Trabalho Final: Premier League

Rodrigo Malta Esteves

16/10/2021

1) Entendimento do problema

O futebol é o esporte coletivo mais popular e praticado não só no Brasil, mas em grande parte do mundo, com grande variedade de análises possíveis para diferentes aspectos que envolvem seu universo. Nesse trabalho estamos interessados em estudar alguns deles.

Já com quase 200 anos de existência, o entendimento sobre como o esporte deve ser praticado mudou constantemente desde o seu surgimento e permanece em constante transformação. Seja em esquema de jogo, preparação física ou funções de cada jogador dentro de campo, cada nova partida representa não só um confronto técnico entre os jogadores das duas equipes, mas uma disputa entre as diferentes filosofias adotadas por cada um dos lados.

Uma das características que sofreu mudanças mais claras durante o período de existência do esporte foram as posições dos jogadores dentro de campo. Algumas posições populares simplesmente pararam de existir, enquanto outras se tornaram cada vez menos populares e foram substituidas dados os novos tipos de esquemas táticos que foram surgindo. Um exemplo são os 'Liberos' ('Sweeper' em inglês), populares nos anos 70 e 80 e representado por jogadores que marcaram época como Franz Beckenbauer e Franco Baresi, são vistos apenas em raras aparições, uma vez que existem em esquemas de jogo que não são mais populares, e o entendimento do que esse jogador faz se tornou confuso para o espectator padrão. Outro exemplo é o 'Segundo atacante' ('Second Striker') que foi resignificado para outras categorias como 'Ponta', 'Extremo' ou 'Falso 9'. Hoje, é comum vermos técnicos com ideais de jogo mais modernas ignorarem as funções pré-definidas dos jogadores que estão no elenco e atribuir a eles novas responsabilidade em campo.

O objetivo desse trabalho é ser capaz de classificar os jogadores, a partir de suas características físicas e técnicas, em clusters que identifiquem padrões sobre suas capacidades para além do que é a sua posição de atuação em campo conhecida e, com isso, determinar se é possível englobá-los em novos grupos.

2) Entendimento dos dados

Os dados referentes às características físicas e técnicas dos atletas são de grande interesse para essa análise. Para isso, escolhi trabalhar com os dados dos jogadores do Fifa 20, referente à temporada 2019-2020, uma vez que esses dados são constantemente atualizados durante a temporada e não estão concluídos ainda para o ano de 2021.

Contendo informações sobre 18,278 jogadores com 104 atributos, o arquivo está disponibilizado pelo kaggle, através do link https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv

Por fim, optaremos por escolher jogadores apenas de uma das ligas nacionais, uma vez que o tipo de jogo executado e nível de cobrança em diferentes paises pode dificultar o entendimento do problema. Utilizaremos apenas jogadores da Premier League (liga do futebol inglês) e apenas os titulares dos times. Além disso, os goleiros foram excluídos da análise por serem um grupo muito homogêneo e que não compartilham características com os demais jogadores de linha.

Descrição das variáveis

O conjunto de dados contém informações descritivas sobre os jogadores (nome, idade, altura, etc) e atributos referentes às capacidades de cada um deles durante o jogo. Esses atributos são divididos em sete categorias:

ataque, habilidade, movimento, força, mentalidade, defesa e atuação como goleiro. Como estamos observando apenas jogadores de linha, essa última categoria será ignorada nas análises.

Os atributos funcionam como notas de 0 até 100. Quanto maior o atributo, maior o peso dele no perfil do jogador.

A seguir uma descrição sobre cada um deles. Para uma descrição mais detalhada consultar https://fifauteam.com/fifa-20-attributes-guide/

- Variáveis descritivas do jogador
  - sofifa_id: ID do jogador
  - short_name: Nome do jogador
  - age: Idade
  - height_cm: Altura em centímetros
  - weight_kg: Peso em kilogramas
  - nacionality: Nacionalidade
  - club: Clube em que atua
  - overall: Nota média do jogaddor
  - potential: Possível nota que o jogador pode alcançar
  - value_eur: Valor do passe em euros
  - wage_eur: Valor do salário do jogador
  - preferred_foot: Pé de preferência (destro ou canhoto)
  - international_reputation: Reputação internacional
  - player_positions: Posições dentro de campo em que o jogador pode jogar
  - team_position: Posição em campo que ocupa no time
  - weak_foot: Quão bem utiliza o pé oposto
  - skill_moves: Quão capaz é de executar movimentos técnicos
  - work_rate: Taxa de esforço que o jogador impõe no ataque e na defesa
  - release_clause_eur: Multa rescisória
- Atributos genéricos do jogador
  - pace: Ritmo
  - shooting: Finalização
  - passing: Passe
  - dribbling: Drible
  - defending: Defesa
  - physic: Físico
- Atributos de ataque do jogador
  - finishing: Abilidade do jogador em marcar gols
  - heading_accuracy: Precisão do jogador quando usando a cabeça para passar, chutar ou limpar a bola
  - short_passing: Precisão do jogador para passes curtos
  - volleys: Abilidade do jogador em performar voleios
- Atributos de habilidade do jogador
  - dribblings: Abilidade do jogador em carregar a bola e passar pelo adversário
  - curve: Abilidade do jogador em fazer a bola curvar quando passando ou chutando
  - fk_accuracy: Precisão do jogador batendo falta
  - long_passing: Precisão do jogador em passes longos
  - ball_control: Abilidade do jogador em controlar a bola
- Atributos de mobilidade do jogador
  - acceleration: Aceleração do jogador
  - sprint_speed: Define a velocidade do jogador correndo

  - agility: Determina quão rapidamente e graciosamente o jogador é em controlar a bola
  - reactions: A velocidade de reação de um jogador às situações ocorrendo ao redor dele
  - balance: A capacidade do jogador em se manter de pé e estável enquanto correndo

- Atributos de força do jogador
  - shot_power: Força do jogador no chute
  - jumping: Abilidade e qualidade do jogador em pular junto com outro jogador
  - stamina: Abilidade do jogador em sustentar esforço físico e mental prolongado
  - strength: A qualidade ou estado de ser fisicamente forte
  - long_shots: Precisão dos chutes de longa distância

- Atributos de mentalidade do jogador
  - aggression: Nível de agressão do jogador
  - interceptions: Capacidade do jogador em interceptar a bola
  - positioning: Define quão bem um jogador é capaz de performar o posicionamento no campo
  - vision: A consciência mental do jogador sobre o posicionamento dos companheiros de equipe, com a finalidade de passar a bola
  - penalties: Precisão do jogador em batidas de pênalti
  - composure: A capacidade de um jogador em se manter calmo e controlar suas frustrações

- Atributos de defesa do jogador
  - marking: Capacidade do jogador em marcar o jogador adversário
  - standing_tackle: Abilidade de performar uma dividida em pé
  - sliding_tackle: Abilidade de performar uma dividida no chão (carrinho)

- Atributos específicos dos goleiros
  - diving: Abilidade do goleiro em fazer a defesa enquanto no ar
  - handling: Frequência que o goleiro pega a bola ao invés de espalmar ou não segurar ela
  - kicking: Comprimento e precisão das reposições do goleiro, tanto com as mãos quanto com os pés
  - positionings: Abilidade do goleiro em se posicionar corretamente quando defendendo chutes. Também afeta como o goleiro responde aos cruzamentos
  - reflexes: Agilidade do goleiro em fazer defesas

A seguir podemos ver a média, o mínimo, o máximo, o primeiro e o terceiro quantil de cada atributo. Chama a atenção que, apesar de variar entre 0 e 100, nenhuma das observações atinjam esses valores extremos. Além disso, os atributos do grupo referente aos goleiros são muito baixos para todos os casos de jogadores não-goleiros.

A idade (age), altura (height_cm) e o peso (weight_kg) são as variáveis que descrevem cada jogador fisicamente. Como estão em escalas diferentes dos demais atributos, deverão ser normalizadas posteriormente. Os demais atributos estão todos na mesma escala e não demandam nenhuma forma de manipulação.

Atributos físicos

```
##       age          height_cm       weight_kg
##  Min.   :19.00   Min.   :163.0   Min.   : 59.0
##  1st Qu.:24.00   1st Qu.:177.0   1st Qu.: 70.0
##  Median :26.00   Median :181.5   Median : 76.0
##  Mean   :26.13   Mean   :182.0   Mean   : 76.2
##  3rd Qu.:28.00   3rd Qu.:188.0   3rd Qu.: 81.0
##  Max.   :35.00   Max.   :199.0   Max.   :100.0
```

Atributos genéricos

```
##       pace          shooting        passing        dribbling
##  Min.   :31.00   Min.   :22.00   Min.   :43.00   Min.   :34.00
##  1st Qu.:62.00   1st Qu.:51.00   1st Qu.:64.00   1st Qu.:68.25
##  Median :71.00   Median :65.00   Median :72.00   Median :74.00
##  Mean   :69.92   Mean   :62.49   Mean   :70.16   Mean   :73.18
##  3rd Qu.:78.00   3rd Qu.:74.00   3rd Qu.:77.00   3rd Qu.:80.00
##  Max.   :94.00   Max.   :91.00   Max.   :92.00   Max.   :92.00
##    defending         physic
##  Min.   :24.00   Min.   :44.00
##  1st Qu.:51.00   1st Qu.:68.00
##  Median :71.00   Median :74.00
##  Mean   :64.44   Mean   :72.39
##  3rd Qu.:78.00   3rd Qu.:78.00
##  Max.   :90.00   Max.   :88.00
```

Atributos de ataque

```
##     crossing       finishing     heading_accuracy short_passing
##  Min.   :20.00   Min.   :19.00   Min.   :30.00    Min.   :58.00
##  1st Qu.:60.00   1st Qu.:46.00   1st Qu.:58.25    1st Qu.:71.00
##  Median :69.00   Median :61.00   Median :68.00    Median :76.00
##  Mean   :66.62   Mean   :59.19   Mean   :66.46    Mean   :75.54
##  3rd Qu.:77.00   3rd Qu.:73.00   3rd Qu.:76.00    3rd Qu.:80.00
##  Max.   :93.00   Max.   :94.00   Max.   :87.00    Max.   :92.00
##     volleys
##  Min.   :16.00
##  1st Qu.:44.00
##  Median :60.00
##  Mean   :56.81
##  3rd Qu.:70.00
##  Max.   :86.00
```

Atributos de habilidade

```
##    dribblings        curve        fk_accuracy     long_passing
##  Min.   :24.00   Min.   :22.00   Min.   :20.00   Min.   :35.00
##  1st Qu.:67.00   1st Qu.:56.00   1st Qu.:45.00   1st Qu.:63.00
##  Median :74.00   Median :68.00   Median :60.00   Median :69.50
##  Mean   :72.44   Mean   :64.62   Mean   :57.78   Mean   :69.25
##  3rd Qu.:80.00   3rd Qu.:76.00   3rd Qu.:70.75   3rd Qu.:76.00
##  Max.   :92.00   Max.   :91.00   Max.   :89.00   Max.   :91.00
##   ball_control
##  Min.   :47.00
##  1st Qu.:72.00
##  Median :77.00
##  Mean   :75.84
##  3rd Qu.:81.00
##  Max.   :92.00
```

Atributos de movimentação

```
##   acceleration    sprint_speed      agility        reactions
##  Min.   :31.00   Min.   :30.00   Min.   :29.00   Min.   :62.00
##  1st Qu.:61.00   1st Qu.:61.25   1st Qu.:62.00   1st Qu.:72.00
##  Median :71.00   Median :72.00   Median :71.00   Median :76.00
##  Mean   :69.71   Mean   :70.07   Mean   :69.47   Mean   :76.62
##  3rd Qu.:79.00   3rd Qu.:78.00   3rd Qu.:78.00   3rd Qu.:81.00
##  Max.   :96.00   Max.   :95.00   Max.   :95.00   Max.   :93.00
##     balance
##  Min.   :26.00
##  1st Qu.:61.00
##  Median :71.00
##  Mean   :68.81
##  3rd Qu.:78.00
##  Max.   :96.00
```

Atributos de força

```
##    shot_power       jumping         stamina         strength
##  Min.   :25.00   Min.   :30.00   Min.   :34.00   Min.   :30.00
##  1st Qu.:64.25   1st Qu.:64.00   1st Qu.:70.00   1st Qu.:64.25
##  Median :74.00   Median :72.00   Median :77.00   Median :72.00
##  Mean   :71.17   Mean   :69.83   Mean   :76.15   Mean   :70.79
##  3rd Qu.:80.00   3rd Qu.:78.00   3rd Qu.:82.00   3rd Qu.:78.00
##  Max.   :91.00   Max.   :92.00   Max.   :97.00   Max.   :95.00
##    long_shots
##  Min.   :14.0
##  1st Qu.:52.0
##  Median :67.0
##  Mean   :62.2
##  3rd Qu.:74.0
##  Max.   :90.0
```

Atributos de mobilidade

```
##    aggression    interceptions    positioning        vision
##  Min.   :31.00   Min.   :18.00   Min.   :22.00   Min.   :25.00
##  1st Qu.:67.25   1st Qu.:49.00   1st Qu.:57.00   1st Qu.:64.00
##  Median :75.00   Median :72.00   Median :68.00   Median :72.00
##  Mean   :71.97   Mean   :63.96   Mean   :65.92   Mean   :69.35
##  3rd Qu.:81.00   3rd Qu.:78.00   3rd Qu.:76.00   3rd Qu.:78.00
##  Max.   :92.00   Max.   :92.00   Max.   :93.00   Max.   :94.00
##    penalties       composure
##  Min.   :24.00   Min.   :54.00
##  1st Qu.:51.25   1st Qu.:71.00
##  Median :60.00   Median :76.00
##  Mean   :60.04   Mean   :75.53
##  3rd Qu.:71.75   3rd Qu.:80.00
##  Max.   :91.00   Max.   :92.00
```

Atributos defensivos

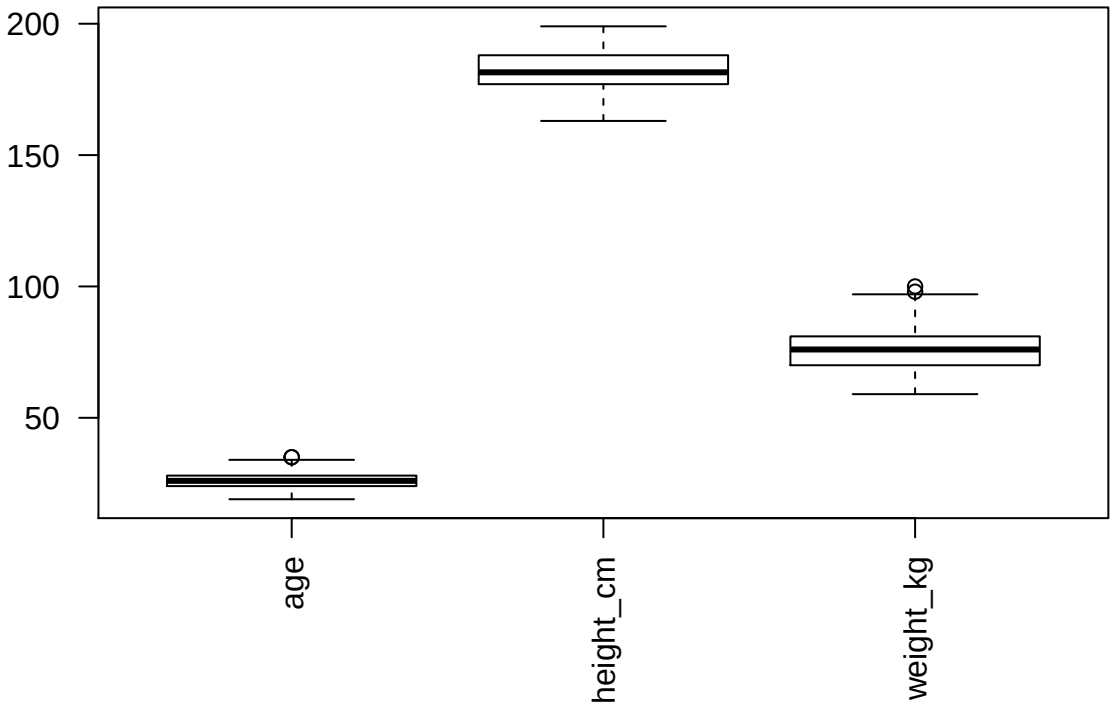
```
##     marking      standing_tackle sliding_tackle
##  Min.   :22.00   Min.   :16.0    Min.   :15.00
##  1st Qu.:50.00   1st Qu.:54.0    1st Qu.:42.50
##  Median :71.00   Median :73.0    Median :70.00
##  Mean   :64.16   Mean   :65.3    Mean   :61.13
##  3rd Qu.:77.00   3rd Qu.:79.0    3rd Qu.:77.00
##  Max.   :91.00   Max.   :92.0    Max.   :88.00
```

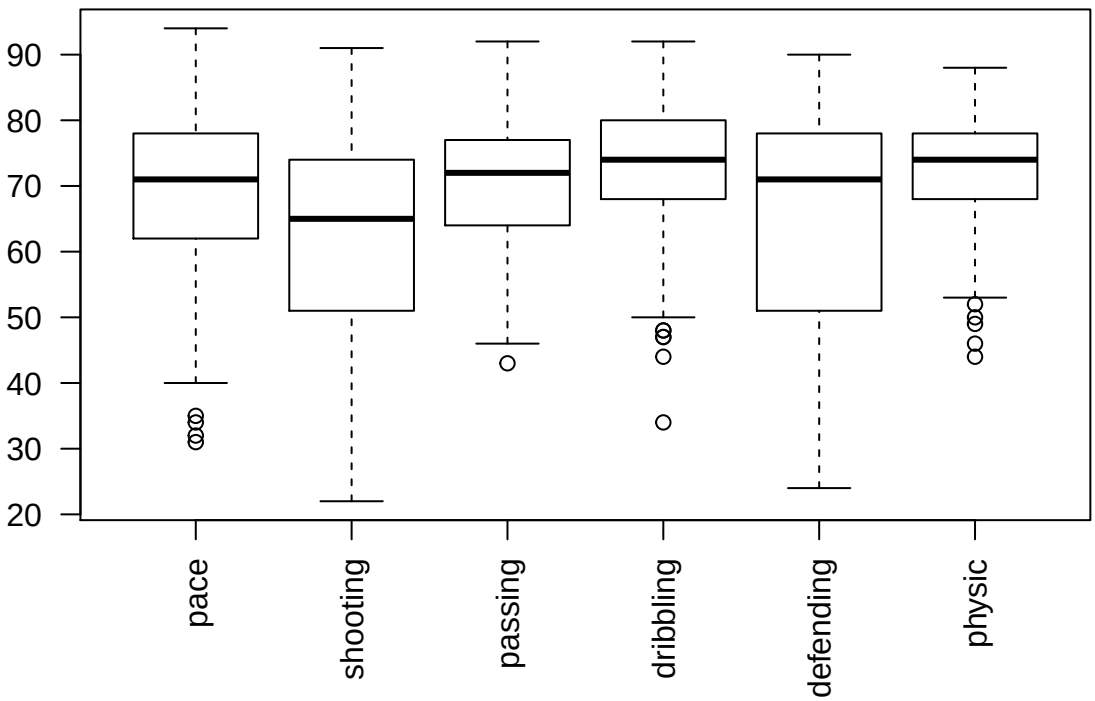
Atributos dos goleiros

```
##      diving        handling        kicking       positionings
##  Min.   : 4.0   Min.   : 5.00   Min.   : 2.00   Min.   : 3.00
##  1st Qu.: 8.0   1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 7.00
##  Median :11.0   Median :11.00   Median :11.00   Median :10.00
##  Mean   :10.8   Mean   :10.85   Mean   :10.63   Mean   :10.13
##  3rd Qu.:14.0   3rd Qu.:13.75   3rd Qu.:13.75   3rd Qu.:13.00
##  Max.   :16.0   Max.   :16.00   Max.   :16.00   Max.   :16.00
##     reflexes
##  Min.   : 3.00
##  1st Qu.: 8.00
##  Median :10.00
##  Mean   :10.07
##  3rd Qu.:12.00
##  Max.   :16.00
```

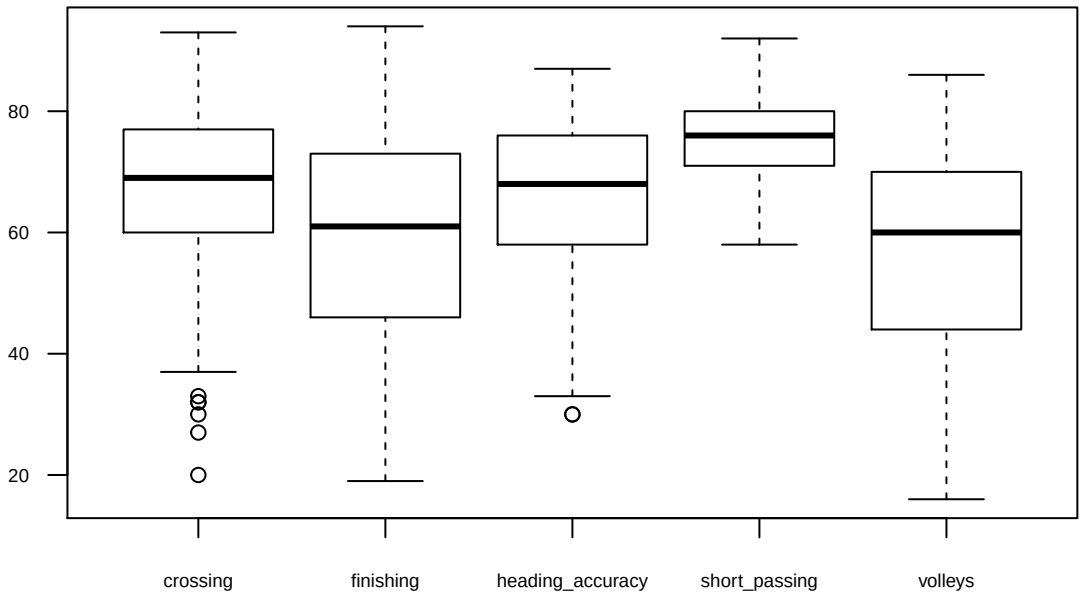
Distribuição de cada atributo

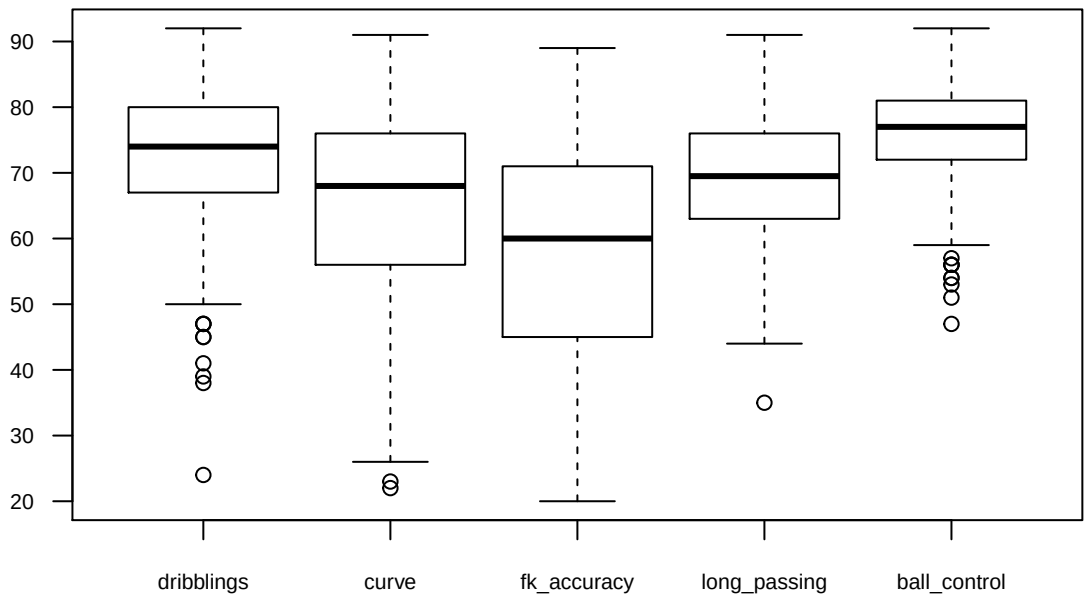
Como esperado, os plots referentes ao peso, altura e idade estão estranhos dada a diferença de escala entre eles e os demais atributos. Esses dados serão colocados na escala de interesse a seguir.

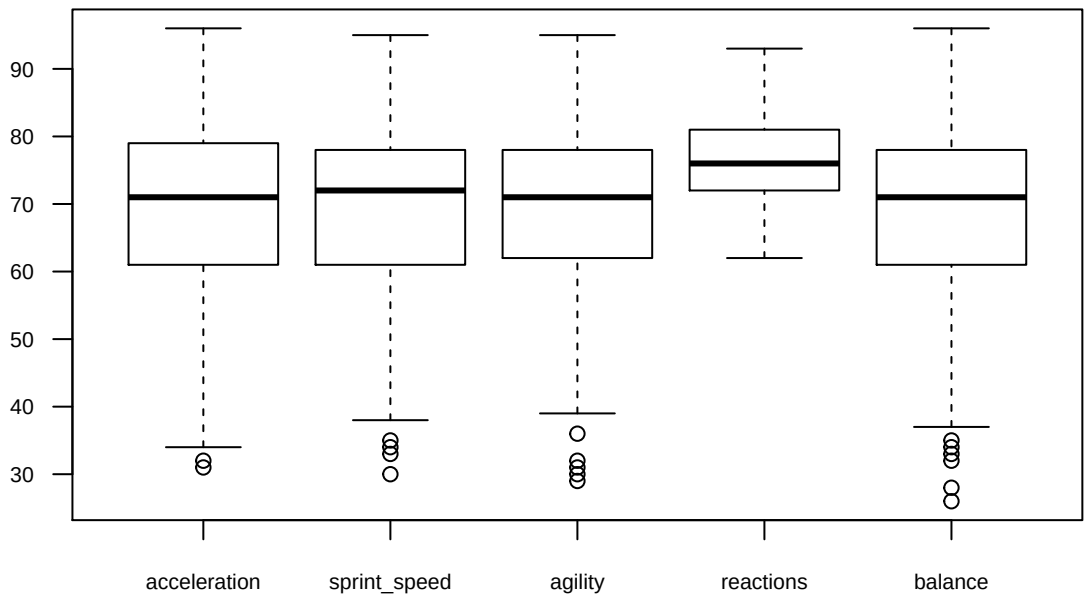

Idade, altura e peso
200
150
100
50
age
height_cm
weight_kg

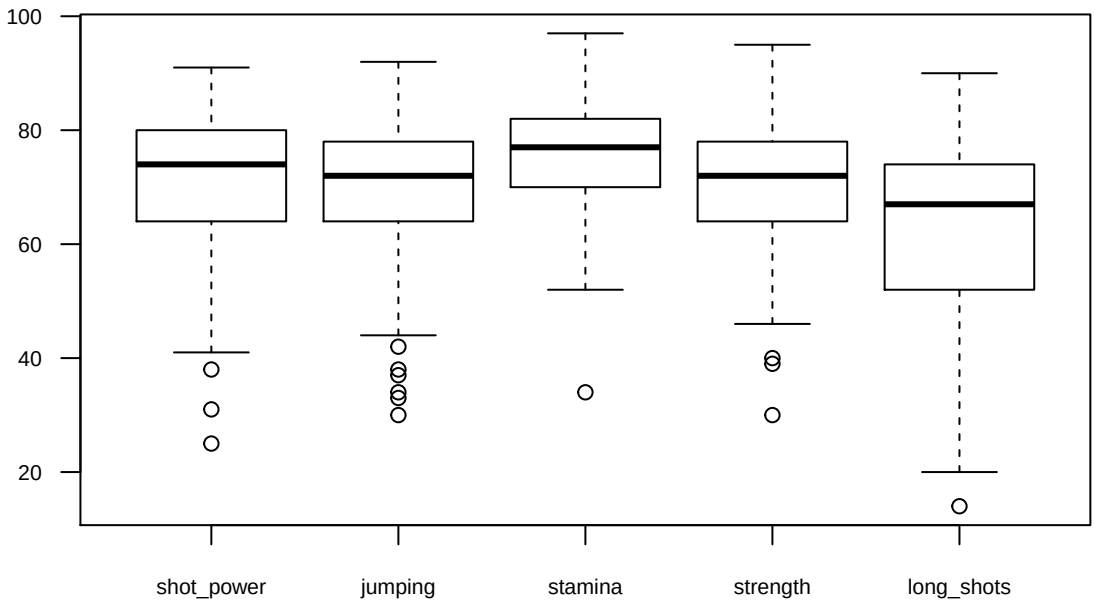

Atributos genéricos
90
80
70
60
50
40
30
20
pace
shooting
passing
dribbling
defending
physic

Atributos de ataque

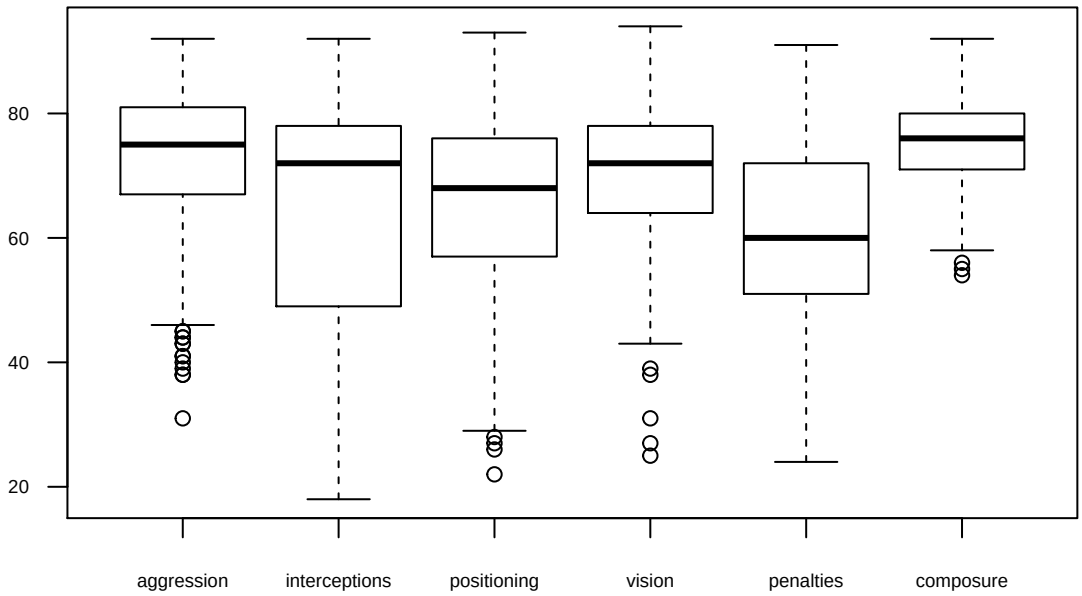


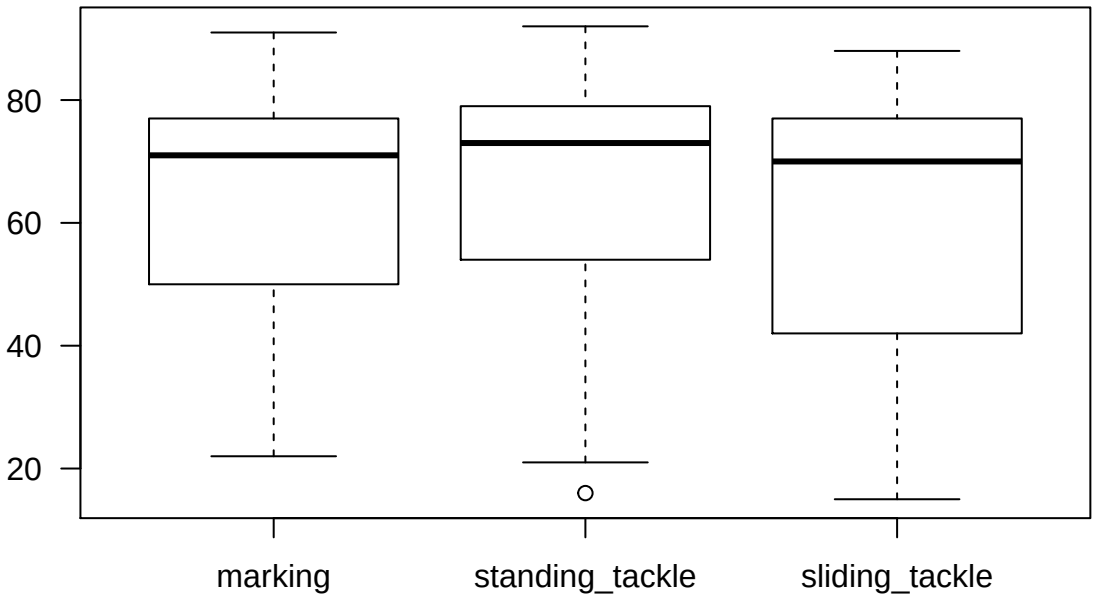

Atributos de habilidade
90
80
70
60
50
40
30
20
dribblings
curve
fk_accuracy
long_passing
ball_control

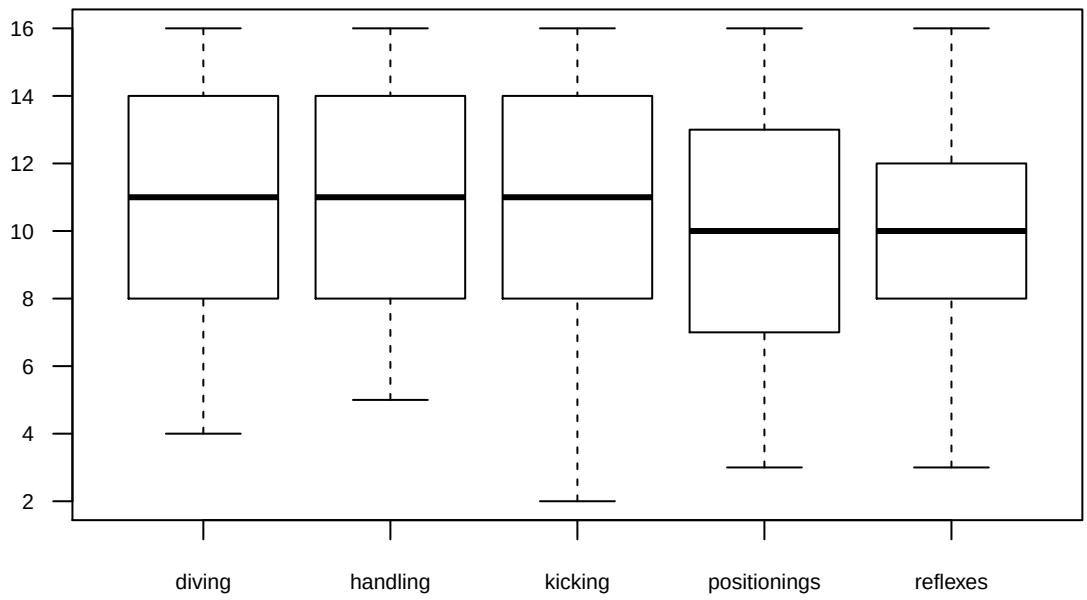

Atributos de mobilidade
90
80
70
60
50
40
30
acceleration
sprint_speed
agility
reactions
balance


Atributos de força
100
80
60
40
20
shot_power
jumping
stamina
strength
long_shots

Atributos de mentalidade




Atributos de defesa
80
60
40
20
marking
standing_tackle
sliding_tackle



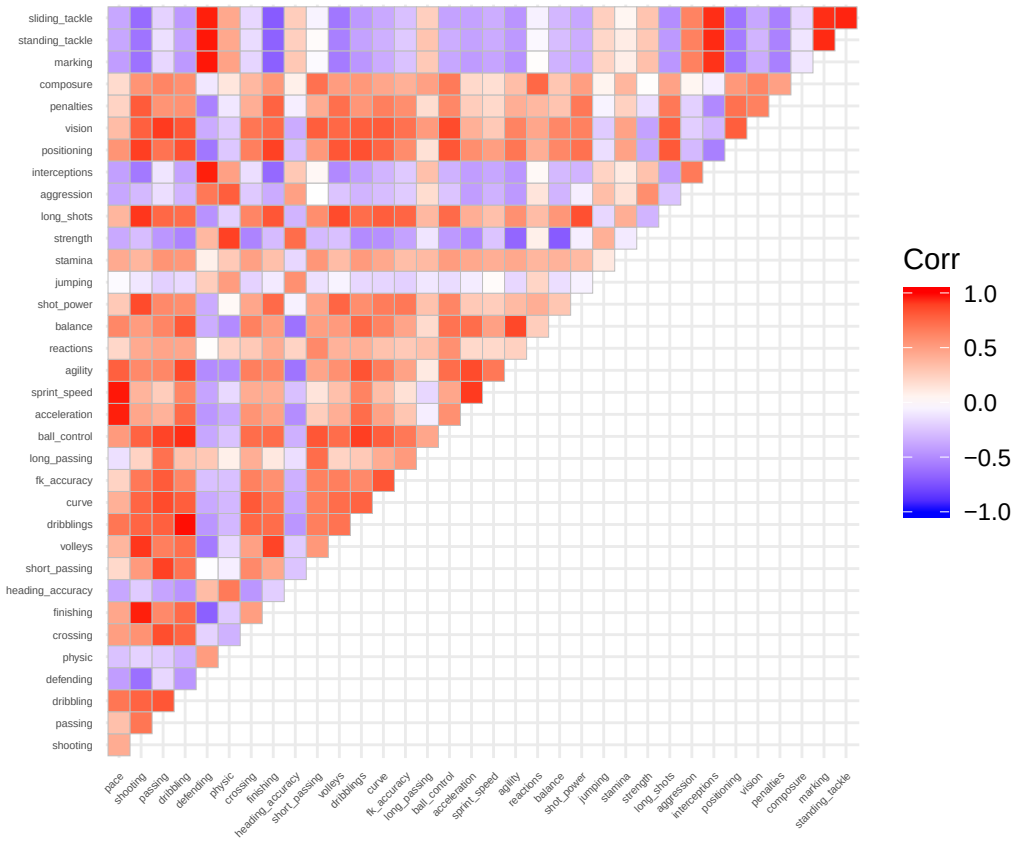
Observando os boxplots anteriores, podemos perceber que os dados possuem alguns outliers, mas nada muito significativo. Como não aparenta haver um peso muito grande nesses ponto, nada será removido.

Todos os dados faltantes foram removidos previamente na mesma etapa em que restringimos as variáveis do nosso conjunto de dados. A maior parte dos NA's ocorria em goleiros. Estes não possuiam atributos de jogadores de linha, prejudicavam os dados para a análise e, portanto, foram excluídos desse estudo.

Para esse conjunto de dados, todos os atributos utilizados são numéricos. O dataset inicial sem modificações possuia dados categoricos relativos à nacionalidade, ser destro ou canhoto, entre outros, mas optei por não utiliza-los.

Correlações

Observa-se alguns pontos muito correlacionados no gráfico. Faz sentido que isso aconteça, uma vez que diferentes atributos são específicos para o mesmo estilo de jogador, ou mesmo pela própria natureza de cada atributos. Por exemplo, velocidade ('sprint_speed') tem alta correlação com a aceleração ('acceleration') e ritmo ('pace'). Apesar disso, são características distintas e serão mantidas no conjunto de dadaos. A redução de dimensionalidade ajudará a descorrelacionar essas variáveis.



3) Preparação dos dados

Apesar de não utilizar dados categóricos na análise, pareceu interessante criar categorias especificas para cada jogador, a fim de simplificar o entendimento do que cada um representa em campo, dado o grande número de posições possíveis.

Para os zagueiros (“CB”,“LCB”,“RCB”), laterais (“RB”,“LB”) e alas(“LWB”, “RWB”) foi criada a categoria ‘defence’, referente aos jogadores de defesa. Para os jogadores de meio de campo (“CM”, “CDM”,“CAM”,“LM”,“RM”, “LAM”, “RAM”, “LCM”, “RCM”, “LDM”, “RDM”) foi criada a categoria ‘midfielder’. As posições restantes de ataque foram designadas como ‘forward’. Além disso, os reservas foram colocados em um classe especial ‘substitutes’ para fácil remoção durante a limpeza dos dados.

Iremos normalizar os valores da idade, altura e peso para estarem entre 0 e 100. Para isso utilizaremos a seguinte expressão:

$zi = (xi \breve{} min(x))/(max(x) \breve{} min(x)) * 100$

onde:

zi: o i-ésimo valor normalizado do conjunto de dados

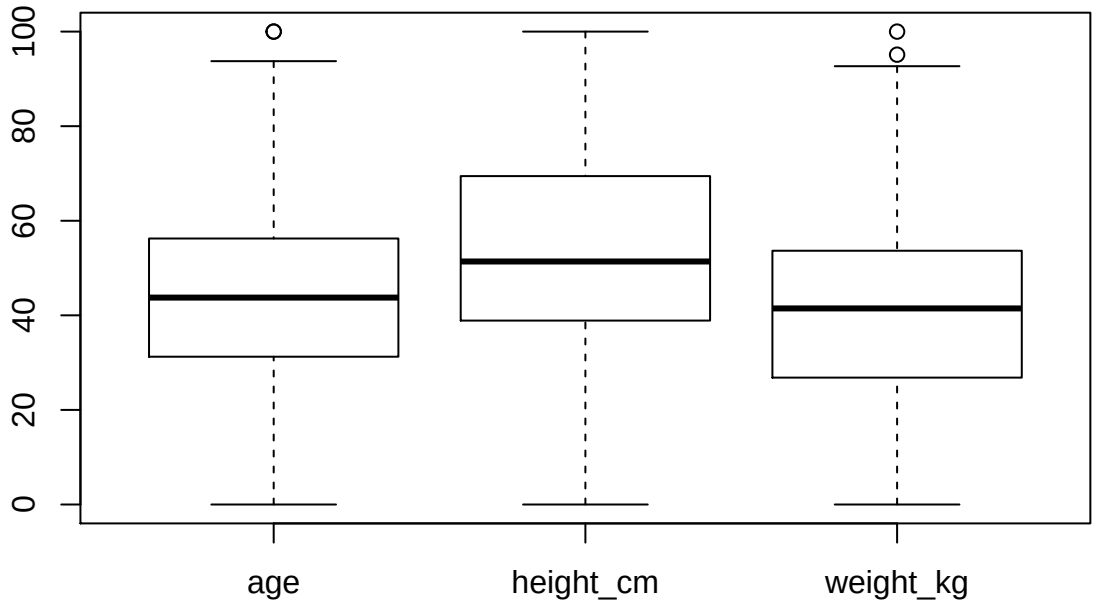
xi: o i-ésimo valor do conjunto de dados

$min(x)$: Valor mínimo do conjunto de dados

$max(x)$: Valor máximo do conjunto de dados

```
##       age           height_cm        weight_kg
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
```

```
##  1st Qu.: 31.25   1st Qu.: 38.89   1st Qu.: 26.83
##  Median : 43.75   Median : 51.39   Median : 41.46
##  Mean   : 44.57   Mean   : 52.79   Mean   : 41.95
##  3rd Qu.: 56.25   3rd Qu.: 69.44   3rd Qu.: 53.66
##  Max.   :100.00   Max.   :100.00   Max.   :100.00
```



Agora esses atributos estão na mesma escala do resto dos dados e poderão ser utilizados para análise.

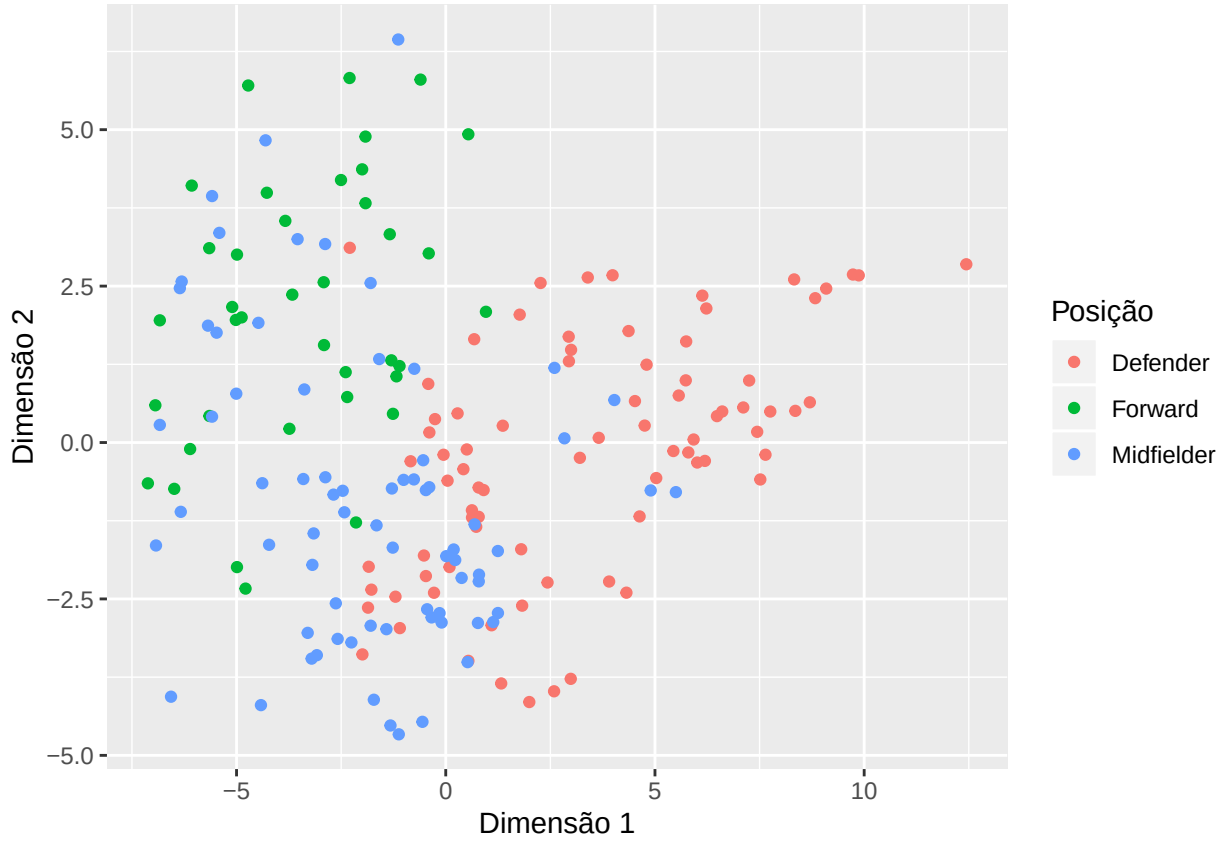
Redução de dimensionalidade

Dada a grande quantidade de variáveis, reduziremos a dimensão dos dados utilizando PCA.

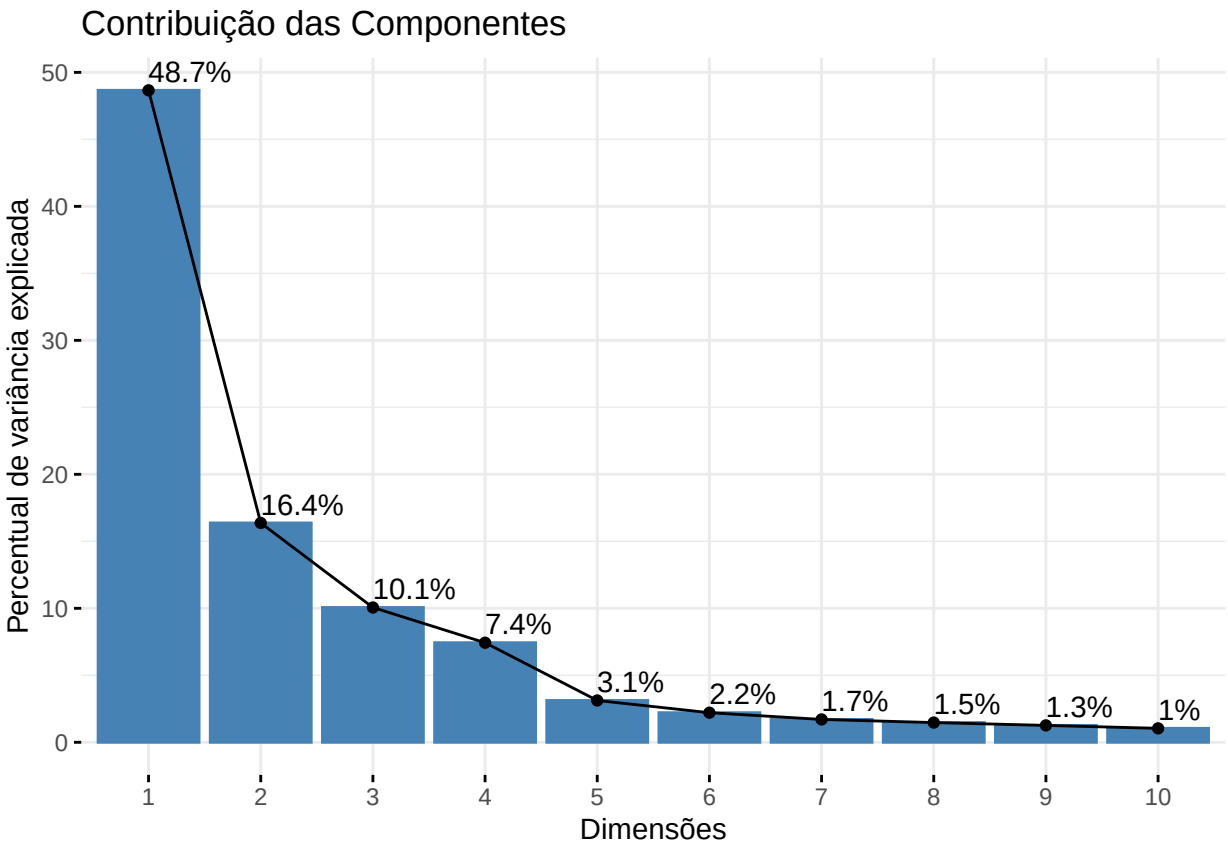
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     4.1268 2.3936 1.8765 1.61290 1.04524 0.8795 0.77264
## Proportion of Variance 0.4866 0.1637 0.1006 0.07433 0.03122 0.0221 0.01706
## Cumulative Proportion  0.4866 0.6503 0.7509 0.82522 0.85643 0.8785 0.89559
##                           PC8     PC9    PC10    PC11    PC12   PC13
## Standard deviation     0.71787 0.66442 0.60374 0.56342 0.50313 0.4806
## Proportion of Variance 0.01472 0.01261 0.01041 0.00907 0.00723 0.0066
## Cumulative Proportion  0.91031 0.92293 0.93334 0.94241 0.94964 0.9562
##                           PC14    PC15    PC16    PC17    PC18    PC19
## Standard deviation     0.44731 0.43429 0.41706 0.37026 0.36092 0.32833
## Proportion of Variance 0.00572 0.00539 0.00497 0.00392 0.00372 0.00308
## Cumulative Proportion  0.96196 0.96735 0.97232 0.97623 0.97996 0.98304
```

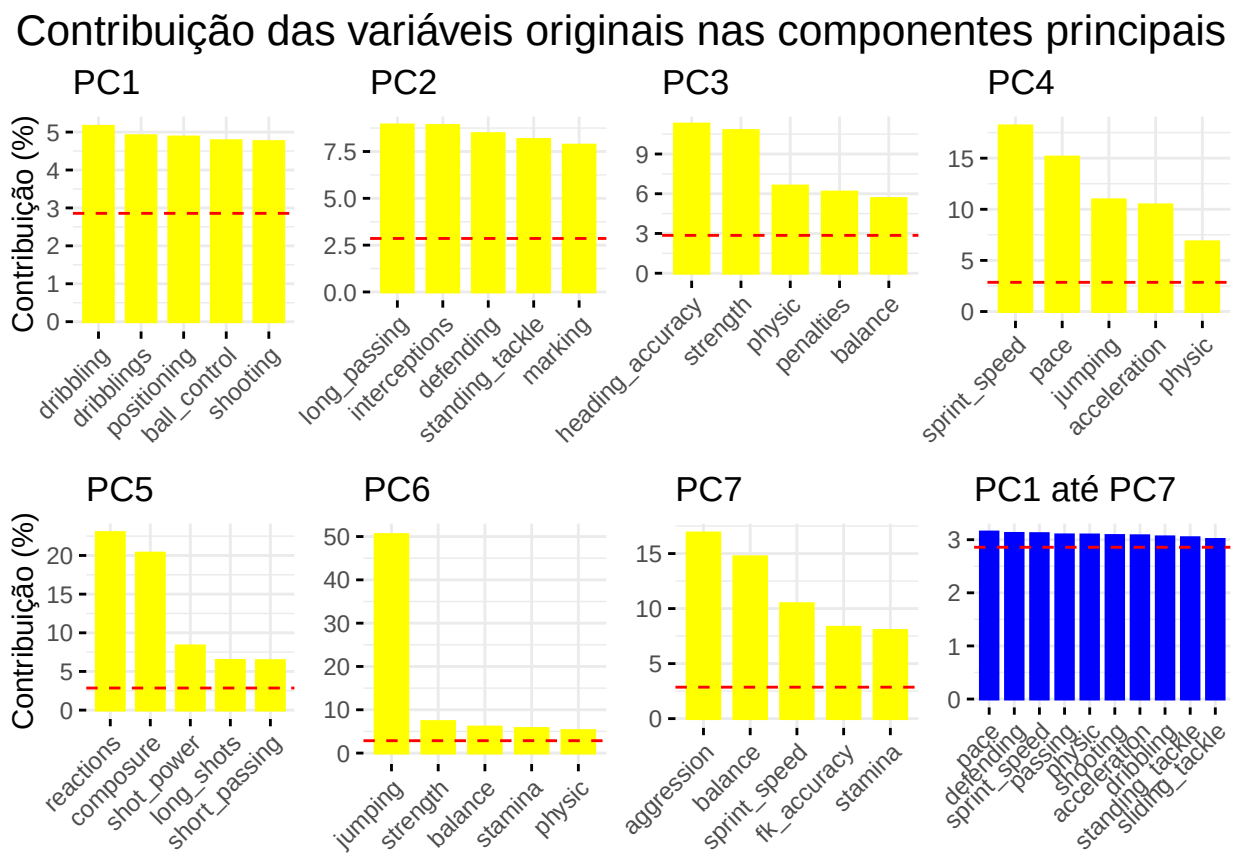
```
##                            PC20    PC21    PC22    PC23    PC24    PC25
## Standard deviation     0.31343 0.29099 0.27703 0.25964 0.25157 0.24221
## Proportion of Variance 0.00281 0.00242 0.00219 0.00193 0.00181 0.00168
## Cumulative Proportion  0.98584 0.98826 0.99045 0.99238 0.99419 0.99587
##                            PC26    PC27   PC28    PC29    PC30    PC31
## Standard deviation     0.22037 0.20352 0.1678 0.15524 0.02627 0.02471
## Proportion of Variance 0.00139 0.00118 0.0008 0.00069 0.00002 0.00002
## Cumulative Proportion  0.99725 0.99844 0.9992 0.99993 0.99995 0.99997
##                            PC32    PC33    PC34    PC35
## Standard deviation     0.02172 0.01623 0.01540 0.01424
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99998 0.99999 0.99999 1.00000
```

Observando as duas primeiras componente principais, podemos perceber que as posições de defesa se distinguem melhor das de ataque e meio campo. Enquanto elas aparecem de forma mais segregada à direita do gráfico, 'forward' e 'midfield' possuem muitos pontos ocupando espaços semelhantes.



Optaremos por trabalho com até 7 componentes principais, explicando aproximadamente 90% da variabilidade. A seguir, os plots das contribuições das variáveis originais para cada dimensão após a redução de dimensionalidade e para o acumulado das sete dimensões.


Contribuição das Componentes
Percentual de variância explicada
Dimensões
48.7%
16.4%
10.1%
7.4%
3.1%
2.2%
1.7%
1.5%
1.3%
1%
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10

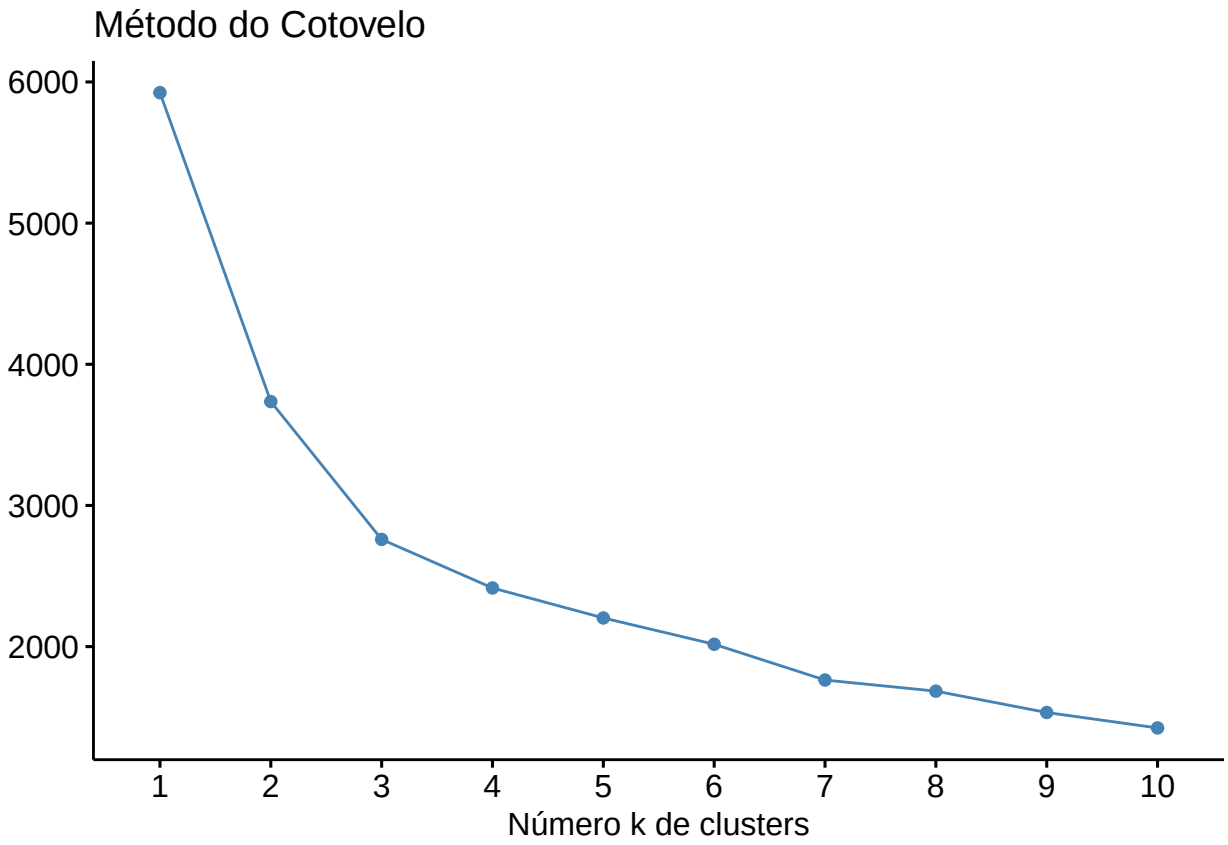


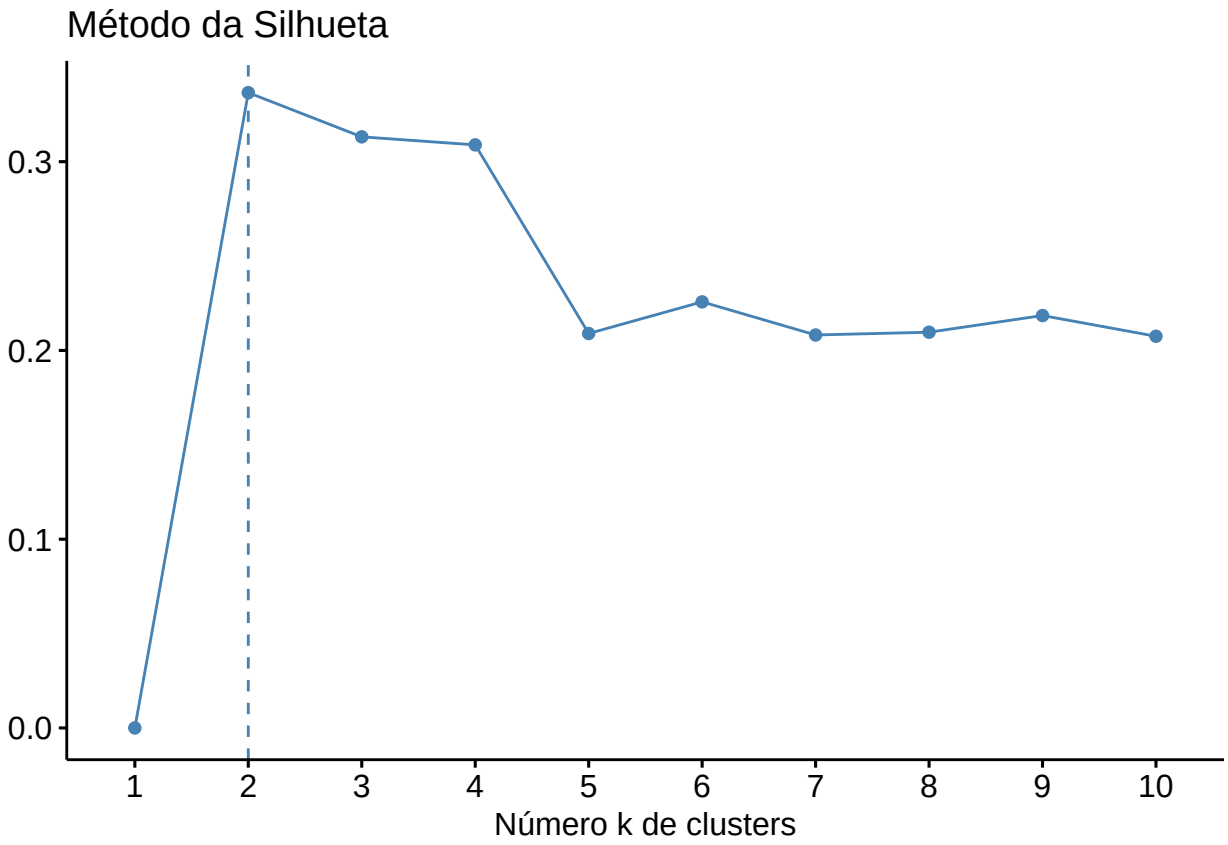
4) Modelagem: escolha dois algoritmos de clusterização

Como sugerido, modelaremos utilizando clusterização k-means e hierárquica.

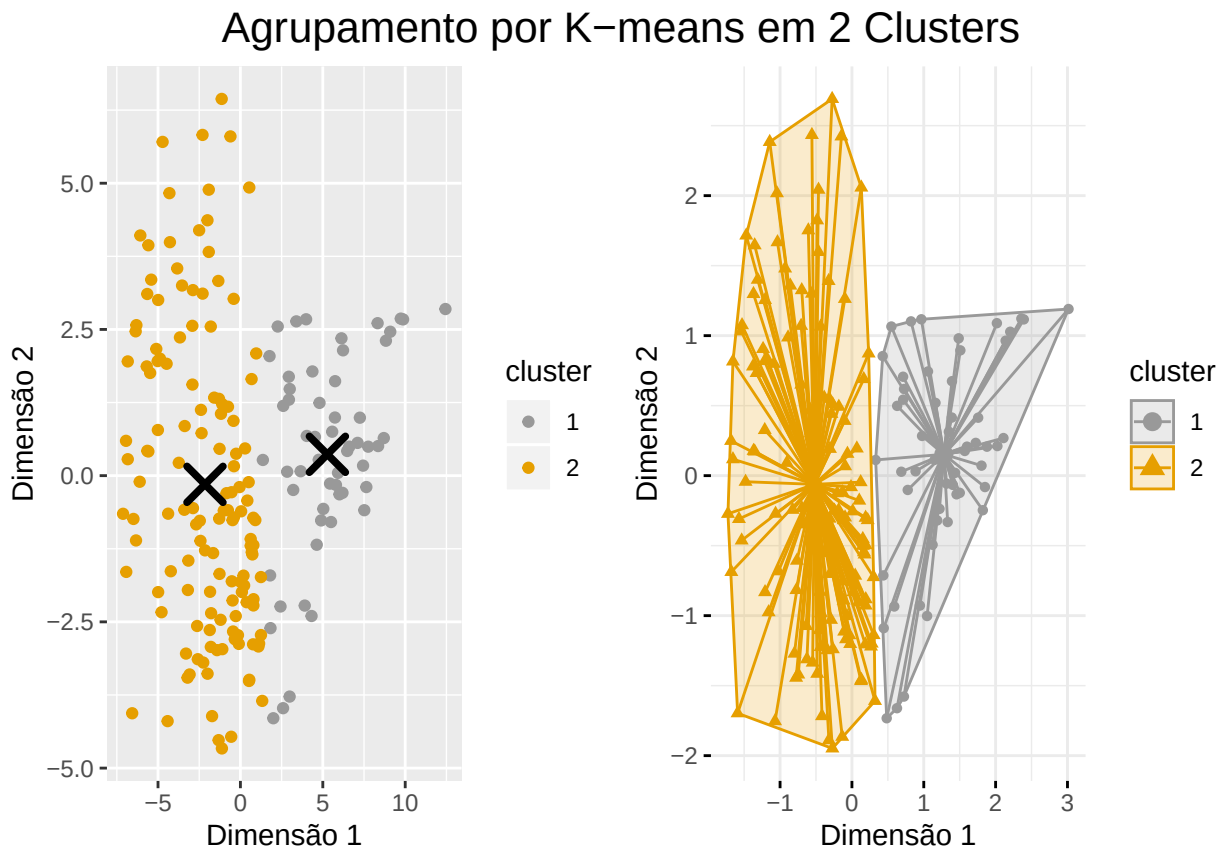
K-means

Os métodos do Cotovelo e da Silhueta apontam para uma separação em dois grupos, mas ainda muito adequada para três ou quatro. Apesar desse diagnóstico, um número pequeno de clusters parece deixar a análise muito redundante. Se imagina que ocorrerá uma divisão entre "jogador bom" e "jogador ruim" para o caso de apenas dois cluster, por exemplo. Para que tenhamos uma abrangência maior,testaremos para um número maior de clusters.

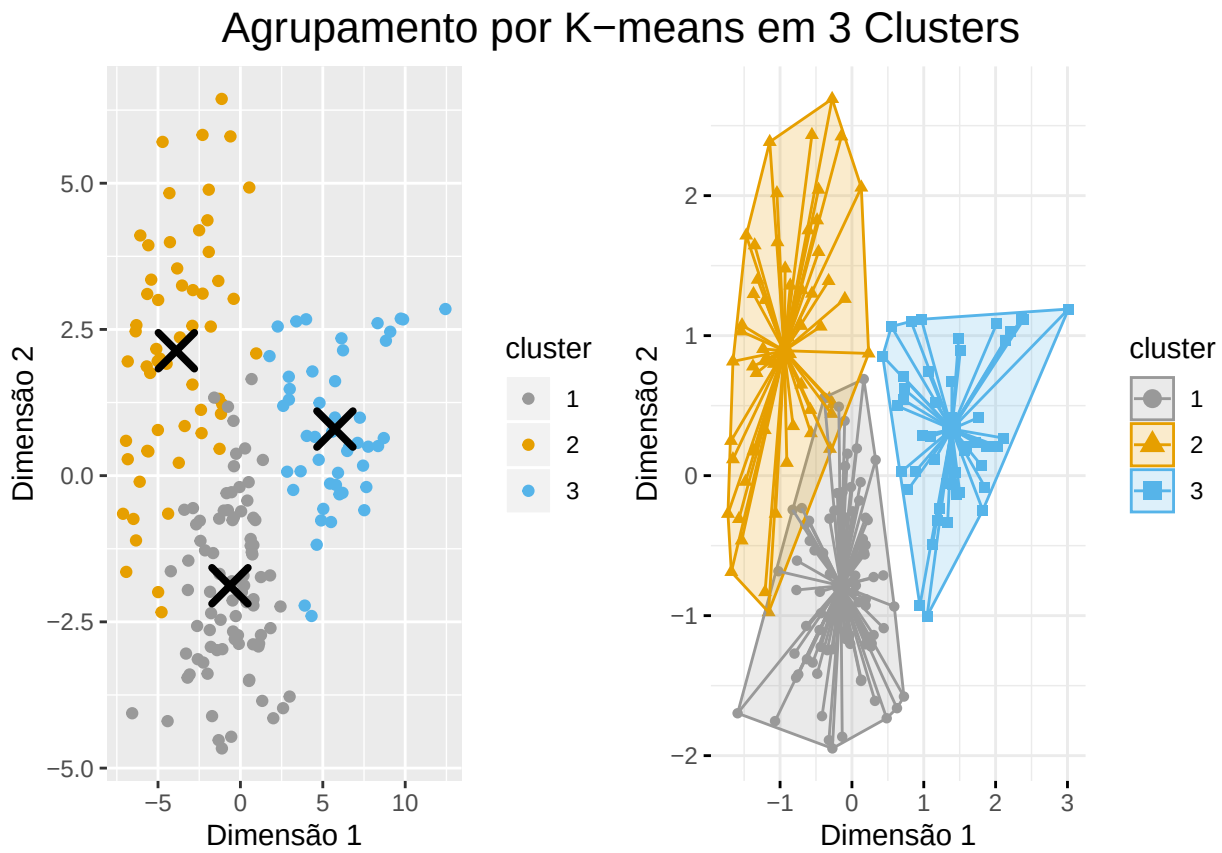

Método do Cotovelo
6000
5000
4000
3000
2000
1
2
3
4
5
6
7
8
9
10
Número k de clusters



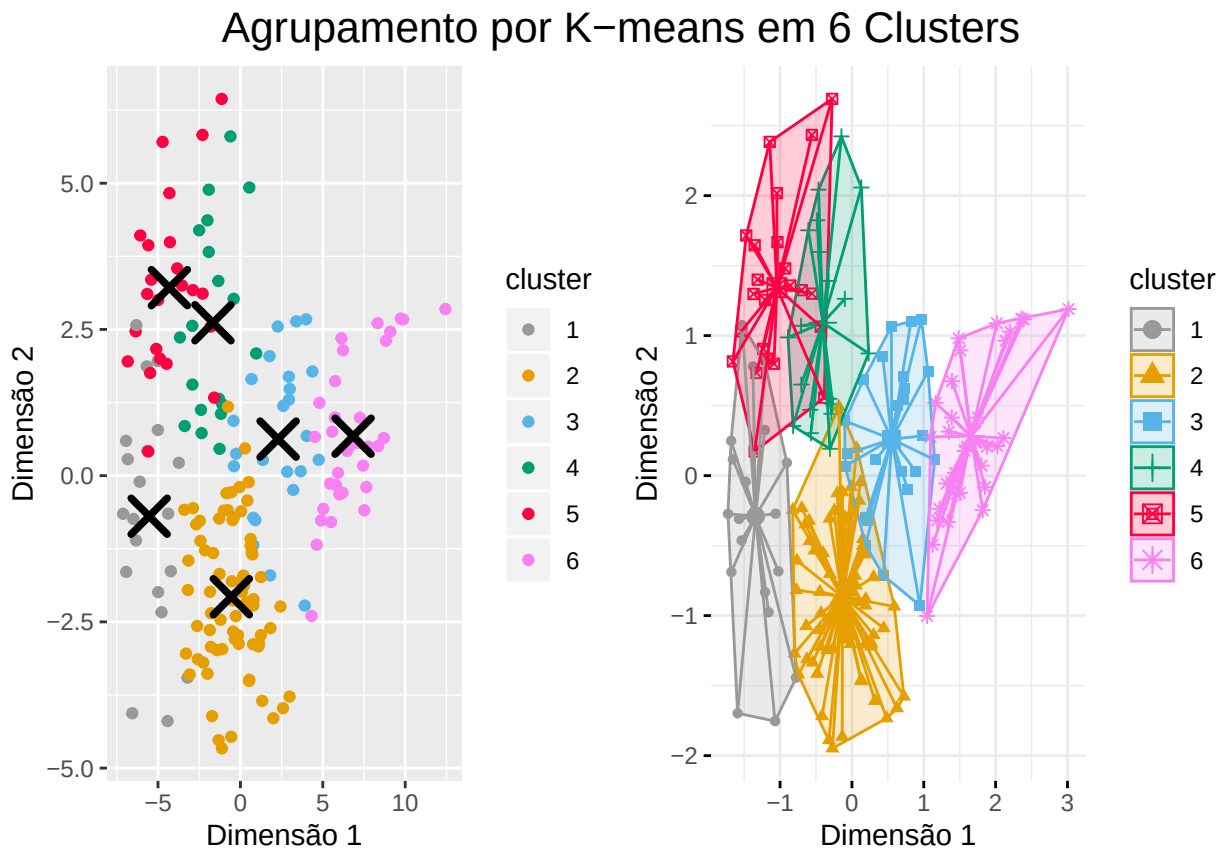
Para dois clusters, como esperado, obtemos uma separação perfeita entre dois grupos. Como já comentado anteriormente, esse tipo de divisão não parece adequado para o tipo de dados que possuimos.



Para a divisão em 3 clusters, ainda há pouca sobreposição. Apesar disso ser positivo teoricamente, talvez também não seja uma boa escolha . Os centróides parecem estar localizados onde se concentram cada uma das posições criadas anteriormente (‘forward’, ‘midfielder’, ‘defender’), logo não parece ser muito informativo.

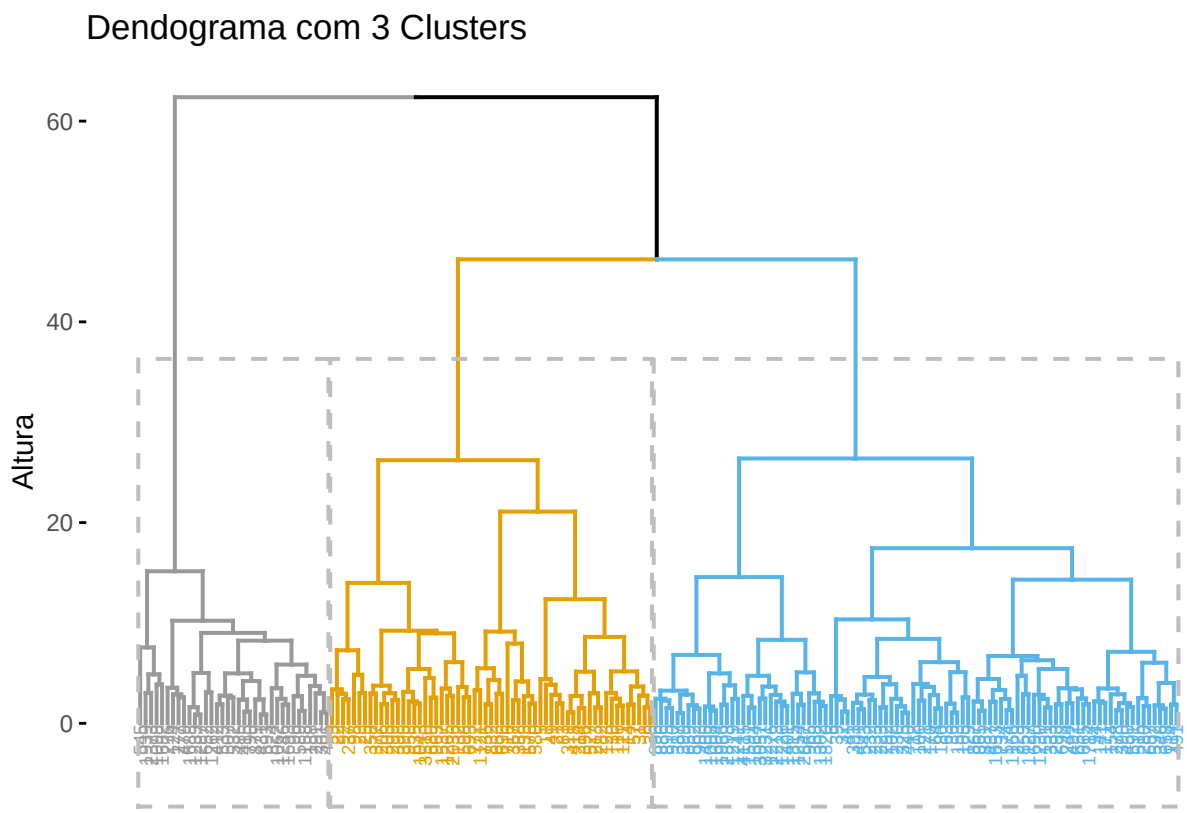


Conforme vamos aumentando o número de clusters vemos progressivamente uma divisão menos clara entre eles. Como vimos na nuvem de pontos das posições anteriormente, há uma sobreposição entre certos tipos de jogadores e talvez não seja tão clara a divisão entre eles. Avaliaremos o número de clusters mais a diante durante o processo de validação.

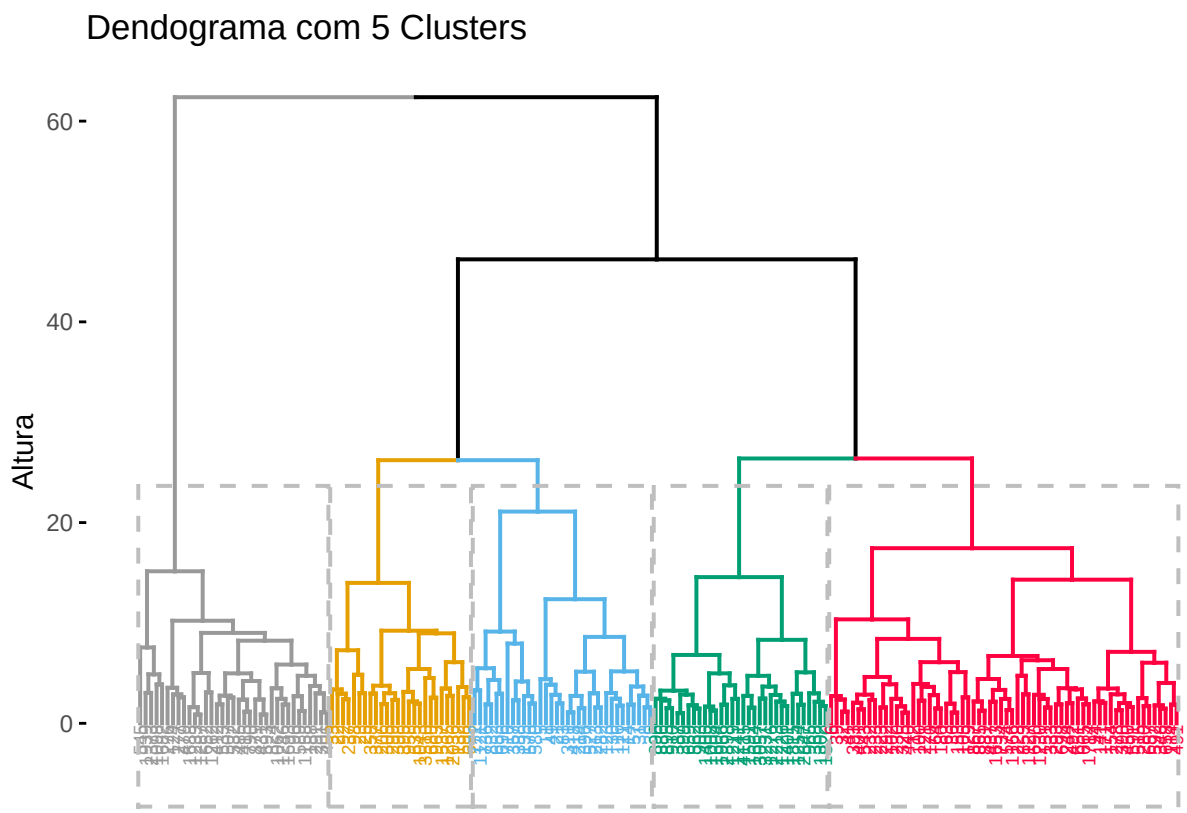


Clusterização hierárquica

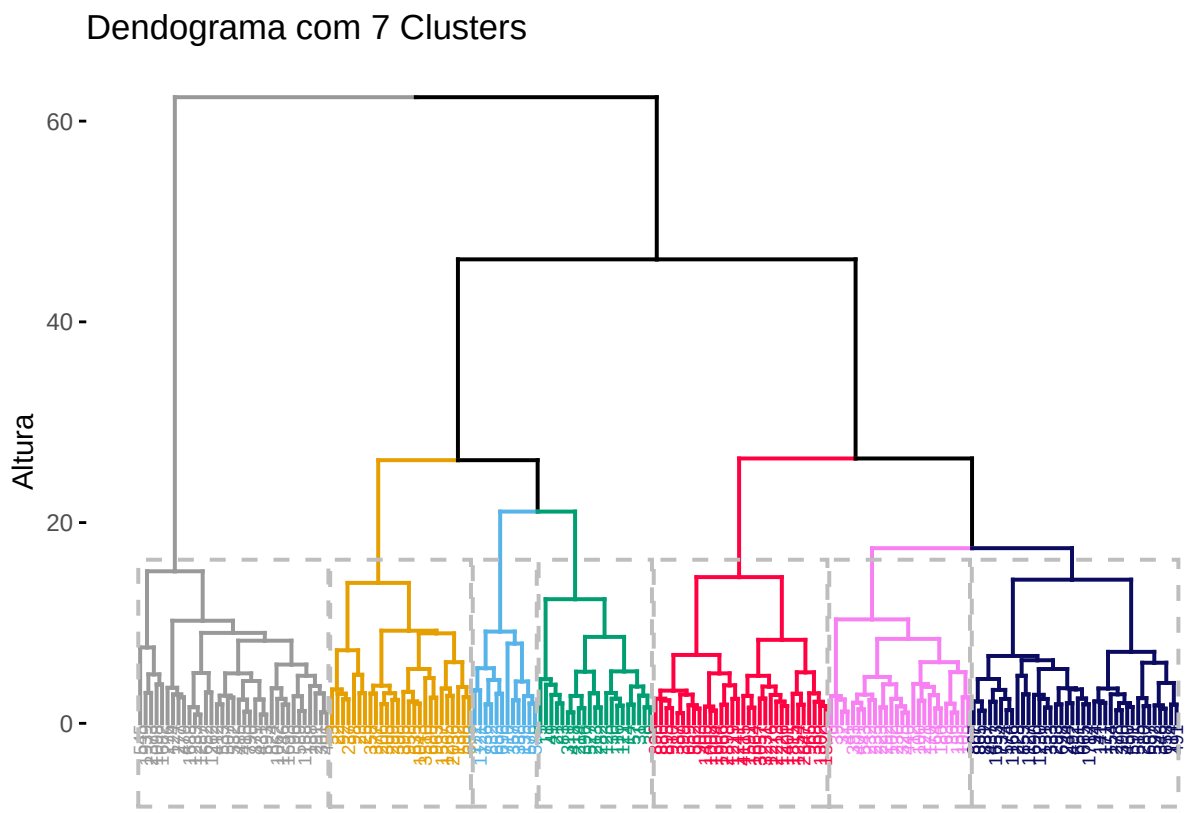
No caso da clusterização hierárquica, observamos que para 3 clusters há um desbalanceamento muito grande entre eles. Sendo um muito mais povoado que os outros dois. Além disso, pela diferença de altura, sabemos que podemos ganhar mais informação optando por mais clusters.



O agrupamento em 5 clusters parece mais adequado que o anterior. Talvez seja possível obter uma divisão ainda melhor, então testaremos para mais clusters.



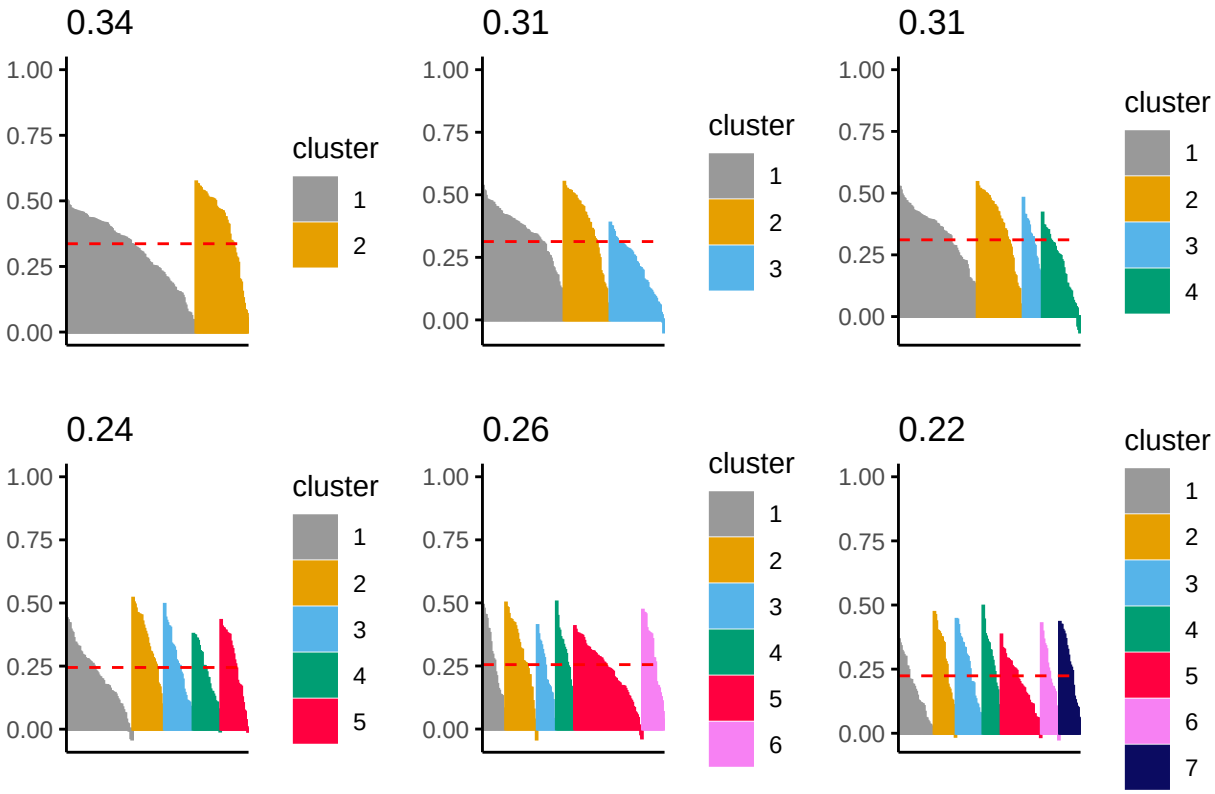
Para divisões maiores parece ocorrer um problema oposto ao da divisão de 3 clusters. Passam a surgir clusters com poucos casos, causando desbalanceamento.



5) Avaliação

Faremos a análise de silhueta para validar os modelos, variando o número de clusters entre 2 e 7, para os casos de clusterização por K-means e clusterização hierárquica.

A avaliação da silhueta para o K-means confirma as recomendações oferecidas pelas análise anteriores de que a divisão em dois, três ou quatro grupos são as mais apropriada, com silhuetas de 0.34, 0.31 e 0.31. Há uma queda nas silhuetas médias para cinco clusters para 0.24, mas um aumento logo depois para 0.26 com seis clusters. É importante observar que para todas as divisões acima de dois clusters, existem observações com silhuetas negativas.

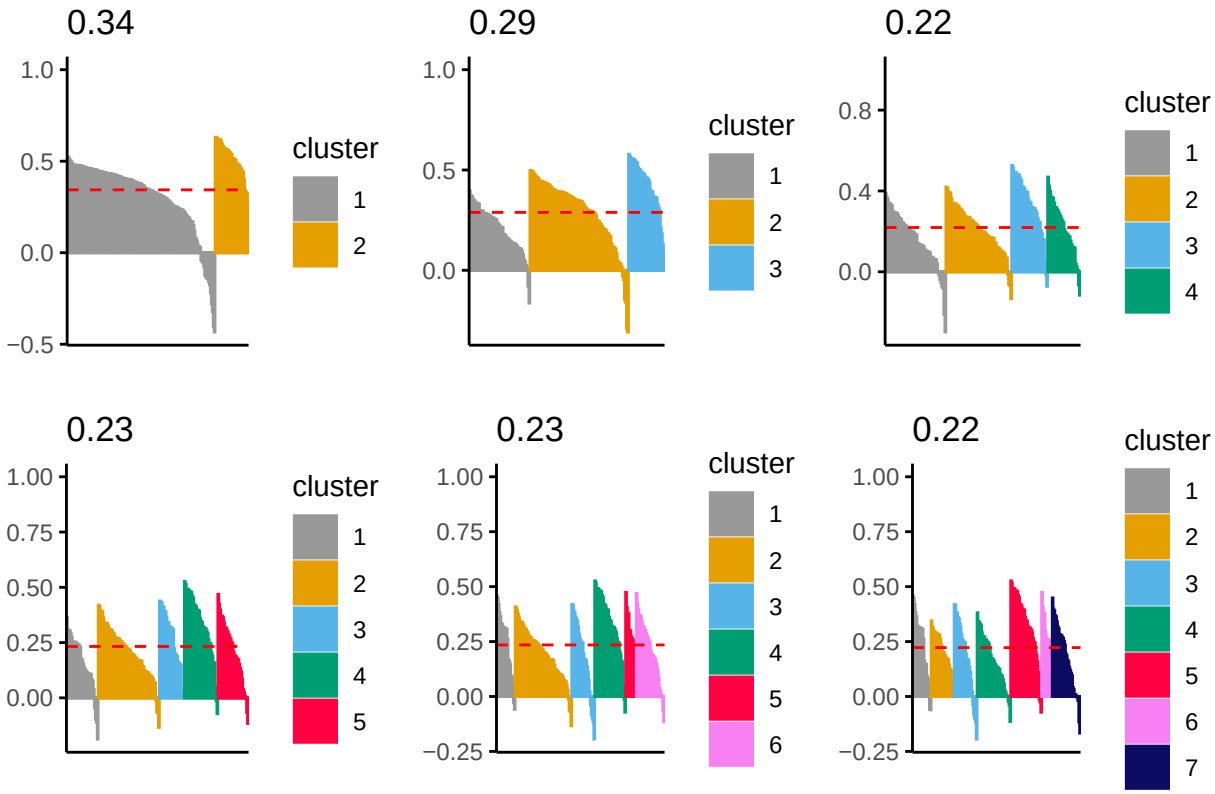



```
##   cluster size ave.sil.width
## 1       1  135          0.31
## 2       2   55          0.40
##   cluster size ave.sil.width
## 1       1  135          0.31
## 2       2   55          0.40
##   cluster size ave.sil.width
## 1       1   85          0.35
## 2       2   48          0.37
## 3       3   57          0.21
##   cluster size ave.sil.width
## 1       1   85          0.35
## 2       2   48          0.37
## 3       3   57          0.21
##   cluster size ave.sil.width
## 1       1   82          0.34
## 2       2   48          0.36
## 3       3   20          0.29
## 4       4   40          0.20
##   cluster size ave.sil.width
## 1       1   82          0.34
## 2       2   48          0.36
## 3       3   20          0.29
## 4       4   40          0.20
##   cluster size ave.sil.width
## 1       1   69          0.21
```

```
## 2       2   33          0.33
## 3       3   30          0.25
## 4       4   29          0.22
## 5       5   29          0.25
##   cluster size ave.sil.width
## 1       1   69          0.21
## 2       2   33          0.33
## 3       3   30          0.25
## 4       4   29          0.22
## 5       5   29          0.25
##   cluster size ave.sil.width
## 1       1   24          0.27
## 2       2   33          0.30
## 3       3   20          0.21
## 4       4   19          0.29
## 5       5   71          0.23
## 6       6   23          0.28
##   cluster size ave.sil.width
## 1       1   24          0.27
## 2       2   33          0.30
## 3       3   20          0.21
## 4       4   19          0.29
## 5       5   71          0.23
## 6       6   23          0.28
##   cluster size ave.sil.width
## 1       1   37          0.17
## 2       2   23          0.27
## 3       3   28          0.24
## 4       4   19          0.28
## 5       5   42          0.19
## 6       6   19          0.22
## 7       7   22          0.26
##   cluster size ave.sil.width
## 1       1   37          0.17
## 2       2   23          0.27
## 3       3   28          0.24
## 4       4   19          0.28
## 5       5   42          0.19
## 6       6   19          0.22
## 7       7   22          0.26
```

Assim como vimos anteriormente, as divisões para dois e três clusters são muito desbalanceados e podemos observar uma silhueta negativa significativa para estes casos. Os valores da silhueta partem de 0.34, para 0.29 e se estabilizam entre 0.22 e 0.23.

Se compararmos com a análise do K-means, a clusterização hierárquica apresenta mais observações negativas.




```
##   cluster size ave.sil.width
## 1       1  155          0.31
## 2       2   35          0.51
##   cluster size ave.sil.width
## 1       1  155          0.31
## 2       2   35          0.51
##   cluster size ave.sil.width
## 1       1   59          0.21
## 2       2   96          0.29
## 3       3   35          0.43
##   cluster size ave.sil.width
## 1       1   59          0.21
## 2       2   96          0.29
## 3       3   35          0.43
##   cluster size ave.sil.width
## 1       1   59          0.16
## 2       2   64          0.21
## 3       3   35          0.33
## 4       4   32          0.22
##   cluster size ave.sil.width
## 1       1   59          0.16
## 2       2   64          0.21
## 3       3   35          0.33
## 4       4   32          0.22
##   cluster size ave.sil.width
## 1       1   33          0.16
```

```
## 2       2   64          0.21
## 3       3   26          0.27
## 4       4   35          0.33
## 5       5   32          0.21
##   cluster size ave.sil.width
## 1       1   33          0.16
## 2       2   64          0.21
## 3       3   26          0.27
## 4       4   35          0.33
## 5       5   32          0.21
##   cluster size ave.sil.width
## 1       1   21          0.25
## 2       2   64          0.20
## 3       3   26          0.18
## 4       4   35          0.33
## 5       5   12          0.29
## 6       6   32          0.21
##   cluster size ave.sil.width
## 1       1   21          0.25
## 2       2   64          0.20
## 3       3   26          0.18
## 4       4   35          0.33
## 5       5   12          0.29
## 6       6   32          0.21
##   cluster size ave.sil.width
## 1       1   21          0.24
## 2       2   26          0.21
## 3       3   26          0.18
## 4       4   38          0.16
## 5       5   35          0.33
## 6       6   12          0.29
## 7       7   32          0.18
##   cluster size ave.sil.width
## 1       1   21          0.24
## 2       2   26          0.21
## 3       3   26          0.18
## 4       4   38          0.16
## 5       5   35          0.33
## 6       6   12          0.29
## 7       7   32          0.18
```

Para os índices a seguir, com a exceção do Davies-Bouldin, quanto maior o valor, melhor o agrupamento.

Os agrupamentos com dois, três ou quatro clusters apresentam índices com resultados melhores que os demais para o K-means, enquanto para o modelo hierárquico os agrupamentos com dois ou três têm melhores indicadores.

Como já comentado anteriormente, um número muito pequeno de clusters tornará a análise pouco abrangente, dada a variedade de posições de um jogador de futebol dentro do campo. Por isso, consideraremos apenas os casos com mais de cinco clusters.

A partir dessa escolha, observa-se que a modelagem por K-means obteve melhores índices de valida que a modelagem hierárquica. O particionamento em seis clusters com K-means apresentou resultados melhores que os demais e portanto será escolhido para os próximos passos como o melhor modelo.

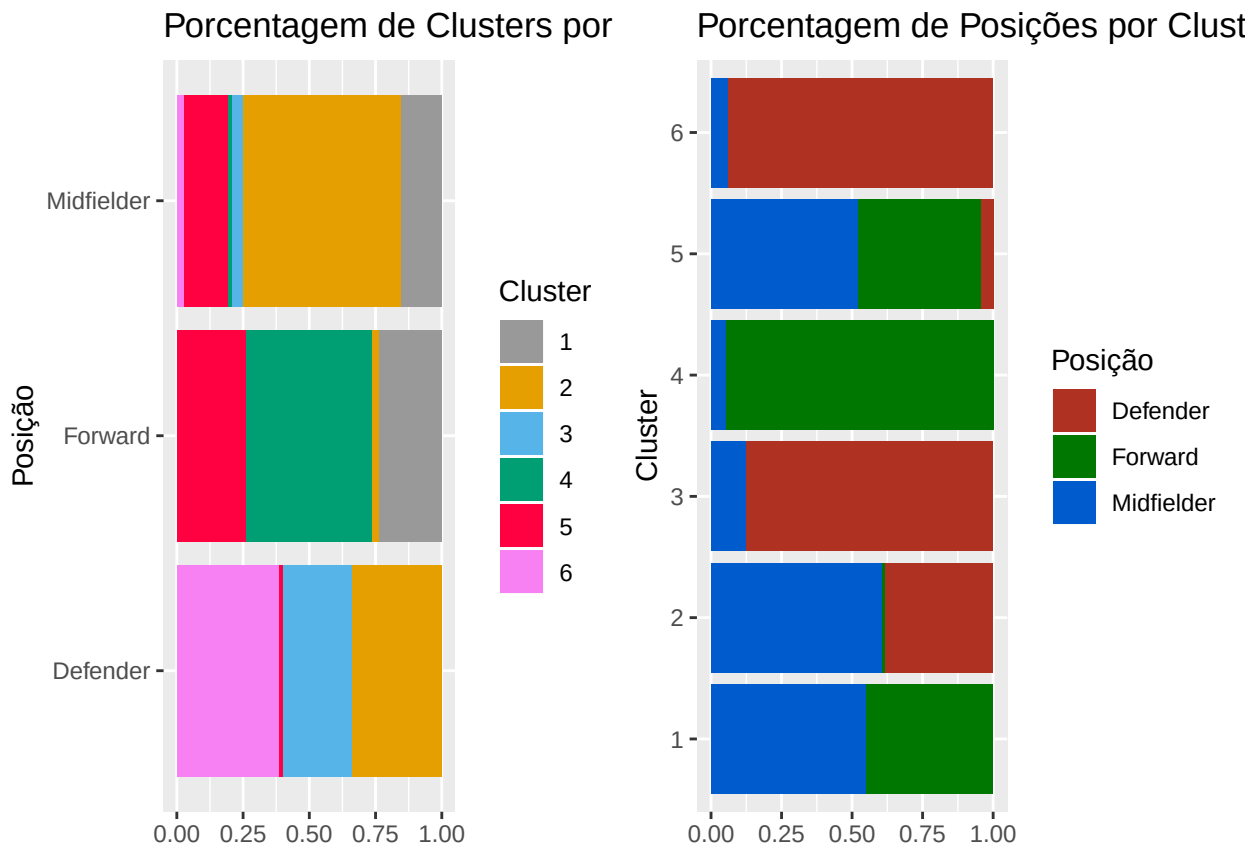
	met	k	sil	d_b	dunn	w_g
1	kmeans	2	0.354	1.090	0.126	0.481
3	kmeans	3	0.310	1.242	0.138	0.446
5	kmeans	4	0.298	1.232	0.115	0.446
7	kmeans	5	0.252	1.306	0.123	0.390
9	kmeans	6	0.262	1.191	0.145	0.401
11	kmeans	7	0.234	1.374	0.137	0.365
2	hclust	2	0.409	0.945	0.122	0.491
4	hclust	3	0.307	1.275	0.128	0.409
6	hclust	4	0.231	1.316	0.126	0.362
8	hclust	5	0.237	1.377	0.128	0.367
10	hclust	6	0.244	1.280	0.143	0.370
12	hclust	7	0.228	1.353	0.129	0.353

Relatório

Divisão das posições por cluster:

```
##
##     Defender Forward Midfielder
##   1        0       9         11
##   2       27       1         43
##   3       21       0          3
##   4        0      18          1
##   5        1      10         12
##   6       31       0          2
```

Percebe-se que algumas posições claramente dominam alguns clusters. Os clusters 3 e 6 são compostos predominantemente por jogadores de defesa, enquanto o cluster 4 é composto por jogadores de ataque e o cluster 2 tem maioria de jogadores de meio campo. Nos outros clusters há um balanceamento maior entre um par de classes.



Através dos resultados abaixo, podemos caracterizar genericamente as funções dos jogadores de cada cluster:

Os jogadores do cluster 1 são os meias e atacantes especialistas em armar o jogo. Se destacam por terem a maior estabilidade emocional (‘composure’), consciência sobre o posicionamento dos seus companheiros de equipe (‘vision’) e capacidade de reagir ao que acontece ao seu redor (‘reactions’). Entretanto, possuem atributos defensivos, relacionados à disputa de bola e agressividade baixos. Jogadores desse cluster são altamente valorizados por possuirem características raras.

Os jogadores do cluster 2 são volantes e meias centrais. Se destacam pela qualidade nos passes curtos (‘short_passing’) e longos (‘long_passing’), pelo condicionamento físico (‘stamina’) e nos seus altos atributos defesivos. Possuem baixos atributos ofensivos e baixa mobilidade. Esse também é cluster com a maior média de idade.

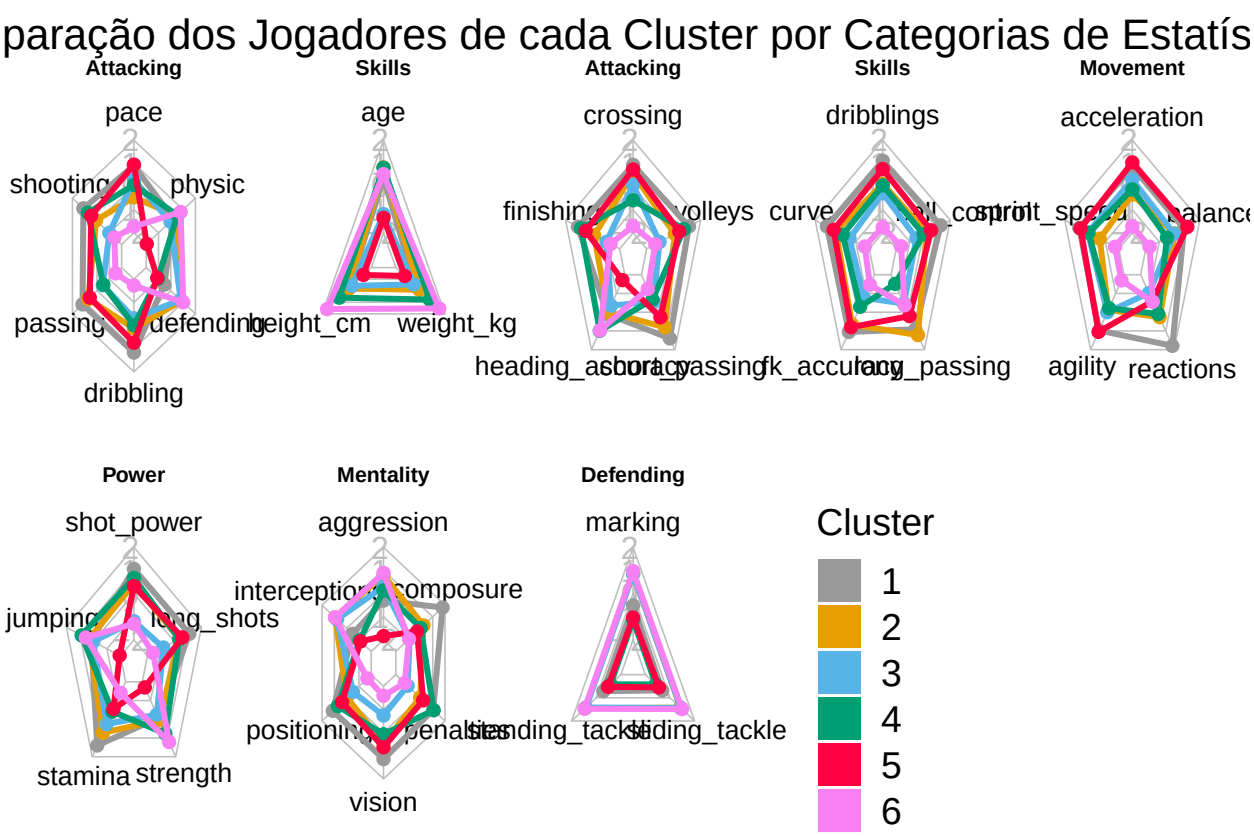
Os jogadores do cluster 3 são laterais e alas. Não possuem nenhum atributo que os defina de forma específica. São jogadores semelhantes aos do cluster 6, mas com maior habilidade, mobilidade, capacidade de cruzmento (‘crossing’) e condicionamento fisico (‘stamina’), porém menos fortes (‘strength’) e com menor capacidade de disputa pelo alto (‘heading_accuracy’). De forma grosseira, seriam um meio termo entre os cluster 5 e 6.

Os jogadores do cluster 4 são os centroavantes. Jogadores desse cluster são mais altos e pesados que os dos outros clusters, excluindo o cluster 2, e se destacam pelos atributos de ataque, possuindo altos níveis de cabeceio (‘heading’), capacidade de finalização (‘finishing’) e voleio (‘volley’). Além disso, têm a maior capacidade de disputa pelo alto (‘jumping’) e são os melhores cobradores de pênaltis (‘penalties’). Contudo, seus atributos de habilidade e movimento são baixos, assim como sua capacidade de toque curto (‘short_passing’) e cruzamento (‘crossing’). Também possuem a menor capacidade de interceptação (‘interceptions’).

Os jogadores do cluster 5 são predominantemente meias e atacantes de velocidade que têm por principal característica jogar pelos lados do campo. São os jogadores mais baixos, leves e jovens entre todos os clusters. Possuem todos os atributos de habilidade e ataque altos, com excessão da precisão no cabeceio (‘heading_accuracy’), assim como os atributos de mobilidade, com excessão da capacidade de reação (‘reactions’).

Os jogadores deste cluster são os fisicamente mais frágeis (‘strength’) e possuem a menor capacidade de disputa pelo alto (‘jumping’) entre todos os clusters. São também os com o menor comportamento agressivo (‘aggression’) e capacidade de interceptação (‘interceptions’).

Os jogadores do cluster 6 são zagueiros. Mais altos e mais pesados que os jogadores dos outros clusters, além de mais disposição para jogar na defesa e atributos físicos maiores de uma forma geral. Têm grande precisão no cabeceio (‘heading_accuracy’), qualidade nas disputas pelo alto (‘jumping’), jogam de forma agressiva (‘aggression’) e têm alta capacidade de interceptação (‘interceptions’). Por outro lado, seus outros atributos chamam a atenção por serem, em geral, muito menores do que os dos jogadores dos outros clusters.



A seguir estão os atletas mais próximos dos centroides para exemplificar o jogador padrão de cada cluster:

	Cluster	short_name
35	1	H. Son
325	2	D. Rose
368	3	A. Wan−Bissaka
311	4	C. Wilson
206	5	Pedro
301	6	M. Keane

Comparando os resultados do K-means com os resultados do Hierarchical clustering

Observando a distribuição de cada cluster percebemos de forma imediata que a ordem de cada um deles é diferentes, mas a distribuição dos jogadores entre os clusters equivalentes é semelhante. Mais especificamente: o cluster 1 do K-means equivale ao 1 do Hierarchical clustering, o cluster 2 K-means equivale ao 2, o cluster 3 equivale ao 6, o cluster 4 equivale ao 5, o cluster 5 equivale ao 3 e o cluster 6 equivale ao 4.

```
##
##     Defender Forward Midfielder
##   1        0       9         11
##   2       27       1         43
##   3       21       0          3
##   4        0      18          1
##   5        1      10         12
##   6       31       0          2
```

```
##
##     Defender Forward Midfielder
##   1        0      12          9
##   2       23       1         40
##   3        1      13         12
##   4       33       0          2
##   5        0      12          0
##   6       23       0          9
```

Alguns jogadores mudam de cluster quando mudamos a forma de clusterização. Por exemplo, se considerarmos o cluster 1 de cada modelo (esse cluster se refere ao mesmo tipo de jogador para ambos os casos) vemos que existem 4 jogadores que estão nele para o K-means, mas não para a H. clustering. Mais especificamente, os jogadores 'C. Eriksen', 'David Silva' e 'Bernardo Silva' são considerados meias armadores pelo K-means, mas meias de velocidade pelo H. clustering, enquanto o 'G. Sigurðsson' é um meia armador pelo K-means, mas um meia central pelo H. clustering.

Existem diversos casos desse tipo entre os clusters com jogadores variando entre eles dependendo do tipo de clusterização. No entanto, é curioso perceber que a maior parte desses jogadores são capazes de desempenhar multiplas funções, apesar de não ser a sua principal posição no time em que atua atualmente. Isso pode ser visto consultando a coluna 'player_positions'.

Conclusão

Fomos capazer de observar a existência de jogadores que, apesar de possuirem a mesma função dentro de campo, se agrupam melhor com jogadores de diferentes posições. Isso fica muito claro quando testamos para diferentes formas de clusterização. Alguns jogadores que não possuem histórico de atuação em outras posições se mostram bem agrupados longe dos clusters mais característicos.

Em uma segunda iteração seria interessante tentar incluir alguma das variáveis categóricas na análise e testar os jogadores das ligas de outros paises para ver como se agrupam. Utilizar outros modelos de clusterização também seria uma possibilidade.

Referências:

https://www.fifplay.com/fifa-20-player-attributes/

https://fifauteam.com/fifa-20-attributes-guide/

https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset?select=players_20.csv