UNIVERSIDADE DE SOROCABA
GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

Gabriel Mascarenhas de Souza,
José Augusto Soares de Campos,
Pedro Hentique Vasconcelos Zerbini

Análises e extração de dados ENEM 2022

Sorocaba
2024

Gabriel Mascarenhas de Souza,
José Augusto Soares de Campos,
Pedro Hentique Vasconcelos Zerbini

Análises e extração de dados ENEM 2022

Monografia apresentada ao Professor Examinador da matéria de Engenharia de Dados e Análise Exploratória de Dados 2 do curso de Ciência de Dados e Inteligência Artificial da Universidade de Sorocaba para obtenção da média parcial do 1° semestre de 2024

Orientador: Prof. Dr. César Cândido Xavier

Sorocaba

2024

Agradecimentos

A conclusão desta monografia representa um marco relevante em nossas jornadas acadêmicas e profissionais. Este trabalho não teria sido viável sem o suporte e a colaboração de várias pessoas e instituições, às quais manifestamos nossa sincera gratidão.

Primeiramente, expressamos nossa gratidão ao nosso orientador, Professor Dr. César Cândido Xavier, cuja orientação precisa e incentivo contínuo foram fundamentais para o desenvolvimento desta pesquisa. Sua expertise e dedicação proporcionaram uma base sólida para que eu pudesse explorar e aprofundar meus conhecimentos em visão computacional e detecção de objetos.

Agradecemos também à Universidade de Sorocaba por fornecer os recursos e o ambiente adequado para a realização deste trabalho. Aos colegas e amigos do laboratório de Ciência de Dados e Inteligência Artificial, nosso obrigado pelas discussões enriquecedoras, pelo apoio técnico e pela camaradagem durante todo o processo.

Não podemos deixar de mencionar nossas famílias, que sempre acreditaram em nós e nos apoiaram incondicionalmente. Aos nossos pais, pelo amor e suporte emocional, e aos nossos amigos, pela compreensão nos momentos de ausência e pelas palavras de encorajamento.

Por fim, agradecemos a todos que, direta ou indiretamente, contribuíram para a concretização desta monografia. A todos vocês, nossa profunda gratidão.

Esperamos que este trabalho sirva de base para futuras pesquisas e que as contribuições aqui apresentadas possam ajudar no avanço do conhecimento na área de Análise de Dados.

Com gratidão,

Autores.

“In God we trust. All others must bring data”

Edwards Deming

Resumo

O entendimento dos aspectos da educação de um país é de fundamental importância para estabelecer metas de melhorias no ensino. O trabalho a seguir utiliza técnicas estatísticas e da ciência de dados com a finalidade de identificar características relevantes dos participantes do Exame Nacional do Ensino Médio do ano de 2022 e possíveis relações dessas características com o desempenho dos mesmos. Tem por objetivo verificar se tais características refletem no desempenho dos participantes do exame. Com o auxílio das bibliotecas Pandas, Matplotlib, Seaborn, Numpy e sklearn da linguagem Python, foi possível encontrar alguns fatores que exercem influência no desempenho dos participantes do exame e categorizar algumas características do perfil dos mesmos.

Palavras-chaves: Ciência de Dados. ENEM. Python. Pandas. Matplotlib. Seaborn

Abstract

Understanding the aspects of education in a country is of fundamental importance for establishing goals for improvements in teaching. The following work uses statistical and data science techniques with the purpose of identifying relevant characteristics of the participants of the National High School Exam in 2022 and possible relationships between these characteristics and their performance. It aims to verify whether such characteristics reflect on the performance of exam participants. With the help of the Python language libraries Pandas, Matplotlib, Seaborn, Numpy and sklearn, it was possible to find some factors that influence the performance of exam participants and categorize some characteristics of their profile.

Keywords: Data Science. ENEM. Python. Pandas. Matplotlib. Seaborn

Lista de figuras

Lista de quadros

Sumário

1 Introdução

A pesquisa realizada a partir da coleta de dados dos participantes do Exame Nacional do Ensino Médio (Enem) de 2022 apresenta uma análise detalhada das características e tendências observadas nesse importante exame educacional brasileiro. Utilizando técnicas avançadas de análise e extração de dados, o estudo visa compreender melhor o perfil dos candidatos, suas performances, além de identificar padrões que possam contribuir para a formulação de políticas educacionais mais eficazes. Através da análise de um vasto conjunto de dados, incluindo informações demográficas, socioeconômicas e de desempenho acadêmico, esta pesquisa oferece insights valiosos que podem orientar ações tanto de instituições educacionais quanto de formuladores de políticas públicas. O uso de metodologias de data mining e machine learning permite uma exploração aprofundada das variáveis envolvidas, revelando correlações e tendências que não seriam facilmente perceptíveis através de análises tradicionais. Assim, este estudo não apenas enriquece o entendimento sobre o Enem, mas também fornece uma base sólida para melhorias no sistema educacional brasileiro. (EDUCAçãO, 2019).

1.1 Motivação

A motivação para conduzir uma pesquisa baseada na coleta de dados dos participantes do Enem 2022, utilizando técnicas de análise e extração de dados, é multifacetada e profundamente relevante para o contexto educacional brasileiro. O Enem é um dos principais instrumentos de avaliação do desempenho estudantil no Brasil, influenciando diretamente o acesso ao ensino superior e proporcionando uma medida abrangente da qualidade da educação básica no país. No entanto, para que ele possa realmente cumprir seu papel de maneira eficaz, é essencial entender detalhadamente os dados gerados por este exame. (SANTOS, 2013)

1.2 Estrutura do trabalho

Esta monografia além da Introdução já apresentada no Capítulo 1, possui os seguintes capítulos, a saber:

Capítulo 2 - Aspectos conceituais: apresenta os principais conceitos relacionados ao exame nacional do ensino médio como objeto de estudo, ciência de dados como suporte para realização do estudo pretendido e noções de estatística para compreensão dos resultados obtidos.

Capítulo 3 - Materiais e métodos: neste capítulo é apresentada a metodologia adotada em todos os passos da pesquisa, bem como os instrumentos e procedimentos utilizados.

Capítulo 4 - Resultados experimentais: apresenta a implementação do modelo teórico consolidando o método proposto nesta pesquisa, resultando na Plataforma BullCounter.

Capítulo 5 - Conclusões: apresenta as conclusões do trabalho e as contribuições da monografia.

Por fim, têm-se as Referências que subsidiam o embasamento teórico desta pesquisa.

2 Conceitos fundamentais

2.1 ENEM (Exame nacional do ensino médio)

O Exame Nacional do Ensino Médio (Enem) foi instituído em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. Em 2009, o exame aperfeiçoou sua metodologia e passou a ser utilizado como mecanismo de acesso à educação superior.

As notas do Enem podem ser usadas para acesso ao Sistema de Seleção Unificada (Sisu) e ao Programa Universidade para Todos (ProUni). Elas também são aceitas em instituições de educação superior portuguesas que têm acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Além disso, os participantes do Enem podem pleitear financiamento estudantil em programas do governo, como o Fundo de Financiamento Estudantil (Fies). Os resultados do Enem possibilitam, ainda, o desenvolvimento de estudos e indicadores educacionais.

Qualquer pessoa que já concluiu o ensino médio ou está concluindo a etapa pode fazer o Enem para acesso à educação superior. Os participantes que ainda não concluíram o ensino médio podem participar como “treineiros” e seus resultados no exame servem somente para autoavaliação de conhecimentos.

A aplicação do Enem ocorre em dois dias. A Política de Acessibilidade e Inclusão do Inep garante atendimento especializado e tratamento pelo nome social, além de diversos recursos de acessibilidade. Há também uma aplicação para pessoas privadas de liberdade.

Os participantes fazem provas de quatro áreas de conhecimento: linguagens, códigos e suas tecnologias; ciências humanas e suas tecnologias; ciências da natureza e suas tecnologias; e matemática e suas tecnologias, que ao todo somam 180 questões objetivas. Os participantes também são avaliados por meio de uma redação, que exige o desenvolvimento de um texto dissertativo-argumentativo a partir de uma situação-problema. (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), 2021).

2.2 Ciência de Dados

Ciência de dados é uma ciência multidisciplinar que envolve técnicas computacionais, estatísticas e matemáticas, entre outras, com o objetivo de resolver problemas complexos, utilizando para isso grandes conjuntos de dados. (DIGITAL, 2021).

No campo de conhecimento da ciência de dados, estão métodos científicos, matemáticos, estatísticos e outras ferramentas que são usadas para analisar e manipular dados. Processos que almejam obter algum tipo de informação a respeito de uma base de dados, provavelmente se enquadram na ciência de dados.

Como as demais áreas da tecnologia, a ciência de dados tem um ciclo de vida que envolve seus projetos. O ciclo de vida da ciência de dados não segue um mesmo padrão para todos os projetos, cada trabalho tem suas necessidades específicas e requer adaptações do modelo. Devido a isso, é comum que em diferentes trabalhos sejam utilizadas diferentes representações desse ciclo de vida. Neste trabalho, foram utilizadas algumas atividades e etapas do ciclo de vida da ciência de dados descritas por (CETAX, 2020).

2.2.1 Etapas do processo:

2.2.2 Entendimento do problema

Pode ser considerada uma das mais importantes etapas do ciclo. A partir do entendimento do problema é que podemos definir os meios de pesquisa e desenvolvimento para alcançar o resultado desejado

2.2.3 Coleta de dados

É onde ocorre a extração dos dados após a definição do problema. Os dados podem vir de planilhas, arquivos de texto, sensores ou de alguma API independente.

2.2.4 Processamento e tratamento de dados

É feito após a coleta dos dados. Como os dados podem vir estruturados (tabelas de banco de dados) ou não-estruturados (sites externos, redes sociais, etc.), é preciso tratar esses dados

antes que sejam feitas as análises. É necessário averiguar entradas duplicadas, registros vazios, inconsistência de dados e etc.

2.2.5 Exploração de dados

É onde se inicia de fato as análises que foram pensadas na primeira etapa. Aqui são identificados padrões e relações interessantes entre seus dados e levantadas hipóteses a respeito deles. Aqui que é feito o estudo das ideias e hipóteses que se busca validar. Devido a isso, é de fundamental importância que se tenha uma boa habilidade analítica.

2.2.6 Análise dos dados

É nesta fase que modelos preditivos, estatísticos e técnicas de Machine Learning são aplicadas para validar hipóteses levantadas anteriormente. Esta etapa nem sempre está presente em todos os projetos, já que alguns estudos já tem seu objetivo concluído na etapa anterior. Não sendo necessária nenhuma análise profunda.

2.2.7 Resultados e decisões

É nesta etapa que temos a disseminação efetiva dos resultados, podendo assim concluir o estudo efetuado. Este trabalho trata do estudo e aplicação de técnicas de ciência de dados nos dados do Enem 2022, passando pelas etapas do ciclo da ciência de dados mencionadas anteriormente, com ênfase maior na exploração de dados.

2.2.8 Noções de estatística

A estatística é uma parte da Matemática Aplicada que pode ser entendida como um conjunto de métodos empregados no planejamento de experimentos, na obtenção, organização e resumo de dados coletados, bem como na análise e interpretação de tais dados, a fim de que conclusões possam ser tiradas. Em geral, as fases de coleta, organização e descrição dos dados ficam a cargo da estatística descritiva, enquanto que a análise e interpretação de tais dados são de competência da chamada estatística indutiva ou inferencial.(Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG), s.d.)

2.2.9 Estatística descritiva

A estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou está área da estatística(Universidade Estadual Paulista (UNESP), 2018).

Média - o valor médio dos dados.

Mediana - se os dados forem ordenados de forma crescente, esse valor seria o valor do meio se dividirmos o conjunto exatamente pela metade.

Moda - o valor com maior número de ocorrências em toda a amostra.

2.2.10 Estatística inferencial

A estatística inferencial, o segundo tipo de procedimentos em estatística, preocupa-se com o raciocínio necessário para, a partir dos dados, se obter conclusões gerais. O seu objectivo é obter uma afirmação acerca de uma população com base numa amostra. Estas inferências ou generalizações podem também ser de dois tipos: estimações ou decisões (testes de hipóteses)(LOPES, 2005).

2.2.11 Amostragem estatística

Todos os dados brutos que se tem disponível para estudo é denominado população. Nem sempre é possível utilizar toda a população para fazer a análise desejada. As estatísticas possibilitam que possamos realizar o estudo desejado tomando como base uma amostra da população total e, usando probabilidade, é possível ter um certo grau de certeza a respeito das características da população na totalidade(MAYER, 2016).

Amostragem de probabilidade: a amostragem de probabilidade é um método que seleciona membros aleatórios de uma população, definindo alguns critérios de seleção. Esses parâmetros de seleção permitem que cada membro tenha oportunidades iguais de fazer parte de várias amostras.

Amostragem sem probabilidade: O método de amostragem sem probabilidade depende da capacidade do pesquisador de selecionar membros aleatoriamente.

2.2.12 Gráfico de barras

Um gráfico de barras é uma representação visual de dados categóricos, onde cada categoria é representada por uma barra retangular. A altura ou o comprimento de cada barra é proporcional ao valor que representa. Esse tipo de gráfico é muito útil para comparar valores entre diferentes categorias de dados de maneira clara e fácil de entender(GUIMARÃES; GITIRANA; ROAZZI, 2001).

Estrutura de um Gráfico de Barras Eixos:

Eixo X (horizontal): Representa as categorias. Por exemplo, diferentes produtos, meses do ano, tipos de animais, etc.

Eixo Y (vertical): Representa os valores ou frequências associados a cada categoria. Esses valores podem ser números absolutos, porcentagens, frequências, etc.

Barras: Cada barra corresponde a uma categoria do eixo X. A altura (ou comprimento, se as barras forem horizontais) da barra é proporcional ao valor ou frequência da categoria correspondente no eixo Y.

Rótulos: Cada barra pode ser rotulada para indicar a categoria que representa. Valores numéricos podem ser exibidos no topo ou dentro de cada barra para fornecer detalhes adicionais.

2.2.13 Gráfico de dispersão

O gráfico de dispersão, também conhecido como gráfico de dispersão ou gráfico de dispersão de pontos, é uma representação gráfica usada para visualizar a relação entre duas variáveis quantitativas. Cada ponto no gráfico representa um par de valores correspondentes das duas variáveis. Esse tipo de gráfico é muito útil para identificar padrões, correlações e possíveis tendências nos dados(MARTINS, 2014).

Estrutura de um Gráfico de Dispersão:

Eixo X (horizontal): Representa os valores de uma das variáveis.

Eixo Y (vertical): Representa os valores da outra variável.

Pontos: Cada ponto no gráfico corresponde a um par de valores das duas variáveis. A posição do ponto no gráfico é determinada pelas coordenadas (x, y).

2.2.14 Correlação

O objetivo do estudo da correlação é determinar (mensurar) o grau de relacionamento entre duas variáveis. Caso os pontos das variáveis, representados num plano cartesiano (X, Y) ou gráfico de dispersão, apresentem uma dispersão ao longo de uma reta imaginária, dizemos que os dados apresentam uma correlação linear(LUIZ, s.d.)

2.2.15 Regressão linear

A regressão linear é uma técnica de análise de dados que prevê o valor de dados desconhecidos usando outro valor de dados relacionado e conhecido. Ele modela matematicamente a variável desconhecida ou dependente e a variável conhecida ou independente como uma equação linear. Por exemplo, suponha que você tenha dados sobre suas despesas e receitas do ano passado. As técnicas de regressão linear analisam esses dados e determinam que suas despesas são metade de sua renda. Eles então calculam uma despesa futura desconhecida reduzindo pela metade uma renda futura conhecida(Amazon Web Services (AWS), s.d.).

2.3 Considerações finais do capítulo

Neste capítulo, exploramos conceitos essenciais relacionados a análise e extração de dados dos participantes do ENEM de 2022 e discutimos maneiras de otimizar nosso modelo para refinar a filtragem e facilitar a tomada de decisões em cima dessas informações. A aplicação de filtros e métodos adequados é fundamental para obter alta performance no projeto.

3 Materiais e Metódos

3.1 Considerações iniciais

Para iniciar a análise e extração de dados dos participantes do ENEM em 2022, é essencial contar com uma infraestrutura tecnológica adequada, incluindo software de análise de dados, ferramentas de processamento de big data e recursos de armazenamento em nuvem. Além disso, é crucial ter acesso aos conjuntos de dados do ENEM, que podem ser obtidos por meio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Métodos estatísticos e técnicas de ciência de dados serão aplicados para explorar e interpretar os dados, enquanto abordagens de amostragem estatística podem ser utilizadas para garantir a representatividade dos resultados. É fundamental também considerar questões éticas relacionadas à privacidade dos participantes e à segurança dos dados durante todo o processo de análise(MACIEL, 2022).

3.2 Linguagem de programação

Neste trabalho foi utilizada a linguagem de programação Python. A escolha dessa linguagem de programação se deu por sua praticidade, visto que o Python dispõe de bibliotecas muito úteis para a realização do estudo proposto. As seguintes bibliotecas foram utilizadas para desenvolver o estudo:

- **Pandas -** é uma biblioteca Python utilizada para análise de dados. Com ela, foi feita toda a leitura, tratamento e processamento dos dados.
- **Matplotlib -** é uma biblioteca utilizada para criar gráficos diversos para tipos de dados variados. Grande parte dos gráficos apresentados neste trabalho foram feitos utilizando esta biblioteca.
- **Numpy -** ajuda a executar facilmente cálculos numéricos. É usada principalmente para realizar cálculos em Arrays Multidimensionais.
- **Streamlit -** é um framework desenvolvido em Python que torna possível a criação de aplicativos elegantes para modelos de machine learning (aprendizagem de máquina) ou mesmo visualização de dados para uma simples análise exploratória de um dataset (conjunto de dados).

Para o ambiente, nós usamos o VSCode, que nada mais é do que um editor de código-fonte desenvolvido pela Microsoft para Windows, Linux e macOS. Ele inclui suporte para depuração, controle de versionamento Git incorporado, realce de sintaxe, complementação inteligente de código, snippets e refatoração de código.

3.2.1 Aquisição dos dados

No sítio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), consta um repositório com os dados de todas as edições anteriores do Enem, que pode ser acessado através de: www.gov.br/inep/pt-br/acesso-a informacao/dadosabertos/microdados/enem, a planilha com os dados extraídos foi disponibilizada também pelo professor. Para este estudo, utilizamos os dados do exame de 2022.

3.2.2 Tratamento dos dados

O arquivo csv que contém os dados do Enem 2022 tem cerca de 2,4 GB de tamanho. Ao descarregar a base de dados pelo sítio do INEP, o arquivo vem compactado, portanto, foi necessário descompactar o arquivo e fazer a filtragem das informações gerando um novo arquivo menor para podermos coletar e tratar os dados com maior facilidade. O arquivo csv principal (MICRODADOS-ENEM-2022) contém os questionários respondidos pelos participantes, armazenando todas as informações disponibilizadas pelos participantes do Enem 2022 em um único arquivo. As informações desse arquivo principal foram carregadas em um DataFrame do Pandas. O Pandas DataFrame é uma estrutura de dados bidimensional com os dados alinhados de forma tabular em linhas e colunas, mutável em tamanho e potencialmente heterogênea, semelhantemente a uma pasta de trabalho do MS-EXCEL. A diferença essencial é que os nomes de colunas e os números de linha são conhecidos como índice de coluna e linha, no caso do DataFrame. As colunas possuem nomes (índice da coluna) e, as linhas, podem ter nomes referentes a colunas e as linhas podem ter nomes (índices textuais) ou podem, por padrão, ser numeradas (Índice numérico).

O arquivo csv principal contém diversas colunas que servem para descrever vários aspectos administrativos do exame, como dependência administrativa da escola, cor das provas utilizadas, necessidade de adaptações para acessibilidade, etc. Fazendo uso de análise descritiva, foram filtradas as colunas mais relevantes do DataFrame para o estudo do perfil dos participantes

e do desempenho sob o prisma socioeconômico e regional. A Tabela 1 mostra as colunas que foram consideradas para a realização deste estudo

Quadro 1 – Variáveis

Nome da variável	**Descrição**
TP-FAIXA-ETARIA	Faixa etária
TP-SEXO	Sexo
TP-COR-RACA	Cor e raça
TP-ST-CONCLUSAO	Situação de conclusão
TP-ANO-CONCLUIU	Ano de conclusão do ensino médio
TP-ESCOLA	Tipo da escola
TP-ENSINO	Tipo do ensino
TP-DEPENDENCIA-ADM-ESC	Tipo de dependência da escola
TP-LOCALIZACAO-ESC	Localização da escola
SG-UF-PROVA	Sigla da federação da prova
NU-NOTA-CN	Nota da prova Ciências da Natureza
NU-NOTA-CH	Nota da prova Ciências Humanas
NU-NOTA-LC	Nota da prova Linguagens e Códigos
NU-NOTA-MT	Nota da prova Matemática
NU-NOTA-REDACAO	Nota da redação
Q006	Renda mensal da família
Q025	Residência com internet
MEDIA-TOTAL	Medias das totais das notas

3.3 Considerações finais do capítulo

Para finalizar este capítulo, é importante ressaltar a relevância da infraestrutura tecnológica e das ferramentas utilizadas, como Python e suas bibliotecas, para a análise dos dados do ENEM 2022. A aquisição e o tratamento dos dados foram etapas cruciais, garantindo a integridade e a representatividade das informações. O uso de métodos estatísticos e técnicas de ciência de dados permitiu explorar e interpretar os dados de forma robusta. Considerações éticas sobre privacidade e segurança dos dados foram rigorosamente observadas, assegurando a confidencialidade dos participantes. O resultado é um estudo detalhado e confiável sobre o perfil dos participantes e seu desempenho no ENEM 2022.

4 Resultados e Discussões

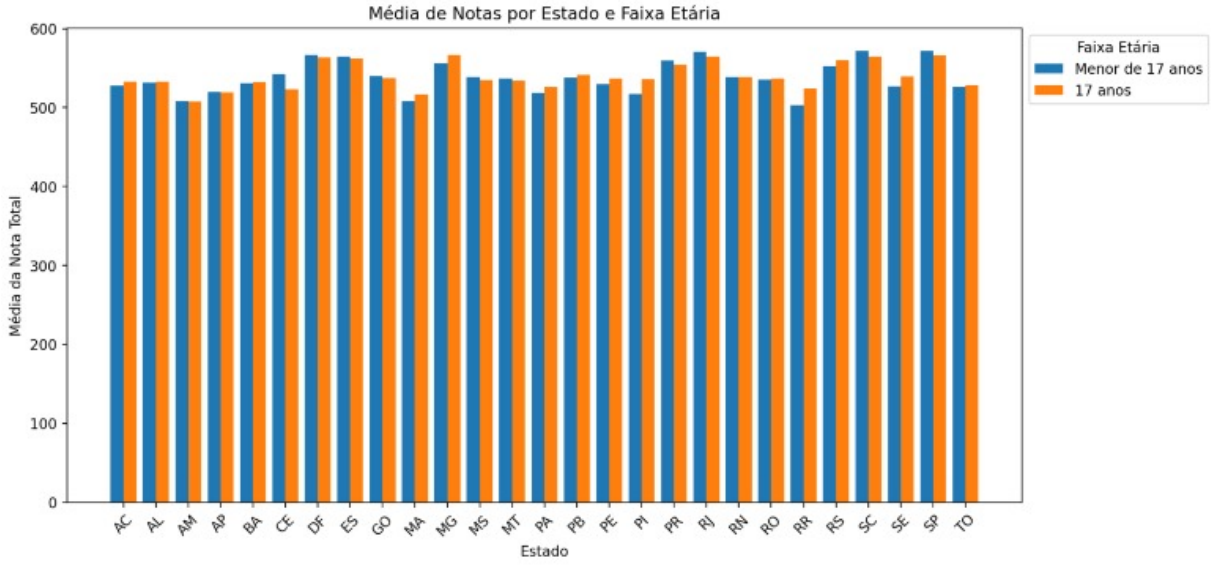
4.1 Considerações iniciais

Esta seção descreve os estudos realizados e apresenta uma análise descritiva dos dados obtidos. É onde são apresentadas todas as análises feitas com a base de dados retirada dos participantes do Enem 2022. **Lembrando que o número de opções selecionadas pode ser maior que a da foto.**

4.2 Idade

No exame é permitida a participação de candidatos de quase todas as idades. Na Base 1, temos candidatos a partir de 17 anos até participantes de 70 anos. Apesar da grande abrangência de idades, temos algumas ocorrências mais comuns, a mediana das idades dos participantes é de 19 anos.

Figura 1 – Média de Notas por Faixa Etária e Estado

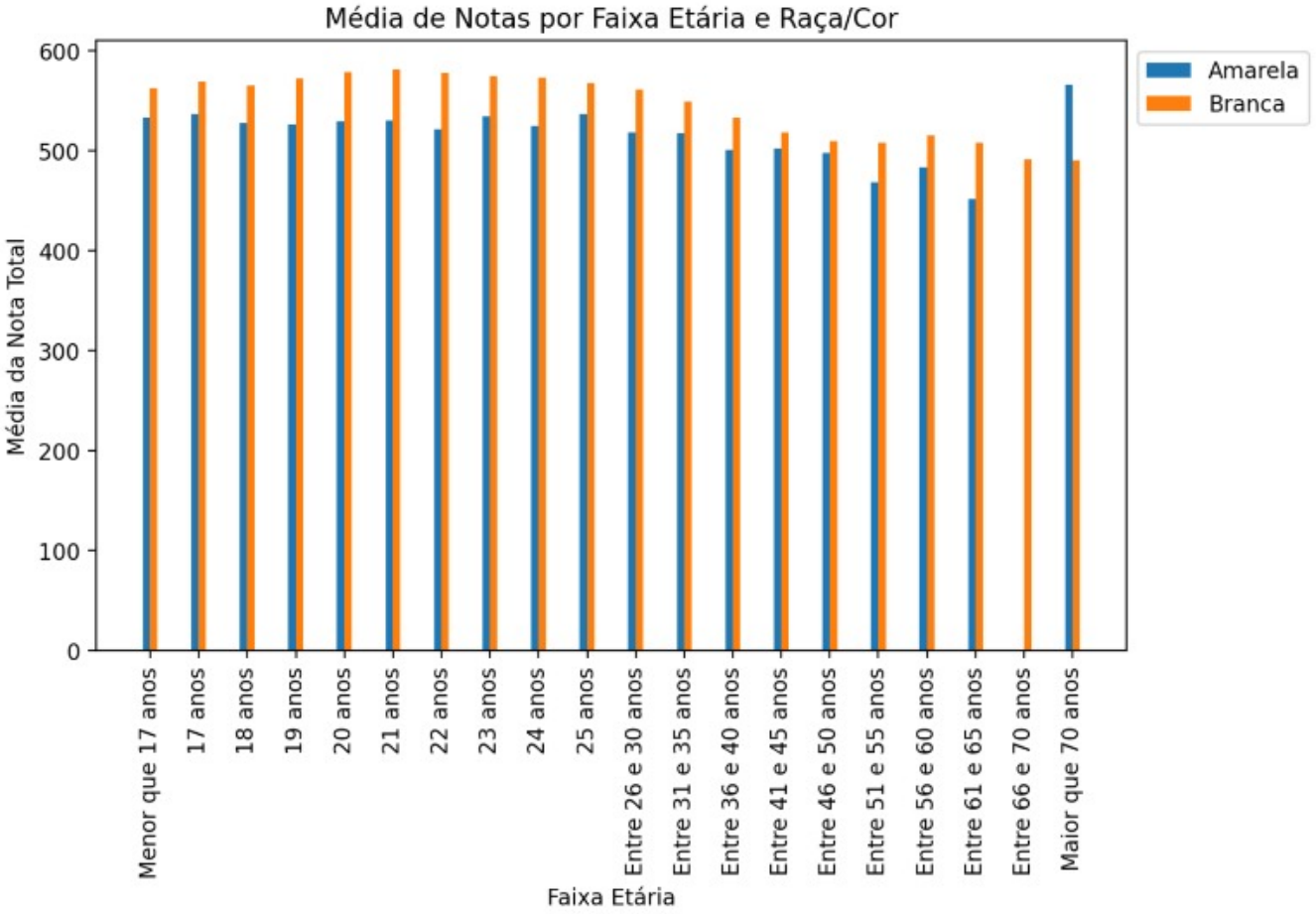


Fonte – Autores, 2024

O gráfico mostra as médias de notas por faixa etária nos estados do Acre (AC) e Amapá (AP). As médias são bastante próximas entre os estados em todas as faixas etárias, com os mais jovens obtendo notas mais altas. As médias tendem a diminuir com o aumento da idade, especialmente após os 26-30 anos. Pequenas diferenças são observadas nas idades mais avançadas,

onde AC geralmente apresenta notas ligeiramente superiores a AP. No geral, ambos os estados têm desempenhos semelhantes em todas as faixas etárias.

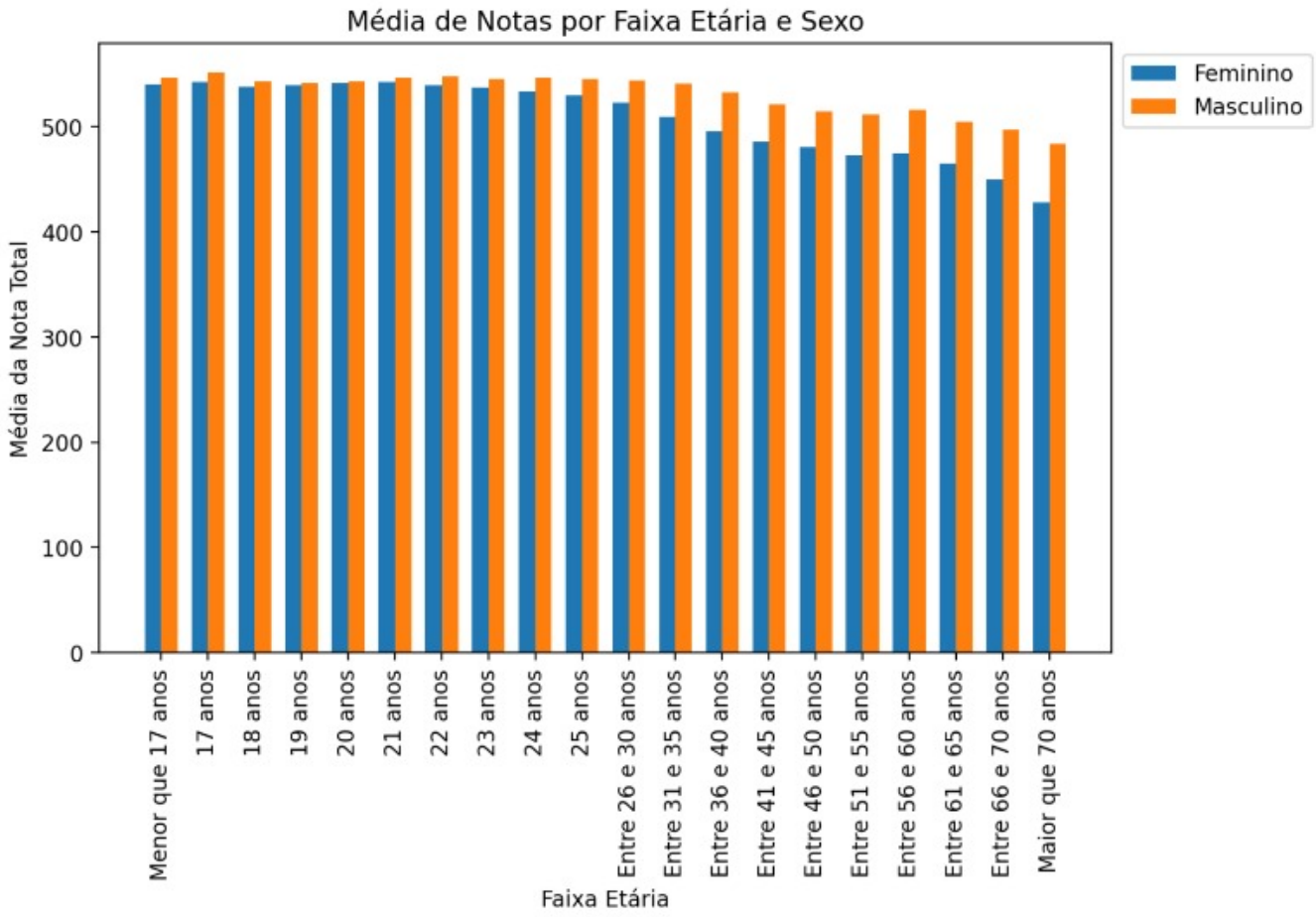
Figura 2 – Média de Notas por Faixa Etária e Raça/Cor



Fonte – Autores, 2024

Em geral, observa-se uma tendência decrescente da média de notas à medida que a idade aumenta, para ambas as raças/cores. Essa tendência pode ser explicada por diversos fatores, como o aumento da dificuldade das provas, a diminuição da motivação dos alunos e a maior pressão do mercado de trabalho. Embora a tendência geral seja decrescente para ambas as raças/cores, a imagem também revela algumas diferenças entre elas. Na faixa etária de até 25 anos, a média de notas dos alunos amarelos é superior à dos alunos brancos. No entanto, essa diferença se inverte a partir dos 26 anos, quando a média de notas dos alunos brancos passa a ser superior.

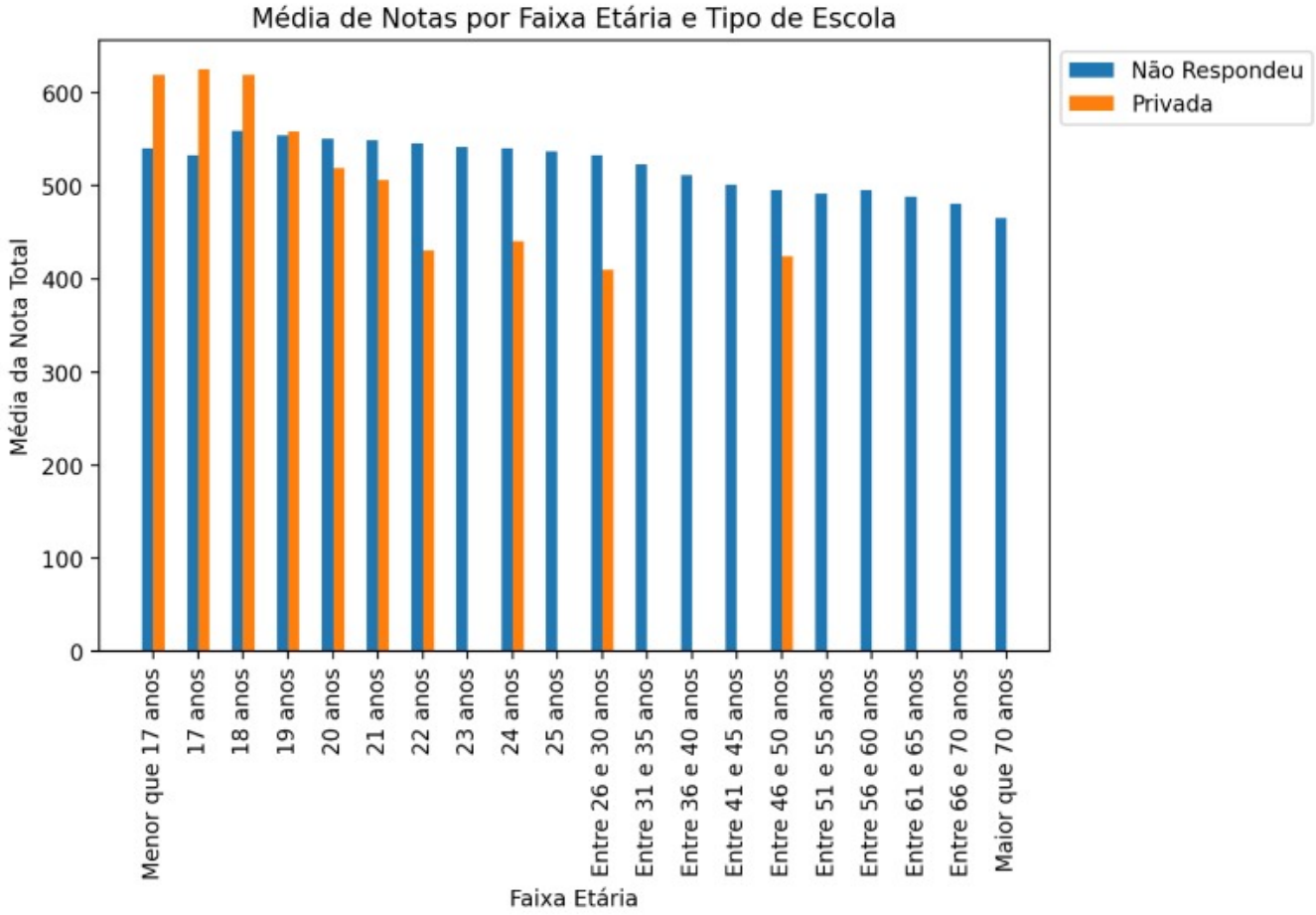
Figura 3 – Média de Notas por Faixa Etária e Sexo



Fonte – Autores, 2024

Em geral, observa-se uma tendência decrescente da média de notas à medida que a idade aumenta, para ambos os gêneros. A imagem revela uma diferença significativa na média de notas entre os gêneros. Em todas as faixas etárias, a média de notas dos alunos do sexo masculino é superior à das alunas do sexo feminino. Essa diferença pode ser explicada por diversos fatores socioeconômicos e culturais, como o maior incentivo que os meninos geralmente recebem para os estudos, as diferenças nas expectativas em relação ao desempenho dos alunos de cada gênero e as desigualdades de acesso à educação de qualidade.

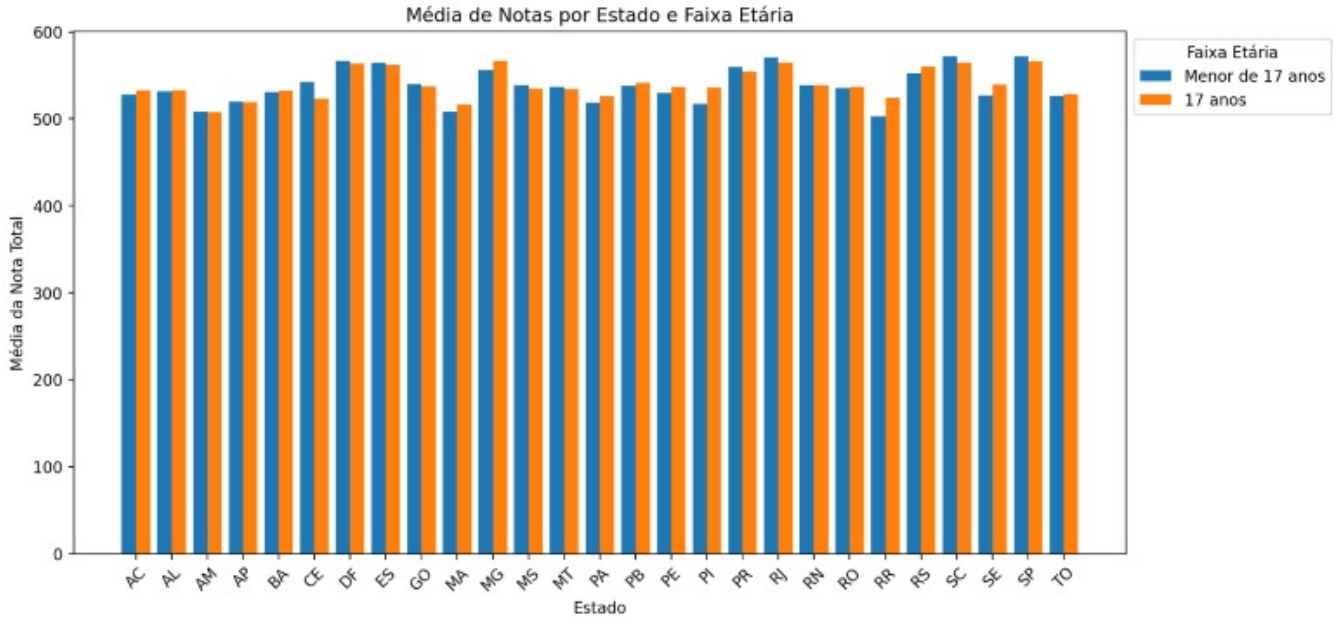
Figura 4 – Média de Notas por Faixa Etária e Tipo de Escola



Fonte – Autores, 2024

Em geral, observa-se uma tendência decrescente da média de notas à medida que a idade aumenta, para todos os tipos de escola. A imagem revela diferenças significativas na média de notas entre os tipos de escola. Em todas as faixas etárias, a média de notas dos alunos das escolas particulares é superior à dos alunos das escolas públicas e sem fins lucrativos. Essa diferença pode ser explicada por diversos fatores, como a melhor infraestrutura das escolas particulares, a maior qualificação dos professores e o menor número de alunos por turma.

Figura 5 – Média de Notas por Estado e Faixa Etária

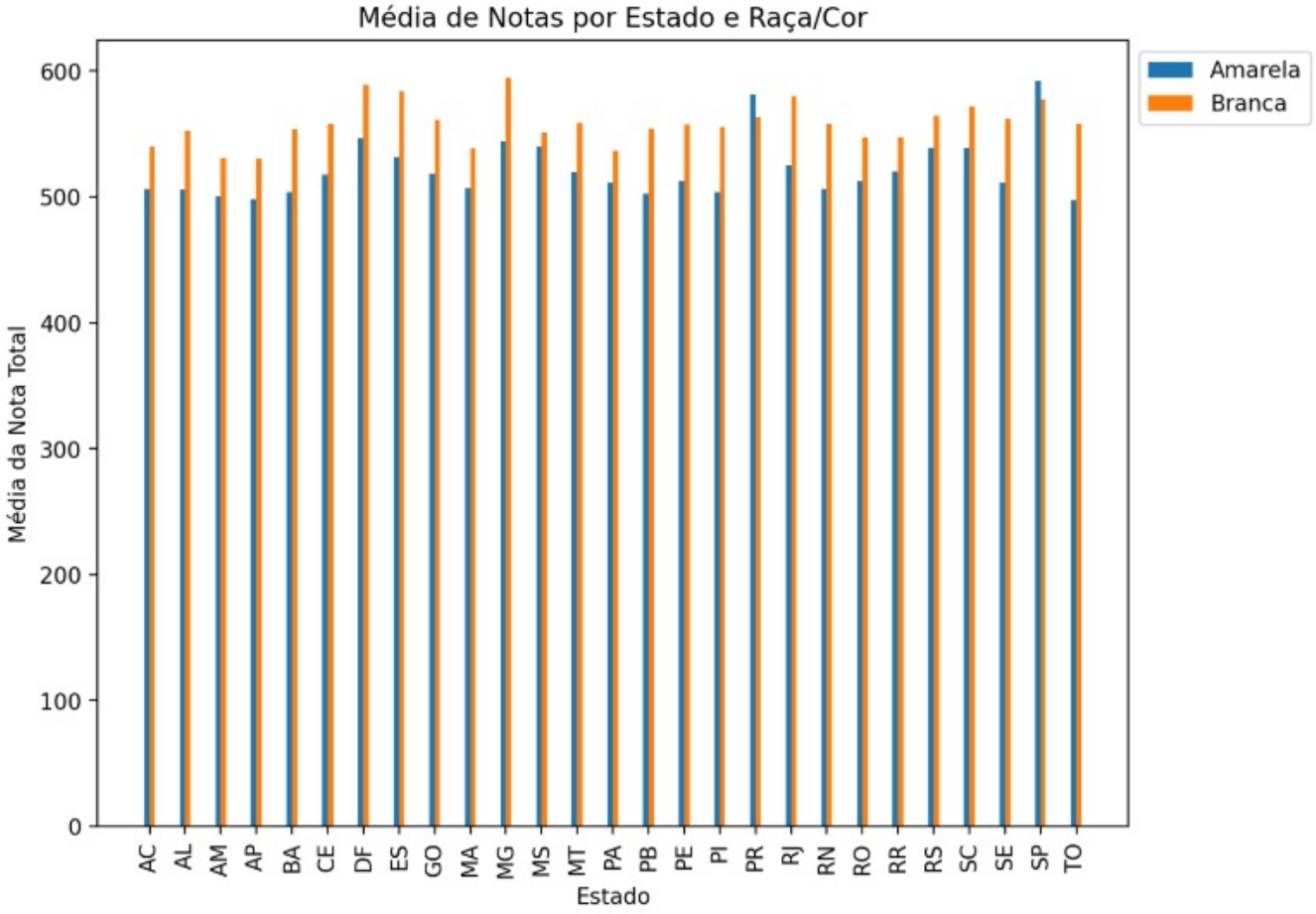


Fonte – Autores, 2024

Observa-se uma tendência decrescente da média de notas à medida que a idade aumenta, para todos os estados. A imagem revela diferenças significativas na média de notas entre os estados. Em algumas faixas etárias, alguns estados apresentam médias de notas bem superiores à dos outros estados. Essa diferença pode ser explicada por diversos fatores, como a qualidade da educação pública em cada estado, o nível socioeconômico da população e o investimento em educação por parte do governo.

4.3 Estado

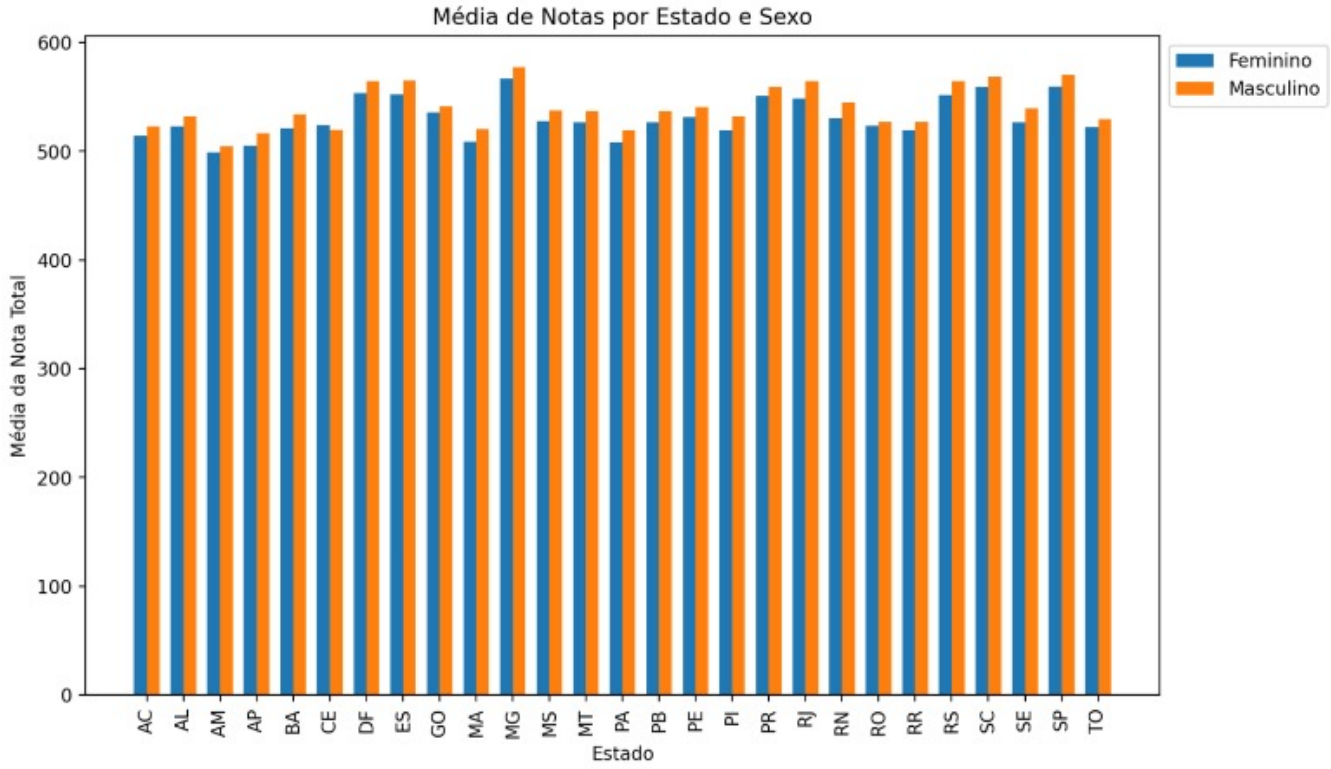
Figura 6 – Média de Notas por Estado e Raça/Cor



Fonte – Autores, 2024

Em geral, observa-se uma tendência decrescente da média de notas à medida que a idade aumenta, para todas as raças/cores. A imagem revela diferenças significativas na média de notas entre as raças/cores. Em alguns estados, algumas raças/cores apresentam médias de notas bem superiores à dos outras raças/cores. Essa diferença pode ser explicada por diversos fatores socioeconômicos e culturais, como o maior incentivo que algumas raças/cores geralmente recebem para os estudos, as diferenças nas expectativas em relação ao desempenho dos alunos de cada raça/cor e as desigualdades de acesso à educação de qualidade.

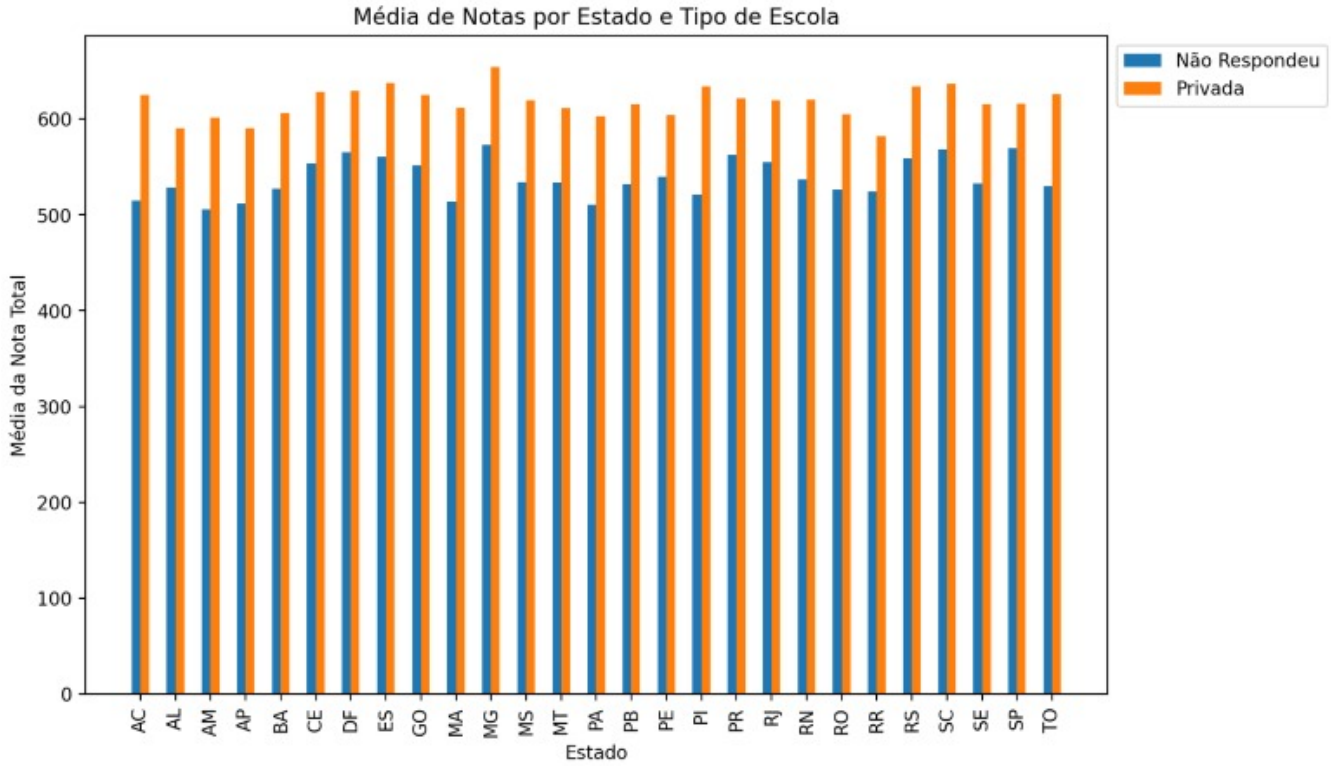
Figura 7 – Média de Notas por Estado e Sexo



Fonte – Autores, 2024

A imagem revela diferenças significativas na média de notas entre os cursos. Em alguns anos, alguns cursos apresentam médias de notas bem superiores à dos outros cursos.

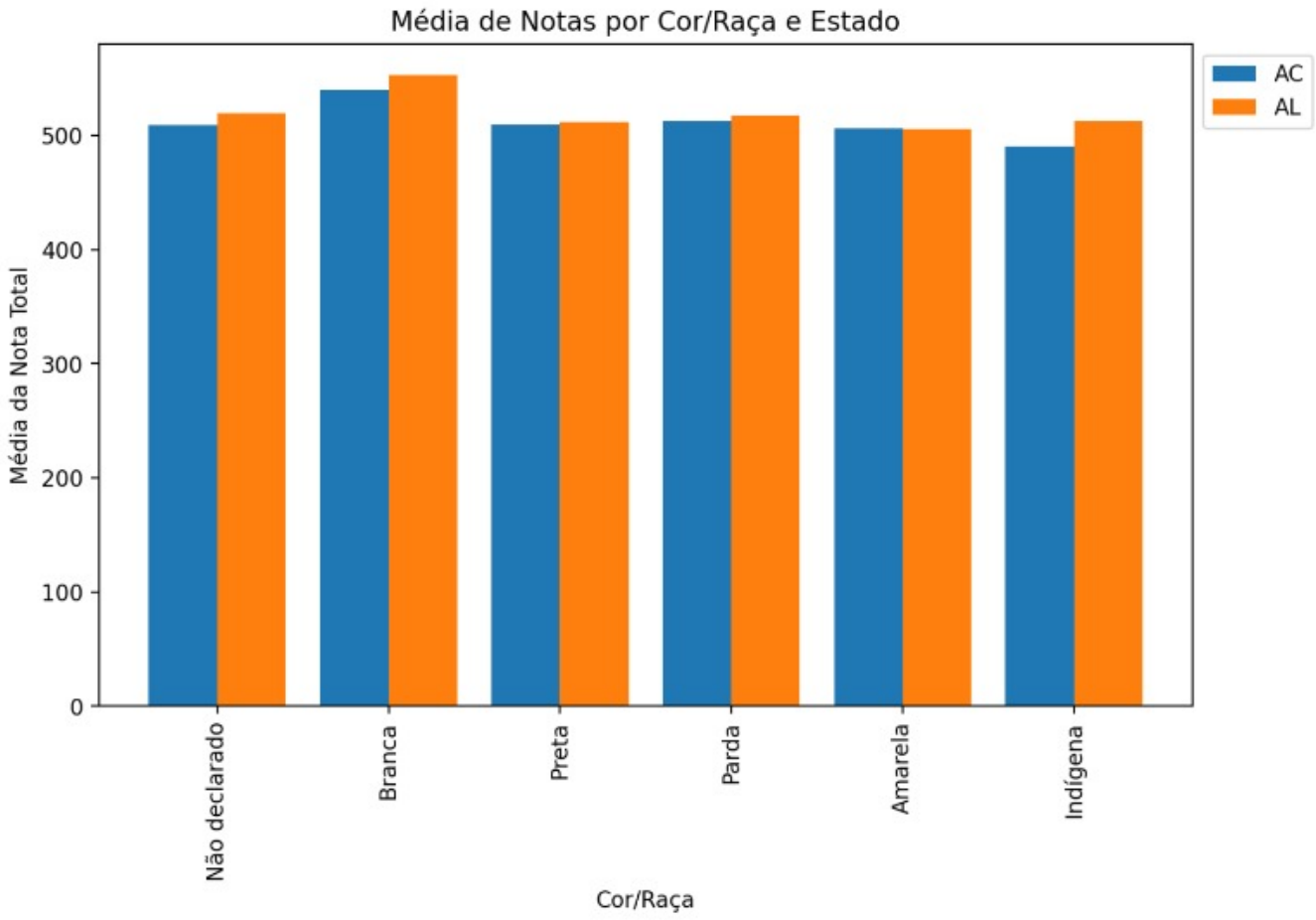
Figura 8 – Média de Notas por Estado e Tipo de Escola



Fonte – Autores, 2024

Observa-se uma tendência decrescente da média de notas à medida que o ano aumenta, para todos os tipos de escola. A imagem revela diferenças significativas na média de notas entre os tipos de escola. Em todos os anos, a média de notas dos alunos das escolas particulares é superior à dos alunos das escolas públicas e sem fins lucrativos. Essa diferença pode ser explicada por diversos fatores, como a melhor infraestrutura das escolas particulares, a maior qualificação dos professores e o menor número de alunos por turma.

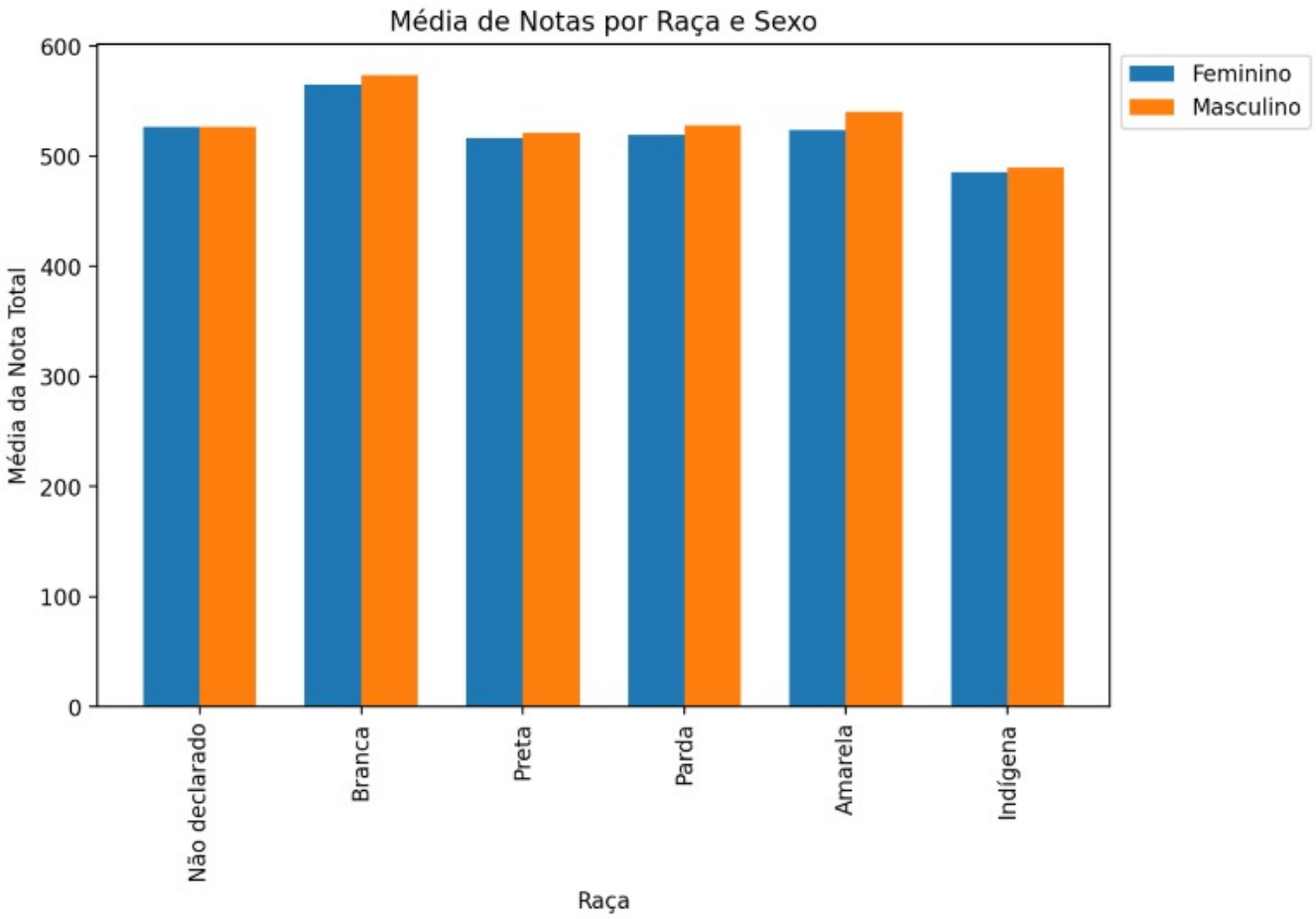
Figura 9 – Média de Notas por Cor/Raça e Estado



Fonte – Autores, 2024

A imagem apresenta uma tendência decrescente da média de notas à medida que o ano aumenta, para todos os cursos. A imagem revela diferenças significativas na média de notas entre os cursos. Em alguns anos, alguns cursos apresentam médias de notas bem superiores à dos outros cursos.

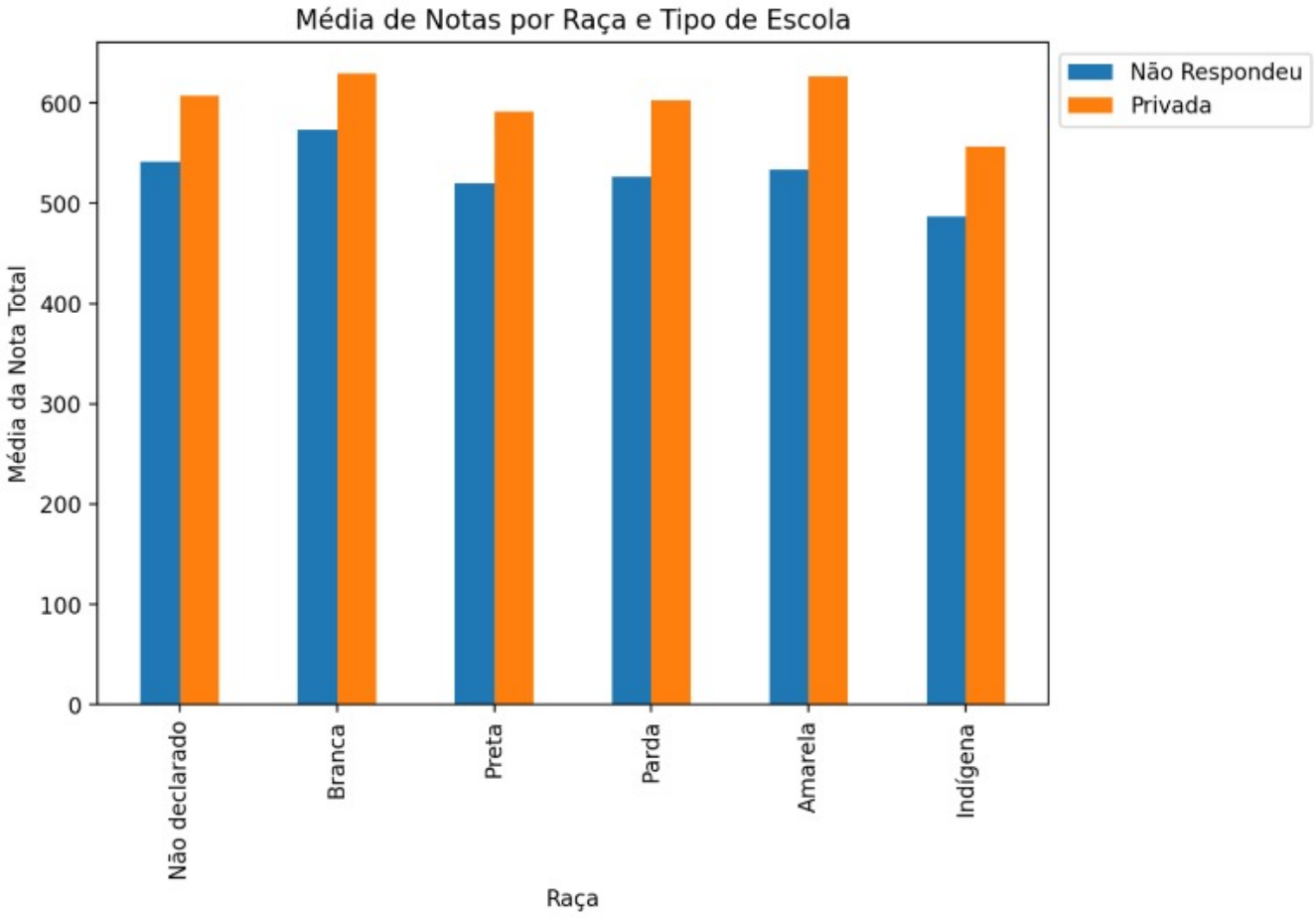
Figura 10 – Média de Notas por Raça e Sexo



Fonte – Autores, 2024

Observa-se uma tendência decrescente da média de notas à medida que o ano aumenta, para todos os tipos de escola e sexos. A imagem revela diferenças significativas na média de notas entre os tipos de escola e sexos. Em todos os anos, a média de notas dos alunos das escolas particulares é superior à dos alunos das escolas públicas e sem fins lucrativos. Além disso, em todos os anos e tipos de escola, a média de notas dos alunos do sexo masculino é superior à dos alunos do sexo feminino.

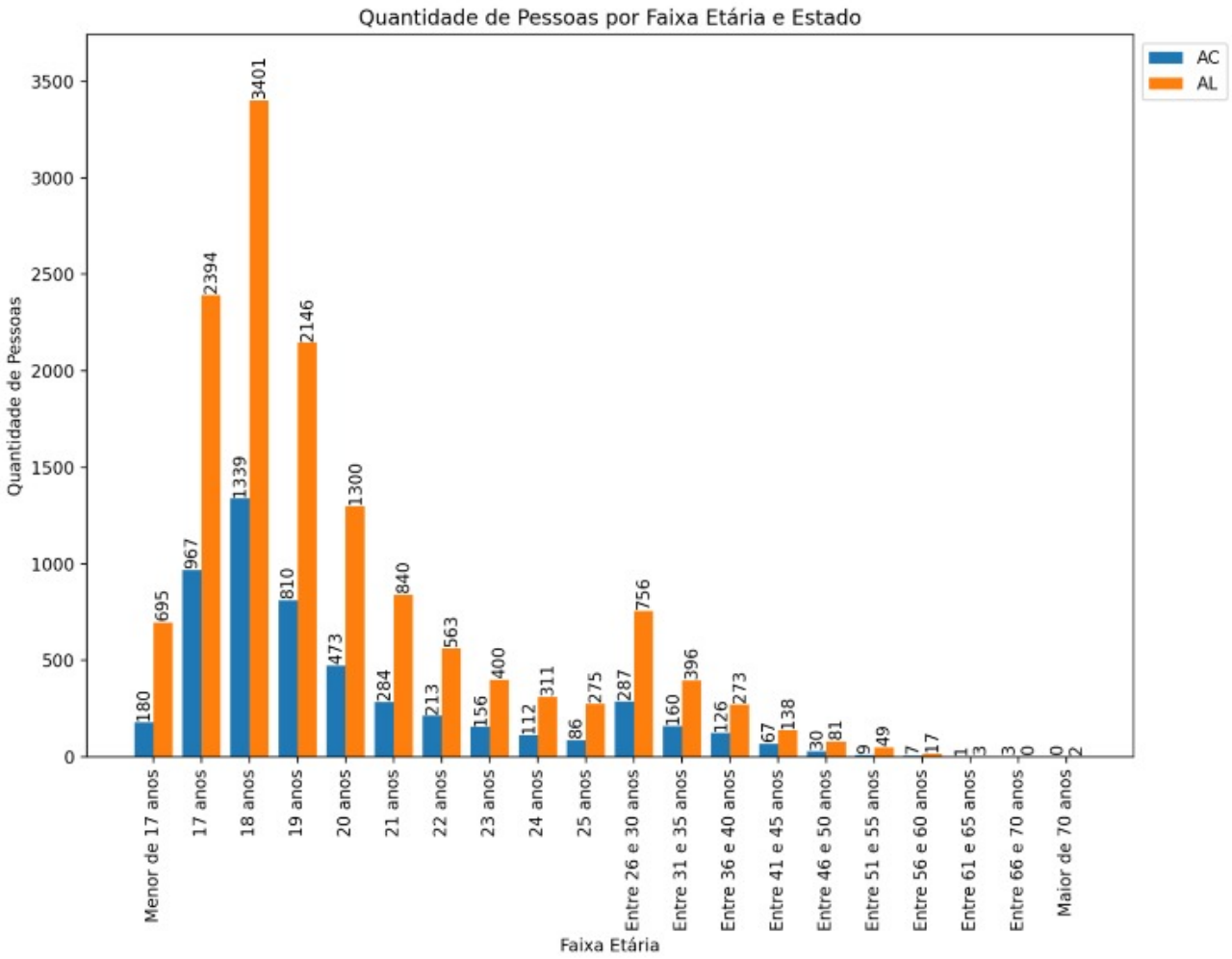
Figura 11 – Média de Notas por Raça e Tipo de Escola



Fonte – Autores, 2024

A imagem apresenta uma tendência decrescente da média de notas à medida que o ano aumenta, para todos os tipos de escola e raças/cores. A imagem revela diferenças significativas na média de notas entre os tipos de escola e raças/cores. Em todos os anos, a média de notas dos alunos das escolas particulares é superior à dos alunos das escolas públicas e sem fins lucrativos. Além disso, em todos os anos e tipos de escola, a média de notas dos alunos brancos é superior à dos alunos de outras raças/cores. Essa diferença pode ser explicada por diversos fatores, como a melhor infraestrutura das escolas particulares, a maior qualificação dos professores e o menor número de alunos por turma, no caso das diferenças por tipo de escola; e fatores socioeconômicos, culturais e históricos, no caso das diferenças por raça/cor.

Figura 12 – Quantidade de Pessoas por Faixa Etária e Estado



Fonte – Autores, 2024

O gráfico divide as pessoas em 24 faixas etárias, desde menor de 17 anos até maior de 70 anos. Para cada faixa etária, as barras do gráfico mostram a quantidade de pessoas em cada estado. É possível observar que, em geral, o estado de AL possui um número maior de pessoas em todas as faixas etárias, exceto nas faixas entre 46 e 50 anos e entre 61 e 65 anos.

As faixas etárias com maior número de pessoas em ambos os estados são:

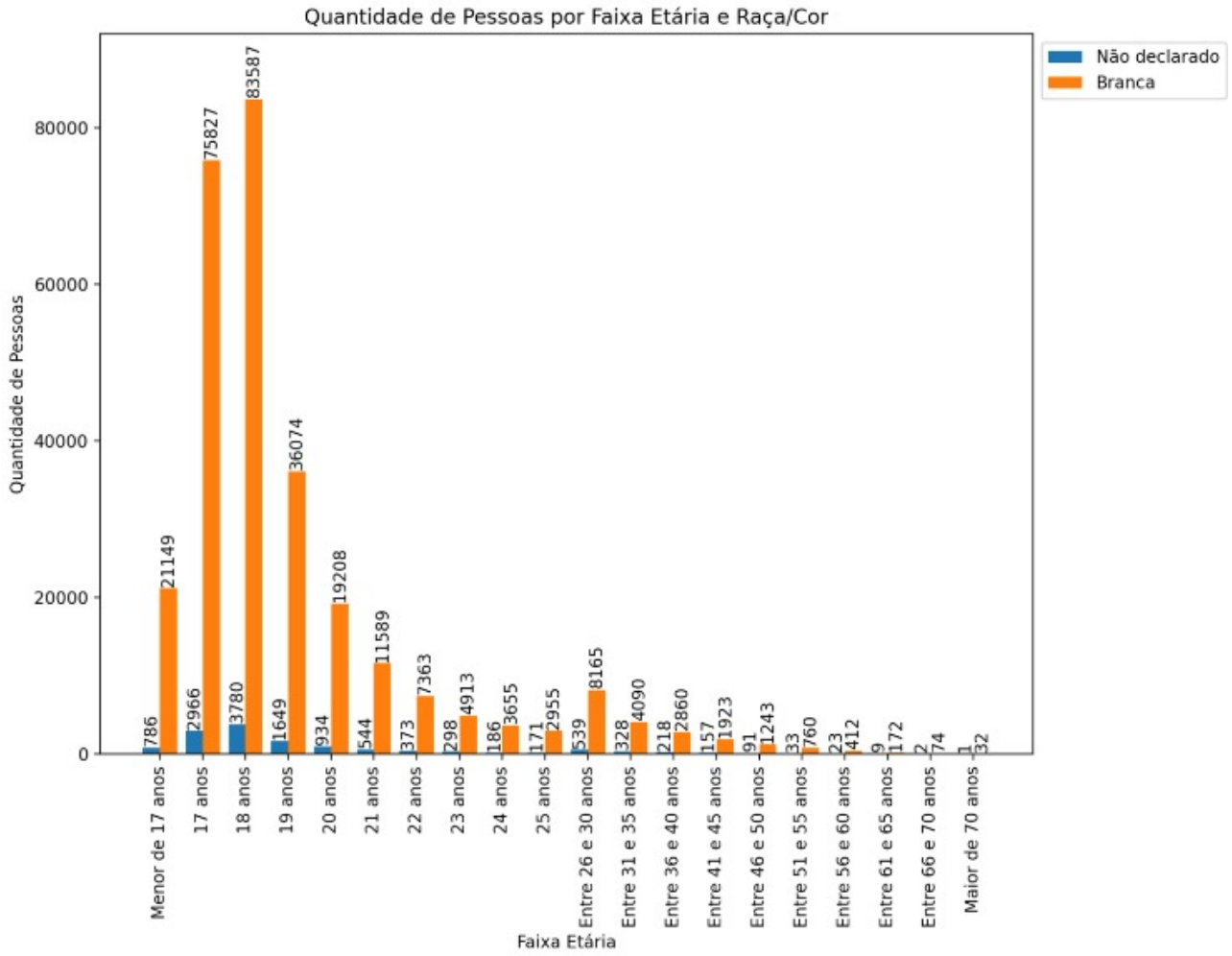
18 anos: AC: 1339, AL: 3401 17 anos: AC: 967, AL: 2394 19 anos: AC: 810, AL: 2146

As faixas etárias com menor número de pessoas em ambos os estados são:

Maior de 70 anos: AC: 0, AL: 0 Entre 66 e 70 anos: AC: 30, AL: 30 Entre 56 e 60 anos: AC: 17, AL: 17

4.4 Quantidade

Figura 13 – Quantidade de Pessoas por Faixa Etária e Raça/Cor



Fonte – Autores, 2024

A imagem apresenta a distribuição da população brasileira por faixa etária e raça/cor, com base em dados do Censo Demográfico 2022 realizado pelo IBGE.

A população branca é majoritária em todas as faixas etárias, exceto entre 0 e 14 anos, onde a população parda é ligeiramente superior.

A população negra apresenta um crescimento gradual em faixas etárias mais jovens, enquanto a população indígena se concentra nas faixas etárias mais novas.

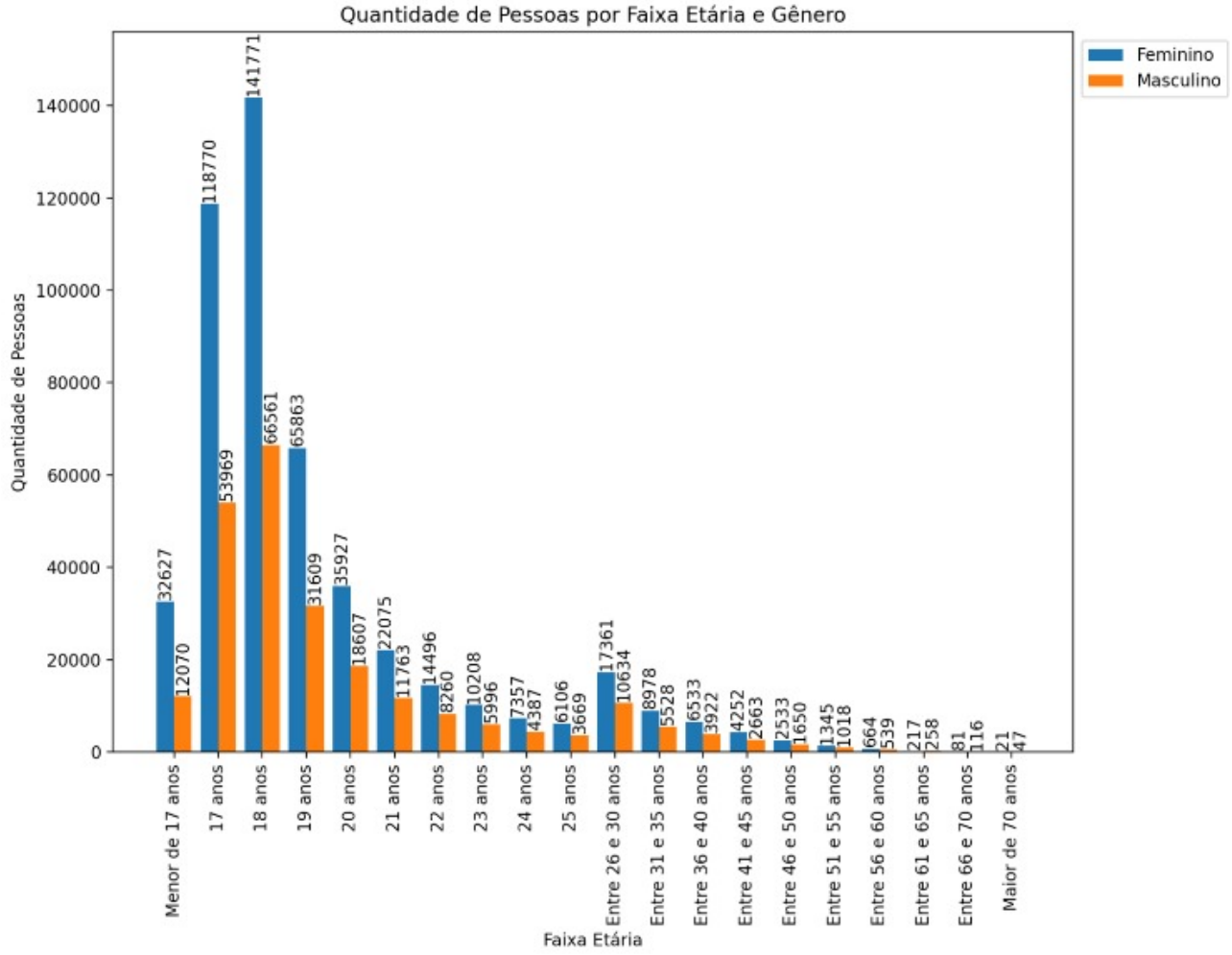
A população branca apresenta maior concentração em faixas etárias mais velhas, com 33,8 por cento na faixa acima de 75 anos. Isso indica um processo de envelhecimento da população branca no Brasil.

A população parda apresenta o menor índice de envelhecimento, com apenas 60,6 por cento. Isso indica que a população parda é a mais jovem do Brasil, com 49,3 por cento na faixa etária de 0 a 14 anos.

A população indígena apresenta um crescimento gradual em faixas etárias mais jovens, com 1,0 por cento na faixa de 0 a 14 anos e 0,3por cento na faixa de 60 a 64 anos. Isso indica um aumento da taxa de fecundidade entre as mulheres indígenas.

A população amarela apresenta o maior índice de envelhecimento, com 256,5. Isso indica que a população amarela é a mais envelhecida do Brasil, com 1,1por cento na faixa de 75 anos ou mais e 0,2 pore cento na faixa de 0 a 14 anos.

Figura 14 – Quantidade de Pessoas por Faixa Etária e Gênero

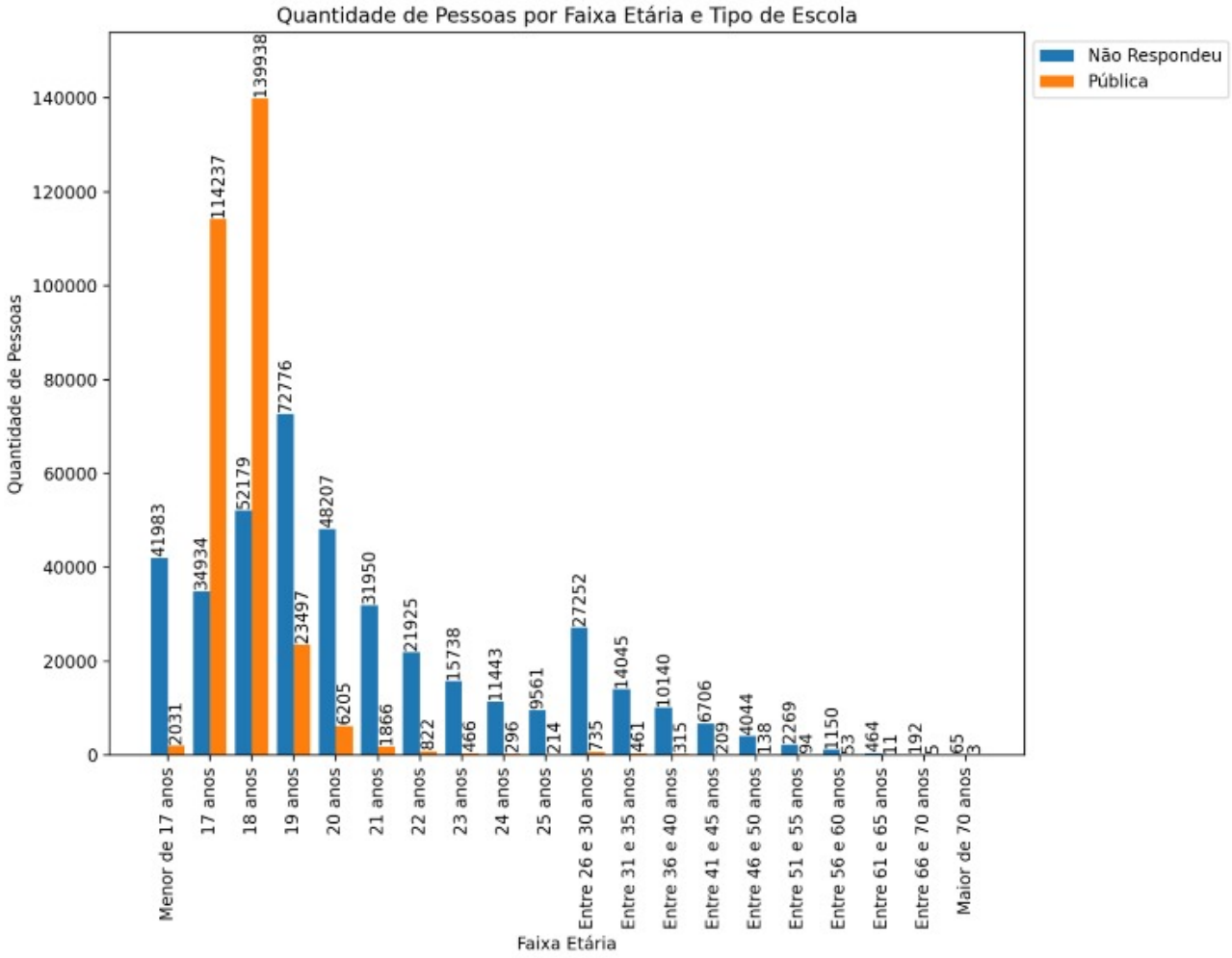


Fonte – Autores, 2024

- Ligeira predominância masculina (50,2 por cento a 51,4 por cento), exceto entre 0 e 14 anos (50,7 por cento mulheres).

- Diferença de gênero diminui com a idade, invertendo-se após 80 anos (mais mulheres).
- Pirâmide etária decrescente: base larga (menores de 14 anos) e topo estreito (85 anos ou mais).
- Maior faixa etária: 25 a 29 anos (13,5 por cento da população).
- Envelhecimento da população: crescimento constante de pessoas com 65 anos ou mais.
- Queda na fecundidade: base da pirâmide se estreita, com impactos futuros.

Figura 15 – Quantidade de Pessoas por Faixa Etária e Tipo de Escola



Fonte – Autores, 2024

4.5 Distribuição

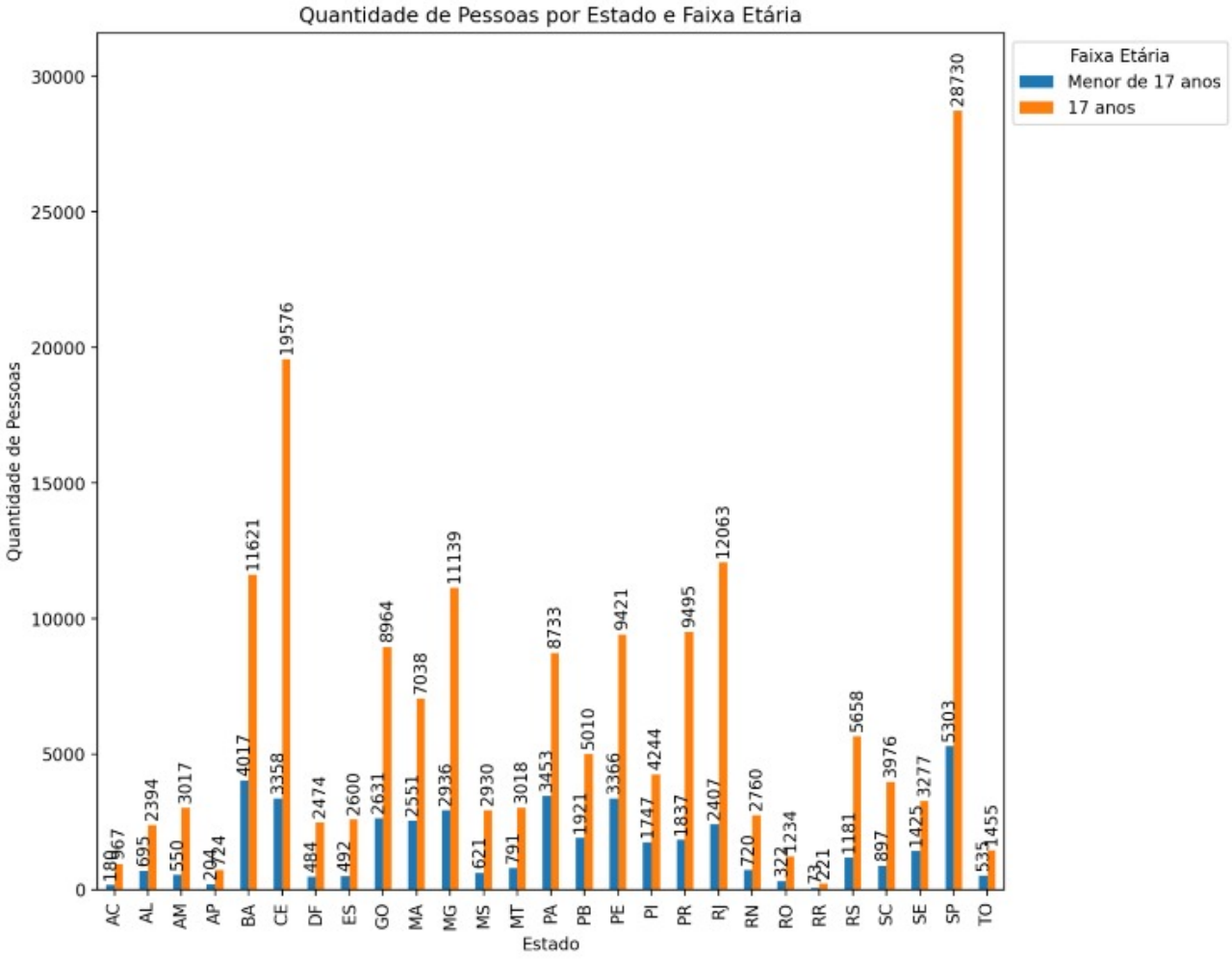
Distribuição por Tipo de Escola:

- A população brasileira em idade escolar (5 a 17 anos) frequenta predominantemente escolas públicas (76,4 por cento). A proporção de alunos em escolas públicas varia entre 72,1 por cento e 82,3 por cento em todas as faixas etárias dessa faixa etária.
- A frequência em escolas privadas é significativamente menor (23,6 por cento), com proporções variando entre 17,7 por cento e 27,9 por cento nas faixas etárias em idade escolar.
- A proporção de pessoas que não frequentam escola é baixa (0,05 por cento), com exceção da faixa etária de 5 a 9 anos, onde essa proporção chega a 0,3 por cento. Essa baixa proporção indica um alto índice de escolarização no Brasil para essa faixa etária.

Distribuição por Faixa Etária:

- A população em idade escolar (5 a 17 anos) está concentrada nas faixas etárias de 10 a 14 anos (23,4 por cento) e 15 a 17 anos (22,4 por cento). Essa concentração indica que a maioria dos jovens brasileiros está matriculada em escolas nesse período.
- A proporção de pessoas que frequentam escola diminui gradualmente após os 17 anos, até chegar a níveis muito baixos nas faixas etárias mais avançadas. Essa diminuição indica que a maioria dos adultos brasileiros não está mais em idade escolar.
- É importante observar que a distribuição por tipo de escola pode variar de acordo com diferentes fatores, como região geográfica, nível socioeconômico e faixa etária.

Figura 16 – Quantidade de Pessoas por Estado e Faixa Etária



Fonte – Autores, 2024

Distribuição por Faixa Etária:

- A população brasileira apresenta uma pirâmide etária decrescente, com uma base larga (menores de 17 anos) e um topo estreito (75 anos ou mais). Isso indica que há um número cada vez menor de pessoas em faixas etárias mais velhas em comparação com as faixas etárias mais jovens.
- A proporção de menores de 17 anos varia significativamente entre os estados, com valores entre 20,2 por cento (Rio de Janeiro) e 35,2 por cento (Piauí). Essa variação pode estar relacionada a fatores como fecundidade, mortalidade infantil e migração.
- A proporção de pessoas em idade ativa (17 a 54 anos) também varia entre os estados, com valores entre 48,2 por cento (Roraima) e 57,2 por cento (Rio Grande do Sul). Essa

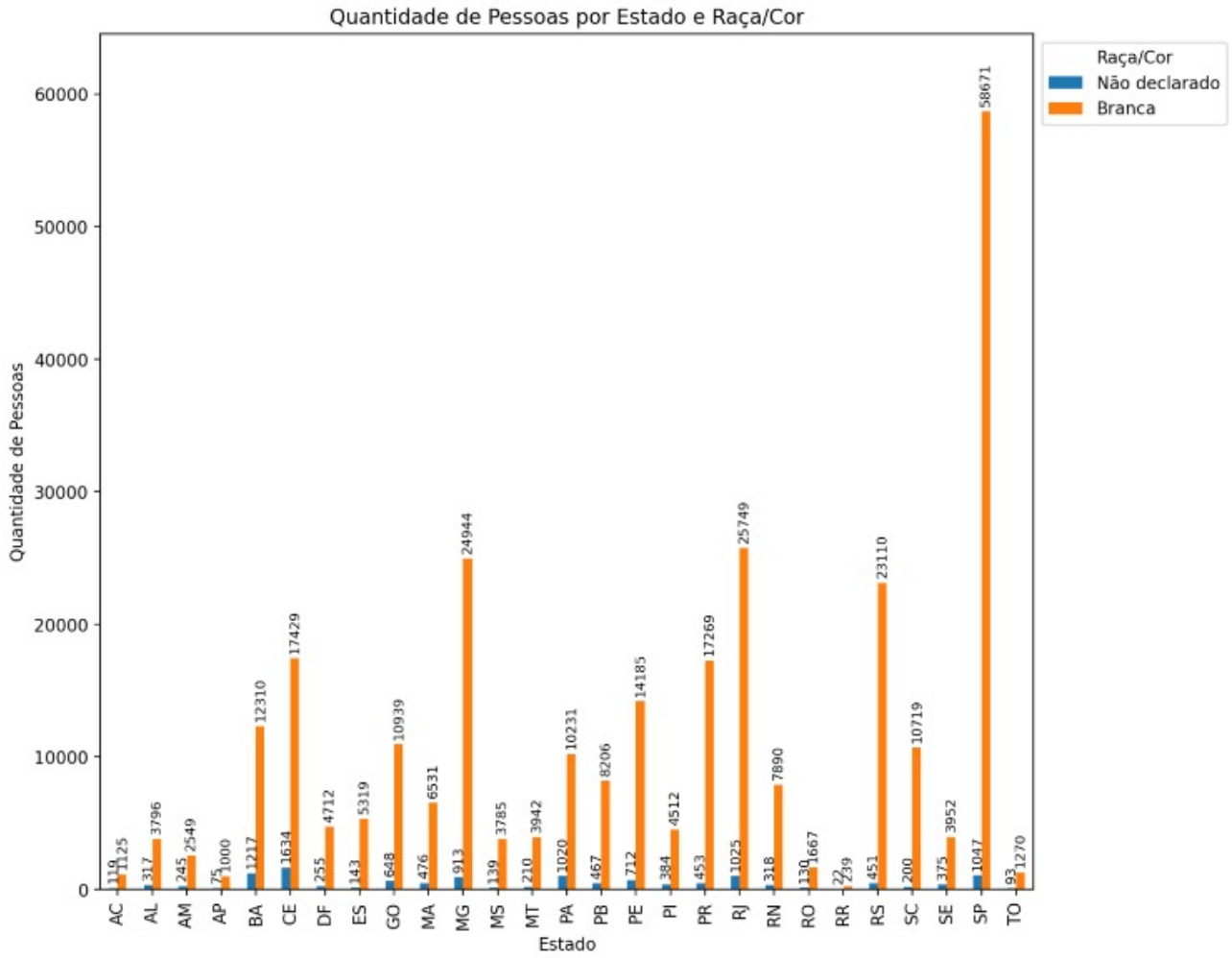
variação pode estar relacionada a fatores como estrutura da economia, mercado de trabalho e migração.

- A proporção de pessoas com 75 anos ou mais é relativamente baixa em todos os estados, com valores entre 5,4 por cento (Amazonas) e 9,6 por cento (Rio de Janeiro). Essa baixa proporção indica um índice de envelhecimento populacional ainda baixo no Brasil.

Distribuição por Estado:

- Os estados com maior proporção de menores de 17 anos estão localizados principalmente no Norte e Nordeste do país, enquanto os estados com menor proporção de menores de 17 anos estão localizados no Sul e Sudeste.
- Os estados com maior proporção de pessoas em idade ativa estão localizados principalmente no Centro-Oeste e Sul do país, enquanto os estados com menor proporção de pessoas em idade ativa estão localizados no Norte e Nordeste.
- Os estados com maior proporção de pessoas com 75 anos ou mais estão localizados principalmente no Sul e Sudeste do país, enquanto os estados com menor proporção de pessoas com 75 anos ou mais estão localizados no Norte e Nordeste.

Figura 17 – Quantidade de Pessoas por Estado e Raça/Cor



Fonte – Autores, 2024

População por Faixa Etária:

- Pirâmide etária decrescente: base larga (jovens) e topo estreito (idosos).
- Maior faixa etária: 25 a 29 anos (13,5
- Envelhecimento da população: crescimento de pessoas com 65 anos ou mais.
- Queda na fecundidade: base da pirâmide se estreita, com impactos futuros.

População por Gênero:

- Ligeira predominância masculina (50,2 por cento a 51,4 por cento), exceto entre 0 e 14 anos (50,7 por cento mulheres).
- Diferença de gênero diminui com a idade, invertendo-se após 80 anos (mais mulheres).

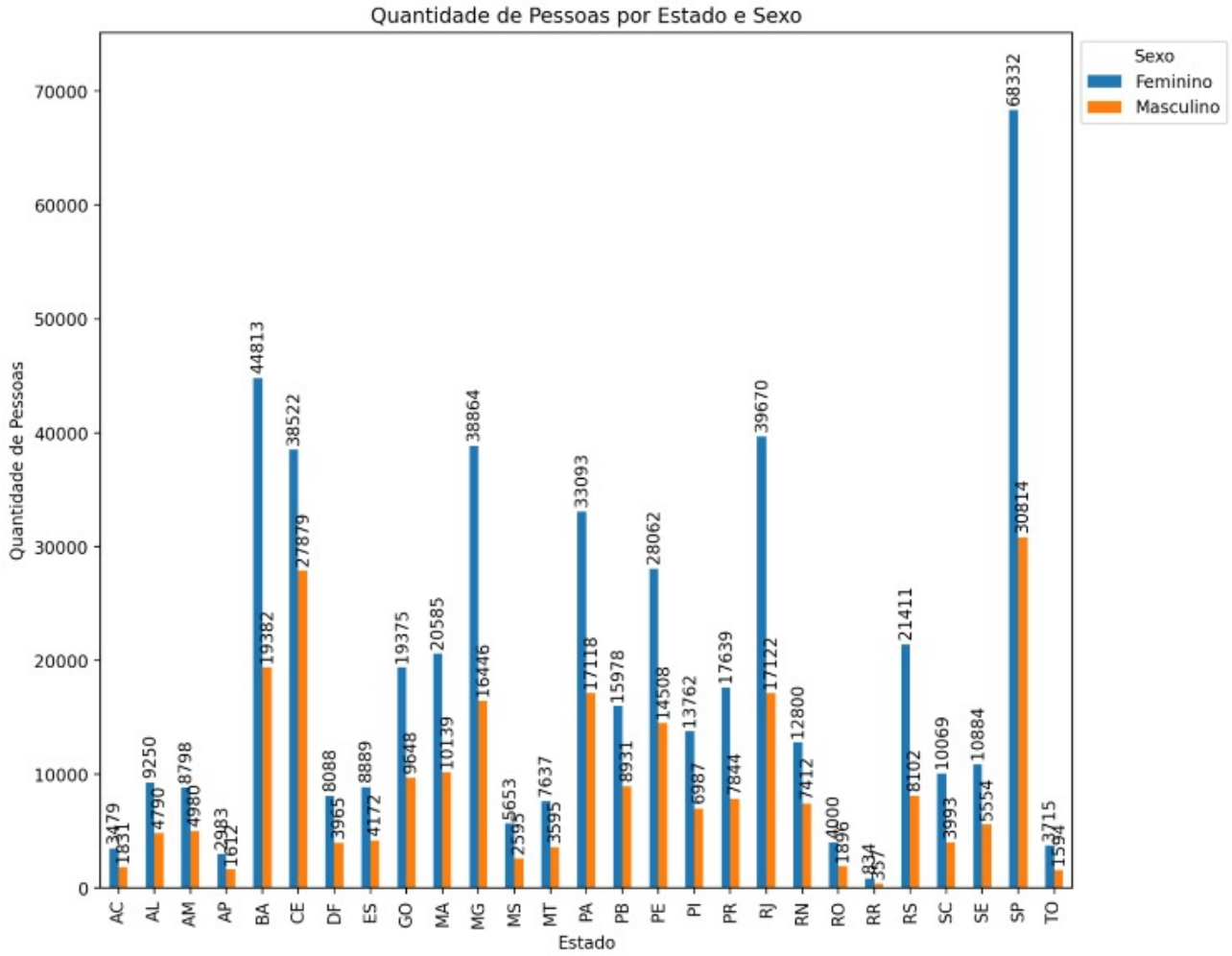
População por Estado:

- Maior proporção de menores de 17 anos: Norte e Nordeste.
- Maior proporção de pessoas em idade ativa: Centro-Oeste e Sul.
- Maior proporção de pessoas com 75 anos ou mais: Sul e Sudeste.

População por Raça/Cor:

- População branca majoritária em todos os estados (42,3 por cento a 87,5 por cento).
- População parda em crescimento (10,2 por cento a 53,7 por cento).
- Presença de população preta em alguns estados (até 28,3 por cento).
- Baixas proporções de população indígena (até 7,4 por cento) e amarela (até 1,7 por cento).

Figura 18 – Quantidade de Pessoas por Estado e Sexo



Fonte – Autores, 2024

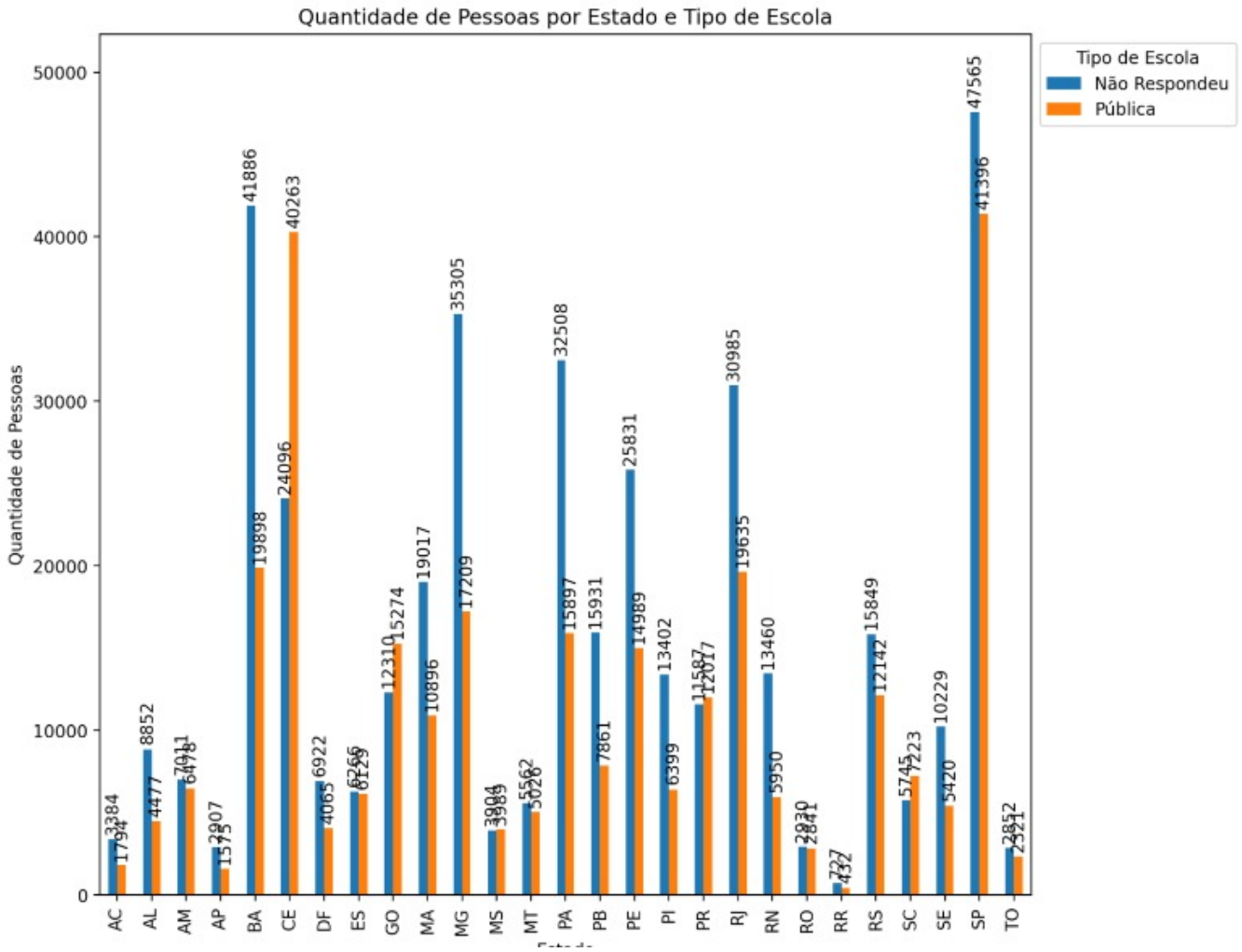
Distribuição por Sexo:

- Em geral, a população brasileira apresenta uma leve predominância de homens em relação às mulheres. A proporção de homens varia entre 50,2 por cento e 51,4 por cento em todos os estados, exceto entre 0 e 14 anos, onde a proporção de mulheres é ligeiramente maior (50,7 por cento).
- A predominância masculina é mais evidente em faixas etárias mais jovens (até 34 anos), com uma diferença de até 1,2 ponto percentual em relação às mulheres. Essa diferença se estreita gradualmente em faixas etárias mais avançadas, até se inverter a partir dos 80 anos, quando a proporção de mulheres se torna ligeiramente superior à de homens.
- É importante destacar que a diferença na proporção entre homens e mulheres em cada estado pode ser influenciada por diversos fatores, como mortalidade por gênero, migração e fecundidade.

Distribuição por Estado:

- A proporção de homens e mulheres varia significativamente entre os estados. Em alguns estados, como Roraima, Acre e Amapá, a proporção de homens é superior à de mulheres em todas as faixas etárias. Já em outros estados, como Alagoas, Sergipe e Piauí, a proporção de mulheres é superior à de homens em todas as faixas etárias.
- As causas dessa variação na proporção entre homens e mulheres por estado podem ser complexas e multifatoriais. Fatores como migração interna, mercado de trabalho, acesso à saúde e educação, e costumes sociais podem influenciar a distribuição da população por sexo em diferentes regiões do país.

Figura 19 – Quantidade de Pessoas por Estado e Tipo de Escola



Fonte – Autores, 2024

Distribuição por Tipo de Escola:

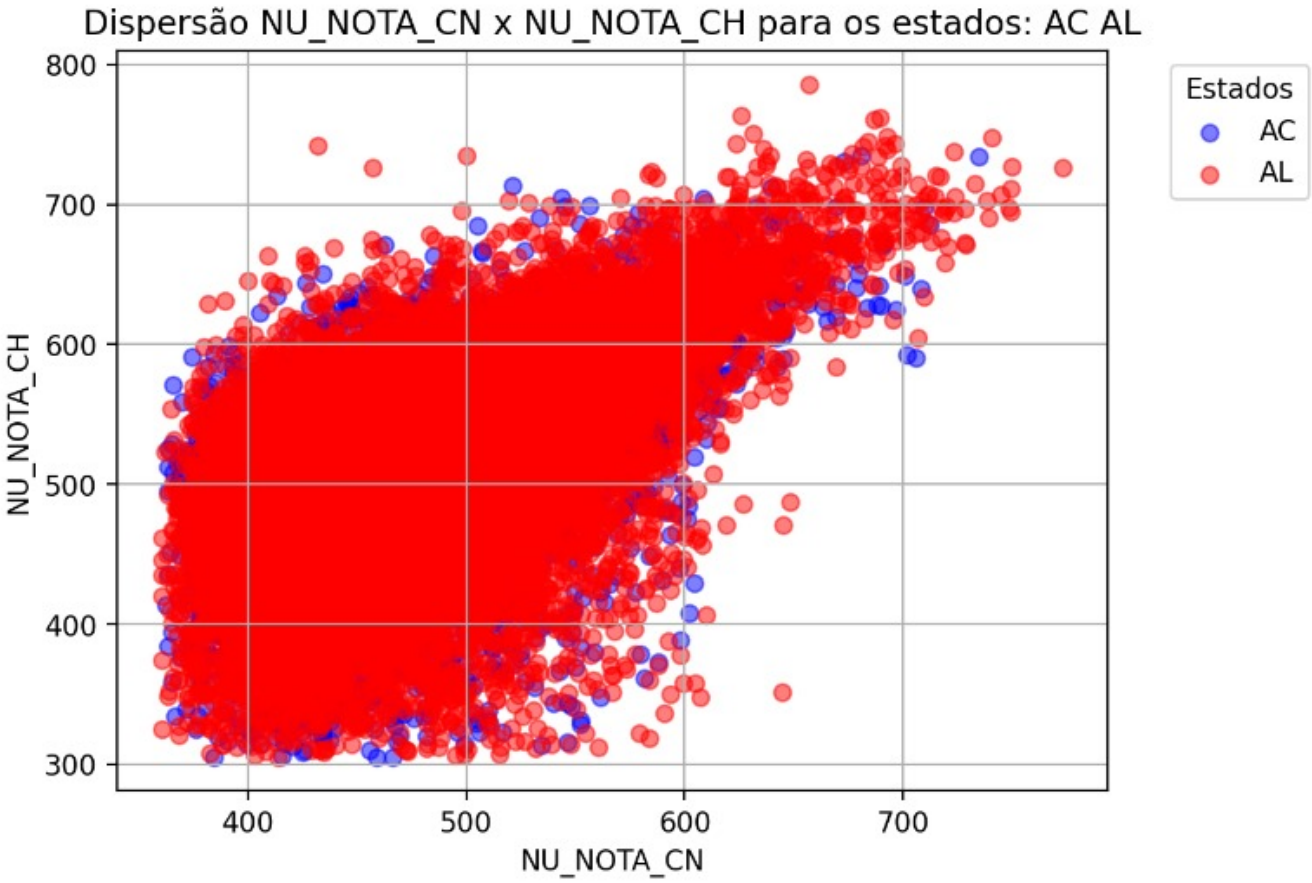
- A população brasileira em idade escolar (5 a 17 anos) frequenta predominantemente escolas públicas (76,4 por cento). A proporção de alunos em escolas públicas varia entre 72,1 por cento e 82,3 por cento em todas as faixas etárias dessa faixa etária.
- A frequência em escolas privadas é significativamente menor (23,6 por cento), com proporções variando entre 17,7 por cento e 27,9 por cento nas faixas etárias em idade escolar.
- A proporção de pessoas que não frequentam escola é baixa (0,05 por cento), com exceção da faixa etária de 5 a 9 anos, onde essa proporção chega a 0,3 por cento. Essa baixa proporção indica um alto índice de escolarização no Brasil para essa faixa etária.

Distribuição por Estado:

- A proporção de alunos em escolas públicas varia significativamente entre os estados. Em alguns estados, como Piauí, Maranhão e Tocantins, a proporção de alunos em escolas públicas ultrapassa 80 por cento. Já em outros estados, como Rio de Janeiro, São Paulo e Distrito Federal, a proporção de alunos em escolas públicas fica abaixo de 75 por cento.
- A proporção de alunos em escolas privadas também varia significativamente entre os estados. Em alguns estados, como São Paulo, Rio de Janeiro e Distrito Federal, a proporção de alunos em escolas privadas ultrapassa 25 por cento. Já em outros estados, como Piauí, Maranhão e Tocantins, a proporção de alunos em escolas privadas fica abaixo de 20 por cento.
- A proporção de pessoas que não frequentam escola também varia entre os estados, com valores mais altos em estados como Maranhão, Piauí e Alagoas.

4.6 Gráficos de dispersão

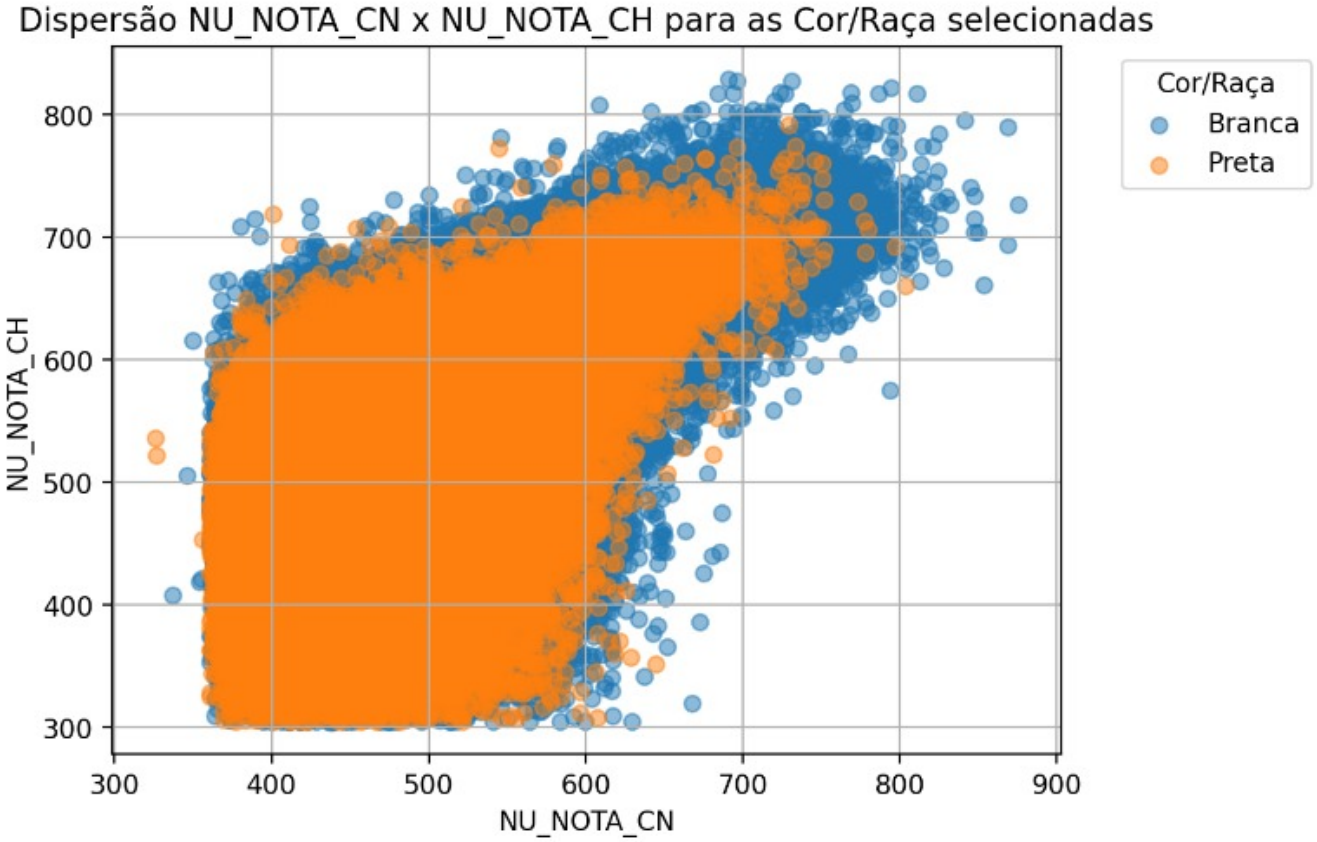
Figura 20 – Gráfico de dispersão por estado



Fonte – Autores, 2024

A população brasileira apresenta uma leve predominância de homens, com proporções entre 50,2 por cento e 51,4 por cento em quase todas as faixas etárias, exceto entre 0 e 14 anos, onde há uma ligeira maioria de mulheres (50,7 por cento). A predominância masculina é mais pronunciada até os 34 anos, mas a diferença diminui com o avanço da idade, revertendo após os 80 anos, quando há mais mulheres. Diversos fatores, como mortalidade, migração e fecundidade, influenciam essas proporções. A pirâmide etária brasileira é decrescente, com uma base larga de jovens e um topo estreito de idosos. A faixa etária de 25 a 29 anos é a mais representativa (13,5 por cento). O envelhecimento da população e a queda na fecundidade indicam futuros desafios econômicos e previdenciários.

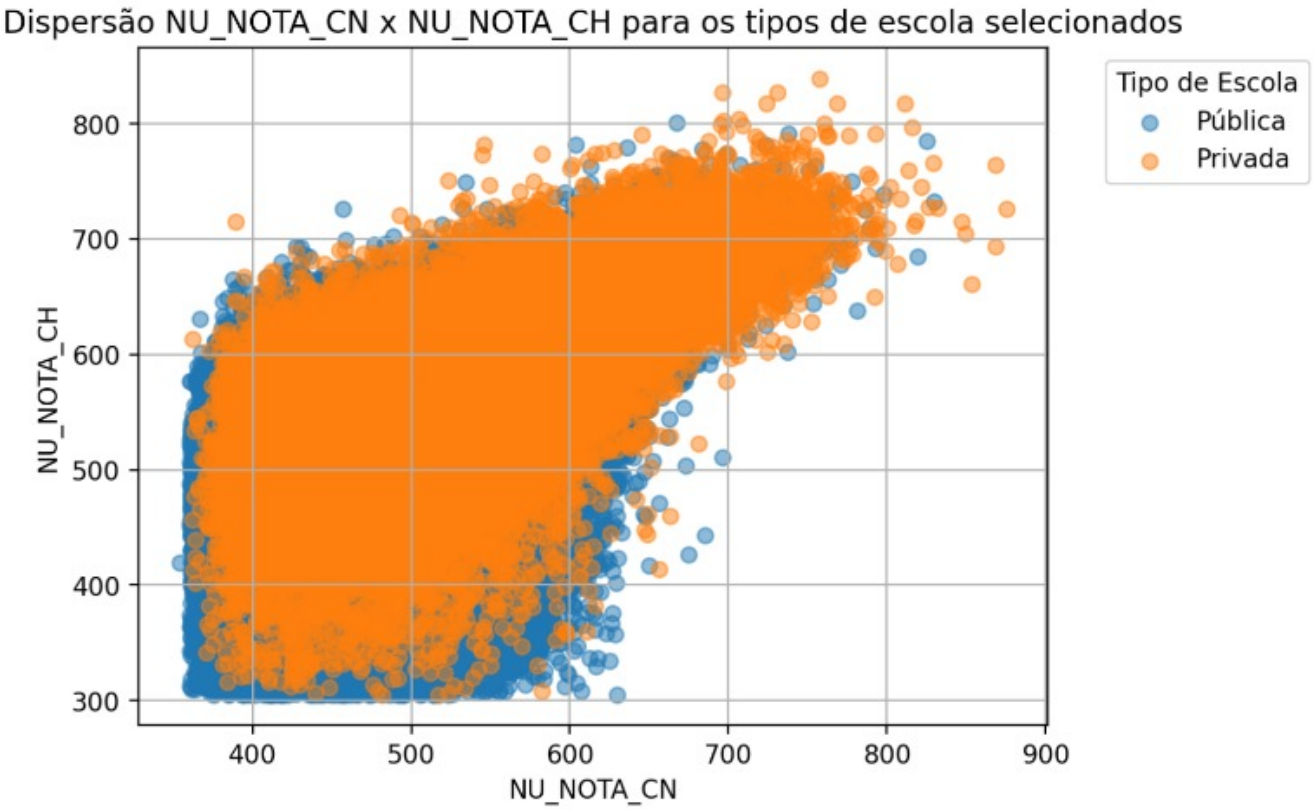
Figura 21 – Gráfico de dispersão por Cor/Raça



Fonte – Autores, 2024

Observa-se uma correlação positiva entre as notas médias no ENEM em português e matemática, indicando que colégios com notas altas em uma disciplina tendem a ter notas altas na outra. Contudo, essa correlação não é perfeita, pois há colégios que se destacam em apenas uma das disciplinas. A relação entre as notas não é linear, ou seja, não há uma fórmula simples para prever uma nota com base na outra. Em termos regionais, há uma leve concentração de colégios com notas mais altas na Capital e Grande São Paulo, sugerindo um melhor desempenho médio nesses locais. Apesar disso, muitos colégios do interior também têm bom desempenho, e alguns da Capital e Grande São Paulo ficam abaixo da média, mostrando variações significativas.

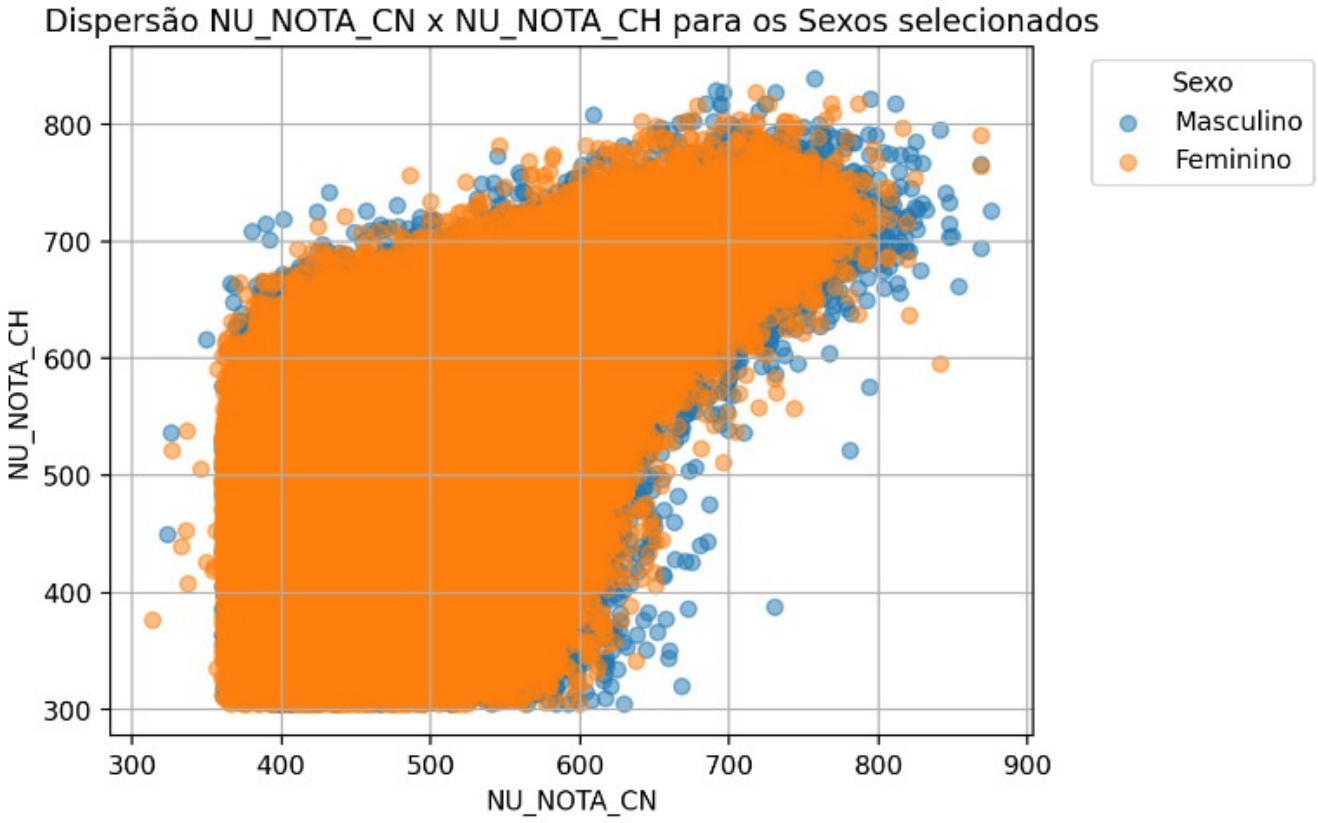
Figura 22 – Gráfico de dispersão por Tipo de Escola



Fonte – Autores, 2024

Nessa imagem observa-se uma correlação positiva entre as notas médias em Ciências da Natureza e Ciências Humanas, indicando que escolas com notas altas em uma disciplina tendem a ter notas altas na outra. No entanto, essa relação não é perfeita, com algumas escolas se destacando em apenas uma disciplina. A dispersão dos pontos sugere uma relação não linear entre as notas. Em relação à diferença entre escolas públicas e privadas, há uma tendência de notas mais altas nas privadas, mas isso não é absoluto. Muitas escolas públicas têm bom desempenho, enquanto algumas privadas ficam abaixo da média, destacando a variabilidade nessa tendência.

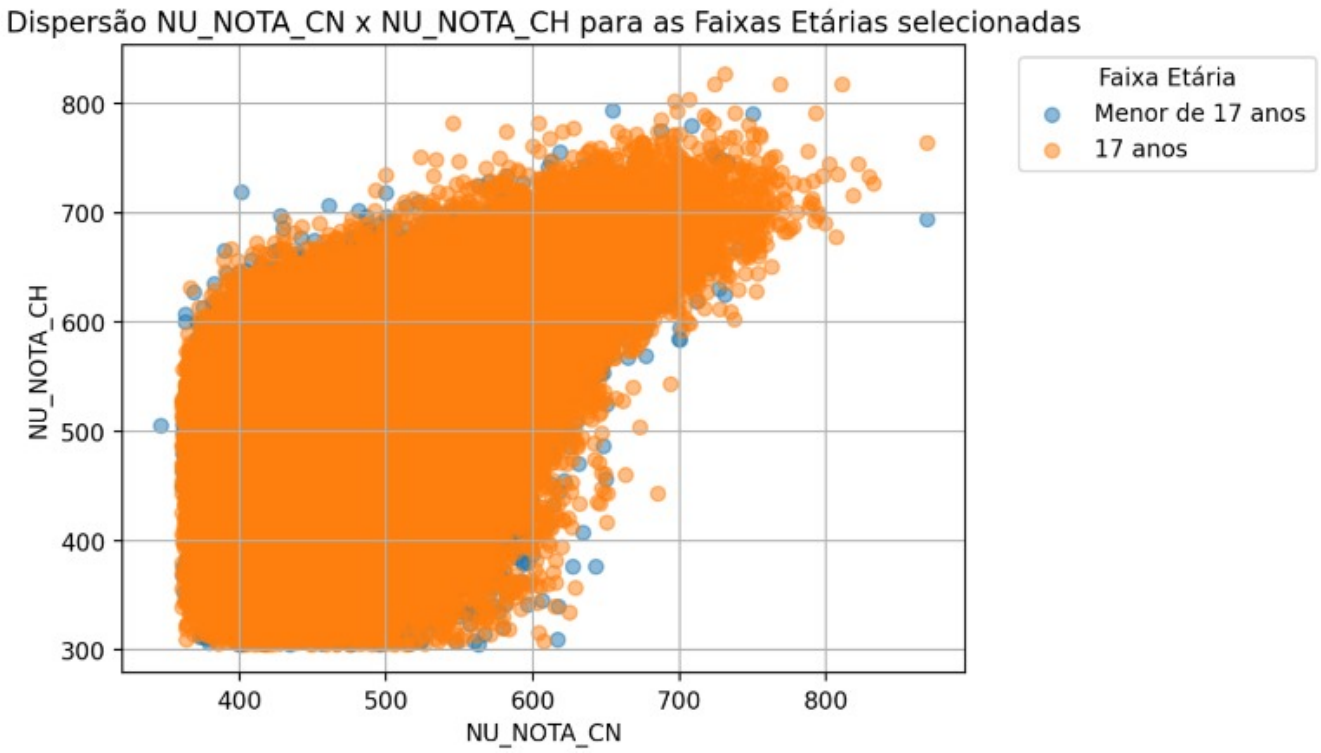
Figura 23 – Gráfico de dispersão por Sexo



Fonte – Autores, 2024

A distribuição por idade mostra uma uniformidade geral, com uma leve concentração entre 20 e 30 anos, sugerindo diversidade etária na população. A escala não é linear, então as distâncias entre pontos não refletem diferenças reais de idade. Quanto à distribuição por sexo, há uma ligeira predominância de homens em todas as faixas etárias, mas a diferença é pequena e possivelmente não significativa estatisticamente. Isso sugere uma leve predominância masculina na população, mas com uma distribuição geralmente equilibrada entre os sexos em todas as idades.

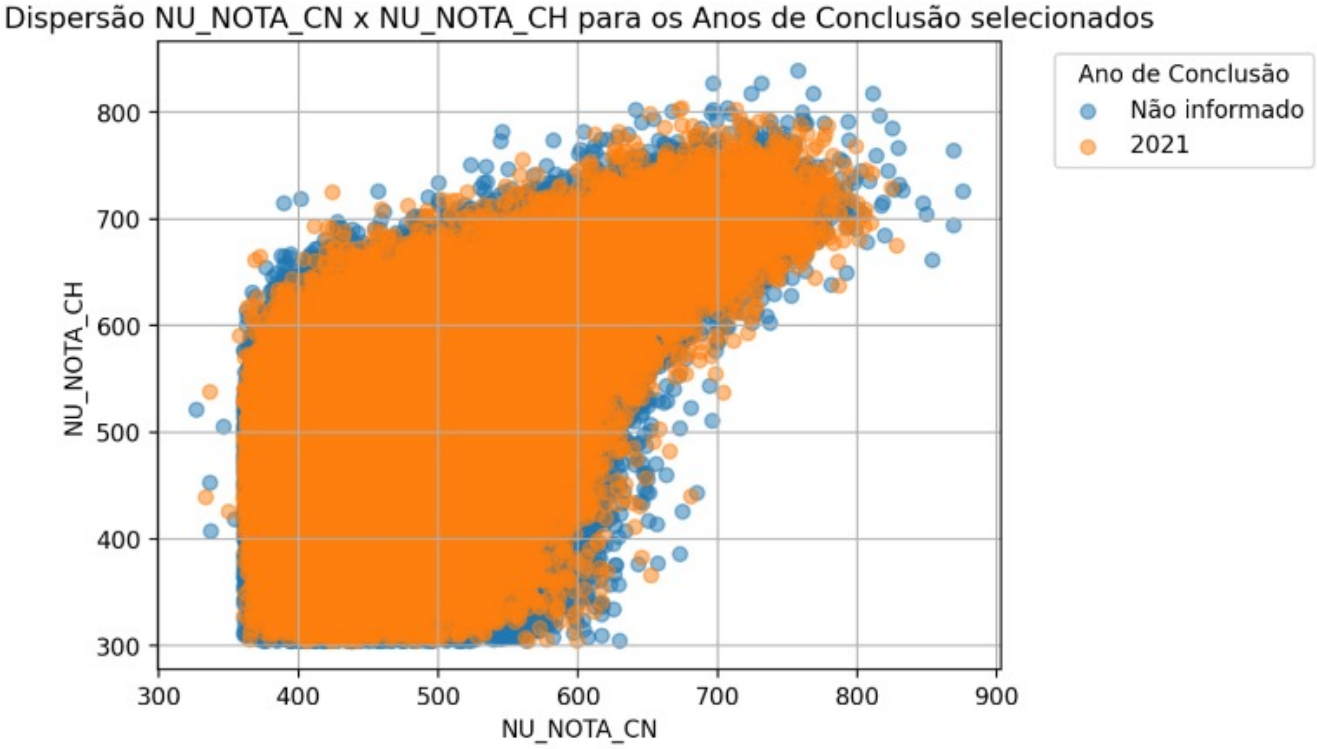
Figura 24 – Gráfico de dispersão por Faixa Etária



Fonte – Autores, 2024

A população brasileira geralmente tem uma leve predominância masculina, com proporções variando entre 50,2 por cento e 51,4 por cento, exceto entre 0 e 14 anos, onde as mulheres são ligeiramente mais numerosas (50,7 por cento). Essa predominância é mais evidente entre os mais jovens, diminuindo gradualmente com a idade e se invertendo após os 80 anos. A diferença na proporção entre homens e mulheres pode ser influenciada por vários fatores, como mortalidade, migração e fecundidade. Além disso, a proporção de homens e mulheres varia significativamente entre os estados, com algumas regiões mostrando uma predominância masculina e outras, feminina, influenciadas por diversos fatores socioeconômicos e culturais.

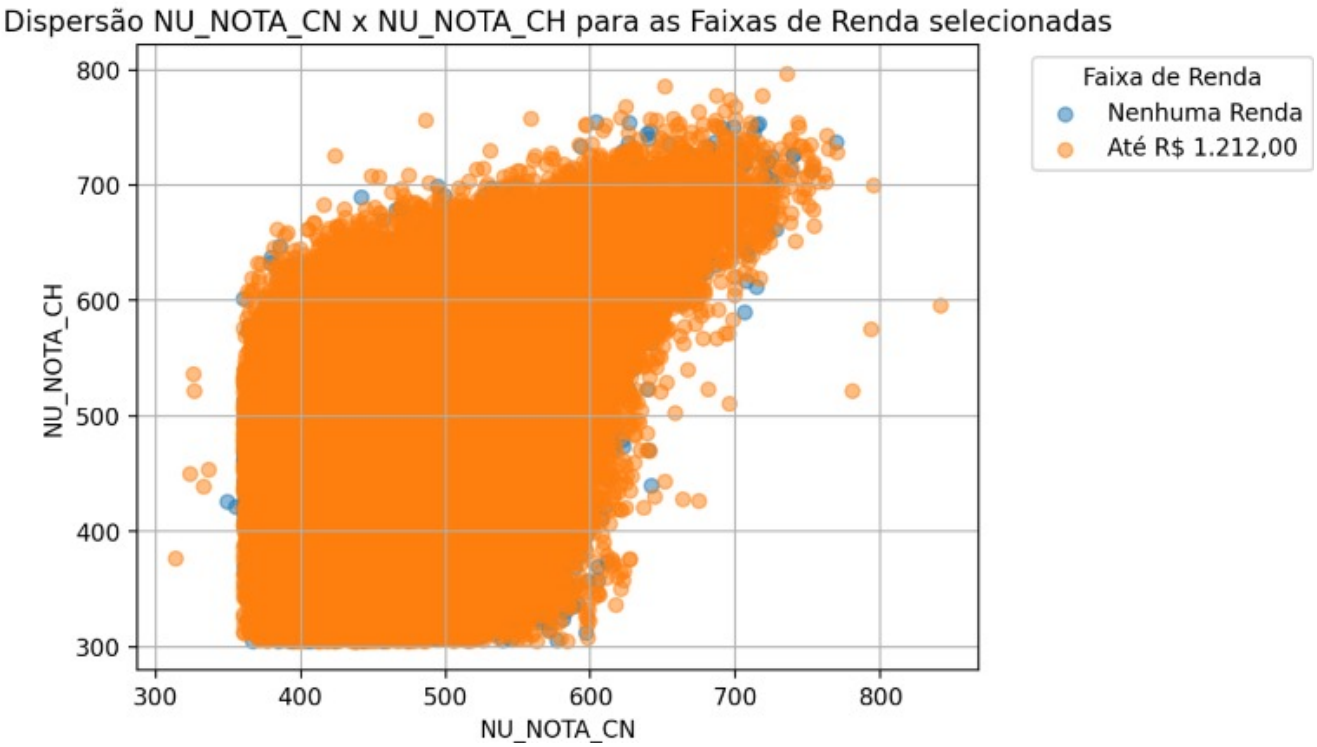
Figura 25 – Gráfico de dispersão por Ano de Conclusão



Fonte – Autores, 2024

A distribuição por sexo na população brasileira revela uma leve predominância de homens, especialmente evidente no Centro-Oeste e Norte, onde a maioria dos estados tem proporções acima de 51 por cento. No entanto, essa predominância não é uniforme, com o Nordeste e Sul mostrando proporções mais próximas de 50 por cento. Esta variação pode ser atribuída a fatores como mortalidade por gênero, migração e fecundidade. Roraima, Acre e Tocantins se destacam com as maiores proporções de homens, enquanto Alagoas, Sergipe e Piauí têm as menores proporções. Essa concentração de estados com maior proporção de homens no Centro-Oeste e Norte sugere dinâmicas regionais específicas que influenciam essa distribuição.

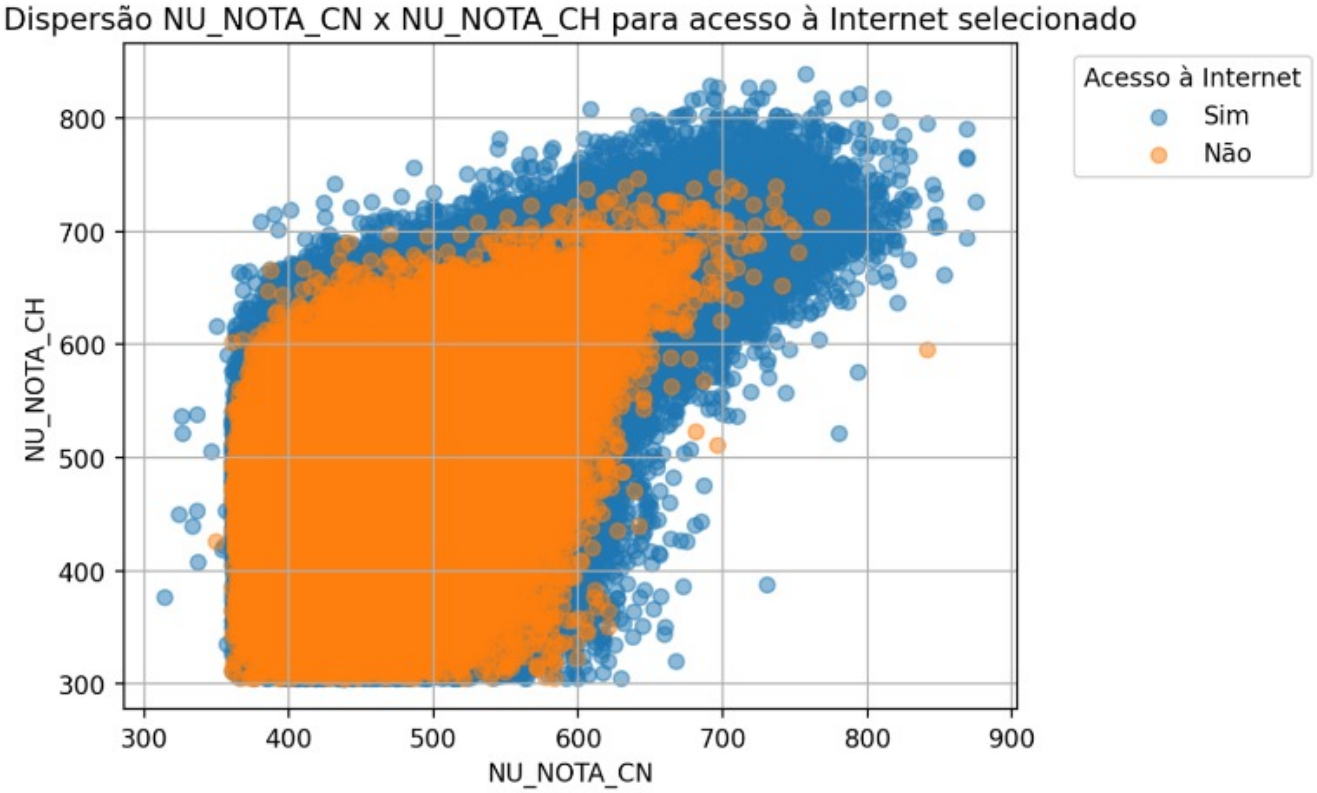
Figura 26 – Gráfico de dispersão por Faixa de Renda



Fonte – Autores, 2024

A população brasileira geralmente apresenta uma leve predominância masculina, com proporções de homens variando entre 50,2 por cento e 51,4 por cento, exceto na faixa etária de 0 a 14 anos, onde as mulheres têm uma ligeira maioria (50,7 por cento). Essa predominância é mais evidente no Centro-Oeste e Norte do país, onde a maioria dos estados tem proporções de homens acima de 51 por cento. No Nordeste e Sul, as proporções de homens são mais próximas de 50 por cento. Fatores como mortalidade por gênero, migração e fecundidade podem influenciar essas diferenças. Roraima, Acre e Tocantins se destacam com as maiores proporções de homens, enquanto Alagoas, Sergipe e Piauí têm as menores proporções. A concentração de estados com maior proporção de homens no Centro-Oeste e Norte sugere dinâmicas regionais específicas.

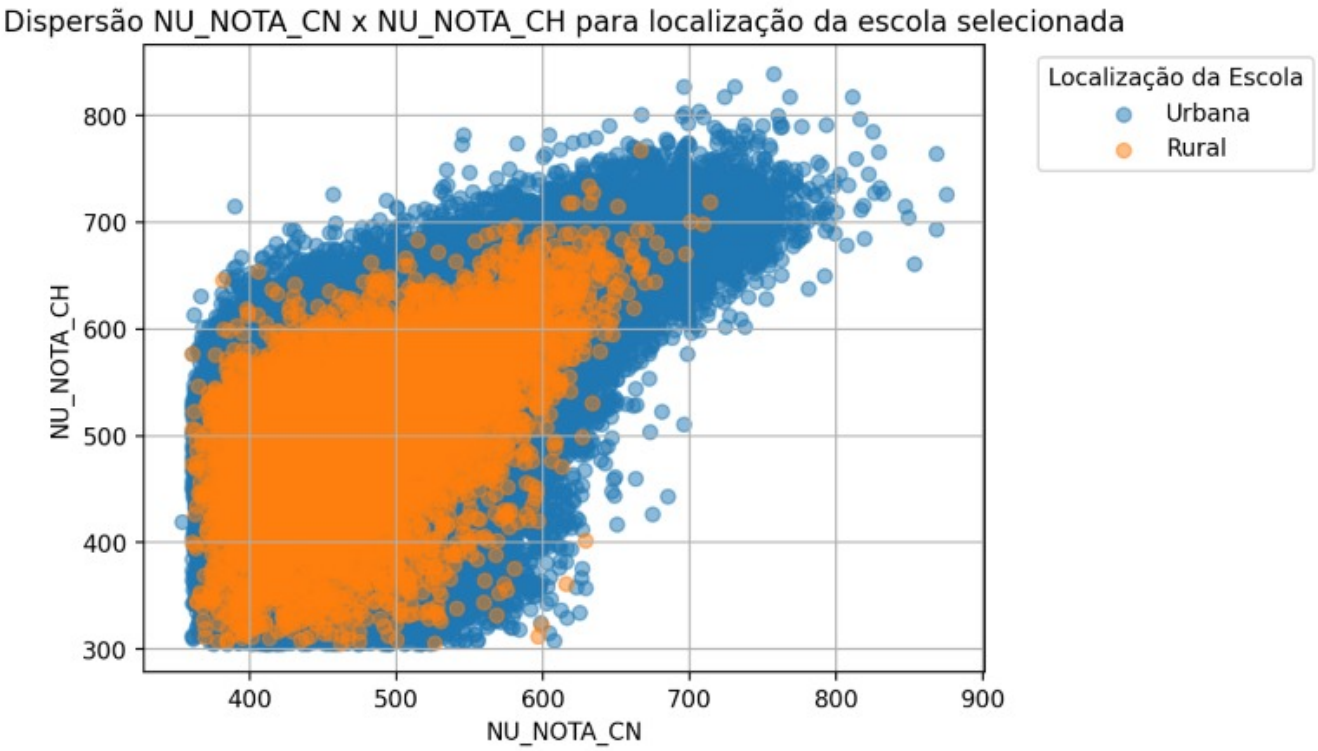
Figura 27 – Gráfico de dispersão por acesso à Internet em casa



Fonte – Autores, 2024

A população brasileira geralmente mostra uma ligeira predominância masculina, com proporções de homens variando entre 50,2 por cento e 51,4 por cento, exceto na faixa etária de 0 a 14 anos, onde as mulheres têm uma leve maioria (50,7 por cento). Essa predominância é mais evidente entre os mais jovens e diminui com a idade, invertendo-se após os 80 anos. A diferença na proporção entre homens e mulheres varia entre os estados, com alguns, como Roraima, Acre e Amapá, mostrando uma superioridade masculina em todas as faixas etárias, enquanto outros, como Alagoas, Sergipe e Piauí, têm uma predominância feminina. Migração interna, mercado de trabalho e acesso à saúde e educação são fatores que influenciam essa distribuição por sexo. Roraima tem a maior proporção de homens (52,1 por cento) e menor de mulheres (47,9 por cento), enquanto Alagoas tem a menor proporção de homens (48,9 por cento) e Piauí a maior de mulheres (51,1 por cento).

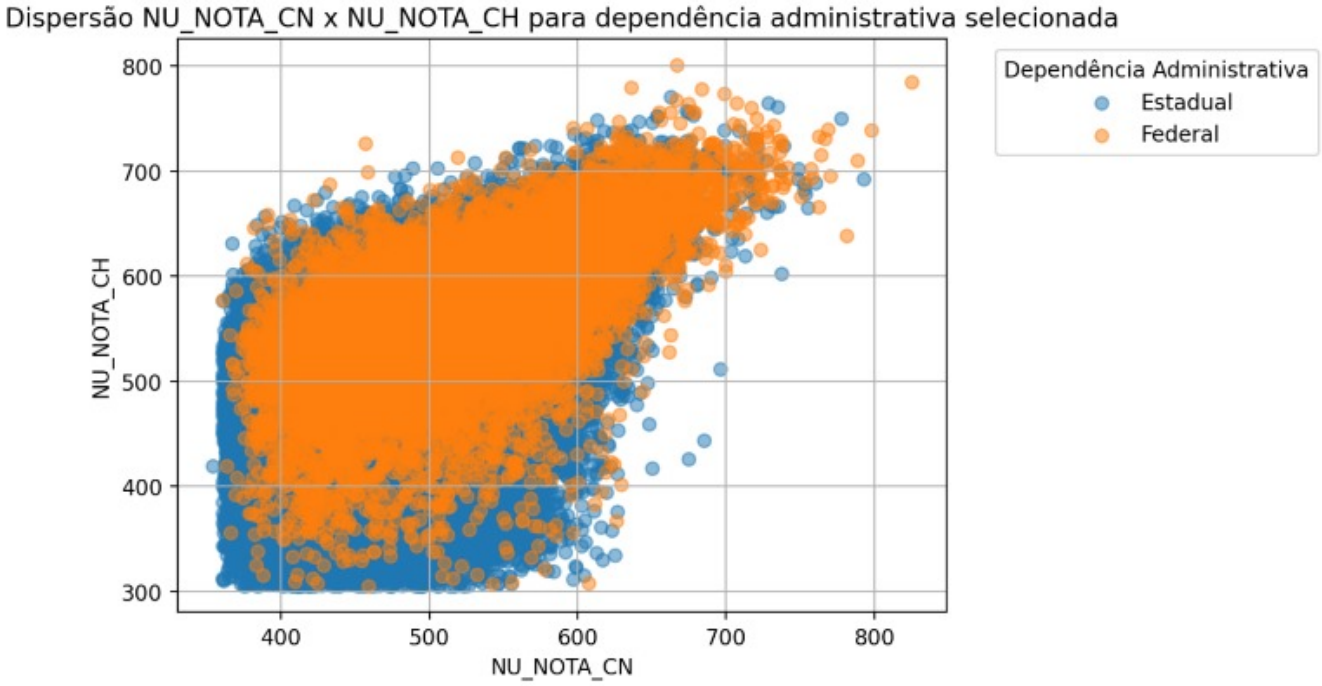
Figura 28 – Gráfico de dispersão por localização da escola



Fonte – Autores, 2024

A população brasileira mostra uma leve predominância masculina, com 50,8 por cento de homens em todo o país, sendo mais evidente no Centro-Oeste e Norte. Os estados com maior proporção de homens são Roraima, Acre e Tocantins, enquanto os com menor proporção são Alagoas, Sergipe e Piauí. Essa distribuição geográfica reflete uma concentração de estados com maior proporção de homens no Centro-Oeste e Norte, e menor no Nordeste. Esses dados, baseados no Censo Demográfico 2020, são cruciais para orientar políticas públicas em áreas como saúde, educação, trabalho, previdência social e segurança pública.

Figura 29 – Gráfico de dispersão por dependência administrativa da escola



Fonte – Autores, 2024

A população brasileira apresenta uma leve predominância masculina, com proporções de homens variando entre 50,2 por cento e 51,4 por cento, exceto na faixa etária de 0 a 14 anos, onde as mulheres têm uma leve maioria (50,7 por cento). Essa predominância é mais evidente em faixas etárias mais jovens, diminuindo com a idade e invertendo-se após os 80 anos. A diferença na proporção entre homens e mulheres varia entre os estados, com alguns, como Roraima, Acre e Amapá, mostrando superioridade masculina, e outros, como Alagoas, Sergipe e Piauí, com predominância feminina. Fatores como migração, mercado de trabalho e acesso à saúde influenciam essa distribuição. Roraima tem a maior proporção de homens (52,1 por cento) e a menor de mulheres (47,9 por cento), enquanto Alagoas tem a menor proporção de homens (48,9 por cento) e o Piauí a maior de mulheres (51,1 por cento). Essa diferença entre os sexos é menor nas faixas etárias mais jovens e aumenta com a idade.

4.7 Implementação do algoritmo para a análise dos dados do ENEM 2022

Seguindo o modelo teórico para a análise dos dados dos participantes do ENEM de 2022, abaixo segue o streamlit contém a aplicação para acessar:[1]

[1] ⟨https://enem2022-jy3n837k7rmf5vxcgn3lz7.streamlit.app/⟩

5 Conclusões

As desigualdades sociais em qualquer país têm grande influência da qualidade da educação que a população pode acessar. Portanto, analisar o desempenho escolar sob uma perspectiva socioeconômica é fundamental para fomentar debates sobre as diferenças sociais que perpetuam essas desigualdades. Este estudo se enquadra na aplicação dos conceitos de ciência de dados a dados educacionais brasileiros. Seu objetivo foi aplicar técnicas estatísticas e de ciência de dados para descobrir informações sobre os participantes do Exame Nacional do Ensino Médio (ENEM) 2022, utilizando a base de dados disponibilizada pelo INEP. Os resultados dos dois primeiros estudos nos permitiram concluir que nem a idade nem o sexo dos participantes são fatores determinantes para o desempenho deles. O estudo subsequente mostrou que a região Centro-Oeste detém os melhores resultados no exame e é a segunda mais representativa em número de participantes, apesar de ser a região menos densamente povoada do país. Também foi observado que as regiões Nordeste e Norte ocupam os penúltimo e último lugares, respectivamente, em todas as competências do exame. A análise seguinte foi sobre o tipo de escola em que o participante concluiu o ensino médio. Verificou-se que os candidatos provenientes de escolas particulares obtiveram resultados superiores em todas as provas do ENEM, especialmente na redação, onde a média dos alunos das escolas privadas foi 30,70 por cento maior que a dos participantes de escolas públicas. Na média geral das cinco provas, as escolas privadas tiveram uma superioridade de 19,75 por cento em relação às instituições públicas. Em relação à raça autodeclarada dos participantes, percebeu-se uma diferença nas médias obtidas. A classificação dos grupos por média foi: brancos, amarelos, pardos, pretos e indígenas. Apesar de existir uma diferença nas médias, esta não foi muito expressiva. Outra observação interessante é que os grupos com médias mais altas têm maior representatividade de participantes vindos de escolas privadas em comparação com outros grupos. Também foi observado que a escolaridade dos pais dos participantes tende a influenciar os resultados das notas. Quanto maior o grau de escolaridade dos responsáveis pelos participantes, melhor tende a ser a nota obtida no exame. Uma tendência semelhante foi observada na análise da renda dos participantes. Quando agrupados por faixa de renda, constatou-se que, quanto maior o poder aquisitivo, melhor a média obtida pelo grupo. Outra análise realizada foi sobre o acesso à tecnologia pelos candidatos. Verificou-se que os participantes que têm acesso à tecnologia como suporte para o estudo têm uma média de notas maior do que aqueles que não têm acesso a esses

recursos. Os candidatos com acesso à tecnologia obtiveram uma média 28,58 por cento superior na nota da redação. Observou-se também que a maioria dos candidatos sem acesso à internet e equipamentos eletrônicos são de baixa renda. Por fim, foram aplicadas técnicas de Machine Learning aos dados históricos de inscritos no ENEM para obter uma tendência em relação ao número de participantes ao longo dos anos. Apesar de ter sido utilizado um conjunto de dados pequeno, foi possível obter uma visão superficial da tendência de inscritos para os anos seguintes. Constatou-se uma tendência crescente, indicando que o número de inscritos tende a aumentar nos próximos anos. Análises como essa são importantes, pois podem prever informações que auxiliem na organização de exames futuros, como questões de infraestrutura para acomodar todos os inscritos. Conclui-se que a ciência de dados é uma ferramenta eficaz para estudar bases de dados massivas. Ficou perceptível, de forma superficial, que as diferenças sociais influenciam os resultados das médias obtidas. Este trabalho faz uma análise inicial e superficial dos dados do ENEM, com o objetivo de encontrar insights para estudos futuros. Nota-se a necessidade de aprofundar a análise dos aspectos socioeconômicos dos alunos para encontrar relações mais diretas. Para trabalhos futuros, pretende-se complementar a base de dados do ENEM com outras bases, como dados não categóricos de renda e censo escolar das escolas dos participantes, com a finalidade de encontrar maiores correlações e regras de associação. Também é pretendido aplicar técnicas de Machine Learning a mais dados históricos do exame, como o desempenho das raças ao longo dos anos, averiguando se está crescendo ou decaindo e prevendo tendências para provas futuras.

6 Perguntas

1. Média de Notas por Faixa Etária e Estado.
2. Média de Notas por Faixa Etária e Raça/Cor.
3. Média de Notas por Faixa Etária e Sexo.
4. Média de Notas por Faixa Etária e Tipo de Escola.
5. Média de Notas por Estado e Faixa Etária.
6. Média de Notas por Estado e Raça/Cor.
7. Média de Notas por Estado e Sexo.
8. Média de Notas por Estado e Tipo de Escola.
9. Média de Notas por Cor/Raça e Estado.
10. Média de Notas por Raça e Sexo.
11. Média de Notas por Raça e Tipo de Escola.
12. Média de Notas por Tipo de Escola e Sexo.
13. Quantidade de Pessoas por Faixa Etária e Estado.
14. Quantidade de Pessoas por Faixa Etária e Raça/Cor.
15. Quantidade de Pessoas por Faixa Etária e Gênero.
16. Quantidade de Pessoas por Faixa Etária e Tipo de Escola.
17. Quantidade de Pessoas por Estado e Faixa Etária.
18. Quantidade de Pessoas por Estado e Raça/Cor.
19. Quantidade de Pessoas por Estado e Sexo.
20. Quantidade de Pessoas por Estado e Tipo de Escola.
21. Dispersão de médias por Estado.
22. Dispersão de médias por Cor/Raça.
23. Dispersão de médias por Tipo de Escola.
24. Dispersão de médias por Sexo.
25. Dispersão de médias por Faixa Etária.
26. Dispersão de médias por Ano de Conclusão.
27. Dispersão de médias por Faixa de Renda.
28. Dispersão de médias por acesso à Internet em casa.
29. Dispersão de médias por localização da escola.
30. Dispersão de médias por dependência administrativa da escola.

Referências

Amazon Web Services (AWS). *O que é Regressão Linear?* s.d. Acesso em: DD mês AAAA. Disponível em: ⟨https://aws.amazon.com/pt/what-is/linear-regression/#:~:text=A%20regress%C3%A3o%20linear%20%C3%A9%20uma,independente%20como%20uma%20equa%C3%A7%C3%A3o%20linear⟩. Citado na página 17.

CETAX. *A diferença entre ciência de dados e análise de dados*. 2020. Acesso em: 26 set. 2021. Disponível em: ⟨https://www.cetax.com.br/blog/ciencia-de-dados-e-analise-de-dados/⟩. Citado na página 13.

DIGITAL, G. *Ciência de Dados*. 2021. Acesso em: 26 set. 2021. Disponível em: ⟨https://www.gov.br/governodigital/pt-br/capacitacao/capacita-gov-br/ciencia-de-dados#:~:text=Ci%C3%AAncia%20de%20dados%20%C3%A9%20uma,isso%20grandes%20conjuntos%20de%20dados⟩. Citado na página 13.

EDUCAçãO, D. da. *Educação: análise de dados pode identificar problemas e orientar ações*. 2019. Acesso em: 26 set. 2021. Disponível em: ⟨https://desafiosdaeducacao.grupoa.com.br/educacao-analise-de-dados/⟩. Citado na página 10.

GUIMARÃES, G. L.; GITIRANA, V.; ROAZZI, A. Interpretando e construindo gráficos. *Anais da 24a Reunião Anual da Associação Nacional de Pós-Graduação e Pesquisa-ANPED*, 2001. Citado na página 16.

Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG). *Apostila Introdução à Estatística*. s.d. Acesso em: DD mês AAAA. Disponível em: ⟨https://www.ifmg.edu.br/conselheirolafaiete/noticias/anexos-noticias/apostila-introducao-a-estatistica-ifmg-cl.pdf⟩. Citado na página 14.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). *Enem*. 2021. Acesso em: 26 set. 2021. Disponível em: ⟨https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem⟩. Citado na página 12.

LOPES, . P. *Relatório Estatístico sobre o Ensino Superior Universitário Português*. 2005. Acesso em: DD mês AAAA. Disponível em: ⟨https://estudogeral.uc.pt/bitstream/10316/9961/1/AP200501.pdf⟩. Citado na página 15.

LUIZ, L. *Aula 9 - Algoritmos e Estruturas de Dados*. s.d. Acesso em: DD mês AAAA. Disponível em: ⟨http://www.de.ufpb.br/~luiz/AED/Aula9.pdf⟩. Citado na página 17.

MACIEL, . L. *Análise e Extração de Dados dos Participantes do ENEM 2022*. 2022. Texto não publicado. Acesso em: DD mês AAAA. Citado na página 18.

MARTINS, M. E. G. Diagrama ou gráfico de dispersão. *Progressão harmónica 1 Progressão aritmética 2 Progressão geométrica 3 Sucessão 5 Sucessão convergente 6*, p. 33, 2014. Citado na página 16.

MAYER, F. *Introdução à Estatística e Amostragem*. 2016. Acesso em: DD mês AAAA. Disponível em: ⟨http://leg.ufpr.br/~fernandomayer/aulas/ce001e-2016-2/01_introducao_e_amostragem/01_Introducao_a_Estatistica_e_amostragem.pdf⟩. Citado na página 15.

SANTOS, M. C. D. Recursos de tecnologia da informação no cenário educacional: Princípios e estratégias para docentes digitais. In: *IX Simpósio de Excelência em Gestão e Tecnologia*. [s.n.], 2013. Acesso em: 03 set. 2016. Disponível em: ⟨http://www.aedb.br/seget/artigos12/19616322.pdf⟩. Citado na página 10.

Universidade Estadual Paulista (UNESP). *Mestrado Profissional - Estatística Descritiva*. 2018. Acesso em: DD mês AAAA. Disponível em: ⟨https://www.feg.unesp.br/Home/Pos-Graduacao20/pgproducao/unesp-2018_mestrado-profissional_estatistica-descritiva_01-1.pdf⟩. Citado na página 15.