TUTORIAL

# MM

# Characteristics of Student Populations in Schools and Universities of the United States

*Allan Joseph Medwick*[1]

## INTRODUCTION

Whether it takes the form of total cost of attendance, graduation rates, evidence of cost effectiveness, or evidence of student learning, information about what happens on college campuses is in high demand. For example, the Higher Education Opportunity Act of 2008 includes a list of required consumer information that must be made available by higher education institutions on the Department of Education's College Navigator website (http://nces.ed.gov/collegenavigator/) that literally runs from paragraph (a) to (z). This is in addition to the many data elements already provided by higher education institutions through the Integrated Postsecondary Education Data System (IPEDS). Increasingly, this information about institutional performance is used not only to describe what is happening at a single campus, but also to facilitate comparisons among a set of peer institutions. While the great diversity among higher education institutions in the United States is one of the county's greatest assets—there is an institution to fit almost every possible type of student (e.g., recent high school graduates, adult learners, distance learners) interested in studying every possible subject (e.g., concrete technology, quantum mechanics, Sanskrit)—it also makes finding an acceptable set of peer institutions particularly difficult. The purpose of this tutorial is to demonstrate how unsupervised data mining methods might be used to determine peer groups based on institutional and student data. It is important to note that while the tutorial includes data from real higher education institutions, the purpose of the tutorial is an examination of the techniques as implemented in StatSoft's *STATISTICA* 9 and not the development of actual peer groups. Institutional names are included in the data set so that readers familiar with higher education institutions can evaluate peer groupings for face validity. Readers interested in learning more about peer groups in higher education are encouraged to consult the references listed in the bibliography at the end of the tutorial.

## INTRODUCTION TO THE DATA

Every year, any higher education institution that participates in Federal Student Aid programs such as Federal Work-Study, Stafford Loans, and Pell Grants must complete all surveys in the Integrated Postsecondary Education Data System (IPEDS). IPEDS is built around a series of

[1] Research Fellow, National Center for Education Statistics, U.S. Department of Education

interrelated surveys to collect institutional-level data in such areas as enrollments, graduation rates, student financial aid, completions, faculty, staff, and finance. The collected data is then made available online through the IPEDS Data Center Website (http://www.nces.ed.gov/ipeds/datacenter).


Figure 1: IPEDS Data Center Website

While IPEDS contains data on over 7,050 institutions, the data in this tutorial was taken from the 2007 IPEDS survey using the sample selection criteria listed below, which resulted in an initial list of 2,163 institutions. While finding appropriate peer groups is a problem shared by all institutions, this sample was designed to be familiar to both higher education and non-higher education audiences alike. A number of the variables have been transformed from how they appear in the Data Center to simplify the tutorial. Handling missing data will not be discussed in this tutorial. Institutions with large amounts of missing data were dropped; the final data set consists of 2,096 observations.

### SAMPLE SELECTION CRITERIA

**Misc:** Title IV participating, First Look Universe

**Sector**: Public, 4-year or above, Private not-for-profit, 4-year or above

**Degree-granting status**: Degree-granting

**Geographical region**: New England, Mid East, Great Lakes, Plains, Southeast, Southwest, Rocky Mountains, Far West

## VARIABLES

The data set for this tutorial consists of 227 variables. A complete list of variable names and labels is provided in Appendix A. There is no target variable in unsupervised data mining. The first five variables include the institution's identification number, name, website, and two different Carnegie Classifications. The next seven variables consist of categorical and dummy variables for such items as size category, urbanization, and land grant status. The next 26 variables represent different aspects of the institution using percentages (e.g., the percentage of women at the institution) and totals (e.g., the total number of degrees conferred). The next 81 variables are dummy variables for religious affiliation[2], region, and state. The final 108 variables (deg_101 to deg_451) reflect the percentage of degrees awarded by award level (i.e., Bachelor's, Master's, Doctoral, and First Professional) and 2-digit Classification of Instructional Programs (CIP) code. For example, in "deg_113", the first digit represents the award level (1 = bachelor's degree) and the last two digits represent the CIP code (13 = education).

## VARIABLE REDUCTION

It can be difficult to make sense of the 108 variables that describe the program offerings at higher education institutions. It can also seriously impact the speed and performance of the clustering algorithm being used. For this example, we will use factor analysis as a data reduction method. After loading the tutorial data set (Tutorial_MM.sta) into *STATISTICA*, start a Factor analysis, which is located on the *Statistics* tab in the *Multi/Exploratory* drop down menu (Figure 2). This opens the Factor Analysis startup panel (Figure 3).

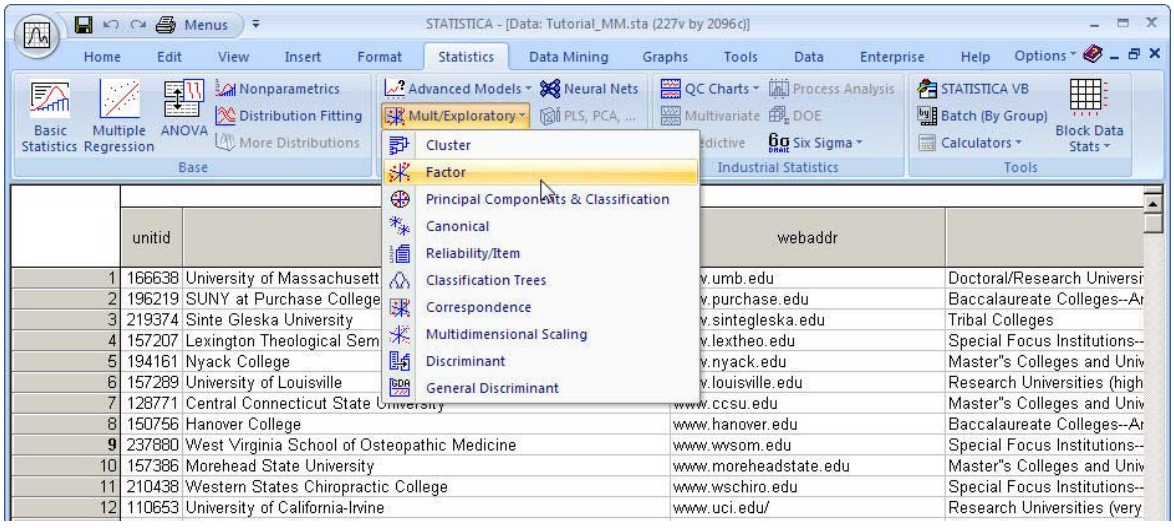



Figure 2: Factor Analysis Menu Selection

[2] A religion needed to be affiliated with at least 10 institutions to have its own indicator variable.

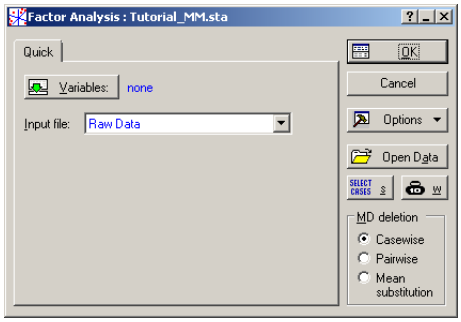



Figure 3: Factor Analysis Startup Panel

Clicking on the Variables button opens a standard variable selection dialog (Figure 4). We will use variables 120 through 227.

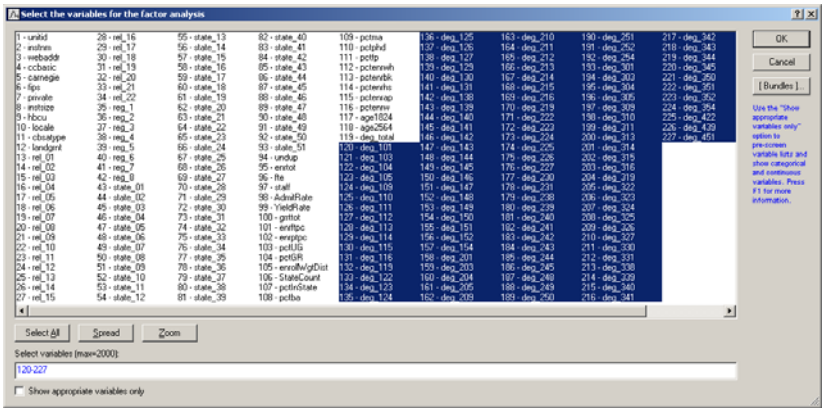



Figure 4: Factor Analysis Variable Selection Dialog

Click OK on both the variable selection dialog and the factor analysis startup panel to open the *Define Method of Factor Extraction* dialog (Figure 5). Click on the *Advanced* tab. We will use *Principle components* as the extraction method with a maximum of 50 factors and a minimum eigenvalue of 1.

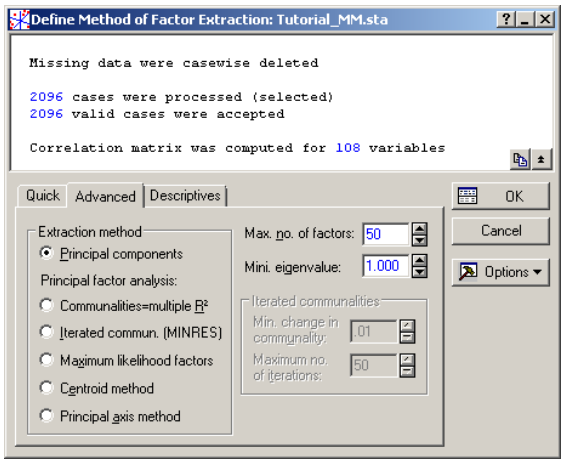



Figure 5: Define Method of Factor Extraction Dialog – Advanced Tab

Clicking OK causes *STATISTICA* to return an error message about the correlation matrix being ill-conditioned and the determinant of the correlation matrix being equal to zero (Figure 6).

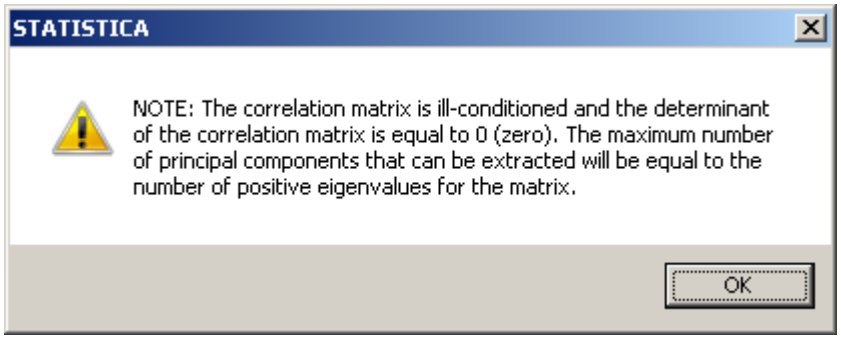



Figure 6: Error Message

Examining the degrees conferred data revealed that some combinations of CIP Code and academic level are so unique that only one institution had a non-zero number in that column, which results in the matrix being "not positive definite." We will try removing any combination for which there are not at least 15 institutions that offer a degree at the level in a particular program. This results in the removal of 14 variables (125, 129, 146, 148, 210, 212, 241, 248, 249, 310, 315, 324, 325, and 341). Return to the variable selection dialog and enter the following without quotes in the *Select Variable (Max 2000)* text box: "120-135 137-138 140-149 151 153-162 164-181 183-186 189-197 199-201 203-206 209-215 217-227".

Click OK on both the variable selection dialog and the factor analysis startup panel to open the *Define Method of Factor Extraction* dialog (Figure 5). Click on the *Advanced* tab. We will use *Principle components* as the extraction method with a maximum of 50 factors and a minimum eigenvalue of 1. *STATISTICA* no longer reports an error and displays the Factor Analysis Results dialog (Figure 7). Click on the *Loadings* tab.

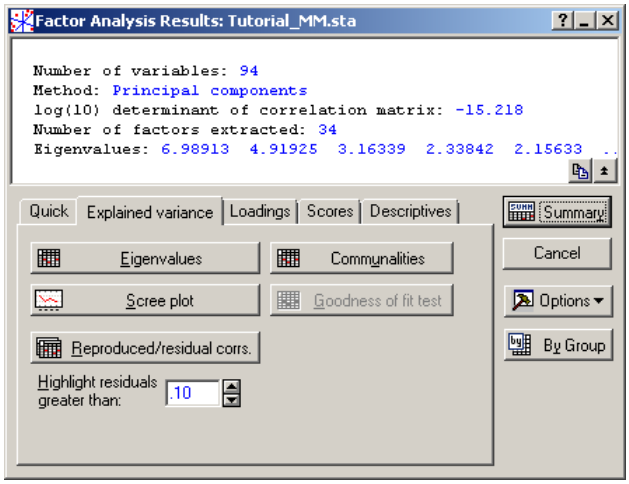



Figure 7: Factor Analysis Results Dialog

From the *Loadings* tab, we can rotate the factors and to try to make better sense of them (Figure 8). Select "Varimax raw" as the factor rotation method and click the *Summary: Factor loadings* button.

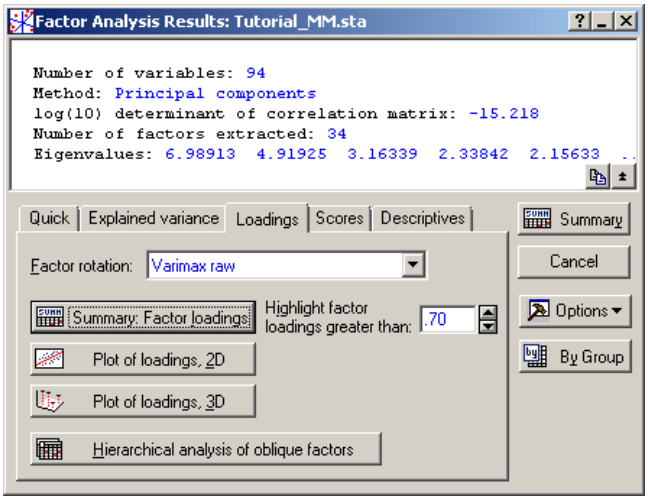



Figure 8: Factor Analysis Results Dialog - Loadings Tab

A results spreadsheet appears containing the loadings for each variable on each of the 34 factors (Figure 9). Notice the red text. This represents factor loadings greater than 0.7. Examine some of the factors using the information on CIP Codes in Appendix A. Notice how Factor 2 has high loadings at the bachelor's degree level for CIP codes 16 (Foreign Languages), 23 (English), 27 (Math and Statistics), 40 (Physical Sciences), 45 (Social Sciences), and 54 (History). This factor might represent undergraduate liberal arts colleges. Factor 4 has high loadings on CIP code 39 (Theology) at the master's and first professional levels.[3] This factor might represent Bible colleges and seminaries.

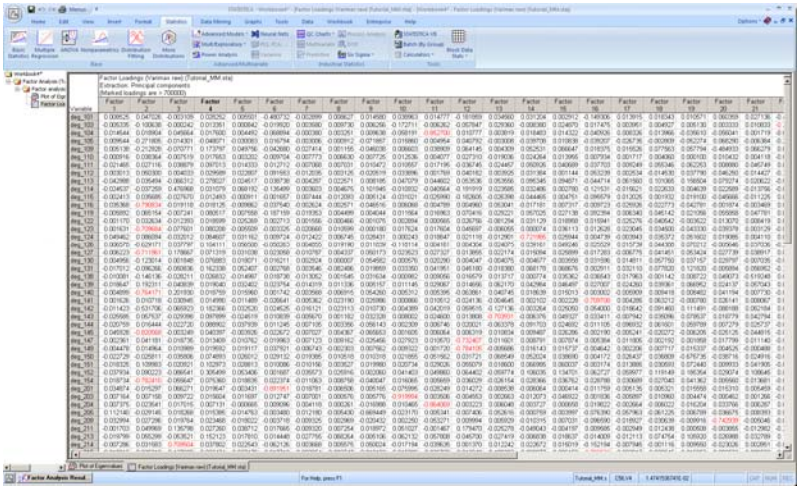


Figure 9: Factor Loadings after Varimax Rotation

[3] Many theology schools offer the Master of Divinity (M.Div.) degree, which is considered a first professional degree since it would allow the graduate to start working as a minister or priest.

The final step of the process is to generate factor scores that can be included in the cluster analysis in the next section. To generate the scores, go to the *Scores* tab of the *Factor Analysis Results dialog* and press *Save factor scores*. The option of selecting which variables to include with the new factor scores now appears. Select variables 1 (unitid) through 119 (deg_total) and click OK.

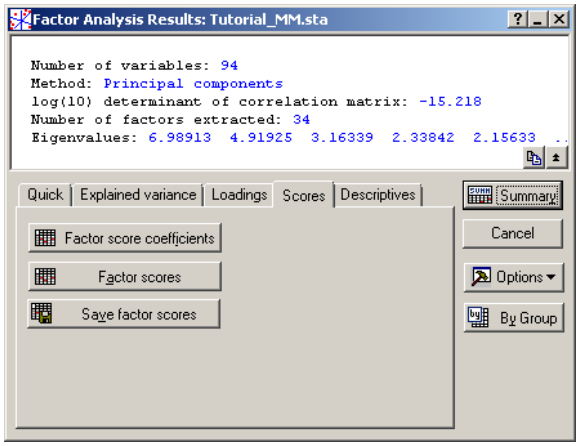



Figure 10: Factor Analysis Results - Scores Tab

## GENERALIZED *K*-MEANS CLUSTERING

Having extracted the essential characteristics of the degree data, we are now ready to move on to finding clusters of similar institutions. *STATISTICA* 9 includes three methods for cluster analysis: cluster analysis, generalized *k*-Means cluster analysis, and generalized EM cluster analysis. For this tutorial, we will examine generalized *k*-Means cluster analysis since it can accommodate categorical variables.

For categorical variables, the highest frequency category becomes the center value for that cluster. So, for example, 41% of the institutions in the sample data set have between 1,000 and 4,999 students, so the center value for the instsize variable will be 2 ("1,000 - 4,999"). It should be noted that for categorical variables all distances can only be 0 (zero) or 1 (one): 0 if the class to which a particular observation belongs is the same as the one that occurs with the greatest frequency in the respective cluster, and 1 if it is different from that class. This caused a problem with the original single variable approach to state, region, and religious affiliation since all of the useful information was lost. Essentially, the variables became indicators of being a New York institution, being located in the Southeast, and not being religiously affiliated. The tutorial data set includes indicator variables for each state, region, and major religious affiliation.

Start a Generalized *k*-Means Clustering analysis by clicking on *Cluster*, which is located on the *Data Mining* tab in the *Clustering/Grouping* section (Figure 11). This opens the Generalized Cluster Analysis startup panel (Figure 12).

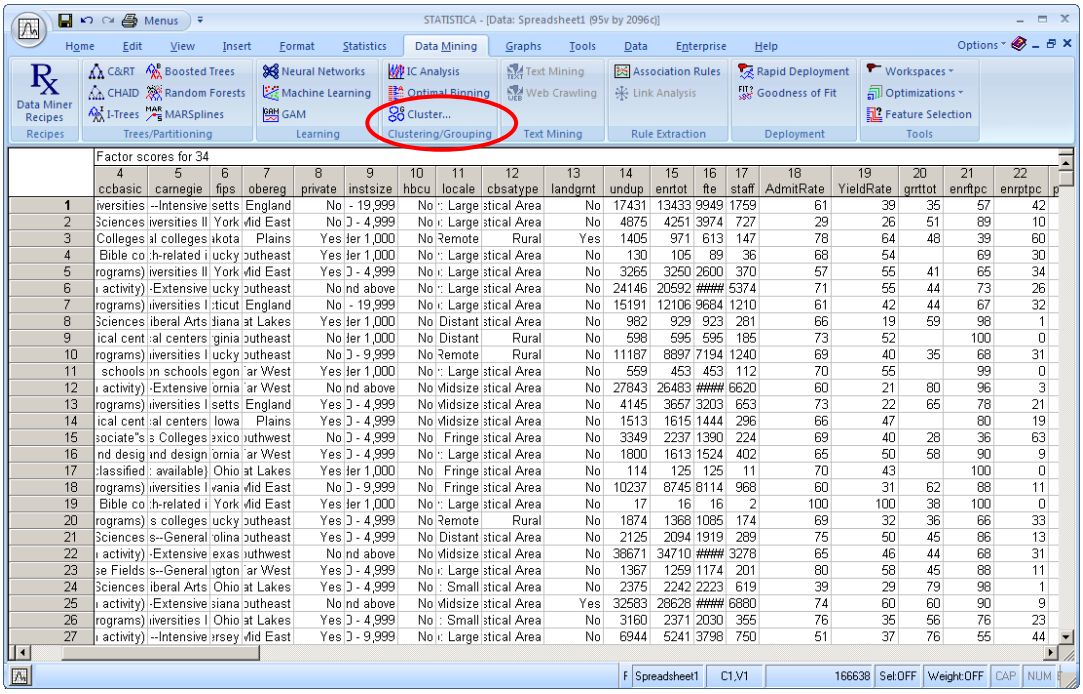



Figure 11: Generalized *k*-Means Clustering

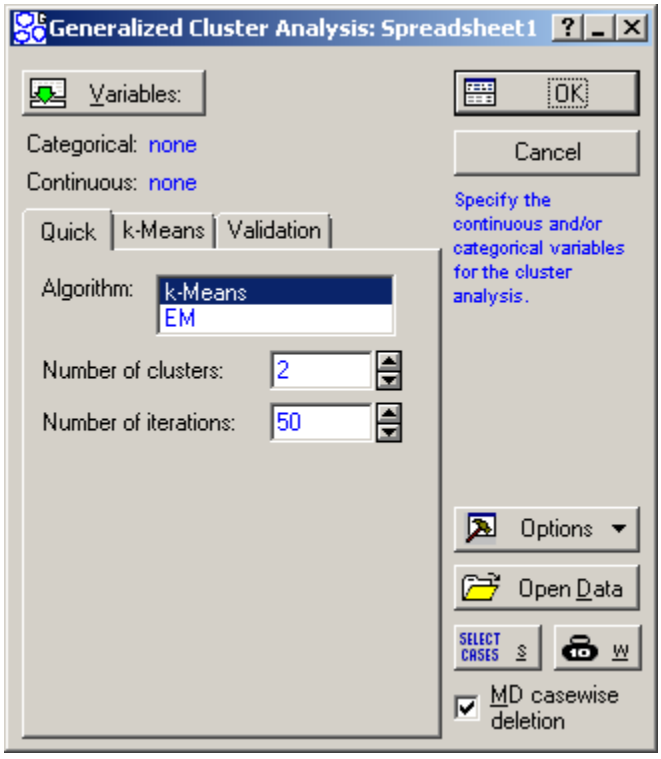



Figure 12: Generalized Cluster Analysis Startup Panel

Clicking on the Variables button opens a variable selection dialog with separate selection panes for categorical and continuous variables (Figure 13).

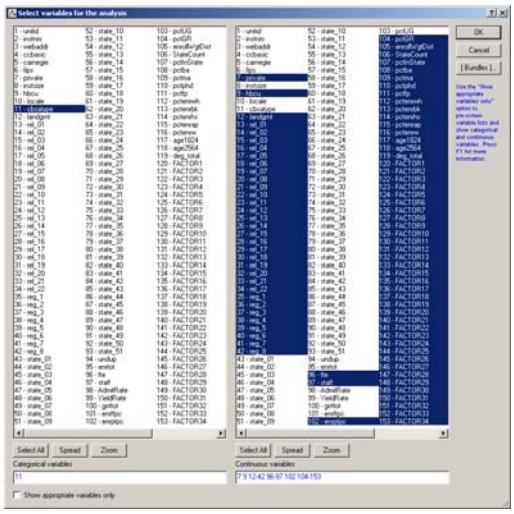


Figure 13: Variable Selection Dialog

Because of the way that *STATISTICA* handles categorical variables in the generalized *k*-Means procedure—forcing the distances to be 0 or 1—it makes some sense to treat the indicator variables as continuous. In this example, the following variables were selected:

**Categorical Variables:** 11
**Continuous Variables:** 7 9 12-42 96-97 102 104-153

On the Generalized Cluster Analysis Startup Dialog (Figure 14), a number of settings were changed. The number of clusters was set to 150. If all of the clusters were the same size, the average peer group size would be 14. The number of iterations was increased to an arbitrarily high number (e.g, 500) to prevent early stopping. The initial cluster centers were selected randomly from *k* observations and standardized city-block (Manhattan) distances were used.

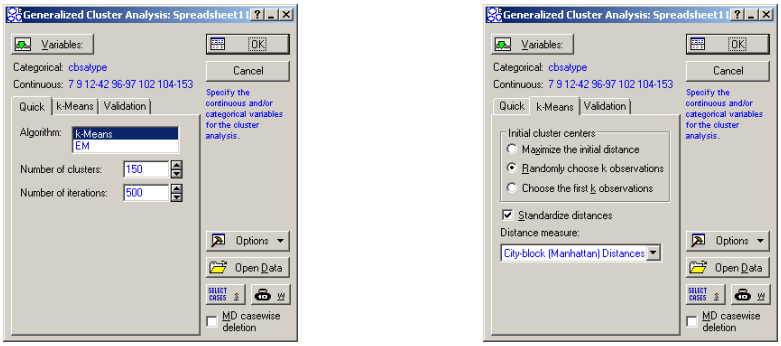



Figure 14: Generalized Cluster Analysis Startup Panel - Quick and *k*-Means Tabs

Please be aware that some combinations of variables may result in an error message about the number of clusters being too large (Figure 14). This seems to happen in combination with the "Maximize the initial distance" option for initial cluster centers. Changing the initial cluster center option to "Randomly choose *k* observations" often helps as does using the "City Block (Manhattan) Distance" option. If not, try reducing the minimum number of clusters or changing the variables used in the model. Do not be afraid to re-run the analysis without any changes as it may also be the result of bad starting values. It can be an extremely frustrating problem.

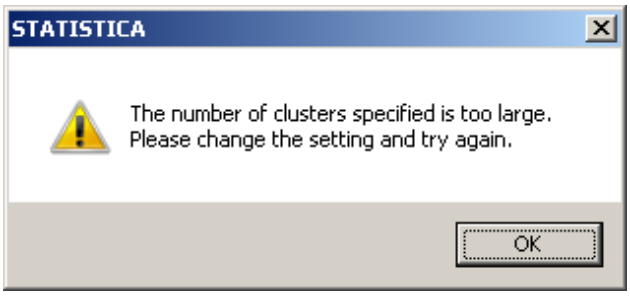



Figure 15: Error Message

Standardized distances were used to prevent variables from affecting the analysis simply based upon how they are scaled. The city-block (Manhattan) distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, in this measure, the effect of single large differences (i.e., outliers) is dampened since they are not squared. After a number of iterations the Results Dialog appears (Figure 15).

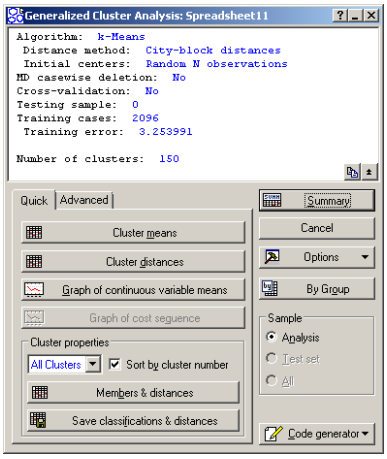



Figure 16: Results Dialog

Check *Sort by cluster number* and click on *Save classifications & distances*. This will open a variable selection window. Select all variables and click OK. The generated clusters are now appended to the end of the original data file. Before moving on to analyzing the results, let's run the same analysis but with v-fold validation. Clicking *Cancel* on the results dialog reopens the Generalized Cluster Analysis Startup Dialog. Click on the *Validation* tab (Figure 16).

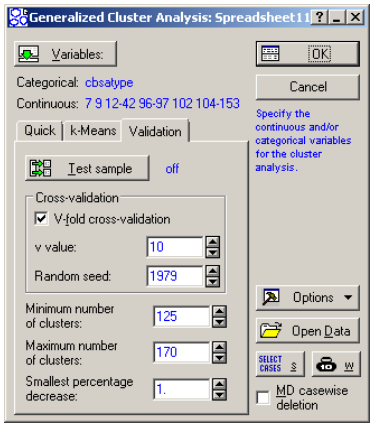



Figure 17: Generalized Cluster Analysis - Validation Tab

Validation is a method for selecting the best number of clusters. In this case, we expect the number of clusters to be between 125 and 170. V-fold cross-validation divides the overall sample into a number of randomly drawn sub-samples. The same type of analysis is then successively applied to the observations belonging to the v-1 folds (i.e., the training sample), and the results of the analyses are applied to sample v (i.e., the testing sample) to determine how well the observations in sample v can be assigned to homogenous clusters using the current cluster solution computed from the v-1 learning samples. The results for the v replications are averaged to yield a single measure of the validity of the model for assigning new observations to clusters. The *Random seed* was set to 1979, so that they analysis can be replicated by tutorial users. The *Smallest percentage decrease* was set to 1. The program will compute cluster solutions for an increasing number of clusters—from the *Minimum number of clusters* to the *Maximum number of clusters*—until the decrease in the average distance of cases to cluster centers between *k* and *k*+1 clusters is less than the percentage specified. The result of the analysis with validation is shown below (Figure 17). Notice that the final number of clusters is 125. As before, check *Sort by cluster number* and click on *Save classifications & distances*. This will open a variable selection window. Select all variables and click OK. The generated clusters are now appended to the end of the original data file.

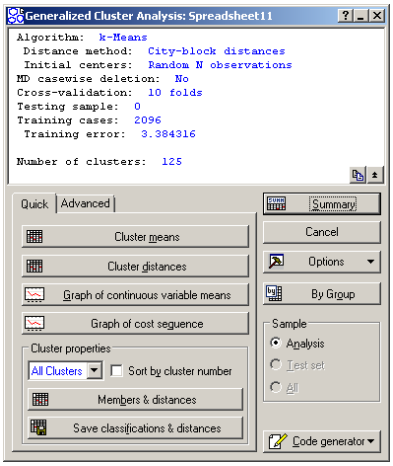



Figure 18: Results Dialog with Validation

In both cases the MD casewise deletion option was unchecked. If an observation is missing data, the *k*-Means algorithm will compute cluster assignments based on the observed data only. Since the goal is to create a peer group for every institution, institutions cannot be dropped for having some missing values, so this option must be unchecked.

## FACE VALIDITY AND OTHER CHECKS

After each completed run, it is useful to examine the generated peer groups of some well-known universities. Four illustrative universities were selected from the cluster analysis with validation: Harvard University, the Curtis Institute of Music, St. Charles Borromeo Seminary, and the University of Maryland - College Park.

*Cluster 50: Harvard University (11 Institutions)*

| Institution Name | Carnegie Classification | State | Private | Institutional Size |
|---|---|---|---|---|
| Harvard University | Research Universities (very high research activity) | Massachusetts | Yes | 20,000 and above |
| Johns Hopkins University | Research Universities (very high research activity) | Maryland | Yes | 10,000 - 19,999 |
| Columbia University in the City of New York | Research Universities (very high research activity) | New York | Yes | 20,000 and above |
| New York University | Research Universities (very high research activity) | New York | Yes | 20,000 and above |
| Yale University | Research Universities (very high research activity) | Connecticut | Yes | 10,000 - 19,999 |
| Boston University | Research Universities (very high research activity) | Massachusetts | Yes | 20,000 and above |
| Georgetown University | Research Universities (very high research activity) | District of Columbia | Yes | 10,000 - 19,999 |
| University of Pennsylvania | Research Universities (very high research activity) | Pennsylvania | Yes | 20,000 and above |
| Washington University in St Louis | Research Universities (very high research activity) | Missouri | Yes | 10,000 - 19,999 |
| George Washington University | Research Universities (high research activity) | District of Columbia | Yes | 20,000 and above |
| University of Chicago | Research Universities (very high research activity) | Illinois | Yes | 10,000 - 19,999 |

Harvard clusters along with 3 of the other 7 Ivy League institutions (Penn, Columbia, and Yale) and with a number of other large, research universities. What is common to all of these universities is the large graduate student population in both academic and professional disciplines.

*Cluster 61: The Curtis Institute of Music (15 Institutions)*

| Institution Name | Carnegie Classification | State | Private | Institutional Size |
|---|---|---|---|---|
| The Curtis Institute of Music | Schools of art, music, and design | Pennsylvania | Yes | Under 1,000 |
| St John's College | Baccalaureate Colleges--Arts & Sciences | Maryland | Yes | Under 1,000 |
| Bryn Mawr College | Baccalaureate Colleges--Arts & Sciences | Pennsylvania | Yes | 1,000 - 4,999 |
| The New School | Doctoral/Research Universities | New York | Yes | 5,000 - 9,999 |
| Sarah Lawrence College | Baccalaureate Colleges--Arts & Sciences | New York | Yes | 1,000 - 4,999 |
| Gallaudet University | Master"s Colleges and Universities (medium programs) | District of Columbia | Yes | 1,000 - 4,999 |
| Bard College | Baccalaureate Colleges--Arts & Sciences | New York | Yes | 1,000 - 4,999 |
| Manhattan School of Music | Schools of art, music, and design | New York | Yes | Under 1,000 |
| Goucher College | Baccalaureate Colleges--Arts & Sciences | Maryland | Yes | 1,000 - 4,999 |
| School of the Art Institute of Chicago | Schools of art, music, and design | Illinois | Yes | 1,000 - 4,999 |
| The University of the Arts | Schools of art, music, and design | Pennsylvania | Yes | 1,000 - 4,999 |
| Maryland Institute College of Art | Schools of art, music, and design | Maryland | Yes | 1,000 - 4,999 |
| Pratt Institute-Main | Schools of art, music, and design | New York | Yes | 1,000 - 4,999 |
| The Juilliard School | Schools of art, music, and design | New York | Yes | 1,000 - 4,999 |
| Pennsylvania Academy of the Fine Arts | Schools of art, music, and design | Pennsylvania | Yes | Under 1,000 |

The Curtis Institute of Music clusters with a number of other preeminent institutions that specialize in the arts. An outlier among this crowd is St. John's College in Maryland, the "great books" college. It could appear here because of its small size, location, and its geographical drawing power.

*Cluster 101: St. Charles Borromeo Seminary (17 Institutions)*

| Institution Name | Carnegie Classification | State | Private | Institutional Size |
|---|---|---|---|---|
| Saint Charles Borromeo Seminary-Overbrook | Theological seminaries, Bible colleges | Pennsylvania | Yes | Under 1,000 |
| New York Medical College | Medical schools and medical centers | New York | Yes | 1,000 - 4,999 |
| Saint Josephs Seminary and College | Theological seminaries, Bible colleges | New York | Yes | Under 1,000 |
| Evangelical Theological Seminary | Theological seminaries, Bible colleges | Pennsylvania | Yes | Under 1,000 |
| Reformed Presbyterian Theological Seminary | Theological seminaries, Bible colleges | Pennsylvania | Yes | Under 1,000 |
| Washington Theological Union | Theological seminaries, Bible colleges | District of Columbia | Yes | Under 1,000 |
| Unification Theological Seminary | Theological seminaries, Bible colleges | New York | Yes | Under 1,000 |
| Wesley Theological Seminary | Theological seminaries, Bible colleges | District of Columbia | Yes | Under 1,000 |
| Westminster Theological Seminary | Theological seminaries, Bible colleges | Pennsylvania | Yes | Under 1,000 |
| Dallas Theological Seminary | Theological seminaries, Bible colleges | Texas | Yes | 1,000 - 4,999 |
| Christ the King Seminary | Theological seminaries, Bible colleges | New York | Yes | Under 1,000 |
| St Bernard's School of Theology and Ministry | Theological seminaries, Bible colleges | New York | Yes | Under 1,000 |
| New Brunswick Theological Seminary | Theological seminaries, Bible colleges | New Jersey | Yes | Under 1,000 |
| St. Mary's Seminary & University | Theological seminaries, Bible colleges | Maryland | Yes | Under 1,000 |
| Washington Bible College-Capital Bible Seminary | Theological seminaries, Bible colleges | Maryland | Yes | Under 1,000 |
| Calvary Baptist Theological Seminary | Theological seminaries, Bible colleges | Pennsylvania | Yes | Under 1,000 |
| Seminary of the Immaculate Conception | Theological seminaries, Bible colleges | New York | Yes | Under 1,000 |

St. Charles Borromeo Seminary is the Roman Catholic seminary for the Archdiocese of Philadelphia. This cluster contains four other Catholic seminaries, including St. Joseph's Seminary and College, Christ the King Seminary, St. Mary's Seminary and College, and the Seminary of the Immaculate Conception. Except for New York Medical College, the remaining institutions are seminaries of other Christian denominations. New York Medical College might be on this list because of its focus on graduate education, size, and location.

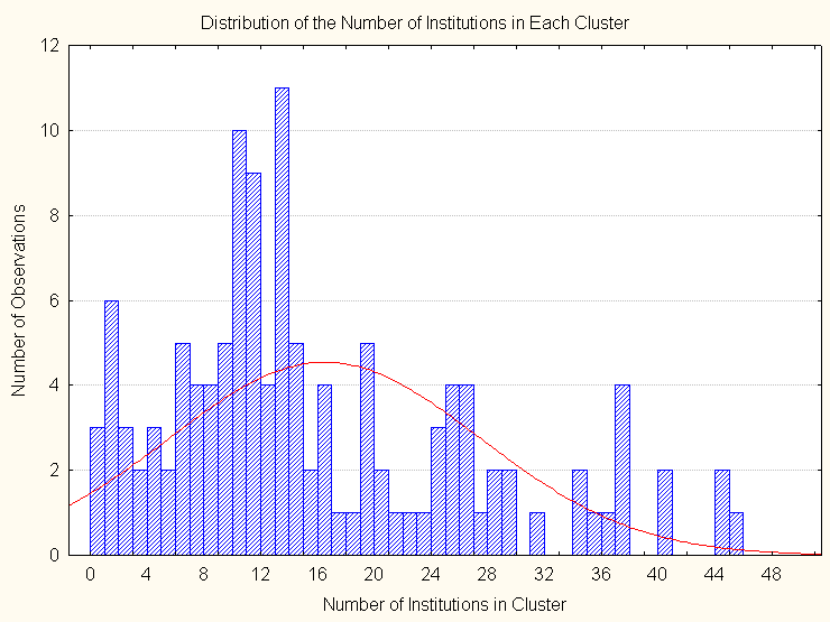
*Cluster 84: University of Maryland - College Park (7 Institutions)*

| Institution Name | Carnegie Classification | State | Private | Institutional Size |
|---|---|---|---|---|
| University of Maryland-College Park | Research Universities (very high research activity) | Maryland | No | 20,000 and above |
| Rutgers University-New Brunswick | Research Universities (very high research activity) | New Jersey | No | 20,000 and above |
| University of Pittsburgh-Pittsburgh Campus | Research Universities (very high research activity) | Pennsylvania | No | 20,000 and above |
| Stony Brook University | Research Universities (very high research activity) | New York | No | 20,000 and above |
| University of Delaware | Research Universities (very high research activity) | Delaware | No | 20,000 and above |
| Temple University | Research Universities (high research activity) | Pennsylvania | No | 20,000 and above |
| University at Buffalo | Research Universities (very high research activity) | New York | No | 20,000 and above |

This cluster contains 7 large, public research universities located in New Jersey, New York, Delaware, and Pennsylvania.

In general, the clustering algorithm created groups of institutions that one might expect to be similar. Sometimes the similarities are not immediately apparent. For example, The New School, a large, doctoral university is listed with the Curtis Institute of Music; however, upon closer examination one sees that The New School includes Parsons The New School for Design, Mannes College The New School for Music, The New School for Drama, and The New School for Jazz and Contemporary Music. Unfortunately, the grouping of other institutions, like New York Medical College's listing among theological schools, defies explanation.

In addition to examining the institutions that comprise the different clusters, it is important to examine the size of the clusters. The average number of institutions in each cluster is 17; however, the size of the clusters ranges from 1 to 46. Closer examination of the data is necessary to determine how to divide larger clusters and combine smaller clusters into more reasonably-sized peer groups.



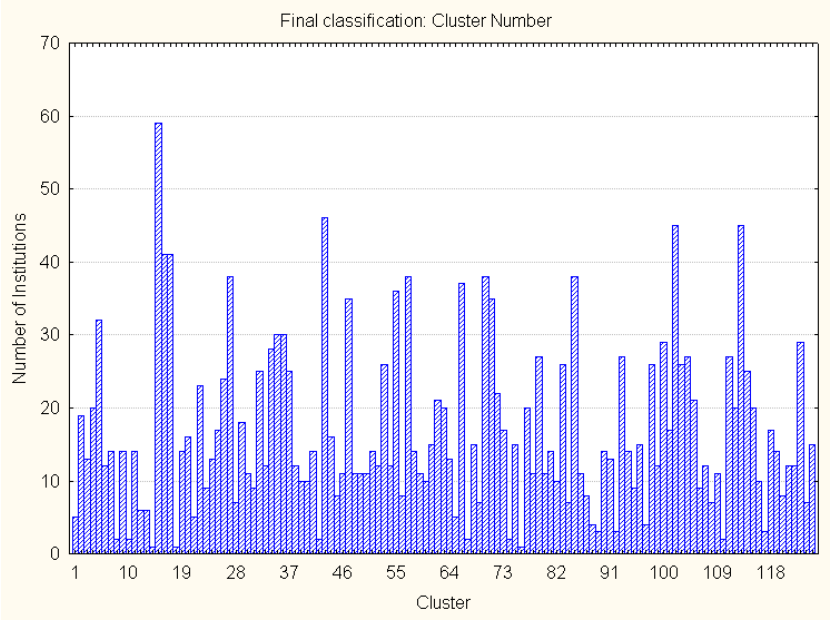




**Figure 19: Number of Institutions by Cluster**

## DATA FILES

The data files for this tutorial are available from the tutorial website in Stata, *STATISTICA*, and SPSS file formats. Files were converted using Stat/Transfer (http://www.stattransfer.com/) and checked by the author for fidelity to the original Stata file. The results file for both analyses presented in this tutorial are also available from the tutorial website:

Tutorial_MM_Sample_Solution_Without_Validation.sta
Tutorial_MM_Sample_Solution_With_Validation.sta

## SUMMARY

The new generalized *k*-Means cluster feature of *STATISTICA* 9 provided some useful insights into this problem without any customization; however, more robust handling of categorical variables and the ability to set minimum and maximum acceptable cluster sizes could greatly improve this tutorial. This can be done using macros and other features of the software. If readers are interested and request such a demonstration, we can possibly post an expanded tutorial showing how to do it on the companion website in the future.

## BIBLIOGRAPHY

Ingram, J. A. (1995). *Using Ipeds Data for Selecting Peer Institutions*. Paper presented at the 35th Annual Forum of the Association for Institutional Research, Boston, MA.

Teeter, D. J., & Christal, M. E. (1987). Establishing Peer Groups: A Comparison of Methodologies. *Planning for Higher Education, 15*(2), 8-17.

Terenzini, P. T., Hartmark, L., Lorang, W. G., & Shirley, R. C. (1980). A Conceptual and Methodological Approach to the Identification of Peer Institutions. *Research in Higher Education, 12*(4), 347-364.

Xu, J. (2008). Using the Ipeds Peer Analysis System in Peer Group Selection. *AIR Professional File, 110*. Retrieved from http://www.airweb.org/page.asp?page=73&apppage=85&id=114.

# APPENDIX A: VARIABLE INFORMATION

**obs:** 2,096
**vars:** 149

**Case Numbers, Institution Names, Website URLs, and Carnegie Classifications**

| Variable Name | Storage Type | Variable Label |
|---|---|---|
| unitid | long | UNITID |
| instnm | str71 | Institution Name |
| webaddr | str91 | Institution's internet website address |
| ccbasic | byte | Carnegie Classification 2005: Basic |
| carnegie | byte | Carnegie Classification 2000 |

**Categorical Variables for Use in Modeling**

| Variable Name | Storage Type | Variable Label |
|---|---|---|
| fips | byte | FIPS state code |
| private | byte | Private Institution |
| instsize | byte | Institution size category |
| hbcu | byte | Historically Black College or University |
| locale | byte | Degree of urbanization (Urban-centric locale) |
| cbsatype | byte | CBSA Type Metropolitan, Micropolitan, or Rural |
| landgrnt | byte | Land Grant Institution |
| rel_01 | byte | No Religious Affiliation |
| rel_02 | byte | Assemblies of God Church |
| rel_03 | byte | Roman Catholic |
| rel_04 | byte | Evangelical Lutheran Church |
| rel_05 | byte | Interdenominational |
| rel_06 | byte | Christian Churches and Churches of Christ |
| rel_07 | byte | American Baptist |
| rel_08 | byte | Baptist |
| rel_09 | byte | Church of the Nazarene |
| rel_10 | byte | Christian Church (Disciples of Christ) |
| rel_11 | byte | Presbyterian Church (USA) |
| rel_12 | byte | Lutheran Church - Missouri Synod |
| rel_13 | byte | United Methodist |
| rel_14 | byte | Protestant Episcopal |
| rel_15 | byte | Churches of Christ |
| rel_16 | byte | Southern Baptist |
| rel_17 | byte | United Church of Christ |
| rel_18 | byte | Other Protestant |
| rel_19 | byte | Jewish |
| rel_20 | byte | Undenominational |
| rel_21 | byte | Seventh Day Adventists |
| rel_22 | byte | Other Religious Affiliation |
| reg_1 | byte | New England |
| reg_2 | byte | Mid East |
| reg_3 | byte | Great Lakes |
| reg_4 | byte | Plains |
| reg_5 | byte | Southeast |
| reg_6 | byte | Southwest |
| reg_7 | byte | Rocky Mountains |
| reg_8 | byte | Far West |
| state_01 | byte | Alabama |
| state_02 | byte | Alaska |
| state_03 | byte | Arizona |
| state_04 | byte | Arkansas |
| state_05 | byte | California |
| state_06 | byte | Colorado |
| state_07 | byte | Connecticut |
| state_08 | byte | Delaware |
| state_09 | byte | District of Columbia |
| state_10 | byte | Florida |
| state_11 | byte | Georgia |

| Variable Name | Storage Type | Variable Label |
|---|---|---|
| state_12 | byte | Hawaii |
| state_13 | byte | Idaho |
| state_14 | byte | Illinois |
| state_15 | byte | Indiana |
| state_16 | byte | Iowa |
| state_17 | byte | Kansas |
| state_18 | byte | Kentucky |
| state_19 | byte | Louisiana |
| state_20 | byte | Maine |
| state_21 | byte | Maryland |
| state_22 | byte | Massachusetts |
| state_23 | byte | Michigan |
| state_24 | byte | Minnesota |
| state_25 | byte | Mississippi |
| state_26 | byte | Missouri |
| state_27 | byte | Montana |
| state_28 | byte | Nebraska |
| state_29 | byte | Nevada |
| state_30 | byte | New Hampshire |
| state_31 | byte | New Jersey |
| state_32 | byte | New Mexico |
| state_33 | byte | New York |
| state_34 | byte | North Carolina |
| state_35 | byte | North Dakota |
| state_36 | byte | Ohio |
| state_37 | byte | Oklahoma |
| state_38 | byte | Oregon |
| state_39 | byte | Pennsylvania |
| state_40 | byte | Rhode Island |
| state_41 | byte | South Carolina |
| state_42 | byte | South Dakota |
| state_43 | byte | Tennessee |
| state_44 | byte | Texas |
| state_45 | byte | Utah |
| state_46 | byte | Vermont |
| state_47 | byte | Virginia |
| state_48 | byte | Washington |
| state_49 | byte | West Virginia |
| state_50 | byte | Wisconsin |
| state_51 | byte | Wyoming |

**Continuous Variables for Use in Modeling**

| Variable Name | Storage Type | Variable Label |
|---|---|---|
| undup | long | 12-month unduplicated headcount, total: Academic year 2006-07 |
| enrtot | long | Total enrollment |
| fte | long | Full-time equivalent enrollment: Fall 2007 |
| staff | int | Total FTE staff |
| AdmitRate | byte | Admissions Rate |
| YieldRate | byte | Yield Rate |
| grrttot | byte | Graduation rate, total cohort * **MISSING VALUES** |
| enrftpc | byte | Percent Full-time enrollment |
| enrptpc | byte | Percent Part-time enrollment |
| pctUG | byte | Percentage Undergraduate |
| pctGR | byte | Percentage Graduate |
| enrollWgtDist | float | Enrollment Weighted Distance * **MISSING VALUES** |
| StateCount | float | Number of States Sending Students to Institution * **MISSING VALUES** |
| pctInState | float | Percentage of Freshmen from In-State * **MISSING VALUES** |
| pctba | float | Percentage Bachelor's Degrees Awarded |
| pctma | float | Percentage Master's Degrees Awarded |
| pctphd | float | Percentage Doctorates Degrees Awarded |
| pctfp | float | Percentage First Professional Degrees Awarded |
| pctenrwh | byte | Percent of total enrollment that are White, non-Hispanic |
| pctenrbk | byte | Percent of total enrollment that are Black, non-Hispanic |
| pctenrhs | byte | Percent of total enrollment that are Hispanic |

| Variable Name | Storage Type | Variable Label |
|---|---|---|
| pctenrap | byte | Percent of total enrollment that are Asian or Pacific Islander |
| pctenrw | byte | Percent of total enrollment that are women |
| age1824 | byte | Percent of undergraduate enrollment 18-24 * **MISSING VALUES** |
| age2564 | byte | Percent of undergraduate enrollment, 25-64 * **MISSING VALUES** |
| deg_total | int | Total Degrees Conferred |
| deg_101 | float | Bachelor's in Agriculture, agriculture operations, and related sciences |
| deg_103 | float | Bachelor's in Natural resources and conservation |
| deg_104 | float | Bachelor's in Architecture and related services |
| deg_105 | float | Bachelor's in Area, ethnic, cultural, and gender studies |
| deg_109 | float | Bachelor's in Communication, journalism, and related programs |
| deg_110 | float | Bachelor's in Communications technologies |
| deg_111 | float | Bachelor's in Computer and information sciences and support services |
| deg_112 | float | Bachelor's in Personal and culinary services |
| deg_113 | float | Bachelor's in Education |
| deg_114 | float | Bachelor's in Engineering |
| deg_115 | float | Bachelor's in Engineering technologies/technicians |
| deg_116 | float | Bachelor's in Foreign languages, literatures, and linguistics |
| deg_119 | float | Bachelor's in Family and consumer sciences/human sciences |
| deg_122 | float | Bachelor's in Legal professions and studies |
| deg_123 | float | Bachelor's in English language and literature/letters |
| deg_124 | float | Bachelor's in Liberal arts and sciences, general studies and humanities |
| deg_125 | float | Bachelor's in Library science |
| deg_126 | float | Bachelor's in Biological and biomedical sciences |
| deg_127 | float | Bachelor's in Mathematics and statistics |
| deg_129 | float | Bachelor's in Military technologies |
| deg_130 | float | Bachelor's in Multi/interdisciplinary studies |
| deg_131 | float | Bachelor's in Parks, recreation, leisure, and fitness studies |
| deg_138 | float | Bachelor's in Philosophy and religious studies |
| deg_139 | float | Bachelor's in Theology and religious vocations |
| deg_140 | float | Bachelor's in Physical sciences |
| deg_141 | float | Bachelor's in Science technologies/technicians |
| deg_142 | float | Bachelor's in Psychology |
| deg_143 | float | Bachelor's in Security and protective services |
| deg_144 | float | Bachelor's in Public administration and social service professions |
| deg_145 | float | Bachelor's in Social sciences |
| deg_146 | float | Bachelor's in Construction trades |
| deg_147 | float | Bachelor's in Mechanic and repair technologies/technicians |
| deg_148 | float | Bachelor's in Precision production |
| deg_149 | float | Bachelor's in Transportation and materials moving |
| deg_150 | float | Bachelor's in Visual and performing arts |
| deg_151 | float | Bachelor's in Health professions and related clinical sciences |
| deg_152 | float | Bachelor's in Business, management, and marketing |
| deg_154 | float | Bachelor's in History |
| deg_201 | float | Master's in Agriculture, agriculture operations, and related sciences |
| deg_203 | float | Master's in Natural resources and conservation |
| deg_204 | float | Master's in Architecture and related services |
| deg_205 | float | Master's in Area, ethnic, cultural, and gender studies |
| deg_209 | float | Master's in Communication, journalism, and related programs |
| deg_210 | float | Master's in Communications technologies/technicians and support services |
| deg_211 | float | Master's in Computer and information sciences and support services |
| deg_212 | float | Master's in Personal and culinary services |
| deg_213 | float | Master's in Education |
| deg_214 | float | Master's in Engineering |
| deg_215 | float | Master's in Engineering technologies/technicians |
| deg_216 | float | Master's in Foreign languages, literatures, and linguistics |
| deg_219 | float | Master's in Family and consumer sciences/human sciences |
| deg_222 | float | Master's in Legal professions and studies |
| deg_223 | float | Master's in English language and literature/letters |
| deg_224 | float | Master's in Liberal arts and sciences, general studies and humanities |
| deg_225 | float | Master's in Library science |
| deg_226 | float | Master's in Biological and biomedical sciences |
| deg_227 | float | Master's in Mathematics and statistics |
| deg_230 | float | Master's in Multi/interdisciplinary studies |
| deg_231 | float | Master's in Parks, recreation, leisure, and fitness studies |
| deg_238 | float | Master's in Philosophy and religious studies |
| deg_239 | float | Master's in Theology and religious vocations |
| deg_240 | float | Master's in Physical sciences |

| Variable Name | Storage Type | Variable Label |
|---|---|---|
| deg_241 | float | Master's in Science technologies/technicians |
| deg_242 | float | Master's in Psychology |
| deg_243 | float | Master's in Security and protective services |
| deg_244 | float | Master's in Public administration and social service professions |
| deg_245 | float | Master's in Social sciences |
| deg_248 | float | Master's in Precision production |
| deg_249 | float | Master's in Transportation and materials moving |
| deg_250 | float | Master's in Visual and performing arts |
| deg_251 | float | Master's in Health professions and related clinical sciences |
| deg_252 | float | Master's in Business, management, and marketing |
| deg_254 | float | Master's in History |
| deg_301 | float | Doctorate in Agriculture, agriculture operations, and related sciences |
| deg_303 | float | Doctorate in Natural resources and conservation |
| deg_304 | float | Doctorate in Architecture and related services |
| deg_305 | float | Doctorate in Area, ethnic, cultural, and gender studies |
| deg_309 | float | Doctorate in Communication, journalism, and related programs |
| deg_310 | float | Doctorate in Communications technologies |
| deg_311 | float | Doctorate in Computer and information sciences and support services |
| deg_313 | float | Doctorate in Education |
| deg_314 | float | Doctorate in Engineering |
| deg_315 | float | Doctorate in Engineering technologies/technicians |
| deg_316 | float | Doctorate in Foreign languages, literatures, and linguistics |
| deg_319 | float | Doctorate in Family and consumer sciences/human sciences |
| deg_322 | float | Doctorate in Legal professions and studies |
| deg_323 | float | Doctorate in English language and literature/letters |
| deg_324 | float | Doctorate in Liberal arts and sciences, general studies and humanities |
| deg_325 | float | Doctorate in Library science |
| deg_326 | float | Doctorate in Biological and biomedical sciences |
| deg_327 | float | Doctorate in Mathematics and statistics |
| deg_330 | float | Doctorate in Multi/interdisciplinary studies |
| deg_331 | float | Doctorate in Parks, recreation, leisure, and fitness studies |
| deg_338 | float | Doctorate in Philosophy and religious studies |
| deg_339 | float | Doctorate in Theology and religious vocations |
| deg_340 | float | Doctorate in Physical sciences |
| deg_341 | float | Doctorate in Science technologies/technicians |
| deg_342 | float | Doctorate in Psychology |
| deg_343 | float | Doctorate in Security and protective services |
| deg_344 | float | Doctorate in Public administration and social service professions |
| deg_345 | float | Doctorate in Social sciences |
| deg_350 | float | Doctorate in Visual and performing arts |
| deg_351 | float | Doctorate in Health professions and related clinical sciences |
| deg_352 | float | Doctorate in Business, management, and marketing |
| deg_354 | float | Doctorate in History |
| deg_422 | float | Professional degree in Legal professions and studies |
| deg_439 | float | Professional degree in Theology and religious vocations |
| deg_451 | float | Professional degree in Health professions and related clinical sciences |