UNSUPERVISED LEARNING

TOTAL MARKS:30 **DURATION: 2 HOURS**

INSTRUCTIONS: -

1. *Candidates should answer all the questions in the same order provided in the question paper.*
2. *Any activity that compromises the integrity of the examination will not be permitted.*
3. *Students should complete the examination within the provided timeline.*
4. *Candidates are expected to check and ensure that the correct answer file (in. ipynb format) is uploaded in LMS.*

DATASET: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories.

ATTRIBUTE INFORMATION:

1. FRESH: annual spending (m.u.) on fresh products
2. MILK: annual spending (m.u.) on milk products
3. GROCERY: annual spending (m.u.) on grocery products
4. FROZEN: annual spending (m.u.) on frozen products
5. DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products
6. DELICATESSEN: annual spending (m.u.) on and delicatessen products
7. CHANNEL: customers Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel
8. REGION: customers Region - Lisbon, Oporto or Other

SECTION A: 5 MARKS

1. Data Understanding (5 marks)
   a. *Read the dataset (tab, csv, xls, txt, inbuilt dataset). What are the number of rows and no. of cols & types of variables (continuous, categorical etc.)? (1 MARK)*
   b. *Calculate five-point summary for numerical variables (1 MARK)*
   c. *Summarize observations for categorical variables – no. of categories, % observations in each category. (1 MARK)*
   d. *Generate the covariance and correlation tables for the data (1 MARK)*
   e. *Create Visualization plots to find the relationship amongst the variables. (1 MARK)*

SECTION B: 10 MARKS

2. How will you decide when to apply PCA based on the correlation? (2 marks)
   Apply PCA on the above dataset and determine the number of PCA components to be used so that 95% of the variance in data is explained by the same. (8 marks)

SECTION C: 15 MARKS

3. Use PCA dimensions to cluster the data. Apply K-means/ Agglomerative clustering based on the data. (15 Marks)
   Some pointers which would help you, but don't be limited by these
   a. *Find the optimal K Value. (3 marks)*
   b. *Apply Clustering and find out if the data points have been clustered correctly using appropriate visualization (6 marks)*

UNSUPERVISED LEARNING

c. Evaluate the clusters formed using appropriate metrics to support the model built. (4 marks)

d. Write down a business interpretation/explanation of the model – which variables are affecting the target the most and explain the relationship. What changes from the base model had the most effect on model performance? (2 marks)