Machine Learning using Python
Exam – Paper 2

[Time: 4 hrs]

[Total Marks: 50]

Part I: Unsupervised Learning [Total Marks - 30]

Given the 'credit_card' dataset, below is the data definition:

1) **CUSTID:** Identification of Credit Card holder (Categorical)

2) **BALANCE:** Balance amount left in their account to make purchases

3) **BALANCEFREQUENCY:** How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)

4) **PURCHASES:** Amount of purchases made from account

5) **ONEOFFPURCHASES:** Maximum purchase amount done in one-go

6) **INSTALLMENTSPURCHASES:** Amount of purchase done in installment

7) **CASHADVANCE:** Cash in advance given by the user

8) **PURCHASESFREQUENCY:** How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)

9) **ONEOFFPURCHASESFREQUENCY:** How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)

10) **PURCHASESINSTALLMENTSFREQUENCY:** How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)

11) **CASHADVANCEFREQUENCY:** How frequently the cash in advance being paid

12) **CASHADVANCETRX:** Number of Transactions made with "Cash in Advanced"

13) **PURCHASESTRX:** Number of purchase transactions made

14) **CREDITLIMIT:** Limit of Credit Card for user

15) **PAYMENTS:** Amount of Payment done by user

16) **MINIMUM_PAYMENTS:** Minimum amount of payments made by user

17) **PRCFULLPAYMENT:** Percent of full payment paid by user

18) **TENURE:** Tenure of credit card service for user

	Perform the following tasks:	**Marks**
Q1.	What does the primary analysis of several categorical features reveal?	**[5]**
Q2.	Perform the following Exploratory Data Analysis tasks: a. Missing Value Analysis b. Outlier Treatment using the Z-score method c. Deal with correlated variables	**[10]**
Q3.	Perform dimensionality reduction using PCA such that the 95% of the variance is explained	**[10]**
Q4.	Find the optimum value of k for k-means clustering using the elbow method. Plot the elbow curve	**[2]**
Q5.	Find the optimum value of k for k-means clustering using the silhouette score method and specify the number of observations in each cluster using a bar plot	**[3]**

Part II: Deep Learning [Total Marks - 20]

The data 'sentiment.csv' contains all information about the tweet, but for this exercise, use the text and sentiment(only positive and negative sentiments) columns. Perform all the necessary data cleaning required and answer the questions below:

	Perform the following tasks:	**Marks**
Q1.	Print the total number of positive and negative sentiments.	**[4]**
Q2.	Build a sequential LSTM model to predict positive and negative sentiments.	**[10]**

Q3. **[6]**

Based on the model, check the sentiment for the following two sentences

a. 'He is a great leader.'
b. 'He is a terrible leader.'