PROJECT REPORT FOR MACHINE LEARNING (PROJECT 1):

UBitName: manishat
personNumber: 50207628

We were assigned five different tasks as a part of the project and the second task deals with calculating the covariance and correlation pairs for the variables given as a part of the data set.

- There are four variables that are to be taken into the consideration for calculation of covariance and correlation matrices. They are CSScore, Research Overhead, Admin Base Pay and Tuition(out-of-state).
- In order to make plots of the pairwise data, there are 6 different plots that are obtained constituting each variable on each axis.
- Plot 1: CSScore vs ResearchOverhead

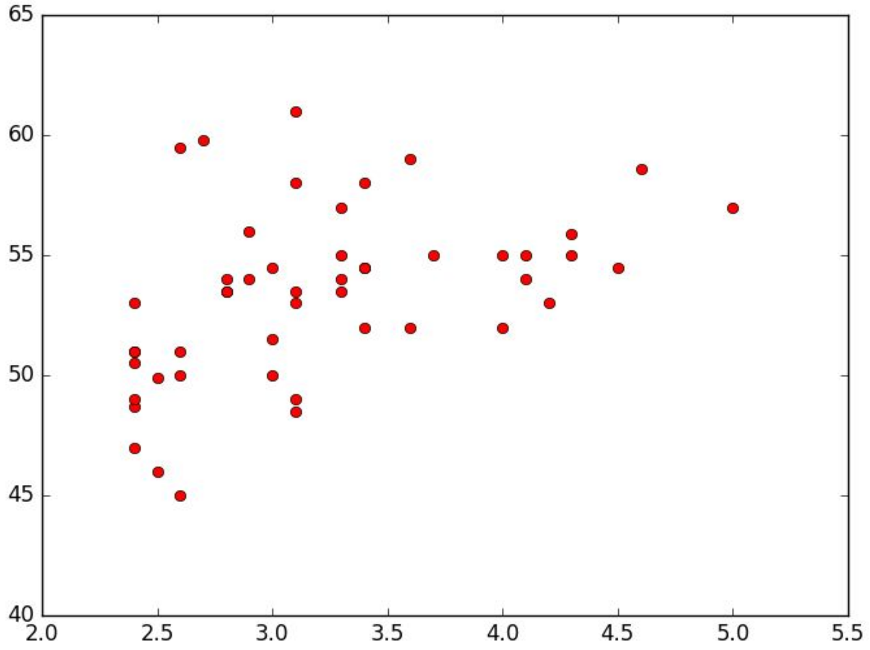
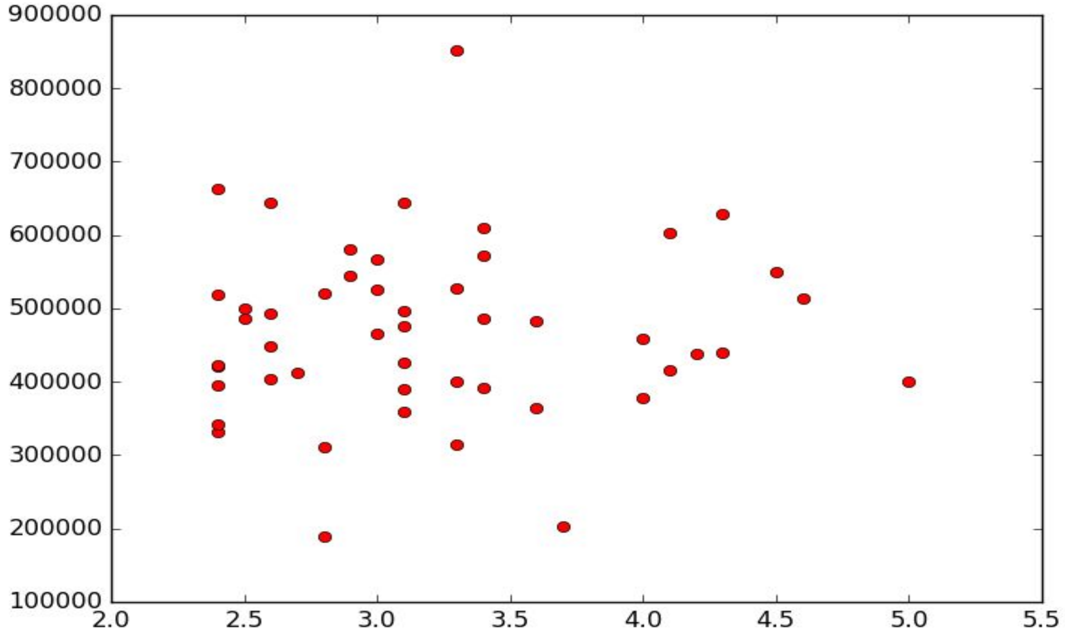


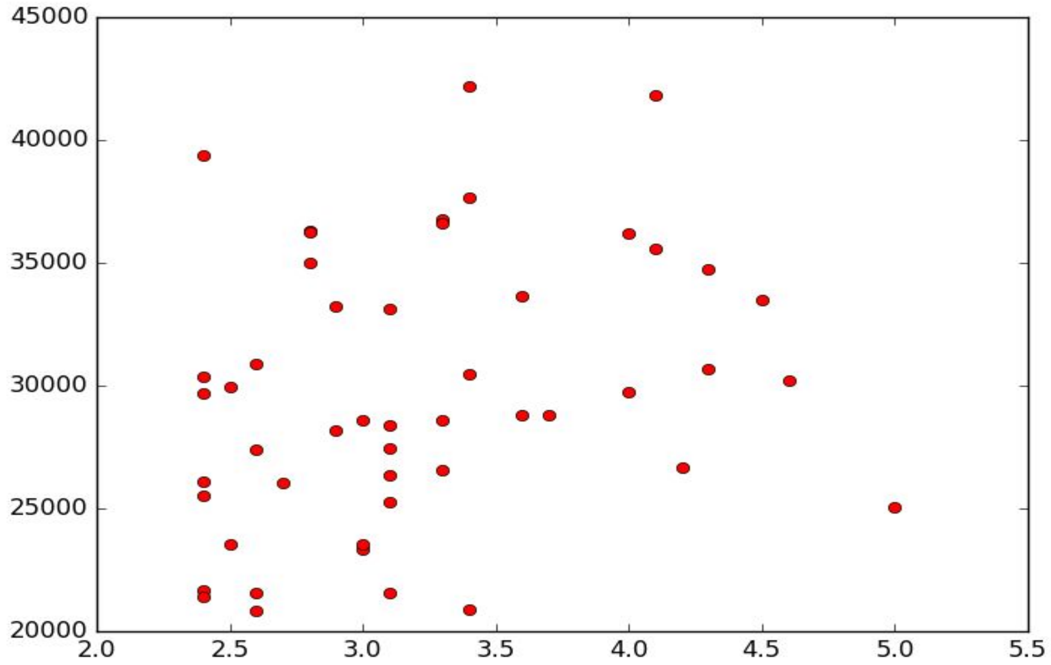
Fig: CSScore(x-axis) vs ResearchOverhead(Y-axis)

- Plot 2: CSScore vs Admin Base Pay



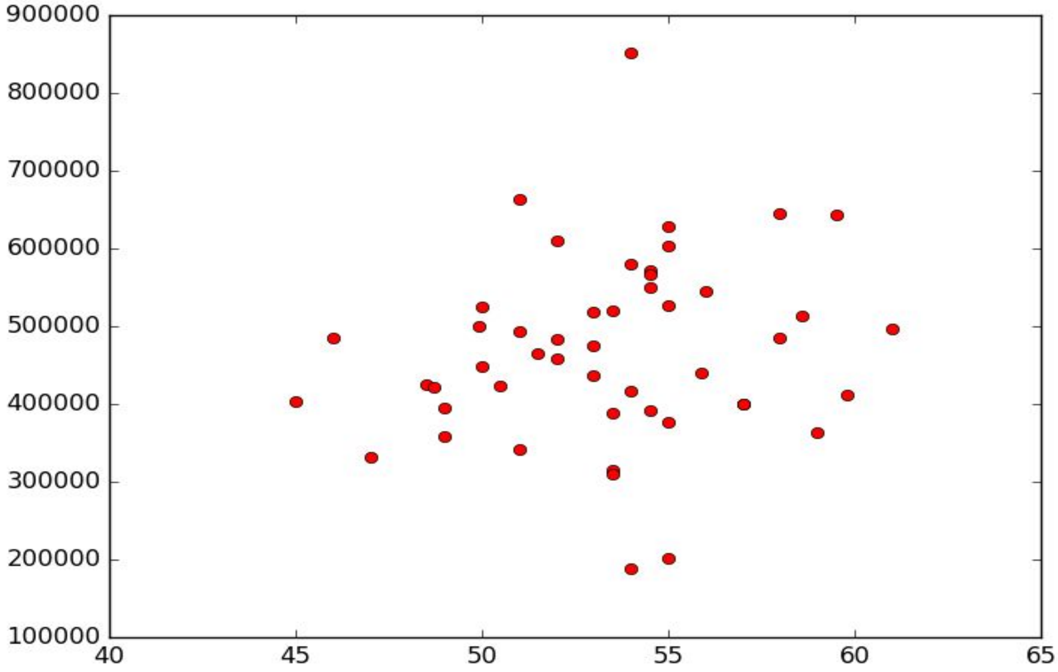
Plot 2 : CSScore (X Axis) vs Admin Base Pay(Y Axis)

- Plot 3: CSScore vs Tuition(out-state)



Plot 3: CSScore vs Tuition(Out-state)

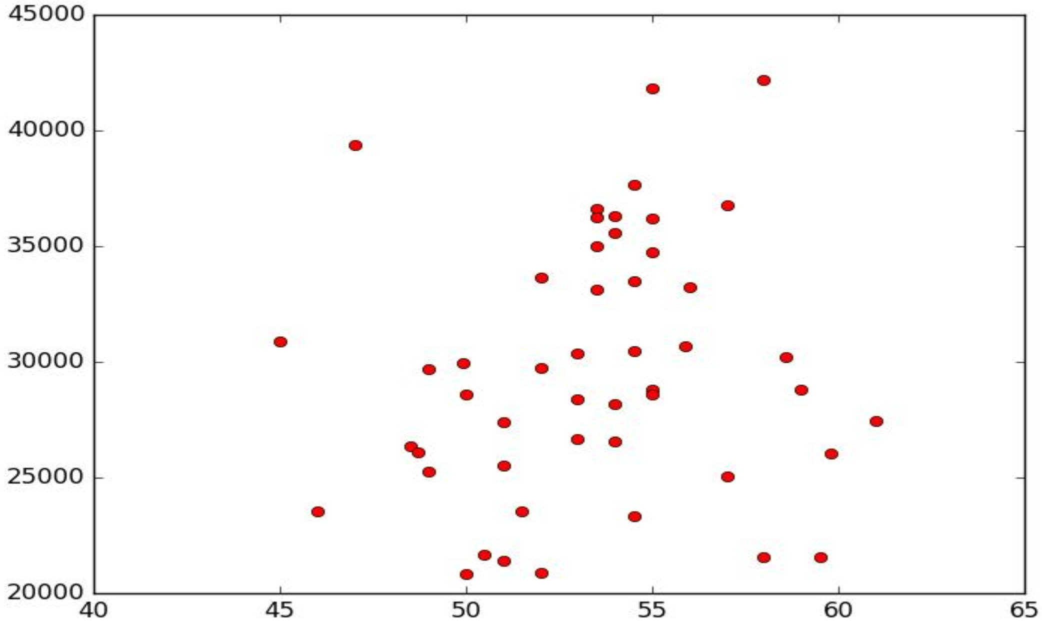
- Plot 4: Research Overhead vs Admin Base Pay:



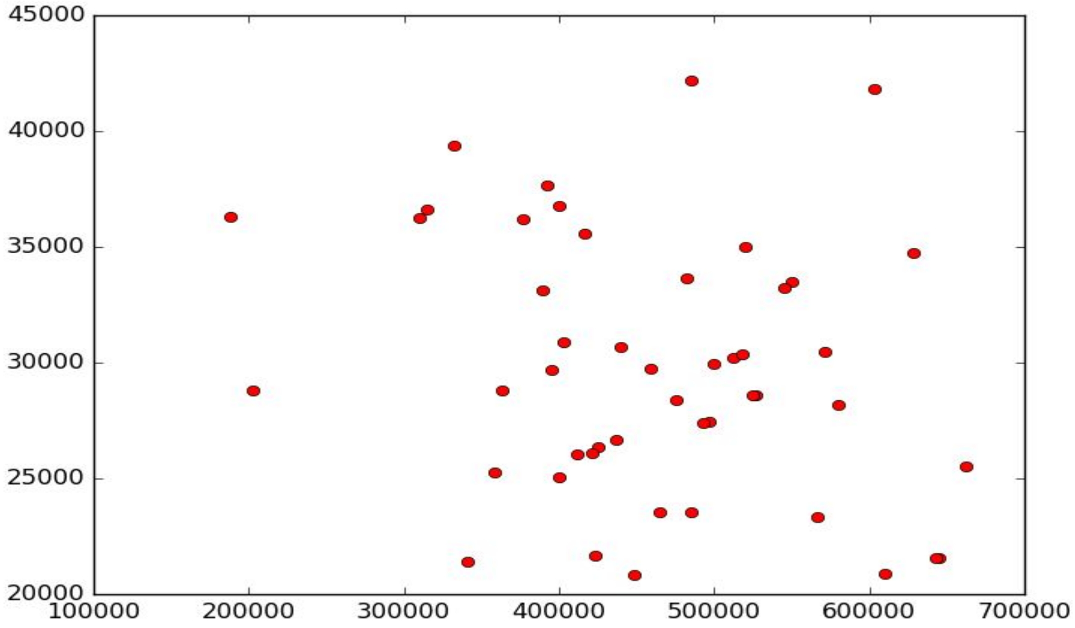
-

Plot 4: Research Overhead vs Admin Base Pay.

- Plot 5: Research Overhead vs Tuition (out-state):



- Plot 6: Admin Base Pay vs Tuition(out-state):



- The correlation matrix is calculated by using the CSScore(X1), Research Overhead(X2), Admin Base Pay(X3) and Tuition(out-state)(X4) as:

```
CorrelationMat =     [[ 1.      0.456  0.048  0.279]
                      [ 0.456  1.      0.165  0.14 ]
                      [ 0.048  0.165  1.     -0.245]
                       [ 0.279  0.14  -0.245  1.    ]]
```

- Here it can be inferred from the matrix that X1 and X2 share the highest correlation and X3 and X4 share the least correlation since CorrelationMat[0][1] has the highest value and CorrelationMat[3][2] has the least value.

TASK4:

- Different dependencies are assumed in between the variables and different probabilities of dependencies are taken to maximise the value of the loglikelihood.

- In task3, we assume the tasks are independent of each other and the computed loglikelihood amounted to -1315(approx)

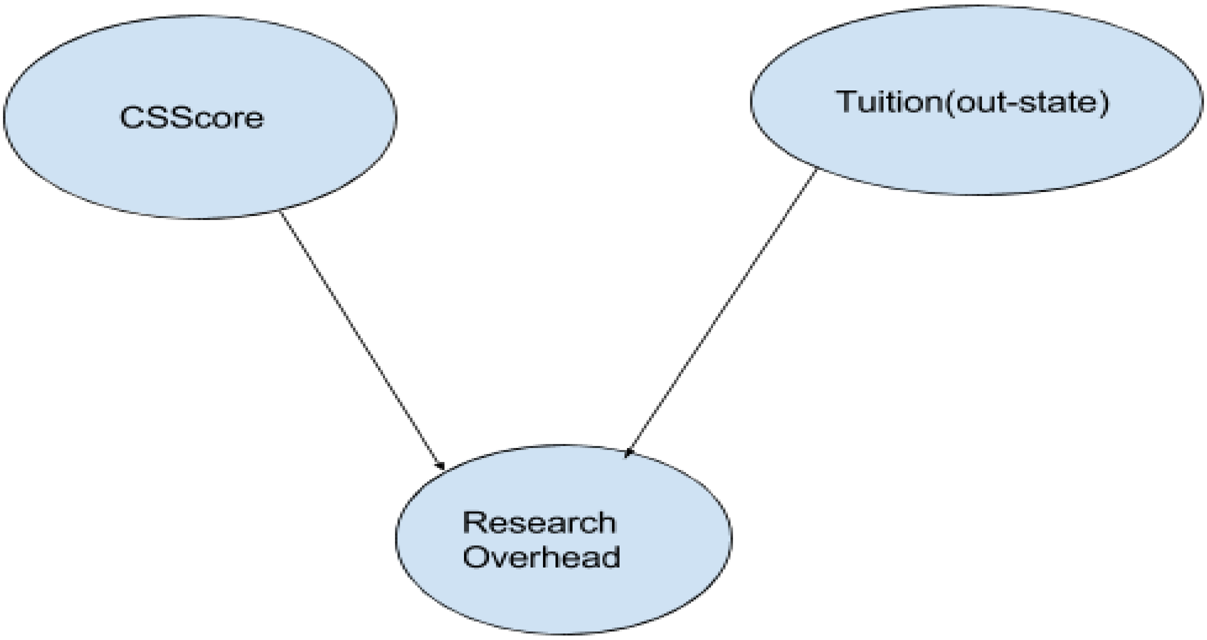
- Here in task4, there has to be atleast one dependency among a set of variables for the graph should be a directed acylic graph

- I assumed different dependencies between variables and the value of the likelihood became maximum when I assumed that X2 was dependent on X1 and X4 (X1 and X4 are the parents of X2).

- The value obtained as a result of this assumption was -1309(approx) which has a greater value than -1315 that was obtained assuming all the variables are independent of each other.

- Other different assumptions made gave me the following results:
- X1 is dependent on X2 and X3 = -1327.9
- X1 is dependent on X3 and X4 = -1326.1
- X1 is dependent on X2 and X4 = -1328.58
- X2 is dependent on X1 and X3 = -1310.73
- X2 is dependent on X3 and X4 = -1313.5
- X3 is dependent on X1 and X2 = -1339.57
- X3 is dependent on X1 and X4 = --1339.49
- X3 is dependent on X2 and X4 = -1340.2
- X4 is dependent on X1 and X2 = -1313.08
- X4 is dependent on X1 and X3 = -1307.03
- X4 is dependent on X2 and X3 = -1305.2
- The obtained BNGraph as a result of my assumption is[[0 1 0 0
- 0 0 0 0
  0 0 0 0
  0 1 0 0]]

Assumption that gave the highest likelihood:


CSScore
Tuition(out-state)
Research Overhead