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	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

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.

Мой запрос:

```
SELECT * FROM `da-nfactorial.covid19.regions`
```

1 – 50 of 233

```
SELECT DISTINCT * FROM `da-nfactorial.covid19.regions`
```

1 – 50 of 230

В таблице regions есть повторяющийся строки.

---

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

Мой запрос:

```
SELECT DISTINCT * FROM `da-nfactorial.covid19.regions`
WHERE LENGTH(iso_code) != 3
```

Row	iso_code	continent	location
1	OWID_KOS	Europe	Kosovo

Есть один, это код страны - Косово

---

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

Мой запрос:

```
SELECT location FROM `da-nfactorial.covid19.regions`
WHERE location LIKE '%Islands%'
```

Row	location
1	Faeroe Islands
2	Solomon Islands
3	Cook Islands
4	Northern Mariana Islands

Results per page: 50 1 – 10 of 10

---

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

Мой запрос:

Я сперва нашел страны в которых есть скобки, потом написал запрос именно для этих стран чтобы проверить написать без скобок.

```
SELECT location FROM `covid19.regions`

where location like '%(%)%'
```

Row	location
1	Micronesia (country)
2	Sint Maarten (Dutch part)

```
SELECT REGEXP_EXTRACT_ALL(location, '^[^(]+') AS location_upd

FROM `da-nfactorial.covid19.regions`

WHERE location LIKE '%(%)%';
```

Row	location_upd
1	Micronesia
2	Sint Maarten

---

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

Мой запрос:

Я бы изменил эти 4 поля на INTEGER.

☐	icu_patients	FLOAT
☐	weekly_icu_admissions	STRING
☐	hosp_patients	STRING
☐	weekly_hosp_admissions	STRING

С помощью запроса:

```
SELECT
  CAST(icu_patients AS INT64) AS icu_patients_upd,
  CAST(weekly_icu_admissions AS INT64) AS weekly_icu_admissions_upd,
  CAST(hosp_patients AS INT64) AS hosp_patients_upd,
  CAST(weekly_hosp_admissions AS INT64) AS weekly_hosp_admissions_upd
FROM `covid19.hospital`
```

---

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

Мой запрос:

```
SELECT
  IFNULL(iso_code,0) as iso_code2,
  IFNULL(date,0) as date2,
  IFNULL(total_cases,0) as total_cases2,
  IFNULL(new_cases,0) as new_cases2,
  IFNULL(total_deaths,0) as total_deaths2,
  IFNULL(new_deaths,0) as new_deaths2,
FROM `covid19.cases`
```

Нет, не стоит менять NULL на 0 так, как типы данных у всех столбцов разные и не являются string, int чтобы можно было везде поставить 0.

Field name	Type
iso_code	STRING
date	DATE
total_cases	FLOAT
new_cases	FLOAT
total_deaths	FLOAT
new_deaths	FLOAT

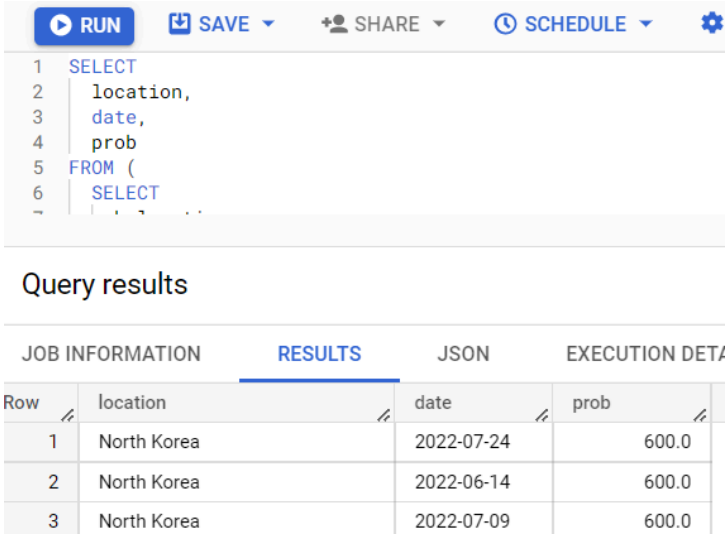
Часть 1

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



Мой запрос:

```
WITH
main_table AS(
  SELECT r.location,
         c.date,
         c.total_deaths/c.total_cases * 100 AS death_prob
  FROM `da-nfactorial.covid19.cases` c
  JOIN `covid19.regions` r
  ON c.iso_code = r.iso_code
  GROUP BY r.location, c.date, c.total_deaths, c.total_cases
)

SELECT * FROM main_table
WHERE death_prob = (SELECT MAX(death_prob) FROM main_table);
```

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

```
FROM (
  SELECT
    location,
    a.iso_code,
    sum(new_cases) as all_cases,
    sum(new_deaths) as all_deaths
  FROM
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

ow	location	all_cases	all_deaths	population	prob_ill	prob_death
1	Slovakia	5162470.0	40438.0	5447622.0	95.0	1.0
2	Faeroe Islands	34658.0	28.0	52888.0	66.0	0.0
3	Cyprus	562911.0	1115.0	896007.0	63.0	0.0

Мой запрос:

```
SELECT r.location,
       SUM(c.total_cases) AS cases_sum,
       SUM(c.total_deaths) AS deaths_sum,
       d.population,
       SUM(c.total_cases)/d.population * 100 as prob_ill,
       SUM(c.total_deaths)/d.population * 100 as prob_dead
FROM `covid19.regions` r
JOIN `covid19.demography` d
ON r.iso_code = d.iso_code
JOIN `covid19.cases` c
ON d.iso_code = c.iso_code and d.date = c.date
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

```
5  SELECT
6    sum(all_cases) as all_cases,
7    sum(all_deaths) as all_deaths,
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECU

ow	all_cases	all_deaths	all_population	prob_ill	prob_death
1	575718205.0	6357185.0	789827581...	7.00000000...	0.0

Ваш запрос:

```
WITH

main_table3 AS(

SELECT c.total_cases AS cases,

       c.total_deaths AS deaths,

       d.population AS population

FROM `covid19.cases` c

JOIN `covid19.demography` d

ON c.iso_code = d.iso_code and c.date = d.date

)


SELECT SUM(cases) AS cases,

       SUM(deaths) AS deaths,

       SUM(population) AS population,

       SUM(cases)/SUM(population) *100 AS prob_ill,

       SUM(deaths)/SUM(population) *100 AS prob_death

FROM main_table3;
```

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

```
select * from (
SELECT
  distinct location,
  FIRST_VALUE(date) OVER (PARTITION BY a.iso_code ORDER BY date ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS first_date,
```

Press Alt+F1

Query results

SAVE RESULTS EXPLO

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

low	location	first_date	first_day	last_date	last_day	diff
1	United States	2020-07-15	9245.0	2022-07-25	4650.0	-4595.0
2	United Kingdom	2020-04-02	1813.0	2022-05-22	175.0	-1638.0

Мой запрос:

```sql
SELECT DISTINCT r.location,
       FIRST_VALUE(h.date) OVER (PARTITION BY h.iso_code ORDER BY h.date
       ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS first_date,
       FIRST_VALUE(h.icu_patients) OVER (PARTITION BY h.iso_code ORDER BY h.date
       ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS first_day,
       LAST_VALUE(h.date) OVER (PARTITION BY h.iso_code ORDER BY h.date
       ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS last_date,
       LAST_VALUE(h.icu_patients) OVER (PARTITION BY h.iso_code ORDER BY h.date
       ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) AS last_dayy,

       LAST_VALUE(h.icu_patients) OVER (PARTITION BY h.iso_code ORDER BY h.date
       ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING) -
       FIRST_VALUE(h.icu_patients) OVER (PARTITION BY h.iso_code ORDER BY h.date
       ROWS BETWEEN UNBOUNDED PRECEDING AND UNBOUNDED FOLLOWING)
       AS diff
FROM `covid19.regions` r
JOIN `covid19.hospital` h
ON r.iso_code = h.iso_code
JOIN `covid19.cases` c
ON h.iso_code = c.iso_code and h.date = c.date
WHERE h.icu_patients = COALESCE(h.icu_patients)
ORDER BY diff;
```

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

```
WHERE
  a.iso_code = "GBR"
and a.date between '2021-01-01' and '2021-12-31'
)
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH PREVIEW

Row	month	new_cases	new_deaths	new_tests	new_vaccinatior	new_icu_patient	new_hosp_patie
1	1	1329021.0	32671.0	17123081.0	7112605.0	1565.0	6008.0
2	2	359777.0	16695.0	16536437.0	11300691.0	-1904.0	-19798.0
3	3	169246.0	3864.0	36071899.0	14569635.0	-1236.0	-9046.0

Мой запрос:

```
WITH
main_table5 AS (
  SELECT c.date AS date_,
         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_vaccs,
         h.icu_patients AS cur_icu_patients,
         LAG(h.icu_patients) OVER (ORDER BY c.date) AS lag_icu_patients,
         h.icu_patients - LAG(h.icu_patients) OVER (ORDER BY c.date) AS new_icu_patients,
         h.hosp_patients AS cur_hosp_patients,
         LAG(h.hosp_patients) OVER (ORDER BY c.date) AS lag_hosp_patients,
         CAST(h.hosp_patients AS float64) - CAST(LAG(h.hosp_patients) OVER (ORDER BY
c.date) AS float64) AS new_hosp_patients
FROM `covid19.cases` c
LEFT JOIN `covid19.tests` t
ON c.iso_code = t.iso_code and c.date = t.date
LEFT JOIN `covid19.vaccinations` v
ON t.iso_code = v.iso_code and t.date = v.date
LEFT JOIN `covid19.hospital` h
```

```
ON v.iso_code = h.iso_code and v.date = h.date
WHERE c.iso_code = 'GBR' AND c.date BETWEEN '2021-01-01' AND '2021-12-31'
GROUP BY c.date, h.icu_patients, h.hosp_patients
ORDER BY c.date
)

SELECT EXTRACT(MONTH FROM date_) AS month,
       SUM(new_cases) AS new_cases,
       SUM(new_deaths) AS new_deaths,
       SUM(new_tests) AS new_tests,
       SUM(new_vaccs) AS new_vaccs,
       SUM(new_icu_patients) AS new_icu_patients,
       SUM(new_hosp_patients) AS new_hosp_patients
FROM main_table5
GROUP BY EXTRACT(MONTH FROM date_)
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

```
select *,
       (case when rel_diff > 0 then "Increase"
             when rel_diff < 0 then "Decrease"
             when rel_diff = 0 then "No change"
             else NULL end) as trend
```

Query results

SAVE

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH | PREVIEW

ow	location	date	new_cases	lagnew_cases	rel_diff	trend
13	Afghanistan	2020-03-07	3.0	0.0	Infinity	Increase
14	Afghanistan	2020-03-08	0.0	3.0	-100.0	Decrease

Мой запрос:

```
WITH

main_table6 AS (

  SELECT r.location,

         c.date,

         c.new_cases,

         LAG(c.new_cases) OVER (PARTITION BY r.location ORDER BY c.date) AS lagnew_cases

FROM `covid19.cases` c

JOIN `covid19.regions` r

ON c.iso_code = r.iso_code

),


secondary_table6 AS (

SELECT *,

       (CASE

          WHEN lagnew_cases IS NULL OR lagnew_cases = 0 THEN 'Infinity'

          ELSE CAST((new_cases-lagnew_cases)/lagnew_cases*100 AS STRING)

          END) AS rel_diff

FROM main_table6

)


SELECT *,

       (CASE

          WHEN rel_diff = 'Infinity' OR CAST(rel_diff AS FLOAT64) > 0 THEN 'Increase'

          WHEN CAST(rel_diff AS FLOAT64) < 0 THEN 'Decrease'
```

```
          WHEN CAST(rel_diff AS FLOAT64) = 0 THEN 'No change'
          ELSE NULL END) AS trend
FROM secondary_table6
WHERE location = 'Afghanistan';
```

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

```
12          where date between '2020-03-20' and '2020-03-30'
13          order by date, new_cases desc ) alldata
14  where rn = 1
15
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTIC

Row	location	date	new_cases	rn
1	United States	2020-03-20	6367.0	1
2	Italy	2020-03-21	6557.0	1

Мой запрос:

```
WITH

main_table7 AS(
  SELECT r.location,
         c.date as date_,
         SUM(c.new_cases) AS new_cases
  FROM `covid19.cases` c
  JOIN `covid19.regions` r
  ON c.iso_code = r.iso_code
  WHERE c.date BETWEEN '2020-03-20' AND '2020-03-30'
  GROUP BY c.date, r.location
  ORDER BY c.date
)


SELECT location,
       date_,
       new_cases,
       DENSE_RANK() OVER (PARTITION BY date_ ORDER BY new_cases) AS rank_
FROM main_table7
WHERE new_cases IN (SELECT FIRST_VALUE(new_cases) OVER (PARTITION BY date_ ORDER BY new_cases DESC) FROM main_table7)
GROUP BY date_, location, new_cases
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

```
4    c.iso_code,
5    c.date,
6    new_deaths,
7    d.population,
8    new_deaths/d.population*100 mort,
```

Мой запрос:

```
SELECT r.location,
       c.date AS date_,
       c.new_deaths,
       d.population,
       (c.new_deaths/d.population) * 100 AS mort,
       DENSE_RANK() OVER (ORDER BY (c.new_deaths/d.population) * 100 DESC) AS rank_
FROM `covid19.cases` c
JOIN `covid19.demography` d
ON c.iso_code = d.iso_code
JOIN `covid19.regions` r
ON d.iso_code = r.iso_code
ORDER BY rank_
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

```
17  SELECT * ,
18  cases * (POWER (Nfact_AVG , 5))AS forecast_for5days,
19  FROM final_table
20  WHERE Country = 'Kazakhstan' and date <= '2021-07-30'
```

Query results SAVE

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH | PREVIEW

Row	Country	date	cases	lag_cases	Nfact	Nfact_AVG	forecast_for5da
1	Kazakhstan	2021-07-30	8304.0	7778.0	1.06762663...	1.65882911...	104302.707...
2	Kazakhstan	2021-07-29	7778.0	7741.0	1.00477974...	1.55206644...	70051.8306...

Мой запрос:

```
WITH
main_table9 AS (
SELECT r.location,
       c.date AS date_,
       c.new_cases AS cases,
       LAG(c.new_cases) OVER (ORDER BY c.date) AS lag_cases,
       (CASE
          WHEN LAG(c.new_cases) OVER (ORDER BY c.date) = 0 THEN NULL
          WHEN LAG(c.new_cases) OVER (ORDER BY c.date) IS NULL THEN NULL
          ELSE c.new_cases/LAG(c.new_cases) OVER (ORDER BY c.date)
          END) AS Nfact
FROM `covid19.cases` c
JOIN `covid19.regions` r
ON c.iso_code = r.iso_code
WHERE c.date <= '2021-07-30' AND r.location = 'Kazakhstan'
ORDER BY c.date DESC
)

SELECT *,
       AVG(Nfact) OVER (ORDER BY date_ DESC ROWS BETWEEN 9 PRECEDING AND CURRENT ROW) AS
Nfact_AVG,
       cases * POWER(AVG(Nfact) OVER (ORDER BY date_ ROWS BETWEEN 9 PRECEDING AND CURRENT
ROW), 5) AS forecast_for5days
FROM main_table9
ORDER BY date_ DESC
```

Часть 2

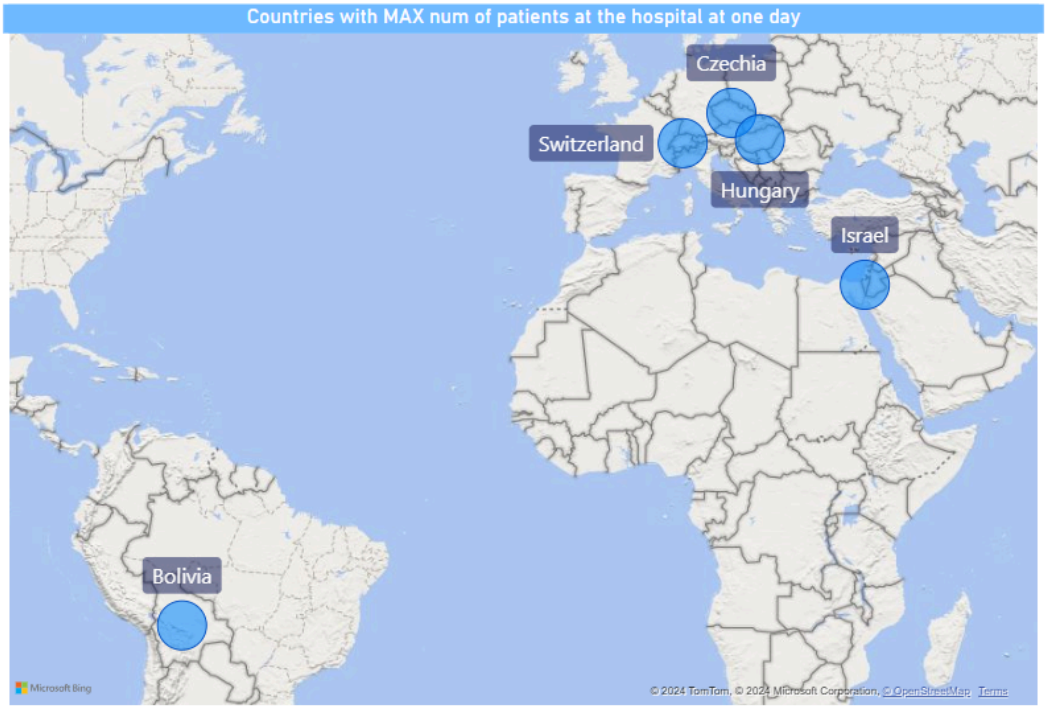
- Определите 4 вопроса о COVID19, на который вы хотите ответить на основе анализа данных.
- Затем напишете SQL-запросы, чтобы получить данные, необходимые для успешного ответа на ваши вопросы.
- Визуализируйте полученные данные (используя гистограммы или другие графики), отвечающие на ваш вопрос.
- Объясните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

Вопрос 1: У какой страны было макс кол-во пациентов в больнице, болевших ковидом в конкретный день?

Мой запрос и объяснение с графиком:

```
SELECT r.location,
       h.date,
       h.hosp_patients
FROM `covid19.regions` r
JOIN `covid19.hospital` h
ON r.iso_code = h.iso_code
WHERE h.hosp_patients = (SELECT MAX(hosp_patients) FROM `covid19.hospital`)
ORDER BY r.location;
```

Row	location	date	hosp_patients
1	Bolivia	2021-07-31	999.0
2	Czechia	2022-04-16	999.0
3	Hungary	2021-10-19	999.0
4	Israel	2020-08-28	999.0
5	Switzerland	2022-07-08	999.0



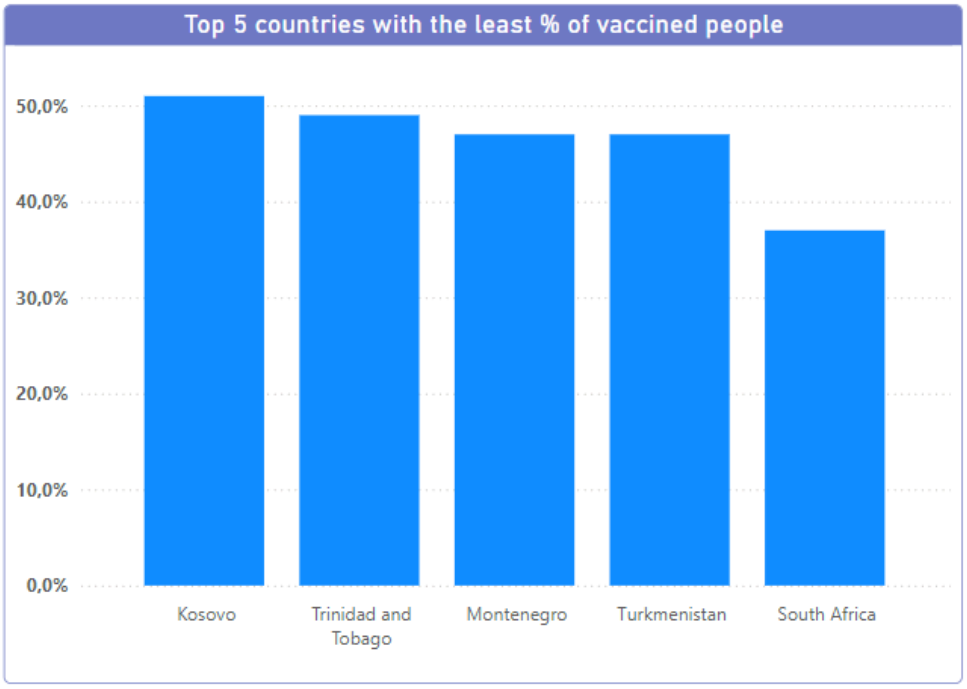
Чтобы найти страну с максимальным количеством пациентов в больнице за один день, я за джойнил таблицы regions and hospital. Через подзапрос сделал проверку чтобы количество пациентов было равно максимальному.

Вопрос 2: **Топ 5 стран в которых наименьшая доля людей которые вакцинированы хотя бы 1 раз**

Мой запрос и объяснение с графиком:

```
SELECT r.location,
       ROUND(SUM(v.people_vaccinated)/d.population * 100) AS vaccinated_part
FROM `covid19.regions` r
JOIN `covid19.vaccinations` v
ON r.iso_code = v.iso_code
JOIN `covid19.demography` d
ON v.iso_code = d.iso_code and v.date = d.date
WHERE v.people_vaccinated = COALESCE(v.people_vaccinated)
GROUP BY r.location, d.population
ORDER BY SUM(v.people_vaccinated)/d.population
LIMIT 5;
```

Row	location	vaccinated_part
1	South Africa	37.0
2	Montenegro	47.0
3	Turkmenistan	47.0
4	Trinidad and Tobago	49.0
5	Kosovo	51.0

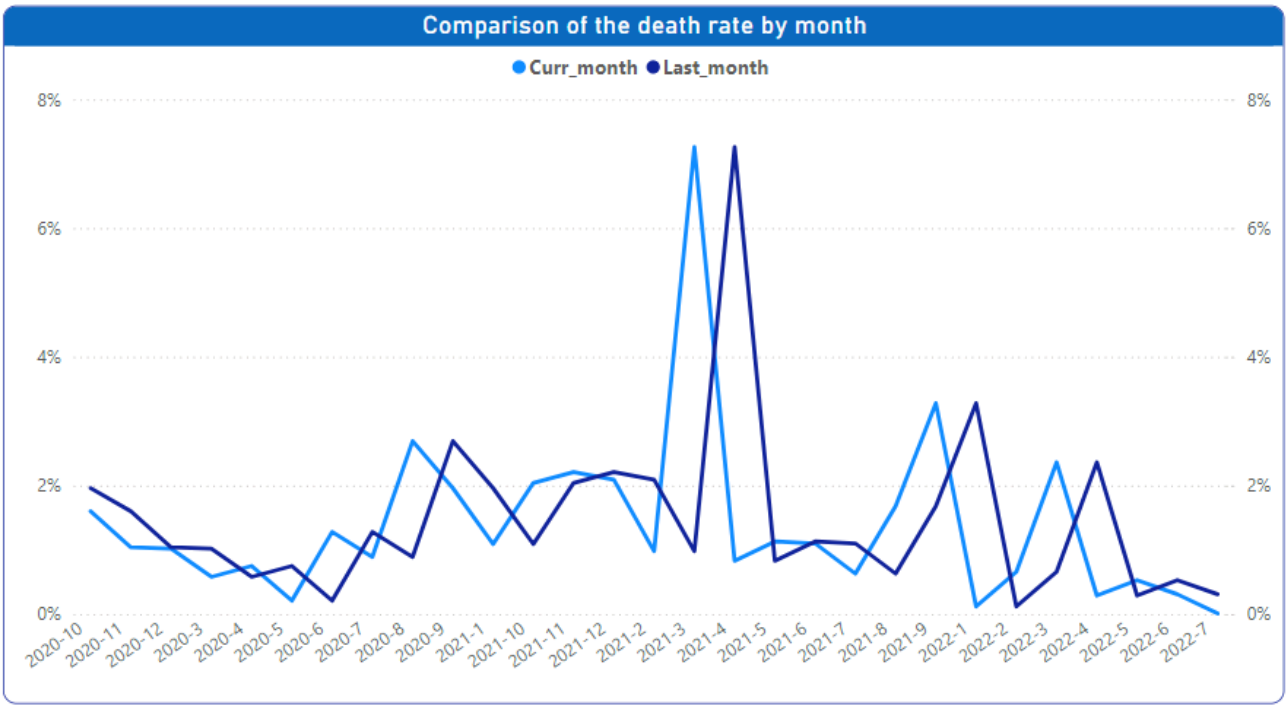


Тут заджойнил 3 таблицы: regions, vaccinations, demography. Долю людей вакцинировавшись хотя бы 1 раз я взял как SUM(v.people_vaccinated)/d.population * 100. Сделал проверку через функцию COALESCE(), чтобы не брать NULL значения. Сгруппировал по странам и отсортировал по доле вакцинировавшихся, и сделал лимит 5 чтобы найти топ 5.

Вопрос 3: **Выведи долю смертности по месяцам и сравни с предыдущими месяцами в Казахстане**

Мой запрос и объяснение с графиком:

```
WITH

main_table33 AS(

SELECT c.date,

       EXTRACT(YEAR FROM c.date) AS year,

       EXTRACT(MONTH FROM c.date) AS month,

       CONCAT(EXTRACT(YEAR FROM c.date), ' - ', EXTRACT(MONTH FROM c.date)) AS year_month,

       c.new_cases,

       c.new_deaths

FROM `covid19.cases` c

JOIN `covid19.regions` r

ON c.iso_code = r.iso_code

WHERE r.location = 'Kazakhstan'

ORDER BY c.date)


SELECT year,

       month,

       SUM(new_cases) AS total_cases,

       SUM(new_deaths) AS total_deaths,

       ROUND(SUM(new_deaths)/SUM(new_cases)*100, 2) AS death_prob,

       ROUND(LAG(SUM(new_deaths)/SUM(new_cases)*100) OVER (ORDER BY year_month), 2) AS
lag_death_prob

FROM main_table33

GROUP BY year_month, year, month

ORDER BY year, month;
```



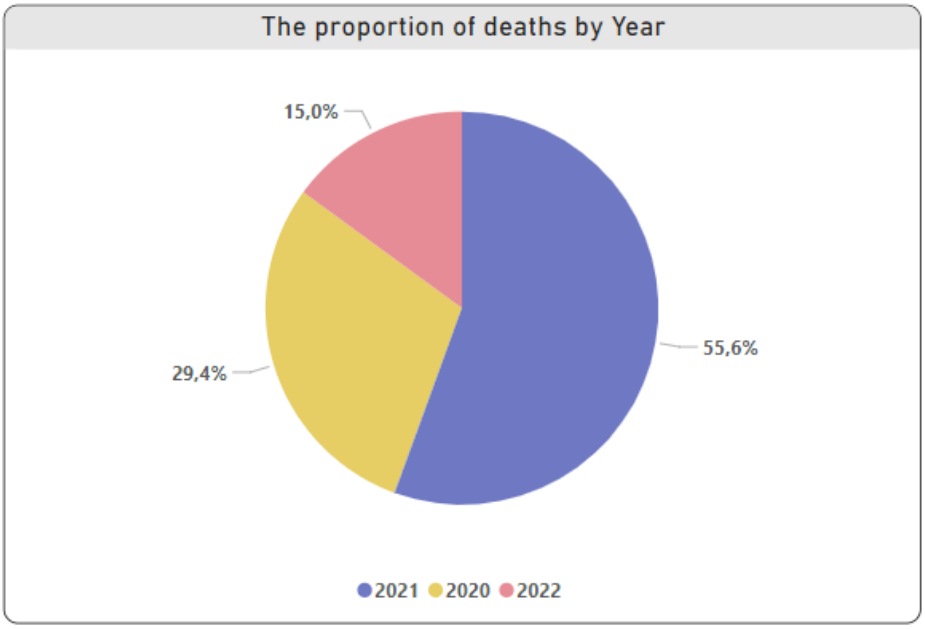
В подзапросе заджойнил regions and cases. Через EXTRACT вытащил год и месяц, создал новую колонку Year_Month через CONCAT и сделал проверку на страну. Потом уже в самом запросе вытащил год и месяц, сумму всех случаев и всех смертей. Потом посчитал долю смертности как SUM(new_deaths)/SUM(new_cases)*100. Потом через функцию LAG вытащил значения за предыдущий месяц.

Вопрос 4: **Выведи долю каждого года по общему кол-ву смертей от ковида во всех странах**

Мой запрос и объяснение с графиком:

```
WITH

main_table44 AS

(

SELECT EXTRACT(YEAR FROM c.date) AS year,

       SUM(c.new_deaths) AS deaths

FROM `covid19.regions` r

JOIN `covid19.cases` c

ON r.iso_code = c.iso_code

GROUP BY EXTRACT(YEAR FROM c.date)

)
```

```sql
SELECT year,
       deaths,
       SUM(deaths) OVER() AS total_deaths,
       ROUND(deaths/SUM(deaths) OVER() *100, 2) AS death_prop
FROM main_table44
ORDER BY year
```



Здесь в подзапросе заджойнил regions and cases. Вытащил год через EXTRACT, вытащил кол-во смертей за каждый год. В запросе посчитал GRAND TOTAL, и разделив на него значения за каждый год нашел долю за каждый год.
