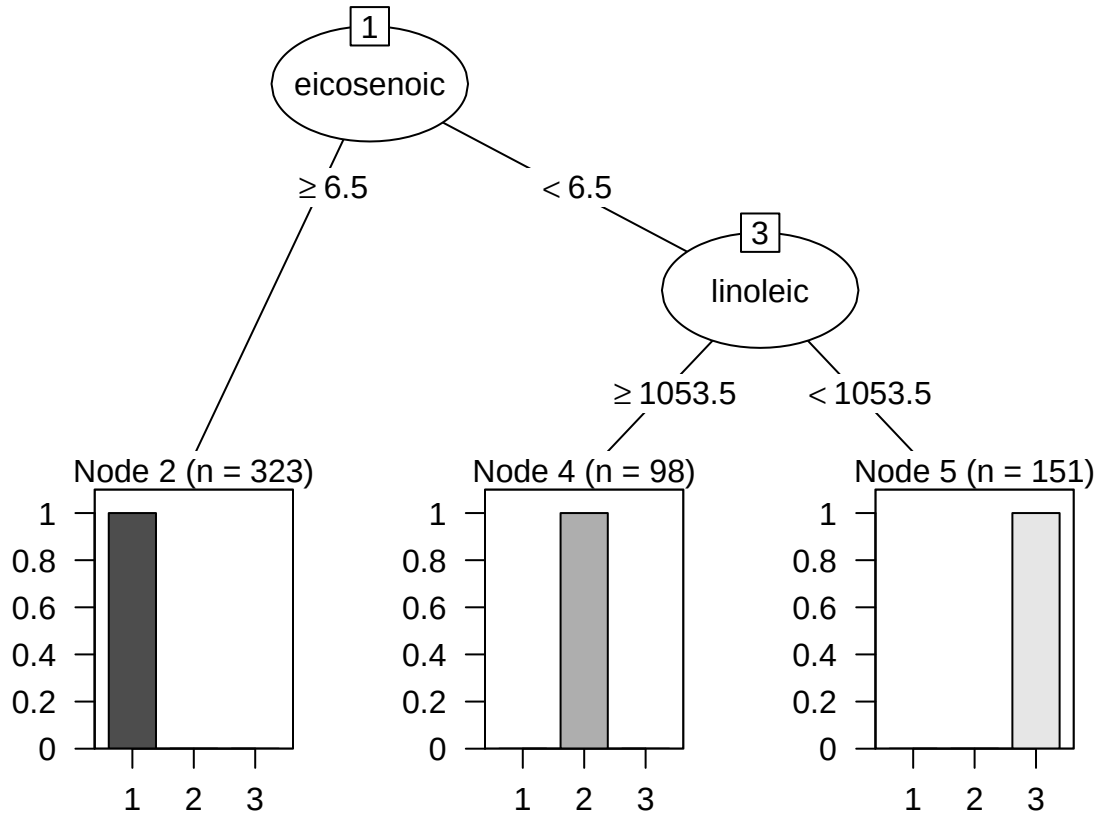
Computer Lab 5

Thomas Zhang

2016 M10 8

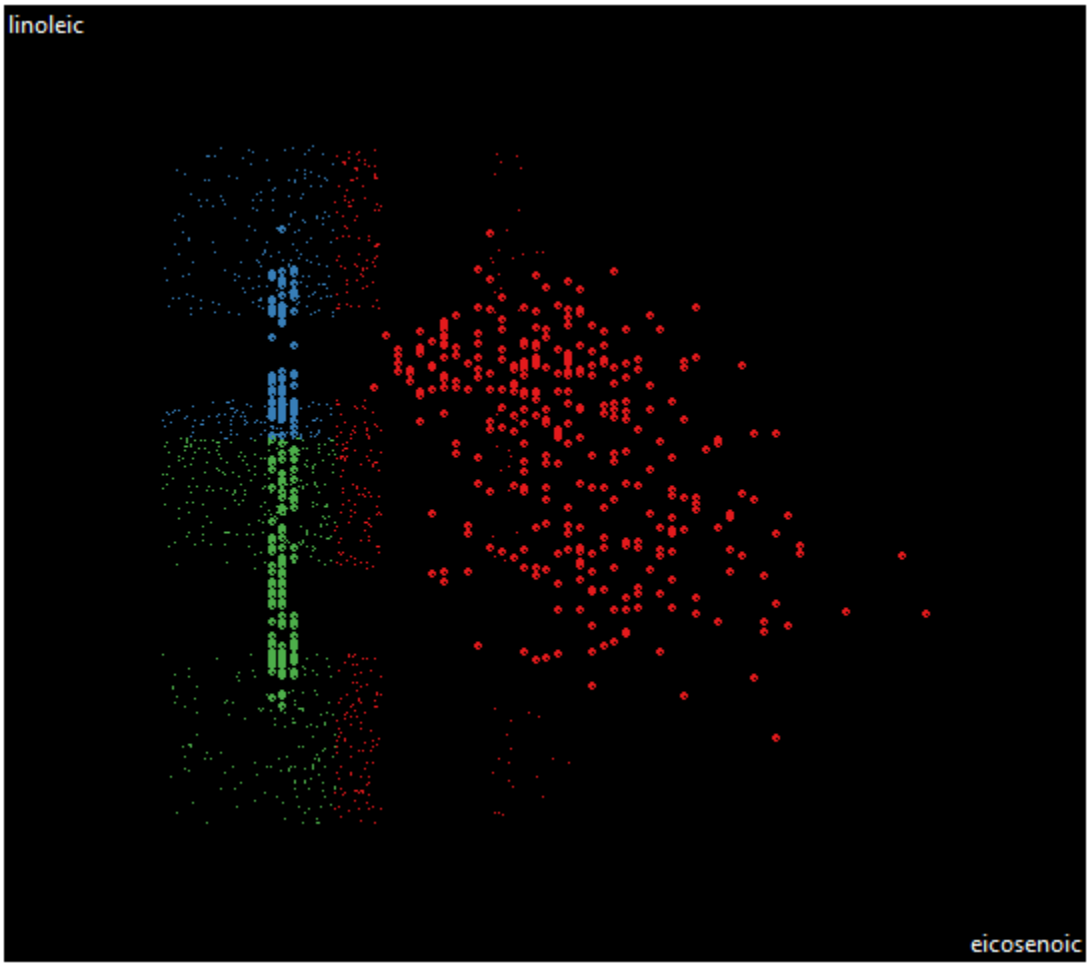
Assignment 1

We start out by importing olive oil data from `olive.csv` into R and try to *"classify oils using decision trees by treating Region as response and all acids as explanatory variables"*. We plot the decision tree below.



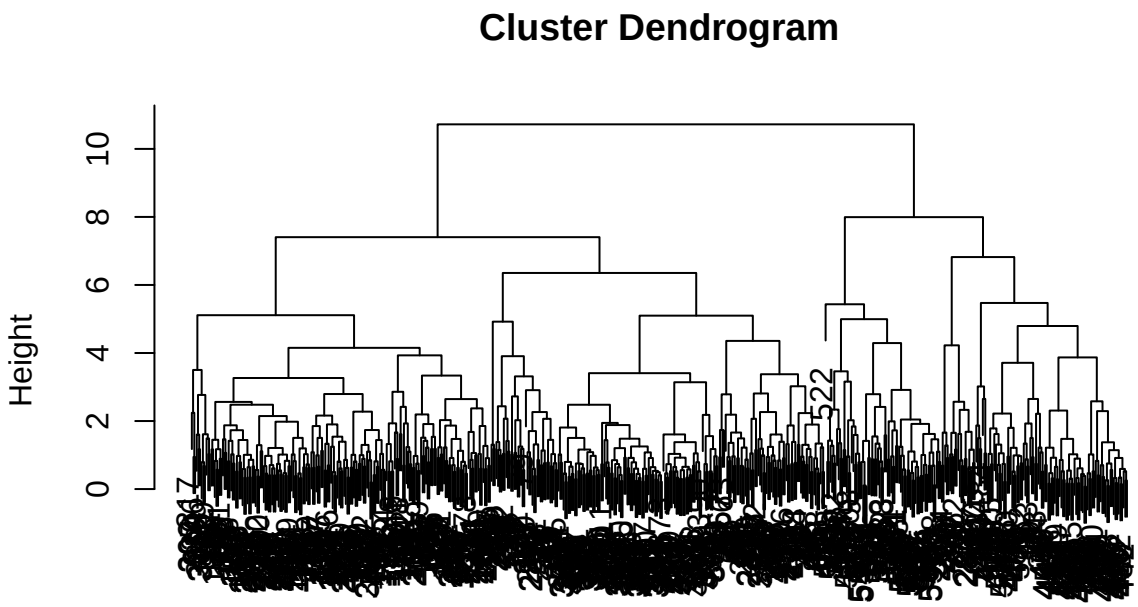
We can see that actually none of the oils were misclassified, meaning that the acid variables used for decision-making were very well chosen. In our case, we used linoleic and eicosenic acids and the tree is very shallow.

Let us use these two acids and make a classifly plot in order to demonstrate the decision boundaries.

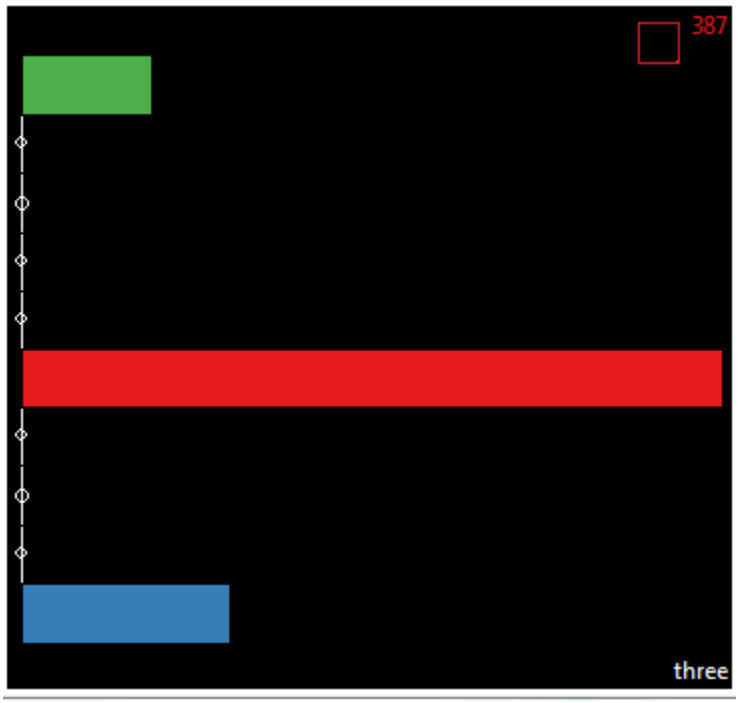


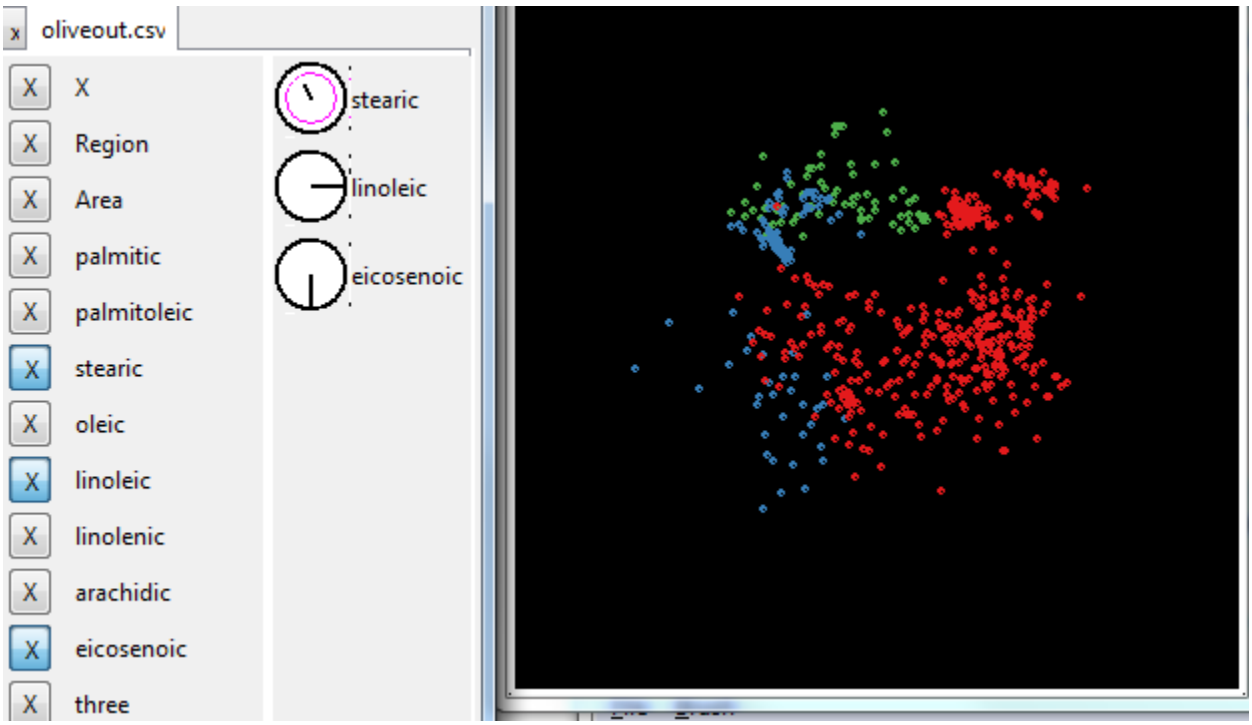
We see that the low and high eicosenic acid content clusters are well-defined, but the separation between low and high linoleic acid content clusters is not easy to detect by clustering, since there is no significant separation between them.

We now scale our data and *"perform a complete-link hierarchical clustering using all acids as explanatory variables."* We plot the resulting dendrogram.

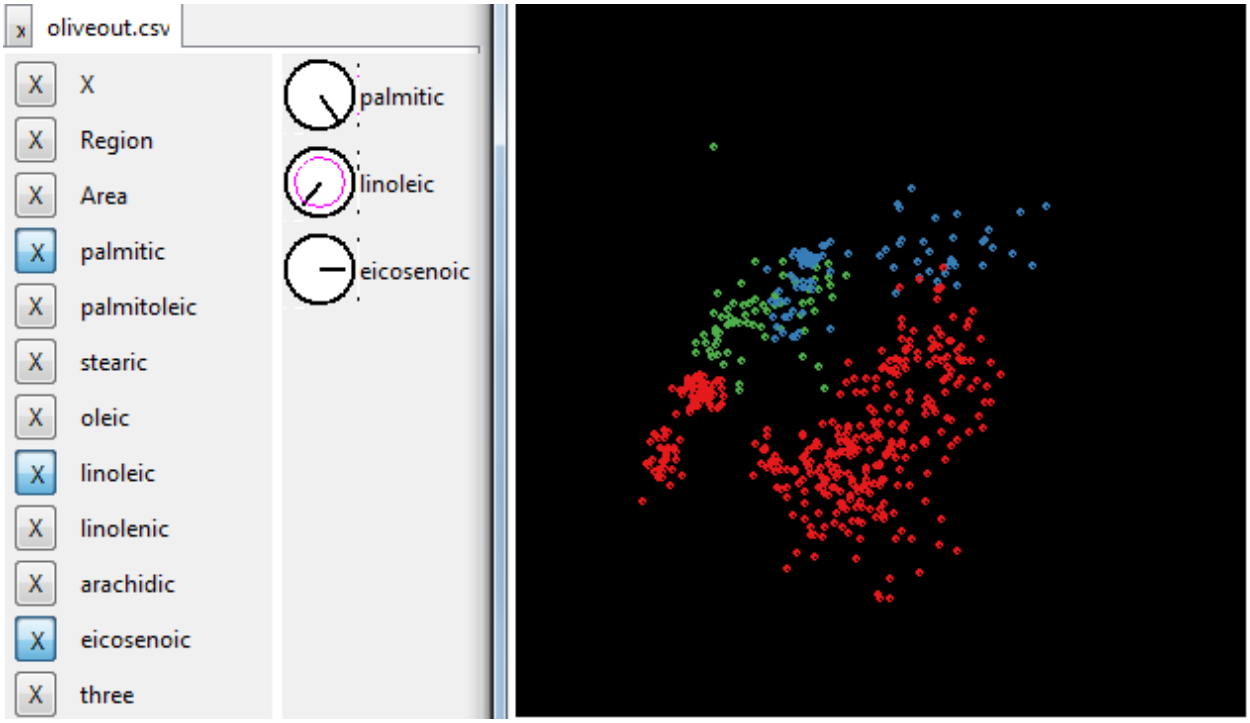


From the picture, I suppose one could say there are about four or five "natural clusters" in the dendrogram. However, the assignment calls for making a cut in the tree at the height where there are three clusters and performing a 2D-tour on the data colored after these three clusters in RGgobi. The first histogram shows the relative number of observations in the three clusters created.

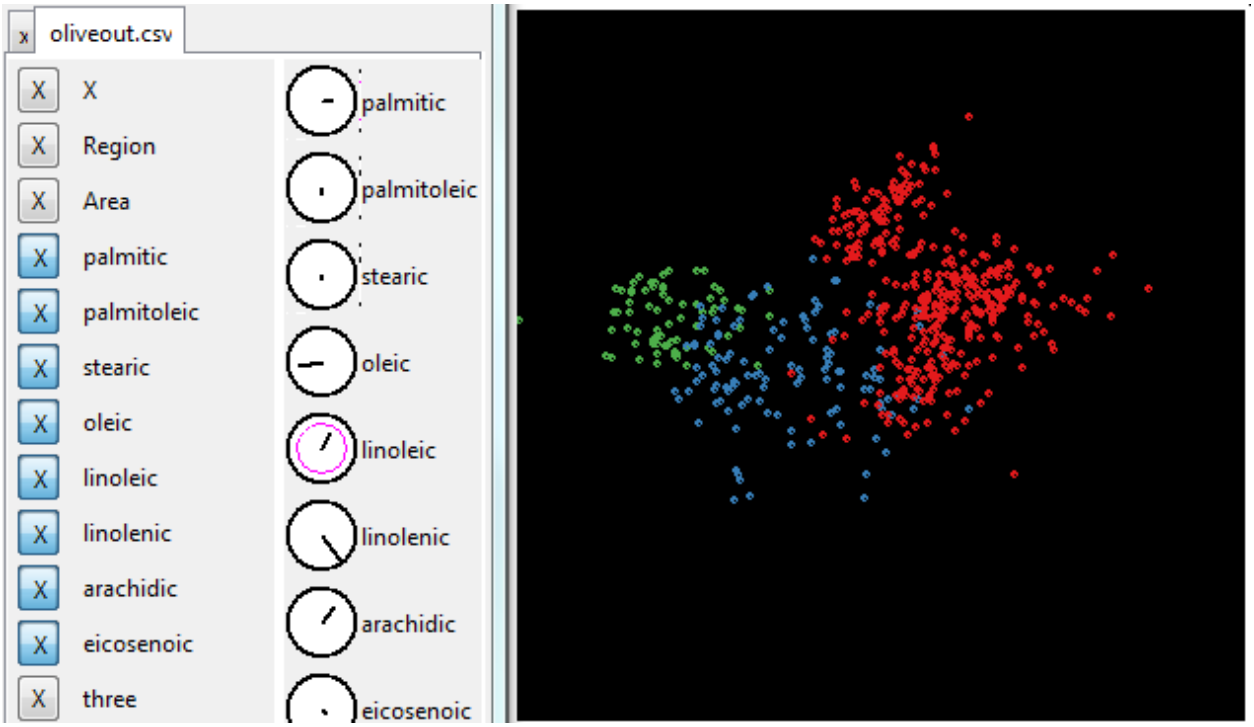




A 2D-tour using three variable acids.



A 2D-tour using three variable acids.

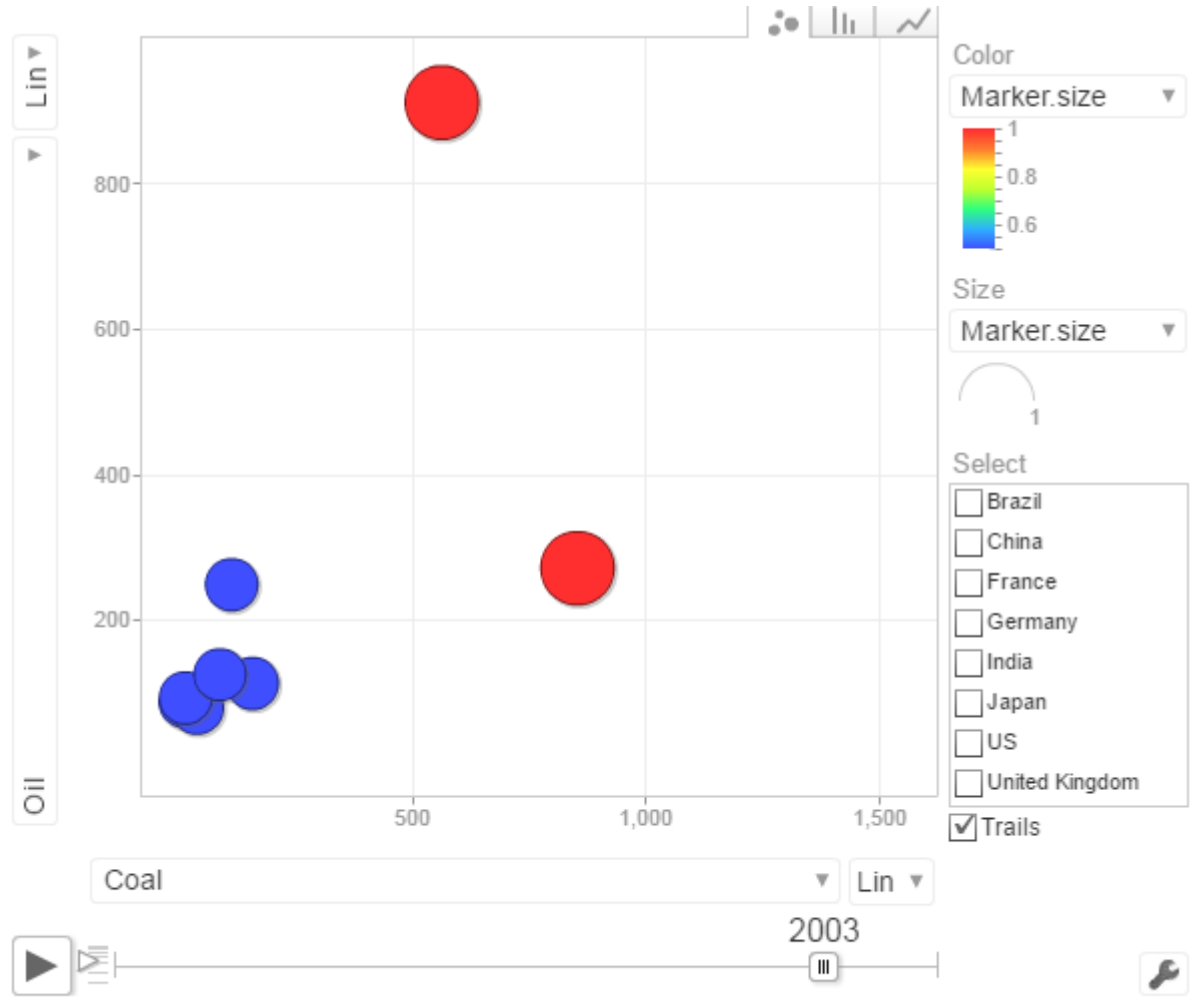


A 2D-tour using all variable acids.

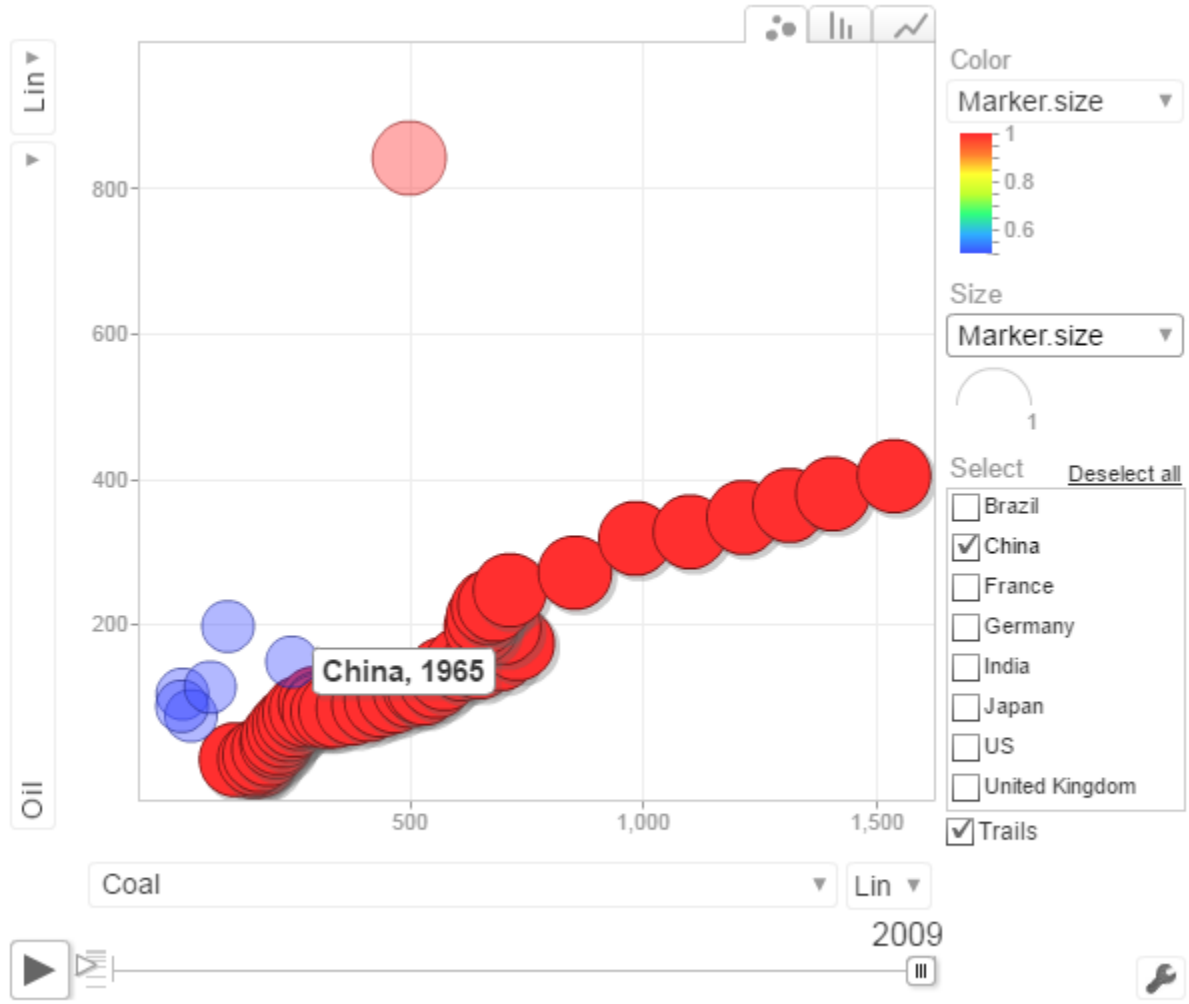
To me, although the result is most likely a proper clustering, this clustering definitely seems less efficient than the one achieved using a decision tree. In fact, one can still clearly make out the distinct clusterings based on linoleic and eicosenic acid content.

Assignment 2

We import the data over total oil and coal consumption in million tonnes oil equivalent in eight large countries from year 1965 to year 2009 from `OilCoal.xls` and create a motion chart using R package `Googlevis`. The large red markers are the U.S. and China, respectively. Now follows the snapshot of starting position and ending position of motion chart, with the China marker traced out in the latter.



At the start of the time series, it seems as if the largest oil consumer was the U.S. The sharp fluctuations downwards in U.S. oil consumption coincides with the inventory-led recession of 73-74 and the escalating inflation countermeasures by the fed around 1981. As the motion chart progresses, one can also make out the dependence of Japan upon oil imports to fuel its economic development boom and later bust during the 70s-00s. The same oil dependence which motivated the entry of Japan into world war two.



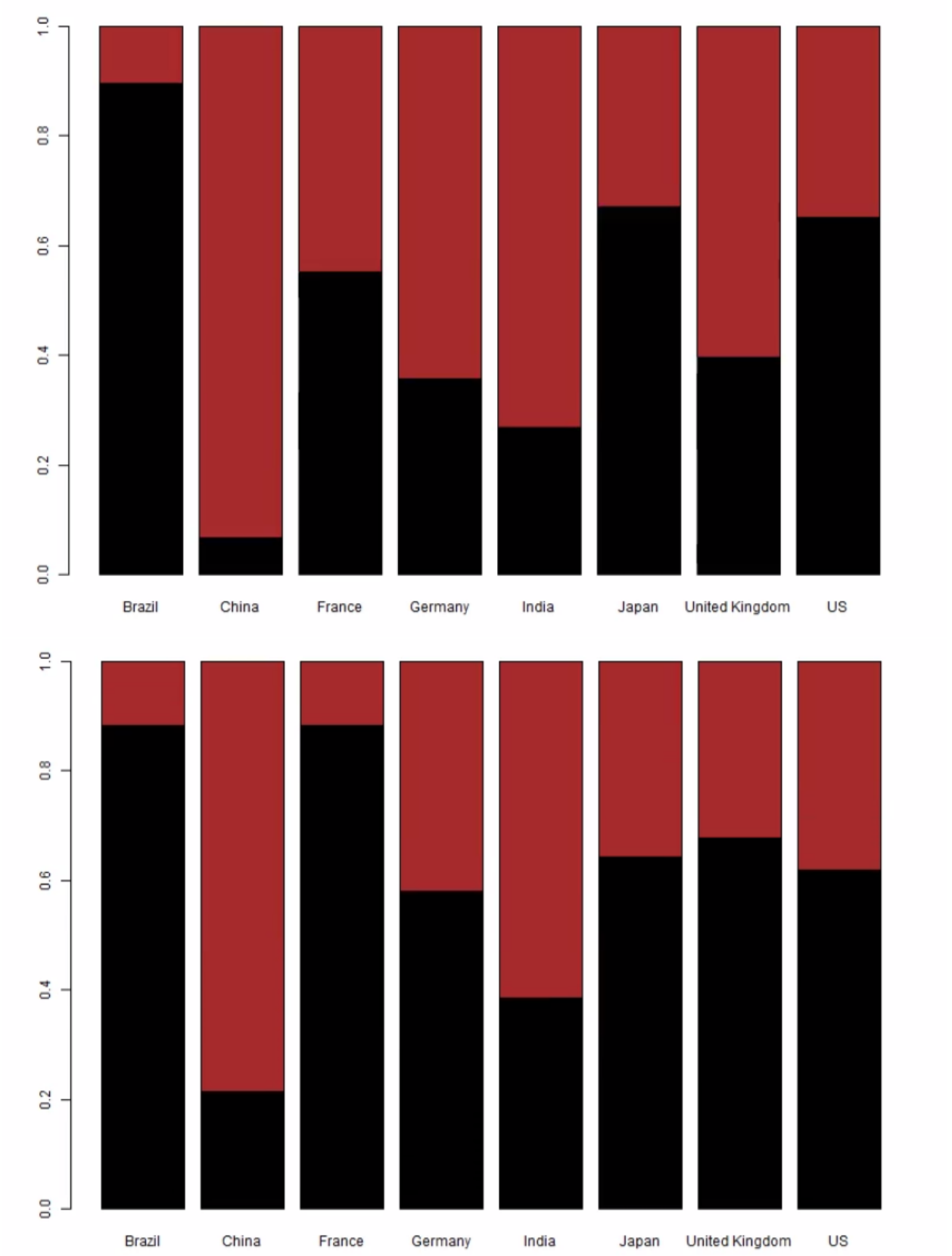
We see that the great rise in oil and coal consumption in China coincides with the country joining the WTO and opening up to FDI, a period when economic growth fueled by fixed asset investment was strong in the country.

France, Germany and the U.K. all developed almost exactly alike over time in terms of oil and coal consumption, which is not surprising given that they were all mature economies and each others' largest trading partners.

I belive it is easier to intuitively grasp the development over time and the sizes of the relative changes in energy consumption in this motionchart compared to a combination of induvidual time series plots per country.

Now we fit a thin plate spline model to the data and choose as the response the percentage of oil consumed out of total oil and coal energy consumption and as the predictors the variables `Year` and `Country`. Then, for each time point, we *"create a bar plot with stacked bars in which one bar corresponds to one country, and each bar is subdivided into percentages of using oil and coal in the country (i.e., each bar has total length 100%)"* and make an animation out of this over time.

Here are the snapshots of said animation at the start and the finish.



Throughout the animation, one can see that the oil/coal consumption ratio for China stays about the same

throughout, despite the previously shown massive increase in both oil and coal consumption. This was not obvious in the motion chart and should count as an advantage. Disadvantage is that it is less interesting to look at an animated barplot compared to an animated motion chart and you will probably not catch the attention of your audience equally well.

R code

```
library(rpart)
library(partykit)
#library(classifly)
#library(rggobi)
library(XLConnect)
library(googleVis)
library(fields)
library(animation)
ani.options(ffmpeg= "C:\\ffmpeg-20161007-c45ba26-win64-static\\bin\\ffmpeg.exe")

olive <-read.csv("olive.csv")
olive$Region <- as.factor(olive$Region)
tree <- rpart(Region ~ palmitic + palmitoleic + stearic +
                oleic + linoleic + linolenic + arachidic +
                eicosenoic, data = olive)
plot(as.party(tree))
#classifly(olive,Region ~ palmitic + palmitoleic + stearic +
#            oleic + linoleic + linolenic + arachidic +
#            eicosenoic,rpart)
olivescaled <- scale(olive[,4:11])
olive_dist <- dist(olivescaled)
olive_dend <- hclust(olive_dist, method = "complete")
plot(olive_dend)
threeclusters <- cutree(olive_dend, k = 3)
olive$three <- as.factor(threeclusters)
#write.csv(olive,file = "oliveout.csv")
wb = loadWorkbook("Oilcoal.xls")
oilcoaldata <- readWorksheet(wb, sheet = 1, header = TRUE)
oilcoaldata$Year <- as.numeric(oilcoaldata$Year)
#hehe <- gvisMotionChart(oilcoaldata,
#                        idvar='Country', timevar='Year',
#                        sizevar = "Marker.size")
#plot(hehe)
oilp <- 100 * (oilcoaldata$Oil / (oilcoaldata$Oil + oilcoaldata$Coal))
preds <- cbind(oilcoaldata$Year,as.factor(oilcoaldata$Country))
thinplate <- Tps(preds,oilp)

#saveVideo({
#  for( i in 1:176 ){
#    oneloop <- predict.Krig(thinplate, x = matrix(nrow = 8, c(rep(1965 + .25 * i,8),
#                                                              1:8)))
#    rownames(oneloop) <- levels(as.factor(oilcoaldata$Country))
#    oneloop <- cbind(oneloop, 100-oneloop)
#    oneprop = prop.table(oneloop,margin=1)
#    barplot(t(oneprop), col=c("black","brown"))
```

```
#  }
#}, video.name = "C:\\Users\\Dator\\Documents\\R_HW\\visualization\\anime.mp4",
#interval = 0.05,ani.width = 800,ani.height = 600
#)
## NA
```