UCSC extension Silicon Valley

Homework 7
Patricia Hoffman, PhD.

For this homework set it is recommended that you work with your project team. If your team has less than four members, you may combine your efforts with another small team. Each person is only required to turn in the answer to one of these problems. You need to agree upon which person will work on which problem. The person who answers the forth question needs to compile the results from his team mates. Before you begin you will need to coordinate which data set to use and how the results will be scored so that the models can be compared with the same method in the answer to the 4th question. Each person who answers the 4th question will be asked to present that answer along with a description of the chosen data set.

For this assignment you are welcome to use any data set that you want. If you choose your own data set, be sure to describe your data set and indicate the source of your data. If you have selected a specific data set for your final project, you may want to consider using that data set for this homework assignment.

An interesting data set to consider for this homework set is the Synthetic Control Chart Time Series Data Set from the University of California at Irvine:
http://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series
As these are unsupervised methods, you must remove the target column before you invoke the clustering method. The targets will be used in scoring the model.

Do each of these problems on first the un-scaled data set and then for the scaled and normalized data set.

1) Use K-means to cluster the data set. How well did K-means work on this data set? What effect did normalizing and scaling the data have on the results?

2) Use Divisive Hierarchal Clustering to cluster the data set. How well did this method work on this data set? What effect did normalizing and scaling the data have on the results?

3) Now use Agglomerative Hierarchal Clustering to cluster this data set. What effect did normalizing and scaling the data have on the results?

4) Compare the three clustering methods. Which one did the best? Which one did the worst?