**THE UNIVERSITY OF HONG KONG**

**DEPARTMENT OF COMPUTER SCIENCE**

**FITE7410 Financial Fraud Analytics**

**Second Semester, 2022-2023**

**Assignment 1**

**(Due Date: Oct 28 (Fri) 23:59)**

**Assessment Criteria:**

* Plagiarism: Please follow the guidelines laid down by our department.
* You are allowed to discuss the assignment with your classmates, however, direct copy and paste is PROHIBITED and would be considered as PLAGIARISM.
* Assignments would be marked based on the logic, presentation and understanding of the problem; not only on accuracy.

1. (25%) Exploratory Data Analysis
   1. Download the following dataset Synthetic Financial Datasets For Fraud Detection from the link below:

https://www.kaggle.com/ealaxi/paysim1

* 1. Using the R package, conduct exploratory analysis of the dataset downloaded

1. (25%) Cluster Analysis
   1. Download the following dataset Credit Card Fraud Detection Data from the link below:

https://www.kaggle.com/mlg-ulb/creditcardfraud

* 1. Using the R package, conduct cluster analysis of the dataset downloaded

NOTE: A sample R script is provided, but you still need to complete the program. Or you can build the model by yourselves and use whatever library you like.

1. (50%) Write a 2-3 pages essay on the following:
   1. Describe the dataset based on the exploratory analysis result, including:
      * Summary description of the dataset
      * Univariate analysis
      * Bi-/Multi-variate analysis
      * Missing data/Outlier analysis
   2. Describe the following based on the k-means clustering results, including:
      * Explain how you find the optimal number “k”, i.e. number of clusters used to build the k-means model
      * Name the clusters found and interpret what each cluster represents
2. Submission on Moodle:
   1. R language script
   2. A **pdf** version report