Spotify Data Analysis

Code

Hide

```
dataset=read.csv('data.csv')
names(dataset)
```

```
 [1] "acousticness"     "danceability"     "duration_ms"
 [4] "energy"           "instrumentalness" "key"
 [7] "liveness"         "loudness"         "mode"
[10] "speechiness"      "tempo"            "time_signature"
[13] "valence"          "target"           "song_title"
[16] "artist"
```

Hide

```
dim(dataset)
```

```
[1] 2017   16
```

Hide

```
str(dataset)
```

```
'data.frame':   2017 obs. of  16 variables:
 $ acousticness    : num  0.0102 0.199 0.0344 0.604 0.18 0.00479 0.0145 0.0202 0.0481 0.00208
...
 $ danceability    : num  0.833 0.743 0.838 0.494 0.678 0.804 0.739 0.266 0.603 0.836 ...
 $ duration_ms     : int  204600 326933 185707 199413 392893 251333 241400 349667 202853 226840
...
 $ energy          : num  0.434 0.359 0.412 0.338 0.561 0.56 0.472 0.348 0.944 0.603 ...
 $ instrumentalness: num  2.19e-02 6.11e-03 2.34e-04 5.10e-01 5.12e-01 0.00 7.27e-06 6.64e-01 0.
00 0.00 ...
 $ key             : int  2 1 2 5 5 8 1 10 11 7 ...
 $ liveness        : num  0.165 0.137 0.159 0.0922 0.439 0.164 0.207 0.16 0.342 0.571 ...
 $ loudness        : num  -8.79 -10.4 -7.15 -15.24 -11.65 ...
 $ mode            : int  1 1 1 1 0 1 1 0 0 1 ...
 $ speechiness     : num  0.431 0.0794 0.289 0.0261 0.0694 0.185 0.156 0.0371 0.347 0.237 ...
 $ tempo           : num  150.1 160.1 75 86.5 174 ...
 $ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
 $ valence         : num  0.286 0.588 0.173 0.23 0.904 0.264 0.308 0.393 0.398 0.386 ...
 $ target          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ song_title      : Factor w/ 1956 levels "'Till I Collapse",..: 1053 1346 1917 1054 1254 1486
319 667 783 409 ...
 $ artist          : Factor w/ 1343 levels "!!!","*NSYNC",..: 455 221 455 97 636 360 360 877 313
521 ...
```

Hide

```
summary(dataset)
```

```
  acousticness         danceability      duration_ms     
 Min.   :0.0000028   Min.   :0.1220   Min.   :  16042  
 1st Qu.:0.0096300   1st Qu.:0.5140   1st Qu.: 200015  
 Median :0.0633000   Median :0.6310   Median : 229261  
 Mean   :0.1875900   Mean   :0.6184   Mean   : 246306  
 3rd Qu.:0.2650000   3rd Qu.:0.7380   3rd Qu.: 270333  
 Max.   :0.9950000   Max.   :0.9840   Max.   :1004627  
                                                         
     energy       instrumentalness         key        
 Min.   :0.0148   Min.   :0.0000000   Min.   : 0.000  
 1st Qu.:0.5630   1st Qu.:0.0000000   1st Qu.: 2.000  
 Median :0.7150   Median :0.0000762   Median : 6.000  
 Mean   :0.6816   Mean   :0.1332855   Mean   : 5.343  
 3rd Qu.:0.8460   3rd Qu.:0.0540000   3rd Qu.: 9.000  
 Max.   :0.9980   Max.   :0.9760000   Max.   :11.000  
                                                        
    liveness         loudness            mode       
 Min.   :0.0188   Min.   :-33.097   Min.   :0.0000  
 1st Qu.:0.0923   1st Qu.: -8.394   1st Qu.:0.0000  
 Median :0.1270   Median : -6.248   Median :1.0000  
 Mean   :0.1908   Mean   : -7.086   Mean   :0.6123  
 3rd Qu.:0.2470   3rd Qu.: -4.746   3rd Qu.:1.0000  
 Max.   :0.9690   Max.   : -0.307   Max.   :1.0000  
                                                      
  speechiness          tempo        time_signature 
 Min.   :0.02310   Min.   : 47.86   Min.   :1.000  
 1st Qu.:0.03750   1st Qu.:100.19   1st Qu.:4.000  
 Median :0.05490   Median :121.43   Median :4.000  
 Mean   :0.09266   Mean   :121.60   Mean   :3.968  
 3rd Qu.:0.10800   3rd Qu.:137.85   3rd Qu.:4.000  
 Max.   :0.81600   Max.   :219.33   Max.   :5.000  
                                                     
    valence           target                 song_title  
 Min.   :0.0348   Min.   :0.0000   Jack            :   3  
 1st Qu.:0.2950   1st Qu.:0.0000   River           :   3  
 Median :0.4920   Median :1.0000   1-800-273-8255:   2  
 Mean   :0.4968   Mean   :0.5057   Acamar          :   2  
 3rd Qu.:0.6910   3rd Qu.:1.0000   Alright         :   2  
 Max.   :0.9920   Max.   :1.0000   Annie           :   2  
                                   (Other)         :2003  
             artist    
 Drake           :  16  
 Rick Ross       :  13  
 Disclosure      :  12  
 Backstreet Boys:  10  
 WALK THE MOON  :  10  
 Crystal Castles:   9  
 (Other)        :1947  
```

Hide

```
sapply(dataset, function(x) sum(is.na(x)))
```

```
    acousticness     danceability      duration_ms 
               0                0                0 
          energy instrumentalness              key 
               0                0                0 
        liveness         loudness             mode 
               0                0                0 
     speechiness            tempo   time_signature 
               0                0                0 
         valence           target       song_title 
               0                0                0 
          artist 
               0 
```

Hide

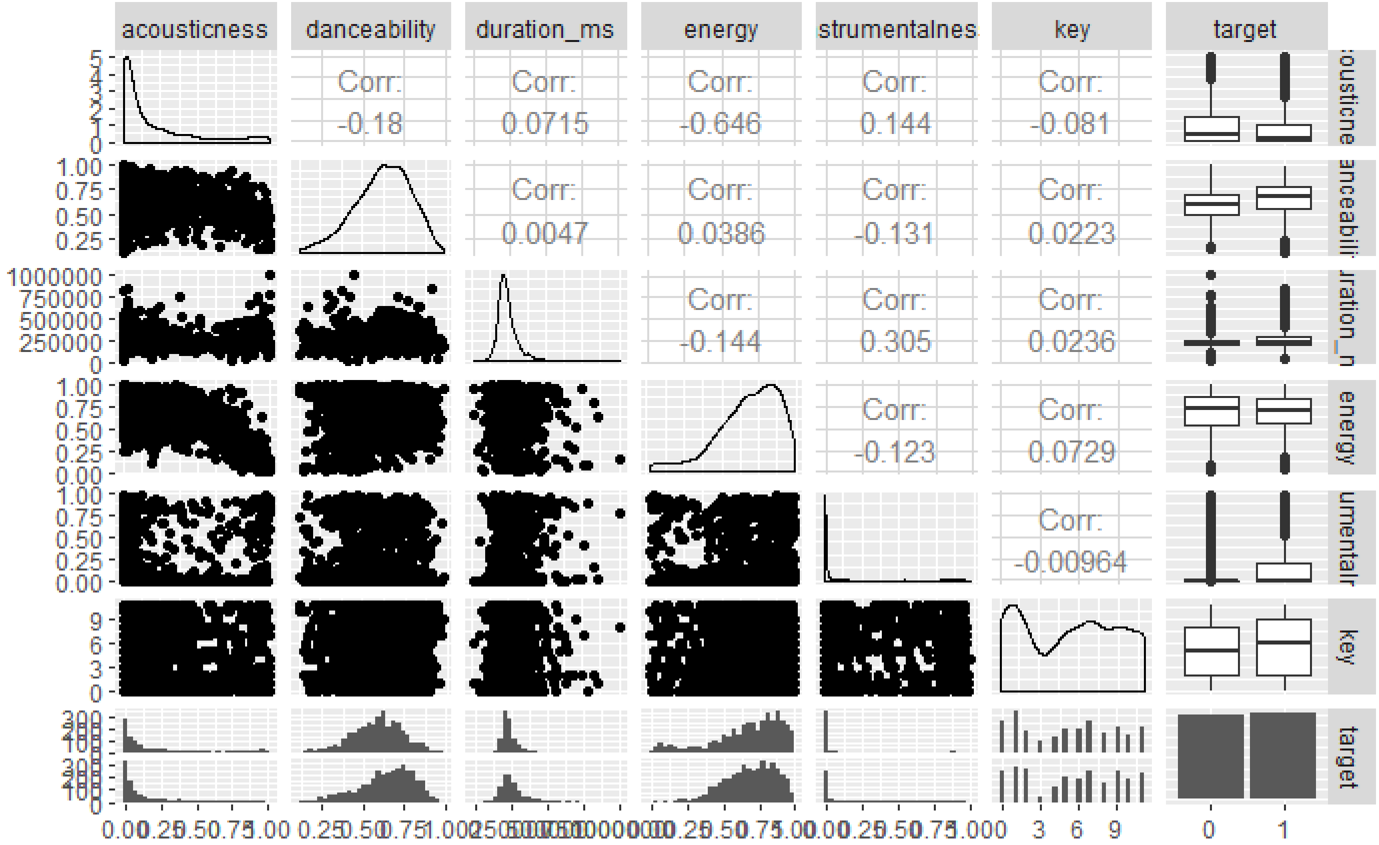
```
table(is.na(dataset))
```

```


FALSE 
32272 
```

Hide

```
dataset$target<-as.factor(dataset$target)
library(GGally)
ggpairs(dataset,columns = c(1:6,14))
```

```
plot: [1,1] [=------------------------------------]  2% est: 0s
plot: [1,2] [==-----------------------------------]  4% est: 2s
plot: [1,3] [==-----------------------------------]  6% est: 2s
plot: [1,4] [===----------------------------------]  8% est: 2s
plot: [1,5] [====---------------------------------] 10% est: 2s
plot: [1,6] [=====--------------------------------] 12% est: 2s
plot: [1,7] [=====--------------------------------] 14% est: 2s
plot: [2,1] [======-------------------------------] 16% est: 2s
plot: [2,2] [=======------------------------------] 18% est: 2s
plot: [2,3] [========-----------------------------] 20% est: 2s
plot: [2,4] [========-----------------------------] 22% est: 2s
plot: [2,5] [=========----------------------------] 24% est: 2s
plot: [2,6] [==========---------------------------] 27% est: 2s
plot: [2,7] [===========--------------------------] 29% est: 2s
plot: [3,1] [===========--------------------------] 31% est: 2s
plot: [3,2] [============-------------------------] 33% est: 2s
plot: [3,3] [=============------------------------] 35% est: 2s
plot: [3,4] [==============-----------------------] 37% est: 2s
plot: [3,5] [==============-----------------------] 39% est: 2s
plot: [3,6] [===============----------------------] 41% est: 2s
plot: [3,7] [================---------------------] 43% est: 2s
plot: [4,1] [=================--------------------] 45% est: 2s
plot: [4,2] [=================--------------------] 47% est: 2s
plot: [4,3] [==================-------------------] 49% est: 2s
plot: [4,4] [===================------------------] 51% est: 1s
plot: [4,5] [====================-----------------] 53% est: 1s
plot: [4,6] [====================-----------------] 55% est: 1s
plot: [4,7] [=====================----------------] 57% est: 1s
plot: [5,1] [======================---------------] 59% est: 1s
plot: [5,2] [=======================--------------] 61% est: 1s
plot: [5,3] [=======================--------------] 63% est: 1s
plot: [5,4] [========================-------------] 65% est: 1s
plot: [5,5] [=========================------------] 67% est: 1s
plot: [5,6] [==========================-----------] 69% est: 1s
plot: [5,7] [==========================-----------] 71% est: 1s
plot: [6,1] [===========================----------] 73% est: 1s
plot: [6,2] [============================---------] 76% est: 1s
plot: [6,3] [=============================--------] 78% est: 1s
plot: [6,4] [=============================--------] 80% est: 1s
plot: [6,5] [==============================-------] 82% est: 1s
plot: [6,6] [===============================------] 84% est: 1s
plot: [6,7] [================================-----] 86% est: 0s
plot: [7,1] [================================-----] 88% est: 0s
plot: [7,2] [=================================----] 90% est: 0s
plot: [7,3] [==================================---] 92% est: 0s
plot: [7,4] [===================================--] 94% est: 0s
plot: [7,5] [===================================--] 96% est: 0s
plot: [7,6] [====================================-] 98% est: 0s
plot: [7,7] [=====================================]100% est: 0s
```

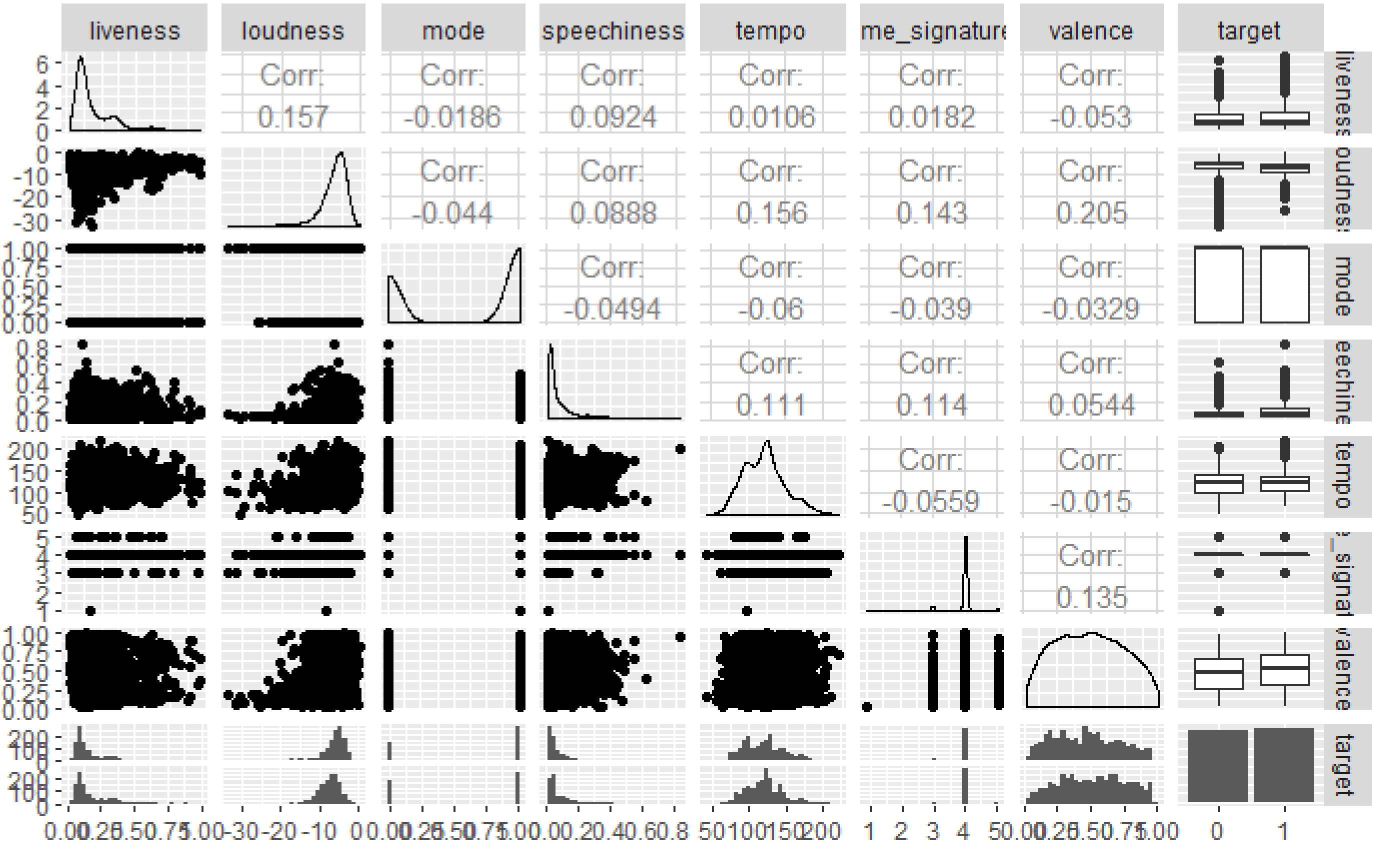


Hide

```
ggpairs(dataset,columns = c(7:13,14))
```

```
plot: [1,1] [=------------------------------------]  2% est: 0s
plot: [1,2] [=------------------------------------]  3% est: 2s
plot: [1,3] [==-----------------------------------]  5% est: 3s
plot: [1,4] [==-----------------------------------]  6% est: 3s
plot: [1,5] [===----------------------------------]  8% est: 3s
plot: [1,6] [===----------------------------------]  9% est: 3s
plot: [1,7] [====---------------------------------] 11% est: 3s
plot: [1,8] [=====--------------------------------] 12% est: 3s
plot: [2,1] [=====--------------------------------] 14% est: 3s
plot: [2,2] [======-------------------------------] 16% est: 3s
plot: [2,3] [======-------------------------------] 17% est: 3s
plot: [2,4] [=======------------------------------] 19% est: 3s
plot: [2,5] [========-----------------------------] 20% est: 3s
plot: [2,6] [========-----------------------------] 22% est: 3s
plot: [2,7] [=========----------------------------] 23% est: 3s
plot: [2,8] [=========----------------------------] 25% est: 3s
plot: [3,1] [==========---------------------------] 27% est: 3s
plot: [3,2] [==========---------------------------] 28% est: 3s
plot: [3,3] [===========--------------------------] 30% est: 3s
plot: [3,4] [============-------------------------] 31% est: 3s
plot: [3,5] [============-------------------------] 33% est: 3s
plot: [3,6] [=============------------------------] 34% est: 3s
plot: [3,7] [=============------------------------] 36% est: 3s
plot: [3,8] [==============-----------------------] 38% est: 2s
plot: [4,1] [==============-----------------------] 39% est: 2s
plot: [4,2] [===============----------------------] 41% est: 2s
plot: [4,3] [================---------------------] 42% est: 2s
plot: [4,4] [================---------------------] 44% est: 2s
plot: [4,5] [=================--------------------] 45% est: 2s
plot: [4,6] [=================--------------------] 47% est: 2s
plot: [4,7] [==================-------------------] 48% est: 2s
plot: [4,8] [==================-------------------] 50% est: 2s
plot: [5,1] [===================------------------] 52% est: 2s
plot: [5,2] [====================-----------------] 53% est: 2s
plot: [5,3] [====================-----------------] 55% est: 2s
plot: [5,4] [=====================----------------] 56% est: 2s
plot: [5,5] [=====================----------------] 58% est: 2s
plot: [5,6] [======================---------------] 59% est: 2s
plot: [5,7] [=======================--------------] 61% est: 2s
plot: [5,8] [=======================--------------] 62% est: 1s
plot: [6,1] [========================-------------] 64% est: 2s
plot: [6,2] [========================-------------] 66% est: 2s
plot: [6,3] [=========================------------] 67% est: 1s
plot: [6,4] [=========================------------] 69% est: 1s
plot: [6,5] [==========================-----------] 70% est: 1s
plot: [6,6] [===========================----------] 72% est: 1s
plot: [6,7] [===========================----------] 73% est: 1s
plot: [6,8] [============================---------] 75% est: 1s
plot: [7,1] [============================---------] 77% est: 1s
plot: [7,2] [=============================--------] 78% est: 1s
plot: [7,3] [=============================--------] 80% est: 1s
plot: [7,4] [==============================-------] 81% est: 1s
```

```
plot: [7,5] [===============================------] 83% est: 1s
plot: [7,6] [===============================------] 84% est: 1s
plot: [7,7] [================================-----] 86% est: 1s
plot: [7,8] [================================-----] 88% est: 1s
plot: [8,1] [=================================----] 89% est: 0s
plot: [8,2] [==================================---] 91% est: 0s
plot: [8,3] [==================================---] 92% est: 0s
plot: [8,4] [===================================--] 94% est: 0s
plot: [8,5] [===================================--] 95% est: 0s
plot: [8,6] [====================================-] 97% est: 0s
plot: [8,7] [====================================-] 98% est: 0s
plot: [8,8] [=====================================]100% est: 0s
```



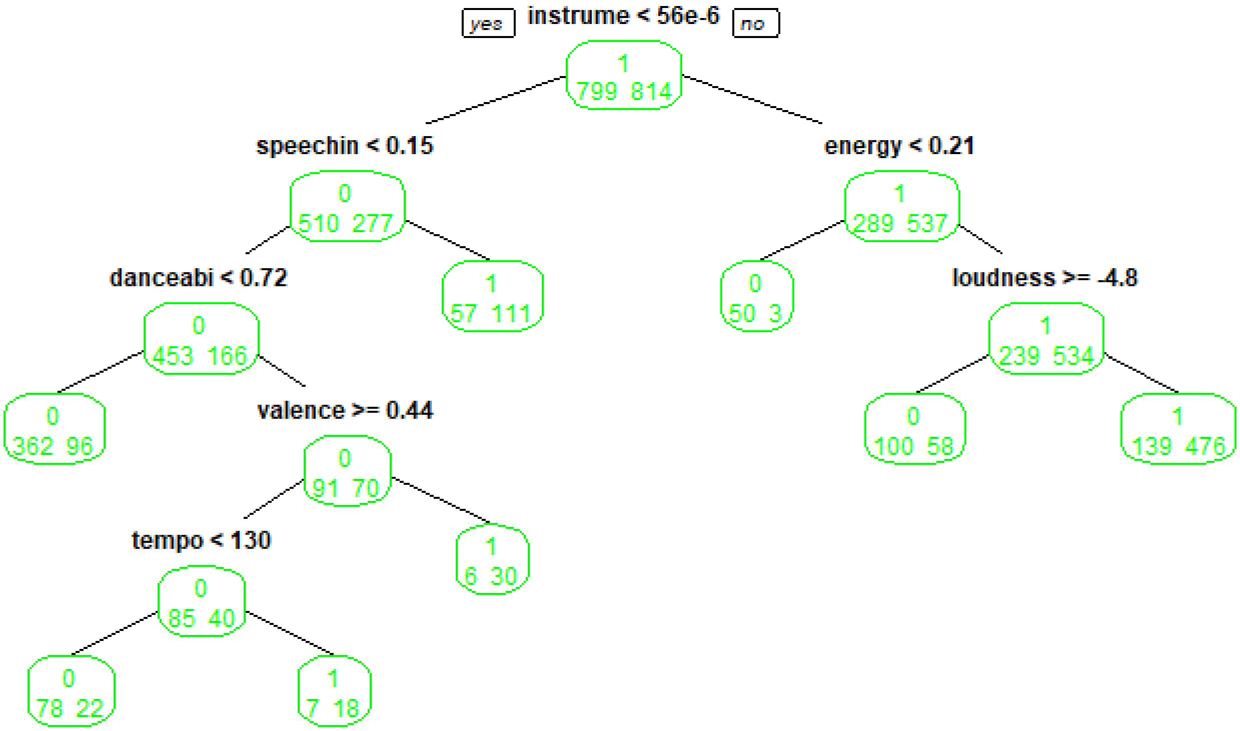
Hide

```
dt<-sort(sample(nrow(dataset),nrow(dataset)*.8))
train<-dataset[dt,]
test<-dataset[-dt,]
library(rpart)
library(rpart.plot)
library(caret)
str(train)
```

```
'data.frame':	1613 obs. of  14 variables:
 $ acousticness    : num  0.0102 0.199 0.0344 0.604 0.18 0.00479 0.0145 0.0202 0.0481 0.00208
...
 $ danceability    : num  0.833 0.743 0.838 0.494 0.678 0.804 0.739 0.266 0.603 0.836 ...
 $ duration_ms     : int  204600 326933 185707 199413 392893 251333 241400 349667 202853 226840
...
 $ energy          : num  0.434 0.359 0.412 0.338 0.561 0.56 0.472 0.348 0.944 0.603 ...
 $ instrumentalness: num  2.19e-02 6.11e-03 2.34e-04 5.10e-01 5.12e-01 0.00 7.27e-06 6.64e-01 0.
00 0.00 ...
 $ key             : int  2 1 2 5 5 8 1 10 11 7 ...
 $ liveness        : num  0.165 0.137 0.159 0.0922 0.439 0.164 0.207 0.16 0.342 0.571 ...
 $ loudness        : num  -8.79 -10.4 -7.15 -15.24 -11.65 ...
 $ mode            : int  1 1 1 1 0 1 1 0 0 1 ...
 $ speechiness     : num  0.431 0.0794 0.289 0.0261 0.0694 0.185 0.156 0.0371 0.347 0.237 ...
 $ tempo           : num  150.1 160.1 75 86.5 174 ...
 $ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
 $ valence         : num  0.286 0.588 0.173 0.23 0.904 0.264 0.308 0.393 0.398 0.386 ...
 $ target          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

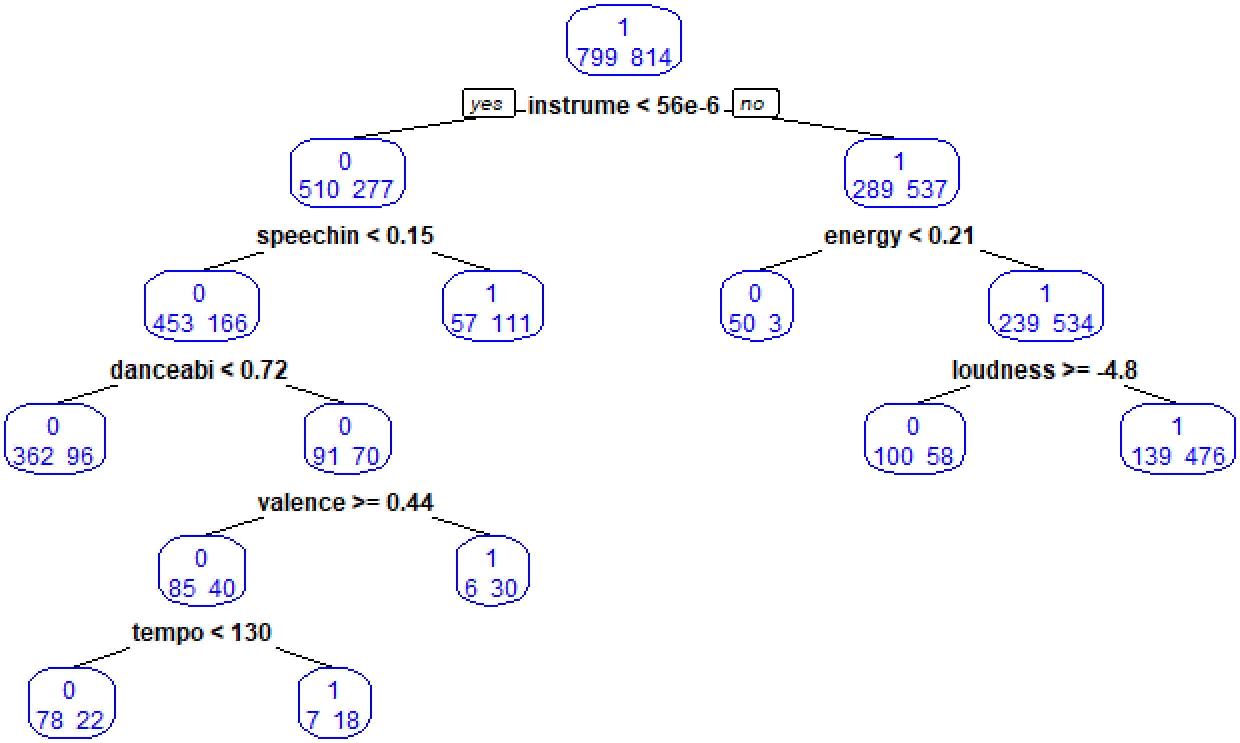
Hide

```
model <- rpart(target~.,data=train)
prp(model, type=1, extra=1, col="green")
```



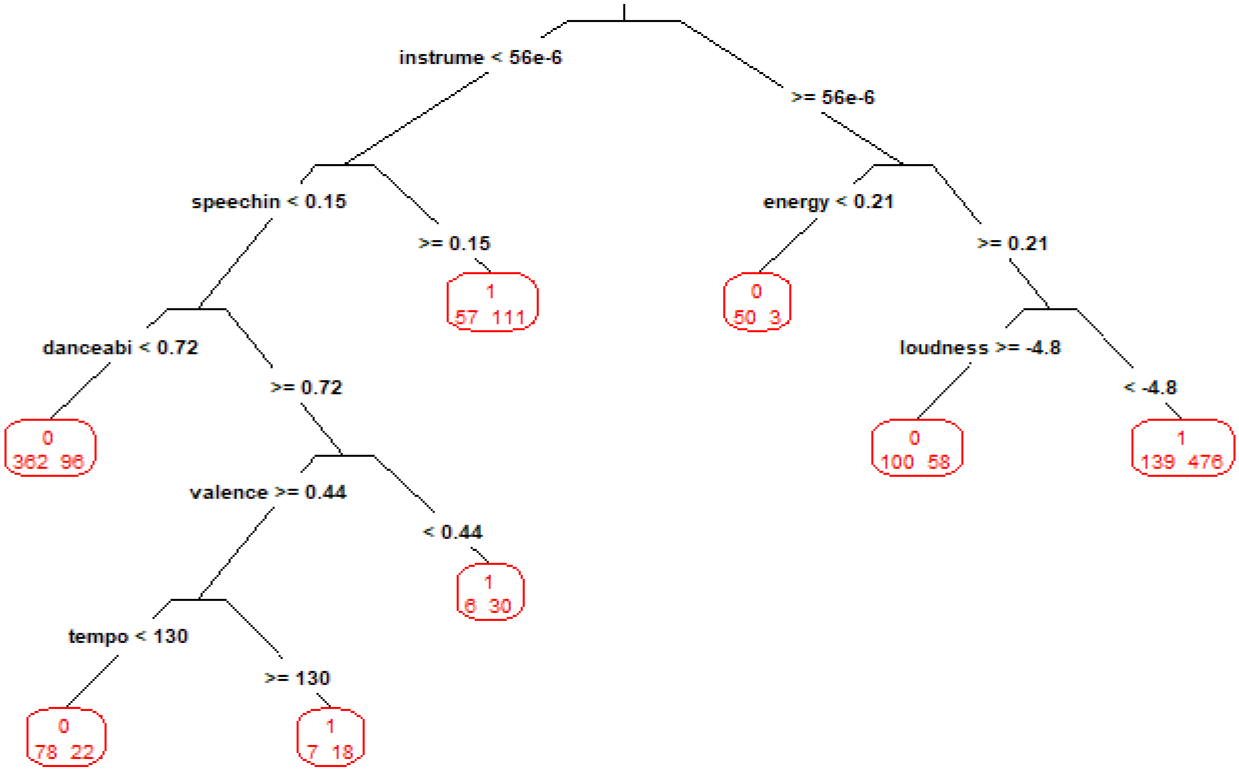
Hide

```
prp(model, type=2, extra=1, col="blue")
```



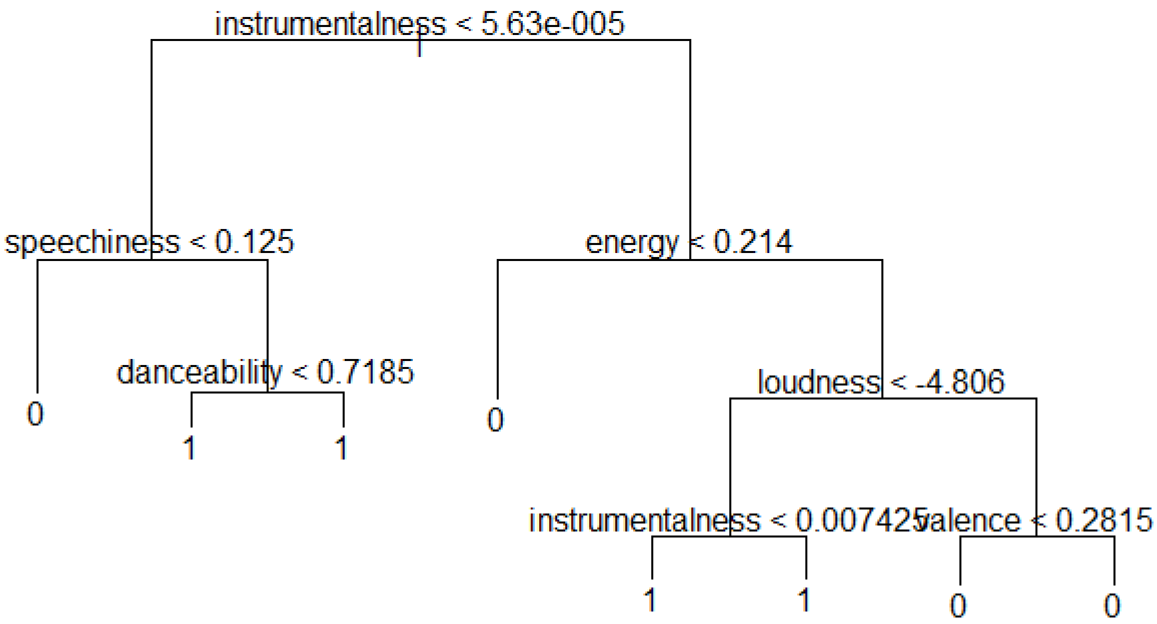
Hide

```
prp(model, type=3, extra=1, col="red")
```



Hide

```
library(tree)
dataset$X<-NULL
dataset$song_title<-NULL
dataset$artist<-NULL
model1<-tree(target~.,train)
plot(model1)
text(model1,pretty=0)
```



Hide

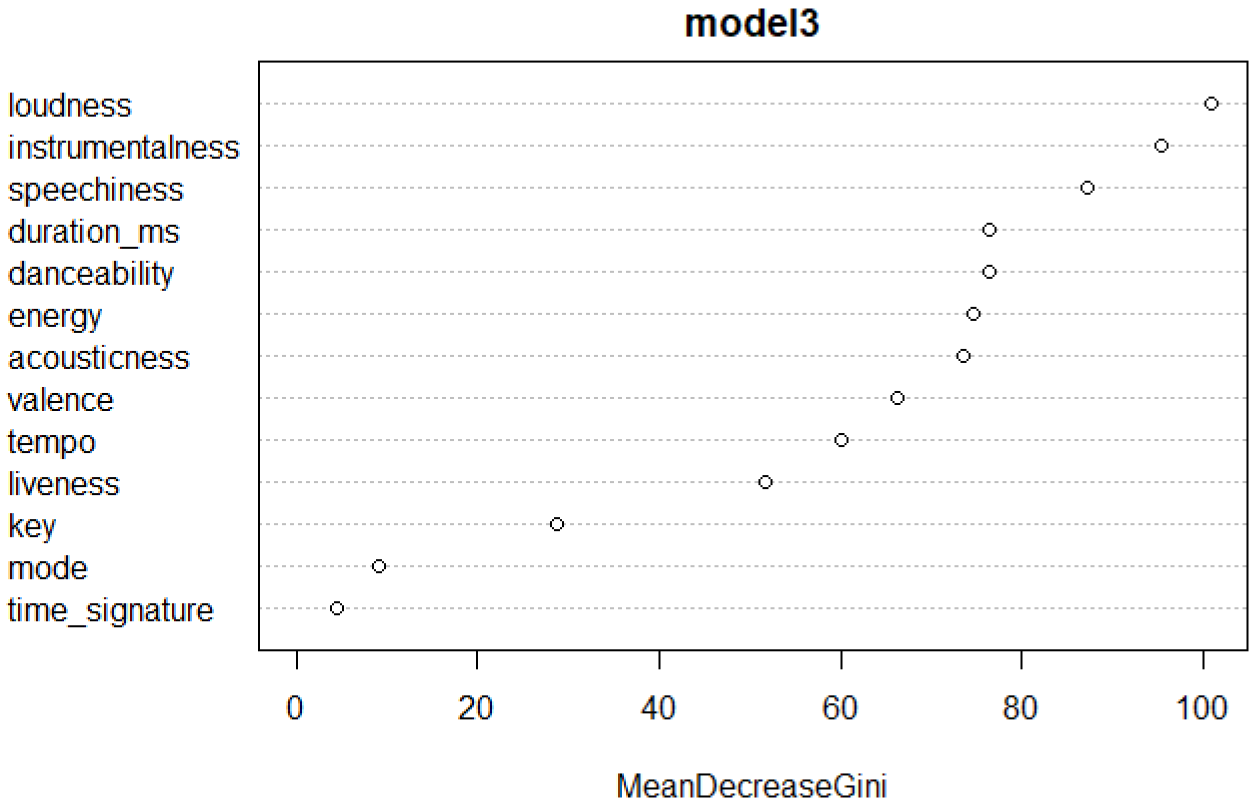
```
str(test)
```

```
'data.frame':	404 obs. of  14 variables:
 $ acousticness    : num  0.604 0.0481 0.019 0.0239 0.00219 0.0516 0.0219 0.297 0.0565 0.00356
...
 $ danceability    : num  0.494 0.603 0.637 0.603 0.781 0.782 0.897 0.722 0.853 0.76 ...
 $ duration_ms     : int  199413 202853 188333 270827 205160 228562 285240 175613 205879 186122
...
 $ energy          : num  0.338 0.944 0.832 0.955 0.795 0.572 0.642 0.823 0.547 0.402 ...
 $ instrumentalness: num  5.10e-01 0.00 5.63e-02 4.51e-02 2.69e-01 0.00 1.31e-06 0.00 0.00 0.00
...
 $ key             : int  5 11 6 1 7 4 2 7 1 8 ...
 $ liveness        : num  0.0922 0.342 0.316 0.119 0.0673 0.33 0.159 0.489 0.341 0.333 ...
 $ loudness        : num  -15.24 -3.63 -6.64 -4.11 -6.76 ...
 $ mode            : int  1 0 1 1 1 0 1 1 1 1 ...
 $ speechiness     : num  0.0261 0.347 0.163 0.0458 0.036 0.0385 0.0534 0.081 0.194 0.164 ...
 $ tempo           : num  86.5 130 100 123.9 110 ...
 $ time_signature  : int  4 4 4 4 4 4 4 4 4 4 ...
 $ valence         : num  0.23 0.398 0.317 0.773 0.795 0.237 0.27 0.855 0.677 0.069 ...
 $ target          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Hide

```
pred <- predict(model, test, type="class")
confusionMatrix(pred, test$target)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 145  43
         1  62 154

               Accuracy : 0.7401
                 95% CI : (0.6944, 0.7822)
    No Information Rate : 0.5124
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.4811
 Mcnemar's Test P-Value : 0.07898

            Sensitivity : 0.7005
            Specificity : 0.7817
         Pos Pred Value : 0.7713
         Neg Pred Value : 0.7130
             Prevalence : 0.5124
         Detection Rate : 0.3589
   Detection Prevalence : 0.4653
      Balanced Accuracy : 0.7411

       'Positive' Class : 0

```


model3
loudness
instrumentalness
speechiness
duration_ms
danceability
energy
acousticness
valence
tempo
liveness
key
mode
time_signature
0
20
40
60
80
100
MeanDecreaseGini