

Universidade Federal de Minas Gerais

Departamento de Ciência da Computação

Trabalho Prático 2 - Notificações de Infectados de Dengue e Zika Introdução a Banco de Dados

Eduardo Birchal - 2024023970

Enzo de Souza Braz - 2024099062

Gabriel Guimarães dos Santos Ricardo - 2024024062

Gabriel Lucas Martins - 2023034900

João Pedro Moreira Smolinski - 2024023996

Github : https://github.com/guimaguima/tp2_ibd

Conteúdo

1	Introdução	2
2	Criação do Banco de Dados	2
2.1	Identificação de Valores e Limpeza dos dados	2
2.2	Modelagem do Banco de Dados	3
2.3	PostgreSQL	4
3	Dificuldades	4
3.1	Inconsistência e Incompletude	5
3.2	Volume de Dados	5
3.3	Referenciadores Errados	5
4	Análise exploratória	5
4.1	Mapas de Infectados	5
4.2	Casos anuais em cada município	10
4.3	Mortalidade	10
4.4	Relação entre Número de Infectados e Habitantes	11
5	Conclusão	13
6	Metadados	13
6.1	Metadados de extração	13
6.2	Dicionário Recriado	13

1 Introdução

Primeiramente, para a escolha dos dados, resolvemos abordar temas de envolvimento geográfico e social. Para o cumprimento da nossa ideia, escolhemos as bases de dados de notificações de casos de infecções de **Dengue** e **Zika** pelo Ministério da Saúde, entre os anos de 2023 e 2025. Dado este fato, verificamos que estas se correlacionavam, como esperado, aos Municípios, por meio do seu código do IBGE. Assim sendo, pegamos municípios e unidades federativas, que correlacionam entre si, oferecidos pelos códigos do IBGE, porém, para questões de análise exploratória mais profunda, correlacionamos estes dados com uma base disponibilizada gratuitamente e aberta a todos no github, a qual adicionava coordenadas nos municípios, tendo como base os dados do IBGE. Para as unidades federativas, para fins estatísticos, acrescentamos os dados populacionais estipulados pelo último do IBGE, de 2024.

As fontes das quais extraímos estão conectadas a seus respectivos hyperlinks (basta clicar no texto):

- Zika: Dados sobre Zika - <https://dados.gov.br/dados/conjuntos-dados/arboviroses-zika-virus>
- Dengue: Dados sobre Dengue - <https://dados.gov.br/dados/conjuntos-dados/arboviroses-dengue>
- Municípios: Dados por Município e UF - <https://www.ibge.gov.br/explica/codigos-dos-municipios.php>
- Base com coordenadas: Dados de Municípios e UF - <https://github.com/kelvins/municipios-brasileiro>
- População estimada por estado: Estimativa IBGE 2024 -
https://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2024/estimativa_dou_2024.pdf

Para a parte dos códigos necessários para a execução/correção dos dados, além de organização do banco de dados e os dados utilizados, colocamos eles no seguinte repositório: https://github.com/guimaguima/tp2_ibd. Nem todos os nossos dados conseguiram ser postados no github, entretanto, no READ.ME provido por nós, está o link do download dos dados restantes necessários.

Em termos de implementações, tínhamos ideias de correlacionar estes fatos com dados de UBS (Unidades Básicas de Saúde) e mortalidade, porém, por razões de dificuldades técnicas, resolvemos retirar esta abordagem, decisão a ser aprofundada na próxima seção. Dessa maneira, fizemos nossas análises nas tabelas supracitadas. Ademais, para detalhes de nossa implementação, para colocar os dados no banco, fazer a limpeza e organizar as visualizações, organizamos por meio de scripts em Python e Jupyter Notebooks, a fim de facilitar as operações necessárias. Por fim, organizamos todos os dados em tabelas usando o gerenciador de banco de dados PostgreSQL.

Para o restante do documento, dividimos em 5 partes: Criação do Banco de Dados, Dificuldades, Análise Exploratória, Conclusão e Metadados. Os metadados estão no final devido a seus grandes tamanhos.

2 Criação do Banco de Dados

2.1 Identificação de Valores e Limpeza dos dados

Para a limpeza das tabelas de Zika e Dengue, foram desenvolvidos scripts em Python que analisaram os tipos de dados e a frequência de valores ausentes em cada coluna. Com base nesses levantamentos e nos dicionários de dados oficiais provindos do Ministério da Saúde, definiu-se o conjunto de colunas a ser mantido, e foram excluídas as linhas que não atendiam às restrições de nulidade nas colunas pré-definidas. Algumas das colunas remanescentes foram renomeadas para aumentar a clareza e manter a consistência com as demais tabelas do banco de dados. Durante todo o processo, foram gerados logs de auditoria detalhando os renomeamentos realizados, a quantidade de registros processados e a qualidade dos dados resultantes.

Vale anotar que também fizemos o mesmo processo para dados de mortalidade do IBGE, mas por uma falta de coerência dos dados, seja pela falta de localização ou pela causa da morte, nós optamos por apenas usar os dados de zika e dengue do Ministério da Saúde, pois os mesmos eram melhores nesse sentido. O mesmo vale para UBS (Unidades Básicas de Saúde), que por mais que existam e se correlacionem aos dados de municípios, possuem muitos conflitos em relação aos dados das notificações de infectados das doenças explicitadas, pois nas notificações havia muitas UBS que na base não estavam registradas.

Com isto em mente, observamos detalhes muito importantes, como o fato de que as tabelas de Zika e Dengue tinham correspondência na maioria dos atributos, com a Dengue possuindo atributos que Zika não tinha, mas Zika sendo totalmente inclusa em Dengue, exceto por variáveis de sistema que resolvemos excluir. Sendo assim, resolvemos unir Zika e Dengue em uma única tabela, a qual marcaria o tipo de cada uma, processo relevante para o posterior diagrama e entidades relacionais observadas por nós. Além disso, retiramos das tabelas as seguintes colunas:

- **Dengue-** hospitaliz, dt_interna, municipio, cs_flgret, flxrecebi, migrado_w.
- **Zika-** in_vincula.
- **Comum a ambos-** id_agrav, sem_not, nu_ano, sg_uf_not, id_region, id_unidade, sem_pr, nu_idade_n, sg_uf, id_rg_resi, id_pais, id_ocupa_n, coufinf, copaisinf, tp_sistema

Vale ressaltar que tais colunas foram excluídas das tabelas pois poderiam representar redundância (como informações sobre unidades federativas quando há municípios), atributos do sistema de notificação (se o dado é migrado de outro sistema, por exemplo), ou atributos que ferem alguma das três primeiras formas normais e que foram julgados como não cruciais para o banco.

Além disso, foram retiradas da base de municípios as colunas siafi_id e ddd, julgadas como não relevantes para o objetivo do banco.

2.2 Modelagem do Banco de Dados

Após a limpeza das tabelas e filtragem de atributos relevantes, optou-se por montar um diagrama de Entidade-Relacionamento, para que a modelagem do sistema ficasse mais fácil posteriormente. O resultado pode ser conferido abaixo, na Figura 3

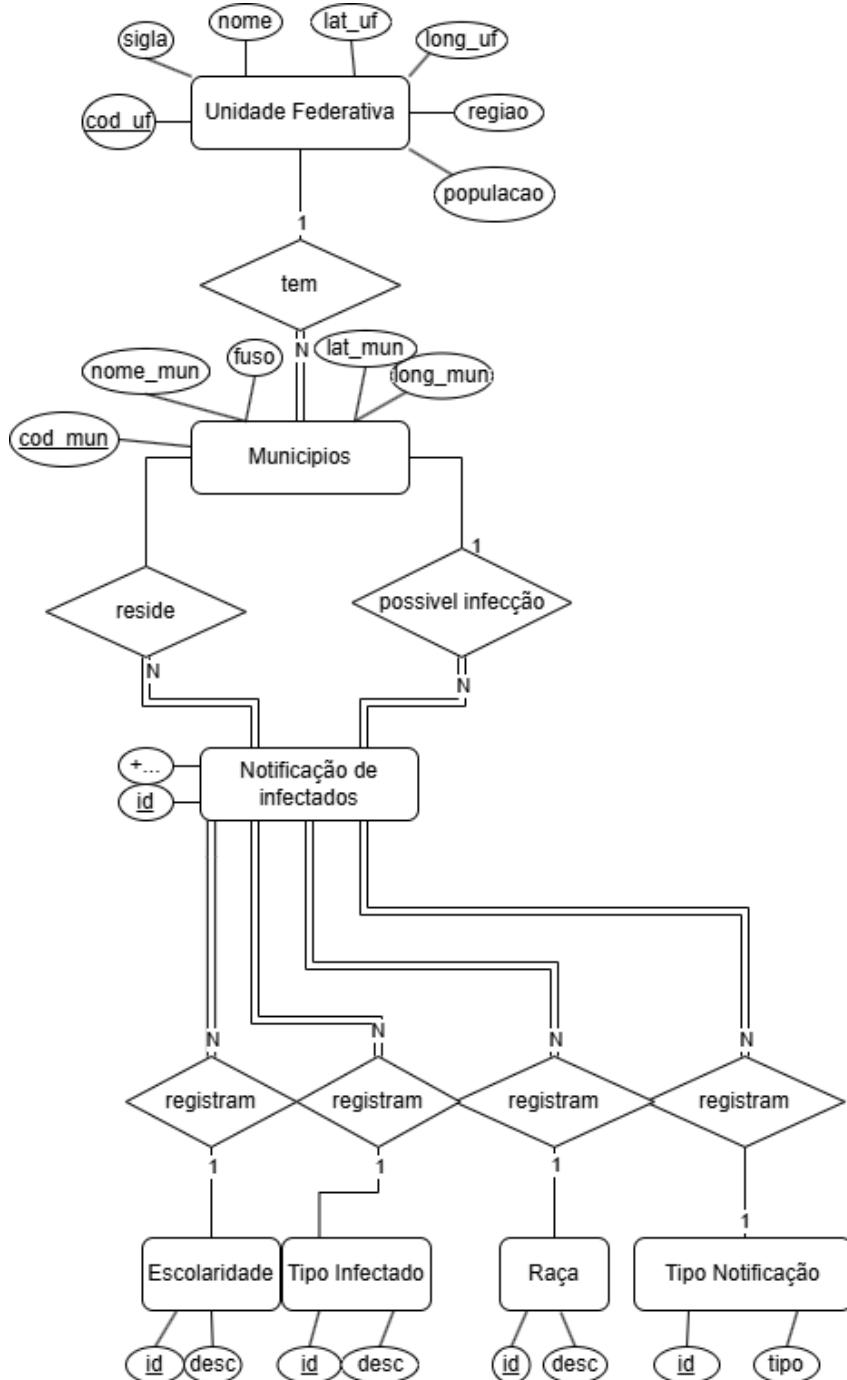


Figura 1: Modelo de Entidade-Relacionamento do Banco

Com base no ER, já é possível verificar os relacionamentos que cada entidade deve ter no fim, sendo que no caso de infectados, como são mais de 100 colunas, foram omitidas. Com base neste ERE descrito, organizamos utilizando o algoritmo de Elmasri Navathe, mapeamos para o ER pelos 8 passos, no caso do nosso ERE, foi utilizado o 1º e o 4º passo, os passos são visíveis nas imagens a seguir:

Entidades Regulares														
Unidade Federativa	cod_nf	sigla	nome	lat_nf	long_nf	regiao	populacao							
Municípios	cod_mun	nome_mun	fuso	lat_mun	long_mun	ddd								
Especialidade	id	desc												
Raça	id	desc												
Tipo Notificação	id	nome												
Tipo Infectado	id	desc												
Notificação de Infectados														
febre	ecido_pept	slim_abdom	slim_hemat	slim_vetar	slim_jiq	slim_pisq	slim_sang	slim_vom	artigia	artrite	auto_imune	cefaleia	clinc_chik	complicia
dt_chik_x1	dt_chik_x2	dt_grav	dt_m1	dt_pnt	dt_soro	dt_viral	epatite	evidencia	exantema	gengivio	grav_ast	grav_corsc	grav_ench	grav_exbre
grav_mic	grav_orgao	grav_pnto	grav_sang	grav_taqi	hematolog	hematura	hepatopat	hipertensa	histopat_n	imunoh_n	laco_n	leucopenia	man_hemor	metro
plasmatico	renal	res_chik1	res_chik2	resul_mst	resul_pnt									

Figura 2: Passo 1: Entidades Fortes

Entidades Regulares														
Unidade Federativa	cod_nf	sigla	nome	lat_nf	long_nf	regiao	populacao							
Municípios	cod_mun	nome_mun	fuso	lat_mun	long_mun	ddd	cod_nf							
Especialidade	id	desc												
Raça	id	desc												
Tipo Notificação	id	nome												
Tipo Infectado	id	desc												
Notificação de Infectados														
febre	ecido_pept	slim_abdom	slim_hemat	slim_vetar	slim_jiq	slim_pisq	slim_sang	slim_vom	artigia	artrite	auto_imune	cefaleia	clinc_chik	complicia
dt_chik_x1	dt_chik_x2	dt_grav	dt_m1	dt_pnt	dt_soro	dt_viral	epatite	evidencia	exantema	gengivio	grav_ast	grav_corsc	grav_ench	grav_exbre
grav_mic	grav_orgao	grav_pnto	grav_sang	grav_taqi	hematolog	hematura	hepatopat	hipertensa	histopat_n	imunoh_n	laco_n	leucopenia	man_hemor	metro
plasmatico	renal	res_chik1	res_chik2	resul_mst	resul_pnt									

Figura 3: Passo 4: Relacionamentos 1:N

Por fim, foi configurado o dicionário a partir do ER, tendo em vista as informações também disponibilizadas pelos links dos quais os dados proveram e o que nossa limpeza nos mostrou, assim sendo, fizemos a tabela de dicionário como no final do documento.

2.3 PostgreSQL

Nessa etapa, utilizando o modelo obtido anteriormente, foi desenvolvido um algoritmo para a criação do SQL a partir de um dicionário CSV. Esse processo envolveu três etapas principais: processamento do CSV para o script, geração de comandos SQL e execução no PostgreSQL para verificar erros. Inicialmente, o algoritmo lê o arquivo CSV, normaliza nomes de tabelas/colunas (convertendo para snake_case e removendo caracteres especiais) e mapeia tipos de dados para equivalentes do PostgreSQL (ex: NUMBER -> NUMERIC, VARCHAR mantido, etc). As principais inovações foram o tratamento inteligente de constraints seguindo o dicionário criado por nós, sendo: valores permitidos são convertidos em CHECK IN, relações FK são detectadas por expressões regulares, e constraints complexas como as condicionais tipo_infec são transformadas em expressões booleanas. A solução final gera um arquivo schema.sql executável e cria o banco.

Após isto, utilizamos de outro script, o qual possui funções específicas para a adição de cada linha dos diferentes csv, iterando sobre cada uma enquanto procura adicioná-la no banco, e caso não coincida com as chaves estrangeiras respectivas de cada caso e tabela, verifica os dados e faz um tratamento para tentar inserir, no caso de haver incompatibilidade de tipagem ou tamanho. Caso não seja possível executar a operação sem perda de informação ou caso referencie chaves estrangeiras que não existam no banco disponibilizado publicamente, esta linha deixa de ser adicionada. Assim garantindo uma adição segura e consistente ao banco, com inúmeros tratamentos de casos que sigam as restrições planejadas por nós para os bancos de dados.

3 Dificuldades

Diversas dificuldades surgiram durante a limpeza dos dados, modulação e criação do banco, a maioria delas herdada dos dados externos.

3.1 Inconsistência e Incompletude

Uma das principais dificuldades foi na tentativa de integrar dados de mortalidade e UBSs. Os dados de mortalidade eram recheados de colunas redundantes e praticamente inteiramente nulas. Enquanto isso, as UBSs estavam incompletas, visto o problema de que muitas notificações referenciavam UBSs inexistentes.

Entretanto, não foi somente nessas duas bases que foram encontrados problemas. Nas bases de notificações de dengue e zika havia muitos problemas em seus dicionários. Diversos atributos tinham seus tipos digitados errados, como VARCHARs no lugar de INTEGERs ou numerações erradas de identificadores, como CHAR(3) para siglas de dois caracteres. Colunas com mais de um tipo de dado também foram encontradas com certa frequência. Além disso, muitos nomes de atributos e restrições eram simplesmente ignorados, campos obrigatórios eram mantidos nulos em grande parte das tuplas e faltavam explicações para o que cada coluna significava de fato.

3.2 Volume de Dados

Outra questão a qual dificultou o andamento do projeto foi o grande volume de dados. Juntando todas as tabelas de notificações, havia mais de 4,6 milhões de tuplas de ocorrências das doenças. Isso acabou atrasando alguns processos de limpeza e construção de banco, já que iterar sobre cada uma das linhas era muito custoso. Além disso, o uso de ferramentas como o GitHub ficava limitado, pois alguns arquivos CSV que foram usados durante todo o processo eram maiores do que o limite das ferramentas permitia.

3.3 Referenciadores Errados

O último grande problema que surgiu foi no identificador das tuplas de municípios. Eram inteiros de 7 dígitos, mas a tabela de notificações os referenciava com apenas 6 dígitos. Foi necessário implementar uma maneira de corresponder apenas aos 6 primeiros dígitos, o que podia levar a algumas inconsistências se existissem dois municípios que se diferenciassem apenas no último dígito. A análise posterior para garantir que não houvesse conflitos acabou sendo bastante custosa e trabalhosa.

4 Análise exploratória

Com o banco preparado, resta fazer algumas análises sobre os dados adquiridos. Fizemos todas as análises utilizando SQL no banco do PostgreSQL, visível nos notebooks presentes em nosso repositório. Resolvemos abordar para analisar principalmente:

- **Mapas de Infectados:** Observar como se comportam a quantidade de infectados por ano ao longo do país, de modo que acreditamos ser razoavelmente distribuído.
- **Casos anuais em cada município:** Observar numericamente as infecções, onde desejamos observar relações com regiões.
- **Mortalidade:** Verificar a taxa de mortalidade e se há locais focos.
- **Relação entre Número de Infectado e Habitantes:** Se há alguma correlação ou parecem dispersos.

A seguir é possível conferir essas análises e suas conclusões.

4.1 Mapas de Infectados

Na nossa base observamos todos os dados de 3 principais anos, 2023 a 2025, sendo, obviamente, o ano de 2025 incompleto perante o restante. Para cada ano obtivemos diferentes valores de infectados, assim resolvemos observar como eles se configuraram no mapa e como estão suas distribuições, no sentido de encontrar outliers. É importante ressaltar que os anos de 2023 e 2024 possuem os dados completos, diferentemente de 2025.

Primeiramente, iremos nos ater aos casos de dengue nos anos supracitados. Neles, fomos capazes de observar todo o país, mapeando os municípios, de maneira que a quantidade de casos ocorridos seja destacada pelo tamanho da bolinha correspondente e pela cor, tal que quanto maior e mais quente a cor, maior a quantidade. Além disso, é marcado com a cor preta no mapa aquele que obteve ao longo do ano a maior quantidade de ocorrências dentro do país durante aquele ano. Por fim, se observa, utilizando um box plot, a presença de *outliers* da quantidade observada por municípios. A seguir os gráficos:

Mapa de Casos de Dengue em 2023

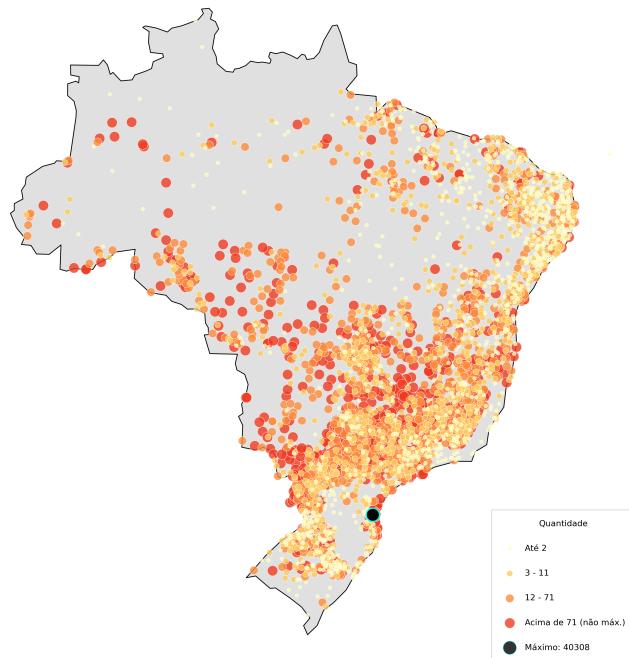


Figura 4: Mapa de Casos de Dengue em 2023.

Mapa de Casos de Dengue em 2024

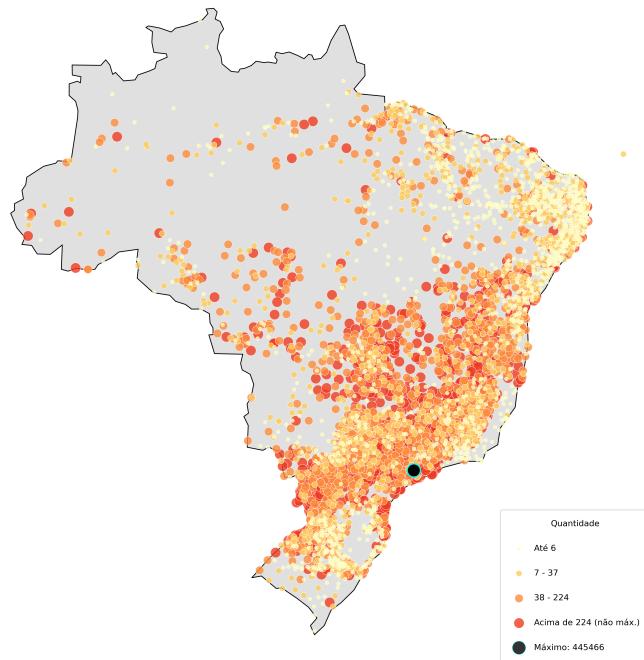


Figura 5: Mapa de Casos de Dengue em 2024.

Mapa de Casos de Dengue em 2025

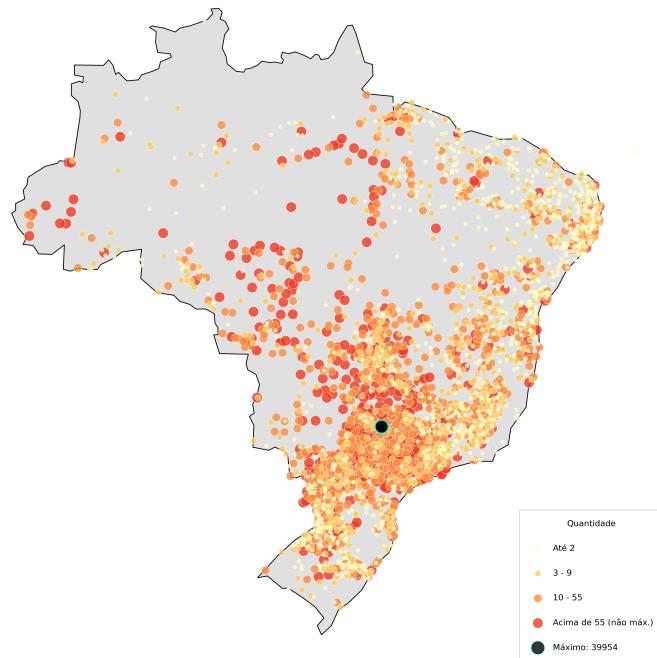


Figura 6: Mapa de Casos de Dengue em 2025.

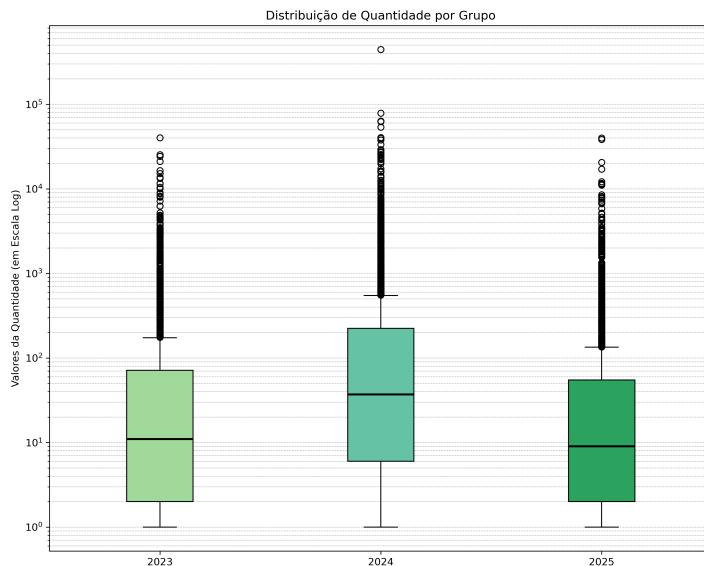


Figura 7: Análise de outliers pelo box plot - Dengue.

Nas figuras percebemos os casos de dengue, baseados nas suas distribuições no mapa, observamos uma concentração no sudeste e sul, de modo que a cidade que mais possui casos normalmente tende a ser destas duas regiões, sendo elas, por ordem de ano, Joinville, São Paulo e São José do Rio Preto. Além disso, com os casos notificados, observamos pequenas quantidades ao longo do restante do país. Por mais que isto aparente dizer que acontecem mais casos de dengue nas regiões Sul e Sudeste, acreditamos que isto é o "Viés do Sobrevivente", onde observamos os danos relatados mas sem observar a infraestrutura social e econômica presente no país, ou seja, acreditamos que isso ocorra pelo fato de outras áreas notificarem poucos casos ocorridos em relação ao total. Dessa maneira, podendo ter ou não relação com as regiões, é possível que os dados sejam insuficientes para pensarmos amplamente sobre o Brasil. Além disso, com a ajuda do box plot, confirmamos que a maior parte dos casos ocorre realmente com poucas infecções relatadas, variando muito pouco de ano a ano.

Para os mapas de Zika, a análise de organização dos mapas é igual a de dengue, sendo eles:

Mapa de Casos de Zika em 2023



Figura 8: Mapa de Casos de Zika em 2023.

Mapa de Casos de Zika em 2024



Figura 9: Mapa de Casos de Zika em 2024.

Mapa de Casos de Zika em 2025



Figura 10: Mapa de Casos de Zika em 2025.

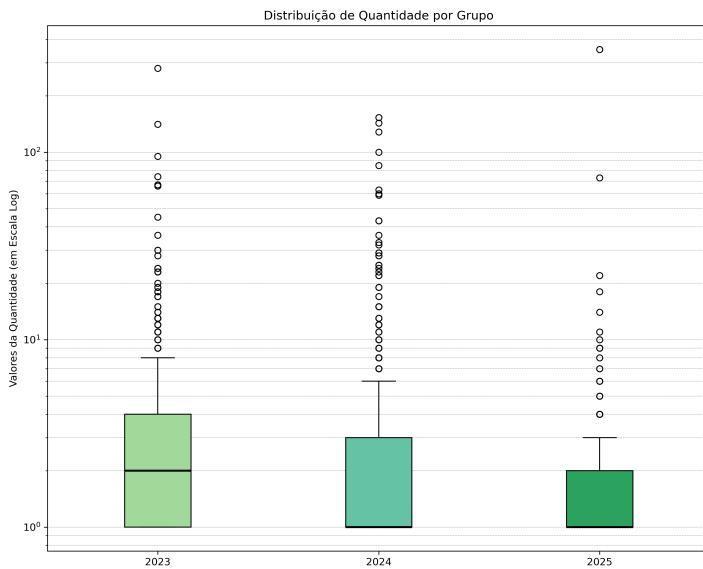


Figura 11: Análise de outliers pelo box plot - Zika.

Entretanto, mesmo mediante a semelhança, acreditamos que os dados revelam outras perspectivas, dado que as ocorrências de notificações de Zika já são pequenas em comparação às de dengue. Diferente da dengue, as notificações ocorrem, em sua maioria, fora do sul e sudeste, tal que as cidades que obtiveram mais casos notificados são, em ordem crescente de ano, Salvador, Ceará-Mirim e Pontes Lacerda. Tendo isto em mente, observamos que em sua maioria são cidades cuja população também é mais vulnerável em relação aos grandes centros urbanos, o que pode revelar como a doença da Zika funciona no Brasil. Mas, ainda assim, é importante levantar a hipótese da não notificação ao longo do território brasileiro, assim podendo aumentar o número de casos. Por fim, com o box plot, vemos que as variações realmente são menores no geral, sem uma grande presença de *outliers*.

4.2 Casos anuais em cada município

Com uma simples consulta SQL, é possível adquirir para cada município o total de casos de Dengue, Zika, ou ambos que foram notificados em um ano específico, ou até mesmo o total. Abaixo será possível ver a Tabela 1 e Tabela 2, que mostram casos anuais e no período inteiro do banco para cada estado e região respectivamente.

Estado	Casos 2025	Casos 2024	Casos 2023	Total
São Paulo	427636	1452291	221462	2101389
Minas Gerais	44185	682046	194702	920933
Paraná	45736	414658	103063	563457
Santa Catarina	4576	206522	89319	300417
Goiás	34442	205182	44425	284049
Rio Grande do Sul	13033	110388	23676	147097
Mato Grosso	13459	30046	19636	63141
Bahia	4172	63694	16513	84379
Mato Grosso do Sul	3159	8918	29322	41399
Pará	5649	11241	2822	19712
Alagoas	789	12068	3445	16302
Ceará	892	7471	6896	15259
Paraíba	1689	5650	4015	11354
Piauí	1409	5145	3474	10028
Rondônia	510	2879	6383	9772
Maranhão	1086	5484	3066	9636
Amazonas	2042	4441	2721	9204
Pernambuco	880	4043	950	5873
Acre	2185	1525	1655	5365
Rio Grande do Norte	316	2831	1116	4263
Sergipe	75	872	1760	2707
Distrito Federal	39	1721	122	1882
Rio de Janeiro	31	255	61	347
Espírito Santo	5	84	47	136
Tocantins	19	38	13	70
Amapá	1	13	4	18
Roraima	1	11	2	14

Tabela 1: Estados e Casos totais 2023-2025

Região	Casos 2025	Casos 2024	Casos 2023	Total
Sudeste	471857	2134676	416272	3022805
Sul	63345	731568	216058	1010971
Centro-Oeste	51099	245867	93505	390471
Nordeste	11308	107258	41235	159801
Norte	10407	20148	13600	44155

Tabela 2: Regiões e Casos totais 2023-2025

Uma informação que salta aos olhos é a baixa quantidade de notificações de infectados na Região Norte e Nordeste. Comparando com o índice de desenvolvimento humano (IDH), 0.683 e 0.659, e com o índice de Gini (que mede a desigualdade de renda), 0.536 e 0.517, das regiões, parece haver alguma contradição. Isso provoca uma reflexão sobre quantos casos de infecção das regiões simplesmente não são notificados ou até mesmo tratados corretamente. Além disso, aprofundando mais na região Nordeste, esta é a segunda região mais populosa do país, somente atrás do Sudeste, o que causa um grau ainda maior de estranheza ao visualizar a tabela.

4.3 Mortalidade

Uma análise que achamos importante realizar era a taxa de mortalidade de infectados, de modo que possamos observar em qual estado a taxa seria maior ou se a quantidade seria razoavelmente grande. Consideramos ambos

os casos, de dengue e zika, entretanto zika adiciona pouco perante o total, cerca de 12 mortes ao longo de todo o território nacional. Sendo assim, a taxa de mortalidade considerando os infectados com dengue ou zika é:

Estado	Total de Óbitos	Total de Infectados	Taxa de Mortalidade (%)
Espírito Santo	2	136	0.014706
Rio de Janeiro	3	347	0.008646
Sergipe	14	2701	0.000518
Maranhão	27	9636	0.002802
Bahia	169	84379	0.002002
Piauí	20	10028	0.001994
Amazonas	18	9204	0.001956
Rondônia	19	9742	0.001951
Rio Grande do Sul	279	147097	0.001897
Mato Grosso do Sul	74	41399	0.001787
Pará	33	19712	0.001674
Mato Grosso	105	63141	0.001663
Paraná	904	563457	0.001603
Distrito Federal	3	1882	0.001594
Pernambuco	9	5923	0.00152
Minas Gerais	1407	920933	0.001528
São Paulo	2803	2101397	0.001334
Goiás	358	284049	0.00126
Santa Catarina	369	300417	0.001228
Alagoas	19	16302	0.001166
Ceará	14	11254	0.001244
Paraíba	10	11554	0.000865
Rio Grande do Norte	3	4263	0.000704
Acre	2	5365	0.000373
Roraima	0	4	0.000000
Amapá	0	18	0.000000
Tocantins	0	70	0.000000

Tabela 3: Taxa de mortalidade percentual por estado

Observamos que as mortes relatadas pelas notificações são um valor ífimo em relação ao total notificado. Isto indicaria uma taxa muito baixa, que em certa medida não parece corresponder aos estipulados pelo Ministério da Saúde, que tendem a 1% no mínimo para ambos os casos. Possivelmente, uma coluna que marcasse a presença de óbito seria melhor do que uma que marcasse a data do óbito para o caso analisado. Concluindo, chegamos a conclusão de que os dados provindos das notificações são insuficientes para observar as mortalidades.

4.4 Relação entre Número de Infectados e Habitantes

Nesta subseção, analisaremos a relação entre o número de infectados em um estado e sua população geral. Essa métrica é boa para encontrar *outliers*, já que um número concreto total de infectados vai ter influência do número de habitantes de um lugar.

Estado	Infectados por 100.000 Habitantes
Santa Catarina	1173.65
Mato Grosso do Sul	1063.54
Minas Gerais	947.92
Paraná	900.56
Goiás	629.56
Mato Grosso	536.70
São Paulo	498.66
Rondônia	403.68
Rio Grande do Sul	217.55
Acre	199.39
Bahia	116.77
Alagoas	110.15
Piauí	106.20
Paraíba	101.01
Sergipe	79.64
Ceará	78.41
Amazonas	69.03
Maranhão	45.24
Pará	34.75
Rio Grande do Norte	33.79
Pernambuco	10.49
Distrito Federal	4.33
Espírito Santo	1.23
Tocantins	0.86
Amapá	0.55
Rio de Janeiro	0.38
Roraima	0.31

Tabela 4: Infectados por 100.000 habitantes em cada estado

Com a tabela de referência, é perceptível que, mesmo o estado de São Paulo dominando em todos os anos no número total de infectados, não aparece nem entre os 5 estados com maior número de infectados por habitante.

Outra informação relevante é o dado referente ao estado de Santa Catarina. O número de infectados por 100 mil habitantes parece muito alto, em uma escala diferente dos demais. Para verificar se tal valor se trata de um *outlier* ou não, vamos usar um gráfico do tipo box plot e ver se o ponto referente ao estado aparece distante dos percentis esperados.

O gráfico pode ser conferido abaixo, Figura 12.

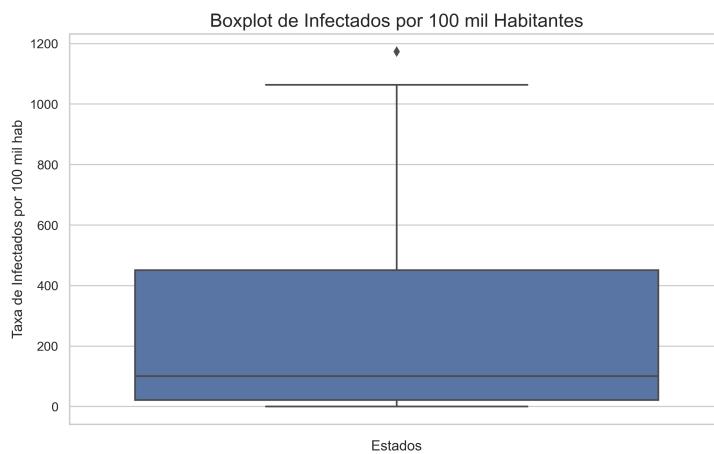


Figura 12: Boxplot de infectados por habitantes

Como é perceptível, há um losango acima do gráfico, representando um *outlier*, que nesse caso é exatamente o estado de Santa Catarina. Ou seja, é possível afirmar que Santa Catarina tem um número de casos de dengue e zika acima do esperado para um estado brasileiro.

5 Conclusão

Com base na análise dos dados de notificações de Dengue e Zika entre 2023 e 2025, o estudo realizado permitiu a criação de um banco de dados consolidado e a extração de *insights* relevantes sobre a distribuição e o impacto dessas arboviroses no Brasil.

A análise exploratória dos dados revelou uma concentração significativa de casos de Dengue nas regiões Sudeste e Sul do país. E, embora estado de São Paulo tenha registrado o maior número absoluto de infectados, a análise proporcional por 100 mil habitantes destacou Santa Catarina, sugerindo uma maior intensidade de transmissão nesses estados.

Outro achado notável foi a baixa quantidade de notificações nas regiões Norte e Nordeste. Considerando que o Nordeste é a segunda região mais populosa do país e que ambas as regiões possuem índices de desenvolvimento humano e de Gini que poderiam sugerir um cenário diferente, levanta-se a hipótese de uma subnotificação expressiva de casos.

Também foram enfrentados diversos desafios durante o trabalho, como a decisão de projeto de não utilizar os dados de mortalidade e UBSs devido a inconsistências em suas publicações, o que acabou limitando a profundidade de algumas análises inicialmente planejadas.

A unificação das bases de dados de Dengue e Zika também se provou eficiente, já que havia uma grande quantidade de atributos correspondentes. Além disso, a utilização do PostgreSQL para a gestão dos dados, juntamente a alguns scripts em Python para a limpeza, tratamento e inserção, garantiu a consistência e integridade do Banco de Dados final. Em suma, o projeto estimulou os alunos a procurarem mais sobre a implementação prática dos bancos de dados, além de aprender a fazer análises estatísticas baseadas neles.

6 Metadados

6.1 Metadados de extração

Para cada dado, de cada fonte diferente, obtemos os seguintes metadados após suas extrações e realizações de alterações por nós:

Tabela 5: Identificação e Origem dos Conjuntos de Dados

Nome do Conjunto	Fonte Principal	Órgão Responsável
Notificação de Infectados (Zika)	Dados.gov	Ministério da Saúde
Notificação de Infectados (Dengue)	Dados.gov	Ministério da Saúde
Municípios e UF	GitHub (kelvins)	Baseado no IBGE
População por UF (Estimativa)	FTP IBGE	IBGE

Tabela 6: Detalhes da Coleta e Frequência de Atualização

Nome do Conjunto	Data de Obtenção	Cobertura	Atualização
Notificação de Infectados (Zika) 2023-2025	31/05/2025	Nacional	A cada 2 semanas
Notificação de Infectados (Dengue) 2023-2025	31/05/2025	Nacional	A cada 2 semanas
Municípios e UF	03/06/2025	Nacional	Conforme o Censo
População por UF (Estimativa)	03/06/2025	Nacional	Anualmente

6.2 Dicionário Recriado

Tabela 7: Dicionário de Dados Completo

Relação	Atributo (Coluna)	Tipo	Nulo?	Único?	Valores Permitidos / Restrição (Chave)
Unidade Federativa	cod_uf	NUMBER(2)	N	S	PK
Unidade Federativa	sigla	CHAR(2)	N	S	
Unidade Federativa	nome	VARCHAR(22)	N	S	
Unidade Federativa	lat_uf	FLOAT	N	N	
Unidade Federativa	long_uf	FLOAT	N	N	
Unidade Federativa	regiao	VARCHAR(12)	N	N	
Unidade Federativa	populacao	INTEGER	N	N	
Municpios	cod_mun	NUMBER(7)	N	S	PK
Municpios	nome_mun	VARCHAR(30)	N	N	
Municpios	fuso	VARCHAR(30)	N	N	
Municpios	lat_mun	FLOAT	N	N	
Municpios	long_mun	FLOAT	N	N	
Municpios	ddd	NUMBER(2)	N	N	
Municpios	cod_uf	NUMBER(2)	N	N	FK(cod_uf) REFERENCES UNIDADE_FEDERATIVA(cod_uf)
Escolaridade	id	INTEGER	N	S	PK
Escolaridade	desc	VARCHAR(50)	N	S	
Raça	id	INTEGER	N	S	PK
Raça	desc	VARCHAR(12)	N	S	
Tipo Notificação	id	INTEGER	N	S	PK
Tipo Notificação	Tipo	VARCHAR(10)	N	S	
Tipo Infectado	id	INTEGER	N	S	PK
Tipo Infectado	desc	VARCHAR(6)	N	S	
Notificação de	id	INTEGER	N	S	PK
Infectados	dt_notific	DATE	N	N	
Notificação de	ano_nasc	NUMBER(4)	N	N	
Infectados	dt_ini_sint	DATE	N	N	
Notificação de	Infectados				

Continua na próxima página

– Continuação da Tabela –

Relação	Atributo	Tipo	Nulo	Único	Valores Permitidos / Restrições
Notificação de Infectados	sexo	CHAR(1)	N	N	M, F (Masculino, Feminino)
Notificação de Infectados	vomito	BOOLEAN	N	N	CHECK
Notificação de Infectados	cs_gestant	NUMBER(1)	S	N	1-1º Trimestre, 2-2º Trimestre, 3-3º Trimestre, 4-Idade gestacional ignorada, 5-Não, 6-Não se aplica, 9-Ignorado
Notificação de Infectados	tpautocto	NUMBER(1)	S	N	0,9-Ignorado, 1-Confirmado, 2-Descartado, 8-Inconclusivo 1-Sim, 2-Não, 3-Indeterminado
Notificação de Infectados	dt_digna	DATE	S	N	
Notificação de Infectados	sorotipo	NUMBER(1)	S	N	1-DEN 1, 2-DEN 2, 3-DEN 3, 4-DEN 4. CHECK(tipo_infec <> 1)
Notificação de Infectados	renal	NUMBER(1)	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	cod_mun_infec	NUMBER(7)	N	N	FK(cod_mun_infec) REFERENCES CES MUNICIPIOS(cod_mun)
Notificação de Infectados	cod_mun_res	NUMBER(7)	N	N	FK(cod_mun_res) REFERENCES MUNICIPIOS(cod_mun)
Notificação de Infectados	tipo_not	INTEGER	N	N	FK(tipo_not) REFERENCES TIPO_NOTIFICACAO(id)
Notificação de Infectados	tipo_infec	INTEGER	N	N	FK(tipo_infec) REFERENCES TIPO_INFECTADO(id)
Notificação de Infectados	raca	INTEGER	N	N	FK(raca) REFERENCES RACA(id)
Notificação de Infectados	escolaridade	INTEGER	N	N	FK(escolaridade) REFERENCES CES_ESCOLARIDADE(id)
Notificação de Infectados	sangramento	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	criterio	NUMBER(1)	S	N	0,9-Ignorado, 1-Laboratório, 2-Clinico epidemiológico, 3-Clinico

Continua na próxima página

– Continuação da Tabela –

Relação	Atributo	Tipo	Nulo	Único	Valores Permitidos / Restrições
Notificação de Infectados	resul_hiv	NUMBER (1)	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não realizado
Notificação de Infectados	doenca_trab	NUMBER (1)	S	N	1-Sim, 2-Não, 9-Ignorado. CHECK(tipo_infec <> 1)
Notificação de Infectados	evolucao	NUMBER (1)	S	N	1-Cura, 2-Óbito pelo agravo, 3-Óbito por outra causa, 9-Ignorado
Notificação de Infectados	dt_obito	DATE	S	N	
Notificação de Infectados	dt_encerra	DATE	S	N	
Notificação de Infectados	duplicid	NUMBER (1)	S	N	1-Não é duplicidade, 2-Duplicidade
Notificação de Infectados	resul_soro	NUMBER (1)	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não realizado
Notificação de Infectados	febre	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	cefaleia	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	mialgia	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	dor_costas	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	conjuntivit	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	artrite	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	artralgia	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	petequias	NUMBER (1)	S	N	1-Sim, 2-Não, 9-Ignorado. CHECK(tipo_infec <> 1)
Notificação de Infectados	leucopenia	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Notificação de Infectados	resul_prnt	NUMBER (1)	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não Realizado
Notificação de Infectados	resul_pcr	NUMBER (1)	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não Realizado

Continua na próxima página

– Continuação da Tabela –

Relação	Atributo	Tipo	Nulo	Único	Valores Permitidos / Restrições
Notificação de Infectados	resul_ns1	NUMBER (1)	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não Realizado
Notificação de Infectados	res_chik_s1	BOOLEAN	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não Realizado
Notificação de Infectados	res_chik_s2	NUMBER (1)	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não Realizado
Notificação de Infectados	hipertensa	BOOLEAN	S	N	1-Reagente, 2-Não Reagente, 3-Inconclusivo, 4-Não Realizado
Notificação de Infectados	diabetes	BOOLEAN	S	N	CHECK(tipo_infec <> 1)
Infectados					CHECK(tipo_infec <> 1)