Olá, mundo! Vamos pensar juntos no que vamos fazer?

A tarefa apresentada é: "entender quais são as regiões pelo Brasil, nas granularidades de Cidade, que apresentaram o maior crescimento na taxa de mortalidade (devido a COVID-19) no primeiro trimestre de 2021 e quais as características em comum dessas cidades que justificam essa taxa de mortalidade crescente que se destaca do restante."

O primeiro passo, óbvio, é ler sempre as instruções e entender o que foi solicitado. Tendo feito isso, puxamos os dados recomendados em https://brasil.io/dataset/covid19/files/. Nesse site, encontramos 4 conjuntos de dados .csv:

- Boletim: Apresenta a origem dos dados apresentados nas demais tabelas, usando as colunas:
    - date (data do boletim); notes (observações sobre o mesmo);
    - state (estado do Brasil cujos dados são apresentados);
    - url (link para acessar o boletim).

  É o banco de dados menos estruturado entre todos, pois algumas informações estão com a linha quebrada no .csv e isso poderá ser um problema dependendo de onde você abra o arquivo.
- Caso: apresenta uma versão resumida dos dados de covid no Brasil. As colunas são:
    - date (data dos dados);
    - state (estado dos dados);
    - city (cidade dos dados);
    - place_type (se aquela linha apresenta dados de cidade ou estado);
    - confirmed (número acumulativo de casos de covid confirmados);
    - deaths (número acumulativo de mortes confirmados);
    - order_for_place (número da linha).
    - is_last (se é a última linha referente àquele tipo de dado, logo o último dia de dado coletado);
    - estimated_population_2019 (população estimada em 2019);
    - estimated_population (população estimada atual);
    - city_ibge_code (código da cidade no ibge);
    - confirmed_per_100k_inhabitants (número de casos confirmados a cada 100 mil habitantes);
    - death_rate (taxa de mortes por casos confirmados).
- Caso_full: possui dados completos de covid no Brasil. As colunas são:
    - city (cidade dos dados);
    - city_ibge_code(código no IBGE);
    - date (data em que a linha foi adicionada - não a do dado em si);
    - epidemiological_week (semana epidemiológica);
    - estimated_population (população estimada atualmente);
    - estimated_population_2019 (população estimada em 2019);
    - is_last (se é a última linha referente àquele tipo de dado, logo o último dia de dado incluído no banco);
    - is_repeated (se o dado já foi adicionado antes, de acordo com a data em last_available_date);
    - last_available_confirmed (número total de quantidade de casos confirmados por último);

    - last_available_confirmed_per_100k_inhabitants (número total de casos confirmados por último a cada 100 mil habitantes);
    - last_available_date (data quando a último dado foi extraído dos dados oficiais);
    - last_available_death_rate (última taxa de mortes total disponível);
    - last_available_deaths (último número de mortes disponível);
    - order_for_place (ordem que o dado foi coletado para o local que ele representa);
    - place_type (se o dado é da cidade ou estado);
    - state (estado dos dados);
    - new_confirmed (número de casos novos confirmados desde o último dado novo incluído);
    - new_deaths (número de mortes novas confirmadas desde o último dado novo incluído)
- obitos_cartorio: apresenta dados de óbitos de acordo com os cartórios, especificando o tipo de morte. As colunas são:
    - date (data do dado);
    - state (estado do dado);
    - epidemiological_week_2019 (semana epidemiológica em 2019);
    - epidemiological_week_2020 (semana epidemiológica em 2020);
    - deaths_indeterminate_2019 (númemero de mortes indeterminadas em 2019);
    - deaths_respiratory_failure_2019 (número de mortes por crise respiratória em 2019);
    - deaths_others_2019 (número de mortes por outros motivos em 2019);
    - deaths_pneumonia_2019 (mortes por pneumonia em 2019);
    - deaths_septicemia_2019 (mortes por septicemia em 2019);
    - deaths_sars_2019 (mortes por sars em 2019);
    - deaths_covid19 (mortes por covid19 em 2020);
    - deaths_indeterminate_2020 (mortes indeterminadas em 2020);
    - deaths_respiratory_failure_2020 (mortes por crise respiratória em 2020);
    - deaths_others_2020 (mortes por outros motivos em 2020);
    - deaths_pneumonia_2020 (mortes por pneumonia em 2020);
    - deaths_septicemia_2020 (mortes por septicemia em 2020);
    - deaths_sars_2020 (mortes por SARS em 2020); deaths_total_2019 (número total de mortes em 2019);
    - deaths_total_2020 (número total de mortes em 2020);
    - new_deaths_indeterminate_2019 (número de mortes novas por motivo indetermado em 2019);
    - new_deaths_respiratory_failure_2019 (número de mortes novas por crise respiratória em 2019);
    - new_deaths_others_2019 (número de mortes novas por outros motivos em 2019);
    - new_deaths_pneumonia_2019 (número de mortes novas por pneumonia em 2019);
    - new_deaths_septicemia_2019 (número de mortes novas por septicemia em 2019);
    - new_deaths_sars_2019 (número de mortes novas por SARS em 2019);
    - new_deaths_covid19 (número de mortes novas por covid 19 em 2020);

- new_deaths_indeterminate_2020(número de mortes novas por motivos indetermados em 2020);
- new_deaths_respiratory_failure_2020 (número de mortes novas por crise respiratória em 2020);
- new_deaths_others_2020 (número de mortes novas por outros motivos em 2020);
- new_deaths_pneumonia_2020(número de mortes novas por pneumonia em 2020);
- new_deaths_septicemia_2020(número de mortes novas por septicemia em 2020)
- new_deaths_sars_2020 (número de mortes novas por SARS em 2020);
- new_deaths_total_2019 (número de novas mortes total em 2019)
- new_deaths_total_2020(número de novas mortes total em 2019).

Podemos ver, pelas tabelas, que a tabela Caso_full é a mais completa para nos apresentar os dados que queremos analisar.

Para trabalhar com os dados, escolhemos o PowerBi, devido a nossa familiaridade com a plataforma. Primeiramente, adicionamos os dados csv, promovemos a primeira linha como título das colunas e editamos os dados da forma necessária:

```
=
Csv.Document(File.Contents("C:\Users\a_bel\Dropbox\freelas\Demo\boletim.csv\caso_full.csv"),[Delimiter=",", Columns=18, Encoding=65001, QuoteStyle=QuoteStyle.None])

= Table.PromoteHeaders(Fonte, [PromoteAllScalars=true])

= Table.TransformColumnTypes(#"Cabeçalhos Promovidos",{{"date", type date}, {"city_ibge_code", Int64.Type}, {"epidemiological_week", Int64.Type}, {"estimated_population", Int64.Type}, {"estimated_population_2019", Int64.Type}, {"is_last", type logical}, {"is_repeated", type logical}, {"last_available_confirmed", Int64.Type}})

= Table.TransformColumnTypes(#"Tipo Alterado", {{"last_available_confirmed_per_100k_inhabitants", type number}}, "en-US")

= Table.TransformColumnTypes(#"Tipo Alterado", {{"last_available_confirmed_per_100k_inhabitants", type number}}, "en-US")

= Table.TransformColumnTypes(#"Tipo Alterado com Localidade",{{"last_available_date", type date}})

= Table.TransformColumnTypes(#"Tipo Alterado1", {{"last_available_death_rate", type number}}, "en-US")

= Table.TransformColumnTypes(#"Tipo Alterado com Localidade1",{{"last_available_deaths", Int64.Type}, {"order_for_place", Int64.Type}, {"new_confirmed", Int64.Type}, {"new_deaths", Int64.Type}})
```

Para trabalhar com os dados, como a granularidade solicitada era de “cidade”, fizemos um filtro na tabela em que:

- Excluímos os dados repetidos (aqueles dados em que o is_repeated = TRUE).
- Excluímos dados consolidados dos estados (place_type = state)
- Excluímos dados sem informação de código da cidade (esses dados poderão ser tanto dados de estado quanto dados em que a cidade não foi adequadamente informada, logo apenas representaria ruído em nossas informações).

A fórmula desse filtro no PowerQuery ficou:

= Table.SelectRows(#"Tipo Alterado2", each ([is_repeated] = false) and ([city] <> "") and ([city_ibge_code] <> null)).

Com tais alterações feitas, subimos os dados para o PowerBi. Para chegarmos à informação de interesse, deveríamos considerar que não estamos buscando o dado para o período todo da pandemia, mas para uma faixa específica de tempo. O grande ponto é que os dados apresentados que precisaríamos usar (a taxa de mortes por casos confirmados ou por número de infectados a cada 100 mil habitantes) não leva em consideração o número de pessoas infectadas num determinado período, apenas o número **total** (acumulado) de mortes ou infectados no dia de inclusão de dados. Assim, o recorte de um determinado momento pode não mostrar os dados adequados, já que o número de infectados a cada 100 mil habitantes vai sempre aumentar ao longo do tempo, superestimando o número de mortes num local num período de queda da taxa, por exemplo, e a taxa de mortes subestimada por não oscilar de forma tão pareada com o momento apresentado. Por esse motivo, optamos por usar os dados das taxas considerando o número de Novos casos confirmados e Novas mortes registradas. Para a taxa de mortes, aplicamos a fórmula:

```
Taxa real de mortes entre infectados = IFERROR(sum('caso_full
(2)'[new_deaths])/sum('caso_full (2)'[new_confirmed]),0)
```

Já para a taxa de novos confirmados por 100 mil habitantes, usamos a fórmula:

```
Taxa real confirmados pela população = sum('caso_full
(2)'[new_confirmed])/AVERAGE('caso_full (2)'[estimated_population])
```

Vale lembrar que, nesse segundo caso, o número deixa de ser em relação a cada 100 mil habitantes, mas se torna a proporção de pessoas da cidade infectadas. Para podermos relativizar a questão do tamanho da população dentro da análise, criamos um filtro que leva em conta o tamanho da população. Isso é importante pois, em cidades muito pequenas tendem a mostrar um dado mais super estimado, que não necessariamente são relevantes quando se olha a informação de uma forma mais macro. Também criamos uma variável com o número de mortes dentro da população da cidade:

```
Taxa real de mortes pela população = IFERROR(sum('caso_full
(2)'[new_deaths])/sum('caso_full (2)'[estimated_population]),0)
```

Para executar as análises, é importante considerar que o Brasil possui um número enorme de cidades. Analisar todas junto seria insano. Assim, criamos filtros (estado e tamanho da cidade) que nos permitem olhar de forma mais específica para cada região.

Considerando de forma macro, porém, podemos afirmar que as cidades cuja situação era mais grave no primeiro trimestre de 2021 são Putinga (RS), Santa Bárbara do Sul (RS), Esperança Nova (PR), São Pedro das Missões (RS), Marema (SC). As 5 cidades são cidades pequenas (menos de 10.000 habitantes), da região sul do país. Essas taxas não são necessariamente altas (a média de novas mortes é baixa nesse período), mas, como essas são cidades muito pequenas, isso tem um impacto alto nas taxas apresentadas.

Visualização dos dados:

Para visualizar essa e outras análises possíveis, basta instalar o PowerBi Desktop na máquina em que se irá visualizar o dado e depois abrir o arquivo pbix que se encontra no link https://www.dropbox.com/s/dz9x8fqp0bnpq9v/Dado.pbix?dl=0 . Caso não se deseje instalar o programa, é possível subir os dados pelo site www.powerbi.com mas, nesse caso, é preciso ter uma conta registrada.

Links de interesse:

Visualização do dado:

https://www.dropbox.com/s/dz9x8fqp0bnpq9v/Dado.pbix?dl=0

Base de dados:

https://www.dropbox.com/s/2ae5f2u359p9dg5/caso_full.zip?dl=0