Universidade de Brasília
Departamento de Estatística

Estatística no Baseball
Uma análise de desempenho dos arremessadores da Liga Principal de Baseball

Amanda Shinkawa Sibin

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

Amanda Shinkawa Sibin

Estatística no Baseball
Uma análise de desempenho dos arremessadores da Liga Principal de Baseball

Orientador(a): Prof. Eduardo Monteiro de Castro Gomes

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
2021

Resumo

O presente trabalho objetiva a análise de desempenho de arremessadores de *baseball* da Liga Principal de Baseball (em inglês, *Major League Baseball*, abreviada como MLB), por meio da criação de um escore de habilidade com base em suas estatísticas disponíveis no site *Baseball Reference*.

Para isso, foi feita uma análise fatorial exploratória, a fim de compreender quais as variáveis mais importantes para descrever a habilidade do arremessador, seguida de um modelo de equações estruturais para criar a pontuação de desempenho final. Posteriormente, 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 habilidade.

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

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Na conjuntura esportiva hodierna, a estatística vêm tomando cada vez mais seu espaço e assumindo sua devida importância no auxílio do processo de tomada de decisões por parte de técnicos e treinadores. No contexto do *baseball*, as estatísticas começaram a serem registradas em 1837, com a anotação do número de corridas anotas por jogador (SCHWARZ, 2004). Após um longo desenvolvimento de técnicas e softwares para aumentar a precisão e complexidade dos dados, várias estatísticas surgiram e foram, também, renunciadas (LAW, 2017). Entretanto, até os dias de hoje, estatísticas básicas como a média de rebatidas e outras que não medem unicamente o esforço do jogador (e sim do time dele como um todo) são utilizadas como critério principal e até mesmo único para determinar a habilidade do jogador. Um exemplo deste problema é o prêmio "Título de Rebatidas", que é destinado ao jogador que possui a maior média de rebatidas, mesmo que essa estatística não leve em consideração diversas jogadas e habilidades do jogador (CASTROVINCE, 2020).

Tendo isso em vista, é essencial que novas formas de se medir desempenho no *baseball* sejam determinadas, principalmente de arremessadores. Ademais, é de extrema importância entender qual o valor que esse jogador agrega ao seu time e quais as habilidades de um arremessador são mais importantes para se diferenciar os melhores jogadores dos demais.

Nesse sentido, este trabalho visa, após realizar a redução de variáveis e a criação de divisões dentro dos arremessadores da MLB do ano de 2020. Assim, espera-se entender como melhor classificá-los com base em sua habilidade. Propõe-se a realização de redução de variáveis por meio de análise fatorial, elaboração de modelos de equações estruturais para entender os traços latentes por trás dos fatores, separação dos times em grupos ou *clusters* com base em proximidade de habilidade e, por fim, obter-se uma nova forma de determinar valor de jogadores, mais estruturada e baseada no máximo de informações possível.

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 Testes de Hipóteses

Testes de Hipóteses são técnicas estatísticas que objetivam verificar a veracidade de determinadas afirmações acerca de uma população feitas com base em uma amostra. Assim, um teste busca fornecer uma metodologia que permita verificar se os dados disponíveis trazem evidências estatísticas suficientes que apoiem ou não uma hipótese formulada (BUSSAB; MORETTIN, 2003).

As hipóteses estatísticas são da forma:

$$\begin{cases} H_0 : \text{Hipótese nula a ser testada} \\ H_1 : \text{Hipótese alternativa, diferente da hipótese nula} \end{cases}$$

Assim, define-se um nível de significância α a ser adotado (neste estudo, 5%), definida como a probabilidade de se cometer um erro do tipo I, ou seja, a probabilidade de se rejeitar a hipótese nula dado que esta é verdadeira (BUSSAB; MORETTIN, 2003).

Para a realização do teste, deve-se calcular uma estatística do teste, que varia de acordo com o teste aplicado e a distribuição da variável ou variáveis a serem testadas.

Em seguida, calcula-se o p-valor (ou probabilidade de significância), calculado como a probabilidade de se observarem valores da estatística do teste iguais ou mais extremos que o observado, sob hipótese nula verdadeira. Caso o nível de significância α seja menor que o p-valor, rejeita-se a hipótese nula. Caso contrário, diz-se que não existem evidências estatísticas suficientes para a rejeição da hipótese nula (BUSSAB; MORETTIN, 2003).

2.1.1 Teste de Correlação de Postos de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que verifica, por meio de postos de variáveis quantitativas ou qualitativas ordinais, o grau de

correlação linear entre duas variáveis (FIELLER; HARTLEY; PEARSON, 1957). Esse coeficiente varia entre os valores -1 e 1 e pode ser calculado pela fórmula:

$$r_{Spearman} = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{2.1.1}$$

Onde:

- d_i é a diferença entre os postos de cada observação de ambas as variáveis;
- n é o número total de observações na amostra.

Por ser um método não paramétrico, não há suposições para o teste, com as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não há correlação de postos entre as variáveis } X \text{ e } Y \ (\rho_{Spearman} = 0) \\ H_1 : \text{Há correlação de postos entre as variáveis } X \text{ e } Y \ (\rho_{Spearman} \neq 0) \end{cases}$$

Ademais, tem-se que $\rho_{Spearman}$ é o coeficiente de correlação de postos populacional (parâmetro a ser testado com base em $r_{Spearman}$) (FIELLER; HARTLEY; PEARSON, 1957).

Para grandes amostras, a estatística do teste $r_{Spearman}$, sob H_0 verdadeira, tem distribuição aproximada pela Normal Padrão (FIELLER; HARTLEY; PEARSON, 1957), tal que:

$$w_p = \frac{z_p}{\sqrt{(n-1)}} \tag{2.1.2}$$

Onde:

- w_p é o quantil de ordem p da distribuição que a estatística do teste segue;
- z_p é o quantil de ordem p da distribuição Normal padrão;
- n é o número total de observações na amostra.

2.1.2 Teste de Correlação Linear de Pearson

O coeficiente de correlação linear de Pearson indica a força e a direção do relacionamento linear entre duas variáveis quantitativas. Possui valores entre -1 e 1, onde o valor -1 representa total correlação linear negativa entre as variáveis e o valor 1 representa total correlação linear positiva entre elas (FIELLER; HARTLEY; PEARSON, 1957). Esse coeficiente pode ser calculado como:

$$r_{Pearson} = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x}) (y_i - \bar{y}) \right]}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \tag{2.1.3}$$

Onde:

- x_i é o i-ésimo valor da variável X;
- y_i é o i-ésimo valor da variável Y;
- $\bar{x}$ é a média dos valores da variável X;
- $\bar{y}$ é a média dos valores da variável Y.

Para o teste de correlação de Pearson, tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \text{Não há correlação linear entre as variáveis } X \text{ e } Y \ (\rho_{Pearson} = 0) \\ H_1 : \text{Há correlação linear entre as variáveis } X \text{ e } Y \ (\rho_{Pearson} \neq 0) \end{cases}$$

Ademais, $\rho_{Pearson}$ é o parâmetro a ser testado (coeficiente de correlação linear populacional). (FIELLER; HARTLEY; PEARSON, 1957). A estatística do teste é dada por:

$$t_{Pearson} = \frac{r_{Pearson}\sqrt{n-2}}{\sqrt{1 - r_{Pearson}^2}} \tag{2.1.4}$$

Assim, sob H_0, $r_{Pearson}$ segue uma distribuição t-Student com $(n-2)$ graus de liberdade (FIELLER; HARTLEY; PEARSON, 1957).

2.1.3 Teste de Dunn

O teste de Dunn tem como objetivo comparar as médias dos grupos 2 a 2, controlando o erro global dos testes. É um teste não paramétrico, ou seja, não exige que as variáveis sigam normalidade (DUNN, 1961). Suas hipóteses são:

$$\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

Seja W_i a soma dos *ranks* do i-ésimo grupo e n_i o número de observações do i-ésimo grupo, e seja $\hat{W}_i$ a média dos *ranks* do i-ésimo grupo, então a estatística do teste para os grupos A e B (DUNN, 1961) é:

$$z_{A,B} = \frac{\hat{W}_A - \hat{W}_B}{\sigma_{A,B}} \tag{2.1.5}$$

Onde:

- $$\sigma_{A,B} = \sqrt{\left[\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r} \tau_s^3 - \tau_s}{12(N-1)}\right]\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{2.1.6}$$
- N é o tamanho total da amostra de todos os grupos;
- r é o número de empates nos *ranks* entre todos os grupos;
- τ_s é o número de observações entre todos os grupos com o s-ésimo *rank* empatado.

2.2 Análise Fatorial

Objetiva a criação de fatores ou construtos, a fim de reduzir a dimensionalidade dos dados e gerar a melhor aproximação possível da matriz de covariâncias (JOHNSON; WICHERN, 2007).

O modelo geral da análise fatorial para o fator i é:

$$X_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + ... + l_{im}F_m + \epsilon_i \tag{2.2.1}$$

Onde:

- X_i é a variável i;
- μ_i é a média populacional da variável i;
- l_{ij}, com $j = 1, 2, ..., m$, são as cargas (*loadings*) da i-ésima variável no j-ésimo fator;
- F_j, com $j = 1, 2, ..., m$, são os j-ésimos fatores;
- ϵ_i é o i-ésimo fator específico associado somente à i-ésima resposta X_i.

Ou, ainda, genericamente:

$$X - \mu = LF + \epsilon \tag{2.2.2}$$

É interessante que cada carga apareça uma vez apenas em cada fator e com valores acima de 0,5 para serem consideradas boas cargas (JOHNSON; WICHERN, 2007).

Em alguns casos, os fatores obtidos são difíceis de serem interpretados e, assim, a solução inicial deve ser rotacionada. Existem dois tipos de rotação: rotação ortogonal (*varimax*), que mantém os fatores não correlacionados e a rotação oblíqua, que torna os fatores correlacionados entre si. O objetivo maior de rotacionar é identificar fatores que possuem variáveis que tenham correlação alta e outros com correlação baixa, ajudando assim na interpretação (JOHNSON; WICHERN, 2007).

Suposições do Modelo

- $$E(F) = 0 \tag{2.2.3}$$
- $$cov(F) = cov(FF') = I \tag{2.2.4}$$
- $$E(\epsilon) = 0 \tag{2.2.5}$$
- $$cov(\epsilon) = cov(\epsilon\epsilon') = \Psi \tag{2.2.6}$$

- $$cov(\epsilon, F) = E(\epsilon F') = 0 \tag{2.2.7}$$

Ademais, sabe-se que a matriz de covariâncias é a soma da matriz de cargas ao quadrado com a matriz diagonal esperança da covariância de ϵ:

$$\Sigma = LL' + \Psi \tag{2.2.8}$$

2.2.1 Rotação de Fatores Varimax

A rotação de fatores corresponde a uma rotação dos eixos de coordenadas, cujo objetivo é verificar algo que não estava muito claro nas cargas originais. O critério de rotação *Varimax* é um uma transformação ortogonal das cargas (JOHNSON; WICHERN, 2007). A função de rotação das novas cargas fatoriais é dada por:

$$\hat{L^*} = \hat{L}T \tag{2.2.9}$$

Com T sendo a transformação a ser determinada. Define-se

$$\tilde{I^*_{ij}} = \hat{I^*_{ij}}/\hat{h_i} \tag{2.2.10}$$

os coeficientes de escala rotacionadas dada pela matriz quadrada das comunalidades. Assim, a transformação ortogonal T é selecionada tal que

$$V = \frac{1}{p}\sum_{j=1}^{m}[\sum_{i=1}^{p}\tilde{I^{*4}_{ij}} - (\sum_{i=1}^{p}\tilde{I^{*2}_{ij}})^2/p] \tag{2.2.11}$$

é o maior possível.

Após selecionada a transformação T, as cargas $\tilde{I^*_{ij}}$ são multiplicadas por $\hat{h_i}$ tal que as comunalidades originais sejam preservadas. Assim, ao maximizar V, os quadrados das cargas são espalhados em cada fator o máximo possível (JOHNSON; WICHERN, 2007).

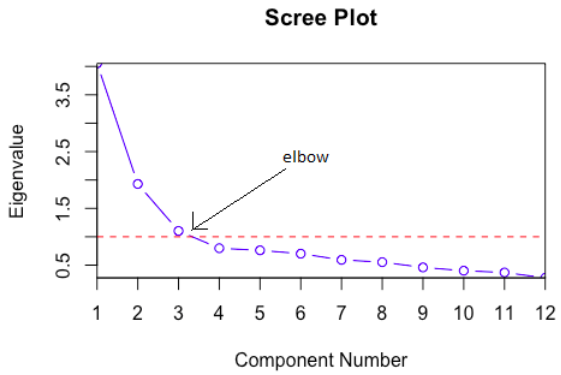
2.2.2 Análise Fatorial Exploratória

Na etapa exploratória da análise fatorial, algumas das formas de se determinar o número ideal de fatores são pelo *scree plot* ou pela análise paralela.

Scree Plot

Uma forma de se obter uma ajuda visual para determinar o número ideal de fatores é o *screeplot*, que coloca no eixo x o número de fatores e no eixo y seu respectivo autovalor (JOHNSON; WICHERN, 2007).

Figura 2.2.1: Exemplo de *ScreePlot*



Fonte: MANGALE (2020)

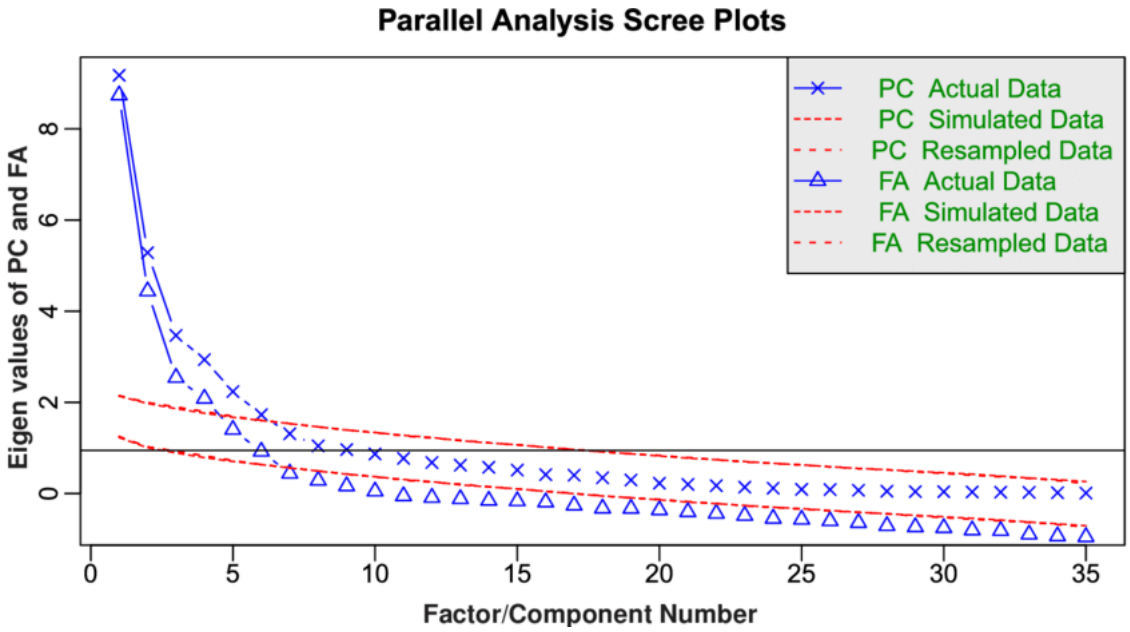
Assim, quando há a formação de um "joelho" no gráfico, existem evidências de que os autovalores depois daquele valor são relativamente pequenos e aproximadamente do mesmo tamanho, indicando que aquele número de fatores fornece uma boa explicação da variância total (JOHNSON; WICHERN, 2007).

Parallel Plot

O gráfico *parallel plot*, da análise paralela, retorna um *scree plot* com duas curvas, uma para valores observados nos dados e outra para valores dos dados simulados.

A análise paralela realiza uma simulação, ou seja, cria dados aleatórios com a mesma quantidade de observações e variáveis dos dados originais. Em seguida, cria uma matriz de correlação dos dados gerados e extrai seus autovalores. Quando os autovalores dos dados aleatórios são maiores que os autovalores dos dados originais, o ganho com aqueles fatores é quase insignificante (GRIS et al., 2018).

Figura 2.2.2: Exemplo de *Parallel Plot*

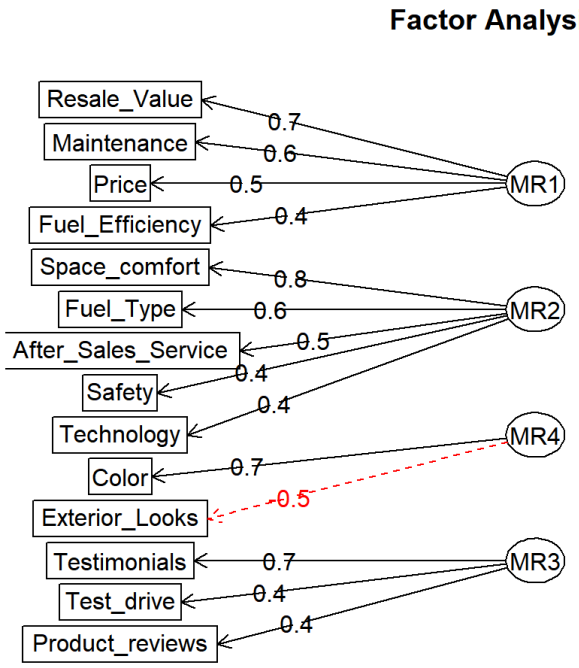


Fonte: GRIS et al. (2018)

Diagrama Fatorial

Para uma análise mais visual de itens e suas devidas correlações com os fatores, o diagrama fatorial pode ser utilizado:

Figura 2.2.3: Exemplo de Diagrama



Fonte: RAJPUT (2018)

2.2.3 Análise Fatorial Confirmatória

A análise fatorial confirmatória objetiva comparar os resultados obtidos na análise fatorial exploratória com a hipótese teórica conhecida, estimando apenas as relações entre fatores e variáveis especificados.

Essa etapa consiste, basicamente, em comparar o(s) modelo(s) obtidos na análise fatorial exploratória com o modelo teórico previamente conhecido, analisando suas medidas de ajuste relativo e absoluto.

Assim, alguns dos índices de ajuste utilizados para avaliar a qualidade do modelo são:

GFI

A estatística Índice de Qualidade de Ajuste ou *Goodness of Fit Index* (GFI) é a estatística Qui-Quadrado utilizada para avaliar a discrepância entre a matriz de covariância da amostra (S) e a matriz de covariância implícita no modelo ajustado (Σ) (CHEVLIN; MILES, 1998). Pode ser calculado pela fórmula:

$$GFI = \sum_i \frac{(S_i - \Sigma_i)^2}{\Sigma_i} \tag{2.2.12}$$

Um modelo bem ajustado deve apresentar, idealmente, valores de GFI maiores que 0,9 (CHEVLIN; MILES, 1998).

RMSEA

O erro quadrático médio de aproximação ou *Root Mean Square Error of Approximation* (RMSEA) é uma medida de erro do modelo que tende a ser sensível a amostras menores que 250 (IACOBUCCI, 2009). Pode ser calculado como:

$$RMSEA = \sqrt{\frac{(\chi^2 - df)}{df(N-1)}} \tag{2.2.13}$$

Onde:

- $$\chi^2 = N[tr(S\Sigma^{-1}) + log|\Sigma| - log|S| - (p+q)] \tag{2.2.14}$$
- df são os graus de liberdade;
- N é o tamanho da amostra;
- S é a matriz de covariâncias da amostra;
- Σ é a matriz de covariâncias da população;
- $p + q$ é a soma do número de variáveis exógenas e endógenas do modelo.

Um bom modelo deve apresentar valores pequenos de RMSEA, preferencialmente próximos ou inferiores a 0,09 (IACOBUCCI, 2009).

TLI

O Índice Tucker-Lewis (TLI) é um índice baseado no princípio de avaliação de ajuste incremental e pode ser calculado como uma adaptação da estatística GFI ou Qui-Quadrado (TUCKER; LEWIS, 1973), pela fórmula:

$$TLI = \frac{\chi^2/gl(ModeloVazio) - \chi^2/gl(ModeloProposto)}{\chi^2/gl(ModeloVazio) - 1} \tag{2.2.15}$$

Assim como o GFI, valores de TLI maiores ou próximos a 0,9 são considerados aceitáveis para um modelo bem ajustado (TUCKER; LEWIS, 1973).

BIC

Para realizar a comparação de diferentes modelos, o Critério de Informação Bayesiano (BIC) pode ser utilizado (VRIEZE, 2012). Pode ser calculado como:

$$BIC = -2ln(\hat{L}) + kln(N) \tag{2.2.16}$$

Onde:

- $\hat{L}$ são os parâmetros do modelo otimizado;
- k é o número de parâmetros;
- N é o número de observações que contribuem para a soma na equação de verossimilhança.

O modelo que obtiver menores valores para BIC é considerado o melhor modelo (VRIEZE, 2012).

CFI

O Índice de Ajuste Comparativo ou *Comparative Fit Index* (CFI) analisa o ajuste do modelo proposto, examinando a discrepância entre os dados e o modelo hipotético. A vantagem desse índice em relação aos demais é que este realiza o ajuste das questões de tamanho da amostra inerentes ao teste qui-quadrado de ajuste do modelo e o índice de ajuste normado (RIGDON, 1996).

$$CFI = 1 - (\lambda_k/\lambda_i) \tag{2.2.17}$$

Onde:

- $$\lambda_k = max(T_k - d_k, 0) \tag{2.2.18}$$
- $$\lambda_i = max(T_i - d_i, T_k - d_k, 0) \tag{2.2.19}$$

- T_i e d_i são a estatística do teste Qui-Quadrado e os graus de liberdade, respectivamente, do modelo base restrito;
- T_k e d_k são a estatística do teste Qui-Quadrado e os graus de liberdade, respectivamente, do modelo da teoria a ser estudada.

O CFI varia de 0 a 1, com valores superiores indicando um modelo melhor ajustado (RIGDON, 1996).

2.2.4 Modelos de Equações Estruturais

Um modelo de equações estruturais é um modelo que especifica o processo por trás da distribuição conjunta de variáveis observadas, por meio de uma expansão de um modelo de regressão linear. É utilizado para estudar as relações entre variáveis latentes e os construtos por trás delas (BIELBY; HAUSER, 1977). Pode ser expressado pelo sistema:

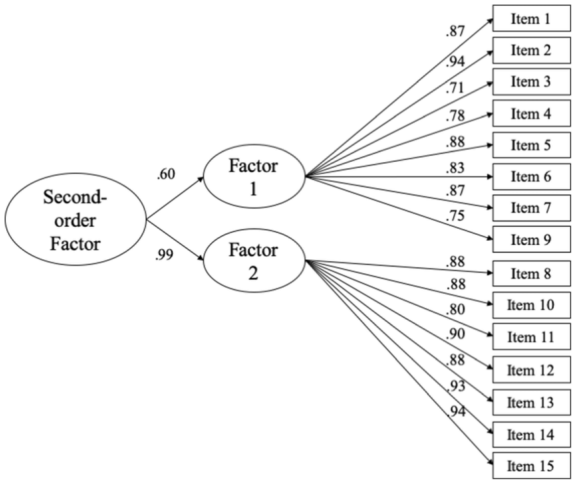
$$\begin{array}{ccccc} \Gamma y_i & = & Bx_i & + & u_i \\ (L \times L)(L \times 1) & & (L \times K)(K \times 1) & & (L \times 1) \end{array} \tag{2.2.20}$$

Onde:

- y_i é a i-ésima observação de L variáveis endógenas a serem calculadas;
- B e Γ são os coeficientes estruturais das variáveis latentes a serem estimados, nesse caso, pelo método da Máxima Verossimilhança;
- x_i é a i-ésima observação de K variáveis exógenas observadas;
- u_i é o distúrbio estrutural da i-ésima observação em cada uma das L equações estruturais.

Assim, constrói-se uma equação para cada fator durante a análise fatorial confirmatória de primeira ordem. Em seguida, realiza-se a análise de segunda ordem, que cria uma nova equação estrutural utilizando-se os fatores como variáveis exógenas (BIELBY; HAUSER, 1977). Um esquema ilustrativo pode ser visto na Figura 2.2.4:

Figura 2.2.4: Exemplo de Modelo Fatorial de Segunda Ordem



Fonte: MORAN; COROIU; KÖRNER (2021)

2.3 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).

2.3.1 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:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{2.3.1}$$

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:

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i| \tag{2.3.2}$$

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 inválidos, 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:

$$D^2 = (\bar{x} - \bar{y})' \sigma^{-1} (\bar{x} - \bar{y}) \tag{2.3.3}$$

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).

2.3.2 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 consistem em separar cada observação em um *cluster* diferente e agrupá-los por distância. Esse procedimento deve ser repetido até que todos os indivíduos encontrem-se em um único agrupamento final (HAIR et al., 2009). A matriz de distâncias é ser definida como:

$$D = \{d_{ik}\} \tag{2.3.4}$$

Onde d_{ik} é a distância entre os indivíduos 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, consiste em agrupar objetos pela maior distância d_{ik} entre eles. (JOHNSON; WICHERN, 2007). A fórmula da distância entre $d_{(UV)W}$ para esse método é dada por:

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\} \tag{2.3.5}$$

Average Linkage

O método *Average Linkage*, também conhecido como método da distância média, opera de forma semelhante ao método anterior. Entretanto, utiliza a média das distâncias como forma de agrupamento (JOHNSON; WICHERN, 2007). A fórmula da distância entre $d_{(UV)W}$ para esse método é dada por:

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W} \tag{2.3.6}$$

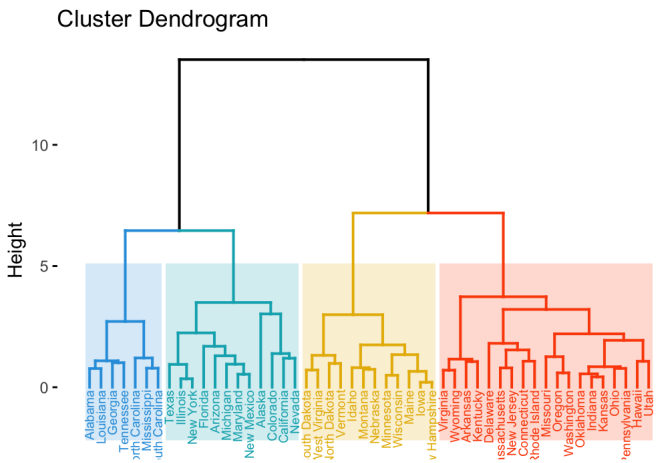
Onde:

- d_{ik} é a distância entre o objeto i no *cluster* (UV) e o objeto k no *cluster* W;
- $N_{(UV)}$ é o número de objetos no *cluster* (UV);
- N_W é o número de objetos no *cluster* W.

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).

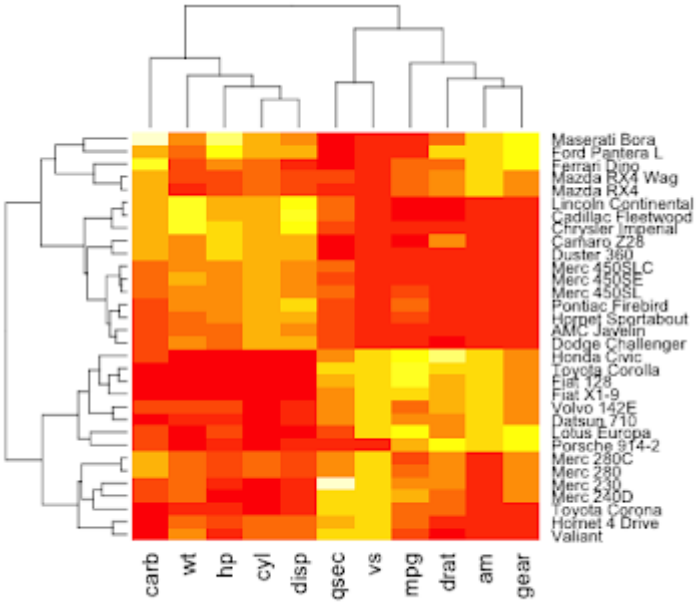
Figura 2.3.1: Exemplo de Dendrograma



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, onde as cores evidenciam os grupos mais similares entre si (BARE, 2011).

Figura 2.3.2: Exemplo de Mapa de Calor



Fonte: BARE (2011)

2.3.3 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 para grande volume de dados. O método mais conhecido para essa separação é o método *k-means* (JOHNSON; WICHERN, 2007).

Método *K-Means*

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

1. Separar os itens em k *clusters* iniciais;

2. Prosseguir com a lista de itens, um a um, mantendo-os em seu *cluster* original ou movendo-os para outro *cluster*, de modo que seu centróide seja mais próximo da média daquele *cluster*. Recalcular o centróide para o *cluster* recebendo aquele item e para o *cluster* perdendo aquele item;

3. Repetir o passo 2 até que todos os itens tenham sido revisitados uma vez cada.

Para p variáveis, as novas p coordenadas do centróide são calculadas pelas fórmulas:

$$\bar{x}_{i,novo} = \frac{n\bar{x}_i + x_{ji}}{n+1}, \text{ se o j-ésimo item é adicionado ao grupo} \tag{2.3.7}$$

$$\bar{x}_{i,novo} = \frac{n\bar{x}_i - x_{ji}}{n-1}, \text{ se o j-ésimo item é removido do grupo} \tag{2.3.8}$$

Em seguida, pode-se utilizar qualquer distância para calcular a proximidade do centróide do item com seus *clusters*, sendo a distância Euclidiana a mais utilizada (JOHNSON; WICHERN, 2007).

2.3.4 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).

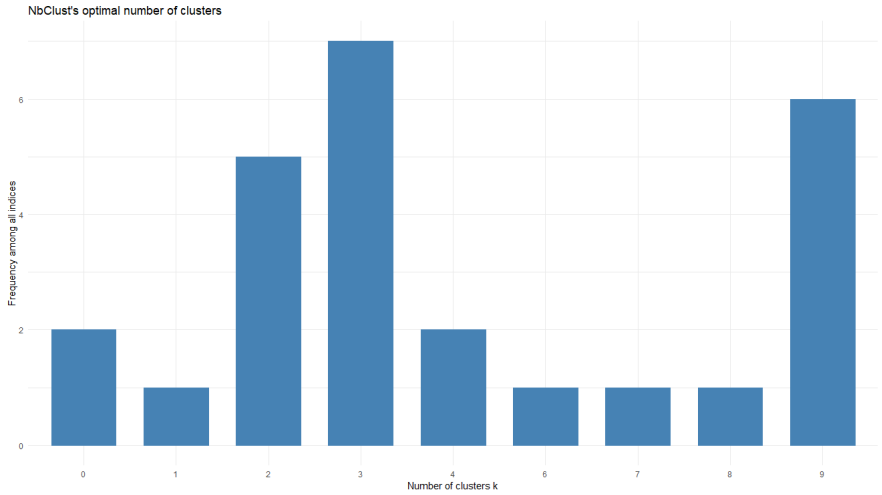
Seguem listados abaixo os índices a serem utilizados, cujos métodos de cálculo podem ser verificados em Charrad et al. (2014).

- 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*, conforme a figura abaixo:

Figura 2.3.3: Exemplo de Gráfico de Índices



Fonte: OLDACH (2019)

3 Metodologia

3.1 Banco de Dados

Os dados que serão utilizados para as conclusões foram retirados do site *Baseball Reference* por meio de *web scrapping* utilizando o pacote *rvest* do *R* e os pacotes *pandas, beautifulsoup4* e *selenium* do *Python*. As informações são referentes à temporada de 2020, com informações sobre as principais estatísticas de todos os arremessadores dos 30 times da MLB, totalizando 713 arremessadores.

Foram coletados 4 bancos de dados, com informações gerais, estatísticas de valor do arremessador, raio de arremesso do jogador e probabilidades de vitória acrescentadas ao time. O banco de dados Geral contém 35 variáveis, o banco Valor contém 17, o banco Raio de Arremesso contém 16 e, por fim, o banco Probabilidade de Vitória contém 21. Todos os bancos foram unidos em um só, totalizando 89 variáveis.

O banco Geral foi obtido pelo *software R*. Já os bancos Valor, Raio de Arremesso e Probabilidade de Vitória, por terem sido gerados dinamicamente, foram obtidos pelo *Python*, transformados em tabelas de formato html e, por fim, convertidos em bancos de dados no *software R*. Os significados das variáveis do banco, separadas por categoria, de acordo com a definição oficial da MLB (MLB, 2021), são traduzidos nas tabelas abaixo:

Tabela 1: Estatísticas Gerais do Jogador - Parte 1

Sigla	**Variável**	**Significado**
Pos	Posição do Jogador	Posição do arremessador: SP (*starting pitcher*, arremessador que inicia o jogo), RP (*relief pitcher*, arremessador que substitui o inicial) ou CL (*closer*, arremessador que termina o jogo)
Name	Nome do Jogador	Nome completo de cada arremessador
Age	Idade	Idade, em anos, do arremessador
W	Vitórias	Número de jogos em que ele foi arremessador oficial quando seu time assumiu a liderança até o final do jogo
L	Derrotas	Número de jogos em que o arremessador permite uma corrida, dando à equipe adversária liderança até o final do jogo
W-L%	Porcentagem de Vitórias	Número de vitórias dividido pela soma de vitórias e derrotas
ERA	Média de Corridas Ganhas	Número de corridas ganhas que um arremessador permite por nove entradas (9 vezes o número de corridas permitidas sobre o número de entradas jogadas)
G	Jogos	Número de jogos arremessados ou jogados
GS	Jogos Começados	Número de jogos iniciados pelo arremessador
GF	Jogos Terminados	Número de jogos finalizados pelo arremessador
CG	Jogos Completos	Número de jogos completos do arremessador, quase sempre 0
SHO	Shutouts	Número de jogos completos em que nenhuma corrida do time adversário foi anotada, quase sempre 0
SV	Saves	Número de jogos em que o arremessador foi substituto e terminou com vitória para o time, mas não configurou-se como uma Vitória (W). Além disso, ele deve ter jogado pelo menos 3 entradas ou entrar no jogo com a corrida de empate passível de ocorrer ou entrar no jogo com uma vantagem de não mais do que 3 corridas e arremessar pelo menos uma entrada
IP	Entradas Arremessadas	Número total de entradas em que ele esteve no jogo
H	Rebatidas Permitidas	Número de rebatidas do time oposto que o arremessador permitiu
R	Corridas Permitidas	Número de corridas anotadas pelo time oposto permitidas pelo arremessador
ER	Corridas Ganhas Permitidas	Número de corridas marcadas contra o arremessador, sem o benefício de um erro ou de uma bola passada
HR	Home Runs Permitidos	Número de Home Runs do time adversário permitidos pelo arremessador

Tabela 2: Estatísticas Gerais do Jogador - Parte 2

Sigla	**Variável**	**Significado**
BB	Andadas Concedidas	Número de vezes em que o arremessador lançou 4 bolas fora da zona de strike, concedendo uma base ao time adversário
IBB	Andadas Intencionais	Número de vezes em que o arremessador concedeu uma base propositalmente
SO	Strikeouts	Número de vezes que o arremessador eliminou um rebatedor
HBP	Atingido por Arremesso	Número de vezes que um arremesso atingiu o rebatedor adversário
BK	Empecilhos	Número de vezes em que um arremessador faz um movimento ilegal no monte que o árbitro considera enganoso para o corredor. Como resultado, qualquer homem na base recebe a próxima base, e o arremesso é descartado por uma bola morta.
WP	Arremessos Selvagens	Número de vezes em que seu arremesso é tão errôneo que o receptor é incapaz de controlá-lo e, como resultado, o(s) corredor(es) de base avançam
BF	Rebatedores Enfrentados	Número total de rebatedores que um arremessador enfrentou
ERA+	Média de Corridas Ganhas Ajustada	Ajuste para o estádio do arremessador, calculado como 100 vezes o log de ERA dividido por ERA
FIP	Arremessos Independentes de Campo	Efetividade do arremessador em prevenir HR, BB, HBP e causar strikes, calculado por 13*HR+3*(BB+HBP)-2*SO + constante, tal que a média de FIP é a mesma que a média de ERA de cada temporada
WHIP	Andadas e Rebatidas por Entrada Arremessada	Soma das andadas e rebatidas de um arremessador, dividida pelo total de entradas arremessadas
H9	Rebatidas Concedidas Ponderadas	9 vezes o número de rebatidas concedidas dividido pelo número de entradas arremessadas
HR9	Home Runs Permitidos Ponderados	9 vezes o número de home runs permitidos dividido pelo número de entradas arremessadas
BB9	Andadas Concedidas Ponderadas	9 vezes o número de andadas concedidas dividido pelo número de entradas arremessadas
SO9	Strikeouts Ponderados	9 vezes o número de strikeouts dividido pelo número de entradas arremessadas
SO/W	Strikeouts por Vitória	Número de strikeouts por andadas concedidas
TEAM	Time	Time do arremessador
Hand	Mão de Arremesso	Mão que o arremessador utiliza: R (direita), L (esquerda) ou B (ambas)

Tabela 3: Estatísticas de Valor do Jogador

Sigla	**Variável**	**Significado**
RA9	Corridas Permitidas a cada 9 Entradas Arremessadas	Semelhante a ERA, mas leva em consideração as corridas não ganhas também
RA9opp	Corridas Permitidas a cada 9 Entradas dos Oponentes	Média de corridas anotadas pelos oponentes desse arremessador, multiplicada por 9 entradas
RA9def	Corridas a cada 9 Entradas Arremessadas Evitadas pela Defesa	Valores negativos significam que a defesa foi abaixo da média, enquanto valores positivos indicam que foi acima
RA9role	Diferença de Corridas a cada 9 Entradas Arremessadas para Arremessadores Iniciais e Arremessadores Substitutos	Mostra quantas corridas por jogo, em média, um arremessador inicial evita a mais que um substituto
RA9extras	Diferença de Corridas a cada 9 Entradas para Entradas Extras	Ajuste do número de corridas esperadas do arremessador que é adicionado ao se ter um jogador na segunda base, para cada vez que o arremessador começa uma entrada extra
PPFp	Fator do Estádio	Ajuste para todos os estádios que o arremessador jogou na temporada
RA9avg	Corridas a Cada 9 Entradas para um Arremessador Mediano	Estimativa de como um arremessador mediano se sairia contra esse arremessador, com essa defesa e nesse estádio, calculada por PPFp/100*(oppRA9-RAdef+RArole)
waaWL%	Vitória-Derrota com Time Mediano	Razão de vitórias por derrotas num time mediano, apenas para os jogos em que o arremessador jogou
162WL%	Vitória-Derrota com Time Mediano na Temporada	Semelhante ao waaWL%, assume esse valor para os jogos em que o arremessador jogou e 0,500 para jogos que ele não jogou
RAA	Corridas Melhores que a Média	Número de corridas que esse jogador é melhor que um jogador mediano, calculado por IP*(RA9avg-RA9)/9
WAA	Vitórias Acima da Média	Número de vitórias que o arremessador acrescenta ao time a mais do que um mediano adicionaria, calculado por (waaWL% - 0.5) vezes o número de jogos
gmLI	Índice de Alavancagem na Entrada do Jogo	Apenas para arremessadores substitutos, é a média do índice de alavancagem de abertura de aparências ajustada pelos rebatedores enfrentados naquela situação
WAAadj	Ajuste de Vitórias Acima da Média	Valor adicional da alavancagem que aquele jogador acrescenta ao time. Para arremessadores substitutos, é meramente WAA*(1+gmLI)/2
WAR	Vitórias Acima da Substituição	Número de vitórias que o arremessador acrescenta ao time a mais do que um substituto adicionaria, calculado pela soma de WAAadj, WAA e o o valor do substituto
RAR	Corridas Melhor que o Nível de Substituição	Número de corridas que esse jogador é melhor que um jogador de substituição
Salary	Salário	Salário anual do arremessador, em dólares
Acquired	Forma Como o Jogador Foi Adquirido	Possui 9 níveis: Recrutamento Amador, Agente Livre Amador, Designado, Agente Livre, Comprado, Recrutamento da Regra 5, Retornado da Regra 5, Trocado e Renúncia

Tabela 4: Estatísticas de Raio de Arremesso do Jogador

Sigla	**Variável**	**Significado**
Ptn%	Vantagem "Platoon"	Proporção de vezes em que um arremessador enfrenta um rebatedor com a mesma lateralidade que ele
HR%	Porcentagem de Home Runs	Número de home runs concedidos sobre todas as aparições do arremessador
SO%	Porcentagem de Strikeouts	Número de strikeouts sobre todas as aparições do arremessador
BB%	Porcentagem de Andadas Concedidas	Número de andadas concedidas sobre todas as aparições do arremessador
SO-BB%	Porcentagem de Strikeouts-Andadas Concedidas	Diferença entre o número de strikeouts e o número de andadas concedidas, dividida por todas as aparições do arremessador
XBH%	Porcentagem de Rebatidas de Base Extra	Proporção de todas as aparições terminando em rebatidas de base extra
X/H%	Porcentagem de Todas as Rebatidas por Bases Extras	Proporção de todas as rebatidas resultando em bases extras
GB/FB	Razão de Bolas Terrestres por Bolas Voadoras	Divisão de todas as bolas terrestres pelas bolas voadoras
GO/AO	Eliminações Baixas por Eliminações Aéreas	Razão entre eliminações feitas com bolas baixas rebatidas e com bolas voadoras
IP%	Porcentagem de Bolas em Jogada	Proporção de todas as jogadas em que a bola foi colocada em jogo
LD%	Linha Reta	Porcentagem de bolas arremessadas que foram rebatidas em linha reta
HR/FB	Porcentagem de Bolas Voadoras que Viraram Home Runs	Proporção de bolas voadoras que configuraram home runs
IF/FB	Porcentagem de Bolas Voadoras Dentro do Campo	Proporção de bolas voadoras que pararam dentro do campo
Opp	Oportunidade de Queimada Dupla	Número de vezes com corredor na primeira base e menos de 2 eliminações
DP	Queimada Dupla	Número de vezes que o arremessador eliminou 2 ou mais corredores em uma única jogada com bola terrestre
%	Razão Queimada Dupla	Proporção de vezes que o arremessador eliminou 2 ou mais corredores em uma única jogada com bola terrestre

Tabela 5: Estatísticas de Probabilidade de Vitória do Jogador

Sigla	Variável	Significado
PtchR	Corridas do Arremessador Ajustadas	Estima a contribuição de um arremessador para o total de corridas anotadas por seu time, com média 0
PtchW	Vitórias do Arremessador Ajustadas	Estima a contribuição de um arremessador para o total de vitórias de seu time, com média 0
WPA+	Probabilidade de Vitória Adicionada	Soma de eventos positivos para esse jogador
WPA-	Probabilidade de Vitória Subtraída	Soma de eventos negativos para esse jogador
WPA	Probabilidade de Vitória Adicionada pelo Arremessador	Número de vitórias que o arremessador acrescenta ao seu time, é dado pela soma de WPA+ e WPA-
Plays	Jogadas Inclusas no Cálculo de WPA	Número de jogadas que foram usadas para o cálculo da WPA
aLI	Índice de Alavancagem Médio	Pressão média que o jogador sentiu na temporada, com valores abaixo de 1 para pouca pressão, iguais a 1 para pressão mediana e acima de 1 para muita pressão
WPA/LI	Vitórias Situacionais	Dado pela soma de WPA divido pelo índice de alavancagem
Clutch	"Garra"	Mede a habilidade do jogador de se adaptar a situações de alta pressão, calculado por WPA/aLI - WPA/LI
cWPA	Probabilidade de Vitória Adicionada pelo Arremessador no Campeonato	Probabilidade de vitórias de campeonato que o arremessador acrescenta ao seu time
cWPA+	Probabilidade de Vitória Adicionada no Campeonato	Soma de eventos positivos para esse jogador, em porcentagem
cWPA-	Probabilidade de Vitória Subtraída no Campeonato	Soma de eventos negativos para esse jogador, em porcentagem
acLI	Índice de Pressão de Campeonato Médio	Pressão média que o jogador sentiu no campeonato, com valores abaixo de 1 para pouca pressão, iguais a 1 para pressão mediana e acima de 1 para muita pressão
cClutch	"Garra"em Campeonato	Mede a habilidade do jogador de se adaptar a situações de alta pressão, calculado por cWPA/aLI - cWPA/LI
RE24	Corridas Base-Fora Salvas	Quantas corridas o arremessador salvou, dada a situação de bases e foras, com média 0
REW	Vitórias Base-Fora Salvas	Número de vitórias acima da média que o jogador vale, com base em seu desempenho
boLI	Alavancagem Base-Fora	Alavancagem base-fora média
RE24/boLI	Corridas Situacionais	Dado pela soma de RE24 dividido pela alavancagem base-fora
LevHi	Jogos Entrados com Alavancagem Alta	Número de jogos entrados em que a alavancagem era maior ou igual a 1,5
LevMd	Jogos Entrados com Alavancagem Média	Número de jogos entrados em que a alavancagem estava entre 0,7 e 1,5
LevLo	Jogos Entrados com Alavancagem Baixa	Número de jogos entrados em que a alavancagem era menor ou igual a 0,7

3.2 Técnicas Utilizadas

Inicialmente, serão removidas as variáveis W-L%, ERA+, SO/W, gmLI, X/H%, GB/FB, GO/AO, HR/FB, IF/FB %, WPA/LI, Clutch e cClutch, uma vez que estas apresentam muitos valores faltantes e pouca informação agregada, o que prejudicaria a construção das análises. Após esse procedimento, obteve-se um banco de dados com 76 variáveis e 711 observações, com apenas os jogadores Mike Brosseau e Brock Holt

excluídos.

Para se obter a classificação, será utilizada a análise fatorial exploratória, a fim de reduzir os dados em menos variáveis, que descrevam a variabilidade dos dados e expliquem o traço latente de habilidade do arremessador. Será utilizado o método dos mínimos resíduos e a rotação *Varimax*, uma vez que esse procedimento facilita a interpretação dos fatores e busca estruturas mais simples de cargas. Em seguida, serão construídos modelos de equações estruturais, 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), GFI (Índice de Qualidade de Ajuste), RMSEA (Erro Quadrático Médio de Aproximação) e BIC (Critério de Informação Bayesiano).

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 *complete linkage* e *average linkage*, com base nas distâncias de Manhattan e de Mahalanobis. 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 arremessador que melhor influenciam em sua habilidade, além de como melhor classificá-los de acordo com semelhanças em suas competências.

4 Resultados

Nesta seção, serão apresentados os resultados obtidos com as análises realizadas ao decorrer do trabalho.

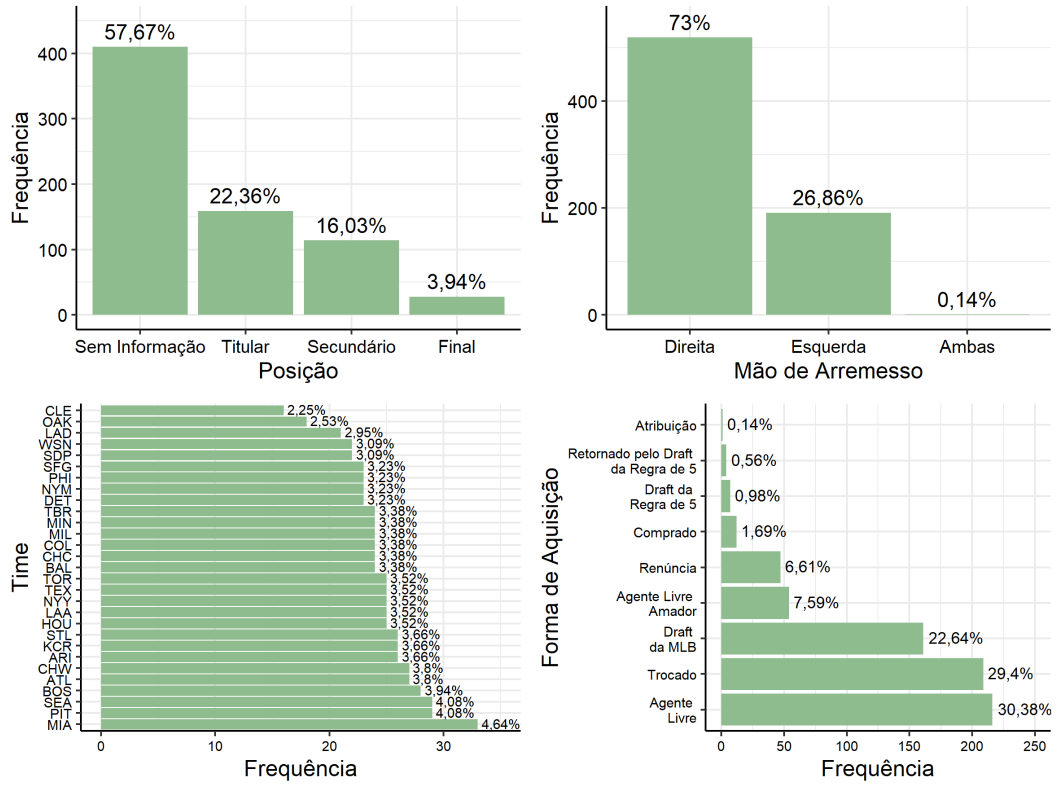
4.1 Análise Descritiva

Inicialmente, foi feita a análise descritiva da amostra, bem como análises de correlações entre as variáveis, a fim de compreender possíveis relações entre elas.

4.1.1 Análise Descritiva Univariada

A fim de compreender as variáveis disponíveis, será realizada a análise descritiva univariada de cada uma das variáveis disponíveis.

Figura 4.1.1: Análise Descritiva das Variáveis Qualitativas



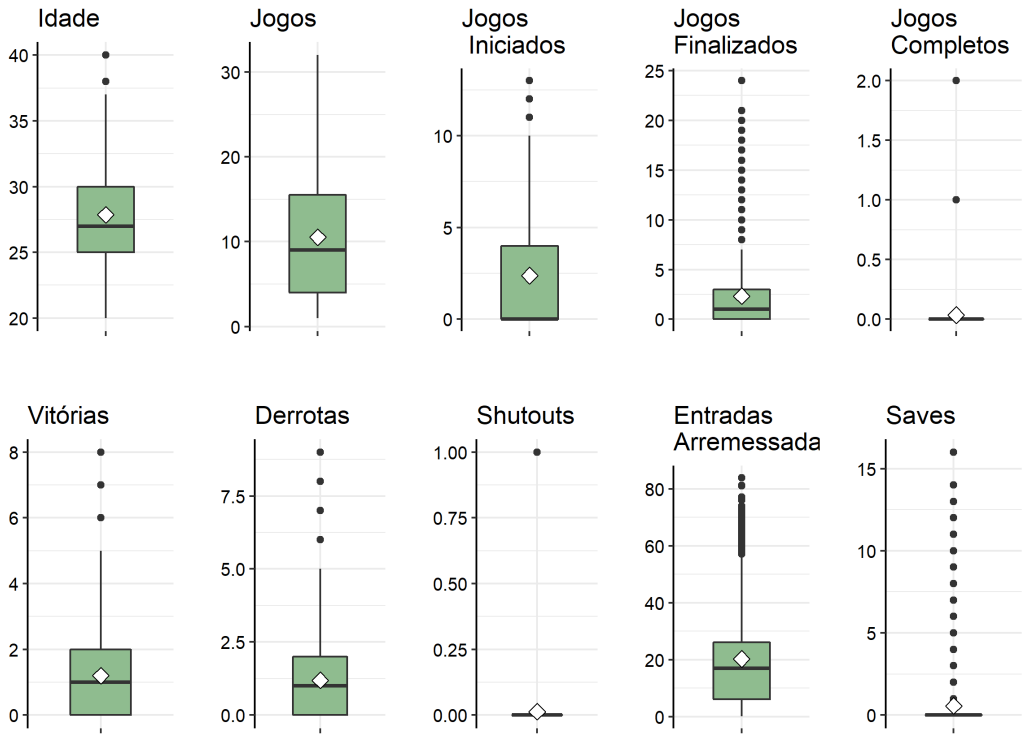
A Figura 4.1.1 mostra o comportamento de todas as variáveis qualitativas do banco. É possível perceber que mais da metade dos jogadores não têm posição definida, enquanto quase um quarto jogam como titulares e os restantes jogam como secundários ou

finais, sendo os jogadores finais a categoria menos frequente. Quanto à mão de arremesso, 73% são destros enquanto 26,86% são canhotos e apenas um jogador (correspondente a 0,14%) arremessa com ambas as mãos.

Em relação ao time, nota-se que Miami Marlins é o time com maior frequência de arremessadores na liga, seguidos por Seattle Mariners e Pittsburgh Pirates, empatados na segunda posição de maior frequência, enquanto o time Cleveland Indians é o time com menos arremessadores. Por fim, quanto à forma de aquisição do jogador, a forma mais comum é agente livre, enquanto a menos comum é por atribuição.

Ademais, será feita a análise univariada das variáveis quantitativas, que segue nas Figuras 4.1.2 a 4.1.8:

Figura 4.1.2: Análise Descritiva das Variáveis Quantitativas

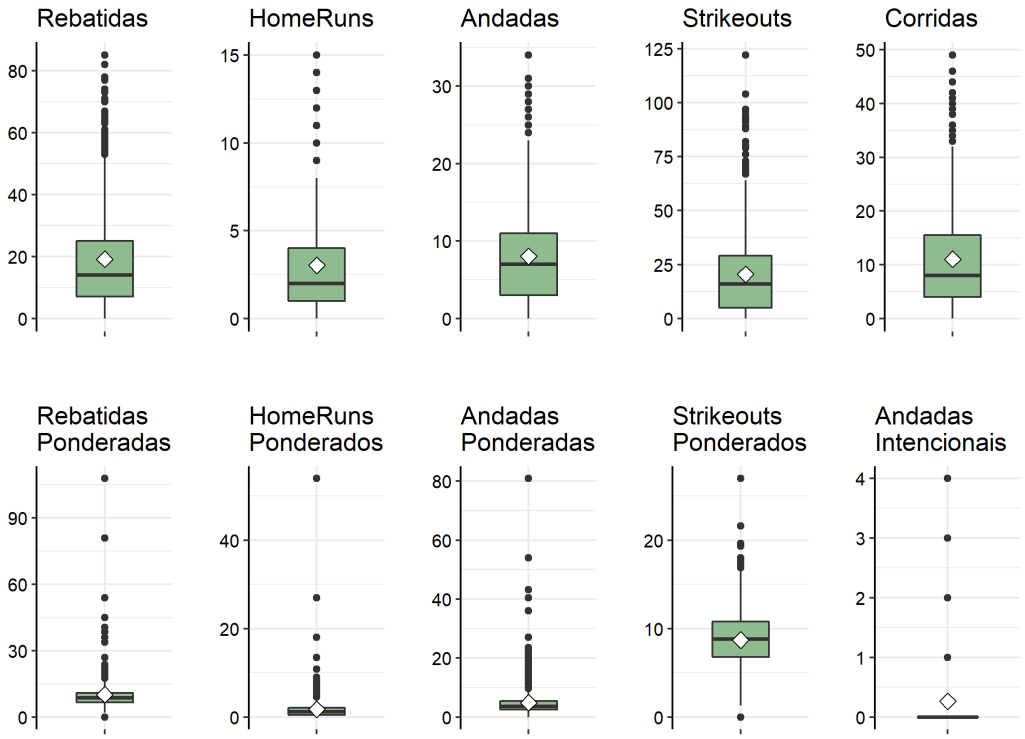


Por meio da Figura 4.1.2, nota-se que apenas a variável Jogos (que representa o número de jogos de cada arremessador em 2020) não apresentou pontos discrepantes, com sua média próxima da mediana, de mais ou menos 10 jogos por jogador. Além disso, as variáveis *Shutouts*, Jogos Completos e Salvamentos (*Saves*) apresentaram suas médias, medianas e quartis bem próximos de 0, com a última apresentando *outliers* de até 16 salvamentos. Nota-se que as variáveis *Shutouts* e Jogos Completos são, quase sempre, iguais a 0, com raras exceções de 1 *shutout* por jogador ou até 2 jogos completos por

jogador, indicando que essas variáveis podem não ser boas indicadoras de desempenho.

Ainda, percebe-se que as variáveis Vitórias e Derrotas apresentaram comportamentos bem semelhantes, com a média e a mediana em torno de 1 vitória ou derrota por arremessador. Esse resultado baixo se deve ao fato de que nem sempre será atribuída uma vitória ou uma derrota ao jogador, mesmo que tenha feito parte de um jogo em que seu time foi vitorioso ou perdedor, uma vez que existem critérios definidos pela MLB para a concessão do número de vitórias ou derrotas de um arremessador. Evidencia-se, também, que cada arremessador joga, em média, cerca de 20 entradas numa temporada, com *outliers* de até mais de 80 entradas e um média de cerca de 2 jogos finalizados e 2 jogos iniciados por jogador, o que, novamente, não podem ser boas variáveis preditoras de desempenho, uma vez que os arremessadores finais e titulares naturalmente tendem a ter maiores números de jogos finalizados e iniciados, respectivamente.

Figura 4.1.3: Análise Descritiva das Variáveis Quantitativas

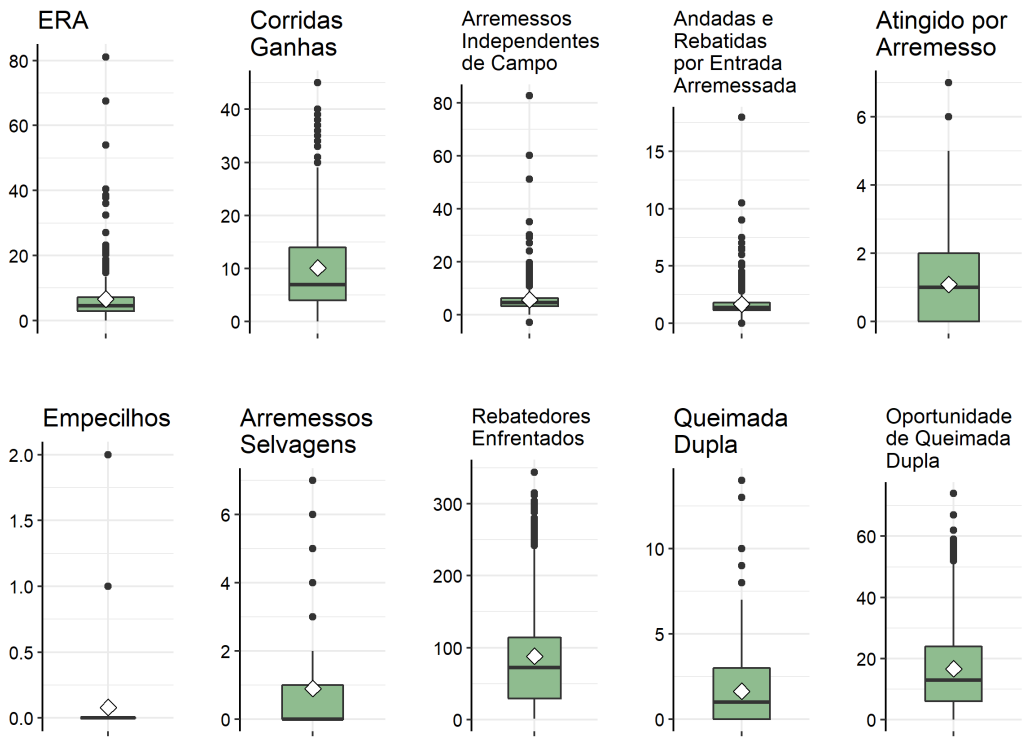


Pela Figura 4.1.3, observa-se que as variáveis Rebatidas Ponderadas, *Home Runs* Ponderados, Andadas Ponderadas e *Strikeouts* Ponderados consistem em uma combinação linear de suas respectivas variáveis (as que se encontram acima delas na imagem) e do número de entradas arremessadas por jogador. Apesar das variáveis ponderadas apresentarem comportamento mais irregular do que as variáveis originais, é evidente que estas

apresentam uma melhor ferramenta de comparação entre os jogadores, uma vez que fornecem variáveis escaladas para cada arremessador, levando em conta os diferentes números de oportunidades que eles obtiveram de permitir rebatidas, *home runs*, andadas e causar *strikeouts*.

Ademais, percebe-se que a variável Andadas Intencionais apresentou mediana e quartis iguais a 0, com *outliers* de no máximo 4 andadas intencionais, indicando que esta talvez não seja uma boa variável preditora de desempenho. Por fim, a variável Corridas apresenta uma média e uma mediana bem próximas de cerca de 10 corridas permitidas por jogador, com pontos discrepantes de até cerca de 50 corridas.

Figura 4.1.4: Análise Descritiva das Variáveis Quantitativas

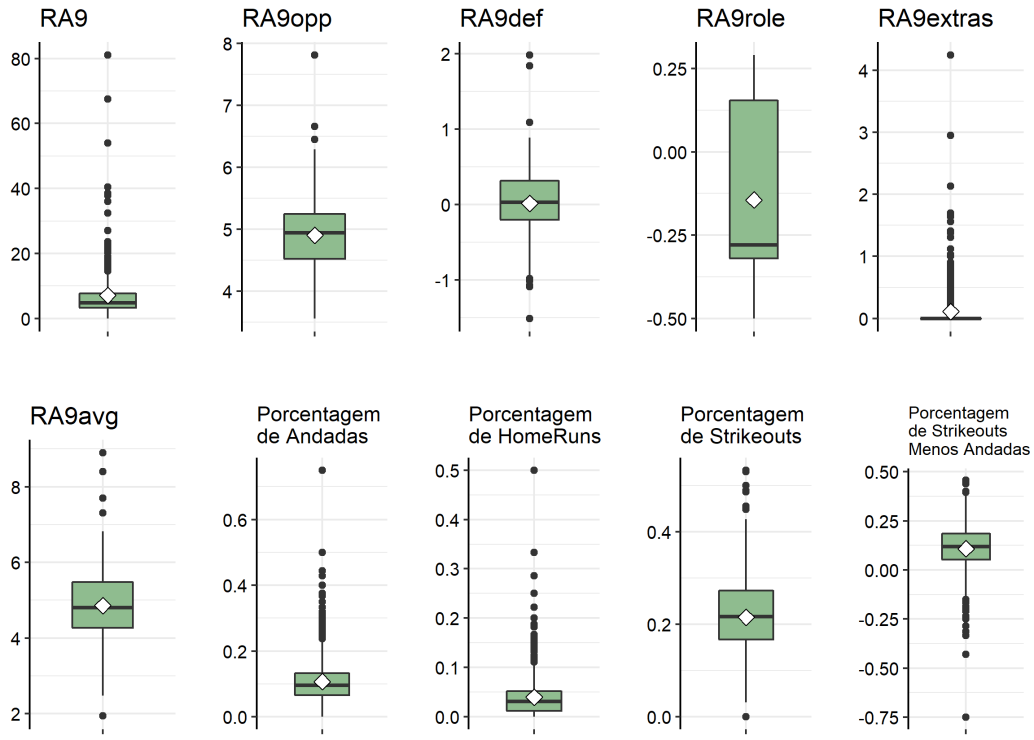


Na Figura 4.1.4, verifica-se que, assim como as variáveis anteriores, a variável ERA é uma versão ponderada das Corridas Ganhas, indicando que apenas uma das variáveis deve ser utilizada para a determinação do desempenho do jogador. Percebe-se, também, comportamento semelhante entre as variáveis FIP (Arremessos Independentes de Campo) e WHIP (Andadas e Rebatidas por Entrada Arremessada), com a presença de muitos *outliers*. Além disso, variáveis como Empecilhos, Atingido por Arremesso e Arremessos Selvagens apresentam indicações de não serem necessárias para descrever a habilidade, uma vez que apresentam até 75% dos seus valores iguais a 0, 2 ou 1, respecti-

vamente, com *outliers* de até 2 empecilhos, 7 atingimentos por arremesso ou 7 arremessos selvagens.

Ainda, é notório que cada arremessador enfrentou, em média, cerca de 100 rebatedores durante a temporada, com alguns *outliers* enfrentando até quase 400 rebatedores. Por fim, nota-se uma média de cerca de 20 oportunidades de queimada dupla por jogador, com *outliers* de no máximo cerca de 80 oportunidades e destas, uma média de cerca de 2 queimadas duplas por arremessador de fato realizadas, com um máximo de cerca de 20 queimadas duplas.

Figura 4.1.5: Análise Descritiva das Variáveis Quantitativas

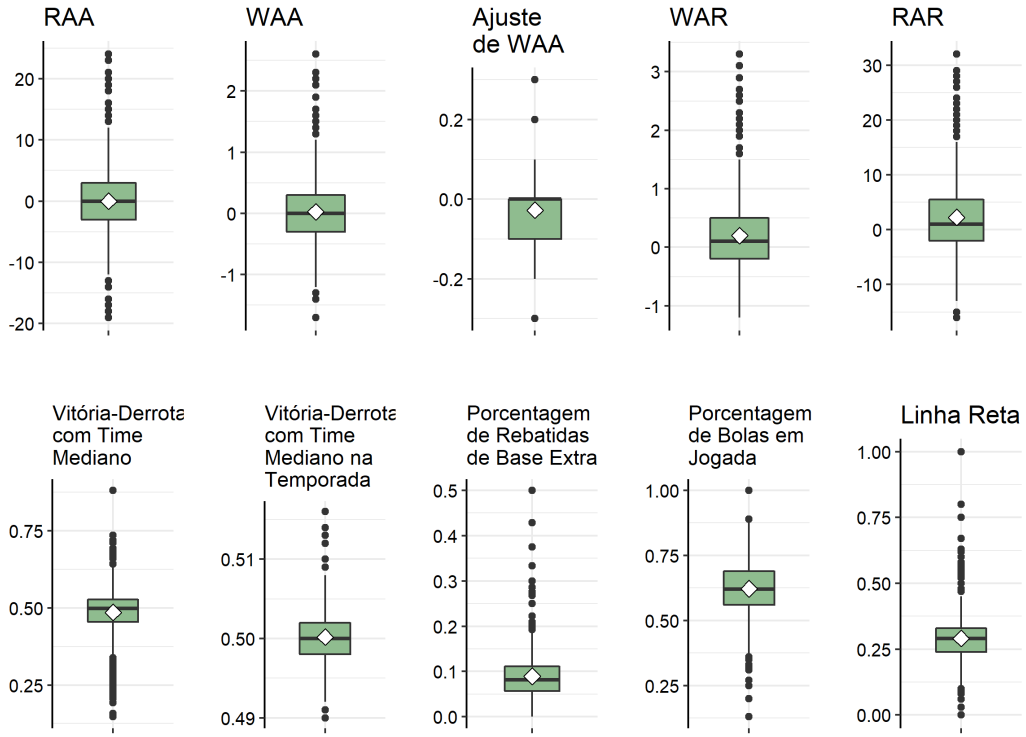


Por meio da Figura 4.1.5, evidencia-se que as variáveis em porcentagem são combinações lineares de outras já apresentadas anteriormente. Assim, é provável que elas apresentem correlação elevada com outras variáveis, podendo causar multicolinearidade no modelo.

Já as variáveis cujos nomes iniciam com "RA9" indicam as corridas a cada 9 entradas que foram permitidas pelo arremessador. É possível perceber que em variáveis como *RA9extras*, por exemplo, 75% dos arremessadores apresentaram resultado igual a 0. Além disso, a variável *RA9role* não apresentou *outliers*, com valores entre -0,5 e cerca de 0,25. A variável *RA9* foi a que apresentou maior discrepância, com valores até 80 mas

com três quartos de seus valores entre 0 e 10.

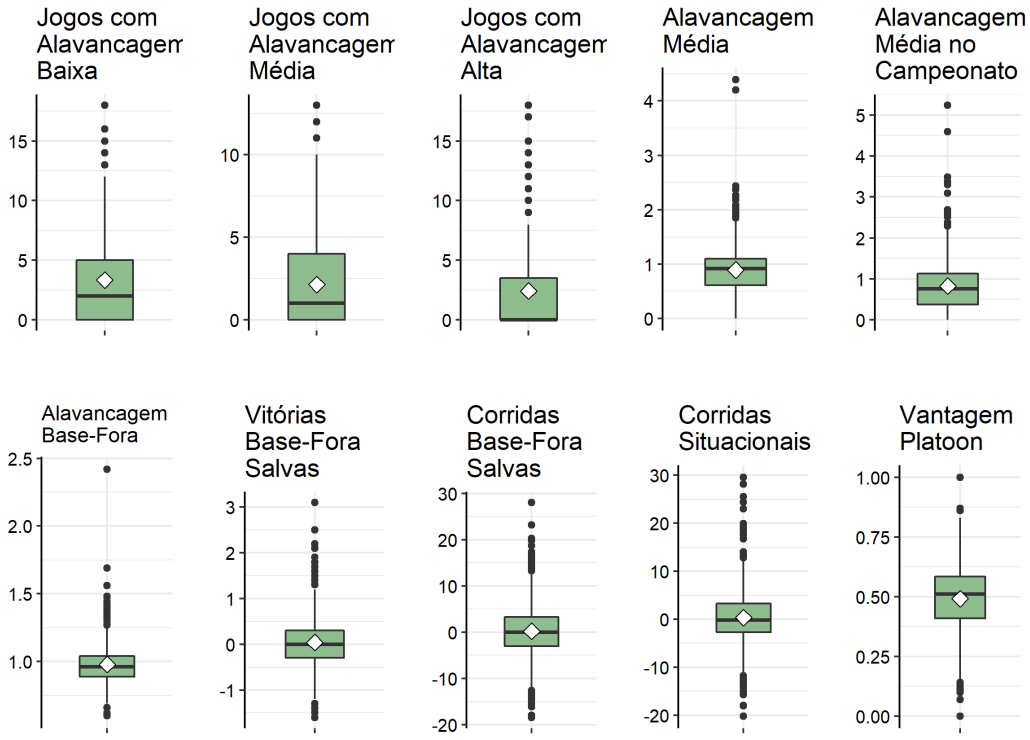
Figura 4.1.6: Análise Descritiva das Variáveis Quantitativas



A Figura 4.1.6 mostra que as variáveis de comparação com jogadores medianos ou com jogadores a nível de substituição (gráficos da linha superior) apresentam comportamento semelhante, com média e mediana em torno de 0 e distribuição relativamente simétrica, apresentando *outliers* acima e abaixo de ambos os limites (com exceção de WAR, que possui valores discrepantes apenas acima do limite superior).

Outrossim, percebe-se que a proporção de vitórias e derrotas com um time mediano, tanto no geral quanto na temporada, apresentou média e mediana próximas de 50%, indicando um equilíbrio entre o número de vitórias e derrotas dos arremessadores. Quanto às variáveis Porcentagem de Bolas em Jogada e Linha Reta, observa-se uma distribuição com média de cerca de 0,625 e 0,3, respectivamente, chegando até a proporção 1. Por fim, para a variável Porcentagem de Rebatidas de Base Extra, nota-se média e quartis bem próximos de 0,1, com *outliers* até 0,5.

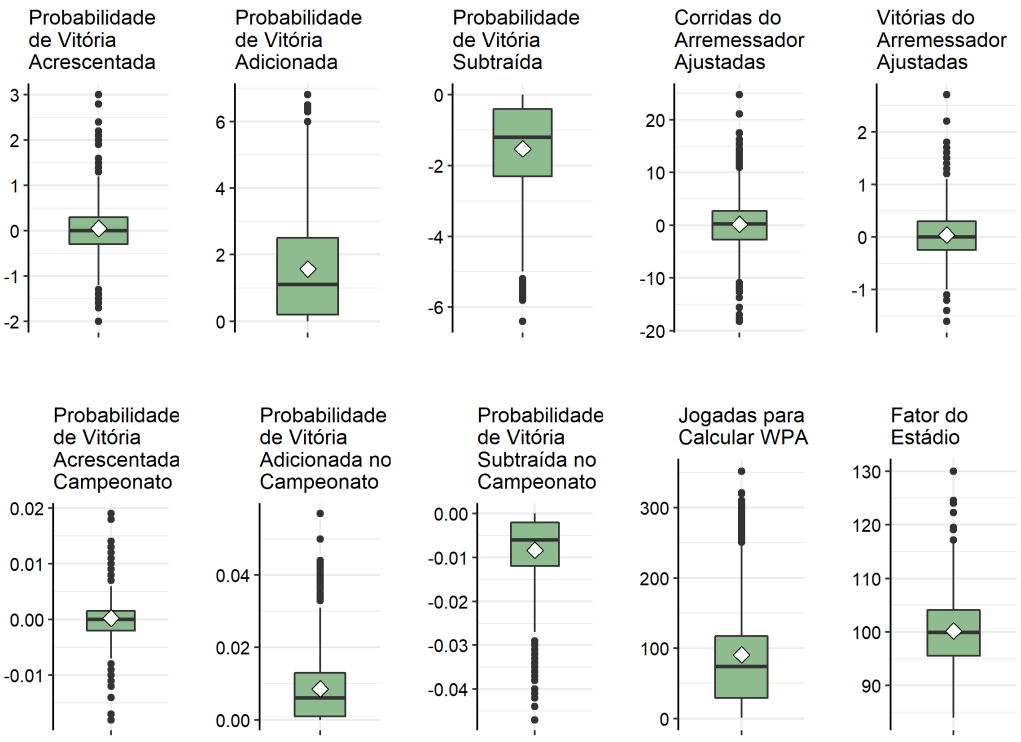
Figura 4.1.7: Análise Descritiva das Variáveis Quantitativas



A partir da Figura 4.1.7, nota-se que as variáveis Jogos com Alavancagem Baixa, Média e Alta apresentam distribuições semelhantes, com 75% de seus valores abaixo de 5 jogos e valor máximo entre 10 e 20 jogos. Percebe-se, também, comportamento semelhante nas variáveis Alavancagem Média, Alavancagem Média no Campeonato e Alavancagem Base-Fora, com média e quartis em torno de 1, variando de 0 até cerca de 5.

Da mesma forma, as variáveis Vitórias Base-Fora Salvas, Corridas Base-Fora Salvas e Corridas Situacionais apresentam, também, comportamentos semelhantes, centrados no 0. Já a variável Vantagem *Platoon*, que mostra a vantagem do rebatedor em relação à lateralidade do arremessador apresentou média e mediana em torno de 0,5, com poucos pontos discrepantes acima do limite superior e muitos abaixo do limite inferior.

Figura 4.1.8: Análise Descritiva das Variáveis Quantitativas

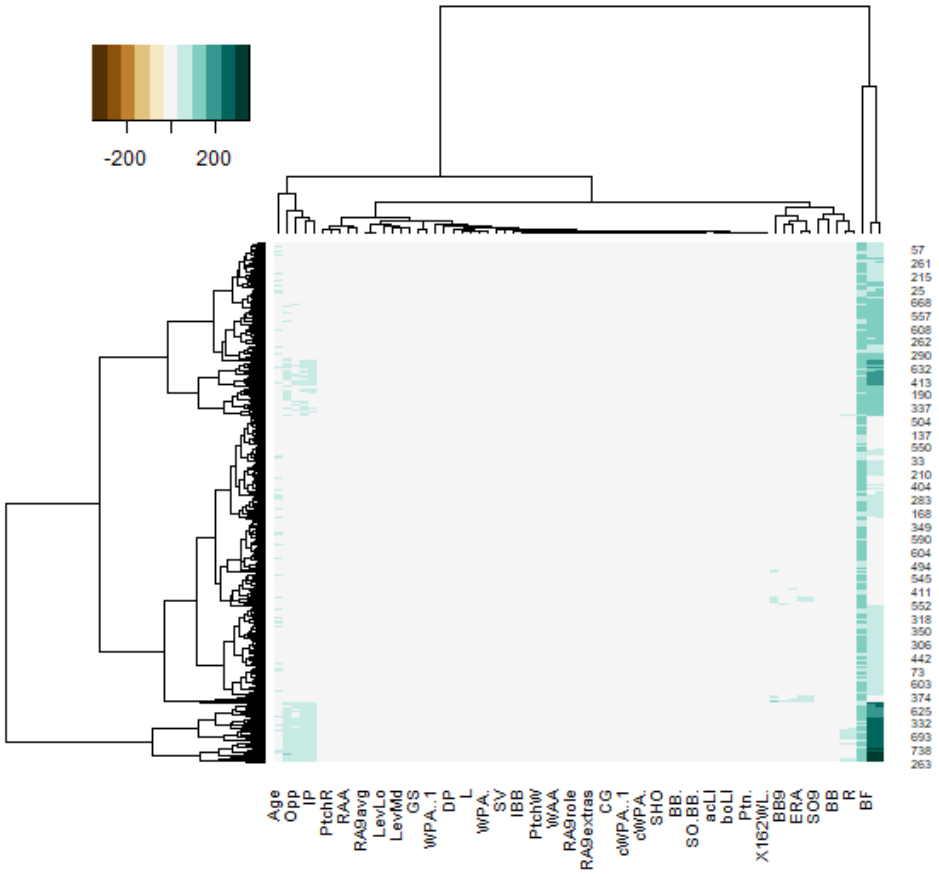


Pela Figura 4.1.8, nota-se que as variáveis Corridas do Arremessador Ajustadas e Vitórias do Arremessador Ajustadas apresentam distribuição semelhante, mas em escalas diferentes. Essas duas variáveis apresentam, também, comportamento similar ao das variáveis Probabilidade de Vitória Acrescentada (WPA) e Probabilidade de Vitória Acrescentada no Campeonato, que por sua vez, são a soma de Probabilidade de Vitória Adicionada com Probabilidade de Vitória Subtraída e Probabilidade de Vitória Adicionada no Campeonato com Probabilidade de Vitória Subtraída no Campeonato, respectivamente. Assim, é provável que estas variáveis apresentem correlação elevada.

Por fim, as variáveis Jogadas para Calcular o WPA e Fator do Estádio apresentam valores superiores, chegando até a cerca de 400 e 130, respectivamente, com a presença de muitos *outliers* acima do limite superior e média em torno de 100.

Uma forma de verificar a escala das variáveis, bem como uma divisão preliminar de *clusters* e redução de dimensionalidade é por meio do mapa de calor. Na Figura 4.1.9, segue o mapa de calor das variáveis:

Figura 4.1.9: Mapa de Calor

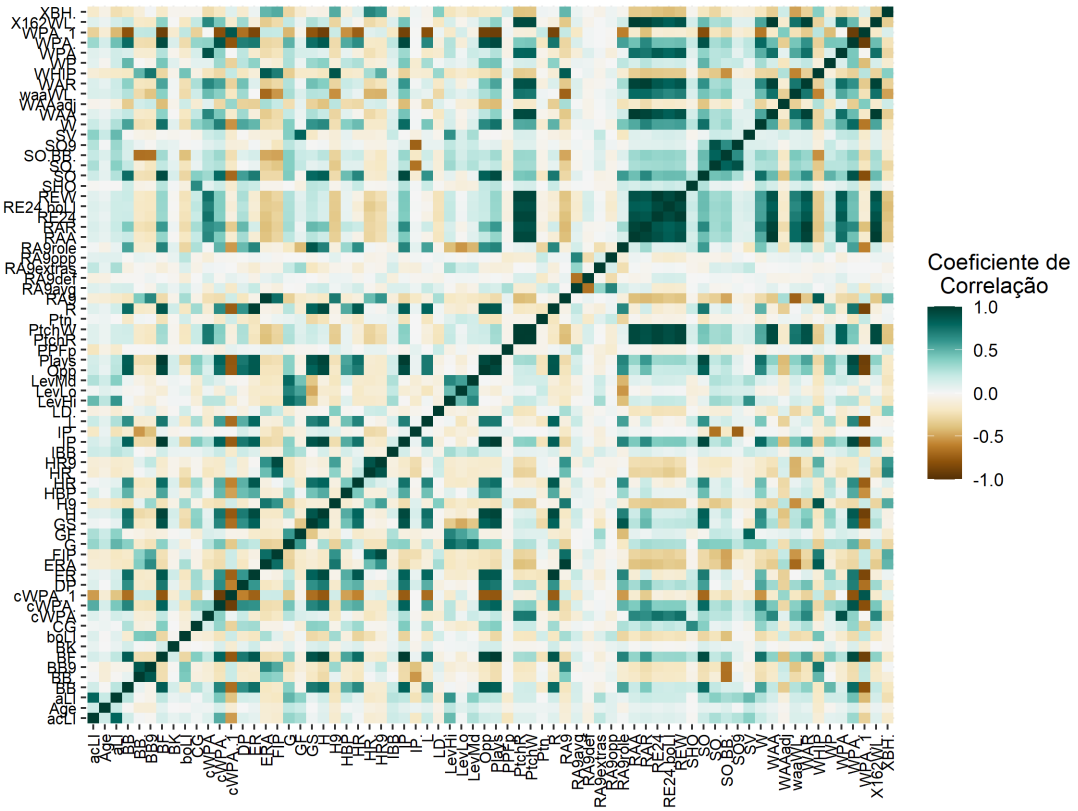


Percebe-se, pela Figura 4.1.9 a semelhança entre variáveis como Idade, Oportunidade de Queimada Dupla e Entradas Arremessadas, que foram classificadas como um mesmo fator, além de variáveis como Rebatedores Enfrentados, Fator do Estádio e Jogadas para Calcular WPA, em outro fator. Nota-se, também, a criação de aproximadamente 3 ou até mesmo 6 *clusters* de arremessadores.

4.1.2 Análise de Correlações

Para se entender melhor as relações entre as variáveis, foram feitos correlogramas que mostram o coeficiente de correlação de Pearson entre as variáveis quantitativas:

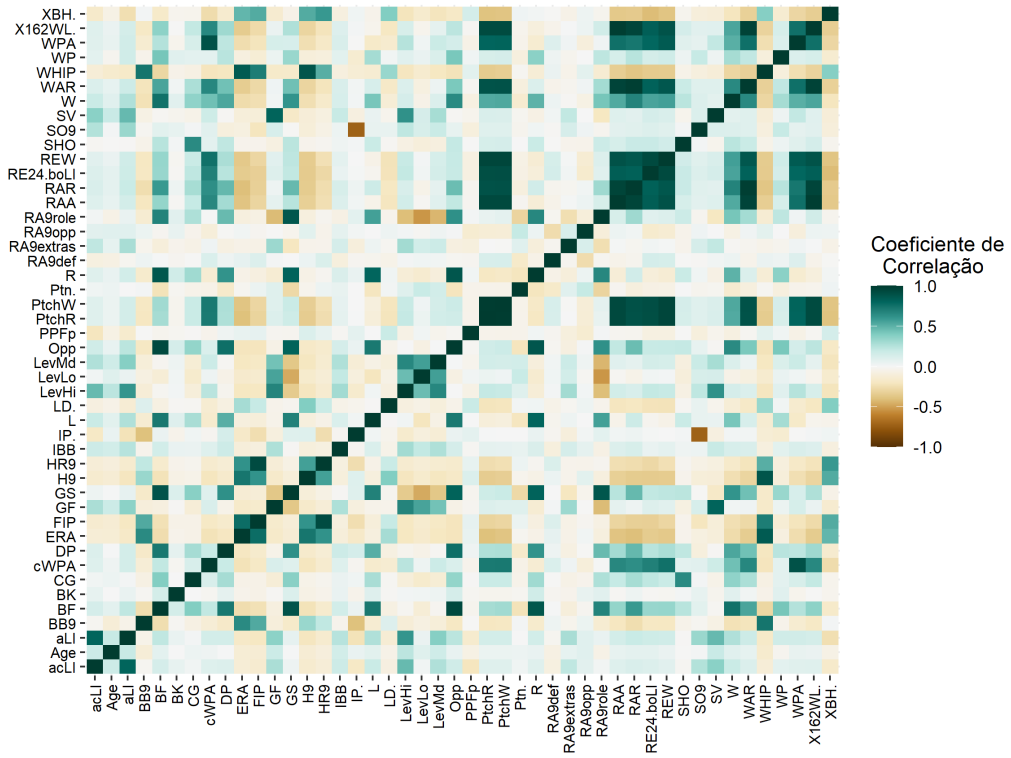
Figura 4.1.10: Correlograma das Variáveis



Pela Figura 4.1.10, evidencia-se a formação de um grande grupo de variáveis fortemente correlacionadas entre si, formado pelas variáveis RAA, RAR, RE24, RE24/boLI e REW, que fazem a comparação do jogador com jogadores a nível de substituição e medem as estatísticas de acordo com a situação Base-Fora do jogo e as variáveis W, waaWL%, PtchW, PtchR, WAA, WAR, WPA, WPA+, cWPA e 162W-L%, que medem, basicamente, o quanto o jogador acrescentou a seu time em termos de corridas ou vitórias estimadas.

Nota-se, também, alta correlação entre as variáveis SO, SO%, SO/BB e SO9, como já se era esperado, uma vez que todas tratam do número de *Shutouts*, mas de maneiras ponderadas ou proporcionais. Além disso, os pares de variáveis PtchR e PtchW, Plays e Opp, HR e HR9, FIP e ERA, BB% e BB9, aLi e acLI e H e GS apresentaram correlações elevadas, sendo que o último par também apresentou correlação elevada com LevLo, LevHi e LevMd, que, por sua vez, apresentaram correlações fortes entre si.

Figura 4.1.11: Correlograma das Variáveis após Remoção



Outro teste foi realizado removendo-se as variáveis que são combinações lineares de outras. As variáveis removidas foram: G, IP, H, ER, HR, BB, SO, HBP, RA9, RA9avg, WAAadj, waaWL%, HR%, SO%, BB%, SO/BB%, Plays, WPA+, WPA-, cWPA+, cWPA-, RE24 e boLI. Após a remoção dessas variáveis, observa-se na Figura 4.1.11 que as variáveis ERA, FIP, WHIP, H9 e HR9 são altamente correlacionadas entre si, bem como R, RA9role, BF, GS e L, além de PtchW e PtchR e, por fim, RAA, RAR, WAR, W, WPA e X162WL%.

4.2 Análise Fatorial

Inicialmente, após a remoção das devidas variáveis que constituíam combinações lineares das demais, foram feitas análises fatoriais com este grupo de 70 variáveis. Após diversas análises, notou-se um padrão de repetição de 24 variáveis, as quais foram selecionadas como as mais importantes, sendo essas: RAR, WAR, REW, PtchR, PtchW, RAA, 162WL%, RE24/boLI, W, ERA, BF, WHIP, H9, Opp, DP, GS, FIP, XBH%, HR9, BB9, RA9role, CG, aLI e WPA. No entanto, optou-se por remover WPA da análise fatorial para que fosse realizada a análise de correlação com o escore de habilidade final a ser calculado, a fim de se obter um parâmetro de avaliação desse escore.

4.2.1 Análise Fatorial Exploratória

Assim, com as 23 variáveis selecionadas, foi realizada a análise de quantidade de fatores necessários, por meio de um *screeplot* e uma análise paralela.

Figura 4.2.1: *Screeplot* com Variáveis Selecionadas

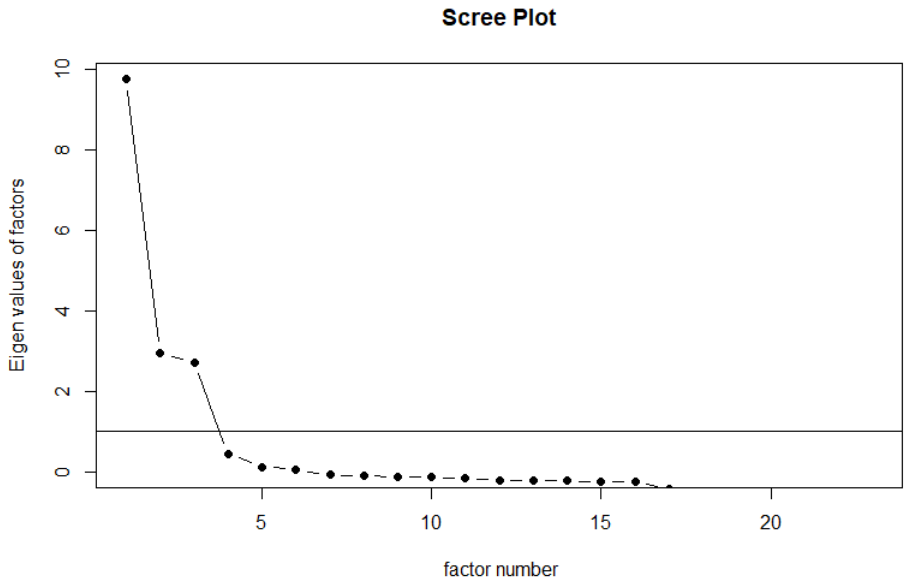
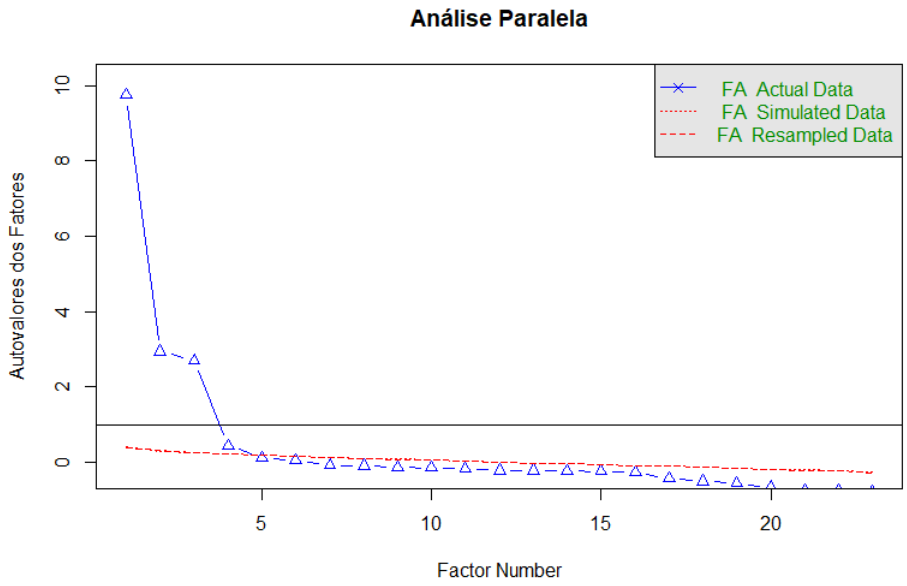


Figura 4.2.2: *Parallel plot* com Variáveis Selecionadas



O *screeplot* sugeriu o uso de 3 fatores, enquanto a análise paralela sugeriu 4 fatores. Assim, os modelos com ambos os números de fatores foram comparados:

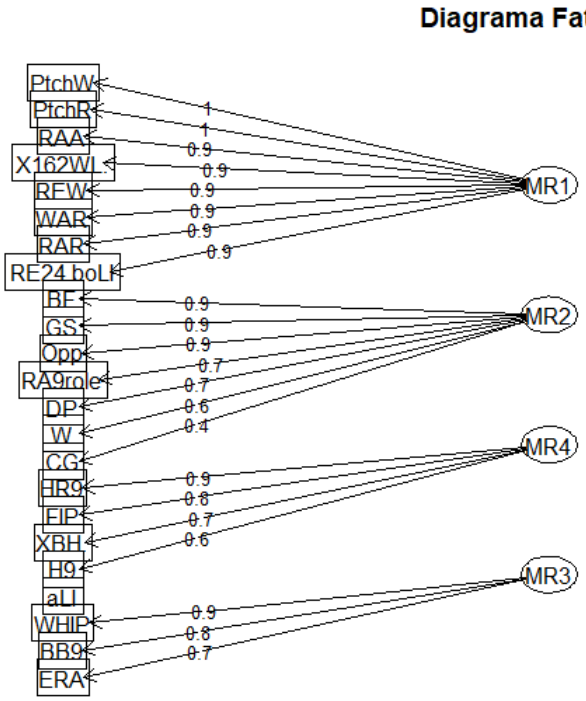
Tabela 6: Medidas de Ajuste para Modelos com Variáveis Selecionadas

Número de Fatores	BIC	RMSEA	TLI	Fit	Explicação da Variância
3	14457,35	0,341	0,42	0,99	71,6%
4	13520,09	0,349	0,394	1	76,3%

De acordo com as medidas de ajuste de cada um dos modelos propostos, foi selecionado o modelo com 4 fatores, uma vez que este possui menor BIC, maior ajuste e maior explicação da variância (apesar de valores menores de TLI e RMSEA, que foram relativamente próximos dos valores do modelo com 3 fatores).

Assim, o modelo selecionado apresenta a seguinte estrutura:

Figura 4.2.3: Diagrama do Modelo com Variáveis Selecionadas



É possível perceber que em todos os fatores as cargas foram superiores a 0,4 (CG no Fator 3). Ademais, a variável aLI não apresentou correlação significante com nenhum fator e foi, portanto, removida no processo de criação do modelo de equações estruturais da análise fatorial confirmatória.

4.2.2 Análise Fatorial Confirmatória

Em seguida, foi realizada a análise fatorial confirmatória de acordo com as correlações indicadas na Figura 4.2.3:

Tabela 7: Medidas de Ajuste para o Modelo de Equações Estruturais

BIC	**RMSEA**	**TLI**	**CFI**
24384,01	0,335	0,497	0,554

Nota-se que após a adoção de equações estruturais, o modelo obtido apresentou melhores medidas de ajuste. Para entender como as variáveis se relacionam com cada fator, seguem as cargas fatoriais obtidas:

Tabela 8: Cargas Fatoriais do Modelo

Variável	**Fator 1**	**Fator 2**	**Fator 3**	**Fator 4**
PtchW	1,00	0	0	0
PtchR	1,00	0	0	0
RAA	1,05	0	0	0
162WL%	1,05	0	0	0
REW	0,96	0	0	0
WAR	1,02	0	0	0
RAR	1,02	0	0	0
RE24/boLI	0,91	0	0	0
BF	0	1,00	0	0
GS	0	0,89	0	0
Opp	0	0,94	0	0
RA9role	0	0,69	0	0
DP	0	0,73	0	0
W	0	0,75	0	0
CG	0	0,39	0	0
HR9	0	0	1,00	0
FIP	0	0	1,10	0
XBH%	0	0	0,53	0
H9	0	0	0,63	0
WHIP	0	0	0	1,00
BB9	0	0	0	0,81
ERA	0	0	0	1,00

Após a análise das variáveis de cada fator, prosseguiu-se com a nomeação dos fatores. No Fator 1, encontram-se majoritariamente variáveis que comparem o arremessador com outro arremessador a nível de substituição ou nível médio. Assim, optou-se por nomear esse fator como "Estatísticas de Comparação". No Fator 2, concentraram-se estatísticas de teor mais descritivo e menos controladas pelo jogador, como número de rebatedores enfrentados, jogos completos, jogos iniciados, queimadas duplas, etc. Portanto, esse fator recebeu o nome de "Estatísticas Descritivas".

O Fator 3, por sua vez, mostrou correlações com variáveis indicadoras de contato do oponente com a bola (quantidade de *homeruns*, rebatidas, etc) e, assim, recebeu o nome de "Estatísticas de Contato". Por fim, o Fator 4 concentrou-se em variáveis que medem a quantidade de andadas, bases ou rebatidas que foram concedidas ao oponente, sendo, assim, chamado de "Estatísticas Concedidas".

Tabela 9: Pesos dos Fatores na Habilidade Final

Fator	**Nome**	**Peso**
Fator 1	Estatísticas de Comparação	1,000
Fator 2	Estatísticas Descritivas	0,860
Fator 3	Estatísticas de Contato	-2,015
Fator 4	Estatísticas Concedidas	-2,458

Nota-se que as Estatísticas Concedidas apresentaram o maior peso absoluto na habilidade final, com peso negativo, enquanto o fator de menor peso na habilidade foram as Estatísticas Descritivas. Conforme esperado, os dois últimos fatores apresentam cargas negativas, uma vez que é esperado de um bom arremessador valores menores para as estatísticas que compõem ambos os fatores.

Assim, o cálculo do escore de habilidade pode ser feito por:

$$\text{Habilidade} = \text{Fator } 1 + 0,860 * \text{Fator } 2 - 2,015 * \text{Fator } 3 - 2,458 * \text{Fator } 4 \qquad (4.2.1)$$

Por fim, foi realizada a análise de correlação entre a habilidade obtida e a variável WPA (Probabilidade de Vitória Adicionada pelo Arremessador), a fim de verificar se o escore calculado é um bom indicador, de fato, da habilidade do jogador.

Tabela 10: Teste de Correlação de Spearman entre WPA e Habilidade

Coeficiente de Correlação	**Estatística do Teste**	**P-Valor**	**Decisão**
0,693	18335797	$< 0,001$	Rejeita H_0

Nota-se que o escore de habilidade apresentou uma correlação relevante com a variável WPA, com uma correlação positiva moderada.

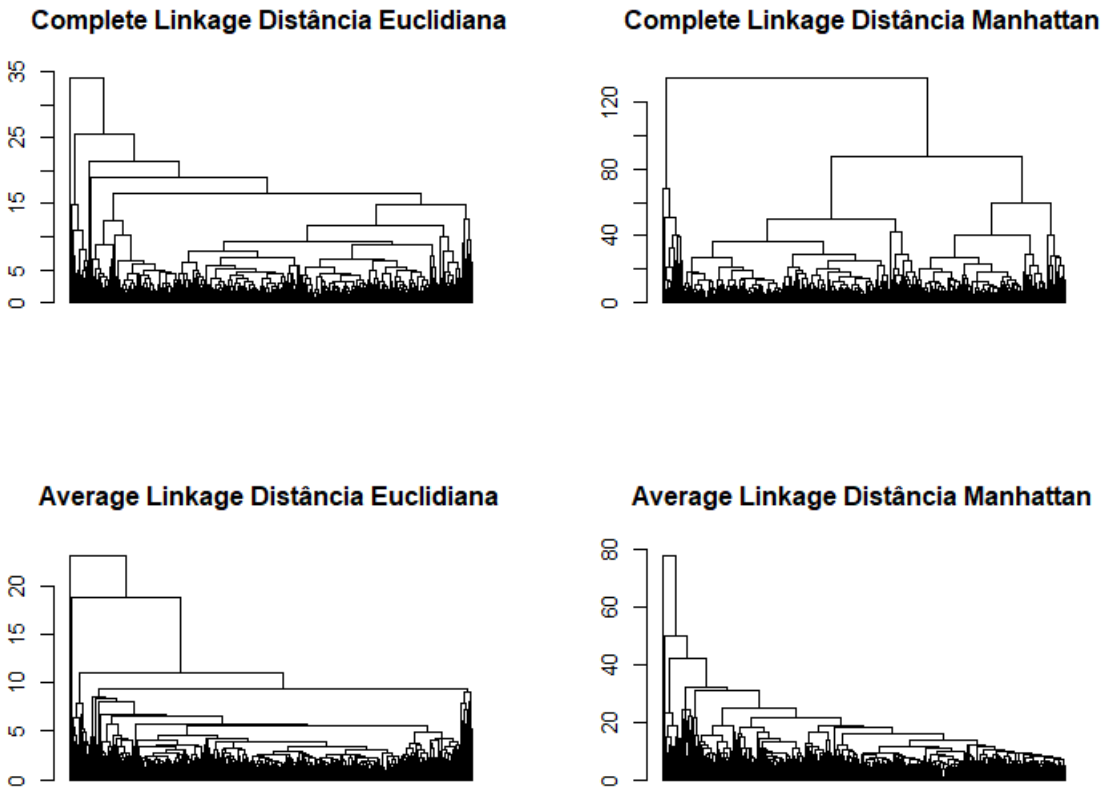
4.3 Análise de Agrupamentos

Para o processo de agrupamento, foram utilizadas apenas as 23 variáveis selecionadas no processo inicial, uma vez que esta combinação forneceu o melhor agrupamento dentre os testados. Após a realização de métodos hierárquicos para verificar o número ideal de *clusters*, o agrupamento final utilizado foi o proposta pelo método não-hierárquico *K-Means*.

4.3.1 Métodos Hierárquicos

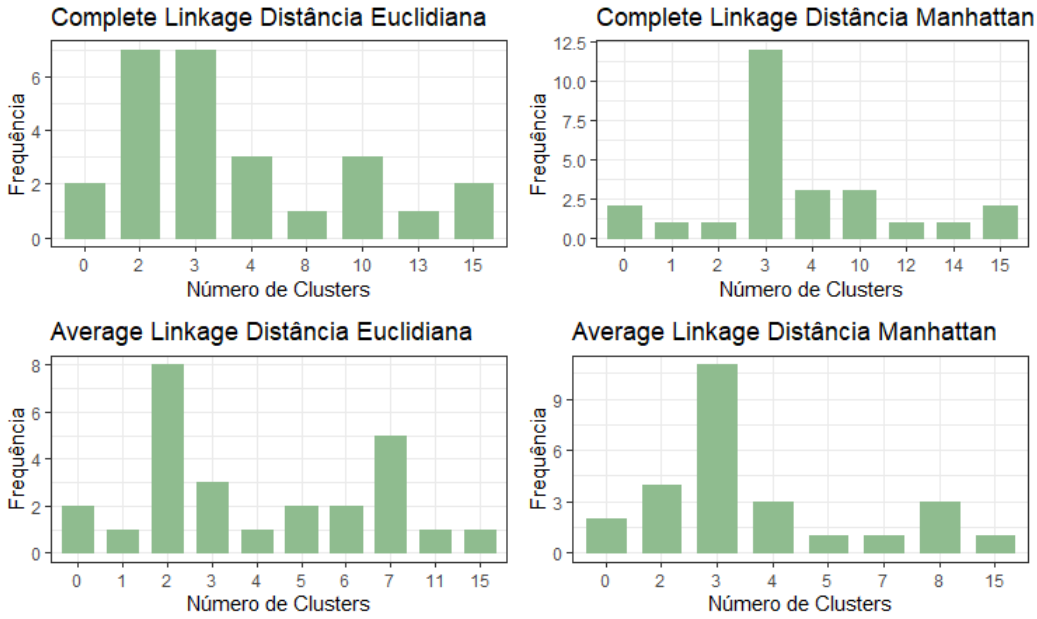
Para iniciar o processo de agrupamento, foram utilizadas as 23 variáveis padronizadas com base em suas distâncias Euclidiana e de Manhattan, por meio dos métodos *Complete* e *Average Linkage*, cujos dendrogramas seguem abaixo:

Figura 4.3.1: Dendrogramas das Variáveis Selecionadas



Nota-se a aparente formação de 2 *clusters* nos dendrogramas analisados. Entretanto, a fim de obter-se o melhor agrupamento possível, foram analisados os índices propostos pelo pacote *NbClust*, cujos resultados seguem no gráfico abaixo:

Figura 4.3.2: Número de *Clusters* Ideal

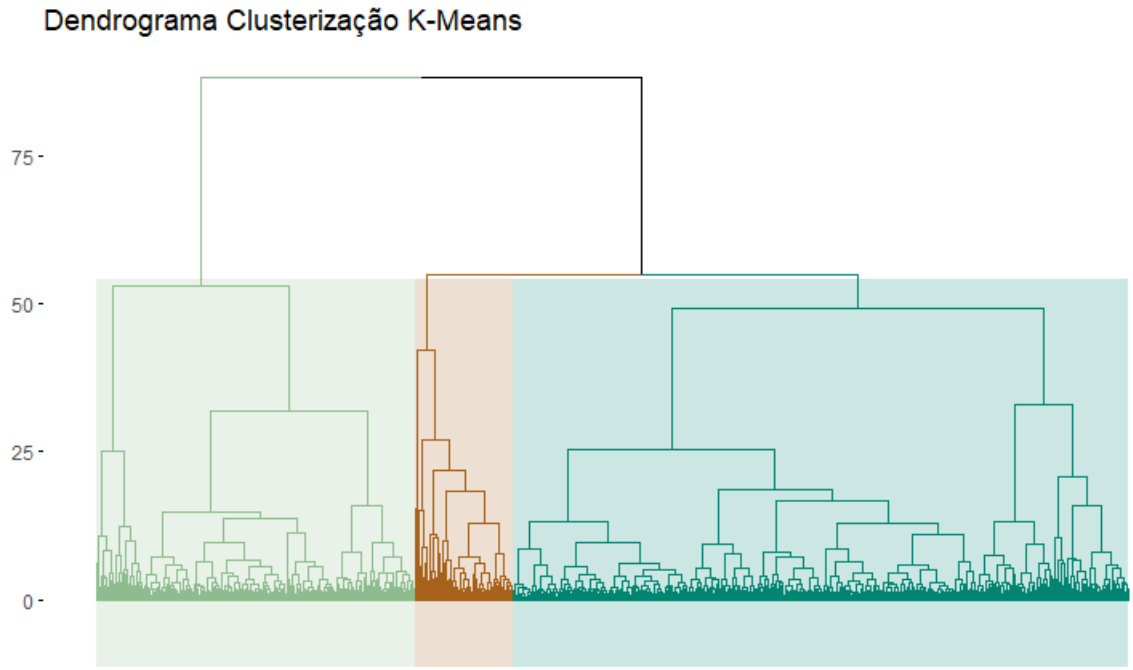


Conforme a maioria dos índices indica, o número ideal de *clusters* a ser adotado será de 3 *clusters* para o processo de agrupamento não-hierárquico *K-Means*.

4.3.2 Métodos Não Hierárquicos

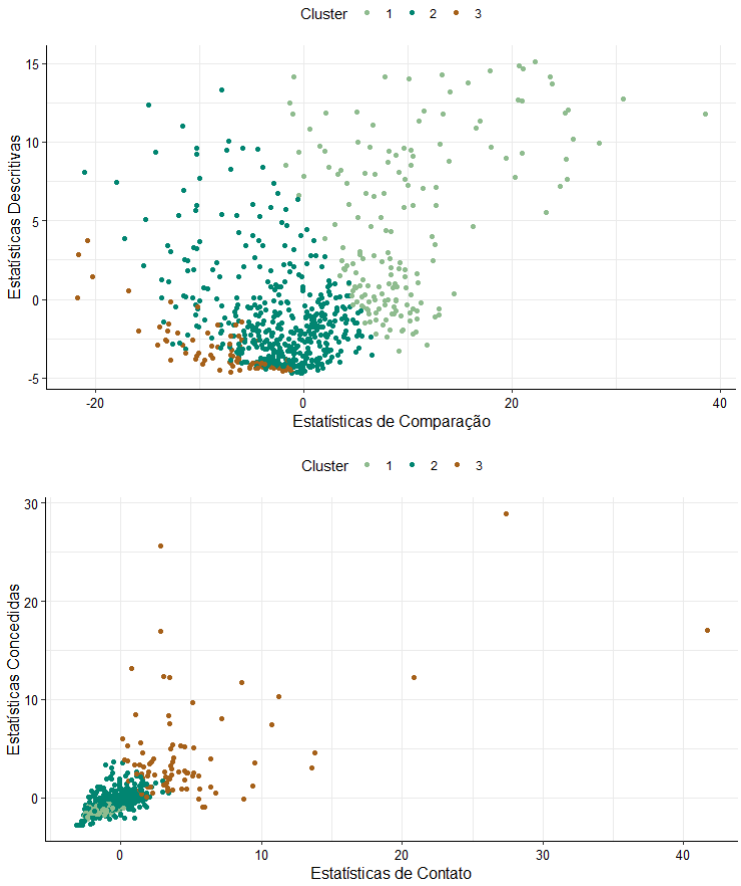
Após definir a quantidade de *clusters* a serem adotados, o dendrograma do agrupamento final segue:

Figura 4.3.3: Dendrograma do Agrupamento *K-Means*



Nota-se a formação de um *cluster* com menos observações (na cor marrom), o qual se encontra pouco distante do segundo grupo (na cor verde escuro). Ambos os grupos apresentaram uma distância maior em relação ao último *cluster*, representado pela cor verde claro. O grupo verde claro (*Cluster 1*) apresentou 174 jogadores, enquanto o grupo verde escuro (*Cluster 2*) possui 453 jogadores e o grupo marrom (*Cluster 3*) apresentou 84 jogadores.

Figura 4.3.4: *Biplots* dos *Clusters*



Ao observar-se a disposição dos grupos de jogadores de acordo com os fatores criados, nota-se que, no geral, arremessadores do grupo 1 tendem a ter Estatísticas de Comparação e Descritivas (Fatores 1 e 2, respectivamente) maiores, enquanto o grupo 3 possui os menores valores e o grupo 2 apresenta valores intermediários.

Em relação às Estatísticas de Contato e Concedidas (Fatores 3 e 4), a situação se inverte: o grupo 3 apresenta os maiores valores, seguido do grupo 2 e, por fim, do grupo 1.

Assim, fica evidente que o *Cluster 1* representa os melhores arremessadores, enquanto o *Cluster 2* abarca os jogadores medianos e, por fim, o *Cluster 3* envolve os piores arremessadores da liga.

4.4 Análises Finais

Após a obtenção do número ideal de *clusters*, serão realizadas algumas análises a fim de compreender o comportamento de cada grupo de jogadores, tal como suas habili-

dades esperadas:

Figura 4.4.1: *Boxplot* da Habilidade por *Cluster*

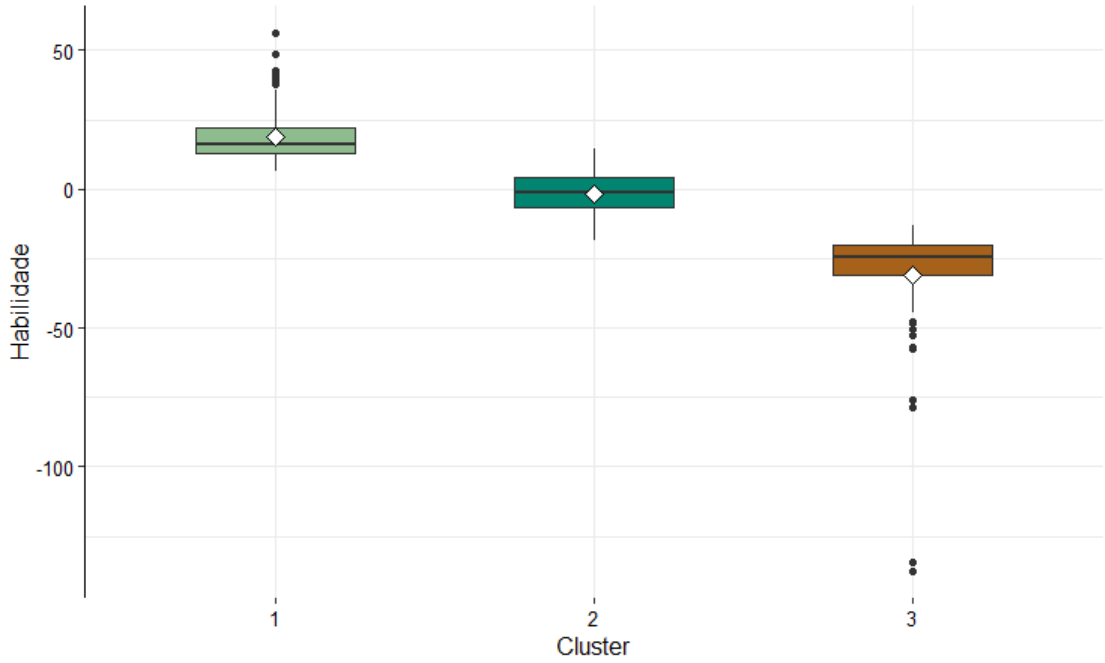


Tabela 11: Teste de Dunn para Escore de Habilidade entre *Clusters*

Comparação	**P-Valor**	**Decisão**
1 e 2	$< 0,001$	Rejeita H_0
1 e 3	$< 0,001$	Rejeita H_0
2 e 3	$< 0,001$	Rejeita H_0

Percebe-se que existe, de fato, diferença significativa de habilidade entre todos os grupos, com o *Cluster 1* sendo o grupo de maior habilidade e o *Cluster 3* sendo o de menor.

Figura 4.4.2: *Boxplot* dos Fatores por *Cluster*

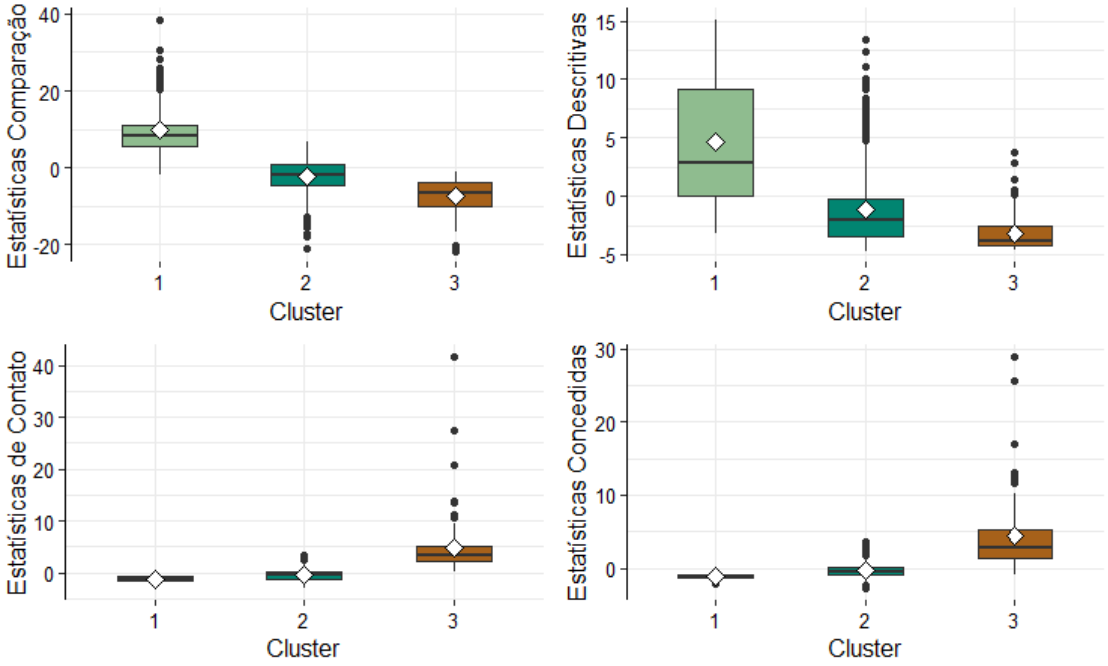
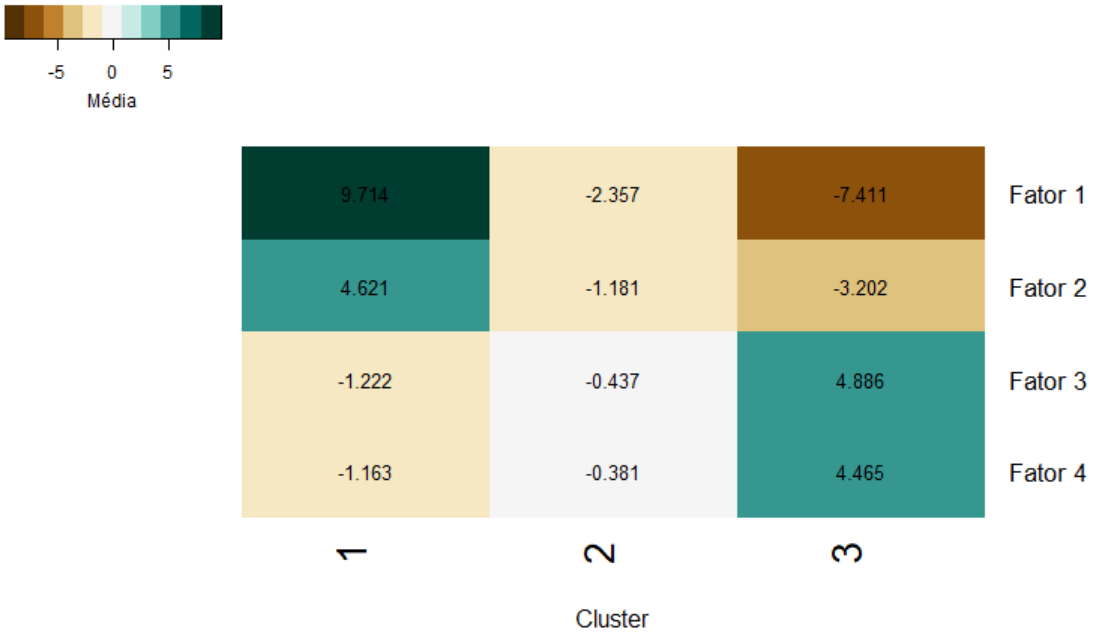


Tabela 12: Teste de Dunn para Fatores entre *Clusters*

Comparação	**Fator 1**	**Fator 2**	**Fator 3**	**Fator 4**	**Decisão**
1 e 2	$< 0,001$	$< 0,001$	$< 0,001$	$< 0,001$	Rejeita H_0
1 e 3	$< 0,001$	$< 0,001$	$< 0,001$	$< 0,001$	Rejeita H_0
2 e 3	$< 0,001$	$< 0,001$	$< 0,001$	$< 0,001$	Rejeita H_0

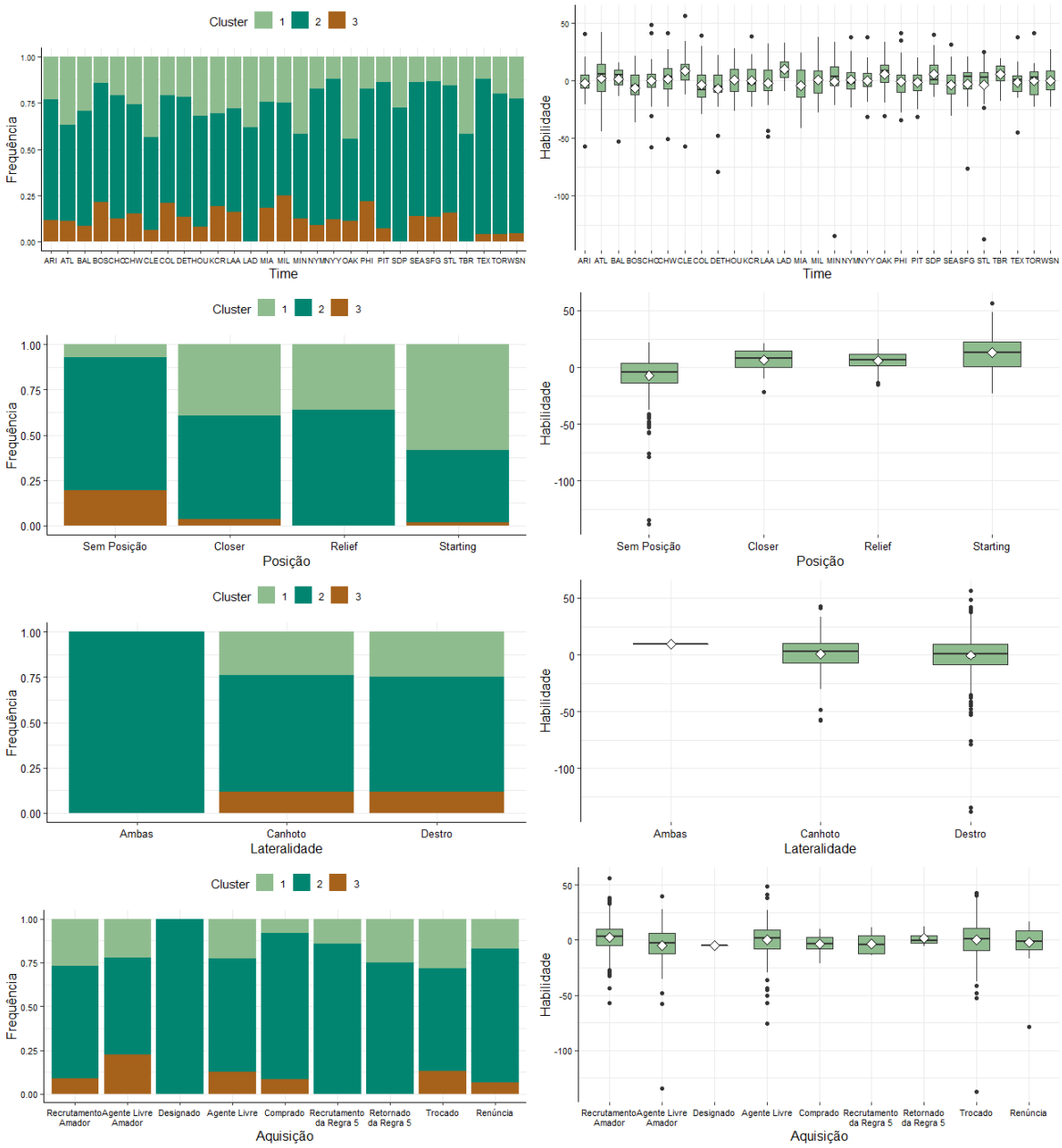
Em relação aos fatores por cada *cluster*, nota-se que, conforme esperado, os grupos de mais habilidade apresentam valores maiores para as Estatísticas de Comparação e Descritivas e, também, valores menores para as Estatísticas de Contato e Concedidas. Ademais, essa diferença é significativa para todos os fatores entre todos os *clusters*.

Figura 4.4.3: Mapa de Calor das Médias dos Fatores por *Cluster*



Percebe-se que o Grupo 1 tem médias positivas para Estatísticas de Comparação e Descritivas, enquanto os Grupos 2 e 3 apresentam valores negativos para essas estatísticas. Para as Estatísticas de Contato e Concedidas, os Grupos 1 e 2 apresentam valores negativos e relativamente próximos entre si, enquanto o Grupo 3 apresenta valores positivos.

Figura 4.4.4: Gráficos de Habilidade e *Clusters* por Time, Posição, Lateralidade e Aquisição



Analisando-se a distribuição da habilidade e dos *clusters* pelas variáveis qualitativas, nota-se que, no geral, os times apresentam proporções semelhantes de arremessadores de cada um dos três grupos, com exceção dos times LAD (*Los Angeles Dodgers*), SDP (*San Diego Padres*) e TBR (*Tampa Bay Rays*), que não possuem nenhum jogador classificado no *Cluster 3*. No geral, o time de maior habilidade é o LAD (time campeão da temporada de 2020), seguido por CLE (*Cleveland Guardians*) e OAK (*Oakland Athletics*), enquanto os times com menor habilidade são DET (*Detroit Tigers*), BOS (*Boston Red Sox*) e MIA (*Miami Marlins*).

Em relação à posição do arremessador, em geral, os Arremessadores Secundários

(*Relief*) não estão classificados no *Cluster 3*, tendo os arremessadores sem posição a maior proporção de jogadores deste *cluster* e a posição Inicial (*Starting*) a posição com maior proporção de jogadores do *cluster* de maior habilidade. Assim, nota-se que, no geral, arremessadores Iniciais possuem a maior habilidade, seguidos dos arremessadores Finais (Closer), arremessadores Secundários e, por fim, os jogadores sem posição fixa.

No que se refere à lateralidade do jogador, nota-se que os jogadores que arremessam com ambas as mãos são classificados no *cluster* de habilidade mediana, enquanto os canhotos e destros apresentam proporções semelhantes de jogadores de cada *cluster*. Apesar da pequena diferença devido à escala do gráfico, os jogadores que arremessam com ambas as mãos possuem habilidade média (9,62) bem superior à habilidade média dos jogadores canhotos (0,78) e destros (-0,30).

Por fim, em relação à forma de aquisição do jogador, nota-se que os jogadores Designados fazem parte apenas do *Cluster 2*. Além disso, jogadores adquiridos por Recrutamento da Regra 5 ou Retornados da Regra 5 não apresentaram nenhum jogador do *Cluster 3*. Ademais, a forma de aquisição com maior proporção de jogadores do *cluster* de menor habilidade foi Agente Livre Amador, enquanto a forma com maior proporção de jogadores mais habilidosos foi Troca. Sendo assim, a forma de aquisição com maior média de habilidade foi Recrutamento Amador, seguido por Retornados da Regra 5 e Troca e as menores médias de habilidade foram Agente Livre Amador, Designado e Recrutamento da Regra 5, respectivamente.

Tabela 13: Top 10 Melhores Jogadores

Jogador	**Habilidade**	**Time**	**Posição**	**Lateralidade**	**Aquisição**
Shane Bieber	56.32	CLE	SP	Destro	Recrutamento Amador
Yu Darvish	48.66	CHC	SP	Destro	Agente Livre
Max Fried	42.57	ATL	SP	Canhoto	Trocado
Kyle Hendricks	41.71	CHC	SP	Destro	Trocado
Hyun Jin Ryu	41.34	TOR	SP	Canhoto	Agente Livre
Zack Wheeler	41.30	PHI	SP	Destro	Agente Livre
Dallas Keuchel	41.29	CHW	SP	Canhoto	Agente Livre
Zac Gallen	40.67	ARI	SP	Destro	Trocado
Dinelson Lamet	40.07	SDP	SP	Destro	Agente Livre Amador
Antonio Senzatela	39.50	COL	SP	Destro	Agente Livre Amador

Analisando-se os 10 melhores arremessadores da liga (todos do *Cluster 1*), se-

gundo a habilidade calculada, nota-se que dois deles são do time CHC (*Chicago Cubs*), todos arremessadores iniciais, 70% são destros, com apenas 30% canhotos e todos foram adquiridos por Agente Livre, Agente Livre Amador, Troca ou Recrutamento Amador. Ademais, a título de curiosidade, o jogador de maior habilidade Shane Bieber foi o jogador que recebeu o prêmio Cy Young de melhor arremessador da liga americana no ano de análise.

5 Conclusão

Após a análise de correlações e remoção de variáveis constituídas de combinações lineares de outras, realizou-se uma primeira análise fatorial, a qual retornou um padrão de 23 variáveis consideradas mais importantes para o estudo: RAR, WAR, REW, PtchR, PtchW, RAA, 162WL%, RE24/boLI, W, ERA, BF, WHIP, H9, Opp, DP, GS, FIP, XBH%, HR9, BB9, RA9role, CG e aLI. A partir deste ponto, foi obtido um modelo com 4 fatores, por meio do método dos mínimos resíduos e com rotação de fatores *Varimax*, cujas cargas fatoriais estão especificadas na Tabela 8. Dessa forma, pode-se obter um escore final de habilidade do arremessador, utilizando-se dos 4 fatores calculados e atribuindo seus devidos pesos, conforme a Equação 4.2.1.

Assim, nota-se que para o Fator 1 (Estatísticas de Comparação), as variáveis de maior peso são RAA e 162WL%, enquanto a de menor peso é RE24/boLI. Para o Fator 2 (Estatísticas Descritivas), o maior peso é da variável BF, enquanto o menor é da variável DP. Já para o Fator 3 (Estatísticas de Contato), o maior peso vem da variável FIP, enquanto o menor é da variável XBH%. E, por fim, em relação ao Fator 4 (Estatísticas Concedidas), as variáveis WHIP e ERA apresentam o mesmo peso, com BB9 apresentando o menor dos pesos. Em relação aos pesos de cada fator na habilidade final, as Estatísticas Concedidas apresentam maior peso absoluto (seu peso foi negativo), enquanto Estatísticas Descritivas apresentaram o menor peso (com valor positivo).

Em seguida, depois do procedimento de obtenção do escore de habilidade, utilizou-se das 23 variáveis selecionadas (em suas formas padronizadas) e realizou-se os agrupamentos hierárquicos *Average* e *Complete Linkage*, com as distâncias Euclidiana e de Manhattan, as quais resultaram na formação de 3 *clusters*. Em seguida, o agrupamento final foi obtido por meio do método não hierárquico *K-Means*. Assim, obtiveram-se 3 *clusters* de jogadores, onde o *Cluster 1* representa os melhores arremessadores, o *Cluster 2* representa jogadores medianos e, por fim, o *Cluster 3* abarca os piores arremessadores da liga.

Por fim, foram realizadas análises descritivas finais com os fatores, o escore de habilidade e os grupos obtidos, a fim de compreender padrões dentre os jogadores. Assim, concluiu-se que, no geral, jogadores de alta habilidade apresentaram valores mais altos de Estatísticas de Comparação e Descritivas e valores mais baixos para Estatísticas de Contato e Estatísticas Concedidas. Ademais, em geral, a maioria dos jogadores de maior habilidade jogam na posição de Arremessador Inicial (*Starting Pitcher*), arremessam com ambas as mãos e foram adquiridos por Recrutamento Amador.

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.

BYUN, K.-W. A study on the viewing experience and performance of professional baseball team: The team performance side and the fan performance side. *International Journal of Advanced Smart Convergence*, v. 10, n. 1, 2021.

CASTROVINCE, A. *A Fan's Guide to Baseball Analytics: Why WAR, WHIP, wOBA, and Other Advanced Sabermetrics Are Essential to Understanding Modern Baseball*. [S.l.]: Sports Publishing LLC, 2020.

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.

CHOU, E. et al. Extreme-k categorical samples problem. 2020.

DUNN, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, v. 56, 1961.

FIELLER, E. C.; HARTLEY, H. O.; PEARSON, E. S. Tests for rank correlation coefficients. i. *Biometrika*, v. 44, 1957.

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.

HIRSCHFELD, G.; BRACHEL, R. von. Improving multiple-group confirmatory factor analysis in r – a tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, v. 19, n. 7, 2014.

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.

KASSAMBARA, A. *Hierarchical Clustering in R: The Essencials*. [S.l.], 2018. Disponível em: ⟨https://www.datanovia.com/en/lessons/examples-of-dendrograms-visualization/⟩. Acesso em: 24 set. 2021.

LAW, K. *Smart Baseball: The Story Behind the Old Stats That Are Ruining the Game, the New Ones That Are Running It, and the Right Way to Think About Baseball*. [S.l.]: William Morrow, 2017.

LINDBERGH, B.; SAWCHIK, T. *The MVP Machine: How Baseball's New Nonconformists Are Using Data to Build Better Players*. [S.l.]: Basic Books, 2019.

MANGALE, S. *Scree Plot*. [S.l.], 2020. Disponível em: ⟨https://sanchitamangale12.medium.com/scree-plot-733ed72c8608⟩. Acesso em: 13 mai. 2022.

MARCOU, C. Investigating major league baseball pitchers and quality of contact through cluster analysis. *Honors Projects*, 2020.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. *Multivariate Analysis*. [S.l.]: Academic Press, 1980.

MLB. *Standard Stats*. [S.l.], 2021. Disponível em: ⟨https://www.mlb.com/glossary/⟩. Acesso em: 10 ago. 2021.

MORAN, C.; COROIU, A.; KöRNER, A. Psychosocial distress in patients with cutaneous melanoma: validation of the skin cancer index (sci). *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*, v. 29, 2021.

OLDACH, M. *10 Tips for Choosing the Optimal Number of Clusters*. [S.l.], 2019. Disponível em: ⟨https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92⟩. Acesso em: 13 mai. 2022.

RAJPUT, P. *Exploratory Factor Analysis*. [S.l.], 2018. Disponível em: ⟨https://rpubs.com/Pun_/Exploratory_factor_Analysis⟩. Acesso em: 13 mai. 2022.

RIGDON, E. E. Cfi versus rmsea: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, Routledge, v. 3, n. 4, p. 369–379, 1996. Disponível em: ⟨https://doi.org/10.1080/10705519609540052⟩.

SCHWARZ, A. *A numbers revolution*. [S.l.], 2004. Disponível em: ⟨https://www.espn.com/mlb/columns/story?columnist=schwarz_alan&id=1835745⟩. Acesso em: 4 ago. 2021.

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.