**PROJETO SEMANTIX – DATA SCIENCE**

# Previsão de Notas em Exames Nacionais

# Estratégias para Educação Inclusiva.

**BASE DE DADOS - ENEM 2023.**

**GUILHERME NICOLAZ RHEIN**

**Julho/2024**

***"A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original".***

*Albert Einstein*

**SUMÁRIO**

## 1 INTRODUÇÃO

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), por intermédio da Diretoria de Avaliação da Educação Básica, em cumprimento da sua missão de desenvolver e disseminar informações sobre os exames e avaliações da educação básica, disponibiliza os Microdados do Enem 2023.

Os resultados do Enem deverão possibilitar:

I - a constituição de parâmetros para a autoavaliação do participante, com vistas à continuidade de sua formação e a sua inserção no mercado de trabalho;

II - a criação de referência nacional para o aperfeiçoamento dos currículos do ensino médio;

III - a utilização do Exame como mecanismo único, alternativo ou complementar para acesso à educação superior, especialmente a ofertada pelas instituições federais de educação superior;

IV - o acesso a programas governamentais de financiamento ou apoio ao estudante da educação;

V - a sua utilização como instrumento de seleção para ingresso nos diferentes setores do mundo do trabalho; e

VI - o desenvolvimento de estudos e indicadores sobre a educação brasileira.

Desde sua primeira edição, em 1998, até 2008, o Enem era realizado anualmente, com a aplicação de uma única prova composta por 63 questões interdisciplinares. Durante esse período, algumas instituições de Ensino Superior passaram a utilizá-lo como instrumento de seleção para o ingresso de seus estudantes.

O Exame, com 180 (cento e oitenta) questões objetivas de múltipla escolha e uma proposta de redação, passou a ser aplicado em dois dias seguidos (sábado e domingo), sendo que no primeiro dia os participantes recebiam um caderno de questões com as provas de Ciências da Natureza e suas Tecnologias e Ciências Humanas e suas Tecnologias e, no segundo, um caderno com as provas de Redação e Linguagens, Códigos e suas Tecnologias e Matemática e suas Tecnologias.

A partir de então, o Enem tornou-se uma das principais vias de acesso ao Ensino Superior público, democratizando as oportunidades e possibilitando a mobilidade acadêmica, além de continuar sendo referência para a autoavaliação dos estudantes. Passou a possibilitar, também, a certificação para conclusão do Ensino Médio, obedecendo às exigências previstas na Lei de Diretrizes e Bases da Educação Nacional (Lei 9.394/1996) para a Educação de Jovens e Adultos.

## 2 COLETA DE DADOS DO PROJETO

A educação é um dos pilares fundamentais para o desenvolvimento de qualquer sociedade. No Brasil, o Exame Nacional do Ensino Médio (ENEM) desempenha um papel crucial na avaliação do desempenho dos estudantes e também na definição de suas oportunidades acadêmicas futuras. No entanto, existe uma grande disparidade entre as notas obtidas por estudantes de diferentes perfis que envolvem diversas variáveis como:

- Renda Familiar
- Regiões do Brasil
- Políticas Públicas
- Grupos socioeconômicos
- Nível de Qualidade de Vida
- Os Diferentes tipos de Acesso à Educação Pública ou Particular.

Esses são apenas alguns dos pontos que podem causar interferência na performance do aluno. Devido a complexidade em reunir todas essas informações, utilizaremos os chamados Microdados disponibilizados pelo INEP, que realiza uma pesquisa aos participantes do Exame Nacional do Ensino Médio.

**Microdados INEP:**
Os microdados do Enem são o menor nível de desagregação de dados recolhidos por meio do exame. Eles atendem a demanda por informações específicas ao disponibilizar as provas, os gabaritos, as informações sobre os itens, as notas e o questionário respondido pelos inscritos no Enem de cunho socioeconômico.

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

**Base de Dados Utilizada:**
Os dados utilizados neste projeto são referentes ao exame aplicado em 2023, caso desejado poderá ser feito Download das informações diretamente no site citado acima, clicando em Microdados Enem 2023.

**Objetivo:**

O projeto propõe a criação de um modelo de machine learning para prever as notas do ENEM com o objetivo de observar o desempenho do estudante por meio de suas notas, assim buscando determinados padrões e insights para melhoria do ensino. A identificação de estudantes que, historicamente, apresentam notas mais baixas permitirá relacionar diversas outras informações que estão atreladas aos dados podendo fornecer diferentes medidas corretivas e maior qualidade de ensino através

dos responsáveis órgãos públicos ou até mesmo como um nível de ferramenta pessoal para ampliar sua consciência situacional e medidas pessoais que possam ser tomadas com objetivo de se destacar.

Ao prever as notas do ENEM, esperamos fornecer informações valiosas que possam ser utilizadas para antecipar as necessidades desses estudantes, possibilitando a implementação de políticas educacionais mais eficazes e condições de ensino melhoradas. O objetivo final deste projeto é contribuir para a redução das desigualdades educacionais no Brasil, oferecendo uma ferramenta poderosa para que possam promover uma educação mais equitativa e inclusiva. Através da análise preditiva, esperamos criar um impacto positivo que se reflete na melhoria do desempenho dos estudantes em grupos vulneráveis e na construção de um sistema educacional mais justo e eficiente.

O objetivo não se limita em apenas fazer a previsão das notas com base nos históricos sociais e econômicos dos alunos, mas se estende para ampliar as percepções sobre determinadas condições que afetam o desempenho do aluno, visando a conscientização para ser possível antecipar suas ações se posicionando sempre um passo à frente. O trabalho busca democratizar a educação, torná-la acessível a todos, fornecendo informações como exemplo, alguns estados que precisam de maior atenção, regiões com menores notas e que deveriam ser pontos de maior atenção do governo com objetivo de trazer melhorias com projetos inovadores de políticas públicas.

**Variáveis selecionadas:**

Podemos verificar que na base de dados o número de colunas é alto, são 76 colunas com informações diversas. Para o nosso modelo iremos selecionar as características mais informativas e relevantes para o problema em questão que será prever a nota das provas realizadas pelos alunos, sendo assim, buscamos encontrar dentro do conjunto de dados informações econômicas do estudante, informações da escola e os valores das notas obtidas. Dentre as colunas, foram selecionadas as seguintes:

- TP_FAIXA_ETARIA – Idade;
- TP_SEXO – Sexo declarado;
- TP_COR_RACA – Cor/raça;
- TP_LOCALIZACAO_ESC – Localização da escola;
- SG_UF_ESC – Sigla da Unidade da Federação da escola;
- TP_DEPENDENCIA_ADM_ESC – Dependência administrativa da escola;
- TP_ESCOLA – Tipo de escola;
- NU_NOTA_CN – Nota da prova de Ciências da Natureza;
- NU_NOTA_CH – Nota da prova de Ciências Humanas;
- NU_NOTA_LC – Nota da prova de Linguagens e Códigos;
- NU_NOTA_MT – Nota da prova de Matemática;

- NU_NOTA_REDACAO – Nota da prova de redação;
- Q001 – Até que série seu pai, ou o homem responsável por você, estudou?
- Q002 – Até que série sua mãe, ou a mulher responsável por você, estudou?
- Q005 – Incluindo você, quantas pessoas moram atualmente em sua residência?
- Q006 – Qual é a renda mensal de sua família?
- Q010 – Na sua residência tem carro?
- Q011 – Na sua residência tem motocicleta?
- Q025 – Na sua residência tem acesso à Internet?

## 3 ANÁLISE DOS DADOS

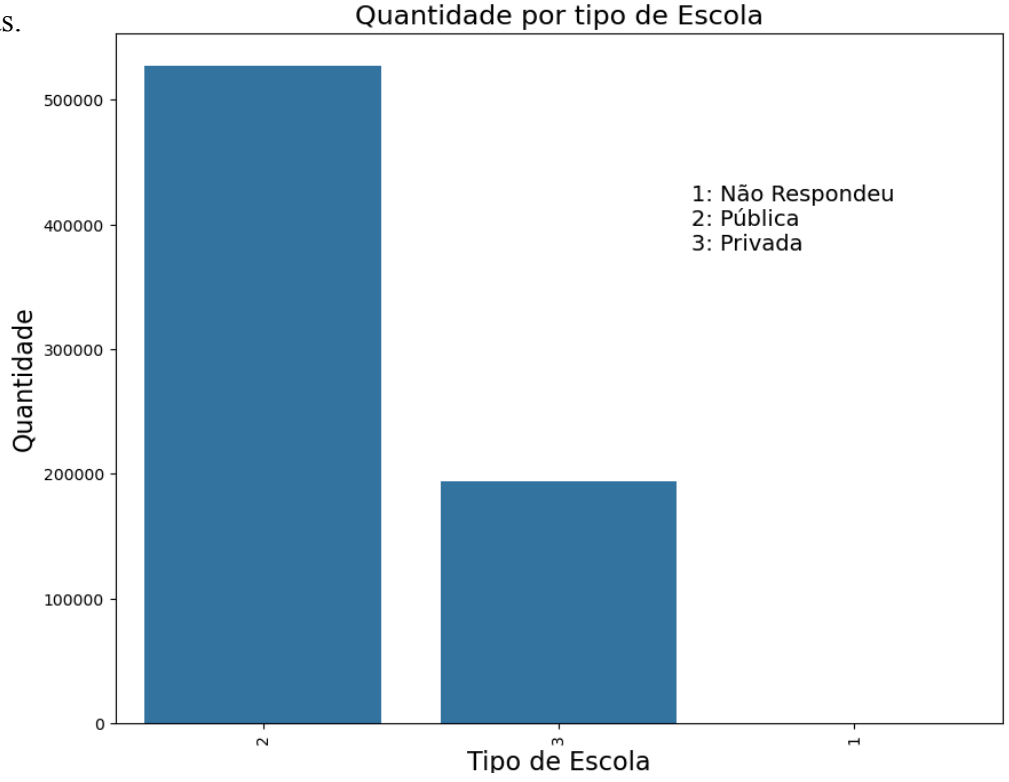
A análise descritiva é crucial para entender os dados antes de avançar para análises mais complexas ou inferenciais. Ela ajuda a identificar padrões, anomalias e características importantes que podem influenciar decisões futuras.

A criação de gráficos e a visualização de dados são cruciais para o entendimento das informações porque transformam dados complexos e muitas vezes abstratos em representações visuais claras e intuitivas.

### 3.1 Análise de variáveis categóricas:

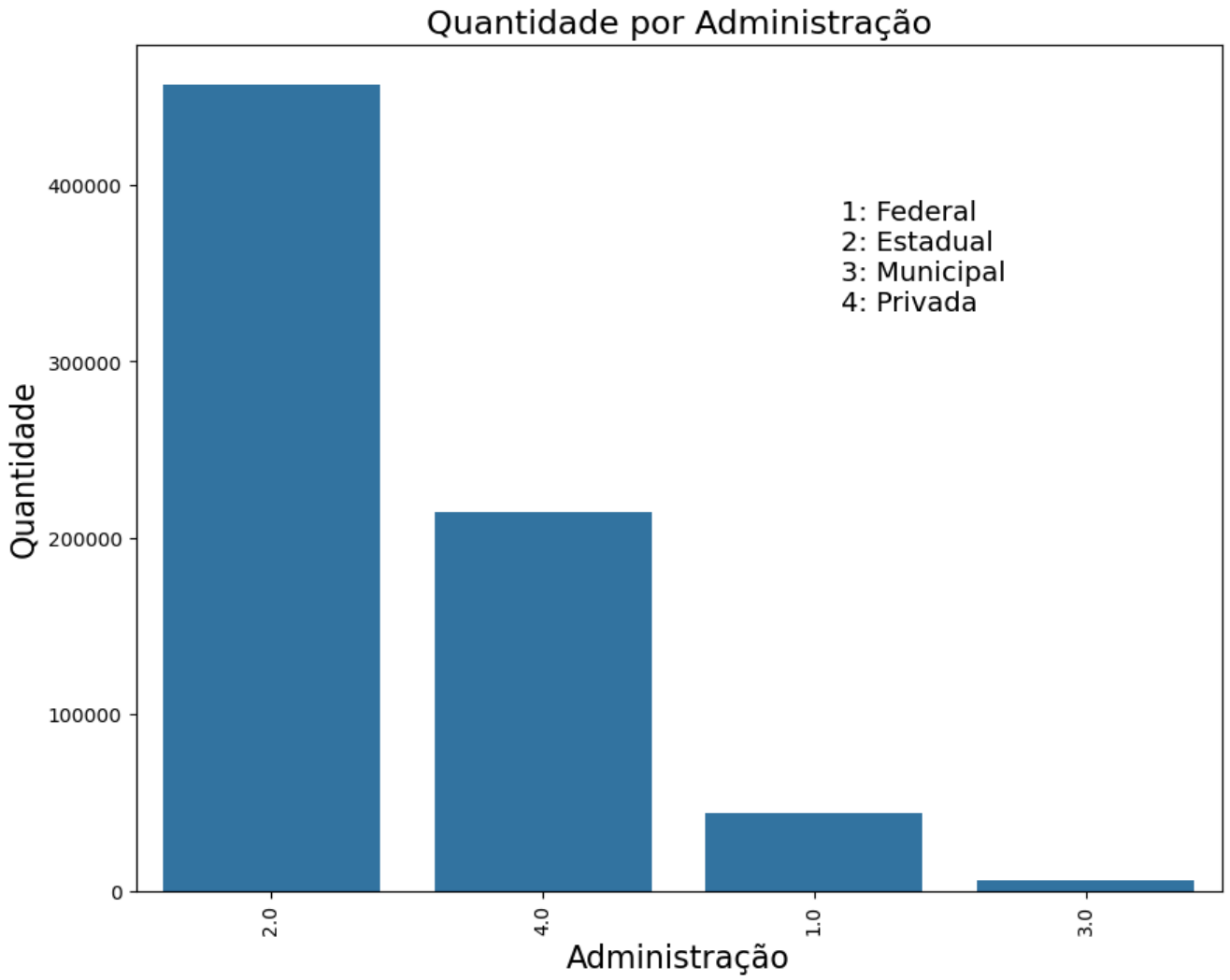
**Quais tipos de escolas aparecem em maior quantidade?**

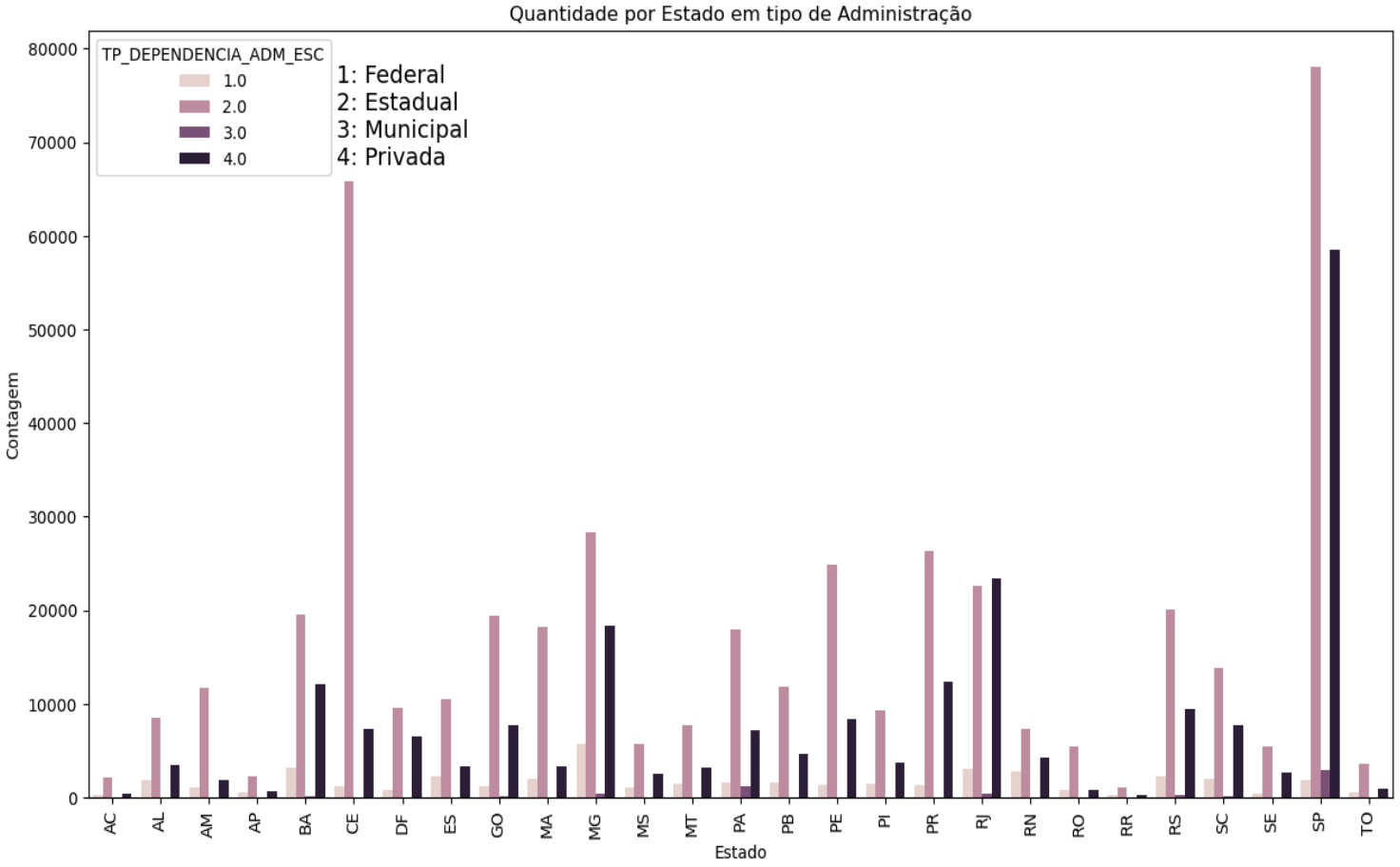
O tipo de escola em maior quantidade é a escola pública, sendo mais que o dobro do número de escolas privadas.



**Quais tipos de administrações aparecem em maior quantidade?**

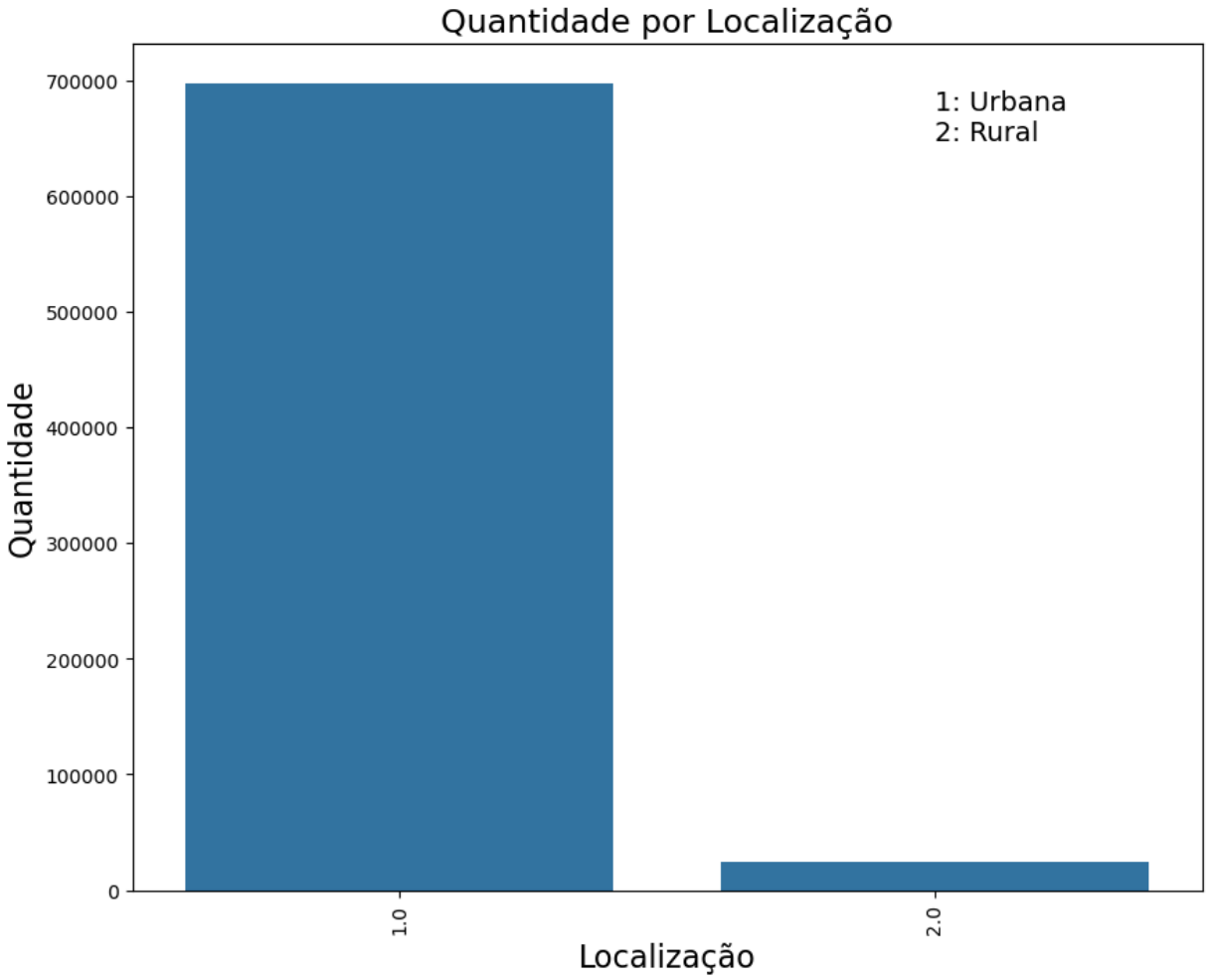
O tipo de administração em maior quantidade é Estadual, seguido da administração Privada. Nesta mesma ordem com menor quantidade seguem administração Federal e Municipal.




Quantidade por Estado em tipo de Administração
TP_DEPENDENCIA_ADM_ESC
1.0
2.0
3.0
4.0
1: Federal
2: Estadual
3: Municipal
4: Privada
Contagem
80000
70000
60000
50000
40000
30000
20000
10000
0
AC
AL
AM
AP
BA
CE
DF
ES
GO
MA
MG
MS
MT
PA
PB
PE
PI
PR
RJ
RN
RO
RR
RS
SC
SE
SP
TO
Estado

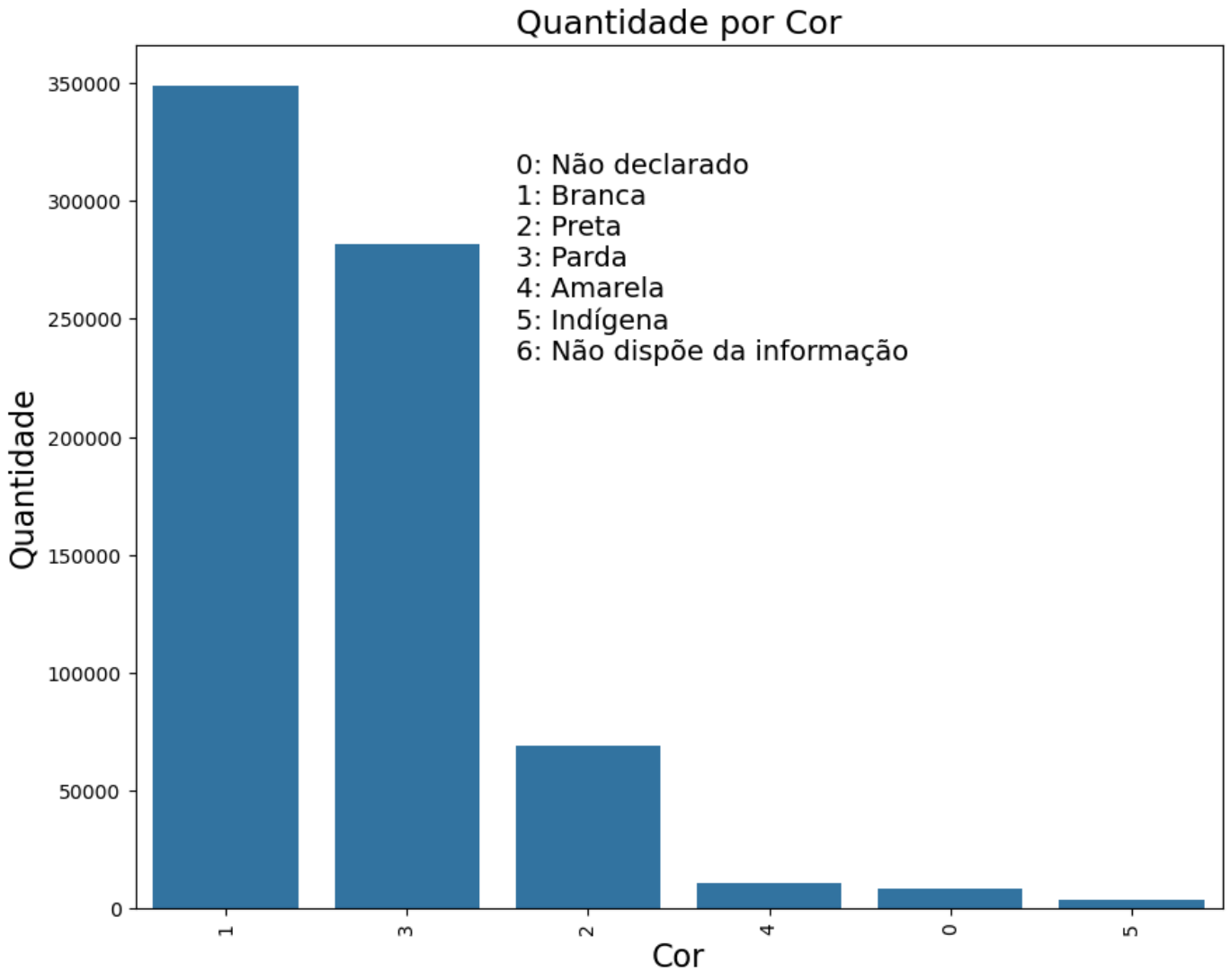
**Quais dos tipos de localização aparecem em maior quantidade?**

Como já era previsto, a quantidade de escolas localizadas na zona urbana é muito maior que as que possuem localização em área rural. Podemos destacar que escolas públicas estão em maior número que escolas privadas, assim como a administração estadual está em maior número. Dentre as regiões escolares, podemos destacar a zona urbana como principal. Podemos também nos perguntar se maior número de escolas sendo públicas realmente consegue entregar uma boa qualidade de ensino, ou será apenas quantidade?



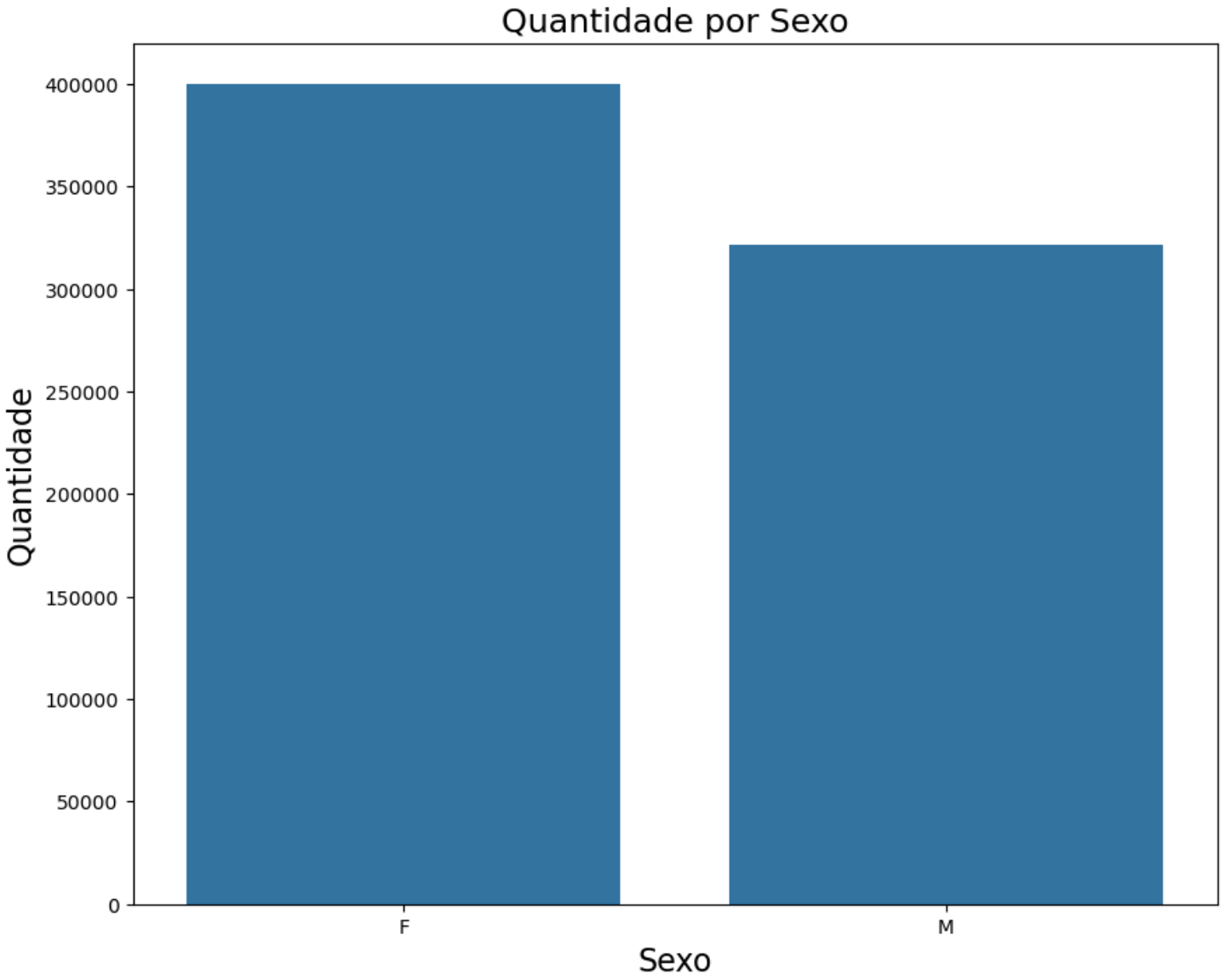
**Quais cores de pele estão em maior quantidade?**

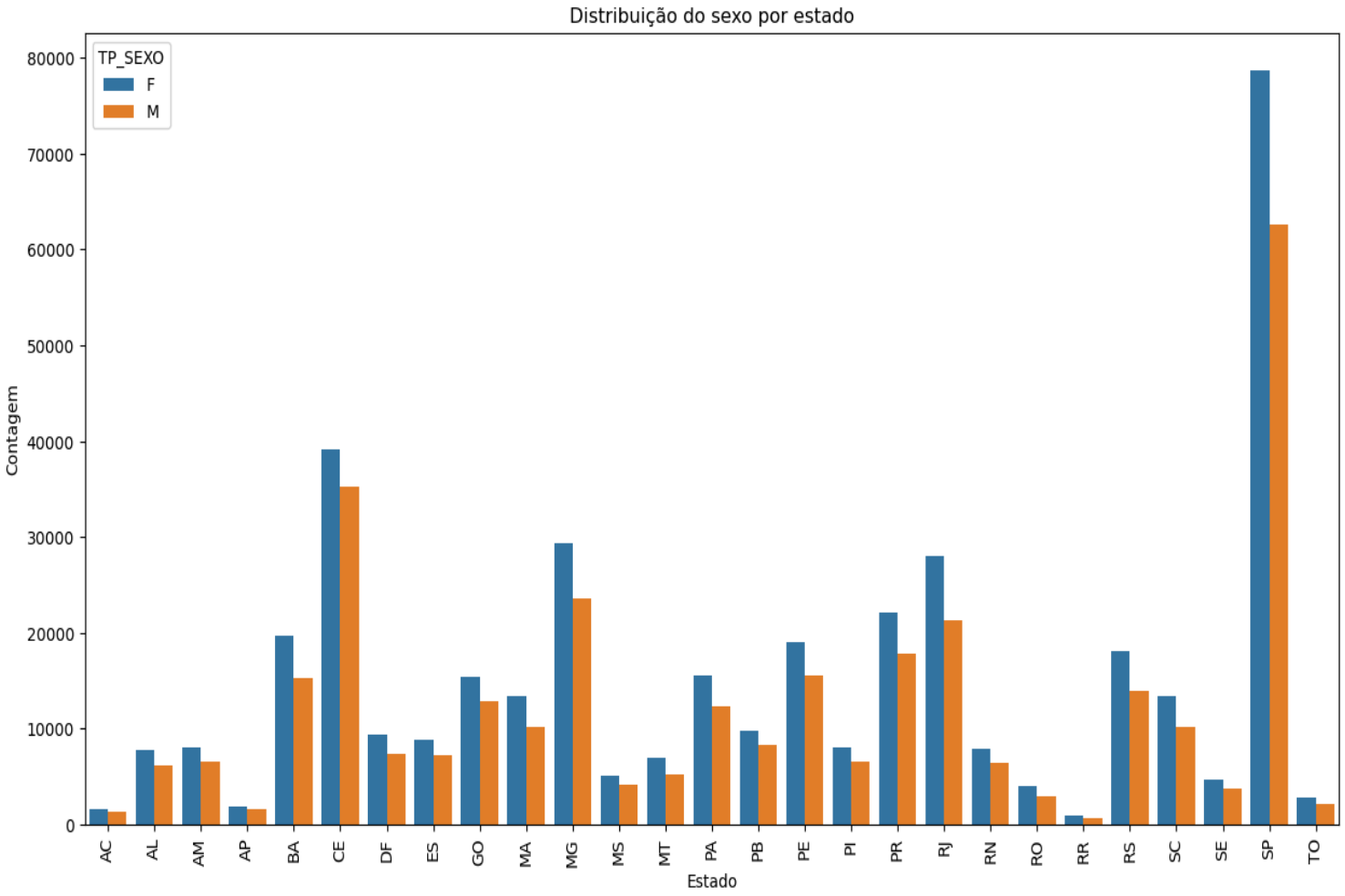
Há uma distribuição desigual entre as diferentes categorias de cor, com uma grande concentração em "Branca" e "Parda". Essa distribuição pode refletir a demografia da população estudada ou poderia ser influenciada por outros fatores como o contexto socioeconômico ou cultural.



**Quais dos Sexos estão em maior quantidade?**

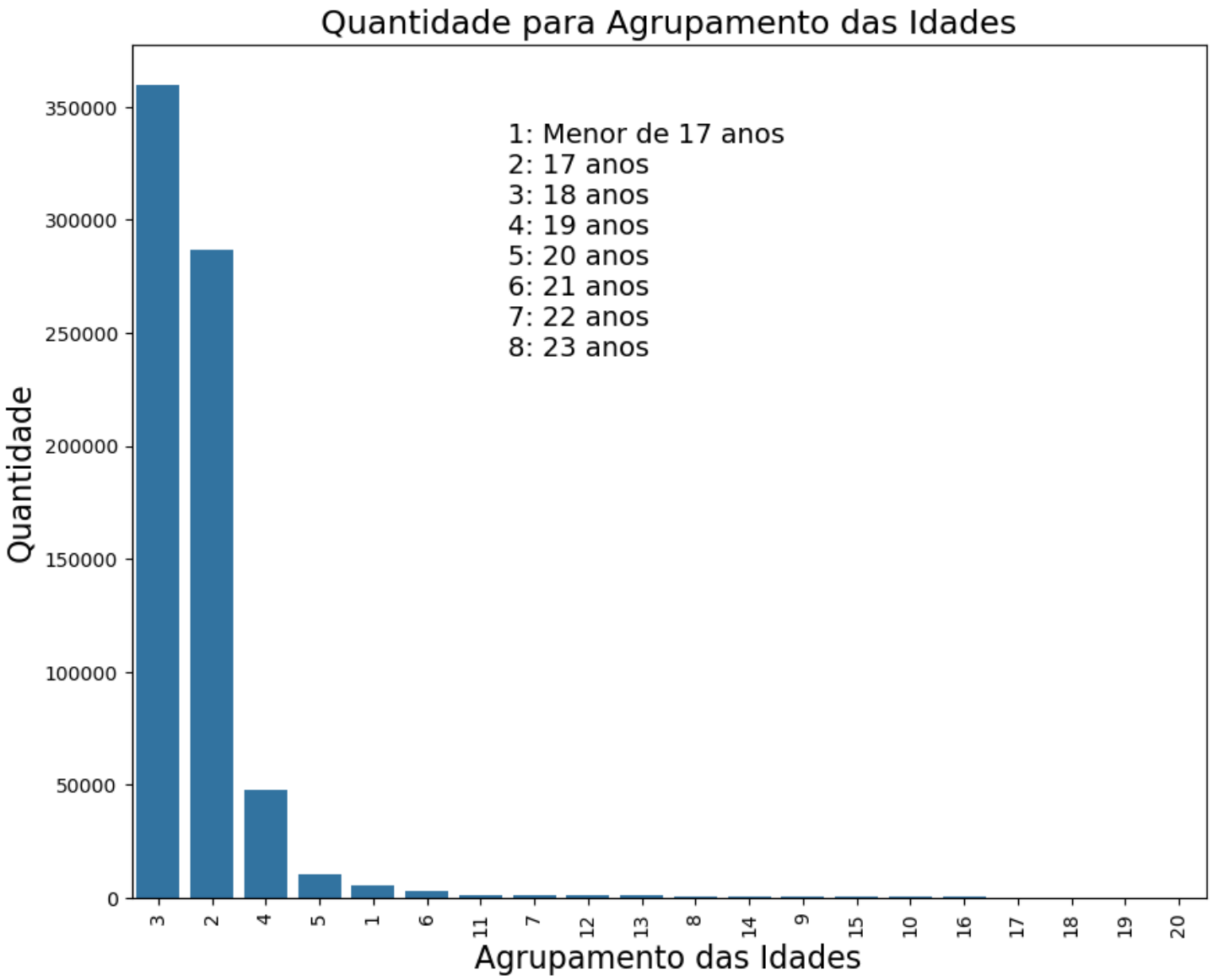
Apesar da quantidade estar próxima, o sexo feminino se destaca apresentando maior número de alunos para realização da prova. Podemos reforçar que trata-se de uma base de dados, uma amostra que ainda passa por limpeza e possivelmente podem haver divergências das informações, não podemos definir como uma verdade absoluta mas que para estes dados obtivemos como resultado o sexo feminino como maior quantidade.




Distribuição do sexo por estado
TP_SEXO
F
M
Contagem
80000
70000
60000
50000
40000
30000
20000
10000
0
AC
AL
AM
AP
BA
CE
DF
ES
GO
MA
MG
MS
MT
PA
PB
PE
PI
PR
RJ
RN
RO
RR
RS
SC
SE
SP
TO
Estado

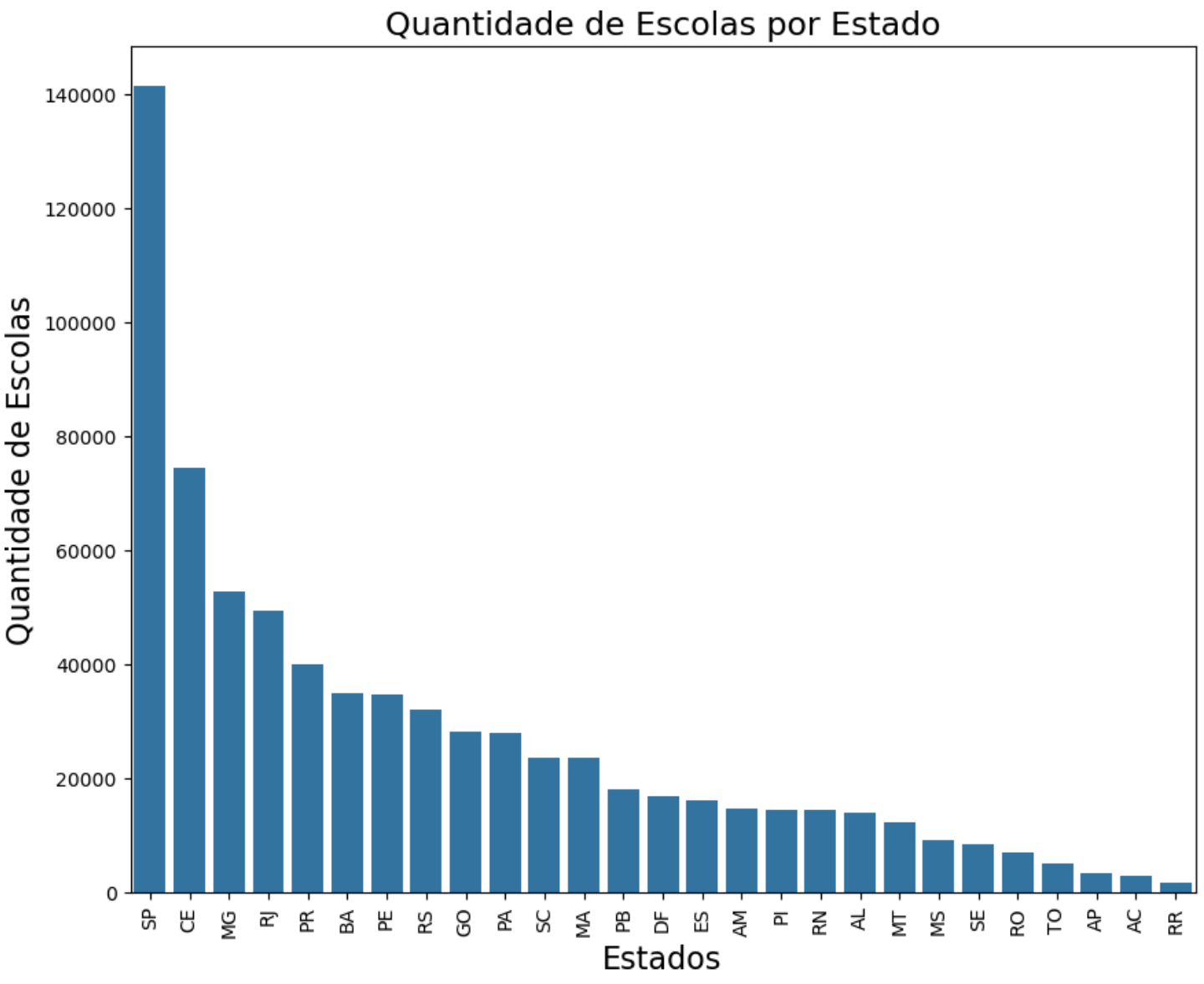
**Quais grupos de idade tem maior quantidade?**

Este gráfico de barras mostra a quantidade de pessoas em diferentes grupos de idade. Há uma concentração clara nas idades de 17 e 18 anos, que são tipicamente as idades dos estudantes que estão concluindo o ensino médio e prestando exames nacionais, como o ENEM no Brasil. As idades acima de 19 anos têm uma representação muito baixa, o que pode indicar que menos indivíduos dessa faixa etária participam do exame ou estão fora da faixa etária típica para tal exame.



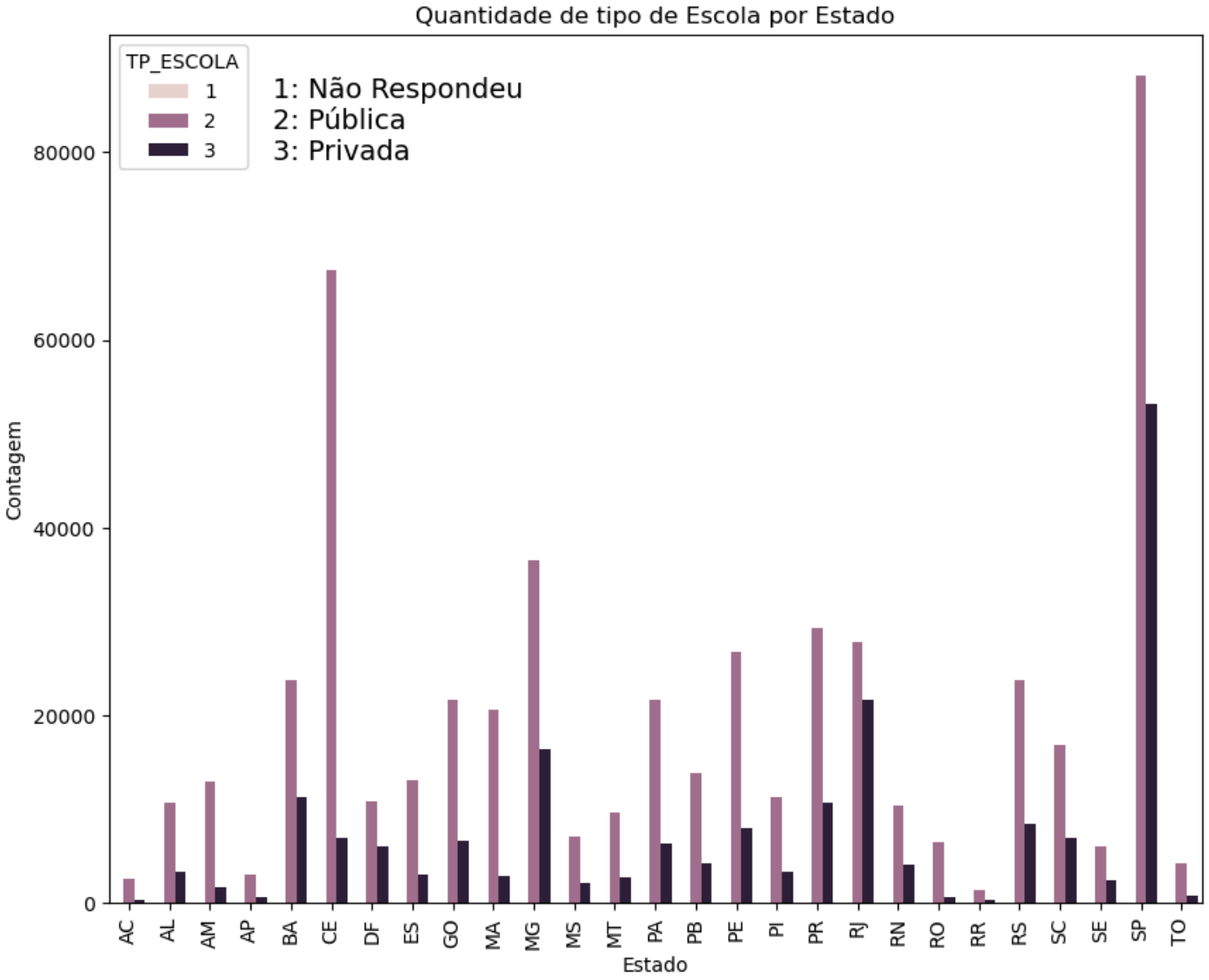
**Qual Estado tem maior número de escolas registradas?**

Pode ser destacado o estado de São Paulo, Ceará e Minas Gerais em questão de maior número de escolas. No estado do Ceará temos em números de habitantes algo próximo dos 8 milhões, já em Minas Gerais valores próximos dos 20 milhões de habitantes. O estado do Ceará se destaca em números de escolas para uma quantidade de habitantes relativamente baixa quando comparado com Minas Gerais com menos escolas para uma população muito mais volumosa. O alto volume de escolas no estado do Ceará permite uma boa qualidade de ensino? Como ficamos quando comparamos Minas Gerais com uma população muito maior e número de escolas menores? Concluiremos esses pontos até o final das análises.



**Como é a distribuição dos tipos de escola por estado?**

Podemos concluir que em todos os estados, existe um maior número de escolas públicas do que privadas, variando apenas na diferença de volumes entre elas. Podemos destacar o Rio de Janeiro como sendo o estado que possui maior número de escolas privadas em comparação ao número de escolas públicas. O Ceará se destaca em suas diferenças entre escolas públicas e privadas, sendo o número de escolas privadas muito pequeno e escolas públicas em quantidades muito elevadas. Em Minas Gerais o número de escolas privadas fica próximo da metade do número de escolas públicas.



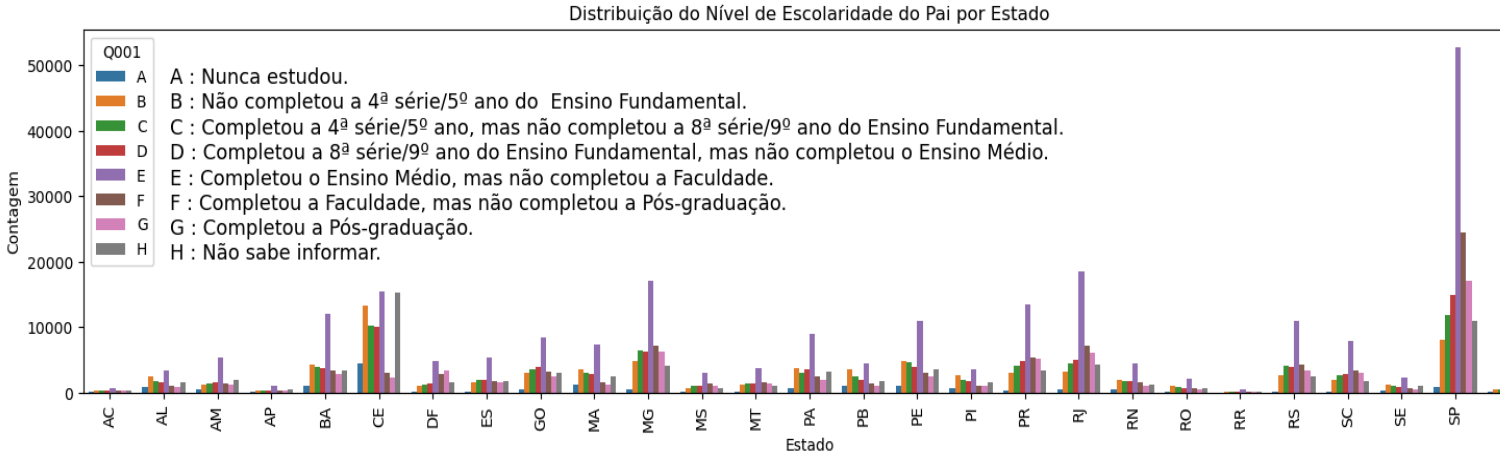

**Como é o nível de escolaridade dos pais ou responsáveis dos participantes da prova?**

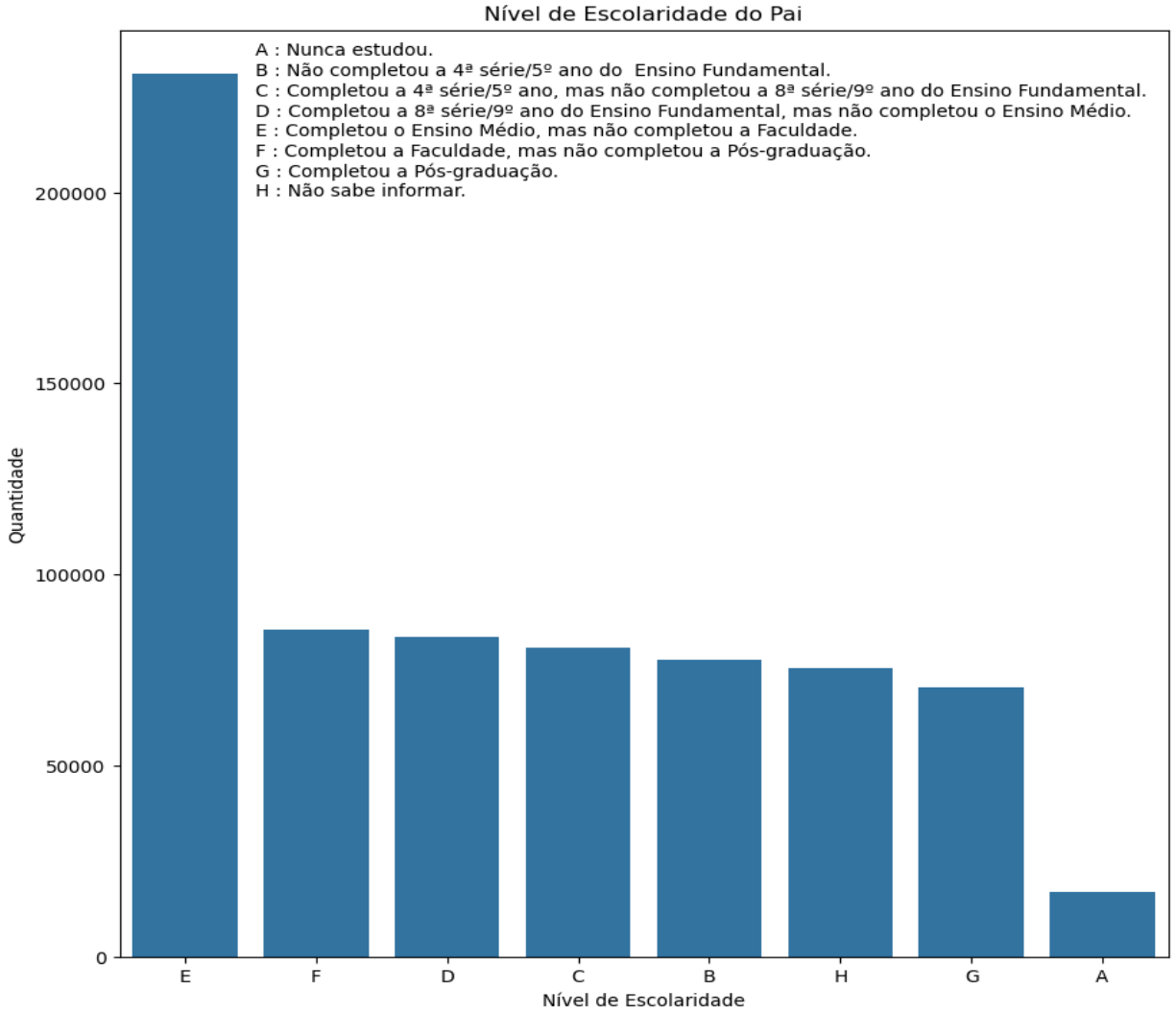
Podemos perceber com os gráficos, que a grande diferença entre Pai e Mãe em relação ao nível de escolaridade se modifica apenas na posição de 'G' é a única que se modifica ficando entre 'H' e 'A' para o Pai e para Mãe ficando entre 'E' e 'F':

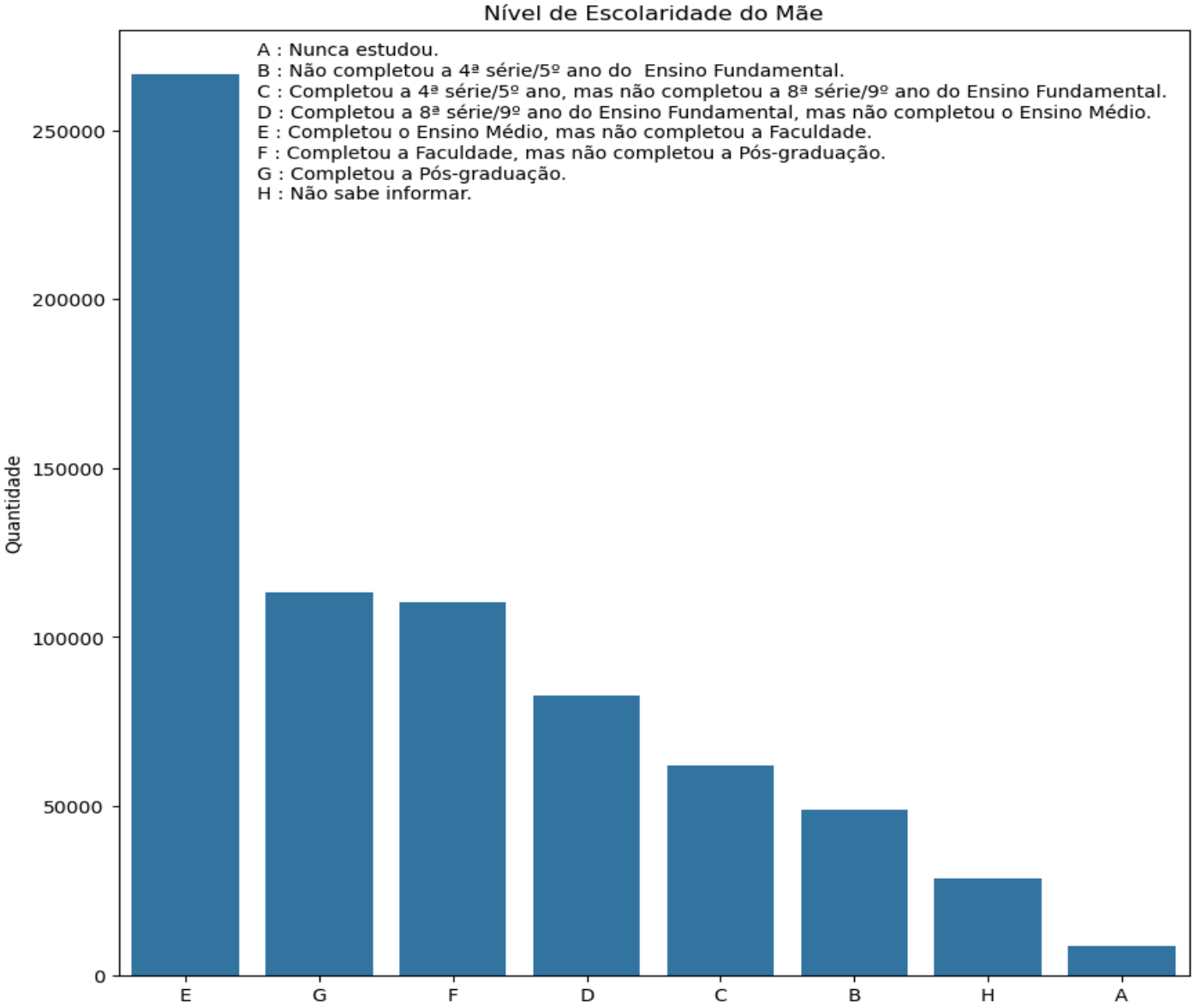
**Pai:**
**E - Completou o Ensino Médio, mas não completou a Faculdade.**
**F - Completou a faculdade, mas não completou a Pós-graduação.**
**D - Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.**
**C - Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.**
**B - Não completou a 4ª série/5º ano do Ensino Fundamental.**
**H - Não sabe informar.**
**G - Completou a Pós-graduação.'**
**A - Nunca estudou.'**

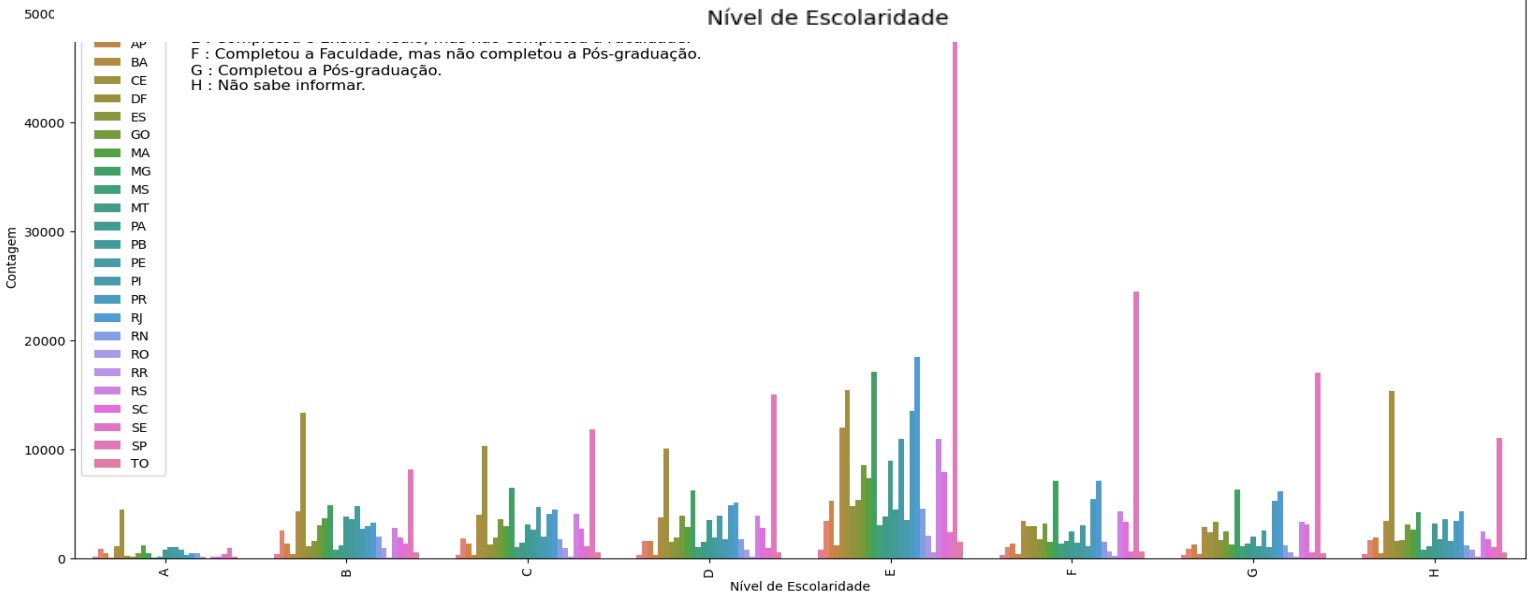
**Mãe:**
**E - Completou o Ensino Médio, mas não completou a Faculdade.**
**G - Completou a Pós-graduação.'**
**F - Completou a faculdade, mas não completou a Pós-graduação.**
**D - Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.**
**C - Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.**
**B - Não completou a 4ª série/5º ano do Ensino Fundamental.**
**H - Não sabe informar.**
**A - Nunca estudou.'**

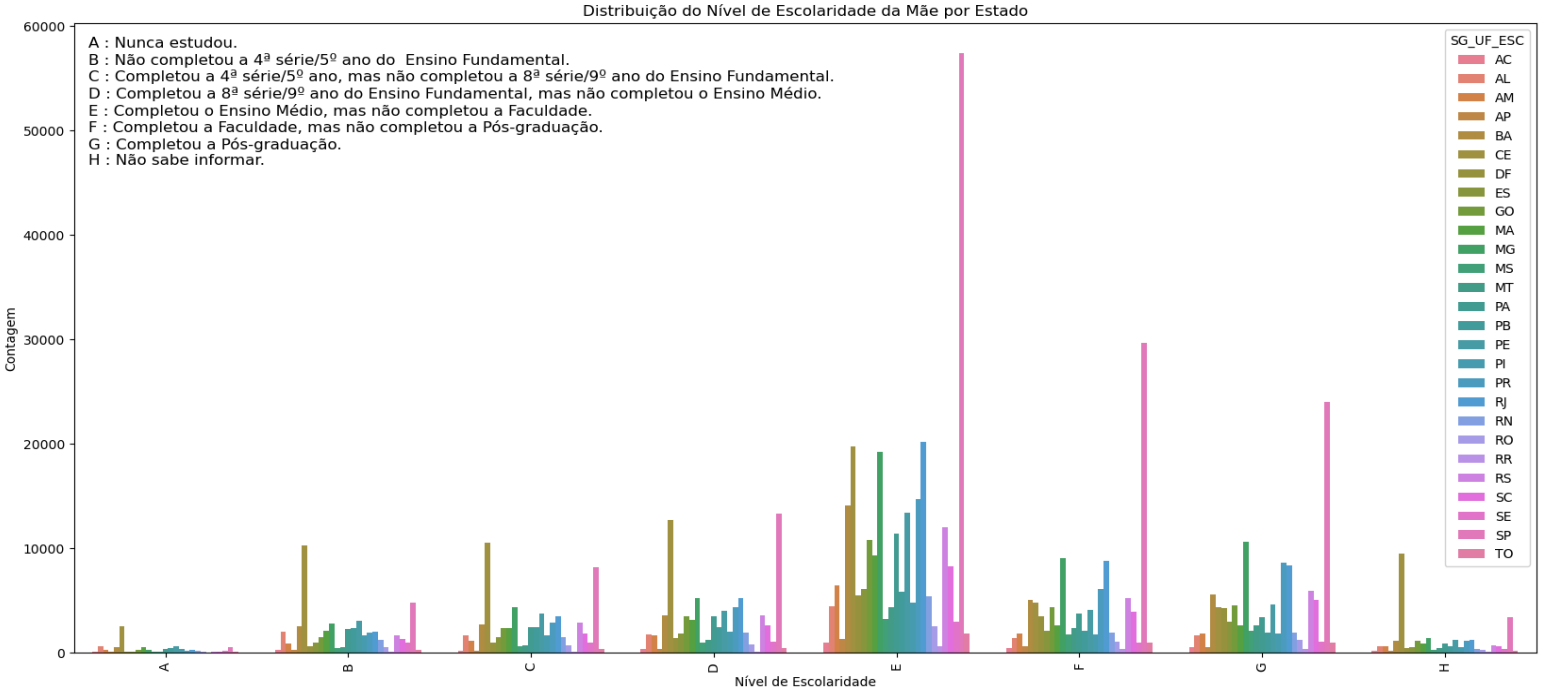
Podemos destacar de forma geral que a definição “E - Completou o Ensino Médio, mas não completou a Faculdade” se destaca dentre as demais definições. De forma preocupante podemos verificar que nas definições que classificam o abandono da escola ainda antes de completar o ensino fundamental é muito elevado principalmente para o sexo masculino no Ceará, Alagoas e na Bahia.

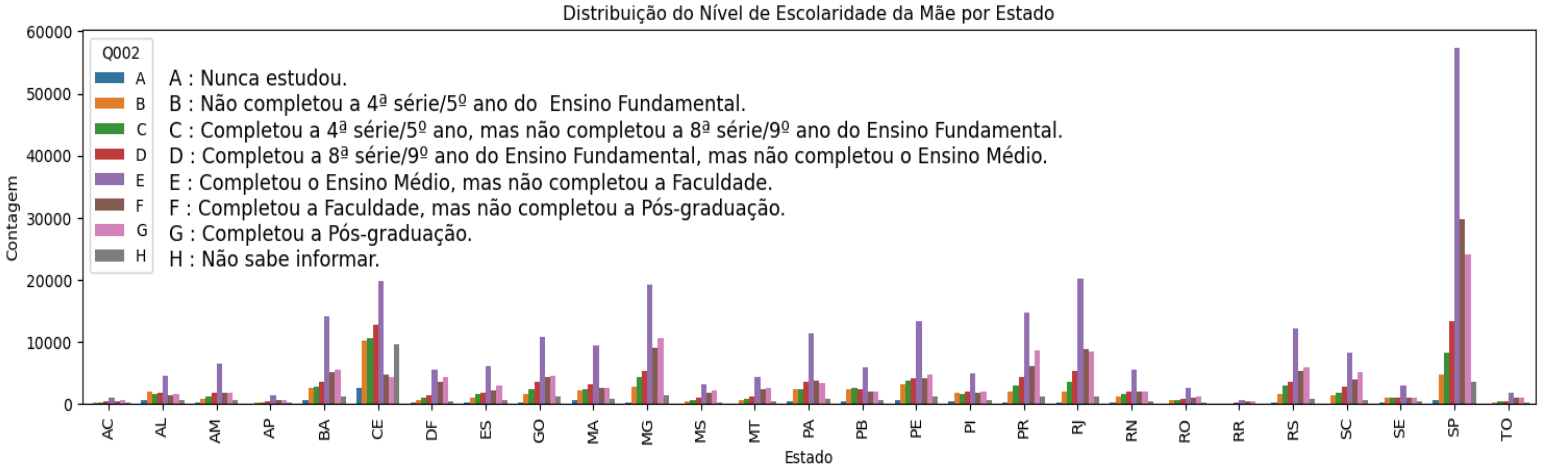



Nível de Escolaridade do Pai
A : Nunca estudou.
B : Não completou a 4ª série/5º ano do  Ensino Fundamental.
C : Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
D : Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
E : Completou o Ensino Médio, mas não completou a Faculdade.
F : Completou a Faculdade, mas não completou a Pós-graduação.
G : Completou a Pós-graduação.
H : Não sabe informar.
Quantidade
0
50000
100000
150000
200000
E
F
D
C
B
H
G
A
Nível de Escolaridade


Nível de Escolaridade do Mãe
A : Nunca estudou.
B : Não completou a 4ª série/5º ano do  Ensino Fundamental.
C : Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
D : Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
E : Completou o Ensino Médio, mas não completou a Faculdade.
F : Completou a Faculdade, mas não completou a Pós-graduação.
G : Completou a Pós-graduação.
H : Não sabe informar.
Quantidade
250000
200000
150000
100000
50000
0
E
G
F
D
C
B
H
A
Nível de Escolaridade

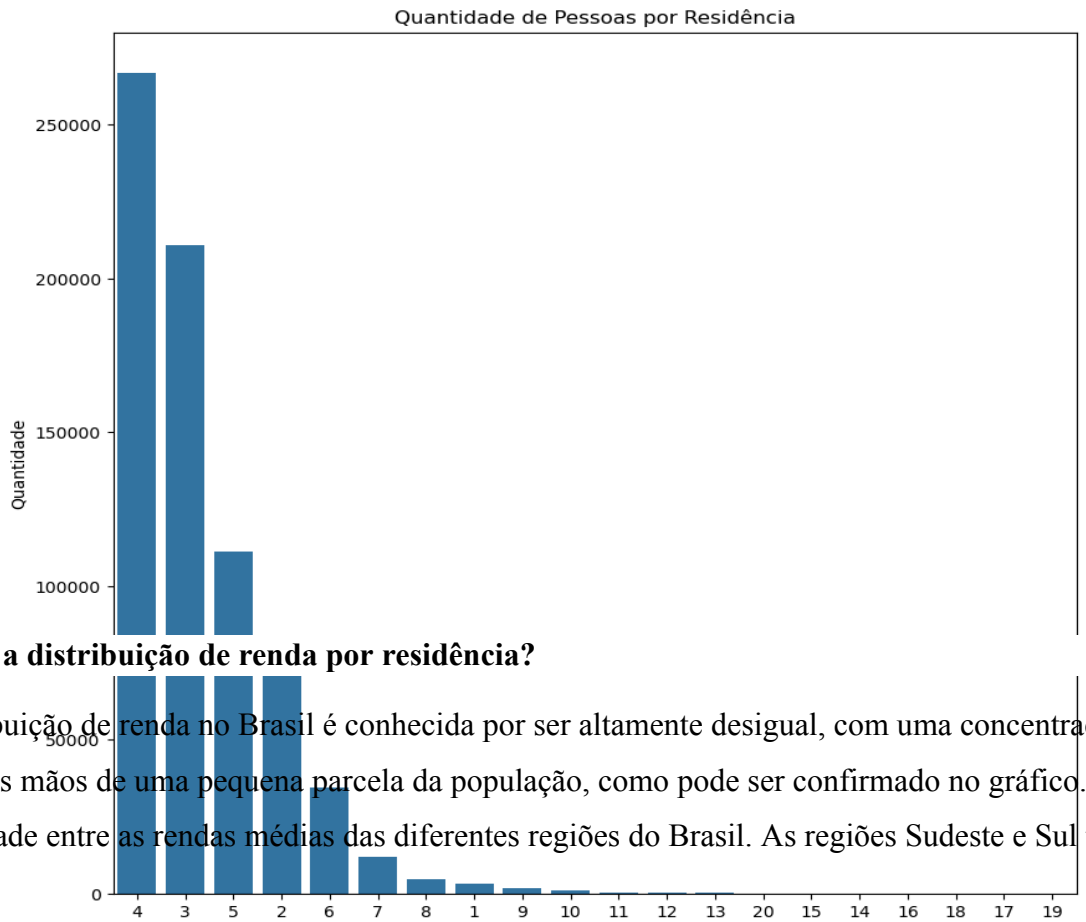

F : Completou a Faculdade, mas não completou a Pós-graduação.
G : Completou a Pós-graduação.
H : Não sabe informar.
AP
BA
CE
DF
ES
GO
MA
MG
MS
MT
PA
PB
PE
PI
PR
RJ
RN
RO
RR
RS
SC
SE
SP
TO
Contagem
40000
30000
20000
10000
0
A
B
C
D
E
F
G
H
Nível de Escolaridade


Distribuição do Nível de Escolaridade da Mãe por Estado
A : Nunca estudou.
B : Não completou a 4ª série/5º ano do Ensino Fundamental.
C : Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
D : Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
E : Completou o Ensino Médio, mas não completou a Faculdade.
F : Completou a Faculdade, mas não completou a Pós-graduação.
G : Completou a Pós-graduação.
H : Não sabe informar.
SG_UF_ESC
AC AL AM AP BA CE DF ES GO MA MG MS MT PA PB PE PI PR RJ RN RO RR RS SC SE SP TO
Contagem
Nível de Escolaridade


Distribuição do Nível de Escolaridade da Mãe por Estado
Q002
A : Nunca estudou.
B : Não completou a 4ª série/5º ano do Ensino Fundamental.
C : Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
D : Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
E : Completou o Ensino Médio, mas não completou a Faculdade.
F : Completou a Faculdade, mas não completou a Pós-graduação.
G : Completou a Pós-graduação.
H : Não sabe informar.
Contagem
Estado

**Qual é a maior frequência para o número de moradores na mesma residência?**

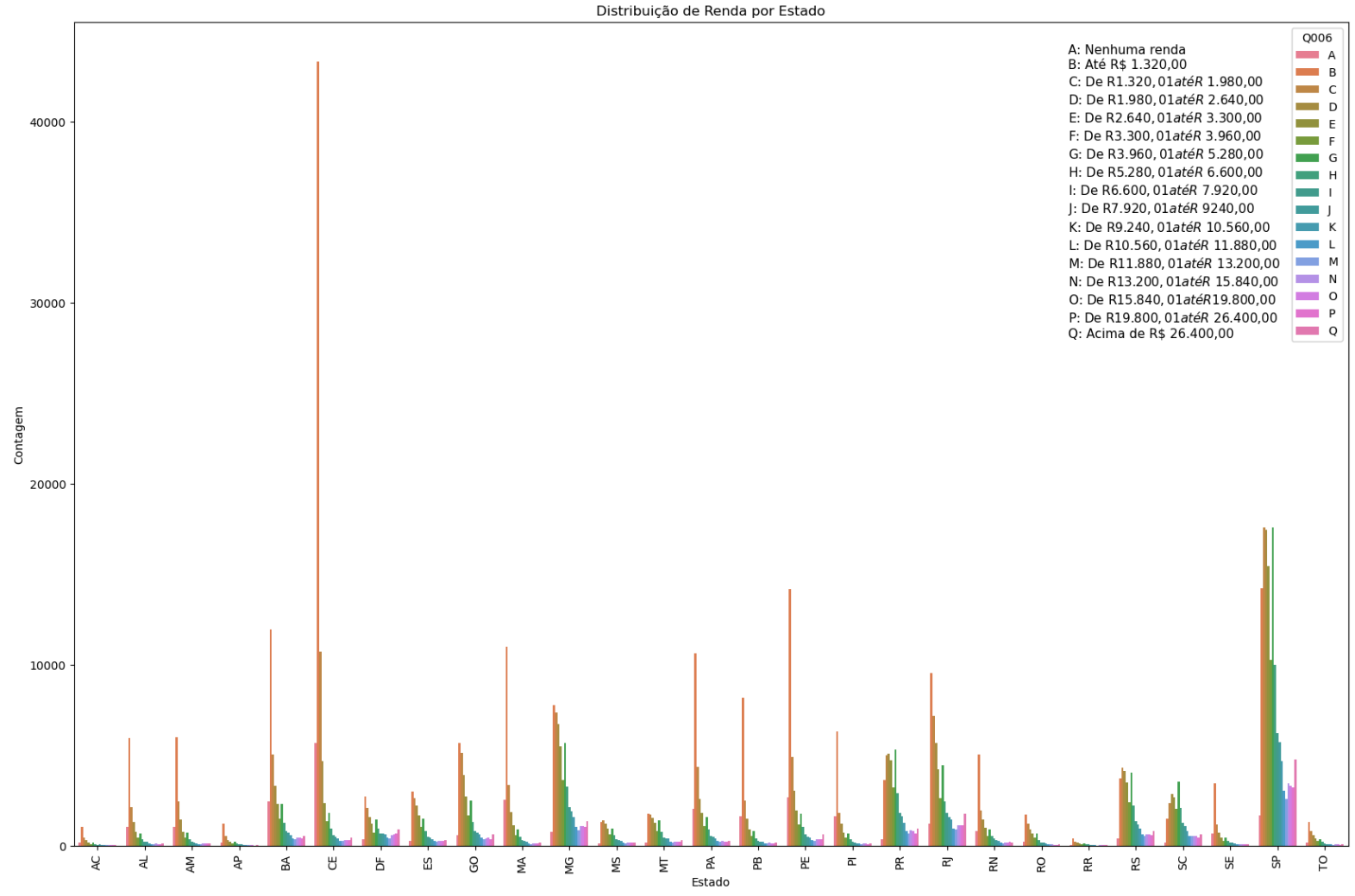
O número de residentes por moradia se destaca com números mais altos para 4 e 3 moradores. Podemos definir como sendo uma casa em que moram os pais e seus dois filhos como a estrutura familiar tipica.

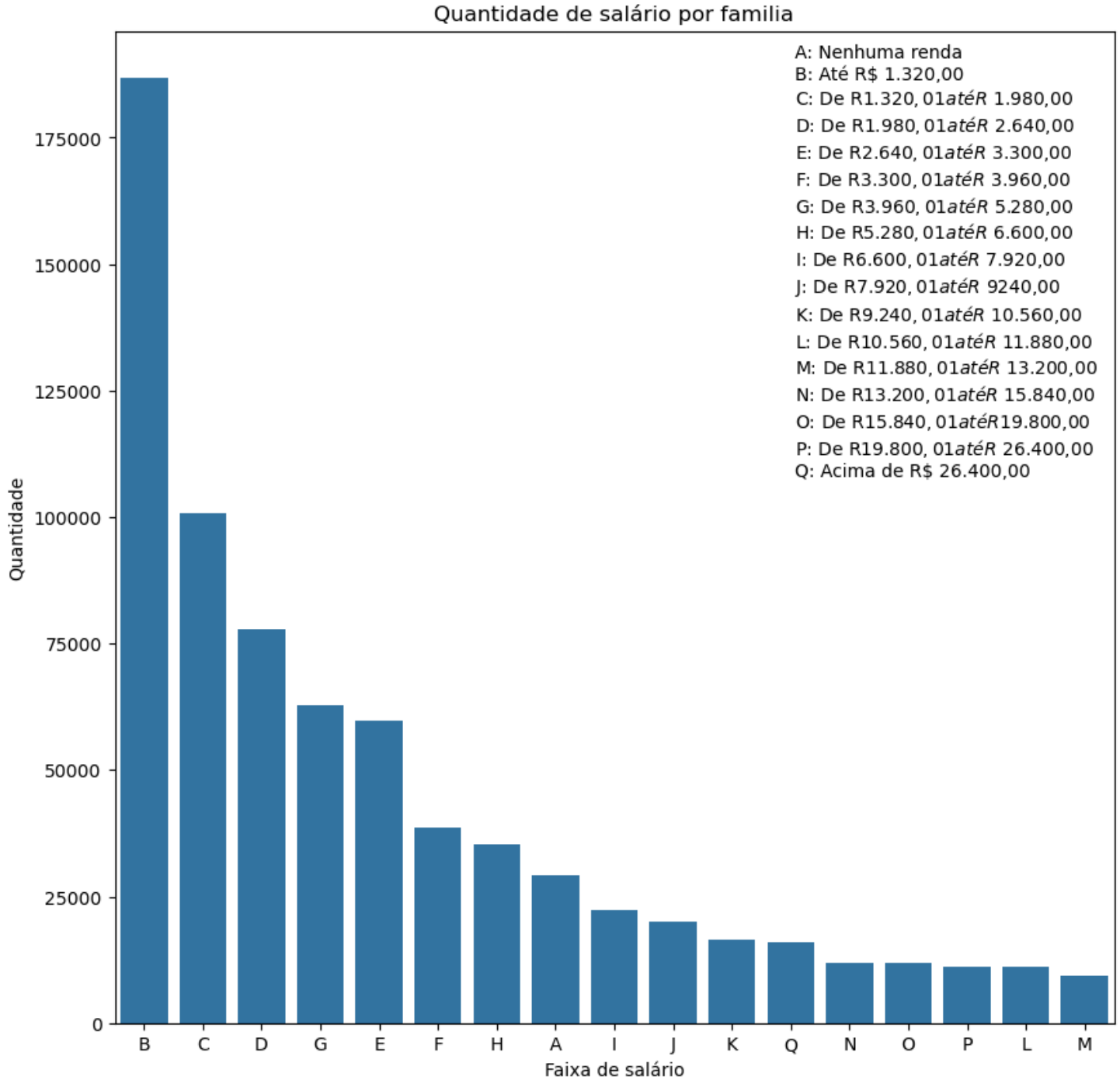



**Como é a distribuição de renda por residência?**

A distribuição de renda no Brasil é conhecida por ser altamente desigual, com uma concentração significativa de renda nas mãos de uma pequena parcela da população, como pode ser confirmado no gráfico. Há uma grande disparidade entre as rendas médias das diferentes regiões do Brasil. As regiões Sudeste e Sul tendem a ter rendas

médias mais altas, enquanto as regiões Norte e Nordeste têm rendas médias mais baixas. As diferenças regionais são marcantes. Estados como São Paulo, Rio de Janeiro e o Distrito Federal apresentam rendas mais elevadas, enquanto estados do Nordeste, como Maranhão e Piauí, possuem rendas significativamente mais baixas.



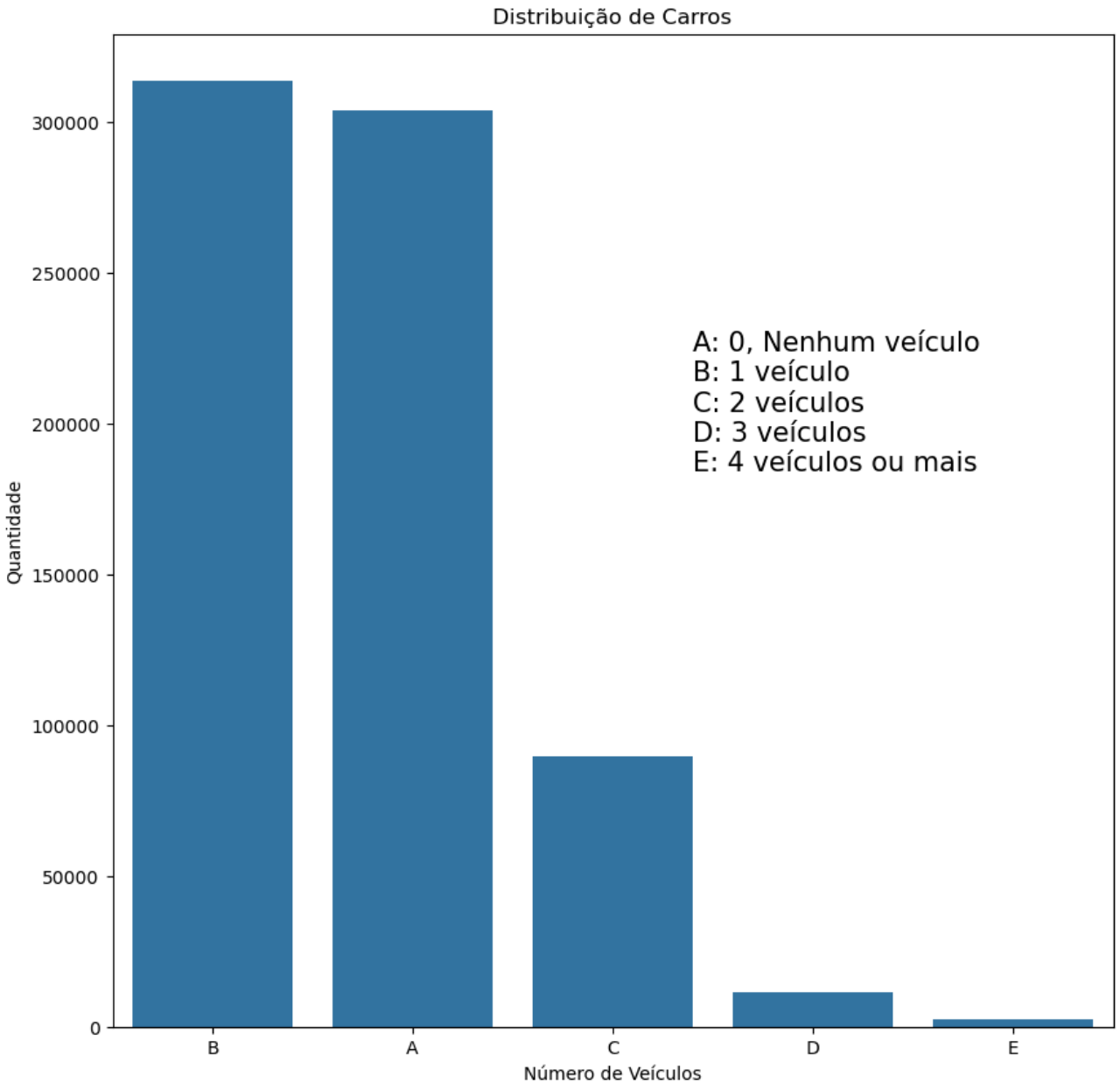

Quantidade de salário por familia
A: Nenhuma renda
B: Até R$ 1.320,00
C: De R1.320,01até R 1.980,00
D: De R1.980,01até R 2.640,00
E: De R2.640,01até R 3.300,00
F: De R3.300,01até R 3.960,00
G: De R3.960,01até R 5.280,00
H: De R5.280,01até R 6.600,00
I: De R6.600,01até R 7.920,00
J: De R7.920,01até R 9240,00
K: De R9.240,01até R 10.560,00
L: De R10.560,01até R 11.880,00
M: De R11.880,01até R 13.200,00
N: De R13.200,01até R 15.840,00
O: De R15.840,01atéR19.800,00
P: De R19.800,01até R 26.400,00
Q: Acima de R$ 26.400,00
Quantidade
0
25000
50000
75000
100000
125000
150000
175000
B
C
D
G
E
F
H
A
I
J
K
Q
N
O
P
L
M
Faixa de salário

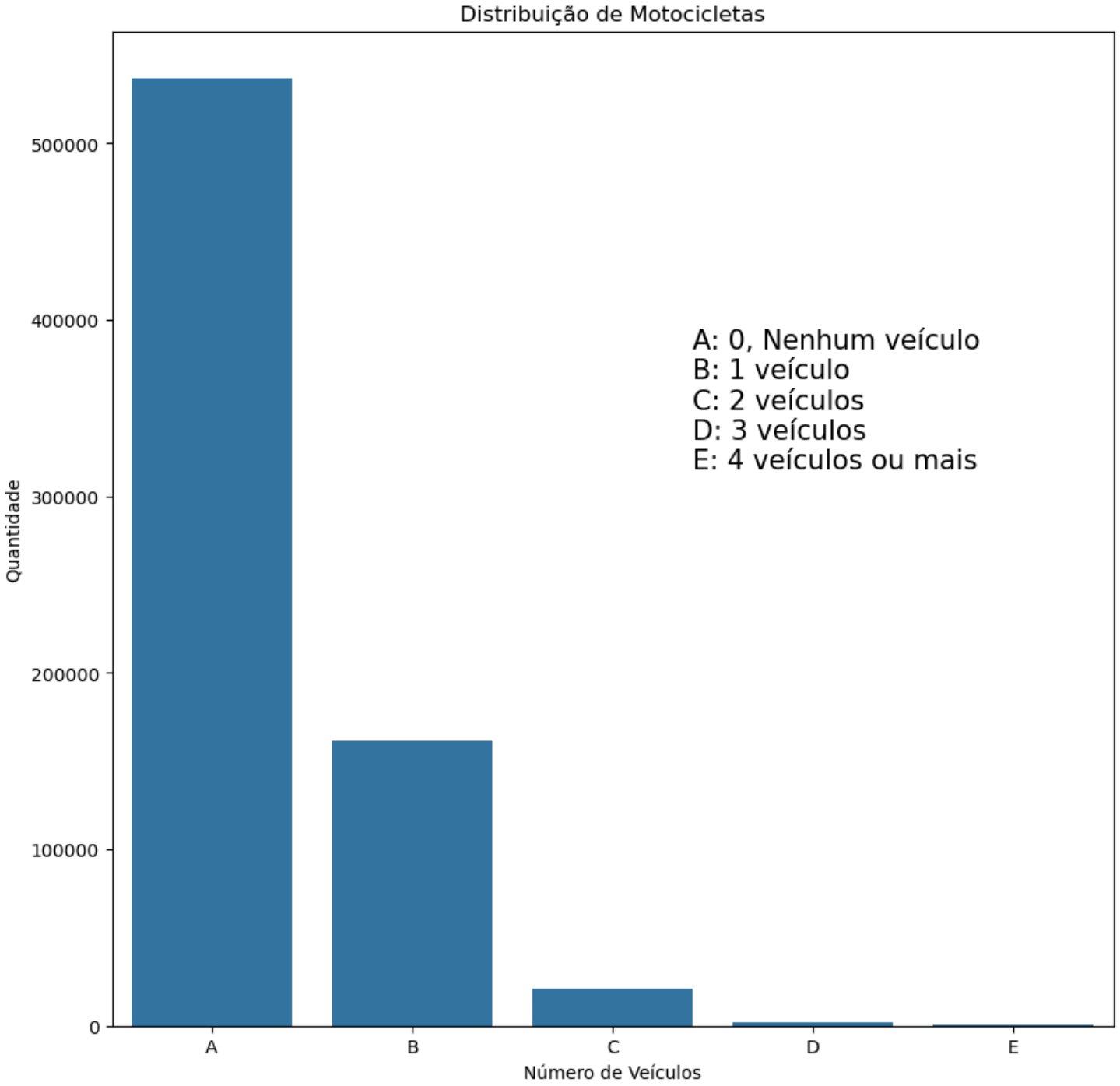
**Como é a distribuição de transporte por residência?**

A desigualdade na posse de veículos entre as famílias no Brasil reflete a ampla disparidade socioeconômica do país.

Uma grande proporção das famílias brasileiras não possui nenhum veículo. Isso é mais comum entre famílias de baixa renda e residentes em áreas urbanas densas, onde o transporte público é mais utilizado, apesar de muitas vezes ser insuficiente ou de baixa qualidade.Muitas famílias possuem apenas um veículo. Esse grupo geralmente inclui famílias de classe média que utilizam o veículo para deslocamentos diários, como trabalho e escola.

Uma parcela menor da população possui dois ou mais veículos. Essas famílias tendem a estar em faixas de renda mais altas e muitas vezes residem em áreas suburbanas ou rurais, onde a dependência de veículos pessoais é maior devido à falta de transporte público adequado.

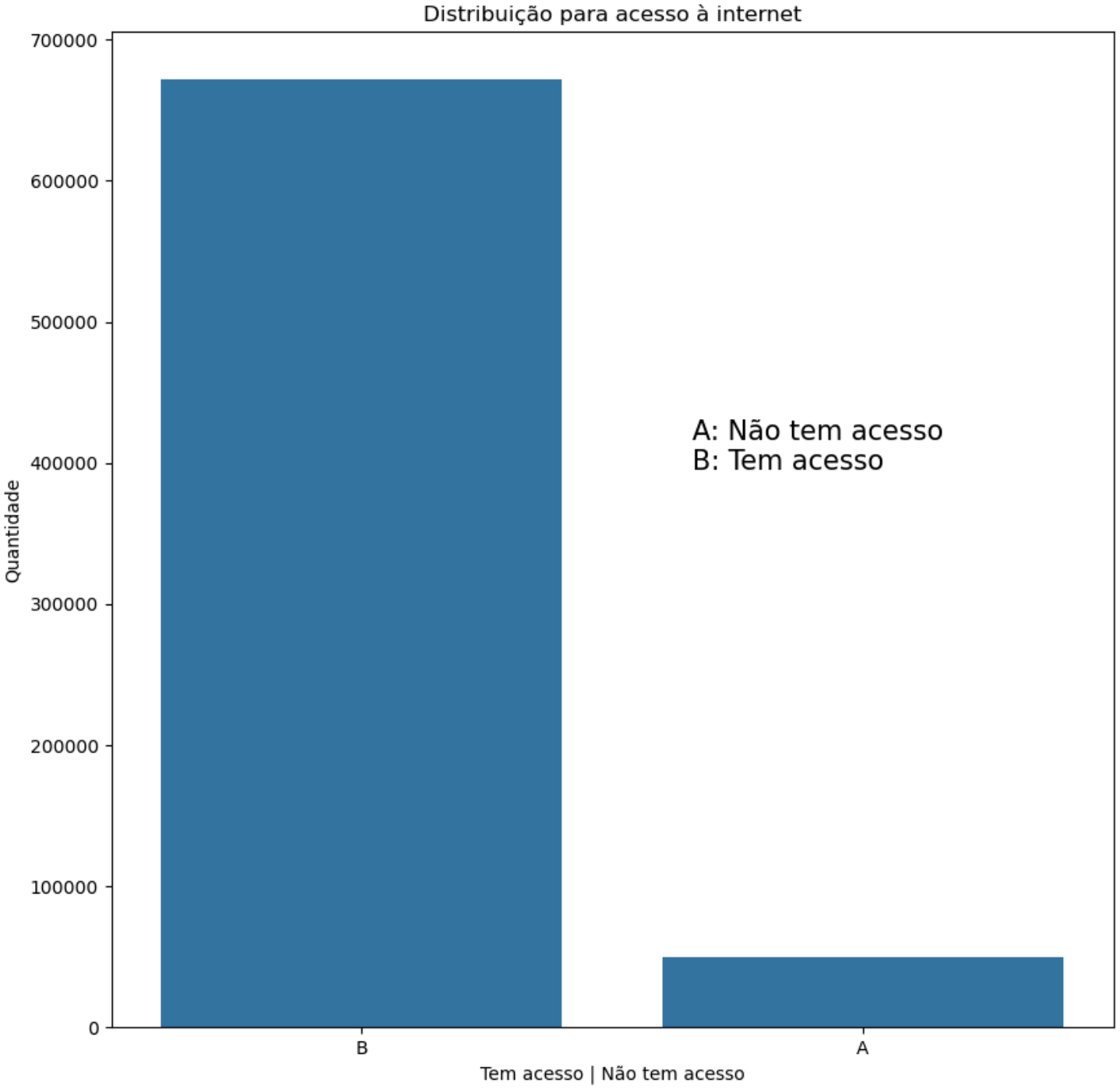



Distribuição de Motocicletas
500000
400000
300000
200000
100000
0
Quantidade
A: 0, Nenhum veículo
B: 1 veículo
C: 2 veículos
D: 3 veículos
E: 4 veículos ou mais
A
B
C
D
E
Número de Veículos

**Como fica o acesso à internet para os alunos?**

O acesso à internet no Brasil tem melhorado significativamente nos últimos anos, mas ainda há disparidades notáveis, especialmente entre diferentes regiões e classes socioeconômicas.

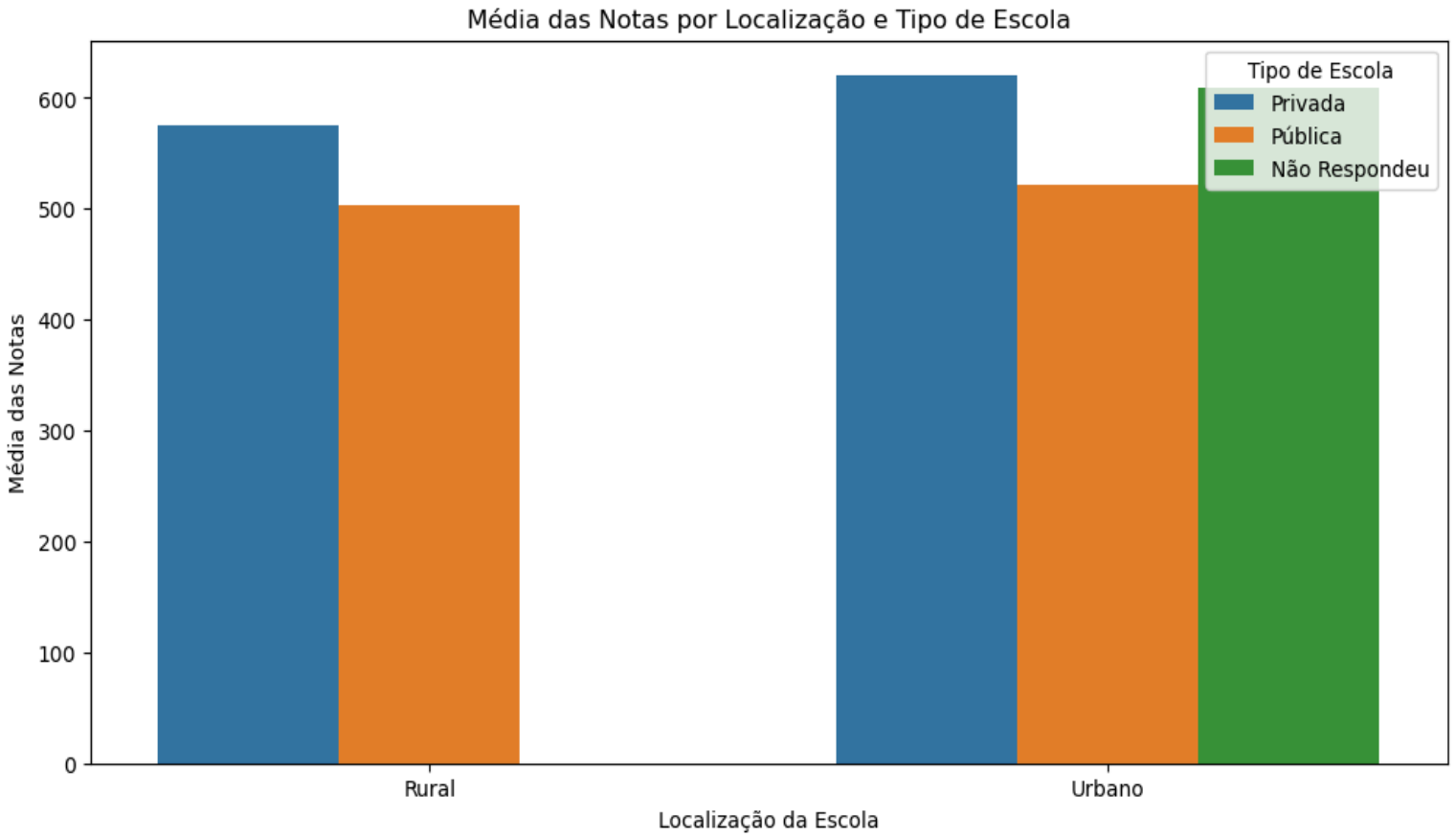
Apesar do progresso significativo, ainda existe uma parcela da população que não tem acesso à internet, principalmente em áreas rurais e entre famílias de baixa renda. Políticas públicas focadas na expansão da infraestrutura de internet, especialmente em áreas remotas e de baixa renda, são essenciais para reduzir essa desigualdade e promover a inclusão digital no Brasil.



### 3.2 Análise de variáveis quantitativas:

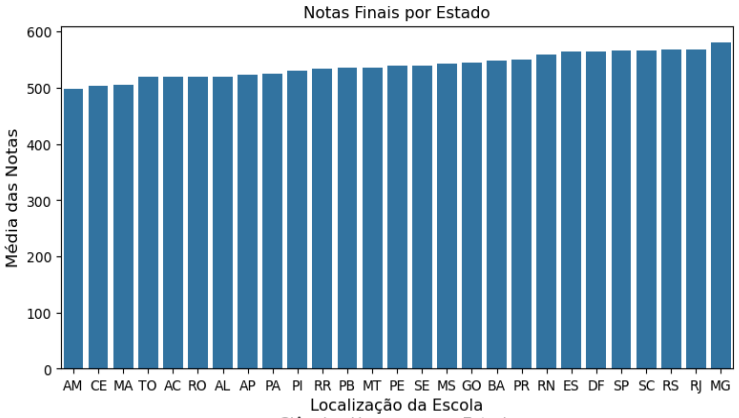
**Como ficam as notas de acordo com a região e o tipo de escola?**

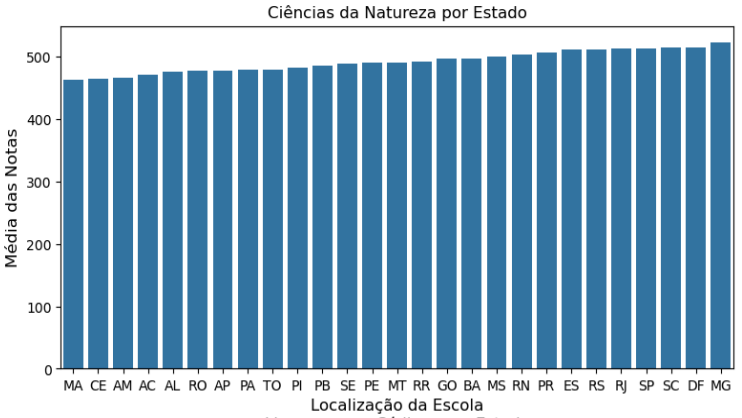
Podemos confirmar que em geral as notas da zona urbana são melhores que em zonas rurais, destacando as escolas privadas como melhores e mais altas notas obtidas.

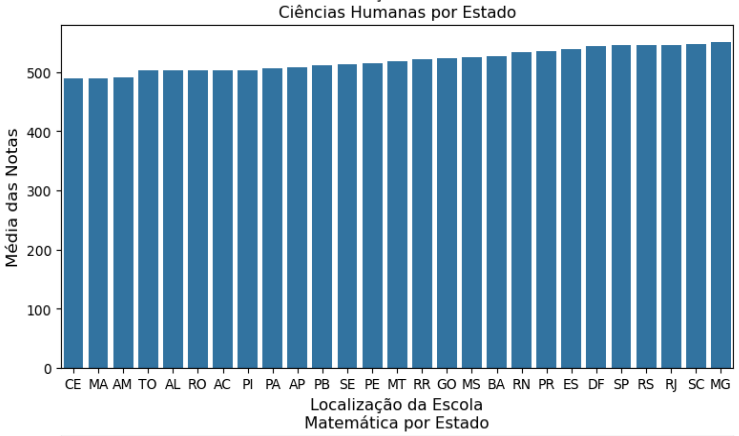


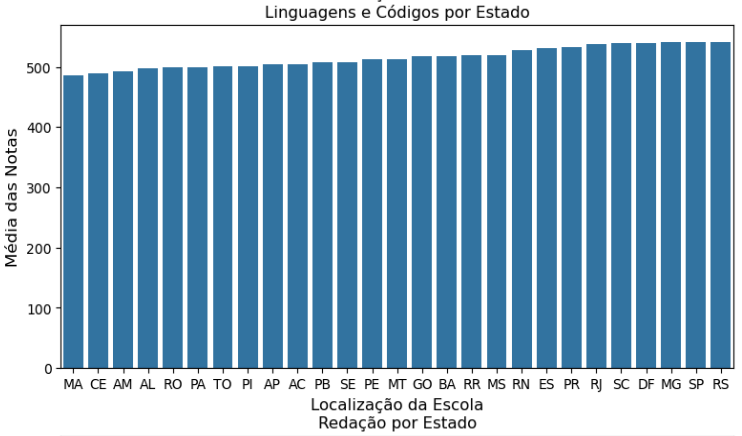
**Vamos definir o valor médio das notas de acordo com cada prova realizada e com a nota por estado:**

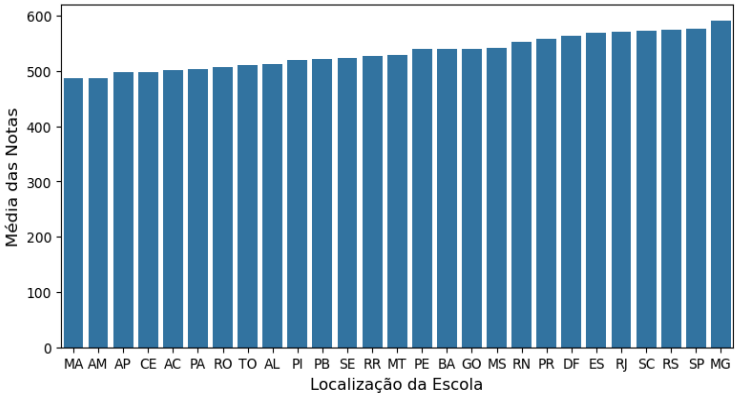
Durante as análises anteriores feitas sobre a quantidade de escolas em relação ao número de população, utilizamos dois exemplos falando sobre o ensino de Minas Gerais e o ensino no Ceará. Aqui podemos concluir no primeiro gráfico “Notas Finais por Estado” que Minas Gerais apesar de mais populosa e com número menor de escolas quando comparado com Ceará, apresenta melhores resultados em quase todas as matérias e apresenta a média de notas finais mais altas. Infelizmente podemos concluir que maior número não significa melhor gestão e qualidade de ensino, pois mesmo que no Ceará exista um grande número de escolas o resultado é um dos piores registrados.

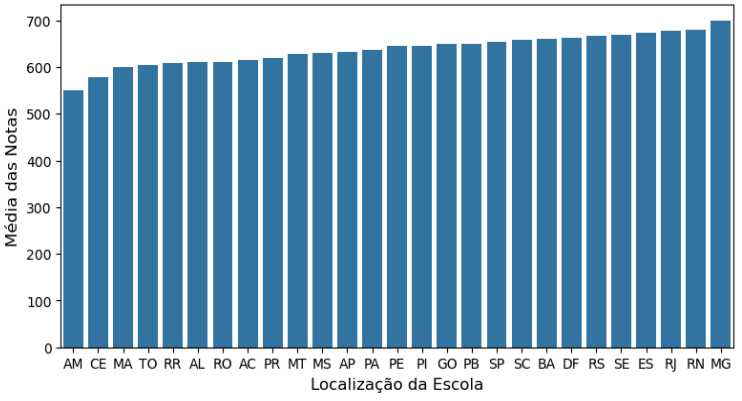


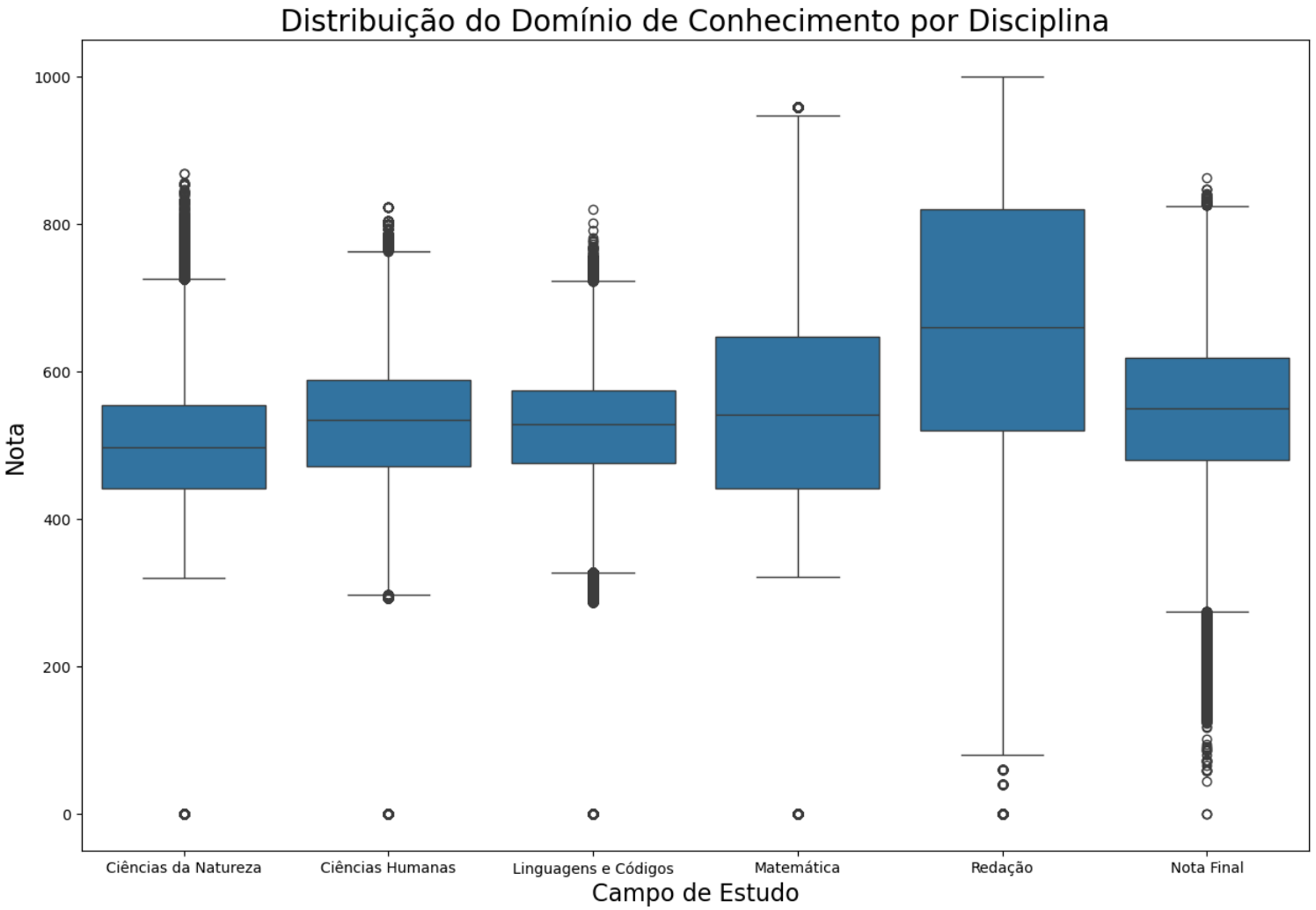
**Como fica a distribuição das notas de acordo com cada disciplina?**

A mediana está muito bem equilibrada entre as disciplinas, indicando uma distribuição simétrica entre os alunos. Podemos verificar que em Redação ocorrem em geral notas maiores e com poucos outliers.

As notas de Ciências da Natureza,Ciências Humanas, Linguagens e Códigos,Nota Final, estão mais concentradas e com menor variabilidade, sugerindo que os alunos têm desempenho similar conforme o formato do "boxplot".

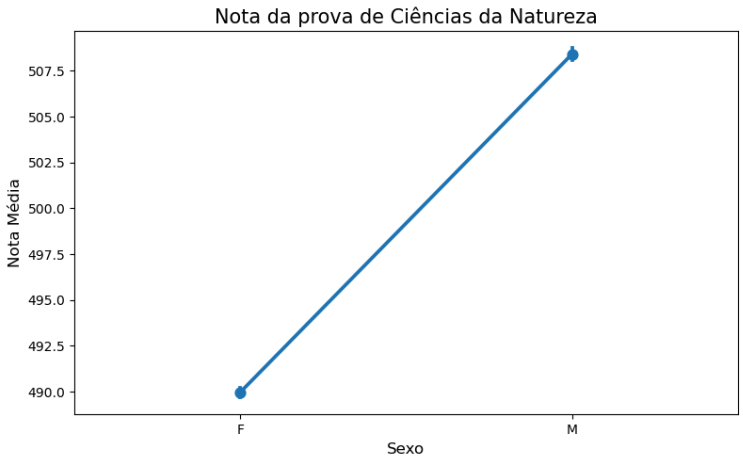
Conseguimos confirmar também que apesar das informações permanecerem concentradas, o número de outliers acaba sendo maior nestes casos.

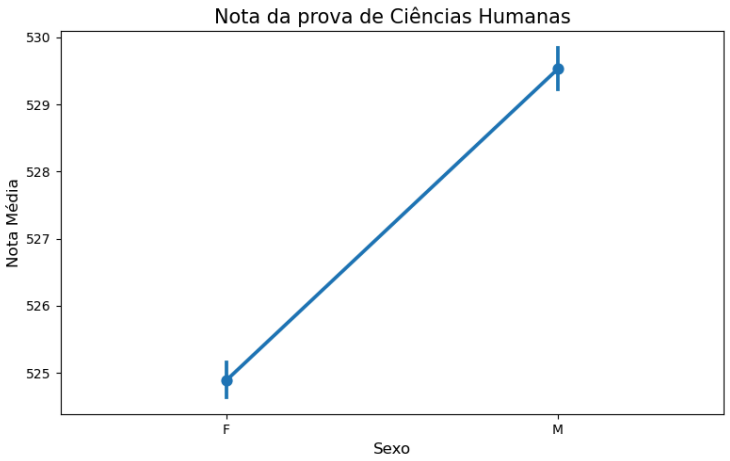
Como oposição à afirmação feita, temos Redação e Matemática, ambas dentre as demais possuem poucos outliers, mas em contrapartida o "boxplot" é maior, o que indica maior variabilidade das notas.

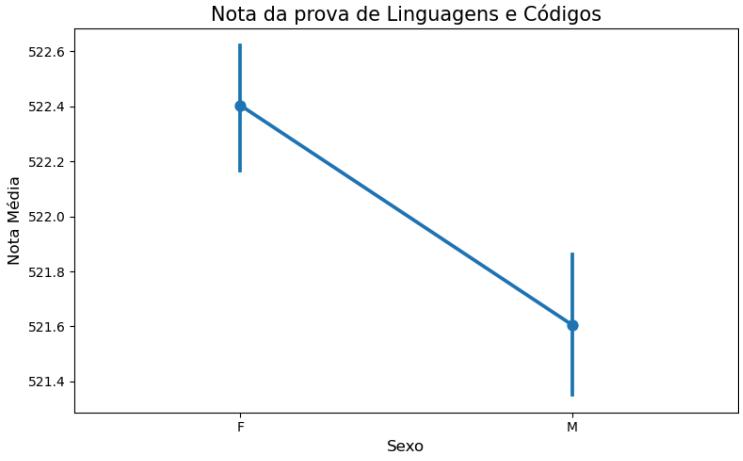


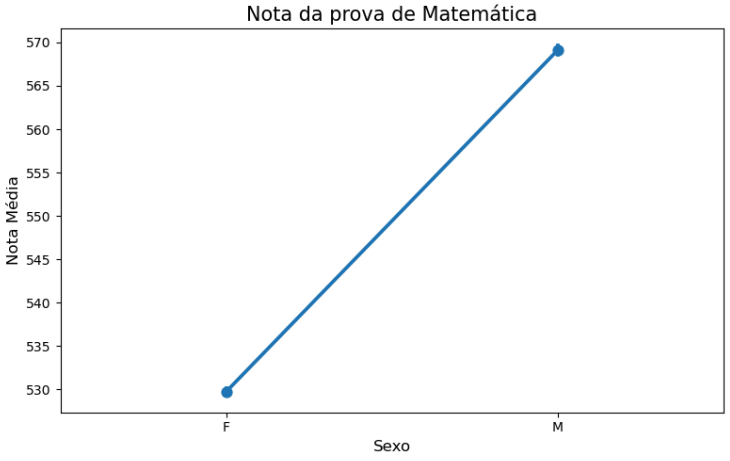
**Como fica a distribuição das notas de acordo com cada tipo de sexo?**

Conforme verificado na imagem, o ponto central em cada grupo no eixo “x” representa a média da variável “y” correspondente. As barras que se formam em cada índice do eixo x no point plot representam os intervalos de confiança para a média dos dados. Eles são úteis para entender a variabilidade e a precisão da estimativa da média, essas barras por padrão medem 95% do intervalo de confiança, o que significa que, se repetíssemos o experimento várias vezes, esperaríamos que a média caísse dentro dessas barras em 95% das vezes.
Podemos confirmar que o sexo masculino tem as melhores notas observando “Nota Final da Prova”.

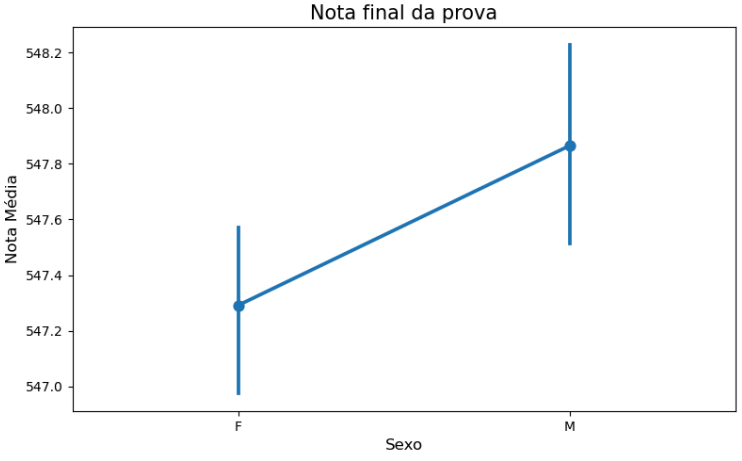


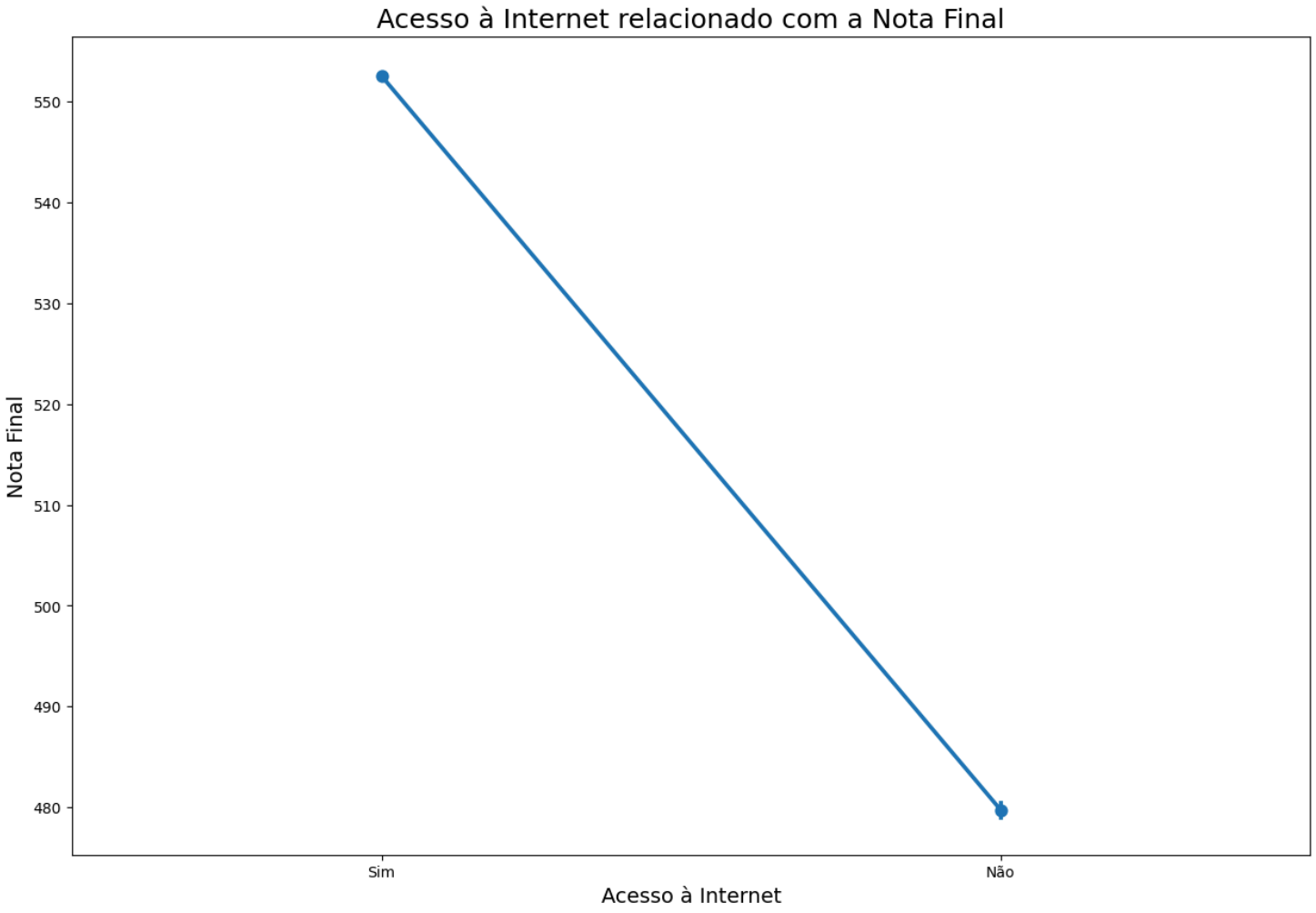




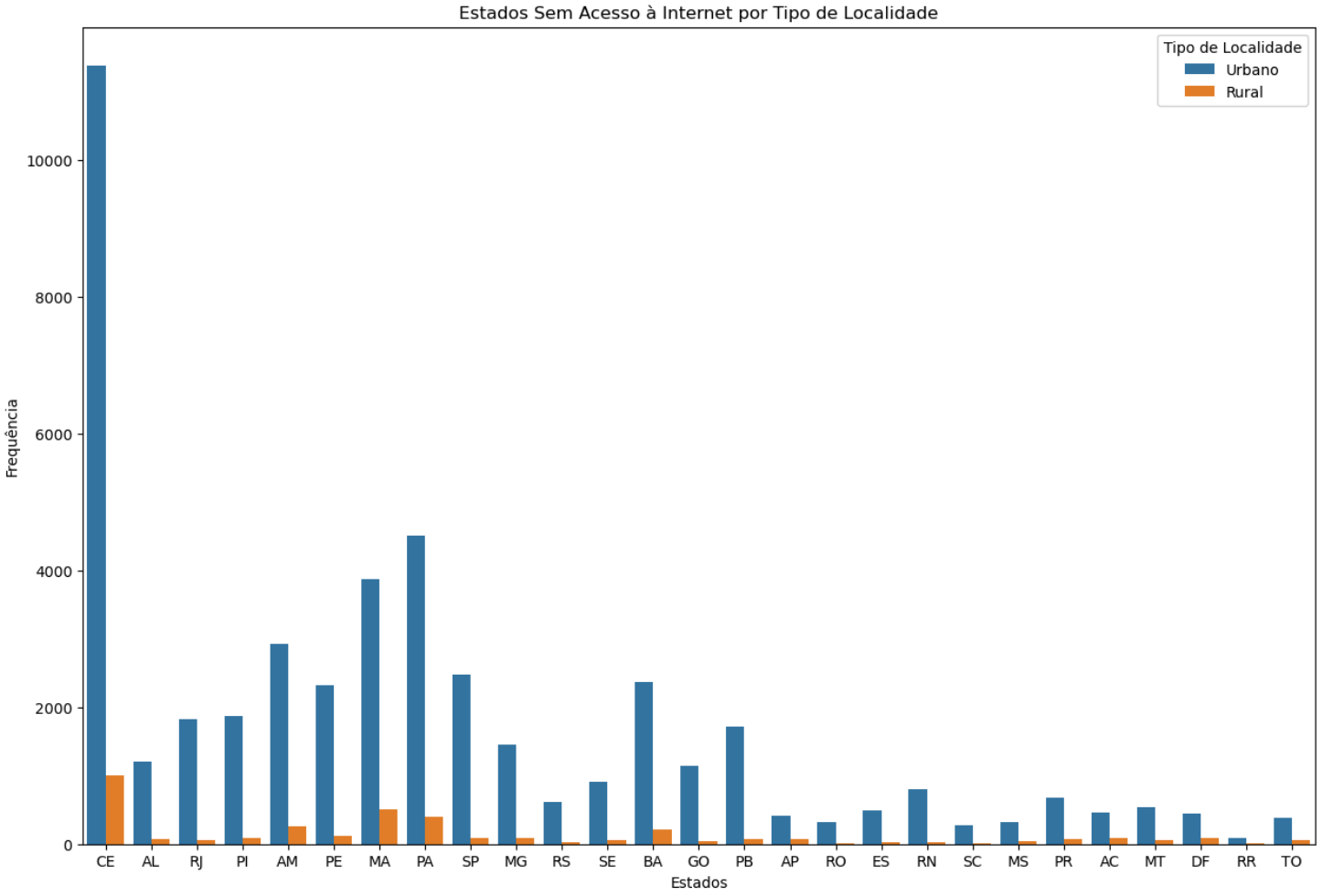
**O acesso à internet causa algum tipo de melhoria nas notas do estudante?**

O acesso à internet tem um impacto significativo no desempenho dos alunos em provas e avaliações educacionais. Vários estudos e dados evidenciam que alunos com acesso à internet tendem a ter melhores notas e desempenho acadêmico comparado àqueles que não possuem esse acesso. Alunos com acesso à internet têm uma vasta gama de recursos educacionais à sua disposição, incluindo vídeos educacionais, tutoriais, livros online, plataformas de aprendizado e ferramentas interativas. Isso facilita o aprendizado autodirigido e o reforço dos conteúdos escolares.

Estudo do IBGE revelou que alunos com acesso regular à internet em casa tiveram um desempenho melhor nas provas do SAEB (Sistema de Avaliação da Educação Básica) comparado àqueles sem acesso. A UNESCO também indicou que o acesso à internet é um fator determinante para a continuidade do aprendizado durante a pandemia, com alunos conectados mostrando menor perda de aprendizado.



Evidentemente, o acesso à internet causa grandes melhorias nas notas dos alunos. Verificando mais detalhes, podemos verificar que a região do Nordeste é uma das maiores regiões afetadas e que causa uma grande redução de oportunidades para os jovens. Necessário ressaltar que o Ceará, que é uma das 27 unidades federativas do Brasil e está situado no norte da Região Nordeste, até o momento possui dados preocupantes para o futuro dos alunos da região.



## 4 MODELAGEM - Machine Learning

Para dar início ao Machine Learning, precisamos ter em mente que todos os dados precisam ser processados e adequados, transformando variáveis categóricas e numéricas.

Podemos realizar assim a visualização das correlações entre as informações:
- Identificação de Relacionamentos Fortes e Fracos: Através da matriz de correlação e do mapa de calor, é possível identificar quais variáveis estão fortemente correlacionadas entre si. Isso ajuda a entender como as variáveis estão relacionadas e quais podem ter impacto significativo em um modelo preditivo.
- Seleção de Variáveis: Nos guiam na seleção das colunas para o modelo, buscando conjuntos eficientes.
- Planejamento: Ao identificar padrões de correlação, podemos planejar análises profundas para investigar relações de causa e efeito entre variáveis específicas.

**Matriz de Correlação:**

A visualização de uma matriz de correlação através de um heatmap pode fornecer insights valiosos sobre a relação entre diferentes variáveis:

**Valor da Correlação:** A correlação entre duas variáveis é medida por um coeficiente que varia de -1 a 1. +1 indica uma correlação positiva perfeita (à medida que uma variável aumenta, a outra também aumenta). -1 indica uma correlação negativa perfeita (à medida que uma variável aumenta, a outra diminui). 0 indica que não há correlação linear entre as variáveis.

**Cores do Gráfico:** O heatmap apresenta diferentes cores que representam diferentes valores de correlação. Tons mais quentes de vermelho indicam uma correlação mais forte, positiva. Tons mais frios indicam uma correlação fraca ou inexistente, negativa.
Interpretações:

**Correlação Positiva Forte:** Se observar valores próximos de 1 entre certas variáveis, significa que essas variáveis têm uma relação linear positiva forte. Por exemplo, se NU_NOTA_MT e NU_NOTA_MÉDIA têm uma correlação alta, pode-se inferir que estudantes com boas notas em Matemática tendem a ter boas notas na média final.

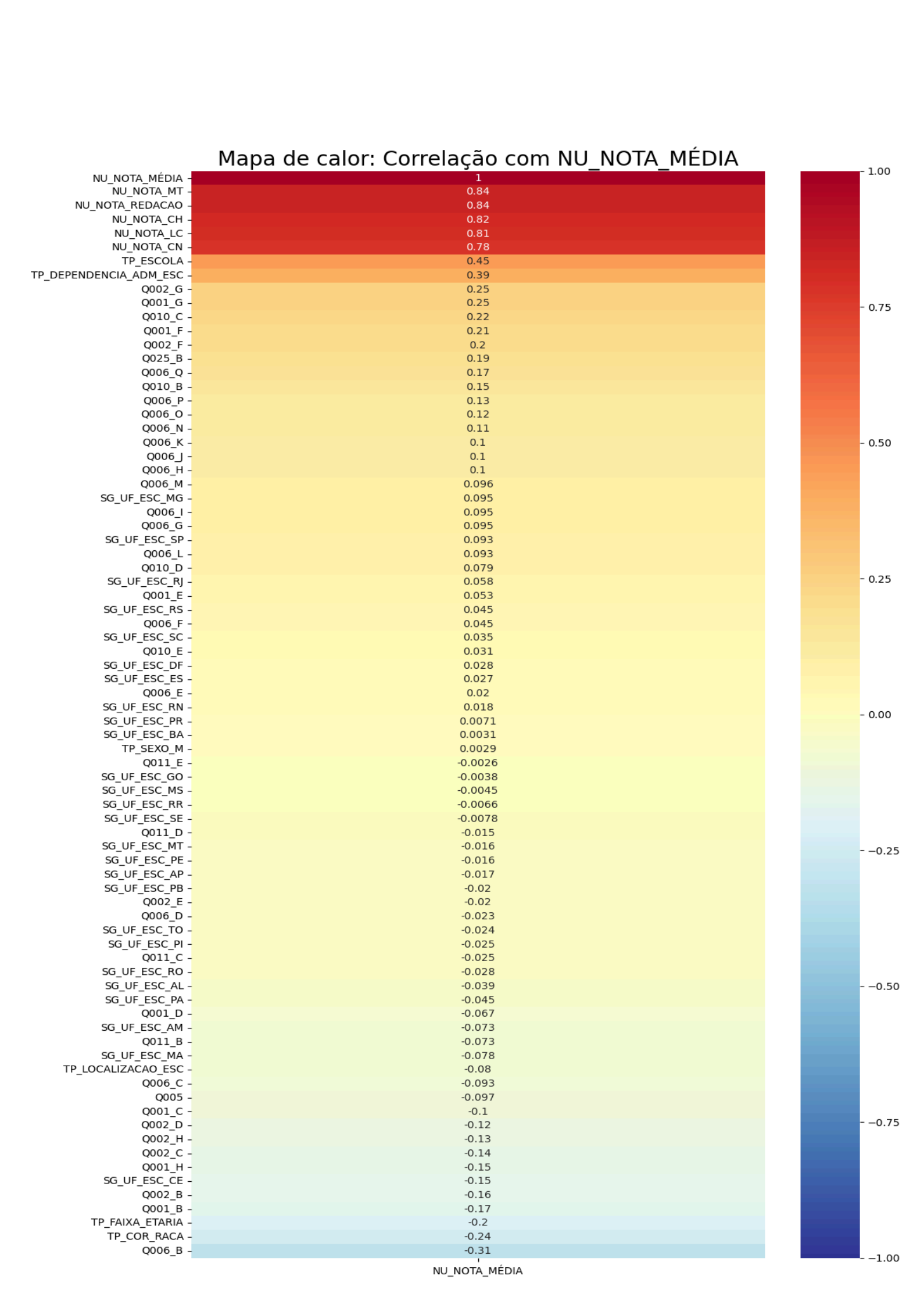
**Correlação Negativa Forte:** Valores próximos de -1 indicam uma forte relação inversa. Por exemplo, se tivesse uma correlação negativa forte entre duas variáveis, isso significaria que quanto maior o valor na variável A, pior seria o valor na variável B.

**Correlação Fraca ou Nula:** Valores próximos de 0 indicam que não há uma relação linear significativa entre as variáveis. Isso sugere que as variáveis não influenciam diretamente umas às outras ou que a relação entre elas não é linear.

Identificar variáveis altamente correlacionadas pode ajudar a escolher quais variáveis incluir em modelos preditivos ou em análises subsequentes. Variáveis altamente correlacionadas podem fornecer informações redundantes.

**Prevenção de Multicolinearidade:** Em modelos de regressão, variáveis altamente correlacionadas podem causar problemas de multicolinearidade, prejudicando a interpretação dos coeficientes. Quando duas ou mais variáveis independentes em um modelo de regressão são altamente correlacionadas, ou seja, elas fornecem informações redundantes sobre a resposta. Isso pode causar problemas na interpretação dos coeficientes do modelo e afetar a precisão das estimativas.

**Neste caso utilizaremos Análise de Componentes Principais (PCA) como solução:** O PCA transforma variáveis correlacionadas em um conjunto de variáveis não correlacionadas chamadas componentes principais. Isso pode ajudar a reduzir a multicolinearidade ao reduzir a dimensionalidade dos dados.


Mapa de calor: Correlação com NU_NOTA_MÉDIA
NU_NOTA_MÉDIA 1
NU_NOTA_MT 0.84
NU_NOTA_REDACAO 0.84
NU_NOTA_CH 0.82
NU_NOTA_LC 0.81
NU_NOTA_CN 0.78
TP_ESCOLA 0.45
TP_DEPENDENCIA_ADM_ESC 0.39
Q002_G 0.25
Q001_G 0.25
Q010_C 0.22
Q001_F 0.21
Q002_F 0.2
Q025_B 0.19
Q006_Q 0.17
Q010_B 0.15
Q006_P 0.13
Q006_O 0.12
Q006_N 0.11
Q006_K 0.1
Q006_J 0.1
Q006_H 0.1
Q006_M 0.096
SG_UF_ESC_MG 0.095
Q006_I 0.095
Q006_G 0.095
SG_UF_ESC_SP 0.093
Q006_L 0.093
Q010_D 0.079
SG_UF_ESC_RJ 0.058
Q001_E 0.053
SG_UF_ESC_RS 0.045
Q006_F 0.045
SG_UF_ESC_SC 0.035
Q010_E 0.031
SG_UF_ESC_DF 0.028
SG_UF_ESC_ES 0.027
Q006_E 0.02
SG_UF_ESC_RN 0.018
SG_UF_ESC_PR 0.0071
SG_UF_ESC_BA 0.0031
TP_SEXO_M 0.0029
Q011_E -0.0026
SG_UF_ESC_GO -0.0038
SG_UF_ESC_MS -0.0045
SG_UF_ESC_RR -0.0066
SG_UF_ESC_SE -0.0078
Q011_D -0.015
SG_UF_ESC_MT -0.016
SG_UF_ESC_PE -0.016
SG_UF_ESC_AP -0.017
SG_UF_ESC_PB -0.02
Q002_E -0.02
Q006_D -0.023
SG_UF_ESC_TO -0.024
SG_UF_ESC_PI -0.025
Q011_C -0.025
SG_UF_ESC_RO -0.028
SG_UF_ESC_AL -0.039
SG_UF_ESC_PA -0.045
Q001_D -0.067
SG_UF_ESC_AM -0.073
Q011_B -0.073
SG_UF_ESC_MA -0.078
TP_LOCALIZACAO_ESC -0.08
Q006_C -0.093
Q005 -0.097
Q001_C -0.1
Q002_D -0.12
Q002_H -0.13
Q002_C -0.14
Q001_H -0.15
SG_UF_ESC_CE -0.15
Q002_B -0.16
Q001_B -0.17
TP_FAIXA_ETARIA -0.2
TP_COR_RACA -0.24
Q006_B -0.31
NU_NOTA_MÉDIA
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

**Processamento dos dados aplicados:**

O processamento de dados é essencial para desenvolvimento do projeto, evitando os dados brutos que frequentemente contêm erros e podem comprometer a qualidade da informação.
Garantir Qualidade: Limpar e corrigir erros, valores ausentes e inconsistências.

Transformar e Normalizar: Diferentes conjuntos de dados podem ter formatos ou unidades diferentes. O processamento de dados permite transformar e normalizar os dados para que possam ser comparados ou combinados de forma eficaz.

Redução de Dimensionalidade: Muitos conjuntos de dados contém mais informações do que o necessário. Técnicas de processamento de dados, como seleção de características ou redução de dimensionalidade, ajudam a simplificar os dados, mantendo apenas as informações relevantes.

Melhoria da Performance de Modelos: Modelos de aprendizado de máquina e algoritmos de análise de dados dependem de dados de alta qualidade. Dados bem processados levam a modelos mais precisos e eficientes.

OneHotEncoder:

- O OneHotEncoder é uma técnica utilizada para converter variáveis categóricas em uma forma que pode ser fornecida a algoritmos de aprendizado de máquina. Ele transforma cada categoria em uma nova coluna binária (0 ou 1).

StandardScaler:

- O StandardScaler padroniza características para que tenham média zero e variância unitária, o que é benéfico para muitos algoritmos de aprendizado de máquina. Ele funciona calculando a média e o desvio padrão de cada característica e transformando os dados de acordo. É uma etapa essencial no pré-processamento de dados para garantir que todas as características contribuem igualmente para o modelo.

ColumnTransformer:

- O ColumnTransformer permite aplicar diferentes transformações a diferentes colunas de um conjunto de dados. É útil quando se trabalha com colunas de diferentes tipos (numéricas e categóricas) e se deseja aplicar transformações específicas a cada tipo.

PCA - Principal Component Analysis | Análise de Componentes Principais:

- Método estatístico que usa uma transformação ortogonal para converter observações de variáveis possivelmente correlacionadas em um conjunto de valores de variáveis linearmente não correlacionadas chamadas componentes principais.
- Reduz a quantidade de variáveis (colunas) dos dados, tornando-os mais simples e manejáveis. Remove variáveis redundantes, mantendo apenas aquelas que contribuem de forma significativa para a

variabilidade dos dados.

- Reduzindo a dimensionalidade, pode-se diminuir a complexidade do modelo, ajudando a prevenir o overfitting (ajuste excessivo) em modelos de aprendizado de máquina.

**Modelo de Machine Learning:**

MSE - Mean Squared Error:

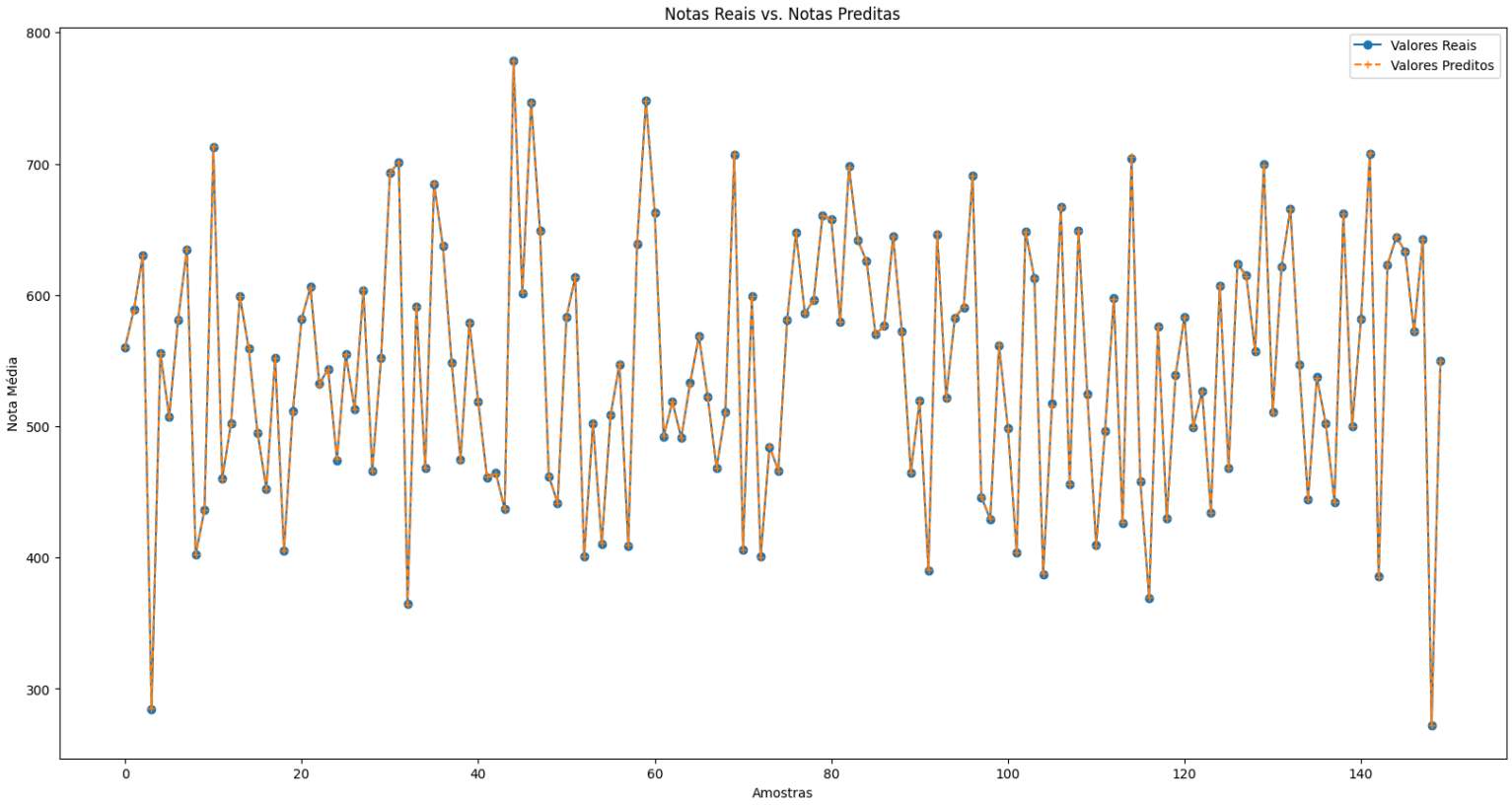
- O MSE é a média dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais, sendo assim, realiza a média do erro ao quadrado do modelo.
- Valores de MSE pequenos indicam um modelo com melhor performance. O valor de 0.14 confirma que a diferença ao quadrado entre os valores previstos e os valores reais é pequena. Concluindo que o modelo está realizando previsões com muita precisão.

$R^2$ Score - Coeficiente de Determinação:

- É a proporção da variância dos dados que é explicada pelo modelo. Ele varia entre 0 e 1, onde 1 indica que o modelo explica perfeitamente toda a variabilidade dos dados, enquanto 0 indica que o modelo não explica nenhuma variabilidade.
- $R^2$ Score de 0.99 é extremamente alto, indicando que o modelo explica quase toda a variabilidade dos dados. Isso sugere que o modelo tem um ajuste muito bom aos dados.

Considerações sobre o Modelo:

- Overfitting: Um $R^2$ Score extremamente alto pode indicar overfitting, especialmente se o modelo foi treinado em um conjunto de dados pequeno. Isso significa que o modelo pode ter se ajustado muito bem aos dados de treinamento, mas pode não generalizar bem para novos dados.



## 5. CONCLUSÃO

A desigualdade é um problema complexo e multifacetado que requer políticas abrangentes e sustentáveis para ser abordado de maneira eficaz. Como já esperado em nosso país, existem grandes níveis de desigualdade em diferentes questões abordadas. Os participantes do Enem vivendo no mesmo país refletem essas características como amostra de um grande número de pessoas representando uma ampla diversidade em termos de idade, origens, escolaridade e classe social, objetivando evoluir em suas carreiras.

A grande maioria dos candidatos é composta por mulheres, principalmente da zona urbana e de diferentes estados, com renda em sua grande maioria sendo considerada muito baixa. Observa-se uma tendência de aumento das notas conforme o acesso da internet se torna comum. Quando verificamos os gráficos com separação feita por estados podemos confirmar que regiões do Nordeste e Norte necessitam de maior atenção para um futuro melhor, ficando claro e objetivo para os tomadores de decisões e principalmente ao governo para que tomem ações direcionadas aos grupos mais vulneráveis.

O tema da educação no Brasil é de enorme relevância, refletindo a diversidade e a extensão do país. A magnitude territorial traz consigo a necessidade de decisões estratégicas e a superação de inúmeros desafios para que o sistema educacional possa realmente florescer. Garantir um acesso mais amplo e equitativo à educação de qualidade é crucial para o desenvolvimento do país. Isso envolve não apenas políticas de inclusão e investimento, mas também a implementação de práticas que atendam às variadas necessidades de diferentes regiões e populações. A prosperidade do sistema educacional brasileiro depende da capacidade de enfrentar essas questões de maneira eficaz e inovadora.

## 6. REFERÊNCIAS BIBLIOGRÁFICAS

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Microdados do Enem 2023. Brasília: Inep, 2024. Disponível em: < https://www.gov.br/inep/ptbr/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 30 abr. 2024.

MELHORIA DAS NOTAS COM ACESSO A INTERNET:
https://extension.okstate.edu/fact-sheets/do-home-computers-internet-access-affect-student-performance.html

https://oapub.org/edu/index.php/ejes/article/view/1892