Capítulo 14

Análise de dados: um estudo do perfil dos participantes do ENEM 2019

Data analysis: a study of the profile of ENEM 2019 participants

https://doi.org/10.56238/sevedi76016v22023-014

Thiago Oliveira de Souza
Graduado
Universidade Federal Rural do Semi-Árido – UFERSA
E-mail: thiagooliveira1450@gmail.com

Angélica Félix de Castro
Doutora
Instituição de atuação atual: Universidade Federal Rural do Semi-Árido - UFERSA
E-mail: angelica@ufersa.edu.br

Amanda Gondim de Oliveira
Doutora
Instituição de atuação atual: Universidade Federal Rural do Semi-Árido - UFERSA
E-mail:amandagondim@ufersa.edu.br

RESUMO
O entendimento dos aspectos da educação de um país é de fundamental impor- tância 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 2019 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-chave: Ciência de Dados, ENEM, Python, Pandas, Matplotlib, Seaborn.

ABSTRACT
Understanding the aspects of education in a country is of fundamental importance to establish goals for improving education. The following work uses statistical and data science techniques in order to identify relevant characteristics of the participants of the National Secondary Education Examination for the year 2019 and possible relationships of these characteristics with their performance. Its purpose is to verify whether such characteris- tics reflect the performance of exam participants. With the help of the Pandas, Matplotlib, Seaborn, Numpy and sklearn libraries of the Python language, it was possible to find some factors that influence the performance of the exam participants and to categorize some characteristics of their profile

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

1 INTRODUÇÃO

Com a internet cada vez mais acessível, a disseminação de grandes quantidades de dados é parte do cotidiano da rede mundial de computadores. Big Data é o termo usado para denominar essa quantidade gigantesca de dados. Com esse volume massivo de dados sendo produzido todos os dias, uma janela se abre para a exploração dos mesmos e a descoberta de informações úteis que estavam mascaradas por trás de todos esses de dados. Esse estudo dos dados denomina-se ciência de dados.

A criação de ferramentas e soluções computacionais que auxiliam a extrair informações e conhecimento de grandes bases de dados tem sido essenciais para grandes organizações, que podem

direcionar seu foco para certas áreas e públicos específicos. O advento da análise de dados não é uma exclusividade de ambientes corporativos, tomando como exemplo a educação. O Big Data pode contribuir para identificar problemas e orientar a ação de gestores, profissionais e governantes - tanto na elaboração de políticas públicas quanto na administração escolar (DESAFIOS DA EDUCAÇÃO, 2019).

Os dados podem ser classificados como estruturados, semiestruturados e não estruturados. Neste trabalho o estudo foi feito utilizando dados estruturados, mais especificamente, dados sobre a educação básica brasileira. O maior repositório de dados sobre a educação do Brasil é do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Em meio aos dados disponíveis no INEP temos informações sobre basicamente todos os níveis de ensino e educação do país. Neste artigo, fez-se o uso dos dados do ENEM do ano de 2019, pois, estes eram os dados mais recentes referentes ao exame, na época do desenvolvimento deste trabalho. A finalidade deste trabalho é traçar o perfil dos participantes do ENEM 2019 e encontrar *insights* que possam relacionar esses perfis ao desempenho dos mesmos.

Na seção 2 é apresentada a fundamentação teórica para a realização do estudo proposto. Na seção 3 estão apresentados os materiais e métodos utilizados neste artigo. Na seção 4 ocorre a discussão dos resultados obtidos. Por fim, na seção 5, as conclusões tiradas a respeito do estudo realizado.

2 BACKGROUND

Nessa seção serão brevemente apresentados: 2.1) O Exame Nacional do Ensino Médio (ENEM) como objeto de estudo; 2.2) ciência de dados como suporte para realização do estudo pretendido; e 2.3) Noções de estatística para compreensão dos resultados obtidos

Exame Nacional do Ensino Médio (ENEM)

"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" (INEP, 2021). Desde 2009 o exame passou também a ser aceito como mecanismo de acesso ao ensino superior, podendo ser complemento a um vestibular ou ser todo o processo seletivo para a faculdade. Sendo assim, o ENEM significa para muitos estudantes uma porta de acesso à educação superior de qualidade e gratuita que é ofertada por instituições públicas pelo Sistema de Seleção Unificada (SISU). Também é possível que os participantes do exame concorram a bolsas de estudo pelo Programa Universidade para Todos (ProUni).

No ato da inscrição, os participantes devem preencher um formulário socioeconômico. Além dos dados socioeconômicos também são disponibilizadas outras informações a respeito dos participantes, como idade, sexo, município de residência, tipo de escola que concluiu o ensino médio, entre outras informações. Os dados desse formulário são disponibilizados pelo INEP e com esses dados foi possível realizar o presente estudo.

Ciência de dados

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 (CETAX, 2022).

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 Gonçalves (2018). Tais etapas são:

- **Entendimento do problema** - pode ser considerada uma das mais importantes etapas do ciclo. A partir do entendimento do problema é que é possível definir os meios de pesquisa e desenvolvimento para alcançar o resultado desejado.
- **Coleta de dados** - 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.
- **Processamento/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.
- **Exploração de dados** - 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. Nessa etapa, é realizado o estudo das ideias e hipóteses que se busca validar. Devido a isso, é de fundamental importância que se tenha uma boa habilidade analítica.
- **Análise profunda de dados** - 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, pois, alguns estudos já tem seu objetivo concluído na etapa anterior; não sendo necessária nenhuma análise profunda.
- **Comunicação de resultados e Feedback** - etapa que se tem 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 2019, passando pelas etapas do ciclo da ciência de dados mencionadas anteriormente, com ênfase maior na exploração de dados.

Noções de Estatística

A estatística lida com coleta, tratamento, análise, interpretação e apresentação de dados numéricos. Ela está presente na maioria dos aspectos da ciência de dados. O campo da estatística é muito amplo, mas nem todos os seus conceitos são obrigatórios para desenvolver estudos com o auxílio da ciência de dados (Vickery, 2021).

Nesta seção, são apresentados alguns conceitos fundamentais que auxiliam na análise e interpretação dos resultados obtidos neste trabalho

Amostragem Estatística

Todos os dados brutos disponíveis para estudo é chamado de população. Nem sempre é possível utilizar toda a população para fazer a análise desejada. As estatísticas possibilitam que seja possível 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.

Suponha-se que se queira ter um panorama da qualidade de ensino dos concluintes do ensino médio no Brasil. A população de estudos desejada seria de todos os concluintes do ensino médio no país. Devido a não ser possível obter todos esses dados, por questões de logística, pode ser utilizada uma amostra que represente toda a população. Desde que essa amostra tenha uma boa representatividade de toda a população, podem ser feitas inferências sobre a população em sua totalidade.

Estatística descritiva

A estatística descritiva auxilia a descrever os dados e compreender suas características. Nesta etapa o objetivo não é formular uma predição ou inferências, é onde são apresentadas descrições da aparência da amostra que se tem. Geralmente as estatísticas descritivas são obtidas a partir dos dados, com médias de tendência central, como:

- **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.

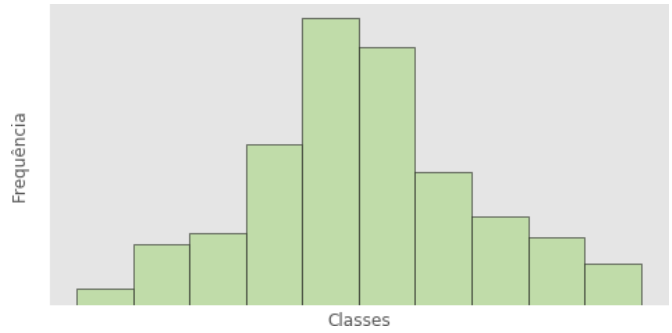
Distribuições

As estatísticas descritivas, apesar de úteis, podem mascarar informações importantes sobre a amostra estudada. Caso um conjunto de dados contenha valores que sejam muito maiores que outros, a média será distorcida e não pode ser considerada uma representação fidedigna dos dados.

Para representar uma distribuição, pode ser feito o uso de um histograma. O histo grama é uma espécie de gráfico de barras que demonstra uma distribuição de frequências. Nesse gráfico, a base das

barras representa uma classe de valores e a altura representa a frequência que o valor de cada classe ocorre. A Figura 1 mostra um exemplo de histograma.

Figura 1: Exemplo de Histograma

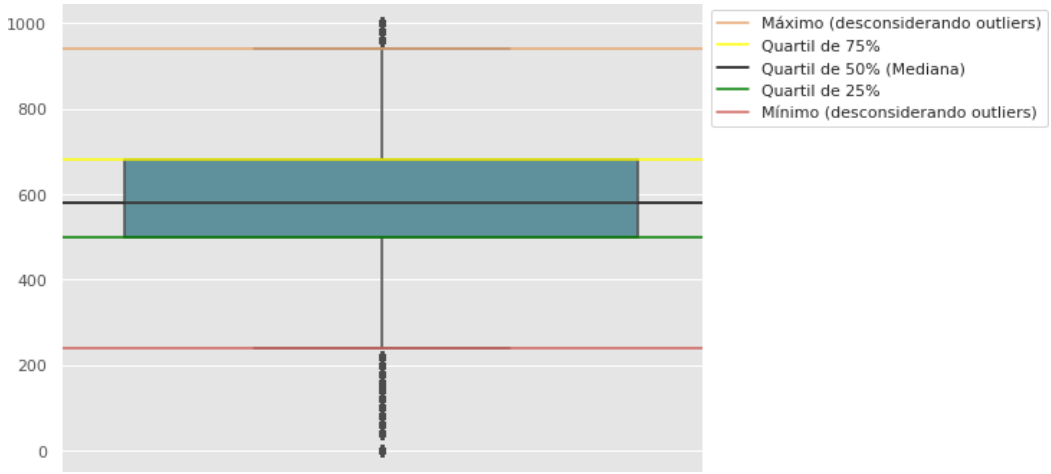


Outro gráfico que pode ser utilizado para analisar distribuição de dados é o *boxplot*. Para explicar o que é um *boxplot*, é necessário compreender o que são os quartis, que são medidas apresentadas nesse gráfico.

Percentil é uma medida posicional que, dada uma amostra ordenada de forma crescente e dividida em cem partes, indica o valor do qual determinado percentual de elementos da amostra são iguais ou inferiores a ele (Oliveira, 2019). Por exemplo, o percentil de 25 de um conjunto ordenado de idades, nos resultaria um valor de idade que indicaria que 25% dos valores desse conjunto são iguais ou inferiores ao valor obtido.

Os quartis apresentados no *boxplot* são equivalentes aos percentis 25, 50 e 75, que representam, respectivamente, o primeiro, segundo e terceiro quartil. O segundo quartil é o percentil 50, ou seja, representa 50% da amostra. Dessa forma, o segundo quartil equivale a mediana do conjunto. O *boxplot* também mostra valores discrepantes do conjunto, chamados de *outliers*. A Figura 2 mostra um exemplo de *bloxplot*.

Figura 2: Exemplo de boxplot



O valor mínimo considerado no *boxplot* é representado pela linha vermelha, valores abaixo desse mínimo são representados como pontos e são os valores atípicos, considerados *outliers*. A linha verde representa o primeiro quartil, que é o quartil de 25%. O que significa que 25% dos valores dessa amostragem são iguais ou inferiores a essa linha verde. A linha preta é equivalente ao quartil de 50% e a mediana dos valores, ou seja, 50% dos valores do conjunto são iguais ou inferiores ao valor dessa linha. A linha amarela é referente ao quartil de 75%, o que indica que 75% dos valores do conjunto são iguais ou inferiores ao valor dessa linha. Por fim, a linha marrom representa o valor máximo do *boxplot*, valores superiores a essa linha são considerados *outliers*. O valor mínimo e o valor máximo são valores considerados adequados para essas faixas, considerando todo o conjunto amostral. Não significa que sejam de fato o menor e o maior valor do conjunto propriamente dito.

Correlação

A correlação é uma técnica estatística que mede as relações entre duas variáveis. A correlação pode ser considerada linear e ser expressa como um número entre +1 e -1. Esse número é chamado de coeficiente de correlação. Quanto mais próximo de +1 ou -1, mais forte a correlação, e quanto mais próximo de 0, mais fraca é a correlação. O valor 0 é considerado uma correlação inexistente. O sinal do coeficiente de correlação indica a forma como as variáveis se correlacionam. Um coeficiente positivo indica que as variáveis crescem ou decrescem na mesma direção. Já um coeficiente negativo, indica que enquanto uma variável cresce a outra decresce.

É interessante ter ciência que a correlação não implica em causa, o fato de existir correlação entre duas variáveis não significa que uma é o motivo de ocorrência da outra.

Regressão Linear

A técnica de regressão linear utiliza o valor de uma variável para fazer uma previsão a respeito de outra variável. Essa variável que se deseja prever é conhecida como variável dependente e a variável que é utilizada para fazer a previsão é chamada de independente (IBM, 2021). Essa forma de análise estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes que melhor preveem o valor da variável dependente. A regressão linear se ajusta a uma linha reta ou superficial que minimiza as discrepâncias entre os valores de saída previstos e reais.

A próxima seção é destinada à apresentação dos materiais e métodos usados na elaboração deste trabalho.

3 MATERIAIS E MÉTODOS

Nesta seção são apresentadas: as ferramentas utilizadas, a obtenção dos dados do ENEM e a descrição e apresentação do tratamento de dados utilizados para realização dos estudos propostos.

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*** **-** biblioteca *Python* utilizada para análise de dados. Com ela, foi feita toda a leitura, tratamento e processamento dos dados.
- ***Matplotlib*** **-** biblioteca utilizada para criar gráficos diversos para tipos de dados variados. Grande parte dos gráficos apresentados neste trabalho foram feitos utilizando esta biblioteca.
- ***Seaborn*** **-** biblioteca que auxilia na criação de gráficos. Costuma ter um *layout* mais apresentável que os gráficos criados pelo *matplotlib*.
- ***Numpy*** **-** ajuda a executar facilmente cálculos numéricos. É usada principalmente para realizar cálculos em *Arrays* Multidimensionais.
- ***Sklearn*** **-** essa biblioteca auxilia especificamente a aplicar técnicas de *Machine Learning* no conjunto de dados desejado.

Para o ambiente de execução, foi utilizado o *Google Colaboratory*, ou simplesmente "Colab". Trata-se de um ambiente de execução virtual que pode ser utilizado através do seu navegador de internet, onde é utilizado processamento e recursos de servidores do Google, o que possibilita que máquinas que não tenham um hardware tão potente possam lidar com bases de dados massivas.

Obtenção 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/dados- abertos/microdados/enem. Para este estudo, utilizamos os dados do exame de 2019.

Tratamento dos dados

O arquivo csv que contém os dados do ENEM 2019 tem cerca de 3,2 GB de tamanho. Ao descarregar a base de dados pelo sítio do INEP, o arquivo vem compactado. Devido ao Colab ser uma propriedade do Google, ele tem integração com alguns do serviços do mesmo, inclusive com o Google Drive. Desse modo, através da biblioteca zipfile do Python, foi possível utilizar a base de dados compactada armazenada no Google Drive sem a necessidade de fazer a descompressão.

O arquivo .csv principal (MICRODADOS_ENEM_2019) contém os questionários respondidos pelos participantes, armazenando todas as informações disponibilizadas pelos participantes do ENEM 2019 em um único arquivo. Essa base de dados contém 5.095.270 linhas e 136 colunas. Essa tabela carrega uma

grande quantidade de dados e erros de execução ao tentar carregar essas informações são comuns, mesmo utilizando um ambiente de execução robusto como o Colab. As informações desse arquivo principal foram carregadas em um *DataFrame* da biblioteca Pandas. De acordo com (VAZ, 2021):

> *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, entre outras. 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.

Tabela 1: Colunas utilizadas no presente estudo.

Nome da Coluna	Descrição
NU_INSCRICAO	Número da Inscrição
CO_MUNICIPIO_RESIDENCIA	Código do município de residencia
NU_IDADE	Idade
TP_SEXO	Sexo
TP_COR_RACA	Cor/raça autodeclarada
TP_ESCOLA	Tipo de escola do Ensino Médio
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? (Some a sua renda com a dos seus familiares.)
Q022	Na sua residência tem telefone celular?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

Como o propósito deste trabalho é averiguar relações entre as características dos participantes e desempenho, foi feita a remoção de entradas vazias do *DataFrame* - ou seja, de inscritos que não participaram de uma ou mais etapas do exame. Após esse recorte inicial, o *DataFrame* principal foi reduzido para 3.701.947 linhas e 17 colunas. Esse *DataFrame* será referenciado como Base 1 na seção seguinte.

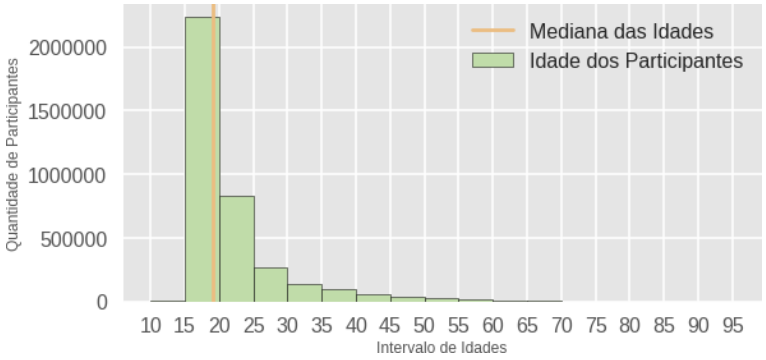
4 RESULTADOS E DISCUSSÕES

Esta seção descreve os estudos realizados e apresenta uma análise descritiva dos dados obtidos. Aqui, estão apresentadas todas as análises feitas com a base de dados dos participantes do ENEM 2019.

Idade

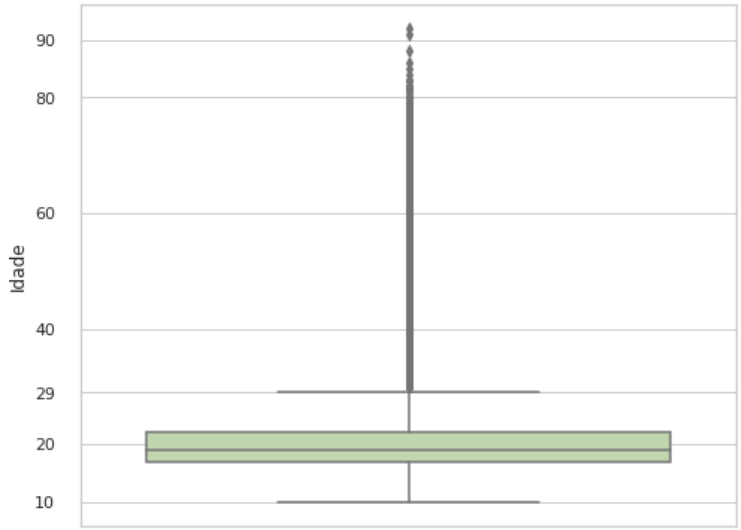
No exame é permitida a participação de candidatos de quase todas as idades. Na Base 1, aparecem candidatos a partir de 10 anos até participantes de 92 anos. Apesar da grande abrangência de idades, existem algumas ocorrências mais comuns, a mediana das idades dos participantes é de 19 anos. Como é mostrado no histograma da Figura 3, pode-se perceber que a maioria dos candidatos (89,82%) estão entre 15 e 29 anos. O intervalo mais representativo é o de participantes entre 15 e 19 anos, contendo 2.232.119 dos candidatos e representando 60,30% de todo o conjunto.

Figura 3: Histograma de idades dos participantes do Enem 2019



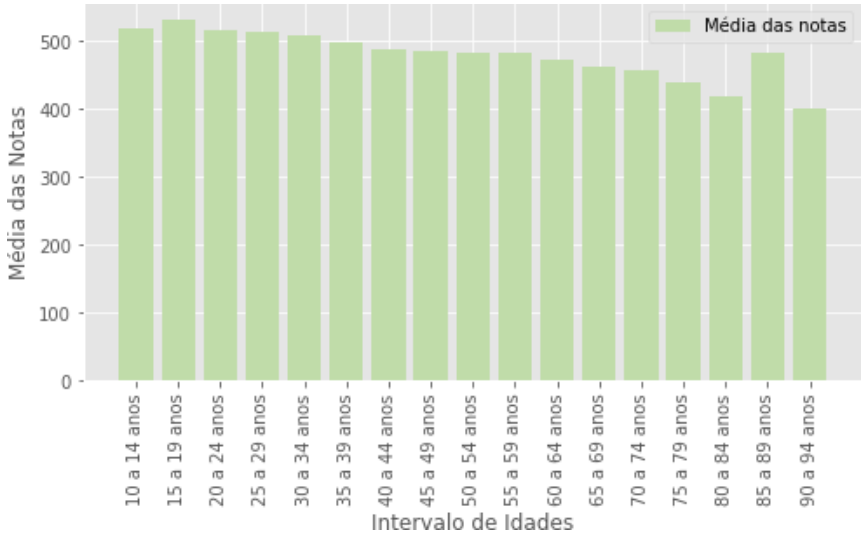
A Figura 4 apresenta um *boxplot* com a distribuição de idades dos participantes. É perceptível que as idades se concentram entre os 15 e 29 anos, como mencionado anteriormente. Valores superiores a 29 anos são consideradas *outliers*.

Figura 4 – Boxplot das idades dos participantes do ENEM 2019



Dado o vasto intervalo de idades, a questão subsequente é se a diferença de idades afeta o resultado no exame de forma determinante. Para este estudo foi feita a média aritmética das cinco notas dos participantes, para se obter uma noção do desempenho geral dos mesmos. Foi considerado o mesmo intervalo de idades da Figura 3, para a averiguação do desempenho dos candidatos. Para cada intervalo de idade foi feita a soma de todas as médias, e em seguida tirada a média aritmética dessa soma (Figura 5).

Figura 5 – Média das notas por idade

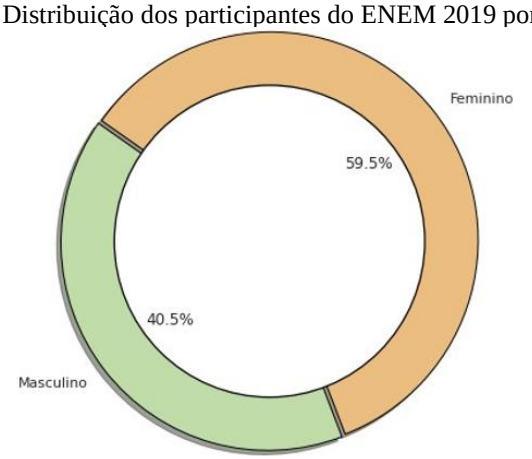


Percebe-se na Figura 5 acima que a idade não tem grande influência na nota obtida no exame. Apesar da densidade de cada intervalo de idades ser diferente, uma constância nos valores da média das notas se mantém para todos eles.

Sexo

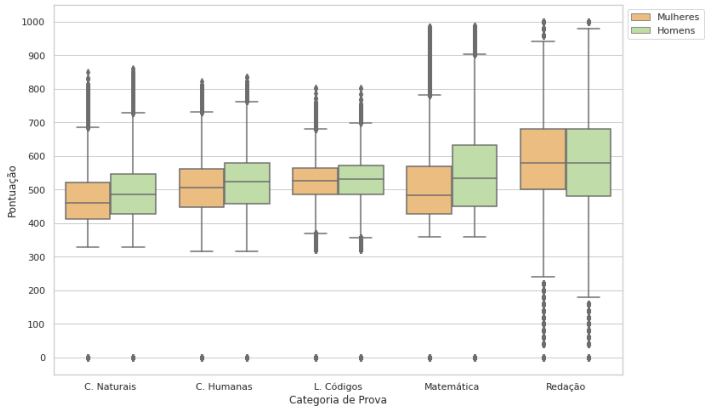
Esta seção analisa o desempenho dos participantes por sexo. Na Base 1, existe um total de 3.701.947 entradas, onde 2.201.184 dessas entradas correspondem a participantes do sexo feminino e 1.500.763 correspondem ao sexo masculino. Na Figura 6, é possível visualizer a distribuição dos candidatos por sexo.

Figura 6 – Distribuição dos participantes do ENEM 2019 por sexo



Os participantes foram divididos em dois grupos, um correspondendo ao sexo feminino e outro ao sexo masculino. A Figura 7 mostra uma representação em *boxplot* das notas nas cinco competências avaliadas no exame.

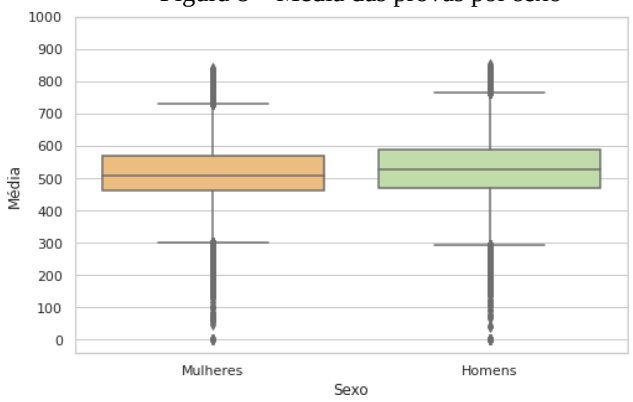
Figura 7 – Desempenho dos participantes por sexo nas cinco provas



É perceptível que nas competências de ciências naturais, ciências humanas e linguagens e códigos os homens tiveram um desempenho levemente superior de forma geral. Para se ter uma noção mais precisa, foi feito o cálculo da média por sexo em cada prova. Em ciências naturais os homens têm uma média 4,12% superior em relação às mulheres, em ciências humanas uma superioridade de 2,71% e em linguagens e códigos uma média 0,42% maior. Em Matemática, pode-se observar que a nota dos participantes do sexo masculino obteve um ápice mais destacável em comparação ao sexo feminino, tendo uma média 8,27% maior. Já na nota da Redação, as mulheres obtiveram uma média 2,52% superior a dos homens.

Para uma visualização mais direta, foi utilizado o cálculo da média aritmética das cinco disciplinas contempladas no Enem. Na Figura 8 aparece um *boxplot* que agrupa os candidatos por sexo e referencia o conjunto de médias que cada um desses grupos obteve no exame.

Figura 8 – Média das provas por sexo

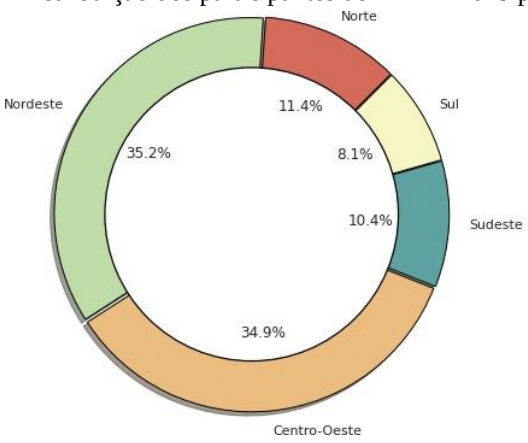


Analisando o *boxplot* com a média das cinco competências do exame, é perceptível que o sexo não é um fator determinante para o desempenho de forma geral do candidato.

Região

Nesta seção é feita a análise da distribuição e desempenho dos participantes pela região em que os mesmos residem. Na Figura 9 é possível visualizer a distribuição de participantes do ENEM 2019 por região do Brasil.

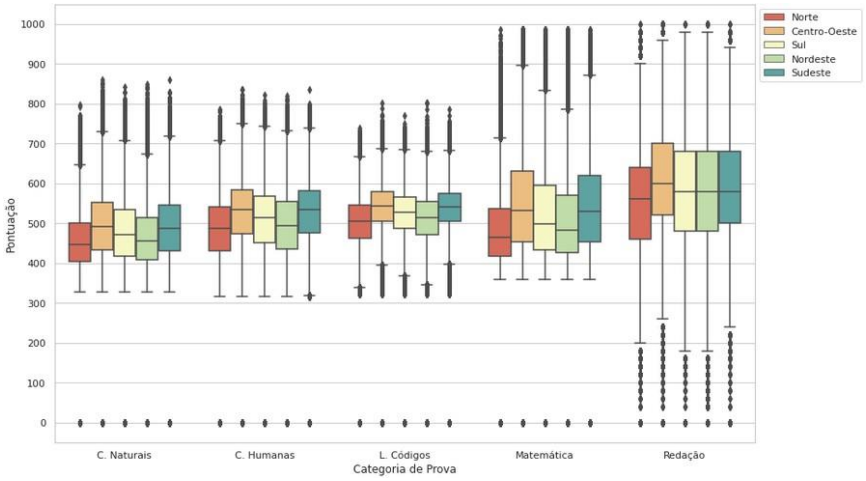
Figura 9 – Distribuição dos participantes do ENEM 2019 por região



Pode-se notar algumas peculiaridades ao analisar essa distribuição. A região Centro-Oeste é a segunda mais representativa em número de participantes, mesmo sendo a região menos densa populacionalmente entre as cinco regiões, de acordo com último censo populacional realizado pelo IBGE (2008). Isso pode ser um reflexo da facilidade de acesso que os estudantes dessa região tem a educação, e pode servir como ferramenta de estudo para melhorias nas demais regiões do país. Um paralelo semelhante ocorre com a região Sudeste. Mesmo tendo o maior número de habitantes do Brasil, é a segunda menos representativa em número de candidatos no ENEM 2019. Isso também proporciona um cenário de estudo para a compreensão da pouca adesão desses alunos em relação as demais regiões.

Em seguida foi feita a análise da nota dos participantes de cada região nas cinco competências do exame. A Figura 10 mostra através de *boxplots* as notas de cada região pela categoria da prova.

Figura 10 – Distribuição de notas por região



Analisando o boxplot da Figura 10, pode-se notar que o Centro-Oeste obtêm a maior mediana em todas as provas, com exceção de ciências humanas, onde fica atrás do sudeste por apenas 0,5 pontos. Também é observável que o norte é detentor da menor mediana das notas dentre as regiões, em todas as provas. Para uma melhor visualização, a Tabela 2 mostra o valor da mediana em cada prova por região.

Tabela 2 – Mediana das notas em cada prova por região

Prova	Centro-Oeste	Sudeste	Sul	Nordeste	Norte
Ciências Naturais	492,2	487,8	470,2	454,9	446,3
Ciências Humanas	533,5	534,0	512,8	494,5	486,1
Linguagens e Códigos	543,5	541,3	527,5	514,2	505,8
Matemática	533,0	529,5	499,1	482,4	464,2
Redação	600,0	580,0	580,0	580,0	560,0

O Índice de Desenvolvimento da Educação Básica (IDEB) foi criado em 2007 pelo Inep com o objetivo de mensurar a qualidade do aprendizado do país e estabelecer metas para melhorias na educação. O IDEB é calculado a partir da taxa de rendimento escolar (aprovação) e médias de desempenho nos exames aplicados pelo INEP. Os índices de aprovação são obtidos a partir do Censo Escolar, que é feito anualmente (MINISTÉRIO DA EDUCAÇÃO, 2018).

No sítio do IDEB é possível encontrar todos os resultados do índice de 2005 até 2019, contendo os indicadores referentes a 4.ª série / 5.º ano, 8.ª série / 9.º ano e 3.ª série. O sítio pode ser acessado através de: http://ideb.inep.gov.br. Como este trabalho faz um estudo do ENEM 2019, foi considerado os dados do IDEB 2019 referente a terceira série do ensino médio. Foi coletada a nota de cada estado e calculada a média do IDEB por região. A Tabela 3 mostra o resultado obtido por cada região.

Tabela 3 – Resultado IDEB 2019 (3.ª série do ensino médio)

Região	IDEB 2019
Sudeste	4,42
Sul	4,36
Centro-Oeste	4,27
Nordeste	3,92
Norte	3,78

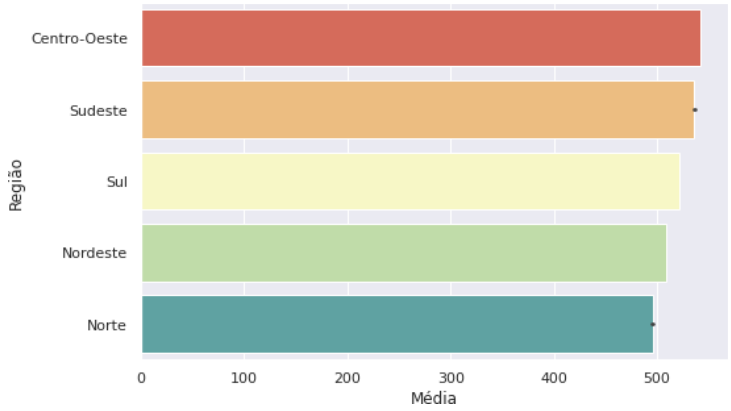
Apesar do Centro-Oeste obter os melhores resultados na prova do ENEM, ele é o terceiro colocado no resultado do IDEB. Já as regiões Nordeste e Norte apresentam um resultado bem fidedigno ao índice. Enquanto o Nordeste foi o penúltimo colocado nos resultados do ENEM e do IDEB, o Norte foi o último colocado em ambos.

Como visto na Figura 10 e na Tabela 2, é notável que o Centro-Oeste tem os melhores resultados, chegando a ter uma diferença de 68,8 pontos do último colocado em matemática. Tendo em vista o desempenho da região Centro-Oeste, outra oportunidade de análise se apresenta. É possível que órgãos e gestores públicos desenvolvam estudos das características de ensino da região que justifiquem esse desempenho superior no exame. O estudo do caso contrário também é válido. A busca por características

em comum do ensino no Norte pode apresentar indicadores do porque ele tem o pior desempenho nas provas dentre as demais regiões.

A Figura 11 mostra a média das cinco provas distribuídas por região. Como visto anteriormente na distribuição individual de cada disciplina, era esperado que a mesma classificação se seguisse. Tem-se: o Centro-Oeste com a maior média das notas, seguido - em ordem, pelo Sudeste, Sul, Nordeste e Norte.

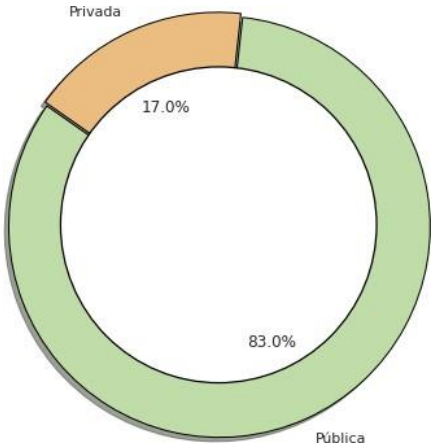
Figura 11 – Média das notas por região



Escola

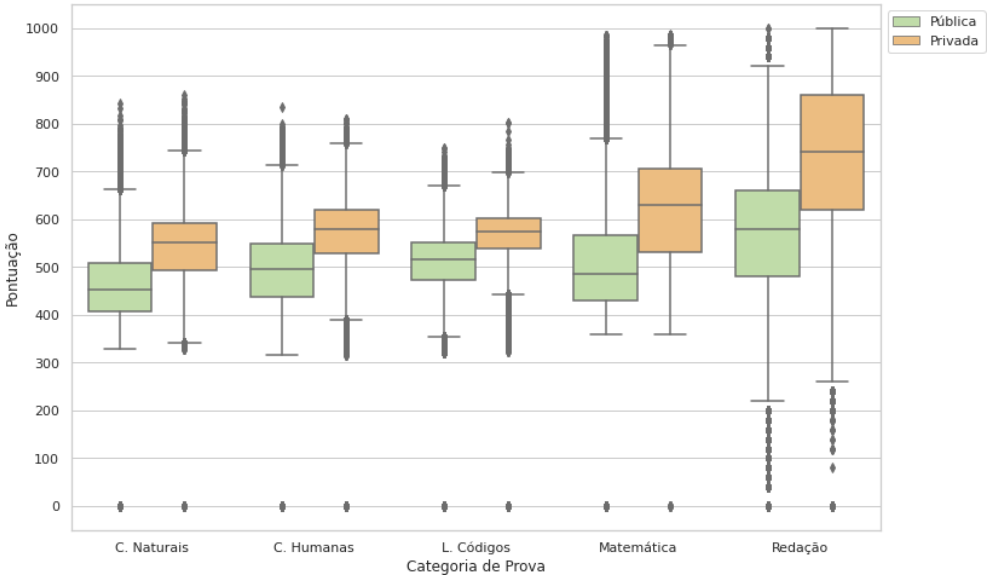
Nesta seção, a distribuição e o desempenho dos participantes pelo tipo de escola em que foi concluído o ensino médio são analisados. Foi feita a exclusão das linhas da Base 1 em que o candidato não respondeu em qual escola cursou o ensino médio. Restaram as classificações: pública, privada e exterior. Em toda a base de dados não houve nenhuma ocorrência de escola do exterior, então a análise se deu entre as escolas públicas e privadas. Após esses recortes, tem-se 1.009.821 de alunos de escolas públicas e 207.299 de escolas privadas. A Figura 12 mostra a porcentagem de representatividade de cada uma dessas redes de ensino no ENEM 2019.

Figura 12 – Rede de ensino dos participantes do ENEM 2019



A seguir, foi feito o estudo das notas obtidas em cada competência para cada rede de ensino. O desempenho de cada escola é observável na Figura 13.

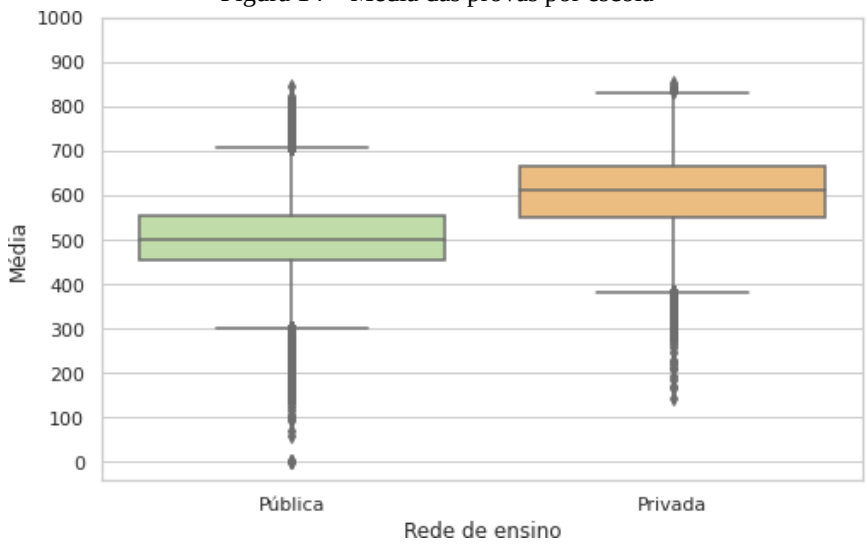
Figura 13 – Distribuição de notas por rede de Ensino



Como é possível reparar, a rede privada obtém os melhores resultados em todas as provas do exame. Foi feito o cálculo da média de cada escola para cada uma das provas contempladas no exame. Em ciências naturais, as escolas privadas obtiveram uma média 17,31% maior que as escolas públicas. Em ciências humanas uma média 15,36% superior. Na prova de linguagens e códigos obteve 11,18% a mais que a rede pública. Em matemática o resultado foi 22,90% superior. Por fim, em redação, as escolas privadas tiveram um resultado mais expressivo, com uma média 30,70% superior em relação as escolas públicas.

A Figura 14 mostra a representação em *boxplot* da média aritmética das cinco provas para cada grupo de escolas.

Figura 14 – Média das provas por escola



Como esperado, é possível ver um reflexo do que foi constatado na Figura 11. Ao analisar os *boxplots*, é notável que as notas mais baixas da rede privada tem quase o mesmo valor que as maiores notas das escolas públicas. O que atesta a grande diferença de desempenho entre as duas redes de ensino. Em média, a rede de ensino privada tem um desempenho 19,75% maior que a pública.

Raça Autodeclarada

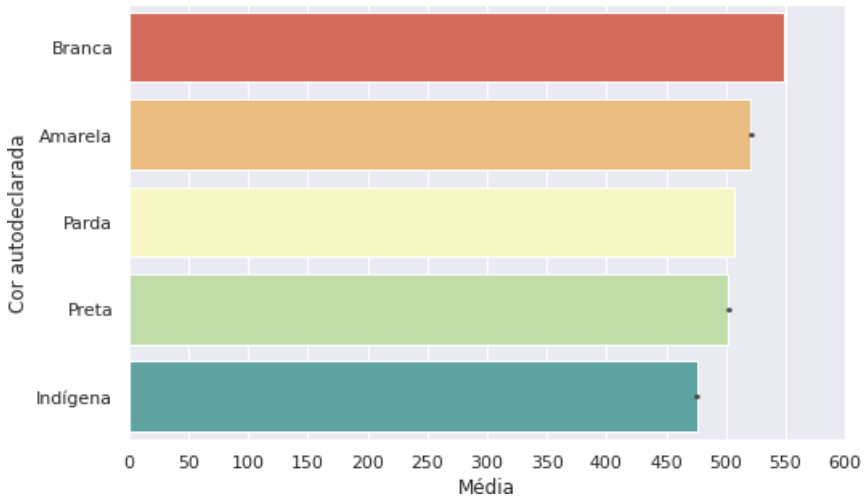
O estudo realizado nesta seção diz respeito à respeito da raça autodeclarada pelos participantes. A Tabela 4 mostra a quantidade e a porcentagem de participantes em cada grupo. As nomenclaturas raciais foram descritas da mesma forma que o formulário disponibilizado pelo INEP.

Tabela 4 – Distribuição dos participantes por raça autodeclarada

Raça	Quantidade	Porcentagem
Parda	1.694.136	45,8%
Branca	1.374.887	37,1%
Preta	453.218	12,2%
Amarela	84.752	2,3%
Não declarada	73.432	2,0%
Indígena	21.522	0,6%

A Figura 15 mostra a média geral obtida por cada grupo de raça. Os candidatos que não declararam nenhuma raça não foram incluídos no gráfico. Os participantes autodeclarados brancos são o segundo grupo mais representativo e são os que obtém as melhores notas. Em seguida vem, em ordem, os amarelos, pardos, negros e indígenas.

Figura 15 – Média das notas por raça autodeclarada



A Tabela 5 mostra a distribuição dessas raças pelo tipo de escola em que foi concluído o ensino médio. É observável uma certa tendência nos grupos por escolas. Por exemplo, os brancos têm uma porcentagem maior de alunos de escolas privadas que os amarelos. Os amarelos, por sua vez, tem uma porcentagem maior de instituições privadas que os pardos e assim suscetivamente. O valor da média das

raças segue essa tendência, quanto maior a representatividade de membros de escolas privadas, maior a média obtida. O que faz sentido levando em consideração a análise feita na seção anterior.

Tabela 5 – Tipo de escola por raça autodeclarada

Escola	Branca	Amarela	Parda	Preta	Indígena
Pública	72,8%	82,2%	89,8%	91,7%	94,3%
Privada	27,2%	17,8%	10,2%	8,3%	5,7%

Ao analisar em conjunto a Figura 15 e a Tabela 5, alguns resultados surpreendem. Por exemplo, os brancos obtiveram em média uma nota 9,16% maior que os pretos, o que não é uma diferença tão discrepante, considerando que 91,7% dos pretos cursaram o terceiro ano do ensino médio em escolas públicas. Como visto na seção anterior, o desempenho dos participantes vindo de escolas públicas costuma ser inferior.

Historicamente, os indígenas e negros são grupos raciais mais marginalizados, em diversos setores, principalmente em oportunidades de acesso à educação. Ao analisar a Figura 15 percebe-se que a diferença de desempenho entre as raças é existente, mas não abissal. Por exemplo, a maior diferença de desempenho é entre brancos e indígenas, onde os brancos possuem uma média de notas 15,36% superiores. Considerando que 94,3% dos indígenas são advindos de escolas públicas, essa diferença não é tão descomunal. O que abre uma oportunidade de estudo do quão melhor os resultados dessas raças menos favorecidas poderiam ser caso todos estivessem em pé de igualdade nas oportunidades de acesso à educação.

Escolaridade dos pais

Nos dados disponibilizados constam dados a respeito da escolaridade dos pais dos participantes. Essas informações são dispostas nas colunas Q001 e Q002, nessas colunas existem as seguintes possibilidades de respostas:

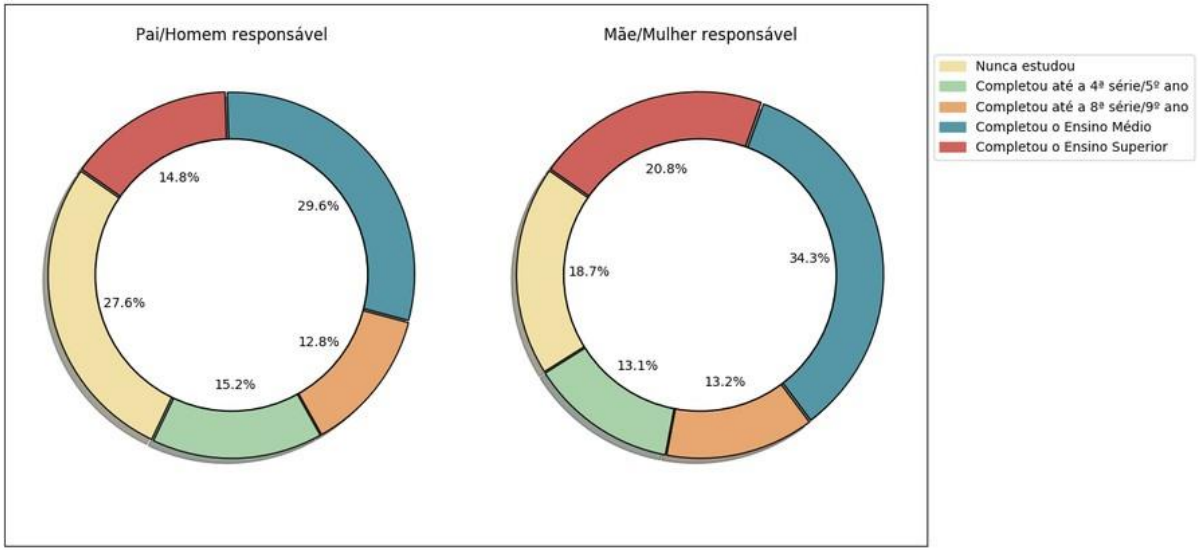
- **A** - Nunca estudou.
- **B** - Não completou a 4ª série/5º ano do Ensino Fundamental.
- **C** - Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
- **D** - Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
- **E** - Completou 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.

Foram desconsideradas as linhas do *dataframe* que continham a resposta H (403.808 linhas). As respostas A e B foram consideradas como sendo apenas uma: Nunca estudou. As respostas F e G foram condensadas em apenas uma, na forma de: Completou o Ensino Superior. Após os ajustes restam as seguintes possíveis respostas:

- Nunca estudou;
- Completou até a 4ª série/5º ano;
- Completou até a 8ª série/9º ano;
- Completou o Ensino Médio; e
- Completou o Ensino Superior.

A Figura 16 mostra a distribuição do grau de escolaridade dos responsáveis dos candidatos do ENEM 2019.

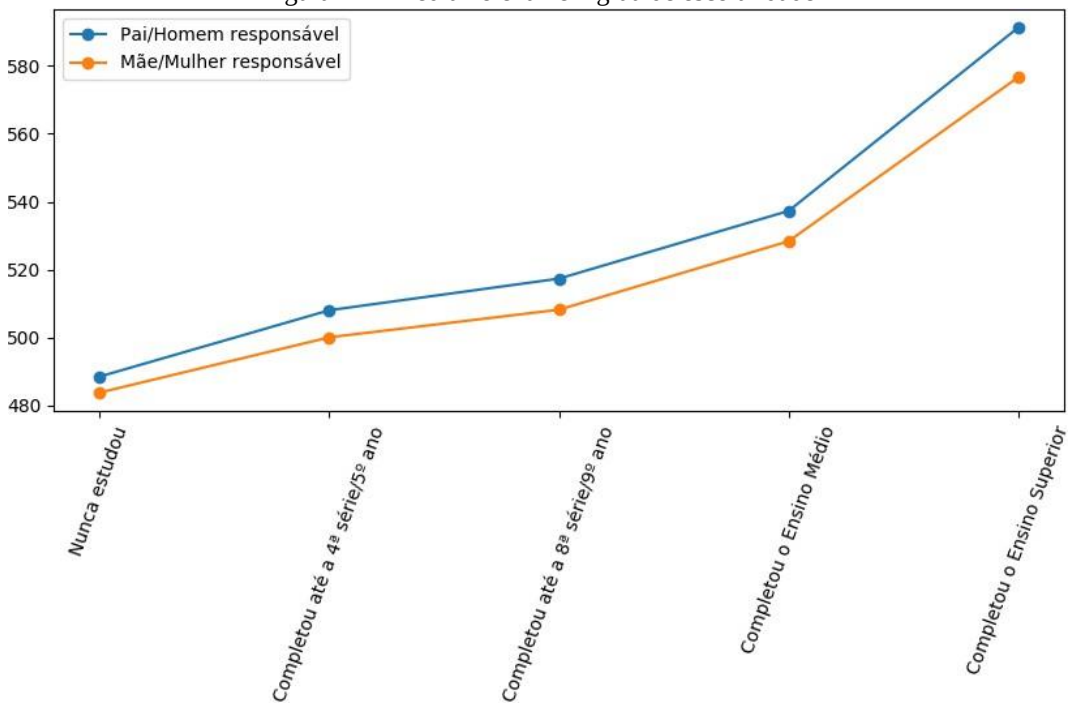
Figura 16 – Distribuição da escolaridade dos responsáveis



É possível notar que há uma diferença considerável entre os responsáveis. Nos graus mais elevados de escolaridade é perceptível que o grupo feminino é mais representativo. Enquanto as porções de pessoas com menor grau de escolaridade são mais representadas por participantes do sexo masculino.

A questão subsequente é se a escolaridade dos responsáveis afeta o desempenho dos participantes. A Figura 17 mostra a média geral no exame de acordo com o grau de escolaridade dos pais.

Figura 17 – Média no exame x grau de escolaridade

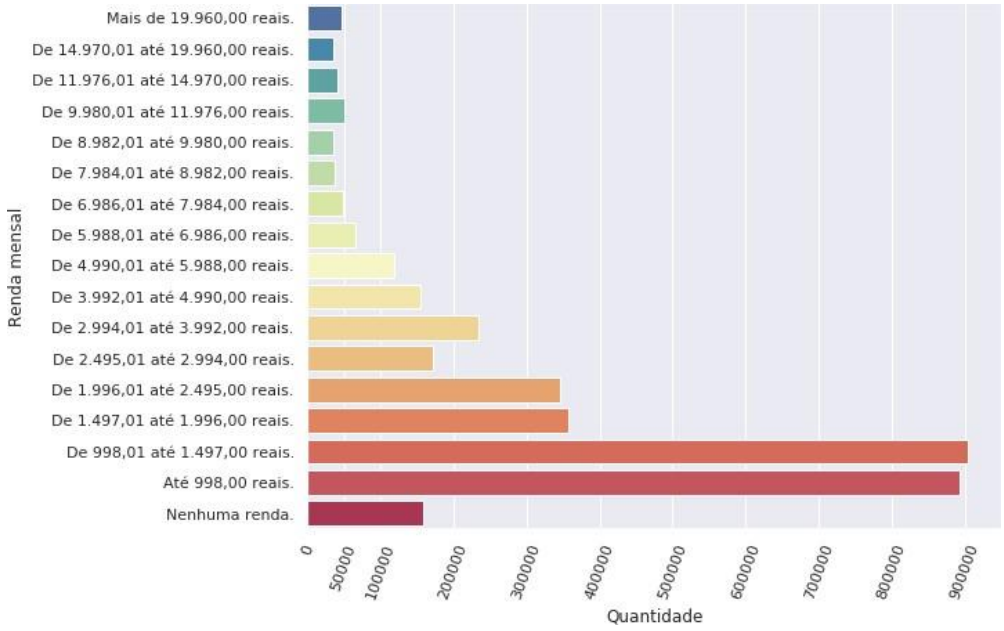


É perceptível que o grupo de estudantes com pais que possuem maior grau de escolaridade tem maiores médias geralmente. É notável que para ambos os sexos há uma tendência de quanto maior o grau de ensino do responsável, maior a nota obtida pelo candidato.

Renda

Esta seção é destinada à análise da renda informada pelos participantes e sua influência no desempenho dos mesmos. A Figura 18 mostra a quantidade de participantes por perfil econômico.

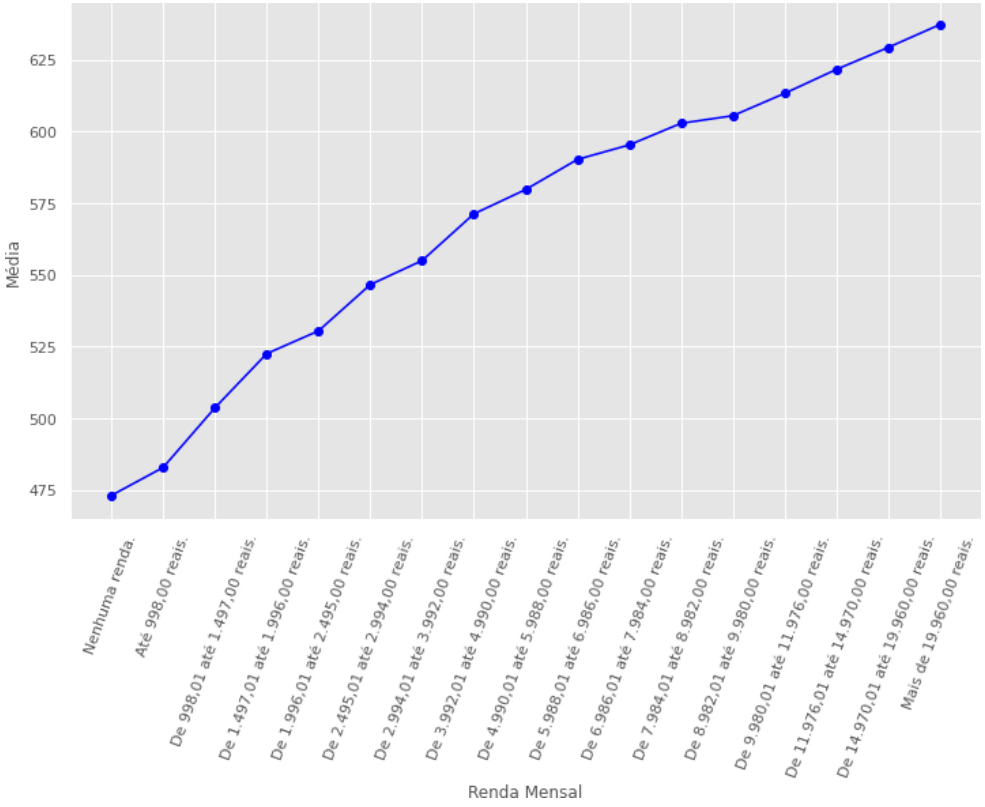
Figura 18 – Quantidade de participantes x renda mensal



Sessenta e dois por cento dos participantes tem uma renda de no máximo R$ 1996,00. Os participantes com maior poder aquisitivo são os que representam as menores porções de candidatos, o que é um reflexo da realidade social brasileira.

A Figura 19 mostra que os participantes com maior poder aquisitivo obtiveram médias melhores que os demais no geral. Porém, essa é uma representação isolada da média no exame de acordo com cada grupo de renda. Para uma averiguação precisa da correlação entre a renda e o desempenho dos participantes, seria preciso dados numéricos diversos da renda dos participantes, mas os dados disponibilizados pelo INEP são dados categóricos. Ao responder o questionário socioeconômico, os participantes não informam o valor exato da renda familiar, é selecionado um intervalo de valores em que sua renda se enquadra. O que torna esse dado categórico e inviabiliza o estudo de correlação entre o desempenho e o poder aquisitivo do participante.

Figura 19 – Média no ENEM 2019 x renda mensal



Acesso à tecnologia

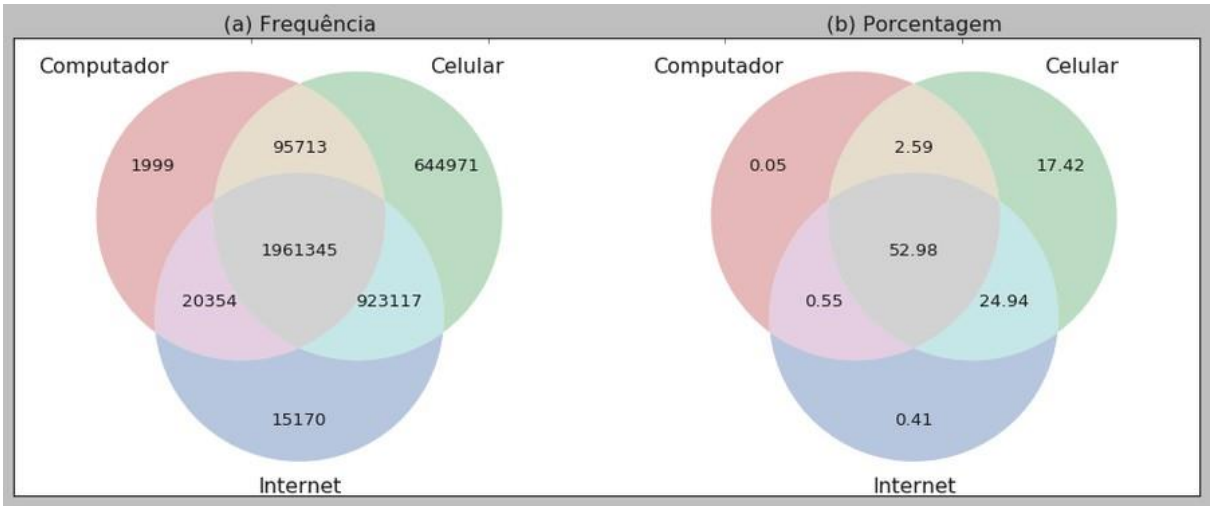
Com a internet cada vez mais acessível, muitas atividades que antes eram feitas por meios físicos, são executadas de forma online. Uma dessas atividades é o estudo. Atualmente tem-se uma vasta gama de conteúdos e cursos digitais que podem auxiliar na jornada de aprendizado. Esta seção faz um estudo das tecnologias que os participantes do ENEM 2019 tem à disposição.

Nos dados disponibilizados pelo INEP, as colunas Q022 e Q024 são referentes, res- pectivamente, a posse de telefone celular e computador pelos participantes. Essas colunas têm quatro possíveis respostas,

uma para caso não possua nenhum desses aparelhos e as demais para detalhar a quantidade. Para o estudo desta seção, foi considerado apenas se o participante tem ou não o equipamento, desconsiderando as quantidades caso tenha. Por fim, a coluna Q025 refere-se ao acesso à internet por parte do candidato.

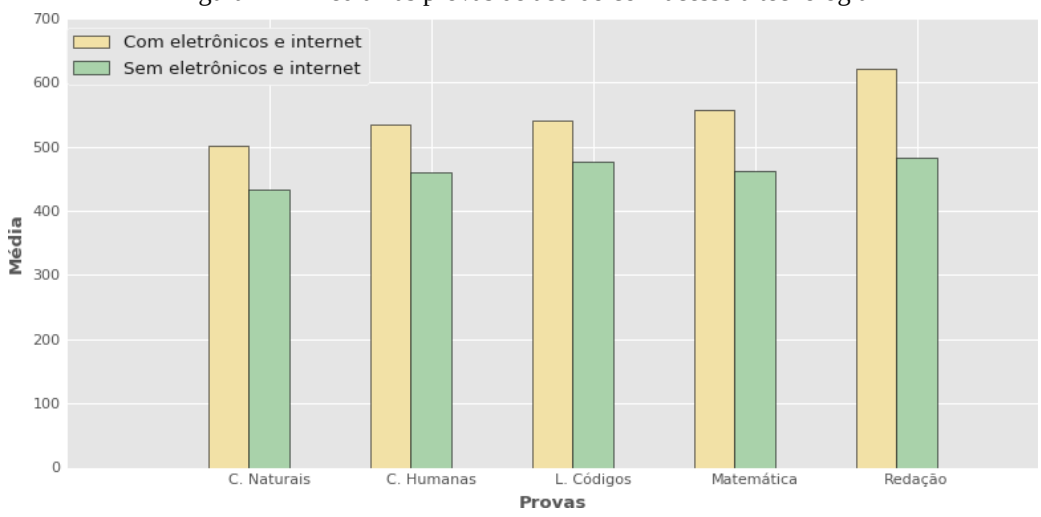
Com fins de visualização, foram criados três grupos: Computador, Celular e Internet. A Figura 20 contém um diagrama de Venn que mostra a quantidade e porcentagem de participantes alocados por grupos.

Figura 20 – Diagrama de Venn



Apenas 39.278 participantes (1,06%) informaram não ter acesso a nenhuma dessas três tecnologias. Apesar de ser uma parte pequena do conjunto, é interessante averiguar se esses participantes obtêm um desempenho semelhante aos demais. Foi feito um comparativo entre os candidatos que possuem acesso as três tecnologias e entre os que não possuem acesso a nenhuma delas. A Figura 21 mostra a média geral obtida em cada uma das provas do exame por esses grupos.

Figura 21 – Média nas provas de acordo com acesso à tecnologia



De uma forma geral, os participantes que tem acesso à eletrônicos e internet tem uma média superior aos demais. Essa diferença atinge o maior valor na média da nota da redação, onde chega a ser 28,58% superior. Noventa e cinco por cento dos participantes que não possuem computador, celular e acesso à internet, tem uma renda de no máximo R$ 1.497,00. Dessa forma, nota-se a importância de um espaço público e gratuito para que alunos que não tenham acesso a essas ferramentas de aprendizado possam se inserir no ambiente de estudo virtual.

Acesso à tecnologia

O ENEM tem um número volátil de participantes ao longo dos anos. Nesta seção é feito o uso de técnicas de *Machine Learning* com os dados referentes a quantidade de participantes por ano para obter uma linha de tendência referente aos inscritos no exame nos próximos anos.

A Tabela 6 contém a quantidade de participantes dos exames de 1998 até 2019. Esses dados foram disponibilizados pelo INEP (2021).

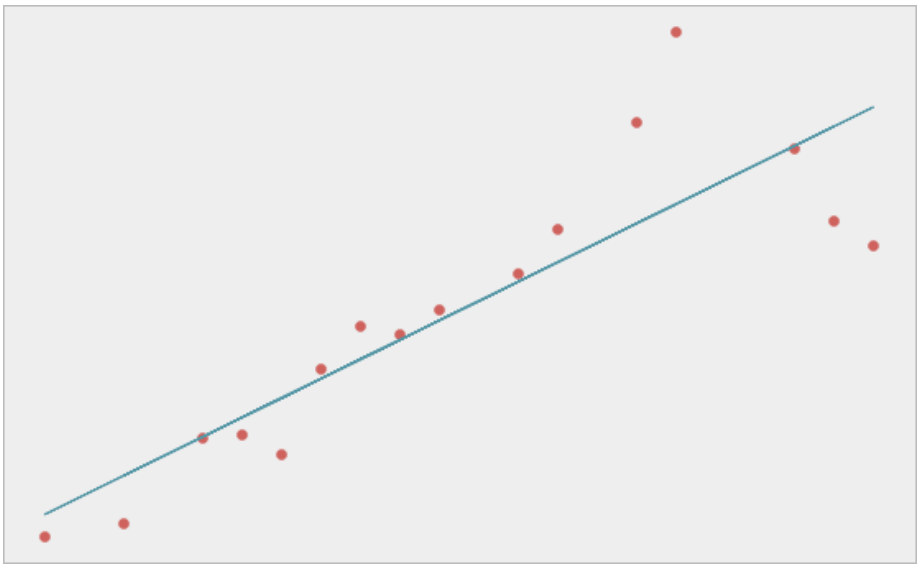
Tabela 6 – Ano x quantidade de participantes

Ano do ENEM	Quantidade de Pessoas	Ano do ENEM	Quantidade de Pessoas
1998	157.221	2009	4.148.720
1999	346.953	2010	4.626.094
2000	390.180	2011	5.380.856
2001	1.624,131	2012	5.791.065
2002	1.829.170	2013	7.173.563
2003	1.882.393	2014	8.722.248
2004	1.552.316	2015	7.746.472
2005	3.004.491	2016	8.627.367
2006	3.742.827	2017	6.731.341
2007	3.584.569	2018	5.513.747
2008	4.018.050	2019	5.095.270

Para obter uma linha de tendência, foi preciso treinar e testar os dados. Para desenvolver este estudo, foram utilizadas as bibliotecas Pandas, Numpy, Matplotlib e Sklearn, todas do Python. Foi necessário importar as bibliotecas necessárias, criar o *DataFrame*, dividir os dados de teste e os de treino, criar o modelo de regressão, treinar o modelo e, por fim, gerar a figura com a linha de tendência após o treino.

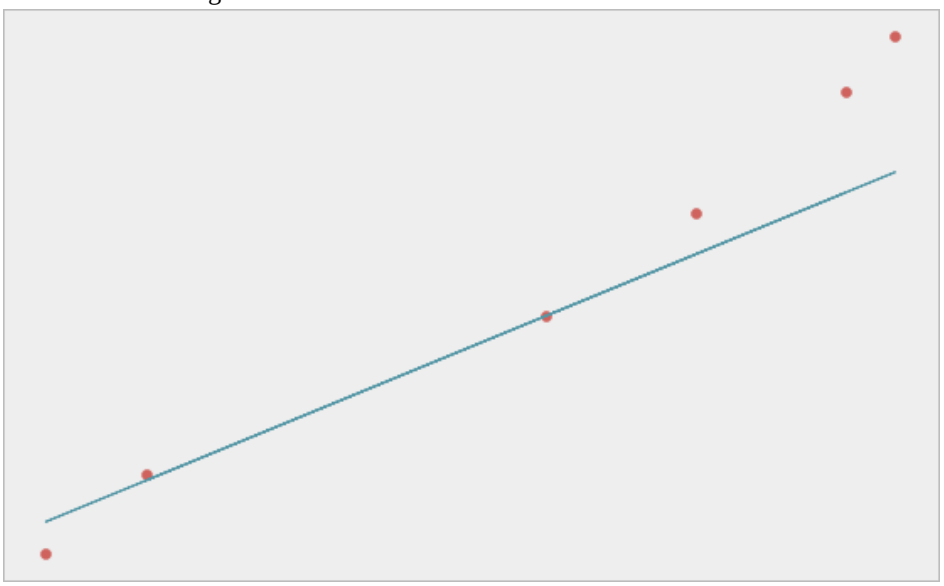
Por padrão, foram reservados 75% dados para treino e 25% para teste. A quantidade de dados disponíveis para este estudo não é tão grande, visto que só existiam dados do ENEM a partir do ano de 1998. O modelo preditivo criado faz uso de regressão linear para obter uma linha de tendência referente ao número de inscritos por ano. Onde o número de participantes por ano é a variável dependente e o ano de realização do exame é a variável independente. A Figura 22 mostra a linha obtida com os dados de treino utilizados.

Figura 22 – Linha de tendência obtida com os dados de treino



O treinamento gerou uma linha crescente. Para dar continuidade no estudo, foram aplicados os dados de teste para averiguação do resultado. Com isso, obtém-se a Figura 23, que apresenta o resultado com os dados de teste.

Figura 23 – Linha de tendência com os dados de teste



É notável que tanto os dados de treino quanto os de teste, geraram uma linha de tendência crescente, o que indica que o número de participantes no ENEM, nos próximos anos, tende a crescer. Para verificar se essa tendência se manteria em outros testes, foi colocado o modelo em execução com os anos de 2021 e 2025. Para o ano de 2021 o modelo resultou em um total de 8.089.014 de participantes e para o ano de 2025 um total de 9.402.842 de inscritos, o que mostra que o modelo segue a tendência crescente que o mesmo previu. Esta é uma análise puramente numérica, o modelo não conta com variáveis como o surgimento da pandemia, que resultou em uma menor adesão às últimas edições do exame.

5 CONCLUSÃO

As desigualdades sociais, em qualquer país, tem grande influência na qualidade da educação que a população tem acesso. Com isso, a análise do desempenho escolar sob um olhar socioeconômico é de grande importância para que se criem debates acerca das diferenças sociais que influenciam na manutenção dessas desigualdades.

Este trabalho está enquadrado na aplicação dos conceitos de ciência de dados em dados educacionais brasileiros. Teve por objetivo a aplicação de técnicas estatísticas e da ciência de dados para a descoberta de conhecimento sobre os participantes do Exame Nacional do Ensino Médio 2019, na base de dados disponibilizada pelo INEP.

O resultado dos dois primeiros estudos permitiu concluir que tanto a idade quanto o sexo dos participantes não são fatores determinísticos para o desempenho dos mesmos. O estudo subsequente mostrou que a região Centro-Oeste detêm os melhores resultados no exame e é a segunda mais representativa em número de participantes, mesmo sendo a região menos densa populacionalmente do país. Foi perceptível também que as regiões Nordeste e Norte ocupam penúltimo e último lugar, respectivamente, em todas as competências do exame.

A análise seguinte foi referente ao tipo de escola em que o participante concluiu o ensino médio. Foi possível verificar que os candidatos provenientes de escolas particulares obtiveram resultados superiores em todas as provas do Enem, com ênfase na redação, onde o grupo de alunos das escolas privadas ficaram com uma média 30,70% maior que os participantes da pública. Na média geral das cinco provas, as escolas privadas tiveram uma superioridade de 19,75% em relação às instituições públicas.

Em relação a raça autodeclarada dos participantes, foi perceptível uma diferença nas médias obtidas. A classificação dos grupos por média obtida foi: brancos, amarelos, pardos, pretos e indígenas. Apesar da diferença de médias ser existente, não foi uma diferença tão expressiva. Outra observação interessante, é que os grupos com maiores médias, tem uma maior representatividade de participantes vindo de escolas privadas em relação aos outros grupos.

Foi observável também que a escolaridade dos pais dos participantes tende a influenciar no resultado obtido nas notas. Quanto maior o grau de escolaridade dos responsáveis dos participantes, maior tende a ser a nota obtida no exame. Uma tendência semelhante foi observada na análise de renda dos participantes. Os candidatos foram agrupados por grupo de renda e foi visto que quanto maior o poder aquisitivo, maior a média obtida pelo grupo.

Outra análise realizada foi a de acesso à tecnologia por parte dos candidatos. Foi visto que participantes quem tem acesso à tecnologia como suporte para o estudo, tem uma média de notas maior que os participantes que não tem acesso à essa facilidade. Os candidatos com acesso à tecnologia obtiveram uma média 28,58% superior na nota da redação. Foi verificado 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* com os dados históricos de inscritos no ENEM, para obter uma linha de tendência em relação ao número de participantes ao longo dos anos. Apesar de ter sido utilizado um conjunto de dados pequeno, foi possível ter uma visão superficial da tendência de inscritos para os anos seguintes. Foi constatada uma linha de tendência crescente, o que indica que o número de pessoas inscritas tende a crescer nos anos seguintes. Análises como essa são importantes, pois, podem prever informações que podem auxiliar na realização de exames futuros, como questões de infraestrutura para adequação de todos os inscritos.

Após a realização do estudo, infere-se que a ciência de dados é uma maneira eficaz de estudar bases de dados massivas. Foi perceptível, de forma rasa, que as diferenças sociais esboçam uma influência no resultado das médias obtidas. Este trabalho faz uma análise inicial e superficial dos dados do ENEM, com o intuito de encontrar *insights* para realização de estudos futuros. Nota-se que é preciso se aprofundar na análise de aspectos socioeconômicos dos alunos, com o intuito de achar relações mais diretas.

Para trabalhos futuros, é pretendido complementar a base de dados do ENEM fazendo uso de 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* em mais dados históricos do exame, como o desempenho das raças ao longo dos anos, averiguar se vem crescendo ou decaindo e ter uma previsão para provas futuras.

REFERÊNCIAS

CETAX. (24 de abril de 2022). A diferença entre ciência de dados e análise de dados. [Blog post]. 2020. Recuperado de: https://www:cetax:com:br/blog/ciencia-de-dados-e-analise-de-dados/.

DESAFIOS DA EDUCAÇÃO. (29 de novembro de 2019). Educação: análise de dados pode identificar problemas e orientar ações. [Portal Eletrônico]. Recuperado de: https://desafiosdaeducacao:grupoa:com:br/ educacao-analise-de-dados/.

Gonçalves, P. (10 de setembro de 2018). Afinal, como se desenvolve um projeto de Data Science? [Blog post]. Recuperado de: https://medium:com/techbloghotmart/afinal-como-se-desenvolve-um- projeto-de-data-science-233472996c34.

IBM. 2021. Regressão linear: Gere previsões usando uma fórmula matemática facilmente interpretada. [Site oficial da IBM]. Recuperado de: https://www:ibm:com/br-pt/analytics/learn/linear-regression.

INEP. Exame Nacional do Ensino Médio (ENEM). 2021. [Site oficial do Governo Federal Brasileiro – Ministério da Educação]. Recuperado de: https://www:gov:br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. (2008*). Contagem da População 2007* (2ª. ed.). Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística – IBGE. ISBN 978-85-240-4004-7. Recuperado de: https://biblioteca:ibge:gov:br/visualizacao/livros/liv93420:pdf.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. 2021. Microdados. [Site oficial do Governo Federal Brasileiro – Ministério da Educação]. Recuperado de: https://www:gov:br/inep/pt-br/acesso-a-informacao/dados- abertos/microdados/enem.

MINISTÉRIO DA EDUCACÃO. 2018. Ideb - Apresentação. [Site oficial do Governo Federal Brasileiro – Ministério da Educação]. Recuperado de: http://portal:mec:gov:br/conheca-o-ideb.

OLIVEIRA, B. 2019. BOXPLOT: COMO INTERPRETAR? [Blog post]. Recuperado de: https://operdata:com:br/blog/como-interpretar-um-boxplot/.

VAZ, M. 2021. Programação Python (Parte 3) - Prof. MARCO VAZ. [Blog post]. Recuperado de: https://www:codingame:com/playgrounds/52723/programacao-python-parte-3---prof-- marco-vaz/pacote-pandas-dataframe.

Vickery, R. (30 de janeiro de 2021). 8 Fundamental Statistical Concepts for Data Science. Recuperado de: https://towardsdatascience:com/8-fundamental-statistical-concepts-for-data-science- 9b4e8a0c6f1c.