Spotify data - Principal Component Analysis

## Introduction

In this report, we will be doing a Principal Component analysis based on Spotify data on popular songs.

A Principal Component Analysis is a technique to analyze large datasets containing multiple dimensions/features. The purpose is to enable visualization of multidimensional data.

PCA identifies the main axes of variance within a data set and allows for easy data exploration to understand the key variables in the data and spot outliers.

The goal is to determine which variables are related or not related to each other.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.7  
## ✔ tidyr 1.1.4 ✔ stringr 1.4.0  
## ✔ readr 2.1.3 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

## Importing the data

We are using a dataset of popular songs on Spotify. In the previous report on pre-processing, descriptive and bivariate statistics, we have described this dataset and made several transformations.

[Link to original dataset](https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset)

setwd("~/Documents/class/stats-final-project/")  
dd <- read.csv("cleaneddata.csv") %>% mutate(  
 time\_signature = as.character(time\_signature)  
)

## Performing the Principal Component Analysis

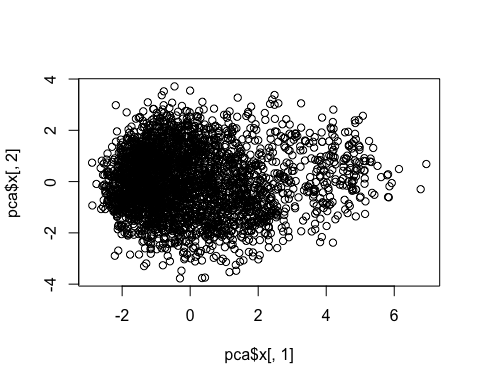
We use the prcomp() function.

This returns 3 things:

1. x => contains the principal components (PCs) for drawing a graph. We will be using the first two columns in x to draw a 2D plot that uses the first 2 PCs. The number of PCs is determined by the number of variables used.
2. sdev
3. rotation

Below is the code to run the prcomp function, and a basic plot of principal component 1 and 2.

# determine which ones are numerical variables  
numerical <- which(sapply(dd,is.numeric))  
# print below to see if the numerical variables are correctly detected  
# numerical  
  
# saving the numerical observations to "dcon"  
dcon <- dd[,numerical]  
# print below to see if variables detected are indeed numerical  
# sapply(dcon,class)  
  
# Now we do a PRINCIPAL COMPONENT ANALYSIS on the numerical variables  
pca <- prcomp(dcon, scale=TRUE, center = TRUE) # centering and scaling true  
pc1 <- pca  
  
plot(pca$x[,1], pca$x[,2])



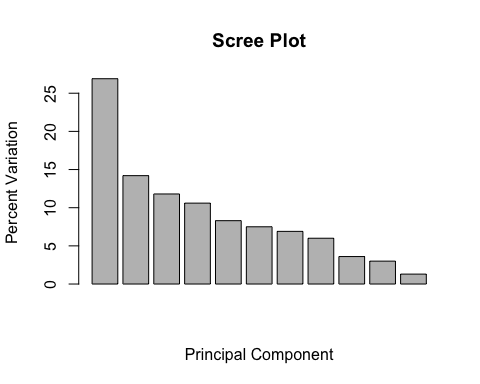
### Scree plot

With the principal component analysis, we can also compute the Scree plot, which displays the variance accounted for by the components.

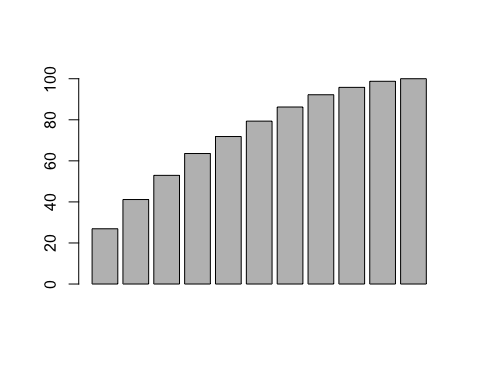
pca.var <- pca$sdev^2  
pca.var.per <- round(pca.var/sum(pca.var)\*100, 1)  
pca.var.per

## [1] 26.9 14.2 11.8 10.6 8.3 7.5 6.9 6.0 3.6 3.0 1.3

barplot(pca.var.per, main="Scree Plot", xlab="Principal Component", ylab="Percent Variation")



#Cummulated Inertia in subspaces, from first principal component to the 11th dimension subspace  
barplot(100\*cumsum(pc1$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2])



percInerAccum<-100\*cumsum(pc1$sdev[1:dim(dcon)[2]]^2)/dim(dcon)[2]  
percInerAccum

## [1] 26.91234 41.15655 52.92813 63.56836 71.84598 79.34463 86.22294  
## [8] 92.19206 95.78665 98.74806 100.00000

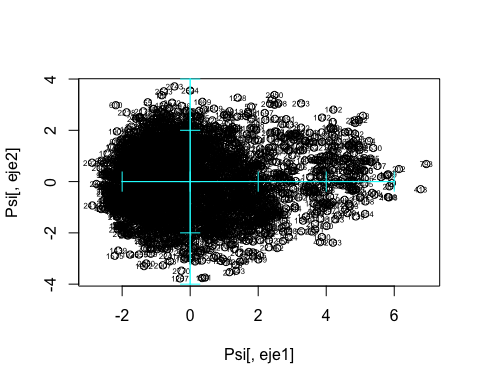
From this we see that Principal component 1 accounts for only 26.9% of the variation. And in order to account for 80% of the variation, we need to include PC 1-7.

Next we will store the eigenvalues, eigenvectors, projections and include PC 1-7.

# SELECTION OF THE SINGIFICANT DIMENSIONS (keep 80% of total inertia)  
nd = 7  
pc1$rotation  
  
# STORAGE OF THE EIGENVALUES, EIGENVECTORS AND PROJECTIONS IN THE nd DIMENSIONS  
Psi = pc1$x[,1:nd]  
Psi  
  
  
# STORAGE OF LABELS FOR INDIVIDUALS AND VARIABLES  
iden = row.names(dcon)  
etiq = names(dcon) # getting names of numerical variables  
ze = rep(0,length(etiq)) # WE WILL NEED THIS VECTOR AFTERWARDS FOR THE GRAPHICS

## Plotting the individuals on PC1 and PC2 axes

# PLOT OF INDIVIDUALS  
  
#select your axis  
#eje1<-2  
eje1<-1  
#eje2<-3  
eje2<-2  
  
plot(Psi[,eje1],Psi[,eje2])  
text(Psi[,eje1],Psi[,eje2],labels=iden, cex=0.5)  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")



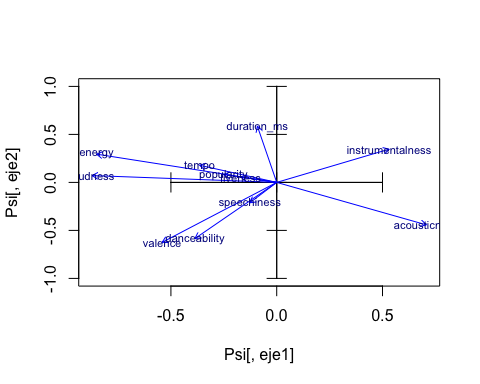
## Plotting projection of variables, PC1 and PC2

We will create a loadings plot to see which variables most influence PC1 and PC2.

#Projection of variables  
  
Phi = cor(dcon,Psi)  
Phi

## PC1 PC2 PC3 PC4 PC5  
## popularity -0.25153061 0.08198524 0.04787668 -0.776358728 0.1270312037  
## duration\_ms -0.09138802 0.58531504 -0.18968378 0.075178804 0.4446524295  
## danceability -0.38602544 -0.58390036 -0.35836660 0.268458136 0.2760524613  
## energy -0.85174751 0.29428633 0.09495404 0.194477832 0.0008212764  
## loudness -0.87255545 0.07126792 0.01756260 -0.062729926 0.0144974134  
## speechiness -0.12714215 -0.21242983 0.60931479 0.424680399 -0.0511402462  
## acousticness 0.70702764 -0.44238120 0.12823351 -0.207996111 -0.0805660637  
## instrumentalness 0.53137677 0.34622971 -0.20001793 0.469699797 -0.0473689265  
## liveness -0.16940196 0.05524380 0.78234881 -0.002361446 0.0622204014  
## valence -0.54122196 -0.62398564 -0.18906447 0.058291474 0.0089392328  
## tempo -0.36450385 0.17835956 -0.20813396 -0.024004157 -0.7777975541  
## PC6 PC7  
## popularity -0.16056296 0.359065114  
## duration\_ms -0.58574289 -0.116380945  
## danceability -0.24189166 0.070197679  
## energy 0.18993894 -0.006828433  
## loudness 0.18207412 0.014498890  
## speechiness -0.26382962 0.542555707  
## acousticness -0.21920809 -0.120653957  
## instrumentalness 0.04368884 0.059121230  
## liveness -0.15271351 -0.479746966  
## valence -0.09771015 -0.252593624  
## tempo -0.41928583 -0.050769168

X<-Phi[,eje1]  
Y<-Phi[,eje2]  
  
#zooms  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(min(X,0),max(X,0)), ylim=c(-1,1))  
axis(side=1, pos= 0, labels = F)  
axis(side=3, pos= 0, labels = F)  
axis(side=2, pos= 0, labels = F)  
axis(side=4, pos= 0, labels = F)  
arrows(ze, ze, X, Y, length = 0.07,col="blue")  
text(X,Y,labels=etiq,col="darkblue", cex=0.7)



### Observations:

Based on the Loadings plot, we see that the variables that most influence PC1 are instrumentalness, energy, loudness, and acousticness.

The variables that most influence PC2 are duration.

From this plot we can also see that these variables may be positively correlated: - Loudness and energy - Valence and danceability - Instrumentalness and acousticness

These variables may be negatively correlated: - Loudness and instrumentalness - Loudness and acousticness - Energy and acousticness - Valence and instrumentalness - Tempo and acousticness

These variables are orthogonal/may not be very related to each other: - Duration vs loudness - Duration vs instrumentalness - Speechiness vs acousticness

We can also try plotting the PC1 against PC3 to see the variables that are not so clearly visible here.

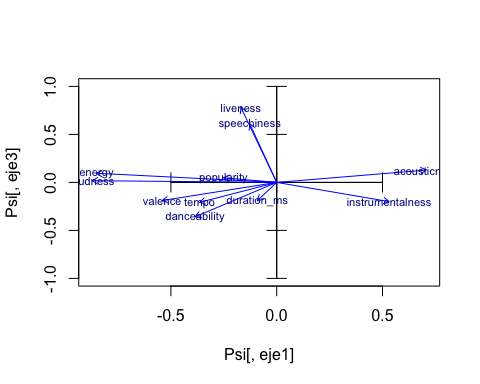
## Plotting projection of variables, PC1 and PC3

Since PC1 and PC2 only accounts for 38% of the variation, we should also plot projection of variables in PC1 and PC3.

#Projection of variables  
  
Phi = cor(dcon,Psi)  
Phi

## PC1 PC2 PC3 PC4 PC5  
## popularity -0.25153061 0.08198524 0.04787668 -0.776358728 0.1270312037  
## duration\_ms -0.09138802 0.58531504 -0.18968378 0.075178804 0.4446524295  
## danceability -0.38602544 -0.58390036 -0.35836660 0.268458136 0.2760524613  
## energy -0.85174751 0.29428633 0.09495404 0.194477832 0.0008212764  
## loudness -0.87255545 0.07126792 0.01756260 -0.062729926 0.0144974134  
## speechiness -0.12714215 -0.21242983 0.60931479 0.424680399 -0.0511402462  
## acousticness 0.70702764 -0.44238120 0.12823351 -0.207996111 -0.0805660637  
## instrumentalness 0.53137677 0.34622971 -0.20001793 0.469699797 -0.0473689265  
## liveness -0.16940196 0.05524380 0.78234881 -0.002361446 0.0622204014  
## valence -0.54122196 -0.62398564 -0.18906447 0.058291474 0.0089392328  
## tempo -0.36450385 0.17835956 -0.20813396 -0.024004157 -0.7777975541  
## PC6 PC7  
## popularity -0.16056296 0.359065114  
## duration\_ms -0.58574289 -0.116380945  
## danceability -0.24189166 0.070197679  
## energy 0.18993894 -0.006828433  
## loudness 0.18207412 0.014498890  
## speechiness -0.26382962 0.542555707  
## acousticness -0.21920809 -0.120653957  
## instrumentalness 0.04368884 0.059121230  
## liveness -0.15271351 -0.479746966  
## valence -0.09771015 -0.252593624  
## tempo -0.41928583 -0.050769168

eje3 <- 3  
  
X<-Phi[,eje1]  
Y<-Phi[,eje3]  
  
#zooms  
plot(Psi[,eje1],Psi[,eje3],type="n",xlim=c(min(X,0),max(X,0)), ylim=c(-1,1))  
axis(side=1, pos= 0, labels = F)  
axis(side=3, pos= 0, labels = F)  
axis(side=2, pos= 0, labels = F)  
axis(side=4, pos= 0, labels = F)  
arrows(ze, ze, X, Y, length = 0.07,col="blue")  
text(X,Y,labels=etiq,col="darkblue", cex=0.7)



Based on the above Loadings plot, we see that the variables that most influence PC3 are popularity, liveness and duration.

It also seems like Popularity and liveness are closely related, while it may be negatively correlated with duration.

We can also see that speechiness is closer to danceability.

## Finding Centroids

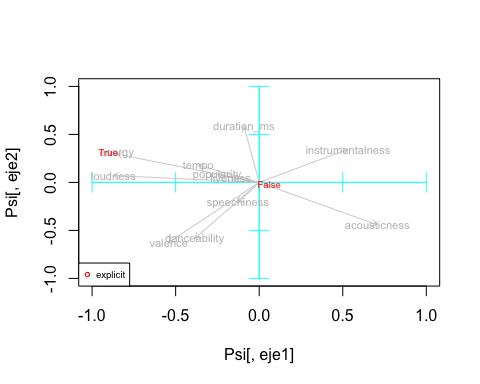
We can also find the centroids of modalities in categorical variables, using the code below.

That looks a bit crowded, so let’s try looking at each categorical variables’ modalities one by one.

### Explicit

Explicit songs/songs that contain swear words are more likely to be energetic. This could be Latino songs or rap songs.

X<-Phi[,eje1]  
Y<-Phi[,eje2]  
  
#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,1), ylim=c(-1,1))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(3)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)



### Key

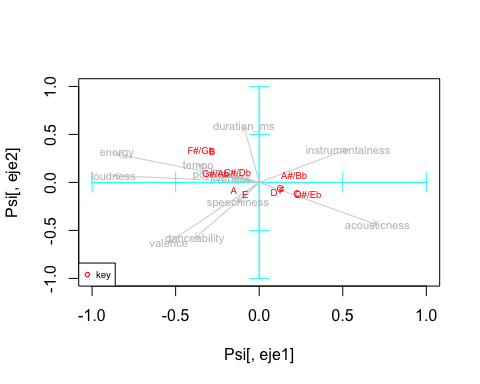
When we plot the modality of keys, it’s interesting because you can see that some keys are associated with energetic music, such as the “black” keys F#, G# and C#.

Some other keys such as the more popular C, D and G are associated with more acoustic music.

E and A are more danceable, speechy songs (probably, children’s music)?

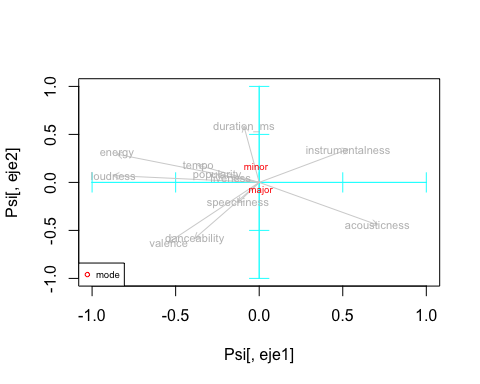
It fits this description that key signatures have music characteristics: “Noisy shouts of joy, laughing pleasure and not yet complete, full delight lies in E Major.”

#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,1), ylim=c(-1,1))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(6)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)



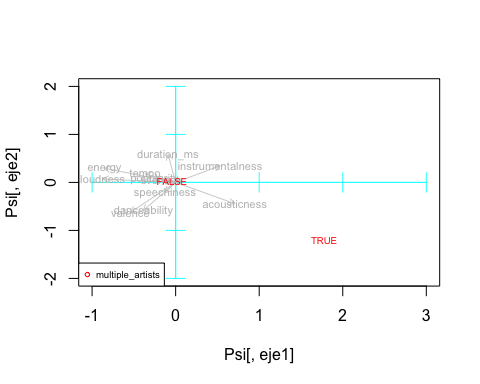
### Mode

#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,1), ylim=c(-1,1))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(8)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)

 ### Multiple artists

This plot shows that songs that have multiple artists are more likely to be acoustic.

#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,3), ylim=c(-2,2))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(17)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)



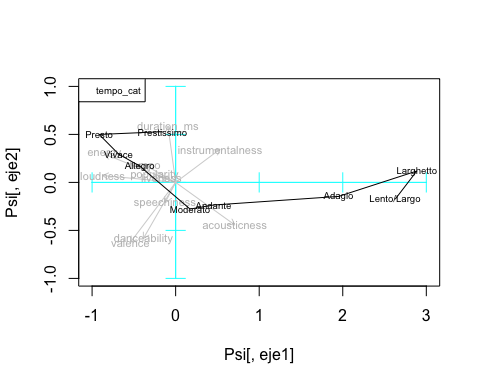
### Tempo category (an ordinal variable)

Slower songs are on the right side, and faster songs are on the left.

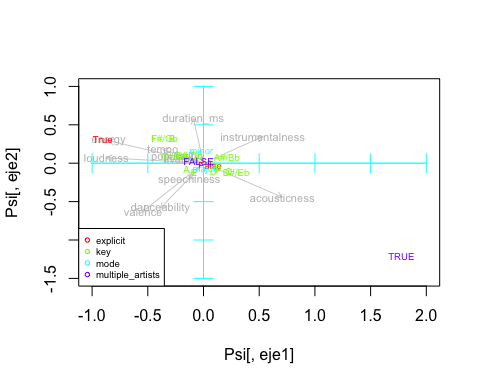
#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,3), ylim=c(-1,1))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
  
#add ordinal qualitative variables. Ensure ordering is the correct  
  
dordi<-c(18)  
  
  
#reorder modalities: when required  
dd[,dordi[1]] <- factor(dd[,dordi[1]], ordered=TRUE, levels= c('Larghissimo','Grave','Lento/Largo','Larghetto','Adagio','Andante','Moderato','Allegro','Vivace','Presto','Prestissimo'))  
levels(dd[,dordi[1]])

## [1] "Larghissimo" "Grave" "Lento/Largo" "Larghetto" "Adagio"   
## [6] "Andante" "Moderato" "Allegro" "Vivace" "Presto"   
## [11] "Prestissimo"

c<-1  
col<-length(dcat)+1  
for(k in dordi){  
 seguentColor<-colors[col]  
 fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
 fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
   
 #points(fdic1,fdic2,pch=16,col=seguentColor, labels=levels(dd[,k]))  
 #connect modalities of qualitative variables  
 lines(fdic1,fdic2,col="#000000")  
 text(fdic1,fdic2,labels=levels(dd[,k]),col=seguentColor, cex=0.6)  
 c<-c+1  
 col<-col+1  
}  
legend("topleft",names(dd)[dordi],pch=19,col=colors[col:col+length(dordi)-1], cex=0.6)

 ### All categorical variables

#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,2), ylim=c(-1.5,1))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(3,6,8,17)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)

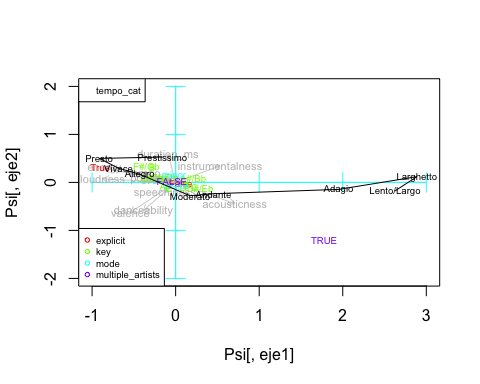


### All categorical variables (including tempo) except genre

#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,3), ylim=c(-2,2))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(3,6,8,17)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)  
  
  
  
#add ordinal qualitative variables. Ensure ordering is the correct  
  
dordi<-c(18)  
  
  
#reorder modalities: when required  
dd[,dordi[1]] <- factor(dd[,dordi[1]], ordered=TRUE, levels= c('Larghissimo','Grave','Lento/Largo','Larghetto','Adagio','Andante','Moderato','Allegro','Vivace','Presto','Prestissimo'))  
levels(dd[,dordi[1]])

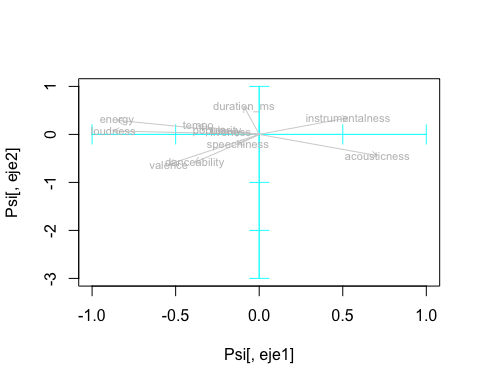
## [1] "Larghissimo" "Grave" "Lento/Largo" "Larghetto" "Adagio"   
## [6] "Andante" "Moderato" "Allegro" "Vivace" "Presto"   
## [11] "Prestissimo"

c<-1  
col<-length(dcat)+1  
for(k in dordi){  
 seguentColor<-colors[col]  
 fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
 fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
   
 #points(fdic1,fdic2,pch=16,col=seguentColor, labels=levels(dd[,k]))  
 #connect modalities of qualitative variables  
 lines(fdic1,fdic2,col="#000000")  
 text(fdic1,fdic2,labels=levels(dd[,k]),col=seguentColor, cex=0.6)  
 c<-c+1  
 col<-col+1  
}  
legend("topleft",names(dd)[dordi],pch=19,col=colors[col:col+length(dordi)-1], cex=0.6)



### All categorical variables together

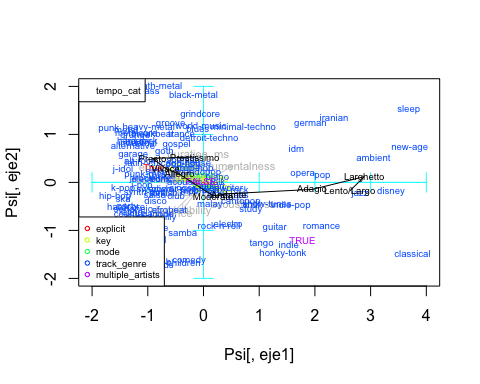
X<-Phi[,eje1]  
Y<-Phi[,eje2]  
  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,1), ylim=c(-3,1))  
#plot(X,Y,type="none",xlim=c(min(X,0),max(X,0)))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)



#all qualitative together  
plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-2,4), ylim=c(-2,2))  
axis(side=1, pos= 0, labels = F, col="cyan")  
axis(side=3, pos= 0, labels = F, col="cyan")  
axis(side=2, pos= 0, labels = F, col="cyan")  
axis(side=4, pos= 0, labels = F, col="cyan")  
arrows(ze, ze, X, Y, length = 0.07,col="lightgray")  
text(X,Y,labels=etiq,col="gray", cex=0.7)  
  
#nominal qualitative variables  
  
dcat<-c(3, 6, 8, 16, 17)  
colors<-rainbow(length(dcat))  
  
c<-1  
for(k in dcat){  
 seguentColor<-colors[c]  
fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
  
text(fdic1,fdic2,labels=levels(factor(dd[,k])),col=seguentColor, cex=0.6)  
c<-c+1  
}  
legend("bottomleft",names(dd)[dcat],pch=1,col=colors, cex=0.6)  
  
#add ordinal qualitative variables. Ensure ordering is the correct  
  
dordi<-c(18)  
  
  
#reorder modalities: when required  
dd[,dordi[1]] <- factor(dd[,dordi[1]], ordered=TRUE, levels= c('Larghissimo','Grave','Lento/Largo','Larghetto','Adagio','Andante','Moderato','Allegro','Vivace','Presto','Prestissimo'))  
levels(dd[,dordi[1]])

## [1] "Larghissimo" "Grave" "Lento/Largo" "Larghetto" "Adagio"   
## [6] "Andante" "Moderato" "Allegro" "Vivace" "Presto"   
## [11] "Prestissimo"

c<-1  
col<-length(dcat)+1  
for(k in dordi){  
 seguentColor<-colors[col]  
 fdic1 = tapply(Psi[,eje1],dd[,k],mean)  
 fdic2 = tapply(Psi[,eje2],dd[,k],mean)   
   
 #points(fdic1,fdic2,pch=16,col=seguentColor, labels=levels(dd[,k]))  
 #connect modalities of qualitative variables  
 lines(fdic1,fdic2,col="#000000")  
 text(fdic1,fdic2,labels=levels(dd[,k]),col=seguentColor, cex=0.6)  
 c<-c+1  
 col<-col+1  
}  
legend("topleft",names(dd)[dordi],pch=19,col=colors[col:col+length(dordi)-1], cex=0.6)



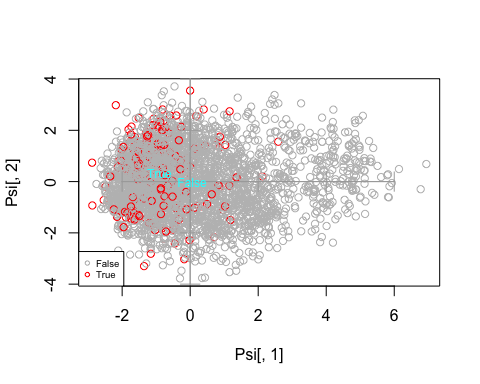
Observations from the plot of centroids above:

1. Seems like the slower songs (Grave, Lento, Larghetto, Adagio) are also on the right side of the plot, which are more acoustic, instrumental songs. Whereas the faster songs are on the left side.
2. Songs that are more loud, energetic and probably contains swear words are related to heavy metal, grunge, goth, genres, or latin/latino songs.
3. Disney, jazz, and classical songs are probably high on the acousticness scale, while new-age, ambient, sleep songs are high on the instrumentalness scale.
4. Songs on the bottom left side are happy, danceable songs, but the genres vary from children songs to songs that are more known to be danceable, such as hip hop, r&b, kpop, disco, salsa, etc.

## Coloring the PCA plot using categorical variables

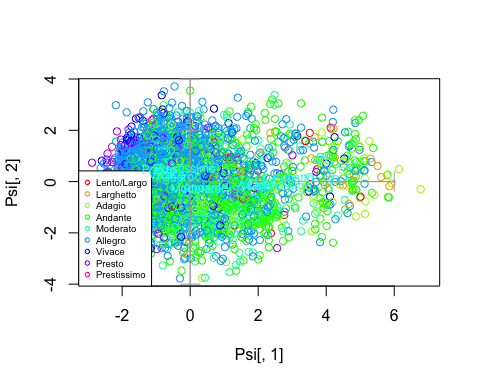
We can also plot all the individuals in PC1 and PC2 as axes, and color-code it by categorical variables. We’ll examine one categorical variable: Explicitness

# PROJECTION OF ILLUSTRATIVE qualitative variables on individuals' map  
varcat=factor(dd[,3])  
plot(Psi[,1],Psi[,2],col=c("grey", "red")[varcat])  
axis(side=1, pos= 0, labels = F, col="darkgray")  
axis(side=3, pos= 0, labels = F, col="darkgray")  
axis(side=2, pos= 0, labels = F, col="darkgray")  
axis(side=4, pos= 0, labels = F, col="darkgray")  
legend("bottomleft",levels(varcat),pch=1,col=c("grey", "red"), cex=0.6)  
  
  
# Overproject THE CDG OF LEVELS OF varcat  
fdic1 = tapply(Psi[,1],varcat,mean)  
fdic2 = tapply(Psi[,2],varcat,mean)   
  
text(fdic1,fdic2,labels=levels(factor(varcat)),col="cyan", cex=0.75)



Although there are not many explicit songs, we can see that the explicit songs tend to be on the left side of PC1.

# PROJECTION OF ILLUSTRATIVE qualitative variables on individuals' map  
varcat=factor(dd[,18])  
plot(Psi[,1],Psi[,2],col=rainbow(9)[varcat])  
axis(side=1, pos= 0, labels = F, col="darkgray")  
axis(side=3, pos= 0, labels = F, col="darkgray")  
axis(side=2, pos= 0, labels = F, col="darkgray")  
axis(side=4, pos= 0, labels = F, col="darkgray")  
legend("bottomleft",levels(varcat),pch=1,col=rainbow(9), cex=0.6)  
  
  
# Overproject THE CDG OF LEVELS OF varcat  
fdic1 = tapply(Psi[,1],varcat,mean)  
fdic2 = tapply(Psi[,2],varcat,mean)   
  
text(fdic1,fdic2,labels=levels(factor(varcat)),col="cyan", cex=0.75)



From the tempo categories plot we see that the left side of PCA is the faster songs, and the right side are the slower songs.