**MACHINE LEARNING FOR DATA SCIENCE**

**Instructions:**

**Use the EM clustering to group the similar slump, flow and compressive strength values from the amount of ingredients. Present your explorations using appropriate model building and visualizations**

**Solution:**

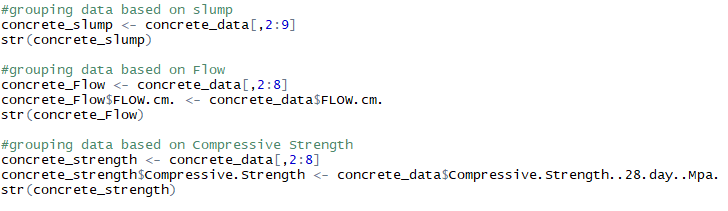
EM clustering is one of the unsupervised clustering algorithms. It finds clusters by first assigning each point, a probability of belonging to each cluster and then re-estimating the probability distribution for each cluster. This process continues until the error has converged at which time the points are all assigned to a given cluster with some degree of uncertainty. This is the iterative process for various counts of clusters. The quality of each model is accessed by lower BIC value (Bayesian Information Criteria) and the model with the number of clusters that maximizes the quality of the model.

In our problem, we are grouping the data based on slump, flow and compressive strength variables with their other ingredients for building their respective models.

**Steps to be followed:**

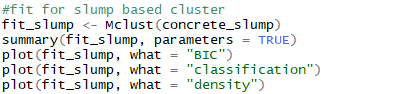
1. Loading the dataset into the workspace and look at the summary and the structure of the data.

2. Now we can segregate the data based on the three specified variables such as SLUMP, FLOW and COMPRESSIVE STRENGTH

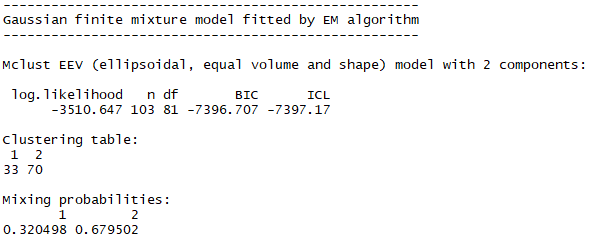


3. Now as the dataset is segregated into three modules, let us apply the EM clustering algorithm on each module to estimate their appropriate model.

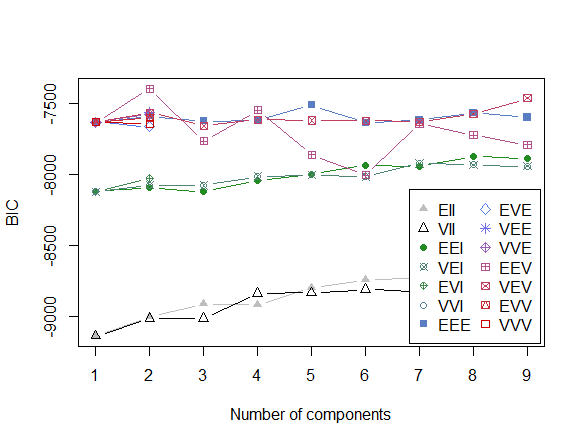
Model for the slump based dataset:

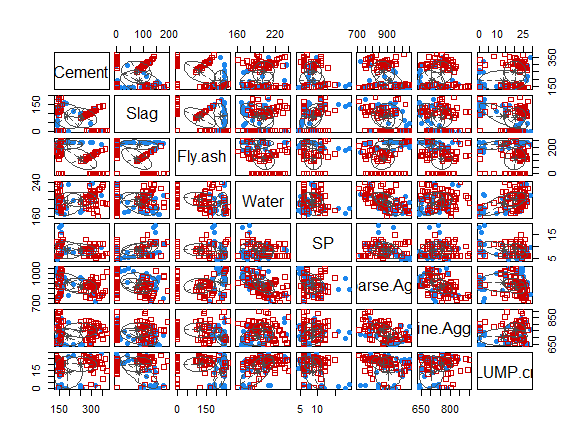


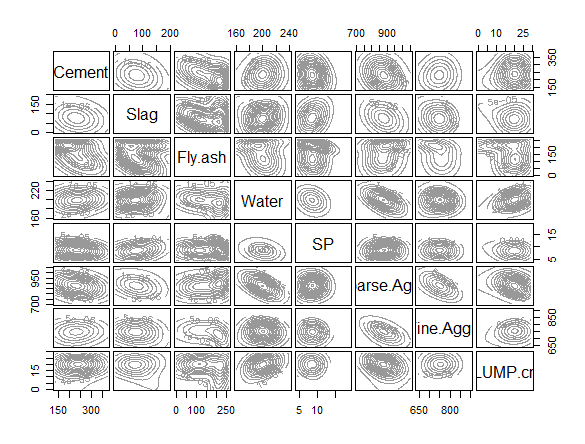
And its output is as below represents the BIC value of -7396.707 and it has formed 2 cluster components with their probability of 0.32 and 0.68 respectively.



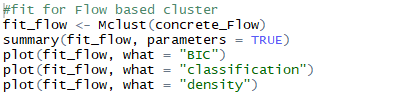
The visualization of the model components such as BIC, classification, density is shown below



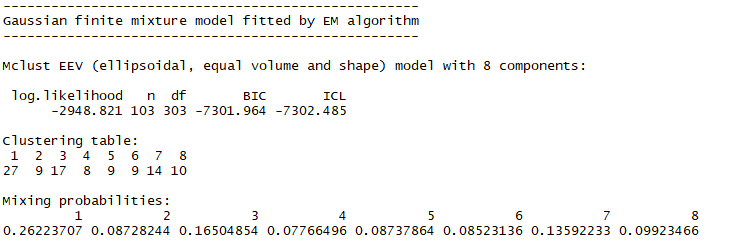




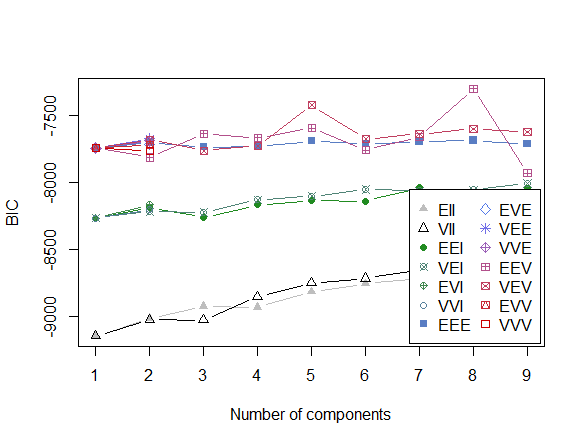
Model for the FLOW based dataset:

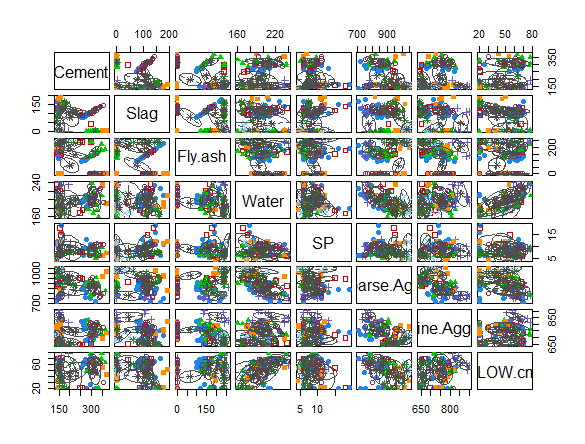


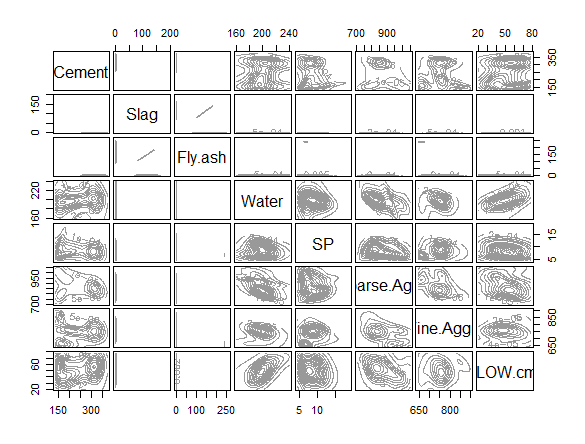
The output of this model fit shows that the data has been split into 8 cluster components with the BIC value of -7301.964 and their corresponding probabilities.



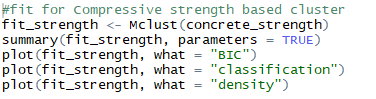
The visualizations of the second model based on the flow grouped clusters for the model components like BIC, Classification and density is as shown below



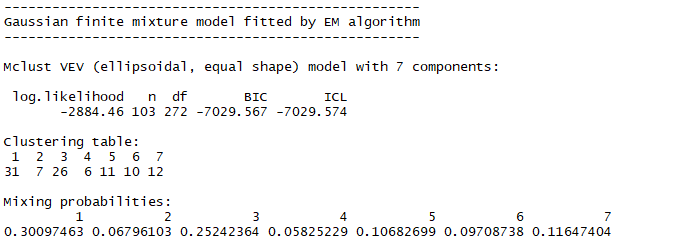




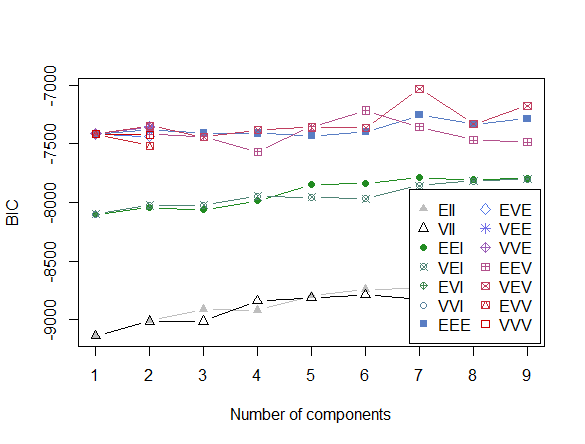
Model for the Compressive Strength based dataset:

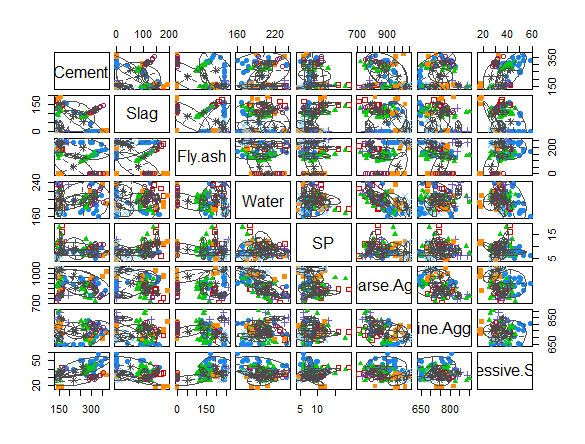


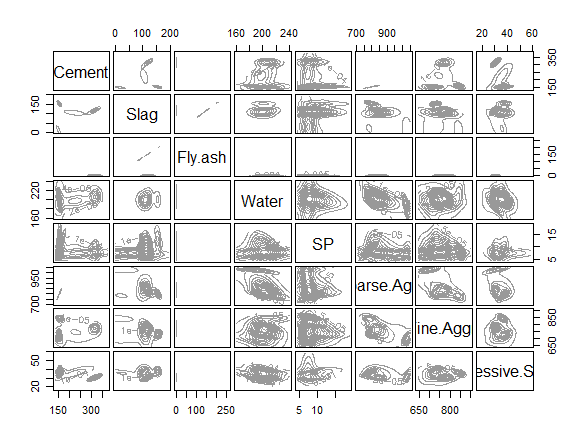
Below output of this model based on the data grouped under compressive strength shows that the data has been split into 7 cluster components with the BIC value of -7029.567 and their corresponding probabilities.



Visualizations of the model with their components:

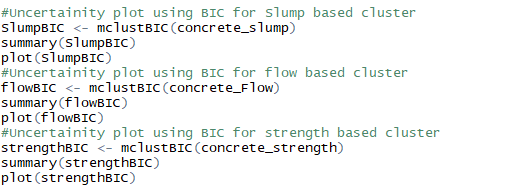






Considering three models BIC value, the data which are grouped with the Slump variable has the lesser BIC value. Always the lower the BIC, the better the quality of the model is.

And the Uncertainty of the model can also be computed using the BIC for EM initialized by model-based hierarchical clustering for parameterized Gaussian mixture models for all three models.



Their summary of the models shows us the different BIC values of the various models created.

