Análise exploratória do Futebol Europeu de 1993 a 2024

Elisa de Oliveira Soares, Gabrielly Esteves Pinheiro Chácara,
João Vitor Tomaz Alves Ferreira, Luiz Eduardo Bravin, Nícolas Mateus Spaniol
Orientador: Rafael de Pinho André

Outubro 2024

Abstract

Neste artigo analisamos dados do futebol europeu, incluindo informações detalhadas sobre partidas, clubes, jogadores e eventos ocorridos em campo extraídos do site Transfermarkt e disponibilizados publicamente no Kaggle. Procuramos responder, usando Python e bibliotecas de visualização e manipulação de dados, à hipóteses desenvolvidas pelo grupo acerca do dataset. Exploramos correlações entre o desempenho de jogadores e seu custo, entre a performance dos clubes em partidas e o fato destes estarem jogando dentro/fora do país e as distribuições dos jogadores por mês de nascimento e de cartões recebidos por posição em campo. Com relação ao mês de nascimento, vimos que a quantidade de jogadores no dataset cai linearmente ao longo do ano, com dezembro tendo cerca de metade dos jogadores de janeiro. A distribuição dos cartões, similarmente, se concentrou nas posições defensivas. Descobrimos que, em média, clubes que vendem e depois compram um mesmo jogador ficam com saldo positivo, e tem o intervalo entre as transações de três anos. Quanto ao desempenho dos clubes dentro e fora do país, não há uma relação direta entre ambos. Em relação aos jogadores com preço fora do comum, os resultados mostraram que estes têm um desempenho levemente superior, mas a diferença em comparação aos demais não é tão significativa.

Palavras-chave: Futebol, Python, visualização de dados, gráfico, estatística, análise de dados

1 Introdução

A análise de dados é bastante relevante em todas as áreas, principalmente no futebol, para analisar previamente qual time tem mais chances de ganhar um jogo, estudar como será definido um campeonato, e diversas outras aplicações. Pensando nisso, formulamos algumas hipóteses sobre tendências no futebol europeu, não com o objetivo de descobrir uma relação existente em alguma delas, mas sim de ampliar a visão sobre cada uma.

Por meio do uso da linguagem Python, foram analisados diferentes aspectos dos times e jogadores, com base em alguns fatores que podem ou não influenciar o desempenho dos mesmos, e elaboramos diferentes tipos de gráficos para facilitar a visualização de cada análise.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta, em cada subseção, uma das hipóteses apresentadas, com o desenvolvimento e o resultado em que chegamos. E, no fim, a Seção 5 discute os resultados e conclusões do trabalho.

2 Desenvolvimento

2.1 Dentre os times visitantes de cada partida, jogar dentro ou fora do país influencia o resultado da partida? (vitória, empate ou derrota)

Na análise dessa hipótese, foram utilizadas as tabelas *"games_csv"* e *"club_games.csv"*. Iniciamos filtrando se a partida era ou não internacional, visto que em todas as partidas internacionais o clube visitante jogou fora de seu país. Após isso separamos os jogos em três tipos: se o time visitante venceu, perdeu ou empatou, e por fim mesclamos as duas informações descobertas para conseguir relacioná-las.

Para uma melhor interpretação dos dados, foi feito um gráfico (Figura 1) com a frequência relativa dos resultados em jogos nacionais ou internacionais, sempre tendo como base o time visitante. Também calculamos o V de Cramer (0.02) para analisar se há ou não uma relação entre as variáveis *"is_international"* e *"result"*.

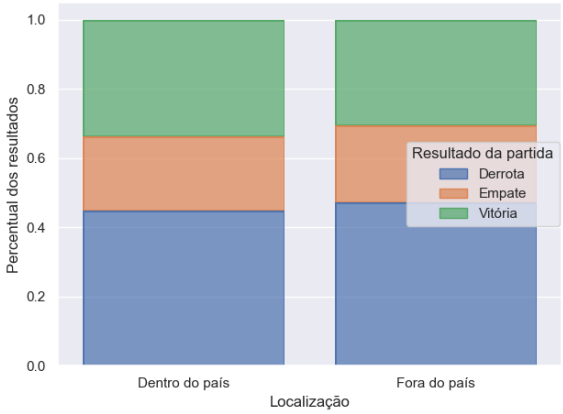

Figure 1: Proporção de resultados internacional ou não. Fonte: elaborado pelo autor.

Por fim, pudemos perceber que há uma baixa associação entre o tipo de jogo (nacional ou internacional) e o resultado. Isto porque os valores observados para ambas as modalidades é muito semelhante e o V de Cramer, que é uma medida que calcula o grau de associação entre duas variáveis, é muito baixo.

2.2 Quais foram as compras de jogadores com melhores e piores custo-benefício registradas?

Na segunda hipótese, o objetivo era identificar quais transferências de jogadores trouxeram maior retorno financeiro aos clubes, avaliando o custo-benefício. Na análise utilizamos variáveis como valor de compra e venda dos jogadores e suas performances em campo, incluindo gols, assistências e cartões recebidos. Esperavamos que jogadores mais caros tivessem um custo-benefício mais estável e um desempenho superior.

Para calcular o custo-benefício, foram ponderadas quatro métricas principais: gols (peso 8), assistências (peso 5), cartões amarelos (peso -1) e cartões vermelhos (peso -3). Esses pesos foram baseados no sistema de pontuação do site "Cartola FC". A análise ajustou os valores de mercado dos jogadores ao longo do tempo, considerando a inflação, para garantir uma comparação justa entre diferentes períodos.

Os dados foram processados utilizando bibliotecas Python como Pandas e Seaborn. Foram descartados registros incompletos e realizados ajustes nos valores de mercado. As tabelas de performances (gols, assistências e cartões) e transferências (valores de compra e venda) foram unidas para gerar a métrica de custo-benefício.

Nome do jogador	Transferência	Clube vendedor	Clube comprador	Custo-benefício
Kylian Mbappé	01/01/2016	Monaco U19	Monaco	1444.8213
Kai Havertz	01/07/2016	Leverkusen U17	Leverkusen	1039.95
James Ward-Prowse	02/07/2012	Southampton U18	Southampton	850.0176
Matthijs de Ligt	01/01/2017	Ajax U21	Ajax	805.4439
Federico Chiesa	01/07/2016	Fiorentina U19	Fiorentina	647.3562
...	...	...	...	...
Morgan Schneiderlin	30/09/2023	Without Club	AE Kifisias	-1.0644
Pedro Tiba	31/08/2015	SC Braga	Real Valladolid	-1.0948
Graeme Shinnie	16/01/2022	Derby	Wigan	-1.1205
Naby Keïta	01/07/2023	Liverpool	Werder Bremen	-1.1613
Mathieu Gorgelin	01/07/2019	Olympique Lyon	Le Havre AC	-1.6626

Figure 2: Tabela com os 5 melhores e 5 piores resultados baseado no custo-benefício calculado. Fonte: elaborado pelo autor.

Os resultados destacaram Kylian Mbappé como o jogador com o melhor custo-benefício na transição entre Monaco U19 e Monaco. Ele superou os demais em desempenho e valorização de mercado. Por outro lado, Mathieu Gorgelin foi considerado o atleta de pior custo-benefício com a transição entre os clubes Olympique Lyon e Le Havre AC. Os Top 5 melhores e piores está listado na Figura 2.

2.3 A posição dos jogadores em campo influencia na quantidade de cartões que estes recebem?

Para analisar essa hipótese, utilizamos as tabelas *"game_events.csv"* e *"game_lineups.csv"*. Foi feita uma tabela com os cartões por posição em que cada cartão, amarelo ou vermelho, foi colocado na posição de acordo com a posição em que o jogador estava atuando no jogo em quelevou o cartão. Posições que tinham uma quantidade insuficiente de dados (menos de 150 cartões em todos os jogos) foram removidas da análise.

Primeiramente, foram filtrados apenas os eventos de recebimento de cartões, e a posição em que o jogador estava atuando no jogo. Foram feitas dois gráficos para uma visualização mais precisa dos dados: um gráfico que mostra a quantidade total de cartões recebidos por cada posição em campo (Figura 3), e um que mostra a proporção de cartões vermelhos e amarelos por posição (Figura 4).

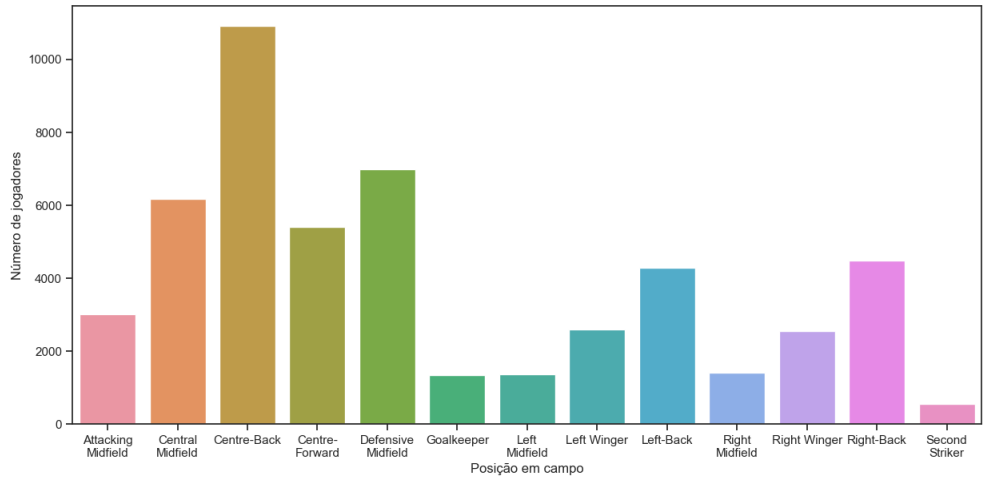

Figure 3: Gráfico de cartões por posição. Fonte: elaborado pelo autor.

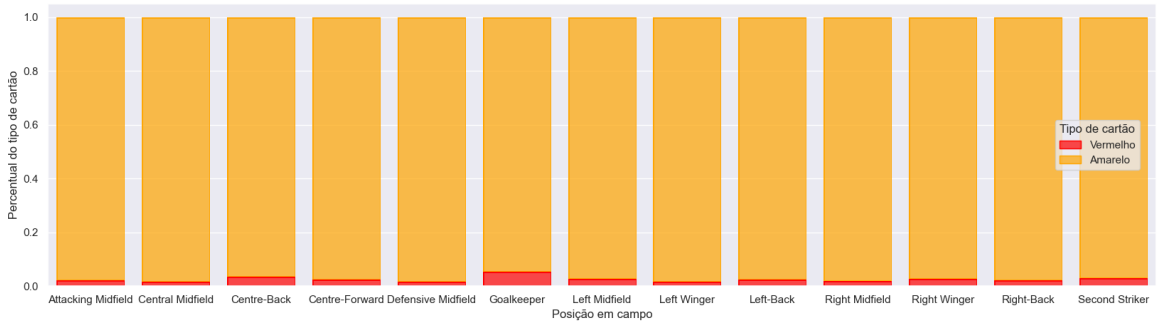

Figure 4: Proporção de cartões por posição. Fonte: elaborado pelo autor.

Enfim, pudemos observar que as posições defensivas, como *Centre-Back*, *Defensive Midfield*, *Left-Back* e *Right-Back* são as que mais tomam cartão. Enquanto isso, as que menos recebem cartões são mais bem espalhadas, contendo pessoas de todas as áreas do jogo, como o *Second Striker* (ataque), *Goalkeeper* (defesa), *Left Midfield* e *Right Midfield* (ambos do meio de campo). Já sobre a proporção de cartões, é possível ver que o goleiro é o que recebe uma maior porcentagem de cartões vermelhos.

2.4 Vendas e compras posteriores de um jogador por um mesmo time costumam gerar lucro para o time?

Para a análise dessa hipótese, usamos as tabelas *"players.csv"* e *"transfers.csv"*. Primeiro, conferimos se um jogador foi vendido por um time e posteriormente comprado pelo mesmo time. Em seguida, analisamos se, após a sequência de transações, o time teve lucro ou prejuízo com esse jogador.

Foram feitos dois gráficos para a análise mais precisa desses dados: Um boxplot com o saldo final do time, para visualizar qual a tendência em operações de compra pós venda (Figura 5), e um boxplot do intervado entre a venda e a compra (Figura 6). Foram calculadas também medidas de tendência, como a mediana do saldo (€ 225.000,00), o desvio padrão do saldo (€ 13.198.979,93), e as médias da idade de venda (23,1), da idade de compra (26,3) e do intervalo entre a venda e a compra (3,2) .

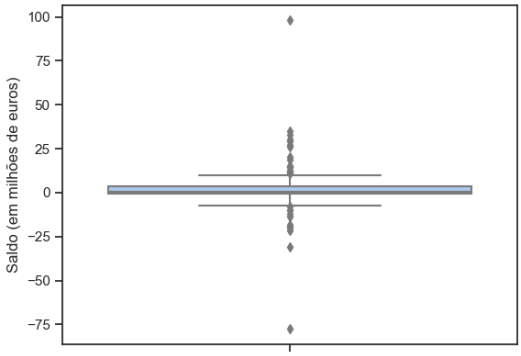

Figure 5: Saldo final do time. Fonte: elaborado pelo autor.

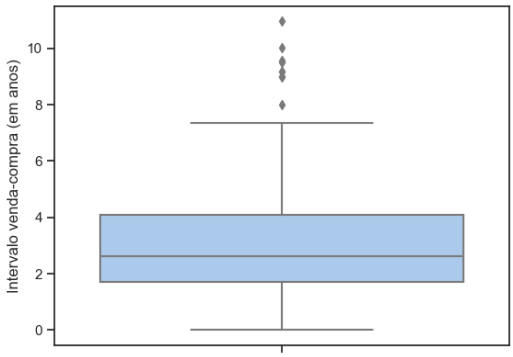

Figure 6: Intervalo entre venda e compra. Fonte: elaborado pelo autor.

Assim, podemos observar que os times costumam ter um pequeno lucro após a venda e posterior compra de jogadores, mas isso não é algo certo, visto que muitos times podem ter pequenos prejuízos, grandes prejuízos ou até mesmo grandes lucros, visto que o desvio padrão é alto.

2.5 Jogadores com preço fora do comum tem o desempenho proporcional?

Com o aumento do número de jogadores com valores de mercado extremamente altos, surgiu a dúvida sobre a relação entre o desempenho em campo e o valor atribuído a esses atletas. Para essa análise, focamos nos jogadores classificados como outliers - com preços de mercado significativamente acima da média - definidos através do limite superior, que é obtido somando o terceiro quartil (Q3) a 1,5 vezes o intervalo interquartil (IQR), conforme descrito por Bussab e Morettin (2017, p. 53).

Esperávamos que esses jogadores apresentassem um desempenho consistentemente superior, justificando seus altos valores. Kylian Mbappé é um exemplo claro, sendo um dos mais caros e com uma performance excelente.

O "desempenho proporcional" foi analisado através da razão entre o desempenho e o valor médio do jogador, buscando identificar se essa relação era equilibrada. A coleta de dados veio das tabelas *"appearances.csv"* e *"player_valuations.csv"*, eliminando linhas incompletas. Identificamos os outliers e avaliamos o desempenho com pesos específicos para diferentes estatísticas, como gols, assistências e cartões, semelhantes aos usados na hipótese 2. Utilizamos gráficos de dispersão para comparar o desempenho com o valor de mercado, aplicando o logaritmo neperiano para facilitar a visualização.

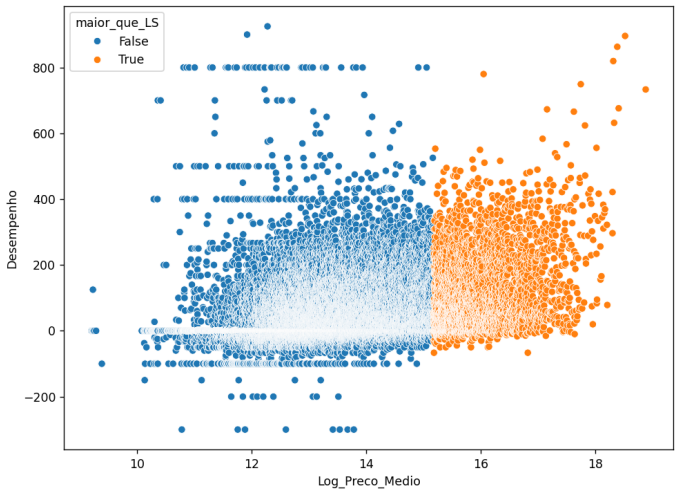

Figure 7: Gráfico de dispersão do logaritmo do preço médio do jogador pelo seu desempenho ao longo da carreira. Os jogadores com preço fora do comum se encontram em laranja, enquanto os restantes em azul. Fonte: elaborado pelo autor.

Os resultados mostraram que, embora alguns jogadores de alto valor tivessem um semelhante desempenho ao dos jogadores com preços mais baixos, eles ainda apresentaram desempenho levemente superior. A correlação entre valor e performance dos outliers foi de aproximadamente 32,2%, o que indica uma relação moderada, mas não tão forte quanto o esperado. Concluímos que, apesar de serem mais bem pagos, a diferença de desempenho entre os jogadores mais caros e os demais não é tão significativa quanto se supunha.

2.6 Pessoas que nascem na primeira metade do ano tem mais chance de se tornarem jogadores profissionais?

O jornalista britânico Malcolm Gladwell, em seu livro *Outliers: The Story of Success (2008)*, escreveu sobre a relação entre o mês de nascimento dos jogadores de Hoquei no Canadá e sua ascenção no esporte, e percebeu que uma tendência dos jogadores profissionais nascerem majoritariamente na primeira metade do ano.

Com base nisso, nos questionamos se esse padrão é válido também para outros tipos de esportes. Para realizar essa análise, utilizamos a coluna *"date_of_birth"* da tabela *"players"* para criar uma nova tabela com a quantidade e a frequência de jogadores que fazem aniversário por mês. Com esses dados, foi feito um gráfico das frequências por mês (Figura 8) e também o cálculo do desvio relativo dos aniversários, que é de 653,32.

A partir da análise do gráfico e do desvio relativo, foi possível perceber que, assim como no Hóquei canadense, há uma relação entre os meses de aniversário mais frequentes no futebol europeu. De acorco com Gladwell, essa relação tem causa na forma que os jogadores são selecionados para os treinos e clube infantis, desde a infância. Essas seleções têm, em sua maioria, limite de idade para se inscrever, e crianças que nasceram e janeiro e dezembro de um mesmo ano apresentam grandes diferenças de desenvolvimento.

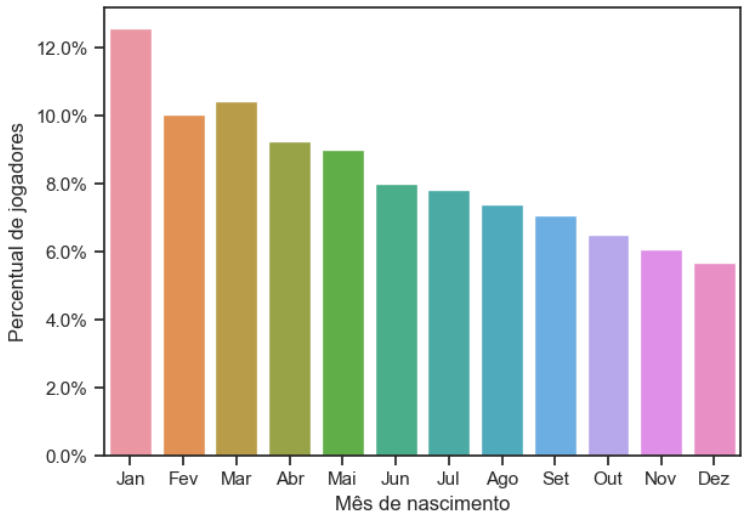

Figure 8: Frequência de jogadores que fazem aniversário por mês. Fonte: elaborado pelo autor.

3 Considerações finais

No processo de analisar o dataset a fim de responder as hipóteses propostas, chegamos não só a conclusões a respeito dessas como também a um comentário a respeito do dataset de futebol utilizado. Pela própria natureza do esporte, a coleta de dados de partidas de futebol atualmente envolve alguma imprecisão e falta de dados, presentes também nesse dataset. Para informações como transferências entre times, dados anteriores a 2010 são escassos e tornam difícil a análise exploratória.

A pesquisa atual pode ser estendida para além do futebol europeu. Embora não tão completos, dados similares de partidas em campeonatos brasileiros podem ser encontrados e a maioria das hipóteses aqui desenvolvidas podem ser respondidas levando em consideração o cenário brasileiro ou de outros países presentes no esporte. A coleta, limpeza e distribuição de mais dados relacionados ao futebol brasileiro também são essenciais no desenvolvimento de mais pesquisas nesta área e no país.

Dito isso, foi possível chegar em resultados concretos para algumas das nossas hipóteses. Na hipótese 2.1, foi possível concluir que não há associação entre o resultado do jogo e onde o mesmo ocorre; já na 2.3, há uma clara associação entre as posições defensivas e o recebimento de cartões. As demais hipóteses foram verificadas de forma mais superficial e não levam a conclusões satisfatórias.

References

GLADWELL, Malcolm. *Outliers: The Story of Success*. 1. ed. Nova York: Little, Brown and Company, 2008.

CARIBOO, David. Player Scores. Kaggle, 2023. Disponível em: https://www.kaggle.com/datasets/davidcariboo/player-scores. Acesso em: 02 out. 2024.

BUSSAB, Wilton de O.; MORETTIN, Pedro A. *Estatística Básica*. 8. ed. São Paulo: Saraiva, 2017.

CARTOLA FC BRASIL. Como funciona o sistema de pontuação do Cartola FC. Disponível em: https://www.cartolafcbrasil.com.br/tutoriais/7/como-funciona-o-sistema-de-pontuacao-do-cartola-fc. Acesso em: 04 out. 2024.