Project Report
Probability distribution and Bayesian networks

Submitted by
Hima Sujani Adike, 50246828
Anjali Sujatha Nair ,50248735
Soumya Venkatesan, 50246599

Summary and Status

In the project, used python to evaluate necessary statistics like mean, variance, co variance and correlation of the given data of university rankings. Also evaluated the univariate and multi variate probability distribution functions of the given data.

Observations

Mean
mu1 = 3.214
mu2 = 53.386
mu3 = 469178.816
mu4 = 29711.959

Variance
var1 = 0.448
var2 = 12.588
var3 = 13900134681.701
var4 = 30727538.733

Standard Deviation
sigma1 = 0.669
sigma2 = 3.548
sigma3 = 117898.832
sigma4 = 5543.243

Co variance matrix :

```
[[0.457, 1.106, 3879.782, 1058.480],
[1.106, 12.850, 70279.376, 2805.789],
[3879.782, 70279.376, 14189720820.903, -163685641.258],
[1058.480, 2805.789, -163685641.258, 31367695.790]]
```

Correlation matrix :

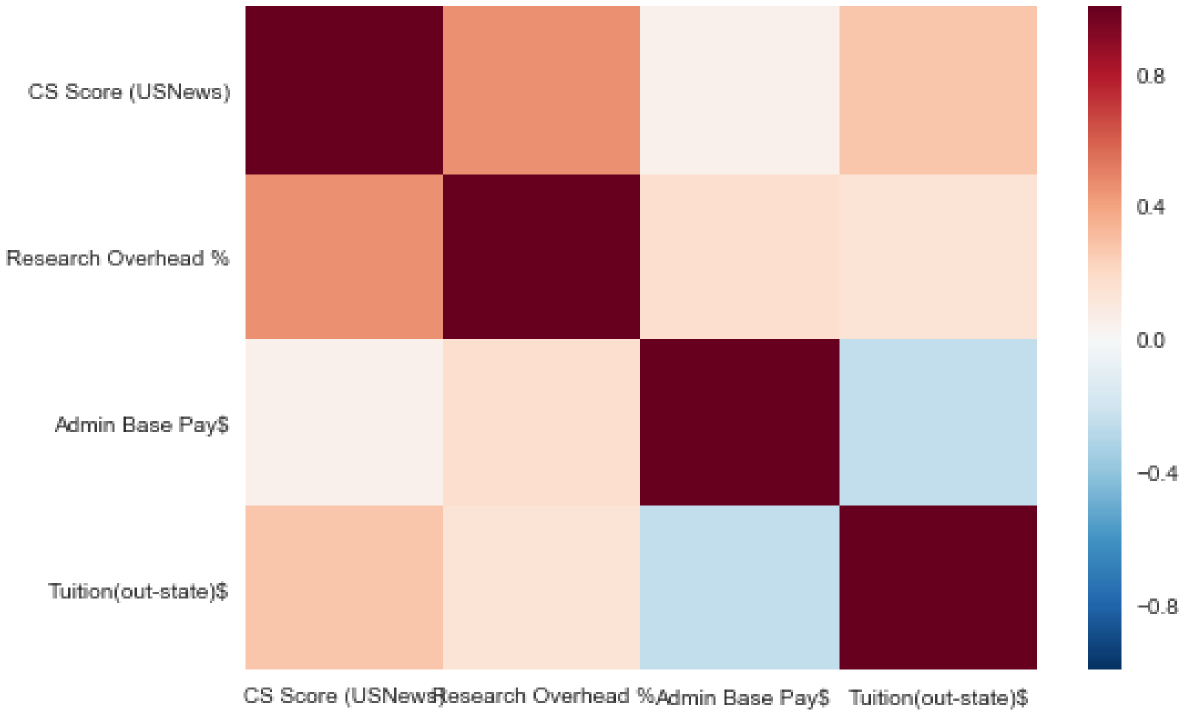
```
[[1.000, 0.456, 0.048, 0.279],
[0.456, 1.000, 0.165, 0.140],
[0.048, 0.165, 1.000, -0.245],
[0.279, 0.140, -0.245, 1.000]]
```

Univariate loglikelihood obtained from calculating formula in code : -1315.099
Univariate loglikelihood obtained from in built function: -1315.099
Multivariate loglikelihood obtained from calculating formula in code: -1304.778
Multivariate loglikelihood obtained from inbuilt function: -1262.327

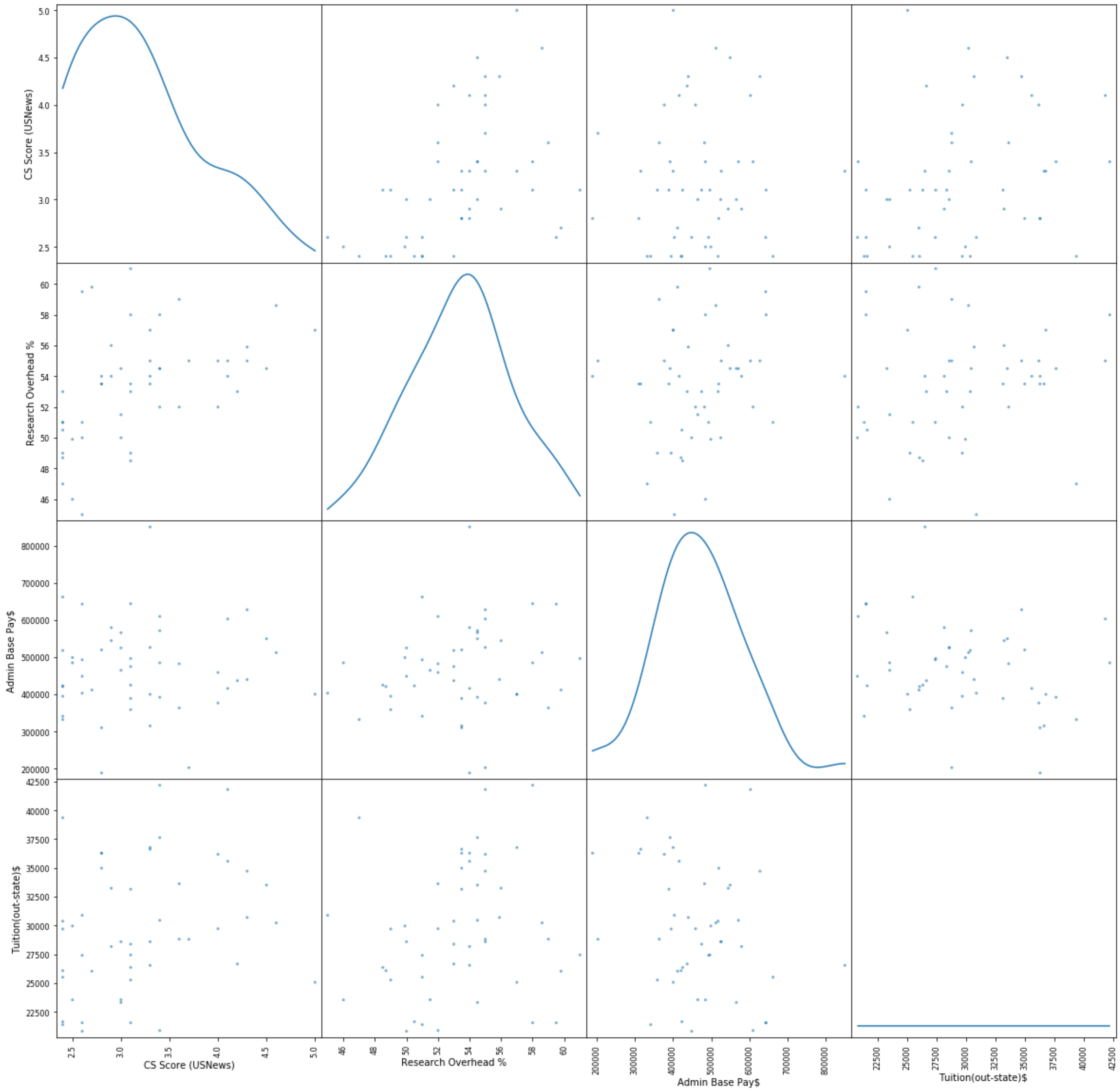
Inference

Correlation - gives us a measure of the strength of the relationship. It takes the value between -1 and 1, a correlation value of –ve indicates a –ve relationship. For every change in value of one type, the magnitude of the other kind moves in an opposite direction. Looking at the obtained correlation matrix, the Admin Base Pay$ and the Tuition(out-state)$ has a strong –ve correlation value of -0.245. Also the CS Score (USNews) and the Research Overhead % has high +ve correlation of 0.456.

Plotting the heat map for correlation matrix



Co-variance – To infer the co variance , pair wise scatter plots were made using scatter_matrix.

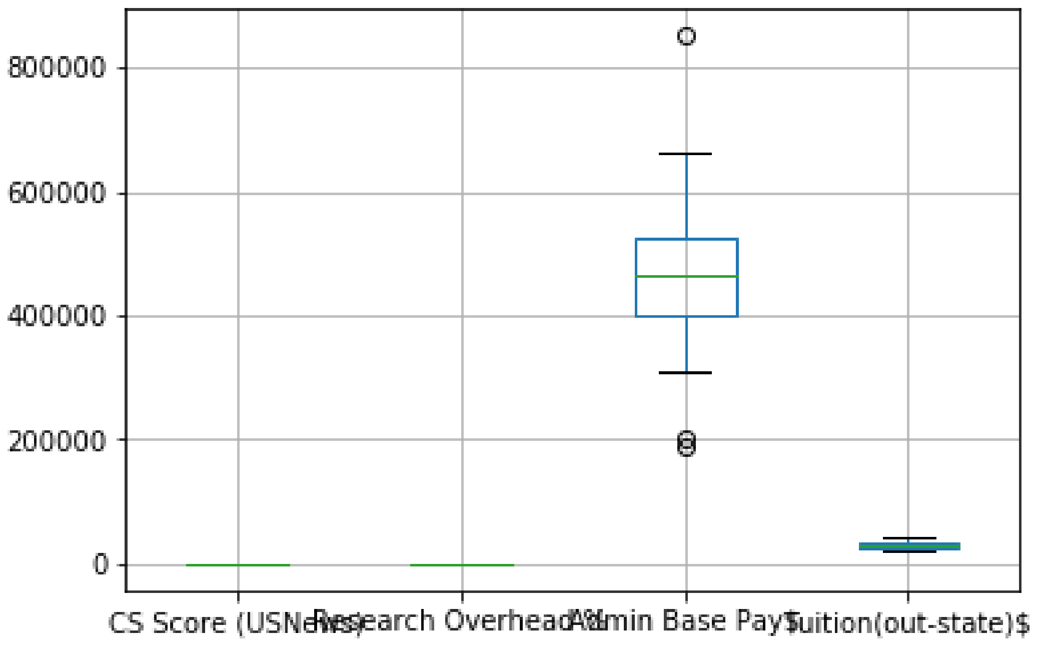


covariance is the extent to which the variance in one variable depends on another variable. Covariances can be positive (both variables move in the same direction), negative (both variables move in different directions), or in the case of no relationship, zero.

In the above plot, for when the value of Research Overhead, increases, the value of CSScore seems to be increasing (scatter plot image(0,1)). Hence there is strong co variance between CSScore and Research Overhead.

Looking at the scatter plot of the Admin Base with other groups, the data does not indicate an increase in magnitude. There is no clear indication shown.

Variance – It gives a measure of how spread out the observations of the variable are from the mean value. Box plot gives a measure of the variance of the dispersion of the variables. Higher the variance more spread out the variables are. The figure below shows the box plot of the four variables.



Log Likelihood :

Loglikelihood indicates the likelihood of the function given a set of data. Since logarithm achieves maximum at the highest point of the function, log likelihood can be used in maximum likelihood estimation. A positive log likelihood means that the likelihood is larger than 1. For the given set of data the log likelihood calculated multivariate as well as univariate is negative number.

Code implemented including the plots :

```
from pandas import read_excel
from numpy import array,mean,var,std,vstack,matrix,corrcoef,cov,set_printoptions
from math import log,exp,pi,sqrt
from scipy.stats import multivariate_normal
from numpy.linalg import inv,det

def getCorrelationMatrix(dataStack):
  return corrcoef(dataStack)

def getCovarianceMatrix(dataStack):
  return cov(dataStack)

def multivaraiateLikelihood():
  loglikelihood=0
  size=(dataList.size/len(X))
  determinant=det(covarianceMat)
  K=(1/sqrt((((2*pi)**len(X))*abs(determinant))))
  for row in range(0,int(size)):
    diff=matrix(dataList[row]-meanList)
    probability=K*exp((-1/2)*diff*inv(covarianceMat)*diff.transpose())
    loglikelihood+=log(probability)
    return round(loglikelihood,3)

def independentLoglikelihood():
    loglikelihood=0
    k=(1/sqrt(2*pi))
    for row in range(0,49):
        probability=1
        for col in range(0,4):
            diff = float((dataList[row][col]-meanList[col])/abs(sigmaList[col]))
            probability *= (1/(sqrt(2*pi)*abs(sigmaList[col])))*exp(-diff*diff/2)
        loglikelihood+=log(probability)
    return round(loglikelihood,3)

data = read_excel('university data.xlsx')
X=['CS Score (USNews)','Research Overhead %','Admin Base Pay$','Tuition(out-state)$']
set_printoptions(formatter={'float': lambda x: "{0:0.3f}".format(x)})
dataList=[]
meanList=[]
varList=[]
sigmaList=[]
i=1
for attr in X:
     attrValues = array(data[attr].dropna())
     locals()['mu{0}'.format(i)] = round(mean(attrValues),3)
     locals()['var{0}'.format(i)] = round(var(attrValues),3)
     locals()['sigma{0}'.format(i)] = round(std(attrValues),3)
     meanList.append(locals()['mu{0}'.format(i)])
     varList.append(locals()['var{0}'.format(i)])
     sigmaList.append(locals()['sigma{0}'.format(i)])
     dataList.append(attrValues)
     i=i+1

dataStack = vstack((dataList))
covarianceMat = getCovarianceMatrix(dataStack)
correlationMat = getCorrelationMatrix(dataStack)
dataList=array(data.loc[:,X].dropna())
print("UBitName = ")
print("personNumber = ")
for i in range(1,4):
   print("mu"+str(i)," = ",locals()['mu{0}'.format(i)])
for i in range(1,4):
   print("var"+str(i)," = ",locals()['var{0}'.format(i)])
for i in range(1,4):
   print("sigma"+str(i)," = ",locals()['sigma{0}'.format(i)])
print("covarianceMat = ",covarianceMat)
print("correlationMat = ",correlationMat)
print("logLikelihood = ",independentLoglikelihood())
print("multivariatelogLikelihood = ", multivaraiateLikelihood())
```

References

http://a.web.umkc.edu/andersonbri/Variance.html

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.multivariate_normal.html
https://www.stata.com/statalist/archive/2007-07/msg00914.html
https://en.wikipedia.org/wiki/Maximum_likelihood_estimation