CSI 5810 (Assignment # 1)

1. In this exercise, you will work with **Vertebral Column Data Set** that you can download from the following link:
http://archive.ics.uci.edu/ml/datasets/Vertebral+Column

   a. Once you have downloaded the data, you will prepare a descriptive summary of the data. The summary should describe the followings in a tabular form:
      i. Means for all features (attributes) for both normal and abnormal classes
      ii. Standard deviations for all features for each class
      iii. Medians for all features for each class.

   b. Next generate scatter plots for all feature pairs

   c. Based on (a) and (b), express your opinion about how well the two classes are separated.

2. This exercise is designed to make you familiar with multivariate normal distribution generation and using the generated data.
   a. Generate 100 3-dimensional vectors that come from a normal distribution with mean vector as $[1\ 2\ 1]^t$ and 3x3 covariance matrix as [4 0.8 -0.3; 0.8 2 0.6; -0.3 0.6 5]
   b. Make scatter plots of x1 vs x2, x1 vs x3, and x2 vs x3. Explain whatever relationships you can gather from these plots
   c. Pick any pair of generated vectors and calculate the Euclidean and Mahalanobis distances between that pair.

3. Consider the following five-dimensional records consisting of attributes 1 to 5.:

5700	12.8	2500	270	25000
1000	10.9	600	10	10000
3400	8.8	1000	10	9000
3800	13.6	1700	140	25000
4000	12.8	1600	140	25000
8200	8.3	2600	60	12000
1200	11.4	400	10	16000
9100	11.5	3300	60	14000
9900	12.5	3400	180	18000
9600	13.7	3600	390	25000
9600	9.6	3300	80	12000
9400	11.4	4000	100	13000

Suppose we are interested in reducing the five-dimensional records to two

dimensions by means of the principal component analysis. List the eigenvalues and eigenvectors obtained via PCA. Determine the reduced representation for all of the records, and plot the reduced representation in form of scatter plot. Reconstruct the original data and compute the reconstruction error.

4. Apply PCA to **Vertebral Column Data Set** and reduce the data to two dimensions [The class labels are not used in PCA]. List all eigenvalues and make a scatter plot of the transformed data. Show transformed normal and abnormal data points in different colors or shapes.

5. Repeat Exercise #3 using t-SNE visualization method. Perform visualization with two perplexity values, 10 and 50. Comment on the results obtained.