TarefaDois

Mauricio Zalamena Bavaresco

2022-09-29

1.Carregue a base de dados e mostre a estrutura do dataset (str()). O arquivo do dataset não pode ser modificado de forma alguma. A leitura deverá tratar qualquer característica do arquivo.

```
setwd("C:\\Users\\Mauricio\\Desktop\\Material\\Atividades\\Computação aplicada")

dados = read.csv("Dry_Bean_Dataset.csv",sep=";",dec=",")
```

```
str(dados)
```

```
## 'data.frame':    13611 obs. of  17 variables:
##  $ Area           : int  28395 28734 29380 30008 30140 30279 30477 30519 30685 30834 ...
##  $ Perimeter      : num  610 638 624 646 620 ...
##  $ MajorAxisLength: num  208 201 213 211 202 ...
##  $ MinorAxisLength: num  174 183 176 183 190 ...
##  $ AspectRation   : num  1.2 1.1 1.21 1.15 1.06 ...
##  $ Eccentricity   : num  0.55 0.412 0.563 0.499 0.334 ...
##  $ ConvexArea     : int  28715 29172 29690 30724 30417 30600 30970 30847 31044 31120 ...
##  $ EquivDiameter  : num  190 191 193 195 196 ...
##  $ Extent         : num  0.764 0.784 0.778 0.783 0.773 ...
##  $ Solidity       : num  0.989 0.985 0.99 0.977 0.991 ...
##  $ roundness      : num  0.958 0.887 0.948 0.904 0.985 ...
##  $ Compactness    : num  0.913 0.954 0.909 0.928 0.971 ...
##  $ ShapeFactor1   : num  0.00733 0.00698 0.00724 0.00702 0.0067 ...
##  $ ShapeFactor2   : num  0.00315 0.00356 0.00305 0.00321 0.00366 ...
##  $ ShapeFactor3   : num  0.834 0.91 0.826 0.862 0.942 ...
##  $ ShapeFactor4   : num  0.999 0.998 0.999 0.994 0.999 ...
##  $ Class          : chr  "SEKER" "SEKER" "SEKER" "SEKER" ...
```

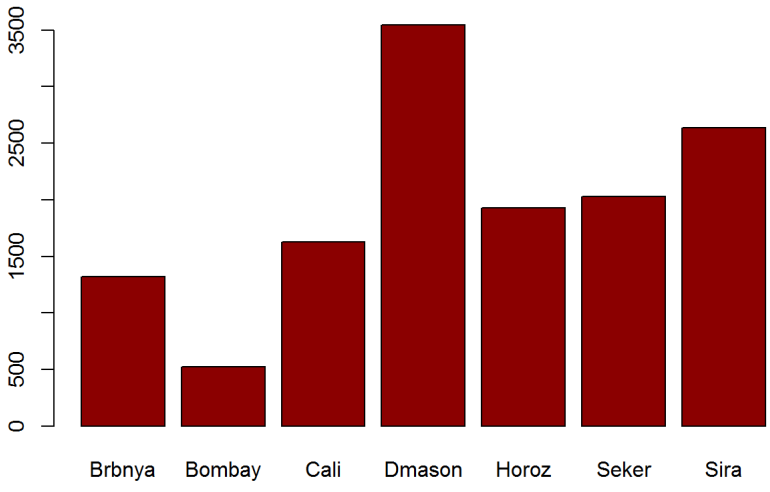
2.Altere a variável do tipo do feijão (Class) para um factor.

```
dados$Class = factor(c(dados$Class))
summary(dados$Class)
```

```
## BARBUNYA   BOMBAY     CALI DERMASON    HOROZ    SEKER     SIRA
##     1322      522     1630     3546     1928     2027     2636
```

3.Plote um gráfico de barras que ilustre as quantidades de cada classe.

```
barplot(summary(dados$Class),names.arg = c("Brbnya", "Bombay", "Cali", "Dmason", "Horoz", "Seker", "Sira"),col="darkred")
```



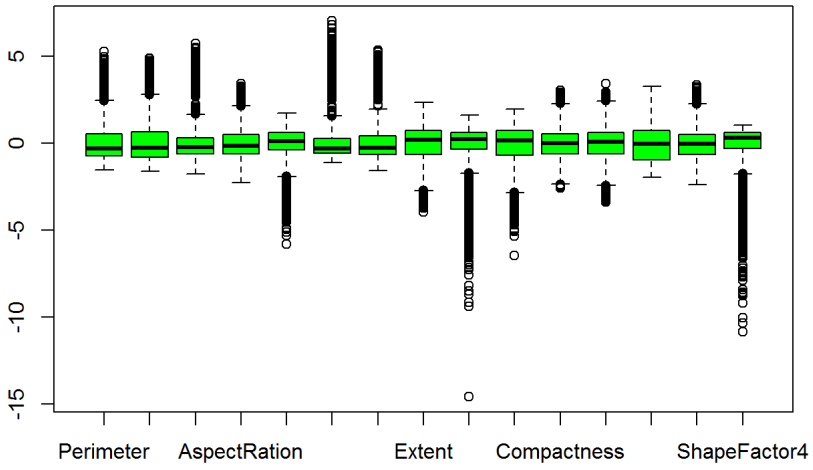
4. Realize a normalização dos dados via Z-score. Plote um boxplot para ilustrar a distribuição de cada variável. Mostre as estatísticas de cada variável (summary).

```
escorez=as.data.frame(lapply(dados[,2:16],function(y)(y-mean(y))/sd(y) ))
```

###ou

```
escorez=as.data.frame(scale(dados[,2:16]))
```

```
boxplot(escorez,col = "green")
```




```
summary(escorez)
```

```
##    Perimeter       MajorAxisLength   MinorAxisLength    AspectRation
##  Min.   :-1.5425   Min.   :-1.5933   Min.   :-1.7736   Min.   :-2.2636
##  1st Qu.:-0.7082   1st Qu.:-0.7800   1st Qu.:-0.5876   1st Qu.:-0.6119
##  Median :-0.2816   Median :-0.2714   Median :-0.2188   Median :-0.1302
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.5690   3rd Qu.: 0.6576   3rd Qu.: 0.3282   3rd Qu.: 0.5021
##  Max.   : 5.2736   Max.   : 4.8862   Max.   : 5.7355   Max.   : 3.4339
##   Eccentricity       ConvexArea      EquivDiameter        Extent
##  Min.   :-5.7819   Min.   :-1.1111   Min.   :-1.5516   Min.   :-3.9607
##  1st Qu.:-0.3801   1st Qu.:-0.5728   1st Qu.:-0.6421   1st Qu.:-0.6336
##  Median : 0.1472   Median :-0.2885   Median :-0.2472   Median : 0.2063
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6475   3rd Qu.: 0.2863   3rd Qu.: 0.4458   3rd Qu.: 0.7562
##  Max.   : 1.7448   Max.   : 7.0359   Max.   : 5.3451   Max.   : 2.3726
##     Solidity          roundness        Compactness       ShapeFactor1
##  Min.   :-14.5689   Min.   :-6.4460   Min.   :-2.5811   Min.   :-3.35603
##  1st Qu.: -0.3160   1st Qu.:-0.6920   1st Qu.:-0.6059   1st Qu.:-0.58838
##  Median :  0.2446   Median : 0.1659   Median : 0.0229   Median : 0.07231
##  Mean   :  0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.:  0.6159   3rd Qu.: 0.7323   3rd Qu.: 0.5575   3rd Qu.: 0.62749
##  Max.   :  1.6167   Max.   : 1.9725   Max.   : 3.0373   Max.   : 3.44642
##   ShapeFactor2       ShapeFactor3       ShapeFactor4
##  Min.   :-1.93292   Min.   :-2.35617   Min.   :-10.8500
##  1st Qu.:-0.94387   1st Qu.:-0.62863   1st Qu.: -0.3116
##  Median :-0.03762   Median :-0.01562   Median :  0.3029
##  Mean   : 0.00000   Mean   : 0.00000   Mean   :  0.0000
##  3rd Qu.: 0.76244   3rd Qu.: 0.52948   3rd Qu.:  0.6457
##  Max.   : 3.27086   Max.   : 3.34535   Max.   :  1.0693
```

5.Realize a seleção de características (correlação). Plote o gráfico de correlação. Liste as características que foram removidas.

```
library(corrplot)
```

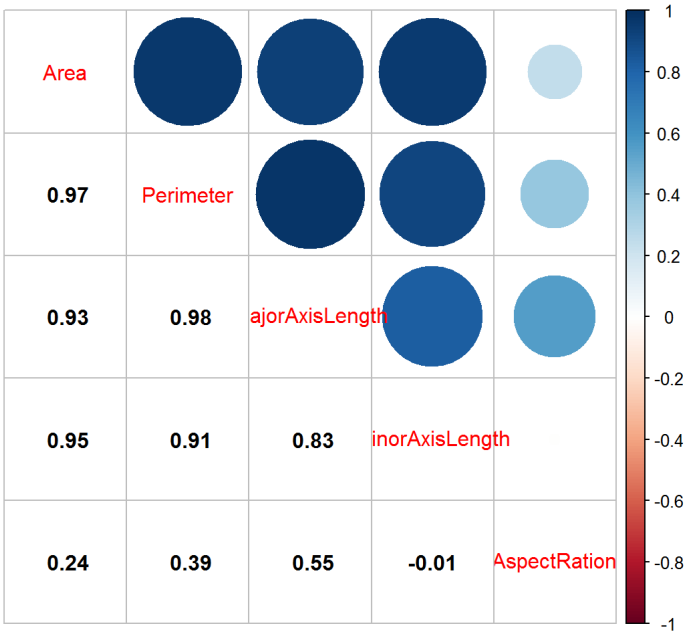
```
## corrplot 0.92 loaded
```

```
library(caret)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
## Carregando pacotes exigidos: lattice
```

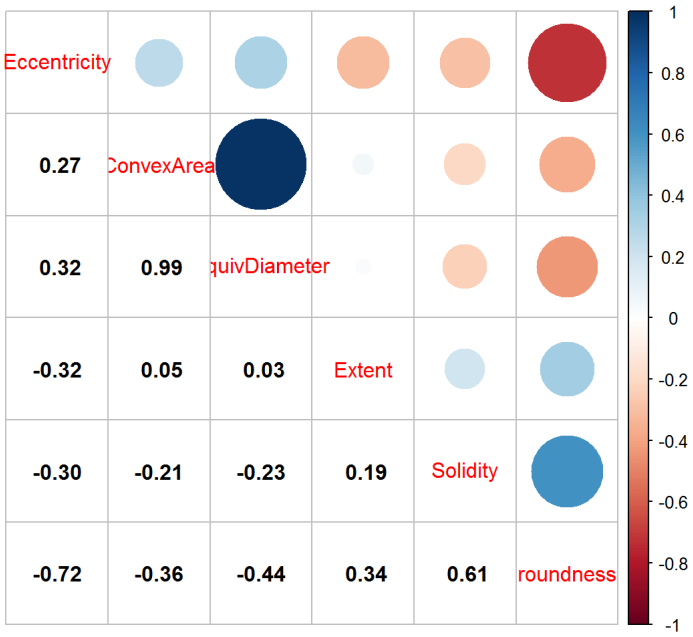
```
dadosCorrelacao = cor(dados[,1:5])
corrplot.mixed(dadosCorrelacao,lower.col = "black")
```




```
correlacaoAlta = findCorrelation(dadosCorrelacao, cutoff=0.95)
print(correlacaoAlta)
```

```
## [1] 3 2 1
```

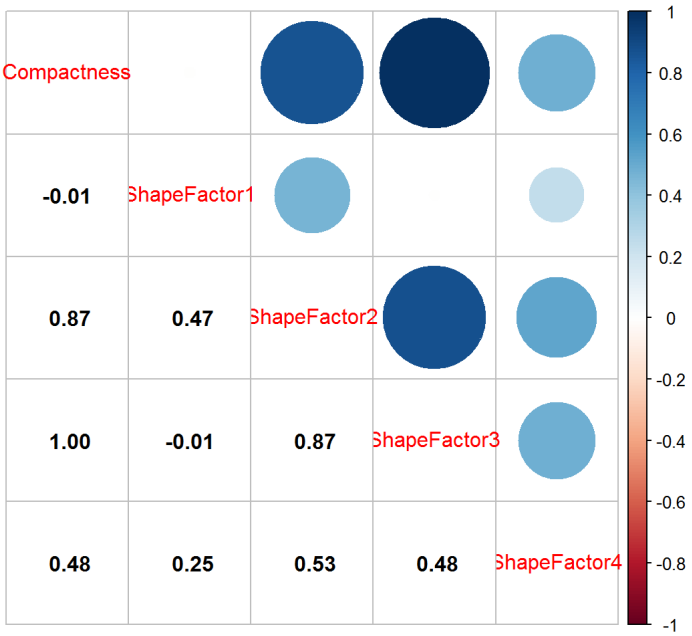
```
dadosCorrelacao = cor(dados[,6:11])
corrplot.mixed(dadosCorrelacao,lower.col = "black")
```




```
correlacaoAlta = findCorrelation(dadosCorrelacao, cutoff=0.95)
print(correlacaoAlta)
```

```
## [1] 3
```

```
dadosCorrelacao = cor(dados[,12:16])
corrplot.mixed(dadosCorrelacao,lower.col = "black")
```

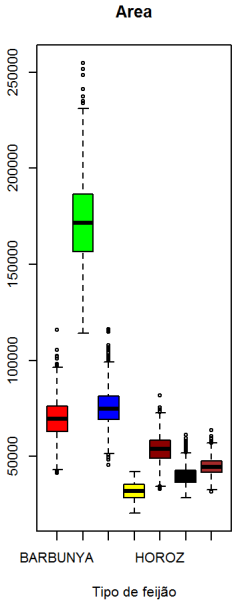


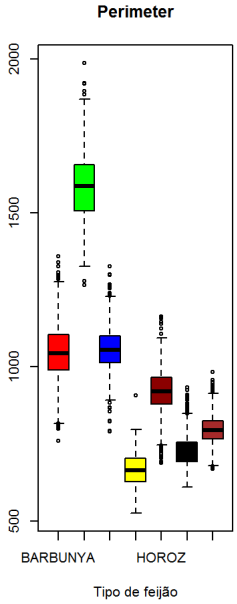

```
correlacaoAlta = findCorrelation(dadosCorrelacao, cutoff=0.95)
print(correlacaoAlta)
```

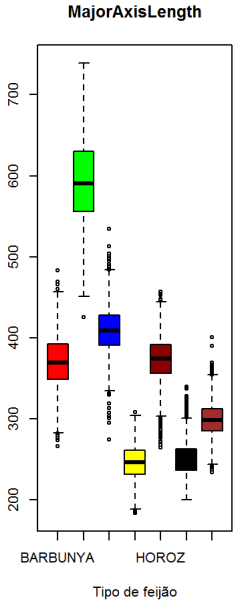
```
## [1] 4
```

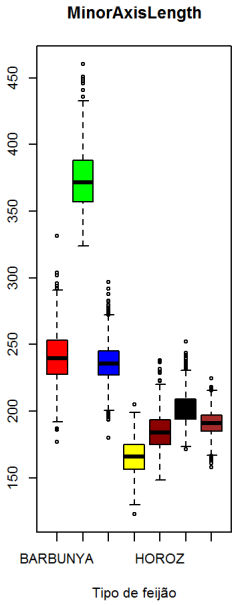
6.Plote um gráfico boxplot ou de densidade por variável x classe (organize em 3 colunas). Discuta qual é a variável que teria maior poder de discriminação? Existe alguma classe que pode ser classificada mais facilmente? Justifique a sua escolha.

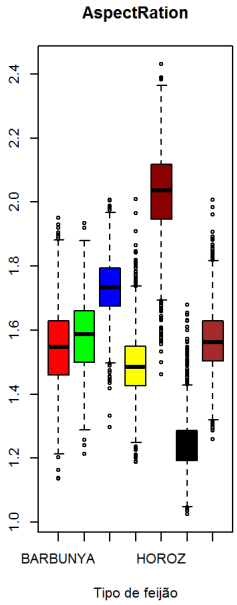
```
cores = c("red","green","blue","yellow","darkred","black","brown")
par(mfrow=c(1,3))
for (i in 1:16) {
boxplot(dados[,i] ~ dados$Class, col=cores, xlab="Tipo de feijão",
ylab="", main=names(dados)[i])
}
```

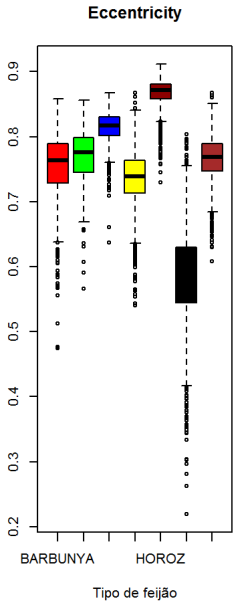

Area
250000
200000
150000
100000
50000
BARBUNYA
HOROZ
Tipo de feijão

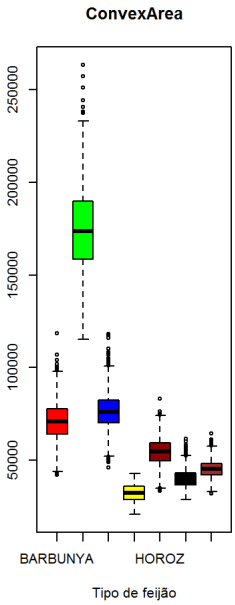

Perimeter
2000
1500
1000
500
BARBUNYA
HOROZ
Tipo de feijão

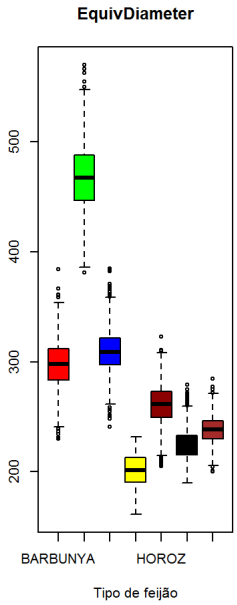

MajorAxisLength
700
600
500
400
300
200
BARBUNYA
HOROZ
Tipo de feijão

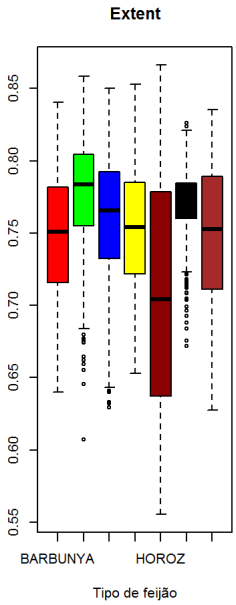

MinorAxisLength
450
400
350
300
250
200
150
BARBUNYA
HOROZ
Tipo de feijão

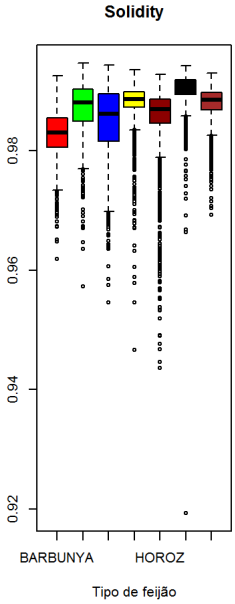

AspectRation
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
BARBUNYA
HOROZ
Tipo de feijão

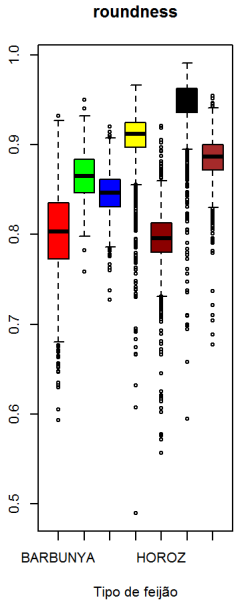

Eccentricity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
BARBUNYA
HOROZ
Tipo de feijão

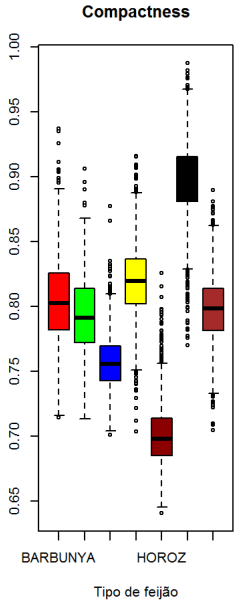

ConvexArea
250000
200000
150000
100000
50000
BARBUNYA
HOROZ
Tipo de feijão

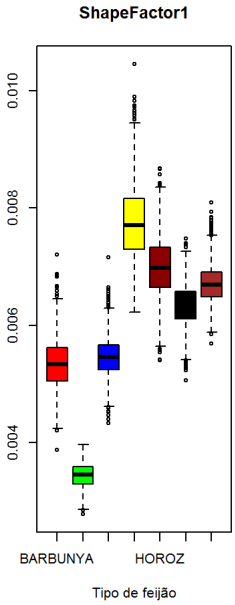

EquivDiameter
500
400
300
200
BARBUNYA
HOROZ
Tipo de feijão

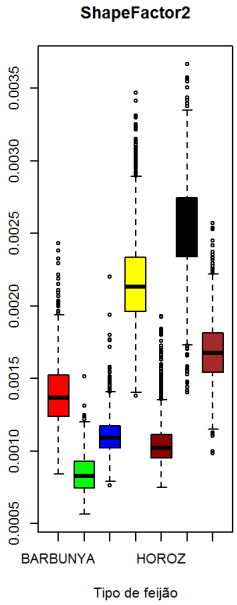

Extent
0.85
0.80
0.75
0.70
0.65
0.60
0.55
BARBUNYA
HOROZ
Tipo de feijão

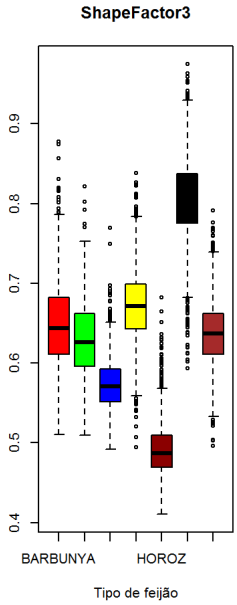


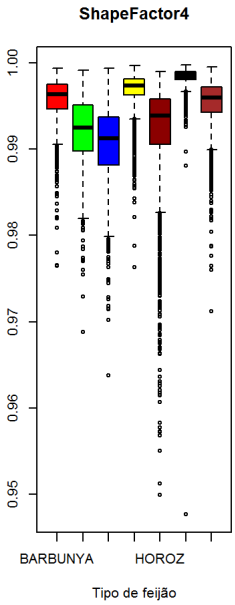












Na variável Área,Perimeter,MajorAxisLength,MinorAxisLength,ConvexArea,EquivDiameter,ShapeFactor1, pode-se visualizar um maior poder de discriminação na classe BOMBAY porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

Na variável AspectRation possuí duas classes com grande poder de discriminação que é a SEKER e HOROZ porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

Na variável Eccentricity a classe que possuí maior poder de discriminação é SEKER porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

Na variável Extent a classe horoz possuí maior poder de discriminação porque a mediana da classe horoz esta fora da caixa das outras

Na variável Solidity a classe SEKER possuí maior poder de discriminação porque a mediana está fora das outras caixas

Na variável roundness a classe SEKER possuí maior poder de discriminação porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

Na variável Compactness as classes Seker e Horoz possuem maior poder de discriminação porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

Na variavél ShapeFactor3 as classes Seker e Horoz possuem maior poder de discriminação porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

Na variável ShapeFactor4 a classe Seker possui o maior poder de discriminação porque a mediana da classe está fora das outras caixas e não sobrepõe as outras caixas.

```
print(levels(dados$Class))
```

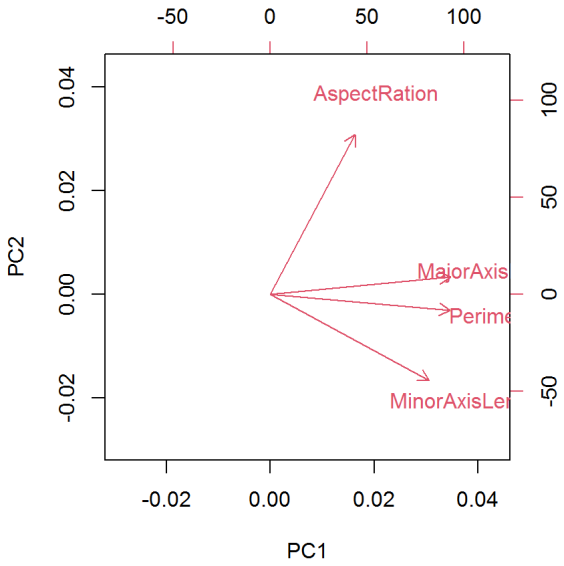
```
## [1] "BARBUNYA" "BOMBAY"   "CALI"     "DERMASON" "HOROZ"    "SEKER"    "SIRA"
```

7.Realize a projeção do dataset utilizando PCA. Explique as características dos componentes principais estimados. O que se pode explicar sobre os componentes principais utilizando o gráfico biplot. Apresente as características básicas (summary) dos dados.

```
pca = prcomp(dados[,2:5], center=TRUE, scale=TRUE)
print(pca)
```

```
## Standard deviations (1, .., p=4):
## [1] 1.72209904 1.01243835 0.07861298 0.05624484
##
## Rotation (n x k) = (4 x 4):
##                        PC1         PC2        PC3         PC4
## Perimeter       0.5770908 -0.08987168  0.8107934  0.03877259
## MajorAxisLength 0.5772849  0.09472639 -0.3657732 -0.72386392
## MinorAxisLength 0.5094353 -0.47175145 -0.4420447  0.56791081
## AspectRation    0.2723673  0.87200949 -0.1158468  0.38986539
```

```
biplot(pca,xlabs = rep("", nrow(dados[,1:16])))
```




```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7221 1.0124 0.07861 0.05624
## Proportion of Variance 0.7414 0.2563 0.00155 0.00079
## Cumulative Proportion  0.7414 0.9977 0.99921 1.00000
```

O Gráfico Biplot é um tipo de gráfico exploratório usado em estatística. As variáveis que estão exibidas no gráfico são as variáveis que são linearmente correlacionadas. E os componentes visualizados pelo summary explicam a variância dos dados em relação ao a cada autovetor.

Analisando os autovetores (biplot), pode-se verificar que:

As variáveis *AspectRation*, *MajorAxisLength*, *Perimeter* e *MinorAxisLength* são as que influenciam mais no componente principal 1. Podem-se dizer que as maiores medidas permitem discriminar melhor as classes.

As variáveis *MajorAxisLength* e *Perimeter* são altamente correlacionadas, pois o ângulo entre elas é muito pequeno.

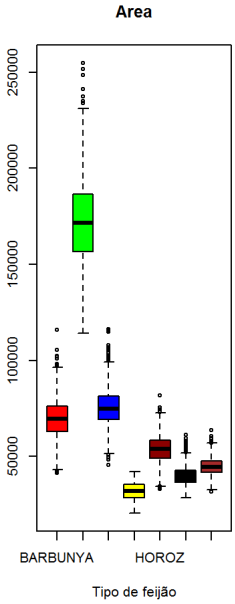
As variáveis *AspectRation* são as que influenciam mais no componente principal 2.

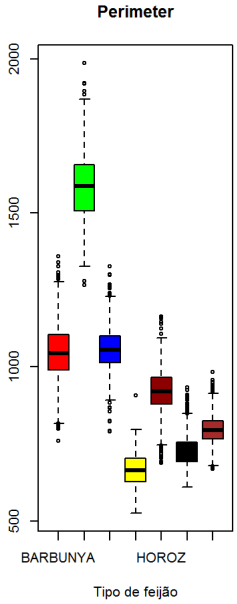
As variáveis *MajorAxisLength*, *Perimeter* e *MinorAxisLength* não são correlacionadas com *AspectRation* porque apresentam um ângulo próximo a 90°.

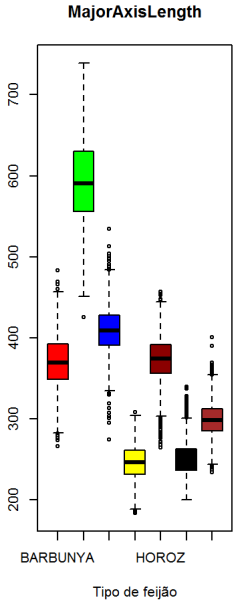
8.Analise o dataset projetado com o auxílio do gráfico de boxplot por classe (igual ao do item 6). Compare com o resultado do item 6. Se quiser, pode gerar um gráfico de espalhamento para auxiliar na explicação.

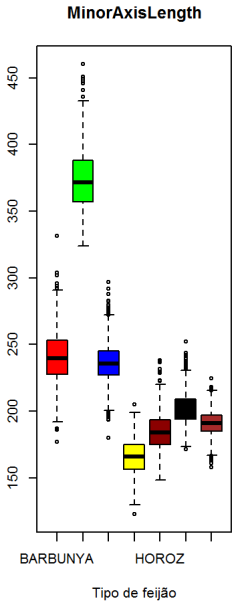
Pode-se observar através do gráfico de espalhamento a correlação entre as variaveis e classes e no item 6 o poder de discriminação.

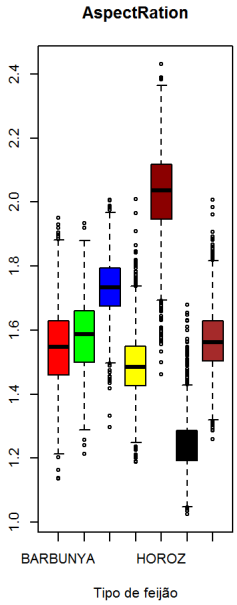
```
cores = c("red","green","blue","yellow","darkred","black","brown")
par(mfrow=c(1,3))
for (i in 1:16) {
boxplot(dados[,i] ~ dados$Class, col=cores, xlab="Tipo de feijão",
ylab="", main=names(dados)[i])
}
```

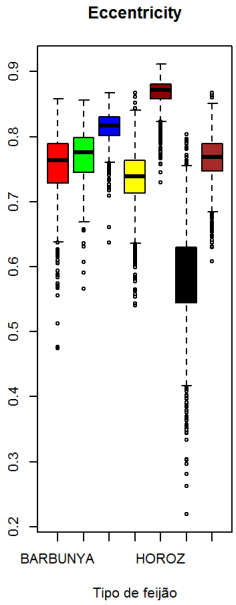

Area
250000
200000
150000
100000
50000
BARBUNYA
HOROZ
Tipo de feijão

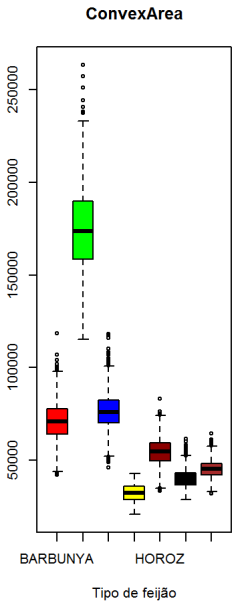

Perimeter
2000
1500
1000
500
BARBUNYA
HOROZ
Tipo de feijão

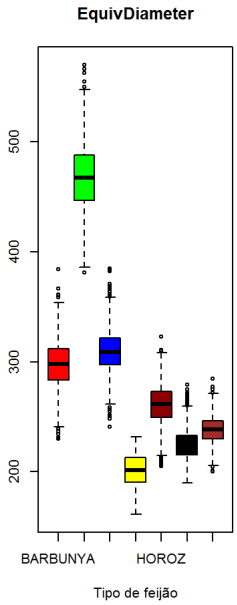

MajorAxisLength
700
600
500
400
300
200
BARBUNYA
HOROZ
Tipo de feijão

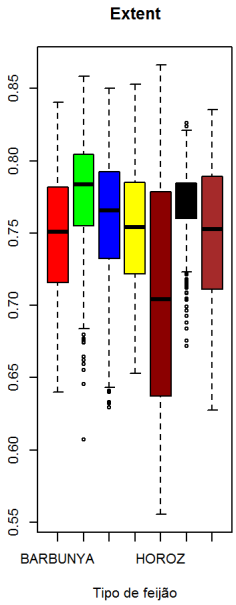

MinorAxisLength
450
400
350
300
250
200
150
BARBUNYA
HOROZ
Tipo de feijão

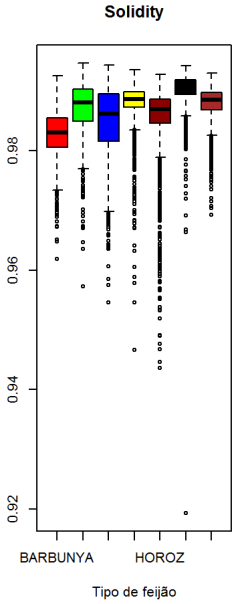

AspectRation
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
BARBUNYA
HOROZ
Tipo de feijão

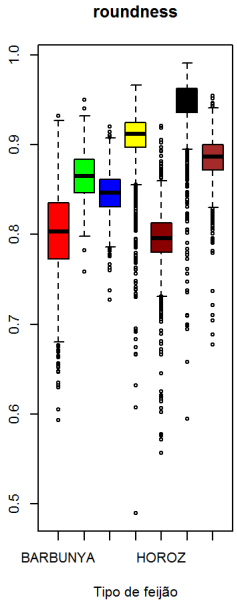

Eccentricity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
BARBUNYA
HOROZ
Tipo de feijão

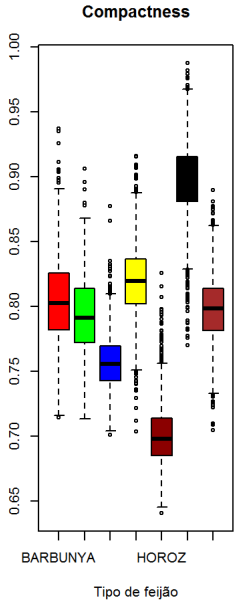

ConvexArea
250000
200000
150000
100000
50000
BARBUNYA
HOROZ
Tipo de feijão

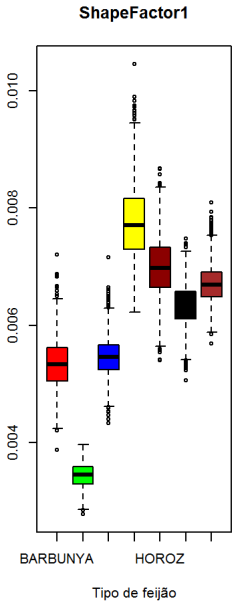

EquivDiameter
500
400
300
200
BARBUNYA
HOROZ
Tipo de feijão

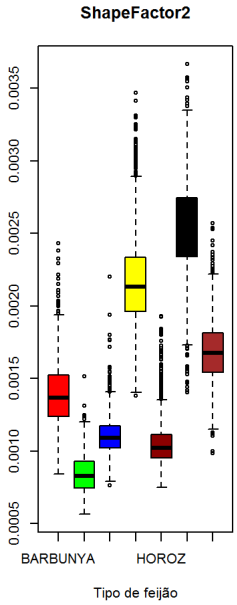

Extent
0.85
0.80
0.75
0.70
0.65
0.60
0.55
BARBUNYA
HOROZ
Tipo de feijão

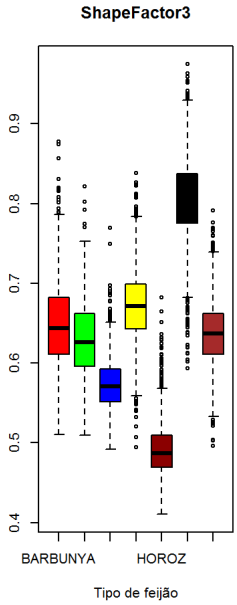


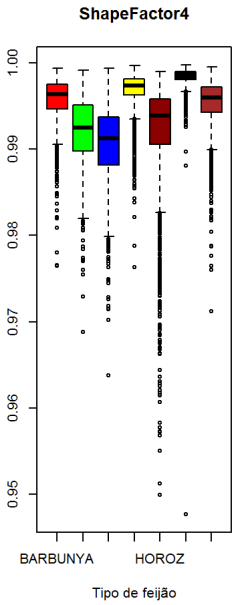








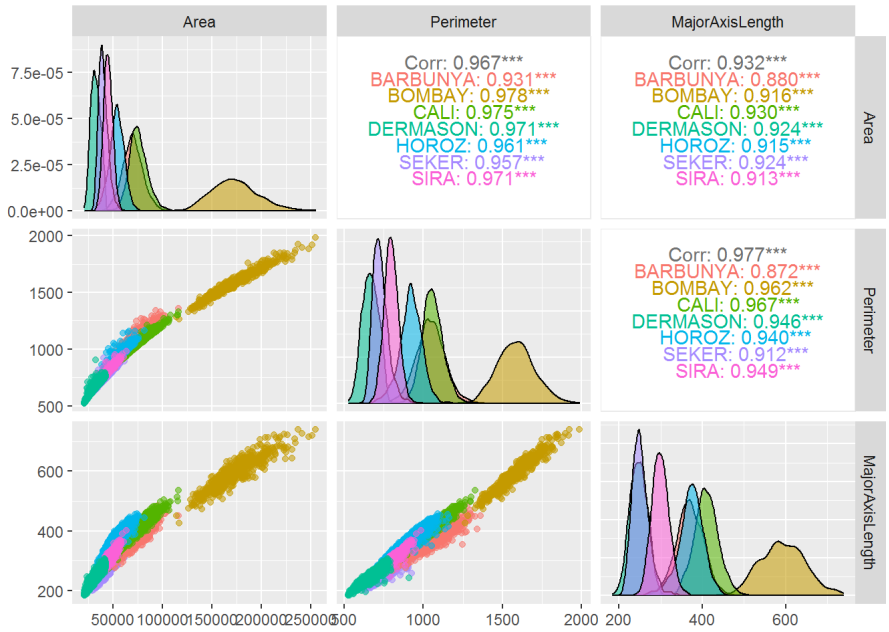




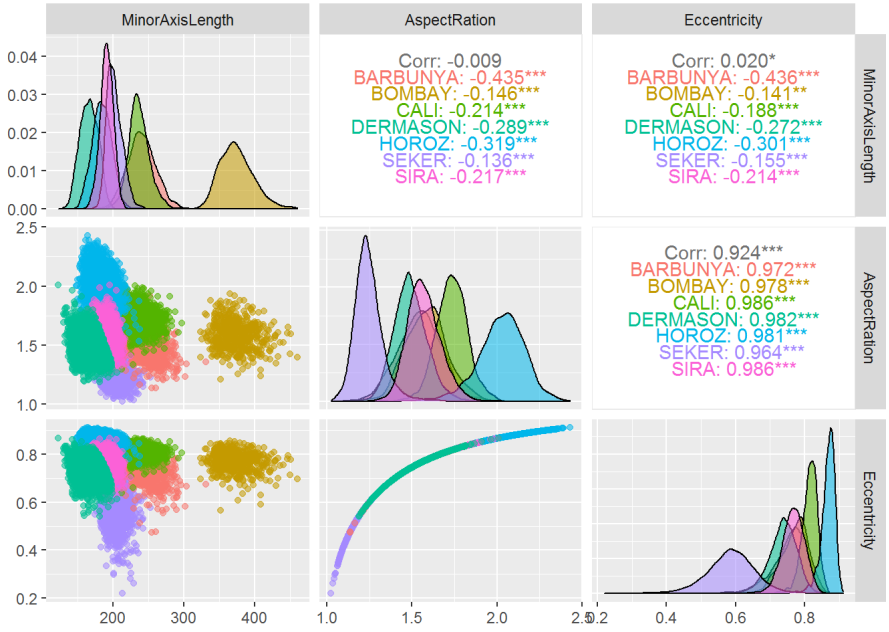

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

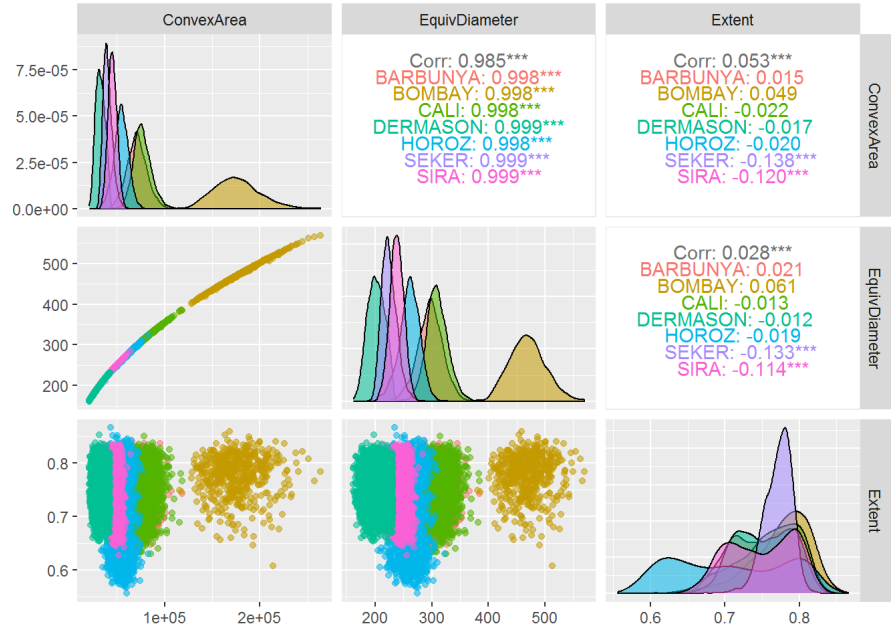
```
ggpairs(dados[,1:3],aes(colour=dados$Class,alpha=0.4))
```



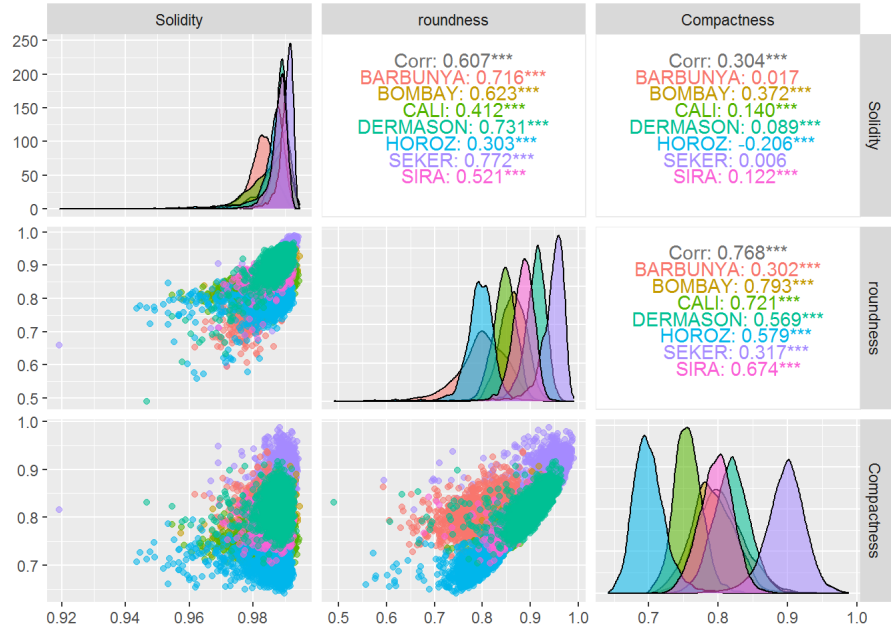

```
ggpairs(dados[,4:6],aes(colour=dados$Class,alpha=0.4))
```



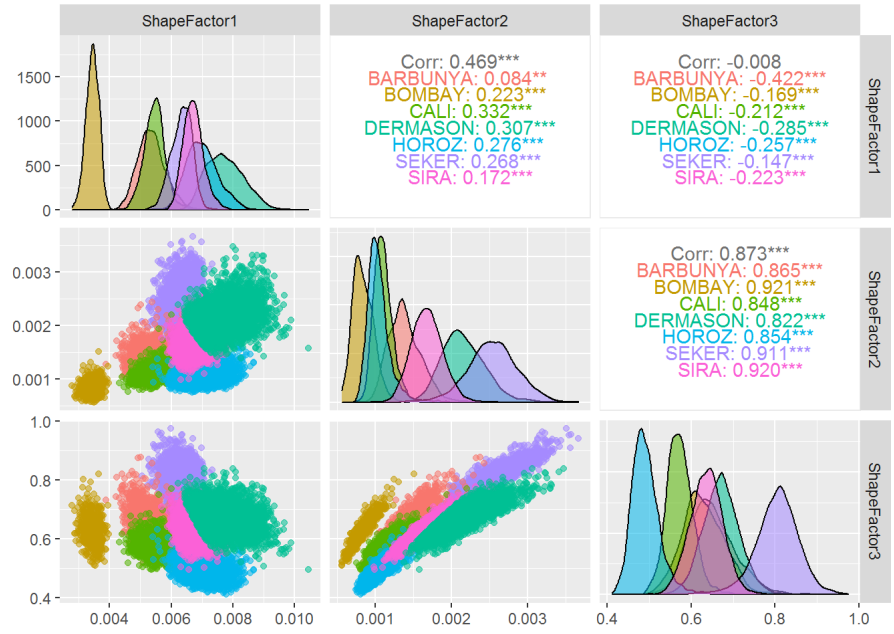

```
ggpairs(dados[,7:9],aes(colour=dados$Class,alpha=0.4))
```




```
ggpairs(dados[,10:12],aes(colour=dados$Class,alpha=0.4))
```

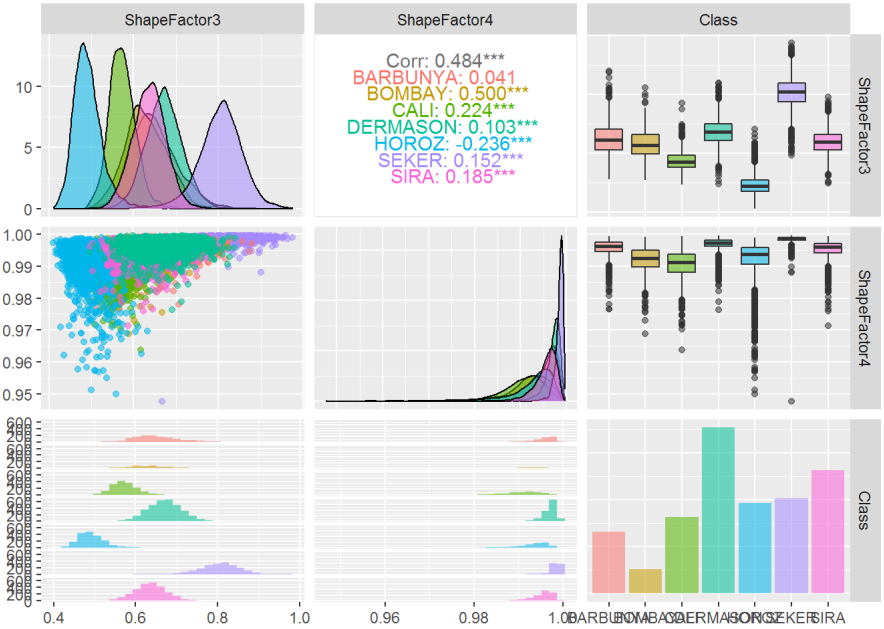



```
ggpairs(dados[,13:15],aes(colour=dados$Class,alpha=0.4))
```



```
ggpairs(dados[,15:17],aes(colour=dados$Class,alpha=0.4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



9.É possível reduzir a dimensionalidade dos dados? Explique como!

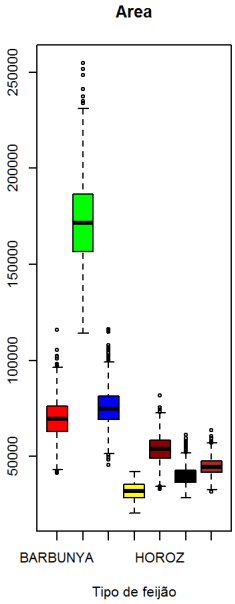
Sim é possível reduzir a dimensionalidade dos dados. A redução pode ser obtida por meio da remoção de informações irrelevantes/redundantes ou uma representação compacta e informativa dos dados originais. O mapeamento das entradas em um espaço original de d dimensões é realizado para um novo espaço com dimensões k (onde k <d), com uma perda mínima de informações.

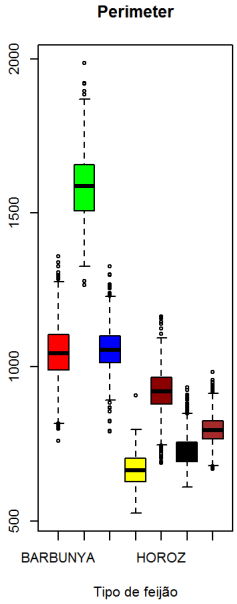
10.Analise o dataset reduzido com o auxílio do gráfico de boxplot por classe (igual ao do item 6). Compare com o resultado do item 6 e do item 8. Se quiser, pode gerar um gráfico de espalhamento para auxiliar na explicação.

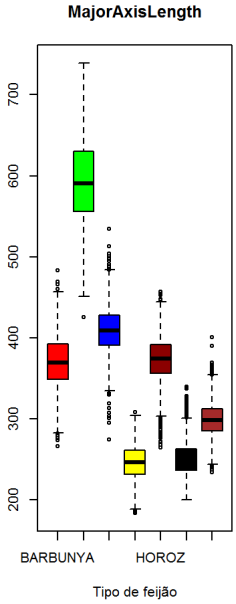
```
nComp = 3
dadosReduzidos = predict(pca, dados)[,1:nComp]
summary(dadosReduzidos)
```

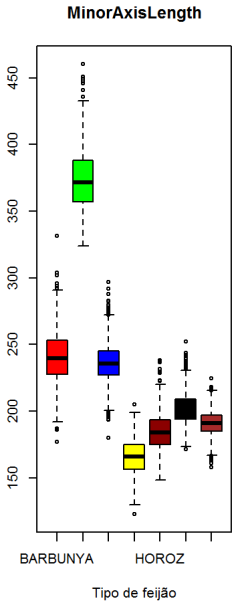
```
##       PC1                PC2                PC3          
##  Min.   :-2.8129   Min.   :-3.4351   Min.   :-0.31683  
##  1st Qu.:-1.2870   1st Qu.:-0.5699   1st Qu.:-0.03773  
##  Median :-0.4688   Median :-0.0226   Median :-0.01526  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.: 1.0646   3rd Qu.: 0.4367   3rd Qu.: 0.01140  
##  Max.   : 8.7164   Max.   : 3.4504   Max.   : 0.90290
```

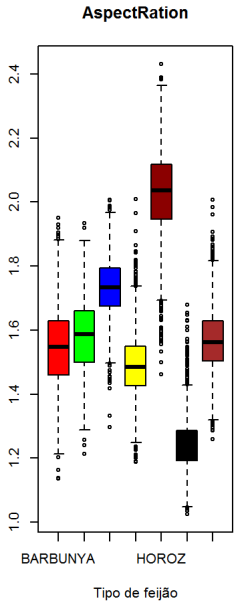
```
cores = c("red","green","blue","yellow","darkred","black","brown")
par(mfrow=c(1,3))
for (i in 1:16) {
boxplot(dados[,i] ~ dados$Class, col=cores, xlab="Tipo de feijão",
ylab="", main=names(dados)[i])
}
```

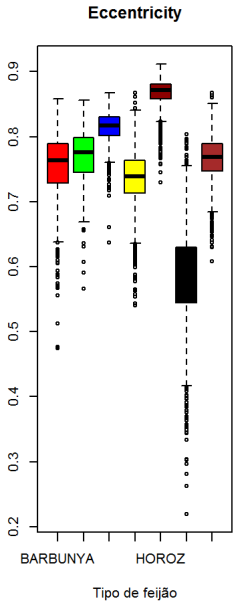

Area
250000
200000
150000
100000
50000
BARBUNYA
HOROZ
Tipo de feijão

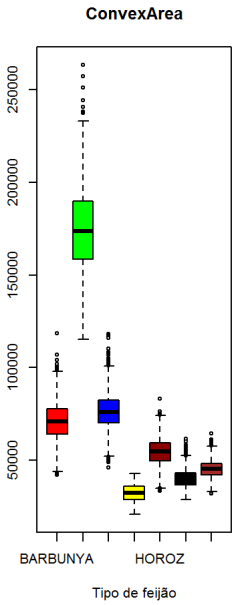

Perimeter
2000
1500
1000
500
BARBUNYA
HOROZ
Tipo de feijão

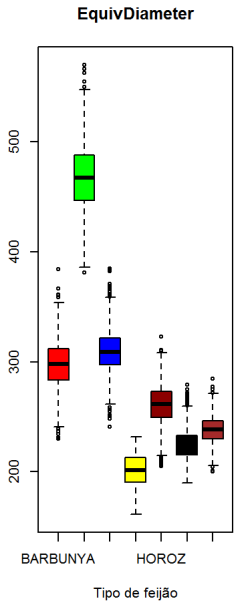

MajorAxisLength
700
600
500
400
300
200
BARBUNYA
HOROZ
Tipo de feijão

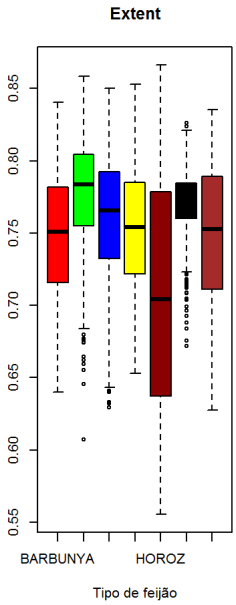

MinorAxisLength
450
400
350
300
250
200
150
BARBUNYA
HOROZ
Tipo de feijão

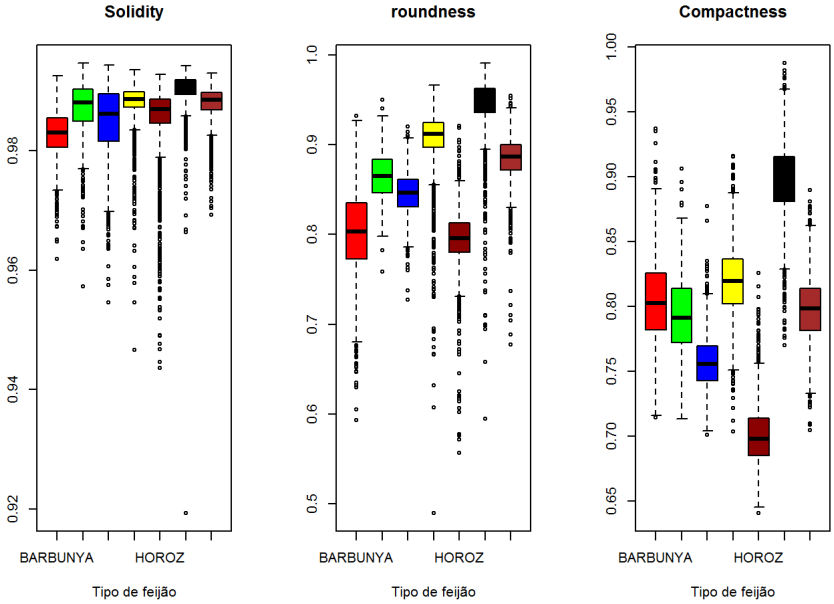

AspectRation
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
BARBUNYA
HOROZ
Tipo de feijão

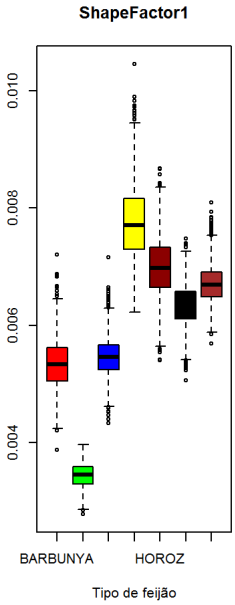

Eccentricity
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
BARBUNYA
HOROZ
Tipo de feijão

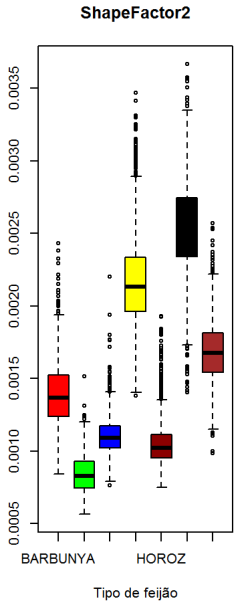

ConvexArea
250000
200000
150000
100000
50000
BARBUNYA
HOROZ
Tipo de feijão

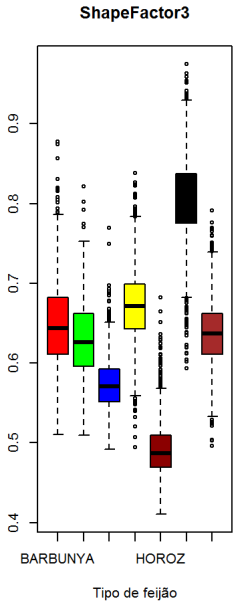

EquivDiameter
500
400
300
200
BARBUNYA
HOROZ
Tipo de feijão

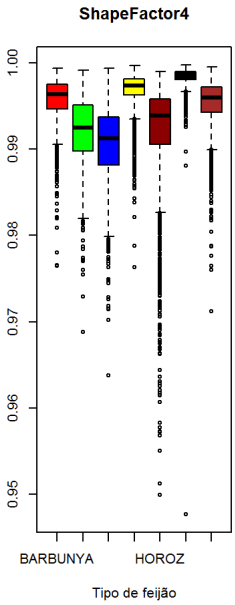

Extent
0.85
0.80
0.75
0.70
0.65
0.60
0.55
BARBUNYA
HOROZ
Tipo de feijão











Pode-se observar que quanto maior o poder de discriminação, possuí também uma alta taxa de correlação. Pode-se perceber que a classe BOMBAY destaca-se das demais.

11. Após ter analisado estas informações, quais considerações você faz sobre este conjunto de dados (ou tarefa)?

Esse conjunto de dados (dataset) DryBeans em complemento a tarefa 1 foi possível observar a relação com os gráficos de espalhamento (correlação) e boxplot (com o poder de discriminação). Foi possível visualizar as informações obtidas, padrões e comportamento das informações através dos gráficos.