Trabalho de Conclusão de Curso

Comparativo entre o campeonato brasileiro de futebol e as principais ligas europeias de futebol: uma análise estatística

Victor Accete Nicácio Placido
victor.accete@gmail.com

Orientador:

Bruno Almeida Pimentel

Maceió, Novembro de 2024

Victor Accete Nicácio Placido

Comparativo entre o campeonato brasileiro de futebol e as principais ligas europeias de futebol: uma análise estatística

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Bruno Almeida Pimentel

Maceió, Novembro de 2024

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Girlaine da Silva Santos – CRB-4 – 1127

P698c Placido, Victor Accete Nicácio.
Comparativo entre o campeonato brasileiro de futebol e as principais ligas europeias de futebol: uma análise estatística/ Victor Accete Nicácio Placido. – 2024.
52 f.: il. color.

Orientador: Bruno Almeida Pimentel.
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação: Bacharelado) - Universidade Federal de Alagoas, Instituto de Computação, Maceió, 2024.

Bibliografia: f. 47- 52.

1.Análise de dados. 2. Futebol. 3. Análise estatística. I. Título.

CDU: 004.67:796.332

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina.

Prof. Dr. Bruno Almeida Pimentel - Orientador
Instituto de Computação
Universidade Federal de Alagoas

Profª. Dra. Roberta Vilhena Vieira Lopes - Examinador
Instituto de Computação
Universidade Federal de Alagoas

Prof. Dr. Evandro de Barros Costa - Examinador
Instituto de Computação
Universidade Federal de Alagoas

Maceió, Novembro de 2024

Agradecimentos

Agradeço ao meu orientador, Prof. Dr. Bruno Pimentel, pela solicitude, disponibilidade, paciência e parceria nesse período de TCC. Suas orientações e direcionamentos foram fundamentais.

Agradeço, desde já, à banca examinadora, pelo *feedback* fornecido e por se dedicarem à correção e avaliação deste trabalho.

Agradeço aos professores que fizeram a diferença em minha graduação, direta ou indiretamente, sobretudo através do e das oportunidades que me foram dadas.

Agradeço a todos os colegas que fiz ao longo do curso, em diversas turmas, pela colaboração e parceria durante todo esse tempo.

Agradeço aos amigos na UFAL em outros cursos, em especial a Pedro César, Jonathas, Zaíne e Hanna. Foi muito bom compartilhar cafezinhos com vocês.

Agradeço a todos os amigos de fora da faculdade, pelo apoio, incentivo, orações e por todos os momentos compartilhados juntos. Em especial, agradeço a todas as pessoas do Pequeno Grupo (PG) liderado por Raphael e Gabi.

Agradeço aos meus pais, Alexandre e Verônica, pelo incentivo, exemplo, suporte, orações e por terem me dado condições de alcançar tantas conquistas.

Agradeço aos meus irmãos, Raphael e Philipi, pela amizade e colaboração; e também às suas esposas, Gabriella e Sara.

Agradeço à minha esposa, Isabelle, por todo o incentivo, compreensão, amizade, paciência e orações em boa parte dessa trajetória. Não sei o que seria de mim sem você.

Por fim, agradeço a Deus, o Pai, por ter me criado e cuidadosamente planejado todos os meus dias.

A Deus, o Filho, por ter me dado o maior presente que alguém pode receber: a Salvação.

A Deus, o Espírito Santo, por ser o professor e conselheiro mais excelente.

Resumo

Apesar do futebol ser o esporte mais popular do Brasil, o campeonato brasileiro de futebol é sub-representado em estudos científicos que analisam aspectos técnicos e táticos do esporte. Estudos que comparam ligas são relativamente comuns, mas eles geralmente focam nas principais ligas europeias. Este trabalho se propõe a analisar como o campeonato brasileiro de futebol se compara, em aspectos técnicos e táticos, às quatro principais ligas europeias de futebol: Premier League da Inglaterra, La Liga da Espanha, Bundesliga da Alemanha e Série A da Itália. Apesar de existir uma grande diferença em receitas e valor de mercado dos elencos, este trabalho explora a oportunidade de analisar se essas diferenças se traduzem em grandes diferenças no que diz respeito a aspectos técnicos e táticos do esporte, avaliando algumas métricas de desempenho notacionais. Além disso, este trabalho analisa também o equilíbrio interno das ligas. Para as análises, foram coletados dados notacionais das ligas e das equipes em cinco temporadas. Os resultados demonstraram que o campeonato brasileiro possui indicadores piores em diversas métricas de desempenho, especialmente ao se considerar apenas os clubes do topo das tabelas de classificação. No entanto, também foi possível observar que algumas métricas foram similares ou até melhores, apesar da discrepância econômica. Também foi possível perceber que o campeonato brasileiro é mais equilibrado do que as principais ligas europeias.

Palavras-chave: estatística, estatística descritiva, análise exploratória de dados, ciência de dados, visualização de dados, análise de futebol, análise esportiva

Abstract

Although football is the most popular sport in Brazil, the brazilian football league is under-represented in scientific studies that analyse technical and tactical aspects of the sport. Studies that compare leagues are relatively common, mas they usually focus on the major european leagues. This work aims to analyse how the brazilian football league compares, in technical and tactical aspects, to the four major european football leagues: English Premier League, Spanish La Liga, German Bundesliga and Italian Serie A. Even though there is a big difference in revenue and market value of the squads, this work explores the opportunity to analyse if these differences translate into big differences concerning technical and tactical aspects of the sport, by analysing a few notational performance metrics. Besides this, this work also analyses the internal balance of the leagues. For the analyses, we collected five seasons of leagues' and squads' notational data. Results showed that the brazilian league has worse indicators in many performance metrics, especially when considering only the top clubs from ranking tables in each season. However, it was also possible to observe that some metrics were similar ou even better, despite the economic discrepancy. It was also possible to see that the brazilian league if more balanced than the major european leagues.

Key-words: statistics, descriptive statistics, exploratory data analysis, data science, data visualization, football analytics, sports analytics

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

xG	Expected Goals (Gols Esperados).
PSxG	Pós-Shot Expected Goals (Gols Esperados Pós-Chute).
npxG	Non-penalty Expected Goals (Gols Esperados Sem Pênaltis).
FIFA	Fédération Internationale de Football Association (Federação Internacional de Futebol).
PSxG-GA	Post-shot Expected Goals minus Goals Allowed (PSxG menos Gols Permitidos).

Conteúdo

1

Introdução

Este capítulo apresenta uma introdução deste trabalho. Primeiro, mostrando a motivação deste trabalho e, em seguida, a sua justificativa. Depois, apresentando os objetivos, tanto o objetivo geral quanto os objetivos específicos. Por fim, será apresentada a estrutura do documento.

1.1 Motivação

O futebol é o esporte mais popular do mundo, espalhado por mais de 200 países [33]. Segundo dados oficiais da FIFA, a final da Copa do Mundo atingiu a marca de 1,5 bilhão de pessoas assistindo ao evento, superando a marca da copa do mundo anterior, que atingiu a marca de 1,12 bilhão em 2018 [6].

Esse esporte é, também, o mais popular do Brasil há décadas e amplamente praticado por brasileiros. Segundo dados de 2017 do IBGE, cerca de 15,3 milhões de brasileiros praticam o futebol [5].

No entanto, apesar da popularidade do esporte no Brasil, o campeonato brasileiro de futebol arrecada significativamente menos que as principais ligas europeias, em especial as quatro maiores: Premier League (Inglaterra), La Liga (Espanha), Bundesliga (Alemanha) e Série A (Itália). Dessas quatro ligas, a que possui a menor arrecadação é a Série A da Itália, com arrecadação de 2,8 bilhões de euros na temporada 2023-2024 [10]. Esse valor representa o dobro da receita de 1,4 bilhão de euros do campeonato brasileiro na temporada de 2023 [7]. A Premier League, por sua vez, foi a liga que mais arrecadou no mundo, com arrecadação de mais de 7 bilhões de euros na temporada 2023-2024 [10], cinco vezes mais que o campeonato brasileiro.

As ligas europeias são amplamente reconhecidas como as mais competitivas do mundo, pela qualidade técnica e alto nível dos atletas. Os atletas mais valiosos do mundo estão nessas ligas. De acordo com dados do Transfermarkt, cada uma das principais quatro ligas da Europa possuem valor de mercado acumulado de 11,74 bilhões de euros (Premier League) [14]; 5,47

bilhões de euros (La Liga) [13]; 4,87 bilhões de euros (Série A da Itália) [15]; 4,49 bilhões de euros (Bundesliga) [11]. Os jogadores do campeonato brasileiro possuem um valor acumulado de 1,65 bilhão de euros [12]. Diante disso, é interessante verificar se essas diferenças de receitas e valor de mercado do elenco se traduzem em diferenças em aspectos técnicos e táticos do esporte.

1.2 Justificativa

Há estudos que investigam com mais profundidade em relação aos investimentos dos clubes brasileiros [34]. No entanto, acerca de estudos relacionados a aspectos técnicos e táticos do futebol, a revisão sistemática de Otero-Saborido et al. [38] observou que a La Liga da Espanha e a Premier League da Inglaterra são as ligas mais estudadas, entre 31 estudos analisados que focavam em aspectos técnicos e táticos. A La Liga foi estudada em 27,91% dos trabalhos e a Premier League em 18,60%. Nesse estudo, não há menção explícita ao campeonato brasileiro, apenas há menção a "outros países", demonstrando que estudos relacionados a aspectos técnicos e táticos não são frequentes em relação ao campeonato brasileiro.

Estudos que comparam dados entre as ligas, em geral, focam nas chamadas *Big 5*, isto é, as cinco maiores ligas europeias: Premier League (Inglaterra), LaLiga (Espanha), Bundesliga (Alemanha), Série A (Itália), e Ligue 1 (França) [32][41][53][31]. Observa-se, então, que estudos que comparam ligas são comuns, mas o campeonato brasileiro acaba sendo sub-representado nesses estudos. Pestana Barros et al. [40] observa que pesquisa sobre a eficiência técnica dos clubes é comum na Europa, mas incomum em outras ligas de futebol. Além disso, Pestana Barros et al. [40] também nota que, embora o futebol brasileiro seja o maior exportador de jogadores internacionais, até o momento em que ele escreveu, nenhum estudo tinha abordado o desempenho do campeonato brasileiro.

Diante de tudo isso, fica clara a sub-representação do campeonato brasileiro de futebol nos estudos, em especial em estudos comparativos com outras ligas. É interessante verificar se a diferença de arrecadação, valor de mercado do elenco e investimento implica em uma diferença significativa em dados de desempenho esportivo.

Essa análise comparativa entre o campeonato brasileiro e as principais ligas europeias é relevante na busca por uma melhor compreensão das diferenças entre as ligas. Este trabalho também possui relevância porque disponibiliza os dados coletados em forma sumarizada. Os dados importados, sumarizados e analisados, bem como as visualizações fornecidas, podem servir de base para estudos futuros que avaliem aspectos sócio-culturais, técnicos e táticos do futebol brasileiro em comparação com as ligas europeias de maior impacto global. Esses dados e análises fornecidos podem ser usados para fundamentar estudos futuros, visto que estudos avaliando aspectos técnicos e táticos do campeonato brasileiro ainda são escassos, como mencionado anteriormente.

1.3 Objetivo

Este trabalho se propõe a investigar se essa diferença de arrecadação e valor de mercado do elenco discutidos anteriormente se traduz em uma diferença significativa em algumas das principais métricas usadas para avaliar o desempenho de times e atletas; além de avaliar o equilíbrio interno de cada uma das cinco ligas.

Ao fim deste trabalho, no capítulo de resultados, serão apresentados os resultados e visualizações para as métricas analisadas. Sendo assim, propõe-se como objetivos específicos:

- Analisar médias das últimas cinco temporadas (começando em 2019 até a temporada que começou em 2023), para cada uma das ligas, para as seguintes métricas: gols, gols esperados (xG), gols esperados sem pênalti (npxG), total de chutes, chutes a gol, gols por chute, defesas de goleiro, percentual de defesas por chutes a gol, gols esperados pós chute menos gols permitidos (psxG-GA), número de passes concluídos, percentual de passes completos, passes progressivos, botes defensivos, confrontos no terço de ataque, erros defensivos que resultaram em uma tentativa de chute do time adversário, número de dribles tentados, dribles certos, dribles que resultaram em uma finalização, cartões (amarelos mais vermelhos), faltas cometidas.

- Analisar as médias das últimas cinco temporadas, para cada uma das ligas, para as mesmas métricas; mas, desta vez, considerando apenas os times de melhor classificação e comparando com os times de pior classificação de cada temporada, a fim de verificar como isso afeta os resultados e avaliar a discrepância entre as equipes do topo e a do fundo da tabela de classificação.

- Analisar a variabilidade, através de histogramas e desvio padrão, de cada métrica para cada uma das ligas, a fim de avaliar o equilíbrio interno de cada campeonato.

É importante notar que, de acordo com Downey [26], para trabalhar com dados, pode-se pensar não apenas no nível de estatística, mas também no nível de contexto. No caso específico deste trabalho, os dados descritos podem não ter as suas razões bem definidas. Por exemplo, o fato de que a liga brasileira possui mais dribles é algo que pode ter fatores sociais e culturais envolvidos como causa [25].

Este trabalho não se propõe a se aprofundar nessas relações do esporte com fatores sociais e culturais. Antes, o objetivo é, através das métricas mencionadas, realizar análise e sumarização dos dados, para investigar se é possível identificar uma diferença significativa entre o campeonato brasileiro de futebol e as principais ligas europeias de futebol, comparando aspectos estatísticos. Ao analisar algumas estatísticas, pretende-se compreender melhor e comparar a dinâmica das competições, fornecendo uma visão um pouco mais detalhada sobre desempenho dos clubes em suas respectivas ligas. Além disso, também é objetivo deste trabalho investigar como essas estatísticas se comparam ao considerar apenas os clubes de melhor classificação

em cada ano, para suas respectivas ligas; e como se comparam ao considerar apenas os clubes de pior classificação em cada ano, em suas respectivas ligas. O objetivo é analisar também a discrepância dessas métricas entre os melhores e os piores clubes, para ter melhor compreensão acerca da competitividade das ligas.

1.4 Estrutura do documento

No capítulo 2, será fornecido um breve contexto sobre diversos conceitos da análise estatística, sobretudo acerca de dados, variáveis, estatística descritiva, visualização de dados e análise exploratória de dados. Também será discutida a ciência de dados aplicada especificamente ao futebol, demonstrando como esse domínio de aplicação cresceu em relevância nos últimos anos.

O capítulo 3 descreverá as ferramentas usadas em todas as etapas deste trabalho. Será descrito o processo para coleta dos dados, bem como uma descrição mais detalhada sobre os conjuntos de dados usados. Este capítulo também abordará os processos de transformação e tratamento de dados adotados para preparar os dados para as etapas subsequentes. Além disso, serão apresentadas as métricas e visualizações adotadas para avaliação, baseadas na fundamentação teórica do capítulo 2, bem como métodos e ferramentas usadas para suas computações.

O capítulo 4 apresentará os resultados por meio de métricas, visualizações gráficas e tabelas para facilitar a compreensão dos dados. Serão apresentados gráficos relevantes para a interpretação dos resultados. Além disso, em cada subseção, será proposta uma breve discussão a respeito dos resultados obtidos.

Por fim, no capítulo 5, será dada uma conclusão deste trabalho, resumindo os resultados apresentados, sumarizando as discussões propostas no capítulo 4 e apresentando algumas limitações deste trabalho. Também serão discutidas oportunidades de trabalhos futuros que podem ser desenvolvidos a partir deste trabalho ou usá-lo como referência.

2

Fundamentação Teórica

Ao longo deste capítulo, será apresentada a fundamentação teórica deste trabalho. Inicialmente, será introduzido o conceito de análises estatísticas e sua importância. Em seguida, o foco estará na descrição dos dados, abordando tipos de dados e de variáveis, além dos tipos de estudos baseados em análises de dados. A partir disso, serão introduzidos os conceitos de estatística descritiva, visualizações de dados comuns e análise exploratória de dados. Por fim, será apresentado um breve panorama sobre estudos relacionados à aplicação da ciência de dados no futebol.

2.1 Análises estatísticas

Dados são coletados e usados em vários domínios de aplicação, incluindo economia, biologia, engenharia, marketing, esportes, entre outros [24][30][37]. No entanto, em diversos casos, opiniões e teorias são baseadas em evidências anedóticas. Evidências anedóticas são evidências baseadas em observações pessoais, coletadas e analisadas de maneira casual e menos estruturada [26]. O futebol, com todo o seu impacto no imaginário popular não apenas brasileiro, mas de todo o mundo, se torna ainda mais sujeito a esse tipo de análise. Diversas opiniões são dadas a respeito de diversos assuntos relacionados a futebol, sem o devido respaldo dos dados.

Essas opiniões podem ser em relação ao nível geral de qualidade e competitividade das ligas de futebol espalhadas pelo mundo. Em uma discussão de 2023 em um tópico no *Reddit*, foi levantada uma discussão com a seguinte pergunta: "Onde a liga brasileira se classificaria entre as ligas europeias?"[16]. Um usuário respondeu, baseado em suas experiências anedóticas, que o campeonato brasileiro ficaria atrás do campeonato holandês e português neste suposto *ranking*, o que gerou discussões subsequentes. No entanto, a empresa de análises esportivas Opta realizou um *ranking* que classificou o campeonato brasileiro como o sexto melhor campeonato de futebol do mundo [19], imediatamente acima da liga portuguesa (em sétimo) e acima da liga holandesa (em décimo-terceiro).

Os dados podem confirmar ou contradizer as opiniões anedóticas [26], há exemplos para ambos os casos. E, em conversas e situações informais, onde dados nem sempre estão disponíveis para serem consultados rapidamente, as anedotas podem ser úteis [26]. No entanto, em muitos casos, é necessário contar com evidências mais confiáveis. As evidências anedóticas estão mais propensas a falhas devido aos seguintes fatores [26]:

- Pequeno número de observações. Em muitas situações, o número de casos observados pelas pessoas é bem restrito, sendo insuficientes para fazer generalizações confiáveis;
- Viés de seleção. Determinadas discussões atraem determinados tipos de pessoas, fazendo com que a seleção para os dados não seja aleatória;
- Viés de confirmação. As pessoas são mais tendenciosas a observar situações que confirmam suas opiniões e hipóteses;
- Inacurácia. Acontecimentos pessoais estão sujeitos a serem mal lembrados, mal representados e mal comunicados.

Por isso, é necessário uma abordagem mais confiável, usando ferramentas da estatística. Estatística é uma ramificação da matemática, que lida com a coleta, organização e interpretação de dados [37]. Processar dados extrai informação que, ao ser processada, extrai conhecimento [37]. Para uma abordagem estatística, duas das ferramentas da estatística que podem ser adotadas são: estatística descritiva e análise exploratória de dados. Para ambos os casos, o ponto de partida são os dados.

2.2 Dados

Os dados são elementos fundamentais da estatística [24]. Organizar os dados é o primeiro passo na maioria das análises [37]. São os dados que contêm a verdade fundamental sobre eventos e experimentos [37]. Por isso, para extração de conhecimento é essencial que os dados capturados sejam relevantes. No entanto, os dados podem aparecer sob diferentes formas.

2.2.1 Dados estruturados, não estruturados e semi-estruturados

Dados podem ser [37]:

- Estruturados;
- Não estruturados;
- Semi-estruturados.

Dados estruturados mantêm uma estrutura uniforme em todas as observações, como por exemplo uma tabela ou um banco de dados relacional. São dados que possuem um formato predefinido [37].

Dados não estruturados, por sua vez, podem aparecer sob várias formas e tamanhos. Texto, imagem, áudio, vídeo, arquivo de documento são exemplos de dados não estruturados [37].

Já os dados semi-estruturados são dados que mantêm alguma estrutura, mas que não obedecem a um modelo tabular. Exemplos de dados semi-estruturados são arquivos XML e JSON [37]. Dados semi-estruturados são úteis quando a estrutura é útil, mas também é necessário que haja a flexibilidade que pode faltar aos dados estruturados [22].

2.2.2 Tipos de variáveis

As variáveis de um conjunto de dados também podem ser divididas, de maneira geral, em dois grupos [37]:

- Variáveis quantitativas;
- Variáveis qualitativas.

Variáveis quantitativas, também chamadas de variáveis numéricas, são variáveis que podem assumir uma gama de valores numéricos e que faz sentido realizar operações matemáticas com esses dados, como por exemplo soma, subtração ou cálculo de média [24]. Uma variável é dita numérica ou quantitativa quando os seus valores diferem em quantidade [29]. Alguns dados, como por exemplo número de CPF, apesar de serem números, não são considerados variáveis numéricas ou quantitativas.

Variáveis quantitativas podem ser discretas ou contínuas [24][29]. Se uma variável possui dois valores em que nenhum valor pode ser assumido entre eles, ela é considerada uma variável discreta [29]. Variáveis contínuas, por sua vez, podem assumir qualquer valor entre dois valores dados [29].

Uma forma proposta para ajudar a diferenciar se uma variável numérica é discreta ou contínua é tentar aplicar o teste do "meio-termo"[29]: se, para todo par de valores de uma variável, um valor entre eles é relevante, então a variável é contínua. Por exemplo, o número de gols que um time marca, não pode ser 0,5 (valor entre 0 e 1), portanto, número de gols que um time marca é uma variável discreta. Por sua vez, a métrica probabilística de gols esperados (xG), é uma métrica que, por definição, pode assumir qualquer valor entre 0 e 1 [36] e, portanto, é uma variável contínua. Em outras palavras, quando uma variável pode ser contada ela é discreta; e quando uma variável precisa ser mensurada, ela é contínua [17].

Variáveis qualitativas são aquelas cujos valores diferem em qualidade [29]. As variáveis qualitativas podem ser nominais ou ordinais [29]. Variáveis nominais são variáveis cujas categorias não envolvem nenhum tipo de ordem, por exemplo, gênero, cor dos olhos, entre outros.

Por sua vez, variáveis ordinais, são variáveis qualitativas cuja ordem importa, como, por exemplo, nível de satisfação.

É importante entender o tipo dos dados e das variáveis, pois isso ajuda a determinar as melhores visualizações, medidas, análises e modelos estatísticos a serem usados [21]. Além disso, do ponto de vista computacional, *softwares* classificam dados por tipos e, em muitos casos, essa classificação é feita de maneira automática; por isso, é útil saber classificar os dados corretamente, como uma forma de sinalizar ao *software* como processar os dados [21].

Uma variável pode ser considerada dependente de outra quando há algum tipo de conexão entre elas [24]. Variáveis são tidas como independentes quando não há nenhum relacionamento evidente entre elas [24]. É importante notar que um par de variáveis é, necessariamente, dependente ou independente; um par de variáveis não pode ser ambos [24].

2.2.3 Estudos observacionais e experimentos

Há, principalmente, duas formas de coletas de dados: estudos observacionais ou experimentos [24]. Estudos observacionais são estudos em que os dados são coletados de maneira que a coleta não afeta a forma como os dados surgem [24], os grupos são formados e os dados são gerados de forma espontânea. Em experimentos, o ambiente é mais controlado, os participantes são escolhidos e atribuídos a grupos. Estudos de análises de desempenho esportivas são, em geral, estudos observacionais. Para este trabalho, todos os dados observados seriam gerados da mesma forma sem nenhum tipo de intervenção.

2.3 Estatística descritiva

Estatística descritiva, como o nome sugere, é usada para ajudar a compreender e descrever conjuntos de dados [26][29][37]. É usada para sumarizar conjuntos de dados e identificar características importantes sobre eles ou extrair respostas. Além de avaliar diferentes formas de visualizar os dados [26].

De acordo com Mukhiya and Ahmed [37], a estatística descritiva lida com a formulação de sumários de dados, de modo que eles possam ser claramente compreendidos. Ainda segundo Mukhiya and Ahmed [37], esses sumários podem ser representações numéricas ou visualizações com gráficos.

2.3.1 Medidas de tendência central e medidas de variabilidade

Há dois tipos de estatística descritiva [37]:

- Medidas de tendência central; e
- Medidas de variabilidade.

As medidas de tendência central buscam resumir valores em um único valor central [29] [24][37]. O objetivo dessas medidas é fornecer uma sumarização de todo o conjunto. O valor sempre é um valor que, de alguma forma, é central ao conjunto. As medidas de tendência central mais comuns são média aritmética, moda e mediana [37], mas existem outras.

A média aritmética de uma lista de números é, simplificadamente, o somatório de todos os números dessa lista, dividido pela quantidade de itens na lista [24][37]. Mais formalmente, a média aritmética $\bar{x}$, é dada por:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

onde:

- $\bar{x}$ representa a média aritmética;
- n é o número total de valores no conjunto;
- x_i representa cada valor individual da lista de números.

É possível ver como a média aritmética busca representar o conjunto de dados ao propor um valor central, por isso é uma medida de tendência central. Apesar disso, nenhuma medida é totalmente precisa na representação de um conjunto de dados. A média aritmética, por exemplo, é uma métrica que não é robusta a valores discrepantes, ao contrário da mediana [24]. E, por esse mesmo motivo, a mediana pode não ser a escolha adequada para representar os dados, caso a intenção seja refletir os valores discrepantes na medida de tendência central.

As medidas de variabilidade, ou medidas de dispersão, buscam quantificar a variação dos valores como, por exemplo, desvio padrão, valores máximos e mínimos [24][37]. Geralmente é usada em conjunto com as medidas de tendência central para obter uma descrição geral dos dados. As medidas de dispersão mais comuns são desvio padrão, variância, valores máximos e mínimos, curtose e assimetria [37]. É possível combinar as medidas de tendência central e de dispersão em uma única medida, criando uma medida composta ou híbrida [50][45].

A distância de uma observação para a sua média é chamada de desvio [24]. O desvio padrão descreve o quão distante da média está uma observação típica [24]. Matematicamente, o desvio padrão é definido como a raiz quadrada da variância, que, por sua vez, é a distância média quadrada de cada observação em relação à média [24].

2.3.2 Visualização de dados

A comunicação dos resultados é uma parte importante da estatística descritiva [30][49] e, por isso, é importante escolher as melhores opções de visualizações para comunicar o que se deseja. Há muitas alternativas de visualizações, mas nem todas são adequadas para representar os dados. De acordo com Wilke [49], uma visualização de dados precisa, em primeiro lugar,

comunicar acuradamente os dados. Mas, ainda assim, a questão estética é importante porque ajuda a transmitir a mensagem [49].

Há diversas formas de sumarizar dados através de visualizações. A visualização a ser escolhida depende do tipo das variáveis e do que se deseja comunicar, o que torna essa escolha uma etapa importante.

O sistema de coordenadas mais amplamente usado para visualização de dados é o Sistema de Coordenadas Cartesiano 2D [49]. Nesse sistema, as posições são especificadas por um valor *x* e um valor *it*, que são eixos ortogonais, conforme a Figura 2.1.

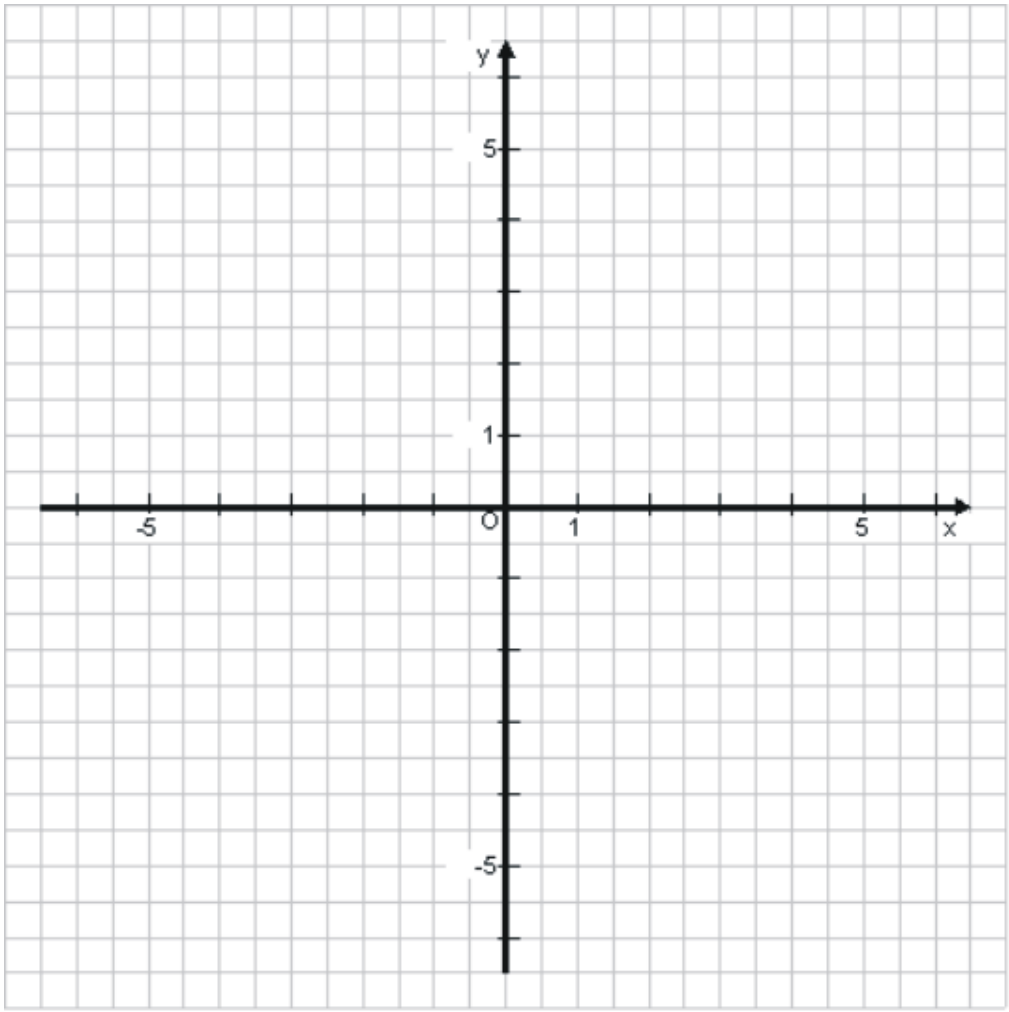

Figura 2.1: Sistema de coordenadas cartesiano 2D. Convencionalmente, o eixo horizontal é chamado eixo *x* e o eixo vertical é o eixo *y* [44]

Usando esse sistema de coordenadas, a maneira mais simples de checar o relacionamento entre um par de variáveis numéricas é o gráfico de dispersão [26]. Um exemplo de gráfico de dispersão está ilustrado na Figura 2.2. Nela, que analisa dados de árvores fictícias, apenas para fins de ilustração, podemos ver que há uma correlação entre o diâmetro da árvore e a altura da árvore, isto é, árvores com diâmetro maior tendem a ser mais altas.

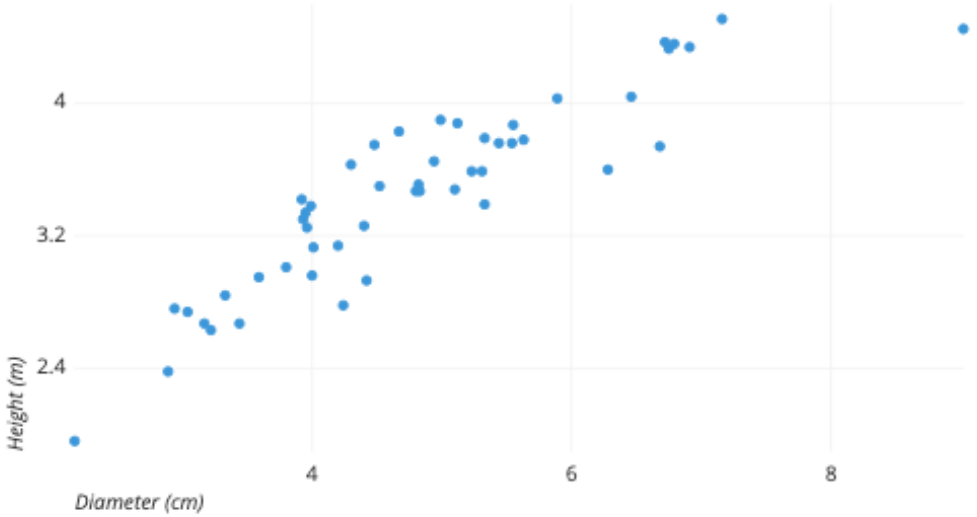

Figura 2.2: Gráfico de dispersão. Com uma amostra de dados de árvores fictícias, apenas para fins de ilustração. O eixo x representa o diâmetro em centímetors e o eixo y representa a altura da árvore em metros [52]

Uma das melhores formas de descrever uma variável numérica é relatar os valores que aparecem no conjunto de dados e a frequência com que eles aparecem. Essa descrição é chamada de distribuição da variável [26]. A forma mais comum de representar uma distribuição são os histogramas [26]. Esse tipo de gráfico permite visualizar a densidade dos dados [24]. Barras mais altas representam em que intervalos os dados são mais comuns.

A Figura 2.3 apresenta um histograma de taxa de juros. Nele, é possível ver quais intervalos de valores são mais frequentes. Essa representação mostra como os dados estão distribuídos. É possível verificar que a maior parte das observações no conjunto de dados são de taxas de juros entre 5% e 10%.

O histograma da Figura 2.3 é um histograma assimétrico à direita, porque a sua "cauda" é mais longa para a direita. Caso a "cauda" fosse mais longa para o lado esquerdo, o histograma seria dito assimétrico à esquerda [24]. Caso as "caudas" sejam semelhantes para cada um dos lados, o histograma é dito ser simétrico [24].

Mas, além de identificar se um histograma é simétrico ou assimétrico, podemos identificar se um histograma é modal, bimodal ou multimodal ao observar os seus picos [24]. Uma distribuição com apenas um pico proeminente, é chamada unimodal; uma distribuição com dois picos proeminentes, por sua vez, é chamada bimodal; e uma distribuição com mais de dois picos proeminentes, é chamada multimodal. A Figura 2.4 ilustra esses três tipos.

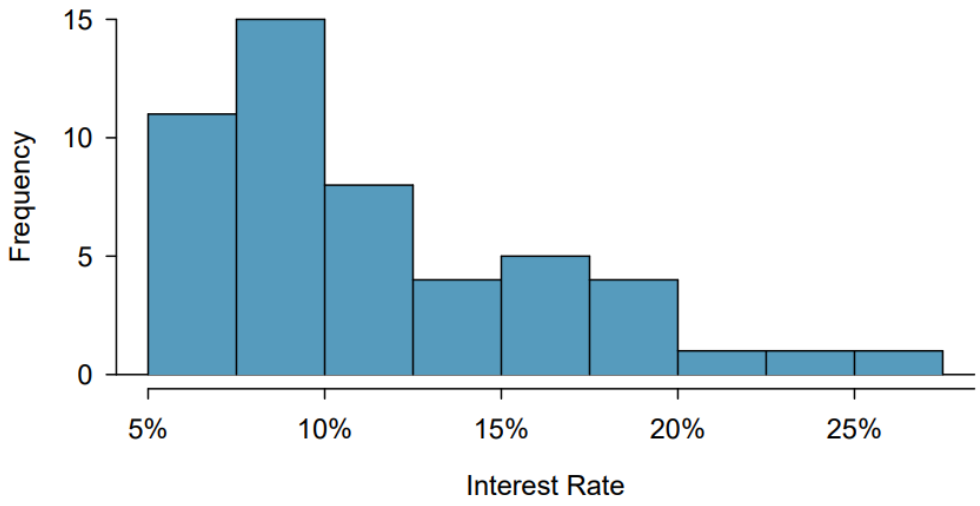

Figura 2.3: Histograma de taxa de juros. No eixo x temos a frequência e no eixo y temos os intervalos das taxas de juros. Este histograma é assimétrico à direita [24]

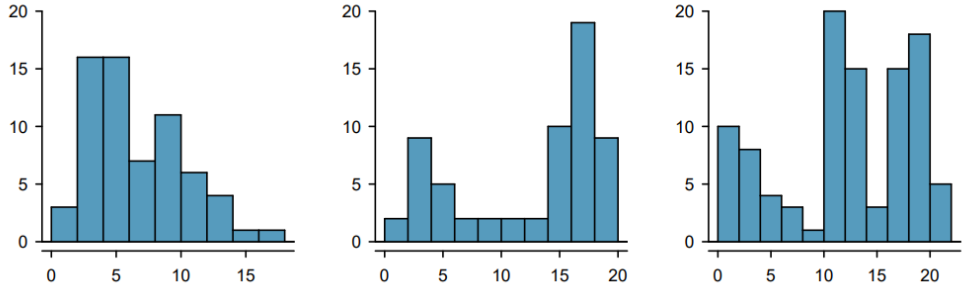

Figura 2.4: Contando apenas picos proeminentes, tem-se, da esquerda para a direita: unimodal, bimodal e multimodal [24]

Caso ao invés de dados numéricos, deseje-se representar dados categórico ou qualitativos, há algumas alternativas de visualizações. Além de tabelas, uma das formas mais comuns de visualização de dados categóricos são os gráficos de barras [24]. Gráficos de barras têm a vantagem de serem muito comuns e, por isso, são mais fáceis de compreender para a audiência a que se deseja comunicar os resultados [30].

A Figura 2.5 ilustra um gráfico de barras para uma única variável categórica. No entanto, é possível projetar gráficos de barras para mais de uma variável, usando o gráfico de barras empilhadas ou o gráfico de barras lado a lado, ilustrados na Figura 2.6.

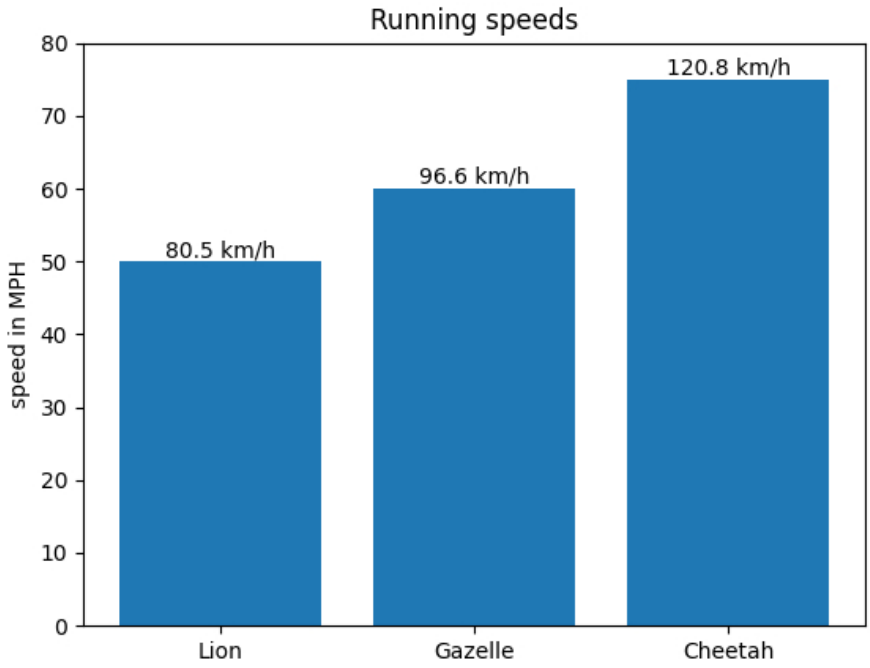

Figura 2.5: Gráfico de velocidade de alguns animais. Da esquerda para a direita: leão, gazela, guepardo [24]

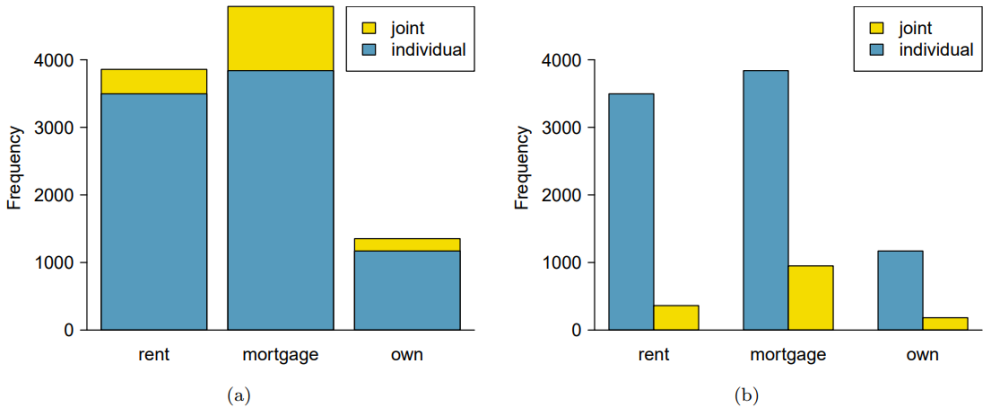

Figura 2.6: (a) Gráfico de barras empilhadas. (b) Gráfico de barras lado a lado [24]

Uma outra forma de visualização de dados categóricos popular são os gráficos de pizza. No entanto, de maneira geral, esse tipo de visualização não é recomendada [24][30][46]. A Figura 2.7 ilustra o porquê de ser um gráfico a ser evitado. Segundo Knaflic [30], quando os segmentos possuem tamanhos parecidos, é difícil determinar qual é maior; e, mesmo quando eles não possuem tamanhos parecidos, é difícil determinar o quanto um segmento é maior que o outro. A Figura 2.8 apresenta uma alternativa, com gráfico de barras, ao gráfico de pizza da Figura 2.7.

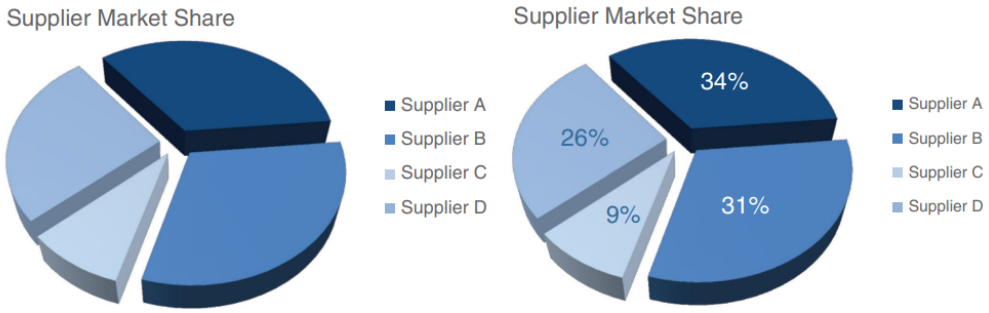

Figura 2.7: (a) Gráfico de pizza sem as anotações. Pode ser difícil estimar qual segmento é maior. (b) Gráfico de pizza com anotações. As anotações ajudam a compreender qual o maior segmento, mas, ainda assim, pode ser difícil discernir o quão maior [30]

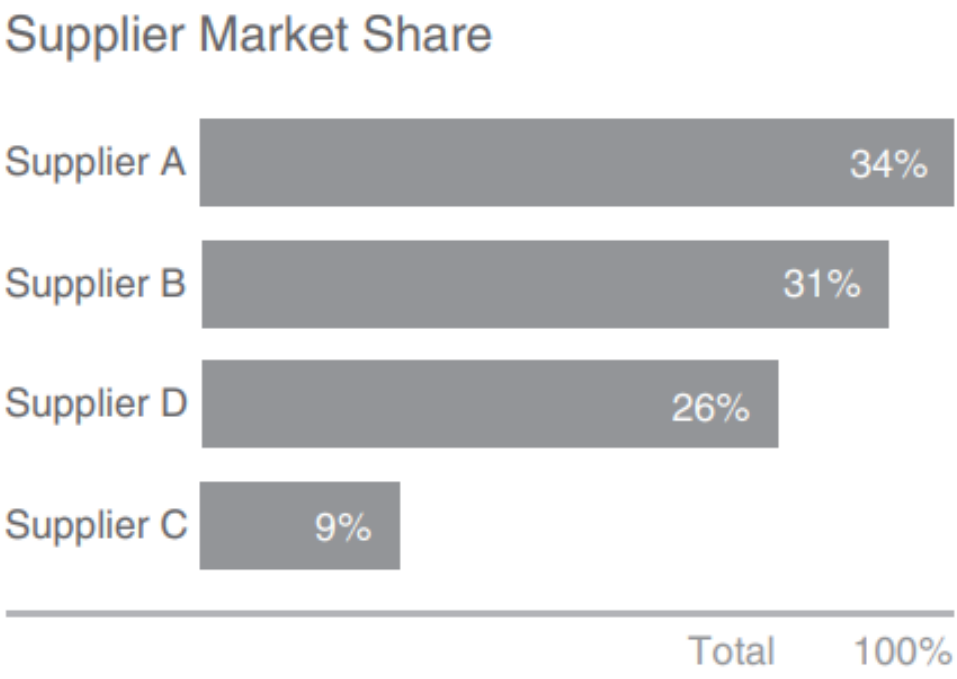

Figura 2.8: Gráficos de barra são mais fáceis de visualizar qual segmento é maior e também a visualizar o quão maior [30]

2.3.3 Análise Exploratória de Dados

A análise exploratória de dados é diferente da estatística descritiva [30][37]. Na análise exploratória, o foco é entender os dados, descobrir e identificar algo que seja considerado interessante [30]. Como por exemplo, buscar padrões e relacionamentos. Para isso, algumas das visualizações já discutidas neste trabalho podem ser empregadas, mas com um propósito diferente. Na análise exploratória, as visualizações são um meio para descobrir algo ao invés de uma ferramenta para apresentar algo [51].

2.4 Análise de Dados no futebol

No que tange à análise de dados aplicada especificamente ao futebol, nota-se que esse domínio de aplicação tem crescido nos últimos anos [35]. Esse crescimento também implicou em um aumento no número de fontes de informação e infraestrutura de análise de dados no contexto do futebol [35]; e também em um aumento no número de estudos relacionados[43]. Tradicionalmente, esportes como futebol não eram analisados com grande foco em dados, métricas e desempenho, pois os jogos eram vistos como oportunidades de demonstrar arte e habilidades individuais [23]. Essa é uma tendência relativamente recente [23][20].

No entanto, medir desempenho de "jogos de bola"(futebol, futebol americano, handebol, etc.) pode ser desafiador, pois nem sempre o time com o melhor resultado foi o time com o melhor desempenho [23]. De acordo com a literatura, "desempenho", no contexto do futebol, pode ser definido como a interação de diferentes fatores táticos, técnicos, mentais e físicos [43].

Uma das formas mais amplamente usadas de se medir esse desempenho, é através da análise notacional [23]. A análise notacional consiste em registrar eventos a fim de que haja registros precisos e objetivos do que aconteceu [23]. Expectadores veem partidas de maneiras diferentes e frequentemente discordam sobre o que aconteceu em determinada partida, especialmente quando a partida já aconteceu há mais tempo. A análise notacional oferece registros mais precisos do que aconteceu de fato, desde que o método de coleta seja confiável [23]. Algumas das métricas que a análise notacional registram são: chutes, passes, dribles, defesas, botes defensivos, interceptações, faltas, entre outras [23].

Além disso, há métricas baseadas em modelos probabilísticos através do uso de aprendizagem de máquina, como por exemplo a métrica *Expected Goals* (xG). Essa métrica avalia a probabilidade de um chute se converter em gol [42]. Essa métrica pode assumir qualquer valor entre 0 e 1, onde 0 representa uma finalização sem nenhuma possibilidade de ser gol e 1 representa uma finalização que é gol certo [36]. Por exemplo, uma finalização com xG de 0,2 significa que, dada as condições da finalização, apenas 2 a cada 10 finalizações resultam em gol. Expected Goals usa aprendizagem de máquina para calcular, levando em consideração algumas características para estimar a probabilidade de uma finalização ser gol [18]. Algumas dessas características são: posição do chute, velocidade do jogador chutando, defensores na frente do chute, posição do goleiro, pressão no jogador chutando, tipo do chute (cabeceio, perna, etc.), entre outras [18]. Essa métrica é bem estabelecida na comunidade de análises esportivas [42].

Derivada do xG, há outra métrica denominada Gols Esperados Pós-Chute (PSxG)[42], que também é chamada de xGOT (Gols Esperados No Alvo). Essa métrica considera o xG apenas após o chute ser realizado e considera apenas chutes ao alvo, chutes para fora assumem o valor 0.

3

Metodologia

Neste capítulo, será abordada a metodologia para o desenvolvimento do trabalho. Primeiro, será descrita a forma como os dados foram coletados e os processos de transformação dos dados. Depois, será descrita a forma como os gráficos e resultados foram gerados. As ferramentas utilizadas para cada etapa também serão descritas.

3.1 Coleta e tratamento de dados

Os dados utilizados foram coletados do FBref.com, um *website* especializado em estatísticas de futebol para times e jogadores de diversos países do mundo. FBref foi criado pela Sports Reference, que também fornece dados de outros esportes, como *Baseball* e Futebol Americano [9].

Para coletar esses dados, foi usada a linguagem de programação *Python* [47] e a biblioteca *Pandas* [39]. A função usada foi a *pandas.read_html*, que, dada uma *string* que representa uma *URL*, são buscadas todas as tabelas no endereço fornecido [4]. O retorno dessa função é uma lista de *pandas dataframes*, que é uma estrutura de dados para dados tabulares[1]. Com essa lista em mãos, foram selecionados os *dataframes* através dos índices, que foram identificados manualmente. Cada índice representa um dataframe, ordenados de acordo com o *html* da *URL* fornecida. Os índices selecionados foram 0, 2, 4, 6, 8, 10, 12, 14, 16, 18 e 22; e eles representam, respectivamente: classificação (*ranking*) da temporada, dados gerais das equipes, dados de goleiro, dados avançados de goleiro, dados de finalizações, dados de passes, dados de tipos de passes, dados de gols e criação de chute, dados de ações defensiva, dados de posse de bola, dados variados. As métricas relacionadas a cada uma desses 11 índices está resumida na tabela 3.1.

Foram utilizados dados das últimas cinco temporadas completas para cada uma das ligas. As ligas escolhidas foram: Campeonato Brasileiro (Brasil), Premier League (Inglaterra), La

Tabelas importadas
Classificação
Equipes
Goleiro
Goleiro (métricas avançadas)
Finalizações
Passes
Tipos de passes
Gols e criação de chutes
Ações defensivas
Posse de bola
Variados

Tabela 3.1: As 11 tabelas importadas possuem dados relacionados a esses itens

Liga (Espanha), Bundesliga (Alemanha), Série A (Itália). Para as ligas europeias, foram usados dados das temporadas 2019-2020, 2020-2021, 2021-2022, 2022-2023 e 2023-2024. Devido ao calendário das temporadas brasileira funcionar de maneira diferente, foram usados dados das temporadas 2019, 2020, 2021, 2022 e 2023.

Inicialmente, foram importadas 11 tabelas contendo variáveis relativas a diversas métricas de desempenho relevantes no esporte, resumidas na tabela 3.1. Essas 11 tabelas foram importadas para cada uma das 5 ligas, totalizando 55 tabelas por temporada. Isso resultou em um total de 275 tabelas importadas (nas 5 temporadas). Somando essas tabelas, foram importadas 5500 linhas e 6175 colunas.

No entanto, para facilitar o desenvolvimento do trabalho, foi realizada uma etapa de limpeza dos dados. Algumas colunas foram removidas dos dados, reduzindo o número total de colunas entre os 275 arquivos de 6175 para 4525 colunas. Os principais dados removidos foram dados que, após uma análise exploratória preliminar, certamente não seriam usados para as análises subsequentes. Também foram removidos dados redundantes, como o número de partidas disputadas, visto que foram usados apenas campeonatos encerrados, o número de partidas jogadas sempre era 38. O número de linhas não foi alterado em nenhum conjunto de dados.

Durante essas etapas, também foram verificados dados faltantes e dados discrepantes (para identificar possíveis erros de preenchimento); e foi checado se a importação considerou corretamente pontos flutuantes. Erros de ponto flutuante não foram encontrados na base de dados em inglês; e também não foram identificados valores anômalos. Em relação a dados faltantes, apenas a variável *Notes* (Notas ou observações) possuía dados faltantes, mas essa coluna foi removida por não ser relevante para a análise. Não foi necessário realizar reduções ou transformações nos dados.

Todas essas tabelas, após essas etapas, foram disponibilizadas em formato de arquivo *csv* em um repositório online [8]. Da Tabela 3.3 à Tabela 3.13, é possível verificar as descrições das variáveis das tabelas disponibilizadas com breves descrições das variáveis. Essas tabelas

# Pl	GA	PKA	FK	CK	OG	PSxG	PSxG/SoT	PsxG+/-	/90	Country
2.5	45.13	4.84	0.95	5.52	1.11	48.05	0.27	4.03	0.11	BR
2.0	54.59	4.28	0.87	6.91	1.98	50.18	0.28	-2.43	-0.07	EN
2.2	48.06	4.91	0.78	5.15	1.33	46.77	0.28	0.04	0.001	SP
2.0	53.55	4.2	0.9	5.96	1.48	49.63	0.30	-2.45	-0.07	GE
2.48	53.69	5.76	1.05	5.93	1.58	49.38	0.26	-2.73	-0.07	IT

Tabela 3.2: Tabela resumida com dados médios de goleiro ao longo das cinco temporadas em cada país (valores arredondados)

disponibilizadas no repositório são após as reduções citadas anteriormente, mas sem tratamentos subsequentes. A estrutura de cada tabela é a mesma para cada liga em cada temporada analisada.

Também foram gerados arquivos *csv* contendo os dados de todas as equipes, em todas as ligas, em todas as cinco temporadas para cada conjunto de métricas, conforme descrito na Tabela 3.1. Esses arquivos não possuem dados médios, mas todas as observações coletadas na fonte original. Para a criação desses arquivos, foram importados os dados para cada temporada e eles foram concatenados em um único *dataframe*, acrescentando apenas colunas referentes à temporada e ao país. A estutura de colunas é semelhante às tabelas da Tabela 3.3 à Tabela 3.13, mas contêm também colunas extras indicando a liga e a temporada corresponentes à observação. Esses dados completos também foram disponibilizados no repositório.

3.2 Cálculos de medidas e criação de visualizações

Após essa etapa de importação e etapas iniciais de redução nos dados, foram gerados novos conjuntos de dados, contendo apenas as médias, ao longo das cinco temporadas, para cada país. A tabela 3.2 mostra, usando os dados de goleiros, como essas tabelas resumidas ficaram; o mesmo foi feito para as 11 tabelas resumidas na tabela 3.1, sempre usando os dados das 5 temporadas. Para calcular as médias, foi usada a média aritmética e o método *mean()* [2] da classe *dataframe* da biblioteca Pandas, que calcula automaticamente as médias das colunas.

Também foram calculadas as médias considerando apenas os quatro clubes de melhor classificação da temporada e calculadas as médias considerando apenas os quatro piores clubes de cada temporada. Para isso, foi usada a tabela de classificação de cada ano para identificar as equipes de melhor ou pior classificação em cada ano; e, em seguida, foram extraídos apenas os dados desses clubes. Essas etapas de cálculo prévio das médias foi adotada para facilitar a criação das visualizações posteriormente.

Todos os gráficos foram criados usando as bibliotecas Matplotlib [28] e Seaborn [48]. Foram escolhidos gráficos de barras, devido à simplicidade e poder de síntese e comparação, como discutido no capítulo Fundamenteção Teórica. Para a criação dos gráficos, foram importados os dados contendo as médias de cada país e selecionadas apenas as colunas relevantes para visu-

alizar as variáveis de interesse, conforme descritas no capítulo Introdução, na seção Objetivo. Os gráficos serão apresentados no capítulo Resultados.

Por fim, a partir dos dados completos importados, tratados e disponibilizados, foram gerados histogramas para analisar variabilidade de cada liga em cada uma das métricas avaliadas. Além dos histogramas, foi calculado o desvio padrão para cada métrica em cada liga, a fim de complementar a visualização de variabilidade dos histogramas. O cálculo dos histogramas foi feito automaticamente através do método *std* [3] da biblioteca Pandas.

Esses histogramas também foram criados usando as bibliotecas Matplotlib e Seaborn. Histogramas foram escolhidos para analisar a variabilidade porque esse tipo de visualização permite analisar a distribuição, simetria dos dados e valores mínimos e máximos em uma única visualização.

Variável	**Descrição**	**Tipo de variável**
Rk	Classificação	Numérica
Squad	Nome da equipe	Categórica
MP	Partidas disputadas	Numérica
W	Vitórias	Numérica
D	Empates	Numérica
L	Derrotas	Numérica
CS	Partidas sem tomar gol	Numérica
CS%	Percentual de partidas sem tomar gol	Numérica
GF	Gols marcados	Numérica
GA	Gols tomados	Numérica
GD	Saldo de gols	Numérica
Pts	Pontos	Numérica
Pts/MP	Pontos por partida	Numérica
xG	Gols esperados	Numérica
xGA	Gols tomados esperados	Numérica
xGD	Saldo de gols esperado	Numérica
xGD/90	Saldo de gols esperado a cada 90 minutos	Numérica
Attendance	Média de público	Numérica

Tabela 3.3: Descrição das variáveis da tabela de classificação do campeonato

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
# Pl	Número de jogadores	Numérica
Age	Idade média	Numérica
Poss	Posse de bola	Numérica
Gls	Gols	Numérica
Ast	Assistências	Numérica
G+A	Gols + Assistências	Numérica
G-PK	Gols exceto de pênalti	Numérica
PK	Gols de pênalti	Numérica
PKatt	Pênaltis cobrados	Numérica
CrdY	Cartões amarelos	Numérica
CrdR	Cartões vermelhos	Numérica
xG	Gols esperados	Numérica
npxG	Gols esperados sem pênaltis	Numérica
xAG	Gols assistidos esperados	Numérica
npxG+xAG	Gols esperados sem pênaltis mais gols assistidos esperados	Numérica
PrgC	Carregadas progressivas	Numérica
PrgP	Passes progressivos	Numérica

Tabela 3.4: Descrição de variáveis da tabela de dados do elenco

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
# Pl	Número de jogadores na posição	Numérica
GA	Gols tomados	Numérica
GA90	Gols tomados a cada 90 minutos	Numérica
SoTA	Chutes no alvo recebidos	Numérica
Saves	Defesas.	Numérica
Save%	Percentual de defesas por chutes no alvo recebidos	Numérica
W	Vitórias	Numérica
D	Empates	Numérica
L	Derrotas	Numérica
CS	Partidas sem tomar gol	Numérica
CS%	Percentual de partidas sem tomar gol	Numérica

Tabela 3.5: Descrição de variáveis da tabela de dados de goleiro

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
# Pl	Número de jogadores na posição	Numérica
GA	Gols tomados.	Numérica
PKA	Penâltis permitidos.	Numérica
FK	Gols de falta tomados	Numérica
CK	Gols de escanteio tomados	Numérica
OG	Gols-contra tomados	Numérica
PSxG	Gols esperados pós-chute	Numérica
PSxG/SoT	Gols esperados pós-chute por chute no gol	Numérica
PSxG+/-	Gols esperados pós-chute menos gols permitidos	Numérica
/90	Gols esperados pós-chute menos gols permitidos a cada 90 minutos	Numérica

Tabela 3.6: Descrição de variáveis da tabela de dados avançados de goleiro

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
Gls	Gols	Numérica
Sh	Finalizações	Numérica
SoT	Finalizações no gol	Numérica
SoT%	Percentual de finalizações no gol	Numérica
Sh/90	Finalizações a cada 90 minutos	Numérica
G/Sh	Gols por finalização	Numérica
G/SoT	Gols por finalização no gol	Numérica
Dist	Distância média das finalizações	Numérica
FK	Finalizações de cobranças de falta	Numérica
PK	Pênaltis convertidos	Numérica
PKatt	Pênaltis cobrados	Numérica
xG	Gols esperados	Numérica
npxG	Gols esperados sem pênalti	Numérica
npxG/Sh	Gols esperados sem pênalti por chute	Numérica
G -xG	Gols menos gols esperados	Numérica
np:G-xG	Gols exceto de pênalti menos gols esperados sem pênaltis	Numérica

Tabela 3.7: Descrição das variáveis dos dados de finalizações

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
Cmp	Passes completos	Numérica
Att	Passes tentados	Numérica
Cmp%	Percentual de passes completos	Numérica
TotDist	Distância total (em jardas) dos passes em qualquer direção	Numérica
PrgDist	Distância total (em jardas) de passes progressivos	Numérica
Ast	Assistências	Numérica
xAG	Gols assistidos esperados	Numérica
xA	Assistências esperadas	Numérica
A-xAG	Assistências menos gols assistidos esperados	Numérica
KP	Passes-chave (que resultam em uma finalização)	Numérica
1/3	Passes no terço final do campo	Numérica
PPA	Passes na área adversária	Numérica
CrsPA	Cruzamentos para a área adversária	Numérica
PrgP	Passes progressivos	Numérica

Tabela 3.8: Descrição das variáveis da tabela de dados de passes

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
Att	Passes tentados.	Numérica
Live	Passes com bola rolando	Numérica
Dead	Passes de bola parada	Numérica
FK	Passes de cobranças de falta	Numérica
TB	Passes que atravessam linhas	Numérica
Sw	Inversões	Numérica
Crs	Cruzamentos	Numérica
TI	Cobranças de lateral	Numérica
CK	Escanteios	Numérica
In	Escanteios com efeito para dentro	Numérica
Out	Escanteios com efeito para fora	Numérica
Str	Escanteios retos	Numérica
Cmp	Passes completos	Numérica
Off	Passes impedidos	Numérica
Blocks	Passes bloqueados	Numérica

Tabela 3.9: Descrição das variáveis da tabela de dados de tipos de passes

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
Tkl	Botes defensivos	Numérica
TklW	Botes vencidos	Numérica
Def 3rd	Botes no terço de defesa	Numérica
Mid 3rd	Botes no terço intermediário	Numérica
Att 3rd	Botes no terço de ataque	Numérica
Tkl	Dribles que receberam botes	Numérica
Att	Dribles em que houve tentativa de bote	Numérica
Tkl%	Percentual de botes que superaram dribles	Numérica
Lost	Botes que não impediram o drible	Numérica
Blocks	Bloqueios de bola ao ficar em sua trajetória	Numérica
Sh	Chutes bloqueados	Numérica
Pass	Passes bloqueados	Numérica
Int	Interceptações	Numérica
Tlk+Int	Botes mais interceptações	Numérica
Clr	Afastamentos de bolas próximas ao gol	Numérica
Err	Erros defensivos que resultaram em uma finalização adversária	Numérica

Tabela 3.10: Descrição das variáveis da tabela de dados de ações defensivas

Variável	Descrição	Tipo de variável
Squad	Nome da equipe	Categórica
SCA	Ações de criação de chutes	Numérica
SCA90	Ações de criação de chutes a cada 90 min	Numérica
PassLive	Passes com bola rolando que resultaram em chute	Numérica
PassDead	Passes de bola parada que resultaram em chute	Numérica
TO	Dribles que resultaram em chutes	Numérica
Sh	Chutes que levaram a outro chute	Numérica
Fld	Faltas recebidas que levaram a um chute	Numérica
Def	Ações defensivas que resultaram em um chute	Numérica
GCA	Ações de criação de gols	Numérica
GCA90	Ações de criação de gols a cada 90 minutos	Numérica
PassLive	Passes com bola rolando que resultaram em gol	Numérica
PassDead	Passes de bola parada que resultaram em gol	Numérica
TO	Dribles que resultaram em gol	Numérica
Sh	Chutes que levaram resultaram em outra finalização que foi gol	Numérica
Fld	Faltas recebidas que levaram a um gol	Numérica
Def	Ações defensivas que resultaram em gol	Numérica

Tabela 3.11: Descrição das variáveis da tabela de dados de gols e chutes

Variável	**Descrição**	**Tipo de variável**
Squad	Nome da equipe	Categórica
Poss	Média de posse	Numérica
Touches	Toques na bola	Numérica
Def Pen	Toques na própria área	Numérica
Def 3rd	Toques no terço de defesa	Numérica
Mid 3rd	Toques no terço intermediário	Numérica
Att 3rd	Toques no terço de ataque	Numérica
Att Pen	Toques na área adversária	Numérica
Live	Toques com bola rolando	Numérica
Att	Dribles tentados	Numérica
Succ	Dribles bem sucedidos	Numérica
Succ%	Percentual de dribles bem sucedidos	Numérica
Tlkd	Botes recebidos durante tentativa de drible	Numérica
Tlkd%	Botes bem sucedidos recebidos durante tentativa de drible	Numérica
Carries	Carregadas de bola	Numérica
TotDist	Distância percorrida carregando a bola	Numérica
PrgDist	Distância progredida carregando a bola	Numérica
PrgC	Carregadas progressivas	Numérica
1/3	Carregadas que entram no terço de ataque	Numérica
CPA	Carregas que entram na área adversária	Numérica
Mis	Perda de controle da bola	Numérica
Dis	Perda de posse por desarme	Numérica
Rec	Passes recebidos	Numérica
PrgR	Passes progressivos recebidos	Numérica

Tabela 3.12: Descrição dos dados da tabela de posse de bola

Variável	**Descrição**	**Tipo de variável**
Squad	Nome da equipe	Categórica
CrdY	Cartões amarelos	Numérica
CrdR	Cartões vermelhos	Numérica
2CrdY	Segundos cartões amarelos	Numérica
Fls	Faltas cometidas	Numérica
Fld	Faltas recebidas	Numérica
Off	Impedimentos	Numérica
Crs	Cruzamentos	Numérica
Int	Interceptações	Numérica
TlkW	Botes defensivos vencidos	Numérica
PKWon	Pênaltis recebidos	Numérica
PKcon	Pênaltis concedidos	Numérica
OG	Gols-contra	Numérica
Recov	Bolas recuperadas	Numérica
Won	Duelos aéreos vencidos	Numérica
Lost	Duelos aéreos perdidos	Numérica
Won%	Percentual de duelos aéreos vencidos	Numérica

Tabela 3.13: Descrição das variáveis da tabela de dados variados

4

Resultados e Discussão

Este capítulo aborda os resultados obtidos. O foco é, sobretudo, apresentar visualizações relevantes sobre algumas métricas de desemepenho, comparando os resultados entre o campeonato brasileiro e as quatro maiores ligas europeias. A partir das visualizações, serão feitos comentários a respeito dos dados e propostas interpretações para eles.

4.1 Métricas de Desempenho

Como já discutido anteriormente, as métricas escolhidas para análise foram: gols, gols esperados (xG), gols esperados sem pênalti (npxG), total de chutes, chutes a gol, gols por chute, defesas de goleiro, percentual de defesas por chutes a gol, gols esperados pós chute menos gols permitidos (psxG-GA), número de passes concluídos, percentual de passes completos, botes defensivos, confrontos no terço de ataque, erros defensivos que resultaram em uma tentativa de chute do time adversário, número de dribles tentados, dribles certos, dribles que resultaram em uma finalização, cartões (amarelos mais vermelhos), faltas cometidas.

Essas métricas foram separadas três a três, para manter a visualização clara ao mesmo tempo em que reúne algumas métricas relacionadas. Cada uma das subseções a seguir diz respeito a um trio de métricas.

4.1.1 Gols (Gls), Gols Esperados (xG) e Gols Esperados Sem Pênalti (npxG)

As primeiras métricas escolhidas para análise foram métricas relacionadas com o objetivo do jogo: gols. A métrica de "Gols" toma os valores totais de gols a favor das equipes, incluindo apenas gols válidos, mesmo que de pênalti. A métrica de "Gols Esperados", também chamada de xG, como já discutido anteriormente, é uma métrica avançada que quantifica a probabilidade de uma finalização ser gol; ela pode assumir valores entre 0 (sem possibilidade) e 1 (gol certo).

Campeonato	Gols	xG	Diferença	Diferença %
Brasileiro	44	48.5	-4.5	9.73%
Inglês	52.6	53	-0.4	0.77%
Espanhol	46.7	47.9	-1.2	2.54%
Alemão	52.1	50.3	+1.8	3.52%
Italiano	52.1	51.2	+0.9	1.74%

Tabela 4.1: Comparativo entre gols marcados e gols esperados

A métrica de "Gols Esperados Sem Pênalti" (npxG) é semelhante ao xG, mas desconsidera cobranças de pênaltis.

A Figura 4.1 apresenta os gráficos com essas métricas. É possível notar que o campeonato brasileiro é o campeonato que menos faz gols, em média, mas não é o campeonato com menos gols esperados. Isso pode significar que o campeonato brasileiro é um campeonato em que há muitas chances de gols, mas a conversão dessas chances está aquém das expectativas. Apesar de que apenas o campeonato alemão apresenta uma diferença positiva entre gols e gols esperados (isto é, a quantidade de gols marcados está além da expectativa), o campeonato brasileiro é o que apresenta a maior diferença negativa. A tabela 4.1 traz um breve resumo dessas diferenças.

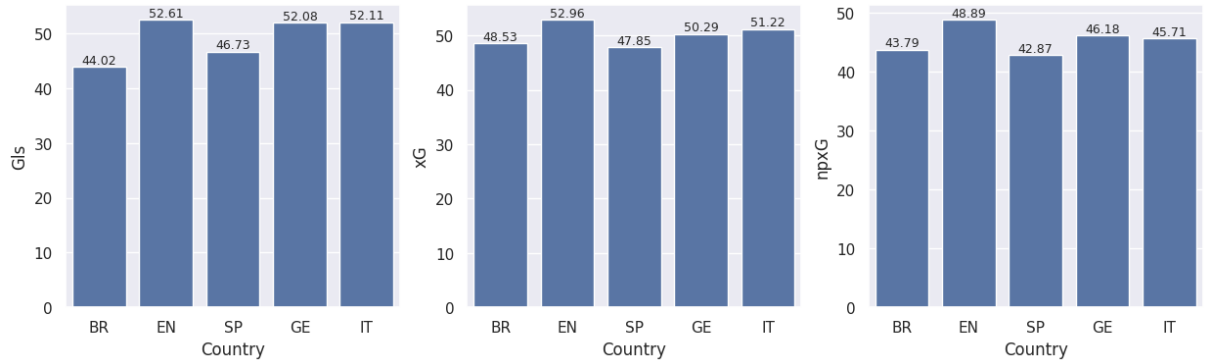

Figura 4.1: Gols a favor (Gls), gols esperados (xG) e gols esperados sem pênalti (npxG)

Uma diferença negativa indica ineficiência ao finalizar. Baseado nesses dados, é possível observar que o campeonato brasileiro, de fato, é menos eficiente ao converter suas chances criadas em gols. Isso pode indicar um desempenho acima do esperado por parte dos goleiros adversários ou baixa qualidade das finalizações dos atacantes. Um xG mais alto também pode ser indicativo de um maior volume de finalizações, mesmo que sem qualidade, que acabam somando ao xG, sem necessariamente impactar os gols marcados de fato.

A Figura 4.2 compara as mesmas métricas de desempenho, mas, desta vez, considerando apenas os dados das quatro equipes de melhor classificação em cada ano e as quatro equipes de pior classificação em cada ano. É possível observar que o campeonato brasileiro apresenta uma menor discrepância entre os dados das equipes do topo e as equipes do fundo da classificação. Essa diferença menor indica que há um equilíbrio maior dentro do próprio campeonato nessas

métricas de desempenho. É possível notar também que se considerar o xG apenas das equipes do fundo, o campeonato brasileiro é o que apresenta o xG mais alto, empatado com a Premier League.

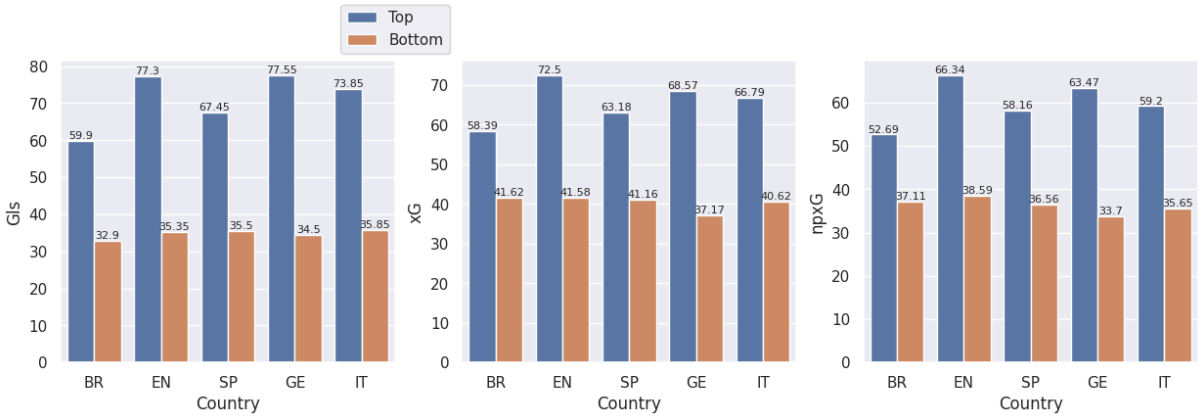

Figura 4.2: Gols a favor (Gls), gols esperados (xG) e gols esperados sem pênalti (npxG). Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano

4.1.2 Total de chutes (Sh), Chutes a gol (SoT) e Gols por chute (G/Sh)

Em relação ao número de chutes, pela Figura 4.3, é possível verificar que o campeonato brasileiro tem um volume de chutes maior do que as ligas europeias. No entanto, como vimos na Figura 4.1, isso não se traduz em um maior número de gols e também não significa que melhores chances foram criadas. A métrica de gols por chute do campeonato brasileiro é a menor entre as ligas analisadas, o que pode indicar um desempenho além do esperado dos goleiros ou uma ineficiência dos finalizadores.

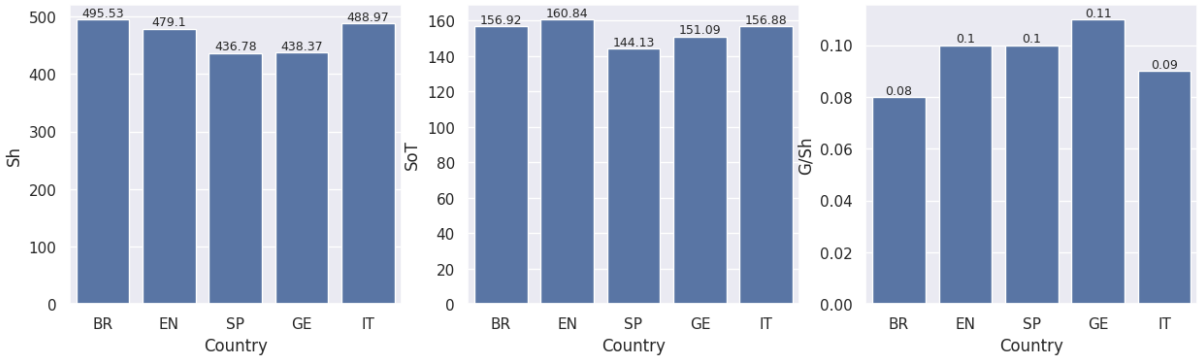

Figura 4.3: Chutes (Sh), chutes a gol (SoT) e gols por chute (G/Sh)

Nos dados que consideram apenas as equipes de melhor e de pior classificação, resumidos na Figura 4.4, é possível verificar que a diferença interna no campeonato brasileiro é a menor na métrica de chutes. No entanto, a diferença cresce quando visualizamos a métrica de Gols por

chute, o que pode ser reflexo também do que pode ser visto na Figura 4.2, em que as melhores equipes são muito mais eficientes em converter chances. Esse mesmo padrão identificado na liga brasileira se repete em todas as ligas, exceto a inglesa; na liga inglesa, a diferença de G/Sh é menor do que a diferença do total de Chutes. Isto é, na liga inglesa, em particular, a diferença de G/Sh entre as melhores e as piores equipes do campeonato é menor.

Ainda pela Figura 4.4, é possível verificar que, entre as piores equipes de cada liga, as equipes da liga brasileira são as que mais finalizam, bem como possuem o maior xG, pela Figura 4.2. Isso revela uma maior ineficiência de seus finalizadores, visto que apresentam o menor número de gols totais e o menor número de gols por chute, mesmo sendo as que mais finalizam.

Figura 4.4: Chutes (Sh), chutes a gol (SoT) e gols por chute (G/Sh). Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano

4.1.3 Análise de goleiros: defesas totais (Saves), percentual de defesas (Save%), Gols Esperados Pós-Chute Menos Gols Permitidos (PSxG-GA)

Agora, serão analisadas algumas métricas relativas ao desempenho dos goleiros. Defesas totais é uma variável numérica discreta que não leva em consideração a qualidade das finalizações ou a dificuldade das defesas; enquanto que o percentual de defesas considera o percentual de chutes a gol defendidos. PSxG-GA toma os valores de gols esperados pós-chute, que foi discutido na seção de Fundamentação Teórica, e subtrai pela quantidade de gols permitidos de fato. Um PSxG-GA positivo significa que o goleiro permitiu menos gols do que era esperado, enquanto que um PSxG negativo significa que o goleiro permitiu mais gols do que era esperado.

A Figura 4.5 apresenta os dados dessas métricas. É possível verificar que o campeonato brasileiro possui um maior número de defesas totais. Isso poderia ser considerado como natural devido ao alto volume de chutes totais, mas a liga brasileira lidera também o percentual de

defesas e a métrica PSxG-GA, que avalia mais especificamente os gols evitados pelos goleiros. É possível notar um alto desempenho dos goleiros do campeonato brasileiro. Nas ligas inglesa, alemã e italiana, os goleiros permitem mais gols do que o esperado.

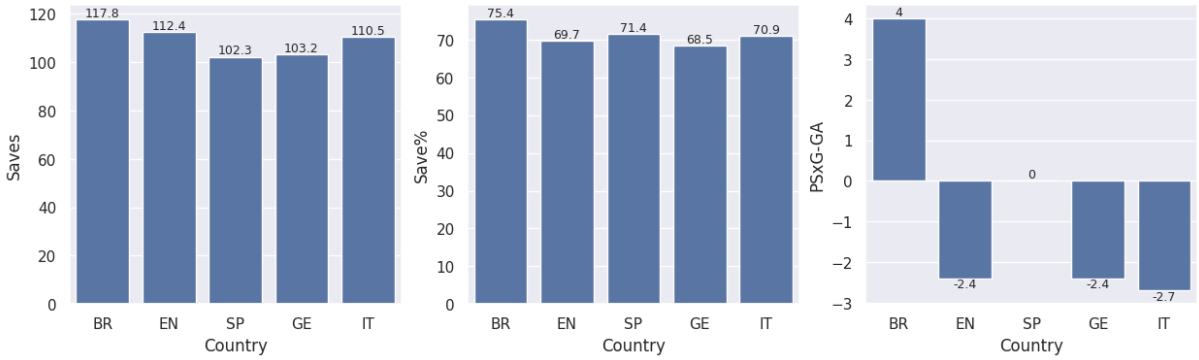

Figura 4.5: Defesas totais (Saves), percentual de defesas (Save%) e gols esperados pós-chute menos gols permitidos (PSxG-GA). Os goleiros da liga brasileira apresentam um bom desempenho nessas métricas

A Figura 4.6 apresenta os dados comparando apenas as equipes de melhor classificação e as de pior classificação em cada ano. Naturalmente, por receberem mais finalizações contra, as equipes de pior classificação fazem mais defesas, mas o percentual de defesas sempre fica abaixo das equipes de melhor classificação, bem como o PSxG-GA, que é sempre bem discrepante, demostrando que os goleiros das melhores equipes de fato evitam mais gols do que o esperado. Pode-se destacar também que a liga brasileira continua liderando as métricas, seja entre as equipes do topo ou as do fundo da tabela de classificação.

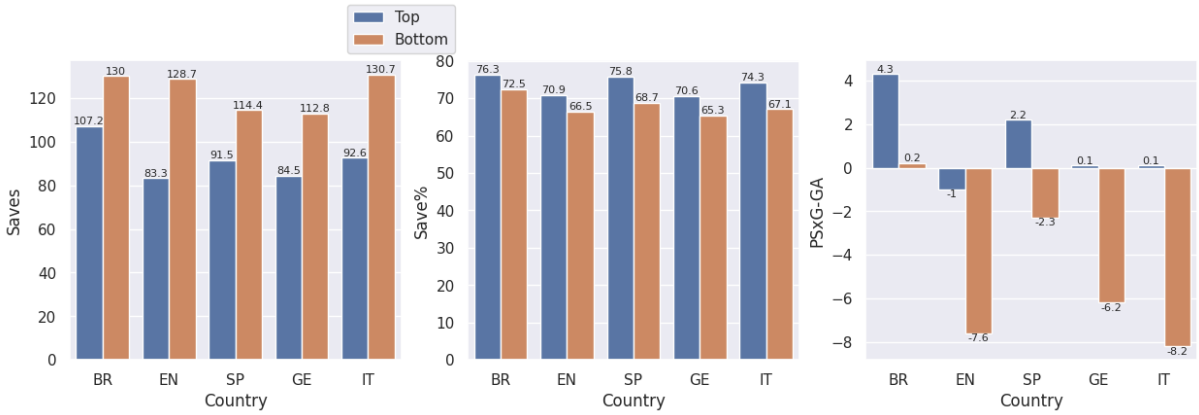

Figura 4.6: Defesas totais (Saves), percentual de defesas 9Save%) e gols esperados pós-chute menos gols permitidos (PSxG-GA). Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano

4.1.4 Passes completos (Completos), percentual de passes completos (Cmp%), passes progressivos (Progressivos)

Agora, serão avaliadas métricas de desempenho relacionadas a passes. As métricas "passes completos" e "percentual de passes completos" não levam em conta o tipo do passe. A métrica de passes progressivos, por sua vez, só contabiliza passes completos que aproximam a bola do gol adversário.

Ao analisar a Figura 4.7, é possível verificar que o percentual de passes completos entre as ligas não varia tanto. Pode-se observar também que, apesar a liga brasileira não ser a liga que mais completa passes, é a liga que mais completa passes progressivos. Isso pode ser devido à qualidade dos passadores, mas também pode ser devido a diferenças táticas das ligas, especialmente o posicionamento das defesas. No geral, é possível ver que, nessas métricas, não há nenhuma grande discrepância que seja particularmente chamativa.

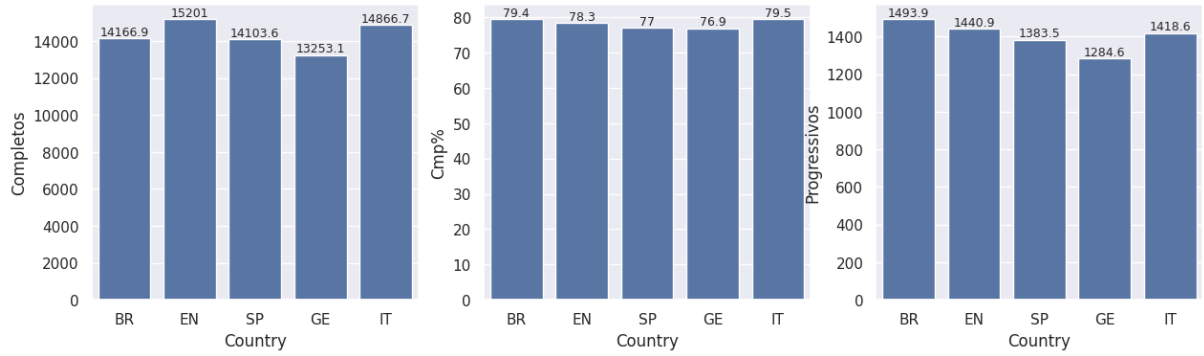

Figura 4.7: Passes completos (Completos), percentual de passes completos (Cmp%), passes progressivos (Progressivos)

Ao analisar as quatro melhores e as quatro piores equipes, conforme a Figura 4.8, é possível notar que o campeonato brasileiro apresenta a menor diferença interna nas três métricas. Também pode-se observar que as quatro piores equipes do campeonato brasileiro, quando comparados às quatro piores equipes das outras ligas, lideram todas as métricas. Isso demonstra um maior equilíbrio interno no campeonato, mas que pode ser causado por diferenças táticas. Na subseção seguinte, serão analisadas métricas defensivas.

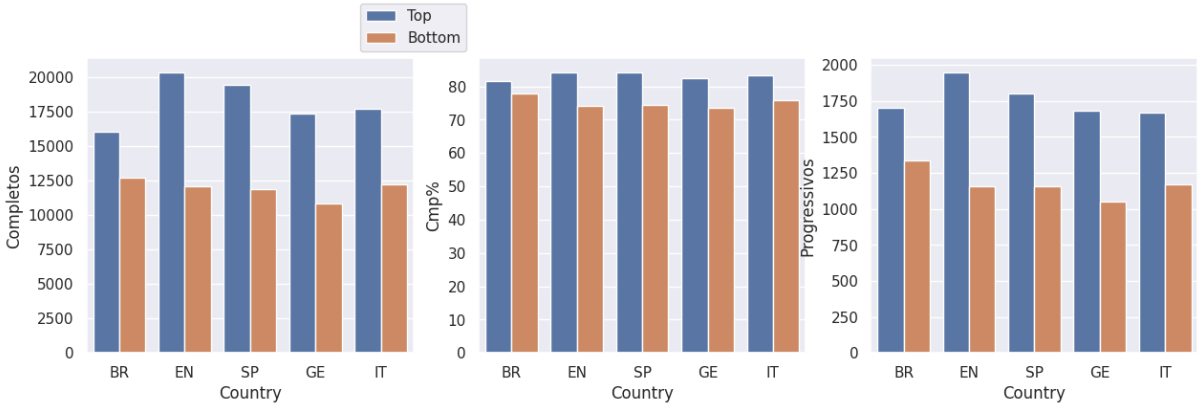

Figura 4.8: Passes completos (Completos), percentual de passes completos (Cmp%), passes progressivos (Progressivos). Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano. Valores das barras foram suprimidos pois são muito grandes

4.1.5 Botes defensivos (Tkl), confrontos no terço de ataque (Att 3rd) e erros defensivos (Err)

A métrica de botes defensivos (Tkl) considera botes em qualquer parte do campo, bem sucedido ou não. Já a métrica de confrontos no terço de ataque, considera apenas confrontos que acontecem no terço de ataque da equipe. É uma métrica que mede o quão agressiva na marcação é uma equipe em porções mais avançadas do campo. A métrica de erros defensivas considera apenas erros que resultaram em uma tentativa de chute do time adversário.

Pela Figura 4.9, é possível observar que o campeonato brasileiro é o segundo em divididas, atrás apenas da liga inglesa. No entanto, ao considerar apenas confrontos no terço de ataque, a liga brasileira cai para a quarta posição nesta métrica, demonstrando que, apesar de apresentar uma marcação combativa, pressão alta não é tão comuns no campeonato brasileiro como é em algumas outras ligas. O campeonato brasileiro também é o que menos apresenta erros defensivos que resultam em tentativas de finalização, o que pode ser reflexo desse baixo número de confrontos no terço de ataque. Essa preterização da marcação alta pode ser devido a aspectos físicos dos atletas, táticos das equipes ou aspectos climáticos.

A liga inglesa se destaca em relação a erros defensivos. Isso pode ser devido ao fato de que é a liga que mais realiza botes defensivos e também a que mais realiza confrontos no terço de ataque (isto é, mais próximo ao gol adversário).

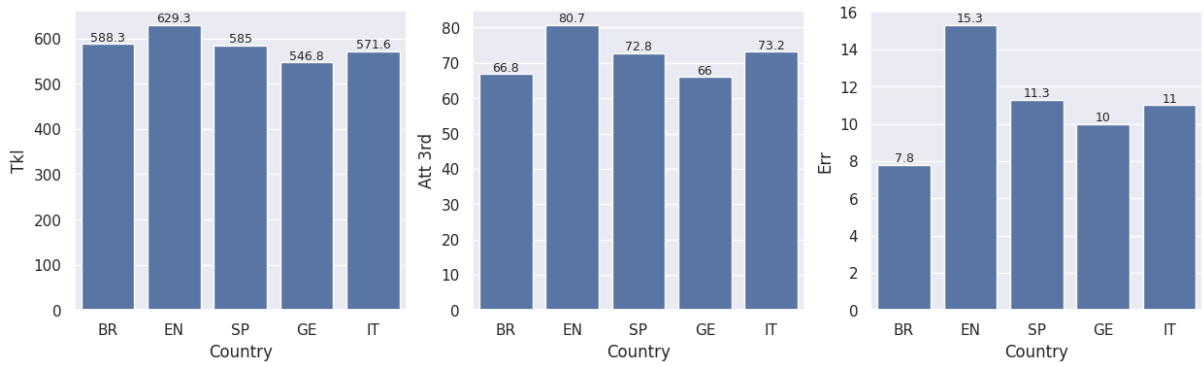

Figura 4.9: Botes defensivos (Tkl), confrontos no terço de ataque (Att 3rd) e erros defensivos que resultaram em chance de gol (Err)

Em relação ao comparativo entre as quatro melhores e as quatro piores equipes, resumido na Figura 4.10, nota-se que apenas nas ligas brasileira e italiana, as equipes de melhor classificação realizam mais botes defensivos que as equipes de pior classificação. Nas outras ligas o que ocorre é o inverso, isto é, as equipes de pior classificação realizam mais botes defensivos do que as equipes de melhor classificação, o que pode ser explicado devido ao fato de que equipes de melhor classificação, em geral, possuem mais posse de bola [27].

Em todas as ligas analisadas, as equipes de melhor classificação realizaram mais confrontos no terço de ataque. No geral, nessas métricas defensivas, não é possível identificar um equilíbrio interno maior no campeonato brasileiro, se comparado com outras ligas, ao analisar essas médias.

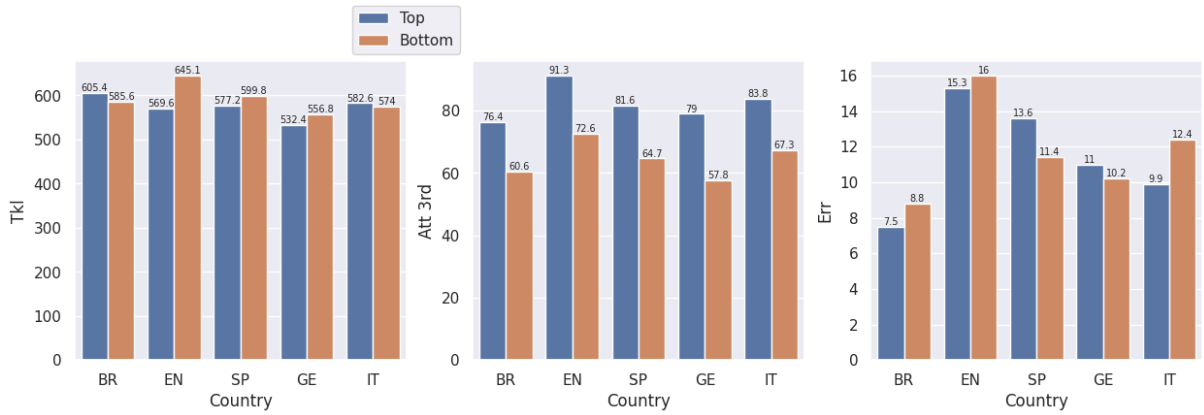

Figura 4.10: Botes defensivos (Tkl), confrontos no terço de ataque (Att 3rd) e erros defensivos que resultaram em chance de gol (Err). Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano. Valores das barras foram suprimidos pois são muito grandes

4.1.6 Tentativas de dribles (Tent. dribles), percentual de dribles certos (Dribles certos %), dribles que resultaram em tentativas de finalização (Dribles tent. chute)

Em relação a dribles, foram analisadas três métricas. Tentativa de dribles, que diz respeito ao número total de tentativas, bem sucedidas ou não; percentual de dribles certos, que usa o cálculo de quantas das tentativas de dribels foram bem sucedidas; e, por fim, dribles que resultaram em tentativas de finalização. Essa última métrica mede a objetividade dos dribles.

Podemos ver os dados resumidos na Figura 4.11. O campeonato brasileiro é o que mais tenta dribles, o que mais acerta dribles e o que tem mais dribles que geram situações de finalização. Ou seja, o campeonato brasileiro lidera em todas essas métricas. O campeonato inglês é o segundo que mais tenta dribles, mas é o penúltimo em percentual de dribles certos.

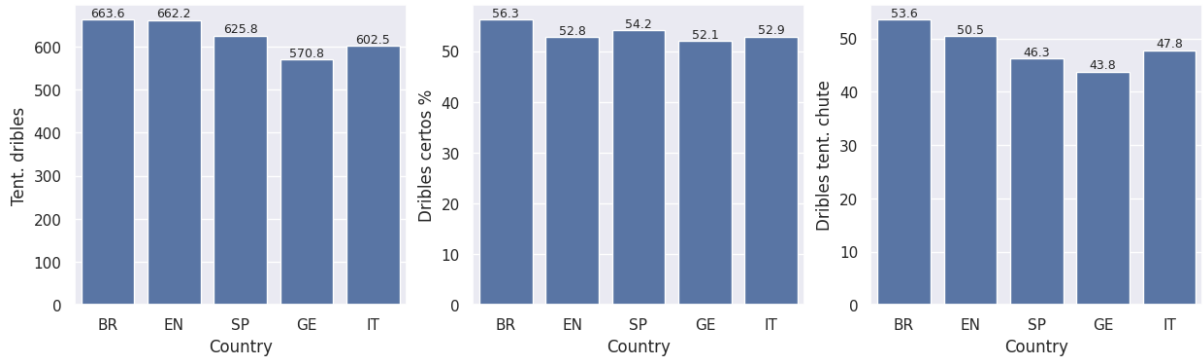

Figura 4.11: Tentativas de dribles (Tent. dribles), percentual de dribles certos (Dribles certos %), dribles que resultaram em tentativas de finalização (Dribles tent. chute)

Analisando os dados considerando apenas as quatro melhores e as quatro piores equipes, pode-se observar, que entre as piores equipes, o campeonato brasileiro ainda lidera todas as métricas. Também é possível ver que o campeonato brasileiro é mais equilibrado em todas as métricas.

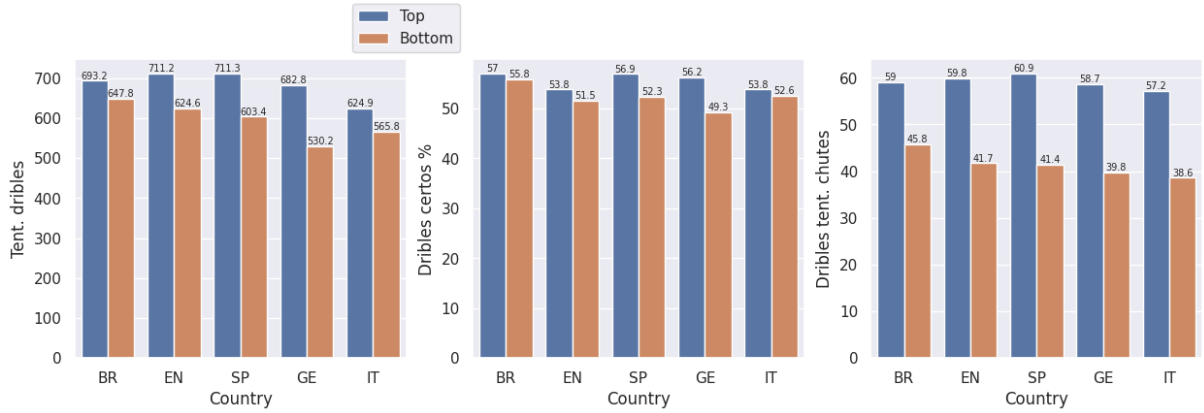

Figura 4.12: Tentativas de dribles (Tent. dribles), percentual de dribles certos (Dribles certos %), dribles que resultaram em tentativas de finalização (Dribles tent. chute). Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano.

4.1.7 Cartões e faltas

A métrica de cartões contabiliza a soma de cartões amarelos e cartões vermelhos. E a métrica de faltas fornece o número de faltas cometidas. Faltas paralisam o jogo e diminuem o tempo de bola rolando nas partidas, o que pode ser considerado uma métrica negativa. Ou seja, números altos nessas métricas podem ser considerados indicadores negativos de desempenho para a equipe e também para o campeonato. Número de cartões quantifica a gravidade das infrações. Essas métricas são particularmente sujeitas a critérios de arbitragem, que podem variar de uma liga para outra.

A Figura 4.13 apresenta uma visualização resumida dessas métricas. Observa-se que o campeonato brasileiro é o campeonato que mais comete faltas e o segundo que mais recebe cartões. Os campeonatos inglês e alemão se destacam positivamente ao apresentarem baixos números para essas duas métricas.

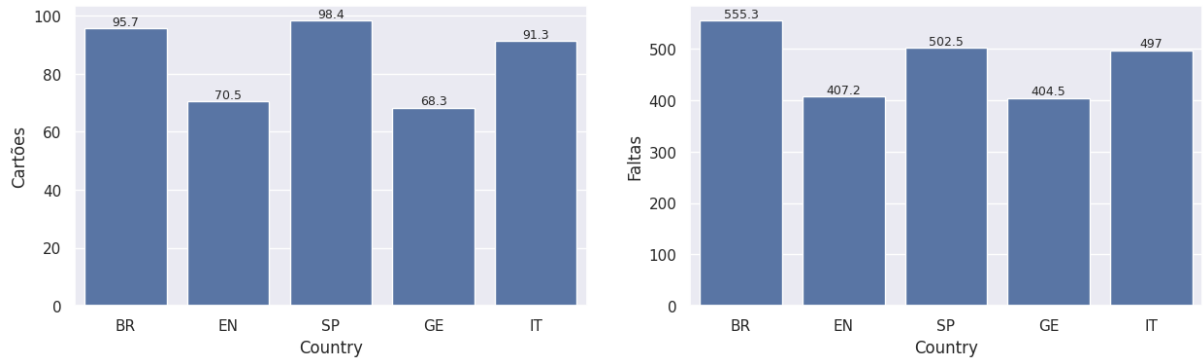

Figura 4.13: Cartões e Faltas

Ao analisar dados comparativos entre as quatro equipes de melhor classificação e as quatro de pior classificação, que podem ser visualizadas na Figura 4.14, é posível observar que em todas as ligas, as equipes de pior classificação cometem mais faltas e tomam mais cartões. No entanto, é possível verificar que o campeonato brasileiro apresenta um maior equilíbrio interno nessas métricas também.

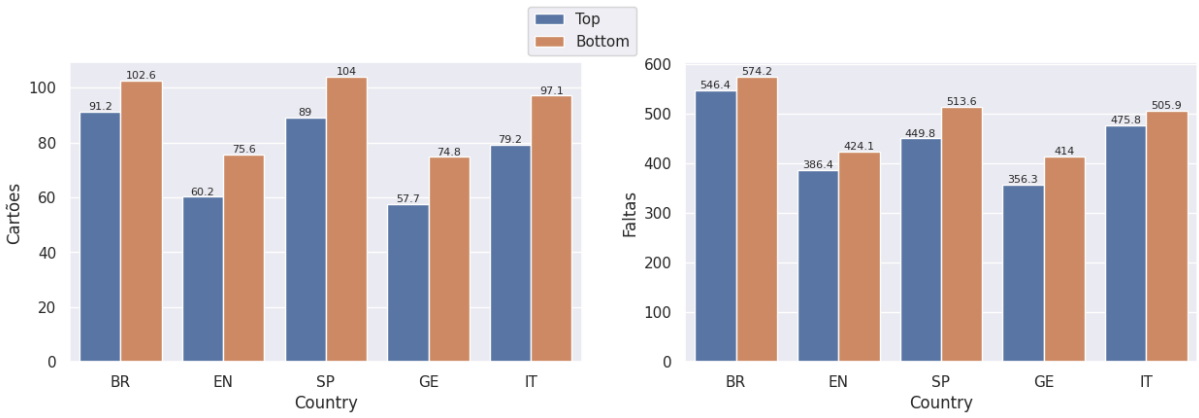

Figura 4.14: Cartões e Faltas. Considerando apenas as quatro equipes de melhor classificação e comparando com as quatro equipes de pior classificação em cada ano.

4.2 Análise de variabilidade

Este seção demonstra alguns resultados relativos à variabilidade das métricas escolhidas. Nas subseções a seguir, serão apresentados histogramas para observar a distribuição dos dados em cada métrica, ao longo de todas as temporadas em análise, para cada uma das ligas. Também é fornecido o Desvio Padrão, que é uma medida de variabilidade, para que seja possível avaliar o quão esparsos são os dados em cada métrica.

4.2.1 Gols (Gls), Gols Esperados (xG) e Gols Esperados Sem Pênalti (npxG)

A Figura 4.15 apresenta os histogramas relativos às métricas Gols, Gols Esperados e Gols Esperados Sem Pênalti. Na figura, é possível observar que todos os histogramas são assimétricos à direita. Isso indica que a ocorrência de valores altos para essas métricas são menos comuns do que ocorrências de valores mais baixos.

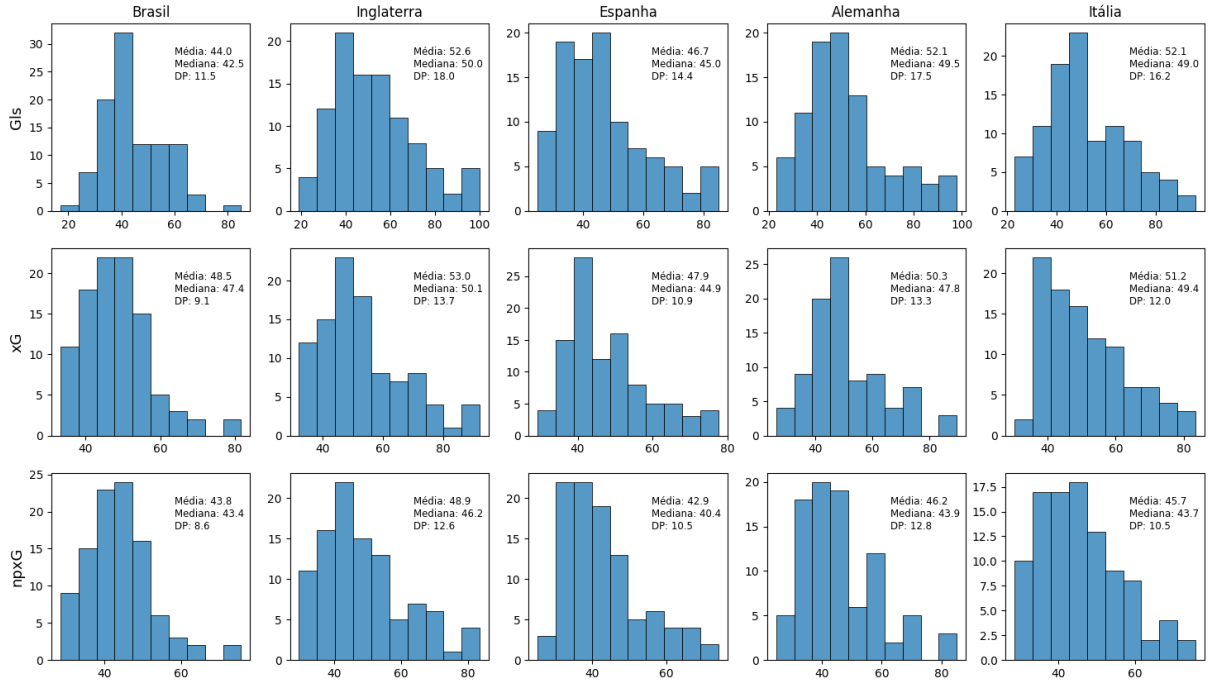

Figura 4.15: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: Dados de gols, gols esperados e gols esperados sem pênalti

Entretanto, pode-se observar que o desvio padrão (DP) do campeonato brasileiro é o menor nas três métricas avaliadas. Isso indica que os valores para cada uma dessas métricas no campeonato brasileiro são menos esparsos, isto é, mais próximos da média.

4.2.2 Total de chutes (Sh), chutes a gol (SoT) e gols por chute (G/Sh)

Os histogramas das métricas Total de chutes, chutes a gol e gols por chute estão apresentados na Figura 4.16. Mais uma vez, é possível observar que o DP do campeonato brasileiro é o menor em todas as métricas. Entretanto, o desvio padrão relativo (quantos por cento da média o desvio padrão representa) revela que para a métrica de Gols por chute, o Desvio Padrão do campeonato italiano é menor, relativamente à média, que é mais alta.

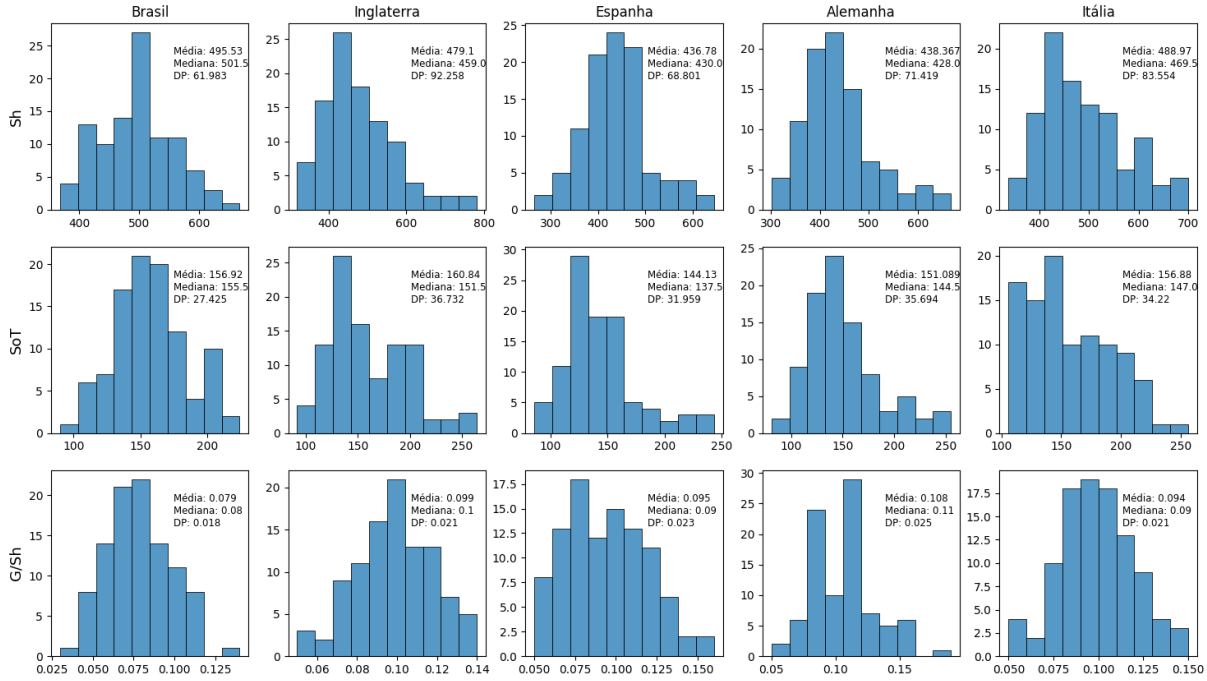

Figura 4.16: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: total de chutes, chutes a gol e gols por chute

Também pode-se observar que, ao contrário do que acontece nas outras ligas, os dados de chutes do campeonato brasileiro são assimétricos à esquerda. Além disso, o histograma de chutes a gol é visivelmente mais simétrico que os histogramas das demais ligas.

4.2.3 Defesas totais (Saves), percentual de defesas (Save%), Gols Esperados Pós-Chute Menos Gols Permitidos (PSxG-GA)

Agora, analisando dados relativos aos goleiros. Para essas métricas, o Desvio Padrão do campeonato brasileiro não é o menor na métrica PSxG-GA. O menor desvio padrão é o do campeonato espanhol. Nas outras duas métricas, o desvio padrão do campeonato brasileiro é o menor.

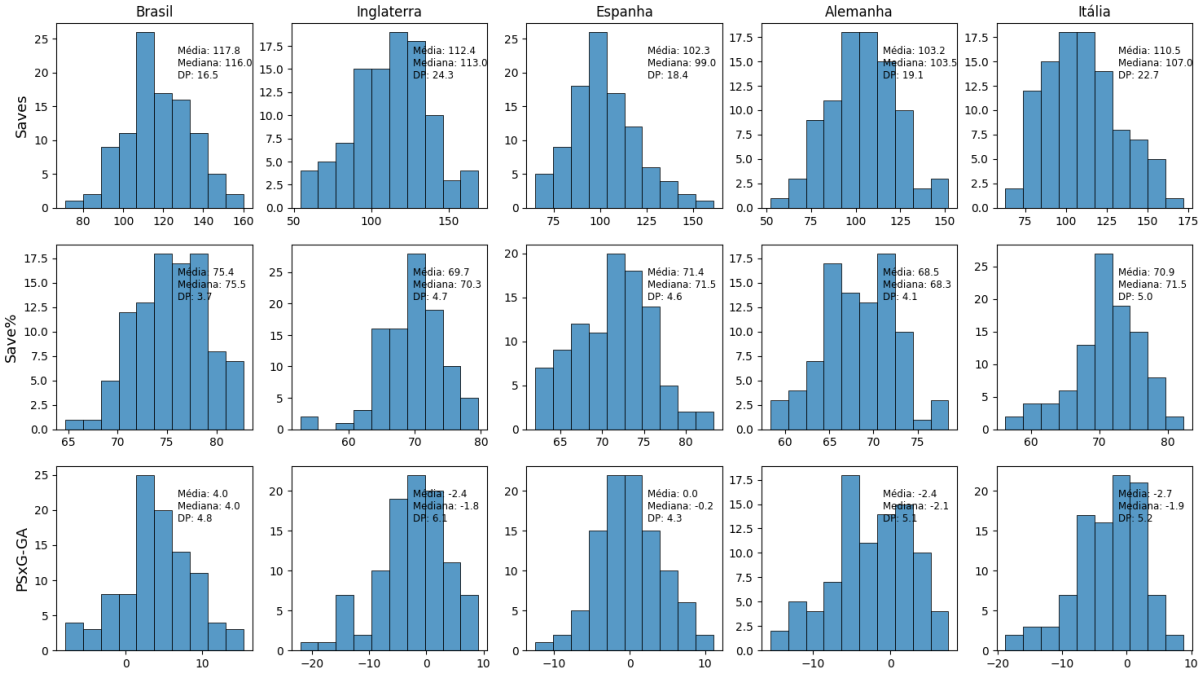

Figura 4.17: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: defesas totais, percentual de defesas, gols esperados pós-chute menos gols permitidos

4.2.4 Passes completos (Completos), Percentual de passes completos (Cmp%), passes progressivos (Progressivos)

Em relação aos dados de passes, pode-se verificar os histogramas na Figura 4.18. Novamente, o campeonato brasileiro apresenta o menor DP nas três métricas, o que significa menor variabilidade nos dados, que pode ser um indicativo de equilibrio interno da liga.

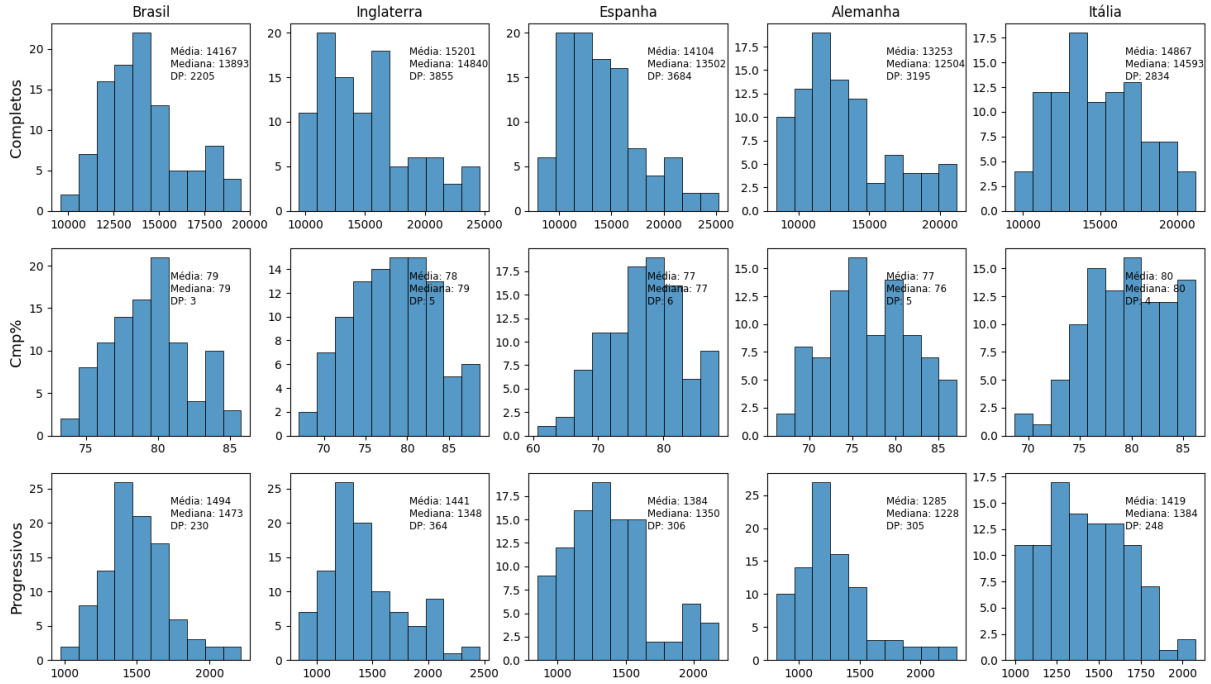

Figura 4.18: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: passes completos, Percentual de passes completos, passes progressivos

4.2.5 Botes defensivos (Tkl), confrontos no terço de ataque (Att 3rd) e erros defensivos (Err)

Em relação às métricas relativas ao desempenho defensivo das equipes, pode-se conferir os histogramas na Figura 4.19. Em relação aos botes defensivos, o campeonato brasileiro possui o segundo menor desvio padrão, atrás apenas da liga alemã. Apesar da liga alemã possuir uma média de botes defensivos menor, mesmo assim o desvio padrão relativo continua sendo o menor. Nas outras duas métricas, o campeonato brasileiro apresenta um desvio padrão menor que as ligas europeias.

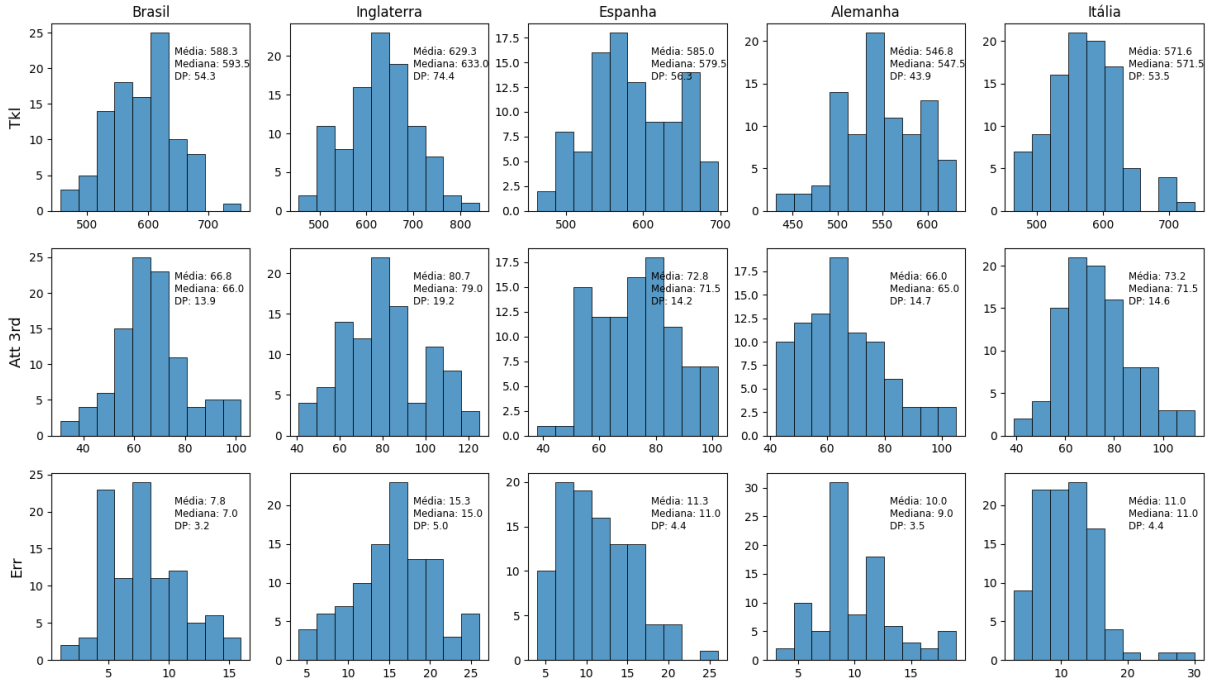

Figura 4.19: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: Botes defensivos, confrontos no terço de ataque e erros defensivos

4.2.6 Tentativas de dribles (Tent. dribles), percentual de dribles certos (Dribles certos %), dribles que resultaram em tentativas de finalização (Dribles tent. chute)

Em relação aos dribles, observa-se os histogramas na Figura 4.20. Na métrica de tentativas de dribles, a liga inglesa possui o menor desvio padrão, isto é, possui uma menor variabilidade, em relação ao campeonato brasileiro. O campeonato italiano possui um DP absoluto menor, mas devido à média menor, o desvio padrão relativo é maior que os campeonatos brasileiro e inglês.

É possível verificar também que o histograma do campeonato brasileiro é assimétrico à direita, enquanto que o histograma da liga inglesa é mais simétrico, mas é levemente assimétrico à esquerda. Isso significa que, no campeonato brasileiro, algumas poucas equipes tentam muito mais dribles que a maioria das equipes. Enquanto que na liga inglesa acontece o inverso: no geral, algumas poucas equipes tentam menos dribles do que a maioria das equipes.

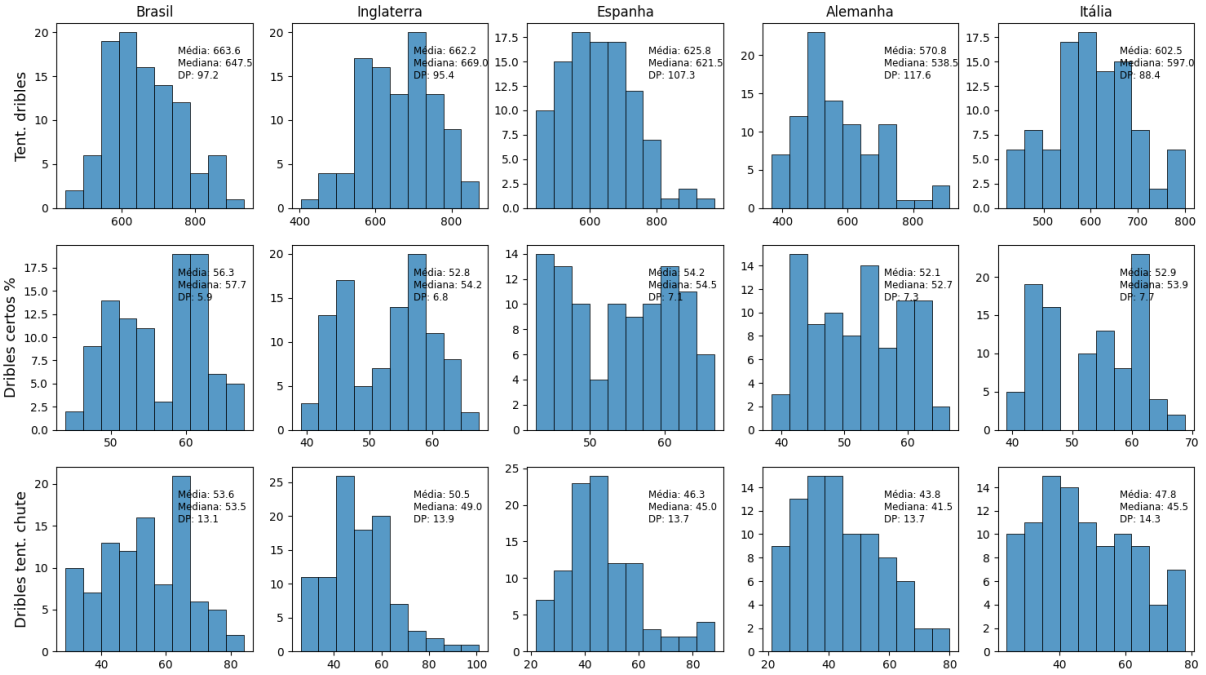

Figura 4.20: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: tentativas de dribles, percentual de dribles certos, dribles que resultaram em tentativas de finalização

Nas outras duas métricas, o desvio padrão do campeonato brasileiro é menor em comparação com as outras ligas. Um comportamento que pode ser observado é que os histogramas da métrica de percentual de dribles certos são multimodais. O fato de isso acontetecer em todas as ligas é algo que chama atenção e pode ser objeto de aprofundamento de estudos.

4.2.7 Cartões e faltas

Os histogramas relativos a infrações, como número de faltas e número de cartões podem ser observados na Figura 4.21. O desvio padrão da métrica de cartões da Alemanha é menor que o do Brasil, mas, ao considerar-se a média, observa-se que o desvio padrão relativo do campeonato brasileiro é menor. O mesmo acontece na métrica de faltas em comparação com as ligas inglesa e alemã, que possuem um DP absoluto menor, mas um DP relativo maior que o campeonato brasileiro.

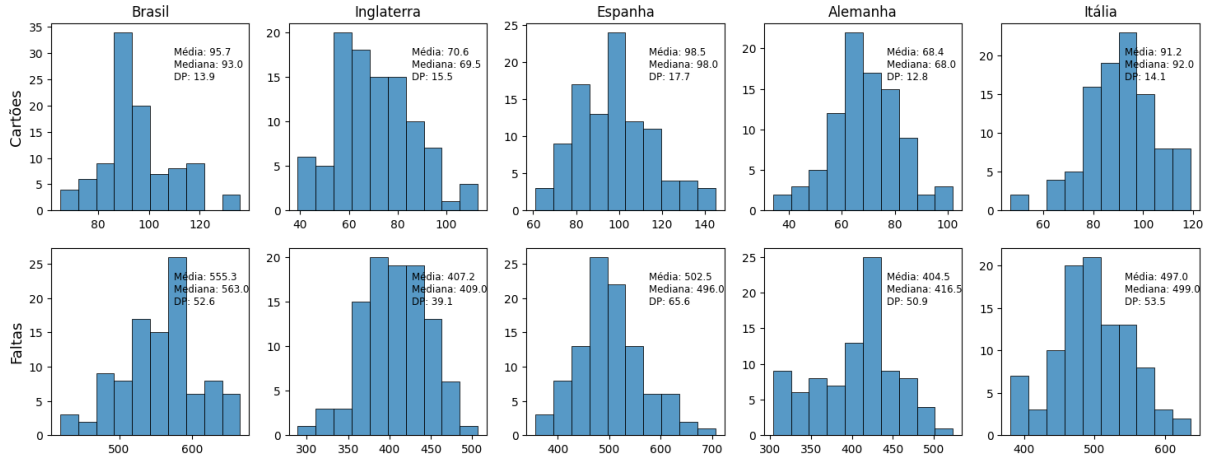

Figura 4.21: Histogramas. Cada coluna representa uma liga, na ordem: Brasil, Inglaterra, Espanha, Alemanha, Itália. Cada fileira representa uma métrica, na ordem: cartões e faltas

5

Conclusão

Neste capítulo, será apresentado um breve resumo de todo o trabalho, inclusive dos resultados obtidos. Além disso, serão comentadas as contribuições e limitações deste trabalho, além de oportunidades para trabalhos futuros.

5.1 Resumo

Uma das formas mais comuns de analisar desempenho de jogadores e equipes de futebol é através da análise notacional. Foi identificado que o campeonato brasileiro ainda não é muito representado em estudos científicos que analisam aspectos técnicos e táticos. Por isso, este trabalho se propôs a coletar e analisar dados do futebol brasileiro e comparar com dados das quatro principais ligas europeias. O objetivo era identificar se a grande disparidade econômica entre as ligas se traduz, necessariamente, em melhores indicadores de desempenho. Diversas métricas que são usadas para análise de desempenho no futebol foram avaliadas por este trabalho, incluindo métricas avançadas, como Gols Esperados (xG). Além disso, também foi avaliado o equilíbrio interno das ligas a partir das métricas analisadas.

Este trabalho importou dados notacionais relacionados do Campeonato Brasileiro e das quatro principais ligas europeias de futebol: Premier League da Inglaterra, La Liga da Espanha, Bundesliga da Alemanha e Série A da Itália. Os dados foram importados do *site* FBRef.com. Foram coletados dados de cada uma das ligas para cinco temporadas: 2019, 2020, 2021, 2022 e 2023 para as temporadas brasileiras; e 2019-2020, 2020-2021, 2021-2022, 2022-2023 e 2023-2024 para as temporadas europeias.

A partir da análise das médias visualizadas através de gráficos de barras, foi possível verificar que o campeonato brasileiro, em diversas métricas, possui indicadores de desempenho piores do que as principais ligas europeias. Ao observar apenas os dados das quatro melhores equipes de cada temporada, é possível verificar que essa discrepância aumenta e que o campeonato brasileiro fica atrás em ainda mais métricas, ao comparar apenas as quatro melhores

equipes de cada temporada para cada uma das cinco ligas. As melhores equipes europeias, de maneira geral, possuem melhores indicadores e apresentam uma diferença maior nas métricas avaliadas.

Entretanto, mesmo com as diferenças de receita e valor de mercado do elenco, o campeonato brasileiro ainda fica a frente em algumas métricas de desempenho, com destaque para o desempenho dos goleiros. Isso acontece especialmente ao considerar a liga como um todo e não apenas os clubes de melhor classificação. A diferença de receitas e valor de mercado da liga nem sempre se traduz em melhores indicadores em diversas métricas de desempenho.

Em relação ao equilíbrio interno das ligas, pode-se verificar, através das visualizações, medidas de variabilidade e medidas de tendência central, que o campeonato brasileiro possui um equilíbrio interno maior que as principais ligas europeias. As análises dos histogramas com o desvio padrão revelaram que os dados do campeonato brasileiro, na maioria das métricas, são menos esparsos e menos assimétricos que os campeonatos europeus. A análise das médias considerando apenas as quatro equipes de melhor classificação e as quatro equipes de pior classificação também revelou que, em geral, a distância entre as melhores equipes e as piores equipes é menor no campeonato brasileiro. Isso corrobora com a tese de que o campeonato brasileiro é mais equilibrado do que as quatro principais ligas europeias.

Em resumo e de maneira geral, pode-se afirmar que as equipes do topo das ligas europeias apresentam indicadores mais positivos que as equipes do topo do campeonato brasileiro. No entanto, essa diferença diminui ao avaliar a liga como um todo. Além disso, o campeonato brasileiro possui um equilíbrio interno maior, apresentando uma menor distância entre as equipes nas métricas analisadas.

5.2 Contribuições

A primeira contribuição deste trabalho foi a disponibilização de conjuntos de dados relacionados a diversas métricas para a análise de desempenho de equipes, tanto em formas mais resumidas como em formas mais extensivas. Esses dados, conforme já descrito, estão disponibilizados publicamente em um repositório online e podem ser usados por outros trabalhos. Isso contribui para abreviar etapas preliminares, como importação e tratamentos de dados, para quaisquer estudos que tenham interesse nessas métricas de desempenho.

Além disso, como discutido no capítulo de Introdução, apesar de haverem diversas pesquisas relacionadas a aspectos técnicos e táticos no futebol, pesquisas com dados de ligas fora da Europa são sub-representadas. Este trabalho também contribui ao estudar dados do futebol brasileiro, que ainda não tem um grande volume de estudos.

Por fim, este trabalho fornece informações e visualizações relacionadas ao futebol brasileiro e quatro das principais ligas europeias. Esses dados e visualizações podem ser consultados e usados como fundamentação para trabalhos futuros.

5.3 Limitações e oportunidades de trabalhos futuros

Uma das limitações deste trabalho se refere à janela de tempo dos dados disponíveis. Dados avançados no futebol brasileiro são mais recentes e, portanto, algumas métricas, como por exemplo o xG, só estão disponíveis no futebol brasileiro a partir de 2019, limitando a janela a cinco anos. Por isso, não foi possível contar com mais dados para aumentar a confiabilidade dos dados, nem fazer uma avaliação temporal considerando um tempo maior, para avaliar tendências e projeções.

Futuramente, a expectativa é de que mais dados avançados estejam disponíveis para análise. Com uma janela de tempo mais longa à disposição, há oportunidade de fazer análises temporais, para identificar tendências nesses dados e fazer projeções para anos futuros usando técnicas de regressão e aprendizagem de máquina. Além disso, essas análises podem ser estendidas para comparar o futebol brasileiro com outras ligas da América do Sul ou mesmo de outros continentes.

Este trabalho também é limitado no que diz respeito às métricas e visualizações usadas para as análises. Outras técnicas estatísticas podem ser empregadas para aprofundar as análises de desempenho e comparações entre as ligas. Pode ser conduzido um teste de hipótese de variância para verificar se há singificância estatística nas diferenças entre as ligas nas métricas avaliadas. Também podem ser empregados coeficientes de igualdade, como o coeficiente de Gini, para aprofundar o estudo em relação ao equilíbrio interno de cada liga. Além disso, podem ser usados coeficiente de assimetria e curtose e outras opções de visualização, como *Boxplot*, para aprofundar as análises.

Referências bibliográficas

[1] pandas.DataFrame &x2014; pandas 2.2.3 documentation — pandas.pydata.org. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html, . [Accessed 24-10-2024].

[2] pandas.DataFrame.mean &x2014; pandas 2.2.3 documentation — pandas.pydata.org. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.mean.html, . [Accessed 02-11-2024].

[3] pandas.DataFrame.std &x2014; pandas 2.2.3 documentation — pandas.pydata.org. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.std.html, . [Accessed 15-11-2024].

[4] pandas.read_html &x2014; pandas 2.2.3 documentation — pandas.pydata.org. https://pandas.pydata.org/docs/reference/api/pandas.read_html.html, . [Accessed 24-10-2024].

[5] Futebol: esporte mais popular no Brasil tem data nacional no mês de julho — www2.camara.leg.br. https://www2.camara.leg.br/a-camara/programas-institucionais/experiencias-presenciais/parlamentojovem/outros-conteudos/projetos-pjb/futebol-esporte-mais-popular-no-brasil-tem-data-nacional-no-mes-de-julho-1. [Accessed 12-10-2024].

[6] Inside FIFA — inside.fifa.com. https://inside.fifa.com/fifa-world-cup-qatar-2022-in-numbers. [Accessed 12-10-2024].

[7] Football Benchmark - Brazilian Football overview: growth, transformation and untapped potential — footballbenchmark.com. https://www.footballbenchmark.com/library/brazilian_football_overview_

growth_transformation_and_untapped_potential#:~:text=The%20aggregate%20operating%20revenue%20of,2023%2C%20despite%20the%20pandemic's%20impact. [Accessed 12-10-2024].

[8] football_leagues_analyses/data at main · victoraccete/football_leagues_analyses — github.com. https://github.com/victoraccete/football_leagues_analyses/tree/main/data. [Accessed 30-10-2024].

[9] Sports Reference | Sports Stats, fast, easy, and up-to-date | Sports-Reference.com — sports-reference.com. https://www.sports-reference.com/?utm_source=fb&utm_medium=sr_xsite&utm_campaign=2023_01_srnav. [Accessed 12-10-2024].

[10] Big Five soccer leagues revenue 2023 | Statista — statista.com. https://www.statista.com/statistics/261218/big-five-european-soccer-leagues-revenue/. [Accessed 12-10-2024].

[11] Bundesliga 24/25 — transfermarkt.com. https://www.transfermarkt.com/bundesliga/startseite/wettbewerb/L1, . [Accessed 12-10-2024].

[12] Campeonato Brasileiro Série A 2024 — transfermarkt.com.br. https://www.transfermarkt.com.br/campeonato-brasileiro-serie-a/startseite/wettbewerb/BRA1/saison_id/, . [Accessed 12-10-2024].

[13] LaLiga 24/25 — transfermarkt.com. https://www.transfermarkt.com/laliga/startseite/wettbewerb/ES1, . [Accessed 12-10-2024].

[14] Premier League 24/25 — transfermarkt.com. https://www.transfermarkt.com/premier-league/startseite/wettbewerb/GB1, . [Accessed 12-10-2024].

[15] Serie A 24/25 — transfermarkt.com. https://www.transfermarkt.com/serie-a/startseite/wettbewerb/IT1, . [Accessed 12-10-2024].

[16] Where would the brazilian league rank amongst European leagues? — reddit.com. https://www.reddit.com/r/football/comments/16vxy76/where_would_the_brazilian_league_rank_amongst/, 2023. [Accessed 15-10-2024].

[17] Zulfiqar Ali and SBala Bhaskar. Basic statistical tools in research and data analysis. *Indian Journal of Anaesthesia*, 60(9):662, 2016. ISSN 0019-5049.

DOI 10.4103/0019-5049.190623. URL http://dx.doi.org/10.4103/0019-5049.190623.

[18] Gabriel Anzer and Pascal Bauer. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3, March 2021. ISSN 2624-9367. **DOI** 10.3389/fspor.2021.624475. URL http://dx.doi.org/10.3389/fspor.2021.624475.

[19] Ryan Benson. The Strongest Leagues in the World: What the Opta Power Rankings Say | Opta Analyst — theanalyst.com. https://theanalyst.com/2024/10/strongest-leagues-world-football-opta-power-rankings. [Accessed 15-10-2024].

[20] Rakshit Bhatnagar and Mridul Babbar. A systematic review of sports analytics. *International Journal of Technology Transfer and Commercialisation*, 19(4):393, 2022. ISSN 1741-5284. **DOI** 10.1504/ijttc.2022.127574. URL http://dx.doi.org/10.1504/IJTTC.2022.127574.

[21] Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists*. O'Reilly Media, Sebastopol, CA, 2 edition, June 2020.

[22] Peter Buneman. Semistructured data. In *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, SIGMOD/PODS97. ACM, May 1997. **DOI** 10.1145/263661.263675. URL http://dx.doi.org/10.1145/263661.263675.

[23] Christopher Carling, A Mark Williams, and Thomas Reilly. *Handbook of soccer match analysis*. Routledge, London, England, March 2005.

[24] Mine Cetinkaya-Rundel, David Diez, and Christopher Barr. *OpenIntro Statistics*. OpenIntro, Inc., fourth edition edition, May 2019. ISBN 978-1943450077.

[25] Silvio Ricardo da Silva and Priscila Augusta Ferreira Campos. The experience of cheering in (so-called) "modern football". In *Football and Social Sciences in Brazil*, pages 471–487. Springer International Publishing, Cham, 2021.

[26] Allen B Downey. *Think Stats 2e*. O'Reilly Media, Sebastopol, CA, October 2014.

[27] Vinicius Martins Farias, Wesley Bierhals Fernandes, Gabriel Gustavo Bergmann, and Eraldo Dos Santos Pinheiro. Relação entre a posse de bola e o resultado de partidas da uefa champions league. *Motricidade*, page Vol. 16 No. 4 (2020): Motricidade, 2020. **DOI** 10.6063/MOTRICIDADE.18382. URL https://revistas.rcaap.pt/motricidade/article/view/18382.

[28] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. **DOI** 10.1109/MCSE.2007.55.

[29] Feroze Kaliyadan and Vinay Kulkarni. Types of variables, descriptive statistics, and sample size. *Indian Dermatology Online Journal*, 10(1):82, 2019. ISSN 2229-5178. **DOI** 10.4103/idoj.idoj$_4$68$_1$8. URL http://dx.doi.org/10.4103/idoj.IDOJ_468_18.

[30] Cole Nussbaumer Knaflic. *Storytelling with data*. John Wiley & Sons, Nashville, TN, October 2015.

[31] Werlayne S. S. Leite. Home advantage: Comparison between the major european football leagues. *Athens Journal of Sports*, 4(1):65–74, February 2017. ISSN 2241-7915. **DOI** 10.30958/ajspo.4.1.4. URL http://dx.doi.org/10.30958/ajspo.4.1.4.

[32] Chunhua Li and Yangqing Zhao. Comparison of goal scoring patterns in "the big five" european football leagues. *Frontiers in Psychology*, 11, January 2021. ISSN 1664-1078. **DOI** 10.3389/fpsyg.2020.619304. URL http://dx.doi.org/10.3389/fpsyg.2020.619304.

[33] Yang Li and Gonzalo Mateos. Networks of international football: communities, evolution and globalization of the game. *Applied Network Science*, 7(1), August 2022. ISSN 2364-8228. **DOI** 10.1007/s41109-022-00498-4. URL http://dx.doi.org/10.1007/s41109-022-00498-4.

[34] Eric Matheus Rocha Lima, Vivian De Oliveira, Vladan Pavlović, Carlos Norberto Fischer, Afonso Antonio Machado, and Ivan Wallan Tertuliano. The influence of expenditures in football industry results: Case study of the brazilian football league. *Management:Journal of Sustainable Business and Management Solutions in Emerging Economies*, 23(1):1, March 2018. ISSN 1820-0222. **DOI** 10.7595/management.fon.2018.0006. URL http://dx.doi.org/10.7595/management.fon.2018.0006.

[35] Lorenzo Lolli, Pascal Bauer, Callum Irving, Daniele Bonanno, Oliver Höner, Warren Gregson, and Valter Di Salvo. Data analytics in the football industry: a survey investigating operational frameworks and practices in professional clubs and national federations from around the world. *Science and Medicine in Football*, page 1–10, May 2024. ISSN 2473-4446. **DOI** 10.1080/24733938.2024.2341837. URL http://dx.doi.org/10.1080/24733938.2024.2341837.

[36] James Mead, Anthony O'Hare, and Paul McMenemy. Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18(4):e0282295,

April 2023. ISSN 1932-6203. **DOI** 10.1371/journal.pone.0282295. URL http://dx.doi.org/10.1371/journal.pone.0282295.

[37] Suresh Kumar Mukhiya and Usman Ahmed. *Hands-On Exploratory Data Analysis with Python*. Packt Publishing, Birmingham, England, March 2020.

[38] Fernando Manuel Otero-Saborido, Rubén D. Aguado-Méndez, Víctor M. Torreblanca-Martínez, and José Antonio González-Jurado. Technical-tactical performance from data providers: A systematic review in regular football leagues. *Sustainability*, 13(18):10167, September 2021. ISSN 2071-1050. **DOI** 10.3390/su131810167. URL http://dx.doi.org/10.3390/su131810167.

[39] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[40] Carlos Pestana Barros, Albert Assaf, and Fabio Sá-Earp. Brazilian football league technical efficiency: A simar and wilson approach. *Journal of Sports Economics*, 11(6): 641–651, January 2010. ISSN 1552-7794. **DOI** 10.1177/1527002509357530. URL http://dx.doi.org/10.1177/1527002509357530.

[41] Girish Ramchandani, Daniel Plumley, Sophie Boyes, and Rob Wilson. A longitudinal and comparative analysis of competitive balance in five european football leagues. *Team Performance Management: An International Journal*, 24(5/6):265–282, August 2018. ISSN 1352-7592. **DOI** 10.1108/tpm-09-2017-0055. URL http://dx.doi.org/10.1108/TPM-09-2017-0055.

[42] Anselmo Ruiz-de Alarcón-Quintero and Blanca De-la Cruz-Torres. An expected goals on target (xGOT) metric as a new metric for analyzing elite soccer player performance. *Data (Basel)*, 9(9):102, August 2024.

[43] Hugo Sarmento, Rui Marcelino, M. Teresa Anguera, Jorge Campaniundefinedo, Nuno Matos, and José Carlos LeitÃo. Match analysis in football: a systematic review. *Journal of Sports Sciences*, 32(20):1831–1843, May 2014. ISSN 1466-447X. **DOI** 10.1080/02640414.2014.898852. URL http://dx.doi.org/10.1080/02640414.2014.898852.

[44] Schlurcher. Sistema de coordenadas cartesiano. https://commons.wikimedia.org/w/index.php?curid=674122, 2017. Obra do próprio, CC BY 4.0.

[45] Mehdi Soltanifar and Francisco J. Santos-Arteaga. Hybrid dea-bwm-kemira approach for multiple attribute decision-making: a weighted analysis perspective. *Soft Computing*,

July 2024. ISSN 1433-7479. **DOI** 10.1007/s00500-024-09933-3. URL http://dx.doi.org/10.1007/s00500-024-09933-3.

[46] Ian Spence. No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics*, 30(4):353–368, December 2005. ISSN 1935-1054. **DOI** 10.3102/10769986030004353. URL http://dx.doi.org/10.3102/10769986030004353.

[47] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[48] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. **DOI** 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

[49] Claus O Wilke. *Fundamentals of data visualization*. O'Reilly Media, Sebastopol, CA, April 2019.

[50] Hong-Bo Xie and Socrates Dokos. A hybrid symplectic principal component analysis and central tendency measure method for detection of determinism in noisy time series with application to mechanomyography. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(2), June 2013. ISSN 1089-7682. **DOI** 10.1063/1.4812287. URL http://dx.doi.org/10.1063/1.4812287.

[51] Nathan Yau. *Data points*. John Wiley & Sons, Nashville, TN, March 2013.

[52] Mike Yi. Mastering Scatter Plots: Visualize Data Correlations — atlassian.com. https://www.atlassian.com/data/charts/what-is-a-scatter-plot. [Accessed 19-10-2024].

[53] Qing Yi, Ryan Groom, Chen Dai, Hongyou Liu, and Miguel Ángel Gómez Ruano. Differences in technical performance of players from 'the big five' european football leagues in the uefa champions league. *Frontiers in Psychology*, 10, December 2019. ISSN 1664-1078. **DOI** 10.3389/fpsyg.2019.02738. URL http://dx.doi.org/10.3389/fpsyg.2019.02738.