PGP AI/ML UNSUPERVISED LEARNING
ASSIGNMENT 1 [Weightage: 6%]
Consider the attached dataset of Credit Card which has the following fields:

- CUSTID : Identification of Credit Card holder (Categorical)
- BALANCE : Balance amount left in their account to make purchases (
- BALANCEFREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- PURCHASES : Amount of purchases made from account
- ONEOFFPURCHASES : Maximum purchase amount done in one-go
- INSTALLMENTSPURCHASES : Amount of purchase done in installment
- CASHADVANCE : Cash in advance given by the user
- PURCHASESFREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
- CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"
- PURCHASESTRX : Numbe of purchase transactions made
- CREDITLIMIT : Limit of Credit Card for user
- PAYMENTS : Amount of Payment done by user
- MINIMUM_PAYMENTS : Minimum amount of payments made by user
- PRCFULLPAYMENT : Percent of full payment paid by user
- TENURE : Tenure of credit card service for user

Questions:
Notes: Write all code related to this assignment in a single jupyter notebook. Implement K-means algorithm from scratch, do not use any predefined ML library function for K-means algorithm. Marks will be deducted if any library function is used for KMeans. Use a single word document to answer questions. Do not unnecessarily leave the print statements in your final submitted notebook.

1. Given that K-means depends on distance metric, it is a convention to normalize the data attributes so that attributes are on the same scale. So, in this first task, normalize all data attributes. [0.5 mark]
2. Write your own code for the K-means algorithm using only two attributes, PURCHASES and CREDITLIMIT. Take K=2. Plot clusters on a scatter plot with X and Y being the two attributes. Color data points belonging to the first cluster with red and the second cluster with blue. Copy the plot diagram in the word document and interpret the output. [1.5 marks]
3. Redo question-2 on different values of K = 3,4,5. For each case, draw the plot of clusters as

stated above. Visualize these plots, copy the plot diagrams in the word document, and comment on which is better clustering (and reasons) based on visualization only. [1.5 marks]
4. Use the code written by you to cluster the data using all the features in the dataset. Take k=5 for this. [1.5 marks]5. Write a few lines as comments in the notebook about the interpretation of the best clusters obtained.
Also write a few statements about how these clusters can be useful. [1 mark]
Deliverables:
One jupyter notebook for Q 1,2,3,4
One word document for Q 2,3,4,5