Serviço Público Federal
Universidade Federal do Pará
Instituto de Ciências Exatas e Naturais
Faculdade de Estatística

Dionisio Alves da Silva Neto
Matrícula: 202007840008

Atividade 3 de Análise Multivariada II:
Análise de Componentes Principais (PCA)

Belém, PA
2022

1. Banco de dados

O heptatlo é composto por sete diferentes provas e nasceu da evolução do pentatlo. As atividades realizadas no circuito compreendem: 100 metros com barreiras, Arremesso de peso, Salto em altura, Salto de 200 metros rasos, Salto em distância, Arremesso de Dardo e 800 metros rasos. Todas as etapas foram consolidadas pela Federação Internacional de Atletismo (IAAF) em 1981 e, a vencedora do circuito é a participante que mais acumula pontos, em dois dias de jogos, para o cálculo do escore final. Em 1988, o heptatlo teve a sua segunda realização dos jogos olímpicos de Seul, capital da Coreia do Sul, tendo como campeã uma das estrelas do atletismo feminino a cidadã americana Jackie-Kersee.

Na base de dados disponibilizada, temos informações de 25 competidores, assim como o desempenho das participantes em cada prova, junto com o seu escore final:

- *hurdles*: resultados de 100 m com barreiras.
- *highjump*: resultados de salto em altura.
- *shot*: resultados de arremesso de peso.
- *run200m*: resultados de 200 m rasos.
- *longjump*: resultados de salto em distância.
- *javelin*: resultados de lançamento de dardos.
- *run800m*: resultados de 800 m rasos.
- *score*: pontuação total.

2. Análise descritiva das variáveis

Antes de qualquer aplicação de modelos estatísticos ou técnicas multivariadas, é de suma importância conhecer o banco de dados através de uma análise descritiva. Desse modo, a **Tabela 1** tem por objetivo apresentar os valores do mínimo, do primeiro quartil (Q1), da mediana, da média, do terceiro quartil (Q3), do máximo e do desvio-padrão padrão para cada variável disponível.

Em análise, percebe-se que o vetor de médias e a mediana apresentam uma discrepância para algumas variáveis. As informações vindas de *shot*, *longjump*, *javelin*, *run800m* e *score* são bem maiores do que o restante das outras variáveis do estudo. Por outro lado, a variabilidade dos valores,

expresso pelo desvio-padrão, está bem controlada, com exceção do escore que apresenta uma grandeza perto de 113,69. Pelos valores do mínimo e máximo para o escore, temos que a menor e maior nota foram 4566 e 7291, respectivamente.

Tabela 1: Valores descritivos para as variáveis do estudo, por estatística.

Estatística	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Mínimo	0,00	1,50	10,00	0,00	4,88	35,70	0,00	4566
Q1	2,35	1,77	12,30	1,38	6,05	39,10	24,90	5746
Mediana	2,67	1,80	12,90	1,78	6,25	41,30	28,70	6137
Média	2,58	1,78	13,10	1,96	6,15	41,50	27,40	6091
Q3	2,95	1,83	14,20	2,69	6,37	44,50	31,20	6351
Máximo	3,73	1,86	16,20	4,05	7,27	47,50	39,20	7291
Desvio-padrão	0,15	0,02	0,30	0,19	0,09	0,71	1,66	113,69

Fonte: Construído pelo autor, 2022.

Na análise gráfica do fenômeno, a **Figura 1** aborda o diagrama de dispersão, gráfico de densidade e correlações lineares para as variáveis do presente trabalho.

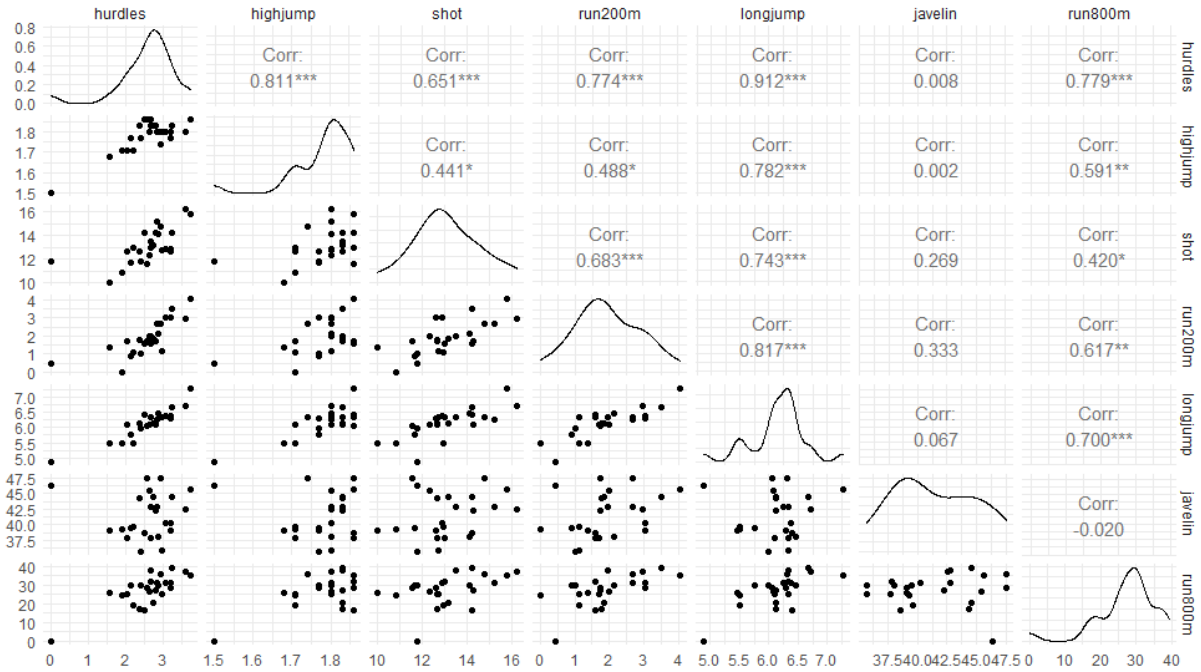
Em primeira instância, pode-se perceber que nem todas as variáveis apresentam uma forte correlação linear (acima de 0,90), com exceção do par *longjump* e *hurdles*, as quais são os salto em distância e 100 metros rasos, nesta respectiva ordem.

Em adição, ao verificar a dispersão entre os pares de variáveis na **Figura 1**, temos que muitos deles apresentam uma nuvem de pontos, o que indica uma baixa associação linear, mas é elucidativo que a variável hurdles (salto em distância) é a que mais tende a formação de uma reta na combinação com outras variáveis.

Para o gráfico de densidade , na diagonal da **Figura 1,** os dados para cada variável mostram não estarem distribuídos assimétricamente. Logo, A variável *hurdles* concentra valores entre 2,5 e 3; a variável *highjump* apresenta dois picos em seu gráfico de densidade, em 1,7 e 1,8; a variável *shot* tem uma alta variabilidade com um centro em 13; a variável *run200m* também apresenta uma alta variabilidade; a variável *longjump* tem duas concentrações em seu

gráfico de densidade, em 5,5 e 5,5; a variável *javelin* não tem um valor central definido e; a variável *run800m* mostra dois valores centrais, em 18 e 30.

Figura 1: Diagrama de dispersão, gráfico de densidade e correlações para as variáveis do estudo.



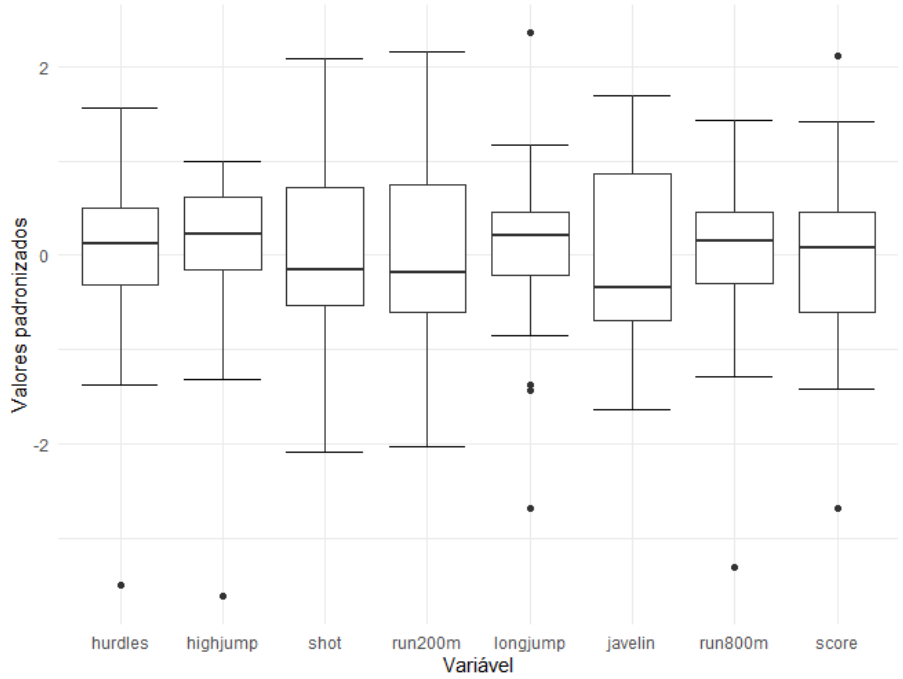
Fonte: Construído pelo autor, 2022.

Como na análise das medidas descritivas de cada variável do estudo percebemos que cada uma pode ter sido medida em diferentes grandezas, é plausível abordar o conjunto geral em uma mesma escala. Portanto, é preciso realizar a padronização de modo que os resultados de qualquer técnica sejam coerentes.

Com resultado da padronização para uma normal padrão, isto é, com média igual a 0 e desvio padrão igual a 1, a **Figura 2** aborda a construção do gráfico boxplot para cada variável do estudo.

Em vista disso, podemos verificar a presença de 4 outliers para os dados padronizados do salto em distância (*longjump*), 2 outliers nos valores padronizados da variável *score* (escore), e um outlier nos dados padronizados das variáveis *hurdles*, *highjump* e *run800m* individualmente. Apesar da padronização realizada, nota-se a presença de uma alta variabilidade em todas medições do heptatlo.

Figura 2: Boxplots para os valores padronizados das variáveis do estudo.

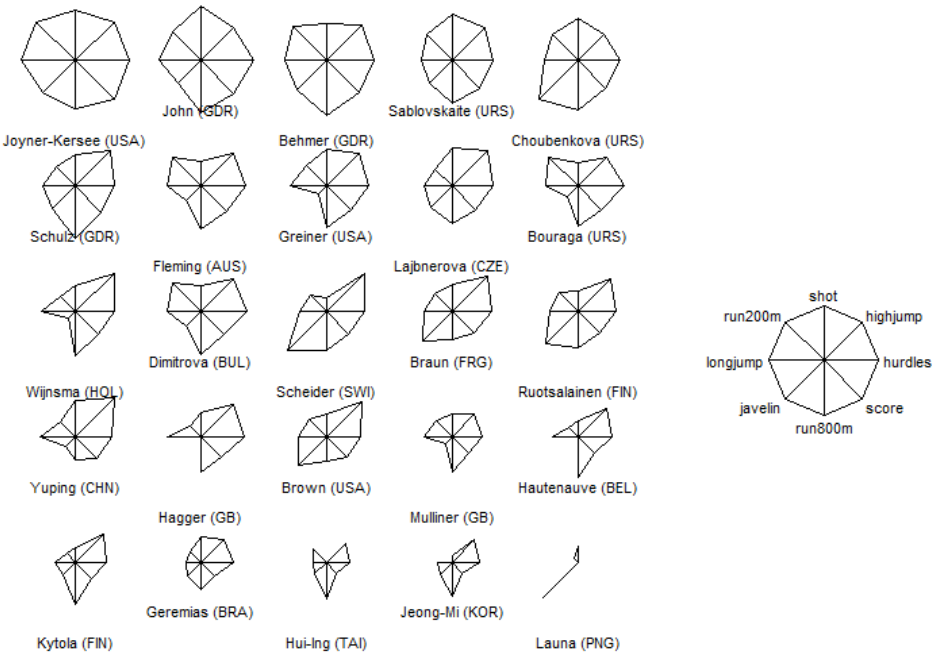


Fonte: Construído pelo autor, 2022.

A **Figura 3** mostra o gráfico de estrela, o qual é uma ferramenta importante para avaliar o desempenho de cada atleta nas fases do heptatlo. Dessa maneira, temos que a estrela mais à direita representa a direção para o desempenho de cada fase do circuito e, quanto mais "puxado" para uma direção, maior é o desempenho da atleta na modalidade.

Assim sendo, é notório o desempenho deslumbrante da vencedora do heptatlo, a americana Joyner-Kersee, ao verificarmos que a sua capacidade em todas as modalidades foi excepcional em comparação às outras competidoras. Acompanhado da vencedora do circuito, vemos que duas participantes da Alemanha (GDR) e outras duas da União Soviética (URS), também tiveram um grande desempenho, como pode ser observado nas 5 primeiras estrelas do gráfico. Em adição, temos que a prova mais distenso para as competidoras foi o salto em distância (*highjump*) e a mais complicada foi o lançamento de dardos (*javelin*).

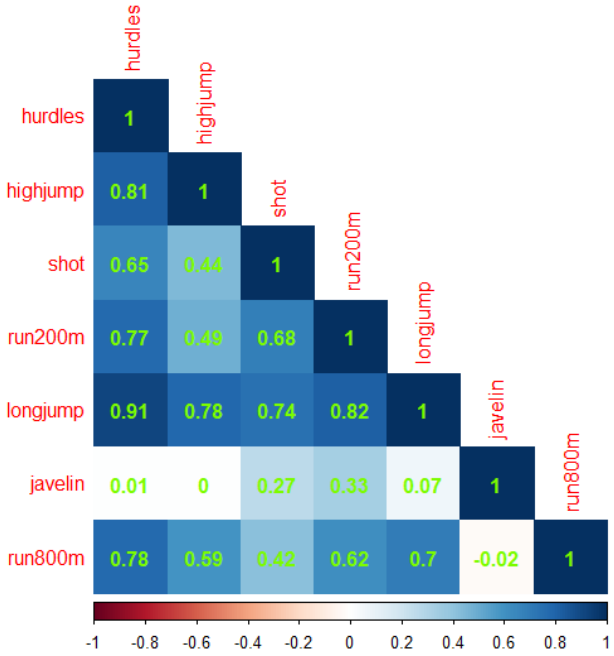
Figura 3: Gráfico de estrela para os indivíduos da base de dados.



Fonte: Construído pelo autor, 2022.

A **Figura 4** mostra o mapa de calor para a matriz de correlações amostrais entre as variáveis do presente trabalho. Com isso, nota-se a ausência de altas correlações no banco de dados, fato que pode afetar a aplicação de técnicas multivariadas, como a análise de componentes principais e a análise fatorial, as quais necessitam deste pressuposto. As correlações estão acima de 0,50, com exceção dos pares que contém a variável *javelin* (lançamento de dardos), a qual apresenta valores bem próximos a zero quando correlacionadas linearmente com as outras modalidades.

Figura 4: Correlograma para as variáveis do estudo.



Fonte: Construído pelo autor, 2022.

3. Análise de Componentes Principais (PCA).

A **Seção 2** deste trabalho, pela Análise Exploratória dos Dados (AED), nos revelou as características peculiares de cada modalidade presente no banco de dados e o desempenho das atletas no heptatlo. Nesta presente secção, iremos aplicar a Análise de Componentes Principais (ACP, ou do inglês, PCA), a qual consiste em uma técnica multivariada para a redução de dimensionalidade e transformação de variáveis.

Um importante direcionamento ao aplicar esta ferramenta é a presença de altos valores para a correlação linear de Pearson. No entanto, na **Figura 4**, concluímos que poucos pares de variáveis apresentam tal característica; mesmo assim, aplicamos o PCA nos dados disponíveis.

A **Tabela 2** apresenta o resumo da importância de cada componente formada a partir das variáveis explicativas de cada modalidade do heptatlo, desconsiderando o escore (*score*), pois o objetivo do projeto é perceber como as modalidades conseguem agregar-se entre si para examinar de forma diferente a participação de cada fase.

Dessa maneira, a **Tabela 2**, mostra a variância de cada componente, a qual é o autovalor descoberto para a matriz de variância e covariâncias, a proporção de variância que cada componente consegue explicar sobre a

variabilidade do banco original e a proporção de variância acumulada, a qual serve para nos informar a soma da explicação da variância a cada nova componente. Nos estudos de Análise de Componentes principais, normalmente, é de preferência utilizar duas, ou até três componentes, que possam explicar acima de 70% da variabilidade do conjunto de dados. Na atual situação, observamos que ao utilizar a primeira e segunda componente, temos cerca de 80,78% de explicação sobre a variabilidade original, tal fato reitera uma das principais funções da PCA sobre trabalhar com menos variáveis do que o banco de dados inicial.

Tabela 2: Importância de cada componente formada.

	C1	C2	C3	C4	C5	C6	C7
Variância	4,4603	1,1943	0,5210	0,4572	0,2453	0,0730	0,0490
Proporção de variância explicada	0,6372	0,1706	0,0744	0,0653	0,03504	0,0104	0,0070
Proporção de variância acumulada	0,6372	0,8078	0,8822	0,9475	0,9826	0,9930	1,0000

Fonte: Construído pelo autor, 2022.

Além dos autovalores, é de suma importância visualizar os autovetores da matriz de variância e covariâncias formada no decorrer da técnica. Nesse viés, a **Tabela 3** mostra os valores dos autovetores padronizados para cada variável; é preciso padronizar tais vetores para realizar a transformação de cada linha do banco de dados original para os valores das novas variáveis (componentes).

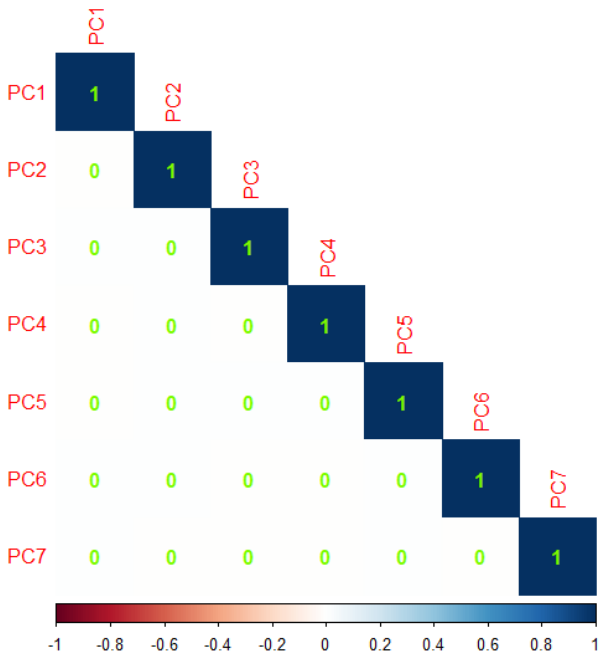
Tabela 3: Autovetores padronizados de cada componente.

	C1	C2	C3	C4	C5	C6	C7
hurdles	-0,9564348	0,17258347	-0,03258984	0,01794393	-0,04702239	-0,21158268	0,08417214
highjump	-0,7966208	0,27110745	-0,26546771	0,45977037	0,00931003	0,02684817	0,08417214
shot	-0,7667860	-0,31627884	0,48808221	0,08405600	0,25339264	-0,01373738	-0,04817382
run200m	-0,8614484	-0,28456217	0,06033788	-0,24413144	-0,32182648	0,00674079	-0,10036204
long jump	-0,9635326	0,06106182	0,10056049	0,07524943	-0,09127255	0,15941736	0,13548751
Javelin	-0,1592590	-0,91984301	-0,34037829	0,08167733	0,06691079	-0,00735780	0,03828377
run800m	-0,791890	0,24533367	-0,28573455	-0,4079912	0,24976209	0,04201591	-0,02176199

Fonte: Construído pelo autor, 2022.

A **Figura 5** aborda o correlograma para as componentes formadas em função do banco de dados heptatlo. Dessa forma, constatamos a propriedade de independência para as novas variáveis formadas, pelo fato da correlação linear entre qualquer par valer zero.

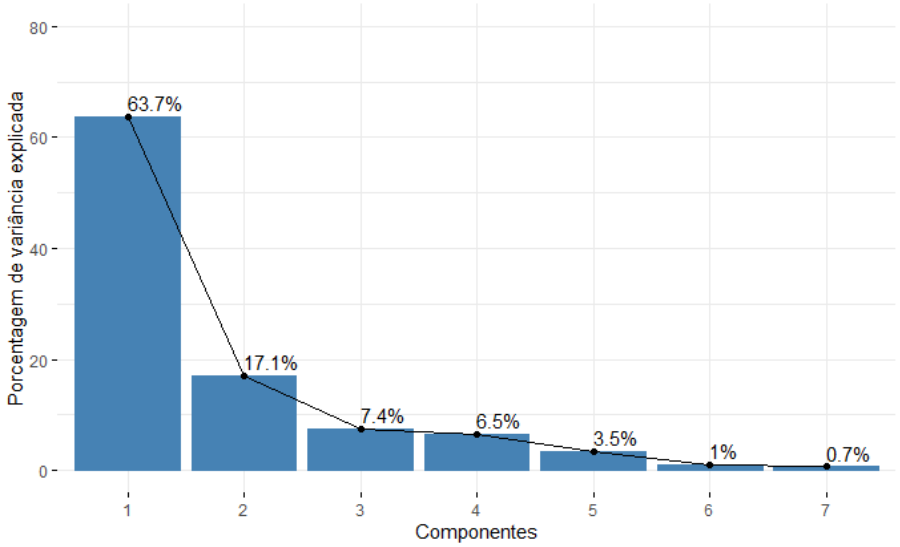
Figura 5: Correlograma entre as componentes principais.



Fonte: Construído pelo autor, 2022.

A **Figura 6** ratifica os resultados observados na **Tabela 2**, ao elucidar o Scree plot para as componentes. De fato, verifica-se que a primeira e a segunda componente explicam 63,70% e 17,10% da variabilidade dos dados originais, nesta respectiva ordem. Também, tem-se a tendência de decrescimento a partir da primeira componente até a última criada.

Figura 6: Scree plot para o grau de importância das componentes formadas.



Fonte: Construído pelo autor, 2022.

Um modo de se avaliar a contribuição de cada variável para a formação da componente, é construindo um mapa de calor para as correlações lineares entre os valores das observações iniciais e os valores para a componente em questão, a qual assume o papel de nova variável.

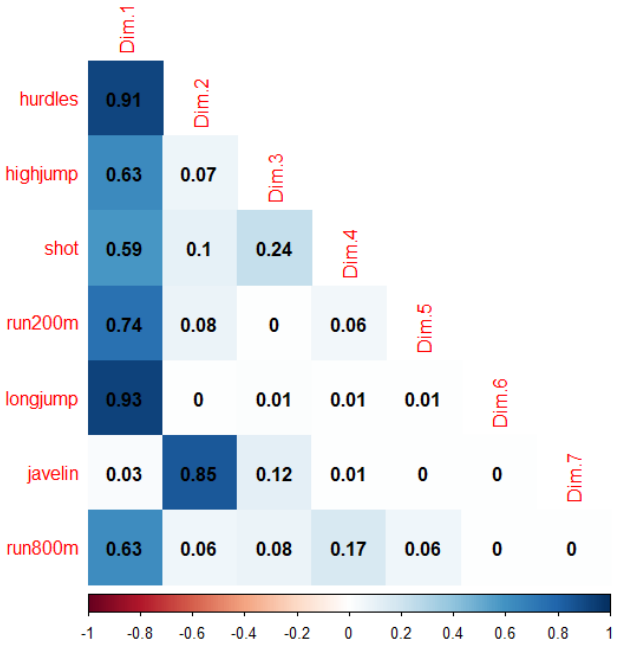
Em vista disso, a **Figura 7** aborda o correlograma entre os valores iniciais das variáveis e os novos formados para as componentes. Logo, podemos concluir que a primeira componente consegue agregar as variáveis *hurdles*, *highjump*, *shot*, *run200m*, *longjump* e *run800m*, em consequência, podemos ver que a nova variável tem a capacidade de explicar conjuntamente a performance das atletas para as modalidades 100 metros rasos, arremesso em peso, 200 metros rasos, salto em distância e 800 metros rasos.

Em segundo plano, a **Figura 7** também mostra que a variável *highjump*, a qual representa o lançamento de dardos, é a única geradora de informação para a formação da segunda componente, estando mais afastada das demais. Este resultado já era uma hipótese ao analisar o gráfico de estrela na **Figura 3**, na qual o lançamento de dardos não era performado com maestria pelas atletas que dominavam as outras modalidades.

No geral, é importante reiterar que se houve maior correlações lineares entre as variáveis, seria possível ter uma maior contribuição para a formação das componentes, em outras palavras, quanto maior é a multicolinearidade do

banco de dados original, maior é a ocorrência de agregação entre as variáveis na componente.

Figura 7: Correlograma entre as variáveis do estudo e as componentes formadas.



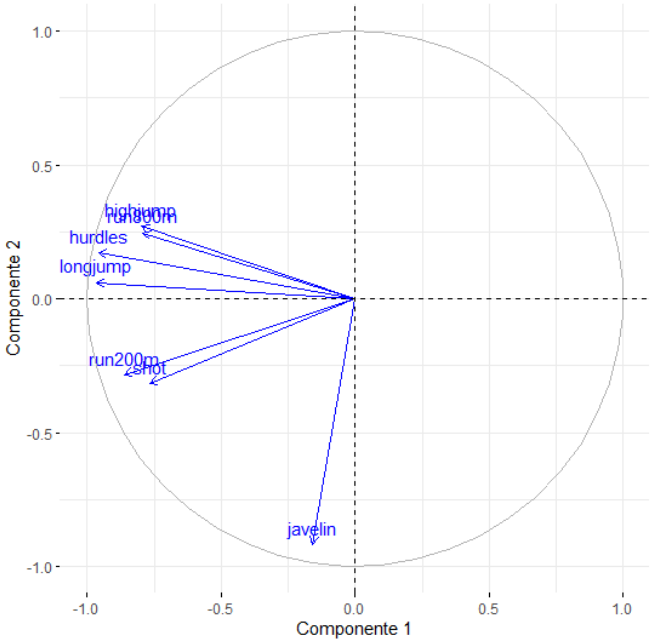
Fonte: Construído pelo autor, 2022.

Como as componentes 1 e 2 conseguem explicar cerca de 80,78% da variabilidade do banco heptatlo, o gráfico das cargas fatoriais representado na **Figura 8** tem por objetivo mostrar a contribuição de cada variável do banco original para a formação da componente. Desse modo, percebe-se que a variável que mais contribui para a formação da primeira componente é a longjump (salto em distância), seguida por *hurdles* (100 metros com barreiras), *run800m* (800 metros rasos), *highjump* (salto em altura) e *shot* (arremesso). A variável *javelin* (lançamento de dardos) é a que encontra-se mais isolada e é o determinante para a formação da segunda componente.

Na visualização entre os ângulos forados entre as cargas fatoriais, percebe-se que a correlação entre *javelin* (lançamento de dardos) e *run800m* (800 metros rasos) é inexistente, pelo ângulo formado entre as suas retas ser de 90 graus. Por outro lado, percebe-se que *highjump* (salto em altura), *run800m* (800 metros rasos), *hurdles* (100 m com barreiras) e *longjump* (salto em distância) estão positivamente correlacionadas, pelo fato dos ângulos entre as suas retas serem bem pequenos; assim como o par *run200m* (200 m rasos)

e *shot* (arremesso de peso), o qual também mostra um baixo valor para o ângulo formado entre tais variáveis.

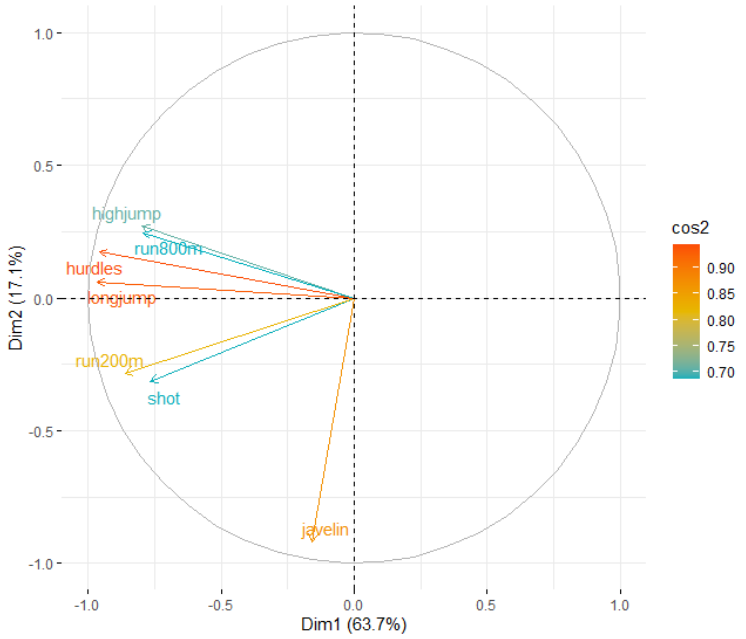
Figura 8: Gráfico das cargas fatoriais para as componentes 1 e 2.



Fonte: Construído pelo autor, 2022.

A **Figura 9** é uma adição à análise das cargas fatoriais visto anteriormente, ao representar em conjunto o grau de importância de cada variável para a sua respectiva componente. Vale ressaltar que os valores de importância variam de 0 a 1 indicando a maior e menor relevância para a construção da componente, nesta respectiva ordem. Destarte, temos que para a componente 1, as variáveis *longjump* (salto em distância) e *hurdles* (100 m com barreiras) são as que apresentam maior contribuição ao ter um grau acima de 0,9. Na segunda componente, temos que a variável *javellin* (lançamento de dardos) apresenta uma participação acima de 0,80.

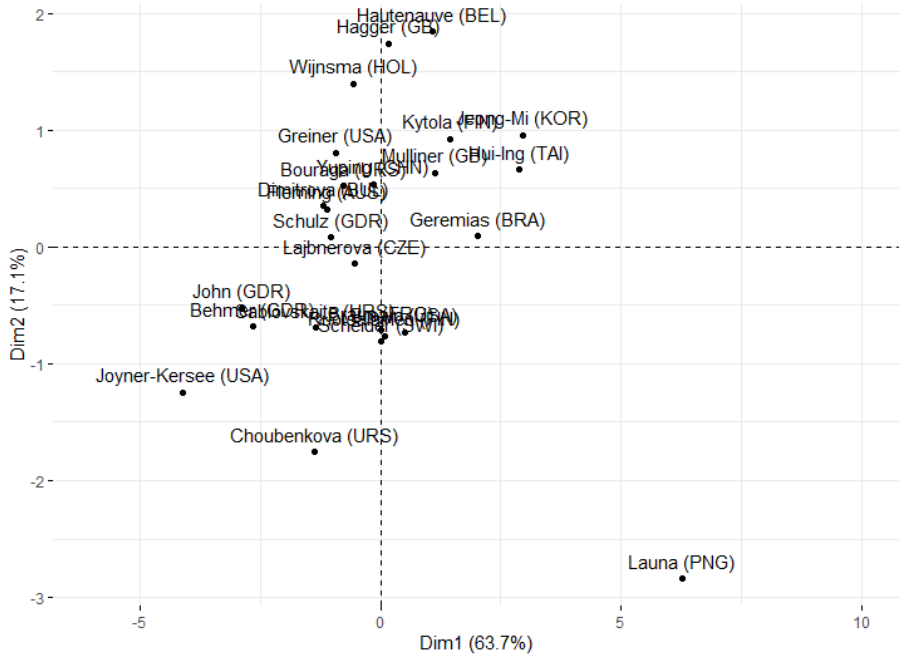
Figura 9: Gráfico das cargas fatoriais para as componentes 1 e 2, com o grau de importância.



Fonte: Construído pelo autor, 2022.

Na **Figura 10**, é possível observar o gráfico de dispersão de cada indivíduo da base de dados para os seus respectivos valores para as duas primeiras componentes principais, esta representação gráfica serve para avaliar grupos, valores discrepantes e tendências. Nota-se a dispersão entre as atletas do heptatlo, com um outlier presente representado pela participante Launa e a formação de um grupo central para várias participantes. A vencedora do circuito, a estadunidense Joyner-Kersee, encontra-se distante de agrupamentos localizando-se mais próximo da candidata Russa Choubenkova, cuja teve um desempenho considerável no heptatlo.

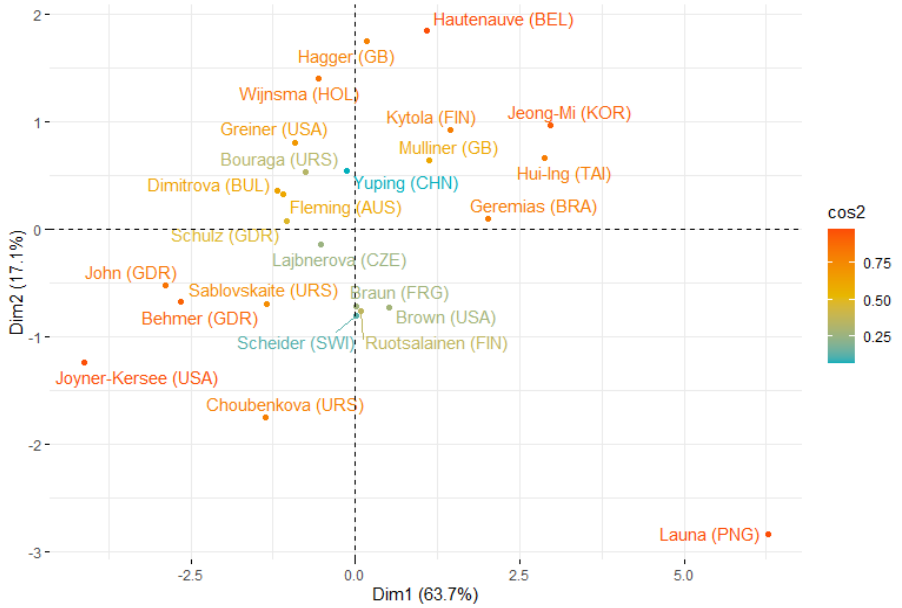
Figura 10: Gráfico de dispersão individual entre as componentes 1 e 2.



Fonte: Construído pelo autor, 2022.

Em complemento, a **Figura 11** mostra o grau de importância de cada participante da base de dados para a formação das componentes. Desse modo, temos que as competidoras Yuping e Schneider geram pouca informação para a composição das componentes. Em contraste, as participantes Launa, Joyner-Kersee e Hautenauve são as que mais geram informações de variabilidade para a criação das componentes.

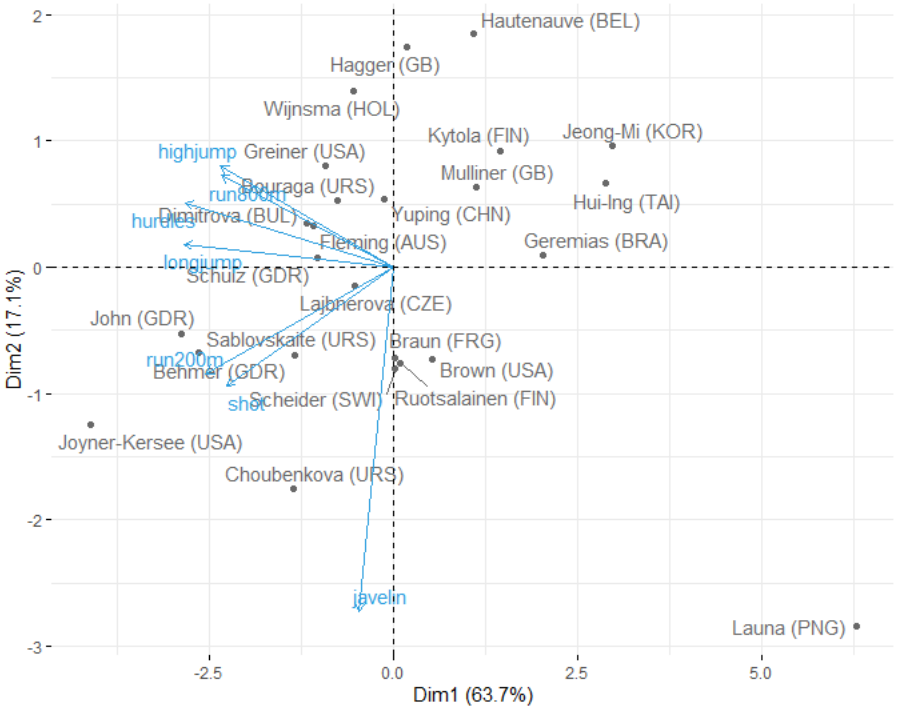
Figura 11: Gráfico de dispersão individual entre as componentes 1 e 2, com o grau de importância.



Fonte: Construído pelo autor, 2022.

O gráfico Biplot, representado pela a aplicação do PCA nos dados do circuito heptatlo, na **Figura 12**, combina as informações das cargas fatoriais com a dispersão do escore das componentes para cada indivíduo da base de dados. Novamente, temos a interpretação de que as setas paralelas são as que trazem contribuição maior para a sua componente; as setas mais longas representam as variáveis que têm a sua variabilidade explicada em sua componente e as os cores individuais servem para observar o desempenho das atletas nas respectivas variáveis agregadas dentro das componentes.

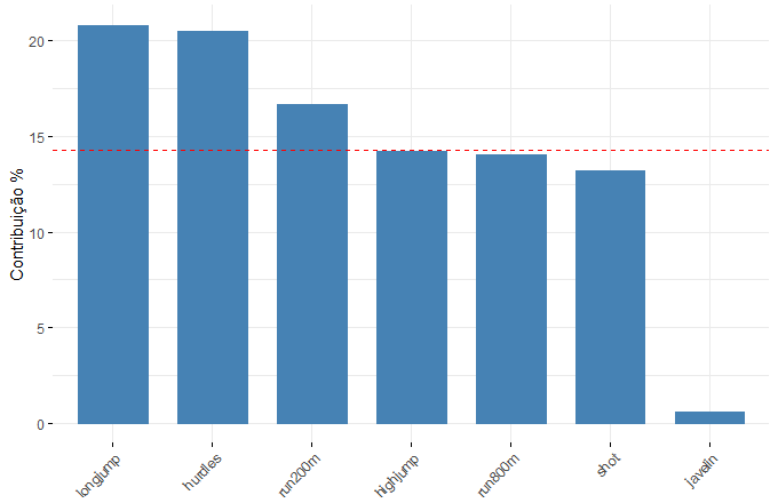
Figura 12: Gráfico de dispersão individual e cargas fatoriais entre as componentes 1 e 2.



Fonte: Construído pelo autor, 2022.

A **Figura 13** aborda um gráfico de barras para evidenciar a contribuição das variáveis originais do banco de dados para a formação da primeira componente. A linha vermelha em destaque serve para mostrar as alturas esperadas para cada barra, se a contribuição fosse uniforme. Para a primeira componente, temos a participação expressiva das variáveis *longjump* (salto em distância) e *hurdles* (100 metros com barreiras), e uma contribuição relevante das variáveis *run200m* (200 metros rasos), *highjump* (salto em altura), *run80m* (800 metros rasos) e *shot* (arremesso de peso).

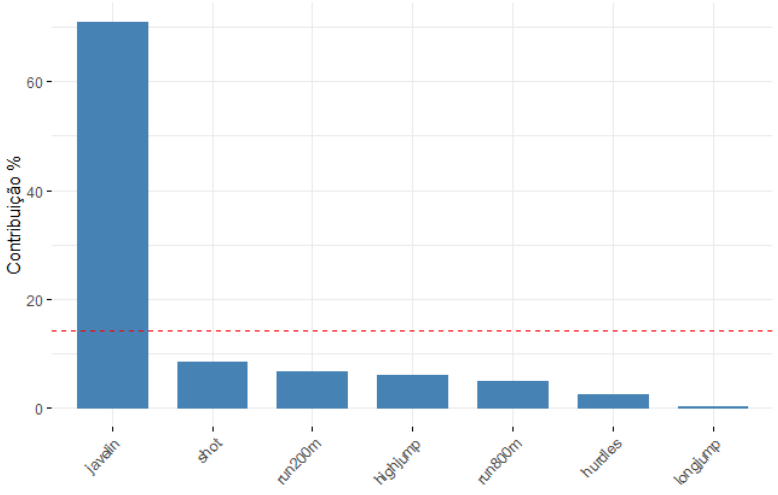
Figura 13: Gráfico de contribuição de cada variável para a formação da primeira componente.



Fonte: Construído pelo autor, 2022.

A **Figura 14** tem por objetivo elucidar a participação das variáveis do estudo para a formação da segunda componente, da mesma forma, a linha em vermelho indica a altura esperada se a contribuição fosse uniforme entre as variáveis. Dessa forma, nota-se que apenas a variável *javellin* (lançamento de dardos) tem uma contribuição alta para a formação desta componente, e as demais variáveis do estudo configuram uma baixa participação.

Figura 14: Gráfico de contribuição de cada variável para a formação da segunda componente.



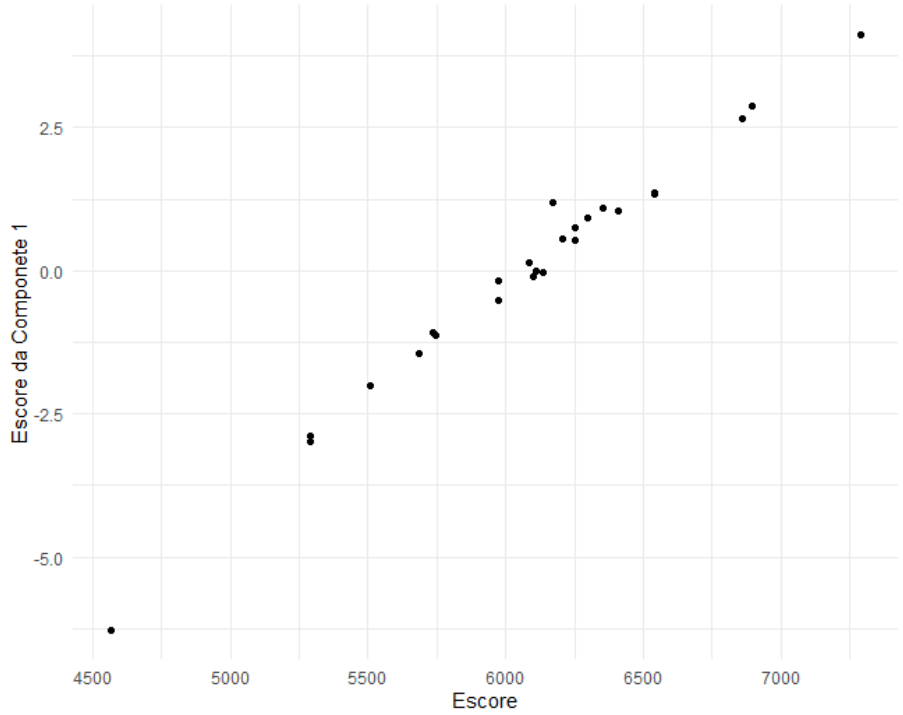
Fonte: Construído pelo autor, 2022.

4. Relação entre as componentes estudadas e o escore do heptatlo.

A fase final deste projeto volta-se a analisar a relação dos escores formados pelas duas componentes principais e o escore real calculado para cada atleta ao final dos dois dias de realização do heptatlo.

Primeiramente, a **Figura 15** ilustra o gráfico de dispersão entre o escore do circuito e o escore gerado pela primeira componente. Deste modo, têm-se a formação de uma reta quase linear entre os pares individuais para as duas variáveis, tal formato deve-se à agregação da informação de variabilidade de 6 das 7 fases do heptatlo.

Figura 15: Gráfico de dispersão entre o escore do heptatlo e os valores da primeira componente.

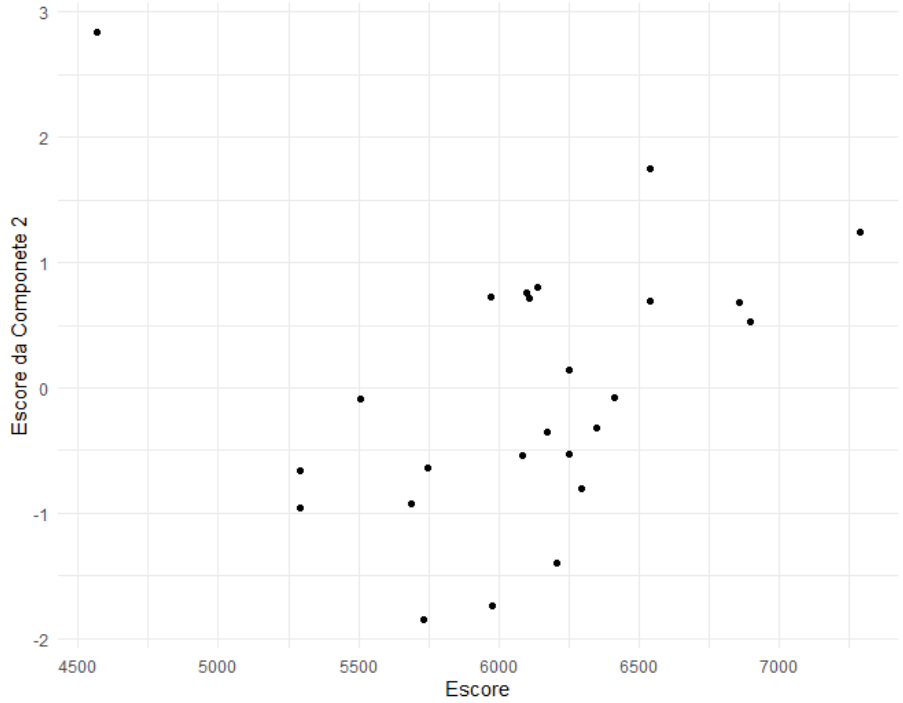


Fonte: Construído pelo autor, 2022.

Em segundo plano, a **Figura 16** aborda o gráfico de dispersão entre o escore total do heptatlo e o escore gerado pela segunda componente. Percebe-se a formação de uma tendência crescente, todavia, a nuvem de dispersão inviabiliza a conclusão de uma relação linear evidente, isto é

explicado pois a segunda componente agrega adequadamente apenas a informação da variabilidade da variável *javellin* (lançamento de dardos).

Figura 16: Gráfico de dispersão entre o escore do heptatlo e os valores da segunda componente.



Fonte: Construído pelo autor, 2022.

Anexo I: Script utilizado na linguagem R.

```
## ------
## Lista 3 de Analise Multivariada II
## Analise de Componentes Principais
## ------


## ---
## Pacotes
## ---

if(!require(pacman)) install.packages("pacman"); library(pacman)
p_load(tidyverse, HSAUR, ggplot2, GGally, corrplot, factoextra, rstatix, psych)


## ---
## Banco de dados
## ---

data("heptathlon")


dados = heptathlon

dados %>% head(10)
dados %>% summary()


## ---
## Objetivos
## ---

# 1. Realizar a Analise de Componentes Principais
# 2. Avaliar os escores das componentes com os escores originais
# 3. Comentar a Matriz de correlcao
# 4. Diagrama de dispersao

plot(heptathlon)

## ---
```

```
## 1. Transformacao dos tempos para que todos representem o maximo
## ---

# hurdles: resultados de 100 m com barreiras --> minimo *
# highjump: resultados de salto em altura   --> maximo
# shot: resultados de arremesso de peso     --> maximo
# run200m: resultados de 200 m rasos        --> minimo *
# longjump: resultados de salto em distância  --> maximo
# javelin: resultados de lançamento de dardos --> maximo
# run800m: resultados de 800 m rasos          --> minimo *
# score: pontuação total

## Temos que transformar as variaveis que indicam o minimo, para que todas
## as variaveis apontem na mesma direcao

dados$hurdles = max(dados$hurdles) - dados$hurdles
dados$run200m = max(dados$run200m) - dados$run200m
dados$run800m = max(dados$run800m) - dados$run800m

options(digits = 5)
dados %>% summary()

describe(dados)

b
## ---
## 2. Estatistica Descritiva
## ---

## desenpenho de cada atleta nas modelidades e no escore
stars(dados, key.loc=c(16,6), cex=0.7, xpd=T, xlim = c(0,17))



## ---
## boxplots
## ---

boxplot(scale(dados))
padronizado = scale(dados)
```

```
padronizado = as.data.frame(padronizado)

ggplot(stack(padronizado), aes(x = ind, y = values)) +
  stat_boxplot(geom ='errorbar') + geom_boxplot() + ylab("Valores padronizados") +
  xlab("Vari?vel")  + theme_minimal()



## ---
## Grafico de correlacao
## ---
cor(dados[,-8]) ## matriz de correlacao

corrplot(cor(dados[,-8]), method = 'color',type = 'lower',
         cl.pos = 'b', addCoef.col = 'lawngreen', xlab ='hhh')



corrplot(cor(scale(dados[,-8])), method = 'color',type = 'lower',
         cl.pos = 'b', addCoef.col = 'lawngreen', xlab ='hhh')


## ---
## Grafico de pares
## ---

ggpairs(dados[,-8]) + theme_minimal()

pairs(dados[,-8], pch = 16) #3 graficos de dispersao


## ---
## relacaoes
## ---

ggpairs(dados[,-8]) + theme_minimal()


## ---
## Analise de Componentes Principais (PCA)
```

```
## ---


pca = prcomp(dados[,-8], scale = TRUE,
             center = TRUE)
pca

summary(pca)

pca$sdev^2


## ---
## Visualiza??o somente dos autovalores
## ---

auto_valores <- get_eigenvalue(pca)
auto_valores

## ---
#Extraindo os resultados do acp para vari?veis
## ---

var <- get_pca_var(pca)
var

## ---
## Correlacoes entre as vari?veis e os PCs
## ---

var$cor

## ---
## coordenadas das vari?veis - correla??o da vari?vel com o CP
## ---

var$coord

## ---
## Contribui??o das vari?veis nos PCs
```

```
## ---

var$contrib

## ---
## Quadrado das correla??es (cos2)- mede a qualidade da representa??o das var?veis no
mapa de calor
## ---
var$cos2


corrplot(var$cos2, method = 'color',type = 'lower',
         cl.pos = 'b', addCoef.col = 'black', xlab ='hhh')


## ---
## Scree plot
## ---

fviz_eig(pca, addlabels = TRUE, main = '', ylim = c(0,80),
         ylab = "Porcentagem de vari?ncia explicada",
         xlab = "Componentes")

## ---
# Grafico de cargas fatoriais
## ---

fviz_pca_var(pca, col.var = "blue",
             xlab = 'Componente 1', ylab = 'Componente 2') + labs(title = '')

## ---
## Grafico de cargas fatoriais com a qualidade das variaveis com ocores
## ---

fviz_pca_var(pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = T) + labs(title = '')

## ---
## Grafico de contribuicao das variaveis na CP1
```

```
## ---
fviz_contrib(pca, choice = "var", axes = 1, top = 10) +
  ylab('Contribui??o %') + labs(title = '')

## ---
## Grafico de contribuicao das variaveis na CP2
## ---
fviz_contrib(pca, choice = "var", axes = 2, top = 10) +
  ylab('Contribui??o %') + labs(title = '')


## ---
## Grafico dos escores individuais
## ---
fviz_pca_ind(pca, xlim = c(-6, 10)) + labs(title = "")

## ---
## Grafico ddos escores individuais com a qualidade dos individuos nas cores
## ---

fviz_pca_ind(pca, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE) + labs(title = "")
## ---
## Biplot
## ---

fviz_pca_biplot(pca, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969" # Individuals color
) + labs(title = "")


## ---
## Relacao do escores dos responentes vs escore do sitema
## ---



ggplot() +
```

```
  geom_point(aes(y = (-1)*pca$x[,1], x = dados$score)) +
  xlab("Escore") + ylab("Escore da Componete 1") + theme_minimal()



ggplot() +
  geom_point(aes(y = (-1)*pca$x[,2], x = dados$score)) +
  xlab("Escore") + ylab("Escore da Componete 2") + theme_minimal()

## ---
## correlacao entre as componentes
## ---

cor(pca$x) ## matriz de correlacao

corrplot(cor(pca$x) , method = 'color',type = 'lower',
         cl.pos = 'b', addCoef.col = 'lawngreen', xlab ='hhh')
```