**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.

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  hospital\_beds\_per\_thousand life\_expectancy | Доля населения, имеющего базовые средства для мытья рук в помещениях Количество больничных коек на 1000 человек  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 не мало “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:   
Проверил столбцы continent and location from regions  
SELECT

r.continent,

r.location,

COUNT(\*)

FROM da-nfactorial.covid19.regions r

GROUP BY continent, location

HAVING COUNT(\*) > 1;

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

Ваш запрос:   
SELECT

r.iso\_code

FROM da-nfactorial.covid19.regions r

where length(r.iso\_code) != 3

or not regexp\_contains(r.iso\_code, r'^[A-Za-z]{3}$')

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

Ваш запрос:   
select

location

from da-nfactorial.covid19.regions

where location like '%Islands%'

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

Ваш запрос:

select

regexp\_replace(location, r'\(.\*\)', '')

from da-nfactorial.covid19.regions

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

Ваш запрос:

hospital отсутствует

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

Ваш запрос: **Чтобы избежать проблема при вычислениях или аналитике**  
select

c.iso\_code,

c.date,

c.total\_cases,

c.new\_cases,

c.total\_deaths,

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

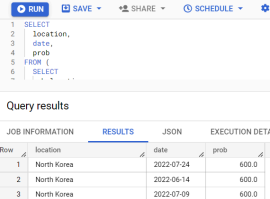
from da-nfactorial.covid19.cases c

**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:   
  
select

r.location,

c.date,

safe\_divide(sum(c.total\_deaths), sum(c.total\_cases)) \* 100 as prob

from da-nfactorial.covid19.cases c

join da-nfactorial.covid19.regions r

on c.iso\_code = r.iso\_code

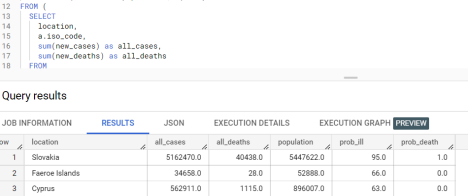
group by r.location, c.date

order by prob desc

**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

r.location,

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

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

d.population,

round(safe\_divide(sum(c.new\_cases), d.population),2) \* 100 as prob\_ill,

round(safe\_divide(sum(c.new\_deaths), d.population),2) \* 100 as prob\_death

from da-nfactorial.covid19.regions as r

join da-nfactorial.covid19.cases as c

on r.iso\_code = c.iso\_code

join da-nfactorial.covid19.demography as d

on c.iso\_code = d.iso\_code

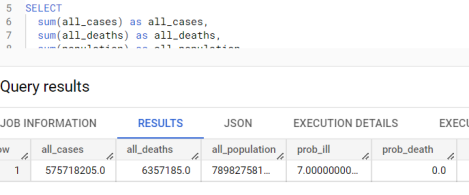
group by r.location, d.population

order by prob\_ill 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 all\_cases,

sum(c.total\_deaths) as all\_deaths,

sum(d.population) as all\_population,

safe\_divide(sum(c.total\_cases), sum(d.population)) \* 100 as prob\_ill,

safe\_divide(sum(c.total\_deaths), sum(d.population)) \* 100 as prob\_death

from da-nfactorial.covid19.cases c

join da-nfactorial.covid19.demography d

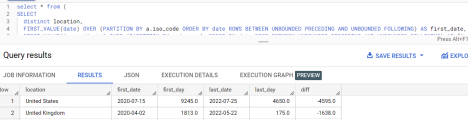
on c.iso\_code = d.iso\_code

and c.date = d.date

**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 first\_last as

(select

r.location,

first\_value(h.date) over (partition by r.location order by h.date asc) as first\_date,

first\_value(h.date) over (partition by r.location order by h.date desc) as last\_date,

coalesce(first\_value(h.icu\_patients) over (partition by r.location order by h.date asc), 0) as first\_day,

coalesce(first\_value(h.icu\_patients) over (partition by r.location order by h.date desc), 0) as last\_day

from da-nfactorial.covid19.regions r

join da-nfactorial.covid19.hospital h

on r.iso\_code = h.iso\_code)

select distinct

location,

first\_date,

last\_date,

first\_day,

last\_day,

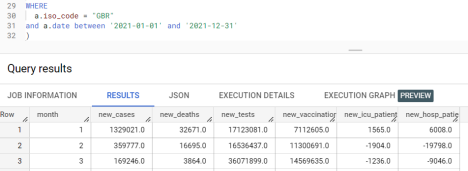
(last\_day - first\_day) as diff

from first\_last

order by diff asc

**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:

select

extract(year from c.date) as year,

extract(month from c.date) as month,

sum(c.new\_cases) as new\_cases,

sum(c.new\_deaths) as new\_deaths,

sum(t.new\_tests) as new\_tests,

sum(v.new\_vaccinations) as new\_vaccinations

from da-nfactorial.covid19.cases as c

join da-nfactorial.covid19.tests as t

on c.iso\_code = t.iso\_code

and c.date = t.date

join da-nfactorial.covid19.vaccinations as v

on t.iso\_code = v.iso\_code

and t.date = v.date

where c.iso\_code = 'GBR' and c.date between '2021-01-01' and '2021-12-31'

group by extract(year from c.date), extract(month from c.date)

order by year, 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 cases\_with\_changes AS (

SELECT

r.location,

c.date,

c.new\_cases,

lag(c.new\_cases, 1, 0) OVER (PARTITION BY r.location ORDER BY c.date) AS lagnew\_cases,

CASE

WHEN lag(c.new\_cases, 1, 0) OVER (PARTITION BY r.location ORDER BY c.date) = 0 THEN c.new\_cases \* 100

ELSE (c.new\_cases - lag(c.new\_cases, 1, 0) OVER (PARTITION BY r.location ORDER BY c.date)) /

lag(c.new\_cases, 1, 0) OVER (PARTITION BY r.location ORDER BY c.date) \* 100

END AS rel\_diff

FROM

da-nfactorial.covid19.regions r

JOIN

da-nfactorial.covid19.cases c

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

)

SELECT

location,

date,

new\_cases,

lagnew\_cases,

rel\_diff,

CASE

WHEN rel\_diff > 0 THEN 'Increase'

WHEN rel\_diff < 0 THEN 'Decrease'

ELSE 'No change'

END AS trend

FROM

cases\_with\_changes

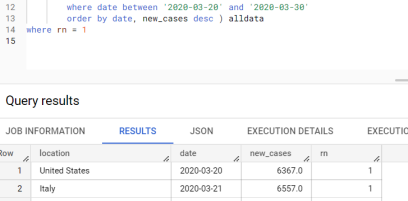
ORDER BY

location, date;

**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:

with ranked\_countries as

(select

r.location,

c.date,

c.new\_cases,

rank() over (partition by c.date order by c.new\_cases desc) as rn

from da-nfactorial.covid19.cases c

join da-nfactorial.covid19.regions r

on c.iso\_code = r.iso\_code

where date between '2020-03-20' and '2020-03-30')

select

location,

date,

new\_cases,

rn

from

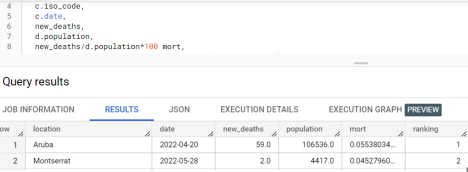
ranked\_countries

where rn = 1

order by date

**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:

select

r.location,

c.date,

c.new\_deaths,

d.population,

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

from da-nfactorial.covid19.regions r

join da-nfactorial.covid19.cases c

on r.iso\_code = c.iso\_code

join da-nfactorial.covid19.demography d

on c.iso\_code = d.iso\_code

order by mort desc

limit 25

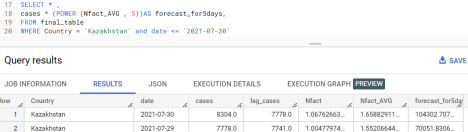
**Вопрос 9: Что ожидать Казахстану следующие 5 дней?**

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 final\_table as

(select

r.location,

c.date,

c.new\_cases,

lag(c.new\_cases, 1, 0) over (partition by r.location order by c.date asc) as lag\_cases,

safe\_divide(c.new\_cases, lag(c.new\_cases, 1, 0) over (partition by r.location order by c.date asc)) as nfact

from da-nfactorial.covid19.regions as r

join da-nfactorial.covid19.cases as c

on r.iso\_code = c.iso\_code

order by c.date asc),

final\_table\_withavg as (

select

location,

date,

new\_cases,

lag\_cases,

nfact,

avg(IF(nfact IS NOT NULL, nfact, 0)) OVER (PARTITION BY location ORDER BY date ASC ROWS BETWEEN 9 PRECEDING AND CURRENT ROW) AS nfact\_avg

from final\_table

)

select

location,

date,

new\_cases,

lag\_cases,

nfact,

nfact\_avg,

new\_cases \* (power (nfact\_avg, 5)) as forecasts\_for5days

from final\_table\_withavg

where location = 'Kazakhstan' and date <= '2021-07-30'

order by date desc

limit 10

**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:**

Ваш запрос и объяснение с графиком:   
Описание: Этот анализ позволяет определить долю населения, полностью вакцинированного против COVID-19, в каждой стране на основании самых актуальных данных. Это помогает выявить лидеров и отстающих в процессе вакцинации, что может быть полезно для оценки эффективности кампаний вакцинации и принятия дальнейших мер.  
  
select

r.location,

d.population,

max(v.people\_fully\_vaccinated) as vaccinated\_people,

safe\_divide(max(v.people\_fully\_vaccinated), d.population) as vaccination\_ratio

from da-nfactorial.covid19.vaccinations v

join da-nfactorial.covid19.regions r

on v.iso\_code = r.iso\_code

join da-nfactorial.covid19.demography d

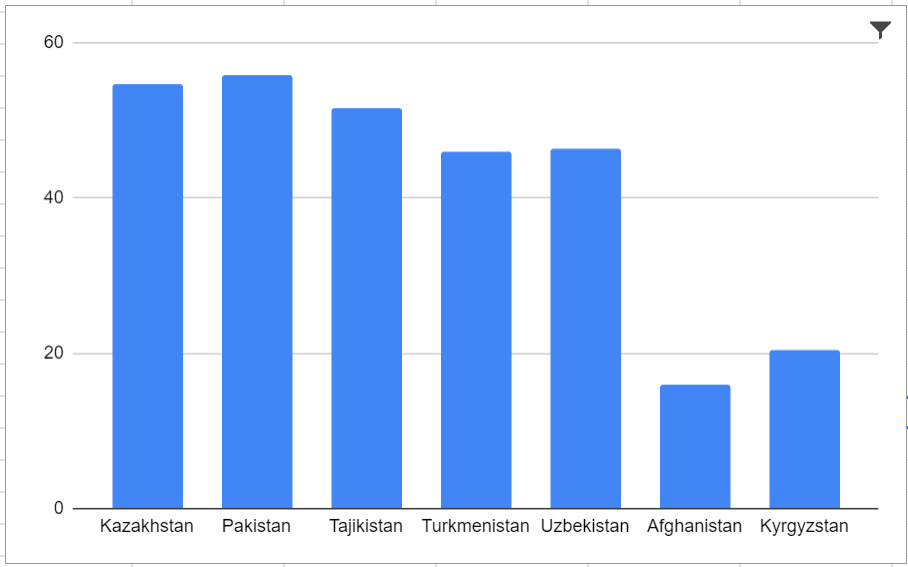
on r.iso\_code = d.iso\_code

group by v.iso\_code, r.location,

d.population

order by vaccination\_ratio desc

**В диаграмме я вывел vaccination\_ratio в странах которые заканчиваются на “stan”**



**Вопрос 2:**

Ваш запрос и объяснение с графиком:

**Описание:** Я сделал запрос, чтобы понять, в каких странах ковид вызвал наибольшие потери. Сравнивая количество смертей, заболевших и продолжительность жизни, я вывел топ стран с самым высоким уровнем смертности. Это помогает мне увидеть, где ситуация была особенно сложной и какие факторы могли повлиять на такие результаты и сравнить насколько они коррелируются с life expectancy

SELECT

r.location,

MAX(c.total\_cases) AS total\_cases,

MAX(c.total\_deaths) AS total\_deaths,

SAFE\_DIVIDE(MAX(c.total\_deaths), MAX(c.total\_cases)) \* 100 AS cfr,

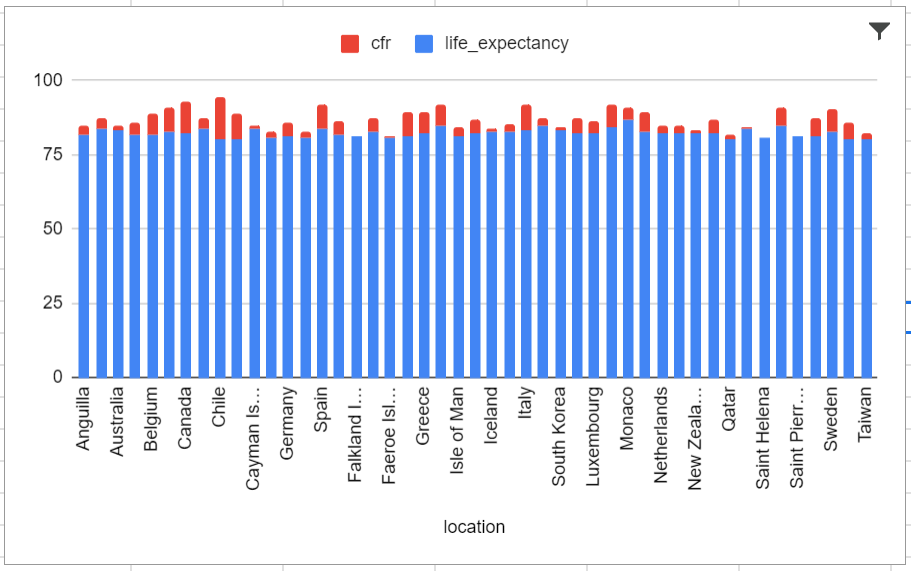
d.life\_expectancy

FROM da-nfactorial.covid19.regions r

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

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

GROUP BY r.location, d.life\_expectancy

order by cfr desc  
  
**В графике я вытаскиваю life\_expectancy and cfr по каждой стране где le больше 80 лет и хочу увидеть насколько они коррелируются  
**

**Вопрос 3:**

Ваш запрос и объяснение с графиком:

**Описание:** Я написал этот запрос, чтобы сравнить количество новых случаев ковида в каждой стране с её средним уровнем за все время. С помощью этого анализа я могу увидеть, в какие дни показатели превышают среднее (или наоборот, ниже), что помогает оценить, насколько ситуация в каждой стране менялась в сравнении с обычным уровнем заболеваемости. Это помогает выявить аномалии и тренды в данных по ковиду.

**with avg\_table as (**

**select**

**r.location,**

**avg(c.new\_cases) as average**

**from da-nfactorial.covid19.cases as c**

**join da-nfactorial.covid19.regions as r**

**on c.iso\_code = r.iso\_code**

**group by r.location**

**)**

**select**

**r.location,**

**c.date,**

**c.new\_cases,**

**average,**

**case**

**when c.new\_cases > average then 'Above'**

**when c.new\_cases < average then 'Below'**

**else 'Average'**

**end as comparison**

**from da-nfactorial.covid19.regions as r**

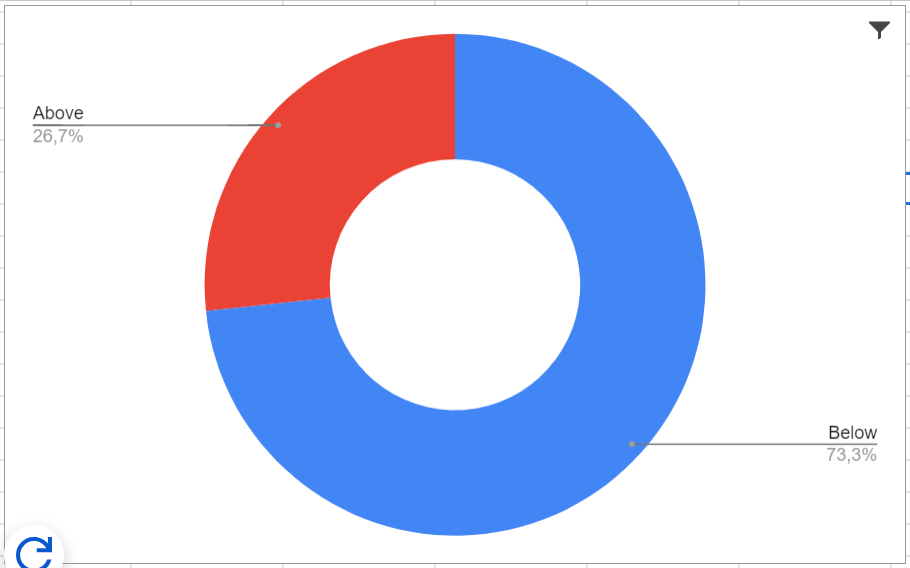
**join da-nfactorial.covid19.cases as c**

**on r.iso\_code = c.iso\_code**

**join avg\_table**

**on r.location = avg\_table.location**

**order by r.location, c.date**

**В графике я посчитал сколько дней было above and below the average в процентном соотношении в Казахстане  
**

**Вопрос 4:**

Ваш запрос и объяснение с графиком:   
  
**Описание:** Я решил собрать данные по количеству случаев COVID-19 и населению по континентам, чтобы сравнить, как эти цифры соотносятся между собой. Сначала я подсчитал максимальное количество случаев по каждой стране, а затем агрегировал данные по континентам, чтобы получить общее количество случаев и населения для каждого континента. Это помогло мне увидеть общую картину и сделать интересные выводы о том, как пандемия повлияла на разные регионы.

with ttl\_table as(

select

r.location,

max(c.total\_cases) as total\_cases

from da-nfactorial.covid19.regions as r

join da-nfactorial.covid19.cases as c

on r.iso\_code = c.iso\_code

group by r.location

)

select

r.continent,

sum(d.population) as population,

sum(total\_cases) as total\_cases

from da-nfactorial.covid19.regions as r

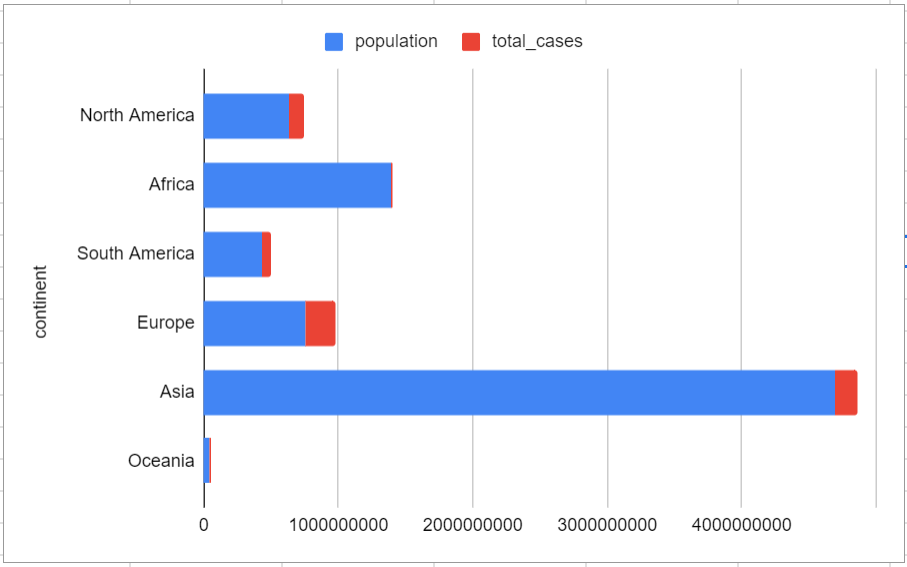
join da-nfactorial.covid19.demography as d

on r.iso\_code = d.iso\_code

join ttl\_table as t

on r.location = t.location

group by r.continent

**В диаграмме показывается соотношение тотал кейсов и население по каждому континенту  
**

Вы закончили свой COVID анализ! Отличная работа!