Statistical Analysis using R

Multivariate analysis: Hierarchical Cluster

Ibnou Dieng
Kayode Fowobaje
Sam Ofodile
Moshood Bakare
Oluwafemi Oyedele

IITA Biometrics Unit

01-02 & 07-09 December 2021

Course overview

1. ~~Short Introduction to R and RStudio~~
2. ~~Preparation of Data for Statistical Analysis~~
3. ~~Data wrangling~~
4. ~~Experimental Designs for Plant Breeding~~
5. ~~ANOVA and MET analysis~~
6. Multivariate analysis
7. Graphics in R with ggplot2

Hierarchical Cluster

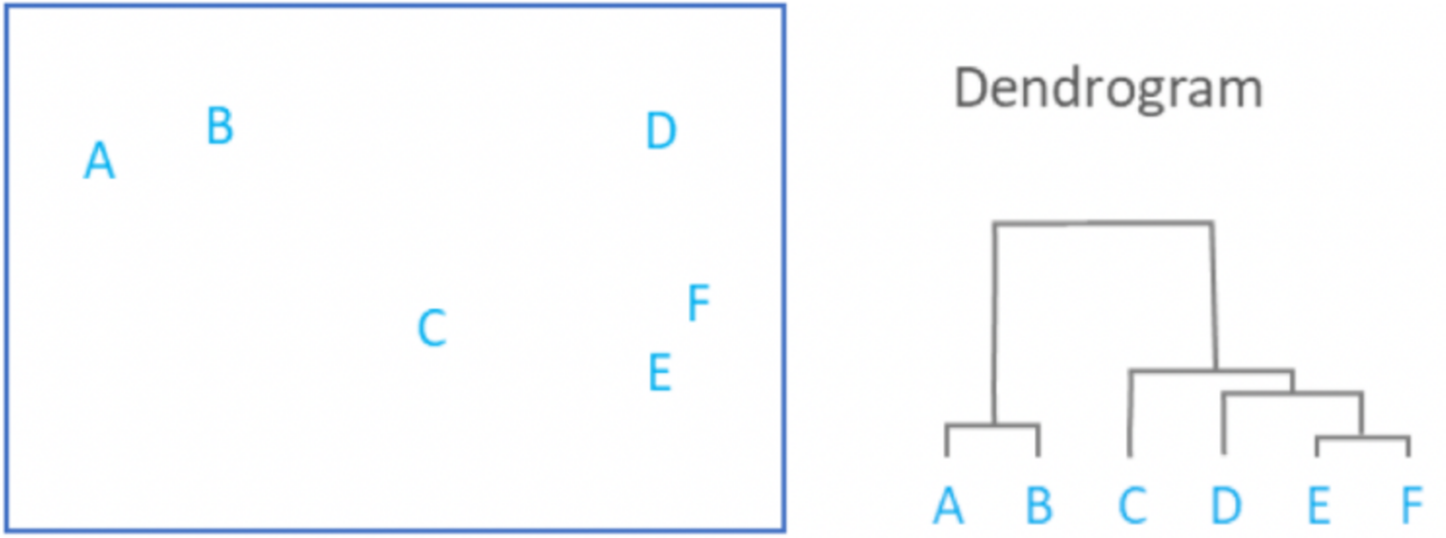
- Hierarchical clustering is an useful approach for exploring multivariate data
- Hierarchical classification algorithms allow a set of individuals to be grouped into subsets or clusters
- Objective: to create coherent clusters, but clearly different from each other
    - individuals in a cluster should be as similar as possible
    - individuals should be as different as possible from cluster to cluster

Hierarchical Cluster

- Two methods of hierarchical clustering
- Agglomerative hierarchical clustering: sequentially grouping similar clusters. At first, it is natural to group the two closest observations together. Afterwards, we can group together either:
  - individuals
  - an individual and a class
  - two classes
- Divisive hierarchical clustering: grouping all the observations into one cluster, and then successively splitting these clusters

Hierarchical Cluster

- A **dendrogram** shows the hierarchical relationship between the clusters:



https://www.displayr.com/what-is-hierarchical-clustering/

Hierarchical Cluster

- To group together we need a criterion
- The similarity between clusters is often calculated using distance metrics such as the Euclidean distance between two clusters. The greater the distance between two clusters, the better
- There are many distance metrics and the choice depends on the type of data:
    - If the data are continuous quantitative, we can use the Euclidean distance or that of Manhattan,
    - If the data is binary (categorical), we can use the Jaccard distance
    - Other distance measurements include Minkowski, Canberra, etc.
- Where there is no theoretical justification for an alternative, the **Euclidean distance** should generally be preferred

Hierarchical Cluster

Some considerations before starting

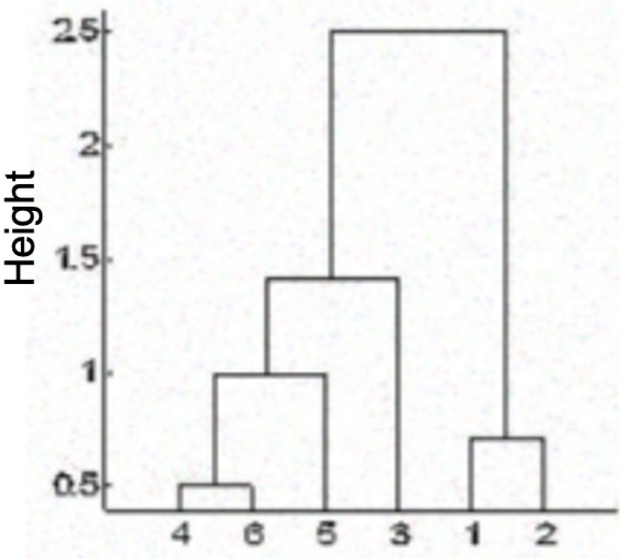
- Standardize the data: the variables to be used for clustering are of different units Standardizing the data (mean zero, unit variance) will ensure that the data is at the same scale. We subtract each data from its mean and divide it by the standard deviation. We can use the `scale ()` function in R
- Important to deal with missing values first. Several ways to deal with these values,
    - delete them
    - impute them with a mean, median, mode or use advanced regression techniques

Hierarchical Cluster

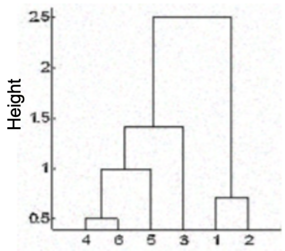
- After the distance metric, we then have to choose a linkage criteria
- There many options: single-linkage, complete-linkage, mean or average-linkage, Ward's method, etc.
- Each of the methods will produce a different dendrogram.
- In practice, we will most often prefer the **Ward method**: Ward's method seeks to minimize intra-class variability and to maximize inter-class variability to obtain the most homogeneous possible clusters

Hierarchical Cluster

- The root of the dendrogram corresponds to the cluster with all the individuals together
- This dendrogram represents a hierarchy of partitions
- The "fusion" distance (**Height**) is indicated on the `y-axis` of the dendrogram



Hierarchical Cluster

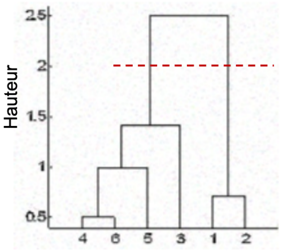


- Step-1: obs. 4 and 6 are combined into a single cluster, say cluster 1, since they were the closest in distance
    - Step-2: they are followed by obs. 1 and 2, which are in cluster 2
    - Step-3: after that obs. 5 was merged into the same cluster 1 followed by 3 resulting in two groups
    - Step-4: the two clusters are merged into a single cluster and this is where the classification process ends

Hierarchical Cluster

- One question of interest is when to stop cluster merging?
- It depends on the domain knowledge of the data
- For example, if we group several varieties of two different species, we already know that we will have to end up with only 2 clusters
- But sometimes we don't have that information *a priori*. Thus, we can use the results of the dendrogram to estimate the number of clusters
- We can cut the tree of the dendrogram with a horizontal line at a height where the line can travel the maximum distance up and down without crossing the melting point
- In this example, it would be between the heights 1.5 and 2.5. This gives two clusters

Hierarchical Cluster



- One question of interest is when to stop cluster merging?
- It depends on the domain knowledge of the data
- For example, if we group several varieties of two different species, we already know that we will have to end up with only 2 clusters
- But sometimes we don't have that information *a priori*. Thus, we can use the results of the dendrogram to estimate the number of clusters
- We can cut the tree of the dendrogram with a horizontal line at a height where we have the maximum distance up and down
- In this example, it would be between the heights 1.5 and 2.5. This gives two clusters

Hierarchical Cluster

Let's Consider the `steptoe.morex.pheno` data from the `agridat` package, multi-environment Trial of barley

- *gen* (genotype)
- *env* (environment)
- *amylase* (alpha amylase), 20 Deg Units
- *diapow* (diastatic power), degree units
- *hddate* (heading date), julian days
- *Lodging*, percent
- *malt* (malt extract), percent
- *height* (plant height), centimeters
- *protein* (grain protein), percent
- *yield* (grain yield), Mt / Ha

Hierarchical Cluster

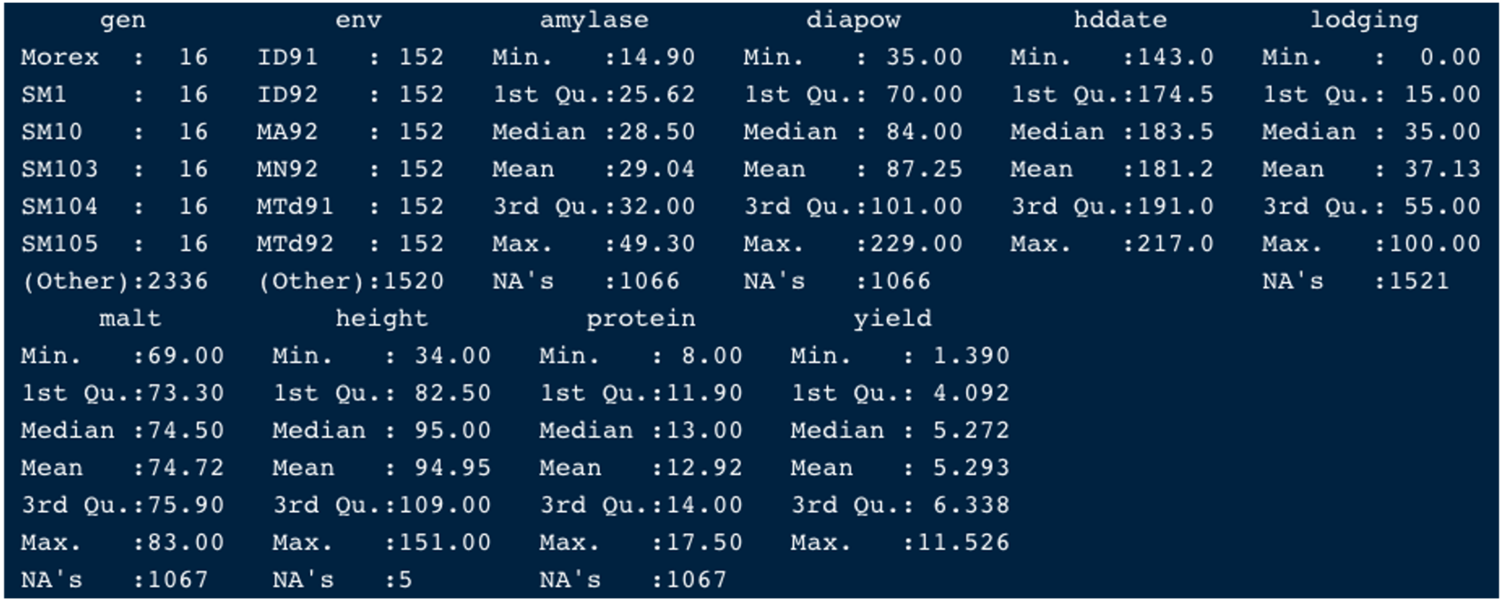
```r
library(tidyverse)
library(agridat)
data(steptoe.morex.pheno)
dat <- steptoe.morex.pheno
dat <- as_tibble(dat)
dat
```

```
## # A tibble: 2,432 x 10
##    gen     env   amylase diapow hddate lodging  malt height protein yield
##    <fct>   <fct>   <dbl>  <int>  <dbl>   <int> <dbl>  <dbl>   <dbl> <dbl>
##  1 Steptoe MN92     22.7     46   150.      NA  73.6   84.5    10.5  5.53
##  2 Steptoe MTi92    30.1     72   178       10  76.5   NA      11.2  8.64
##  3 Steptoe MTd92    26.7     78   165       15  74.5   75.5    13.4  5.90
##  4 Steptoe ID91     26.2     74   179       NA  74.1  111      12.1  8.63
##  5 Steptoe OR91     19.6     62   191       NA  71.5   90      11.7  5.34
##  6 Steptoe WA91     23.6     54   181       NA  73.8  112      10     6.27
##  7 Steptoe MTi91    21       62   181       NA  70.8   98      12     4.10
##  8 Steptoe MTd91    NA       NA   181       NA  NA     82      NA     7.07
##  9 Steptoe NY92     NA       NA   176        0  NA     77.5    NA     6.05
## 10 Steptoe ON92     NA       NA   198       50  NA     95      NA     3.70
## # … with 2,422 more rows
```

Hierarchical Cluster

- Let's look at the summary of the data

```
summary(dat)
```



```
      gen             env           amylase           diapow           hddate          lodging
 Morex    :  16   ID91    : 152   Min.   :14.90   Min.   : 35.00   Min.   :143.0   Min.   :  0.00
 SM1      :  16   ID92    : 152   1st Qu.:25.62   1st Qu.: 70.00   1st Qu.:174.5   1st Qu.: 15.00
 SM10     :  16   MA92    : 152   Median :28.50   Median : 84.00   Median :183.5   Median : 35.00
 SM103    :  16   MN92    : 152   Mean   :29.04   Mean   : 87.25   Mean   :181.2   Mean   : 37.13
 SM104    :  16   MTd91   : 152   3rd Qu.:32.00   3rd Qu.:101.00   3rd Qu.:191.0   3rd Qu.: 55.00
 SM105    :  16   MTd92   : 152   Max.   :49.30   Max.   :229.00   Max.   :217.0   Max.   :100.00
 (Other):2336     (Other):1520    NA's   :1066    NA's   :1066                     NA's   :1521
      malt           height           protein          yield
 Min.   :69.00   Min.   : 34.00   Min.   : 8.00   Min.   : 1.390
 1st Qu.:73.30   1st Qu.: 82.50   1st Qu.:11.90   1st Qu.: 4.092
 Median :74.50   Median : 95.00   Median :13.00   Median : 5.272
 Mean   :74.72   Mean   : 94.95   Mean   :12.92   Mean   : 5.293
 3rd Qu.:75.90   3rd Qu.:109.00   3rd Qu.:14.00   3rd Qu.: 6.338
 Max.   :83.00   Max.   :151.00   Max.   :17.50   Max.   :11.526
 NA's   :1067    NA's   :5        NA's   :1067
```

Transforming African Agriculture

Hierarchical Cluster

- We have data from 17 environments

```
levels(dat$env)
```

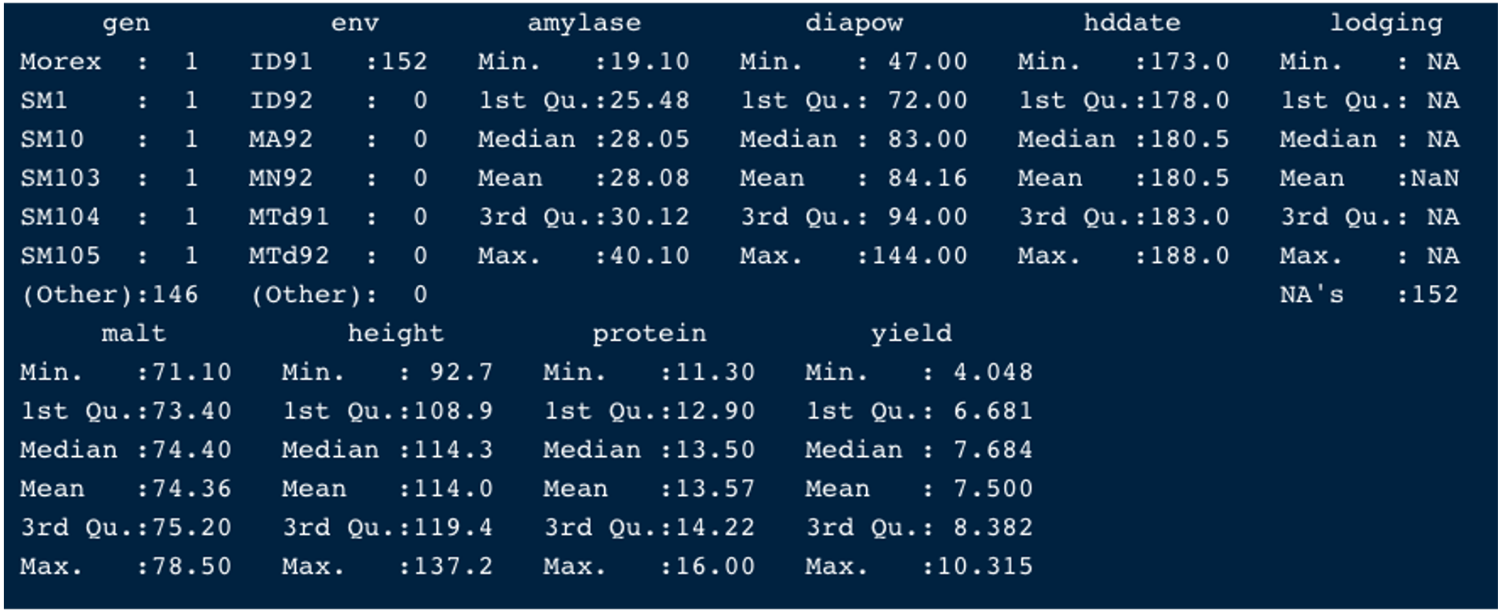
```
##  [1] "ID91"  "ID92"  "MA92"  "MN92"  "MTd91" "MTd92" "MTi91" "MTi92" "NY92
## [11] "OR91"  "SKg92" "SKk92" "SKo92" "WA91"  "WA92"
```

- Let's consider only the env **ID91**

```
dat.ID91 <- dat %>%
  filter(env=="ID91")
```

Hierarchical Cluster

```
summary(dat.ID91)
```



```
     gen            env          amylase          diapow           hddate           lodging     
 Morex    :  1   ID91    :152   Min.   :19.10   Min.   : 47.00   Min.   :173.0   Min.   : NA  
 SM1      :  1   ID92    :  0   1st Qu.:25.48   1st Qu.: 72.00   1st Qu.:178.0   1st Qu.: NA  
 SM10     :  1   MA92    :  0   Median :28.05   Median : 83.00   Median :180.5   Median : NA  
 SM103    :  1   MN92    :  0   Mean   :28.08   Mean   : 84.16   Mean   :180.5   Mean   :NaN  
 SM104    :  1   MTd91   :  0   3rd Qu.:30.12   3rd Qu.: 94.00   3rd Qu.:183.0   3rd Qu.: NA  
 SM105    :  1   MTd92   :  0   Max.   :40.10   Max.   :144.00   Max.   :188.0   Max.   : NA  
 (Other):146   (Other):  0                                                        NA's   :152  
      malt           height         protein          yield       
 Min.   :71.10   Min.   : 92.7   Min.   :11.30   Min.   : 4.048  
 1st Qu.:73.40   1st Qu.:108.9   1st Qu.:12.90   1st Qu.: 6.681  
 Median :74.40   Median :114.3   Median :13.50   Median : 7.684  
 Mean   :74.36   Mean   :114.0   Mean   :13.57   Mean   : 7.500  
 3rd Qu.:75.20   3rd Qu.:119.4   3rd Qu.:14.22   3rd Qu.: 8.382  
 Max.   :78.50   Max.   :137.2   Max.   :16.00   Max.   :10.315  
```

- `lodging` is missing for all genotyps in this environment

Hierarchical Cluster

- We exclude `lodging` in the dataset

```
dat.ID91 <- dat.ID91 %>%
  select(-lodging)
names(dat.ID91)
```

```
## [1] "gen"      "env"      "amylase" "diapow"   "hddate"   "malt"     "height"
## [9] "yield"
```

- No more missing data, but it's a good practice to use the function `na.omit()` and exclude any missing data the dataset may have

```
dat.ID91 <- na.omit(dat.ID91)
```

- Let's store the names of the genotypes in a vector `dat.ID91.label`

```
dat.ID91.label <- dat.ID91$gen
```

Hierarchical Cluster

- We delete the gen and env columns from the dataset to only keep the quantitative variables

```
dat.ID91$gen <- NULL
dat.ID91$env <- NULL
dat.ID91
```

```
## # A tibble: 152 x 7
##    amylase diapow hddate  malt height protein yield
##      <dbl>  <int>  <dbl> <dbl>  <dbl>   <dbl> <dbl>
##  1    26.2     74   179   74.1   111     12.1  8.63
##  2    36.2     97   180   76.7   116     14.8  7.95
##  3    32.8     83   183   74.4   117.    14.3  6.10
##  4    24.9     81   178.  72.3   113     14    7.20
##  5    30.2     69   180   72.9   117.    13.8  6.01
##  6    30.3     99   184.  73.2   119.    14.9  8.09
##  7    33.4     63   178   73.5   104.    13.1  7.98
##  8    25.5     74   186.  71.8   130.    14.6  5.93
##  9    26.8     78   185   75     124.    12.8  8.72
## 10    28.3     85   179   71.3   119.    14.1  6.48
## # … with 142 more rows
```

Hierarchical Cluster

- The data have different units. We standardize them using the function `scale()`

```
dat.ID91.sc <- scale(dat.ID91)
summary(dat.ID91.sc)
```

```
##     amylase              diapow             hddate              malt
##  Min.   :-2.323173   Min.   :-2.13347   Min.   :-2.13124   Min.   :-2.3946
##  1st Qu.:-0.674579   1st Qu.:-0.69806   1st Qu.:-0.71724   1st Qu.:-0.7054
##  Median :-0.008677   Median :-0.06648   Median :-0.01023   Median : 0.0289
##  Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.527924   3rd Qu.: 0.56510   3rd Qu.: 0.69677   3rd Qu.: 0.6165
##  Max.   : 3.107488   Max.   : 3.43592   Max.   : 2.11078   Max.   : 3.0401
##      height            protein              yield
##  Min.   :-2.38931   Min.   :-2.44921   Min.   :-2.7274
##  1st Qu.:-0.57607   1st Qu.:-0.72269   1st Qu.:-0.6471
##  Median : 0.02835   Median :-0.07525   Median : 0.1461
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.59919   3rd Qu.: 0.70708   3rd Qu.: 0.6974
##  Max.   : 2.59152   Max.   : 2.62243   Max.   : 2.2246
```

IITA
Transforming African Agriculture

Hierarchical Cluster

- The distance metric is calculated using the Euclidian distance

```
dist.ID91 <- dist(dat.ID91.sc, method = 'euclidean')
```

- We perform the hierarchical classification with the `hclust ()` function and specify the method

```
hclust.ID91 <- hclust(dist.ID91, method = 'ward.D2')
```

Hierarchical Cluster

- We can visualize the dendrogram using the `plot()` funnction

```
plot(hclust.ID91)
```

Hierarchical Cluster

- We can visualize the three clusters with different colors by using the `color_branches ()` function of the `dendextend` package

```
library(dendextend)
dend.ID91 <- as.dendrogram(hclust.ID91)
col.dend.ID91 <- color_branches(dend.ID91, k = 3)
```

Hierarchical Cluster

- We can visualize the dendrogram

```
plot(col.dend.ID91)
```

Hierarchical Cluster

- We can get the groups to which the genotypes belong by specifying the number of clusters

decide the number of clusters to go with

```
cut.ID91 <- cutree(hclust.ID91, k = 3)
```

- We add the group names of the genotypes to the original data

```
dat.ID91 <- dat.ID91 %>%
  mutate(cluster = cut.ID91)
```

Hierarchical Cluster

```
dat.ID91
```

```
## # A tibble: 152 x 8
##    amylase diapow hddate  malt height protein yield cluster
##      <dbl>  <int>  <dbl> <dbl>  <dbl>   <dbl> <dbl>   <int>
##  1    26.2     74    179   74.1   111     12.1  8.63       1
##  2    36.2     97    180   76.7   116     14.8  7.95       2
##  3    32.8     83    183   74.4   117.    14.3  6.10       1
##  4    24.9     81    178.  72.3   113     14    7.20       1
##  5    30.2     69    180   72.9   117.    13.8  6.01       1
##  6    30.3     99    184.  73.2   119.    14.9  8.09       3
##  7    33.4     63    178   73.5   104.    13.1  7.98       1
##  8    25.5     74    186.  71.8   130.    14.6  5.93       3
##  9    26.8     78    185   75     124.    12.8  8.72       3
## 10    28.3     85    179   71.3   119.    14.1  6.48       1
## # … with 142 more rows
```

Hierarchical Cluster

- We can aggregate the data based on groups

```
dat.ID91.summary <- dat.ID91 %>%
  group_by(cluster) %>%
  summarize(
    amylase=mean(amylase, na.rm=TRUE),
    diapow=mean(diapow, na.rm=TRUE),
    hddate=mean(hddate, na.rm=TRUE),
    malt=mean(malt, na.rm=TRUE),
    height=mean(height, na.rm=TRUE),
    protein=mean(protein, na.rm=TRUE),
    yield=mean(yield, na.rm=TRUE),
    Nobs=n()
  )
dat.ID91.summary
```

```
## # A tibble: 3 x 9
##   cluster amylase diapow hddate  malt height protein yield  Nobs
##     <int>   <dbl>  <dbl>  <dbl> <dbl>  <dbl>   <dbl> <dbl> <int>
## 1       1    26.6   75.7   179.  74.0   110.    13.3  7.71    86
## 2       2    33.9  110.    179.  74.6   115.    14.8  6.24    17
## 3       3    28.6   89.9   184.  74.9   121.    13.5  7.56    49
```