**CLUSTER ANALYSIS PROJECT IN SAS**

**TABLE OF CONTENTS:**

1. Setting up the Library in SAS
2. Outliers Checking and Treatment
3. Missing Value Checking and Treatment
4. Factor analysis (removing multi collinearity)
5. Standardization
6. Cluster formation
7. Cluster Details

**1. SETTING UP THE LIBRARY IN SAS:**

/\*Setting Library\*/

LIBNAME Cluster "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project";

**RUN**;

/\*To Export SAS Data to Excel FIle\*/

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\ClusterData.XLS";

**PROC** **PRINT** DATA = Cluster.CLUSTER\_DATA;

**RUN**;ODS HTML CLOSE;

1. *Cleaning/Sanity Check*
   1. *Outliers – Valid Value (Maximum 1% of Total data) – Floor Them/Cap them, if same observation have outliers across multiple variables remove the observation*
      1. *Extremely Value list from Proc Univariate (2-3% on either side)*
      2. *Scatter Plot*
      3. *Min and 1% Values/Max and 99% Values is very different*
      4. *Mean and Median is very different*
      5. *Thumb Rules* 
         1. *Normal Distribution (Mean ± 3\* Standard Deviation)*
         2. *Not Normal (Q1 – 1.5\*(Q3-Q1), Q3 + 1.5\*(Q3-Q1)*
   2. *Invalid Values – Make them Valid*
   3. *Missing Values – NMISS Option*
      1. *Mean of Non Missing Values*
      2. *Grouped Means*
      3. *Regression*

**2. OUTLIER TREATMENT:**

Check 15% of either side (Low and High) for Outliers. Check if the difference between Mean and Median is substantial. Also, use scatter plot to find the outliers. Outlier treatment by using *PROC MEANS* and *PROC UNIVARIATE*

/\*Checking Outliers\*/

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\Outliers.XLS";

**PROC** **MEANS** DATA = Cluster.CLUSTER\_DATA MIN MAX

MEAN MEDIAN Q1 Q3 STD;**RUN**;ODS HTML CLOSE;

/\*Univariate Test to check outliers exist or not\*/

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\UnivariateTest.XLS";

**PROC** **UNIVARIATE** DATA = Cluster.CLUSTER\_DATA NEXTROBS = **15**;

**RUN**;ODS HTML CLOSE;

*Outliers exist for following Variables:*

*VARIABLES (Higher Outlier Capping)*

*TOTAL\_DAILY\_ORDER\_NET\_AMT\_12 (80925)*

*AVG\_DAILY\_ORDER\_NET\_AMT\_12 (11554.2)*

*TOTAL\_LINE\_ITEM\_QTY\_12 (80)*

*AVG\_LINE\_ITEM\_QTY\_12 (5)*

*TOTAL\_DOLLAR\_DISCOUNT\_12 (536197)*

*AVG\_DOLLAR\_DISCOUNT\_12 (21945.75)*

*SIZE\_NET\_SALES (1.905025)*

*SIZE\_QTY (1.650989)*

*SIZE\_DOLLAR\_DISC (2.585263)*

*RATIO\_DISC\_SALES (6.88889)*

*med\_DAILY\_ORDER\_NET\_AMT\_12 (22500)*

*med\_LINE\_ITEM\_QTY\_12 (12)*

*med\_DOLLAR\_DISCOUNT\_12 (21945.75)*

*count (31)*

*sum\_value\_12 (498465)*

/\*Outliers Treatment\*/

**DATA** Cluster.CLUSTER\_DATA;

SET Cluster.CLUSTER\_ DATA;

IF TOTAL\_DAILY\_ORDER\_NET\_AMT\_12>**80925** THEN TOTAL\_DAILY\_ORDER\_NET\_AMT\_12=**80925**;

IF AVG\_DAILY\_ORDER\_NET\_AMT\_12>**16600** THEN AVG\_DAILY\_ORDER\_NET\_AMT\_12=**16600**;

IF TOTAL\_LINE\_ITEM\_QTY\_12>**80** THEN TOTAL\_LINE\_ITEM\_QTY\_12=**80**;

IF AVG\_LINE\_ITEM\_QTY\_12>**5** THEN AVG\_LINE\_ITEM\_QTY\_12=**5**;

IF TOTAL\_DOLLAR\_DISCOUNT\_12>**307920** THEN TOTAL\_DOLLAR\_DISCOUNT\_12=**536197**;

IF AVG\_DOLLAR\_DISCOUNT\_12>**21945.75** THEN AVG\_DOLLAR\_DISCOUNT\_12=**21945.75**;

IF SIZE\_NET\_SALES>**1.905025** THEN SIZE\_NET\_SALES=**1.905025**;

IF SIZE\_QTY>**1.650989** THEN SIZE\_QTY=**1.650989**;

IF SIZE\_DOLLAR\_DISC>**2.585263** THEN SIZE\_DOLLAR\_DISC=**2.585263**;

IF RATIO\_DISC\_SALES>**6.88889** THEN RATIO\_DISC\_SALES=**6.88889**;

IF med\_DAILY\_ORDER\_NET\_AMT\_12>**22500** THEN med\_DAILY\_ORDER\_NET\_AMT\_12=**22500**;

IF med\_LINE\_ITEM\_QTY\_12>**6** THEN med\_LINE\_ITEM\_QTY\_12=**6**;

IF med\_DOLLAR\_DISCOUNT\_12>**21945.75** THEN med\_DOLLAR\_DISCOUNT\_12=**21945.75**;

IF count>**31** THEN count=**31**;

IF sum\_value\_12>**498465** THEN sum\_value\_12=**498465**;

**RUN**;

*Note: If a particular row has multiple outliers across variables, then delete the observation. Rows deleted: 48,58,68,184, 279, 616, 696,698,*

**3. MISSING VALUE TREATMENT:**

/\*Missing Value Checking\*/

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\Missing Value.XLS";

**PROC** **MEANS** DATA = Cluster.CLUSTER\_DATA Nmiss;

**RUN**;

ODS HTML CLOSE;

*Note: The given data doesn’t have any missing values.*

**4. FACTOR ANALYSIS (REMOVING MULTI COLLINEARITY):**

The multicollinearity issue is reduced using PROC FACTOR method in SAS. Factors are linear combination of variables. Each factor has some part of each variable into it. Factors will be rotated in N-Dimensional plane till they become mutually orthogonal and hence independent from each other. Factors thus generated has no multicollinearity. By Factor Analysis, we consider those factors which explain 95% of total variation with data.

/\*FACTOR ANALYSIS\*/

**PROC** **FACTOR**:

NFACT: NO OF FACTORS GENERALLY KEPT SAME AS THE NUMBER OF VARIABLES

METHOD: PRIN (SHORT FORM FOR PRINCIPAL COMPONENT ANALYSIS)

ROTATE : VARIMAX (ROTATION METHOD ENSURING ORTHOGANILITY OF FACTORS)

OUT : OUTPUT FILE WHICH STORE THE FACTOR SCORES;

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\FactorAnalysis.XLS";

**PROC** **FACTOR** DATA = Cluster.CLUSTER\_DATA NFACT = **18**

METHOD = PRIN ROTATE = VARIMAX OUT = Cluster.FAC;

VAR

med\_Actual\_price\_12

AVG\_Actual\_price\_12

TOTAL\_DAILY\_ORDER\_NET\_AMT\_12

AVG\_DAILY\_ORDER\_NET\_AMT\_12

TOTAL\_LINE\_ITEM\_QTY\_12

AVG\_LINE\_ITEM\_QTY\_12

TOTAL\_DOLLAR\_DISCOUNT\_12

AVG\_DOLLAR\_DISCOUNT\_12

SIZE\_NET\_SALES

SIZE\_QTY

SIZE\_DOLLAR\_DISC

RATIO\_DISC\_SALES

med\_DAILY\_ORDER\_NET\_AMT\_12

med\_LINE\_ITEM\_QTY\_12

med\_DOLLAR\_DISCOUNT\_12

MONTH\_SINCE\_LAST\_TRANSACTION

count

sum\_value\_12

;

**RUN**;

**QUIT**;

ODS HTML CLOSE;

After generating the excel file with all the details of Factor Analysis, look into the Eigen values of the Factors. Select the Factors with Eigen values more than 1. Check how many factors are giving 95% variation in total (Cumulative Proportion).

After selecting factors with highest Eigen values, we need to check which are the most important variables. We need to go to the last generated table (after all the rotations) and check for factor loadings. Choose the variables for all the factors separately which has highest factor loadings. This activity we need to perform for all the selected factors and then select the one with relatively higher loading.

/\*To check the correlation between Factors\*/

**PROC** **CORR** data = Cluster.fac;

var

Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7

Factor8 Factor9 Factor10 Factor11 Factor12

Factor13 Factor14 Factor15 Factor16 Factor17 Factor18

;**RUN**;

**5. STANDARDIZATION:**

In statistics, standardized variables are variables that have been standardized to have a [mean](http://www.statisticshowto.com/mean/)of 0 and a [standard deviation](http://www.statisticshowto.com/probability-and-statistics/standard-deviation/)of 1. The variables are rescaled using the [z-score formula](http://www.statisticshowto.com/probability-and-statistics/z-score/). **Standardizing makes it easier to compare scores**, even if those scores were measured on different scales. It also makes it easier to read results from [regression analysis](http://www.statisticshowto.com/probability-and-statistics/regression-analysis/) and ensures that all variables contribute to a scale when added together.

/\*STANDARDISATION\*/

/\*Only selected variables from factor analysis\*/

/\*to be standardised\*/

**PROC** **STANDARD** DATA = Cluster.CLUSTER\_Copy2

MEAN = **0** STD = **1** OUT = Cluster.Standard;

VAR

/\*med\_Actual\_price\_12\*/

AVG\_Actual\_price\_12

TOTAL\_DAILY\_ORDER\_NET\_AMT\_12

/\*TOTAL\_LINE\_ITEM\_QTY\_12\*/

/\*TOTAL\_DAILY\_ORDER\_NET\_AMT\_12\*/

/\*AVG\_LINE\_ITEM\_QTY\_12\*/

/\*TOTAL\_DOLLAR\_DISCOUNT\_12\*/

AVG\_DOLLAR\_DISCOUNT\_12

/\*SIZE\_NET\_SALES \*/

/\*SIZE\_QTY\*/

/\*SIZE\_DOLLAR\_DISC \*/

/\*RATIO\_DISC\_SALES\*/

/\*med\_LINE\_ITEM\_QTY\_12\*/

/\*med\_DOLLAR\_DISCOUNT\_12\*/

MONTH\_SINCE\_LAST\_TRANSACTION

/\*count\*/

sum\_value\_12

;

**RUN**;**quit**;

**6. 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.

/\*CLUSTER FORMATION\*/

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\ClusterFormation.XLS";

**PROC** **FASTCLUS** DATA = Cluster.Standard MAXC = **50**

MAXITER = **150** DELETE = **36** OUT = Cluster.CLOUT;

VAR

/\*med\_Actual\_price\_12\*/

AVG\_Actual\_price\_12

TOTAL\_DAILY\_ORDER\_NET\_AMT\_12

/\*TOTAL\_LINE\_ITEM\_QTY\_12\*/

/\*TOTAL\_DAILY\_ORDER\_NET\_AMT\_12\*/

/\*AVG\_LINE\_ITEM\_QTY\_12\*/

/\*TOTAL\_DOLLAR\_DISCOUNT\_12\*/

AVG\_DOLLAR\_DISCOUNT\_12

/\*SIZE\_NET\_SALES \*/

/\*SIZE\_QTY\*/

/\*SIZE\_DOLLAR\_DISC \*/

/\*RATIO\_DISC\_SALES\*/

/\*med\_LINE\_ITEM\_QTY\_12\*/

/\*med\_DOLLAR\_DISCOUNT\_12\*/

MONTH\_SINCE\_LAST\_TRANSACTION

/\*count\*/

sum\_value\_12

;

**QUIT**;

ODS HTML CLOSE;

If Cluster checks are not satisfied, we need to delete the variable with lowest R-Square and we need to perform the Standardisation step with selected variables.

**7. CLUSTER DETAILS:**

Code snippet to get cluster details and mean of each cluster w.r.t each variable.

/\*GETTING CLUSTER MEMBERSHIP

VARIABLE IN THE ORIGINAL FILE\*/

/\*Merge Cluster Column with Original Data\*/

**PROC** **SQL**;

CREATE TABLE Cluster.FIN

AS SELECT X.\*,Y.CLUSTER

FROM Cluster.CLUSTER\_DATA X,Cluster.CLOUT Y

WHERE X.CUST\_ID = Y.CUST\_ID;

**QUIT**;

/\*Cluster mean w.r.t each variable\*/

ODS HTML FILE = "C:\SAS\_101\CASE STUDIES\Cluster Analysis Project\ClusterMEAN.XLS";

**PROC** **MEANS** DATA = Cluster.FIN MEAN;

VAR

AVG\_Actual\_price\_12

TOTAL\_DAILY\_ORDER\_NET\_AMT\_12

AVG\_DOLLAR\_DISCOUNT\_12

MONTH\_SINCE\_LAST\_TRANSACTION

sum\_value\_12;

CLASS CLUSTER;

**RUN**;

ODS HTML CLOSE;