# Lecture des données

data = read.table("GSE53626\_series\_matrix2.txt", sep='\t', header=T, comment.char="!", row.names=1, na.string="NC")

nationality = as.factor(t(read.table("GSE53626\_series\_matrix2.txt", skip=42, nrows=1))[-1])  
table(nationality)

length(nationality)

## [1] 158

library(stringr)  
natio = as.factor(unlist(lapply(nationality, str\_sub, 14)))  
table(natio)

## natio  
## Bulgarian EastSicilian Greek Moroccan Norwegian   
## 18 10 17 15 15   
## Polish SouthItalian Spanish Tunisian UAE   
## 11 18 18 12 14   
## WestSicilian   
## 10

sexe = as.factor(t(read.table("GSE53626\_series\_matrix2.txt", skip=41, nrows=1))[-1])  
table(sexe)

## sexe  
## gender: female gender: male   
## 4 154

length(sexe)

## [1] 158

# Gestion des données manquantes

data2 = data[rowSums(is.na(data)) == 0, ]

**ACP**

library(FactoMineR)

res.pca = PCA(X = t(data2))

library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

ind = get\_pca\_ind(res.pca)

data\_ind = ind$coord[,1:2]  
data\_ind = cbind(data\_ind, natio)

library(ggplot2)  
  
data\_ind = as.data.frame(data\_ind)  
ggplot(data\_ind, aes(x = data\_ind[,1],y = data\_ind[,2], colour = factor(data\_ind[,3]))) + geom\_point() + scale\_color\_manual(name="Nationalités", labels = c("Bulgarian", "East-Sicilian", "Greek", "Moroccan", "Norwegian", "Polish", "South-Italian", "Spanish", "Tunisian", "UAE", "West-Sicilian"), values = rainbow(11)) + xlab("Dim.1") + ylab("Dim.2")

var <- get\_pca\_var(res.pca)

# Contributions des variables à PC1  
fviz\_contrib(res.pca, choice = "var", axes = 1, top = 500)

# Contributions des variables à PC2  
fviz\_contrib(res.pca, choice = "var", axes = 2, top = 500)

# Clustering: k-means

res.kmeans = kmeans(x = t(data2), centers = 11)

plot\_grid(  
ggplot(data\_ind, aes(x = data\_ind[,1],y = data\_ind[,2])) + geom\_point(colour = factor(data\_ind[,3])) + xlab("Dim.1") + ylab("Dim.2") + ggtitle("Groupes nationality"),  
ggplot(data\_ind, aes(x = data\_ind[,1],y = data\_ind[,2])) + geom\_point(colour = factor(res.kmeans$cluster)) + xlab("Dim.1") + ylab("Dim.2") + ggtitle("Groupes kmeans"))