Sumário

Objetivos

Compreender o perfil da população que se inscreveu no Exame Nacional do Ensino Médio (ENEM) no ano de 2023.

Avaliar possíveis fatores sociodemográficos e/ou econômicos que podem estar associados a menor e maior nota. Usando como parâmetro de comparação os dados do perfil da população.

Perguntas

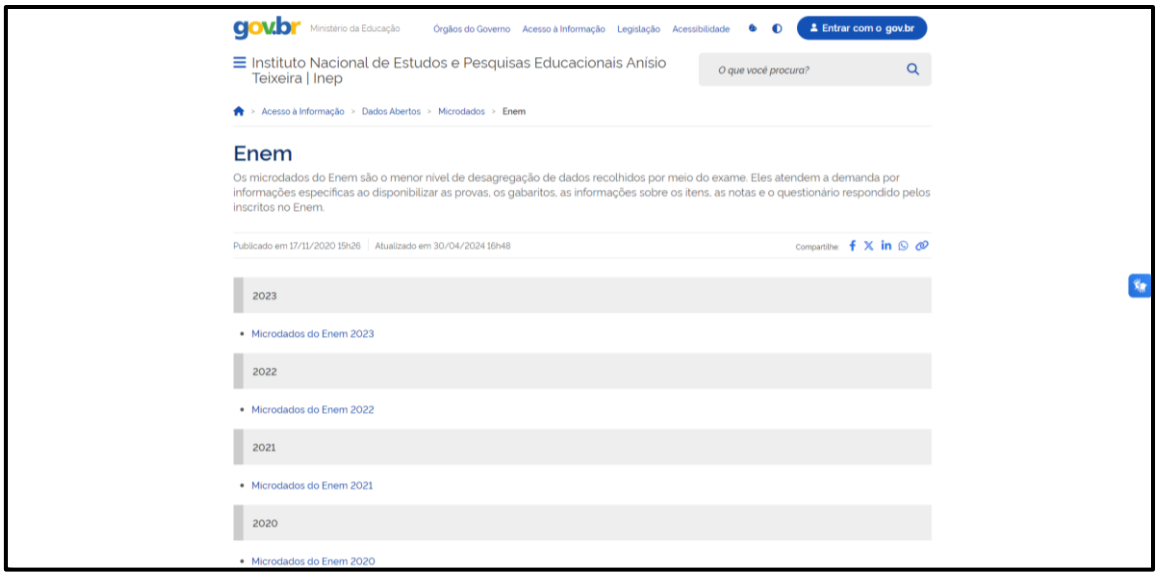
- Qual o perfil das pessoas que se inscreveram no ENEM em 2023?
- Há diferenças sociodemográficas entre aqueles que fizeram a prova e aqueles que perderam?
- Há diferença na pontuação entre estudantes por tipo de escola (i.e., Privada e Pública)?
- Qual a competência da redação é mais fácil (maior média), e mais difícil (menor média)?
- Quais as características gerais sociodemográficas entre aqueles que tiveram uma nota de 2 Desvios Padrões acima da média em:
    - Humanas?
    - Exatas?
    - Redação?
    - Matemática?
    - Linguagem e Códigos?

Busca pelos dados

O autor tem um interesse pessoal por dados educacionais. Para compreender na prática a importância da engenharia de dados e da linguagem SQL em grandes conjuntos de dados, o autor optou por buscar datasets mais densos. Nesse caso, o dataset escolhido foi referente aos Microdados abertos disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), mais especificamente os dados referentes ao Exame Nacional do Ensino Médio (ENEM) do ano de 2023.

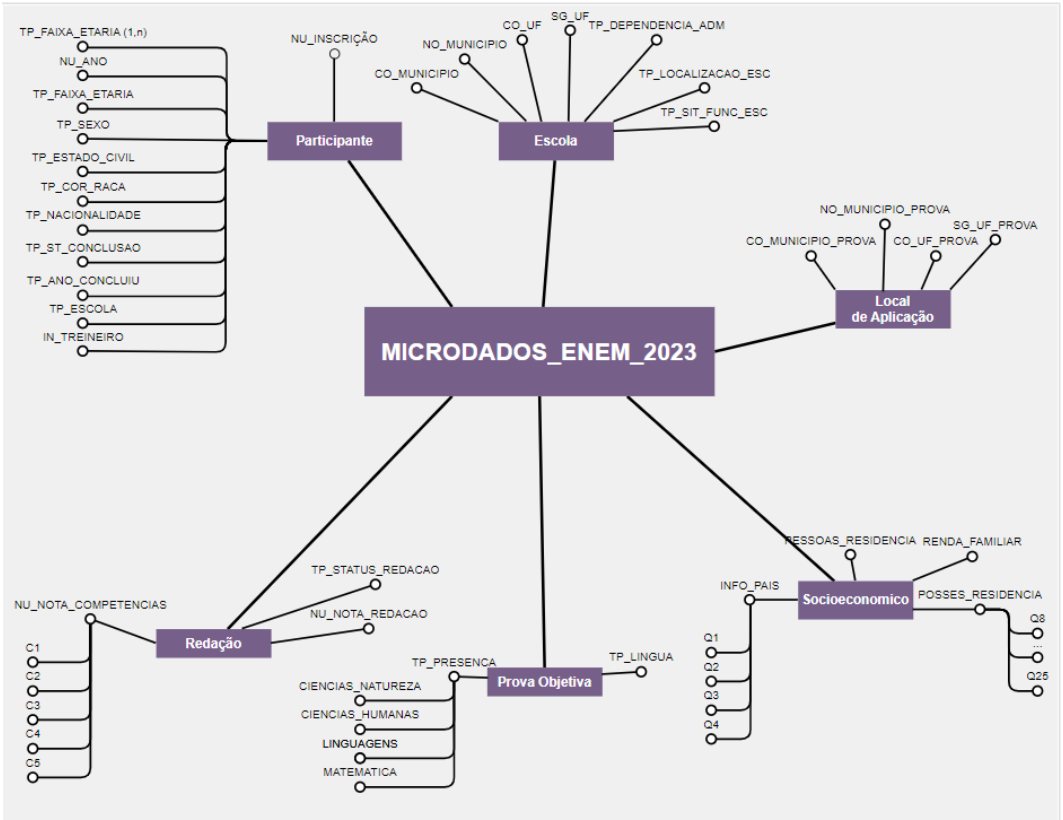
Coleta

As informações foram coletadas por meio do banco de dados educacionais de acesso aberto do Inep, denominado “Microdados ENEM”. Esse site possui dados dos exames de 1998 a 2023. Para este projeto, foi escolhido apenas os dados referentes ao exame mais recente.
https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem



Modelagem

Trata-se de uma tabela flat, porém, este conjunto de dados possui vários atributos (>50), estes podem ser agrupados por características específicas. Com isso em vista, uma representação gráfica foi elaborada para melhor compreensão das informações disponíveis no dataset.



NOTA: atributos com "..." são indicadores de sequência, onde "1, ..., 5" = "1, 2, 3, 4, 5"

1. Categoria dos dados

Os dados referentes ao ENEM 2023 são categorizados em:

a) **Informações dos inscritos**, todas categóricas, inclusive idade, que que é declarada em forma codificada (i.e., 1: 17 anos: 1, 2: 18 anos) e em intervalos (5: 30 a 32 anos).
b) **Informações da escola do inscrito**, estado da escola, unidade administrativa (ex: Municipal, Federal, Privada)
c) **O local de aplicação da prova**. Cidade, estado, região.
d) **Informações referentes ao exame**, separados entre as Provas objetivas (Ciências da Natureza, Matemática, Linguagens e Códigos e Ciência Humanas) e a Redação. Aqui temos informações sobre Presença, Gabarito do aluno, gabarito de correção e notas. As notas são os únicos dados numéricos. Tanto das provas objetivas (0 a 1.000) quanto da Redação, a nota da redação é dividida por nota total (0 a 1.000) e nota por competência (0 a 200).
e) Por último temos **mais dados sociodemográficos** do participante. O diferencial é que estes são mais associados a renda e posses da família do inscrito, sendo mais relevante para caracterização **socioeconômica**.

Vale ressaltar que, com exceção dos dados socioeconômicos (e) , todas as variáveis categóricas (nominais e ordinais) são codificadas em forma numérica (ex: Branco = 0, Pardo = 1, ...). Além disso, algumas variáveis que poderiam ser ordinais se tornaram nominais por conta a alternativa “Não respondeu” ou “Não informado” dentro das opções de resposta.

2. Dicionário/Catálogo de dados

2.1 Dados Gerais

Participante

Informações gerais referentes a dados sociodemográficos do participante. Como sexo, etnia, nacionalidade, tipo de escola (i.e., pública e privada)

Variável	Descrição	Tipo	Intervalo
NU_INSCRICAO	Identificador, nº de inscrição	Nominal	-
NU_ANO	Ano do ENEM, útil para comparações por ano	Contínua	-
TP_FAIXA_ETARIA	Idade, organizada em intervalos (ex:1 = <17; 11 = entre 26 e 30 anos)	Ordinal	<17 a >70
TP_SEXO	Sexo autodeclarado	Binária	F ou M
TP_ESTADO_CIVIL	Estado civil	Nominal	-
TP_COR_RACA	Etnia autodeclarada	Nominal	Branco, Pardo, ...
TP_NACIONALIDADE	Origem do inscrito	Nominal	Brasileiro, Brasileiro naturalizado, ... Estrangeiro
TP_ST_CONCLUSAO	Situação do Ensino Médio	Nominal	Concluído, cursando ... concluindo depois de 2023
TP_ANO_CONCLUIU	Ano de conclusão do Ensino Médio	Nominal	NI, 2022, ..., <2007
TP_ESCOLA	Tipo da escola	Nominal	NI, pública, privada
TP_ENSINO	Tipo de instituição em que concluiu o ensino médio	Binário	Regular, Educação Especial
TP_TREINEIRO	Realizou a prova para treinar	Binário	Não, sim

NI: Não informado

Escola

Informações sobre a escola do participante

Variável	Descrição	Tipo	Intervalo
CO_MUNICIPIO_ESC	Código do município da escola	Nominal	-
NO_MUNICIPIO_ESP	Nome do município da escola	Nominal	-
CO_UF_ESC	Código da Unidade da Federação	Nominal	-
SG_UF_ESC	Sigla da Unidade da Federação	Nominal	-
TP_DEPENDENCIA_ADM_ESC	Dependência Administrativa	Nominal	Federal, Estadual, Municipal, privada
TP_LOCALIZACAO_ESC	Tipo da localização	Binária	Rural, Urbana
TP_SIT_FUNC_FUNC	Situação de funcionamento	Numérica	Em atividade, ... paralisada

NI: Não informado

Local de prova

Informações sobre o local em que o participante realizou/realizaria o exame.

Variável	Descrição	Tipo	Intervalo
CO_MUNICIPIO_PROVA	Código do município da escola	Nominal codificada	1º dígito: Região 1º e 2º dígitos: UF 3º, 4º, 5º e 6º dígitos: Município 7º dígito: dígito verificador
NO_MUNICIPIO_PROVA	Nome da cidade da escola	Nominal	-
CO_UF_PROVA	Código de Unidade da Federação	Nominal	-
SG_UF_PROVA	Sigla da Unidade de Federação	Nominal	AL, SP, ..., AC

2.2 Dados referentes a prova

Prova objetiva

Informações sobre notas, presença, gabarito marcado pelo aluno, gabarito correto e língua escolhida (espanhol ou inglês). Com exceção do atributo “TP_LINGUA” todos possuem 4 variações, que são referentes a CN – Ciências da Natureza, CH – Ciências Humanas, MT – Matemática e LC – Linguagens e Códigos. Para evitar redundâncias no catálogo, o termo entre parênteses (área) será utilizado para representar as quatro áreas (CN, CH, MT, LC) em cada atributo.

Variável	**Descrição**	**Tipo**	**Intervalo**
TP_PRESENCA_(área)	Situação da presença durante o exame	Nominal	Faltou, presente, eliminado
CO_PROVA_(área)	Tipo da prova, indica adaptação	Nominal	Azul, amarela, ..., laranja (Braile)
NU_NOTA_(área)	Nota referente ao inscrito na área	Discreta	0 – 1000
TX_RESPOSTAS_(área)	Alternativas escolhidas pelo aluno	Nominal	A,B,C,D,E,D,E,A ...
TX_GABARITO_(área)	Alternativas corretas do exame	Nominal	A,B,C,D,E,D,E,A ...
TP_LINGUA	Língua estrangeira escolhida pelo inscrito	Binária	Inglês, Espanhol

Redação

Características da redação do aluno inscrito, apresenta a situação da redação até as notas por competências e nota final (somatório das competências).

Variável	Descrição	Tipo	Intervalo de resposta
TP_STATUS_REDACAO	Situação da redação do participante	Nominal	1) Sem problemas 2) Anulada 3)Cópia Texto Motivador 4) Em Branco 6)Fuga ao tema 7)Não atendimento ao Tipo textual 8) Texto insuficiente 9)Parte desconectada
NU_NOTA_COMP1	Demonstrar domínio da modalidade escrita formal da Língua Portuguesa	Contínua	0.0 – 200.0
NU_NOTA_COMP2	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.	Contínua	0.0 – 200.0
NU_NOTA_COMP3	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.	Contínua	0.0 – 200.0
NU_NOTA_COMP4	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação. Nota da competência	Contínua	0.0 – 200.0
NU_NOTA_COMP5	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.	Contínua	0.0 – 200.0
NU_NOTA_REDACAO	Soma das competências anteriores	Contínua	0.0 – 1000.0

2.3 Dados Socioeconômicos

Os dados socioeconômicos foram classificados pelo número da pergunta, indo de 'Q001' a 'Q025'.

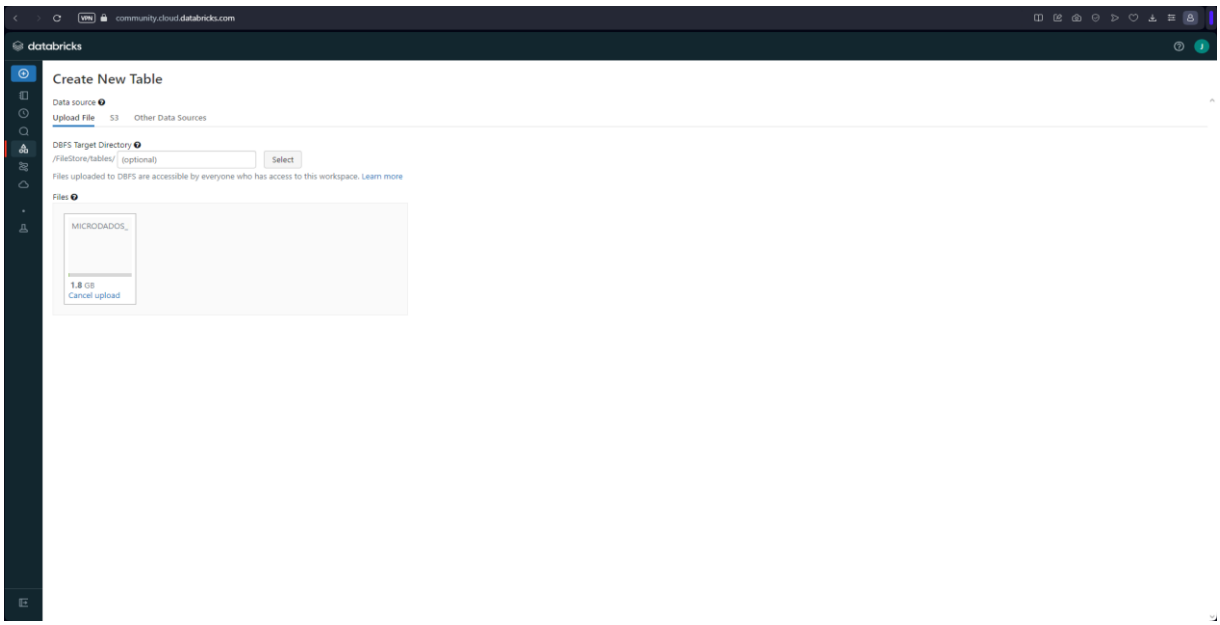
Variável	Descrição	Tipo	Intervalo de resposta
Q001 a Q004	Formação e Ocupação dos pais/responsáveis	Nominal	-
Q005	Quantidade de pessoas que moram com o inscrito	Discreta	1 – 20
Q006	Renda mensal familiar	Ordinal	a) (Nenhuma renda) até q) (Acima de R$:26.400,00)
Q007 a Q024	Posse de itens em casa (ex. quantidade de automóveis em casa, quantidade de geladeiras em casa)	Ordinal	0, 1, 2, >3
Q025	"Na sua residência tem acesso à internet?"	Binária	a) "Sim" e b) "Não"

3. Alinhagem dos dados

Como os dados vieram em uma tabela flat, então não foi realizada uma modelagem complexa dos dados.

Carga

Após o download do dataset via site do Inep, foi feito o upload dos dados na plataforma *databricks*.

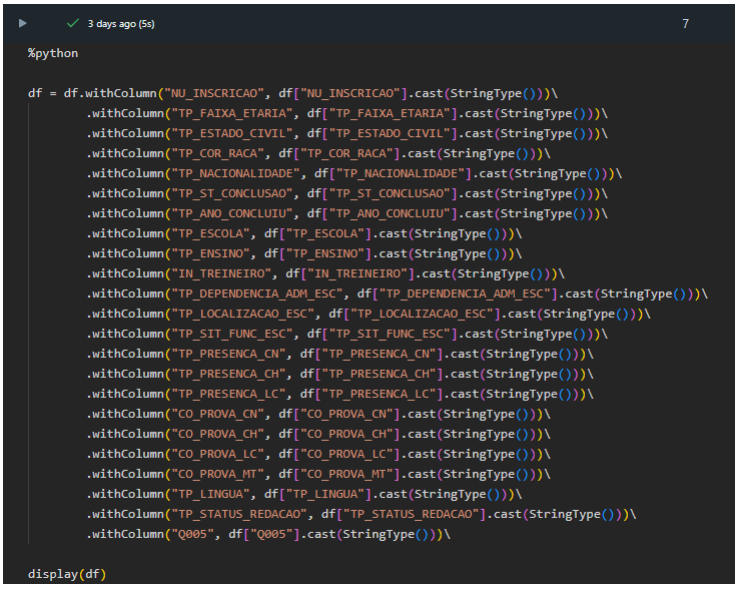


Realizado o upload, a biblioteca PySpark foi utilizada para criar o dataset

1. Pré processamento

1.1 Correção de tipologia dos dados

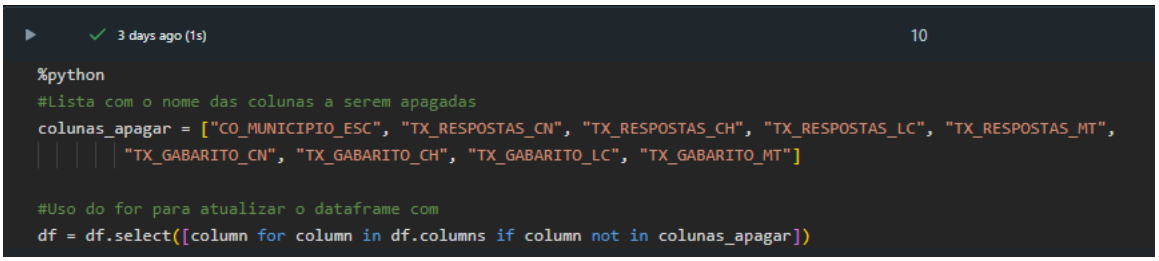
Alguns dados vieram com a tipologia incorreta, por conta da codificação do conjunto de dados citada anteriormente, (ex: 2 = 17 anos), classificando a maioria das variáveis categóricas como numéricas. Para solucionar este problema, foi usado o método '.cast' para corrigir este problema.



```
3 days ago (5s)                                                    7

%python

df = df.withColumn("NU_INSCRICAO", df["NU_INSCRICAO"].cast(StringType()))\
        .withColumn("TP_FAIXA_ETARIA", df["TP_FAIXA_ETARIA"].cast(StringType()))\
        .withColumn("TP_ESTADO_CIVIL", df["TP_ESTADO_CIVIL"].cast(StringType()))\
        .withColumn("TP_COR_RACA", df["TP_COR_RACA"].cast(StringType()))\
        .withColumn("TP_NACIONALIDADE", df["TP_NACIONALIDADE"].cast(StringType()))\
        .withColumn("TP_ST_CONCLUSAO", df["TP_ST_CONCLUSAO"].cast(StringType()))\
        .withColumn("TP_ANO_CONCLUIU", df["TP_ANO_CONCLUIU"].cast(StringType()))\
        .withColumn("TP_ESCOLA", df["TP_ESCOLA"].cast(StringType()))\
        .withColumn("TP_ENSINO", df["TP_ENSINO"].cast(StringType()))\
        .withColumn("IN_TREINEIRO", df["IN_TREINEIRO"].cast(StringType()))\
        .withColumn("TP_DEPENDENCIA_ADM_ESC", df["TP_DEPENDENCIA_ADM_ESC"].cast(StringType()))\
        .withColumn("TP_LOCALIZACAO_ESC", df["TP_LOCALIZACAO_ESC"].cast(StringType()))\
        .withColumn("TP_SIT_FUNC_ESC", df["TP_SIT_FUNC_ESC"].cast(StringType()))\
        .withColumn("TP_PRESENCA_CN", df["TP_PRESENCA_CN"].cast(StringType()))\
        .withColumn("TP_PRESENCA_CH", df["TP_PRESENCA_CH"].cast(StringType()))\
        .withColumn("TP_PRESENCA_LC", df["TP_PRESENCA_LC"].cast(StringType()))\
        .withColumn("CO_PROVA_CN", df["CO_PROVA_CN"].cast(StringType()))\
        .withColumn("CO_PROVA_CH", df["CO_PROVA_CH"].cast(StringType()))\
        .withColumn("CO_PROVA_LC", df["CO_PROVA_LC"].cast(StringType()))\
        .withColumn("CO_PROVA_MT", df["CO_PROVA_MT"].cast(StringType()))\
        .withColumn("TP_LINGUA", df["TP_LINGUA"].cast(StringType()))\
        .withColumn("TP_STATUS_REDACAO", df["TP_STATUS_REDACAO"].cast(StringType()))\
        .withColumn("Q005", df["Q005"].cast(StringType()))\

display(df)
```

1.2 Redução da dimensionalidade

As perguntas propostas não necessitam das informações de gabarito para serem respondidas. Para tornar as consultas mais eficientes, alguns atributos foram removidos do dataset.



```
3 days ago (1s)                                                    10

%python
#Lista com o nome das colunas a serem apagadas
colunas_apagar = ["CO_MUNICIPIO_ESC", "TX_RESPOSTAS_CN", "TX_RESPOSTAS_CH", "TX_RESPOSTAS_LC", "TX_RESPOSTAS_MT",
                  "TX_GABARITO_CN", "TX_GABARITO_CH", "TX_GABARITO_LC", "TX_GABARITO_MT"]

#Uso do for para atualizar o dataframe com
df = df.select([column for column in df.columns if column not in colunas_apagar])
```

Análise

1. Qualidade dos dados

Trata-se de um conjunto de dados estruturados em uma tabela flat, arquivo em csv previamente tratado e sem perda de informações, apesar da existência de nulls.

Pontos fortes

Os dados são de boa qualidade em maioria, os valores faltantes também passam informação em alguns casos. A exemplo temos os valores faltantes nas notas, indicando que o sujeito não obteve a nota por falta, atraso ou infração durante a prova. Caso tivesse errado todas as questões, o zero seria registrado.

Limitações

Alguns atributos possuíam como alternativa a ausência de resposta ao item (e.g., “prefiro não informar”), dificultando um pouco a caracterização da população em alguns casos. A exemplo temos o atributo referente ao tipo de escola, relevante para uma das perguntas propostas, onde 64,4% da população optou por escolher a alternativa “prefiro não responder”.

2. Solução do Problema

2.1 Introdução

Recapitulando, o problema em questão se refere a dois objetivos: (1) compreender mais a fundo o perfil da população de inscritos no Exame Nacional do Ensino Médio (ENEM) do ano de 2023, e (2) avaliar se existem diferenças relevantes entre subgrupos dessa população. Assim podemos ter um norte para avaliar fatores que poderiam estar associados a melhor ou pior desempenho na prova em comparação com o perfil geral. As perguntas propostas foram desenvolvidas com estes dois objetivos em mente.

Após a caracterização do perfil da população geral de inscritos, foram realizadas comparações entre grupos (ex. perfil de faltantes x perfil da pop. Geral). Como o perfil da população será o parâmetro de comparação, as diferenças serão avaliadas de duas formas:

a) Alteração na variável mais frequente (ex. na população geral, a maioria possui 18 anos, na população de escolas privadas a maioria possui 17 anos).
b) alteração na proporção da variável mais frequente (na população geral 50% dos inscritos são autodeclarados pardos, na população de escola pública, 70% são autodeclarados pardos, ou seja, diferença de 20% em proporção).

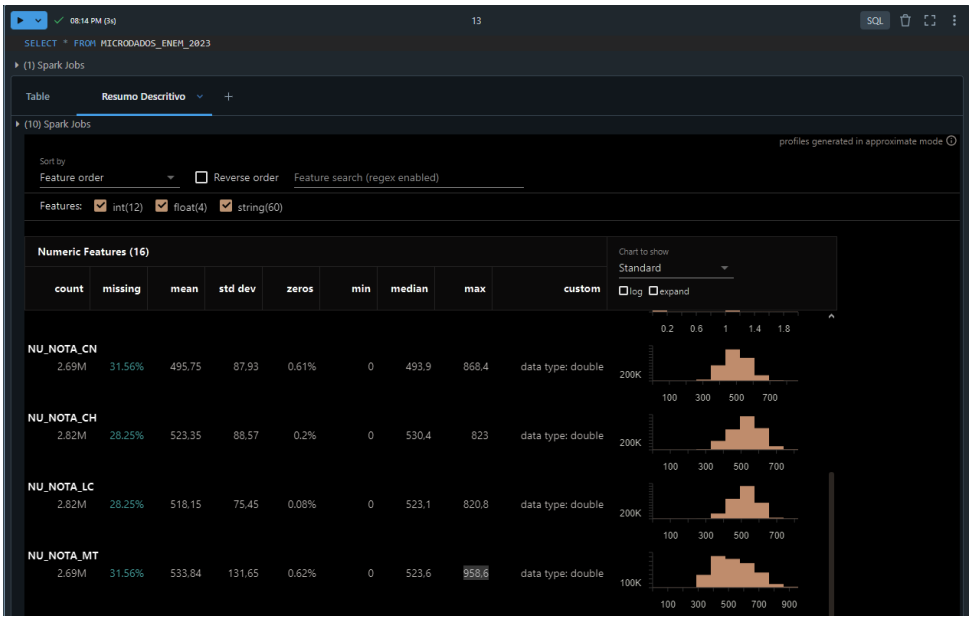
Vale ressaltar que as análises de diferenças entre grupos específicos em comparação com a população geral usarão valores ≥ 10% como ponto de corte para classificar uma diferença como significativa.

2.2 Perfil da população geral

Atributos específicos foram selecionados para compreender a amostra de forma mais específica, estes serão usados para fazer avaliação de diferença entre grupos. Os atributos foram selecionados com base na ideia de caracterização demográfica (etnia, idade, estado civil, sexo e tipo da escola que frequenta), caracterização econômica usando indicadores (Renda mensal familiar, ter ou não empregada e ter ou não internet). Por último, as notas em si, tanto das provas objetivas quanto o total da redação, totalizando 8 atributos sociodemográficos e 5 atributos referentes a nota. As notas são também informações descritivas para o objetivo 1, mas são a nossa variável dependente para o objetivo 2

As estatísticas descritivas gerais da população foram adquiridas por meio da aba “Data Profile” e por meio da visualização de após buscar todos os atributos da tabela mediante:

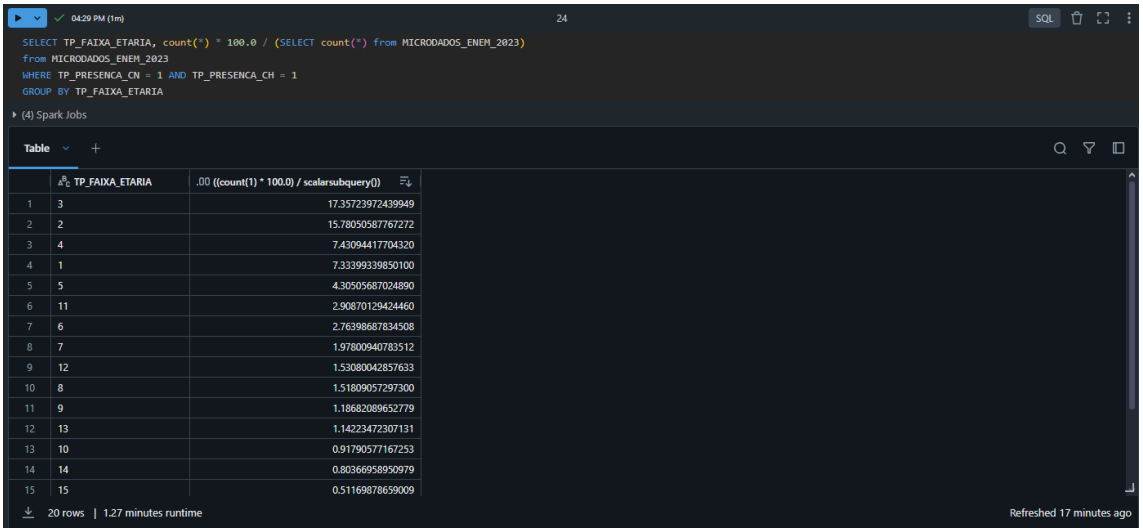
```
SELECT *
FROM MICRODADOS_ENEM_2023
```



Depois de compreender melhor a distribuição geral dos dados no conjunto, foi tomada a decisão de quais atributos seriam escolhidos para caracterizar a amostra, pois apesar da redução da dimensionalidade no pre processamento, ainda sobraram muitos atributos.

Uma alternativa para avaliar proporção de variáveis num dado atributo seria dividir pelo total, como no exemplo a seguir:

Neste exemplo podemos avaliar a proporção da faixa etária



A priori, temos os dados gerais da população. Estes dados servirão para comparação posterior entre subgrupos, sob o objetivo de avaliar possíveis diferenças. Os atributos escolhidos para caracterização são os que aparecem nas tabelas a seguir, referentes a informações sociodemográficas, socioeconômicas e de nota.

Área	Mais frequente	Proporção	Frequência
Faixa Etária	18 anos	23.01%	905K
Sexo	Feminino	70.4%	2.41M
Estado Civil	Solteiro(a)	88.76%	3.49M
Etnia	Parda	43.4%	1.71M
Tipo da escola	Pública*	29.65%	1.1M
Renda mensal Familiar	Até R$: 1.320	31.65%	1.25M
Tem empregada?	Não	91.8%	3.61M
Tem acesso à internet?	Sim	90.45%	3.56M

*M = Milhões; K = Milhares; * = Desconsiderando os que optaram por não relatar (64,38%, 2.5M)*

Tratando-se das notas por área de conhecimento, a maior quantidade de ausências foi nas provas do segundo dia (Ciências da natureza e Matemática), além disso a área do conhecimento com maior desvio no tipo de distribuição foi matemática, que apresentou uma distribuição assimétrica

Área	% sem realização	Média	% de Zeros	Maior nota
Ciências da Natureza	31,56%	495,75 (±87,93)	0.61%	868,4
Matemática	31,56%	533,84 (±131,65)	0.62%	958,6
Ciências Humanas	28,25%	523,35 (±88,57)	0.2%	823
Linguagens e Códigos	28,25%	518,15(±75,45)	0.08%	820,8

A Nota da redação será avaliada em seu total para comparação de perfis, porém um dado interessante foi a proporção zeros na competência 5, que se divergiu do restante. Esta competência se refere a "Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.".

Notas da pop. Geral (Redação)

Área	% sem realização	Média	Desvio Padrão	% de Zeros	Maior nota
Nota Geral	28,25%	617,8	±214,62	4,17%	1.000
Competência 1	28,25%	121,55	±35,65	4,18%	200
Competência 2	28,25%	139,3	±51,62	4,17%	200
Competência 3	28,25%	118,51	±43,41	4,19%	200
Competência 4	28,25%	129,8	±43,77	4,19%	200
Competência 5	28,25%	108,65	±61,6	**10,27%**	200

Resumo geral

Nota-se que a maior parte da população é autodeclarada parda () (Atributos específicos foram selecionados para compreender a amostra de forma mais específica, estes serão usados para fazer avaliação de diferença entre grupos. Os atributos selecionados foram

2.3 Há diferenças sociodemográficas entre aqueles que fizeram a prova e aqueles que perderam?

Aqui vemos os atributos sociodemográficos/econômicos daqueles que perderam as provas, foi utilizado o WHERE com os atributos que avaliam presença em provas que foram aplicadas em dias diferentes (Ciências da Natureza e Ciências Humanas). O total de pessoas que faltaram, em média, foi de aproximada mente 1,2M, de pessoas presentes foi aproximadamente 2,7M, estes valores foram retirados com base na média aritmética de faltas e presenças no primeiro e segundo dia. Também em média, 3.000 pessoas foram eliminadas durante a execução da prova. O enfoque será dado aos dois primeiros.

Linha em SQL para consulta:

```sql
SELECT TP_FAIXA_ETARIA, TP_SEXO, TP_ESTADO_CIVIL, TP_COR_RACA,
TP_ESCOLA, Q006, Q007, Q025
FROM MICRODADOS_ENEM_2023
WHERE TP_PRESENCA_CN = 0 AND TP_PRESENCA_CH = 0
```

```sql
SELECT TP_FAIXA_ETARIA, TP_SEXO, TP_ESTADO_CIVIL, TP_COR_RACA, TP_ESCOLA,
Q006, Q007, Q025
FROM MICRODADOS_ENEM_2023
WHERE TP_PRESENCA_CN = 1 AND TP_PRESENCA_CH = 1
```

Na tabela a baixo temos os dados sociodemográficos, com adição da coluna “Diferença” que aponta a proporção (em %) dos faltantes/presentes em comparação (subtraído) com a tabela geral. Amarelo indica que foi menor em comparação com o geral e verde indica maior em comparação com o geral.

Não houveram mudanças quanto a variável mais frequente. A diferença mais expressiva foi quanto a proporção de autodeclarados pardos, que no grupo presentes (62,28%) foi maior que na população geral (43,40%).

Área	Mais frequente	Proporção	Diferença (Geral)
Faltantes (média aritmética = 1.173.015)			
Faixa Etária	18 anos	17,04	-5,97
Sexo	Feminino	61,06	-9,34
Estado Civil	Solteiro(a)	82,88	-5,88
Etnia	Parda	47,45	4,05
Tipo da escola	Pública*	26,21	-3,44
Renda mensal Familiar	Até R$: 1.320	37,64	5,99
Tem empregada?	Não	94,34	2,54
Tem acesso à internet?	Sim	87,58	-2,87

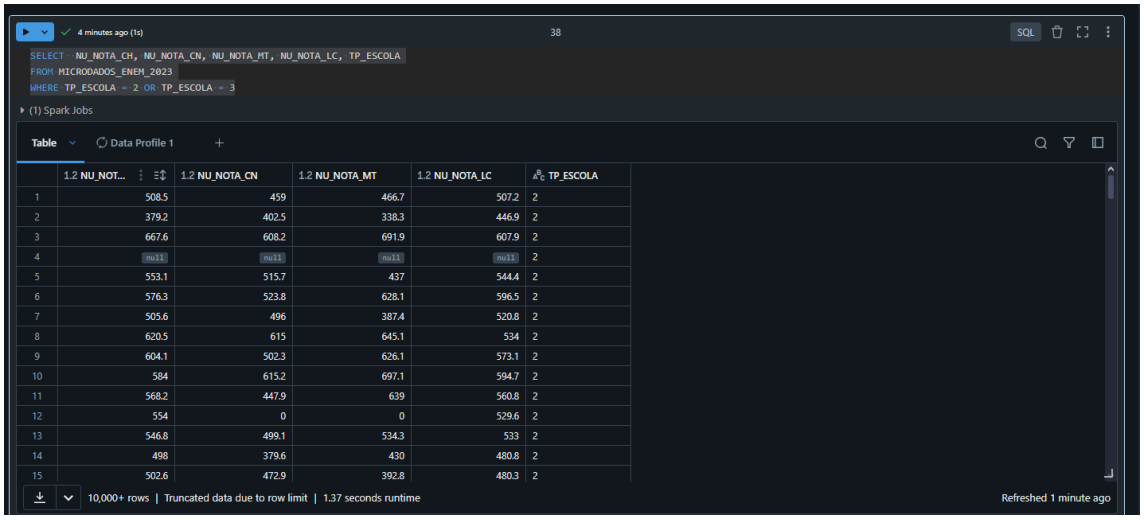
Presentes (média aritmética = 2.757.535)			
Faixa Etária	18 anos	17,4	-5,61
Sexo	Feminino	61,38	-9,02
Estado Civil	Solteiro(a)	88,89	0,13
Etnia	Parda	62,28	18,88
Tipo da escola	Pública*	30,97	1,32
Renda mensal Familiar	Até R$: 1.320	29,09	-2,56
Tem empregada?	Não	90,82	-0,98
Tem acesso à internet?	Sim	91,7	1,25

** = A instituição selecionada foi a segunda mais frequente, já que a mais frequente foi a ausência de resposta*

2.4 Qual a diferença de pontuação entre estudantes de escolas públicas se comparados a estudantes de escolas privadas?

A consulta para busca dos dados desses dois grupos foi a seguinte:

```
SELECT NU_NOTA_CH, NU_NOTA_CN, NU_NOTA_MT, NU_NOTA_LC, TP_ESCOLA
FROM MICRODADOS_ENEM_2023
WHERE TP_ESCOLA = 2 OR TP_ESCOLA = 3
```



	1.2 NU_NOT...	1.2 NU_NOTA_CN	1.2 NU_NOTA_MT	1.2 NU_NOTA_LC	TP_ESCOLA
1	508.5	459	466.7	507.2	2
2	379.2	402.5	338.3	446.9	2
3	667.6	608.2	691.9	607.9	2
4	null	null	null	null	2
5	553.1	515.7	437	544.4	2
6	576.3	523.8	628.1	596.5	2
7	505.6	496	387.4	520.8	2
8	620.5	615	645.1	534	2
9	604.1	502.3	626.1	573.1	2
10	584	615.2	697.1	594.7	2
11	568.2	447.9	639	560.8	2
12	554	0	0	529.6	2
13	546.8	499.1	534.3	533	2
14	498	379.6	430	480.8	2
15	502.6	472.9	392.8	480.3	2

Nota-se que quando comparamos o tipo de escola, algumas diferenças maiores surgem. Quando comparamos com a média da população de inscritos, vemos que temos mais pessoas de 18 anos em escolas públicas, o restante diverge, mas não de forma muito significativa. Enquanto os estudantes de escolas privadas parecem começar o ENEM já mais cedo, aos 17 anos.

A maioria (64,4%) optou por não declarar o tipo de escola que frequenta.

Distribuição sociodemográfica por tipo de escola

Área	Mais frequente	Proporção	Diferença
Pública (n = 1.166.540)			
Faixa Etária	18 anos	52,79	29,78
Sexo	Feminino	70,75	0,35
Estado Civil	Solteiro(a)	90,37	1,61
Etnia	Pardo(a)	46,57	3,17
Tipo da escola	-	-	-
Renda mensal Familiar	Até R$ 1.320	37,91	6,26
Tem empregada?	Não	95,32	3,52
Tem acesso à internet?	Sim	88,89	-1,56
Privada (n = 234.639)			
Faixa Etária	17 anos	50,53	27,52

Sexo	Feminino	54,18	-16,22
Estado Civil	Solteiro(a)	95,62	6,86
Etnia	Branca(o)	68,02	24,62
Tipo da escola	-	-	-
Renda mensal Familiar	De R$ 3.960,01 a R$ 5.280,00	12,44	-19,21
Tem empregada?	Não	71,35	-20,45
Tem acesso à internet?	Sim	99,13	8,68

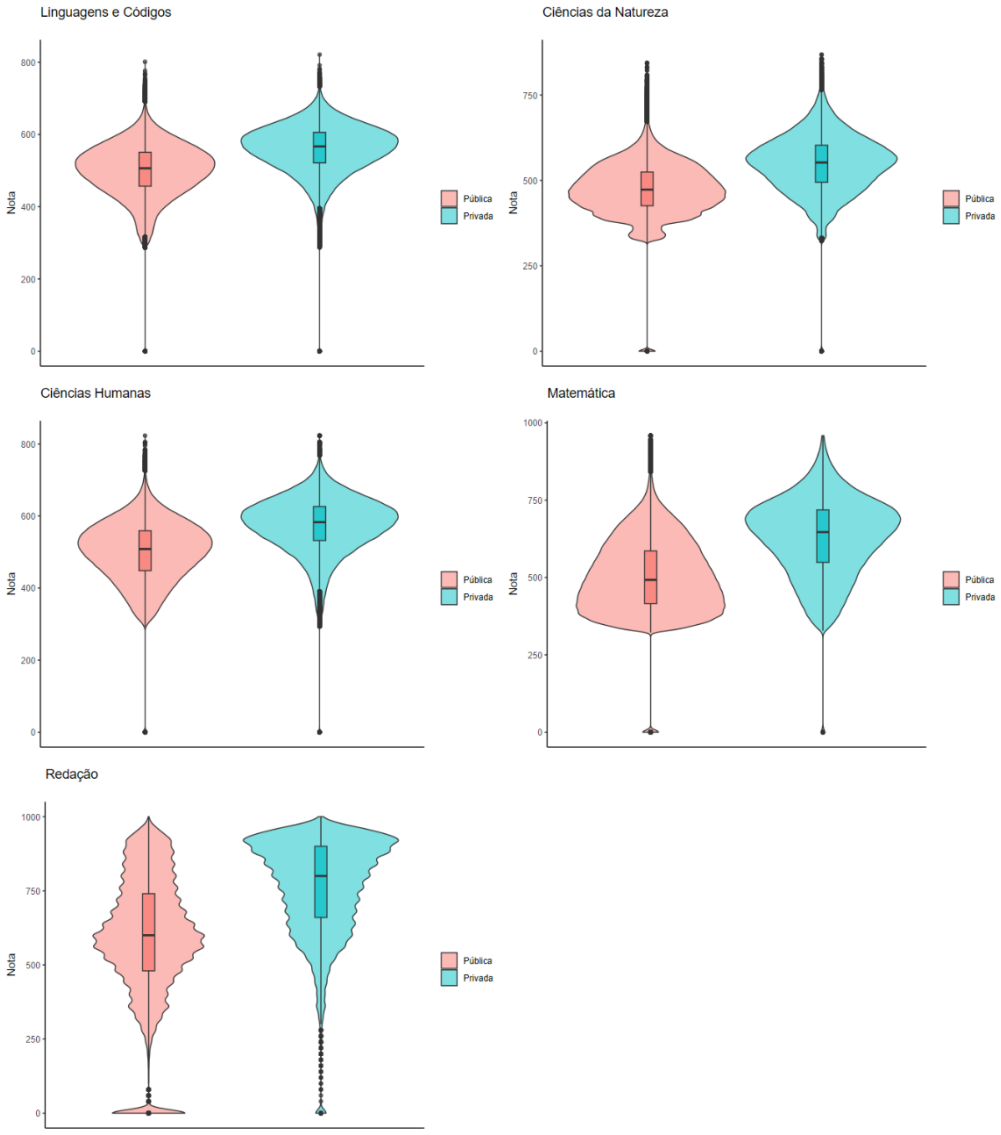
Em comparação a população geral, fica evidente que as notas dos estudantes de escolas públicas tiveram menor média em todas as áreas. Enquanto os estudantes de escolas privadas apresentaram um resultado inverso, com notas superiores a população geral em todas as áreas. Fica evidente então que há uma disparidade muito grande ainda em termos de qualidade de ensino público quando comparado ao ensino privado. Porém, tendo em vista a maior renda da população de escolas privadas, pode ser possível afirmar que os estudantes com maior renda familiar teriam acesso a cursos preparatórios. Ou seja, há a possibilidade de que a nota seja maior por conta da renda.

Notas por tipo de escola

Área	**Média**	**Desv. Pad**	**Diferença**
Pública (n = 1.166.540)			
Ciências da Natureza	473,52	±77,73	-22,23
Linguagens e Códigos	500,55	±71,59	-17,60
Ciências Humanas	501,22	±82,37	-22,13
Matemática	503,98	±116,13	-29,86
Redação	583,75	±219,73	-34,05
Privada (n = 234.619)			
Ciências da Natureza	548,47	±83,99	52,72
Linguagens e Códigos	559,75	±71,59	41,60
Ciências Humanas	573,92	±77,41	50,57
Matemática	630,77	±124,61	96,93
Redação	762,63	±163,99	144,83

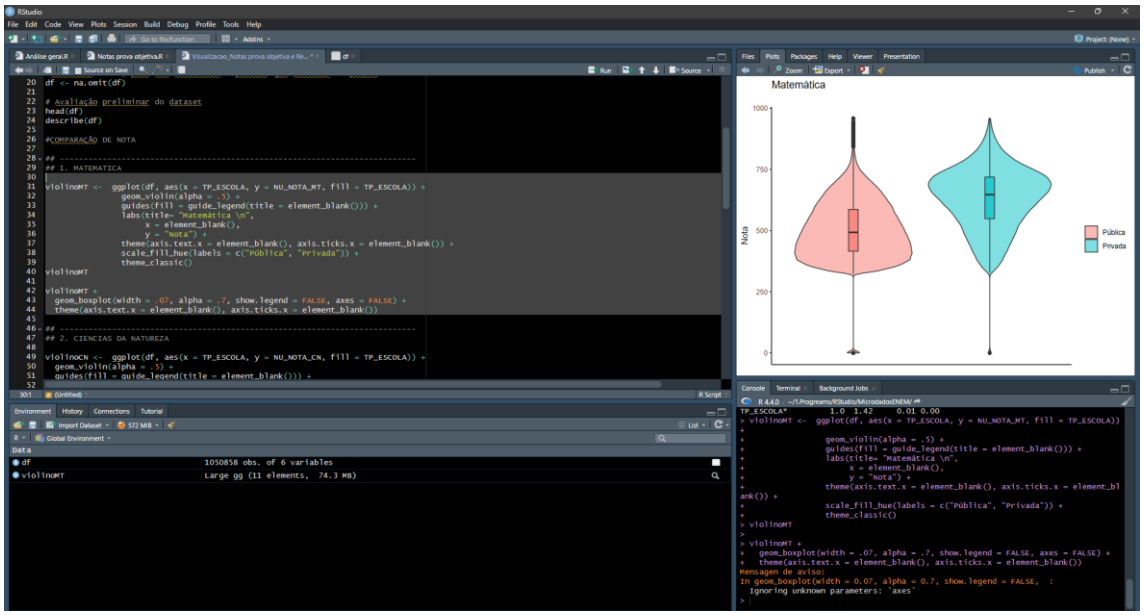
Como houveram diferenças significativas entre o tipo de escola quando comparada a população geral, é interessante avaliar as diferenças ao comparar os dois grupos entre si. Essa diferença foi visualizada por meio de *boxplots* e *violinplots,* elaborados no programa R, usando a IDE RStudio, o script se encontra no GitHub.

Nessas visualizações fica ainda mais notável a disparidade de notas, principalmente em matemática e na quantidade de zeros em redação.

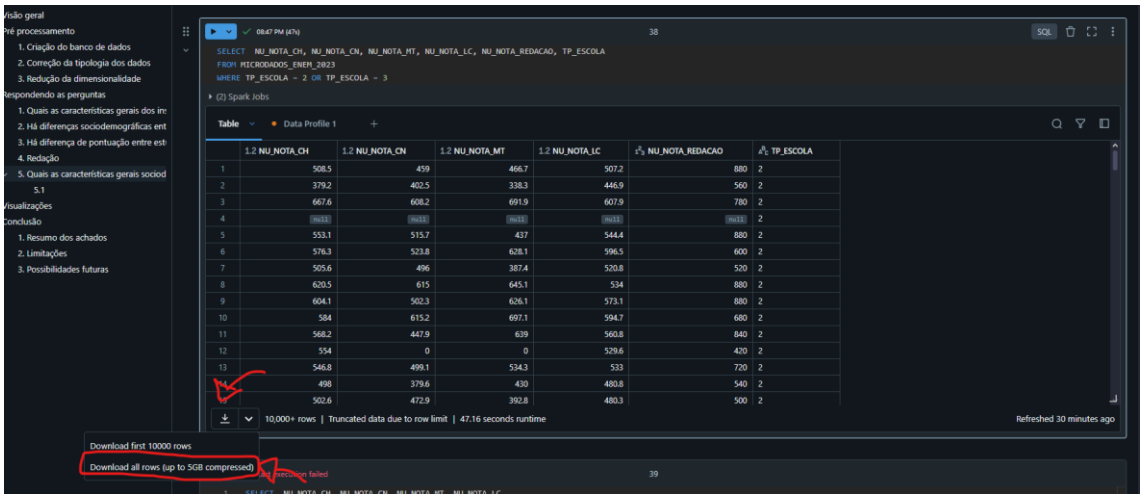


Print referente ao script base para visualização das informações, foi usado o R Localmente por problemas de lentidão para uso do R no databricks, para resolver este problema, a consulta era feita e em sequência era realizado o download da tabela com

apenas as informações necessárias



Como nesse caso:



2.5 Redação

Competência mais fácil: **Competência 2** (Média = 139 ±51,62) possuindo uma prevalência de zeros semelhante as outras competências (4.17%).

A competência 2 envolve "compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa."

Competência mais difícil: **Competência 5** (Média = 108,65 ±61,6) possuindo a maior prevalência de zeros (10.27%) se comparada a outras competências. A competência 5 envolve "elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos."

2.6 Caracterização daqueles com nota inferior e superior à média por área

Tentamos compreender se há maiores diferenças entre o perfil destes com nota a baixo da média e acima, como critério para classificação, foi escolhido que a baixo e acima da média seriam aqueles com um total de pontuação de dois desvios padrão. Esse critério parece relevante já que a distribuição visual das notas tem aparência semelhante a normal.

Área	Nota e DP	+2 DP	- 2 DP
Ciências Humanas	523,35 ±88,57	700,49	346,21
Linguagens e Códigos	518,15 ±75,45	669,05	367,25
Matemática	533,84 ±131,65	797,14	270,54
Ciências da natureza	495,75 ±87,93	671,61	319,89

DP = Desvio Padrão;

O conjunto de tabelas a baixo evidencia em vermelho aquelas características que divergiram do geral em termos de variável mais frequente. Temos a área, a nota de corte e o quantitativo de pessoas que tiveram nota maior ou igual a nota de corte estabelecida A maior disparidade foi em matemática, onde a maioria declarou ter renda mensal familiar acima de R$24.400.

Área	Mais frequente
Ciências Humanas | Nota: 700,49 |N:10.000	
Faixa Etária	18 anos
Sexo	Masculino
Estado Civil	Solteiro(a)
Etnia	Branca
Tipo da escola	Privada
Renda mensal Familiar	R$ 3.960,01 aR$ 5.280,00
Tem empregada?	Não
Tem acesso à internet?	Sim
Ciências da Natureza |Nota: 671,61 |N: 60.165	
Faixa Etária	18 anos
Sexo	Masculino
Estado Civil	Solteiro(a)
Etnia	Branca
Tipo da escola	Privada
Renda mensal Familiar	R$ 3.960,01 aR$ 5.280,00
Tem empregada?	Não
Tem acesso à internet?	Sim

Área	Mais frequente
Matemática |Nota : 797,14 |N: 59.709	
Faixa Etária	18 anos
Sexo	Masculino
Estado Civil	Solteiro(a)
Etnia	Branca
Tipo da escola	Privada
Renda mensal Familiar	Acima de R$ 26.400,00
Tem empregada?	Não
Tem acesso à internet?	Sim
Linguagens e Códigos |Nota: 669,05 |N: 34.698	
Faixa Etária	18 anos
Sexo	Feminino
Estado Civil	Solteiro(a)
Etnia	Branca
Tipo da escola	Privada
Renda mensal Familiar	R$ 3.960,01 aR$ 5.280,00
Tem empregada?	Não
Tem acesso à internet?	Sim

Área	Mais frequente
Redação |Nota : 900 |N: 214.354	
Faixa Etária	18 anos
Sexo	Feminino
Estado Civil	Solteiro(a)
Etnia	Branca
Tipo da escola	Privada
Renda mensal Familiar	Até R$: 1.320
Tem empregada?	Não
Tem acesso à internet?	Sim

Fica evidente mais uma vez que Etnia atodeclarada e tipo da escola são variáveis relevantes para prever maiores notas. Em todas as áreas a escola foi privada.

Conclusão

A priori, vemos que o perfil da estudante inscrita no ENEM é Parda, com renda familiar de ate um salário mínimo, tem 18 anos, é solteira, majoritariamente do sexo feminino, de escola pública (entre os que declararam o tipo de escola) e tem acesso à internet possivelmente condizente com os dados do censo populacional brasileiro. Podemos concluir que não houveram diferenças significativas que nos ajudassem a compreender melhor o que poderia influenciar as chances de perder a prova.

Na hora de tentar compreender o que poderia influenciar nas maiores notas, ficou claro que o tipo da escola e a renda foram variáveis relevantes para garantir melhores desfechos. Tendo como caso mais extremo a renda daqueles que tiveram maiores notas em matemática, com a maioria tendo renda familiar acima de 24 mil. É possível afirmar que com base nos dados do ENEM de 2023, a oportunidade de ingressar nas universidades públicas é barrada pela renda do participante, afinal, sem dinheiro para uma escola privada e cursinhos, resta ingressar em uma universidade privada. O recorte de etnia autodeclarada parece ser um pouco mais complexo, onde vemos que ser branco autodeclarado também está associado a maiores notas, talvez haja alguma associação entre etnia autodeclarada e renda que possa justificar, porem fica fora do escopo deste relatório.

Tratando-se da redação, vimos que a competência mais difícil e que mais garante zeros foi a competência 5, que envolve a proposta de solução e respeito aos direitos humanos. Uma possível hipótese para avaliar se é possível dissociar quem feriu direitos humanos ou só não elaborou uma proposta relevante seria avaliar os zeros. Infringir direitos humanos já garante perder essas competências, então zeros poderiam se associar a infração dos Direitos Humanos, enquanto menores notas (p. ex: <100) poderiam se associar a propostas de intervenção pouco relevantes.

Autoavaliação

Em geral, tempo considerável foi dedicado ao projeto, mas a tentativa de organizar o método para elaborar as respostas válidas para as perguntas tomou mais tempo que o previsto. Por conta disso, percebe-se que talvez mais tópicos aprendidos durante a disciplina poderiam ter sido contemplados no projeto. Apesar disso, a importância da engenharia de dados para a profissão de Cientista de dados foi muito bem percebida pelo autor, em termos de organização e governança de dados para garantir que toda uma equipe poderá tirar proveito das informações, Camadas Bronze, Silver e Gold, a depender do publico alvo dos dados organizados, e afins. Quanto aos objetivos delineados antes do começo do processo, acredita-se que estes foram cumpridos de forma satisfatória no decorrer do relatório

A maior dificuldade acabou envolvendo a compreensão e seleção de atributos para caracterizar a população, tendo em vista a riqueza de informações do conjunto de dados. Em sequência, a dificuldade veio em forma de sumarizar de um jeito não tão cansativo as informações, evitando ao máximo redundâncias ao comunicar os achados mais relevantes

Algumas propostas foram pensadas para melhorar o projeto futuramente:

- Separar as categorias de dados em diferentes esquemas usando um Modelo de Entidades e Relacionamentos. O número de inscrição poderia ser a chave primária/estrangeira. Seria interessante avaliar se há alteração no desempenho das consultas. Caso houvesse, talvez fosse possível adicionar Exames de anos anteriores.
    - Ex[1]: Como tem sido a média e desvio padrão de notas por áreas do conhecimento nos ultimos 10 anos?
    - Ex[2]: O ENEM tem ficado mais ou menos acessivel com o passar do tempo: (avaliar prevalência de tipo de escola do aluno inscrito nos ultimos 10 anos,s e a prevalencia de escolas particulares aumentou, possívelmente tem ficado menos acessível, seria necessário comparar com a prevalencia geral da população por tipo de escola junto a renda familiar)
- Decodificar o dataset e avaliar alteração disso no desempenho da consulta. Seria possível ver se valeria a pena abrir mão de desempenho em prol de celeridade na compreensão do dataset, já que não seria necessário checar constantemente o dicionário.
- Alterar atributo “faixa etária” e transformar em <17, 17, a 22 e >25
- Realizar testes de hipótese estatísticos por meio de analises robustas (G de Hedge para diferença de médias entre dois grupos, ANOVAS para diferenças entre >2 grupos)