Universidade Federal de Pernambuco - Centro de Informática
Sistemas de Informação

Mateus Ribeiro de Albuquerque

Análise descritiva, estatística e espacial sobre SCZ em Pernambuco no período 2015–2024

O Problema Nomao

Recife
2025

Contextualização

Este projeto realiza uma análise exploratória, temporal e espacial dos casos da Síndrome Congênita associada ao Vírus Zika (SCZ) notificados no estado de **Pernambuco, Brasil**, no período de 2015 a 2024.

OBS: Não é necessário mas é recomendado que o Notebook disponível em: https://github.com/MateusRiba/Analise-SCZ seja usado de apoio a leitura do dataset, principalmente caso haja o interesse de visualizar os códigos que permitem retirar as informações aqui demonstradas. Neste há também muitos dos insights aqui apresentados.

Objetivos:

1. Preparação de Dados (DataTransforming + DataMerging + Main)
2. Exploração de Dados
3. Análise Temporal
4. Mapas
5. Testes de Autocorrelação Espacial
6. Identificação de Clusters
7. Modelagem (Regressão Espacial/Contagem)

Etapa 1 - Preparação e Unificação dos Dados

Os arquivos brutos disponibilizados pelo sistema oficial estavam originalmente no formato .dbc, um tipo de arquivo compactado derivado de DBF (dBase), utilizado bastante em sistemas de informação em saúde no Brasil. Para permitir que esses dados fossem lidos e processados de forma universal dentro do ecossistema analítico em Python, foi desenvolvido um **conversor robusto em R** capaz de **transformar arquivos .dbc em .csv** (formato amplamente compatível e legível). O objetivo principal dessa etapa foi descompactar, ler e converter os arquivos de entrada, garantindo sua integridade e padronização para as etapas seguintes de análise.
Essa etapa é fundamental para o projeto, pois padroniza a entrada de dados para o pipeline em **Python**, que **trabalha exclusivamente com formatos .csv e .parquet.** o CSV garante ampla compatibilidade com outras ferramentas.

O próximo passo consistiu na **padronização** e **unificação** das bases de dados originais referentes às notificações da Síndrome Congênita associada ao Zika Vírus (SCZ) em Pernambuco, abrangendo o período de 2015 a 2024. Isso pode ser feito após a retirada dos dados por meio do site do GOV.
Os arquivos de entrada, agora em formato .csv, estavam distribuídos em subpastas dentro do diretório Data/processed/, representando diferentes anos e formatos de coleta.
Para viabilizar uma análise integrada e consistente, foi desenvolvido um script em Python denominado DataMerging.py, responsável por **Localizar automaticamente todos os arquivos CSV dentro do diretório especificado** (inclusive em subpastas), utilizando varredura recursiva (Path.rglob), **Ler os arquivos de forma robusta**, testando múltiplas

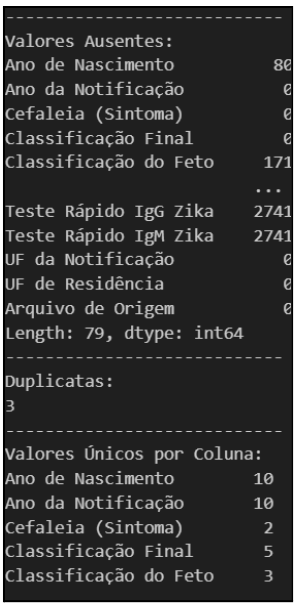
codificações (utf-8 e latin1) e padronizando nomes de colunas e valores ausentes, **Adicionar uma coluna auxiliar __source_file, que identifica a origem de cada registro** (nome do arquivo de origem), garantindo rastreabilidade dos dados após a unificação e mais importante: **Concatenar todas as tabelas em um único DataFrame**, gerando uma base consolidada denominada **dados_unificados**, contendo todos os registros do período de estudo.

Ao término da execução, o script exibe o número total de linhas e colunas resultantes, além de salvar os arquivos finais no diretório Data/processed/.

Etapa 2 - Organização e Insights Preliminares

Agora, via Notebook, foram realizadas importações de bibliotecas necessárias para a análise e a organização inicial dos dados.

Seguem imagens retiradas do notebook:



Estrutura geral do dataset

- Dimensão: linhas = número total de registros notificados; colunas = variáveis clínicas, laboratoriais, demográficas e administrativas.

- Padronização: os nomes foram **mapeados para rótulos legíveis** (ex.: **ANO_NASC** → **“Ano de Nascimento”**, **MICROCEFAL** → **“Microcefalia”**, etc.), o que facilita leitura de tabelas e gráficos.

- Datas: campos Data de Nascimento, Data da Notificação, etc., são críticos para recortes temporais e construção de séries; **já estão em tipo data** (devido a primeira preparação feita na etapa 1**)** no pipeline Python.

- Integridade:
    - Valores ausentes (NaN) foram inspecionados. Alguns grupos de variáveis (especialmente exames) têm falta de preenchimento relevante (ex.: “Exame de Ultrassonografia”, “Resultados STORCH/Zika”).

    - Duplicatas: checadas; se detectado, tratado (ex.: manter primeiro registro ou consolidar por “Código Sequencial do Registro”).

A completude desigual entre campos indica necessidade de filtros ou imputações cuidadosas nos próximos passos (principalmente para relatórios e gráficos comparativos).

Visão rápida do Perfil demográfico e clínico

- Idade da Gestante (campo “Idade da Gestante”):

    - Média ~ 25,7 anos, mediana 26 anos; variação ampla (mín. ~13, máx. ~45).

Distribuição concentrada em adultas jovens, com presença de gestação na adolescência e idades mais avançadas — faixas relevantes para estratificações futuras (riscos e desfechos).

- Classificação do Feto (“Classificação do Feto”):

    - Predomínio de A termo (37–41s) e parcela importante Pré-termo.

Implica olhar para relação entre prematuridade e desfechos (microcefalia, óbito, alterações no SNC).

- Tipo de Gravidez (“Tipo de Gravidez”):

    - Majoritariamente Única, poucas múltiplas.

Implica pouco impacto de gêmeos no agregado; ainda assim vale ajustar por esse fator em modelos.

- Sintomas na gestação (Cefaleia, Febre, Dor articular/muscular, Conjuntivite, Edema, Prurido):

    - Médias próximas a “2 = Não” após mapeamento; ou seja, baixa frequência de relato de sintomas nos registros.

Interpretação: pode refletir subregistro de sintomas, diferença de protocolo (notificação/acolhimento) ou foco do fluxo no desfecho neonatal (microcefalia). Não inferir ausência clínica apenas a partir do banco.

- Microcefalia (“Microcefalia”):

    - A variável foi categórica (e não apenas binária). Nos primeiros cortes, observa-se parcela relevante de “Microcefalia apenas” ou “Microcefalia com outras alterações”.

Essa distinção será útil na análise temporal (tipos ao longo do tempo) e na espacial (distribuição de fenótipos por município).

Exames e resultados laboratoriais

- Ultrassonografia / Tomografia / Ressonância:

    - Muitas linhas com “Não realizado” / “Ignorado”.
    - Resultado para Zika e STORCH (sífilis, toxoplasmose, citomegalovírus, herpes) também apresentam alto não preenchimento.

Isso limita análises etiológicas diretas sem critérios de completude. Possivelmente não uteis para o modelo geográfico.

- Perímetro cefálico / Comprimento / Peso ao nascer:

    - Perímetro cefálico com média ~ 34 cm está compatível com neonatos a termo; valores extremos podem sinalizar microcefalia/macrocefalia e devem ser conferidos (outliers e coerência com “Classificação do Feto”).
    - Peso/Comprimento: úteis para boxplots por classificação final (Confirmado/Descartado) e por tipo de microcefalia, testando diferenças de distribuição.

Em seguida, aplicou-se um mapeamento categórico abrangente (category_map) nas variáveis codificadas. O objetivo foi substituir códigos por descrições de significado, por exemplo:

- Tipo de Notificação:
  1 → “Recém-nascido com microcefalia (≤ 28 dias)”; **2 → “Criança com microcefalia… (> 28 dias)”**; **3 → “Feto com alterações do SNC”**; etc.

- Microcefalia:
  1 → “Microcefalia apenas”; **2 → “Microcefalia com alteração do SNC”**; **3 → “Microcefalia com outras alterações congênitas”**; **4 → “Alterações congênitas sem microcefalia”**.

- Sintomas e exames (Cefaleia, Conjuntivite, Dor Articular, PCR/IgG/IgM Zika, STORCH e seus desdobramentos no RN):
  códigos 1/2/3/4/9 mapeados para “Sim/Não/Indeterminado/Não realizado/Ignorado” conforme o dicionário.

- Classificação Final, Etiologia, Regiões, Sexo, Status da Notificação/Óbito, Classificação do Feto, Tipo de Detecção etc.:
  todos receberam rótulos textuais padronizados.

Pontos de atenção no mapeamento:

- As colunas foram convertidas para string antes de mapear, garantindo que valores como 1, 1.0 ou 1 em string fossem tratados de forma uniforme.

- Em casos específicos (ex.: “Classificação do Feto”), houve chave ‘1.0’ no dicionário para cobrir entradas importadas como float; a função de mapeamento primeiro converte a série para str, garantindo a correspondência correta.

- Qualquer valor sem correspondência é mantido (não vira NaN), preservando a integridade.

O dataset mapeado foi salvo como dados_unificados_mapeado_renomeado.csv, pronto para visualização e gráficos (sem “gírias” numéricas internas).
Por fim, para manter o dataset **enxuto e focado** na análise exploratória e visual, foi removido:

- **“Código Sequencial do Registro”**: identificador técnico, não agrega insight estatístico.

- **“Data de Nascimento da Mãe”**: a **idade da gestante** já está disponível e é mais informativa para o objetivos; manter a data traria pouca utilidade e maior sensibilidade.

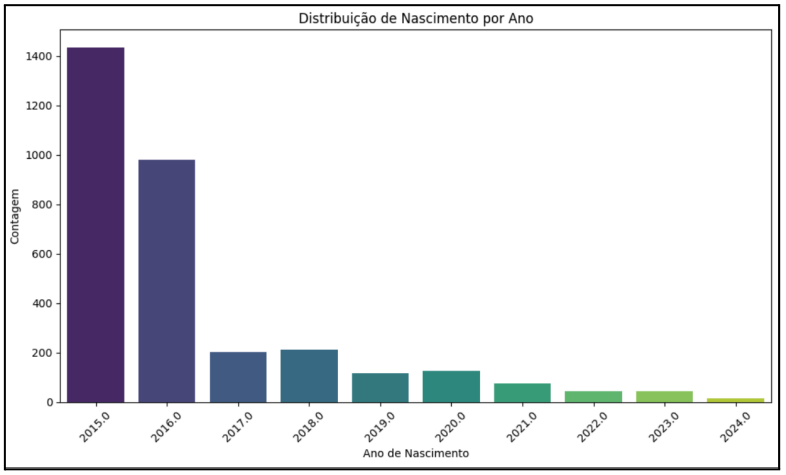
Essa limpeza reduz o ruído e o risco de interpretações indevidas, sem perda de informação analítica relevante.

Etapa 3 - Exploração dos Dados

O próximo passo, mais direto, foi a elaboração de gráficos para o melhor entendimento da situação e compreensão das informações que podem ser obtidas com o dataset. Nesta secção do relatório, gráficos extraídos por meio das bibliotecas plotly, seaborn e matplotlib serão demonstrados e interpretados.

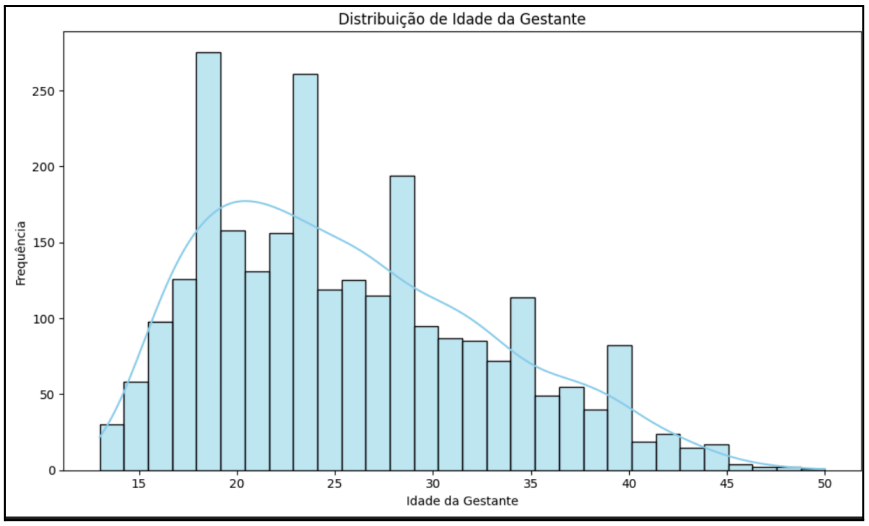
Histogramas:

1.



A grande concentração de casos no ano de 2015 sugere uma alta incidência de microcefalia e outras condições relacionadas ao Zika, provavelmente devido à epidemia de Zika nesse período. Isso pode indicar um pico de casos associados ao vírus no período crítico de infecção, o que foi muito documentado no Brasil. Após 2015, os casos diminuem consideravelmente, com poucos casos em anos subsequentes. Isso pode refletir uma diminuição na incidência do vírus ou uma melhoria no controle das infecções.

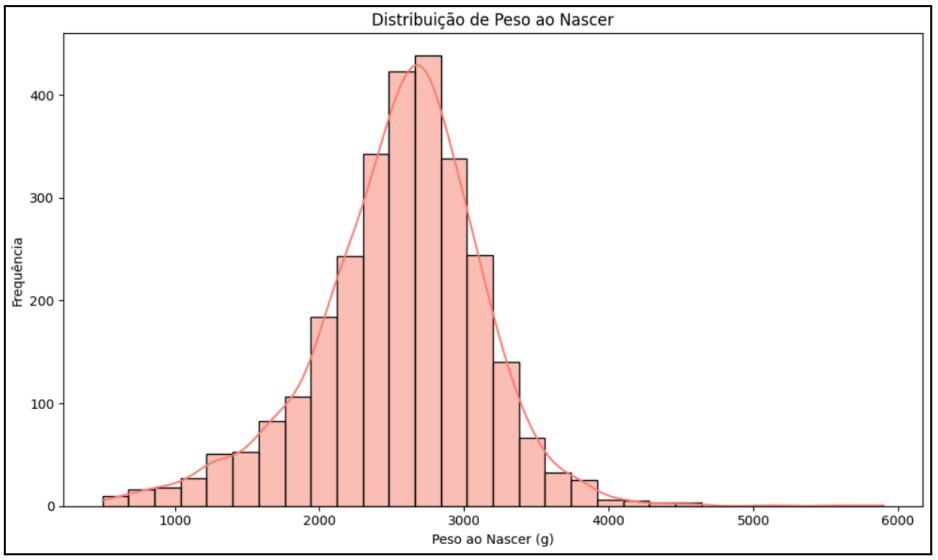
2.



A maioria das gestantes está na faixa etária entre 15 e 30 anos, com uma concentração maior de casos entre 20 e 25 anos. Nada disso é necessariamente uma indicação para a microcefalia, visto que, a maioria das gestantes em geral se concentra nessa faixa. Porém,

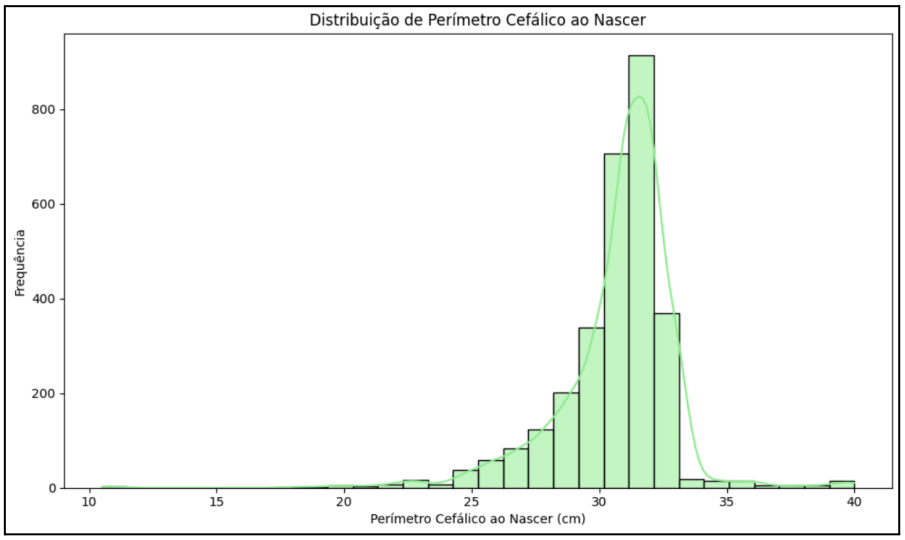
A cauda direita (maior idade) pode sugerir que, embora a maioria das gestantes seja jovem, também há um número considerável de gestantes com idades mais avançadas**, o que pode ser relevante para a situação.

3.



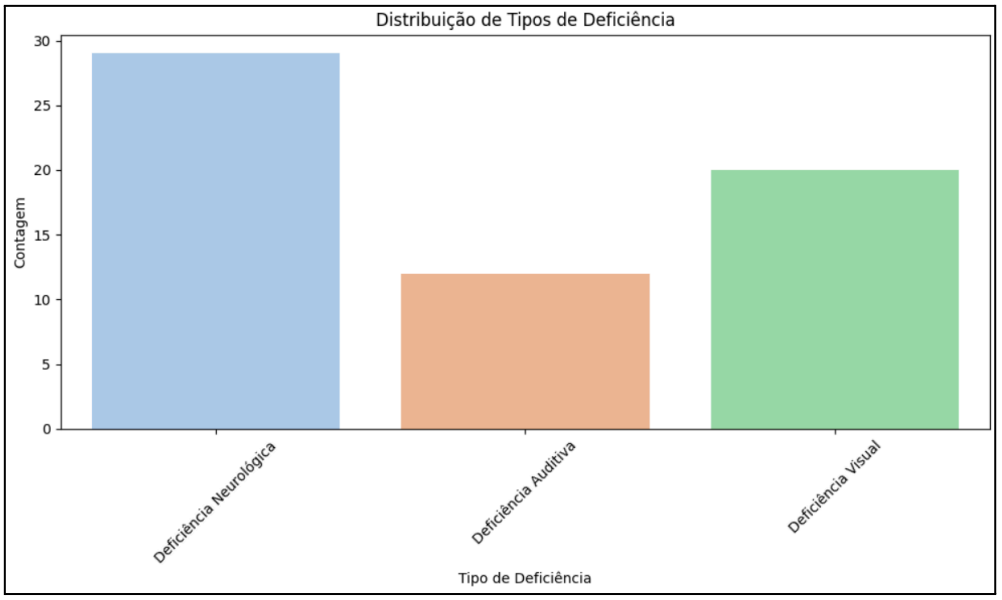
O gráfico de peso ao nascer mostra uma distribuição normalmente inclinada para a direita, com a maioria dos bebês nascendo com peso entre 2500g e 3500g. No entanto, a presença de outliers pode indicar casos de bebês com baixo peso ao nascer, o que pode ser indicativo de prematuridade ou outras complicações relacionadas à infecção por Zika. Porém, a maioria dos nascimentos foi saudável em termos de peso.

4.



A distribuição do perímetro cefálico também segue um padrão normal, mas com uma cauda à esquerda. A maioria dos bebês tem um perímetro cefálico saudável (cerca de 34-35 cm), mas a presença de uma **cauda mais à esquerda indica a presença de bebês com microcefalia**.

5.



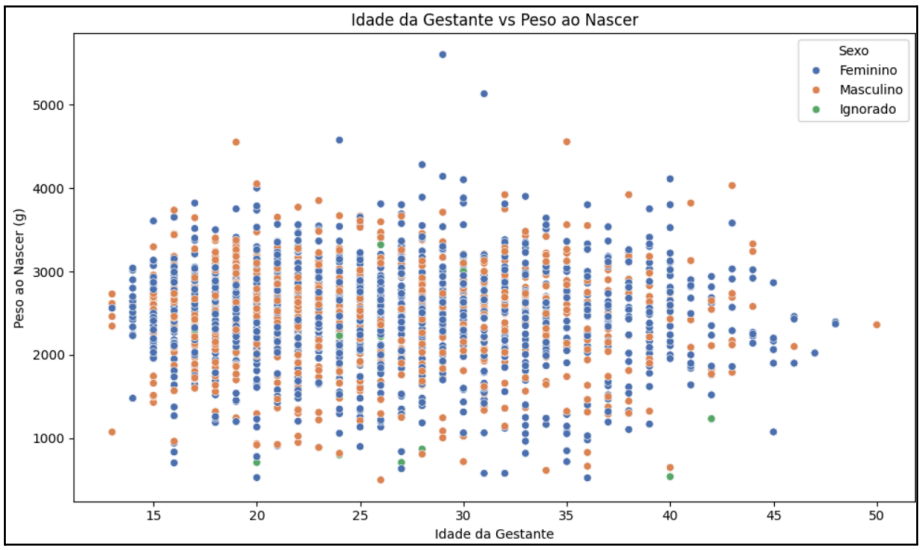
A maior parte dos casos com deficiência atrelada está associada a deficiência neurológica, como esperado devido à natureza da Síndrome Congênita da Zika (SCZ), que pode resultar em comprometimento neurológico (como microcefalia e outras alterações cerebrais). Esse gráfico mostra que a deficiência neurológica é a mais prevalente entre os casos analisados
A deficiência visual também tem uma presença significativa nos casos, embora não tão alta quanto a deficiência neurológica. Isso pode estar relacionado a danos ao sistema nervoso central causado pela infecção, que muitas vezes afeta a visão
A deficiência auditiva tem menos prevalência em comparação com as outras deficiências. Esse valor é mais baixo, sugerindo que embora a infecção por Zika possa afetar a audição em alguns casos, essa **não é a condição mais comum** entre os casos de microcefalia então, há uma consistência nos dados.

Tabela estatística baseada na Idade da Gestante

Idade da Gestante	mean	var	std
13	2.2975	0.375987	0.613178
14	2.538158	0.120273	0.346804
15	2.496945	0.20568	0.453519
16	2.52843	0.34128	0.584192
17	2.568522	0.190333	0.436271
18	2.5094	0.241593	0.491521
19	2.537872	0.280271	0.529406
20	2.5651	0.369324	0.607721
21	2.544054	0.28311	0.532081
22	2.466238	0.316394	0.562489

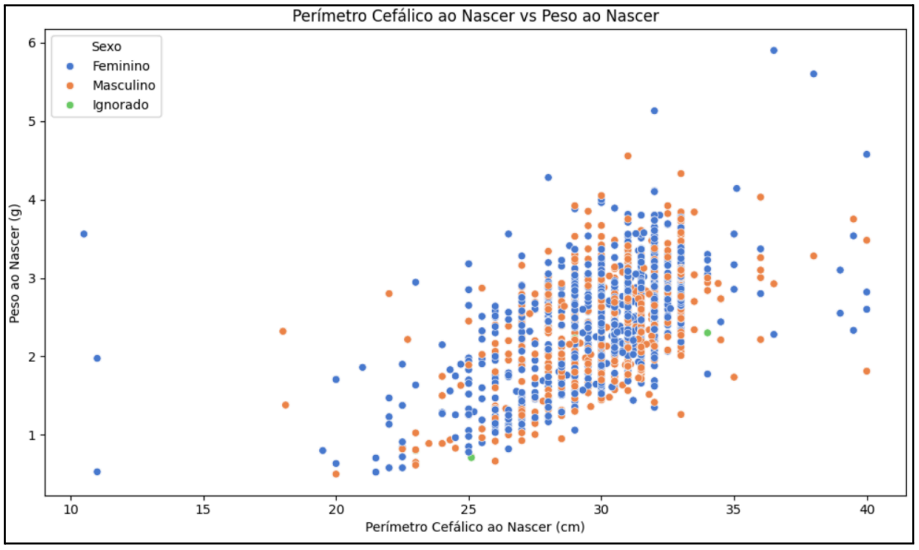
Gráficos de Dispersão

6.



A Relação entre idade da gestante e peso ao nascer é visível neste gráfico de dispersão. Pode-se observar uma distribuição mais densa de pontos na faixa etária 20-35 anos, com a maioria dos bebês nascendo dentro da faixa normal de peso.
Gestantes muito jovens (13–16 anos) e mais velhas (acima de 40) tendem a apresentar médias de peso ao nascer um pouco menores, sugerindo maior risco de baixo peso. O desvio padrão é maior em algumas faixas (ex: 35–36 anos), indicando maior dispersão dos pesos ao nascer nessas idades, possivelmente por maior heterogeneidade clínica ou social. Idades com poucos casos (ex: 47, 48, 50) têm variância e desvio padrão muito baixos ou nulos, indicando poucos registros e menor confiabilidade estatística. Faixa de 20–30 anos apresenta médias mais altas e menor variabilidade, reforçando que é o grupo de menor risco para baixo peso ao nascer.

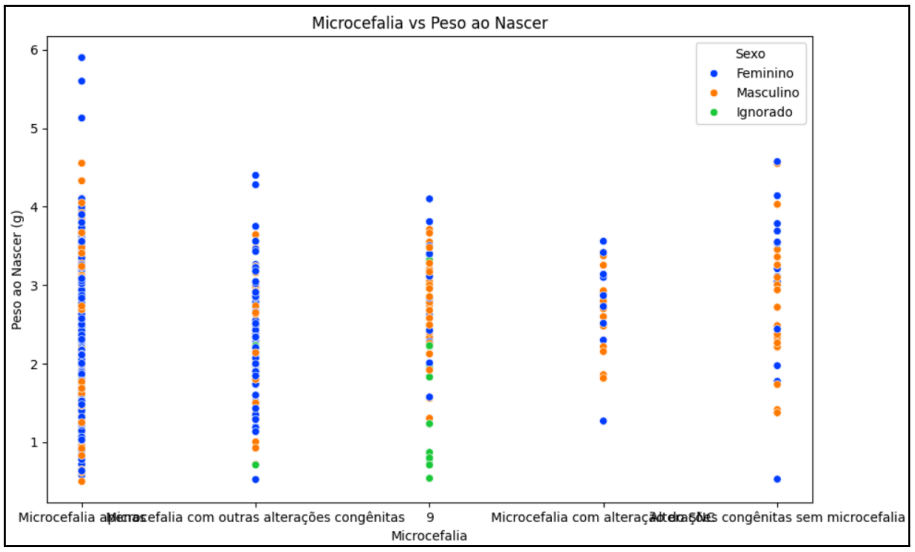
7.



Existe uma correlação positiva entre o perímetro cefálico e o peso ao nascer. Bebês com perímetros cefálicos maiores tendem a ter mais peso ao nascer, o que é esperado, já que microcefalia é frequentemente associada a bebês com peso mais baixo.

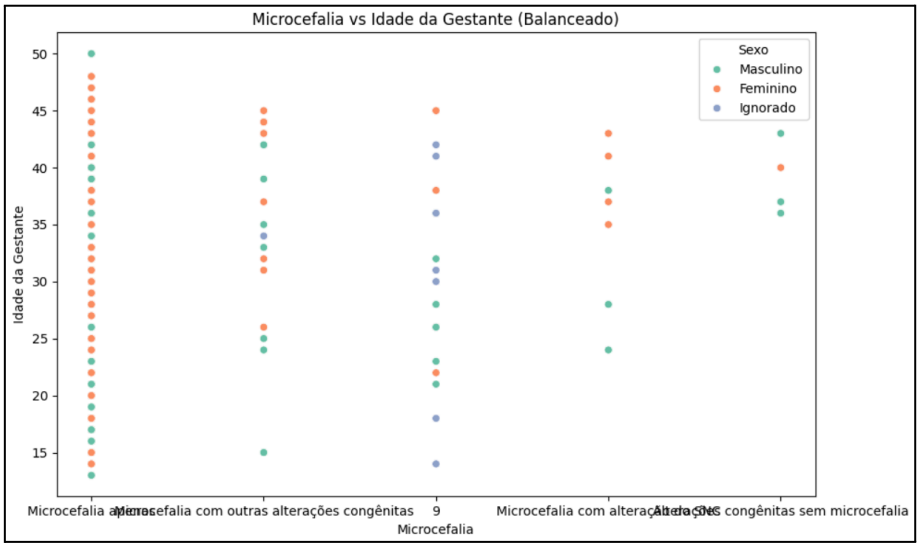
O gráfico também sugere que a maior parte dos bebês apresenta perímetros cefálicos dentro da faixa normal, com alguns casos mais extremos, possivelmente associados à microcefalia.

8.



A microcefalia está claramente associada a bebês com menor peso ao nascer. Como esperado, bebês com microcefalia apresentam peso ao nascer mais baixo (Talvez a exceção daqueles com alteração no SNC).
Este gráfico ajuda a confirmar a relação entre microcefalia e baixo peso e pode servir para identificar a gravidade da condição.

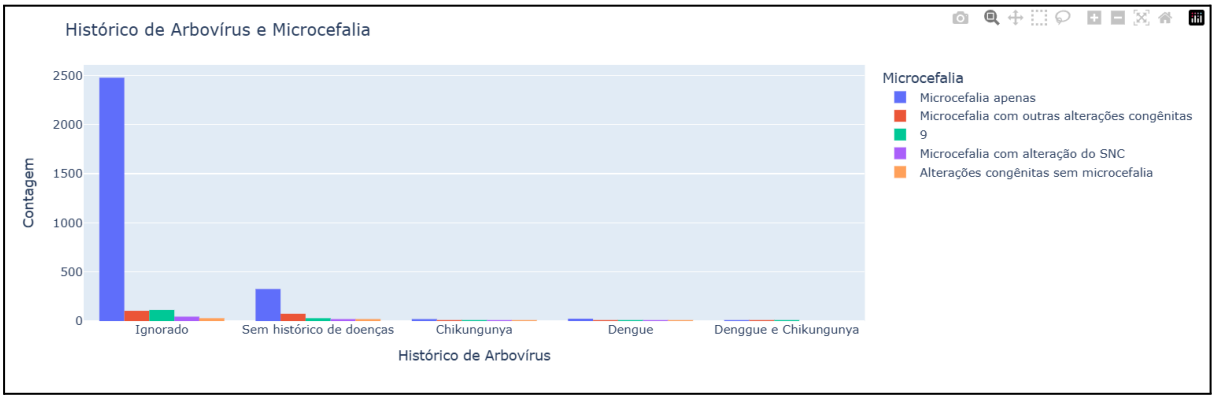
9.

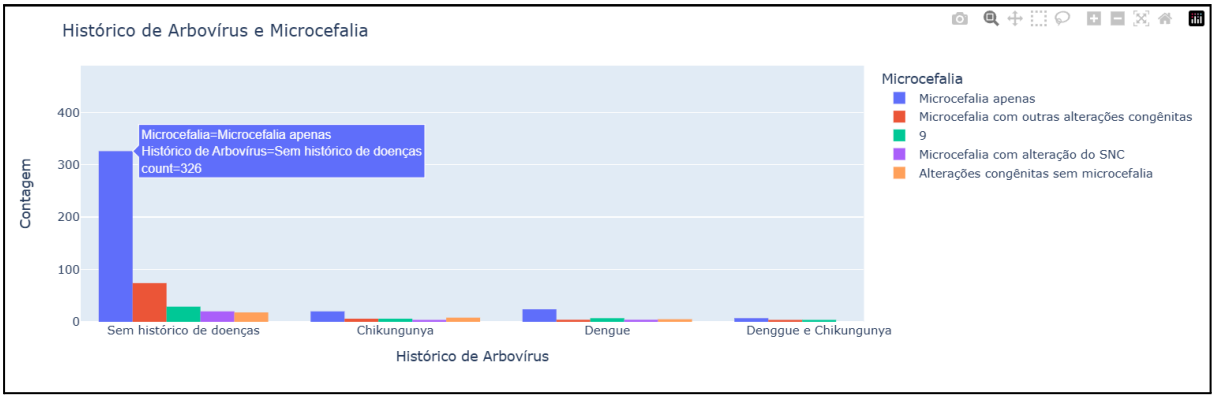


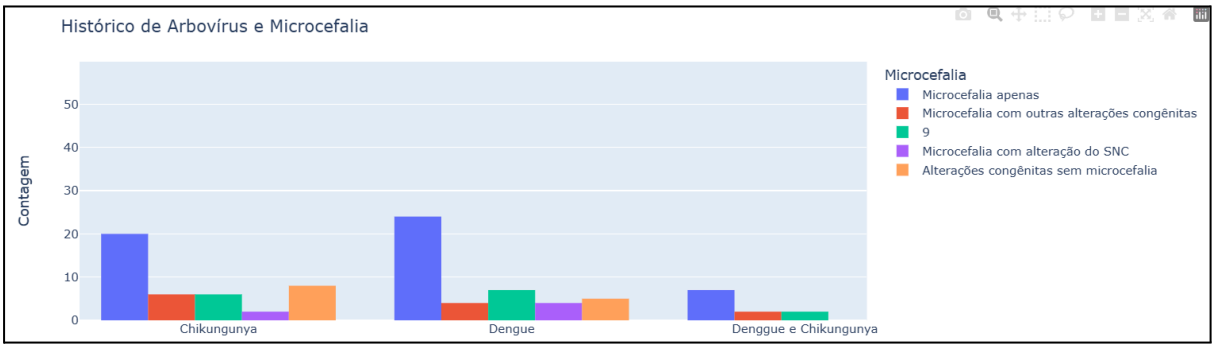
Microcefalia é observada em todas as faixas etárias, mas não há uma relação direta clara entre a idade da gestante e a ocorrência de microcefalia, segundo o gráfico.
Mesmo assim, podemos ver que a microcefalia está mais distribuída entre as gestantes de 20 a 40 anos. Isso pode indicar que a idade materna não é o único fator de risco para a microcefalia, já que casos também são encontrados em gestantes mais jovens e mais

velhas. Tambem pode indicar que, a base de dados desbalanceada aponta mais casos de microcefalia no geral na classe majoritária. Gestações mais velhas não necessariamente estão associadas a um aumento em casos de microcefalia. Percebemos porem que, com o Dataset Balanceado (Ou seja, para cada 10 casos em cada idade) é visivel que para idades mais elevadas, há mais casos de microcefalia com outras alterações congênitas e microcefalia com alteração do SNC.
OBS: O gráfico demonstra valores balanceados ou seja, cada classe (idade) foi limitada a uma amostra de 10 valores apenas, demonstrando assim uma melhor distribuição proporcional.

10.





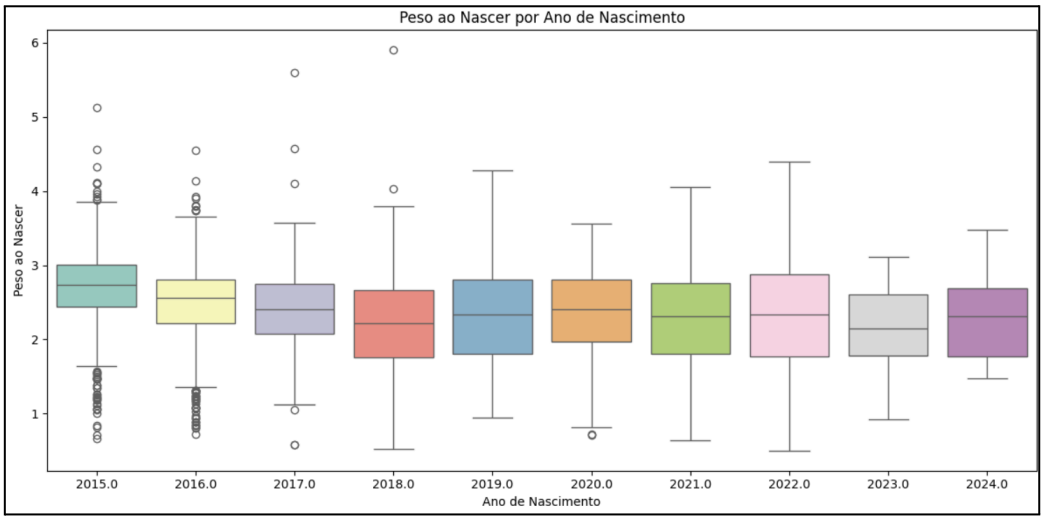


No primeiro gráfico, observa-se que a imensa maioria dos registros aparece na categoria **“Ignorado”**, indicando que, em grande parte das notificações, não houve informação preenchida sobre o histórico prévio de infecção por arbovírus (como Zika, Dengue ou Chikungunya). Esse dado ressalta uma **limitação** importante do sistema de vigilância, a subnotificação ou ausência de coleta dessa variável. Isso reduz a capacidade de estabelecer correlações diretas entre histórico infeccioso e desfecho fetal.

Entre os registros com informação disponível (segundo gráfico), destaca-se a categoria **“Sem histórico de doenças**”, que representa o maior número de casos com informação válida. Dentro desse grupo, predomina amplamente a categoria “Microcefalia apenas”, seguida de “Microcefalia com outras alterações congênitas”. Isso sugere que, mesmo entre as gestantes que não relataram infecção prévia por arbovírus, houve ocorrência significativa de microcefalia que pode estar associado ao fenômeno amplamente relatado durante a epidemia de 2015–2016.
No terceiro gráfico, ao focar especificamente nas categorias Chikungunya, Dengue e Coinfecção (Dengue e Chikungunya), percebe-se um número pequeno, mas não nulo, de casos associados à microcefalia. Nesses grupos, a predominância ainda é de “Microcefalia apenas”, com algumas ocorrências de “Microcefalia com outras alterações congênitas” e “Alterações congênitas sem microcefalia”. Esses resultados indicam que, embora a Zika tenha sido o principal agente etiológico da síndrome congênita, outros arbovírus também podem ter contribuído em menor escala ou atuado no impacto dos casos.

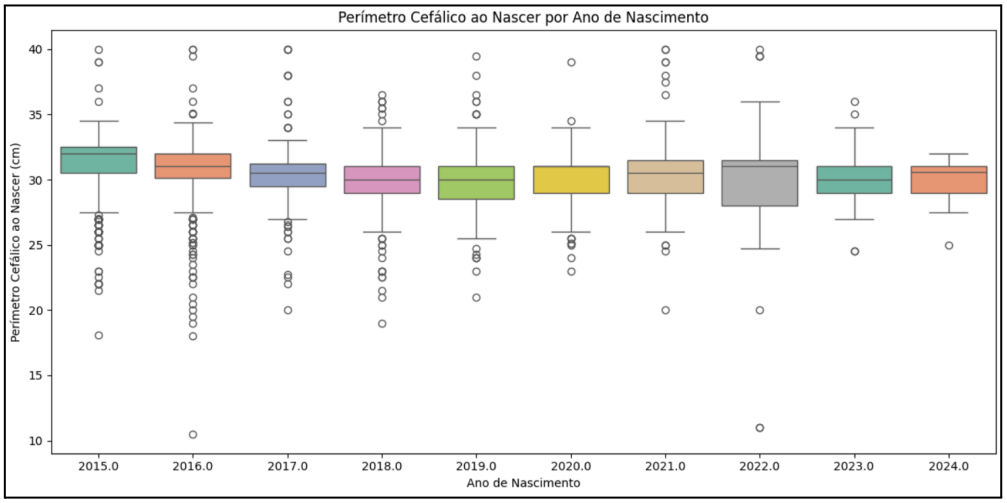
Boxplots (Gráficos de Caixa)

11.



O peso ao nascer permanece relativamente estável ao longo dos anos, porém com uma quantidade significativa de outliers baixos nos anos de 2015 e 2016 demonstrando claramente uma conexão com a epidemia de Zika do ano relacionado.

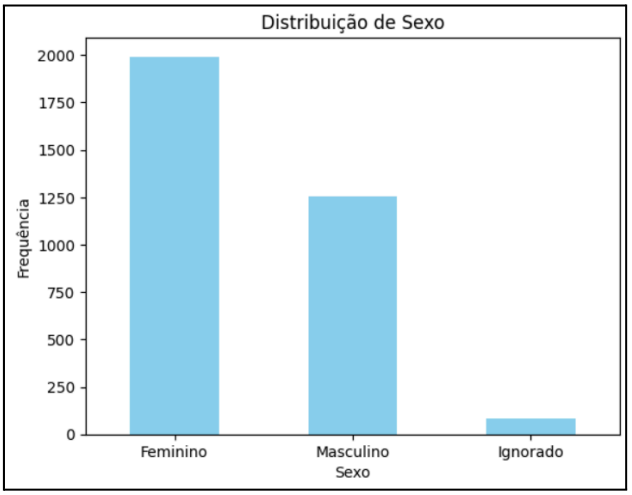
12.



Em 2015/2016 anos apresentam **maior quantidade de outliers (valores fora do padrão)** abaixo de 2 kg, o que sugere maior incidência de baixo peso possivelmente tambem reflexo direto da epidemia de Zika e da síndrome congênita que atingiu gestantes nesses anos. 2015 e 2016 possuem muitos outliers abaixo de 28 cm no segundo gráfico, sugerindo altos índices de microcefalia nesses períodos. Nos anos seguintes a distribuição se torna mais concentrada, com menos casos fora da faixa normal. Isso indica redução da ocorrência de microcefalia nos anos pós-epidemia. Outliers altos podem representar variações naturais (bebês maiores), não necessariamente anômalas.

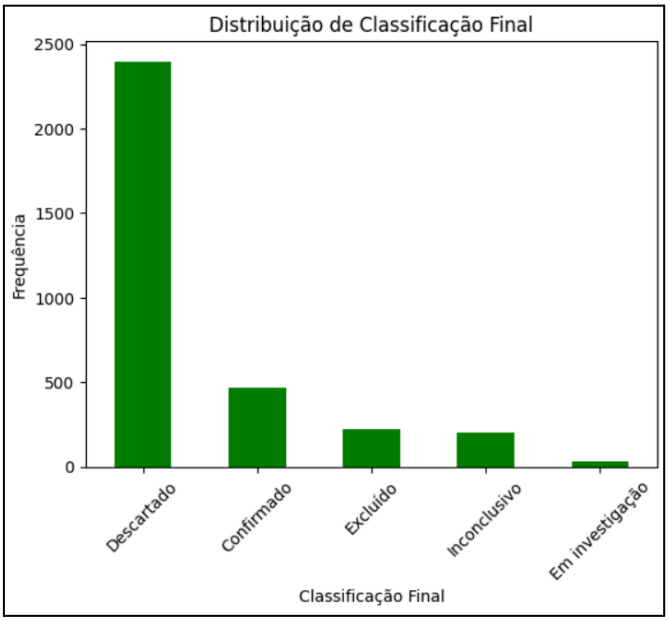
Análise de Variáveis Categóricas

13.



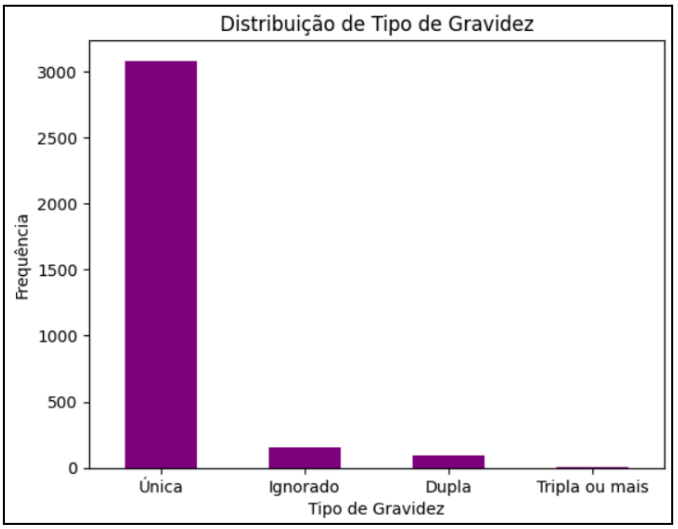
A frequência de recém-nascidos do sexo feminino, masculino e ignorado.
A maior parte dos registros corresponde a bebês do sexo feminino (~2000 casos), superando os masculinos (~1250).
Essa diferença pode estar relacionada a viés de registro, diferenças biológicas ou simplesmente amostragem.
A variável sexo está bem distribuída e completa, podendo ser usada em análises cruzadas com microcefalia, peso ao nascer e tipo de gravidez.

14.



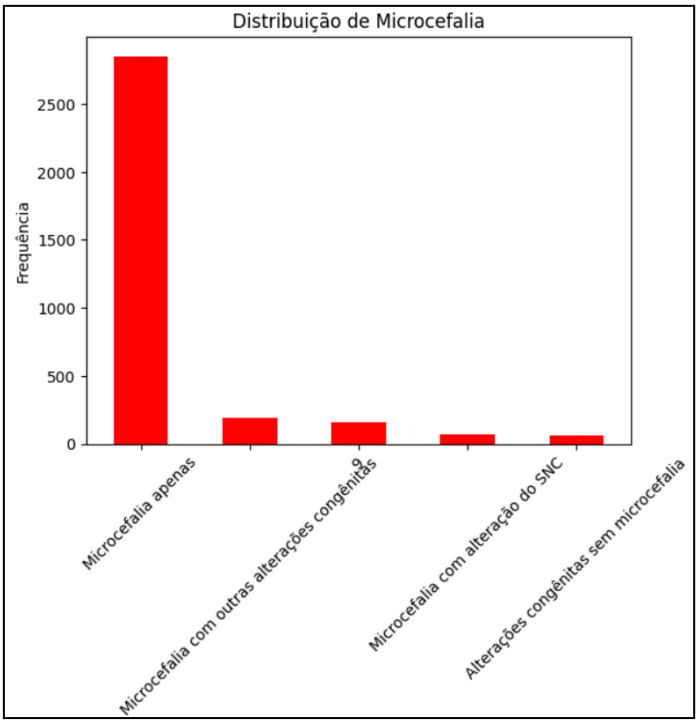
A situação final da notificação de SCZ (confirmado, descartado, inconclusivo etc.).
Descartados representam a grande maioria dos casos (~2400).
Confirmados aparecem em número muito menor (~500), seguidos por excluídos e inconclusivos, com proporções próximas.
Em investigação são poucos, provavelmente casos mais recentes ou sem desfecho confirmado até a data do registro.
A epidemia pode ter levado a um **grande número de notificações preventivas, mas poucos casos confirmados**.
Isso reflete o cenário epidemiológico do pós-2015, quando o **sistema de vigilância passou a registrar com cautela todos os casos suspeitos**, mesmo sem confirmação laboratorial.

15.



Os tipos de gestação (única, dupla, tripla, ignorada).
A grande maioria das gestações é única (~3000 casos), o que é esperado, gestações múltiplas são raras.
Há poucos registros de duplas e triplas, o que está dentro da normalidade biológica.

16.

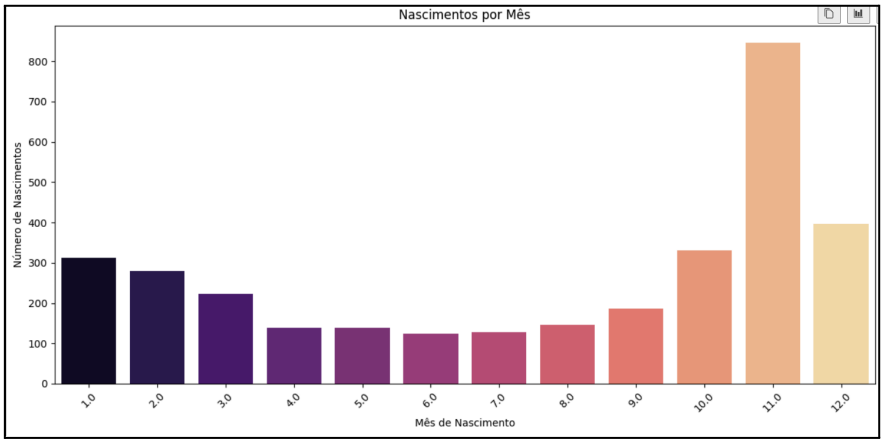


O gráfico evidencia que **a síndrome congênita associada ao Zika em Pernambuco teve como principal manifestação clínica a microcefalia isolada**, enquanto formas associadas a outras alterações neurológicas ou congênitas foram bem menos frequentes. Esse padrão é coerente com o perfil clínico predominante durante o surto de 2015–2016..

Etapa 4 - Análise Temporal

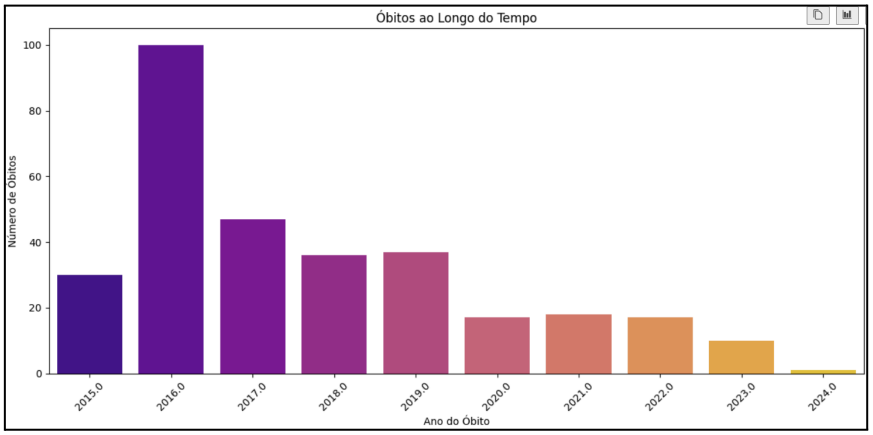
O próximo passo é composto pela conversão de valores temporais (Data de Notificação, Data de Nascimento, Data de óbito) para valores mensais e anuais para uma verificação da evolução temporal dos dados.

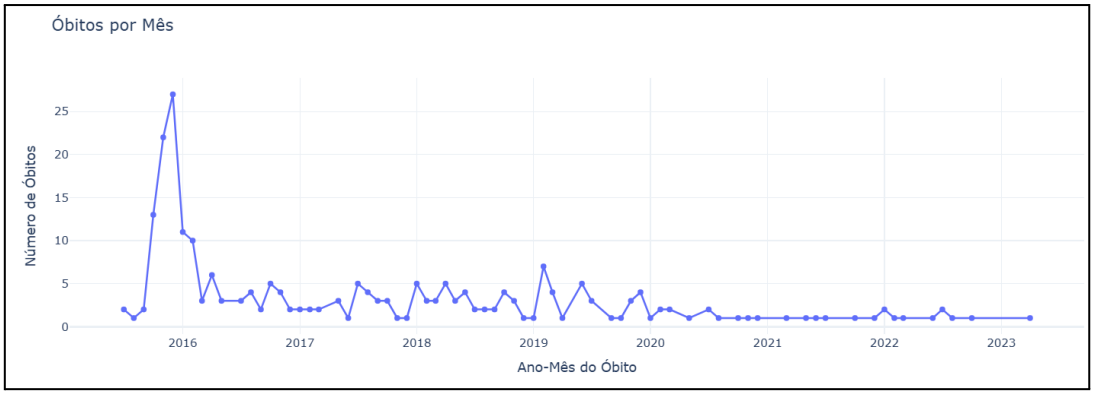
17.



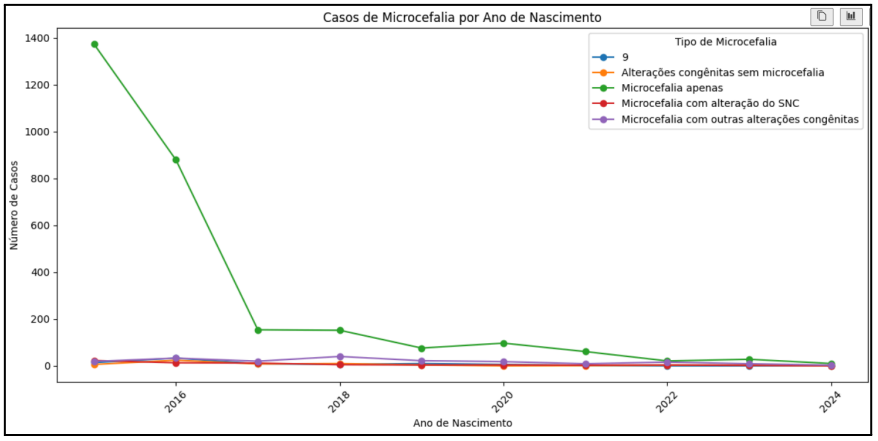
O mês de novembro (11) se destaca fortemente, com um pico acentuado de nascimentos seguido por dezembro.
Entre abril e agosto, há um vale pronunciado, com menor número de registros.
Esse padrão sazonal pode refletir tanto a variação natural da fecundidade quanto picos de notificações associadas ao surto de Zika.
O aumento no fim do ano (novembro/dezembro) coincide com o período em que os casos de microcefalia começaram a ser detectados com mais frequência em 2015, sugerindo impacto direto da epidemia sobre os nascimentos.
A queda nos meses intermediários possivelmente reflete menor número de gestações iniciadas durante o auge do surto.

18.





	Nascimentos	Óbitos
2015	1435	30
2016	982	100
2017	203	47
2018	213	36
2019	116	37
2020	125	17
2021	76	18
2022	44	17
2023	43	10
2024	14	1



Realizando uma separação temporal das tendências que a tabela e os 2 gráficos apresentados na aba 18 temos que:

2015–2016: pico epidêmico
O número de nascimentos disparou em 2015, acompanhado por aumento expressivo de óbitos em 2016.
Esse padrão reflete o ápice da epidemia de Zika, quando houve o maior registro de microcefalia e de mortalidade neonatal.

2017–2019: estabilização e queda
Queda drástica nos nascimentos notificados e redução progressiva dos óbitos.

Isso sugere o **controle gradual da epidemia**, associado ao aumento da imunidade populacional e às medidas preventivas.

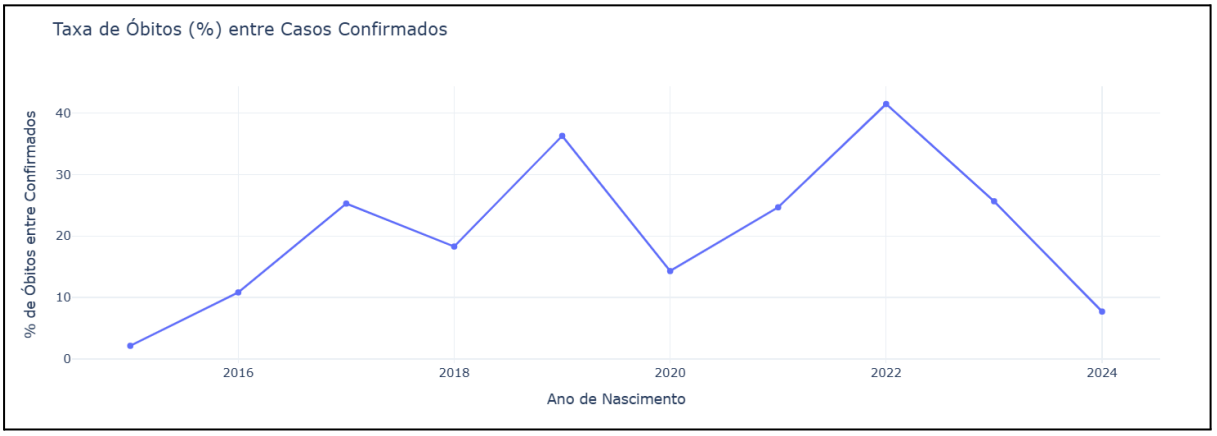
2020–2024: fase endêmica / residual
Casos tornam-se esporádicos, com números baixos e estáveis.
A taxa de óbitos caiu proporcionalmente, indicando **melhor manejo clínico dos casos** e provável subnotificação residual.

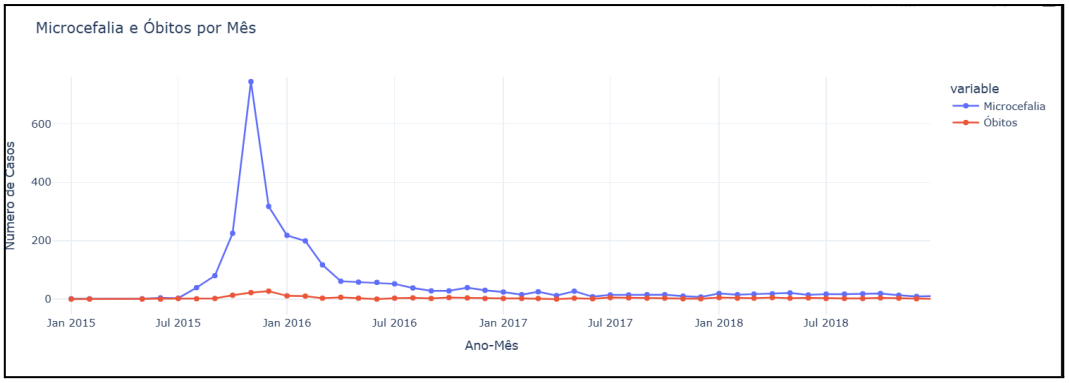
OBS: A partir de 2017, os valores se mantêm baixos e dispersos, com pequenas flutuações ocasionais, refletindo possivelmente casos residuais ou esporádicos detectados após o surto. Possivelmente é válido uma investigação de alguns pequenos picos em 2019- 2020 mas não necessariamente existe uma razão fixa.

Nos anos seguintes — especialmente de 2020 em diante — o gráfico mostra quase ausência de óbitos, indicando o controle epidemiológico da síndrome e a redução drástica da circulação do vírus Zika.

20.

Ano de Nascimento	Classificação Final: Confirmado	Descartado	Em investigação	Excluído	Inconclusivo
2015	352	967	0	63	53
2016	91	778	0	68	45
2017	7	154	1	17	24
2018	4	160	0	20	29
2019	3	95	0	7	11
2020	0	102	2	11	10
2021	0	57	0	8	11
2022	0	25	5	9	5
2023	0	26	15	2	0
2024	0	2	11	1	0





As duas curvas (azul = microcefalia; vermelho = óbitos) seguem o mesmo padrão temporal, com picos quase sobrepostos entre final de 2015 e início de 2016.
O número de casos de microcefalia chega a mais de 700 notificações mensais no auge, enquanto os óbitos acompanham o aumento, mas em menor magnitude.
Após 2017, ambas as curvas se mantêm praticamente em linha de base, sem novos picos.

Essa correspondência temporal, de maneira esperada, indica forte relação causal entre o surto de Zika e o aumento súbito de microcefalias e óbitos neonatais.
A defasagem leve (microcefalia primeiro, óbitos logo depois) reflete a progressão natural do desfecho clínico, bebês diagnosticados com microcefalia apresentando maior risco de morte nos meses subsequentes.
O desaparecimento simultâneo dos dois fenômenos reforça que a intervenção em saúde pública (controle do mosquito, vigilância, diagnóstico) foi efetiva.

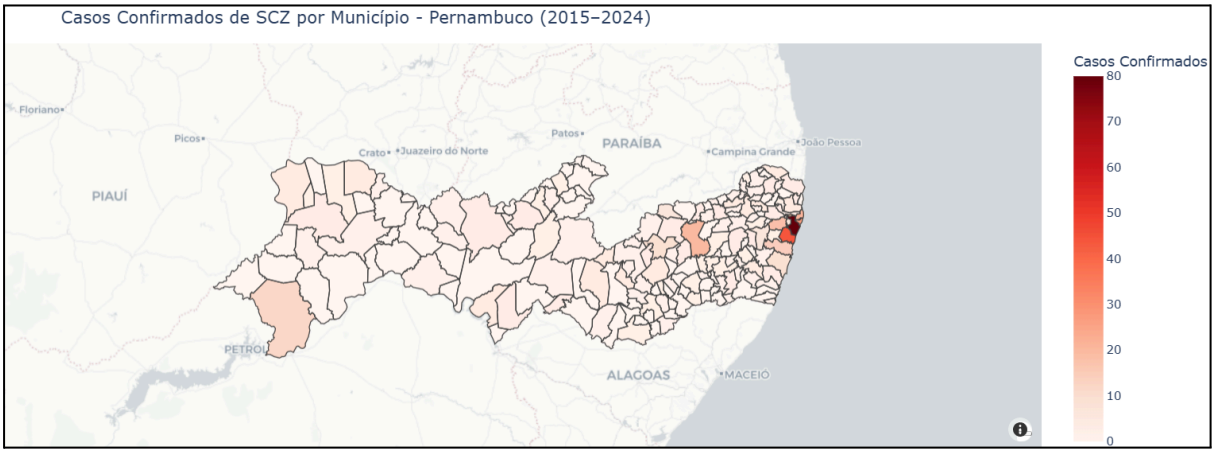
Etapa 4 - Distribuição Espacial de Casos

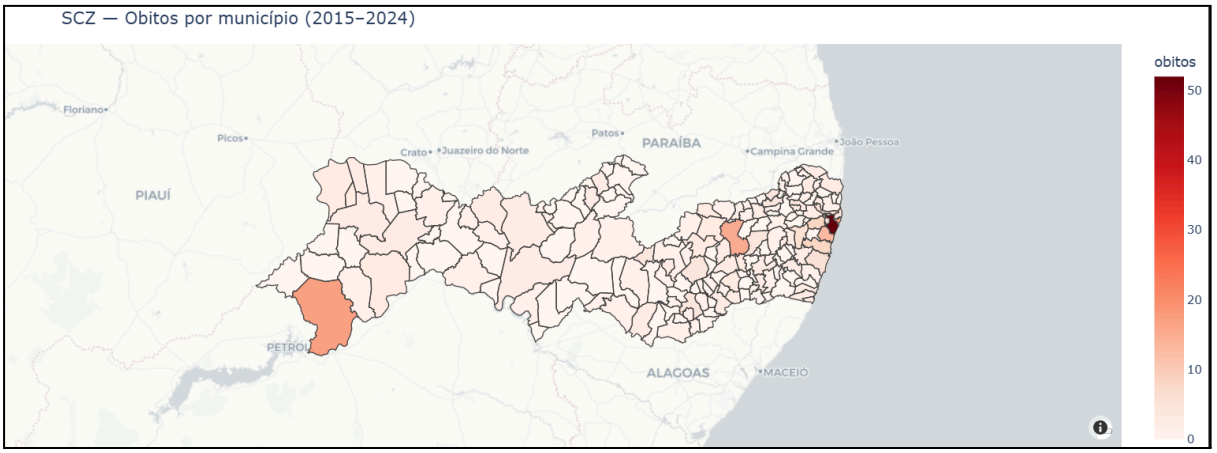
Para compreender onde a SCZ se concentrou em Pernambuco, integrou-se a base tabular às geometrias municipais do estado (malha IBGE/SEPLAG 2024). A chave espacial foi padronizada para 6 dígitos (CD_MUN6), garantindo o casamento entre o shapefile e a coluna “Código Município de Residência” do dataset. A partir dessa união produziu-se um GeoDataFrame único com a geometria de cada município e um conjunto de métricas epidemiológicas agregadas para 2015–2024:

- notificados: número total de registros;
- confirmados / descartados;
- óbitos;
- microcefalia (casos com qualquer forma de microcefalia);
- proporção de confirmados (%) = confirmados / notificados × 100;
- taxa de óbito entre confirmados (%) = óbitos / confirmados × 100 (com proteção contra divisão por zero).

Com esse GeoDataFrame, gerou-se diferentes visualizações (interativas no notebook):

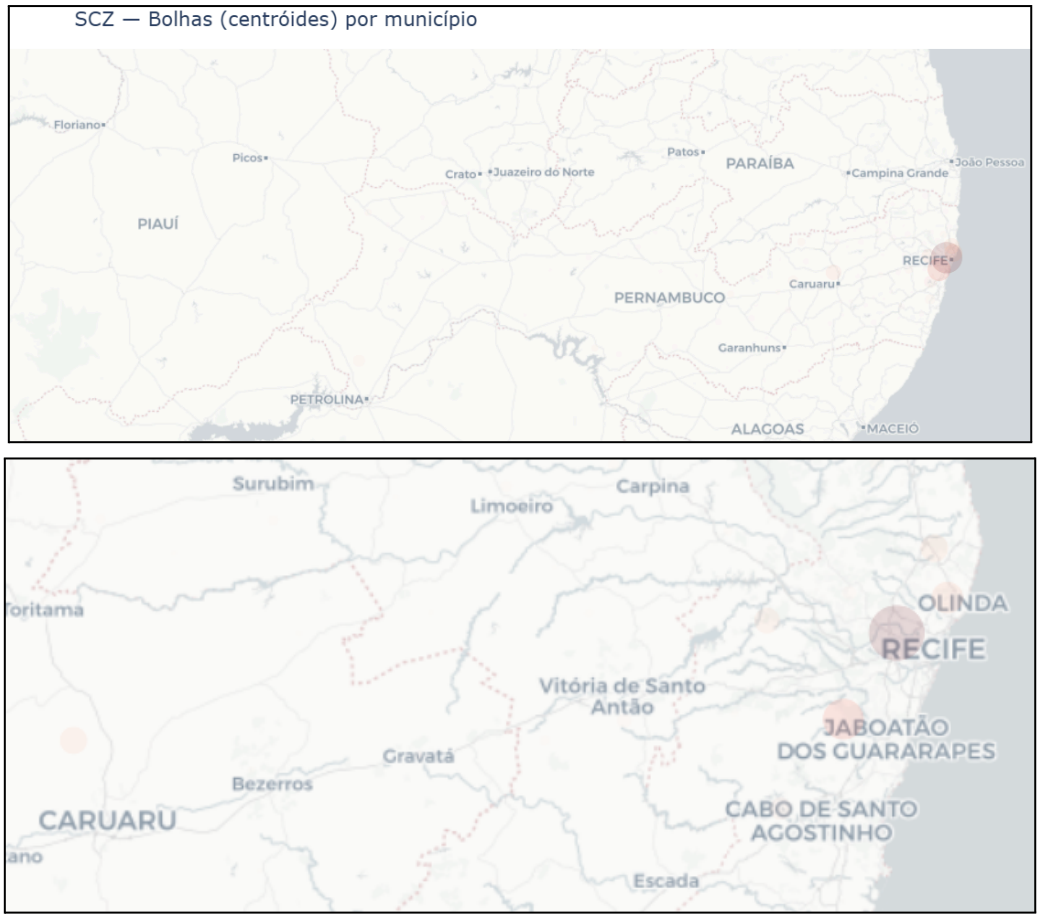
(1) Mapas coropléticos interativos
Com Plotly Mapbox para preencher cada município conforme a intensidade da variável escolhida (p. ex., “confirmados”, “óbitos”, “microcefalia”). O tooltip exibe o nome do município e indicadores, o que facilita comparar rapidamente áreas com carga maior de doença. Em paralelo, um Folium estilizado foi preparado para navegação leve (cores graduais, destaque ao passar o mouse e rótulos percentuais formatados).





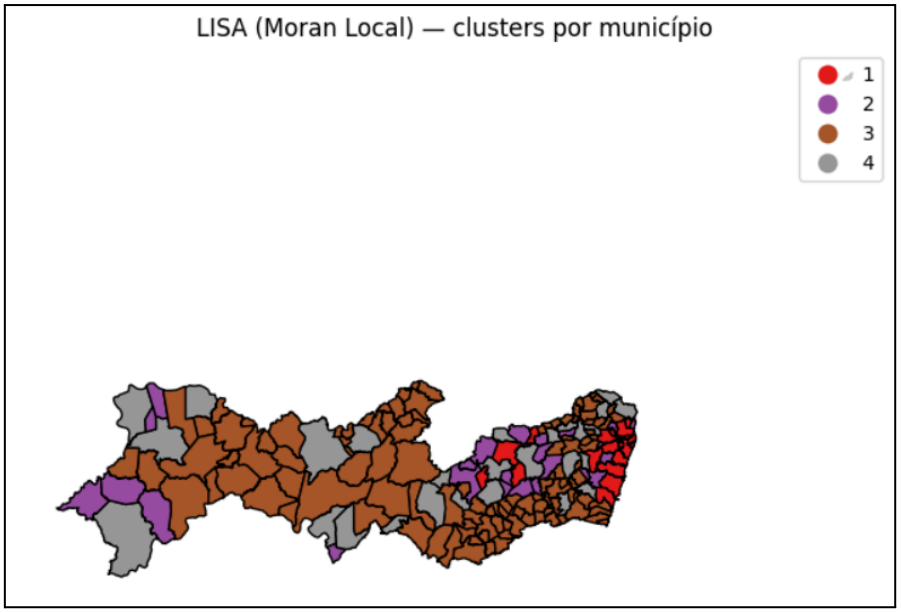
(2) Mapa de bolhas

Para enfatizar “quantidade” sem depender do preenchimento da área, foi plotado centróides municipais com o tamanho proporcional aos confirmados. Isso evidencia polos regionais mesmo quando municípios vizinhos têm áreas muito distintas. (Os centróides foram calculados em CRS projetado para evitar distorções.)



(Melhor visualização interativamente).

(3) Estatística espacial (autocorrelação)
Para além da leitura visual, avaliou-se a presença de dependência espacial com Moran's I global e, localmente, com LISA (Moran Local) e Getis-Ord Gi*. O resultado aponta **autocorrelação positiva** e estatisticamente significativa, isto é, **municípios tendem a se parecer com seus vizinhos em termos de número de casos**.



O que os mapas mostram

Hotspots na faixa litorânea e Zona da Mata (leste do estado).
A Região Metropolitana do Recife (RMR) e municípios da Zona da Mata aparecem de forma consistente como **áreas de alta–alta (High–High) no LISA e com tons mais intensos nos coropléticos** (tanto para confirmados quanto para óbitos). Isso é coerente com o período epidêmico de 2015–2016: **maior densidade populacional, maior circulação do vetor e melhor capacidade diagnóstica/notificadora favorecem a detecção e, infelizmente, a carga de doenças e desfechos graves**.

Coldspots no Sertão e extremo oeste.
Predominam padrões baixa–baixa (Low–Low), indicando menor incidência persistente e fraco contágio espacial. Parte desse comportamento pode refletir características demográficas (dispersão populacional), ambientais (menor favorabilidade ao vetor) e de acesso/fluxo assistencial.

Outliers e áreas de transição no Agreste/Sertão Central.
Alguns municípios aparecem como High–Low (valores altos cercados de vizinhos baixos) ou Low–High (valores baixos cercados de vizinhos altos). Esses pontos chamam atenção para **investigações locais**: surtos pontuais, subnotificação em vizinhos, rotas de atendimento que concentram diagnóstico no polo regional, entre outras hipóteses.
Óbitos acompanham o gradiente dos casos.
O padrão espacial dos óbitos espelha, em grande parte, o dos casos confirmados: maior concentração na RMR e municípios próximos e ocorrências relevantes em polos do interior (ex.: Petrolina). Isso pode indicar tanto maior gravidade onde há mais casos quanto melhor capacidade de registro do desfecho (linhas de cuidado e hospitais de referência).

Em suma, a análise espacial confirma um núcleo de alta concentração de SCZ na RMR e Zona da Mata, principalmente no auge epidêmico de 2015–2016, com gradiente decrescente em direção ao interior e pontos isolados sugerindo surtos locais ou efeitos do sistema de notificação. Essa leitura sustenta decisões de vigilância e pesquisa operacional (priorização de áreas, alocação de recursos e investigações de campo) e prepara o terreno para a etapa seguinte do projeto modelagem espacial onde incorporamos formalmente essa dependência entre vizinhos para melhorar a explicação e previsão do fenômeno.

Etapa 5 - Modelagem de Regressão Espacial

Para iniciar a etapa final é necessário primeiramente compreender cada modelo utilizado no geodataframe e o motivo para o uso de tal. Foi definido a variável “confirmados” como a **dependente** (Valor de y). As variáveis explicativas foram:

- 'notificados'
- 'prop_confirmados_%'
- 'taxa_obito_confirmados_%'

OBS: Para modelos de regressão espacial (cujo foco é análise e entendimento e não previsão) não é necessário dividir o dataset em treino e teste, visto que não há necessidade de acerto de nada.

Seguem os modelos utilizados:

1.Regressão Linear Clássica (OLS) - (Extra)

O modelo Ordinary Least Squares (OLS) é o ponto de partida para avaliar as relações entre uma variável dependente e um conjunto de variáveis explicativas. Ele **assume independência espacial** ou seja, considera que cada município é uma **observação isolada**, sem influência dos seus vizinhos.

O Modelo tenta responder “Quais fatores ajudam a explicar o número de casos confirmados de SCZ nos municípios de Pernambuco?”

A equação geral é:

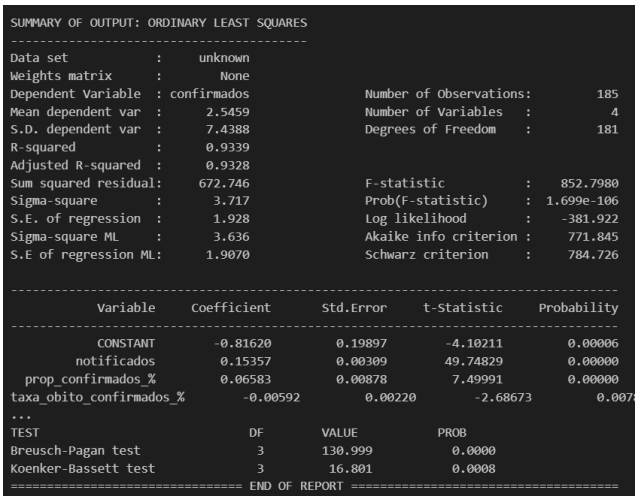
$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

onde:

- y = número de casos confirmados;
- X1 = número de notificações;
- X2 = proporção de confirmados (%);
- X3 = taxa de óbitos entre confirmados (%);

- ε = erro aleatório (suposto independente).

Resultados obtidos:



```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :     unknown
Weights matrix      :        None
Dependent Variable  : confirmados                Number of Observations:         185
Mean dependent var  :      2.5459                Number of Variables   :           4
S.D. dependent var  :      7.4388                Degrees of Freedom    :         181
R-squared           :      0.9339
Adjusted R-squared  :      0.9328
Sum squared residual:     672.746                F-statistic           :    852.7980
Sigma-square        :       3.717                Prob(F-statistic)     :  1.699e-106
S.E. of regression  :       1.928                Log likelihood        :    -381.922
Sigma-square ML     :       3.636                Akaike info criterion :     771.845
S.E of regression ML:      1.9070                Schwarz criterion     :     784.726

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT        -0.81620         0.19897        -4.10211         0.00006
         notificados         0.15357         0.00309        49.74829         0.00000
  prop_confirmados_%         0.06583         0.00878         7.49991         0.00000
taxa_obito_confirmados_%        -0.00592         0.00220        -2.68673         0.007
...
TEST                             DF        VALUE           PROB
Breusch-Pagan test                3        130.999          0.0000
Koenker-Bassett test              3         16.801          0.0008
================================ END OF REPORT =====================================
```

- R^2 = 0,9339 → o modelo **explica cerca de 93% da variação no número de casos** confirmados.

- Notificados (+) e Proporção de confirmados (+) apresentaram forte associação positiva, o que é **esperado**: quanto mais notificações e melhor capacidade de confirmação, mais casos identificados.

- Taxa de óbitos (-) teve efeito negativo, sugerindo que **locais com maior letalidade podem registrar menos confirmações** — possivelmente por subnotificação ou gravidade dos casos.

- Breusch–Pagan e Koenker–Bassett ($p < 0,001$) indicaram heterocedasticidade, ou seja, os resíduos **não têm variância constante** entre os municípios.

A análise dos resíduos mostrou padrão espacial persistente — **o Moran's I** foi **-0,03**, o que **indica leve autocorrelação espacial** (ainda que pequena), violando o pressuposto de independência do modelo OLS.
Essa constatação é útil para explicar o motivo de usar uma modelagem espacial, mesmo que, não tenha sido requisitado, é interessante comparar com a regressão linear não espacial.

2. Modelo Spatial Lag (SAR)

O **Spatial Lag Model (SAR)** introduz um termo espacial que considera que o valor de y em um município **é influenciado pelos valores de y em seus vizinhos**.
Em outras palavras, ele incorpora o **efeito de difusão espacial.**

forma geral:

$$y = \rho W y + X\beta + \varepsilon$$

onde:

- Wy representa a média ponderada dos valores de y nos municípios vizinhos (definidos pela matriz de pesos espaciais W);
- ρ é o **coeficiente de autocorrelação espacial** (mede o quanto o valor de um município depende dos vizinhos);
- Xβ são os efeitos das variáveis explicativas tradicionais.

Resultados:

```
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  : confirmados               Number of Observations:         185
Mean dependent var  :      2.5459               Number of Variables   :           5
S.D. dependent var  :      7.4388               Degrees of Freedom    :         180
Pseudo R-squared    :      0.9532
Spatial Pseudo R-squared:  0.9554
Log likelihood      :   -350.9453
Sigma-square ML     :      2.5773               Akaike info criterion :     711.891
S.E of regression   :      1.6054               Schwarz criterion     :     727.992

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT        -1.12895         0.17090        -6.60584         0.00000
          notificados         0.13991         0.00298        46.89125         0.00000
  prop_confirmados_%          0.05967         0.00734         8.13206         0.00000
taxa_obito_confirmados_%        -0.00517         0.00184        -2.81473         0.00488
       W_confirmados          0.21362         0.02514         8.49743         0.00000
------------------------------------------------------------------------------------
...
          notificados          0.1399          0.0380          0.1779
  prop_confirmados_%           0.0597          0.0162          0.0759
taxa_obito_confirmados_%         -0.0052         -0.0014         -0.0066
================================ END OF REPORT =====================================
```

- Pseudo R^2 = 0,9532, superior ao OLS (93% → 95%), mostrando melhor ajuste.
- AIC = 711,9 (menor que 771,8 do OLS) confirma ganho de qualidade.
- Coeficiente espacial ρ = 0,21 ($p < 0,001$) → presença de dependência espacial positiva: **municípios com altos números de casos tendem a estar próximos de outros com valores igualmente altos**.

- Todas as variáveis explicativas permaneceram significativas e com **sinais coerentes** com o OLS.

O SAR captura o contágio territorial ou difusão espacial da SCZ — ou seja, o risco não é isolado, mas **compartilhado entre áreas vizinhas**. Esse tipo de padrão é comum em **doenças transmitidas por vetores (Como é o caso da Zika)**, em que mobilidade, condições ambientais e fronteiras administrativas porosas facilitam a propagação.
3.Modelo Spatial Error (SEM)

O Spatial Error Model (SEM) parte de uma lógica diferente, não assume que o número de casos de um município depende dos vizinhos, mas que os **fatores não observados** (os erros) são espacialmente correlacionados.
Ou seja, há características regionais (socioeconômicas, ambientais ou estruturais) que afetam grupos de municípios próximos de forma similar, mas não foram medidas diretamente.

$$y = X\beta + \varepsilon, \quad \varepsilon = \lambda W\varepsilon + \xi$$

onde:

- λ é o **coeficiente de autocorrelação nos erros**;
- Wε representa a influência dos erros dos vizinhos;
- ξ é o ruído aleatório.

Resultados obtidos:

```
SUMMARY OF OUTPUT: ML SPATIAL ERROR (METHOD = full)
---------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  : confirmados                Number of Observations:         185
Mean dependent var  :      2.5459                Number of Variables   :           4
S.D. dependent var  :      7.4388                Degrees of Freedom    :         181
Pseudo R-squared    :      0.9339
Log likelihood      :   -380.9211
Sigma-square ML     :      3.5313                Akaike info criterion :     769.842
S.E of regression   :      1.8792                Schwarz criterion     :     782.724

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT        -1.04351         0.17012        -6.13400         0.00000
          notificados         0.16086         0.00269        59.90593         0.00000
  prop_confirmados_%          0.06923         0.00837         8.27577         0.00000
taxa_obito_confirmados_%        -0.00551         0.00217        -2.53493         0.01125
              lambda        -0.31657         0.12027        -2.63220         0.00848
------------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

- Pseudo R^2 = 0,9339, semelhante ao OLS, mas com correção de dependência nos resíduos.
- λ = -0,316 (p = 0,008) → indica que há padrões regionais inversos: áreas com erro positivo estão cercadas por áreas com erro negativo, refletindo heterogeneidade regional não capturada por XXX.

Apesar do ajuste levemente melhor nos resíduos, o desempenho global foi inferior ao SAR.

O SEM mostra que **existem fatores regionais não mensurados (como acesso à saúde, condições socioeconômicas ou intensidade de vigilância)** que geram correlação nos resíduos, mas **não há um padrão de difusão direta de casos entre municípios**.

Comparativo dos modelos

A comparação mostra que o modelo **Spatial Lag (SAR)** apresentou o melhor desempenho estatístico e teórico.
O coeficiente espacial positivo indica difusão territorial: municípios próximos tendem a compartilhar padrões de incidência semelhantes. Isso reforça o diagnóstico de que a SCZ não é aleatória nem localmente isolada, mas segue uma lógica regional de contágio, coerente com a biologia do vetor e a dinâmica populacional de Pernambuco.

Portanto, o modelo espacial permite capturar não apenas os efeitos das variáveis locais, mas também o efeito de propagação entre municípios, fornecendo uma base sólida para políticas de vigilância que considerem o território como rede interdependente, e não apenas como um conjunto de unidades administrativas independentes.

Métrica	OLS	Spatial Lag (SAR)	Spatial Error (SEM)
R^2 / Pseudo R^2	0,9339	**0,9532**	0,9339
AIC	771,8	**711,9**	769,8
Coeficiente espacial	N/A	**ρ = +0,21 (p<0,001)**	**λ = -0,32 (p=0,008)**
Interpretação	Relações locais	Difusão espacial ativa (vizinhos influenciam)	Dependência nos erros (fatores regionais não observados)
Melhor ajuste	Não	**Spatial Lag**	Não (Mas util)

Conclusão

Síntese dos achados

Foco temporal do surto. As séries mensais mostram pico agudo de notificações, microcefalia e óbitos entre o 2º semestre de 2015 e início de 2016, seguido de queda sustentada até níveis residuais a partir de 2018.

Padrão espacial **não aleatório**. Mapas coropléticos e análises LISA/Getis-Ord indicam hotspots na Região Metropolitana do Recife e Zona da Mata, com coldspots predominando no Sertão. Há outliers alto-baixo em áreas de transição do Agreste/Sertão Central.

Indicadores operacionais. Municípios com maior número de notificações apresentam, em média, mais casos confirmados e maior proporção de confirmações, sugerindo melhor capacidade de vigilância/diagnóstico nesses territórios.

Modelagem.

OLS explica ~93% da variância dos confirmados, mas apresenta heterocedasticidade e resíduos com dependência espacial. (Extra, não espacial)

Spatial Lag (SAR) melhora o ajuste (Pseudo $R^2 \approx 0,95$; AIC menor) e revela efeito de vizinhança positivo ($\rho \approx 0,21$; $p<0,001$) — **municípios com muitos casos tendem a estar cercados por municípios com muitos casos**.
Spatial Error (SEM) indica **dependência nos resíduos** ($\lambda<0$; $p<0,01$), mas desempenho global inferior ao SAR.
Em contexto de difusão territorial, o **SAR é mais aderente** ao fenômeno observado.

Implicações

A presença de clusters alto-alto reforça a necessidade de arranjos intermunicipais (RMR/Zona da Mata) para controle vetorial, comunicação de risco e coordenação de leitos/diagnóstico.

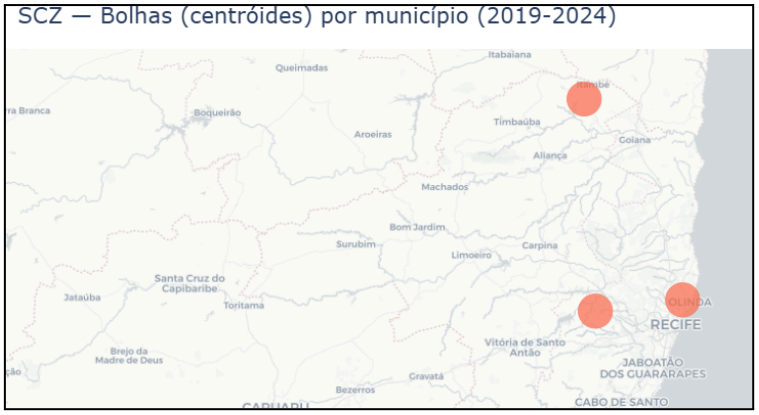
Capacitação e cobertura diagnóstica. O gradiente confirmados↔notificados sugere ganhos ao fortalecer a notificação ativa e o diagnóstico laboratorial em municípios com baixa captação (coldspots).

Monitoramento contínuo. Mesmo com declínio pós-2016, a manutenção de linhas de base e painéis trimestrais é prudente para detecção de reemergência ou micro-surtos locais.

Limitações

Completude desigual. Elevadas proporções de “não realizado/ignorado” em exames (STORCH/Zika, imagem) reduzem a potência de análises etiológicas finas.
Contagens, não taxas. Os mapas agregam contagens; a interpretação de risco exige padronização por população/nascidos vivos.

Agregação temporal ampla. A consolidação 2015–2024 é útil para visão histórica, porém mapas por janelas (2015–2016; 2017–2019; 2020–2024) podem esclarecer mudanças de padrão. Exemplo:



Variáveis omitidas. Fatores ambientais, socioeconômicos e de infraestrutura de saúde não foram incorporados; sua ausência **provavelmente explica parte da estrutura espacial capturada pelo erro** (SEM).

Sugestão:
Seria interessante o desenvolvimento de um dashboard interativo com filtros por ano, tipo de microcefalia e desfecho.

Conclusão definitiva

A SCZ em Pernambuco apresentou núcleo espacial persistente na RMR/Zona da Mata, com declínio acentuado após 2016 e dependência espacial positiva evidenciada pelos modelos e testes de autocorrelação. Em termos operacionais, os achados sustentam estratégias regionais coordenadas, manutenção de capacidade diagnóstica e monitoramento contínuo com indicadores padronizados. Do ponto de vista analítico, a incorporação de denominadores populacionais, covariáveis ambientais e modelos espaço-temporais constitui o caminho natural para consolidar um sistema de vigilância analítica mais sensível e preditivo.