DATA-445 IN-CLASS ASSIGNMENT NOVEMBER 10, 2021

1. (3 points) k-means algorithm can be used for which of the following?
   (a) clustering
   (b) feature engineering
   (c) All of the above
   (d) None of the above

2. (3 points) The goal for k-means cost function is to _____ squared error function where error function represents distance between data points and cluster centroid.
   (a) minimize
   (b) maximize
   (c) it depends
   (d) All of the above
   (e) None of the above

3. Consider the `customers.csv` datafile. This file contains information related to customers' activity on a company website. Below are the description of the variables.

   - `ID:` customer ID
   - `Visit_Time:` The number of visits to the company's website in a given month.
   - `Average_Expense:` The average amount of money that the customer has spend.
   - `Sex:` gender of the customer (0: female, 1: male).
   - `Age:` age of the customer.

   In Python, answer the following:

   (a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `customers`.
   (b) (3 points) Using the appropriate Python commands, remove the `ID` variable.
   (c) (5 points) Using the appropriate standardization formula, put all the variables on the same scale. *Hint:* Notice that `Sex` is a 0-1 variable.
   (d) (6 points) Using the `KMeans` function from the `sklearn.cluster` library, cluster the customers into four clusters. Make sure you use standardized variables as the inputs in the k-means algorithms, append the cluster labels to the `customers` data-frame, and use `n_init = 20` in the `KMeans` function.
   (e) (5 points) Describe each of the clusters from part (d).