TCC R Notebook

Code

Analise exploratória: Continuação.

Introdução.

Este *notebook* investiga a base de dados de propriedades acústicas *Rvoice_fix.csv*, derivada da primeira parte deste estudo.

Hide

```
#install.packages('Amelia')
#install.packages('corrplot')
#install.packages('caret')
#install.packages('ggplot2')
```

Carrega pacote com os dados usados no teste.

Hide

```
library(mlbench)
library(e1071)
library(lattice)
library(Amelia)
library(corrplot)
library(caret)
datasetvoice2 = read.csv("C:\\Users\\jorge\\Desktop\\TCC\\tcc_to_git\\tcc\\baseDados\\Rvoice_fix.csv",sep=',',header=T)
#================================================
# Mostrar dados
#================================================
#View(head(datasetvoice))
#View(tail(datasetvoice))
#print(head(datasetvoice))
```

Verificando alguns dados.

Hide

```
#datasetvoice2$X <- NULL
datasetvoice = datasetvoice2
head(datasetvoice, n=10)
```

	meanfreq <dbl>	sd <dbl>	median <dbl>	Q25 <dbl>	Q75 <dbl>	IQR <dbl>	skew <dbl>	kurt <dbl>	sp.ent <dbl>
1	0.1725574	0.06424127	0.1768932	0.12108928	0.2278422	0.1090547	1.906048	6.450221	0.8933694
2	0.1725574	0.06731003	0.1768932	0.12108928	0.2278422	0.1090547	1.906048	6.450221	0.8921932
3	0.1725574	0.06354869	0.1768932	0.12108928	0.2278422	0.1232070	1.906048	6.450221	0.9185527
4	0.1512281	0.06121566	0.1580112	0.09658173	0.2079553	0.1113735	1.232831	4.177296	0.9633225
5	0.1351204	0.06276914	0.1246562	0.07872022	0.2060449	0.1273247	1.101174	4.333713	0.9719551
6	0.1327864	0.06276914	0.1190898	0.06795799	0.2095916	0.1090547	1.932562	8.308895	0.9631813
7	0.1507623	0.06160811	0.1601064	0.09289894	0.2057181	0.1128191	1.530643	5.987498	0.9675731
8	0.1605143	0.06160811	0.1443368	0.11053217	0.2319619	0.1214297	1.397156	4.766611	0.9592546
9	0.1422394	0.06160811	0.1385874	0.08820628	0.2085874	0.1203812	1.099746	4.070284	0.9707229
10	0.1343288	0.06276914	0.1214513	0.07557999	0.2019571	0.1263771	1.190368	4.787310	0.9752461

1-10 of 10 rows | 1-10 of 21 columns

Verifica a dimensão dos dados (linhas, colunas)

Hide

```
dim(datasetvoice)
```

```
[1] 3168   21
```

Verifica os tipos de dados de cada atributo método 1.

Hide

```
sapply(datasetvoice, class)
```

```
 meanfreq        sd    median       Q25       Q75       IQR      skew      kurt    sp.ent       sfm      mode  centroid   me
anfun    minfun    maxfun   meandom
"numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "num
eric" "numeric" "numeric" "numeric"
   mindom    maxdom   dfrange   modindx     label
"numeric" "numeric" "numeric" "numeric"  "factor"
```

Verifica os tipos de dados de cada atributo método 2.

Hide

```
str(datasetvoice)
```

```
'data.frame':   3168 obs. of  21 variables:
 $ meanfreq: num  0.173 0.173 0.173 0.151 0.135 ...
 $ sd      : num  0.0642 0.0673 0.0635 0.0612 0.0628 ...
 $ median  : num  0.177 0.177 0.177 0.158 0.125 ...
 $ Q25     : num  0.1211 0.1211 0.1211 0.0966 0.0787 ...
 $ Q75     : num  0.228 0.228 0.228 0.208 0.206 ...
 $ IQR     : num  0.109 0.109 0.123 0.111 0.127 ...
 $ skew    : num  1.91 1.91 1.91 1.23 1.1 ...
 $ kurt    : num  6.45 6.45 6.45 4.18 4.33 ...
 $ sp.ent  : num  0.893 0.892 0.919 0.963 0.972 ...
 $ sfm     : num  0.492 0.514 0.479 0.727 0.784 ...
 $ mode    : num  0 0 0 0.0839 0.1043 ...
 $ centroid: num  0.173 0.173 0.173 0.151 0.135 ...
 $ meanfun : num  0.0843 0.1079 0.0987 0.089 0.1064 ...
 $ minfun  : num  0.0157 0.0158 0.0157 0.0178 0.0169 ...
 $ maxfun  : num  0.276 0.274 0.271 0.274 0.275 ...
 $ meandom : num  0.00781 0.00901 0.00799 0.2015 0.71281 ...
 $ mindom  : num  0.00781 0.00781 0.00781 0.00781 0.00781 ...
 $ maxdom  : num  0.00781 0.05469 0.01562 0.5625 5.48438 ...
 $ dfrange : num  0 0.04688 0.00781 0.55469 5.47656 ...
 $ modindx : num  0.133 0.125 0.125 0.13 0.125 ...
 $ label   : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
```

Estatística descritiva.

Hide

```
summary(datasetvoice)
```

```
    meanfreq            sd              median            Q25               Q75             IQR              skew
kurt            sp.ent
 Min.   :0.1103   Min.   :0.01836   Min.   :0.08067   Min.   :0.06776   Min.   :0.1578   Min.   :0.01456   Min.   :0.6923
Min.   : 2.210   Min.   :0.7387
 1st Qu.:0.1680   1st Qu.:0.04197   1st Qu.:0.17263   1st Qu.:0.12165   1st Qu.:0.2094   1st Qu.:0.04033   1st Qu.:1.6618
1st Qu.: 5.711   1st Qu.:0.8629
 Median :0.1865   Median :0.05953   Median :0.19125   Median :0.14935   Median :0.2263   Median :0.07508   Median :1.9060
Median : 6.450   Median :0.9026
 Mean   :0.1843   Mean   :0.05483   Mean   :0.18877   Mean   :0.15189   Mean   :0.2257   Mean   :0.07427   Mean   :2.0683
Mean   : 7.398   Mean   :0.8958
 3rd Qu.:0.1991   3rd Qu.:0.06245   3rd Qu.:0.21062   3rd Qu.:0.18193   3rd Qu.:0.2437   3rd Qu.:0.10905   3rd Qu.:2.4283
3rd Qu.: 9.158   3rd Qu.:0.9287
 Max.   :0.2511   Max.   :0.09606   Max.   :0.26122   Max.   :0.23178   Max.   :0.2735   Max.   :0.13200   Max.   :4.1249
Max.   :16.053   Max.   :0.9820
      sfm              mode           centroid         meanfun            minfun             maxfun          meandom
mindom             maxdom
 Min.   :0.03688   Min.   :0.0000   Min.   :0.1103   Min.   :0.07025   Min.   :0.009775   Min.   :0.2697   Min.   :0.007812
Min.   :0.004883   Min.   : 0.007812
 1st Qu.:0.25804   1st Qu.:0.1480   1st Qu.:0.1680   1st Qu.:0.11692   1st Qu.:0.018223   1st Qu.:0.2739   1st Qu.:0.419828
1st Qu.:0.007812   1st Qu.: 2.070312
 Median :0.39634   Median :0.1956   Median :0.1865   Median :0.14050   Median :0.043340   Median :0.2752   Median :0.759524
Median :0.023438   Median : 4.953125
 Mean   :0.40822   Mean   :0.1764   Mean   :0.1843   Mean   :0.14283   Mean   :0.035703   Mean   :0.2755   Mean   :0.823085
Mean   :0.040827   Mean   : 4.922102
 3rd Qu.:0.53368   3rd Qu.:0.2193   3rd Qu.:0.1991   3rd Qu.:0.16967   3rd Qu.:0.047856   3rd Qu.:0.2775   3rd Qu.:1.167568
3rd Qu.:0.023438   3rd Qu.: 6.984375
 Max.   :0.84294   Max.   :0.2800   Max.   :0.2511   Max.   :0.21726   Max.   :0.091743   Max.   :0.2791   Max.   :2.591580
Max.   :0.281250   Max.   :17.343750
    dfrange          modindx           label
 Min.   : 0.000   Min.   :0.06108   female:1584
 1st Qu.: 2.045   1st Qu.:0.11167   male  :1584
 Median : 4.922   Median :0.12587
 Mean   : 4.870   Mean   :0.12384
 3rd Qu.: 6.906   3rd Qu.:0.13526
 Max.   :17.320   Max.   :0.18534
```

Distribuição das classes.

Hide

```
y <- datasetvoice$label
cbind(freq=table(y), percentage=prop.table(table(y))*100)
```

```
       freq percentage
female 1584         50
male   1584         50
```

Desvio padrão.

Hide

```
sapply(datasetvoice[,1:20], sd)
```

```
   meanfreq          sd      median         Q25         Q75         IQR        skew        kurt      sp.ent         sfm
mode    centroid     meanfun
0.025579860 0.013947331 0.031509174 0.036211949 0.021551118 0.036408918 0.635557030 2.746345998 0.044617526 0.177521105 0.06
6645690 0.025579860 0.031740970
     minfun      maxfun     meandom      mindom      maxdom     dfrange     modindx
0.015903587 0.002263936 0.516908567 0.056160053 3.262845264 3.261823988 0.023245084
```

Skew.

Hide

```
skew <- apply(datasetvoice[,1:20], 2, skewness)
print(skew)
```

```
   meanfreq          sd      median         Q25         Q75         IQR        skew        kurt      sp.ent         sfm
mode    centroid     meanfun
-0.20860871 -0.10565308 -0.59258876  0.04452280 -0.25129521 -0.02571567  0.70331007  0.99197716 -0.45738916  0.33963572 -1.1
4683518 -0.20860871  0.04511686
     minfun      maxfun     meandom      mindom      maxdom     dfrange     modindx
 0.16214834 -0.09879414  0.57694742  2.14563423  0.30101478  0.30356447 -0.32528125
```

Correlação.

Hide

```
correlacao <- cor(datasetvoice[,1:20])
print(correlacao)
```

```
           meanfreq          sd     median         Q25         Q75         IQR         skew         kurt     sp.ent        s
fm       mode   centroid     meanfun
meanfreq  1.0000000 -0.53907703  0.8789897  0.74189791  0.62694238 -0.44394912  0.161847369  0.145889426 -0.6300724 -0.74136
38  0.6450026  1.0000000  0.58034385
sd       -0.5390770  1.00000000 -0.3621163 -0.50317312 -0.08281425  0.49773507 -0.375903950 -0.336718814  0.8001013  0.78341
24 -0.2078177 -0.5390770 -0.41062245
median    0.8789897 -0.36211626  1.0000000  0.64952918  0.63357624 -0.36788424  0.124161528  0.114531752 -0.5285481 -0.63600
70  0.7145617  0.8789897  0.51976772
Q25       0.7418979 -0.50317312  0.6495292  1.00000000  0.26264251 -0.83221975  0.414358412  0.421511407 -0.6339032 -0.56752
72  0.5479519  0.7418979  0.86956247
Q75       0.6269424 -0.08281425  0.6335762  0.26264251  1.00000000  0.16141858 -0.221184691 -0.235552575 -0.2004399 -0.38374
22  0.4201490  0.6269424  0.12781336
IQR      -0.4439491  0.49773507 -0.3678842 -0.83221975  0.16141858  1.00000000 -0.573220375 -0.593070850  0.5700820  0.40259
12 -0.3633326 -0.4439491 -0.83801756
skew      0.1618474 -0.37590395  0.1241615  0.41435841 -0.22118469 -0.57322038  1.000000000  0.829802441 -0.5219561 -0.29852
61  0.1021200  0.1618474  0.45718271
kurt      0.1458894 -0.33671881  0.1145318  0.42151141 -0.23555258 -0.59307085  0.829802441  1.000000000 -0.4581040 -0.25024
21  0.1110559  0.1458894  0.46951283
sp.ent   -0.6300724  0.80010126 -0.5285481 -0.63390325 -0.20043993  0.57008196 -0.521956133 -0.458104006  1.0000000  0.87386
84 -0.3207432 -0.6300724 -0.54722468
sfm      -0.7413638  0.78341240 -0.6360070 -0.56752723 -0.38374217  0.40259120 -0.298526096 -0.250242120  0.8738684  1.00000
00 -0.4005448 -0.7413638 -0.42289526
mode      0.6450026 -0.20781767  0.7145617  0.54795195  0.42014899 -0.36333255  0.102120001  0.111055927 -0.3207432 -0.40054
48  1.0000000  0.6450026  0.50020343
centroid  1.0000000 -0.53907703  0.8789897  0.74189791  0.62694238 -0.44394912  0.161847369  0.145889426 -0.6300724 -0.74136
38  0.6450026  1.0000000  0.58034385
meanfun   0.5803439 -0.41062245  0.5197677  0.86956247  0.12781336 -0.83801756  0.457182706  0.469512834 -0.5472247 -0.42289
53  0.5002034  0.5803439  1.00000000
minfun    0.4471326 -0.29873109  0.4023625  0.29174581  0.32225074 -0.12708623 -0.020103358 -0.042859314 -0.3358209 -0.41607
56  0.3734247  0.4471326  0.29209514
maxfun    0.3664865 -0.23145759  0.3328092  0.29163406  0.28487030 -0.15744765  0.027513203  0.024123794 -0.2439865 -0.28330
59  0.2766606  0.3664865  0.26004771
meandom   0.5262278 -0.41669089  0.4537282  0.36324552  0.34705342 -0.20521020  0.001488233 -0.004482325 -0.3209281 -0.43235
78  0.3957602  0.5262278  0.26985842
mindom    0.2367805 -0.44145626  0.1729833  0.35778873 -0.10308377 -0.45958668  0.343609438  0.324337240 -0.4175981 -0.34241
71  0.1666245  0.2367805  0.38124573
maxdom    0.5206479 -0.45074506  0.4483106  0.38084700  0.33338827 -0.21101601  0.033699992  0.033735697 -0.3520423 -0.44704
93  0.3881853  0.5206479  0.28349615
dfrange   0.5173670 -0.44490680  0.4455382  0.37765749  0.33461447 -0.20626188  0.029621721  0.029955283 -0.3463337 -0.44162
66  0.3855417  0.5173670  0.28052545
modindx  -0.1492124  0.12171752 -0.1437588 -0.09790834 -0.12761877  0.03321638 -0.025162134 -0.031207964  0.1344716  0.16678
79 -0.1066373 -0.1492124 -0.07268063
              minfun      maxfun      meandom      mindom      maxdom     dfrange     modindx
meanfreq  0.44713265  0.36648651  0.526227844  0.23678048  0.52064785  0.51736705 -0.14921239
sd       -0.29873109 -0.23145759 -0.416690891 -0.44145626 -0.45074506 -0.44490680  0.12171752
median    0.40236249  0.33280917  0.453728209  0.17298333  0.44831065  0.44553819 -0.14375876
Q25       0.29174581  0.29163406  0.363245525  0.35778873  0.38084700  0.37765749 -0.09790834
Q75       0.32225074  0.28487030  0.347053424 -0.10308377  0.33338827  0.33461447 -0.12761877
IQR      -0.12708623 -0.15744765 -0.205210197 -0.45958668 -0.21101601 -0.20626188  0.03321638
skew     -0.02010336  0.02751320  0.001488233  0.34360944  0.03369999  0.02962172 -0.02516213
kurt     -0.04285931  0.02412379 -0.004482325  0.32433724  0.03373570  0.02995528 -0.03120796
sp.ent   -0.33582090 -0.24398650 -0.320928061 -0.41759807 -0.35204226 -0.34633373  0.13447159
sfm      -0.41607557 -0.28330592 -0.432357755 -0.34241713 -0.44704928 -0.44162659  0.16678789
mode      0.37342468  0.27666059  0.395760236  0.16662445  0.38818527  0.38554172 -0.10663726
centroid  0.44713265  0.36648651  0.526227844  0.23678048  0.52064785  0.51736705 -0.14921239
meanfun   0.29209514  0.26004771  0.269858425  0.38124573  0.28349615  0.28052545 -0.07268063
minfun    1.00000000  0.38194102  0.500841540  0.08712209  0.44981697  0.44960960 -0.16075475
maxfun    0.38194102  1.00000000  0.380625820  0.04105787  0.36172587  0.36319426 -0.16064144
meandom   0.50084154  0.38062582  1.000000000  0.14457892  0.82012582  0.81841385 -0.04039630
mindom    0.08712209  0.04105787  0.144578916  1.00000000  0.10221439  0.08660828  0.01816070
maxdom    0.44981697  0.36172587  0.820125823  0.10221439  1.00000000  0.99981447 -0.24011166
dfrange   0.44960960  0.36319426  0.818413853  0.08660828  0.99981447  1.00000000 -0.24158354
modindx  -0.16075475 -0.16064144 -0.040396297  0.01816070 -0.24011166 -0.24158354  1.00000000
```

Histograma (univariado).

Hide

```
par(mfrow=c(5,4))
for(i in 1:20) {
  hist(datasetvoice[,i], main=names(datasetvoice)[i])
}
```

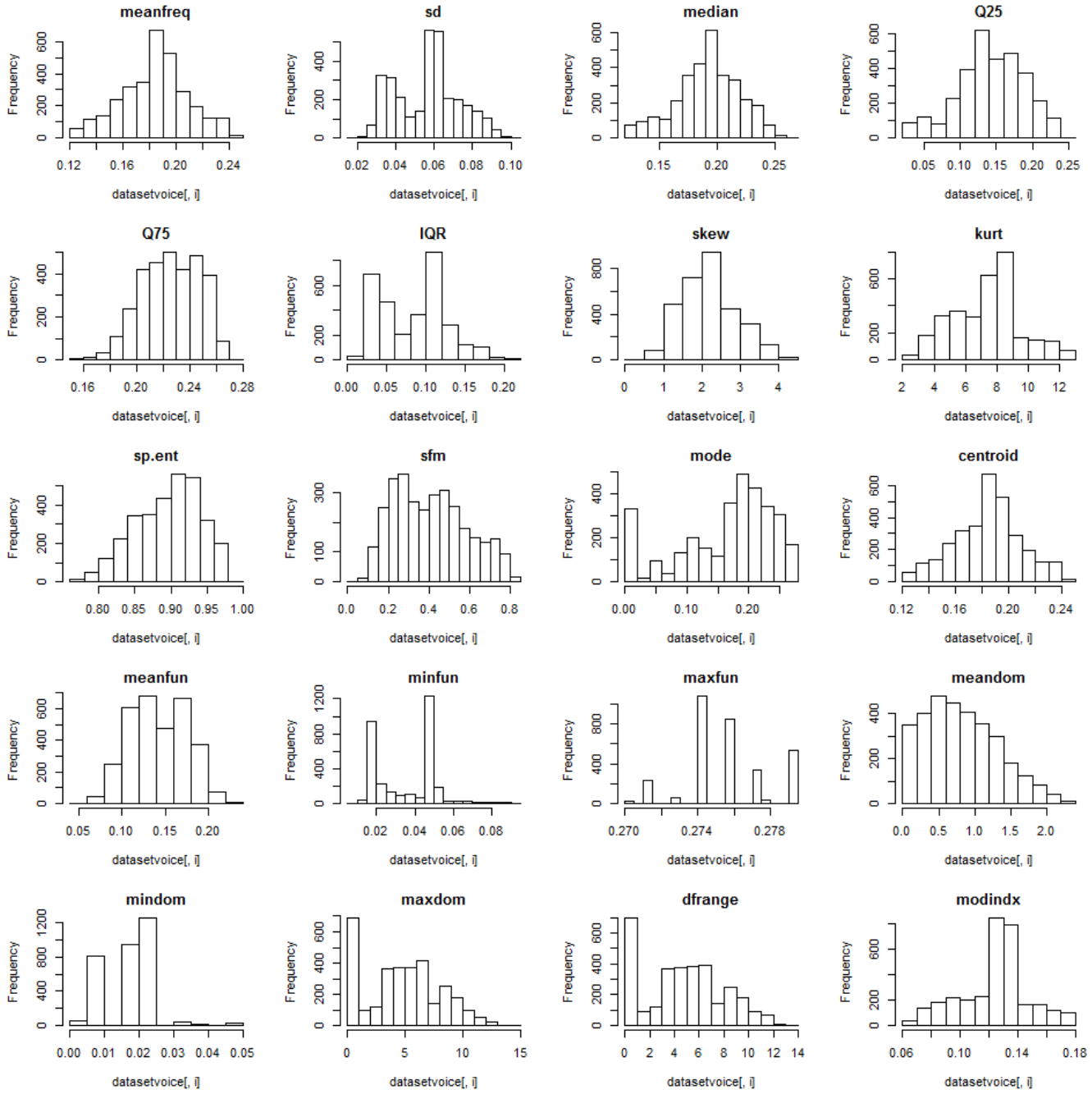
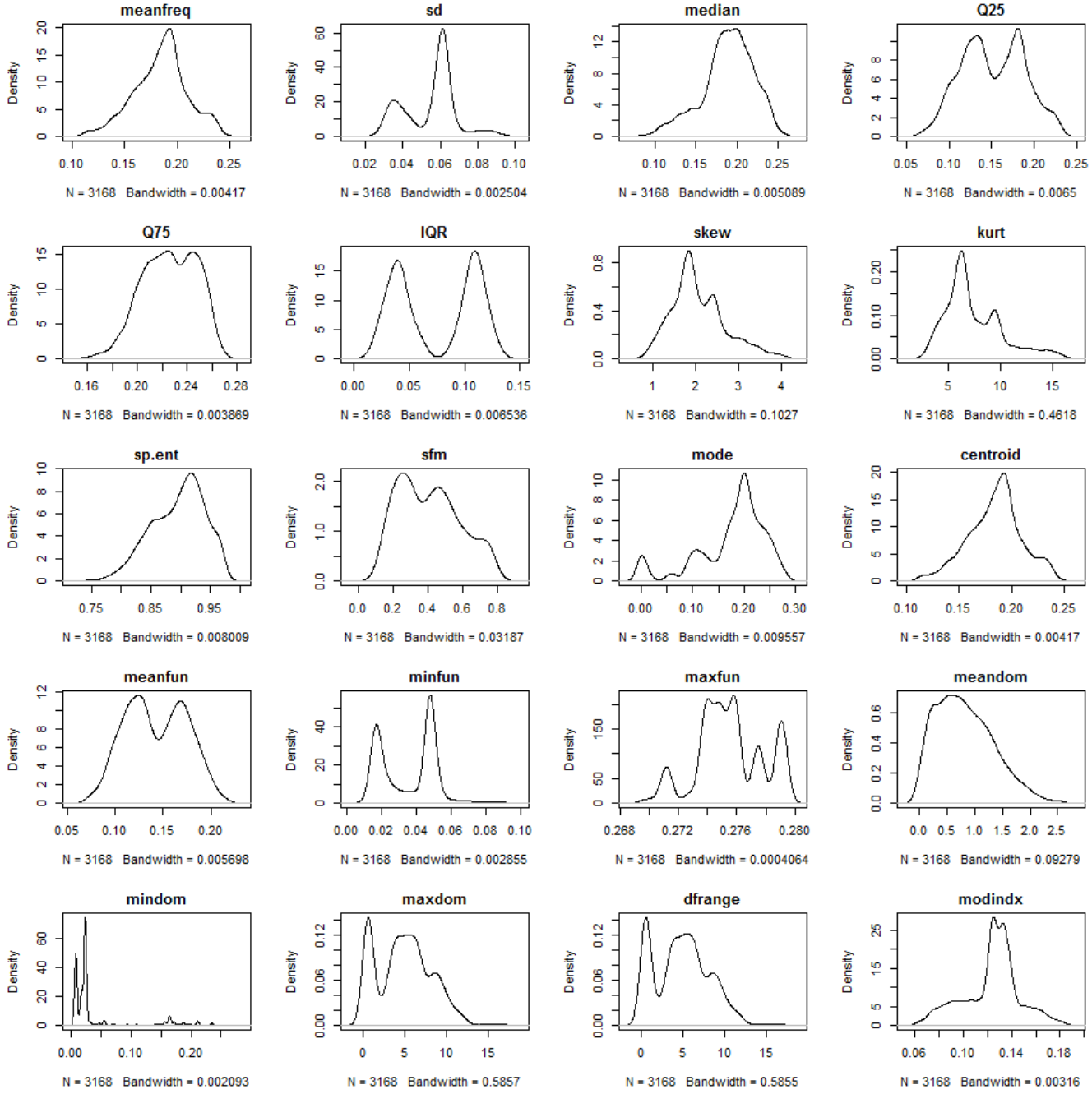


Gráfico de densidade (univariado).

Hide

```
par(mfrow=c(5,4))
for(i in 1:20) {
  plot(density(datasetvoice[,i]), main=names(datasetvoice)[i])
}
```



Boxplot e Whisker (univariado).

Hide

```
par(mfrow=c(5,4))
for(i in 1:20) {
  boxplot(datasetvoice[,i], main=names(datasetvoice)[i])
}
```

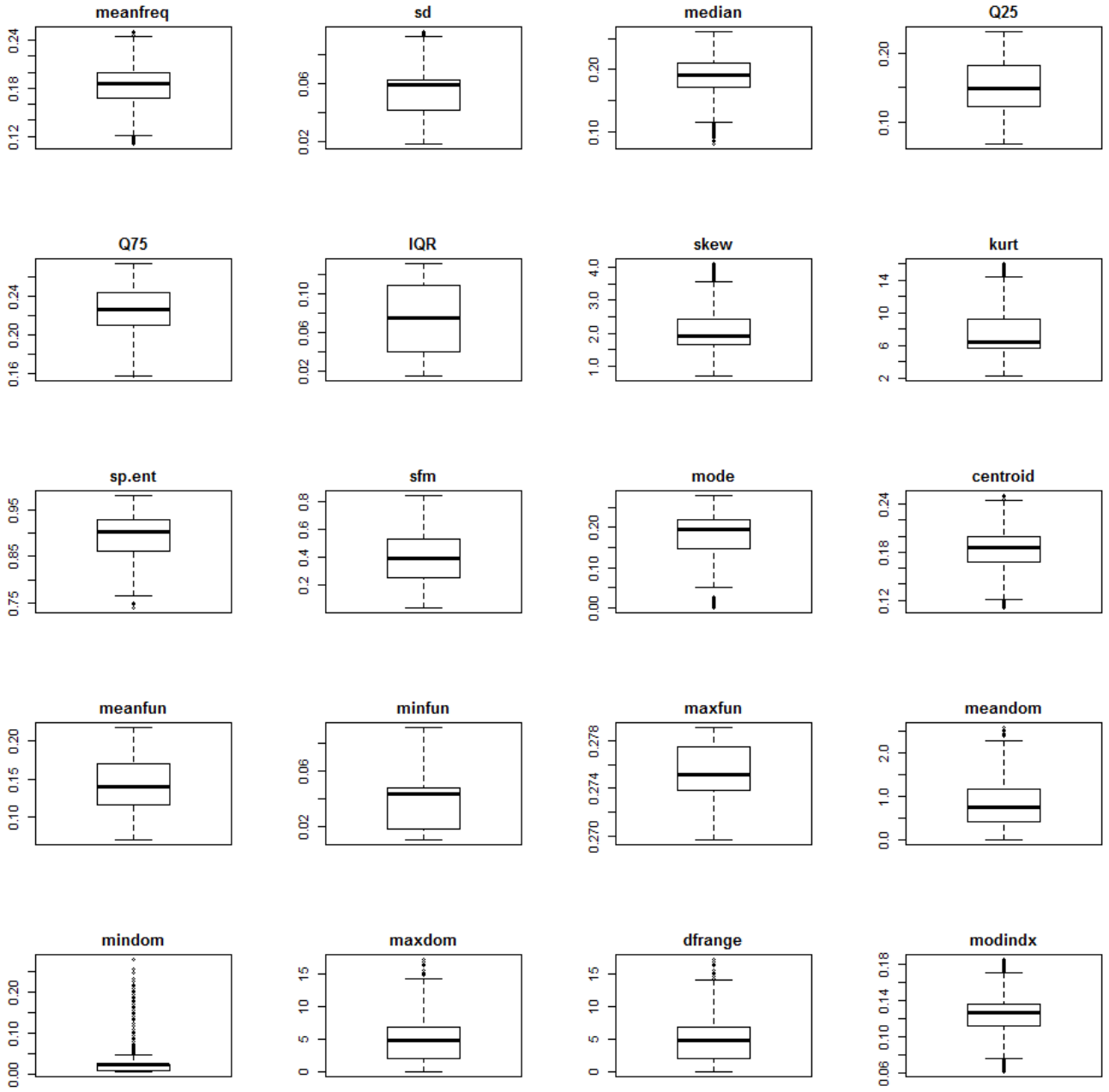
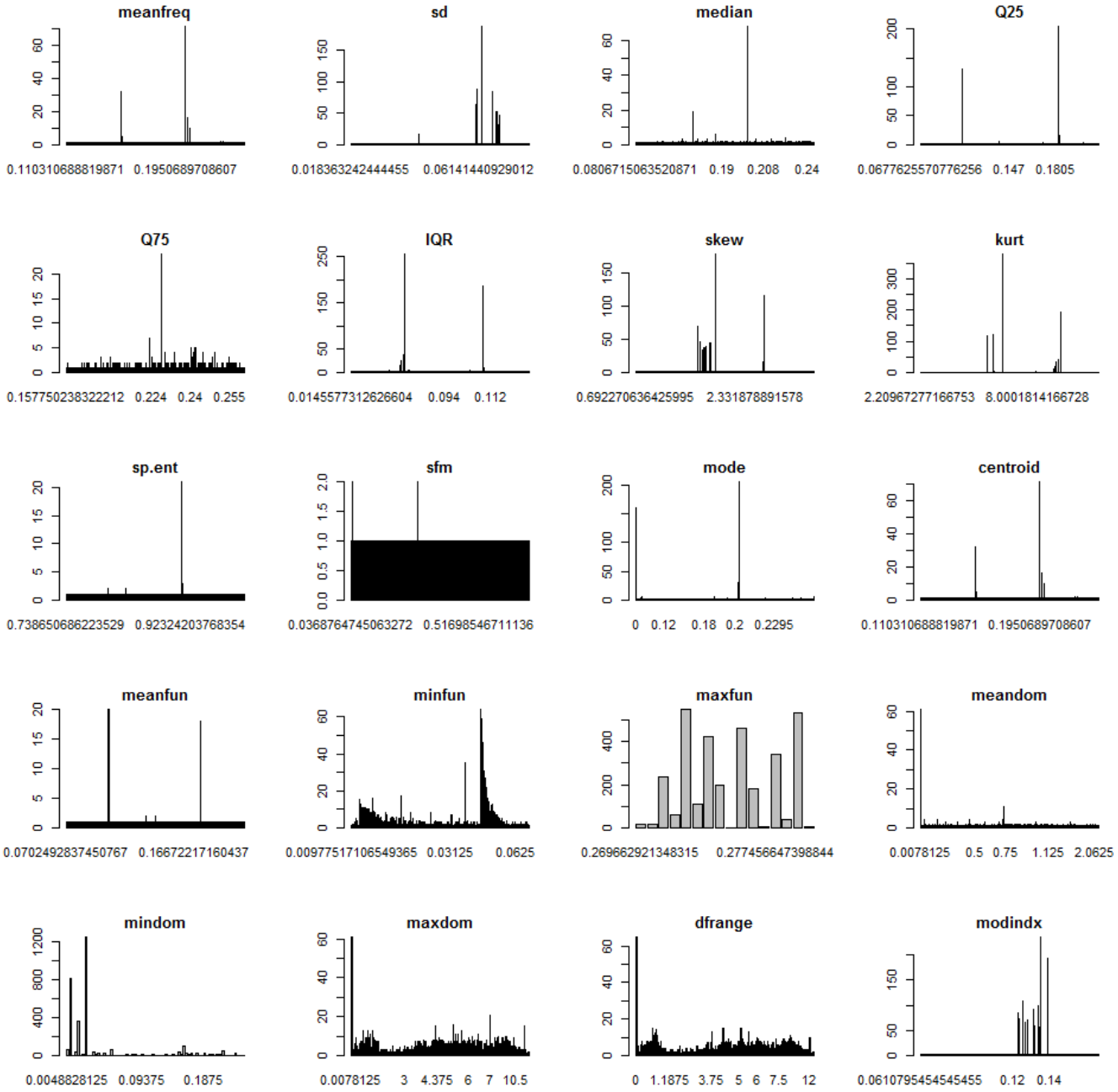


Gráfico de barras.

Hide

```
par(mfrow=c(5,4))
for(i in 1:20) {
  counts <- table(datasetvoice[,i])
  name <- names(datasetvoice)[i]
  barplot(counts, main=name)
}
```



Mapa de valores ausentes (univariado). #

`{r fig.width = 10, fig.height = 10} #par(mfrow=c(1,1)) #datasetvoice(Soybean) #missmap(Soybean, col=c("black", "grey"), legend=FALSE)`

Gráfico de correlação (multivariado)

Hide

```
correlacao <- cor(datasetvoice[,1:20])
cores <- colorRampPalette(c("red", "white", "blue"))
corrplot(correlacao, order="AOE", method="square", col=cores(20), tl.srt=45, tl.cex=0.75, tl.col="black")
corrplot(correlacao, add=TRUE, type="lower", method="number", order="AOE", col="black", diag=FALSE, tl.pos="n", cl.pos="n",
 number.cex=0.75)
```

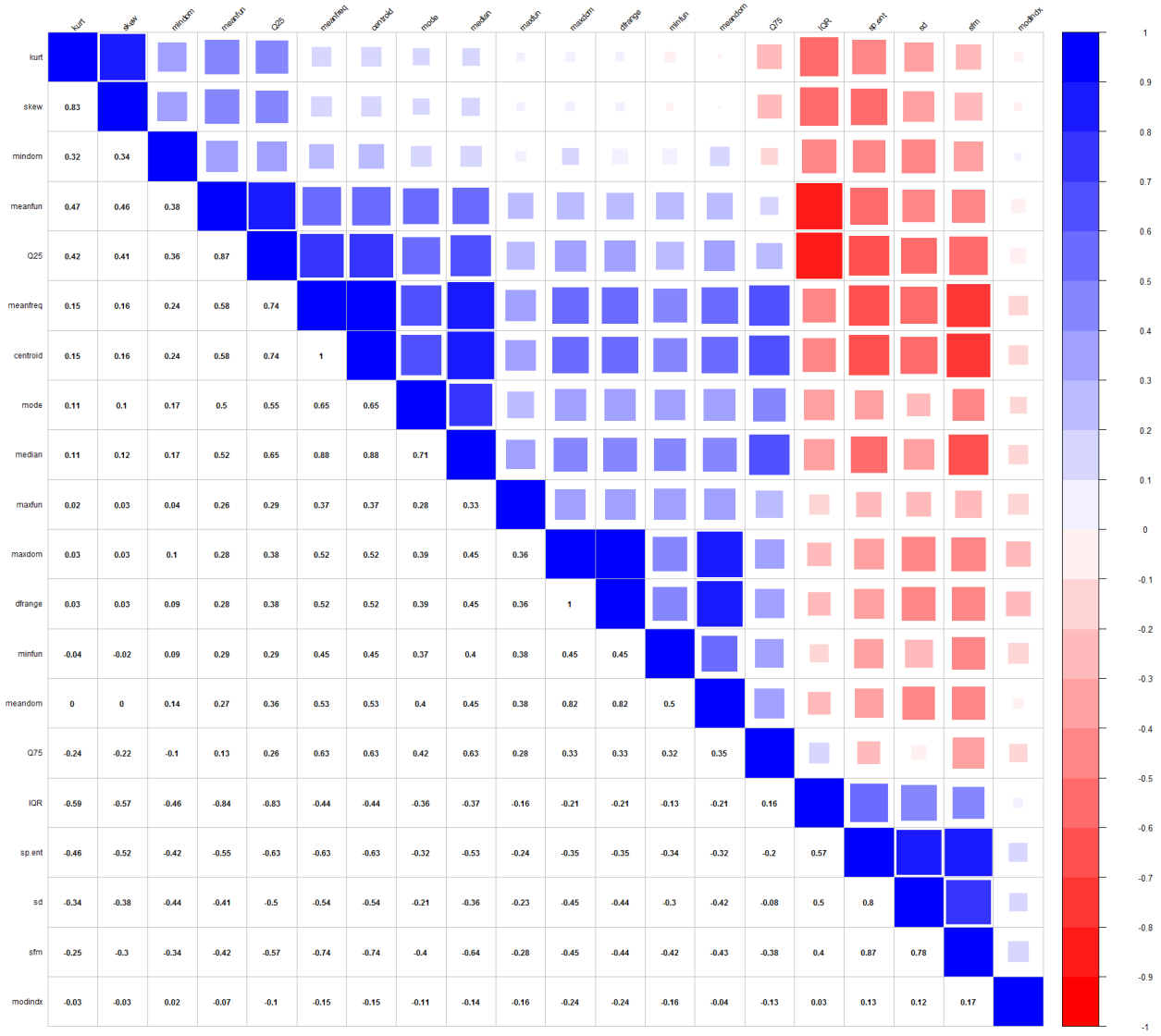
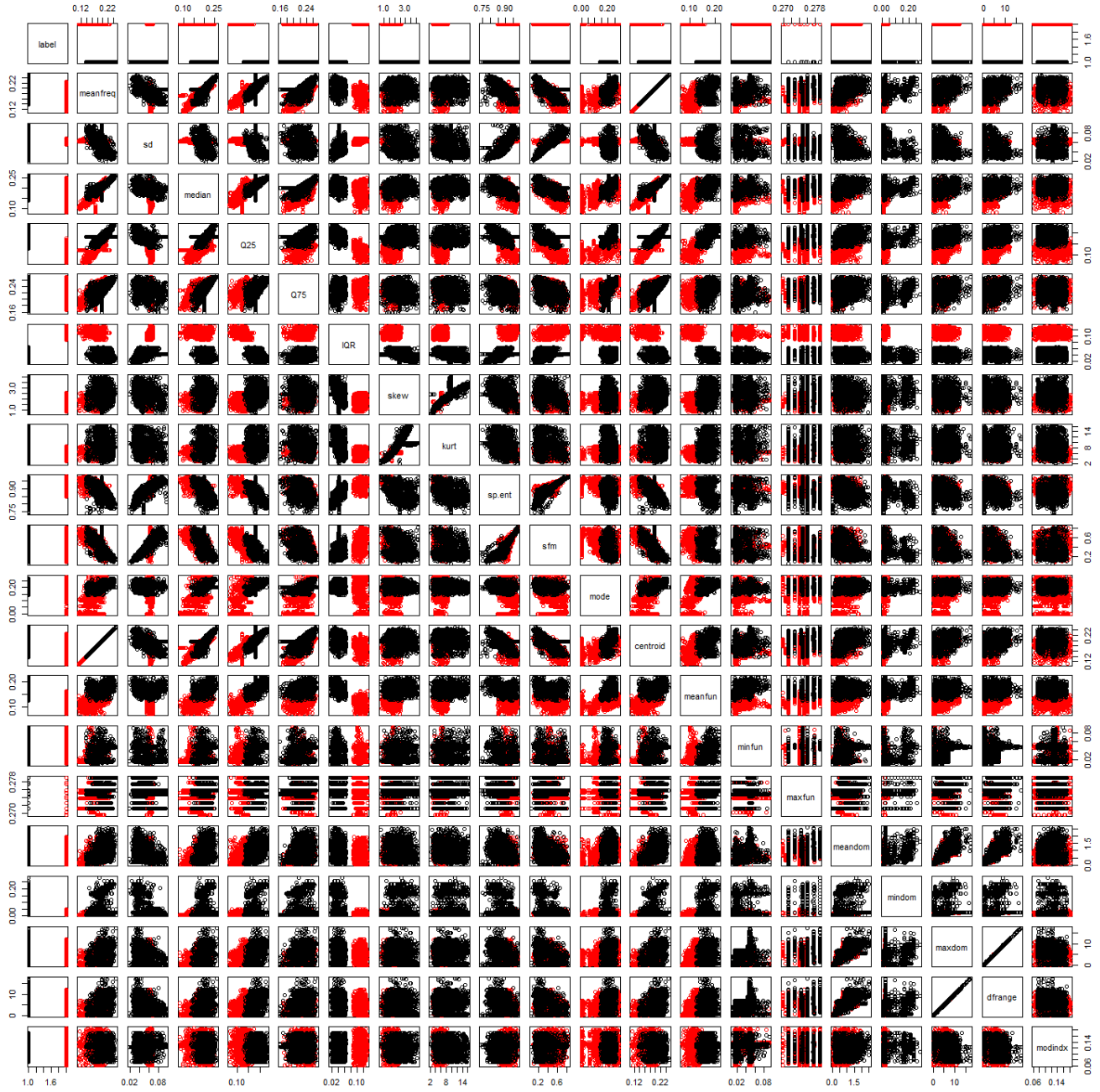


Gráfico de dispersão por classe (multivariado).

Hide

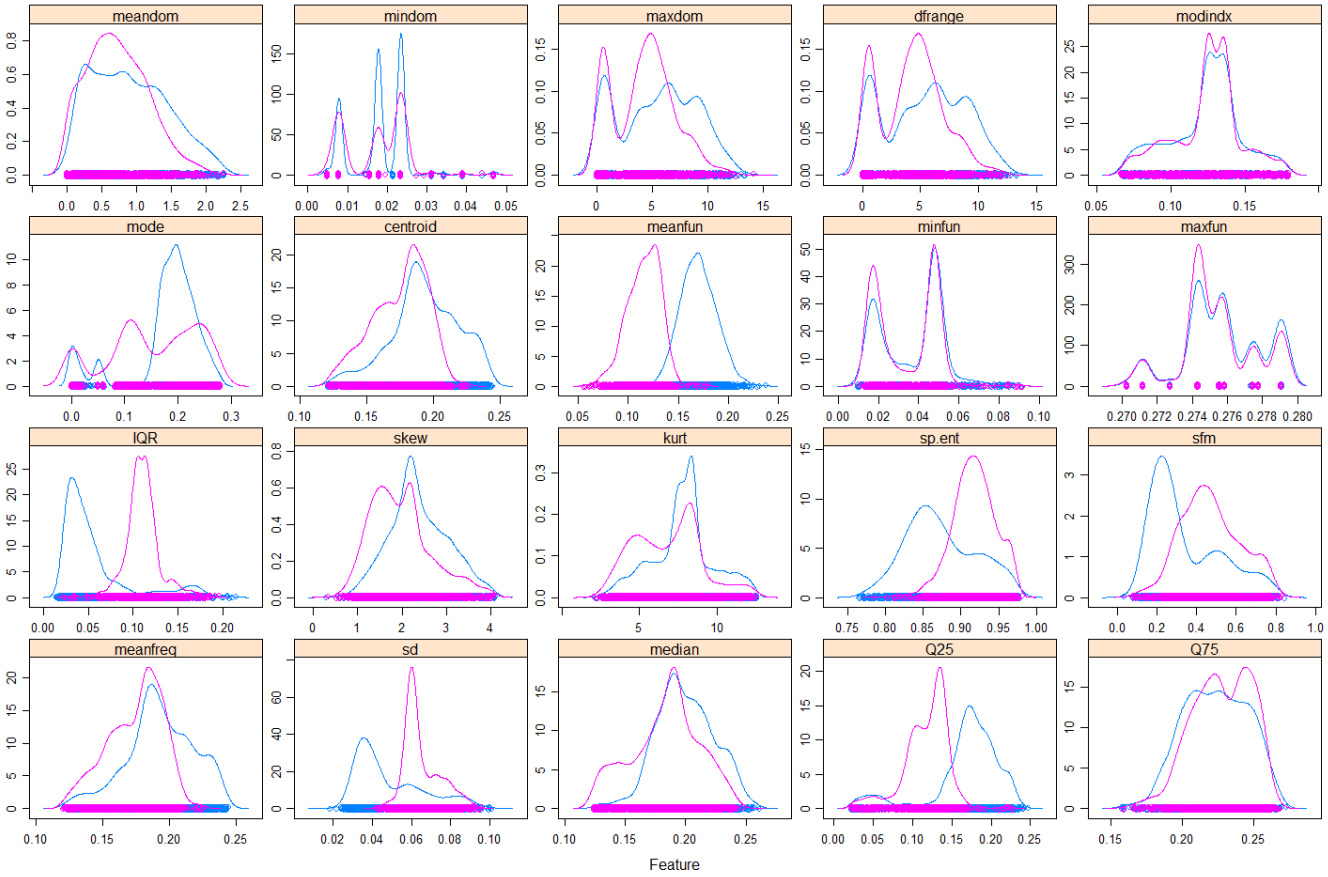
```
pairs(label~., data=datasetvoice, col=datasetvoice$label)
```



Gr?fico de densidade por classe (multivariado).

Hide

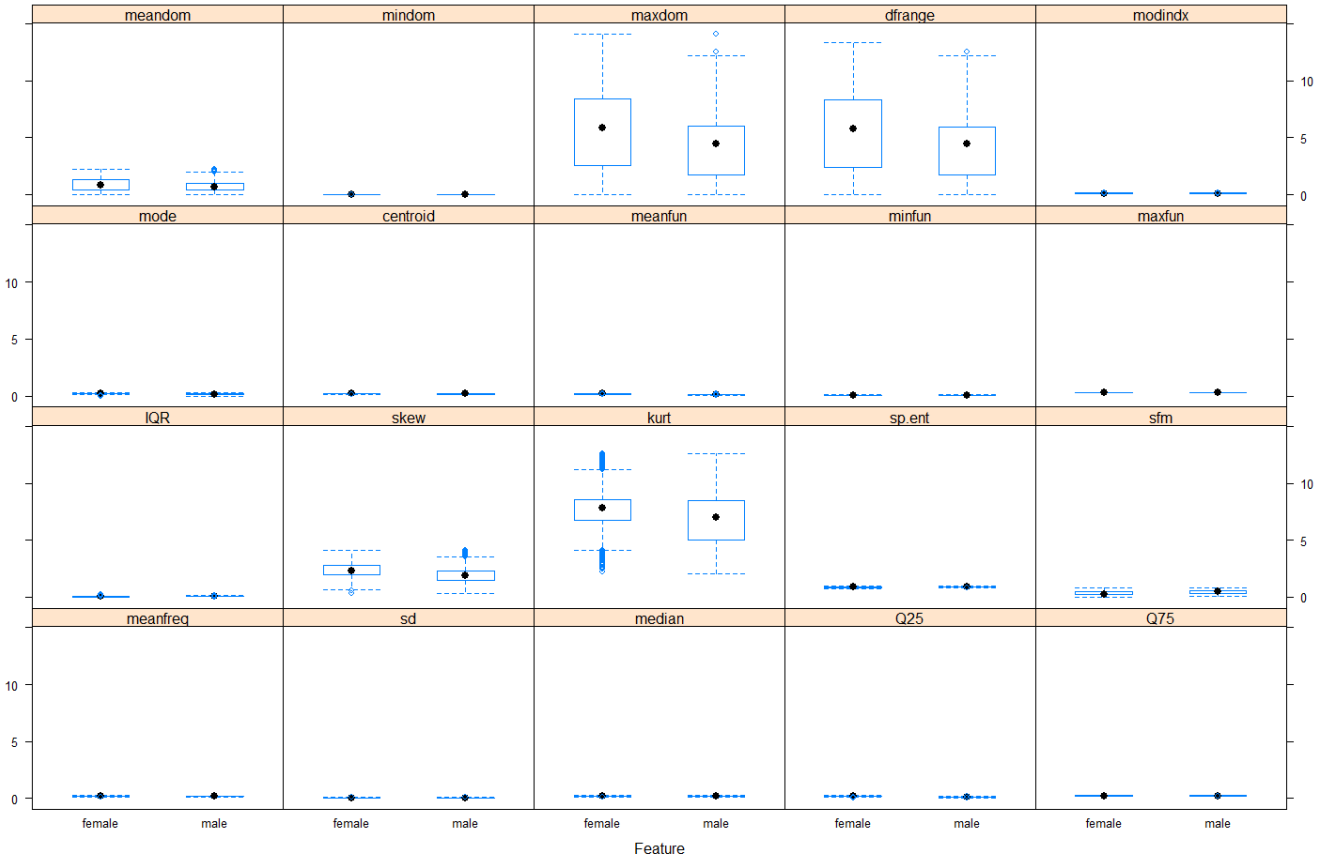
```
x <- datasetvoice[,1:20]
y <- datasetvoice[,21]
scales <- list(x=list(relation="free"), y=list(relation="free"))
featurePlot(x=x, y=y, plot="density", scales=scales)
```



Boxplot por classe (multivariado)

Hide

```
x <- datasetvoice[,1:20]
y <- datasetvoice[,21]
featurePlot(x=x, y=y, plot="box")
```



Fim da analise