

Projeto Final: Saúde Pública

Introdução à Ciência de Dados

Bruna Saturnino de Carvalho
Gabriell Laurentino Ferreira Bispo
João Vitor de Carvalho Silva
Pedro Henrique Egito Aguiar
Rafael Araujo Magesty

Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

1. Introdução

A análise de dados de saúde pública é fundamental para avaliar a eficiência de políticas públicas e identificar fatores que influenciam diretamente os resultados de indicadores essenciais, como mortalidade infantil, cobertura vacinal e disponibilidade de médicos por habitante. Este trabalho utiliza dados regionais para explorar a relação entre indicadores de saúde, como percentual de pré-natal adequado, taxa de médicos e gastos em saúde per capita. Por meio de técnicas de visualização e análise estatística, buscamos compreender padrões e possíveis correlações que subsidiem a formulação de políticas mais eficazes.

O foco principal desta análise é investigar como diferentes variáveis, como gastos regionais em saúde e indicadores de cobertura de serviços, estão relacionadas a resultados de saúde pública. Por meio de gráficos e modelos estatísticos, responderemos a perguntas específicas e identificar relações significativas entre os dados, com base em um conjunto de informações de abrangência nacional.

Todos os notebooks utilizados para gerar os resultados aqui presentes podem ser encontrados em: <https://github.com/brunasaturnino/Projeto-ICD>

2. Motivação

A saúde pública no Brasil enfrenta desafios complexos, incluindo desigualdades regionais, limitações orçamentárias e demandas crescentes por serviços de qualidade. Apesar

de avanços significativos, como o aumento da cobertura da Estratégia de Saúde da Família e a ampliação do acesso a serviços básicos, ainda existem lacunas críticas em muitas áreas, especialmente em regiões com menor capacidade econômica.

Este trabalho é motivado pela necessidade de compreender melhor como diferentes fatores interagem no contexto da saúde pública. A relação entre gastos em saúde e indicadores, como a cobertura vacinal e a taxa de médicos, é particularmente relevante para avaliar a eficácia do investimento público. Além disso, analisar o impacto do percentual de pré-natal adequado e de outros fatores no desempenho de indicadores críticos pode ajudar a priorizar políticas de alocação de recursos e estratégias de intervenção.

3. Perguntas Principais

1. Há uma correlação entre o número de famílias beneficiadas pelo Bolsa Família e a taxa de mortalidade nos municípios?
2. Existe uma relação estatística entre o número de médicos e a taxa de mortalidade nos municípios?
3. Como as despesas totais de saúde per capita influenciam a cobertura da atenção básica?
4. Qual é o impacto do gasto estadual com saúde per capita na taxa de mortalidade infantil?
5. Municípios com maiores despesas em saúde municipal tendem a ter mais médicos?

4. Dados Utilizados

Para a realização deste projeto, foi necessário utilizar dados de domínio público. Nesta seção iremos entrar em mais detalhes sobre a base utilizada.

IEPS Data

Descrição: O banco de dados utilizado neste trabalho é composto por informações regionais sobre indicadores de saúde no Brasil, abrangendo aspectos relacionados à cobertura de serviços, mortalidade, internações, infraestrutura hospitalar, profissionais de saúde, e gastos públicos em saúde. A base de dados é estruturada em forma tabular, onde cada linha representa uma região específica (como uma região de saúde ou estado) e as colunas contêm variáveis descritivas e quantitativas relacionadas ao ano de análise.

Fonte: IEPS Data

5. Limpeza dos Dados

As bases de dados selecionadas possuem alguns dados nos quais não estamos interessados e que podemos remover para melhorar o desempenho e facilitar o trabalho. Além disso, algumas estão faltando informações ou possuem uma formatação não muito adequada

6. Métodos e modelos utilizados

No projeto, foram utilizados diversos métodos e modelos para análise dos dados de saúde pública. A coleta de dados foi realizada a partir de bases públicas, como a IEPS Data e a Base dos Dados, contendo informações sobre infraestrutura hospitalar, profissionais de saúde, gastos públicos e indicadores epidemiológicos. Em seguida, foi feita a limpeza de dados, que incluiu a remoção de colunas irrelevantes, tratamento de valores ausentes com substituições pela mediana, média ou moda, e padronização dos formatos para compatibilidade com os modelos estatísticos.

A análise exploratória de dados (EDA) foi conduzida por meio da geração de estatísticas descritivas e gráficos, como histogramas e gráficos de dispersão, para identificar padrões e possíveis relações entre variáveis. Além disso, foram aplicados testes de hipótese para avaliar diferenças estatisticamente significativas entre grupos de municípios e estados.

Para modelagem estatística, foram utilizadas regressão linear e regressão polinomial para explorar a relação entre despesas per capita em saúde e indicadores de cobertura de serviços. Além disso, foram calculados coeficientes de correlação estatística para medir a associação entre diferentes variáveis, e métodos de inferência estatística, como intervalos de confiança, foram empregados para avaliar a robustez dos resultados.

7. Análise Exploratória

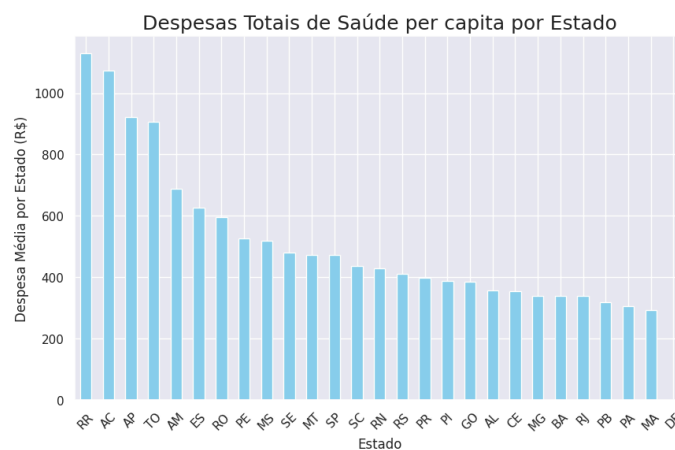


Imagem 1: Gráfico de despesas totais de saúde per capita no Estado

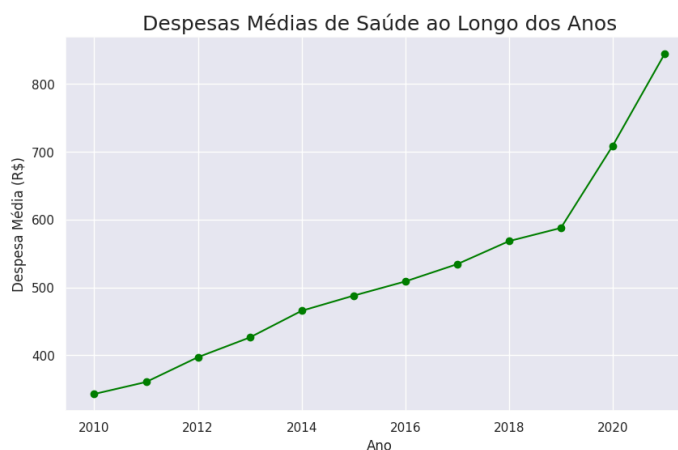


Imagem 2: Gráfico de despesas médias de saúde ao longo dos anos

Os dois gráficos mostram que, embora haja crescimento nos investimentos em saúde ao longo dos anos, a distribuição desses recursos varia significativamente entre os estados. Isso indica desigualdades regionais no financiamento da saúde, o que pode impactar o acesso e a qualidade dos serviços disponíveis para a população em diferentes partes do país.

Municípios com maior gasto em saúde per capita apresentam maior cobertura de Estratégia de Saúde da Família?

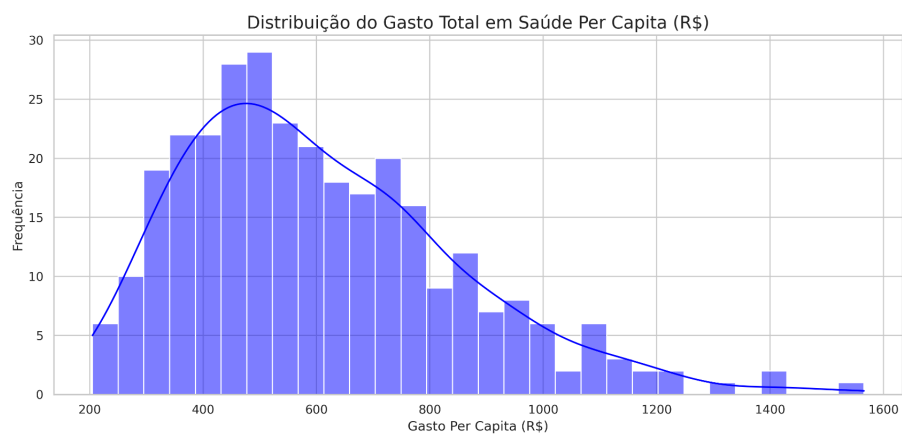


Imagem 3: Gráfico de Distribuição do Gasto Total em Saúde Per Capita

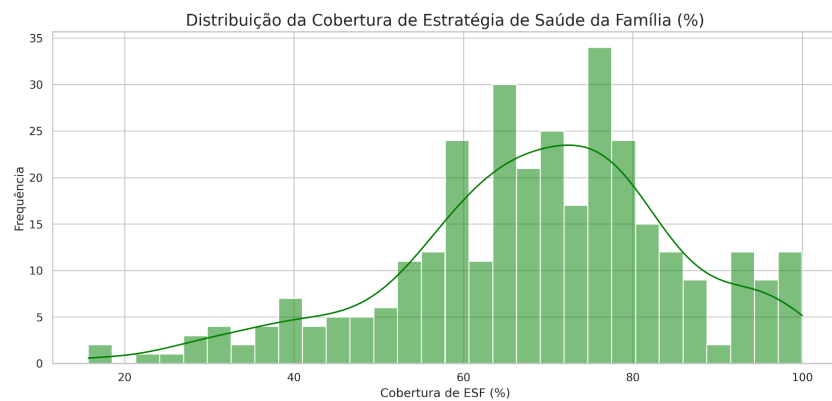


Imagem 4: Gráfico de Distribuição da Cobertura Estratégica de Saúde da Família

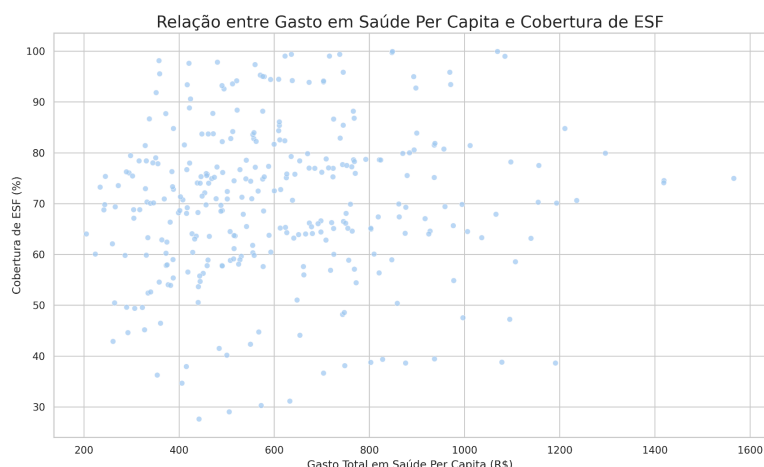


Imagem 5: Gráfico da relação entre gasto em saúde per capita e cobertura ESF

Com base nesses gráficos gerados, percebe-se que, embora o gasto em saúde per capita possa refletir maior investimento em infraestrutura ou serviços de saúde, ele não apresenta uma relação significativa com a cobertura da Estratégia de Saúde da Família (ESF), conforme indicado pela baixa correlação (0.08). Isso sugere que aumentar o orçamento per capita não garante, por si só, uma maior cobertura da ESF. Fatores como a eficiência na aplicação dos recursos, as prioridades locais de saúde, limitações logísticas ou estruturais, e a capacidade de expansão dos serviços de saúde desempenham um papel mais relevante. Para ampliar a cobertura da ESF, é necessário focar em estratégias específicas de gestão e planejamento regional que vão além do simples aumento de gastos.

Qual a relação entre o gasto estadual com saúde per capita e a taxa de mortalidade infantil?

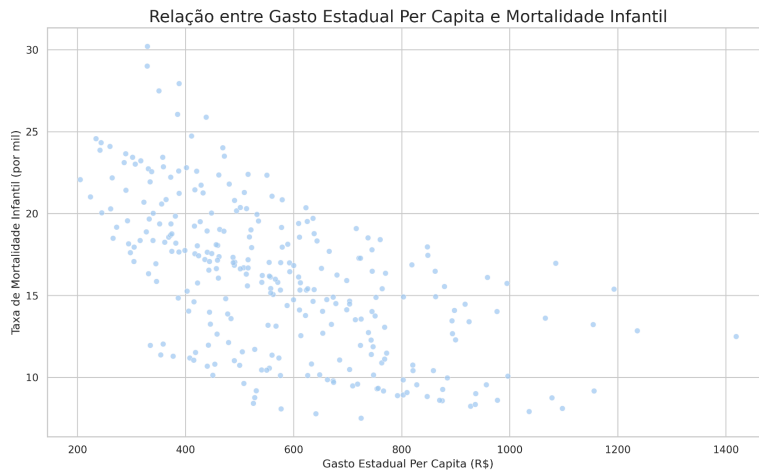


Imagem 6: Gráfico da Relação entre Gasto Estadual per Capita e Mortalidade Infantil

O gráfico ilustra a relação entre o gasto estadual per capita em saúde e a taxa de mortalidade infantil (por mil nascidos vivos). Observa-se uma tendência negativa, indicando que, em geral, estados com maiores gastos per capita apresentam menores taxas de mortalidade infantil. A relação inversa sugere que maiores investimentos em saúde por parte dos estados podem contribuir para melhores condições de saúde infantil, refletidas em menores taxas de mortalidade. Apesar da tendência, há uma certa dispersão dos dados, indicando que outros fatores, além do gasto em saúde, também influenciam a mortalidade infantil. Isso pode incluir questões de infraestrutura, políticas públicas, educação e acesso aos serviços de saúde. Dessa maneira, este resultado reforça a importância do investimento em saúde como um dos fatores determinantes para a redução da mortalidade infantil, embora políticas complementares sejam essenciais para maximizar os impactos desses investimentos.

Existe uma diferença significativa na cobertura de agentes comunitários de saúde em municípios com maior despesa própria municipal em saúde?

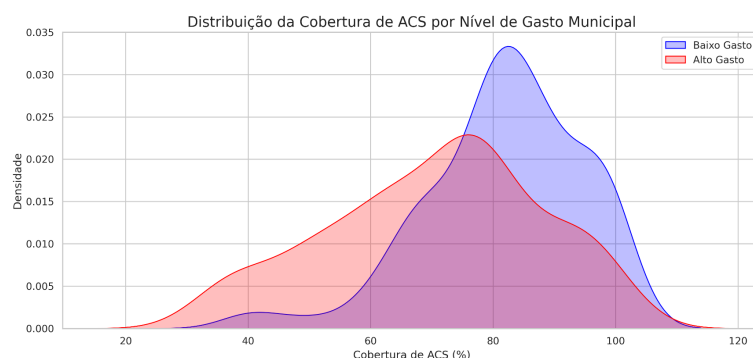


Imagem 7: Gráfico de distribuição da Cobertura ACS por Nível de Gasto Municipal

Existe uma diferença significativa na cobertura de ACS entre municípios com maior e menor despesa própria em saúde. Municípios com maior gasto apresentam, em média, uma cobertura maior de ACS, como indicado pela curva de densidade deslocada para a direita (maiores porcentagens de cobertura) em relação aos municípios com menor gasto, cuja curva está mais concentrada em valores mais baixos de cobertura. Isso sugere que um maior investimento municipal em saúde está associado a uma maior capacidade de expandir a cobertura de ACS, possivelmente devido ao maior financiamento para contratar agentes comunitários e implementar estratégias de atenção primária mais abrangentes.

Municípios com maior despesa per capita em saúde municipal possuem maior número de médicos por 1.000 habitantes?

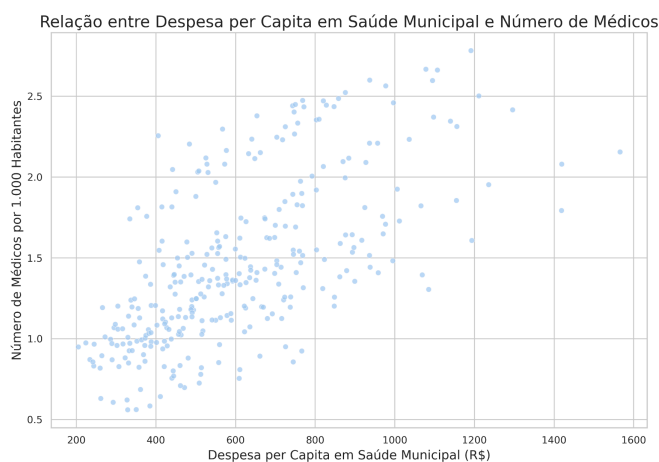


Imagem 8: Gráfico de relação entre despesa per capita de saúde municipal e número de médico

A partir do gráfico apresentado, pode-se observar uma correlação moderada positiva (correlação = 0.65) entre a despesa per capita em saúde municipal (R\$) e o número de médicos por 1.000 habitantes. Isso significa que, em geral, municípios que gastam mais em saúde por habitante tendem a ter mais médicos por 1.000 habitantes.

Municípios com maior gasto em saúde per capita possuem maior cobertura vacinal?

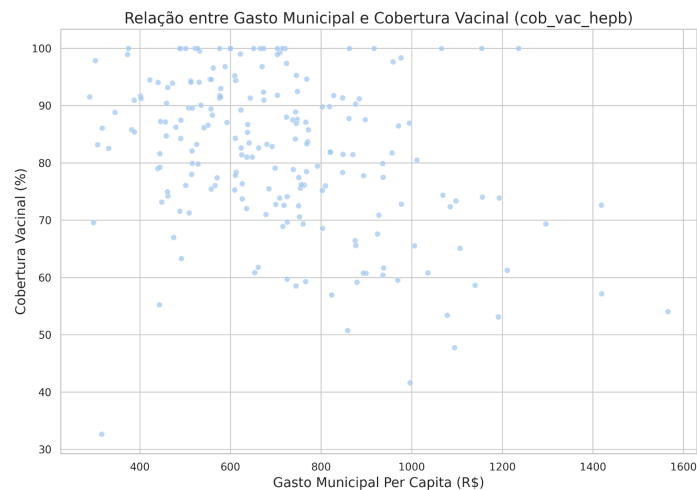


Imagem 9: Gráfico de relação entre gasto municipal e cobertura vacinal de Hepatite B

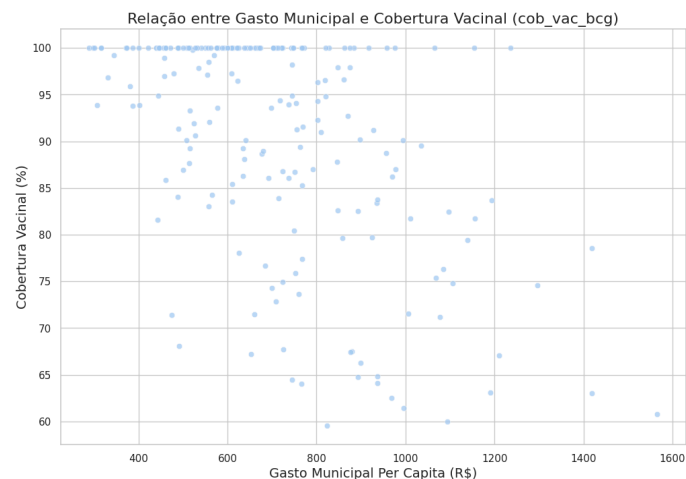


Imagem 10: Gráfico de relação entre gasto municipal e cobertura vacinal de BCG

A análise do gráfico indica não haver uma correlação direta significativa entre o gasto municipal per capita em saúde e a cobertura vacinal de BCG. A cobertura vacinal parece ser alta na maior parte dos municípios, independentemente do nível de gasto per capita. Isso sugere que a cobertura de BCG pode ser influenciada por outros fatores, como políticas nacionais de imunização, infraestrutura básica de saúde ou campanhas de vacinação.

Qual a relação entre o percentual de mulheres com 7 ou mais consultas pré-natal e a taxa de médicos por 1.000 habitantes?

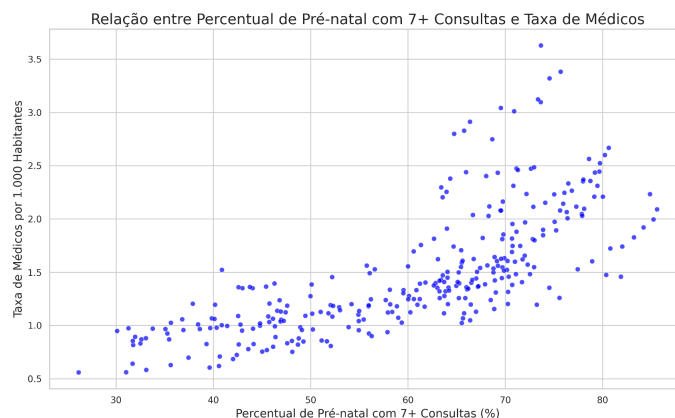


Imagem 11: Gráfico de relação entre porcentagem de Nascidos Vivos com 7 ou Mais Consultas Pré-Natal e taxa de médicos

O gráfico mostra uma relação positiva entre o percentual de mulheres com 7 ou mais consultas de pré-natal e a taxa de médicos por 1.000 habitantes. Essa tendência sugere que, em municípios com maior disponibilidade de médicos, é mais provável que mulheres grávidas realizem um número adequado de consultas pré-natal. A taxa de médicos por 1.000 habitantes parece estar associada a uma maior cobertura e acesso aos serviços de saúde para gestantes, refletido no percentual elevado de consultas pré-natal. Há certa variabilidade nos dados, indicando que outros fatores, como políticas locais de saúde, acesso a unidades de saúde e programas de incentivo ao pré-natal, também podem influenciar a realização de consultas.

A maior presença de médicos é um fator importante para garantir a realização adequada de consultas pré-natal, mas melhorias na infraestrutura e nas políticas públicas de saúde são igualmente cruciais para alcançar resultados consistentes em todas as regiões.

8. Testes de Hipótese e Intervalo de Confiança

8.1. Teste de Hipótese com Correlação de Pearson

Formulação das Hipóteses:

Hipótese nula (H_0): Não há correlação linear entre as variáveis ($\rho = 0$).

Hipótese alternativa (H_1): Há correlação linear entre as variáveis ($\rho \neq 0$).

O Coeficiente de Pearson (r) é calculado para medir a força e a direção da relação linear entre as variáveis. O p-valor associado ao coeficiente de Pearson indica a probabilidade de observar uma correlação tão extrema quanto a calculada, assumindo que a hipótese nula é verdadeira.

Assim, a estratégia adotada é a seguinte: se o p-valor for menor que o nível de significância (geralmente 0,05), rejeita-se a hipótese nula, concluindo que a correlação é estatisticamente significativa. Caso contrário, não se rejeita a hipótese nula.

8.1.1. Teste 1

Hipótese nula (H_0): Não há relação linear entre o número de famílias beneficiadas pelo Bolsa Família e a taxa de mortalidade ($\rho = 0$).

Hipótese alternativa (H_1): Há uma relação linear entre o número de famílias beneficiadas pelo Bolsa Família e a taxa de mortalidade ($\rho \neq 0$).

Correlation between number of families that receive BF and infant mortality rate

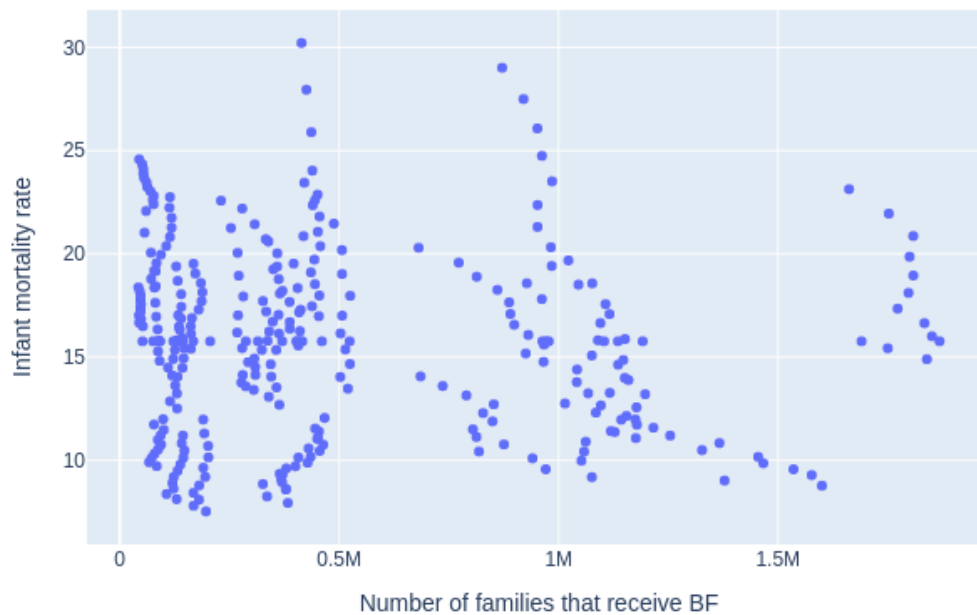


Imagem 12: Gráfico de correlação entre número de famílias apoiadas pelo Bolsa Família e taxa de mortalidade infantil

Nesta análise, obtemos os seguintes coeficientes:

Correlação: -0,07637703233834087

P-valor: 0,1702262391639119

À primeira vista, não foi identificada uma correlação aparente entre o número de famílias beneficiadas pelo Bolsa Família e a taxa de mortalidade nos municípios analisados. Para investigar melhor essa relação, propõe-se separar os municípios por estados e realizar uma análise mais granular.

Correlation between number of families that receive BF and infant mortality rate

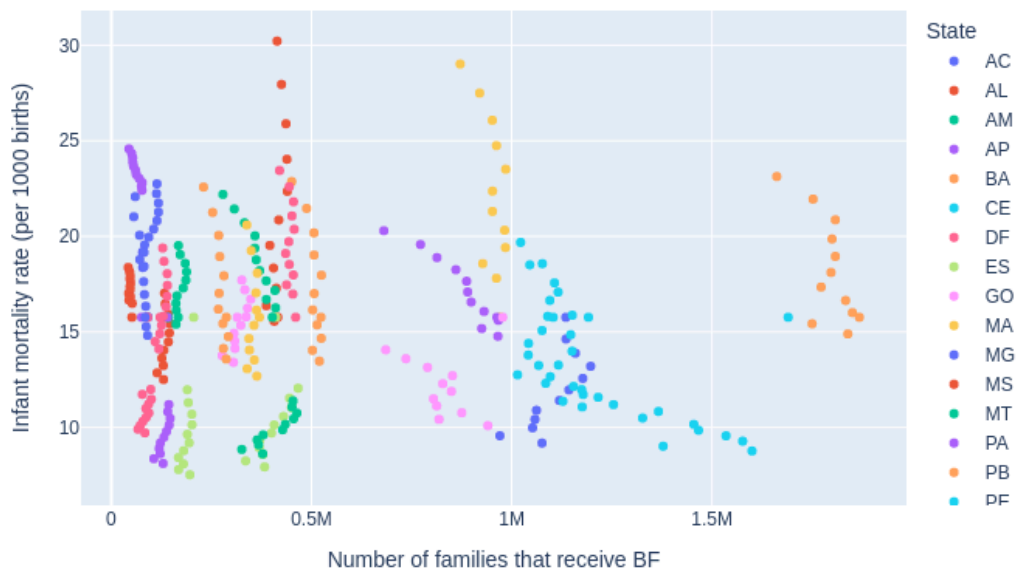


Imagem 13: Gráfico de correlação entre número de famílias apoiadas pelo Bolsa Família e taxa de mortalidade infantil, agrupada por estados

Observou-se a formação de clusters. Isso sugere haver padrões regionais na relação entre o número de famílias beneficiadas pelo Bolsa Família e a taxa de mortalidade. Vamos focar no estado de São Paulo como exemplo para uma análise mais detalhada.

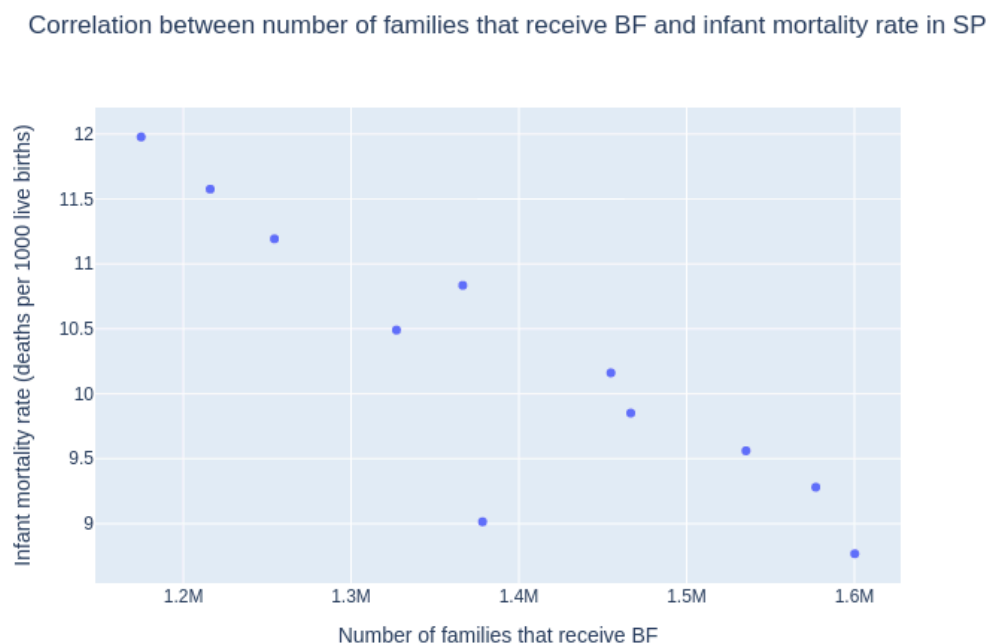


Imagem 14: Gráfico de correlação entre número de famílias apoiadas pelo Bolsa Família e taxa de mortalidade infantil em SP

Correlação: -0,8911680619534786

P-valor: 0,00023106386214497837

Após realizar a análise dos municípios do estado de São Paulo, observou-se uma correlação significativa entre o número de famílias beneficiadas pelo Bolsa Família e a taxa de mortalidade. Isso indica haver uma relação estatisticamente relevante entre essas duas variáveis no estado.

8.1.2. Teste 2

Hipótese nula (H_0): Não há relação entre o número de médicos e a taxa de mortalidade nos municípios ($\rho = 0$).

Hipótese alternativa (H_1): Há uma relação negativa entre o número de médicos e a taxa de mortalidade nos municípios ($\rho < 0$).

Correlation between number of medical doctors and mortality rate

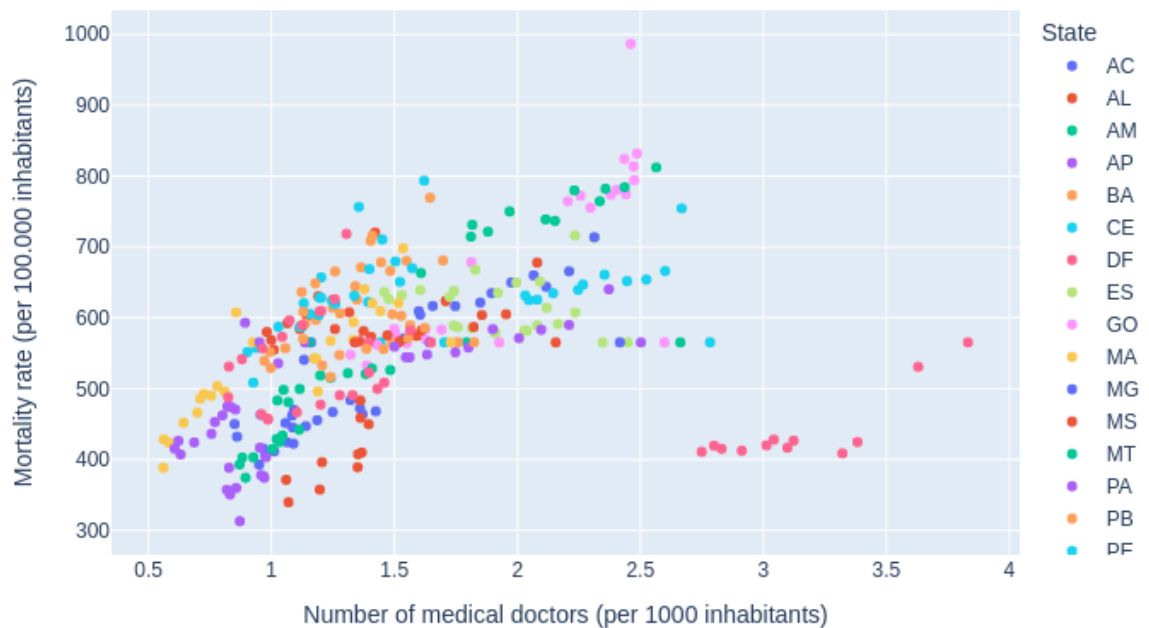


Imagem 15: Gráfico de correlação entre número de médicos e taxa de mortalidade, agrupada por estados

Correlação: 0,43788442313582865

P-value: 1,3071410617845005e-16

Após realizar o teste de hipótese para avaliar a relação entre o número de médicos e a taxa de mortalidade nos municípios, observou-se uma correlação positiva. Esse resultado indica que, nos municípios analisados, mais médicos está associado a uma maior taxa de mortalidade.

9. Regressão

A aplicação da regressão no projeto foi explorar a relação entre despesas totais de saúde e um indicador de cobertura.

A partir dos dados que obtivemos, foram escolhidas duas regressões para serem utilizadas na análise feita. Linear e Polinomial.

Para a análise feita nessa etapa do trabalho, foi realizada uma nova limpeza dos dados originais do banco de dados. Optou-se por limpá-los de modo que, ao invés de substituir os valores faltantes nas colunas financeiras por zero, eles foram substituídos pela mediana dos valores originais diferentes de NaN. Para as outras colunas, a substituição foi feita pela média (em colunas numéricas) ou moda (em colunas categóricas).

Para chegar nos modelos de regressão que seriam utilizados para analisar os dados, foi observada a distribuição das variáveis

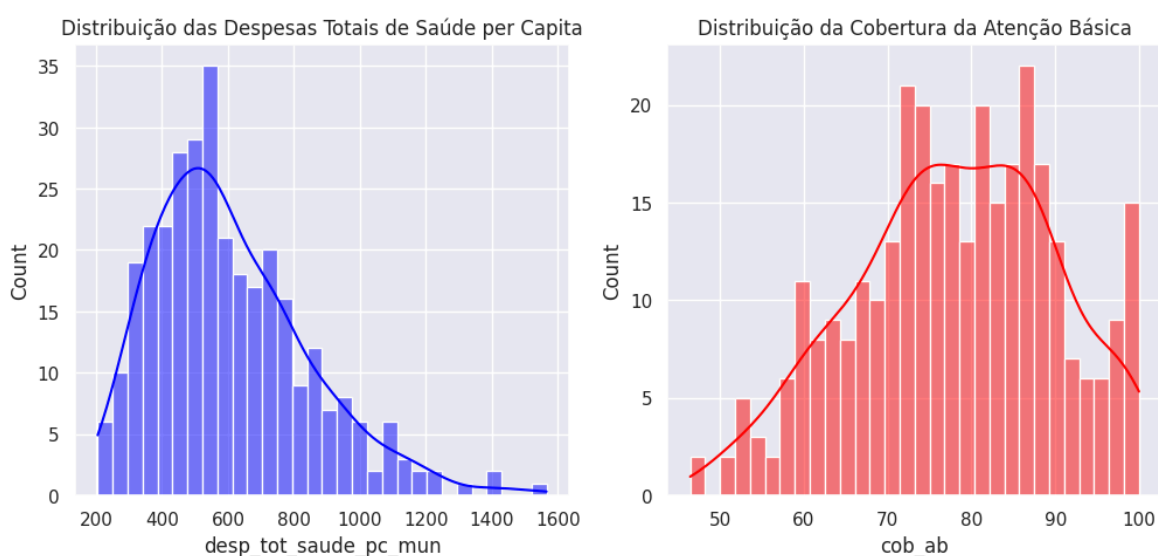


Imagem 16: Distribuição das variáveis de Despesas Totais de Saúde per Capita e Cobertura da Atenção Básica.

A partir dessa distribuição, foi decidido que seriam usados os modelos de regressão linear e regressão polinomial.

9.1. Regressão Linear

A regressão linear é uma técnica estatística usada para prever o valor de uma variável com base no valor de outra variável.

Foi feita então a regressão linear das variáveis apresentadas que produziu como resultado um coeficiente = 0,01, um Intercepto = 74,76, o Erro Quadrático Médio: 152,87 e o $R^2 = 0.00$. Veja abaixo o gráfico produzido:

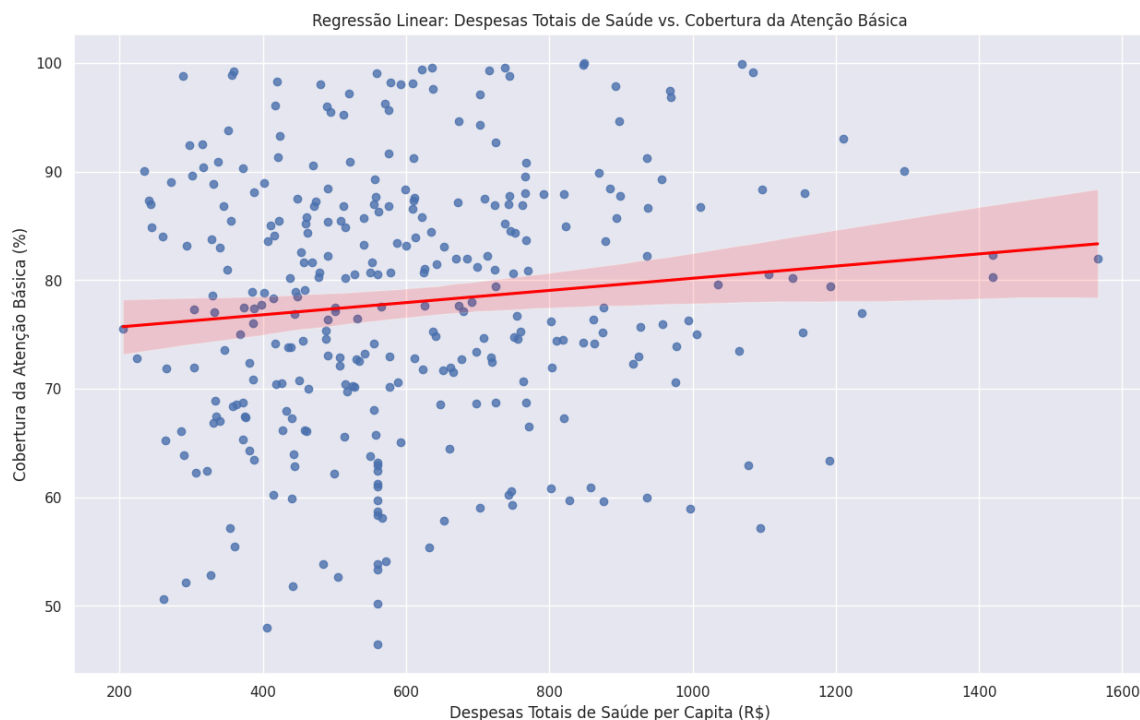


Imagem 17: Gráfico de Regressão Linear

Podemos observar pela distribuição dos dados no gráfico que a cobertura de atenção básica tende a crescer conforme crescem as despesas totais de saúde per capita, mesmo que lentamente se nos guiarmos pela linha da regressão.

Apesar de fazer sentido, a regressão linear pode não representar da melhor maneira esse conjunto de dados que está sendo analisado. Isso fica, inclusive, visualmente claro, já que a distribuição dos pontos no gráfico não representa ou lembra uma linha.

Com isso, foi introduzida no projeto a regressão polinomial.

9.2. Regressão Polinomial

A regressão polinomial é uma técnica de análise de regressão que modela a relação não linear entre uma variável dependente e uma variável independente.

A regressão polinomial foi aplicada no nosso conjunto de dados, obtendo os seguintes resultados: coeficientes = [0, 8,6179158, -0,38845024], intercepto = 39,21, Erro Quadrático Médio: 155,23 e R^2 : -0.01. Veja abaixo o gráfico produzido:

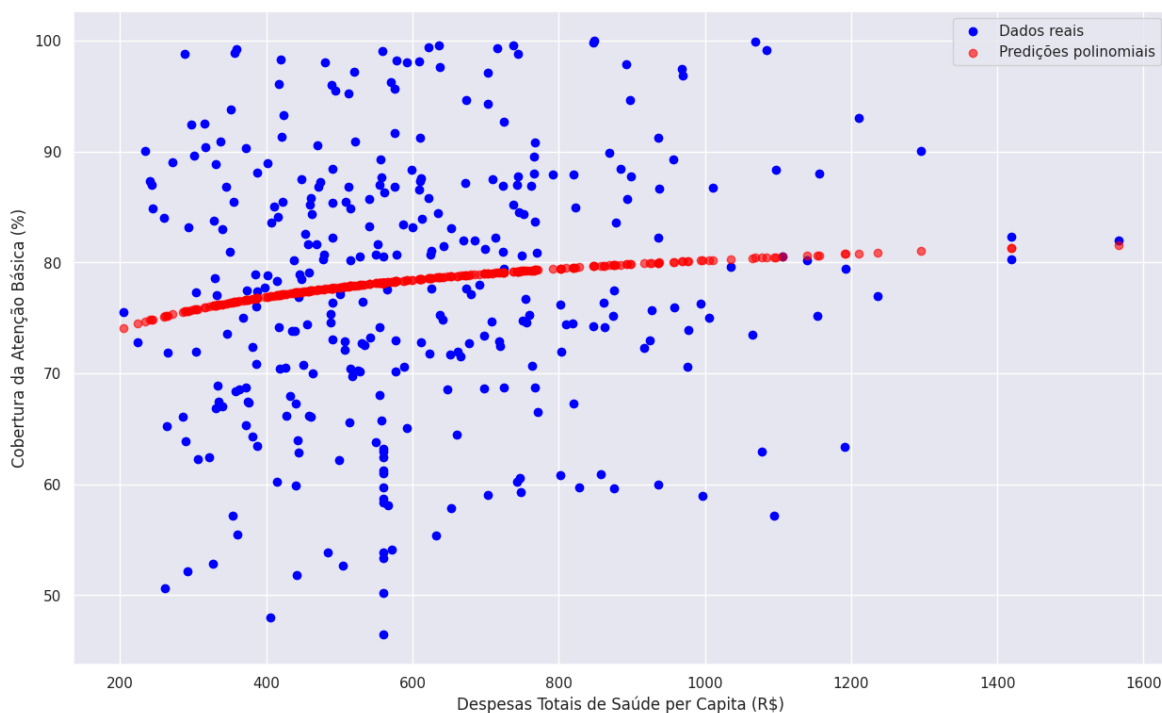


Imagem 18: Gráfico de Regressão Polinomial

É possível perceber que a linha traçada pela regressão polinomial se assemelha a uma reta, contrariando a hipótese que ela representaria os dados de uma forma diferente do que a regressão linear.

Existem alguns detalhes que podem ser melhor identificados, como o crescimento maior da cobertura a partir do aumento das despesas em seus níveis iniciais. Após as despesas de saúde per capita atingirem 400 reais, a linha tende a se estabilizar, mas continua apresentando um crescimento, ainda que baixo.

9.3. Previsões

Com base nos resultados obtidos e nos gráficos gerados pelos modelos de regressão linear e polinomial, é possível observar uma relação entre as despesas totais de saúde per capita e a cobertura da atenção básica. Os dados sugerem que o aumento nas despesas está associado a um crescimento na cobertura, embora esse aumento seja marginal, como evidenciado pela inclinação reduzida da linha na regressão linear e pela estabilização da curva no modelo polinomial após certo ponto, em torno de R\$ 400 de despesas per capita. Isso indica que o impacto inicial das despesas sobre a cobertura é mais significativo em

níveis baixos de investimento, mas tende a diminuir à medida que os valores per capita aumentam.

Além disso, os baixos valores de R^2 nos modelos aplicados (0,00 na regressão linear e -0,01 na polinomial) reforçam que as despesas per capita explicam apenas uma fração mínima da variação na cobertura de atenção básica. Isso sugere que outros fatores, como características socioeconômicas, infraestrutura de saúde e políticas públicas, desempenham um papel mais relevante nessa dinâmica. Dessa forma, embora seja possível prever que um aumento nos gastos resulte em uma melhora na cobertura, essa relação apresenta retornos decrescentes e depende de um direcionamento eficiente dos recursos.

Outro ponto a ser destacado é a estabilidade da cobertura em regiões com altos níveis de despesa per capita, com variações positivas pouco expressivas. Isso reforça a necessidade de repensar a alocação de recursos, já que mudanças estruturais ou políticas específicas podem ser mais eficazes do que aumentos indiscriminados nos investimentos. Assim, os resultados indicam que, para alcançar níveis mais elevados de cobertura, será necessário não apenas aumentar as despesas, mas também explorar estratégias complementares que considerem os diversos fatores que influenciam a atenção básica.

9.4. Discussão

Foi possível observar inúmeros pontos dispersos pelo gráfico que estão distantes das linhas de regressão. Com isso, podemos ter um indício de que há outros fatores além das despesas que influenciam a cobertura da atenção básica.

A estabilização da regressão polinomial e a baixa inclinação da regressão linear pode sugerir que aumentar os gastos além de um nível específico (entre 600 e 800 reais) pode não gerar grandes melhorias na cobertura.

Apesar dos resultados encontrados, a limpeza dos dados para a execução da regressão e o formato que os dados foram apresentados foram um dos grandes desafios da implementação desta seção do Projeto. Como foi mencionado, a limpeza teve de ser realizada novamente para que houvesse uma diminuição de outliers e a para que as regressões fossem mais acuradas. Além de que, as bibliotecas usadas para regressão não aceitam valores NaN, consideravelmente presentes nas colunas financeiras do banco de dados.

10. Conclusão

Os resultados obtidos nesta análise reforçam a importância de investir de forma estratégica em saúde pública para reduzir desigualdades regionais e melhorar os resultados gerais. Foi identificado que, embora haja certa correlação entre indicadores como gastos per capita e cobertura de serviços de saúde, outros fatores estruturais e socioeconômicos também desempenham papéis significativos.

A análise sugere que políticas voltadas à ampliação da cobertura de pré-natal e da infraestrutura médica podem ter impactos diretos na melhoria de indicadores de saúde. No entanto, a variabilidade regional demonstra que uma abordagem única não será suficiente: é necessário considerar as especificidades locais para maximizar os resultados. Por fim, este trabalho destaca a importância de integrar dados de diferentes dimensões para uma análise mais completa e informada, contribuindo para a tomada de decisões baseada em evidências.

11. Referências

BASE DOS DADOS. **Indicadores de saúde pública no Brasil por regiões e estados.**

Disponível em:

<https://basedosdados.org/dataset/bbc6df97-08ff-4bb4-819d-6e5d85a80392?table=480e7218-8c65-4ae8-85a9-deffe37c1e68>. Acesso em: 24 jan. 2025.

SATURNINO, Bruna. **Projeto ICD.** Disponível em:

<https://github.com/brunasaturnino/Projeto-ICD>. Acesso em: 24 jan. 2025.

<https://www.ibm.com/br-pt/topics/linear-regression>

<https://developers.google.com/machine-learning/crash-course/linear-regression?hl=pt-br>