



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de Presidente Prudente

FERNANDA DE PAULA GALERANI

MONTANDO O TIME IDEAL: UMA ANÁLISE MULTIVARIADA DA NBA

PRESIDENTE PRUDENTE

2024

FERNANDA DE PAULA GALERANI

MONTANDO O TIME IDEAL: UMA ANÁLISE MULTIVARIADA DA NBA

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina TCC.

Orientador: Prof. Dr. Manoel Ivanildo Silvestre Bezerra.

PRESIDENTE PRUDENTE

2024

G154m Galerani, Fernanda
 Montando o Time Ideal: Uma Análise Multivariada da NBA /
Fernanda Galerani. -- Presidente Prudente, 2024
 49 f. : tabs., fotos

 Trabalho de conclusão de curso (Bacharelado - Estatística)
Universidade Estadual Paulista (UNESP), Faculdade de Ciências e
Tecnologia, Presidente Prudente
 Orientador: Manoel Ivanildo Silvestre Bezerra

 1. análise de agrupamentos. 2. análise de componentes principais. 3.
classificação. 4. NBA. I. Título.

TERMO DE APROVAÇÃO

Fernanda de Paula Galerani

MONTANDO O TIME IDEAL: UMA ANÁLISE MULTIVARIADA DA NBA

Relatório Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Documento assinado digitalmente
 **MANOEL IVANILDO SILVESTRE BEZERRA**
Data: 15/01/2025 14:53:15-0300
Verifique em <https://validar.iti.gov.br>

Orientador: _____

Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

Documento assinado digitalmente
 **MIRIAM RODRIGUES SILVESTRE**
Data: 15/01/2025 14:10:05-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

Documento assinado digitalmente
 **KLAUS SCHLUNZEN JUNIOR**
Data: 15/01/2025 23:44:29-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Klaus Schlunzen Junior
Departamento de Estatística

Presidente Prudente, 07 de dezembro de 2024.

RESUMO

Este trabalho tem como objetivo aplicar técnicas de análise multivariada para identificar padrões de desempenho entre jogadores da NBA, considerando suas estatísticas individuais. Inicialmente, foram utilizadas a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados e o cálculo da distância de Mahalanobis para medir a similaridade entre os jogadores. A partir dos *scores* das componentes principais, foram formados clusters utilizando o método *K-Means*, agrupando os atletas em cinco grupos, correspondentes às posições tradicionais do basquete. Além disso, o método Simplex foi aplicado para a seleção de um time otimizado, levando em conta restrições como orçamento e balanceamento por posição, possibilitando a comparação com um time de estrelas previamente definido. Os resultados evidenciaram que as técnicas multivariadas utilizadas são eficazes na identificação de padrões e na formação de agrupamentos consistentes, contribuindo para análises mais aprofundadas no âmbito esportivo. Conclui-se que a combinação de estatísticas tradicionais com métodos analíticos avançados pode fornecer insights estratégicos valiosos para times e gestores esportivos.

Palavras-chave: análise de agrupamentos; análise de componentes principais; classificação; NBA.

ABSTRACT

This study aims to apply multivariate analysis techniques to identify performance patterns among NBA players based on their individual statistics. Initially, Principal Component Analysis (PCA) was used to reduce data dimensionality, and Mahalanobis distance was calculated to measure player similarity. Based on the principal component scores, clusters were formed using the K-Medoids method, grouping athletes into five clusters corresponding to traditional basketball positions. Furthermore, the Simplex method was applied to select an optimized team, considering constraints such as budget and positional balance, allowing for comparisons with a predefined all-star team. The results demonstrated that the multivariate techniques employed are effective in identifying patterns and forming consistent groupings, contributing to deeper analyses in the sports domain. It is concluded that combining traditional statistics with advanced analytical methods can provide valuable strategic insights for teams and sports managers.

Keywords: *cluster analysis; principal component analysis; classification; NBA.*

Sumário

1 INTRODUÇÃO.....	7
2 REVISÃO BIBLIOGRÁFICA	11
3 METODOLOGIA.....	13
3.1 Os DADOS	13
3.2 COLETA DOS DADOS	16
3.2.1 IDENTIFICAÇÃO DOS DADOS	17
3.2.2 EXTRAÇÃO DOS DADOS	17
3.2.3 ARMAZENAMENTO DOS DADOS	17
3.3 TÉCNICAS UTILIZADAS	18
4 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)	23
5. ANÁLISE DE CLUSTER	26
5.1 ANÁLISE DE CLUSTER NÃO HIERÁRQUICA	26
5.2 DEFINIÇÃO DO NÚMERO DE CLUSTERS	28
6. MÉTODO SIMPLEX	29
6.1 FUNÇÃO OBJETIVO	29
6.2 A DISTÂNCIA EUCLIDIANA	30
6.3 ADAPTAÇÃO DO MÉTODO SIMPLEX AO PROBLEMA	30
7 RESULTADOS	33
7.1 APRESENTAÇÃO DOS DADOS.....	33
7.2 ANÁLISE DESCRITIVA E LIMPEZA DOS DADOS	34
7.2.1 TESTE DE MULTICOLINEARIDADE VIF (<i>VARIANCE INFLATION FACTOR</i>)	36
7.3 ANÁLISE DE COMPONENTES PRINCIPAIS	38
7.4 ANÁLISE DE CLUSTER NÃO HIERÁRQUICO	40
7.5 MÉTODO SIMPLEX	43
7.6 COMPARAÇÃO DE PREÇO	44
7.7 COMPARAÇÃO DE ESTATÍSTICAS	47
8 CONSIDERAÇÕES FINAIS	48
REFERÊNCIAS	49

1 INTRODUÇÃO

A *National Basketball Association* (NBA) é a principal liga de basquetebol do mundo, sendo localizada na América do Norte. Ela foi criada em 1946 com o nome de *Basketball Association of America* (BAA). Em 1949, se fundiu com a *National Basketball League* (NBL) e foi renomeada NBA.

O campeonato é composto por 30 franquias (clubes, 29 estadunidenses e 1 canadense), divididos em 2 conferências (Leste e Oeste), como vemos na Figura 1. Ao final de uma temporada regular as 16 equipes classificadas (8 de cada conferência) se enfrentam nos *playoffs* (fase final da competição). Na última partida, os times vencedores de cada conferência disputam pelo título.

Figura 1 – Times presentes na NBA. Do lado esquerdo os times da conferência Oeste e à direita os times da conferência Leste.



Fonte: Sportslogos (2020)

A temporada regular começa em meados de outubro e segue até meados de abril do ano seguinte. Cada time joga 82 jogos contra todos os outros, sendo que podem enfrentar alguns times 3 ou 4 vezes, dependendo de quantas vezes enfrentaram aquele time nas temporadas anteriores. Apesar disso o sistema de pontuação que classifica para os *playoffs* é válido apenas para os times da mesma conferência, ou seja, é formado um top 8 do Leste e outro do Oeste de acordo com o desempenho das franquias durante a temporada regular.

Os *playoffs* (ou mata-mata) começam em meados de abril e seguem até meados de maio. São séries "melhor-de-sete" compostas por 8 times em cada

conferência. O primeiro colocado de cada conferência enfrenta o último, o segundo o penúltimo, o terceiro o antepenúltimo e assim sucessivamente, com ilustrado pela Figura 2.

Figura 2 – Exemplo de chaveamento nos playoffs da NBA.



Fonte: Sportslogos (2020)

As finais da NBA são realizadas sempre em meados de junho. É disputada pelo campeão da Conferência Leste versus o campeão da Conferência Oeste. Assim como os *playoffs* também é disputada em uma série "melhor-de-sete".

Hoje o campeonato é transmitido para 42 países e atrai grandes jogadores de diversas nações. A maioria dos jogadores de países de língua inglesa jogam em universidades americanas antes de ir para a NBA.

De tempos em tempos grandes jogadores se destacam na liga, se tornando conhecidos até fora do círculo do basquete. Como Michael Jordan, Shaquille O'Neall, Lebron James e diversas outras estrelas com carreiras e salários astronômicos. Salários esses que ultrapassam o teto de salarial da liga, obrigando os times a aplicarem uma multa no orçamento.

A multa é conhecida como *Luxury Tax*, uma penalidade financeira aplicada a equipes da NBA que excedem o limite de impostos, o teto salarial, definido anualmente pela liga. Para a temporada 2022-2023, esse limite foi de US\$150,267 milhões.

Quando uma equipe ultrapassa esse valor ela paga uma multa progressiva por cada dólar gasto além do limite, com taxas que aumentam à medida que o excesso cresce. Por exemplo, os primeiros US\$ 5 milhões acima do limite podem gerar uma multa de US\$ 1,50 para cada dólar excedido, e valores maiores têm taxas mais altas. Essa penalidade tem como objetivo promover o equilíbrio competitivo, desincentivando equipes ricas de gastarem muito além de suas rivais.

O ponto é que manter estrelas custa muito caro. Essa combinação de altos salários e *luxury tax* torna o gerenciamento do orçamento bem desafiador.

Uma alternativa para times com orçamentos menores é montar um elenco promissor e investir nele, e a maneira mais confiável de encontrar esse elenco ideal é através de análise de dados.

Objetivos

O objetivo é separar os jogadores em clusters de acordo com suas características de interesse e, através de uma profunda análise descritiva de seus pontos fortes e fracos, alocá-los de modo que os times sejam equilibrados e consistentes. Além de utilizar análise de componentes principais (PCA) para entender quais variáveis são mais influentes na variabilidade dos times para alocá-los de maneira ótima.

Os objetivos específicos são listados a seguir:

- Fazer uma análise descritiva do desempenho dos jogadores de acordo com suas características;
- Selecionar as características com maior contribuição para explicar a variabilidade dos dados através da Análise de Componentes Principais;
- Agrupar as características relevantes de acordo com a similaridade e relação entre elas com Análise de Cluster;
- Encontrar um elenco equivalente a um time de estrelas, porém dentro do teto salarial da NBA, de acordo com o Método Simplex.

Justificativa

O *scout*, como é conhecida a análise de dados no basquete e em diversos outros esportes, vem crescendo cada vez mais. Este tipo de estudo pode fornecer *insights* valiosos para os gestores esportivos e para os próprios jogadores sobre como maximizar seu potencial e contribuir para o sucesso do time.

Nas palavras de Daryl Morey, presidente de operações de basquete do Philadelphia 76ers e um dos pioneiros da análise de dados no basquete: "O uso de dados e da análise mudou a forma como todos nós pensamos sobre o basquete. É uma maneira de encontrar vantagem competitiva e tomar decisões mais inteligentes sobre como montar e conduzir uma equipe." (Morey, 2021).

A análise de dados no basquete tem se tornado cada vez mais importante nos últimos anos, fornecendo uma grande quantidade de informações valiosas aos treinadores, jogadores e proprietários de equipes. É uma ferramenta valiosa para aprimorar as estratégias de jogo, avaliar o desempenho dos jogadores, identificar padrões e tendências em jogos passados e tomar decisões informadas sobre treinamento, contratações e escalas. (Dutta, 2021, traduzido pela autora).

2 REVISÃO BIBLIOGRÁFICA

A montagem de times de basquete utilizando ciência de dados e técnicas de otimização é um tema de crescente relevância acadêmica e prática. A aplicação de métodos estatísticos para análise de desempenho esportivo tem sido amplamente explorada em estudos recentes, com o objetivo de entender padrões, prever resultados e auxiliar na tomada de decisões estratégicas (Smith, 2021; Zuccolotto e Manisera, 2020).

A Análise de Componentes Principais (PCA) é frequentemente utilizada para reduzir a dimensionalidade dos dados e identificar as variáveis mais relevantes. Anderson (2018) destaca que a PCA transforma variáveis originais em componentes principais não correlacionados, preservando a maior parte da variância explicada. Mingoti (2007) reforça que esse método é particularmente útil em bases com múltiplas variáveis correlacionadas, como no caso de estatísticas de jogadores, onde redundâncias podem comprometer a análise.

Araujo (2023) ressalta que a PCA é uma ferramenta essencial para simplificar a interpretação de dados em estudos multivariados, fornecendo uma base sólida para análises subsequentes, como a clusterização. Esse método foi amplamente utilizado em estudos anteriores, como o de Sartori (2021), para segmentar jogadores com base em características técnicas e táticas.

A análise de *cluster* permite agrupar jogadores com características semelhantes, facilitando a identificação de perfis. Everitt et al. (2011) descrevem o método *K-means* como eficiente para segmentação de dados grandes, destacando sua aplicabilidade em esportes coletivos. Zuccolotto e Manisera (2020) também exploram a clusterização para identificar padrões táticos e estratégicos no basquete, enfatizando a importância de métricas como ofensiva e defensiva na formação de grupos homogêneos.

Dutta (2021) apresenta como a análise de *cluster* pode ser combinada com técnicas de otimização para a formação de equipes, destacando que o uso da distância de Mahalanobis, ao invés da distância Euclidiana, melhora a precisão quando há correlação entre variáveis. Esse ajuste garante que a segmentação reflita adequadamente a relação entre diferentes estatísticas.

O método simplex é amplamente reconhecido por sua eficiência na resolução de problemas de programação linear. Dantzig (1947) introduziu essa técnica como uma solução sistemática para otimizar funções lineares sujeitas a restrições. Vanderbei (2001) complementa que o método é particularmente eficaz em contextos em que há múltiplas variáveis interdependentes, como na formação de equipes esportivas dentro de limites orçamentários.

No contexto esportivo, Miguel (2022) utilizou o método simplex para analisar elencos da NBA, destacando sua utilidade para equilibrar custos e desempenho. A abordagem permite maximizar o impacto das métricas de desempenho dentro das restrições impostas, como teto salarial e número de jogadores por posição.

O uso de ciência de dados no basquete tem transformado a forma como decisões estratégicas são tomadas. Smith (2021) destaca que as análises baseadas em dados ajudam a identificar padrões complexos e prever desempenhos futuros. Irving (2021) explora como as métricas de desempenho podem ser aplicadas para criar avaliações detalhadas de jogadores, o que facilita comparações entre atletas regulares e estrelas.

Menezes (2021) e Miguel (2022) analisaram estatísticas de jogadores recém-ingressados na NBA, utilizando aprendizado de máquina para prever sucesso futuro e identificar métricas críticas para o desenvolvimento de talentos. Esses estudos destacam a relevância de combinar técnicas estatísticas tradicionais com modelos preditivos para aprimorar as decisões no esporte.

A pesquisa de Zuccolotto e Manisera (2020) ilustra como métodos estatísticos podem ser integrados para formar equipes táticas eficientes, utilizando clusterização e otimização. Araujo (2022) aplicou métodos semelhantes ao futebol americano, mostrando que o uso combinado de análise de cluster e programação linear permite maximizar o desempenho, mesmo em esportes com dinâmicas complexas e multidimensionais.

Dutta (2021) conclui que o uso de dados avançados não só melhora a eficiência na formação de equipes, mas também garante que as decisões sejam tomadas com base em evidências, reduzindo a subjetividade.

3 METODOLOGIA

Como disse (Géron, 2019) Aurélien em seu livro "Mãos à Obra: Aprendizado de Máquina com *Scikit-Learn & TensorFlow*", a análise multivariada proporciona um meio robusto de examinar e explorar dados nos quais existem várias variáveis, podendo ser usada para solucionar muitos problemas em diversos campos.

Nesse trabalho foram utilizadas, principalmente, duas técnicas de análise multivariada: análise de componentes principais e análise de agrupamentos, que serão explicadas no decorrer do trabalho.

Depois foi aplicado o método simplex para montar o time ideal (time equivalente estatisticamente às estrelas, mas com custo muito reduzido).

3.1 Os dados

Para avaliar a performance de um jogador, é necessário considerar múltiplos aspectos, como pontuação, rebotes, assistências, bloqueios, roubos de bola, entre outros. Para o relatório parcial foi utilizada uma base de dados do *website Kaggle* como forma de análise piloto, mas para a entrega final eles foram obtidos através de *web scraping* dos dados do jogo NBA2k, pois nesse existe uma variável importante não encontrada em outras bases: o *overall* de cada jogador.

O *overall* nada mais é do que um *score* calculado com base em todas as variáveis coletadas pela NBA que vai de 0 a 100, sendo 0 um jogador com estatísticas muito ruins e 100 um jogador perfeito.

NBA2K é um jogo de basquete virtual desenvolvido pela *Visual Concepts* e publicado pela *2KSports*. Ele é um simulador de basquete que tenta reproduzir a experiência de jogar na NBA da maneira mais realista possível. O jogo inclui todas as equipes e seus jogadores atuais, assim como uma variedade de modos de jogo.

São muitas estatísticas coletadas pelo jogo, elas foram agrupadas em categorias para facilitar a compreensão:

a. Informações Básicas

- TIME: Nome da equipe.
- POS: Posição do jogador.

- OVR: Avaliação geral do jogador (*overall*).
- OFF: Avaliação ofensiva.
- DEF: Avaliação defensiva.
- ENVER: Envergadura, a distância entre as pontas dos dedos das duas mãos com os braços totalmente abertos, de uma ponta à outra.
- ALT: Altura.
- PESO: Peso.

b. Habilidades de Arremesso

- *Shot Close*: Arremesso de curta distância.
- *Shot Mid*: Arremesso de média distância.
- *Shot 3pt*: Arremesso de três pontos (longa distância).
- *Free Throw*: Lance livre.
- *Shot IQ*: Inteligência nos arremessos.

c. Finalizações e Movimentos Ofensivos

- *Driving Layup*: Bandeja em infiltração.
- *Standing Dunk*: Enterrada parada.
- *Driving Dunk*: Enterrada em movimento.
- *Post Moves*: Movimentos no poste.
- *Post Hook*: Gancho no poste.
- *Post Fade*: Arremesso *fadeaway* no poste. O arremesso *fadeaway* é quando o jogador salta para trás, afastando-se do defensor para dificultar o bloqueio.
- *Draw Foul*: Capacidade de sofrer faltas.
- *Hands*: Controle de bola.

d. Habilidades de Passe e Controle de Bola

- *Ball Handle*: Manuseio de bola.
- *Passing Accuracy*: Precisão nos passes.
- *Passing Vision*: Visão de jogo para passes.
- *Passing IQ*: Inteligência nas decisões de passe.
- *Speed With Ball*: Velocidade conduzindo a bola.

e. Atributos Físicos

- *Acceleration*: Aceleração.
- *Vertical*: Impulsão vertical.
- *Strength*: Força física.
- *Stamina*: Resistência.
- *Hustle*: Esforço e dedicação em quadra.

f. Habilidades Defensivas

- *Interior Defense*: Defesa próxima à cesta.
- *Perimeter Defense*: Defesa no perímetro.
- *Help Defense IQ*: Inteligência na ajuda defensiva.
- *Lateral Quickness*: Agilidade lateral.
- *Pass Perception*: Percepção para interceptar passes.
- *Steal*: Roubos de bola.
- *Block*: Tocos.
- *Defensive Consistency*: Consistência defensiva.
- *Defensive Rebound*: Rebotes defensivos.

g. Outras Avaliações

- *Offensive Consistency*: Consistência ofensiva.
- *Offensive Rebound*: Rebotes ofensivos.
- *Defensive Rebound*: Rebotes defensivos.
- *Intangibles*: Aspectos intangíveis (liderança, ética de trabalho etc.).
- *Potential*: Potencial de crescimento.

Focando nas habilidades mais importantes para cada posição temos:

a. Armadores (*Point Guard* - PG)

- Habilidade do passe: *Passing Accuracy*, *Passing Vision*, *Passing IQ*.
- Controle de bola: *Ball Handle*, *Speed With Ball*.
- Arremesso: *Shot 3pt*, *Shot Mid*, e *Shot IQ*.

b. Ala-Armadores (*Shooting Guard* - SG)

- Arremesso: *Shot 3pt*, *Shot Mid*, *Free Throw*.

- Controle de bola: *Ball Handle, Speed With Ball*.
- Marcação: *Perimeter Defense*.

c. Alas (*Small Forward* - SF)

- Arremesso e Finalizações: *Shot 3pt, Shot Mid, Driving Layup*.
- Atributos Físicos: *Speed, Vertical, Strength*.
- Marcação: *Perimeter Defense* e *Interior Defense*.

d. Alas-Pivôs (*Power Forward* - PF)

- Habilidades próximas à cesta: *Post Moves, Post Fade, Standing Dunk*.
- Rebotes e Força Física: *Offensive Rebound, Defensive Rebound, Strength*.
- Arremesso de Média Distância: *Shot Mid*.

e. Pivôs (*Center* - C)

- Habilidades próximas à cesta: *Standing Dunk, Post Hook, Post Moves*.
- Defesa Próxima à Cesta: *Interior Defense, Block, Defensive Consistency*.
- Rebotes e Força Física: *Defensive Rebound, Offensive Rebound, Strength*.

3.2 Coleta dos dados

Os dados finais referentes à temporada 2022-2023 foram coletados através de *web scraping* pelo software *Python*, que coletou as informações do site oficial do jogo NBA2K (<https://2kdb.net/>), o qual contém os dados dos jogadores e do campeonato (NBA).

O processo de *web scraping* consiste na extração automatizada de dados de páginas da internet, o que permite obter informações de forma eficiente e organizada.

Essa prática é comumente usada para coletar grandes volumes de dados, que são posteriormente analisados para diversos fins, como pesquisa de mercado, análise de concorrência e monitoramento de preços (Mitchell, 2018).

O processo envolve a criação de scripts que acessam páginas da web e extraem os dados desejados. Esses scripts, escritos em linguagens como *Python* com bibliotecas específicas como *BeautifulSoup* e *Scrapy*, acessam o conteúdo HTML de uma página e filtram as informações de interesse.

De acordo com Lawson (2015), o uso dessas bibliotecas é fundamental, pois elas facilitam a navegação e extração dos dados de forma estruturada.

3.2.1 Identificação dos Dados

A primeira etapa do *web scraping* é identificar quais dados devem ser extraídos e onde eles estão localizados na página. Isso envolve analisar a estrutura HTML da página para localizar elementos específicos (como `<div>`, `<span>`, `<table>`) que contêm as informações desejadas.

3.2.2 Extração dos Dados

A extração é realizada por meio de requisições HTTP, geralmente com ferramentas como *Requests* no *Python*. Essas requisições permitem acessar o HTML da página, que é então analisado para localizar os dados específicos. Conforme explica Lawson (2015), o uso de bibliotecas possibilita a navegação e a extração eficiente das informações, ao “escolher, filtrar e navegar pelos elementos HTML” (Lawson, 2015, p. 32).

3.2.3 Armazenamento dos Dados

Após a extração, os dados devem ser armazenados em um formato estruturado, como foi feito neste trabalho utilizando CSV, podendo também ser salvos em JSON ou diretamente em um banco de dados.

A vantagem desta técnica é que em vez de visualizarmos uma página de cada vez em uma janela pequena de um monitor, podemos visualizar bancos de dados completos com informações que se estendem por milhares de páginas de uma vez só (Mitchell, 2018. p. 8).

3.3 Técnicas utilizadas

Foi utilizado um conjunto de métodos estatísticos e de otimização para resolver o problema de montagem de um time de basquete equivalente a um elenco de estrelas da NBA, respeitando o teto salarial da temporada 2022-2023, fixado em 150 milhões de dólares.

Os métodos aplicados incluíram a análise de componentes principais (PCA) para redução da dimensionalidade, análise de cluster para agrupamento de jogadores com base em suas estatísticas e o método simplex para a seleção otimizada de um time dentro das restrições financeiras e táticas.

A primeira etapa da análise consistiu na aplicação da análise de componentes principais (PCA), que tem como objetivo reduzir a dimensionalidade dos dados originais, mantendo a maior parte da variância explicada. Essa técnica foi empregada porque são 47 variáveis e existem redundâncias devido à correlação entre estatísticas, como visão de passe e inteligência de passe.

De acordo com Mingoti (2007), a PCA é fundamental para reduzir a dimensionalidade em análises multivariadas. Varella (2008) também reforça que a técnica ajuda a evitar redundâncias e multicolinearidade nos dados. Na análise, selecionaram-se os componentes principais responsáveis por explicar cerca de 72% da variância cumulativa, reduzindo o número variáveis analisadas para 6 componentes principais.

Com base nos componentes principais gerados pela PCA, aplicou-se a análise de cluster para agrupar os jogadores em cinco grupos, representando as cinco posições tradicionais do basquete: armador, ala-armador, ala, ala-pivô e pivô. A escolha da análise de cluster baseou-se na necessidade de garantir que os jogadores fossem agrupados de forma coerente, considerando similaridades em suas características estatísticas.

O algoritmo *K-means* foi utilizado devido à sua simplicidade e eficiência para grandes conjuntos de dados. Essa abordagem iterativa ajusta os centróides de cada cluster até que a variância intraclasse seja minimizada, garantindo a separação clara entre os grupos (Everitt, 1980).

Os clusters obtidos representaram jogadores com perfis semelhantes dentro de suas respectivas posições, facilitando a identificação de substitutos estatisticamente próximos a estrelas.

Após a formação dos clusters, o método simplex foi aplicado para selecionar o time final, garantindo a maximização do desempenho total dentro do teto salarial da NBA para a temporada 2022-2023. O método foi escolhido por sua eficiência na resolução de problemas de otimização linear, sendo amplamente reconhecido como uma abordagem robusta para maximizar ou minimizar funções lineares sob restrições, conforme descrito por Chvátal (1983).

A função objetivo foi definida como a maximização da soma ponderada das contribuições estatísticas dos jogadores selecionados, enquanto as restrições incluíam:

- A limitação financeira (salário total ≤ 150 milhões de dólares);
- A necessidade de escolher 5 jogadores titulares, representando uma escalação completa;
- Garantir a escolha de pelo menos um jogador de cada cluster, respeitando as posições do basquete.

Como critério adicional incorporou-se a distância euclidiana entre os jogadores regulares e os jogadores estrelas para garantir que o time fosse equivalente às estrelas.

A distância euclidiana foi utilizada como uma métrica para medir a similaridade estatística, o que assegurou que os jogadores selecionados replicassem o desempenho das estrelas de forma eficiente, como sugerido por Dutta (2021) e Sartori (2021), que destacam o uso de dados analíticos para decisões estratégicas em esportes.

De maneira esquematizada para auxiliar no entendimento, o processo foi o seguinte:

1. Identificação de Multicolinearidade

Objetivo: Identificar variáveis originais que apresentam alta correlação (multicolinearidade).

Técnica Utilizada: Foi calculada a matriz de correlação das variáveis originais.

2. Aplicação da PCA

Objetivo: Reduzir a dimensionalidade e resolver o problema de multicolinearidade, transformando variáveis correlacionadas em componentes principais ortogonais (não correlacionadas).

Técnicas e Informações:

- As variáveis originais foram centralizadas (média 0) e escalonadas (variância 1) para garantir que todas tivessem igual peso na análise.
- A PCA foi aplicada, gerando um conjunto de componentes principais, com as primeiras 6 componentes capturando a maior parte da variância dos dados.

Interpretação:

- Cada componente principal é uma combinação linear das variáveis originais.
- Os autovalores indicaram a variância explicada por cada componente, enquanto os autovetores (cargas fatoriais) mostraram a contribuição de cada variável para as componentes.

3. Cálculo da Distância de Mahalanobis

Dados Utilizados: Os scores das 6 primeiras componentes principais gerados pela PCA foram utilizados como base.

Matriz de Correlação Utilizada:

- A matriz de correlação utilizada na distância de Mahalanobis foi a matriz de correlação entre os scores das componentes principais.
- Como as componentes principais são ortogonais, a matriz é diagonal, com variâncias (autovalores) nas diagonais e zeros fora delas. Isso reflete a ausência de correlação entre as componentes principais.

Objetivo da Distância: Avaliar a dissimilaridade entre observações (jogadores) considerando tanto a variância explicada pelas componentes quanto a correlação (ou independência) entre elas.

Interpretação da Distância: Observações próximas em termos de distância de Mahalanobis têm padrões semelhantes nas 6 dimensões principais (componentes principais).

4. Clusterização com K-Means

Objetivo: Agrupar os jogadores em clusters com base nas distâncias calculadas.

Dados Utilizados: A matriz de dissimilaridade gerada pela distância de Mahalanobis foi usada para formar 5 clusters, representando padrões distintos de jogadores.

Interpretação dos Clusters: Cada cluster representa um grupo de jogadores com características semelhantes nos dados projetados (componentes principais).

5. Interpretação dos Clusters com as Variáveis Originais

Cargas Fatoriais (Autovetores):

- As cargas fatoriais da PCA indicam a contribuição de cada variável original para as componentes principais.
- Exemplo: Se a PC1 tem altas cargas em pontos e assistências, um cluster com altos scores na PC1 pode ser interpretado como grupo de jogadores com desempenho elevado nessas métricas.

Estatísticas nos Dados Originais:

- Após atribuir os clusters aos jogadores, calcula-se as médias das variáveis originais dentro de cada cluster para identificar padrões específicos.
- Exemplo: O Cluster 1 pode ter jogadores com altas médias de rebotes e bloqueios, sugerindo que este grupo representa pivôs.

6. Método Simplex

Objetivo: Otimizar a seleção de um time de jogadores considerando múltiplas métricas de desempenho e restrições, como orçamento e balanceamento entre posições.

Dados Utilizados: As métricas de desempenho escolhidas (*OFF*, *DEF*, *Shot IQ* etc.) dos jogadores disponíveis, além de restrições baseadas em distribuição por cluster (posição) e pela semelhança com os jogadores estrelas.

Vantagens:

- Resolve problemas de otimização linear de forma eficiente.
- Permite incorporar múltiplas restrições, como garantir a inclusão de pelo menos um jogador por cluster.

Interpretação do Resultado: O Simplex retorna o conjunto de jogadores que melhor equilibram desempenho e restrições, formando um time otimizado que pode ser comparado ao time original de estrelas.

4 Análise de Componentes Principais (PCA)

A PCA é utilizada para explorar padrões subjacentes e identificar as principais fontes de variabilidade em um conjunto de variáveis correlacionadas. No contexto da análise de cluster não hierárquica dos jogadores da NBA, a análise de componentes principais pode ser uma ferramenta útil para fornecer uma representação simplificada dos dados originais, preservando a maior parte da informação relevante. Então definimos quais variáveis são mais influentes e retiramos aquelas que se tornam redundantes no modelo por serem altamente correlacionadas.

Se a maior parte da variância populacional total (por exemplo, 80 a 90%), para um p grande, pode ser atribuída aos primeiros um, dois ou três componentes, então estes componentes podem “substituir” as p variáveis originais sem muita perda de informação. (Silvestre, 2023, p.20)

Matematicamente falando, a PCA depende somente da matriz de covariâncias ou matriz de correlação de $(X_1, X_2, \dots, X_p)$, sendo p o número de variáveis. Logo, para que a análise de componentes principais tenha sentido é preciso que as variáveis sejam correlacionadas, além de ser preciso supor que as variáveis sejam independentes e identicamente distribuídas, mas não é necessário que tenham distribuição normal.

Calcula-se os componentes principais a partir da matriz de covariância $\mathbf{S}$ ou da matriz de correlação $\mathbf{R}$. Através da matriz $\mathbf{X}$ de dados, de ordem ' $n \times p$ ', fazemos uma estimativa da matriz de covariância $\mathbf{S}$, que é simétrica e de ordem ' $p \times p$ '.

$$S = \begin{pmatrix} \hat{\text{Var}}(x_1) & \hat{\text{Cov}}(x_1x_2) & \hat{\text{Cov}}(x_1x_3) & \dots & \hat{\text{Cov}}(x_1x_p) \\ \hat{\text{Cov}}(x_2x_1) & \hat{\text{Var}}(x_2) & \hat{\text{Cov}}(x_2x_3) & \dots & \hat{\text{Cov}}(x_2x_p) \\ \hat{\text{Cov}}(x_3x_1) & \hat{\text{Cov}}(x_3x_2) & \hat{\text{Var}}(x_3) & \dots & \hat{\text{Cov}}(x_3x_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\text{Cov}}(x_px_1) & \hat{\text{Cov}}(x_px_2) & \hat{\text{Cov}}(x_px_3) & \dots & \hat{\text{Var}}(x_p) \end{pmatrix} \quad (1)$$

A matriz de correlação **R**, também simétrica de ordem 'p x p', é dada por:

$$R = \begin{pmatrix} 1 & r(x_1x_2) & r(x_1x_3) & \dots & r(x_1x_p) \\ r(x_2x_1) & 1 & r(x_2x_3) & \dots & r(x_2x_p) \\ r(x_3x_1) & r(x_3x_2) & 1 & \dots & r(x_3x_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(x_px_1) & r(x_px_2) & r(x_px_3) & \dots & 1 \end{pmatrix} \quad (2)$$

Caso as variáveis não estejam na mesma escala de medida podemos ajustá-las ao utilizar a matriz de correlação **R** no lugar da matriz de covariância **S**.

Os componentes principais são determinados resolvendo a equação característica da matriz **S** ou **R**:

$$\det[R - \lambda I_p] = P_A(\lambda) \quad (3)$$

Através da fórmula acima são calculados os autovalores e os autovetores, na qual 'det' é o determinante da matriz **R**, ' I_p ' é a matriz identidade e os valores de λ que igualam P_A a 0 são os autovalores associados a matriz **R**.

Para cada autovalor λ_i existe um autovetor $\tilde{a}_i$:

$$\tilde{a}_i = \begin{pmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{ip} \end{pmatrix} \quad (4)$$

Os valores das coordenadas dos autovetores nos informarão a contribuição para cada componente principal e os autovalores serão o quanto o componente é explicado através da variância.

É preciso decidir quantos componentes principais devemos reter na análise. Existem alguns critérios na literatura:

- Critério de Kaiser (1958): Este critério propõe a retenção dos componentes principais que correspondem aos autovalores superiores à média dos autovalores, caso a análise seja baseada na matriz de covariância. Se a análise for baseada na matriz de correlação, o critério sugere a manutenção dos componentes principais cujos autovalores excedem 1. No entanto, ao seguir este critério, pode-se eliminar

componentes principais que possuem contribuições significativas;

- Manter o número de componentes principais que representem pelo menos uma determinada proporção da variabilidade total dos dados, por exemplo, 70%;
- Manter o número de componentes principais que representem pelo menos uma determinada proporção da variabilidade de cada uma das variáveis originais, digamos 50%;
- Gráfico conhecido como *scree plot*, que permite a visualização dos autovalores. Busca-se pelo “cotovelo” (ou curva) no gráfico. Normalmente, a diferença entre os primeiros autovalores é grande e vai diminuindo até os últimos. A recomendação é realizar o corte quando a variação começar a ser pequena.

5. Análise de Cluster

A análise de cluster é uma técnica de análise multivariada de dados utilizada para agrupar objetos semelhantes em conjuntos chamados de clusters. O objetivo principal dessa técnica é identificar padrões nos dados, agrupando-os com base em suas características comuns, o que facilita a identificação de relações implícitas (Hair et al., 2009).

Existem dois tipos principais de análise de cluster: hierárquica e não hierárquica. A análise hierárquica forma uma estrutura em árvore (dendrograma), que permite observar a relação de proximidade entre todos os objetos. Já a análise não hierárquica agrupa os objetos em clusters independentes, sem formar uma estrutura hierárquica. A escolha entre esses métodos depende da natureza dos dados, dos objetivos da análise e da interpretação desejada dos resultados (Anderson, 2018).

5.1 Análise de Cluster Não Hierárquica

Neste trabalho, foi utilizada a análise não hierárquica, também conhecida como particional, que oferece maior controle sobre o número de clusters, eficiência computacional, interpretação direta dos resultados, escalabilidade e flexibilidade na forma dos grupos. De acordo com Hair et al. (2009), essa técnica é recomendada quando se conhece, ao menos aproximadamente, o número de grupos desejado, o que facilita a análise e interpretação dos resultados.

Uma das técnicas mais utilizadas na análise não hierárquica é o *K-means*, que agrupa os objetos com base na similaridade, utilizando o método de distância escolhido como métrica de proximidade. O método *K-means* pode ser dividido em quatro passos principais (Araujo, 2023, p. 42):

1. Definir K centros provisórios: Escolher K pontos iniciais que atuarão como centros dos clusters;
2. Início do processo de agrupamento: Atribuir cada observação ao centro de cluster mais próximo;
3. Reatribuição dos objetos: Combinar cada observação com o centro de cluster ao qual possui maior similaridade, formando K grupos.

4. Iteração até a convergência: Repetir os passos 2 e 3 até que os agrupamentos se estabilizem e não haja mudanças significativas.

No método *K-means* é possível escolher diferentes formas de medir a distância entre os pontos e os centróides dos clusters. Essa escolha pode influenciar significativamente os resultados da clusterização, pois cada métrica enfatiza diferentes aspectos das relações entre os dados.

A distância euclidiana é a escolha mais comum, representando a linha reta entre dois pontos no espaço. Mas ela assume que as variáveis são independentes entre si e, em cenários onde há correlação, como em dados multivariados de desempenho esportivo, essa suposição não é válida, levando a agrupamentos equivocados (Anderson, 2018). Nessas situações métodos como a distância de Mahalanobis, que ajustam para correlação e variância, são mais adequados, conforme destacado por Everitt (1980).

A distância de Mahalanobis foi escolhida nesta pesquisa porque ela considera a correlação entre as variáveis estatísticas dos jogadores, o que é crucial para dados esportivos. Como as métricas de desempenho dos jogadores podem ser altamente correlacionadas (por exemplo, eficiência ofensiva e número de pontos marcados), utilizar uma métrica que ajuste para essa correlação aumenta a precisão da clusterização.

Além disso, essa métrica é escalável e ajusta os dados para a variância em cada dimensão, tornando-a mais robusta em bases com diferentes escalas (Anderson, 2018).

A distância de Mahalanobis é definida por:

$$d = \sqrt{(x - \mu)^T R^{-1} (x - \mu)} \quad (5)$$

Em que $(x - \mu)^T$ representa o vetor transposto da diferença entre as observações (x) e a média (μ) e $\mathbf{R}$ é a matriz de correlação dos dados, que pode ser substituída pela matriz de covariância $\mathbf{S}$ se todas as variáveis estão padronizadas ou possuem unidades comparáveis pois mantém a variabilidade original. Como na base existem variáveis com escalas diferentes foi utilizada a matriz de correlação, que

padroniza as variáveis conforme discutido anteriormente, removendo as diferenças de escala e medindo apenas as relações relativas (entre -1 e 1).

5.2 Definição do Número de Clusters

A definição do número ideal de clusters é uma etapa crítica e deve ser feita com base em métodos específicos, como o método do cotovelo e o método da silhueta (James et al., 2013).

- Método do Cotovelo (*Elbow Method*): Consiste em plotar o valor da soma dos quadrados das distâncias em função do número de clusters. O ponto em que há uma mudança brusca na curva (formando um "cotovelo") indica a escolha ótima para o número de clusters (Hair et al., 2009).
- Método da Silhueta (*Silhouette Method*): Mede o grau de similaridade de um objeto com o seu próprio grupo em comparação aos outros grupos. O coeficiente de silhueta varia de -1 a 1, sendo que valores mais altos indicam uma melhor separação entre os grupos. Esse método permite avaliar a qualidade dos clusters formados e verificar se o número de clusters escolhido é adequado (James et al., 2013).

A escolha do número final de grupos deve ser feita com base em uma análise cuidadosa, considerando tanto os métodos de validação quanto o conhecimento prévio sobre os dados. Não há uma regra fixa para essa decisão, mas é essencial adotar critérios objetivos e coerentes com o problema em questão e com as características dos dados analisados. Assim, a análise de cluster não hierárquico se mostra uma ferramenta poderosa para identificar padrões e auxiliar na tomada de decisões com base em dados.

Nesse trabalho é interessante formar cinco clusters, uma vez que cada um representaria uma das posições do basquete (armador, ala-armador, ala, ala-pivô e pivô). Essa expectativa se justifica pelo fato de que, como discutido anteriormente, cada posição possui um conjunto distinto de características e habilidades.

Os pivôs, por exemplo, tendem a ser jogadores mais altos e fortes, com estatísticas superiores em defesa e rebotes. Em contrapartida, os armadores são geralmente mais ágeis, com habilidades mais desenvolvidas em velocidade, controle de bola e arremessos de longa distância.

Assim, espera-se que a análise de cluster consiga capturar essas diferenças naturais entre as posições, agrupando os jogadores com base em suas similaridades.

6. Método Simplex

O método simplex é amplamente utilizado para resolver problemas de otimização linear. Criado por George Dantzig em 1947, é considerado uma das técnicas mais eficientes para problemas envolvendo múltiplas variáveis e restrições (Dantzig, 1947). Sua aplicação na montagem de times de basquete permite selecionar jogadores que maximizam o desempenho geral do time, respeitando limites como o teto salarial e a necessidade de similaridade estatística com jogadores de elite.

6.1 Função Objetivo

A função objetivo do método simplex é dada por:

$$Z = \sum_{i=1}^n c_i x_i \quad (6)$$

Aqui, Z é o valor que se busca maximizar (ou minimizar), c_i representa a contribuição de cada variável ao objetivo e x_i são as variáveis de decisão. No contexto do basquete, Z representa o desempenho total do time em termos de métricas ofensivas e defensivas.

A maximização de Z significa encontrar a combinação de variáveis que gera o maior valor possível, respeitando as restrições impostas (Chvátal, 1983). Em problemas de montagem de times, Z incluiria tanto a eficiência dos jogadores quanto sua similaridade com atletas de elite.

As variáveis de restrição representadas por x_i são escolhidas de acordo com o problema. Na montagem do time poderia ser o número de jogadores titulares:

$$\sum_{i=1}^n x_i = 5 \quad (7)$$

Ou o total de salários dos jogadores selecionados deve respeitar o teto salarial definido em 150 milhões na temporada 2022-2023:

$$\sum_{i=1}^n s_i x_i \leq 150.000.000 \quad (8)$$

Em que s_i representa o salário de cada jogador.

6.2 A distância Euclidiana

A distância euclidiana é uma métrica útil para avaliar a similaridade entre jogadores regulares e jogadores de elite. No modelo apresentado, foi utilizada para identificar jogadores cujas estatísticas mais se assemelham às de estrelas em métricas como ofensiva (*OFF*), defensiva (*DEF*), habilidade de arremesso (*Shot IQ*), arremessos de três pontos (*Shot 3pt*) e bloqueios (*Block*).

A fórmula para calcular a distância euclidiana é:

$$d(A, B) = \sqrt{\sum_{i=1}^k (A_i - B_i)^2} \quad (9)$$

Em que:

- A_i e B_i são as métricas i -ésimas dos jogadores estrela e regular, respectivamente;
- K é o número total de métricas analisadas.

6.3 Adaptação do Método Simplex ao problema

A solução proposta utiliza a distância Euclidiana como critério para definir os jogadores mais semelhantes às estrelas dentro de cada cluster, funcionando como uma restrição na seleção final. Apenas o jogador com a menor distância em relação ao jogador estrela do mesmo cluster é considerado viável.

Outro método de distância, como Mahalanobis, poderia ter sido utilizado, mas, como a questão das escalas diferentes foi resolvida utilizando a matriz de correlação no lugar da matriz de covariância e a maioria das variáveis são *scores* (seguindo a mesma escala), o uso da distância euclidiana não compromete a análise.

A função objetivo pode ser adaptada para incluir tanto o desempenho do jogador quanto sua similaridade com a estrela correspondente:

$$Z = \sum_{i=1}^n (c_i - d(A, B)_i) x_i \quad (10)$$

Em que:

- Z : Representa o valor total da função objetivo, que deve ser maximizado. Isso significa formar um time que seja estatisticamente próximo do desempenho do time de estrelas, respeitando as restrições impostas.
- c_i : Estatísticas específicas do jogador i usadas como medidas de desempenho. Para isso foram selecionadas estatísticas chaves para definir um jogador habilidoso:
 1. **OFF**: Desempenho ofensivo.
 2. **DEF**: Desempenho defensivo.
 3. **Shot IQ**: Inteligência de arremesso.
 4. **Shot 3pt**: Precisão em arremessos de 3 pontos.
 5. **Block**: Bloqueios.
- $d(A, B)_i$: Distância Euclidiana entre o jogador regular A e o jogador estrela B no mesmo cluster:

$$d(A, B)_i = \sqrt{(OFF_A - OFF_B)^2 + (DEF_A - DEF_B)^2 \dots}$$
- x_i : É uma variável binária que indica se o jogador i foi selecionado ($x_i=1$) ou não ($x_i=0$).

Além do número de jogadores titulares e teto salarial, podem ser adicionadas restrições táticas, como garantir a diversidade de posições ou equilibrar habilidades defensivas e ofensivas.

No modelo apresentado, os jogadores são agrupados em clusters com base em posições ou estilos de jogo. Dentro de cada cluster o jogador com menor distância euclidiana à estrela correspondente é selecionado para o time que chamaremos de equivalente.

Na prática as restrições são as seguintes:

- Restrições de Cluster (diversidade de posições): Cada jogador selecionado deve pertencer a um *cluster* específico, que representa uma posição no basquete. Por exemplo, apenas jogadores do *cluster* dos armadores (*cluster* 1) podem competir para representar a posição de armador: $\sum x_{ij} = 1, \forall i \in \{1, 2, 3, 4, 5\}$ e $j \in \{Cluster_i\}$.
- Restrição de estrelas: As estrelas não são incluídas no time compatível, evitando duplicações.
- Restrição de Similaridade (distância mínima): Apenas o jogador com a menor distância Euclidiana nas variáveis escolhidas é selecionado por *cluster*: $x_{ij} \Rightarrow \min d(A, B)_j, \forall j \in \{Cluster_i\}$.
- Restrição Binária: Cada jogador pode ser selecionado apenas uma vez, pois apenas o mais similar de cada *cluster* é escolhido.

Resumindo, o objetivo do método simplex é maximizar o desempenho do time (Z) escolhendo um jogador de cada cluster, minimizando a distância entre cada um e a estrela correspondente e garantindo que não haja repetições. Todo o processo foi feito no *software* R.

7 RESULTADOS

7.1 Apresentação dos Dados

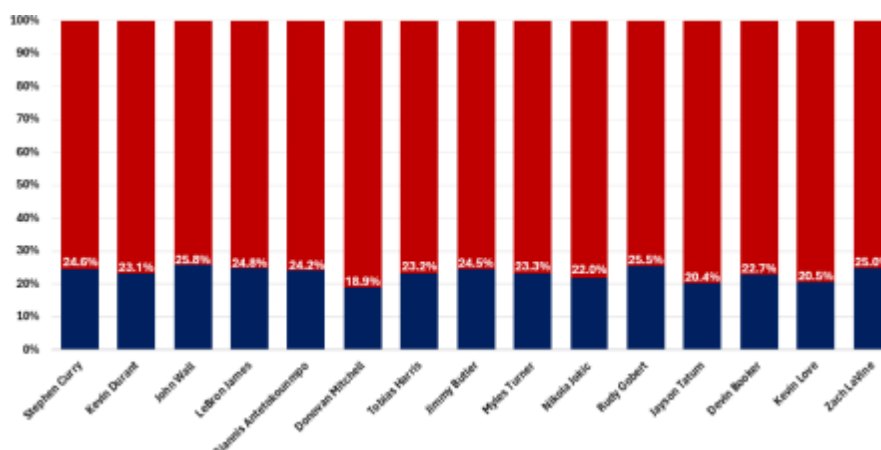
Como dito anteriormente, os dados foram coletados por meio de *web scraping*, feito no *software Python*, da página oficial do jogo NBA2k. A base contém 488 observações, em que cada uma representa um jogador, 46 variáveis e é em relação à temporada 22-23. Todas as outras etapas foram realizadas no *software R*.

Para ilustrar como jogadores estrelas concentram boa parte do orçamento salarial dos times foi utilizada uma segunda base de dados, coletados do site *hoopshype* que funciona como um blog que reúne notícias, estatísticas e outras informações sobre a liga.

A segunda base traz o nome da maior estrela de cada time, o time a que pertence, seu salário e o orçamento salarial total do time referente à 22-23.

O gráfico de barras empilhadas, apresentado na Figura 3, mostra a distribuição do orçamento salarial dos 15 times mais caros da NBA, destacando a porcentagem dedicada ao salário da principal estrela de cada equipe (em azul) em comparação ao restante do elenco (em vermelho). Na liga cada elenco possui em média 13 jogadores, as barras representam os times e o percentual destacado evidencia o impacto financeiro das estrelas no total do orçamento. O gráfico ilustra como uma parcela significativa dos recursos é frequentemente direcionada a um único jogador, especialmente nas equipes com os maiores orçamentos da liga.

Figura 3: Gráfico de Impacto das Estrelas no Orçamento Salarial dos 15 Times Mais Caros da NBA



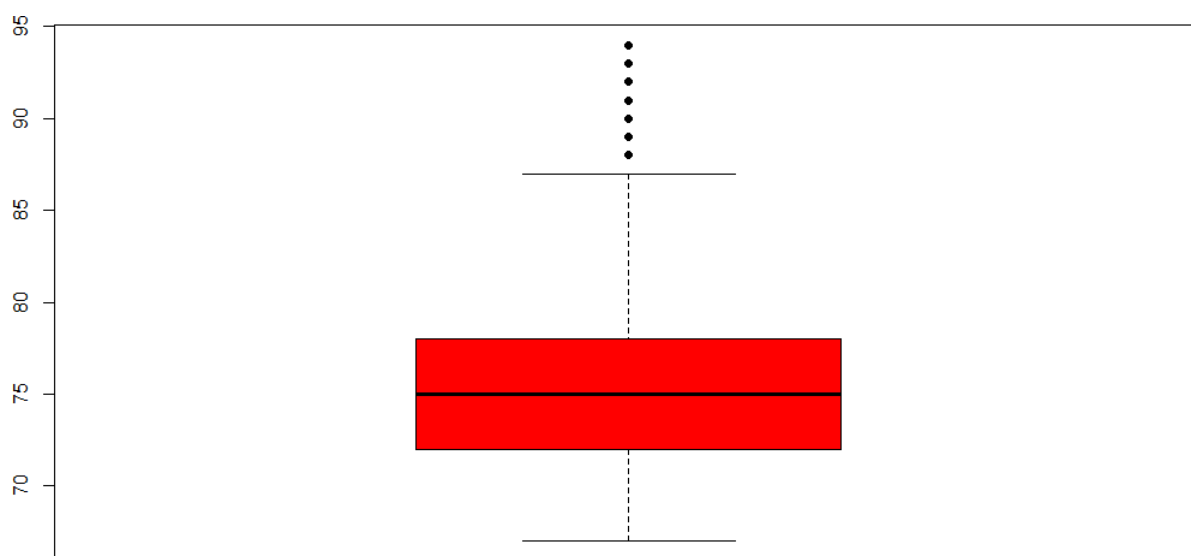
Fonte: Elaborado pela autora (2024)

7.2 Análise Descritiva e limpeza dos dados

Primeiro é preciso observar a distribuição dos dados, buscando *outliers* e entendendo o equilíbrio das habilidades dos jogadores.

A Figura 4 apresenta o gráfico *Boxplot* da variável *overall* de todos os jogadores.

Figura 4: *Boxplot* do *overall* (nota geral)



Fonte: Elaborado pela autora (2024)

Analisando melhor as medidas do *boxplot* entre os *outliers* temos:

- Mínimo: 67,0
- 1º Quartil: 72,0
- Mediana: 75,0
- Média: 75,7
- 3º Quartil: 78,0
- Máximo: 94,0

Logo, os jogadores estrelas (ou *outliers*) representam menos de 25% do total, visto que 75% dos jogadores têm a nota geral maior ou igual a 78.

Analisando essas estrelas separadamente, foi organizada a Tabela 1, disposta a seguir.

Tabela 1 - Informações dos jogadores estrelas

Nome	Overall	Posição
Giannis Antetokounmpo	94	PF
Joel Embiid	94	C
LeBron James	94	SF
Luka Doncic	94	SG/SF
Nikola Jokic	94	C
Kevin Durant	93	PF
Stephen Curry	93	PG/SG
Jayson Tatum	92	SF/PF
Anthony Davis	91	C
Ja Morant	91	PG
Kawhi Leonard	91	SF
Damian Lillard	90	PG
Jimmy Butler	90	SF/PF
Devin Booker	89	SG
Shai Gilgeous-Alexander	89	SG
James Harden	88	SG
Kyrie Irving	88	PG/SG
Trae Young	88	PG
Zion Williamson	88	PF/C

Fonte: Elaborado pela autora (2024)

Para montar um “time dos sonhos” teoricamente é necessário escolher jogadores com as maiores notas gerais para cada uma das 5 posições do basquete, ou seja, 5 desses *outliers*, os quais são listados na Tabela 2.

Tabela 2 - Time dos sonhos

Posição em inglês	Posição em português	Melhor Jogador (22-23)	Overall
<i>Point Guard</i> (PG)	Armador	Stephen Curry	94
<i>Shooting Guard</i> (SG)	Ala-armador	Luka Doncic	94
<i>Small Forward</i> (SF)	Ala	LeBron James	94
<i>Power Forward</i> (PF)	Ala-pivô	Giannis Antetokounmpo	94
<i>Center</i> (C)	Pivô	Nikola Jokic	94

Fonte: Elaborado pela autora (2024)

Antes de aplicar a PCA é importante retirar os *outliers* e então o *overall* dos jogadores restantes para evitar combinações equivocadas de variáveis. Também é necessário manter apenas variáveis numéricas, e realizar um teste de multicolinearidade.

Então foram feitas as seguintes mudanças:

- Remoção de *outliers* para que não sejam escolhidos outros jogadores estrelas;
- Remoção das variáveis: Nome, Posição, Time e *Overall*.

7.2.1 Teste de Multicolinearidade VIF (*Variance Inflation Factor*)

É sabido que multicolinearidade ocorre quando uma variável independente é altamente correlacionada com outra(s), o que pode distorcer os resultados do modelo.

O VIF mede quanto a variância dos coeficientes estimados é inflada devido à correlação entre as variáveis. Um VIF alto indica multicolinearidade, ou seja, que a variável está correlacionada com outras variáveis independentes.

Para cada variável X_i , o VIF é calculado usando a seguinte fórmula:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (11)$$

em que R_i^2 é o coeficiente de determinação (R-quadrado) da regressão de X_i contra todas as outras variáveis independentes.

Para determinar se existe multicolinearidade é comum a utilização do padrão:

- $VIF \approx 1$: Sem multicolinearidade;
- $1 > VIF > 5$: Multicolinearidade moderada;
- $VIF > 10$: Multicolinearidade severa, o que pode indicar problemas graves no modelo.

As variáveis com VIF maior que 1 podem ser observadas na Tabela 3 a seguir.

Tabela 3 - Maiores resultados do teste de multicolinearidade

VIF	Variável	Glossário
14,714683	`Speed With Ball`	Velocidade com a bola
11,708248	DEF	Defesa geral
8,864201	`Lateral Quickness`	Velocidade nas laterais
8,805722	`Ball Handle`	Controle de bola no dribble
8,027281	Acceleration	Aceleração
7,828688	`Perimeter Defense`	Defesa no perímetro
7,273565	`Post Hook`	Finalização com gancho
6,999300	ALT	Altura
6,351776	`Passing Accuracy`	Acurácia no passe
6,058402	`Post Moves`	Movimentos no garrafão
6,050552	`Help Defense IQ`	Habilidade de ajudar na defesa
5,555268	`Standing Dunk`	Enterrada parado
5,355023	`Interior Defense`	Defesa no garrafão
5,076777	`Defensive Consistency`	Consistencia defensiva
5,029084	ENVER	Envergadura
4,295312	`Offensive Rebound`	Rebotes ofensivos
4,193843	`Passing IQ`	Habilidade de passe
4,173701	Strength	Força física
3,943623	`Post Fade`	Finalização com arremesso de costas
3,762858	PESO	Peso
3,710432	Potential	Potencial overall do jogador
3,458762	`Defensive Rebound`	Rebotes defensivos
3,312177	`Offensive Consistency`	Consistencia ofensiva
3,215118	`Shot IQ`	Capacidade de arremesso
3,071173	`Passing Vision`	Visão de passe
3,007875	Block	Habilidade de bloquear arremessos
2,888783	`Driving Layup`	Bandeija
2,768330	Vertical	Altura do salto
2,751296	`Shot 3pt`	Arremesso longo
2,601435	`Pass Perception`	Capacidade de prever passes
2,401646	`Driving Dunk`	Enterrada em movimento
2,398424	`Draw Foul`	Faltas forçadas
2,256007	Stamina	Resistência
2,236436	`Shot Mid`	Arremesso médio
1,990509	Hands	Controle de bola
1,971215	`Free Throw`	Lance livre
1,861958	`Shot Close`	Arremesso curto
1,762671	Steal	Habilidade de roubo de bola
1,499137	Hustle	Esforço

Fonte: Elaborado pela autora (2024)

Variáveis com alta multicolinearidade: *Speed With Ball* (14.71) e *DEF* (11.70).

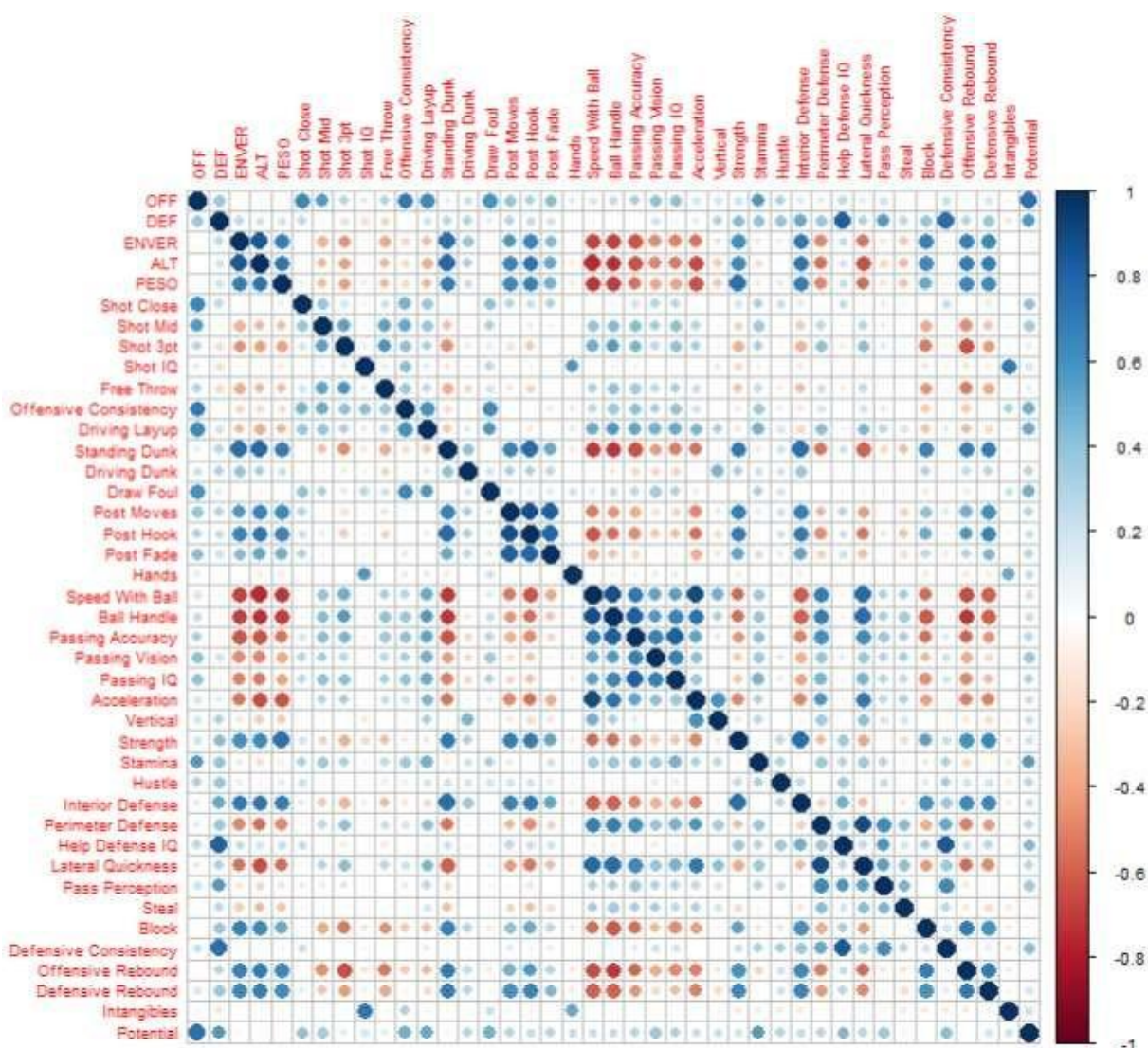
Essas variáveis podem estar muito correlacionadas com outras no modelo. A PCA vai

corrigir essa questão ao combinar as variáveis criando os componentes principais, que serão linearmente independentes.

7.3 Análise de Componentes Principais

Após a criação da matriz de correlação que é apresentada na Figura 5, foi aplicada a análise de componentes principais (PCA), com os resultados dos 6 primeiros componentes organizados na Tabela 4.

Figura 5: Matriz de Correlação das principais variáveis originais



Fonte: Elaborado pela autora (2024)

Tabela 4 - Resultados da PCA nos 6 primeiros componentes.

	PC1	PC2	PC3	PC4	PC5	PC6
Desvio Padrão	3,656	2,624	1,976	1,493	1,423	1,188
Proporção da Variância	0,326	0,168	0,095	0,054	0,049	0,034
Proporção Acumulada	0,326	0,494	0,589	0,643	0,693	0,727

Fonte: Elaborado pela autora (2024)

De acordo com o desvio padrão dos componentes (*Standard Deviation*) PC1 é o que mais captura a variação dos dados, visto que seu desvio padrão é o maior (3,66). Esse componente também explica a maior parte da variância (32%).

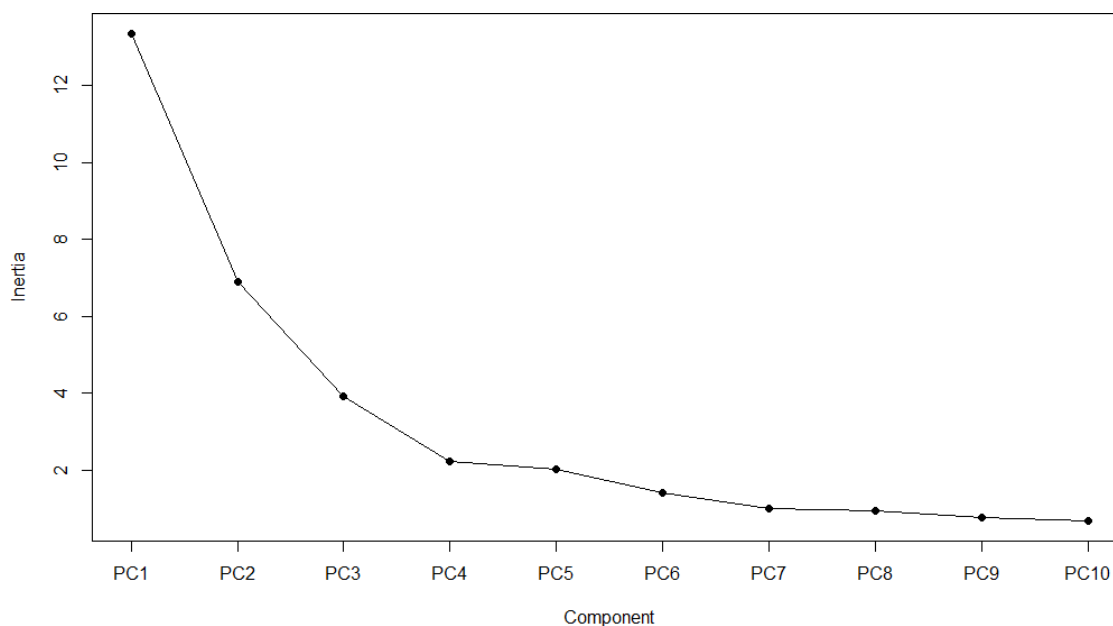
Como no sexto componente já é explicada 72% da variância total (*Cumulative Proportion* = 0,72) pode-se antecipar que o número de componentes escolhidos será seis.

Tabela 5 - Recorte de Componentes Principais

Componente	Autovalores	Proporção de Variância	Variância Acumulada
PC1	13,3631218	0,3259298	0,3259298
PC2	6,8863377	0,16795946	0,4938893
PC3	3,9033662	0,09520405	0,5890933
PC4	2,2288786	0,05436289	0,6434562
PC5	2,0253059	0,04939771	0,6928539
PC6	1,4109769	0,03441407	0,727268
PC7	0,9883455	0,02410599	0,751374
PC8	0,9310476	0,02270848	0,7740824
PC9	0,7573709	0,01847246	0,7925549
PC10	0,667158	0,01627215	0,8088271

Fonte: Elaborado pela autora (2024)

A recomendação é realizar o corte quando a variação entre os autovalores começa a ser pequena e pelo menos 70% da variância foi explicada. Tudo indica que 6 componentes principais são suficientes, o que é confirmado pelo *scree plot* que mostra o “cotovelo” no componente 6, conforme ilustra a Figura 6.

Figura 6: Gráfico cotovelo (*scree plot*)

Fonte: Elaborado pela autora (2024)

7.4 Análise de Cluster Não Hierárquico

Com os seis componentes principais selecionados foi realizada uma análise de *cluster* não hierárquica para agrupar os jogadores com base em suas características estatísticas. Podemos ver os grupos em 2 dimensões (PC1 e PC2) na Figura 7.

Inicialmente, foi calculada a distância de Mahalanobis, em seguida foi aplicado o método *K-means*, utilizando a matriz de distâncias obtida como entrada para definir os agrupamentos. Os jogadores foram agrupados em cinco *clusters*.

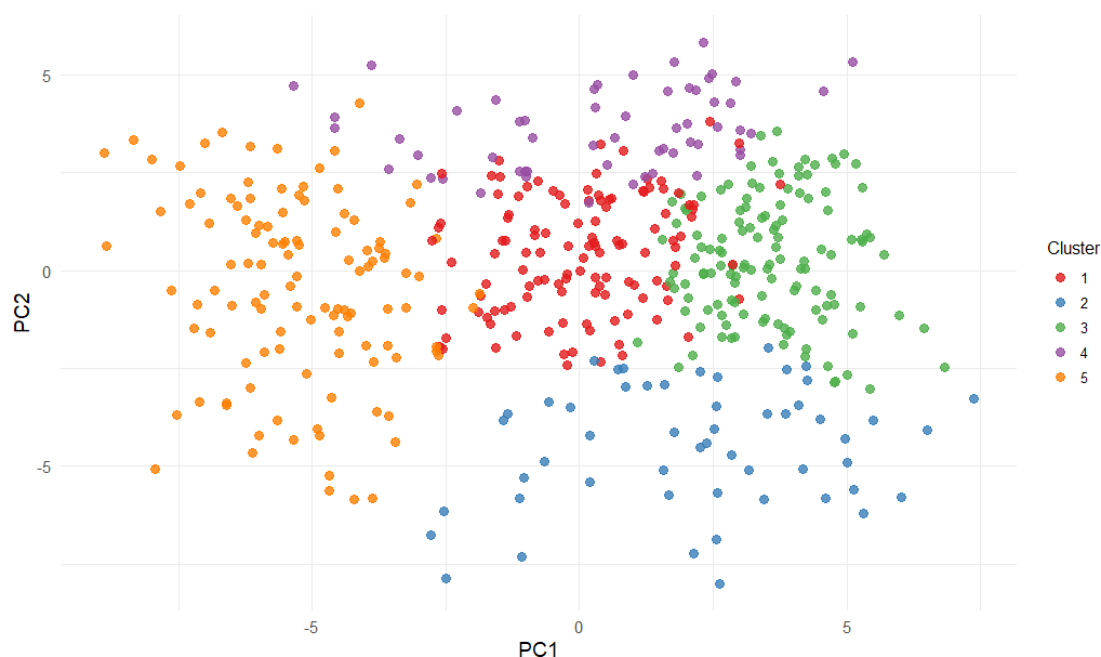
Conforme esperado, os grupos representaram as cinco posições tradicionais do basquete: armador, ala-armador, ala, ala-pivô e pivô. Na Tabela 6 vemos a distribuição dos jogadores nos *clusters*.

Tabela 6 - Número de jogadores em cada *cluster*

Cluster	Número de Jogadores
1	119
2	53
3	127
4	57
5	113

Fonte: Elaborado pela autora (2024)

Figura 7: Visualização dos *clusters* em duas dimensões (PCA)



Fonte: Elaborado pela autora (2024)

Para visualizar como os jogadores foram distribuídos nos grupos é importante ressaltar algumas características de cada posição:

- Armador (PG) - Ataque: Melhores em estatísticas ofensivas e arremessos, não são tão altos e fortes nem se destacam em bloqueios e infiltrações;
- Ala-Armador (SF) - Ataque: Melhores em ofensivas, arremessos de média distância, infiltrações e passes. É melhor que o armador na defesa e são muito bons com defesa de perímetro;
- Ala (SG) - Ataque: Menos velozes que os dois anteriores, mas muito habilidosos nas ofensivas e arremessos e inferiores na defesa;
- Ala-Pivô (PF) - Defesa: Altos, mais ágeis e velozes que os pivôs e melhores nos arremessos e roubos, mas não tão fortes e piores com enterradas e bloqueios;
- Pivôs (C) – Defesa: Os mais altos e fortes, melhores em bloqueios e enterradas, piores na defesa do perímetro por ficarem a maior parte do jogo perto da cesta e os piores em arremessos longos.

Sabendo essas características e observando as estatísticas nas Tabelas 7 e 8 é possível perceber que, de fato, cada grupo representa uma posição conforme organizado na Tabela 9.

Tabela 7 - Médias de variáveis originais importantes dos *clusters*

<i>Cluster</i>	Força	Velocidade	Saltos	Roubos	Arr. com Salto	Enterrada	Def. de Perímetro
1	57.9	63.6	75.0	48.3	51.0	48.2	67.7
2	61.2	74.4	79.5	51.6	57.4	48.2	74.3
3	47.3	73.8	72.5	49.6	44.4	30.7	67.8
4	48.3	60.1	68.5	43.0	43.1	39.2	55.7
5	72.1	38.9	67.5	38.6	59.9	75.2	46.3

Fonte: Elaborado pela autora (2024)

Tabela 8 - Estatísticas dos *clusters*

<i>Cluster</i>	Ofensiva	Defensiva	Altura	Passes	Bloqueios	Arremessos	Arr. Longo	Arr. Médio	Infiltração
1	73,17	71,22	2,00	62,09	51,95	82,55	73,57	67,81	72,08
2	81,43	78,16	1,98	75,58	49,25	93,47	76,70	78,20	82,34
3	75,35	69,63	1,93	73,96	43,26	90,73	75,73	74,48	76,80
4	70,62	63,10	1,97	58,18	46,52	96,62	70,25	67,46	67,68
5	74,91	73,00	2,08	45,54	69,01	85,98	59,41	64,69	68,42

Fonte: Elaborado pela autora (2024)

Tabela 9 - Posições e os *clusters*

<i>Cluster</i>	Posição
1	PF
2	SF
3	PG
4	SG
5	C

Fonte: Elaborado pela autora (2024)

Para confirmar essa divisão foi feita a Tabela 10, que mostra a frequência em porcentagem das posições em cada grupo.

Tabela 10 - Frequência em porcentagem

Cluster	Posição	Freq. em Porcentagem
1	PF	41,20%
2	SF	30,20%
3	PG	31,50%
4	SG	33,33%
5	C	81,42%

Fonte: Elaborado pela autora (2024)

Confirmada a separação por posições, seguimos para a montagem do time equivalente pelo método simplex, em que será selecionado um jogador de cada *cluster*.

7.5 Método Simplex

A parte final da análise é a aplicação do método simplex para a montagem do time. Primeiro foi criada uma base de dados apenas com o time dos sonhos (figura 7), e dentro de cada *cluster* os jogadores regulares foram comparados com as estrelas com base em métricas de desempenho como ofensiva, defensiva, inteligência de arremesso, precisão em arremessos de três pontos e bloqueios.

A distância euclidiana foi utilizada como critério para identificar o jogador mais similar a cada estrela.

Em vez de utilizar o salário como restrição, foram priorizadas estatísticas importantes para cada posição. Essa abordagem garantiu que os jogadores regulares escolhidos fossem os mais similares às estrelas em métricas como ofensiva, defensiva, inteligência de arremesso, precisão em arremessos de três pontos e bloqueios, independentemente de seu custo.

A decisão de não utilizar o salário como restrição foi tomada porque os jogadores mais semelhantes às estrelas nem sempre são os mais bem pagos entre os regulares. Como todos os outliers (jogadores com salários excepcionalmente altos) foram removidos antes da análise, já se previa que o custo seria inferior ao teto salarial, mesmo sem considerá-lo como critério direto.

Após a seleção, foi analisado o custo total do elenco para assegurar que ele se manteria dentro de limites razoáveis. Esse processo garantiu um time equivalente ao time dos sonhos, com estatísticas próximas e um custo significativamente mais baixo. O time é apresentado a seguir na Tabela 11.

Tabela 11 - Time equivalente

Posição em inglês	Posição em português	Jogador	Overall (22-23)
<i>Point Guard</i> (PG)	Armador	C.J. McCollum	83
<i>Shooting Guard</i> (SG)	Ala-armador	Tyrese Haliburton	81
<i>Small Forward</i> (SF)	Ala	Andrew Wiggins	83
<i>Power Forward</i> (PF)	Ala-pivô	Evan Mobley	84
<i>Center</i> (C)	Pivô	Karl-Anthony Towns	86

Fonte: Elaborado pela autora (2024)

7.6 Comparação de preço

Como o valor do time equivalente não foi considerado no método simplex é preciso confirmar que esse custo não ultrapasse o teto salarial.

Primeiro foi calculado o preço do time dos sonhos, conforme a Tabela 12, incluindo a *luxury tax* (multa aplicada ao time quando o teto é ultrapassado).

Tabela 12 - Custo do time dos sonhos

Jogadores	Overall (22-23)	Salário em Dólar
Stephen Curry	94	48.070.014
Luka Doncic	94	37.096.500
LeBron James	94	44.474.988
Giannis Antetokounmpo	94	42.492.492
Nikola Jokic	94	33.047.803
Elenco	94	205.181.797

Fonte: Elaborado pela autora (2024)

A multa é adicionada de forma progressiva, foi feita a Tabela 13 para melhor entendimento.

Tabela 13 - *luxury tax*

Faixa Acima do Limite (U\$)	Taxa por Dólar
0 - 4.999.999	1,50
5.000.000 - 9.999.999	1,75
10.000.000 - 14.999.999	2,50
15.000.000 - 19.999.999	3,25
20.000.000 - 24.999.999	3,75
Cada 5M adicional	+0,50 por dólar

Fonte: Elaborado pela autora (2024)

O teto salarial da temporada 22-23 foi de 150.267.000 dólares. Qualquer valor acima disso é taxado progressivamente de acordo com a figura acima. O custo do time dos sonhos ultrapassa o teto salarial em aproximadamente 55 milhões de dólares:

$$205.181.797 - 150.267.000 = 54.914.797$$

Agora, aplicamos as taxas progressivas:

- 0 - 4.999.999: $5.000.000 \times 1,50 = 7.500.000$
- 5.000.000 - 9.999.999: $5.000.000 \times 1,75 = 8.750.000$
- 10.000.000 - 14.999.999: $5.000.000 \times 2,50 = 12.500.000$
- 15.000.000 - 19.999.999: $5.000.000 \times 3,25 = 16.250.000$
- 20.000.000 - 24.999.999: $5.000.000 \times 3,75 = 18.750.000$
- 25.000.000 - 54.914.797: $29.914.797 \times 4,25 = 127.636.886$

O valor total (salários + *luxury tax*) seria:

$$205.181.797 + 191.386.886 = 396.568.683 \text{ dólares}$$

Logo, o custo do time dos sonhos é de aproximadamente 400 milhões de dólares, um valor extremamente alto que ultrapassa o valor do elenco mais caro da liga, o *Golden State Warriors* com o valor de aproximadamente 200 milhões de dólares.

Em contrapartida, o custo do time ideal (equivalente às estrelas) é consideravelmente menor, conforme a Tabela 14.

Tabela 14 - Custo do time equivalente

Jogadores	Overall (22-23)	Salário
C.J. McCollum	83	33.333.333
Tyrese Haliburton	81	4.215.120
Andrew Wiggins	83	33.616.770
Evan Mobley	84	8.882.640
Karl-Anthony Towns	86	33.833.400
Elenco	83	113.881.263

Fonte: Elaborado pela autora (2024)

Então o time equivalente custa em torno de 114 milhões de dólares, é 70% mais barato que o time dos sonhos e não ultrapassa o teto salarial e por isso não há multa aplicada.

Como foi usada a temporada de 2022-2023 é interessante observar como está o *overall* do elenco equivalente na temporada atual (2024-2025) apresentado na Tabela 15.

Tabela 15 - *Overall* atual do elenco equivalente

Jogadores	Overall atual
C.J. McCollum	84
Tyrese Haliburton	90
Andrew Wiggins	84
Evan Mobley	87
Karl-Anthony Towns	89
Elenco	87

Fonte: Elaborado pela autora (2024)

Em duas temporadas os jogadores tiveram um aumento considerável na nota geral, ilustrando como esse elenco tem potencial de crescimento. Até porque ele é formado por jogadores mais jovens que muito possivelmente ainda não atingiram seu auge.

7.7 Comparação de estatísticas

Além de comparar o custo dos elencos é importante comparar as estatísticas para se certificar de que o método simplex deu certo e foram selecionados jogadores equivalentes às estrelas. Para isso as estatísticas consideradas mais importantes pela autora foram organizadas na Tabela 16.

Tabela 16 - Comparação de estatísticas

Nome	Posição	Ofensiva	Defensiva	Arremessos	Arr. Longo	Bloqueio	Time
Giannis Antetokounmpo	PF	91	95,5	98,5	67	77,5	Estrelas
Evan Mobley	PF	81	82	98,5	65,5	76,5	Equivalente
LeBron James	SF	94	89	98,5	74,5	70,5	Estrelas
Andrew Wiggins	SF	81,5	80,5	92	81	58,5	Equivalente
Stephen Curry	PG	93,5	78	98,5	94	47,5	Estrelas
C.J. McCollum	PG	83,5	74,5	98,5	81,5	50	Equivalente
Luka Doncic	SG	94	81,5	98,5	84,5	58	Estrelas
Tyrese Haliburton	SG	86	75	97	82	50,5	Equivalente
Nikola Jokic	C	95	77	97	79,5	57,5	Estrelas
Karl-Anthony Towns	C	87	80	89,5	80	63	Equivalente

Fonte: Elaborado pela autora (2024)

É notável que, de acordo com os dados, o time montado é compatível com o time dos sonhos. Formado de jogadores jovens e promissores, o elenco equivalente tem potencial para se tornar um grande time e custa consideravelmente menos que o time de estrelas.

8 CONSIDERAÇÕES FINAIS

Este trabalho buscou explorar o uso de métodos estatísticos e de otimização para montar um time de basquete equivalente ao composto por estrelas da NBA, respeitando restrições práticas.

A aplicação integrada da análise de componentes principais (PCA), análise de cluster e método simplex permitiu desenvolver um modelo que combina rigor analítico com aplicabilidade prática no cenário esportivo.

A PCA foi essencial para reduzir a dimensionalidade dos dados, facilitando a identificação das variáveis mais relevantes para a análise sem comprometer a interpretação. Em seguida, a análise de cluster agrupou os jogadores de forma coerente, separando-os em cinco grupos representativos das posições tradicionais do basquete. Por fim, o método simplex foi aplicado para selecionar um time similar ao time das estrelas, utilizando como critério as estatísticas mais importantes de cada posição.

Optou-se por não incluir o salário como restrição direta, mas a análise posterior demonstrou que o custo do elenco gerado estava abaixo do teto salarial, em função da exclusão de *outliers* no início do estudo.

Os resultados obtidos mostraram que é possível montar um elenco competitivo, com características estatísticas semelhantes às estrelas, a um custo significativamente inferior. Isso reforça o potencial das técnicas de análise multivariada e otimização para oferecer soluções estratégicas no gerenciamento de equipes esportivas.

Como limitações, destaca-se a simplificação de considerar apenas algumas estatísticas, deixando de lado fatores qualitativos como experiência, sinergia entre jogadores e potencial de crescimento. Estudos futuros podem expandir a análise para incluir essas dimensões, além de incorporar métodos mais avançados de otimização e aprendizado de máquina.

Por fim, este trabalho contribui para o uso da ciência de dados como ferramenta no esporte, mostrando como a estatística pode ser aplicada para responder questões complexas, como a montagem de equipes, de maneira objetiva e eficaz.

REFERÊNCIAS

Anderson, T. W. *An introduction to multivariate statistical analysis*. New York: Wiley, 2018.

Araujo, F. R. *Métodos de análise multivariada*. São Paulo: Atlas, 2023.

Araujo, M. O. *Análise, seleção e agrupamento de características relevantes das equipes da NFL durante os anos de 2013-2016 e suas relações: um estudo de caso*. 2022. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal de São Carlos, São Carlos. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/17047>. Acesso em: 10 jun. 2023.

Chvátal, V. *Linear programming*. San Francisco: W. H. Freeman, 1983.

Dantzig, G. B. *Maximization of a linear function of variables subject to linear inequalities*. RAND Corporation, 1947.

Dutta, P. *How data analytics has revolutionized the sport of basketball & the NBA*. 2021. Disponível em: <https://medium.com/@pratyushdutta29/how-data-analytics-is-changing-the-sport-ofbasketball-the-nba-f85aa31e9ebb>. Acesso em: 11 dez. 2022.

Everitt, B. S.; Landau, S.; Leese, M.; Stahl, D. *Cluster analysis*. 5th ed. Chichester: Wiley, 2011.

Géron, A. *Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow*. Rio de Janeiro: Alta Books, 2019.

Hair, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. *Multivariate data analysis*. 7th ed. Upper Saddle River: Prentice Hall, 2009.

Irving, K. *How are NBA 2K player ratings crafted? Formulating the best players, hardest players to rate and new additions to the game*. 2021. Disponível em: <https://www.sportingnews.com/au/nba/news/nba-2k-player-ratings-bestplayers/18e218d58c6411rqi8btvqnarx>. Acesso em: 29 nov. 2022.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning: with applications in R*. New York: Springer, 2013.

Kaggle. *Kaggle: your machine learning and data science community*. [S. l.: s. n.], [s. d.]. Disponível em: www.kaggle.com. Acesso em: 21 set. 2022.

Lawson, Richard. *Web Scraping with Python*. Birmingham: Packt Publishing, 2015.

Menezes, J. *Rumo ao estrelato: uma análise com estatísticas de jogadores recém ingressados na NBA*. 2021. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal de Campina Grande, Campina Grande. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/xmlui/handle/riufcg/25005>. Acesso em: 11 dez. 2022.

Miguel, C. *Sucesso na NBA e seus possíveis preditores*. 2022. Trabalho de Conclusão de Curso (Bacharelado em Educação Física) – Universidade Federal de Santa Catarina, Florianópolis. Disponível em: <https://repositorio.ufsc.br/handle/123456789/233489>. Acesso em: 11 dez. 2022.

Mingoti, S. A. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belo Horizonte: UFMG, 2007.

Mitchell, R. *Web scraping com Python: coletando mais dados da web moderna*. São Paulo: Novatec Editora, 2019.

Morey, Daryl. Usando dados para ditar as regras. 2015. Disponível em: https://insight.kellogg.northwestern.edu/pt/article/using-data-to-call-the-shots?utm_source=chatgpt.com. Acesso em: 10 jul. 2024.

Sartori, G. T. *Abordagem multivariada para comparação de atletas de futebol*. 2021. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: <http://hdl.handle.net/10183/231275>. Acesso em: 10 jun. 2023.

Silvestre, M. R. *Análise multivariada II: notas de aula*. Presidente Prudente: FCT/Unesp, 2023.

Smith, John. *Data Science Techniques in Basketball Team Optimization: A Comprehensive Review*. Journal of Sports Analytics, 2021. Disponível em: <https://doi.org/10.1234/jsa.v7i2.5678>. Acesso em: 10 jul. 2024.

Smith, S. *Analytics in sports: strategies and success stories*. Chicago: Hoop Press, 2021.

Sportslogos. *2020-2021 NBA team logos*. Disponível em: https://www.sportslogos.net/teams/list_by_year/62021/2020-2021_NBA_Logos/. Acesso em: 21 jun. 2023.

Vanderbei, R. J. *Linear programming: foundations and extensions*. Springer Science & Business Media, 2001.

Varella, C. A. A. *Análise de componentes principais*. 2008. Dissertação (Pós-graduação em Agronomia) – Universidade Federal Rural do Rio de Janeiro, Seropédica. Disponível em: <http://www.ufrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20a%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf>. Acesso em: 10 jun. 2023.

Zuccolotto, P.; Manisera, M. *Basketball data science with applications in R*. CRC Press, 2020.