Real Data - Iris data set

III.1. Data description

In this section, we will analyze the three clustering techniques on real data set. We chose the classical the well known data set "Iris". The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. (This discription was copied from the **UCI website**

Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class

Predicted attribute: class of iris plant:

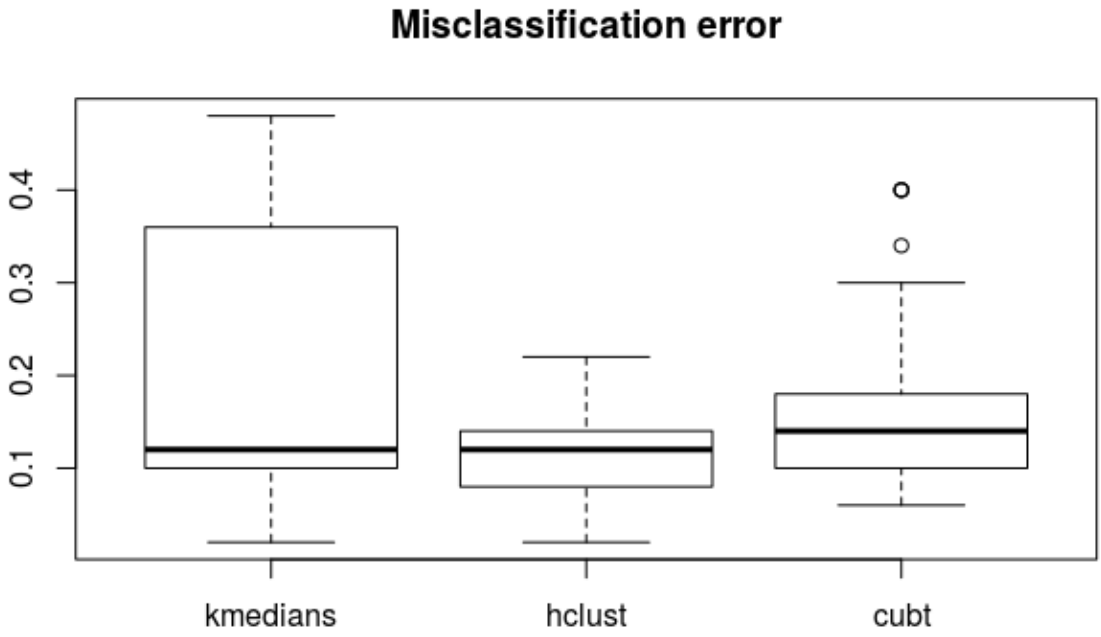
- Iris Setosa
- Iris Versicolour
- Iris Virginica

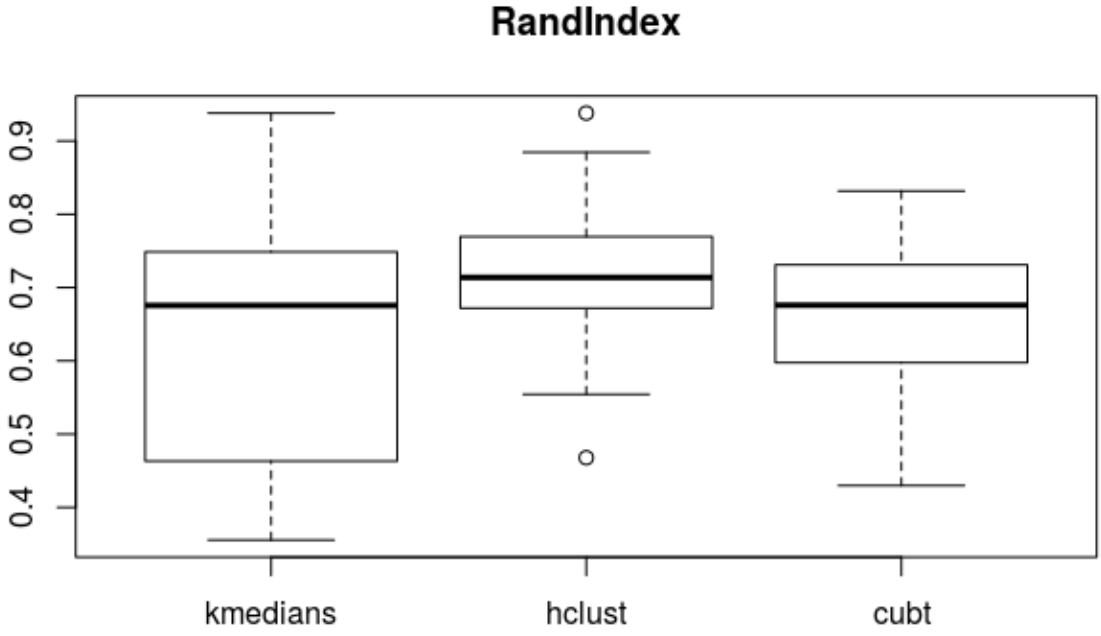
III.2. Computation of Misclassification Error and Rand Index

As previously, we will compute these value $M = 100$ times. After analyzing the boxplots, we can observe that for this data set, none of the type of clustering is significantly better thant the other.

```
data(iris)
dataset <- iris
M <- 100
randind.km <- NULL
randind.hc <- NULL
randind.cubt <- NULL

error.km <- NULL
error.hc <- NULL
error.cubt <- NULL

n <- nrow(iris)

for (i in 1:M){
  train.index <- sample(1:nrow(dataset),size = floor(2/3 * nrow(dataset)))
  x.train <- dataset[train.index,1:ncol(dataset)-1]
  x.test <- dataset[-train.index,1:ncol(dataset)-1]
  y.train <- dataset[train.index,ncol(dataset)]
  y.test <- dataset[-train.index,ncol(dataset)]

  ## Kmedians
  km <- kcca(x.train, k=3, kccaFamily("kmedians"))
  pred <- predict(km, newdata = x.test)

  randind.km <- c(randind.km, randIndex(pred,y.test))
  error.km <- c(error.km, error(pred , y.test, print = FALSE)[1])
```

```
  ## Hclust
  x.train.dist <- dist(x.train,method = "manhattan")
  x.test.dist <- dist(x.test, method = "manhattan")
  hc <- hclust(x.train.dist,  "ward.D")  # Using "manhattan" for "categorical variables"
  # hcd <- as.dendrogram(hc)
  # pred <- predict.clus(as.numeric(hc$labels), x.train,x.test)
  cc <- as.kcca(hc,x.train,k=3)
  pred <- predict(cc,x.test)

  randind.hc <- c(randind.hc, randIndex(pred,y.test))
  error.hc <- c(error.hc, error(pred , y.test, print = FALSE)[1])


  ## CUBT

  cubt.clustering.max <- cubt(as.matrix(x.train))
  cubt.joined <- join.cubt(cubt.clustering.max,x.train,nclass=3)

  x.test <- as.matrix(x.test)
  pred<-predict.cubt(cubt.joined, x.test, minsplit = 4, type = "class")
  randind.cubt <- c(randind.cubt, randIndex(factor(pred),y.test))
  error.cubt <- c(error.cubt, error(pred , y.test, print = FALSE)[1])

}

## Plotting misclassification errors and rand index boxplot
errors.models <- list(error.km,error.hc,error.cubt)
randindex.models <- list(randind.km,randind.hc,randind.cubt)

names(errors.models) <- c('kmedians','hclust','cubt')
names(randindex.models) <- c('kmedians','hclust','cubt')

boxplot(errors.models,main="Misclassification error")
boxplot(randindex.models,main="RandIndex")
```


Misclassification error
0.1
0.2
0.3
0.4
kmedians
hclust
cubt


RandIndex
0.4
0.5
0.6
0.7
0.8
0.9
kmedians
hclust
cubt