CIS 662: Assignment 5

Instruction: This is an individual **in-class** assignment. You are welcome to look anything up online but do NOT collaborate with your classmates or get help from anyone outside. You need to enter the responses to the questions that are given online on Blackboard (but we are also listing them below so you know the whole context). In addition to responding to the online questions on Blackboard, upload a PDF of the python code which includes the run outputs.

1. **Get data and pre-process**: The datasets are from the ISLR text book on various colleges in the US. What are uploaded are subsets of that so it is easier to visualize while plotting. See the PDF document for pages from ISLR that describe the various columns.

   (a) Read the CSV file `College_subset_islr.csv` onto your code. In your code output make sure you print the data frame header. The **main** columns we are going to use as features are starting with the fourth, "Apps". Do NOT consider "id", "College", and "Private" for the main analysis. *How many rows and main columns are in the data frame*?

   (b) Compute the mean and standard deviation of all the main columns. *What is the mean and standard deviation of the number of new students enrolled*?

   (c) Graduation rate is something a lot of policy makers are concerned about. With which other variable (i.e. not counting "graduation rate") is graduation rate most correlated? *Name the feature that is most correlated with "graduation rate" (other than itself) and what is the correlation value*?

   (d) Scale each of the main features using the standard scaler. Then use PCA to create $k = 2$ principal component vectors. Plot a graph so that the ID of each college is displayed next to the scatter plot associated with the two principal components. *What is the ID of the college in the southwest corner of the graph (it has the most negative first component), and what is the name of that college?*

   (e) Now let us look at some close ones. *From the graph state which college is closest to Texas A&M University*. Also, *is Dartmouth closer on the graph to Johns Hopkins or to Syracuse University*?

   (f) Except for Princeton, the other seven Ivy League Colleges are in the set: Brown, Columbia, Cornell, Dartmouth, Harvard, University of Pennsylvania, and Yale. *Six of them are clustered relatively close to each other, which is the one that is away*? *Also, is Georgia Institute of Technology (Georgia Tech) closer to the Ivy League cluster of 6 or closer to Syracuse*?

2. Let us now see where MIT (Massachusetts Institute of Technology) falls in this. That college is not in the list and you have to use the CSV file `MIT_from_islr.csv` and read it into your code.

   (a) As a first step use the standard scaler and transform the data by fitting in the mean and standard deviation of the previous set. Then use the PCA to transform the columns into $k = 2$ components. *What were the transformed values for MIT in the two components*?

(b) Use a distance function to determine how far MIT is from each of the other colleges in the graph. *Using the minimum distance say which was the closest college to MIT on the graph, and how far was it*?

3. Redo the code now selecting $k = 3$, i.e. three principal components. You do **not** have to display the 3D graph, but you must change the distance values for MIT.

   (a) *Did the first two component values change when you went from* $k = 2$ *to* $k = 3$?

   (b) *For the case* $k = 3$ *which college was closest in distance to MIT*?

   (c) *Now how far is the closest school to MIT for* $k = 3$?

   (d) *Say in a sentence or two what your thoughts were of visualizing and understanding similarity using PCA.*