CDA 532 Homework 4

UB ID: 50413349

UB Name: Sgudemar

Q1)

First, we shall load our data as shown below, and we shall use summary function to obtain statistical overview of our dataset. We use the help() command to get the summary oof each variable's meaning and its correlation.

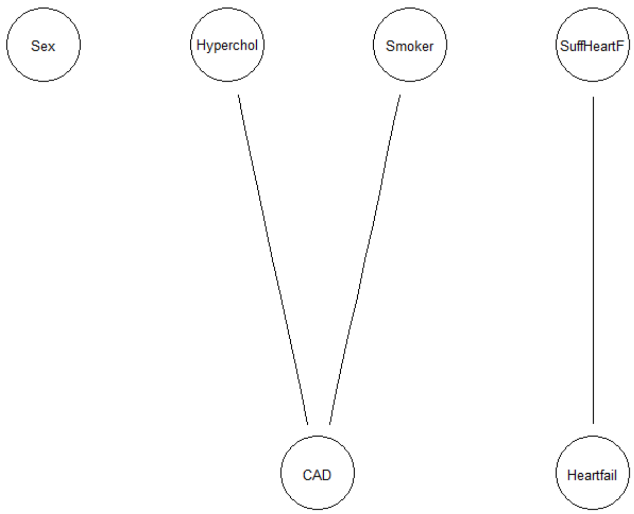
```
> #Q1)
> library(ggm)
> library(gRbase)
> library(bnlearn)
> library(gRain)
> library(dagitty)
> data(cad1)
> summary(cad1)
     Sex           AngPec             AMI         QWave          QWavecode          STcode       STchange   SuffHeartF  Hypertrophi
 Female: 47   Atypical: 30   Definite  : 63   No :153   Nonusable: 13   Nonusable: 79   No :133   No :167    No :172
 Male  :189   None    : 85   NotCertain:173   Yes: 83   Usable   :223   Usable   :157   Yes:103   Yes: 69    Yes: 64
              Typical :121
 Hyperchol Smoker    Inherit    Heartfail  CAD
 No :108   No : 51   No :162   No :177   No :129
 Yes:128   Yes:185   Yes: 74   Yes: 59   Yes:107

> head(cad1)
     Sex   AngPec        AMI QWave QWavecode    STcode STchange SuffHeartF Hypertrophi Hyperchol Smoker Inherit Heartfail
1   Male     None NotCertain    No    Usable    Usable       No         No          No        No     No      No        No
2   Male Atypical NotCertain    No    Usable    Usable       No         No          No        No     No      No        No
3 Female     None   Definite    No    Usable    Usable       No         No          No        No     No      No        No
4   Male     None NotCertain    No    Usable Nonusable       No         No          No        No     No      No        No
5   Male     None NotCertain    No    Usable Nonusable       No         No          No        No     No      No        No
6   Male     None NotCertain    No    Usable Nonusable       No         No          No        No     No      No        No
  CAD
1  No
2  No
3  No
4  No
5  No
6  No
>
> #To infer knowledge regarding the dataset.
> help(cad1)
```

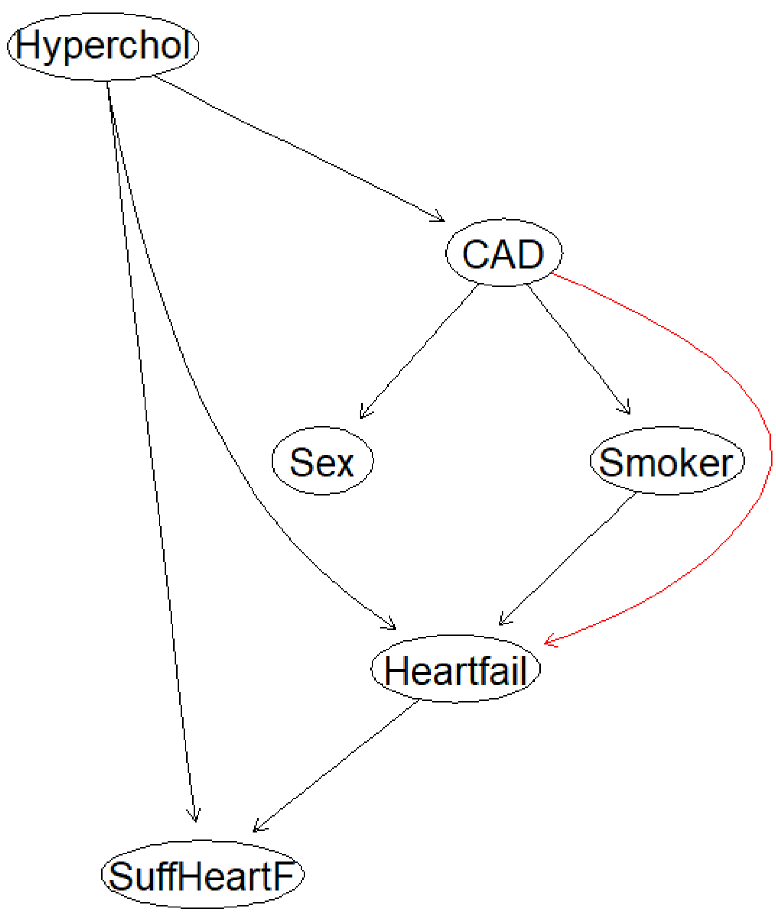
a) Here we perform the two distinct types of structure learning. The first one gives us an undirected graph whereas the second gives us a fully directed graph with the Dag constraints as shown below.

```
###################################################
#a)

#Create a smaller dataset on which we shall perform structural learning
df <- subset(cad1, select = c(Sex,Hyperchol,SuffHeartF, Smoker,Heartfail,CAD))
#use grow Shrink algorithm constrained based Structure learning algorithm
bn_gs_Clrn <- gs(df)
bn_gs_Clrn

  Bayesian network learned via Constraint-based methods

  model:
    [undirected graph]
  nodes:                                 6
  arcs:                                  3
    undirected arcs:                     3
    directed arcs:                       0
  average markov blanket size:           1.00
  average neighbourhood size:            1.00
  average branching factor:              0.00

  learning algorithm:                    Grow-Shrink
  conditional independence test:         Mutual Information (disc.)
  alpha threshold:                       0.05
  tests used in the learning procedure:  68

#hierarchical clustering Score-based methods
bn_hc_Sclrn <- hc(df, score = "aic")
bn_hc_Sclrn

  Bayesian network learned via Score-based methods

  model:
   [Hyperchol][CAD|Hyperchol][Sex|CAD][Smoker|CAD][Heartfail|Hyperchol:Smoker:CAD][SuffHeartF|Hyperchol:Heartfail]
  nodes:                                 6
  arcs:                                  8
    undirected arcs:                     0
    directed arcs:                       8
  average markov blanket size:           3.00
  average neighbourhood size:            2.67
  average branching factor:              1.33

  learning algorithm:                    Hill-Climbing
  score:                                 AIC (disc.)
  penalization coefficient:              1
  tests used in the learning procedure:  65
  optimized:                             TRUE
```

We also compare them side by side.

```
> compare(bn_gs_Clrn, bn_hc_Sclrn
$tp
[1] 0

$fp
[1] 8

$fn
[1] 3
```


Sex
Hyperchol
Smoker
SuffHeartF
CAD
Heartfail


Hyperchol
CAD
Sex
Smoker
Heartfail
SuffHeartF

b) First we fit our Hierarchical Bayesian network model and obtain all the conditional probabilities of the variables as shown below.

```
> #finding out the conditional probability tables (CPTs) at each node
> fittedbn <- bn.fit(bn_hc_Sclrn, data = df)
> fittedbn

  Bayesian network parameters

  Parameters of node Sex (multinomial distribution)

Conditional probability table:

        CAD
Sex              No       Yes
  Female 0.2635659 0.1214953
  Male   0.7364341 0.8785047

  Parameters of node Hyperchol (multinomial distribution)

Conditional probability table:
        No       Yes
0.4576271 0.5423729

  Parameters of node SuffHeartF (multinomial distribution)

Conditional probability table:

, , Heartfail = No

          Hyperchol
SuffHeartF        No       Yes
       No  0.7027027 0.5825243
       Yes 0.2972973 0.4174757

, , Heartfail = Yes

          Hyperchol
SuffHeartF        No       Yes
       No  1.0000000 0.8400000
       Yes 0.0000000 0.1600000


  Parameters of node Smoker (multinomial distribution)

Conditional probability table:

      CAD
Smoker        No       Yes
   No  0.3100775 0.1028037
   Yes 0.6899225 0.8971963

  Parameters of node Heartfail (multinomial distribution)

Conditional probability table:

, , Smoker = No, CAD = No

         Hyperchol
Heartfail         No         Yes
```

Based on the above structure we shall move forward to create a direct acyclic graph.

```
> #Creating a Directed Acyclic Graph
> df_dag <- dag(c("CAD","Hyperchol"),c("SuffHeartF","Hyperchol"),c("Heartfail","Hyperchol"),c("Sex","CAD"),c("Smoker","CAD"), c("H
eartfail","Smoker"),c("SuffHeartF", "Heartfail") )
```

Below we check the d-separation between the nodes. In the first combination we see that neither node are parent or child to one another hence we false whereas the other combinations true as there is direct co relation between each node involved.

```
> #Identifying d-separations in the DAG
> dSep(as(df_dag, "matrix"),first = "Sex", second="SuffHeartF", cond = "Hyperchol")
[1] FALSE
> dSep(as(cad1_dag, "matrix"),first = "Smoker", second="CAD", cond = "Heartfail")
[1] TRUE
> dSep(as(cad1_dag, "matrix"),first = "Sex", second="SuffHeartF", cond = "CAD")
[1] TRUE
>
```

Next, we shall use extractCPT to the DAG structure we created after viewing the Bayesian network and find out the corresponding conditional probabilities. We see that we obtain a similar conditional probability table as above.

```
> #Create CPT
> cpt <- extractCPT(df, df_dag, smooth = 0.5)
> cpt
$CAD
     Hyperchol
CAD          No       Yes
  No  0.7477064 0.3759690
  Yes 0.2522936 0.6240310

$Hyperchol
Hyperchol
       No       Yes
0.4578059 0.5421941

$SuffHeartF
, , Heartfail = No

          Hyperchol
SuffHeartF         No        Yes
       No  0.70000000 0.58173077
       Yes 0.30000000 0.41826923

, , Heartfail = Yes

          Hyperchol
SuffHeartF         No        Yes
       No  0.98571429 0.82692308
       Yes 0.01428571 0.17307692


$Heartfail
, , Smoker = No

         Hyperchol
Heartfail        No       Yes
      No  0.7258065 0.4772727
      Yes 0.2741935 0.5227273

, , Smoker = Yes

         Hyperchol
Heartfail        No       Yes
      No  0.6645570 0.8657407
      Yes 0.3354430 0.1342593


$Sex
        CAD
Sex             No       Yes
  Female 0.2653846 0.1250000
  Male   0.7346154 0.8750000

$Smoker
      CAD
Smoker        No       Yes
   No  0.3115385 0.1064815
```

c) Now we shall use compileCPT to compile the extractedCPT tables and then use querygrain to build a network model accordingly.

```
> #c)
> ## Build the network
> #creating conditional probability tables
> ctable <- compileCPT(cpt)
> ctable
cpt_spec with probabilities:
 P( CAD | Hyperchol )
 P( Hyperchol )
 P( SuffHeartF | Hyperchol Heartfail )
 P( Heartfail | Hyperchol Smoker )
 P( Sex | CAD )
 P( Smoker | CAD )
> grn1 <- grain(ctable)
> querygrain(grn1, nodes=c("CAD", "Heartfail"), type="marginal")
$CAD
CAD
       No       Yes
0.5461526 0.4538474

$Heartfail
Heartfail
       No       Yes
0.7422568 0.2577432

> summary(grn1)
Independence network: Compiled: TRUE Propagated: FALSE
 Nodes : Named chr [1:6] "CAD" "Hyperchol" "SuffHeartF" "Heartfail" "Sex" "Smoker"
 - attr(*, "names")= chr [1:6] "CAD" "Hyperchol" "SuffHeartF" "Heartfail" ...
 Number of cliques:                 4
 Maximal clique size:               3
 Maximal state space in cliques:    8
>
```

c) Now we observe evidence based one female with Hypercholesterolemia.

```
> #likelihood evidence
> evd <- setFinding(grn1, flist=list(c("Sex", "Female"), c("Hyperchol", "yes")))
> getFinding(evd)
  nodes is.hard.evidence hard.state
1   Sex             TRUE     Female
```

Now based on the above new findings we see that the probability distribution of occurrence of heart failure as shown below. We observe that 27& there less probability of heart failure, and 28& of of CAD occuring.

```
> querygrain(evd, nodes=c("CAD", "Heartfail"), type="marginal")
$CAD
CAD
       No       Yes
0.7186962 0.2813038

$Heartfail
Heartfail
       No       Yes
0.7286385 0.2713615

>
```

d) Now we shall create a new dataset with simulated values a shown below.

```
> #d)Creating dataset
> new_df <- simulate(evd, n = 100 , seed = NULL)
> summary(new_df)
  CAD      Hyperchol  SuffHeartF  Heartfail      Sex        Smoker
 No :74    No :52     No :61      No :74     Female:100   No :27
 Yes:26    Yes:48     Yes:39      Yes:26     Male  :  0   Yes:73
> head(new_df)
  CAD Hyperchol SuffHeartF Heartfail    Sex Smoker
1  No        No        Yes        No Female    Yes
2  No        No         No       Yes Female    Yes
3 Yes        No         No        No Female    Yes
4  No       Yes         No       Yes Female     No
5  No        No         No        No Female    Yes
6  No        No         No        No Female    Yes
> typeof(new_df)
[1] "list"
```

Now we shall present this new data set in a tabular form using the table function.

```
> #new data in a table
> table(new_df)
, , SuffHeartF = No, Heartfail = No, Sex = Female, Smoker = No

     Hyperchol
CAD   No Yes
  No   4   4
  Yes  0   0

, , SuffHeartF = Yes, Heartfail = No, Sex = Female, Smoker = No

     Hyperchol
CAD   No Yes
  No   6   2
  Yes  0   2

, , SuffHeartF = No, Heartfail = Yes, Sex = Female, Smoker = No

     Hyperchol
CAD   No Yes
  No   1   5
  Yes  0   1

, , SuffHeartF = Yes, Heartfail = Yes, Sex = Female, Smoker = No

     Hyperchol
CAD   No Yes
  No   0   2
  Yes  0   0

, , SuffHeartF = No, Heartfail = No, Sex = Male, Smoker = No

     Hyperchol
CAD   No Yes
  No   0   0
  Yes  0   0

, , SuffHeartF = Yes, Heartfail = No, Sex = Male, Smoker = No

     Hyperchol
CAD   No Yes
  No   0   0
  Yes  0   0

, , SuffHeartF = No, Heartfail = Yes, Sex = Male, Smoker = No

     Hyperchol
CAD   No Yes
  No   0   0
  Yes  0   0

, , SuffHeartF = Yes, Heartfail = Yes, Sex = Male, Smoker = No

     Hyperchol
CAD   No Yes
  No   0   0
  Yes  0   0
```

Now we shall perform predictions based on the above generated model with respect to smoker and CAD values.

```
> #finding probability of heart-failure and coronary artery disease
> pred <- predict(grn1, response = c("Smoker","CAD"), newdata= new_df,se=TRUE, vcov.=hccm)
> pred
$pred
$pred$Smoker
  [1] "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "No"  "Yes" "Yes"
 [20] "Yes" "Yes" "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"
 [39] "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"
 [58] "Yes" "No"  "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No"  "Yes" "No"  "Yes"
 [77] "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "No"  "Yes" "Yes" "No"  "Yes"
 [96] "Yes" "Yes" "Yes" "Yes" "Yes"

$pred$CAD
  [1] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 [24] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 [47] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 [70] "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No" "No"
 [93] "No" "No" "No" "No" "No" "No" "No" "No"


$pEvidence
  [1] 0.021537642 0.033009473 0.050254499 0.017562186 0.050254499 0.050254499 0.021537642 0.017562186 0.050254499
 [10] 0.050254499 0.043719055 0.050254499 0.031434362 0.003675806 0.050254499 0.043719055 0.017562186 0.050254499
 [19] 0.033009473 0.050254499 0.031434362 0.031434362 0.021537642 0.021537642 0.017562186 0.031434362 0.031434362
 [28] 0.031434362 0.031434362 0.050254499 0.021537642 0.033009473 0.021537642 0.021537642 0.021537642 0.033009473
 [37] 0.031434362 0.043719055 0.050254499 0.021537642 0.043719055 0.043719055 0.050254499 0.021537642 0.033009473
 [46] 0.033009473 0.050254499 0.033009473 0.050254499 0.031434362 0.033009473 0.043719055 0.031434362 0.043719055
 [55] 0.050254499 0.031434362 0.021537642 0.043719055 0.017562186 0.033009473 0.043719055 0.043719055 0.031434362
 [64] 0.033009473 0.031434362 0.021537642 0.050254499 0.031434362 0.031434362 0.033009473 0.033009473 0.021537642
 [73] 0.017562186 0.031434362 0.017562186 0.021537642 0.031434362 0.031434362 0.043719055 0.050254499 0.031434362
 [82] 0.043719055 0.050254499 0.017562186 0.033009473 0.050254499 0.043719055 0.017562186 0.021537642 0.021537642
 [91] 0.003675806 0.050254499 0.033009473 0.003675806 0.050254499 0.043719055 0.050254499 0.043719055 0.043719055
[100] 0.031434362

>
```

Based on the predictions obtained above we shall now find a new Bayesian hierarchical structure learning model and then fit it into bn.fit () function to obtain total conditional probability distribution of the new modified dataset.

```
> bn_newdf <- hc(new_df, score = "aic")
Warning message:
In check.data(x) :
  variable Sex has levels that are not observed in the da
> fbn_newdf <- bn.fit(bn_newdf, data = df)
> fbn_newdf

  Bayesian network parameters

  Parameters of node CAD (multinomial distribution)

Conditional probability table:
        No       Yes
0.5466102 0.4533898

  Parameters of node Hyperchol (multinomial distribution)

Conditional probability table:

         CAD
Hyperchol        No       Yes
      No  0.6279070 0.2523364
      Yes 0.3720930 0.7476636

  Parameters of node SuffHeartF (multinomial distributior

Conditional probability table:

, , Heartfail = No

          Hyperchol
SuffHeartF        No       Yes
       No  0.7027027 0.5825243
       Yes 0.2972973 0.4174757

, , Heartfail = Yes

          Hyperchol
SuffHeartF        No       Yes
       No  1.0000000 0.8400000
       Yes 0.0000000 0.1600000


  Parameters of node Heartfail (multinomial distribution)

Conditional probability table:
   No  Yes
0.75 0.25

  Parameters of node Sex (multinomial distribution)

Conditional probability table:
    Female      Male
0.1991525 0.8008475

  Parameters of node Smoker (multinomial distribution)
```

Q2) First we shall load our titanic dataset as shown below and use the summarize tool to get a statistical overview of our dataset.

```
> library(dplyr)
> library(ggplot2)
> library(rpart)
> library(rpart.plot)
> titanic <- read.csv(file = "titanic.csv", header = TRUE, sep = ",")
> summary(titanic)
    Survived          Pclass           Name               Sex                 Age         Siblings.Spouses.Aboard
 Min.   :0.0000   Min.   :1.000   Length:887         Length:887         Min.   : 0.42   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.25   1st Qu.:0.0000
 Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00   Median :0.0000
 Mean   :0.3856   Mean   :2.306                                         Mean   :29.47   Mean   :0.5254
 3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :3.000                                         Max.   :80.00   Max.   :8.0000
 Parents.Children.Aboard      Fare
 Min.   :0.0000          Min.   :  0.000
 1st Qu.:0.0000          1st Qu.:  7.925
 Median :0.0000          Median : 14.454
 Mean   :0.3833          Mean   : 32.305
 3rd Qu.:0.0000          3rd Qu.: 31.137
 Max.   :6.0000          Max.   :512.329
```

Next we shall represent the data in a tabular format and use the head() function to view the first 5 tuples of the dataframe.

```
> #representing data set in a table format
> tbl_df(titanic)
# A tibble: 887 x 8
   Survived Pclass Name                                               Sex      Age Siblings.Spouses.… Parents.Childre…  Fare
      <int>  <int> <chr>                                              <chr>  <dbl>              <int>            <int> <dbl>
 1        0      3 Mr. Owen Harris Braund                             male      22                  1                0  7.25
 2        1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female    38                  1                0 71.3
 3        1      3 Miss. Laina Heikkinen                              female    26                  0                0  7.92
 4        1      1 Mrs. Jacques Heath (Lily May Peel) Futrelle        female    35                  1                0 53.1
 5        0      3 Mr. William Henry Allen                            male      35                  0                0  8.05
 6        0      3 Mr. James Moran                                    male      27                  0                0  8.46
 7        0      1 Mr. Timothy J McCarthy                             male      54                  0                0 51.9
 8        0      3 Master. Gosta Leonard Palsson                      male       2                  3                1 21.1
 9        1      3 Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson   female    27                  0                2 11.1
10        1      2 Mrs. Nicholas (Adele Achem) Nasser                 female    14                  1                0 30.1
# … with 877 more rows
> head(titanic)
  Survived Pclass                                               Name    Sex Age Siblings.Spouses.Aboard
1        0      3                             Mr. Owen Harris Braund   male  22                       1
2        1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38                       1
3        1      3                              Miss. Laina Heikkinen female  26                       0
4        1      1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35                       1
5        0      3                            Mr. William Henry Allen   male  35                       0
6        0      3                                    Mr. James Moran   male  27                       0
  Parents.Children.Aboard    Fare
1                       0  7.2500
2                       0 71.2833
3                       0  7.9250
4                       0 53.1000
5                       0  8.0500
6                       0  8.4583
> 
```

Next, we shall use the tally() function to get the total number of passengers on board the ship.

```
> #total number of passengers
> totalpassengers <- tally(titanic)
> totalpassengers
    n
1 887
```

Here we observe that there are around 887 passengers on board.

Now we are going to factor in our age column. We now classify it into child and adult

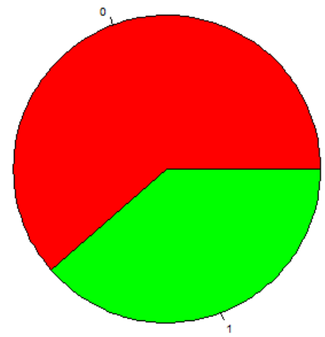
```
> #factoring into child and adult
> ind = which(is.na(titanic$Age))
> titanic[["Age"]]=replace(titanic$Age,c(ind),28)
> titanic[["Age"]]=ifelse(titanic$Age<18,"Child","Adult")
> titanic$Age=factor(titanic$Age,levels = c("Child","Adult"))
> 
> #viewing the changed result
> head(titanic)
  Survived Pclass                                               Name    Sex   Age Siblings.Spouses.Aboard
1        0      3                             Mr. Owen Harris Braund   male Adult                       1
2        1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female Adult                       1
3        1      3                              Miss. Laina Heikkinen female Adult                       0
4        1      1        Mrs. Jacques Heath (Lily May Peel) Futrelle female Adult                       1
5        0      3                            Mr. William Henry Allen   male Adult                       0
6        0      3                                    Mr. James Moran   male Adult                       0
  Parents.Children.Aboard    Fare
1                       0  7.2500
2                       0 71.2833
3                       0  7.9250
4                       0 53.1000
5                       0  8.0500
6                       0  8.4583
> 
```

Now we shall perform our extrapolatory analysis on our dataset. We partition the data set into 2 parts based on the survival. We also predict the percentage of survivors in the data set, and we see that only 39% survived.

```
> ####################################################
> #Analysis of Survived
>
> #Survived passeneger subset
> survivors <-  subset(titanic, Survived == 1)
> head(survivors)
   Survived Pclass                                              Name    Sex   Age Siblings.Spouses.Aboard
2         1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female Adult                       1
3         1      3                             Miss. Laina Heikkinen female Adult                       0
4         1      1       Mrs. Jacques Heath (Lily May Peel) Futrelle female Adult                       1
9         1      3  Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson female Adult                       0
10        1      2                Mrs. Nicholas (Adele Achem) Nasser female Child                       1
11        1      3                    Miss. Marguerite Rut Sandstrom female Child                       1
   Parents.Children.Aboard    Fare
2                        0 71.2833
3                        0  7.9250
4                        0 53.1000
9                        2 11.1333
10                       0 30.0708
11                       1 16.7000
> #dead passenger subset
> dead <-  subset(titanic, Survived == 0)
> head(dead)
   Survived Pclass                           Name  Sex   Age Siblings.Spouses.Aboard Parents.Children.Aboard    Fare
1         0      3         Mr. Owen Harris Braund male Adult                       1                       0  7.2500
5         0      3       Mr. William Henry Allen male Adult                       0                       0  8.0500
6         0      3               Mr. James Moran male Adult                       0                       0  8.4583
7         0      1        Mr. Timothy J McCarthy male Adult                       0                       0 51.8625
8         0      3 Master. Gosta Leonard Palsson male Child                       3                       1 21.0750
13        0      3 Mr. William Henry Saundercock male Adult                       0                       0  8.0500
> total_survivors<- tally(survivors)
> surv_perc <- round((total_survivors/totalpassengers)*100)
> surv_perc
   n
1 39
> pie(table(titanic$Survived),main = "PIE CHART FOR SURVIVAL RATE with 0 representing dead and 1 alive",
+     col = c('red','green'),cex=0.6)
> ####################################################
> 
```

Below is the pie chart distribution of survival of the passengers.

PIE CHART FOR SURVIVAL RATE with 0 representing dead and 1 alive

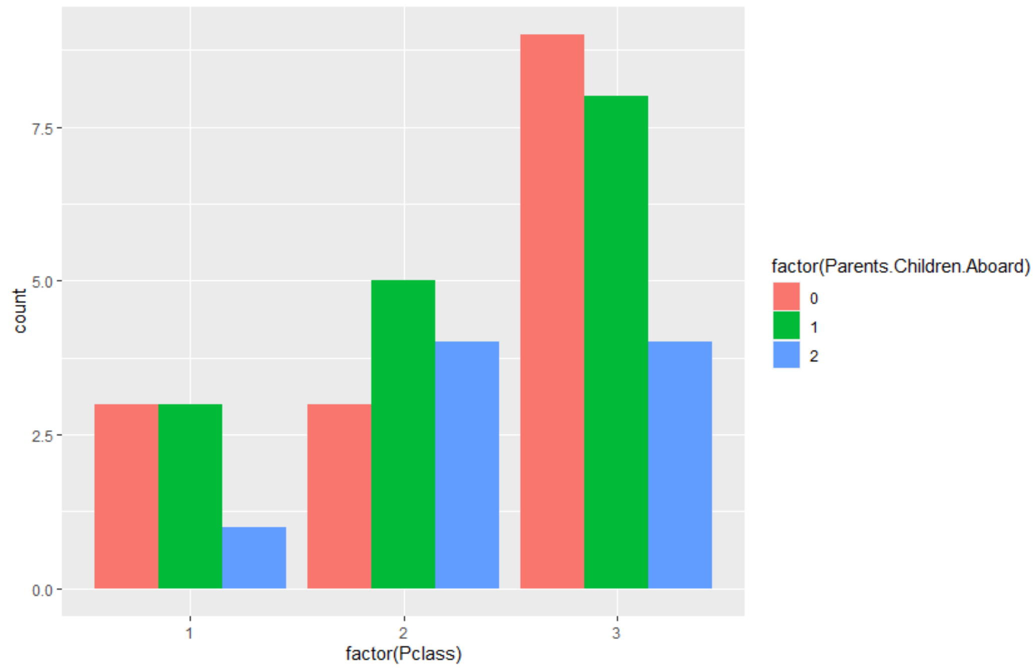


Next, we further classify our dataset of survivors into a partition where only women and children have survived. We see that only 5% of the total passengers survived belong to this class.

```
> #Analysis of  women and children surviving
> women_child_survived <- subset(survivors,  Sex == "female" & Age == "Child")
> head(women_child_survived)
   Survived Pclass                              Name    Sex   Age Siblings.Spouses.Aboard Parents.Children.Aboard
10        1      2      Mrs. Nicholas (Adele Achem) Nasser female Child                       1                       0
11        1      3          Miss. Marguerite Rut Sandstrom female Child                       1                       1
23        1      3                    Miss. Anna McGowan female Child                       0                       0
40        1      3             Miss. Jamila Nicola-Yarred female Child                       1                       0
43        1      2 Miss. Simonne Marie Anne Andree Laroche female Child                       1                       2
58        1      2              Miss. Constance Mirium West female Child                       1                       2
      Fare
10 30.0708
11 16.7000
23  8.0292
40 11.2417
43 41.5792
58 27.7500
> tot_WCS <- tally(women_child_survived)
> WCS_perc <- round((tot_WCS/totalpassengers)*100)
> WCS_perc
  n
1 5
```

Below we view the distribution of the Sex corresponding to demographics.

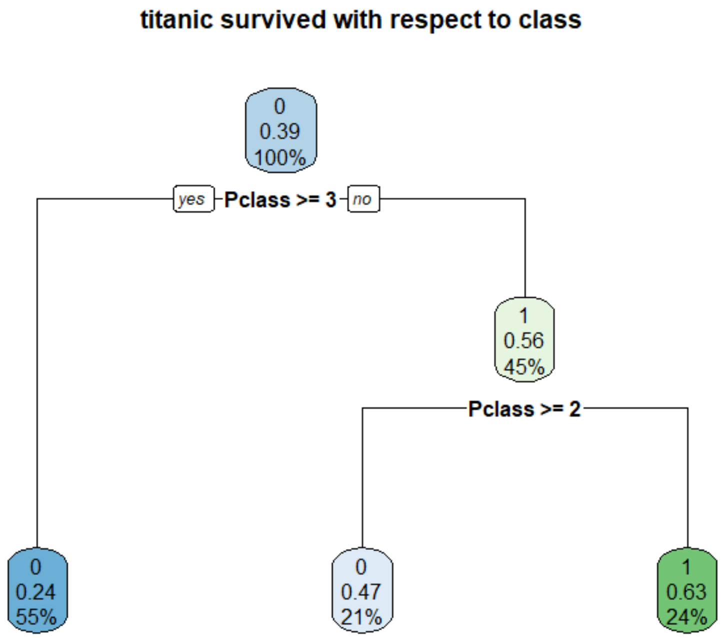
```
R 4.2.0 · C:/Users/sgudemar/Desktop/spring semester/SDM 2/SDM2/
> ################################################
> #distribution of sex with respect to class
> ggplot(titanic,aes(x=factor(Pclass),fill=factor(Sex)))+
+   geom_bar(position="dodge")
> #Graph of mother and child surviving with respect to class
> ggplot(women_child_survived,aes(x=factor(Pclass),fill=factor(Parents.Children.Aboard)))+
+   geom_bar(position="dodge")
> ################################################
> 
```



Now we shall view the probability distribution of survivors with respect to their demographics

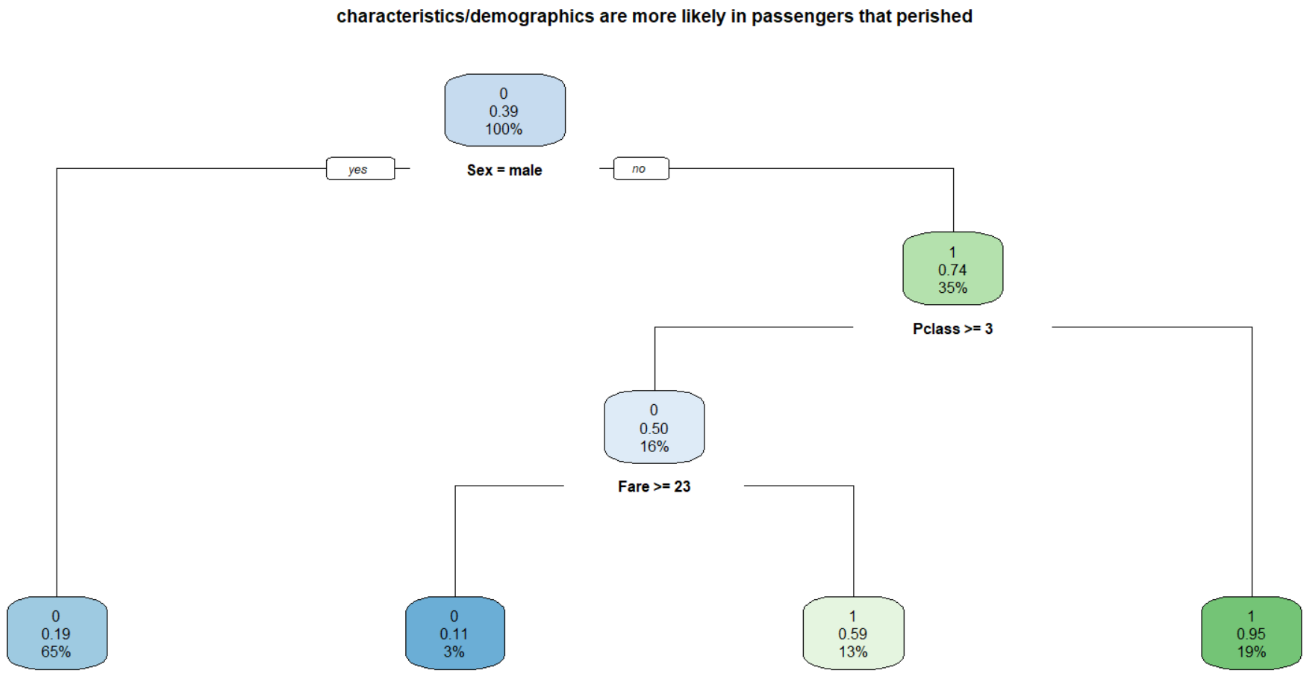
```
tree_1<-rpart(Survived ~ Pclass, data = titanic, method = "class",cp = .02)
rpart.plot(tree_2,main = "titanic survived with respect to class")
```

Here we see that the lowest demographic region has only a survival rate of 24% with respect to total survivors whereas 56% belong to class 1 and 2. In class 1 and 2, we see that 47% have survived, that is only 21% of the total passengers belong to this class. The Survivors of class 1 constitute only 24% of the total survivors.



Below we see the characteristics in passengers that perished. We see that only 19% of the male passenger survived, which is nothing but 65% of the survivors being male. Below are the distributions

```
> tree_2 <- rpart(Survived ~ Pclass + Sex + Age + Fare, data = titanic, method = "class",cp = .02)
> x11()
> rpart.plot(tree_2,main = " characteristics/demographics are more likely in passengers that perished ")
>
```



We finally see the probability of a class 3 male and a class 1 female surviving. We see that 55% of male belonging to class 1 survive whereas only 24% of women in class 3 can survive.

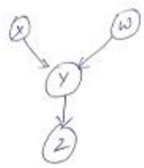
```
#Probability of Rose and Jack Surviving
tree_3 <- rpart(Survived ~ Pclass + Age , data = titanic, method = "class",cp = .02)
rpart.plot(tree_3,main = "titanic survived with respect to class")
```

Q3)

UBNAME : sgudemar
UB PERSON ID : 50413349

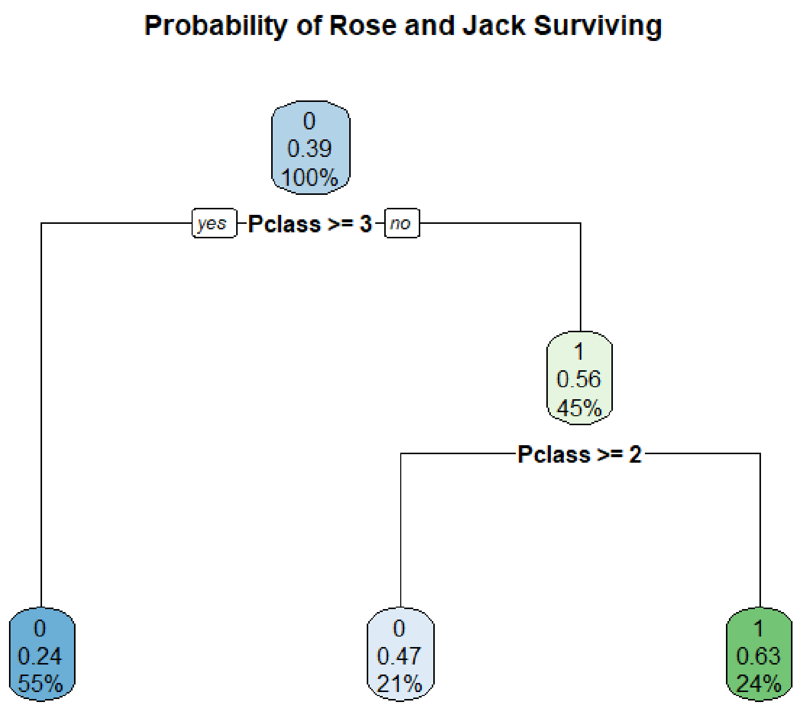
Homework - 4

Q3).
The Bayesian Network can be constructed as shown below:



The above shown networks satisfies all the given local markov assumptions:
W node is independent of node X as the nodes are not connected in the above network, this satisfies the assumption (1).
Node W is dependent on ~~x~~ Z if the node X is observed as there is a dependency flow from the node W to Z via Y, this satisfies the assumption (2)
The node W is dependent on Y, satisfying assumption (4)
The node X is also dependent on Y, this fulfills the assumption (5.)

$$W \perp X \quad — (1)$$
$$W \not\perp Z \mid X \quad — (2)$$
$$Z \perp W \mid Y \quad — (3)$$
$$W \not\perp Y \quad — (4)$$
$$X \not\perp Y \quad — (5)$$
$$W \not\perp X \mid Z \quad — (6)$$
$$X \perp Z \mid W, Y \quad — (7)$$



Q4) First let us load the data set into a data frame variable df. Next, we shall view the data using the head function.

We shall move forward to perform various plots such as corplot which gives us correlation between the variables.

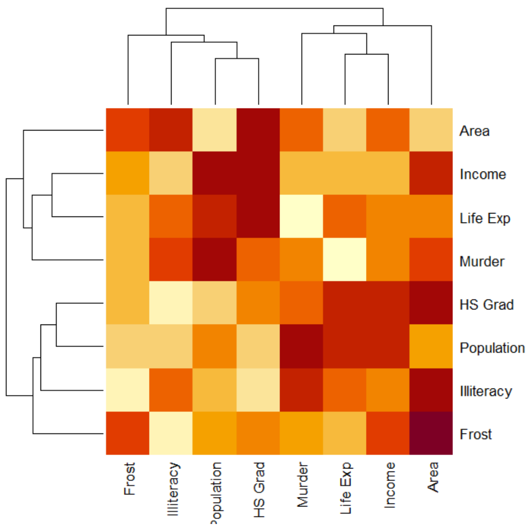
Next, we shall also scale the dataset and perform pca analysis.

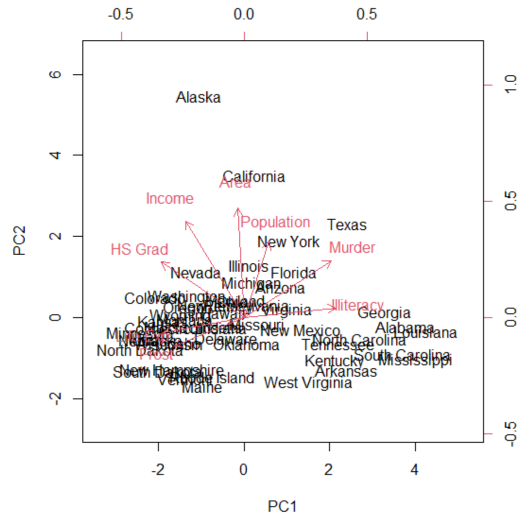
```
> df <- state.x77
> head(df)
           Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
>
> names(df)
NULL
>
> df=(df)
>
>
> M <- cor(df)
> x11()
> corrplot(M)
>
>
> fit.pca <- prcomp(scale(df))
> xlim_1 <- min(fit.pca$x[,1])-1
> xlim_2 <- max(fit.pca$x[,1])+1
> ylim_1 <- min(fit.pca$x[,2])-1
> ylim_2 <- max(fit.pca$x[,2])+1
>
> head(df)
           Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
>
```

Next we shall view the biplots after we fit the PCA analysis.

```
x11()
biplot(fit.pca, choices = c(1,2), scale = 0, xlim = c(xlim_1, xlim_2), ylim = c(ylim_1, ylim_2))

S.body <- cov.wt(state_data, method = "ML")
PC.body <- cov2pcor(S.body$cov)
diag(PC.body) <- 0

x11()
heatmap(PC.body)
```

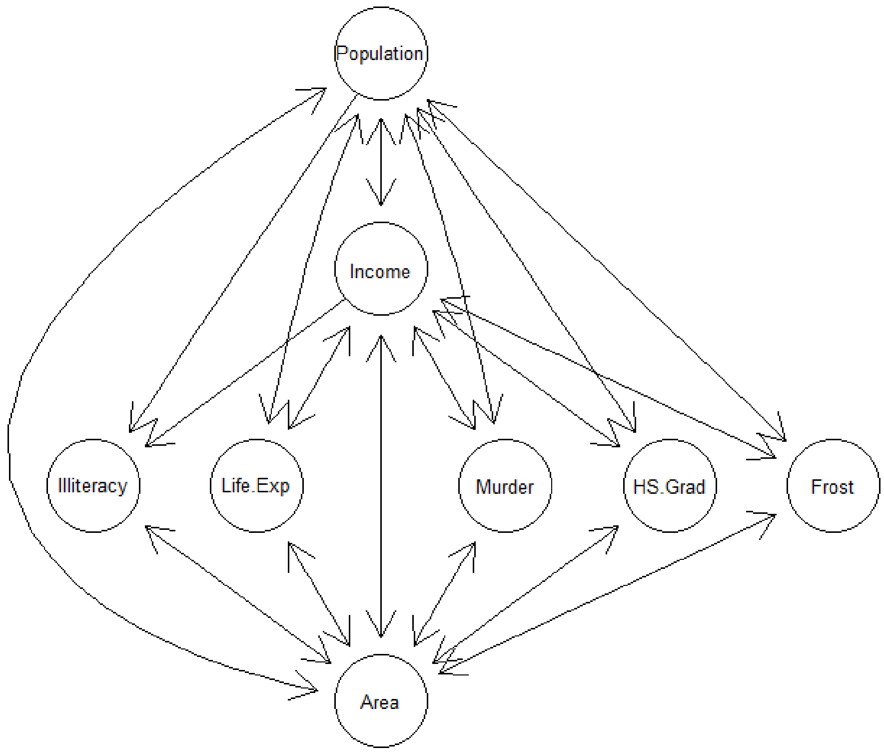
Below we shall see the corresponding heat map





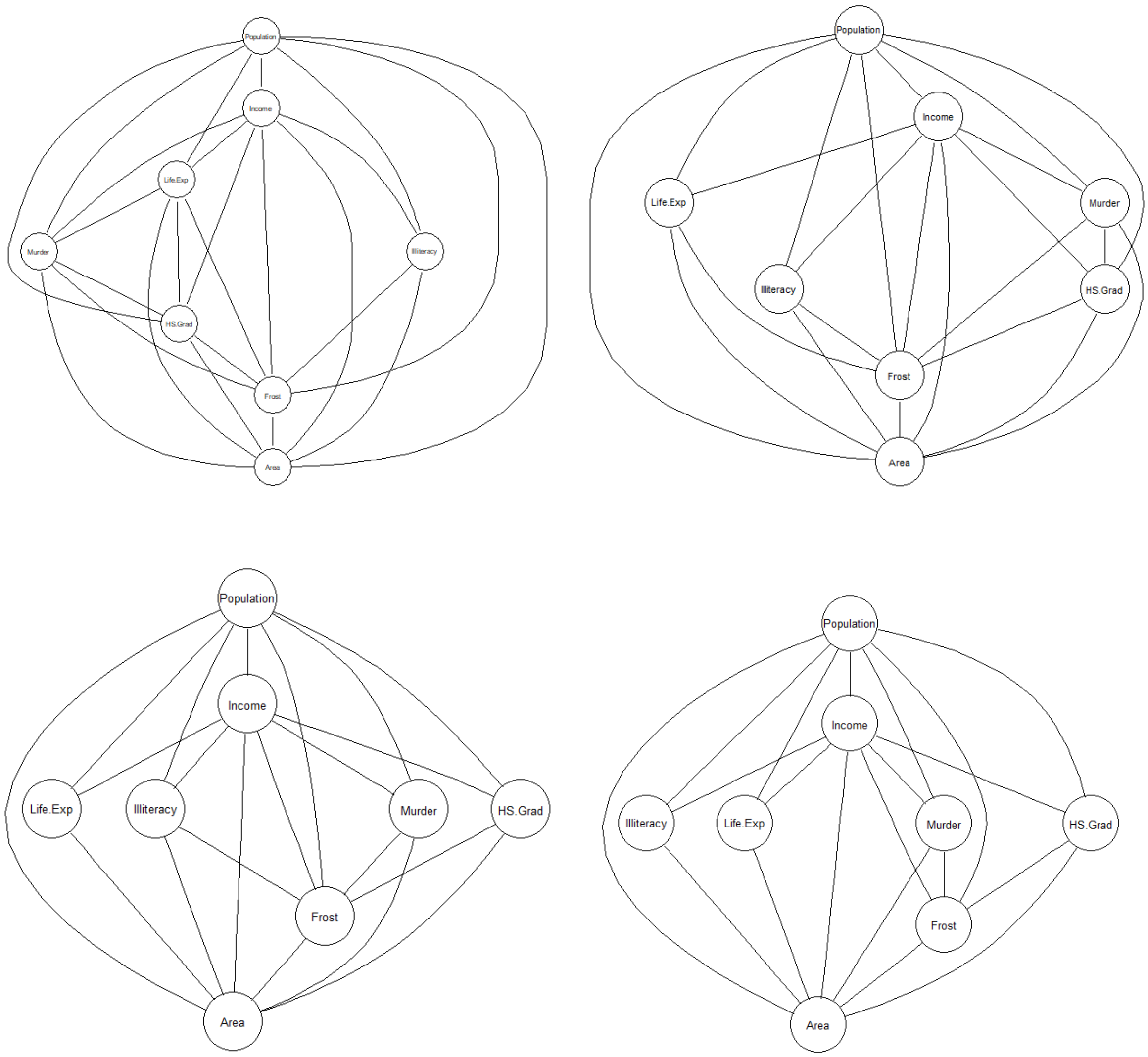
Now we shall use the glasso() function to estimate a sparse inverse covariance matrix and use the names() function to view the names of the prepared covariance matrix's variables. Now we create a DAG as shown below and visualize it using the plot function.

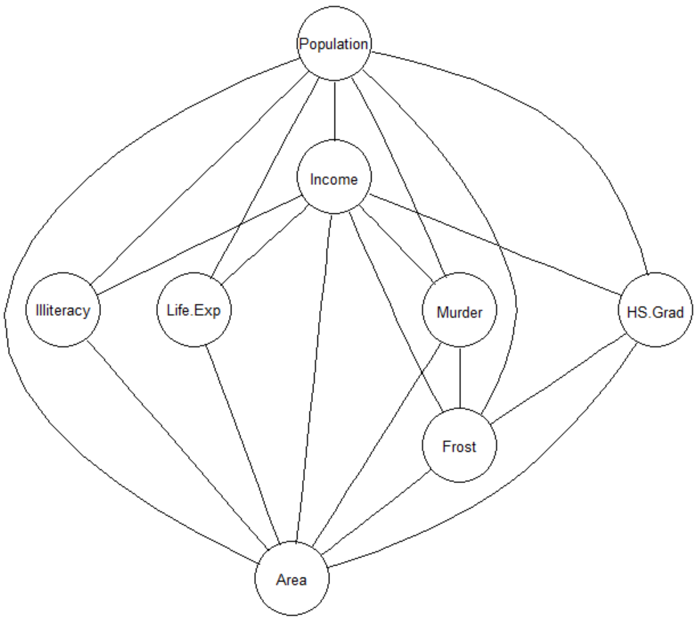
```
> # Estimate a single graph
> S <- S.body$cov
> m0.lasso <- glasso(S, rho = 100)
> names(m0.lasso)
[1] "w"       "wi"       "loglik"  "errflag" "approx"  "del"      "niter"
> my.edges <- m0.lasso$wi != 0
> diag(my.edges) <- 0
> g.lasso <- as(my.edges, "graphNEL")
> nodes(g.lasso) <- names(data.frame(state_data))
>
> x11()
> plot(g.lasso)
>
```


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

```
Error in library(geneplotter) : there is no package
> graphics.off()
> my_rhos <- c(2,5,10,15,25,50)
> m0.lasso <- glassopath(s, rho = my_rhos)
m
[1] 1
m
[1] 2
m
[1] 3
m
[1] 4
m
[1] 5
m
[1] 6
m
[1] 7
m
[1] 8
rho=
[1] 50
rho=
[1] 25
rho=
[1] 15
rho=
[1] 10
rho=
[1] 5
rho=
[1] 2
> for (i in 1:length(my_rhos)){
+   my.edges <- m0.lasso$wi[ , , i] != 0
+   diag(my.edges) <- 0
+   g.lasso <- as(my.edges, "graphNEL")
+   nodes(g.lasso) <- names(data.frame(df))
+
+   x11()
+   plot(g.lasso)
+ }
```

Below we see the combination of all undirected graphs with respect to correlation of each nodes, to view how each node interact among each other respectively based on the list of my_rhos list shown above.


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



Next, we shall load the state data set again as shown below,

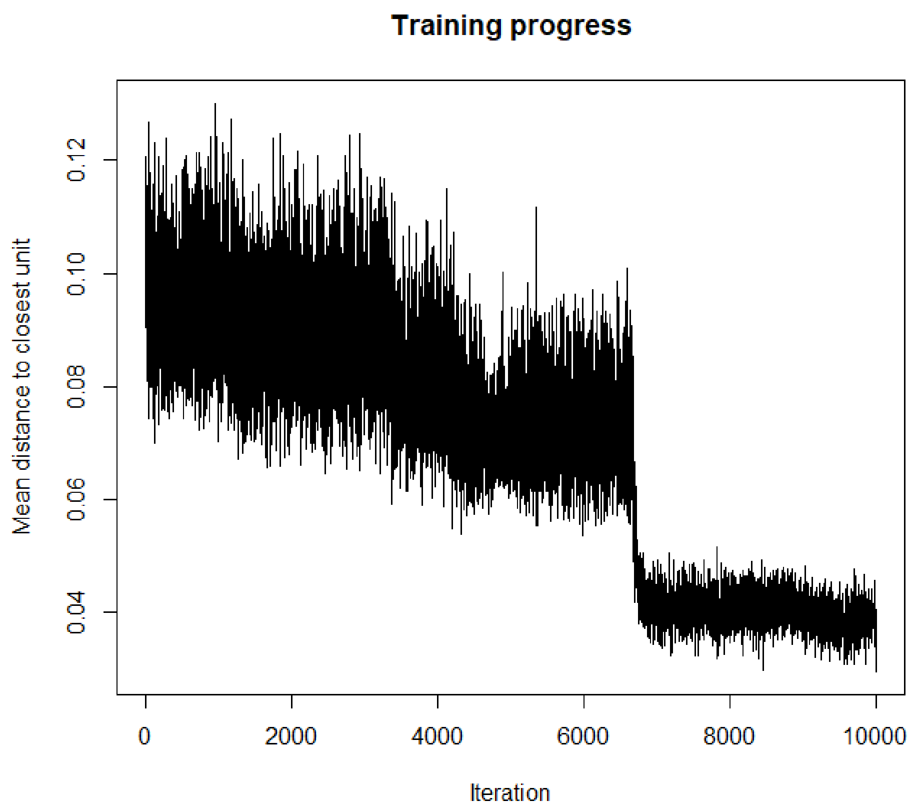
```
> data(state)
> head(state.x77)
           Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
>
> df= state.x77
> df=scale(state.x77)
> |
```

We then load the Kohonen library and create a SOM (Self organizing Maps) using the function somgrid() and plot the grid accordingly.

```
> library(kohonen)
>
> set.seed(100)
> som_grid = somgrid(xdim = 5, ydim=5, topo = "hexagonal")
> df.som = som(df,grid = som_grid,rlen = 10000)
>
> x11()
> plot(df.som)
Warning message:
In par(opar) : argument 1 does not name a graphical parameter
>
> codes = df.som$codes[[1]]
>
```

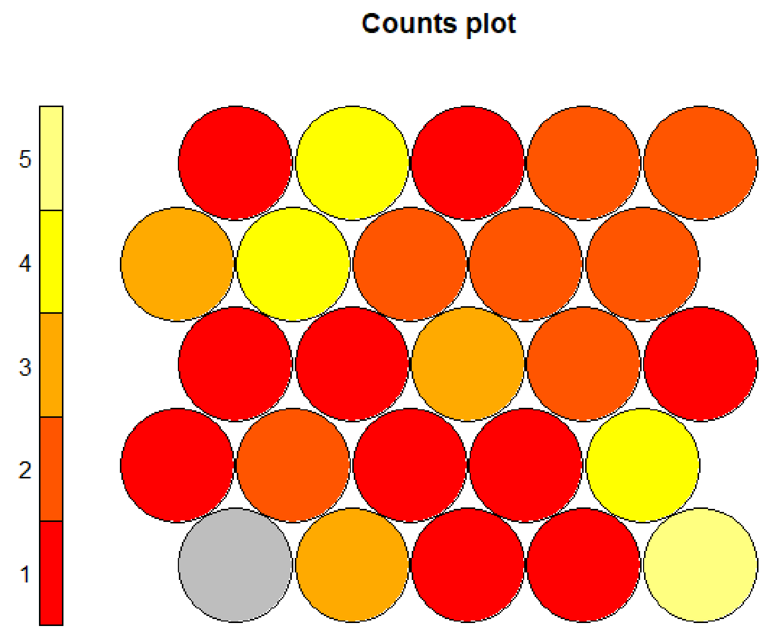
We get the SOM's representation of distribution of the variables present in our dataset.

Below we shall observe the training progress required which represent the changes over the number of iterations.



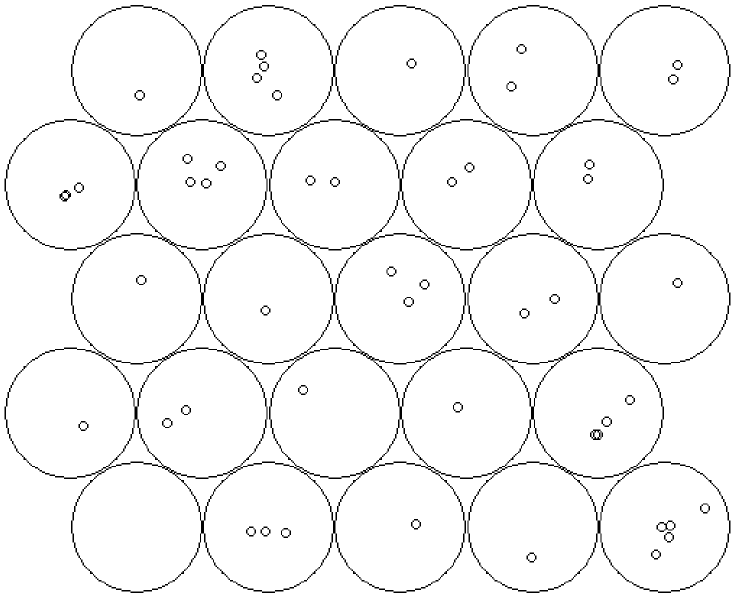
Next we shall find the plot of the Counts in the plot.

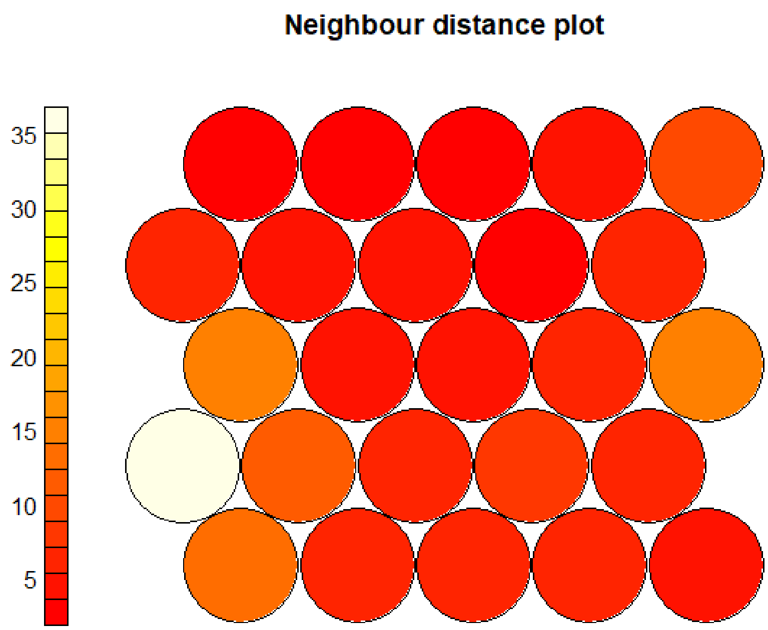
```
x11()
plot(df.som,type = "count")
```




```
x11()
plot(df.som,type = "mapping")
x11()
plot(df.som,type = "dist.neighbours")
```

Mapping plot

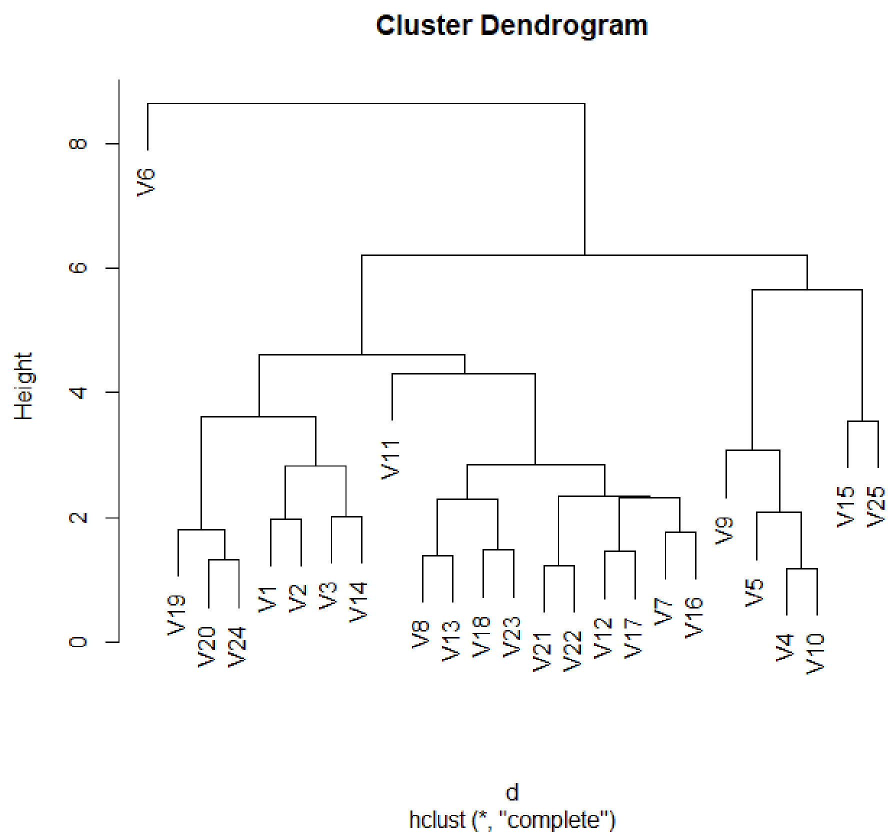




Next, we shall find the distances between each node and form a hierarchical cluster.

```
d = dist(codes)
hq = hclust(d)
x11()
plot(hq)
```

Below we shall plot the dendrogram



Finally, we shall cut the dendrogram tree as shown below with k value of 4.

We shall then plot the SOM after the above operation

```
smc = cutree(hq,k = 4)


my_pal = c("red","blue","green","pink")
my_bhcol = my_pal[smc]

x11()
plot(df.som,type="mapping",col = "black",bgcol = my_bhcol)
add.cluster.boundaries(df.som,smc)

```

We can now obtain distinction between the nodes on how different penalties compliment with the fitted SOM model.

Mapping plot