CSE 574: Introduction to Machine Learning (Fall 2017)

Project 1: Report

Probability Distributions and Bayesian Networks

September 25, 2017

Group members:

1. pbisht2 - 50247429
2. rakeshsi - 50249135

Task 1

Result

mu1 = 3.214

mu2 = 53.386

mu3 = 469178.816

mu4 = 29711.959

var1 = 0.448

var2 = 12.588

var3 = 13900134681.7

var4 = 30727538.733

sigma1 = 0.669

sigma2 = 3.548

sigma3 = 117898.832

sigma4 = 5543.243

Approach

We read the input excel file using pandas and stored the data in a pandas dataFrame. From the dataFrame the columns which are not required were dropped. The remaining dataFrame was converted into a matrix after dropping na values.

Using numpy, we calculated mean, variance and standard deviation from the generated matrix. All the values were rounded up to 3 significant digits during display.

Task 2

Result

covarianceMat =

[[4.57000000e-01 1.10600000e+00 3.87978200e+03 1.05848000e+03]

[1.10600000e+00 1.28500000e+01 7.02793760e+04 2.80578900e+03]

[3.87978200e+03 7.02793760e+04 1.41897208e+10 -1.63685641e+08]

[1.05848000e+03 2.80578900e+03 -1.63685641e+08 3.13676958e+07]]

correlationMat =

[[1. 0.456 0.048 0.279]

[0.456 1. 0.165 0.14]

[0.048 0.165 1. -0.245]

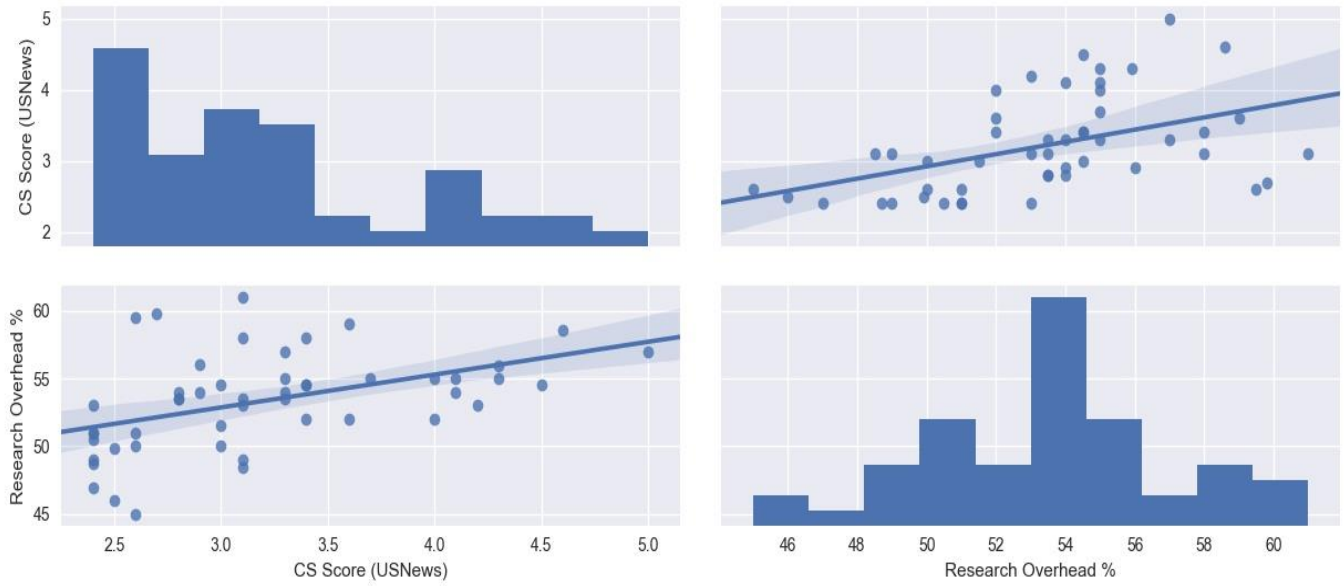
[0.279 0.14 -0.245 1.]]

Approach

We used numpy's cov, corrcoef methods to get the covarianceMat and correlationMat matrices respectively. We used seaborn library's pairplot method to plot pairwise data.

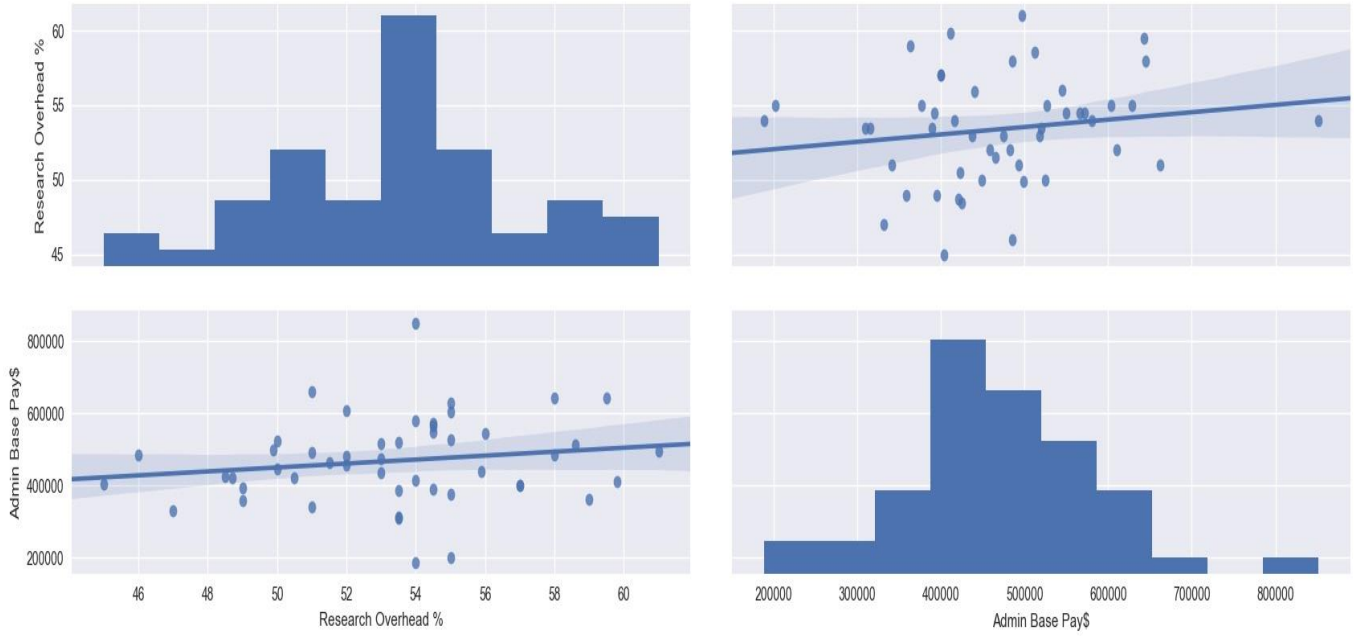
CS Score and Research Overhead are the most correlated variable pair due to highest magnitude of correlation while CS Score and Admin Base pay are least correlated variable pair due to lowest magnitude of correlation.

CS Score and Research Overhead pairwise plot



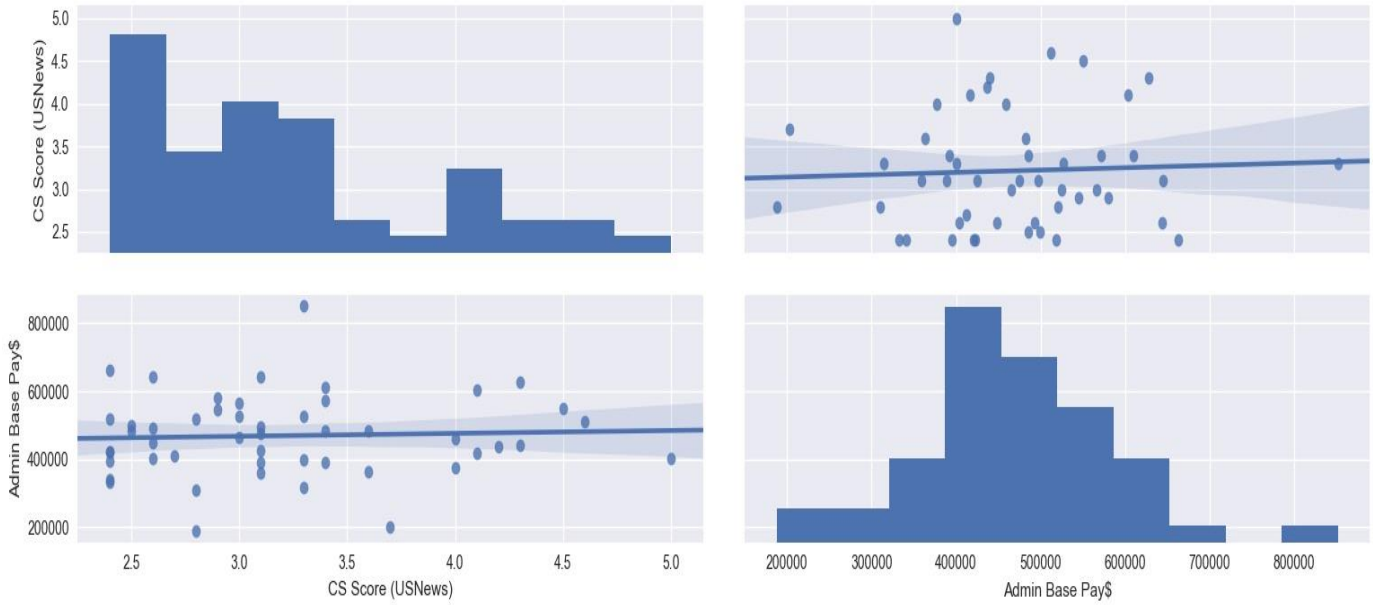
We can analyze that the properties – CS Score and Research Overhead have positive correlation and the linear regression line emphasizes the same.

Research Overhead and Admin Base Pay pairwise plot



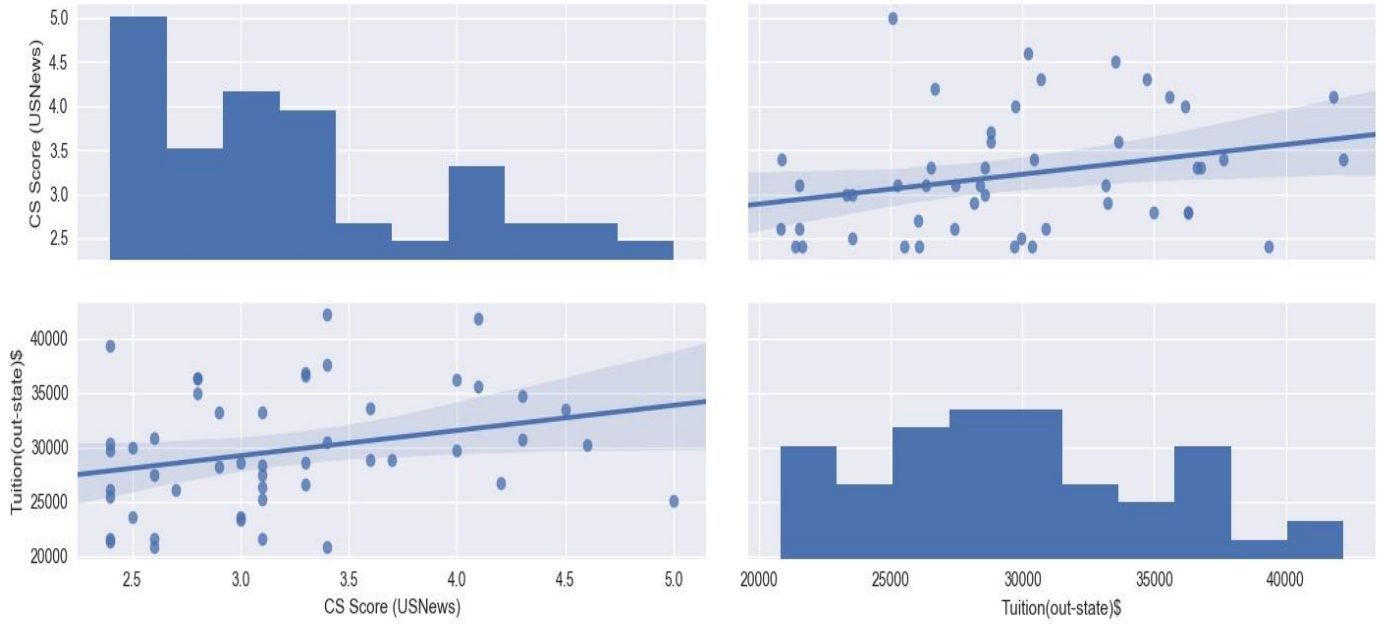
We can analyze that the properties - research and admin_base_pay have positive correlation and the linear regression line emphasizes the same.

Admin Base Pay and CS Score pairwise plot



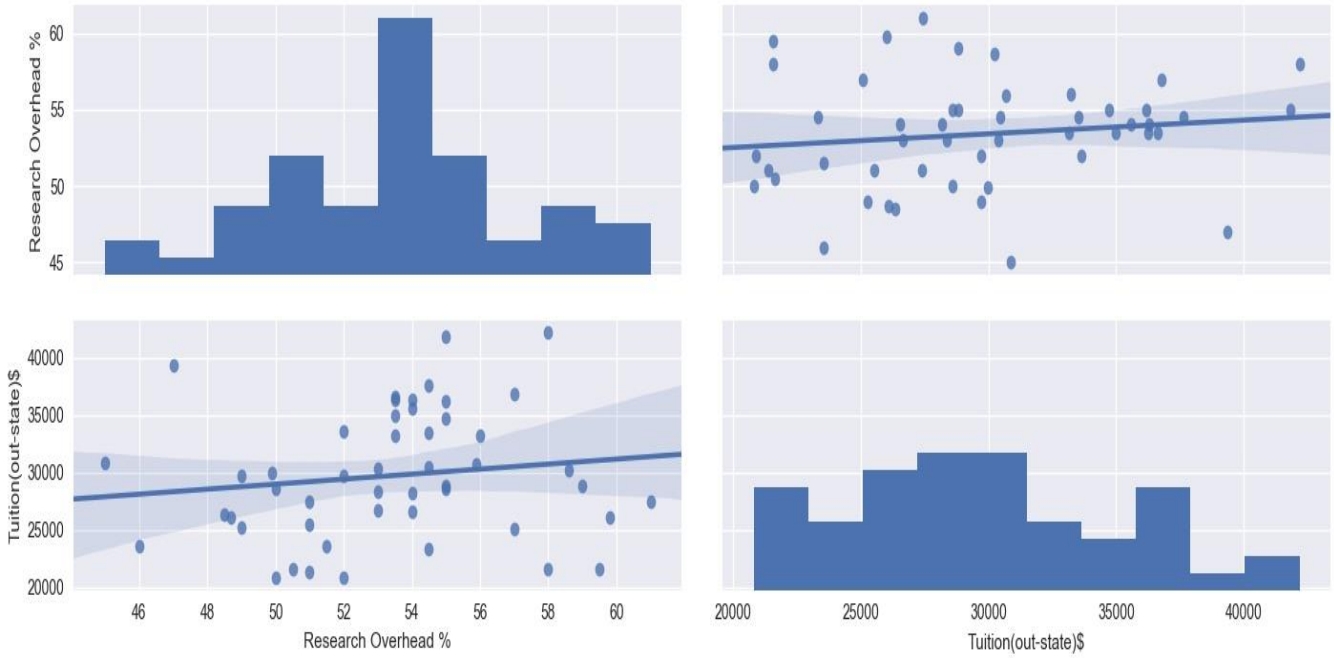
We can analyze that the properties – Admin Base Pay and CS Score have positive correlation and the linear regression line emphasizes the same.

CS Score and Tuition pairwise plot



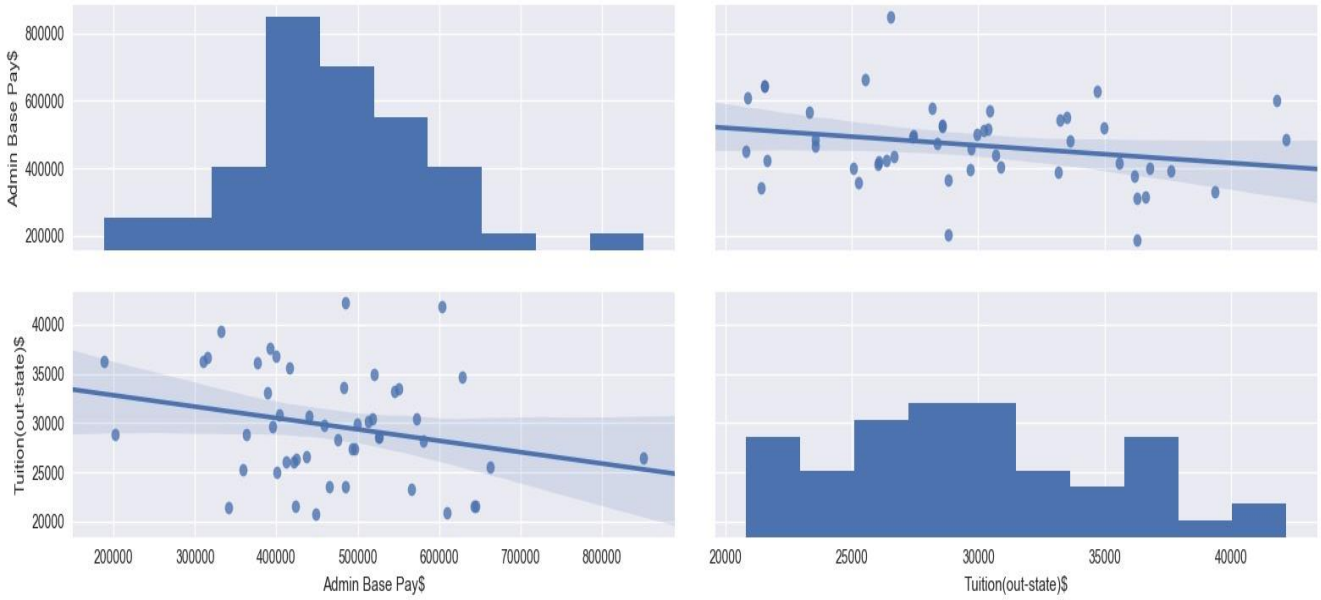
We can analyze that the properties – Admin Base Pay and CS Score have positive correlation and the linear regression line emphasizes the same.

Research Overhead and Tuition pairwise plot



We can analyze that the properties - tuition and Research Overhead have positive correlation and the linear regression line emphasizes the same.

Admin Base Pay and Tuition pairwise plot



We can analyze that the properties – Tuition and Admin Base Pay have negative correlation and the linear regression line emphasizes the same.

Task 3

Result

logLikelihood = -1315.099

logLikelihood Multivariate = -1304.777

Approach

Univariate logLikelihood: We have calculated pdf for each element of the dataset assuming them as independent variables. We have calculated likelihood for each of the dataset. Then we have 49 likelihoods across 49 data set. Then to calculate logLikelihood, we took log of these values and took sum of all the values.

Multivariate logLikelihood: We have calculated pdf for each row vector of the dataset assuming their properties as dependent variables. Then to calculate logLikelihood, we took log of these pdf values and took sum of all the values.

Output

```
D:\Anaconda3\python.exe D:/Users/Rakesh/PycharmProjects/proj1code/main.py
UBitName = pbisht2
personNumber = 50247429
UBitName = rakeshsi
personNumber = 50249135
mu1 = 3.214
mu2 = 53.386
mu3 = 469178.816
mu4 = 29711.959
var1 = 0.448
var2 = 12.588
var3 = 13900134681.7
var4 = 30727538.733
sigma1 = 0.669
sigma2 = 3.548
sigma3 = 117898.832
sigma4 = 5543.243
covarianceMat =
 [[  4.57000000e-01   1.10600000e+00   3.87978200e+03   1.05848000e+03]
 [  1.10600000e+00   1.28500000e+01   7.02793760e+04   2.80578900e+03]
 [  3.87978200e+03   7.02793760e+04   1.41897208e+10  -1.63685641e+08]
 [  1.05848000e+03   2.80578900e+03  -1.63685641e+08   3.13676958e+07]]
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.   ]]
logLikelihood = -1315.099
logLikelihood MultiVariate = -1304.777

Process finished with exit code 0
```