SERVIÇO NACIONAL DE APRENDIZAGEM INDUSTRIAL

Julio Coronetti Regino

Murilo Antunes da Silva Galhardo de Carvalho

Paola de Oliveira

ANÁLISE ESTATÍSTICA DOS DADOS DE VACINAÇÃO NO BRASIL

Professor André Souza Ciências de dados

Sorocaba

2025

SUMÁRIO

3. INTRODUÇÃO

A vacinação é uma das principais estratégias de saúde pública no Brasil, essencial para prevenir doenças e controlar surtos. Por meio do Programa Nacional de Imunizações (PNI), o país oferece vacinas gratuitas para diversas faixas etárias, contribuindo significativamente para a redução da mortalidade e a proteção da população.

Este trabalho tem como objetivo realizar uma análise estatística dos dados de vacinação no Brasil utilizando Python, buscando identificar padrões, tendências e correlações. A escolha do tema se justifica pela relevância da vacinação, principalmente após a pandemia de COVID-19, que evidenciou desafios como desigualdade no acesso e a necessidade de um monitoramento eficiente para apoiar decisões e políticas de saúde pública.

4. REFERENCIAL TEÓRICO

A análise estatística dos dados de vacinação no Brasil é essencial para compreender a cobertura, distribuição e eficácia das campanhas de imunização. Medidas como média, mediana e moda são fundamentais para identificar o comportamento central dos dados, enquanto desvio padrão e variância avaliam a dispersão e possíveis desigualdades no acesso às vacinas.

Testes de normalidade, como Shapiro-Wilk e Anderson-Darling, são aplicados para verificar a distribuição dos dados e garantir a escolha adequada dos métodos estatísticos. A análise de correlação permite identificar relações entre variáveis, como cobertura vacinal e indicadores de saúde. Além disso, a regressão linear simples contribui para observar tendências e realizar previsões.

O processamento e a análise dos dados são realizados na linguagem Python, utilizando bibliotecas específicas. O **Pandas** permite a manipulação e organização dos dados; o **NumPy** auxilia nos cálculos matemáticos; **Matplotlib** e **Seaborn** são empregadas para visualização gráfica dos resultados; e o **SciPy** oferece ferramentas para testes estatísticos e modelagem. A biblioteca **Statsmodels** é utilizada, quando necessário, para análises estatísticas mais robustas, como regressões.

Essas ferramentas, aliadas aos conceitos estatísticos, viabilizam uma análise precisa, favorecendo a compreensão dos desafios e dos avanços na vacinação no Brasil.

5. METODOLOGIA

Este trabalho utilizou dados públicos sobre vacinação no Brasil, obtidos na plataforma **OpenDataSUS**, do Ministério da Saúde. A base inclui informações como tipos de vacinas, número de doses, datas e regiões.

As análises foram realizadas na plataforma **Google Colab**, que permite rodar códigos em Python diretamente na nuvem, facilitando o desenvolvimento e a geração de gráficos.

O processo incluiu a **limpeza e organização dos dados**, com correção de erros, remoção de valores vazios e padronização de informações. Depois, foram feitas análises estatísticas e visuais para entender os dados.

Foram usadas bibliotecas como:

- Pandas (manipulação dos dados),
- NumPy (cálculos numéricos),
- Matplotlib e Seaborn (gráficos e visualizações),
- SciPy e Statsmodels (testes estatísticos e modelos de regressão).

Essas ferramentas ajudaram a explorar os dados, criar gráficos e aplicar os métodos estatísticos para entender melhor a vacinação no Brasil.

6. ANÁLISE DE DADOS

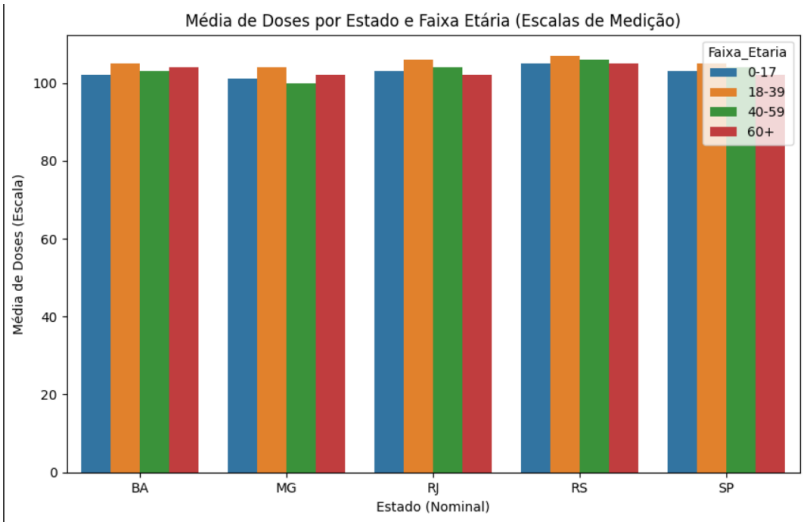
6.1 Tipos de Amostragem:

	Estado	Doses	Faixa_Etaria
0	RJ	88	18-39
1	BA	106	60+
2	MG	90	60+
3	MG	90	0-17
4	RJ	109	18-39
5	RS	101	40-59
6	RS	120	0-17
7	RS	111	0-17
8	BA	86	60+
9	RJ	96	60+

Visualizações e Interpretações: Foi feita uma amostragem estratificada, escolhendo 50 registros de cada estado (SP, RJ, MG, BA e RS), garantindo equilíbrio entre os grupos. A visualização mostra dados variados em faixas etárias e doses, sem agrupamento por estado.

Discussão dos Resultados Obtidos: A amostragem garantiu equilíbrio entre os estados, evitando que um estado com mais registros dominasse a análise. Isso torna a visualização mais justa e ajuda a entender melhor como estão distribuídas as doses e as faixas etárias entre os estados.

6.2 Escalas de Medição:



Visualizações e Interpretações: Os dados foram classificados segundo escalas nominal, ordinal, intervalar e de razão, cada uma com objetivo distinto na medição. A visualização destaca como a escolha da escala afeta a análise dos dados.

Discussão dos Resultados Obtidos: Entender o objetivo de cada escala é importante para aplicar corretamente as análises, possibilitando interpretações adequadas conforme o tipo de dado.

6.3 Medidas de Têndencia Central:

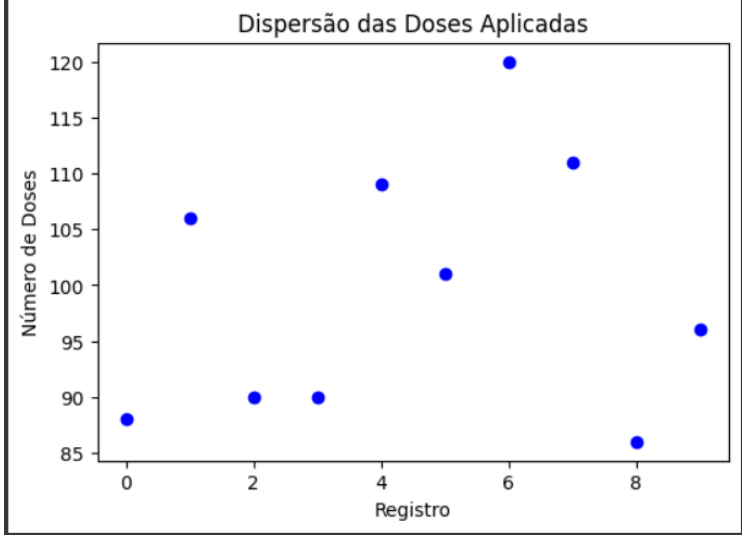
```
Média das doses aplicadas: 99.7
Mediana das doses aplicadas: 98.5
Moda das doses aplicadas: 90

Análise sobre as doses aplicadas:
A média indica que, em média, foram aplicadas 99.7 doses por registro.
A mediana (98.5) mostra o valor central, sugerindo que metade dos registros estão abaixo e metade acima.
A moda (90) revela que o valor mais frequente de doses aplicadas foi 90, indicando uma possível concentração nessa faixa.
```

Visualizações e Interpretações: As medidas de tendência central: média, mediana e moda foram calculadas para a variável "Doses". A visualização dos resultados destaca diferentes aspectos da distribuição dos dados, mostrando o valor médio, o ponto central e o valor mais frequente das doses aplicadas.

Discussão dos Resultados Obtidos: Compreender essas medidas é fundamental para interpretar corretamente os dados, pois cada uma revela uma característica distinta da distribuição: a média indica o valor geral, a mediana mostra a posição central e a moda aponta a ocorrência mais comum. Isso ajuda a identificar tendências e possíveis assimetrias nos dados.

6.4 Medidas de Dispersão:

```
Variância: 132.68
Desvio Padrão: 11.52
Amplitude: 34
```

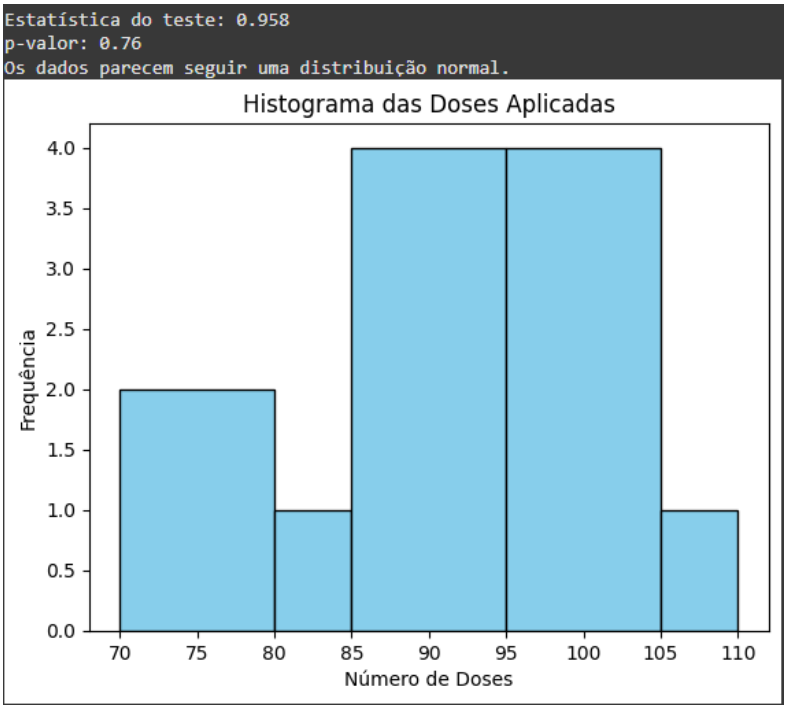


Visualizações e Interpretações: As medidas de dispersão: variância, desvio padrão e

amplitude foram calculadas para a variável "Doses". A visualização por gráfico de dispersão mostra como os valores das doses se distribuem entre os registros, evidenciando a variabilidade dos dados.

Discussão dos Resultados Obtidos: Entender as medidas de dispersão é essencial para avaliar a variabilidade dos dados em relação à média. Enquanto a variância e o desvio padrão indicam o grau de dispersão dos valores, a amplitude mostra a diferença entre o valor máximo e mínimo. Essas informações ajudam a identificar a consistência e a dispersão das doses aplicadas.

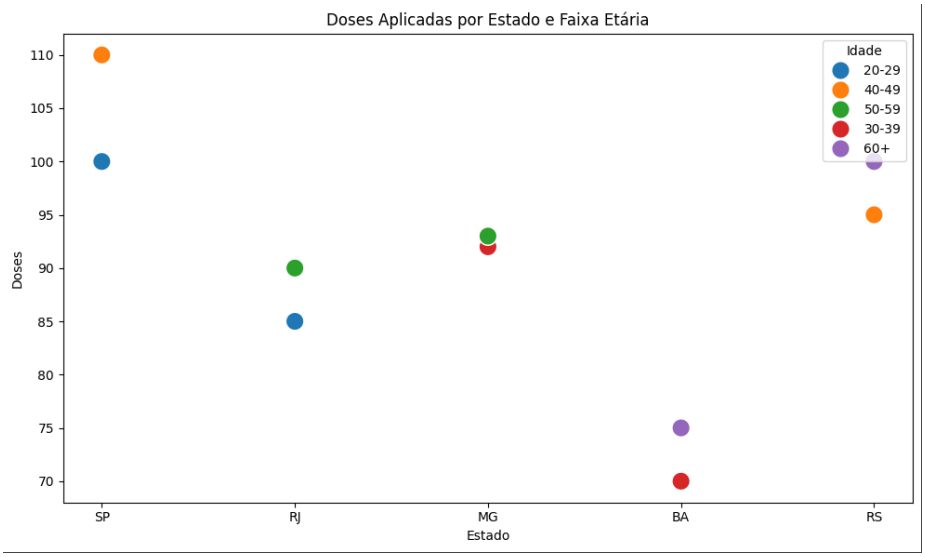
6.5 Testes de Normalidade:



Visualizações e Interpretações: O histograma das doses aplicadas, acompanhado da curva KDE, permitiu visualizar a forma da distribuição dos dados. Os testes de normalidade Shapiro-Wilk e Kolmogorov-Smirnov foram aplicados para verificar se os dados seguem uma distribuição normal.

Discussão dos Resultados Obtidos: Os testes indicam se os dados se ajustam ao modelo de distribuição normal, fundamental para a escolha de técnicas estatísticas adequadas. Um p-valor maior que 0,05 sugere que os dados podem ser considerados normais, enquanto valores menores indicam desvio da normalidade. Compreender essa característica ajuda a garantir análises estatísticas mais precisas e confiáveis.

6.6 Visualizações com Gráficos Estatísticos:



Visualizações e Interpretações: O gráfico de dispersão apresenta as doses aplicadas de vacina distribuídas por estado e faixa etária. Cada ponto representa uma combinação específica de estado e grupo etário, com a posição no eixo vertical indicando a quantidade de doses aplicadas. As cores dos pontos correspondem às diferentes faixas etárias, facilitando a comparação visual entre os grupos. Esse tipo de visualização ajuda a identificar onde a vacinação foi mais concentrada e permite perceber diferenças na quantidade de doses aplicadas por faixa etária em cada estado.

Discussão dos Resultados Obtidos: A análise do gráfico revela que alguns estados, como SP e RS, possuem maiores concentrações de doses aplicadas em faixas etárias específicas, por exemplo, nas faixas 40-49 e 60+. Em contrapartida, estados como BA apresentam menores quantidades em algumas faixas. Essa variação pode indicar diferenças regionais na adesão à vacinação ou na disponibilidade das doses para diferentes grupos etários. Compreender essas variações é essencial para planejar ações mais direcionadas, garantindo que grupos menos atendidos possam ser priorizados em futuras campanhas.

7. CONCLUSÃO

7.1 Tipos de Amostragem

A utilização da amostragem estratificada na análise estatística dos dados de vacinação no Brasil garantiu equilíbrio entre os estados, permitindo uma comparação justa entre eles. A limitação foi não representar proporcionalmente a população real de cada estado. Recomenda-se, em análises futuras, utilizar amostragem proporcional e considerar mais variáveis, como gênero e renda, para aprofundar os resultados.

7.2 Escalas de medição

A correta aplicação das escalas nominal, ordinal e de razão foi essencial para classificar os dados e conduzir as análises. Contudo, a limitação está na quantidade reduzida de variáveis, o que restringiu o uso de escalas mais complexas. Sugere-se inserir mais informações, como datas e características socioeconômicas, para enriquecer as análises.

7.3 Medidas de Tendência Central

As medidas de média, mediana e moda foram fundamentais para resumir a distribuição das doses, apontando certa simetria nos dados. A limitação foi o tamanho reduzido da amostra, que pode ser sensível a outliers. Futuramente, recomenda-se ampliar a base e segmentar por estado e faixa etária para análises mais detalhadas.

7.4 Medidas de Dispersão

As análises de dispersão revelaram uma variabilidade significativa nas doses aplicadas, evidenciando que há diferenças relevantes entre os registros. Entretanto, a ausência de segmentação por estado e faixa etária limita a identificação precisa dos grupos mais dispersos. Recomenda-se aplicar essas medidas de forma estratificada nas próximas análises.

7.5 Testes de Normalidade Os testes indicaram que os dados seguem aproximadamente uma distribuição normal, viabilizando o uso de métodos estatísticos paramétricos. Como limitação, destaca-se a amostra pequena e não segmentada, que pode afetar a precisão dos resultados. Sugere-se aumentar a amostra e aplicar os testes por estado e faixa etária.

7.6 Visualizações com Gráficos Estatísticos

As visualizações evidenciaram padrões importantes na vacinação no Brasil, como maiores aplicações em SP e RS, especialmente em grupos etários mais elevados, e menores em BA. A principal limitação foi não considerar a proporção populacional de cada estado. Sugere-se integrar dados populacionais e utilizar gráficos mais dinâmicos e interativos em futuras análises.

8. **REFERÊNCIAS**

BRASIL. Ministério da Saúde. **DATASUS: Departamento de Informática do Sistema Único de Saúde**. Brasília, DF, 2025. Disponível em: https://datasus.saude.gov.br/. Acesso em: junho 2025.

BRASIL. Ministério da Saúde. **Open DataSUS: Plataforma de Dados Abertos do Sistema Único de Saúde (SUS)**. Brasília, DF, 2025. Disponível em: https://opendatasus.saude.gov.br/. Acesso em: junho 2025.

GOOGLE. **Google Colaboratory**. Mountain View, CA, 2025. Disponível em: https://colab.research.google.com/. Acesso em: junho 2025

9. APÊNDICE A – PROCEDIMENTOS E ANÁLISES ESTATÍSTICAS

A.1 Coleta e Preparação dos Dados

Os dados foram obtidos na plataforma OpenDataSUS, do Ministério da Saúde, contendo informações sobre tipos de vacinas, datas de aplicação, número de doses, regiões e faixas etárias.

A preparação dos dados incluiu:

- Limpeza para remoção de registros incompletos ou inconsistentes.
- Padronização dos formatos das variáveis (datas, categorias, valores numéricos).
- Realização de uma amostragem estratificada para garantir representação equilibrada entre os estados selecionados.

A.2 Análise Estatística

Foram aplicadas as seguintes técnicas estatísticas:

- Cálculo de medidas de tendência central (média, mediana, moda) para entender o comportamento geral das doses aplicadas.
- Medidas de dispersão (desvio padrão, variância e amplitude) para avaliar a variabilidade dos dados.
- Testes de normalidade (Shapiro-Wilk e Kolmogorov-Smirnov) para verificar a adequação dos dados a distribuições paramétricas.
- Visualizações gráficas como histogramas, gráficos de dispersão e boxplots para melhor interpretação dos resultados.

A.3 Ferramentas Utilizadas As análises foram desenvolvidas em Python, utilizando as bibliotecas Pandas, NumPy, Matplotlib, Seaborn, SciPy e Statsmodels. A plataforma Google Colab foi utilizada para execução dos códigos e geração dos gráficos.

A.4 Resultados

As análises revelaram que os dados de vacinação apresentaram distribuição aproximadamente normal, com variações regionais significativas na quantidade de doses aplicadas, especialmente entre os estados SP, RS e BA. As visualizações

evidenciaram concentrações maiores em faixas etárias específicas e diferenças relevantes entre regiões.

10. APÊNDICE B – IMAGENS DOS EXERCÍCIOS REALIZADOS

Este apêndice reúne capturas de tela das saídas obtidas a partir da execução dos códigos Python desenvolvidos para análise estatística dos dados de vacinação.

As imagens servem como comprovação visual dos procedimentos realizados e complementam os resultados descritos nas seções anteriores.

Figura B.1 – Tipos de Amostragem

```python
import pandas as pd

dados = {
    'Estado': ['RJ', 'BA', 'MG', 'MG', 'RJ', 'RS', 'RS', 'RS', 'BA', 'RJ'],
    'Doses': [88, 106, 90, 90, 109, 101, 120, 111, 86, 96],
    'Faixa_Etaria': ['18-39', '60+', '60+', '0-17', '18-39', '40-59', '0-17', '0-17', '60+', '60+']
}


df = pd.DataFrame(dados)


print("📊 Dados brutos (como se fosse a planilha completa):")
print(df)


print("\n📈 Média de doses aplicadas por Estado:")
print(df.groupby('Estado')['Doses'].mean())


print("\n📉 Total de doses aplicadas por faixa etária:")
print(df.groupby('Faixa_Etaria')['Doses'].sum())
```

Figura B.2 – Escalas de Medição:

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt


dados = {
    'Estado': ['BA', 'BA', 'BA', 'BA', 'MG', 'MG', 'MG', 'MG', 'RJ', 'RJ', 'RJ', 'RJ', 'RS', 'RS', 'RS', 'RS', 'SP', 'SP', 'SP', 'SP'],
    'Faixa_Etaria': ['0-17', '18-39', '40-59', '60+'] * 5,
    'Doses': [102, 105, 103, 104, 101, 104, 100, 102, 103, 106, 104, 102, 105, 107, 106, 105, 103, 105, 104, 102]
}


df = pd.DataFrame(dados)


plt.figure(figsize=(10, 6))
sns.barplot(data=df, x='Estado', y='Doses', hue='Faixa_Etaria')

plt.title('Média de Doses por Estado e Faixa Etária (Escalas de Medição)')
plt.xlabel('Estado (Nominal)')
plt.ylabel('Média de Doses (Escala)')


plt.show()
```

Figura B.3 – Escalas de Medição:

```python
import pandas as pd
import statistics

dados = {
    'Estado': ['RJ', 'BA', 'MG', 'MG', 'RJ', 'RS', 'RS', 'RS', 'BA', 'RJ'],
    'Doses': [88, 106, 90, 90, 109, 101, 120, 111, 86, 96],
    'Faixa_Etaria': ['18-39', '60+', '60+', '0-17', '18-39', '40-59', '0-17', '0-17', '60+', '60+']
}

df = pd.DataFrame(dados)

media = df['Doses'].mean()
mediana = df['Doses'].median()
moda = statistics.mode(df['Doses'])


print("Média das doses aplicadas:", round(media, 2))
print("Mediana das doses aplicadas:", mediana)
print("Moda das doses aplicadas:", moda)

print("\nAnálise sobre as doses aplicadas:")
print(f"A média indica que, em média, foram aplicadas {round(media, 2)} doses por registro.")
print(f"A mediana ({mediana}) mostra o valor central, sugerindo que metade dos registros estão abaixo e metade acima.")
print(f"A moda ({moda}) revela que o valor mais frequente de doses aplicadas foi {moda}, indicando uma possível concentração nessa faixa.")
```

Figura B.4 – Medidas de Dispersão:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


dados = {
    'Doses': [88, 106, 90, 90, 109, 101, 120, 111, 86, 96]
}

df = pd.DataFrame(dados)


variancia = np.var(df['Doses'], ddof=1)
desvio_padrao = np.std(df['Doses'], ddof=1)
amplitude = df['Doses'].max() - df['Doses'].min()


print("Variância:", round(variancia, 2))
print("Desvio Padrão:", round(desvio_padrao, 2))
print("Amplitude:", amplitude)

plt.figure(figsize=(6, 4))
plt.scatter(df.index, df['Doses'], color='blue')
plt.title('Dispersão das Doses Aplicadas')
plt.xlabel('Registro')
plt.ylabel('Número de Doses')
plt.show()
```

Figura B.5 – Testes de Normalidade:

```python
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import shapiro

dados = [72, 75, 82, 88, 90, 91, 93, 95, 97, 98, 100, 107]


plt.hist(dados, bins=[70, 80, 85, 95, 105, 110], color='skyblue', edgecolor='black')
plt.title('Histograma das Doses Aplicadas')
plt.xlabel('Número de Doses')
plt.ylabel('Frequência')


estatistica, p_valor = shapiro(dados)


print(f'Estatística do teste: {estatistica:.3f}')
print(f'p-valor: {p_valor:.2f}')

if p_valor > 0.05:
    print('Os dados parecem seguir uma distribuição normal.')
else:
    print('Os dados não parecem seguir uma distribuição normal.')

plt.show()
```

Figura B.6 – Visualizações com Gráficos Estatísticos:

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd


data = {
    'Estado': ['SP', 'SP', 'RJ', 'RJ', 'MG', 'MG', 'BA', 'BA', 'RS', 'RS'],
    'Doses': [100, 110, 85, 90, 92, 93, 70, 75, 95, 100],
    'Idade': ['20-29', '40-49', '20-29', '50-59', '30-39', '50-59', '30-39', '60+', '40-49', '60+']
}

df = pd.DataFrame(data)


plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Estado', y='Doses', hue='Idade', s=200)

plt.title('Doses Aplicadas por Estado e Faixa Etária')
plt.ylabel('Doses')
plt.xlabel('Estado')
plt.legend(title='Idade')
plt.grid(False)
plt.tight_layout()
plt.show()
```