EST0133 - Projeto I

Marcus Nunes

10 de Novembro de 2021

Instruções

- A data limite de entrega é 22/11/2021, às 23:59, via SIGAA
- O R é o único software permitido para coleta, limpeza e análise dos dados
- O trabalho deve ser feito em R Markdown, utilizando o arquivo `modelo.Rmd`
- Renomeie-o para `NomeSobrenome.Rmd`, em que `Nome` é o seu primeiro nome e `Sobrenome` é um de seus sobrenomes
- Envie conjuntamente seus arquivos `NomeSobrenome.Rmd` e `NomeSobrenome.pdf` para avaliação em um arquivo chamado `NomeSobrenome.zip`
- Respostas em arquivos que não estejam nos formatos Rmd e pdf não serão consideradas
- Identifique corretamente os eixos dos gráficos produzidos
- Respostas numeradas incorretamente não serão corrigidas
- Respostas com códigos supérfluos para a sua resolução, como pacotes desnecessários para as análises realizadas, terão pontuação descontada
- Não é permitido reportar resultados como capturas de tela (*screenshots*), a menos que sejam figuras

Parte I: Clusterização

O objetivo da primeira análise é identificar padrões em um conjunto de dados de jogadores de futebol presentes no game FIFA 2022. O arquivo `fifa.csv` possui informacões a respeito de 505 jogadores selecionados aleatoriamente. A descrição das colunas presentes no conjunto de dados está a seguir. A partir da coluna `crossing`, os valores possíveis para os atributos variam de 1 a 100[1]

- `id`: código de identificação do jogador (aleatório)
- `age`: idade (anos)
- `height`: altura (cm)
- `weight`: peso (kg)
- `club_number`: número da camisa no clube
- `national_number`: número da camisa na seleção
- `crossing`: cruzamento
- `finishing`: finalização
- `heading_accuracy`: precisão no cabeceio
- `short_passing`: qualidade do passe curto
- `volleys`: voleio
- `dribbling`: drible
- `curve`: efeito na bola ao chutar
- `fk_accuracy`: precisão na cobrança de faltas
- `long_passing`: qualidade do passe longo
- `ball_control`: controle da bola
- `acceleration`: aceleração
- `sprint_speed`: velocidade de corrida
- `agility`: agilidade
- `reactions`: reflexos
- `balance`: equilíbrio
- `shot_power`: força no chute
- `jumping`: pulo
- `stamina`: vigor físico
- `strength`: força física
- `long_shots`: qualidade dos chutes longos
- `aggression`: disposição para marcação
- `interceptions`: interceptação de passes
- `positioning`: posicionamento no campo
- `vision`: visão de jogo
- `penalties`: precisão na cobrança de pênaltis
- `composure`: habilidade de se manter calmo durante a partida
- `marking`: marcação
- `standing_tackle`: roubar a bola em pé
- `sliding_tackle`: roubar a bola com um carrinho
- `gk_diving`: habilidade do goleiro em pular para realizar uma defesa
- `gk_handling`: habilidade do goleiro em manusear e segurar a bola
- `gk_kicking`: habilidade do goleiro em chutar a bola
- `gk_positioning`: posicionamento do goleiro
- `gk_reflexes`: reflexos do goleiro

O objetivo desta análise é separar estes 505 jogadores em grupos similares, a partir das suas características físicas.

[1] Isto é, estas colunas podem, teoricamente, atingir quaisquer valores entre 1 e 100, mas não necessariamente algum jogador os tenha atingido nesta versão do jogo.

Questão 1

(*10 pontos*) Leia o conjunto de dados dentro do R em um objeto chamado `fifa` e verifique suas variáveis através dos comandos `head` e `summary`. Em um primeiro momento, sem realizar análise quantitativa alguma, é possível afirmar se existem uma ou mais variáveis deste conjunto de dados que deveriam ficar de fora da análise? Em particular, quais variáveis não irão acrescentar informação válida ao objetivo que desejamos completar? Justifique as suas escolhas. Em seguida, retire esta(s) variável(is) do conjunto de dados antes de prosseguir com a sua análise.

```
library(tidyverse)
theme_set(theme_bw())
library(ggfortify)
library(GGally)

fifa <- read_csv("dados/fifa.csv")

head(fifa)
```

```
## # A tibble: 6 x 40
##   id         age height weight club_number national_number crossing finishing
##   <chr>    <dbl>  <dbl>  <dbl>       <dbl>           <dbl>    <dbl>     <dbl>
## 1 njp4VXhW    32    185     81           9               9       71        95
## 2 PwrC9Vjm    36    187     83           7               7       87        95
## 3 F5MSKfZR    22    182     73           7              10       78        93
## 4 j29kjcUa    27    188     89          10               9       80        94
## 5 O769tLXh    33    185     81           9              19       75        90
## 6 D1CbYnrx    28    191     94           9               9       73        92
## # ... with 32 more variables: heading_accuracy <dbl>, short_passing <dbl>,
## #   volleys <dbl>, dribbling <dbl>, curve <dbl>, fk_accuracy <dbl>,
## #   long_passing <dbl>, ball_control <dbl>, acceleration <dbl>,
## #   sprint_speed <dbl>, agility <dbl>, reactions <dbl>, balance <dbl>,
## #   shot_power <dbl>, jumping <dbl>, stamina <dbl>, strength <dbl>,
## #   long_shots <dbl>, aggression <dbl>, interceptions <dbl>, positioning <dbl>,
## #   vision <dbl>, penalties <dbl>, composure <dbl>, marking <dbl>, ...
```

```
summary(fifa)
```

```
##       id                 age            height          weight
##  Length:505         Min.   :18.00   Min.   :167.0   Min.   : 61.00
##  Class :character   1st Qu.:26.00   1st Qu.:183.0   1st Qu.: 77.00
##  Mode  :character   Median :29.00   Median :187.0   Median : 82.00
##                     Mean   :28.96   Mean   :186.6   Mean   : 81.38
##                     3rd Qu.:32.00   3rd Qu.:191.0   3rd Qu.: 86.00
##                     Max.   :43.00   Max.   :201.0   Max.   :103.00
##
##   club_number    national_number    crossing       finishing
##  Min.   : 1.00   Min.   : 1.00   Min.   : 8.00   Min.   : 5.00
##  1st Qu.: 7.00   1st Qu.: 7.00   1st Qu.:14.00   1st Qu.:13.00
##  Median :10.00   Median :11.50   Median :37.00   Median :74.00
##  Mean   :15.41   Mean   :11.68   Mean   :38.87   Mean   :47.81
##  3rd Qu.:21.00   3rd Qu.:18.75   3rd Qu.:64.00   3rd Qu.:80.00
##  Max.   :99.00   Max.   :26.00   Max.   :87.00   Max.   :95.00
##  NA's   :4       NA's   :391
##  heading_accuracy short_passing      volleys       dribbling
##  Min.   : 7.00    Min.   :11.00   Min.   : 5.0   Min.   : 7.00
##  1st Qu.:14.00    1st Qu.:32.00   1st Qu.:13.0   1st Qu.:16.00
```

```
##  Median :63.00    Median :64.00   Median :64.0   Median :67.00
##  Mean   :46.31    Mean   :53.56   Mean   :45.2   Mean   :47.17
##  3rd Qu.:77.00    3rd Qu.:73.00   3rd Qu.:75.0   3rd Qu.:76.00
##  Max.   :93.00    Max.   :86.00   Max.   :90.0   Max.   :93.00
##
##      curve        fk_accuracy     long_passing    ball_control   acceleration
##  Min.   : 8.00   Min.   : 7.00   Min.   :12.00   Min.   : 9.0   Min.   :17.00
##  1st Qu.:15.00   1st Qu.:14.00   1st Qu.:31.00   1st Qu.:23.0   1st Qu.:45.00
##  Median :47.00   Median :33.00   Median :47.00   Median :70.0   Median :59.00
##  Mean   :42.41   Mean   :36.92   Mean   :45.94   Mean   :51.5   Mean   :60.11
##  3rd Qu.:68.00   3rd Qu.:59.00   3rd Qu.:61.00   3rd Qu.:77.0   3rd Qu.:76.00
##  Max.   :86.00   Max.   :86.00   Max.   :86.00   Max.   :91.0   Max.   :97.00
##
##   sprint_speed     agility        reactions        balance        shot_power
##  Min.   :18.0   Min.   :19.00   Min.   :53.00   Min.   :20.00   Min.   :32.0
##  1st Qu.:46.0   1st Qu.:47.00   1st Qu.:72.00   1st Qu.:44.00   1st Qu.:54.0
##  Median :60.0   Median :62.00   Median :75.00   Median :56.00   Median :73.0
##  Mean   :61.1   Mean   :60.14   Mean   :75.59   Mean   :56.65   Mean   :67.8
##  3rd Qu.:77.0   3rd Qu.:73.00   3rd Qu.:79.00   3rd Qu.:70.00   3rd Qu.:80.0
##  Max.   :97.0   Max.   :93.00   Max.   :94.00   Max.   :92.00   Max.   :94.0
##
##     jumping         stamina         strength       long_shots     aggression
##  Min.   :31.00   Min.   :15.00   Min.   :33.00   Min.   : 4.0   Min.   :11.00
##  1st Qu.:65.00   1st Qu.:35.00   1st Qu.:67.00   1st Qu.:14.0   1st Qu.:29.00
##  Median :72.00   Median :52.00   Median :73.00   Median :63.0   Median :45.00
##  Mean   :70.48   Mean   :54.32   Mean   :72.67   Mean   :44.4   Mean   :49.09
##  3rd Qu.:78.00   3rd Qu.:75.00   3rd Qu.:80.00   3rd Qu.:73.0   3rd Qu.:69.00
##  Max.   :95.00   Max.   :91.00   Max.   :95.00   Max.   :93.0   Max.   :90.00
##
##  interceptions    positioning        vision       penalties       composure
##  Min.   : 6.00   Min.   : 4.00   Min.   :11.0   Min.   :10.00   Min.   :22.00
##  1st Qu.:19.00   1st Qu.:12.00   1st Qu.:49.0   1st Qu.:22.00   1st Qu.:58.00
##  Median :24.00   Median :73.00   Median :63.0   Median :60.00   Median :67.00
##  Mean   :27.05   Mean   :47.74   Mean   :59.6   Mean   :48.95   Mean   :65.59
##  3rd Qu.:34.00   3rd Qu.:80.00   3rd Qu.:70.0   3rd Qu.:74.00   3rd Qu.:76.00
##  Max.   :66.00   Max.   :95.00   Max.   :87.0   Max.   :92.00   Max.   :95.00
##
##     marking      standing_tackle sliding_tackle    gk_diving
##  Min.   : 5.00   Min.   : 7.00   Min.   : 8.00   Min.   : 1.00
##  1st Qu.:17.00   1st Qu.:14.00   1st Qu.:14.00   1st Qu.: 9.00
##  Median :25.00   Median :20.00   Median :19.00   Median :16.00
##  Mean   :26.32   Mean   :24.76   Mean   :21.66   Mean   :42.86
##  3rd Qu.:35.00   3rd Qu.:33.00   3rd Qu.:27.00   3rd Qu.:77.00
##  Max.   :64.00   Max.   :72.00   Max.   :66.00   Max.   :91.00
##
##   gk_handling      gk_kicking    gk_positioning   gk_reflexes
##  Min.   : 1.00   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:11.00   1st Qu.:10.00   1st Qu.:10.00   1st Qu.:10.00
##  Median :16.00   Median :16.00   Median :16.00   Median :16.00
##  Mean   :41.91   Mean   :40.28   Mean   :42.57   Mean   :43.66
##  3rd Qu.:74.00   3rd Qu.:72.00   3rd Qu.:76.00   3rd Qu.:79.00
##  Max.   :92.00   Max.   :93.00   Max.   :92.00   Max.   :90.00
##
```

```r
fifa <-
  fifa %>%
  select(-id, -club_number, -national_number)
```

As variáveis `id`, `club_number` e `national_number` foram retiradas porque não adicionam informação na análise. O código de identificação do jogador, por ser determinado aleatoriamente, não influencia na sua posição dentro do campo. Além disso, os números das camisas, seja no clube ou na seleção, também não influenciam na posição do jogador, embora em alguns casos sejam derivadas a partir da posição do jogador em campo.

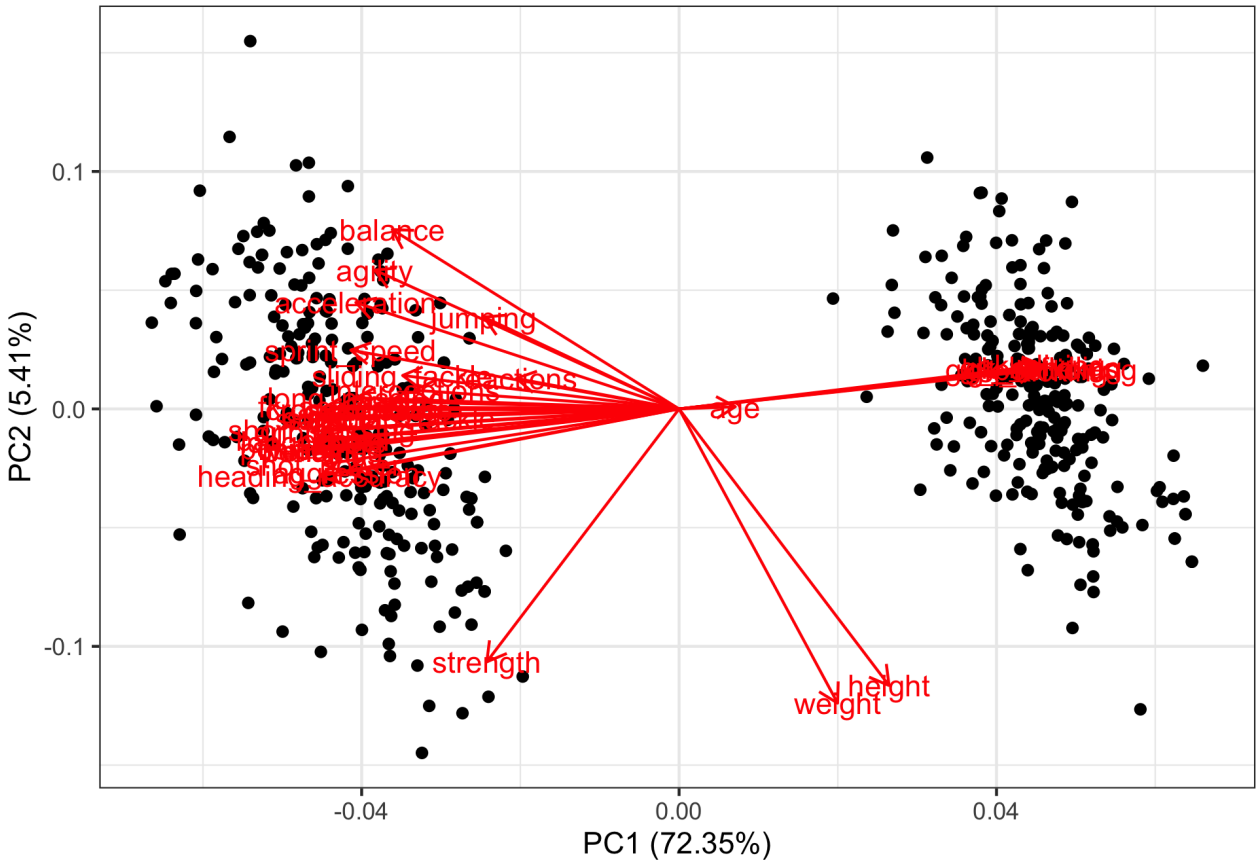
Questão 2

(*10 pontos*) Faça a PCA para este conjunto de dados, interpretando o resultado obtido. Essa interpretação deve ser numérica e visual.

```r
fifa_pca <- prcomp(fifa, center = TRUE, scale. = TRUE)

summary(fifa_pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     5.1740 1.41440 1.17579 1.06455 0.98600 0.8981 0.71037
## Proportion of Variance 0.7235 0.05407 0.03736 0.03063 0.02628 0.0218 0.01364
## Cumulative Proportion  0.7235 0.77759 0.81496 0.84558 0.87186 0.8937 0.90730
##                            PC8    PC9    PC10    PC11    PC12   PC13    PC14
## Standard deviation     0.66472 0.6173 0.55026 0.54893 0.52553 0.4978 0.45572
## Proportion of Variance 0.01194 0.0103 0.00818 0.00814 0.00746 0.0067 0.00561
## Cumulative Proportion  0.91924 0.9295 0.93772 0.94586 0.95333 0.9600 0.96564
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.41858 0.40990 0.39239 0.37243 0.31781 0.29721 0.28252
## Proportion of Variance 0.00474 0.00454 0.00416 0.00375 0.00273 0.00239 0.00216
## Cumulative Proportion  0.97037 0.97491 0.97908 0.98282 0.98555 0.98794 0.99010
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.25623 0.23749 0.20932 0.19799 0.18675 0.15108 0.14610
## Proportion of Variance 0.00177 0.00152 0.00118 0.00106 0.00094 0.00062 0.00058
## Cumulative Proportion  0.99187 0.99340 0.99458 0.99564 0.99658 0.99720 0.99778
##                           PC29    PC30   PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.14388 0.10846 0.1047 0.09250 0.08836 0.08535 0.07859
## Proportion of Variance 0.00056 0.00032 0.0003 0.00023 0.00021 0.00020 0.00017
## Cumulative Proportion  0.99834 0.99866 0.9990 0.99918 0.99939 0.99959 0.99976
##                           PC36    PC37
## Standard deviation     0.07042 0.06339
## Proportion of Variance 0.00013 0.00011
## Cumulative Proportion  0.99989 1.00000
```

```r
autoplot(fifa_pca,
         data = fifa,
         loadings = TRUE,
         loadings.label = TRUE) +
  theme_bw()
```




```
fifa_pca$rotation %>%
  as_tibble() %>%
  bind_cols(habilidades = names(fifa)) %>%
  relocate(habilidades) %>%
  filter(PC1 > 0)
```

```
## # A tibble: 8 x 38
##   habilidades      PC1      PC2     PC3      PC4     PC5     PC6     PC7     PC8
##   <chr>          <dbl>    <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 age           0.0291  0.00384 -0.608   0.129   -0.471   0.473  -0.0364  0.232
## 2 height        0.109  -0.478   -0.0199 -0.0285   0.183  -0.0966 -0.145   0.298
## 3 weight        0.0820 -0.509   -0.130   0.0953   0.214   0.0806 -0.438   0.130
## 4 gk_diving     0.188   0.0756  -0.118   0.00765  0.112  -0.0518 -0.0371 -0.0325
## 5 gk_handling   0.187   0.0723  -0.125   0.00489  0.105  -0.0450 -0.0522 -0.0331
## 6 gk_kicking    0.186   0.0671  -0.127  -0.00482  0.112  -0.0543 -0.0736 -0.0789
## 7 gk_positioni~ 0.188   0.0692  -0.133   0.00308  0.0978 -0.0403 -0.0451 -0.0306
## 8 gk_reflexes   0.188   0.0715  -0.117   0.00326  0.106  -0.0495 -0.0435 -0.0226
## # ... with 29 more variables: PC9 <dbl>, PC10 <dbl>, PC11 <dbl>, PC12 <dbl>,
## #   PC13 <dbl>, PC14 <dbl>, PC15 <dbl>, PC16 <dbl>, PC17 <dbl>, PC18 <dbl>,
## #   PC19 <dbl>, PC20 <dbl>, PC21 <dbl>, PC22 <dbl>, PC23 <dbl>, PC24 <dbl>,
## #   PC25 <dbl>, PC26 <dbl>, PC27 <dbl>, PC28 <dbl>, PC29 <dbl>, PC30 <dbl>,
## #   PC31 <dbl>, PC32 <dbl>, PC33 <dbl>, PC34 <dbl>, PC35 <dbl>, PC36 <dbl>,
## #   PC37 <dbl>
```

Após a transformação nos dados, encontramos uma PCA que concentra 77,76% da variabilidade dos dados nas duas primeiras componentes principais. São necessárias pelo menos as três primeiras componentes principais para que mais de 80% da variância seja explicada.

Ao fazer o gráfico da PCA, é possível encontrar dois clusters de dados, localizados à esquerda e à direita no gráfico. É possível ver que as habilidades cuja PC1 são maiores do que zero são idade, altura e peso, além de todas aquelas relacionadas diretamente com a posição de goleiro.

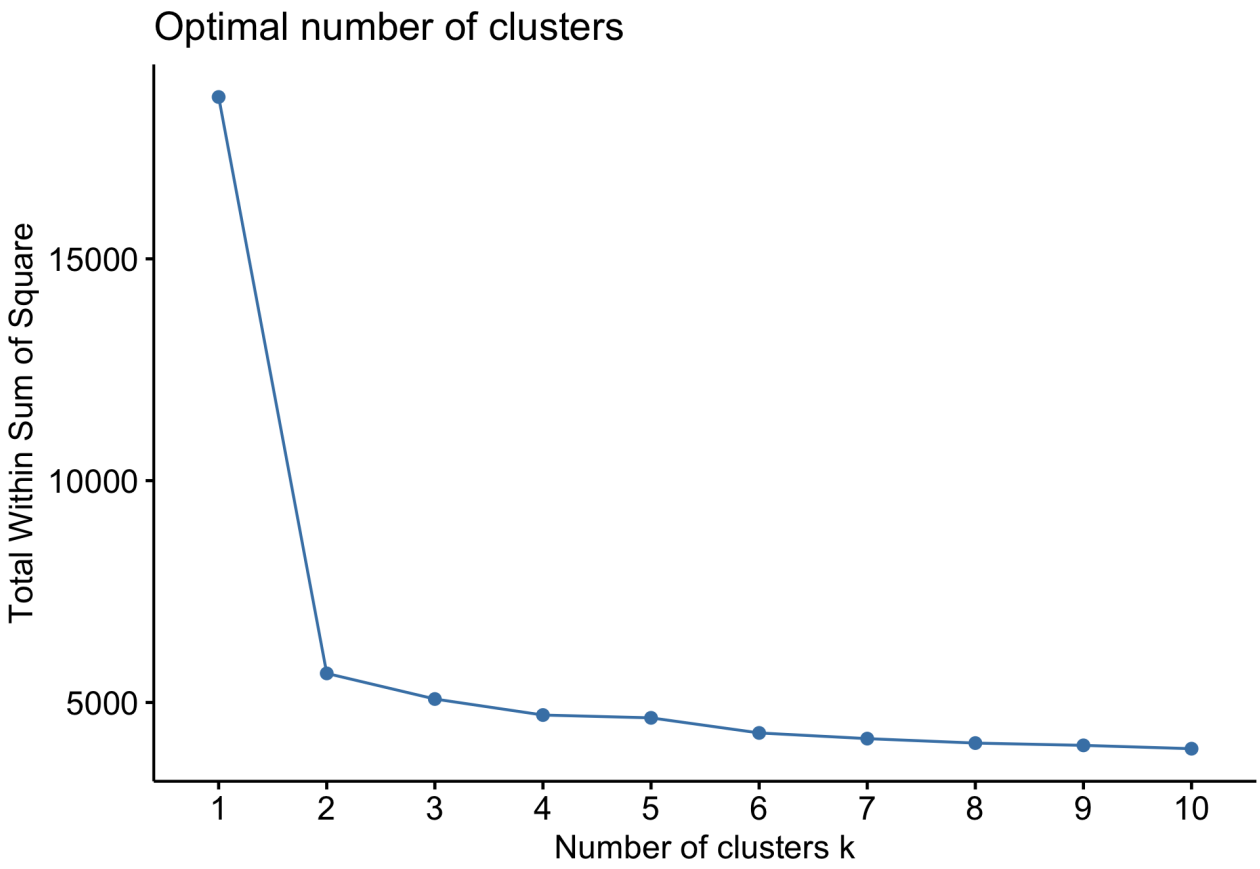
Questão 3

(*10 pontos*) Os jogadores presentes no conjunto de dados foram selecionados a partir de um determinado número de posições no campo. Estas posições foram simplificadas a ponto de serem apenas Goleiro (G), Defesa (D), Meio-Campo (M) e Ataque (A). Não necessariamente todas elas estão representadas no conjunto de dados fornecido. Qual é o número de clusters que os métodos do cotovelo e da silhueta sugerem para estes dados?

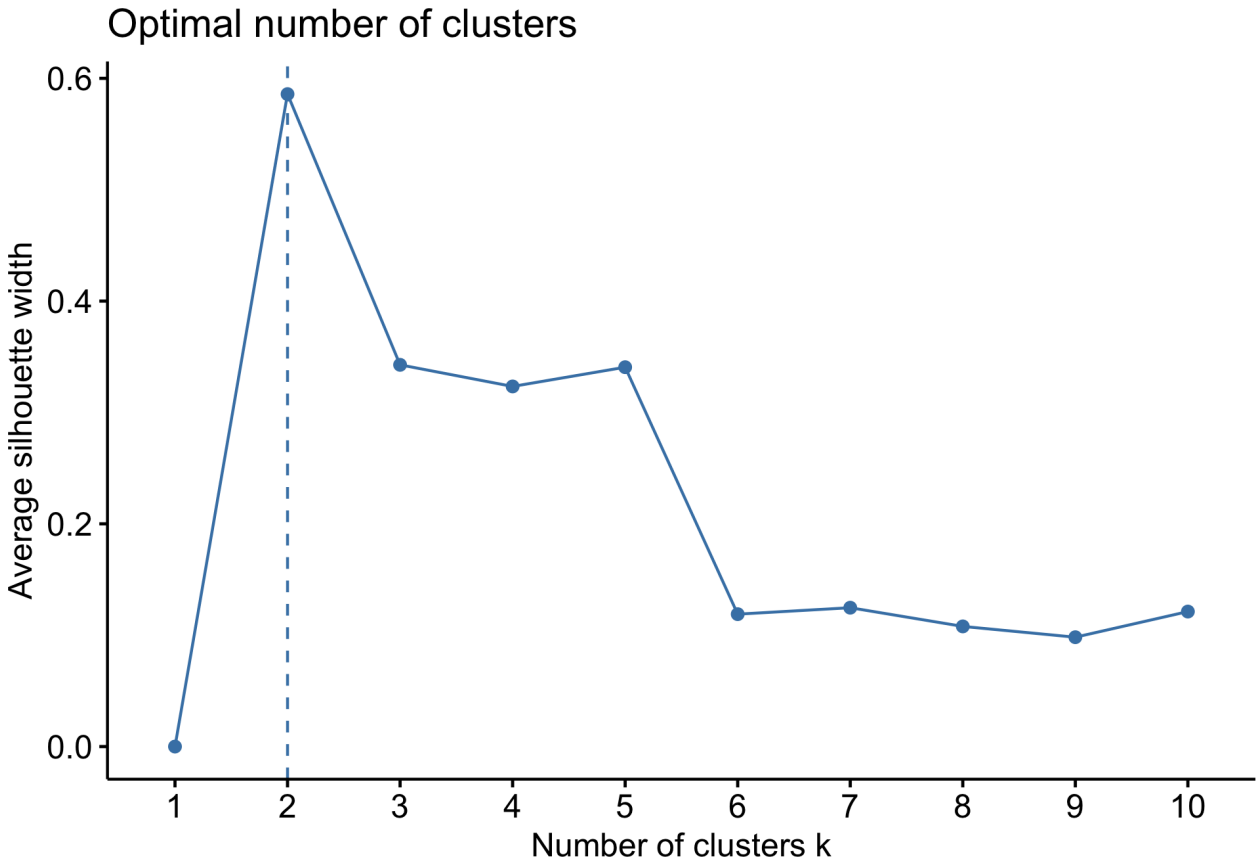
```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(NbClust)

fifa_cs <- as_tibble(scale(fifa))

fviz_nbclust(fifa_cs, kmeans, method = "wss")
```




```
fviz_nbclust(fifa_cs, kmeans, method = "silhouette")
```



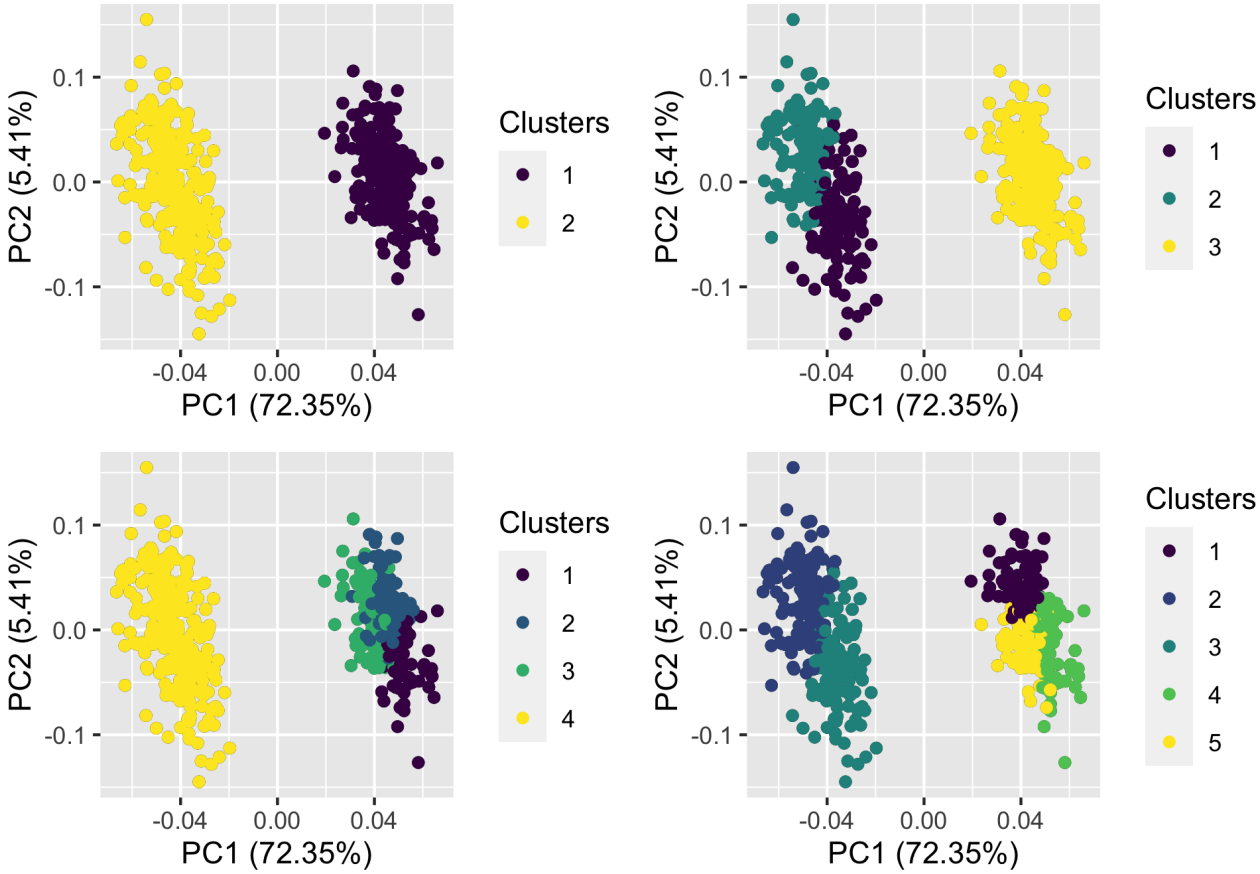
Ambos os métodos sugerem dois clusters para os dados. O método do cotovelo, por exemplo, sofre uma quebra abrupta de 1 para dois clusters, e estabiliza logo em seguida. O método da silhueta indica o máximo em 2 clusters. Portanto, ambos os métodos sugerem a divisão em dois grupos.

Questão 4

(*10 pontos*) Combine os resultados das questões 2 e 3, isto é, visualize o resultado da PCA colorindo os pontos de acordo com a divisão em clusters pelo k-means com 2, 3, 4 e 5 grupos, sem inserir os loadings. Estas visualizações alteram aquilo concluído na Questão 3?

```
# semente aleatoria fixada para reprodutibilidade

set.seed(1011)

# criacao das clusterizacoes

cluster_2 <- kmeans(scale(fifa_cs), 2)$cluster

g2 <- autoplot(fifa_pca) +
  geom_point(aes(colour = as.factor(cluster_2))) +
  labs(colour = "Clusters") +
  scale_colour_viridis_d()

cluster_3 <- kmeans(scale(fifa_cs), 3)$cluster

g3 <- autoplot(fifa_pca) +
  geom_point(aes(colour = as.factor(cluster_3))) +
```

```
  labs(colour = "Clusters") +
  scale_colour_viridis_d()

cluster_4 <- kmeans(scale(fifa_cs), 4)$cluster

g4 <- autoplot(fifa_pca) +
  geom_point(aes(colour = as.factor(cluster_4))) +
  labs(colour = "Clusters") +
  scale_colour_viridis_d()

cluster_5 <- kmeans(scale(fifa_cs), 5)$cluster

g5 <- autoplot(fifa_pca) +
  geom_point(aes(colour = as.factor(cluster_5))) +
  labs(colour = "Clusters") +
  scale_colour_viridis_d()

library(gridExtra)

grid.arrange(g2, g3, g4, g5)
```



Graficamente, 2 clusters parece ser a melhor solução para esse problema, concordando com aquilo que os métodos do cotovelo e silhueta sugerem. As outras divisões acabam parecendo artificiais, com divisões que parecem não existir na prática.

Questão 5

(*10 pontos*) Coloque as informações sobre a clusterização final no conjunto de dados original, sem as transformações aplicadas. Calcule a média, por cluster, das variáveis consideradas no estudo. Determine qual cluster equivale a cada posição no campo. Lembre-se que as posições no campo disponíveis para clusterizar os jogadores são

- Goleiro (G)
- Defesa (D)
- Meio-Campo (M)
- Ataque (A)

Não é necessário utilizar todos as posições na sua resposta. Algumas delas podem sobrar (ou não). Justifique a sua resposta.

```
fifa_final <- bind_cols(fifa,
                        cluster = cluster_2)

fifa_final %>%
  group_by(cluster) %>%
  summarise_all(mean) %>%
  glimpse()
```

```
## Rows: 2
## Columns: 38
## $ cluster          <int> 1, 2
## $ age              <dbl> 29.75410, 28.21073
## $ height           <dbl> 189.7910, 183.5747
## $ weight           <dbl> 84.08607, 78.85057
## $ crossing         <dbl> 14.93852, 61.24904
## $ finishing        <dbl> 13.00000, 80.36015
## $ heading_accuracy <dbl> 14.64754, 75.91571
## $ short_passing    <dbl> 32.89754, 72.88506
## $ volleys          <dbl> 13.55738, 74.78161
## $ dribbling        <dbl> 15.99590, 76.30651
## $ curve            <dbl> 16.12295, 66.99234
## $ fk_accuracy      <dbl> 15.04098, 57.37165
## $ long_passing     <dbl> 31.66393, 59.27969
## $ ball_control     <dbl> 23.43852, 77.74330
## $ acceleration     <dbl> 45.09836, 74.14176
## $ sprint_speed     <dbl> 45.19672, 75.97318
## $ agility          <dbl> 47.18852, 72.24521
## $ reactions        <dbl> 73.39754, 77.64368
## $ balance          <dbl> 45.09836, 67.45594
## $ shot_power       <dbl> 54.31967, 80.40613
## $ jumping          <dbl> 64.96721, 75.63218
## $ stamina          <dbl> 34.68033, 72.68966
## $ strength         <dbl> 66.80738, 78.14176
## $ long_shots       <dbl> 13.89754, 72.92337
## $ aggression       <dbl> 29.31148, 67.58621
## $ interceptions    <dbl> 19.57787, 34.04215
## $ positioning      <dbl> 12.48361, 80.69732
## $ vision           <dbl> 48.95492, 69.54406
## $ penalties        <dbl> 22.28689, 73.87356
## $ composure        <dbl> 54.70082, 75.77395
## $ marking          <dbl> 17.47541, 34.58238
```

```
## $ standing_tackle  <dbl> 14.94672, 33.93103
## $ sliding_tackle   <dbl> 14.35656, 28.47893
## $ gk_diving        <dbl> 77.96311, 10.04215
## $ gk_handling      <dbl> 75.12295, 10.86973
## $ gk_kicking       <dbl> 72.23361, 10.39847
## $ gk_positioning   <dbl> 77.00410, 10.37165
## $ gk_reflexes      <dbl> 79.45082, 10.19540
```

É possível perceber que, no cluster 1, a média das variáveis relacionadas às habilidades de goleiros (`gk_diving`, `gk_handling`, `gk_kicking`, `gk_positioning` e `gk_reflexes`) são bastante altas, na faixa entre 70 e 80. Isso nos sugere que o cluster 1 é formado por goleiros.

Por outro lado, as variáveis acima de 80 para o outro cluster são `finishing`, `shot_power` e `positioning`, todas relacionadas principalmente a habilidades de atacantes. Portanto, os jogadores que formam o cluster 2 são atacantes.

Parte II: Webscraping

A página *List of countries and dependencies by population* da Wikipedia - https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population - exibe uma lista de países de acordo com a sua população atual ou estimativa mais recente. Além disso, a página *List of countries by population in 2015* - https://en.wikipedia.org/wiki/List_of_countries_by_population_in_2015 - exibe uma lista análoga, referente ao ano de 2015. O objetivo deste problema é juntar os dados destas duas páginas e analisar o crescimento populacional de acordo com as informações coletadas.

Questão 6

(*10 pontos*) Baixe a tabela principal disponível na página *List of countries and dependencies by population*. Mantenha apenas as colunas referentes ao país, região e população no final, renomeando-as como `pais`, `regiao` e `populacao`. Deixe a tabela limpa e pronta para ser analisada.

```
library(rvest)
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(tidyverse)
theme_set(theme_bw())
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
url_populacao_atual <- "https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population"

populacao_atual <-
  read_html(url_populacao_atual) %>%
  html_table()

populacao_atual <- populacao_atual[[1]] %>%
  clean_names()

populacao_atual <-
  populacao_atual  %>%
  filter(region != "World") %>%
  select(pais = country_or_dependent_territory,
         regiao = region,
         populacao = population) %>%
  mutate(pais = str_replace_all(pais, "\\[.*\\]", ""),
         populacao = as.numeric(str_replace_all(populacao, ",", "")))
```

Questão 7

(*10 pontos*) Baixe a tabela principal disponível na página *List of countries by population in 2015*. Mantenha apenas as colunas referentes ao país, area e população no final, renomeando-as como `pais`, `area` e `populacao`.

Deixe a tabela limpa e pronta para ser analisada.

```
url_populacao_2015 <- "https://en.wikipedia.org/wiki/List_of_countries_by_population_in_2015"

populacao_2015 <-
  read_html(url_populacao_2015) %>%
  html_table()

populacao_2015 <- populacao_2015[[2]] %>%
  clean_names()

populacao_2015 <-
  populacao_2015  %>%
  filter(country_territory != "World") %>%
  head(-1) %>%
  select(pais = country_territory,
         area = area_km2_1,
         populacao = population2015_un_estimate) %>%
  mutate(pais = str_replace_all(pais, "\\[.*\\]", ""),
         area = as.numeric(str_replace_all(area, ",", "")),
         populacao = as.numeric(str_replace_all(populacao, ",", "")))
```

Questão 8

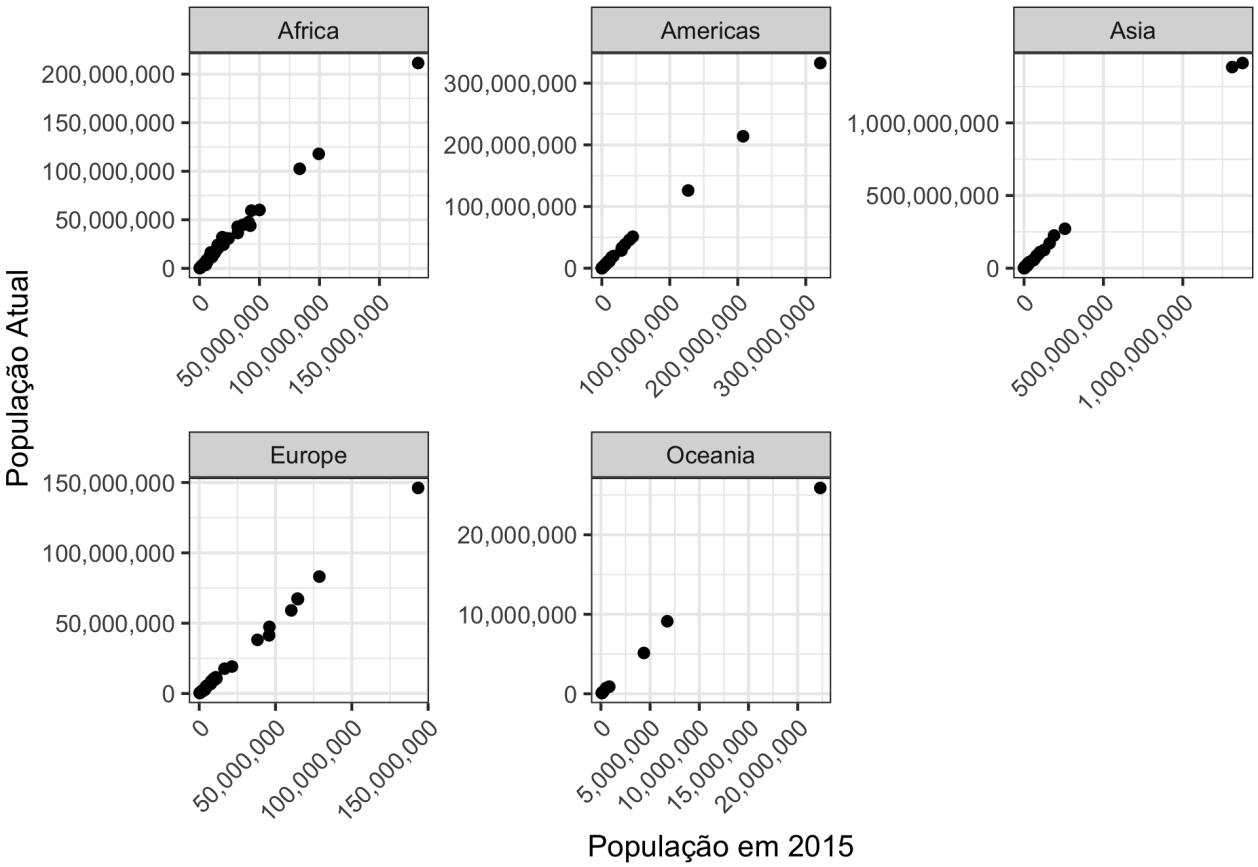
(*10 pontos*) A quantidade de países disponível na tabela mais atual é diferente da quantidade disponível na tabela de 2015. Com isso em mente, crie um objeto chamado `tabela_final`, juntando os dados das duas tabelas e mantendo na tabela final apenas os países com dados de população disponíveis para os dois anos. Note que a tabela final deverá ter 5 colunas.

```
tabela_final <-
  populacao_atual %>%
  left_join(populacao_2015, by = "pais") %>%
  select(pais, regiao, area,
         populacao_atual = populacao.x,
         populacao_2015 = populacao.y) %>%
  filter(!is.na(populacao_2015))
```

Questão 9

(*10 pontos*) Crie gráficos adequados para comparar a evolução da população atual em relação à população de 2015 por região. Os devem ser fáceis de interpretar, tanto em relação à utilização da área gráfica quanto às legendas. Descreva o que é possível perceber nesta visualização.

```
ggplot(tabela_final, aes(x = populacao_2015, y = populacao_atual)) +
  geom_point() +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  facet_wrap(~ regiao, scales = "free") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  labs(x = "População em 2015", y = "População Atual")
```



Aparentemente há uma relação linear no crescimento da população dos países entre 2015 e os dados mais atuais. Aparentemente não há, entre os países analisados, algum cujo crescimento populacional tenha variado muito em relação à tendência observada em sua região.

Questão 10

(*10 pontos*) Calcule a variação percentual entre os dados mais atuais em relação a 2015. Coloque em ordem decrescente a média destas variações por região, informando qual região teve os países com populações que mais cresceram, em média, durante este intervalo de tempo. Qual foi a variação média entre os países desta região?

```
tabela_final %>%
  mutate(variacao = (populacao_atual - populacao_2015)/populacao_2015*100) %>%
  group_by(regiao) %>%
  summarise(media = mean(variacao)) %>%
  arrange(desc(media))
```

```
## # A tibble: 5 x 2
##   regiao    media
##   <chr>     <dbl>
## 1 Africa   26.0
## 2 Asia     17.1
## 3 Oceania  16.3
## 4 Americas 10.1
## 5 Europe   -0.326
```

Com 26%, a África foi a região na qual a média de variação da população dos países mais aumentou de 2015 para cá.