**Programming for Big Data**

**Assignment 3**

**Due date : 05-06-24 Due date : 12-06-24**

**Instructions:**

This is a group base assignment, no more than 3 members can be the part of a group. Only one member has to submit this on portal by mentioning name and roll numbers of all 3 members.

Provided is a dataset of **Online Retail** and you have to apply **K-means** on it.keeping in view the following points:

**Overview**

[Online retail is a transnational data set](https://archive.ics.uci.edu/ml/datasets/online+retail) which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

**Business Goal**

We aim to segment the Customers based on RFM so that the company can target its customers efficiently.

RFM stands for: Recency, Frequency,Monetary

The steps are broadly divided into:

[Step 1: Reading and Understanding the Data](https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#1)

[Step 2: Data Cleansing](https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#2)

[Step 3: Data Preparation](https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#3)

[Step 4: Model Building](https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#4)

[Step 5: Final Analysis](https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#5)

* Drop records having missing values
* Customer ID should be treated as string
* In data Preparation:

We are going to analysis the Customers based on below 3 factors:[¶](https://www.kaggle.com/code/hellbuoy/online-retail-k-means-hierarchical-clustering#We-are-going-to-analysis-the-Customers-based-on-below-3-factors:)

R (Recency): Number of days since last purchase

F (Frequency): Number of tracsactions

M (Monetary): Total amount of transactions (revenue contributed)

* Compute amount spent by a customer **hint(use groupby)**
* Frequency of a customer
* Recency of a customer **hint ( Convert to datetime to proper datatype and Compute the maximum date to know the last transaction date but extract days only)**
* After this merge these two dataframes to get these final column:

columns = ['CustomerID', 'Amount', 'Frequency', 'Recency']

* now apply model having 4 clusters
* and apply Silhouette Analysis

Note:

silhouette score=(p−q)/max(p,q)

𝑝 is the mean distance to the points in the nearest cluster that the data point is not a part of

𝑞 is the mean intra-cluster distance to all the points in its own cluster.

* The value of the silhouette score range lies between -1 to 1.
* A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
* A score closer to -1 indicates that the data point is not similar to the data points in its cluster.