MTH 443: Lab Problem Set 4

Dataset #1: The "Wine dataset" (source: UCI Machine Learning Repository) (Wine_data.csv) gives data that are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different wineries. The analysis determined the quantities of 13 constituents found in each of the three types of wines, these are 1) Alcohol, 2) Malic acid, 3) Ash, 4)Alcalinity of ash, 5) Magnesium, 6) Total phenols, 7) Flavanoids, 8) Nonflavanoid phenols, 9)Proanthocyanins, 10)Color intensity, 11)Hue, 12)OD280/OD315 of diluted wines and 13)Proline.

(a) Ignoring the Type variable, representing the winery, obtain the clustering of the wine samples based on the features using;

(i) Hierarchical clustering with complete and single linkage (plot the dendograms);

(ii) K-means method with K=3 (obtain clusters associated with at least 5 initial cluster configurations, more if required).

(b) Calculate total within cluster point scatter around respective cluster centroids for comparing the clusters detected in (a) and (b), keeping the number of clusters same.

(c) Validate the results of your clustering exercise with the "Type" variable after you have obtained the clusters.

Dataset #2: The dataset (CrimesOnWomenData.csv) for this exercise is an utterly depressing dataset recording the crimes committed against women in various states of India between 2001 to 2021. The source of this *kaggle* dataset is *https://data.gov.in*. It has state-wise data of India on the following crimes committed against women between 2001 to 2021.

Rape	No. of Rape Cases
K&A	Kidnapping and Abduction cases
DD	Dowry Deaths
AoW	Assault against Women
AoM	Assault against Modesty of Women
DV	Domestic Violence
WT	Women Trafficking

(a) Obtain year-wise hierarchical clustering of the states with average and single linkage. Plot the dendogram.

(b) Obtain year-wise clusters using K-means method with K=5 (obtain clusters associated with at least 5 initial cluster configurations, more if required).

(c) Compare hierarchical and K-means clusters of states for same number of clusters using within cluster point scatter measure. Also compute the between cluster point scatter measure.

(d) Do you observe any change in clustering patterns/composition over the years?