DAT16 SF: HOMEWORK 7 ASSIGNMENT

Assigned: Wednesday, September 9, 2015
Due: Monday, September 14, 2015, before class
Review Due: Wednesday, September 16, 2015, before class

The purpose of this homework is to review what we've learned about Clustering.

HOMEWORK QUESTIONS

DUE MONDAY:

1. Use the dataset of wholesale customers: http://archive.ics.uci.edu/ml/datasets/Wholesale+customers
2. Import the dataset, check for normalization, missing values etc.
3. Use the k-means algorithm to find clusters in the data
4. Plot the Silhouette coefficient as a function of the number of clusters. What's the ideal value for k? Why?
5. Read about other clustering techniques here: http://scikit-learn.org/stable/modules/clustering.html
6. Pick your favorite clustering technique and repeat search for clusters. Do you get a different number of clusters?

BONUS POINTS:

Go through the code in this example and make sure you understand it: http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

DUE WEDNESDAY:

1. Go to your new assigned review-buddy's repo
2. Read through your buddy's ipython notebook and make sure you understand what he/she is doing.
3. Open an issue in his/her repo and write comments on the things you don't understand and on the things you like in his/her code.
4. Quote the instructors in the comments so that we get notified about the open issue