

Cluster Analysis

Objective of Cluster analysis is to combine observations into groups or clusters such that groups formed are homogeneous (similar) within the group and heterogeneous (different) from other groups on some variables.

Applications

- Brand manager is interested in identifying groups of customers which are alike with respect to buying pattern.
- Researcher is interested in identifying stores that can be used for test marketing purpose.
- Investor would like to identify groups of stocks that have similar returns while others that have different pattern of returns.
- Tourism managers needs to know which media markets are likely to bring “best” return for their efforts.

Data Requirement

- Measurement on two or more variables which are used for clustering purpose.
- All variables have similar measurement scales and at least have ordinal characteristics.
- If measured variables are nominal scaled (e.g. brand choice), then similarity measures based on nominal association should be used.

• Steps in Cluster Analysis

1. Decide on measure of dissimilarity or similarity.
2. Decide on technique to combine observations.
3. Method to confirm clusters.
4. Interpretation of Clusters.

- **Measure of dissimilarity or similarity**

- The Euclidean distance (D_{ij}) between two points (i and j) with p dimension is squared difference on each dimension and then summed and taken square root. Formally,

$$D_{ij} = \left[\sum_{k=1}^p (X_{ik} - X_{jk})^2 \right]^{\frac{1}{2}}.$$

Euclidean distance is a special case of a more general metric called Minkowski and it is given by

$$D_{ij} = \left[\sum_{k=1}^p (|X_{ik} - X_{jk}|)^n \right]^{\frac{1}{n}},$$

and $| \cdot |$ are used to denote absolute difference. When n is 1, such distance measure is called city block distance metric.

- Canberra metric measure computes absolute difference between two observations and then divides by total distance. That is,

$$D_{ij} = \sqrt{\sum_{k=1}^p \left(\frac{|X_{ik} - X_{jk}|}{X_{ik} + X_{jk}} \right)^2}.$$

- The Mahalanobis distance (M_{ij}) is computed by pre-and post-multiplying by difference between two observations to the covariance matrix for the sample. That is,

$$M_{ij} = \left[\sum_{l=1}^p \sum_{k=1}^p (X_{ik} - X_{jk})(\mathbf{S}_{kl})^{-1}(X_{il} - X_{jl}) \right]^{\frac{1}{2}}.$$

- Measures of association, such as correlation coefficient or Spearman rank order correlation. For example, correlation coefficient between two observations i and j and over a set of p variables is given by

$$A_{ij} = \frac{\sum_{k=1}^p (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{\sqrt{\left(\sum_{k=1}^p (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^p (X_{kj} - \bar{X}_j)^2 \right)}}.$$

Unlike other distance measures, A_{ij} is bounded between -1 and 1 and it is considered similarity measure. That is, the higher the value, more similar are two individuals.

- When all the variables take 0 and 1 values, we can create several similarity indices. Consider two individuals i and j and both take values 0 and 1 for various variables. Then 2×2 table with cell frequencies can be used to construct various similarity measures.

		Individual i		
		1	0	
Individual j	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

In this table a is number times both individuals have value of 1. Similarly b is number times individual i had zero value while individual j had value of 1.

Following are some of similarity indices proposed in the literature.

Simple matching	$\frac{a + d}{a + b + c + d}$
Czekanowski, Sorensen and Dice	$\frac{2a}{2a + b + c}$
Hamman	$\frac{a + d - (b + c)}{a + b + c + d}$
Phi or Correlation	$\frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$
Russell and Rao	$\frac{a}{a + b + c + d}$
Rogers and Tanimoto	$\frac{a + d}{a + 2b + 2c + d}$
Yule	$\frac{ad - bc}{ad + bc}$
Ochiai	$\frac{a}{\sqrt{(a + b)(a + c)}}$

- **Methods to combine observations.**

There are many methods to combine observations. We will review five such methods below.

- In the *Centroid* method, each individual or group is replaced by an *average* of two or more observations.

- The *Ward's* method tries to minimize the total within-cluster sums of squares from the mid-point of two observations or clusters.
 - The Average-linkage method the distance between two clusters is obtained by taking the *average distance* between all pairs of observations in the two clusters.
 - The Nearest-Neighbour or Single linkage method requires to compute the distance between two clusters as the *minimum* of the distance between all possible pairs of observations in the two clusters.
 - The Farthest-Neighbour or Complete linkage method requires to compute the distance between two clusters as the *maximum* of the distance between all possible pairs of observations in the two clusters.
- **Method to confirm clusters.**
 - *Hierarchical* Cluster Solutions depend upon statistical procedure to suggest number of clusters, while
 - *Nonhierarchical* Cluster solutions require one to pre-specify number clusters with their centroids.
 - **Interpretation of Clusters.**
 - Reliability or consistency of clusters.
 - External validity of clusters.
 - Predictive validity of clusters.

Illustrative Example

Following observations were obtained from *Advertising Age's* homepage. It contained advertising expenditure for major brands in the U. S. For illustrative purpose, I have selected nine movie studios. We will first investigate centroid method and then Ward's method. Note that Following Table indicates that movie studios use television (TV) as their primary advertising medium while secondary medium is either newspaper (**news**) or magazines (**MAG**). Outdoor (**OUTDOOR**) and and radio (**RADIO**) account for less than 2% of advertising spending. Note that Disney with its three studios account for \$254 millions in spending which slightly less than third of total spending by all studios combined.

Movie Studios and Their Advertising Spending

Brand	Ownership	Total Spending in (\$'000)	% Spending Allocated to					Code
			Magazine	Newspaper	Outdoor	TV	Radio	
Buena Vista	Disney	145,419	3.59%	15.93%	0.28%	78.92%	1.28%	M1
Columbia	Sony	138,434	21.26%	12.75%	1.08%	63.92%	0.98%	M2
Paramount	Viacom	100,168	1.90%	20.09%	0.98%	75.93%	1.11%	M3
Universal	Seagram	93,580	2.42%	15.46%	1.33%	77.72%	3.06%	M4
Warner	Time Warner	92,228	8.04%	22.64%	0.16%	68.00%	1.16%	M5
20th Century Fox	News Corp.	72,583	1.61%	21.26%	0.41%	76.57%	0.14%	M6
Disney	Disney	59,527	21.22%	3.98%	1.92%	71.07%	1.80%	M7
Miramax	Disney	49,260	1.28%	40.00%	0.00%	57.87%	0.85%	M8
Tri-Star	Sony	39,811	0.03%	23.93%	0.22%	74.97%	0.85%	M9

Let us compute dissimilarity or distance matrix using Euclidean metric. That is to compute distance between Buena Vista and Columbia (D_{12}), we would calculate following:

$$\begin{aligned}
 D_{12} &= [(.0359 - 0.2126)^2 + (0.1593 - 0.1275)^2 + (0.0028 - 0.0108)^2 \\
 &\quad + (0.7892 - 0.6392)^2 + (0.0128 - 0.0098)^2]^{\frac{1}{2}} \\
 &= \sqrt{0.0312 + 0.0010 + 0.00006 + 0.0225 + 0.000009} \\
 &= 0.2341
 \end{aligned}$$

We continue this process for all pairs and obtain following distance matrix. Intuitively larger the distance between two studios, farther apart they would be in terms of advertising spending allocation.

Distance between Pair of Movie Studios									
Studio	Buena	Columbia	Paramount	Universal	Warner	Fox	Disney	Miramax	Tri-Star
Buena	0.0000								
Columbia	0.2341	0.0000							
Paramount	0.0544	0.2394	0.0000						
Universal	0.0270	0.2360	0.0537	0.0000					
Warner	0.1357	0.1703	0.1039	0.1351	0.0000				
Fox	0.0626	0.2489	0.0177	0.0671	0.1085	0.0000			
Disney	0.2276	0.1137	0.2564	0.2305	0.2313	0.2680	0.0000		
Miramax	0.3207	0.3435	0.2692	0.3169	0.2121	0.2649	0.4329	0.0000	
Tri-Star	0.0962	0.2644	0.0446	0.0955	0.1071	0.0357	0.2943	0.2351	0.0000

Since Paramount and 20th Century Fox are very similar in their advertising budget allocation (or the lowest distance among available distances), we would form first cluster by grouping these two observations. After combining these observation, we would have one less observation. This would result in following revised observations (I have included clustered observations for reference purpose).

Revised Data after forming Cluster 1					
MAG	NEWSP	OUTDOOR	TV	RADIO	Brands
0.0161	0.2126	0.0041	0.7657	0.0014	20th Century Fox
0.0190	0.2009	0.0098	0.7593	0.0111	Paramount
0.0176	0.2067	0.0070	0.7625	0.0063	Cluster 1
0.0359	0.1593	0.0028	0.7892	0.0128	Buena Vista
0.2126	0.1275	0.0108	0.6392	0.0098	Columbia
0.0242	0.1546	0.0133	0.7772	0.0306	Universal
0.0804	0.2264	0.0016	0.6800	0.0116	Warner
0.2122	0.0398	0.0192	0.7107	0.0180	Disney
0.0128	0.4000	0.0000	0.5787	0.0085	Miramax
0.0003	0.2393	0.0022	0.7497	0.0085	Tri-Star

We repeat computation of distance matrix for this revised data matrix. Of course now we will

have one less observation. A resulting distance matrix is represented below.

	Cluster 1	Buena	Columbia	Universal	Warner	Disney	Miramax	Tri-Star
Cluster 1	0.0000							
Buena	0.0579	0.0000						
Columbia	0.2440	0.2341	0.0000					
Universal	0.0601	0.0270	0.2360	0.0000				
Warner	0.1058	0.1357	0.1703	0.1351	0.0000			
Disney	0.2621	0.2276	0.1137	0.2305	0.2313	0.0000		
Miramax	0.2669	0.3207	0.3435	0.3169	0.2121	0.4329	0.0000	
Tri-Star	0.0394	0.0962	0.2644	0.0955	0.1071	0.2943	0.2351	0.0000

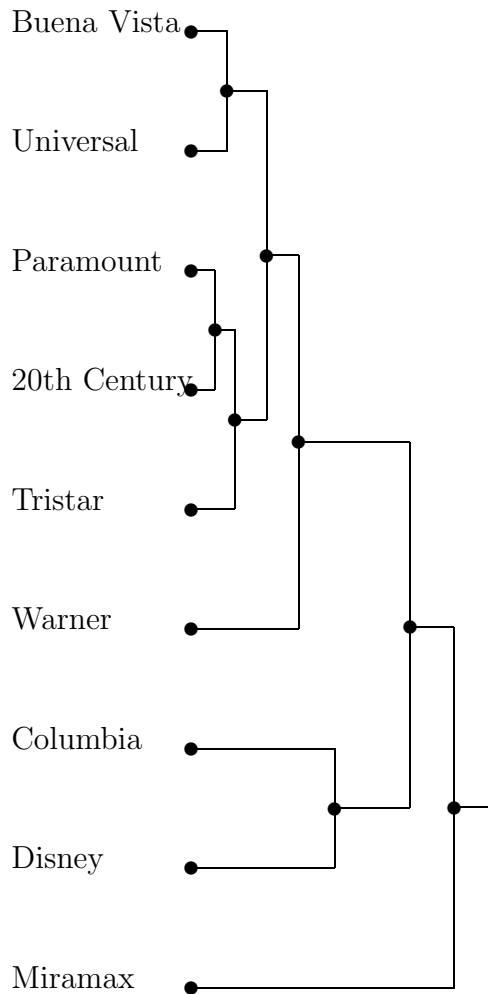
Among the paired distances, note that Universal and Buena Vista has the lowest distance. Our second cluster will be formed by combining these two studios. Revised data after combining these two observations is shown below.

MAG	NEWSP	OUTDOOR	TV	RADIO	Brands
0.0161	0.2126	0.0041	0.7657	0.0014	20th Century Fox
0.0190	0.2009	0.0098	0.7593	0.0111	Paramount
0.0176	0.2067	0.0070	0.7625	0.0063	Cluster 1
0.0359	0.1593	0.0028	0.7892	0.0128	Buena Vista
0.0242	0.1546	0.0133	0.7772	0.0306	Universal
0.0300	0.1570	0.0081	0.7832	0.0217	Cluster 2
0.2126	0.1275	0.0108	0.6392	0.0098	Columbia
0.0804	0.2264	0.0016	0.6800	0.0116	Warner
0.2122	0.0398	0.0192	0.7107	0.0180	Disney
0.0128	0.4000	0.0000	0.5787	0.0085	Miramax
0.0003	0.2393	0.0022	0.7497	0.0085	Tri-Star

Note that at this stage we have 7 observations (5 original observations and 2 clusters). We repeat distance computation and obtain 7×7 distance matrix.

	Cluster 1	Cluster 2	Columbia	Warner	Disney	Miramax	Tri-Star
Cluster 1	0.0000						
Cluster 2	0.0634	0.0000					
Columbia	0.2489	0.2347	0.0000				
Warner	0.1085	0.1348	0.1703	0.0000			
Disney	0.2680	0.2287	0.1137	0.2313	0.0000		
Miramax	0.2649	0.3185	0.3435	0.2121	0.4329	0.0000	
Tri-Star	0.0357	0.0949	0.2644	0.1071	0.2943	0.2351	0.0000

At this stage, we would add Tri-star to Cluster 1 or Paramount, 20th Century Fox and Tri-star would form cluster 3. Revised data at this stage would have two clusters (one with three members and another with two members) and four observations do not belong to any groups.



Revised Data after forming Cluster 3					
MAG	NEWSP	OUTDOOR	TV	RADIO	Brands
0.0161	0.2126	0.0041	0.7657	0.0014	20th Century Fox
0.0190	0.2009	0.0098	0.7593	0.0111	Paramount
0.0003	0.2393	0.0022	0.7497	0.0085	Tri-Star
0.0118	0.2176	0.0054	0.7582	0.0070	Cluster 3
0.0359	0.1593	0.0028	0.7892	0.0128	Buena Vista
0.0242	0.1546	0.0133	0.7772	0.0306	Universal
0.0300	0.1570	0.0081	0.7832	0.0217	Cluster 2
0.2126	0.1275	0.0108	0.6392	0.0098	Columbia
0.0804	0.2264	0.0016	0.6800	0.0116	Warner
0.2122	0.0398	0.0192	0.7107	0.0180	Disney
0.0128	0.4000	0.0000	0.5787	0.0085	Miramax

Continuing these two steps, computing distances and then finding two observations with minimum distance, then computing average for newly formed cluster would eventually form following “tree” like structure. Note that cluster formed by Buena Vista, Universal, Paramount, 20th Century Fox, Tri-star spend their three fourths of budget on television advertising and remaining one fourth of money is spent on newspaper advertising. Disney and Columbia are similar and spend relatively higher proportion of money of magazine advertising. Warner does not fit in either groups but it is somewhat closer to the bigger group. Miramax is unique, it allocates its advertising between television and newspaper with relatively higher allocation to newspaper (40%).

Clustering of Movie Studios by Centroid Method

SAS Input to Compute Centroid Method of Clustering

```

data movad ;
input mag news outdoor tv radio total brand $ Owner $;
datalines;
0.0359 0.1593 0.0028 0.7892 0.0128 145418.90 Buena_Vista Disney
0.2126 0.1275 0.0108 0.6392 0.0098 138434.20 Columbia Sony
0.0190 0.2009 0.0098 0.7593 0.0111 100168.00 Paramount Viacom
0.0242 0.1546 0.0133 0.7772 0.0306 93580.10 Universal Seagram
0.0804 0.2264 0.0016 0.6800 0.0116 92228.00 Warner Time_Warner
0.0161 0.2126 0.0041 0.7657 0.0014 72583.40 20th_Century_Fox News_Corp
0.2122 0.0398 0.0192 0.7107 0.0180 59526.90 Disney Disney
0.0128 0.4000 0.0000 0.5787 0.0085 49260.00 Miramax Disney
0.0003 0.2393 0.0022 0.7497 0.0085 39811.20 Tri-Star Sony
;;;

```

```
proc cluster method=centroid outtree=cendata nonorm rmsstd data=movad ;
var mag news outdoor tv radio;
id brand;
copy total;
run;
proc tree data=cendata;
run;
```

SAS output

Root-Mean-Square Total-Sample Standard Deviation = 0.066316

Number of Clusters	--Clusters Joined--	Frequency of New Cluster	RMS of New Cluster	STD of New Cluster	Centroid Distance	Tie
8	Paramoun 20th_Cen	2	0.005593	0.017687		
7	Buena_Vi Universa	2	0.008544	0.027020		
6	CL8 Tri-Star	3	0.010903	0.039351		
5	CL7 CL6	5	0.019214	0.069700		
4	CL5 Warner	6	0.026782	0.112510		
3	Columbia Disney	2	0.035974	0.113761		
2	CL4 CL3	8	0.054506	0.230331		
1	CL2 Miramax	9	0.066316	0.284464		

Name of Observation or Cluster						
B	U	P	2	T	C	M
u	n	a	0	r	o	
e	i	r	t	i	W	l
n	v	a	h	-	a	u
a	e	m	-	S	r	m
-	r	o	C	t	n	b
V	s	u	e	a	e	i
i	a	n	n	r	r	a
						y
						x

0.3 +	XX	.
	XX	.
	XX	.
	XX	.
D	XX	.
i	+XX	.
0.25	XX	.
s	XX	.
t	XX	.
a	XX	.
n	XX	.
c	XX	XXXXXX
e	XX	XXXXXX
	XX	XXXXXX
B	+XX	XXXXXX
0.2	XX	XXXXXX
e	XX	XXXXXX
t	XX	XXXXXX
w	XX	XXXXXX
e	XX	XXXXXX
e	XX	XXXXXX
n	XX	XXXXXX
	XX	XXXXXX
C	+XX	XXXXXX
0.15	XX	XXXXXX
l	XX	XXXXXX
u	XX	XXXXXX
s	XX	XXXXXX
t	XX	XXXXXX
e	XX	XXXXXX
r	XX	XXXXXX
	XX	.
C	+XX	.
0.1	XX	.
e	XX	.
n	XX	.
t	XX	.
r	XX	.
o	XX	.
i	XXXXXXXXXX XXXXXXXXXXXXXXXX	.
d	XXXXXXXXXX XXXXXXXXXXXXXXXX	.
s	+XXXXXXXXXX XXXXXXXXXXXXXXXX	.
0.05	XXXXXXXXXX XXXXXXXXXXXXXXXX	.
	XXXXXXXXXX XXXXXXXXXXXXXXXX	.
	XXXXXXXXXX XXXXXXXX	.
	XXXXXXXXXX XXXXXXXX	.
	. . XXXXXXXX	.

0 +.

Ward's Method of Clustering

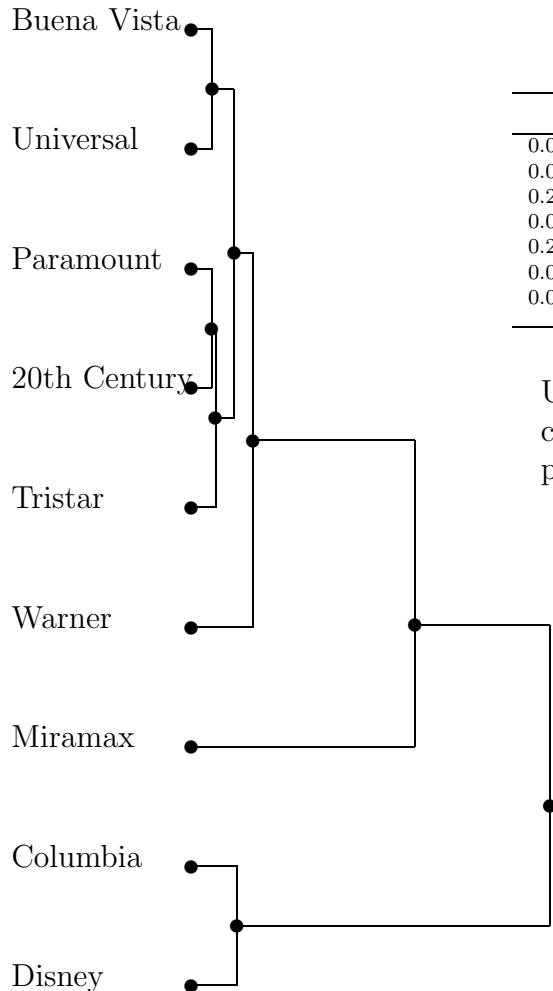
For this method, we need to compute average for each pair of observations and then compute between cluster sums of squares. Following table illustrates various paired sums of square calculations.

Means						
MAG	NEWSP	OUTDOOR	TV	RADIO	Brands	
0.1242	0.1434	0.0068	0.7142	0.0113	0.0274	M1 & M2
0.0274	0.1801	0.0063	0.7742	0.0119	0.0015	M1 & M3
0.0300	0.1570	0.0081	0.7832	0.0217	0.0004	M1 & M4 ²
0.0581	0.1929	0.0022	0.7346	0.0122	0.0092	M1 & M5
0.0260	0.1860	0.0035	0.7774	0.0071	0.0020	M1 & M6
0.1240	0.0996	0.0110	0.7499	0.0154	0.0259	M1 & M7
0.0243	0.2797	0.0014	0.6839	0.0107	0.0514	M1 & M8
0.0181	0.1993	0.0025	0.7694	0.0106	0.0046	M1 & M9
0.1158	0.1642	0.0103	0.6993	0.0105	0.0287	M2 & M3
0.1184	0.1411	0.0121	0.7082	0.0202	0.0279	M2 & M4
0.1465	0.1770	0.0062	0.6596	0.0107	0.0145	M2 & M5
0.1144	0.1701	0.0075	0.7024	0.0056	0.0310	M2 & M6
0.2124	0.0837	0.0150	0.6750	0.0139	0.0065	M2 & M7
0.1127	0.2638	0.0054	0.6089	0.0092	0.0590	M2 & M8
0.1064	0.1834	0.0065	0.6945	0.0092	0.0349	M2 & M9
0.0216	0.1777	0.0116	0.7682	0.0208	0.0014	M3 & M4
0.0497	0.2136	0.0057	0.7196	0.0113	0.0054	M3 & M5
0.0176	0.2067	0.0070	0.7625	0.0063	0.0002	M3 & M6 ¹
0.1156	0.1203	0.0145	0.7350	0.0146	0.0329	M3 & M7
0.0159	0.3005	0.0049	0.6690	0.0098	0.0362	M3 & M8
0.0096	0.2201	0.0060	0.7545	0.0098	0.0010	M3 & M9
0.0523	0.1905	0.0075	0.7286	0.0211	0.0091	M4 & M5
0.0202	0.1836	0.0087	0.7714	0.0160	0.0022	M4 & M6
0.1182	0.0972	0.0163	0.7440	0.0243	0.0266	M4 & M7
0.0185	0.2773	0.0067	0.6779	0.0196	0.0502	M4 & M8
0.0122	0.1970	0.0078	0.7635	0.0195	0.0046	M4 & M9
0.0483	0.2195	0.0029	0.7228	0.0065	0.0059	M5 & M6
0.1463	0.1331	0.0104	0.6953	0.0148	0.0267	M5 & M7
0.0466	0.3132	0.0008	0.6293	0.0101	0.0225	M5 & M8
0.0403	0.2329	0.0019	0.7148	0.0101	0.0057	M5 & M9
0.1142	0.1262	0.0117	0.7382	0.0097	0.0359	M6 & M7
0.0145	0.3063	0.0021	0.6722	0.0050	0.0351	M6 & M8
0.0082	0.2260	0.0032	0.7577	0.0050	0.0006	M6 & M9
0.1125	0.2199	0.0096	0.6447	0.0133	0.0937	M7 & M8
0.1062	0.1396	0.0107	0.7302	0.0133	0.0433	M7 & M9
0.0065	0.3197	0.0011	0.6642	0.0085	0.0276	M8 & M9

¹ Cluster 1 members

² Cluster 2 members

Since sums of squares are the lowest for M3 (Paramount) and M6 (20th Century Fox), first cluster will formed by combining these two observations. Note that second lowest sums of squares is for M1 (Buena Vista) and M4 (Universal) second cluster will be formed combining these two observations. By forming these two clusters, we have reduced observation set by 2. We would repeat process of computing means and sums of squares for those reduced set of observations.



Revised Data after forming Clusters 1 & 2

MAG	NEWSP	OUTDOOR	TV	RADIO	Brands
0.0176	0.2067	0.0070	0.7625	0.0063	Cluster 1
0.0300	0.1570	0.0081	0.7832	0.0217	Cluster 2
0.2126	0.1275	0.0108	0.6392	0.0098	Columbia
0.0804	0.2264	0.0016	0.6800	0.0116	Warner
0.2122	0.0398	0.0192	0.7107	0.0180	Disney
0.0128	0.4000	0.0000	0.5787	0.0085	Miramax
0.0003	0.2393	0.0022	0.7497	0.0085	Tri-Star

Using these seven observations, we would compute combinations of means and sums of squares for each possible combination (see table below).

Means

MAG	NEWSP	OUTDOOR	TV	RADIO	Sum of Sq	Brands
0.0238	0.1819	0.0075	0.7728	0.0140	0.0038	M3, M6 & M1, M4
0.0891	0.1626	0.0078	0.7292	0.0112	0.0412	M3, M6 & M2
0.0385	0.2133	0.0052	0.7350	0.0080	0.0076	M3, M6 & M5
0.0842	0.1357	0.0122	0.7512	0.0167	0.0469	M3, M6 & M7
0.0160	0.2712	0.0046	0.7012	0.0070	0.0476	M3, M6 & M8
0.0118	0.2176	0.0054	0.7582	0.0070	0.0012	M3, M6 & M9 ³
0.0909	0.1472	0.0090	0.7352	0.0178	0.0371	M1, M4 & M2
0.0468	0.1801	0.0059	0.7488	0.0183	0.0125	M1, M4 & M5
0.0908	0.1179	0.0118	0.7590	0.0205	0.0352	M1, M4 & M7
0.0243	0.2380	0.0054	0.7150	0.0173	0.0680	M1, M4 & M8
0.0201	0.1844	0.0061	0.7720	0.0173	0.0064	M1, M4 & M9
0.1465	0.1770	0.0062	0.6596	0.0107	0.0145	M2 & M5
0.2124	0.0837	0.0150	0.6750	0.0139	0.0065	M2 & M7
0.1127	0.2638	0.0054	0.6089	0.0092	0.0590	M2 & M8
0.1064	0.1834	0.0065	0.6945	0.0092	0.0349	M2 & M9
0.1463	0.1331	0.0104	0.6953	0.0148	0.0267	M5 & M7
0.0466	0.3132	0.0008	0.6293	0.0101	0.0225	M5 & M8
0.0403	0.2329	0.0019	0.7148	0.0101	0.0057	M5 & M9
0.1125	0.2199	0.0096	0.6447	0.0133	0.0937	M7 & M8
0.1062	0.1396	0.0107	0.7302	0.0133	0.0433	M7 & M9
0.0065	0.3197	0.0011	0.6642	0.0085	0.0276	M8 & M9

³ Cluster 3 members

Tree Diagram for Movie Studios by Ward's Method

Note that cluster 3 will be formed by combining adding M9 (Tri-star) to cluster 1 (Buena Vista and 20th Century Fox). Moreover, we also could form cluster 4 by

combining M2 (Columbia) and M7 (Disney). We could continue this process and combine all observations into one cluster. The resulting tree structure is presented below.

SAS Output from Ward's Method of Clustering

Use Method=ward in PROC Cluster to obtain Ward's method.

Ward's Minimum Variance Cluster Analysis
 Root-Mean-Square Total-Sample Standard Deviation = 0.066316

NCL	-Clusters Joined-	FREQ	RMS STD	SPRSQ	RSQ	BSS	T i e
8	Paramoun 20th_Cen	2	0.00559	0.000889	0.99911	0.000156	
7	Buena_Vi Universa	2	0.00854	0.002075	0.99704	0.000365	
6	CL8 Tri-Star	3	0.01090	0.005869	0.99117	0.001032	
5	CL7 CL6	5	0.01921	0.033140	0.95803	0.005830	
4	Columbia Disney	2	0.03597	0.036784	0.92124	0.006471	
3	CL5 Warner	6	0.02678	0.059967	0.86128	0.010549	


```

options ls=75 ps=60 nocenter nodate;
data recog;
input id @4 (a1-a5) (5*1.) @10 (b1-b6) (6*1.) @17 (c1-c7) (7*1.)
      @25 (d1-d4) (4*1.) @30 (e1-e5) (5*1.) group;
datalines;
09 00100 001000 1111111 1111 00100 1
11 00000 011000 0101111 0111 00000 1
10 00010 000000 1111011 1111 00000 1
07 00000 000000 1111111 1101 00000 1
12 10000 100000 1111111 1111 00000 1
08 00000 001100 0111111 1011 00000 1
13 00000 000000 0111111 0111 00000 1

06 11110 111011 0000010 0000 00000 2
03 11111 110111 0000000 0000 01000 2
04 11111 011111 0010000 0000 00000 2
02 11111 111111 0000100 0000 00000 2
01 10111 111111 0000000 0001 00000 2
05 10111 111111 0000000 0000 10000 2

15 00000 000010 1000000 1111 01111 3
18 10000 000000 0100000 0111 11111 3
14 00010 100000 0100000 1111 11111 3
19 00000 100010 0100000 0111 11111 3
16 00100 000000 0000010 1111 11111 3
17 00000 010000 0000000 1011 11111 3

20 11111 000000 0001000 1111 11111 4
22 11111 000000 0001000 1011 11101 4
29 11101 000000 0000000 1111 00011 4
21 11011 000001 0000000 1111 11111 4
23 01111 000000 0010000 1111 10111 4
24 11011 000000 0000000 1111 10111 4

25 11111 100100 0000000 1111 10000 5
30 10111 100000 0000000 1110 00000 5
28 11111 000000 1000000 1111 00000 5
26 11111 000000 0001000 1111 00000 5
27 01111 000000 0000000 1111 10000 5
;;;

/* compute distance matrix */

data distjacc(type=distance);
  array dj(*) dj1-dj30;
  retain dj1-dj30 .;          /* initialize to missing values */
  do row=1 to 30;             /* loop over rows of distance matrix */
    set recog point=row;     /* read row resp */
    array indi(*) a1-a5 b1-b6
              c1-c7 d1-d4 e1-e5; /* declare arrays after */
    array indj(*) var1-var27;
    do g=1 to 27;             /* Need to use same obs. */
      indj(g)=indi(g);
    end;
    do col=1 to row;         /* loop over columns of distance matrix */
      set recog(drop=id) point=col;
      a = 0;                  /* Compute a, b, c, and d, First initialize */
      b = 0;
      c = 0;
      d = 0;
      do g=1 to 27;
        if indi(g) eq 1 and indj(g) eq 1 then a = a + 1;

```

```

        if indi(g) eq 1 and indj(g) eq 0 then b = b + 1;
        if indi(g) eq 0 and indj(g) eq 1 then c = c + 1;
        if indi(g) eq 0 and indj(g) eq 0 then d = d + 1;
    end;
/* convert to distance Measure and Then Dissimilarity - Correlation */
/*      dj(col) = (-a*d + b*c)/sqrt((a + b)*(a + c)*(b + d)*(c + d)); */
/* Czekanowski et al Similarity Measure */
/*      dj(col) = 1 - (2*a)/(2*a + b + c); */
/* Simple Matching */
      dj(col) = 1 - (a + d)/(a + b + c + d);
    end;
    output;          /* output a row of the distance matrix */
end;
stop;
/* stop statement is needed because set statement
      uses point= option */
keep id dj1-dj30;   /* keep only the id and distance matrix */
run;
proc print data=distjacc(obs=10);
    id id; var dj1-dj10;
    title2 'First 10 obs';
run;
title2;

proc cluster data=distjacc method=ward
    outtree=tree nosquare nonorm ;
id id;
var dj1-dj30;
run;

proc tree data=tree n=5 out=out noprint;
id id;
run;
/*          Combine Datasets to Crosstab results */
proc sort data=out; /* Sort output dataset by id */
by id;
run;

proc sort data = recog; /* Sort original dataset */
by id;

data clus;
merge recog out;
by id;
/* Do cross tabulation to check groups */
proc freq;
tables cluster*group / nopercnt nocol norow;
run;

```

Partial SAS Output

Let us look at distance matrix based on observations with id less than or equal to 10. In this instance we are using similarity measure based on “Simple matching”.

Dissimilarity for 10 observations

ID	1	2	3	4	5	6	7	8	9	10
1	0.000									
2	0.111	0.000								
3	0.148	0.111	0.000							
4	0.148	0.111	0.148	0.000						
5	0.074	0.111	0.148	0.148	0.000					
6	0.185	0.148	0.185	0.185	0.185	0.000				
7	0.704	0.741	0.778	0.704	0.778	0.667	0.000			
8	0.593	0.630	0.741	0.593	0.667	0.630	0.185	0.000		
9	0.704	0.741	0.852	0.704	0.778	0.667	0.148	0.185	0.000	
10	0.667	0.778	0.741	0.667	0.741	0.630	0.111	0.222	0.185	0.000

Note that first six observation belong to group 2 and the last four observations in above table belong to group 1 and not surprising degree of similarity (dissimilarity) is higher (lower) within the groups and similarity (dissimilarity) is lower (higher) between the groups. Similar pattern is evident when correlation is used as measure of dissimilarity.

Ward's Minimum Variance Cluster Analysis

NCL	---Clusters	Joined---	Group	FREQ	SPRSQ	RSQ	BSS	T i e
29			2	2	0.005833	0.99417	0.037037	T 1 5
28			4	2	0.005833	0.98833	0.037037	T 20 22
27			4	2	0.005833	0.98250	0.037037	T 21 24
26			5	2	0.005833	0.97667	0.037037	T 26 28
25			2	2	0.008750	0.96792	0.055556	T 2 3
24			1	2	0.008750	0.95917	0.055556	T 7 10
23			1	2	0.008750	0.95042	0.055556	T 11 13
22			3	2	0.008750	0.94167	0.055556	T 14 19
21			5	2	0.008750	0.93292	0.055556	T 25 27
20 CL25			2	3	0.010694	0.92222	0.067901	T 4
19 CL24			1	3	0.010694	0.91153	0.067901	T 12
18 CL22			3	3	0.010694	0.90084	0.067901	18
17			3	2	0.011666	0.88917	0.074074	16 17
16 CL21	CL26		5	4	0.013125	0.87605	0.083333	
15 CL27			4	3	0.013611	0.86243	0.086420	23
*14 CL16			5	5	0.014291	0.84814	0.090741	30
13 CL29	CL20		2	5	0.014388	0.83375	0.091358	
12			1	2	0.014583	0.81917	0.092593	T 8 9
11	CL17		3	3	0.015555	0.80362	0.098765	15
*10 CL13			2	6	0.016722	0.78689	0.106173	6
9 CL15			4	4	0.017013	0.76988	0.108025	29
8 CL12	CL23		1	4	0.017500	0.75238	0.111111	
*7 CL28	CL9		4	6	0.020902	0.73148	0.132716	
*6 CL19	CL8		1	7	0.023055	0.70842	0.146384	
*5 CL18	CL11		6	6	0.024305	0.68412	0.154321	
4 CL7	CL14		11	0.053895	0.63022	0.342200		
3 CL5	CL4		17	0.119247	0.51098	0.757147		
2 CL6	CL3		24	0.244742	0.26624	1.553960		
1 CL10	CL2		30	0.266236	0.00000	1.690432		

If we closely examined the last of clusters, we would find that “Simple matching” and Ward’s method does produce 100% correct classification.

Cluster Summary

Cluster	Included individuals	Original Groups
5	18, 14, 19, 16, 17 and 15.	3
6	12, 8, 9, 11, 13, 7, and 10.	1
7	20, 22, 29, 23, 21 and 24.	5
10	30, 25, 27, 26 and 28	4
14	6, 4, 2, 3, 1, and 5	2

Centroid Hierarchical Cluster Analysis

Number of Clusters	-----Clusters Joined-----	Group	Frequency of New Cluster	Squared Centroid Distance	Tie	
29	1	5	(2)	2	0.074074	T
28	20	22	(4)	2	0.074074	T
27	21	24	(4)	2	0.074074	T
26	26	28	(5)	2	0.074074	
25	CL29	2	(2)	3	0.092593	T
24	CL26	27	(5)	3	0.092593	T
23	CL25	3	(2)	4	0.102881	T
22	CL23	4	(2)	5	0.094907	
21	25	CL24	(5)	4	0.102881	
20	7	10	(1)	2	0.111111	T
19	CL20	12	(1)	3	0.101852	T
18	CL19	13	(1)	4	0.094650	
17	14	19	(3)	2	0.111111	T
16	CL17	18	(3)	3	0.101852	
15	CL28	CL27	(4)	4	0.111111	
14	CL21	30	(5)	5	0.113426	
13	CL15	23	(4)	5	0.120370	
12	CL18	9	(1)	5	0.127315	
*11	CL22	6	(2)	6	0.127407	
*10	CL14	29	(5)	6	0.131852	Classified incorrectly
9	CL13	CL10		11	0.129095	Difficult to interpret
8	CL12	8	(1)	6	0.133333	
*7	CL8	11	(1)	7	0.141975	
6	16	17	(3)	2	0.148148	
5	CL16	CL6	(3)	5	0.119342	
*4	CL5	15	(3)	6	0.131852	
3	CL4	CL9		17	0.195023	
2	CL7	CL3		24	0.313404	
1	CL11	CL2		30	0.352173	

When we used centroid method to combine observations, we classified 80% observations correctly. Our error rate is much higher than we would like to and it is because cluster 9 is combination of group 4 and 5. In the table on the next page, we indicate effectiveness of various measures of similarity and various means of grouping observations in their ability to classify observations in five groups.

Cluster Summary

Cluster	Included individuals	Original Groups
4	15, 16, 17, 18, 14 and 19.	3
7	11, 8, 9, 13, 12, 10, and 7.	1
9	29, 30, 25, 27, 26 and 28. 23, 21, 24, 20 and 22	5 & 4
11	6, 4, 2, 3, 2, and 1	2

**Effect of Clustering Algorithm and
Similarity Measures on Classification Accuracy**

Clustering Algorithm	Similarity Measure			
	Correlation	Simple Matching	Czekanowski et al.	Russell and Rao
Single	24	23	24	24
Ward's	30	30	29	29
Centroid	30	24	25	26
Average	30	29	30	30
Complete	30	29	25	26
Median	30	24	25	26

In each cell, we have 30 observations
to be classified in five groups

Based on above dataset results, we would conclude that Ward's and the average methods perform comparatively well in classifying observations. It also cautions that it might be good practice to conduct cluster analysis using one or more algorithms as well as using one or more dissimilarity measures, since with real datasets, we do not know group memberships.

NAME: 1993 New Car Data
TYPE: Sample
SIZE: 93 observations, 26 variables

DESCRIPTIVE ABSTRACT:
Specifications are given for 93 new car models for the 1993 year.
Several measures are given to evaluate price, mpg ratings, engine size,
body size, and features.

SOURCES:
Consumer Reports: The 1993 Cars Annual Auto Issue (April 1993),
Yonkers, NY: Consumers Union.
PACE New Car & Truck 1993 Buying Guide (1993), Milwaukee, WI: Pace
Publications Inc.

VARIABLE DESCRIPTIONS:
Columns
1 14 Manufacturer
15 29 Model
30 36 Type
 Small, Sporty, Compact, Midsize, Large as defined in the
 Consumer Reports article
38 41 Minimum Price (in \$1,000) Price for basic version of this model
43 46 Midrange Price (in \$1,000) Average of Min and Max prices
48 51 Maximum Price (in \$1,000) Price for a premium version
53 54 City MPG (miles per gallon by EPA rating)
56 57 Highway MPG
59 59 Air Bags standard
 0 = none, 1 = driver only, 2 = driver & passenger
61 61 Drive train type

```

        0 = rear wheel drive
        1 = front wheel drive
        2 = all wheel drive
63 63 Number of cylinders
65 67 Engine size (liters)
69 71 Horsepower (maximum)
73 76 RPM (revs per minute at maximum horsepower)
77 80 Engine revolutions per mile (in highest gear)
82 82 Manual transmission available
        0 = No, 1 = Yes
84 87 Fuel tank capacity (gallons)
89 89 Passenger capacity (persons)
91 93 Length (inches)
95 97 Wheelbase (inches)
99 100 Width (inches)
102103 Uturn space (feet)
105108 Rear seat room (inches)
110111 Luggage capacity (cu. ft.)
113117 Weight (pounds)
119119 Domestic?
        0 = nonU.S. manufacturer, 1 = U.S. manufacturer

```

Values are aligned and delimited by blanks.
Missing values are denoted with *.

SPECIAL NOTES:

The only missing values are for CYLINDERS in the rotary engine Mazda RX7, REAR SEAT room for the twoseaters (Corvette and RX7), and LUGGAGE capacity for the vans and twoseaters.

WEIGHT is taken from the _Consumer Reports_ data and includes a full fuel tank, automatic transmission (if available), and air conditioning.

STORY BEHIND THE DATA:

Cars were selected at random from among 1993 passenger car models that were listed in both the _Consumer Reports_ issue and the _PACE Buying Guide_. Pickup trucks and Sport/Utility vehicles were eliminated due to incomplete information in the _Consumer Reports_ source. Duplicate models (e.g., Dodge Shadow and Plymouth Sundance) were listed at most once.

SAS Input for Automobile Example

```

options nocenter nodate ps = 70 ls =80 nonumber ;
title1 "Hierarchical cluster analysis of New Cars of 1993";
data newcar;
  infile "a:\newcar.dat";
input  manuf $ 19 model $ 1525 type $ 3036 minprc midprc maxprc citympg highmpg
      air_bag drvtrn numcyl engsize hrsepwr rpmmax rpmgear trans
      tankcap passcap length wheel width u_turn rearseat lugcap weight
      domest;
/*  Commands for hierarchical clustering                               */
proc cluster noprint method=ward nonorm out=tree;
id model ;
  var numcyl engsize hrsepwr rpmmax rpmgear tankcap passcap length
      wheel width u_turn rearseat weight ;
run;
/*  Proc Tree is used to retrieve 4cluster solution                   */
proc tree data=tree out=clus4 nclusters=4 noprint;
id model;

```

```

copy numcyl engsize hrsepwr rpmmx rpmgear tankcap passcap length
  wheel width u_turn rearseat weight ;
run;
/* Use Proc FREQ to get cluster sizes */
proc freq data=clus4;
tables cluster;
run;

/* Obtain Cluster means using Proc Tabulate */
proc tabulate data=clus4;
class cluster;
var numcyl engsize hrsepwr rpmmx rpmgear tankcap passcap length
  wheel width u_turn rearseat weight;
table (numcyl engsize hrsepwr rpmmx rpmgear tankcap passcap length
  wheel width u_turn rearseat weight)*(mean std),cluster;
run;

```

SAS Output

Hierarchical cluster analysis of New Cars of 1993

CLUSTER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	25	27.5	25	27.5
2	28	30.8	53	58.2
3	17	18.7	70	76.9
4	21	23.1	91	100.0

Frequency Missing = 2

Cluster Means

Variable	1	2	3	4
NUMCYL	6.20	3.96	4.00	5.48
ENGSIZE	3.72	1.75	2.13	3.00
HRSEPWR	163.28	108.18	103.59	188.05
RPMMAX	4600.00	5760.71	5029.41	5609.52
RPMGEAR	1820.80	2813.04	2469.41	2231.19
TANKCAP	19.25	13.63	15.24	18.47
PASSCAP	6.08	4.54	5.00	5.00
LENGTH	195.88	170.07	180.29	188.86
WHEEL	111.12	97.71	101.88	106.14
WIDTH	73.76	65.79	67.88	69.95
U_TURN	42.12	35.89	38.71	39.38
REARSEAT	30.00	26.43	26.65	28.07
WEIGHT	3687.00	2467.50	2741.47	3411.19
Examples	Crown Victoria	Tercel	Cavalier	Volvo
Size	Big cars or minivans	Small cars	Compact or mid size	Mid size
Cylinders	6 or 8	4	4	4 or 6