relatorio

October 31, 2025

1 RELATÓRIO ENTREGA 2

Entregue à disciplina Introdução à Ciência de Dados - DCC UFMG - Segundo semestre de 2025, Belo Horizonte

Integrantes:

Daniel Costa (2024006064)

Isaac Reyes (2025050342)

João Araldi (2024005475)

Pedro Luiz (2024006129)

1.1 I. Introdução e contexto:

O relatório detalha as etapas da análise exploratória de dados e a aplicação de Testes de Hipótese e Intervalos de Confiança sobre o dataset dos jogos da Premier League 2019-2025. O Objetivo central do projeto é utilizar a análise estatística para extrair informações sobre o desempenho das equipes, o fator casa e a relação das odds das casas de apostas com os resultados finais das partidas.

1.2 II. Descrição da Análisa Exploratória

Linha de pensamento durante o Processo: - A Análise começou com a verificação da consistência do dataset, que contém 2280 partidas e 84 colunas, mesclando as features originais e novas features para descrever. Foi identificado uma pequena quantidade de valores faltantes (10), o que mostra uma robustez na coleta dos dados do dataset. Esses valores estavam nas colunas de odds de gols (P>2.5 e P<2.5), que, conforme o plano, foram ignoradas por não apresentarem risco aos estudos.

- No trabalho, as odds foram convertidas em probabilidades normalizadas: As odds são a forma como as casas de apostas expressam a probabilidade de um evento ocorrer, mas invertida.A fórmula para converter uma odd (O) em uma probabilidade implícita (P) é:

$$\text{Probabilidade Implícita}(P) = \frac{1}{\text{Odd}(O)}$$

A soma das probabilidades implícitas de todos os resultados possíveis (vitória da casa, empate, vitória do visitante) em uma casa de apostas nunca será 100% (ou 1.0). Ela será sempre maior que 100%.

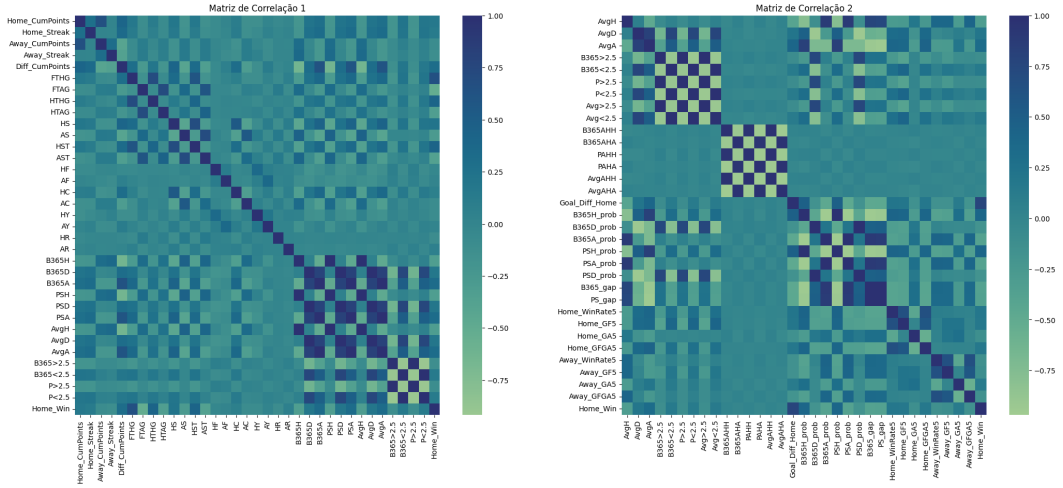
$$\text{Soma Total} = P_{\text{Casa}} + P_{\text{Empate}} + P_{\text{Fora}} > 1.0$$

Essa diferença acima de 1.0 é a margem de lucro da casa de apostas, ou overround. Para análises estatísticas e modelagem, é necessário que as probabilidades se somem a 1.0, simulando um mercado "justo" sem a margem da casa. A normalização é o processo de ajustar essas probabilidades implícitas para que a soma delas seja exatamente 1.0 (ou 100%).

$$\text{Probabilidade Normalizada} = \frac{P_{\text{Resultado}}}{\sum P_{\text{Total}}}$$

O resultado final são as colunas `B365H_prob_norm`, `B365D_prob_norm`, `B365A_prob_norm`, etc. Essas são as probabilidades que nosso modelo estatístico utilizará, representando a chance real de cada resultado ocorrer, excluindo a margem da casa de apostas.

- Foram criadas features cruciais para a análise, como `Goal_Diff_Home` (diferença de gols: mandante - visitante) e `Home_Win` (variável binária para vitória do mandante). Além disso, foram calculados os pontos acumulados (`Home_CumPoints`, `Away_CumPoints`, `Diff_CumPoints`) e o momentum dos times (`Home_Streak`, `Away_Streak`) com base nas últimas partidas, incluindo métricas de ataque/defesa (razão GF/GA).
- A imagem a seguir é uma Matriz de Correlação plotada em um heatmap entre as features numéricas do dataset, obtidas também durante a fase de análise exploratória.



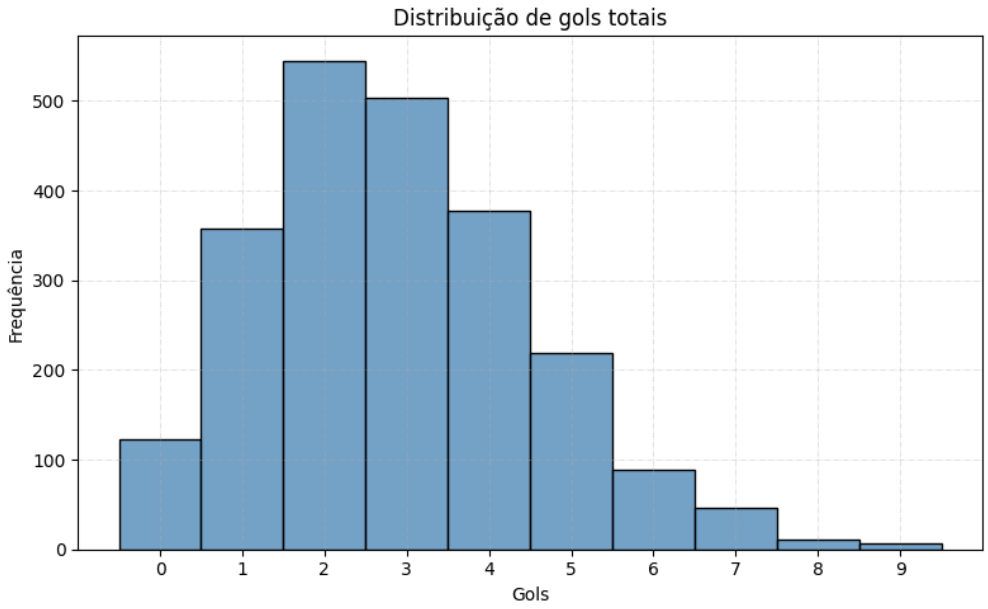
- O que é mais interessante notar é a correlação das variáveis novas criadas com `Home_Win`, indicando possíveis preditores para modelos de regressão.
  - `B365H_prob`: 40,3%
  - `PSH_prob`: 40,3%
  - `Diff_CumPoints`: 30,6%
  - `Home_WinRate5`: 17,9%
  - `Home_GFGA5`: 16,3%
  - `Home_CumPoints`: 16,2%
  - `Home_Streak`: 13,6%

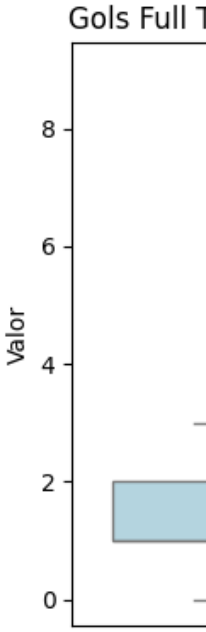
A etapa seguinte focou na visualização das distribuições de gols e fator casa

1.3 III Efeito das Variáveis Dependentes

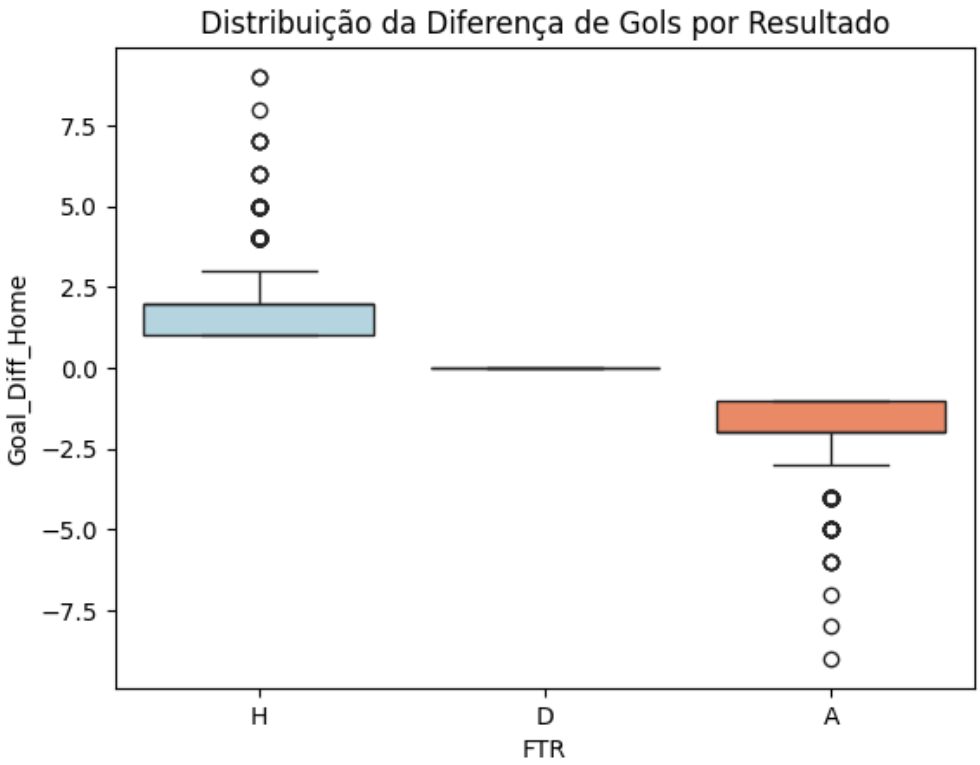
- Pergunta: Qual a distribuição e o impacto do Fator Casa em gols?
  - A mediana de gols para ambos os times é de aproximadamente 1 gol.
  - As distribuições são assimétricas, concentradas em valores baixos, com a maioria dos jogos terminando com 0 a 3 gols.

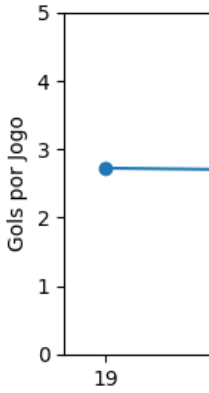
– O time mandante possui uma média de 1.53 gols, enquanto o visitante possui de 1.24 gols, sugerindo uma leve vantagem do mandante.



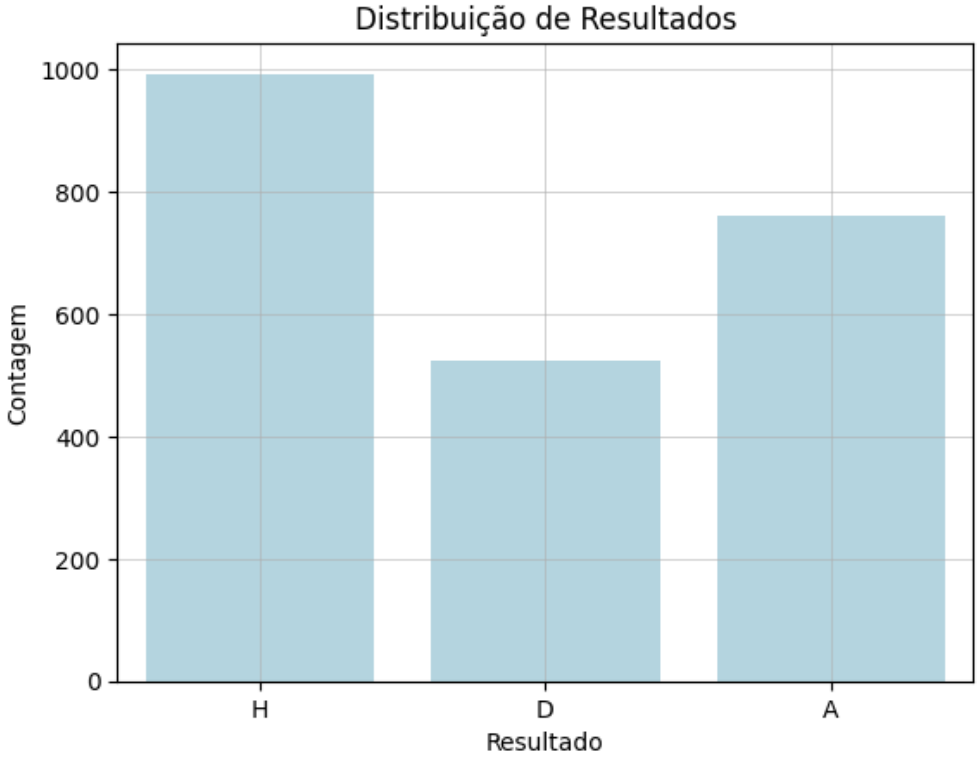


– Fator casa: A frequência de vitórias dos mandantes é de 44,5%
– A diferença média de gols é de 0.29
– O grafico da boxplot da `Goal_Diff_Home` mostra claramente a vantagem do time da casa, com a mediana das vitórias fora de casa sendo de -2 gols e a das vitórias em casa sendo de 2 gols. A mediana dos empates é 0, como esperado.
– A média de gols por jogo oscila, mas se mantem geralmente acima de 2.7 gols por jogo, subindo constantemente ao longo das temporadas.





– Os times mandantes demonstram uma média ligeiramente maior de chutes a gol (13.78 vs. 11.28) e chutes a gol no alvo (4.72 vs. 3.94), confirmando que a vantagem de jogar em casa se reflete na agressividade ofensiva.



– Podemos perceber também na distribuição de resultados que o time da casa tem uma

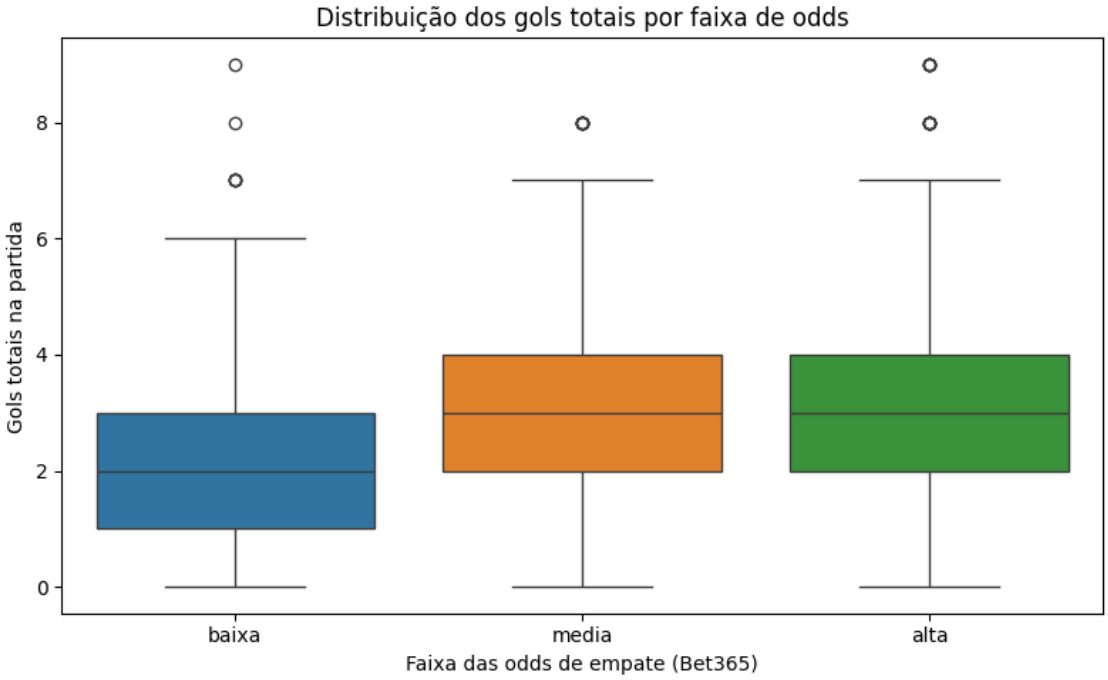
tendência histórica a ganhar mais partidas.

1.4 IV Testes de Hipótese e Intervalos de Confiança

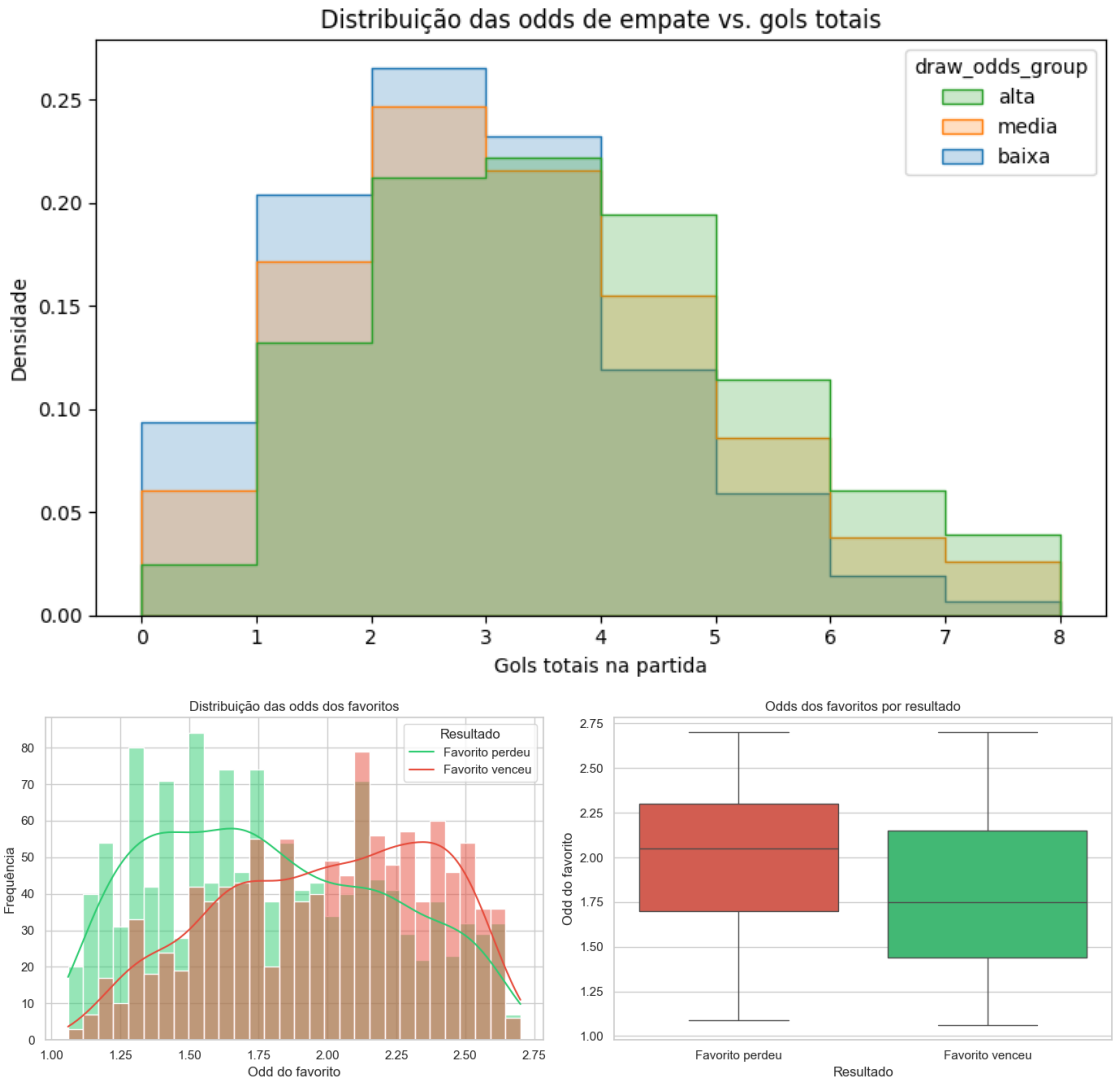
Esta seção detalha o plano, as Hipóteses Nulas (H0) e os resultados estatístico dos quatro testes conduzidos. ### **1. Teste de Proporção: taxa de Sucesso dos Favoritos.** - O objetivo deste teste foi verificar se a taxa de vitória dos times favoritos (definidos pela menor odd pré-jogo) na Premier League é compatível com a referência teórica de **70%**.

Hipótese Nula (H_0): A proporção de vitórias dos favoritos é maior ou igual a 70% ($\pi \geq 0.70$). **Hipótese Alternativa (H_1):** A proporção de vitórias dos favoritos é inferior a 70% ($\pi < 0.70$), sinalizando falhas relevantes nas expectativas das casas. > A referência de **70%** para a taxa de sucesso dos favoritos é comumente citada em literatura de apostas esportivas e análises de mercado de odds. Essa taxa reflete a expectativa de que, em ligas competitivas de futebol, o favorito (definido pelas menores odds) deveria vencer cerca de 7 em cada 10 partidas para que as casas de apostas mantenham margens de lucro consistentes e os apostadores profissionais consigam identificar valor.

Modus operandis: - Utilizou-se um Teste Z para proporção única, adequado para amostras grandes; Partidas onde houve empate na odd mínima (23) foram removidas para definir um favorito claro. - Nível de significância $\alpha = 0.05$, com a hipótese alternativa "smaller" para testar se a proporção observada é menor que 0.7. **Proporção Observada ($\hat{\pi}$):** 53.9% **Intervalo de Confiança (95%):** [51.9%, 55.9%] **Estatística Z:** -15.839 **P-valor:** 0.00000 **Decisão: Rejeitar**

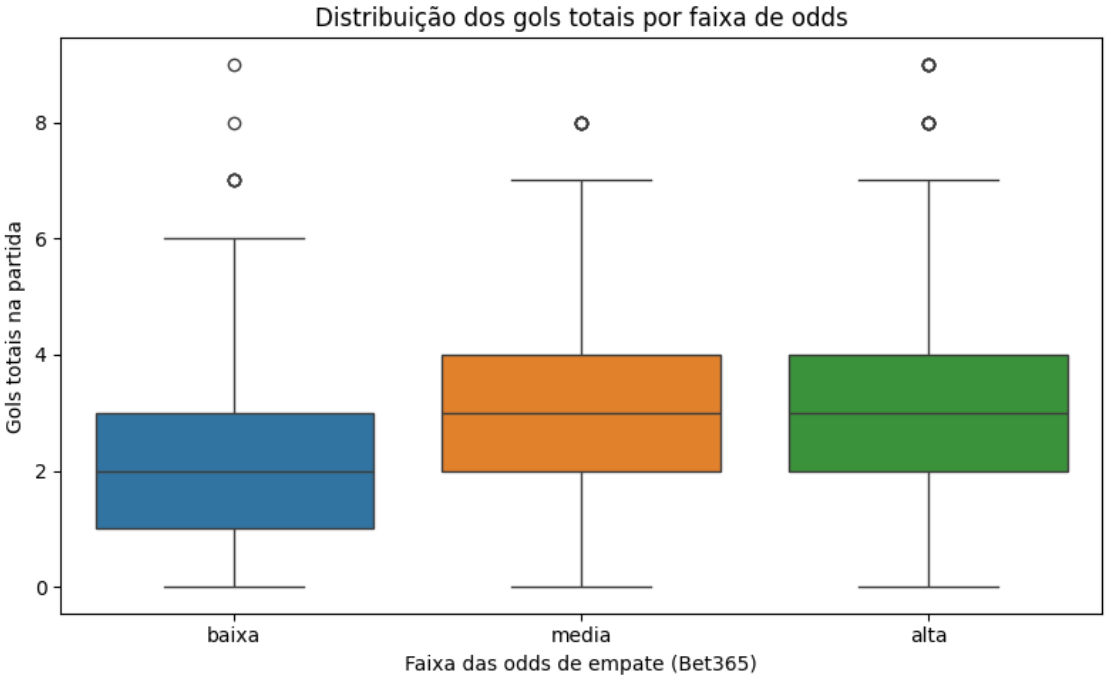


H_0



Conclusão: O teste Z rejeitou categoricamente a hipótese nula. A taxa real de sucesso dos favoritos na Premier League está em torno de 54%, significativamente abaixo da referência de 70%, o que confirma a alta competitividade e imprevisibilidade da liga inglesa. Esta taxa empírica deve ser o novo baseline de sucesso para análises futuras.

2. Análise de Cluster: Disparidade de Resultados por Estilo de Jogo. - O teste avaliou se times com perfis estatísticos semelhantes (definidos por clustering K-Means em K=2) apresentavam médias de gols em casa estatisticamente indistinguíveis. **Hipótese Nula** (H_0):As distribuições de gols do mandante não diferem significativamente entre as equipes classificadas no mesmo cluster de estilo de jogo. **Hipótese Alternativa** (H_1):Mesmo com perfil estatístico semelhante, existem diferenças significativas de gols ou resultados entre times. **Modus operandis:** - As estatísticas ofensivas, disciplinares e de odds médias por equipe e temporada foram consolidadas e padronizadas. - O algoritmo K-Means foi aplicado. O número ideal de clusters foi $\mathbf{K = 2}$, determinado pelo maior Silhouette Score (0.269). O Cluster 0 foi caracterizado como "Ofensivo/Favorito" e o Cluster 1 como "Defensivo/Azarão".



- O Teste de Kruskal-Wallis (Análise de Variância não-paramétrica) foi usado para comparar as médias de gols dentro de cada cluster. **Resultados: Cluster 0 (Ofensivo/Favorito):** 2.038, Teste Kruskal-Wallis (p-valor): p = 2.42e-06 **Cluster 1 (Defensivo/Azarão):** 1.243, Teste Kruskal-Wallis (p-valor): p = 0.0072 **Decisão: Rejeitar** H_0 (para ambos) - **Conclusão:** A rejeição de H_0 demonstra que, mesmo em grupos de "estilos de jogo" semelhantes, a variação de gols entre as equipes é estatisticamente significativa. A performance em campo é, portanto, influenciada por fatores de momentum ou matchups que não são totalmente capturados pelas médias sazonais utilizadas no clustering.

3. Teste de Distribuição (H4): Odds Altas de Empate e Gols Totais

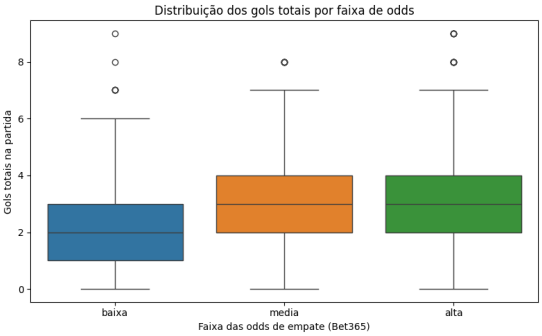
- O teste analisou se as partidas com odds de empate (B365D) mais altas (indicando um favoritismo claro) tinham uma distribuição de gols totais diferente daquelas com odds baixas (jogos mais equilibrados).
- **Hipótese Nula (H0):** A distribuição de gols totais é a mesma para todas as faixas de odds de empate.
- **Modus Operandi:** As partidas foram categorizadas em três faixas ('Baixa', 'Média', 'Alta') com base nos quartis da distribuição de odds de empate. Foi usado o Teste de Kruskal-Wallis para comparar a distribuição de Gols Totais. O Teste Qui-Quadrado foi aplicado para verificar a independência entre a faixa de odds e a ocorrência de empate.

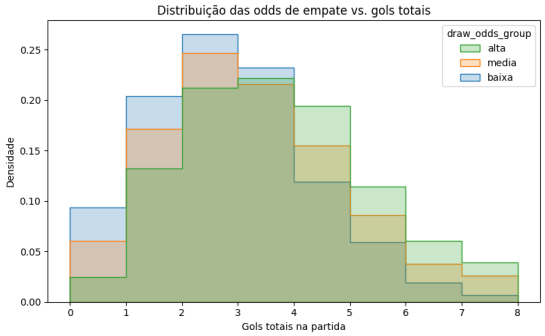
Métrica	Odds Baixas (Equilibrado)	Odds Altas (Desequilibrado)
Odds Média (B365D)	3.287	5.751
Gols Totais Médios	2.375	3.267
Taxa de Empates	29.5%	16.2%

Teste Kruskal-Wallis (p-valor): p = 8.45e-20 (0.0000)

Teste Qui-Quadrado (independência vs. faixa de odds): p 0.0000

Decisão: **Rejeitar H0**





- Conclusão: A rejeição de H0 é clara. Partidas desequilibradas (Odds Altas) resultam em uma média de Gols Totais quase um gol acima da média dos jogos equilibrados. O teste Qui-Quadrado também rejeitou a independência (p 0.0000), validando que a taxa de empates varia significativamente entre as faixas de odds. A odd de empate é um forte proxy para o potencial ofensivo do jogo.