# Goal

The aim of this exercise is to explore Probability Distributions and Bayesian Networks. The data-set in use is data of U.S. universities obtained from :

X1 (CS ranking score) = <http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/>

X2 (research Overload) = Portion of research grants obtained from each university's website

X3 (Admin base salary) = http://chronicle.com/factfile/ec-2015/#id=table_public_2014

X4 (Out of state tuition) = http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/

X5 (Number of CS grads in Fall 2015) = Mostly from each college's website

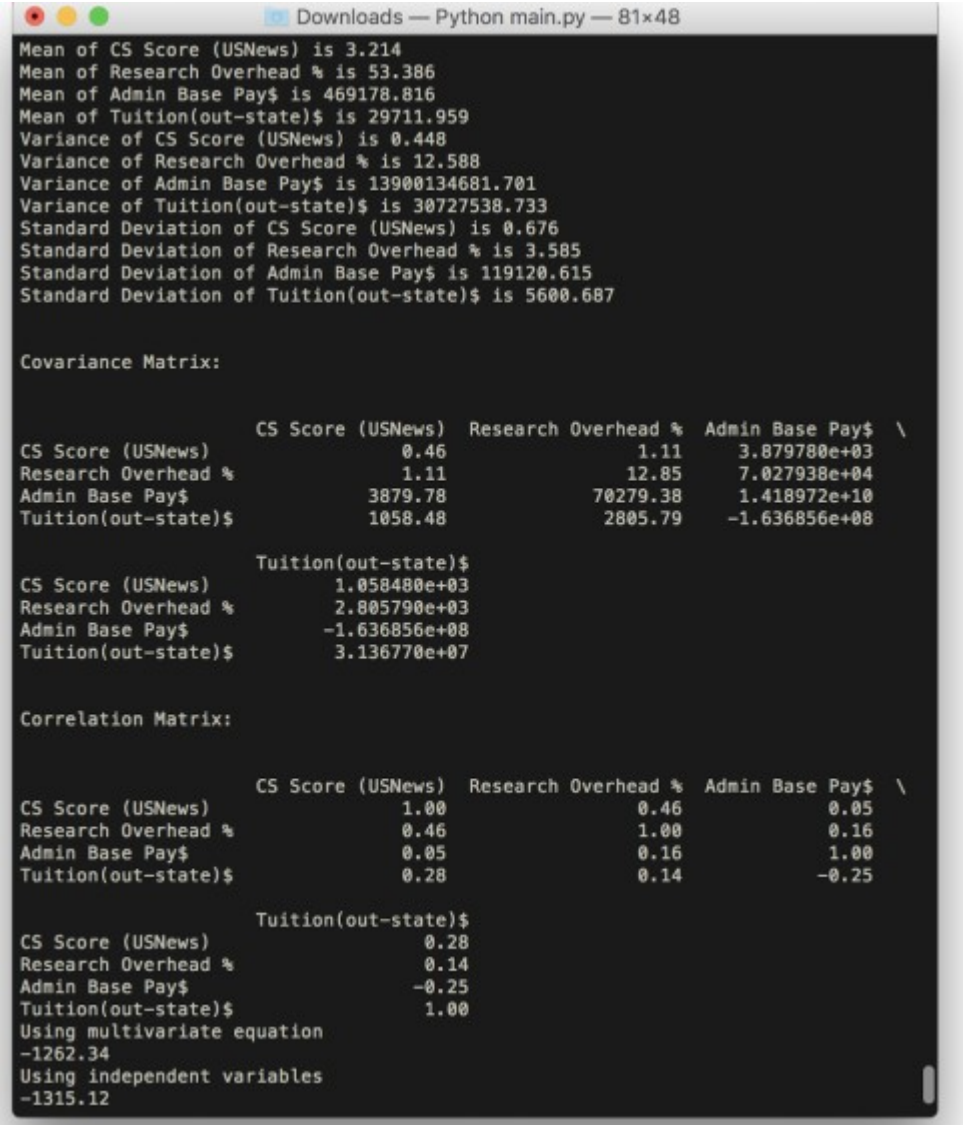
It is a multivariate data-set with four variables and a fifth with some missing values. The missing values are provided separately in UniversityData.xls.

We will use Bayesian network tp determine some interesting conditional probabilities.

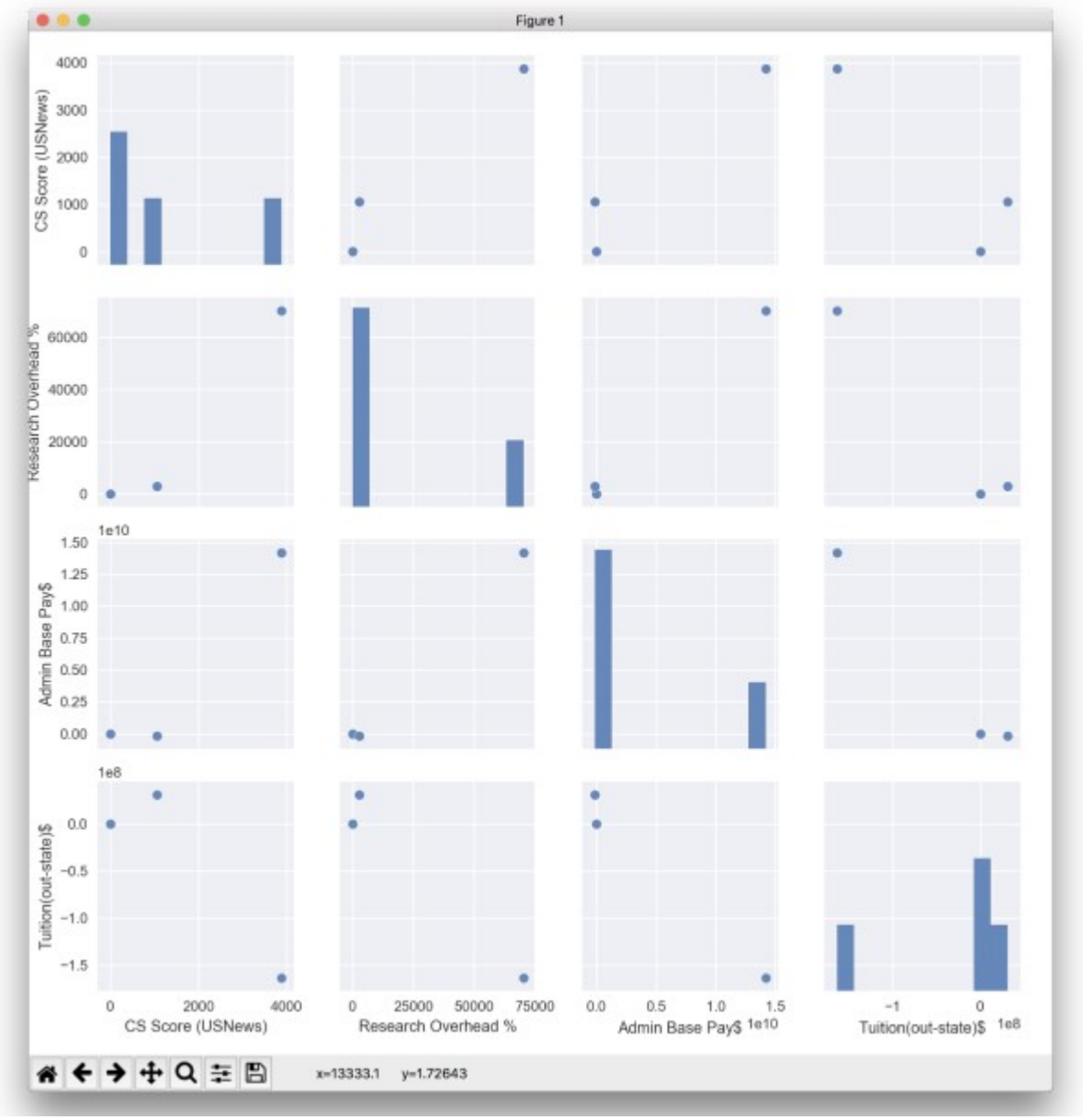
# Introduction

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.

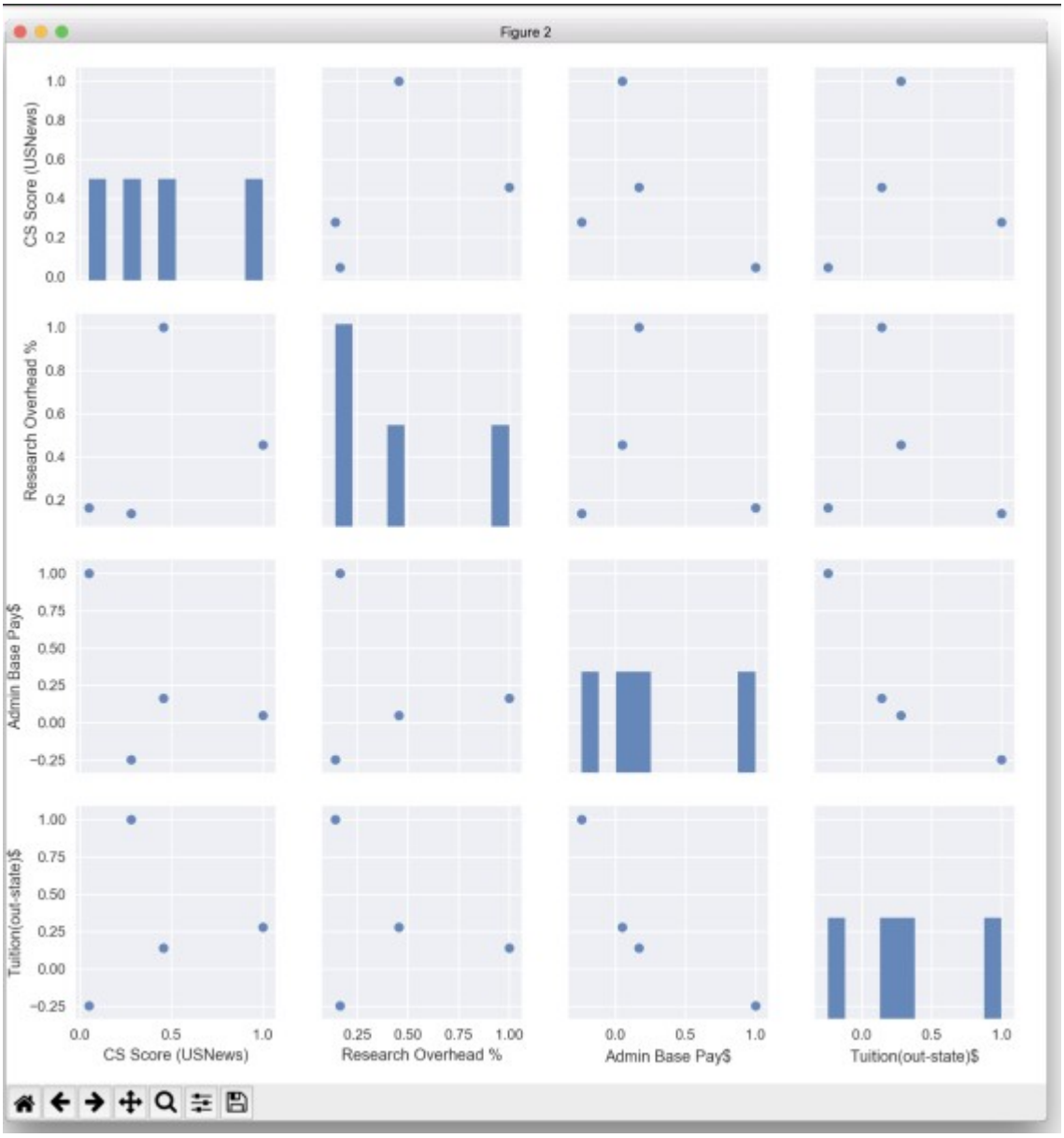
# Results




*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.