UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

Gianpedro Robertto Mella Brigante

ANÁLISE MULTIVARIADA APLICADA NA CONSTRUÇÃO DE SCORES DE RENDIMENTO DOS PRINCIPAIS JOGADORES DO FUTEBOL MUNDIAL

Presidente Prudente

2021/2022

Gianpedro Robertto Mella Brigante

ANÁLISE MULTIVARIADA APLICADA NA CONSTRUÇÃO DE SCORES DE RENDIMENTO DOS PRINCIPAIS JOGADORES DO FUTEBOL MUNDIAL

Relatório final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientadora:Profa. Dra. Miriam Rodrigues Silvestre

Presidente Prudente

2021/2022

B854a

Brigante, Gianpedro Robertto Mella

Análise multivariada aplicada na construção de scores de rendimento dos principais jogadores do futebol mundial / Gianpedro Robertto Mella Brigante. -- Presidente Prudente, 2022

60 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientadora: Miriam Rodrigues Silvestre

1. Estatística. 2. Análise multivariada. 3. Análise de agrupamentos. 4. Análise de componentes principais. 5. Rendimento esportivo. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Gianpedro Robertto Mella Brigante

ANÁLISE MULTIVARIADA APLICADA NA CONSTRUÇÃO DE SCORES DE RENDIMENTO DOS PRINCIPAIS JOGADORES DO FUTEBOL MUNDIAL

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

Orientador: _____

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

_____

Prof. Dr. Edilson Ferreira Flores
Departamento de Estatística

_____

Profa. Dra. Silvely Nogueira de Almeida Salomão Néia
Departamento de Estatística

Presidente Prudente, 17 de março de 2022.

Agradecimentos

Gostaria de agradecer a minha família que sempre me apoiou em minhas decisões e me ajudou nos momentos difíceis. Também gostaria de fazer uma homenagem ao Prof. Dr. Luiz Carlos Benini pois no começo dessa jornada era meu orientador. Sou muito grato a Profa. Dra. Miriam Silvestre que mesmo na orientação de muitos alunos conseguiu dar prioridade para a minha pessoa e ajudar nessa caminhada, e por ultimo queria registrar um forte abraço para a Profa. Dra. Vera Tomazella que está me ajudando muito nesses períodos de responsabilidades.

Resumo

Neste trabalho foram adotados dois métodos de análise multivariada, a Análise de Componentes Principais e a Análise de Agrupamentos, com o intúito de analisar o rendimento, tendo em consideração variáveis relacionadas ao ataque dos principais atletas dos cinco maiores campeonatos nacionais, sendo eles: o Campeonato Brasileirão Série A, La Liga, Série A Italia, Premier League e France Ligue.

O método de Análise de Componentes Principais foi usado para diminuir a quantidade de variáveis e simplificar a interpretação dos jogadores, além de proporcionar os scores de rendimento de cada atleta analisado. Esta aplicação foi muito efetiva pois conseguiu extrair cerca de 84% da informação de oito variaveis correlacionadas em duas novas variaveis não correlacionadas. Com esse modelo também foi possível fazer gráficos Biplot que ajudaram a identificar os jogadores que mais se destacaram em cada variavel devido aos scores obtidos. Também foi analisado o desempenho dos jogadores por campeonato possibilitando a comparação dessas competições estudadas.

Após a obtenção dos scores de rendimento foi usado um método de agrupamento denominado Método de Ward, que agrupa os indivíduos (atletas) conforme suas proximidades de acordo com os dados, depois a qualidade desses grupos foram observadas pelo gráfico da silhueta que possibilita ver se o jogador está bem alocado em seu grupo. Fazendo o agrupamento levando em conta somente os scores das duas componentes escolhidas percebeu-se que não foi possível encontrar uma forte estrutura dos grupos, mas os grupos foram condizentes às interpretações obtidas nos gráficos Biplot.

Palavras-chave: *Futebol, Análise Multivariada, Score, Rendimento, Componentes Principais, Método de Ward.*

Abstract

In this work, two methods of multivariate analysis were adopted, Principal Component Analysis and Cluster Analysis, with the aim of analyzing the performance, taking into account variables related to the attack of the main athletes of the five biggest national championships, namely: the Brasileirão Serie A, La Liga, Serie A Italia, Premier League and France Ligue.

The Principal Component Analysis method was used to reduce the number of variables and simplify the players' interpretation, in addition to providing the performance scores of each analyzed athlete. This application was very effective as it managed to extract about 84% of the information from eight correlated variables into two new uncorrelated variables. With this model it was also possible to make Biplot graphs that helped to identify the players who stood out the most in each variable due to the scores obtained. The performance of players by championship was also analyzed, allowing the comparison of these studied competitions.

After obtaining the performance scores, a grouping method called Ward's Method was used, which groups the individuals (athletes) according to their proximity according to the data, then the quality of these groups was observed by the silhouette graph that makes it possible to see if the player is well placed in his group. Making the grouping taking into account only the scores of the two components chosen, it was noticed that it was not possible to find a strong structure of the groups, but the groups were consistent with the interpretations obtained in the Biplot graphs.

Keywords: *Soccer, Multivariate Analysis, Score, Performance, Principal Components, Ward's Method.*

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

O futebol é um esporte muito popular que atrai qualquer tipo de pessoa e por conta disso, muitos dedicam suas vidas para se beneficiarem desse mundo esportivo. Seu investimento é de grande valor em todo mundo, gerando uma grande quantidade de empregos como jogadores, empresários, entre outros (LEONCINI, 2001). Além disso, o futebol também é um meio de inclusão social, pois é notável a presença de integrantes de várias etnias e classes sociais fazendo parte desses eventos.

A modalidade esportiva vem evoluindo ao decorrer do tempo e hoje possui características como grande quantidade de táticas de jogo e muita exigência física, técnica e psicológica (FERNANDES, 1994). Os dados também contribuíram para o progresso do esporte disponibilizando interpretações científicas para os clubes, trazendo inovação aos treinamentos e nas análises de rendimento dos atletas (VENDITE et al., 2003).

De acordo com Godik (1996) os primeiros registros de análises das ações individuais técnico-tática foram apresentados em 1936, onde era proposto que em cada jogo fosse necessário anotar a quantidade de passes e outras táticas de jogo referente à efetividade da defesa e do ataque de cada jogador. A análise da qualidade das atuações dos jogos é multiforme, com parâmetros registrados e uma das formas de registros mais utilizadas é chamada de scout que possibilita mensurar o desempenho coletivo e individual.

Portanto, devido a todos esses avanços é notável o aumento do nível de competitividade dos torneios de futebol destacando a necessidade de se ter cada vez mais maneiras de medir o rendimento dos atletas. Logo neste trabalho serão apresentadas técnicas de análise multivariada, mais especificamente, Análise de componentes principais e Análises de agrupamento, dos scouts e algumas outras variáveis relacionadas ao ataque coletadas dos principais jogadores, segundo a imprensa mundial, das cinco maiores ligas de futebol, a fim de medir o desempenho de cada jogador e comparar os resultados de cada liga.

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo principal deste trabalho é aplicar através de um banco de dados, relacionado ao futebol, as técnicas de análise multivariada, denominadas análise de componentes principais e análise de agrupamentos pelo método de Ward, a fim de medir e comparar o rendimento dos atletas.

1.1.2 Objetivos Específicos

1. Estudar as técnicas de componentes principais e método de agrupamento de Ward em suas aplicabilidades;

2. Entender como funciona o rendimento dos atletas;

3. Aplicar a metodologia existente para dados reais voltado ao desempenho dos jogadores de futebol;

4. Obter uma classificação de jogadores de acordo com o desempenho, através da aplicação das técnicas multivariadas;

5. Comparar os resultados para diferentes ligas.

1.2 Justificativa

O futebol é um esporte muito rentável para seus patrocinadores e para as emissoras de televisão, sendo alvo de muito investimento por ambos os lados, esse fato fez com que a modalidade ficasse cada vez mais acessível possibilitando qualquer torcedor ter informações sobre seu time ou campeonato (VENDITE et al., 2005). Com isso, plataformas de apostas online vem desafiando fanáticos pelo esporte propondo diferentes tipos de apostas como por exemplo arriscando qual jogador vai ter um melhor rendimento na partida, e saber analisar dados extraídos desses jogadores pode ser útil.

O foco é dado para as variáveis ofensivas, ou seja, variáveis que comprovam o rendimento de cada jogador no ataque. Pois os jogadores ofensivos são de maior importância

por serem mais caros, responsáveis pela vitória do time, e terem maiores chances de ganhar premiações importantes como melhor do campeonato trazendo mais visibilidade e lucro para o clube.

Além de ser valioso para os torcedores, saber interpretar dados no esporte também é essencial para a comissão técnica dos clubes, pois serve de auxílio na hora do treinamento e para montar as táticas de cada jogo (VENDITE et al., 2003).

É sabido que análise de desempenho é muito importante para o esporte e para uma parte da população, portanto esse trabalho será realizado para contribuir com o desenvolvimento e evolução do esporte.

2 Referencial Teórico

2.1 Análise Multivariada

Quando acontece um evento, seja ele natural ou social, pode-se dizer que diversas variáveis possuem influência sobre este. Portanto, para uma boa compreensão de um acontecimento é necessário ter a pretensão de conhecer a realidade e de interpretar os fenômenos baseados no conhecimento das variáveis importantes. Para isso, é necessário controlar, manipular e medir as variáveis que são consideradas relevantes ao entendimento do fenômeno analisado (VICINI, 2005).

Sendo assim, os objetivos gerais, para os quais a análise multivariada contribui são:

- Redução de dados ou simplificação estrutural, sem perda de informações valiosas e facilitando as interpretações;
- Ordenação e agrupamento. Os agrupamentos de objetos ou variáveis similares, baseados em dados;
- Procura de dependência entre variáveis;
- Predição e construção de hipóteses.

A análise multivariada é uma grande área, na qual até os estatísticos mais experientes seguem com cuidado, devido a estes métodos serem um ramo recente da ciência. Já se descobriu muito sobre esta técnica estatística, mas muito ainda está para ser descoberto (MAGNUSSON; MOURÃO, 2003).

Os métodos multivariados são escolhidos de acordo com os objetivos da pesquisa, pois sabe-se que a análise multivariada é uma análise exploratória de dados. Resumidamente o estudo multivariado tem como propósito interpretar o novo conjunto de variáveis e conseguir traduzir as informações que estão sendo reveladas, que até então não eram percebidas por estarem em um espaço dimensional maior do que três.

2.2 Análise de Componentes Principais

A Análise de Componentes Principais é uma técnica matemática da análise multivariada, que gera interpretações de um grande número de dados disponíveis. Permite, também, a identificação das medidas responsáveis pelas maiores variações entre os resultados, sem muitas perdas de informações. Além de transformar um conjunto original de variáveis em outro conjunto que são os componentes principais (CP) de dimensões equivalentes (HONGYU et al., 2016).

A técnica multivariada é uma modelagem da estrutura de covariância que foi desenvolvida por Pearson(1901, apud HONGYU et al., 2016) e aprimorada por Hotelling(1933, 1936, apud HONGYU et al., 2016) que usou com o propósito determinado de analisar as estruturas de correlação. A técnica estatística transforma, linearmente, um conjunto original de variáveis correlacionadas entre si em um conjunto consideravelmente menor de variáveis não correlacionadas que contém a maior parte da informação do conjunto original. Essa transformação também permite eliminar algumas variáveis originais que contribui com pouca informação.

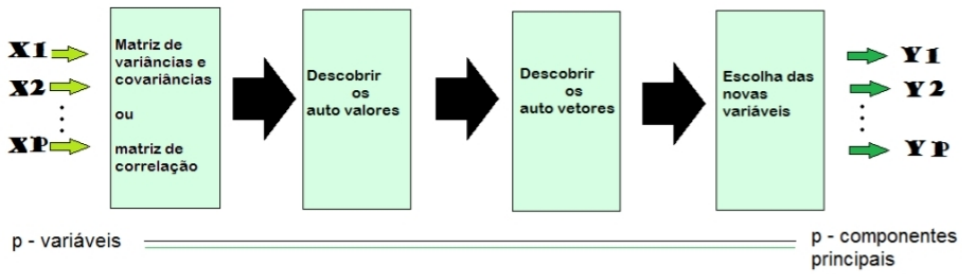
Primordialmente, o objetivo da ACP foi o de encontrar linhas e planos que melhor se moldassem a um conjunto de pontos em um espaço p-dimensional (PEARSON, 1901 apud VICINI, 2005). O objetivo da análise de componentes principais é o de explicar a estrutura da variância e covariância de um vetor aleatório, composto de p-variáveis aleatórias, através de combinações lineares das variáveis iniciais. Essas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si (SANDANIELO, 2008).

Muitos pesquisadores têm utilizado a análise de componentes principais para resolver problemas como da multicolinearidade em regressão linear, para realizar a modelagem da interação entre fatores em experimentos sem repetição e até mesmo para análise de

rendimento.

Mas esse modelo tem algumas desvantagens que são: a sensibilidade a outliers, não recomendada quando se tem muitos zeros na matriz e dados ausentes. A ACP também não é recomendada quando se obtém mais variáveis do que unidades amostrais. Ao reduzir o número de variáveis, há perda da informação de variabilidade das variáveis originais. Na prática, o algoritmo baseia-se na matriz de variância-covariância, ou na matriz de correlação, de onde são extraídos os autovalores e os autovetores. Para a determinação das componentes principais, é necessário calcular a matriz de variância-covariância (Σ), ou a matriz de correlação (R), encontrar os autovalores e os autovetores e, por fim, escrever as combinações lineares, que serão as novas variáveis, denominadas de componentes principais, sendo que cada componente principal é uma combinação linear de todas as variáveis originais (HONGYU et al., 2016). A figura a seguir explica os passos da aplicação da ACP.

Figura 1 – Passos para seguir na aplicação da ACP



Fonte : Elaborada pelo Autor (2021)

Então considerando um conjunto de variáveis correlacionadas entre si, escritas por $X_1, X_2, \ldots, X_p$, com matriz de variância e covariância dada por Σ. Dessa vez, considerando uma sequência de variáveis aleatórias $Y_1, \ldots, Y_p$ de tal maneira que as novas variáveis Y_i são combinações lineares das variáveis X_i Considerando, também, o vetor aleatório $\mathbf{X}^\top = (X_1, X_2, \ldots, X_p)$ com a matriz de covariância Σ. Com $\mathbf{a}_1^\top = (a_{i1}, \ldots, a_{ip}), \forall_i = 1, \ldots, p$ sendo um vetor de constantes. Dado que a matriz Σ é igual a:

$$\sum = \begin{pmatrix} \mathrm{Var}\,(\mathrm{X}_1) & \mathrm{Cov}\,(\mathrm{X}_1, \mathrm{X}_2) & \cdots & \mathrm{Cov}\,(\mathrm{X}_1, \mathrm{X}_\mathrm{p}) \\ \mathrm{Cov}\,(\mathrm{X}_2, \mathrm{X}_1) & \mathrm{Var}\,(\mathrm{X}_2) & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}\,(\mathrm{X}_\mathrm{p}, \mathrm{X}_1) & \ldots & \ldots & \mathrm{Var}\,(\mathrm{X}_\mathrm{p}) \end{pmatrix}$$

Agora as combinações lineares são representadas por:

$$\mathrm{Y}_1 = \mathbf{a}_1^\top \mathbf{X} = a_{11}\mathrm{X}_1 + \mathrm{a}_{12}\mathrm{X}_2 + \ldots + \mathrm{a}_{1p}\mathrm{X}_p$$

$$\vdots$$

$$\mathrm{Y}_p = \mathbf{a}_p^\top \mathbf{X} = \mathrm{a}_{p1}\mathrm{X}_1 + \mathrm{a}_{p2}\mathrm{X}_2 + \ldots + \mathrm{a}_{pp}\mathrm{X}_p$$

Logo a variância de Y_i e a covariância de Y_i e Y_k são, respectivamente, escritas nas formas a seguir:

$$\begin{aligned} &\mathrm{Var}\,(\mathrm{Y_i}) = \mathbf{a}_\mathrm{i}^\top \Sigma \mathbf{a}_\mathrm{i}, && \mathrm{i} = 1, 2, \ldots, \mathrm{p}; \\ &\mathrm{Cov}\,(\mathrm{Y_i}, \mathrm{Y}_k) = \mathbf{a}_\mathrm{i}^\top \sum \mathbf{a}_k, && \mathrm{k} = 1, 2, \ldots, \mathrm{p}, \mathrm{i} \neq \mathrm{k} \end{aligned}$$

Para descobrir as componentes principais Y_1, ..., Y_p, é preciso determinar os valores de $\mathbf{a}_\mathrm{i}^\top, \mathrm{i} = 1, 2, \ldots, \mathrm{p}$ que possuem as maiores variâncias possíveis de Y_i e que fazem as covariâncias serem nulas.

Desse modo, Σ sendo a matriz de variâncias e covariâncias do vetor p-variado X. Os autovalores e autovetores normalizados de Σ são exposto como $(\lambda_i, e_i), i = 1, \ldots, p$ em que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ e a i-ésima componente principal é dada por:

$$\mathrm{Y}_i = \mathrm{e}_i^\top \mathrm{X} = \mathrm{e}_{i1}\mathrm{X}_1 + \ldots + \mathrm{e}_{ip}\mathrm{X}_p;$$

sendo $\mathbf{e}_i^\top = (e_{i1}, \ldots, e_{ip})$.

Sabendo que a matriz Σ é simétrica tem-se que $\mathbf{e}_i^\top e_i = 1$ e $\mathbf{e}_i^\top e_j = 0, i \neq j$ em que $\mathbf{e}_i$ é o autovetor relacionado ao i-ésimo autovalor. Assim é notável que a Var de Y_1 definida como:

$$\mathrm{Var}\,(\mathrm{Y}_1) = \max_{\mathrm{a}_1 \neq 0} \frac{\mathrm{a}_1^\top \sum \mathrm{a}_1}{\mathrm{a}_1^\top \mathrm{a}_1} = \frac{\boldsymbol{e}_1^\top \Sigma \mathrm{e}_1}{e_1^\top \mathrm{e}} = \lambda_1,$$

em outras palavras a primeira componente principal é obtida a partir da combinação linear $\mathbf{a_i}^\top \mathbf{X}$ que maximiza a variância de Y_1 sendo que $\mathbf{a_i}^\top \mathbf{a_i} = 1$. Analogamente tem-se que:

$$\mathrm{Var}\left(\mathrm{Y_k}\right) = \max_{\mathrm{a_k} \neq 0 \perp \mathrm{a_1}, \ldots, \mathrm{a_{k-1}}} \frac{\mathrm{a_k^\top} \sum \mathrm{a_k}}{\mathrm{a_k^\top a_k}} = \frac{\mathrm{e_k^\top} \sum e_k}{\mathrm{e_k^\top e}} = \lambda_k,$$

ou seja, o k-ésimo componente principal se origina da combinação linear $\mathbf{a}_k^\top \mathbf{X}$ que maximiza a variância de $\mathrm{Y_k}$ atribuído à $\mathbf{a_k}^\top \mathbf{a_k} = 1$,e a variância entre $\mathrm{Y_k}$ e $\mathrm{Y_i}$ é zero para $\mathrm{i} < \mathrm{k}$.

Dessa maneira tem-se que:

$$\mathrm{Var}\left(\mathrm{Y_i}\right) = \lambda_\mathrm{i}, \mathrm{i} = 1, \ldots, \mathrm{p} \text{ e } \mathrm{Cov}\ \left(\mathrm{Y_i}, \mathrm{Y}_j\right) = 0, \forall_\mathrm{i} \neq \mathrm{j}.$$

Interpretando a equação acima é notável que os componentes principais têm variância igual aos autovalores e são não variáveis aleatórias não correlacionadas.

Então percebendo que $\mathbf{X}^\top = (\mathrm{X}_1, \ldots, \mathrm{X}_p)$ são variáveis aleatórias com matriz de variância e covariância Σ com autovalores e autovetores dados pelo par $(\lambda_i, e_i), \forall_i = 1, \ldots, \mathrm{p.}$, a soma das variâncias de cada variável é igual à soma das variâncias das componentes principais, ou seja:

$$\sum_{\mathrm{i}=1}^{p} \mathrm{Var}\left(\mathrm{X}_i\right) = \sum_{\mathrm{i}=1}^{p} \mathrm{Var}\left(\mathrm{Y_i}\right) = \sum_{i=1}^{p} \lambda_i.$$

Pois sabe-se que o traço de uma matriz é definido pela soma dos elementos da diagonal principal de uma matriz quadrada. Fazendo a decomposição espectral de Σ obtemse

$$\mathrm{tr}(\Sigma) = \mathrm{tr}\left(\Gamma\Lambda\Gamma^\top\right) = \mathrm{tr}\left(\Lambda\Gamma\Gamma^\top\right) = \mathrm{tr}(\Lambda) = \sum_{\mathrm{i}=1}^{p} \lambda_\mathrm{i},$$

em que $\Gamma\Gamma^\top = \mathbf{I}_p, \Lambda$ é a matriz diagonal dos autovalores e Γ é a matriz dos autovetores por coluna.

Tem-se que a variância populacional total é definida como a soma das variâncias das variáveis aleatórias. Como efeito , definindo a proporção da variância populacional total (PVPT) da i-ésima componente como:

$$\mathrm{PVPT_i} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}, \mathrm{i} = 1, \ldots, \mathrm{p.}$$

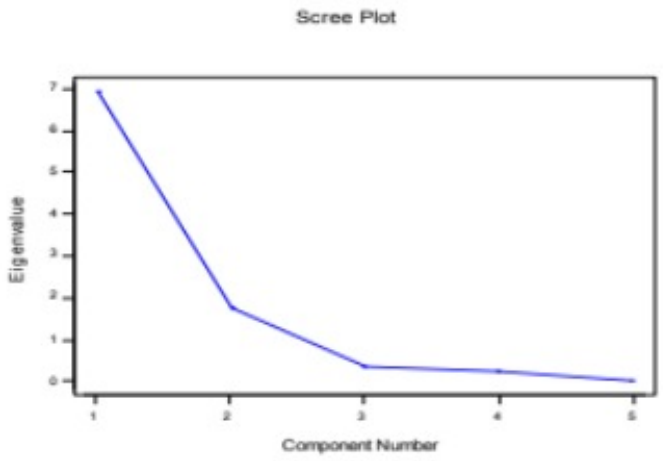
Tendo em vista que o principal objetivo de obter as componentes principais é reduzir

a dimensão das variáveis aleatórias, geralmente é usado a PVPT acumulada até a i-esima componente, se esta medida estiver entre 70% e 90% então é adotado somente as i primeiras componentes principais. Dependendo do estudo, a porcentagem explicada pela PVPT pode ser alterada com base nas quantidades de variáveis, no tamanho da amostra ou na interpretação das componentes principais, dessa forma a escolha depende das informações dadas pelo pesquisador. Por outro lado, os últimos componentes principais serão responsáveis por direções que não estão associadas a muita variabilidade (SILVA, 2018).

Outro método que é usado pelos pesquisadores para selecionar os componentes principais (CPs) que explicam a maior parte da variação dos dados é o critério de Kaiser (KAISER, 1958 apud HONGYU et al., 2016). Este critério obtém CPs com valores próprios maiores do que um ($\lambda_i \geq 1$) e é usado quando a análise é feita pela matriz de correlação.

Outra maneira de escolher os Componentes principais é pelo gráfico denominado por "screeplot". No eixo horizontal (eixo x) é obtida as componentes principais e no eixo vertical (eixo y) os autovalores de cada componente ordenados do maior para o menor. Para encontrar a quantidade ideal de componentes é observado um "cotovelo" (decaimento significativo) no gráfico, então é escolhida a quantidade de componentes até o valor dos autovalores serem baixos, ou seja, próximo de zero.

Figura 2 – Gráfico screeplot com 5 componentes principais



Fonte: Silvestre (2021)

Para interpretar os resultados obtidos dos componentes principais é preciso observar

a correlação entre o i-ésimo componente e a k-ésima variável, seguindo a forma a seguir :

$$\rho(Y_i, X_k) = \frac{e_{i,k}\sqrt{\lambda_i}}{\sqrt{\text{Var}\,(X_{k,k})}} i, k = 1, 2, \ldots, p.$$

Lembrando que as correlações são as medidas padronizadas da relação entre duas variáveis e indica a força e a direção do relacionamento linear entre duas variáveis aleatórias. Sabe-se que embora as correlações das variáveis com os componentes principais geralmente ajudem a interpretar o componente, elas medem apenas a contribuição uni-variada de um indivíduo X_k para um componente Y_i portanto, a interpretação com base nos autovalores podem ser diferentes das obtidas com as correlações (SILVA, 2018).

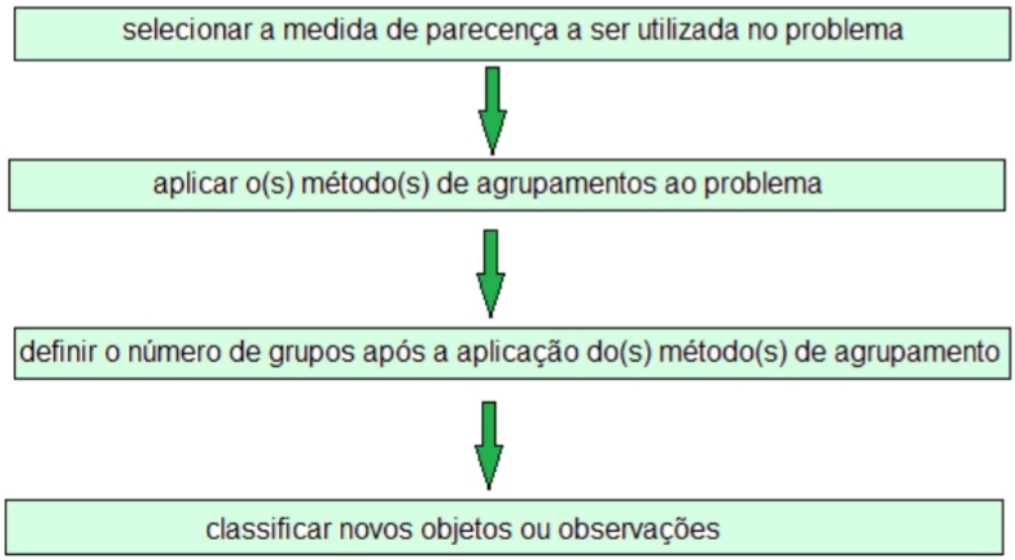
2.3 Análise de Agrupamentos

A análise de agrupamentos ou análise de cluster tem como objetivo dividir os elementos de uma amostra ou população em grupos, de uma maneira que os elementos de um mesmo grupo sejam semelhantes entre si e os elementos de grupos diferentes sejam diversos em relação às variáveis de interesse, então é possível dizer que os métodos de análise de agrupamentos são procedimentos de estatística multivariada que tentam organizar um conjunto de indivíduos para os quais são conhecidas informações detalhadas, em grupos relativamente homogêneos (AFFONSO; TACHIBANA, 2011).

A análise de cluster é bem popular por conta da sua aplicação em diversas áreas. Existem aplicações em data mining, onde a organização de grandes conjuntos de dados torna a análise estatística fácil e mais eficiente; em pesquisa de mercado na identificação de diferentes perfis de consumidores, na construção de estratos na amostragem estratificada, na identificação das variáveis mais importantes na descrição de um determinado fenômeno (SILVESTRE, 2021).

A estrutura da análise de agrupamentos se expressa em quatro itens:

Figura 3 – Itens para se basear na análise de agrupamentos



Fonte: Elaborada pelo autor (2021)

É possível dividir as técnicas de análise de agrupamentos em hierárquicos e não hierárquicos. O método hierárquico propõe que se defina uma matriz que informa a "similaridade" ou a "dissimilaridade" entre os elementos do estudo e sua vantagem é de possibilitar a construção de um gráfico denominado dendrograma, que mostra todas as fases do processo de agrupamento, já o método não hierárquico não armazena as informações básicas do histórico dos agrupamentos, mas tem a vantagem de não ser necessária a construção de uma matriz de semelhança para a sua execução sendo mais eficaz em grandes conjuntos de dados.

Nesse trabalho, a análise de agrupamento visou reunir os atletas que apresentaram os desempenhos mais similares durante seus devidos campeonatos e por meio desta análise verificar quais características cada grupo criado possui fazendo uso de um método hierárquico, então para a construção da matriz de distância será utilizado o método de ward. As medidas de distância representam a similaridade, que é representada pela proximidade entre as observações ao longo das variáveis.

2.3.1 Distância Euclidiana

A distância euclidiana é a medida de distância mais frequentemente usada quando todas as variáveis são quantitativas, como no caso deste trabalho, e é utilizada para

calcular medidas específicas.

Considerando o caso mais simples, no qual existem n indivíduos, onde cada um dos quais possuem valores para p variáveis, a distância euclidiana entre eles é obtida mediante o teorema de Pitágoras, para um espaço multidimensional.

$\mathbf{X_i} = \left(\mathrm{X_{i1}}, \ldots, \mathrm{X_{i}}_p\right)'$ o vetor de observações do objeto i, i = 1, ..., n , sendo que $\mathbf{x_{ij}}$ representa o valor assumido pela variável j no objeto i. Portanto como a definição da distância euclidiana entre os objetos i e k é derivada da ideia de distância existente entre dois pontos no espaço então ela será da seguinte forma:

$$d_{ik} = \sqrt{(\mathbf{x}_i - \mathbf{x}_k)'(\mathbf{x}_i - \mathbf{x}_k)} = \sqrt{\sum_{j=1}^{p} (X_{ij} - X_{kj})^2}$$

ou

$$d_{ik} = \left[\sum_{j=1}^{p} (X_{ij} - X_{kj})^2\right]^{1/2}.$$

Quanto mais próximo de zero for a distância euclidiana, mais similares ou seja, mais idênticos são os objetos comparados.

2.3.2 Método de Ward

O método de Ward equivale a um procedimento de agrupamento hierárquico onde a medida de similaridade usada para juntar os agrupamentos é calculada como a soma de quadrados entre os dois agrupamentos feita sobre todas as variáveis.

Seidel et al. (2008) explica que este método costuma a resultar em agrupamentos de tamanhos aproximadamente iguais por conta da sua minimização de variação interna, ou seja, em cada estágio combinam-se os dois agrupamentos que apresentarem menor aumento na soma global de quadrados dentro dos agrupamentos.

O método de Ward é definido da seguinte forma:

- Inicialmente cada elemento é considerado como um único grupo, então é obtido n grupos tendo apenas um indivíduo cada.

- Em cada passo do algoritmo é calculada a soma de quadrados dentro de cada grupo. Esta soma é o quadrado da distância euclidiana de cada elemento pertencente ao

grupo em relação ao vetor de médias desse grupo, sendo da seguinte forma: $\mathbf{SS_i} = \sum_{j=1}^{n_i} \left(X_{ij} - \overline{\mathbf{X}}_{i.}\right) \cdot \left(X_{ij} - \overline{\mathbf{X}}_{i.}\right)$. Sabendo que , n é o número de elementos no grupo $\mathbf{G_i}$ quando se está no passo k do processo de agrupamento, $\mathbf{X_{ij}}$ é o vetor de observações do j-ésimo elemento que pertence ao i-ésimo grupo, $\mathbf{X_i}$. é o centroide do grupo $\mathbf{G_i}$ e $\mathbf{SS_i}$ é a soma de quadrados correspondente ao grupo $\mathbf{G_i}$. No passo k, a soma de quadrados total dentro dos grupos é definida como : $\mathbf{SSR} = \sum_{\mathbf{i=1}}^{\mathbf{g_k}} \mathbf{SS_i}$, onde k e g é o número de grupos existentes quando se está no passo k.

A distância entre os grupos $\mathbf{G_l}$ e $\mathbf{G_i}$ é definida por:

$$d\left[G_l, G_i\right] = \frac{n_l n_i}{n_l + n_i} \left(\overline{\mathbf{X}}_{l.} - \overline{\mathbf{X}}_{i.}\right) \cdot \left(\overline{\mathbf{X}}_{l.} - \overline{\mathbf{X}}_{i.}\right).$$

Essa distância é a soma de quadrados entre os grupos $\mathbf{G_l}$ e $\mathbf{G_i}$. Em cada etapa do algoritmo, os dois grupos que minimizam a distância acima são combinados. Essa medida é a diferença entre o valor de **SSR** depois e antes de combinar os grupos $\mathbf{G_l}$ e $\mathbf{G_i}$ num único grupo. Portanto, em cada etapa do agrupamento, o método de Ward combina os dois grupos que resultam no menor valor de **SSR**.

2.3.3 Gráfico da Silhueta

O gráfico de Silhueta será usado para verificar a qualidade dos agrupamentos . Esse método que foi proposto por (ROUSSEEUW, 1987), também, ajuda na compreensão e interpretação dos grupos gerados, além de auxiliar na tomada de decisão.

Então, Vale (2005) explica que tendo A como um agrupamento ao qual o objeto i pertence, a dissimilaridade média do objeto i em relação a todos os demais objetos do mesmo grupo A é dada por:

$$a(i) = \frac{1}{|\mathbf{A}| - 1} \sum_{j \in A, j \neq i} d(i, j),$$

sendo que $|\mathbf{A}|$ é o número total de objetos existentes no grupo A e $d(i, j)$ representa a dissimilaridade entre os objetos i e j.

Dessa vez, admita outro agrupamento C diferente de A. A dissimilaridade média do

objeto i para todos os objetos de C é calculada por meio de:

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j),$$

onde de tal maneira que, $|C|$ é o total de objetos do grupo C e $d(i, j)$ é a dissimilaridade entre os objetos i e j.

A menor distância de dissimilaridade entre o objeto i e um grupo A será dada por:

$$b(i) = mind_{C \neq A}(i, C).$$

Considere como B o grupo C que contem a menor distância definida em $b(i)$.Esse grupo é considerado vizinho do objeto i e é o segundo melhor grupo para esse objeto. O valor da silhueta do objeto i é dada por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

Observando a equação acima é observado que o valor de s(i) está entre $-\mathbf{1}$ e 1. Sua interpretação se dá em :

- $s(i) \sim 1$ indica que o objeto i está bem classificado no grupo A;
- $s(i) \sim 0$ indica que o objeto i está entre os grupos A e B;
- $s(i) \sim -\mathbf{1}$ indica que o objeto i está mal classificado no grupo A, e que está mais perto do grupo B do que de A.

Sendo que o gráfico da silhueta do agrupamento A é representado por todos os objetos pertencentes ao agrupamento A em ordem decrescente. Quanto mais próximo de 1, melhor foi a qualidade do agrupamento.

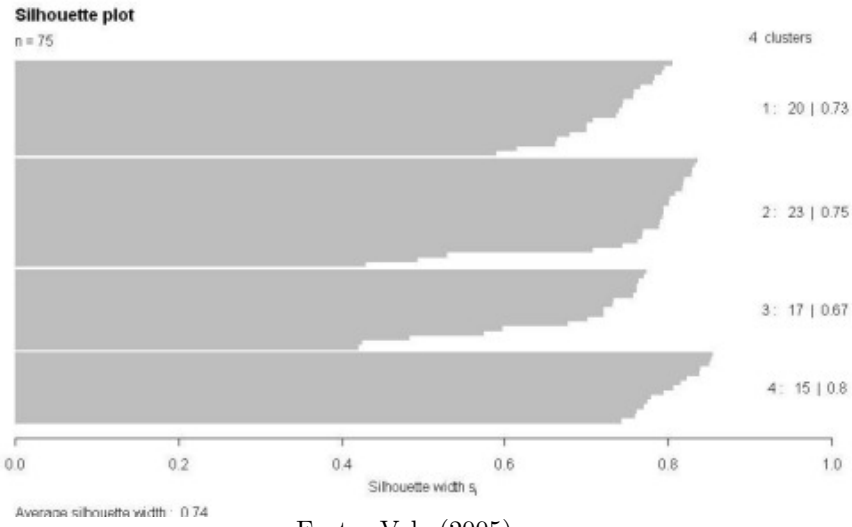
Os valores da silhueta são interpretados da maneira a seguir:

- se $s(i)$ estiver entre 0,71 − 1,00 significa que uma estrutura forte de agrupamento foi encontrada;
- se $s(i)$ estiver entre 0,51 − 0,70 significa que uma estrutura razoável de agrupamento foi encontrada;

- se $s(i)$ estiver entre 0,26 − 0,50 significa que a estrutura de agrupamento é fraca e pode ser superficial, logo é aconselhável o uso de outros métodos para esses dados;
- se $s(i)$ for menor ou igual a 0,25 significa que não teve estrutura de agrupamento entre os indivíduos.

A Figura 4 ilustra um gráfico de silhueta com 4 agrupamentos com 75 indivíduos. Cada grupo possui seu valor da silhueta que julga a qualidade do agrupamento, nesse caso todos os 4 agrupamentos foram bem montados.

Figura 4 – Gráfico de Silhueta



Fonte: Vale (2005)

3 Caracterização dos Dados

3.1 Banco de dados

O banco de dados utilizado para a realização do estudo foi retirado do site infogol.com, este site é sobre apostas futebolísticas e ajuda os usuários com algumas estatísticas de cada jogador atraindo, assim, mais público.

As competições escolhidas foram a La liga, Primiere ligue, Seria A da Itália, France Ligue e o Brasileirão pois esses campeonatos tem o mesmo número de jogos e possuem o mesmo modelo de organização. Dessa maneira é possível fazer as comparações entre os jogadores.

Este conjunto de dados contém 20 jogadores mais famosos de cada campeonato totalizando 100 atletas e suas variáveis foram coletadas durante a temporada de 2019. Essa data foi escolhida pois foi a última temporada que não obteve empecilhos, até então, causados pelo Coronavírus.

As variáveis de cada jogador são :

- Matches Played -> partidas jogadas
- Substitution -> Substituições feitas durante todo o campeonato.
- Mins -> minutos jogados durante todo o campeonato
- Goals -> Gols feitos durante todo o campeonato
- Shots -> Total de chutes feitos durante todo o campeonato.
- OnTarget -> Total de chutes no gol durante todo o campeonato.
- Shots Per Avg Match -> É o valor médio de chutes para o jogador numa média por partidas (incluindo acréscimos)
- On Target Per Avg Match -> É o valor médio de chutes no gol para o jogador numa média por partidas (incluindo acréscimos)

3.2 Análise descritiva

Na Tabela 1 é obtida a representação de 5 jogadores com o intuito de ilustrar como os dados se apresentam. Além dessas variáveis, para cada jogador também tem o campeonato que o mesmo atua. Lembrando que no total o banco de dados possui 100 jogadores.

Tabela 1 – Representação dos dados

Nomes	**Matc.P.**	**Subst.**	**Mins**	**Goals**	**Shots**	**OnTa.**	**Sh.P**	**OT.P**
Lionel M.	32.00	1.00	3067	25.00	159.00	68.00	4.93	2.11
Cristiano R.	33.00	0.00	3127	31.00	208.00	79.00	6.32	2.40
Neymar Jr.	15.00	0.00	1396	12.00	71.00	36.00	4.83	2.45
Gabriel J.	21.00	13.00	2209	14.00	101.00	46.00	4.34	1.98
Dudu	34.00	2.00	3313	9.00	95.00	38.00	2.72	1.09

Fonte: Elaborada pelo autor (2021)

Com a finalidade de compreender melhor o comportamento dos dados serão apresentadas algumas medidas descritivas de tendência central, tais como valor mínimo, mediana, média e valor máximo de cada variável considerando todo o conjunto de dados.

Tabela 2 – Mínimo, máximo e medidas de tendência central de cada variável

Variável	**Mínimo**	**Mediana**	**Média**	**Máximo**
Partidas jogadas	9.00	28.00	26.75	37.00
Subistituição	0.00	3.00	4.14	23.00
Total de minutos jogados	1077.00	2622.00	2531.00	3474.00
Total de gols	7.00	12.00	13.00	36.00
Total de chutes	28.00	80.00	78.97	208.00
Chutes no gol	14.00	32.00	32.88	79.00
Média de chutes por jogo	1.20	2.86	3.00	6.32
Média de chutes no gol por jogo	0.52	1.17	1.26	2.89

Fonte: Elaborada pelo autor (2021)

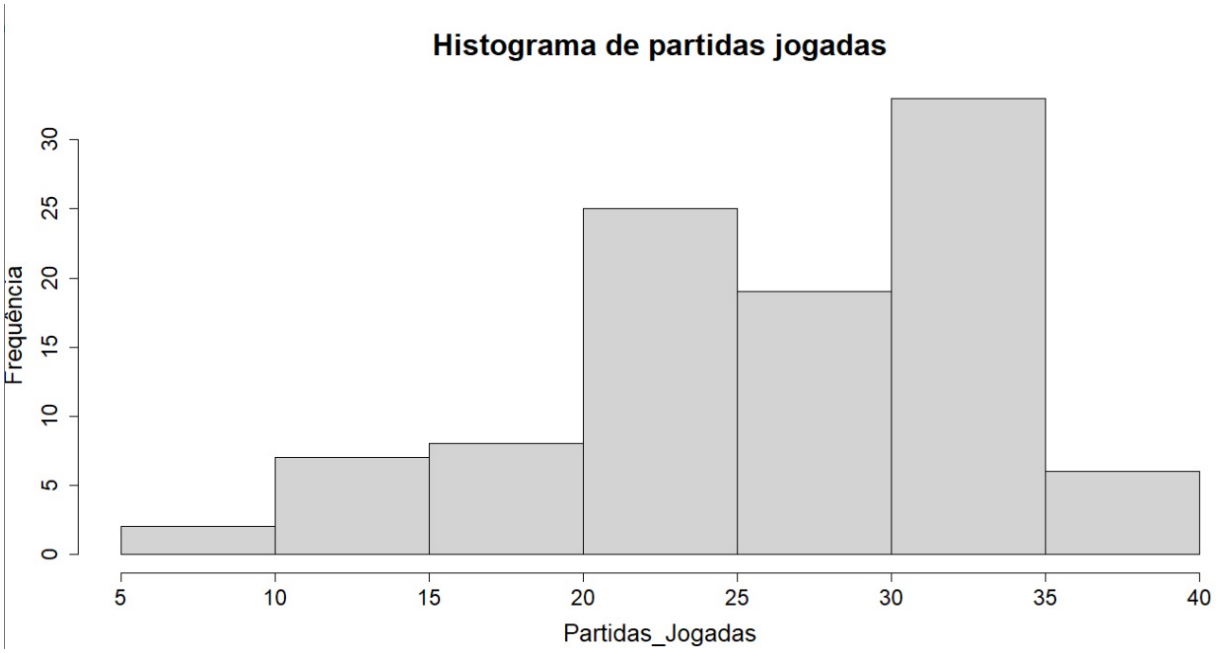
Para facilitar a visualização do comportamento de cada variável são apresentados gráficos de histogramas. O histograma é muito importante para análises estatísticas, ele é um gráfico que mostra a distribuição de acontecimentos registrados.

3.2.1 Histogramas das variáveis

Analisando o histograma das partidas jogadas apresentadas na Figura 5, é possível observar que a maioria dos jogadores disputam 30 a 35 partidas durante todo o campe-

onato. Pode-se dizer que aquele jogador que tem mais jogos disputados tende a ter um melhor desempenho pois possui mais chances de demonstrar seu trabalho.

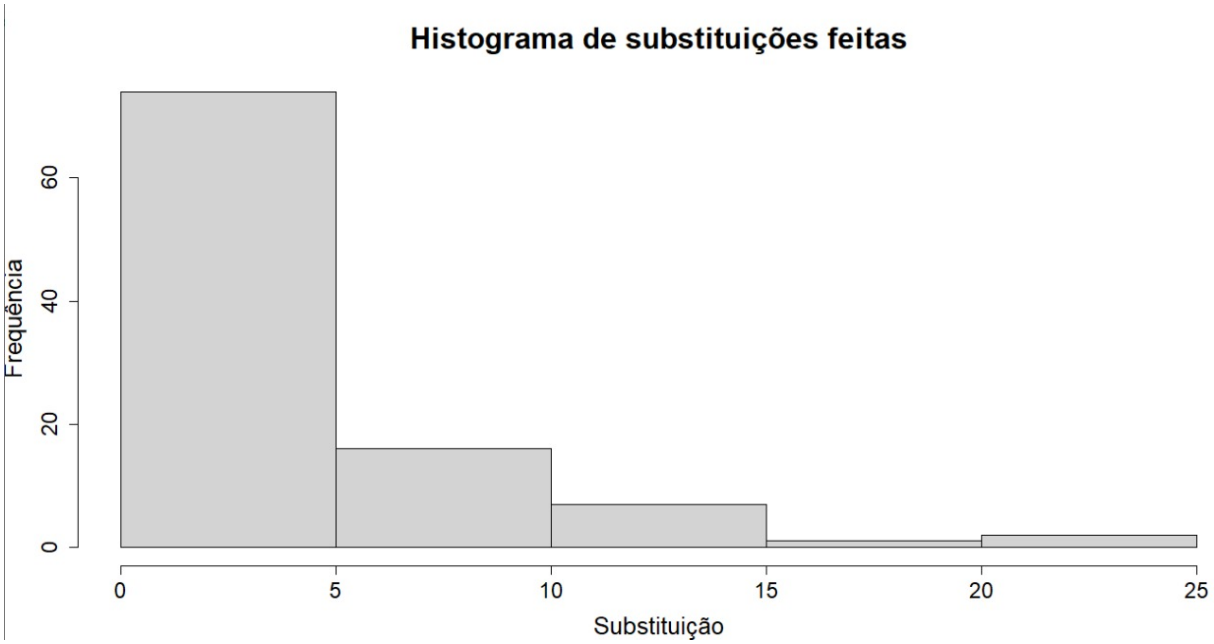
Figura 5 – Histograma de partidas jogadas



Fonte: Elaborada pelo autor (2021)

Observando a Figura 6 é visto que a maioria dos jogadores não são substituidos ou são substituidos poucas vezes ocorrendo de zero a cinco vezes, barra de maior frequência no histograma, durante o campeonato. Mas nos dados também possui aqueles que foram substituidos muitas vezes, cerca de 10 a 15 trocas, isso é muito maléfico para o desempenho do jogador uma vez que durante os jogos as substituições são limitadas.

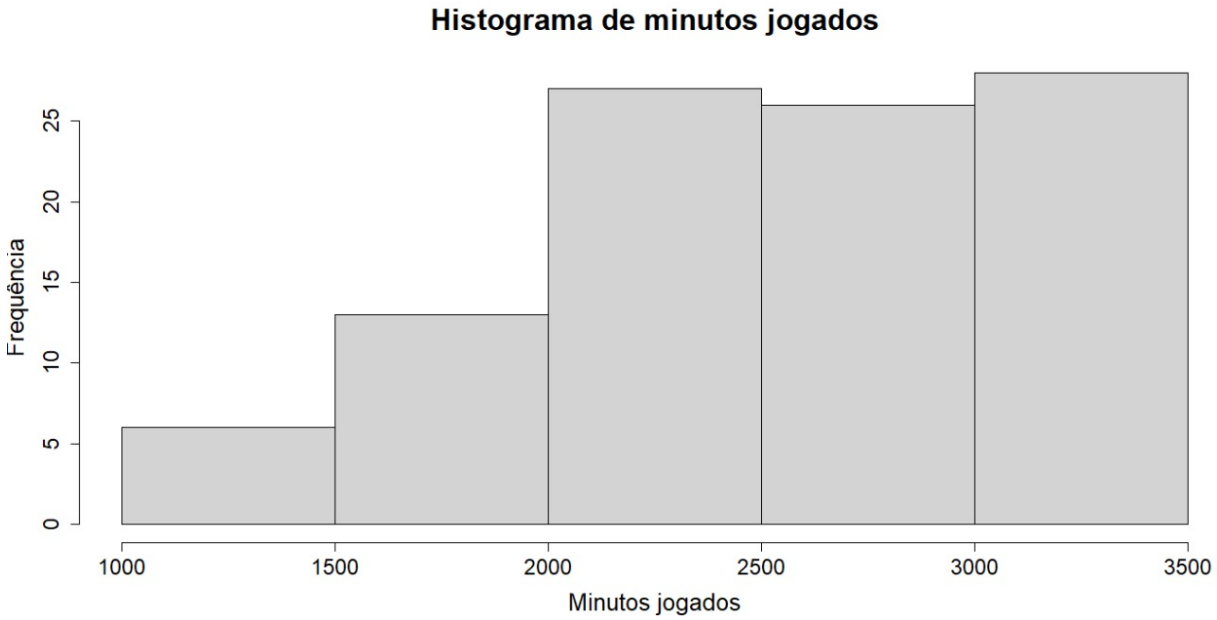
Figura 6 – Histograma de substituições ocorridas



Fonte: Elaborada pelo autor (2021)

Observando as informações da Figura 7, pode-se notar que a maioria dos atletas jogam de 3.000 a 3.500 minutos durante toda competição, mas também possui uma granden frequencia entre 2.000 e 2.500 minutos. O jogador que possui muitos minutos jogados significa que foi substituido poucas vezes então a interpretação fica inversamente porporcional ao do histograma das substituições, portanto aquele que possui mais minutos jogados tende a ter um desempenho melhor.

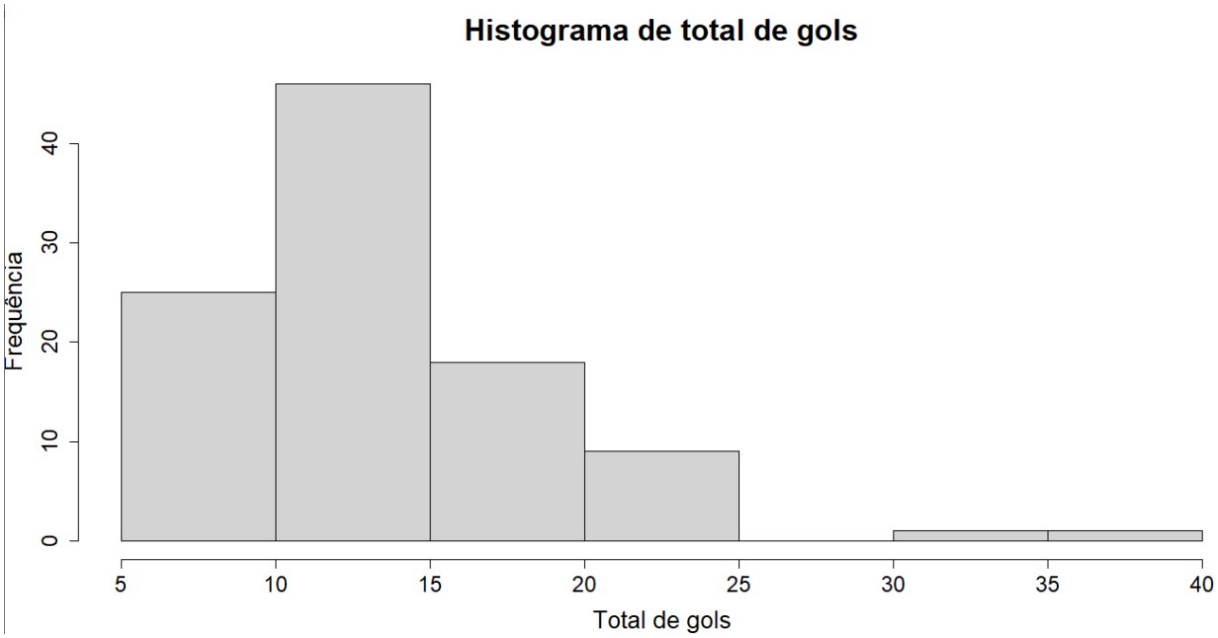
Figura 7 – Histograma de minutos jogados



Fonte: Elaborada pelo autor (2021)

Na interpretação da Figura 8 observa-se que a maioria dos jogadores fizeram de 10 a 15 gols durante seus devidos compeonatos. A segunda maior frequência é dada pelo intervalo de 5 a 10 gols. Também é observado que possuem alguns jogadores que se destacaram em quesito de números de gols chegando a marcar mais de 35 gols. O número de gols é a variável mais importante no rendimento dos jogadores, logo aquele que possui mais gols terá um score muito favorável.

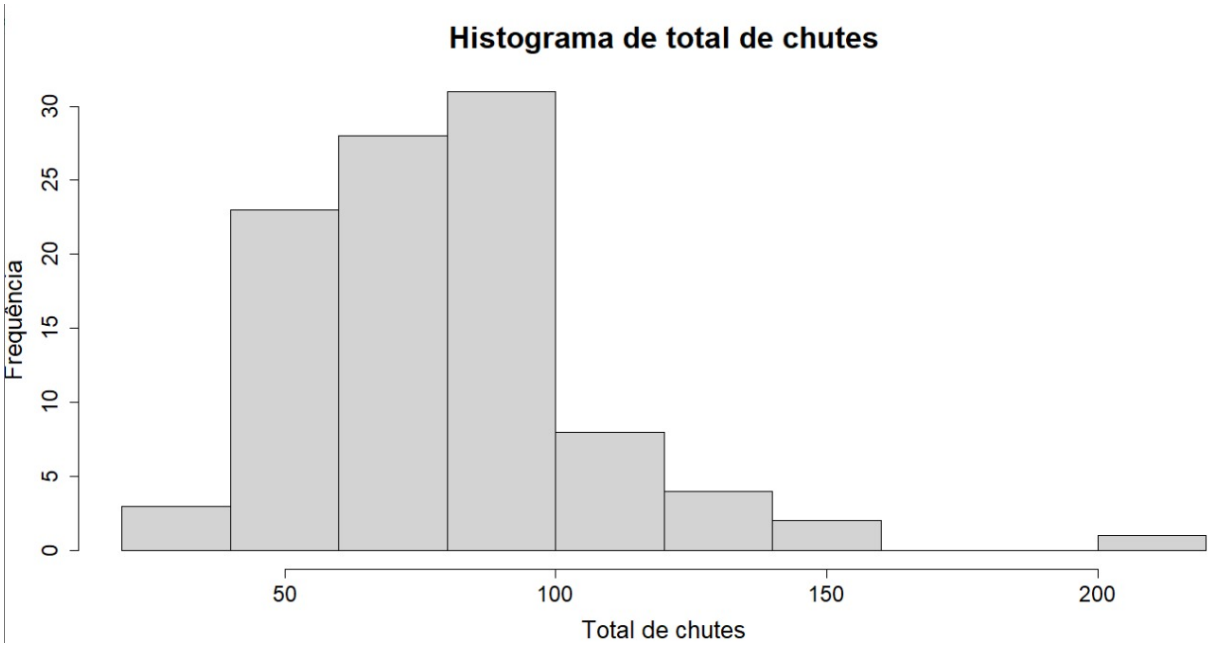
Figura 8 – Histograma de gols feitos



Fonte: Elaborada pelo autor (2021)

Observando o histograma de total de chutes é visivel que a maioria dos jogadores chutam num intervalo de 50 a 100 chutes por competição, e tem alguns destaques de atletas que chutaram mais de 200 vezes. Aqueles jogadores que chutam com maior frequencia tende a ter um rendimento melhor pois a chance de marcar um gol aumenta, mas nessa variável possui um contraponto nesse quesito pois se o jogador chuta muito e não marca para seu time o mesmo possuirá um rendimento muito debilitado.

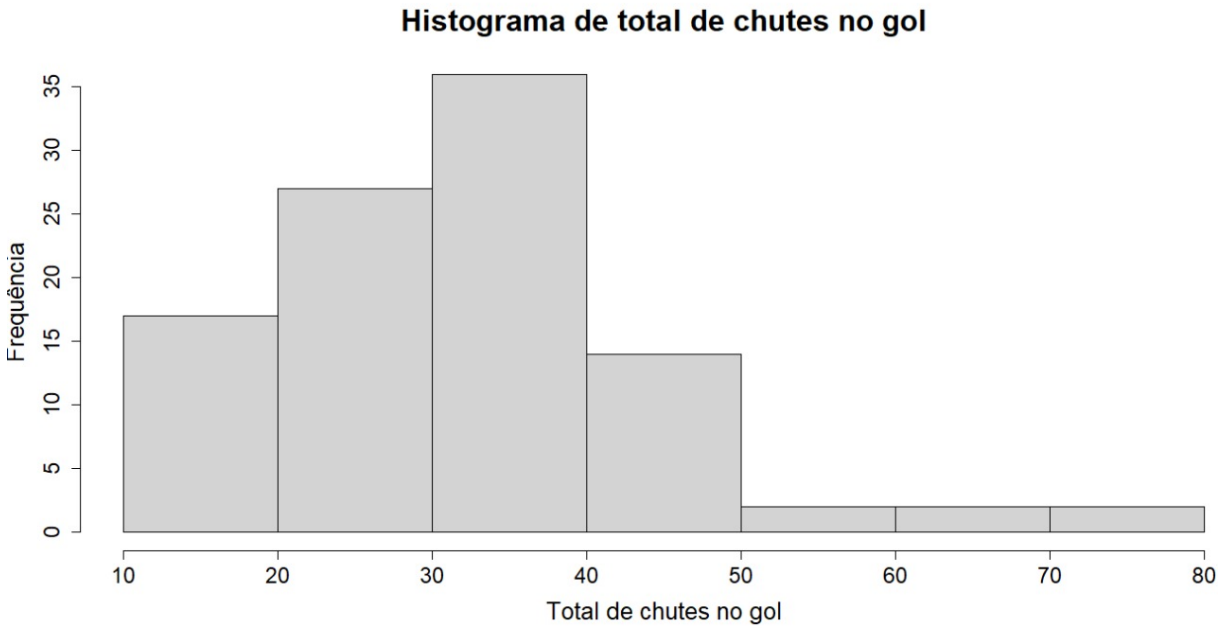
Figura 9 – Histograma do total de chutes



Fonte: Elaborada pelo autor (2021)

Tem-se que a maioria dos jogadores chutam, com mais frequencia, cerca de 20 a 50 vezes a bola em diração ao gol durante todo o campeonato, e sempre tendo alguns jogadores destaques que chutem mais 70 vezes na direção do gol tendo mais chances de marcar para seu time, se observado na Figura 10. Esta variável assim como a variável Chutes precisa de cautela para sua interpretação pois se o jogador chuta muito em direção ao gol mas não marca, esse fator, com certeza, diminuirá seu rendimento.

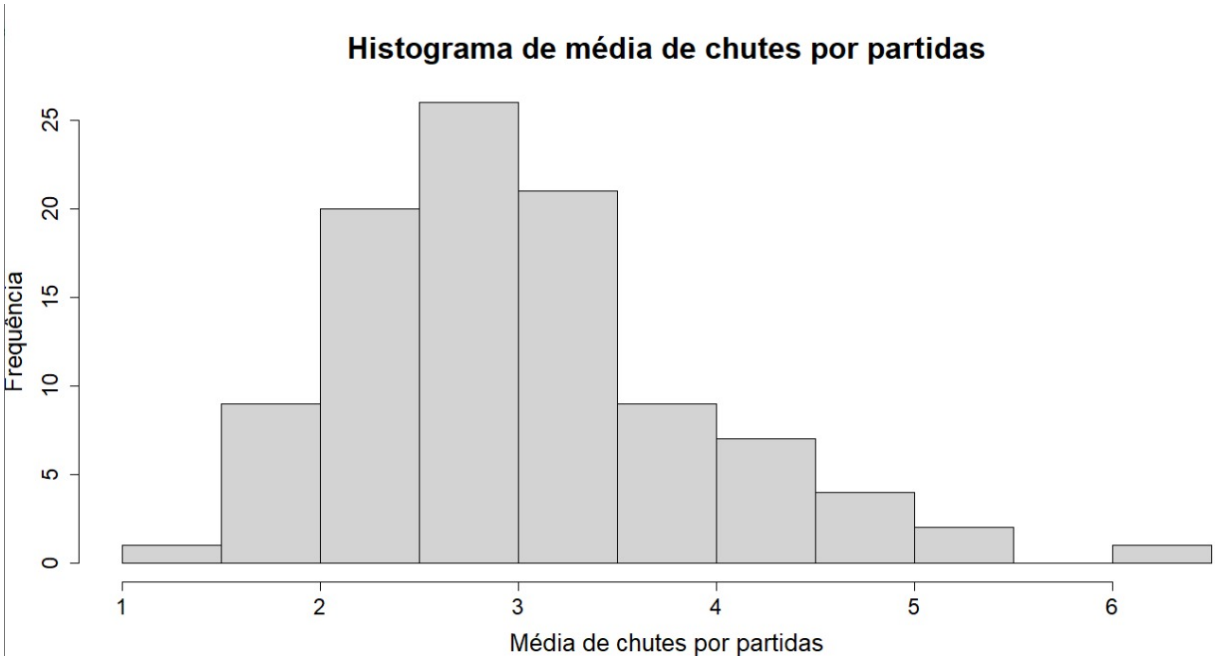
Figura 10 – Histograma de chutes na direção do gol



Fonte: Elaborada pelo autor (2021)

A fim de encontrar novas variáveis que possam influenciar no desempenho de um jogador de futebol, foi feita a coleta de dados sobre a média de chutes numa média por partidas de cada jogador. Analisando a Figura 11 tem-se que a maioria dos jogadores chutam em média cerca de duas a três vezes por partida, sendo que desses chutes nem todos são na direção do gol.

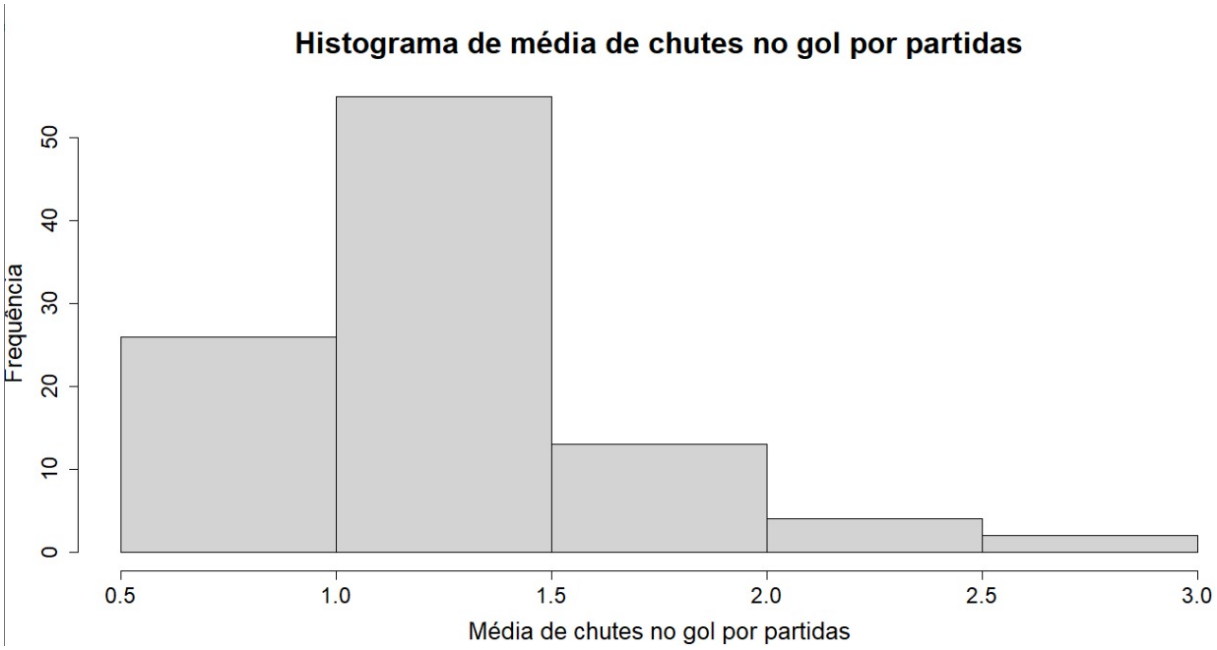
Figura 11 – Histograma do valor médio de chutes numa média por partida



Fonte: Elaborada pelo autor (2021)

Seguindo a mesma linha de inovação de variáveis, também foram coletados os valores médios de chutes na direção do gol de cada jogador com o intuito de medir o rendimento em relação a finalização do atleta. Fazendo a observação do gráfico a maior frequencia esta entre um chute e um chute e meio na direção do gol por partida. Tendo alguns destaques que possuem uma média de até 3 chutes na direção do gol por partida.

Figura 12 – Histograma do valor médio de chutes na direção do gol numa média por partida



Fonte: Elaborada pelo autor (2021)

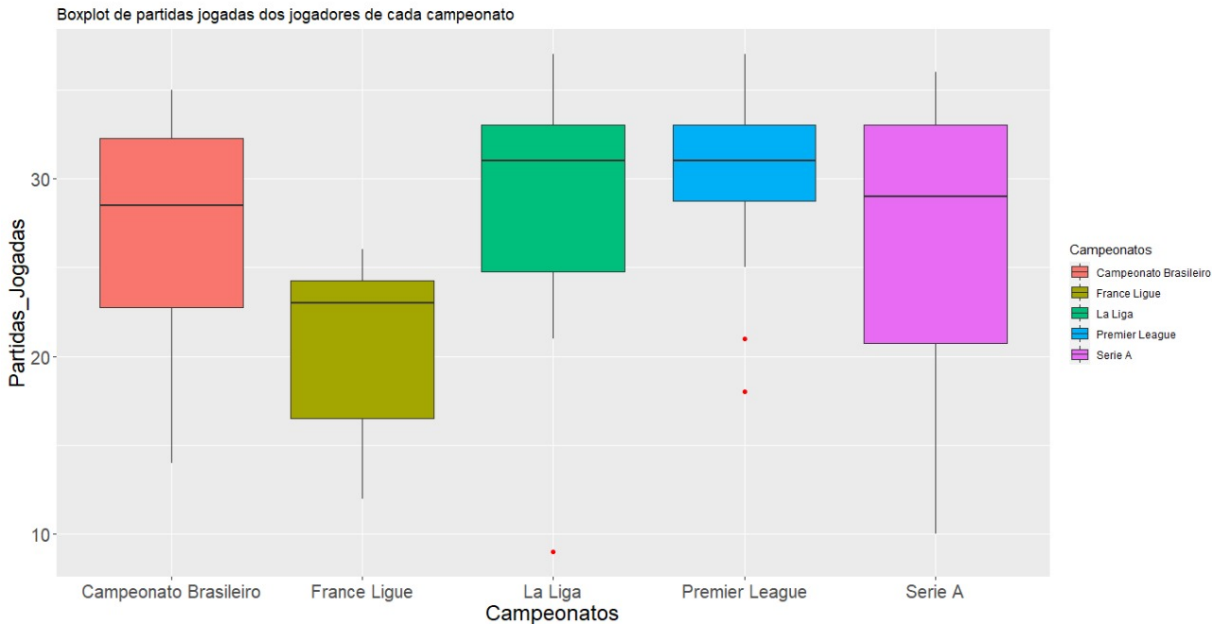
3.2.2 Comparações das variáveis entre os campeonatos escolhidos

Essa seção tem como objetivo fazer algumas comparações das variáveis levando em consideração cada campeonato. Essas primeiras análises e comparações darão uma direção do que os dados podem estar representando antes das aplicações dos modelos sugeridos.

Os gráficos boxplot que foram feitos permitem visualizar a distribuição e valores discrepantes dos dados, fornecendo assim um modo complementar para obter uma perspectiva sobre o caráter dos dados, além de ajudar nas comparações. O boxplot é formado pelo valor mínimo, valor máximo, primeiro quartil, segundo quartil e terceiro quartil.

Nesses boxplots da Figura 13 é notado que os jogadores da France Ligue são os que possuem menos número de jogos disputados. Já os jogadores do Brasileirão, La Liga, Premier League e Serie A da Italia apresentam um número de partidas jogadas parecidos. Esses boxplots estão representando bem os dados, pois apenas a La Liga e a Premier League possuem out liers. Também é observado que os dados da Premier League tem a menor variabilidade das demais competições.

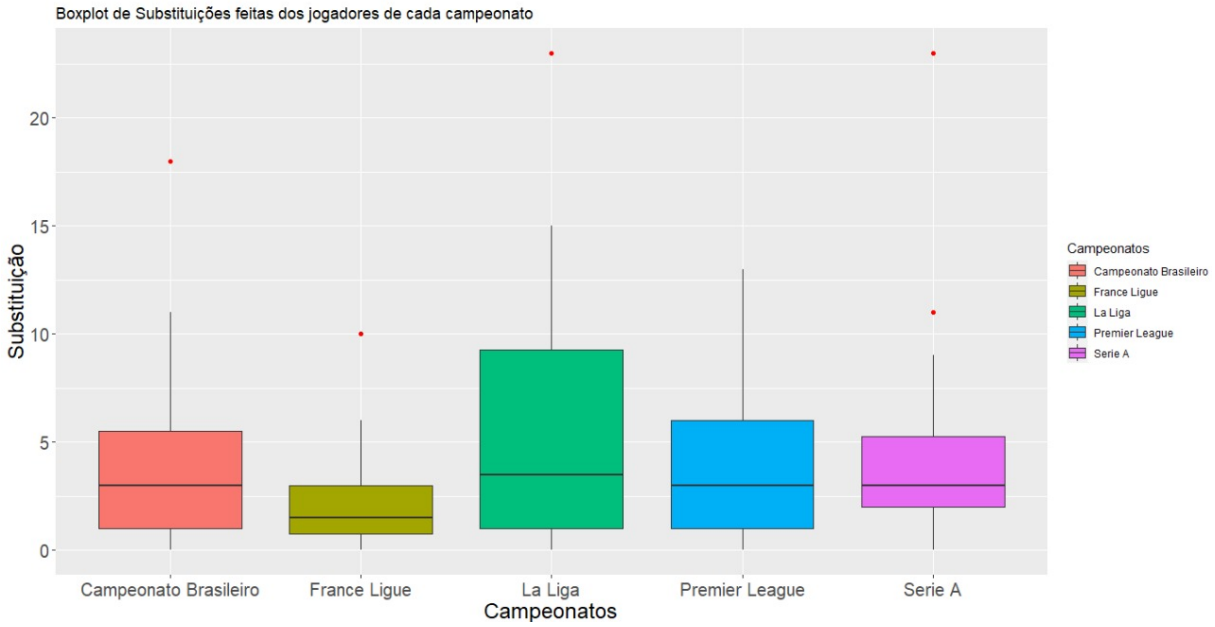
Figura 13 – Boxplot de partidas jogadas dos jogadores de cada campeonato



Fonte: Elaborada pelo autor (2021)

Nos gráficos boxplots da Figura 14 referente as substituições é percebido que os jogadores de todas as competições tem substituições pareadas. Esses boxplots estão representando bem os dados uma vez que possuem apenas poucos outliers. Dando um adendo para um jogador da La Liga e um da Serie A da Italia que foram substituidos mais de 20 vezes influenciando negativamente para seus devidos desempenhos.

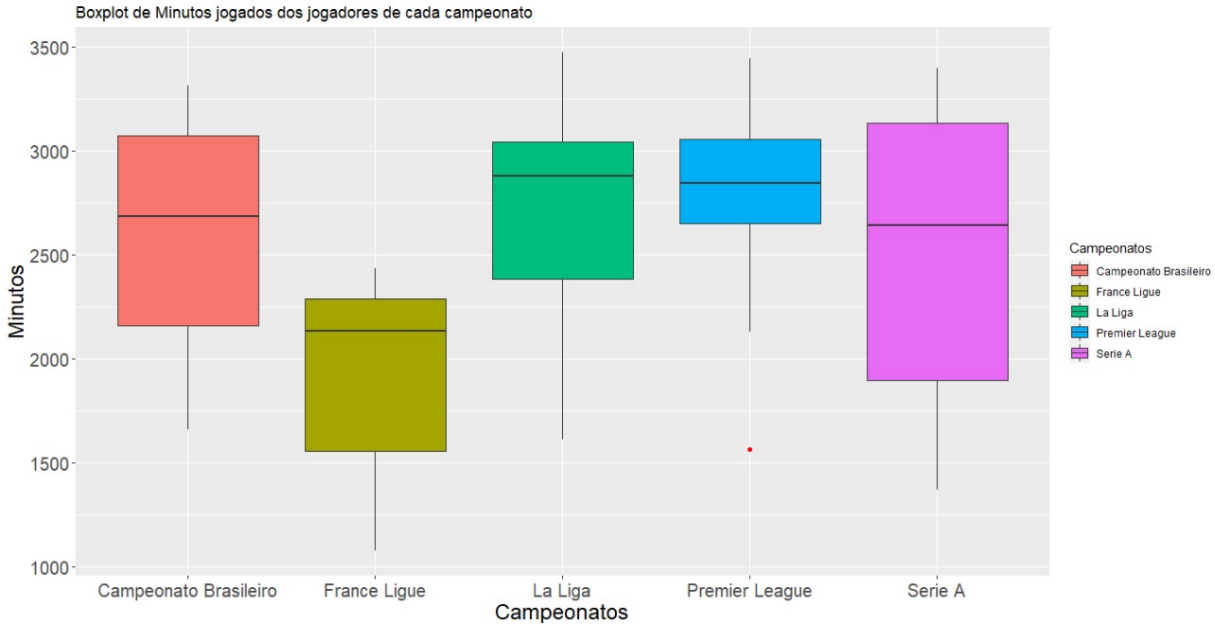
Figura 14 – Boxplot de substituições dos jogadores de cada campeonato



Fonte: Elaborada pelo autor (2021)

Tem-se que os boxplots da Figura 15 estão representando bem os dados sendo que apenas na Premier League possui um out lier. Interpretando as comparações no número de partidas jogadas, os jogadores da France Ligue também possuem menos minutos jogados enquanto os atletas das demais competições possuem um número parecido de minutos em ação.

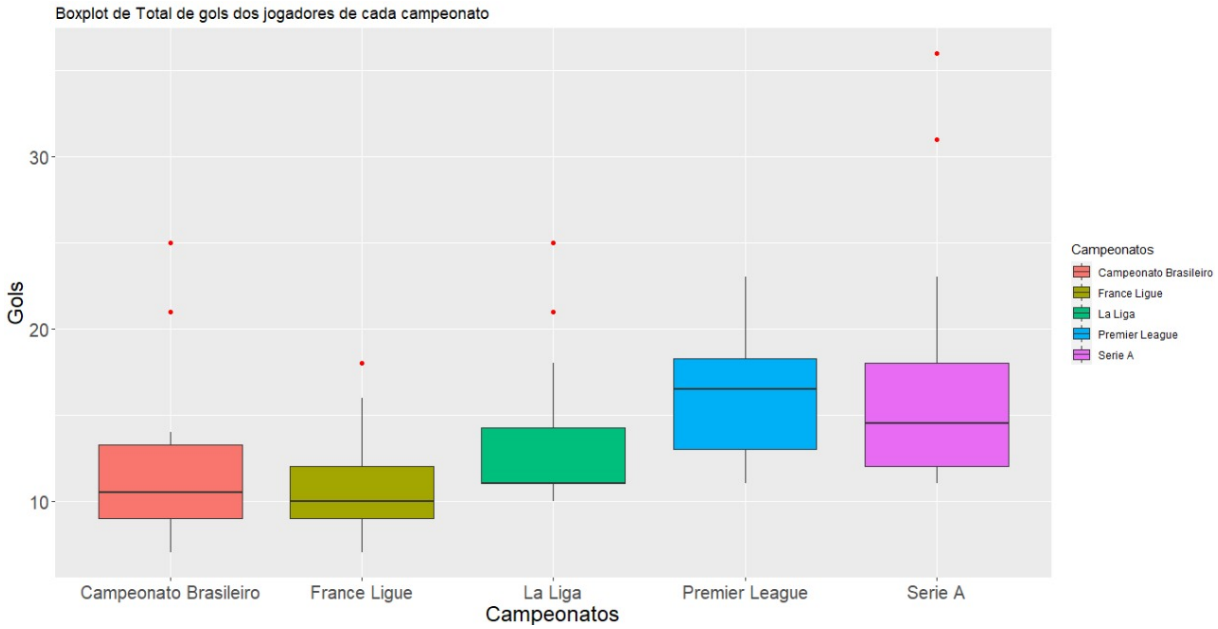
Figura 15 – Boxplot de minutos jogados dos jogadores de cada campeonato



Fonte: Elaborada pelo autor (2021)

Nesses gráficos da Figura 16 possuem alguns outliers mas no geral os dados estão sendo bem representados. Tem-se que a Liga que tem os jogadores com maior números de gols é a Premier League, mas todas as competições estão pareadas nesse quesito. Deixando em evidencia o campeonato Serie A da Itália que possui dois jogadores que fizeram mais de 30 gols no campeonato, impactando positivamente no desempenho dos mesmos.

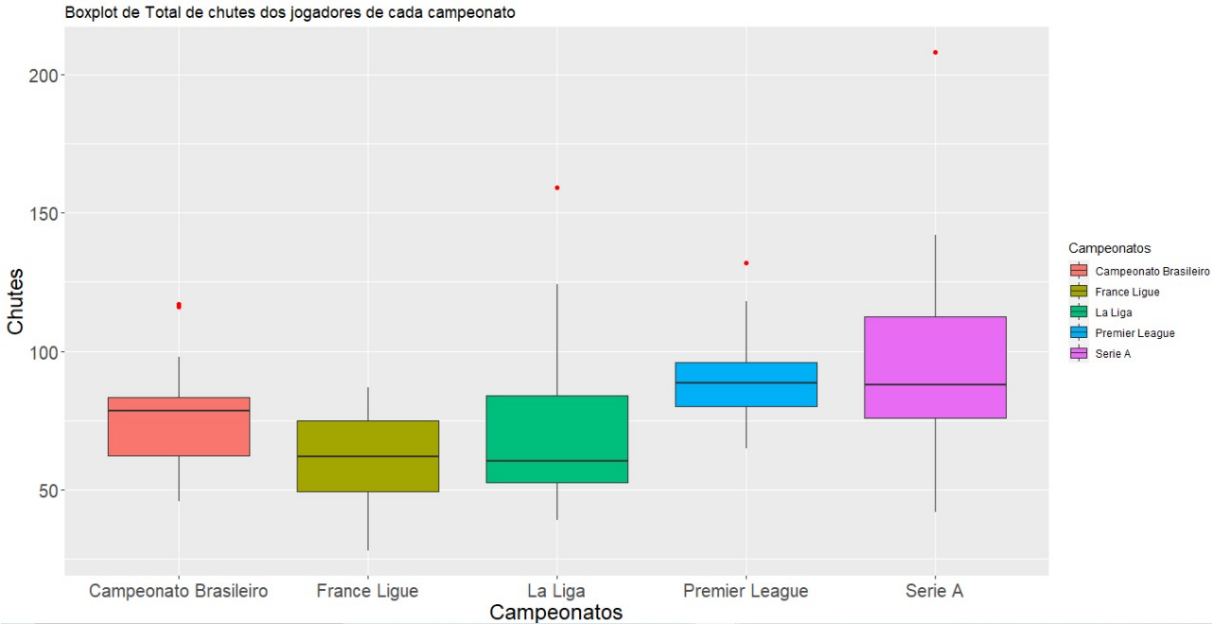
Figura 16 – Boxplot de gols feitos dos jogadores de cada campeonato



Fonte: Elaborada pelo autor (2021)

Os boxplots referente aos totais de chutes dos jogadores em suas devidas competições também estão sendo bem representados pelos dados, dando evidência na Serie A da Italia tendo um jogador com mais de 200 chutes e na La liga tendo um jogador que teve mais de 150 chutes totais.

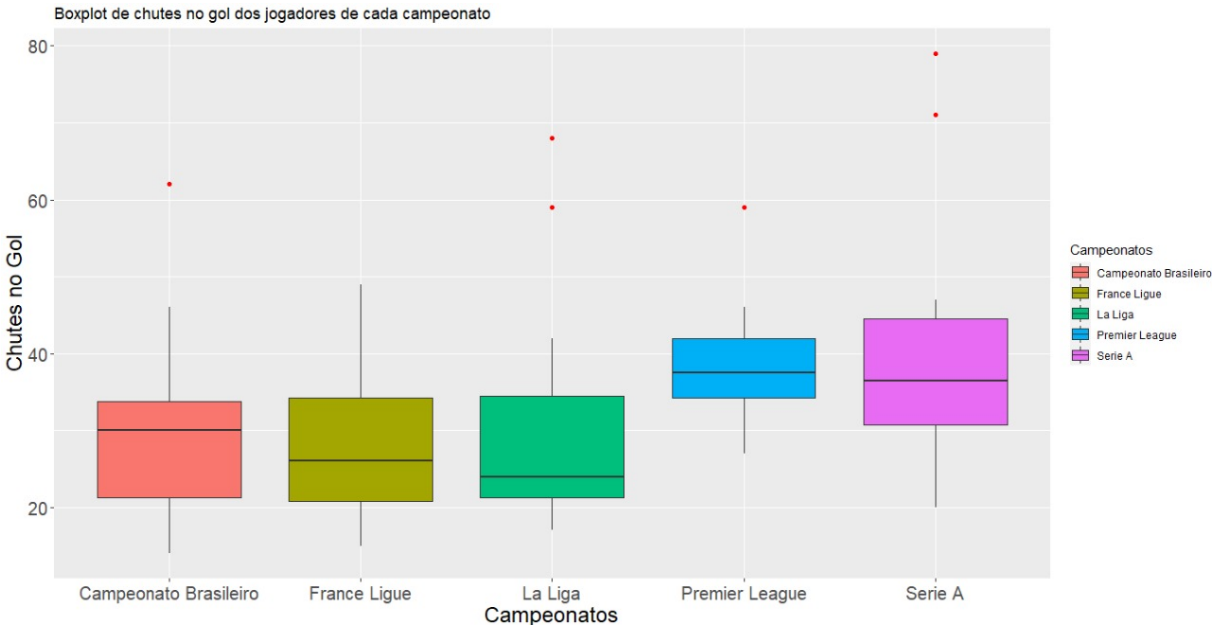
Figura 17 – Boxplot de chutes dos jogadores de cada campeonato



Fonte: Elaborada pelo autor (2021)

A Figura 18 nos apresenta os gráficos de boxplot dos chutes na direção do gol de cada jogador separados por competição. Apenas o boxplot da Liga da França não possuem out liers logo é o boxplot que melhor representa seus dados, os boxplots dos campeonatos restantes apresentam apenas no máximo dois pontos discrepantes portanto os dados estão sendo razoavelmente bem representados. Dando mais importância a um jogador do Brasileirão, um da La Liga e dois da Série A da Itália que finalizaram em direção ao gol mais de 60 vezes.

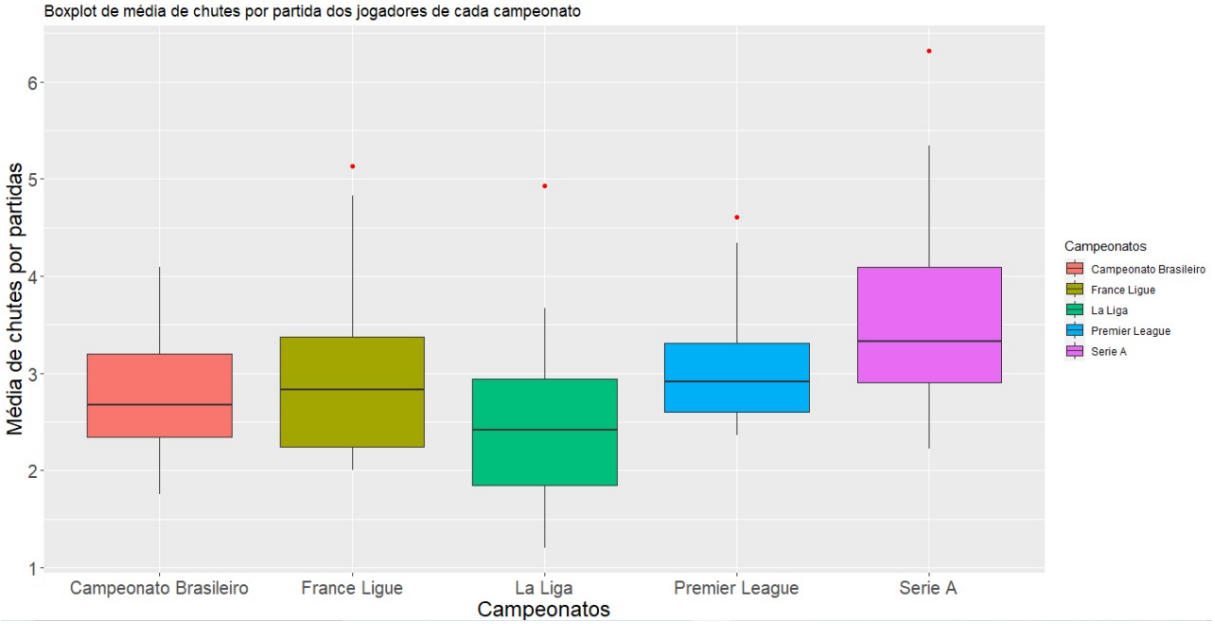
Figura 18 – Boxplot de chutes na direção do gol dos jogadores de cada campeonato



Fonte: Elaborada pelo autor (2021)

Já nos boxplots da Figura 19 é observada comparações referente à média de chutes de um jogador. Diferente dos demais, essa medida leva em consideração o desempenho durante a partida. Logo, fazendo a leitura dos gráficos percebe-se que apenas o boxplot do Brasileirão não possui out liers já os demais, possuem um jogador que se desaca, dando créditos aos atletas da Ligue 1 France e da Série A da Itália que tiveram o valor médio por partida de mais de 5 chutes.

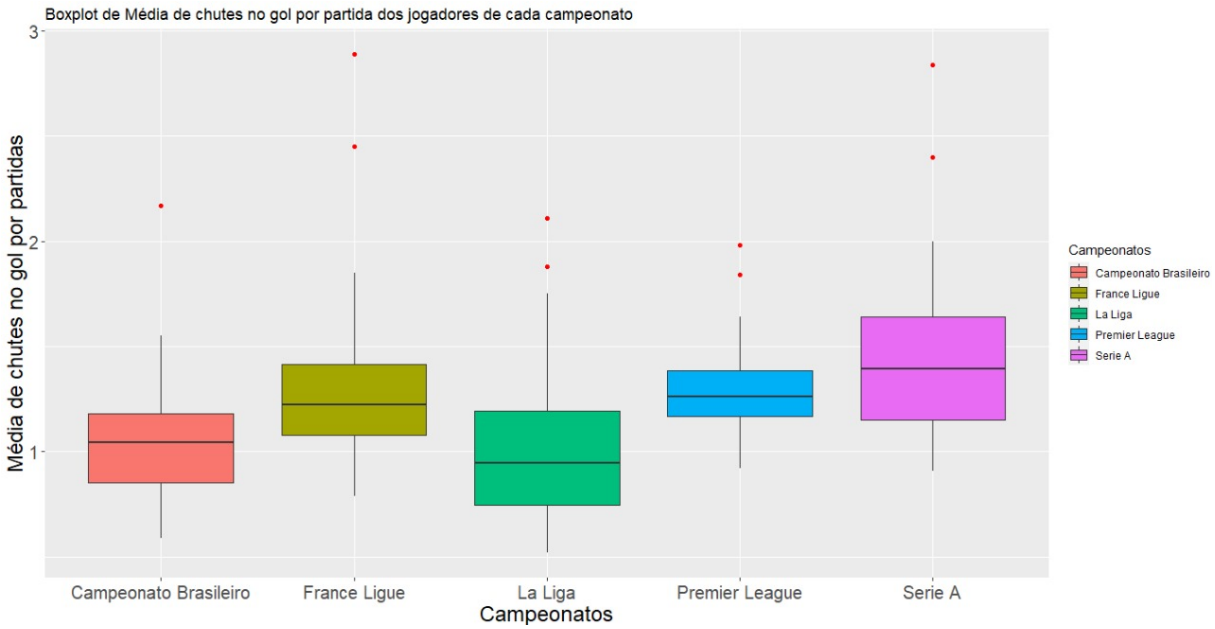
Figura 19 – Boxplot do valor médio de chutes dos jogadores de cada campeonato numa média por partida



Fonte: Elaborada pelo autor (2021)

Nesses últimos boxplots da Figura 20 o intuito é comparar a qualidade das finalizações dos jogadores por cada liga. Nesses gráficos é obtido a presença de pontos atípicos em todas as competições, esses pontos atípicos refletem um bom desempenho dos jogadores aos demais de suas respectivas ligas, pois quanto mais chutes ao gol um atleta tiver mais chances de marcar o mesmo possui. Deixando em evidência um jogador do brasileirão, dois da Liga da França um da La liga, e dois da Série A da Itália que tiveram em média mais de dois chutes ao gol por partida. Sendo a Série A da Itália o campeonato com maior média de finalizações por partidas.

Figura 20 – Boxplot do valor médio de chutes na direção do gol dos jogadores de cada campeonato numa média por partida



Fonte: Elaborada pelo autor (2021)

4 Aplicação do modelo de Componentes principais e resultados

Dada a matriz de covariância que é obtida na Figura 21 é possível observar que existe uma discrepância bem grande entre as variáveis, uma vez que as variáveis estão em escalas diferentes portanto a variância delas também se encontram diferentes.

Figura 21 – Matriz de covariância das variáveis

	Partidas_Jogadas	Substituição	Minutos	Gols	Chutes	Chutes no Gol	Média de chutes por partidas	Média de chutes no gol por partidas
Partidas_Jogadas	46.4318182	-19.4191919	3917.06313	13.906566	96.729798	35.434343	-1.3375758	-0.8512121
Substituição	-19.4191919	21.9195960	-1321.94242	-4.899596	-37.470505	-15.871919	0.2549293	0.1368202
Minutos	3917.0631313	-1321.9424242	350496.35101	1244.307576	8862.953030	3232.660606	-113.5258586	-73.5850909
Gols	13.9065657	-4.8995960	1244.30758	25.581717	98.690606	51.095758	2.0131111	1.1895677
Chutes	96.7297980	-37.4705051	8862.95303	98.690606	767.423333	307.430707	16.5172121	6.3575717
Chutes no Gol	35.4343434	-15.8719192	3232.66061	51.095758	307.430707	147.662222	7.1765253	3.7417010
Média de chutes por partidas	-1.3375758	0.2549293	-113.52586	2.013111	16.517212	7.176525	0.7824788	0.3504075
Média de chutes no gol por partidas	-0.8512121	0.1368202	-73.58509	1.189568	6.357572	3.741701	0.3504075	0.1976947

Fonte: Elaborada pelo autor (2021)

Quando isso ocorre é preciso padronizar os dados, para isso é usada a matriz de

correlação que é apresentada na Figura 22. Com a matriz de correlação é notado que existe, de fato, correlação entre algumas variáveis, sendo que a maior correlação está entre as variáveis 'Minutos jogados' e 'Partidas jogadas', já a segunda maior correlação está entre 'Total de chutes' e 'Total de chutes na direção do gol'.

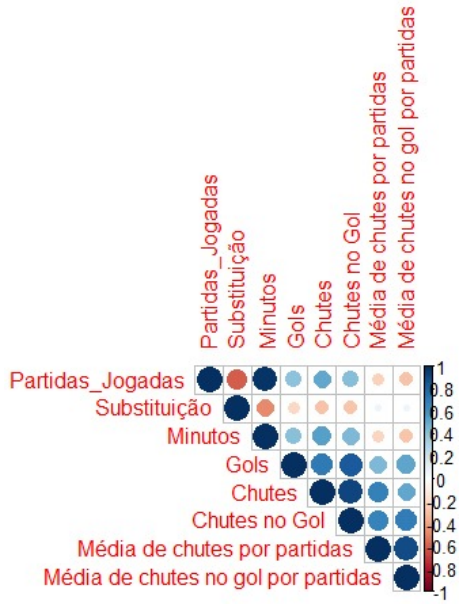
Figura 22 – Matriz de correlação das variáveis

	Partidas_Jogadas	Substituição	Minutos	Gols	Chutes	Chutes no Gol	Média de chutes por partidas	Média de chutes no gol por partidas
Partidas_Jogadas	1.0000000	-0.60870523	0.9709815	0.4035034	0.5124305	0.4279390	-0.22190866	-0.28095220
Substituição	-0.6087052	1.00000000	-0.4769299	-0.2069089	-0.2889055	-0.2789835	0.06155555	0.06572588
Minutos	0.9709815	-0.47692986	1.0000000	0.4155480	0.5404048	0.4493489	-0.21677878	-0.27954425
Gols	0.4035034	-0.20690889	0.4155480	1.0000000	0.7043580	0.8313527	0.44995244	0.52896482
Chutes	0.5124305	-0.28890548	0.5404048	0.7043580	1.0000000	0.9132613	0.67403579	0.51615037
Chutes no Gol	0.4279390	-0.27896347	0.4493489	0.8313527	0.9132613	1.0000000	0.66764136	0.69252724
Média de chutes por partidas	-0.2219087	0.06155555	-0.2167788	0.4499524	0.6740358	0.6676414	1.00000000	0.89092174
Média de chutes no gol por partidas	-0.2809522	0.06572588	-0.2795442	0.5289648	0.5161504	0.6925272	0.89092174	1.00000000

Fonte: Elaborada pelo autor (2021)

A Figura 23 auxilia na observação das variáveis que possuem uma forte correlação, se o circulo for grande e azul significa que as variáveis tem forte correlação positiva, se o circulo for grande e vermelho as variáveis possuem uma correlação forte e inversamente proporcional e se forem pequenas e brancas não possuem nenhum tipo de correlação.

Figura 23 – Gráfico de correlações



Fonte: Elaborada pelo autor (2021)

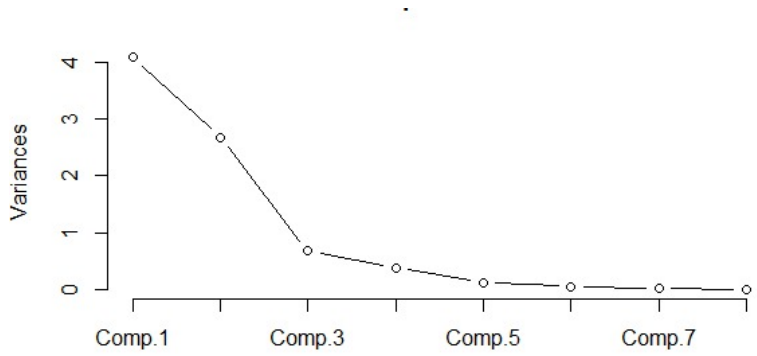
A implementação do modelo de componentes principais é dada encontrando os autovalores e a proporção da variância explicada por cada autovalor, lembrando que o valor da componente principal é o mesmo do autovalor. Nos dados do estudo percebe-se que os dois primeiros componentes explicam cerca de 84% da variabilidade total, também analisando os valores dos componentes pelo método de Kaiser (1958) é notável que os dois primeiros componentes são maiores do que um, já no gráfico screplot nota-se um decaimento brusco a partir do segundo para o terceiro componente . Portanto devido as análises feitas com o auxílio do gráfico screeplot é evidente que os dois primeiros componentes principais resumem efetivamente a variância amostral total e podem ser utilizados para o estudo do conjunto de dados.

Tabela 3 – Valores das Componentes principais e proporção da variância explicada

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
2.021	1.635	0.821	0.620	0.345	0.220	0.118	0.053
0.510	0.334	0.084	0.048	0.014	0.006	0.001	0.0009
0.510	0.844	0.929	0.977	0.991	0.997	0.999	1.000

Fonte: Elaborada pelo autor (2021)

Figura 24 – Gráfico screeplot



Fonte: Elaborada pelo autor (2021)

Com o objetivo de se compreender a importância de cada variável na construção dos dois componentes foram calculados os coeficientes de ponderação, que são os autovetores, e a correlação entre as variáveis originais e os componentes principais. Com a seleção de dois componentes principais, a redução da dimensão de 8 variáveis originais para 2 componentes principais é bem satisfatório. Portanto decidiu-se utilizar unicamente os dois primeiros componentes principais para a composição das equações CP1 e CP2, sendo que o primeiro componente representa 51% e o segundo representa 33% da variabilidade total.

Tabela 4 – Autovetores e Correlações dos dois primeiro componentes

Variáveis	Autovetores		Correlação	
	Comp.1	Comp.2	Comp.1	Comp.2
Partidas jogadas	0.278	0.496	0.561	0.811
Subistituição	-0.201	-0.329	-0.405	-0.537
Total de minutos jogados	0.279	0.478	0.563	0.780
Total de gols	0.419	-0.037	0.845	-0.060
Total de chutes	0.468	-0.020	0.945	-0.032
Total de chutes no gol	0.481	-0.086	0.971	-0.141
Média de chutes por jogo	0.308	-0.439	0.622	-0.7181
Média de chutes no gol por jogo	0.292	-0.464	0.590	-0.758

Fonte: Elaborada pelo autor (2021)

Logo, foram obtidas duas equações:

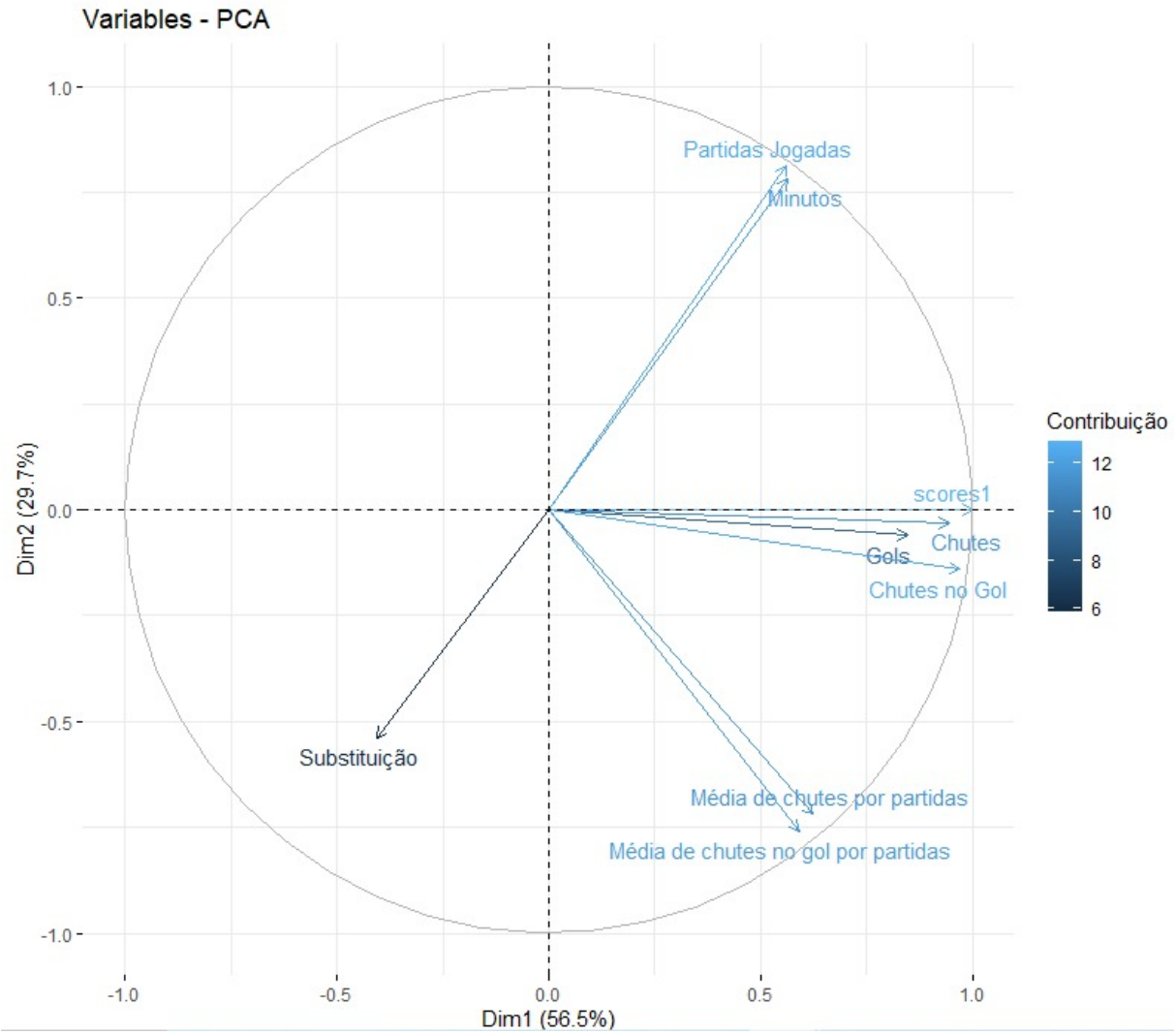
$$CP1 = 0.278\mathrm{X}_1 - 0.201\mathrm{X}_2 + 0.279\mathrm{X}_3 + 0.419\mathrm{X}_4 + 0.468\mathrm{X}_5 + 0.481\mathrm{X}_6 + 0.308\mathrm{X}_7 + 0.292\mathrm{X}_8$$

$$CP2 = 0.496\mathrm{X}_1 - 0.329\mathrm{X}_2 + 0.478\mathrm{X}_3 - 0.037\mathrm{X}_4 - 0.020\mathrm{X}_5 - 0.086\mathrm{X}_6 - 0.439\mathrm{X}_7 - 0.464\mathrm{X}_8$$

De acordo com a equação CP1 e com a Tabela 4, no primeiro componente destacaram-se as variáveis Gols, Totais de chutes, e Totais de chutes na direção do gol as duas sendo correlacionadas com a variável chutes na direção do gol, sendo essa última a melhor representada no primeiro componente. Portanto esse componente pode ser chamado de componente de finalização e qualidade da finalização. Desta forma o jogador, que tiver um valor alto nesse componente significa que o mesmo finalizou diversas vezes na direção do gol e fez vários gols.

Já pela segunda equação e a Tabela 4, no segundo componente destacaram-se as variáveis Números de partidas jogadas, Minutos jogados, Média de chutes por partida e Média de chutes no gol por partida, com a variável Partidas jogadas sendo a mais representada pelo segundo componente. É evidente que existe um contraste entre as variáveis Partidas jogadas e Minutos jogados com as variáveis Média de chutes por partida e Média de chutes no gol por partida analisado pelo angulo de 90 graus que elas formam no gráfico Biplot da Figura 25, ou seja, esse segundo componente pode ser chamado de componente contraste entre partidas jogadas, minutos jogados e média de chutes por partida, média de chutes no gol por partida. Outra observação notada é sobre a variável substituição, percebe-se que ela não é bem representada por nenhuma das duas componentes selecionadas nesses casos é aconselhável estudar as variáveis que se encontram na mesma situação. O gráfico Biplot CP1 $\times$ CP2 sobre as variáveis ajuda na confirmação da interpretação obtida da Tabela 4, quanto mais próxima uma variável for do círculo de correlações, mais representada será pelo componente. As variáveis próximas ao centro do gráfico são menos importantes para os primeiros componentes.

Figura 25 – Biplot CP1xCP2



Fonte: Elaborada pelo autor (2021)

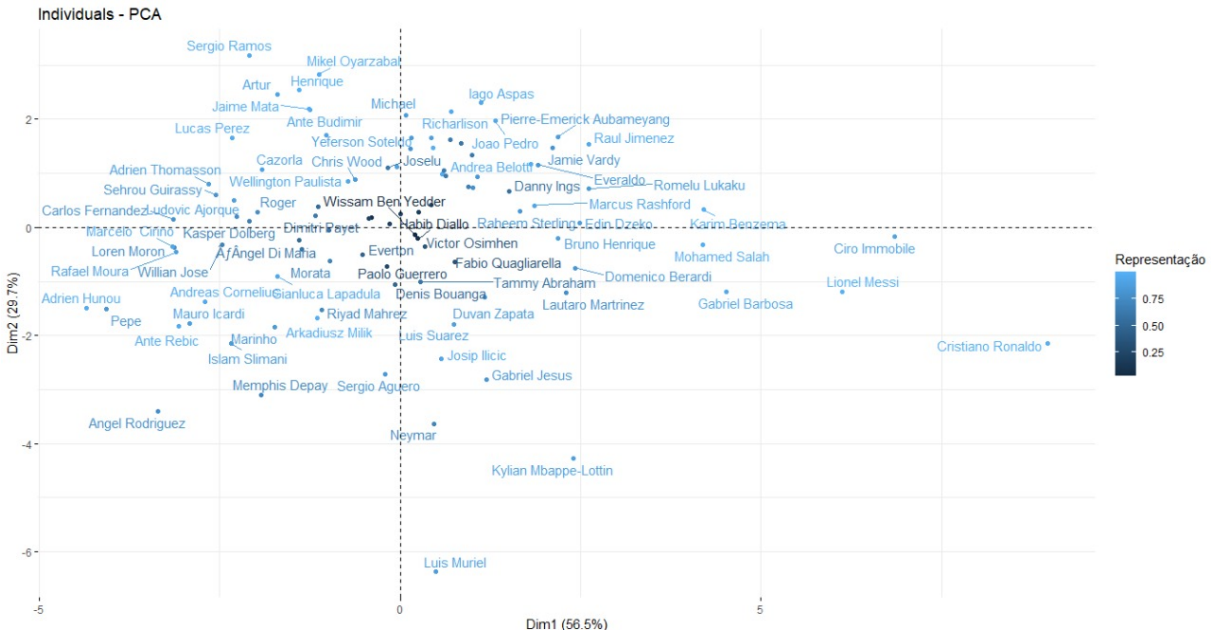
Com a ajuda das equações de CP1 e CP2 são obtidos os scores de cada jogador relacionado aos respectivos componentes que é interpretado com as correlações das variáveis. Os scores são obtidos multiplicando o valor de cada variável obtida dos jogadores com os autovetores, os dois valores do CP1 e CP2 formam coordenadas para um gráfico denominado de Biplot CP1 × CP2 com os indivíduos (jogadores). Este tipo de gráfico é importante pois possibilita visualizar qual jogador se sobressaiu de acordo com os componente dos demais.

Analisando a Figura 26, que é o Biplot CP1xCP2 com os jogadores sobre as variáveis e as equações CP1 e CP2, pode-se concluir que, de acordo com os dados de seus respectivos campeonatos e com a Análise de Componentes Principais, Cristiano Ronaldo, Ciro Immo-

bile, Lionel Messi, Mohamed Salah, Karim Benzema e Gabriel Barbosa possuem maiores desempenhos no quesito de finalizações, qualidade de finalizações e principalmente sobre maiores números de gols pela CP1 se caracterizando como jogadores atacantes e decisivos. Adrien Hounou, Pepe, Angel Rodriguez, Islam Slimani foram alguns dos jogadores que apresentaram menores números de desempenho de finalizações, qualidade das finalizações e gols se caracterizando como jogadores de defesa ou de criação.

E pela CP2, conclui-se que Serigo Ramos, Mikael Oyarzabal, Henrique, Iago Aspas e Richarlison são alguns dos jogadores que tiveram maiores números de ocorrências de jogos disputados, minutos jogados e menores desempenho nas variáveis média de chutes por partidas e média de chutes no gol por partida, se caracterizando como jogadores de defesa ou de criação. Já os jogadores como Sergio Aguero, Memphis Depay, Neymar, Kylain Mbappe e Luis Muriel apresentaram maiores números de média de chutes por jogos e média de chutes no gol por jogos, mas apresentam baixos numeros de jogos disputados e poucos minutos jogados.

Figura 26 – Biplot CP1xCP2 com os jogadores



Fonte: Elaborada pelo autor (2021)

4.1 Análise exploratória dos scores

Nessa seção será apresentado algumas análises exploratórias dos scores obtidos pelos jogadores. Com o auxílio da Tabela 5 e dos gráficos boxplot apresentados pela Figura

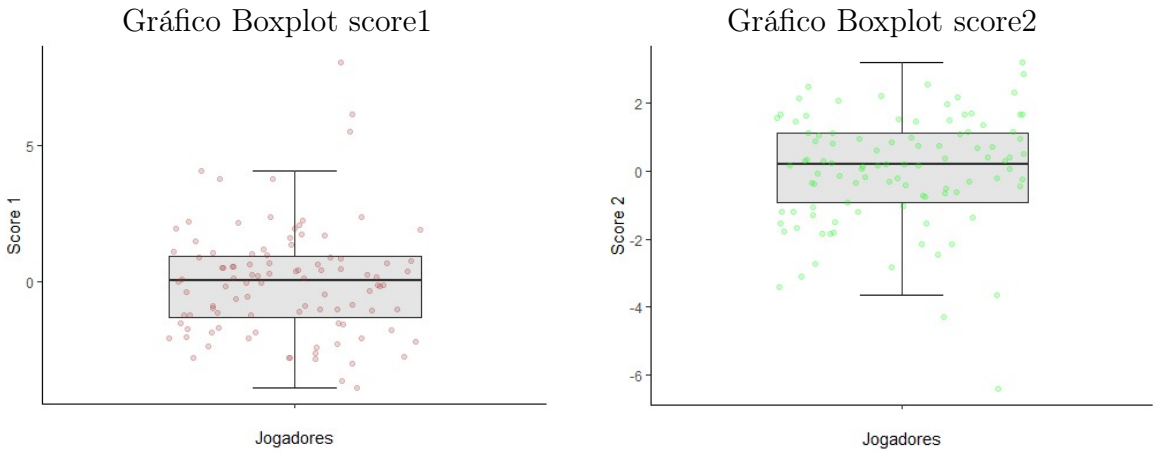
27 é analisado que os scores estão localizados, em sua maioria, perto do valor zero tanto para o componente um quanto para o componente dois. Também é notado que o boxplot do score 1 indica a existência de assimetria à esquerda, com pouca variabilidade e com alguns outliers encontrados acima do limite superior, e para a análise do boxplot do score 2 percebe-se que os dados possuem uma distribuição que também é assimétrica à esquerda e com pouca variabilidade possuindo alguns outliers encontrados abaixo do limite inferior. Logo fica evidente que, de acordo com os dados, a maioria dos valores estão concentrados no zero mas para o score do primeiro componente é possível encontrar, com mais facilidade, jogadores que se destacaram nos quesitos de numero gols e qualidade da finalização do chute e para a segunda componente é evidente que encontre mais jogadores que não possuiram um bom número de partidas jogadas e minutos jogados mas conseguiram um bom resultado em médias de chutes por partida e médias de chutes no gol por partidas.

Tabela 5 – Medidas de dispersão dos scores

	Comp.1	Comp.2
Min.	-3.91341	-6.4042
1st Qu.	-1.32688	-0.9409
Mediana	0.03898	0.1918
Média	0.00000	0.0000
3rd Qu	0.92069	1.1141
Max.	8.07915	3.2013

Fonte: Elaborada pelo autor (2021)

Figura 27 – Gráficos Boxplot dos scores 1 e 2 para comparação



Fonte: Elaborada pelo autor (2021)

4.1.1 Análise exploratória dos scores serparados por campeonatos

Fazendo os mesmos tipos de análises feitas anteriormente só que seraparando os jogadores pelos seus respectivos campeonatos é possível encontrar as seguintes evidências com a ajuda dos gráficos boxplot da Figura 28 e Figura 29:

- Para os jogadores do Campeonato Brasileirão Série A foi percebido que os rendimentos para a componente de finalização e qualidade da finalização não foram muito satisfatório pois a maioria deles possuiram um score infeiror a zero, porém para os scores da componente de contraste entre partidas jogadas, minutos jogados e média de chutes por partida, média de chutes no gol por partida, esses mesmo jogadores possuiram um bom desempenho se caracterizando por jogadores que mais disputaram partidas pelo seu time.

- Para os jogadores da France Ligue é notado que os rendimentos para a componente da finalização e qualidade de finalização não foram satisfatórios uma vez que a maioria deles possuiram scores inferiores a zero, e para os scores da componente de contraste entre partidas jogadas, minutos jogados e média de chutes por partida, média de chutes no gol por partida, esses mesmos jogadores tiveram scores bastante negativos se caracterizando como jogadores que não possuem finalizações de qualidade, não disputam muitas partidas pelo seu time mas tem um boa média de chutes por partidas jogadas.

- Já os jogadores da La Liga é observado que os rendimentos para a componente de finalização e qualidade da finalização não foram satisfatórios pois a maioria deles estão com scores inferiores a zero, porém para os scores da componente de contraste entre partidas jogadas, minutos jogados e média de chutes por partida, média de chutes no gol por partida, os mesmos apresentam um bom desempenho se caracterizando como jogadores que mais disputaram partidas pelo seu time.

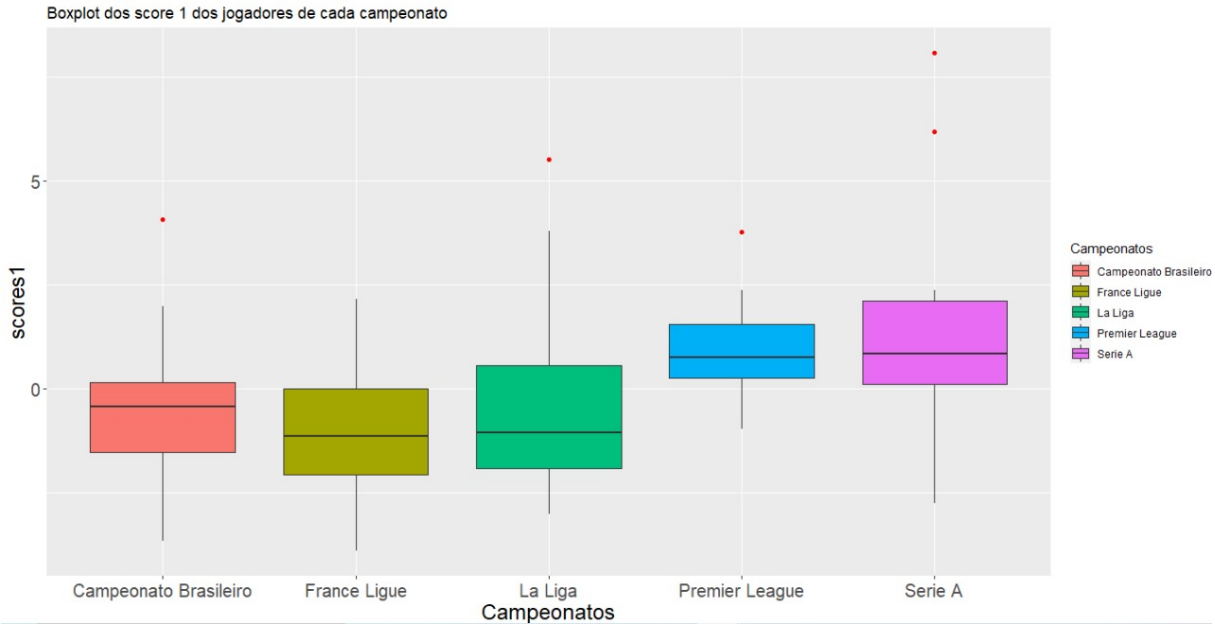
- Os jogadores da Primiere League apresentaram um bom desempenho para a componente da finalização e qualidade de finalização. Foram satisfatórios uma vez que a maioria deles possuiram scores superiores a zero, e para os resultados da componente de contraste entre partidas jogadas, minutos jogados e média de chutes por partida, média de chutes no gol por partida, eles também tiveram, em sua maioria,

scores superiores a zero se caracterizando como jogadores que disputaram bastante partidas e minutos para seus respectivos times e que tiveram boas finalizações com muitos gols.

- E para os atletas da Série A da Italia foi observado um bom desempenho na componente de finalização e qualidade da finalização pois seus respectivos scores encontram-se maiores que zero, ja no componente de contraste a maioria dos jogadores não possuem scores maiores que zero se caracterizando como jogadores que não disputam muitos jogos pelos seus times mas possuem muitas finalizações de qualidade e muitos gols.

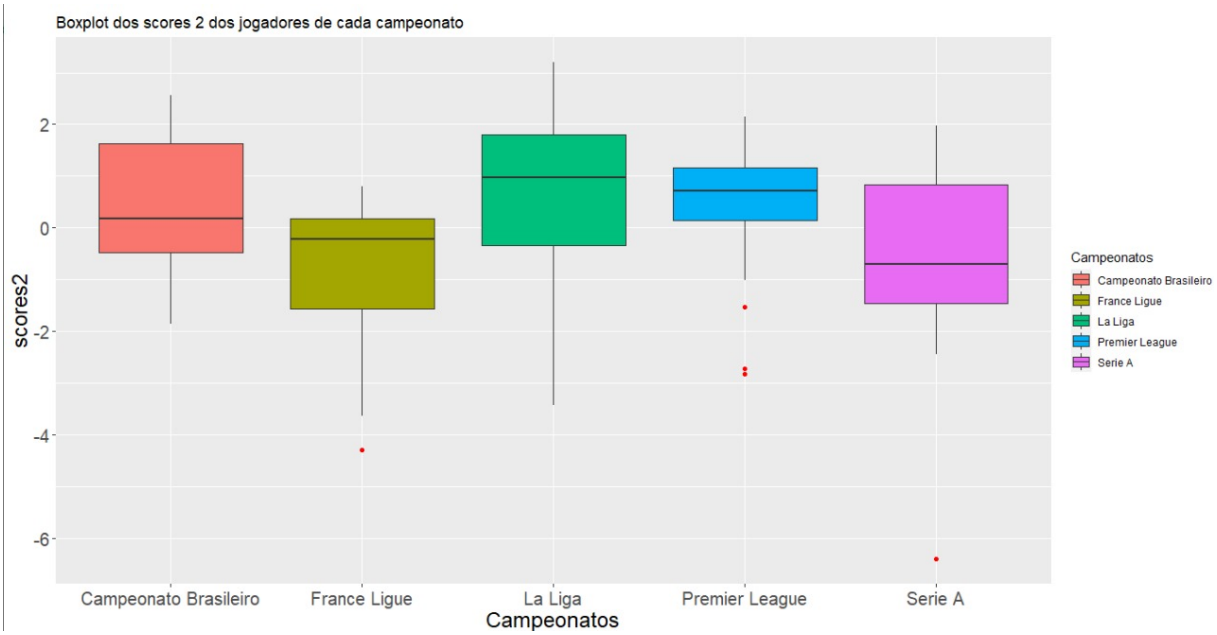
De uma maneira geral é percebido que os jogadores do Campeonato Brasileiro Serie A, La Liga e Primiere League são caracterizados por possuirem mais jogos e minutos jogados do que as demais competições. Já os atletas da Primiere League e Série A da Itália tem a característica de jogadores com boas finalizações e elevado número de gols comparados com os demais campeonatos.

Figura 28 – Gráficos boxplot do score 1 para comparação separados por camponato



Fonte: Elaborada pelo autor (2021)

Figura 29 – Gráficos boxplot do score 2 para comparação separados por camponato

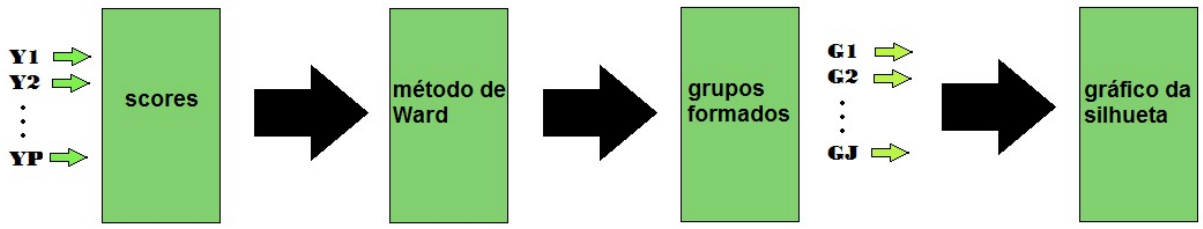


Fonte: Elaborada pelo autor (2021)

5 Aplicação da análise de agrupamentos pelo método de Ward e resultados

O método de agrupamento será feito com o objetivo de formar grupos de jogadores que possuem as mesmas características, com a medida de distância sendo a Distância Euclidiana explicada no referencial teórico. Os passos são mensurados na Figura 30.

Figura 30 – Passos para realização e interpretação dos agrupamentos



Fonte: Elaborada pelo autor (2021)

5.1 Agrupando pelos score das componentes 1 e 2

Esses agrupamentos serão feitos levando em consideração os scores das componentes 1 e 2 encontrados anteriormente como variaveis de entrada, ou seja seram agrupados dada

a proximidade de suas características em relação aos valores dos scores.

Ao aplicar o método é observado a qualidade dos agrupamentos dado pelo gráfico da silhueta que nos proporciona descobrir se, de fato, existe uma forte estrutura.

Os valores da silhueta são dados na Tabela 6 junto com o número de grupos formados.

Tabela 6 – Médias por grupos e Média geral de cada agrupamento

Ngrupos	m. por grupo	m. geral
2grupos	G1=0.4, G2=0.2	0.3
3grupos	G1=0.2, G2=0.5, G3=0.3	0.3
4grupos	G1=0.3, G2=0.5, G3=0.4, G4=0.25	0.35
5grupos	G1=0.27, G2=0.44, G3=0.36, G4=0.46, G5=0.14	0.33
6grupos	G1=0.25, G2=0.44, G3=0.3, G4=0.46, G5=0.44, G6=0.38	0.36
7grupos	G1=0.2, G2=0.44, G3=0.5, G4=0.7, G5=0.4, G6=0.35, G7=0.4	0.38

Fonte: Elaborada pelo autor (2021)

Portanto ao interpretar a Tabela 6 nota-se que é encontrada uma estrutura superficial de agrupamento e é aconselhável o uso de outros métodos para esses dados. O maior valor da silhueta foi encontrado no agrupamento com 7 grupos com uma média geral de 0.38.

Figura 31 – Grupos formados pelo Método de Ward com duas componentes

Fonte: Elaborada pelo autor (2021)

Os agrupamentos estão disponibilizados na Figura 31. Pode-se concluir que os grupos

G1 e G5 são os que mais tem jogadores alocados, já os grupos G2, G4 e G7 são os que possuem menos atletas. Calculando as médias de cada grupo levando em consideração os scores é possível interpretar as características de cada grupo.

Tabela 7 – Média das duas componentes de cada grupo

Grupo	Comp.1	Comp.2
G1	-0.08989576	-0.2073458
G2	5.23514504	-0.7865796
G3	-2.29279061	0.3013564
G4	-1.26636869	2.4547819
G5	1.11725841	1.2065253
G6	-2.35056629	-1.9819932
G7	0.73999323	-3.7248972

Fonte: Elaborada pelo autor (2021)

Observando as médias calculadas de cada grupo referente aos scores é possível analisar os seguintes fatos:

- No grupo G1 estão alocados os atletas que não foram bem representados por nenhuma das componentes encontradas, ou seja, são aqueles pontos próximo ao zero no gráfico Biplot.

- No agrupamento G2 estão os jogadores que mais se destacaram na primeira componente, logo foram os jogadores que mais fizeram gol e que mais finalizaram durante o comapeonato. e são os localizados mais a direita no gráfico Biplot.

- O grupo G3 são os atletas que não tiveram um bom desempenho no componente de finalização e qualidade das finalizações tendo poucos chutes e poucos gols feitos. Esses jogadores também não foram bem representados pelo componente de contraste entre minutos jogados, partidas jogadas e média de chutes por jogo. Logo são os jogadores que estão mais ao centro e a esquerda do gráfico Biplot.

- Para o grupo G4 os jogadores alocados são aqueles que obtiveram um valor baixo no quesito de número de gols e chutes no gol mas foram o que mais disputaram partidas e tiveram mais minutos jogados.

- No agrupamento G5 são os jogadores que alem de possuirem um número razoavelmente bom no quesito de finalizações e gols também tiveram um número bom de partidas e minutos jogados.

- Para o grupo G6 os jogadores alocados possuem altos números de médias de chutes por jogo e poucos minutos disputados, também possuem baixa qualidade de finalizações tendo poucos gols devido seus baixos scores nos respectivos componentes.

- Neste último agrupamento G7 sobraram os jogadores que possuem um número razoável na qualidade da finalização e no números de gols, também possuem poucos jogos disputados mas são os jogadores que tem maiores valores de chutes médios por partidas criando bastante oportunidades de gols. Esses jogadores são os que se encontram mais abaixo no gráfico Biplot.

Portanto é possível afirmar que mesmo com uma estrutura razoável de agrupamento os grupos não fogem do que foi visualizado e interpretado nos gráficos Biplot.

5.2 Agrupando pelos dados padronizados

Fazendo o agrupamento pelo método de Ward com os dados padrozinados sem a aplicação do modelo de componentes principais temos que os valores obtidos do gráfico da Silhueta são próximos de zero indicando que não foram encontradas fortes indícios de grupos formandos, podendo concluir que os jogadores possuiram desempenho parecido ou que esse método não é apropriado para esse tipo de classificação.

6 Conclusão

A aplicação do modelo de Componentes Principais foi efetiva, uma vez que, conseguiu diminuir o número de dimensões sem muita perda de informação. Transformando oito variáveis correlacionadas em duas novas variáveis independentes com cerca de 84% da variabilidade total.

Este modelo proporcionou duas componentes, a primeira se caracterizando por ser a componente de Finalizações e Qualidade de finalizações. Já a segunda se caracterizou por ser uma componente de contraste entre Partidas jogadas, Minutos jogados e Média de chutes por partidas, Média de chutes no gol por partidas. Esse segundo componente pode parecer contraditório pois quanto mais tempo um jogador tem mais chances o mesmo tem de chutar no gol, porém no futebol, principalmente no Brasil, existe uma tática de jogo em que o time começa o jogo atacando o máximo possível e quando executa um gol coloca todo o time na defesa, sendo assim o número de finalizações vai decaindo ao longo do tempo e nesse banco de dados essa informação persistiu.

Outra interpretação dada foi da importância da variável Substituição que não foi bem representada por nenhum dos dois componentes, contribuindo pouco para a análise feita destes dados.

Também foi obtido scores do rendimento de cada jogador levando em conta os componentes criados das variáveis escolhidas e com o auxilio dos gráficos Biplot foi verificado quais jogadores se sobressairam dos demais de acordo com as características encontradas. Depois foram realizadas as comparações dos rendimentos por campeonato e observou-se que os atletas do Campeonato Brasileiro Serie A, La Liga e Primiere League tiveram uma maior ocorrência de jogos disputados e minutos jogados, já os atletas da Série A da Itália e da Premier League foram os que mais se destacaram em quesito de qualidade da

finalização e gols feitos dos demais campeonatos comparados.

Nos agrupamentos formados pelos componentes 1 e 2 que representam cerca de 84% da variabilidade total dos dados é possível analisar que o método de Ward não conseguiu obter uma boa estrutura dos grupos possuindo baixos valores no gráfico da silhueta, sendo o maior valor dado por 0.38 com 7 grupos formados sendo que os grupos G1 e G7 os que possuem mais jogadores e os grupos G2, G4 e G7 os que tem menos atletas.

Já fazendo os agrupamentos com todos os dados padronizados o desempenho no gráfico da Silhueta é menor do que com os dois primeiros componentes encontrado pelo método de componentes principais.

O intuito do trabalho foi de fazer as análises com os principais jogadores, logo está incluso na base dados atletas de diversas posições como atacantes, meias, zagueiros, laterias. Se o foco fosse dado para coletar os dados apenas de uma única posição os resultados poderiam ser diferentes principalmente nos agrupamentos efetuados.

Referências

AFFONSO, J. J.; TACHIBANA, V. M. Análise multivariada aplicada em dados de futebol–campeonato brasileiro de 2011. 2011.

FERNANDES, J. L. Futebol: ciência, arte ou-sorte!: treinamento para profissionais-alto rendimento: preparação física, técnica, tática e avaliação. *São Paulo: EPU*, 1994.

GODIK, M. Preparacao dos futebolistas de alto nivel. *Rio de Janeiro, Editora Grupo Palestra Sport*, 1996.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. de O. Análise de componentes principais: resumo teórico, aplicação e interpretação. *E&S Engineering and science*, v. 5, n. 1, p. 83–90, 2016.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, Springer, v. 23, n. 3, p. 187–200, 1958.

LEONCINI, M. P. *Entendendo o negócio futebol: um estudo sobre a transformação do modelo de gestão estratégica nos clubes de futebol.* Tese (Doutorado) — Universidade de São Paulo, 2001.

MAGNUSSON, W. E.; MOURÃO, G. Estatística sem matemática. Editora Planta, 2003.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, Elsevier, v. 20, p. 53–65, 1987.

SANDANIELO, V. L. *Emprego de técnicas estatísticas na construção de índices de desenvolvimento sustentável aplicados a assentamentos rurais.* 2008. 159 f. Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agronômicas de Botucatu, 2008. Disponível em: http://hdl.handle.net/11449/101703. Acesso em: 23 set. 2021.

SEIDEL, E. J. et al. Comparação entre o método ward e o método k-médias no agrupamento de produtores de leite. *Ciência e Natura*, v. 30, n. 1, p. 07–15, 2008.

SILVA, T. M. *Análise de Componentes Principais e Suas Aplicações*. 2018. Monografia (Bacharel em Estatística), UFRN (Universidade Federal do Rio Grande do Norte), Natal, Brazil.

SILVESTRE, M. R. *Notas de aula de Análise Multivariada II.* . 2021. Presidente Prudente: [s. n.], 2021. Apostila disponível no ambiente virtual de apoio disciplina Análise Multivariada II da FCT/Unesp.

VALE, M. N. do. *Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos*. Tese (Doutorado) — PUC-Rio, 2005.

VENDITE, C. C.; VENDITE, L. L.; MORAES, A. d. Scout no futebol: uma ferramenta para a imprensa esportiva. In: *XXVIII Congresso Brasileiro de Ciências da Comunicação*. [S.l.: s.n.], 2005. v. 1.

VENDITE, L. L.; MORAES, A. C. de; VENDITE, C. C. Scout no futebol: uma análise estatística. *Conexões*, v. 1, n. 2, p. 183–194, 2003.

VICINI, L. *Análise Multivariada: da teoria à prática.* 2005. Monografia (Especialização) - Universidade Federal de Santa Maria, Centro de Ciências Naturais Exatas, Programa de Pós-Graduação em Estatística e Modelagem Quantitativa, RS, 2005. Disponível em: https://repositorio.ufsm.br/handle/1/18058. Acesso em: 23 set. 2021.