Clustering

28-29/05

Clustering is a form of **unsupervised learning**.

In cluster analysis, observations are placed into groups in such a way that the observations in each group are similar in some way, and different from the other groups. It is hoped that these groups can be interpreted in some meaningful way.

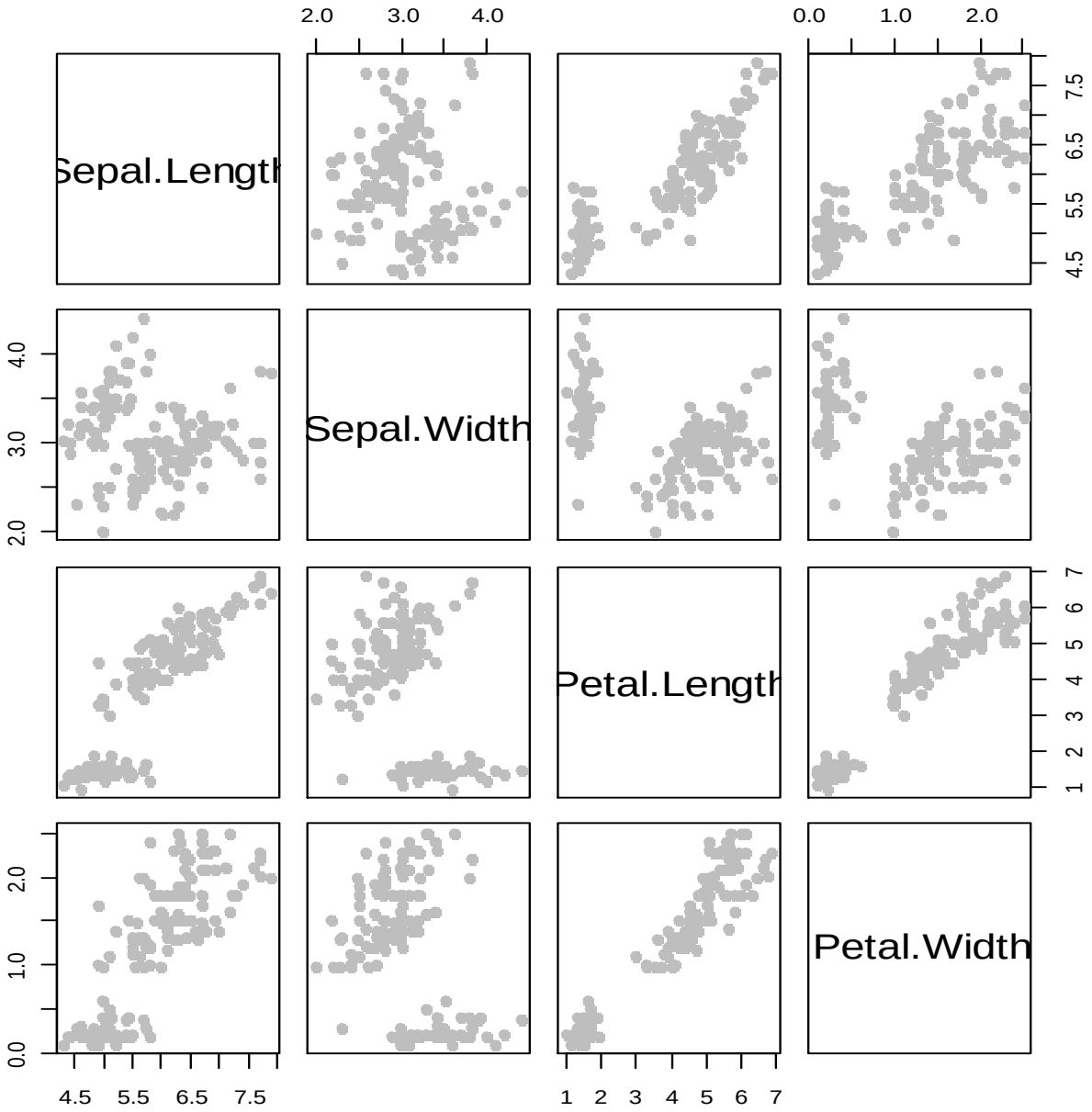
If this sounds vague and wishy-washy … it is!

Why do it?

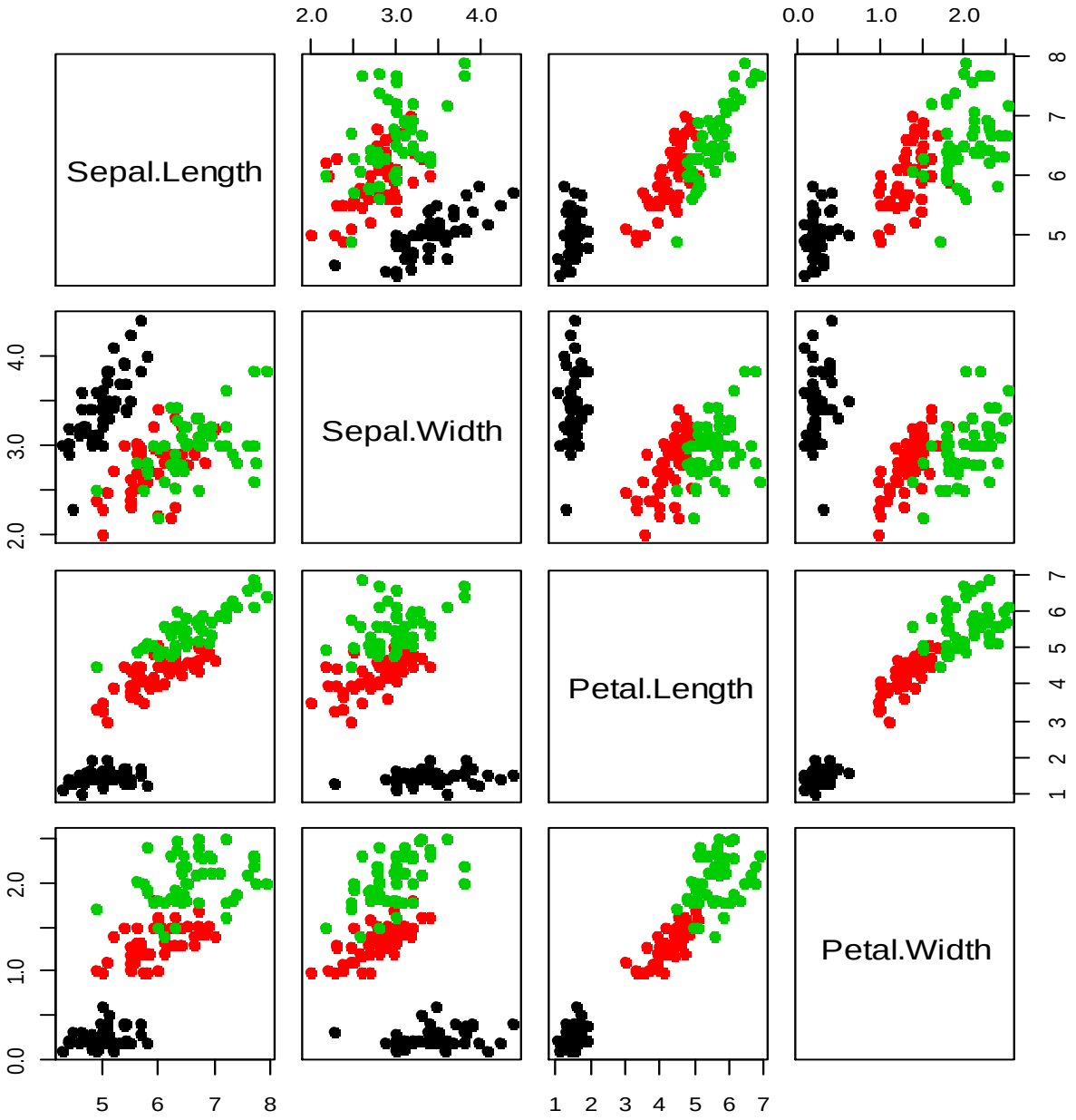
Clustering problems come up when you try to understand data. Examples:

- Market segmentation in advertising.
- Association rules for supermarkets.
- Communities in criminal networks.
- Grouping documents into topics.

Just like PCA, the results of clustering are very much open to interpretation. There is no guarantee that the clusters found will be meaningful or that they reflect properties of the *population (test data)* rather than the *sample (training data.)*



How many clusters do you think there are?


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0
3.0
4.0
0.0
1.0
2.0
5
6
7
8
2.0
3.0
4.0
1
2
3
4
5
6
7
0.0
1.0
2.0
5
6
7
8
1
2
3
4
5
6
7

Strategy: find some measure of how good a clustering is. Try all possible assignments of data to clusters, and choose the one that maximises the measure.

Problem: there are lots of ways of assigning data to clusters. You can't examine all of them.

Solution: use some sort of systematic method to find a reasonably good clustering, with no guarantee that it will be the best.

k-means clustering (Bell Labs, 1957)

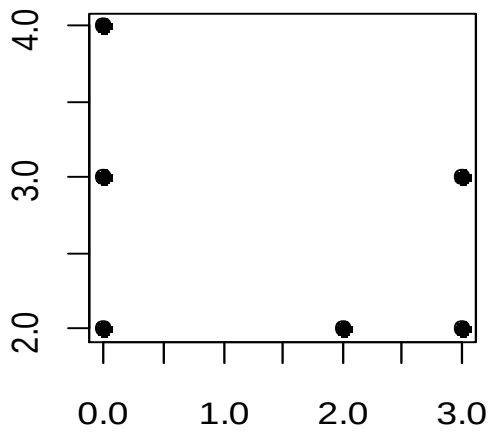
- Number of clusters, k, fixed in advance.
- Randomly assign data to clusters.
- *(1) Calculate the mean of each cluster.*
- *(2) Assign each data point to the cluster with the nearest mean.*
- Repeat (1) & (2) until convergence.
- [Recommended: do the whole thing many times with repeated random assignments.]

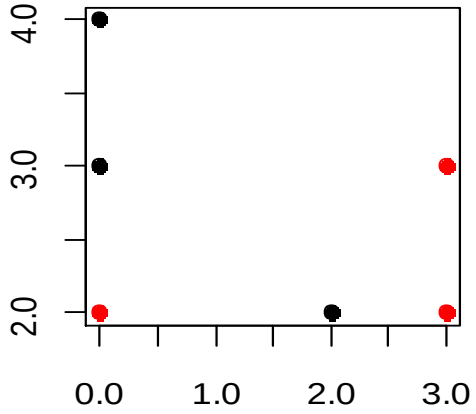
Simple example, by hand:

Data: (2,2), (3, 3), (3, 2), (0, 3), (0, 2), (0, 4)

Choose k=2.

Randomly assign points to clusters.



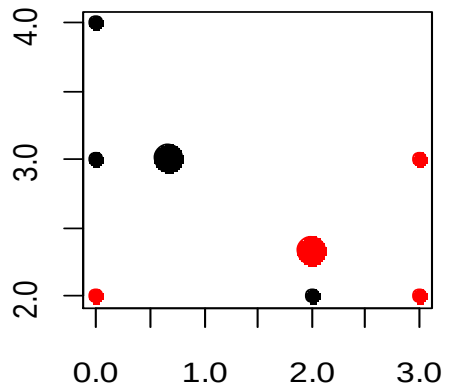


Data: (2,2), (3, 3), (3, 2), (0, 3), (0, 2), (0, 4)

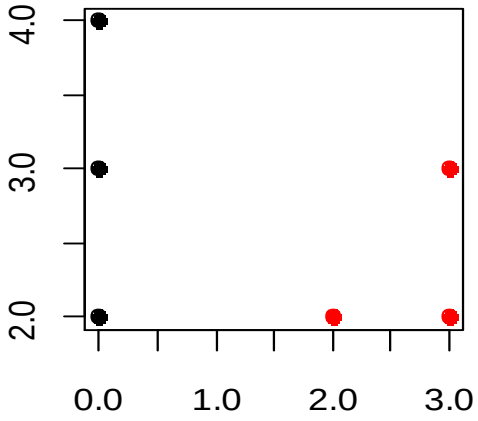
Cluster means:

((2,2)+(0,3)+(0,4))/3 = (2/3,3)

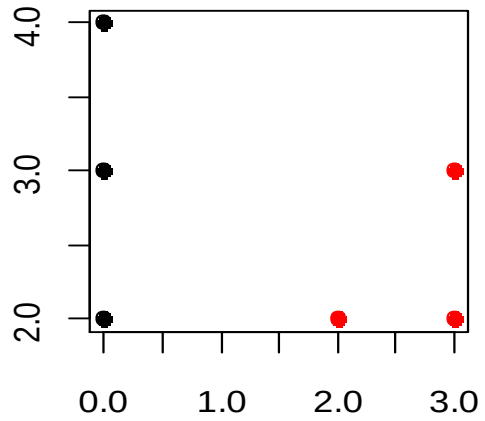
((3,3)+(3,2)+(0,2))/3 = (2, 7/3)



Assign each point to nearest cluster mean:

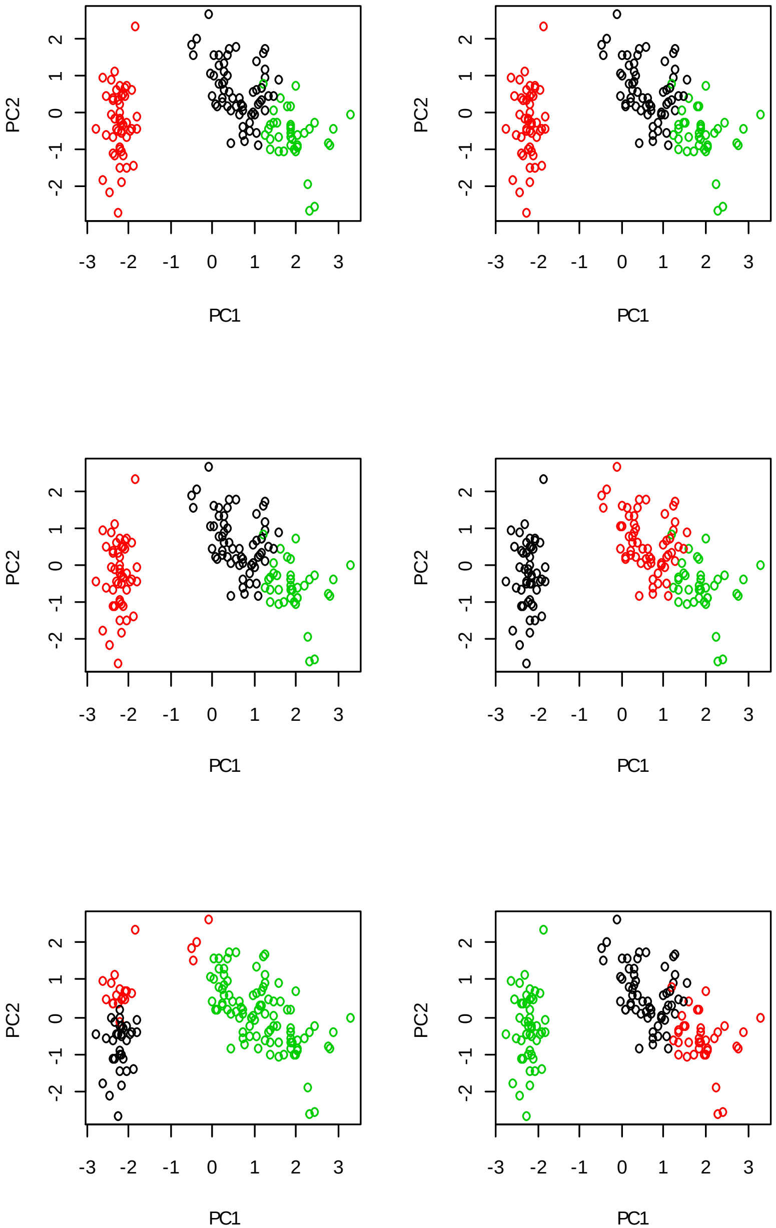


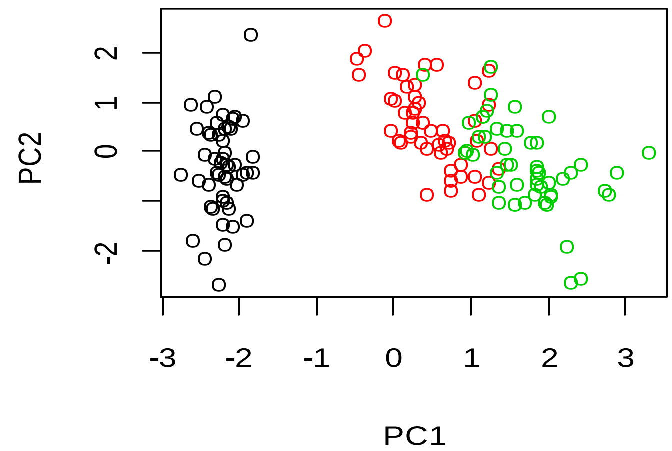
Repeat and you get the same clustering, so the algorithm has converged. End result:



This algorithm finds a **local minimum** of the quality function

$$\sum\sum(\boldsymbol{x}_i - \boldsymbol{\mu}_i)^2$$





Actual species for comparison. Note that the colours are irrelevant.

Type	*FFD*	*SPR*	*RGF*	*PLF*	*SLF*	*CAR*
FH-1	82	1.468	3.30	0.166	0.10	0
FJ-1	89	1.605	3.64	0.154	0.10	0
F-86A	101	2.168	4.87	0.177	2.90	1
F9F-2	107	2.054	4.72	0.275	1.10	0
F-94A	115	2.467	4.11	0.298	1.00	1
F3D-1	122	1.294	3.75	0.150	0.90	0
F-89A	127	2.183	3.97	0.000	2.40	1
XF10F-1	137	2.426	4.65	0.117	1.80	0
F9F-6	147	2.607	3.84	0.155	2.30	0
F-100A	166	4.567	4.92	0.138	3.20	1
F4D-1	174	4.588	3.82	0.249	3.50	0
F11F-1	175	3.618	4.32	0.143	2.80	0
F-101A	177	5.855	4.53	0.172	2.50	1
F3H-2	184	2.898	4.48	0.178	3.00	0
F-102A	187	3.880	5.39	0.101	3.00	1
F-8A	189	0.455	4.99	0.008	2.64	0
F-104B	194	8.088	4.50	0.251	2.70	1
F-105B	197	6.502	5.20	0.366	2.90	1
YF-107A	201	6.081	5.65	0.106	2.90	1
F-106A	204	7.105	5.40	0.089	3.20	1
F-4B	255	8.548	4.20	0.222	2.90	0
F-111A	328	6.321	6.45	0.187	2.00	1

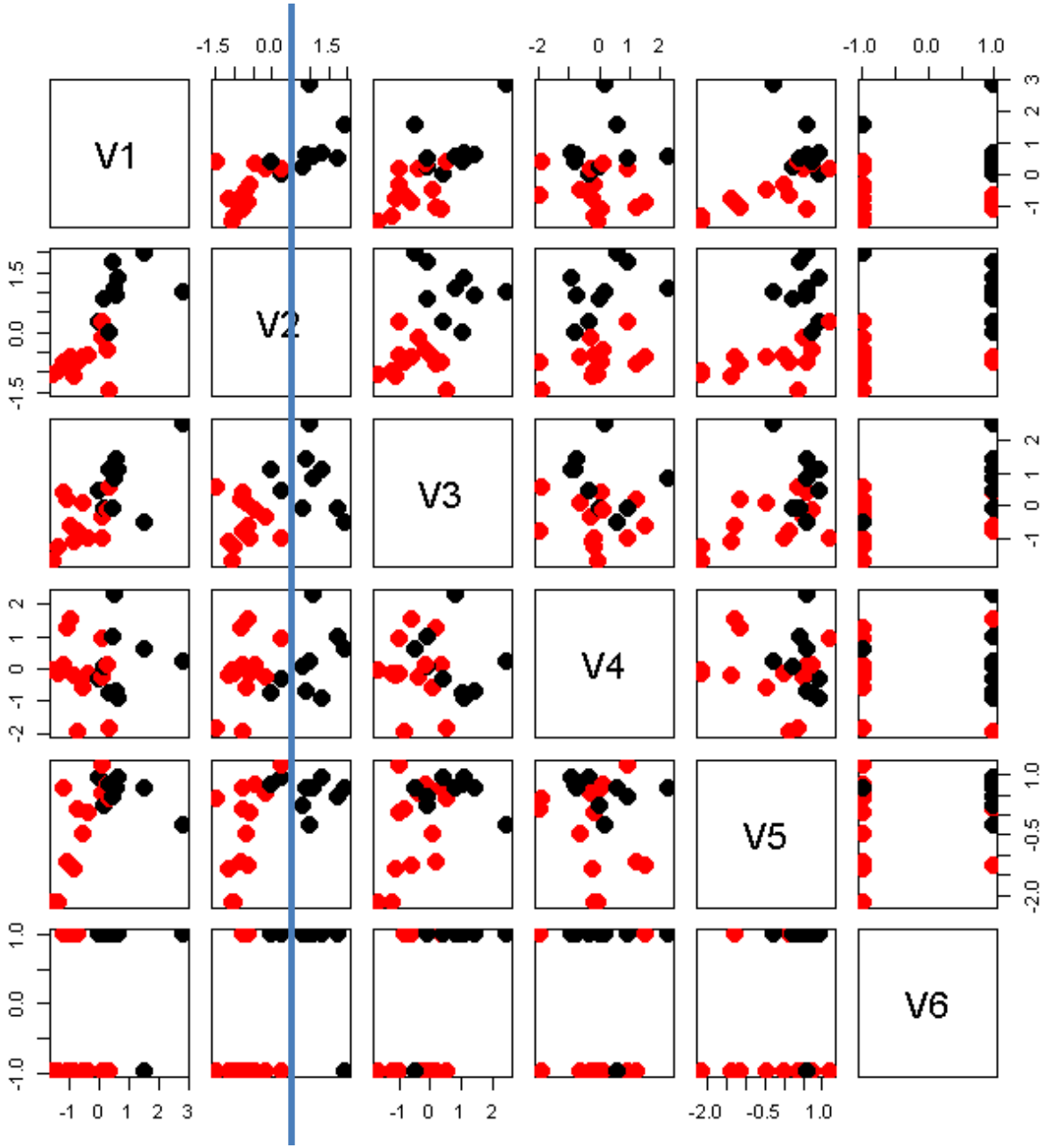
FFD first flight date, in months after January 1940
SPR specific power, proportional to power per unit weight
RGF flight range factor
PLF payload as a fraction of gross weight of aircraft
SLF sustained load factor
CAR a binary variable which takes the value 1 if the aircraft can land on a carrier, 0 otherwise.

V1
V2
V3
V4
V5
V6

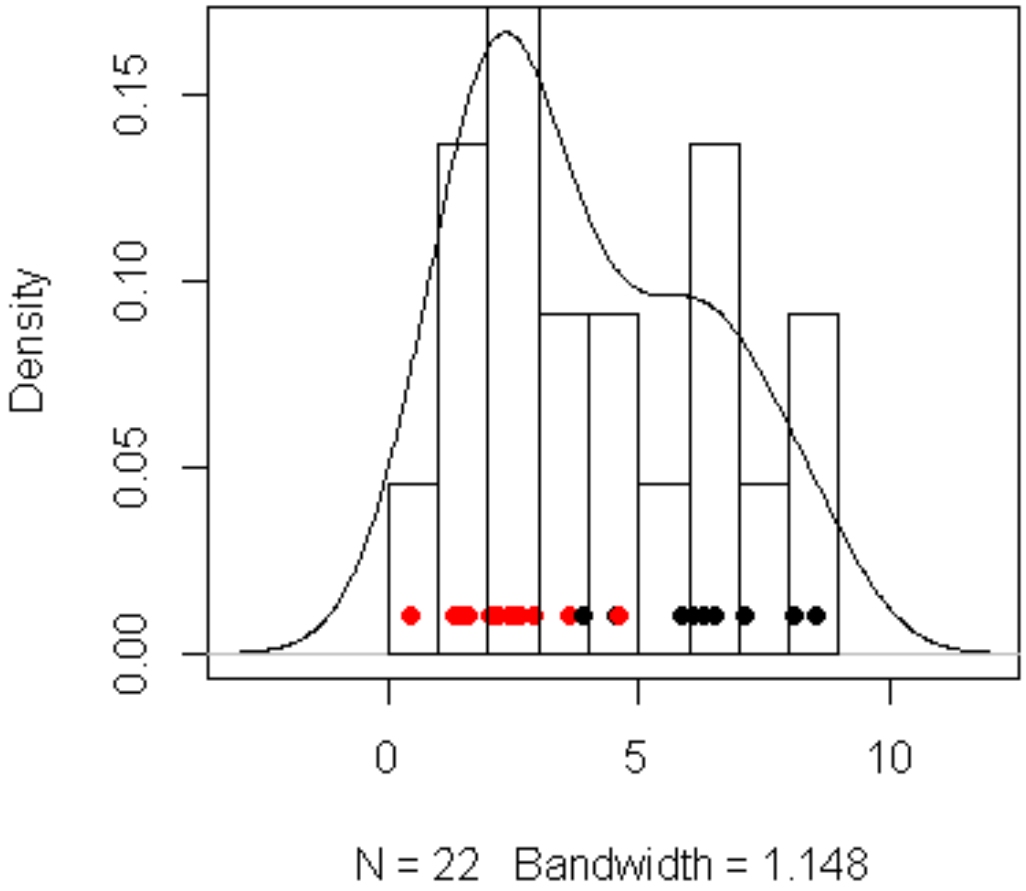
PC2
PC1

pr2
Variances

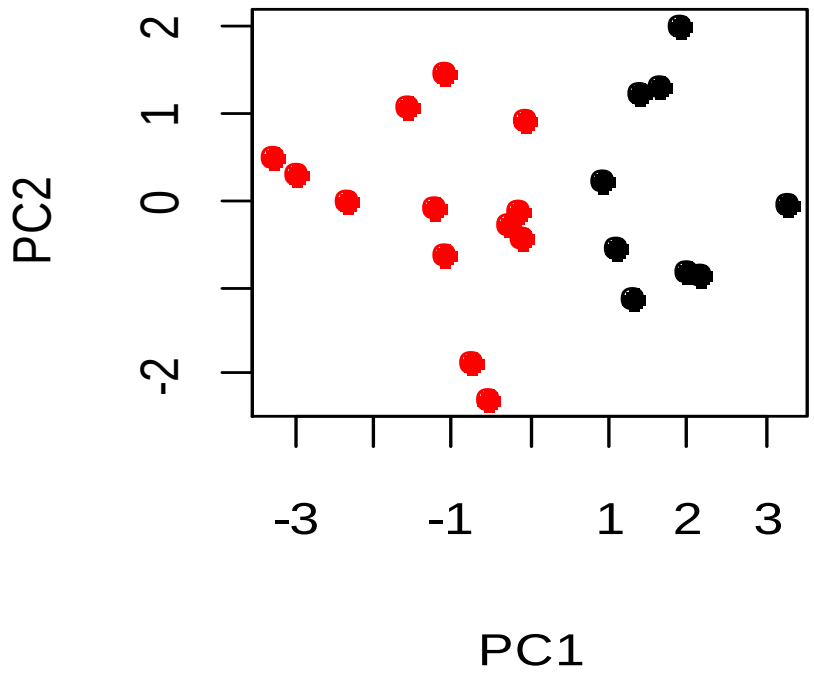
```
cl2 <-
kmeans(jets2, 2,
nstart=50)

pairs(jets2,
col=cl2$cluster,
pch=19, cex=2)
```



SPR



Discussion: is this meaningful in any way?



	PC1	PC2
V1	0.487	-0.008
V2	0.476	0.342
V3	0.463	-0.223
V4	0.055	0.889
V5	0.422	-0.201
V6	0.374	-0.054

Whisky data (after Gloukhov 2013): 86 obs of 13 variables

```
## Distillery Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey
## 4 Ardbeg 4 1 4 4 0 0 2 0 ..
## 22 Caol Ila 3 1 4 2 1 0 2 0 ..
## 24 Clynelish 3 2 3 3 1 0 2 0 ..
## 58 Lagavulin 4 1 4 4 1 0 1 2 ..
## 59 Laphroig 4 2 4 4 1 0 0 1 ..
```

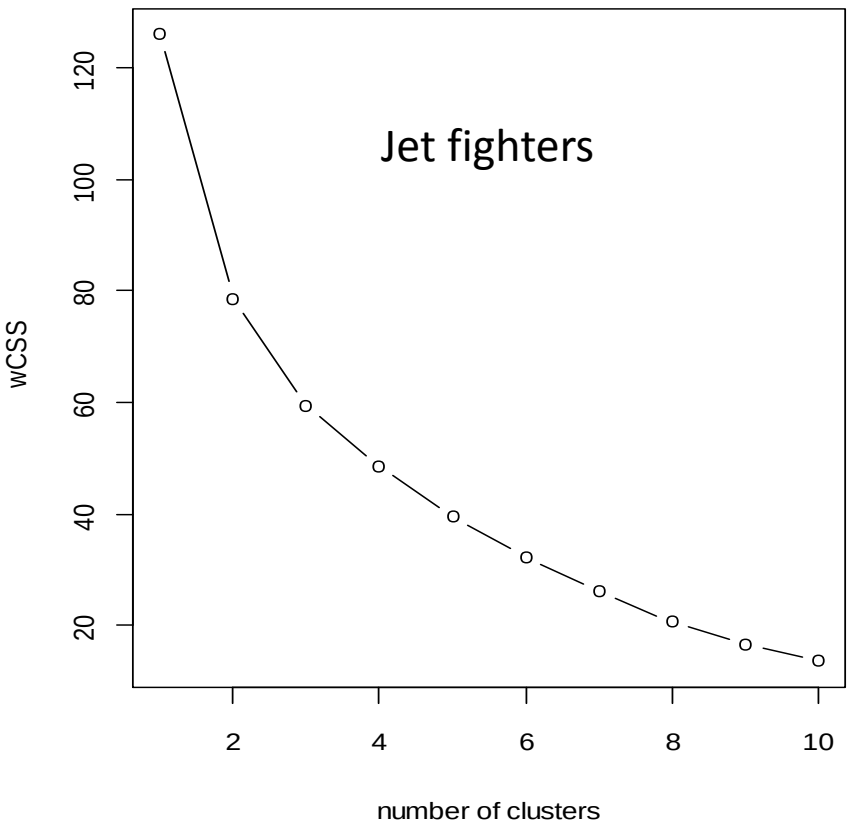
Cluster by flavour profile. (Scale all variables to [0,1] and use Euclidean diatance)

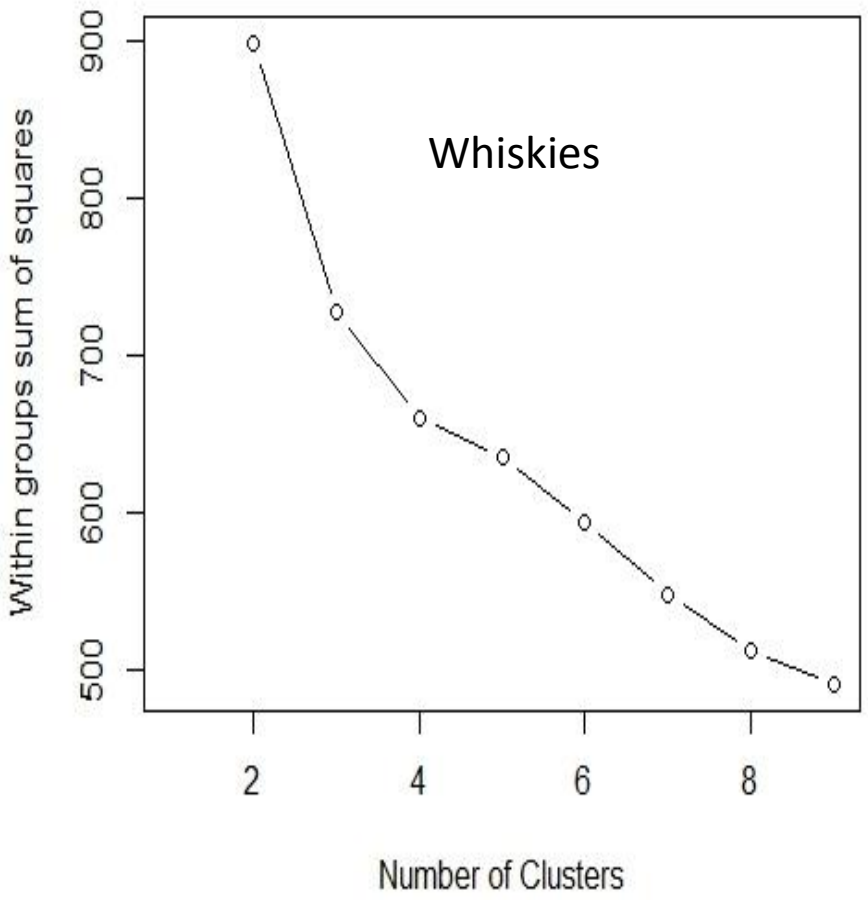
Typical whisky from each cluster:

```
##      Distillery Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey
## 42 Glenallachie    1         3     1         0       0     1     1     0
## 70 RoyalBrackla    2         3     2         1       1     1     2     1
## 1     Aberfeldy    2         2     2         0       0     2     1     2
## 4        Ardbeg    4         1     4         4       0     0     2     0
##    Nutty Malty Fruity Floral   Postcode Latitude Longitude fit.cluster
## 42     1     2      2      2   AB38 9LR   326490    841240           1
## 70     0     2      3      2   IV12 5QY   286040    851320           2
## 1      2     2      2      2  \tPH15 2EB   286580    749680           3
## 4      1     2      1      0  \tPA42 7EB   141560    646220           4
```

How to choose the number of clusters?

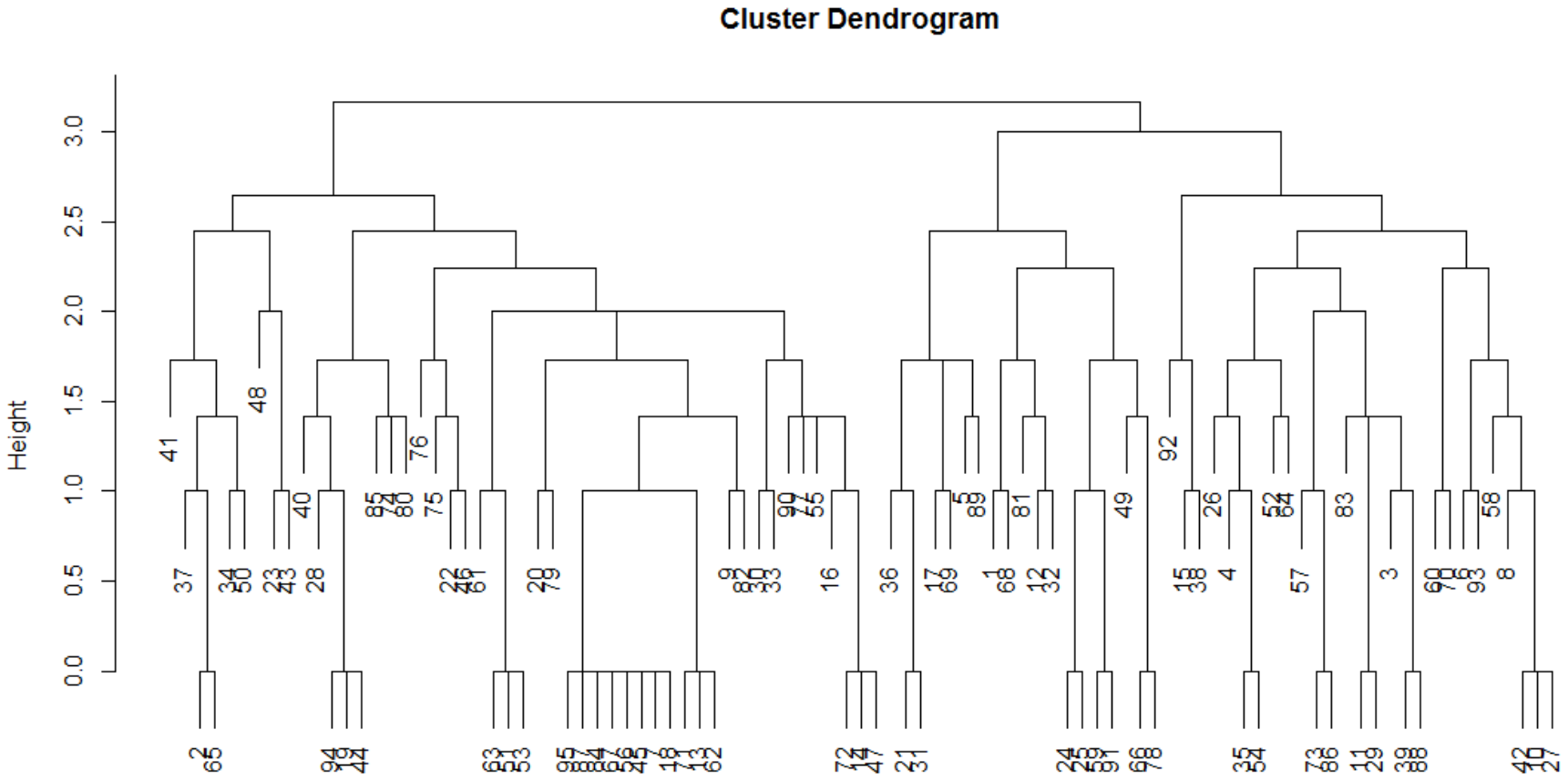
This is difficult. The wCSS (within-cluster sum of squares) goes down as the number of clusters increases. It is recommended to try various numbers of clusters and stop when wCSS begins to decrease less rapidly (called an “elbow”.)


Jet fighters
wCSS
20
40
60
80
100
120
2
4
6
8
10
number of clusters


Whiskies
Within groups sum of squares
500
600
700
800
900
2
4
6
8
Number of Clusters

Hierarchical clustering or **agglomerative clustering** is another form of clustering. It is intuitively simpler than k-means but takes more computation and therefore cannot be used easily in large problems.

- Start with each observation in one individual cluster (ie *N* clusters in total)
- Combine two clusters if they are close to each other
- Continue, combining each cluster with its closest neighbour *(agglomeration)*, until everything is in one big cluster.



The output might look like this. The image is called a **dendrogram .** It shows at which point observations and clusters have been merged together.

What don't we know yet?

How to merge two clusters

There are several options for merging two clusters, and it is not obvious which one is best.

- Single linkage: distance between clusters A and B is

$$d(A,B) = min_{x \epsilon A, y \epsilon B} d(x,y)$$

- Complete linkage: distance between clusters A and B is

$$d(A,B) = max_{x \epsilon A, y \epsilon B} d(x,y)$$

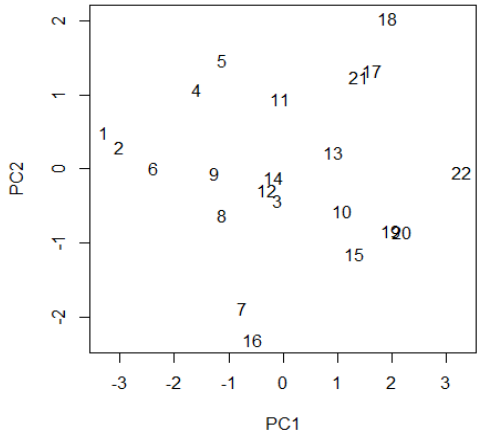
- Average linkage: distance between clusters A and B is

$$d(A,B) = mean_{x \epsilon A, y \epsilon B} d(x,y)$$

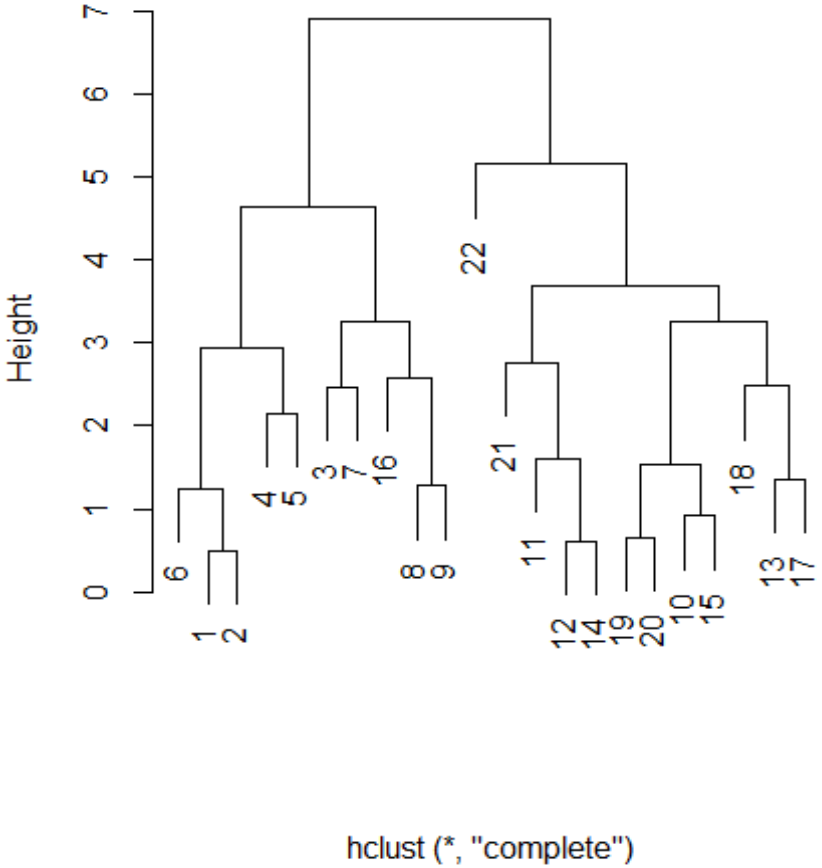
You merge cluster X with the closest cluster Y, using whatever of the three measures is chosen.

But it makes a difference! *[Explain at board]* When in doubt, complete linkage is the best one to use because it gives you "round" clusters. (However, in some cases you might be looking for stringy clusters.)

`jets` example:

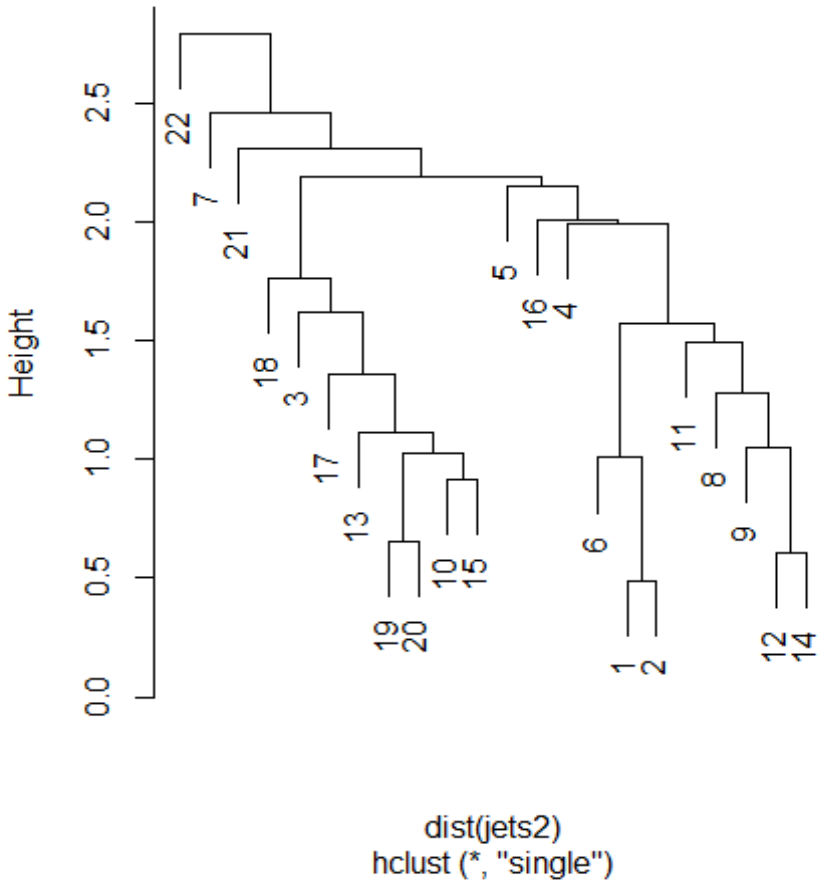


Complete

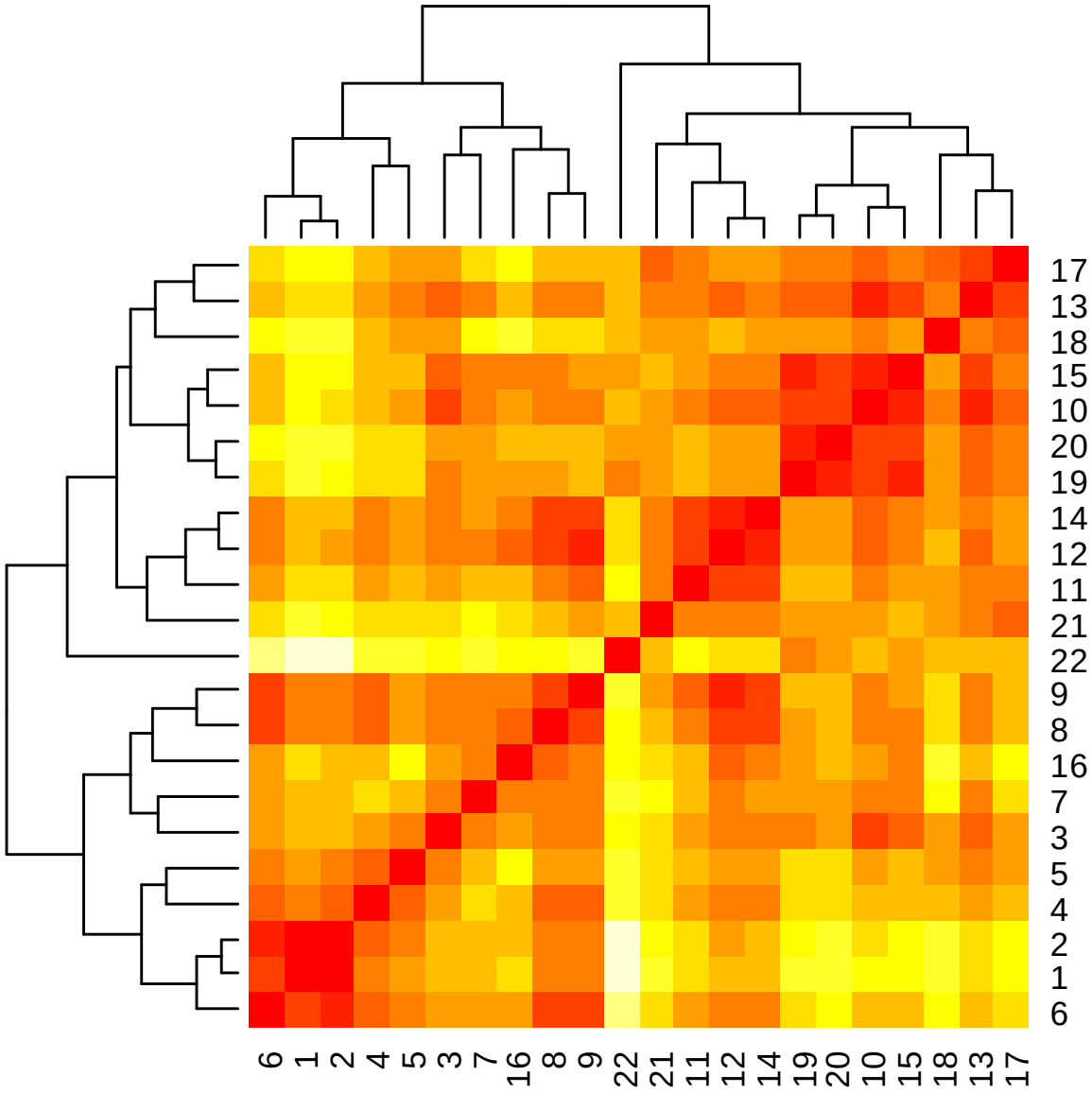


hclust (*, "complete")

Single



dist(jets2)
hclust (*, "single")


17
13
18
15
10
20
19
14
12
11
21
22
9
8
16
7
3
5
4
2
1
6
6 1 2 4 5 3 7 16 8 9 22 21 11 12 14 19 20 10 15 18 13 17

Example: clustering text documents. Commonly used in, e.g.

- Clustering blogs (political opinion)
- Natural language processing
- Search engines. *("If you are interested in X, you may also be interested in Y.")*

Simple approach: convert a document into a **bag of words** and treat these as vectors in a high-dimensional Euclidean space.

essays - Notepad

File Edit Format View Help

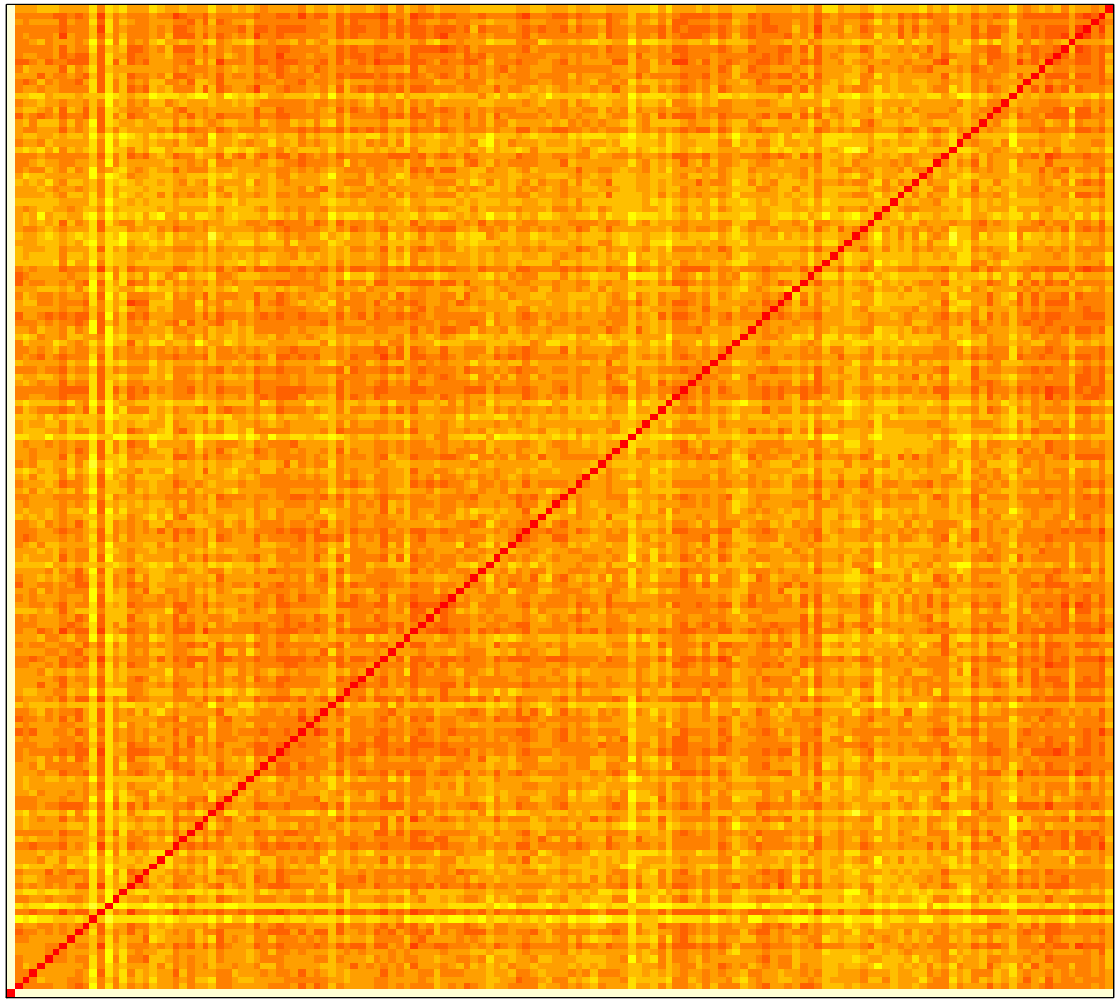
1.essay
2.I support the idea behind Badger Connect, but I am strongly against it as a mandatory requirement. The idea behind
ver under the impression that it would be mandatory, and that I do not believe is fair to students. I do not believe
3.I believe that The University of Wisconsin-Madison does not need this BadgerConnect service. There are already mar
my floor that are racially and ethnically different than myself. Therefore, I believe that the BadgerConnect service
4.Although I agree with the idea of finding a way to connect black and white students on the UW-campus and creating
. Creating an organization specifically focusing on inter-racial friendship and understanding might feel forced and
5.While BadgerConnect can implement many positive outcomes in the interaction between students of different ethnic b
ent that aims at increasing racial interaction, without forcing it. In all, despite the intended beneficial goal of
6.I feel that the BadgerConnect program would be a very beneficial program to students and the University as a whole
7.I feel like BadgerConnect will be a great resource for the students at UW-Madison. Diversity is still a very large
mily closer together.
8.BadgerConnect is a program that will help students understand one another better. It will help a students understa
teract to gain a better outlook on someone elses life. There are many differences within each person, but the one th
ith people they would never have thought to of.
9.It seems like a good idea and concept with numerous positives. However, I think that there could also be some nega
stereotypes. If there was a chance to pick teams for a sporting event, I think that the black students would be pick
10.I believe that having a student service called BadgerConnect would be a great opportunity to have on campus. Thro
nd a fun way to interact with others. I also think it would be a good idea to get the athletic teams here on campus
11.BadgerConnect sounds like a good idea from the outside looking in. There would be some problems with this idea th
nt. The second reason that BadgerConnect would not be a good idea is based on the University of Wisconsin-Madison it
12.Against badgers connect: - I think that there are a lot of other bonding organizations that already implemented t
13.I think BadgerConnect would be a great opportunity for students to get to know the life and cultures of their pee
making up this beautiful campus and an opportunity like the BadgerConnect events would improve the student relations
epted and it is important everyone knows it. Bringing together black and white students is an opportunity to learn a
14.In my opinion, I consider that student service will bring lots of benefits. Students can participant in some acti
15.The UW does a great job of organizing groups that bring the universities student body closer together. I believe
16.I don't think that Badger Connect is necessary because people going to these events will most likely go with thei
17.I believe that BadgerConnect would be a great opportunity for students to be able to meet new people and have fur
dividual. I agree that when people have groups of friends, they tend to stick together and not branch out very much,
18.The University of Wisconsin-Madison is considering implementing the program BadgerConnect, to bring together blac
ernative from the usual classroom setting for earning ethnic studies credits; making it more of an application of et
19.I think that BadgerConnect could be a good option for the students at the University of Wisconsin-Madison. It wou
20.Making new friends is usually one of the biggest issues that concerns most incoming students. Moving away from ho
21.UW-Madison is already very diverse, creating connections between races in everyday activities like class, intramu
or not between the different races. Therefore, BadgerConnect will not be an effective student group.
22.I think that making such a big deal out trying to integrate black and white students will make it even more awkwa
23.UW Madison has always stressed the importance of bringing together students from different backgrounds to enrich
was put here because of their academic success and achievement in high school, and I think BadgerConnect is a great
24.The University of Wisconsin-Madison needs to create a service called BadgerConnect to integrate students of diffe
of different ethnic backgrounds and nationalities. Often times students see a person of a different background and i
25.Being a freshman new to the UW-Madison campus and all the resources that it has to offer to competitive, hard-wor
w and become more common. As well as advocating for this program to be created, I also encourage that this program b
26.I agree with the idea that students of different races should connect together and form friendship, but I don't t
27.BadgerConnect is another attempt by the university to bring different types of students, from different walks of

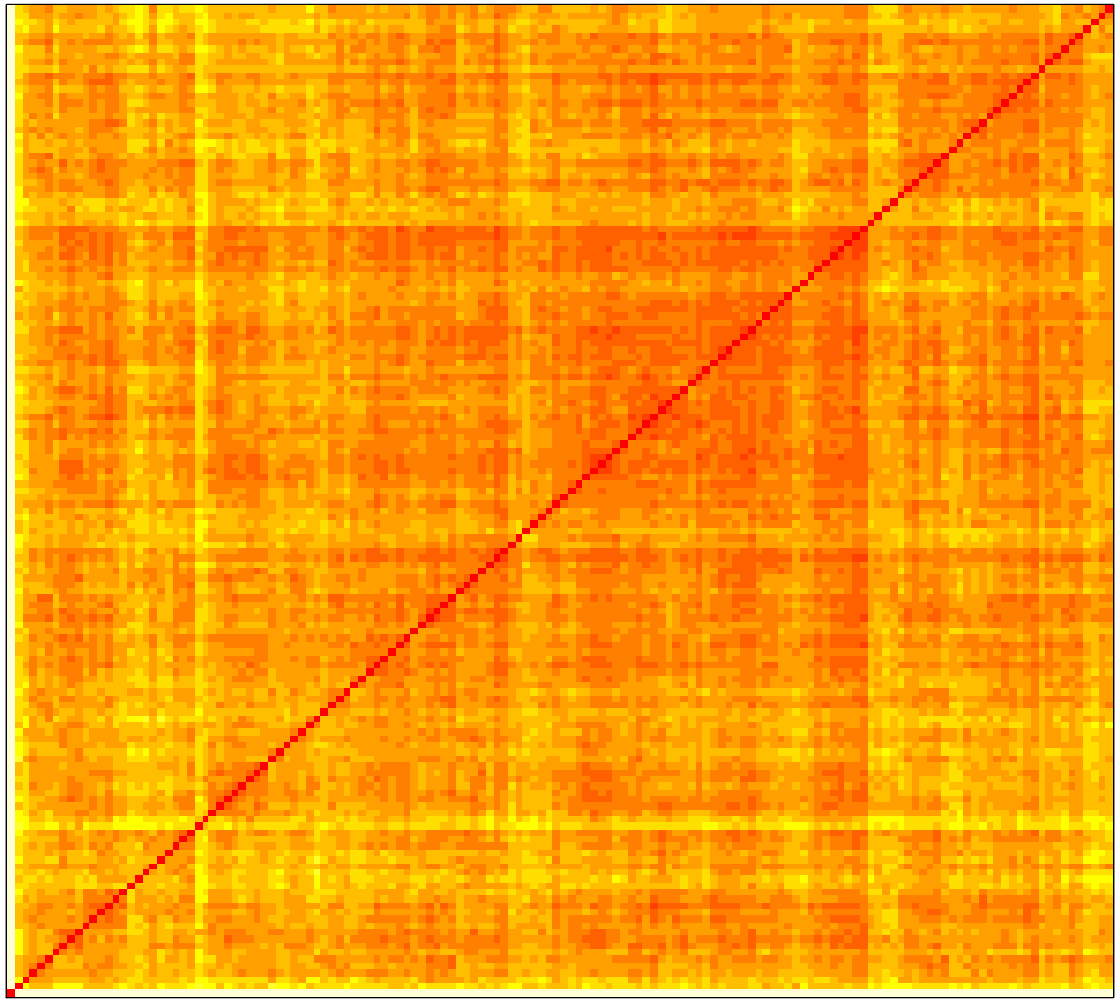
Ln 1, Col 1

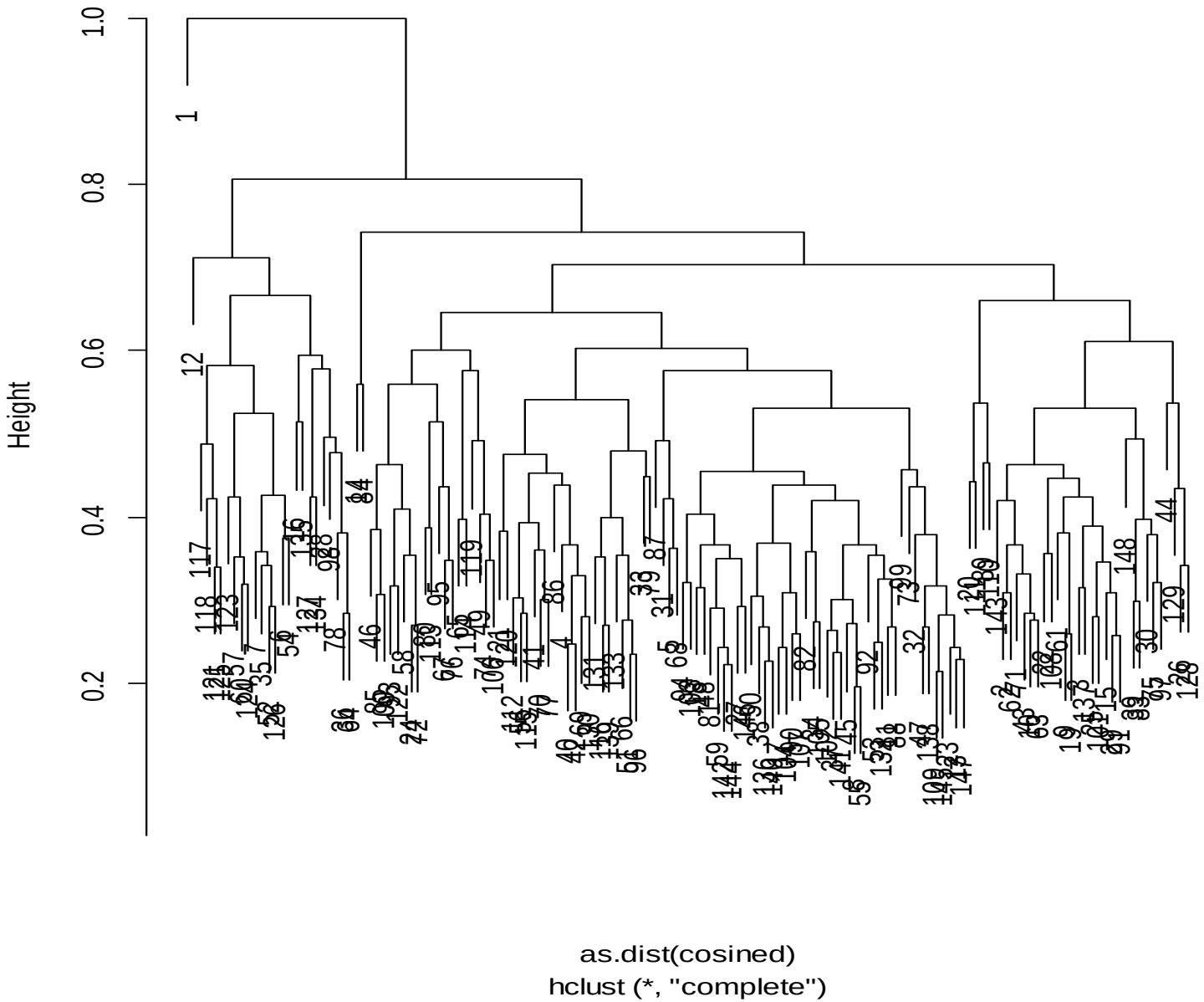
	above	absolutely	absolve	absorb	abuse	academics	accept	accepted	access	accomplish
[1,]	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0
[7,]	0	0	0	0	0	0	0	0	0	0
[8,]	0	0	0	0	0	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	0	0

After removing common words like “a” and “the”, we have a 148 x 1934 matrix. Usually, we would weight the word counts, but let us ignore this step. We can measure the distance between two documents using *cosine similarity*.

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$






Cluster Dendrogram
Height
1.0
0.8
0.6
0.4
0.2
as.dist(cosined)
hclust (*, "complete")

"badgerconnect is a program that will help students understand one another better it will help a students understand where each person comes from and how they react to certain things and how to be aware of the other people around them first off badgerconnect will allow students to gain a better understanding of the people around them on campus there is a diverse amount of students from all over the place whether they are asian african american caucasiannative american and etc and it is important to know that there are these kind of people that help make up the campus in addition to recognizing these people on campus the program will help us as students to learn to be non judgmental towards each other because of the different cultural backgrounds that we come from this program will also create a safe environment for everyone too for example someone with and asian cultural background may choose to do a certain task differently from a caucasian student and the program will allow for students to interact to gain a better outlook on someone elses life there are many differences within each person but the one thing that makes us similar is the fact that we are all students within the uwmadison community and it is imperative that we know who each other are secondly it is also important to be aware of the people around us we want to be knowledgeable about how our university is different from others it is essential that we stay connected within our university to help one another out no matter what and this can happen by knowing that there are people of different cultural backgrounds that know of things that you potentially dont and they could help you awareness is always the first step to making a change and knowing the people around you just helps you gain such a better understanding overall badgerconnect is a great program idea to help students connect with one another on a different level which is the cultural level it will help students gain awareness and be able to create new friendships with people they would never have thought to of“

"being that uw madison is such a large campus it can be hard for students to meet and connect with other people when students do being to meet people in their dorms or in classes it seems that they generally spend time with people who are similar to themselves and have the same background uw madison is a very diverse campus and part of the benefits of attending school here is for students to be able to experience new things it would be beneficial to students if the university could help to facilitate an environment where people could connect with one another build meaningful friendships and learn from people of different backgrounds truly connecting with others can be a difficult task for students when they first arrive on campus because it is intimidating to be alone on such a large campus the first instinct for a person is to find someone that is similar to him or her because it will keep him or her feeling safe and in their comfort zone however the majority of learning does not take place in the classroom and one of the most important parts of going to college is to learn that stepping outside of your comfort zone can be one of the most educational experiences possible the badgerconnect program could be an effective way for the university to help students take the first step in learning to challenge themselves and meet new different people it is especially important for black and white students to connect with one another when it comes to going out into the real world and getting a job students need to realize that they will be working with people from all different backgrounds and walks of life it is important that the students are given opportunities to connect with different people in college so that they are prepared for a job also giving students the opportunity to talk to different types of people than who they might normally talk to could be an amazing learning experience learning about different cultural traditions family styles and religions could help the students to better understand other people and to be more accepting in general the badgerconnect program would be extremely helpful in helping students to get the out of classroom experience necessary to work in the real world and to better understand others"