**SAS CLUSTERING ASSIGNMENT**

For this assignment, use “powerusagel.xlsx” dataset uploaded in canvas. This dataset contains measurements of electric power consumption in one household with a one-minute sampling rate for a day of a typical household.

This dataset has 5 attributes as described below:

1. **Time** in format hh:mm:ss

2. **global\_active\_power**: household global minute-averaged active power (in kilowatt)

3. **global\_reactive\_power**: household global minute-averaged reactive power (in kilowatt)

4. **voltage**: minute-averaged voltage (in volt)

5. **global\_intensity**: household global minute-averaged current intensity (in ampere)

Using this data, user desires to divide data into different clusters to find further insights about power consumption statistics

a) Perform K-means clustering on powerusage data with k=6 between global\_active\_power, global\_reactive\_power and global\_intensity variables

b) How many iterations are performed? What are initial and final centroids for Cluster 2?

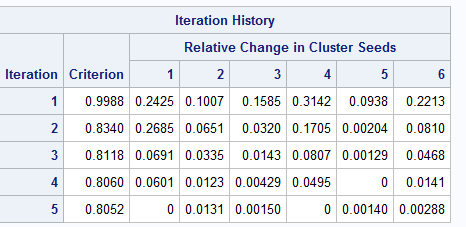
c) Report “Cubic Clustering Criterion” of this clustering? Does this value suggests a good quality Clustering? If not, suggest what can be done to improve the quality of clustering with proper reasoning.

d) Perform K-means clustering with K=6 (after steps suggested in c) and comment on the quality of clustering.

e) What are the similarity and dissimilarity measures in SAS output for Cluster 4? (Hint: Look into Cluster Summary table).

**Response:**

1. Ran **proc** **fastclus** on powerusage data with maxclusters=6 and max iterations = 10 between var global\_active\_power global\_reactive\_power global\_intensity
2. 5 iterations were performed

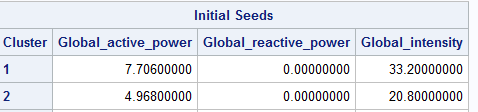


Initial centroids of cluster 2 are:

global\_active\_power : 4.968

global\_reactive\_power : 0.00

global\_intensity : 20.80

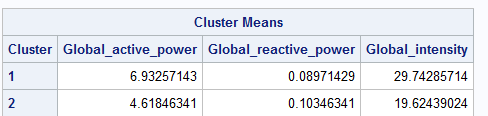


Final centroids of cluster 2 are:

global\_active\_power : 4.61846341

global\_reactive\_power : 0.10346341

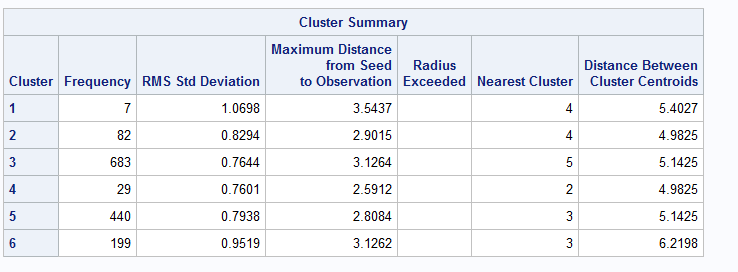
global\_intensity : 19.62439024



1. Cubic clustering criterion is 1.268 which means it is not a good cluster and indicates potential clusters.



Also, from the cluster summary table we can see that the clusters are not uniformly distributed as there are few cluster with very less frequency (like cluster 1 and 4) and few with large frequency (Cluster 3 and cluster 5) which may indicate the presence of outliers.

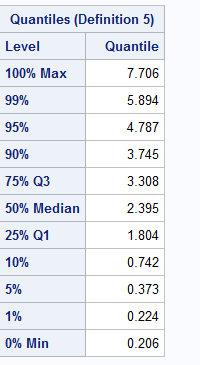


Ways to improve the Cubic clustering criterion are:

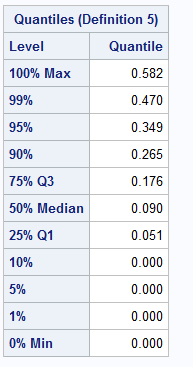
1. Identify and remove outliers: Outliers are observations that fall below Q1 - 1.5(IQR) or above Q3 + 1.5(IQR). Here, IQR = Q3 - Q1. I can either remove all the values for the variables above 1.5IQR or cap large values at 99th percentile after plotting the boxplots.
2. Check for Multi-collinearity and remove variables if they are highly correlated as the inputs should be relatively independent and some variables can get a higher weight than others when the variables used in clustering are highly correlated.
3. Run the clustering proc by increasing or decreasing the number of clusters to identify the optimum number of clusters based on CCC.
4. Since it is asked to run the model with K=6, I will not be using #3 above.
5. Box plot for each of the variables and remove outliers based on a 1.5IQR

There are no observations below Q1-1.5IQR for any of the variables, so will just remove upper fence outliers greater than Q3+1.5IQR

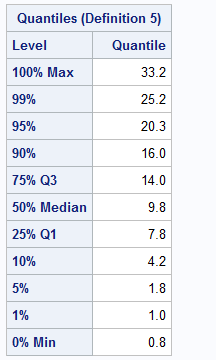
global\_active\_power : Q3+1.5IQR = 5.564



global\_reactive\_power : Q3+1.5IQR = 0.3635



global\_intensity : Q3+1.5IQR = 23.3



/\*removing outliers based on the output from box plot\*/

**data** mod\_powerusage; set powerusage;

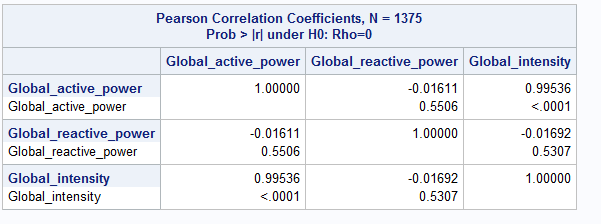
if global\_active\_power > **5.570** then delete;

if global\_reactive\_power > **0.3700** then delete;

if global\_intensity > **23.4** then delete;

**run**;

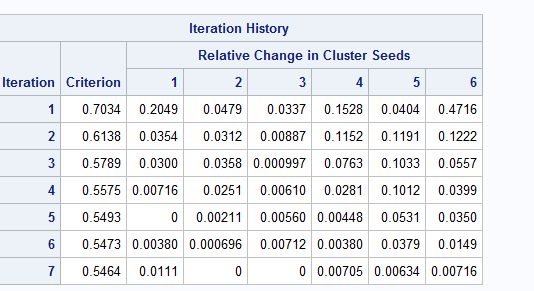
1. Check for multi-collinearity by using proc corr between global\_active\_power global\_reactive\_power global\_intensity



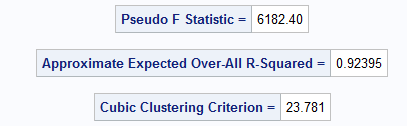
From the table we can see that global\_active\_power and global\_intensity are highly correlated (99.5% correlation) and need to keep any one of them. Since it is asked to run the model with all the three variables I will not be removing any.

First run the clustering after removing the outliers for all the three variables.

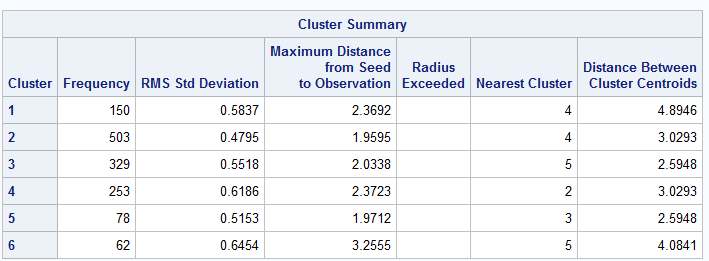
7 Iterations were performed this time.



CCC, R-squared and Pseudo F-stat all looks good after removing the outliers



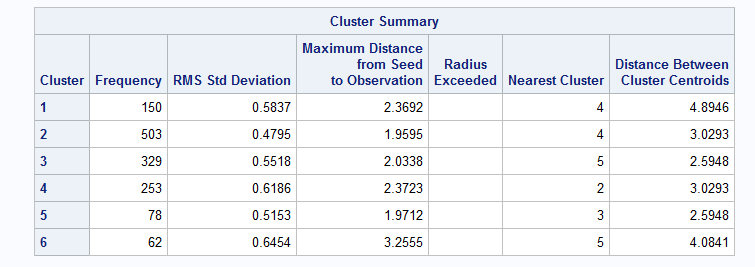
The cluster distribution also looks ok.



1. From the Cluster Summary table we can see that the nearest cluster to cluster 4 is Cluster 2 and it also the nearest cluster to Cluster 1.

As, small distance equals high similarity / low dissimilarity, we can say that cluster 4 is similar to cluster 2 and 1.

And, great distance equals low similarity / high dissimilarity, we can say that cluster 4 is dissimilar to clusters 3, 5 and 6.



SAS Code:

/\* Import data from folder\*/

**proc** **import** datafile='G:\My Drive\DSBA-6201 Data\SAS Demo\power\_usage.xlsx'

DBMS=xlsx out = powerusage replace;

/\* Performing K-means clustering on powerusage data with k=6

between global\_active\_power, global\_reactive\_power and global\_intensity variables\*/

**proc** **fastclus** data=powerusage maxclusters=**6** out=clust MAXITER=**10**;

var global\_active\_power global\_reactive\_power global\_intensity;

**run**;

/\* finding outliers \*/

**proc** **univariate** data=powerusage normal plot;

var global\_active\_power global\_reactive\_power global\_intensity;

**run**;

/\*removing outliers based on the output from box plot\*/

**data** mod\_powerusage; set powerusage;

if global\_active\_power > **5.570** then delete;

if global\_reactive\_power > **0.3700** then delete;

if global\_intensity > **23.4** then delete;

**run**;

**proc** **print** data = mod\_powerusage;

**run**;

/\* Check for Multi-collinearity \*/

**proc** **corr** data=mod\_powerusage;

var global\_active\_power global\_reactive\_power global\_intensity;

**run**;

/\* Performing K-means clustering on powerusage data with k=6

between global\_active\_power, global\_reactive\_power and global\_intensity variables

after removing outliers\*/

**proc** **fastclus** data=mod\_powerusage maxclusters=**6** out=clust MAXITER=**10**;

var global\_active\_power global\_reactive\_power global\_intensity ;

**run**;