INSTITUTO FEDERAL DE SÃO PAULO
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Felipe Dalbosco Paludo SP3123766

Projeto - Análise exploratória de dados

ESP1A5 - Estatística e Probabilidade

Profª. Josceli

São Paulo

2025

1.Introdução

A análise de dados emergiu como uma ferramenta indispensável para a formulação de políticas de saúde pública baseadas em evidências. A compreensão dos desfechos de saúde na população requer uma investigação que transcenda os fatores biológicos, englobando também os determinantes sociais e ambientais. Dentre estes, a qualidade da infraestrutura urbana como saneamento, pavimentação e iluminação, representa um fator crucial que pode influenciar diretamente os padrões de morbidade e mortalidade. Este trabalho se propõe a investigar quantitativamente esta relação no contexto brasileiro, valendo-se de dados abertos governamentais para explorar as conexões entre o ambiente de vida e os registros de óbito.

Para a presente investigação, foram empregados dois conjuntos de dados primários: os microdados do Sistema de Informações sobre Mortalidade (SIM) para o ano de 2022, fornecidos pelo DATASUS, e os dados agregados sobre o Entorno dos Domicílios, oriundos do Censo Demográfico de 2022. Reconhecendo o desafio de conectar diretamente estas duas fontes, foi desenvolvida uma estratégia metodológica robusta. A partir dos dados do Censo, foi construído um "Índice de Infraestrutura" para classificar 186 concentrações urbanas do Brasil. Esta classificação permitiu a segmentação da análise de mortalidade, possibilitando uma comparação direta e estatisticamente fundamentada entre áreas com distintos níveis de desenvolvimento de infraestrutura.

O objetivo central deste estudo é verificar a existência de diferenças estatisticamente significativas nos padrões de mortalidade em áreas urbanas com níveis de infraestrutura contrastantes. Para alcançar este objetivo, a análise foi guiada pelas seguintes perguntas de pesquisa:

1. Qual é o perfil descritivo da mortalidade no Brasil, considerando idade, sexo e principais grupos de causas de óbito?
2. A idade média de óbito apresenta uma diferença estatisticamente significativa entre as cidades classificadas com "melhor" e "pior" infraestrutura?
3. Existe uma associação estatística entre o nível de infraestrutura de uma localidade e os tipos de causas de morte mais prevalentes?
4. Como os padrões de mortalidade e o índice de infraestrutura se distribuem geograficamente pelo território brasileiro?

2.Metodologia

Toda a análise de dados deste projeto foi conduzida utilizando a linguagem de programação Python em um ambiente interativo Google Colab. Foram empregadas as seguintes bibliotecas de código aberto, que são padrão na área de Ciência de Dados: Pandas para a manipulação e estruturação dos dados; NumPy para operações numéricas; Matplotlib e Seaborn para a geração de gráficos estáticos; SciPy para a execução dos testes de hipótese estatística; e Geopandas para a criação e manipulação das visualizações geoespaciais.

O estudo utilizou duas fontes primárias de dados públicos. A primeira consiste nos microdados do Sistema de Informações sobre Mortalidade (SIM) para o ano de 2022, obtidos através da plataforma DATASUS do Ministério da Saúde. A segunda fonte são os dados agregados sobre o Entorno dos Domicílios, resultantes do Censo Demográfico de 2022 e disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). O processo de pré-processamento foi uma etapa crítica, envolvendo a seleção de variáveis de interesse de ambos os datasets, a decodificação da variável 'IDADE' do SIM para uma representação em anos, a remoção de registros com dados essenciais ausentes e a padronização dos códigos de município para um formato de 6 dígitos. A base de mortalidade foi, por fim, enriquecida com nomes de municípios e siglas das unidades da federação através da junção com tabelas de consulta.

Uma etapa central da metodologia foi a criação de um "Índice de Infraestrutura" para conectar os dois datasets. A partir dos dados do Censo, variáveis como "Via pavimentada - Existe" e "Existência de iluminação pública - Existe" foram normalizadas pelo total de domicílios de cada localidade, resultando em um indicador percentual para cada característica. O índice final para cada uma das 186 concentrações urbanas foi calculado como a média simples desses percentuais. Subsequentemente, utilizando a mediana do índice como ponto de corte, as localidades foram classificadas em dois grupos ("Melhor Infraestrutura" e "Pior Infraestrutura"), que serviram como base para toda a análise estatística comparativa.

A abordagem analítica foi multifacetada. Inicialmente, a estatística descritiva foi utilizada para resumir os dados através de medidas de tendência central e dispersão, bem como distribuições de frequência. A seguir, a análise de probabilidade, incluindo o cálculo de probabilidades condicionais, foi empregada para estimar a chance de ocorrência de determinados eventos. O núcleo do trabalho consistiu na inferência estatística, onde foram aplicados o Teste T para Amostras Independentes para comparar as médias de idade entre diferentes grupos (por sexo e por nível de infraestrutura) e o Teste Qui-Quadrado de Independência para verificar a associação entre o nível de infraestrutura e os tipos de causas de morte. Por fim, a análise geoespacial foi realizada através da criação de mapas para visualizar a distribuição espacial das variáveis de interesse no território brasileiro.

3. Resultados e Análise:

3.1 Análise Descritiva e Perfil da Mortalidade

A análise inicial dos dados de mortalidade visa caracterizar o perfil geral dos óbitos registrados no Brasil em 2022. Foram investigadas as distribuições de idade, sexo e as principais causas de morte para estabelecer uma base de compreensão antes das análises inferenciais.

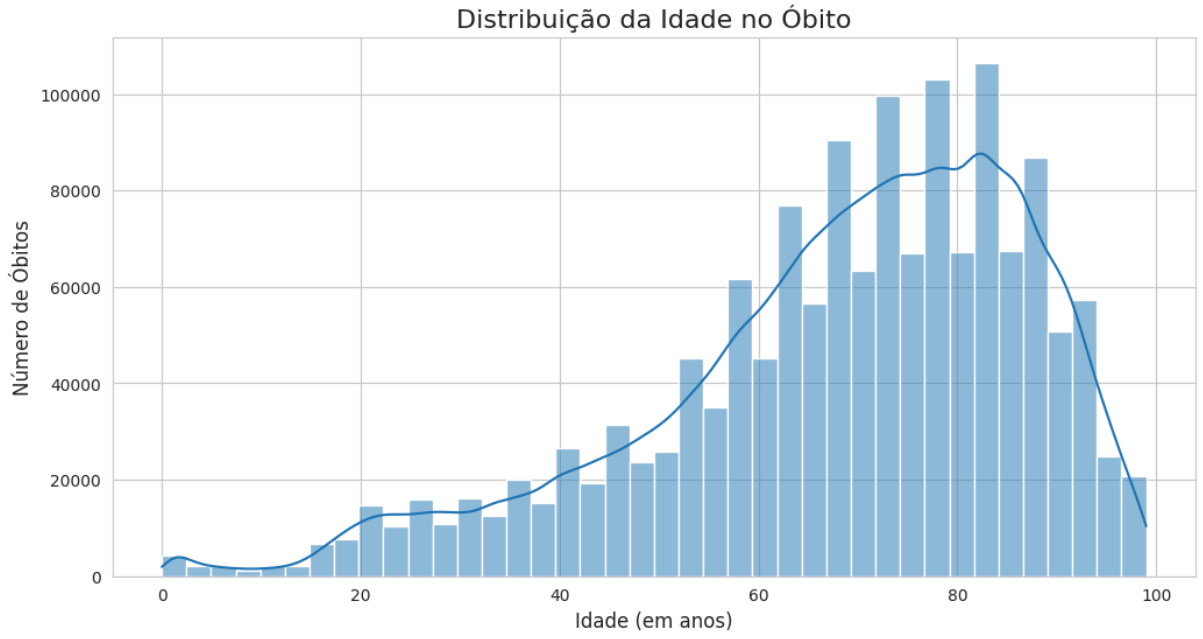
As estatísticas descritivas para a idade dos indivíduos no momento do óbito são apresentadas na Tabela 1.

Estatística	Valor
Contagem	1.496.396
Média	68,05
Desvio Padrão	19,24
Mínimo	0,00
25º Percentil	58,00
Mediana	72,00
75º Percentil	83,00
Máximo	99,00

[Tabela 1 - Estatísticas Descritivas da Idade no Óbito]

A Tabela 1 demonstra que a idade média de óbito na amostra analisada foi de 68,05 anos, com uma mediana de 72 anos. A proximidade entre a média e a mediana sugere uma distribuição com leve assimetria. O desvio padrão de 19,24 anos indica uma dispersão considerável dos dados de idade.

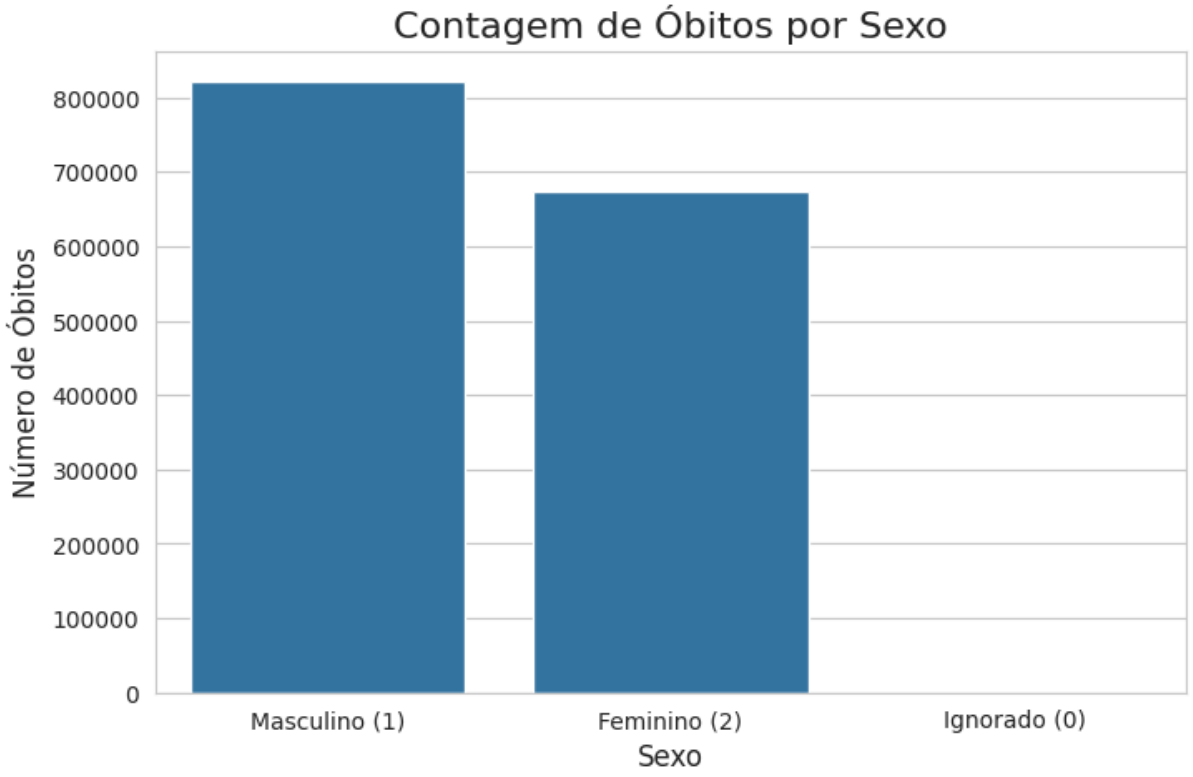
A distribuição das idades é melhor visualizada no histograma da Figura 1.



[Figura 1 - Distribuição da Idade no Óbito]

A Figura 1 confirma visualmente a concentração de mortes em idades mais avançadas. Nota-se um aumento progressivo no número de óbitos com o avançar da idade, com um pico proeminente na faixa etária de 70 a 80 anos, seguido de um declínio natural para as idades subsequentes.

A análise por sexo, apresentada na Figura 2, revela as diferenças na contagem de óbitos entre os sexos.



[Figura 2 - Contagem de Óbitos por Sexo]

Observa-se na Figura 2 um número maior de óbitos registrados para o sexo masculino em comparação com o feminino. De acordo com a análise de probabilidade realizada posteriormente, a chance de um óbito registrado ser do sexo masculino é de 54.91%.

A Tabela 2 detalha os principais grupos de causas de morte, classificados pelo capítulo da Classificação Internacional de Doenças (CID-10).

DESCRIÇÃO_CAUSA	Contagem de Óbitos	Percentual (%)
Doenças do Aparelho Circulatório	395.945	26,46%
Neoplasias (Tumores)	238.032	15,91%

Doenças do Aparelho Respiratório	171.658	11,47%
Causas Externas de Morbidade	150.112	10,03%
Doenças Infecciosas e Parasitárias	128.986	8,62%
Outras	113.361	7,58%
Doenças Endócrinas, Nutricionais e Metabólicas	92.686	6,19%

[Tabela 2 - Principais Grupos de Causas de Morte]

A Tabela 2 evidencia que as doenças do aparelho circulatório (Capítulo I) são a causa mais frequente de mortalidade na amostra, representando 26.46% do total de óbitos. Em seguida, destacam-se as neoplasias (Capítulo C) e as doenças do aparelho respiratório (Capítulo J), delineando as principais cargas de doença na população estudada.

3.2 Análise Comparativa da Mortalidade por Nível de Infraestrutura

Nesta seção, são apresentados os resultados da análise inferencial que investiga a relação entre o nível de infraestrutura das cidades, derivado dos dados do Censo, e os padrões de mortalidade registrados no SIM. Foram realizados múltiplos testes de hipótese para explorar esta conexão sob diferentes perspectivas.

O primeiro teste buscou verificar se a idade média de óbito difere entre os grupos de cidades com melhor e pior infraestrutura. As hipóteses para este teste foram definidas como:

- Hipótese Nula (H0): A idade média de óbito é igual em ambos os grupos.
- Hipótese Alternativa (H1): A idade média de óbito é diferente entre os grupos.

A Tabela 3 resume os resultados encontrados nesta análise.

Grupo	**Número de Óbitos**	**Idade Média (anos)**
Melhor Infraestrutura	297.547	69.69
Pior Infraestrutura	284.319	67.77

[Tabela 3 - Resultados do Teste T: Comparativo da Idade Média de Óbito por Nível de Infraestrutura]

Análise Estatística:
Estatística t = 38.4839
Valor-p = 0.000000

O grupo de cidades com melhor infraestrutura apresentou uma idade média de óbito de 69,69 anos, enquanto o grupo com pior infraestrutura registrou uma média de 67,77 anos. O Teste T para amostras independentes resultou em um valor-p de 0,000000. Como este valor é menor que o nível de significância de 0.05, a hipótese nula é rejeitada. Portanto, conclui-se que existe uma diferença estatisticamente significativa na idade média de óbito entre os dois grupos de cidades.

A seguir, investigou-se se o nível de infraestrutura estava associado ao perfil das causas de morte, utilizando um Teste Qui-Quadrado (x^2) de Independência. As hipóteses testadas foram:

- Hipótese Nula (H0): Não há associação entre o nível de infraestrutura e o grupo da causa do óbito.
- Hipótese Alternativa (H1): Existe uma associação entre as duas variáveis.

A Tabela de Contingência (Tabela 4) abaixo resume as contagens de óbitos observadas para cada combinação de grupo, focando nos cinco principais grupos de causas de morte.

Análise da Distribuição de Óbitos por Grupo de Causa e Nível de Infraestrutura:

GRUPO_INFRA	B	C	E	I	J
Melhor Infra	19731	53155	15125	80041	33783
Pior Infra	17140	48204	17781	70528	31292

[Tabela 4 - Relação entre Nível de Infraestrutura e Grupos de Causas de Morte]

Estatística Qui-Quadrado (χ^2): 598.2623
Valor-p: 0.000000

O Teste Qui-Quadrado produziu um valor-p de 0,000000. Sendo este valor menor que α=0.05, a hipótese nula de independência é rejeitada. Este resultado indica que existe uma associação estatisticamente significativa entre o nível de infraestrutura de uma cidade e o perfil das causas de morte.

Por fim, para aprofundar a análise, foi realizado um Teste T específico para o grupo de "Causas Externas" (acidentes, violência, etc.), investigando se a idade média das vítimas deste tipo de óbito varia entre os grupos de infraestrutura. A Tabela 5 resume os achados.

Grupo	Nº de Óbitos	Idade Média (anos)
Melhor Infraestrutura	25.075	49.85
Pior Infraestrutura	29.702	44.24

[Tabela 5:Resultados do Teste T - Idade Média de Óbito por Causas Externas]

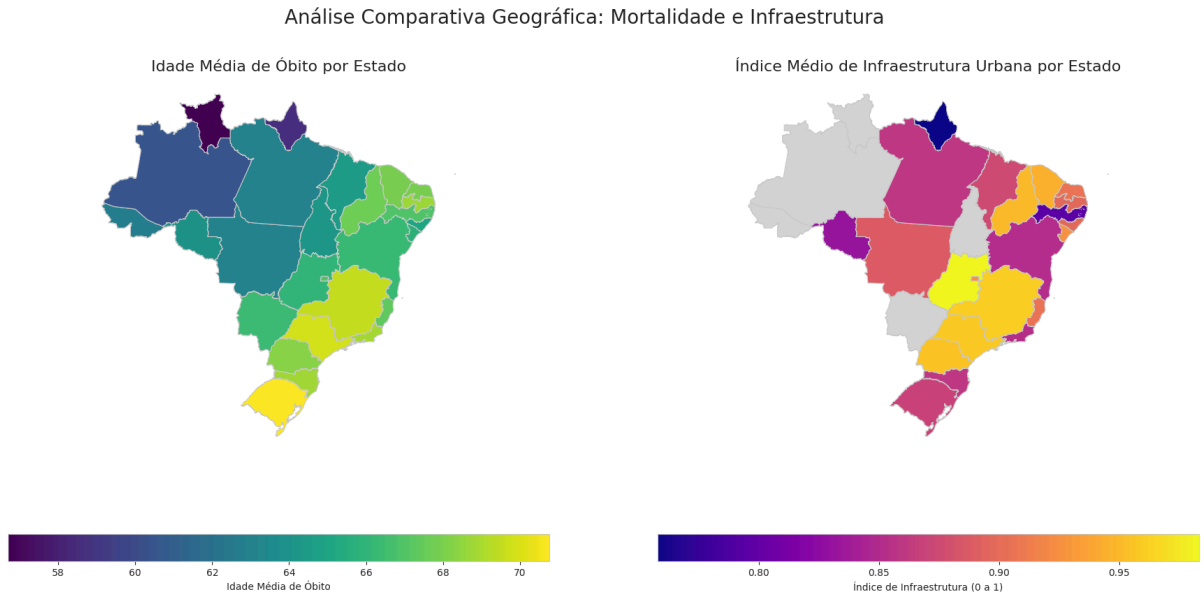
Estatística t: 28.1458

Valor-p: 0.000000

Focando apenas nas vítimas de causas externas, a idade média de óbito foi de 49,85 anos no grupo de melhor infraestrutura e 44,24 anos no grupo de pior infraestrutura. O valor-p resultante do Teste T para esta subpopulação foi de 0,000000. Com base neste resultado, é possível afirmar, com 95% de confiança, que a idade média das vítimas de causas externas difere significativamente entre os dois tipos de localidade.

3.3 Análise Geoespacial Comparativa

Para complementar a análise estatística e explorar a dimensão espacial dos dados, foi realizada uma análise geoespacial. O objetivo foi visualizar a distribuição da idade média de óbito e do índice de infraestrutura no território brasileiro, permitindo a identificação de possíveis padrões regionais e correlações visuais.



[Figura 3 - Mapas Comparativos da Idade Média de Óbito (esquerda) e do Índice Médio de Infraestrutura (direita) por UF]

A Figura 3 apresenta a análise geográfica comparativa. O mapa à esquerda ilustra a idade média de óbito, onde as cores mais quentes indicam uma maior longevidade média. É

possível observar um claro padrão regional: os estados pertencentes às regiões Sul e Sudeste do país consistentemente apresentam as maiores idades médias de óbito. Em contrapartida, os estados das regiões Norte e Nordeste exibem, em geral, uma idade média de morte inferior, indicada pelas cores mais frias.

O mapa à direita, por sua vez, exibe o índice médio de infraestrutura, calculado a partir dos dados do Censo. Nota-se um padrão geográfico notavelmente similar ao do mapa de mortalidade. Os mesmos estados do Sul e Sudeste que registraram maior idade média de óbito são também aqueles que apresentam os maiores índices de Infraestrutura.

A forte semelhança visual entre os dois mapas sugere uma aparente correlação positiva em nível estadual: localidades com melhor infraestrutura tendem a apresentar uma idade média de óbito mais elevada. Esta observação geoespacial reforça e contextualiza os resultados do Teste T (apresentados na Seção 3.2), que já haviam apontado uma diferença estatisticamente significativa entre os grupos de cidades com infraestrutura distinta.

4. Discussão

Os resultados apresentados neste estudo mostram uma conexão clara e estatisticamente significativa sobre a relação entre infraestrutura urbana e mortalidade no Brasil. A análise estatística demonstrou uma diferença significativa na idade média de óbito entre áreas com distintos níveis de infraestrutura, um achado que foi visualmente corroborado pela análise geoespacial. A aparente correlação positiva entre um maior índice de infraestrutura e uma maior longevidade média sugere que o investimento em infraestrutura urbana — como saneamento, pavimentação e acesso a serviços — pode atuar como um fator protetor para a saúde da população, contribuindo para uma redução da mortalidade prematura.

Ao aprofundar a análise, um dos achados mais interessantes foi que o Teste Qui-Quadrado, ao indicar uma associação entre o nível de infraestrutura e o perfil das causas de morte, sugere que o ambiente urbano não apenas se relaciona com "quando" se morre (idade), mas também com "o porquê". O teste focado em "Causas Externas" reforçou essa ideia, mostrando que a idade média das vítimas de acidentes e violência é significativamente menor em áreas com piores condições de infraestrutura. Isso pode estar pragmaticamente ligado a fatores como piores condições de tráfego, menor iluminação pública ou outras vulnerabilidades sociais que expõem populações mais jovens a riscos.

É fundamental, contudo, reconhecer as limitações deste estudo. A análise de infraestrutura baseou-se em uma amostra de 186 concentrações urbanas, não representando a totalidade do território nacional e resultando na ausência de dados para estados como Amazonas, Mato Grosso e Piauí nos mapas comparativos. Adicionalmente, o "Índice de Infraestrutura" criado é uma métrica agregada e simplificada. Por fim, as associações encontradas são estatísticas e não estabelecem uma relação de causalidade direta. Outros fatores de confusão, como renda, acesso a serviços de saúde e hábitos de vida, que não foram controlados nesta análise, certamente desempenham um papel relevante nos desfechos de mortalidade.

5. Conclusão

Este trabalho demonstrou, através de uma abordagem quantitativa multifacetada, a existência de uma conexão estatisticamente significativa e geograficamente visível entre a qualidade da infraestrutura urbana e os padrões de mortalidade no Brasil. Áreas com melhores indicadores de infraestrutura apresentaram, em média, uma idade de óbito superior e um perfil de causas de morte distinto, mesmo quando analisando subgrupos específicos como as vítimas de causas externas.

Os achados reforçam a importância de políticas públicas integradas que considerem o planejamento urbano como um pilar da saúde pública. A análise de dados, como a realizada neste projeto, serve como uma ferramenta poderosa para identificar disparidades e informar decisões estratégicas que visem à promoção da saúde e à redução das desigualdades. Sugere-se que estudos futuros possam aprofundar esta investigação utilizando dados em nível municipal para todo o país e incluindo outras variáveis socioeconômicas para a construção de um modelo explicativo mais completo.