Practical Session 7 : clustering

Live selling is becoming increasingly popular in Asian countries. Small vendors can now reach a wider audience and connect with many customers.

We try to discover different categories of customers using clustering techniques. To this end, we shall use the following dataset that has been obtained from UCI ML Repository :
`https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand`

Exercise 1 : Preprocessing

1. Import the dataset
2. Check shape of the dataset and preview the dataset using `head()`
3. View summary of dataset using `info()`.
4. Are there some missing values ? Remove the columns corresponding to these missing values. View a summary of the dataset after removal
5. View the statistical summary of numerical variables using the function `describe()`

There are 3 categorical variables in the dataset. In Exercise 2, we shall explore them one by one

Exercise 2 : Basic data exploration

1. We first explore the variable `status_id`
   (i) View the labels in the variable
   (ii) view how many different types of variables are there
   (iii) What is your conclusion ? Is this variable useful for the analysis ?
2. We now explore the variable `status_published`
   (i) View the labels in the variable
   (ii) view how many different types of variables are there
   (iii) What is your conclusion ? Is this variable useful for the analysis ?
3. Finally, we explore the variable `status_type`. View the labels in the variable and view how many different types of variables are there. Is this variable useful for the analysis ?
4. Remove unuseful variables. View the summary of dataset again
5. Convert the categorical variable `status_type` into integers. View the summary of dataset again

We shall now perform clustering

Exercise 3 : Clustering

1. Rescale the features using `MinMaxScaler()`
2. Perform k means with two clusters
3. We shall now investigate our output model
   (i) Display the centroids
   (ii) Calculate the inertia of the model

We now want to find the optimal number of clusters.

Exercise 4 : optimal number of clusters

1. Perform k means for $p = 1, \ldots, 10$ clusters. Calculate each time inertia. What is the best choice for the number of clusters
2. We shall now evaluate the choice of number of clusters using true classes related to `status_type`. We shall then set
   `y = df['status_type']`
   (i) Perform k means with 3 clusters. Compare the true labels and predicted ones
   (ii) Same with 4 clusters. What is your conclusion ?