



Trabalho de Estatística

Davi Lucas de Jesus Caetano

Henrique Hudson Maia

Camila de Souza Oliveira

Juliana Roncaglione

Bevilaqua Câmara

Isabella Rodrigues de Paulo

Paolla Pinheiro Pacheco

Relatório Base de dados

Niterói
2024

Introdução

Neste relatório, apresentamos uma análise estatística de dados obtidos a partir de uma pesquisa realizada em duas escolas (Gabriel Pereira e Mousinho da Silva) no curso de Português, a fim de encontrar tendências entre diferentes situações experienciadas pelos alunos. Entender padrões e as características que afetam o desempenho acadêmico é essencial para a elaboração de um ensino mais eficaz, e maior suporte aos alunos, atenuando possíveis diferenças socioeconômicas no ensino.

Metodologia

A base de dados pertence a duas escolas de Portugal, que usam um cálculo de notas diferente do utilizado no Brasil. A variável mais importante a ser analisada é a nota final (G3) que varia de 0 até 20, sendo 10 a aprovação e 20 a nota máxima e essa variável é o cerne do relatório. A manipulação dos dados foi feita no programa RStudio, com os códigos dos scripts em anexo. No total, foram escolhidas das 33 variáveis, 16 para o exercício das análises. As variáveis foram separadas em grupos com o objetivo de avaliar melhor os fatores por diferentes áreas da vida do aluno. As variáveis foram separadas nos grupos:

Educação
Família
Região/Situação econômica
Vida social/Saúde

Ao longo do relatório foram utilizados boxplots, gráficos de barra e tabelas para melhor visualização dos efeitos das variáveis junto do Grau de Associação (R^2), medida usada para quantificar relação entre a variável em questão e a Nota final, sendo sempre medido em (%), e considerado fraco (0% - 1%), moderado (1% - 3%) e forte (> 3%) nesta análise.

É válido ressaltar que a associação não indica necessariamente uma relação de causalidade entre as variáveis, apenas a relação entre elas.

Também é importante dizer que as análises são apenas observações gerais da base de dados, a Nota final depende de múltiplas variáveis que influenciam a Nota ao mesmo tempo podendo haver falhas de análises observacionais.

Resultados

Família

Dentro do grupo Família, foram analisadas o Suporte Escolar, (se o aluno teve ajuda adicional em português), e o Suporte Familiar (se teve algum parente ajudando nos estudos). Com o objetivo de saber se a ajuda externa que o aluno tem fora da aula normal, influencia nas notas finais.

As tabelas a seguir resumem as médias dos alunos que possuem ou não Apoio Escolar e possuem ou não Suporte Familiar:

Suporte Escolar	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Sim	11,28	2,30	10,60%	68
Não	11,97	3,33	89,40%	572

Tabela 1.1: Resumo sobre Suporte Escolar x Média de Notas

Suporte Familiar	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Sim	12,04	2,97	61,40%	393
Não	11,67	3,62	38,60%	247

Tabela 1.2: Resumo sobre Suporte Familiar x Média de Notas

Baseado nas tabelas, é possível reparar que embora exista a relação entre o Suporte Familiar e Suporte Escolar com as Notas, esta parece ser fraca, sendo necessária uma análise sobre o Grau de associação para exemplificar melhor esta relação.

Foi feito o Grau de Associação entre as Variáveis e a Nota final para quantificar a relação. Foram encontrados os valores de 0,34% entre o Suporte Escolar que indicam uma relação fraca, e a Nota final, e sendo 0,14% entre o Suporte Familiar e a Nota final, também fraca.

Os Boxplots a seguir demonstram melhor como essa variação ocorre para cada mudança nas variáveis em comparação com as médias:

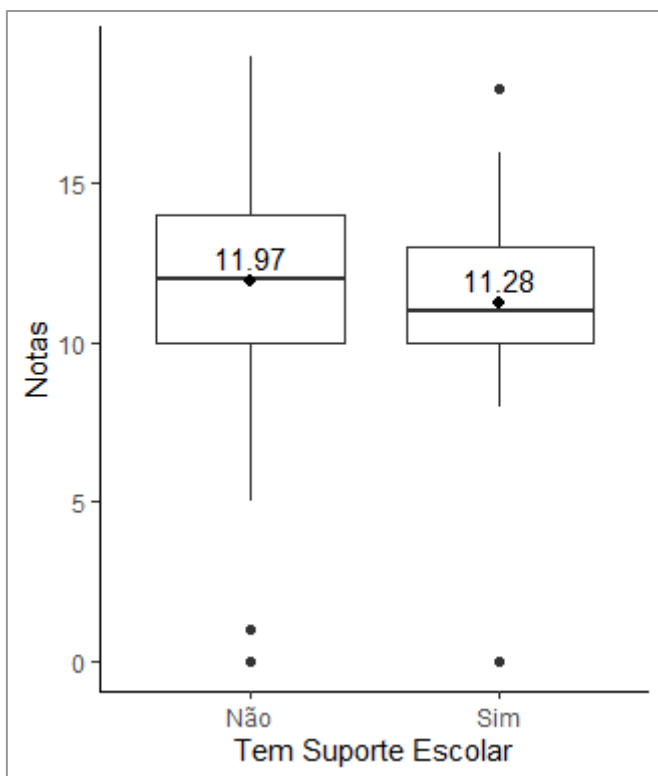


Figura 1.1: Gráfico da Distribuição de Notas pelo Suporte Escolar

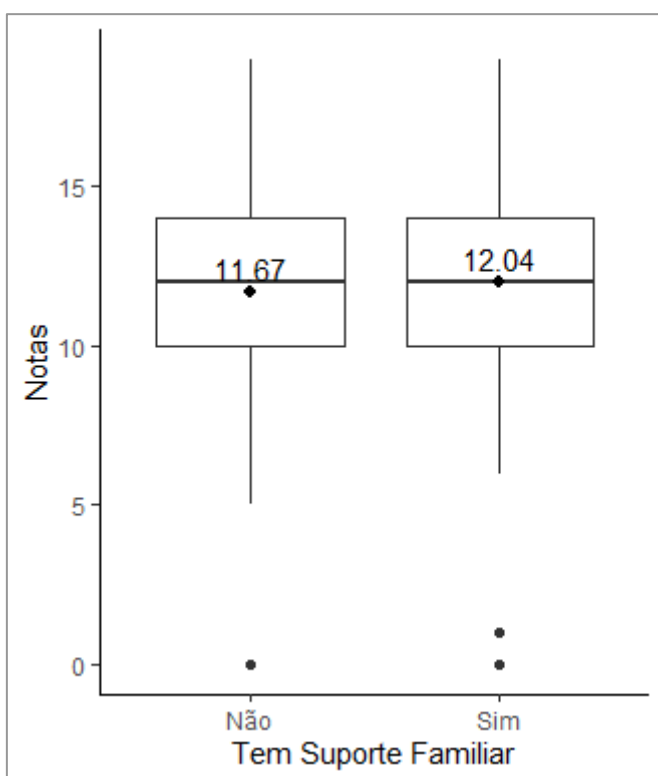


Figura 1.2: Gráfico da Distribuição das Notas por Suporte Familiar

Portanto é vista uma ligeira variação na média dentro dos grupos, mas esta não se distancia da média global mesmo com a diferença de proporção, fator que pode indicar que o impacto das duas variáveis parece não afetar tanto a nota.

Região/Situação econômica

Dentro deste grupo, foram analisadas a disponibilidade de Internet, a Localidade do aluno (Rural ou Urbano), o Tempo de Estudo, o Tempo de Deslocamento até as escolas e a preferência ou não do aluno de querer fazer Ensino Superior com o objetivo de entender os impactos dessas variáveis na Nota final.

Em primeiro momento foi analisado o impacto na nota baseado na presença ou falta de Internet e da Localidade dos alunos como mostra a seguir:

Possui Internet	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
Sim	12,16	3,12	76,71%	491
Não	11,00	3,46	23,29%	149

Tabela 2.1: Resumo sobre Internet x Média de Notas

Onde Vive	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
Rural	11,05	3,16	30,02%	194
Urbano	12,26	2,99	69,98%	446

Tabela 2.2: Tabela resumo sobre Localidade x Média de Notas

Percebe-se que existe uma diferença considerável nas proporções de ambas as tabelas indicando uma maioria vivendo na região urbana e tendo internet.

É visto também que quem possui internet tende a ter maior nota e quem vive em meio urbano também tende a tirar maiores notas. Embora exista uma diferença nessas proporções quanto ao número de alunos, a hipótese pode ser quantificada através do grau de associação entre as variáveis e a nota.

Foi feito o Grau de Associação entre as variáveis Internet e Localidade com as notas, buscando quantificar o quão forte é essa associação. Foi achado um valor de: 2,11% entre a Internet e a Nota, e 2,78% entre a Localidade e a Nota, ambos os valores considerados moderados.

Portanto, considerando os Graus de Associação entre as variáveis e as informações vistas nas tabelas, nota-se que a associação das variáveis e a nota demonstram algum efeito da ambientação do aluno e dos recursos disponíveis diretamente no valor das notas.

Em segundo momento, foram analisados o Tempo de Deslocamento do aluno até a escola (menor que 15 minutos, 15 a 30 minutos, 30 minutos a 1 hora, maior que 1 hora) e o Tempo de Estudo (menos que 2 horas, 2 a 5 horas, 5 a 10 horas, mais que 10 horas) e como essas variáveis influenciam na nota final dos alunos.

A tabela a seguir resume os valores das médias, e proporções dos alunos em cada categoria para cada variável a ser analisada:

Tempo de Deslocamento	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
< 15 min	12,24	3,13	56,25%	360
15 - 30 min	11,56	3,42	32,81%	210
30min - 1 hora	11,16	3,27	8,44%	54
> 1 hora	10,87	1,99	2,50%	16

Tabela 2.3: Resumo sobre Tempo de Deslocamento x Média de Notas

Tempo de Estudo	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
< 2 horas	10,85	3,22	56,25%	211
2 - 5 horas	12,07	3,25	32,81%	301
5 - 10 horas	13,24	2,53	8,44%	34
> 10 horas	13,08	3,07	2,50%	94

Tabela 2.4: Resumo sobre Tempo de Estudo x Média de Notas

Depreende-se uma relação entre as duas variáveis e a nota final, sendo diretamente proporcional o tempo de estudo e a Nota e inversamente proporcional o tempo de deslocamento e a Nota como observado nas tabelas.

Foi feito o Cálculo entre o Grau de associação das duas variáveis com a Nota para medir o quanto de impacto elas possuem na Nota final. Foram encontrados os valores de 1,27% entre o Tempo de Deslocamento e a Nota, e 6,41% entre o Tempo de Estudo e a Nota.

Isto denota uma relação forte com relação ao Tempo de Estudo e uma relação moderada entre o Deslocamento do aluno.

As tabelas demonstram que existe uma variação da nota dentro dos grupos de cada variável, porém, as notas médias e o desvio padrão seguem próximos do global. Logo, essa variação parece ser relevante. Os Gráficos a seguir resumem o que foi dito anteriormente neste segmento demonstrando as variações de cada fator junto das medias de cada categoria:

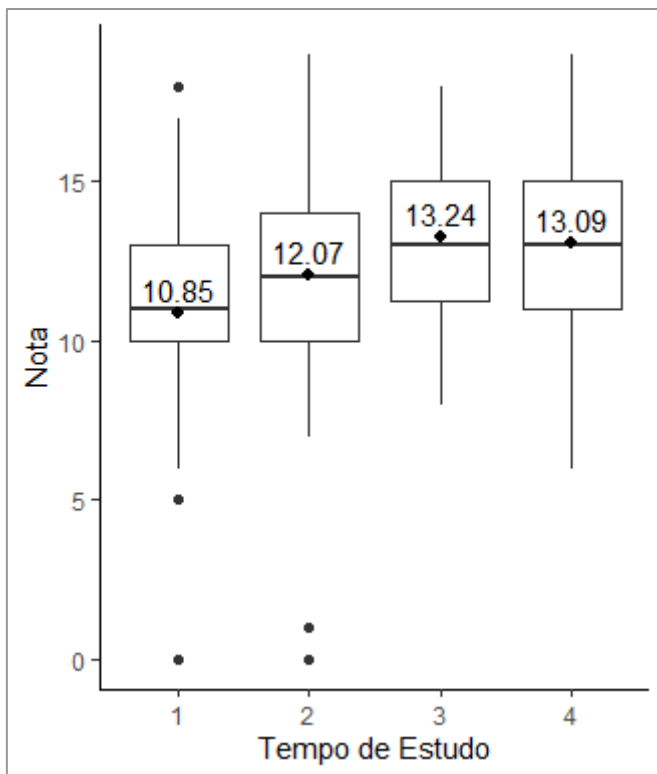


Figura 2.1:Distribuição das Notas por Tempo de Estudo

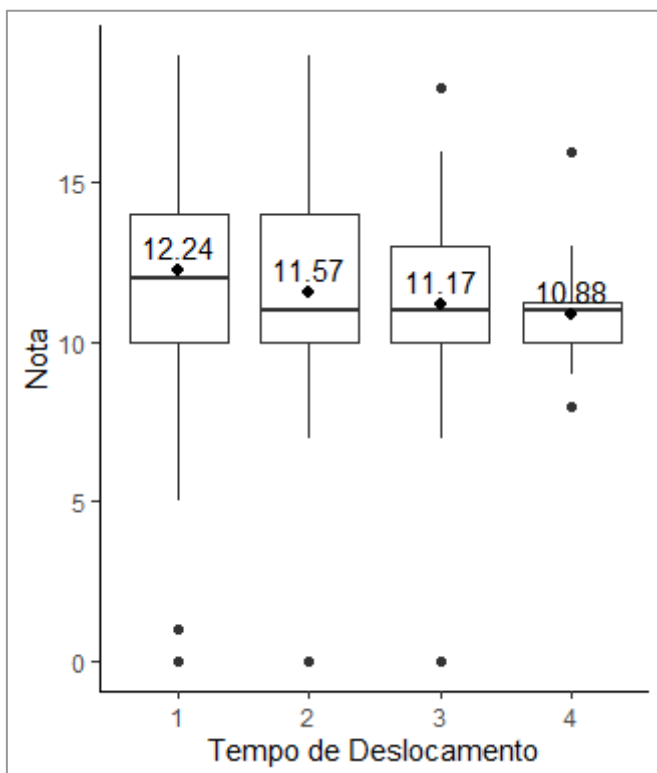


Figura 2.2:Distribuição das Notas por Tempo de Deslocamento

Vida social

Nesse grupo foram analisadas dentre as variáveis numa escala de 1-5, sendo 5 o mais alto, com o objetivo de saber se a vida social do aluno interfere na nota. Para isso foi analisado o Consumo de Álcool nos finais de semana (1-5), Consumo de Álcool diário (1-5), Relacionamento (Namoram ou Não namoram), se saem com amigos (1-5), Tempo livre após a aula (1-5) e o Estado atual de Saúde (Muito ruim, Ruim, Moderado, Bom, Muito bom)

Em primeiro momento foram feitas análises sobre o consumo de álcool diário e semanal, para descobrir se as quantidades de álcool ingeridas sugerem um impacto nas notas. As tabelas a seguir resumem os valores dos fatores:

Consumo de Álcool Diário	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
Muito Baixo	12,29	3,12	69,38%	444
Baixo	11,38	3,33	18,75%	120
Moderado	11,02	2,14	6,56%	42
Alto	8,94	5,12	2,66%	17
Muito Alto	10,23	2,94	2,66%	17

Tabela 3.1: Resumo sobre Consumo de Álcool Diariamente x Média de Notas

Consumo de Álcool Semanal	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
Muito Baixo	12,36	3,08	37,81%	242
Baixo	12,26	3,09	23,28%	149
Moderado	11,66	3,65	18,44%	118
Alto	10,97	2,76	13,44%	86
Muito Alto	10,55	3,58	7,03%	45

Tabela 3.2: Resumo sobre Consumo de Álcool Semanalmente x Média de Notas

Como visto nas tabelas, o consumo de Álcool Diário parece confuso dado o alto desvio padrão e valores discrepantes em algumas classes. Então foi feito um Boxplot que ajudasse a visualizar e identificar esses valores que pudessem atrapalhar na análise:

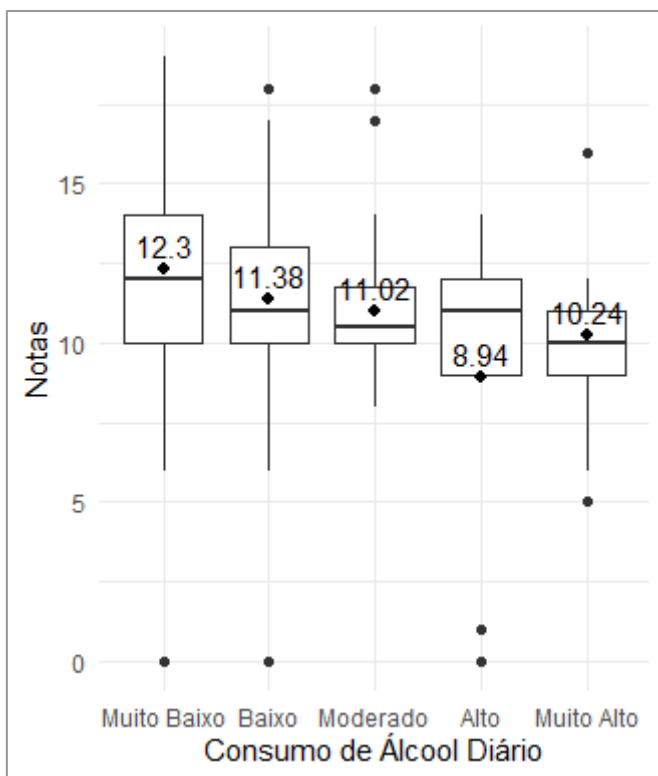


Figura 3.1: Distribuição das Notas pelo Consumo de Álcool Diário

Após a retirada dos valores discrepantes, a tabela foi refeita com informações mais precisas e correspondentes com a análise:

Consumo de Álcool Diário	Médias de Notas	Proporção (em %)	Total
Muito Baixo	12,29	70,50%	444
Baixo	11,38	18,93%	120
Moderado	10,70	6,31%	40
Alto	8,94	2,05%	13
Muito Alto	9,78	2,21%	14

Tabela 3.3: Tabela ajustada sobre Consumo de Álcool Diário x Média de Notas

Essa mudança indica que não parece haver relação entre o consumo de álcool diário e as notas, mesmo com grande variação nos diferentes níveis de consumo.

Foi calculado o Grau de associação entre as variáveis e a nota a fim de quantificar o impacto delas na nota final. Foram encontrados valores de 4,14% para o Consumo de Álcool Diário e 2,82% para o Consumo de Álcool Semanal, sendo consideradas associações forte e normal respectivamente. Isso indica que o Grau de associação nesse caso pode não ser uma boa métrica para resumir o impacto que as variáveis possuem na nota pela alta variação da proporção nas categorias.

Em segundo momento, foi analisado se os alunos namoram, e as Saídas com amigos, com o objetivo de saber se o tempo não gasto em estudo influencia na nota final

As tabelas a seguir resumem as informações colhidas pela análise dos dados:

Namora	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Sim	11,49	3,58	36,72%	235
Não	12,12	3,01	63,28%	405

Tabela 3.4: Resumo sobre Relacionamento do aluno x Média de Notas

Saídas com amigos	Médias de Notas	Desvio Padrão	Proporção (em %)	Total
Muito Baixo	10,70	3,88	7,34%	47
Baixo	12,66	3,19	22,19%	142
Moderado	12,14	2,92	31,56%	202
Alto	11,99	2,81	21,88%	140
Muito Alto	10,82	3,70	17,03%	109

Tabela 3.5: Resumo sobre Saídas dos Alunos x Média de Notas

Analisando as tabelas, ambas variáveis relacionadas com a nota não parecem ter relação forte, apresentam, porém, ligeira variação dentro de suas próprias definições, mas não afetam tanto a nota final.

Por último foi analisado o Tempo Livre e o Estado de Saúde do aluno. Presume-se que quem tenha mais tempo livre tenha pior nota e quem tem um pior estado de saúde também tenha pior nota.

As tabelas a seguir resumem as informações dos fatores e demonstram as variações para cada grupo:

Tempo livre após a aula	Médias	Desvio Padrão	Proporção (em %)	Total
Muito Baixo	11,65	2,47	6,87%	44
Baixo	12,68	3,49	16,25%	104
Moderado	12,06	3,05	38,75%	248
Alto	11,72	3,07	27,50%	176
Muito Alto	10,69	3,98	10,63%	68

Tabela 3.6: Resumo sobre Relacionamento do aluno x Média de Notas

Estado de Saúde	Médias	Desvio Padrão	Proporção (em %)	Total
Muito Ruim	11,65	3,26	13,91%	89
Ruim	12,16	3,34	11,71%	75
Moderado	11,83	3,17	18,91%	121
Bom	12,33	3,00	16,56%	106
Muito Bom	11,46	3,30	38,91%	249

Tabela 3.7: Resumo sobre Saúde do aluno x Média de Notas

Portanto, é visto que não existe relação entre Tempo Livre, Estado de Saúde do aluno comparado com as Notas, dado que as médias permanecem bem próximas a média global.

Educação

Dentro do grupo Educação, foram analisadas as Escolas Gabriel Pereira (GP) e Mousinho da Silva (MS), Sexo (Masculino e Feminino), preferência por Ensino Superior, e se o aluno teve aulas extras pagas.

As tabelas a seguir resumem os valores encontrados das variáveis citadas e seus impactos nas notas:

Escola	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Gabriel Pereira	12,57	2,63	64,69%	414
Mousinho da Silva	10,65	3,83	35,31%	226
Total	11,89	3,24	100%	640

Tabela 4.1: Resumo sobre Escola x Média de Notas

Sexo	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Masculino	11,39	3,33	40,94%	262
Feminino	12,24	3,14	59,06%	378

Tabela 4.2: Resumo sobre Sexo x Média de Notas

Ensino Superior	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Quer	12,26	3,07	89,53%	573
Não quer	8,74	2,99	10,47%	67

Tabela 4.3: Resumo sobre Preferência de Ensino Superior x Média de Notas

Aulas Extras Pagas	Média de Notas	Desvio Padrão	Proporção (em %)	Total
Teve	11,18	3,26	5,94%	38
Não teve	11,94	2,88	94,06%	602

Tabela 4.4: Resumo sobre Aulas Particulares pagas x Média de Notas

Ao olhar as tabelas, depreende-se que as variáveis que mais impactam na nota é a diferença entre as Escolas e a escolha ou não de Ensino Superior. A diferença de proporções pode indicar o motivo das variações nas médias, mas é necessário quantificar a relação para provar a hipótese.

Foi feito o Grau de Associação entre essas variáveis e a nota final, com o objetivo de mensurar o impacto delas na nota final dos alunos. A variável Escola possui um grau de associação de 7,89% e a escolha de ensino superior possui 10,9%, ambas relacionadas a nota final. As variáveis possuem forte associação com a nota final, mas não necessariamente causam o resultado direto na nota.

Foi feita também uma análise sobre as notas 0 da base de dados, para identificar possíveis relações entre variáveis e a nota mínima:

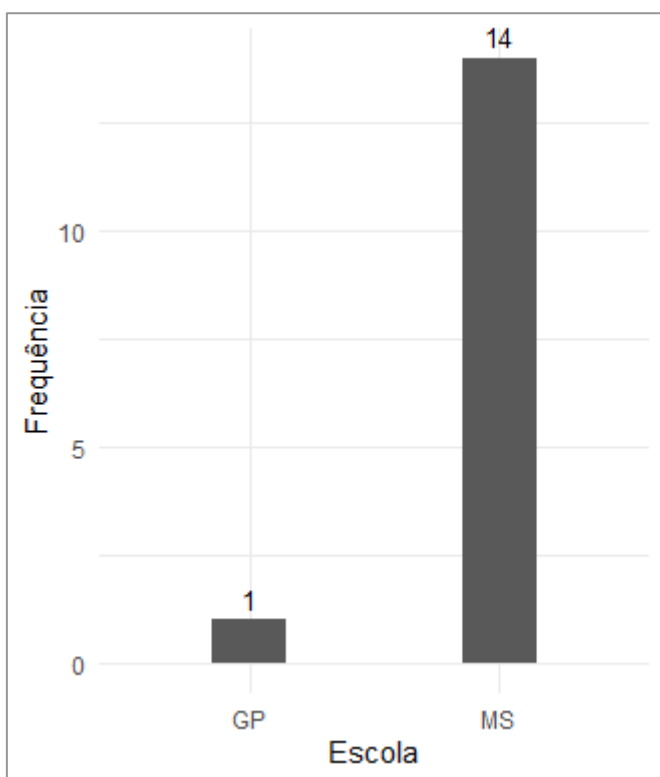


Figura 4.1: Gráfico da Frequência de Notas 0 por Escola

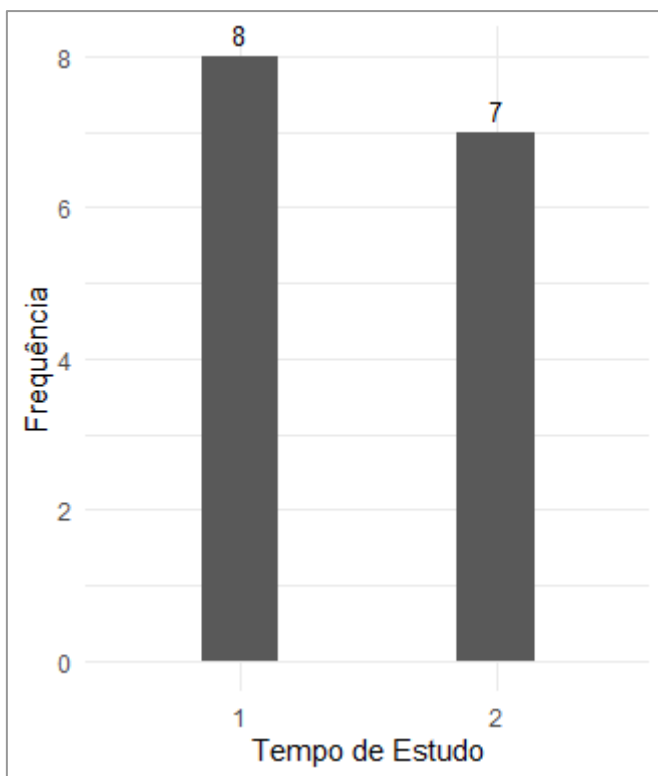


Figura 4.2:Gráfico da Frequência de Notas 0 por Tempo de Estudo

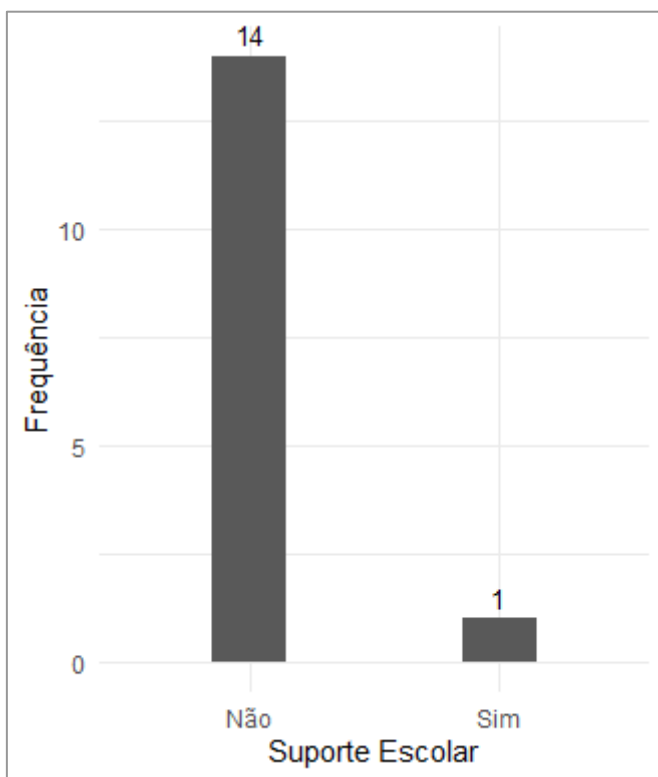


Figura 4.3:Gráfico da Frequência de Notas 0 Suporte Escolar

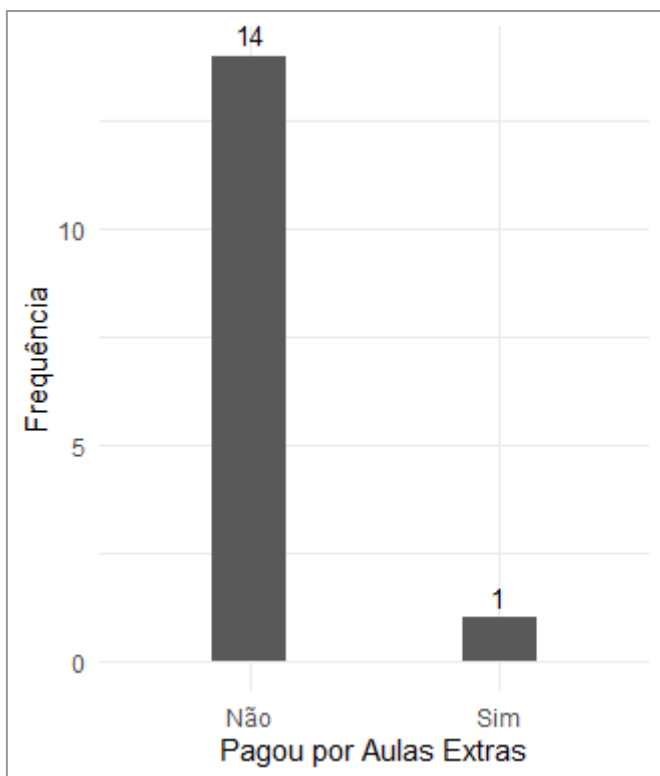


Figura 4.4: Gráfico da Frequência de Notas 0 por Aulas Extras Pagas

Em resumo, os alunos que tiraram nota 0 foram os que menos tiveram acesso a estudo extra, e existe uma disparidade tanto na média quanto nas notas 0, com relação à Escola Mousinho da Silva, que possui menores notas no geral.

Conclusão

Este relatório teve como premissa, a análise do desempenho acadêmico de alunos de duas escolas, em português, e os motivos das diferentes variações na nota final. A divisão de variáveis por grupos separou os efeitos socioeconômicos em vários âmbitos da vida dos estudantes.

Foi possível notar durante o relatório que a maioria tem uma relação com a Nota final de alguma forma, e para quantificar essa relação foi medido com o Grau de Associação (R^2), embora nem sempre a associação forte indique impacto direto na nota.

Na divisão dos grupos de variáveis foi possível notar que, quanto mais recursos o aluno tiver, mais próximo morar da escola e mais tempo dedicado ao estudo, tendem a ser melhores suas notas. E foi visto que o âmbito social do estudante de forma geral não parece afetar tanto a Nota final mesmo com a associação considerada forte e moderada entre as variáveis e a Nota.

Foi também encontrada uma possível disparidade do ensino entre as Escolas, dada a diferenças nas médias e na proporção de alunos, indicando que a Escola Gabriel Pereira pode obter melhores resultados acadêmicos por motivos não indicados na base de dados.

Considerando a limitação de dados socioeconômicos detalhados, uma análise mais precisa e com maior quantidade de dados seria necessária para entender melhor os motivos que possam afetar o desempenho dos alunos, visto que, existe multiplicidade de fatores tendo influência ao mesmo tempo na Nota final, e as análises foram feitas apenas observando e filtrando a base de dados sem modificação das informações coletadas.

De forma geral, os alunos que possuem melhores condições de vida e estudam mais, tendem a tirar maior nota. Esses achados sugerem a necessidade de melhores estratégias de suporte para os alunos com o objetivo de atenuar os impactos que as diferenças econômicas e sociais possam ter nas notas dos alunos e assim melhorar a educação nas escolas.