###### 

###### **Data Mining**

**Information Systems Department**

**Faculty of Computers and Information**

###### **Cairo University**

**Assignment 2**

**Clustering**

**Instructions:**

Assignment should be done individually; copies will be graded to zero.

1. You should choose only 1 problem out of the 2 given problems to solve.
2. Total grade is 5 marks.
3. No late submissions are allowed.
4. The submission will be on classroom.
5. Deadline will be on 23/4 until 11:55 pm

**Problem 1**

**Description:**

* Consider users ratings from Google reviews dataset in the Review\_ratings.xlsx file. It contains the ratings of 1000 users on attractions from 24 categories across Europe.
* Write a program in any programming language you prefer to group the users based on the similarity of user ratings.
* You should use k-means algorithm to cluster the users to k clusters
* Number of clusters (k) will be provided from the user as an input.
* Initial centroid should be choosing randomly.
* You should use Euclidean distance as your distance function.
* You should detect outlier data (if exists).
* The final output of your program should show k lists of users and show outlier user’s records.

**Problem 2**

**Description:**

* Consider crime dataset in crime\_data.csv file, it contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
* Write a program in any programming language you prefer to group the US states based on the similarity of their crimes.
* You should use k-means algorithm to cluster the states to k clusters
* Number of clusters (k) will be provided from the user as an input.
* Initial centroid should be choosing randomly.
* You should use Manhattan distance as your distance function.
* You should detect outlier data (if exists).
* The final output of your program should show k lists of states and show outlier product’s records.