Título

1) Coleta de dados

A Prova Brasil é uma avaliação diagnóstica realizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep/MEC) com o intuito de avaliar os alunos de ensino básico do quinto ano e do nono ano do Ensino Fundamental. Assim, tem o objetivo de aferir a qualidade do ensino oferecido pelo sistema educacional brasileiro a partir de testes padronizados e questionários socioeconômicos.

A prova é realizada em escolas públicas e particulares. Pode-se declarar que a população alvo são os próprios alunos do quinto e nono anos. Entretanto, não temos o mesmo grupo como população acessível, visto que apenas escolas com mais de 20 alunos matriculados nesses anos são avaliadas, perdendo assim uma parte da população alvo.

As principais características levantadas incluem nome, idade, nota das provas (Portugues e Matemática), escolaridade do pai e da mãe, sexo, entre outras variáveis. Esses dados são coletados através da prova e de questionários feitos pelos alunos.

2) Variáveis

Segue uma tabela com todas as variáveis presentes nas amostras recebidas pelos alunos de matrícula 232014825, 232014905 e 232014880.

Variável	Tipo de variável	Escala	Descrição da variável
Ano	Quantitativa discreta	Nominal	Ano de realização da SAEB
Região	Qualitativa nominal	Nominal	Região onde é localizada a escola
Unidade Fedarativa	Qualitativa nominal	Nominal	UF da localização da escola
Município	Qualitativa nominal	Nominal	Município onde é localizada a escola
Área	Qualitativa nominal	Nominal	Área de localização do estudante
Dependência administrativa	Qualitativa nominal	Nominal	Âmbito de subordinação da escola
Localização	Qualitativa nominal	Nominal	Localização da escola
Nota em LP	Quantitativa contínua	Intervalar	Nota na prova de língua portuguesa
Nota em MAT	Quantitativa contínua	Intervalar	Nota na prova de matemática
Sexo	Qualitativa nominal	Nominal	Sexo do estudante
Cor	Qualitativa nominal	Nominal	Raça/etnia do estudante
Mês de Nasc.	Qualitativa nominal	Nominal	Mês de nascimento
Ano de Nasc.	Quantitativa discreta	Nominal	Ano de nascimento
Computador	Qualitativa ordinal	Razão	Se o aluno tem computador em casa
Esc. Mãe	Qualitativa ordinal	Ordinal	Escolaridade da mãe
Esc. Pai	Qualitativa ordinal	Ordinal	Escolaridade do pai
Tempo de tela	Qualitativa ordinal	Ordinal	Quanto tempo o estudante gasta em telas eletrônicas
Afazeres dom.	Qualitativa ordinal	Ordinal	Tempo gasto fazendo afazeres domésticos
Trabalho	Qualitativa nominal	Nominal	Se o aluno trabalha ou não
Perspectivas	Qualitativa nominal	Nominal	Perspectivas do futuro

Tabela 1: Explicação das Variáveis

3) Variáveis categóricas

As variáveis escolhidas para análise foram UF, sexo, cor, área e trabalho. Todas são variáveis qualitativas nominais.

UF

A variável UF se refere à unidade federativa da localização da escola. Na amostra, existem estudantes de escolas de todas as 27 UFs do país.

Unidade Federativa	Frequência absoluta	Frequência relativa
São Paulo	277	0,185
Minas Gerais	143	0,095
Bahia	113	0,075
Rio De Janeiro	95	0,063
Ceará	90	0,060
Pará	84	0,056
Pernambuco	78	0,052
Paraná	77	0,051
Rio Grande Do Sul	68	0,045
Maranhão	65	0,043
Santa Catarina	57	0,038
Goiás	49	0,033
Amazônas	39	0,026
Paraíba	33	0,022
Espírito Santo	30	0,020
Mato Grosso	30	0,020
Alagoas	28	0,019
Distrito Federal	27	0,018
Piauí	25	0,017
Rio Grande Do Norte	20	0,013
Mato Grosso Do Sul	17	0,011
Tocantins	15	0,010
Acre	11	0,007
Rondônia	11	0,007
Sergipe	11	0,007
Amapá	4	0,003
Roraima	3	0,002

Tabela 2: Quantidade de alunos por unidade federativa

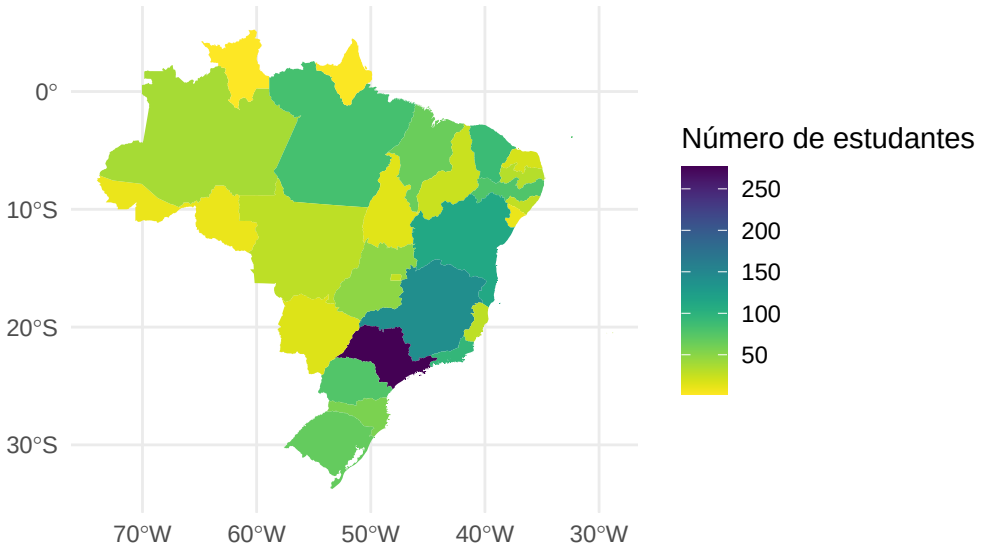

Gráfico 1: Mapa de calor dos estudantes por UF

A partir das informações mostradas pelo mapa e pela tabela, observa-se que as unidades federativas com maior número de estudantes que realizaram a prova são São Paulo, Minas Gerais e Bahia, concentrando 35,53% dos estudantes. Já os estados com menos estudantes foram Roraima e Amapá, com 3 e 4 estudantes, respectivamente.

Cor

A variável cor representa a etnia de cada estudante. É qualitativa nominal e está dividida em seis categorias: branca, parda, preta, indígena, amarela e não declarada. Segue uma tabela e um gráfico sua distribuição:

Cor	Frequência absoluta	Frequência relativa
Parda	690	48
Branca	396	27
Preta	175	12
Não quero declarar	92	6
Indígena	50	3
Amarela	49	3

Tabela 3: Quantidade de alunos pela cor

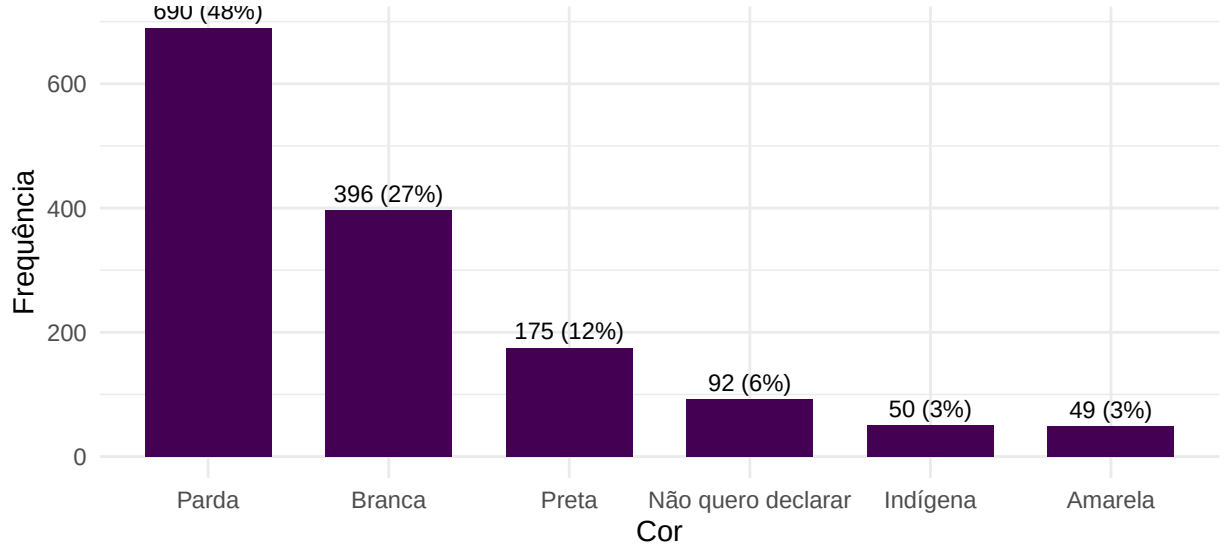

Gráfico 2: Gráfico de barras das etnias

Com base nos dados apresentados, é possível notar que a maioria dos estudantes são pardos, representando 48% do total de estudantes. As etnias com menor número de estudantes foram indígenas e amarelos, representando, juntas, apenas 6% dos estudantes.

Área

A variável área representa a região de localização da escola do estudante, na capital ou no interior. Segue uma tabela e um gráfico com sua distribuição:

Área	Frequência absoluta	Frequência relativa
Capital	255	0,17
Interior	1245	0,83

Tabela 4: Quantidade de alunos por área

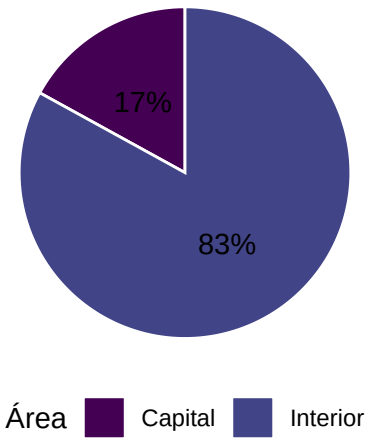

Gráfico 3: Gráfico de setores das áreas

A partir do gráfico, fica claro que há uma proporção muito maior de estudantes de escolas localizadas em cidades de interior do que em cidades de capital.

Sexo

A variável sexo indica o gênero dos estudantes que realizaram a prova. A seguir, é apresentada sua distribuição:

Sexo	Frequência absoluta	Frequência relativa
Feminino	756	0,52
Masculino	695	0,48

Tabela 5: Quantidade de alunos pelo sexo

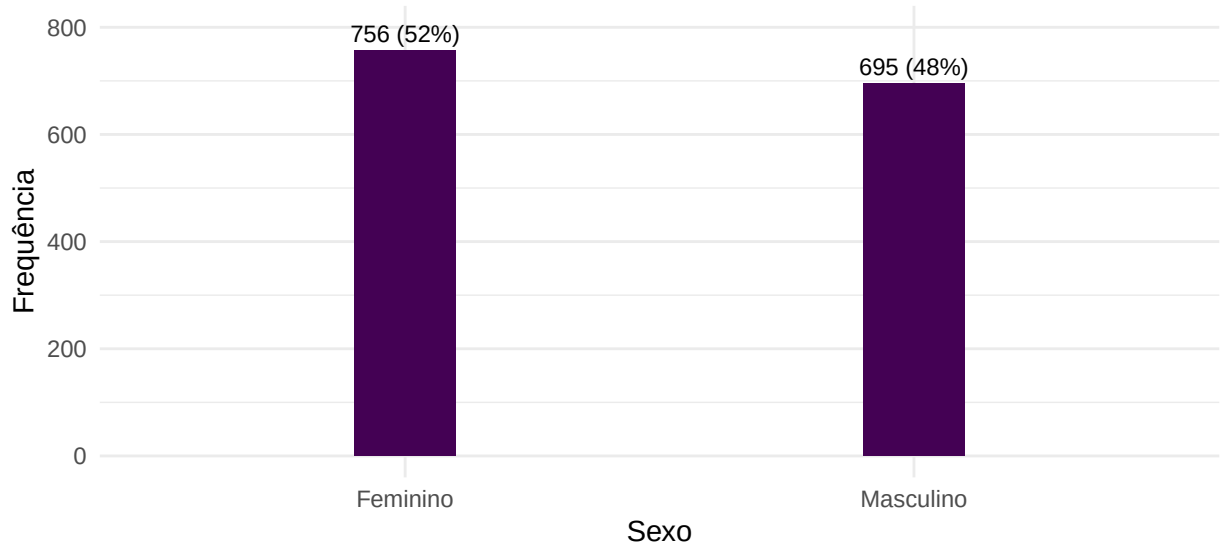

Gráfico 4: Gráfico de barra do sexo

Com base nas informações apresentadas, nota-se que o número de meninas e de meninos que realizaram a prova é semelhante, embora a proporção de meninas seja maior, representando 52% do total de estudantes.

Trabalho

A variável trabalho indica se o estudante trabalha ou não. A seguir, observa-se sua distribuição:

Trabalho	Frequência absoluta	Frequência relativa
Não	1268	0,87
Sim	193	0,13

Tabela 6: Quantidade de alunos pelo trabalho

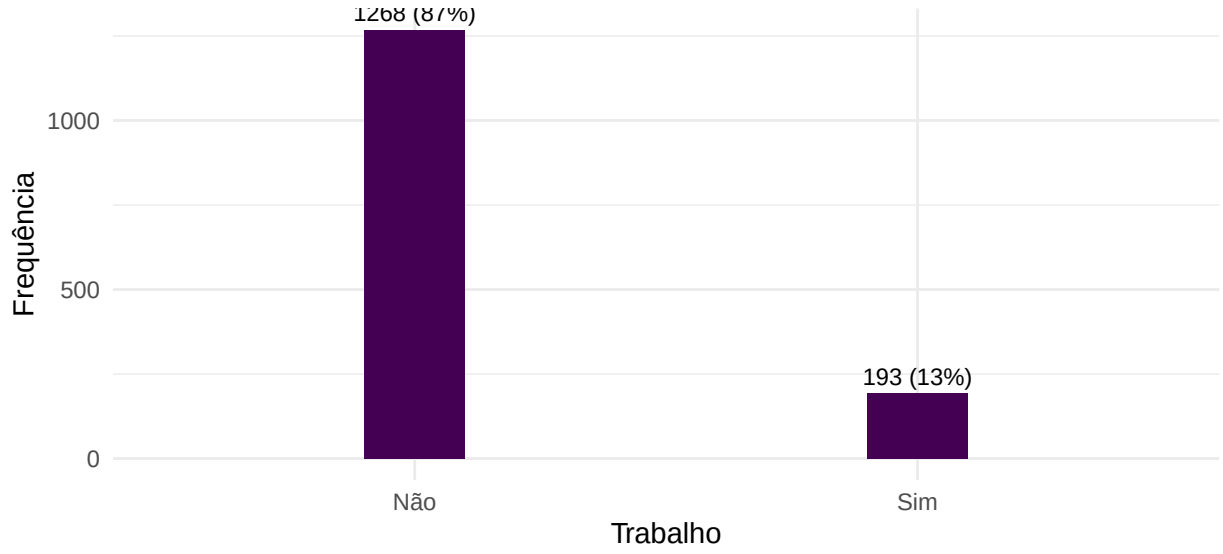

Gráfico 5: Gráfico de barras sobre trabalho

Analisando os dados apresentados, constata-se que 193 estudantes realizam algum tipo de trabalho, o que representa 13% do total.

4) Variáveis quantitativas

Proeficiência do aluno em Língua Portuguesa

A variável que mede a proeficiência do aluno em Língua Portuguesa é sua nota neste exame. É uma variável com uma escala intervalar criada pelos orgãos responsáveis pela avaliação. Ao fazermos uma tabela de sua distribuição e medidas-resumo, temos:

Tabela 7: Distribuição da nota dos alunos

Intervalo de Classe	Frequência
135 [——) **157**	27
157 [——) **179**	96
179 [——) **201**	111
201 [——) **222**	162
222 [——) **244**	238
244 [——) **266**	238
266 [——) **287**	238
287 [——) **309**	200
309 [——) **331**	114
331 [——) **352**	59
352 [——) **374**	17
Total	1500

Tabela 8: Medidas resumo

Média	Mediana	Erro padrão	Coef. de Variação	Coef. de Assimetria de Pearson	Curtose
253,529	255,934	48,805	0,193	-2,817	-0,571

Através da distribuição de frêquencias observamos uma alta concentração nos intervalos de nota 222 e 309. Intervalo que pega 60,86% das observações da amostra.

Ademais, ao calcular medidas de posição, como a média e a mediana da pontuação desse exame temos: 253,5 e 255,9, respectivamente. Por essas medidas serem muito próximas, suspeitamos de uma simetria, que ao calcular o coeficiente(de assimetria de Pearson), obtemos -2,8, um valor muito próximo a 0 que corroboram nossa suspeita de uma distribuição simétrica.

Calculando também medidas de dispersão como o erro padrão, temos que as notas variam, em média, ao redor de 48,8. Ao calcular o coeficiente de variação,que resultou em 0,19 numa escala de 0 a 1, observamos uma baixa variabilidade.

Finalizando com uma medida de curtose muito próxima de 0(-0,57), temos uma distribuição mesocúrtica na nossa amostra.

Entretando é necessária também a visualização gráfica dos dados. Ao fazer o gráfico histograma da nossa amostra de notas de Língua Portuguesa:

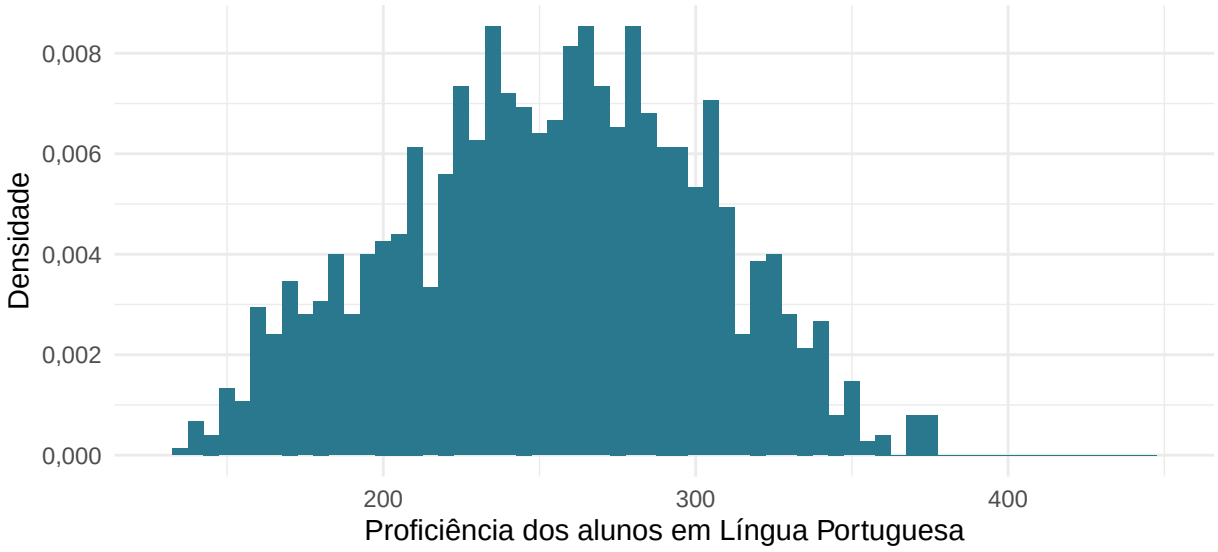

Gráfico 6: Histograma das notas de Língua Portuguesa

Analisando o histograma, temos que os valores se concentram próximos da média e poucas ocorrências nas pontas. Observamos também certa normalidade dos dados, visto que ele é perto da simetria e visto as outras características. Podemos também refazer o gráfico, mas dessa vez com uma linha caso os dados realmente tivessem normalidade usando a média e variância da nota de Língua Portuguesa.

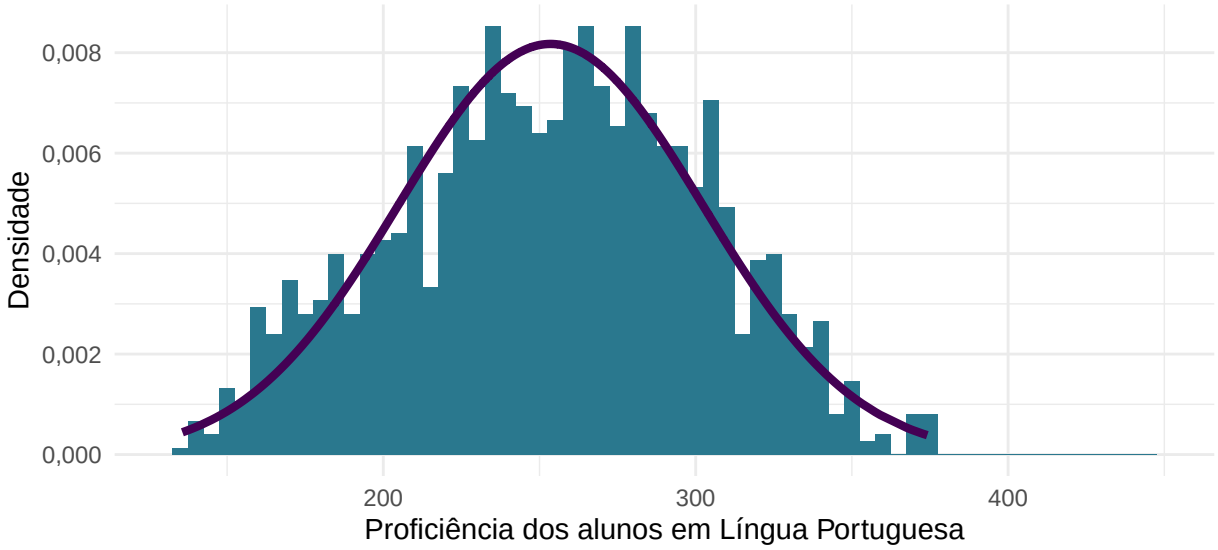

Gráfico 7: Histograma das notas de Língua Portuguesa com linha da normal

De fato, até que os dados amostrais coborram uma distribuição normal.

Como gráfico complementar para nossa análise, podemos usar o box-plot:

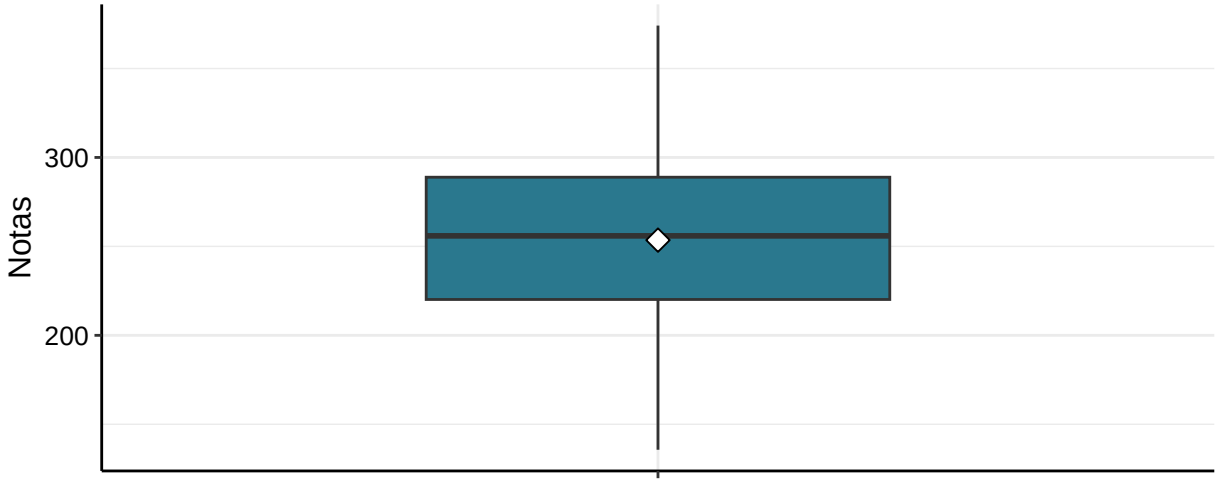

Gráfico 8: Box-plot das notas de LP

Pelo box-plot podemos observar que nossos dados são comportados e de fato, não observamos nenhum outlier, valores extremos e com alta distância do centro(média e mediana) da distribuição.

Proeficiência do aluno em Matemática

Tabela 9: Distribuição das notas dos alunos

Intervalo de Classe	Frequência
133 [——) **160**	37
160 [——) **187**	108
187 [——) **214**	201
214 [——) **240**	299
240 [——) **267**	290
267 [——) **294**	255
294 [——) **321**	193
321 [——) **347**	83
347 [——) **374**	28
374 [——) **401**	4
401 [——) **428**	2
Total	1500

Tabela 10: Medidas resumo

Média	Mediana	Erro padrão	Coef. de Variação	Coef. de Assimetria de Pearson	Curtose
250,859	249,162	48,681	0,194	3,786	-0,374

Por meio da distribuição de frequência das notas de matemática, conseguimos notar grande concentração entre as notas 187 e 294, intervalo que contém a proeficiencia de 1044(69,5%) dos estudantes.

Calculando medidas de posição, detectamos valores parecidos com as avaliações de Língua Portuguesa. Para matemática temos a media como 250,1 e mediana como 249,1. Da mesma maneira que anteriormente, podemos suspeitar de uma simetria. Ao verificar o coeficiente de assimetria de Pearson temos novamente um valor próximo de 0(3,78), dessa maneira é menos simétrico que as notas de Língua Portuguesa.

Além disso, em relação as medidas de dispersão, calculamos a variância, que resultou no erro padrão de 48,6. Computando o coeficiente de variação temos novamente 0,19, valor muito próximo de 0, indicando uma distribuição mais comportada.

Por fim, em relação à curtose, obtemos o coeficiente de excesso de curtose como -0,37, valor perto de 0. Dessa maneira podemos afirmar que como a distribuição das notas de Língua Portuguesa, a de matemática também é mesocúrtica.

A fim de complementar nossa compreensão dos dados, construimos alguns gráficos.

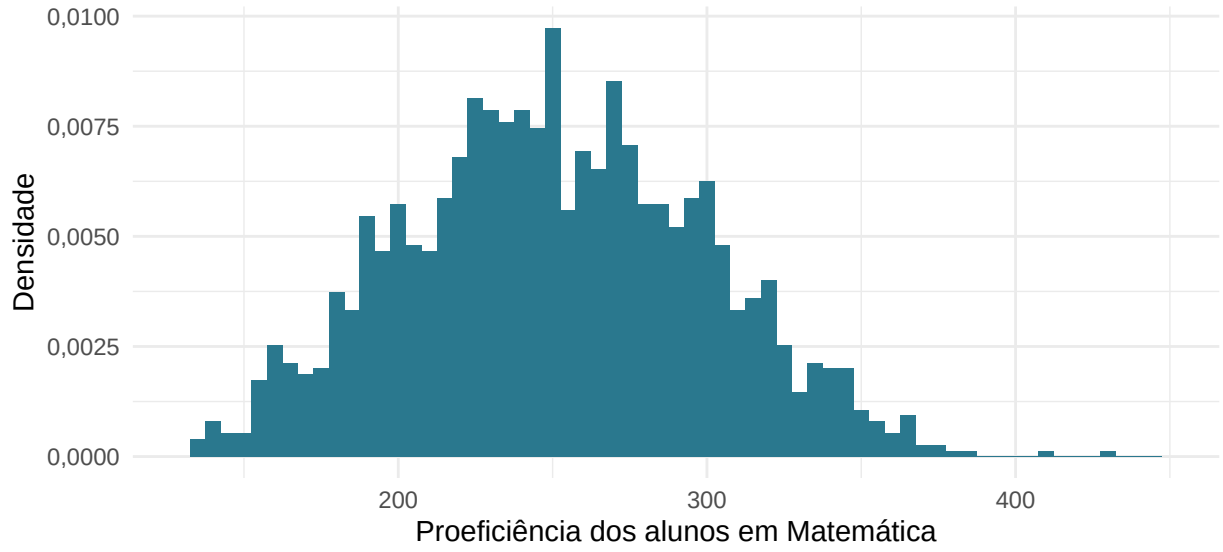

Gráfico 9: Histograma das notas de matemática com linha da normal

Interpretando o histograma, temos alta concentração de valores próximos a média e poucos valores nos extremos. Como o gráfico também parece simétrico, temos suspeitas de normalidade. Ademais, é possível observar duas barras em valores muito mais altos que os demais. Da mesma forma que fizemos com os dados da nota para Língua Portuguesa, podemos também para as notas da avaliação de Matemática, traçar uma linha para mostrar a normalidade, como a variância e média observada desses dados amostrais.

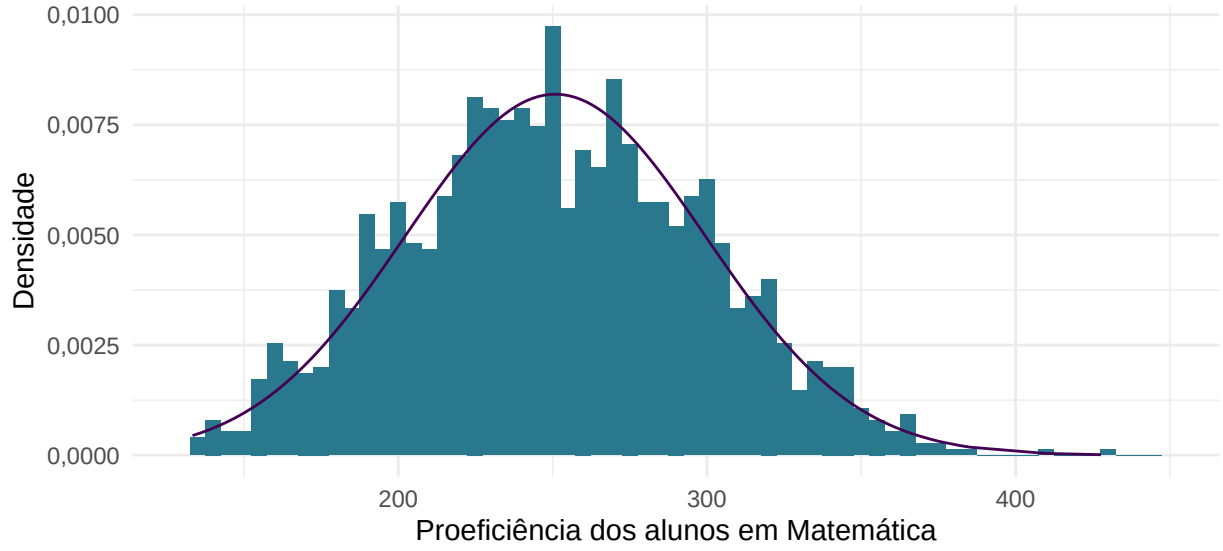

Gráfico 10: Histograma das notas de matemática

Novamente, temos os dados se encaixando bem nesse molde da normal.

Construindo um box-plot para concluir nossa análise, obtemos:

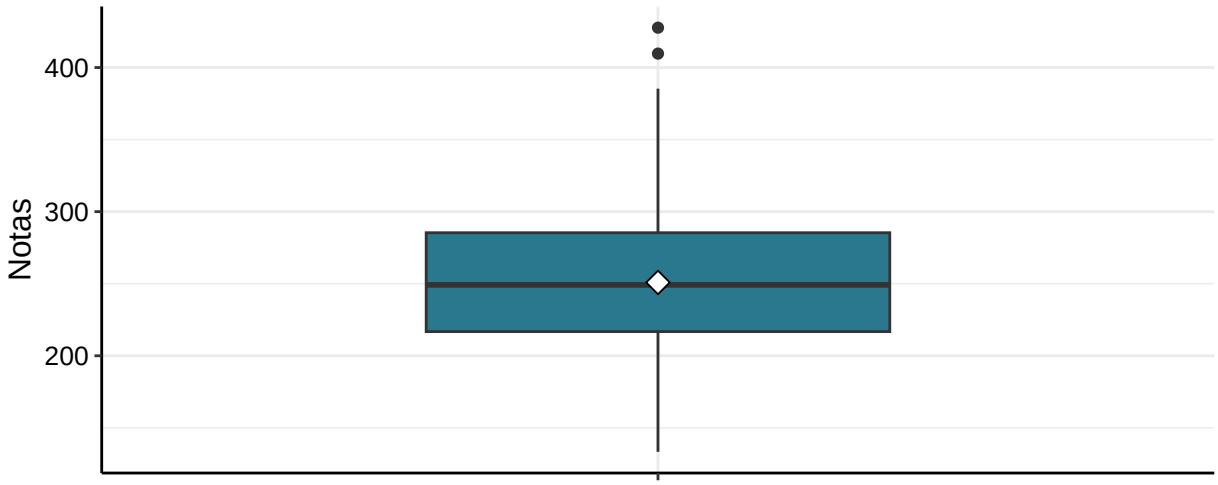

Gráfico 11: Box-plot das notas de Matemática

Este box-plot nos mostra de primeiro glance, dois outliers, sendo eles 427,65 e 409,65, notas que são incrivelmente maiores que a média dos demais. Esse gráfico também mostra que fora esses dois, os dodos se mostram mais comportados ainda, visto o intervalo interquartil(altura da caixa) menor.