**QUESTION 1**

1. Use the R code that you programmed to do the ch 6 homework assignment of the MMM book (the video tutorial is available on youtube)

You need to modify the existing code slightly (or add a few lines) to answer the following questions (note: NO significant modification is necessary).

Make sure you set the seed to **7** immediately before you create the k-Means model with **K = 4 (4 clusters, namely, cluster 1, 2, 3 and 4)**

For your reference:  the data set (**ch6 Exercise Data Set**) contains information about at-risk students in grades 9 through 12. The attributes are as follows:

* + Student\_ID: A randomly assigned three-digit number used to uniquely identify each student.
  + First\_Name: The student’s given name.
  + Last\_Name: The student’s family name.
  + Gender: A binary code indicating female (0) or male (1).
  + Tardies: The number of times the student was late to class in the most recent academic term.
  + Absences: The number of times the student was absent from class in the most recent academic term.
  + Suspension: A binary code indicating that the student has not (0) or has (1) been suspended (temporarily excluded from attending school due to behavior problems). This code does not indicate how many suspensions the student has, only if the student has ever been suspended.
  + Expulsion: A binary code indicating that the student has not (0) or has (1) been expelled (permanently excluded from attending school due to behavior problems). This code does not indicate how many expulsions the student has, only if the student has ever been expelled.
  + Arrest: The number of times the student has been arrested by law enforcement.

Part A)

 1) In which cluster is Student ID 802 found?  [5 points]

  Your answer:

2) Overall, which cluster would you label “Critical/highest Risk”? Note that there are 4 clusters, namely, cluster1, 2,3, and 4.  [5 points]

Your asnwer:

3) In the cluster #1, how many students have not been arrested and suspended (neither arrested nor suspended)? [10 points]

    Which R expression did you use to answer this question?

Your answer:

Part B) Please execute the code with K = 3 (3 clusters) and examine the result. Which K value do you think is better? K=3 or K=4? Justify your answer by comparing two results (when K= 3 and K=4) and providing **two** **numerical evidences**. [13 points]

 Your asnwer: