Relatório – Case Técnico de Dados Nexodata

Demanda: O time de epidemiologia da Nexodata fez o envio de datasets e quer saber quais são as regiões pelo Brasil, na granularidade de cidade, que apresentam o maior crescimento na taxa de mortalidade por COVID-19 no primeiro trimestre de 2021. E quer saber quais as características em comum dessas cidades que justifiquem tal variação da taxa de mortalidade.

Compreendido a demanda, as etapas para realizar a análise requisitada serão divididas da seguinte forma:

- 1. Entendimento dos datasets disponibilizados e suas características

- 2. Tratamento dos dados

- 3. Análise exploratória

- 4. Interpretação dos resultados

- 5. Repositório no GITHUB

1 Entendimento dos datasets disponibilizados e suas características

Será utilizado o editor de código Jupyter Notebook (linguagem Python) para a análise inicial dos datasets e as bibliotecas os, pandas e gzip. O script 'Tratamento.jpynb' deve estar na mesma pasta onde os datasets se encontram para seu devido funcionamento.

Os datasets podem ser acessados através do link https://brasil.io/dataset/covid19/files/, onde nele contém 4 arquivos '.csv', que se encontram compactados no formato '.gz':

- 'boletim': dataset com o histórico de origem das informações coletadas para a elaboração dos outros datasets, cujas colunas são:
    - date: data de publicação do boletim
    - notes: comentários sobre a publicação
    - state: estado do Brasil ao qual a publicação se refere
    - url: link de acesso do boletim

Para fins de análise, este dataset não possuí nenhuma aplicação prática e objetiva para responder a demanda requisitada, portanto ele será desconsiderado da análise.

- 'caso': dataset com informações resumidas de casos de COVID-19 no Brasil, cujas colunas são:
    - date: data de coleta dos dados
    - state: estado do Brasil dos dados
    - city: cidade do Brasil dos dados
    - place_type: o tipo do local, se é uma cidade ou estado
    - confirmed: quantidade acumulada de casos confirmados de COVID-19
    - deaths: quantidade acumulada de mortes por COVID-19
    - order_for_place: numeração da linha
    - is_last: informação booleana que distingue a última informação coletada da data dos dados

    - estimated_population_2019: população estimada do ano de 2019
    - estimated_population: população estimada do ano corrente
    - city_ibge_code: código IBGE da cidade
    - confirmed_per_100k_inhabitants: número de casos confirmados por 100 mil habitantes
    - death_rate: taxa de mortalidade

- 'caso_full': dataset com informações completas de casos de COVID-19 no Brasil, cujas colunas são:
    - city: cidade do Brasil dos dados
    - city_ibge_code: código IBGE da cidade
    - date: data de coleta dos dados
    - epidemiological_week: semana epidemiológica da coleta dos dados
    - estimated_population: população estimada do ano corrente
    - estimated_population_2019: população estimada do ano de 2019
    - is_last: informação booleana que distingue a última informação coletada da data dos dados
    - is_repeated: informação booleana que define se o dado está sendo repetido
    - last_available_confirmed: quantidade acumulada de últimos casos confirmados de COVID-19
    - last_available_confirmed_per_100k_inhabitants: quantidade acumulada de últimos casos confirmados por 100 mil habitantes
    - last_available_date: última data da coleta dos dados
    - last_available_death_rate: última taxa de mortalidade da coleta dos dados
    - last_available_deaths: última quantidade acumulada de mortes da coleta dos dados
    - order_for_place: numeração da linha
    - place_type: o tipo do local, se é uma cidade ou estado
    - state: estado do Brasil dos dados
    - new_confirmed: número de novos casos confirmados de COVID-19
    - new_deaths: número de novas mortes por COVID-19

Devido ao fato do dataset 'caso_full' ter dados mais completos, contendo o período de 2021 requisitado na demanda, ele será utilizado para a análise definitiva. Sendo assim, o dataset 'caso' será desconsiderado.

- 'obito_cartorio': dataset com informações, de acordo com os cartórios, do tipo e volume das causas de óbito por COVID-19, cujas colunas são:
    - date: data de coleta dos dados
    - state: estado do Brasil dos dados
    - epidemiological_week_2019: semana epidemiológica da coleta dos dados de 2019
    - epidemiological_week_2020: semana epidemiológica da coleta dos dados de 2020
    - deaths_indeterminate_2019: número acumulado de mortes indeterminadas em 2019
    - deaths_respiratory_failure_2019: número acumulado de mortes por falha respiratória em 2019

- deaths_others_2019: número acumulado de mortes por outros motivos em 2019
- deaths_pneumonia_2019: número acumulado de mortes por pneumonia em 2019
- deaths_septicemia_2019: número acumulado de mortes por sepse em 2019
- deaths_sars_2019: número acumulado de mortes por SARS-CoV-2 em 2019
- deaths_covid19: número acumulado de mortes por COVID-19
- deaths_indeterminate_2020: número acumulado de mortes indeterminadas em 2020
- deaths_respiratory_failure_2020: número acumulado de mortes por falha respiratória em 2020
- deaths_others_2020: número acumulado de mortes por outros motivos em 2020
- deaths_pneumonia_2020: número acumulado de mortes por pneumonia em 2020
- deaths_septicemia_2020: número acumulado de mortes por sepse em 2020
- deaths_sars_2020: número acumulado de mortes por SARS-CoV-2 em 2020
- deaths_total_2019: número acumulado de mortes total em 2019
- deaths_total_2020: número acumulado de mortes total em 2020
- new_deaths_indeterminate_2019: número de novas mortes indeterminadas em 2019
- new_deaths_respiratory_failure_2019: número de novas mortes por falha respiratória em 2019
- new_deaths_others_2019: número de novas mortes por outros motivos em 2019
- new_deaths_pneumonia_2019: número de novas mortes por pneumonia em 2019
- new_deaths_septicemia_2019: número de novas mortes por sepse em 2019
- new_deaths_sars_2019: número de novas mortes por SARS-CoV-2 em 2019
- new_deaths_covid19: número de novas mortes por COVID-19
- new_deaths_indeterminate_2020: número de novas mortes indeterminadas em 2020
- new_deaths_respiratory_failure_2020: número de novas mortes por falha respiratória em 2020
- new_deaths_others_2020: número de novas mortes por outros motivos em 2020
- new_deaths_pneumonia_2020: número de novas mortes por pneumonia em 2020
- new_deaths_septicemia_2020: número de novas mortes por sepse em 2020
- new_deaths_sars_2020: número de novas mortes por SARS-CoV-2 em 2020
- new_deaths_total_2019: número de novas mortes totais em 2019
- new_deaths_total_2020: número de novas mortes totais em 2020

Como o dataset ‘obito_cartorio’ possui somente dados de datas entre 2019 e 2020, ele não poderá ser utilizado para tirar conclusões definitivas para o período de 2021, porém ainda pode ser usado para determinar somente tendências do tipo de causa de morte em regiões de cada estado.

2 Tratamento dos dados

Continuando com o uso do editor de código Jupyter Notebook, o tratamento de dados foi somente realizado no dataset ‘caso_full’.

De início, foi criado uma variável df_caso_full que recebe o dataframe descompactado com o uso da biblioteca pandas. Em seguida, foram retiradas as linhas onde na coluna ‘is_repeated’ continham a informação booleana ‘True’, que mostram que os dados são repetidos, portanto não há proposito de mantê-los na tabela. A linha de código utilizada foi:

```
df_caso_full = df_caso_full.drop(df_caso_full[df_caso_full['is_repeated'] == True].index)
```

E logo após a execução da linha de código acima, a coluna foi excluída por não ter informações analíticas pertinentes. A linha de código usada foi:

```
df_caso_full = df_caso_full.drop('is_repeated', axis=1)
```

Como a granularidade deve ser de cidade, a coluna ‘city’ não pode conter nenhum campo vazio. Para realizar a exclusão de todas as linhas sem o campo preenchido, a linha de código usada foi:

```
df_caso_full = df_caso_full.dropna(subset=['city'])
```

Feito o entendimento anterior dos datasets, algumas colunas acabam não sendo pertinentes à análise, como as colunas:

- ‘city_ibge_code’ (informação sem importância analítica)
- ‘is_last’ (não há necessidade de saber qual a última data de coleta dos dados)
- ‘last_available_date’ (é exatamente igual a coluna ‘date’)
- ‘place_type’ (como a análise será somente de cidades, esta coluna não tem mais propósito)
- ‘estimated_population_2019’ (estimativas de anos diferentes de 2021 não são pertinentes à análise)

Portanto estas colunas foram excluídas, e a linha de código utilizada foi:

```
colunas_drop = ['city_ibge_code', 'is_last', 'last_available_date', 'place_type',
'estimated_population_2019']

for coluna in colunas_drop:
    df_caso_full = df_caso_full.drop(coluna, axis=1)
```

Com o tratamento dos dados contidos na variável ‘df_caso_full’, será feito a extração compactada do dataframe com a extensão ‘.csv.gz’, usando a linha de código:

```
df_caso_full.to_csv('novo_caso_full.csv.gz', sep=',', index=False, compression='gzip')
```

*Nesta etapa é importante se atentar ao separador ‘sep=’,’’, pois, dependendo do idioma das plataformas de visualização (português ou inglês), o separador pode ser confundido com o separador de decimais (‘,’ em português e ‘.’ em inglês). Recomenda-se trocar o separador para ‘sep=’;’’, caso não saiba qual o idioma da plataforma de visualização.

Por fim, agora o dataset tratado e salvo como 'novo_caso_full.csv.gz' está pronto para a análise exploratória.

3 Análise exploratória

A ferramenta escolhida para visualização foi o Power BI, por ser a ferramenta que possuo maior conhecimento e por ter dashboards visuais limpos.

Para acrescentar o arquivo 'novo_caso_full.csv.gz', sem precisar descompactar e manter o arquivo 'leve', deve-se acrescentar uma nova *Blank Query*, depois abrir o Advanced Editor e colar o seguinte código:

```
let
   #"novo_caso_full csv" = let
   Source = Binary.Decompress(
    File.Contents(
     "C:\Users\[local completo de onde encontra-se o arquivo 'novo_caso_full.csv.gz']"),
      Compression.GZip),
   #"Imported CSV" = Csv.Document(Source,
    [Delimiter=",",Encoding=1252])
in
   Source
```

Depois é só dar 'Ok'/'Done' e a tabela já poderá ser visualizada.

Em seguida foi criado uma tabela calendário auxiliar, com o propósito de cobrir todo o intervalo de datas existente no dataset. Para tal foi criado um outro *Blank Query* e criado as seguintes variáveis:

```
= List.Min(novo_caso_full[date])                #com nome DataMin
= List.Max(novo_caso_full[date])                #com nome DataMax
= Duration.Days(DataMax-DataMin)+1              #com nome QtdDias
= List.Dates(DataMin, QtdDias, #duration(1,0,0,0))
```

Depois basta clicar em 'Para a tabela', no canto esquerdo superior e a tabela estará pronta para ser correlacionada no Model View.

Como somente o primeiro trimestre de 2021 será analisado, para reduzir o tamanho do arquivo e deixar as manipulações mais rápidas, o período entre 01/01/2021 e 31/03/2021 foi filtrado e o restante foi deletado.

Devido ao fato de existirem muitas cidades no Brasil, seria praticamente inviável e improdutivo a análise particular de cada uma. E há o fato de existirem cidades muito pequenas e que podem acabar superestimando métricas baseadas na população de cada cidade, portanto foi determinado o uso das seguintes métricas/medidas:

- Total Mortes: total de mortes por COVID-19

```
Total Mortes = SUM(novo_caso_full[new_deaths])
```

- Total Casos Confirmados: total de casos confirmados de COVID-19

```
Total Casos Confirmados = SUMX(novo_caso_full, novo_caso_full[new_confirmed])
```

- Taxa de Mortalidade: taxa de mortalidade real do COVID-19

```
Taxa de Mortalidade =
IFERROR(SUM(novo_caso_full[new_deaths])/SUM(novo_caso_full[new_confirmed]), 0)
```

- População Estimada por Cidade: população estimada não cumulativa, que pode ser filtrada por cidade

```
População Estimada por Cidade =
SUMX(VALUES(novo_caso_full[city]),AVERAGE(novo_caso_full[estimated_population]))
```

- Taxa Mortes por População: taxa de mortes real pela população, que pode ser filtrada por cidade

```
Taxa Mortes por População = IFERROR(sum(novo_caso_full[new_deaths])/[População
Estimada por Cidade],0)
```

- Taxa Confirmados por População: taxa de casos confirmados real de COVID-19, que pode ser filtrada por cidade

```
Taxa Confirmados por População =
IFERROR(SUM(novo_caso_full[new_confirmed])/[População Estimada por Cidade],0)
```

- Taxa de Mortalidade Dia Anterior: para cálculo posterior da variação da taxa de mortalidade

```
Taxa de Mortalidade Dia Anterior = IFERROR(CALCULATE([Taxa de
Mortalidade],DATEADD('Calendário'[Data],-1,DAY)), 0)
```

- Delta Taxa de Mortalidade: diferença entre a taxa de mortalidade do dia seguinte com o anterior

```
Delta Taxa de Mortalidade = [Taxa de Mortalidade]-[Taxa de Mortalidade Dia
Anterior]
```

- Variação Taxa de Mortalidade: calcula a variação percentual da taxa de mortalidade

```
Variação Taxa de Mortalidade = IFERROR([Delta Taxa de Mortalidade]/[Taxa de
Mortalidade Dia Anterior],0)
```

- População Estimada Absoluta: para cálculo posterior da taxa mortes por população total

```
Populaçao Estimada Absoluta = CALCULATE([População Estimada por
Cidade],ALL('novo_caso_full'))
```

- Taxa Mortes por População Total: taxa de mortes pela população total, que pode ser filtrada por cidade

```
Taxa Mortes por População Total =
IFERROR(sum(novo_caso_full[new_deaths])/[Populaçao Estimada Absoluta],0)
```

Com essas métricas é possível criar filtros de análises que dão relatividade tanto para casos em que as cidades forem pequenas, quanto para cidades grandes. Logo, para se basear no percentual populacional de cada cidade, foi definido a filtragem por cidades com maiores taxas de casos confirmados de COVID-19 (‘Taxa Confirmados por População’) e cidades com maiores taxas de mortes (‘Taxa Mortes por População’), assim é possível verificar em quais cidades os casos de COVID-19 estão sendo mais agravantes.

Como a demanda é verificar quais cidades apresentam o maior crescimento na taxa de mortalidade, o filtro ‘Variação da Taxa de Mortalidade’ será usado para observar quais cidades tiveram maior crescimento no período do primeiro trimestre de 2021. Mas para tal análise, será feita primeiro uma visão macro de todas as cidades e depois outra visão macro de todos os estados, para focar a visão micro entre os 5 estados com maiores variações da taxa de mortalidade.

4 Interpretação dos resultados

Focando em responder as perguntas da demanda, a interpretação será por tópicos:

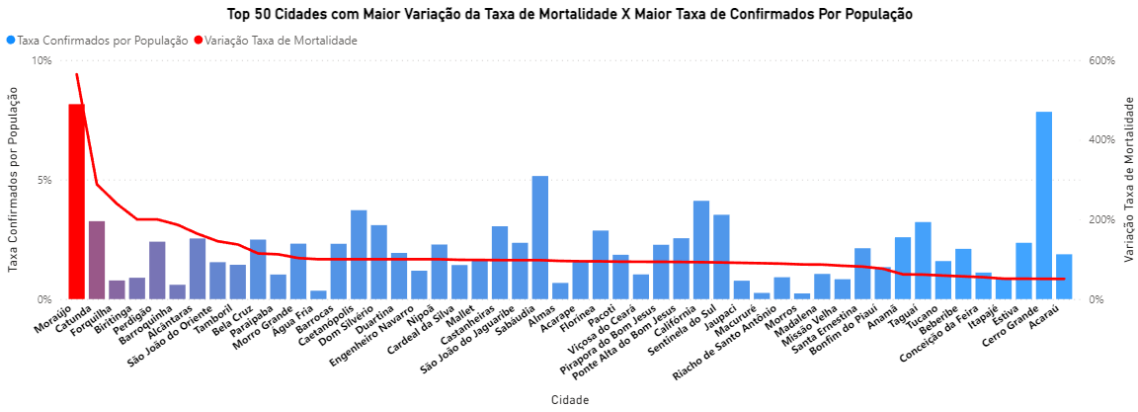
- **1.** Quais cidade apresentam maior crescimento na taxa de mortalidade por COVID-19 no primeiro trimestre de 2021

    - **1.1** Visão macro (todas as cidades)

    - **1.2** Visão micro (todas as cidades dos top 5 estados com maior variação)

- **2.** Quais as características em comum das cidades que justificam tal variação da taxa de mortalidade que se destaca das outras

1.1:



Na visão macro, o gráfico acima mostra ordenadamente as 50 cidades com os maiores crescimentos na taxa de mortalidade no período do primeiro trimestre de 2021 (representado pela linha passante sobre as barras), e a altura das barras representa a taxa de casos confirmados de COVID-19. Destas 50 cidades o valor da variação da taxa de mortalidade vai de 564,8% (em Moraújo no Ceará) até 50,58% (em Acaraú no Ceará). Neste gráfico a maioria das cidades possuem menos de 50 mil habitantes (de acordo com o IBGE classifica-se cidades pequenas aquelas com contingente populacional de até 50 mil habitantes), logo a variação da taxa de mortalidade afetou muito mais as cidades pequenas. Isso pode se dever ao fato de que, em cidades pequenas, as redes de serviços, sejam de quaisquer áreas, são muito limitadas. E isso faz com que os cidadãos destas cidades tenham que buscar estes serviços em outros municípios aos quais se conectam. E como a contaminação do COVID-19 teve início nos grandes centros urbanos, este deslocamento, principalmente por vias urbanas, acabou gerando uma disseminação do vírus de forma regional tardia, porém bastante intensa.

1.2:

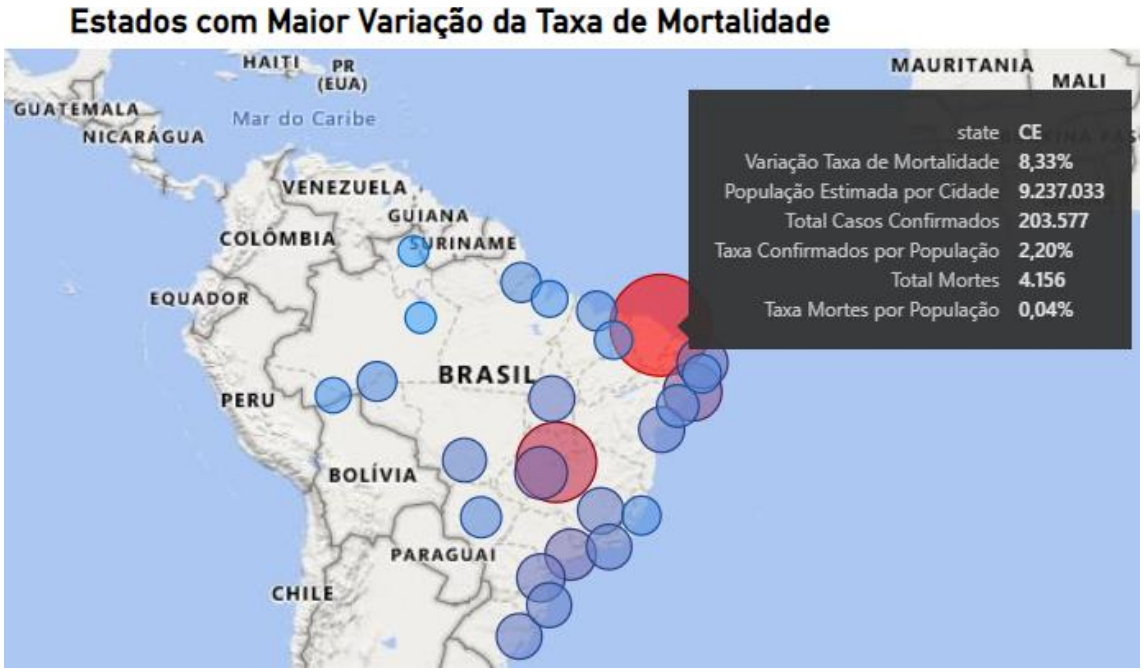
Na visão micro, o gráfico acima mostra a mesma ordenação que o gráfico de visão macro, porém com o uso do filtro de estados com maior variação na taxa de mortalidade, os quais são: Ceará, Distrito Federal, Alagoas, Goiás, São Paulo. Ainda assim a maioria das cidades contidas neste gráfico são cidades pequenas, portanto a mesma justificativa do gráfico da visão macro serve para o gráfico micro.

2.:

A característica principal que as cidades com maior variação na taxa de mortalidade possuem em comum é a de todas serem cidades pequenas. Então por conta da sazonalidade e

temporalidade da disseminação do vírus, por vias urbanas em busca de serviços em outros municípios, no primeiro trimestre de 2021 as cidades pequenas começaram a ter crescimentos acentuados na taxa de mortalidade.

Comentários: claramente que algumas regiões possuem características particulares, como, por exemplo, neste período, no Ceará, a cidade de Fortaleza estava se tornando o hub principal de voos internacionais da região do Nordeste e isso impactou bastante o número de contaminações nas cidades locais e periféricas. E isso se mostra bastante factível observando a variação da taxa de mortes por estado, como demonstrado no gráfico abaixo.



Todavia, o motivo da característica principal da variação da taxa de mortalidade ter sido grande nas cidades desta análise, é por serem cidades pequenas. E isto também acompanhado pelo fato da sazonalidade, do período do primeiro trimestre de 2021, e temporalidade da disseminação do vírus, de cidades urbanas grandes para médias e, por fim, pequenas.

5 Repositório no GITHUB

Caso queira somente visualizar o relatório gerado no Power BI, basta ter o Power BI Desktop instalado, baixar o arquivo do link https://drive.google.com/file/d/1FsTztSvqxBBCr9M2PsSLPFtvroecBf5R/view?usp=sharing e abrir o arquivo.

Os datasets podem ser acessados pelo link https://brasil.io/dataset/covid19/files/

O repositório criado no link https://github.com/IgorVillela/Case-Nexodata-COVID19 contém os arquivos:

- Link.txt: link para baixar os datasets
- Novo_caso_full – Relatório.pbix: relatório feito na plataforma de visualização Power BI
- Relatório.pdf
- Tratamento.jpynb: script utilizado para o tratamento dos dados