HW Bonus 07: k-Means See the associated Dropbox for due date.
Thomas Kinsman

Homework is to be programmed only in one of the following languages. No other languages will be accepted. Please limit yourself exclusively to: Java, Python, Matlab, or R. The last three have good native graphics and plotting support.

Assume that the grader has no knowledge of the language or API calls, but can read comments. Use prolific block comments before each section of code, or complicated function call to explain what the code does, and why you are using it. Put your name and date in the comments at the heading of the program.

Create one directory named HW07_LastName_FirstName.
Put everything in it. Zip up the entire directory. Hand in one zip file named HW07_LastName_FirstName.zip so that when it becomes unzipped, we recover your directory.

Hand in IN A ZIP FILE:

1. Your write-up including your results, **HW05_LastName_FirstName_kMeans.pdf**
2. Your well commented code, **HW05_LastName_FirstName_kMeans…**

You are provided with a file of training data. This data has only three attributes to select from.

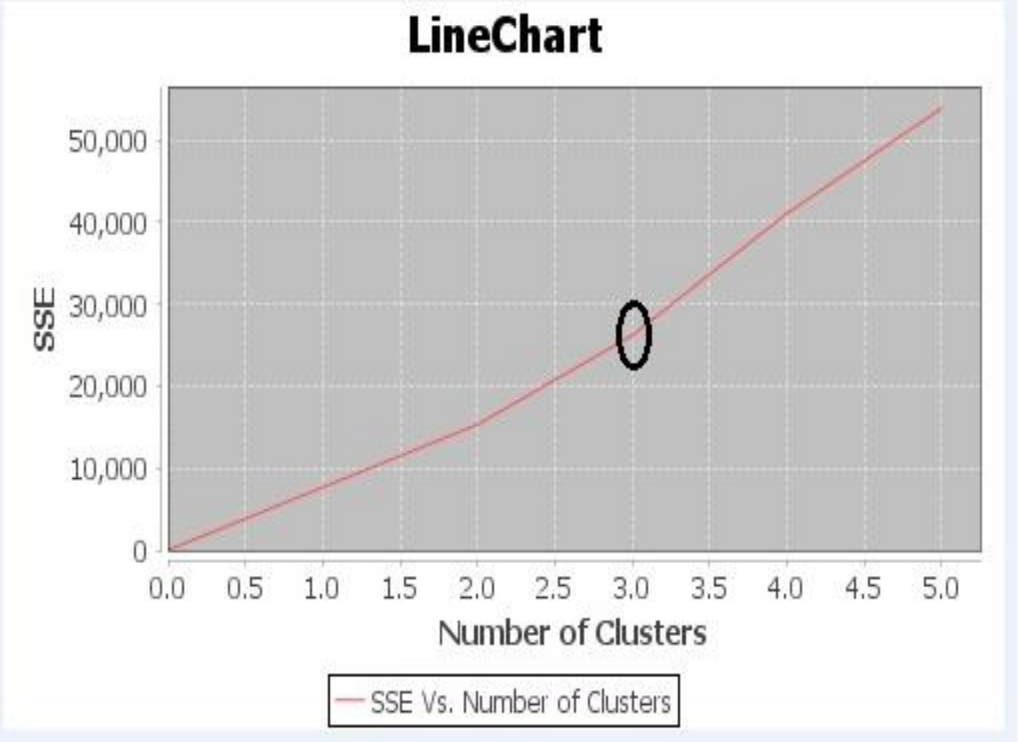
IMPLEMENT k-MEANS CLUSTERING IN BASIC CODE. (4)

1. The data file is provided.
2. The data points represents stars in space. The three attributes are (X,Y,Z) coordinates. The units are "Astronomical Units". Your job is to identify the galaxies in this star field using k-Means clustering, and your own smarts.

3. **Write-UP Should Include:**
   a. What distance metric did you use? **Use the Euclidean Distance.** Why is this a good distance metric to use? **K means clustering works on minimizing the inter cluster distance between points. In order to do this , we need to reassign the initial centroids selected to some of the clusters (by finding the centroid after initial assignment based on random points). As we might not have any specific point to consider after first iteration (centroid might not be a point) , using Euclidean distance gives us a better chance to minimize the inter cluster distance.**

   b. What prototype did you use for a cluster?
   Use a center of mass as the center of a cluster.

c. Plot the SSE versus K **FROM K = 1 to 20.** (2) (Do not use Excel.)
Identify the knee in the graph.

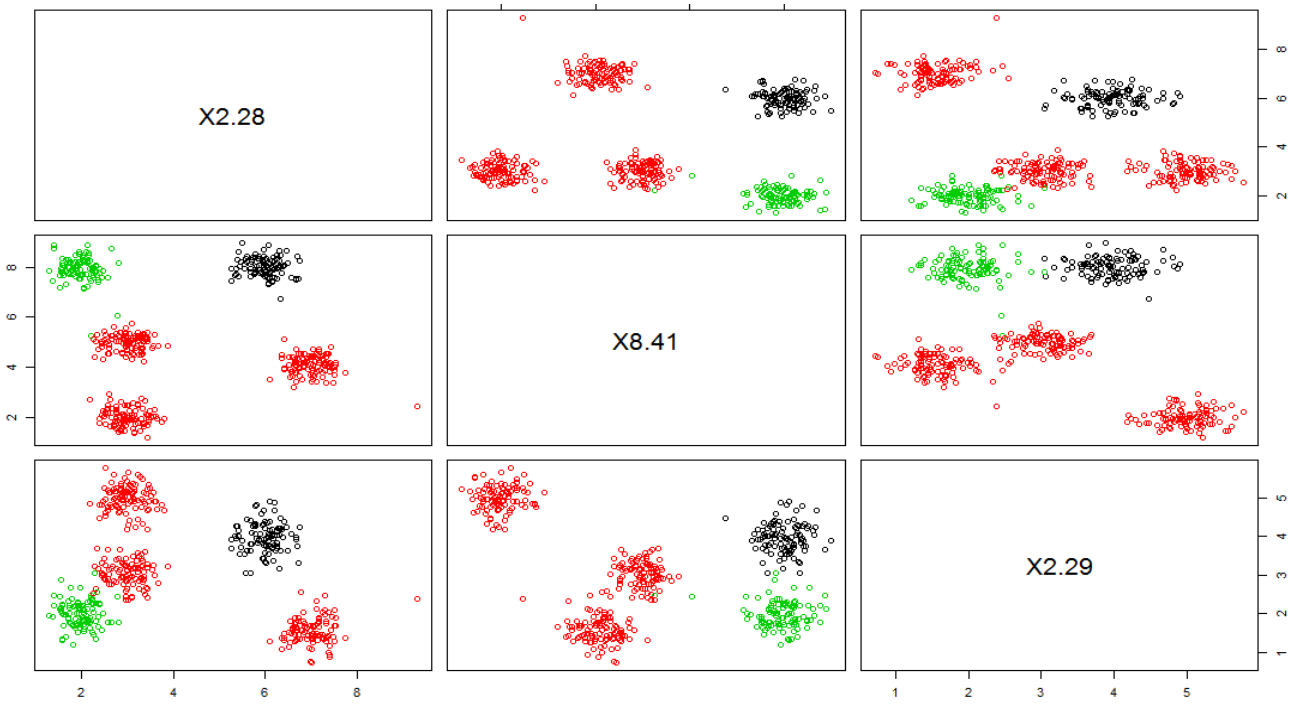


d. What value of K did you select? What was the associated SSE? (1)
Reason behind selecting this value of K (though the graph is plotted exactly reverse than expected, somehow could not debug the reason) is 3 because from that point onwards, the SSE was increasing exponentially. So this combination of SSE and number of clusters would be the best one to select.

e. Sort the clusters from smallest to largest. (2) For each cluster, 1 to K report:
   i. How many data points were in it?
   ii. What was the center of mass of the cluster?

Cluster Seed Point Selected	Number of Points	Centroid Value
1	32	(2.394,7.964,2.357)
101	270	(2.724,4.652,3.459)
201	199	(6.487,6.004,2.751)

f. Assuming K is under 11, plot each cluster in a different namespace color, starting with: red, green, blue, yellow, magenta, black, gray, brown, orange, pink, and cyan. (Skip white, and use cyan last.) Plot this on an (X,Y) axis only, with Z coming out of the paper at us. (1)
(Do not use Excel. **Plotted using R**




```
setwd("E:/BDA Homework/HW07-K-Means")
Data<-read.csv("HW_07_KMEANS_DATA_v300.csv")
clusters<-kmeans(data,3)
plot(data,col=clusters$cluster)
```