JANEIRO DE 2025

PROJETO DE ANÁLISE DE DADOS

Este projeto tem como objetivo analisar os microdados do ENEM 2023, fornecendo insights sobre o desempenho dos candidatos em diferentes áreas do conhecimento.

enem2023

CADERNO 1 AZUL

ATENÇÃO: transcreva no espaço apropriado do seu CARTÃO-RESPOSTA, com sua caligrafia usual, considerando as letras maiúsculas e minúsculas, a seguinte frase:

Autor: Lilian Rondan

LEIA ATENTAMENTE AS INSTRUÇÕES SEGUINTES:

Este Projeto contém 10 páginas com análises dispostas da seguinte maneira:

1. Visão Geral – Apresentação do objetivo da análise, destacando os principais aspectos abordados.
2. Ferramentas e Bibliotecas – Listagem dos recursos utilizados.
3. Estrutura da Análise – Explicação das etapas de carregamento, inspeção e tratamento dos dados.
4. Distribuição das Notas – Visualização da frequência das notas, com interpretação dos padrões observados.
5. Correlação entre Áreas de Conhecimento – Análise das relações entre as notas por meio de um mapa de calor.
6. Desempenho por Sexo – Comparação das médias de notas entre homens e mulheres.
7. Desempenho por Tipo de Escola – Comparação das notas de alunos de escolas públicas e privadas.
8. Desempenho por Região Geográfica – Análise das médias das notas por região..
9. Análise da Nota Final – Correlação com as diferentes áreas do exame.
10. Considerações Finais – Reflexões sobre os resultados encontrados.
11. Referências – Lista de fontes utilizadas.

Repositório: GitHub

Visão Geral

Este projeto tem como objetivo analisar os microdados do ENEM 2023, fornecendo insights sobre o desempenho dos candidatos em diferentes áreas do conhecimento, como Matemática, Linguagens e Redação. A análise inclui a distribuição das notas, correlação entre as áreas, e o impacto de variáveis como sexo, tipo de escola e região geográfica no desempenho dos alunos.

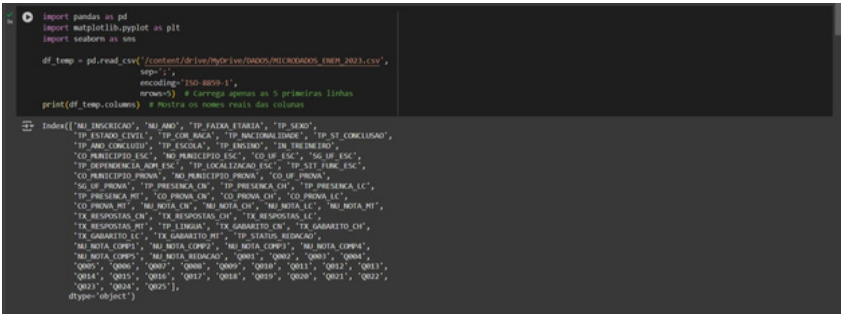
Ferramentas e Bibliotecas:

Python – Linguagem de programação principal
Pandas e NumPy – Para manipulação e análise de dados
Matplotlib e Seaborn – Para visualização de dados estatísticos
Google Colab – Ambiente para execução do código e análise
Google Drive – Para armazenamento dos dados

Estrutura do Projeto

1. Carregamento dos Dados

Os dados são carregados a partir de um arquivo CSV armazenado no Google Drive. Apenas as primeiras 5 linhas são carregadas inicialmente para inspeção das colunas.



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

df_temp = pd.read_csv('/content/drive/MyDrive/DADOS/MICRODADOS_ENEM_2023.csv',
                      sep=';',
                      encoding='ISO-8859-1',
                      nrows=5)  # Carrega apenas as 5 primeiras linhas
print(df_temp.columns)  # Mostra os nomes reais das colunas
```

```
Index(['NU_INSCRICAO', 'NU_ANO', 'TP_FAIXA_ETARIA', 'TP_SEXO',
       'TP_ESTADO_CIVIL', 'TP_COR_RACA', 'TP_NACIONALIDADE', 'TP_ST_CONCLUSAO',
       'TP_ANO_CONCLUIU', 'TP_ESCOLA', 'TP_ENSINO', 'IN_TREINEIRO',
       'CO_MUNICIPIO_ESC', 'NO_MUNICIPIO_ESC', 'CO_UF_ESC', 'SG_UF_ESC',
       'TP_DEPENDENCIA_ADM_ESC', 'TP_LOCALIZACAO_ESC', 'TP_SIT_FUNC_ESC',
       'CO_MUNICIPIO_PROVA', 'NO_MUNICIPIO_PROVA', 'CO_UF_PROVA',
       'SG_UF_PROVA', 'TP_PRESENCA_CN', 'TP_PRESENCA_CH', 'TP_PRESENCA_LC',
       'TP_PRESENCA_MT', 'CO_PROVA_CN', 'CO_PROVA_CH', 'CO_PROVA_LC',
       'CO_PROVA_MT', 'NU_NOTA_CN', 'NU_NOTA_CH', 'NU_NOTA_LC', 'NU_NOTA_MT',
       'TX_RESPOSTAS_CN', 'TX_RESPOSTAS_CH', 'TX_RESPOSTAS_LC',
       'TX_RESPOSTAS_MT', 'TP_LINGUA', 'TX_GABARITO_CN', 'TX_GABARITO_CH',
       'TX_GABARITO_LC', 'TX_GABARITO_MT', 'TP_STATUS_REDACAO',
       'NU_NOTA_COMP1', 'NU_NOTA_COMP2', 'NU_NOTA_COMP3', 'NU_NOTA_COMP4',
       'NU_NOTA_COMP5', 'NU_NOTA_REDACAO', 'Q001', 'Q002', 'Q003', 'Q004',
       'Q005', 'Q006', 'Q007', 'Q008', 'Q009', 'Q010', 'Q011', 'Q012', 'Q013',
       'Q014', 'Q015', 'Q016', 'Q017', 'Q018', 'Q019', 'Q020', 'Q021', 'Q022',
       'Q023', 'Q024', 'Q025'],
      dtype='object')
```

2. Configuração e Carregamento Completo dos Dados

Após inspecionar as colunas, o conjunto completo de dados é carregado, selecionando apenas as colunas relevantes para a análise.

```
[ ] # Configurar estilo dos gráficos
    sns.set(style="whitegrid")

    # Caminho do arquivo
    caminho_arquivo = '/content/drive/MyDrive/DADOS/MICRODADOS_ENEM_2023.csv'

    # Definição das colunas necessárias
    colunas = [
        'NU_INSCRICAO', 'TP_SEXO', 'TP_ESCOLA',
        'NU_NOTA_MT', 'NU_NOTA_LC', 'NU_NOTA_REDACAO', 'SG_UF_ESC', 'SG_UF_PROVA'
    ]

    # Carregar os dados
    df = pd.read_csv(caminho_arquivo, sep=';', usecols=colunas, encoding='ISO-8859-1')
```

3. Renomeação e Limpeza dos Dados

As colunas são renomeadas para facilitar a análise e os valores são substituídos por categorias mais legíveis.

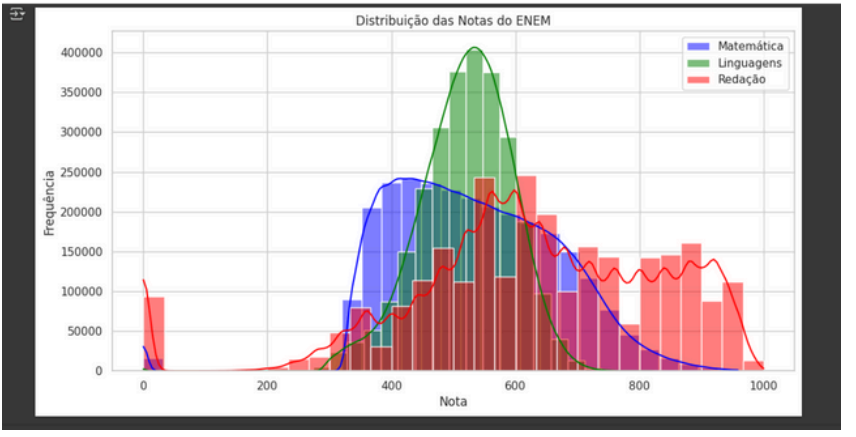
```
    # Renomear colunas para facilitar
    df.rename(columns={
        'TP_SEXO': 'Sexo',
        'TP_ESCOLA': 'Tipo_Escola',
        'NU_NOTA_MT': 'Nota_Matematica',
        'NU_NOTA_LC': 'Nota_Linguagens',
        'NU_NOTA_REDACAO': 'Nota_Redacao',
        'SG_UF_ESC': 'UF',
        'SG_UF_PROVA': 'UF_Prova'

    }, inplace=True)

[11] # Substituir códigos por valores mais simples
    df['Sexo'].replace({1: 'Masculino', 2: 'Feminino'}, inplace=True)
    df['Tipo_Escola'].replace({1: 'Não Respondeu', 2: 'Pública', 3: 'Privada'}, inplace=True)

[12] # Remover valores nulos nas notas
    df.dropna(subset=['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao'], inplace=True)

    # Estatísticas básicas das notas
    print(df[['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao']].describe())
```

4. Análise Exploratória dos Dados

4.1. Distribuição das Notas

A distribuição das notas é visualizada através de histogramas.

```
# 1. Distribuição das notas
plt.figure(figsize=(12, 6))
sns.histplot(df['Nota_Matematica'], bins=30, kde=True, color='blue', label="Matemática")
sns.histplot(df['Nota_Linguagens'], bins=30, kde=True, color='green', label="Linguagens")
sns.histplot(df['Nota_Redacao'], bins=30, kde=True, color='red', label="Redação")
plt.legend()
plt.title("Distribuição das Notas do ENEM")
plt.xlabel("Nota")
plt.ylabel("Frequência")
plt.show()
```



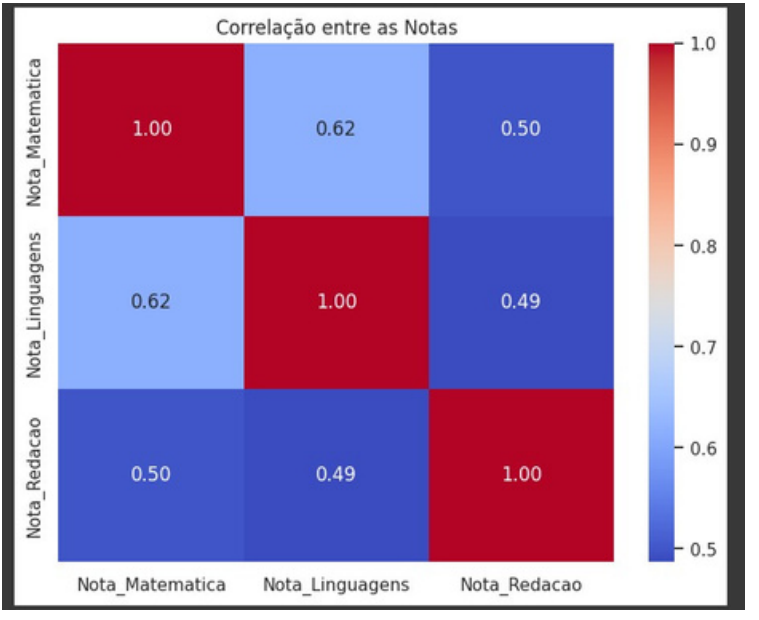
Interpretação

Linguagens (Maior frequência em torno de 500) A maior parte dos alunos está concentrada em torno de 500 pontos em Linguagens.
Isso pode indicar que essa é a nota média ou comum entre os participantes do ENEM. Pode sugerir que, para muitos alunos, a compreensão e produção de textos em Linguagens não foi particularmente difícil nem fácil, sendo uma prova mais equilibrada. Matemática (Maior frequência abaixo de 500) A maioria dos alunos tem notas abaixo de 500 em Matemática, o que pode ser um
reflexo das dificuldades maiores com a matéria. Isso pode sugerir que muitos alunos não têm um bom desempenho em Matemática, o que é comum em muitas avaliações devido à natureza das questões, que envolvem mais raciocínio lógico e resolução de problemas. Redação (Oscilação grande, mas fica acima de 500) As notas de Redação variam bastante, mas a maioria está acima de 500, o que pode
indicar que muitos alunos conseguiram apresentar uma boa argumentação e estrutura textual. A oscilação pode ser causada por diversidade nas habilidades de escrita dos alunos, já que a redação avalia tanto a coerência e coesão quanto a capacidade de argumentação.

4.2. Correlação entre as Notas

A correlação entre as notas é analisada através de um mapa de calor.

```
[14] # 2. Correlação entre as notas
     plt.figure(figsize=(6, 6))
     sns.heatmap(df[['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao']].corr(), annot=True, cmap="coolwarm", fmt=".2f")
     plt.title("Correlação entre as Notas")
     plt.show()
```



Interpretação

Matemática e Linguagens: A relação mais forte entre essas duas áreas pode ser atribuída ao fato de que ambas exigem habilidades de raciocínio lógico e interpretação.

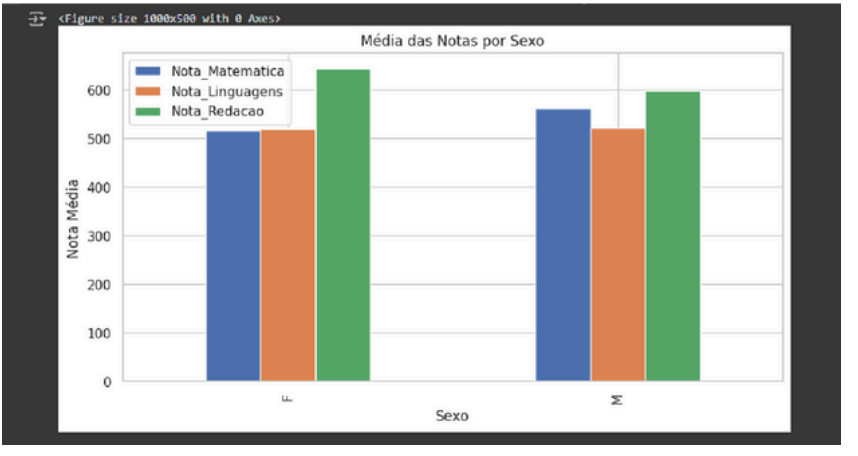
Matemática e Redação: A relação mais fraca aqui é compreensível, pois as habilidades necessárias para se destacar em Matemática (pensamento abstrato, cálculo, resolução de problemas) são bastante diferentes das necessárias para a Redação (criatividade, organização de ideias, domínio da língua escrita).

Linguagens e Redação: A relação moderada entre essas duas áreas é esperada, já que ambas envolvem competências relacionadas à leitura e à escrita. A capacidade de compreender textos complexos e expressar ideias de forma clara e coerente é crucial tanto em Linguagens quanto em Redação. No entanto, a Redação também exige criatividade e originalidade que pode não ser tão diretamente avaliada em Linguagens.

4.3. Média das Notas por Sexo

A média das notas é comparada por sexo através de um gráfico de colunas.

```
[16] # 3. Média das notas por Sexo
     plt.figure(figsize=(10, 5))
     df.groupby("Sexo")[['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao']].mean().plot(kind='bar', figsize=(10,5))
     plt.title("Média das Notas por Sexo")
     plt.ylabel("Nota Média")
     plt.show()
```



Interpretação

Mulheres demonstram um desempenho superior em áreas que exigem habilidades linguísticas, como Redação e Linguagens, refletindo uma possível maior afinidade com a expressão verbal e escrita.

A diferença em Linguagens e Redação entre os sexos é menos notável, mas as mulheres ainda mostram uma leve vantagem em ambas as áreas.

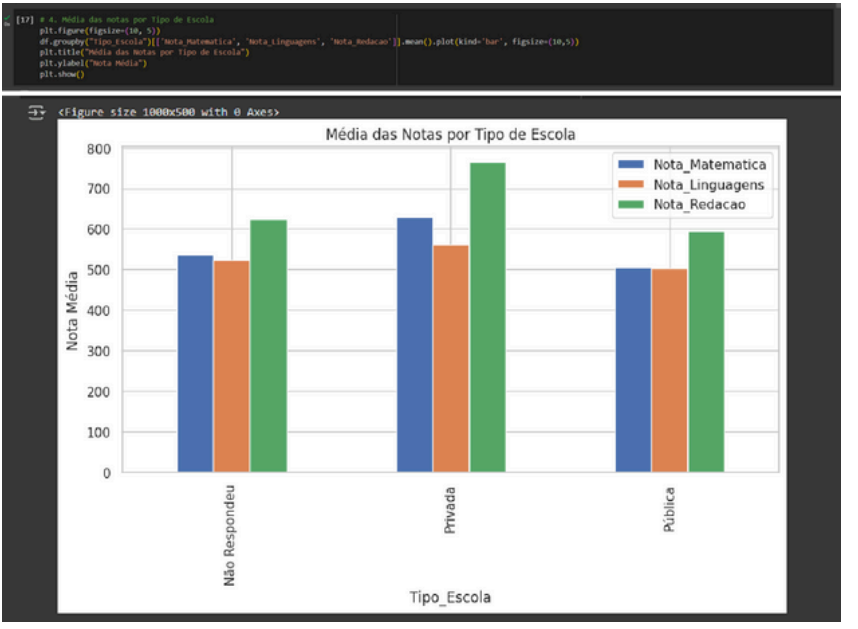
Homens têm uma superioridade mais expressiva em Matemática, sugerindo que a área de raciocínio lógico e analítico é um ponto forte para eles, com uma diferença de desempenho mais acentuada em comparação com as mulheres.

Muitas vezes, as mulheres têm uma facilidade maior com palavras, seja escrevendo ou interpretando textos. Talvez porque, desde cedo, são incentivadas a se comunicar mais e a expressar emoções.

Já os homens costumam se destacar mais em Matemática, talvez porque são mais estimulados a lidar com números e lógica desde pequenos, como em brincadeiras que envolvem construção e estratégia.

4.4. Média das Notas por Tipo de Escola

A média das notas é comparada por Tipo de Escola através de um gráfico de colunas.



Interpretação

Nas escolas privadas, as notas em Linguagens, Matemática e Redação são significativamente mais altas.

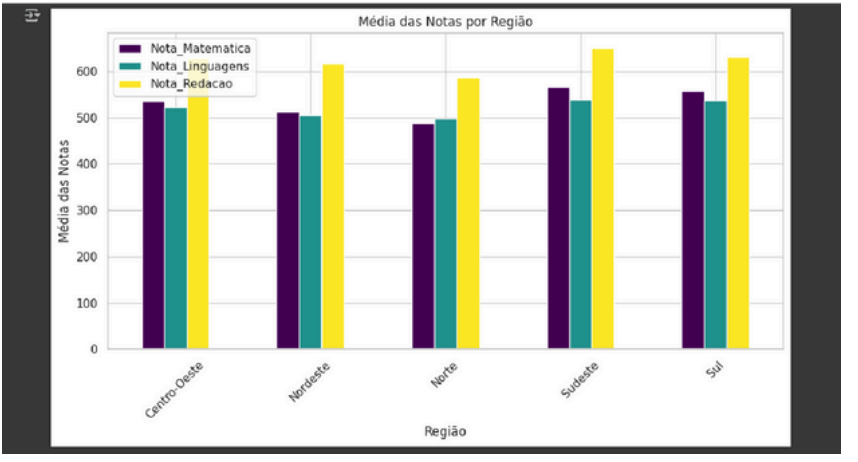
Nas escolas públicas, o desempenho é bem inferior, principalmente em Matemática e Redação, onde a diferença para as privadas é gritante. As notas em Linguagens também são bem mais baixas, mas não chegam a ser tão extremas quanto nas outras áreas.

Esses resultados mostram uma desigualdade clara entre as escolas públicas e privadas, com as privadas se destacando devido a melhores condições de ensino, enquanto as públicas enfrentam sérios desafios que afetam o desempenho dos alunos.

4.5. Média das Notas por Região

A média das notas é analisada por região geográfica através de um gráfico de colunas.

```
[18] # 5. Médias por Região
     # Mapeando os estados para as regiões
     regioes = {
         'AC': 'Norte', 'AL': 'Nordeste', 'AM': 'Norte', 'AP': 'Norte', 'BA': 'Nordeste', 'CE': 'Nordeste',
         'DF': 'Centro-Oeste', 'ES': 'Sudeste', 'GO': 'Centro-Oeste', 'MA': 'Nordeste', 'MG': 'Sudeste',
         'MS': 'Centro-Oeste', 'MT': 'Centro-Oeste', 'PA': 'Norte', 'PB': 'Nordeste', 'PE': 'Nordeste',
         'PI': 'Nordeste', 'PR': 'Sul', 'RJ': 'Sudeste', 'RN': 'Nordeste', 'RO': 'Norte', 'RR': 'Norte',
         'RS': 'Sul', 'SC': 'Sul', 'SE': 'Nordeste', 'SP': 'Sudeste', 'TO': 'Norte'
     }

     # Adicionando a coluna 'Regiao'
     df['Regiao'] = df['UF_Prova'].map(regioes)

     # Calculando a média das notas por região
     notas_por_regiao = df.groupby('Regiao')[['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao']].mean()

     # Plotando o gráfico de colunas
     notas_por_regiao.plot(kind='bar', figsize=(10, 6), cmap='viridis')

     # Ajustando o gráfico
     plt.title('Média das Notas por Região')
     plt.xlabel('Região')
     plt.ylabel('Média das Notas')
     plt.xticks(rotation=45)
     plt.tight_layout()

     # Exibindo o gráfico
     plt.show()
```



Interpretação

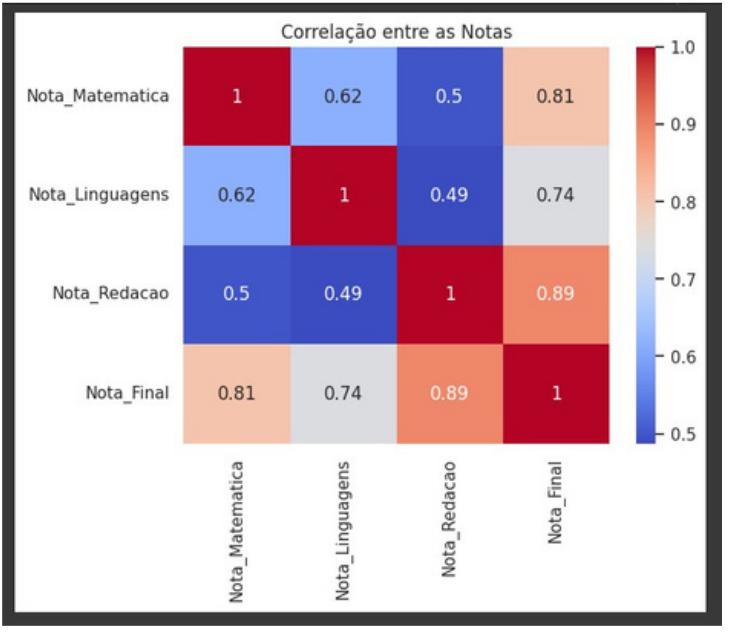
O gráfico mostra que as regiões Norte e Nordeste têm as menores médias de notas no ENEM, enquanto o Sudeste lidera, seguido pelo Sul e Centro-Oeste, que estão em um nível intermediário.

Essa distribuição reflete a desigualdade educacional no Brasil, com as regiões mais desenvolvidas apresentando melhores desempenhos. A discrepância mostra a necessidade de políticas públicas focadas na melhoria da educação nas regiões Norte e Nordeste.

5. Análise de Correlação com a Nota Final

A nota final é calculada como a média das notas de Matemática, Linguagens e Redação, e sua correlação com as demais notas é analisada através de um mapa de calor.

```
[19] # 6. Impacto da redação na nota final
     # Criar a variável de Nota Final
     df['Nota_Final'] = df[['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao']].mean(axis=1)

     # Gráfico
     sns.heatmap(df[['Nota_Matematica', 'Nota_Linguagens', 'Nota_Redacao', 'Nota_Final']].corr(), annot=True, cmap='coolwarm')
     plt.title('Correlação entre as Notas')
     plt.show()
```



Interpretação

Alto valor de correlação sugere que a nota de redação pode ser um fator determinante para o desempenho final, especialmente considerando que ela possui um peso considerável no exame. Para um desempenho final elevado, é provável que o candidato também tenha se saído bem na redação.

Conclusão

As análises mostram que o desempenho no ENEM é influenciado por fatores como sexo, tipo de escola e região geográfica. Mulheres se destacam em Linguagens e Redação, enquanto homens têm melhor desempenho em Matemática. Escolas privadas e regiões mais desenvolvidas apresentam notas mais altas, evidenciando desigualdades educacionais. A Redação tem um impacto significativo na nota final, destacando sua importância no exame. Esses insights podem orientar políticas públicas e estratégias educacionais para melhorar o desempenho dos alunos.

Referências

Documentação do Pandas: https://pandas.pydata.org/pandas-docs/stable/

Documentação do Matplotlib: https://matplotlib.org/stable/contents.html

Documentação do Seaborn: https://seaborn.pydata.org/

Microdados do ENEM: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem