Universidade de Brasília
Departamento de Estatística

Estatística no Basquete
Análise de desempenho dos jogadores da Liga Nacional de Basquetebol dos Estados Unidos

Tiago Gonçalves de Sampaio Alves

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Tiago Gonçalves de Sampaio Alves

Estatística no Basquete
Análise de desempenho dos jogadores da Liga Nacional de Basquetebol dos Estados Unidos

Orientador(a): Prof(a). André Cançado
Coorientador(a): Prof(a).

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Agradecimentos

- Agradeço à minha família, especialmente meu pai, que me inspirou a entrar na estatística, minha mãe, que sempre me deu ótimos conselhos, e minha irmã, a quem eu sempre quis ser um exemplo;
- Agradeço aos meus professores e amigos do curso, por sempre me inspirarem a ser melhor e compartilharem conhecimento e formação ao longo desses anos;
- Agradeço ao professor André Cançado, por ter aceitado participar desse trabalho e além de revisar, contribuir com novas ideias para que seja exemplar;
- Agradeço à minha colega Amanda Shinkawa, que inspirou esse trabalho (SHINKAWA AMANDA E MONTEIRO, 2022) e me mostrou que nosso limite vai até aonde nossos olhos podem ver;

Resumo

O presente relatório tem como finalidade a análise de desempenho dos jogadores de basquetebol da *National Basketball Association*, abreviada por NBA e principal liga de basquete profissional dos Estados Unidos, através da exploração da correlação entre as variáveis coletadas e estudo das mesmas para tentar diferenciar o desempenho entre os jogadores.

Construiu-se uma análise fatorial exploratória, com objetivo de compreender quais as variáveis mais importantes para descrever o desempenho do jogador, e logo após um modelo de equações estruturais para criar a pontuação de desempenho final. Por último, foram feitas as clusterizações hierárquicas *Average* e *Complete Linkage*, a fim de determinar o número ideal de *clusters* para a clusterização não-hierárquica final *K-Means*, para dividir os arremessadores em grupos de acordo com seus diferentes níveis de desempenho.

Palavras-chaves: análise fatorial, análise fatorial confirmatória, análise fatorial exploratória, basquetebol, clusterização, clusterização hierárquica, clusterização não-hierárquica, equações estruturais

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

O basquete é um esporte coletivo e dinâmico, fundado em 1891 em Massachusetts, nos Estados Unidos, que vem dominando o cenário global. É um esporte dinâmico e efêmero, sendo atualizado constantemente ao longo dos anos pela própria comunidade. Seu principal público se dá pela *National Basketball Association* dos Estados Unidos ou Liga Nacional de Basquete, também denominada NBA, e é a liga de basquete mais competitiva do mundo, abrangendo times com jogadores de todos os continentes. Por conta da popularização do esporte, existe uma presença cada vez maior de jogadores internacionais na Liga, por conta da sua visibilidade, retorno monetário, qualidade de vida e muitos outros motivos.

Por conta da popularização do esporte e crescimento massivo na competitividade da Liga, vem-se tornando cada vez mais necessária a adoção de técnicas mais detalhadas para o treinamento dos jogadores, para a escolha de prêmios e para que a equipe técnica de cada equipe saiba os pontos fortes e a se desenvolver de seus jogadores, além dos pontos fracos das equipes adversárias. Objetivando a implementação e desenvolvimento de técnicas cada vez mais específicas ao esporte, veio também a necessidade da captação de cada vez mais informações e estatísticas dos jogadores e das equipes, objetivando potencializar o desenvolvimento da equipe dentro do jogo e das temporadas.

Na atual conjuntura, são coletados diversos dados das equipes e de seus jogadores, permitindo a exposição mais detalhada de estatísticas e maximização de habilidades específicas dos jogadores, além da construção em seus pontos a se desenvolver. Tal cenário, que permite fácil acesso a uma enorme quantidade de informação sobre os jogadores e suas equipes, abre caminhos para numerosas análises que nas décadas antecedentes não seriam possíveis.

A partir de dados coletados, esta pesquisa objetiva traduzí-los por meio de técnicas de estatística multivariada e modelagem para entender se a criação e publicação desses dados são aproveitadas pela metodologia utilizada, e se conseguem diferenciar jogadores com desempenhos distintos, se consegue criar *clusters* que categorizam jogadores considerados como melhores no mesmo patamar, além de coincidir com resultados de premiações da NBA, como o prêmio *Michael Jordan Trophy*, mais conhecido como o troféu de *Most Valuable Player*, ou jogador mais valioso.

2 Referencial Teórico

Nesta seção, serão descritas as técnicas estatísticas utilizadas neste relatório, a fim de promover melhor compreensão acerca das análises aplicadas.

2.1 Análise de Agrupamentos

A análise de agrupamentos (ou clusterização) é um método primitivo que consiste em agrupar objetos com base em suas similaridades ou distâncias (proximidade), de modo que cada objeto é semelhante aos demais objetos de seu *cluster* e diferente dos objetos de outros grupos. O objetivo é classificar em um mesmo grupo objetos cuja distância entre si seja mínima e maximizar a distância desses objetos para os de outros grupos (HAIR et al., 2009).

2.1.1 Distâncias

Para a determinação de similaridades, a fim de se formar grupos, medidas de distância são muito utilizadas. Algumas das mais conhecidas e que serão utilizadas neste estudo são as distâncias Euclidiana, de Manhattan e de Mahalanobis (HAIR et al., 2009).

Distância Euclidiana

A distância euclidiana é, na prática, obtida pelo cálculo do comprimento da hipotenusa de um triângulo retângulo (HAIR et al., 2009). Seu cálculo pode ser realizado pela fórmula

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}. \tag{2.1.1}$$

A distância euclidiana é a mais utilizada para análise de agrupamentos, uma vez que não há conhecimento prévio dos grupos a serem formados, não sendo necessário assim, ter conhecimento das variâncias e covariâncias das amostras (JOHNSON; WICHERN, 2007).

Distância de Manhattan

A distância de Manhattan emprega a soma das diferenças absolutas das variáveis, ou seja, os dois lados de um triângulo retângulo em vez da hipotenusa (JOHNSON; WICHERN, 2007). Pode ser calculada pela fórmula

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i|, \tag{2.1.2}$$

onde

- x_i é a i-ésima observação para a variável x;
- y_i é a i-ésima observação para a variável y;
- p é o número de dimensões.

Apesar de ser de cálculo mais simples, essa medida pode conduzir a agrupamentos espúrios caso as variáveis sejam altamente correlacionadas (HAIR et al., 2009).

Distância de Mahalanobis

A distância de Mahalanobis consiste em calcular as distâncias após a ponderação igual das variáveis (MARDIA; KENT; BIBBY, 1980). Sendo assim, ela utiliza variáveis padronizadas em seu cálculo, que pode ser feito pela fórmula

$$D^2 = (\bar{x} - \bar{y})' \sigma^{-1} (\bar{x} - \bar{y}), \tag{2.1.3}$$

onde

- $\bar{x}$ é o vetor média $[\bar{x_1}, \bar{x_2}, ..., \bar{x_p}]$;
- $\bar{y}$ é o vetor média $[\bar{y_1}, \bar{y_2}, ..., \bar{y_p}]$;
- σ^{-1} é a matriz de covariâncias.

Basicamente, consiste na distância Euclidiana para variáveis escaladas. Quando as medidas são muito correlacionadas, essa é a distância mais adequada, uma vez que ela realiza o ajuste das correlações (HAIR et al., 2009).

2.1.2 Métodos Hierárquicos

Para criar as aglomerações em *clusters*, é necessário agrupar os elementos de modo que se obtenha a melhor combinação possível. Entretanto, como seria impossível analisar todas as possibilidades de agrupamento, métodos de análise de agrupamentos conhecidos como hierárquicos foram desenvolvidos para fornecer as combinações mais "razoáveis" (JOHNSON; WICHERN, 2007).

Os métodos hierárquicos se dividem em aglomerativos e divisivos. Aqui trataremos apenas de métodos aglomerativos.

Os métodos hierárquicos aglomerativos consistem em, inicialmente, separar cada observação em um *cluster* diferente. A partir daí, em cada iteração, os dois *clusters* mais próximos são agrupados em um único *cluster*. Esse procedimento é repetido até que todos os indivíduos encontrem-se em um único agrupamento final (HAIR et al., 2009). O que difere os métodos é a forma com que as distâncias entre diferentes *clusters* são calculadas.

onde d_{ik} é a distância entre os elementos i e k (JOHNSON; WICHERN, 2007). Ademais, define-se a distância entre o *cluster* formado pelos elementos U e V e o elemento W como $d_{(UV)W}$ (JOHNSON; WICHERN, 2007).

Complete Linkage

O método *Complete Linkage*, também conhecido como método da máxima distância ou do vizinho mais distante, considera que a distância entre dois *clusters* A e B é dada pela maior distância d_{ik} tal que o elemento $i \in A$ e o elemento $k \in B$. (JOHNSON; WICHERN, 2007).

Average Linkage

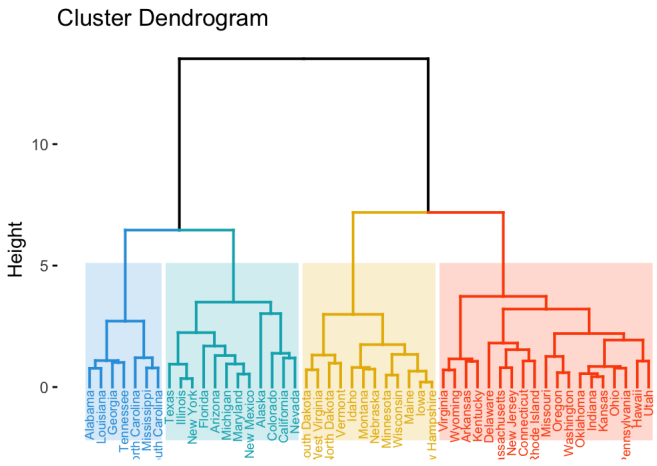
No método *Average Linkage*, também conhecido como método da distância média, a distância entre dois *clusters* A e B é definida como a média de todas as distâncias entre elementos de A e B. (JOHNSON; WICHERN, 2007).

Dendrograma

Uma forma de verificar como os elementos se dividem nos *clusters* é o dendrograma, que consiste em uma representação gráfica dos resultados de um procedimento

hierárquico de análise de agrupamentos. Cada elemento é colocado em um eixo e a distância entre elementos é colocada no outro eixo. Assim, o processo de análise de agrupamentos é repetido até que um único *cluster* seja formado ao final (HAIR et al., 2009).

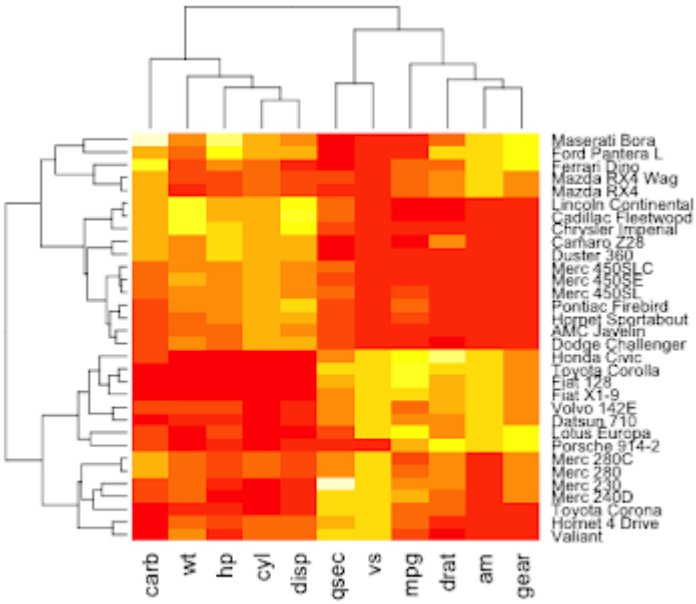
Figura 1: Exemplo de Dendrograma



Fonte: KASSAMBARA (2018)

Outra possibilidade de visualização é a combinação de um dendrograma com um mapa de calor, onde as observações se localizam em um eixo e as variáveis no outro, sendo que as cores evidenciam os grupos mais similares entre si (BARE, 2011).

Figura 2: Exemplo de Mapa de Calor



Fonte: BARE (2011)

2.1.3 Métodos Não-Hierárquicos

Os métodos não-hierárquicos de agrupamento necessitam de um número k de *clusters* determinados previamente ou por meio de métodos hierárquicos de análise de agrupamentos. Esses métodos não necessitam da matriz de distâncias e os dados básicos não precisam ser armazenados durante o processo de formação. Por esses motivos, esses métodos são mais indicados do que os hierárquicos quando o volume de dados é grande. O método não-hierárquico mais conhecido é o método *k-means* (JOHNSON; WICHERN, 2007).

Método *K-Means*

O algoritmo de análise de agrupamentos não-hierárquico *K-Means* consiste nos seguintes passos:

1. Inicie escolhendo k centroides, que são pontos no espaço de variáveis.
2. Associe cada elemento do conjunto de dados ao centroide mais próximo. Os elementos associados ao mesmo centroide formam um *cluster*.
3. Para cada um dos k *clusters* recalcula-se o centroide como a média de seus elementos.
4. Repita os passos 2 e 3 até que, de uma iteração para outra, a composição dos *clusters* não se altere.

No passo 1, a escolha inicial dos k centroides pode ser feita escolhendo-se ao acaso, por exemplo, k elementos distintos do conjunto de dados, ou gerando-se k pontos ao acaso no espaço de variáveis.

O método k-means deve receber como parâmetro o número k de *clusters*. Na próxima seção iremos abordar formas de determinar esse valor.

2.1.4 Fomas de Determinar o Número de Clusters

Para obter-se uma análise de agrupamentos mais precisa, por meio de métodos como *K-Means*, é necessário estabelecer-se o número de *clusters*. Para tal, alguns índices e métodos gráficos foram propostos para auxiliar na determinação do melhor número de *clusters* possível (CHARRAD et al., 2014).

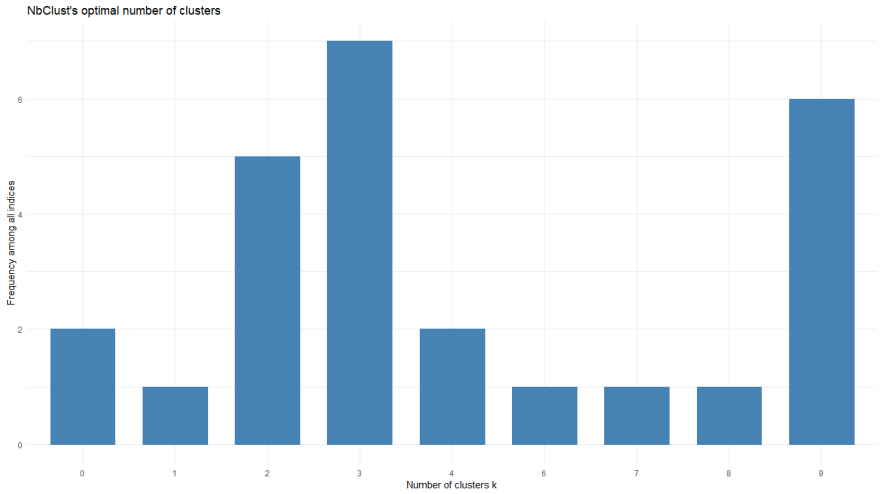
Seguem listados abaixo os índices a serem utilizados, cujos métodos de cálculo podem ser verificados em Charrad et al. (2014).

- Índice CH;
- Índice Duda;
- Índice Pseudot2;
- Índice C;
- Índice Beale;
- Índice CCC;
- Índice Ponto Bisserial;
- Índice DB;
- Índice Frey;
- Índice Hartigan;
- Índice Ratkowsky;
- Índice Scott;
- Índice Marriot;
- Índice Ball;
- Índice Trcovw;
- Índice Tracew;
- Índice Friedman;
- Índice McClain;
- Índice Rubin;
- Índice KL;
- Índice Silhouette;
- Índice D;

- Índice Dunn;
- Índice Hubert;
- Índice SD;
- Índice SDbw.

Os resultados de todos os índices acima podem ser sumarizados em um único gráfico de barras, onde o número de *clusters* mais frequente pode ser utilizado como o número ideal de *clusters*. No exemplo da Figura 3, o número k de *clusters* seria definido como 3, que foi o valor mais frequente.

Figura 3: Exemplo de Gráfico de Índices



Fonte: OLDACH (2019)

3 Metodologia

3.1 Conjunto de dados

Os bancos de dados utilizados nas análises e interpretação dos resultados foram obtidos do site *Basketball Reference*, por meio do uso de técnicas de *Web Scrapping* do pacote *rvest* do *software* estatístico *RStudio*, versão 4.2.0. Os dados são referentes a todos os jogadores das temporadas de 1982-1983 até a última temporada finalizada, dos anos 2021-2022. Um único jogador aparece várias vezes no banco, uma vez para cada temporada que participou, totalizando 21374 linhas, sendo cada linha um jogador de uma temporada.

Foram coletados 3 tabelas de dados, com estatísticas gerais totais dos jogadores, com estatísticas avançadas dos times, e com estatísticas avançadas dos jogadores. A primeira possui 31 colunas, a segunda 29 e a última planilha, 27. Serão selecionadas algumas variáveis da segunda e terceira planilha, que serão adicionadas na primeira planilha a partir do time do jogador, isso objetivando unificar os dados somente na primeira planilha.

Tabela 1: Estatísticas do Jogador - Banco unificado

Sigla	**Variável**
Rk	Identificação do jogador
Player	Nome do jogador
Pos	Posição do jogador: PG (armador), SG (Ala), SF (Ala), PF (Ala-pivô), C (Pivô)
Age	Idade do jogador
Tm	Time do jogador
G	Jogos jogados
GS	Jogos jogados como titular
MP	Minutos jogados
FG	Gols de campo feitos
FGA	Tentativas de gols de campo
FG%	Porcentagem de gols de campo
3P	Cestas de 3 pontos
3PA	Tentativas de cestas de 3 pontos
3P%	Porcenagem de cestas de 3 pontos
2P	Cestas de 2 pontos
2PA	Tentativas de cestas de 2 pontos
2P%	Porcenagem de cestas de 2 pontos
eFG	Porcentagem de gols de campo ajustada
FT	Lances livres feitos
FTA	Tentativas de lances livres
FT%	Porcentagem de lances livres
ORB	Rebotes ofensivos
DRB	Rebotes defensivos
TRB	Rebotes
AST	Assistências
STL	Roubos de bola
BLK	Bloqueios de bola
TOV	Entregas de bola
PF	Faltas cometidas
PTS	Pontos na temporada
year	Ano final da temporada
W	Vitórias na temporada
L	Derrotas na temporada
PER	Nota de eficiência do jogador
TS%	Porcentagem de arremesso verdadeira
OWS	Aumento de vitórias ao time ao ter o jogador no ataque
DWS	Aumento de vitórias ao time ao ter o jogador na defesa
WS	Aumento de vitórias ao time ao ter o jogador no elenco
OBPM	Pontos ofensivos contribuídos acima de um jogador de desempenho mediano a cada 100 posses em uma equipe de desempenho regular
DBPM	Pontos defensivos contribuídos acima de um jogador de desempenho mediano a cada 100 posses em uma equipe de desempenho regular
BPM	Pontos contribuídos acima de um jogador de desempenho mediano a cada 100 posses em uma equipe de desempenho regular
USG%	Medida da porcentagem das jogadas do time que um jogador utiliza quando está em quadra
VORP	Pontos a cada 100 posses de time que um jogador contribuiu acima do nível de um jogador de substituição em 82 jogos de um time de desempenho regular

3.2 Técnicas utilizadas

Após selecionadas as variáveis e unificadas as informações, totalizando 51 colunas, será feita análise descritiva e de correlações das variáveis utilizando gráficos *Boxplot*, Correlogramas e Mapas de calor (BUSSAB; MORETTIN, 2003). Logo após, será realizada a análise fatorial exploratória (JOHNSON; WICHERN, 2007), objetivando reduzir dimensionalmente os dados, ao mesmo tempo que traduzam a variabilidade dos dados e expliquem o traço latente de habilidade do jogador. Em seguida, serão construídos modelos de equações estruturais (BIELBY; HAUSER, 1977), a fim de verificar os resultados

obtidos na análise fatorial exploratória, obtendo-se o modelo que melhor se ajuste aos dados, utilizando critérios de ajuste como o TLI (Índice de Tucker Lewis) (TUCKER; LEWIS, 1973), GFI (Índice de Qualidade de Ajuste) (CHEVLIN; MILES, 1998), RMSEA (Erro Quadrático Médio de Aproximação) (IACOBUCCI, 2009) e BIC (Critério de Informação Bayesiano) (VRIEZE, 2012).

Após esse procedimento, os novos fatores obtidos serão utilizados para separar os jogadores por similaridade, criando *clusters* que forneçam melhor visibilidade quanto às características que melhor descrevem um bom jogador. Serão utilizadas as técnicas hierárquicas com base nas distâncias. Serão avaliados índices e métodos gráficos para determinar o número ideal de *clusters*. Em seguida, será realizada a análise de agrupamentos final pelo método não hierárquico *K-Means*, uma vez que esse método é o mais adequado para grandes volumes de dados. Por fim, serão realizadas novas análises descritivas a fim de entender as características mais discriminantes para cada *cluster*.

Assim, após o desenvolvimento dos devidos modelos e técnicas descritos, espera-se compreender quais são as características de um jogador que mais influenciam em sua habilidade, além de como melhor classificá-los de acordo com semelhanças em suas competências.

4 Resultados

Nesta seção, serão apresentados abaixo os resultados das análises realizadas e suas respectivas interpretações.

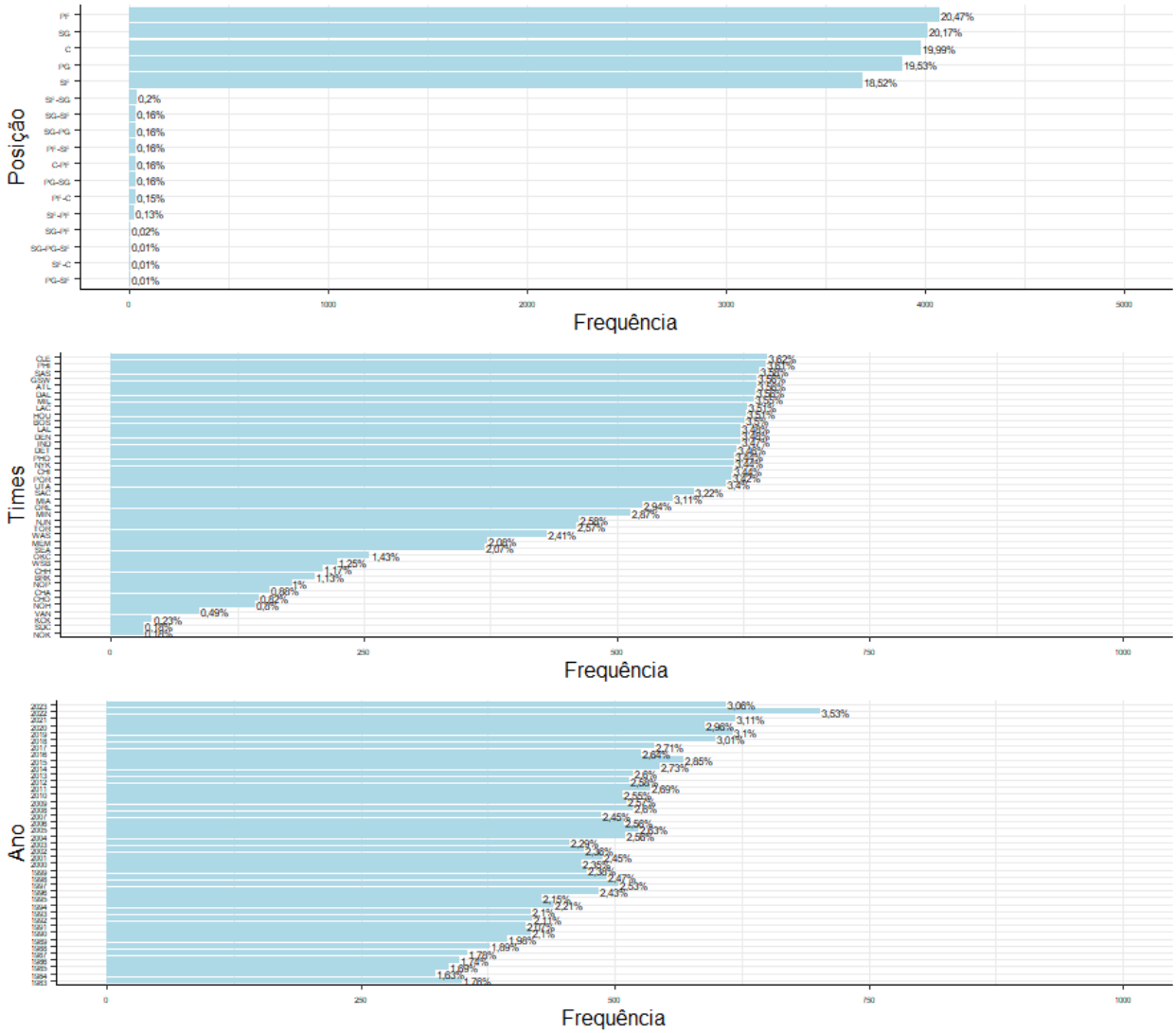
4.1 Análise Descritiva

Primeiramente, fez-se a análise descritiva da amostra, tanto univariada como correlações, objetivando conhecer melhor a amostra de dados obtida e suas relações.

4.1.1 Análise Descritiva Univariada

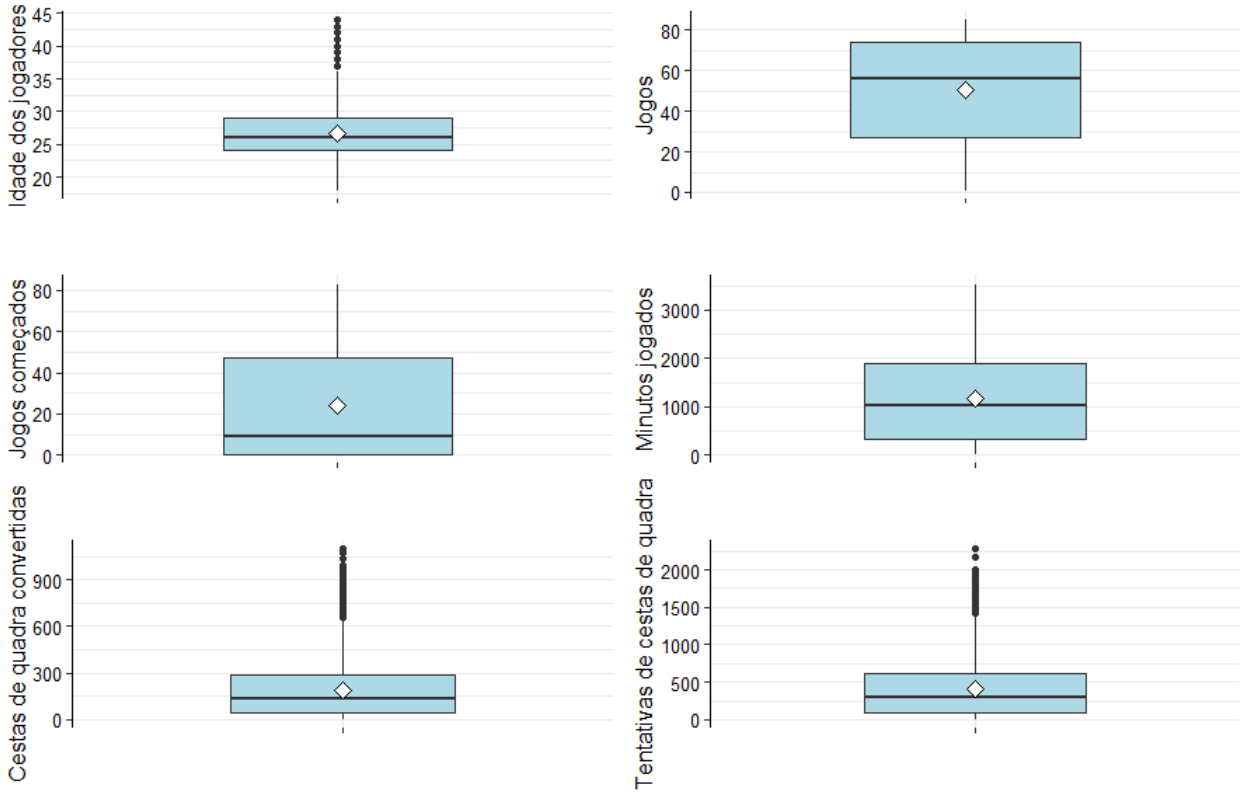
Buscando entender como está distribuída cada variável estudada, foi realizada a análise descritiva de cada variável presente no banco de dados.

Figura 4: Análise Descritiva das Variáveis Qualitativas



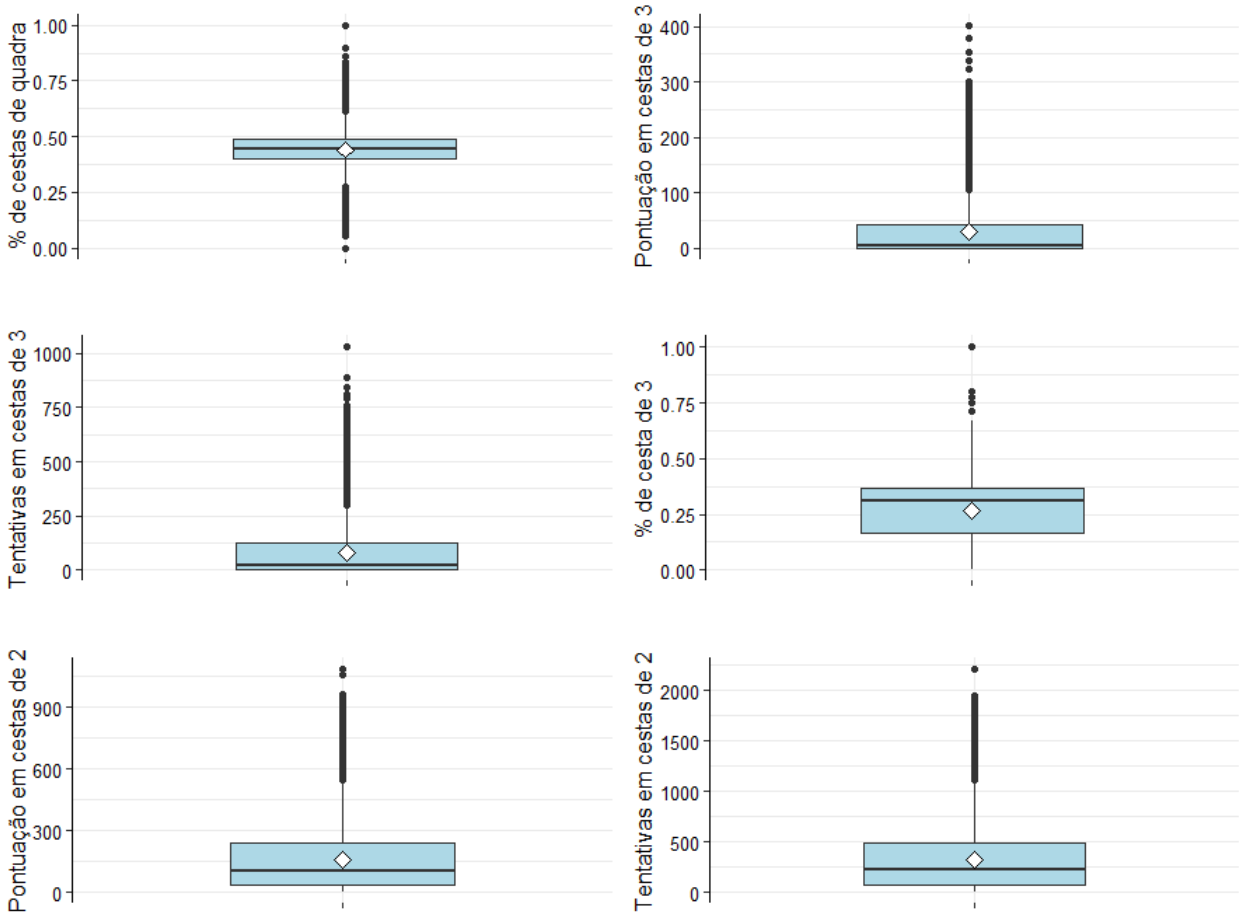
A figura 4 refere-se às variáveis qualitativas presentes no banco de dados. Percebe-se que as posições mais frequentes dos jogadores são as cinco principais e únicas, Ala-pivô, Ala-armador, Pivô , Armador, Ala, sendo o restante posições mistas. Quanto aos times, os mais frequentes são os com mais jogadores e que existem desde o ano de 1983, ano que começa a amostra, os menos frequentes são times que acabaram após essa data ou começaram bem depois dela. Os anos no gráfico são referentes às temporadas dos dados, é perceptível que houve um aumento do número de jogadores desde os anos de início da amostra.

Figura 5: Análise Descritiva das Variáveis Quantitativas



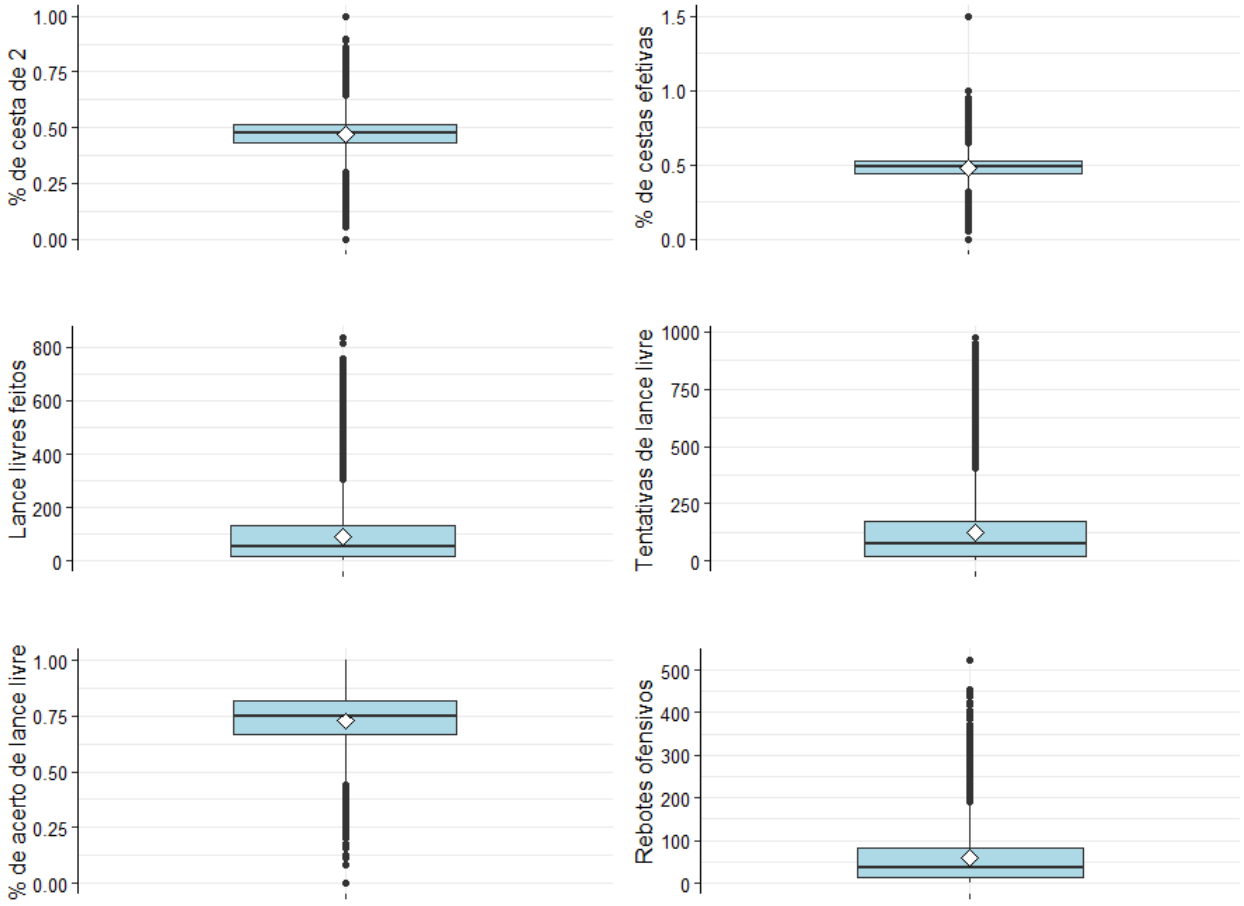
Através da figura 5, nota-se que a Idade dos Jogadores possui uma média próxima da mediana, que possuem valores próximos de 27 anos e com alguns *outliers* acima de 37 anos. Quanto à variável Jogos, não há *outliers* e os jogadores jogam, em média, 50 jogos por temporada de 82 jogos. Entretando, a média de Jogos Começados se aproxima de 30 jogos e é bem acima da mediana da variável, que atinge o valor de 10 jogos. A variável Minutos Jogados também não apresenta valores extremos e a média de minutos jogados, que atinge cerca de 1500 minutos, se aproxima da mediana em valor pouco acima de 1000 minutos. Os jogadores também acertam em média quase 300 cestas por temporada, porém essa variável agora possui um enorme número de valores extremos, que atingem desde 600 cestas até mais de mil cestas por temporada. Agora, quanto às Tentativas de Cestas de Quadra, em média são quase 500 arremessos realizados, porém com *outliers* que atingem mais de 2000 tentativas de arremessos.

Figura 6: Análise Descritiva das Variáveis Quantitativas



Por meio da figura 6, observa-se a Porcentagem de Acerto de Cestas de Quadra dos jogadores, que beira os 50% e diversos valores extremos, que variam entre 0% e 25% além de pouco acima de 60% até 100%. Em Pontuação em Cestas de 3 Pontos, os jogadores possuem média de 28 cestas de 3 pontos na temporada, porém o número de *outliers* existente mostra jogadores que atingem desde 100 até 400 cestas de 3 na temporada. O mesmo acontece com as Tentativas em Cestas de 3 Pontos, onde a média beira as 100 tentativas por temporada, entretando existem valores extremos que passam de 300 tentativas e chegam até mais de 1000 tentativas. A Porcentagem de Cestas de 3 dos jogadores possui média pouco acima de 25% e com poucos *outliers*, que vão de 60% até 100% de acerto. Na Pontuação em Cestas de 2 Pontos, os atletas possuem média pouco acima de 150 cestas de 2 pontos, mas o número de valores extremos na amostra expõe jogadores que superam 600 até mais de 1000 cestas de 2 em 82 jogos da temporada. É paralelo com as Tentativas em Cestas de 2 Pontos, onde a média ultrapassa 300 por temporada, porém existem *outliers* que superam 1000 e chegam até mais de 2000 tentativas.

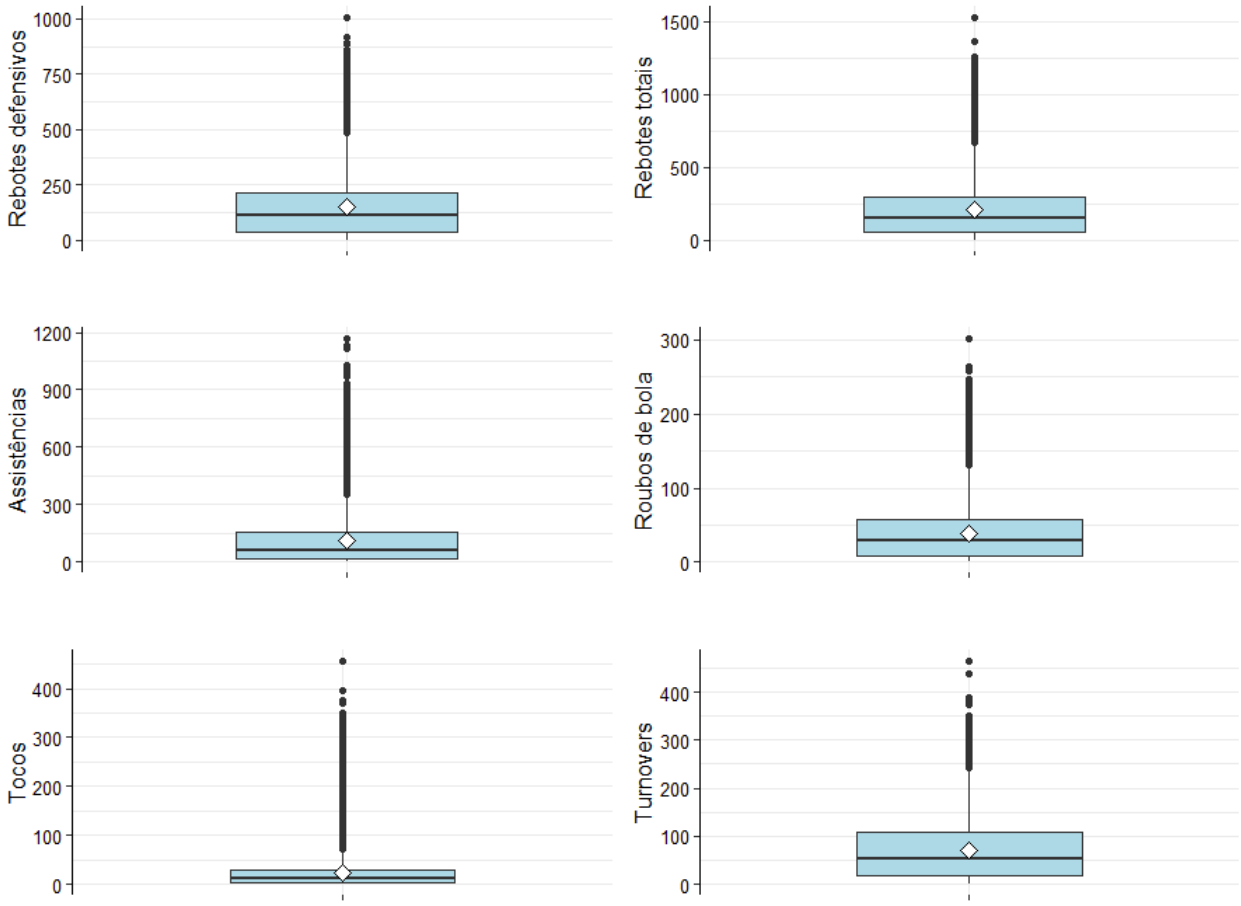
Figura 7: Análise Descritiva das Variáveis Quantitativas



Ao ler a figura 7, interpreta-se da variável Porcentagem de Cesta de 2 Pontos que sua média se aproxima dos 50%, possuindo valores extremos superiores e inferiores, que atingem desde 0% até 100%, respectivamente, comportamento similar a mesma variável, porém referente às cestas de 3 pontos. Paralelamente, Porcentagem de Cestas Efetivas possui comportamento parecido, porém com um *outlier* de valor 1,5, causado pela ponderação feita entre cestas de 2 e 3 pontos na efetividade. Quanto aos Lance Livres Feitos, cada jogador realiza aproximadamente 100 lances livres por temporada em média, porém com a existência de atletas que ultrapassam dos 300 lances livres e chegam até a superar 800 lances livres por temporada. A perspectiva de Tentativas de Lance Livre segue distribuição parecida, porém com média pouco abaixo de 125 lances livres tentados por temporada e valores extremos que se aproximam de 1000 tentativas. A Porcentagem de Acerto de Lance Livre tem média perto de 75%, não possui *outliers* acima do limite superior, somente abaixo do limite inferior e com porcentagens que vão desde abaixo de 50% até 0%. Sobre os Rebotes Ofensivos dos jogadores, também seguem uma distribuição homogênea àquelas dos lances livres, com média de aproximadamente 50 rebotes ofensivos

por temporada, sem valores extremos abaixo do limite inferior e com valores extremos acima do limite superior que variam desde 200 até pouco mais de 500 rebotes ofensivos por temporada.

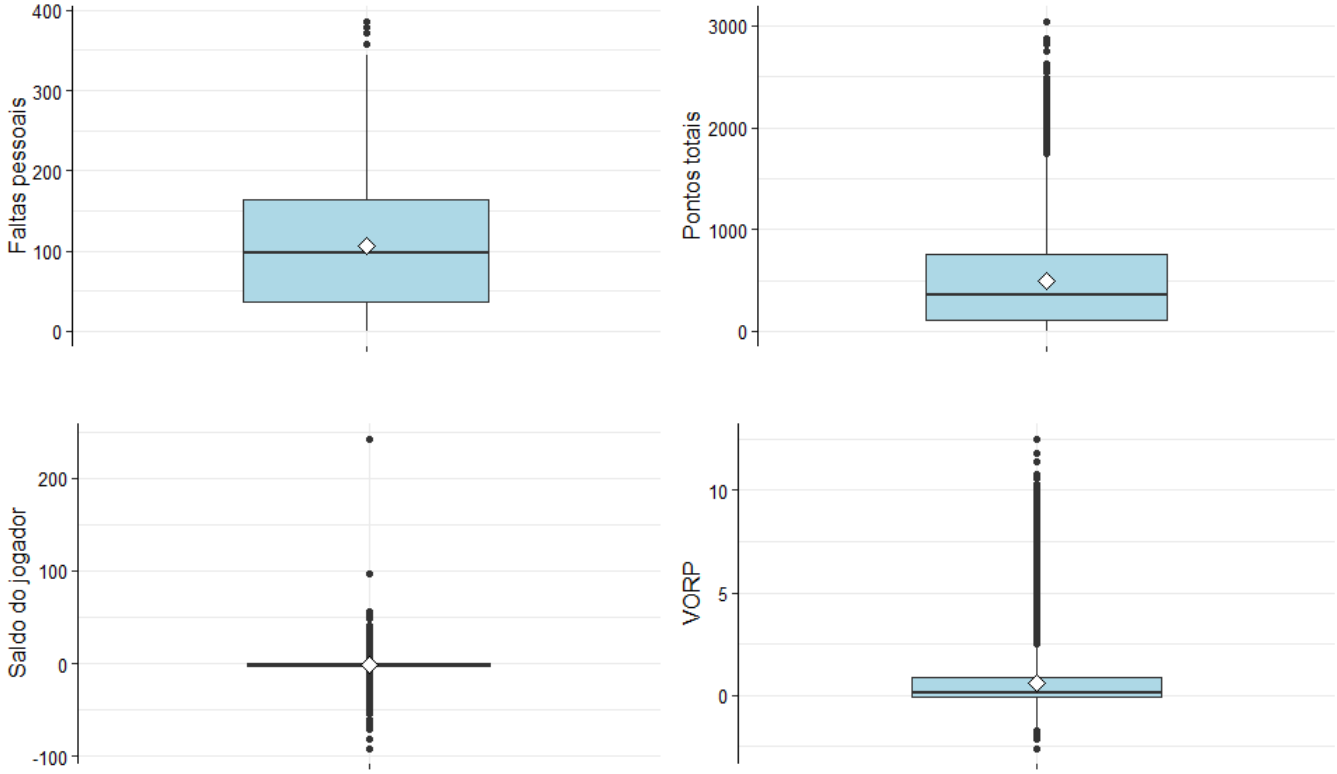
Figura 8: Análise Descritiva das Variáveis Quantitativas



Correspondendo à figura 8, todas as variáveis presentes nessa figura têm distribuição assimétrica à direita ou positiva. Os Rebotes Defensivos dos jogadores possuem média pouco maior que 125 rebotes defensivos por temporada e valores extremos que chegam até os 1000 rebotes, a mesma lógica segue para os Rebotes Totais, que têm média de 200 rebotes e com *outliers* que atingem 1500 rebotes, importante relembrar que a variável Rebotes Totais é a soma de Rebotes Ofensivos e Rebotes Defensivos. Os atletas de basquete ultrapassam 100 assistências na temporada em média, com valores extremos acima do limite superior que se aproximam de 1200 assistências. Os Roubos de Bola acontecem, em média, pouco menos de 50 vezes por jogador em uma temporada, porém com ocorrências que atingem até 300 roubos de bola. Quanto aos tocos, os atletas realizam, em média, aproximadamente 25 tocos por temporada, porém valores extremos passam

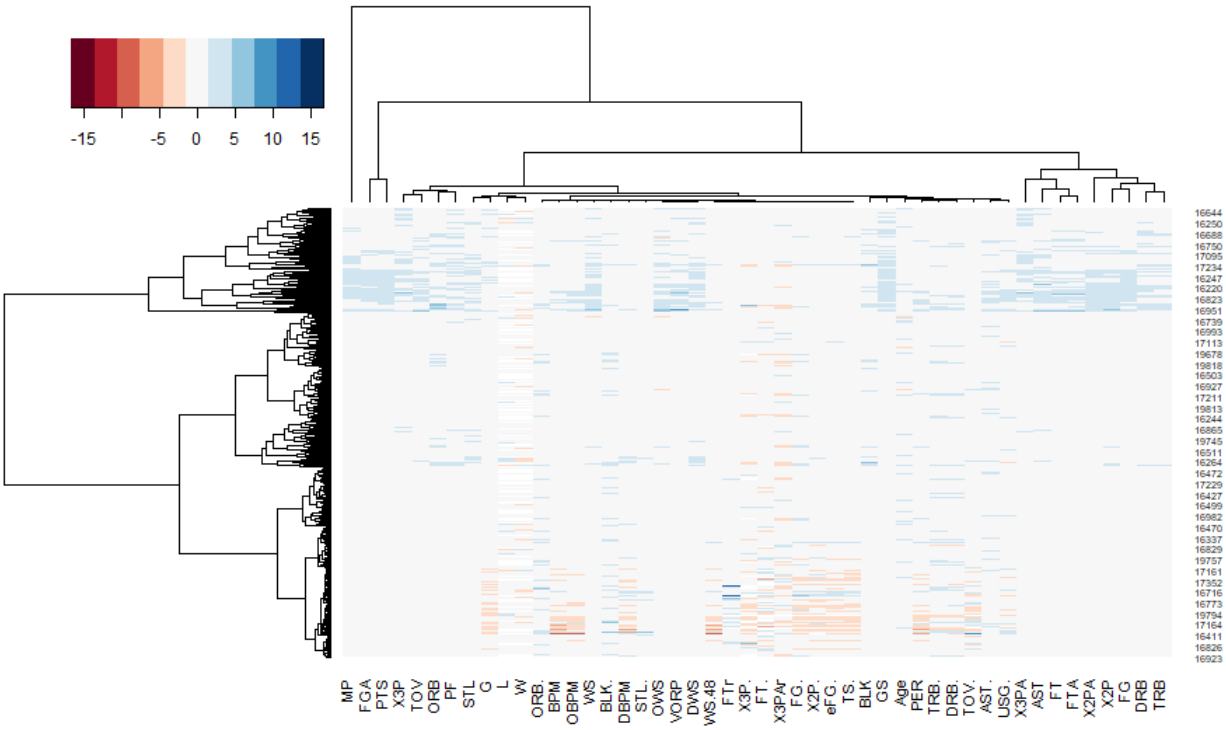
dos 300 tocos por temporada e um atleta chegou a passar de 400 tocos realizados na temporada. Já sobre os *Turnovers* ou Inversões de Posse de Bola, os atletas são responsáveis por inversões de posse cerca de 70 vezes por temporada, entretando *outliers* mostram que alguns jogadores ultrapassaram 200 *turnovers* e chegaram a passar de 400 na temporada.

Figura 9: Análise Descritiva das Variáveis Quantitativas



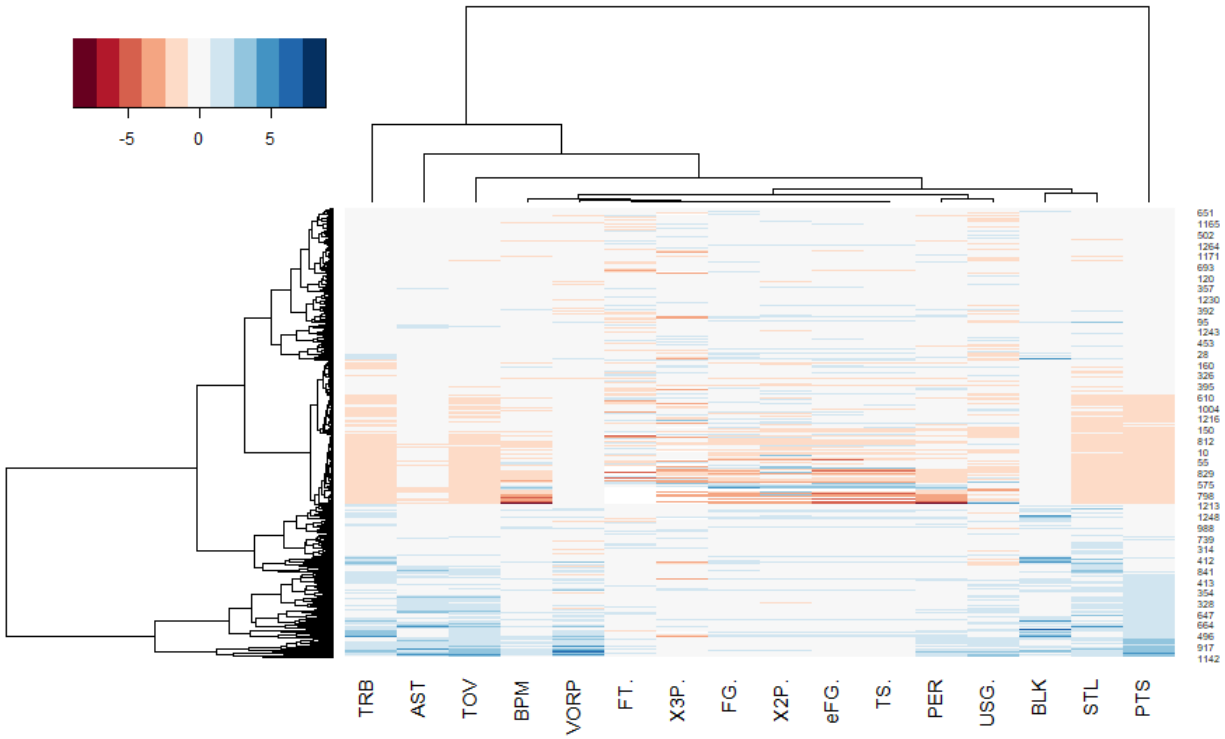
A figura 9 faz representação da distribuição das variáveis Faltas Pessoais e Pontos Totais. Sobre Faltas Pessoais, os jogadores fazem em média 100 faltas por temporada, alguns *outliers* que ultrapassam as 350 faltas pessoais por temporada e a média sendo pouco maior que a mediana indicam uma leve assimetria à direita (ou positiva) na distribuição dessa variável. Acerca da variável Pontos Totais, vê-se que os atletas fazem, em média, mais de 500 pontos por temporada, porém com inúmeros valores extremos que atingem 2000 pontos e chegam a superar 3000 pontos na temporada, o que resulta em também uma distribuição assimétrica positiva, mas mais acentuada. O Saldo do Jogador é uma variável com média centrada no zero e *outliers* variando desde -100 até 100, com apenas uma observação ultrapassando 200 pontos. Sobre a variável de comparação entre um jogador titular e um jogador de substituição, V.O.R.P., a média também é próxima de zero, porém acima da mediana, com poucos valores extremos negativos e majoritariamente valores extremos acima do limite superior, que chegam a ultrapassar os 10 pontos.

Figura 10: Mapa de Calor das Variáveis



Percebe-se pela figura 10 a junção das variáveis utilizadas em 2 ou até 4 fatores de acordo com as semelhanças entre si. Em um fator, as variáveis Tentativas de Três Pontos e de Dois Pontos, Assitências, Lances Livres Feitos, Lances Livres, Porcentagem de Cestas de Dois Pontos, Cestas de Quadra e Rebotes foram classificadas como semelhantes. Um fator seria formado somente pela variável Minutos Jogados, o terceiro pela dupla de variáveis Tentativas de Cestas de Quadra e Pontos, o quarto fator seria formado pelo conjunto de variáveis dado desde Inversões de Posse de Bola até Tentativas de Cesta de Três Pontos.

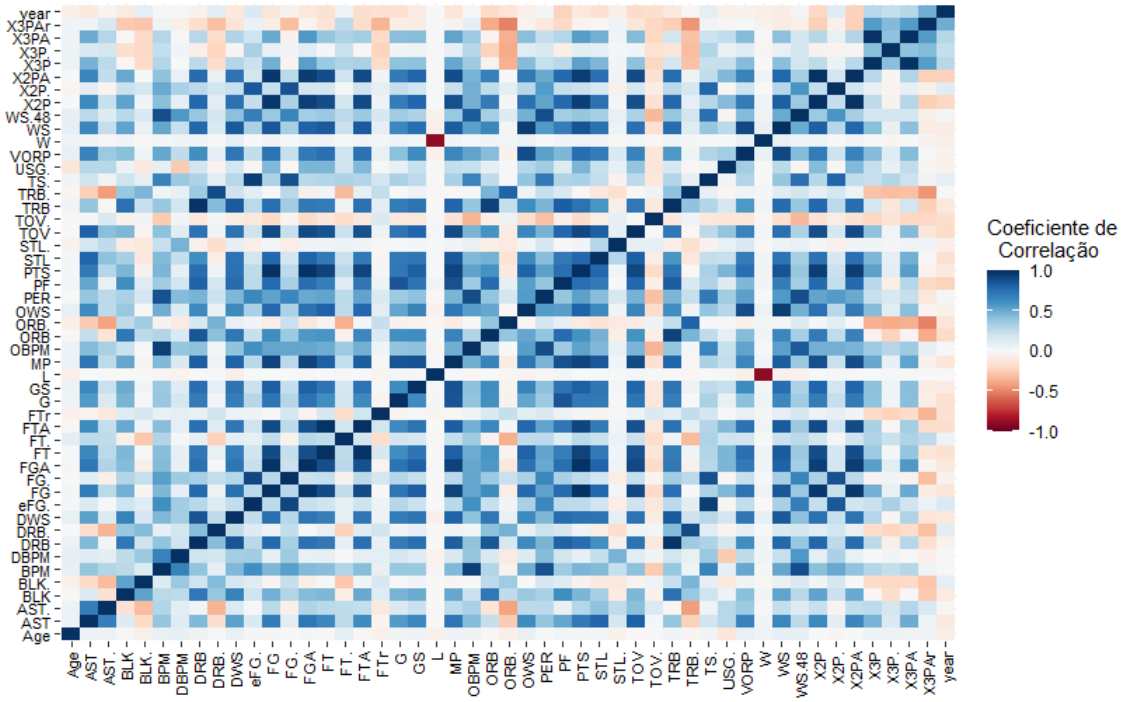
Figura 11: Mapa de Calor das Variáveis Selecionadas



A partir da figura 11, que utiliza um recorte das variáveis apresentadas na figura 10, também mostra a possibilidade da utilização de 2 a 4 fatores, porém com correlações mais fortes entre as variáveis. Três fatores seriam formados individualmente pelas variáveis Rebotes Totais, Assistências e Pontos, o quarto e último fator seria formado pelas variáveis restantes, desde Saldo do Jogador até Roubos de Bola.

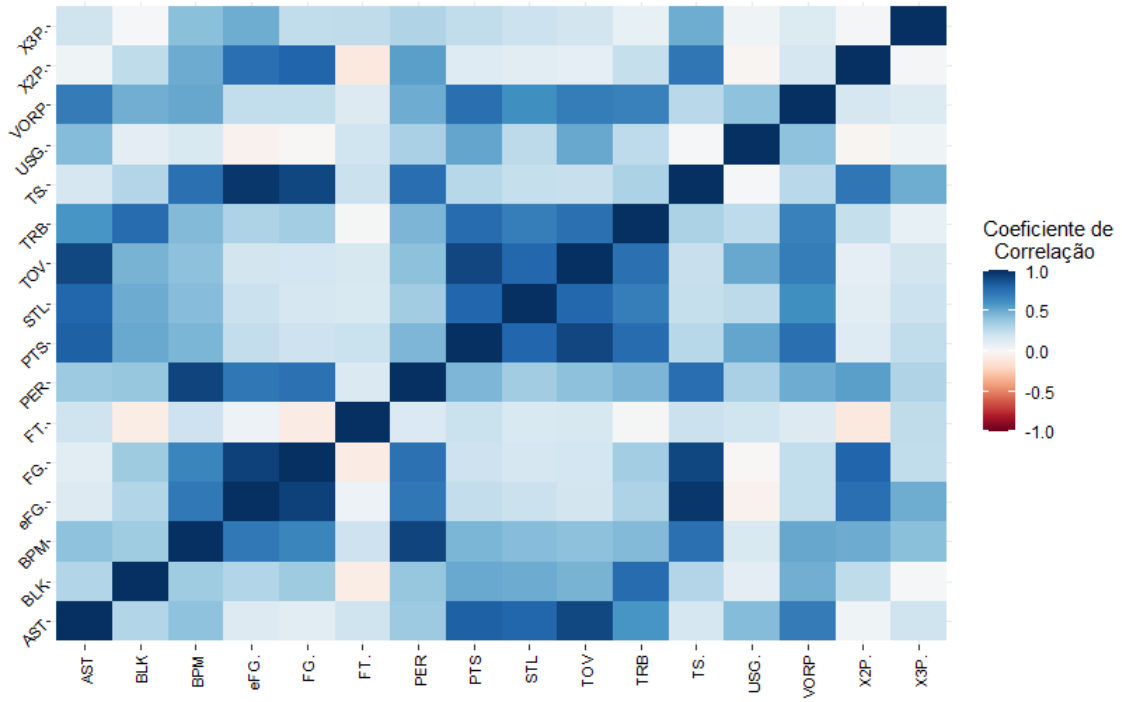
4.1.2 Análise de Correlações

Figura 12: Análise de Correlações das Variáveis



Pela figura 12, nota-se majoritariamente correlações positivas entre as variáveis. Percebe-se alguns grupos de variáveis altamente correlacionadas entre si, as variáveis X3P, X3P%, X3PA, X3PAr se agrupam e dizem respeito aos arremessos de 3 pontos. Outro grupo de variáveis fortemente correlacionadas são OWS, PER, PF, PTS e STL que medem a efetividade do jogador e feitos que acontecem principalmente em situação de contra-ataque.

Figura 13: Análise de Correlações das Variáveis Selecionadas



Tomando percepção de análises realizadas e testes feitos, foram retiradas variáveis que são combinações lineares umas das outras e filtradas as principais variáveis do estudo. Restaram: X3P%, X2P%, VORP, USG%, TS%, TRB, TOV, STL, PTS, PER, FT%, FG%, eFG%, BPM, BLK e AST, totalizando dezesseis variáveis. Com isso, nota-se maior correlação entre as variáveis e também majoritariamente positivas. As variáveis PTS, STL. TOV e TRB estão fortemente correlacionadas entre si, assim como as variáveis BPM eFG% e FG% também estão.

4.2 Análise Fatorial

A princípio, fez-se uma análise fatorial exploratória utilizando os anos das temporadas de 2021 e 2022 como base e todas as 48 variáveis numéricas presentes no banco, porém depois de múltiplas análises e leituras de resultados, foram retiradas as variáveis que eram combinações lineares ou que compunham outras variáveis. Ao fim, resultou no uso de 17 variáveis principais ou selecionadas, sendo elas: FG%, X3P%, X2P%, eFG%, FT%, TRB, AST, STL, BLK, TOV, PTS, PER, TS%, USG%, BPM, WS e VORP. Contudo, a variável WS foi retirada do grupo de variáveis selecionadas para que fosse usada para cálculo da correlação com o escore final de habilidade a ser construído, a fim de validar o escore de habilidade.

4.2.1 Análise Fatorial Exploratória

Tendo em vista as 16 variáveis selecionadas do banco de dados, previamente à análise fatorial exploratória, realizaram-se um *Scree plot* e um *Parallel Analysis plot* para determinar o melhor número de fatores para os dados (GRIS et al., 2018).

Figura 14: *Screeplot* das Variáveis Selecionadas

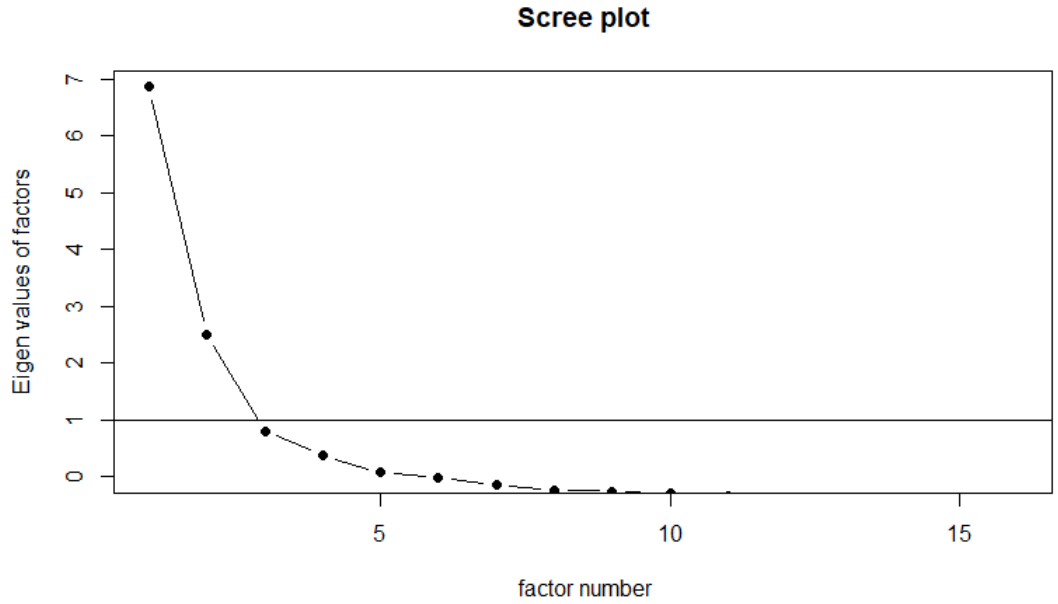
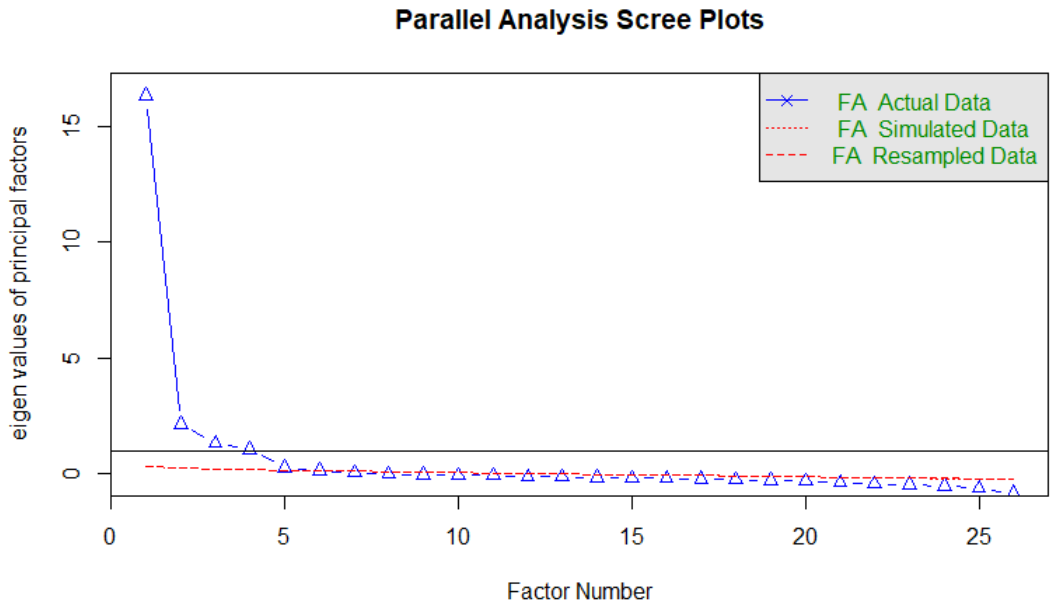


Figura 15: *Parallel plot* das Variáveis Selecionadas



Enquanto o *Scree plot* recomendou o uso de 2 ou até 3 fatores, o *Parallel Analysis plot* sugeriu que fossem construídos 4 ou 5 fatores. Então, foram construídos modelos com diversos fatores e suas medidas foram comparadas:

Tabela 2: Medidas de Ajuste para Modelos com Variáveis Selecionadas

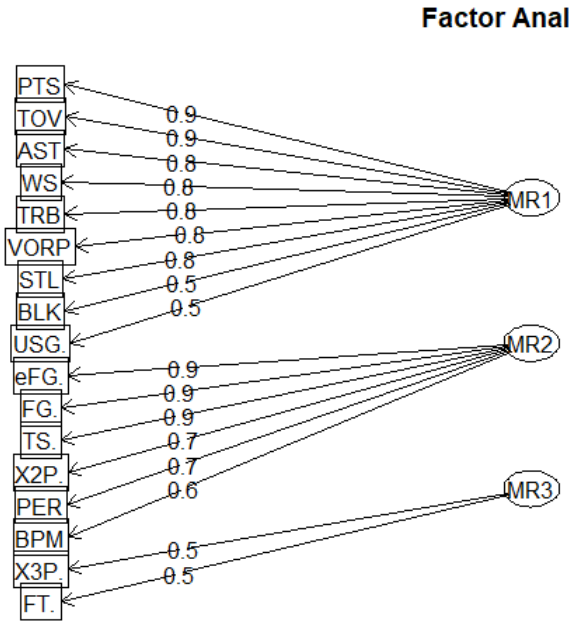
Número de Fatores	**BIC**	**RMSEA**	**TLI**	**Fit**	**Explicação da Variância**
2 fatores	6790,551	0,265	0,571	0,923	62%
3 fatores	5253,864	0,255	0,604	0,955	69%
4 fatores	3915,311	0,243	0,639	0,970	75%
5 fatores	2744,182	0,228	0,682	0,976	78%
6 fatores	1772,749	0,209	0,731	0,981	82%
7 fatores	753,543	0,165	0,832	0,993	88%

De acordo com a recomendação dos *plots* feitos e com as medidas dos modelos na tabela 2, o modelo escolhido foi o com 3 fatores, apesar de que o modelo fatorial com 4 fatores possui de fato melhor BIC, as medidas restantes RMSEA, TLI e Ajuste foram próximas e não houve grande adição na explicação da variância, além de que o 4º fator seria constituído somente da variável USG%, o que busca-se evitar durante a análise fatorial. Ademais, as medidas do modelo fatorial de 7 fatores apresentam bons valores, apesar do *Scree plot* não recomendar esse número de fatores, o modelo foi levado à análise

fatorial confirmatória para que fosse investigada a hipótese de *overfitting* ou não desse modelo fatorial.

Assim, o modelo selecionado apresenta a seguinte estrutura:

Figura 16: Diagrama Fatorial da Análise Fatorial Exploratória



Verifica-se que todas as cargas dos fatores são superiores a 0,5 (BLK e USG% no fator 1, X3P% e FT% no fator 3). Importante notar que a variável TOV, apesar de ser algo negativo para o atleta em um jogo, está com carga positiva e de valor 0,9, isto se dá pela sua correlação com as outras variáveis presentes, um jogador que conduz mais a bola faz mais pontos e arma mais jogadas para o time consequentemente possui a bola por mais tempo, e comete mais inversões de posse.

4.2.2 Análise Fatorial Confirmatória

Em seguida, foi realizada a análise fatorial confirmatória, utilizando os dados dos atletas dos anos de 2019 e 2020 e seguindo as correlações mostradas na figura 16.

Tabela 3: Medidas de Ajuste para o Modelo de Equações Estruturais

Fatores	BIC	RMSEA	TLI	CFI
3	31324,235	0,261	0,585	0,650
7	16581,981	0,260	0,576	0,643

Houve melhora nas medidas de ajuste do modelo após a utilização do modelo de equações estruturais. Tendo em vista as medidas de ajuste do modelo com 7 fatores, nota-se que realmente se tratava de um *overfitting* nos dados, já que, mesmo com mais parâmetros, os dois modelos obtiveram medidas muito próximas fora o BIC e a partir do princípio da parcimônia e para evitar *overfitting*, o modelo com 3 fatores continua sendo o modelo selecionado. Com o objetivo de melhor entender como as variáveis e os fatores se relacionam, seguem as cargas fatoriais obtidas:

Figura 17: Diagrama Fatorial da Análise Fatorial Confirmatória

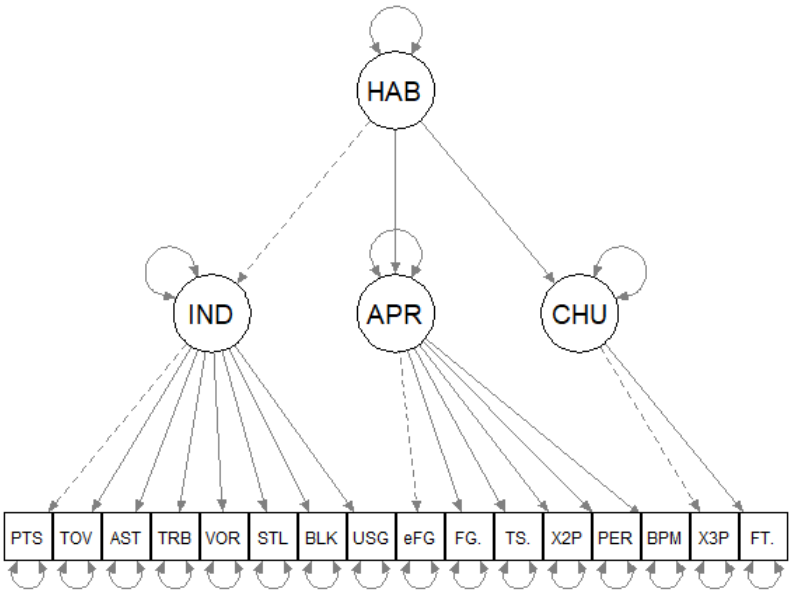


Tabela 4: Cargas Fatoriais do Modelo

Variável	**Fator 1**	**Fator 2**	**Fator 3**
PTS	0,949	0,000	0,000
TOV	0,970	0,000	0,000
AST	0,880	0,000	0,000
TRB	0,777	0,000	0,000
VORP	0,775	0,000	0,000
STL	0,822	0,000	0,000
BLK	0,521	0,000	0,000
USG%	0,552	0,000	0,000
eFG%	0,000	0,973	0,000
FG%	0,000	0,898	0,000
TS%	0,000	0,972	0,000
X2P%	0,000	0,706	0,000
PER	0,000	0,714	0,000
BPM	0,000	0,711	0,000
X3P%	0,000	0,000	0,965
FT%	0,000	0,000	0,235

A posteriori da análise das variáveis e dos fatores, seguiu-se para a nomeação dos próprios. O Fator 1 é constituído majoritariamente de variáveis que são as mais valiosas para um jogador, individualmente, dentro de um jogo, sendo então chamado de Estatísticas Individuais. No Fator 2, são realçadas variáveis que medem a eficiência de um jogador e possibilitam comparar seus acertos com o total de tentativas, sendo chamado de Estatísticas de Aproveitamento. Por último, o Fator 3 é composto por variáveis que indicam principalmente a qualidade do arremesso do jogador, sendo então chamado de Estatísticas de Chute.

Tabela 5: Pesos dos Fatores na Habilidade Final

Fator	**Nome**	**Peso**
Fator 1	Estatísticas Individuais	1,000
Fator 2	Estatísticas de Aproveitamento	1,922
Fator 3	Estatísticas de Chute	1,310

Nota-se que as Estatísticas de Aproveitamento obtiveram maior peso para o cálculo do escore de Habilidade Final, sendo essas que medem a eficiência do jogador como também possibilitam comparações com outros jogadores, todas as cargas se apre-

sentam positivas.

Dessa forma, o cálculo do escore de Habilidade pode ser feito por:

$$\text{Habilidade} = 1,000 * \text{Fator } 1 + 1,922 * \text{Fator } 21,310 * \text{Fator } 3 \tag{4.2.1}$$

Finalmente, realizou-se a análise de correlação entre a habilidade calculada e a variável WS (vitórias adicionadas ao time quando esse jogador entra no time) buscando verificar se o escore calculado é um bom indicador, de fato, da habilidade do atleta.

Tabela 6: Teste de Correlação de Spearman entre WS e Habilidade

Coeficiente de Correlação	**Estatística do Teste**	**P-Valor**	**Decisão**
0,874	33376694	$< 0,001$	Rejeita H_0

Percebe-se que o escore de Habilidade e a variável WS possuem de fato correlação forte e positiva.

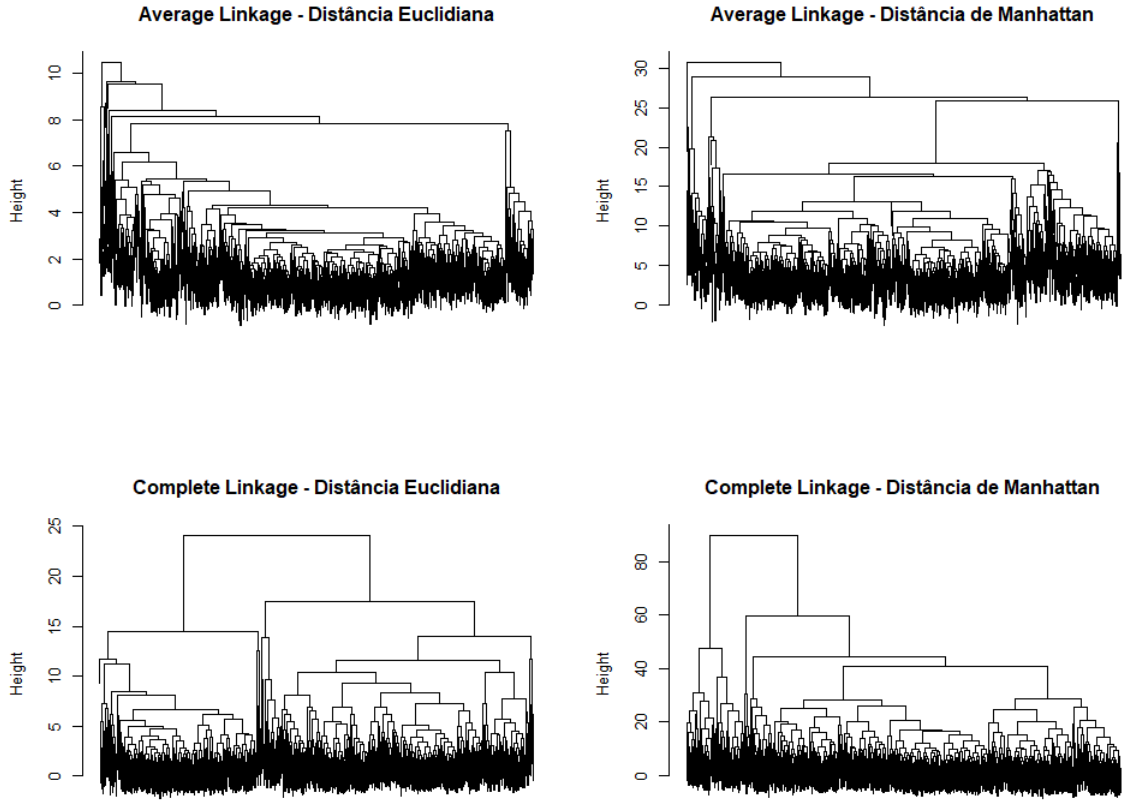
4.3 Análise de Agrupamento

Objetivando encontrar grupos de jogadores, suas características e como estão divididos, foram utilizadas as 16 variáveis selecionadas do banco de dados na análise de agrupamentos pois elas fornecem o melhor agrupamento, também foi com ela que foram construídos os modelos fatoriais. Depois da aplicação de métodos hierárquicos para determinar o melhor número de *clusters*, o método de agrupamento final utilizado foi o método não-hierárquico *K-means*.

4.3.1 Métodos Hierárquicos

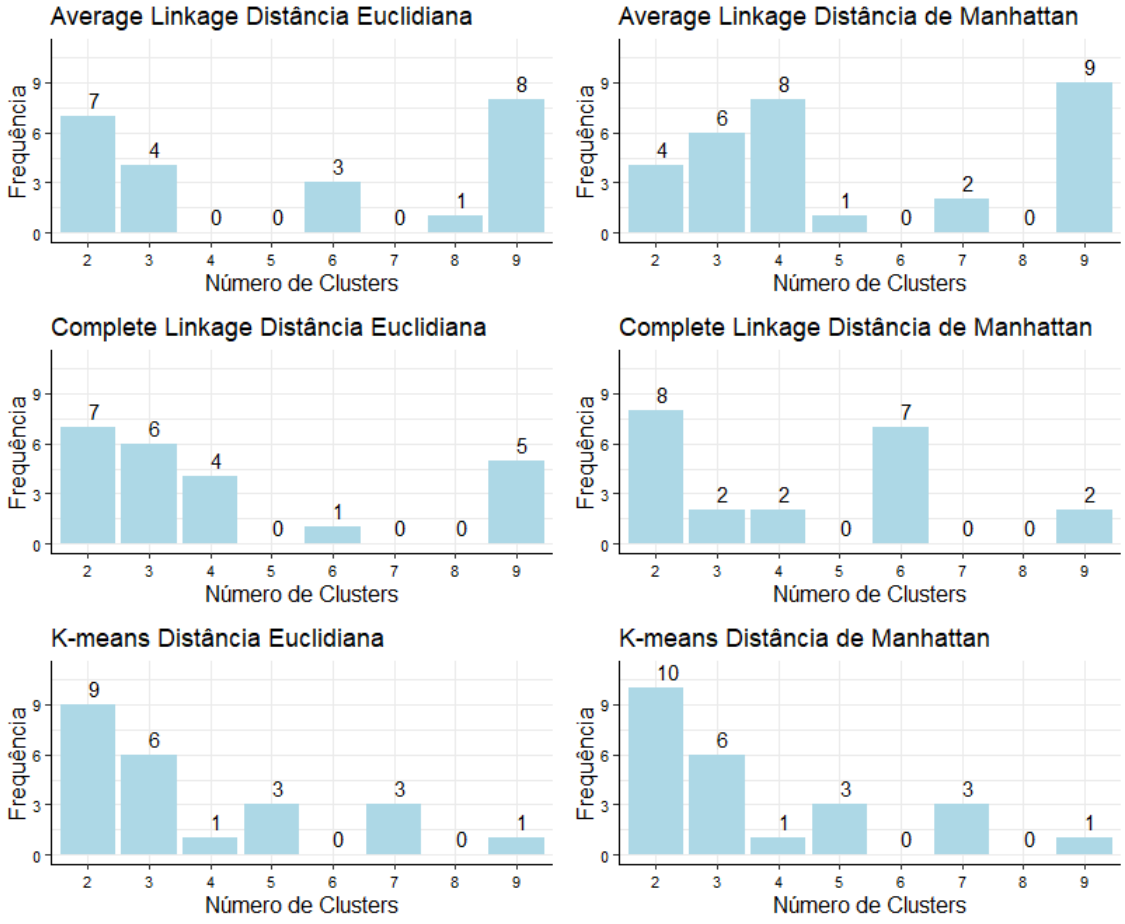
Ao começo da análise, as 16 variáveis selecionadas foram padronizadas e então as distâncias Euclidianas e de Manhattan das observações foram calculadas usando os métodos *Average Linkage* e *Complete Linkage*, totalizando quatro gráficos. Os dendrogramas seguem abaixo:

Figura 18: Dendrogramas das Variáveis Selecionadas



Analisa-se a partir da figura 18 a formação de 2 clusters, principalmente em ambas as distâncias na metodologia *Complete Linkage*. Contudo, objetivando o melhor agrupamento possível, também foram calculados os índices propostos pelo pacote NbClust, cujos resultados seguem no gráfico abaixo.

Figura 19: Gráfico de Barras do Número de Clusters Ideal

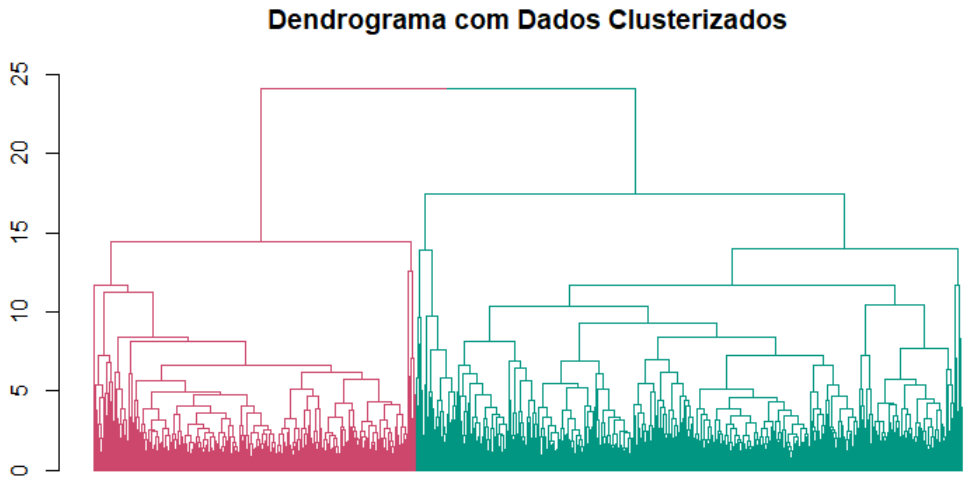


Vê-se a recomendação do uso de 2 *clusters* em 4 dos 6 gráficos de barras demonstrados na figura 19, sendo das metodologias *Complete Linkage* e *K-means*. O número de agrupamentos a ser adotado para os dados será de 2 grupos nos agrupamentos não-hierárquicos.

4.3.2 Métodos Não Hierárquicos

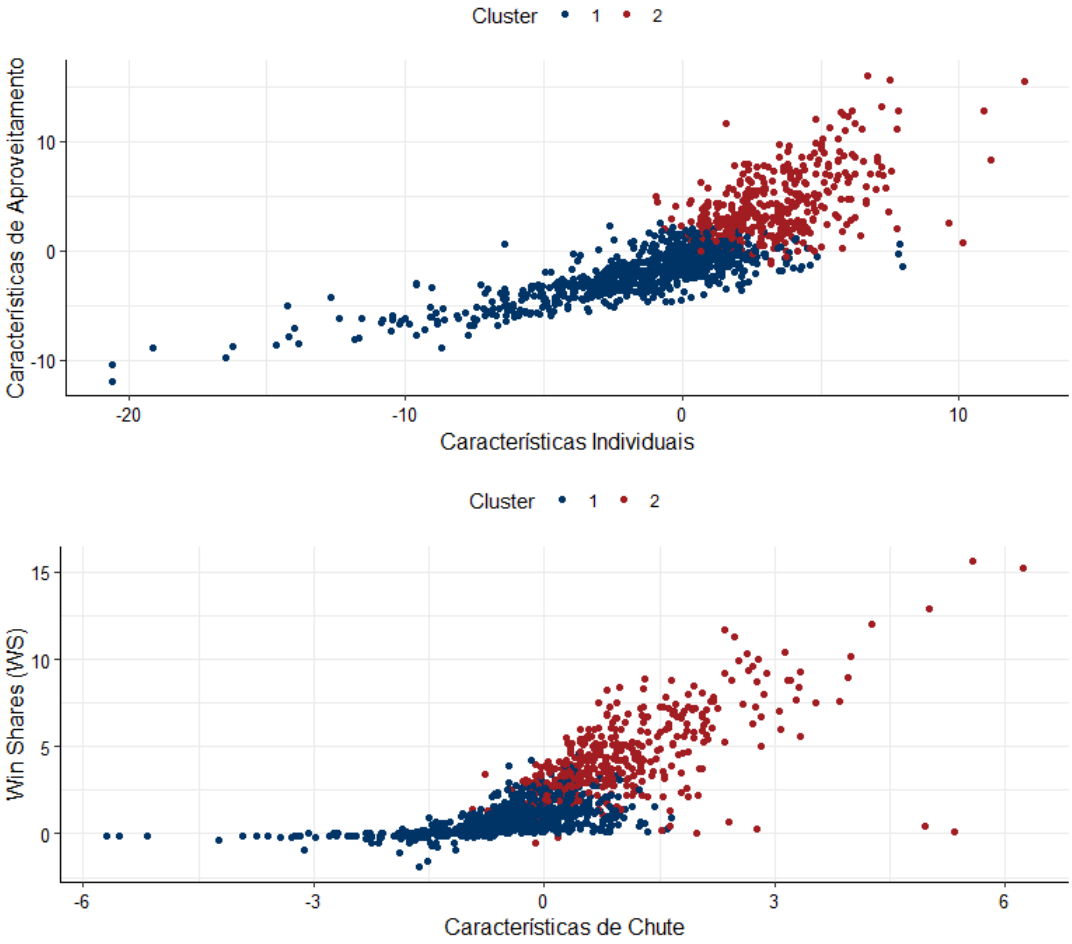
Fixado o número de *clusters* a serem agrupados, construiu-se o dendrograma final abaixo:

Figura 20: Dendrograma Clusterização K-means



Nota-se que o *cluster* 1 (cor verde) é composto por mais observações, 774 atletas, que o *cluster* 2 (cor vermelha) é formado por 395 jogadores, pouco mais da metade.

Figura 21: *Biplot* dos Clusters



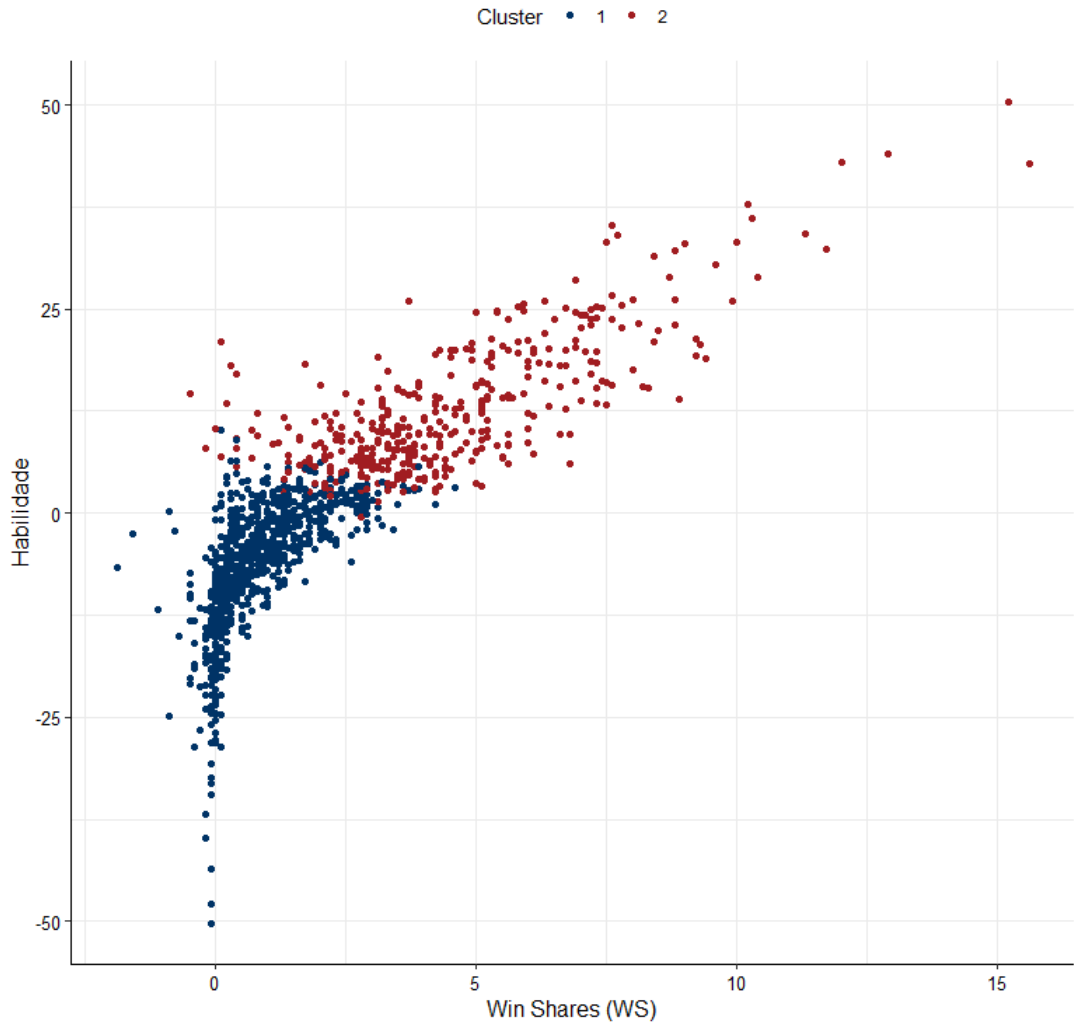
Analisando a figura 21 e a relação entre os fatores e a variável WS que ela apresenta, verifica-se que os jogadores do *cluster* 2 possuem, majoritariamente, valores superiores nos escores de Características individuais (fator 1) e Características de aproveitamento (fator 2) e é um grupo com maior variabilidade. Já o *cluster* 1 é composto de jogadores que não possuem valores tão altos nos escores das Características individuais e de aproveitamento, além de ser um grupo com variabilidade menor ao se concentrarem muito mais ao redor de sua média.

Os jogadores do *cluster* 2 também possuem valores de escore de Características de chute (fator 3) positivos em sua maioria, além de estarem mais concentrados em valores de WS bem acima de zero. Enquanto isso, os atletas do *cluster* 1 têm valores de Característica de chute negativos em sua maioria e *win shares* próximas de zero.

4.4 Análises Finais

Agrupados os jogadores em *clusters*, foram realizadas análises para compreender esses grupos dentro das variáveis utilizadas, dos próprios fatores e também do escore de Habilidade construído.

Figura 22: Gráfico de Dispersão da Habilidade por *Win Share*



Com base na figura 22, é notória a concentração do *cluster* 1 em valores do escore de habilidade abaixo de 0, além de também, em sua totalidade, estar abaixo de 5 *win shares* e ser um grupo com variabilidade muito menor, ou seja, mais compacto. Na perspectiva do *cluster* 2, seus atletas estão majoritariamente acima de 0 de escore de habilidade e se concentram acima de 2,5 *win shares*, porém ao contrário do *cluster* 1, possuem maior variabilidade e estão mais espaçados entre si.

Figura 23: Boxplots dos Clusters pelos Fatores e *Win Share*

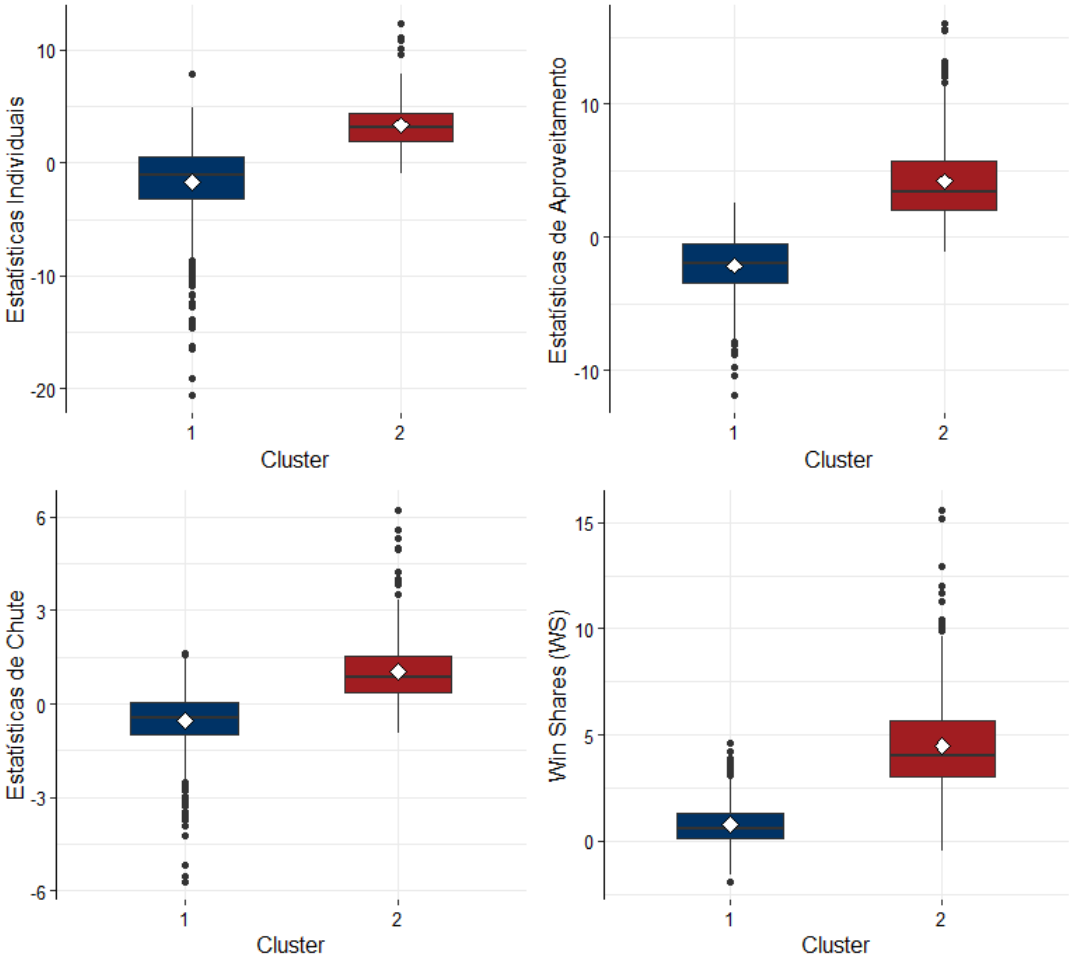


Tabela 7: Média dos Fatores por Cluster

Variável	**Cluster 1**	**Cluster 2**
Estatísticas Individuais	-1,71	3,34
Estatísticas de Aproveitamento	-2,13	4,18
Estatísticas de Chute	-0,53	1,05
Win Share	0,82	4,49

Tabela 8: Teste de Kruskal-Wallis para Fatores entre *Clusters*

Comparação	**Fator 1**	**Fator 2**	**Fator 3**	**WS**	**Decisão**
1 e 2	$< 0,001$	$< 0,001$	$< 0,001$	$< 0,001$	Rejeita H_0

Com a leitura da figura 23 e das tabelas 7 e 8, analisa-se que em todas as estatísticas (fatores) e na variável WS, o *cluster* 2 possui médias superiores a do *cluster* 1, comprovado pelos p-valores dos testes de Kruskal-Wallis realizados. Motrando que atletas

do segundo grupo possuem estatísticas médias melhores que a dos jogadores do primeiro grupo.

Figura 24: Boxplot da Habilidade por Cluster

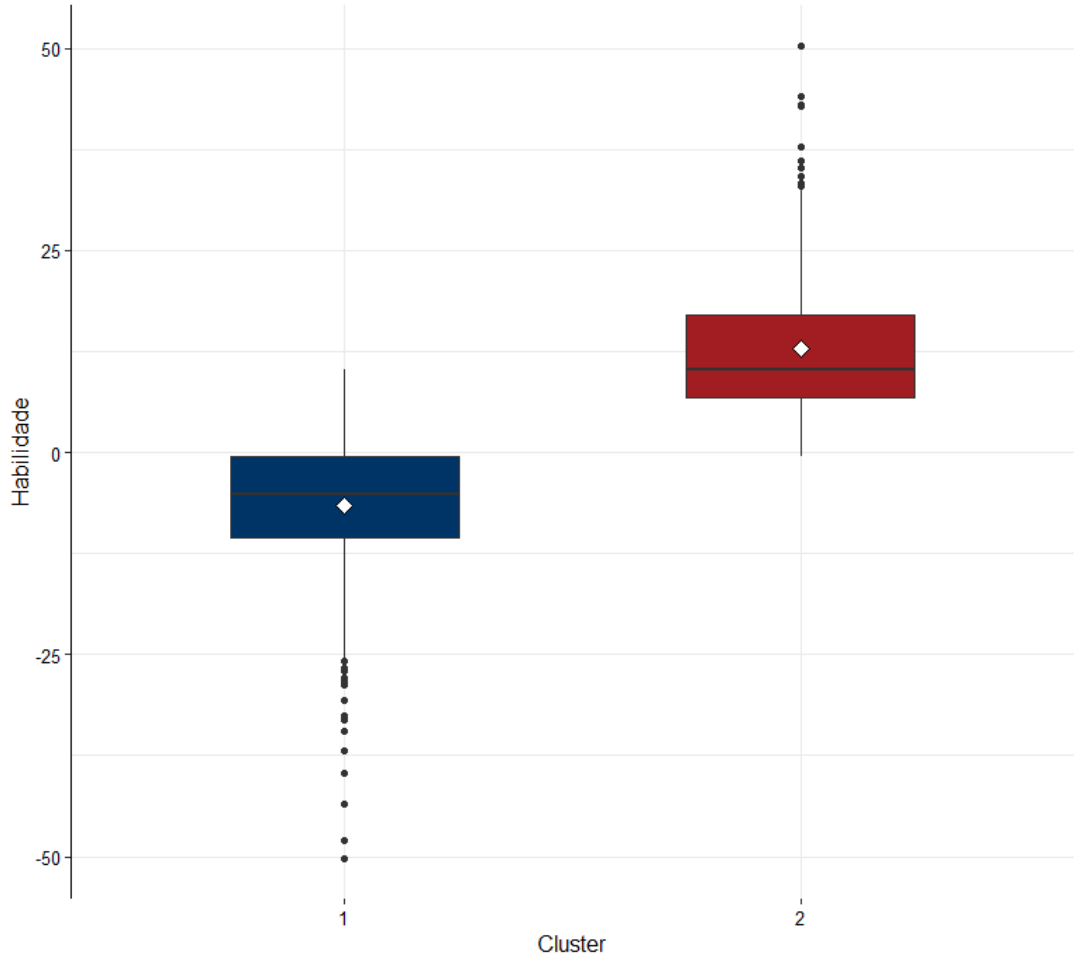


Tabela 9: Média da Habilidade por Cluster

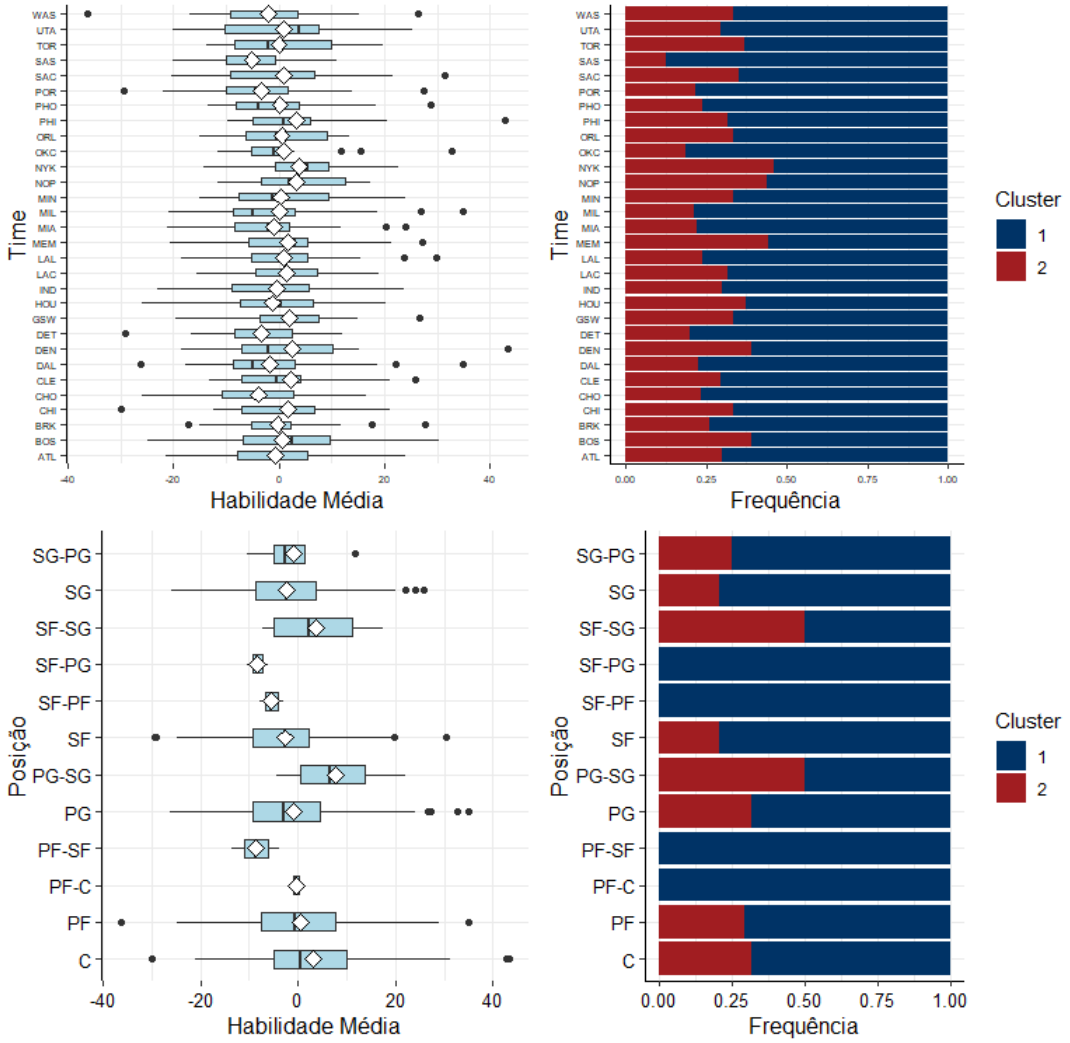
Variável	**Cluster 1**	**Cluster 2**
Habilidade	-6,51	12,76

Tabela 10: Teste de Kruskal-Wallis para Escore de Habilidade entre *Clusters*

Comparação	**P-Valor**	**Decisão**
1 e 2	$< 0,001$	Rejeita H_0

Nota-se a partir da figura 24 e tabelas 9 e 10 que também no escore de habilidade o *cluster* 2 possui média superior ao *cluster* 1, comprovada a diferença pelo teste de Kruskall-Wallis entre os grupos.

Figura 25: Gráficos de Habilidade e Cluster por Time e Posição da temporada 2022-2023



Ao ver a distribuição da habilidade e dos *clusters* dentro dos times dos jogadores, nota-se o *Santo Antonio Spurs* (SAS) como sendo o time com menor habilidade média, assim como maior número número de jogadores do *cluster* 1, grupo com menor média de habilidade. Outros times que apresentam número baixo de atletas do *cluster* 2 foram *Oklahoma City Thunder* (OKC), *Detroit Pistons* (DET) e *Charlotte Hornets*, times que não se classificaram para o mata-mata da temporada e que também apresentaram baixas médias de habilidade.

Sob a lente da variável posição, 4 delas que não apresentam jogadores do *cluster* 2: SF-PG, SF-PF, PF-SF e PF-C, o que reflete nas habilidades médias dessas posições, que são as mais baixas dentre todas. A posição com maior número de atletas do segundo grupo, SF-SG, também é a posição com maior habilidade média. Outras posições que se destacaram tanto na distribuição da habilidade quanto dos *clusters* foram as posições PG-SG e PG. Com isso, nota-se que as posições que têm mais contato com a bola, como

PG (armador) e derivadas dela, possuem mais habilidade, no geral.

Tabela 11: Últimos 10 Jogadores eleitos Mais Valiosos e 10 Jogadores mais Habilidosos por ano de acordo com o Modelo treinado com os dados de cada ano

Jogador Mais Valioso	**Habilidade**	**Jogador Mais Habilidoso**	**Habilidade**	**Temporada**
Joel Embiid (1º)	36,39	Joel Embiid	36,39	2022-2023
Nikola Jokić (1º)	45,70	Nikola Jokić	45,70	2021-2022
Nikola Jokić (1º)	43,87	Nikola Jokić	43,87	2020-2021
Giannis Antetokounmpo (2º)	38,63	James Harden	41,11	2019-2020
Giannis Antetokounmpo (2º)	43,84	James Harden	49,55	2018-2019
James Harden (3º)	36,22	LeBron James	41,35	2017-2018
Russel Westbrook (1º)	44,67	Russell Westbrook	44,67	2016-2017
Stephen Curry (1º)	48,36	Stephen Curry	48,36	2015-2016
Stephen Curry (2º)	42,55	James Harden	46,65	2014-2015
Kevin Durant (1º)	50,17	Kevin Durant	50,17	2013-2014

A tabela 11 mostra os jogadores mais valiosos, ou MVPs, dos últimos 10 anos, além dos jogadores com maior habilidade de cada ano. Com sua leitura, nota-se que o atleta com maior escore de habilidade final calculado foi o mesmo atleta que foi MVP naquela temporada em 6 dos 10 anos analisados, enquanto que nas temporadas onde o resultado não foi compatível, o jogador mais valioso seria o 2º ou 3º jogador mais habilidoso pelo modelo.

5 Conclusão

Seguidamente da análise descritiva e de correlações entre as variáveis, foram retiradas as mesmas formadas por combinações lineares ou que compunham outras variáveis. Ao total, 17 variáveis permaneceram no estudo: FG%, X3P%, X2P%, eFG%, FT%, TRB, AST, STL, BLK, TOV, PTS, PER, TS%, USG%, BPM, WS e VORP, onde a variável WS não foi levada à análise fatorial para futura validação do próprio modelo. A partir dessas 16 variáveis selecionadas, formou-se um modelo fatorial com 3 fatores através do método dos mínimos resíduos e com rotação de fatores *Varimax*, os valores das respectivas cargas fatoriais estão descritos na tabela 4. Com isso, construiu-se um escore de habilidade do atleta, mostrado pela equação 4.2.1, que utiliza dos 3 fatores e atribui pesos a cada um deles.

Para o Fator 1, nomeado Estatísticas Individuais, as variáveis com maior carga nesse fator são PTS e TOV, e de menor peso as variáveis BLK e USG%. No Fator 2, chamado Estatísticas de Aproveitamento, as variáveis mais significativas são eFG% e TS%, enquanto as de menor relevância são X2P% e BPM. No terceiro Fator, batizado como Estatísticas de Chute, é composto somente por duas variáveis, onde X3P% possui o maior peso e FT% o menor. Na construção da Habilidade Final, todos os fatores possuem pesos positivos, dos quais as Estatísticas de Aproveitamento possuem maior peso, já as Estatísticas Individuais contribuem com o menor peso na composição do escore de habilidade.

Logo após a análise fatorial, o mesmo grupo de 16 variáveis selecionadas e padronizadas foi utilizado para análise de agrupamento, inicialmente para verificar o número ideal de *clusters* por meio de métodos hierárquicos *Average* e *Complete linkage* com as distâncias Euclidiana e de Manhattan. O número ideal de grupos a dividirem os dados foi de 2 *clusters*, o método utilizado para agrupar os dados foi o método não hierárquico *K-Means* que dividiu os dados no *Cluster* 1, composto por mais jogadores e menos habilidosos, e no *Cluster* 2, constituído por menos e mais habilidosos atletas dadas as variáveis.

Por último, realizaram-se análises descritivas com os fatores e o escore de habilidade construídos, como também com a variável WS reservada anteriormente, que apresentou forte correlação positiva e, apesar de não participar do agrupamento *K-Means*, também conseguiu separar muito bem os *clusters* de jogadores junto à Habilidade Final. Com base as análises, gráficos e tabelas realizados, concluiu-se que os atletas que jogam mais próximos da bola, em posições como PG e PG-SG são, em média, mais habilidosos que os demais, apesar de que nas últimas 5 temporadas, os jogadores mais valiosos

(MVPs) foram pivôs (C), e o atleta de basquete com maior pontuação de habilidade nos últimos 3 anos coincidiu com o jogador mais valioso da temporada, acertando em um total 6 dos últimos 10 MVPs das temporadas.

Referências

BARE, C. *Drawing heatmaps in R*. [S.l.], 2011. Disponível em: ⟨https://www.r-bloggers.com/2011/06/drawing-heatmaps-in-r/⟩. Acesso em: 24 set. 2021.

BIELBY, W. T.; HAUSER, R. M. Structural equation models. *Annual Review of Sociology*, v. 3, 1977.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. 5. ed. [S.l.]: Saraiva, 2003.

CHARRAD, M. et al. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, n. 6, 2014.

CHEVLIN, M.; MILES, J. Effects of sample size, model specification and factor loadings on the gfi in confirmatory factor analysis. *Personality and Individual Differences*, 1998.

GRIS, K. et al. Exhaustive behavioral profile assay to detect genotype differences between wild-type, inflammasome-deficient, and nlrp12 knock-out mice. *AIMS Medical Science*, v. 5, p. 238–251, 05 2018.

HAIR, J. F. et al. *Análise Multivariada de Dados*. [S.l.]: Bookman, 2009.

IACOBUCCI, D. Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, v. 20, 2009.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. [S.l.]: Pearson, 2007.

SHINKAWA AMANDA E MONTEIRO, E. Estatística no baseball: Uma análise de desempenho dos arremessadores da liga principal de baseball. 2022.

TUCKER, L.; LEWIS, C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, v. 38, 1973.

VRIEZE, S. I. Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Prychol Methods*, v. 17 (2), 2012.