Cluster Analysis I

Mike Strube

October 24, 2018

1 Preliminaries

In this section, the RStudio workspace and console panes are cleared of old output, variables, and other miscellaneous debris. Packages are loaded and any required data files are retrieved.

```
options(replace.assign = TRUE, width = 65, digits = 4, scipen = 4, fig.width = 4,
    fig.height = 4)
# Clear the workspace and console.
rm(list = ls(all = TRUE))
cat("\f")
```

```
# Turn off showing of significance asterisks.
options(show.signif.stars = F)
# Set the contrast option; important for ANOVAs.
options(contrasts = c("contr.sum", "contr.poly"))
how_long <- Sys.time()
set.seed(123)
library(knitr)
```

```
library(psych)
library(ggplot2)

##
## Attaching package:  'ggplot2'
## The following objects are masked from 'package:psych':
##
##     %+%, alpha

library(MASS)
library(sciplot)
library(ggplot2)
library(vegan)

## Warning:  package 'vegan' was built under R version 3.5.1
## Loading required package:  permute
## Warning:  package 'permute' was built under R version 3.5.1
## Loading required package:  lattice
## This is vegan 2.5-2

library(smacof)
```

```
## Warning:  package ’smacof’ was built under R version 3.5.1
## Loading required package:  plotrix
##
## Attaching package:  ’plotrix’
## The following object is masked from ’package:psych’:
##
##    rescale
##
## Attaching package:  ’smacof’
## The following object is masked from ’package:base’:
##
##    transform
```

```
library(ape)
library(ade4)
```

```
## Warning:  package ’ade4’ was built under R version 3.5.1
```

```
library(scatterplot3d)
library(cluster)
library(factoextra)
```

```
## Warning:  package ’factoextra’ was built under R version 3.5.1
## Welcome!  Related Books:  ‘Practical Guide To Cluster Analysis in R‘ at https://goo.gl/13EFCZ
```

```
library(ggdendro)
```

```
## Warning:  package ’ggdendro’ was built under R version 3.5.1
```

2 Simple Case

Let's begin with a very simple problem: 5 cases that are arranged in two-dimensional space for which Euclidean distance is easily visualized. To emphasize the distances a bit we can use squared Euclidean distance.

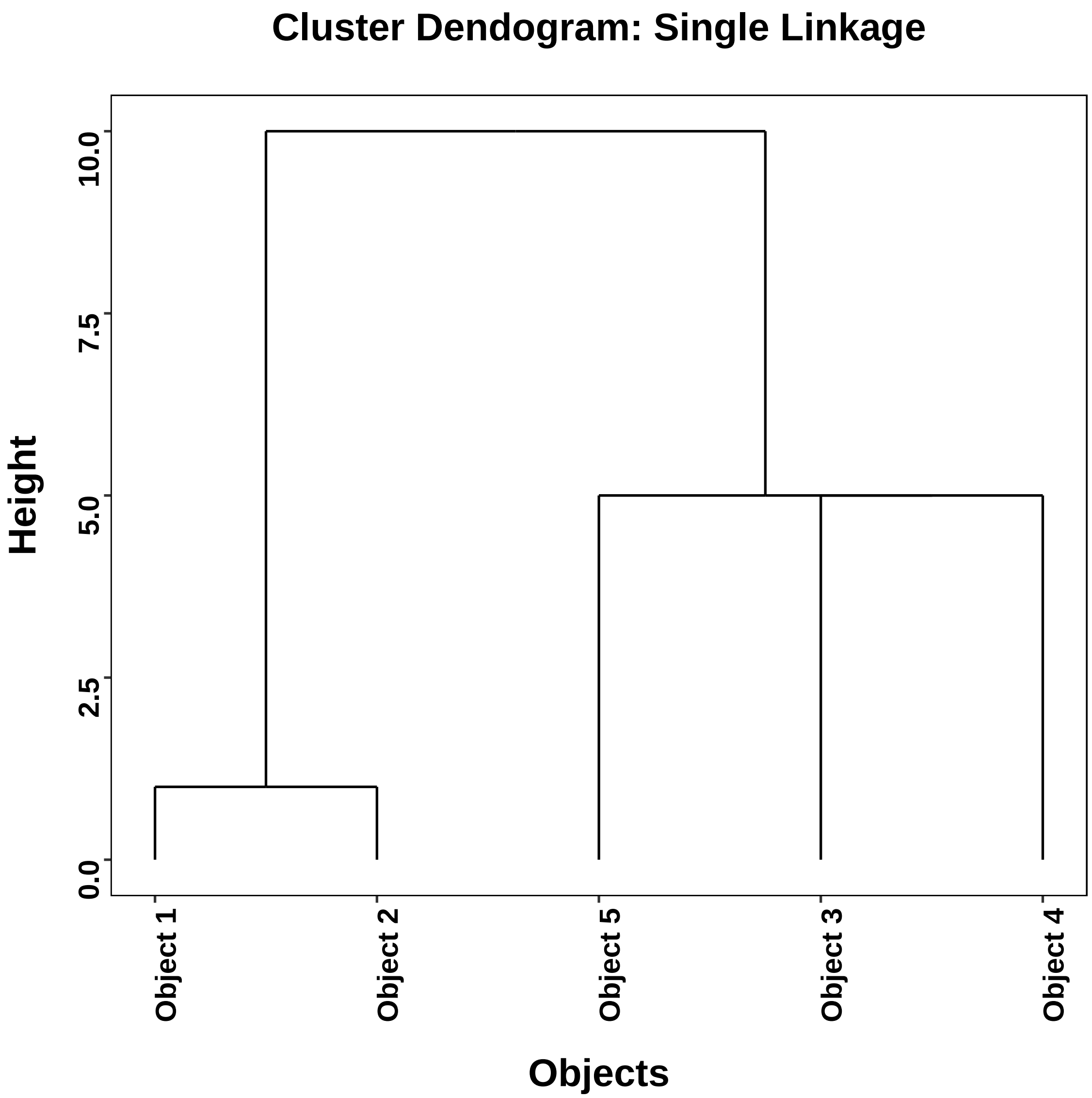
```
Data <- matrix(c(2, 3, 4, 6, 7, 2, 2, 5, 4, 6), nrow = 5, ncol = 2,
    byrow = FALSE)
row.names(Data) <- c("Object 1", "Object 2", "Object 3", "Object 4",
    "Object 5")
Data

##          [,1] [,2]
## Object 1    2    2
## Object 2    3    2
## Object 3    4    5
## Object 4    6    4
## Object 5    7    6

Data_Dist <- (dist(Data, method = "euclidean"))^2
Data_Dist

##          Object 1 Object 2 Object 3 Object 4
## Object 2        1
## Object 3       13       10
## Object 4       20       13        5
## Object 5       41       32       10        5
```

2.1 Single Linkage

The dissimilarity between two clusters (A and B) is the minimum of all possible distances between the cases in Cluster A and the cases in Cluster B.

```
hc_1 <- hclust(Data_Dist, method = "single")
hc_1$merge

##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3   -4
## [3,]   -5    2
## [4,]    1    3

hc_1$height

## [1]  1  5  5 10
```

```
ggdendrogram(hc_1, theme_dendro = FALSE, size = 4) + xlab("Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 12, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Cluster Dendogram: Single Linkage")
```

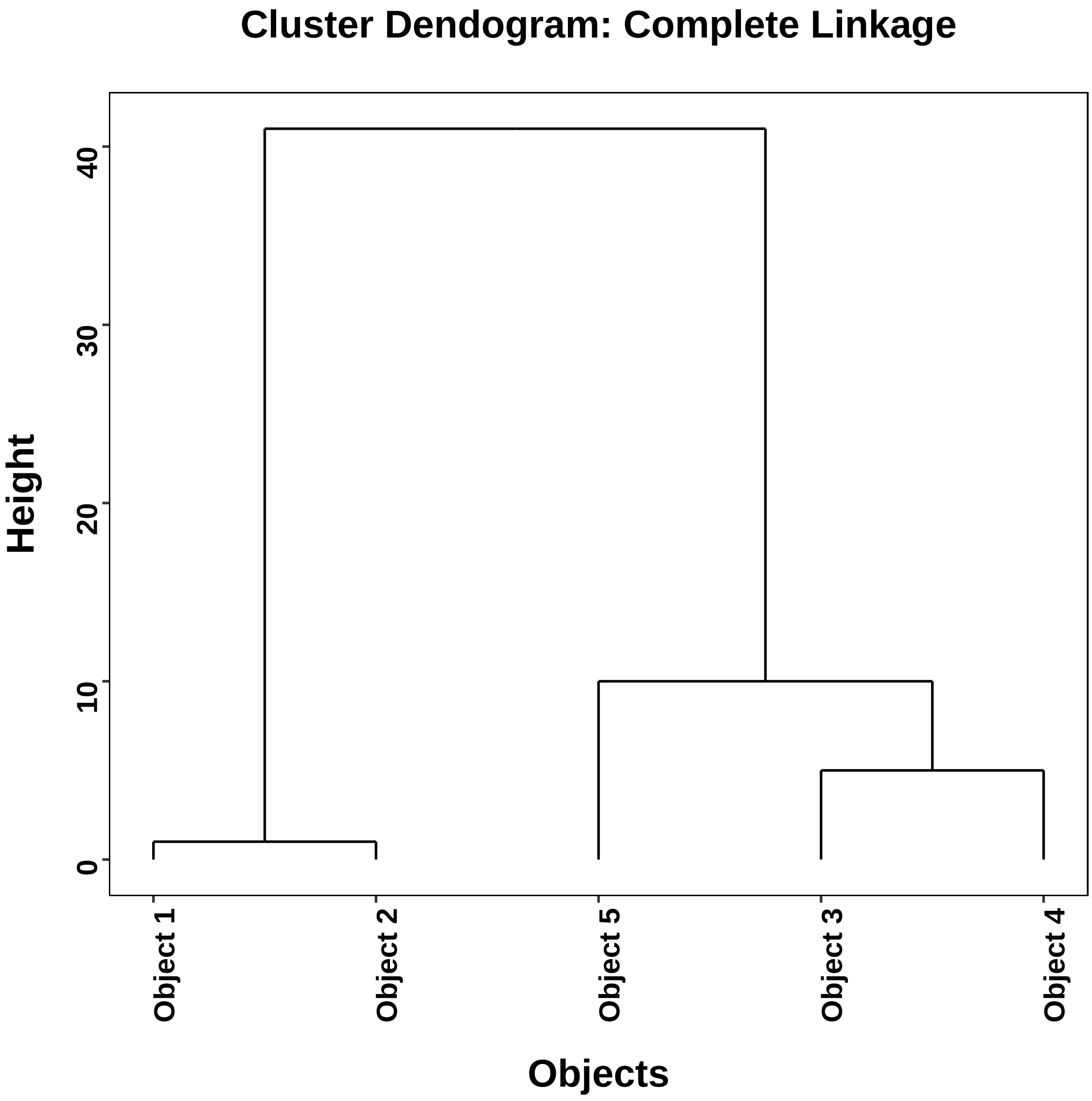

Cluster Dendogram: Single Linkage
Height
10.0
7.5
5.0
2.5
0.0
Object 1
Object 2
Object 5
Object 3
Object 4
Objects

2.2 Complete Linkage

The dissimilarity between two clusters (A and B) is the maximum of all possible distances between the cases in Cluster A and the cases in Cluster B.

```
hc_2 <- hclust(Data_Dist, method = "complete")
hc_2$merge

##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3   -4
## [3,]   -5    2
## [4,]    1    3

hc_2$height

## [1]  1  5 10 41
```

```
ggdendrogram(hc_2, theme_dendro = FALSE) + xlab("Objects") + ylab("Height") +
    theme(text = element_text(size = 14, family = "sans", color = "black",
        face = "bold"), axis.text.y = element_text(colour = "black",
        size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
        size = 12, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
        0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
        15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
        axis.line.y = element_blank(), plot.title = element_text(size = 16,
            face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
        panel.background = element_rect(fill = "white", linetype = 1,
            color = "black"), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
        plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
        legend.title = element_blank()) + ggtitle("Cluster Dendogram: Complete Linkage")
```

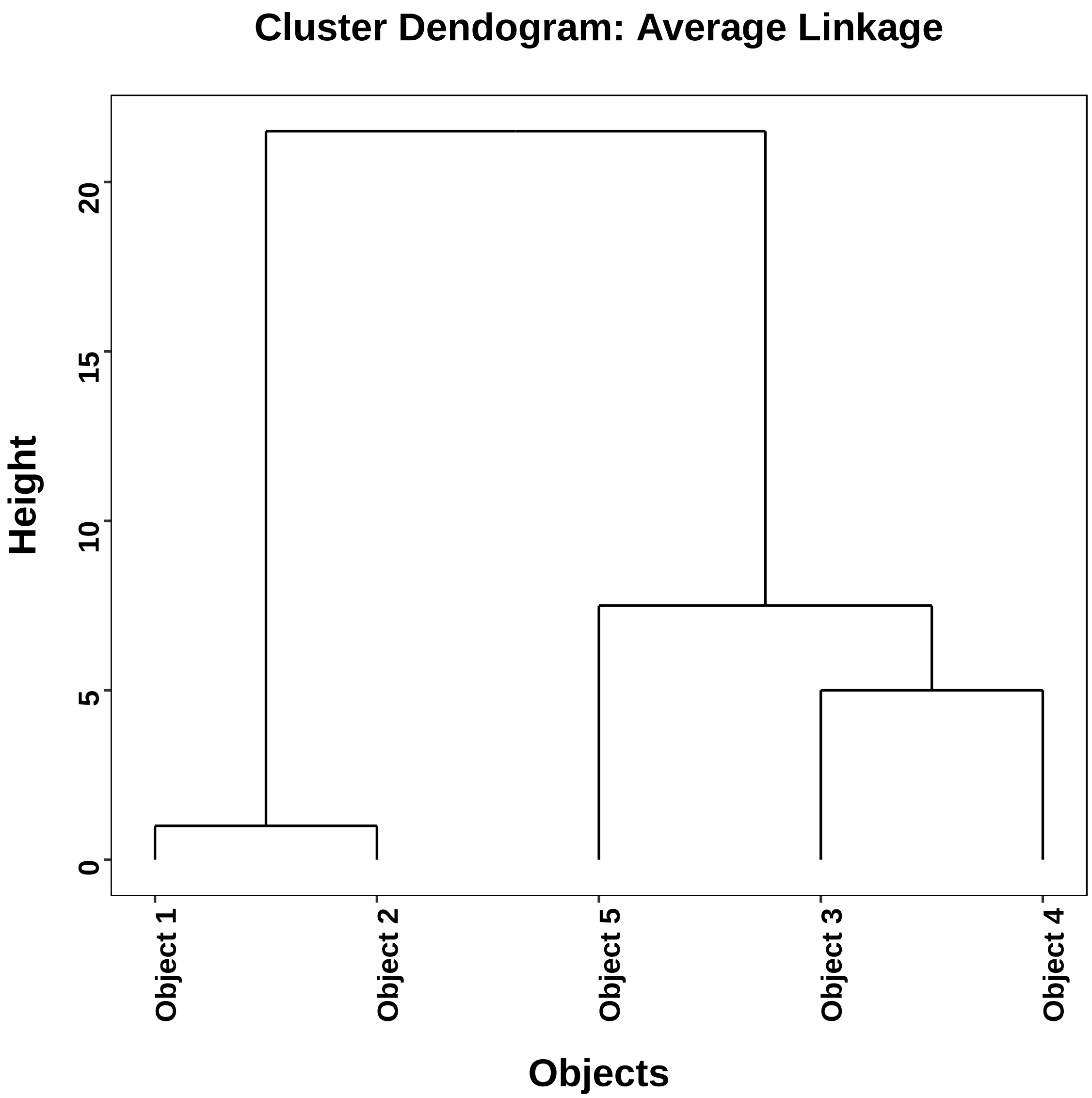

Cluster Dendogram: Complete Linkage
Height
0
10
20
30
40
Object 1
Object 2
Object 5
Object 3
Object 4
Objects

2.3 Average Linkage

The dissimilarity between two clusters (A and B) is the average of all possible distances between the cases in Cluster A and the cases in Cluster B.

```
hc_3 <- hclust(Data_Dist, method = "average")
hc_3$merge

##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3   -4
## [3,]   -5    2
## [4,]    1    3

hc_3$height

## [1]  1.0  5.0  7.5 21.5
```

```
ggdendrogram(hc_3, theme_dendro = FALSE) + xlab("Objects") + ylab("Height") +
    theme(text = element_text(size = 14, family = "sans", color = "black",
        face = "bold"), axis.text.y = element_text(colour = "black",
        size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
        size = 12, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
        0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
        15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
        axis.line.y = element_blank(), plot.title = element_text(size = 16,
            face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
        panel.background = element_rect(fill = "white", linetype = 1,
            color = "black"), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
        plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
        legend.title = element_blank()) + ggtitle("Cluster Dendogram: Average Linkage")
```

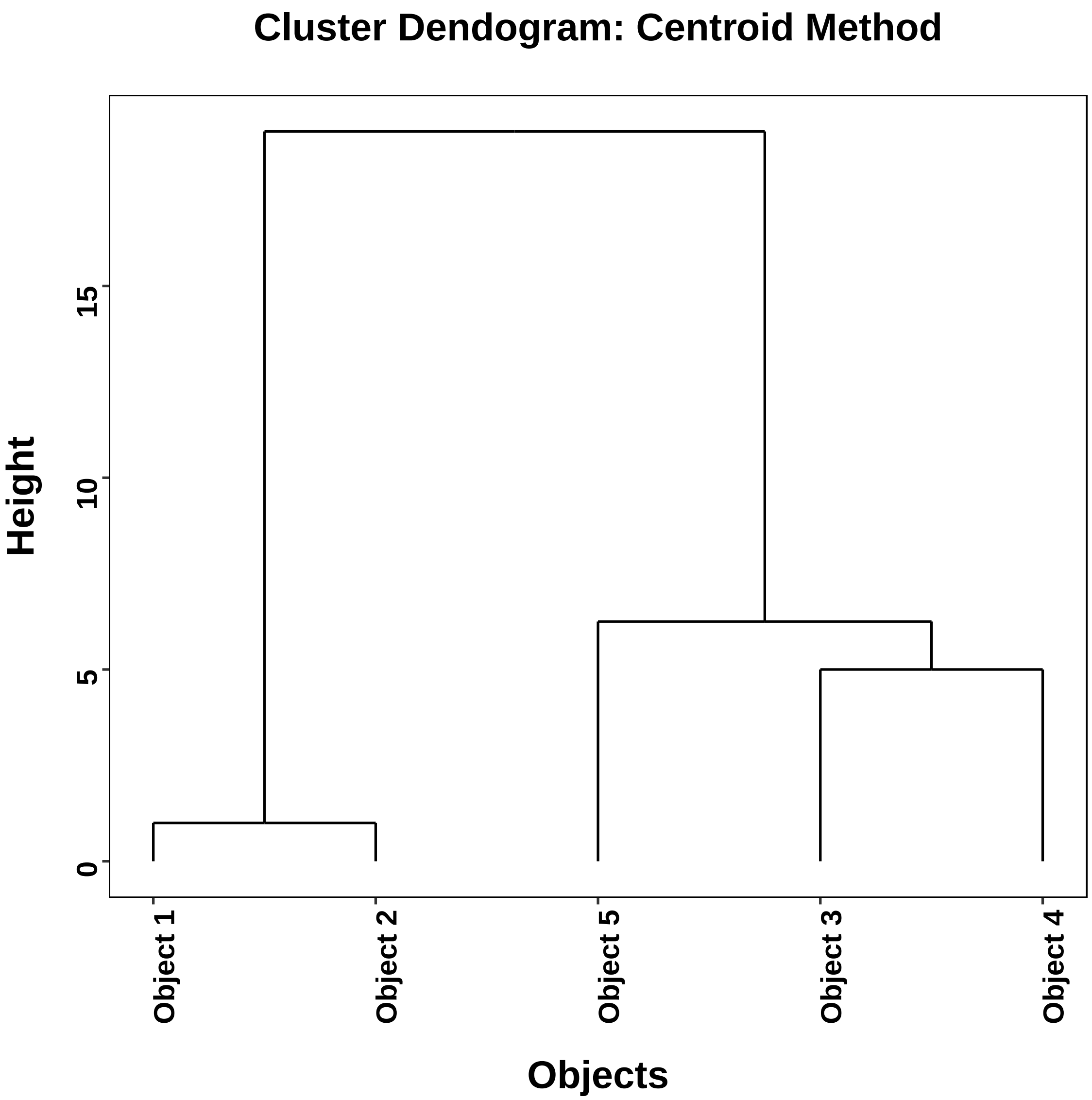

Cluster Dendogram: Average Linkage
Height
0
5
10
15
20
Object 1
Object 2
Object 5
Object 3
Object 4
Objects

2.4 Centroid Method

The dissimilarity between two clusters (A and B) is the distance between the centroid for the cases in Cluster A and the centroid for the cases in Cluster B.

```
hc_4 <- hclust(Data_Dist, method = "centroid")
hc_4$merge

##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3   -4
## [3,]   -5    2
## [4,]    1    3

hc_4$height

## [1]  1.00  5.00  6.25 19.03
```

```
ggdendrogram(hc_4, theme_dendro = FALSE) + xlab("Objects") + ylab("Height") +
    theme(text = element_text(size = 14, family = "sans", color = "black",
        face = "bold"), axis.text.y = element_text(colour = "black",
        size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
        size = 12, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
        0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
        15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
        axis.line.y = element_blank(), plot.title = element_text(size = 16,
            face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
        panel.background = element_rect(fill = "white", linetype = 1,
            color = "black"), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
        plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
        legend.title = element_blank()) + ggtitle("Cluster Dendogram: Centroid Method")
```


Cluster Dendogram: Centroid Method
Height
0
5
10
15
Object 1
Object 2
Object 5
Object 3
Object 4
Objects

2.5 Ward's Method

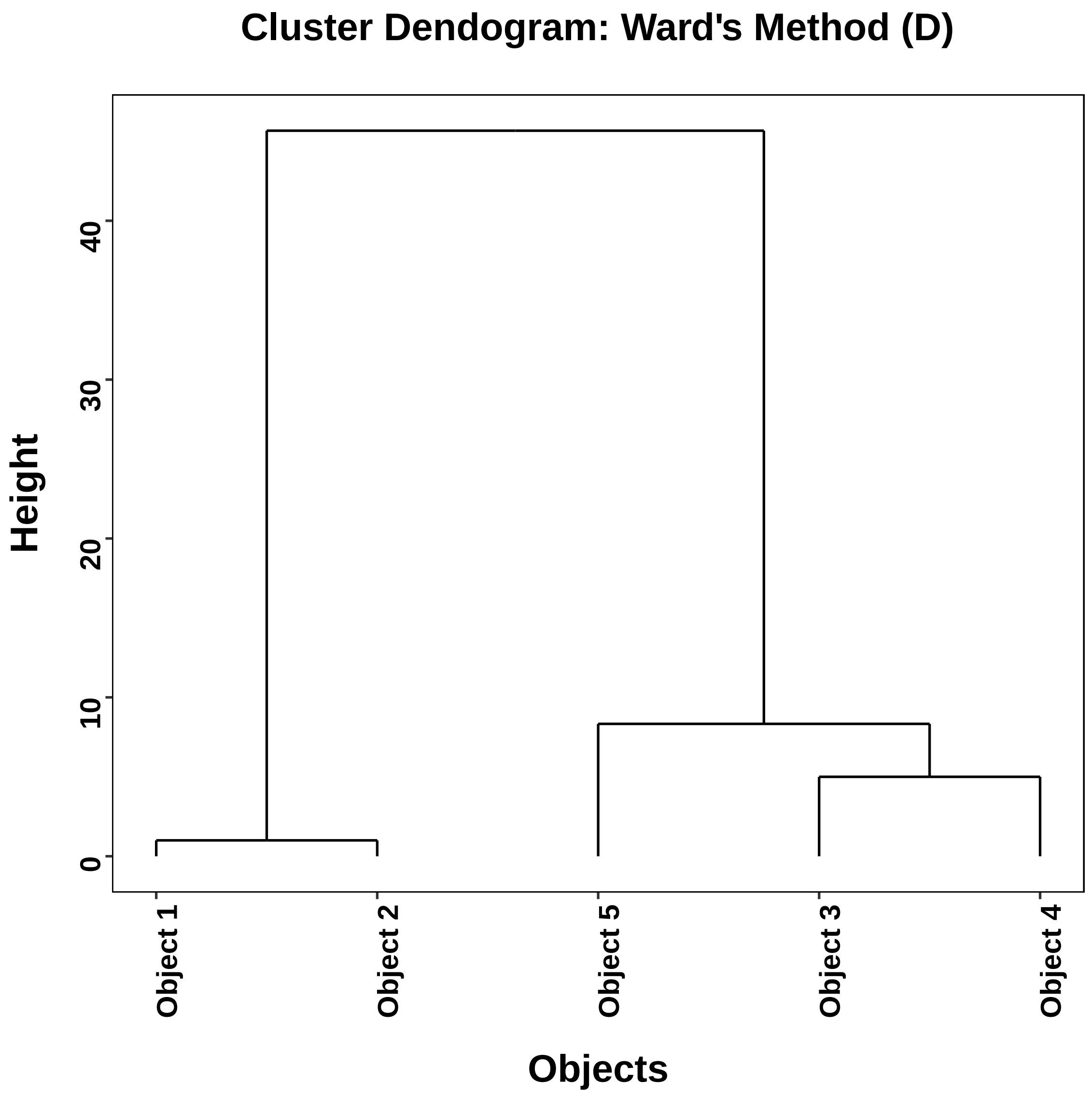
The dissimilarity between two clusters (A and B) is the loss of information from joining the clusters, measured by the increase in error sum of squares.

The sum of squares for a cluster is the sum of squared deviations of each case from the centroid for the cluster. The error sum of squares is the total of these for all clusters. The two clusters among all possible combinations that have the minimum increase in error sum of squares are joined.

Two versions are available. The Ward D method should be chosen if squared Euclidean distances are used. the Ward D2 method will produce the traditional Ward solution starting from Euclidean distances.

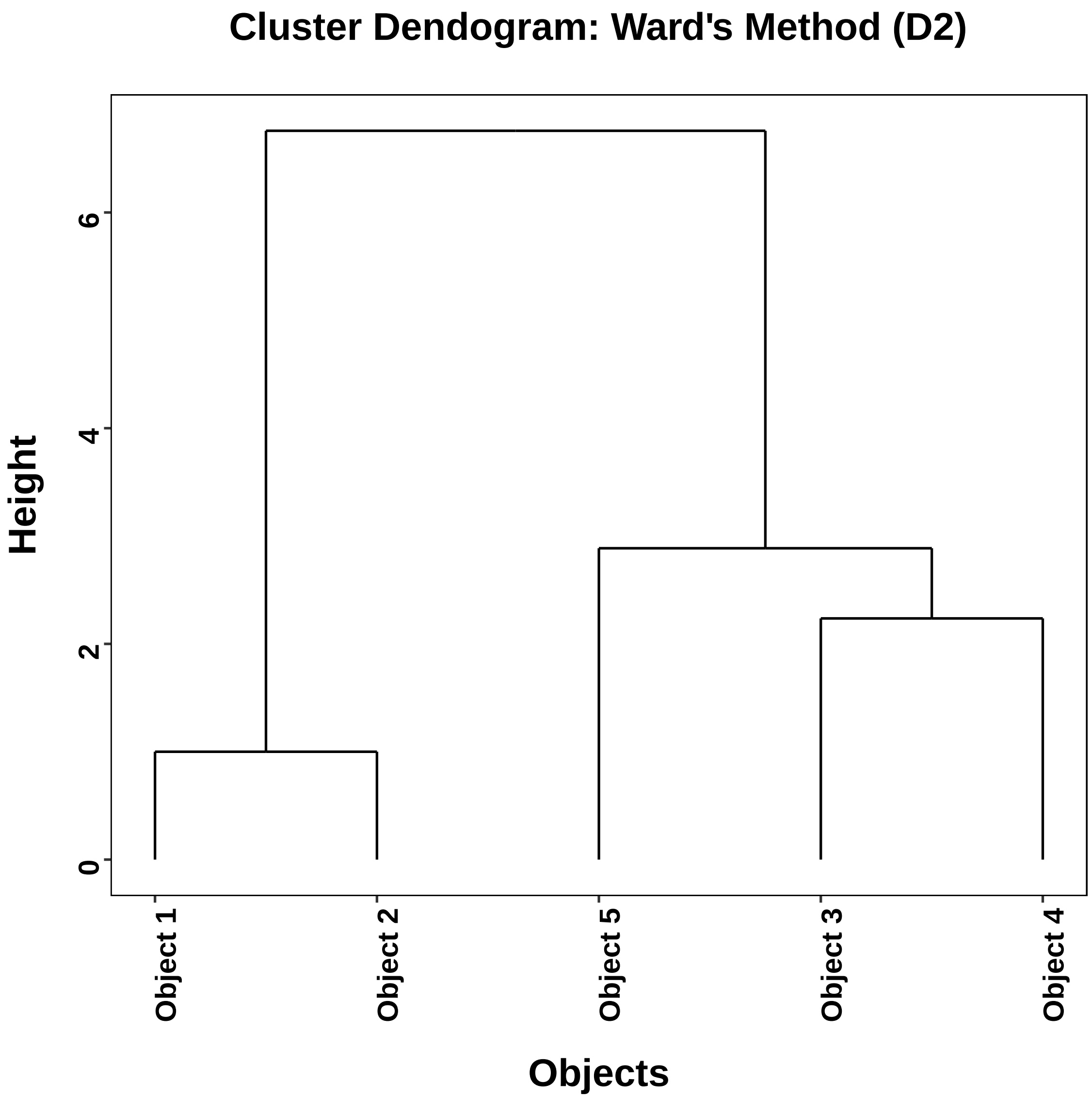
```
hc_5 <- hclust(Data_Dist, method = "ward.D")
hc_5$merge

##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3   -4
## [3,]   -5    2
## [4,]    1    3

hc_5$height

## [1]  1.000  5.000  8.333 45.667
```

```
ggdendrogram(hc_5, theme_dendro = FALSE) + xlab("Objects") + ylab("Height") +
    theme(text = element_text(size = 14, family = "sans", color = "black",
        face = "bold"), axis.text.y = element_text(colour = "black",
        size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
        size = 12, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
        0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
        15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
        axis.line.y = element_blank(), plot.title = element_text(size = 16,
            face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
        panel.background = element_rect(fill = "white", linetype = 1,
            color = "black"), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
        plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
        legend.title = element_blank()) + ggtitle("Cluster Dendogram: Ward's Method (D)")
```



```
Data_Dist <- (dist(Data, method = "euclidean"))
hc_5 <- hclust(Data_Dist, method = "ward.D2")
hc_5$merge

##      [,1] [,2]
## [1,]   -1   -2
## [2,]   -3   -4
## [3,]   -5    2
## [4,]    1    3

hc_5$height

## [1] 1.000 2.236 2.887 6.758
```

```
ggdendrogram(hc_5, theme_dendro = FALSE) + xlab("Objects") + ylab("Height") +
    theme(text = element_text(size = 14, family = "sans", color = "black",
        face = "bold"), axis.text.y = element_text(colour = "black",
        size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
        size = 12, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
        0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
        15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
        axis.line.y = element_blank(), plot.title = element_text(size = 16,
            face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
        panel.background = element_rect(fill = "white", linetype = 1,
            color = "black"), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
        plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
        legend.title = element_blank()) + ggtitle("Cluster Dendogram: Ward's Method (D2)")
```



3 Iris Data

An important question is how well the different clustering methods can recover a group structure when it is known in advance. That can lend insight into the ability of the methods to identify any group structure when that structure is not known in advance.

A classic test data set was introduced by R. A. Fisher (1936): three species of iris, varying in their petal length, petal width, sepal length, and sepal width.

```
# Get the drug use data from the working directory.
setwd("C:\\Courses\\Psychology 516\\PowerPoint\\2018")
Iris <- read.table("iris.csv", sep = ",", header = TRUE)
Iris <- as.data.frame(Iris)

Iris$Species[Iris$Species == "1"] <- "Setosa"
Iris$Species[Iris$Species == "2"] <- "Versicolor"
Iris$Species[Iris$Species == "3"] <- "Virginica"

Iris

##     Sepal_Length Sepal_Width Petal_Length Petal_Width    Species
## 1             50          33           14           2     Setosa
## 2             64          28           56          22  Virginica
## 3             65          28           46          15 Versicolor
## 4             67          31           56          24  Virginica
## 5             63          28           51          15  Virginica
## 6             46          34           14           3     Setosa
## 7             69          31           51          23  Virginica
## 8             62          22           45          15 Versicolor
## 9             59          32           48          18 Versicolor
## 10            46          36           10           2     Setosa
## 11            61          30           46          14 Versicolor
## 12            60          27           51          16 Versicolor
## 13            65          30           52          20  Virginica
## 14            56          25           39          11 Versicolor
## 15            65          30           55          18  Virginica
## 16            58          27           51          19  Virginica
## 17            68          32           59          23  Virginica
## 18            51          33           17           5     Setosa
## 19            57          28           45          13 Versicolor
## 20            62          34           54          23  Virginica
## 21            77          38           67          22  Virginica
## 22            63          33           47          16 Versicolor
## 23            67          33           57          25  Virginica
## 24            76          30           66          21  Virginica
## 25            49          25           45          17  Virginica
## 26            55          35           13           2     Setosa
## 27            67          30           52          23  Virginica
## 28            70          32           47          14 Versicolor
## 29            64          32           45          15 Versicolor
## 30            61          28           40          13 Versicolor
## 31            48          31           16           2     Setosa
## 32            59          30           51          18  Virginica
## 33            55          24           38          11 Versicolor
## 34            63          25           50          19  Virginica
## 35            64          32          53          23  Virginica
## 36            52          34          14           2     Setosa
## 37            49          36          14           1     Setosa
## 38            54          30          45          15 Versicolor
## 39            79          38          64          20  Virginica
## 40            44          32          13           2     Setosa
## 41            67          33          57          21  Virginica
## 42            50          35          16           6     Setosa
## 43            58          26          40          12 Versicolor
## 44            44          30          13           2     Setosa
## 45            77          28          67          20  Virginica
## 46            63          27          49          18  Virginica
## 47            47          32          16           2     Setosa
## 48            55          26          44          12 Versicolor
## 49            50          23          33          10 Versicolor
## 50            72          32          60          18  Virginica
## 51            48          30          14           3     Setosa
## 52            51          38          16           2     Setosa
## 53            61          30          49          18  Virginica
## 54            48          34          19           2     Setosa
## 55            50          30          16           2     Setosa
## 56            50          32          12           2     Setosa
## 57            61          26          56          14  Virginica
## 58            64          28          56          21  Virginica
## 59            43          30          11           1     Setosa
## 60            58          40          12           2     Setosa
## 61            51          38          19           4     Setosa
## 62            67          31          44          14 Versicolor
## 63            62          28          48          18  Virginica
## 64            49          30          14           2     Setosa
## 65            51          35          14           2     Setosa
## 66            56          30          45          15 Versicolor
## 67            58          27          41          10 Versicolor
## 68            50          34          16           4     Setosa
## 69            46          32          14           2     Setosa
## 70            60          29          45          15 Versicolor
## 71            57          26          35          10 Versicolor
## 72            57          44          15           4     Setosa
## 73            50          36          14           2     Setosa
## 74            77          30          61          23  Virginica
## 75            63          34          56          24  Virginica
## 76            58          27          51          19  Virginica
## 77            57          29          42          13 Versicolor
## 78            72          30          58          16  Virginica
## 79            54          34          15           4     Setosa
## 80            52          41          15           1     Setosa
## 81            71          30          59          21  Virginica
## 82            64          31          55          18  Virginica
## 83            60          30          48          18  Virginica
## 84            63          29          56          18  Virginica
## 85            49          24          33          10 Versicolor
## 86            56          27          42          13 Versicolor
## 87            57          30          42          12 Versicolor
## 88            55          42          14           2     Setosa
## 89            49          31           15           2     Setosa
## 90            77          26           69          23  Virginica
## 91            60          22           50          15  Virginica
## 92            54          39           17           4     Setosa
## 93            66          29           46          13 Versicolor
## 94            52          27           39          14 Versicolor
## 95            60          34           45          16 Versicolor
## 96            50          34           15           2     Setosa
## 97            44          29           14           2     Setosa
## 98            50          20           35          10 Versicolor
## 99            55          24           37          10 Versicolor
## 100           58          27           39          12 Versicolor
## 101           47          32           13           2     Setosa
## 102           46          31           15           2     Setosa
## 103           69          32           57          23  Virginica
## 104           62          29           43          13 Versicolor
## 105           74          28           61          19  Virginica
## 106           59          30           42          15 Versicolor
## 107           51          34           15           2     Setosa
## 108           50          35           13           3     Setosa
## 109           56          28           49          20  Virginica
## 110           60          22           40          10 Versicolor
## 111           73          29           63          18  Virginica
## 112           67          25           58          18  Virginica
## 113           49          31           15           1     Setosa
## 114           67          31           47          15 Versicolor
## 115           63          23           44          13 Versicolor
## 116           54          37           15           2     Setosa
## 117           56          30           41          13 Versicolor
## 118           63          25           49          15 Versicolor
## 119           61          28           47          12 Versicolor
## 120           64          29           43          13 Versicolor
## 121           51          25           30          11 Versicolor
## 122           57          28           41          13 Versicolor
## 123           65          30           58          22  Virginica
## 124           69          31           54          21  Virginica
## 125           54          39           13           4     Setosa
## 126           51          35           14           3     Setosa
## 127           72          36           61          25  Virginica
## 128           65          32           51          20  Virginica
## 129           61          29           47          14 Versicolor
## 130           56          29           36          13 Versicolor
## 131           69          31           49          15 Versicolor
## 132           64          27           53          19  Virginica
## 133           68          30           55          21  Virginica
## 134           55          25           40          13 Versicolor
## 135           48          34           16           2     Setosa
## 136           48          30           14           1     Setosa
## 137           45          23           13           3     Setosa
## 138           57          25           50          20  Virginica
## 139           57          38           17           3     Setosa
## 140           51          38           15           3     Setosa
## 141           55          23           40          13 Versicolor
## 142           66          30           44          14 Versicolor
## 143           68          28          48          14 Versicolor
## 144           54          34          17           2     Setosa
## 145           51          37          15           4     Setosa
## 146           52          35          15           2     Setosa
## 147           58          28          51          24  Virginica
## 148           67          30          50          17 Versicolor
## 149           63          33          60          25  Virginica
## 150           53          37          15           2     Setosa
```

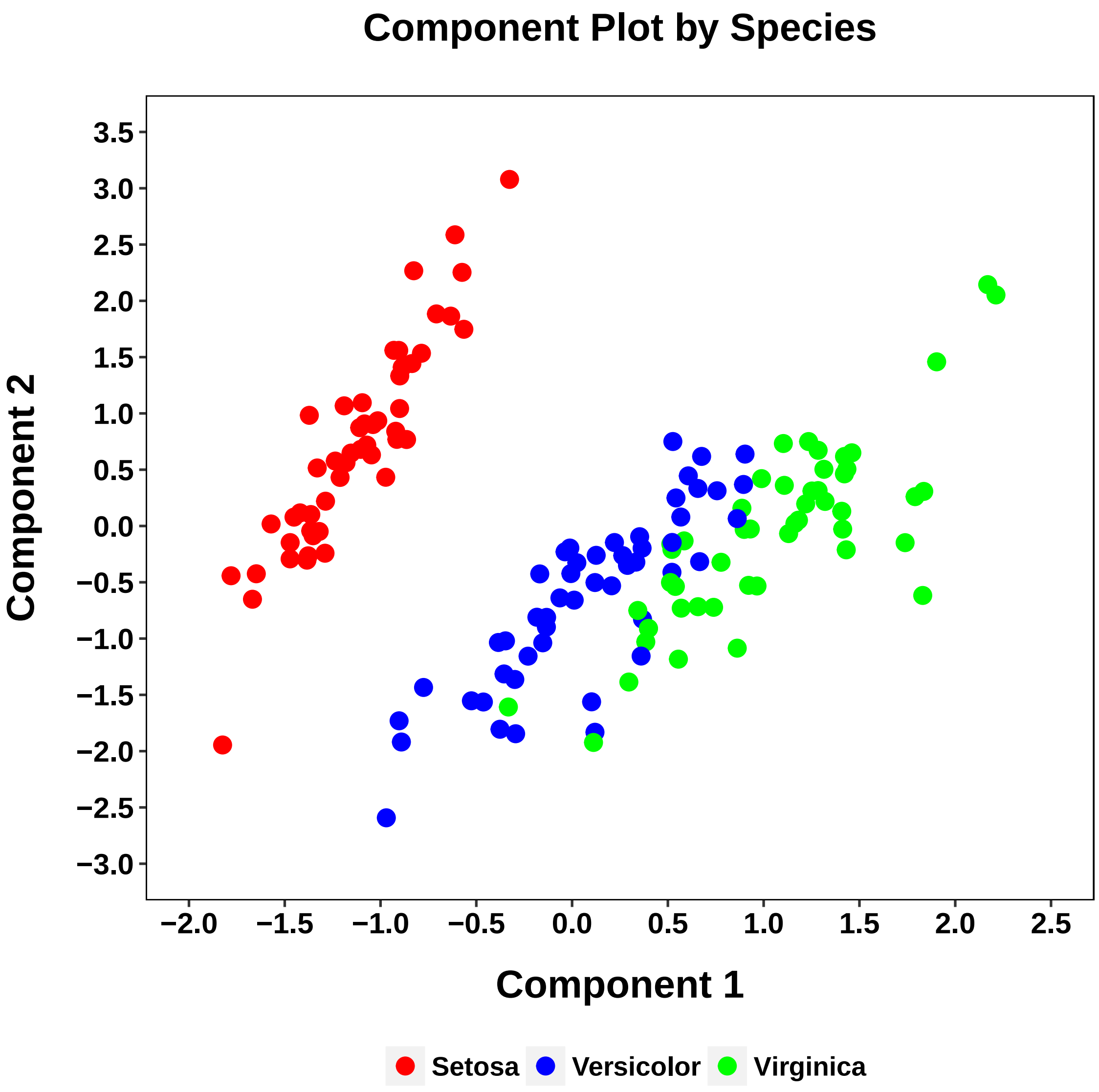
3.1 Descriptives and ANOVA

The basic features of the data can be seen in descriptive information and analyses of variance. A principal component analysis also provides a convenient way to capture and display most of the important variation in the data.

```
describeBy(Iris[, c(1:4)], group = Iris$Species, digits = 2)

##
##  Descriptive statistics by group
## group: Setosa
##              vars  n  mean   sd median trimmed  mad min max
## Sepal_Length    1 50 50.06 3.52     50   50.02 2.97  43  58
## Sepal_Width     2 50 34.28 3.79     34   34.15 3.71  23  44
## Petal_Length    3 50 14.62 1.74     15   14.60 1.48  10  19
## Petal_Width     4 50  2.46 1.05      2    2.38 0.00   1   6
##              range skew kurtosis   se
## Sepal_Length    15 0.11    -0.45 0.50
## Sepal_Width     21 0.04     0.60 0.54
## Petal_Length     9 0.10     0.65 0.25
## Petal_Width      5 1.18     1.26 0.15
## ------------------------------------------------
## group: Versicolor
##              vars  n  mean   sd median trimmed  mad min max
## Sepal_Length    1 50 59.36 5.16   59.0   59.38 5.19  49  70
## Sepal_Width     2 50 27.70 3.14   28.0   27.80 2.97  20  34
## Petal_Length    3 50 42.60 4.70   43.5   42.92 5.19  30  51
## Petal_Width     4 50 13.26 1.98   13.0   13.25 2.22  10  18
##              range  skew kurtosis   se
## Sepal_Length    21  0.10    -0.69 0.73
## Sepal_Width     14 -0.34    -0.55 0.44
## Petal_Length    21 -0.57    -0.19 0.66
## Petal_Width      8 -0.03    -0.59 0.28
## ------------------------------------------------
## group: Virginica
##              vars  n  mean   sd median trimmed  mad min max
## Sepal_Length    1 50 65.88 6.36   65.0   65.72 5.93  49  79
## Sepal_Width     2 50 29.74 3.22   30.0   29.62 2.97  22  38
## Petal_Length    3 50 55.52 5.52   55.5   55.10 6.67  45  69
## Petal_Width     4 50 20.26 2.75   20.0   20.32 2.97  14  25
##              range  skew kurtosis   se
## Sepal_Length    30  0.11    -0.20 0.90
## Sepal_Width     16  0.34     0.38 0.46
## Petal_Length    24  0.52    -0.37 0.78
## Petal_Width     11 -0.12    -0.75 0.39

# Check if the species are different in their sepal and petal
# measurements.
anova(aov(Iris$Sepal_Length ~ as.factor(Species), data = Iris))

## Analysis of Variance Table
##
## Response: Iris$Sepal_Length
##                     Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Species)   2   6321    3161     119 <2e-16
## Residuals          147   3896      27
```

```
anova(aov(Iris$Sepal_Width ~ as.factor(Species), data = Iris))

## Analysis of Variance Table
##
## Response: Iris$Sepal_Width
##                      Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Species)   2   1134     567    49.2 <2e-16
## Residuals          147   1696      12

anova(aov(Iris$Petal_Length ~ as.factor(Species), data = Iris))

## Analysis of Variance Table
##
## Response: Iris$Petal_Length
##                      Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Species)   2  43710   21855    1180 <2e-16
## Residuals          147   2722      19

anova(aov(Iris$Petal_Width ~ as.factor(Species), data = Iris))

## Analysis of Variance Table
##
## Response: Iris$Petal_Width
##                      Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Species)   2   8041    4021     960 <2e-16
## Residuals          147    616       4

# Use PCA to show potential clustering along two dimensions.
PCA <- principal(Iris[, 1:4], nfactors = 2, rotate = "varimax", scores = TRUE)
PCA

## Principal Components Analysis
## Call: principal(r = Iris[, 1:4], nfactors = 2, rotate = "varimax",
##     scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                 RC1   RC2   h2     u2 com
## Sepal_Length  0.96  0.05 0.92 0.0774 1.0
## Sepal_Width  -0.14  0.98 0.99 0.0091 1.0
## Petal_Length  0.94 -0.30 0.98 0.0163 1.2
## Petal_Width   0.93 -0.26 0.94 0.0647 1.2
##
##                         RC1  RC2
## SS loadings           2.70 1.13
## Proportion Var        0.68 0.28
## Cumulative Var        0.68 0.96
## Proportion Explained  0.71 0.29
## Cumulative Proportion 0.71 1.00
##
## Mean item complexity =  1.1
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.03
##  with the empirical chi square  1.72  with prob <  NA
##
## Fit based upon off diagonal values = 1

Iris <- cbind(Iris, PCA$scores)
```

```r
ggplot(Iris, aes(x = RC1, y = RC2, color = factor(Species))) + geom_point(shape = 19,
    size = 3) + scale_color_manual(values = c("red", "blue", "green")) +
    scale_y_continuous(breaks = c(seq(-3, 3.5, 0.5))) + scale_x_continuous(breaks = c(seq(-2,
    2.5, 0.5))) + coord_cartesian(xlim = c(-2, 2.5), ylim = c(-3,
    3.5)) + xlab("Component 1") + ylab("Component 2") + theme(text = element_text(size = 14,
    family = "sans", color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 12, face = "bold", angle = 0), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16), axis.line.x = element_blank(), axis.line.y = element_blank(),
    plot.title = element_text(size = 16, face = "bold", margin = margin(0,
        0, 20, 0), hjust = 0.5), panel.background = element_rect(fill = "white",
        linetype = 1, color = "black"), panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(), plot.background = element_rect(fill = "white"),
    plot.margin = unit(c(1, 1, 1, 1), "cm"), legend.position = "bottom",
    legend.title = element_blank()) + ggtitle("Component Plot by Species")
```


Component Plot by Species
Component 2
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Component 1
● Setosa
● Versicolor
● Virginica

3.2 Distance Calculation

```
# Use Euclidean distance for subsequent clustering.
Iris_Dist <- dist(Iris[, 1:4], method = "euclidean")
```

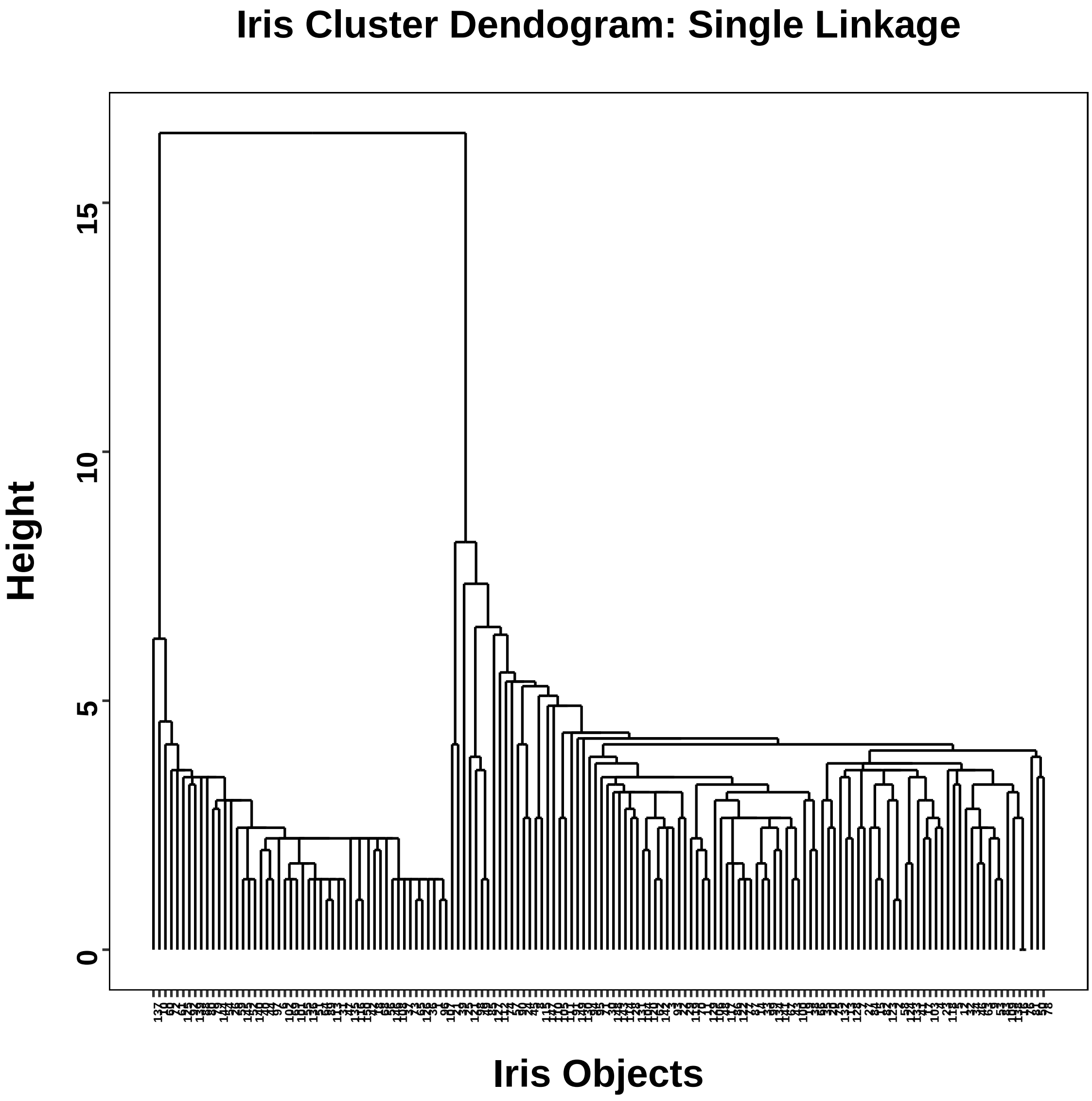
3.3 Clustering Methods

Each of the clustering methods can be applied to the iris data. Given the known structure of the data, the ability to recover the three species will help identify clustering methods that may be particularly useful.

3.3.1 Single Lingage

```
hc_1 <- hclust(Iris_Dist, method = "single")

clustnumber <- cutree(hc_1, k = 3)
Iris_Class <- as.data.frame(cbind(clustnumber, Iris$Species))
names(Iris_Class) <- c("Cluster", "Species")
table(Iris_Class$Species)

##
##     Setosa Versicolor  Virginica
##         50         50         50

table(Iris_Class$Cluster)

##
##  1  2  3
## 50 98  2

table(Iris_Class$Species, Iris_Class$Cluster)

##
##               1  2  3
##   Setosa     50  0  0
##   Versicolor  0 50  0
##   Virginica   0 48  2
```

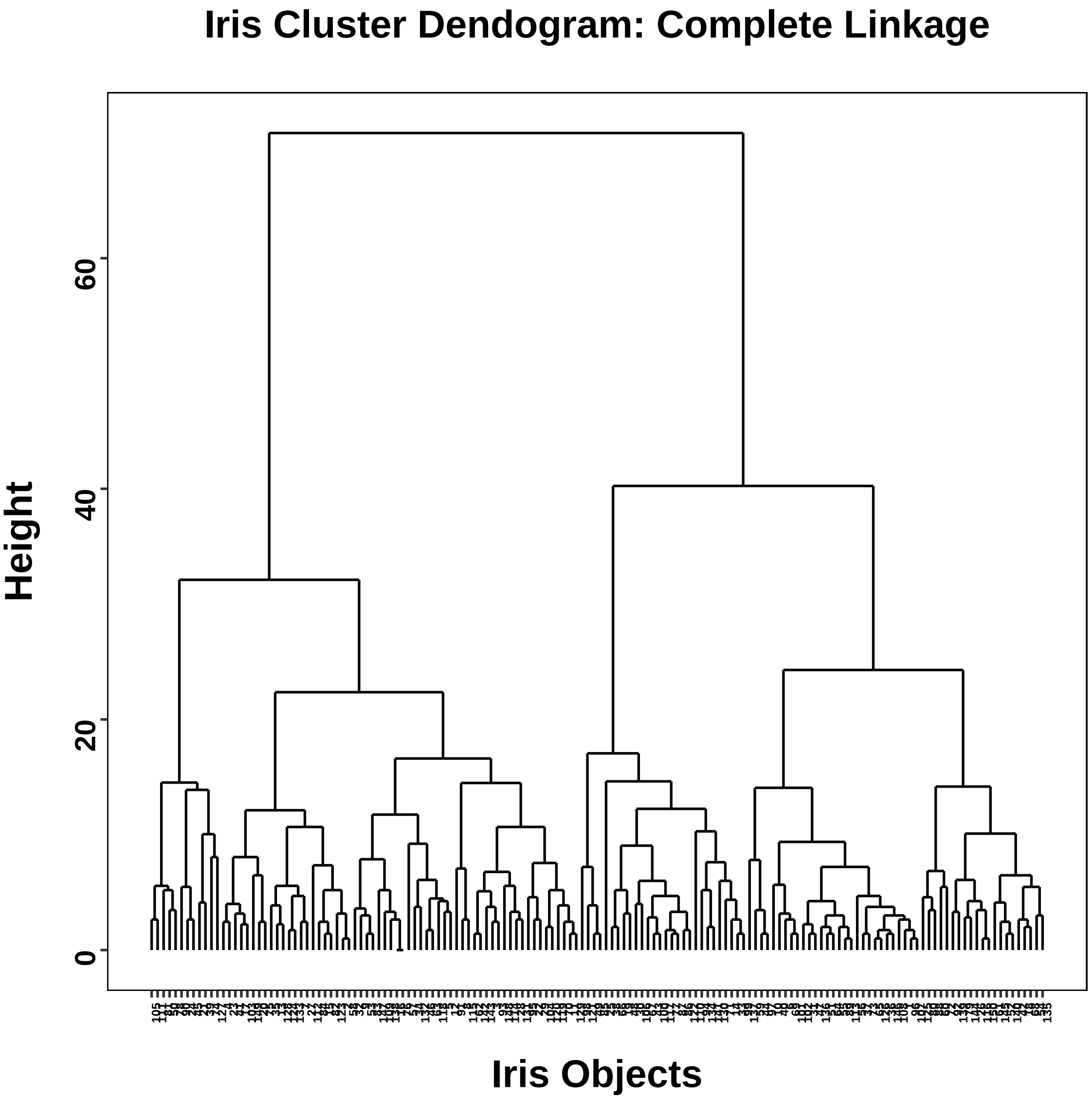
```
ggdendrogram(hc_1, theme_dendro = FALSE) + xlab("Iris Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 5, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Iris Cluster Dendogram: Single Linkage")
```


Iris Cluster Dendogram: Single Linkage
Height
0
5
10
15
Iris Objects

3.3.2 Complete Lingage

```
hc_2 <- hclust(Iris_Dist, method = "complete")

clustnumber <- cutree(hc_2, k = 3)
Iris_Class <- as.data.frame(cbind(clustnumber, Iris$Species))
names(Iris_Class) <- c("Cluster", "Species")
table(Iris_Class$Species)

##
##     Setosa Versicolor  Virginica
##         50         50         50

table(Iris_Class$Cluster)

##
##  1  2  3
## 50 72 28

table(Iris_Class$Species, Iris_Class$Cluster)

##
##                1  2  3
##   Setosa     50  0  0
##   Versicolor  0 23 27
##   Virginica   0 49  1
```

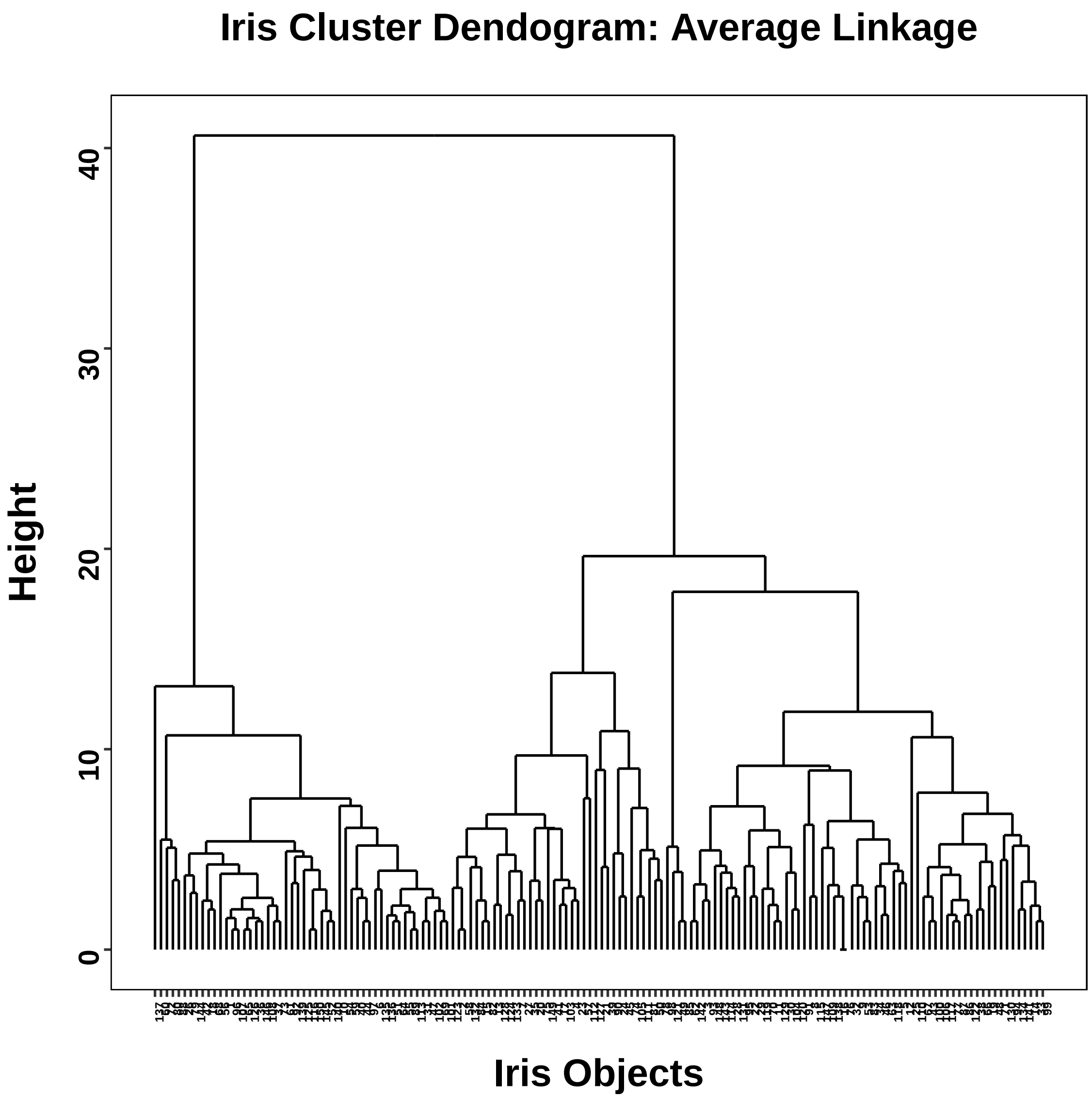
```
ggdendrogram(hc_2, theme_dendro = FALSE) + xlab("Iris Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 5, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Iris Cluster Dendogram: Complete Linkage")
```


Iris Cluster Dendogram: Complete Linkage
Height
0
20
40
60
Iris Objects

3.3.3 Average Linkage

```
hc_3 <- hclust(Iris_Dist, method = "average")

clustnumber <- cutree(hc_3, k = 3)
Iris_Class <- as.data.frame(cbind(clustnumber, Iris$Species))
names(Iris_Class) <- c("Cluster", "Species")
table(Iris_Class$Species)

##
##     Setosa Versicolor  Virginica
##         50         50         50

table(Iris_Class$Cluster)

##
##  1  2  3
## 50 36 64

table(Iris_Class$Species, Iris_Class$Cluster)

##
##                1  2  3
##   Setosa      50  0  0
##   Versicolor   0  0 50
##   Virginica    0 36 14
```

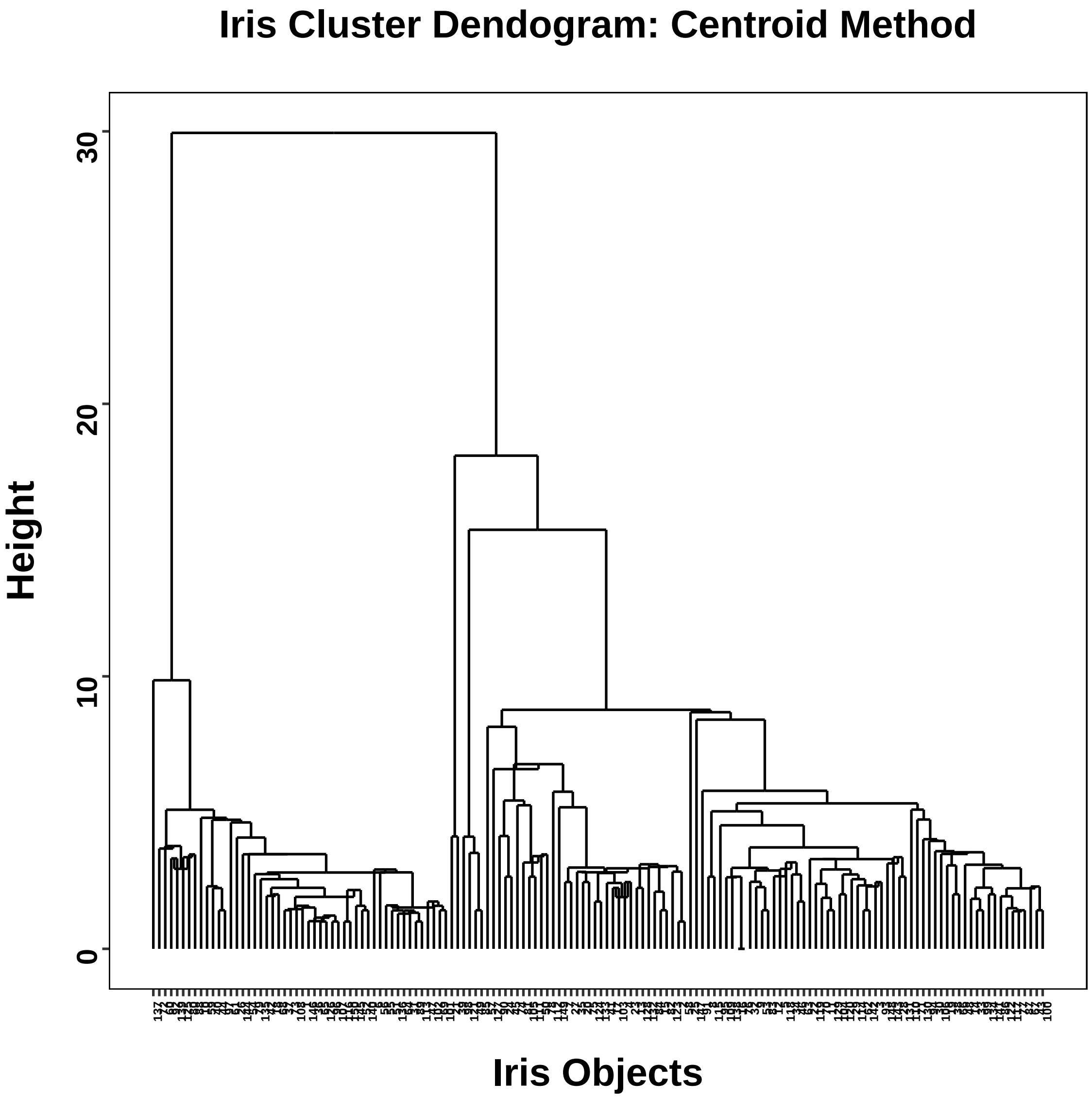
```
ggdendrogram(hc_3, theme_dendro = FALSE) + xlab("Iris Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 5, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Iris Cluster Dendogram: Average Linkage")
```


Iris Cluster Dendogram: Average Linkage
Height
0
10
20
30
40
Iris Objects

3.3.4 Centroid Method

```
hc_4 <- hclust(Iris_Dist, method = "centroid")

clustnumber <- cutree(hc_4, k = 3)
Iris_Class <- as.data.frame(cbind(clustnumber, Iris$Species))
names(Iris_Class) <- c("Cluster", "Species")
table(Iris_Class$Species)

## 
##     Setosa Versicolor  Virginica 
##         50         50         50

table(Iris_Class$Cluster)

## 
##  1  2  3 
## 50 98  2

table(Iris_Class$Species, Iris_Class$Cluster)

##             
##               1  2  3
##   Setosa     50  0  0
##   Versicolor  0 50  0
##   Virginica   0 48  2
```

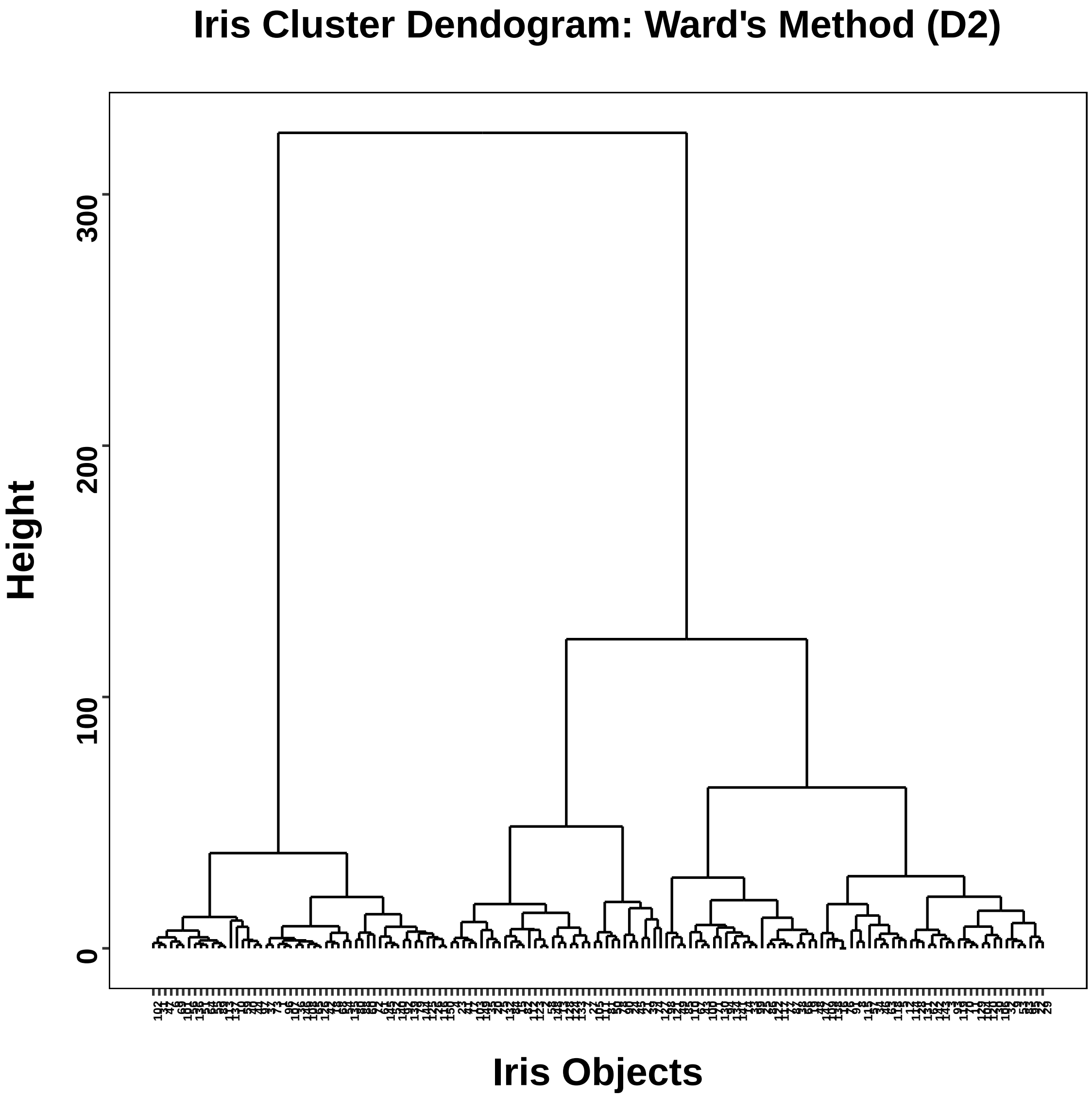
```
ggdendrogram(hc_4, theme_dendro = FALSE) + xlab("Iris Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 5, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Iris Cluster Dendogram: Centroid Method")
```


Iris Cluster Dendogram: Centroid Method
Height
0
10
20
30
Iris Objects

3.3.5 Ward's Method (D2)

```
hc_5 <- hclust(Iris_Dist, method = "ward.D2")

clustnumber <- cutree(hc_5, k = 3)
Iris_Class <- as.data.frame(cbind(clustnumber, Iris$Species))
names(Iris_Class) <- c("Cluster", "Species")
table(Iris_Class$Species)

## 
##     Setosa Versicolor  Virginica 
##         50         50         50

table(Iris_Class$Cluster)

## 
##  1  2  3 
## 50 36 64

table(Iris_Class$Species, Iris_Class$Cluster)

##             
##                1  2  3
##   Setosa      50  0  0
##   Versicolor   0  1 49
##   Virginica    0 35 15
```

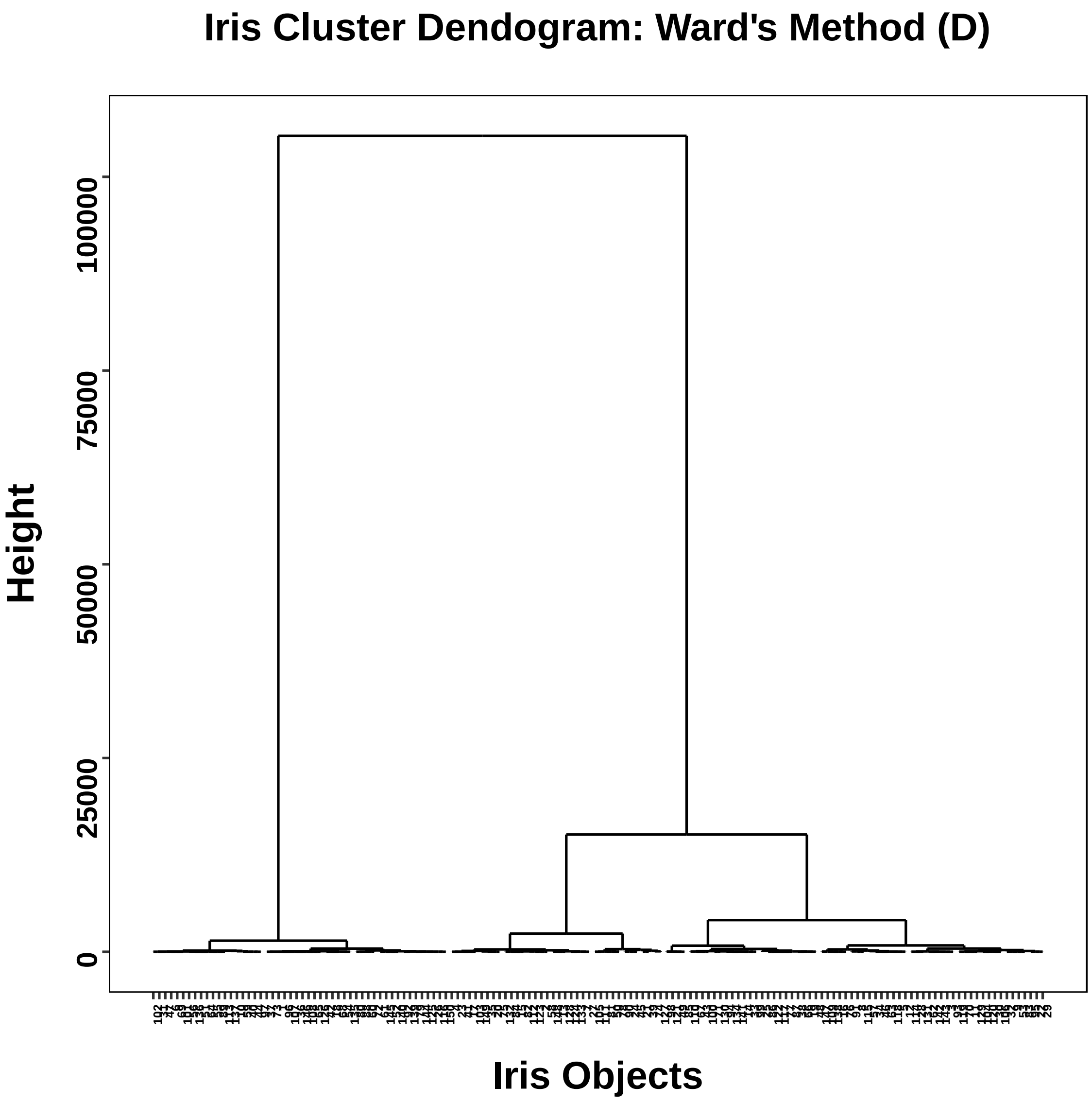
```
ggdendrogram(hc_5, theme_dendro = FALSE) + xlab("Iris Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 5, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Iris Cluster Dendogram: Ward's Method (D2)")
```


Iris Cluster Dendogram: Ward's Method (D2)
Height
0
100
200
300
Iris Objects

3.3.6 Ward's Method (D)

```r
Iris_Dist <- dist(Iris[, 1:4], method = "euclidean")^2

hc_5 <- hclust(Iris_Dist, method = "ward.D")

clustnumber <- cutree(hc_5, k = 3)
Iris_Class <- as.data.frame(cbind(clustnumber, Iris$Species))
names(Iris_Class) <- c("Cluster", "Species")
table(Iris_Class$Species)

##
##     Setosa Versicolor  Virginica
##         50         50         50

table(Iris_Class$Cluster)

##
##  1  2  3
## 50 36 64

table(Iris_Class$Species, Iris_Class$Cluster)

##
##                 1  2  3
##   Setosa       50  0  0
##   Versicolor    0  1 49
##   Virginica     0 35 15
```

```r
ggdendrogram(hc_5, theme_dendro = FALSE) + xlab("Iris Objects") +
    ylab("Height") + theme(text = element_text(size = 14, family = "sans",
    color = "black", face = "bold"), axis.text.y = element_text(colour = "black",
    size = 12, face = "bold"), axis.text.x = element_text(colour = "black",
    size = 5, face = "bold", angle = 90), axis.title.x = element_text(margin = margin(15,
    0, 0, 0), size = 16), axis.title.y = element_text(margin = margin(0,
    15, 0, 0), size = 16, angle = 90), axis.line.x = element_blank(),
    axis.line.y = element_blank(), plot.title = element_text(size = 16,
        face = "bold", margin = margin(0, 0, 20, 0), hjust = 0.5),
    panel.background = element_rect(fill = "white", linetype = 1,
        color = "black"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
    plot.background = element_rect(fill = "white"), plot.margin = unit(c(1,
        1, 1, 1), "cm"), legend.position = "bottom", legend.title = element_blank()) +
    ggtitle("Iris Cluster Dendogram: Ward's Method (D)")
```


Iris Cluster Dendogram: Ward's Method (D)
Height
100000
75000
50000
25000
0
Iris Objects