HW05: 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 had 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.

Hand in:

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

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

IMPLEMENT k-MEANS CLUSTERING.

1. (½ pts) Read through the entire homework, and estimate how long it will take to do this homework before you start the homework. Again, this is for your education. Don't cheat. Write it down before you start coding.

2. **Comments and code indicate full understanding of algorithm and mechanism.** (3)

3. The provided data file is named HW_KMEANS_DATA_v015.csv. It is completely synthetic data, based on some statistical modeling.

   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.

4. **Write-UP Should Include:**
   a. Did you do any pre-processing? (½)
   b. Did you do any noise cleaning?
      For example, you might want to ignore any point that has less than, say 10 (for example) other points within one astronomical unit of it. This might remove some stray noise. (½)
   c. What distance metric did you use? (½)
   d. What prototype did you use for a cluster? (½)
   e. Plot the SSE versus K. (1) (Do not use Excel.)
   f. What value of K did you select? What was the associated SSE? (1)
   g. Sort the clusters from smallest to largest. For each cluster, 1 to K, how many data points were assigned to it. (Sort and print smallest to largest). (1)
   h. Assuming K is under 11, plot each cluster in a different name space 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.)
   i. What was the hardest part of getting all this working?
      Did anything go wrong? (½)

5. (½) How many hours did this actually take?

6. (2 pts) BONUS:
   Black holes are most noticeable by their absence of other stars near them. Suppose that there was a black hole hidden in the data, where would the center of it be located?

   Any star too close to the black hole would be absorbed.