DESAFIO FINAL SEMANTIX 2021 – BIG DATA SCIENCE

RELATÓRIO ESCRITO: ANÁLISE DOS DADOS DO EXAME NACIONAL DO ENSINO MÉDIO DE 2019

Dezembro/2021

DANIEL DANTAS DO AMARAL RAMOS
daniel.dantas.a.r@gmail.com

SUMÁRIO

1 INTRODUÇÃO

Em 1998 o governo federal do Brasil criou o Exame Nacional do Ensino Médio (ENEM) como um instrumento para avaliar o desempenho dos estudantes no término da educação básica. A partir de 2009 medidas governamentais estimularam o uso do ENEM não apenas como um processo de avaliação do Ensino Médio, mas como forma de acesso ao ensino superior no Brasil. O Sistema de Seleção Unificada (Sisu) passou a operar em larga escala no processo de alocação dos candidatos às vagas (SILVEIRA; BARBOSA; SILVA, 2015).

O Enem é composto de 180 questões, distribuídas em quatro provas objetivas: Ciências Humanas e suas Tecnologias (História, Geografia, Filosofia e Sociologia); Ciências da Natureza e suas Tecnologias (Química, Física e Biologia); Linguagens, Códigos e suas Tecnologias (Língua Portuguesa, Literatura, Língua Estrangeira - Inglês ou Espanhol, Artes, Educação Física e Tecnologias da Informação e Comunicação); Matemática e suas Tecnologias (Matemática). Além disso, os alunos devem fazer uma redação e responder a um questionário socioeconômico-cultural. O INEP disponibiliza todas as respostas do questionário bem como as notas em cada uma das áreas de conhecimento por meio dos micro dados do Enem.

A partir das análises dos dados produzidos pela aplicação desse exame, é possível observar o desempenho tanto do estudante quanto das instituições e, assim, calcular indicadores de qualidade que, dentro de um contexto, oportunizarão decisões de melhorias do processo de ensino e aprendizagem (LIMA et al., 2019).

Com esta análise objetiva-se identificar padrões, tendências e oportunidades para habilitar tomadores de decisão (governantes, executivos, coordenadores e demais) com insumos para a tomada de decisões baseadas em dados (mais assertivas e ágeis) na busca de melhoria na educação nacional.

2 MATERIAIS E MÉTODOS

O universo de estudo é composto pelos alunos que prestaram o Exame Nacional do Ensino Médio (Enem) em 2019. Ao realizar a prova, estes devem responder um questionário com diversas perguntas de cunho social e econômico e estas respostas foram armazenadas e disponibilizadas ao público por meio dos micro dados do Enem (https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem), juntamente com outros dados referentes à prova. Ao total, foram disponibilizadas 136 features, passando por dados sobre o participante, a escola, sobre atendimentos e recursos especializados e específicos, o local de aplicação, sobre a prova em si e um questionário socioeconômico, dos quais foram selecionados para serem utilizados no estudo, os seguintes:

- SG_UF_RESIDENCIA – Estado de residência;
- NU_IDADE – Idade do aluno;
- TP_SEXO – Sexo declarado pelo aluno;
- TP_COR_RACA – Cor/raça declarada pelo aluno;
- TP_ESCOLA - Tipo de escola (pública ou privada);
- TP_DEPENDENCIA_ADM_ESC – Dependência administrativa da escola;
- TP_LOCALIZACAO_ESC – Se a escola é rural ou urbana;
- NU_NOTA_CN – Nota em Ciências da Natureza;
- NU_NOTA_CH – Nota em Ciências Humanas;
- NU_NOTA_LC – Nota em Linguagens e Códigos;
- NU_NOTA_MT – Nota em Matemática;
- NU_NOTA_REDACAO – Nota em Redaçaõ;
- Q001 – Escolaridade do pai;
- Q002 – Escolaridade da mãe;
- Q005 – Quantas pessoas moram na residência do aluno;
- Q006 – Renda mensal da família;
- Q025 – Acesso à internet;

Essas variáveis foram apresentadas através de tabela no formato CSV e foram avaliadas com relação a medidas descritivas, com a identificação de padrões, tendências e oportunidades com o objetivo de traçar o perfil dos alunos prestadores do Enem 2019 para auxiliar tomadores de decisão públicos e privados em ações no âmbito educacional.

3 ANÁLISE DESCRITIVA DOS DADOS

O conjunto de dados analisados consiste de informações de 5095270 alunos, com 136 *features* (colunas) com informações diferentes, de 27 estados e 5570 municípios diferentes, com 82 idades diferentes relatadas, divididos entre brasileiros, naturalizados e estrangeiros, com representantes de cada uma das 5 cores/raças oficiais.

Tamanha diversidade encontrada nos dados corrobora com o que (SILVEIRA; BARBOSA; SILVA, 2015) classificou como um dos aspectos positivos do Enem: A mobilidade estudantil para instituições de ensino superior no Brasil. Esta mobilidade possibilita que sujeitos oriundos de regiões menos desenvolvidas desloquem-se para outras mais desenvolvidas, além de criar lideranças e mão de obra especializada nas regiões não desenvolvidas. Todavia esta mobilidade pode, segundo o autor, ser comprometida pelo fato de que os estudantes dos estados mais ricos conseguem melhores notas e conseguem mais facilmente ocupar vagas em outros estados.

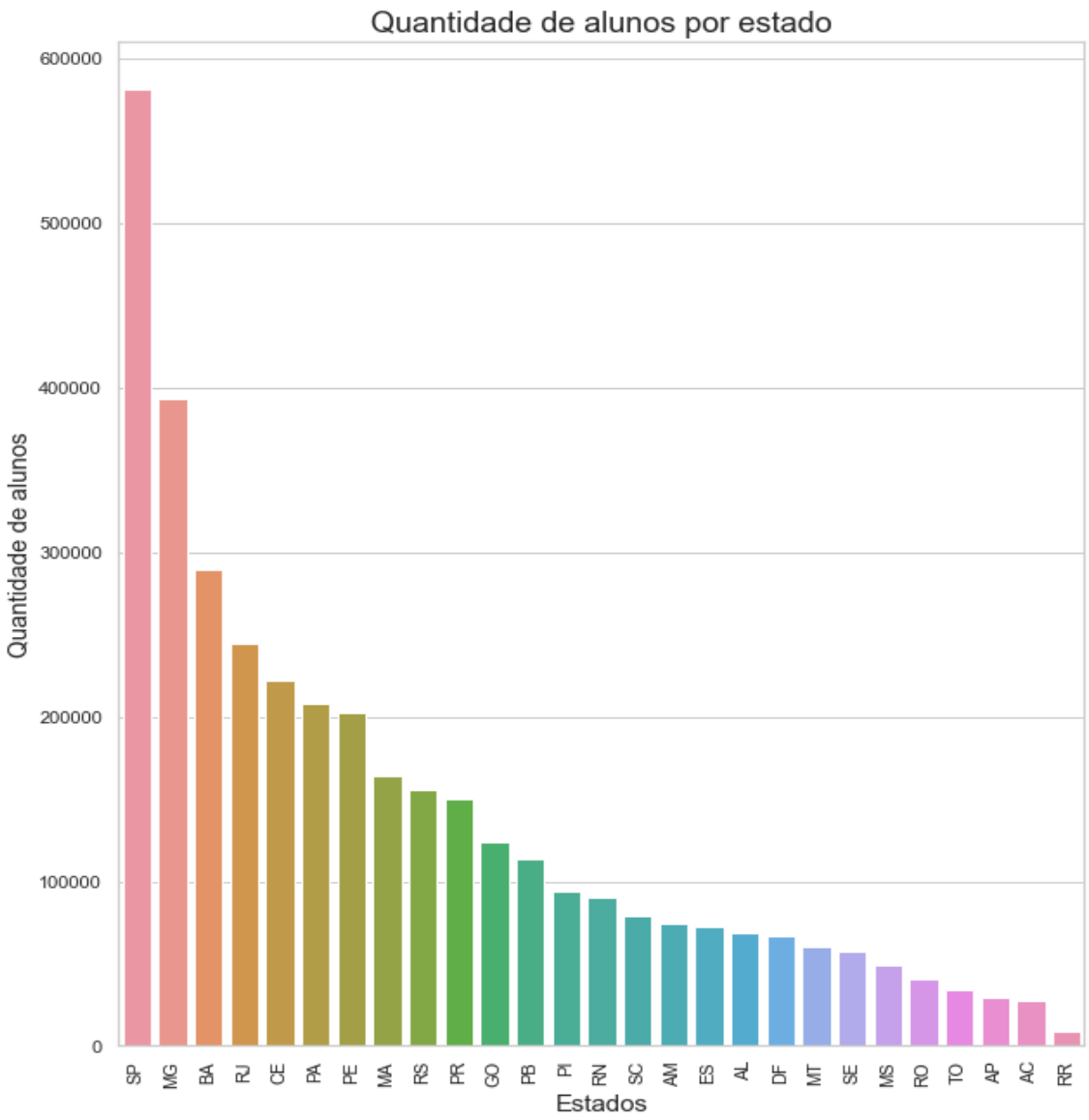
Um conjunto de dados tão grande possui, naturalmente, alguns valores faltantes e nulos, principalmente. Para tais valores, foram feitas etapas de processamento e limpeza de dados que serão explicitadas nos seus respectivos itens dentro do *Jupyter Notebook*.

Para iniciar nosso entendimento sobre os dados, vamos analisar e visualizar as variáveis categóricas e numéricas e após isso fazer e responder perguntas que suportem nossas suposições.

3.1 Análise de variáveis categóricas

Vamos visualizar cada uma das variáveis categóricas e entender como se comportam as características dos candidatos no Enem 2019.

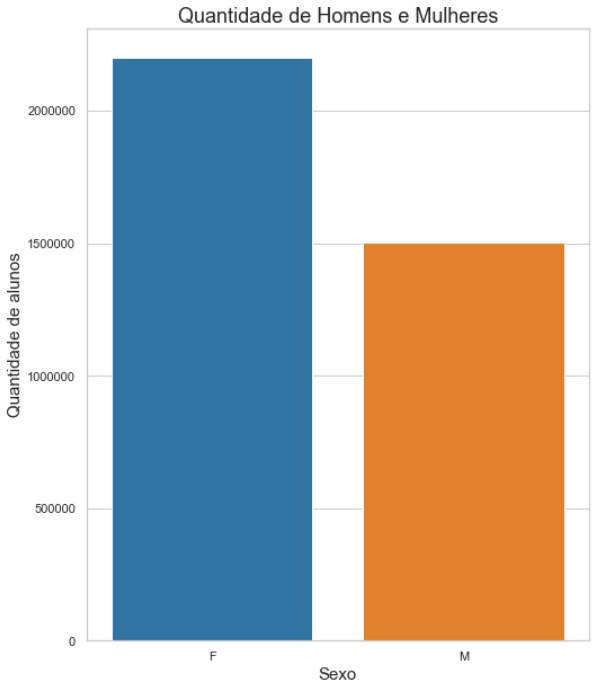
Figura 1 – Quantidade de alunos por estado



Fonte: Autor

Percebemos na figura 1 que a região Sudeste, representada principalmente pelo estado de São Paulo, conta com a maioria dos representantes na prova do Enem 2019, mais especificamente 28,96%. Note que alguns estados como Bahia, Ceará e Pará se destacam possuindo uma quantidade considerável de participantes.

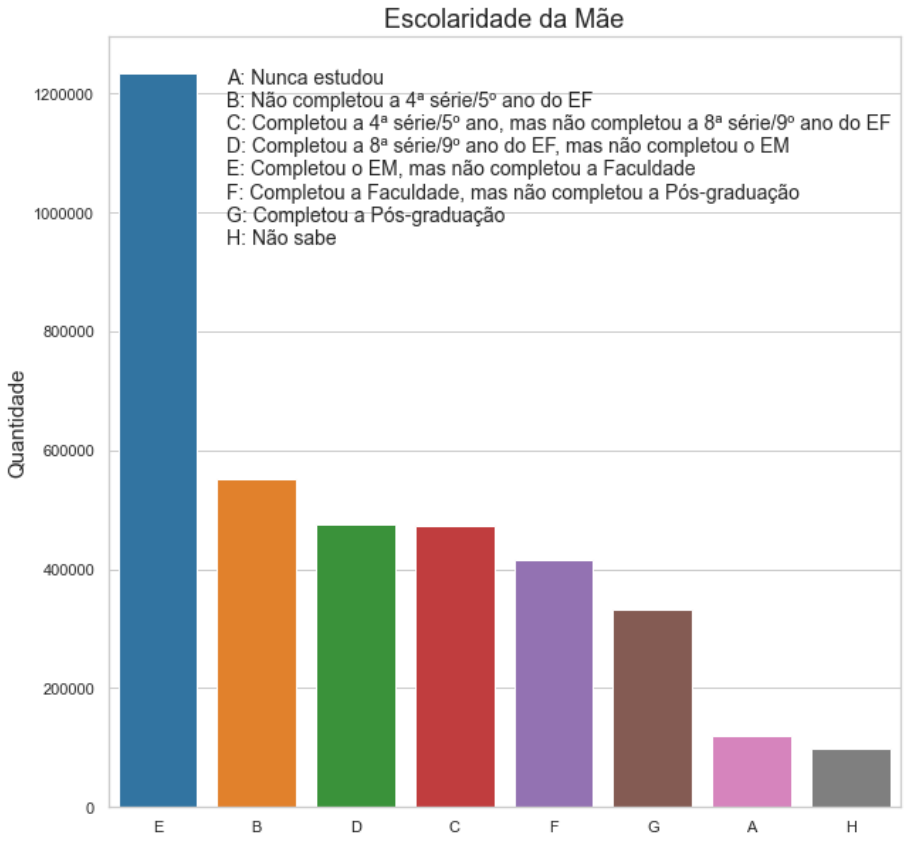
Figura 2 – Quantidade de alunos por estado



Fonte: Autor

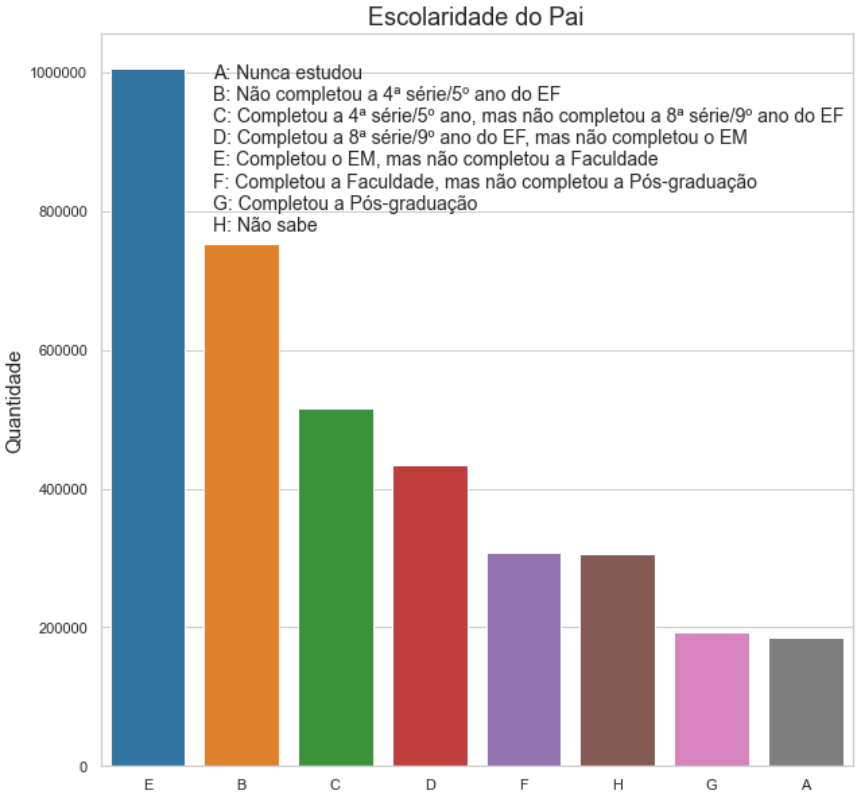
Percebemos que a maioria dos participantes é mulher, com 58.19% das presenças na prova.

Figura 3 – Escolaridade da Mãe



Fonte: Autor

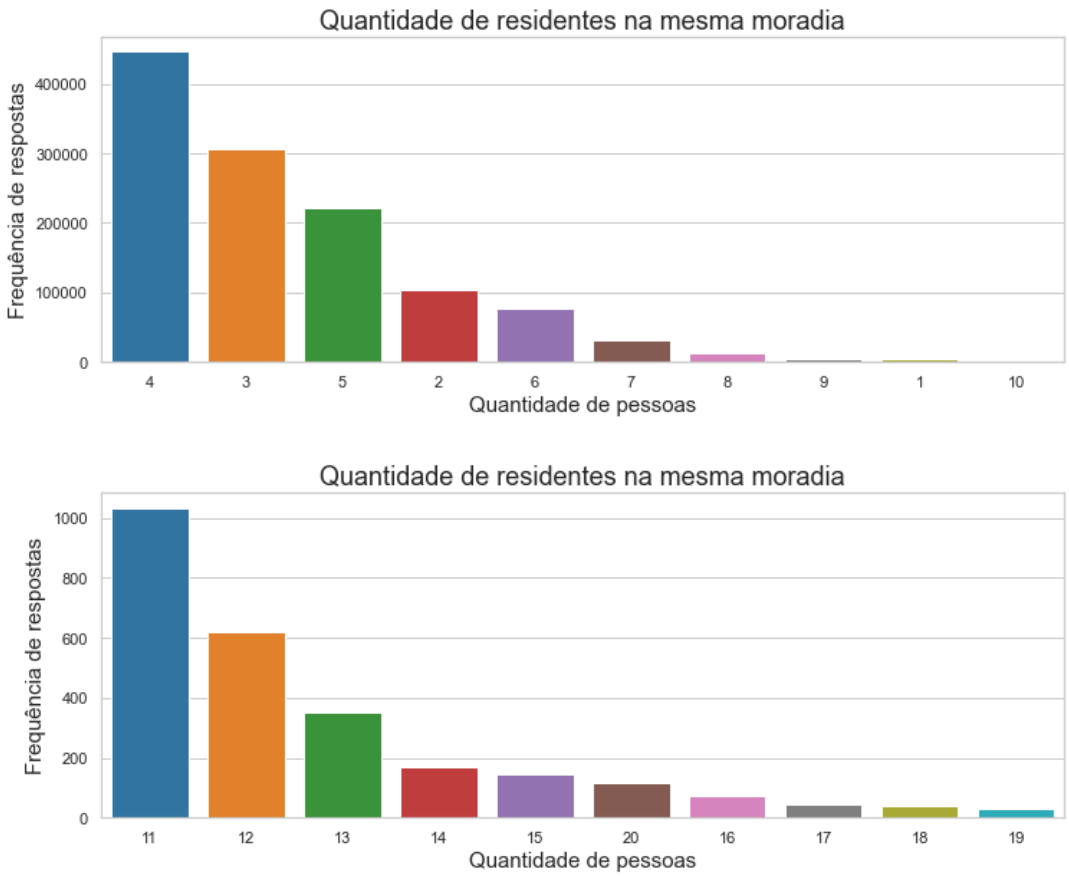
Figura 4 – Escolaridade do Pai



Fonte: Autor

Analisando ambos os gráficos, percebemos que a maioria dos candidatos do Enem possui pais com o Ensino Médio completo, o que reflete na renda da família. Um ponto interessante de se observar é que mais candidatos não sabem a escolaridade do pai do que da mãe, o que pode ser explicado por alguns fatores sociais no Brasil, como a taxa de mortalidade dos homens ser maior do que das mulheres, ou até o índice de abandono dos pais de suas famílias.

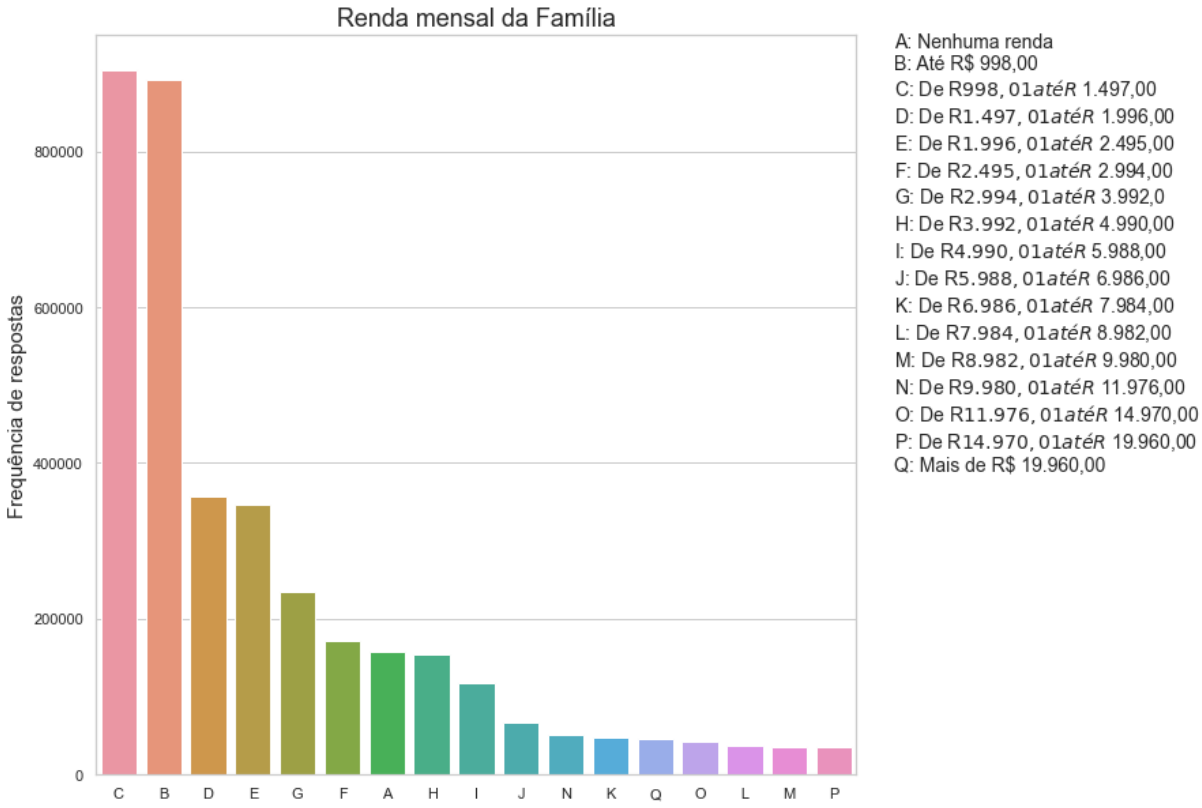
Figura 5 – Residentes na mesma moradia



Fonte: Autor

Para melhor visualização da quantidade de residentes na mesma moradia, optou-se por separar em 2 grupos, de 0 a 10 residentes e de 11 a 19. O gráfico de cima deixa claro que a maioria dos candidatos vive em uma casa com 4 pessoas (o aluno e mais 3). O gráfico de baixo nos mostra valores muito menores de frequência de resposta, onde a resposta “11” apareceu em torno de 1000 vezes. Ou seja, em torno de 1000 alunos vivem em uma casa com mais 10 pessoas (11 ao total). Os dados são consistentes com informações advindas do IBGE (https://brasilemsintese.ibge.gov.br/populacao/taxas-de-fecundidade-total.html), que nos fala que a taxa de fecundidade em 2002 era de 2,26 filhos por mulher. Escolhemos 2002 pois, como a maioria dos candidatos no Enem de 2019 tem 17 anos, a taxa de fecundidade de 2002 nos dá uma boa estimativa da quantidade de filhos que os casais tiveram.

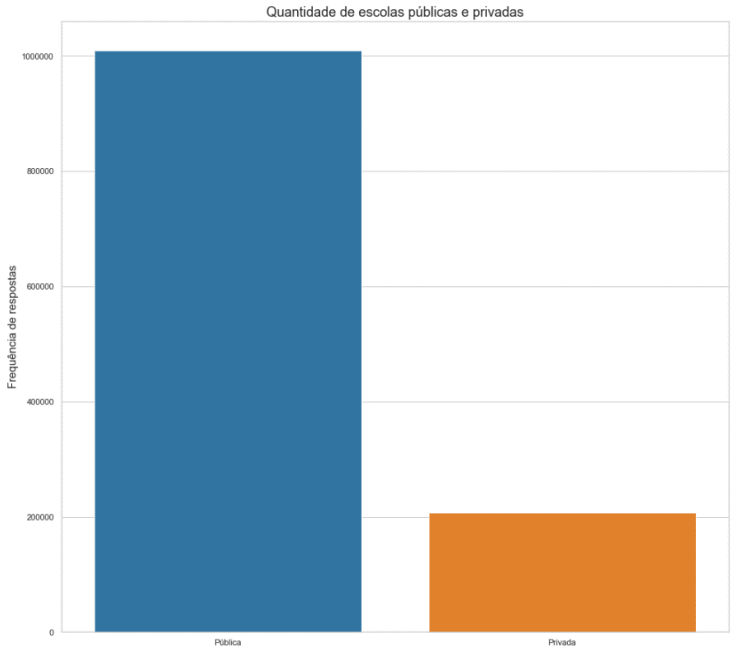
Figura 6 – Renda mensal das famílias



Fonte: Autor

Este é um gráfico que nos dá uma noção clara da imensa desigualdade social que existe no Brasil (em 2019), com a maioria da população estando nas condições B ou C (mais precisamente, 48% das respostas).

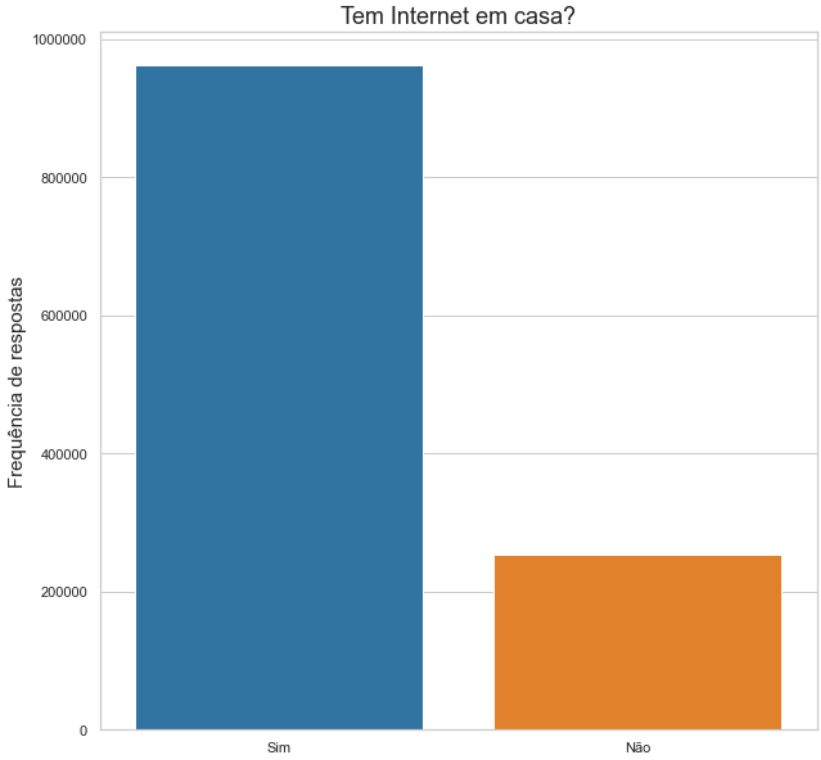
Figura 7 – Escolas públicas e privadas



Fonte: Autor

Já a figura 7 nos mostra que a maioria das escolas das quais os candidatos provém são públicas. Este dado nos levanta a pergunta: Será que as escolas públicas conseguem competir com as privadas? Como está o desempenho desses 2 tipos de escola?

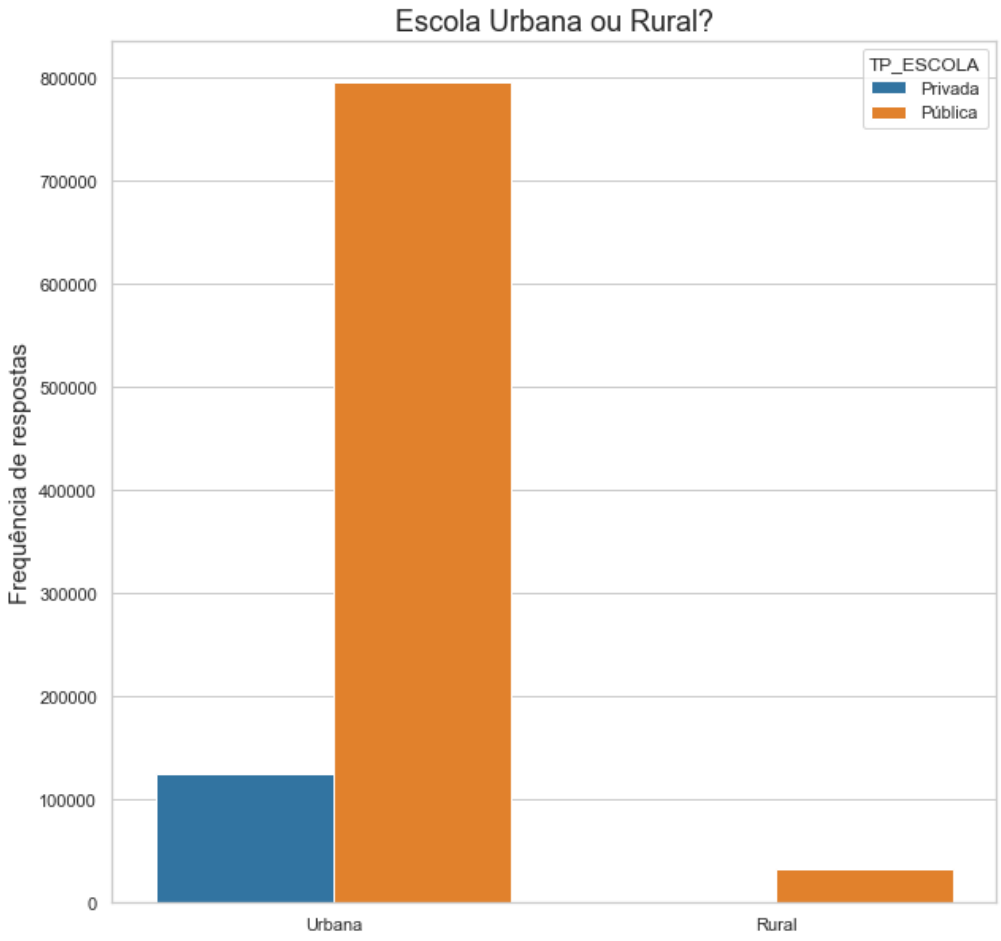
Figura 8 – Escolas públicas e privadas



Fonte: Autor

Este gráfico nos mostra que a maioria dos alunos tem acesso à internet em suas casas, mesmo uma parte considerável ainda carecendo deste recurso. Isso pode ser explicado pelo grande avanço da infraestrutura de comunicação e democratização de aparelhos e internet móvel no Brasil dos últimos anos.

Figura 9 – Escolas Urbanas e Rurais



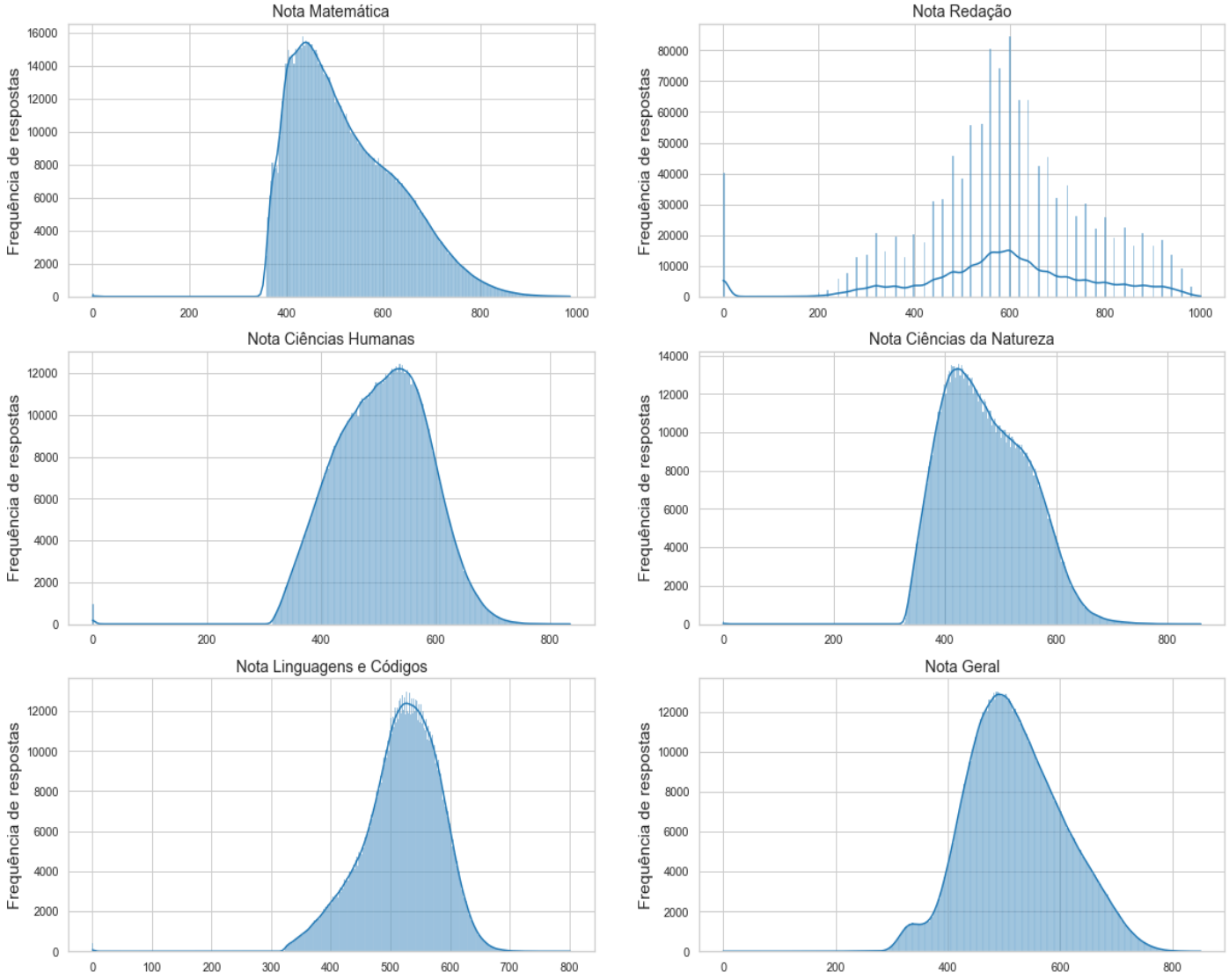
Fonte: Autor

A análise do gráfico e dos dados nos permite ver que as escolas urbanas representam 96.53% dos nossos dados. Elas tem médias melhores do que as escolas rurais em todas as áreas do conhecimento, por exemplo em matemática, onde as escolas urbanas possuem média de 637 (privada) e 507 (pública) enquanto que para as escolas rurais, a média de matemática foi de 552 (privada) e 496 (pública).

3.2 Análise de variáveis quantitativas

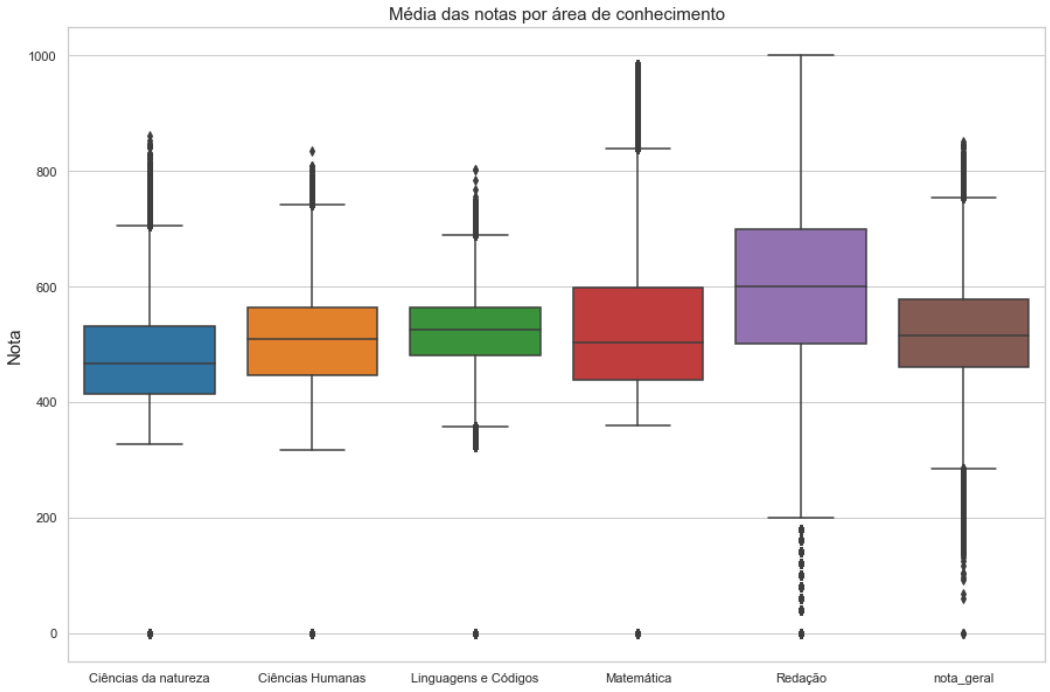
Em resumo, as variáveis quantitativas que iremos analisar são as notas dos candidatos nas diferentes áreas de conhecimento do Enem.

Figura 10 – Notas nas áreas de conhecimento



Fonte: Autor

Figura 11 – Boxplot: Áreas de conhecimento



Fonte: Autor

Note na figura 10 que, todas as distribuições tem curvas próximas da normal, exceto pelas notas de redação. Com conhecimento de área, é possível explicar este fenômeno da seguinte maneira: Diferentemente das outras notas, que são calculadas segundo o TRI (Teoria da resposta ao item), a nota da redação é calculada segundo correção de profissionais individuais, e suas notas são todas computadas em intervalos de 20 em 20 pontos, ex: 540, 560, 580, 600.

Note que a média é a curva que mais se aproxima de uma normal, o que pode ser descrito pelo Teorema central do limite, que diz que quanto maior o tamanho da amostra, a distribuição amostral da sua média tende a uma distribuição normal. Já que temos uma amostra bem grande, podemos ver esse teorema na prática. (TEOREMA CENTRAL DO LIMITE – WIKIPÉDIA, A ENCICLOPÉDIA LIVRE, [*s. d.*])

A figura 11 nos dá a noção visual da tabela 1 a seguir:

Tabela 1 – Estatísticas básicas das notas

	Ciências da natureza	Ciências Humanas	Linguagens e Códigos	Matemática	Redação	Nota geral
Quantidade	1217154,000	1217154,000	1217154,000	1217154,000	1217154,000	1217154,000
Média	474,332	505,987	519,045	523,982	583,566	521,382
Desvio padrão	75,366	80,258	63,742	108,201	190,282	85,845
min	0,000	0,000	0,000	0,000	0,000	0,000
25%	414,300	447,000	481,500	437,200	500,000	460,980
50%	466,100	509,000	524,800	502,000	600,000	514,140
75%	530,400	564,600	564,300	597,600	700,000	578,480
max	860,900	835,100	801,700	985,500	1000,000	850,820

Fonte: Autor

A tabela 1 e figura 11 nos mostram que a área de conhecimento que os alunos obtiveram a maior média foi a redação, o que talvez se deva ao fato de que a nota da redação possa chegar até 1000 pontos. Todavia, foi redação também que obteve o maior desvio padrão, de 190, o que evidencia que as notas na redação são bastante discrepantes, com alunos obtendo notas muito altas enquanto outros obtém notas muito baixas, o que será evidenciado mais a frente. Já a menor média foi a de ciências da natureza (474,33), que conta também com o menor desvio padrão (75,36), mostrando que a tendência dos alunos como um todo é não obter boas médias nesta área do conhecimento.

O menor desvio padrão foi de 63,7 em Linguagens e códigos, isso significa que as notas nessa área do conhecimento são menos díspares. Como esperado, as notas mínimas em todas as áreas do conhecimento foram 0 (zero), e a nota máxima de redação e matemática foram as maiores, seguindo o definido pelo INEP.

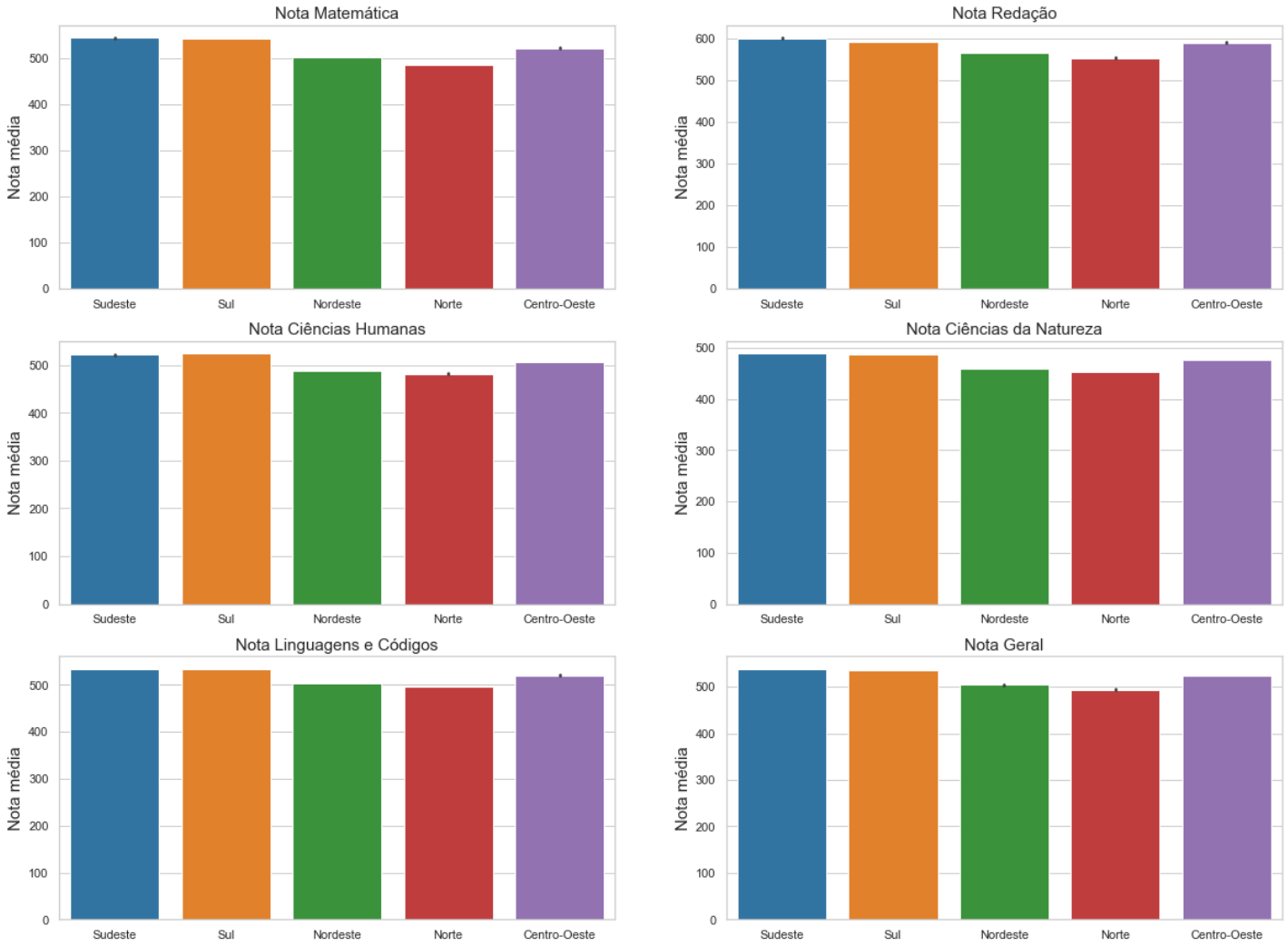
4 PERGUNTAS

Para nos guiar em nossa análise, agora que obtivemos uma melhor noção dos dados visualizando suas *features*, vamos redigir perguntas e basear nossas análises nestas. Tais perguntas tem por objetivo identificar algum padrão e/ou oportunidade, investigar alguma característica ou suposição e apresentar possíveis direções para os órgãos públicos aos quais esta análise será dirigida.

4.1 Quais regiões obtiveram as melhores notas?

Nosso objetivo com esta pergunta é enxergar e analisar as diferenças entre as regiões oficiais do Brasil, para identificar algum item de atenção que pode ser direcionado aos órgãos competentes, já que sabemos haver diferenças socioeconômicas que podem significar diferentes desempenhos.

Figura 12 – Notas por região

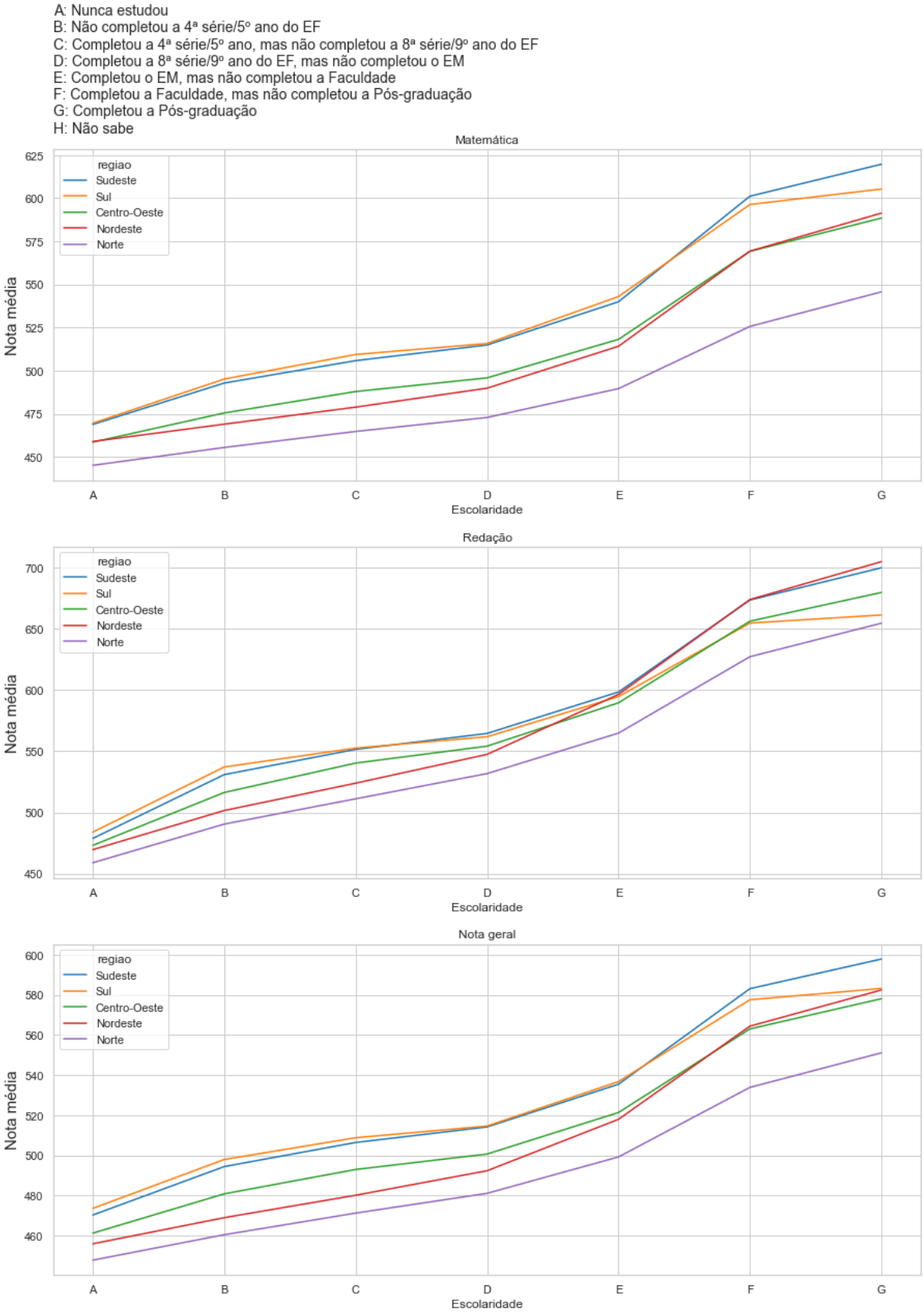


Fonte: Autor

A análise dos gráficos evidencia a diferença entre as regiões brasileiras, onde a região sudeste obteve as melhores médias e a região norte as piores, em todas as áreas de conhecimento. Segundo (GUIMARÃES; BRITO; SANTOS, 2020) a distribuição de renda é desproporcional entre as regiões, situação constituída historicamente mas que ainda hoje reflete a realidade do cenário nacional, tendo como fatores causadores desde a colonização e exploração de terras, o plano de desenvolvimento do país até a industrialização desigual. Nesse contexto, o eixo Sul-Sudeste se sobressaiu e reflete hoje melhor qualidade econômica e social, o que influencia nas maiores médias de seus alunos em todas as áreas do conhecimento.

4.2 Ter pais letrados influencia na nota?

Figura 13 – Notas por escolaridade



Fonte: Autor

Com esta pergunta, queremos entender como a dinâmica social de escolaridade dos pais influencia no comportamento, nível social e notas dos filhos no Enem, para entender quais grupos precisam de mais ajuda/amparo do governo, por exemplo.

Com a análise do gráfico, fica evidente que quanto maior a escolaridade da mãe, maior a nota dos candidatos. O que pode ser explicado pelo maior acesso ao letramento e a atividades e itens culturais pelos candidatos, frequentemente relacionado também ao maior nível econômico, o que permite mais acesso à informação e às melhores escolas para os alunos.

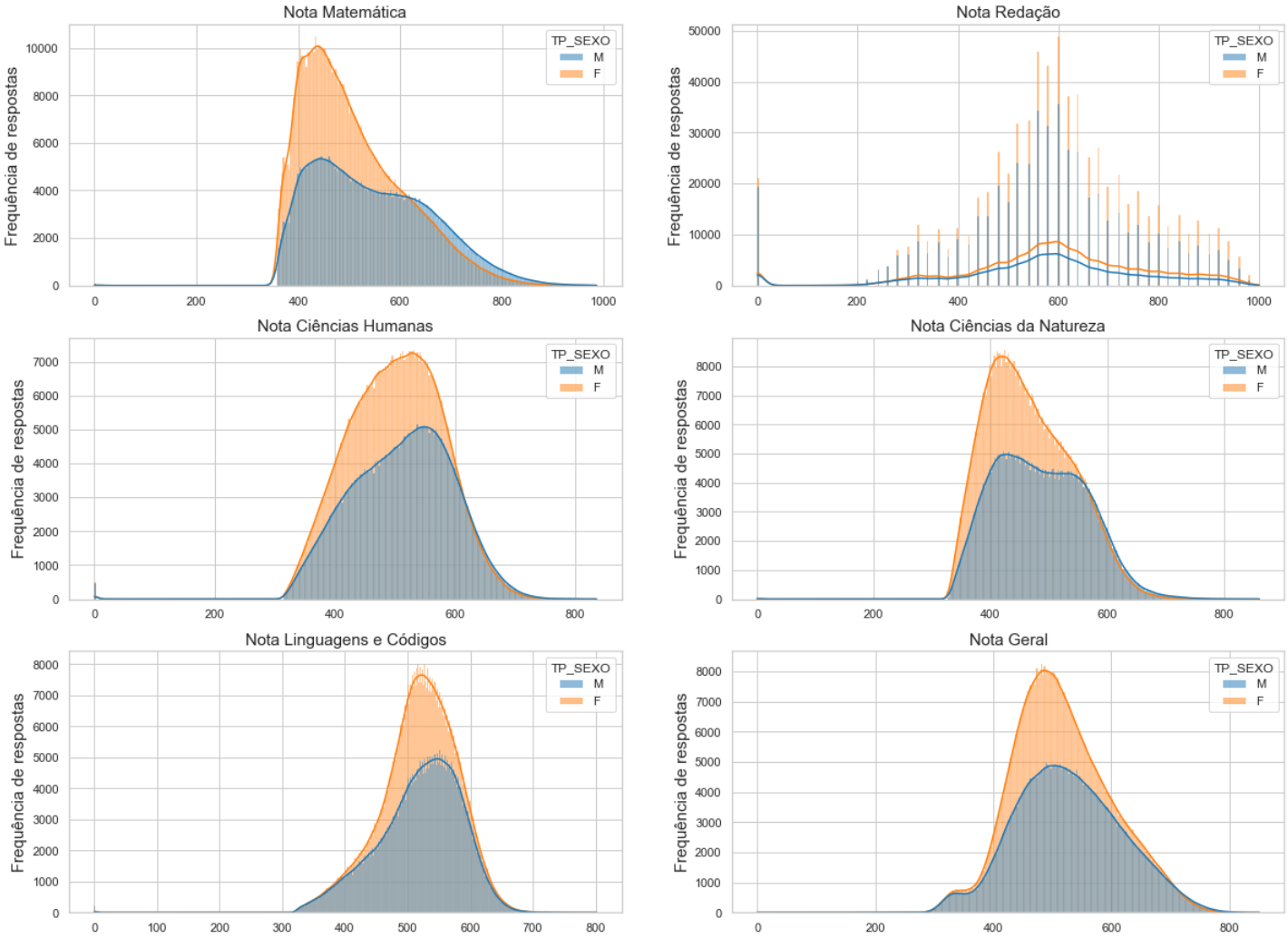
Segundo o site governamental (CRIANÇA LIVRE DE TRABALHO INFANTIL, [*s. d.*]), algumas das causas para o trabalho infantil acontecer são pobreza, má qualidade da educação e questões culturais. Vemos que nas classes mais desfavorecidas (A a D), o nível socioeconômico baixo, a pobreza e a necessidade de se sustentar pode ser uma causa para a evasão (parcial ou total) dos candidatos da escola e consequente nota baixa.

Note que, mesmo nas classes mais favorecidas, a diferença regional ainda é clara, com os alunos da região sudeste tendo notas gerais maiores que as demais regiões, mesmo com o mesmo nível de escolaridade. Por exemplo, para pessoas com mães que completaram a pós graduação, a média geral foi de aproximadamente 600 na região sudeste e 550 na região norte.

4.3 Há diferenças nas notas de homens e mulheres?

Nosso objetivo com esta pergunta é entender se a diferença de sexo é um fator determinístico para um melhor desempenho ou não para guiar possíveis condutas e programas governamentais.

Figura 14 – Notas de Homens e Mulheres



Fonte: Autor

Tabela 2 – Notas médias por sexo

_SEXO	NOTA_CN	NOTA_CH	NOTA_LC	NOTA_REDACAO	NOTA_MT	Nota geral
Feminino	466,24	501,41	518,85	591,31	507,63	517,09
Masculino	485,58	512,35	519,30	572,77	546,73	527,35

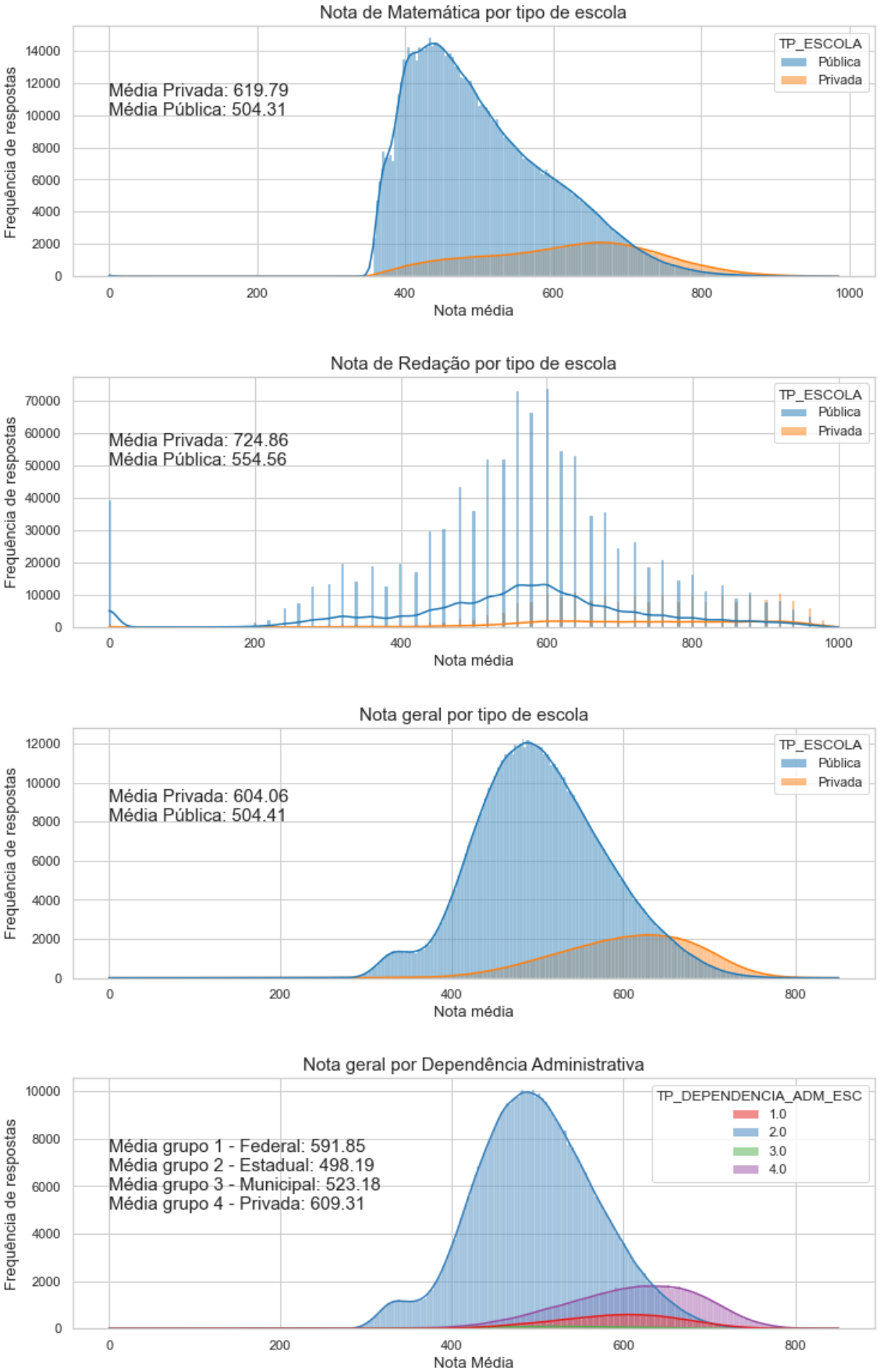
Fonte: Autor

As mulheres são maioria no Enem, todavia como evidenciado pela figura 14 e tabela 2, sua média nas áreas de conhecimento são menores que as dos homens, principalmente nas áreas de matemática (546,7) e ciências da natureza (485,5), mas se sobressaindo na nota de redação (591,31), onde conseguiram uma média maior. Tais discrepâncias podem ser entendidas por desigualdades de gênero e fatores sociais, como a baixa expectativa em relação ao desempenho das meninas nas notas de exatas.

4.4 Há diferença entre escolas públicas e privadas?

Com esta pergunta queremos saber se o tipo de escola influencia na educação e capacidade de tirar boas notas do aluno, para que este dado seja visível e de fácil entendimento para os tomadores de decisão, já que a instituição de ensino é um fator muito importante para a formação dos estudantes.

Figura 15 – Notas por tipo de escola



Fonte: Autor

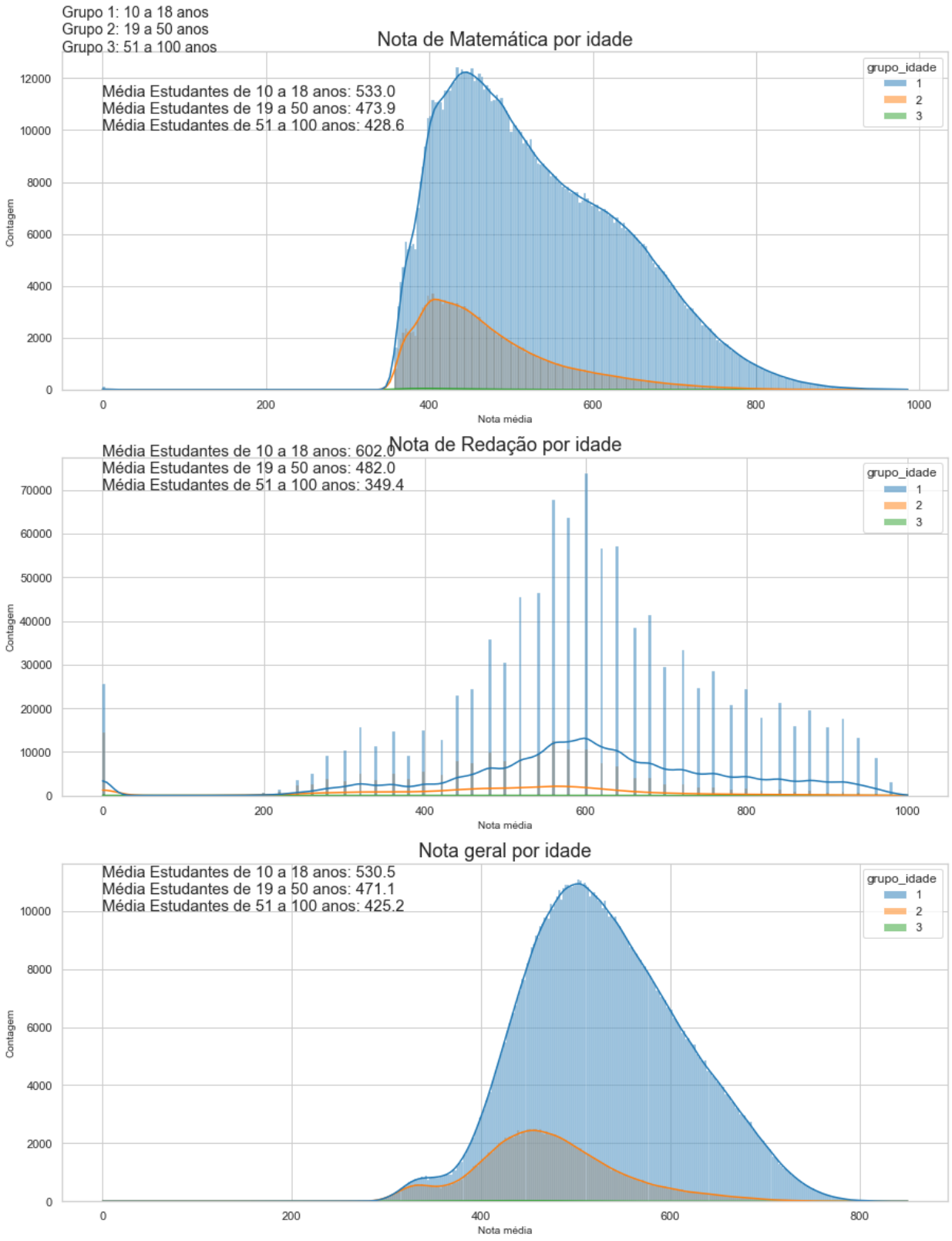
Com a análise dos gráficos fica evidente que as escolas particulares possuem médias (604 pontos) muito superiores em relação as escolas públicas (504 pontos). Estas observações podem ser explicadas, segundo (FEIJÓ; FRANÇA, 2021), por fatores como: composição socioeconômicas entre os alunos, o ambiente familiar, a renda, a escolaridade dos pais e o *background* dos colegas de classe, já que esse estudante vai interagir diretamente com seus colegas, características essas que divergem de escolas públicas e privadas.

Além disso, dentre as escolas públicas, fica evidente que as escolas com a sua administração na esfera federal possuem melhores notas (591 pontos), quase se equiparando com as notas das escolas particulares, em relação às notas de escolas municipais (523 pontos) e estaduais (498 pontos). Isso pode ser explicado, dentre outros fatores, pois as escolas federais podem possuir algum tipo de convênio com universidades, centros de educação tecnológica voltado para o mercado, dentre outros programas educacionais que recebem verba diretamente do governo federal.

4.5 A idade influencia na nota?

Com esta pergunta, pretendemos analisar se os candidatos mais velhos estão capacitados a fazer o Enem tão bem quando os mais novos, e isso é necessário para avaliar a capacidade de reinserção nas instituições de ensino de adultos que podem, por exemplo, ter passado por alguma instituição de EJA (Educação de Jovens e Adultos).

Figura 16 – Notas por idade



Fonte: Autor

Dentre os candidatos, a pessoa de maior idade foi de 94 anos e a de menor idade foi a de 10 anos! Este dado pode estar errado, todavia existiram 6 pessoas com 10 anos, o que dilui esta possibilidade. 50% ou menos dos candidatos possuíam 19 anos e 75% dos candidatos possuíam 24 anos ou menos.

Da análise dos dados, vemos que a idade é sim um fator que influencia na nota do candidato, o que tem explicações sociais, já que os candidatos que irão utilizar o Enem para adentrar em uma universidade estão, em média, na faixa etária de 17 a 18 anos. Portanto, é esta faixa etária que mais estuda e se prepara ativamente para o exame. Claro que pessoas de outras faixas etárias estudam e se preparam para o exame, mas a sua quantidade é muito reduzida, já que na fase adulta, as pessoas tem outras preocupações e não podem dedicar todo o seu tempo aos estudos. Ainda vemos que existem bastante pessoas mais novas fazendo o Enem, o que pode ser elucidado lembrando do fato de que que estas pessoas irão fazer o Enem no futuro e fazem o exame como um teste, uma preparação, são os chamados treineiros.

Após tratar nossos dados, fazemos uma separação em 3 grupos para facilitar a análise, o grupo 1 é composto por alunos com idade até 18 anos, o grupo 2 por alunos com idade entre 19 e 50 anos e o grupo 3 com alunos de idades entre 51 a 100 anos. A média geral do grupo 2 e 3 (471,1 e 425,2) é menor do que a do grupo 1 (530,5), o que comprova que estas faixas etárias tem mais dificuldades em adentrar instituições de ensino. Em termos de decisão, estes dados podem servir para tomadores de decisão eu cuidam da educação de adultos, os orientando a melhoras seu planejamento, após uma análise mais específica.

4.6 A renda familiar influencia na nota?

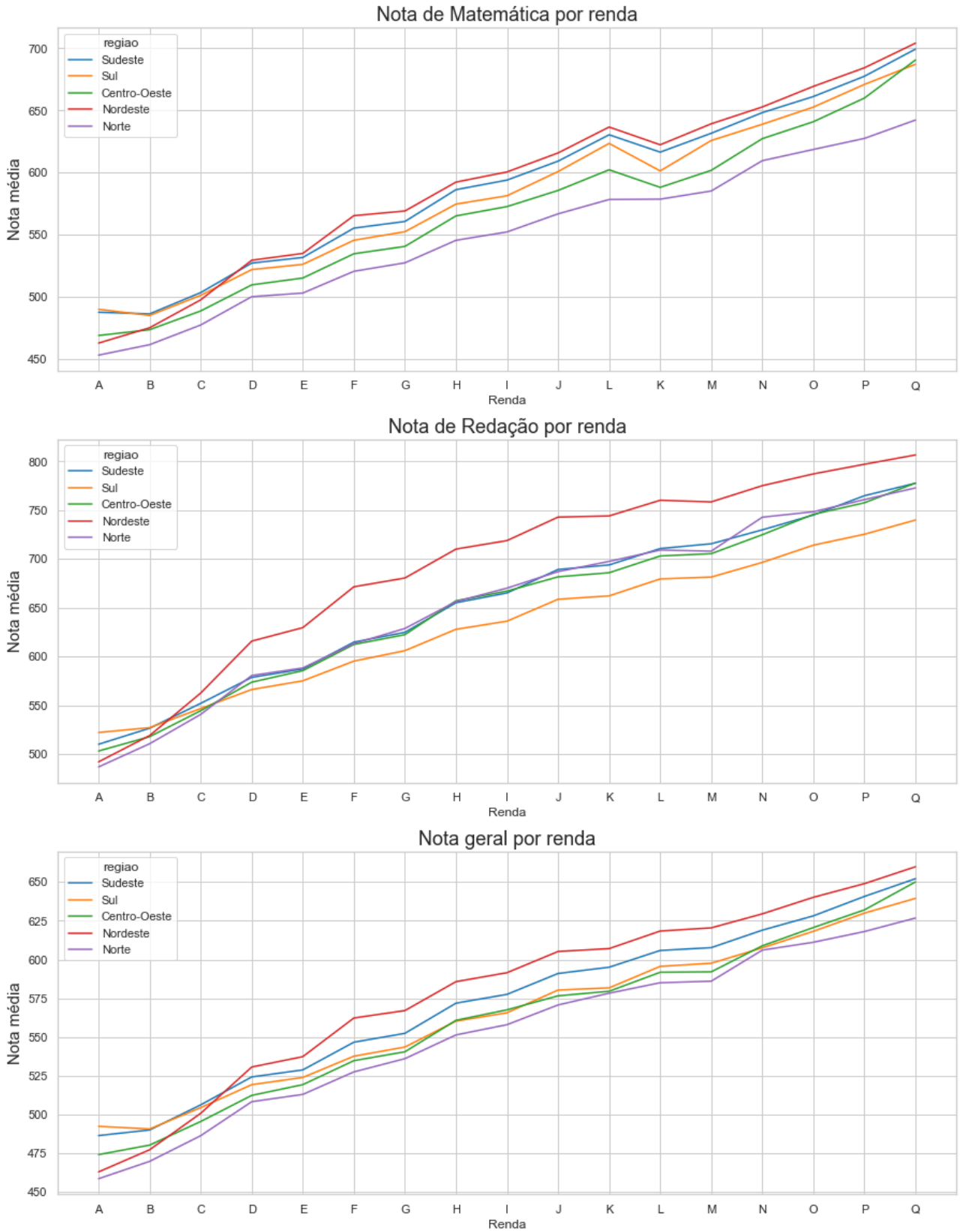
Com esta pergunta, nosso objetivo é evidenciar e comprovar se a renda das familias é um fator importante na nota do Enem. Já que supomos este ser um fator importante para determinar o nível sociocultural de um aluno e, portanto, suas habilidades cognitivas para resolução de questões, interpretação e escrita. A figura 18 segue a legenda da figura 17.

Figura 17 – Legenda

```
A: Nenhuma renda
B: Até R$ 998,00
C: De R$ 998,01 até R$ 1.497,00
D: De R$ 1.497,01 até R$ 1.996,00
E: De R$ 1.996,01 até R$ 2.495,00
F: De R$ 2.495,01 até R$ 2.994,00
G: De R$ 2.994,01 até R$ 3.992,0
H: De R$ 3.992,01 até R$ 4.990,00
I: De R$ 4.990,01 até R$ 5.988,00
J: De R$ 5.988,01 até R$ 6.986,00
K: De R$ 6.986,01 até R$ 7.984,00
L: De R$ 7.984,01 até R$ 8.982,00
M: De R$ 8.982,01 até R$ 9.980,00
N: De R$ 9.980,01 até R$ 11.976,00
O: De R$ 11.976,01 até R$ 14.970,00
P: De R$ 14.970,01 até R$ 19.960,00
Q: Mais de R$ 19.960,00
```

Fonte: Autor

Figura 18 – Notas por renda



Fonte: Autor

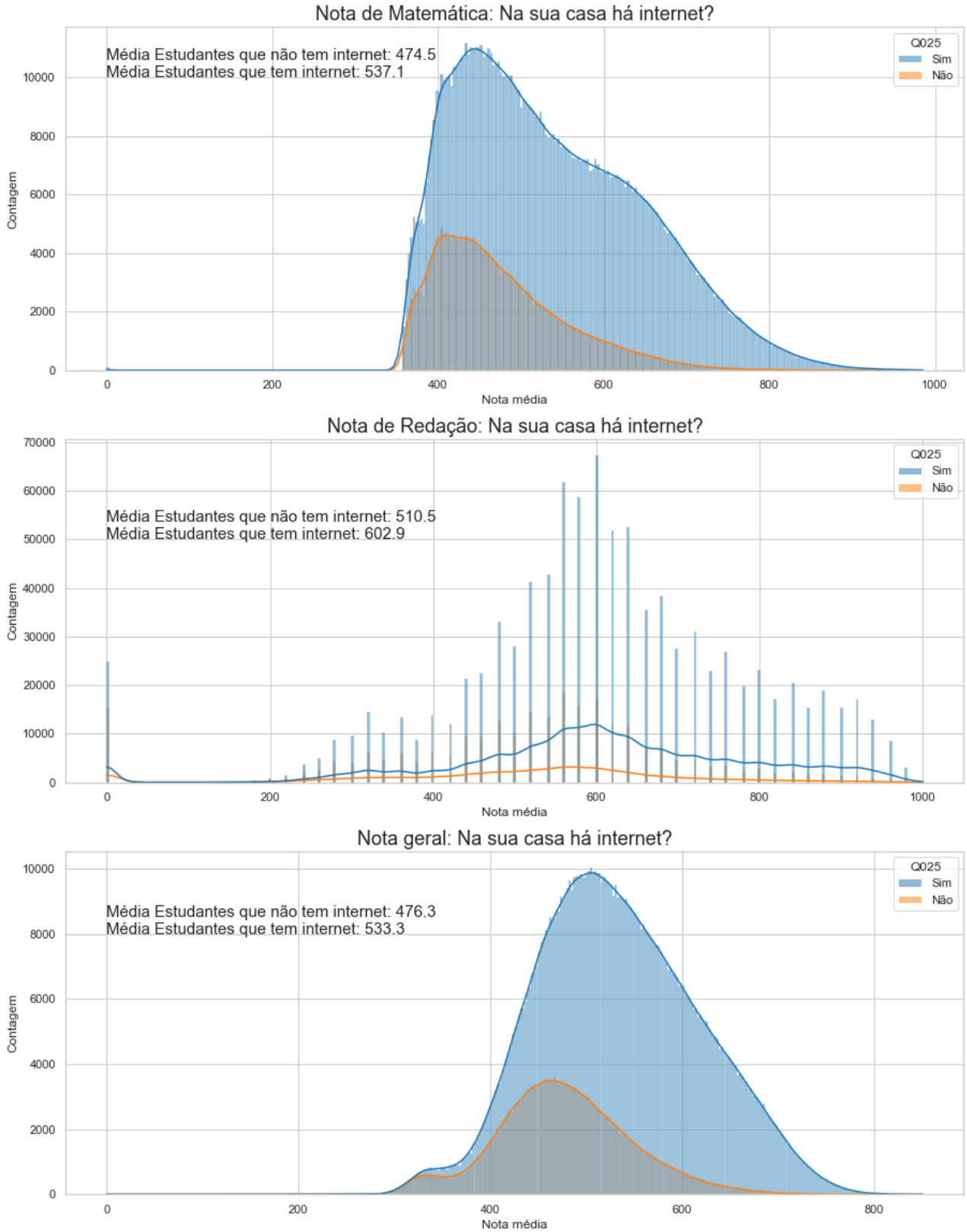
Os gráficos comprovam nossa suposição e mostram ainda mais. Ainda aqui é notório a relação quase linear que a renda das famílias tem com a nota dos candidatos, afinal de contas, escolaridade e renda estão conectados.

Note que, mesmo subdividido por região, a expectativa inicial (assim como aconteceu em nível de escolaridade da mãe) era que a região sudeste obtivesse as melhores notas e a região norte as piores. Todavia, o que vemos agora é um destaque da região Nordeste, que supera as demais, com média geral de aproximadamente 675 para famílias de renda maior que R$ 19.000, enquanto que para a mesma renda, a região sudeste, por exemplo, obteve média geral de aproximadamente 650 pontos. Esta constatação gera dúvidas e possíveis oportunidades de análises. Por que, quando comparado por escolaridade a região sudeste se destaca e comparado por renda a região nordeste se destaca? As pessoas de maior renda no nordeste tem acesso a melhores escolas? Tais perguntas podem e devem ser respondidas com a ajuda de outros bancos de dados com informações mais específicas.

4.7 Acesso à internet influencia na nota?

Com esta pergunta, queremos verificar e evidenciar a importância que a internet tem no processo educacional de hoje em dia, não só no quesito de acesso a aulas e a material didático, como também no acesso a cultura, história e outros quesitos importantes para a formação de opinião e senso crítico (que ajuda em provas como Linguagens, Ciências humanas e redação, por exemplo).

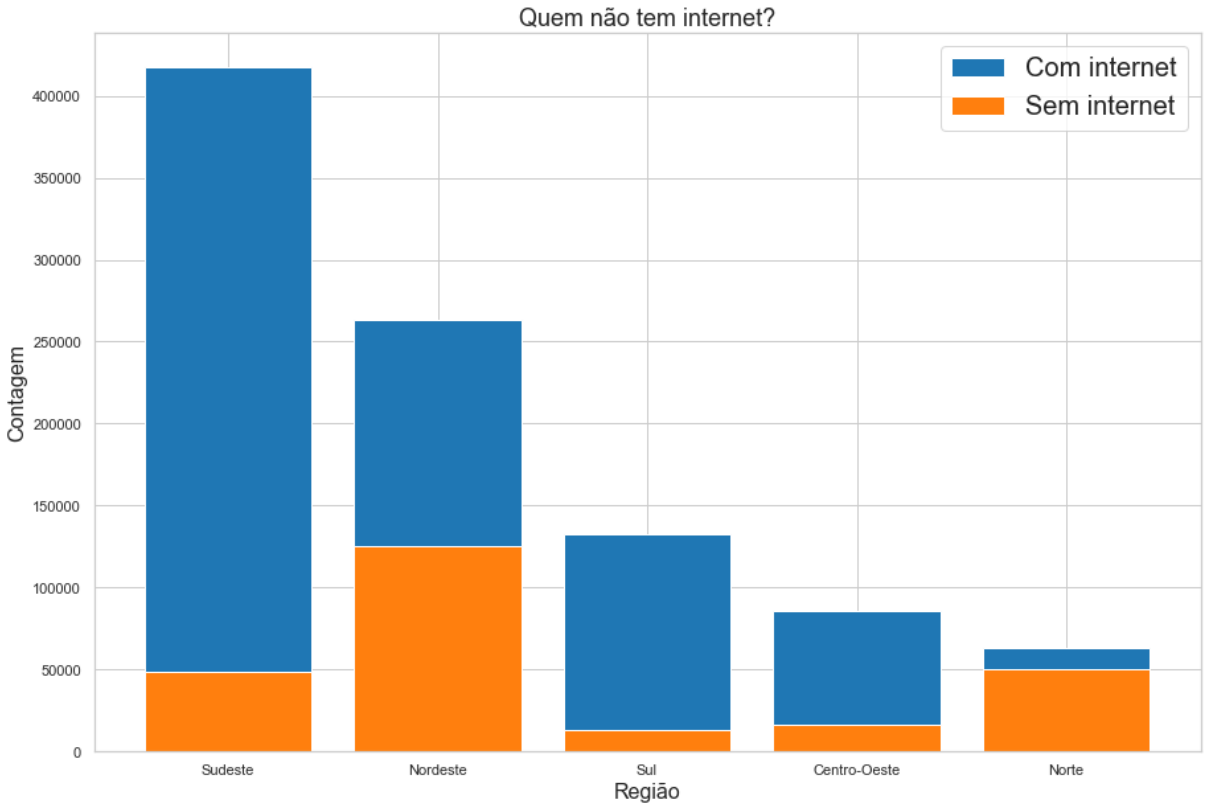
Figura 19 – Notas por acesso à internet



Fonte: Autor

Com a análise dos gráficos da figura, fica evidente que há uma enorme diferença entre os alunos que possuem internet (com média geral de 533) e os que não possuem (com média geral de 476). Afinal, o acesso à internet possibilita os alunos a contrair mais informação e cultura, bem como possibilita o acompanhamento de aulas e conteúdos gratuitos que ajudam na formação escolar. Mas, quem são os alunos que não tem acesso à internet?

Figura 20 – Acesso à internet por região?



Fonte: Autor

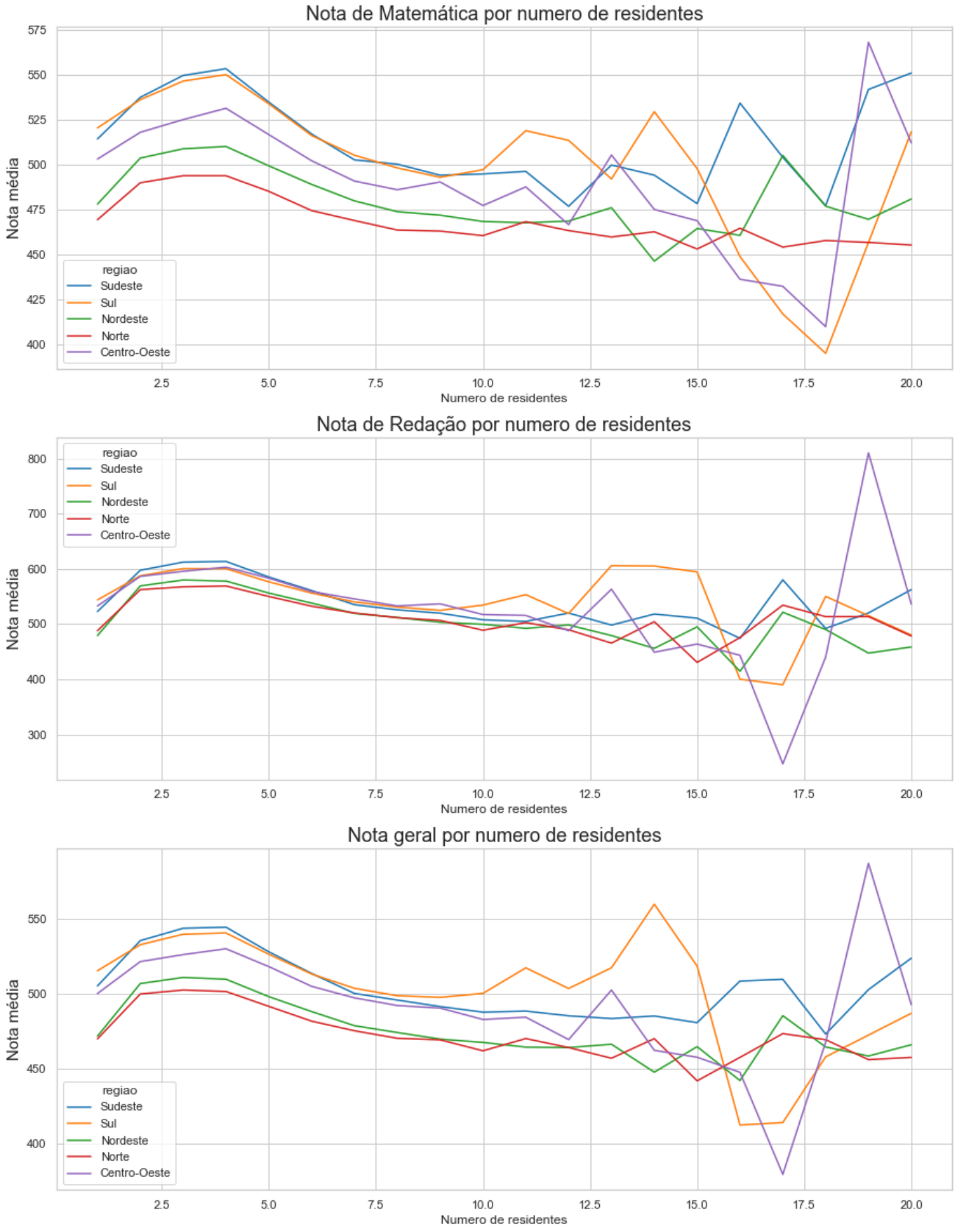
Analisando o gráfico e as informações geradas, vemos que proporcionalmente, a região norte está em estado crítico comparada a outras regiões. 8,8% dos alunos não tem acesso à internet na região sul, 10,4% na região sudeste, 16,17% na região centro-oeste, 32.33% na região nordeste e 44,4% na região norte, o que tem o potencial de gerar um enorme déficit educacional nos alunos dessa região.

Essa diferença deve ser levada em consideração por tomadores de decisão ao se criar e fomentar políticas públicas educacionais, tendo como um dos aspectos fundamentais o acesso à internet pelos alunos.

4.8 A quantidade de residentes na casa influencia na nota?

Com esta pergunta pretendemos verificar e evidenciar um reflexo da desigualdade de renda e entre regiões, que é o maior índice de fertilidade em regiões mais pobres, o que pode levar ao trabalho infantil e outros reflexos. Chamando a atenção de órgãos público para a criação de políticas voltadas a esta temática, como a criação de creches, escolas, dentro outras.

Figura 21 – Notas por quantidade de residentes



Fonte: Autor

O esperado era que quanto mais residentes, menor a nota nas provas, o que foi verificado até em torno de 12 residentes. Todavia, a partir de 12 residentes em uma mesma casa, os dados não se comportam da maneira esperada e nós vemos um aumento nas médias das notas em algumas regiões e acentuados altos e baixos entre as regiões. Destaque para a região Centro-Oeste, que atinge um pico positivo na média das notas quando o número de residentes chega a 19. Esta visualização pode representar possíveis oportunidades de estudo e análises mais aprofundadas pelos órgãos competentes e interessados: A partir de 12 residentes são respostas de alunos que vivem em repúblicas? Há muitas cidades universitárias na região Centro-Oeste que comportam estudantes de ensino médio? Estas e outras perguntas podem revelas padrões interessantes para os tomadores de decisão.

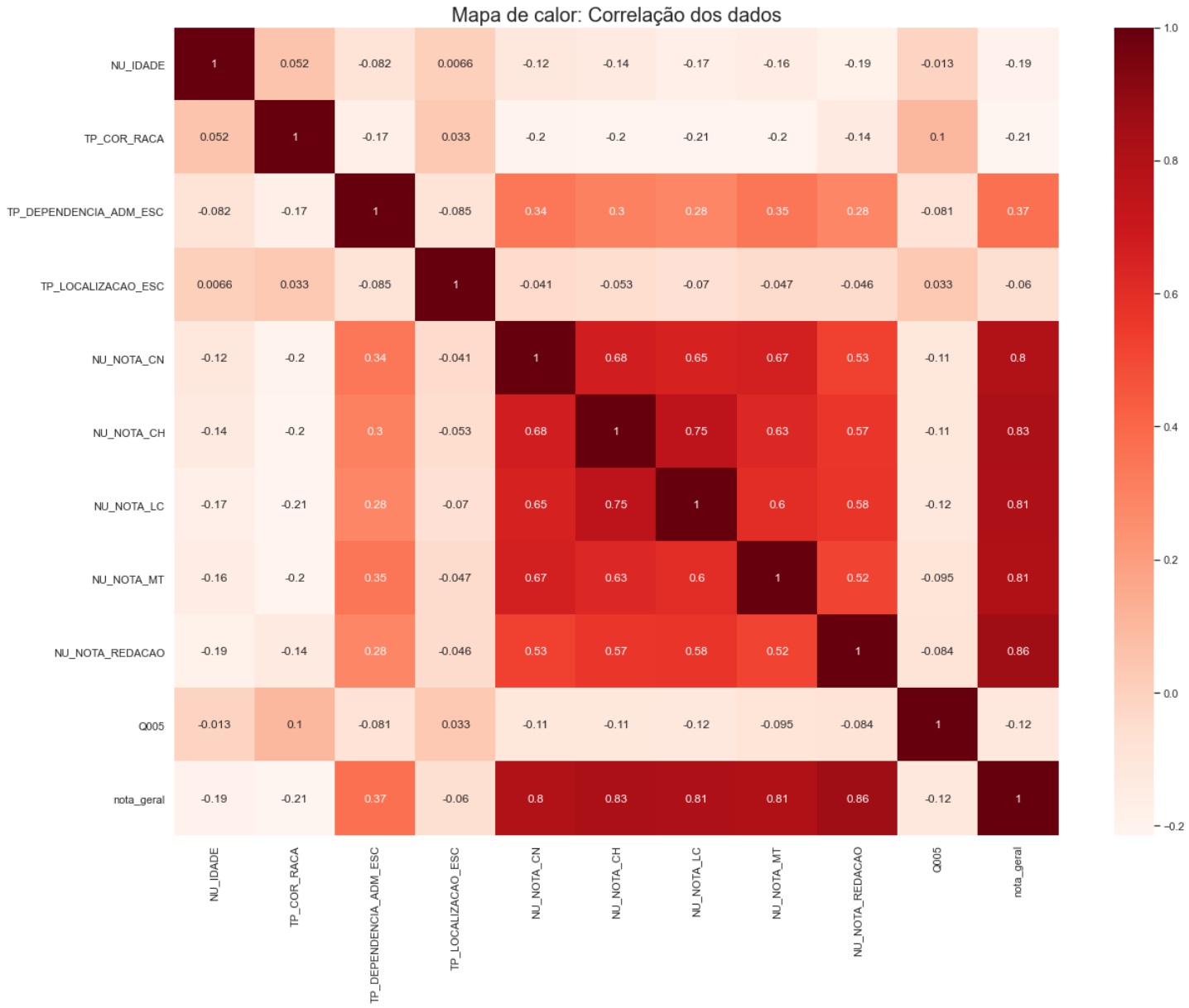
5. ANÁLISES GERAIS

Vamos utilizar outros recursos para compreender melhor as nuances e características dessa base de dados do Enem 2019.

5.1 Correlação

A correlação é uma análise descritiva que mede se há e qual é o grau de dependência de duas variáveis, ela serve para confirmarmos alguma suspeita de algum evento dentro do nosso banco de dados. Especificamente, queremos ver quais variáveis tem mais correlação com a nota geral do Enem, bem como que tipo de relação cada uma das variáveis guarda entre si.

Figura 22 – Matriz de correlação



Fonte: Autor

Note que quando se trata das notas nas áreas de conhecimento individuais (CH, CN, LC, Matemática e Redação) a correlação entre elas é considerada forte, sendo a maior de todas a correlação entre a nota de Linguagens e Códigos com a nota de Ciências Humanas, com uma correlação de 0.75. Isso significa que caso o aluno consiga uma nota alta em uma dessas áreas, ele tende a conseguir nota alta na outra também. Isso pode ser explicado pelo fato de que são provas com conteúdos e habilidades próximas (assuntos culturais, artísticos, históricos, habilidades como interpretação de texto, leitura e escrita).

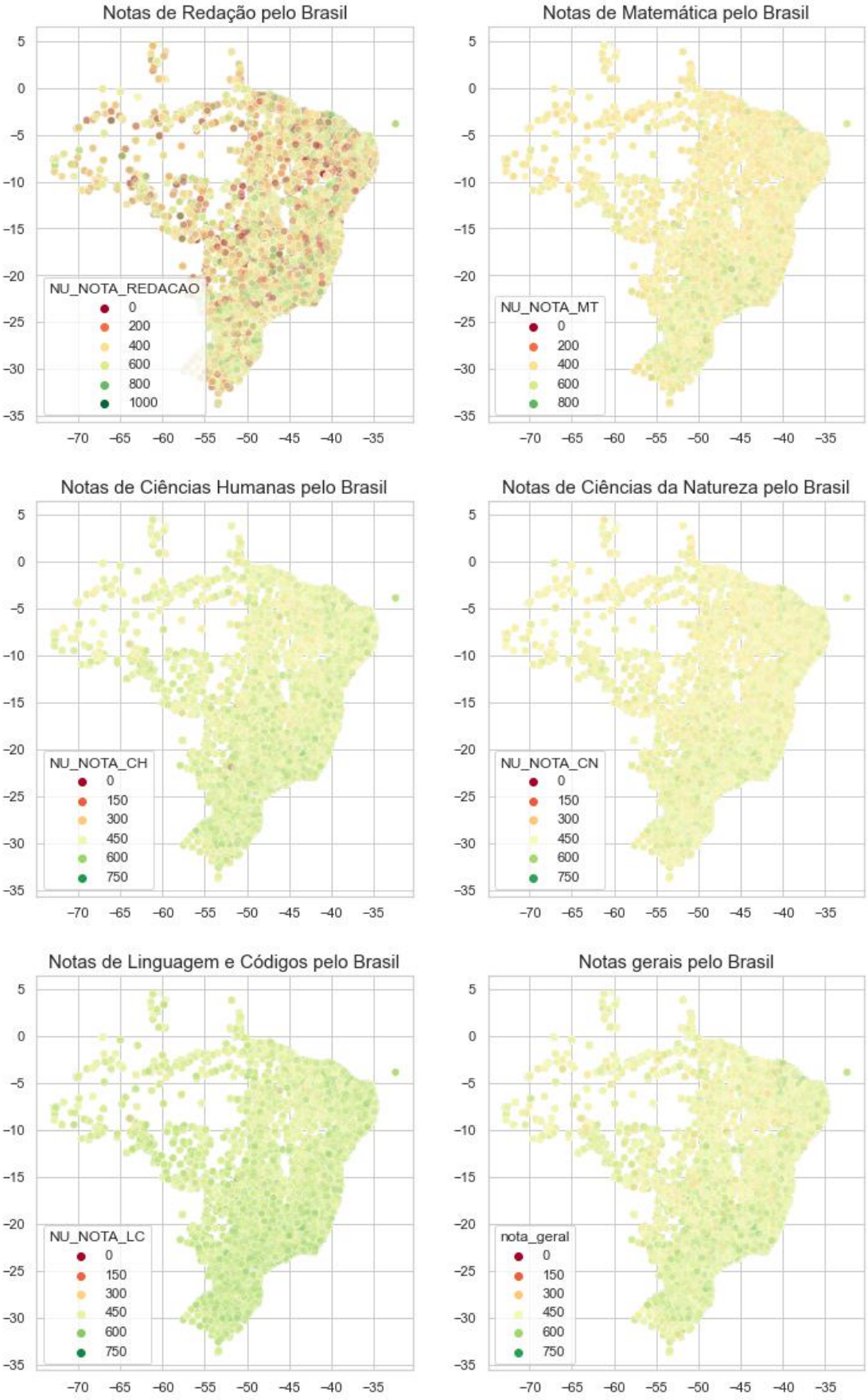
Note que, dentre as provas, a que possui maior correlação com a nota geral do candidato é a de redação, com impressionantes 0.86 de correlação.

Nos quesitos socioeconômicos, vemos que a dependência administrativa (TP_DEPENDENCIA_ADM_ESC) da escola (municipal, estadual, federal ou privada) tem uma correlação moderada de 0.37, indicando novamente que dependendo de onde o candidato cursou o ensino médio, ele pode ter mais chances de conseguir uma nota maior. Já o fato de a escola ser urbana ou rural não tem uma correlação nem moderada, contrariando o senso popular.

5.2 Visualização espacial

Para analisar as informações de maneira mais visual, vamos utilizar outra base de dados que relaciona o município dos candidatos com sua localização geográfica. Assim, esperamos conseguir aumentar nosso entendimento de quais cidades do Brasil tem mais notas baixas e altas.

Figura 23 – Localização espacial das notas



Fonte: Autor

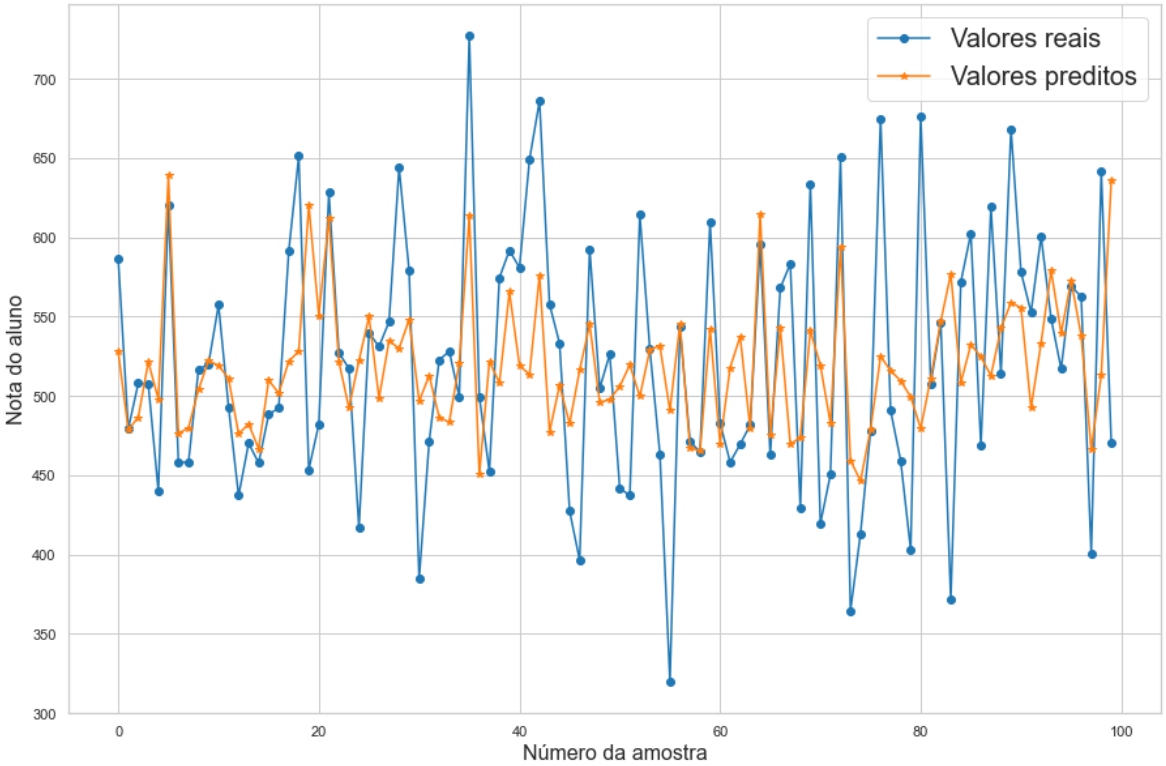
Analisando os gráficos, percebemos que ciências humanas e linguagens apresentam muito menos pontos vermelhos e amarelos do que matemática e ciências da natureza. De maneira geral, a região sudeste e sul concentram mais pontos verdes que as demais regiões.

É interessante evidenciar que na região norte (amazônica) há bastante espaços em branco, ou seja, sem cidades, já que é floresta densa, o que de certa forma dialoga com o fato de 44% dos alunos dessa região não ter internet pela falta de infraestrutura. Note que nas notas de redação, há muito mais pontos vermelho escuro do que nas outras áreas de conhecimento, o que já havia sido evidenciado no item 3.2 com o alto desvio padrão desta prova.

6. MACHINE LEARNING

Foi desenvolvido um modelo de aprendizado de máquina utilizando-se bibliotecas do Sklearn para executar as etapas de processamento e aprendizagem. Foi desenvolvido um modelo utilizando a técnica do ElasticNet, que é uma combinação das regressões Lasso e Ridge. Uma parte do resultado do modelo pode ser visto na figura 24 abaixo, onde está sendo apresentado de maneira visual o resultado das predições das primeiras 100 amostras na base de dados de teste, juntamente com os valores reais (os quais deveriam ser previstos).

Figura 24 – Predição das primeiras 100 amostras – Modelo 1



Fonte: Autor

O modelo obteve r^2 de 0.341, o que é considerado baixo para prever a nota de um candidato.

O erro quadrático médio (MSE) pega a diferença entre o valor predito pelo modelo e o valor real, eleva o resultado ao quadrado e divide pelo número de elementos preditos. O MSE foi de 4887 e se mostrou alto, mesmo em uma tarefa para prever algo com não tanta precisão.

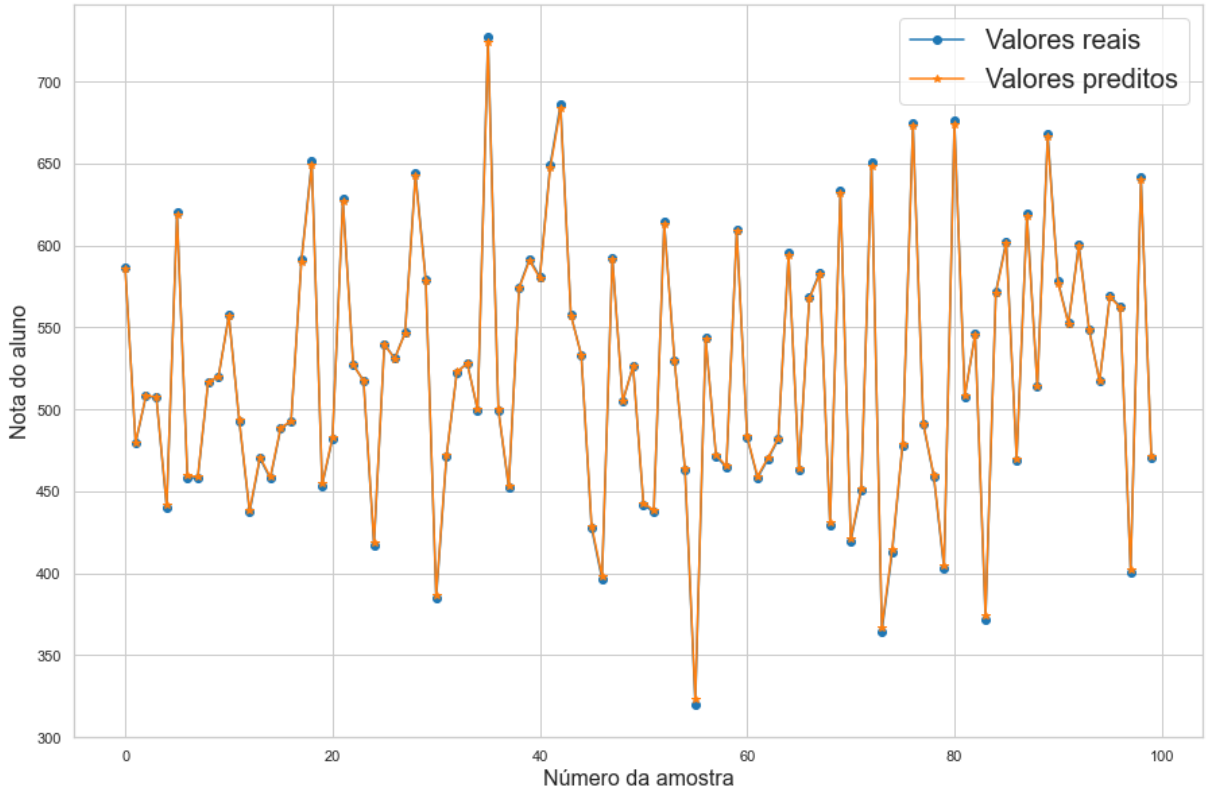
A raiz do erro médio quadrático (RMSE) executa o que o nome diz, faz a raiz do MSE e foi de 69. Isso se dá, pois a unidade de medida do MSE é de difícil interpretação (já que está elevada ao quadrado). No RMSE, a interpretação melhora, já que agora temos a mesma unidade de medida. Em outras palavras, com um RMSE de 69 aproximadamente, queremos dizer que a média do erro de previsão que o nosso modelo obteve foi de 69 pontos na prova do Enem, o que é considerável, principalmente para testes e início de planejamento de campanhas públicas.

Este modelo, juntamente com o aprendizado obtido lendo este relatório, pode ser utilizado por tomadores de decisão, políticos, professores e outros interessados para simular as possíveis notas de seus alunos e obter insights, como: Quais cursos possuem a nota de corte adequada, que regiões possivelmente obterão notas menores, quais alunos priorizar e suportar mais, que tipo de investimento é necessário e em qual parte da cidade, etc.

Note que, este modelo possui métricas não tão boas, todavia optou-se por manter ele nesta análise para mostrar como é complicada a tarefa de prever a nota de um candidato somente com base em dados socioeconômicos dos candidatos, as que saberemos antes dos alunos fazerem a prova e assim replicamos uma situação real em que um tomador de decisão tem exatamente essas mesmas informações. Mesmo assim, como evidenciado na figura 24, a estimativa se mostra satisfatória ao menos para iniciar os processos de planejamentos de políticas públicas. Para que o modelo tenha melhor desempenho, se faz necessário mais pesquisas e análises sobre quais variáveis e tipos de dados podem ser incluídos.

Caso queiramos utilizar as notas nas áreas específicas para estimar a nota final do candidato, podemos as incluir em nosso modelo e obteremos o resultado da figura 25, um modelo com r^2 de 0.99, MSE de 1.47 e RMSE de 1.21, o que é considerado extremamente bom, e pode ainda ser utilizado por estudantes que queiram ter um vislumbre de sua nota no Enem para saber qual faculdade, curso e nota de corte podem optar.

Figura 25 – Predição das primeiras 100 amostras



Fonte: Autor

7. CONCLUSÃO

Podemos concluir que os candidatos participantes do Enem 2019 possuem idades, *backgrounds*, raça/cor, escolaridade e classe sociais muito diferentes, reforçando mais uma vez o discurso oficial do Enem como instrumento de democratização e porta de entrada ao ensino superior, contribuindo com mais diversidade, um fator importante para o sucesso de sistemas educacionais e econômicos (SILVEIRA; BARBOSA; SILVA, 2015).

Os candidatos são compostos por sua maioria por mulheres e possuem média geral de 521,38 pontos, valor que varia bastante quando analisamos os candidatos com base em sua escolaridade, renda, estado de origem, tipo de escola e outros aspectos socioeconômicos.

Algumas tendências, como o aumento das notas na medida do aumento da renda, outras descobertas, como a diferença de escolaridade ser fator determinante para a nota, dão material de análise e pesquisa para que tomadores de decisões da esfera pública e privada pensem e executem ações para os públicos certos. Nesse cenário, o modelo de regressão se mostra como uma boa estimativa e pode guiar previsões e estudos iniciais sobre características que fazem um aluno conseguir boas notas.

De maneira geral, percebe-se mais necessidade de investimentos em educação nas regiões mais desfavorecidas (Norte e Nordeste), principalmente para os alunos provenientes de escolas públicas (municipais e estaduais) e que venham de famílias com baixa escolaridade. Faz-se necessário mais pesquisas para entender como e o porquê de mesmo com a mesma escolaridade, as notas médias sejam menores nas regiões desfavorecidas, bem como é necessário entender como a maior renda da família consegue superar essa dificuldade, mesmo nas regiões desfavorecidas.

Portanto, sabendo-se que este assunto é de grande relevância para o setor educacional brasileiro, visto que ainda existem muitas barreiras a serem vencidas, esta análise se mostra como um ponto de partida para trabalhos e pesquisas mais aprofundadas e específicas no futuro.

REFERÊNCIAS BIBLIOGRÁFICAS

CRIANÇA LIVRE DE TRABALHO INFANTIL. [*S. l.*], [*s. d.*]. Disponível em: https://livredetrabalhoinfantil.org.br/. Acesso em: 1 dez. 2021.

FEIJÓ, Janaína Rodrigues; FRANÇA, João Mário Santos de. Diferencial de desempenho entre jovens das escolas públicas e privadas. **Estudos Econômicos (São Paulo)**, [*s. l.*], v. 51, p. 373–408, 2021.

GUIMARÃES, André Rodrigues; BRITO, Cristiane de Sousa; SANTOS, José Almir Brito dos. EXPANSÃO E FINANCIAMENTO DA PÓS-GRADUAÇÃO E DESIGUALDADE REGIONAL NO BRASIL (2002-2018). **Práxis Educacional**, [*s. l.*], v. 16, n. 41, p. 47–71, 2020.

LIMA, Priscila da Silva Neves *et al.* Análise de dados do Enade e Enem: uma revisão sistemática da literatura. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, [*s. l.*], v. 24, p. 89–107, 2019.

SILVEIRA, Fernando Lang da; BARBOSA, Marcia Cristina Bernardes; SILVA, Roberto da. Exame Nacional do Ensino Médio (ENEM): Uma análise crítica. **Revista Brasileira de Ensino de Física**, [*s. l.*], v. 37, 2015. Disponível em: http://www.scielo.br/j/rbef/a/TpSdTxpHR3XBgFttPmgmyPF/?lang=pt. Acesso em: 29 nov. 2021.

TEOREMA CENTRAL DO LIMITE – WIKIPÉDIA, A ENCICLOPÉDIA LIVRE. [*S. l.*], [*s. d.*]. Disponível em: https://pt.wikipedia.org/wiki/Teorema_central_do_limite. Acesso em: 5 dez. 2021.