Machine Learning CSE 574

Project 1

Team

Sunil Umasankar (UBITName = suniluma, personNumber = 50249002)

Prajna Bhandary (UBITName = prajnaga, personNumber = 50244304)

Abhishek Subramaniam (UBITName = a45, personNumber = 50244979)

Overview

This project intends to understand the given university data by the use of probabilistic methods. The data consists of 49 universities and their values to some common variables. Particularly, 4 variables were given importance. These are "CS Score, Research Overhead, Admin Base Pay, and Tuition Out State".

We started with importing the data using Pandas, a Python Library which helps us to get the data in the form of a table. This project has three tasks each with their own set of calculations and inference. These are discussed in detail below.

Task 1

Task 1 of this project is to find the mean, variance and standard deviation of the 4 variables.

To do this, we used NumPy functions on the data imported from Pandas to calculate the sample mean, variance and standard deviations of each of the variables.

We observed their values to be the following:

---

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

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

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

These values are in the same order as they were given in the overview. We can then use these basic values to compute more results in the further tasks.

Task 2

Task 2 of this project is to calculate the covariance and correlation matrix for the four variables. Scatter plot of the pairwise data should also be plotted.

Both the matrices can be calculated from inbuilt NumPy functions namely *numpy.cov()* and *numpy.corrcoef().* These functions require the input to matrix of the form 4*x where x is the number of sample data we have. In this case, the excel has 49 rows, so we need to pass a 4*49 matrix.

Thus, we converted the DataFrame in Pandas to a NumPy matrix and took the transpose to get the desired matrix. We then passed it to these functions to get two 4*4 matrices namely the covariance and correlation matrices.

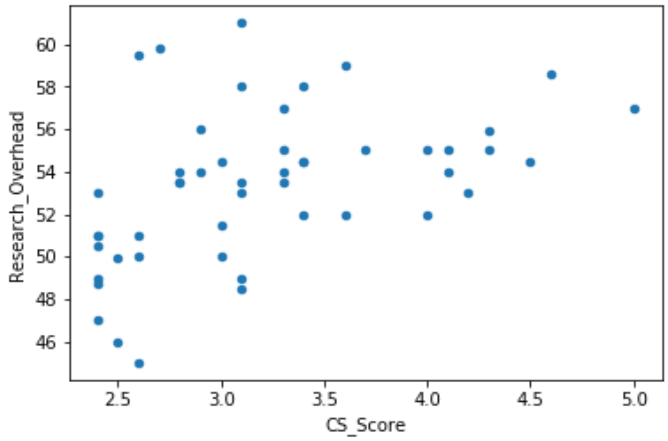
We observed their values to be the following:

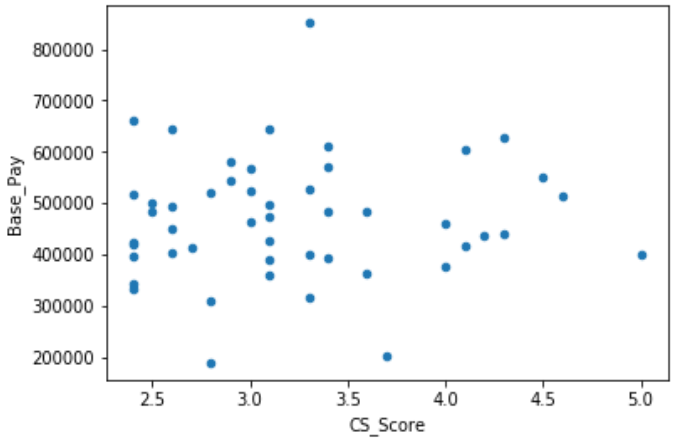
Covariance	**CS Score**	**Research Overhead**	**Admin Base Pay**	**Tuition Out State**
CS Score	0.457	1.106	3879.782	1058.480
Research Overhead	1.106	12.850	70279.376	2805.789
Admin Base Pay	3879.782	70279.376	14189720820.903	-163685641.258
Tuition Out State	1058.480	2805.789	-163685641.258	31367695.790

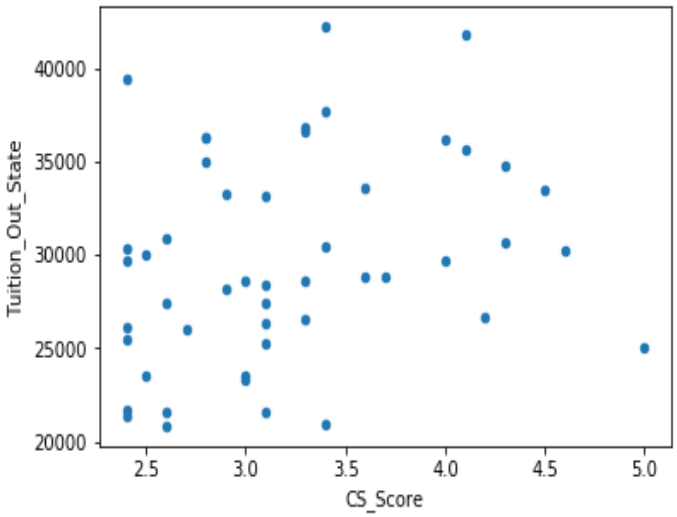
Correlation	**CS Score**	**Research Overhead**	**Admin Base Pay**	**Tuition Out State**
CS Score	1.000	0.456	0.048	0.279
Research Overhead	0.456	1.000	0.165	0.140
Admin Base Pay	0.048	0.165	1.000	-0.245
Tuition Out State	0.279	0.140	-0.245	1.000

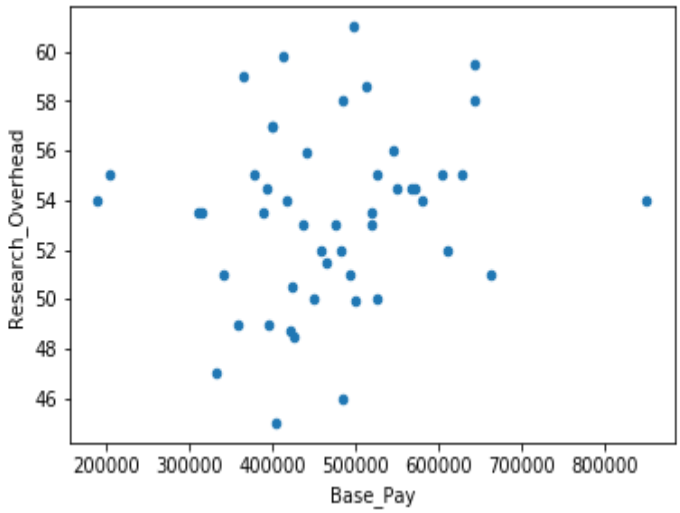
We also used Pandas scatterplot to plot a pairwise relation between each variable taken into consideration.

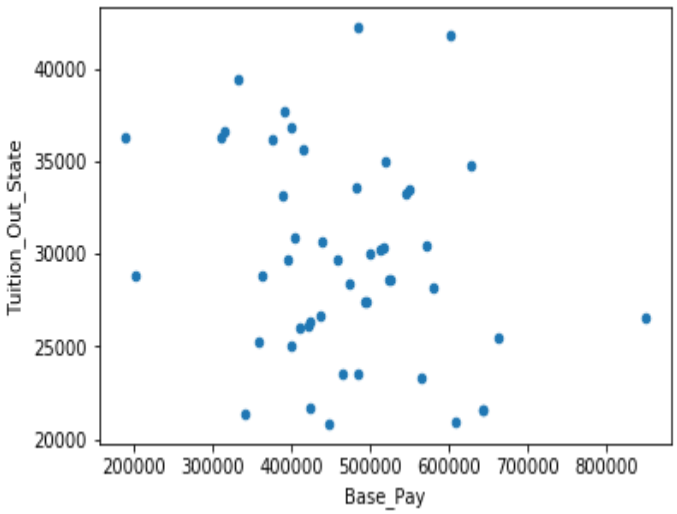
We observed the graphs to be the following:

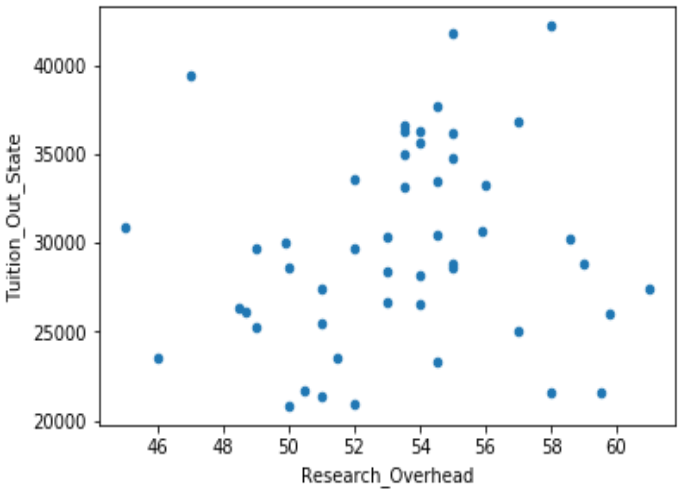












From the graphs and the correlation matrix, we can infer that the variables that are correlated the most are CS Score and Research Overhead, followed by CS Score and Tuition Out State, and the variables that are the least correlated are Research Overhead and Tuition Out State. Considering this from a logical point of view the data proves that, in general, the better the institution's rankings, the higher its research overhead and out of state tuition. Also, the amount of research overhead a university accrues, in general, has no relation with the amount of out of state tuition fee charged by the university.

Task 3:

Task 3 of this project is to calculate the loglikelihood for the variables.

To calculate log likelihood, we first assumed the variables to be independent. We then computed the probability distribution function (pdf) values for each variable present. Since the data consists of 49 samples, we got 49 pdf values for each variable. We multiplied these pdfs for each row to get a cumulative of 49 pdfs. Finally, we took log of each pdf and added them to get the log likelihood. This is the log likelihood assuming the variables are independent. We found its value to be –1315.099.

Now, we assumed the variables to be dependent on each other and calculated the multivariate log likelihood. For calculating this, we first computed the multivariate pdf from the formula given. For each sample we found its multivariate pdf and took log of the value. We added these log values to get the multivariate log likelihood and found its value to be –1304.778.

This proves that, the variables are dependent on each other because, the value of dependent log likelihood is higher than its independent counterpart.

logLikelihood = -1315.099

multivariatelogLikelihood = -1304.778

Code Output:

```
Group members

UBITName = suniluma
personNumber = 50249002

UBITName = a45
personNumber = 50244979

UBITName = prajnaga
personNumber = 50244304

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

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

sigma1 = 0.669
sigma2 = 3.548
sigma3 = 117898.832
sigma4 = 5543.243
```

```
covarianceMat =
[[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]]

correlationMat =
[[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]]

logLikelihood = -1315.099
MultivariatelogLikelihood = -1304.778
```