Spotify Dataset Clustering - Profiling Analysis

2023-01-05

# Spotify Dataset Clustering - Profiling Analysis

This report is a continuation of the clustering.Rmd process. The SpotifyClustersData.csv was sourced from that dataset.

# Replace with your own working directory if needed  
WD <- "C:/Users/Joseph/Documents/Codes/2022/mvtec-2022/finalproject/spotify-statistics-final/clustering"  
setwd(WD)  
dd <- read.table("data/SpotifyClustersData.csv",header=T, sep=",", dec='.');  
head(dd, 10)

## loudness energy acousticness instrumentalness valence cluster  
## 1 -5.274 0.854 0.049800 2.53e-04 0.699 1  
## 2 -3.464 0.855 0.003410 4.28e-02 0.609 1  
## 3 -5.308 0.617 0.171000 0.00e+00 0.759 1  
## 4 -13.106 0.235 0.917000 2.64e-01 0.382 2  
## 5 -10.344 0.714 0.000726 9.08e-01 0.269 3  
## 6 -6.901 0.771 0.001190 2.61e-02 0.743 1  
## 7 -2.976 0.992 0.004000 2.73e-05 0.200 1  
## 8 -13.509 0.573 0.906000 8.60e-01 0.650 4  
## 9 -9.089 0.445 0.990000 5.38e-01 0.192 4  
## 10 -7.280 0.397 0.556000 0.00e+00 0.307 1

As you may notice, the new cluster variable was generated from dissimilarity Matrix / Gower’s Distance from previous script.

names(dd)

## [1] "loudness" "energy" "acousticness" "instrumentalness"  
## [5] "valence" "cluster"

attach(dd)

Calculate the test value of variable Xnum for all modalities of factor P

ValorTestXnum <- function(Xnum,P){  
 # Freq dis of factors  
 nk <- as.vector(table(P));   
 n <- sum(nk);  
   
 # Averages vs groups  
 xk <- tapply(Xnum,P,mean);  
   
 # Test values  
 txk <- (xk-mean(Xnum))/(sd(Xnum)\*sqrt((n-nk)/(n\*nk)));  
   
 # P-values  
 pxk <- pt(txk,n-1,lower.tail=F);  
 for(c in 1:length(levels(as.factor(P)))){if (pxk[c]>0.5){pxk[c]<-1-pxk[c]}}  
 return (pxk)  
}  
  
ValorTestXquali <- function(P,Xquali){  
 taula <- table(P,Xquali);  
 n <- sum(taula);   
 pk <- apply(taula,1,sum)/n;  
 pj <- apply(taula,2,sum)/n;  
 pf <- taula/(n\*pk);  
 pjm <- matrix(data=pj,nrow=dim(pf)[1],ncol=dim(pf)[2], byrow=TRUE);   
 dpf <- pf - pjm;   
 dvt <- sqrt(((1-pk)/(n\*pk))%\*%t(pj\*(1-pj)));  
   
 # And there are divisions equal to 0 woman NA and it doesn't work  
 zkj <- dpf  
 zkj[dpf!=0]<-dpf[dpf!=0]/dvt[dpf!=0];   
 pzkj <- pnorm(zkj,lower.tail=F);  
 for(c in 1:length(levels(as.factor(P)))){for (s in 1:length(levels(Xquali))){if (pzkj[c,s]> 0.5){pzkj[c,s]<-1- pzkj[c,s]}}}  
 return (list(rowpf=pf,vtest=zkj,pval=pzkj))  
}

# dades contain the dataset  
dades <- dd  
  
K<-dim(dades)[2]  
par(ask=TRUE)

Using the added column ‘cluster’ which we’ll continue use here as c2

# P must contain the class variable  
c2 <- dd$cluster  
P <- c2  
  
nc <- length(levels(factor(P)))  
nc

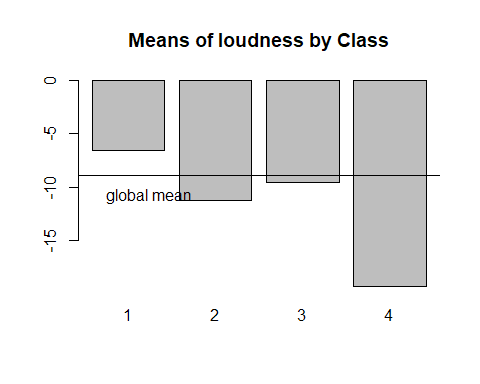
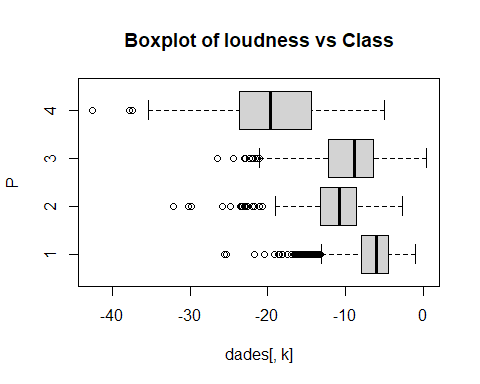
## [1] 4

pvalk <- matrix(data=0,nrow=nc,ncol=K, dimnames=list(levels(P),names(dades)))  
nameP <- "Class"  
n <- dim(dades)[1]

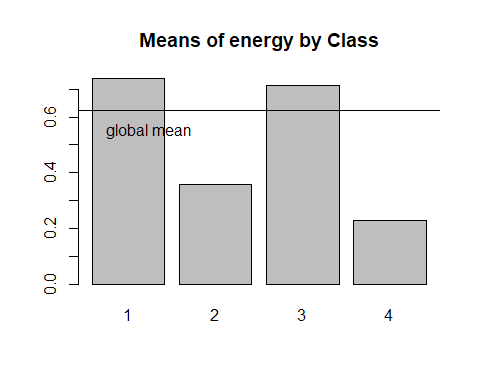
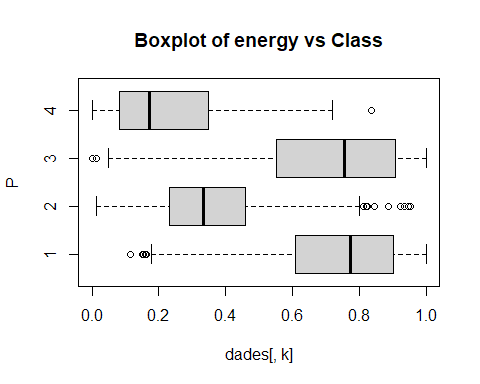
### Loop through the variable dataset to profile and visualize each

for(k in 1:K){  
 if (is.numeric(dades[,k])){   
 print(paste("Analysis by classes of the Variable:", names(dades)[k]))  
   
 boxplot(dades[,k]~P, main=paste("Boxplot of", names(dades)[k], "vs", nameP ), horizontal=TRUE)  
   
 barplot(tapply(dades[[k]], P, mean),main=paste("Means of", names(dades)[k], "by", nameP ))  
 abline(h=mean(dades[[k]]))  
 legend(0,mean(dades[[k]]),"global mean",bty="n")  
 print("Statistics per groups:")  
 for(s in levels(as.factor(P))) {print(summary(dades[P==s,k]))}  
 o<-oneway.test(dades[,k]~P)  
 print(paste("p-value ANOVA:", o$p.value))  
 kw<-kruskal.test(dades[,k]~P)  
 print(paste("p-value Kruskal-Wallis:", kw$p.value))  
 pvalk[,k]<-ValorTestXnum(dades[,k], P)  
 print("p-values ValorsTest: ")  
 print(pvalk[,k])   
 }else{  
 if(class(dd[,k])=="Date"){  
 print(summary(dd[,k]))  
 print(sd(dd[,k]))  
   
 # Decide breaks: weeks, months, quarters...  
 hist(dd[,k],breaks="weeks")  
   
 }else{  
 # Qualitatives  
 print(paste("Variable", names(dades)[k]))  
 table<-table(P,dades[,k])  
  
 rowperc<-prop.table(table,1)  
   
 colperc<-prop.table(table,2)  
   
 # Observe why the variable is true or false. It identifies the type of logic  
 # This one has no levels, therefore, coercion was prevented  
 dades[,k]<-as.factor(dades[,k])  
   
 marg <- table(as.factor(P))/n  
 print(append("Categories=",levels(as.factor(dades[,k]))))  
   
 # From next plots, select one of them according to your practical case  
 plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))  
 paleta<-rainbow(length(levels(dades[,k])))  
 for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }  
   
 # With legend  
 plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))  
 paleta<-rainbow(length(levels(dades[,k])))  
 for(c in 1:length(levels(dades[,k]))){lines(colperc[,c],col=paleta[c]) }  
 legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)  
   
 # Conditioned to classes  
 print(append("Categories=",levels(dades[,k])))  
 plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))  
 paleta<-rainbow(length(levels(dades[,k])))  
 for(c in 1:length(levels(dades[,k]))){lines(rowperc[,c],col=paleta[c]) }  
   
 # With legend  
 plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]))  
 paleta<-rainbow(length(levels(dades[,k])))  
 for(c in 1:length(levels(dades[,k]))){lines(rowperc[,c],col=paleta[c]) }  
 legend("topright", levels(dades[,k]), col=paleta, lty=2, cex=0.6)  
   
 # With abcisses axis variable  
 marg <-table(dades[,k])/n  
 print(append("Categories=",levels(dades[,k])))  
 plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)  
 paleta<-rainbow(length(levels(as.factor(P))))  
 for(c in 1:length(levels(as.factor(P)))){lines(rowperc[c,],col=paleta[c]) }  
   
 # With legend  
 plot(marg,type="l",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)  
 for(c in 1:length(levels(as.factor(P)))){lines(rowperc[c,],col=paleta[c])}  
 legend("topright", levels(as.factor(P)), col=paleta, lty=2, cex=0.6)  
   
 # Conditioned to column  
 plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)  
 paleta<-rainbow(length(levels(as.factor(P))))  
 for(c in 1:length(levels(as.factor(P)))){lines(colperc[c,],col=paleta[c]) }  
   
 # With legend  
 plot(marg,type="n",ylim=c(0,1),main=paste("Prop. of pos & neg by",names(dades)[k]), las=3)  
 for(c in 1:length(levels(as.factor(P)))){lines(colperc[c,],col=paleta[c])}  
 legend("topright", levels(as.factor(P)), col=paleta, lty=2, cex=0.6)  
   
 table<-table(dades[,k],P)  
 print("Cross Table:")  
 print(table)  
 print("Conditional distributions column :")  
 print(colperc)  
   
 # Stacked bar charts   
 paleta<-rainbow(length(levels(dades[,k])))  
 barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta )  
   
 barplot(table(dades[,k], as.factor(P)), beside=FALSE,col=paleta )  
 legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)  
   
 # Attached bar charts  
 barplot(table(dades[,k], as.factor(P)), beside=TRUE,col=paleta )  
   
 barplot(table(dades[,k], as.factor(P)), beside=TRUE,col=paleta)  
 legend("topright",levels(as.factor(dades[,k])),pch=1,cex=0.5, col=paleta)  
   
 print("Square Chi test: ")  
 print(chisq.test(dades[,k], as.factor(P)))  
   
 print("Values test:")  
 print( ValorTestXquali(P,dades[,k]))  
 # Calculate the pvalues of quali  
 }  
 }  
}

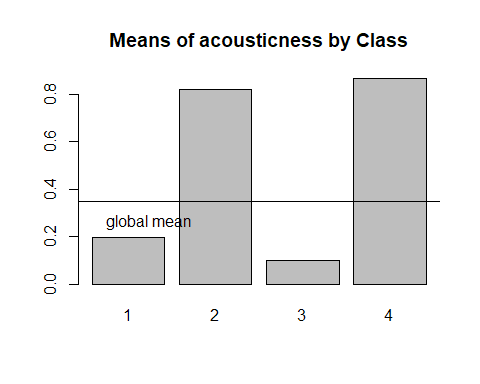
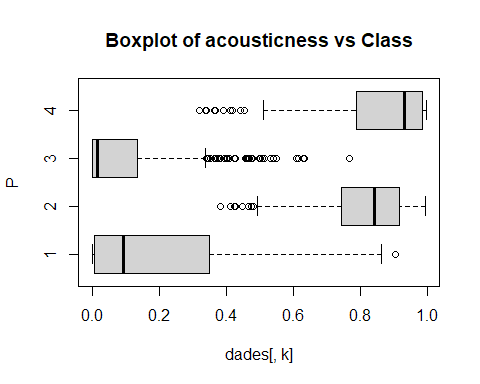
## [1] "Analysis by classes of the Variable: loudness"



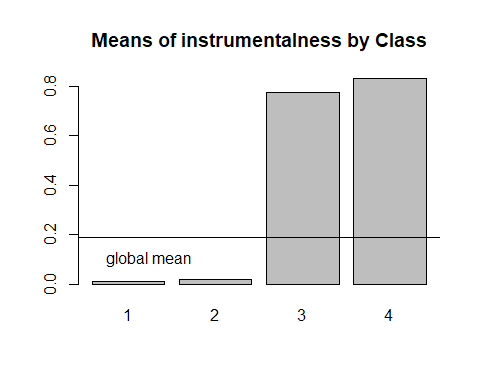
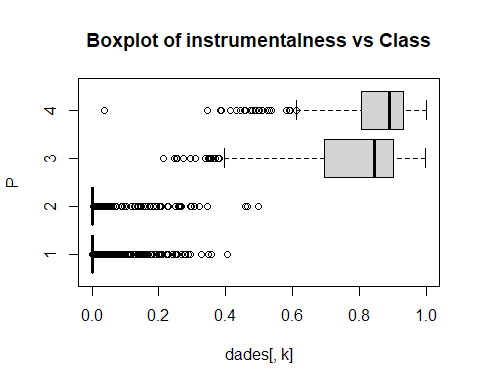
## [1] "Statistics per groups:"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -25.641 -7.987 -6.035 -6.583 -4.519 -0.958   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -32.108 -13.261 -10.819 -11.236 -8.603 -2.689   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -26.527 -12.243 -8.899 -9.548 -6.371 0.377   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -42.631 -23.609 -19.674 -19.321 -14.454 -5.044   
## [1] "p-value ANOVA: 2.17489695592098e-173"  
## [1] "p-value Kruskal-Wallis: 2.45213703529884e-239"  
## [1] "p-values ValorsTest: "  
## [1] 3.451881e-166 0.000000e+00 3.369412e-03 0.000000e+00  
## [1] "Analysis by classes of the Variable: energy"



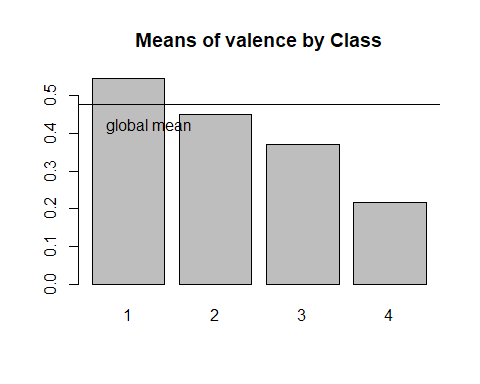
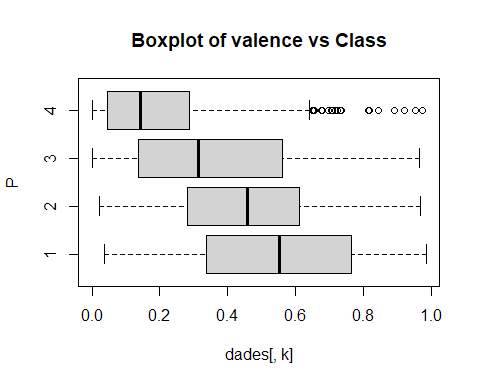
## [1] "Statistics per groups:"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1140 0.6070 0.7710 0.7381 0.9000 0.9990   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0125 0.2315 0.3325 0.3562 0.4590 0.9530   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000242 0.551500 0.755500 0.714677 0.906000 0.999000   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00108 0.08330 0.17200 0.22836 0.34800 0.83600   
## [1] "p-value ANOVA: 1.24124380786274e-274"  
## [1] "p-value Kruskal-Wallis: 1.11360835084654e-278"  
## [1] "p-values ValorsTest: "  
## [1] 5.555962e-161 0.000000e+00 1.102107e-13 0.000000e+00  
## [1] "Analysis by classes of the Variable: acousticness"



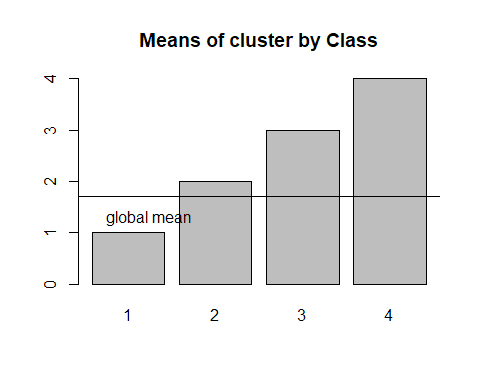
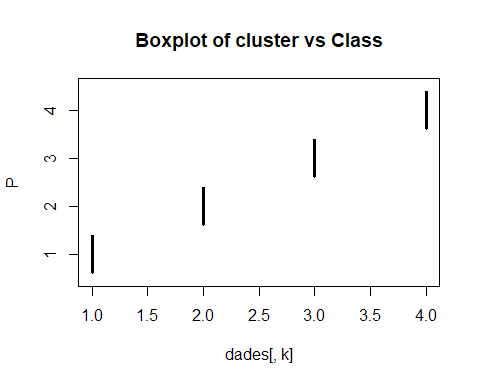
## [1] "Statistics per groups:"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000001 0.006440 0.093600 0.197365 0.348500 0.905000   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.3830 0.7438 0.8410 0.8189 0.9170 0.9950   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000011 0.0003835 0.0147000 0.0970514 0.1352500 0.7660000   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.3190 0.7870 0.9310 0.8646 0.9850 0.9960   
## [1] "p-value ANOVA: 0"  
## [1] "p-value Kruskal-Wallis: 0"  
## [1] "p-values ValorsTest: "  
## [1] 0.000000e+00 6.393416e-212 0.000000e+00 2.450066e-123  
## [1] "Analysis by classes of the Variable: instrumentalness"



## [1] "Statistics per groups:"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000000 0.0000000 0.0000049 0.0121457 0.0006955 0.4060000   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000000 0.000000 0.000006 0.019713 0.001240 0.498000   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2130 0.6940 0.8440 0.7773 0.9010 0.9990   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0379 0.8050 0.8900 0.8312 0.9330 1.0000   
## [1] "p-value ANOVA: 0"  
## [1] "p-value Kruskal-Wallis: 0"  
## [1] "p-values ValorsTest: "  
## [1] 0.000000e+00 0.000000e+00 1.726791e-258 6.493344e-192  
## [1] "Analysis by classes of the Variable: valence"



## [1] "Statistics per groups:"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0366 0.3370 0.5530 0.5458 0.7635 0.9850   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0219 0.2830 0.4585 0.4514 0.6085 0.9680   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1378 0.3130 0.3702 0.5613 0.9640   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0454 0.1430 0.2167 0.2860 0.9730   
## [1] "p-value ANOVA: 6.3023780909863e-92"  
## [1] "p-value Kruskal-Wallis: 4.24544276547456e-94"  
## [1] "p-values ValorsTest: "  
## [1] 7.842021e-65 7.847820e-03 0.000000e+00 0.000000e+00  
## [1] "Analysis by classes of the Variable: cluster"



## [1] "Statistics per groups:"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1 1 1 1 1 1   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2 2 2 2 2 2   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3 3 3 3 3 3   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4 4 4 4 4 4   
## [1] "p-value ANOVA: NaN"  
## [1] "p-value Kruskal-Wallis: 0"  
## [1] "p-values ValorsTest: "  
## [1] 0.000000e+00 8.169678e-14 4.324152e-157 1.458199e-264

## Descriptions for every variable

### What patterns can we see for each class

* Class 1 (black) - Most loudest and energetic
* Class 2 (red) - Most acoustic
* Class 3 (green) - Most energetic and instrumental
* Class 4 (blue) - Most acoustic and instrumental

Happiest to saddest songs - Class 1 > 2 > 3 > 4

# Descriptors of the most significant classes. Add infoboxes  
for (c in 1:length(levels(as.factor(P)))) {  
 if(!is.na(levels(as.factor(P))[c])){  
 print(paste("P.values per class:",levels(as.factor(P))[c]));  
 print(sort(pvalk[c,]), digits=3)   
 }  
}

## [1] "P.values per class: 1"  
## acousticness instrumentalness cluster loudness   
## 0.00e+00 0.00e+00 0.00e+00 3.45e-166   
## energy valence   
## 5.56e-161 7.84e-65   
## [1] "P.values per class: 2"  
## loudness energy instrumentalness acousticness   
## 0.00e+00 0.00e+00 0.00e+00 6.39e-212   
## cluster valence   
## 8.17e-14 7.85e-03   
## [1] "P.values per class: 3"  
## acousticness valence instrumentalness cluster   
## 0.00e+00 0.00e+00 1.73e-258 4.32e-157   
## energy loudness   
## 1.10e-13 3.37e-03   
## [1] "P.values per class: 4"  
## loudness energy valence cluster   
## 0.00e+00 0.00e+00 0.00e+00 1.46e-264   
## instrumentalness acousticness   
## 6.49e-192 2.45e-123

Add the information of the modalities of qualitative to the list of pvalues and make global ordering