EAS 507 : Statistical Data Mining II

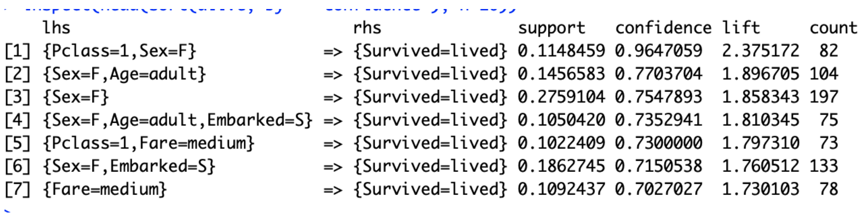
Homework 4

Shubham Sharma (Person No.: 50290293, Class No. 43)

May 18, 2019

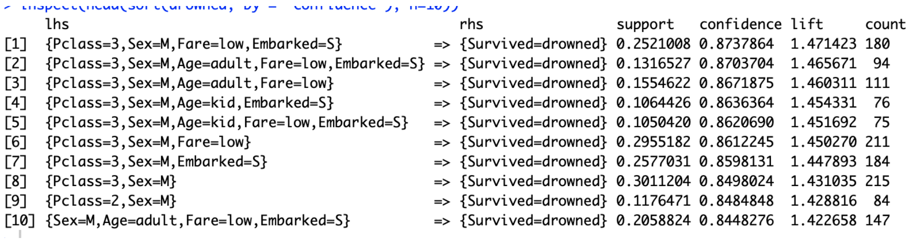
Solution 2: The Titanic data set was used to analyze if the policy of women and children first was successfully followed. Association rules were used because it helps us to see the important features with conditional probabilities. Survival = 0 or 1 was the target aimed for.

We ran apriori algorithm on the data and filtered the rules with Lift $1 forRHSSurvived = 0andSurvived = 1$.



```
    lhs                              rhs              support   confidence lift     count
[1] {Pclass=1,Sex=F}              => {Survived=lived} 0.1148459 0.9647059  2.375172  82
[2] {Sex=F,Age=adult}             => {Survived=lived} 0.1456583 0.7703704  1.896705 104
[3] {Sex=F}                       => {Survived=lived} 0.2759104 0.7547893  1.858343 197
[4] {Sex=F,Age=adult,Embarked=S} => {Survived=lived} 0.1050420 0.7352941  1.810345  75
[5] {Pclass=1,Fare=medium}        => {Survived=lived} 0.1022409 0.7300000  1.797310  73
[6] {Sex=F,Embarked=S}            => {Survived=lived} 0.1862745 0.7150538  1.760512 133
[7] {Fare=medium}                 => {Survived=lived} 0.1092437 0.7027027  1.730103  78
```

Figure 1: People who survived



```
     lhs                                           rhs                support   confidence lift     count
[1]  {Pclass=3,Sex=M,Fare=low,Embarked=S}           => {Survived=drowned} 0.2521008 0.8737864  1.471423 180
[2]  {Pclass=3,Sex=M,Age=adult,Fare=low,Embarked=S} => {Survived=drowned} 0.1316527 0.8703704  1.465671  94
[3]  {Pclass=3,Sex=M,Age=adult,Fare=low}            => {Survived=drowned} 0.1554622 0.8671875  1.460311 111
[4]  {Pclass=3,Sex=M,Age=kid,Embarked=S}            => {Survived=drowned} 0.1064426 0.8636364  1.454331  76
[5]  {Pclass=3,Sex=M,Age=kid,Fare=low,Embarked=S}   => {Survived=drowned} 0.1050420 0.8620690  1.451692  75
[6]  {Pclass=3,Sex=M,Fare=low}                      => {Survived=drowned} 0.2955182 0.8612245  1.450270 211
[7]  {Pclass=3,Sex=M,Embarked=S}                    => {Survived=drowned} 0.2577031 0.8598131  1.447893 184
[8]  {Pclass=3,Sex=M}                               => {Survived=drowned} 0.3011204 0.8498024  1.431035 215
[9]  {Pclass=2,Sex=M}                               => {Survived=drowned} 0.1176471 0.8484848  1.428816  84
[10] {Sex=M,Age=adult,Fare=low,Embarked=S}          => {Survived=drowned} 0.2058824 0.8448276  1.422658 147
```

Figure 2: People who did not survived

Solution 3: The Webgraph A is shown in figure **??**, webgraph B in figure **??** and the page-rank vectors for damping factors: 0.05, 0.25, 0.50, 0.75, 0.95 is shown in figure 5.

We see in figure 5 that the ranks we get in the page rank vectors align with the graph for damping factors 0.05 and 0.25. Page C is crucial as per the webgraph, however in the the vectors D has more loading than C which is counter intuitive. We observe that the damping factor keeps up with the graph upto a certain point after which less important pages gain more importance. This is due to the random clicks which is highly likely to occur.

Figure 6 shows that webpage C is of the highest rank, followed by A and B. Since the webpages D, E, F, G and H are the leaf nodes, i.e. no incoming nodes, their ranks are exactly the same and the lowest.

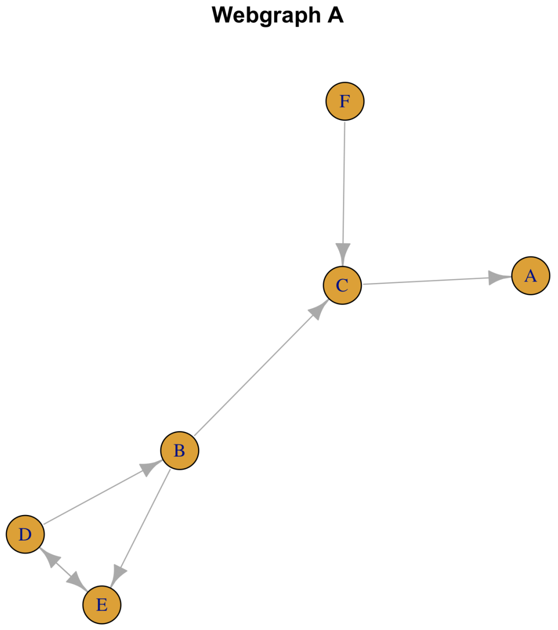


Figure 3: Webgraph A

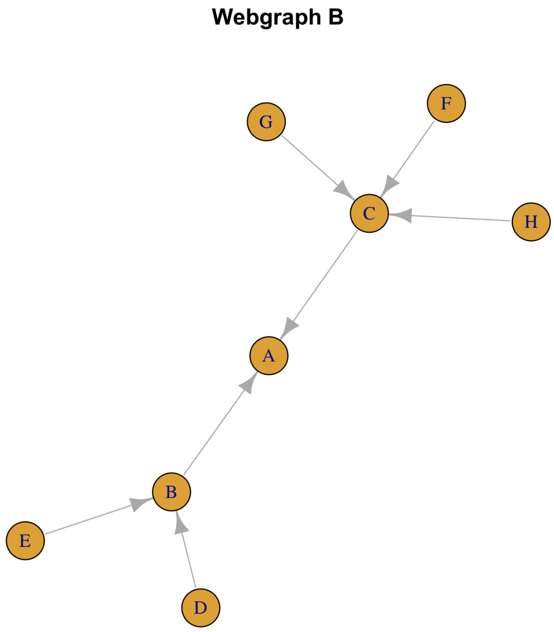
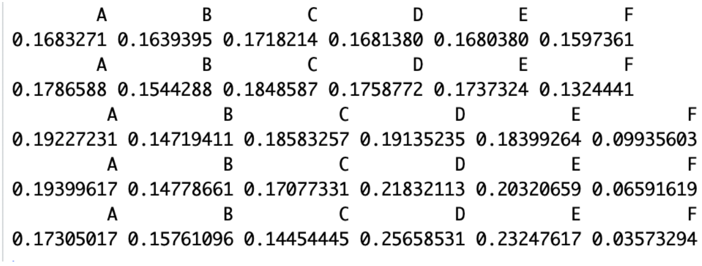
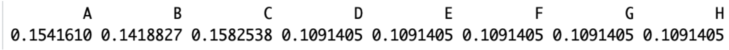


Figure 4: Webgraph B



```
        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 
0.19399617 0.14778661 0.17077331 0.21832113 0.20320659 0.06591619 
         A          B          C          D          E          F 
0.17305017 0.15761096 0.14454445 0.25658531 0.23247617 0.03573294 
```

Figure 5: Pagerank vectors : part a



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

Figure 6: Pagerank vectors : part b

Question 4: Data released from the US department of Commerce, Bureau of the Census is available in R. Build a Gaussian Graphical Model using the Graphical Lasso for the 8 predictors (Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. What do you find for different penalties, and how does it compliment (and/or contradict) a model fit with SOM?

The dataset *state.x*77 has 50 observations of 8 variables which is data related to the 50 states of the United States of America.

The plots for Graphical Lasso using different penalties (rho value) is shown in figures **??**, 7, 5, 8 and 6.

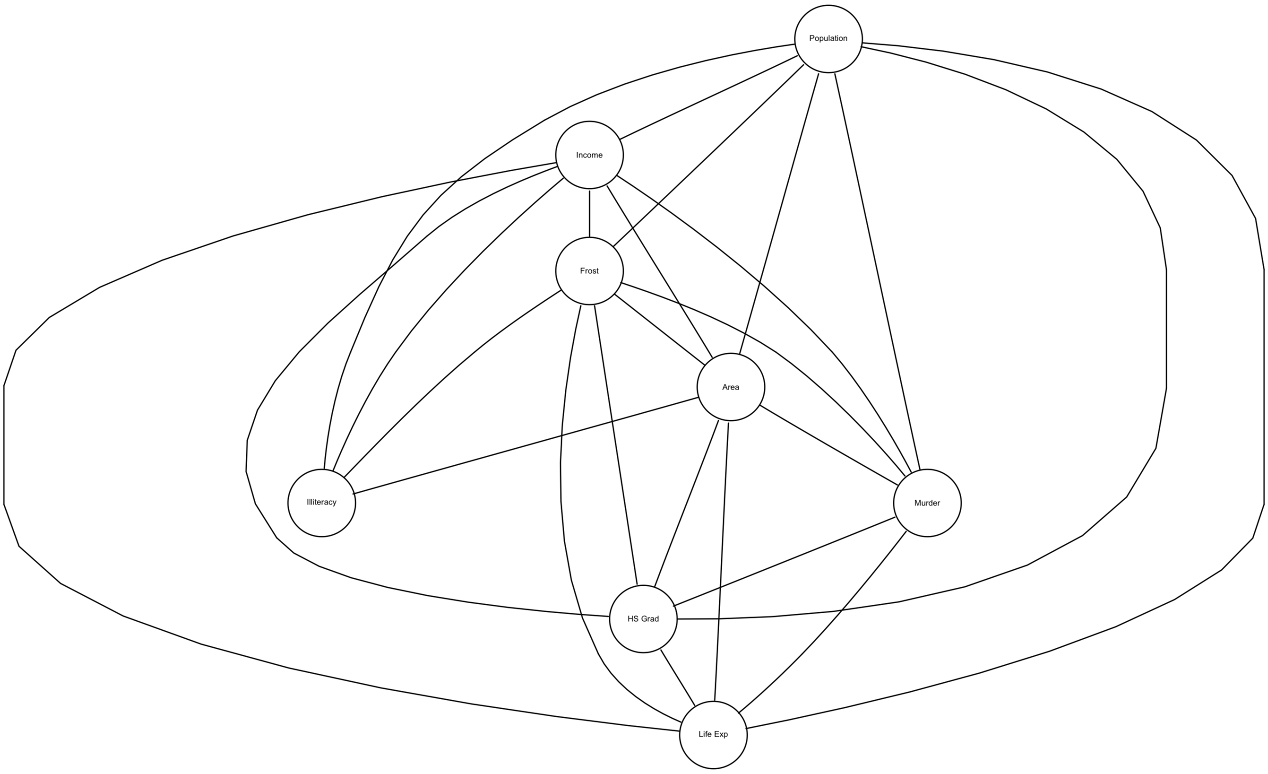


Figure 7: Plot for Graphical Lasso for $rho = 2$

We observe a general relationship that exists between different variables over different values of rho in the shown plots.

As is clear we see that the connection of Life expectancy and Frost decreases as we increase the penalty. Also, HS Grad and Murder get disconnected as we increase the penalty. The variables Population, Income and Area retain their high connectivity to other nodes (7 in number) throughout the process of increasing the penalty and act as hub nodes. Also, Illiteracy has same connectivity to other nodes (4 in number) throughout the process.

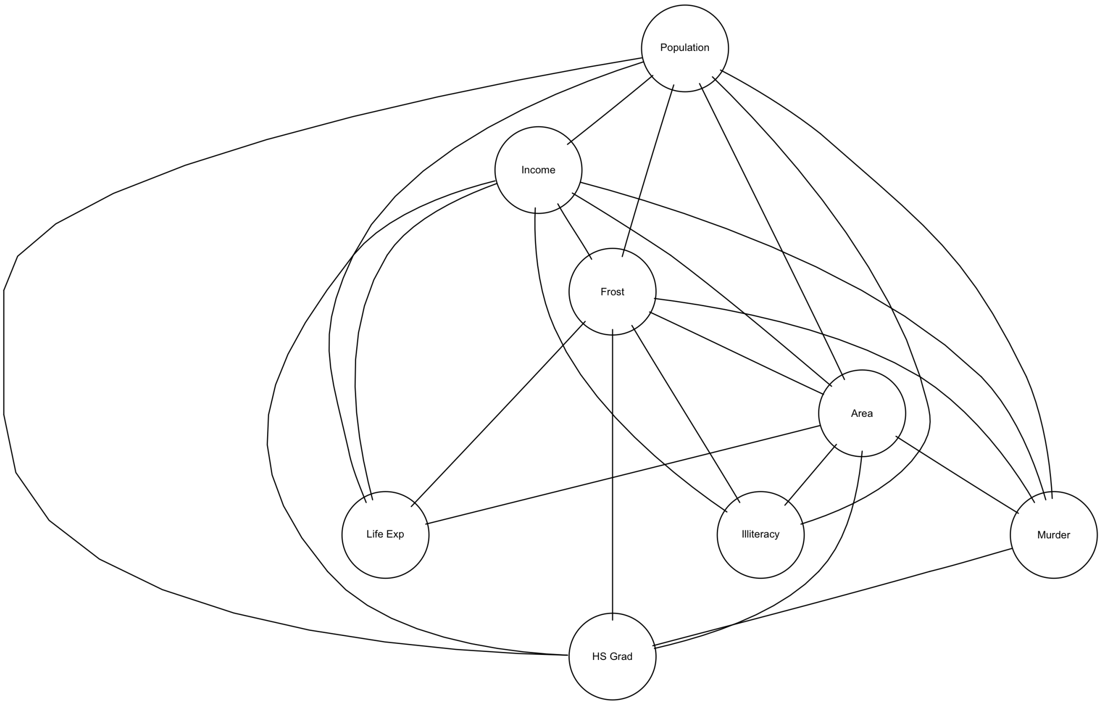


Figure 8: Plot for Graphical Lasso for $rho = 4$

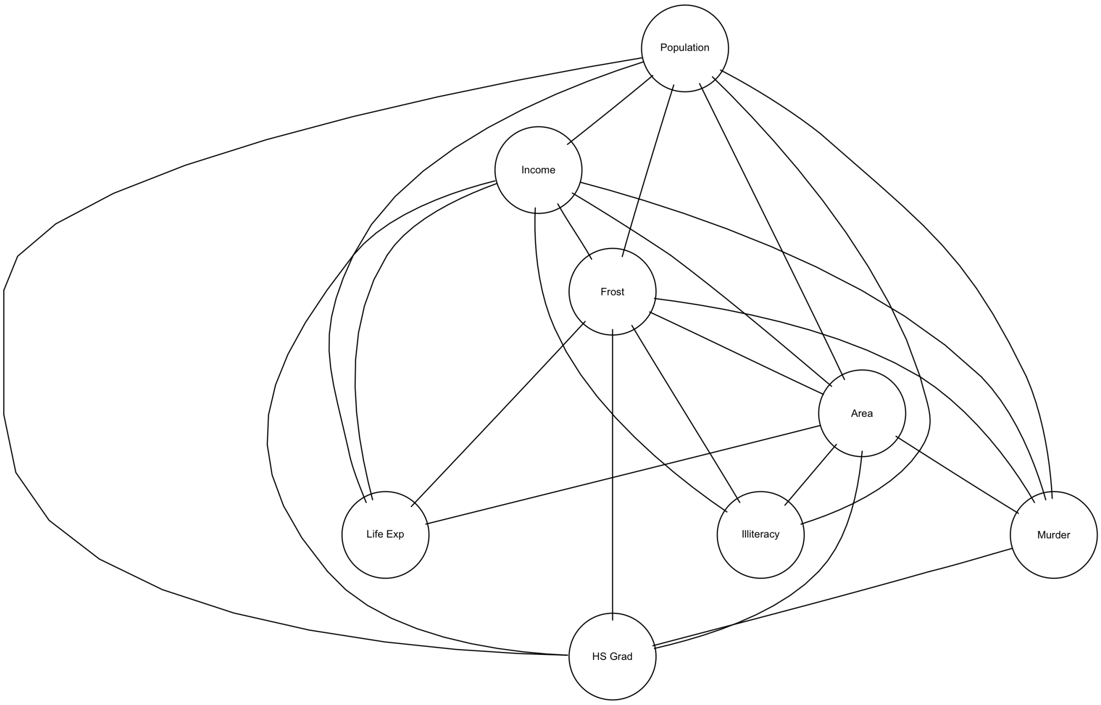


Figure 9: Plot for Graphical Lasso for $rho = 6$

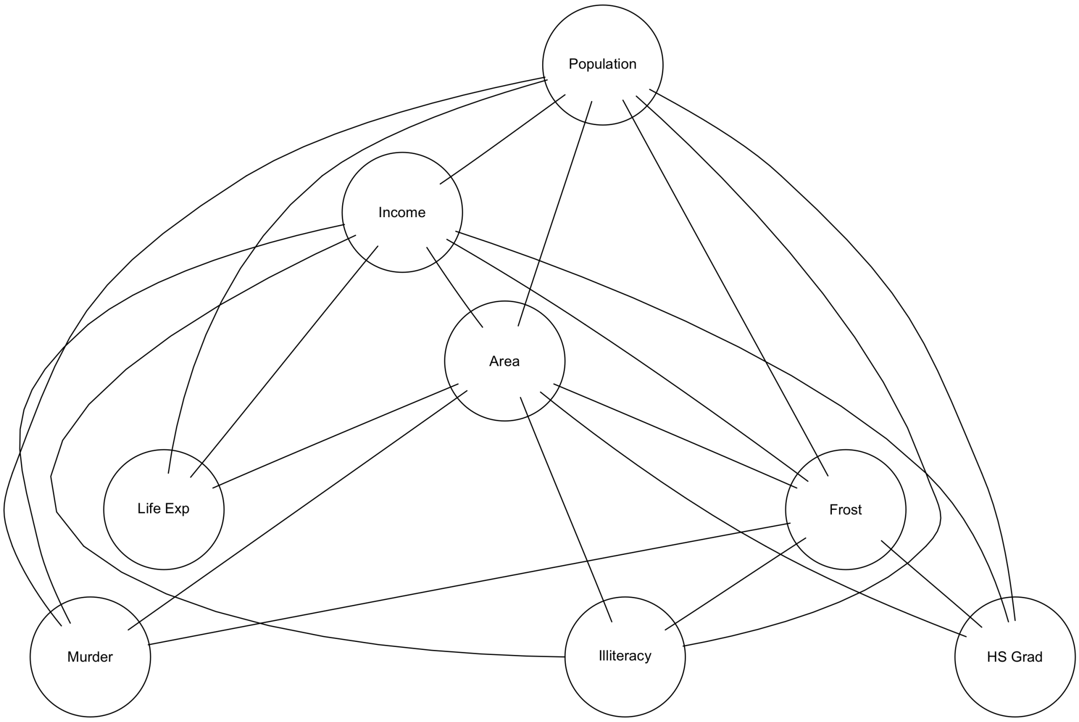


Figure 10: Plot for Graphical Lasso for $rho = 10$

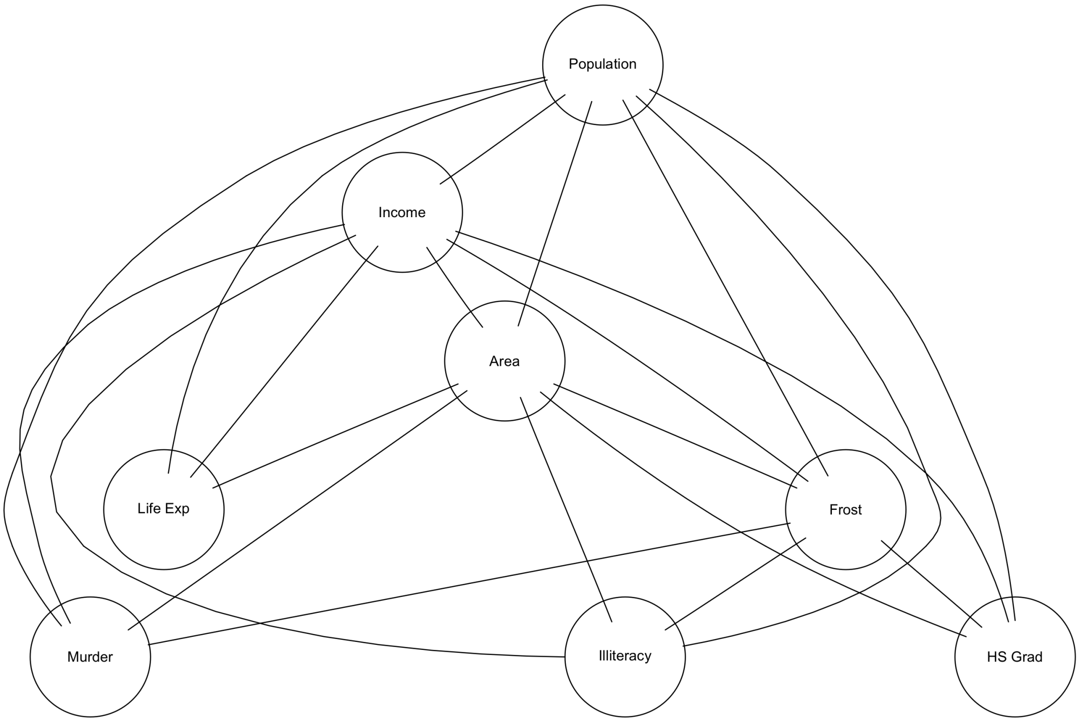


Figure 11: Plot for Graphical Lasso for $rho = 15$

Solution 5: The minimum distance from every point in the cluster is used for single linkage. For complete linkage, The maximum distance from each points in the cluster is used. For average linkage, the mean of the maximum and the minimum distances from each point is calculated.

The dendrograms obtained can be seen from Figures 12, 13 and 14.

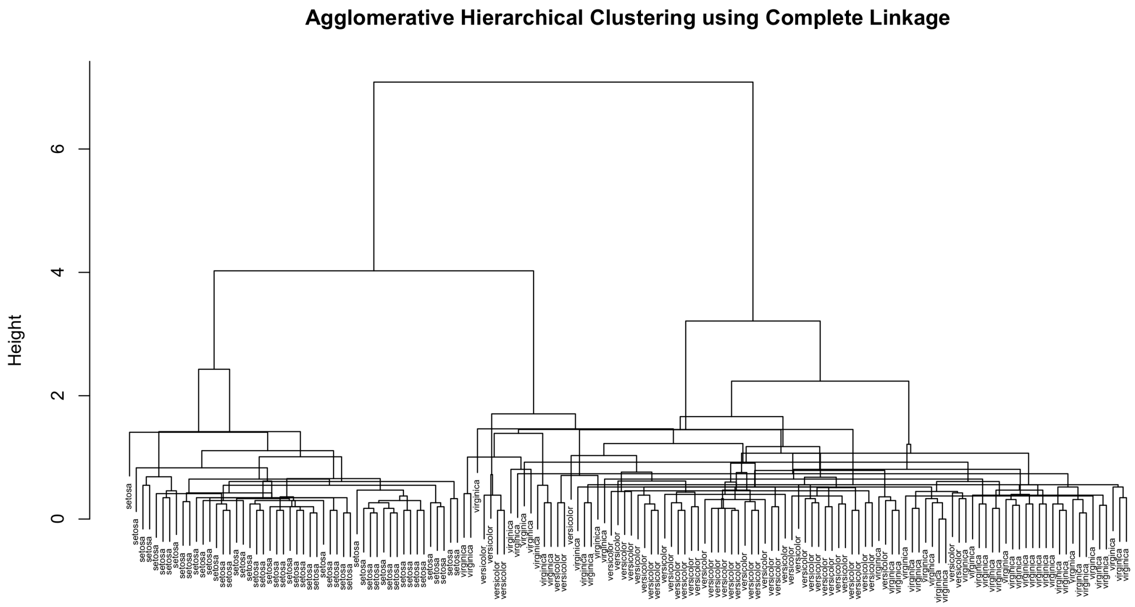


Figure 12:

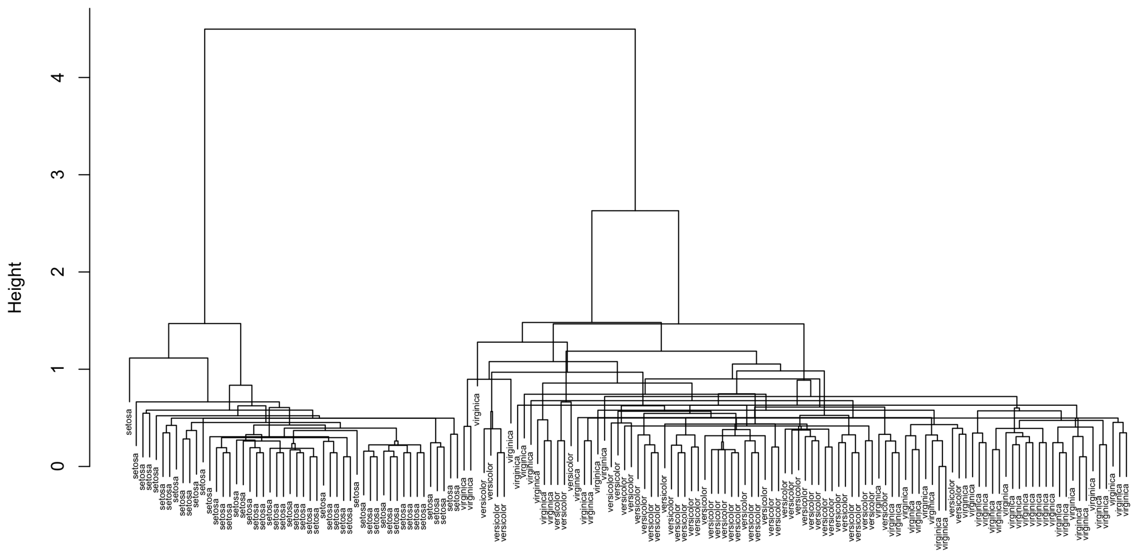

Agglomerative Hierarchical Clustering using Average Linkage

Figure 13:

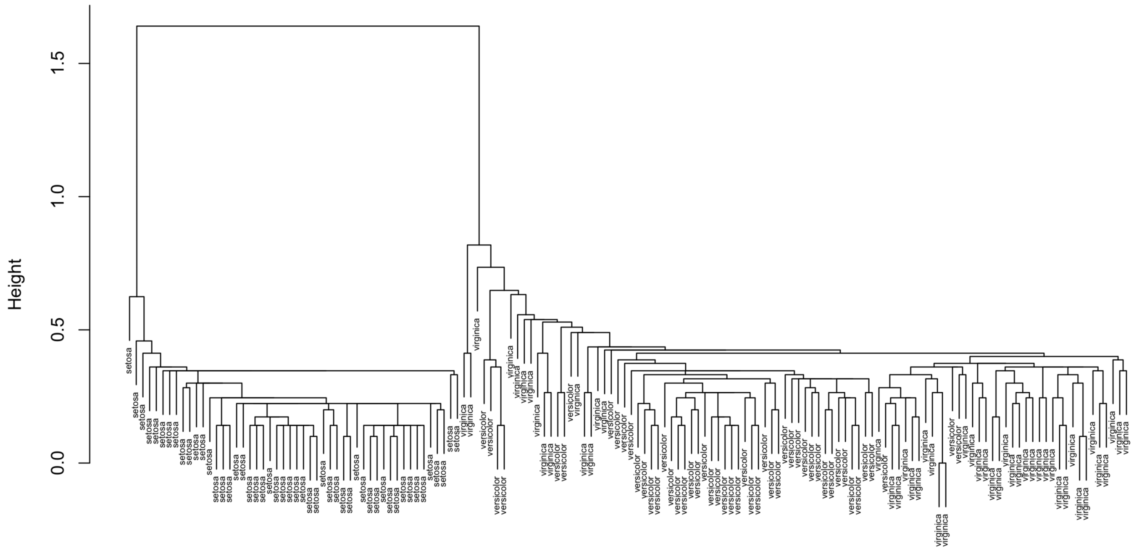

Agglomerative Hierarchical Clustering using Single Linkage

Figure 14: