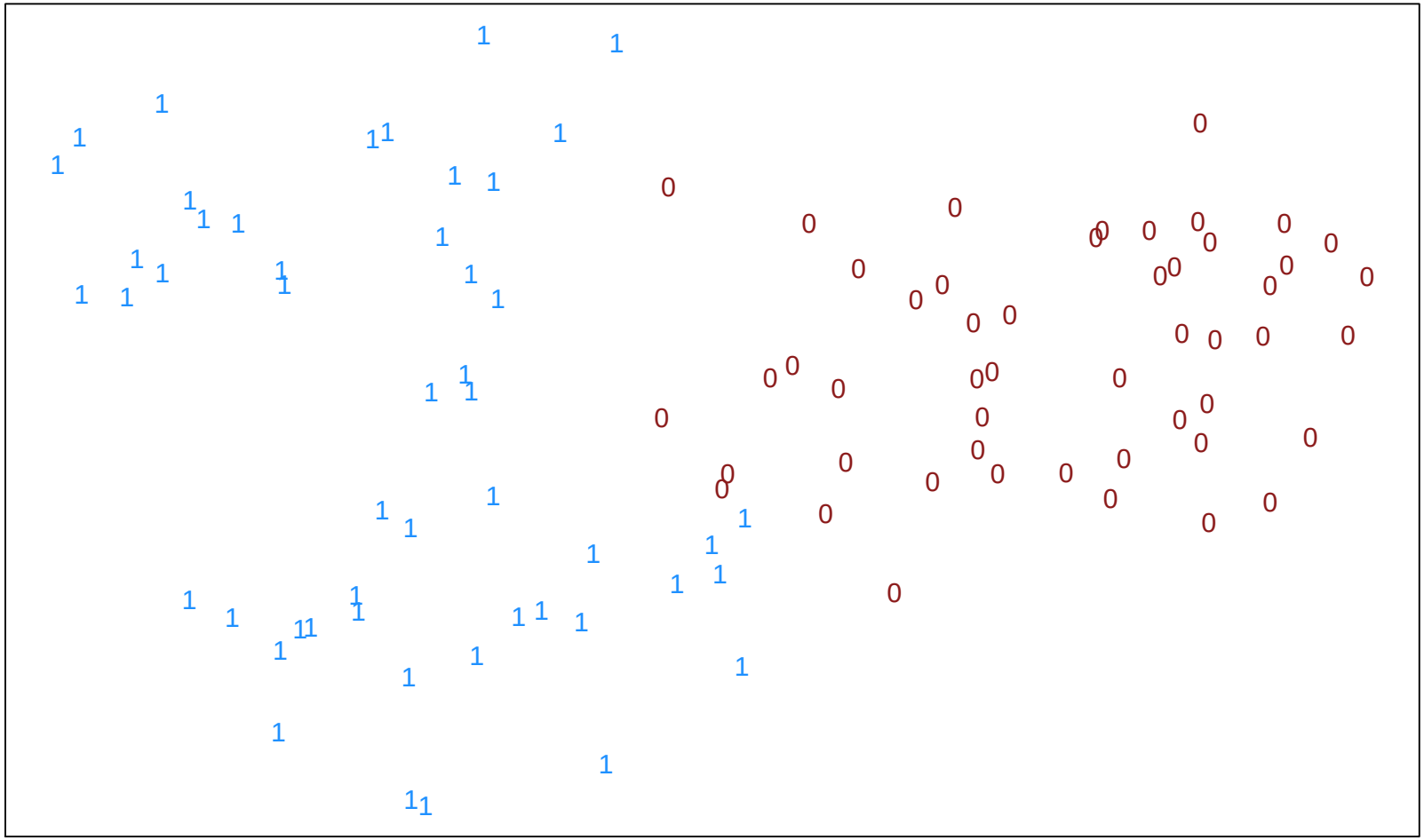
STAT 408 - Statistical Learning Clustering

April 3, 2018

Unsupervised Learning

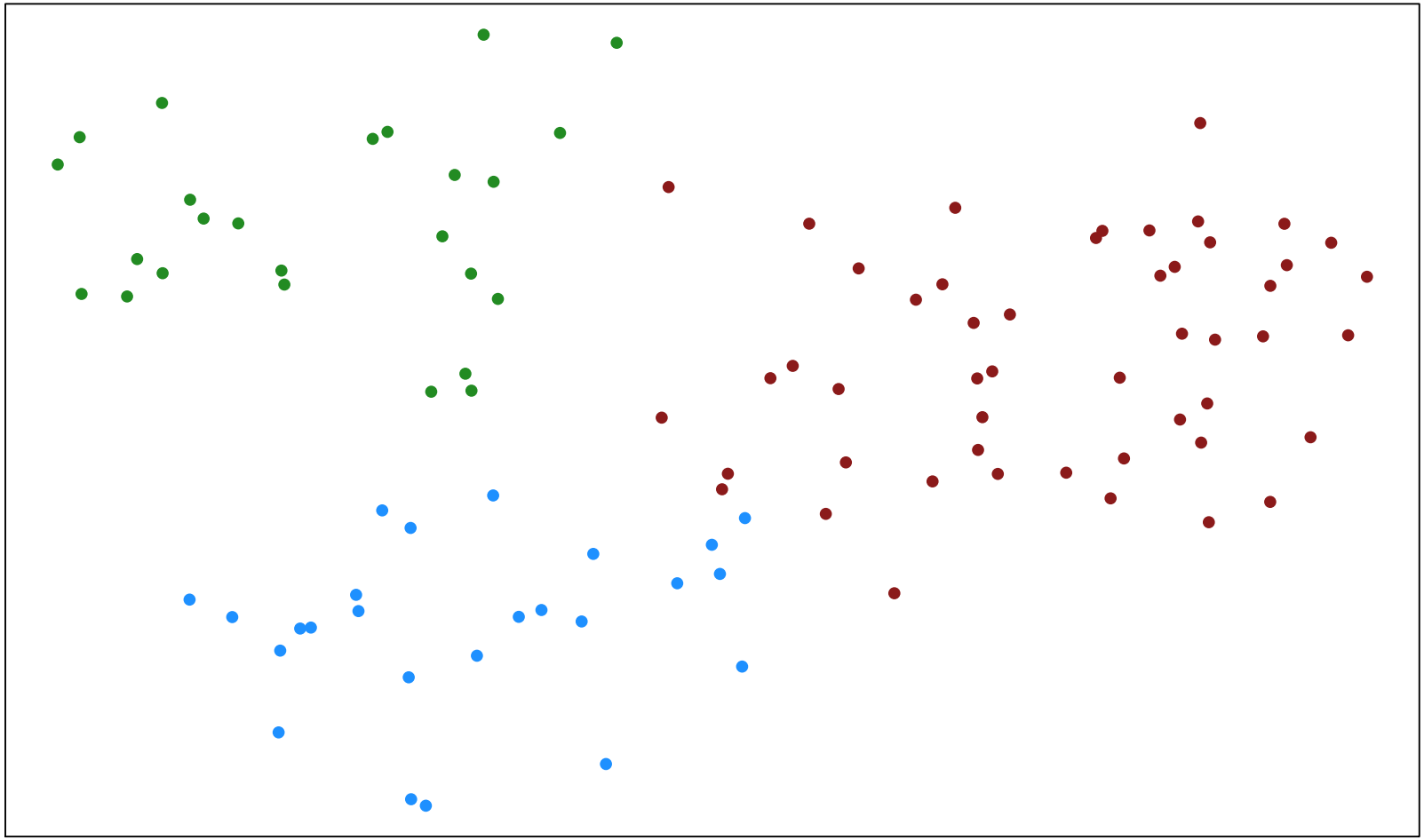
Supervised vs. Unsupervised Learning

Supervised

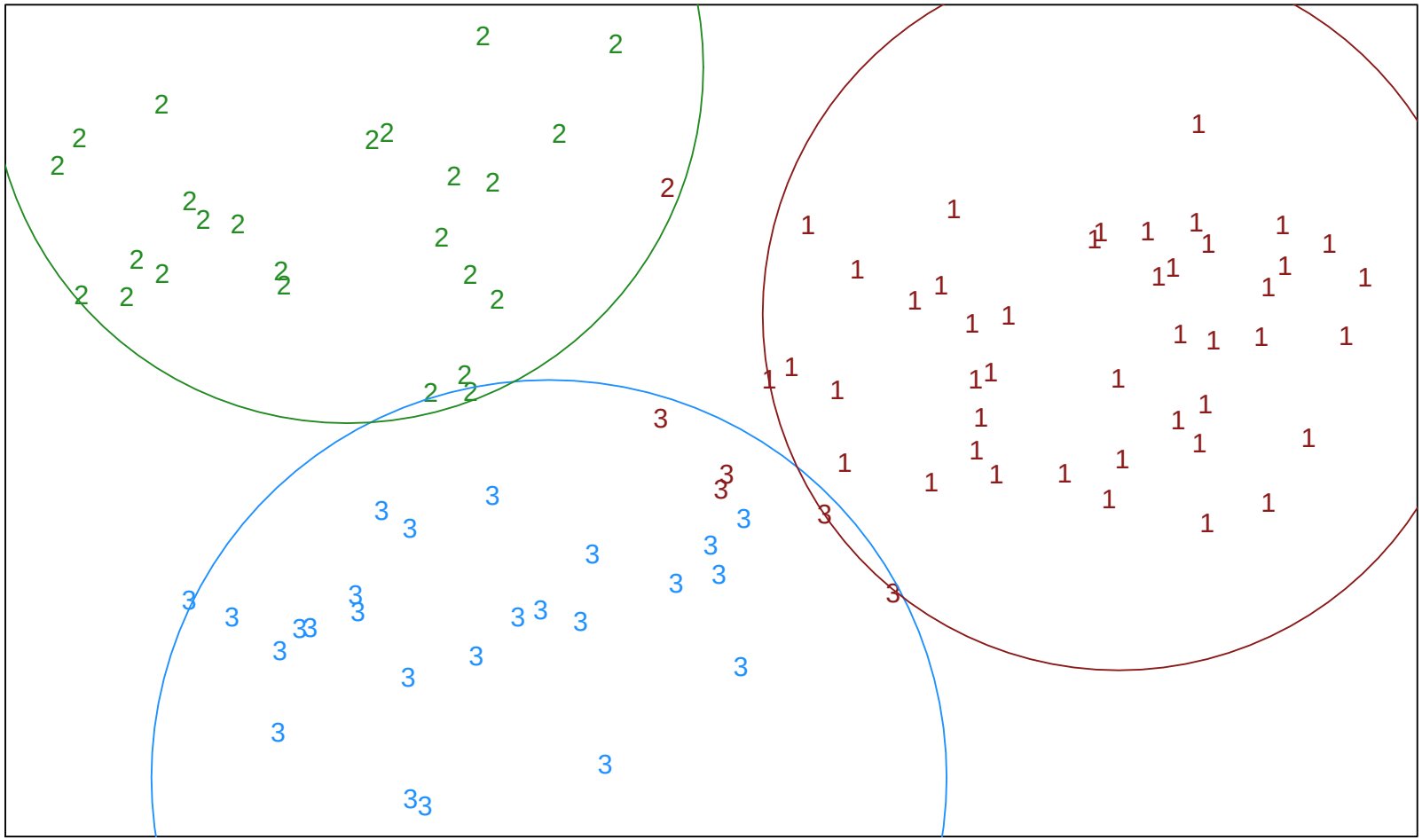


Unsupervised - How many clusters?

Unsupervised



K-MEANS CLUSTERING



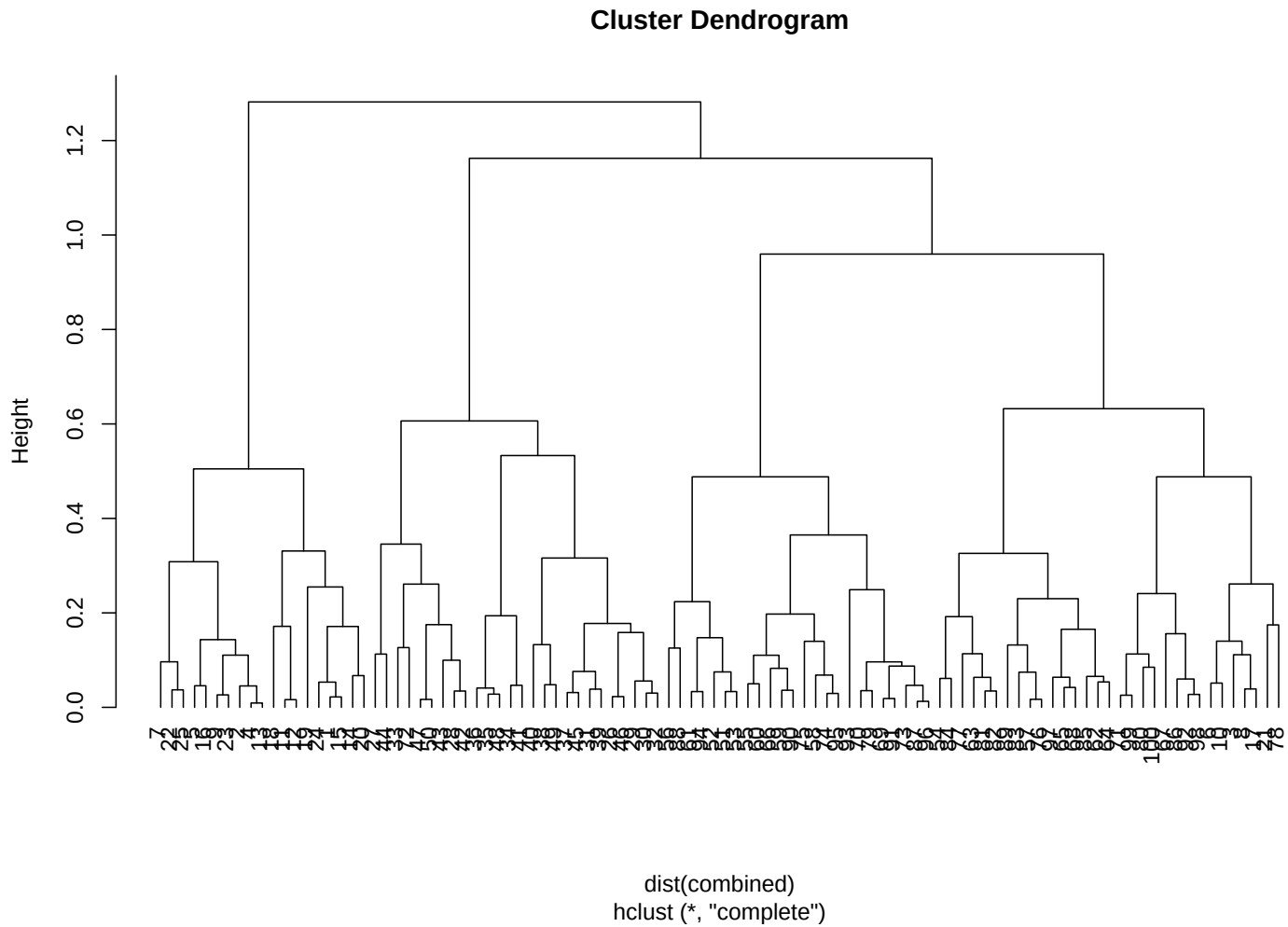
K-MEANS CLUSTERING

```
## K-means clustering with 3 clusters of sizes 44, 26, 30
##
## Cluster means:
##         [,1]      [,2]
## 1 0.7693055 0.6026478
## 2 0.1329025 0.8018608
## 3 0.2844651 0.1830810
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2
##  [36] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3
##  [71] 3 2 1 1 1 1 1 3 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1
##
## Within cluster sum of squares by cluster:
## [1] 1.915896 1.175133 1.417926
##  (between_SS / total_SS =  75.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

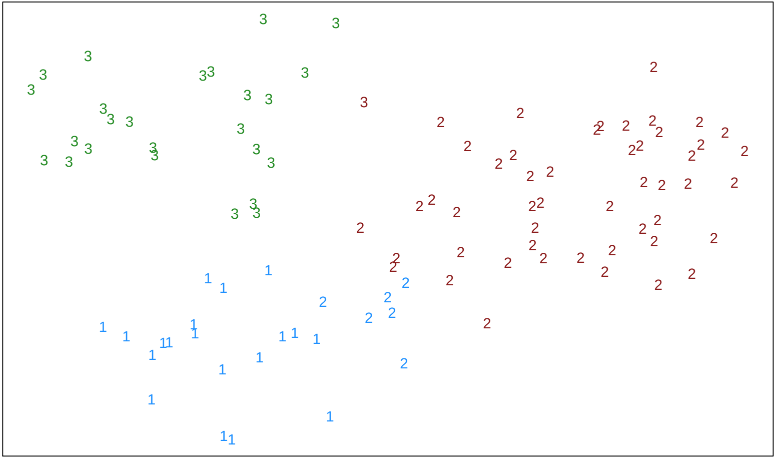
K-MEANS CLUSTERING - CODE


```
km <- kmeans(combined, 3)
plot(combined,type='n',axes=F, xlab='',ylab='')
box()
points(combined,pch=as.character(km$cluster),
       col=c(rep('dodgerblue',25),
             rep('forestgreen',25),
             rep('firebrick4',50)))
draw.circle(.31,-0.1,.335, border='dodgerblue')
draw.circle(.79,.65,.3, border='firebrick4')
draw.circle(.14,1.05,.3, border='forestgreen')
```

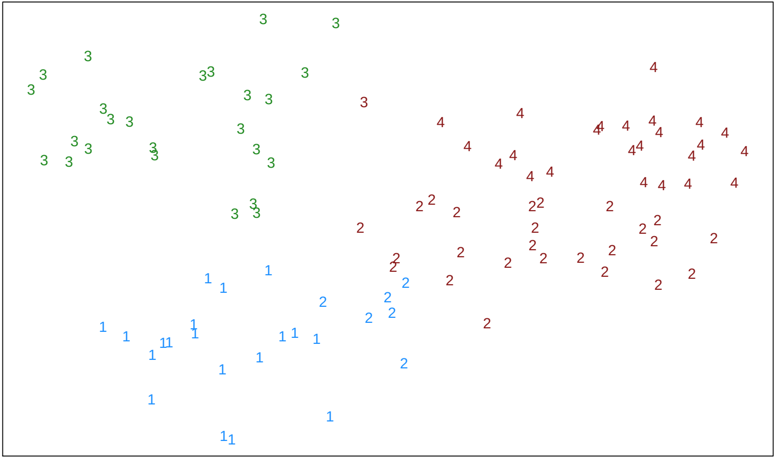
Hierarchical clustering



Hierarchical clustering - with 3 clusters



Hierarchical clustering - with 4 clusters



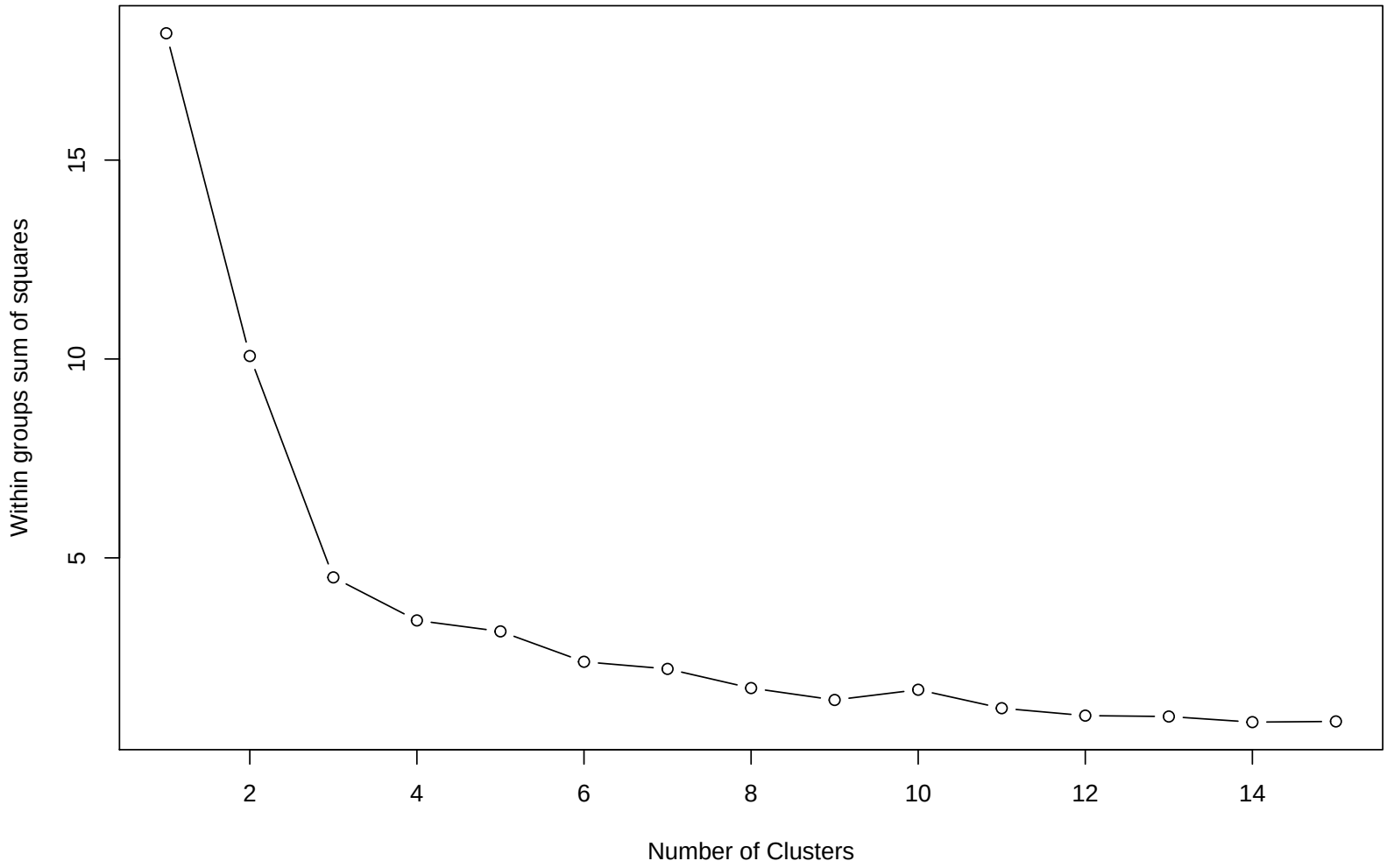
Hierarchical clustering - code


```
hc <- hclust(dist(combined))
plot(hc, hang=-1)
plot(combined,type='n',axes=F, xlab='',ylab='')
box()
points(combined,pch=as.character(cutree(hc,4)),
       col=c(rep('dodgerblue',25),
             rep('forestgreen',25),
             rep('firebrick4',50)))
```

How to choose the number of clusters?

Given these plots that we have seen, how do we choose the *appropriate* number of clusters?

How to choose the number of clusters? - Scree plot



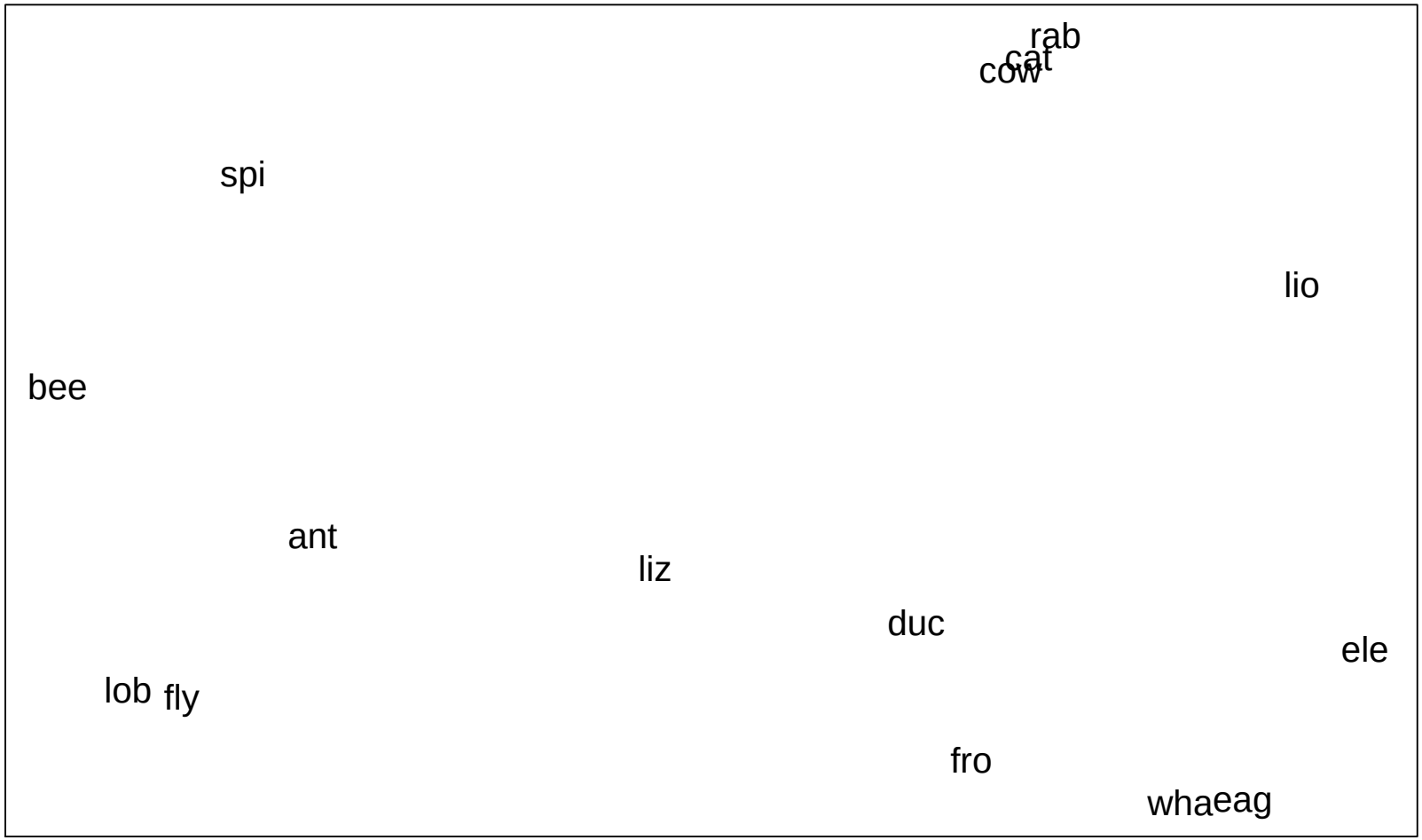
Scree plot - code


```
wss <- rep(0,15)
for (i in 1:15) {
  wss[i] <- sum(kmeans(combined,centers=i)$withinss)
}
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")
```

Data with more than 2 dimensions

	warm-blooded	can fly	vertebrate	endangered	have hair
ant	No	No	No	No	No
bee	No	Yes	No	No	Yes
cat	Yes	No	Yes	No	Yes
cow	Yes	No	Yes	No	Yes
duc	Yes	Yes	Yes	No	No
eag	Yes	Yes	Yes	Yes	No
ele	Yes	No	Yes	Yes	No
fly	No	Yes	No	No	No
fro	No	No	Yes	Yes	No
lio	Yes	No	Yes	Yes	Yes
liz	No	No	Yes	No	No
lob	No	No	No	No	No
rab	Yes	No	Yes	No	Yes
spi	No	No	No	No	Yes
wha	Yes	No	Yes	Yes	No

Multidimensional Scaling



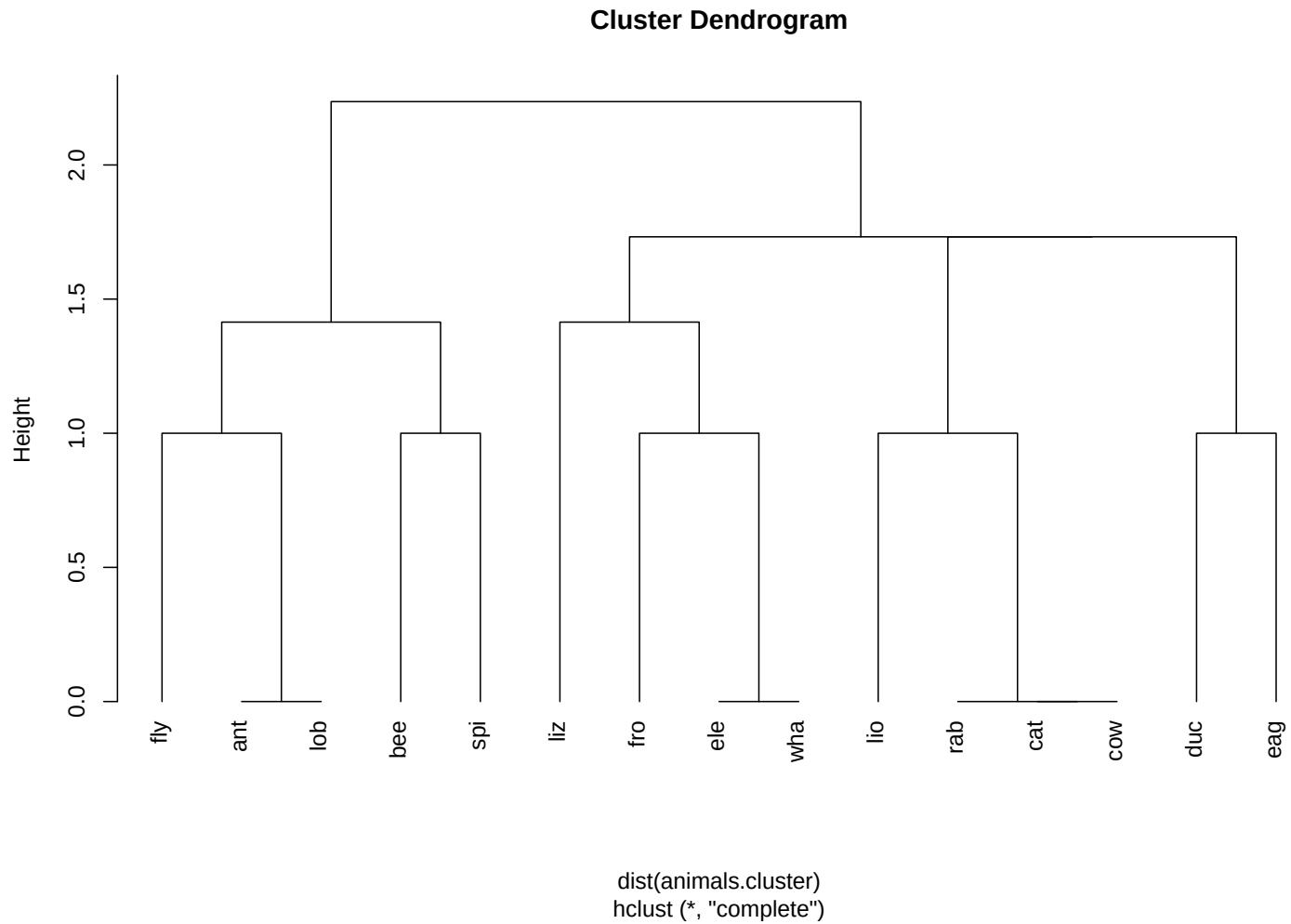
MDS - Code


```
animals <- cluster::animals

colnames(animals) <- c("warm-blooded",
                       "can fly",
                       "vertebrate",
                       "endangered",
                       "live in groups",
                       "have hair")
animals.cluster <- animals[,-(5)]
animals.cluster <- animals.cluster[-c(4,5,12,16,18),]
animals.cluster[10,4] <- 2
animals.cluster[14,4] <- 1

d <- dist(animals.cluster)
fit <- cmdscale(d, k=2)
fit.jitter <- fit + runif(nrow(fit*2),-.15,.15)
plot(fit.jitter[,1], fit.jitter[,2], xlab="", ylab="",   main="",   type
box()
text(fit.jitter[,1], fit.jitter[,2], labels = row.names(animals.cluster)
```

Hierarchical Clustering of Animals



Lecture Exercise: Clustering Zoo Animals

Use the dataset create below for the following questions.

```
zoo.data <- read.csv('http://www.math.montana.edu/ahoegh/teaching/stat408/datasets/ZooClean.csv'
rownames(zoo.data) <- zoo.data[,1]
zoo.data <- zoo.data[,-1]
```

- Use multidimensional scaling to visualize the data in two dimensions.
- What are two animals that are very similar and two that are very different?
- Create a hierachical clustering object for this dataset.

Why are a leopard and raccoon clustered together for any cluster size?

- Now add colors corresponding to four different clusters to your MDS plot.

Interpret what each of the four clusters correspond to.