1. **INTRODUCTION**

In 1974, data were collected by *Motor Trend* magazine for 32 models of cars produced in 1973 and 1974. The car make and models ranged from a standard Honda Civic to a Maserati with over 250 horsepower. If data grouping is possible by make and model, results will provide additional insight useful for deeper analysis, such as optimizing price bands or developing marketing campaigns for certain types of vehicles.

Cluster analysis is useful to determine what features or attributes of the vehicles are similar and group each vehicle accordingly. Further, a successful cluster analysis that yields meaningful groups may be used as a predictive tool since it is cognitively easier to determine outcomes from items with similar properties.

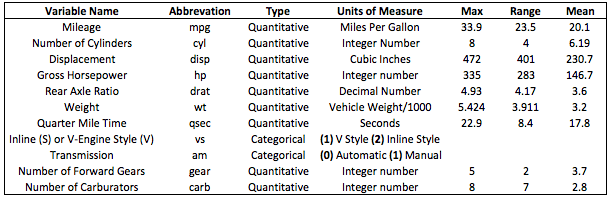
Cluster algorithms are a form of data analysis that seeks to find underlying structure in the data than may not be obvious by other methods. The dataset in this report will be subjected to a variety of cluster methods to determine the optimal number of meaningful clusters, compare the outcomes and synergy between different methods and interpret the meaning and value of the clusters with regard to the dataset. The results of the analysis could be used for many different purposes, ranging from marketing to manufacturing planning.

1. **EXPLORATORY DATA ANALYSIS**

The *Motor Trend* data set includes 32 observations uniquely identified by the make and model of a car in 1973 and 1974. Popular vehicles like the Toyota Corolla and Honda Civic, luxury vehicles such as the Lincoln Continental and Cadillac Fleetwood, and exotic vehicles like the Maserati Bora are all part of the dataset.

Associated with each observation are eleven mixed type variables to be used for cluster analysis. Full exposition to the variables can be found in Table 1:

**Table #1** - Variable Listing

****

Both quantitative variables and categorical factors are considered in the analysis. The integrity of the cluster analysis should be compromised not lead to false conclusions, even though Euclidean distance, our choice for a similarity metric, works best with quantitative variables.

Due to the diversity of makes and models in the *Motor Trend* dataset, clusters are likely to be distinct and easy to detect. For instance, before performing the analysis, it is anticipated that the Honda Civic will be in a different cluster than the Maserati Bora. These are very different vehicles and the cluster analysis should confirm this or provide a correlation that is not obvious that reasonably brings them together.

1. **ASSUMPTIONS AND PREREQUISITES**

Unlike many statistical methods, no prior knowledge of data groupings or structure is assumed in cluster analysis. However, it is assumed there is *potential* structure in the data. Therefore, analysis for all motor trend vehicles seeks to find clusters of vehicles based on the previously mentioned variables. In addition, based on a small number of observations (n=32), the goal is to find an optimal number of clusters to effectively interpret the membership in each cluster. Because of the small sample size, we seek no more than five clusters to describe groupings of the data.

Further, the range of all variables is not egregiously different, however, measurements are on different scales. For instance, number of cylinders ranges from four to eight, while gross horsepower tops out at 335. In order to truly re-scale, data would need to be different by orders of magnitude. However, displacement is measured in cubic inches while horsepower is measured in number of horses (an integer). Therefore, in an effort to be thorough and account for any scale differences, standardized variables will be employed to account for unit measurement differences across variables.

1. **CLUSTER ANALYSIS**

Without any pre-defined knowledge or training data, we wish to find structure in the *Motor Trend* cars data to group car makes and models and interpret the groupings in a meaningful way. In order to find an optimal number of clusters, multiple algorithmic approaches will be considered on a standardized *Motor Trend* data set with mean of 1 and standard deviation of 0. Both hierarchical and partitioning clustering methods will be utilized to discover the optimal number of clusters that best represents the data and underlying structure.

Criteria for judging clustering effectiveness include the pseudo-F statistic, pseudo-t statistic, the cubic clustering criteria (CCC) and eigenvalues from the correlation or covariance matrix. Euclidean distance is utilized as the standard distance measurement in each of the hierarchical clustering algorithms and centroid distances in partitioning methods. Further, only agglomerative methods are considered for hierarchical clustering.

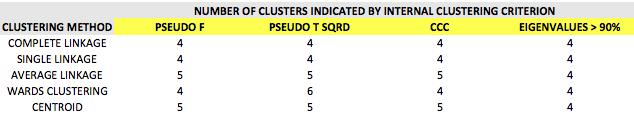
The ultimate goal of the analysis is to find a consensus result for the minimum number of clusters required to interpret data structure. Only significant variables, as a result of a one-way ANOVA on each cluster, will be interpreted.

* 1. **CLUSTER SELECTION**

Table 2 presents summary results from four different clustering criteria: The Pseudo F, Pseudo T-Squared, Cubic Clustering Criterion and the eigenvalues from the correlation matrix associated with each clustering method.

For clarification, the Pseudo T-Squared criterion represents an index quantifying the difference between two merged clusters. Peaks in the Pseudo F statistic ensure mean vectors of the clusters are different, indicating good separation of clusters. The goal is to minimize the Pseudo T-squared (minimize the distance between merged clusters), while maximizing the Pseudo F-Squared [1, 2]. Further, maximizing the Cubic Clustering Criterion and finding an acceptable level of correlation or covariance explained by eigenvalues associated with each cluster can also help determine cluster counts. In this case, for cumulative variation or correlation explained by eigenvalues, 90 percent explanation is the cut off. In each clustering method in table 2, four clusters explain 90 percent of total correlation or total covariance.

**Table 2 – Cluster Criteria Results**



While there is no definitive measure or rule for determining the best number of clusters, the numerical clustering criteria are helpful in pointing to a reasonable number of optimal clusters. If the internal clustering criterion tend to agree for one or more methodologies, and visual inspection via dendrogram show compact branches and a long trunk for cluster groupings, the clustering solution is likely a good one. Further, too many clusters will lead to difficulty in interpretation of individual clusters. Therefore, in order to minimize the number of clusters on a small data set such as the *Motor Trend* cars dataset and ensure we are consistent with the results from clustering criterion, four clusters will be used.

Other internal clustering criteria not used include the Root Mean Square Standard Deviation (RMSSTD) and **R**2. These criteria are measures of standard deviation and homogeneity in the clusters and these values tend to move predictably with the number of clusters. For example, **R**2 will increase as the number of clusters increases because variance within each cluster can more easily be explained with tighter, but smaller groupings, and this effect is analogous to adding more predictor variables to a linear regression equation. A high **R**2 squared indicating better homogeneity within the clusters is not necessarily meaningful if it if hides the underlying relationships in the data set by involving too many clusters. Balance and reason from the researcher must prevail to capture a “best fit”.

The RMSSTD parameter gives some insight about the within cluster standard deviation. However, as our data is standardized to account for scale differences, the RMSSTD is not practical in this situation.

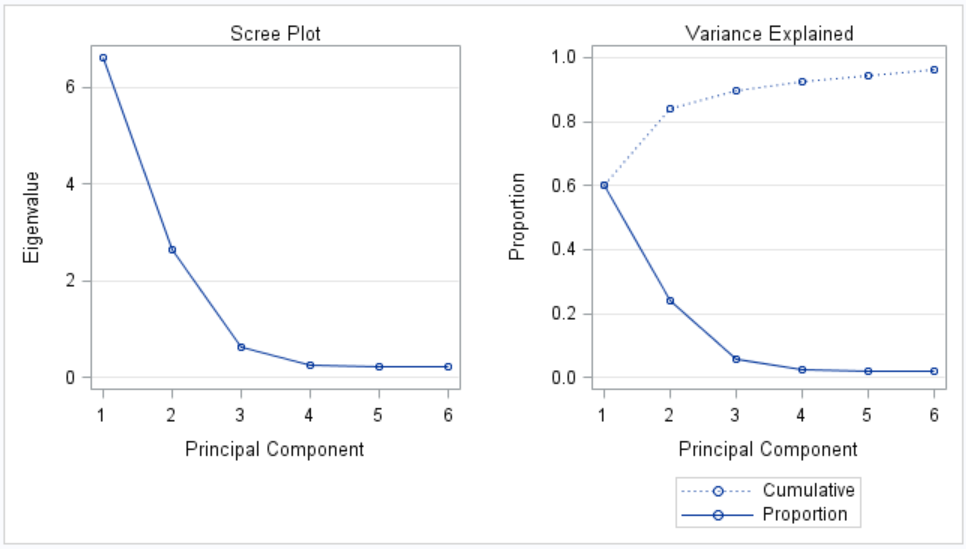
Both Ward’s Method and Complete Linkage result in acceptable dendrograms, with short branches defining distinct clusters and long trunks, while maximizing Pseudo F statistics, indicating good clustering:

|  |  |
| --- | --- |
| **COMPLETE LINKAGE** | **WARD’S METHOD** |
| **F: 27.3**  Macintosh HD:Users:patrickcorynichols:Desktop:Screen Shot 2015-12-10 at 4.52.54 PM.png | **F: 27.3**  Macintosh HD:Users:patrickcorynichols:Desktop:Screen Shot 2015-12-10 at 4.52.00 PM.png |



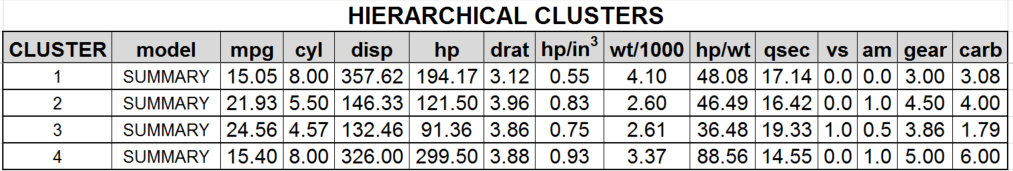
Visual analysis and clustering criteria lead to the decision that four clusters are adequate and meaningful for this data set. Because distinct, short branch clusters are apparent with Ward’s method, additional interpretation will be executed on clusters associated with Ward’s method.

Additionally, a principal component analysis was performed for the purposes of comparing the number of components to the number of clusters as a back check to the cluster analysis. The scree plot (Fig 2) shows that four components will explain 92.3% (>90%) of the variation and while the component membership is slightly different, the use of four components to adequately represent the data set affirms that four clusters is also likely the right number for this data set.



1. **CLUSTER INTERPRETATION UNDER HIERARCHICAL MODEL**

Cluster analysis is intended to show correlations between the members of a data set that may not otherwise be obvious. Simply ending the analysis at finding the number of clusters is inadequate. Further interpretation of the cluster makeup is required to truly make the analysis valuable. To understand what variables are critical to the cluster analysis for the *Motor Trend* cars dataset and determine the makeup of each cluster, a One-Way ANOVA is executed with the clusters obtained from hierarchical analysis as the grouping or class variable. Statistical significance of each variable is critical in determining the makeup of each cluster. All variables show significance at an experiment-wise error rate of 0.0045. Appendix Figure 1 shows individual p-values for each variable as well as the detailed membership in each cluster.

Summary data for the clusters:

In reviewing the summary data of the clusters from the hierarchical algorithm, it becomes easy to see the theme of each cluster.

**Cluster 1** is characterized by large engines (disp), high horsepower, greater weight and fewer gears in the transmission. All of these factors, contribute to the low mileage in this group. The car models that are members of this group fit observations.

**Cluster 2** is made up of RX4’s, a Ferrari and a Porsche all of which are peppy sports cars that are light weight (about 50% of cluster 1) and they get more horsepower per cubic inch of displacement and they have more gears (4.5 avg) to make better use of available power. More gears usually means that the mileage is improved because the engine experiences a smaller operating speed range. These cars are fun to drive and have reasonable acceleration as evidenced by the elapsed times for the quarter mile.

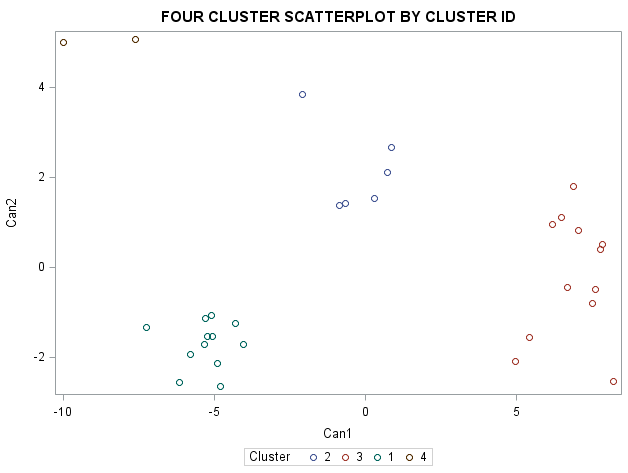
**Cluster 3** is the economy group. These cars have the highest mileage, smallest engines, lowest horsepower and the slowest times for the quarter mile. The car models in this group are the ones known for fuel economy.

**Cluster 4** is a racer group. The horsepower to weight ratio is nearly double the next highest group, the engines are large and the horsepower is high. All ingredients for top end cars. The Ford Pantera and Maserati are well known super cars.

**5.1)** **CLUSTER INTERPRETATION WITH K-MEANS**

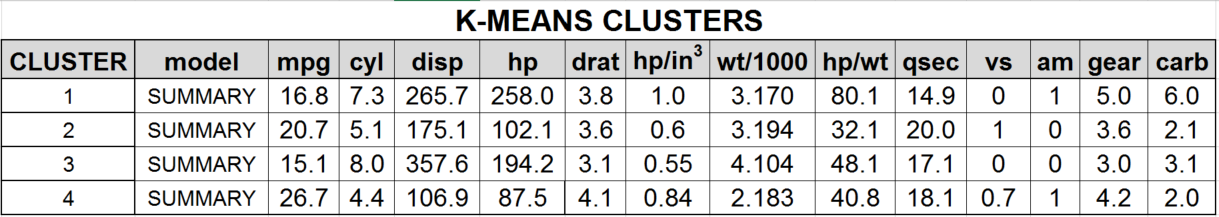
Utilizing the standardized *Motor Trend* cars dataset, a K-Means partitioning analysis is carried out on the *Motor Trend* cars dataset with four and five clusters identified prior to analysis.

Based on 1000 repetitions, four clusters show a superior Pseudo F (27) and higher R-Squared (0.75) than five clusters. More importantly, four clusters will let us interpret the clusters more appropriately. Plotting the four clusters using a random selection for cluster means results in a very distinct set of clusters seen in figure 3:





Summary data for the K-Means clusters:



In reviewing the summary data of the clusters from the K-Means algorithm, it becomes easy to see the theme of each cluster.

**Cluster 1** is a racer group. The horsepower to weigh ratio is nearly double the next highest group, quarter mile elapsed times are the lowest, the engines are large, the horsepower is high and the transmission is loaded with lots of gears to optimize engine power. All ingredients for top end cars. The Ford Pantera, Masarati and Ferrari are well known super cars.

**Cluster 2** is characterized by midsize sedans with good mileage (second highest) and modest engine sizes (disp). Quarter mile elapsed times are the slowest of the four cluster and come with automatic transmissions. These are family cars and this is consistent with observations.

**Cluster 3** is characterized by large engines (disp), high horsepower, greater weight and fewer gears in the transmission. All of these factors, contribute to the low mileage in this group. The car models that are members of this group fit observations and this group is very similar to cluster 1 that was assembled in the hierarchical algorithm.

**Cluster 4** is the economy group. These cars have the highest mileage, smallest engines, lowest horsepower and the slowest times for the quarter mile. The car models in this group are the ones known for fuel economy.

1. **CONCLUSION:**

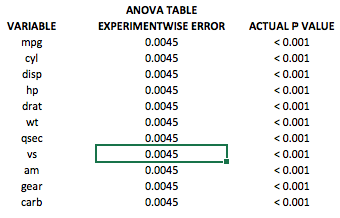
For the Motor Trend cars data set, two clustering algorithms were applied and while the membership of each cluster varies slightly the themes and rationale for the groupings is similar. Mileage, engine displacement and horsepower seem to have evolved as the variables that were detected as the most dominant by the two algorithms. The Ferrari was included with the Ford Pantera and Masarati in the K-Means method and inspection of the detailed membership tables in the appendix, reveals that the Ferrari has attributes that place between two clusters. A case for each assignment can be made but not with a great deal of conviction.

The aforementioned example makes the point that the researcher would be well served to apply several cluster methods to a data set to improve knowledge about the data, understand the salient points that may have been used for a particular cluster assignment and evaluate the reasonableness of the results with regard to the research question. Cluster analysis is very good tool for exploratory data analysis whose results can influence the selection of more precise statistical methods and assist in understanding of those outcomes.

For future study, other variables (such as price or number of passengers) could be added to better detect patterns in the data and sharpen the cluster assignments.

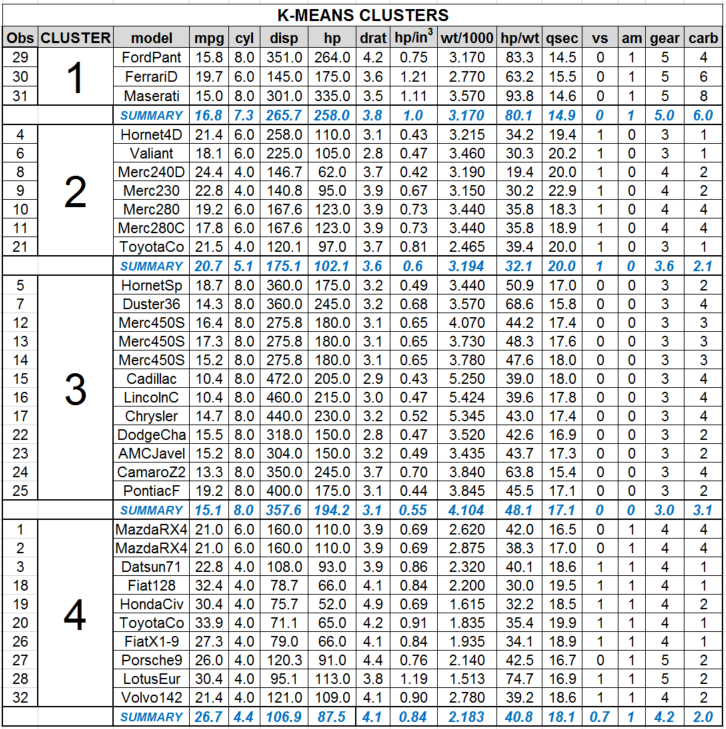
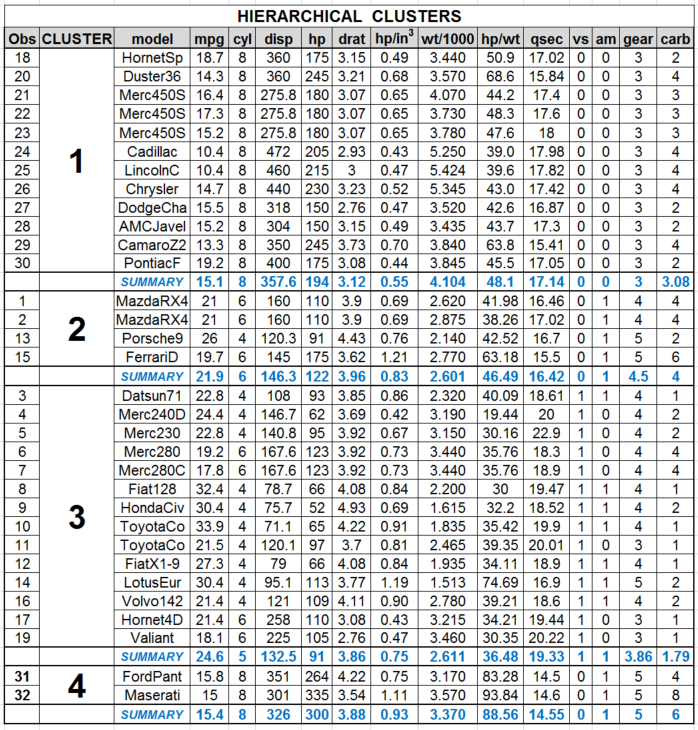
**APPENDIX**

**1 ANOVA TABLE**



**2 Detailed cluster membership from each algorithm**

Note that the cluster numbers are not related between the two tables. The summary data in blue will aid in matching the groupings. For example, based on displacement (disp), cluster 3 in the K-Means table aligns with cluster 1 in the Hierarchical table.



**REFERENCES**

1 <http://cda.psych.uiuc.edu/multivariate_fall_2012/systat_cluster_manual.pdf>

2 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848336/>

**SAS CODE**

/\* PROJECT 3 - CARS CLUSTERING - HIERARCHICAL APPROACH - WILL DO PARTITIONING LATER \*/

**data** cars;

input model $ mpg cyl disp hp drat wt qsec vs am gear carb;

datalines;

MazdaRX4 21 6 160 110 3.9 2.62 16.46 0 1 4 4

MazdaRX4Wag 21 6 160 110 3.9 2.875 17.02 0 1 4 4

Datsun710 22.8 4 108 93 3.85 2.32 18.61 1 1 4 1

Hornet4Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1

HornetSportabout 18.7 8 360 175 3.15 3.44 17.02 0 0 3 2

Valiant 18.1 6 225 105 2.76 3.46 20.22 1 0 3 1

Duster360 14.3 8 360 245 3.21 3.57 15.84 0 0 3 4

Merc240D 24.4 4 146.7 62 3.69 3.19 20 1 0 4 2

Merc230 22.8 4 140.8 95 3.92 3.15 22.9 1 0 4 2

Merc280 19.2 6 167.6 123 3.92 3.44 18.3 1 0 4 4

Merc280C 17.8 6 167.6 123 3.92 3.44 18.9 1 0 4 4

Merc450SE 16.4 8 275.8 180 3.07 4.07 17.4 0 0 3 3

Merc450SL 17.3 8 275.8 180 3.07 3.73 17.6 0 0 3 3

Merc450SLC 15.2 8 275.8 180 3.07 3.78 18 0 0 3 3

CadillacFleetwood 10.4 8 472 205 2.93 5.25 17.98 0 0 3 4

LincolnContinental 10.4 8 460 215 3 5.424 17.82 0 0 3 4

ChryslerImperial 14.7 8 440 230 3.23 5.345 17.42 0 0 3 4

Fiat128 32.4 4 78.7 66 4.08 2.2 19.47 1 1 4 1

HondaCivic 30.4 4 75.7 52 4.93 1.615 18.52 1 1 4 2

ToyotaCorolla 33.9 4 71.1 65 4.22 1.835 19.9 1 1 4 1

ToyotaCorona 21.5 4 120.1 97 3.7 2.465 20.01 1 0 3 1

DodgeChallenger 15.5 8 318 150 2.76 3.52 16.87 0 0 3 2

AMCJavelin 15.2 8 304 150 3.15 3.435 17.3 0 0 3 2

CamaroZ28 13.3 8 350 245 3.73 3.84 15.41 0 0 3 4

PontiacFirebird 19.2 8 400 175 3.08 3.845 17.05 0 0 3 2

FiatX1-9 27.3 4 79 66 4.08 1.935 18.9 1 1 4 1

Porsche914-2 26 4 120.3 91 4.43 2.14 16.7 0 1 5 2

LotusEuropa 30.4 4 95.1 113 3.77 1.513 16.9 1 1 5 2

FordPanteraL 15.8 8 351 264 4.22 3.17 14.5 0 1 5 4

FerrariDino 19.7 6 145 175 3.62 2.77 15.5 0 1 5 6

MaseratiBora 15 8 301 335 3.54 3.57 14.6 0 1 5 8

Volvo142E 21.4 4 121 109 4.11 2.78 18.6 1 1 4 2

;

**run**;

**data** cars2;

set cars;

id= TRIM('R'||\_n\_);

**run**;

/\* STANDARDIZE DATA \*/

**PROC** **STANDARD** data = cars2 out=stddata

MEAN = **0**

STD = **1**;

VAR mpg cyl disp hp drat wt qsec vs am gear carb;

**RUN**;

/\* latent clustering as part of two stage clustering to account for categoricals - NOT covered in class \*/

**proc** **varclus** data=stddata maxclusters=**3**;

var mpg cyl disp hp drat wt qsec vs am gear carb;

**run**;

/\* COMPLETE - Sweet Spot: 4 Clusters\*/

title 'Hierarchical - Complete Method';

**proc** **cluster** method=complete outtree=clust1 data=stddata rsquare rmsstd pseudo ccc;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

/\* WARD'S \*/

title 'Hierarchical - Wards Method';

**proc** **cluster** method=ward outtree=clust1 data=stddata rsquare rmsstd pseudo ccc;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

/\* CENTROID \*/

title 'Hierarchical - CENTROID Method';

**proc** **cluster** method=centroid outtree=clust1 data=stddata rsquare rmsstd pseudo ccc;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

/\* SINGLE \*/

title 'Hierarchical - SINGLE Method';

**proc** **cluster** method=single outtree=clust1 data=stddata rsquare rmsstd pseudo ccc;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

/\* AVERAGE \*/

title 'Hierarchical - AVERAGE Method';

**proc** **cluster** method=average outtree=clust1 data=stddata rsquare rmsstd pseudo ccc;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

/\* optimal clusters according to hierarchical methods is about 3 \*/

title 'Dendrogram for Optimal Clusters';

**proc** **tree** horizontal nclusters=**4** data = clust1 out=clust2;

id id;

**run**;

**proc** **sort** data = stddata;

by id;

**run**;

**proc** **sort** data = clust2;

by id;

**run**;

/\* combine data for variable hypothesis testing \*/

**data** combine;

merge stddata clust2;

by id;

**run**;

/\* test it out \*/

**proc** **glm** data=combine;

class cluster;

model mpg cyl disp hp drat wt qsec vs am gear carb = cluster;

means cluster;

**run**;

**quit**;

**proc** **sort** data = combine;

by cluster;

**run**;

**proc** **print** data=combine;

**run**;

/\* NOW TO PARTITIONING: K-MEANS \*/

title 'Four Clusters';

**proc** **fastclus** maxclusters=**4** replace=random maxiter=**10000** list distance data = stddata out=clusters summary;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

**proc** **fastclus** maxclusters=**4** radius=**3.5** maxiter=**1000** summary list distance data = stddata out=clusters;

var mpg cyl disp hp drat wt qsec vs am gear carb;

id id;

**run**;

/\* set up for plots \*/

**proc** **candisc** out=can data=clusters;

var mpg cyl disp hp drat wt qsec vs am gear carb;

class cluster;

title 'FOUR CLUSTER SCATTERPLOT BY CLUSTER ID';

**proc** **sgplot** data=can;

scatter y=can2 x=can1/group=cluster;

**run**;

**quit**;

**proc** **sort** data=clusters;

by cluster;

**run**;

**proc** **print** data = clusters;

**run**;

**PROC** **MEANS** data=cars2 mean min max range stddev;

VAR disp;

**run**;