Assignment : MTech KE Unit 5 – Group Exercise – PCA/FACTOR/Cluster Analysis– 10%

Due Date : 3rd March, 2018 : **Please mention the name of the group members in your submission.** FT & PT groups should be formed separately. Follow the group structure already formed for other KE assignments

The output can be in 1) Word or 2) PowerPoint or 3) PDF but complete with all the details of methodology followed as described in the next slide.

1. Download from the net an appropriate data set (data sets with many explanatory i.e. **X** variables) suitable for significant dimension reduction (PCA/Factor Analysis). Dataset has to be either from Singapore or student's home country (acceptable as long as it is from this region). Try to find the data so that it involves some prediction problem i.e. it has some **Y** variable as well. With this dataset solve the following problems :

   a) Perform the dimension reduction using PCA/Factor Analysis. Select the reduced number of orthogonal variables (using eigen value criteria as taught in the class) and give them suitable names.

   b) Use these new variables to perform suitable number of clusters (maximum 6) complete with their description or profile.

   c) Perform a regression analysis using the reduced number of orthogonal variables found **in a)** following proper regression steps and comment on your prediction quality. There is no need to spilt into development and validation sample for this regression exercise.

Assignment 1 : MTech EBAC Unit 5 – Group Exercise – PCA/FACTOR/Cluster Analysis– contd.

The final report (no more than 10 pages) should contain,

1. All the necessary assumptions, steps followed to reduce the dimension, comment on the variability in the data explained by the selected orthogonal variables.

2. The rationale behind naming the orthogonal variables.

3. The description of clustering results in terms of why certain number of clusters were chosen.

4. The steps followed in regression analysis and comment on the strength of the model and limitations in the findings

Choice of appropriate data set – 2 Marks

Part a) **– 3 Marks**

Part b) **– 3 Marks**

Part c) **– 2 Marks**

Total **10 Marks**