Project 1 Report

Chaitanya Chembolu – 50248987 Kautuk R Desai - 50247648

Problem Statement

The project aims to evaluate the basic statistics related to probability theory that help us to understand basic machine learning concepts.

These statistics include finding the mean, standard deviation, variance of multiple variables given in the dataset.

Further, we find covariance, and correlation of the variables in order to find the log likelihood for two different cases:

i. Considering the variables as dependent and
ii. Independent of each other.

The given dataset is about the computer science program rankings of select universities from sources such as US News, The Chronicle of Higher Education. The dataset consists of the following variables, CS Score, Research Overhead, Admin Base Pay and Tuition with 49 samples each.

Approach

In order to find mean, variance, and standard deviation we use different functions from numpy package. We then calculate the covariance (considering the variables to be dependent, independent) and correlation between the variables.

While finding the variance and covariance, we use numpy functions (.var(), .cov()) with ddof (Delta Degrees of Freedom) as 1. The default implementation of variance (and covariance) uses ddof=0 which gives, $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$. Where σ ² is the variance, and μ is the mean of the variable.

But with ddof = 1, we get the denominator of variance expression as N-1. The expression is shown below,

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}[x(i) - \mu]^2$$

To calculate the log likelihood of the data considering the variables as independent, we followed two approaches

i. Using the formula for Gaussian distribution of independent variables and implementing it in python.
ii. Using the multivariate_normal.logpdf() function from scipy package. To compute, we use the covariance matrix for independent variables.

Results & Observations

As per the approach mentioned above, the following results were obtained

Statistical Parameter	CS Score	Research Overhead	Admin Base Pay	Tuition
Mean: μ	3.214	53.386	469178.816	29711.959
Variance: σ ²	0.475	12.85	1.4190e+10	3.1368e+07
Standard Deviation: σ	0.676	3.585	119120.615	5600.687

The covariance matrix (dependent variables):

	CS Score	Research Overhead	Admin Base pay	Tuition
CS Score	4.5800e-01	1.1060e+00	3.8798e+03	1.0585e+03
Research Overhead	1.1060e+00	1.2850e+01	7.0279e+04	2.8058e+03
Admin Base pay	3.8798e+03	7.0279e+04	1.4190e+10	-1.6369e+08
Tuition	1.0585e+03	2.8058e+03	-1.6369e+08	3.1368e+07

The corresponding correlation matrix:

	CS Score	Research Overhead	Admin Base pay	Tuition
CS Score	1	0.456	0.048	0.279
Research Overhead	0.456	1	0.165	0.14
Admin Base pay	0.048	0.165	1	-0.245
Tuition	0.279	0.14	-0.245	1

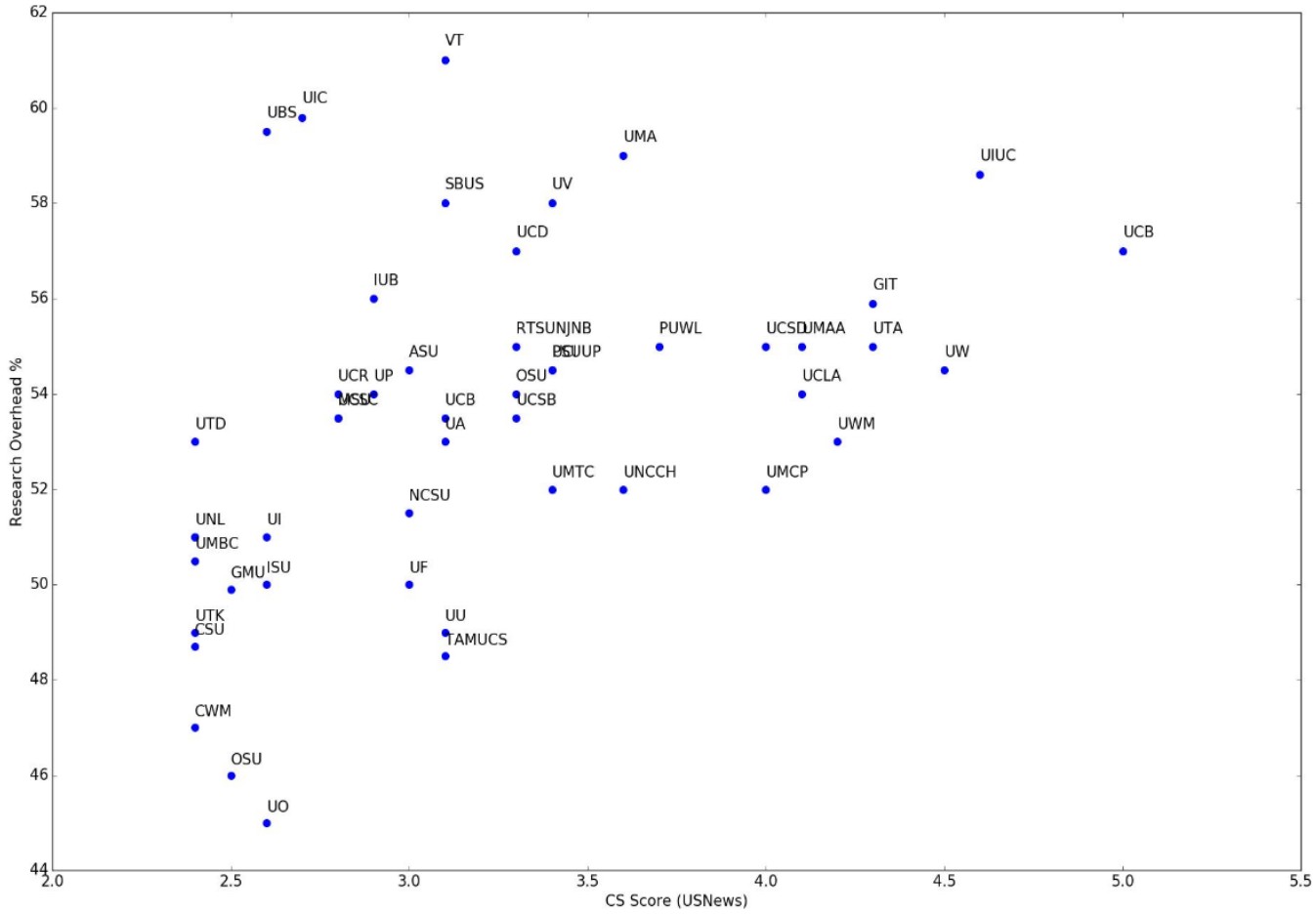
The Covariance matrix (independent variables):

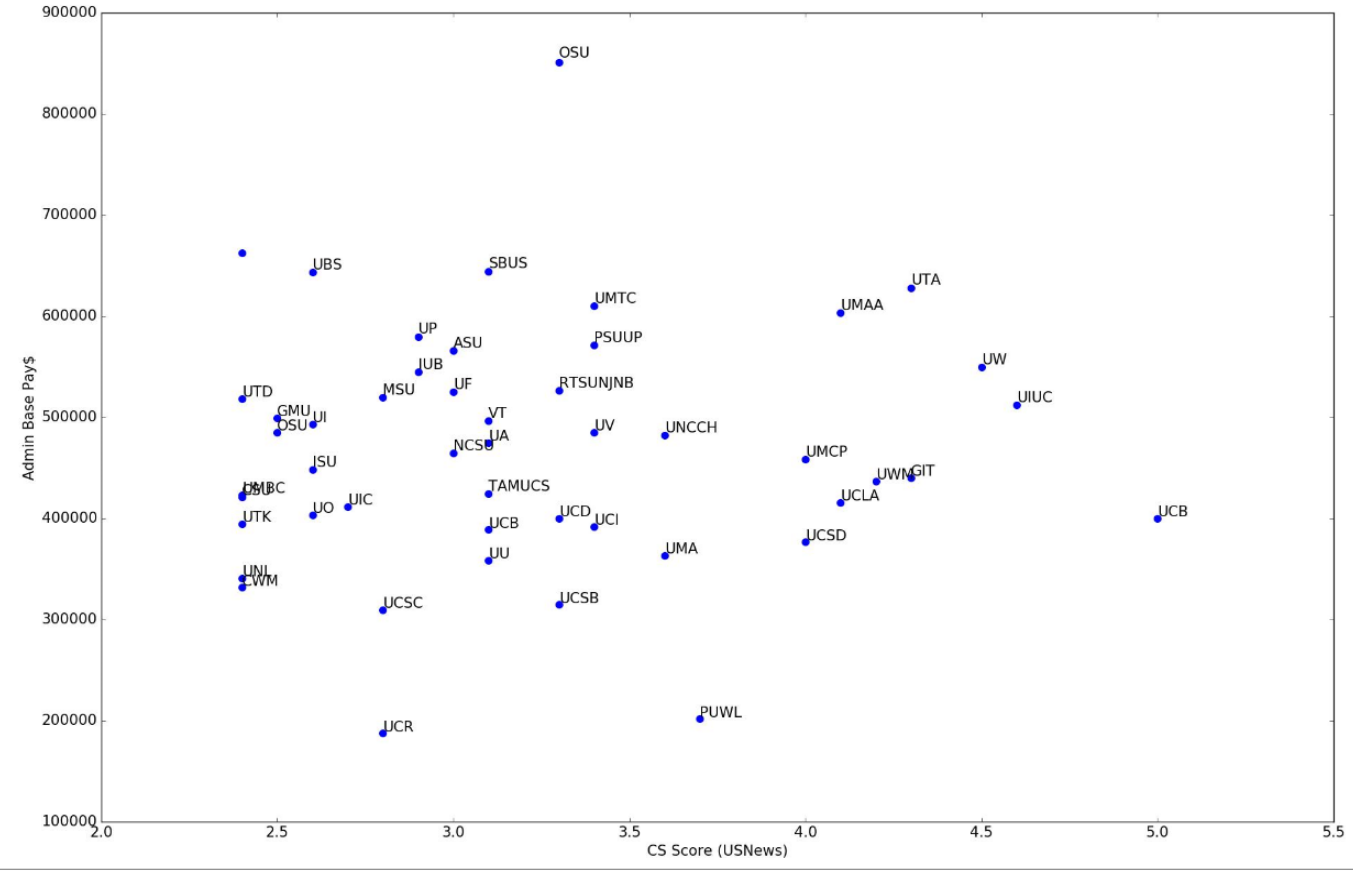
	CS Score	Research Overhead	Admin Base pay	Tuition
CS Score	4.5800e-01	0	0	0
Research Overhead	0	1.2850e+01	0	0
Admin Base pay	0	0	1.4190e+10	0
Tuition	0	0	0	3.1368e+07

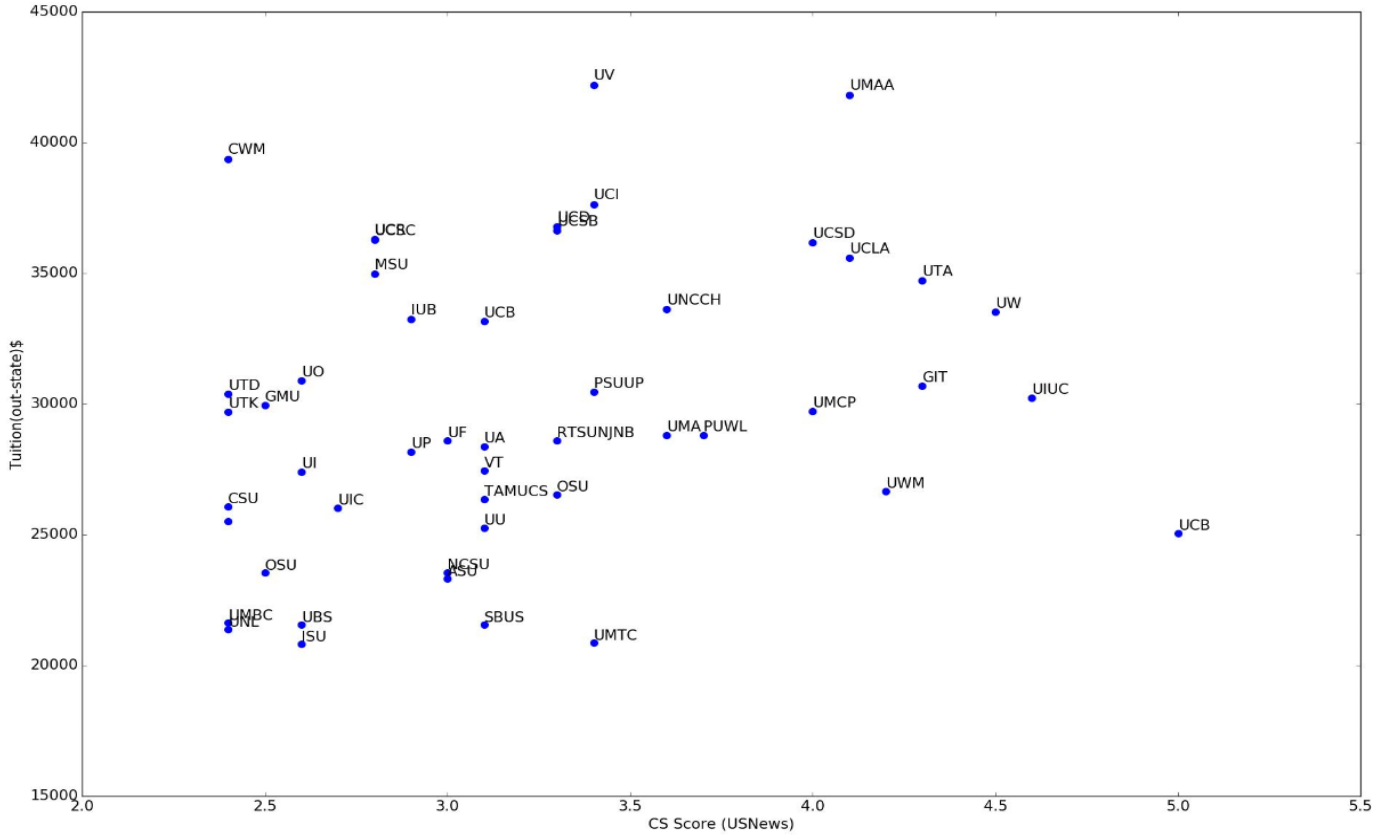
The log likelihood of the given dataset:

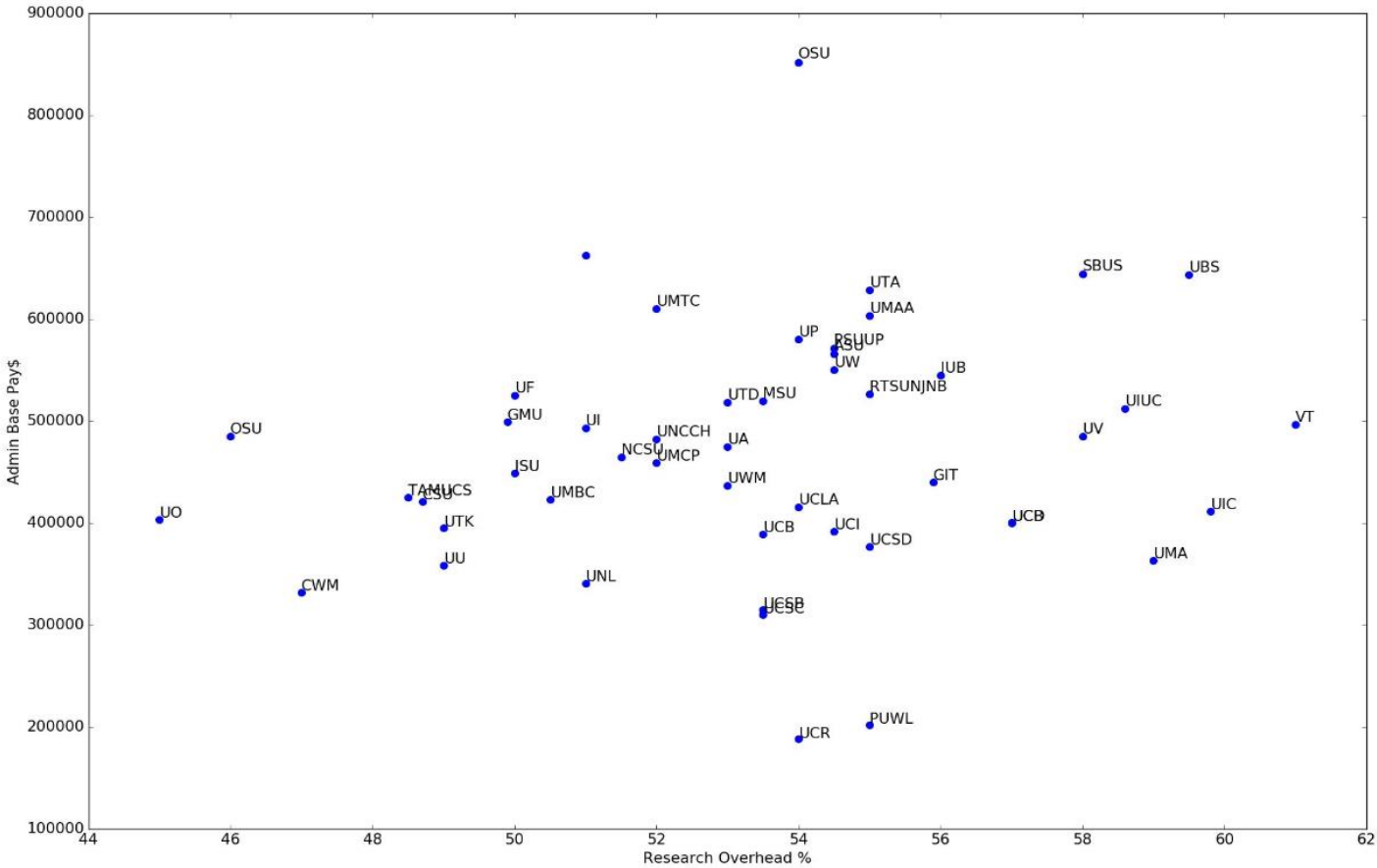
Variable Relation	Compute Method	Log Likelihood value
Independent	Formula: loglikelyhood = $\Sigma_{i=1}^{N}$log p(x_i) Where, $p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]$	-1315.119
Independent	Function: >> multivariate_normal.logpdf()	-1265.25
Dependent	Function: >> multivariate_normal.logpdf()	-1262.327

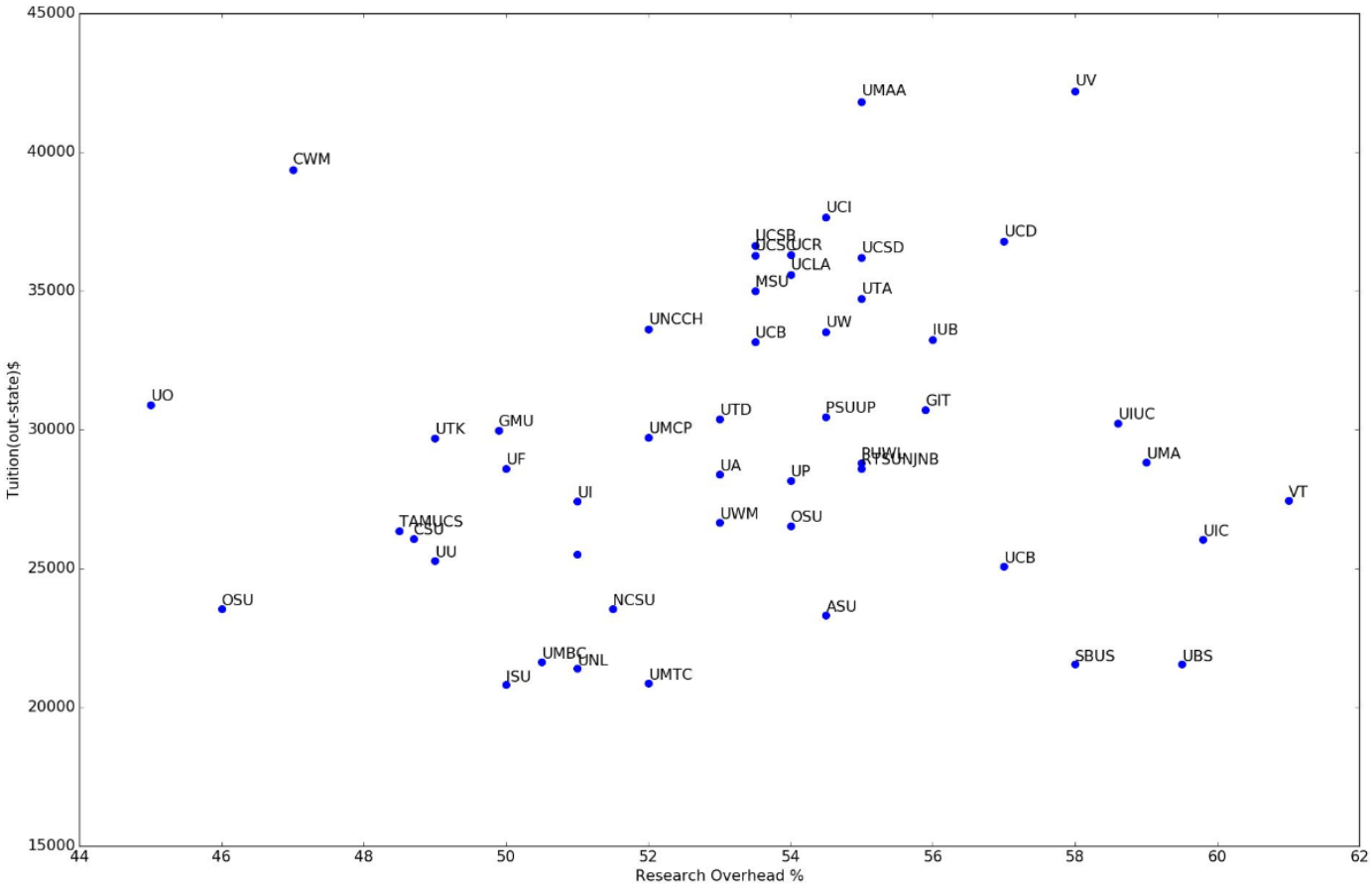
The pairwise data showing the label associated with each data point of the variables is shown below,

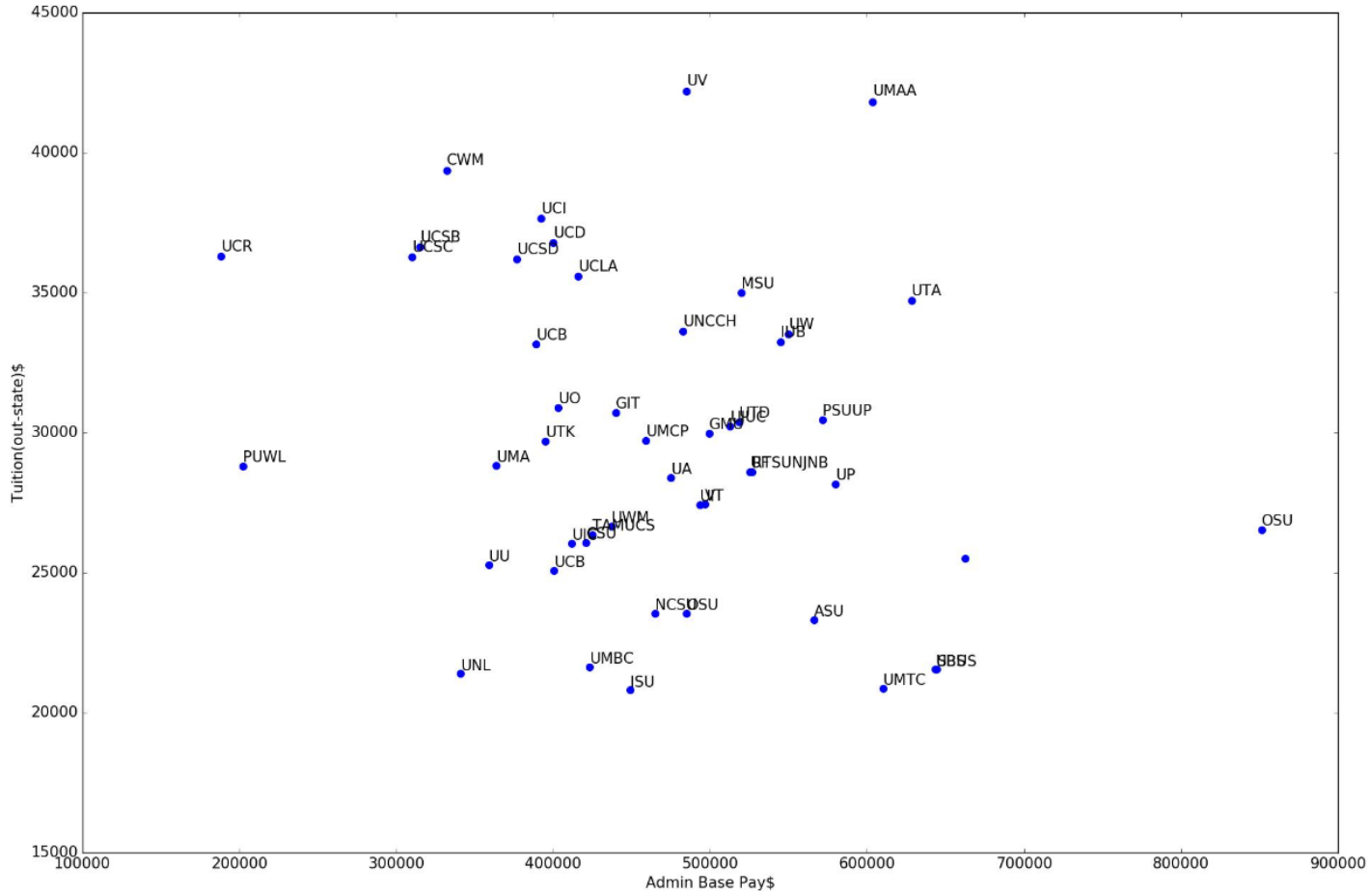

Research Overhead %
CS Score (USNews)
VT
UIC
UBS
UMA
UIUC
SBUS
UV
UCD
UCB
IUB
GIT
RTSUNJNB
PUWL
UCSD
UMAA
UTA
ASU
UW
UCR
UP
OSU
UCLA
UCB
UCSB
UTD
UA
UWM
UMTC
UNCCH
UMCP
NCSU
UNL
UI
UMBC
GMU
ISU
UF
UTK
CSU
UU
TAMUCS
CWM
OSU
UO


Admin Base Pay$
CS Score (USNews)
OSU
UBS
SBUS
UTA
UMTC
UMAA
UP
ASU
PSUUP
UW
IUB
UTD
MSU
UF
RTSUNJNB
UIUC
GMU
UI
VT
OSU
UV
UNCCH
UA
NCSU
UMCP
ISU
UWM
GIT
TAMUCS
UCLA
UIC
UO
UCD
UCI
UCB
UTK
UCSD
UMA
UU
UCSC
UCSB
PUWL
UCR


Tuition(out-state)$
CS Score (USNews)
45000
40000
35000
30000
25000
20000
15000
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
UV
UMAA
CWM
UCI
UCSD
UCLA
MSU
UTA
UNCCH
UW
IUB
UCB
UO
GIT
UTD
PSUUP
UIUC
UTK
GMU
UMCP
UP
UF
UA
RTSUNJNB
UMA
PUWL
UI
VT
UWM
CSU
UIC
TAMUCS
OSU
UU
OSU
NCSU
UMBC
UBS
SBUS
UNL
ISU
UMTC


Admin Base Pay$
Research Overhead %
900000
800000
700000
600000
500000
400000
300000
200000
100000
44
46
48
50
52
54
56
58
60
62
OSU
SBUS
UBS
UTA
UMTC
UMAA
UP
PSUUP
NCSU
UW
IUB
UF
UTD
MSU
RTSUNJNB
UIUC
GMU
UI
VT
OSU
UNCCH
UV
UA
NCSU
UMCP
ISU
GIT
UWM
TAMUCS
UMBC
UCLA
UIC
UO
UTK
UCB
UCI
UCSD
UMA
UU
UNL
CWM
PUWL
UCR


Tuition(out-state)$
Research Overhead %
45000
40000
35000
30000
25000
20000
15000
44
46
48
50
52
54
56
58
60
62
UMAA
UV
CWM
UCI
UCSB
UCSC
UCR
UCSD
UCD
UCLA
MSU
UTA
UNCCH
UW
UCB
IUB
UO
UTK
GMU
UMCP
UTD
PSUUP
GIT
UIUC
UF
UA
UP
UMA
UI
VT
TAMUCS
CSU
UWM
OSU
UIC
UU
UCB
OSU
NCSU
ASU
UMBC
UNL
SBUS
UBS
JSU
UMTC


Tuition(out-state)$
Admin Base Pay$
45000
40000
35000
30000
25000
20000
15000
100000
200000
300000
400000
500000
600000
700000
800000
900000
UV
UMAA
CWM
UCI
UCD
UCR
UCSB
UCSC
UCSD
UCLA
MSU
UTA
UNCCH
UW
IUB
UCB
UO
GIT
UTD
PSUUP
UTK
UMCP
PUWL
UMA
UA
UP
VT
OSU
UU
UCB
NCSU
OSU
ASU
UNL
UMBC
JSU
UMTC

Observation & Conclusion

From the pairwise data plots and correlation matrix, we can infer that CS Score and Research Overhead are the most correlated pair whereas CS Score and Admin Base Pay are the least correlated pair. Also, the log likelihood of the data is higher when the variables are dependent compared to when variables are independent.