Probability Distributions and Bayesian Networks

CSE 574 – Machine Learning

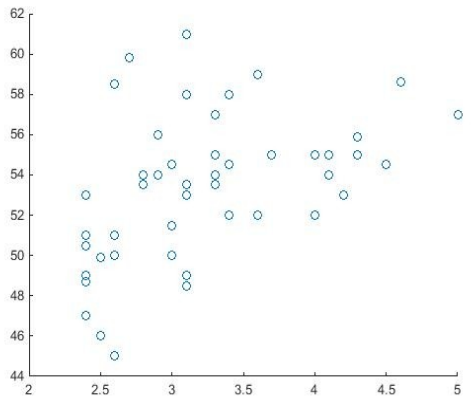
Karanjeet Singh

1. Problem Introduction

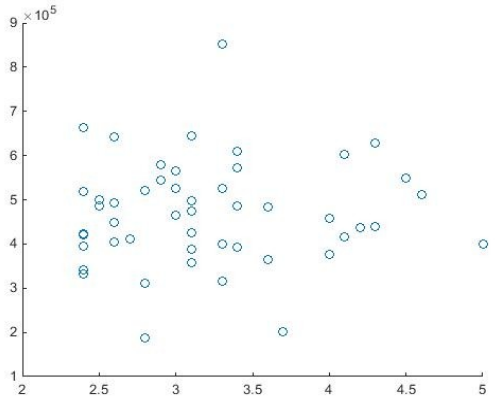
This project concerns probability distributions of several variables and to find the optimal Bayesian Network correlating them with the highest joint probability distribution. Before getting to this we also find the basic statistical properties of the variables, namely: Mean, Variance, Standard Deviation of each variable and a covariance and correlation matrix to analyze the behavior of the variable and try to find a correlation among them. We also compute the log-likelihood of the variables assuming them to be independent and mutually exclusive. This gives us a benchmark to optimize the variables correlation. We use the exhaustive technology to find the best Bayesian Network which returns the best log-likelihood. Thus, we are trying to find the best BN representation of the variables in agreement with the data provided. We will be using Matlab explicitly to evaluate the statistics.

2. Dataset

The data was provided from http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/ top-science-schools/computer-science-rankings. This data represents some facts about some top universities in US. It has the ranking of the University, research overhead, admin base pay and tuition fee. We will find the most optimal Bayesian Network to represent these parameters and find a co-relation among them. What factor affect the school ranking directly and indirectly should be the motive.



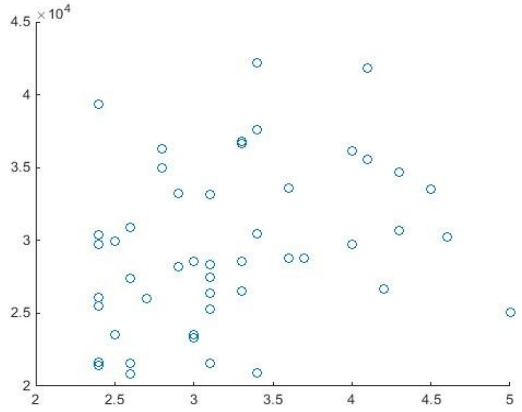
Representation of CSScore with ResearchOverhead



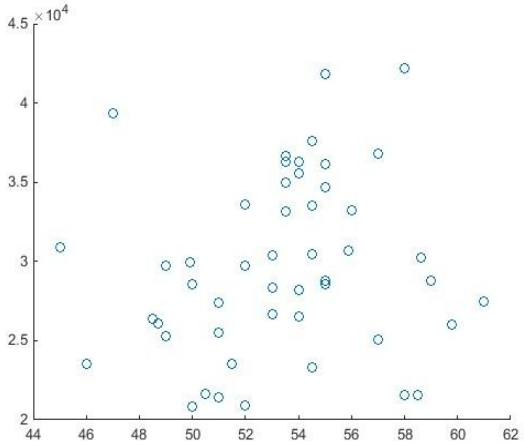
Representation of CSScore with AdminBasePay

We can see the trend the first figure, as the score of the school increases, the research overhead also tends to increases.
In the second figure, the AdminBasePay is somewhat constant with the CSScore, but we need to evaluate the relation between the both.

The thrid figure shows the relation between the CSScore and the TuitionFee. We can derive that the fee tends to increase with the the score.
Fourth figure depicting the graph between ResearchOverhead and TuitionFee shows that the as research overhead increases, tuition fee increases.

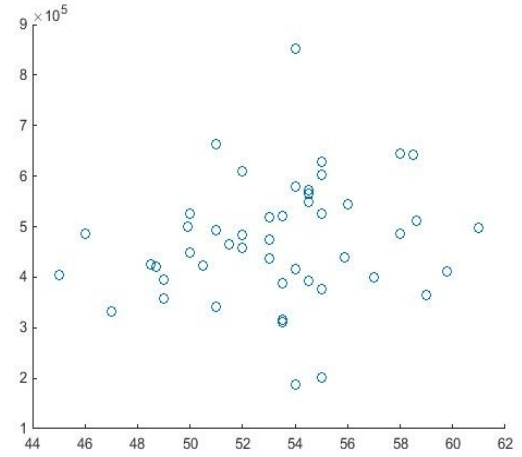


Representation of CSScore with TuitionFee

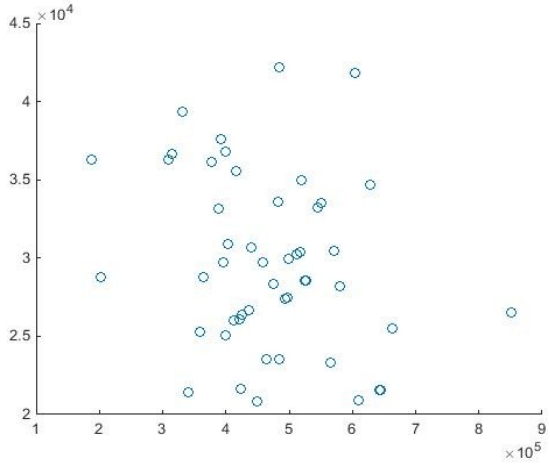


Representation of ResearchOverhead with TuitionFee

In the observations done below, we see that, Research Overhead is proportional to the AdminBasePay. As the pay increases, the researchoverhead increases is the general trend here.
In the graph between AdminPay and Tuitionfee, we see that although the plot is distributed in all whole range for a given AdminPay but we see small clusters and the position of these clusters goes higher as we move to the right of the x axis.



Representation of ResearchOverhead with AdminBaseBay

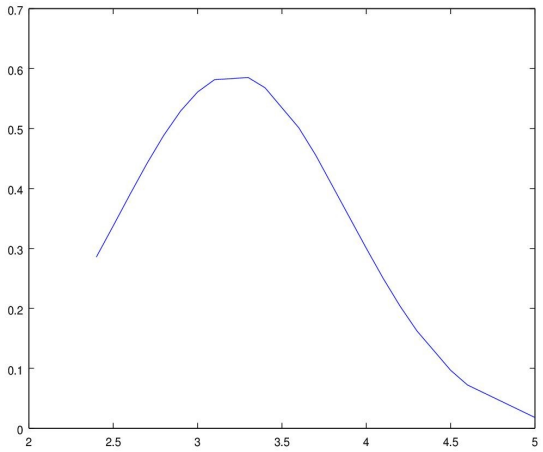


Representation of AdminBasePay with TuitionFees

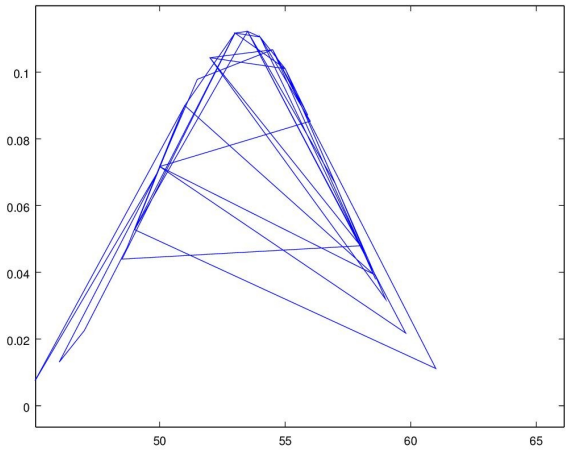
2. Features and Processing

Firstly we find the mean, variance and standard variation for each of the random variable to help in further processing and to get more insight into the data. We also calculate the covariance matrix and correlational matrix for the random variables.

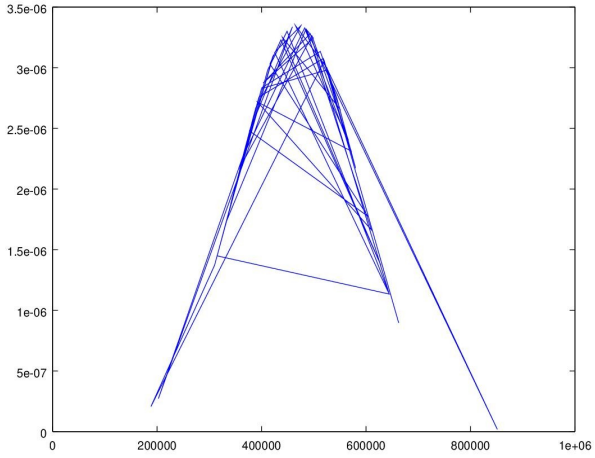
The normal distribution of the variables is as follows:



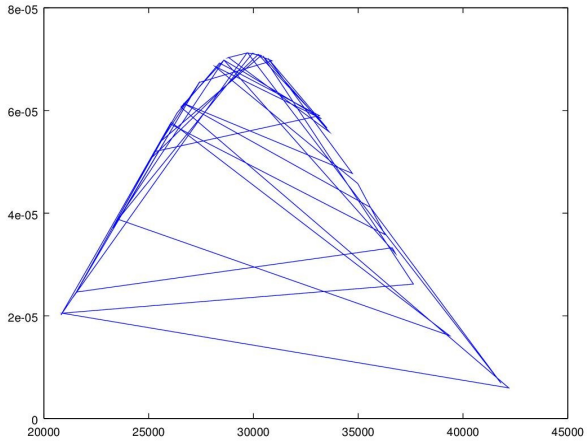
representation of distribution of CSScore



representation of distribution of ResearchOverhead



representation of distribution of AdminBasePay



representation of distribution of TuitionFees

To calculate the mean, variance, sigma, correlation-matrix, covariance matrix of the given variables, we firstly read the xls file in Matlab.

Using,

```
xls_file = xlsread('university data.xlsx');
```

Let,

```
x=[xls_file(:,3),xls_file(:,4),xls_file(:,5),xls_file(:,6)];
```

to find mean,

```
total_mean=mean(x);
mu1=total_mean(1);
mu2=total_mean(2);
mu3=total_mean(3);
mu4=total_mean(4);
```

to calculate variance:

```
total_var=var(x);
var1=total_var(1);
var2=total_var(2);
var3=total_var(3);
var4=total_var(4);
```

to calculate sigma:

```
total_sigma = std(x);
sigma1=total_sigma(1);
sigma2=total_sigma(2);
sigma3=total_sigma(3);
sigma4=total_sigma(4);
```

```
covarianceMat=cov(x);
correlationMat=corrcoef(x);
```

```
correlationMat =
   1.000000   0.465531   0.048153   0.279412
   0.465531   1.000000   0.157530   0.149591
   0.048153   0.157530   1.000000  -0.245348
   0.279412   0.149591  -0.245348   1.000000
```

We can see that according to the correlationMat, CSScore is most related to ResearchOverhead.

In the next step we calculate the log-likelihood of the variables assuming they are not related to each other. Thus,

$$P(A,B,C,D) = P(A).P(B).P(C).P(D)$$

when all are independent of each other.

We find the probability of each of variables using the normpdf function in MATLAB. We find the normpdf and then we take log and sum of the column to find the loglikelyhood for that column.

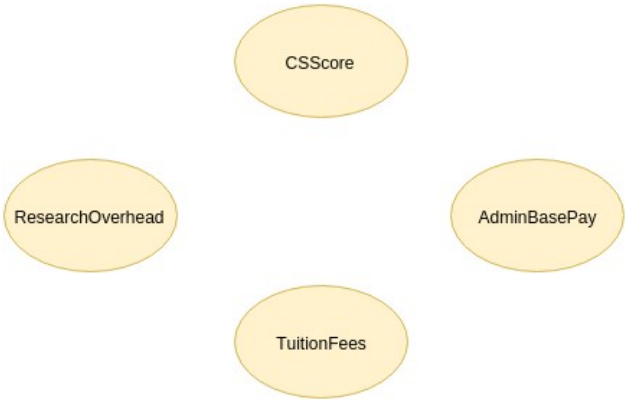
log1=sum(log(normpdf(x(:,1),total_mean(1),total_sigma(1))));

similarly we calculate log2,log3,log4 and add them to find the log-likelyhood of the variables assuming all the variables are independent of each other.

logLikelihood = log1 + log2 + log3 + log4;

this is the probility when none of the variables are related and thus, this is the least optimal probability.

Variables independent of each other,
loglikelyhood = -1314.7



To find the most optimal loglikelyhood, we will use the Exhaustive technique, we firstly find all the possible matrices having values only 0s and 1s which can form Directed Acyclic Graphs (DAG)s. Here the 4 variables will give us 548 possible matrices, thus 548 Bayesian Networks. For each of these matrices we find the loglikelyhood and find the matrix which returns the highest value.

There actually are 2^16 combinations but as BNs can only be DAGs, no nodes can be cyclic. We cannot have cyclic subnodes, ie 1,2,3 nodes being cyclic. So, we find iterations of all the matrices possible and we test them with this condition in the iteration. If the matrix passes the test, we find the loglikelyhood of the matrix. During the iteration, we store the loglikelyhood in a list and we also store all the combinations of the matrices which pass the test.

Then, using the index of the max(loglikelyhood), we find the index of the matrix. This matrix will be the most optimized Bayesian Network and the likelyhood of this BN will be the most optimized loglikelyhood.

To find the loglikelyhood,

for each column in the matrix we check the position of 1s in the column, these are the dependencies of that variable. To find the loglikelyhood of the column,

$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)}$$

P(A|B,C) = loglikelyhood of A,B,C – loglikelyhood of B,C

We calculate the loglikelyhood of the numerator and subtract it by the loglikelyhood of the denominator. To do this we need to find the loglikelyhood of multiple variables.

If s is a 0,1 matrix which is a DAG, then to find the 1s in the first column we do, **col_1_depd=find(s(:,1));**

similarly we do for other columns.

If the col_X_depd is a null set, then in that case the variable is not dependent on any other variable for that given matrix.

col_1=[1;col_1_depd]; this is to merge the column no. with the dependencies which will be used to find the numerator for the loglikelyhood.

We use the mvnpdf function to find the probability for more than one variables.

numerator_prob = sum(log(mvnpdf(x(:,col_1),total_mean(col_1),covarianceMat(col_1,col_1)))))

If **col_1_depd == NULL**

If there were no dependencies for that particular variable, then the loglikelyhood of that variable which we had calculated earlier using normpdf will be the probability of that variable.

prob1=log1 (we had calculated log1,
log1=sum(log(normpdf(x(:,1),total_mean(1),total_sigma(1))));)

If **col_1_depd !=NULL**

prob1=(sum(log(mvnpdf(x(:,col_1),total_mean(col_1),covarianceMat(col_1,col_1))))) –(sum(log(mvnpdf(x(:,col_1_depd),total_mean(col_1_depd),covarianceMat(col_1_depd,col_1_depd)))))

prob1 gives us the the probability for the first column, we do the same as above for rest of the columns and find prob2, prob3 and prob4. Lastly, we add the probabilities of each column to find the loglikelyhood for the given matrix. Thus,

total_prob=prob1+prob2+prob3+prob4;

For each of the matrices, we save the total_prob in a list and find the maximum one. This will return the optimal loglikelyhood.

BnlogLikelihood=max(r);

if we had saved the DAGs too in a list say, bigmat, then the index of the optimal loglikelyhood = index of the optimized Bayesian Network.

index=find(r==max(r),1);

BNgraph=bigmat{index};

After doing the above calculations for all the DAGs we find that,

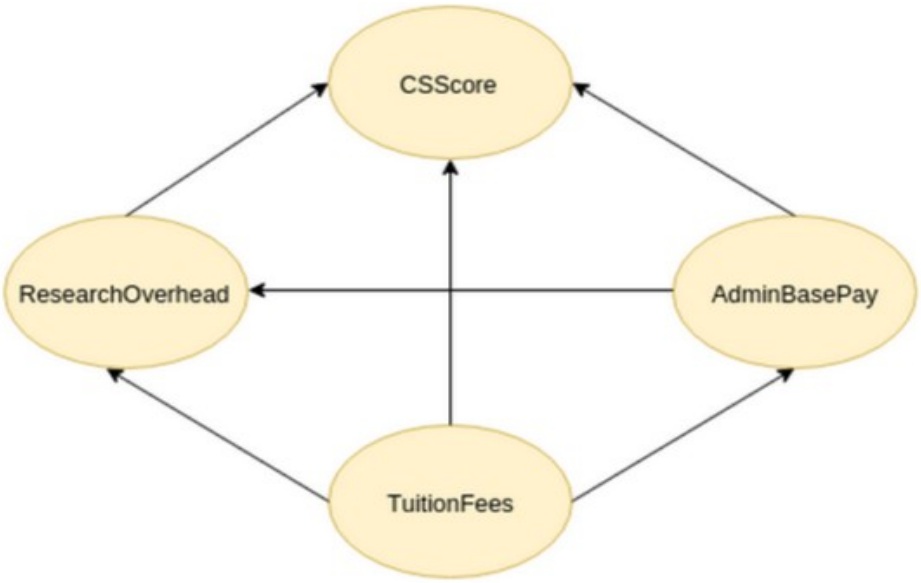
BnlogLikelihood: **-1304.1**

Bngraph:

0 0 0 0
1 0 0 0
1 1 0 0
1 1 1 0

Thus, the most optimal Bayeian Network has conditional probabilities,

P (CSScore | ResearchOverhead, AdminBasePay, TuitionFees)
P (ResearchOverhead | AdminBasePay, TuitionFees)
P (AdminBasePay | TuitionFees)
P (TuitionFees)

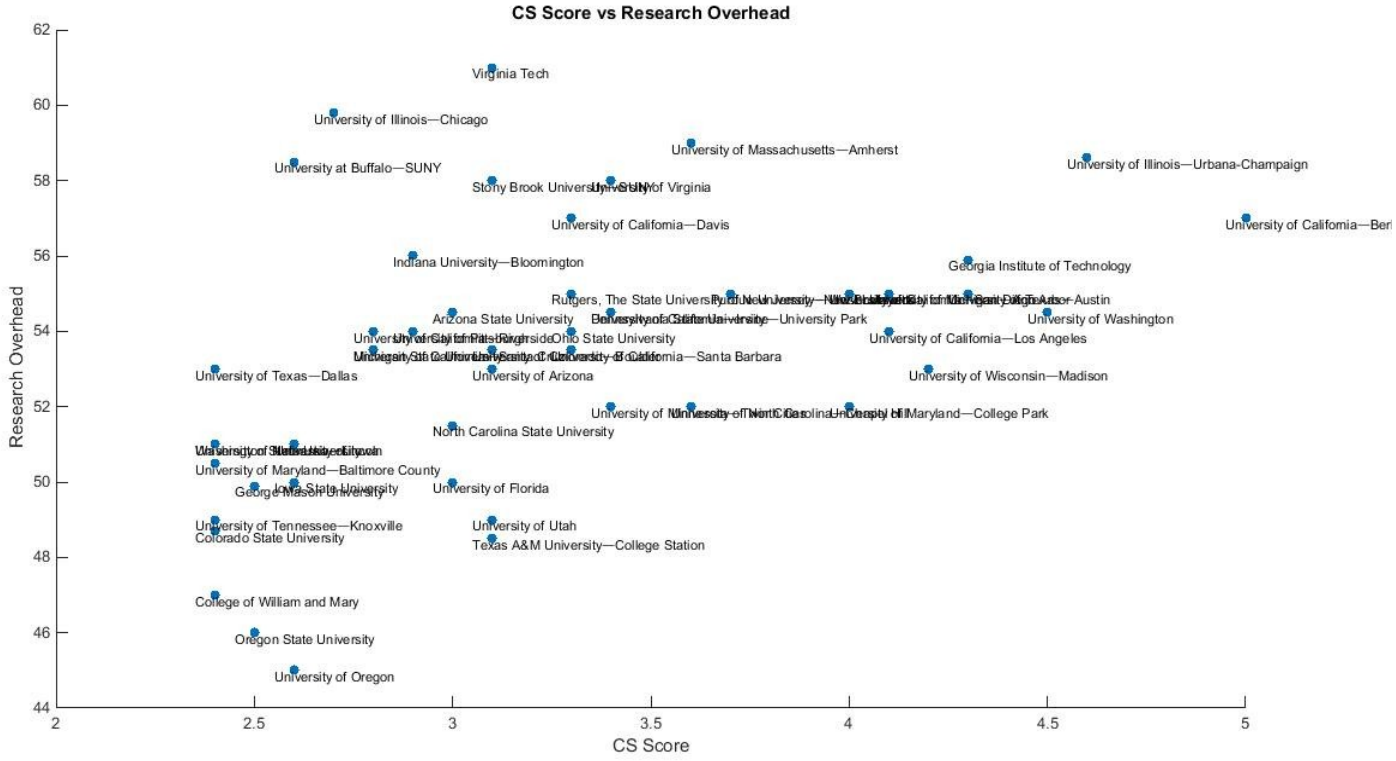


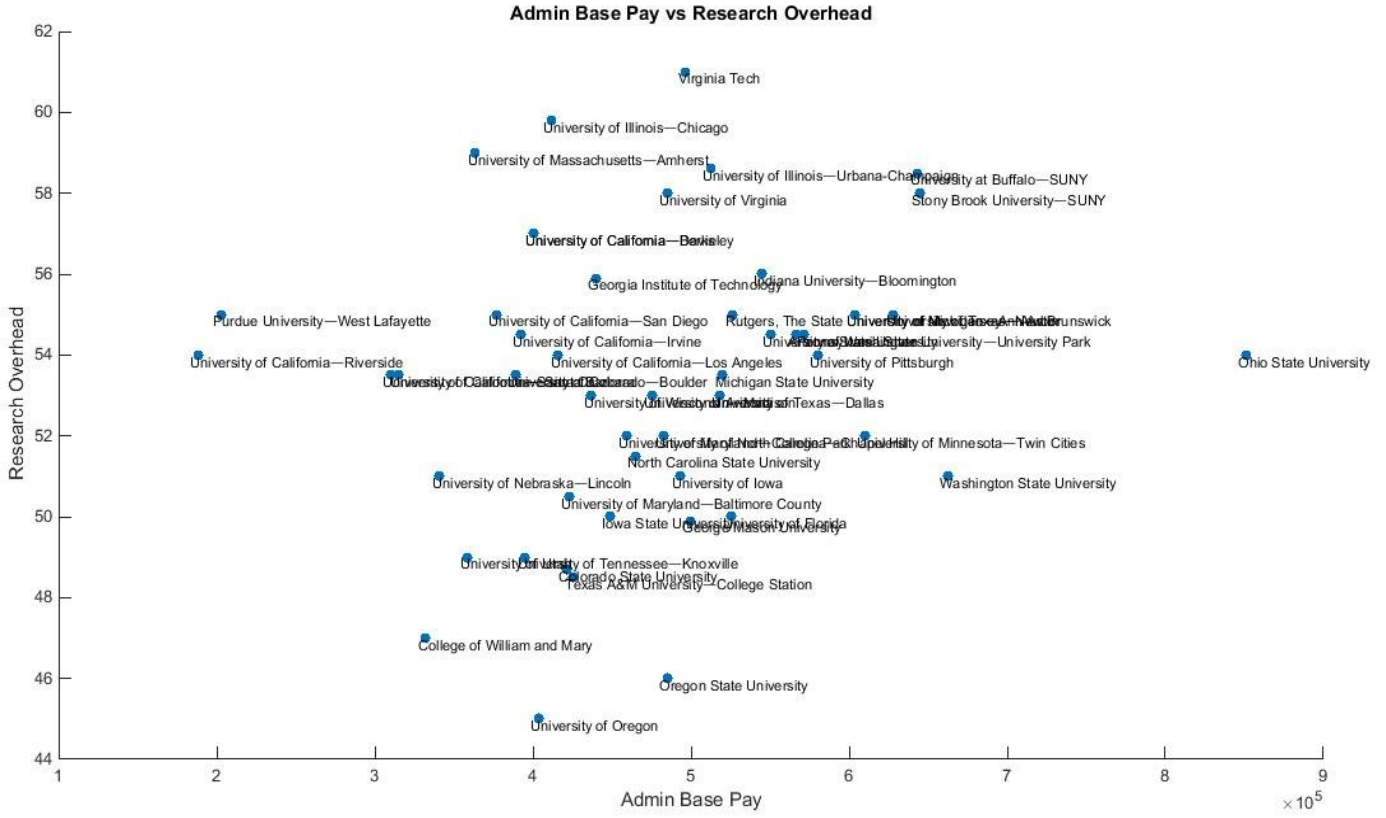
Bngraph

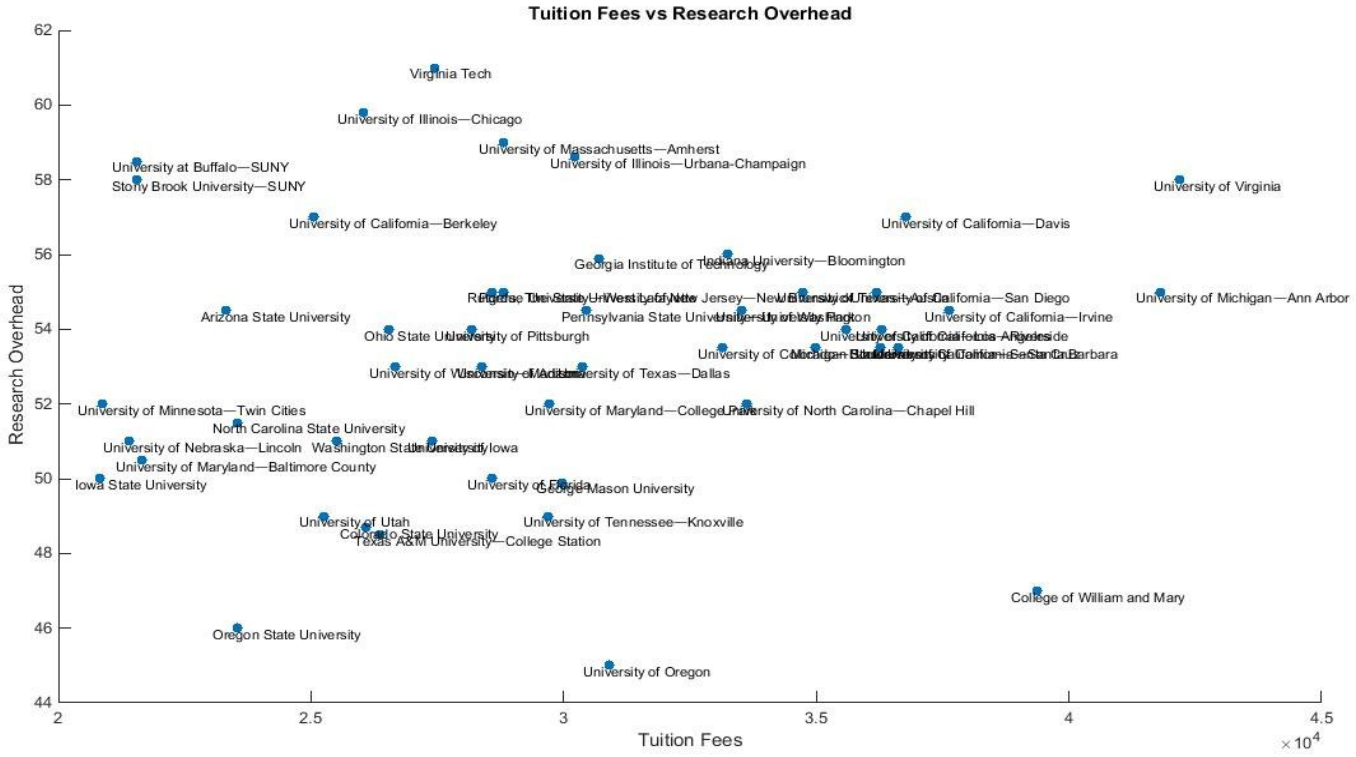
3. Conclusion:

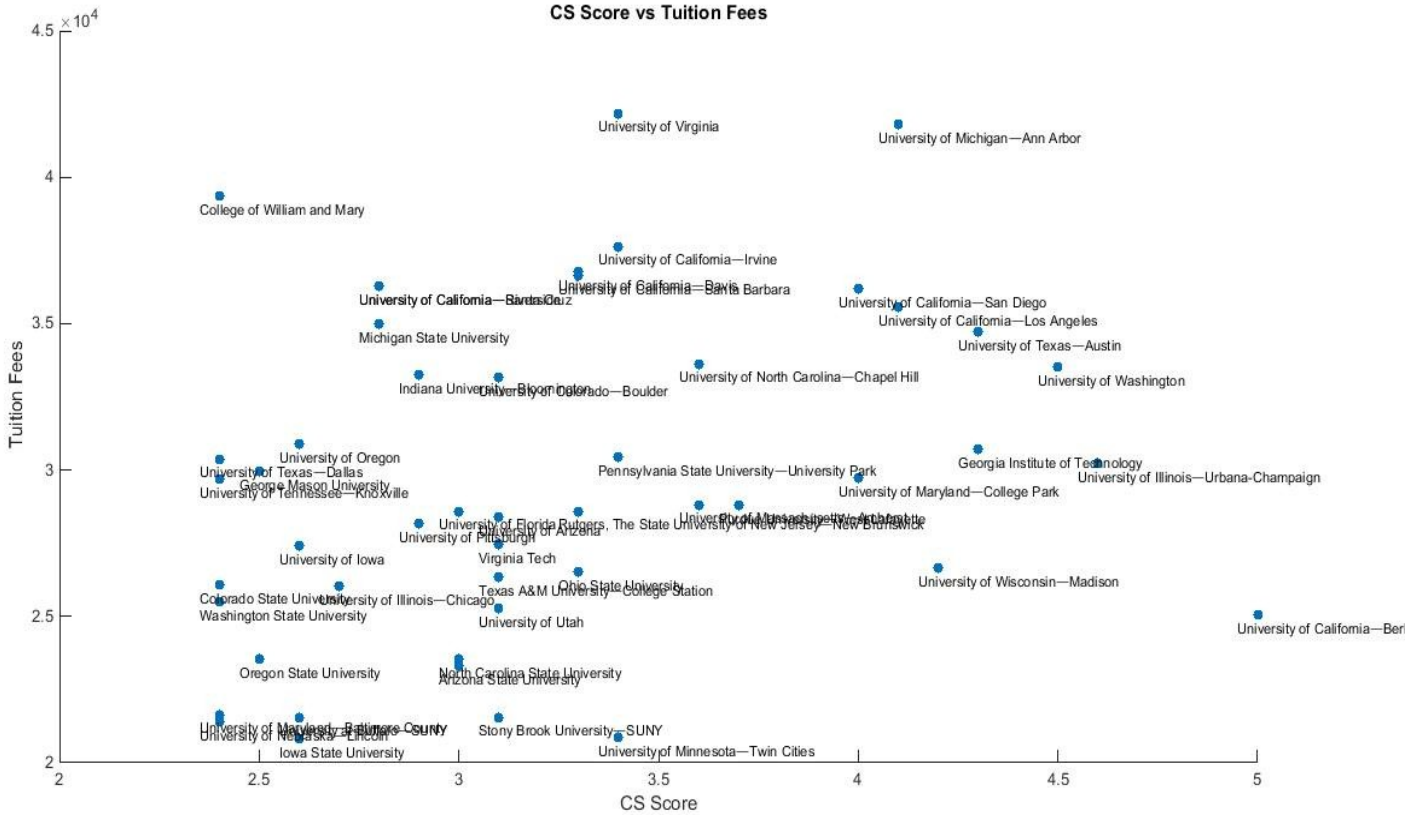
CSScore is dependent on ResearchOverhead, AdminBasePay, TuitionFees.
ResearchOverhead's is dependent on AdminBasePay and TuitionFees.
AdminBasePay is dependent on TuitionFees.

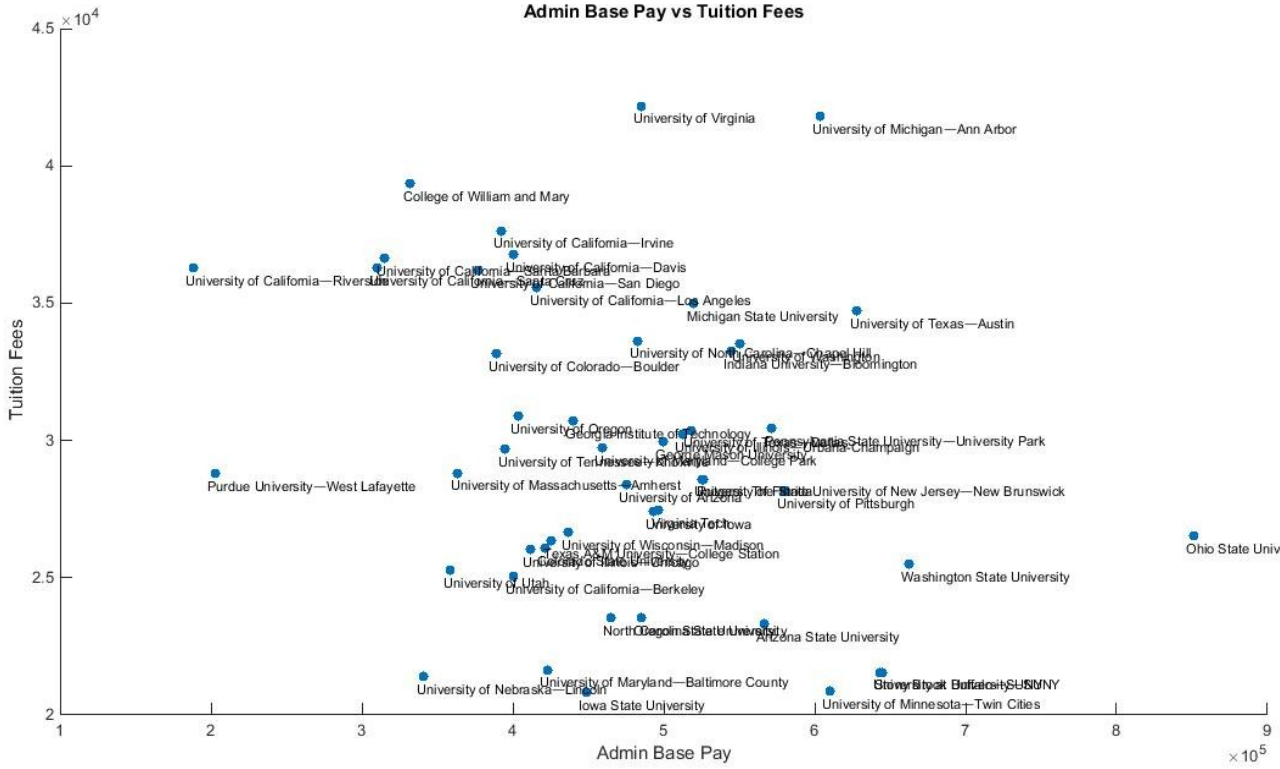
So, If a school has a higher tuition fees then AdminBasePay is better, if so then, ResearchOverhead is also higher and all these factors lead to a higher CSScore.

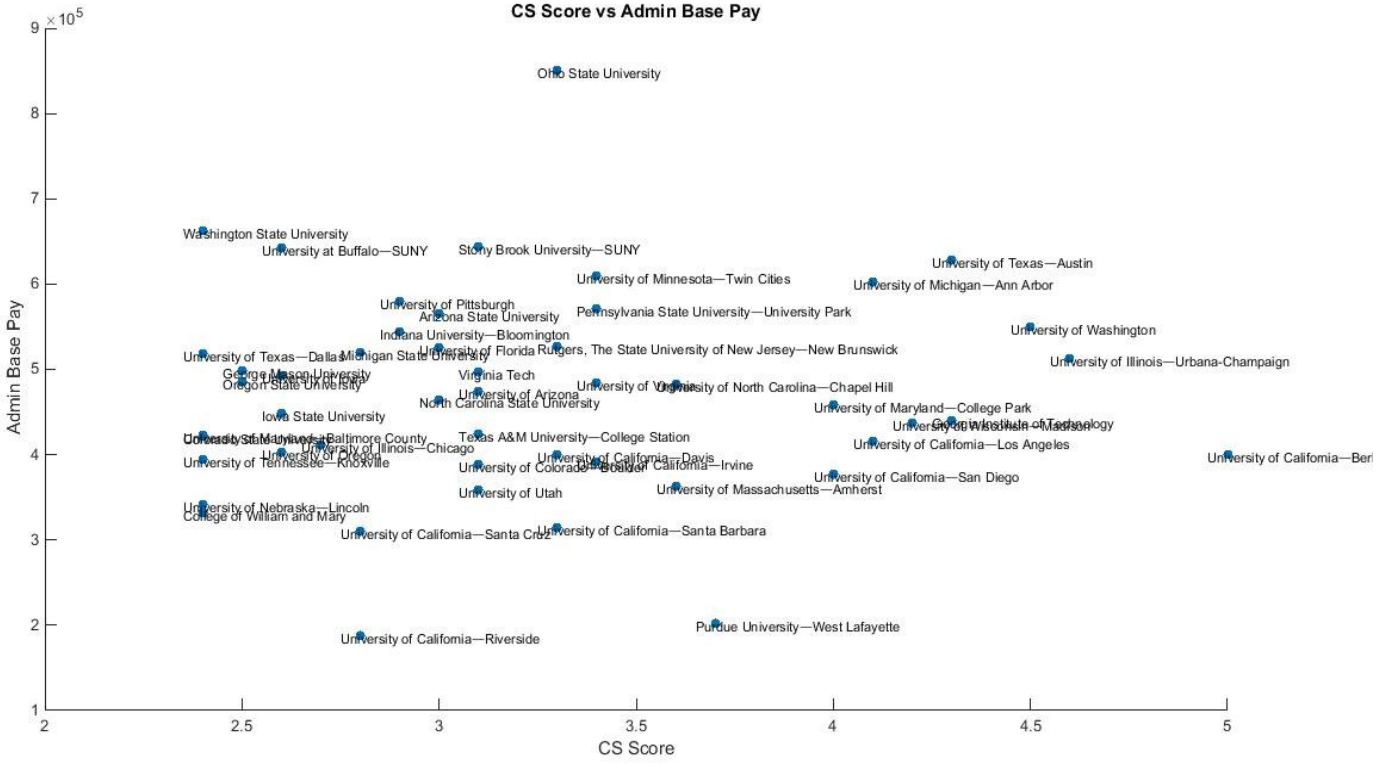

CS Score vs Research Overhead
Research Overhead
CS Score
Virginia Tech
University of Illinois—Chicago
University of Massachusetts—Amherst
University at Buffalo—SUNY
University of Illinois—Urbana-Champaign
Stony Brook University—SUNY
University of Virginia
University of California—Davis
Indiana University—Bloomington
Georgia Institute of Technology
Arizona State University
University of Washington
Ohio State University
University of California—Los Angeles
University of California—Santa Barbara
University of Texas—Dallas
University of Arizona
University of Wisconsin—Madison
University of Maryland—College Park
North Carolina State University
University of Maryland—Baltimore County
George Mason University
University of Florida
University of Tennessee—Knoxville
Colorado State University
University of Utah
Texas A&M University—College Station
College of William and Mary
Oregon State University
University of Oregon


Admin Base Pay vs Research Overhead
Research Overhead
Admin Base Pay
Virginia Tech
University of Illinois—Chicago
University of Massachusetts—Amherst
University of Illinois—Urbana-Champaign
University at Buffalo—SUNY
University of Virginia
Stony Brook University—SUNY
Georgia Institute of Technology
Indiana University—Bloomington
Purdue University—West Lafayette
University of California—San Diego
University of California—Irvine
University of California—Riverside
University of California—Los Angeles
University of Pittsburgh
Ohio State University
Michigan State University
University of Minnesota—Twin Cities
North Carolina State University
University of Nebraska—Lincoln
University of Iowa
Washington State University
University of Maryland—Baltimore County
George Mason University
University of Florida
University of Tennessee—Knoxville
Colorado State University
Texas A&M University—College Station
College of William and Mary
Oregon State University
University of Oregon


Tuition Fees vs Research Overhead
Research Overhead
Tuition Fees
×10^4
Virginia Tech
University of Illinois—Chicago
University of Massachusetts—Amherst
University at Buffalo—SUNY
University of Illinois—Urbana-Champaign
Stony Brook University—SUNY
University of Virginia
University of California—Berkeley
University of California—Davis
Georgia Institute of Technology
Indiana University—Bloomington
University of Michigan—Ann Arbor
Arizona State University
University of California—San Diego
Pennsylvania State University—University Park
University of California—Irvine
Ohio State University
University of Pittsburgh
University of Texas—Dallas
University of Minnesota—Twin Cities
University of Maryland—College Park
University of North Carolina—Chapel Hill
North Carolina State University
University of Nebraska—Lincoln
Iowa State University
University of Maryland—Baltimore County
University of Florida
George Mason University
University of Utah
University of Tennessee—Knoxville
Colorado State University
Texas A&M University—College Station
College of William and Mary
Oregon State University
University of Oregon


CS Score vs Tuition Fees
Tuition Fees
×10^4
CS Score
University of Virginia
University of Michigan—Ann Arbor
College of William and Mary
University of California—Irvine
University of California—Davis
University of California—Santa Barbara
University of California—San Diego
University of California—Los Angeles
Michigan State University
University of Texas—Austin
Indiana University—Bloomington
University of Colorado—Boulder
University of North Carolina—Chapel Hill
University of Washington
University of Oregon
University of Texas—Dallas
George Mason University
University of Tennessee—Knoxville
Pennsylvania State University—University Park
Georgia Institute of Technology
University of Illinois—Urbana-Champaign
University of Maryland—College Park
University of Florida
University of Pittsburgh
University of Iowa
Virginia Tech
University of Wisconsin—Madison
Colorado State University
University of Illinois—Chicago
Texas A&M University—College Station
Washington State University
University of Utah
Oregon State University
North Carolina State University
Arizona State University
Stony Brook University—SUNY
Iowa State University
University of Minnesota—Twin Cities


Admin Base Pay vs Tuition Fees
Tuition Fees
Admin Base Pay
×10^4
×10^5
University of Virginia
University of Michigan—Ann Arbor
College of William and Mary
University of California—Irvine
University of California—Davis
University of California—San Diego
University of California—Los Angeles
Michigan State University
University of Texas—Austin
University of Colorado—Boulder
Indiana University—Bloomington
Purdue University—West Lafayette
University of Massachusetts—Amherst
University of Pittsburgh
Ohio State University
Washington State University
University of Utah
University of California—Berkeley
Arizona State University
University of Nebraska—Lincoln
University of Maryland—Baltimore County
Iowa State University
University of Minnesota—Twin Cities


CS Score vs Admin Base Pay
Admin Base Pay
CS Score
×10^5
Ohio State University
Washington State University
University at Buffalo—SUNY
Stony Brook University—SUNY
University of Texas—Austin
University of Minnesota—Twin Cities
University of Michigan—Ann Arbor
University of Pittsburgh
Arizona State University
Pennsylvania State University—University Park
University of Washington
Indiana University—Bloomington
University of Texas—Dallas
Michigan State University
University of Florida
Rutgers, The State University of New Jersey—New Brunswick
University of Illinois—Urbana-Champaign
George Mason University
Virginia Tech
Oregon State University
University of Arizona
University of North Carolina—Chapel Hill
North Carolina State University
University of Maryland—College Park
Iowa State University
University of Wisconsin—Madison
University of Illinois—Chicago
Texas A&M University—College Station
University of California—Los Angeles
University of California—Davis
University of Tennessee—Knoxville
University of Colorado—Boulder
University of California—Irvine
University of California—San Diego
University of Utah
University of Massachusetts—Amherst
University of Nebraska—Lincoln
College of William and Mary
University of California—Santa Cruz
University of California—Santa Barbara
Purdue University—West Lafayette
University of California—Riverside