unesp

Relatório de Estatística Aplicada

Nome: Maria Alice Gonçalves, Eduardo Conde

Curso/Turma: BSI 022

Data de Entrega: 25/06/2025

1. Introdução e Objetivo do Trabalho

O objetivo do presente trabalho é apresentar todo o processo de interpretação de dados, aplicando conhecimentos de análise exploratória e técnicas preditivas, além da descrição das etapas de limpeza e tratamento de dados.

2. Descrição do Conjunto de Dados Utilizado

O conjunto de dados utilizados foi adquirido pela plataforma *Kaggle*, que disponibiliza centenas de milhares de datasets para serem utilizados por pesquisadores.

Os dados se referem ao desempenho acadêmicos de estudantes dos Estados Unidos, bem como informações de perfil pessoal, como gênero, etnia, formação dos pais e tipo de almoço recebido (padrão ou com preço reduzido), que indica o padrão socioeconômico do estudante.

As colunas com dados disponibilizadas são:

- *gender*: gênero (variável qualitativa nominal);
- *race/ethnicity*: raça/etnia (variável qualitativa nominal);
- *parental level of education*: nível da educação parental (variável qualitativa ordinal);
- *lunch*: almoço (variável qualitativa nominal);
- *test preparation course*: curso de preparação para provas (variável qualitativa nominal);
- *math score*: nota matemática (variável quantitativa contínua);
- *reading score*: nota leitura (variável quantitativa contínua);
- *writing score*: nota escrita (variável quantitativa contínua).

3. Etapas de Limpeza e Tratamento dos Dados

O tratamento de dados é uma etapa fundamental para a análise estatística, pois garante a qualidade, consistência e integridade das informações utilizadas. Dados brutos frequentemente contêm erros, valores ausentes, duplicidades ou inconsistências que, se não forem devidamente tratados, podem comprometer a validade dos resultados obtidos. Ademais,

o tratamento adequado permite organizar e padronizar os dados, facilitando a aplicação de técnicas estatísticas e a interpretação dos resultados. Assim, essa etapa é essencial para que as conclusões da análise sejam confiáveis e representem com precisão a realidade estudada.

Nesta análise, a etapa de tratamento e limpeza dos dados não foi realizada, pois foi utilizada uma base de dados previamente estruturada e tratada. A fonte dos dados já apresentava informações organizadas, sem valores ausentes ou inconsistências aparentes, o que dispensou intervenções adicionais nesta etapa. Dessa forma, foi possível iniciar diretamente a fase de análise exploratória e estatística, com a garantia de que os dados estavam prontos para uso.

4. Análise exploratória (estatística descritiva e gráficos)

A partir dos dados disponibilizados, é possível, inicialmente, obtermos informações como média, desvio padrão, valores mínimos e máximos por meio da função *describe*. A função gera estatísticas descritivas resumidas de colunas numéricas ou categóricas de um DataFrame.

Estatísticas Descritivas Gerais:

	matemática	leitura	escrita
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

Ademais, é possível fazer análises mais específicas, como a comparação das notas de cada disciplina entre os alunos de sexos diferentes, feita de forma simples utilizando a função *mean*, após aplicado um *groupby* no *dataframe*.

Médias das Notas por Gênero:

gênero	matemática	leitura	escrita
feminino	63.633205	72.608108	72.467181
masculino	68.728216	65.473029	63.311203

Utilizando a mesma lógica, podemos utilizar o campo que indica se o aluno participou de algum curso preparatório. Assim podemos comparar as notas e verificar os benefícios que a formação pode trazer ao estudante:

Médias das Notas por Curso Preparatório:

curso de preparação para as provas	matemática	leitura	escrita
realizado	69.695531	73.893855	74.418994
não	64.077882	66.534268	64.504673

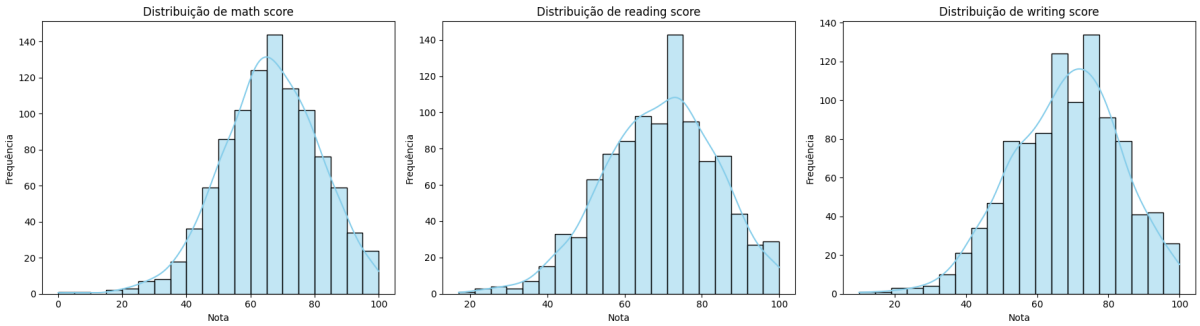

Figura 1: Distribuição das notas

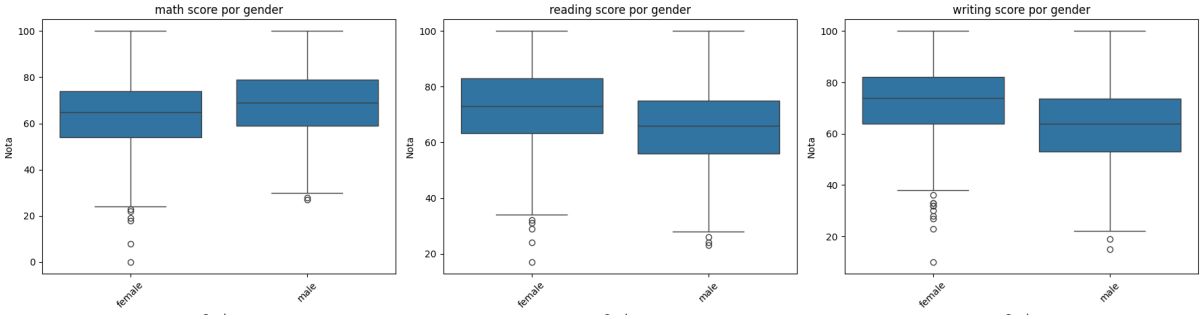

Figura 2: BoxPlot de notas por gênero

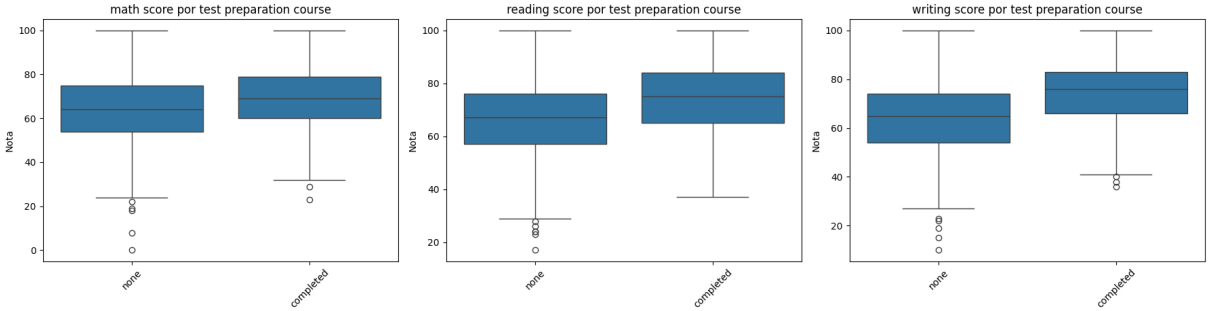

Figura 3: BoxPlot de curso de preparação para provas

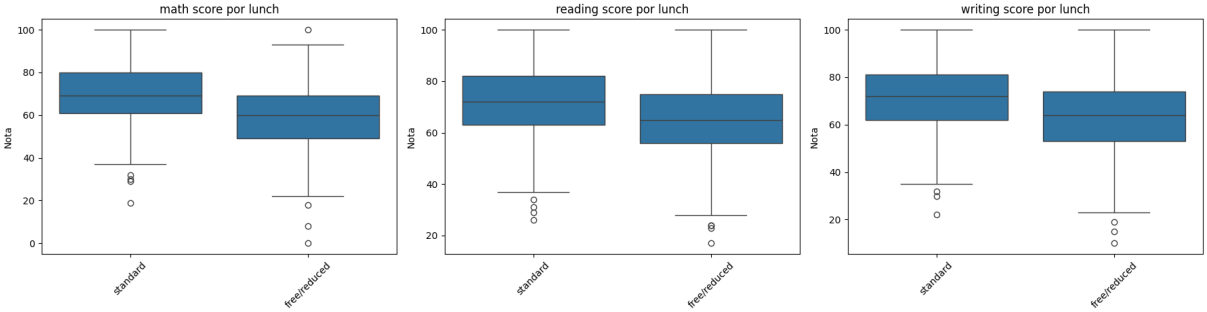

Figura 4: BoxPlot de almoço

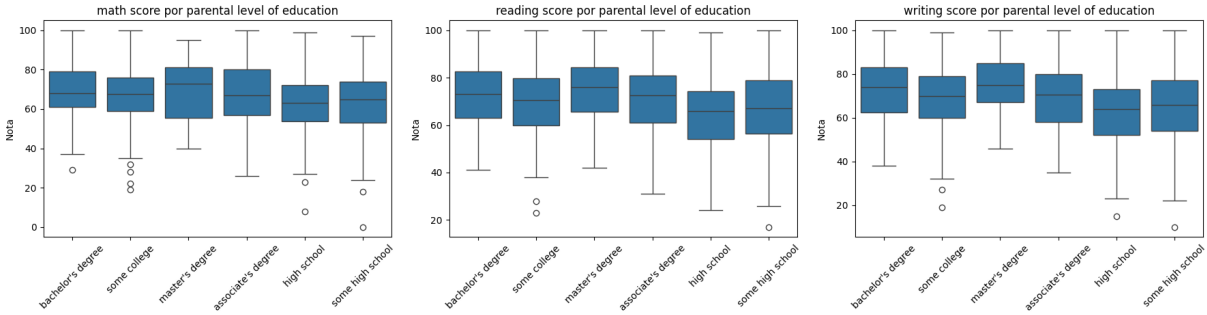

Figura 5: BoxPlot do nível de educação

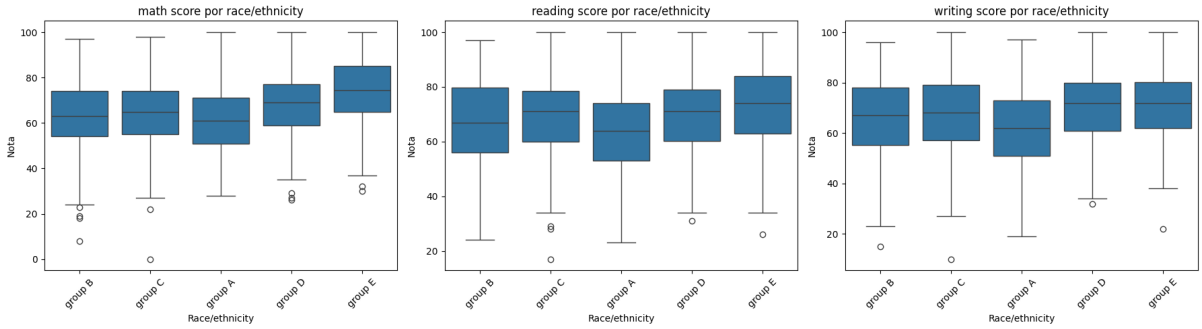

Figura 6: BoxPlot de raça/etnia

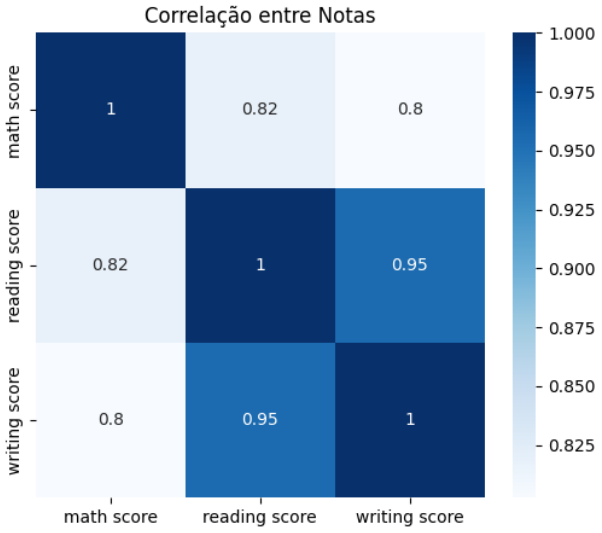

Figura 7: Correlação entre as notas

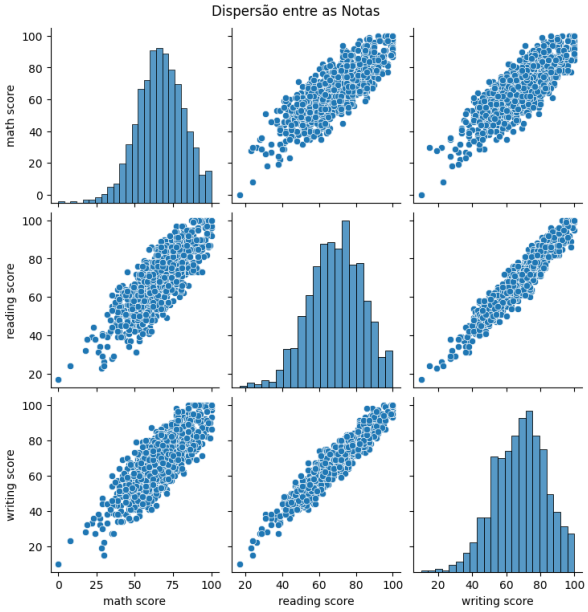

Figura 8: Dispersão entre as notas

5. Aplicação da técnica estatística ou preditiva

0.1 Figura 1

Para a geração dos gráficos histograma com a distribuição das notas, primeiro percorre-se a lista das três avaliações existentes com o *enumerate*. Após isso, é necessário indicar o *dataframe* para ser utilizado, bem como a distância indicada na legenda (parâmetro *bins*, indicado como 20 no exemplo)

```
plt.figure(figsize=(18, 5))
for i, subject in enumerate(['math score', 'reading score', 'writing score']):
    plt.subplot(1, 3, i+1)
    sns.histplot(df[subject], kde=True, bins=20, color='skyblue')
    plt.title(f'Distribuição de {subject}')
    plt.xlabel('Nota')
    plt.ylabel('Frequência')
plt.tight_layout()
plt.show()
```

0.2 Figuras 2, 3 e 4, 5 e 6

Para facilitar o entendimento por parte daquele que lê o código, foi desenvolvida a função `plot_boxplots_by_category`, que executa o *plot* do gráfico com base na categoria informada como parâmetro. Os dados do *dataframe* são informados por meio do comando `sns.boxplot(data=df, x=category, y=subject)`.

```
def plot_boxplots_by_category(category):
    plt.figure(figsize=(18, 5))
    for i, subject in enumerate(['math score', 'reading score', 'writing score']):
        plt.subplot(1, 3, i+1)
```

```
        sns.boxplot(data=df, x=category, y=subject)
        plt.title(f'{subject} por {category}')
        plt.xlabel(category.capitalize())
        plt.ylabel('Nota')
        plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()

# Comparações
plot_boxplots_by_category('gender')
plot_boxplots_by_category('test preparation course')
plot_boxplots_by_category('lunch')
plot_boxplots_by_category('parental level of education')
plot_boxplots_by_category('race/ethnicity')
```

0.3 Figuras 7 e 8

A Figura 7 apresenta a correlação entre as notas. Esse tipo de gráfico é possíve graças à função *heatmap*, da biblioteca *seaborn*, em que são especificados os dados do *dataframe* e alguns parâmetros para estilização.

```
plt.figure(figsize=(6, 5))
sns.heatmap(df[['math score', 'reading score', 'writing score']].corr(), annot=True, c
plt.title('Correlação entre Notas')
plt.show()
```

Já a figura 8, que apresenta a dispersão entre as notas, é gerada por meio da função *pairplot*.

```
sns.pairplot(df[['math score', 'reading score', 'writing score']])
plt.suptitle('Dispersão entre as Notas', y=1.02)
plt.show()
```

6. Discussão dos resultados obtidos

Em relação à Figura 1, as médias das notas em matemática, leitura e escrita giram em torno de 66 a 69 pontos, com desvios padrão em torno de 15. Isso indica uma distribuição relativamente normal, mas com alta dispersão, ou seja, existem tanto alunos com desempenho excelente quanto alunos com grandes dificuldades.

Já em relação à Figura 2, meninas apresentaram notas superiores em leitura e escrita. Meninos tiveram ligeira vantagem em matemática, embora com maior variação nas notas. Essa diferença sugere que pode haver influência de fatores culturais, expectativas sociais e métodos de ensino nas áreas de linguagem e exatas.

A Figura 3 nos indica que alunos que completaram o curso preparatório tiveram médias mais altas em todas as disciplinas, sugerindo que intervenções educacionais específicas (como aulas extras ou revisão de conteúdo) podem ser eficazes para melhorar o desempenho.

Na Figura 4, alunos com almoço padrão (standard) — geralmente de famílias com melhor condição socioeconômica — tiveram desempenho significativamente melhor. Já aqueles que recebem almoço gratuito ou reduzido (free/reduced) apresentaram médias mais baixas, o que sugere uma relação entre vulnerabilidade econômica e dificuldade escolar.

Em relação à Figura 5, há uma correlação clara entre o nível educacional dos pais e o desempenho dos filhos: alunos cujos pais têm nível superior ou mestrado obtêm as melhores médias. Aqueles com pais que não completaram o ensino médio têm, em média, desempenho inferior.

Nas figuras 7 e 8, as notas de matemática, leitura e escrita apresentam alta correlação entre si, o que significa que um aluno com bom desempenho em uma disciplina tende a ir bem nas outras também. Isso pode refletir habilidades cognitivas gerais, hábitos de estudo e motivação.

Ao compararmos, em especial, as notas nas avaliações de leitura e escrita, vemos que a correlação nas notas é ainda mais forte para disciplinas e áreas do conhecimento que têm mais entre si.

7. Considerações finais

Em síntese, os dados analisados revelam padrões importantes sobre o desempenho escolar dos alunos, com destaque para a influência de fatores como gênero, participação em programas educacionais, condição socioeconômica e escolaridade dos pais. Além disso, a forte correlação entre as notas em disciplinas específicas sugere que o desempenho acadêmico está relacionado a uma gama específica de habilidades que se relacionam a proporcionar vantagens em assuntos específicos.