Descrição do projeto

- Qual é a descrição do problema?
 - O objetivo é entender se tivemos uma queda na taxa de vacinação no nosso país, a partir de dados coletados no site do <u>datasus</u>.
- Você tem premissas ou hipóteses sobre o problema? Quais.
 - Hipótese que houve uma queda em todas as regiões do país.

Principais aprendizados

Trabalhar com dados reais sempre traz uma complexidade a mais, isso é parte do dia a dia de um engenheiro de dados ou DBA, tratar com os dados abertos do governo mostrou como temos um trabalho maior para podermos ter acesso às informações de nosso país.

A estrutura que os dados são fornecidos são boas, mas limitadas, o que me gerou uma repetição de ações para gerar um dataframe com 10 anos de cobertura vacinal.

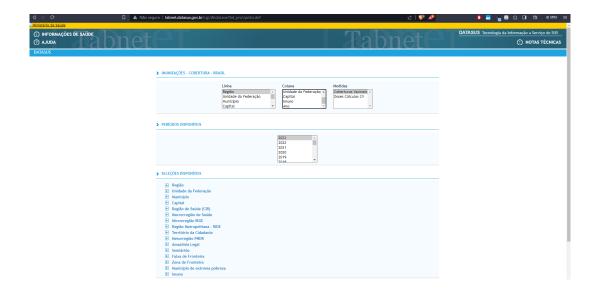
O tipo dos dados, gerou um pouco de problema, mesmo mudando no início do pipeline de varchar para integer acabava gerando uma tabela vazia. Um ponto a tentar é alterar vírgula por ponto no pipeline que na época não tentei.

Parte da conexão entre alguns produtos da Microsoft não é tão fácil para quem é iniciante, tive que pesquisar bem como criar a liberação do Pureview acessar o banco de dados como administrador.

A criação de pipelines no data factory foi muito simples e intuitiva, a visualização do pipeline como blocos facilita bastante esse processo.

Busca dos dados

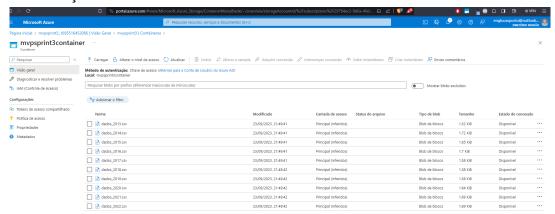
Selecionei dados sobre imunizações disponibilizadas pelo <u>Datasus</u> (http://tabnet.datasus.gov.br/cgi/dhdat.exe?bd_pni/cpnibr.def).



Coleta dos dados

Para coletar os mesmos dados, deve-se selecionar em linha: Região, coluna: Imuno. Porém uma dificuldade que foi imposta é de apenas conseguir detalhamento extraindo ano a ano, logo foram realizadas 10 extrações, de 2013 a 2022.

A inserção no store da azure foi realizada manualmente.

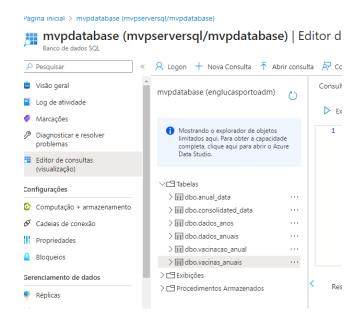


Modelagem

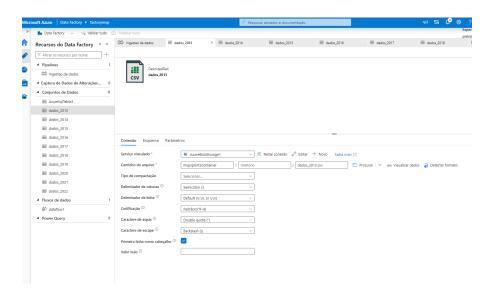
Como trabalharemos com apenas uma tabela neste projeto, não farei a criação de um modelo delta ou estrela. Para fins de catálogo foi utilizado o Microsoft Pureview que será mostrado após a carga, onde já foi explicado como foi criado o banco de dados.

Carga

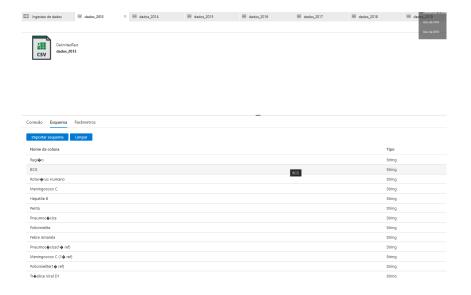
- Antes de realizar a carga, tive que criar um Banco de Dados SQL do Azure, nomeado de mvpdatabase.



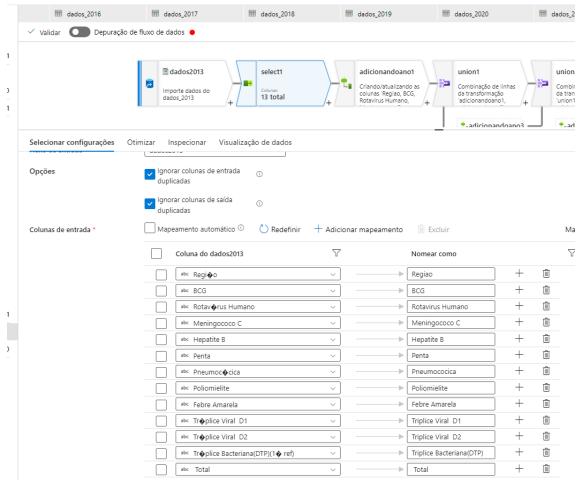
- Como visto na seção de Coleta dos Dados, inseri os dados em um blob na Azure.
- Na próxima etapa foi realizada a carga de todos eles no Data Factory.



Foi importado o esquema do próprio arquivo:

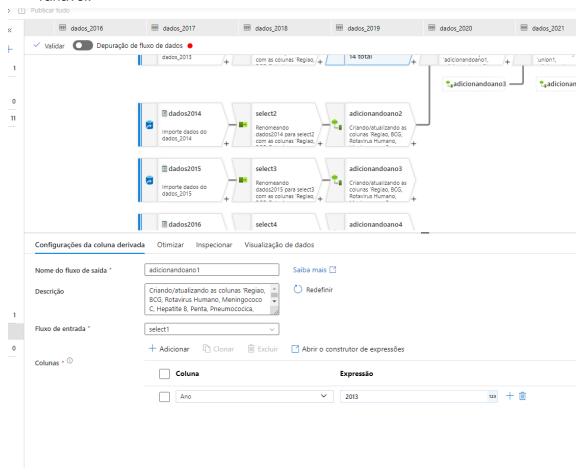


 Após a subida dos dados foi iniciado o Data Flow, onde selecionei apenas as colunas que queria trazer como informação relevante e alterei os nomes de colunas com erro de codificação.



Cada etapa de select foi executada da mesma forma para os 10 arquivos.

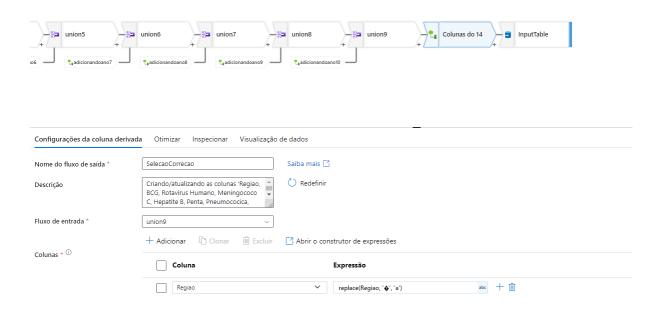
 Após select, adicionou uma coluna ano para cada arquivo, dando assim a clareza dos dados, já que no momento da extração não consegui deixar o ano como variável.



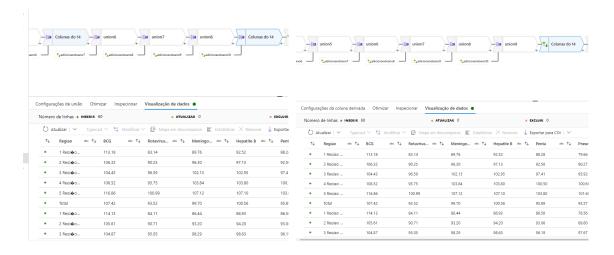
 Após esta etapa foi necessário realizar um union, que basicamente foi um append de todas as tabelas, como cada tabela é um data source diferente, tive que realizar 9 unions diferentes.



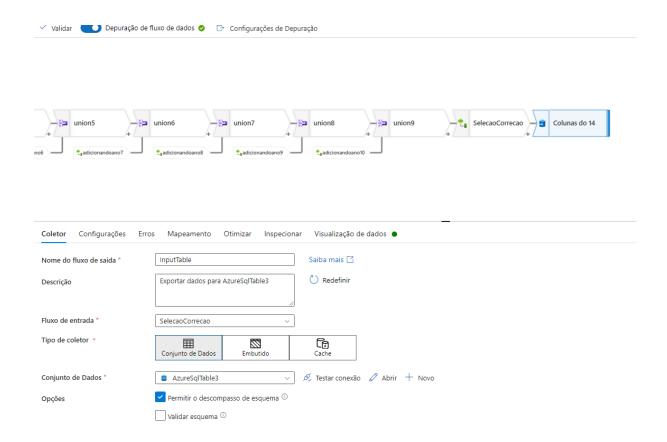
 Após esta etapa fiz a correção do erro de digitação de Região, alterando o dado da coluna Região.



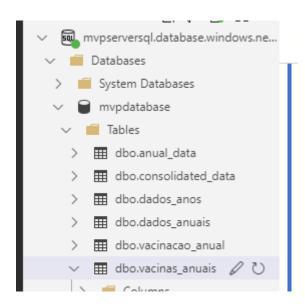
Abaixo eu mostro como era e como ficou a coluna, na esquerda trago como estava e após a correção.



- Por último, foi popular a tabela dbo.vacinas_anuais.



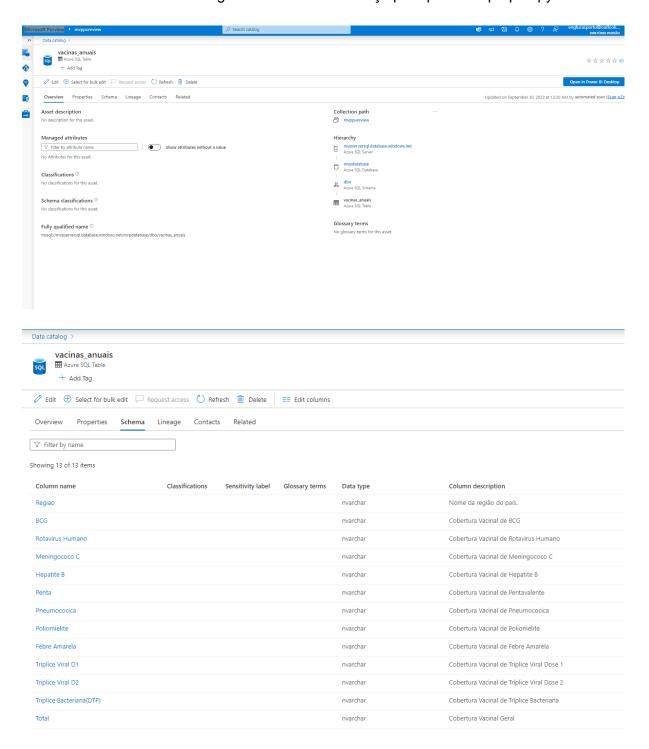
Que nos retorna uma tabela no nosso ambiente de banco de dados.



Catálogo de Dados

Como parte da organização dos dados e gestão do mesmo, foi utilizado o Microsoft Pureview como catálogo dos dados.

Primeiro realizamos a conexão com o banco de dados e realizamos o scan para conseguir trazer as informações desse banco de dados. Assim, podemos descrever o que cada coluna significa e validar o estilo dos dados, como durante o trabalho a vírgula que está presente no dado bruto foi importante para o banco de dados entender a informação, mantive as colunas como string e realizamos a mudança para ponto no próprio python.



Análise

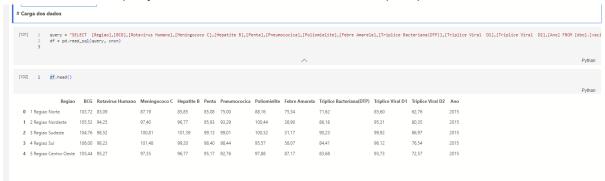
Qualidade de dados:

A base de dados continha conforme mencionado na carga, problemas de configuração, logo trazia dados com problema de escrita, foi necessário na etapa de Transformação executada no Data Factory alterar tanto os erros nas colunas quanto nas linhas. Esses problemas são facilmente trabalhados nas ferramentas, não afetando nossa qualidade dos dados finais.

Solução do Problema:

Conforme nossa proposta inicial, queríamos ver se tínhamos um problema de queda na vacinação, essa parte do trabalho foi realizada em um notebook em python na Azure Data Studio.

Foi realizada a importação das bibliotecas e carga dos dados via python, a carga dos dados é uma query no nosso banco de dados bem simples para trazer todos os dados.



Alterei o formato dos dados, já que os valores das coberturas vacinais estavam como string, foi alterada a vírgula por ponto e alterado o tipo de dado para float.

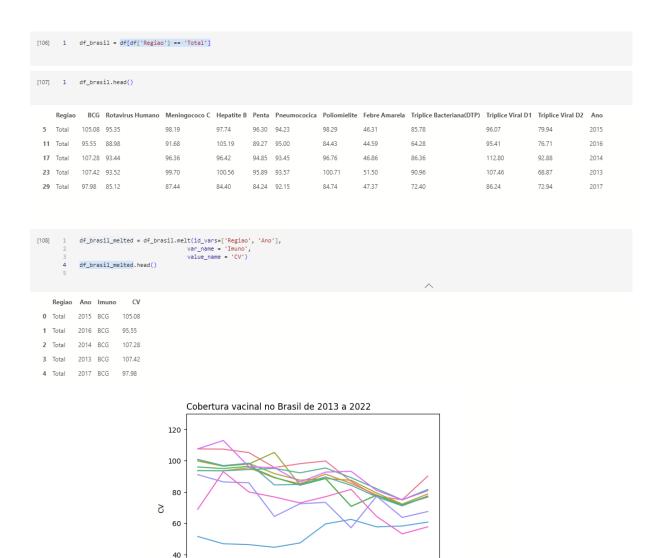
```
[103] 1 df.replace(',', '.', regex=True, implace=True)

Python

[105] 1 columas = ['8CG', 'Rotavirus Humano', 'Meningococo C', 'Hepatite B',
2 'Penta', 'Penunococica', 'Molionielite', 'Febre Amarela',
3 'Triplice Bacterianma(DTP)', 'Triplice Viral DI', 'Triplic
```

Houve queda na cobertura vacinal no Brasil?

Para responder a primeira hipótese se havia uma queda no Brasil na cobertura vacinal, separei os dados apenas do total, que corresponde ao país, e plotei os dados para analisar as séries temporais.



E o resultado que temos é, sim tivemos queda em quase todas as vacinas analisadas, com exceção da Febre amarela. Como temos pouco tempo para analisar a tendência final, podemos ver que as vacinas começaram a ter um retorno no aumento da cobertura vacinal em 2022. Pode ser um bom sinal da recuperação de nossa histórica boa política pública.

2018

2020

2022

— Triplice Bacteriana(DTP)

— Triplice Viral D1

— Triplice Viral D2

A queda foi igual em todas as regiões do país?

2014

Meningococo C

Hepatite B

2016

Rotavirus Humano — Pneumococica

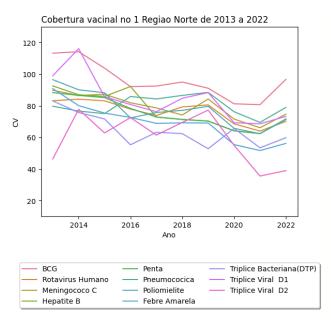
— Poliomielite

--- Febre Amarela

Separei os dados apenas das regiões para realizar o tratamento e verificar como estávamos com a série temporal.

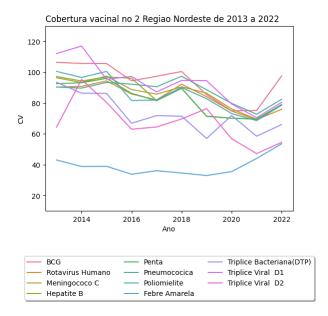
Podemos olhar por região:

Região Norte



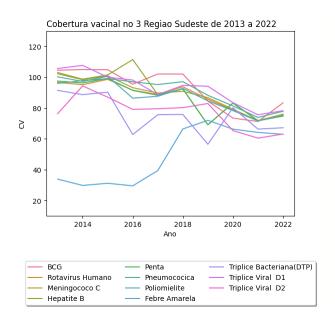
- Uma diferença que vemos de primeira é que nesta região apesar da queda, sempre teve a cobertura vacinal acima do que o consolidado do Brasil, porém nos últimos anos ficou próximo da média nacional.
- Porém todas as vacinas seguiram o padrão de queda do Brasil.

Região Nordeste

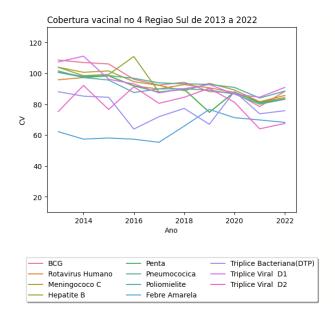


- Não tivemos um crescimento para a febre amarela no mesmo período que no Brasil como um todo, mas nos últimos anos começou a subir.
- Porém todas as outras vacinas seguiram o padrão de queda do Brasil.

Região Sudeste

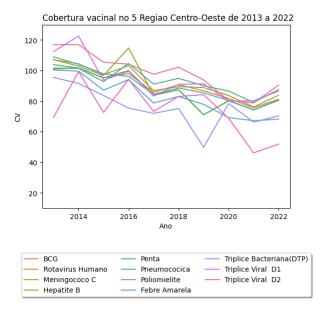


- Tivemos um crescimento acentuado na de febre amarela, muito acima do restante do país.
- Porém todas as outras vacinas seguiram o padrão de queda do Brasil.



- Tivemos um crescimento e estabilização próximo dos 80% para febre amarela.
- Porém todas as outras vacinas seguiram o padrão de queda do Brasil, porém não tão grande, mantendo próximo de 90% a maioria das vacinas.

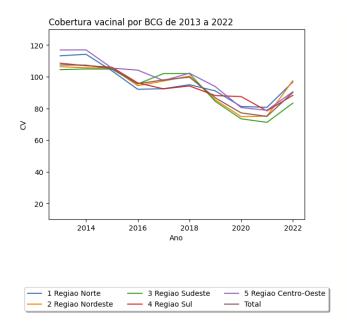
Região Centro-Oeste



- A cobertura vacinal de febre amarela n\u00e3o teve uma queda muito acentuada, mantendo at\u00e9 pr\u00f3ximo de 80%
- Porém todas as outras vacinas seguiram o padrão de queda do Brasil.

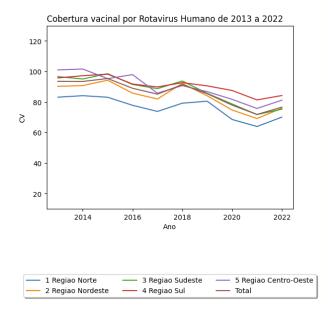
Quais vacinas tivemos mais quedas?

BCG:



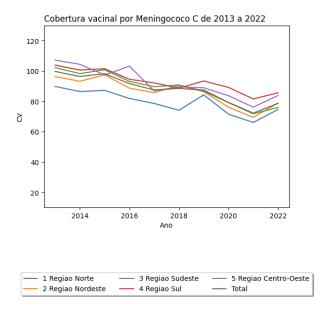
- Previne a tuberculose, a queda na cobertura vacinal, é preocupante, pois em 2021 tivemos 5 mil mortes, conforme <u>notícia</u> da USP.
- Todas as regiões tiveram queda parecida, tendo o Sudeste com a menor cobertura vacinal.

Rotavírus Humano:



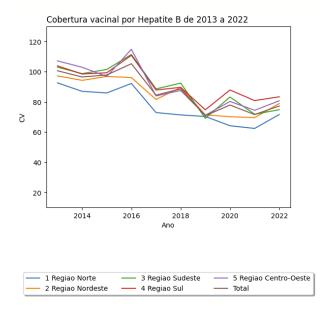
- Previne doenças diarreicas agudas, a queda foi igual em todo país, porém a região Norte nunca teve patamar nos últimos 10 anos de 100% de cobertura
- Um <u>estudo</u> com a participação Fiocruz mostra que esta doença pode ter causado a morte de 200 mil por ano, no período de 2017 a 2019.

Meningococo C:



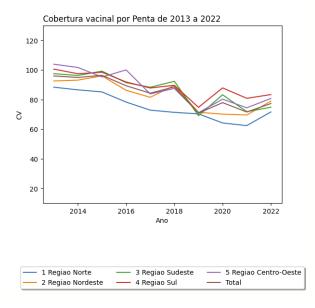
- Previne a meningite tipo C, a queda foi igual em todo país, porém a região Norte nunca teve patamar nos últimos 10 anos de 100% de cobertura.
- Apesar da ligeira subida em 2022, tivemos 3 surtos desta doença registrados, conforme informe da Fiocruz.

Hepatite B:



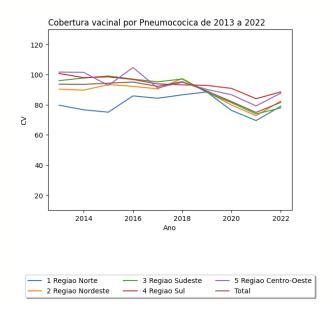
- Previne a hepatite B, a queda foi igual em todo país, tendo a maior parte das regiões chegando próximo dos 80% de cobertura vacinal
- Esta doença teve predominância na região Sul e Sudeste em 2022, conforme informado pelo Brasil 61.

Pentavalente:



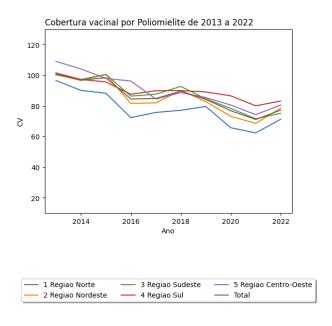
- Previne cinco doenças, que são elas: difteria, tétano, coqueluche, hepatite B e contra a bactéria haemophilus influenza tipo b.
- Queda em todas regiões, tendo a região Norte a menor cobertura dos últimos 10 anos em 2021.

Pneumocócica:



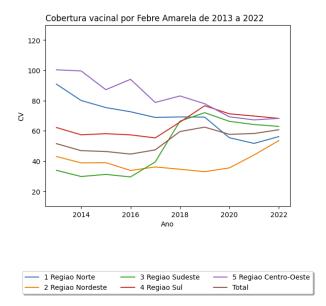
- Previne contra doenças como pneumonia, meningite e otite.
- Queda em todas regiões, tendo a Região Norte se aproximado da média nacional.

Poliomielite:



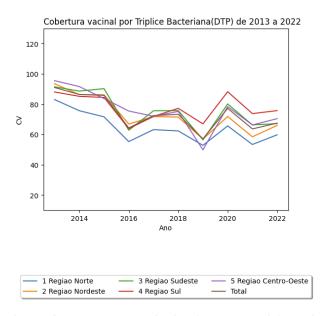
- Previne contra a poliomielite, a região norte teve a menor cobertura em 2021.
- Estamos sem esta doença no país há <u>34 anos</u>, apesar de termos diminuído a cobertura vacinal, porém isso representa um risco caso tenhamos algum caso externo entrando no país.

Febre amarela:



- Previne contra a febre amarela, a região Centro-Oeste e Norte possuíam as maiores coberturas, tendo as outras regiões iniciado um avanço na cobertura vacinal.
- A subida principalmente na região Sul e Sudeste da vacinação vem de termos tido casos, conforme temos elencado <u>aqui</u>, <u>aqui</u> e <u>aqui</u>.
- Originalmente por ser uma doença mais presente nas regiões rurais e amazônicas, sendo <u>endêmica</u> na região amazônica, sempre tivemos uma maior cobertura no Centro-Oeste e Norte do país.

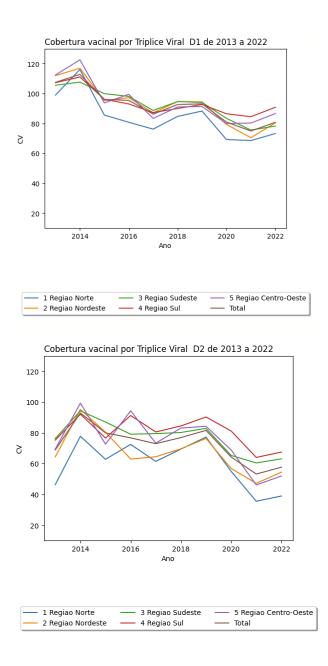
Tríplice Bacteriana (DTP):



- Previne a difteria, o tétano e a coqueluche (ou pertussis), muito indicada para reforço do esquema básico da pentavalente.

- Todas as regiões tiveram queda, mas a recuperação começou em 2020 diferente das outras vacinas.

Tríplice Viral:



- Previne a sarampo, coqueluche e rubéola, para dose 1 tivemos uma queda igual em todos as regiões.
- A dose 2 tem uma queda maior, porém temos outras vacinas que podem ser aplicadas para imunização também.
- Ponto importante que temos nessa notícia da Fiocruz:

"O Brasil recebeu, no segundo semestre de 2016, o certificado de eliminação do sarampo da Organização Panamericana de Saúde (Opas). Desde 2001, não havia registro de casos da doença originados no Brasil. Entre 2013 e 2015, ocorreram surtos relacionados à importação, sendo que o

maior número de casos foi registrado nos estados de Pernambuco e Ceará. Após a implementação de medidas de prevenção e controle, como intensificação vacinal, campanhas de seguimento, bloqueio vacinal, varredura e monitoramento rápido de cobertura vacinal, a transmissão foi interrompida."