для мытья рук в помещениях |
| hospital\_beds\_per\_thousand | Количество больничных коек на 1000 человек |
| life\_expectancy | Ожидаемая продолжительность жизни при рождении в 2019 году |
| human\_development\_index | Составной индекс, измеряющий средние достижения по трем основным параметрам человеческого развития: долгая и здоровая жизнь, знания и достойный уровень жизни |

*Hospital*

| **Параметр** | **Описание** |
| --- | --- |
| icu\_patients | Количество пациентов с COVID-19 в отделениях интенсивной терапии в данный день |
| weekly\_icu\_admissions | Количество пациентов с COVID-19, впервые поступивших в отделения интенсивной терапии за данную неделю (отчетная дата и предшествующие 6 дней) |
| hosp\_patients | Количество пациентов с COVID-19 в больнице в данный день |
| weekly\_hosp\_admissions | Количество пациентов с COVID-19, впервые поступивших в больницы за данную неделю (отчетная дата и предшествующие 6 дней) |

*Regions*

| **Параметр** | **Описание** |
| --- | --- |
| continent | Континент |
| location | Страна |

*Tests*

| **Параметр** | **Описание** |
| --- | --- |
| total\_tests | Общее количество тестов на COVID-19 |
| new\_tests | Новые тесты на COVID-19 |
| positive\_rate | Доля положительных тестов на COVID-19, указанная как среднее за 7 дней |
| tests\_units | Единицы, используемые странами для предоставления данных о тестировании |

*Vaccinations*

| **Параметр** | **Описание** |
| --- | --- |
| total\_vaccinations | Общее количество введенных доз вакцины против COVID-19 |
| people\_vaccinated | Общее количество людей, которые получили минимум одну дозу вакцины |
| people\_fully\_vaccinated | Общее количество людей, которые получили все дозы, предусмотренные первоначальным протоколом вакцинации |
| total\_boosters | Общее количество введенных бустерных доз вакцины против COVID-19 |
| new\_vaccinations | Количество новых доз вакцины введенных против COVID-19 |

Изучите таблицы и данные в таблицах перед тем как начать работать над проектом.

**Инструкции по сдаче проекта**

* После выполнения задания нажмите на кнопку "+ Добавить или создать" в Google Classroom и сохранить файл Google sheets + Google docs
* После загрузки файлов нажмите на кнопку "Сдать".
* Ожидайте обратную связь от менторов.

Давайте перейдем к выполнению задания! Обязательно сохраните ваши запросы для каждого вопроса в этом файле.

**Часть 0**

В данной базе данных немало “грязных данных”. Давайте воспользуемся функциями для чистки данных, чтобы исправить ситуацию.

1. Проверьте, нет ли повторяющихся строк в таблицах (менять данные в таблице не нужно). Поделитесь запросом проверки одной таблицы.

Ваш запрос:

SELECT iso\_code, date, COUNT(\*)

FROM `da-nfactorial.covid19.cases`

GROUP BY iso\_code, date

HAVING COUNT(\*) > 1;

1. iso\_code должен состоять из трех букв. Есть ли в наборе iso\_code, который не соответствует данному критерию?

Ваш запрос:

SELECT iso\_code

FROM `da-nfactorial.covid19.cases`

WHERE LENGTH(iso\_code) != 3;

1. Нам нужно узнать включили ли в наш набор данных острова. Найдите все названия стран в котором есть “Islands”.

Ваш запрос:

SELECT DISTINCT location

FROM `da-nfactorial.covid19.regions`

WHERE location LIKE '%Islands%';

1. Мы хотим убрать текст в скобках в названиях стран. Напишите запрос, который поможет нам с этой задачей.

Ваш запрос:

UPDATE `da-nfactorial.covid19.regions`

SET location = TRIM(REGEXP\_REPLACE(location, '\\(.\*\\)', ''))

WHERE location LIKE '%(%';

1. Посмотрите на типы данных в hospital. Что бы вы изменили и каким запросом бы воспользовались?

Ваш запрос:

| 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 |
| --- |

1. Давайте заменим все NULL значения в cases на 0. Правильно ли мы делаем меняя значения на 0? Почему?

| SELECT COUNT(\*) AS null\_count  FROM da-nfactorial.covid19.cases  WHERE total\_cases IS NULL; |
| --- |

Ваш запрос: помогает узнать, сколько строк в таблице имеют пустое значение в столбце total cases. Это поможет понять, насколько данные полные. Замена пустых значений на 0 может быть разумной, если пусто означает отсутствие информации и это нужно учитывать при анализе.

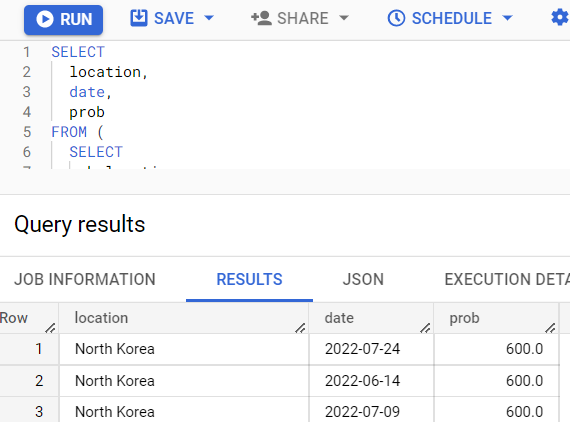
**Часть 1**

**Вопрос 1: В какой стране вероятность смерти инфицированного человека была самой высокой?**

Вероятность смерти инфицированного человека = (количество смертей \ количество подтвержденных случаев) \* 100

Предоставьте название страны, дату наблюдения, вероятность смерти инфицированного человека. Если есть несколько строк с наибольшей вероятностью смерти инфицированного, верните все строки.

*Hint: часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

WITH DeathRates AS (

SELECT

C.iso\_code AS Country,

C.date,

(C.total\_deaths / C.total\_cases) \* 100 AS DeathProbability

FROM

da-nfactorial.covid19.cases C

WHERE

C.total\_cases > 0

)

SELECT \*

FROM DeathRates

WHERE DeathProbability = (

SELECT MAX(DeathProbability) FROM DeathRates

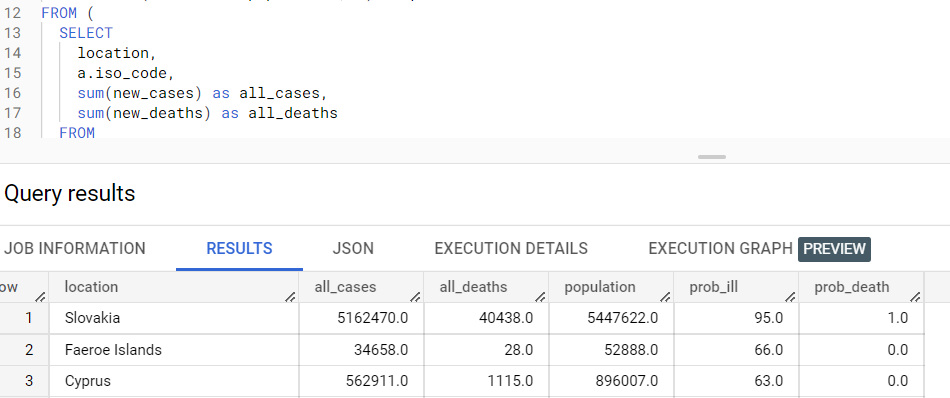
);

**Вопрос 2: Какова доля зараженного населения и доля населения умершего от COVID-19 для каждой страны?**

Доля зараженного населения страны = (общее количество подтвержденных случаев / численность населения) \* 100

Предоставьте название страны, общее количество подтвержденных случаев, общее количество смертей, численность населения, доля зараженного населения страны и доля населения страны умершего от COVID-19. Страна с наибольшей долей зараженного населения должна отображаться первой.

*Hint: часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

SELECT

D.iso\_code AS Country,

SUM(C.total\_cases) AS TotalCases,

SUM(C.total\_deaths) AS TotalDeaths,

D.population AS Population,

(SUM(C.total\_cases) / D.population) \* 100 AS InfectionRate,

(SUM(C.total\_deaths) / D.population) \* 100 AS DeathRate

FROM

da-nfactorial.covid19.cases C

JOIN

da-nfactorial.covid19.demography D ON C.iso\_code = D.iso\_code

GROUP BY

D.iso\_code, D.population

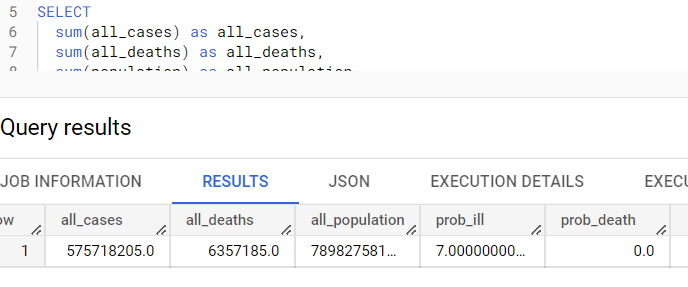
ORDER BY

InfectionRate DESC;

**Вопрос 3: Какова доля зараженного населения и доля населения умершего от COVID-19 в мире?**

Предоставьте общее количество подтвержденных случаев по всему миру, общее количество смертей, численность населения в мире, доля зараженного населения и доля населения умершего от COVID-19.

*Hint: часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

SELECT

SUM(C.total\_cases) AS GlobalTotalCases,

SUM(C.total\_deaths) AS GlobalTotalDeaths,

SUM(D.population) AS GlobalPopulation,

(SUM(C.total\_cases) / SUM(D.population)) \* 100 AS GlobalInfectionRate,

(SUM(C.total\_deaths) / SUM(D.population)) \* 100 AS GlobalDeathRate

FROM

da-nfactorial.covid19.cases C

JOIN

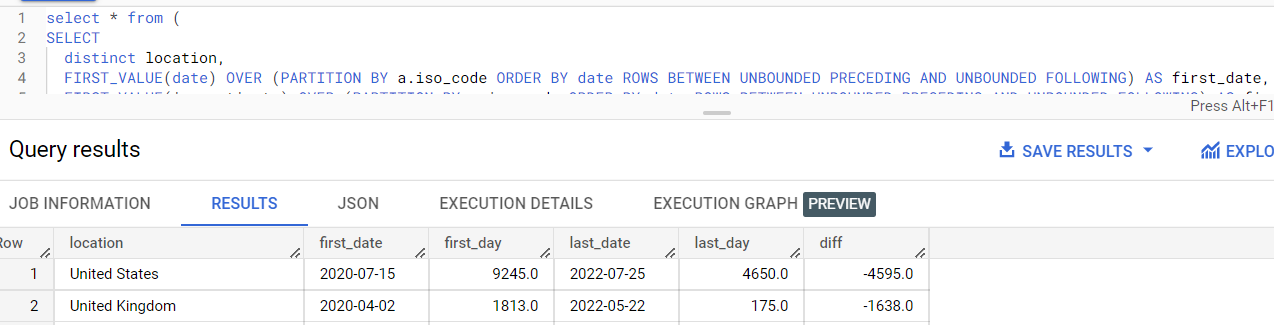
da-nfactorial.covid19.demography D ON C.iso\_code = D.iso\_code;

**Вопрос 4: Какие страны хорошо справились с лечением?**

В рамках этого проекта страна хорошо справилась с лечением если последнее зафиксированное количество пациентов в неотложке (icu patients) меньше чем в первом наблюдении.

Предоставьте названия стран, первую дату наблюдения количество пациентов в неотложке в наборе данных, последнюю дату наблюдения количество пациентов в неотложке в наборе данных и разницу в количестве пациентов.

*Hint: часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос;

WITH ICUData AS (

SELECT

iso\_code AS Country,

date,

icu\_patients,

FIRST\_VALUE(date) OVER (PARTITION BY iso\_code ORDER BY date) AS FirstObservationDate,

FIRST\_VALUE(icu\_patients) OVER (PARTITION BY iso\_code ORDER BY date) AS FirstICUPatients,

LAST\_VALUE(date) OVER (PARTITION BY iso\_code ORDER BY date

ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS LastObservationDate,

LAST\_VALUE(icu\_patients) OVER (PARTITION BY iso\_code ORDER BY date

ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS LastICUPatients

FROM

da-nfactorial.covid19.hospital

)

SELECT

Country,

FirstObservationDate,

FirstICUPatients,

LastObservationDate,

LastICUPatients,

(FirstICUPatients - LastICUPatients) AS Difference

FROM

ICUData

GROUP BY

Country, FirstObservationDate, FirstICUPatients, LastObservationDate, LastICUPatients

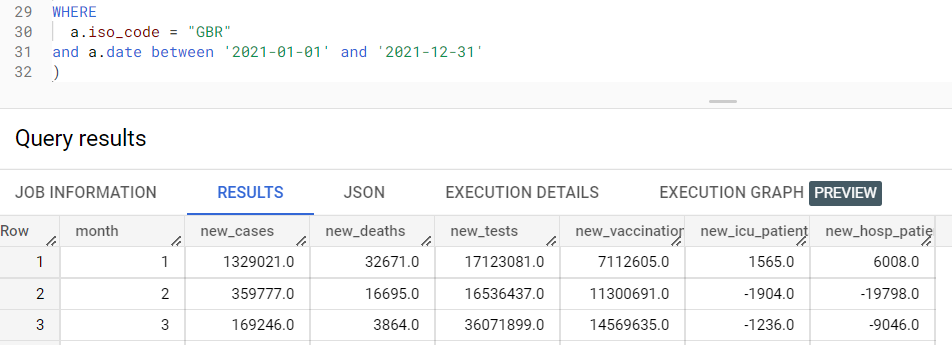
HAVING

LastICUPatients < FirstICUPatients;

**Вопрос 5: Как Великобритания справлялась с COVID-19?**

Предоставьте данные о новых подтвержденных случаях и смертей, о количестве новых тестов и новых доз вакцин, о количество пациентов, впервые поступивших в больницы и в отделения интенсивной терапии по месяцам. (Можно сделать по всем 3 годам, или взять один показательный год)

*Hint: часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

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

**Вопрос 6: Как менялось количество новых подтвержденных случаев на ежедневной основе внутри стран?**

Чтобы ответить на этот вопрос, воспользуйтесь относительным изменением.

Относительное изменение = (новые случаи - новые случаи в предыдущий день) / новые случаи в предыдущий день \* 100

Предоставьте названия стран, дату наблюдения, новые подтвержденные случаи, новые случаи в предыдущий день, относительное изменение. Также добавьте столбец trend, который будет содержать следующую информацию:

* ‘Increase’, если относительное изменение положительное;
* ‘Decrease’, если относительное изменение отрицательное;
* ‘No change’, если нет изменении.

*Hint:часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

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

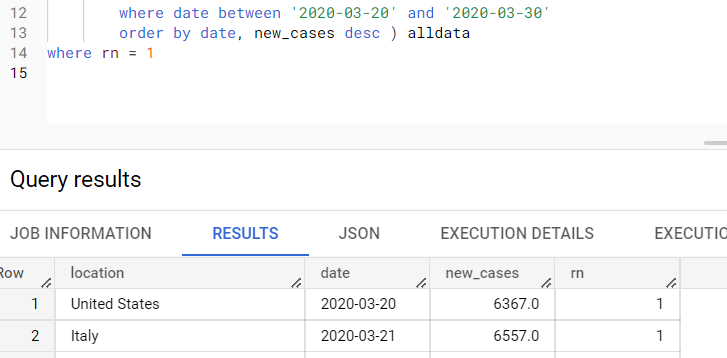
FROM Table1

**Вопрос 7: В каких странах зафиксированы наибольшее количество подтвержденных случаев в период с 20 марта по 30 марта 2020 года?**

Мы хотим, чтобы страна с наибольшим количеством подтвержденных случаев в определенный день имела ранг 1, вторая по величине — ранг 2 и так далее. Вы должны найти топ-1 страну для каждого дня в период с 20 по 30 марта.

Предоставьте данные о названии стран, дату наблюдения, новые подтвержденные случаи (можно вывести ранк, чтобы проверить что вы выбрали только топ-1 стран).

*Hint:часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос: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

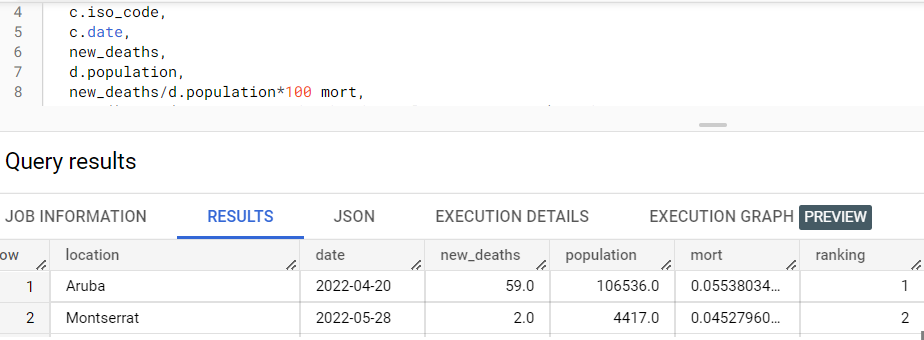
ORDER BY date ASC

**Вопрос 8: Какие 25 стран имели наибольшую смертность во время COVID-19?**

Смертность = (новые смерти / численность населения) \* 100

Предоставьте данные о названии стран, дату наблюдения, новые смерти, численность населения, и уровень смертности.

*Hint:часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

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: Что ожидать Казахстану следующие 5 дней?**

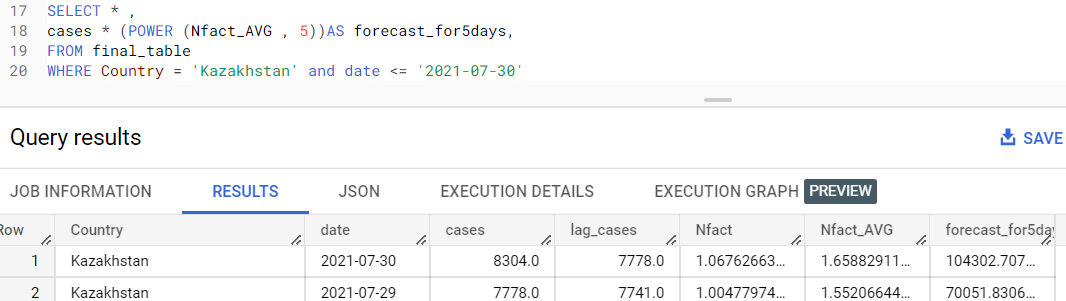
Фактор роста пандемии между двумя днями можно рассчитать, разделив количество подтвержденных случаев за определенный день на количество подтвержденных случаев за предыдущий день.

Фактор роста пандемии за день N = количество случаев в день N / количество случаев в день N-1

Для более точного результата фактором роста пандемии возьмите среднее значение за последние 10 дней.

Количество случаев через N дней = (количество случаев сегодня) \* (фактор роста пандемии) ^ N

*Hint: часть кода показанная в подсказке - не обязательно начало запроса*



Ваш запрос:

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)

**Часть 2**

* Определите 4 вопроса о COVID19, на который вы хотите ответить на основе анализа данных.
* Затем напишете SQL-запросы, чтобы получить данные, необходимые для успешного ответа на ваши вопросы.
* Визуализируйте полученные данные (используя гистограммы или другие графики), отвечающие на ваш вопрос.
* Объясните ответ в 1-2 предложениях.
* Вопросы, которые вы задаете, зависят от вас, но **два запроса должны содержать аналитические функции**.

**Вопрос 1:**

### **Какова средняя смертность от COVID-19 по регионам?**

### **Какова средняя продолжительность жизни по странам и её влияние на смертность от COVID-19?**

Ваш запрос и объяснение с графиком:

### WITH MortalityAndLifeExpectancy AS (SELECT r.location AS Country, d.life\_expectancy AS LifeExpectancy,

### (SUM(c.new\_deaths) / NULLIF(SUM(c.total\_cases), 0)) \* 100 AS MortalityRate

### FROM

### da-nfactorial.covid19.cases c

### JOIN

### da-nfactorial.covid19.regions r ON c.iso\_code = r.iso\_code

### JOIN

### da-nfactorial.covid19.demography d ON c.iso\_code = d.iso\_code

### GROUP BY

### r.location, d.life\_expectancy)

### SELECT

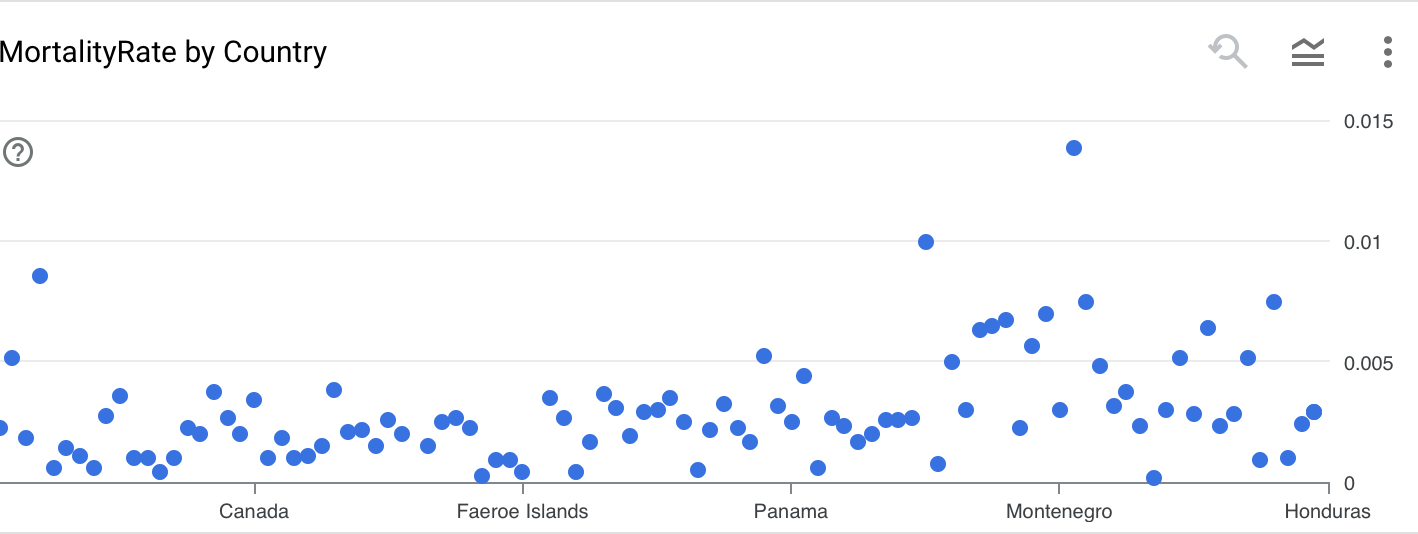
### Country, LifeExpectancy,MortalityRate

### FROM

### MortalityAndLifeExpectancy

### ORDER BY

### LifeExpectancy DESC;



На графике представлена информация о смертности от COVID-19 для различных стран, где ось X соответствует названиям стран, а ось Y — уровню смертности (в процентах) от COVID-19.

**Вопрос 2:**

### **Какова динамика новых случаев и смертей по месяцам в определенной стране?**

### 

Ваш запрос и объяснение с графиком:

SELECT

DATE\_TRUNC(c.date, MONTH) AS Month,

SUM(c.new\_cases) AS NewCases,

SUM(c.new\_deaths) AS NewDeaths

FROM

da-nfactorial.covid19.cases c

WHERE

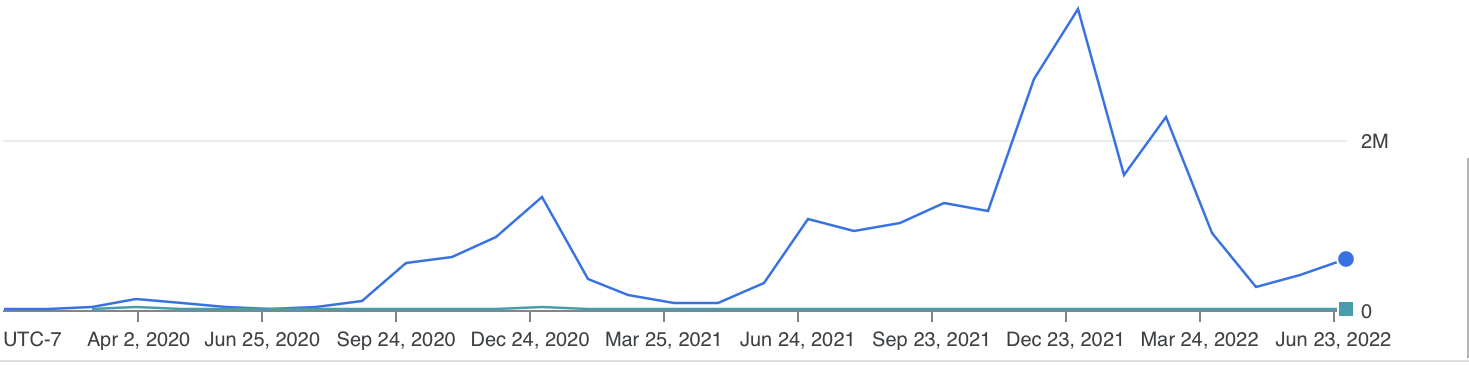
c.iso\_code = 'GBR'

GROUP BY

Month

ORDER BY

Month;



Линейный график, показывающий изменения в новых случаях и смертях по месяцам.

**Вопрос 3:**

### **Какие страны имеют наибольшее количество случаев COVID-19 на душу населения?**

Ваш запрос и объяснение с графиком:

SELECT

r.location AS Country,

SUM(c.total\_cases) AS TotalCases,

d.population,

(SUM(c.total\_cases) / d.population) \* 100 AS CasesPerPopulation

FROM

da-nfactorial.covid19.cases c

JOIN

da-nfactorial.covid19.regions r ON c.iso\_code = r.iso\_code

JOIN

da-nfactorial.covid19.demography d ON c.iso\_code = d.iso\_code

GROUP BY

r.location, d.population

ORDER BY

CasesPerPopulation DESC;

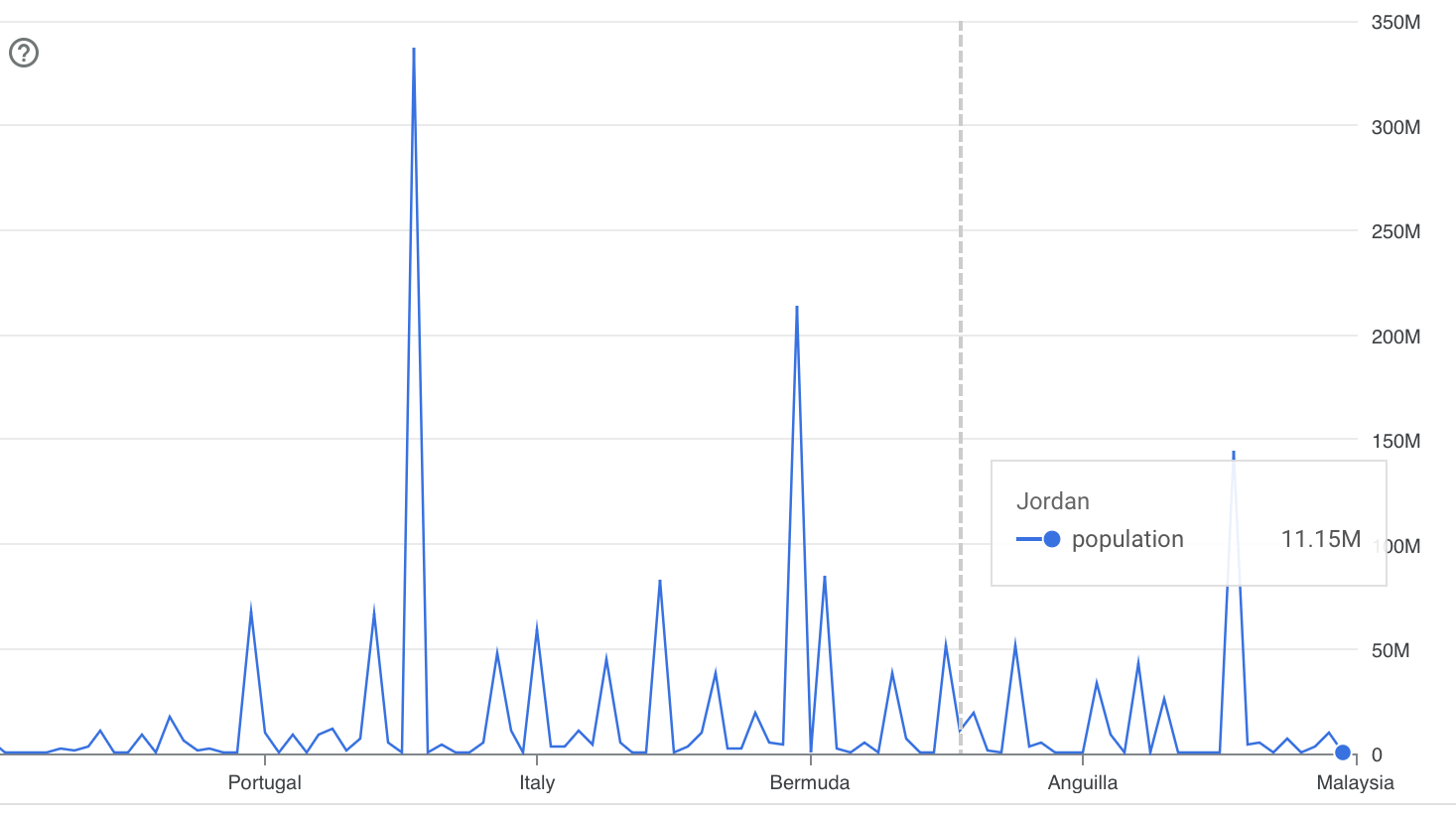


График визуализирует, как количество случаев COVID-19 варьируется по странам относительно их населения. Это важно для понимания масштабов пандемии и разработки стратегий реагирования на COVID-19, а также для анализа влияния различных факторов на распространение вируса.

**Вопрос 4:**

**В какой момент в странах был самый высокий процент полностью вакцинированного населения?**

Ваш запрос и объяснение с графиком:

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

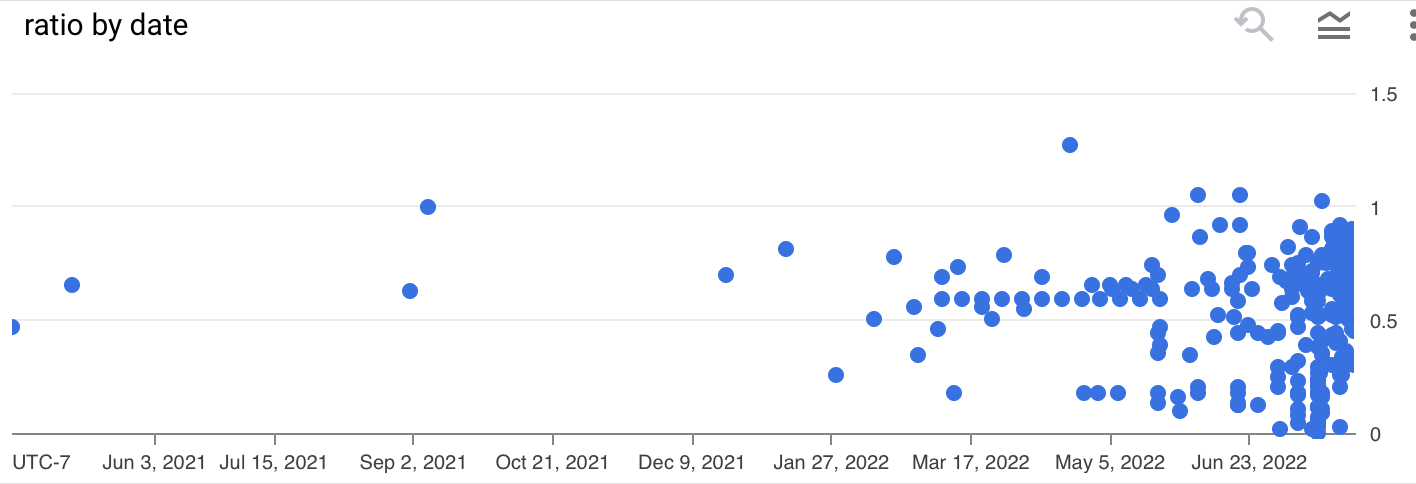
ON v.iso\_code = d.iso\_code

GROUP BY 1,2,3,4

ORDER BY 1 ASC)))

WHERE ratio\_ranking = 1

ORDER BY 1 ASC



Этот график рассеяния показывает, что большинство стран достигли наивысшего уровня вакцинации ближе к концу периода. Это связано с тем, что вакцины стали доступны только после появления вируса, а процесс вакцинации требует времени.

Вы закончили свой COVID анализ! Отличная работа!