Assignment

Implement k-means clustering. Your objects are weeks and your feature set is (μ, σ) for your weeks. We will choose the optimal number of clusters using year 1 and year 2 data. We apply the classifier to our data and will examine the composition of your clusters

Questions:

1. take $k = 3$ and use k-means sklearn library routing for k-means (random initialization and use the defaults). Take $k = 1, 2, \ldots 7, 8$ and compute the distortion vs. k. Use the "knee" method to find out the best k.

2. for this optimal k, examine your clusters and for each cluster compute the percentage of "green" and "red" weeks in that cluster.

3. does your k-means clustering find any "pure" clusters (percent of red or green weeks in a cluster is more than, say, 90% of all weeks in that cluster)?