UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

O basquete é um jogo de defesa ou um jogo de ataque? *Outliers* multivariados podem elucidar essa questão

Luiz Claudio Diniz

Ouro Preto-MG
2022

Luiz Cláudio Diniz

O basquete é um jogo de defesa ou um jogo de ataque? *Outliers* multivariados podem elucidar essa questão

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador: Anderson Ribeiro Duarte

Coorientador: Josino José Barbosa

Ouro Preto

2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

D585b Diniz, Luiz Claudio.
O basquete é um jogo de defesa ou um jogo de ataque? Outliers multivariados podem elucidar essa questão. [manuscrito] / Luiz Claudio Diniz. - 2022.
41 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Anderson Ribeiro Duarte.
Coorientador: Dr. Josino Barbosa Barbosa.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Estatística .

1. Basketball. 2. Defense. 3. Offense. 4. NBA. 5. Multivarariate outliers. I. Barbosa, Josino Barbosa. II. Duarte, Anderson Ribeiro. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 31

Bibliotecário(a) Responsável: Luciana De Oliveira - SIAPE: 1.937.800

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Luiz Cláudio Diniz

O basquete é um jogo de defesa ou um jogo de ataque? Outliers multivariados podem elucidar essa questão

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 25 de outubro de 2022

Membros da banca

Dr Anderson Ribeiro Duarte - Orientador (Universidade Federal de Ouro Preto)
Dr Josino José Barbosa - Co-orientador (Universidade Federal de Ouro Preto)
Dr Fernando Luiz Pereira de Oliveira (Universidade Federal de Ouro Preto)
Ms Gabriel Lima de Souza (Universidade Federal de Ouro Preto)

Professor Dr. Anderson Ribeiro Duarte, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 25/10/2022

Documento assinado eletronicamente por **Anderson Ribeiro Duarte**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 31/10/2022, às 16:45, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0420158** e o código CRC **6B5D1624**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

Agradecimentos

Agradeço primeiramente a Deus, pela vida e oportunidades nela vivenciadas, nos mantendo com força e persistência para realização de nossos objetivos, mesmo diante das dificuldades.

Ao Professor Doutor Anderson Ribeiro Duarte, meu orientador, por toda atenção e dedicação em suas orientações e sem o qual esse trabalho não seria possível. Também agradeço de forma especial ao Doutor Josino José Barbosa coorientador deste trabalho pela sua contribuição e disponibilidade.

Aos meus quatro amigos irmãos, de curso, (Vinicius Starlino, Ronan Souza, Rogéria Gerçóssimo e Jussara Pereira) os quais fazem parte desta conquista sendo fundamentais nesta trajetória.

Aos meus dois anjos da guarda, minha irmã Leticia Daniela de Oliveira que passou a maior parte desta trajetória acadêmica ao meu lado, quando trouxe ainda mais alegria à minha vida, me motivando a me tornar um ser humano melhor. E a Doutora Natalia Mattar Cantagali por ser um espelho e exemplo e que sempre me incentivou a buscar conhecimento e me cativou a ter pensamento crítico e personalidade nas minhas decisões.

Em especial, gostaria de agradecer à minha mãe Elizabete Saturnino Diniz (em memória), por ser meu exemplo de força e determinação, por todas as sábias palavras e ensinamentos.

Resumo

Nos dias atuais, grande interesse é dedicado para diversas modalidades esportivas de alto rendimento. Valores financeiros bastante vultosos são envolvidos nesses esportes. Tudo isso desperta grande interesse na busca por escolhas baseadas em análise estatística para estabelecer estratégias esportivas para as equipes de alta performance. Particularmente, no basquete, a competição da liga dos Estados Unidos tem um apelo enorme. Este fato é verificado tanto pelo alto nível de qualidade verificado na prática do jogo quanto nos valores financeiros surpreendentemente elevados envolvidos. Um dilema importante é discutido acerca do basquete, as equipes devem dar privilégio para as estratégias defensivas ou ofensivas para ampliar sua expectativa por bons resultados? No jargão popular dos envolvidos com o basquete, a pergunta se resume em: "O basquete é um jogo de defesa ou um jogo de ataque?". As ferramentas estatísticas têm grande importância na tentativa de elucidar esta dúvida. Esta investigação tem por alvo alcançar este objetivo através de técnicas para detecção de valores *outliers*. No mundo da Ciência dos Dados, os *outliers* são definidos como valores extremos ou incomuns em um conjunto de dados. Um *outlier* é um valor específico que parece ser suficientemente diferente do restante dos elementos do conjunto de dados e podem ocorrer por uma série de fatores, como casualidade, erro humano, ou manipulação deliberada, entre outros. Os procedimentos de detecção de *outliers* tem como base análises estatísticas para determinar se elementos de um dado conjunto de dados se comportam como observações extremas. Existem diversos métodos para detecção em um conjunto de dados univariados, mas as técnicas para dados multivariados ainda estão em franca evolução na literatura. Diante disso, esse estudo tem por objetivo apresentar a metodologia *Data-driven Cluster Analysis Method* (DDCAM), uma técnica de detecção de *outliers* multivariados baseada em análise de agrupamentos. Este estudo recorre à metodologia para investigar se equipes de melhor desempenho esportiva na liga de basquete dos Estados Unidos são equipes que se comportam como *outliers* de defesa, ou então como *outliers* de ataque. A análise ocorre através de variáveis específicas inerentes à prática do basquete em nível de alta performace.

Palavras-chave: Basketball, Defense, Offense, NBA, multivariate *outliers*.

Abstract

Nowadays, great interest is devoted to various high-performance sports. Entirely substantial financial values are involved in these sports. All this arouses great interest in the search for choices based on statistical analysis to establish sports strategies for high-performance teams. Particularly in basketball, US league competition has vast appeal. This fact is verified both by the high quality demonstrated in the game practice and the surprisingly high financial values involved. A relevant quandary is discussed about basketball. Should teams prioritize defensive or offensive strategies to increase their expectation for the best results? In the popular terminology of those involved with basketball, the question boils down to this: "Is basketball a defensive game or an offensive game?". Statistical tools are of great importance in an attempt to elucidate this doubt. This investigation aims to achieve this objective through techniques for detecting outliers values. In Data Science, outliers are defined as extreme or unusual values in a dataset. An outlier is a specific value that appears to be sufficiently different from the rest of the elements in the dataset. It can be due to several factors, such as chance, human error, or deliberate manipulation. Outliers detection procedures are based on statistical analysis to determine whether elements of a given data set behave as extreme observations. Several methods for detection in a univariate dataset, but techniques for multivariate data are still evolving in the literature. Therefore, this study aims to present the *Data-driven Cluster Analysis Method* (DDCAM) methodology, a technique for detecting multivariate outliers based on cluster analysis. This study makes use of the method to investigate whether teams with the best sports performance in the US basketball league are teams that behave as defensive outliers or as offensive outliers. The analysis takes place through specific variables inherent to basketball practice at a high-performance level.

Keywords: Basketball, Defensive, Offensive, NBA, multivariate outliers.

Lista de ilustrações

Figura 1 – Boxplot dos pontos convertidos por jogo. 15
Figura 2 – Boxplot da média de arremessos tentados. 17
Figura 3 – Boxplot do percentual de arremessos certos. 18
Figura 4 – Arremesso de três pontos. 19
Figura 5 – Boxplot da média de arremessos de três pontos tentados 20
Figura 6 – Boxplot do percentual de arremessos de três pontos certos. 21
Figura 7 – Lance livre. 22
Figura 8 – Boxplot da média de lances livres tentados. 23
Figura 9 – Boxplot do percentual de lances livres certos. 24
Figura 10 – Boxplot do número de assistências. 26
Figura 11 – Boxplot do número de erros em posse de bola 27
Figura 12 – Boxplot dos rebotes defensivos por jogo. 29
Figura 13 – Ação de rebote. 30
Figura 14 – Boxplot do total de rebotes. 30
Figura 15 – Ação de toco. 31
Figura 16 – Boxplot do número de tocos. 32
Figura 17 – Boxplot do número de faltas. 34
Figura 18 – Relação entre as equipes classificadas aos *playoffs*. 36

Lista de tabelas

Tabela 1 – Equipes *outliers* verificadas univariadamente para a variável Pontos por jogo. 15
Tabela 2 – Equipes *outliers* verificadas univariadamente para a variável Média de arremessos tentados. 17
Tabela 3 – Equipes *outliers* verificadas univariadamente para a variável Percentual de arremessos certos. 19
Tabela 4 – Equipes *outliers* verificadas univariadamente para a variável Média de arremessos de três pontos tentados. 20
Tabela 5 – Equipes *outliers* verificadas univariadamente para a variável Percentual de arremessos de três pontos certos. 22
Tabela 6 – Equipes *outliers* verificadas univariadamente para a variável Média de lances livres tentados. 23
Tabela 7 – Equipes *outliers* verificadas univariadamente para a variável Percentual de lances livres certos. 25
Tabela 8 – Equipes *outliers* verificadas univariadamente para a variável Número de assistências. 26
Tabela 9 – Equipes *outliers* verificadas univariadamente para a variável Número de erros em posse de bola. 28
Tabela 10 – Equipes *outliers* verificadas univariadamente para a variável rebotes defensivos por jogo. 29
Tabela 11 – Equipes *outliers* verificadas univariadamente para a variável Total de rebotes. 31
Tabela 12 – Equipes *outliers* verificadas univariadamente para a variável Número de tocos. 32
Tabela 13 – Equipes *outliers* verificadas univariadamente para a variável Número de faltas. 34
Tabela 14 – Equipes *outliers* verificadas por temporada da NBA. 35

Sumário

1 INTRODUÇÃO 1
1.1 Motivação 2
1.2 Objetivos 2
1.2.1 Objetivos Gerais 2
1.2.2 Objetivos Específicos 2

2 FUNDAMENTAÇÃO TEÓRICA 3

3 ABORDAGEM DO PROBLEMA E ASPECTOS METODOLÓGICOS 9
3.1 Estimação do valor δ 10
3.1.1 Processo de Refinamento - I 11
3.1.2 Processo de Refinamento - II 11
3.1.3 Busca pelo Valor Adequado k 11

4 RESULTADOS ALCANÇADOS 13
4.1 População de Dados do Estudo 13
4.2 Análise Descritiva de Dados 14
4.2.1 Variáveis de Ataque 14
4.2.1.1 Pontos Convertidos 14
4.2.1.2 Média de Arremessos Tentados 16
4.2.1.3 Percentual de Arremessos Certos 18
4.2.1.4 Média de Arremessos de Três Pontos Tentados 19
4.2.1.5 Percentual de Arremessos de Três Pontos Certos 21
4.2.1.6 Média de Lances Livres Tentados 22
4.2.1.7 Percentual de Lances Livres Certos 24
4.2.1.8 Número de Assistências 25
4.2.1.9 Número de Erros em Posse de Bola 27
4.2.2 Variáveis de Defesa 28
4.2.2.1 Número de Rebotes Defensivos 28
4.2.2.2 Total de Rebotes 29
4.2.2.3 Número de Tocos 31
4.2.2.4 Número de Faltas Cometidas por Jogo 33
4.3 Investigação multivariada 35

5 CONSIDERAÇÕES FINAIS 37

REFERÊNCIAS 39

1 Introdução

Em um mundo cada dia mais dinâmico, a geração e tratamento de dados é algo recorrente e intrínseco aos processos que cercam a sociedade. Dessa forma, os conjuntos de dados estão em toda parte e em grandes quantidades, estes conjuntos abrangem quase todos os assuntos de interesse prático. Compreender o sentido de toda essa informação é uma necessidade pungente da sociedade atual. Em particular, a presença de observações aberrantes nestes conjuntos de dados, usualmente denominadas como *outliers* tende a dificultar a análise estatística.

De uma forma direta, porém enxuta, valores *outliers* são valores que se distanciam demasiadamente da maioria das observações. Estes valores podem estar presentes em amostras de grande ou pequena dimensão. Como mencionado, podem interferir diretamente ou indiretamente nas análises estatísticas. Em geral, essa interferência tende a gerar estimativas enviesadas ou modelos que produzem predições distorcidas ou até mesmo errôneas.

De acordo com o domínio de aplicação, os *outliers* podem ser nomeados como observações suspeitas, exceções, peculiaridades, contaminantes. É possível citar alguns exemplos em que processos para detectar outliers são usados:

- descoberta de fraudes em análise de dados de cartões de crédito, nestes casos, comportamentos atípicos de compras podem ser considerados como observações suspeitas;

- análise de dados geoquímicos em que os *outliers* geralmente são derivados de processos secundários que podem revelar, por exemplo, uma contaminação;

- identificação de fraudes em recursos públicos, nas cotas para o exercício de atividade parlamentar é possivel visualizar valores peculiares e valores fora dos padrões que podem identificar um abuso de gastos ou mal versação do dinheiro público.

Estes são apenas alguns exemplos dentre tantos outros do cotidiano. Por outro lado, a definição sobre determinar se determinada informação, em uma base de dados, é ou não um valor *outlier* não é nada trivial. Isso decorre da subjetividade da interpretação desse problema de identificação. Apesar dessa dificuldade, a detecção de valores atípicos em inúmeras aplicações é um processo indispensável e valioso.

1.1 Motivação

Toda essa aplicabilidade ligada à investigação da presença de valores *outliers* transforma este assunto em um objeto de pesquisa demasiadamente instigante. A existência de técnicas específicas para identificação de valores *outliers* e a possibilidade de estudo e aplicação dessas técnicas é a motivação principal da proposição deste estudo.

A recente técnica *Data-driven Cluster Analysis Method* (DDCAM) proposta por Duarte, Barbosa, Martins e Oliveira (2022) [1] tem mostrado potencial significativo nos procedimentos de detecção de *outliers* multivariados. Este apresentará a referida técnica e uma aplicação em dados reais. Os dados em estudo são originários do ambiente esportivo, mais precisamente do basquete. O basquete é um esporte que se apresenta na vanguarda da tomada de decisões baseada em análise estatística. Em particular, as decisões sobre estratégias defensivas e ofensivas são de enorme interesse no basquete. A aplicação de detecção de *outliers* em variáveis específicas ligadas às estratégias de ataque e defesa podem fornecer informações de grande valia no processo decisório do basquete.

1.2 Objetivos

1.2.1 Objetivos Gerais

De uma forma mais generalista, os objetivos deste estudo são apresentar uma conceitualização do termo *outlier*, o que já foi inicializado nessa introdução. Discorrer sobre metodologias para deteção de *outliers* e apresentar uma aplicação efetiva do processo de identificação de *outliers* em dados reais esportivos do basquete.

1.2.2 Objetivos Específicos

- apresentar revisão bibliográfica atualizada acerca das investigações sobre valores *outliers*;
- descrever a metodologia de detecção de valores *outliers* denominada DDCAM;
- ilustrar a utilização da estratégia através da aplicação em dados reais esportivos do basquete.

2 Fundamentação Teórica

Uma observação aparentemente inconsistente ao ser comparada ao restante de um conjunto de dados é definida como um *outlier* segundo Hawkins (1980) [2]. Já Barnett e Lewis (1994) [3], afirmam que observações extremamente desviadas do conjunto de dados, que levantam suspeição tendem a ser geradas por mecanismos distintos da estrutura dos dados.

A conceitualização de valores *outliers* univariados é menos complexa. Por outro lado, o conceito de *outlier* multivariado é mais sofisticado, tem respeito com distância entre subespaços $k-$dimensionais definidos pelas variáveis em estudo. É importante ressaltar que observações podem configurar *outliers* multivariados e não configurarem *outliers* univaridados se avaliadas unidimensionalmente [4].

Diversos autores abordaram aplicações contendo metodologias para a detecção de *outliers* multivariados. Informações relevantes com respeito à características raras, porém existentes, quanto aos dados sob investigação poodem impactar sobremaneira as análises envolvidas. Será apresentada aqui uma revisão bibliográfica acerca de estudos e metodologias que circundam a área de pesquisa que aborda os *outliers* multivariados.

Rousseeuw e Zomeren (1990) [5] abordam a detecção de *outliers* através de distâncias robustas de Mahalanobis. O procedimento é baseado em estimação robusta com base no método de *minimum volume elemental ellipse* (MVE). O estimador do MVE pode ser obtido com a utilização do elipsoide de menor volume capaz de cobrir pelo menos k pontos do conjunto amostral, em que $^n/_2 < k < n$.

Filzmoser, Garrett e Reimann (2005) [6] apresentam uma abordagem de detecção de *outliers* multivariados em dados geoquímicos. A importância desta correta identificação destaca-se na geoquímica, porque os outliers são geralmente observações resultantes de um processo secundário e não valores extremos da distribuição base. Amostras em que os valores analíticos são derivados de um processo secundário podem ser mineralização ou contaminação e não necessariamente precisam ser altos ou baixos em relação a todos os valores de uma variável em um conjunto de dados. Para a análise dos dados a técnica utilizada foi o estimador robusto com base no método de *Minimum covariance determinant* (MCD).

Diferentemente da técnica baseada no MCD, que encontra um subconjunto de pontos tal que o volume do elipsoide é definido pela covariância da amostra desses pontos é minimizada, o método baseado no MVE busca de forma diretas o elipsoide de volume mínimo, basicamente o algoritmo encontra um elipsoide que inclui o conjunto de dados que está centrado em torno da origem e tem o menor volume. Uma

desvantagem deste método se deve ao fato dele identificar um grande número de *outliers*.

Filzmoser (2005) [7] apresenta uma comparação entre três métodos para a identificação de outliers multivariados, o método baseado em MVE [5]; o método de Becker e Gather (1999) [8], uma extensão multivariada do método de Davies e Gather (1993) [9] e o método baseado em MCD [6]. Os três são métodos baseados na distância robusta de Mahalanobis. Os experimentos foram conduzidos em dados simulados em duas condições (dados com baixa dimensão e dados de alta dimensão). Como ponderação para as conclusões dos experimentos destacam que simulações mostram que o desempenho dos três métodos é determinado principalmente pelo desempenho do estimador MCD. Especialmente os experimentos com dados de alta dimensão refletem as limitações do MCD para identificar porcentagens mais altas de *outliers*.

Leys, Klein, Dominicy e Ley (2018) [10] tratam da importância da correta identificação de *outliers* multivariados e apresenta uma crítica ao método de detecção usando a distância básica de Mahalanobis. A crítica decorre desse indicador usar a média da amostra multivariada e a matriz de covariância que são particularmente sensíveis aos *outliers*.

Li, Deng, Li e Jiang (2019) [11] trazem uma abordagem de detecção de *outliers* com foco em dados com grandes dimensões (*big data*). É proposta a distância de Mahalanobis melhorada, com base em um estimador *Rocke* mais robusto sob dados de alta dimensionalidade. A motivação da proposta é atribuída pelo problema da distância clássica de Mahalanobis ser extremamente afetada pelos *outliers*. Outro problema se deve ao fato do estimador MCD, baseado na distância robusta de Mahalanobis, poder detectar discrepâncias com mais precisão em dados de baixa dimensão, mas capaz de produzir grande desvio em dados de alta dimensão. Desta forma o método de detecção baseado no estimador *Rocke* se mostrou bastante eficaz em dados de alta dimensão.

Veloso e Cirillo (2016) [12] abordam o emprego de simulações de Monte Carlo na construção de um teste de significância aplicado na identificação de *outliers* utilizando o método de componentes principais. Para a validação, o método foi empregado em uma série temporal multivariada em relação às taxas de volume de vendas no estado de Minas Gerais, obtidas em diferentes segmentos de mercado. A aplicação do teste de significância para o coeficiente de curtose associado ao primeiro componente principal e identificação de *outliers* pelos componentes principais foi bem promissora e os outliers detectados correspondiam a picos sazonais em todos os segmentos avaliados no estudo. Embora o método não propõe nenhuma nova forma de detecção de outliers, visto que a metodologia de detecção de outliers via análise de componentes principais já ser bastante difundida, o método através do teste de significância sugere quais componentes principais são significativos na geração dos valores discrepantes. Essa contribuição se

torna útil devido que os componentes são sensíveis a *outliers*, e sua estimação pode ser influenciada, também existe o problema de quais componentes principais serão considerados.

Kamalov e Leung (2020) [13] abordam um método de detecção de *outliers* em dados multivariados através da análise de componentes principais. O PCOut é um procedimento para identificar que utiliza propriedades inerentes da decomposição de componentes principais. A principal vantagem do método proposto em relação aos existentes como os baseados em MVE e MCD é seu baixo custo computacional e aplicação em dados de grandes dimensões. O método apresenta um desempenho eficaz para dados de alta dimensionalidade e também bons resultados para detecção de valores discrepantes em dimensão inferiores.

Ghorbani (2019) [14] apresenta uma revisão de algumas ferramentas para detecção de *outliers* em ambientes univariados tais como o *box-plot* e o gráfico *QQ-plot*. Para ambientes com dados multivariados correlacionados. Afirma, ainda que a distância de Mahalanobis é mais adequada que a distância euclidiana porque leva em consideração a correlação entre as variáveis. O estudo mostra como as distâncias de Mahalanobis podem ser usadas como ferramentas para identificar *outliers* multivariados e também como é mais vantajoso usar a distância robusta de Mahalanobis visto que este método e menos afetado por dados discrepantes.

Zhu et al. (2017) [15] apresentam problemas enfrentados por algoritmos de detecção de *outliers* que trabalham em tempo real aplicado às rotas de GPS. O estudo tem como objetivo detectar as trajetórias anômalas com a ajuda do conjunto de dados de trajetórias históricas e das rotas populares, levando em consideração as anormalidades espaciais e temporais simultaneamente. Como proposta, os autores apresentam o algoritmo *time-dependent popular routes based* (TPRO). Este método utiliza um algoritmo de detecção de *outliers* multivariados em tempo real, que contém uma etapa de pré-processamento *off-line* e outra etapa de detecção *on-line*.

Kutsuna e Yamamoto (2017) [16] abordam um método para detecção de *outliers* através de diagramas de decisão binários. A densidade *leave-one-out* é proposta como uma nova medida para detecção de *outliers*, que é definida como a razão entre o número de elementos de dados dentro de uma região e o volume da região após a remoção de um dado focalizado. Inicialmente o método foi aplicado em um conjunto de dados sintéticos e de um repositório de dados para estudos de *machine learning*. Os resultados mostram que o método pode lidar com um grande conjunto de dados de forma eficiente, porque a complexidade do tempo é quase linear.

Luo et al. (2018) [17] apresentam um método de identificação de *outliers* multivariados baseado em variogramas para vetores de imagens médicas. O variograma é uma ferramenta geoestatística simples e informativa para caracterizar a dependência espa-

cial de processos estocásticos. O método apresentado usa o variograma para estimar a dependência espacial dos vetores deslocamento. Uma vez que a correlação espacial dos vetores atípicos se comporta de forma diferente dos vetores de deslocamento válidos, eles podem ser identificados e classificados como *outliers* de forma eficiente no variograma.

Barbosa, Pereira, Oliveira (2018) [18] apresentam uma técnica de detecção de *outliers* multivariados baseada em análise agrupamento e compara essa técnica com o método de identificação de outliers via Distância de Mahalanobis em dados simulados via método de Monte Carlo e a técnica de mistura de distribuições normais multivariadas Os resultados apresentados mostram que o método proposto foi superior ao método baseado na distância de Mahalanobis tanto para sensibilidade quanto para especificidade mostrando uma capacidade superior de identificar os *outliers* e os não *outliers*.

Wang et al. (2019) [19] apresentam um novo modelo de detecção de *outliers* multivariados que usa um gráfico virtual, denominado *virtual outlier score* (VOS). O modelo proposto constrói um gráfico de similaridade através dos primeiros k vizinhos semelhantes de cada objeto e introduz um acoplamento de nó virtual com uma coleção de arestas virtuais para gerar um grafo virtual. A seguir, um processo de passeio aleatório de Markov personalizado é executado no grafo virtual, sob a pressuposição de que um potencial valor discrepante deve receber mais peso para ser visitado. O método foi aplicado em um conjunto de dados sintéticos e reais e mostrou uma melhoria sobre alguns algoritmos de última geração sob as medidas de ROC.

Wang e Mao (2019) [20] afirmam que para um processo industrial, é de grande importância garantir a qualidade do produto e controlar o desempenho simultaneamente. Neste contexto, a correta identificação de valores e observações atípicas se faz de extrema importância. O estudo aborda a identificação de *outliers* baseado em modelos de processo gaussianos. Através de seleções específicas de função média, função de covariância, função de verossimilhança e método de inferência, foram desenvolvidos três algoritmos de detecção de *outliers* baseados na regressão de processo gaussiana e um baseado na classificação de processo. Como resultado e comparado com os métodos tradicionais de detecção, o modelo proposto tem menos suposições e é mais adequado para processos industriais.

Wahid e Rao (2019) [21] apresenta um algoritmo de detecção de *outliers* baseado em distância usando o clássico algoritmo *particle swarm optimization* (PSO). Neste algoritmo, é atribuido um grau de possbilidade de ser um *outlier* a cada dado através da soma das distâncias entre os pontos e seu conjunto de k-vizinhos mais próximos. O PSO é usado para detectar subespaços periféricos em que podem existir *outliers*. O algoritmo PSO lida bem em conjuntos de dados com grandes dimensões.

Lejeune, Mothe, Soubki e Teste (2020) [22] abordam a detecção de valores atípicos em dados funcionais multivariados (dados funcionais multivariados referem-se a uma população de funções multivariadas geradas por um sistema envolvendo parâmetros dinâmicos dependendo de variáveis contínuas, por exemplo, séries temporais multivariadas). A identificação de valores discrepantes nesse ambiente pode ser desafiadora porque tanto o comportamento individual dos parâmetros quanto a correlação dinâmica entre eles são importantes. A abordagem proposta para a identificação de *outliers* foi efetivada através de um método pelo qual diferentes características periféricas são capturadas com base em funções de mapeamento de geometria diferencial.

Barbosa, Duarte, e Martins (2020) [23] apresentam uma comparação entre as técnicas de detecção baseadas na distância de Mahalanobis (baseados em MCD e MVE), além de uma metodologia baseada da técnica de análise de agrupamento [18]. O estudo encontra possibilidades para melhorias na metodologia baseada em análise de agrupamento. O processo de detecção é bastante dependente da escolha do número de grupos.

van Zoest, Stein e Hoek (2018) [24] apresentam uma abordagem de detecção de *outliers* para uma rede de sensores de monitoramento de ar da cidade de Eindhoven na Holanda. A rede de sensores utilizada está espalhada por 35 pontos distintos da cidade. Cada caixa de ar mede material particulado, ozônio (O^3) e/ou dióxido de nitrogênio (NO^2). Para a análise dos dados e a detecção de possíveis valores atípicos, as observações em um intervalo de um ano são divididas em 16 classes espaço-temporais. A detecção de valores discrepantes é baseada na verificação se um valor de concentração observado está dentro de um determinado intervalo de confiança. A abordagem de detecção de valores discrepantes se baseia em análises gráficas e a determinação de um intervalo de confiança de 99,7% para o nível médio de NO^2 no ar.

Tabjula, Kanakambaran, Rajagopal e Srinivasan (2021) [25] consideraram a importância da Saúde Estrutural (SHM) de ativos de infraestrutura tais como pontes, torres de transmissão e de transportes tais como aeronaves e navios. É, portanto de suma importância o monitoramento dessas estruturas para minimizar e coibir falhas catastróficas. Diante disso, os autores propõem o monitoramento através da geração, propagação e detecção de ondas guiadas por ultra-som. Para a detecção de defeitos foi proposta uma abordagem de detecção de *outliers* em duas etapas para identificar as regiões potencialmente defeituosas na amostra. As etapas principais envolvidas nesta abordagem são a seleção esparsa aleatória dos pontos de detecção, seguida por um processo de detecção de *outliers* multivariados de duas etapas com base no limiar e no cálculo do desvio absoluto mediano.

Diversos estudos acerca do problema de detecção de *outliers* são apresentados nesta revisão. Por fim, o estudo de Duarte, Barbosa, Martins e Oliveira (2022) [1]

apresenta a proposição DDCAM para uma estratégia guiada pelos dados para conduzir o procedimento de detecção de *outliers* multivariados e será o objeto central desse estudo. O método carrega especificidades únicas que serão apresentadas com mais detalhes nas apresentações metodologicas desse estudo.

3 Abordagem do Problema e Aspectos Metodológicos

Com o interesse específico em analisar dados esportivos por meio de técnica de detecção de valores extremos, o método DDCAM [1] será detalhado. O DDCAM parte de duas ideias principais: o ajuste *a priori* baseado nas informações do usuário da metodologia, e em informações intrínsecas contidas na própria base de dados, essas premissas justificam a nomenclatura *Data-driven*.

A estratégia DDCAM é proposta com base em um método não hierárquico de análise de *clusters*. Este método é caracterizado pela necessidade inicial de definir um número prévio de grupos, para produzir alguma partição de seus elementos. A proposição parte da técnica predecessora *Cluster Analysis Method* (CAM) [18].

No procedimento de análise de *clusters*, para a construção de k grupos, o método CAM utiliza o procedimento de análise de agrupamentos $k-$médias. O objetivo é estabelecer agrupamentos por semelhança entre as observações da base de dados. O método de agrupamento $k-$médias é iniciado pela escolha aleatória de k elementos que serão definidos como centroides. Um fato importante é que a estratégia de agrupamentos é diretamente impactada pelo valor de k (o número de centroides escolhidos). Na metodologia CAM, este valor é escolhido de forma completamente arbitrária.

Já a metodologia posterior, o DDCAM [1], utiliza do clássico método de agrupamento de Ward para a geração dos centroides iniciais do método $k-$médias. Esse cuidado é suficiente para garantir que, para o mesmo conjunto de dados, em duas realizações distintas, sempre será obtida a mesma partição.

Posterior ao particionamento dos dados, o objetivo passa a ser estabelecer a distância entre os centroides dos agrupamentos e a mediana referente ao conjunto de dados. O interesse está em verificar se existem agrupamentos suficientemente afastados da mediana por meio de distância euclidiana. Quanto mais distante um centroide de um determinado agrupamento estiver, em relação à mediana, maior o potencial para que os elementos do agrupamento deste referido centroide se comportem como valores *outliers*.

O método CAM considera o desvio padrão (s_c) entre os centroides de agrupamentos e a mediana ($\tilde{X}$) do conjunto completo de dados para detectar discrepâncias multivariadas, caso à distância euclidiana entre o centroide do agrupamento e a mediana exceder ao limiar de $2{,}5 \times s_c$, esses grupos serão considerados compostos por valores *outliers*.

A abordagem DDCAM [1] é proposta como uma evolução do método CAM, o procedimento é executado em 4 estágios iniciais, denominados:

- estimação do valor δ;
- processo de refinamento – I;
- processo de refinamento – II;
- busca pelo valor adequado k.

Após a execução destes estágios, o procedimento CAM com a parametrização estabelecida pelos estágios anteriores é conduzido.

3.1 Estimação do valor δ

Em um problema no qual se busca a detecção de possíveis valores *outliers*, uma premissa básica é determinar uma quantidade máxima de valores parece ser admissível enquanto valores extremos. Dito de outra forma, parece razoável afirmar que em um contexto usual, admitir que uma proporção muito elevada de elementos possam ser definidos por *outliers* é inadequado. Se essa proporção é muito elevada, significa que são valores não raros, ou seja, não condizentes com a definição usual de valores extremos. O método DDCAM nomina tal limiar de proporção por δ e apresenta uma estratégia de determinar uma estimativa eficiente para δ através do próprio conjunto de dados. O principal objetivo desta etapa é estimar δ, o que corresponde ao problema de delimitar, qual a quantidade de *outliers* parece ser aceitável para a base de dados. A análise univariada do conjunto de dados será utilizada para fornecer um estimador para δ.

Sejam $\bar{X}_i$ e s_i, respectivamente, a média e o desvio-padrão amostrais pertencente ao vetor $(x_{1i}, x_{2i}, \ldots, x_{ni})$ de n observações da i-ésima variável sob investigação. O estimador $\hat{\xi}_i$ é definido como a proporção de observações do vetor $(x_{1i}, x_{2i}, \ldots, x_{ni})$ que está a uma distância maior que η desvios padrão amostrais s_i da média $\bar{X}_i$.

Posteriormente é definido o estimador $\hat{\delta}$, dado pela equação 3.1, que toma o máximo entre as estimativas univariadas $\hat{\xi}_i$. Trata-se de uma estratégia conservadora, seu conservadorismo está associado em sempre majorar estimativas univariadas, o que conduz para admitir um volume mais flexível para os possíveis valores *outliers*.

$$\hat{\delta} = \max_{\substack{1 \leqslant i \leqslant p \\ i \in \mathbb{N}}} (\hat{\xi}_i) \ . \tag{3.1}$$

3.1.1 Processo de Refinamento - I

Nesta etapa, é necessário definir o número k de agrupamentos. O DDCAM propõe investigar diversos valores possíveis, até um valor máximo denominado k_{max} definido pela divisão da quantidade total dos dados por log dessa quantidade, ou seja, o número máximo de grupos permitidos pelo método será igual ao valor atribuído para k_{max}. Dessa forma, inicialmente o método pressupõe que o valor adequado de k seja um valor pertencente ao conjunto $\mathcal{K} = \{2, 3, \ldots, k_{max}\}$

O procedimento começa por investigar o agrupamento gerado para $k = 2$ e posteriormente verifica se o menor (em quantidade de elementos) grupo possui no máximo $\hat{\delta} \times n$ elementos. Se essa condição é aceita, o agrupamento com $k = 2$ é dito ser um agrupamento válido, caso contrário, o agrupamento com $k = 2$ será dito inválido e não será investigado. O procedimento descrito será repetido para $k = 3$ e assim subsequencialmente até $k = k_{max}$. Ao final desse processo, haverá um conjunto $\mathcal{K}$ refinado, aqui nominado $\mathcal{K}_1$, tal que $\mathcal{K}_1 \subseteq \mathcal{K}$.

3.1.2 Processo de Refinamento - II

Após o primeiro refinamento, um segundo refinamento é realizado. Para cada valor $k \in \mathcal{K}_1$, é verificado se existe algum agrupamento cujo centroide esteja afastado da mediana mais que $\phi \times s_c$ (o valor ϕ é pré-fixado pelo usuário do método) e que **simultaneamente** possua no máximo $\hat{\delta} \times n$ elementos. Se essa condição não for satisfeita, esse valor k também deverá ser desprezado da investigação. Assim, será produzido um segundo conjunto de valores k ainda mais refinado, denominado $\mathcal{K}_2$ com $\mathcal{K}_2 \subseteq \mathcal{K}_1 \subseteq \mathcal{K}$.

Após o primeiro processo de refinamento, se $\mathcal{K}_1 = \emptyset$, então o método conclui que não existem observações *outliers*. Analogamente, se $\mathcal{K}_1 \neq \emptyset$, mas $\mathcal{K}_2 = \emptyset$ então o método também conclui que não existem valores *outliers* no conjunto de dados investigado.

3.1.3 Busca pelo Valor Adequado k

A existência de possíveis valores *outliers*, ou seja, $\mathcal{K}_2 \neq \emptyset$ demanda a escolha pelo valor $k \in \mathcal{K}_2$ mais adequado para o procedimento de detecção de *outliers*. Visto de outra forma, procurar a escolha k capaz de fornecer os resultados mais adequados no procedimento de detecção. A escolha do valor mais adequado para k proposta no método DDCAM se baseia no critério de informação Bayesiano (BIC).

O critério de informação Bayesiano (BIC), também conhecido como critério de informação de Schwarz (1978) [26], é uma medida bastante difundida para seleção de modelos. Essa medida penaliza modelos mais complexos, com excesso de parâmetros a fim de evitar a possibilidade de sobreajuste. Este critério baseia-se na função de verossi-

milhança e o modelo escolhido é aquele com menor valor de BIC. Dada a padronização prévia dos dados envolvidos, a expressão $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\boldsymbol{x}_{ij}-\boldsymbol{c}_j)'(\boldsymbol{x}_{ij}-\boldsymbol{c}_j)$ representa a soma de quadrados residual (RSS) e o valor do critério de informação Bayesiano pode ser reescrito pela equação 3.2:

$$BIC(k) = RSS + \log(n) \times k \times p\ . \tag{3.2}$$

Em suma, quando $\mathcal{K}_2 \neq \emptyset$, o valor k é definido pork^*, o valor que minimiza o critério de informação Bayesiano restrito aos valores k que satisfazem as condições de $\mathcal{K}_2$, ou seja, dado pela equação 3.3:

$$k^* = \arg\min_{k\in\mathcal{K}_2} BIC(k)\ . \tag{3.3}$$

4 Resultados Alcançados

Os procedimentos de detecção de valores *outliers* tem sido usados para detectar padrões anômalos em diversos setores. Detecção de fraudes financeiras, controle de processos produtivos, entre outros. Por outro lado, a investigação esportiva tem investido cada vez mais em procedimentos estatísticos, os métodos de detecção de valores *outliers* também podem agregar informações nestes propósitos. Verificar se padrões anômalos em competidores ou equipes podem ser usados na detecção de fraudes, como uso de dopagem e outros tipos de contravenções são bastante relevantes. Outra frente, é a detecção de padrões anômalos em determinadas variáveis esportivas que possa identificar a capacidade de oferecer um melhor desempenho. Em algumas situações, estas identificações podem ocorrer até mesmo antes do resultado final e determinar alguma correção para o esportista ou a equipe envolvida.

4.1 População de Dados do Estudo

Particularmente o basquete é um dos esportes que tem demonstrado interesse crescente pelas informações estatísticas. O campeonato de maior atrativo do basquete é a liga dos Estados Unidos (inclui também equipes do Canadá). Essa liga é mundialmente conhecida como NBA (*National Basketball Association*). Este estudo coletou dados de diversas variáveis inerentes ao jogo de basquete nas cinco últimas temporadas do campeonato da NBA.

O campeonato é composto por uma longa fase inicial de confrontos entre as equipes, denominada temporada regular e posteriormente uma fase decisiva denominada pós-temporada ou *playoffs*. Os estudos de desempenho das equipes, muitas vezes, se baseiam na temporada regular. Isso porque as equipes praticam o mesmo número de partidas sob iguais condições, o objetivo máximo nessa etapa é alcançar a classificação para os *playoffs*. A avaliação de estatísticas da pós-temporada é quase intratável para comparações maiores entre equipes, uma vez que nem todas as equipes participam e os participantes são eliminados ao longo da disputa (são confrontos eliminatórios).

Diante disso, este estudo analisou dados apenas das temporadas regulares. Diversas variáveis foram avaliadas entre as equipes disputantes e as variáveis foram divididas em dois grupos, variáveis de ataque e variáveis de defesa. As variáveis avaliadas como variáveis de ataque foram: pontos convertidos (PTS), média de arremessos tentados por jogo (FGA), percentual de arremessos certos (FG%), média de arremessos de três pontos tentados (3PA), percentual de arremessos de três pontos certos (3P%), média de lances livres tentados por jogo (FTA), percentual de lances livres certos (FT%),

número de assistências por jogo (AST) e número de erros em posse de bola por jogo (TO). Já as variáveis avaliadas como variáveis de defesa foram: número de rebotes defensivos (DR), número total de rebotes (REB), roubos de bola por jogo (STL), tocos por jogo (BLK), número de faltas por jogo (PF). Inicialmente, uma análise unidimensional descritiva acerca das variáveis de estudo foi conduzida.

O basquete é um esporte que foi criado em 1891 por um pastor canadense presbiteriano e também professor de educação física da Associação Cristã de Moços, o senhor James Naismith. O esporte é mundialmente conhecido como *basketball*, que em tradução livre significa bola no cesto. Trata-se de um dos esportes mais populares do mundo. O jogo é disputado por duas equipes de cinco jogadores que tem por objetivo central, fazer uma bola passar por dentro do cesto disposto nas extremidades do campo. O conjunto de regras é bastante detalhista e não é objeto central desse estudo, mas não é objeto dessa investigação, diversas referências detalham bem as regras, como, por exemplo, em Phelps (2011) [27].

4.2 Análise Descritiva de Dados

Particularmente aqui, trataremos como já mencionado anteriormente, de dados específicos da liga de basquete dos Estados Unidos, a NBA. Os dados de desempenho das equipes na liga NBA são dados de domínio público, podem ser obtidos em diversos caminhos eletrônicos. Particularmente para este estudo, o *dataset* foi produzido através da metodologia *web scraping* no endereço eletrônico `https://www.espn.com.br/nba/estatisticas/time/_/temporada/` por meio do software estatístico R [28]. O conjunto de dados foi composto pelos dados das últimas cinco temporadas regulares disputadas na liga NBA, as temporadas 17/18, 18/19, 19/20, 20/21 e 21/22.

4.2.1 Variáveis de Ataque

4.2.1.1 Pontos Convertidos

Os pontos convertidos definem o resultado final para cada uma jogada ofensiva executada pelas equipes que em disputa de um jogo de basquete. Obter uma pontuação no basquete significa acertar a bola na cesta, esta pontuação pode ter um valor de três ou dois pontos a depender de onde o jogador arremessou a bola. A equipe vencedora é aquela que obtiver a maior pontuação ao fim da partida. Diante disso, as estatísticas ligadas ao número de pontos obtidos por cada equipe, são estatísticas ligadas à ofensividade das equipes.

A Figura 1 avalia os pontos convertidos ao longo das temporadas em estudo para todas as equipes disputantes do torneio. A representação divide por temporadas,

além disso, para cada temporada, estão separadas as equipes que se classificaram para a fase decisiva de *playoffs* e as equipes que não se classificaram para a fase derradeira do torneio.

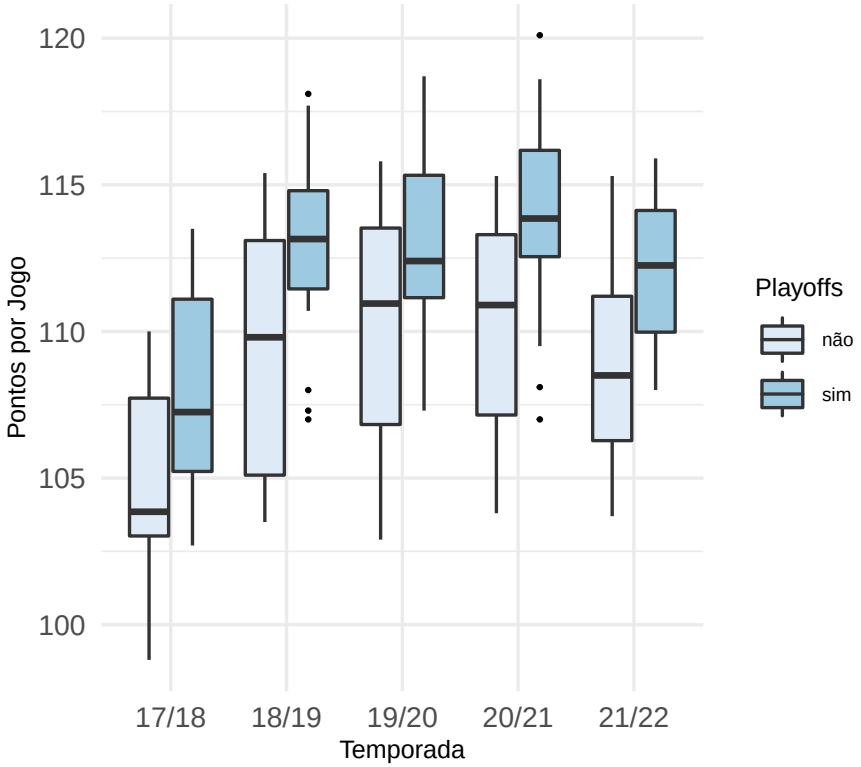

Figura 1 – Boxplot dos pontos convertidos por jogo.

De forma objetiva, observa-se um crescimento da pontuação em todas as temporadas das equipes classificadas aos *playoffs* com respeito às demais. Com exeção da temporada 17/18, nota-se uma redução no intervalo interquartílico dos boxplots o que aparentemente demostra um maior equilíbrio entre as equipes que passaram aos *playoffs*. Este fato faz com que equipes que pontuaram menos apareçam destacadas como *outliers* na cauda inferior dos boxplots. Neste contexto, é possível observar três equipes na temporada 18/19 e duas equipes na temporada 20/21. O equivalente vale para a cauda superior em que destacam-se uma equipe na temporada 18/19 e outra na temporada 20/21. Na Tabela 1 estas equipes *outliers* são apresentadas nominalmente.

Tabela 1 – Equipes *outliers* verificadas univariadamente para a variável Pontos por jogo.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
18/19	Milwaukee Bucks*	Indiana Pacers* Orlando Magic* Detroit Pistons*
20/21	Milwaukee Bucks*	Miami Heat* New York Knicks*

* equipes classificadas aos *playoffs*.

Na investigação dessa variável, *outliers* nos extremos superiores são provenientes de valores que estão a dois desvios padrão acima da média e nos extremos inferiores são provenientes de valores que estão abaixo de dois desvios padrão da média. Esse procedimento será o mesmo ao longo do estudo das demais variáveis envolvidas nesse estudo.

4.2.1.2 Média de Arremessos Tentados

Como discutido anteriormente, o objetivo de uma ação ofensiva é a obtenção dos pontos. Para tanto, o arremesso, ou chute na terminologia usual dos jogadores de basquete, é preponderante. O arremesso é a ação final de uma jogada ofensiva, é o momento no qual, ao realizar um movimento coordenado de braços e pernas, o jogador lança a bola em direção à tabela, com o intuito de pontuar no jogo. A média de arremessos e composta pelo número de arremessos dividido pelo número de jogos de cada equipe.

No basquete existem 4 tipos (mecânicas) de arremesso: a bandeja, o lance livre, o *jump*, e a enterrada. A bandeja, que é um dos mais comuns, é considerada um dos formatos de arremesso de mais fácel execução dentre os arremessos no basquete. Seu nome decorre do movimento manual de segurar a bola próxima à cesta semelhante ao movimento de carregar a bola em uma bandeja. O lance livre, arremesso parado e sem ação dos oponentes, essa situação de arremesso é resultante da marcação de faltas durante um movimento de arremesso. O atleta que sofreu a falta tem direito a cobrar o lance livre através de um arremesso executado de uma linha a 4,6 metros da tabela, cada lance livre vale um ponto. O *jump* no qual ao chegar a um ponto de equilíbrio, o jogador salta e, no ponto mais alto de sua trajetória no ar, solta a bola com uma das mãos. A enterrada é vista com um dos lances mais impactantes do basquete, ela acontece quando o jogador salta em direção à tabela e coloca a bola diretamente dentro do aro, se apoiando no mesmo, com extrema firmeza.

A média de arremessos nesse estudo, é composta pelo soma do número de arremessos durante a temporada dividido pelo número de jogos de cada equipe durante a temporada. A Figura 2 avalia essa variável ao longo das temporadas em estudo. Novamente nessa análise, os gráficos separam as equipes entre equipes classificadas aos *playoffs* e equipes não classificadas, este procedimento será repetido na análise das demais variáveis dessa investigação. É possível observar que as equipes que passaram aos *playoffs*, em média, obtiveram um rendimento bem similar aos times que não foram aos *playoffs*. Como esta variável representa a média de arremessos, que obtiveram pontuação ou não, trata-se de uma variável que mensura mais o volume de ataque (ofensividade) que propriamente a produtividade desse ataque. Uma particularidade fica por conta do grande número de *outliers* principalmente nas temporadas 17/18 e

20/21.

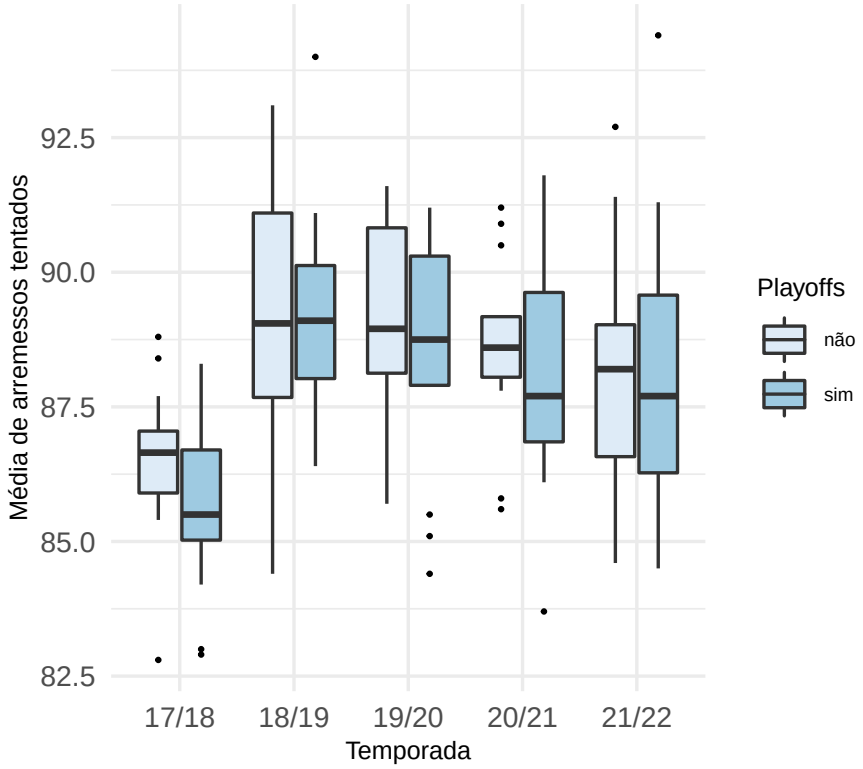

Figura 2 – Boxplot da média de arremessos tentados.

As equipes *outliers* que foram verificadas na análise gráfica da Figura 2 estão apresentadas nominalmente na Tabela 2.

Tabela 2 – Equipes *outliers* verificadas univariadamente para a variável Média de arremessos tentados.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Los Angeles Lakers Chicago Bulls	Memphis Grizzlies Milwaukee Bucks* Utah Jazz*
18/19	Oklahoma City Thunder*	
19/20		Miami Heat* Utah Jazz* Oklahoma City Thunder*
20/21	Indiana Pacers* Minnesota Timberwolves* San Antonio Spurs*	Detroit Pistons Cleveland Cavaliers Miami Heat*
21/22	San Antonio Spurs* Memphis Grizzlies	

* equipes classificadas aos *playoffs*.

4.2.1.3 Percentual de Arremessos Certos

Esta variável mede quantitativamente o percentual de arremessos que efetivamente foram convertidos em pontuação. A porcentagem de aproveitamento desses arremessos mostra quão eficiente foi a equipe ou até mesmo um jogador em particular quando a análise é individualizada. É importante observar que, equipes que tentam muitos arremessos acabam por perder um pouco de sua efetividade em acertos por arremessos. Porém, existem dois caminhos, muitos arremessos para ter margens para os erros, ou menos arremessos, para buscar arremessos mais seguros e certeiros. Cada equipe acaba por estabelecer suas estratégias de acordo com as valências individuais de seus atletas.

A Figura 3 avalia essa variável ao longo das temporadas em estudo. É possível observar a superioridade das equipes que se classificaram aos *playoffs*, em todas as temporadas essas equipes obtiveram um maior aproveitamento para o percentual de arremessos corretos. De uma forma direta, uma efetivididade maior neste quesito significa uma maior probabilidade de um time passar para os *playoffs*. Uma análise de correlação indica uma correlação positiva da ordem de 0,62 desta variável com a variável Pontos convertidos.

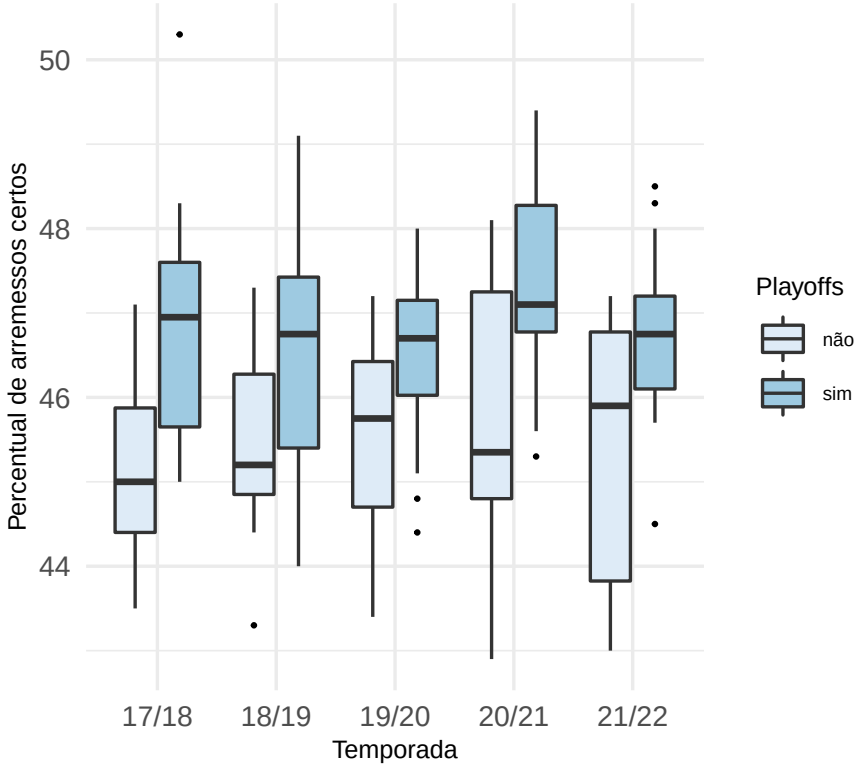

Figura 3 – Boxplot do percentual de arremessos certos.

Outra observação decorrente da Figura 3 é o baixo número de *outliers* verificados para a variável, e estes estão presentes em maior número na cauda inferior da distribuição. Na Tabela 3 observa-se nominalmente estas equipes.

Tabela 3 – Equipes *outliers* verificadas univariadamente para a variável Percentual de arremessos certos.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Golden State Warriors*	
18/19		New York Knicks
19/20		Brooklyn Nets* Orlando Magic*
20/21		Portland Trail Blazers*
21/22	Phoenix Suns* Denver Nuggets*	Toronto Raptors*

* equipes classificadas aos *playoffs*.

4.2.1.4 Média de Arremessos de Três Pontos Tentados

No basquete, a linha dos três pontos é uma linha em formato de arco presente na quadra. A distância da cesta até a linha de três pontos varia de acordo com o nível da competição: na NBA, o arco fica a 7,24 m do centro da cesta. Para um arremesso de três pontos valer o jogador tem que estar antes dessa linha, este arremesso e considerado o mais difícil e exige uma melhor mecânica de arremesso dos jogadores (veja a figura 4).

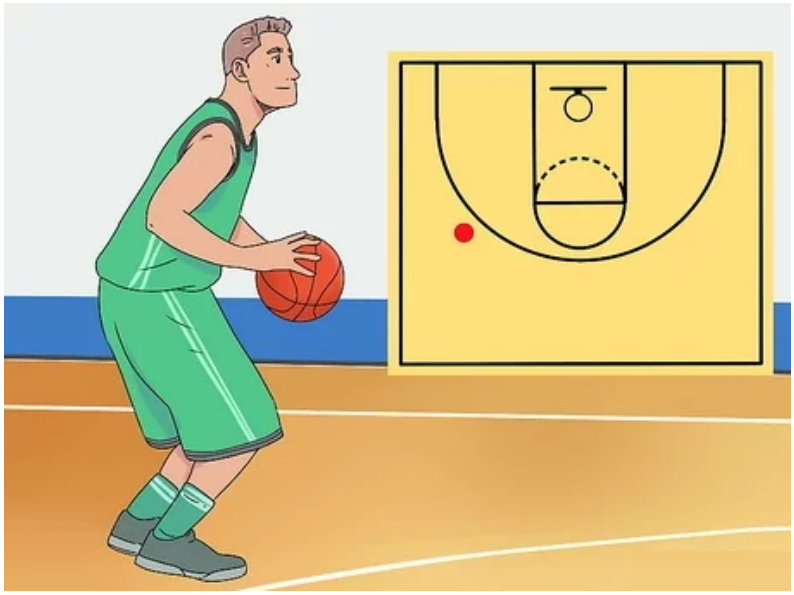

Figura 4 – Arremesso de três pontos.
fonte: `https://pt.wikihow.com/Jogar-Basquete`

A média de arremessos de três pontos tentados é composta pela soma de tentativas de arremessos de três pontos executado dividido pelo número de jogos realizado por cada equipe.

A Figura 5 avalia essa variável ao longo das temporadas em estudo. A análise gráfica mostra que as equipes que passaram aos *playoffs*, em média, obtiveram um aproveitamento ligeiramente superior ao desempenho verificado para as equipes que não foram promovidas aos *playoffs*. Como esta variável representa a média de arremes-

sos, novamente é preciso relativizar. Estão considerados os arremessos que obtiveram pontuação ou não, trata-se de uma métrica de ofensividade e não de produtividade.

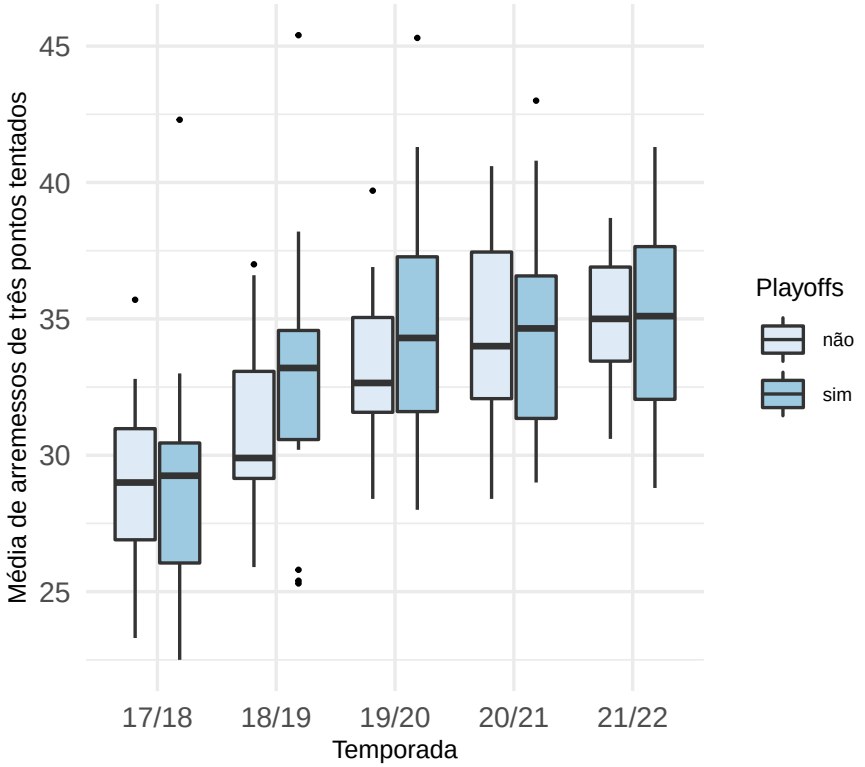

Figura 5 – Boxplot da média de arremessos de três pontos tentados

Os arremessos não ocorrem de forma completamente casual, as equipes que arremessam mais vezes da linha de três pontos, usualmente tem isso por estratégia. Trata-se de um planejamento prévio para maximizar as virtudes de seus próprios atletas e também minimizar os talentos das equipes adversárias. Na Tabela 4 as equipes *outliers* são representadas.

Tabela 4 – Equipes *outliers* verificadas univariadamente para a variável Média de arremessos de três pontos tentados.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Brooklyn Nets Houston Rockets*	
18/19	Atlanta Hawks Houston Rockets*	LA Clippers San Antonio Spurs Indiana Pacers
19/20	Minnesota Timberwolves Houston Rockets*	
20/21	Utah Jazz*	

* equipes classificadas aos *playoffs*.

Ao verificar uma correlação direta da variável Média de arremessos de três pontos tentados com a variável Pontos convertidos, observou-se um valor de correlação da ordem de 0,44. No gráfico observou-se um grande número de *outliers* na direção da cauda superior da distribuição, nota-se também que estes *outliers* encontram-se bem afastados da média.

4.2.1.5 Percentual de Arremessos de Três Pontos Certos

Esta variável mensura quantitativamente o percentual de arremessos executados da linha de três pontos que efetivamente resultaram em cesta, ou seja, efetivamente convertidos e resultantes de três pontos. A porcentagem de aproveitamento desses arremessos mostra a eficiência de uma equipe nos lances de três pontos, uma métrica importante dado o alto valor em pontos destes arremessos.

A Figura 6 avalia essa variável ao longo de todas temporadas deste estudo. O gráfico mostra que as equipes que passaram aos *playoffs* obtiveram um aproveitamento superior aos times que não foram aos *playoffs*. Este fato foi observado para todas as temporadas sob investigação.

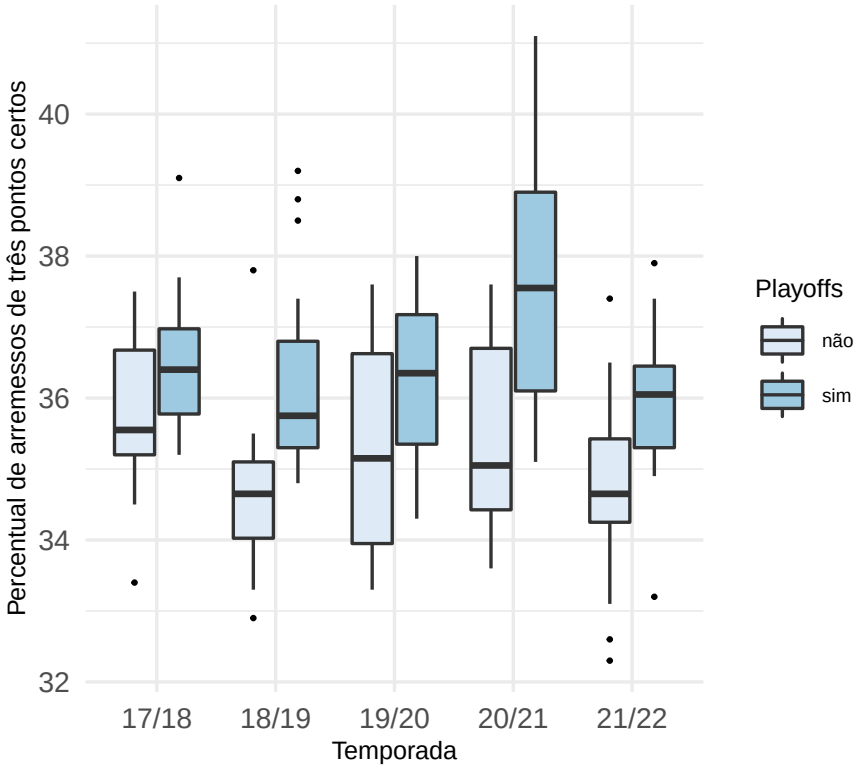

Figura 6 – Boxplot do percentual de arremessos de três pontos certos.

Importante apresentar destaque para as temporadas 18/19 em que três equipes *outliers* conseguiram um aproveitamento superior aos trinta e oito por cento, e temporada 20/21 em que a mediana de sucesso nos arremessos de três pontos ficou próxima

a trinta e sete por cento e o primeiro quartil já foi superior aos trinta e seis por cento. A Tabela 5 apresenta nominalmente as equipes *outliers*.

Tabela 5 – Equipes *outliers* verificadas univariadamente para a variável Percentual de arremessos de três pontos certos.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Golden State Warriors*	Phoenix Suns
18/19	Sacramento Kings* Golden State Warriors* LA Clippers San Antonio Spurs*	Phoenix Suns
21/22	LA Clippers Miami Heat*	Detroit Pistons Oklahoma City Thunder New Orleans Pelicans*

* equipes classificadas aos *playoffs*.

4.2.1.6 Média de Lances Livres Tentados

No basquete, o lance livre ocorre quando o jogador sofre falta no ato de arremessar. O árbitro então assinala a infração e determina a cobrança de lances livres. A equipe que sofreu a falta é beneficiada com o arremesso do lançamento livre, em que o jogador que sofreu a falta, obrigatoriamente, deve arremessar a bola na cesta sempre da mesma posição, atrás de uma linha, chamada de "*foul line*", que fica a uma distância de 5,8 metros da linha de fundo e 4,6 metros da tabela, e dentro de um tempo de 5 segundos (veja a figura 7).

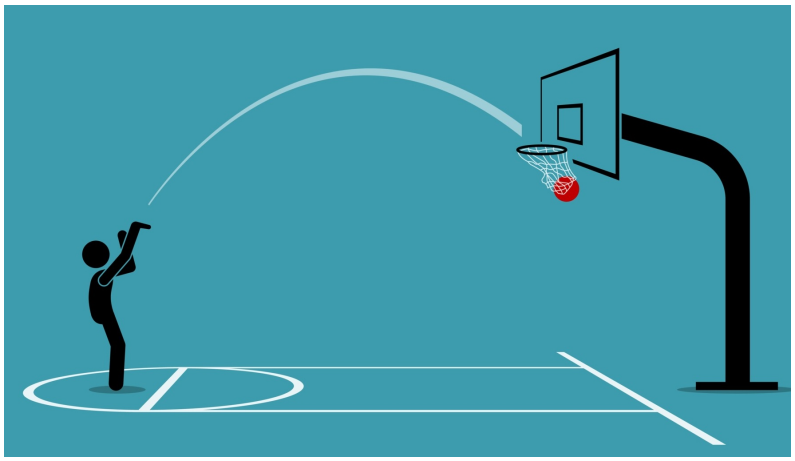

Figura 7 – Lance livre.
fonte: `https://pt.vecteezy.com/arte-vetorial/2225455-homem-atirando-uma-bola-de-basquete-em-um-aro-e-marcando-a-linha-de-lance-livre`

No ato do lance livre, os demais jogadores colocam-se nos espaços da área restritiva para a disputa dos rebotes. Caso o jogador acerte a cesta, ele marca 1 ponto. Se o último lance livre de uma sequência não for convertido, jogadores de ambas equipes

podem disputar o rebote pela posse de bola. A Figura 8 avalia essa variável ao longo das temporadas em estudo.

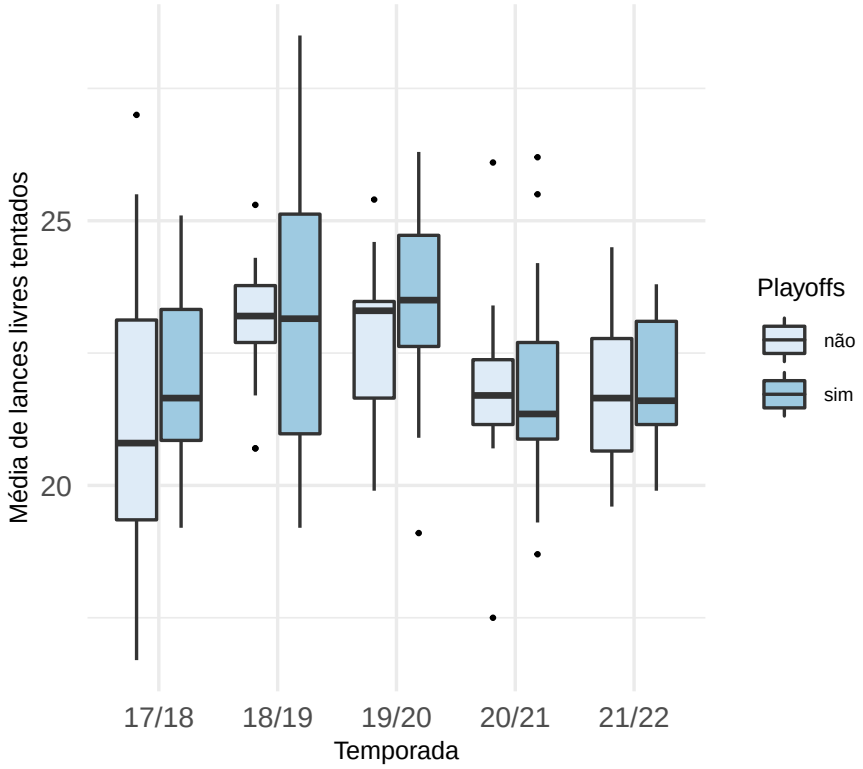

Figura 8 – Boxplot da média de lances livres tentados.

Observa-se um equilíbrio entre os times que passaram aos *playoffs* e os que não se classificaram. Em todas as temporadas, as medianas dos dois grupos, classificados e não classificados aos *playoffs* ficaram bem próximas. Durante a temporada 17/18 os times que não se classificaram aos *playoffs* tiverem uma grande variação para a média de arremessos tentados, fato também observado para os times que se classificaram aos *playoffs* na temporada 18/19. Para as temporadas em que a amplitude interquartílica é menor, nota-se o aparecimento de algumas equipes textitoutliers, tanto na cauda superior quanto na cauda inferior das distribuições. A Tabela 6 mostra nominalmente estas equipes.

Tabela 6 – Equipes *outliers* verificadas univariadamente para a variável Média de lances livres tentados.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Charlotte Hornets*	
18/19	Dallas Mavericks*	Cleveland Cavaliers*
19/20	Minnesota Timberwolves*	Indiana Pacers
20/21	New Orleans Pelicans Washington Wizards* Philadelphia 76ers*	Chicago Bulls Phoenix Suns*

* equipes classificadas aos *playoffs*.

4.2.1.7 Percentual de Lances Livres Certos

Esta variável mede quantitativamente o percentual de lances livres que efetivamente foram convertidos em pontos. Quanto maior esse número, maior e melhor é a eficiência de uma equipe ou mesmo um jogador neste fundamento do basquete. A Figura 9 avalia essa variável ao longo das temporadas em estudo.

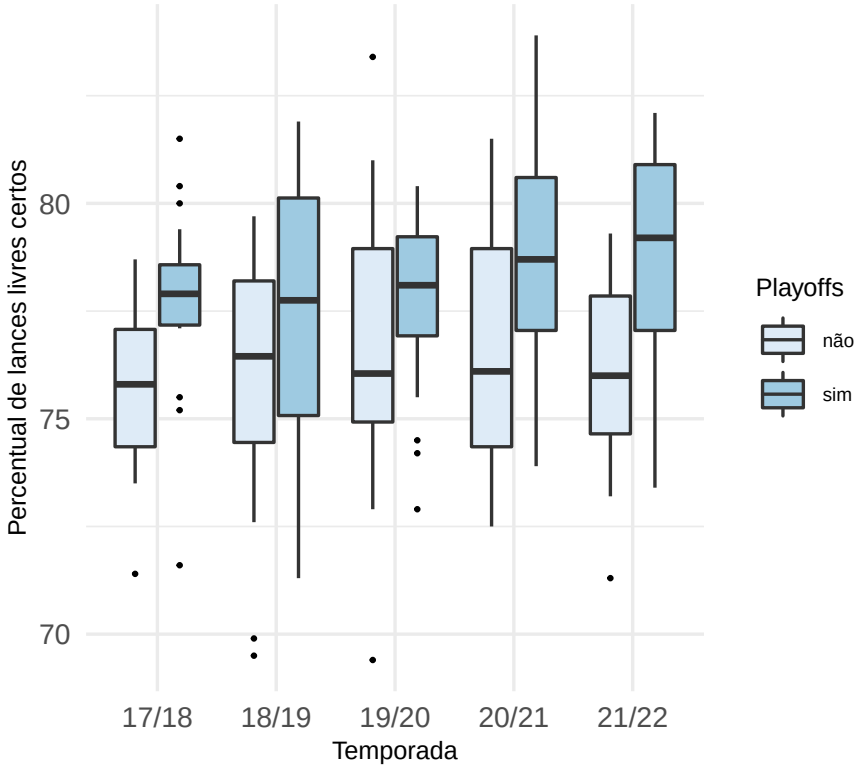

Figura 9 – Boxplot do percentual de lances livres certos.

É possivel observar a discrepância das equipes que não passaram aos *playoffs* no aproveitamento dos lances livres. As equipes que não passaram aos *playoffs* obtiveram um aproveitamento médio próximo aos 75%, enquanto as equipes que foram aos *playoffs* tiveram aproveitamento médio entre 78% e 79%.

Além disso, é possível observar um grande número de *outliers* no extremo inferior das distribuições. Embora os aproveitamentos observados ainda continuem perto da casa dos 70%. O destaque para os *outliers* no extremo inferior se deve ao fato que toda vez que um jogador erra o lance livre, ele oferece à equipe adversária a possibilidade de rebote e consequentemente um contra-ataque.

Todas as equipes *outliers* verificadas na análise gráfica da Figura 9 estão representadas nominalmente na tabela 7.

Tabela 7 – Equipes *outliers* verificadas univariadamente para a variável Percentual de lances livres certos.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Golden State Warriors* Minnesota Timberwolves* Portland Trail Blazers*	Los Angeles Lakers Philadelphia 76ers* Oklahoma City Thunder* Miami Heat*
18/19		Los Angeles Lakers Miami Heat
19/20	Phoenix Suns*	New York Knicks Milwaukee Bucks* Los Angeles Lakers* Brooklyn Nets*
21/22		Houston Rockets

* equipes classificadas aos *playoffs*.

4.2.1.8 Número de Assistências

A assistência é um passe final de uma jogada que resulta em cesta. Ela é bastante valorizada por contribuir diretamente para a cesta. O jogador que recebe a bola fica em uma posição privilegiada para a execução do arremesso. O número de assistências, entretanto, deve ser observado em conjunto com o de número de *turnovers* (termo usual da NBA), na prática, as perdas de posse de bola, comumente chamados de desperdícios de bola.

O motivo para esse formato de análise é que muitas das assistências tentadas podem ser conseguidas "forçando" passes improváveis. Por outro lado, essa ação forçada pode acabar por gerar também muitos desperdícios de bola que tendem a virar contra-ataques perigosos. Em geral, é muito melhor perder a bola através de um arremesso ruim que bateu no aro do que errando um passe, pois erros de passe viram contra-ataques com maior facilidade, isso se deve ao fato do percurso adversário até a cesta, usualmente ser mais curto nessas situações.

A Figura 10 avalia essa variável ao longo das temporadas em estudo. A análise da imagem gráfica ilustra algum equilíbrio entre as equipes classificadas aos *playoffs* e as demais equipes que não se classificaram. Durante as temporadas 17/18 e 20/21, as equipes que não se classificaram aos *playoffs* foram, em mediana, ligeiramente superiores, enquanto que nas temporadas 18/19, 19/20 e 21/20 as equipes que foram elevadas aos *playoffs* mantiveram a hegemonia no número de assistências. Outro fato apresentado é um número significativo de equipes *outliers* nos dois extremos da distribuição. Isso claramente mostra que algumas equipes foram bem assertivas neste fundamento, enquanto outras ficaram bem aquém da média.

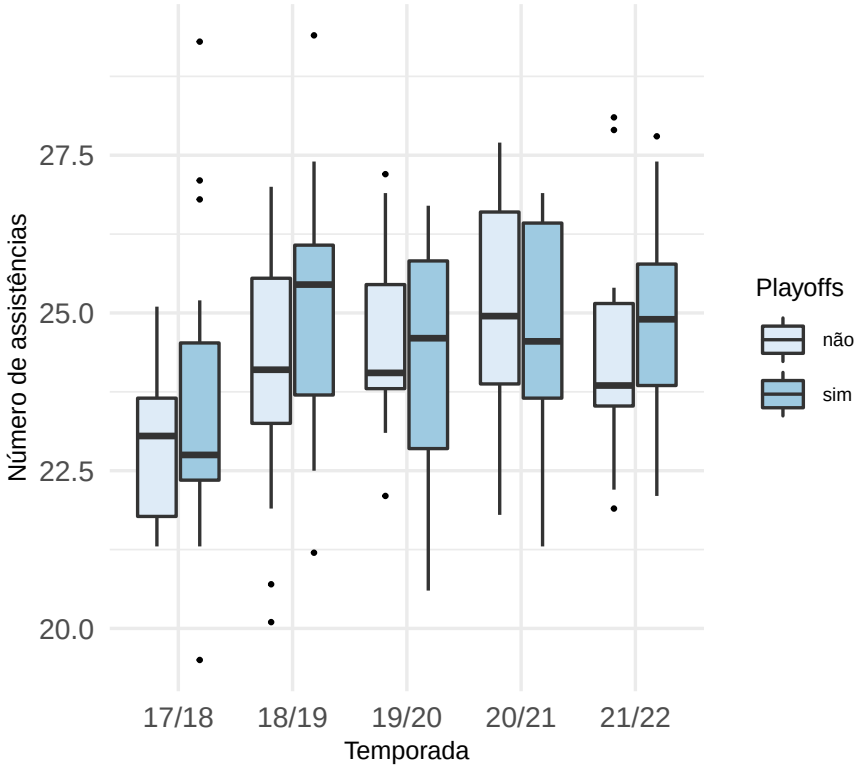

Figura 10 – Boxplot do número de assistências.

Uma curiosidade desta variável vem da correlação com as variáveis Percentual de arremessos certos em que a medida de correlação fica em 0,60 e Percentual de arremessos de três pontos certos em que a correlação verificada é de 0,18, o que deixa subentendido que a maioria das assistências resulta em pontuação simples. A Tabela 8 mostra nominalmente estas equipes *outliers*.

Tabela 8 – Equipes *outliers* verificadas univariadamente para a variável Número de assistências.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Golden State Warriors* New Orleans Pelicans* Philadelphia 76ers*	Portland Trail Blazers*
18/19	Golden State Warriors*	New York Knicks Cleveland Cavaliers Houston Rockets*
19/20	Phoenix Suns	New York Knicks
21/22	Charlotte Hornets San Antonio Spurs Denver Nuggets*	New York Knicks

* equipes classificadas aos *playoffs*.

4.2.1.9 Número de Erros em Posse de Bola

Número de erros em posse de bola, ou popularmente nos Estados Unidos o *turnover*, é um erro, um desperdício de bola, é a perda da posse que não tenha acontecido numa tentativa de arremesso devido a um passe ruim ou má condução da bola, ou mesmo devido a uma falta ofensiva. Como exemplo, um time que cometeu 20 *turnovers* em um jogo é aquele que por 20 vezes cometeu algum tipo de erro e ficou sem a posse de bola, seja por ter jogado a bola para fora, ter estourado o limite de cronômetro sw 24 segundos de posse de bola, seja por ter tido a bola roubada, ou até mesmo por ter cometido uma falta de ataque. De certa forma, um alto índice de *turnovers* pode afetar consideravelmente o desempenho de uma equipe.

A Figura 11 avalia essa variável ao longo das temporadas. O gráfico mostra que as equipes que foram aos *playoffs* em média erraram menos e obtiveram um melhor desempenho comparado com as equipes que não alcançaram a classificação para os *playoffs*. Outro fator que pode ser verificado através da análise gráfica são os *outliers* presentes em todas as temporadas, tanto no extremo inferior quanto no extremo superior da distribuição. Aqui cabe um adendo, os *outliers* no extremo inferior dos gráficos boxplot significam equipes que tiveram um melhor desempenho e foram mais assertivas.

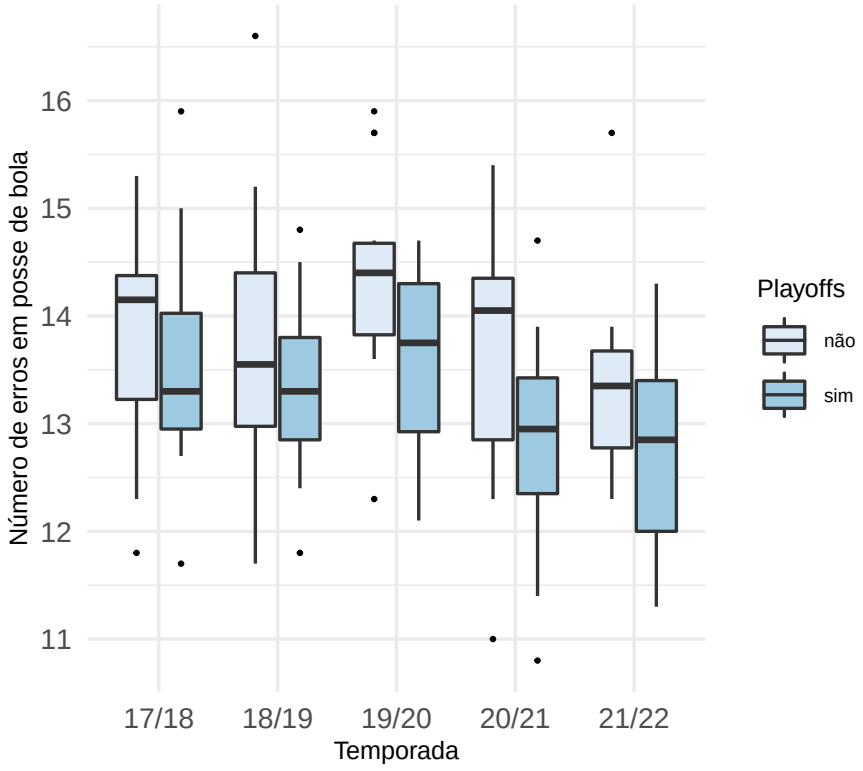

Figura 11 – Boxplot do número de erros em posse de bola

Baseando na análise da Figura 11, as equipes *outliers* verificadas para o número de *turnovers* estão elencadas por temporada na Tabela 9.

Tabela 9 – Equipes *outliers* verificadas univariadamente para a variável Número de erros em posse de bola.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Philadelphia 76ers*	Dallas Mavericks Minnesota Timberwolves*
18/19	Atlanta Hawks Utah Jazz*	San Antonio Spurs*
19/20	New Orleans Pelicans San Antonio Spurs	Cleveland Cavaliers
20/21	Portland Trail Blazers	San Antonio Spurs Los Angeles Lakers*
21/22	Houston Rockets	

* equipes classificadas aos *playoffs*.

4.2.2 Variáveis de Defesa

4.2.2.1 Número de Rebotes Defensivos

Uma das valências mais importantes do jogo de basquete são os rebotes. Dado que o jogo consistena busca pela ação de arremessar a bola em busca da obtenção da cesta, por muitas vezes esse objetivo não é alcançado em sua plenitude, e bola retorna ao jogo e pode ser recolhida por qualquer um dos jogadores. Quando uma tentativa de arremesso não é bem sucedida, jogadores de ambas as equipes buscam retomar a posse de bola a fim de realizar uma nova finalização ou iniciar a armação de uma jogada de ataque, capturar a bola nessas condições caracteriza o rebote no basquete. Usualmente, os atletas são extremamente preparados para executar bem as ações de rebote, dada a enorme importância desse fundamento para o resultado final do jogo de basquete. Os rebotes capturados pela equipe que se encontra em situação de defesa, são ditos rebotes defensivos. Ao passo que os rebotes obtidos pela equipe em situação de ataque são ditos rebotes ofensivos.

A Figura 12 avalia o número de rebotes defensivos para as equipes ao longo de todas temporadas desse estudo. É possível observar que na temporada 17/18, as equipes que se classificaram aos *playoffs* não apresentaram um rendimento extremamente superior às demais equipes. Porém nas demais temporadas essa diferença foi bastante evidente. Apenas nas temporadas 18/19, 19/20 e 20/21 e apenas dentre as equipes classificadas aos *playoffs* foram verificadas equipes *outliers* na direção da cauda superior da distribuição. É possível verificar que os *outliers* das caudas superiores estão bem mais afastados da faixa central que os *outliers* das caudas inferiores. Este efeito de assimetria é mais evidente nas distribuições que incluem somente as equipes classificadas aos *playoffs*.

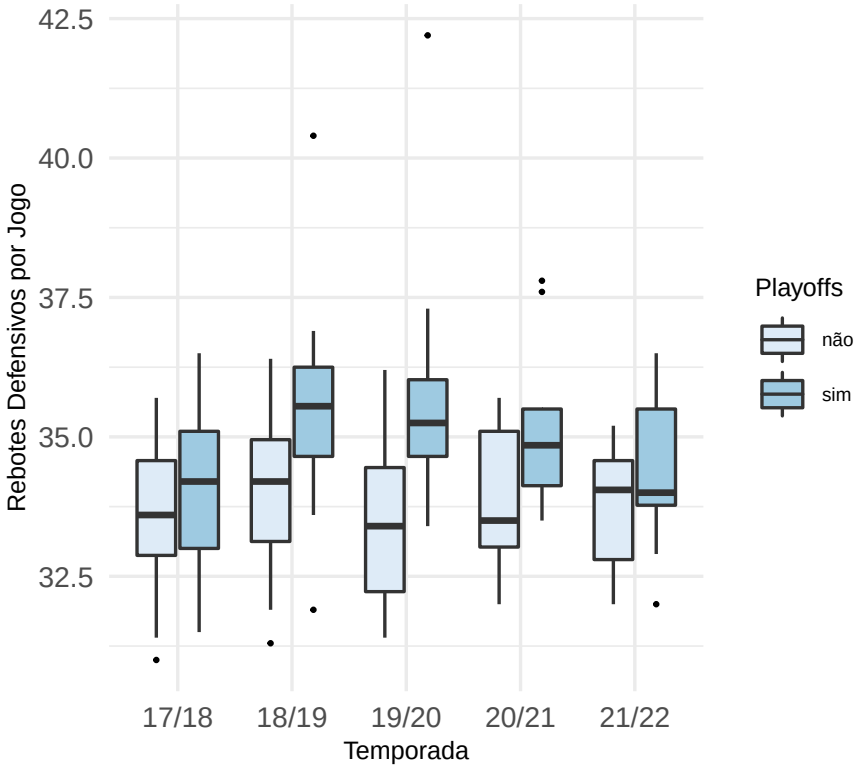

Figura 12 – Boxplot dos rebotes defensivos por jogo.

As equipes *outliers* verificadas através da análise gráfica são apresentadas nominalmente através da Tabela 10.

Tabela 10 – Equipes *outliers* verificadas univariadamente para a variável rebotes defensivos por jogo.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18		Memphis Grizzlies
18/19	Milwaukee Bucks*	Phoenix Suns* Cleveland Cavaliers
19/20	Milwaukee Bucks*	
20/21	Milwaukee Bucks* Utah Jazz*	
21/22		Detroit Pistons*

* equipes classificadas aos *playoffs*.

4.2.2.2 Total de Rebotes

Como já mencionado, o rebote é um dos fundamentos mais importantes do basquete, pois influência diretamente na dinâmica do jogo. Ao conquistar um rebote de ataque, o jogador está geralmente posicionado próximo à tabela, o que facilita a tentativa de arremesso a partir de uma posição bastante privilegiada, já no caso de um

rebote de defesa, o time tem a possibilidade de armar uma jogada de contra-ataque (veja a figura 13).

Figura 13 – Ação de rebote.
fonte: `https://pt.wikihow.com/Jogar-Basquete`

A variável Total de rebotes é a soma dos rebotes defensivos e ofensivos realizados pelas equipes. A Figura 14 ilustra o comportamento dos times ao longo das temporadas para essa variável.

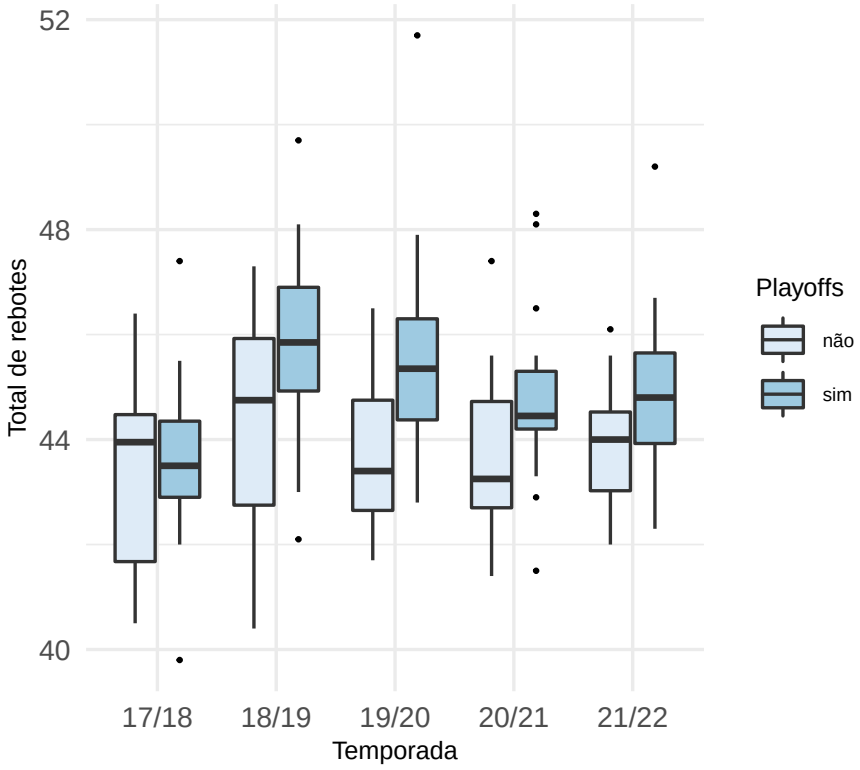

Figura 14 – Boxplot do total de rebotes.

É possível visualizar um melhor aproveitamento das equipes que se classificaram aos *playoffs*, com exceção da temporada 17/18. Também é possível observar equipes *outliers* nos extremos superiores e inferiores. O destaque para os *outiliers* ocorre pelo fato da variável Total de rebotes apresentar uma correlação significativa de 0,49 com a variável Pontos convertidos e uma correlação significativa de 0,53 com a variável Média de arremessos tentados. Isso confirma que apesar de ser uma variável de defesa, possui influência direta na partida, e as equipes que tem o domínio desta variável podem apresentar melhores resultados. A tabela 11 mostra nominalmente as equipes *outliers*.

Tabela 11 – Equipes *outliers* verificadas univariadamente para a variável Total de rebotes.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Philadelphia 76ers*	Milwaukee Bucks*
18/19	Milwaukee Bucks*	Houston Rockets*
19/20	Milwaukee Bucks*	
20/21	New Orleans Pelicans Milwaukee Bucks* Utah Jazz* Memphis Grizzlies*	Phoenix Suns* Miami Heat*
21/22	New York Knicks Memphis Grizzlies*	

* equipes classificadas aos *playoffs*.

4.2.2.3 Número de Tocos

O toco é uma ação de defesa em que um jogador busca impedir o arremesso de um adversário ou desvia o curso de uma bola ainda na parte ascendente do movimento parabólico afim de evitar uma cesta da equipe adversária. (veja a figura 15).

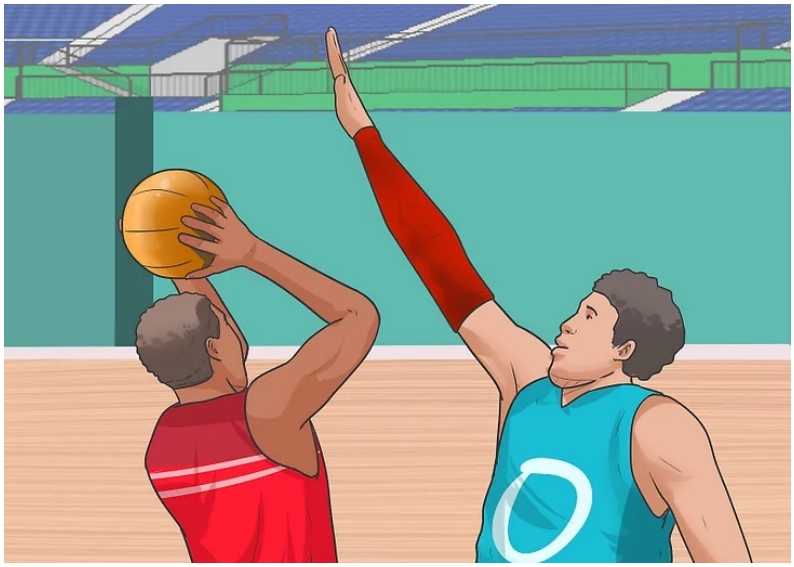

Figura 15 – Ação de toco.
fonte: `https://pt.wikihow.com/Jogar-Basquete`

No movimento de toco, o defensor se projeta verticalmente e, com uma ou duas mãos, bate na bola com força suficiente para alterar a sua direção. A Figura 16 ilustra o comportamento dos times durante as temporadas para a variável Tocos por jogo.

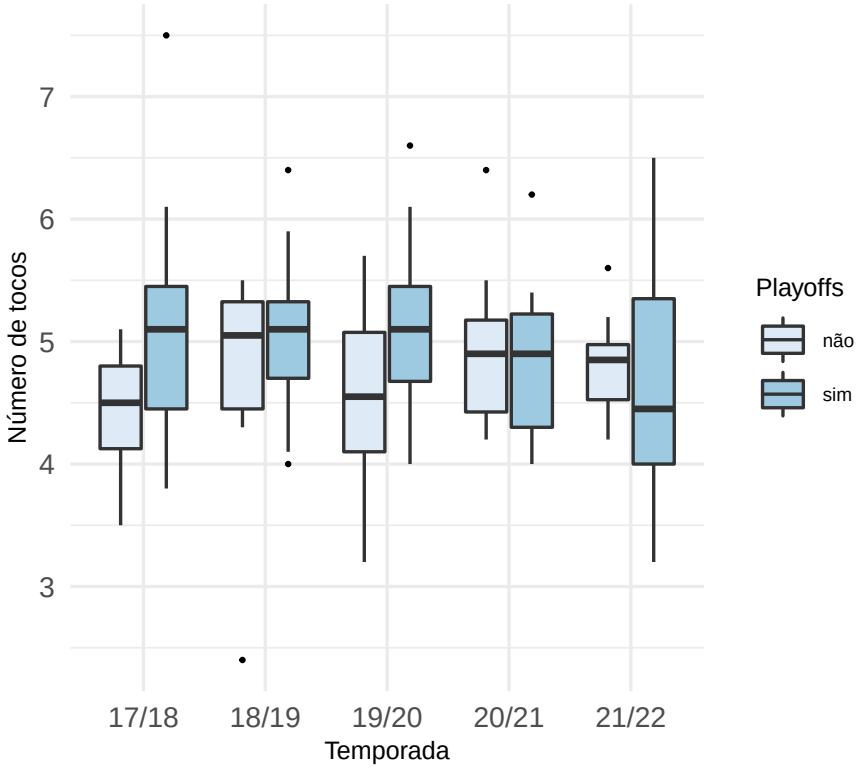

Figura 16 – Boxplot do número de tocos.

A verificação gráfica apresenta um baixo número de equipes *outliers*, a maior parte concentrada no extremo superior da distribuição. A tabela 12 mostra nominalmente estas equipes *outliers*.

Tabela 12 – Equipes *outliers* verificadas univariadamente para a variável Número de tocos.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Golden State Warriors*	
18/19	Golden State Warriors*	Minnesota Timberwolves Detroit Pistons*
19/20	Los Angeles Lakers*	
20/21	Indiana Pacers Philadelphia 76ers*	
21/22	Philadelphia 76ers	

* equipes classificadas aos *playoffs*.

Ainda na análise gráfica da Figura 16, é possível observar um equilíbrio nas temporadas 18/19, 19/20 e 20/21 entre as equipes que foram aos *playoffs* e as que não

se classificaram. Durante a temporada 17/18 os times que foram aos *playoffs* se saíram melhor neste fundamento, já na temporada 21/22 as equipes que não se classificaram obtiveram um aproveitamento ligeiramente superior. Além disso, é notório uma grande amplitude interquatílica na temporada 21/22 para os times que se classificaram aos *playoffs*, o que significa que na temporada 21/22 ocorreu uma grande variabilidade neste fundamento.

4.2.2.4 Número de Faltas Cometidas por Jogo

No basquete existem diferentes tipos de infrações que levam à marcação de faltas pela arbitragem. Os tipos de falta são: pessoais, técnicas, antidesportivas e desqualificantes. A ocorrência de um contato ilegal entre os atletas acarreta na falta pessoal. Quando o comportamento do jogador atrapalha o desenvolvimento da partida, a arbitragem pode pode apontar uma falta técnica. A ocorrência de contato entre atletas de forma alheia aos padrões do jogo, os árbitros aplicam a falta antidesportiva. Por fim, quando há algum ato de violência praticado por um atleta, ou no caso de briga entre dois ou mais atletas pode ser determinada uma falta desqualificante. Durante a partida, cada jogador possui um limite de 5 faltas pessoais. Após o cometimento da quinta falta, o jogador é excluído da partida. Usualmente as faltas como recurso defensivo ao longo do jogo.

Particularmente para o campeonato da NBA, algumas adaptações às regras mundiais são acopladas. As faltas pessoais seguem o mesmo formato, mas são acrescidas as faltas de disputa de posse de bola, ocorrem quando nenhuma das equipes estabeleceu definitivamente a posse de bola. A bola é considerada sem posse efetiva de quaisquer das equipes. Faltas nesse caso envolvem contato físico ilegal ao tentar conseguir a bola de basquete descontrolada. A penalidade para este tipo de falta são os lances livres ou perda de posse de bola. Também são caracterizadas as faltas antidesportivas e desqualificantes, por faltas flagrantes de nível 1 e 2. As faltas flagrantes de nível 1 são as faltas que envolvem contato desnecessário. A penalidade para este tipo de falta é o lance livre para o adversário além da posse de bola. As faltas flagrantes de nível 2 são as faltas com contato desnecessário e excessivo. Os árbitros realizam uma revisão instantânea do jogo (via monitor) para determinar se um ato se qualifica para a falta flagrante de nível 2, e em caso afirmativo, o jogador que fez a falta recebe uma multa financeira e uma expulsão automática do jogo. Além disso, a equipe adversária recebe os lances livres e a posse da bola.

Cada equipe também possui um limite de 5 faltas (faltas coletivas/faltas de equipe) por quarto de jogo. A partir daí, a equipe que sofre a falta tem direito à execução de dois lances livres. A Figura 17 ilustra o comportamento dos times durante as temporadas para a variável Número de Faltas.

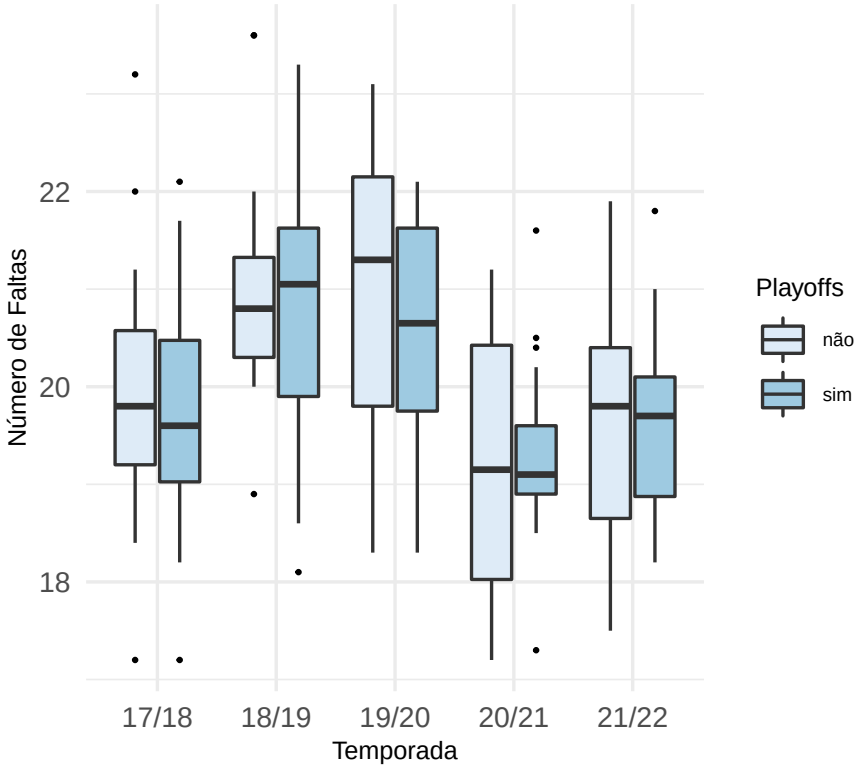

Figura 17 – Boxplot do número de faltas.

A análise gráfica possibilita observar que nas temporadas 17/18, 18/19, 20/21 e 21/22 as equipes que se classificara aos *playoffs* e aquelas que não obtiveram a classificação, em média, cometeram o mesmo número de faltas. Na temporada 19/20 os times classificados se saíram melhor e cometeram menos faltas. O destaque fica para os *outliers* localizados no extremo inferior da distribuição, visto que o maior número de faltas rende a equipe adversária a chance de pontuar mais nos lances livres. A Tabela 13 mostra as equipes *outliers* identificadas.

Tabela 13 – Equipes *outliers* verificadas univariadamente para a variável Número de faltas.

Temporada	Equipes *outliers*	
	Extremos superiores	Extremos inferiores
17/18	Phoenix Suns Memphis Grizzlies Fhiladelphia 76ers*	Charlotte Hornets San Antonio Spurs*
18/19	Phoenix Suns	Orlando Magic San Antonio Spurs*
20/21	Washington Wizards* Boston Celtics* Detroit Pistons*	Milwaukee Bucks*
21/22	Minnesota Timberwolves*	

* equipes classificadas aos *playoffs*.

4.3 Investigação multivariada

Historicamente sempre se discute se o basquete é um jogo de ataque ou um jogo de defesa? Em outras palavras, quem vence mais? As melhores defesas, ou os melhores ataques? Este estudo buscou verificar a existência de *outliers* em estatísticas de defesa e também em estatísticas de ataque. Com o interesse em determinar se mais equipes com defesas discrepantes em relação às demais que usualmente se classificam aos *playoffs*, ou se equipes com ataques discrepantes em relação às demais que usualmente se classificam aos *playoffs*.

A literatura tende a afirmar que o resultado das partidas não está ligado exclusivamente à ofensividade ou a qualidade da defesa. Por outro lado, ao discutir os resultados de forma coloquial, sem uma análise de dados, os envolvidos com o basquete tem uma tendência de acreditar na supremacia das grandes defesas. Essa investigação, como dito outrora, busca verificar através de uma técnica inovadora a possível existência de alguma associação aparente entre as equipes *outliers* no que tange às variáveis de defesa e de ataque com as equipes com melhor desempenho esportivo nos campeonatos.

Os dados analisados se referem ao campeonato da NBA nas temporadas de 2017/18 até 2021/22. Particularmente serão obtidas as equipes *outliers* para o subgrupo das variáveis de defesa e também para o subgrupo das variáveis de ataque. A busca por equipes *outliers* através do banco de dados multivariado será executada através da técnica DDCAM. Uma descrição bastante detalhada da técnica pode ser obtida no Capítulo 3. A Tabela 14 apresenta o resultado do procedimento de detecção de *outliers*.

Tabela 14 – Equipes *outliers* verificadas por temporada da NBA.

Temporada	*outliers* por subgrupo	
	Ataque	Defesa
2017/18	Golden State Warriors*	Milwaukee Bucks* Memphis Grizzlies
2018/19	Philadelphia 76ers* LA Clippers*	Houston Rockets* Phoenix Suns Memphis Grizzlies
2019/20	Golden State Warriors Orlando Magic	Milwaukee Bucks*
2020/21	San Antonio Spurs New York Knicks Orlando Magic	Washington Wizards* Golden State Warriors* Philadelphia 76ers* Memphis Grizzlies*
2021/22	Houston Rockets New York Knicks	Memphis Grizzlies*

* equipes classificadas aos *playoffs*.

É possível observar em uma análise inicial superficial uma maior presença de

equipes classificadas para a disputa dos *playoffs* dentre as equipes *outliers* de defesa que dentre as equipes *outliers* de ataque. Apenas a temporada 2018/19 apresentou uma superioridade clara dos ataques quanto aos classificados para os *playoffs*. Na temporada 2017/18 ocorreu um equilíbrio e nas demais temporadas uma clara superioridade das defesas. A Figura 18 apresenta uma visão desse resultado.

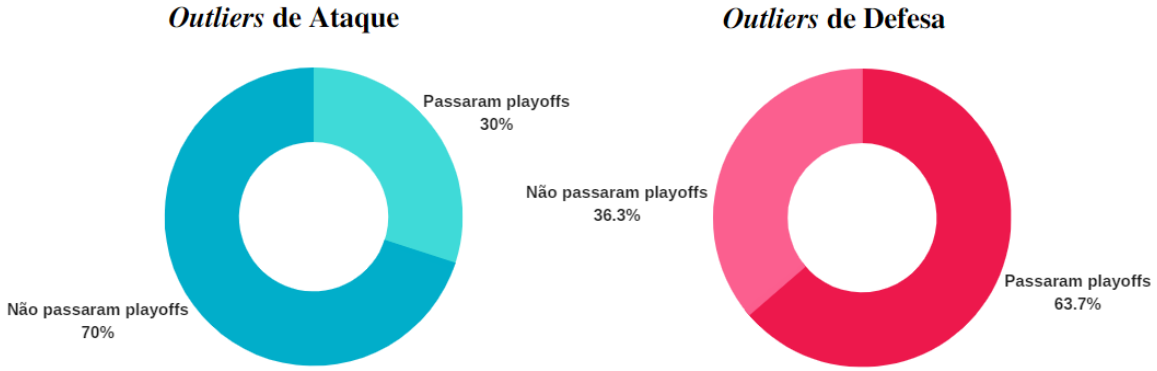

Figura 18 – Relação entre as equipes classificadas aos *playoffs*.

Em uma análise agregada, para as cinco temporadas, 30% das equipes discrepantes em ataque foram aos *playoffs* e 70% não. Por outro lado, 63,6% das equipes discrepantes em defesa foram aos *playoffs*, ao passo que apenas 36,3% das equipes discrepantes em defesa não foram aos *playoffs*.

Obviamente este não é um resultado peremptório para determinar que a defesa é mais importante no jogo de basquete, mas é um indicativo bem subsidiado que as defesas apresentam alguma tendência de melhor desempenho.

Diante disso, duas constatações específicas vem a tona, primeiro, de fato o método DDCAM é hábil no processo de detecção de *outliers* multivariados. Em sequência, um segundo fato, as equipes que destoam em defesa na NBA são preponderantemente mais favoritas que as demais equipes, até mesmo que as equipes que destoam em ataque. Claro que esta é uma investigação ainda incipiente, porém extremamente motivadora. Isso faz com que estudos futuros mais profundos sejam instigantes e também necessários para subsidiar estes resultados.

5 Considerações Finais

A utilização das informações baseadas na série histórica de dados em diferentes áreas tem alcançado grande penetração em diversos setores. O aumento da capacidade computacional permite a manipulação de conjuntos de dados com um volume de informações cada vez mais extenso. Este contexto tem impulsionado bastante a área atualmente denominada Ciência de Dados. A Ciência de Dados é uma área interdisciplinar, voltada para o estudo e a análise de dados, com interesse na extração de conhecimento e *insights* para tomadas de decisão.

A apresentação deste estudo se caracterizou exatamente nessa linha de conduta. Este estudo utilizou de uma técnica de análise de dados e produz uma percepção específica da informação contida nos dados para subsidiar a tomada de decisão em uma outra área, ou seja, na efetiva prática da interdisciplinaridade. Neste caso particular, o *insight* direcionou as decisões da área esportiva através do basquete.

A ferramenta particular de análise de dados utilizada neste estudo foi um procedimento para a identificação e o tratamento de valores *outliers* que é um tema bastante estudado e explorado, e tem uma enorme gama de aplicações. A importância do tema motivou a comunidade científica ao desenvolvimento de diversas técnicas, voltadas para diferentes tipos de contextos.

O desenvolvimento do presente estudo, teve um conjunto importante de objetivos que foram efetivamente alcançados. Foi apresentada uma revisão de literatura bastante abrangente e atualizada acerca de metodologias diversas para procedimentos de identificação de valores *outliers*. Um método específico e inovador, o método DD-CAM, foi apresentado com riqueza de detalhes técnicos e de aplicabilidade. Além disso, o estudo apresentou um banco de dados específico, com informações de área esportiva (dados da liga de basquete dos Estados Unidos, a NBA). Uma análise descritiva abrangente foi realizada para o referido banco de dados. Por fim, a contribuição inédita deste estudo, a busca pela associação entre os resultados desportivos do basquete com predominante bom desempenho em quesitos defensivos, isso através de métodos de detecção de equipes *outliers*.

Os resultados apresentados neste estudo mostram que durante a temporada classificatória da NBA, os times que são mais consistentes e mais eficientes em variáveis específicas de defesa tem uma maior propenção de se classificarem aos *playoffs* decisivos. Essa informação pode ser utilizada para nortear equipes e comissões técnicas para criarem e abordarem estratégias que permitam o aproveitamento destas informações estatísticas.

Este estudo, ainda apresenta capilaridade suficiente para propostas futuras de continuidade. Existe, por exemplo, a possibilidade de aplicação do método DDCAM para dados detalhados de cada equipe, ou até mesmo de um jogador em específico. Este estudo seria com o interesse em encontrar naquela equipe ou jogador qual variável analisada pode destacar este atleta como um atleta diferenciado nessa particular valência. E assim, aproveitar esta informação em prol de toda a equipe. Uma outra possibilidade de estudo próximo seria a aplicação do método DDCAM para dados na liga de basquete brasileira, o NBB (o Novo Basquete Brasil), que de uma forma similar poderia trazer informações para ajudar aos técnicos e equipes dessa liga.

Referências

[1] Duarte, Anderson Ribeiro, Josino José Barbosa, Helgem Souza Ribeiro Martins e fernando Luiz Pereira Oliveira: *Data-driven cluster analysis method: a novel outliers detection method in multivariate data*. (*submitted paper*)( ):1–25, 2022. Citado 4 vezes nas páginas 2, 7, 9 e 10.

[2] Hawkins, Douglas: *Identification of Outliers*, volume 11. Chapman and Hall, 1980. Citado na página 3.

[3] Barnett, Vic e Toby Lewis: *Outliers in Statistical Data*. John Wiley & Sons, 1994. Citado na página 3.

[4] Jolliffe, Ian e Jorge Cadima: *Principal component analysis: a review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202, 2016. Citado na página 3.

[5] Rousseeuw, Peter e Berti Van Zomeren: *Unmasking Multivariate Outliers and Leverage Points*. Journal of the American Statistical Association, 85(411):633–639, 1990. Citado 2 vezes nas páginas 3 e 4.

[6] Filzmoser, Peter, Robert Garrett e Clemens Reimann: *Multivariate Outlier Detection in Exploration Geochemistry*. Computers & Geosciences, 31(5):579–587, 2005. Citado 2 vezes nas páginas 3 e 4.

[7] Filzmoser, Peter: *Identification of Multivariate Outliers: a Performance Study*. Austrian Journal of Statistics, 34(2):127–138, 2005. Citado na página 4.

[8] Becker, Claudia e Ursula Gather: *The masking breakdown point of multivariate outlier identification rules*. Journal of the American Statistical Association, 94(447):947–955, 1999. Citado na página 4.

[9] Davies, Laurie e Ursula Gather: *The identification of multiple outliers*. Journal of the American Statistical Association, 88(423):782–792, 1993. Citado na página 4.

[10] Leys, Christophe, Olivier Klein, Yves Dominicy e Christophe Ley: *Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance*. Journal of experimental social psychology, 74:150–156, 2018. Citado na página 4.

[11] Li, Xu, Songren Deng, Lifang Li e Yunchuan Jiang: *Outlier detection based on robust mahalanobis distance and its application*. Open Journal of Statistics, 9(1):15–26, 2019. Citado na página 4.

[12] Veloso, Manoel Vitor Souza e Marcelo Angelo Cirillo: *Principal components in the discrimination of outliers: A study in simulation sample data corrected by Pearson's and Yates´ s chisquare distance*. Acta Scientiarum. Technology, 38(2):193–200, 2016. Citado na página 4.

[13] Kamalov, Firuz e Ho Hon Leung: *Outlier detection in high dimensional data*. Journal of Information & Knowledge Management, 19(01):2040013, 2020. Citado na página 5.

[14] Ghorbani, Hamid: *Mahalanobis distance and its application for detecting multivariate outliers*. Facta Univ Ser Math Inform, 34(3):583–95, 2019. Citado na página 5.

[15] Zhu, Jie, Wei Jiang, An Liu, Guanfeng Liu e Lei Zhao: *Effective and efficient trajectory outlier detection based on time-dependent popular route*. World Wide Web, 20(1):111–134, 2017. Citado na página 5.

[16] Kutsuna, Takuro e Akihiro Yamamoto: *Outlier detection using binary decision diagrams*. Data Mining and Knowledge Discovery, 31(2):548–572, 2017. Citado na página 5.

[17] Luo, Jie, Sarah Frisken, Ines Machado, Miaomiao Zhang, Steve Pieper, Polina Golland, Matthew Toews, Prashin Unadkat, Alireza Sedghi, Haoyin Zhou, Alireza Mehrtash, Frank Preiswerk, Cheng Chieh Cheng, Alexandra Golby, Masashi Sugiyama e William Wells III: *Using the variogram for vector outlier screening: application to feature-based image registration*. International Journal of Computer Assisted Radiology and Surgery, 13(12):1871–1880, 2018. Citado na página 5.

[18] Barbosa, Josino José, Tiago Martins Pereira e Fernando Luiz Pereira Oliveira: *Uma proposta para identificação de outliers multivariados*. Ciência & Natura, 40:e40 1–8, 2018. Citado 3 vezes nas páginas 6, 7 e 9.

[19] Wang, Chao, Zhen Liu, Hui Gao e Yan Fu: *VOS: A new outlier detection model using virtual graph*. Knowledge-Based Systems, 185:104907, 2019. Citado na página 6.

[20] Wang, Biao e Zhizhong Mao: *Outlier detection based on Gaussian process with application to industrial processes*. Applied Soft Computing, 76:505–516, 2019. Citado na página 6.

[21] Wahid, Abdul e Annavarapu Chandra Sekhara Rao: *A distance-based outlier detection using particle swarm optimization technique*. Em *Information and Communication Technology for Competitive Strategies*, páginas 633–643. Springer, 2019. Citado na página 6.

[22] Lejeune, Clément, Josiane Mothe, Adil Soubki e Olivier Teste: *Shape-based outlier detection in multivariate functional data*. Knowledge-Based Systems, página 105960, 2020. Citado na página 7.

[23] Barbosa, Josino José e Helgem Souza Ribeiro Duarte, Anderson Ribeiro: *A Performance Evaluation in Multivariate Outliers Identification Methods*. Ciência & Natura, 42:e16 1–14, 2020. Citado na página 7.

[24] Van Zoest, Vera, Alfred Stein e Gerard Hoek: *Outlier Detection in Urban Air Quality Sensor Networks*. Water, Air, & Soil Pollution, 229(4):111, 2018. Citado na página 7.

[25] Tabjula, Jagadeeshwar, Srijith Kanakambaran, Sheetal Kalyani, Prabhu Rajagopal e Balaji Srinivasan: *Outlier analysis for defect detection using sparse sampling in guided wave structural health monitoring*. Structural Control and Health Monitoring, página e2690, 2021. Citado na página 7.

[26] Schwarz, Gideon: *Estimating the Dimension of a Model*. The Annals of Statistics, 6(2):461–464, 1978. Citado na página 11.

[27] Phelps, Richard: *Basketball for dummies*. John Wiley & Sons, 2011. Citado na página 14.

[28] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. `https://www.R-project.org/`. Citado na página 14.