TCC R Notebook

Analise exploratória.

Introdução.

Este Jupyter Notebook investiga a base de dados de propriedades acústicas disponível no site http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/
Objetivo da investigação é determinar as chances de algum algoritmo para detecção de gênero, seja por estatística tradicional ou por meio técnicas machine learning e redes neurais, possibilitando a implantação em dispositivos embarcados de baixo custo de memoria e processamento restrito, para utilização de mídias inteligentes e interativas em lojas em moda.

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

Carrega pacote com os dados usados no teste.

```
library(mlbench)
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.4.4
```

```
library(lattice)
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 3.4.4
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.4.4
```

```
## corrplot 0.84 loaded
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
datasetvoice = read.csv("C:\\Users\\jorge\\Desktop\\TCC\\00-PRATICA\\04-Resultados TCC\\baseDados\\Tvoi

#================================================
# Mostrar dados
#================================================
#View(head(datasetvoice))
#View(tail(datasetvoice))
#print(head(datasetvoice))
```

Verificando alguns dados.

```
head(datasetvoice, n=10)
```

```
##      meanfreq         sd     median         Q25        Q75        IQR
## 1  0.05978098 0.06424127 0.03202691 0.015071489 0.09019344 0.07512195
## 2  0.06600874 0.06731003 0.04022873 0.019413867 0.09266619 0.07325232
## 3  0.07731550 0.08382942 0.03671846 0.008701057 0.13190802 0.12320696
## 4  0.15122809 0.07211059 0.15801119 0.096581728 0.20795525 0.11137352
## 5  0.13512039 0.07914610 0.12465623 0.078720218 0.20604493 0.12732471
## 6  0.13278641 0.07955687 0.11908985 0.067957993 0.20959160 0.14163361
## 7  0.15076233 0.07446321 0.16010638 0.092898936 0.20571809 0.11281915
## 8  0.16051433 0.07676688 0.14433678 0.110532168 0.23196187 0.12142971
## 9  0.14223942 0.07801846 0.13858744 0.088206278 0.20858744 0.12038117
## 10 0.13432878 0.08035003 0.12145135 0.075579989 0.20195712 0.12637713
##         skew        kurt    sp.ent       sfm       mode   centroid
## 1  12.863462  274.402906 0.8933694 0.4919178 0.00000000 0.05978098
## 2  22.423285  634.613855 0.8921932 0.5137238 0.00000000 0.06600874
## 3  30.757155 1024.927705 0.8463891 0.4789050 0.00000000 0.07731550
## 4   1.232831    4.177296 0.9633225 0.7272318 0.08387819 0.15122809
## 5   1.101174    4.333713 0.9719551 0.7835681 0.10426140 0.13512039
## 6   1.932562    8.308895 0.9631813 0.7383070 0.11255543 0.13278641
## 7   1.530643    5.987498 0.9675731 0.7626377 0.08619681 0.15076233
## 8   1.397156    4.766611 0.9592546 0.7198579 0.12832407 0.16051433
## 9   1.099746    4.070284 0.9707229 0.7709921 0.21910314 0.14223942
## 10  1.190368    4.787310 0.9752461 0.8045053 0.01169874 0.13432878
##       meanfun     minfun    maxfun     meandom    mindom    maxdom
## 1  0.08427911 0.01570167 0.2758621 0.007812500 0.0078125 0.0078125
## 2  0.10793655 0.01582591 0.2500000 0.009014423 0.0078125 0.0546875
## 3  0.09870626 0.01565558 0.2711864 0.007990057 0.0078125 0.0156250
## 4  0.08896485 0.01779755 0.2500000 0.201497396 0.0078125 0.5625000
## 5  0.10639784 0.01693122 0.2666667 0.712812500 0.0078125 5.4843750
## 6  0.11013192 0.01711230 0.2539683 0.298221983 0.0078125 2.7265625
## 7  0.10594452 0.02622951 0.2666667 0.479619565 0.0078125 5.3125000
## 8  0.09305243 0.01775805 0.1441441 0.301339286 0.0078125 0.5390625
## 9  0.09672895 0.01795735 0.2500000 0.336476293 0.0078125 2.1640625
## 10 0.10588093 0.01930036 0.2622951 0.340364583 0.0156250 4.6953125
##      dfrange    modindx label
## 1  0.0000000 0.00000000  male
## 2  0.0468750 0.05263158  male
## 3  0.0078125 0.04651163  male
## 4  0.5546875 0.24711908  male
## 5  5.4765625 0.20827389  male
## 6  2.7187500 0.12515964  male
## 7  5.3046875 0.12399186  male
## 8  0.5312500 0.28393665  male
## 9  2.1562500 0.14827202  male
## 10 4.6796875 0.08991998  male
```

Verifica a dimensão dos dados (linhas, colunas)

```
dim(datasetvoice)
```

```
## [1] 3168   21
```

Verifica os tipos de dados de cada atributo método 1.

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

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

Verifica os tipos de dados de cada atributo método 2.

```
str(datasetvoice)
```

```
## 'data.frame':    3168 obs. of  21 variables:
##  $ meanfreq: num  0.0598 0.066 0.0773 0.1512 0.1351 ...
##  $ sd      : num  0.0642 0.0673 0.0838 0.0721 0.0791 ...
##  $ median  : num  0.032 0.0402 0.0367 0.158 0.1247 ...
##  $ Q25     : num  0.0151 0.0194 0.0087 0.0966 0.0787 ...
##  $ Q75     : num  0.0902 0.0927 0.1319 0.208 0.206 ...
##  $ IQR     : num  0.0751 0.0733 0.1232 0.1114 0.1273 ...
##  $ skew    : num  12.86 22.42 30.76 1.23 1.1 ...
##  $ kurt    : num  274.4 634.61 1024.93 4.18 4.33 ...
##  $ sp.ent  : num  0.893 0.892 0.846 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  0.0598 0.066 0.0773 0.1512 0.1351 ...
##  $ 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 0.25 0.271 0.25 0.267 ...
##  $ 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 0.0526 0.0465 0.2471 0.2083 ...
##  $ label   : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
```

Estatística descritiva.

```
summary(datasetvoice)
```

```
##     meanfreq             sd              median             Q25
##  Min.   :0.03936   Min.   :0.01836   Min.   :0.01097   Min.   :0.0002288
##  1st Qu.:0.16366   1st Qu.:0.04195   1st Qu.:0.16959   1st Qu.:0.1110865
##  Median :0.18484   Median :0.05916   Median :0.19003   Median :0.1402864
##  Mean   :0.18091   Mean   :0.05713   Mean   :0.18562   Mean   :0.1404556
##  3rd Qu.:0.19915   3rd Qu.:0.06702   3rd Qu.:0.21062   3rd Qu.:0.1759388
##  Max.   :0.25112   Max.   :0.11527   Max.   :0.26122   Max.   :0.2473469
##       Q75               IQR               skew              kurt
##  Min.   :0.04295   Min.   :0.01456   Min.   : 0.1417   Min.   :   2.068
##  1st Qu.:0.20875   1st Qu.:0.04256   1st Qu.: 1.6496   1st Qu.:   5.670
##  Median :0.22568   Median :0.09428   Median : 2.1971   Median :   8.319
##  Mean   :0.22476   Mean   :0.08431   Mean   : 3.1402   Mean   :  36.569
##  3rd Qu.:0.24366   3rd Qu.:0.11418   3rd Qu.: 2.9317   3rd Qu.:  13.649
##  Max.   :0.27347   Max.   :0.25223   Max.   :34.7255   Max.   :1309.613
##      sp.ent            sfm               mode           centroid
##  Min.   :0.7387   Min.   :0.03688   Min.   :0.0000   Min.   :0.03936
```

```
##  1st Qu.:0.8618   1st Qu.:0.25804   1st Qu.:0.1180   1st Qu.:0.16366
##  Median :0.9018   Median :0.39634   Median :0.1866   Median :0.18484
##  Mean   :0.8951   Mean   :0.40822   Mean   :0.1653   Mean   :0.18091
##  3rd Qu.:0.9287   3rd Qu.:0.53368   3rd Qu.:0.2211   3rd Qu.:0.19915
##  Max.   :0.9820   Max.   :0.84294   Max.   :0.2800   Max.   :0.25112
##     meanfun            minfun            maxfun          meandom
##  Min.   :0.05557   Min.   :0.009775   Min.   :0.1031   Min.   :0.007812
##  1st Qu.:0.11700   1st Qu.:0.018223   1st Qu.:0.2540   1st Qu.:0.419828
##  Median :0.14052   Median :0.046110   Median :0.2712   Median :0.765795
##  Mean   :0.14281   Mean   :0.036802   Mean   :0.2588   Mean   :0.829211
##  3rd Qu.:0.16958   3rd Qu.:0.047904   3rd Qu.:0.2775   3rd Qu.:1.177166
##  Max.   :0.23764   Max.   :0.204082   Max.   :0.2791   Max.   :2.957682
##      mindom             maxdom            dfrange          modindx
##  Min.   :0.004883   Min.   : 0.007812   Min.   : 0.000   Min.   :0.00000
##  1st Qu.:0.007812   1st Qu.: 2.070312   1st Qu.: 2.045   1st Qu.:0.09977
##  Median :0.023438   Median : 4.992188   Median : 4.945   Median :0.13936
##  Mean   :0.052647   Mean   : 5.047277   Mean   : 4.995   Mean   :0.17375
##  3rd Qu.:0.070312   3rd Qu.: 7.007812   3rd Qu.: 6.992   3rd Qu.:0.20918
##  Max.   :0.458984   Max.   :21.867188   Max.   :21.844   Max.   :0.93237
##     label
##  female:1584
##  male  :1584
##
##
##
##
```

Distribuição das classes.

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

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

Desvio padrão.

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

```
##     meanfreq           sd       median          Q25          Q75
##   0.02991784   0.01665225   0.03636015   0.04867972   0.02363928
##          IQR         skew         kurt       sp.ent          sfm
##   0.04278305   4.24052871 134.92866124   0.04497952   0.17752111
##         mode     centroid      meanfun      minfun       maxfun
##   0.07720301   0.02991784   0.03230443   0.01921995   0.03007731
##      meandom      mindom       maxdom     dfrange      modindx
##   0.52520503   0.06329948   3.52115661   3.52003912   0.11945439
```

Skew.

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

```
##    meanfreq          sd      median         Q25         Q75         IQR
## -0.61691065  0.13678669 -1.01182579 -0.49041194 -0.89945843  0.29515265
##        skew        kurt      sp.ent         sfm        mode    centroid
##  4.92864347  5.86702644 -0.43052599  0.33963572 -0.83644332 -0.61691065
```

```
##     meanfun      minfun      maxfun     meandom      mindom      maxdom
##  0.03910363  1.87622592 -2.23641539  0.61044394  1.65954109  0.72550141
##     dfrange     modindx
##  0.72757157  2.06238013
```

Correlação.

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

```
##            meanfreq         sd     median        Q25          Q75
## meanfreq  1.0000000 -0.7390388  0.9254454  0.9114163  0.740996718
## sd       -0.7390388  1.0000000 -0.5626026 -0.8469309 -0.161075841
## median    0.9254454 -0.5626026  1.0000000  0.7749216  0.731849232
## Q25       0.9114163 -0.8469309  0.7749216  1.0000000  0.477139811
## Q75       0.7409967 -0.1610758  0.7318492  0.4771398  1.000000000
## IQR      -0.6276051  0.8746603 -0.4773520 -0.8741890  0.009635774
## skew     -0.3223269  0.3145970 -0.2574071 -0.3194753 -0.206338932
## kurt     -0.3160356  0.3462409 -0.2433816 -0.3501824 -0.148880617
## sp.ent   -0.6012025  0.7166200 -0.5020049 -0.6481258 -0.174905239
## sfm      -0.7843323  0.8380865 -0.6616899 -0.7668745 -0.378198373
## mode      0.6877152 -0.5291500  0.6774327  0.5912770  0.486857375
## centroid  1.0000000 -0.7390388  0.9254454  0.9114163  0.740996718
## meanfun   0.4608444 -0.4662815  0.4149093  0.5450351  0.155090956
## minfun    0.3839368 -0.3456089  0.3376019  0.3209943  0.258002476
## maxfun    0.2740041 -0.1296619  0.2513280  0.1998407  0.285583560
## meandom   0.5366661 -0.4827262  0.4559427  0.4674028  0.359180617
## mindom    0.2292610 -0.3576670  0.1911687  0.3022549 -0.023750103
## maxdom    0.5195277 -0.4822778  0.4389190  0.4596832  0.335114045
## dfrange   0.5155699 -0.4759991  0.4356207  0.4543938  0.335647521
## modindx  -0.2169787  0.1226597 -0.2132975 -0.1413774 -0.216474678
##                  IQR        skew        kurt     sp.ent         sfm
## meanfreq -0.627605054 -0.32232693 -0.31603555 -0.6012025 -0.78433231
## sd        0.874660319  0.31459695  0.34624087  0.7166200  0.83808650
## median   -0.477352003 -0.25740709 -0.24338163 -0.5020049 -0.66168990
## Q25      -0.874188990 -0.31947531 -0.35018239 -0.6481258 -0.76687452
## Q75       0.009635774 -0.20633893 -0.14888062 -0.1749052 -0.37819837
## IQR       1.000000000  0.24949748  0.31618474  0.6408132  0.66360146
## skew      0.249497476  1.00000000  0.97702046 -0.1954592  0.07969407
## kurt      0.316184735  0.97702046  1.00000000 -0.1276436  0.10988403
## sp.ent    0.640813242 -0.19545924 -0.12764358  1.0000000  0.86641084
## sfm       0.663601458  0.07969407  0.10988403  0.8664108  1.00000000
## mode     -0.403763599 -0.43485906 -0.40672189 -0.3252985 -0.48591287
## centroid -0.627605054 -0.32232693 -0.31603555 -0.6012025 -0.78433231
## meanfun  -0.534461948 -0.16766801 -0.19455985 -0.5131937 -0.42106568
## minfun   -0.222679719 -0.21695429 -0.20320141 -0.3058260 -0.36210032
## maxfun   -0.069588302 -0.08086107 -0.04566725 -0.1207380 -0.19236944
## meandom  -0.333362476 -0.33684839 -0.30323357 -0.2935624 -0.42844249
## mindom   -0.357036676 -0.06160765 -0.10331264 -0.2948689 -0.28959288
## maxdom   -0.337876663 -0.30565086 -0.27450011 -0.3242531 -0.43664879
## dfrange  -0.331563477 -0.30464003 -0.27272943 -0.3190536 -0.43157977
## modindx   0.041252438 -0.16932471 -0.20553932  0.1980743  0.21147723
##                mode   centroid     meanfun       minfun      maxfun
## meanfreq  0.6877152  1.0000000  0.46084440  0.383936793  0.27400407
## sd       -0.5291500 -0.7390388 -0.46628148 -0.345608905 -0.12966188
```

```
## median    0.6774327  0.9254454  0.41490926  0.337601923  0.25132802
## Q25       0.5912770  0.9114163  0.54503508  0.320994291  0.19984072
## Q75       0.4868574  0.7409967  0.15509096  0.258002476  0.28558356
## IQR      -0.4037636 -0.6276051 -0.53446195 -0.222679719 -0.06958830
## skew     -0.4348591 -0.3223269 -0.16766801 -0.216954285 -0.08086107
## kurt     -0.4067219 -0.3160356 -0.19455985 -0.203201414 -0.04566725
## sp.ent   -0.3252985 -0.6012025 -0.51319368 -0.305826013 -0.12073798
## sfm      -0.4859129 -0.7843323 -0.42106568 -0.362100316 -0.19236944
## mode      1.0000000  0.6877152  0.32477126  0.385467306  0.17232879
## centroid  0.6877152  1.0000000  0.46084440  0.383936793  0.27400407
## meanfun   0.3247713  0.4608444  1.00000000  0.339386726  0.31195050
## minfun    0.3854673  0.3839368  0.33938673  1.000000000  0.21398718
## maxfun    0.1723288  0.2740041  0.31195050  0.213987182  1.00000000
## meandom   0.4914794  0.5366661  0.27083961  0.375979020  0.33755275
## mindom    0.1981496  0.2292610  0.16216251  0.082015330 -0.24342566
## maxdom    0.4771867  0.5195277  0.27798214  0.317860109  0.35539024
## dfrange   0.4737750  0.5155699  0.27515429  0.316486170  0.35988049
## modindx  -0.1823435 -0.2169787 -0.05485794  0.002041973 -0.36302924
##              meandom      mindom      maxdom     dfrange      modindx
## meanfreq  0.53666606  0.22926100  0.51952765  0.51556987 -0.216978748
## sd       -0.48272620 -0.35766702 -0.48227782 -0.47599914  0.122659705
## median    0.45594268  0.19116867  0.43891903  0.43562066 -0.213297510
## Q25       0.46740280  0.30225493  0.45968325  0.45439385 -0.141377375
## Q75       0.35918062 -0.02375010  0.33511405  0.33564752 -0.216474678
## IQR      -0.33336248 -0.35703668 -0.33787666 -0.33156348  0.041252438
## skew     -0.33684839 -0.06160765 -0.30565086 -0.30464003 -0.169324710
## kurt     -0.30323357 -0.10331264 -0.27450011 -0.27272943 -0.205539321
## sp.ent   -0.29356241 -0.29486887 -0.32425314 -0.31905357  0.198074268
## sfm      -0.42844249 -0.28959288 -0.43664879 -0.43157977  0.211477226
## mode      0.49147940  0.19814956  0.47718671  0.47377496 -0.182343536
## centroid  0.53666606  0.22926100  0.51952765  0.51556987 -0.216978748
## meanfun   0.27083961  0.16216251  0.27798214  0.27515429 -0.054857943
## minfun    0.37597902  0.08201533  0.31786011  0.31648617  0.002041973
## maxfun    0.33755275 -0.24342566  0.35539024  0.35988049 -0.363029240
## meandom   1.00000000  0.09965605  0.81283770  0.81130367 -0.180954102
## mindom    0.09965605  1.00000000  0.02663969  0.00866554  0.200212223
## maxdom    0.81283770  0.02663969  1.00000000  0.99983841 -0.425531023
## dfrange   0.81130367  0.00866554  0.99983841  1.00000000 -0.429266452
## modindx  -0.18095410  0.20021222 -0.42553102 -0.42926645  1.000000000
```

Histograma (univariado).

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

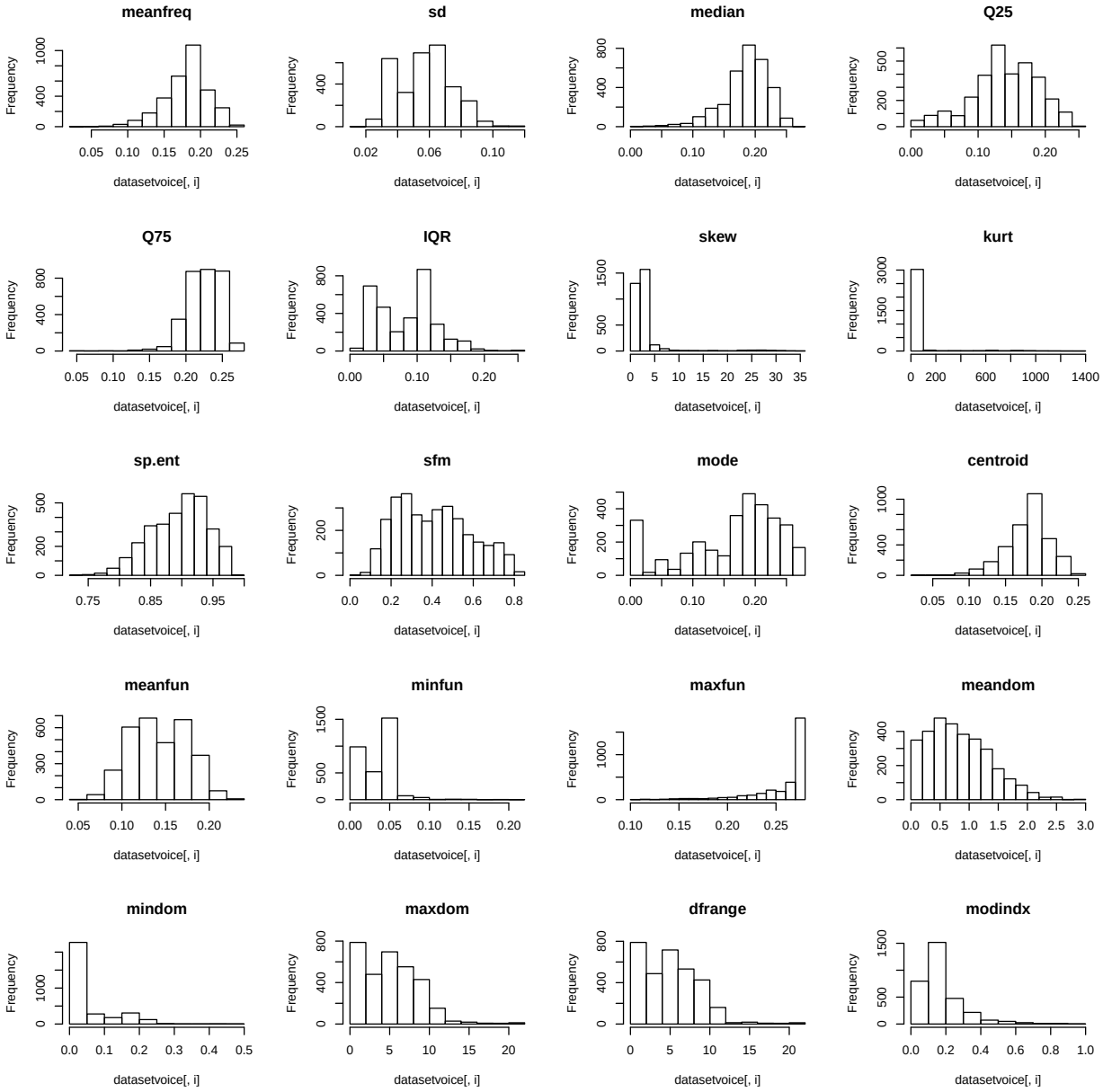
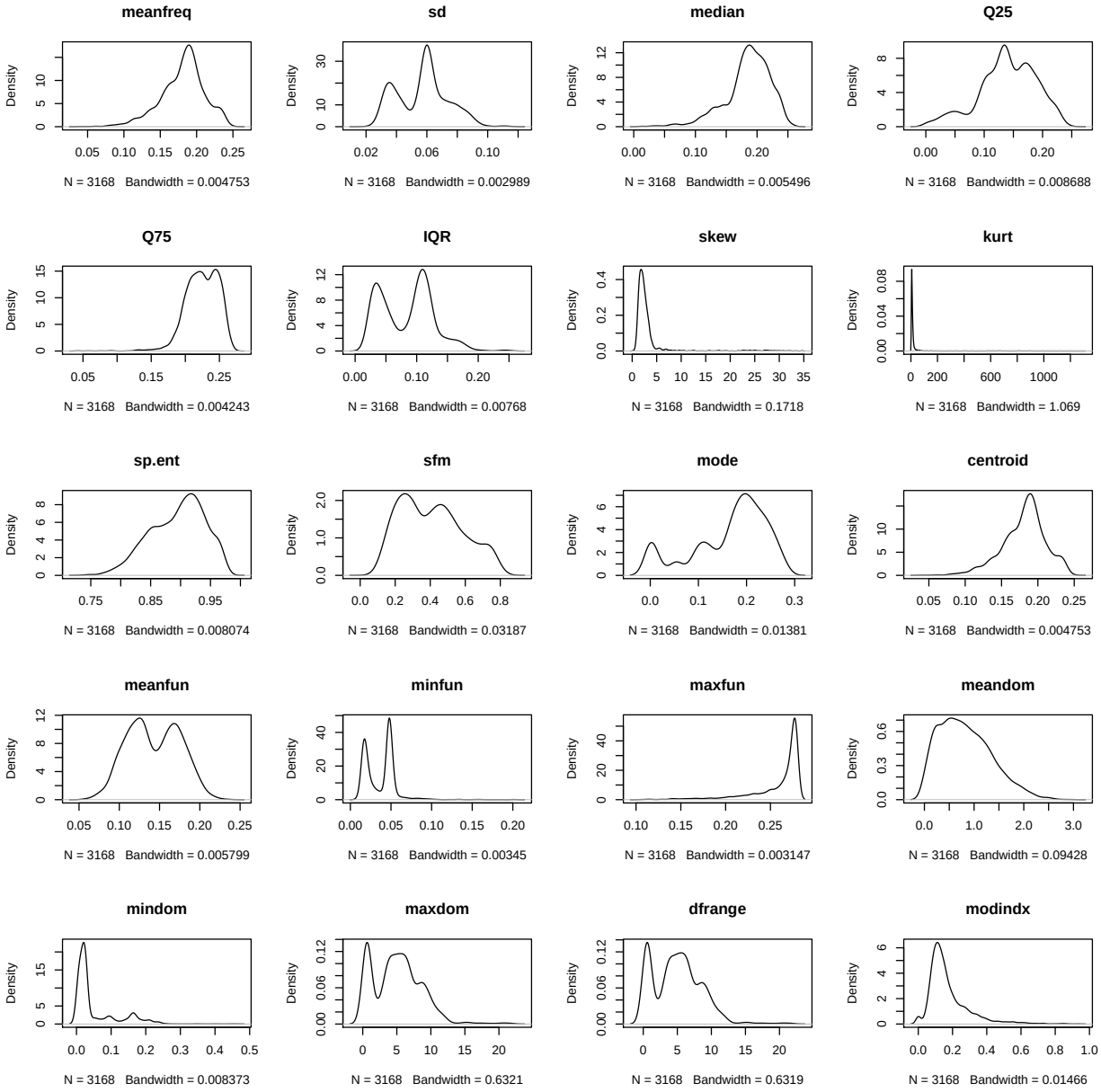


Gráfico de densidade (univariado).

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



Boxplot e Whisker (univariado).

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

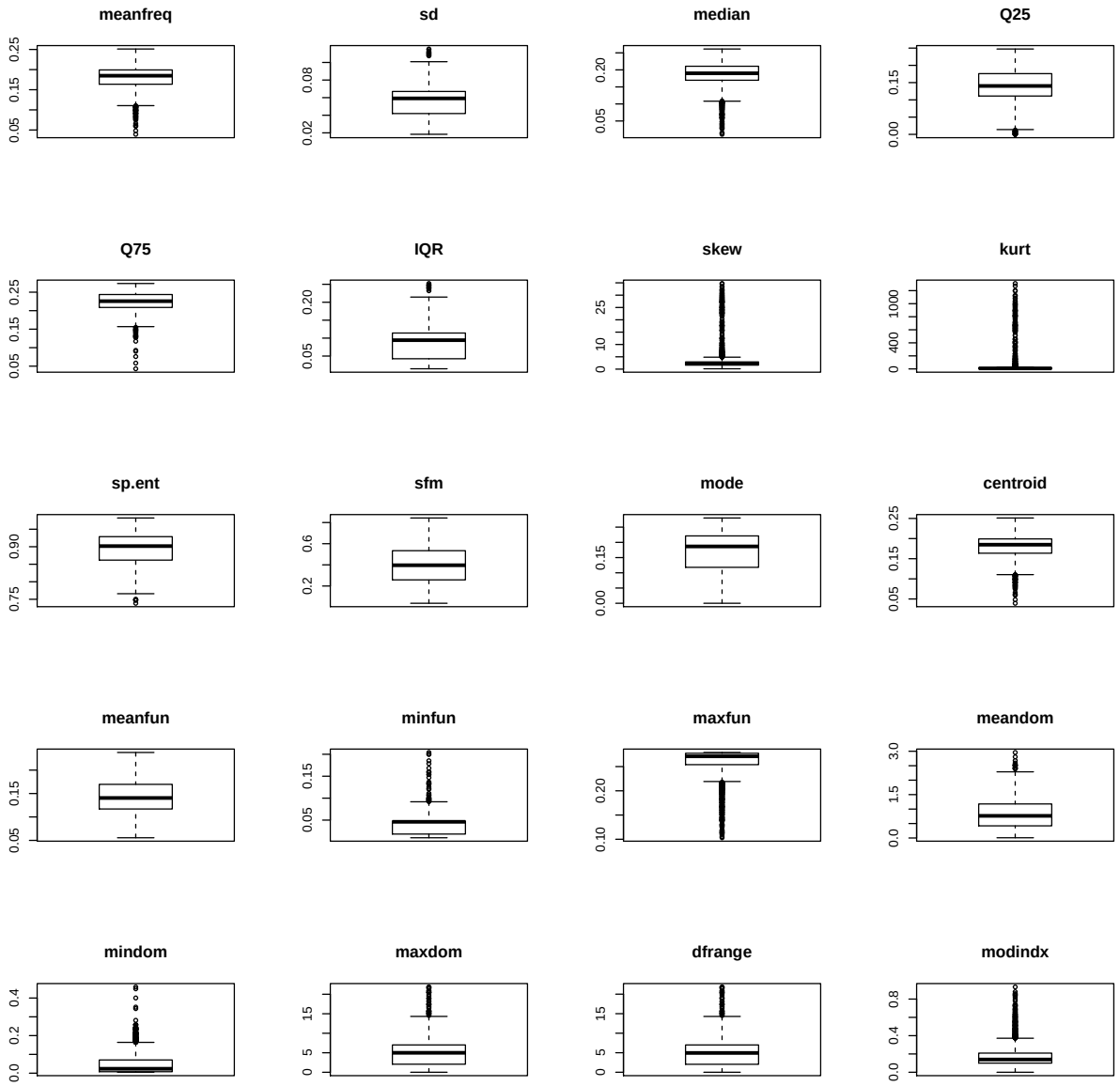
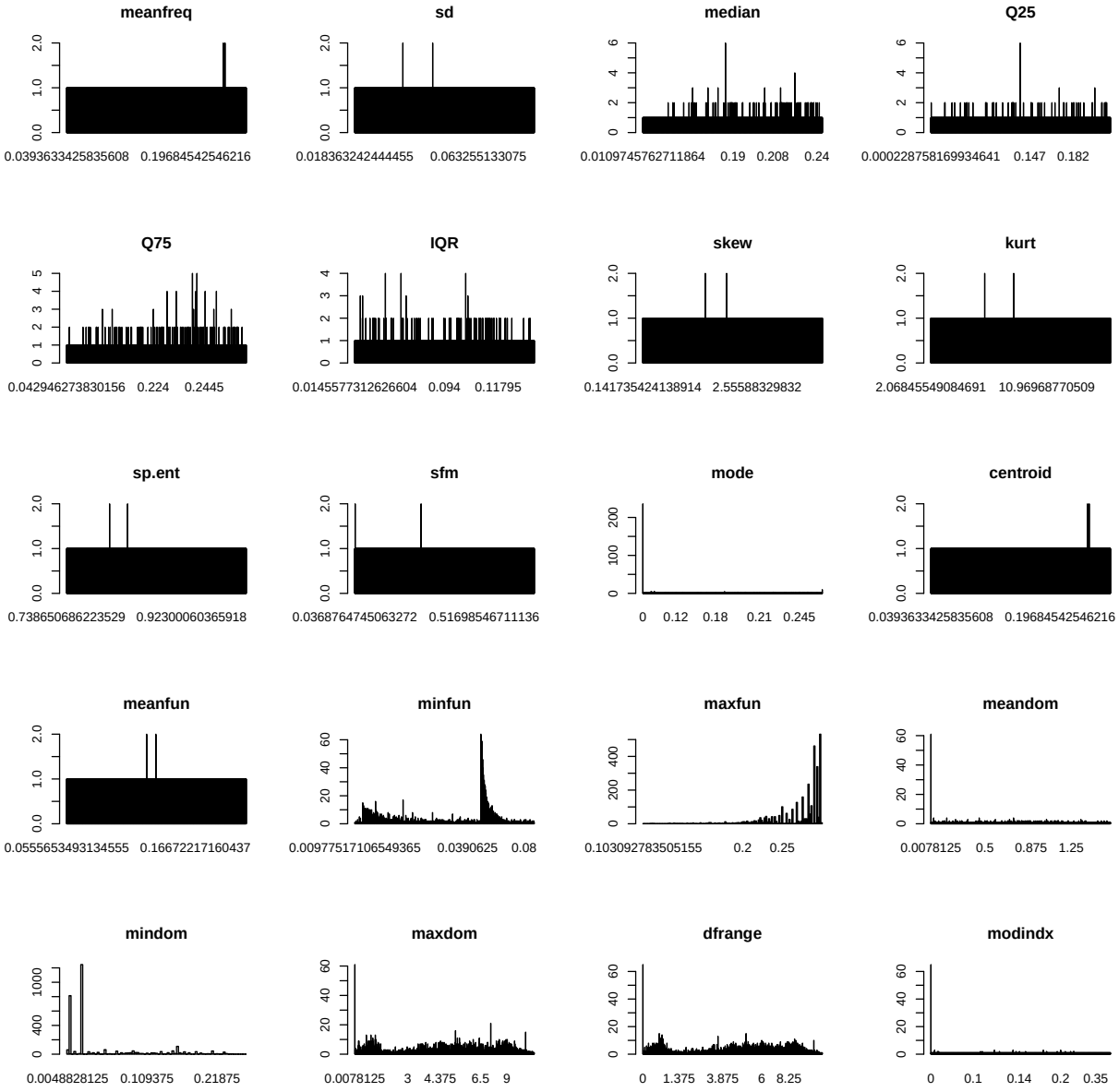


Gráfico de barras.

```
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)

```
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.
```

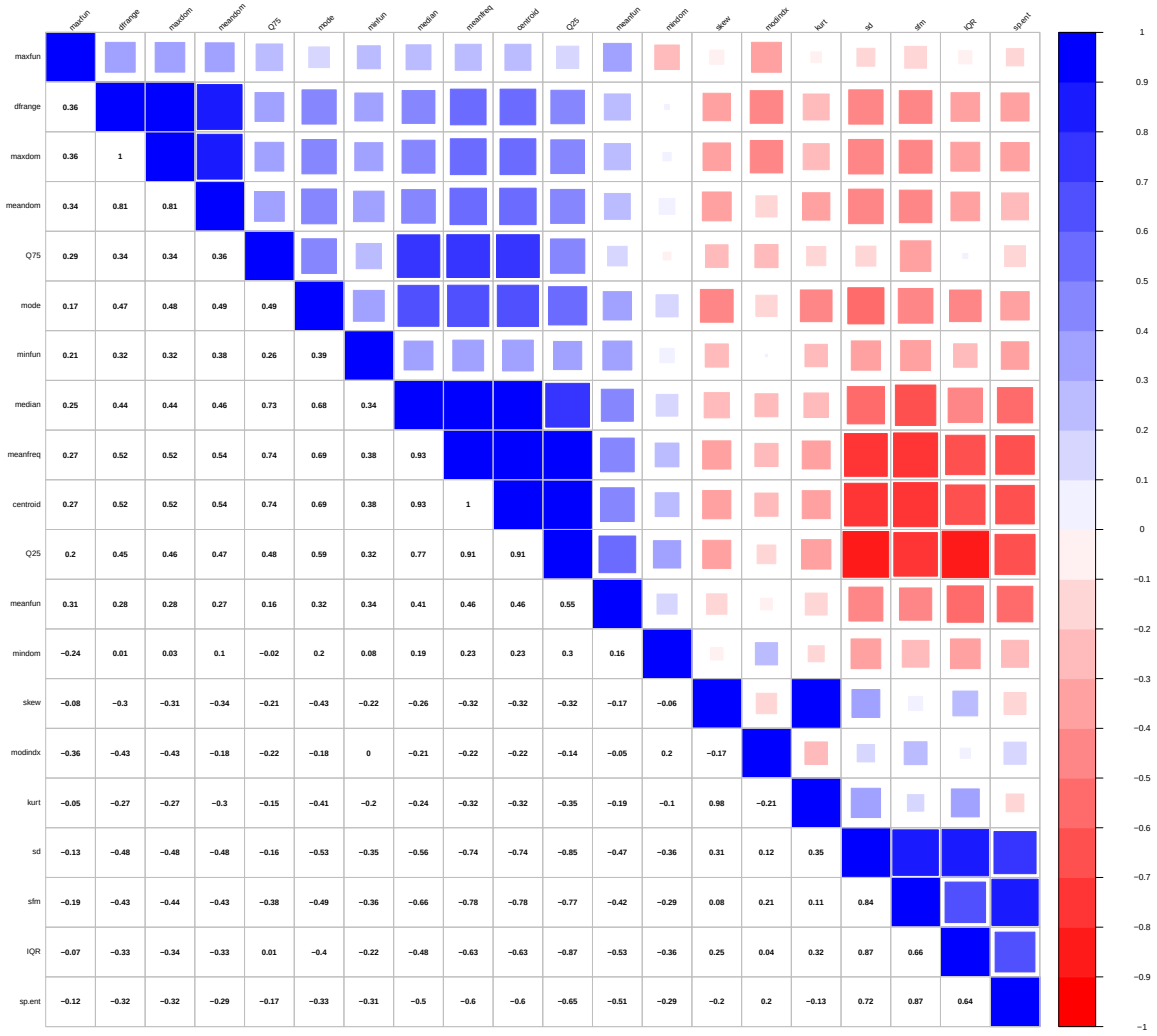
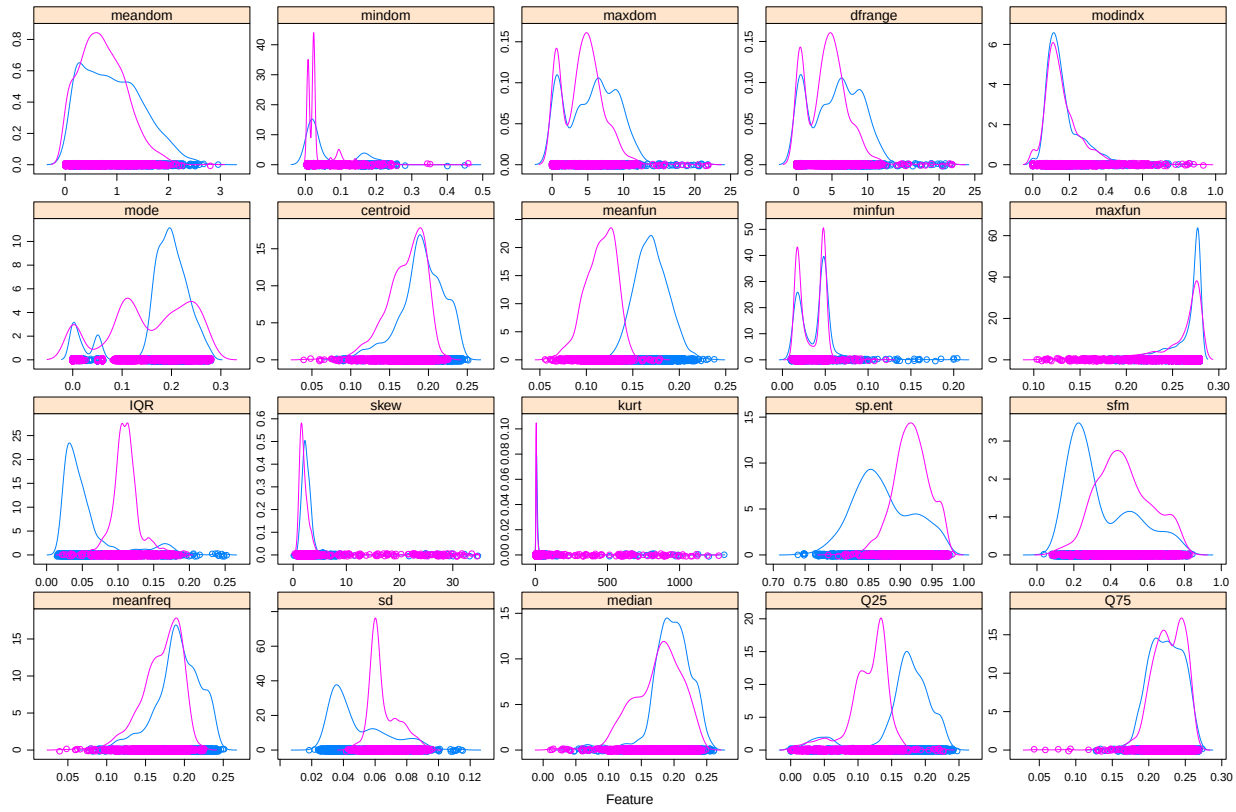


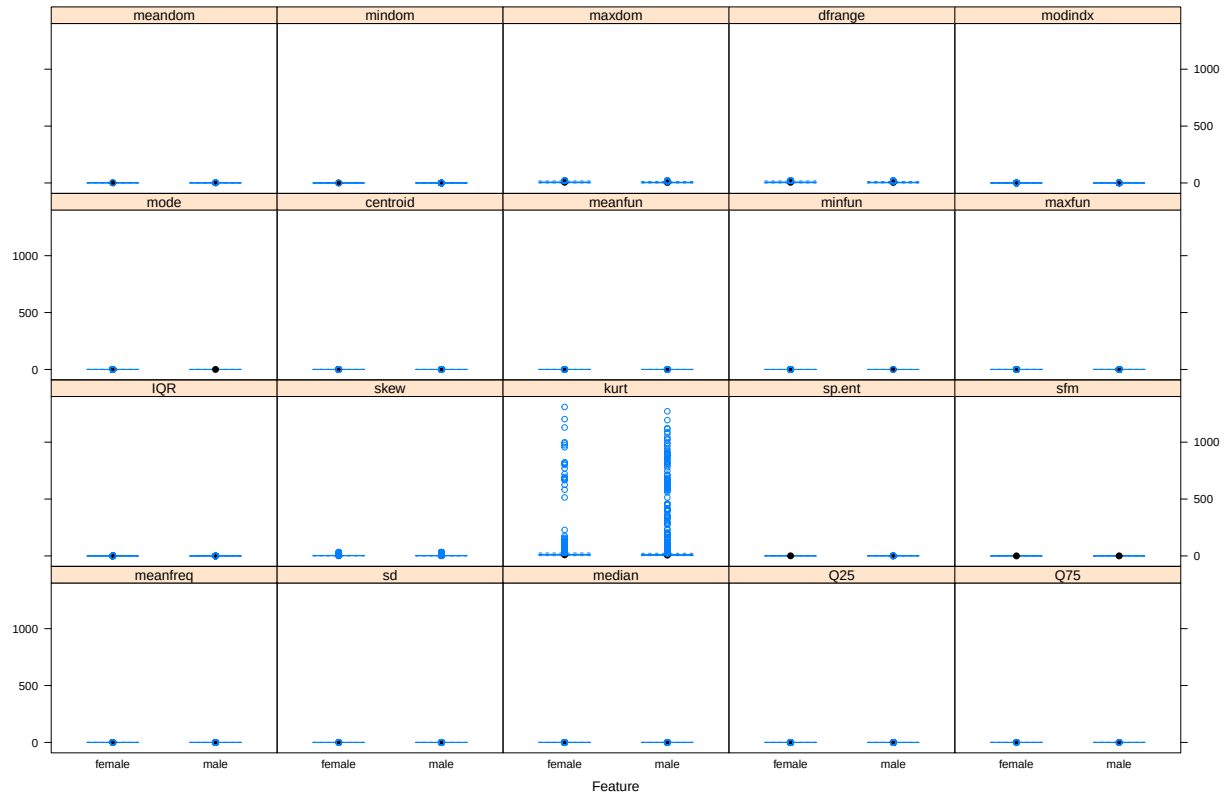
Gráfico de dispersão (multivariado).

```
pairs(datasetvoice)
```



Boxplot por classe (multivariado)

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



Fim da analise