CSI 5810 (Assignment # 1)

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

   Once you have downloaded the data, you will prepare a data visualization report along the lines of visualization done for the Boston Housing data. Feel free to provide any additional visualization that might help in better understanding of the data. Write a paragraph about what characteristics of the data you see via visualization.

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 [5 0.8 -0.3; 0.8 3 0.6; -0.3 0.6 4]
   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 5 pairs of generated vectors and calculate the Euclidean and the Mahalanobis distances between those pairs

3. This exercise is designed to make you familiar with IPUMS USA data source. Go to http://usa.ipums.org, click on IPUMS Registration and Login and apply for access. You will need an account to get data. You will select 1960 1% sample and use the following variables to prepare your data extract: Marital status, Sex, Relate, Age, and Employment status. You will specify the csv format for your extract. Once you have downloaded the data, you will read the data into a data frame and visualize the age distribution of working men and the age distribution of working women. You will also calculate the % of household headed by women.

4. You will perform this exercise using the PCA-Exercise data posted on the course page.

   Suppose we are interested in reducing the six-dimensional records to two dimensions by means of 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 the form of a scatter plot. Reconstruct the original data and compute the reconstruction error.

5. In this exercise, you will apply PCA to the Spoken Arabic Digit Dataset at the following link:
   https://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit

and reduce the train 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 data points for any digit pair of your choice in different colors or shapes.

6. Repeat Exercise #5 using t-SNE visualization method to visualize the entire train data set. Comment on the results obtained.

Note: The submission should be in the form of a single PDF document. Submission in any other format will not be graded.