INTRODUCTION TO MACHINE LEARNING

PROJECT 1

PROJECT REPORT

GROUP MEMBERS

NITHISH SHOKEEN (UB Person Number: 50247681)

MAHALAKSHMI PADMA SRI HARSHA MADDU (UB Person Number: 50246769)

CHARANYA SUDHARSANAN (UB person Number: 50245956)

The following **Libraries** are called to perform various functions in this project

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- from pandas.tools.plotting
- import scatter_matrix
- from scipy import stats, integrate

Question 1:

We first created a dataframe and imported the excel dataset 'University data'. We created another dataframe 'df1' to import the first four columns namely '(CS Score, Research Overhead, Admin Base Pay, Tuition)'

We arrived at the mean, variance and standard deviation by calling the below functions of the pandas library.

Mean = mu1 = dataframe[column_ name].mean()

Variance = var1 = dataframe[column_ name].var()

Standard deviation = sigma1 = dataframe[column_ name].std()

Output 1:

mu1 = 3.214

mu2 = 53.386

mu3 = 469178.816

mu4 = 29711.959

var1 = 0.457

var2 = 12.85

var3 = 14189720820.903

var4 = 31367695.79

sigma1 = 0.676

sigma2 = 3.585

sigma3 = 119120.615

sigma4 = 5600.687

Question 2

The covariance and correlation matrix is computed using below functions:

Df.cov() -> To compute the covariance matrix

Df.corr() -> To compute the correlation matrix

The Resultant **Covariance matrix** obtained is as below (**Output Snap Shot 2**)

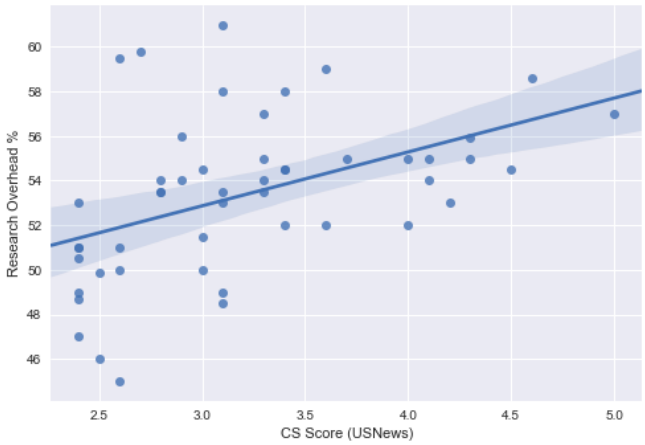
```
                      CS Score (USNews)  Research Overhead %  Admin Base Pay$  \
CS Score (USNews)              0.457500             1.105625     3.879782e+03
Research Overhead %            1.105625            12.850417     7.027938e+04
Admin Base Pay$             3879.781845         70279.376488     1.418972e+10
Tuition(out-state)$         1058.479762          2805.788988    -1.636856e+08

                      Tuition(out-state)$
CS Score (USNews)           1.058480e+03
Research Overhead %         2.805789e+03
Admin Base Pay$            -1.636856e+08
Tuition(out-state)$         3.136770e+07
```

The Resultant **Correlation Matrix** of the four columns are as follows (**Output Snap Shot 3**):

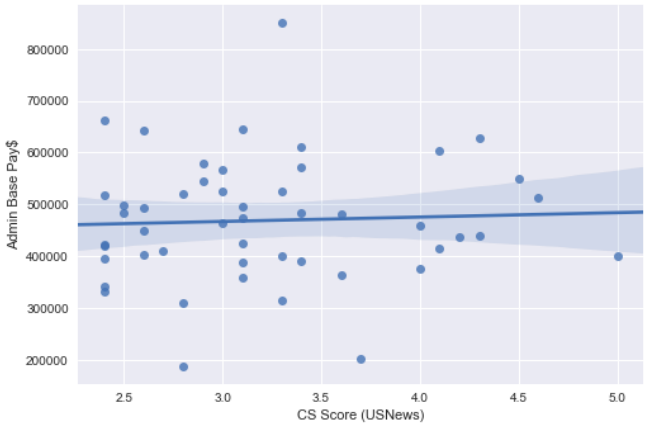
```
                        CS Score (USNews)  Research Overhead %  Admin Base Pay$  \
CS Score (USNews)                1.000000             0.455988         0.048153
Research Overhead %              0.455988             1.000000         0.164582
Admin Base Pay$                  0.048153             0.164582         1.000000
Tuition(out-state)$              0.279412             0.139751        -0.245348

                        Tuition(out-state)$
CS Score (USNews)                 0.279412
Research Overhead %               0.139751
Admin Base Pay$                  -0.245348
Tuition(out-state)$               1.000000
```

Scatter Plots for each combination of the variables:

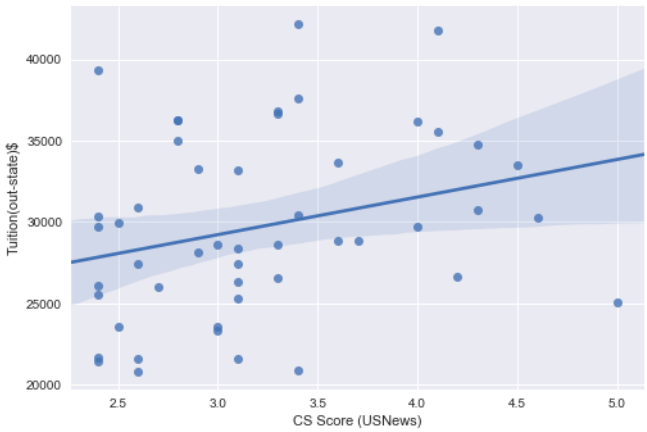
Scatter Plot for CS Score (US News) VS Research Overhead %:



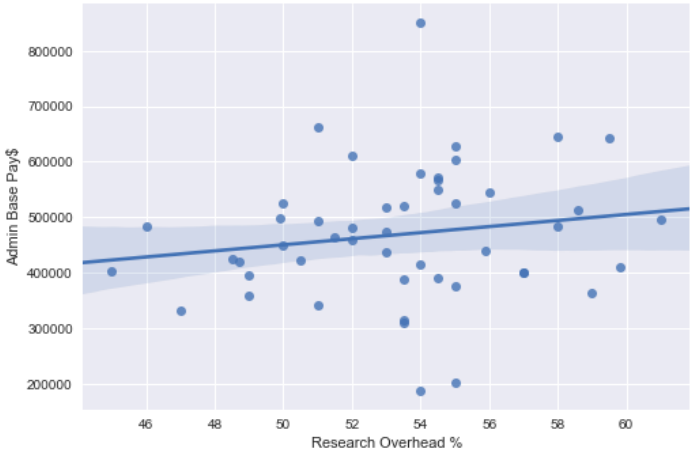
Scatter Plot for CS Score (US News) VS Admin Base Pay$:



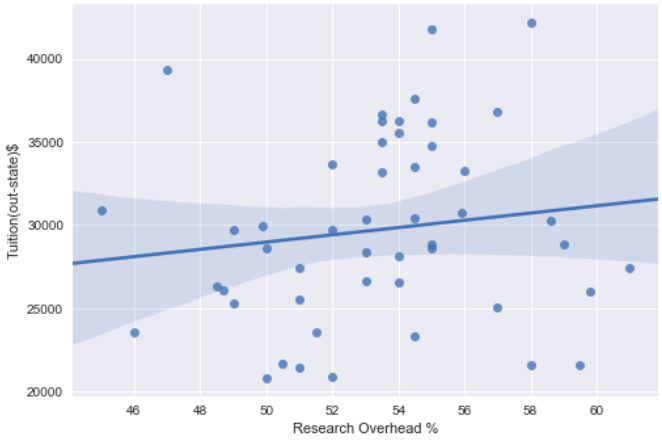
Scatter Plot for CS Score (USNews) % VS Tuition(out-state)$:



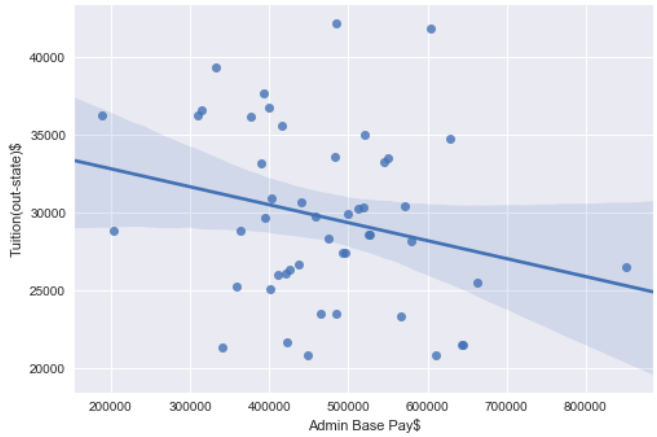
Scatter Plot for Research Overhead % VS Admin Base Pay$:



Scatter Plot for Research Overhead % VS Tuition(out-state)$:



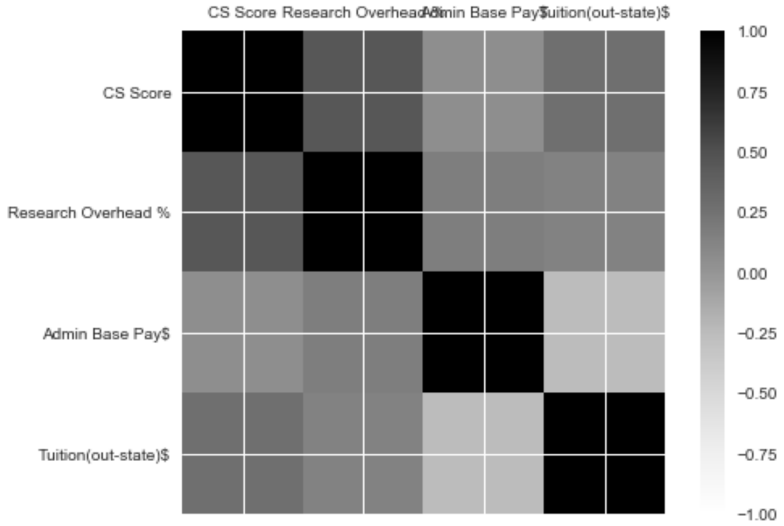
Scatter Plot for Admin Base Pay$ VS Tuition(out-state)$:



To Create a **ScatterMatrix**, we've used the functions matshow(). A colorbar has been sub-plotted to show the differences in correlation via different colors

The following ScatterMatrix has been created to show the correlation between four variables (CS Score, Research Overhead, Admin Base Pay, Tuition):

Output Snapshot 4(Scatter Matrix to show correlation between the variable pairs):



From the above Correlation Matrix as well as the ScatterMatrix it is evident that the most correlated variable pair is (CS Score (US News), Research Overhead %) and the Least Correlated Variable pair is (CS Score (US News), Admin Base Pay).

Following Interpretations have been made from the Correlation Matrix as well as the Scatter Plot:

- The Correlation Matrix shows how close two variables is to being linear.
- The Main diagonal has a value 1 i.e., they are perfectly positively correlated as the variable CS Score has been plotted against itself as so on.
- The sign tells us whether they are positively correlated or negatively Correlated.
- The order of Correlation between various variable pairs are as follows (Descending from highly correlated to least correlated):
    - {CS Score, Research Overhead}
    - {CS Score, Tuition}
    - {Admin Base Pay, Tuition}
    - {Admin Base Pay, Research Overhead}
    - {Tuition, Research overhead}
    - {CS Score, Admin Base Pay}

Question 3:

a. **Loglikelihood (Univariate):**

Computed using the function: **stats.norm.pdf()**

Normal pdf was computed for every cell in a given variable using the above function . Hence the function was called within a loop 49 times and 49 normal pdfs were calculated for each variable . All the 49 normal pdfs were multiplied to get the normal pdf of the corresponding variable .

Math.Log() is called to compute the logarithmic value of normal pdfs. Finally four values one corresponding to each of the variable is arrived at.

The loglikelihood is calculated by adding up the four values obtained.

Output Snapshot 5:

```
logLikelihood(Univariate) = -1315.119482706607
```

b. **Log Likelihood(Multi Variate) :**

Computed using the function **scipy.stats.multivariate_normal.pdf().**

Parameters passed :

df1.iloc[itr,:] → Row consisting of all the four variables

Mean vector → Consists of mu1,mu2,mu3,mu4

Covariance matrix

After computing multivariate normal pdf , logarithmic value is calculated using Math.log(). Each of the 49 log values are added up to find the log likelihood.

Output Snapshot 6:

```
logLikelihood (Multivariate) = -1262.327200282979
```