

Aluno(a)			RA:
Curso	Ciência da Computação	Ano:	
Disciplina	Ciência de Dados		
Professor	Prof. Eduardo Pena		
Nota			
EDA e Visualização			
<p>Orientações gerais:</p> <p>1 - Todos os materiais (códigos fontes, documentos, diagramas, etc) deverão ser entregues em um arquivo .zip identificado (Nome/RA/Ano).</p> <p>2 - A interpretação das questões é parte do processo de avaliação.</p>			

Laboratório Prático: EDA e Viz do Campeonato Brasileiro

Dataset: futebol.csv

Entrega: Notebook + PDF (2 slides com imagens e resultados de teste estatístico) + Apresentação (Em sala)

Individual ou duplas

Contexto

Você é um(a) analista de dados de uma consultoria esportiva. Analise os dados do Campeonato Brasileiro e outros (8.453 jogos, 2003-2024) para extrair insights que orientem decisões sobre clubes e jogos.

1 Tarefas

1.1 Tarefa 1: Estatísticas Descritivas e Visualizações Básicas

Desenvolva o notebook mostrando:

1. Distribuição de gols:

- Crie coluna `gols_total = gols_mandante + gols_visitante`
- Histograma da distribuição de gols totais
- Calcule: média, mediana, moda, desvio padrão, IQR

2. Mando de campo:

- Boxplots: gols mandante vs visitante

- Identifique e interprete outliers
- Compare medidas de tendência central

3. Público:

- Gráfico de barras: 10 times que mais jogaram em casa
- Estatísticas descritivas do público (dados não-nulos)

Responda:

- A distribuição de gols é simétrica ou assimétrica? O que isso indica?
- Há vantagem numérica do mando de campo?
- Qual a relação entre público e gols?

Entrega: 3 visualizações + tabela de estatísticas.

1.2 Tarefa 2: Teste de Hipótese

Contexto: Quantificar estatisticamente se a “vantagem do mando de campo” existe.

Hipótese: “Times mandantes marcam significativamente mais gols que visitantes”

Desenvolva o notebook mostrando:

1. As hipóteses:

- H_0 : $\mu_{mandante} = \mu_{visitante}$ (sem diferença)
- H_1 : $\mu_{mandante} > \mu_{visitante}$ (mandantes marcam mais)

2. Execute teste t:

- Use `scipy.stats.ttest_ind()`
- Nível de significância: $\alpha = 0,05$
- Calcule e interprete valor-p

3. Analise resultados:

- Tamanho do efeito (diferença das médias)
- Significância estatística vs relevância prática
- Visualização ilustrando o resultado

Responda:

- O valor-p permite rejeitar H_0 ?
- Qual a diferença prática em gols?
- O resultado é significativo E relevante?

Entrega: Código do teste + interpretação + visualização.

1.3 Tarefa 3: Investigação Livre + Pitch

Contexto: Você é o lead analyst! Formule sua hipótese e apresente descobertas para a diretoria.

Faça:

1. Escolha sua hipótese:

- Explore o dataset e formule pergunta interessante
- Use qualquer variável ou combinação
- Mantenha relevância para o negócio

2. Investigue:

- Pelo menos 2 visualizações diferentes das Tarefas 1-2
- **1 gráfico caprichado** aplicando princípios de design visual
- 1 teste estatístico apropriado
- Interprete no contexto prático

3. Prepare pitch (2 min):

- Slide 1: Hipótese + gráfico caprichado + **justifique o princípio de viz usado**
- Slide 2: Teste + resultados + conclusão

Entrega: Código + 2 slides PDF (com seu nome) + pitch 2 min + **justificativa do princípio de visualização**.

Princípios de visualização para aplicar:

- | | |
|---------------------------------------|--|
| • Clareza e simplicidade | • Escolha do tipo de gráfico adequado |
| • Hierarquia visual (cores, tamanhos) | • Integridade gráfica (escalas honestas) |
| • Paleta de cores apropriada | • Acessibilidade (contraste, legibilidade) |

Entrega e Avaliação

Formato:

- Notebook .ipynb com 3 tarefas organizadas
- PDF com 2 slides (Tarefa 3) contendo seu nome

Bibliotecas recomendadas:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

Dicas:

- Sempre rotule eixos e adicione títulos
- Significância \neq relevância prática
- Nome obrigatório nos slides da Tarefa 3
- **Tarefa 3: Explique na apresentação QUAL princípio de viz você aplicou e POR QUÊ**
- Considere: paleta colorblind-friendly, contraste, hierarquia visual

Apresentações: Máximo de 2 min cada, ordem por sorteio, professor compila PDFs em um só. No caso de duplas, cada aluno(a) fala por um minuto.