Homework 4: Unsupervised Machine Learning **CSCE 633**
Due: 11:59pm on November 28, 2023

Instructions for homework submission
a) For each question, please explain your thought process, results, and observations in a markdown cell after the code cells. Please do not just include your code without justification.
b) **You can use any available libraries for this homework.**
c) Please start early :)
d) Total: 100 points

The goal of this homework is to identify a personalized marketing strategy based on customer data. An important step to achieve this is to conduct customer segmentation via customer behavior measures pertaining to credit card expenditures and payments. The data for this homework is uploaded on CANVAS (data.csv) and includes the usage behavior of approximately 9000 active credit card holders during the span of 6 months. Each row of the data corresponds to a customer and the columns include the following information:

1. CUST ID: Identification of Credit Card holder
2. BALANCE: Balance amount left in their account to make purchases
3. BALANCE FREQUENCY: How frequently the Balance is updated
4. PURCHASES: Amount of purchases made from account
5. ONEOFF PURCHASES: Maximum purchase amount done in one-go (i.e., in a single attempt)
6. INSTALLMENTS PURCHASES: Amount of purchase done in installment
7. CASH ADVANCE: Cash in advance given by the user
8. PURCHASES FREQUENCY: Frequency of purchases
9. ONEOFFPURCHASESFREQUENCY : Frequently of purchases happening in one-go
10. PURCHASESINSTALLMENTSFREQUENCY : Frequency of purchases happening in installments
11. CASHADVANCEFREQUENCY : Frequency of cash being paid in advance
12. CASHADVANCETRX: Number of transactions made with "Cash in Advanced"
13. PURCHASES TRX: Number of purchase transactions being made
14. CREDIT LIMIT: Credit card limit for the user
15. PAYMENTS: Amount of payment made by the user
16. MINIMUM PAYMENTS: Minimum amount of payments made by the user
17. PRCFULLPAYMENT: Percent of full payment made by the user
18. TENURE: Tenure of credit card service for the user

(a) (10 points) Data exploration. Plot the histograms of variables 2-18 in the data (i.e., 17 histograms total). Provide a brief discussion on your intuition regarding the variables and the resulting histograms.

(b) (20 points) Data exploration. Compute the Pearson's correlation between all pairs of variables 2-18 (i.e., all variables except from the customer ID). Assign the resulting correlation values in a 17x17 matrix C, whose (i; j) element represents the correlation value between variables i and j, i.e., C(i; j) = corr(i; j). Visualize the resulting matrix C with a heatmap and discuss potential associations between the considered variables. Note: You can use the 'heatmap' function from 'seaborn'.

(c) (30 points) K-Means Clustering. Use the K-Means clustering algorithm to cluster participants based on variables 2-16. Experiment with different number of clusters K and use the elbow method to identify the optimal number of clusters K^* based on the data. Using K^*, report the number of users that were assigned to each cluster, the centroid of each cluster (i.e., average value of each feature per cluster), and the scatter of each cluster (i.e., average distance of each sample of the cluster to the cluster centroid). Discuss your findings in association to users' percent of full payment (variable 17) and tenure of credit card service (variable 18). Note: You can use the *sklearn.cluster.KMeans* function. Consider feature normalization to avoid artificially assigning higher importance to features of a larger range.

(d) (10 points) K-Means Clustering. Repeat question (c) using a different combination of features at the input of K-Means, informed by your findings in questions (a) and (b). Please discuss your findings. Note: Consider removing highly skewed features and/or keeping features that are not correlated with each other.

(e) (30 points) Gaussian mixture models. Use the Gaussian Mixture Models (GMMs) to cluster participants based on the subset of variables that you have identified from question (d). The number of Gaussian mixtures can be approximately equal to the optimal number of clusters K^* found by (d). Report the mean vector and covariance matrix for each Gaussian and discuss your findings. Compute the log-likelihood of each sample belonging to the GMM. Plot and discuss the histogram of the resulting log-likelihood values. Note: You can use the s*klearn.mixture.GaussianMixture* function to conduct the GMM clustering and the *sklearn.mixture.score* samples to compute the log-likelihood of each data sample. You can use a heatmap to visualize the covariance matrices of the GMM, instead of printing their actual values.