|  |  |  |
| --- | --- | --- |
| https://www.universitiesegypt.com/ImageHandler.ashx?Id=12743&SS=45d51611eb2f67a4fbfc98c93ac6c147 | **Information Systems Department**  **Faculty of Computers and Artificial Intelligence Cairo University** | E:\My Documents\Cairo University\عام\FCI new logo.bmp |

**IS422 DATA MINING**

**Assignment II: Clustering**

**Instructions:**

1. Assignment is an individual task; copies or any other method of cheating will be graded to -3.
2. Solve 1 problem. If you solve both, the higher grade will be considered.
3. The assignment’s total grade is 3 marks.
4. The deadline for this assignment will be on 19/4 on Classroom; no late submissions are allowed.
5. Discussion will be during the office hours of Eng. Noura Hassan and Eng. Esraa Salah. The discussion schedule will be posted soon.
6. Your program should include a simple GUI.
7. The interface should enable the user to select the percentage of the data needed to be read from the input file, e.g., if the file contains 100 records and the user needs to read 70% of the file then the analysis should be done on 70 records only.
8. Your program should enable the user to select the file needed to be clustered.
9. The initial centroid should be chosen randomly.
10. Your program should detect the outliers (if they exist).
11. Using the programming language of your choice, write a program with the following specifications:
    1. Inputs:
       1. A file with a set of records (e.g., Excel, text, or CSV file) – (The file attached).
       2. The percentage of the data needed to be read from the input file.
       3. Number of clusters - K - will be provided from the user as an input.
    2. Outputs
       1. The final output of your program should show the content of each cluster and show outlier records.

**Problem 1**

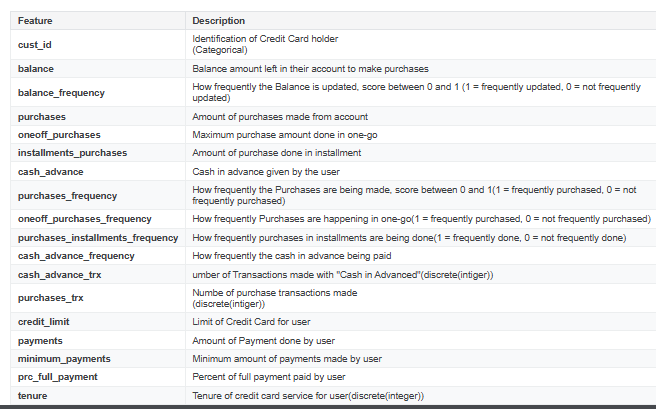
# Description:

* Consider this dataset of Supermarket customers in the file SS2025\_Clustering\_SuperMarketCustomers.CSV attached. The file contains some basic data about supermarket customers like Customer ID, age, gender, annual income and spending score. Spending Score is an attribute assigned to the customer based on some defined parameters like customer behavior and purchasing data.
* Write a program in any programming language of your choice to group customers based on the similarity of their attributes.
* You should use the k-means algorithm to cluster customers into k clusters.
* Number of clusters (k) will be provided from the user as an input.
* Initial centroid should be chosen randomly.
* You should use Euclidean distance as your distance function.
* You should detect outlier data (if it exists).
* The final output of your program should show k lists of customers and show outlier customer records.

**Problem 2**

# Description:

* Consider the bank user credit card interactions dataset in the file: SS2025\_Clustering\_CreditCardData.csv about credit card holders’ usage.
* Dataset Description:



* You are required to implement k-means clustering in any programming language you prefer to find intrinsic groups of customers within the dataset that display the same spending behavior.
* Notice that the customer can be segmented by their spending habits (e.g., high spenders, low spenders, installment buyers, or cash-heavy users) which can be assumed from attributes like: PURCHASES (Total amount spent on purchases), ONEOFF\_PURCHASES (Large one-time purchases), INSTALLMENTS\_PURCHASES (Purchases made in installments), CASH\_ADVANCE (Total amount withdrawn as cash), CREDIT\_LIMIT (Maximum credit available), and PURCHASESFREQUENCY (How often the customer makes purchases)
* Number of clusters (k) will be provided from the user as an input.
* Initial centroid should be chosen randomly.
* You should use Manhattan distance as your distance function.
* You should detect outlier data (if it exists).
* The final output of your program should show k lists of users and show outlier records.