**Cluster Formation:**

With the standardized cluster formation is done. K – Means Clustering is done using

PROC FASTCLUS DATA = C.STD MAXC = 50 (maximum number of clusters to start with)

MAXITER = 150 (maximum Number of Iterations) DELETE = 36 (Minimum Observation per Cluster – generally kept at 5% of total observation) OUT = C.CLOUT (saves the cluster membership variable);

After the 1st cut we alter the variable set till the following checks of optimality for cluster are fulfilled.

**Checks for cluster**

1. **Individual R-Squared** >= 0.25. Every variable used in cluster formation generates an R-Squared. This measures the worth of the variable in the cluster formation. The final model must have only those variables for which R-Squared >= 0.25
2. **Overall R-Squared** >= 0.5. This measure the overall goodness of fit of the model and should be >= 0.5.
3. **Approximate Expected Overall R-Squared** > =0.3 (This is the R-Squared which the model would have generated if there was no Multicollinearity among the variables used in cluster formation). It should be >= 0.3.
4. **The Difference Between Overall R-Squared and Approximate Expected Overall R-Squared** Should Not Be Greater Than 0.2. A higher difference indicates unacceptable amount of Multicollinearity among the variables used in cluster formation.
5. **RMS Standard Dev** < =1.4. This is a measure of within cluster homogeneity. It should be < =1.4 for each cluster. A higher value for any cluster indicates presence of outliers in that cluster.
6. **Distance between Cluster Centroids** >= 1.4. This is a measure of across cluster heterogeneity. The distance between centroid of any cluster with that of the nearest cluster should be >= 1.4
7. **Number of Clusters** Should Be Between 4 and 15.
8. **Percentage of Frequency** in Each Cluster Should Be < = 35.