Para o relatório de AED do Projeto Integrador 1, utilize a estrutura de itens abaixo quando aplicável.

1. Título e Informações Preliminares

- **Título:** Análise Exploratória de Dados - Performance Física de Alunos (PROESP)
- **Autor(es) e Data:** Pedro dos Santos Garcia, Matheus Serra Lourenço, Danilo Pereira Peixoto, Vinícius Machado de Assunção - 03 de Novembro de 2025

2. Introdução e Objetivos

- **Contexto:** Esta análise investiga o perfil e a performance física de uma amostra de 100 alunos, com base em 10 variáveis de desempenho físico e antropométricas.
- **Objetivos:** O foco principal é compreender a estrutura dos dados, analisar a distribuição demográfica da amostra (idade, gênero) e investigar as correlações existentes entre as diferentes métricas de performance.
- **Fonte dos Dados:** Base_Alunos_Imaginarios_PROESP (1).xlsx.

3. Descrição dos Dados (Visão Geral)

- **Estrutura do Conjunto de Dados:**
    - **Número de observações (linhas):** 100
    - **Número de variáveis (colunas):** 12
    - **Visão das primeiras linhas dos dados:**

ID_Aluno	Idade	Gênero	IMC	RCE	Flexibilidade	Potência_MMII	Potência_MS	Velocidade	Agilidade	Abdominal	Resistência
1	17	Masculino	28,2	0,46	26,6	134,2	30	5,51	13,18	36	1380
2	15	Masculino	22,1	0,43	23,5	157,1	28,2	5,59	13,32	37	1805
3	16	Feminino	20,4	0,39	31,8	106,2	23,3	4,53	12,51	29	1523
4	16	Feminino	19,2	0,45	36,8	81,8	25,4	6	12,63	35	1207
5	15	Masculino	24,3	0,41	29,2	148,5	25,7	5,49	12,82	42	1767

    - **Visão das últimas linhas dos dados:**

96	15	Masculino	23,2	0,42	32,2	155,6	27,1	4,83	12,84	34	1679
97	16	Feminino	22,1	0,39	32,2	116,5	27,5	5,51	12,75	21	1058
98	15	Masculino	17,7	0,5	26,9	160,3	31,1	5,24	12,4	25	1750
99	16	Feminino	17,6	0,46	24,2	111,1	19,6	5,57	13,78	38	1459
100	17	Feminino	18	0,47	29,3	102	32,1	5,69	13,03	24	1479

    - **Tipo de dados de cada variável (numérica, categórica, data, etc.):** Os dados estão estruturados de forma coesa. A variável `Gênero` é do tipo `object` (texto), enquanto as demais são numéricas (`int64` ou `float64`), adequadas para análise quantitativa.
- **Descrição das Variáveis:**
    - **ID_Aluno:** Identificador único do aluno.
    - **Idade:** Idade do aluno (em anos).
    - **Gênero:** Gênero do aluno (Masculino, Feminino).
    - **IMC:** Índice de Massa Corporal.
    - **RCE:** Relação Cintura-Estatura.
    - **Flexibilidade:** Resultado do teste de flexibilidade (cm).
    - **Potência_MMII:** Potência de Membros Inferiores (salto horizontal, cm).
    - **Potência_MS:** Potência de Membros Superiores (arremesso de medicine ball, m).
    - **Velocidade:** Resultado do teste de velocidade (corrida, segundos).
    - **Agilidade:** Resultado do teste de agilidade (shuttle run, segundos).
    - **Abdominal:** Resultado do teste de resistência abdominal (contagem).
    - **Resistência:** Resultado do teste de resistência cardiorrespiratória (metros).

4. Qualidade dos Dados e Pré-processamento

Esta seção detalha os problemas encontrados e como eles foram tratados.

- **Valores Ausentes (Missing Data):** Não foram encontrados valores ausentes em nenhuma coluna. O conjunto de dados está completo.
- **Valores Duplicados:** Não foram encontradas linhas duplicadas (Número de linhas duplicadas: 0).
- **Inconsistências e Erros:** Os tipos de dados são consistentes. `altura`, `peso` e `imc` são `float64`, enquanto `indice` e `indiceIMC` são `int64`, o que é apropriado.
- **Outliers (Valores Atípicos):**A análise das estatísticas descritivas (abaixo) e dos boxplots gerados na análise anterior (`boxplots_numericos_univariados.png`) indica que os dados não possuem outliers extremos que justifiquem tratamento.Os valores médios (ex: IMC 21.67, Idade 16.06) e os intervalos (mín/máx) são plausíveis para a amostra de alunos adolescentes.

5. Análise Exploratória

É a parte principal do relatório, onde os *insights* são apresentados, geralmente com o uso intenso de visualizações (gráficos) e estatísticas, mas pode ser textual.

5.1. Análise Univariada (Variáveis Individuais)

- **Variáveis Categóricas (Gênero e Idade):**
  - **Gênero:** A distribuição é equilibrada (Gráfico: `distribuicao_genero.png`).
    - Feminino: 51 (51.0%)
    - Masculino: 49 (49.0%)
  - **Idade:** A amostra é composta por alunos de 15, 16 e 17 anos, com leve predominância de 16 anos (Gráfico: `distribuicao_idade.png`).
    - 15 anos: 29 (29.0%)
    - 16 anos: 36 (36.0%)
    - 17 anos: 35 (35.0%)
- **Variáveis Numéricas:**
  - As estatísticas descritivas encontram-se na Seção 4.
  - Os histogramas (`histogramas_numericos.png`) indicam que a maioria das variáveis de performance (ex: IMC, Flexibilidade, Potências) segue uma distribuição aproximadamente normal, sem assimetrias pronunciadas.

5.2. Análise Bivariada e Multivariada (Relações entre Variáveis)

- **Relações Numérica vs. Numérica:**
  - O heatmap de correlação (`heatmap_correlacao.png`) destaca as seguintes relações:
  - **Correlação Positiva Moderada (r=0.40):** Entre `Potência_MMII` e `Potência_MS`. Indica que alunos com maior potência de membros inferiores tendem a ter, também, maior potência de membros superiores.
  - **Correlação Negativa Fraca (r=-0.29):** Entre `Flexibilidade` e `Potência_MS`.
  - **Correlações Fracas:** A variável `Idade` não demonstrou correlações fortes com as métricas de performance, sugerindo pouca variação de desempenho entre 15 e 17 anos nesta amostra.

- **Relações Categórica vs. Numérica:**
  - Esta análise revela as distinções mais significativas (Gráficos: `boxplots_metricas_por_genero.png`):
  - **Masculino:** Apresenta médias superiores em testes de potência, velocidade e resistência: `Potência_MMII` (148.9 vs 114.7), `Potência_MS` (29.6 vs 22.0), `Resistência` (1607m vs 1420m) e `Abdominal` (36.9 vs 34.1). Também registram tempos médios menores (melhor performance) em `Velocidade` (5.14s vs 5.27s) e `Agilidade` (12.92s vs 13.41s).
  - **Feminino:** Apresenta média superior significativa no teste de `Flexibilidade` (27.6 cm vs 23.2 cm).

    - `IMC` e `RCE` médios foram ligeiramente superiores no grupo masculino.

- **Relações Categórica vs. Numérica (Performance por Idade):**
    - As médias de performance entre as três faixas etárias (15, 16 e 17 anos) são notavelmente similares, indicando pouca variação no desempenho relacionada à maturação nesse intervalo de idade.
- **Relações Categórica vs. Categórica (Gênero vs Idade):**
    - **A tabela de contingência (Gráfico: `barras_idade_genero.png`) ilustra a distribuição de gênero dentro de cada faixa etária.**

6. Descobertas e *Insights* Principais

- **Resumo das Descobertas:**
    - **Qualidade dos Dados:** O conjunto de dados é de alta qualidade, caracterizando-se por estar completo (sem valores nulos) e consistente (sem duplicatas).
    - **Amostra Equilibrada:** A amostra de 100 alunos está demograficamente bem distribuída em termos de Gênero (51% F, 49% M) e Idade (15-17 anos).
    - **Gênero como Fator Diferenciador:** O Gênero emergiu como o principal fator de diferenciação na performance física. Alunos do sexo masculino apresentaram, em média, melhores resultados em testes de potência (MMII, MS) e resistência (cardiorrespiratória, abdominal), enquanto alunas do sexo feminino apresentaram resultados superiores em flexibilidade.
    - **Associação de Força:** Identificou-se uma correlação positiva moderada (r=0.40) entre a potência de membros inferiores e superiores.
    - **Impacto da Idade:** Na faixa etária analisada (15-17 anos), a idade não se mostrou um fator de correlação forte com as métricas de performance.

7. Conclusão e Próximos Passos

Conclusão: A análise exploratória demonstrou que, para esta amostra de alunos, o gênero é o principal preditor de performance física, com distinções claras entre os grupos. Os dados são robustos, limpos e consistentes, permitindo análises futuras. A

ausência de variação significativa de performance por idade (15-17) é um achado relevante.

Recomendações para a Próxima Etapa:

- **Clusterização (Segmentação):** Sugere-se a aplicação de algoritmos de clusterização (ex: K-Means) para identificar perfis de alunos (ex: "Perfil Velocista/Potente", "Perfil Resistente", "Perfil Flexível"), independentemente do gênero.
- **Modelagem Preditiva:** Os dados estão prontos para modelagem. Recomenda-se a criação de um modelo de regressão para prever uma variável-alvo (ex: `Resistência`) com base nas demais métricas (ex: `IMC`, `Potência_MMII`, `Agilidade`).
- **Padronização (Z-Score):** Para criar um "Escore de Performance Geral" que combine múltiplas métricas, recomenda-se a padronização das variáveis (cálculo de Z-Score) para normalizar as diferentes escalas e unidades de medição.
- (27.04) da amostra encontra-se acima do limite de peso normal (25.0). Os dados são robustos e limpos, permitindo conclusões claras sobre a distribuição do peso.
- **Recomendações para a Próxima Etapa:**
    - **Modelagem Preditiva:** Os dados estão limpos e prontos para modelagem. Pode-se tentar prever o `indiceIMC` (nível de obesidade) com base na `altura` e `peso`.
    - **Enriquecimento de Dados:** O conjunto de dados atual é limitado (altura, peso). Para insights mais profundos sobre políticas de saúde, sugere-se coletar ou cruzar esses dados com variáveis demográficas (como Gênero, Idade, Região do Brasil) e de estilo de vida (Ex: prática de atividade física, tabagismo) que também estão presentes na pesquisa Vigitel original.