# HW3

Separate the HW2 data set (the same data set you used for your HW2) into a training set (80%) and a test set (20%).
Use an appropriate distance measure, to determine nearest neighbors, and to group individuals in the training set, based on all the 2017-2021 citation columns in the data set.
What is the right number of clusters for this problem? Why?

For each of the test data, find the nearest cluster centroid and place the test data into that cluster.
Tabulate the following predictions for the 2022 citation numbers for the test set, using the average difference magnitude to evaluate them:

(1) same as the 2022 citation number of the nearest neighbor from the training set;
(2) same as the point nearest the cluster centroid;
(3) average of all others from the training set in the same cluster.
Draw conclusions from the comparison.

Your submission should include:
1. code and predictions.
2. report_HW3.pdf (explaining your results and your conclusions)

Due: 10/18/2023