

Objetivo

Este trabalho tem como objetivo ajudar na resposta de perguntas referentes ao impacto que o Covid19 ocasionou no mundo, através de informações como mortes ocasionadas pelo Covid19, número de pessoas que conseguiram se curar e números de novas pessoas infectadas.

Perguntas que desejo responder:

- Qual os top 10 países que tiveram o maior número de novos casos?
- Qual o período com o maior número de novos casos?
- Qual a proporção de novos casos de 19, mortes por Covid19 e recuperação do Covid19 durante um período?
- Qual a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 10 países?
- Qual o comportamento de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 4 países com maiores números de novos casos por em um período de tempo?
- Qual o número de novos casos por continente?
- Qual a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 no Brasil?
- Qual o número de pessoas recuperadas que tomaram a vacina contra Covid19?

Coleta

Para conseguir cumprir o objetivo deste trabalho foi necessário buscar um dataset que tivesse os dados para responder as perguntas, através do site <https://www.kaggle.com/> consegui o dataset que me ajudou com as informações.

Utilizei o dataset

https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=full_grouped.csv.

Depois de obter o dataset fiz a importação dele no <https://community.cloud.databricks.com> através da opção DBFS.

Como usei a opção DBFS:

1. Cliquei a opção catalog
2. Apertei na aba DBFS
3. Apertei em UPLOAD
4. Selecionei o arquivo que estava no meu computador.

Modelagem

Para o modelagem utilizei um CSV extraído do <https://www.kaggle.com/>, fazendo o upload desse CSV no <https://community.cloud.databricks.com>, depois de fazer o upload utilizei Python para criar uma tabela com os dados desse csv no databricks.

As outras tabelas foram populadas e crias através do SQL

Dados armazenados no CSV:

- Lista de Países
- Lista de Continentes
- Data
- Número de mortes
- Número de pessoas recuperadas
- Número de pessoas infectadas
- Número de pessoas ainda infectadas
- Número de novos casos de Covid19
- Número de novas mortes
- Número de novas pessoas recuperadas

Período dos dados : 21/01/2020 até 26/07/2020

Catálogo de dados:

tbGoldCovid19Dataset		
Coluna	Tipo	Descrição
Data	STRING	Data que foi coletado o dado informado
País	STRING	País onde foi coletado o dado
Confirmados	STRING	Número de pessoas infectadas
Mortes	STRING	Número de mortes
Recuperados	STRING	Número de pessoas recuperadas
Ativos	STRING	Número de pessoas ainda infectadas
Novos_casos	STRING	Número de novos casos de Covid19
Novas_mortes	STRING	Número de novas mortes
Novos_recuperados	STRING	Número de novas pessoas recuperadas
Continente	STRING	Continente onde foi coletado o dado

tbSilverCovid19Dataset		
Coluna	Tipo	Descrição
Data	STRING	Data que foi coletado o dado informado
País	STRING	País onde foi coletado o dado
Confirmados	STRING	Número de pessoas infectadas
Mortes	STRING	Número de mortes
Recuperados	STRING	Número de pessoas recuperadas

Ativos	STRING	Número de pessoas ainda infectadas
Novos_casos	STRING	Número de novos casos de Covid19
Novas_mortes	STRING	Número de novas mortes
Novos_recuperados	STRING	Número de novas pessoas recuperadas
Continente	STRING	Continente onde foi coletado o dado

full_grouped_base_dataset_csv		
Coluna	Tipo	Descrição
Date	STRING	Data que foi coletado o dado informado
Country_Region	STRING	País onde foi coletado o dado
Confirmed	STRING	Número de pessoas infectadas
Deaths	STRING	Número de mortes
Recovered	STRING	Número de pessoas recuperadas
Active	STRING	Número de pessoas ainda infectadas
New_cases	STRING	Número de novos casos de Covid19
New_deaths	STRING	Número de novas mortes
New_recovered	STRING	Número de novas pessoas recuperadas
WHO_Region	STRING	Continente onde foi coletado o dado

De Para de mudança de nomenclatura			
Tabela	Coluna com nome antes	Tabela	Coluna com nome depois
full_grouped_base_dataset_csv	Date	tbSilverCovid19Dataset	Data
full_grouped_base_dataset_csv	Country_Region	tbSilverCovid19Dataset	País
full_grouped_base_dataset_csv	Confirmed	tbSilverCovid19Dataset	Confirmados
full_grouped_base_dataset_csv	Deaths	tbSilverCovid19Dataset	Mortes
full_grouped_base_dataset_csv	Recovered	tbSilverCovid19Dataset	Recuperados
full_grouped_base_dataset_csv	Active	tbSilverCovid19Dataset	Ativos
full_grouped_base_dataset_csv	New_cases	tbSilverCovid19Dataset	Novos_casos
full_grouped_base_dataset_csv	New_deaths	tbSilverCovid19Dataset	Novas_mortes

full_grouped_base_dataset_csv	New_recovered	tbSilverCovid19Dataset	Novos_recuperados
full_grouped_base_dataset_csv	WHO_Region	tbSilverCovid19Dataset	Continente

Carga

A primeira carga foi feita através de um CSV extraído do <https://www.kaggle.com/>, fazendo o upload desse CSV no <https://community.cloud.databricks.com>, depois de fazer o upload utilizei Python para criar uma tabela com os dados desse csv no databricks.

```

▶ 03:30 PM (53s) 1

# File location and type
file_location = "/FileStore/full_grouped_base_dataset_csv.csv"
file_type = "csv"

# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
df = spark.read.format(file_type) \
    .option("inferSchema", infer_schema) \
    .option("header", first_row_is_header) \
    .option("sep", delimiter) \
    .load(file_location)

display(df)

temp_table_name = "full_grouped_base_dataset_csv"

df.createOrReplaceTempView(temp_table_name)

df.write.mode("overwrite").saveAsTable("full_grouped_base_dataset_csv")

```

Imagem: Criação da tabela full_grouped_base_dataset_csv com os dados do csv.

A carga da tabela tbSilverCovid19Dataset foi feita através de um "insert select" da tabela full_grouped_base_dataset_csv originada do CSV.

```
▶ 03:30 PM (11s)

%sql
Insert into default.tbSilverCovidDataset
(
Data
,Pais
,Confirmados
,Mortes
,Recuperados
,Ativos
,Novos_casos
,Novas_mortes
,Novos_recuperados
,Continente
)
select
Date
,Country_Region
,Confirmed
,Deaths
,Recovered
,Active
,New_cases
,New_deaths
,New_recovered
,WHO_Region
from default.full_grouped_base_dataset_csv
```

Imagem: Carga da tabela tbSilverCovid19Dataset

A carga da tabela tbGoldCovid19Dataset foi feita através de um “insert select” da tabela tbSilverCovid19Dataset utilizando um filtro para tirar informações que só continham valor 0 nas colunas de número.

```
▶ 03:30 PM (5s) 7

%sql
Insert into default.tbGoldCovidDataset
(
Data
,Pais
,Confirmados
,Mortes
,Recuperados
,Ativos
,Novos_casos
,Novas_mortes
,Novos_recuperados
,Continente
)
select
Data
,Pais
,Confirmados
,Mortes
,Recuperados
,Ativos
,Novos_casos
,Novas_mortes
,Novos_recuperados
,Continente
from default.tbSilverCovidDataset
where
Confirmados > 0
or Mortes > 0
or Recuperados > 0
or Ativos > 0
or Novos_casos > 0
or Novas_mortes > 0
or Novos_recuperados > 0
```

Imagem: Carga da tabela tbGoldCovid19Dataset

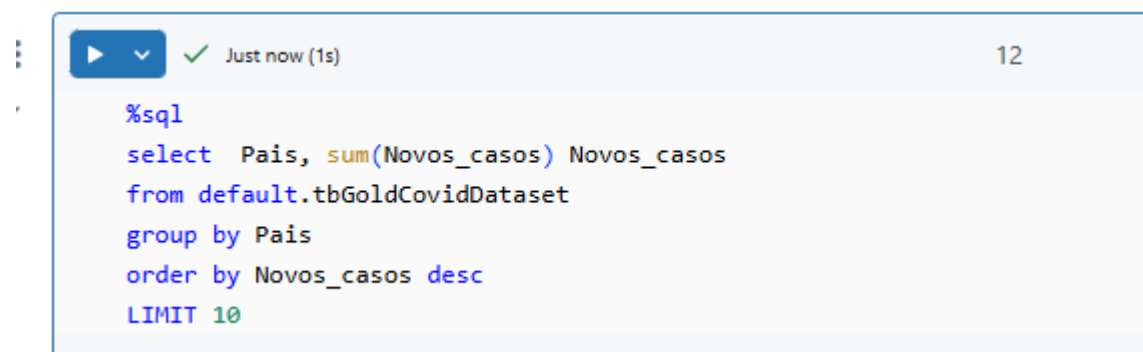
Análise

Após analisar os dados do csv não obtive problema com a qualidade dos dados, as colunas possuíam corretamente suas respectivas informações, por exemplo, coluna de país mostravam países, colunas de números mostravam números.

Depois de toda a modelagem e o carregamento das tabelas iniciei a análise dos dados para responder meu objetivo, respondendo às seguintes perguntas:

Qual os top 10 países que tiveram o maior número de novos casos ?

Resposta:US, Brasil, Índia, Rússia, Peru, Chile, South America, Mexico, United Kingdom e Iran.



```
%sql
select Pais, sum(Novos_casos) Novos_casos
from default.tbGoldCovidDataset
group by Pais
order by Novos_casos desc
LIMIT 10
```

Just now (1s) 12

Imagem: Query que trás os top 10 países com o maior número de novos casos.

Table Visualization 1 +		
	A ^B _C Pais	1.2 Novos_casos
1	US	3576156
2	Brazil	2046328
3	India	1039084
4	Russia	751612
5	Peru	341586
6	Chile	326439
7	South Africa	324221
8	Mexico	324041
9	United Kingdo...	294116
10	Iran	269440

Imagem: Resultado da query que trás os top 10 países com o maior número de novos casos.

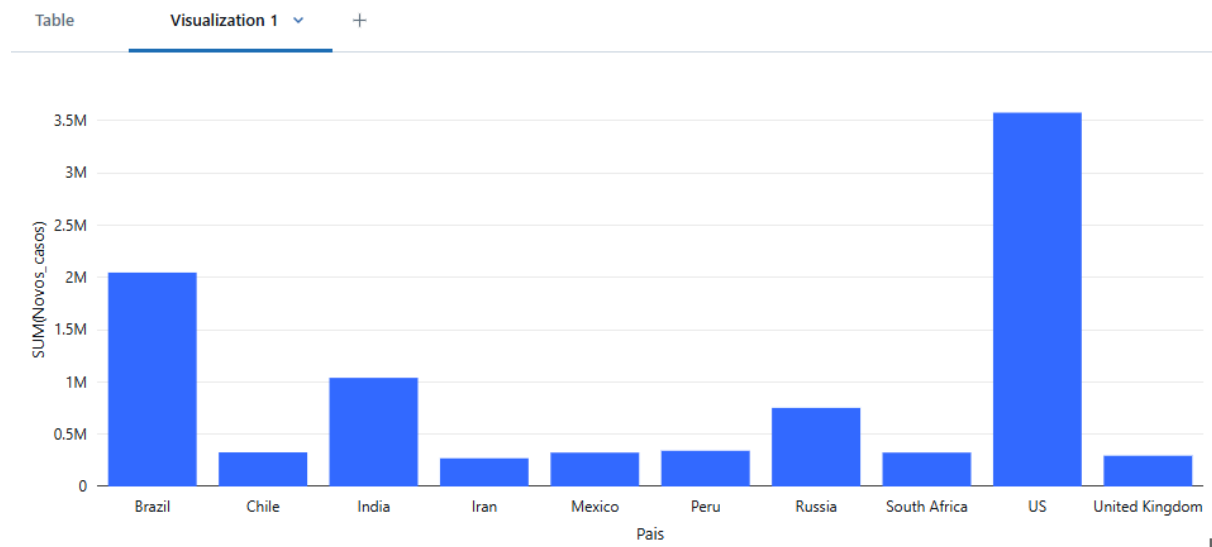


Imagem: Gráfico do resultado da query que trás os top 10 países com o maior número de novos casos.

Qual o período com o maior número de novos casos?

Resposta: o período com maior número de novos casos é em Junho de 2025.

```

%sql
select month(data) Mes, sum(Novos_casos) Novos_casos
from default.tbGoldCovidDataset
group by month(data)
order by month(data) asc

```

Imagem: Query que trás o período com o maior número de novos casos.

	1.2.3 Mes	1.2 Novos_casos
1	1	9372
2	2	75379
3	3	786064
4	4	2412383
5	5	2921042
6	6	4265801
7	7	3473901

Imagem: resultado que trás o período com o maior número de novos casos.

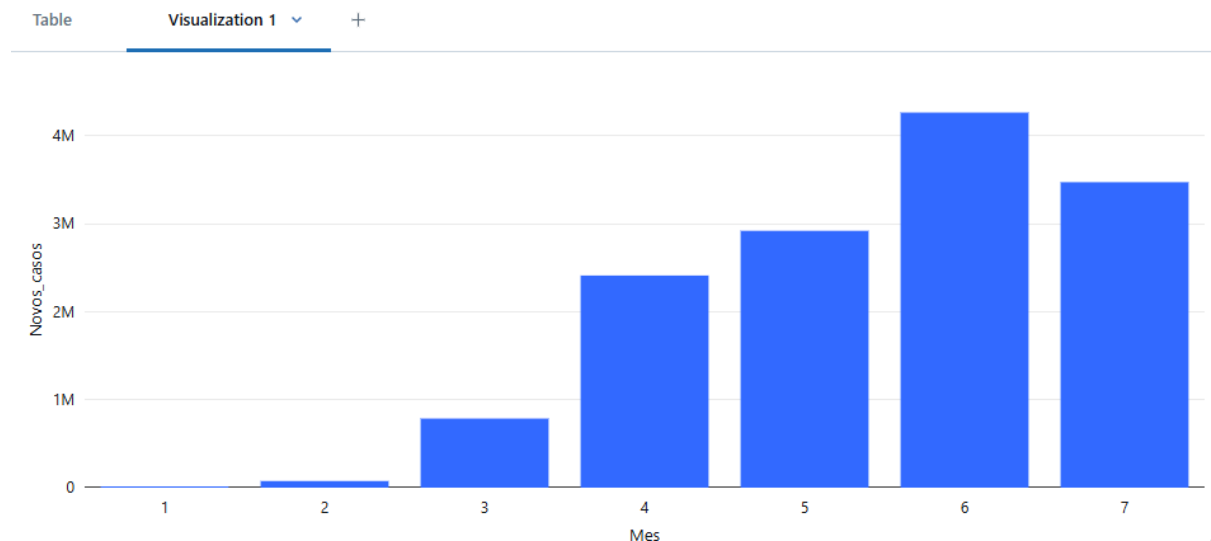


Imagem: Gráfico da query que trás o período com o maior número de novos casos.

Qual a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 durante um período?

Resposta:Depois de analisar podemos ver que conforme o período foi passando o número de pessoas recuperada foi aumentando mais que o número de novos casos.

03:47 PM (1s) 11

```
%sql
select month(data) Mes, sum(Novos_casos) Novos_casos, sum(Novas_mortes) Novas_mortes, sum(Novos_recuperados) Novos_recuperados
from default.tbGoldCovidDataset
group by month(data)
order by month(data) asc
```

Imagem: Query da proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 durante um período.

Table	Visualization 1			
	1.2 Mes	1.2 Novos_casos	1.2 Novas_mortes	1.2 Novos_recuperados
1	1	9372	196	191
2	2	75379	2723	38095
3	3	786064	41542	135760
4	4	2412383	190226	815542
5	5	2921042	138902	1595973
6	6	4265801	137604	2695870
7	7	3473901	82573	2429906

Imagem: Resultado da query da proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 durante um período.

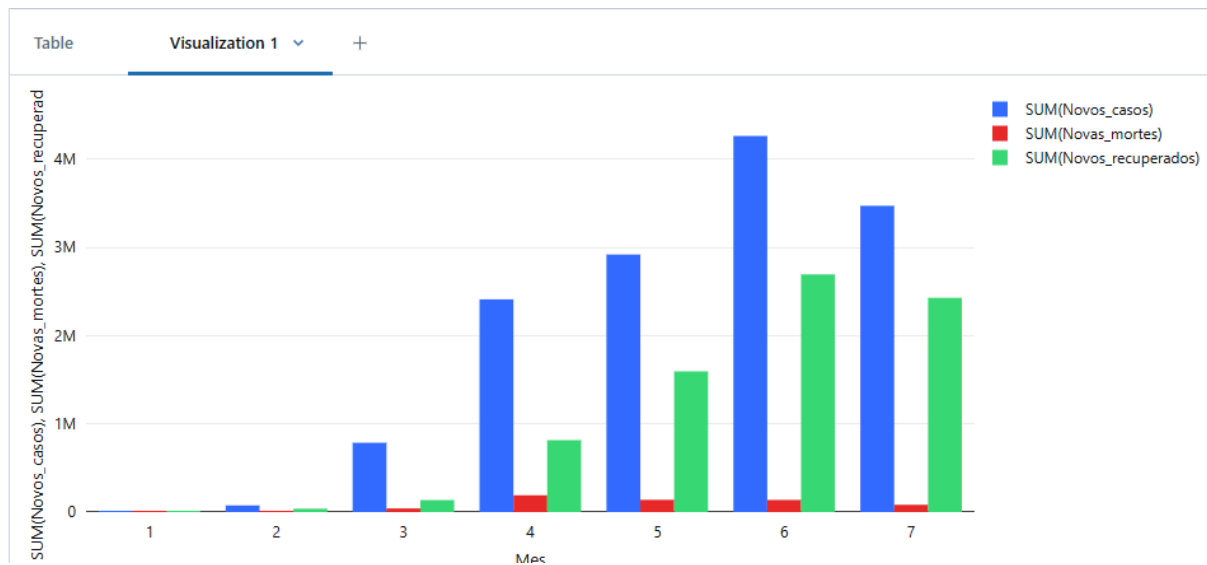


Imagem: Gráfico do resultado da query da proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 durante um período.

Qual a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 10 países?

Resposta: Depois de analisar podemos ver que em alguns países o número de novos casos foi muito maior que o número de pessoas recuperadas, enquanto em outros os valores ficaram um pouco mais próximos.

```

5 minutes ago (1s) 13

%sql
select Pais, sum(Novos_casos) Novos_casos, sum(Novas_mortes) Novas_mortes , sum(Novos_recuperados) Novos_recuperados
from default.tbGoldCovidDataset
group by pais
order by Novos_casos desc
LIMIT 10;

```

Imagem:Query da proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 10 países.

Table Visualization 1 +

	A ^B C Pais	1.2 Novos_casos	1.2 Novas_mortes	1.2 Novos_recuperados
1	US	3576156	138358	1090645
2	Brazil	2046328	77851	1428520
3	India	1039084	26273	653751
4	Russia	751612	11920	530801
5	Peru	341586	12615	230994
6	Chile	326439	8347	296814
7	South Africa	324221	4669	165591
8	Mexico	324041	37574	257681
9	United Kingdo...	294116	45204	1403
10	Iran	269440	13791	232873

Imagem:Resultado da query da proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 10 países.

Table Visualization 1 +

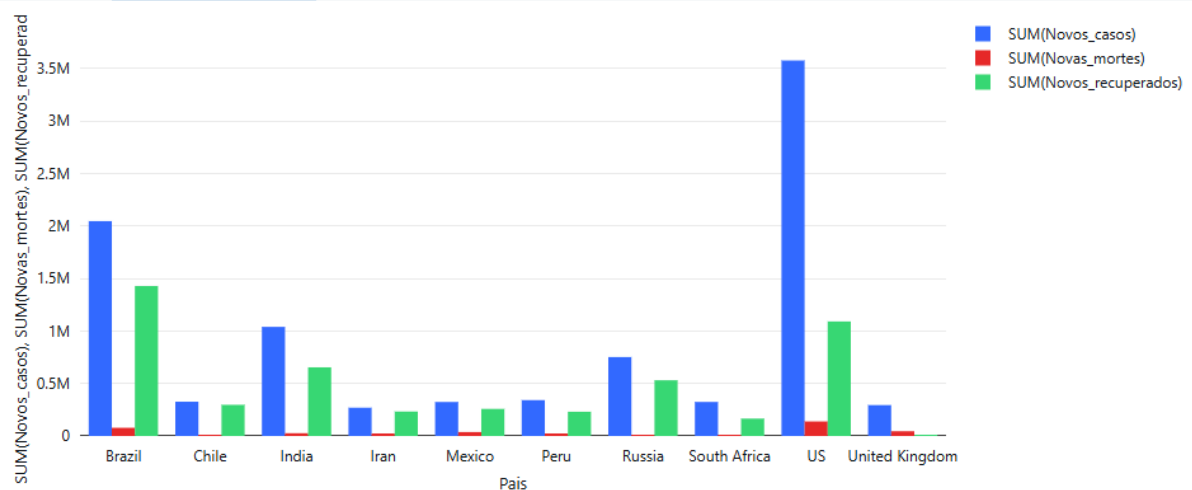


Imagem:Gráfico do resultado da query da proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 10 países.

Qual o comportamento de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 4 países com maiores números de novos casos por em um período de tempo?

Resposta: Depois de analisar podemos ver que os os países menos os US conforme foi passando o período o número de pessoas recuperados foi chegando mais próximo do número de novos casos.

```

%sql
Select month(data),pais,sum(Novos_casos) Novos_casos, sum(Novas_mortes) Novas_mortes , sum(Novos_recuperados)
Novos_recuperados from
default.tbGoldCovidDataset
where Pais in (
    select Pais from
    (
        select Pais, sum(Novos_casos) Novos_casos
        from default.tbGoldCovidDataset
        group by Pais
        order by Novos_casos desc
        LIMIT 4
    )
)
group by month(data),pais
order by month(data) desc

```

Imagem: Query do comportamento de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 4 países com maiores números de novos casos por em um período de tempo.

	1.2 month(data)	1.2 pais	1.2 Novos_casos	1.2 Novas_mortes	1.2 Novos_recuperados
1	7	India	453603	8873	305839
2	7	Russia	104683	2614	118828
3	7	Brazil	644287	18257	640202
4	7	US	939743	10926	370014
5	6	US	837290	22068	275873
6	6	Brazil	887192	30280	581763
7	6	India	394872	11992	256060
8	6	Russia	241086	4613	240090
9	5	Russia	299345	3620	160264
10	5	US	726457	41108	290811
11	5	India	155746	4254	82784
12	5	Brazil	427662	23308	170620
13	4	India	33466	1119	8945
14	4	Russia	104161	1056	11498
15	4	US	883943	58651	146923

Imagem: Resultado da query do comportamento de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 4 países com maiores números de novos casos por em um período de tempo (Parte 1).

	1.3 month(data)	1.2 pais	1.2 Novos_casos	1.2 Novas_mortes	1.2 Novos_recuperados
13	4	India	33466	1119	8945
14	4	Russia	104161	1056	11498
15	4	US	883943	58651	146923
16	4	Brazil	81470	5805	35808
17	3	US	188700	5604	7017
18	3	Brazil	5715	201	127
19	3	India	1394	35	120
20	3	Russia	2335	17	119
21	2	Russia	0	0	2
22	2	US	17	1	7
23	2	India	2	0	3
24	2	Brazil	2	0	0
25	1	Russia	2	0	0
26	1	US	6	0	0
27	1	India	1	0	0

Imagem: Resultado da query do comportamento de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 dos top 4 países com maiores números de novos casos por em um período de tempo (Parte 2).

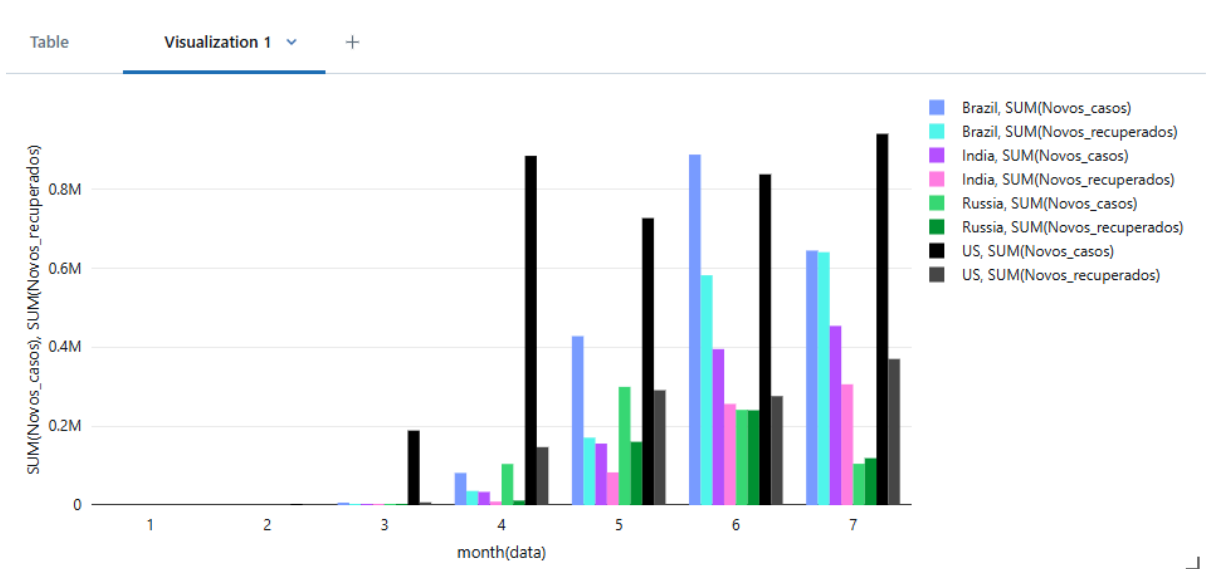


Imagem: Gráfico do resultado da query do comportamento de novos casos de Covid19 e recuperação do Covid19 dos top 4 países com maiores números de novos casos por em um período de tempo.

Qual o número de novos casos por continente?

Resposta: Depois de analisar podemos ver que o continente com o maior número de novos casos foi a América.

```
Just now (1s) 15 SC
%sql
select Continente, sum(Novos_casos) Novos_casos
from default.tbGoldCovidDataset
group by Continente
order by Novos_casos desc
LIMIT 10
```

Imagem: Query do número de novos casos por continente.

Table Visualization 1 +		
	Continente	1.2 Novos_casos
1	Americas	7376844
2	Europe	3073304
3	Eastern Mediterrane...	1349193
4	South-East Asia	1348186
5	Africa	542799
6	Western Pacific	253616

Imagem: Resultado da query do número de novos casos por continente.

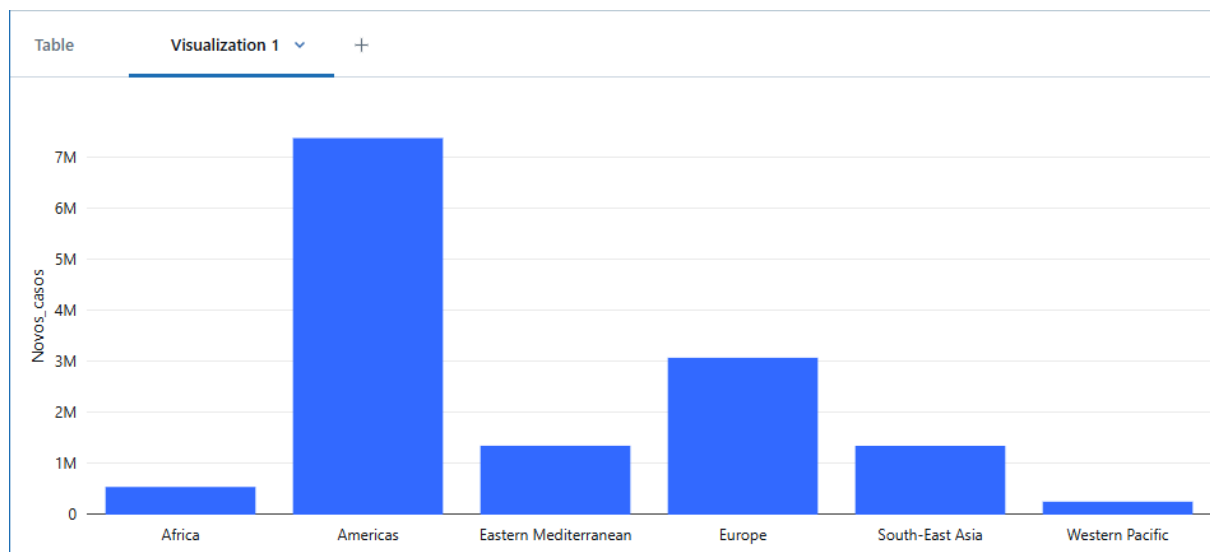


Imagem: Gráfico do resultado da query do número de novos casos por continente.

Qual a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 no Brasil?

Resposta: Depois de analisar podemos ver que no início do período o teve mais mais novos casos do que pessoas recuperadas, no final do período os valores quase se igualaram.

```

%sql
select month(data), sum(Novos_casos) Novos_casos, sum(Novas_mortes) Novas_mortes , sum(Novos_recuperados) Novos_recuperados
from default.tbGoldCovidDataset
where pais = 'Brazil'
group by month(data)
order by month(data)

```

Imagem: Query a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 no Brasil.

	1.2 month(data)	1.2 Novos_casos	1.2 Novas_mortes	1.2 Novos_recuperados
1	2	2	0	0
2	3	5715	201	127
3	4	81470	5805	35808
4	5	427662	23308	170620
5	6	887192	30280	581763
6	7	644287	18257	640202

Imagem: Resultado da query a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 no Brasil.

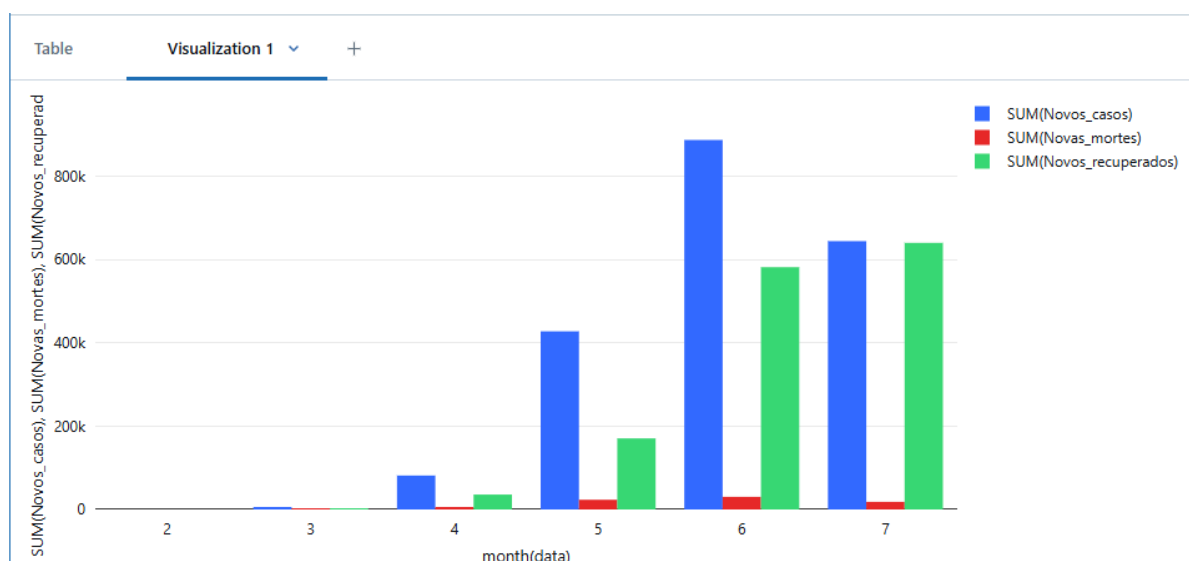


Imagem: Gráfico do resultado da query a proporção de novos casos de Covid19, mortes por Covid19 e recuperação do Covid19 no Brasil.

Autoavaliação

Infelizmente não consegui responder minhas perguntas completamente, meus dados só tinham informações de 21/01/2020 até 26/07/2020, um período muito curto para fazer uma análise completa do Covid19 no mundo, não possuía o primeiro e último mês completos podendo mascarar uma análise dos meses, como por exemplo a pergunta “Qual o período com o maior número de novos casos?” pelo gráfico mostrou que foi junho mas será que se julho estivesse com mês fechado seria junho.

Meu dataset não possuía informações de vacinados e não consegui responder a pergunta “Qual o número de pessoas recuperadas que tomaram a vacina contra Covid19?”