Midterm 1

This exam is open-book. You may use your scripts, textbook, notes, homework assignments, Python. You may NOT use anything else including search engines, on-line chatting, phones, etc. You may NOT discuss with anyone.

Problem 1

Data: data.txt
Script: P1.py
Type your name in the script to agree with the UNC Honor Code Pledge.

To investigate weight loss programs, 500 volunteers participated in 5 different fitness programs (100 people per program), each lasting for 12 months. There were female and male participants in each program. The heights and weights were measured at the beginning the 12 month period. The weights were measured again at the end of each month.

The data is stored in text file data.txt, containing 500 lines. The first 100 lines are the participants from program 1; the second 100 lines are from program 2; …; the last 100 lines are from program 5.

The first column is the gender of each participant (0: female and 1: male). The second column is the height in meters. The third column is the weight in kilograms measured before the program. The next 12 columns are the weights in kilograms measured after each of the 12 months.

Question 1. (15 points) How many female participants were there in this investigation?

Question 2. (15 points) Plot the histogram of female heights with 10 bins using `pylab.hist()` function.

Question 3. (20 points) The body mass index (BMI) is a measure for human body shape. BMI is defined as the individual's body mass (kg) divided by the square of their height (m).

$$BMI = \frac{mass}{height^2} \quad \left(\frac{\mathrm{kg}}{\mathrm{m}^2}\right)$$

For each program calculate the average BMI of the 100 participants for each month (no need to print them out). You will need to use the last 13 columns in the data. Plot five average-BMI-to-month curves for the five programs in a figure.

Problem 2

Data: Carolina.csv and Opponent.csv
Script: P2.py
Type your name in the script to agree with the UNC Honor Code Pledge.

In this problem, you will be analyzing data from recent Carolina Men's basketball games (23 games). The file Carolina.csv contains game data for Carolina team, and Opponent.csv contains game data for Carolina's opponents. The first row specifies the data name for each column. The next 23 rows are from the 23 games. The same rows in the two files correspond to the same game. You can view the data by Excel (with better column alignment) or by any text editors.

Question 1. (20 points) Compute scores Carolina and Opponents earned for the recent 23 games. Then create two 1-D arrays called `CarolinaScore` and `OpponentScore` to store the scores for Carolina and Opponents, respectively. Plot score-to-game curves for Carolina and Opponents in one figure. Use blue color for Carolina and green for Opponents.
Hint: Game scores = 2*two-point made + 3*three-point made + free-throw made.

Question 2. (10 points) Create a 1-D array named `CarolinaVictory` and store 1 to represent Carolina victory and -1 for Carolina defeat. Print out the array `CarolinaVictory`.
Hint: When `CarolinaScore` is higher than `OpponentScore`, it's a Carolina victory. There is no tie game in college basketball rules.

Question 3. (20 points) Correlation coefficient can be used to measure the relationship between two 1-D arrays with the same size. It is a scalar value between -1 and 1, inclusively. Higher correlation coefficient indicates better correlation. The correlation coefficient C between two 1-D arrays X and Y is defined as

$$C(X,Y) = \frac{\sum_{i=1}^{n}[(X_i-\bar{X})\cdot(Y_i-\bar{Y})]}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

where X_i is the i$^{\text{th}}$ value in array X, and Y_i is the i$^{\text{th}}$ value in array Y;
$\bar{X}$ is the mean value in array X, and $\bar{Y}$ is the mean value in array Y;
n is the size of the 1-D array.

Print out the correlation coefficients between the array `CarolinaVictory` (X) and the following five *feature* arrays (Y). The pound sign # denotes "the number of".

0. #2-point made by Carolina
1. #3-point made by Carolina
2. Offensive power: #Offensive rebound by Carolina – #Defensive rebound by Opponent
3. Defensive power: #Defensive rebound by Carolina – #Offensive rebound by Opponent
4. #Assist by Carolina

Determine which feature among the five listed above is most important (having the highest correlation coefficient) to Carolina victory. Print out the feature index (0-4).

Hint: You can compute the correlation coefficients directly from the formula above **(5 bonus points)**; or, alternatively, you can use NumPy function `corrcoef`. The output of `corrcoef` function is a 2-D array (2-by-2). Use the value at the first row and the second column in the 2-D array as the final correlation coefficient.