CLUSTERING

Introduction

Cluster analysis is mainly used for segmentation. Cluster Analysis represents finding similarities between data on the basis of the characteristics found in the data and grouping similar data objects into clusters. It is an unsupervised learning technique (No dependent variable).

Quality of Clustering

A good clustering method produces high quality clusters with **minimum intra-cluster distance** (high similarity within the cluster) and **maximum inter-class distance** (low similarity between two clusters).

Data Preparation before cluster analysis.

- Adequate Sample Size
- Remove outliers

1. *Adequate Sample Size*
   Sufficient size is needed to ensure representativeness of the population and its underlying structure, particularly small groups within the population.
   Minimum group sizes are based on the relevance of each group to the research question and the confidence needed in characterizing that group.

2. *Remove outliers / Percentile Capping*
   Outliers are observations that fall below Q1 - 1.5(IQR) or above Q3 + 1.5(IQR). Here,
   IQR = Q3 - Q1. Another method to handle outliers is to cap large values at 99th percentile.

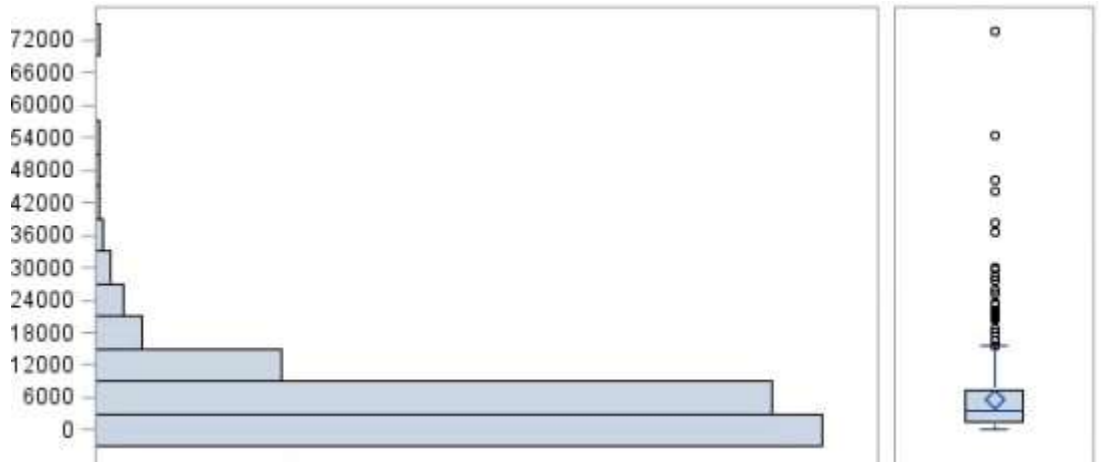
The "wholesale_customers.xls" data will be used. The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

Attribute Information:

1) FRESH: annual spending (m.u.) on fresh products (Continuous);
2) MILK: annual spending (m.u.) on milk products (Continuous);
3) GROCERY: annual spending (m.u.) on grocery products (Continuous);
4) FROZEN: annual spending (m.u.) on frozen products (Continuous)
5) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products(Continuous)
6) REGION: 3 Regions coded as 1,2,3(nominal)

Our aim is to segment customer based on their spending on 3 items named Milk, Detergents paper, Frozen different items.

Before performing clustering, outliers must be removed, else clustering may get biased results. There are many criteria for selecting outliers. Detecting outliers from Boxplots is one of them. As we are using only 3 variables for clustering, we will delete outliers from these 3 variables. Below is the boxplot for milk variable. There are so many outlier's above upper fence. We will delete outliers which are above upper fence (approximately 15000 here)



Similarly, for frozen and detergents_paper, delete values above 8000 and 10000 respectively.

Output Interpretation:

Both the MAXCLUSTERS= and MAXITER= options are set in the PROC FASTCLUS statement. MAXITIER corresponds to maximum number of iterations to find the optimal centroid.

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=5 Maxiter=20 Converge=0.0

Initial Seeds			
Cluster	Frozen	Milk	Detergents_Paper
1	950.00000	577.00000	4762.00000
2	529.00000	8323.00000	93.00000
3	662.00000	14982.00000	3891.00000
4	7849.00000	2209.00000	210.00000
5	937.00000	7677.00000	9836.00000

Minimum Distance Between Initial Seeds = 7667.123

Below displays the number of observations in each cluster (frequency) and the root mean squared standard deviation. The next two columns display the largest Euclidean distance from the cluster seed to any observation within the cluster and the number of the nearest cluster. The last column of the table displays the distance between the centroid of the nearest cluster and the centroid of the current cluster. A centroid is the point having coordinates that are the means of all the observations in the cluster. The pseudo ***F*** statistic, approximate expected overall R square, and cubic clustering criterion (CCC) are listed at the bottom of the figure. You can compare values of these statistics by running PROC FASTCLUS with different values for the MAXCLUSTERS=n.

Values of the cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters, but they should be taken with caution; large negative values can indicate outliers. Here, CCC value is 8.915 which indicates that these are good clusters.

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	170	1013.8	4265.8		4	3821.1
2	51	1614.1	6335.0		5	4163.8
3	32	1556.4	4540.0		5	5323.5
4	60	1262.4	5257.4		1	3821.1
5	51	1450.6	5318.8		2	4163.8

Statistics for Variables

Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
Frozen	1806	1155	0.594998	1.469124
Milk	3425	1452	0.822368	4.629630
Detergents_Paper	2451	1179	0.771147	3.369612
OVER-ALL	2646	1269	0.772419	3.394039

Pseudo F Statistic =	304.61
Approximate Expected Over-All R-Squared =	0.70715
Cubic Clustering Criterion =	8.915

Cluster Means

Cluster	Frozen	Milk	Detergents_Paper
1	1282.70588	1977.98824	746.56471
2	1617.96078	7309.07843	1497.80392
3	1501.59375	11540.87500	6542.68750
4	5089.13333	2301.08333	660.20000
5	1246.05882	6321.68627	5525.68627

Cluster Standard Deviations

Cluster	Frozen	Milk	Detergents_Paper
1	835.056145	1191.137167	983.617944
2	1609.277154	1973.993534	1153.118346
3	1164.202284	1545.167070	1877.376746
4	1376.593911	1511.050833	776.212879
5	1249.928693	1504.557430	1576.856639

- The final table indicate the means and standard deviations of variables in each cluster.

It is useful to study the clusters further. One method is to look at a frequency tabulation of the clusters with other classification variables. We cross tabulate the empirical clusters with the variable *region*. This method enables to study clusters in different regions.

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Region by CLUSTER						
		CLUSTER(Cluster)					
	Region(Region)	1	2	3	4	5	Total
	1	25 6.87 37.31 14.71	13 3.57 19.40 25.49	5 1.37 7.46 15.63	16 4.40 23.88 26.67	8 2.20 11.94 15.69	67 18.41
	2	19 5.22 54.29 11.18	1 0.27 2.86 1.96	6 1.65 17.14 18.75	6 1.65 17.14 10.00	3 0.82 8.57 5.88	35 9.62
	3	126 34.62 48.09 74.12	37 10.16 14.12 72.55	21 5.77 8.02 65.63	38 10.44 14.50 63.33	40 10.99 15.27 78.43	262 71.98
	Total	170 46.70	51 14.01	32 8.79	60 16.48	51 14.01	364 100.00