ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

Mining Data through Clustering in SAS Studio

Dr. Reena Hooda

Assistant Professor, Department of CSE, Indira Gandhi University Meerpur (Rewari). Haryana, India

ABSTRACT: Clustering is one of the major methods of data mining can be merged with the other methodslike KNNto get the precise results. Natural clustering is unsupervised method where a user has no control over the number of clusters generated. However, a user can make it supervised through the specification of number of clusters required, number of variables on which clustering has to be done and the selection of variables to be included in the output and results of the coding as done in SAS Studio. In a simple way, clustering is the grouping of the objects in such a way that members or the objects of a group are much similar to each other than the members or the objects of other group or class or the partition. The grouping may be done on the basis of distance measure say Euclidean distance between members. The application of clustering is useful in study the data from different perceptions helpful in research or to select the target area for production or marketing etc. The present application is emphasizing on the analyzing the current data through grouping and representation in various ways and applicability of clustering in SAS studio to get the results and user defined outputs via procedural language by means of predefined keywords and variable assignments to create clusters from given data source and their graphical representation.

KEYWORDS*:* Cluster, SAS Studio, Plot, Library, criterion.

I. INTRODUCTION

SAS studio has inbuilt facility to create clusters via various predefined methods and arguments and generated the reports automatically in form of tables and diagram. The data source for the clustering is *mining* data. [3]SAS procedures for clustering arefocused to disjoint or hierarchical clusters from coordinate data, distance data, or a correlation or covariance matrix. [4]Proc Cluster computes Euclidean distance [1] The various methods of clustering in SAS Studio are Average linkage , Median, Single linkage, Density linkage including KNN methods, Centroid, MCQ, MED, Wards methods etc. [1] [5][6][7][8][9] [10]
Data mining includes various methods of clustering like Fuzzy logic, k-Nearest Neighbor Method, Decision trees, Clustering and Neural Networks etc. Clustering is one of the most popular techniques of data mining that includes the grouping of the objects based on some similarity measure and to view the data in different perspectives. Advantage of this technique is that in addition to the grouping of the data, it can also be used as a base with other methods of mining, for instance k-Nearest Neighbor, Fuzzy Logic;all can be implemented after performing clustering. After basic understanding of the SAS mechanics, presentation of the data, the current paper highlighted the coding part to show the methodology to create different clusters and view the data including the pictorial representation.

II. LITERATURE SURVEY

SAS (Statistical Analysis System) is anemergent application with little available literature. Most of the source is the documentation or the SAS support, by which present paper highlighted the simpler way to create clusters and view data from different perspectives. The documentation part and SAS support provide the basic methodology and coding to help the user in creating and maintaining own database or to upload a database from other platform, statically represent the data to meet the user's requirements. User can apply various operations and also able to store as a template for the future use.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

III. CREATING CLUSTERS IN SAS STUDIO AND RESULTS

The present application applied average linkage method of clustering. Before applying the method, the csv mining file has to be imported in SAS Studio, the code also include an output data file work.mining23 stored in the library. If it is already exists it will be overwrite with this new uploaded data file. The code is:

```
proc import datafile="/home/reenah20130/mining.csv"
out=work.mining23 replace;
run;
```

After creating the data source file named *mining23in* the library, the procedure is required to mention the cluster, input data, method of clustering and output data clusters, name of variables are required to be included in the results. The Proc stands for the procedure name cluster where cluster is a predefined procedure in SAS. The user defined procedure is given and results are shown in Fig. 1.The present application uses the average linkage method of clustering as discussed.

```
proc cluster data=work.mining23  noprint  method=average outtree=work.rcluster;
var YEAR        QTR     PRO     POW;
run;
proc cluster data=work.mining23 method=ward ccc pseudo rmsstd print=10outtree=work.rcluster
plots=den(height=rsq);
var YEAR        QTR     PRO     POW;
idid       ; run;
```

The CLUSTER Procedure
Average Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	200.558379	169.635715	0.8507	0.8507
2	30.922665	27.845239	0.1312	0.9819
3	3.077425	1.880043	0.0131	0.9949
4	1.197382		0.0051	1.0000

Root-Mean-Square Total-Sample Standard Deviation	7.677172

Root-Mean-Square Distance Between Observations	21.71432

Cluster History					
Number of Clusters	Clusters Joined		Freq	Norm RMS Distance	Tie
110	OB85	OB86	2	0.0471	
109	OB74	OB75	2	0.0471	
108	OB102	OB103	2	0.0487	
107	OB94	OB95	2	0.0487	
106	OB3	OB4	2	0.0512	
105	OB98	OB99	2	0.0516	
104	OB65	OB66	2	0.055	
103	OB79	OB80	2	0.0596	
102	OB91	OB92	2	0.0608	
101	OB7	OB8	2	0.0617	
100	OB17	OB18	2	0.062	
99	OB70	OB71	2	0.064	
98	OB10	OB11	2	0.0648	
97	OB64	OB67	2	0.0658	
96	OB20	OB24	2	0.0667	
95	OB89	OB90	2	0.0669	
94	OB109	OB110	2	0.0684	
93	OB33	OB34	2	0.0684	
25	CL57	OB12	6	0.2039	
24	OB63	CL36	11	0.2056	
23	CL34	CL44	12	0.2144	
22	OB37	CL26	4	0.2358	
21	CL40	CL25	12	0.2375	
20	CL50	CL27	22	0.238	
19	CL31	CL41	6	0.2386	
18	CL30	CL60	7	0.257	
17	CL32	CL82	8	0.2578	
16	CL20	OB87	23	0.2866	
15	CL19	OB25	7	0.2969	
14	CL37	CL28	8	0.3038	
13	CL17	CL18	15	0.3053	
12	CL22	CL33	11	0.3226	
11	CL24	CL16	34	0.3373	
10	CL21	CL51	18	0.3561	
9	CL23	CL72	15	0.3633	
8	CL12	CL29	13	0.3821	
7	CL15	OB43	8	0.4185	
6	CL11	CL9	49	0.4417	
5	CL13	CL8	28	0.4553	
4	CL7	CL14	16	0.5506	
3	CL4	CL5	44	0.7518	
2	CL3	CL6	93	0.8947	
1	CL10	CL2	111	1.5243	

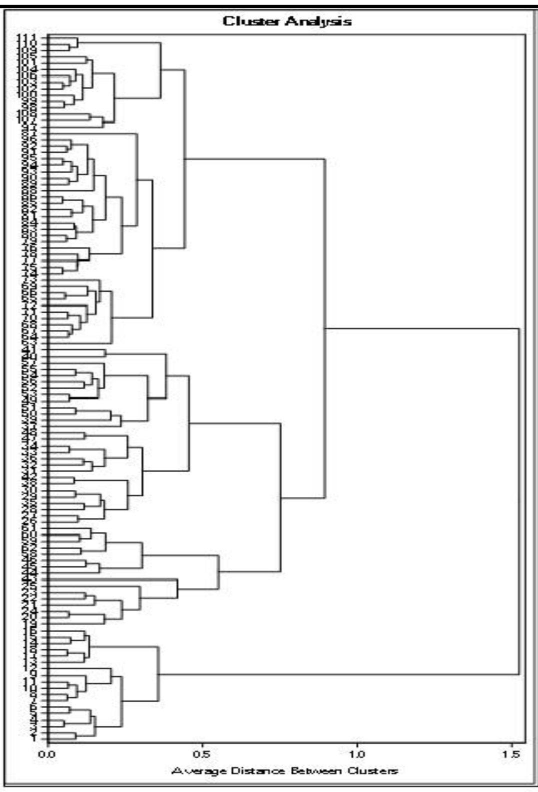

Fig.1: Shows the clusters and graphical representation using Average distance method.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

The output data shows the various columns generated after clustering. Fig. 1 shows the resultant tables covariance matrix, root-mean-square total-sample standard deviation, root-mean-square distance between observations, cluster history like numberof clusters, clusters joined, frequency, norm RMS distance and tie along with the titled graph. The user defined code written for the clustering shows the unsupervised clustering where all the data, number of columns, rows and their graphical representation is automaticallygenerated. A user has to just define the output file name for the clustering that will be stored in the SAS library. The output window as shown in Fig. 3 also shows the name of the output file. The way data is grouped is shown graphically by the Studio that forms a tree shown in Fig. 1. However, SAS Studio provide a flexibilityto the user in output data set is that number of columns can be increased or decreased just deselecting the columns on the left side of the table. Similar way the order of the columns can also be changed in output table, the output of the clustering is shown in Fig. 2. The following code has been generated automatically by SAS Studio to get the output.

```
PROC SQL;
CREATE TABLE WORK.query AS
SELECT _NAME_ , _PARENT_ , _FREQ_ , _HEIGHT_ , _SPRSQ_ , 'YEAR'n , QTR , PRO , POW , _DIST_ ,
_AVLINK_ FROM WORK.RCLUSTER ORDER BY sortkey(_PARENT_, "en_US");
RUN;
QUIT;
PROC DATASETS NOLIST NODETAILS;
CONTENTS DATA=WORK.query OUT=WORK.details;
RUN; PROC PRINT DATA=WORK.details; RUN
```

User can also specified the criterion for the clusters like CCC, PSEUDO, RMSSTD that are useful for estimating the total clusters in the data, here CCC stands for cubic clustering criterion in which values greater than 2 or 3 show good clusters, values between 0 and 2 specify possible clusters and negative values of the CCC may shows outliers. [2] PSEUDO shows pseudo *F* that specifies relatively large values point to good numbers of clusters and to interpret t^2 statistics as in the Fig. 3, find from right to left the first largest value and go back one step upward right, this value define the good clustering level; whereas RMSSTD displays the root mean square standard deviation of each cluster in SAS Studio. [2] print=10 shows the number of clusters will be included in Cluster History. In the code, plots=den(height=rsq) dispalys adendrogram with R square, id is like a primary key in SQL that required unique values in the column mentioned in code. *Id* statement in the code defines id variableas the Y axis variable in the dendrogram and in the output data set rcluster too as shown in Fig. 3.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

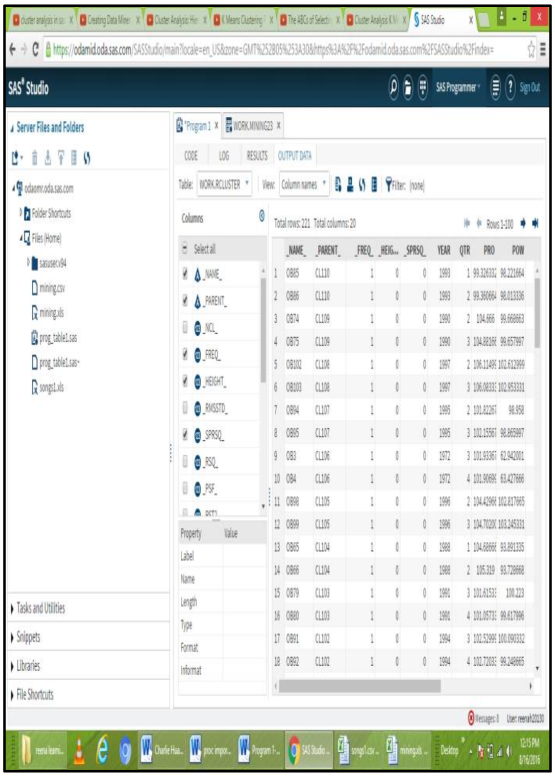

Fig. 2:Output data after Clustering

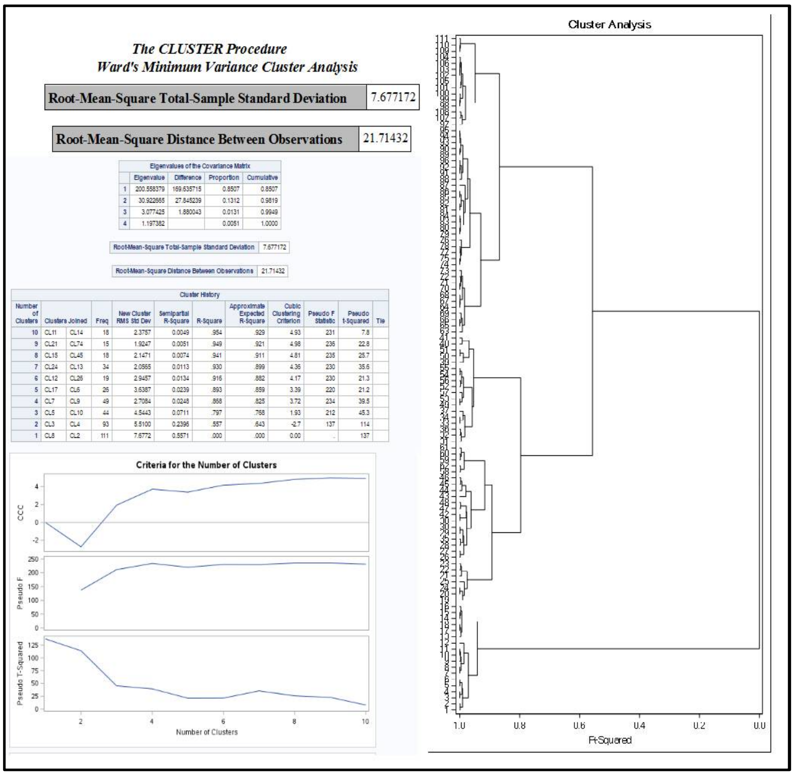

Fig.3: Criteria for the Clusters and Dendrogram using Average Distance Method

To get the flexibility to define the number of clusters required in output data and to avoid the unnecessary clustering a new keyword nclusters has been used to defined the number of clusters i.e. 5 in this program and code is given and output table is shown in Fig.4.

```
proc tree data=work.rclusternclusters=5
out=work.finalcluster;
copy year qtr pro;
run;

proc plot;
plotqtr*pro=cluster;
run;
quit;
```

The resultant table in Fig.4 shows three more columns other than the mention variables in the code. One is a _NAME_ i.e. unique name to each value in resultant table plus two more columns titled CLUSTER and CLUSNAM. The CLUSTER columns contains the maximum value 5 that shows the maximum number of clusters generated and CLUSNAM shows the corresponding clusters names of each group. Plotting of clusters is shown in Fig. 6 for two variables QTR and PRO.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

NAME	YEAR	QTR	PRO	CLUSTER	CLUSNAME
OB85	1993	1	99.32633	1	CL6
OB86	1993	2	99.36066	1	CL6
OB74	1990	2	104.666	1	CL6
OB75	1990	3	104.8817	1	CL6
OB102	1997	2	106.115	1	CL6
OB103	1997	3	106.0833	1	CL6
OB94	1995	2	101.8227	1	CL6
OB95	1995	3	102.1557	1	CL6
OB3	1972	3	101.9337	2	CL10
OB4	1972	4	101.907	2	CL10
OB98	1996	2	104.4297	1	CL6
OB99	1996	3	104.702	1	CL6
OB65	1988	1	104.6867	1	CL6
OB66	1988	2	105.319	1	CL6
OB79	1991	3	101.6153	1	CL6
OB80	1991	4	101.0573	1	CL6
OB91	1994	3	102.53	1	CL6
OB92	1994	4	102.7203	1	CL6
OB7	1973	3	103.4387	2	CL10
OB8	1973	4	103.5537	2	CL10
OB17	1976	1	99.716	2	CL10
OB18	1976	2	99.62434	2	CL10
OB70	1989	2	104.3623	1	CL6
OB71	1989	3	103.4327	1	CL6
OB10	1974	2	102.8033	2	CL10
OB11	1974	3	102.0883	2	CL10
OB64	1987	4	104.8687	1	CL6
OB67	1988	3	104.945	1	CL6
OB20	1976	4	101.602	3	CL7
OB24	1977	4	102.4227	3	CL7
OB89	1994	1	101.9497	1	CL6
OB90	1994	2	102.6707	1	CL6
OB109	1999	1	97.61633	1	CL6
OB110	1999	2	97.08	1	CL6
OB33	1980	1	112.9473	4	CL5
OB34	1980	2	111.8667	4	CL5
OB49	1984	1	112.273	4	CL5
OB53	1985	1	111.4307	4	CL5
OB5	1973	1	101.158	2	CL10
OB6	1973	2	100.9233	2	CL10
OB106	1998	2	105.2387	1	CL6
OB14	1975	2	99.31533	2	CL10
OB15	1975	3	98.14267	2	CL10
OB96	1995	4	101.8843	1	CL6
OB81	1992	1	99.733	1	CL6
OB82	1992	2	100.1437	1	CL6
OB93	1995	1	102.404	1	CL6
OB68	1988	4	103.8717	1	CL6
OB100	1996	4	103.7727	1	CL6
OB83	1992	3	99.89733	1	CL6
OB84	1992	4	100.2263	1	CL6
OB38	1981	2	111.6513	4	CL5
OB42	1982	2	113.063	4	CL5
OB29	1979	1	106.4057	4	CL5
OB30	1979	2	107.9507	4	CL5
OB104	1997	4	105.6997	1	CL6
OB54	1985	2	112.1543	4	CL5
OB55	1985	3	110.5113	4	CL5
OB50	1984	2	115.0793	4	CL5
OB51	1984	3	116.263	4	CL5
OB1	1972	1	100.3657	2	CL10
OB2	1972	2	101.1983	2	CL10
OB77	1991	1	104.4583	1	CL6
OB111	1999	3	98.09567	1	CL6
OB78	1991	2	103.169	1	CL6
OB26	1978	2	109.755	4	CL5
OB27	1978	3	109.564	4	CL5
OB59	1986	3	99.502	5	CL14
OB60	1986	4	99.994	5	CL14
OB58	1986	2	102.088	5	CL14
OB62	1987	2	100.79	5	CL14
OB32	1979	4	110.149	4	CL5
OB36	1980	4	111.5467	4	CL5
OB28	1978	4	109.3247	4	CL5
OB35	1980	3	109.596	4	CL5
OB52	1984	4	111.8573	4	CL5
OB56	1985	4	109.737	4	CL5
OB13	1975	1	100.7713	2	CL10
OB47	1983	3	107.716	4	CL5
OB48	1983	4	109.7503	4	CL5
OB16	1975	4	99.765	2	CL10
OB22	1977	2	104.8183	3	CL7
OB23	1977	3	104.4	3	CL7
OB9	1974	1	103.0503	2	CL10
OB45	1983	1	104.8507	5	CL14

Fig. 5: Resultant 5 Clusters.

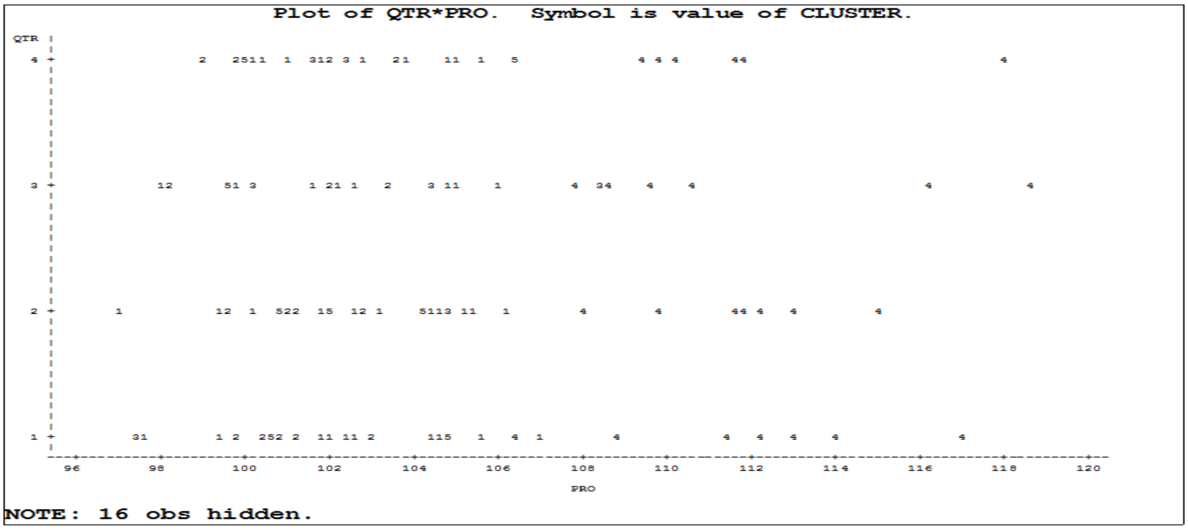

Fig.6: Plotting of QTR and PRO.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 4, Issue 12, December 2016

IV. CONCLUSION

Creating clusters on the basis of different variables through suitable methods like average, centroid, wards, and median etc. as well as finding the correlation between data say Pearson Correlation coefficient, provides a great flexibility to user to view the data and gain knowledge about the patterns in the data quickly, with minimum level of coding, basic knowledge by way of SAS Studio. One can perform analysis without looking for the outside expert or software support, data can be categorized, copied for future use, reports can be generated including graphical dispersal of the data. The number of clusters can be increased or decreasedjust similar to the variables as done in the current application where mining.csv data is classified with natural clustering including leveled clustering to view data with different outlooks and to show the correlation between variable like PRO and POW. Further in present application input 5 specified the user defined number of clusters for comprising the variables QTR, YEAR and PRO and created with their unique names and grouping of values into different partitions and are plotted using PROC Plot. The main advantage of clustering in SAS studio is that the criterion for the clusters can be specified viaCCC, PSEUDO, RMSSTDto estimate the good number of clusters, possible clusters, outliers and clustering level. The further scope of the work may roll towards the discovery of hidden values to aid in estimating unknown values using the SAS Studio.

REFERENCES

1. https://support.sas.com/rnd/app/stat/procedures/cluster.html
2. https://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_cluster_gettingstarted.htm
3. http://www.principlesofeconometrics.com/excel.htm
4. SAS Documentation, SAS/STAT® 9.2 User's Guide Introduction to Clustering Procedures. https://support.sas.com/documentation/cdl/en/statugclustering/61759/PDF/default/statugclustering.pdf
5. https://en.wikipedia.org/wiki/SAS_(software)
6. http://support.sas.com/training/tutorial/studio/create-table-csv-file.html
7. http://support.sas.com/software/products/sasstudio/
8. http://www2.sas.com/proceedings/sugi31/099-31.pdf
9. http://support.sas.com/documentation/onlinedoc/sasstudio/
10. Michael A. Monaco, Marie Dexter, Jennifer Tamburro, 'Introduction to SAS® Studio'. Paper SAS302-2014. Retrieved from: http://support.sas.com/resources/papers/proceedings14/SAS302-2014.pdf

BIOGRAPHY

Reena Hooda is Assistant Professor in the Department of Computer Science & Engineering, Indira Gandhi University Meerpur (Rewari) Haryana, India. She received herMaster of Computer Application (MCA) degree in 2005,MBA in 2007 and Ph.D in 2012 from MDU Rohtak, Haryana (India). Her research interests are DBMS, Computer Networks and Data Mining