Trabalho 3

Daniel Krügel

2023-06-10

Questão 1

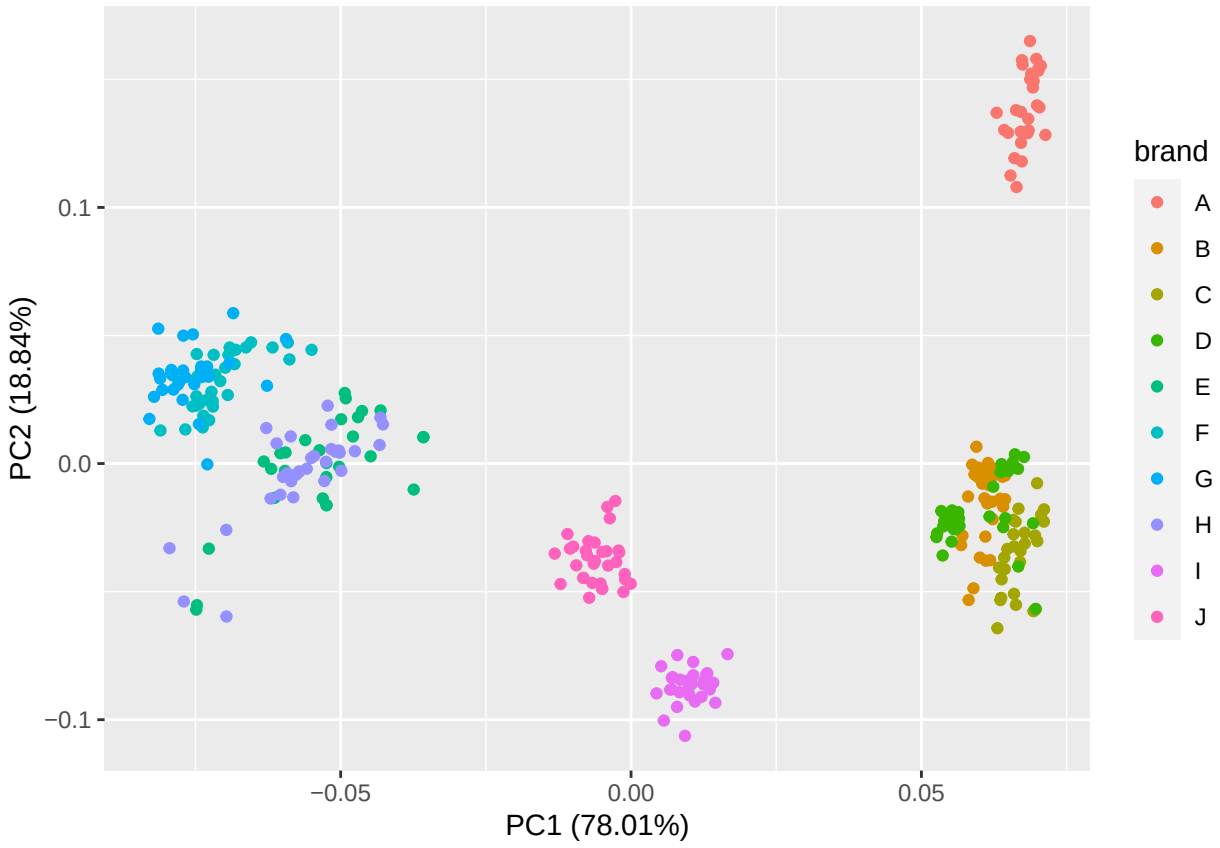
```
#Carregando os dados

pizza = read.csv(
  file = "http://leg.ufpr.br/~lucambio/CE090/20231S/pizza.csv",
  header = TRUE, sep = ",")
```

```
resultado <- prcomp(pizza[,-c(1,2)])
# calcula as Componentes Principais
summary(resultado)
```

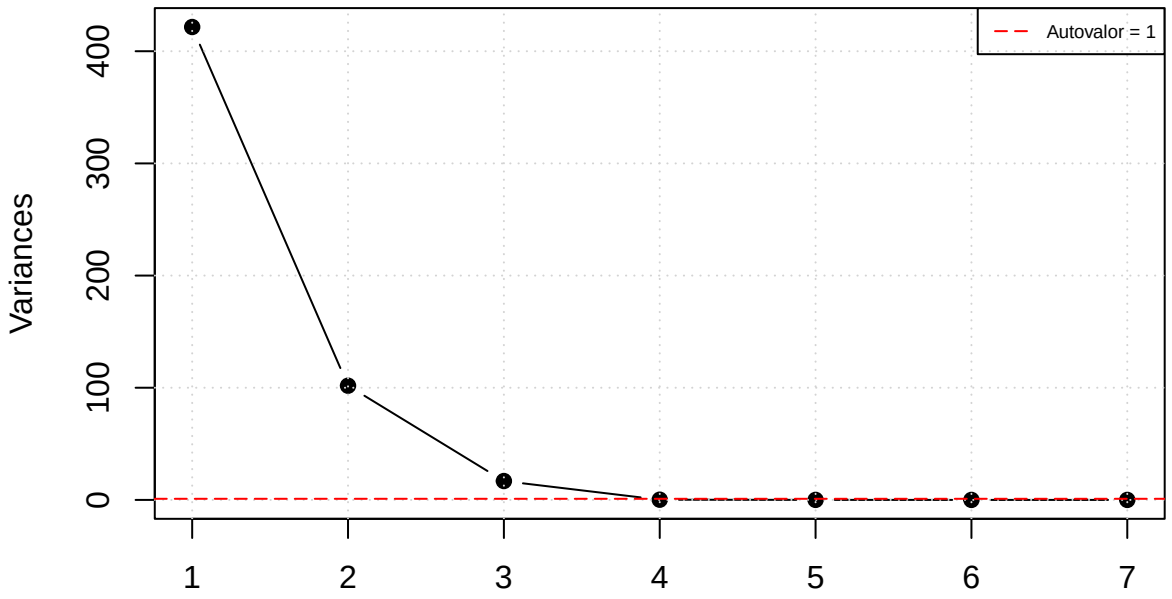
```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     20.5326 10.0906 4.10186 0.42945 0.09151 0.03221 0.01562
## Proportion of Variance  0.7801  0.1884 0.03113 0.00034 0.00002 0.00000 0.00000
## Cumulative Proportion   0.7801  0.9685 0.99964 0.99998 1.00000 1.00000 1.00000
```

Incluiria até a terceira componente ja que ela explica em torno de 0.99 da variância dos dados

```
pca.plot <- autoplot(resultado, data = pizza, colour = 'brand')
pca.plot
```



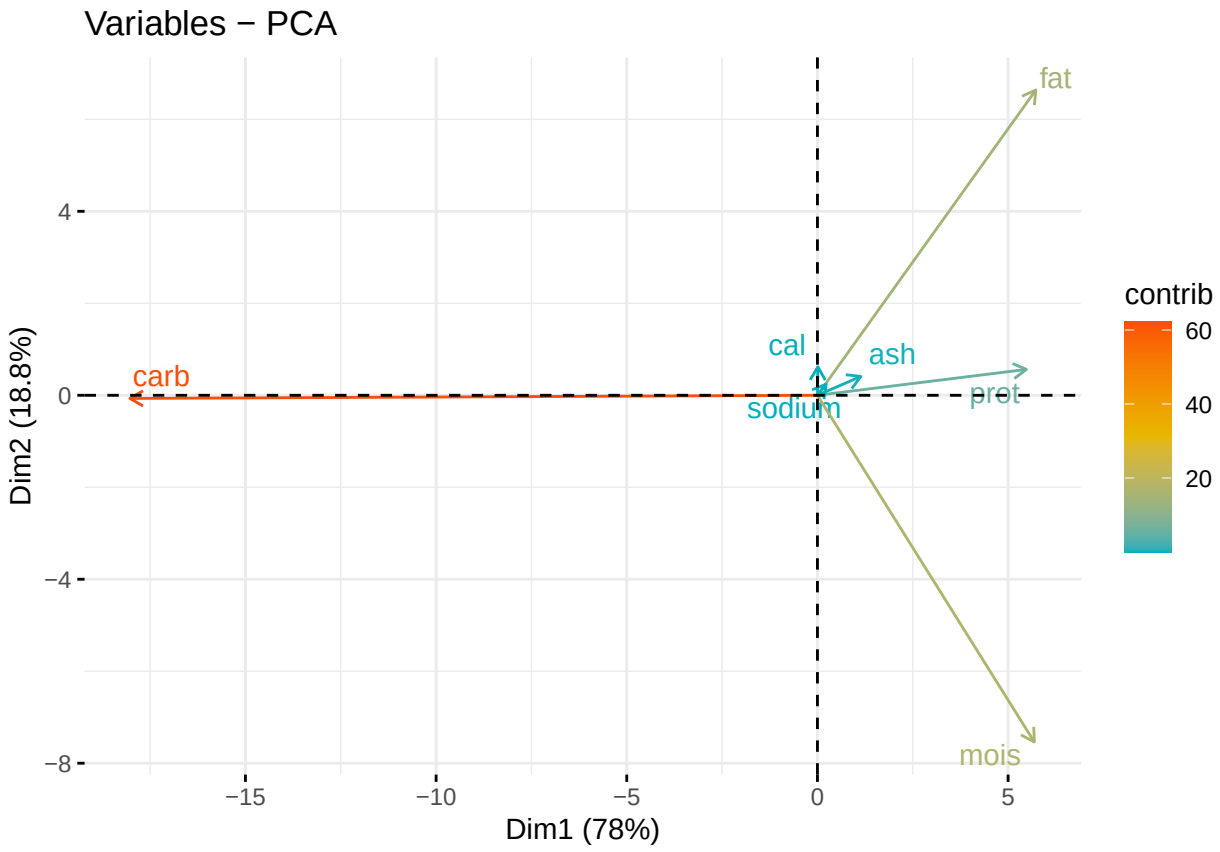

```
screeplot(resultado, type = "l", npcs = 7, main = "Gráfico das 7 PCs", pch = 19)
box()
grid()
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Autovalor = 1"), col=c("red"), lty=5, cex=0.6)
```



O screenplot confirma a minha teoria de que a inclusão dos 3 PC é mais do que o suficiente.

b)

```
factoextra::fviz_pca_var(resultado,
                         col.var = "contrib", # Cor por contribuições para o PC
                         gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
                         repel = TRUE     #Evite sobreposição de texto
)
```




```
resultado$rotation[,1:3]
```

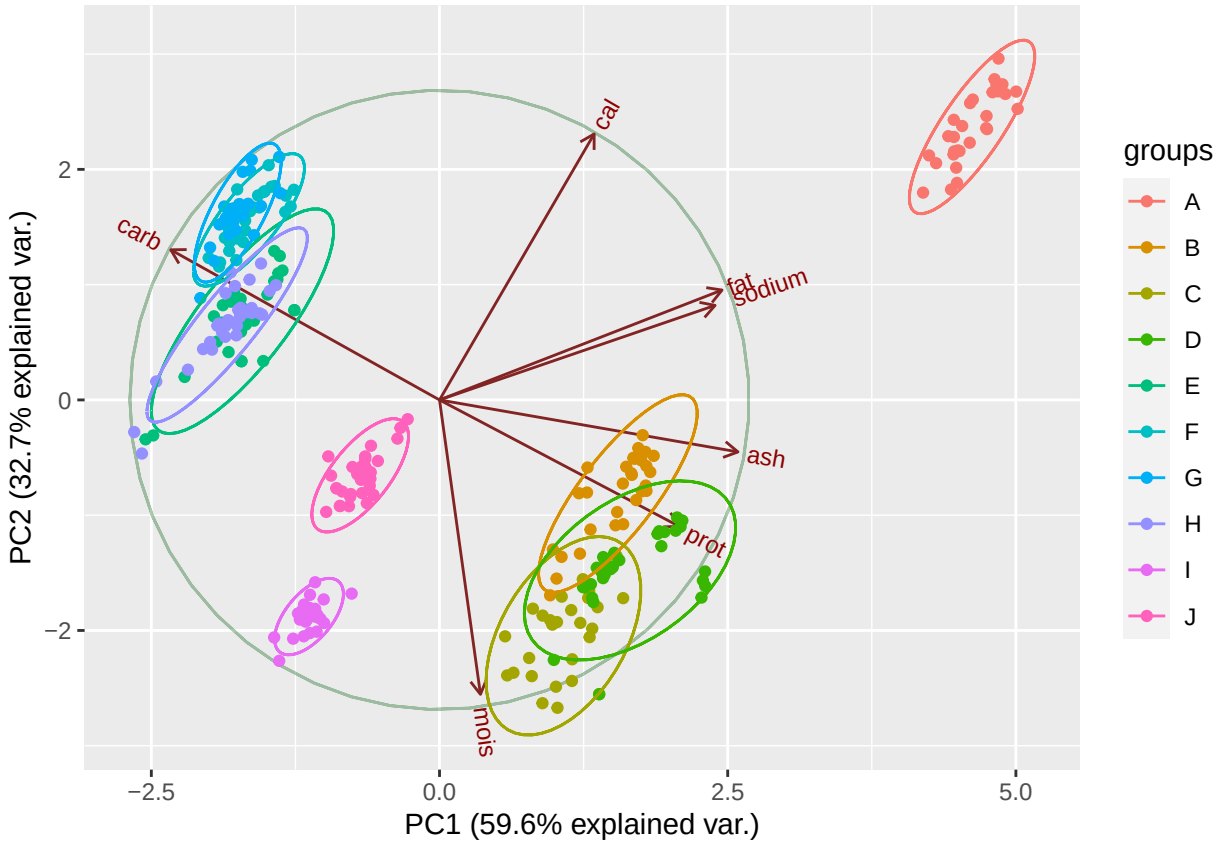
```
##                 PC1          PC2         PC3
## mois    0.2769634265 -0.747073681  0.35201618
## prot    0.2669414573  0.055732948 -0.80971797
## fat     0.2789335590  0.657845306  0.46797644
## ash     0.0554340960  0.040604210 -0.02222528
## sodium  0.0111416057  0.023813760  0.02624469
## carb   -0.8780843639 -0.006817551  0.01246929
## cal     0.0006032876  0.061253828  0.01006227
```

O PC1 representa pizzas pobres em carboidratos e rica em umidade, proteínas e gorduras. Enquanto o PC2 representa pizzas secas mas ricas em gordura. Já o PC3 representa pizzas com poucas proteínas mas gordurosas e úmidas.

c)

```
resultado_normalizado <- prcomp(pizza[,-c(1,2)], scale. = T, center = T) # calcula as Componentes Princ

ggbiplot::ggbiplot(resultado_normalizado,
         obs.scale = 1,
         var.scale = 1,
         groups = as.factor(pizza$brand),
         ellipse = TRUE,
```

```
        circle = TRUE,
        ellipse.prob = 0.95)
```



d)

Nível de crocância não é uma das variáveis avaliadas neste estudo, estarei avaliando a crocância com base na quantidade de água (umidade)

Gordurosa e crocante: Marca A Macia e pouco gordurosa: Marca I e J Equilibrada: Marca B e J

Questão 2

a)

```
exemplo = read.csv(
  file = "http://leg.ufpr.br/~lucambio/CE090/20231S/NTRforW.csv",
  sep = ",")

# transformando os dados

exemplo[,6:8] <- exemplo[,6:8] * 60
```

b)

```
Fatorial2 <- factanal(exemplo[,2:8], factors = 2)
Fatorial2
```

```
##
## Call:
## factanal(x = exemplo[, 2:8], factors = 2)
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6    X7
## 0.092 0.131 0.346 0.057 0.514 0.335 0.332
##
## Loadings:
##    Factor1 Factor2
## X1 0.442   0.844
## X2 0.438   0.823
## X3 0.622   0.517
## X4 0.885   0.399
## X5 0.585   0.379
## X6 0.714   0.395
## X7 0.680   0.454
##
##                Factor1 Factor2
## SS loadings      2.873   2.321
## Proportion Var   0.410   0.332
## Cumulative Var   0.410   0.742
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 6.25 on 8 degrees of freedom.
## The p-value is 0.62
```

```
Fatorial3 <- factanal(exemplo[,2:8], factors = 3)
Fatorial3
```

```
##
## Call:
## factanal(x = exemplo[, 2:8], factors = 3)
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6    X7
## 0.092 0.131 0.211 0.080 0.465 0.315 0.242
##
## Loadings:
##    Factor1 Factor2 Factor3
## X1 0.420   0.796   0.311
## X2 0.392   0.775   0.340
## X3 0.349   0.438   0.690
## X4 0.714   0.320   0.555
## X5 0.624   0.321   0.207
## X6 0.667   0.317   0.373
## X7 0.749   0.386   0.221
```

```
##
##                Factor1 Factor2 Factor3
## SS loadings      2.357   1.880   1.228
## Proportion Var   0.337   0.269   0.175
## Cumulative Var   0.337   0.605   0.781
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 0.09 on 3 degrees of freedom.
## The p-value is 0.993
```

Inicialmente, temos dois modelos interessantes para análise: um com dois fatores e outro com três fatores. O modelo com dois fatores apresenta um valor de p de 0.62, indicando que não há evidências suficientes para rejeitar a hipótese nula de que dois fatores são adequados. Por outro lado, o modelo com três fatores possui um carregamento SS loading maior que 1, o que, de acordo com a regra de Kaiser, sugere que vale a pena mantê-lo.

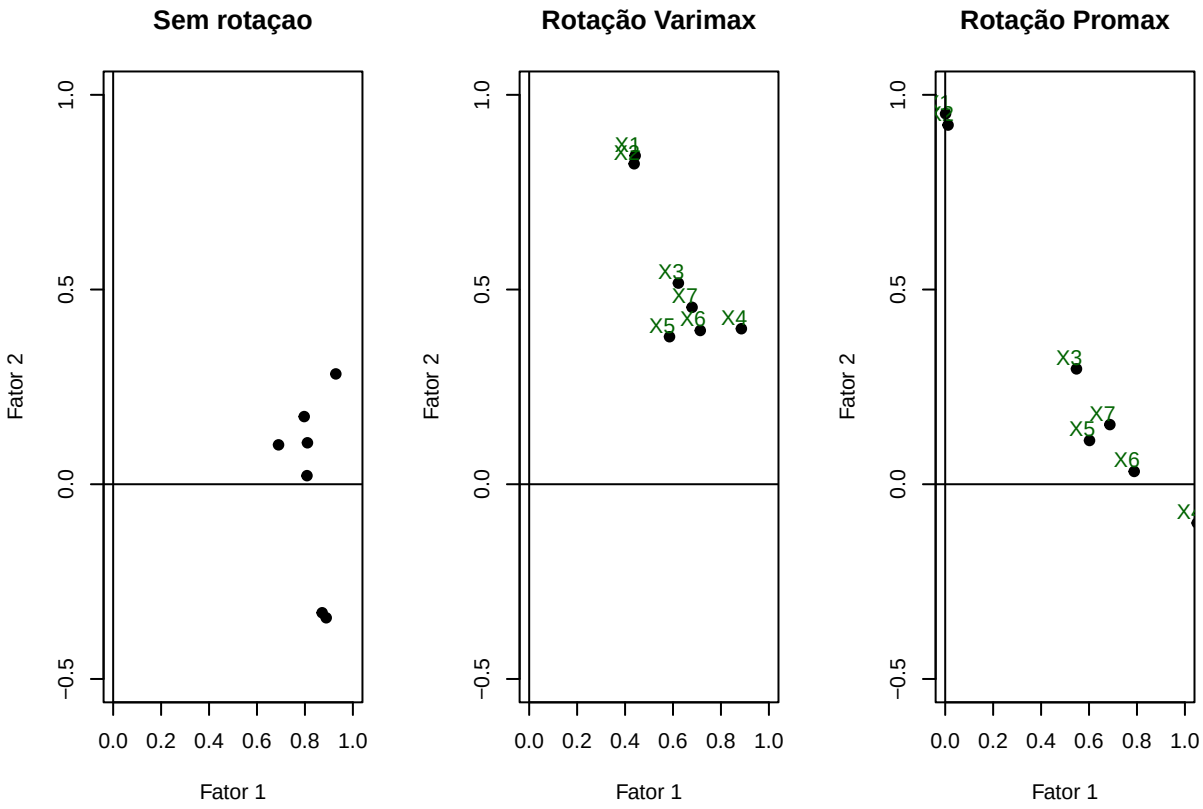
Considerando esses resultados, a escolha do modelo dependerá da análise realizada. Nesse caso, optaremos pelo modelo mais simples, com apenas dois fatores.

A interpretação dos fatores é a seguinte: o Fator 1 tem maior influência nas corridas de longa duração, enquanto o Fator 2 tem maior influência nas corridas de curta duração.

c)

```r
Fatorial2_none <- factanal(exemplo[,2:8], factors = 2, rotation = "none")
Fatorial2_varimax <- factanal(exemplo[,2:8], factors = 2, rotation = "varimax")
Fatorial2_promax <- factanal(exemplo[,2:8], factors = 2, rotation = "promax")

par(mfrow = c(1,3))
plot(Fatorial2_none$loadings[,1],
     Fatorial2_none$loadings[,2],
     xlab = "Fator 1",
     ylab = "Fator 2",
     ylim = c(-0.5,1),
     xlim = c(0,1),
     pch = 19,
     main = "Sem rotaçao")
abline(h = 0, v = 0)

plot(Fatorial2_varimax$loadings[,1],
     Fatorial2_varimax$loadings[,2],
     xlab = "Fator 1",
     ylab = "Fator 2",
     ylim = c(-0.5,1),
     xlim = c(0,1),
     pch = 19,
     main = "Rotação Varimax")

text(Fatorial2_varimax$loadings[,1]-0.03,
     Fatorial2_varimax$loadings[,2]+0.03,
     colnames(exemplo[,2:8]),
     col="darkgreen")
abline(h = 0, v = 0)
```

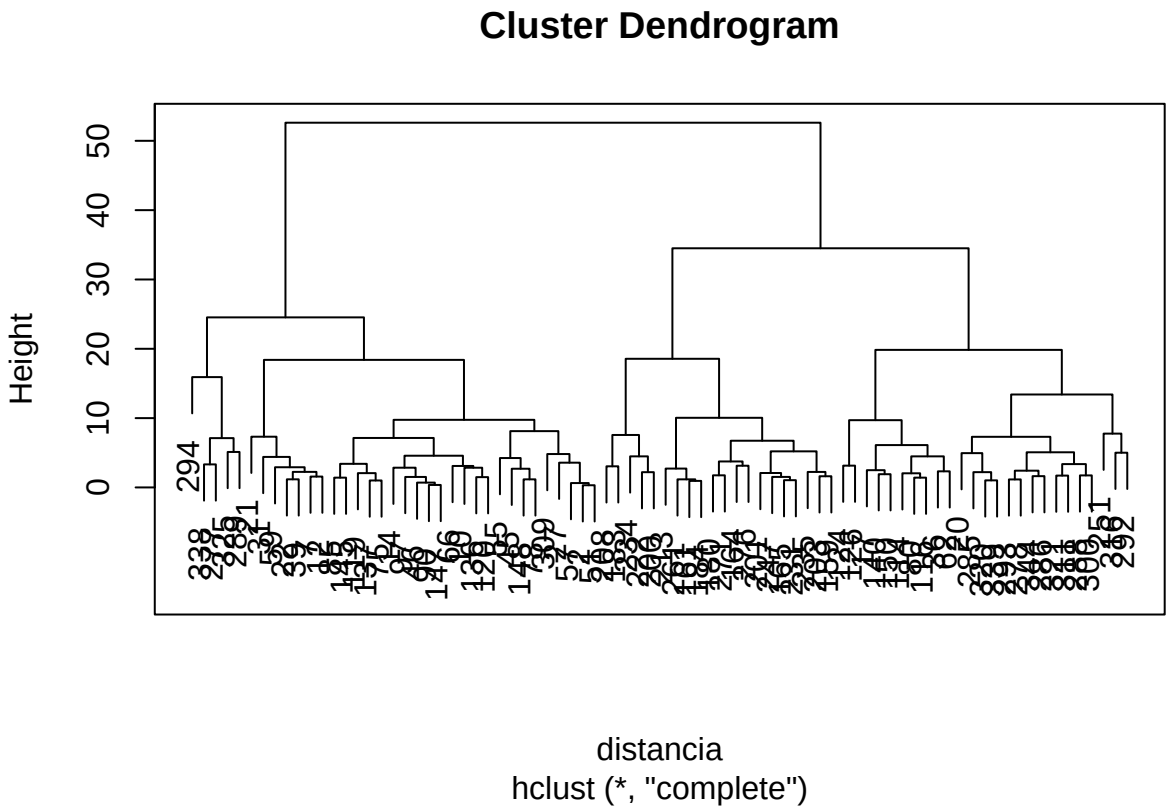
```
plot(Fatorial2_promax$loadings[,1],
     Fatorial2_promax$loadings[,2],
     xlab = "Fator 1",
     ylab = "Fator 2",
     ylim = c(-0.5,1),
     xlim = c(0,1),
     pch = 19,
     main = "Rotação Promax")
text(Fatorial2_promax$loadings[,1]-0.03,
     Fatorial2_promax$loadings[,2]+0.03,
     colnames(exemplo[,2:8]),
     col="darkgreen")
abline(h = 0, v = 0)
```



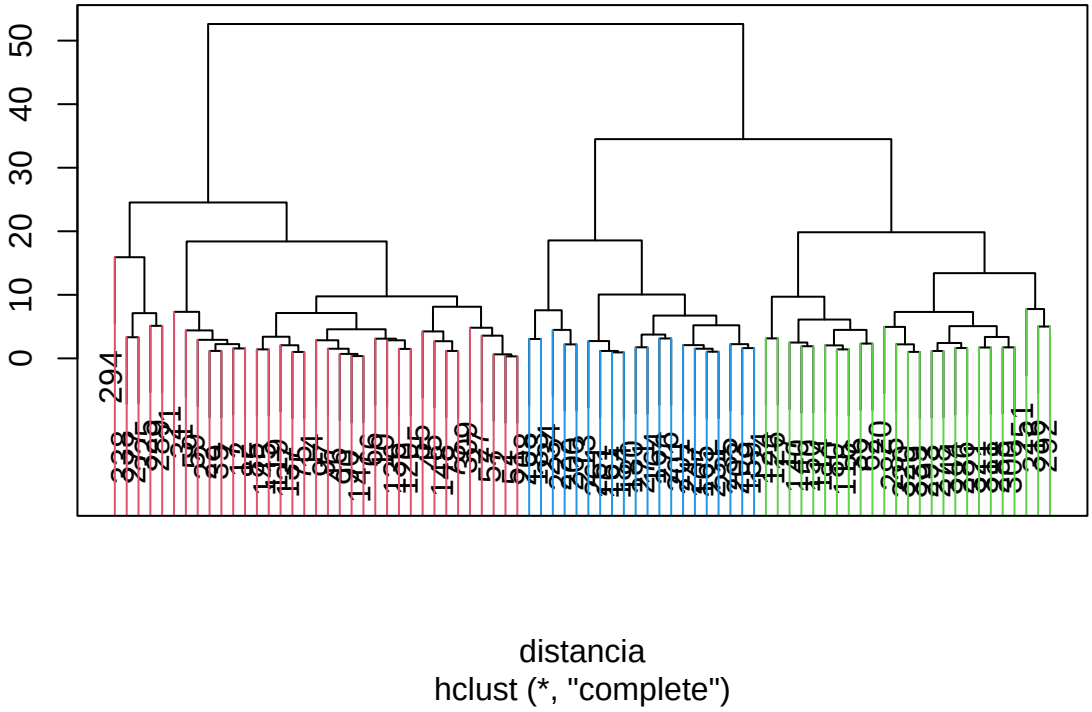
Questão 3

```
penguins = read.csv(
  file = "http://leg.ufpr.br/~lucambio/CE090/20231S/penguins.csv",
  header = TRUE, sep = ",")

bicos <- penguins[,4:6]

# Amostragem
set.seed(2203)
```

```r
amostra <- penguins[sample(nrow(penguins), 80, replace=FALSE),]
distancia <- dist(bicos[amostra$X,], method = "euclidean")

# Criação dos cluster
hbloco  <- hclust(distancia)
plot(hbloco)
box()
```




```r
# Colorindo os blocos
y = cutree(hbloco, 3)
ColorDendrogram(hbloco, y = y, labels = names(y), main = "Espécies de pinguins",
                branchlength = 80)
box()
```



Colorindo os blocos com o número de espécies presentes no estudo facilita a visualização dos clusters através da árvore, então sim, é possível separar as espécies utilizando somente informação dos bicos