



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANTONIO MÁDSON ROCHA

ANÁLISE DE DADOS: PERFIL E DESEMPENHO DOS PARTICIPANTES DAS
EDIÇÕES DO ENEM 2019 A 2022 SOB A PERSPECTIVA DA COVID-19

MOSSORÓ
2023

ANTONIO MÁDSON ROCHA

**ANÁLISE DE DADOS: PERFIL E DESEMPENHO DOS PARTICIPANTES DAS
EDIÇÕES DO ENEM 2019 A 2022 SOB A PERSPECTIVA DA COVID-19**

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Angélica Felix de Castro, Prof. Dra.

Co-orientadora: Amanda Gondim de Oliveira, Prof. Dra.

MOSSORÓ

2023

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

R672a Rocha, Antonio Mádsen.
Análise de dados: perfil e desempenho dos
participantes das edições do ENEM 2019 a 2022 sob
a perspectiva da COVID-19 / Antonio Mádsen Rocha.
- 2023.
53 f. : il.

Orientadora: Angélica Félix de Castro.
Coorientadora: Amanda Gondim de Oliveira.
Monografia (graduação) - Universidade Federal
Rural do Semi-árido, Curso de Ciência da
Computação, 2023.

1. ENEM. 2. COVID-19. 3. Ciência de Dados. 4.
Python. I. Castro, Angélica Félix de, orient. II.
Oliveira, Amanda Gondim de, co-orient. III.
Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade
com AACR2 e os dados fornecidos pelo(a) autor(a).
Biblioteca Campus Mossoró / Setor de Informação e Referência
Bibliotecária: Keina Cristina Santos Sousa e Silva
CRB: 15/120

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

Antonio Mádsen Rocha

Análise de dados: perfil e desempenho dos participantes das edições do ENEM
2019 a 2022 sob a perspectiva da COVID-19

Monografia apresentada a Universidade
Federal Rural do Semi-Árido como
requisito para obtenção do título de
Bacharel em Ciência da Computação.

Defendida em: 09/10/2023.

BANCA EXAMINADORA

Documento assinado digitalmente
 **ANGELICA FELIX DE CASTRO**
Data: 18/10/2023 20:03:25-0300
Verifique em <https://validar.iti.gov.br>

Angélica Felix de Castro, Prof. Dra. (UFERSA)
Presidente e orientadora

Documento assinado digitalmente
 **AMANDA GONDIM DE OLIVEIRA**
Data: 18/10/2023 17:11:18-0300
Verifique em <https://validar.iti.gov.br>

Amanda Gondim de Oliveira, Prof. Dra. (UFERSA)
Co-orientadora

Documento assinado digitalmente
 **LENARDO CHAVES E SILVA**
Data: 18/10/2023 21:12:10-0300
Verifique em <https://validar.iti.gov.br>

Lenardo Chaves e Silva, Prof. Dr. (UFERSA)
Membro Examinador

Documento assinado digitalmente
 **SEBASTIAO EMIDIO ALVES FILHO**
Data: 20/10/2023 10:07:34-0300
Verifique em <https://validar.iti.gov.br>

Sebastião Emídio Alves Filho, Prof. Dr. (UERN)
Membro Examinador

AGRADECIMENTOS

À Deus, em todos os seus conceitos e formas, por tudo isso ter se tornado possível.

À minha família, em especial aos meus pais, por todo o apoio fornecido nesta longa jornada, por serem fonte de inspiração e por mostrar que a vida é cheia de possibilidades.

Às professoras Dra. Angélica Félix e Dra. Amanda Gondim pela orientação e encorajamento dados para realização deste trabalho.

À todo o corpo docente que fez parte da minha formação, integrando novos conhecimentos.

Aos funcionários da Fundação Guimarães Duque por confiar no meu trabalho e tornar minhas manhãs mais tranquilas.

Aos moradores da Casa 17 da Vila Acadêmica da UFERSA, pelo acolhimento e as trocas de experiências de vida.

Aos meus amigos próximos, que estão comigo há bastante tempo e também, os que vieram ao longo desta trajetória, pelos puxões de orelha, conversas aleatórias e por tornarem meus dias mais coloridos.

RESUMO

A COVID-19 teve altos índices de óbitos no Brasil desde seu início em março de 2020, quando se tornou uma pandemia. Apesar dos desafios enfrentados desde então em todos os setores da sociedade, é possível ter um resultado do impacto da COVID-19 no desempenho dos candidatos no Exame Nacional do Ensino Médio (ENEM) de 2019 a 2022, pois, este último, foi o ano em que menos apresentou índices de óbitos. Visto isto, o objetivo deste trabalho é aplicar técnicas da ciência de dados para averiguar se houve ou não impacto da doença no desempenho dos participantes, realizando um estudo comparativo entre as edições mencionadas. Analisar, também, a possível mudança nos perfis dos candidatos, bem como os diversos grupos que estão inseridos. Para isso, foram utilizadas bibliotecas da linguagem *Python*, como *Pandas*, *Numpy*, *Matplotlib* e *Seaborn* para extrair informações do conjunto de dados e representá-los por meio de gráficos. Dos resultados encontrados, é possível notar um impacto no que diz respeito a quantidade de participantes por raça autodeclarada, no qual, de 2019 a 2021, a maior queda foi dos povos indígenas de 51,03% e a menor, 28,76%, dos candidatos que se autodeclararam brancos. De modo geral, a pandemia não impactou o desempenho dos candidatos, no qual de 2019 a 2022, houve um aumento de 508,39 pontos para 516,72 pontos, uma diferença de 2,97% na média dos participantes do exame.

Palavras-chave: ENEM. COVID-19. Ciência de Dados. Python.

ABSTRACT

COVID-19 has had high death rates in Brazil since its beginning in March 2020, when it became a pandemic. Despite the challenges faced since then in all sectors of society, it is possible to have a result of the impact of COVID-19 on the performance of candidates in the Exame Nacional do Ensino Médio (ENEM) from 2019 to 2022, as the latter was the year in which there were fewer death rates. Given this, the objective of this work is to apply data science techniques to determine whether there was an impact of the disease on the participants' performance by carrying out a comparative study between the editions mentioned. Also analyze the possible changes in the profiles of candidates as well as the different groups that are included. For this, Python language libraries such as Pandas, Numpy, Matplotlib and Seaborn extract information from the data set and represent it through graphs. From the results found, it is possible to notice an impact with regard to the number of participants by self-declared race, in which, from 2019 to 2021, the biggest drop was for indigenous peoples, at 51.03%, and the smallest, 28.76%, of candidates who declare themselves white. In general, the pandemic did not impact candidates performance. In general, the pandemic did not impact student performance; from 2019 to 2022, there was an increase from 508.39 points to 516.72 points, a difference of 2.97% in the average of exam participants.

Keywords: ENEM. COVID-19. Data Science. Python.

LISTA DE FIGURAS

Figura 1	– Exemplo de boxplot	15
Figura 2	– Exemplo da aplicação do algoritmo de K-médias.....	19
Figura 3	– Número de participantes por intervalo de idade e ano de aplicação	22
Figura 4	– Boxplot da faixa-etária dos participantes por edição do ENEM.....	23
Figura 5	– Médias das cinco disciplinas por intervalo de idade e edição do ENEM	24
Figura 6	– Distribuição dos participantes por sexo e edição do ENEM	25
Figura 7	– Desempenho dos participantes por sexo e edição do ENEM	26
Figura 8	– Desempenho dos participantes na disciplina de Ciências da Natureza de acordo com o sexo e edição do ENEM	27
Figura 9	– Desempenho dos participantes na disciplina de Ciências Humanas de acordo com o sexo e edição do ENEM	28
Figura 10	– Desempenho dos participantes na disciplina de Matemática de acordo com o sexo e edição do ENEM	29
Figura 11	– Desempenho dos participantes na disciplina de Linguagens e Códigos de acordo com o sexo e edição do ENEM	30
Figura 12	– Desempenho dos participantes na Redação de acordo com o sexo e edição do ENEM.....	31
Figura 13	– Distribuição dos participantes por Região nas edições de 2019 e 2020 do ENEM.....	32
Figura 14	– Distribuição dos participantes por Região nas edições 2021 e 2022.....	33
Figura 15	– Médias dos participantes por região e ano de edição do ENEM.....	34
Figura 16	– Percentual dos participantes por rede de ensino e edição do ENEM.....	35
Figura 17	– Média dos participantes nas disciplinas de Ciências Humanas e da Natureza por rede de ensino e edição do ENEM.....	35

Figura 18	– Média dos participantes nas disciplinas de Matemática e Linguagens por rede de ensino e edição do ENEM	36
Figura 19	– Média dos participantes na Redação por rede de ensino e edição do ENEM.....	37
Figura 20	– Desempenho médio dos participantes por edição do ENEM	38
Figura 21	– Número de candidatos por raça autodeclarada e edição do ENEM.....	39
Figura 22	– Distribuição dos participantes por escolaridade dos responsáveis de acordo com a edição do ENEM	41
Figura 23	– Média dos participantes por escolaridade dos responsáveis de acordo com a edição do ENEM	42
Figura 24	– Número de candidatos por renda mensal de cada edição do ENEM.....	43
Figura 25	– Médias dos candidatos por renda mensal de cada edição do ENEM.....	44
Figura 26	– Percentual de distribuição dos participantes com acesso à tecnologia nas edições 2019 e 2020 do ENEM.....	45
Figura 27	– Percentual de distribuição dos participantes com acesso à tecnologia nas edições 2021 e 2022 do ENEM.....	46

LISTA DE TABELAS

Tabela 1	– Descrição das colunas utilizadas.....	20
Tabela 2	– Percentagem do número de óbitos causados pela COVID-19 distribuídos por região no Brasil	31
Tabela 3	– Desempenho médio e a frequência percentual de raças autodeclaradas por edição do ENEM	38
Tabela 4	– Percentual da distribuição dos participantes com raças autodeclaradas de acordo com rede de ensino por edição do ENEM.....	39
Tabela 5	– Média dos participantes por disciplina de acordo com o acesso às tecnologias nas edições no ENEM.....	46
Tabela 6	– Média dos participantes por grupo no ENEM.....	47

LISTA DE ABREVIATURAS E SIGLAS

ENEM	Exame Nacional do Ensino Médio
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
OMS	Organização Mundial da Saúde
ESPII	Emergência de Saúde Pública de Importância Internacional
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
EaD	Ensino à Distância
SISU	Sistema de Seleção Unificada
ProUni	Programa Universidade para Todos
FIES	Fundo de Financiamento Estudantil
LGPD	Lei Geral de Proteção de Dados
MPF	Ministério Público Federal
IBGE	Instituto Brasileiro de Geografia e Estatística

SUMÁRIO

1 Introdução.....	11
2 Fundamentação Teórica	12
2.1 Fundamentos básicos da estatística	12
2.1.2 Estatística Descritiva	12
2.1.3 Variável, População e Amostra	12
2.1.4 Distribuição e as Medidas Estatísticas	13
2.1.5 Medidas de Tendência Central.....	13
2.1.6 Representação Gráfica.....	14
2.1.7 Boxplot	14
2.2 A pandemia da COVID-19.....	15
2.3 O ENEM e a LGPD	15
2.4 Ciência de Dados	16
2.5 Aprendizagem de Máquina.....	17
2.5.1 Clusterização.....	18
3 Materiais e Métodos	19
3.1 Python para Análise de Dados	19
3.2 Obtenção e Tratamento dos Dados	20
4 Resultado e Discussões	22
4.1 Faixa-etária	22
4.2 Frequência e desempenho por sexo	24
4.3 Região	31
4.4 Rede de Ensino	34
4.5 Raça Autodeclarada	38
4.6 Escolaridade dos pais ou responsáveis	40
4.7 Renda mensal	43
4.8 Acesso à Tecnologia	44
4.9 Clusterização.....	47
5 Conclusão	48
REFERÊNCIAS BIBLIOGRÁFICAS	50

1 Introdução

Com a tecnologia cada dia mais acessível e presente na sociedade, o estudo analítico dos dados gerados a partir da utilização dessas tecnologias torna-se uma demanda importante e oferece diversas possibilidades para o processo de desenvolvimento contínuo de pesquisa e inovação. A Ciência de Dados, principal ferramenta para produzir essas possibilidades, viabiliza a extração de informações relevantes para a composição de *insights*, construção de modelos de aprendizagem de máquina e análises exploratórias de dados que auxiliam na tomada de decisões dentro de uma organização.

No contexto educacional, a aplicação de técnicas, métodos e outras práticas educacionais auxiliam no ensino-aprendizagem do aluno e o resultado da eficiência é medida através de métricas pré-estabelecidas para avaliar o desempenho dos estudantes no tocante às metodologias utilizadas. O Exame Nacional do Ensino Médio (ENEM), além de ser requisito no ingresso à Educação Superior, tem o objetivo de mensurar o desempenho escolar dos estudantes na finalização da Educação Básica: o Ensino Médio. Tendo em vista uma média 6.805.012 de inscritos anualmente, o ENEM, bem como outros sistemas de avaliação da educação básica, é uma fonte importante de dados que permite entender o cenário educacional no Brasil.

Para este trabalho, foram utilizados os microdados¹ – conjunto de informações estruturadas – do ENEM disponibilizados pelo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) das edições de 2019, 2020, 2021 e 2022, sendo esta última a edição marcada pelo baixo número de notificações da COVID-19 (OMS, 2023), em comparação com os anos anteriores, e a mais recente com o objetivo de analisar o impacto da pandemia no desempenho dos participantes para, consequentemente, realizar um estudo comparativo entre as edições.

Na Seção 2, é apresentada a fundamentação teórica com os principais conceitos utilizados no presente trabalho. Na Seção 3 é abordada a metodologia usada para realizar o estudo. Na Seção 4 são apresentados os resultados e discussões, e na Seção 5, as conclusões.

¹ Disponível em: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem>
Acesso em: 12 jul, 2023.

2 Fundamentação Teórica

2.1 Fundamentos básicos da estatística

Para Neto (2002) e Cavalcanti (2006), a Estatística é uma parte da matemática aplicada e também uma ciência meio, que tem por objetivo extrair conclusões por meio da organização, análise, interpretação e apresentação das informações coletadas.

Dessa forma, é possível formalizar os resultados numéricos através de representações gráficas. A coleta, organização e a descrição dos dados fazem parte da Estatística Descritiva, ao passo que a análise e a interpretação desses dados são compostas pela Estatística Indutiva (Crespo, 2009). Mediante ao exposto neste trabalho, será discutido apenas o conceito de estatística descritiva.

2.1.2 Estatística Descritiva

A estatística descritiva é a parte da estatística que descreve e analisa um determinado grupo, sem tirar quaisquer conclusões deste, a partir de um todo ou uma amostra (Spiegel e Stephens, 2009). Uma vez possuindo os resultados observados, consoante a Neto (2002), o próximo passo é organizar as informações contidas nesse resultado, e, para isso, é preciso definir quais as variáveis que serão trabalhadas.

2.1.3 Variável, População e Amostra

Para Spiegel e Stephens (2009), a variável é um símbolo que assume qualquer um de um conjunto de valores atribuídos a ela. Ainda de acordo com Neto (2002), as variáveis podem ser qualitativas ou quantitativas. É qualitativa quando se tratar de uma classificação categórica que se diferencia por algum atributo não numérico, e quantitativa quando seus valores forem evidenciados por um número.

Um conjunto que possui características em comum, ou seja, que é ou faz parte de um objeto de estudo, é denominado população ou universo. Em muitos casos, é inviável a utilização de toda a população. Dessa forma, é possível usar apenas uma parte da população, a amostra (Crespo, 2009). Por exemplo, tem-se uma população composta por todos os alunos da rede pública de ensino, a amostra é o subconjunto deste objeto.

2.1.4 Distribuição e as Medidas Estatísticas

Ao se resumir um grande número de dados brutos – não numericamente organizados –, costuma-se frequentemente distribuí-los em classes ou categorias e determinar o número de indivíduos pertencentes a cada uma das classes. Uma distribuição mostra de certa forma como estão representados os valores em uma determinada área através de sua concentração (Spiegel e Stephens, 2009).

Na Figura 1, é possível ver um exemplo de como pode ser visualizada a representação gráfica da distribuição dos dados no Boxplot, uma das formas de retratar essa distribuição.

Geralmente, é preciso sintetizar as características das distribuições por meio das quantidades. E estas são denominadas de medidas da distribuição de frequências, pelo fato de quantificar aspectos importantes. Dentre essas medidas, destacam-se as medidas de posição constituídas pelas medidas de tendência central (Spiegel e Stephens, 2009).

2.1.5 Medidas de Tendência Central

Ainda, para Spiegel e Stephens (2009), as medidas de tendência central são assim chamadas pois os dados tendem a se agrupar em torno dos valores centrais. Dito isso, subdividem-se em:

- **Média:** quociente da divisão da soma dos valores da variável pelo número total deles;
- **Mediana:** valor que é situado no meio de um conjunto ordenado;
- **Moda:** valor que ocorre com maior frequência.
- **Separatrizes:** distribuição dos valores em partes iguais.

As separatrizes são medidas utilizadas para entender a distribuição da população (Cavalcanti, 2006). Engloba:

- **Mediana:** conforme explicado anteriormente;
- **Quartis:** divisão de um conjunto em quatro partes iguais.
- **Decis:** divisão de um conjunto em dez partes iguais;
- **Percentis:** divisão em cem partes iguais.

2.1.6 Representação Gráfica

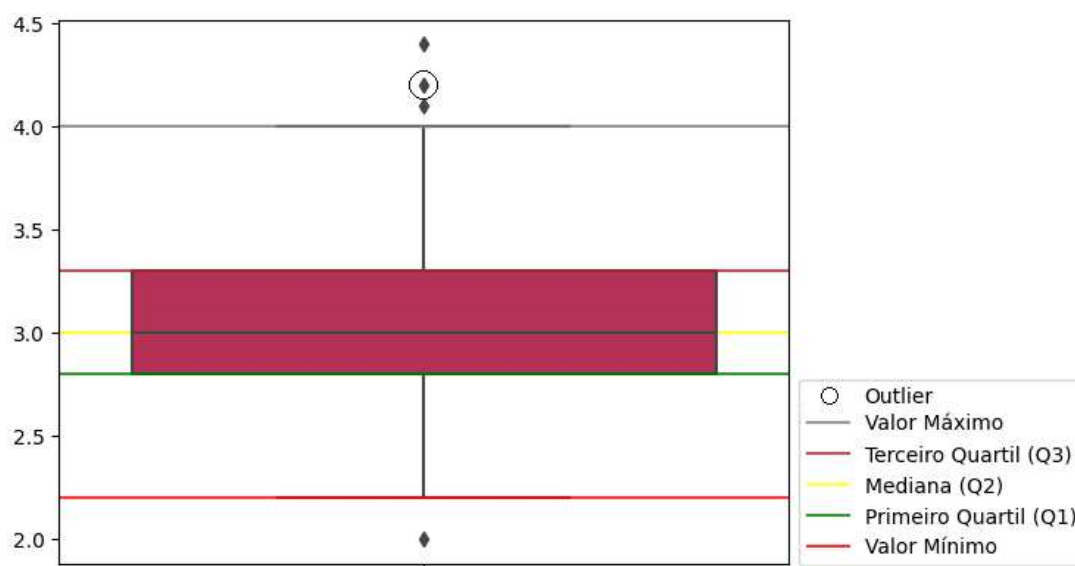
A última etapa da estatística descritiva é a que compreende a representação gráfica dos dados após coletados e organizados. Para Crespo (2009), o objetivo dessa representação é produzir para um terceiro, uma impressão mais rápida e viva do fenômeno em estudo, já que os gráficos falam mais rápido à compreensão. A representação gráfica de um fenômeno deve atender a alguns requisitos para ser de fato útil: simplicidade, clareza e veracidade.

Neste trabalho, serão apresentados alguns gráficos de fácil leitura, porém o diagrama de caixas (do inglês, boxplot) necessita de um conhecimento a mais por ter uma complexidade maior para, assim, melhor extrair informações.

2.1.7 Boxplot

O objetivo do Boxplot é representar a distribuição dos dados de uma determinada variável e identificar as seguintes estatísticas: valor mínimo e máximo dentro do intervalo proposto na representação; primeiro e terceiro quartis que significa, respectivamente, 25% e 75% dos dados do conjunto possuem o valor do eixo y; e a mediana ou, ainda, segundo quartil, que equivale a 50%. Há também a possibilidade da existência de *outliers* – pontos discrepantes – como apresentado na Figura 1.

Figura 1 – Exemplo de boxplot



Fonte: Elaborada pelo autor.

2.2 A pandemia da COVID-19

O termo pandemia está ligado à distribuição geográfica de uma determinada doença. A COVID-19 foi caracterizada em março de 2020 como pandemia pela Organização Mundial da Saúde (OMS, 2020).

A COVID-19 é uma doença infecciosa causada pela coronavírus SARS-CoV-2 e tem como principais sintomas febre, cansaço e tosse seca. Outros sintomas menos comuns e que podem afetar alguns pacientes são: perda de paladar ou olfato, congestão nasal, conjuntivite, dor de garganta, dor de cabeça, dores nos músculos ou juntas, diferentes tipos de erupção cutânea, náusea ou vômito, diarreia, calafrios ou tonturas (OMS, 2020).

Vários casos de pneumonia apareceram na cidade de Wuhan, província de Hubei, inserida na República Popular da China faltando apenas um dia para o ano de 2020. Os alertas tratavam-se de um novo tipo de coronavírus que ainda não havia sido identificado (OMS, 2020).

Em 30 de janeiro de 2020, a OMS expôs que o surto do novo coronavírus atingiu o nível elevado de alerta na Emergência de Saúde Pública de Importância Internacional (ESPII). Dentre as possibilidades de transmissão da doença, estaria o contato próximo entre indivíduos, e como medida de restrição para evitar o contágio foi adotado o *lockdown*, do inglês, confinamento (OMS, 2020).

Além do setor econômico, o sistema educacional também foi afetado, o que causou o fechamento de instituições de ensino. Para contornar a situação, a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO) recomendou a utilização de programas de Ensino à Distância (EaD) e plataformas educacionais abertas para que dessa forma fossem minimizados os efeitos da paralisação educacional (UNESCO, 2020; SOUTO, 2020).

2.3 O ENEM e a LGPD

O Exame Nacional do Ensino Médio (ENEM) é um exame anual que desde 2009 tem como finalidade aferir o desempenho dos estudantes ao final da educação básica, atribuindo-lhes uma pontuação para concorrer a vagas de ensino profissionalizante e superior em diversas instituições no Brasil, através do Sistema de Seleção Unificada (SISU) e do Programa Universidade para Todos (ProUni), e mais

de 50 instituições de ensino superior em Portugal. É possível ainda pleitear financiamento estudantil pelo Fundo de Financiamento Estudantil (FIES) (INEP, 2023).

Aplicada em dois domingos consecutivos, avaliação é composta por quatro áreas de conhecimento: linguagens, códigos e suas tecnologias; ciência humanas e suas tecnologias; ciências da natureza e suas tecnologias; e matemática e suas tecnologias que ao todo totalizam 180 questões objetivas, somadas a uma Redação de cunho dissertativo-argumentativo dado uma situação-problema (INEP, 2023).

Devido a pandemia, houve um atraso na aplicação do ENEM 2020, sendo este aplicado em janeiro de 2021. O ENEM 2021 aconteceu normalmente no mesmo ano (INEP, 2022).

Após a Lei 13.709/2018, Lei Geral de Proteção de Dados (LGPD), ter entrado em vigor em setembro de 2020, as bases de dados do INEP passaram por uma reformulação nos dados com o intuito de impedir o reconhecimento de indivíduos munido de tecnologias que permitem o acesso a essas informações. A LGPD assegura os direitos fundamentais de liberdade e privacidade de cada pessoa construindo, assim, um cenário de segurança jurídica, com a padronização de procedimentos que assegurem a proteção aos dados de todo cidadão brasileiro utilizando como base os parâmetros internacionais existentes (MPF, 2023).

2.4 Ciência de Dados

A Ciência de Dados é uma área interdisciplinar constituída pelo campo de estudo da Estatística e da Ciência da Computação. A princípio o termo ciência de dados foi proposto por Cleveland (2001), em seu trabalho, como sugestão da união de programação com a estatística, objetivando a facilitação das análises estatísticas mediante a um grande volume de dados. Sabe-se que, atualmente, esta ciência evoluiu e dessa forma compreende modelos, processos e procedimentos voltados ao ciclo de vida dos dados. Com isto, é possível extrair conhecimento e produzir novas informações através de diversas ferramentas que constituem esta área (Rautenberg e do Carmo, 2019).

Ao desenvolver um projeto de ciência de dados, é necessário se atentar às suas devidas particularidades. Para Gonçalves (2018), cada projeto possui suas especificidades e, dessa maneira, as etapas do desenvolvimento de um projeto de

dados é único para cada objetivo, porém segue um padrão de início, meio e fim. Este trabalho seguirá o passo a passo proposto por Gonçalves (2018):

- **Captar o problema:** a base para todo o projeto de pesquisa. É nesta etapa que são definidos o objeto de estudo, as variáveis que deve-se focar e as estratégias estabelecidas para extrair os dados;

- **Coleta de dados:** esta etapa compreende a definição da fonte dos dados, se são dados provenientes de banco de dados, planilhas (dados internos) ou dados retirados de bases públicas ou pagas (dados externos); do tipo de dados que compõem o objeto de estudo, no caso de dados estruturados (Tabelas bidimensionais) ou não estruturados (redes sociais, sites externos);

- **Procedimentos para tratamento e transformação dos dados:** ideal que os dados não tenham nenhuma inconsistência, dados duplicados, faltantes, inválidos, e que sejam de fácil leitura;

- **Análise exploratória dos dados:** utilização de ferramentas da Ciência de Dados para extrair informações e *insights* para compor o estudo em questão;

- **Análise profunda dos dados:** apesar de grande parte dos projetos pularem esta etapa, pelo fato de ter encontrado as soluções que satisfaçam os objetivos traçados na primeira etapa, ainda assim é possível agregar ainda mais às informações extraídas, aplicando técnicas de aprendizagem de máquina (*machine learning*);

- **Feedback:** apresentação das informações encontradas.

A contexto das etapas apresentadas acima, este trabalho tomará como objeto de estudo os dados públicos externos e estruturados das bases de dados do ENEM 2019, 2020, 2021, 2022.

2.5 Aprendizagem de Máquina

A aprendizagem de máquina, ou *Machine Learning*, tornou-se uma das ferramentas mais utilizadas da inteligência artificial, onde é possível a resolução de problemas complexos e/ou a manipulação de grande volume de dados através da execução de tarefas mediante a experiência adquirida do histórico de execuções da máquina. Para Lenz et.al, pode-se subdividir este ramo em quatro abordagens de aprendizado: supervisionado, não supervisionado, semissupervisionado e por reforço (Lenz et al. apud Mitchell, 2020).

No aprendizado supervisionado, existe uma classificação prévia da rotulação dos dados para que dessa forma a máquina realize a predição de novos dados de entrada. O aprendizado não supervisionado identifica dados que apresentam correlações entre si ou não, e separa as instâncias em grupos diferentes (Lenz et al. apud FACELI et al., 2020).

Consoante a Witten e Frank, o aprendizado semissupervisionado é o caso híbrido que ocorre quando o algoritmo pertencente a essa abordagem realiza tarefas de classificação, entretanto, aplicado a dados não rotulados completamente. E o aprendizado por reforço, para Russel e Norvig, equipara-se a uma espécie de adestramento, onde os parâmetros de um determinado modelo são reforçados de acordo com os resultados adquiridos e então são aplicadas métricas para definir o grau de qualidade das decisões tomadas e, por fim, decidir se o modelo será recompensado ou punido (Lenz et al. apud Witten e Frank; Russel e Norvig, 2020).

O presente trabalho adota a abordagem de aprendizado não supervisionado para segmentar a pontuação de desempenho dos participantes do ENEM 2019 a 2022. Para isso, será utilizado a técnica de clusterização e o algoritmo K-médias.

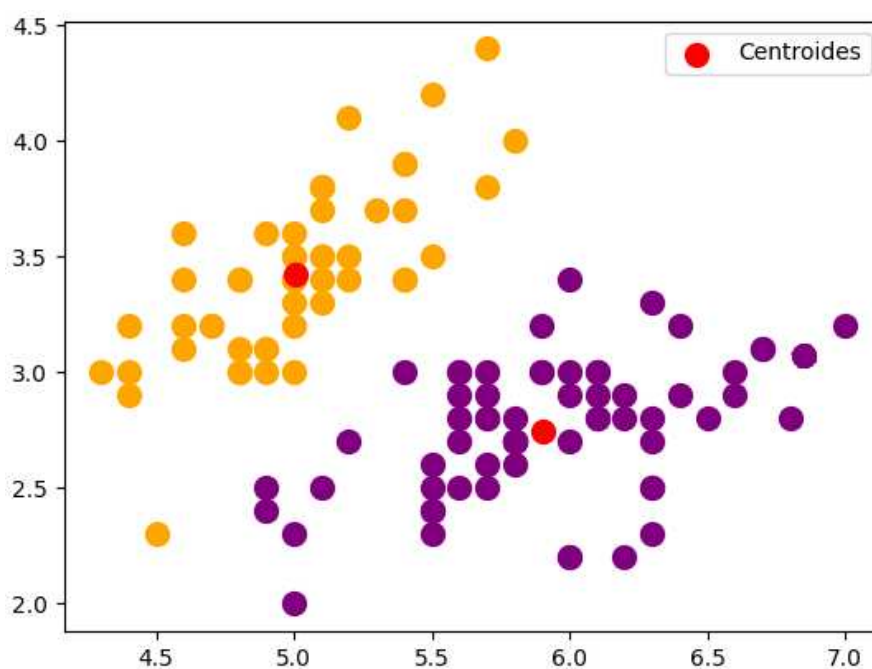
2.5.1 Clusterização

A clusterização – agrupamento (do inglês, *clustering*) – é uma técnica de aprendizado não supervisionado que consiste na separação dos dados em grupos (em inglês, *clusters*) os quais possuem valores próximos entre os atributos de entrada. A maneira como ocorre a divisão desses grupos varia de acordo com o algoritmo utilizado (Lenz et. al, 2020).

2.5.2 Algoritmo K-médias

O K-médias (em inglês, *K-means*) é um algoritmo de aprendizado não supervisionado, que utiliza de medidas de centro, ou centroides, como ponto de referência para separar em grupos, assim definido na figura 2. A divisão do conjunto de dados em *clusters* (do inglês, grupos) é realizada por meio de métodos que definem a quantidade precisa de grupos. (Lenz et. al, 2020).

Figura 2 – Exemplo da aplicação do algoritmo de K-médias



Fonte: Elaborado pelo autor.

Após a definição da quantidade ideal de grupos, a classificação desses dados para cada K grupo é feita comparando a distância entre cada valor (Lenz et. al, 2020).

3 Materiais e Métodos

Esta seção discorre das ferramentas utilizadas neste trabalho que perpassa desde a origem e obtenção dos dados, processos de tratamento de cada base de dados até a análise e extração de informações para compor o estudo idealizado.

3.1 Python para Análise de Dados

A linguagem de programação de alto nível e código aberto, o *Python*, possui diversas facetas e utilidades, o que a torna para uso geral, tanto na implementação de softwares com diferentes graus de complexidades, como para o foco principal deste trabalho, a análise exploratória.

Para este propósito, o *Python* dispõe de algumas bibliotecas utilizadas neste trabalho foram: *Pandas*, para leitura, processamento e manipulação dos dados; *Matplotlib*² e *Seaborn*³, geração e manipulação de gráficos; *Numpy*⁴, apoio na

² Disponível em: <<https://matplotlib.org>> Acesso em: 12 jul, 2023.

³ Disponível em: <<https://seaborn.pydata.org>> Acesso em: 12 jul, 2023.

⁴ Disponível em: <<https://numpy.org>> Acesso em: 12 jul, 2023.

construção de cálculos numéricos; *Sklearn*⁵, utilização dos modelos de *Machine Learning*.

O *Google Colab*⁶ foi a plataforma utilizada para desenvolver todo o estudo, nele é oferecido um ambiente prático e interativo, possibilitando o uso de recursos de RAM de até 12.7 GB para o plano gratuito.

3.2 Obtenção e Tratamento dos Dados

Os dados são provenientes da base de dados do INEP. Foram coletados os dados dos questionários socioeconômicos das edições do ENEM de 2019, 2020, 2021 e 2022, a partir do arquivo de microdados com formato .csv, para realizar o presente estudo. A somatória do tamanho de armazenamento de todas as edições resultou em 7,51 GB, o que representa um total de 17.744.217 inscritos. O total de colunas para cada edição é de 76 colunas, para realizar a leitura de todos esses dados, foi necessário utilizar a biblioteca *zipfile* do *Python* para selecionar as colunas utilizadas. Após este processo, foi criado um outro *DataFrame* para comportar todos esses dados.

Um requisito para a realização do estudo foi a presença do participante nos dois dias de aplicação do exame, sendo assim não foram considerados os dados dos faltantes dessas edições. Com isto, foi possível reduzir consideravelmente o tamanho do conjunto inicial dos dados. A título de informação, restaram 10.873.521 inscritos na base única composta pelas quatro edições sob estudo, um total de 5,4GB de armazenamento.

Na Tabela 1, estão apresentados os dados das colunas utilizadas no estudo.

Tabela 1 – Descrição das colunas utilizadas

Nome da Variável	Descrição
NU_ANO	Ano do Enem
TP_FAIXA_ETARIA	Faixa etária
TP_SEXO	Sexo
TP_COR_RACA	Cor/raça autodeclarada
TP_ESCOLA	Tipo de escola do Ensino Médio

⁵ Disponível em: <<https://scikit-learn.org/stable>> Acesso em: 12 jul, 2023.

⁶ Disponível em: <<https://colab.google>> Acesso em: 12 jul, 2023.

SG_UF_PROVA	Sigla da Unidade da Federação da escola
NU_NOTA_CN	Nota da prova de Ciências da Natureza
NU_NOTA_CH	Nota da prova de Ciências Humanas
NU_NOTA_LC	Nota da prova de Linguagens e Códigos
NU_NOTA_MT	Nota da prova de Matemática
NU_NOTA_REDACAO	Nota da prova de Redação
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q006	Qual é a renda mensal de sua família (incluindo o candidato)?
Q022	Na sua residência tem telefone celular?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

Fonte: Elaborada pelo autor com base em De Souza, 2021.

Após a seleção das colunas, foi inserida uma nova coluna denominada “Região” de acordo com cada UF do participante. A base geral contendo os dados das edições mencionadas será nomeada como Base 1.

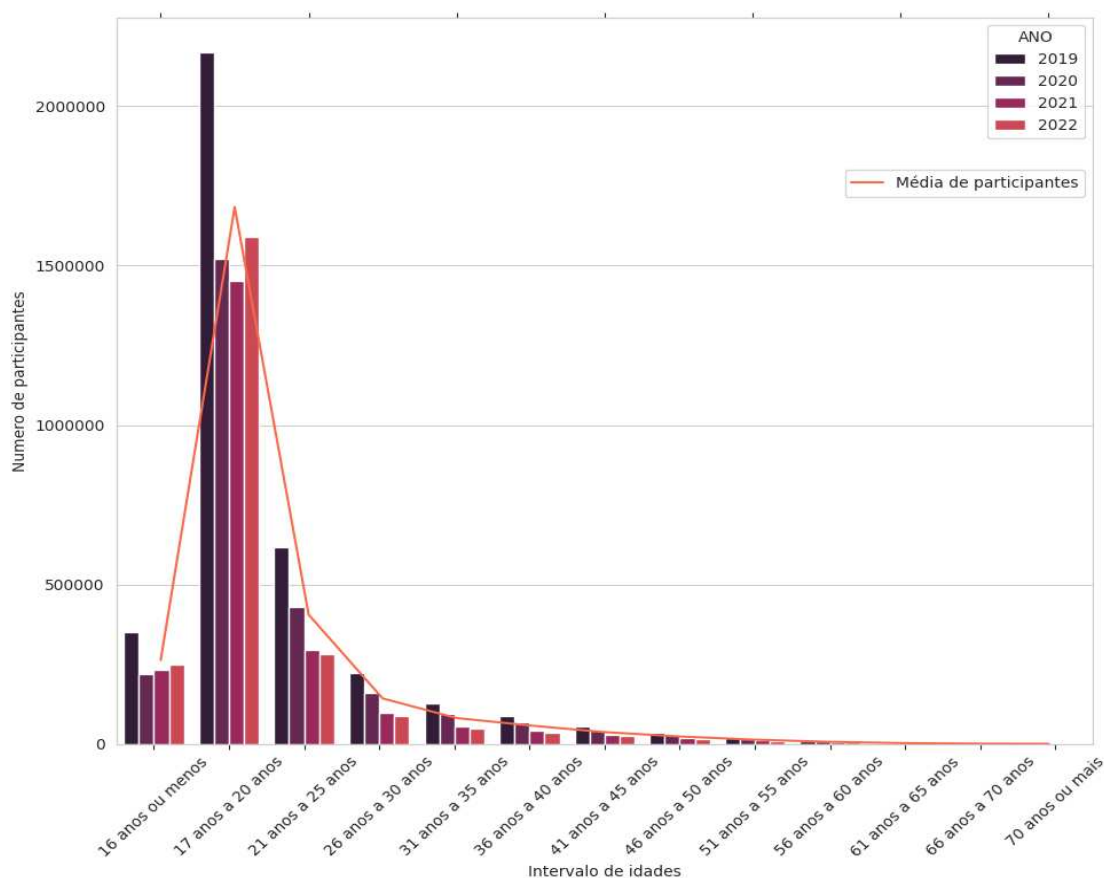
4 Resultado e Discussões

Nesta seção, será apresentado o resultado da análise de dados utilizando os dados das edições de 2019 a 2022 do ENEM.

4.1 Faixa-etária

Como visto, o conjunto de dados do ENEM, disponibilizado pelo INEP, passou por uma reestruturação para impedir a identificação dos participantes de acordo com as normas previstas na Lei nº 13.709, de 14 de agosto de 2018 – a LGPD. Dessa forma, a coluna **TP_FAIXA_ETARIA** foi reformulada e é representada não completamente pela idade, mas também por intervalos. A partir dos 17 anos até os 25 anos é possível constatar a idade. Sendo assim, para tornar uniforme aos demais intervalos dispostos na base de dados do ENEM, foi dividido em mais dois intervalos que compreende dos 17 anos aos 20 anos e dos 21 anos aos 25 anos como mostrado na Figura 3.

Figura 3 – Número de participantes por intervalo de idade e ano de aplicação

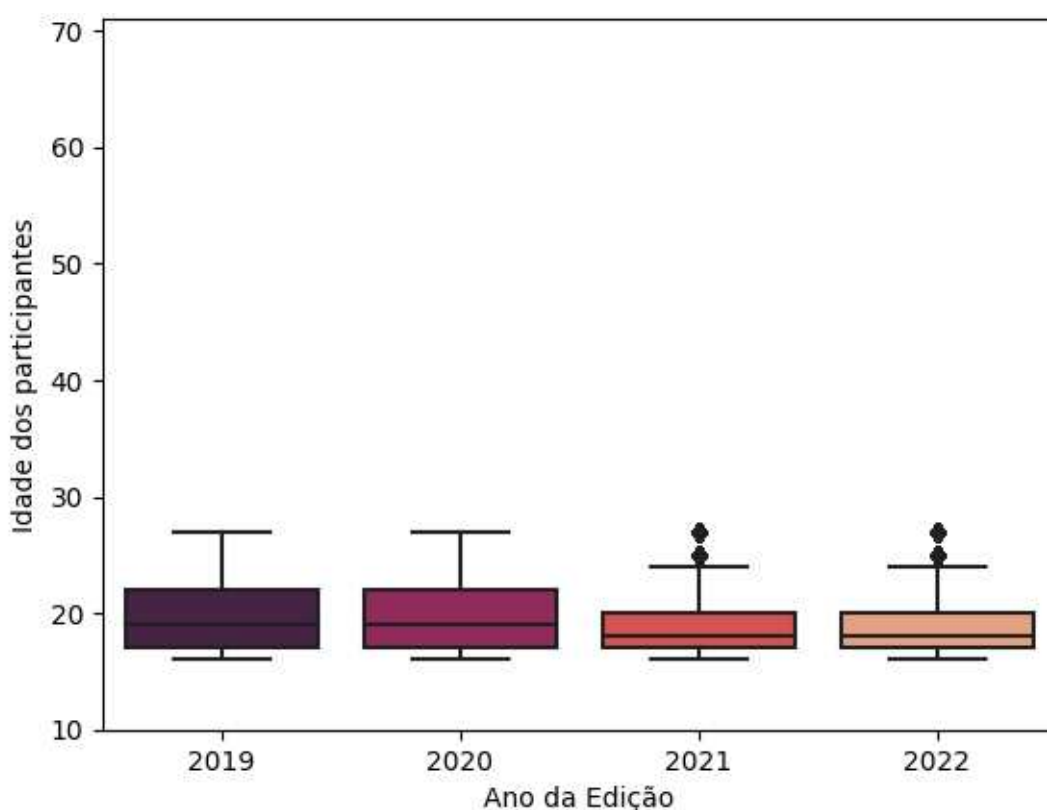


Fonte: Elaborado pelo autor.

É perceptível que o intervalo que compreende dos 17 anos aos 20 anos possui a maior quantidade de participantes de todas as edições, representando exatamente 61,94% do total de candidatos. A média desse intervalo é de 1.683.815 participantes.

No intervalo de 16 anos ou menos, apesar da queda no número de candidatos de 2019 para 2020, nota-se um aumento gradativo da edição de 2020 a 2022, mostrando uma certa recuperação também vista no intervalo seguinte a partir de 2022, um aumento de 4,45% em comparação com a edição de 2020. Este aumento entre as edições não é visto nos demais intervalos, porém nota-se que quanto maior a idade do participante, menor é a frequência por intervalo. No boxplot da Figura 4, pode-se atestar a distribuição das idades sem a aplicação dos intervalos dos 17 anos aos 25 anos.

Figura 4 – Boxplot da faixa-etária dos participantes por edição do ENEM



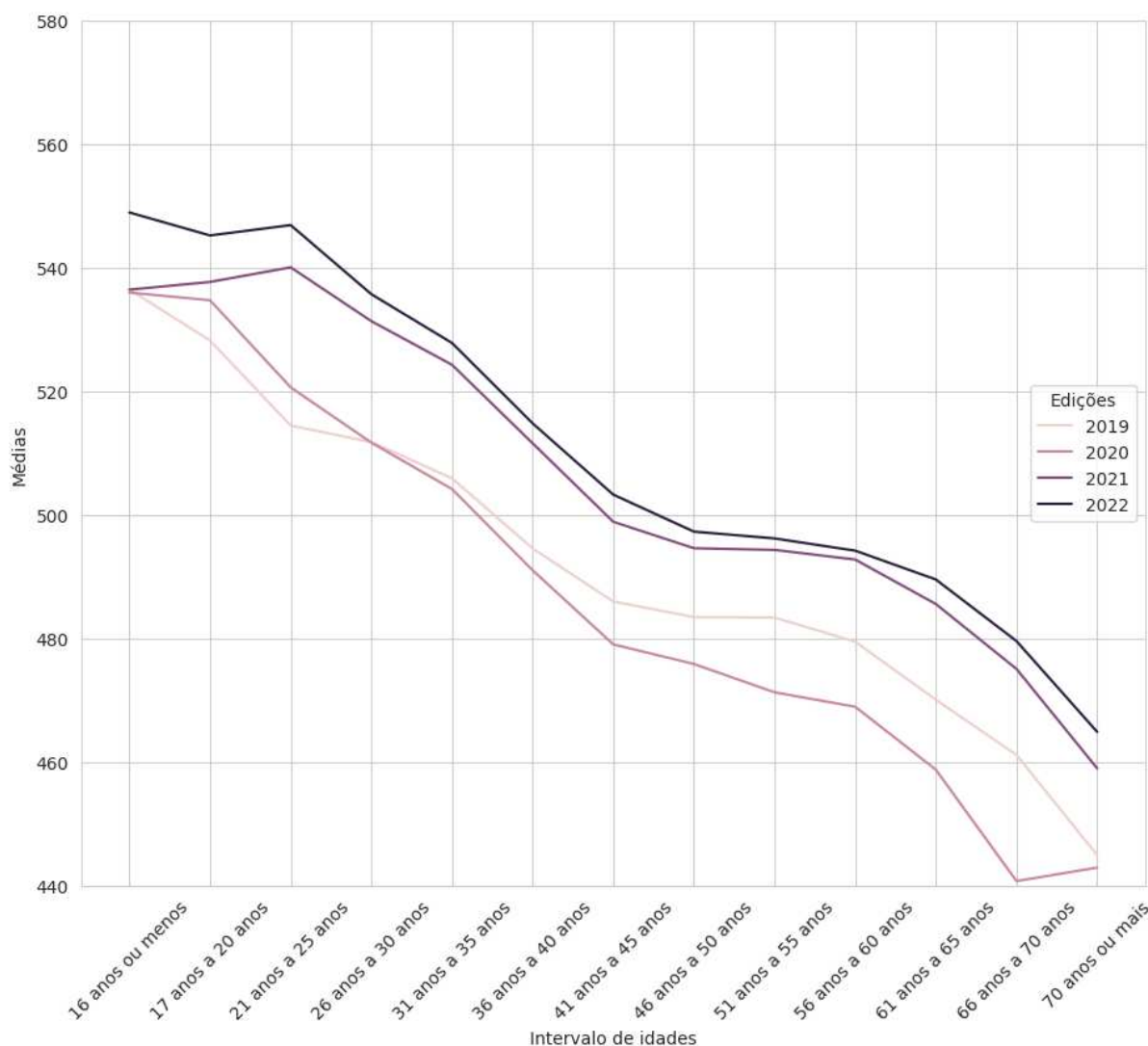
Fonte: Elaborado pelo autor.

A mediana, ou seja, 50% das idades nas edições de 2019 e 2020 foram de 19 anos, enquanto 2021 e 2022, baixou para 18 anos.

Na Figura 5, é observável que o desempenho na edição do ENEM de 2022 foi maior em comparação às edições anteriores. Em contrapartida da edição de 2020,

que foi aplicado em janeiro de 2021 em razão da pandemia da COVID-19, onde mostra uma queda na pontuação, no qual as médias, a partir do intervalo de 26 anos a 30 anos, ficaram abaixo da edição de 2019, antes da pandemia. Em geral, destaca-se ainda que quanto maior a idade do participante, menor é o seu desempenho.

Figura 5 – Médias das cinco disciplinas por intervalo de idade e edição do ENEM

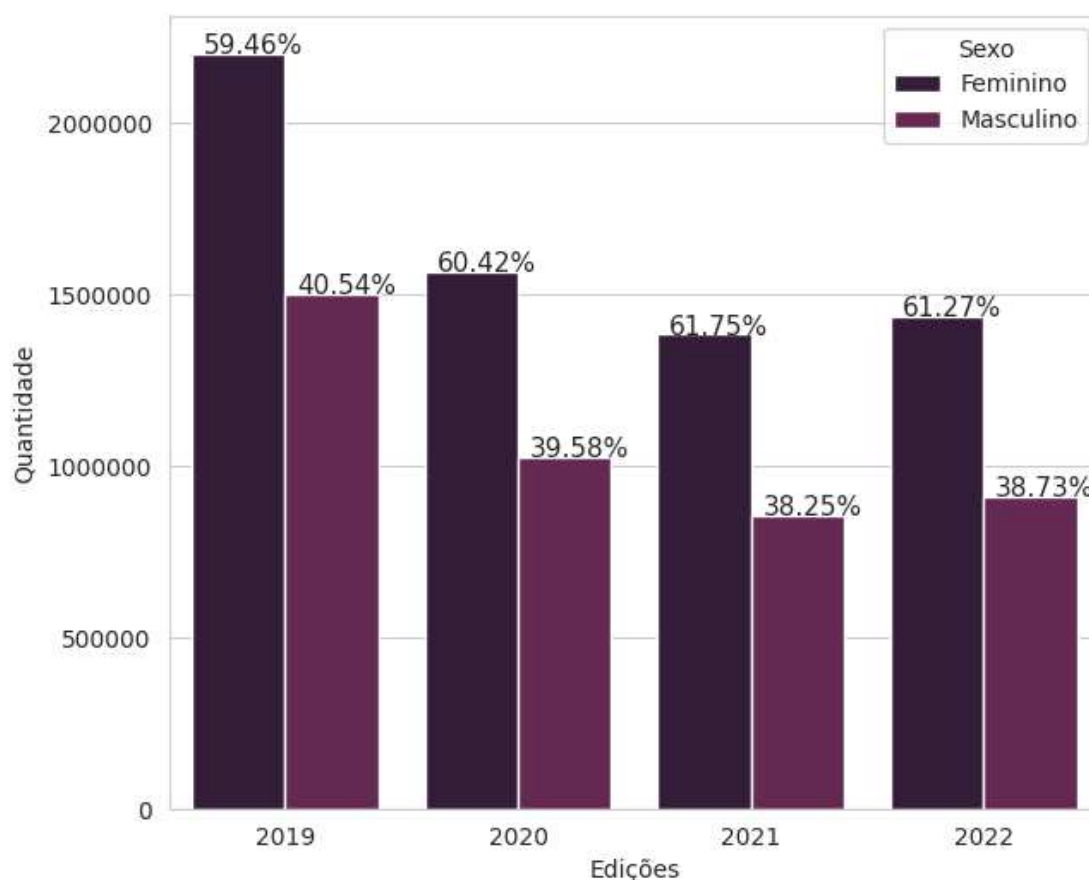


Fonte: Elaborado pelo autor.

4.2 Frequência e desempenho por sexo

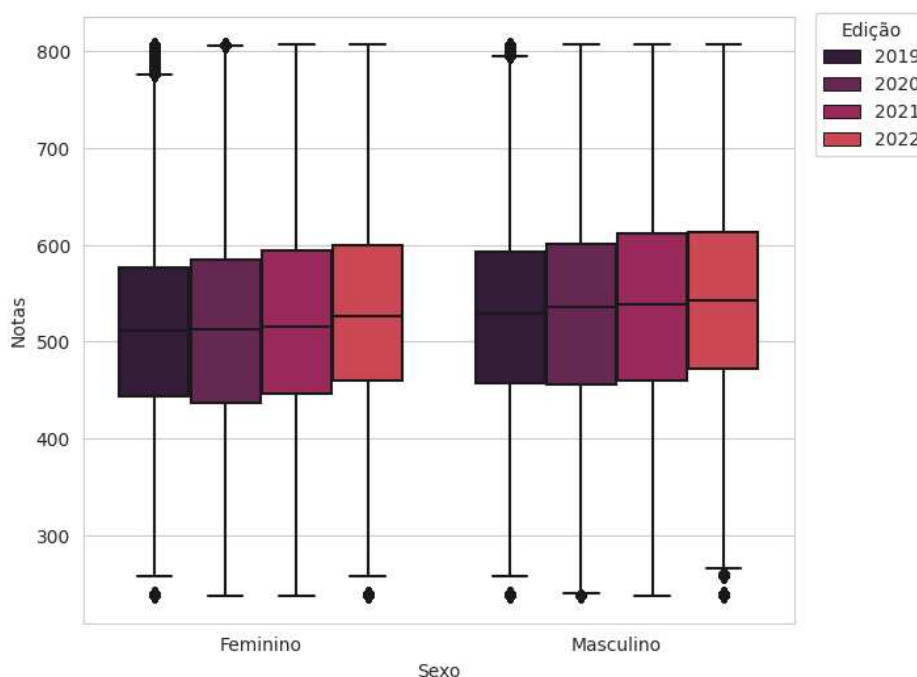
A diferença entre a frequência das pessoas que se identificam como feminino para o masculino de acordo com o total de cada edição é contrastante como mostrado na Figura 6. A edição de 2021 foi que teve a maior predominância de pessoas do sexo feminino, enquanto a edição de 2019, maior quantidade de pessoas do sexo masculino em comparação com as demais edições.

Figura 6 – Distribuição dos participantes por sexo e edição do ENEM



Fonte: Elaborado pelo autor.

Em relação ao desempenho médio dos participantes por sexo, as pessoas do sexo masculino possuem uma média um pouco mais elevada que o sexo feminino como é mostrado na Figura 7. É notório, para ambos os sexos, que a média de desempenho cresce ao longo das edições e que 50% dos participantes adquiriram uma pontuação acima de 500 pontos.

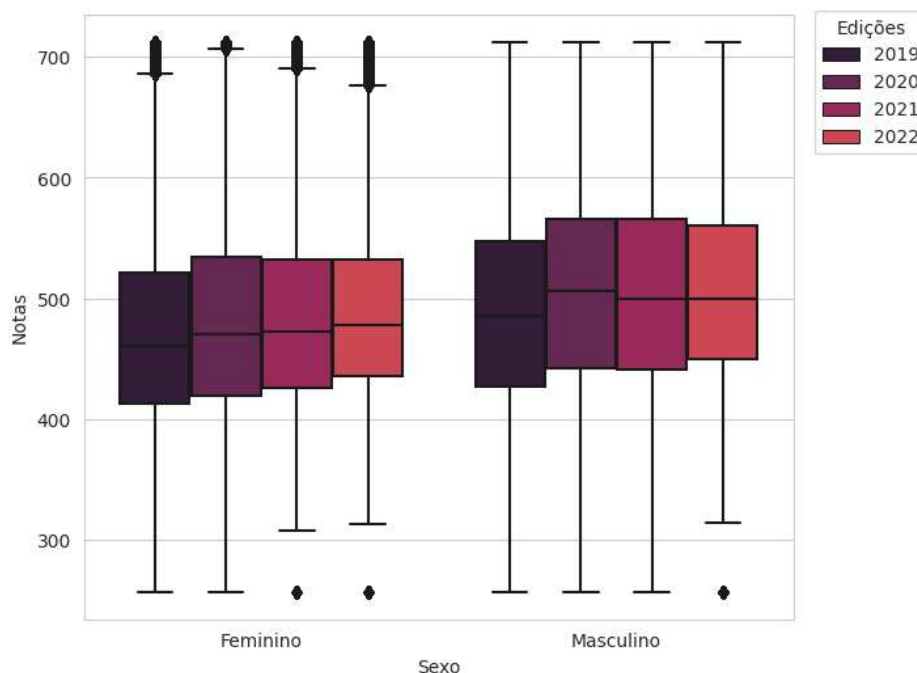
Figura 7 – Desempenho dos participantes por sexo e edição do ENEM

Fonte: Elaborado pelo autor.

Haja vista a possibilidade de adquirir de 0 a 1000 na prova de redação, esse intervalo é divergente para cada disciplina, pois o cálculo de desempenho é de acordo com a dificuldade das questões selecionadas em cada edição. Dessa forma, foi averiguado a distribuição do desempenho dos participantes em cada disciplina.

Na Figura 8, verifica-se que as notas do sexo feminino na disciplina Ciências da Natureza apresentam um crescimento na mediana ao longo das provas, isso significa que 50% dos participantes em cada edição teve um aumento gradativo em seu desempenho. Já os participantes do sexo masculino, possuem uma maior variabilidade na pontuação. Por meio do boxplot percebe-se ainda que 75% dos candidatos do sexo masculino nas edições de 2020 e 2021 apresentaram uma mediana de aproximadamente 566 pontos nesta disciplina.

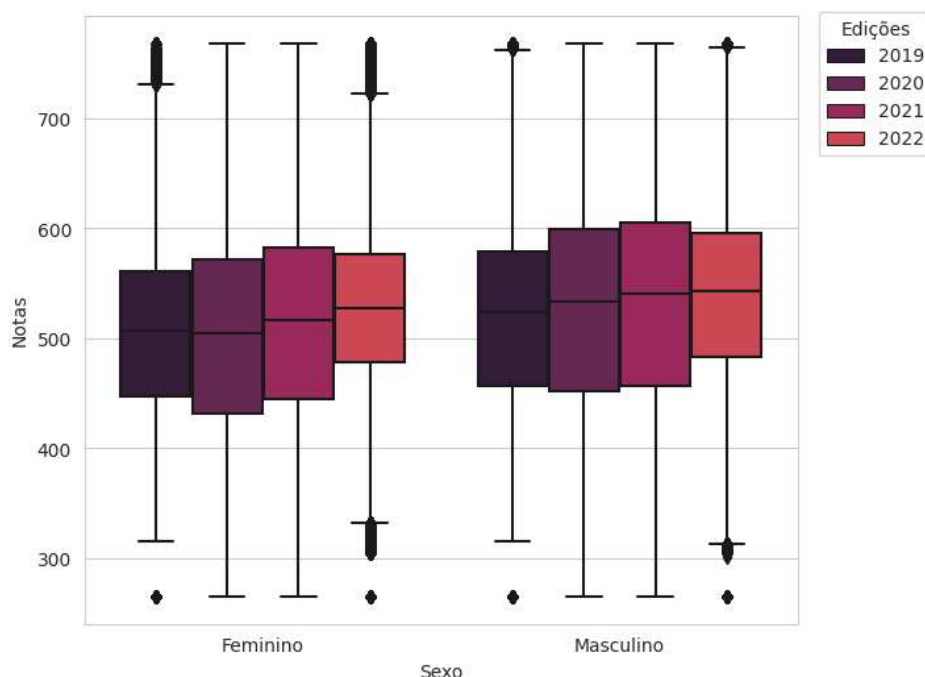
Figura 8 – Desempenho dos participantes na disciplina de Ciências da Natureza de acordo com o sexo e edição do ENEM



Fonte: Elaborado pelo autor.

Na Figura 9, identifica-se uma semelhança no desempenho da disciplina de Ciências Humanas, onde 25% dos candidatos de ambos o sexo, possuem uma mediana similar. A maior diferença das medianas entre os sexos dos participantes é de 29 pontos na edição de 2020, porém esse intervalo diminuiu. Na edição de 2022, a diferença foi de 18,7 pontos, apenas 1.1 pontos a menos que a diferença registrada na edição de 2019, antes da pandemia da COVID-19.

Figura 9 – Desempenho dos participantes na disciplina de Ciências Humanas de acordo com o sexo e edição do ENEM

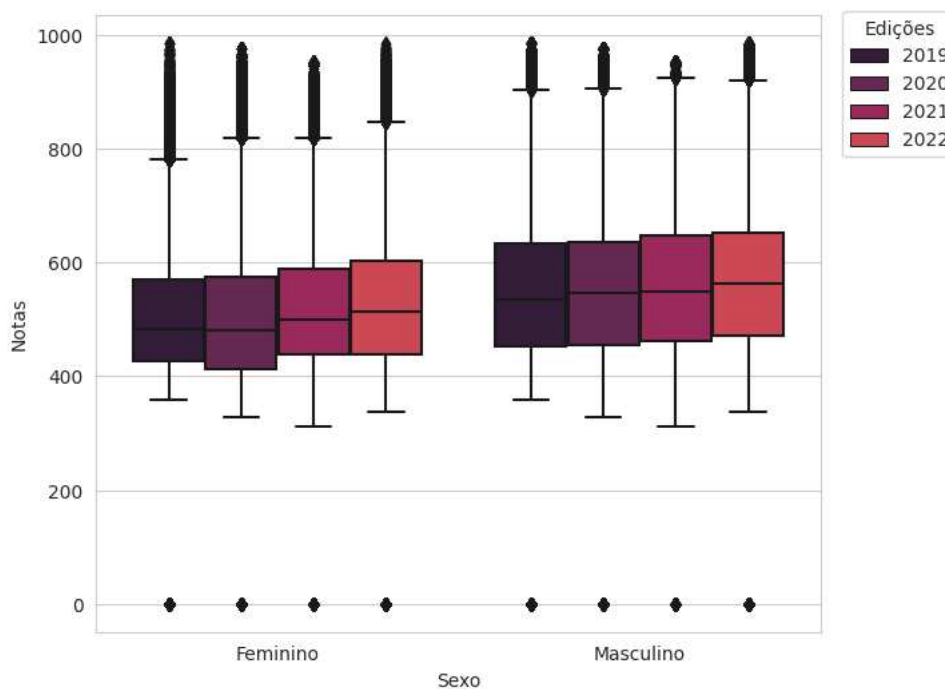


Fonte: Elaborado pelo autor.

A distribuição das notas na disciplina de Matemática e suas tecnologias também apresenta semelhança no que se refere ao desempenho dos participantes como mostrado na Figura 10. Além do mais, 75% dos candidatos do sexo feminino tiveram uma melhora de 5,45% ao longo das edições, já os do sexo masculino, obtiveram 2,90%.

Além disso, ainda na Figura 10, há uma grande disposição de *outliers* – valores discrepantes – no decurso das edições. A razão para este fenômeno se dá pelo grande desvio padrão das notas da disciplina em cada edição. Na edição de 2019 para o sexo feminino, por exemplo, o desvio padrão registrado é de 99.94, isso significa que existe uma grande variação de notas nesse conjunto, pois quanto maior o desvio padrão, maior a variabilidade em um determinado conjunto.

Figura 10 – Desempenho dos participantes na disciplina de Matemática de acordo com o sexo e edição do ENEM

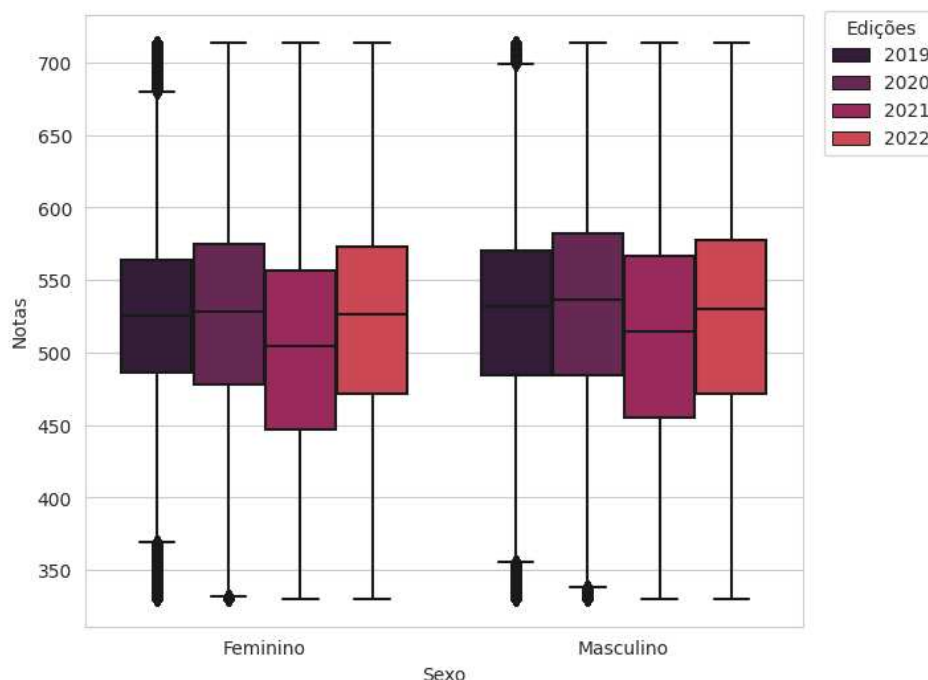


Fonte: Elaborado pelo autor.

Analisando a Figura 11, observa-se uma semelhança no desempenho dos participantes, independente do sexo, na disciplina de Linguagens e Códigos com 50% dos candidatos obtendo um desempenho entre 500 e 550 pontos. Cerca de 25% dos candidatos (primeiro quartil) obtiveram um menor desempenho na edição de 2021 onde foi registrada uma pontuação abaixo de 456 pontos.

Ainda na Figura 11, é perceptível uma diminuição no desempenho para ambos os sexos na edição de 2021. De 2019 a 2021, para 25% dos candidatos do sexo feminino houve uma queda de -7,94%, enquanto que o sexo masculino apresentou queda de -6,15%.

Figura 11 – Desempenho dos participantes na disciplina de Linguagens e Códigos de acordo com o sexo e edição do ENEM

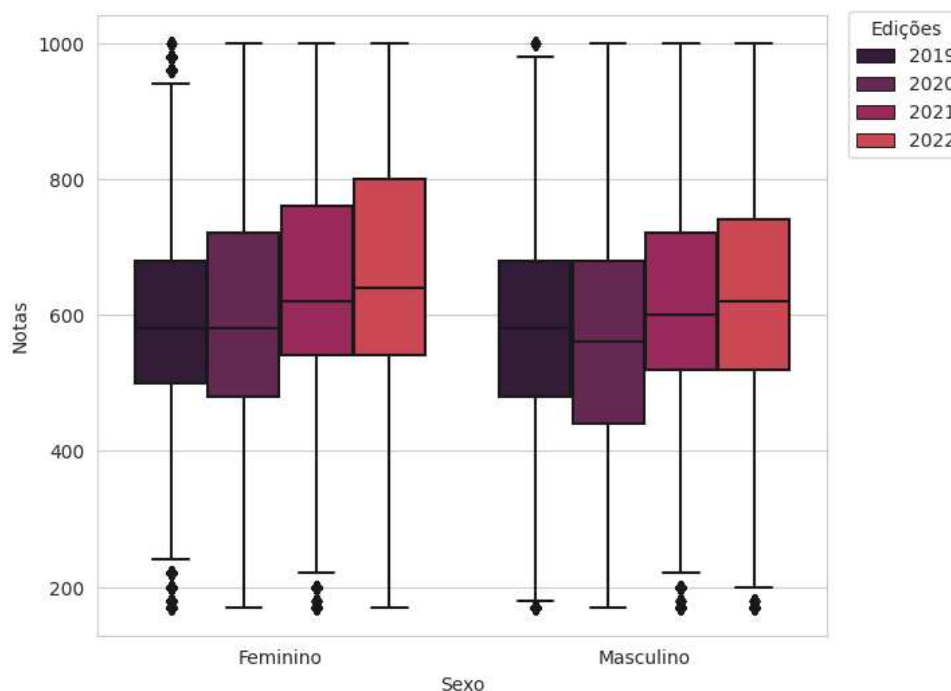


Fonte: Elaborado pelo autor.

Como citado anteriormente, a redação é a única possibilidade do participante conseguir 1000 pontos. No decorrer das edições de 2019 a 2022, apenas 131 candidatos alcançaram a pontuação máxima. Dentre estes, 35 candidatos eram do sexo masculino e 96 do sexo feminino. A Figura 12 demonstra que os candidatos do sexo feminino possuem um desempenho maior comparado aos do sexo masculino.

Houve um aumento de 15% na nota da redação nas edições de 2019 para 2022 em relação a 75% do público feminino. Haja vista que na última edição do ENEM realizado, 2022, essa percentagem dos candidatos do sexo feminino obtiveram uma pontuação de 800 pontos, a maior até então. Enquanto, os candidatos do sexo masculino, neste intervalo, obtiveram aumento de 8,11%. A redação é a única prova em que os participantes do sexo feminino tiveram o desempenho maior comparado aos participantes do sexo masculino em outras disciplinas.

Figura 12 – Desempenho dos participantes na Redação de acordo com o sexo e edição do ENEM



Fonte: Elaborado pelo autor.

4.3 Região

Segundo os primeiros resultados do Instituto Brasileiro de Geografia e Estatística (IBGE), o Brasil possui, aproximadamente, 203 milhões de habitantes. A região com maior população é a Sudeste, apresentando 41,78% do total. Em seguida, as regiões: Nordeste (26,91%); Sul (14,74%); Norte (8,54%); e Centro-Oeste (8,02%).

De acordo com as Figuras 13 e 14, que mostram a distribuição dos participantes por região e edição do ENEM, a maior predominância de participantes são das regiões Sudeste e Nordeste. A região Centro-Oeste tendo o menor índice populacional, também possui a menor distribuição no exame.

Na Tabela 2, é apresentada a percentagem da distribuição do número de óbitos causados pela COVID-19, que de acordo com Brasil (2023), ao todo, durante os anos de 2020 a 2022, foram 693.853 mortes.

Tabela 2 - Percentagem do número de óbitos causados pela COVID-19 distribuídos por região no Brasil

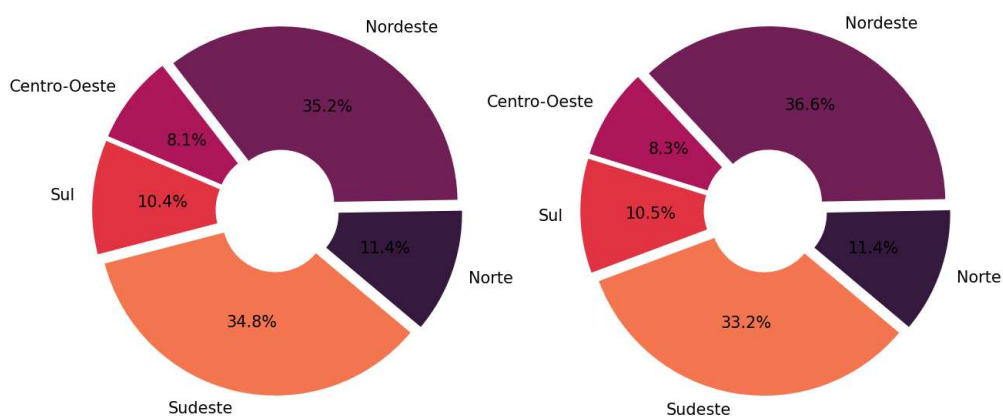
Região	Ano		
	2020	2021	2022
Sudeste	12,86%	29,61%	5,57%
Nordeste	6,88%	10,42%	1,99%

Sul	3,18%	10,87%	1,77%
Norte	2,60%	4,25%	0,55%
Centro-Oeste	2,57%	5,98%	0,90%

Fonte: Elaborado pelo autor.

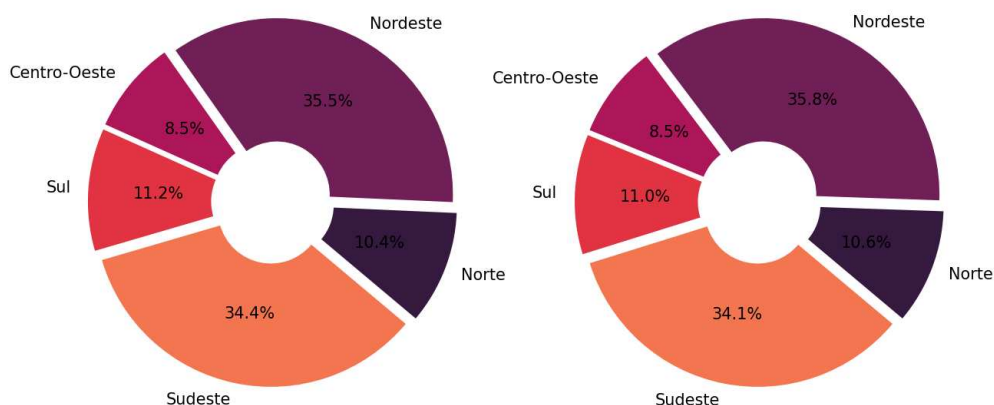
Apesar do aumento no número de óbitos de 2020 a 2021 em decorrência da pandemia da COVID-19 como mostrado na Tabela 2, não houve mudança discrepante na distribuição por região entre as percentagens da edição do ENEM 2020, aplicado em janeiro de 2021, e o ENEM 2021, realizado em novembro do mesmo ano.

Figura 13 – Distribuição dos participantes por Região nas edições de 2019 e 2020 do ENEM



Fonte: Elaborado pelo autor.

Figura 14 – Distribuição dos participantes por Região nas edições 2021 e 2022

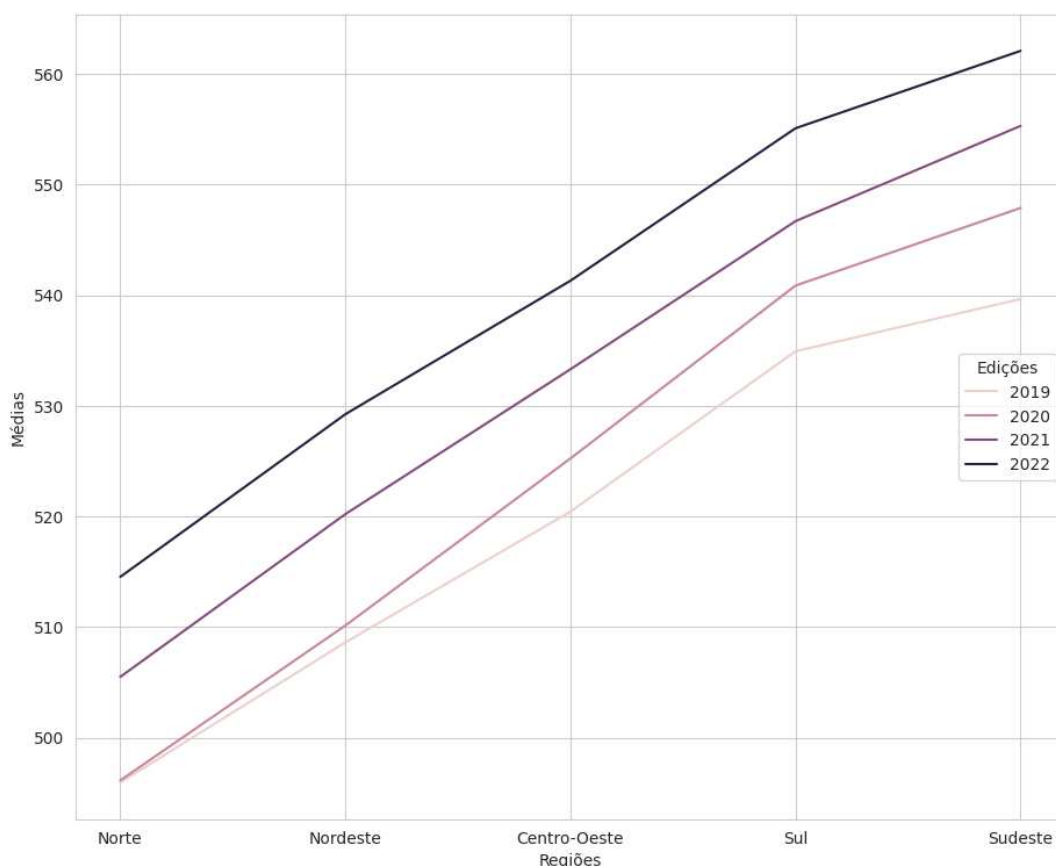


Fonte: Elaborado pelo autor.

Ainda, na Figura 13, se tratando da comparação distributiva do ano de 2019 para o de 2022, presente na Figura 14, as regiões que apresentaram um aumento foram apenas Centro-Oeste, Nordeste e Sul. As demais regiões, Sudeste e Norte, diminuiram 0,7% e 0,8%, respectivamente.

Em relação à média dos participantes mostrada na Figura 15, a região Norte, apesar de ter um aumento de 3,61% de 2019 a 2022, detém a menor média, dentre as regiões, em todas as edições do ENEM.

Figura 15 – Médias dos participantes por região e ano de edição do ENEM



Fonte: Elaborado pelo autor.

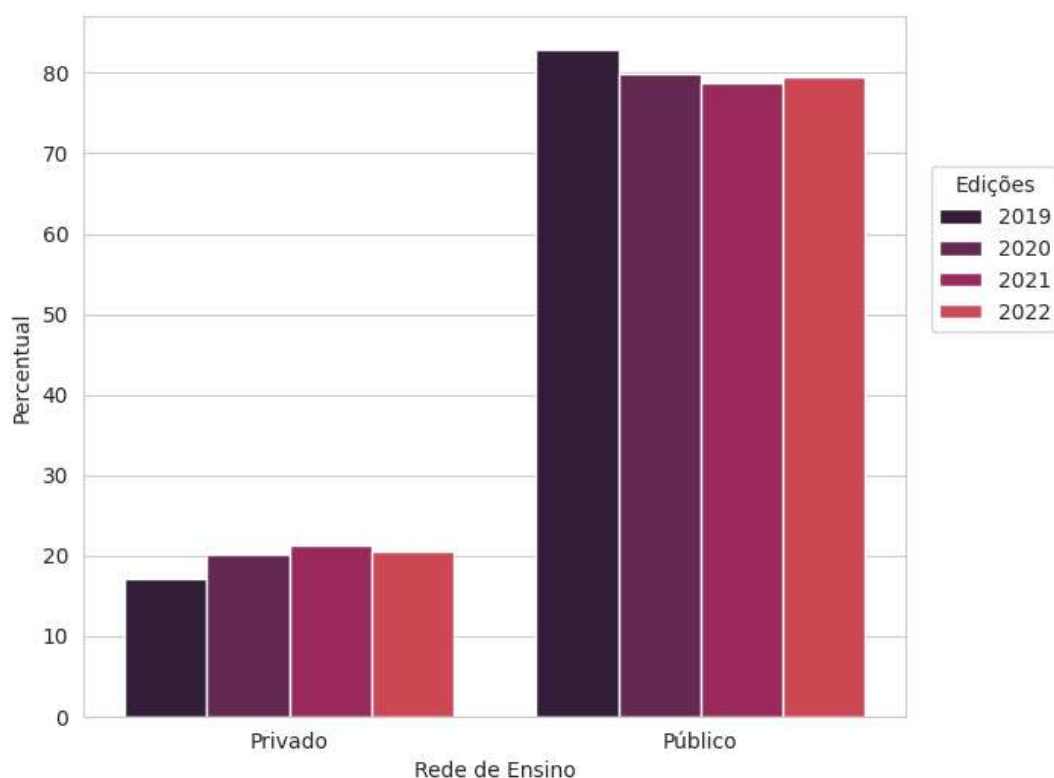
A região Sudeste possui as maiores médias em todas as edições. E em cada aplicação do exame, nota-se uma diferença média de 8,74% da região com o menor desempenho em comparação com a maior. Todavia, é visto que há um crescimento no decorrer das edições do ENEM.

4.4 Rede de Ensino

Nesta seção, foram excluídos dados ausentes de 7.037.269 participantes. Dessa forma, apenas 3.836.252 candidatos foram utilizados para o estudo.

Como mostrado na Figura 16, há uma disparidade no número de participantes da rede de ensino pública.

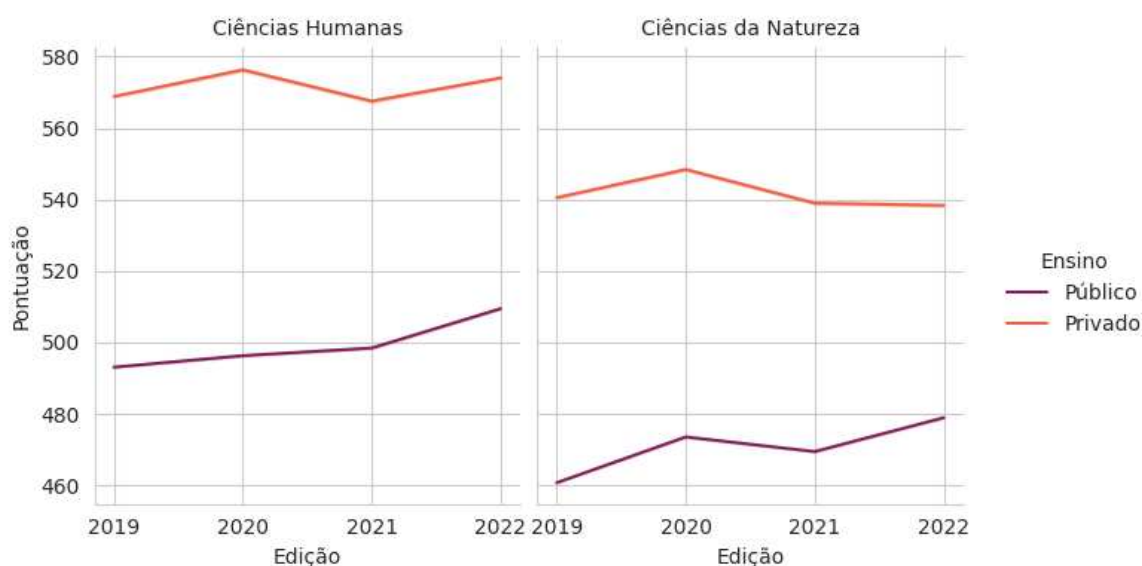
Figura 16 – Percentual dos participantes por rede de ensino e edição do ENEM



Fonte: Elaborado pelo autor.

Apesar do percentual discrepante de participantes por rede de ensino, o desempenho nas disciplinas retrata um cenário diferente. O ensino privado possui as melhores médias no exame em todas as edições como apresentado nas Figuras 17, 18 e 19.

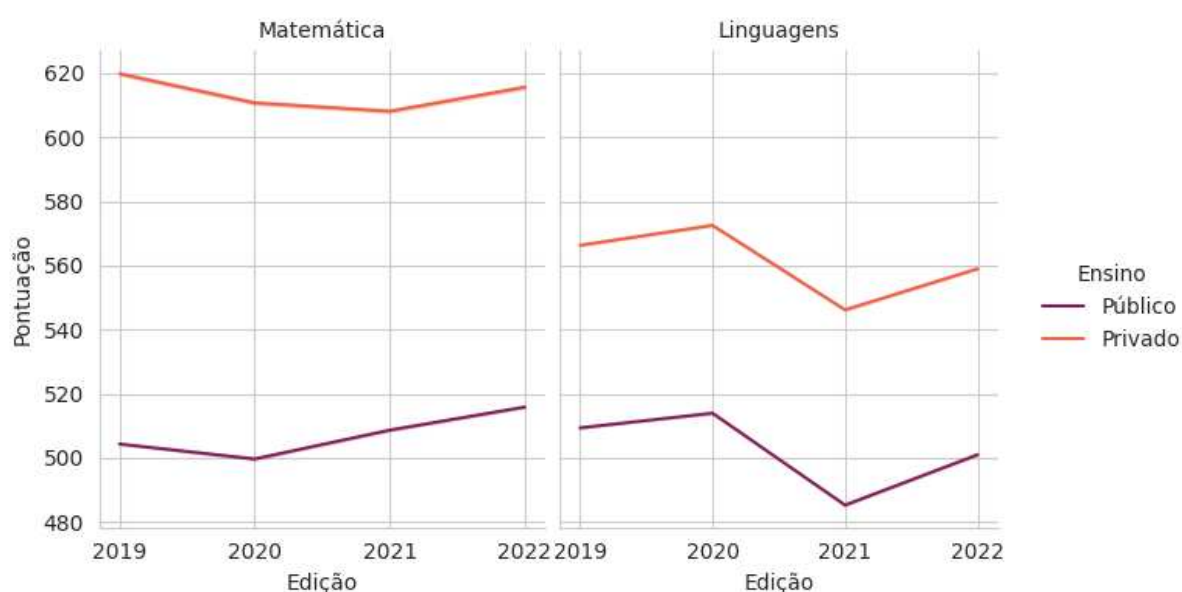
Figura 17 – Média dos participantes nas disciplinas de Ciências Humanas e da Natureza por rede de ensino e edição do ENEM



Fonte: Elaborado pelo autor.

Na Figura 17 é possível ver um aumento de 3,21% e 3,80% no desempenho por parte dos candidatos de escola pública nas disciplinas de Ciências Humanas e Ciências da Natureza, respectivamente. Na Figura 18, a diferença entre as notas de Matemática é ainda maior que as disciplinas de Ciências Humanas e da Natureza (ver Figura 17).

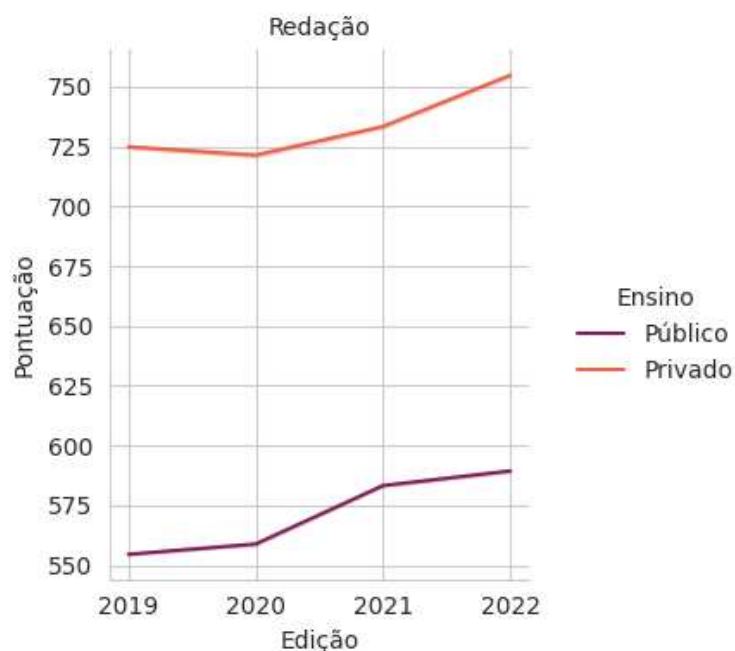
Figura 18 – Média dos participantes nas disciplinas de Matemática e Linguagens por rede de ensino e edição do ENEM



Fonte: Elaborado pelo autor.

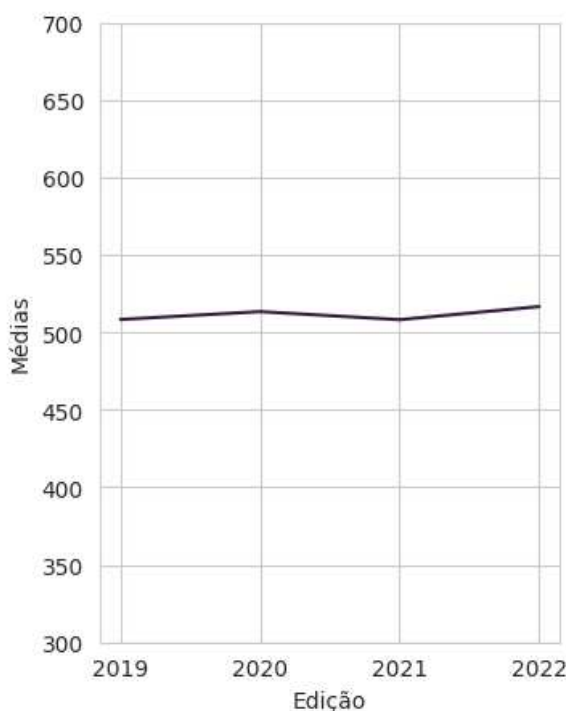
Na disciplina de Linguagens é apresentada certa semelhança nas médias como visto no gráfico da Figura 18. É perceptível uma diferença de 56 pontos do ensino privado para o público.

O desempenho na prova dissertativa-argumentativa não apresentou nenhum impacto negativo. Percebe-se um aumento das notas de redação em ambas as redes de ensino. Uma diferença, vista na Figura 19, de 165,38 pontos na média da rede privada para a pública em 2022.

Figura 19 – Média dos participantes na Redação por rede de ensino e edição do ENEM

Fonte: Elaborado pelo autor.

Embora, os candidatos da rede de ensino privado tenham um melhor desempenho na prova do ENEM do que os candidatos no ensino público, a média geral, independentemente da rede de ensino, demonstra uma constância como é visto na Figura 20. De 2019 a 2022, houve um aumento de 2,97% na média dos participantes do exame.

Figura 20 – Desempenho médio dos participantes por edição do ENEM

Fonte: Elaborado pelo autor.

4.5 Raça Autodeclarada

Para o estudo desta seção, foram considerados apenas os participantes que declararam alguma raça. Isto posto, os candidatos que não declararam nenhuma raça, um total de 208.037, foram excluídos.

Na Tabela 3 é mostrado a média e a quantidade percentual de participantes por raça autodeclarada.

Tabela 3 – Desempenho médio e a frequência percentual de raças autodeclaradas por edição do ENEM

Edições	Raças									
	Pardo		Branco		Preto		Amarelo		Indígena	
	Freq(%)	Média	Freq(%)	Média	Freq(%)	Média	Freq(%)	Média	Freq(%)	Média
2019	45,76%	503,71	37,14%	548,34	12,24%	499,56	2,29%	520,21	0,58%	477,20
2020	45,17%	506,93	38,04%	557,43	12,10%	500,62	2,22%	525,37	0,54%	471,95
2021	40,79%	505,97	43,76%	554,67	11,02%	503,06	2,01%	525,22	0,47%	467,50
2022	41,23%	516,36	43,77%	563,58	10,91%	512,81	1,87%	532,29	0,48%	481,36

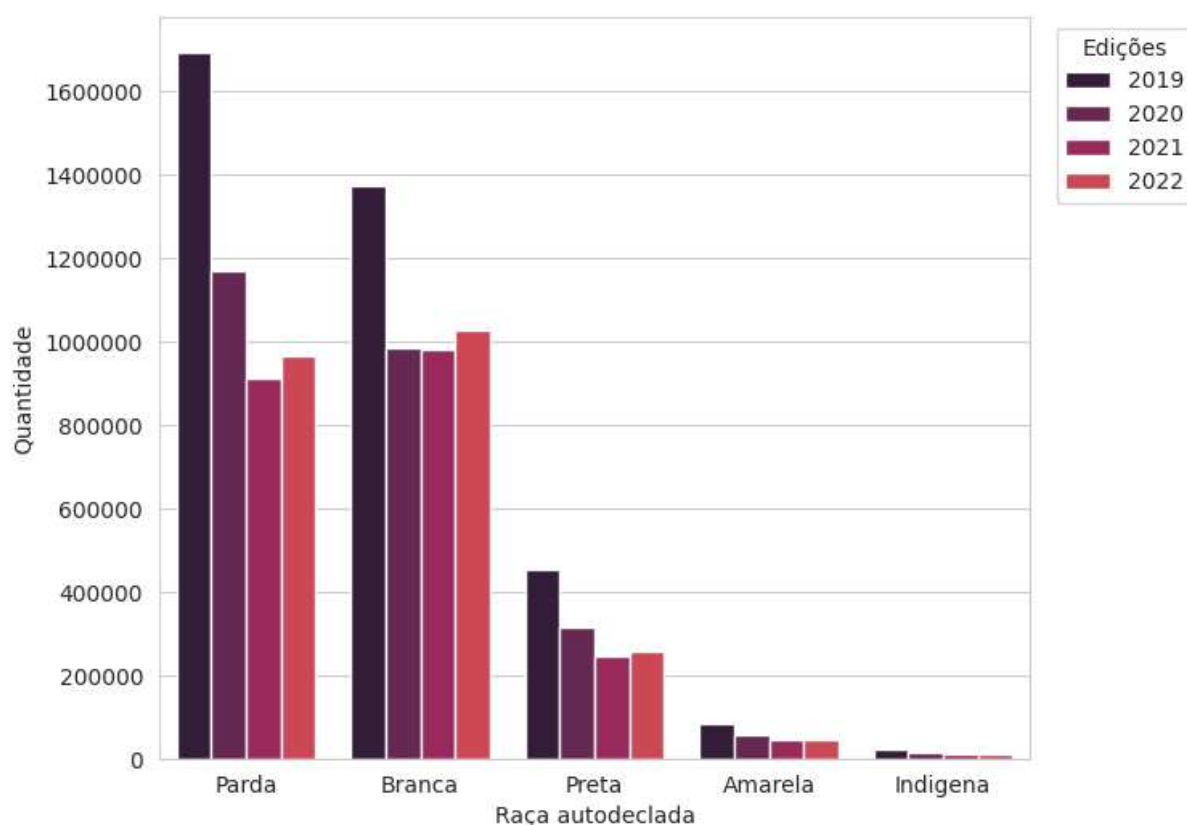
Fonte: Elaborada pelo autor.

Ao longo das edições, a frequência de pessoas autodeclaradas brancas aumentou em 6,63%, em contrapartida, percebe-se uma diminuição nas demais raças. Em relação ao desempenho, a média de todas as raças apresentam um crescimento, sendo a raça branca com as maiores médias e as indígenas com as

menores, apresentando uma diferença, na edição de 2022 por exemplo, de 82,22 pontos.

Na Figura 21, é possível ver com nitidez a distribuição do número de candidatos por raça autodeclarada.

Figura 21 – Número de candidatos por raça autodeclarada e edição do ENEM



Fonte: Elaborado pelo autor.

Apesar da queda perceptível na distribuição em todas as raças visto na Figura 21, a raça que teve maior impacto durante as edições de 2019 a 2021, no que diz respeito a sua frequência, foi a indígena na qual apresentou uma queda de 51,03%, em seguida, os amarelos (46,97%); pardos (46,12%); pretos (45,60%), e, por último, os brancos com a menor taxa (28,76%). Na Tabela 4, é mostrado a quantidade de candidatos por raça de acordo com a rede de ensino.

Tabela 4 – Percentual da distribuição dos participantes com raças autodeclaradas de acordo com rede de ensino por edição do ENEM

Edições	Raças									
	Pardo		Branco		Preta		Amarelo		Indígena	
	Freq(%)		Freq(%)		Freq(%)		Freq(%)		Freq(%)	
	Privado	Público	Privado	Público	Privado	Público	Privado	Público	Privado	Público

2019	4,49%	39,64%	10,80%	28,91%	0,93%	10,24%	0,41%	1,88%	0,03%	0,57%
2020	5,12%	37,60%	13,00%	29,01%	1,02%	9,40%	0,47%	1,74%	0,03%	0,52%
2021	4,76%	35,22%	14,54%	30,76%	1,05%	9,13%	0,43%	1,54%	0,02%	0,45%
2022	4,74%	36,04%	14,03%	30,84%	1,07%	9,17%	0,39%	1,39%	0,02%	0,48%

Fonte: Elaborado pelo autor.

Observa-se, na Tabela 4, que há uma percentagem maior de participantes que se autodeclararam pardos e que concluíram ou estão concluindo o ensino médio em rede pública, em comparação com as demais raças, e uma percentagem maior de brancos na rede privada. A raça indígena é a que possui os menores índices, independente da rede de ensino.

A partir da próxima seção, compreende os estudos do questionário socioeconômico. Vale salientar que para analisar os dados de alguns pontos do questionário, foi necessário excluir os dados faltantes. Portanto, foram excluídos 27.378, ficando um total de 10.846.143 participantes.

4.6 Escolaridade dos pais ou responsáveis

Nesta seção, analisa-se a participação dos candidatos segundo escolaridade dos pais ou responsáveis. Para isto, precisou-se juntar as duas colunas Q001 e Q002 do questionário socioeconômico que significa, respectivamente, a escolaridade de pais e/ou homens responsáveis e mães e/ou mulheres responsáveis. As respostas para estas perguntas são as seguintes:

- A – Nunca estudou;
- B – Não completou a 4ª série/5º ano do Ensino Fundamental;
- C – Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental;
- D – Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio;
- E – Completou o Ensino Médio, mas não completou a Faculdade;
- F – Completou a Faculdade, mas não completou a Pós-graduação;
- G – Completou a Pós-graduação;
- H – Não sei.

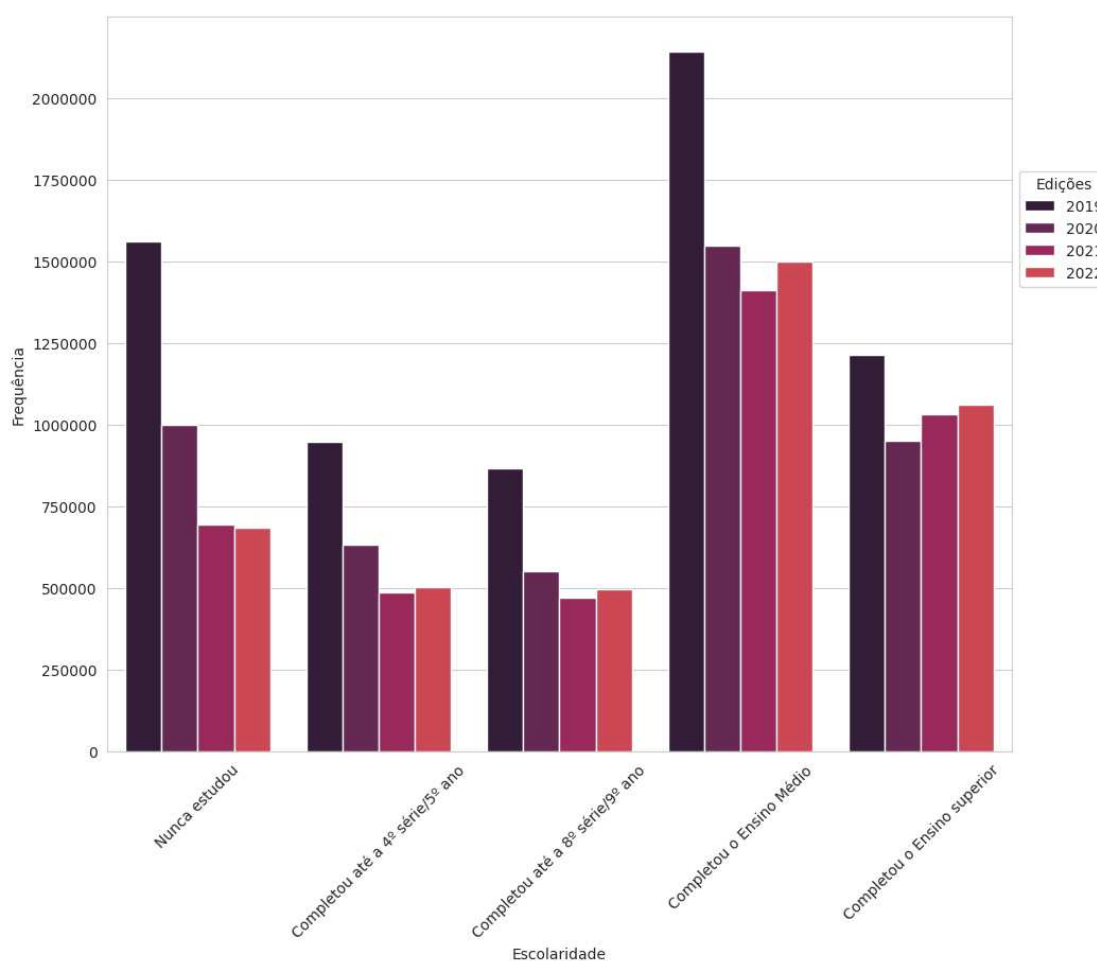
Assim como De Souza (2021), foi reduzido o número de opções para facilitar o estudo. Sendo assim, foram desconsideradas as pessoas que não sabem a escolaridade dos pais (item H) e consideradas os demais itens da seguinte forma:

- Nunca estudou: A e B;

- Completou a 4ª série/5º ano: C;
- Completou a 8ª série/9º ano: D;
- Completou o Ensino Médio: E;
- Completou o Ensino Superior: F e G.

Na Figura 22 é apresentada a distribuição dos participantes de acordo com a escolaridade de seus responsáveis.

Figura 22 – Distribuição dos participantes por escolaridade dos responsáveis de acordo com a edição do ENEM



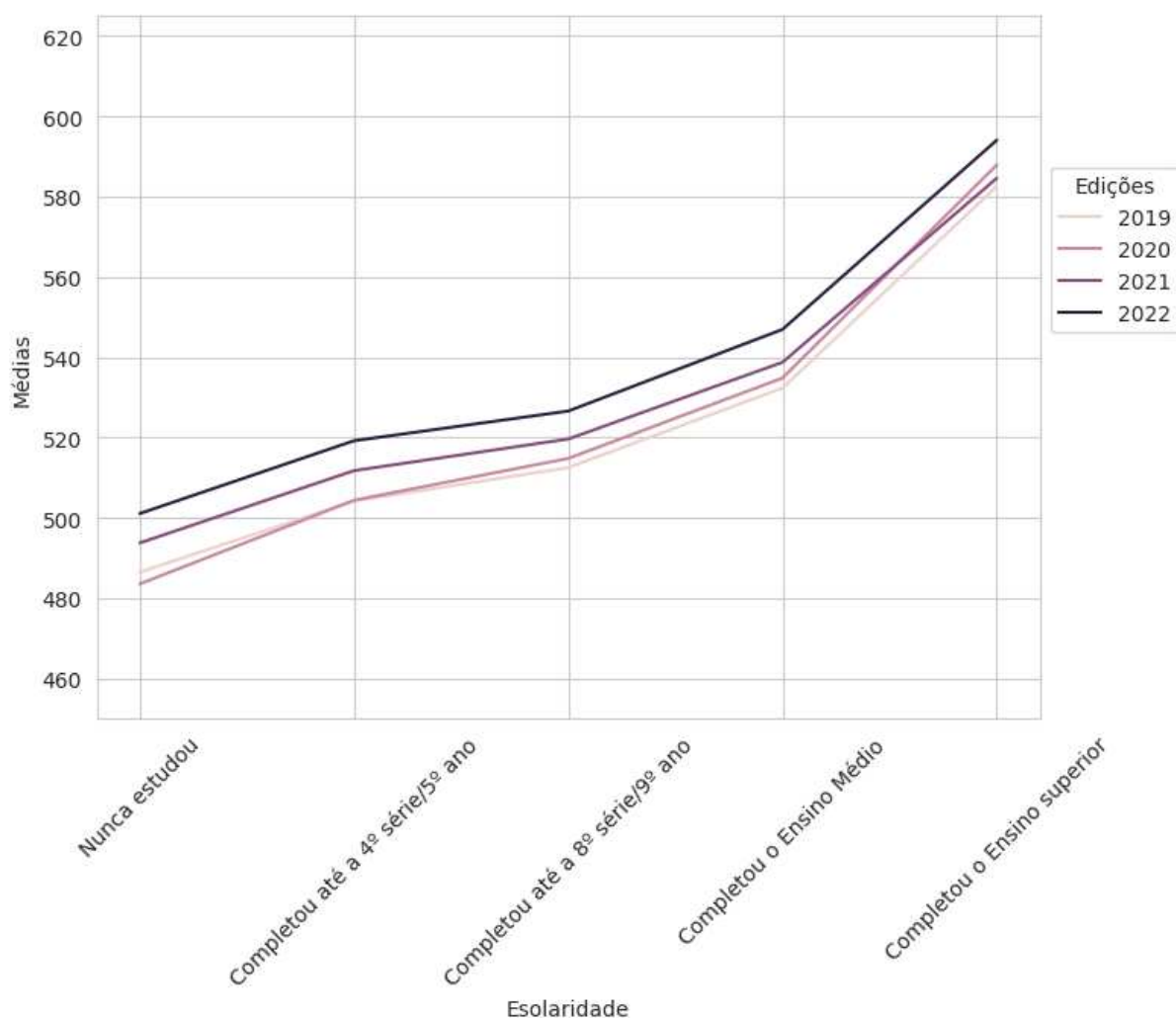
Fonte: Elaborado pelo autor.

De acordo com o gráfico da figura 22, nota-se uma maior frequência dos participantes com pais ou responsáveis que completaram o ensino médio no decorrer das edições destacadas e uma menor para os que completaram até a 8ª série/9º ano. Da edição do ano de 2019 para a de 2020, houve uma queda em todos os grupos. O grupo que completou o ensino superior teve o impacto de -21,62% no número de participantes. Já os que possuem o ensino fundamental completo tiveram -36,20% em

relação a 2019. É perceptível também, que o grupo dos que nunca estudaram não apresentou o mesmo crescimento de outros grupos na edição de 2022. Em 2022, em comparação com 2019, o impacto nesse grupo foi de -56,02% dos participantes.

Na Figura 23, é representada a média dos participantes de acordo com a escolaridade dos responsáveis.

Figura 23 – Média dos participantes por escolaridade dos responsáveis de acordo com a edição do ENEM



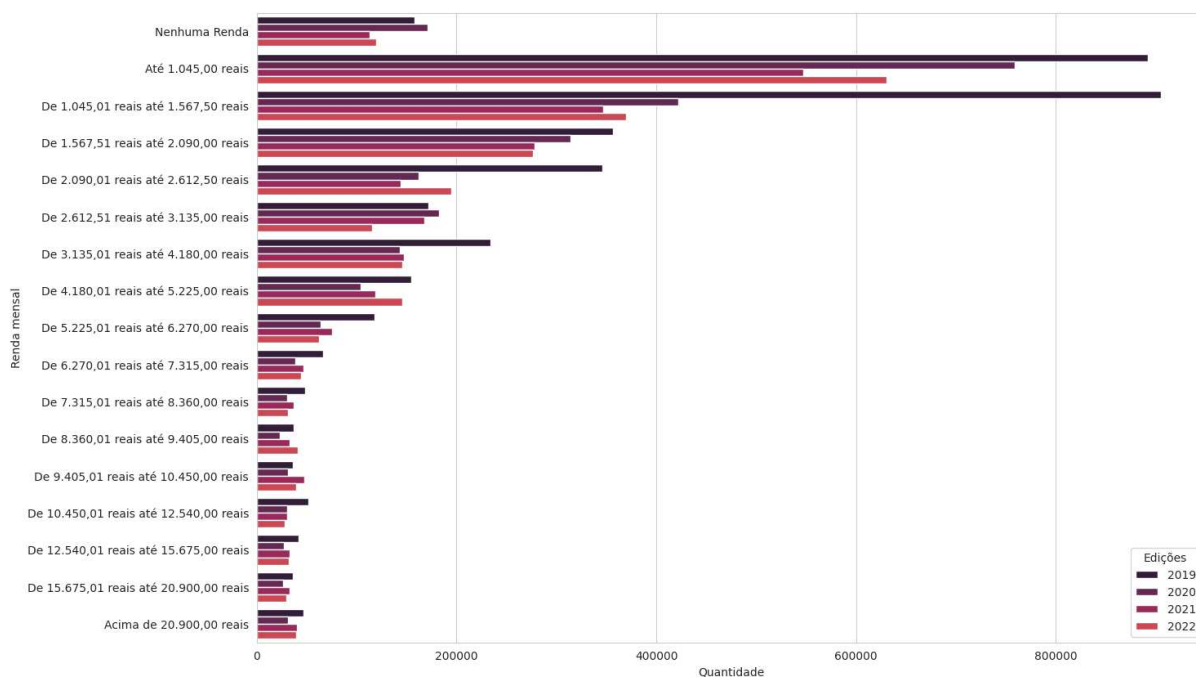
Fonte: Elaborado pelo autor.

É perceptível no gráfico, que quanto maior a escolaridade dos pais ou responsáveis, maior será a média do candidato. Todos os grupos tiveram um aumento na média. Na edição de 2022, a mais recente, a diferença entre os dois grupos é de 15,66% de uma média para outra, ou seja, 93 pontos na média.

4.7 Renda mensal

Nesta seção, estuda-se, por grupos, a renda mensal dos participantes das edições do ENEM. Na Figura 24 é apresentado o número de candidatos de acordo com a renda mensal.

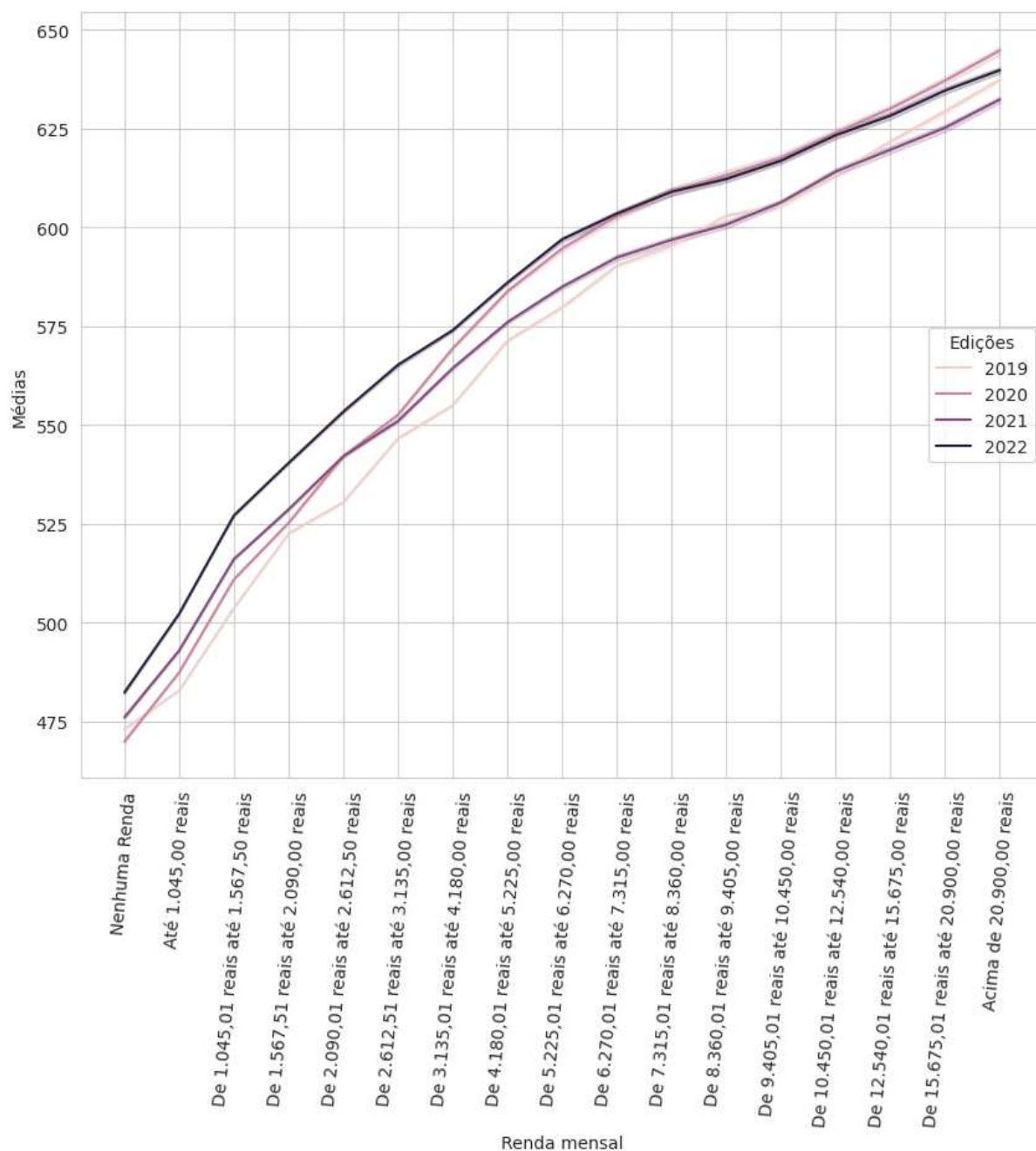
Figura 24 – Número de candidatos por renda mensal de cada edição do ENEM



Fonte: Elaborado pelo autor.

O grupo de participantes com a renda mensal de até R\$ 1.045,00 possui a maior frequência em todas as edições, como é apresentado na Figura 24. Já o grupo que sofreu o maior impacto, foi o de R\$ 1.045,01 até R\$ 1.567,50 com uma baixa - 53,38% da edição de 2019 para a de 2020. Os grupos com renda mensal a partir de R\$ 4.180,01 não sofreram nenhuma alteração discrepante em comparação com os grupos de renda inferior.

Na Figura 25, está representada a média dos candidatos de acordo com a renda mensal.

Figura 25 – Médias dos candidatos por renda mensal de cada edição do ENEM

Fonte: Elaborado pelo autor.

No decorrer das edições, é perceptível um aumento na média em cada grupo como mostrado na Figura 25, porém não há uma diferença discrepante entre as edições. Ademais, quanto maior a renda, melhor é o desempenho dos participantes.

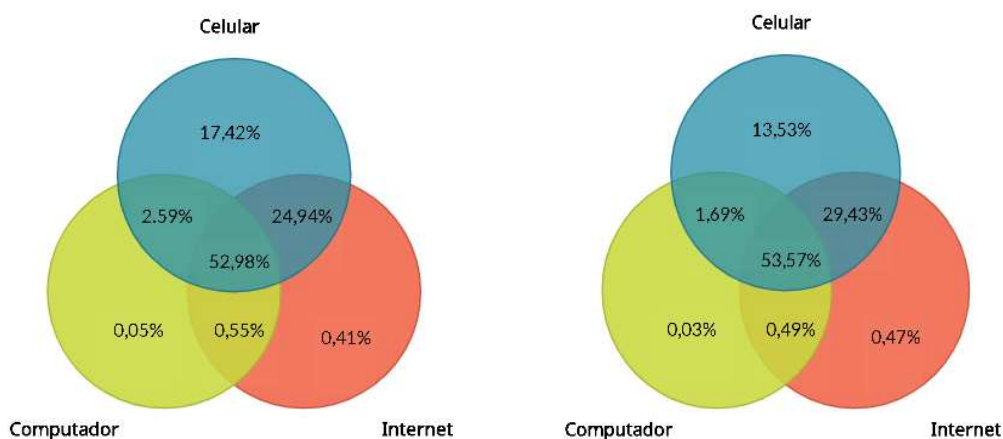
4.8 Acesso à Tecnologia

Para a realização da análise desta seção, foram consideradas as colunas Q022, Q024 e Q025 que significam, respectivamente, a posse de celulares,

computadores e internet pelos participantes. As colunas Q022 e Q024 são numéricas, ou seja, variam de acordo com a quantidade de aparelhos. Assim sendo, foi considerado apenas as repostas sim ou não, independente da quantidade ou não de dispositivos que o candidato apresentar.

Com a adoção da medida restritiva para conter os avanços da COVID-19, houve uma necessidade ainda maior do uso das tecnologias. Da edição de 2019 para 2021, teve um aumento de 8,38% para os participantes que utilizaram computador, celular e internet, como visto nas Figuras 26 e 27.

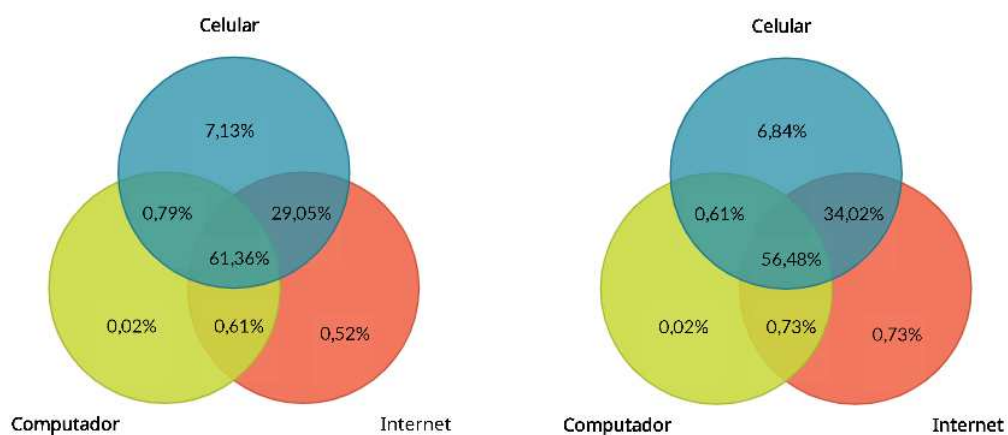
Figura 26 – Percentual de distribuição dos participantes com acesso à tecnologia nas edições 2019 e 2020 do ENEM



Fonte: Elaborado pelo autor.

É perceptível também a diminuição dos participantes que não possuem acesso à internet. O dispositivo celular, por exemplo, apresentou uma queda na utilização somente do aparelho, ao longo das edições, de -10,58%. Essa redução é visível inclusive para os candidatos que apresentam apenas computadores ou os dois juntos.

Figura 27 – Percentual de distribuição dos participantes com acesso à tecnologia nas edições 2021 e 2022 do ENEM



Fonte: Elaborado pelo autor.

Em relação aos participantes que não possuem nenhum tipo de tecnologia, somam um total de 84.512, ou seja, 0,78% de todas as edições.

Na Tabela 5 é mostrado a média dos participantes de acordo com o acesso ou não às tecnologias para cada disciplina de cada edição.

Tabela 5 – Média dos participantes por disciplina de acordo com o acesso às tecnologias nas edições no ENEM

Edições	Disciplinas									
	Ciências da Natureza		Ciências Humanas		Linguagens		Matemática		Redação	
	Com Acesso	Sem Acesso	Com Acesso	Sem Acesso	Com Acesso	Sem Acesso	Com Acesso	Sem Acesso	Com Acesso	Sem Acesso
2019	478,36	433,32	510,42	460,00	522,89	476,05	523,85	461,61	580,80	483,61
2020	490,92	440,86	514,65	454,26	526,46	470,05	521,35	445,29	582,32	483,31
2021	492,17	440,23	523,52	458,36	505,21	441,09	535,58	462,16	623,13	510,29
2022	496,25	455,67	530,28	475,83	520,23	458,10	543,05	465,90	629,63	510,17

Fonte: Elaborado pelo autor.

É nítido que os participantes com acesso às tecnologias possuem um desempenho maior em todas as edições.

4.9 Clusterização

Para esta seção, foram utilizadas apenas as notas dos participantes e suas respectivas médias para identificar possíveis grupos. Utilizando o método do cotovelo (no inglês, *elbow method*), foram identificados três possíveis grupos: baixo, médio e alto desempenhos como mostrado na Tabela 6.

Tabela 6 – Média dos participantes por grupo no ENEM

Grupos	Disciplinas				
	Ciências da Natureza	Ciências Humanas	Linguagens	Matemática	Redação
Baixo	436,76	448,96	458,22	453,45	297,04
Médio	461,77	492,40	502,69	486,01	583,87
Alto	559,72	599,20	580,10	644,55	788,78

Fonte: Elaborado pelo autor.

Apesar do resultado ideal do agrupamento, o algoritmo foi testado com outros valores, 4 e 5, para verificar se seria possível a formação de outros grupos, porém apresentarem resultados semelhantes.

De acordo com a Tabela 6, é visto que a diferença mais discrepante é na prova de redação, do grupo de notas baixas para as médias, há um intervalo de 49,13%, já do médio para o alto, 25,98%.

5 Conclusão

A pandemia da COVID-19 trouxe novos desafios para todos os setores do país. No setor educacional, a utilização em massa do Ensino à Distância que antes era decidido pelo aluno, durante a pandemia, foi obrigatório como forma de impedir o contágio da doença nas instituições de ensino.

O presente trabalho apresentou a aplicação de conceitos e técnicas de Ciência de Dados na base de dados das edições de 2019 a 2022 do ENEM. O objetivo foi verificar o impacto da COVID-19 no desempenho dos participantes, visto que o ano de 2022, comparado aos anteriores, foi o ano em que teve menos notificações e óbitos pela doença. Além do mais, este trabalho também buscou visualizar a possível mudança nos perfis dos candidatos, averiguando se houve alteração tendo como base a edição de 2019.

No estudo realizado com as faixas-etárias dos participantes, foi visto que há uma frequência maior no grupo dos candidatos que possuem de 17 a 20 anos, idade que costuma-se finalizar o ensino médio, do que os demais. Em relação à média por intervalo de idade, há um padrão que mostra que quanto maior a idade do participante, menor a sua média.

Com relação ao sexo, há uma presença maior de pessoas do sexo feminino do que masculino em todas as edições. A única queda aparente na média dos participantes foi na disciplina de Linguagens e Códigos da edição de 2019 para 2021, em que 25%, de ambos os sexos, apresentaram uma queda de -7,94% e -6,15%, respectivamente, para o sexo feminino e masculino. A redação foi a única prova em que os participantes do sexo feminino apresentaram um resultado melhor que o masculino, onde obtiveram um aumento de 15% na nota da redação em 75% do público feminino no decorrer das provas.

Sobre os estudos da distribuição por região e rede de ensino dos participantes, não houve discrepâncias no resultado ao longo das edições. No que se refere ao estudo das raças autodeclaradas em relação a rede de ensino, a rede pública possui uma percentagem maior de pardos, e na rede privada, maior número de brancos, sendo os indígenas os que possuem o menor número em ambas as redes de ensino. Ainda, em relação as raças indígenas e brancas, o número de participantes de 2019

a 2021 sofreu um impacto, respectivamente, de 51,03% e 28,76%, sendo esta a menor dentre as raças.

De acordo com o estudo da escolaridade dos pais ou responsáveis e estudo da renda mensal dos participantes, é perceptível o padrão que os dados fornecem, onde quanto maior o grau de formação dos responsáveis pelos participantes, melhor é o desempenho deles, e quanto maior a renda, melhor a pontuação. Com relação ao estudo dos participantes com acesso às tecnologias, é possível perceber um aumento no uso de dispositivos com internet e, paralelamente, um desempenho melhor dos candidatos.

Ainda, para fim de estudos, foi utilizada a técnica de clusterização com intuito de encontrar segmentos de dados que apresentem algum padrão. O resultado da aplicação do algoritmo de *K-means* encontrou três grupos com diferentes pontuações. Dessa maneira foi elencado em baixo, médio e alta levando em consideração o intervalo de 0 a 1000 como referência. O intervalo maior com as médias foi da prova de redação, onde a diferença do grupo com notas mais baixas para as mais altas é de 62,34%.

Tendo em vista os estudos realizados no presente trabalho, é perceptível que a pandemia da COVID-19, ao longo das edições de 2019 a 2022, não apresentou impacto algum em relação ao desempenho dos estudantes. No entanto, é possível constatar que houve comportamentos distintos em relação aos perfis dos participantes. Dessa maneira, é preciso investigar as características separadamente para obter um maior entendimento dessas informações.

Para trabalhos futuros, objetiva-se a agregação de outros bancos de dados de outras instituições para ter resultados completos. Uma outra possibilidade, é o estudo dos grupos socioeconômicos e dos participantes faltantes, para compreender de acordo com os seus perfis, o que pode motivar o candidato a faltar um ou os dois dias da prova.

REFERÊNCIAS BIBLIOGRÁFICAS

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Enem**. 2023. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>. Acesso em: 26 set. 2023.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). **Enem**. 2023. Disponível em: <https://www.gov.br/inep/pt-br/area-de-informacao/dados-abertos/microdados/enem>. Acesso em: 12 jul. 2023.

BRASIL. Ministério Público Federal (MPF). **O que é a LGPD?** 2023. Disponível em: <https://www.mpf.mp.br/servicos/lgpd/o-que-e-a-lgpd>. Acesso em: 27 set. 2023.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). **População**. 2023. Disponível em: <https://cidades.ibge.gov.br/brasil/panorama>. Acesso em: 30 set. 2023.

CAVALCANTI, Gilmara Alves; **Apostila de Estatística**. 2006. Notas de aula.

CLEVELAND, William S. **Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics**. International Statistical Review / Revue Internationale de Statistique, vol. 69, no. 1, 2001, pp. 21–26. JSTOR. Disponível em: <https://doi.org/10.2307/1403527>. Acesso em: 25 set. 2023.

CRESPINO, Antônio Arnot; **Estatística Fácil**. 19ª edição. São Paulo: Saraiva, 2009.

DE SOUZA, Thiago Oliveira. **Análise de dados: um estudo do perfil dos participantes do ENEM 2019**. 2021. 46p. Monografia (Bacharelado em Ciência da Computação) – Universidade Federal Rural do Semi-Árido, Mossoró, 2021. Disponível em: /. Acesso em 12 jul. 2023.

DEVMEDIA. **Data Mining na Prática: Algoritmo K-Means**. Disponível em: <https://www.devmedia.com.br/data-mining-na-pratica-algoritmo-k-means/4584>. Acesso em 27 set. 2023.

GONÇALVES, Pollyana. **Afinal, como se desenvolve um projeto de Data Science?** 2018. Disponível em: <https://medium.com/techbloghotmart/afinal-como-se-desenvolve-umprojeto-de-data-science-233472996c34>. Acesso em: 25 set. 2023.

LENZ, Maikon L.; NEUMANN, Fabiano B.; SANTARELLI, Rodrigo; SALVADOR, Douglas. **Fundamentos de aprendizagem de máquina**. Porto Alegre: Grupo S. A, 2020. E-book. ISBN 9786556900902. Disponível em: <https://app.minhabiblioteca.com.br/#/books/9786556900902/>. Acesso em: 28 set. 2023.

SOUTO, X. M. COVID-19: aspectos gerais e implicações globais. **Recital - Revista de Educação, Ciência e Tecnologia de Almenara/MG**, v. 2, n. 1, p. 12–36, 2020. Disponível em: <https://recital.almenara.ifnmg.edu.br/index.php/recital/article/view/90>. Acesso em: 27 set. 2023.

COSTA NETO, Pedro Luiz de Oliveira; **Estatística**. 3ª edição. São Paulo: Blucher, 2002.

OMS. Organização Mundial de Saúde. **Histórico da pandemia de COVID-19**. 2023. Disponível em: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19>. Acesso em: 28 set. 2023.

RAUTENBERG, Sandro; DO CARMO, Paulo Ricardo Viviurka. **Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão**. Brazilian Journal of Information Studies: Research Trends. 2019. p.56-p.67. ISSN 1981-1640. Disponível em: <https://www.brapci.inf.br/index.php/res/download/112105>. Acesso em: 25 set. 2023.

SPIEGEL, Murray R.; STEPHENS, Larry J. **Estatística**. 4ª edição. Porto Alegre: Bookman, 2009.