DADOS INEP

Análise Exploratória de Dados do ENEM

Vinícius dos Santos Moreira

Disciplina: Fundamentos de Banco de Dados / PPCA / UnB

2024.2

DADOS INEP

Análise Exploratória de Dados do ENEM

Introdução

O Exame Nacional do Ensino Médio (Enem) tem como objetivo principal avaliar, ao final do ensino médio, o domínio dos princípios científicos e tecnológicos para a produção moderna e o conhecimento das formas contemporâneas de linguagem. Os resultados do Enem servem para autoavaliação, melhoria dos currículos do ensino médio, acesso à educação superior, programas de financiamento estudantil, ingresso no mercado de trabalho e desenvolvimento de estudos sobre a educação brasileira.

Implementado em 1998, o Enem passou por uma reformulação metodológica em 2009, que estruturou as Matrizes de Referência por competências em quatro áreas do conhecimento: Linguagens, Códigos e suas Tecnologias; Matemática e suas Tecnologias; Ciências Humanas e suas Tecnologias; e Ciências da Natureza e suas Tecnologias. O exame, composto por 180 questões objetivas e uma redação, é aplicado em dois dias.

Os microdados do Enem 2023, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), incluem informações gerais sobre a realização das provas, caracterização dos participantes e escolas, notas das provas objetivas e da redação. Esses dados são disponibilizados em formato ".csv" e acompanhados de dicionários de variáveis e inputs para softwares estatísticos.

Este estudo tem como objetivo realizar uma análise exploratória dos microdados do Enem 2023, buscando identificar padrões e tendências no desempenho dos participantes. A análise considerará as quatro áreas de conhecimento avaliadas no exame, bem como o perfil socioeconômico dos participantes, utilizando os dados do questionário respondido pelos inscritos. A Teoria de Resposta ao Item (TRI), utilizada no cálculo das notas do Enem, será

abordada para entender como os parâmetros dos itens e o padrão de respostas influenciam a proficiência dos participantes.

Modelo de Dados Relacional

Um modelo de dados relacional para representar os microdados do ENEM 2023 pode ser estruturado em torno de cinco tabelas principais, complementadas por tabelas de apoio para detalhar aspectos específicos dos dados.

- **Tabela de Participantes:** Esta tabela é o núcleo do modelo, contendo informações sobre cada participante do ENEM 2023. Cada linha representa um participante único, e as colunas incluem dados de caracterização do participante, do tipo de escola (pública ou privada) e de ensino (regular ou especial) que ele declarou ter frequentado.
- **Tabela de Notas das Provas Objetivas**
- **Tabela de Notas da Prova de Redação**.
- **Tabela de Escolas**
- **Tabela de Locais de Provas**

Para complementar estas tabelas principais, podem ser criadas tabelas de apoio ou de domínio, como:

- **Tabela de Gabaritos:** Contém os gabaritos oficiais de cada caderno de prova, permitindo a comparação com as respostas dos participantes.
- **Tabela de Faixa Etária:**
- **Tabela de Tipo de Dependência Adminsitrativa:**

Este modelo relacional permite a análise exploratória dos dados do ENEM 2023, possibilitando a identificação de padrões e tendências no desempenho dos participantes, a avaliação da

qualidade dos itens da prova e a análise do impacto de fatores socioeconômicos no desempenho dos estudantes.

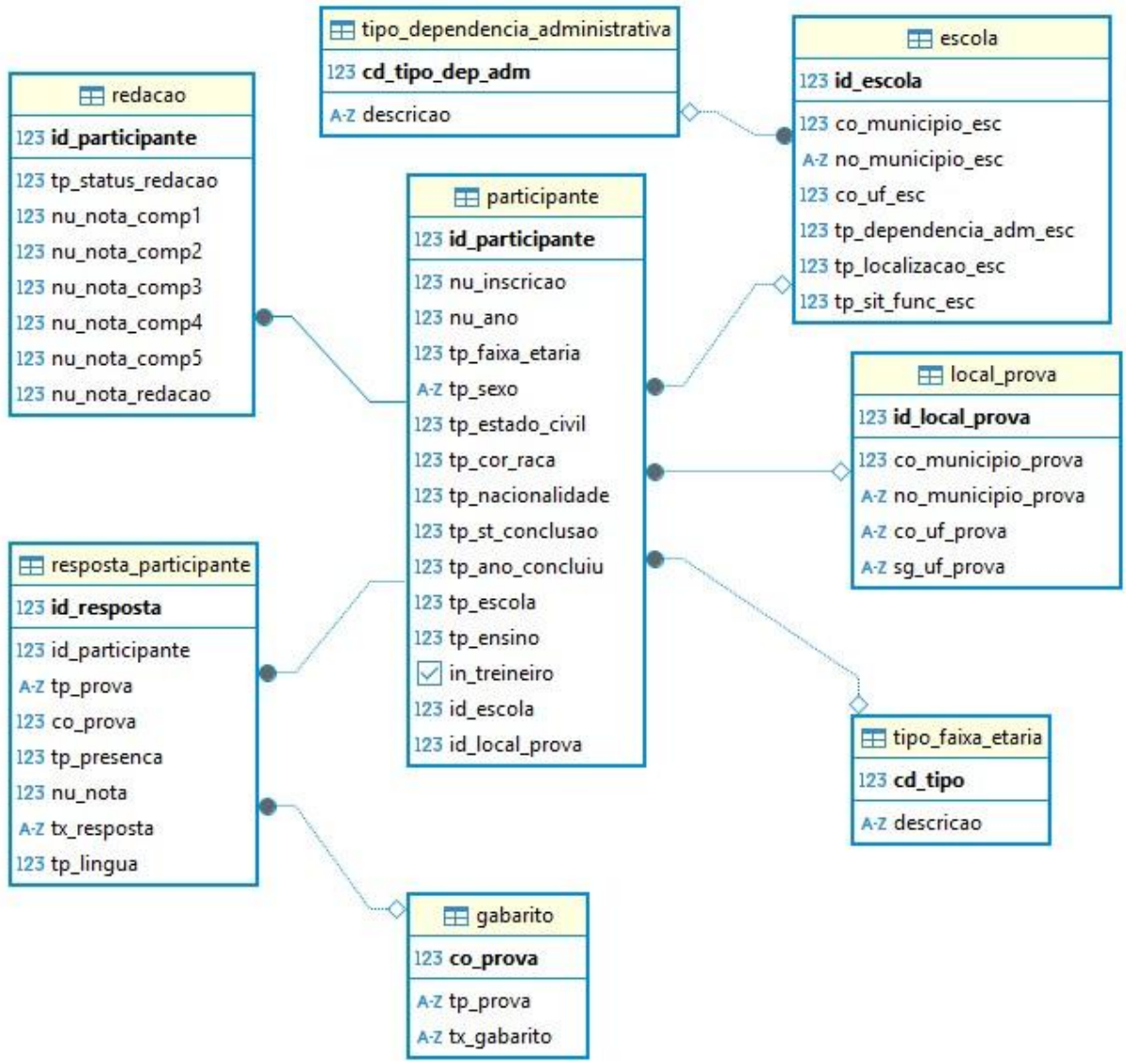

Figura 1.Diagrama do Medelo Entidade Relacionamento

Processo ETL

Com base nos microdados do ENEM 2023, um processo de ETL (Extração, Transformação e Carga) para carregar os dados do arquivo microdados_enem_2023.csv para um banco de dados PostgreSQL versão 17.3 pôde ser estruturado da seguinte forma:

- **Extração**:

- Os microdados do ENEM 2023 estão disponíveis em formato **.csv** (valores separados por ponto e vírgula).
- O arquivo principal, MICRODADOS_ENEM_2023.csv, contém os questionários respondidos pelos participantes, reunindo informações sobre a realização das provas, características dos participantes e das escolas, e as notas das provas objetivas e da redação.

- **Transformação**
  - **Anonimização:** Os microdados já estão anonimizados, sem informações que permitam a identificação direta dos participantes. O número de inscrição é substituído por uma máscara sequencial.
  - **Filtragem e seleção de dados:** O Inep excluiu da base de dados alguns registros de participantes que realizaram provas específicas (códigos 1230, 1305, 1310, 1317 e 1318 para Ciências da Natureza; 1200, 1275, 1280, 1311 e 1312 para Ciências Humanas; 1210, 1285, 1290, 1313 e 1314 para Linguagens e Códigos; e 1220, 1295, 1300, 1315 e 1316 para Matemática) devido ao pequeno número de participantes, o que poderia permitir a identificação indevida. Um participante com cálculo de resultado alterado por decisão judicial também foi excluído.
  - **Simplificação:** O Inep adotou um modelo simplificado de microdados a partir de 2020, replicado nas edições posteriores, para eliminar variáveis que facilitem a identificação indevida. As seguintes alterações foram feitas:
    - Exclusão da variável CO_ESCOLA.
    - Exclusão de informações referentes aos pedidos de atendimento especializado e específico, recursos de atendimento especializado e específico para a realização da prova.

        - Substituição da variável NU_IDADE por TP_FAIXA_ETARIA.
        - Exclusão de informações referentes aos municípios de nascimento e residência do participante.
    - **Estruturação dos dados:** Os dados precisam ser estruturados para corresponder ao esquema do banco de dados PostgreSQL. Isso envolveu a conversão do tipo de dado da coluna NU_INSCRICAO originária, criação de novas colunas para chaves primárias e estrangeiras nas tabelas PARTICIPANTE, ESCOLA, REDACAO, LOCAL_PROVA e segregação de colunas de notas existentes na tabela matriz para unificação numa única tabela RESPOSTA_PARTICIPANTE.
- **Carga**
    - **Criação do banco de dados e tabelas:** No PostgreSQL, foi criado um banco de dados e uma tabela correspondente para armazenar os microdados do ENEM. Foram definidos os tipos de dados das colunas de acordo com o dicionário de dados fornecido pelo INEP.
    - **Importação dos dados:** Foi utilizada a rotina de importação de dados do gerenciador universal de banco de dados DBeaver, versão 24.3.5. Após a importação dos dados para a tabela MICRODADOS_ENEM, foram construídas rotinas para hidratação das tabelas do modelo relacional do projeto, conforme disponível no ANEXO I.

Exploração dos Dados

Participantes por Tipo de Escola

"*1. Quais os totais de participantes por tipo de escola, excluindo-se os treineiros?*"

```sql
-- 1. projetar total de participantes por tipo de escola excluindo-se os treineiros
select
    case tp_escola
        when 1 then 'Não respondeu'
        when 2 then 'Pública'
        when 3 then 'Privada'
        else 'Não informado'
    end as tipo_escola,
    COUNT(*) as quantidade
from participante
where
    not in_treineiro
group by
    tp_escola;
```

Figura 2. Consulta estruturada (SQL) da questão 1

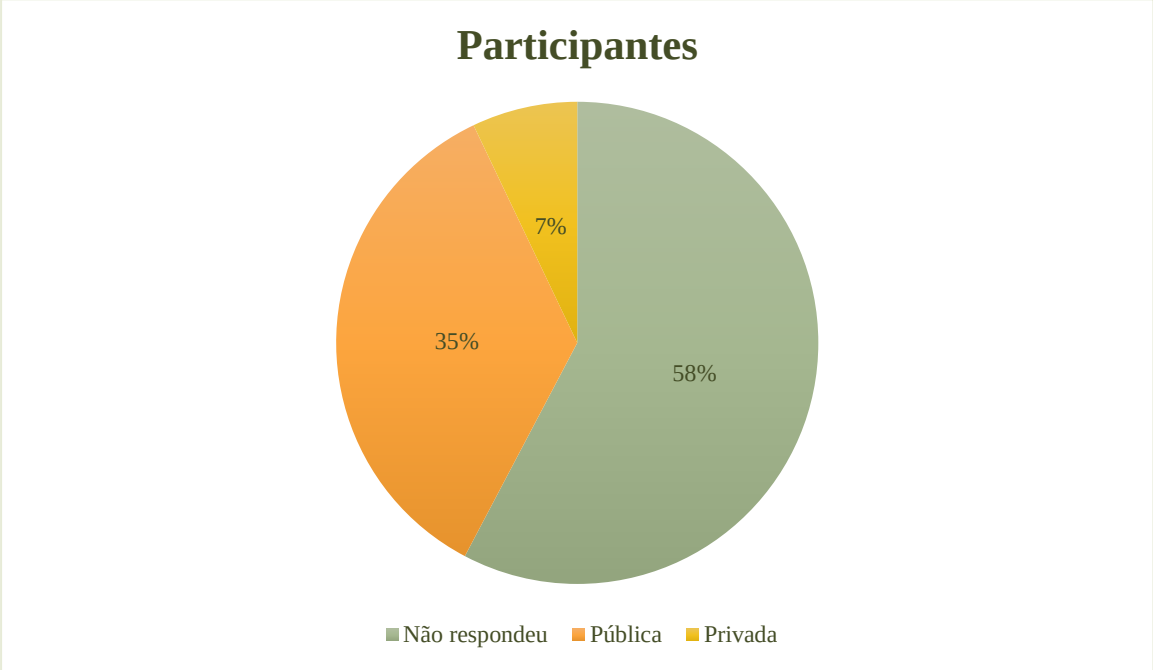

Figura 3. Gráfico de participantes por tipo de escola.

Municípios com maiores médias

"*2. Quais os 10 municípios com maior média de notas?*"

```sql
-- 2. Selecionar os 10 municípios com maior média de notas
with
    pontuacao_participante as (
        select p.id_participante, sum(coalesce(nu_nota, 0)) as nota_final
        from
            participante p
            inner join resposta_participante rp on (
                rp.id_participante = p.id_participante
            )
        where
            rp.tp_presenca = 1
            and not p.in_treineiro
        group by
            p.id_participante
    )
select e.no_municipio_esc, avg(pp.nota_final) as media_nota
from
    participante p
    inner join pontuacao_participante pp on (
        p.id_participante = pp.id_participante
    )
    inner join escola e on (p.id_escola = e.id_escola)
group by
    e.no_municipio_esc
order by avg(pp.nota_final) desc
limit 10;
```

Figura 4. Consulta Estruturada (SQL) da segunda questão.

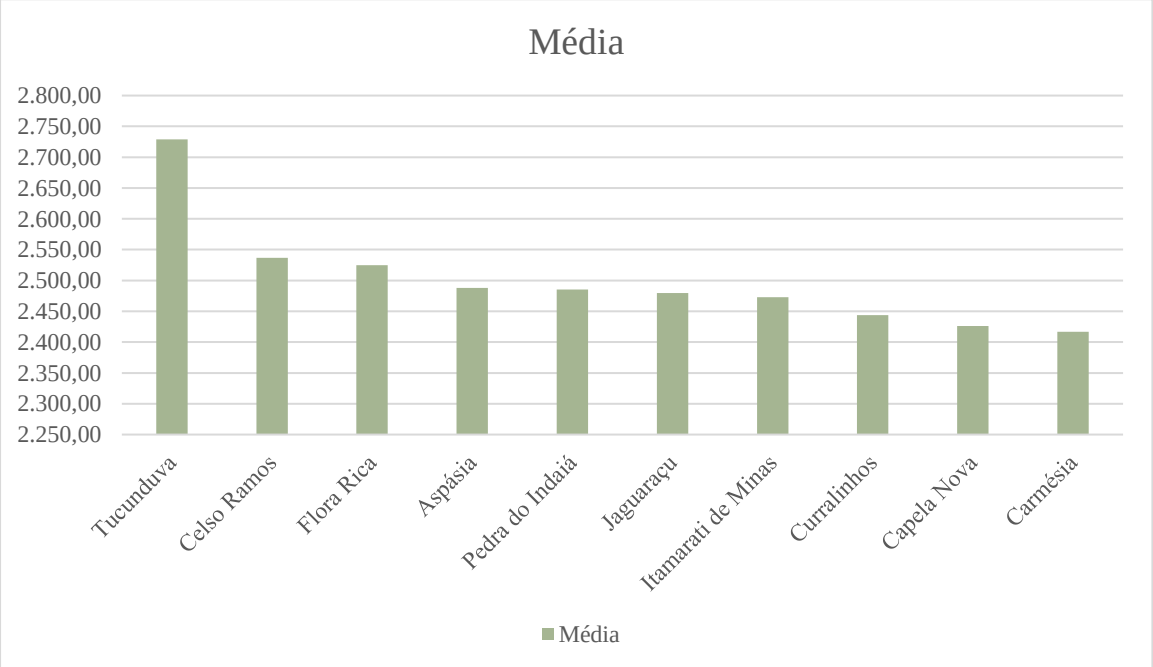

Figura 5. Gráfico dos 10 municípios com as maiores médias nas provas objetivas.

Tabela 1. Os 10 municípios com maiores médias nas provas objetivas

Município	Média
Tucunduva	2.728,85
Celso Ramos	2.536,80
Flora Rica	2.524,97
Aspásia	2.487,90
Pedra do Indaiá	2.485,30
Jaguaraçu	2.479,77
Itamarati de Minas	2.472,80
Curralinhos	2.443,55
Capela Nova	2.426,28
Carmésia	2.416,90

Locais de Prova com Maior Número de Abstenções

"3. *Quais são os locais de prova com maior número de abstenções?*"

```sql
-- 3. projetar os locais de prova com maior índice de abstenção
select lp.no_municipio_prova, lp.sg_uf_prova, count(*) as quantidade
from
    participante p
    inner join local_prova lp on (
        p.id_local_prova = lp.id_local_prova
    )
    inner join resposta_participante rp on (
        rp.id_participante = p.id_participante
        and rp.tp_prova = 'CN'
        and rp.tp_presenca = 0
    )
group by
    lp.no_municipio_prova,
    lp.sg_uf_prova
order by count(*) desc
limit 10;
```

Figura 6. Consulta estruturada (SQL) da terceira questão.

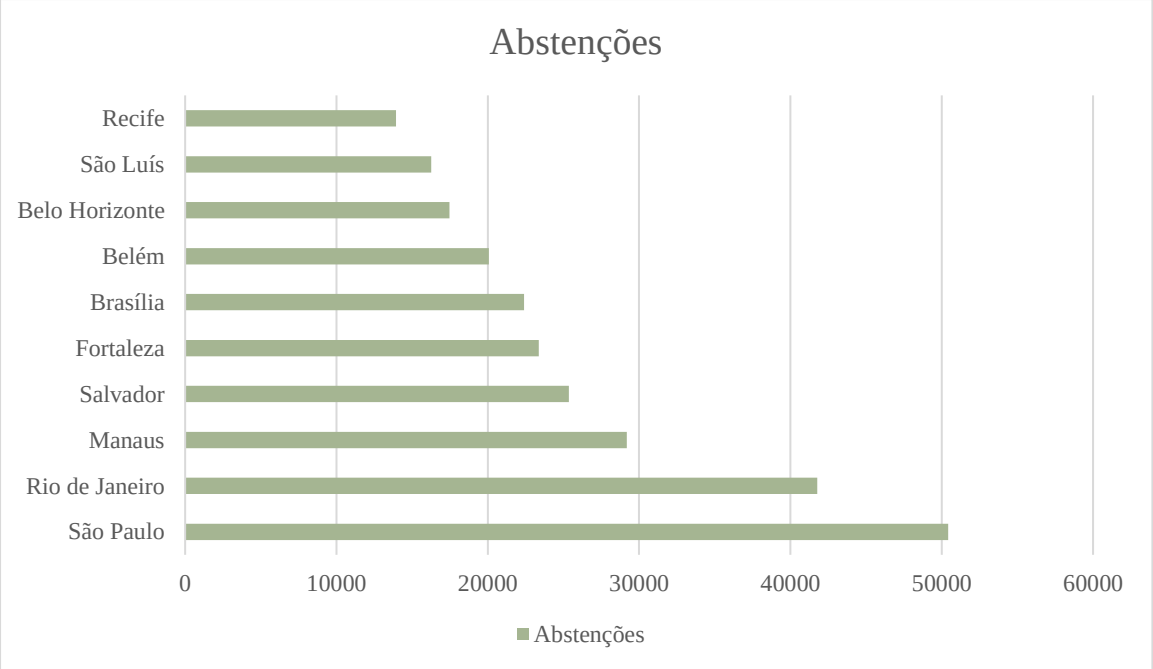

Figura 7. Gráfico de abstenções por mumicípio

Tabela 2. Quantidades de absentenções por município.

Município	Abstenções
São Paulo	50424
Rio de Janeiro	41773
Manaus	29189
Salvador	25360
Fortaleza	23378
Brasília	22403
Belém	20065
Belo Horizonte	17459
São Luís	16263
Recife	13933

Municípios com Melhores Desempenhos em Redação

"4. Quais são os municípios com maiores médias de redação por esfera de escola?"

```sql
-- 4. projetar as maiores médias de redação por município e tipo de escola
select
    e.no_municipio_esc as municipio,
    e.co_uf_esc as uf,
    case e.tp_dependencia_adm_esc
        when 1 then 'Federal'
        when 2 then 'Estadual'
        when 3 then 'Municipal'
        when 4 then 'Privada'
        else 'Não informado'
    end as esfera,
    round(avg(rp.nu_nota_redacao), 2) as media_redacao
from
    participante p
    inner join redacao rp on (
        rp.id_participante = p.id_participante
    )
    inner join escola e on (p.id_escola = e.id_escola)
where
    not p.in_treineiro
    and rp.nu_nota_redacao is not null
group by
    e.co_uf_esc,
    e.no_municipio_esc,
    e.tp_dependencia_adm_esc
order by avg(rp.nu_nota_redacao) desc
limit 10;
```

Figura 8. Consulta estruturada (SQL) dos municípios escolares com maiores médas de redação por esfera.

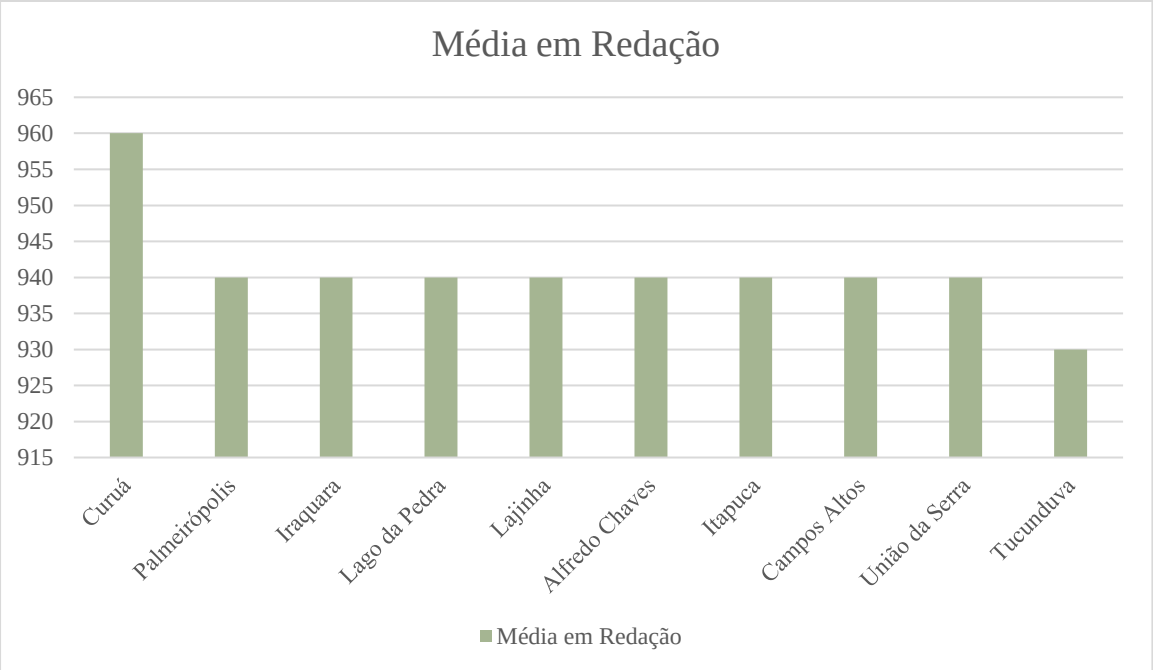

Figura 9. Gráfico de município das erscolas com maiores médias em redação.

Tabela 3. Municípios das escolas com maiores médias em redação

Município	UF	Média em Redação
Curuá	PA	960
Palmeirópolis	TO	940
Iraquara	BA	940
Lago da Pedra	MA	940
Lajinha	MG	940
Alfredo Chaves	ES	940
Itapuca	RS	940
Campos Altos	MG	940
União da Serra	RS	940
Tucunduva	RS	930

Faixas Etárias com Maiores Médias

"5. Quais são as faixas etárias com maiores médias?"

```sql
-- 5. projetar as maiores médias por faixa etária
with
    nota_total_participante as (
        select p.id_participante, sum(
                coalesce(rp.nu_nota, 0) + coalesce(r.nu_nota_redacao, 0)
            ) as nota_final
        from
            participante p
            left join resposta_participante rp on (
                rp.id_participante = p.id_participante
                and rp.tp_presenca = 1
            )
            left join redacao r on (
                r.id_participante = p.id_participante
            )
        where
            not p.in_treineiro
        group by
            p.id_participante
    )
select tfe.descricao, round(avg(ntp.nota_final)) as media_nota
from
    participante p
    inner join tipo_faixa_etaria tfe on (
        p.tp_faixa_etaria = tfe.cd_tipo
    )
    inner join nota_total_participante ntp on (
        p.id_participante = ntp.id_participante
    )
where
    not p.in_treineiro
group by
    tfe.descricao
order by avg(ntp.nota_final) desc
limit 10;
```

Figura 10. Consulta estruturada (SQL) da quinta questão.

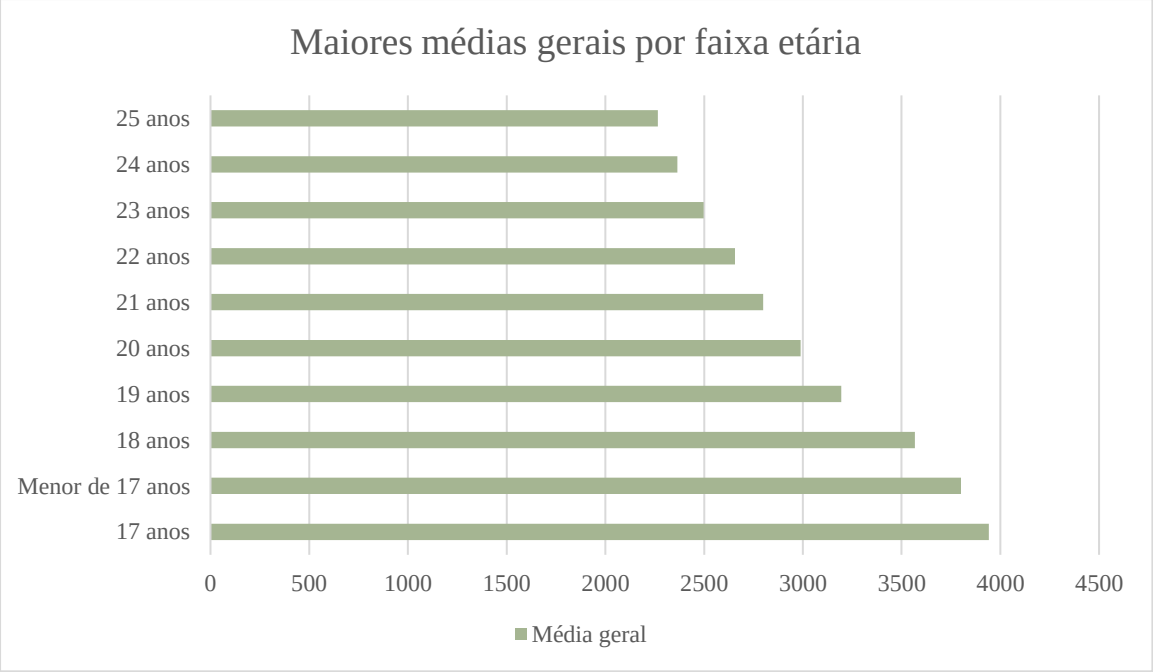

Figura 11. Gráfico das médias gerais por faixa etária

Tabela 4. Médias gerais por faixa etária

Faixa etária	Média geral
17 anos	3942
Menor de 17 anos	3800
18 anos	3567
19 anos	3195
20 anos	2989
21 anos	2799
22 anos	2656
23 anos	2498
24 anos	2364
25 anos	2266

Visão

Foi construída uma visão materializada com os dados estatísticos das médias das notas gerais por faixa etária para facilitar as avaliações de dados agregados.

```sql
create MATERIALIZED view if not exists inep.estatisticas_participante as
with
    nota_total_participante as (
        select p.id_participante, sum(
                coalesce(rp.nu_nota, 0) + coalesce(r.nu_nota_redacao, 0)
            ) as nota_final
        from
            participante p
            left join resposta_participante rp on (
                rp.id_participante = p.id_participante
                and rp.tp_presenca = 1
            )
            left join redacao r on (
                r.id_participante = p.id_participante
            )
        where
            not p.in_treineiro
        group by
            p.id_participante
    )
select
    p.tp_faixa_etaria,
    count(*) as quantidade,
    round(avg(ntp.nota_final)) as media_nota,
    stddev(ntp.nota_final) as desvio_padrao,
    min(ntp.nota_final) as menor_nota,
    max(ntp.nota_final) as maior_nota,
    percentile_cont(0.25) within group (order by ntp.nota_final) as primeiro_quartil,
    percentile_cont(0.5) within group (order by ntp.nota_final) as mediana,
    percentile_cont(0.75) within group (order by ntp.nota_final) as terceiro_quartil
from
    participante p
    inner join nota_total_participante ntp on (
        p.id_participante = ntp.id_participante
    )
group by
    p.tp_faixa_etaria;
```

Figura 12. Código estruturado (DDL) de criação da visão materializada.

Tabela 5. Dados estatísticos por faixa etária.

descricao	quantidade	media	Desvio padrão	mínima	máxima	1ºquartil	mediana	3ºquartil
Menor de 17 anos	19961	38000	20370	0	69310	25600	44250	53650
17 anos	497159	39420	19740	0	71760	33690	44960	53700
18 anos	878894	35670	21200	0	72530	19390	42510	51590
19 anos	426390	31950	22860	0	71800	0	39830	50270
20 anos	266041	29890	23410	0	71510	0	37990	48970
21 anos	182869	27990	23720	0	71030	0	36030	47840
22 anos	137534	26560	23720	0	71180	0	34070	46830
23 anos	111569	24980	23530	0	70810	0	31080	45700
24 anos	91174	23640	23230	0	71240	0	25330	44660
25 anos	72968	22660	22960	0	71410	0	19850	43930

descricao	quantidade	media	Desvio padrão	mínima	máxima	1ºquartil	mediana	3ºquartil
Entre 26 e 30 anos	245671	20850	22350	0	71020	0	0	42480
Entre 31 e 35 anos	133003	19660	21640	0	70970	0	0	41260
Entre 36 e 40 anos	96663	19450	20950	0	70670	0	0	40310
Entre 41 e 45 anos	66905	19090	20310	0	68690	0	9230	39330
Entre 46 e 50 anos	40652	19490	19950	0	68810	0	16400	39010
Entre 51 e 55 anos	24532	19550	19730	0	68330	0	16950	38870
Entre 56 e 60 anos	13403	20020	19640	0	69210	0	18400	38940
Entre 61 e 65 anos	5484	20530	19070	0	65500	0	21750	38320
Entre 66 e 70 anos	2150	19820	18490	0	64370	0	19190	37020
Maior de 70 anos	866	18770	17820	0	58160	0	17670	35640

Funções

Foram criadas 2 funções para aprofundamento da análise exploratória: uma para cálculo da nota total do participante e outra para a comparação de notas considerando-se as faixas etárias dos participantes.

Função de cálculo da nota

A função de cálculo da nota recebe o valor do código de inscrição como parâmetro e retorna o somatório das notas das provas objetivas e da redação para o cálculo da nota geral do participante.

```sql
-- -- função para cálculo do total da nota
create or replace function calcular_nota_total(p_nu_inscricao bigint)
returns numeric as $$
declare
    l_nota_total numeric;
begin
    select sum(coalesce(rp.nu_nota, 0) + coalesce(r.nu_nota_redacao, 0))
    into l_nota_total
    from
        participante p
        left join resposta_participante rp on (
            rp.id_participante = p.id_participante
            and rp.tp_presenca = 1
        )
        left join redacao r on (
            r.id_participante = p.id_participante
        )
    where
        p.nu_inscricao = p_nu_inscricao;

    return l_nota_total;
end;
$$ language plpgsql;
```

Figura 13. Código plpgsql da função calcular_nota_total

A chamada à função *calcular_nota_total(210060925335)* retorna o valor do somatório das notas do participante cujo código para o número de inscrição seja igual a 210060925335. Para este caso, o valor retornado será igual a 5.902,90.

Função comparativa de notas

A função comparativa de notas realiza a comparação relativa entre duas notas ajustando-as através do cálculo do Z-Score:

$$Z = \frac{x - \mu}{\sigma}$$

Onde, Z é a pontuação a ser calculada, x é o valor da variável analisada, μ é a média e σ o desvio padrão.

```sql
-- função para cálculo do z-score
create or replace function comparar_participantes(
    p_nu_inscrição_a bigint,
    p_nu_inscrição_b bigint
) returns numeric as $$
declare
    l_participante_a participante%rowtype;
    l_participante_b participante%rowtype;
    l_nota_total_a numeric;
    l_nota_total_b numeric;
    l_stats_a estatisticas_participante%rowtype;
    l_stats_b estatisticas_participante%rowtype;
    l_media_b numeric;
    l_z_a numeric;
    l_z_b numeric;
begin
    select * into l_participante_a
    from participante
    where nu_inscricao = p_nu_inscrição_a;

    select * into l_participante_b
    from participante
    where nu_inscricao = p_nu_inscrição_b;

    if l_participante_a.tp_faixa_etaria = l_participante_b.tp_faixa_etaria then
        return calcular_nota_total(l_participante_a.nu_inscricao) - calcular_nota_total(l_participante_b.nu_inscricao);
    else
        select * into l_stats_a from estatisticas_participante where tp_faixa_etaria = l_participante_a.tp_faixa_etaria;
        select * into l_stats_b from estatisticas_participante where tp_faixa_etaria = l_participante_b.tp_faixa_etaria;

        l_z_a := (calcular_nota_total(l_participante_a.nu_inscricao) - l_stats_a.media_nota) / l_stats_a.desvio_padrao;
        l_z_b := (calcular_nota_total(l_participante_b.nu_inscricao) - l_stats_b.media_nota) / l_stats_b.desvio_padrao;
        return l_z_a - l_z_b;
    end if;
end;
$$ language plpgsql;
```

Figura 14. Código plpgsql da função comparativa de notas.

Como resultado, espera-se um retorno da função igual a zero quando as notas dos participantes A e B forem consideradas iguais, maior que zero quando a nota de A for maior que B e menor que zero quando a nota de B for maior que A.

Para exemplificar o uso dessa função, quando se comparam as notas dos participantes A e B, inscritos sob os números 210060925335 e 210060978194, cujas faixas etárias são respectivamente “17 anos” e “entre 56 e 60 anos”, obtemos que B teve um desempenho relativamente melhor que A, ou seja, a função retorna um valor negativo. Ainda que a nota de A seja igual a 5.902,90 pontos e de B seja igual a 5.070,20 pontos.

Trigger na Tabela de Notas

O gatilho (*trigger*) abaixo foi implementado para evitar a edição ou exclusão de dados da tabela de notas do participante

```
-- Criar a função
create or replace function evitar_update_delete_resposta_participante()
returns trigger as $$
begin
    raise exception 'Operação não permitida: Não é possível atualizar ou excluir registros na tabela resposta_participante';
    return null;
end;
$$ language plpgsql;

-- Criar a trigger
create trigger trg_evitar_update_delete_resposta_participante
before update or delete on resposta_participante
for each row
execute function evitar_update_delete_resposta_participante();
```

Figura 15. Código da trigger de negação de alterações e exclusões na tabela de notas.

Caso o usuário de banco, mesmo com permissão de update e delete, tente alterar um registro da tabela de notas indevidamente, a trigger lançará uma exceção impedindo a execução do comando, conforme mensagem de erro abaixo:

> Erro SQL [P0001]: ERROR: Operação não permitida: Não é possível atualizar ou excluir registros na tabela resposta_participante

Consultas em Álgebra Relacional

Seleção com múltiplas condições

A expressão em álgebra relacional abaixo seleciona os participantes cuja faixa etária é igual a 17 anos, não é treineiro e oriundo de escola pública.

- σ (tp_faixa_etaria = 2 ∧ in_treineiro = false ∧ tp_escola = 2) (participante)

Seleção e Projeção

A expressão em álgebra relacional abaixo projeta a soma das notas (nu_nota) dos registros na tabela resposta_participante onde tp_prova seja igual a matemática ('MT') e tp_presenca seja igual a “Presente na prova” (1).

- SUM(π (nu_nota)(σ (tp_prova = 'MT' ∧ tp_presenca = 1)(resposta_participante)))

Seleção, Projeção e Junção

A expressão em álgebra relacional abaixo projeta a descrição da faixa etária e o número de inscrição da junção entre participantes e tipos de faixas etárias com seleção dos participantes cujos números de inscrição estejam no conjunto {210060925335, 210060978194}.

- π(tfe.descricao, p.nu_inscricao) (σ(p.nu_inscricao ∈ {210060925335, 210060978194}) (participante p ⋈ (tfe.cd_tipo = p.tp_faixa_etaria) tipo_faixa_etaria tfe))

Referências

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2024). *Microdados do Enem 2023*. Recuperado em 22 de fevereiro de 2025, de https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.