Análise de Dados - Desempenho de Estudantes

Esse projeto tem como objetivo investigar os fatores que mais corroboram para notas de estudantes. Para tanto, tomou-se os dados do diretório público: Student Performance Factors (Insights into Student Performance and Contributing Factors), disponível na plataforma Kaggle. O Método escolhido foi a regressão linear haja vista que almeja-se detalhar se a presença de determida condição explica uma maior/ menor nota (dummies), ou se um incremento em determinado valor tem efeito positivo/ negativo sobre a nota (variáveis númericas).

No que tange aos dados disponíveis, a base de dados contempla tanto variáveis categóricas e ordenadas como variáveis numéricas, vide:

Atributo	Descrição
Hours_Studied	Número de horas dedicadas ao estudo por semana.
Attendance	Percentual de aulas frequentadas.
Parental_Involvement	Nível de envolvimento dos pais na educação do aluno (Baixo, Médio, Alto).
Access_to_Resources	Disponibilidade de recursos educacionais (Baixo, Médio, Alto).
Extracurricular_Activities	Participação em atividades extracurriculares (Sim, Não).
Sleep_Hours	Número médio de horas de sono por noite.
Previous_Scores	Notas de exames anteriores.
Motivation_Level	Nível de motivação do aluno (Baixo, Médio, Alto).
Internet_Access	Disponibilidade de acesso à internet (Sim, Não).
Tutoring_Sessions	Número de sessões de tutoria frequentadas por mês.
Family_Income	Nível de renda familiar (Baixo, Médio, Alto).
Teacher_Quality	Qualidade dos professores (Baixa, Média, Alta).
School_Type	Tipo de escola frequentada (Pública, Privada).
Peer_Influence	Influência dos colegas no desempenho acadêmico (Positiva, Neutra, Negativa).
Physical_Activity	Número médio de horas de atividade física por semana.
Learning_Disabilities	Presença de dificuldades de aprendizado (Sim, Não).
Parental_Education_Level	Nível de educação mais alto dos pais (Ensino Médio, Faculdade, Pós-graduação).

Atributo	Descrição
Distance_from_Home	Distância de casa para a escola (Perto, Moderada, Longe).
Gender	Sexo do aluno (Masculino, Feminino).
Exam_Score	Nota final do exame.

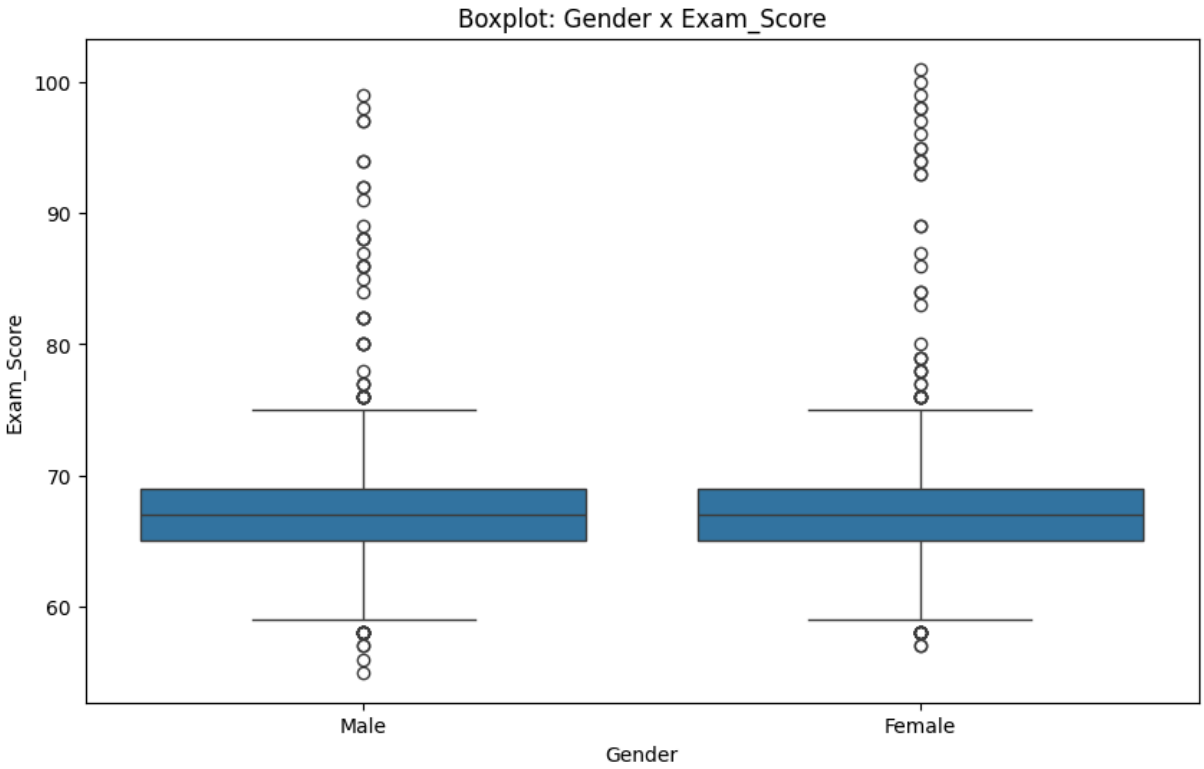
1.Análise Exploratória

O heatmap de correlações indica que as variáveis com maior potencial preditivo são: **horas de estudo, presença em aulas, notas anteriores e sessões de tutoria.**

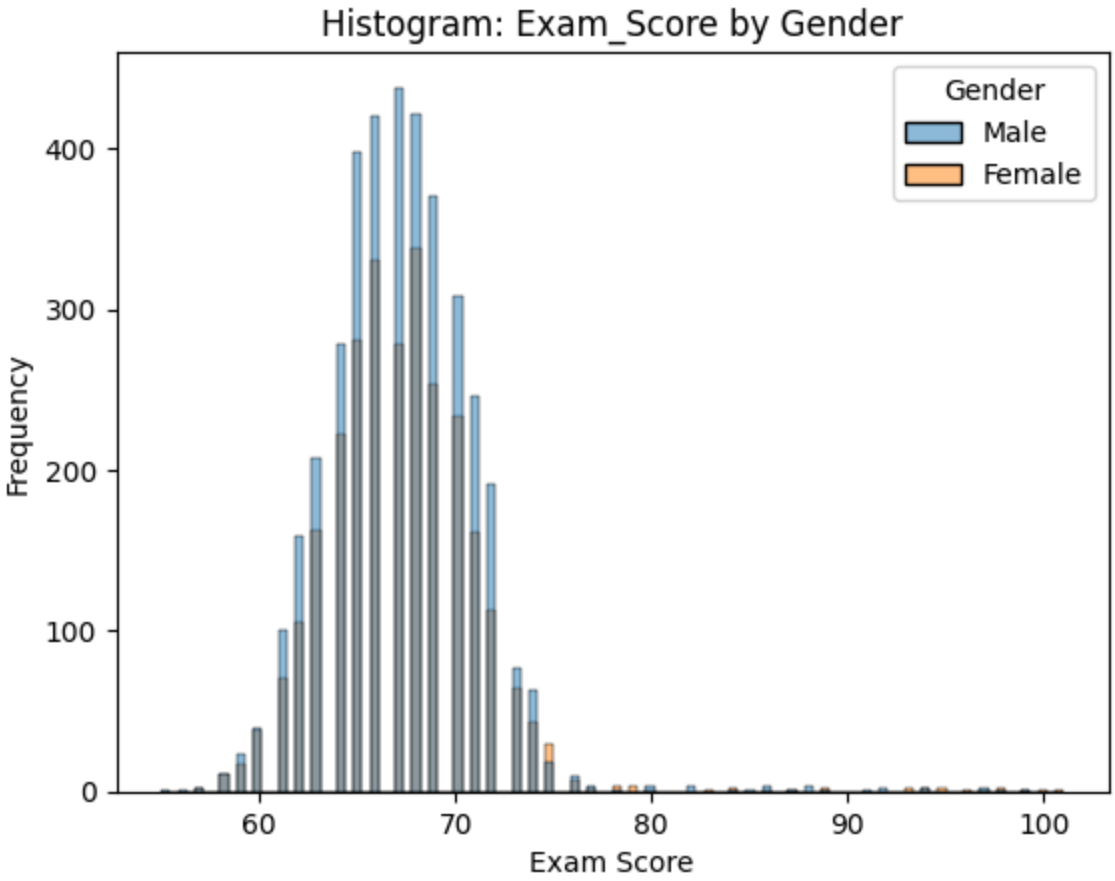
Categorias

- **Dados Demográficos:**
  - Gênero

- **Fatores Acadêmicos:**
  - Horas de Estudo
  - Percentual de Presença
  - Notas Anteriores
  - Nível de Motivação
  - Sessões de Tutoria
  - Tipo de Escola
  - Influência dos Colegas
  - Dificuldades de Aprendizagem
  - Qualidade do Professor
- **Fatores Socioeconômicos e Familiares:**
  - Nível de Renda Familiar
  - Nível de Educação dos Pais
  - Acesso a Recursos
  - Envolvimento dos Pais
  - Acesso à Internet
  - Distância de Casa para a Escola
- **Estilo de Vida:**
  - Horas de Sono
  - Atividades Extracurriculares
  - Atividade Física

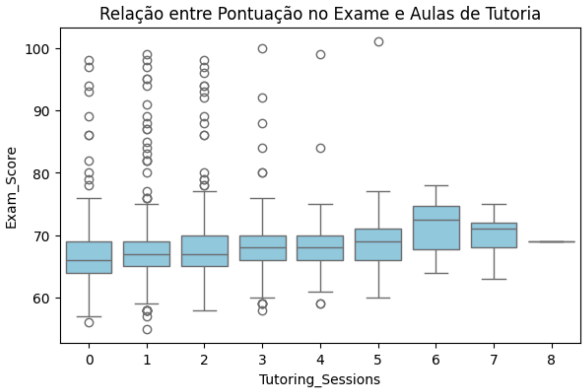


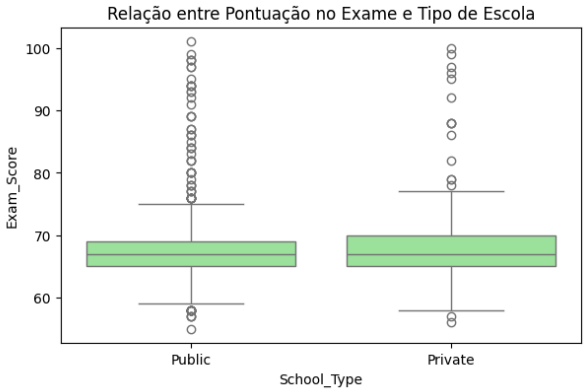
Gender	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Female	63.0	64.0	65.0	66.0	67.0	68.0	69.0	70.0	72.0
Male	63.0	64.0	65.0	66.0	67.0	68.0	69.0	70.0	72.0

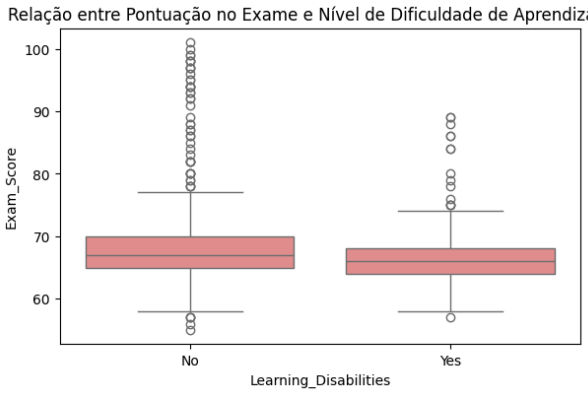


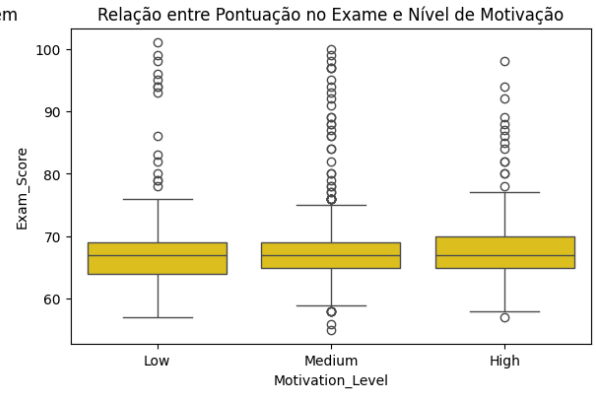
Fatores Demográficos

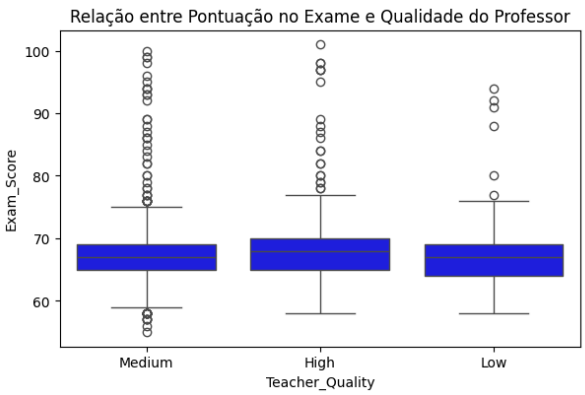
A tabela mostra que não há diferença significativa nas notas entre homens e mulheres em nenhum dos quantis. Ambos os gêneros apresentam uma progressão semelhante nos quantis, com as notas aumentando gradualmente do quantil 0,1 ao 0,9. Essa similaridade sugere que o gênero não tem um impacto significativo no desempenho nas provas, pelo menos não de acordo com os dados fornecidos.

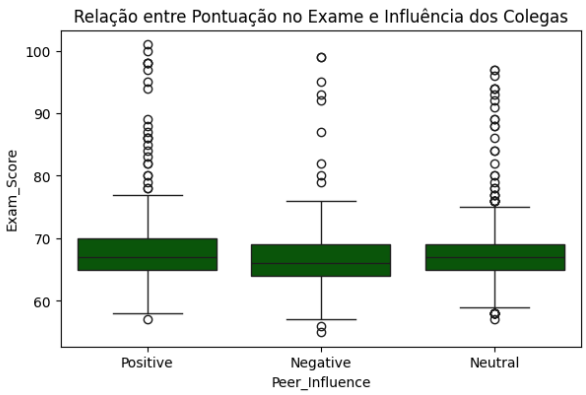

Relação entre Pontuação no Exame e Aulas de Tutoria
Exam_Score
Tutoring_Sessions
0
1
2
3
4
5
6
7
8
60
70
80
90
100


Relação entre Pontuação no Exame e Tipo de Escola
Exam_Score
School_Type
Public
Private
60
70
80
90
100


Relação entre Pontuação no Exame e Nível de Dificuldade de Aprendizagem
Exam_Score
Learning_Disabilities
No
Yes
60
70
80
90
100


Relação entre Pontuação no Exame e Nível de Motivação
Exam_Score
Motivation_Level
Low
Medium
High
60
70
80
90
100


Relação entre Pontuação no Exame e Qualidade do Professor
Exam_Score
Teacher_Quality
Medium
High
Low
60
70
80
90
100


Relação entre Pontuação no Exame e Influência dos Colegas
Exam_Score
Peer_Influence
Positive
Negative
Neutral
60
70
80
90
100

```
                    0.25  0.50   0.75
Tutoring_Sessions
0                  64.00  66.0  69.00
1                  65.00  67.0  69.00
2                  65.00  67.0  70.00
3                  66.00  68.0  70.00
4                  66.00  68.0  70.00
5                  66.00  69.0  71.00
6                  67.75  72.5  74.75
7                  68.00  71.0  72.00
8                  69.00  69.0  69.00

             0.25  0.50  0.75
School_Type
Private      65.0  67.0  70.0
Public       65.0  67.0  69.0

                        0.25  0.50  0.75
Learning_Disabilities
No                      65.0  67.0  70.0
Yes                     64.0  66.0  68.0

                   0.25  0.50  0.75
Motivation_Level
High               65.0  67.0  70.0
Low                64.0  67.0  69.0
Medium             65.0  67.0  69.0

                  0.25  0.50  0.75
Teacher_Quality
High              65.0  68.0  70.0
Low               64.0  67.0  69.0
Medium            65.0  67.0  69.0

                 0.25  0.50  0.75
Peer_Influence
Negative         64.0  66.0  69.0
Neutral          65.0  67.0  69.0
Positive         65.0  67.0  70.0
```

Fatores acadêmicos que influenciam as notas

Sessões de Tutoria

- O aumento do número de sessões de tutoria está geralmente associado a notas mais altas, especialmente no quantil 0,75 (percentil 75).
- No entanto, essa relação não é estritamente linear. O maior aumento nas notas parece ocorrer entre 0 e 6 sessões, com um aumento menos pronunciado após esse ponto.

Tipo de Escola

- Alunos de escolas particulares tendem a ter notas ligeiramente mais altas no quantil 0,75 em comparação com alunos de escolas públicas.
- As notas nos quantis 0,25 e 0,50 são semelhantes entre os tipos de escola.

Dificuldades de Aprendizagem

- Alunos sem dificuldades de aprendizado apresentam notas mais altas em todos os quantis em comparação com alunos com dificuldades de aprendizado.
- Essa diferença é consistente em todos os três quantis.

Nível de Motivação

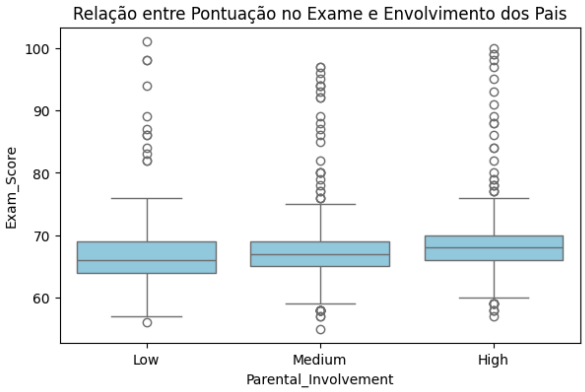
- Alunos com alto nível de motivação tendem a ter notas mais altas no quantil 0,75 em comparação com alunos com baixo ou médio nível de motivação.
- As notas nos quantis 0,25 e 0,50 são semelhantes entre os diferentes níveis de motivação.

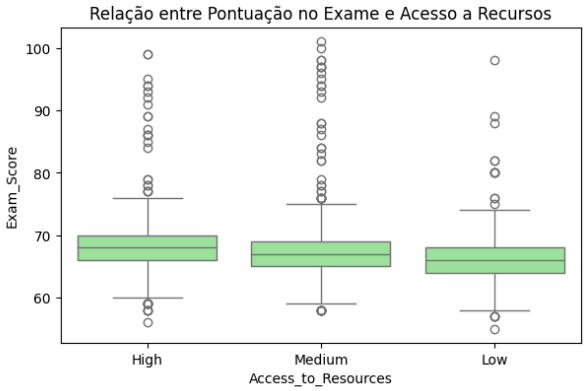
Qualidade do Professor

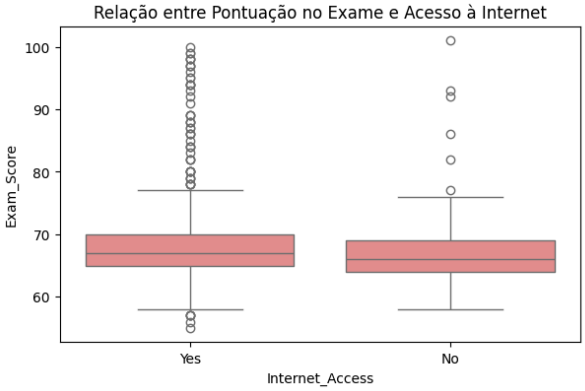
- Alunos que avaliaram a qualidade do professor como alta tendem a ter notas mais altas no quantil 0,50 e 0,75 em comparação com alunos que avaliaram a qualidade do professor como baixa ou média.
- As notas no quantil 0,25 são semelhantes entre os diferentes níveis de qualidade do professor.

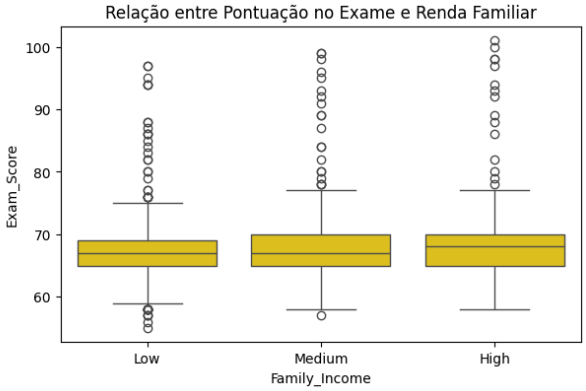
Influência dos Colegas

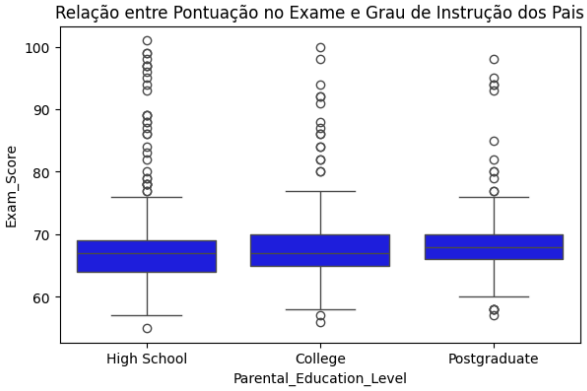
- Alunos que relatam influência positiva dos colegas tendem a ter notas mais altas no quantil 0,75 em comparação com alunos que relatam influência negativa ou neutra.
- As notas nos quantis 0,25 e 0,50 são semelhantes entre os diferentes tipos de influência dos colegas.

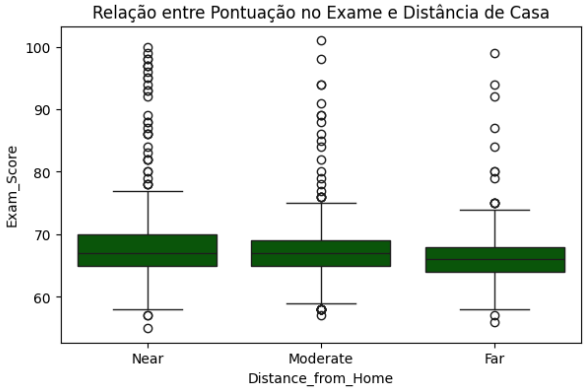

Relação entre Pontuação no Exame e Envolvimento dos Pais
Exam_Score
Low
Medium
High
Parental_Involvement


Relação entre Pontuação no Exame e Acesso a Recursos
Exam_Score
High
Medium
Low
Access_to_Resources


Relação entre Pontuação no Exame e Acesso à Internet
Exam_Score
Yes
No
Internet_Access


Relação entre Pontuação no Exame e Renda Familiar
Exam_Score
Low
Medium
High
Family_Income


Relação entre Pontuação no Exame e Grau de Instrução dos Pais
Exam_Score
High School
College
Postgraduate
Parental_Education_Level


Relação entre Pontuação no Exame e Distância de Casa
Exam_Score
Near
Moderate
Far
Distance_from_Home

```
                      0.25  0.50  0.75
Parental_Involvement
High                  66.0  68.0  70.0
Low                   64.0  66.0  69.0
Medium                65.0  67.0  69.0

                     0.25  0.50  0.75
Access_to_Resources
High                 66.0  68.0  70.0
Low                  64.0  66.0  68.0
Medium               65.0  67.0  69.0

                   0.25  0.50  0.75
Internet_Access
No                 64.0  66.0  69.0
Yes                65.0  67.0  70.0

                0.25  0.50  0.75
Family_Income
High            65.0  68.0  70.0
Low             65.0  67.0  69.0
Medium          65.0  67.0  70.0

                           0.25  0.50  0.75
Parental_Education_Level
College                    65.0  67.0  70.0
High School                64.0  67.0  69.0
Postgraduate               66.0  68.0  70.0

                     0.25  0.50  0.75
Distance_from_Home
Far                  64.0  66.0  68.0
Moderate             65.0  67.0  69.0
Near                 65.0  67.0  70.0
```

Fatores Socioeconômicos e Familiares que influenciam as notas

Envolvimento Parental

- Maior envolvimento dos pais está associado a notas mais altas em todos os quantis, especialmente no quantil 0,75.
- Alunos com baixo envolvimento parental tendem a ter as notas mais baixas, enquanto aqueles com alto envolvimento parental tendem a ter as notas mais altas.

Acesso a Recursos

- O acesso a recursos mostra um padrão semelhante ao envolvimento parental.
- Alunos com alto acesso a recursos têm notas mais altas em todos os quantis, especialmente no quantil 0,75.
- Alunos com baixo acesso a recursos têm as notas mais baixas.

Acesso à Internet

- Alunos com acesso à internet tendem a ter notas mais altas em todos os quantis, especialmente no quantil 0,75, em comparação com aqueles sem acesso.

Renda Familiar

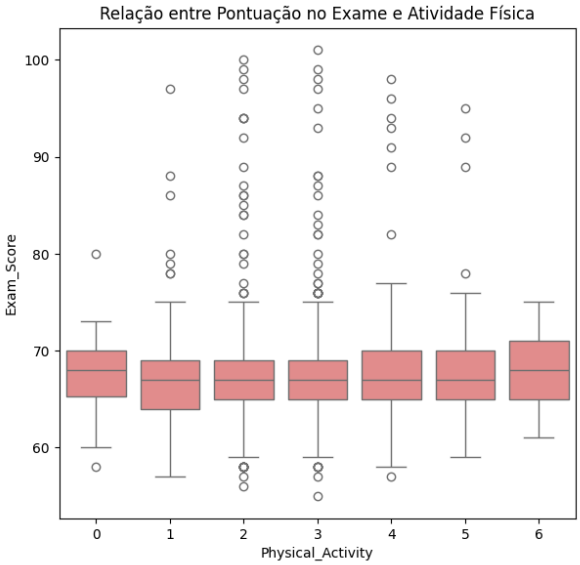
- Renda familiar alta está associada a notas mais altas no quantil 0,50 e 0,75.
- No entanto, a diferença entre os níveis de renda é menos pronunciada no quantil 0,25.

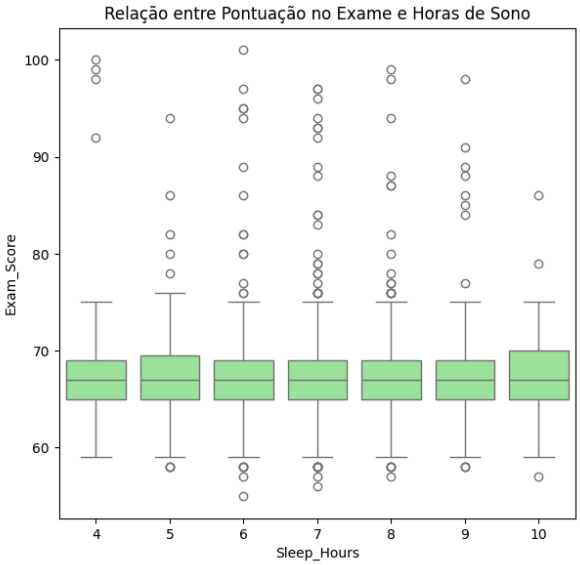
Nível de Educação Parental

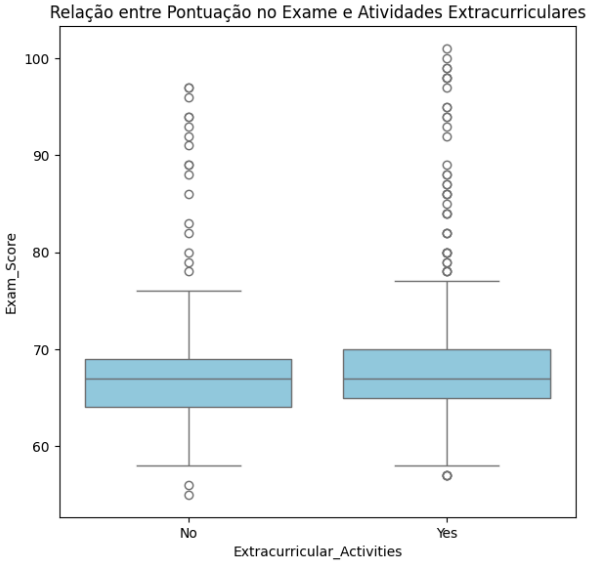
- O nível de educação parental "Pós-graduação" está associado a notas mais altas no quantil 0,25 e 0,75.
- Alunos cujos pais possuem apenas o ensino médio tendem a ter as notas mais baixas no quantil 0,25.

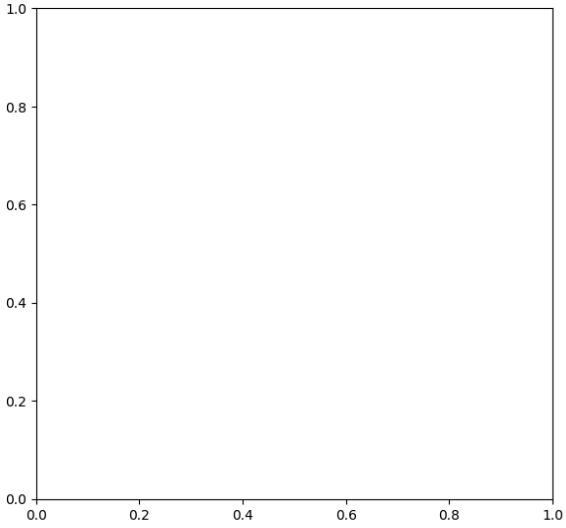
Distância de Casa

- Morar perto da escola está associado a notas mais altas no quantil 0,75.
- Alunos que moram longe da escola tendem a ter as notas mais baixas em todos os quantis.


Relação entre Pontuação no Exame e Atividade Física
Exam_Score
Physical_Activity
0
1
2
3
4
5
6
60
70
80
90
100


Relação entre Pontuação no Exame e Horas de Sono
Exam_Score
Sleep_Hours
4
5
6
7
8
9
10
60
70
80
90
100


Relação entre Pontuação no Exame e Atividades Extracurriculares
Exam_Score
Extracurricular_Activities
No
Yes
60
70
80
90
100


0.0
0.2
0.4
0.6
0.8
1.0

```
                            0.25  0.50  0.75
Extracurricular_Activities
No                          64.0  67.0  69.0
Yes                         65.0  67.0  70.0

             0.25  0.50  0.75
Sleep_Hours
4            65.0  67.0  69.0
5            65.0  67.0  69.5
6            65.0  67.0  69.0
7            65.0  67.0  69.0
8            65.0  67.0  69.0
9            65.0  67.0  69.0
10           65.0  67.0  70.0

                     0.25  0.50  0.75
Physical_Activity
0                   65.25  68.0  70.0
1                   64.00  67.0  69.0
2                   65.00  67.0  69.0
3                   65.00  67.0  69.0
4                   65.00  67.0  70.0
5                   65.00  67.0  70.0
6                   65.00  68.0  71.0
```

Fatores de Estilo de Vida que influenciam as notas

Atividades Extracurriculares

- Participar de atividades extracurriculares tem um impacto positivo, ainda que pequeno, nas notas.
- Alunos que participam de atividades extracurriculares tendem a ter notas ligeiramente maiores no quantil 0,75 em comparação com aqueles que não participam.
- Essa diferença sugere que atividades extracurriculares podem contribuir para o desenvolvimento de habilidades e conhecimentos que auxiliam no desempenho acadêmico.

Horas de Sono

- A relação entre horas de sono e notas não é linear.
- Dormir 10 horas está associado a notas ligeiramente maiores no quantil 0,75.
- No entanto, não há muita diferença nas notas entre aqueles que dormem entre 4 e 9 horas.

- É importante notar que essa análise não considera a qualidade do sono, que também pode influenciar o desempenho acadêmico.

Atividade Física

- A atividade física parece ter um impacto positivo nas notas, especialmente para aqueles que praticam com maior frequência.
- Alunos que praticam 6 horas de atividade física por semana tendem a ter as maiores notas no quantil 0,75.
- No entanto, o aumento nas notas não é linear, com o maior salto ocorrendo entre 0 e 6 horas de atividade física semanal.

2. Dados

2.1. Valores nulos

2.2. Variáveis numéricas

2.3. Variáveis categóricas

```
Unique values for Parental_Involvement: ['Low' 'Medium' 'High']
Unique values for Access_to_Resources: ['High' 'Medium' 'Low']
Unique values for Extracurricular_Activities: ['No' 'Yes']
Unique values for Motivation_Level: ['Low' 'Medium' 'High']
Unique values for Internet_Access: ['Yes' 'No']
Unique values for Family_Income: ['Low' 'Medium' 'High']
Unique values for Teacher_Quality: ['Medium' 'High' 'Low']
Unique values for School_Type: ['Public' 'Private']
Unique values for Peer_Influence: ['Positive' 'Negative' 'Neutral']
Unique values for Learning_Disabilities: ['No' 'Yes']
Unique values for Parental_Education_Level: ['High School' 'College' 'Postgr
aduate']
Unique values for Distance_from_Home: ['Near' 'Moderate' 'Far']
Unique values for Gender: ['Male' 'Female']
```

2.5. Treino/Teste

Separando a base de dados em treino e teste utilizando uma proporção de 2/3 para treino e 1/3 para testes.

Out[ ]:

	Hours_Studied	Attendance	Sleep_Hours	Previous_Scores	Tutoring_Ses
6227	22	75	6	54	
1729	31	90	7	61	
1123	13	88	9	61	
2200	31	91	6	82	
205	14	82	8	65	

5 rows × 22 columns

Out[ ]: (4252, 22)

Out[ ]:

	Exam_Score
6227	68
1729	72
1123	66
2200	74
205	65

dtype: int64

Out[ ]:

	Exam_Score
3293	65
2322	71
6276	64
490	66
1459	68

dtype: int64

3. Modelagem

3.1. Treino

```
const                                   34.374491
Hours_Studied                            0.291542
Attendance                               0.198376
Sleep_Hours                              0.000491
Previous_Scores                          0.049389
Tutoring_Sessions                        0.491943
Physical_Activity                        0.185544
Parental_Involvement_num                 0.967175
Access_to_Resources_num                  1.021033
Motivation_Level_num                     0.583417
Family_Income_num                        0.543847
Teacher_Quality_num                      0.556857
Extracurricular_Activities_Yes           0.589570
Internet_Access_Yes                      0.966224
School_Type_Public                       0.020294
Peer_Influence_Neutral                   0.485958
Peer_Influence_Positive                  1.062080
Learning_Disabilities_Yes               -0.789120
Parental_Education_Level_High School    -0.472761
Parental_Education_Level_Postgraduate    0.527055
Distance_from_Home_Moderate              0.476787
Distance_from_Home_Near                  1.053270
Gender_Male                             -0.044334
dtype: float64
```

```
const                                   0.000000e+00
Hours_Studied                           0.000000e+00
Attendance                              0.000000e+00
Sleep_Hours                             9.834942e-01
Previous_Scores                         2.517635e-89
Tutoring_Sessions                       8.246374e-68
Physical_Activity                       3.612277e-08
Parental_Involvement_num                4.945460e-81
Access_to_Resources_num                 1.262183e-89
Motivation_Level_num                    1.943706e-31
Family_Income_num                       6.884203e-31
Teacher_Quality_num                     6.071359e-22
Extracurricular_Activities_Yes          8.949158e-17
Internet_Access_Yes                     4.519353e-13
School_Type_Public                      7.878558e-01
Peer_Influence_Neutral                  2.167424e-07
Peer_Influence_Positive                 1.631322e-29
Learning_Disabilities_Yes               3.941716e-12
Parental_Education_Level_High School    3.588010e-09
Parental_Education_Level_Postgraduate   1.061842e-07
Distance_from_Home_Moderate             1.373322e-04
Distance_from_Home_Near                 3.479786e-19
Gender_Male                             5.267077e-01
dtype: float64
```

Fatores que Influenciam as Notas dos Exames

Fatores Acadêmicos

Esta categoria tem um alto impacto nas notas dos exames, com 90% das variáveis sendo estatisticamente significativas.

- Variáveis como **Horas_Studied**, **Attendance**, **Previous_Scores**, **Motivation_Level_num**, **Tutoring_Sessions**, **Peer_Influence**, **Learning_Disabilities_Yes** e **Teacher_Quality_num** mostraram-se significativas.
- **School_Type_Public** não foi significativo, sugerindo que o tipo de escola (pública ou privada) pode não ter um efeito direto nas notas.

Fatores Socioeconômicos e Familiares

Todas as variáveis nesta categoria são altamente significativas.

- **Family_Income_num**, **Parental_Education_Level**, **Access_to_Resources_num**, **Parental_Involvement_num**, **Internet_Access_Yes** e **Distance_from_Home** demonstraram influenciar as notas dos exames.
- Isso sugere que o contexto socioeconômico e familiar do aluno desempenha um papel crucial no desempenho acadêmico.

Estilo de Vida

- **Extracurricular_Activities_Yes** e **Physical_Activity** são significativas, indicando que a participação em atividades extracurriculares e a prática de atividades físicas podem ter um efeito positivo nas notas.
- **Sleep_Hours** não foi significativo neste estudo.

Dados Demográficos

- **Gender_Male** não foi significativo, sugerindo que o gênero pode não ter um impacto direto nas notas dos exames.

Resultados da Regressão: Impacto das Variáveis nas Notas dos Exames

Variáveis Numéricas

- **Hours_Studied:** Um aumento de 1 hora de estudo está associado a um aumento de 0.291542 pontos na nota do exame.
- **Attendance:** Um aumento de 1 ponto percentual na frequência está associado a um aumento de 0.198376 pontos na nota do exame.
- **Previous_Scores:** Um aumento de 1 ponto nas notas anteriores está associado a um aumento de 0.049389 pontos na nota do exame.
- **Access_to_Resources_num:** Um aumento de 1 unidade na escala de acesso a recursos está associado a um aumento de 1.02103 pontos na nota do exame.
- **Parental_Involvement_num:** Um aumento de 1 unidade na escala de envolvimento dos pais está associado a um aumento de 0.967175 pontos na nota do exame.
- **Motivation_Level_num:** Um aumento de 1 unidade na escala de nível de motivação está associado a um aumento de 0.583417 pontos na nota do exame.
- **Teacher_Quality_num:** Um aumento de 1 unidade na escala de qualidade do professor está associado a um aumento de 0.556857 pontos na nota do exame.
- **Family_Income_num:** Um aumento de 1 unidade na escala de renda familiar está associado a um aumento de 0.543847 pontos na nota do exame.

Variáveis Categóricas

- **Extracurricular_Activities_Yes:** Participar de atividades extracurriculares está associado a um aumento de 0.58957 pontos na nota do exame, em comparação a não participar.
- **Internet_Access_Yes:** Ter acesso à internet está associado a um aumento de 0.966224 pontos na nota do exame, em comparação a não ter acesso.
- **Peer_Influence_Neutral:** Ter influência neutra dos colegas está associado a um aumento de 0.485958 pontos na nota do exame, em comparação a ter influência negativa (que é a categoria omitida).
- **Peer_Influence_Positive:** Ter influência positiva dos colegas está associado a um aumento de 1.06208 pontos na nota do exame, em comparação a ter influência negativa (que é a categoria omitida).
- **Learning_Disabilities_Yes:** Ter dificuldades de aprendizado está associado a uma diminuição de 0.78912 pontos na nota do exame, em comparação a não ter dificuldades.
- **Parental_Education_Level:**
  - Ter pais com nível de educação "High School" está associado a uma diminuição de 0.472761 pontos na nota do exame, em comparação a ter pais com nível de educação "College" (que é a categoria omitida).

    - Ter pais com nível de educação "Postgraduate" está associado a um aumento de 0.527055 pontos na nota do exame, em comparação a ter pais com nível de educação "College" (que é a categoria omitida).
- **Distance_from_Home:**
    - Morar a uma distância "Moderate" da escola está associado a um aumento de 0.476787 pontos na nota do exame, em comparação a morar "Far" (que é a categoria omitida).
    - Morar "Near" da escola está associado a um aumento de 1.05327 pontos na nota do exame, em comparação a morar "Far" (que é a categoria omitida).

3.2. Avaliação

```
1.6554851886259283
```

O RMSE é uma medida da precisão do modelo de regressão. Ele indica o desvio médio entre os valores previstos pelo modelo e os valores reais. Quanto menor o RMSE, melhor o modelo se ajusta aos dados. Um RMSE de 1.655 sugere que, em média, as previsões do modelo estão desviando 1.655 unidades do valor real da variável alvo.