Lab #5
Data Science I

These problems are adapted from the textbook An Introduction to Statistical Learning

1. In this problem, you will generate simulated data, and then perform K-means clustering on the data.
    a. Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.
    b. Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels (Hint: use pd.crosstab to make a contingency table of the results)?
    c. Perform K-means clustering with K = 2. Describe your results.
    d. Now perform K-means clustering with K = 4 and describe your results.
    e. Perform K-means clustering with K = 3 on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (c)? Explain.
2. Use the gene expression data in 'Ch10Ex11.csv'. This is gene expression data that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.
    a. Load the data – note that there is no header so you will need to read this in without one.
    b. Apply hierarchical clustering to the **samples**. Use all combinations of correlation and Euclidian distance with single, average and complete linkage. Plot the dendrograms. Do the genes separate the samples into the two groups? Do the results change with distance and linkage? How?

Rubric:

Code Style (5 points) Is code organized well and commented?

Submission (5 points) Was the lab submitted as an html document on Canvas? Does the html document contain a link for a GitHub repository that contains your code?

Participation (5 points) When the instructor or TA came into the breakout room were you working together and screen sharing.

Correctness (35 points, 5 points per each sub question)