Paper 2436-2018

Using Cluster Analysis to Maximize Workplace Design Effectiveness

Renae K. Rich, HDR Inc.

ABSTRACT

While it is generally accepted in the design industry that work spaces should be thoughtfully planned around the workers that occupy them and the types of work they do, available space, required design standards, and budget limit the number of unique workspace options that are feasible. By applying a cluster analysis method to survey data related to the type of work individuals do, job roles are categorized into work style groups with distinct workplace environmental needs and characteristics. These work style profiles are then used to inform and develop design strategies to best support the various types of workers in an organization, within spatial and budgetary parameters.

This presentation outlines the considerations for selecting and combining variables for analysis, exploration and creation of the clustering pattern, investigation of unique cluster characteristics, as well as visualization techniques related to this method. SAS/STAT® procedures used include PROC FASTCLUS, PROC CANDISC, and PROC GLM. PROC SGPLOT, a SAS ODS Graphic procedure, is used for producing high-quality visualizations.

INTRODUCTION

In developing workplace solutions for an organization, a primary goal is to create spaces that allow employees to complete their jobs effectively and efficiently. However, available space, required design standards, and budget constraints limit the number of unique workspace alternatives, such as individual workstations or private offices, that can be included in the design. In order to select the best combination of workspace options and determine the number and type of workers who should occupy each, this innovative analytic approach classifies employees into a limited number of groups based on information about how they actually work.

To accomplish this, survey data is collected related to how workers allocate their time and to what degree they work individually or with others, in addition to their job role within the organization. Based on this information, a cluster analysis is performed to identify groups with unique work style characteristics that could translate to differences in workspace definition. Finally, as the surveys are typically anonymous, job roles linked to the work style groups allow for understanding, projection to non-respondents, and application of design recommendations. Often, when asked directly, employees have preconceived opinions about the type of workspace they would prefer. By instead analyzing data about what individuals actually do at work, this process helps inform design recommendations objectively to support such work effectively.

This paper outlines the steps in SAS/STAT® to utilize survey data to develop groups of observations that have a distinctive set of characteristics using cluster analysis, along with suggestions for the types and formats of variables to consider, and model exploration and visualization techniques. While this method was developed in order to address a specific need in the field of workplace strategy and design, it can be applied anywhere a large number of people or objects need to be grouped according to their similarities.

SELECT AND PREPARE VARIABLES

To perform the cluster analysis, use variables that are hypothesized to be related to the construct in question and vary across the observations. In this case, variables include the type of work people do and the way they spend their time at work. You can select any number of variables for inclusion in the clustering method. It is important to consider need to combine or transform the variables, depending upon the situation.

VARIABLE SELECTION

Variable selection depends on the questions included in the survey and the construct that is being analyzed. Ideally, the survey questions have been developed and validated with this analysis in mind to cover the points of greatest interest. In this paper, variables are related to the types of work people do and their time in various activities. Some examples of variables that were included in the analysis include the following: number of direct reports, time spent in individual focused work, time spent collaborating, time spent away from the primary workspace, ability to work remotely or in alternate spaces, and work-required needs for visual and/or acoustic privacy.

VARIABLE TRANSFORMATION

You should consider whether variables should be transformed before including them in the cluster analysis. It may be necessary to combine multiple variables to create a higher-level score. In this example, we summed results from specific questions related to the time spent in various types of collaboration, including scheduled conferences, impromptu face-to-face encounters, and virtual meetings, to calculate a “total time spent collaborating” value.

If variables are measured on different unit scales, variables with large variances could have a greater effect on the resulting clusters than those with small variances. If variables are measured on different scales, you should consider standardizing your values before performing the cluster analysis. This can be done in SAS using the STANDARD procedure as follows:

```
proc standard data=mydata out=stand mean=0 std=1;
  var myvariables;
run;
```

In the VAR statement, “myvariables” is a list of all variables that are to be standardized and used for the cluster analysis. In the PROC statement, specify the desired distribution of the transformed variables to have mean equal to zero and standard deviation equal to one using the MEAN= and STD= options. The OUT= option creates a new dataset “stand” that will subsequently be used for analysis.

CLUSTER ANALYSIS

A cluster analysis groups observations, in this case individual survey respondents, with similar responses across the selected variables and is based on their proximity to, or distance from, each other. Respondents within a cluster will be more similar to each other than to respondents in the other clusters. As outlined above, this technique is useful to segment the population into groups with different preferences or characteristics and to use the information in applying targeted strategies that are tailored to each group.

Since many surveys contain a large number of responses, which make hierarchical clustering methods unreasonable, nonhierarchical clustering using a *K*-means model is used for this analysis. Nonhierarchical clustering methods can be applied to larger data sets than hierarchical techniques because a matrix of all pairwise distances, or similarities, does not need to be found, and the calculations for assigning respondents to clusters at each step do not need to be stored.

K-MEANS CLUSTERING METHOD

K-means clustering is a nonhierarchical method that, given a set of observations for n respondents $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$, where each is a p-dimensional vector, attempts to partition the observations into k clusters so as to minimize the inter-cluster sums of squares. As shown in Figure 1, first, the respondents are either portioned into *k* initial clusters with the centroid for cluster i as the mean of all vectors of observations in that cluster ($\boldsymbol{c}_i = \overline{\boldsymbol{x}}_i$) or an initial *k* seed points are selected to form the initial nuclei of clusters. Next, each respondent is assigned (or reassigned) to the cluster whose centroid is nearest, usually based on Euclidian distance, and the centroid is recalculated for the cluster receiving the new item and the cluster losing the item. Smaller distance indicates that respondents are more similar on the variables considered. This process is repeated until no more reassignments take place based on a pre-defined level of convergence (Johnson & Wichern, 2007).

The FASTCLUS procedure in SAS/STAT runs a *k*-means clustering method by default:

```
proc fastclus data=stand out=clust maxclusters=5 maxiter=100;
  var myvariables;
run;
```

The standardized variables created above are used by specifying the “stand” data set in the DATA= option. Also, in the PROC FASTCLUS statement, the OUT= option specifies a SAS dataset that contains the original data as well as the cluster assignments that will be used later to perform some diagnostics and further exploration of the clusters. Use the MAXCLUSTERS= option to specify the maximum number of clusters in the analysis and the MAXITER= option for the maximum number of iterations. The variables selected for inclusion in the cluster analysis are listed as “myvariables” in the VAR statement.

The final assignment of items to clusters is, to some extent, dependent on the initial partition or initial selection of seed points. If a SEED= option is specified in the PROC FASTCLUS statement, that data set will be used to select initial seed clusters and must contain the same variables used in the cluster analysis. Otherwise, initial seeds are selected from the DATA= data set.

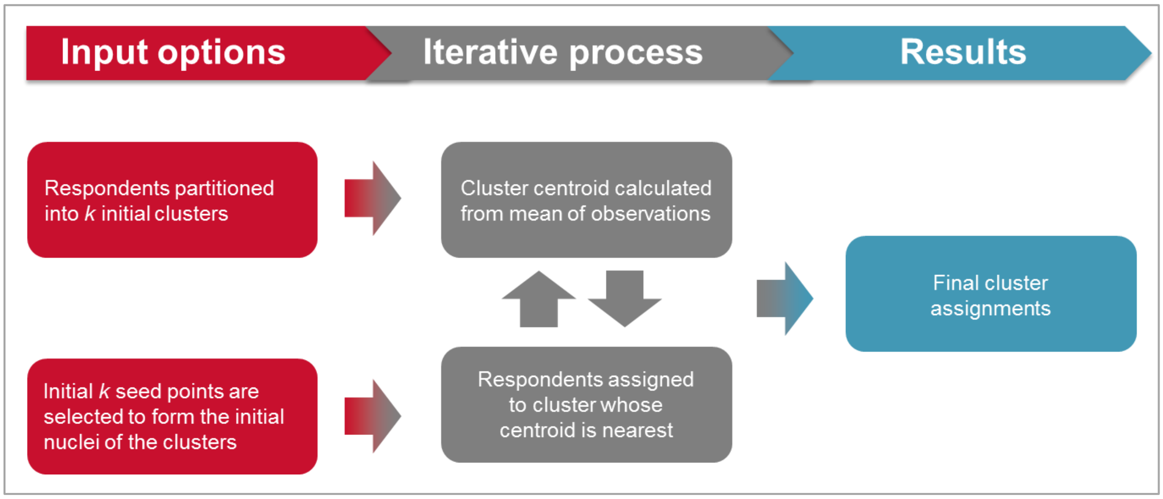

Input options
Iterative process
Results
Respondents partitioned into *k* initial clusters
Initial *k* seed points are selected to form the initial nuclei of the clusters
Cluster centroid calculated from mean of observations
Respondents assigned to cluster whose centroid is nearest
Final cluster assignments

Figure 1. *K*-means clustering method process.

Investigate Clustering Pattern & Determine Number of Clusters

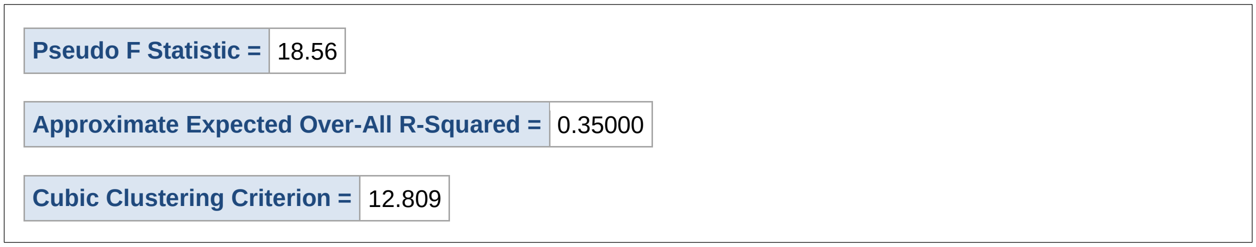
The ideal number of clusters is selected by the analyst, and is somewhat of an exploratory process to find clusters that are well separated and interpretable. Two methods presented here are an evaluation of the Calinski Harabasz (CH) index, which describes the ratio of between-cluster variance to within-cluster variance, and a plot of canonical discriminants, which will visually show the clustering patterns if any exist.

Calinski-Harabasz Index

The CH index is a ratio that compares the between and within cluster variability in an attempt to determine the best number of clusters to include in a cluster analysis. For k clusters, the CH index is the ratio of the between and within sums of squares, divided by their respective degrees of freedom, such that

$$CH_k = \frac{SSB/(k-1)}{SSW/(n-k)}$$

This ratio is analogous to the overall *F*-statistic in a one-way ANOVA, and thus, is sometimes referred to as a Pseudo *F*-statistic, like it is in the SAS output (See Output 1.) within PROC FASTCLUS.



```
Pseudo F Statistic = 18.56

Approximate Expected Over-All R-Squared = 0.35000

Cubic Clustering Criterion = 12.809
```

Output 1. Psuedo F Statistic from PROC FASTCLUS output.

The CH is generally calculated for several values of k, and can be helpful in deciding the “best number” of groups based on the following interpretations:

- If CH_k increases monotonically as k increases, then no reasonably better partition of the points exists than as individuals and there is no cluster structure.
- If CH_k decreases monotonically as k increases, then the points are uniformly distributed in space, and there is a hierarchical structure of clustering, or no natural number of clusters.
- If CH_k rises to a local maximum or at least has a comparatively rapid increase, the value of k at which that occurs will be the best choice for the number of clusters. If there are several local maxima, the smallest value of k amongst those will be the most economical choice (Calinski & Harabasz, 1974).

The CH index across a range of number of clusters is shown in Figure 2. Here there is a global maximum is at five (and a very close value at four) clusters, indicating that the data does follow a cluster structure, with five, or possibly four, clusters.

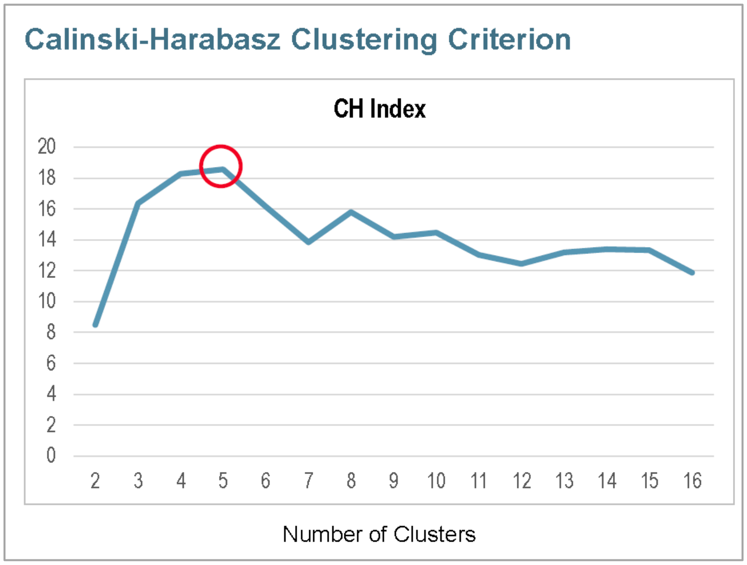


Figure 2. CH Index by number of clusters with global maximum at five clusters.

Plot of Canonical Discriminants

Canonical discriminants are linear combinations of the variables that maximize separation between the clusters. The canonical discriminants $\boldsymbol{l}$ are found to maximize the ratio of between group sums of squares to within group sums of squares, such that

$$\frac{SSB}{SSW} = \frac{\boldsymbol{l}'\left[\sum_{i=1}^{k}(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})'\right]\boldsymbol{l}}{\boldsymbol{l}'\left[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)'\right]\boldsymbol{l}} = \frac{\boldsymbol{l}'H\boldsymbol{l}}{\boldsymbol{l}'E\boldsymbol{l}}$$

where $\boldsymbol{x}_{ij}$ is the vector of p observations for respondent j in cluster i.

The first canonical discriminant maximizes the above ratio, the second canonical discriminant maximizes the ratio under the constraint that it is uncorrelated with the first, and so on. The first two canonical discriminant scores for each respondent can be plotted against each other as a visual representation of the separation between groups (see Figure 3.) (Johnson & Wichern, 2007).

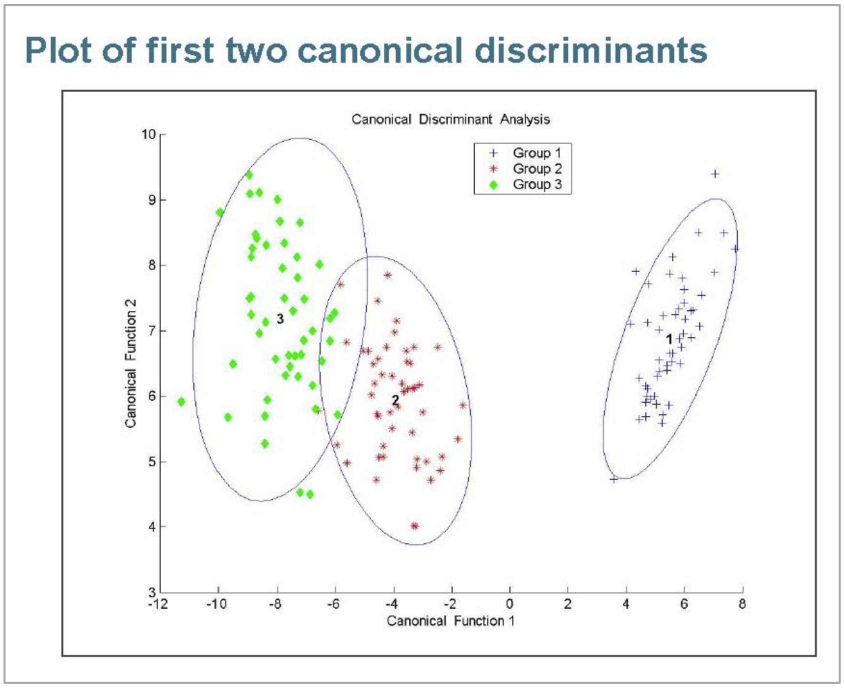
The canonical discriminants of the cluster data are found and plotted using the CANDISC and SGPLOT procedures:

```
PROC CANDISC data=clust out=can noprint;
    class Cluster;
    var myvariables;

PROC SGPLOT data=can;
    scatter y=Can2 x=Can1 / group=Cluster;
run;
```

The CANDISC procedure uses the cluster data set output from the FASTCLUS procedure and produces a new data set specified using the OUT= option. The same list of variables is specified in the VAR statement with the new Cluster variable specified in the CLASS statement.

The SGPLOT procedure then uses the data set output by PROC CANDISC to create a scatter plot. Can1 and Can2 are entered as the variables on the x and y axes, and Cluster is specified with the GROUP= option.



Plot of first two canonical discriminants

Figure 3. Plot of first two canonical discriminants showing a clustering pattern.

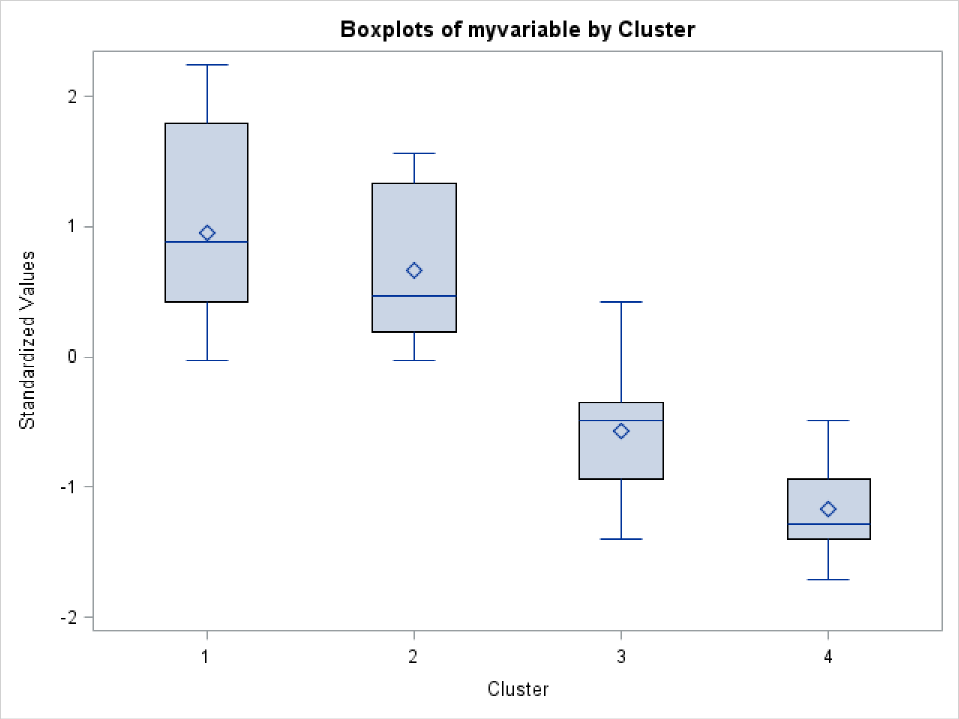
EXPLORE BETWEEN CLUSTER VARIATION BY VARIABLE

Once a clustering pattern is established, you can further explore the differences in distribution between the clusters by variable in order to understand each cluster's unique characteristics.

Visually Compare Clusters

To compare clusters visually, create box plots, as shown in Output 2, with the output data set from PROC FASTCLUS using the SGPLOT procedure:

```
proc sgplot data=clust noautolegend;
title 'Boxplots of myvariable by Cluster';
    label myvariable = 'Standardized Values' cluster = 'Cluster';
    vbox myvariable / category=cluster;
    xaxis discreteorder=data;
run;
```



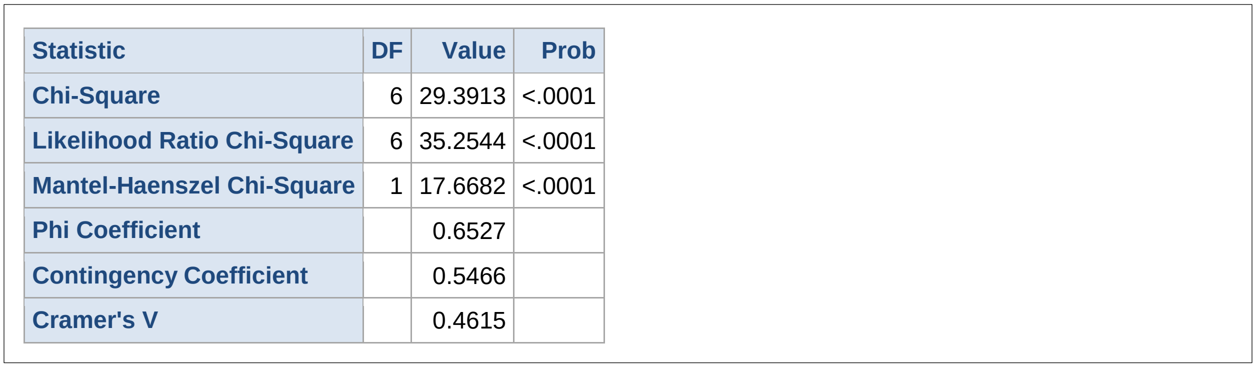
Output 2. Boxplots of selected variable by cluster.

Test for Differences Between Clusters

For categorical variables, chi-square tests for differences in factor levels between clusters are performed using the FREQ procedure:

```
PROC FREQ data=clust;
TITLE 'Test for differences in factor variable frequencies between
clusters';
    TABLE Cluster * myvariable / chisq;
run;
```

The factor variable of interest is specified as “myvariable” in the TABLE statement. The CHISQ option in the TABLE statement will produce the chi-square tests, as shown in Output 3.



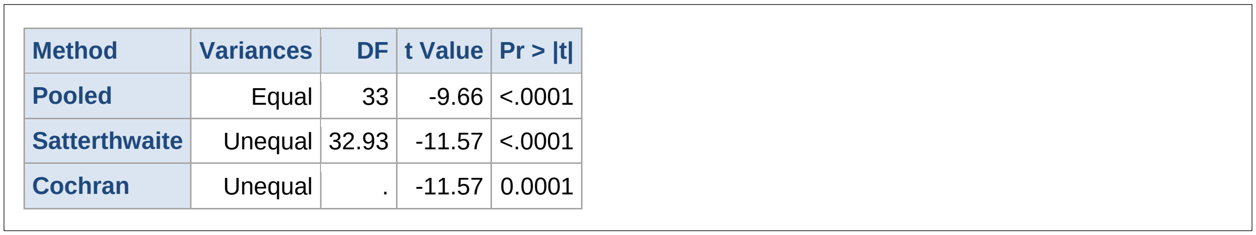
Statistic	DF	Value	Prob
Chi-Square	6	29.3913	<.0001
Likelihood Ratio Chi-Square	6	35.2544	<.0001
Mantel-Haenszel Chi-Square	1	17.6682	<.0001
Phi Coefficient		0.6527	
Contingency Coefficient		0.5466	
Cramer's V		0.4615	

Output 3. Output from PROC FREQ with the CHISQ option.

You can use the TTEST procedure to test for differences in cluster means between pairs of clusters:

```
title 'test for differences in cluster means between clusters 1 and 2';
proc ttest data=clust cochran;
    where (cluster=1 or cluster=2);
    class cluster;
    var myvariable;
run;
```

Specify the variable of interest in the VAR statement and the two clusters to compare in the WHERE statement. If the group variances are unequal, you can include the COCHRAN option in the PROC TTEST statement, which gives the Cochran approximation in addition to the Satterthwaite approximation (included by default), as shown in Output 4.



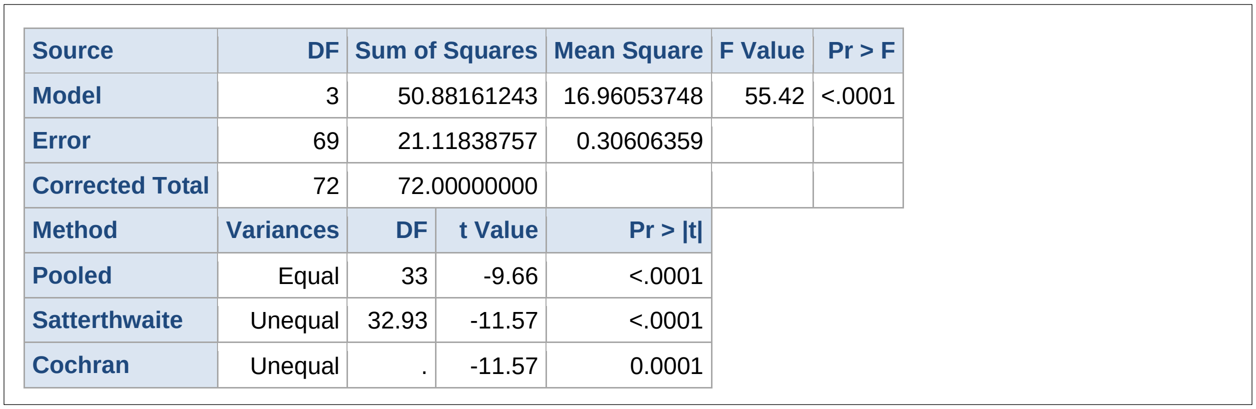
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	33	-9.66	<.0001
Satterthwaite	Unequal	32.93	-11.57	<.0001
Cochran	Unequal	.	-11.57	0.0001

Output 4. Output from PROC TTEST with the COCHRAN option.

You can also test for a cluster effect, or a difference in means across all clusters using the GLM procedure:

```
proc glm data=clust;
    class cluster;
    model myvariable=cluster;
    lsmeans cluster / pdiff adjust=tukey;
run;
```

A significant p-value for the overall F-test in the ANOVA table, as shown in Output 5, indicates that the mean of at least of least one cluster is unequal to the others.

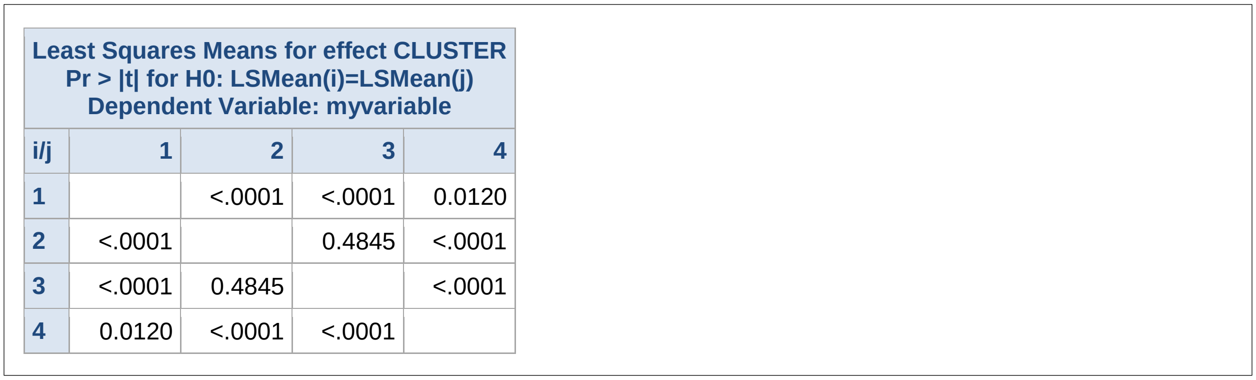


Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	50.88161243	16.96053748	55.42	<.0001
Error	69	21.11838757	0.30606359		
Corrected Total	72	72.00000000			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	33	-9.66	<.0001
Satterthwaite	Unequal	32.93	-11.57	<.0001
Cochran	Unequal	.	-11.57	0.0001

Output 5. ANOVA table output from PROC GLM.

Including the LSMEANS statement will calculate the least squares means and the PDIFF option within the LSMEANS statement tests for differences between all pairwise clusters, shown in Output 6. When ADJUST=TUKEY is specified and data are unbalanced, the p-values are adjusted for multiplicity based on Tukey-Kramer method (Kramer, 1956).



Least Squares Means for effect CLUSTER
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: myvariable

i/j	1	2	3	4
1		<.0001	<.0001	0.0120
2	<.0001		0.4845	<.0001
3	<.0001	0.4845		<.0001
4	0.0120	<.0001	<.0001	

Output 6. Tests for equal least squares means by pairwise cluster with Tukey-Kramer adjustment.

GROUP ASSIGNMENT AND CHARACTERISTICS

MAP JOB ROLES INTO CLUSTERS

Once you identify key characteristics of each cluster, analyze job roles for the best fit to create work style groups. Consider any differences within job roles (e.g., supervisors vs. non-supervisors) that should be categorized into separate groups. Table 1 summarizes this information in a color-coded manner, which makes it easy to identify the similarities and differences between the groups.

Work Style Group	Group 1	Group 2	Group 3	Group 4
Work Style Characteristics	High focus work	High focus work	Low focus work	Low focus work
	Low collaborating	Low collaborating	High collaborating	High collaborating
	Low privacy needs	High privacy needs	Low privacy needs	High privacy needs
Job Roles	Administrative Assistant IT Specialist	Program Manager Registered Nurse	Business Analyst Operations Manager Project Manager	Administrator Medical Director

Table 1. Table of work style group characteristics and job roles within groups.

EXPLORE OTHER SIMILARITIES WITHIN GROUPS

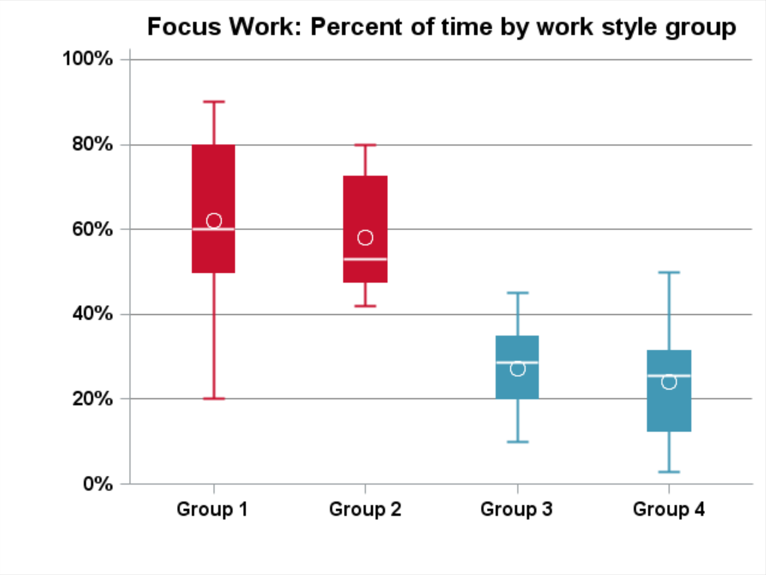
After your job roles are coded into a work-style group variable, you might also consider exploring any other similarities within (or differences across) the work style groups by apply the same techniques explained above to other variables of interest and comparing across groups.

PRESENTATION-READY PLOTS OF GROUP CHARACTERISTICS

The following code can be copied and altered to produced presentation-ready graphics using the SGPLOT procedure depicting the distribution of a variable by work style group, as shown in Output 7:

```
proc sgplot data=wsgroups noautolegend noborder;
  title font=arial bold height=1.75 'Focus Work: Percent of time by
    cluster';
  styleattrs datacolors=(cxc8102e cx4298b5) datacontrastcolors=(cxc8102e
    cx4298b5) datasymbols=(circle);
  vbox perctime_focus / category=wsgroup group=focus_group
    boxwidth=.4 meanattrs=(symbol=circle color=white size=13)
    medianattrs=(pattern=solid color=white thickness=2)
    whiskerattrs=(thickness=2);
  xaxis discreteorder=data label=" " labelattrs=(color=black family=arial
    size=14 weight=bold)
    values=("1" "2" "3" "4") valuesdisplay=("Group 1" "Group 2" "Group 3"
    "Group 4") valueattrs=(color=black family=arial size=12 weight=bold);
  yaxis min=0 offsetmin=0 max=1 grid gridattrs=(color=grey thickness=1)
    label=" " labelattrs=(color=black family=arial size=14 weight=bold)
    valueattrs=(color=black family=arial size=12 weight=bold)
    valueshalign=("center");
run;
```

As compared to the default SAS output shown above in Output 2, all plot text has been customized to fit the style of the presentation it was created for and group box plots are color coded to match the levels presented in Table 1. Since box plots can be difficult to reproduce in Microsoft Excel© or design software, having a way to create presentation-ready graphic in SAS can save time and effort.



Output 7. Presentation-ready plot of variable distribution by group.

TRANSLATE PREFERENCES AND NEEDS INTO STRATEGIES

Finally, once job roles are categorized into work style groups and group characteristics are defined and understood, the information can be used by designers and workplace strategists to develop design the most appropriate design strategies for the mix of employees in an organization.

CONCLUSION

While this method was demonstrated with a specific goal of delineating and distinguishing work style groups for workplace design, this idea can be applied to a variety of situations where data is collected at the observation level, but needs to be categorized for a limited number of solutions to be applied.

REFERENCES

Calinski, T. and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics, 3(1),* 1-27.

Johnson, R. A. and Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis*. 6th ed. Prentice Hall: New York.

Kramer, C. Y. (1956), “Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications,” Biometrics, 12, 307–310.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Renae Rich
HDR, Inc.
402-399-4811
renae.rich@hdrinc.com
www.hdrinc.com/services/research