VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABHINAV VIJAYAKUMAR

19BCE1311

CSE3506 – ESSENTIALS OF DATA ANALYTICS
LAB-8

DR. LAKSHMI PATHI JAKKAMPUTI (L21 + L22)

Tasks for Week-8: Hierarchical Clustering

Aim: To understand the following operations/functions on 'USArrests' data and perform similar operations on 'iris' dataset based on given instructions.

Algorithm:

1. Removing all the values from the global environment

2. Set the working directory to the dataset where we store by using setwd().

3. To see the dataset use view() function.

4. By using scale function, we scale the data and store it in another variable.

5. Using dist function we find the Euclidean distances for the scaled data.

6. By using the Euclidean distances and hclust function we can create and then plot the hierarchical clustering dendogram.

7. By using cutree we divide the elements of the dendogram into k number of clusters (k=4 in our case).

8. Then, by using rect.hclust function we can divide the dendogram into k clusters (k=4 in our case) , i.e. create k rectangular divisions/borders in the dendogram

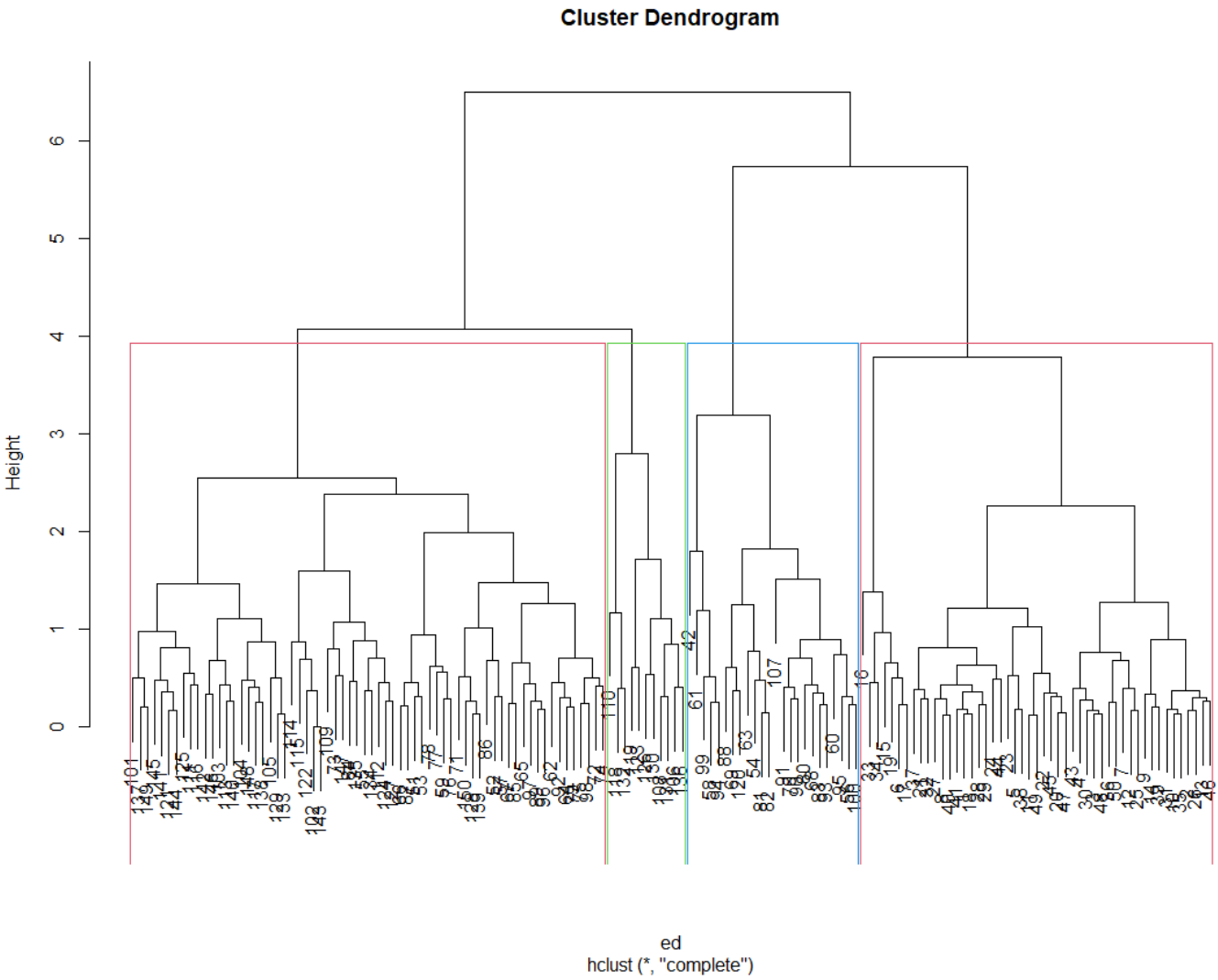
Result:

For iris.csv

Cluster:

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

Cluster Dendogram:

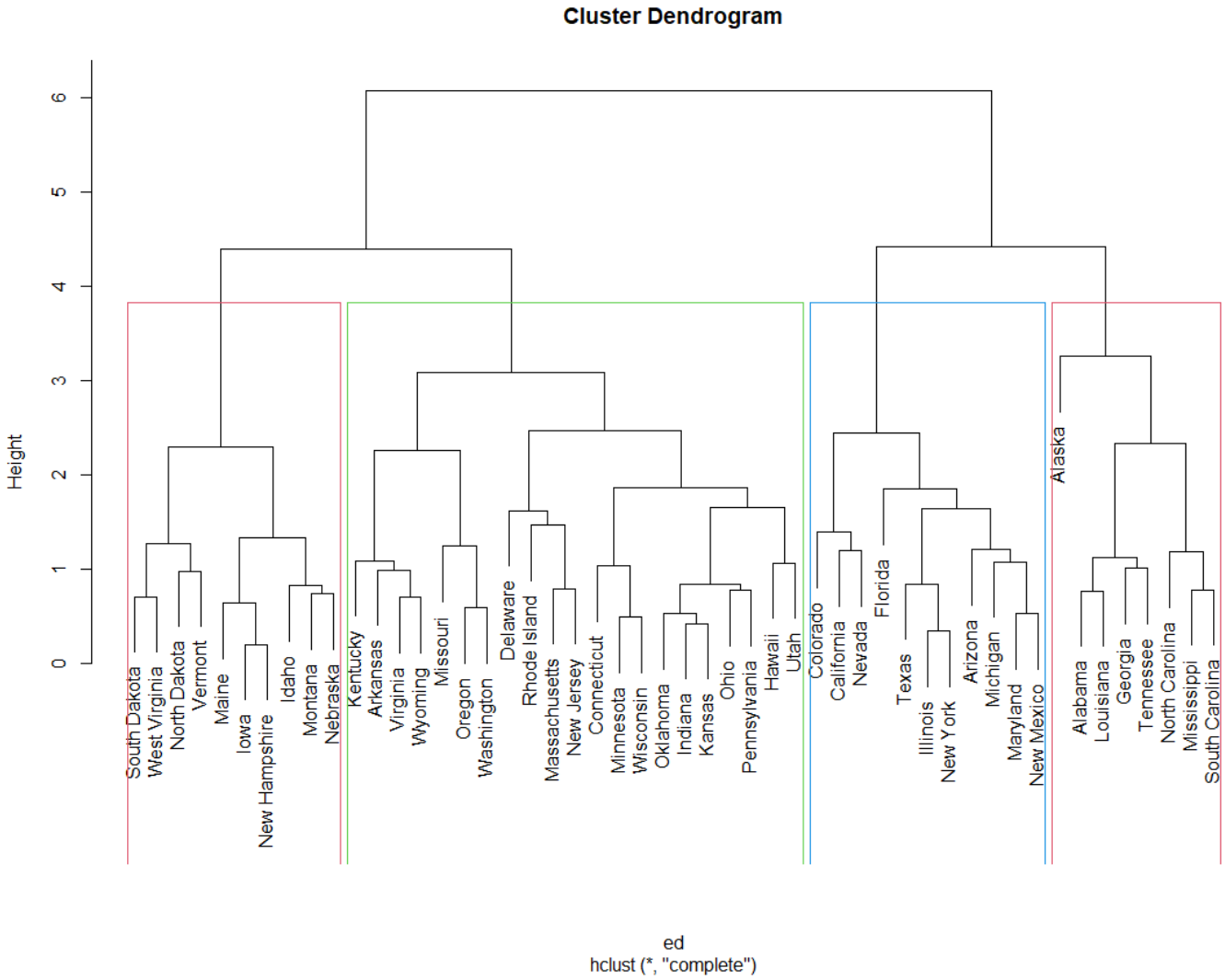


For USArrests.csv

Cluster:

```
> cluster
       Alabama         Alaska        Arizona       Arkansas     California       Colorado    Connecticut       Delaware
             1              1              2              3              2              2              3              3
       Florida        Georgia         Hawaii          Idaho       Illinois        Indiana           Iowa         Kansas
             2              1              3              4              2              3              4              3
      Kentucky      Louisiana          Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi
             3              1              4              2              3              2              3              1
      Missouri        Montana       Nebraska         Nevada New Hampshire     New Jersey     New Mexico       New York
             3              4              4              2              4              3              2              2
North Carolina   North Dakota           Ohio       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
             1              4              3              3              3              3              3              1
  South Dakota      Tennessee          Texas           Utah        Vermont       Virginia     Washington  West Virginia
             4              1              2              3              4              3              3              4
     Wisconsin        Wyoming
             3              3
```

Cluster Dendogram:



Program:

For iris.csv

```
rm(list=ls())
setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem 6/EDA/LAB/LAB 8")

data <- read.csv("iris.csv",row.names=1)
View(data)

df <- scale(data)
View(df)

ed <- dist(df, method = 'euclidean')
```

```
hierClust <- hclust(ed, method = 'complete')
plot(hierClust)

cluster <- cutree(hierClust, k = 4)
cluster

rect.hclust(hierClust, k = 4, border = 2:4)
```

For USArrests.csv

```
rm(list=ls())
setwd("C:/Users/Abhinav Vijayakumar/Desktop/VIT Academics/Sem 6/EDA/LAB/LAB 8")

data <- read.csv("USArrests.csv",row.names=1)
View(data)

df <- scale(data)
View(df)

ed <- dist(df, method = 'euclidean')
hierClust <- hclust(ed, method = 'complete')
plot(hierClust)

cluster <- cutree(hierClust, k = 4)
cluster

rect.hclust(hierClust, k = 4, border = 2:4)
```