Project 1
Project Report

Pranav Sankhe (UBITname: pranavgi UB Person No: 50244956)

Varun Shijo (UBITname: varunshi UB Person No: 50244968)

Introduction

Using Python and libraries like pandas, numpy, scipy, scikit-learn to process the given dataset we could glean metrics like the mean, variance, standard deviation of variables.

We also analyze the significance of parameters like correlation, covariance, etc.

Also by calculating the log likelihood we can estimate the goodness of fit of the joint probability density of the four variables we are concerned with.

Results


```
Mean of CS Score (USNews) is 3.214
Mean of Research Overhead % is 53.386
Mean of Admin Base Pay$ is 469178.816
Mean of Tuition(out-state)$ is 29711.959
Variance of CS Score (USNews) is 0.448
Variance of Research Overhead % is 12.588
Variance of Admin Base Pay$ is 13900134681.701
Variance of Tuition(out-state)$ is 30727538.733
Standard Deviation of CS Score (USNews) is 0.676
Standard Deviation of Research Overhead % is 3.585
Standard Deviation of Admin Base Pay$ is 119120.615
Standard Deviation of Tuition(out-state)$ is 5600.687


Covariance Matrix:


                    CS Score (USNews)  Research Overhead %  Admin Base Pay$  \
CS Score (USNews)                0.46                 1.11     3.879780e+03
Research Overhead %              1.11                12.85     7.027938e+04
Admin Base Pay$               3879.78             70279.38     1.418972e+10
Tuition(out-state)$           1058.48              2805.79    -1.636856e+08

                    Tuition(out-state)$
CS Score (USNews)          1.058480e+03
Research Overhead %        2.805790e+03
Admin Base Pay$           -1.636856e+08
Tuition(out-state)$        3.136770e+07


Correlation Matrix:


                    CS Score (USNews)  Research Overhead %  Admin Base Pay$  \
CS Score (USNews)                1.00                 0.46             0.05
Research Overhead %              0.46                 1.00             0.16
Admin Base Pay$                  0.05                 0.16             1.00
Tuition(out-state)$              0.28                 0.14            -0.25

                    Tuition(out-state)$
CS Score (USNews)                  0.28
Research Overhead %                0.14
Admin Base Pay$                   -0.25
Tuition(out-state)$                1.00
Using multivariate equation
-1262.34
Using independent variables
-1315.12
```
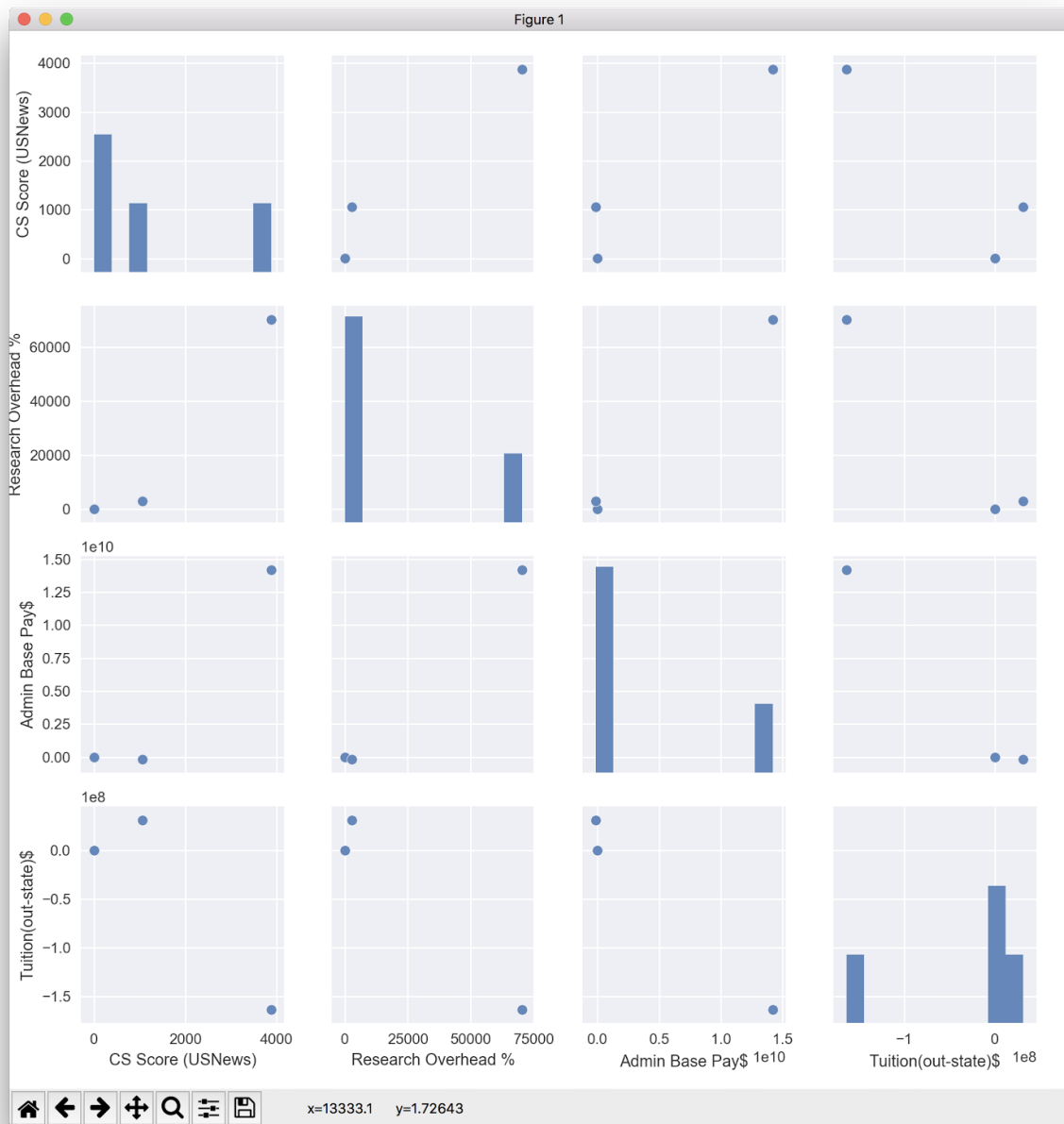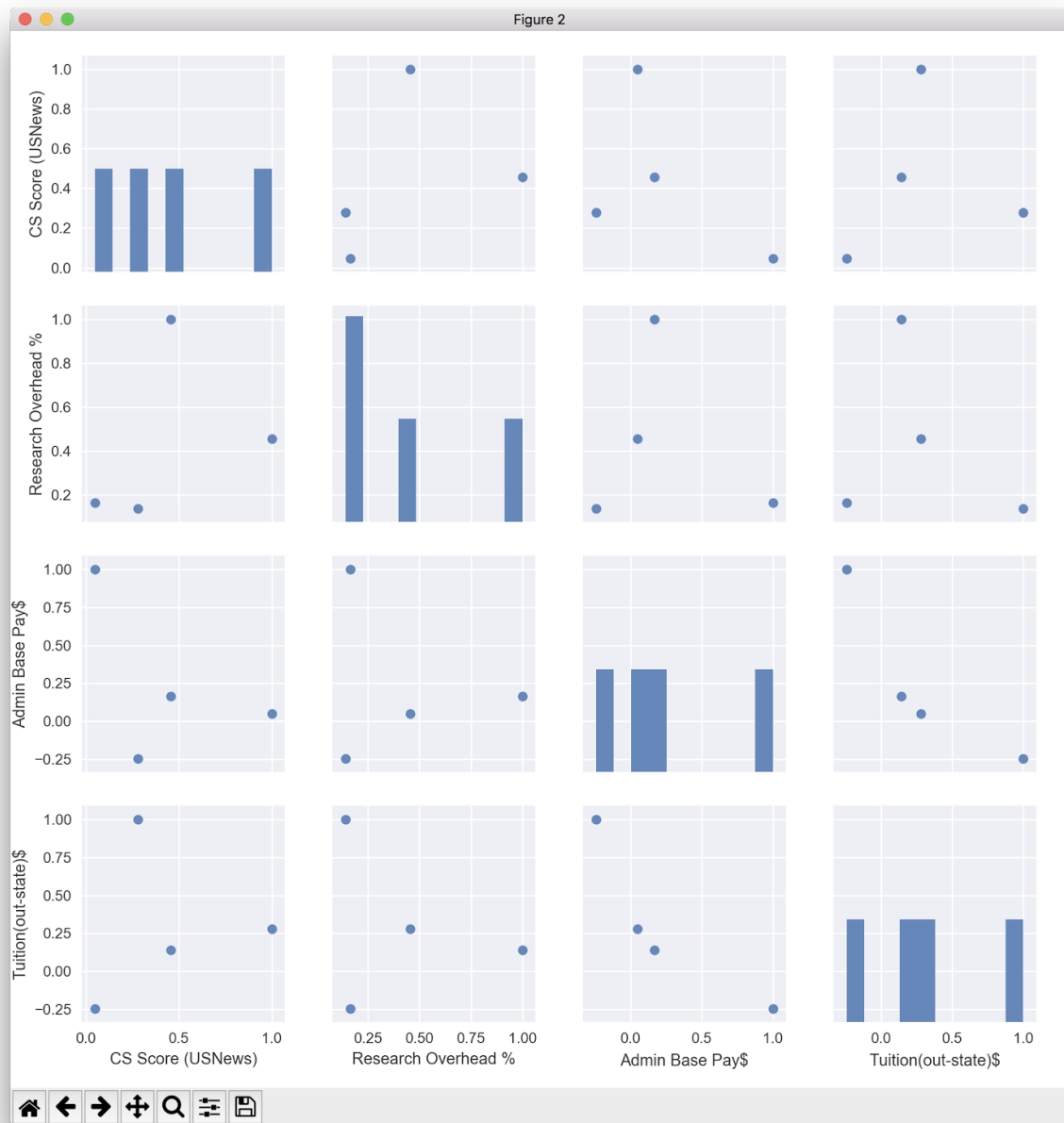
Results of the script

Seaborn visualization of Covariance Matrix

Pairwise Plot of Correlation Matrix

Interpretation

1. The obtained **mean** value of CS score signify that the Universities scoring greater than 3.3 can be safely categorized as good universities for CS course.

2. Similarly, universities with research overhead % greater than the **mean** value of 53 can be regarded as research oriented universities. Also, the tuition of universities lower than $29,711 are affordable. Students can benefit with such vital information, so as to decide which universities to choose during their university application process.

3. **Variance** helps student gauge the impact of various criteria while looking for a suitable university. As the low variation in CS ranking and research overhead coupled with much greater variance in tuition suggest that the student must aim for a university that provides education at lower tuitions as most of the most of the universities are abreast in rest of the criteria.

4. **Standard deviation** on the other hand would help an admin choose which university would pay him the best. Universities tend to show a great deviation from the mean payment, thus smartly choosing the university would yield the best payment.

5. We can observe a high **correlation** between the values of CS Score and Research overhead indicating that universities with high CS score tend to high research spendings.

6. **Covariance matrix** shows us that the universities that invest heavily on research tend to have higher admin base pay variations. Thus, someone seeking employment at such university must consider this deductions.

7. **Log-likelihood** value, is used to evaluate goodness-of-fit for data models. In the case of our dataset the negative value denotes that we have a large dataset. For further use of the likelihood we must apply it to different representative functions and the highest likelihood would signify best fit.

Citations

https://pandas.pydata.org/pandas-docs/

https://docs.scipy.org/doc/numpy-1.13.0/reference/

https://seaborn.pydata.org/

http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf

https://scipy.github.io/devdocs/