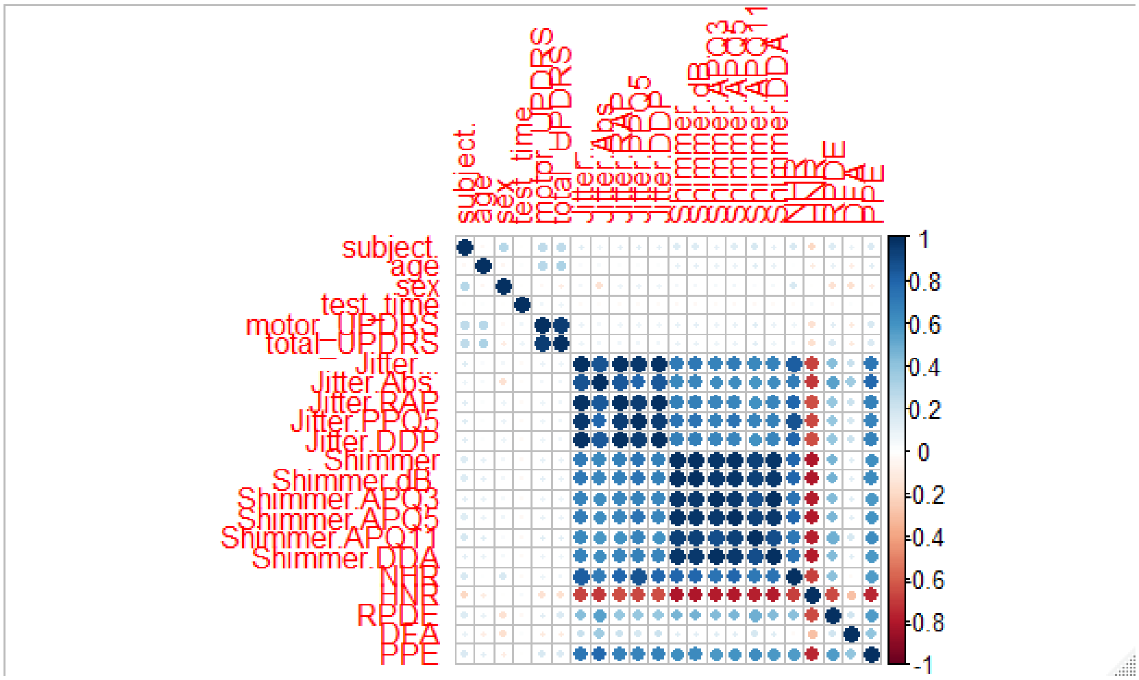
EAS 507: STATISTICAL DATA MINING II

Report for Homework #4

Kishore Ravisankar
UBIT: kravisan

1.

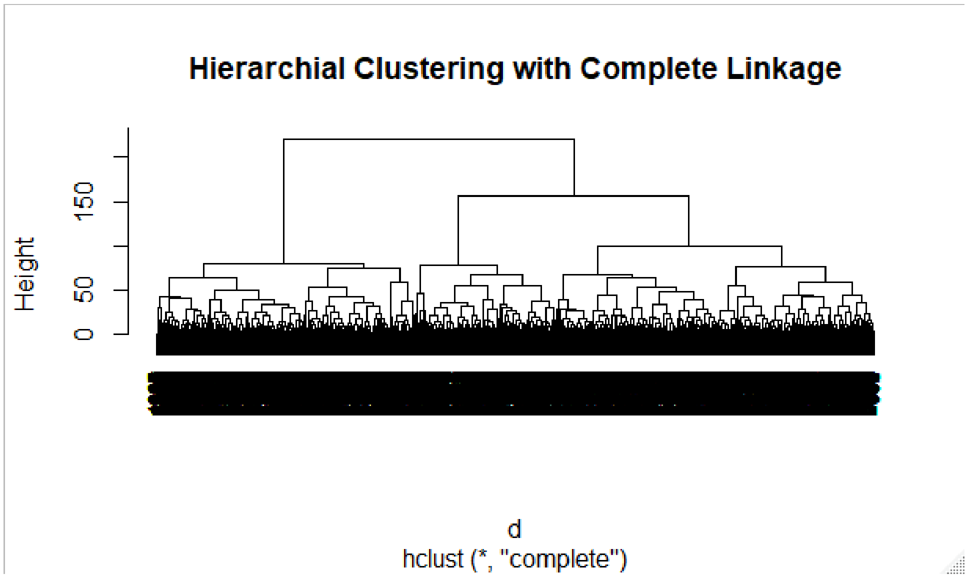
Highly correlated variables were removed by plotting a correlation matrix.



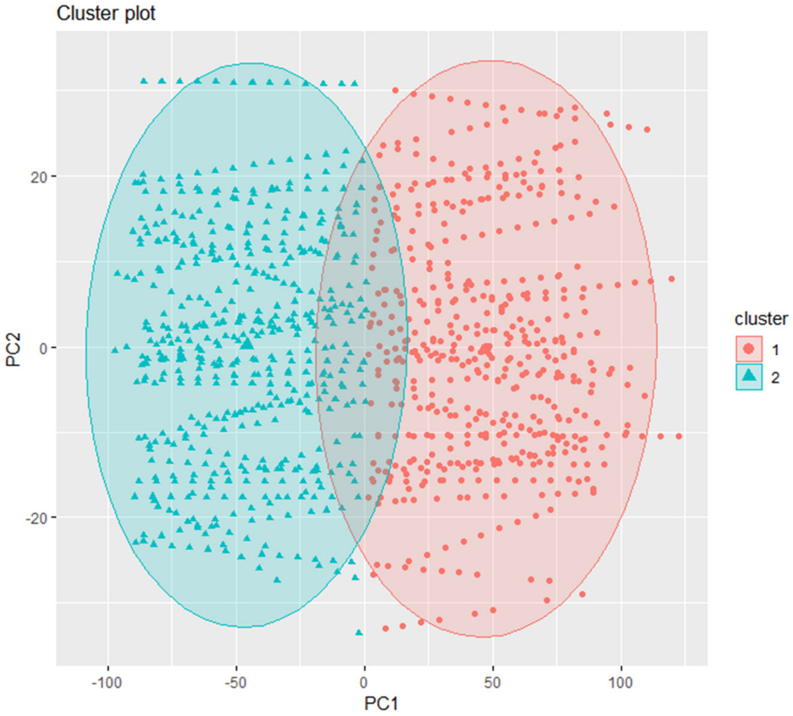
a. The data was clustered using hierarchical clustering and k-means clustering.

The graphs are obtained from the code.

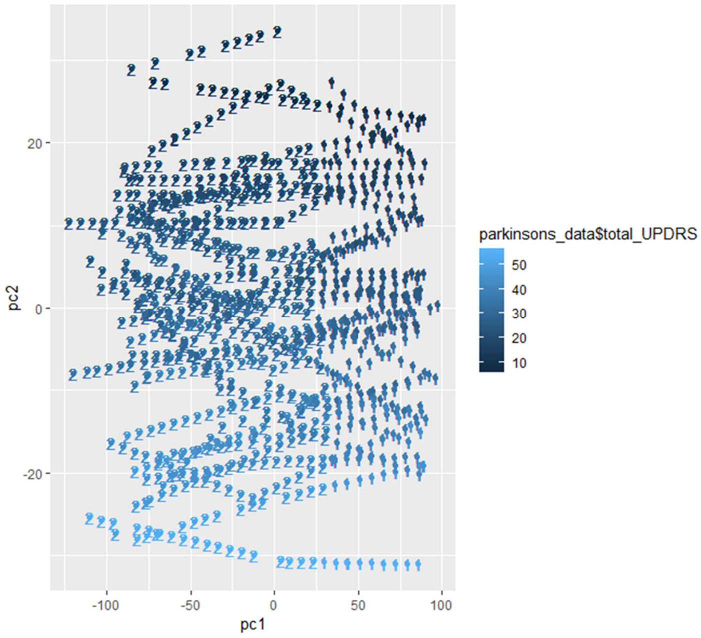
Hierarchical clustering:



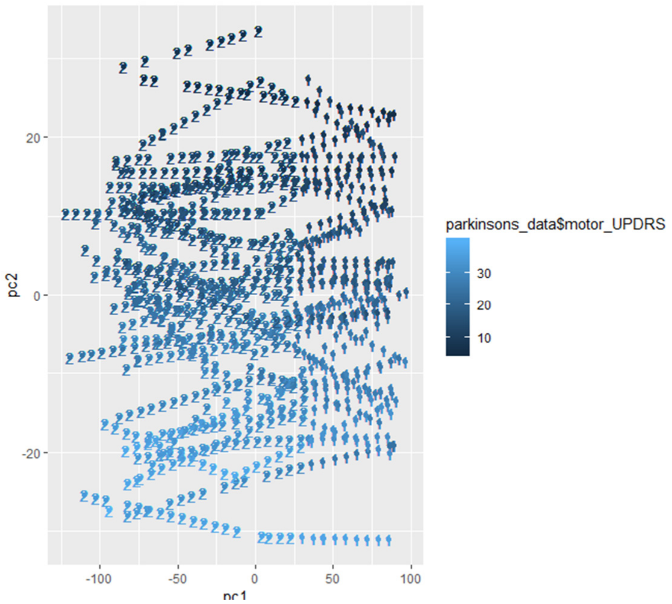
K-means clustering:



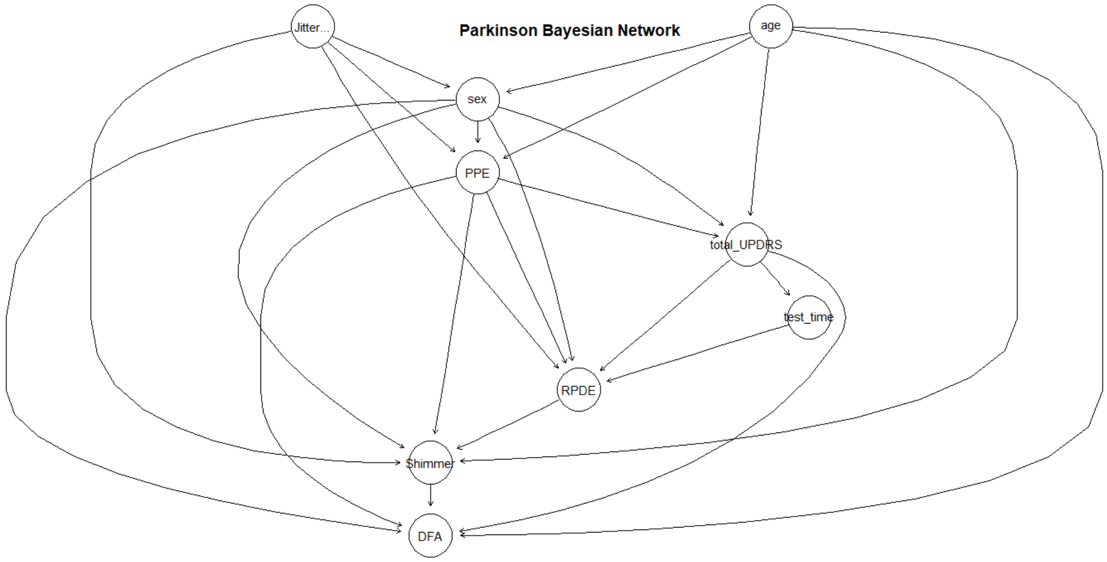
Capturing of total_UPDRS:



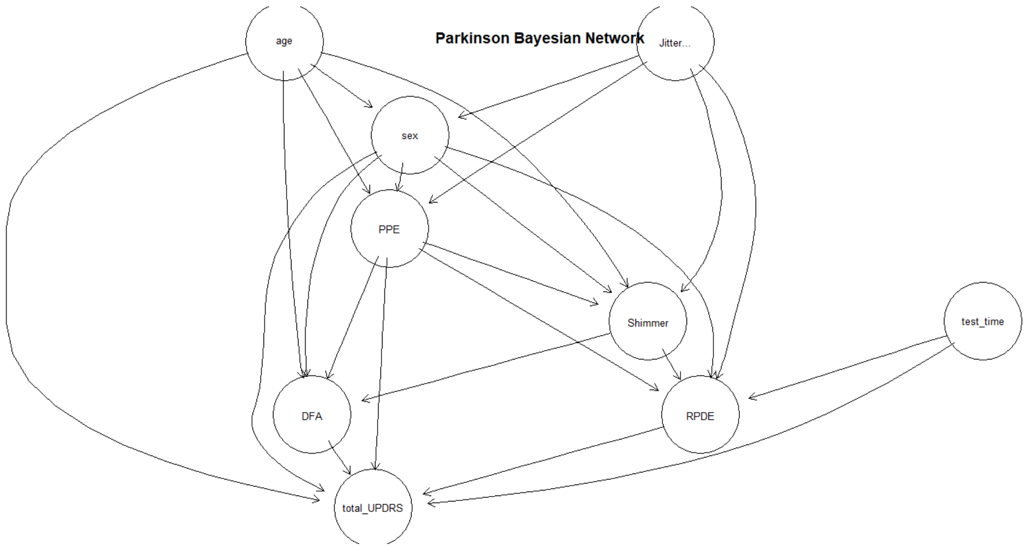
Capturing of motor_UPDRS:



b. A Bayesian network was learnt from the variables, and motor_UPDRS was forced to be at the bottom of the network. The initially learnt network is as follows:



total_UPDRS is forced to be at the bottom of the network. The resulting graph is as follows:



The queries to characterise Jitter variable was made from the generated CPT tables.

Jitter(%) with evidence:

(0,0.006] (0.006,1]

0.4965521 0.5034479

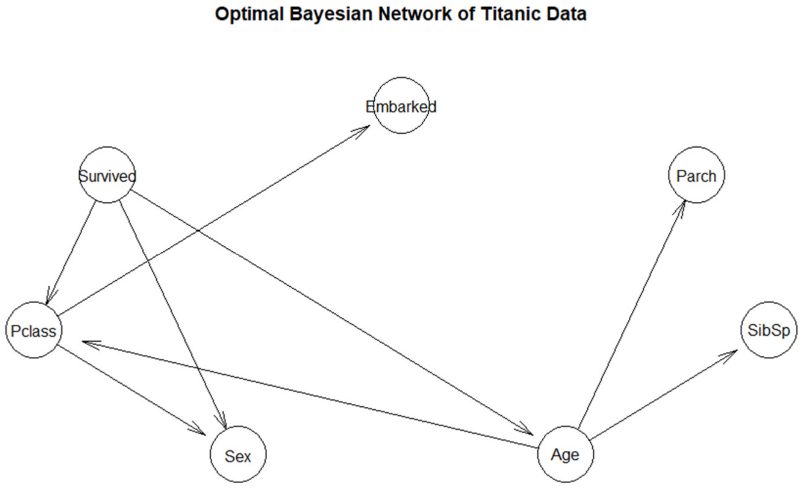
Jitter(%) without evidence:

(0,0.006] (0.006,1]

0.6685671 0.3314329

2.

In this exercise, we use a probabilistic graphical model to analyse the Titanic dataset. Using the hill climbing algorithm in the 'bnlearn' package in R, we learn the variables and create the following network.



The probability of survival of women is 0.7643033 and probability of survival of children is 0.513832. Based on these probabilities, we can say that it has not been very successful in case of children.

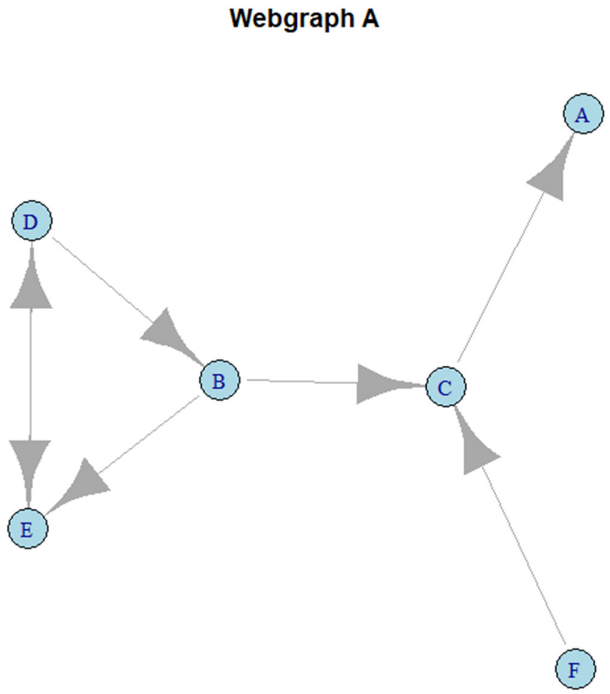
The probability of Rose surviving the disaster is 0.9531881. The probability of Jack not surviving the disaster is 0.8441151.

These results are in line with the movie Titanic.

3.

To compute the PageRank vectors, we first construct the graphs in R using the 'igraph' library in R. After the graphs are constructed, the page.rank function is used to compute the vectors.

a.



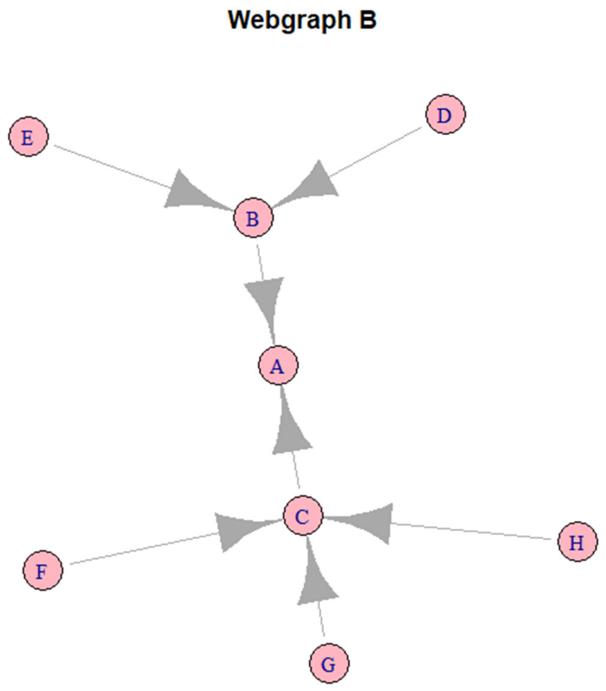
We see that C and E should be equally important, due to the number of incoming arrows.

The PageRank vectors for the various damping constants are tabulated below.

Node	PageRank vector for different damping constant values				
	0.05	0.25	0.50	0.75	0.95
A	0.1683271	0.1786588	0.19227231	0.19399617	0.17305017
B	0.1639395	0.1544288	0.14719411	0.14778661	0.15761096
C	0.1718214	0.1848587	0.18583257	0.17077331	0.14454445
D	0.1681380	0.1758772	0.19135235	0.21832113	0.25658531
E	0.1680380	0.1737324	0.18399264	0.20320659	0.23247617
F	0.1597361	0.1324441	0.09935603	0.06591619	0.03573294

When the damping constant is low, the PageRank values are almost the same. But when it is very high, we can see that D has the highest PageRank. This is not in line with our initial observation.

b.



We see that C is the most influential node, based on the incoming arrows. The PageRank vector for a damping constant value of 0.15 is as follows:

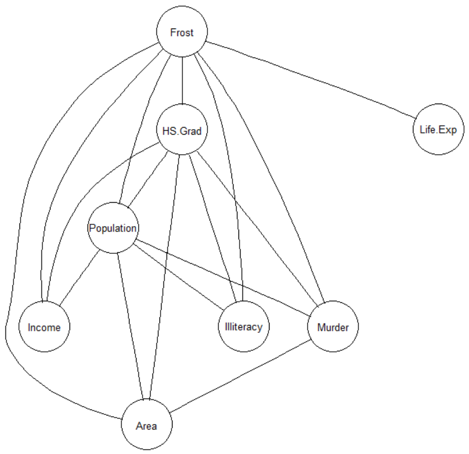
Node	PageRank Vector
A	0.1541610
B	0.1418827
C	0.1582538
D	0.1091405
E	0.1091405
F	0.1091405
G	0.1091405
H	0.1091405

In the above case, C has the highest PageRank value, and is the most important node. This is in line with our initial observation.

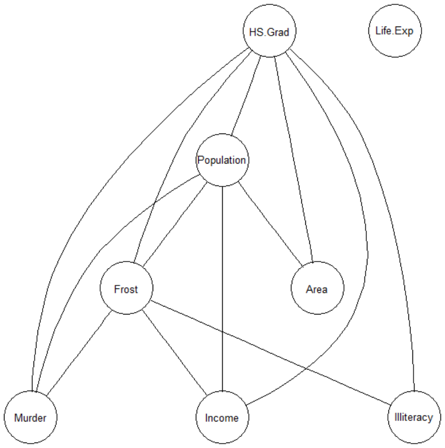
4.

The Gaussian Graphical Models for various rho values are as follows:

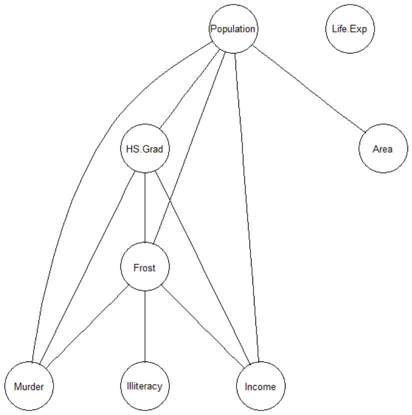
GGM for rho = 1



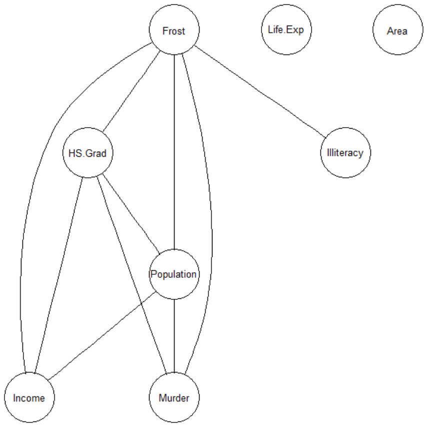
GGM for rho = 2



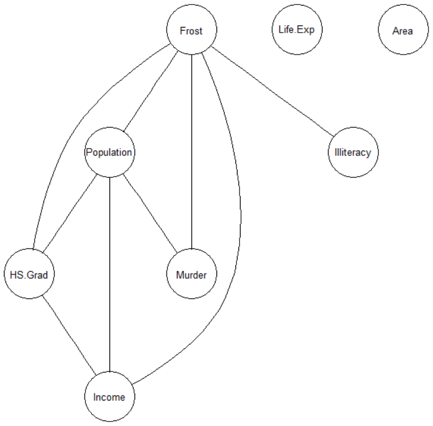
GGM for rho = 4



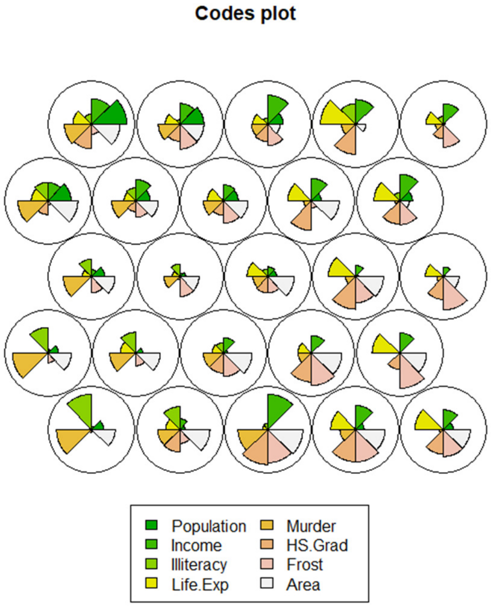
GGM for rho = 8



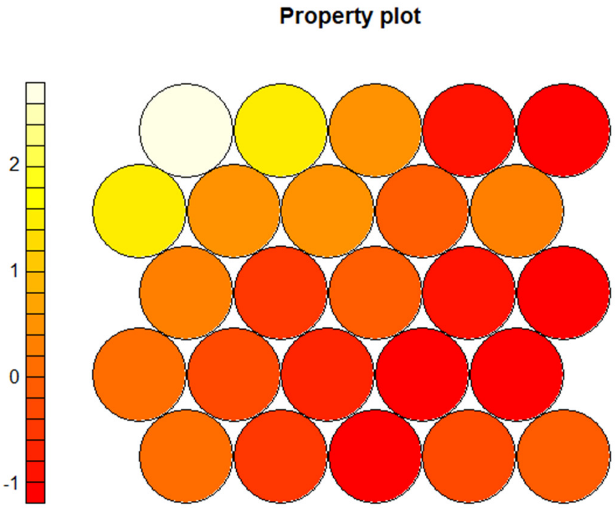
GGM for rho = 10



The SOM for this census data is as follows.



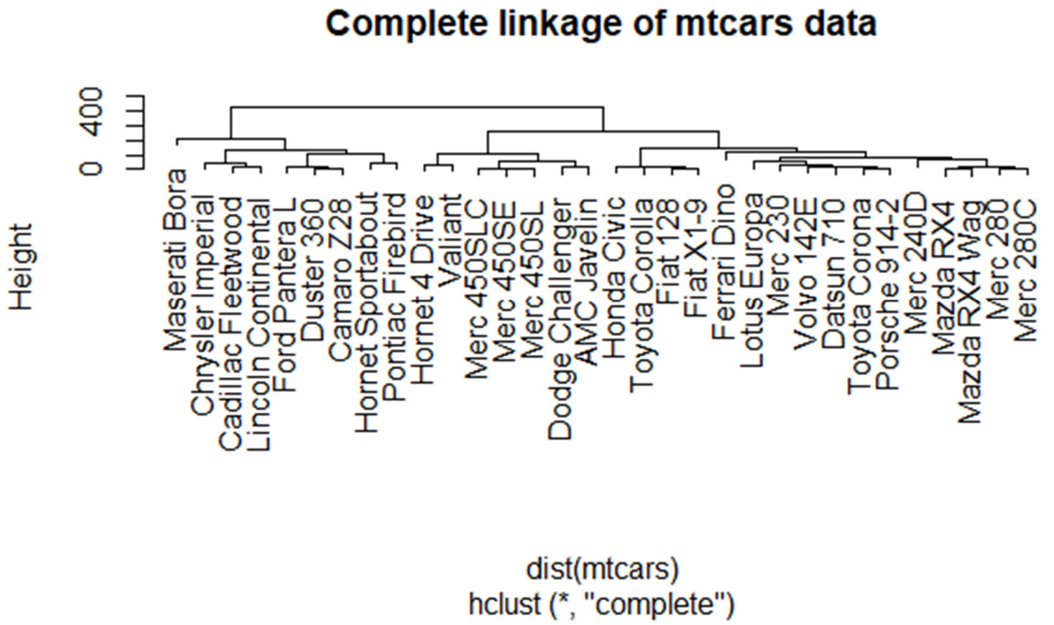
Property plot:

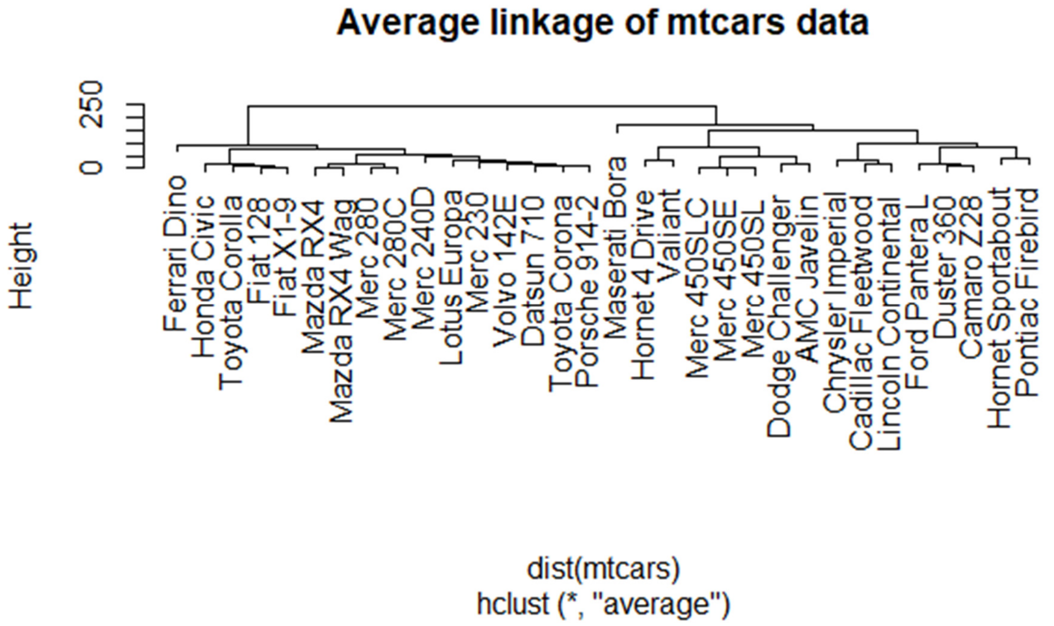


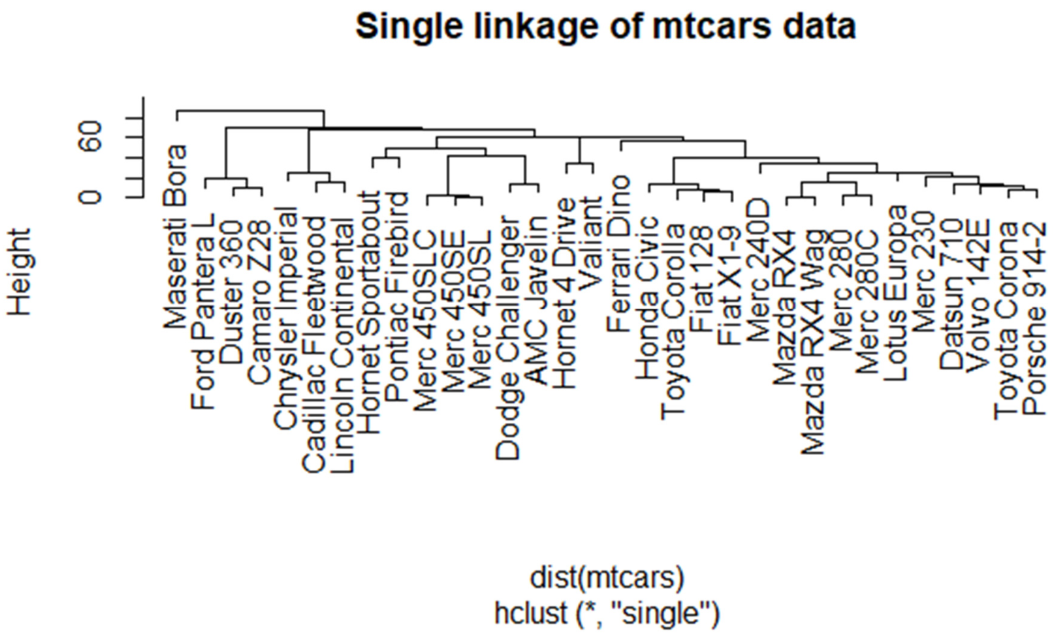
We see that increasing the rho values increases sparsity in the graphical model. We also notice that population is an important feature. We observe the same in the SOM as well.

5.

Custom functions were written in R to perform hierarchical clustering. In this exercise, the mtcars dataset is used. The dendrograms obtained on performing hierarchical clustering are as follows.






Single linkage of mtcars data
Height
0
60
Maserati Bora
Ford Pantera L
Duster 360
Camaro Z28
Chrysler Imperial
Cadillac Fleetwood
Lincoln Continental
Hornet Sportabout
Pontiac Firebird
Merc 450SLC
Merc 450SE
Merc 450SL
Dodge Challenger
AMC Javelin
Hornet 4 Drive
Valiant
Ferrari Dino
Honda Civic
Toyota Corolla
Fiat 128
Fiat X1-9
Merc 240D
Mazda RX4
Mazda RX4 Wag
Merc 280
Merc 280C
Lotus Europa
Merc 230
Datsun 710
Volvo 142E
Toyota Corona
Porsche 914-2
dist(mtcars)
hclust (*, "single")