Assignment 5
Due: Wednesday, October 2
6.5 pts

This assignment will use a file called 'Assignment5.csv' for its input. There is data for 10 patients who either have a disease ('D') or do not ('H' for healthy). Fifty genes were measure for each of these patients. These measurements are either 0 or 1 indicating that the gene is not expressed or is expressed in the patient. Are ultimate question is to identify a gene that can classify a patient as 'D' or 'H'.

1. **(0.5 pts)** Given a measurement of 0 or 1 for 50 genes, how many different profiles (variations of 0/1) for these genes are possible? (Show how you arrived at this number.)
1.1258999e+15
Given that there are only 2 options for each variable and there are 50 variables the number of possibilities would be 2^50th

2. **(2 pts)** Let's consider just the first 6 patients in the data sheet. Three have the disease and three are healthy. Is there a ***single gene*** that can distinguish these two cohorts? Write Python code to read in the spreadsheet and identify gene(s) that are expressed in one cohort (meaning that all members of the cohort have the gene expressed [1]) and not expressed in the other cohort. Write to the screen the following output for each single gene that can distinguish these two cohorts:

```
Gene 100: Disease 1; Healthy 0
Gene 209: Disease 0; Healthy 1
```

This example output would tell me that Gene 100 is expressed in individuals <u>with</u> the disease and is not expressed in individuals <u>without</u> the disease, and Gene 209 is expressed in individuals <u>without</u> the disease and is expressed in individuals <u>with</u> the disease. (Submit Python code a file named YOURLASTNAME_5.2.py.)

3. **(1 pt)** Using the genes identified in question 2, test your hypothesis that this gene(s) distinguishes the two cohorts by considering the data from Patients 7-10. How good of a *classifier* is this gene(s)? Report your findings using the same format as #2. Remember, you need only check the genes identified in question 2. (Submit Python code a file named YOURLASTNAME_5.3.py.)

4. **(0.5 pts)** You've come to learn that your results for the gene(s) identified in #3 are in fact an artifact of bad data entry! Yikes! So there is no single gene that can distinguish the two cohorts. Remove this gene(s) from your data set. Now, another research group has published a paper saying that this disease is the result of up to 3 different genes meaning that if you have expression for Gene A or Gene B or Gene C, you have the disease. Given the data set that we have, how many different combinations of 3 genes are there? (Show how you arrived at this number.)

49 choose 3 or 18424

5. **(2.5 pts)** Using the full data set (all 10 patients) identify Gene triplets that distinguish the two cohorts. Note, while each patient in the disease cohort will have expression for Gene A or Gene B or Gene C, the individuals of the healthy cohort can have expression of one or more of these genes, but unlike the disease cohort, all of the healthy cohort will not express one or more of these genes. Write to the screen the number of gene triplets found to distinguish the

two cohorts, followed by the gene triplets and their expression in disease/healthy. For example:

```
55
Gene 100, 101, 103: Disease 1; Healthy 0
Gene 209, 210, 233: Disease 0; Healthy 1
```

Submit Python code a file named YOURLASTNAME_5.5.py. *Also include in a comment block at the beginning of your code a brief description of the approach you took (0.5 pts of total points).*