MAY 17, 2019

FINAL HOMEWORK

STATISTICAL DATA MINING II

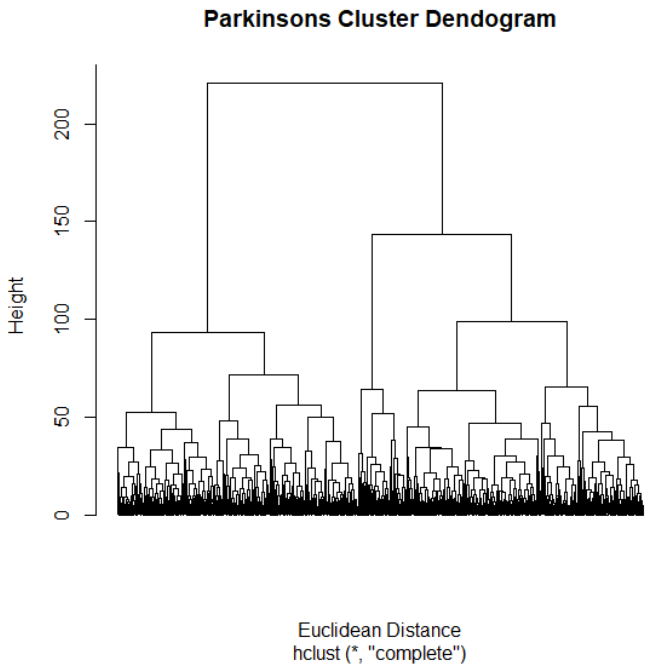
RAVI TEJA SUNKARA
50292191

Question 1

Part A

The clustering methods that will be used to cluster this dataset is hierarchical clustering and k-means clustering.

Heirarchical clustering:

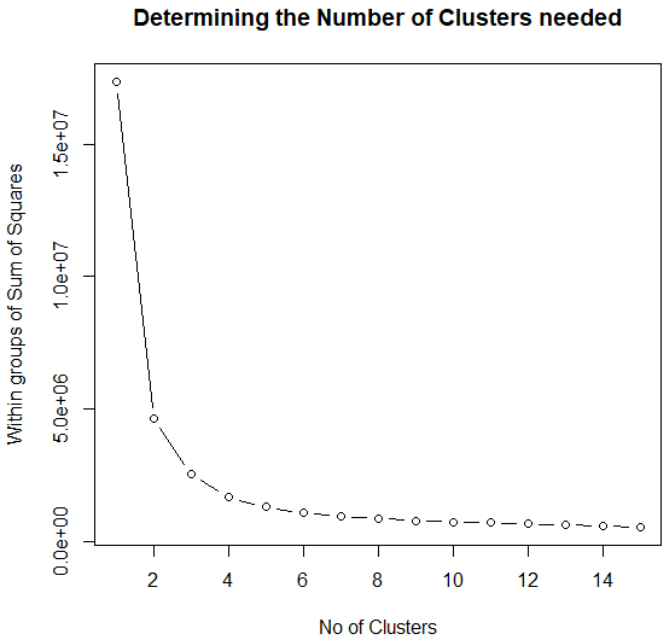


Measuring the capture of data:

```
> table(clustcut, total_UPDRS)
        total_UPDRS
clustcut (0,10] (10,20] (20,30] (30,40] (40,50] (50,60]
      1      23     104     222     187      64      22
      2      10     166     186     173      49      12
      3      49     185     338     342     132      36
      4      59     148     172     221     116      15
      5      11      18     170      95      86      33
      6      17      88     370     175     167      42
      7       1      79     212      57      57      33
      8       2     156     112      98      24       0
      9      19     141     229     156      92       6
      10      0       0      54      24       6      14
```

```
> table(clustcut, motor_UPDRS)
        motor_UPDRS
clustcut (0,10] (10,20] (20,30] (30,40]
      1      58     225     256      83
      2      65     232     250      49
      3      84     401     389     208
      4      83     253     251     144
      5      17     140     130     126
      6      46     343     285     185
      7      14     181     166      78
      8      46     185     134      27
      9      54     291     219      79
      10      0      42      36      20
```

K-means:



The optimal number of clusters look like 3.

Capture of data:

```
> table(kmean3$cluster, total_UPDRS)
   total_UPDRS
    (0,10] (10,20] (20,30] (30,40] (40,50] (50,60]
1     113      342      569      523      305       54
2      54      502      716      634      205       58
3      24      241      780      371      283      101
```

```
> table(kmean4$cluster, motor_UPDRS)
   motor_UPDRS
    (0,10] (10,20] (20,30] (30,40]
1      53      558      486      241
2     149      586      569      161
3     137      589      493      328
4     128      560      568      269
```

Part B

We are removing 'motor_UPDRS' and are forcing the variable to be at the bottom.

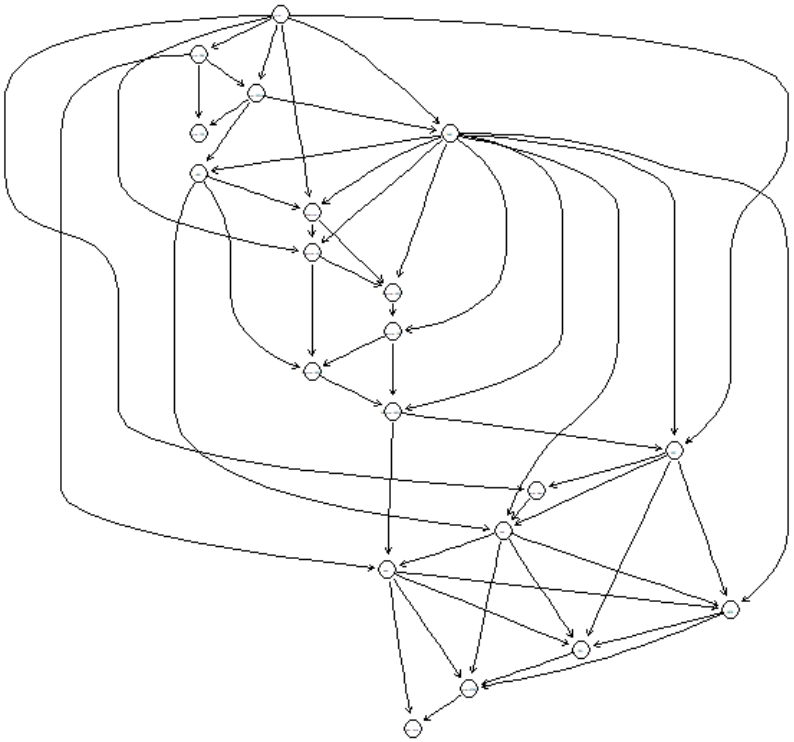


Fig: Graph without forcing the variable ‘total_UPDRS’ to the bottom.

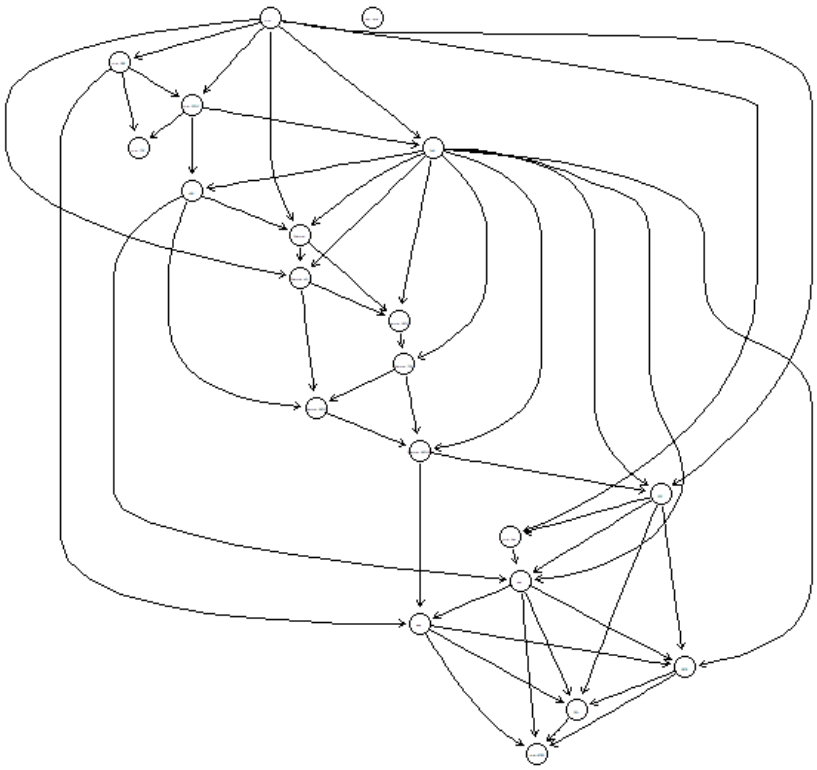


Fig: Graph with ‘total_UPDRS’ at the bottom

Using the Bayesian Network to characterize 'Jitter' related variables for a new patient with out any evidence and with evidence can be seen below.

```
> no_ev
$Jitter.RAP
Jitter.RAP
 (0,0.006]  (0.006,1]
0.93368717 0.06631283

$Jitter...
Jitter...
 (0,0.01]  (0.01,1]
0.9044418 0.0955582

$Jitter.Abs.
Jitter.Abs.
(0,8e-05] (8e-05,1]
0.8985813 0.1014187

$Jitter.PPQ5
Jitter.PPQ5
 (0,0.007]  (0.007,1]
0.95145754 0.04854246

$Jitter.DDP
Jitter.DDP
  (0,0.02]   (0.02,1]
0.94520458 0.05479542
```

```
> with_ev
$Jitter.RAP
Jitter.RAP
 (0,0.006]  (0.006,1]
0.95948939 0.04051061

$Jitter...
Jitter...
  (0,0.01]   (0.01,1]
0.90291041 0.09708959

$Jitter.Abs.
Jitter.Abs.
(0,8e-05] (8e-05,1]
0.8668996 0.1331004

$Jitter.PPQ5
Jitter.PPQ5
 (0,0.007]  (0.007,1]
0.97222362 0.02777638

$Jitter.DDP
Jitter.DDP
  (0,0.02]   (0.02,1]
0.96767553 0.03232447
```

Question 2

This dataset was collected by the British Board of Trade to investigate the sinking. Features of the surviving passengers was analysed to see if women and children were evacuated first.

For survived passengers, there are a set of 129 rules. The support threshold was taken as 0.01 to accommodate more rules and the confidence of 0.6.

```
set of 128 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5  6
 8 32 50 32  6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   4.000   3.969   5.000   6.000

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.01053   Min.   :0.6029   Min.   :1.003   Min.   : 11.0
 1st Qu.:0.01627   1st Qu.:0.7128   1st Qu.:1.186   1st Qu.: 17.0
 Median :0.03589   Median :0.7876   Median :1.311   Median : 37.5
 Mean   :0.11359   Mean   :0.7992   Mean   :1.330   Mean   :118.7
 3rd Qu.:0.13541   3rd Qu.:0.8965   3rd Qu.:1.492   3rd Qu.:141.5
 Max.   :0.53971   Max.   :0.9697   Max.   :1.614   Max.   :564.0

mining info:
                data ntransactions support confidence
 titanic_data_matrix          1045    0.01        0.6
```

```
     lhs                                                  rhs           support    confidence lift     count
[1]  {Pclass=2,Sex=male,Age=adult,SibSp=1}                => {Survived=yes} 0.03062201 0.9696970  1.613588  32
[2]  {Pclass=2,Sex=male,Age=adult}                        => {Survived=yes} 0.12918660 0.9642857  1.604584 135
[3]  {Pclass=2,Sex=male,Age=adult,SibSp=0}                => {Survived=yes} 0.09282297 0.9603960  1.598111  97
[4]  {Pclass=2,Sex=male,Age=adult,Parch=0}                => {Survived=yes} 0.11578947 0.9603175  1.597980 121
[5]  {Pclass=2,Sex=male,Age=adult,SibSp=0,Parch=0}        => {Survived=yes} 0.08995215 0.9591837  1.596094  94
[6]  {Pclass=2,Sex=male,SibSp=1,Parch=0}                  => {Survived=yes} 0.02200957 0.9583333  1.594679  23
[7]  {Pclass=2,Sex=male,Age=adult,SibSp=1,Parch=0}        => {Survived=yes} 0.02105263 0.9565217  1.591664  22
[8]  {Pclass=2,Sex=male,Parch=0}                          => {Survived=yes} 0.12057416 0.9545455  1.588376 126
[9]  {Pclass=2,Sex=male,SibSp=0,Parch=0}                  => {Survived=yes} 0.09377990 0.9514563  1.583235  98
[10] {Sex=male,Age=adult,SibSp=2}                         => {Survived=yes} 0.01531100 0.9411765  1.566130  16
```

We can see that the rules with high confidence or lift, that is above 90%, of the surviving passengers describe very less about the Women and Children being evacuated first or survived.

The next ten rules were also inspected in order to check whether there is any other evidence that can be captured to support the survived passengers.

```
[11] {Pclass=3,Sex=male,Age=adult,Parch=1}        => {Survived=yes} 0.01435407 0.9375000  1.560012  15
[12] {Pclass=2,Sex=male,SibSp=0}                  => {Survived=yes} 0.09760766 0.9357798  1.557150 102
[13] {Sex=male,SibSp=4}                           => {Survived=yes} 0.01339713 0.9333333  1.553079  14
[14] {Pclass=3,Sex=male,SibSp=4}                  => {Survived=yes} 0.01339713 0.9333333  1.553079  14
[15] {Sex=male,SibSp=2,Parch=0}                   => {Survived=yes} 0.01339713 0.9333333  1.553079  14
[16] {Sex=male,Age=adult,SibSp=2,Parch=0}         => {Survived=yes} 0.01339713 0.9333333  1.553079  14
[17] {Sex=male,Age=child,SibSp=4}                 => {Survived=yes} 0.01148325 0.9230769  1.536012  12
[18] {Pclass=3,Sex=male,Age=child,SibSp=4}        => {Survived=yes} 0.01148325 0.9230769  1.536012  12
[19] {Sex=male,Age=adult,SibSp=1,Parch=1}         => {Survived=yes} 0.02296651 0.9230769  1.536012  24
[20] {Pclass=3,Sex=male,Age=adult,SibSp=1}        => {Survived=yes} 0.03253589 0.9189189  1.529093  34
```

Summary:

There are a set of 112 rules for passengers who did not survive the disaster. The support threshold was taken as 0.01 and confidence as 0.6.

```
set of 112 rules

rule length distribution (lhs + rhs):sizes
 2  3  4  5  6
 1 25 49 30  7

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   4.000   4.000   4.152   5.000   6.000

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.01053   Min.   :0.6000   Min.   :1.504   Min.   : 11.00
 1st Qu.:0.01603   1st Qu.:0.7195   1st Qu.:1.803   1st Qu.: 16.75
 Median :0.02871   Median :0.8562   Median :2.146   Median : 30.00
 Mean   :0.04711   Mean   :0.8323   Mean   :2.086   Mean   : 49.23
 3rd Qu.:0.05478   3rd Qu.:0.9353   3rd Qu.:2.344   3rd Qu.: 57.25
 Max.   :0.31005   Max.   :1.0000   Max.   :2.506   Max.   :324.00

mining info:
               data ntransactions support confidence
 titanic_data_matrix          1045    0.01        0.6
```

	lhs		rhs	support	confidence	lift	count
[1]	{Sex=female,Age=senior}	=>	{Survived=no}	0.01052632	1	2.505995	11
[2]	{Pclass=2,Sex=female,Parch=2}	=>	{Survived=no}	0.01722488	1	2.505995	18
[3]	{Pclass=2,Sex=female,Age=child}	=>	{Survived=no}	0.01339713	1	2.505995	14
[4]	{Pclass=1,Sex=female,Parch=1}	=>	{Survived=no}	0.02679426	1	2.505995	28
[5]	{Pclass=2,Sex=female,Age=adult,Parch=2}	=>	{Survived=no}	0.01148325	1	2.505995	12
[6]	{Pclass=1,Sex=female,SibSp=1,Parch=1}	=>	{Survived=no}	0.01148325	1	2.505995	12
[7]	{Pclass=1,Sex=female,SibSp=0,Parch=1}	=>	{Survived=no}	0.01531100	1	2.505995	16
[8]	{Pclass=1,Sex=female,Age=adult,Parch=1}	=>	{Survived=no}	0.02583732	1	2.505995	27
[9]	{Pclass=1,Sex=female,SibSp=1,Parch=0}	=>	{Survived=no}	0.03444976	1	2.505995	36
[10]	{Pclass=1,Sex=female,Age=adult,SibSp=1,Parch=1}	=>	{Survived=no}	0.01052632	1	2.505995	11

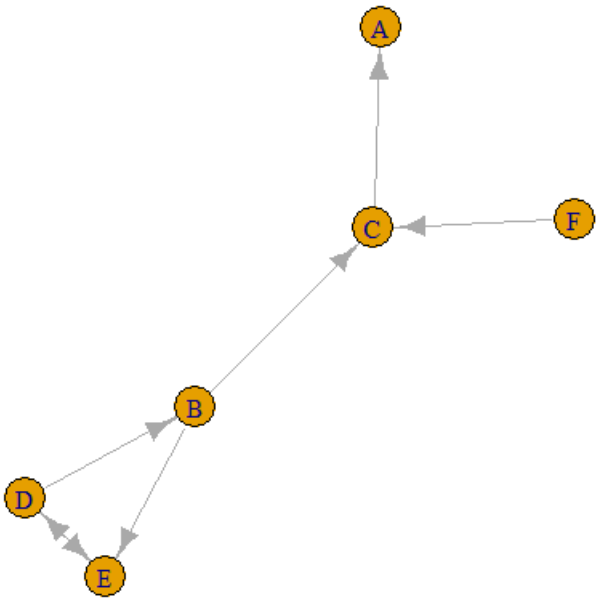
All the above displayed rules produce a confidence of 100% for certain aspects of dead passengers.

[11]	{Pclass=1,Sex=female,Age=adult,SibSp=0,Parch=1}	=>	{Survived=no}	0.01531100	1.0000000	2.505995	16
[12]	{Pclass=1,Sex=female,Age=adult,SibSp=1,Parch=0}	=>	{Survived=no}	0.03062201	1.0000000	2.505995	32
[13]	{Pclass=1,Sex=female,Parch=0}	=>	{Survived=no}	0.08229665	0.9885057	2.477191	86
[14]	{Pclass=1,Sex=female,Age=adult,Parch=0}	=>	{Survived=no}	0.07655502	0.9876543	2.475057	80
[15]	{Pclass=1,Sex=female,SibSp=0}	=>	{Survived=no}	0.06602871	0.9857143	2.470195	69
[16]	{Pclass=1,Sex=female,Age=adult,SibSp=0}	=>	{Survived=no}	0.06315789	0.9850746	2.468592	66
[17]	{Pclass=1,Sex=female,Age=adult}	=>	{Survived=no}	0.11578947	0.9837398	2.465247	121
[18]	{Pclass=1,Sex=female,Age=adult,SibSp=1}	=>	{Survived=no}	0.04497608	0.9791667	2.453787	47
[19]	{Pclass=1,Sex=female,SibSp=0,Parch=0}	=>	{Survived=no}	0.04497608	0.9791667	2.453787	47
[20]	{Pclass=1,Sex=female,Age=adult,SibSp=0,Parch=0}	=>	{Survived=no}	0.04306220	0.9782609	2.451517	45

These two evidences give enough evidence that Women and Children were not the first to be evacuated from the ship.

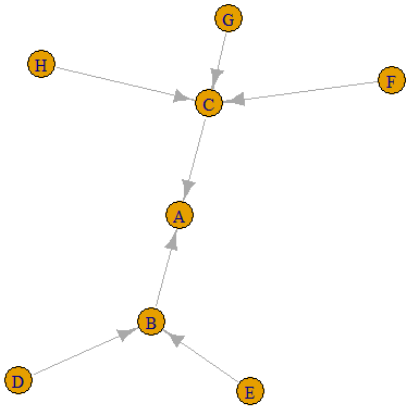
Question 3

Part A




```
        A         B         C         D         E         F
0.1683271 0.1639395 0.1718214 0.1681380 0.1680380 0.1597361
        A         B         C         D         E         F
0.1786588 0.1544288 0.1848587 0.1758772 0.1737324 0.1324441
        A         B         C         D         E         F
0.19227231 0.14719411 0.18583257 0.19135235 0.18399264 0.09935603
        A         B         C         D         E         F
0.19540224 0.14684475 0.17515044 0.21156522 0.19824042 0.07279693
        A         B         C         D         E         F
0.16458660 0.16116326 0.13713449 0.26735313 0.24093907 0.02882346
```

As damping factor increases, page rank value is more sensitive to the number of incoming links. For F point, it can be seen that the rank value goes on decreasing as number of incoming links reduces. It holds more importance to the number of outgoing links than incoming links. C has 2 incoming and 1 outgoing link, but the rank value is decreasing. But for A the rank value decreases for very high damping factor. As expected, its value increases with the damping factor expect that case.

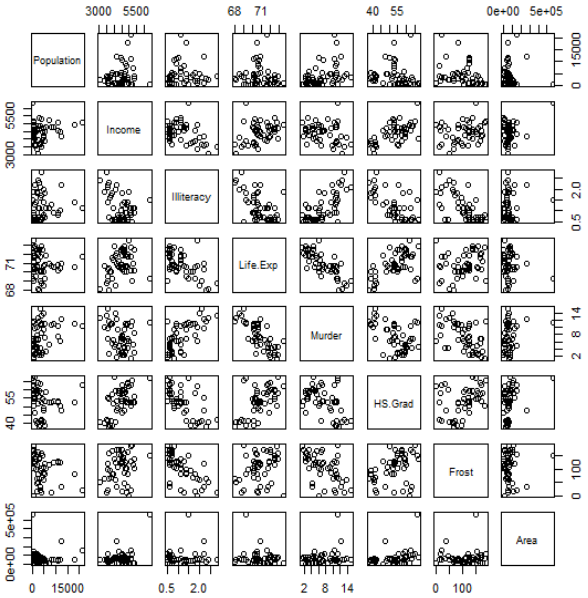



```
        A         B         C         D         E         F         G         H
0.1541610 0.1418827 0.1582538 0.1091405 0.1091405 0.1091405 0.1091405 0.1091405
```

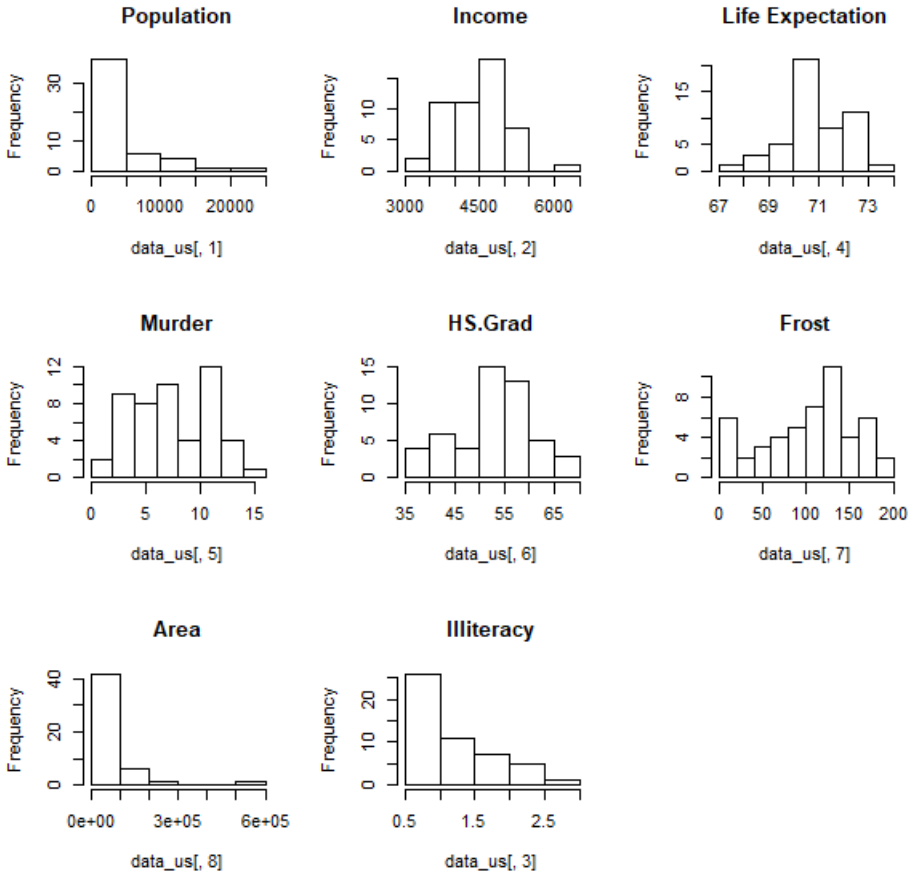
As it can be seen from the graph, C has most number of incoming links that is 3 incoming links. Therefore, as expected it has maximum rank value. Even though it is pointing to A, A has only 2 incoming links. So the number of incoming links dominate over number of outgoing links.

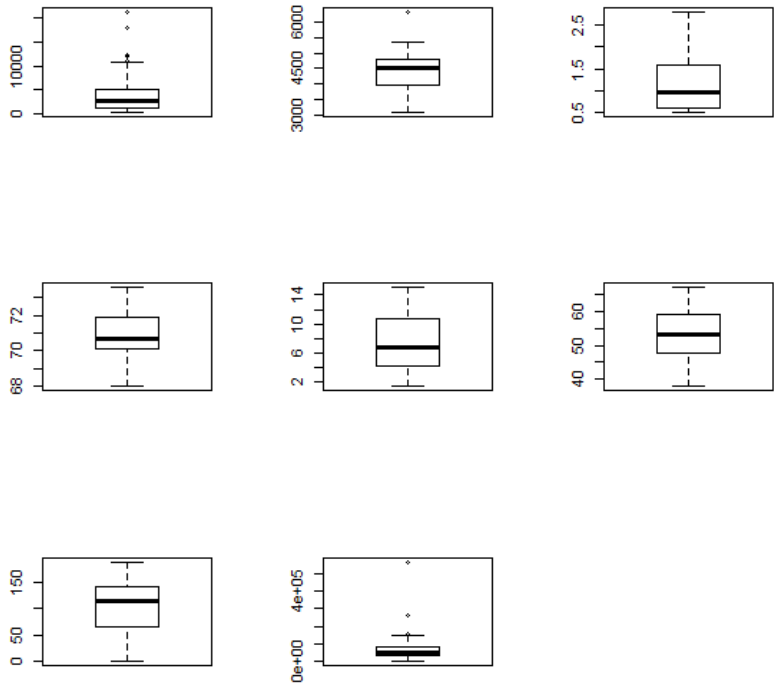
Question 4

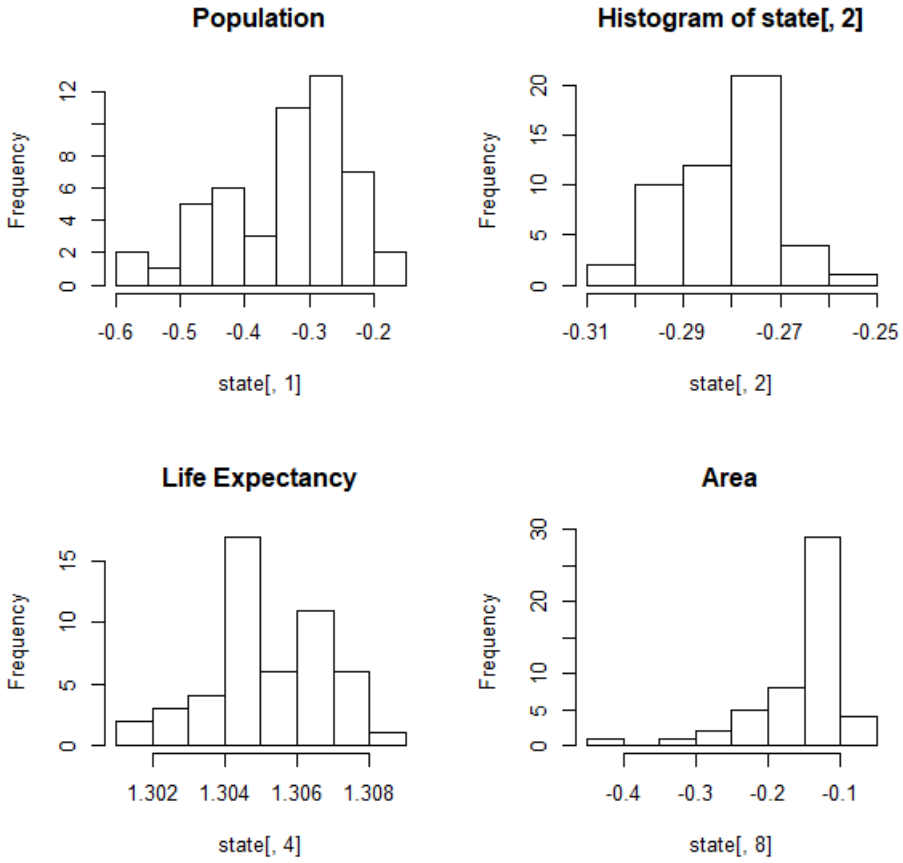
The data set describes the Census of the 50 states in the United States in the 1970s. There are 8 variables or features such as Population, Income, Illiteracy, Life Expectancy, Murder, HS Grad, Frost and Area.



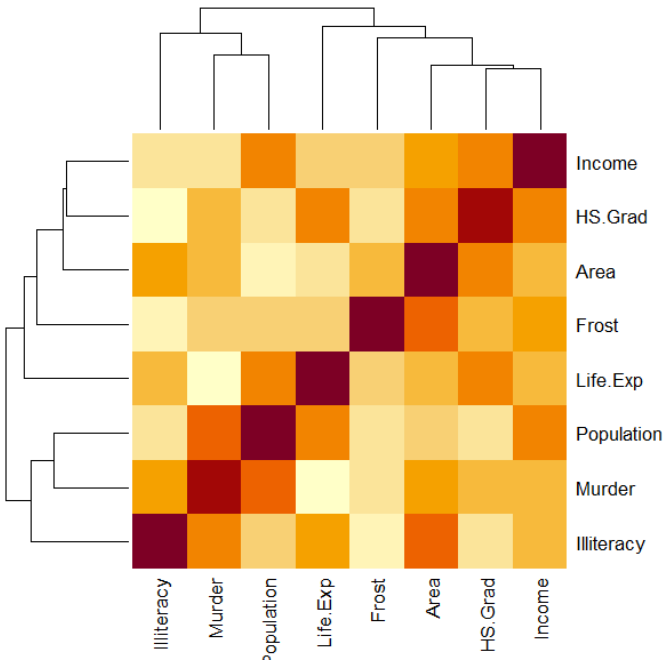
Here are some interesting plots obtained from the data.



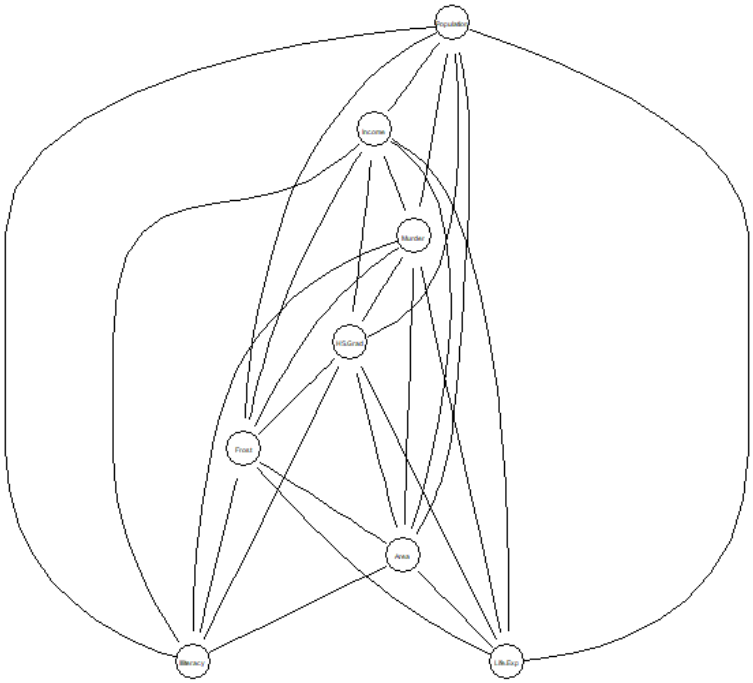




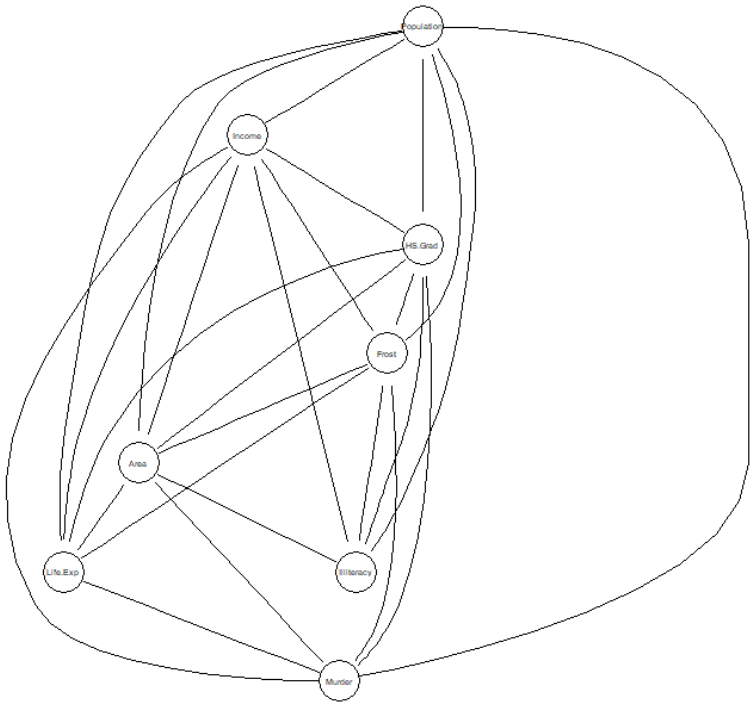
The partial covariance and covariance were obtained for the data and a heatmap was created.



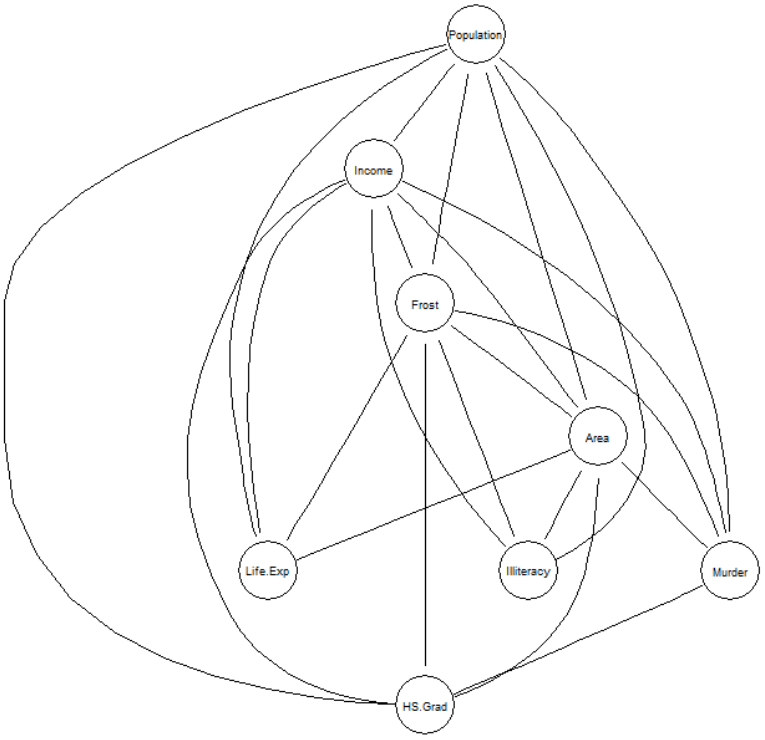
We can infer some useful results from the heatmap. It can be seen that Life Expectancy and Murder have high negative correlation, which is an obvious relationship.



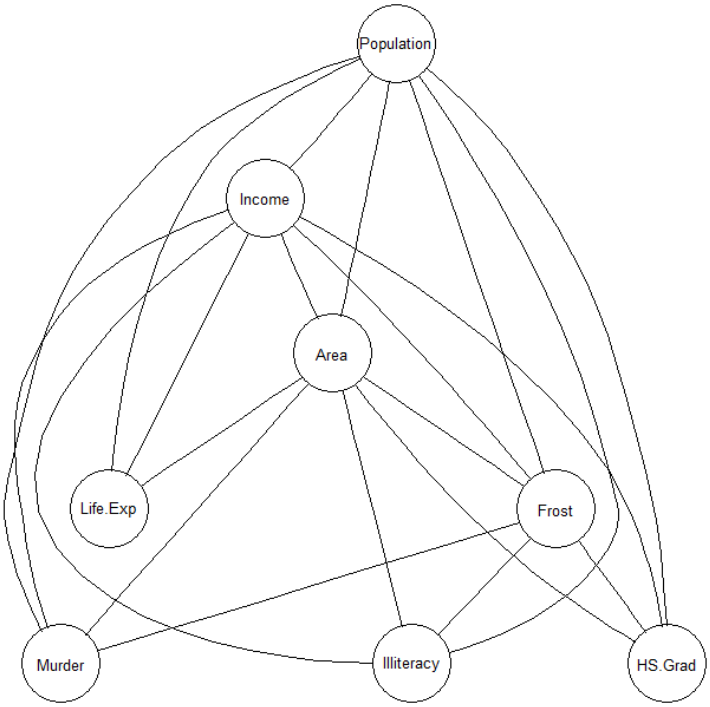
Rho = 1



Rho = 5

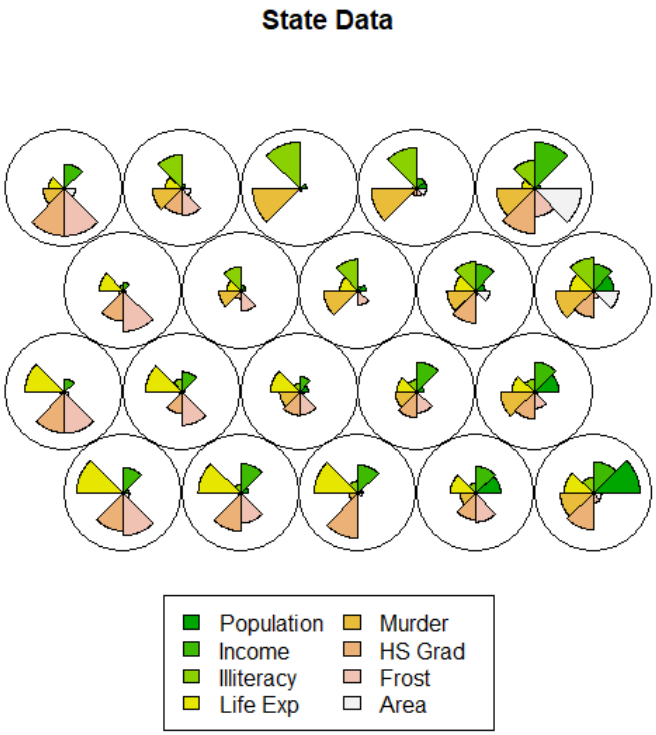

Population
Income
Frost
Area
Life.Exp
Illiteracy
Murder
HS.Grad

Rho: 10


Population
Income
Area
Life.Exp
Frost
Murder
Illiteracy
HS.Grad

We can see correlation goes on decreasing. But some relations remain the same for all the values of rho.

Correlation heatmap generated using glasso model shows equivalence with the graphical models generated.



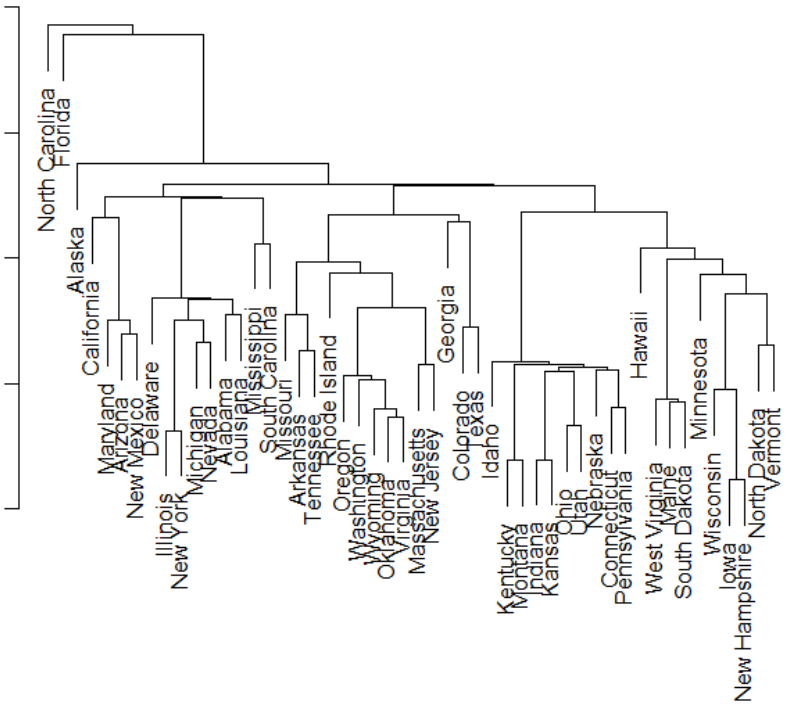
For grid size 20, we see same correlation as that of Glasso model. So results of som compliments results of glasso.

Correlation heatmap generated using glasso model shows equivalence with the graphical models generated.

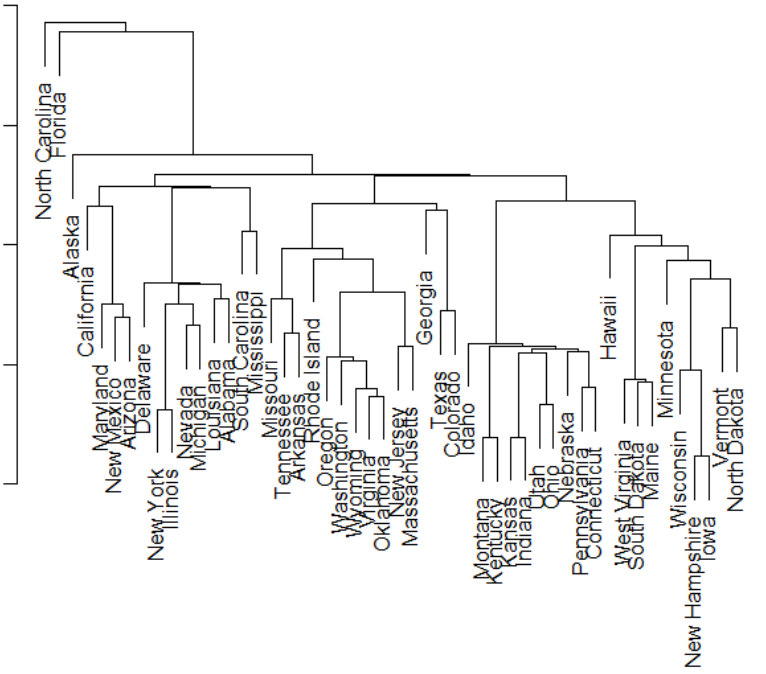
Question 5

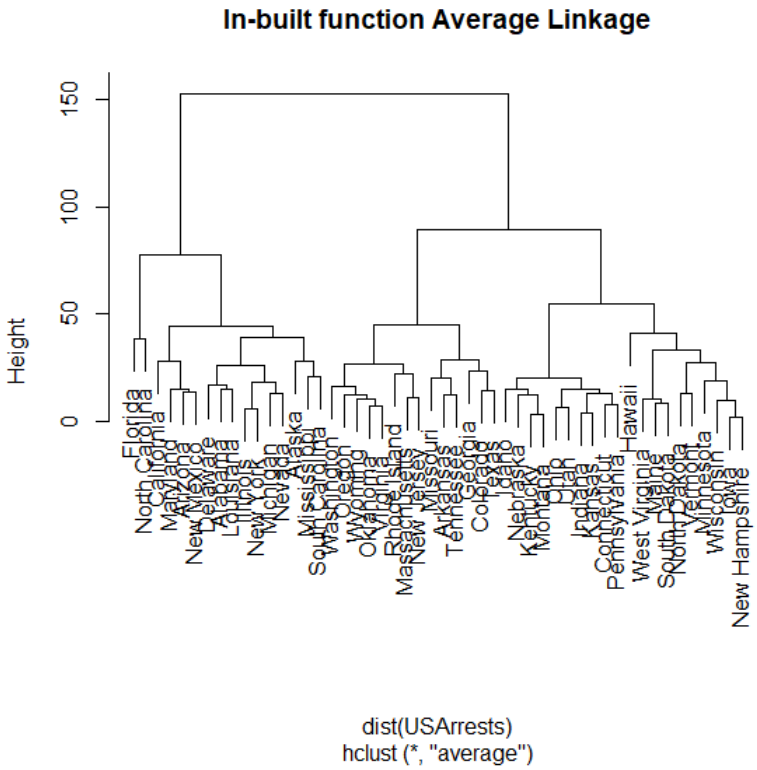
Comparison between In-built and Our Defined Functions. From the below results we can see that both are same. Therefore, we have successfully implemented the single, complete and average linkage agglomerative clustering algorithms from scratch in R.

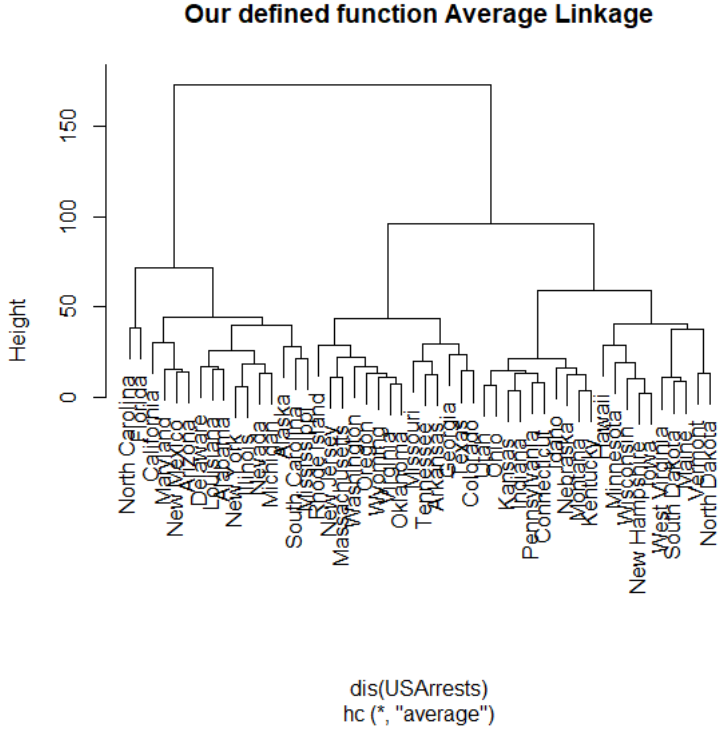
In-built function Single Linkage



Our defined function Single Linkage




In-built function Average Linkage
Height
0
50
100
150
dist(USArrests)
hclust (*, "average")


Our defined function Average Linkage
Height
0
50
100
150
dis(USArrests)
hc (*, "average")