**Homework 6:**

Let’s still use the AdultUCI dataset about whether a person’s income will exceed $50K/yr. Let’s apply a *k-means* cluster analysis on this dataset. Since the *k-means* approach only applies to a dataset with numeric variables, we need to remove all the categorical variables.

By the way, *k-modes* is used for clustering categorical variables. It defines clusters based on the number of matching categories between data points. (This is in contrast to the more well-known k-means algorithm, which clusters numeric data based on Euclidean distance.) The *k-prototypes* algorithm combines *k-modes* and *k-means* and is able to cluster mixed numeric / categorical data. Both *k-modes* and *k-prototype* are documented in the paper below. If you are interested, please read the paper. It is very easy to follow.

Huang, Z.: Clustering large data sets with mixed numeric and categorical values, Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, Singapore, pp. 21-34, 1997.

Back to this Homework 6, we will apply the k-means approach, please finish the following tasks:

1. Remove all the categorical features. You need to retain the target variable as the ground truth to compare your *k-means* model prediction.

2. Normalize all the numeric variables.

3. Perform k-means cluster analysis based on the Euclidean distance.

4. Please check the average within cluster distances for different k values (from k =1 to k = 10).

5. If we decide that k takes the value of 2 in order to match the two ground truth groups (small income and big income), please report the k-means model accuracy.