

Análise Exploratória de Dados

Base de dados: Students Performance in Exams

Trabalho Substitutivo — Análise Exploratória de Dados (2026.1)

Entrega: 14/07/2026

1. Introdução

Este relatório apresenta uma análise exploratória de dados (AED) realizada sobre a base *Students Performance in Exams*, que reúne informações sobre o desempenho de 1.000 estudantes em três provas padronizadas (matemática, leitura e escrita), além de características demográficas e socioeconômicas de cada aluno. O objetivo do trabalho é descrever a estrutura da base, tratar eventuais problemas de qualidade dos dados, explorar o comportamento individual das variáveis (análise univariada) e investigar relações entre elas (análise bivariada), de modo a identificar padrões relevantes que possam apoiar conclusões sobre os fatores associados ao desempenho acadêmico dos estudantes.

2. Descrição da Base de Dados

A base de dados **Students Performance in Exams** é amplamente utilizada em exercícios de ciência de dados e reúne notas de estudantes do ensino médio em três disciplinas, associadas a variáveis de contexto socioeconômico e educacional. Cada linha representa um estudante e cada coluna uma característica observada.

2.1 Número de observações e variáveis

A base contém **1.000 observações** (estudantes) e **8 variáveis**.

2.2 Classificação das variáveis

Variável	Descrição	Tipo
gender	Gênero do estudante	Qualitativa nominal
race/ethnicity	Grupo étnico (A a E)	Qualitativa nominal
parental level of education	Escolaridade dos pais	Qualitativa ordinal
lunch	Tipo de refeição escolar	Qualitativa nominal
test preparation course	Curso preparatório concluído	Qualitativa nominal
math score	Nota de matemática (0–100)	Quantitativa discreta
reading score	Nota de leitura (0–100)	Quantitativa discreta
writing score	Nota de escrita (0–100)	Quantitativa discreta

As três notas são tecnicamente *quantitativas discretas*, pois assumem apenas valores inteiros entre 0 e 100 (resultado de uma contagem de acertos convertida em escala percentual), embora na prática sejam frequentemente tratadas como contínuas devido à grande amplitude de valores possíveis. A variável *parental level of education* é ordinal porque seus níveis seguem uma progressão natural (ensino médio incompleto → ensino médio → cursando faculdade → tecnólogo → bacharelado → mestrado).

3. Tratamento Inicial dos Dados

3.1 Valores ausentes

Foi verificada a existência de valores ausentes (*missing values*) em todas as colunas por meio da contagem de células nulas. O resultado indicou **zero valores ausentes** em qualquer uma das 8 variáveis, ou seja, a base está completa. Também não foram encontradas linhas duplicadas.

3.2 Inconsistências

Foi verificado o domínio de valores de cada variável categórica (categorias únicas) e o intervalo das variáveis numéricas. Não foram identificadas categorias inconsistentes (grafias duplicadas, espaços extras, etc.) nem valores fora do intervalo esperado de 0 a 100 nas notas — o valor mínimo observado foi 0 (em matemática) e o máximo 100, ambos fisicamente plausíveis para uma prova.

3.3 Valores extremos (outliers)

A identificação de outliers foi feita pelo critério do intervalo interquartil (IQR), em que são considerados extremos os valores abaixo de $Q1 - 1,5 \times IQR$ ou acima de $Q3 + 1,5 \times IQR$. Foram encontrados outliers apenas no lado inferior das três distribuições de notas, correspondendo a estudantes com desempenho muito abaixo da média:

Variável	Limite inferior (IQR)	Nº de outliers	Valores
math score	27,0	8	0, 8, 18, 19, 22, 23, 24, 26
reading score	29,0	6	17, 23, 24, 24, 26, 28
writing score	25,9	5	10, 15, 19, 22, 23

Destaca-se o caso de uma aluna que obteve simultaneamente as três notas mais baixas da base (0 em matemática, 17 em leitura e 10 em escrita), o que sugere uma situação atípica (possível ausência na prova, dificuldade pontual ou erro de registro). Como esses valores são plausíveis dentro da escala de 0 a 100 e não há evidência de erro de digitação, optou-se por **não remover** os outliers da base, apenas sinalizá-los, já que representam casos reais de baixo desempenho e sua retirada poderia mascarar informações relevantes para a análise (por exemplo, a relação entre baixo desempenho e curso preparatório não realizado).

4. Análise Univariada

Foram selecionadas duas variáveis quantitativas (*math score* e *writing score*) e duas variáveis qualitativas (*gender* e *test preparation course*) para a análise univariada.

4.1 Variáveis quantitativas

Variável	Média	Desvio padrão	Mediana	Assimetria (skew)
math score	66,09	15,16	66,0	-0,28
writing score	68,05	15,20	69,0	-0,29

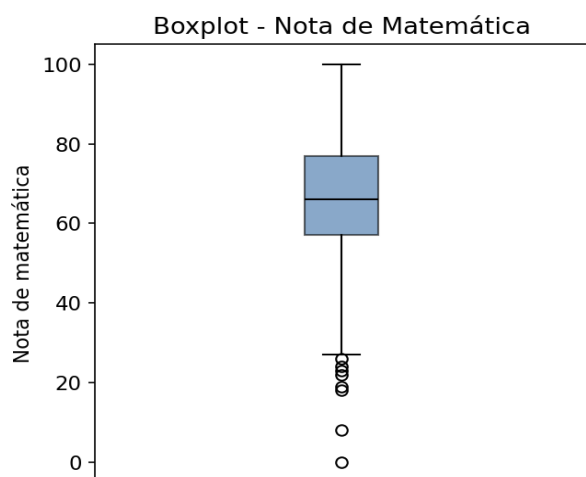


Figura 1 — Boxplot da nota de matemática.

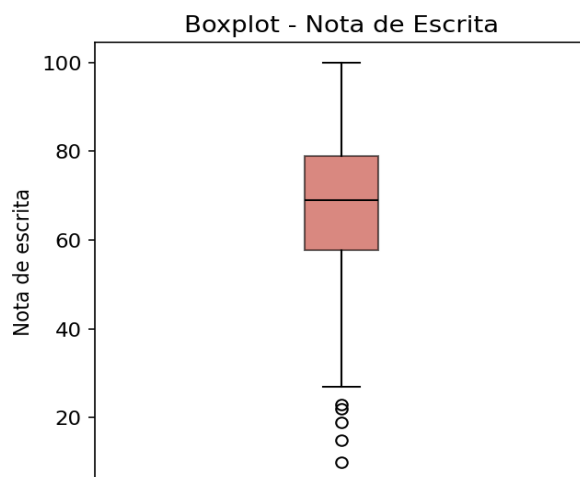


Figura 2 — Boxplot da nota de escrita.

A nota média de matemática (66,1) é ligeiramente inferior à de escrita (68,1), com desvios padrão semelhantes (em torno de 15 pontos), indicando dispersão moderada em ambas. Os coeficientes de assimetria próximos de -0,3 mostram uma leve assimetria à esquerda (cauda mais longa para notas baixas), o que é coerente com os boxplots: a mediana está próxima do centro da caixa, mas há mais outliers no extremo inferior do que no superior — refletindo os poucos alunos com desempenho muito abaixo da média identificados na seção anterior.

4.2 Variáveis qualitativas

gender	Frequência absoluta	Frequência relativa
female	518	51,8%
male	482	48,2%

test preparation course	Frequência absoluta	Frequência relativa
none	642	64,2%
completed	358	35,8%

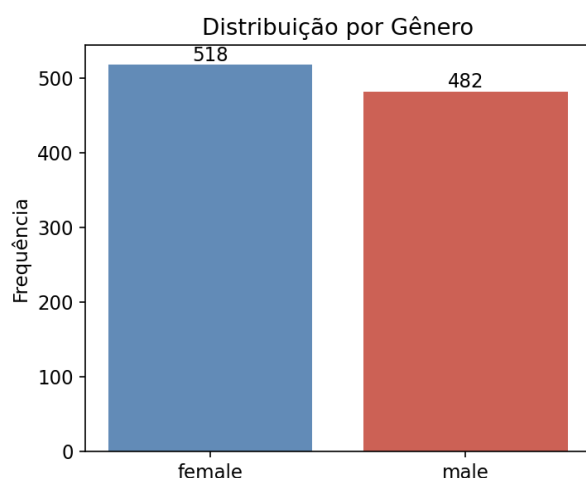


Figura 3 — Distribuição por gênero.

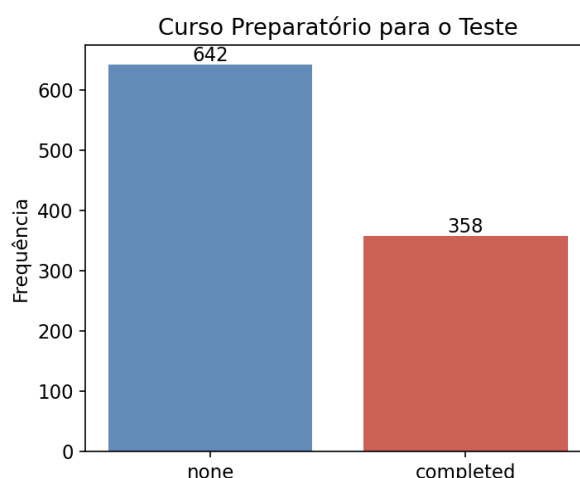


Figura 4 — Distribuição por curso preparatório.

A base é praticamente equilibrada entre os gêneros (51,8% mulheres e 48,2% homens). Já em relação ao curso preparatório, a maioria dos estudantes (64,2%) **não** realizou o curso de preparação para o teste, enquanto apenas 35,8% o completaram — um desbalanceamento relevante que será explorado na análise bivariada.

5. Análise Bivariada

5.1 Matemática × Escrita (quantitativa × quantitativa)

Figura 5 — Dispersão entre nota de matemática e nota de escrita, com reta de tendência.

O coeficiente de correlação de Pearson entre as notas de matemática e escrita é $r = 0,80$, indicando uma associação linear positiva forte: alunos com boas notas em matemática tendem a apresentar boas notas em escrita também. Isso sugere que o desempenho acadêmico está mais associado a características gerais do aluno (dedicação, preparo, contexto) do que a habilidades específicas de cada disciplina isoladamente. Vale notar que a correlação entre leitura e escrita é ainda mais forte ($r = 0,95$), o que é esperado por serem habilidades linguísticas próximas.

5.2 Curso preparatório × Nota de matemática (qualitativa × quantitativa)

Figura 6 — Boxplot da nota de matemática segundo a realização do curso preparatório.

Curso preparatório	Média	Desvio padrão	n
Completo	69,70	14,44	358
Não fez	64,08	15,19	642

Estudantes que completaram o curso preparatório apresentam, em média, **5,6 pontos a mais** em matemática (69,7 contra 64,1) e essa diferença é ainda maior em escrita (74,4 contra 64,5, uma diferença de quase 10 pontos — não mostrado no gráfico, mas calculado na mesma base). Isso indica que a preparação prévia está associada a um ganho consistente de desempenho em todas as provas, sendo um dos padrões mais claros encontrados na base.

5.3 Gênero × Curso preparatório (qualitativa × qualitativa)

Figura 7 — Proporção de realização do curso preparatório, por gênero.

Gênero	Completo (%)	Não fez (%)
Feminino	35,5%	64,5%
Masculino	36,1%	63,9%

As proporções de realização do curso preparatório são praticamente idênticas entre os gêneros (35,5% entre mulheres e 36,1% entre homens). O teste qui-quadrado de independência resultou em $p \approx 0,90$, muito acima de 0,05, confirmando que **não há associação estatisticamente significativa** entre gênero e realização do curso preparatório — ou seja, essas duas variáveis são praticamente independentes nesta amostra. Isso é interessante porque, apesar disso, homens e mulheres apresentam desempenhos médios diferentes por disciplina: os meninos têm média maior em matemática (68,7 vs. 63,6) e as meninas têm média maior em escrita (72,5 vs. 63,3), diferenças que não são explicadas pelo curso preparatório, mas por outros fatores não capturados nesta base.

5.4 Complemento: escolaridade dos pais × nota de matemática

Figura 8 — Média de matemática por nível de escolaridade dos pais.

Observa-se uma tendência geral de aumento da nota média de matemática conforme aumenta o nível de escolaridade dos pais: de 62,1 pontos, o valor mais baixo, registrado entre alunos cujos pais concluíram apenas o ensino médio (*high school*), até 69,7 pontos entre os que têm pai ou mãe com mestrado (*master's degree*) — uma diferença de aproximadamente 7,6 pontos entre os extremos. O padrão não é perfeitamente monotônico: chama atenção o fato de os filhos de pais que **não concluíram** o ensino médio (*some high school*, média 63,5) apresentarem nota ligeiramente **maior** do que os filhos de pais que concluíram o ensino médio (*high school*, média 62,1) — uma inversão pequena (1,4 ponto), possivelmente dentro da margem de ruído amostral, que não compromete a tendência geral de associação positiva entre o capital educacional da família e o desempenho do estudante observada a partir de "some college" em diante.

6. Discussão dos Resultados

a) Principais padrões encontrados

- As três notas (matemática, leitura e escrita) são fortemente correlacionadas entre si, especialmente leitura e escrita ($r = 0,95$).
- Concluir o curso preparatório está associado a ganhos consistentes de desempenho em todas as disciplinas (cerca de 5 a 10 pontos a mais, em média).
- Existe uma diferença de desempenho por gênero: homens têm médias maiores em matemática e mulheres em escrita/leitura, mesmo com taxas de curso preparatório equivalentes entre os grupos.
- O tipo de refeição escolar (lunch) também está associado ao desempenho: alunos com refeição padrão (standard) têm média de matemática de 70,0 contra 58,9 dos que recebem refeição gratuita/reduzida, um proxy indireto de condição socioeconômica.
- Maior escolaridade dos pais está associada, em média, a notas mais altas dos filhos.

b) Variáveis com forte assimetria

As três variáveis de nota apresentam assimetria leve e negativa (entre -0,26 e -0,29), ou seja, uma cauda discretamente mais longa em direção às notas baixas. Não se trata, porém, de uma assimetria forte — os valores estão próximos de zero, o que é coerente com distribuições relativamente simétricas em torno da média, com leve puxão para baixo causado pelos outliers de baixo desempenho identificados na Seção 3.3.

c) Outliers relevantes

Sim: foram identificados de 5 a 8 outliers por variável de nota, todos no extremo inferior, com destaque para um estudante com nota 0 em matemática e notas muito baixas nas demais provas simultaneamente. Esses casos são relevantes para identificar alunos em situação de risco acadêmico, mas não foram removidos da análise por representarem desempenho real (e não erro de medição).

d) Relações interessantes entre variáveis

A relação mais forte e consistente encontrada foi entre a realização do curso preparatório e o desempenho nas provas — um fator sobre o qual a escola/aluno tem controle direto, ao contrário de variáveis demográficas fixas. Também chama atenção a relação entre tipo de refeição escolar

(um proxy de renda familiar) e desempenho, sugerindo que fatores socioeconômicos têm peso relevante nos resultados.

e) Informações úteis para tomada de decisão

- Incentivar e ampliar o acesso ao curso preparatório pode ser uma intervenção custo-efetiva para elevar o desempenho médio dos alunos, já que a diferença observada é grande e atinge todas as disciplinas.
- A associação entre tipo de refeição (proxy de renda) e desempenho reforça a importância de políticas de apoio socioeconômico (merenda, bolsas, material didático) como estratégia de equidade educacional.
- Diferenças de desempenho por gênero entre disciplinas podem orientar ações pedagógicas direcionadas (ex.: reforço em escrita para meninos, em matemática para meninas), sem que isso implique determinismo, mas sim uma leitura de tendência estatística da amostra.

7. Conclusão

A análise exploratória da base *Students Performance in Exams* mostrou uma base de dados completa, sem valores ausentes ou inconsistências relevantes, com poucos outliers pontuais no extremo inferior das notas. As três notas analisadas (matemática, leitura e escrita) apresentam distribuições aproximadamente simétricas e fortemente correlacionadas entre si, sugerindo que o desempenho acadêmico tende a ser um fenômeno geral do aluno, e não isolado por disciplina.

O achado mais robusto do trabalho é a associação positiva entre a realização do curso preparatório e notas mais altas em todas as disciplinas, aliado à evidência de que fatores socioeconômicos (tipo de refeição escolar, escolaridade dos pais) também se relacionam com o desempenho. Já a variável gênero não está associada à realização do curso preparatório, mas mostra diferenças relevantes de desempenho por disciplina.

Limitações da base

- A base não informa a instituição, período ou local dos estudantes, o que impede controlar fatores contextuais (qualidade da escola, região, ano letivo).
- As variáveis são majoritariamente categóricas amplas (ex.: apenas 'group A' a 'group E' para etnia, sem detalhamento), o que limita a profundidade da interpretação.
- Trata-se de dados observacionais e transversais (um único ponto no tempo), portanto as relações encontradas são associações e não podem ser interpretadas como causais.

Possíveis análises futuras

- Aplicar modelos de regressão múltipla para estimar o efeito conjunto e controlado de cada variável sobre as notas.
- Investigar interações entre variáveis (ex.: efeito do curso preparatório pode variar conforme o tipo de refeição escolar).
- Se disponíveis, incorporar dados longitudinais (mais de um período de avaliação) para analisar evolução do desempenho ao longo do tempo.