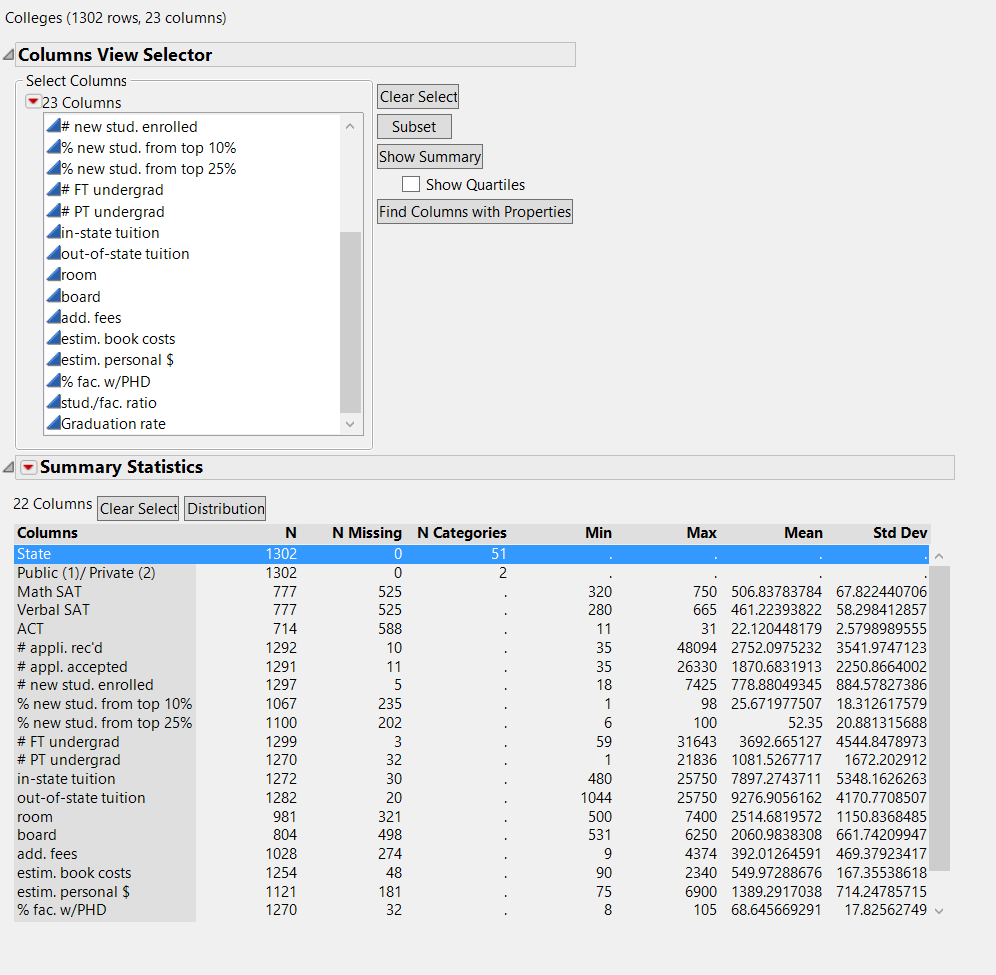
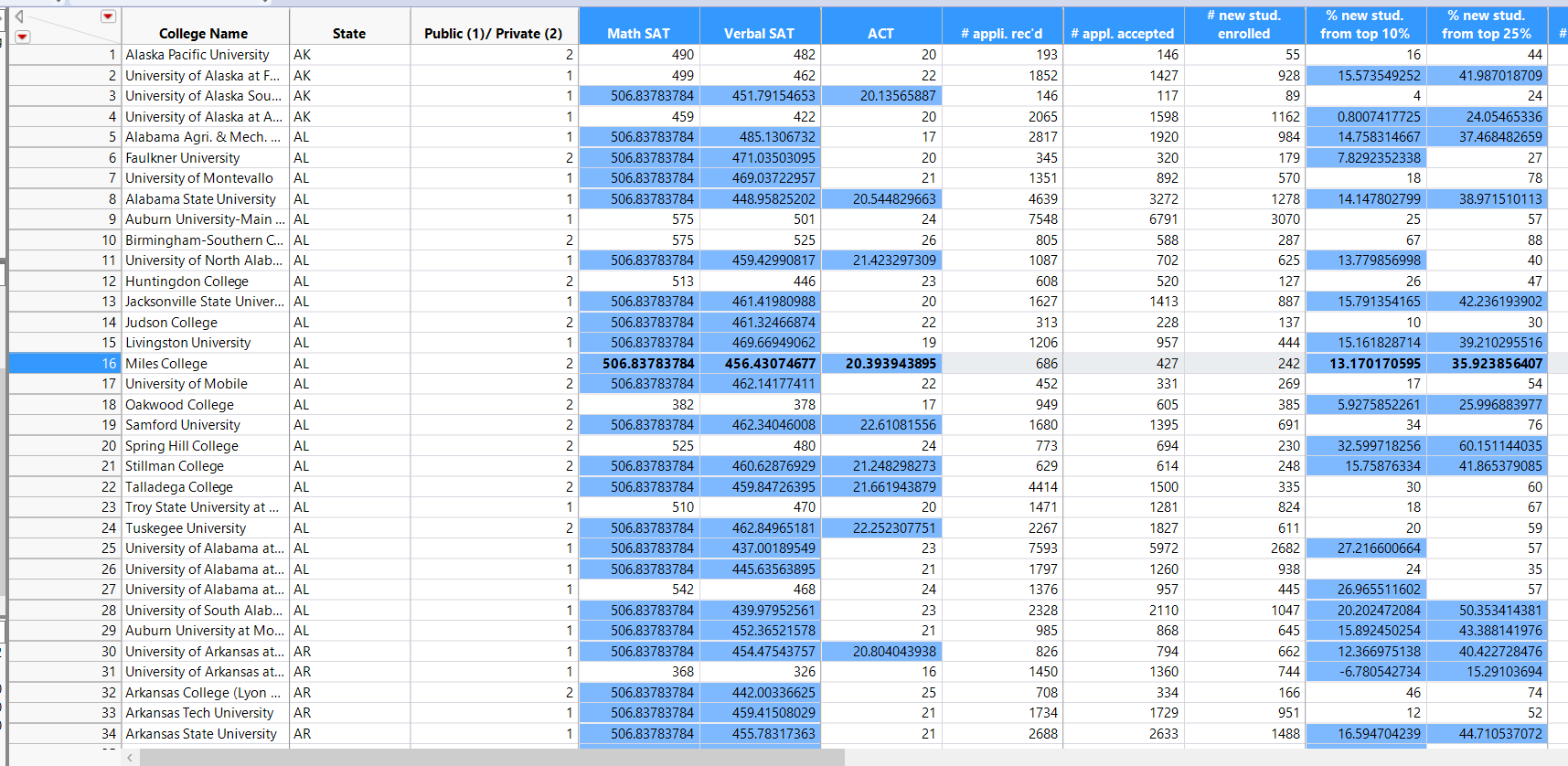
**OPIM 5604 B15 – Predictive Modeling Assignment Meghana Kasula (Net ID=mek15120)**

*“The work contained and presented here is my work and my work alone.”*

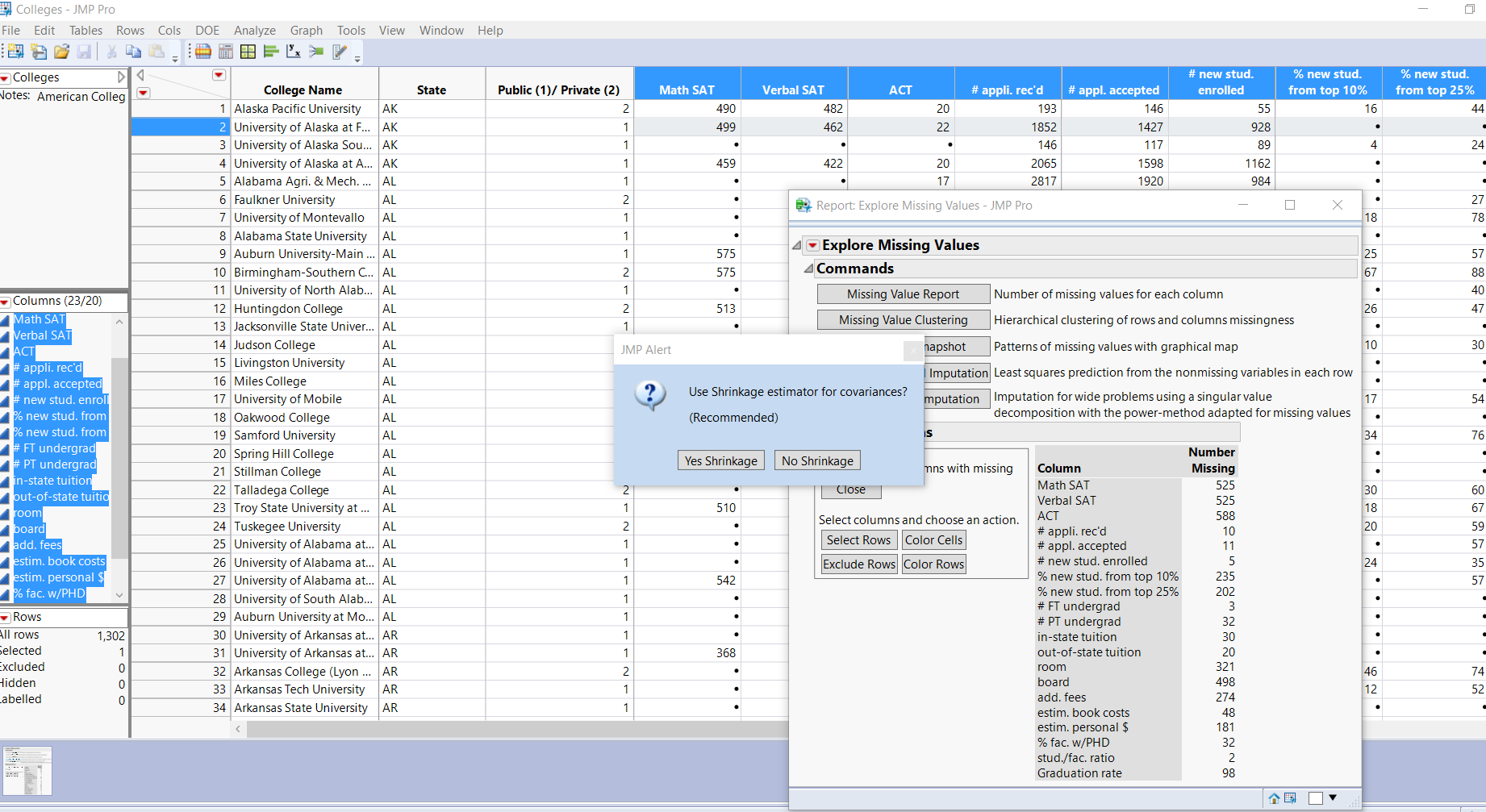
**14.4 University Rankings.**

**a. Use the *Columns Viewer* to produce numeric summaries of all of the variables. Note that many observations are missing some measurements. Clustering methods in JMP omit records with missing values. For the purposes of this exercise, we will assume that the values are missing purely at random (not that this may in fact not be the case). So our first goal is to estimate (impute) these missing values.**

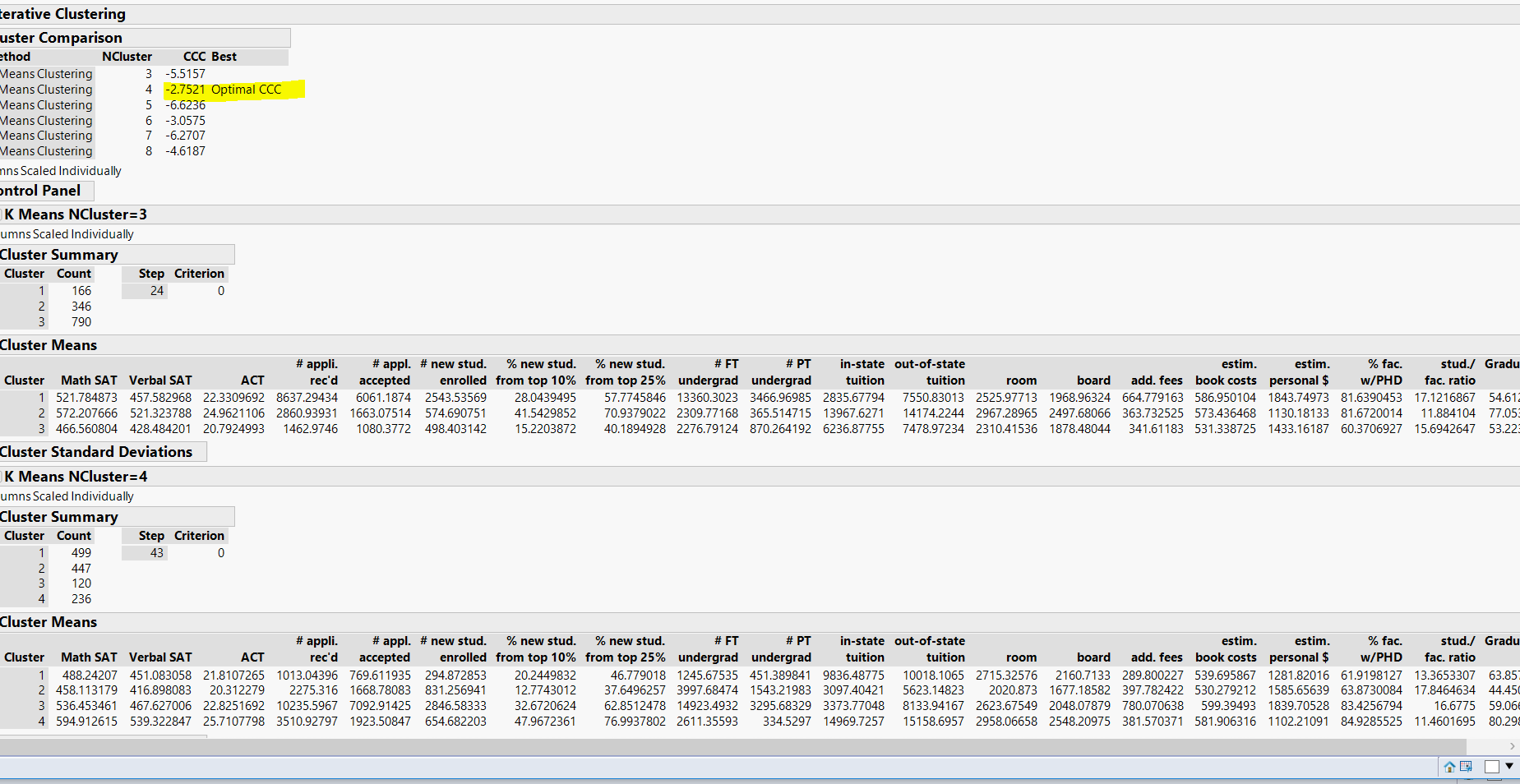
Here, we can see the number of missing values in each. Most are continuous values.

**b. Select all of the continuous columns, and go to *Cols > Modeling Utilities >Explore Missing Values*. Select the option *Multivariate Normal Imputation*,and click *Yes Shrinkage*. Multivariate normal imputation uses least squares regression to predict the missing values from the nonmissing variables in each row. The imputed values will be highlighted in the data table. Save the data table with imputed values under a new name.**

Using the imputation method, we successfully filled the missing values with normalizations.



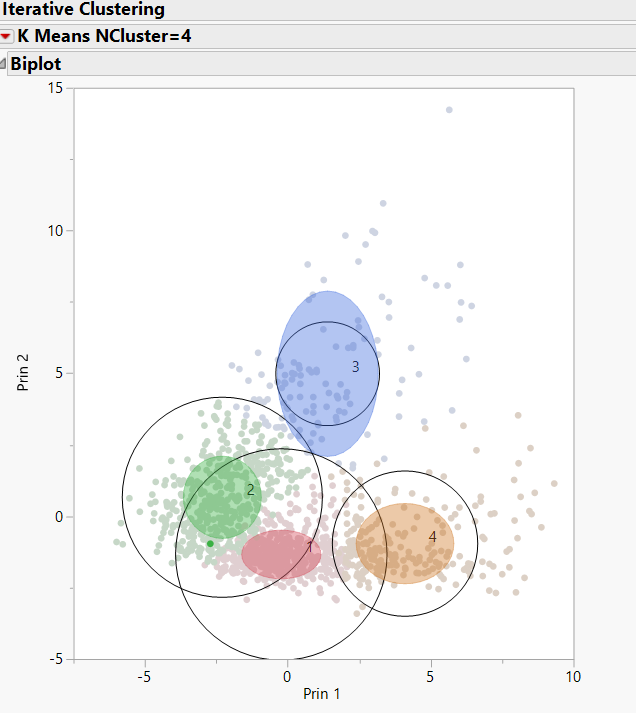
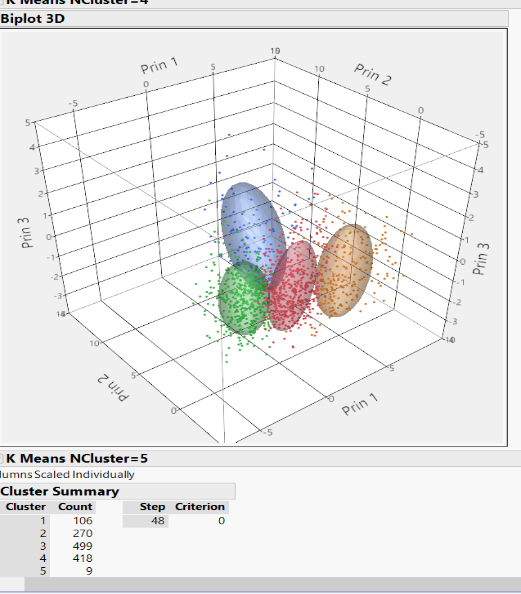
**c. Use *k*-means clustering, with only the continuous variables. Use College Name as the label. Form clusters starting with 3 and ending with 8. What is the optimal number of clusters based on the Cubic Clustering Criterion (under *Cluster Comparison*)?**

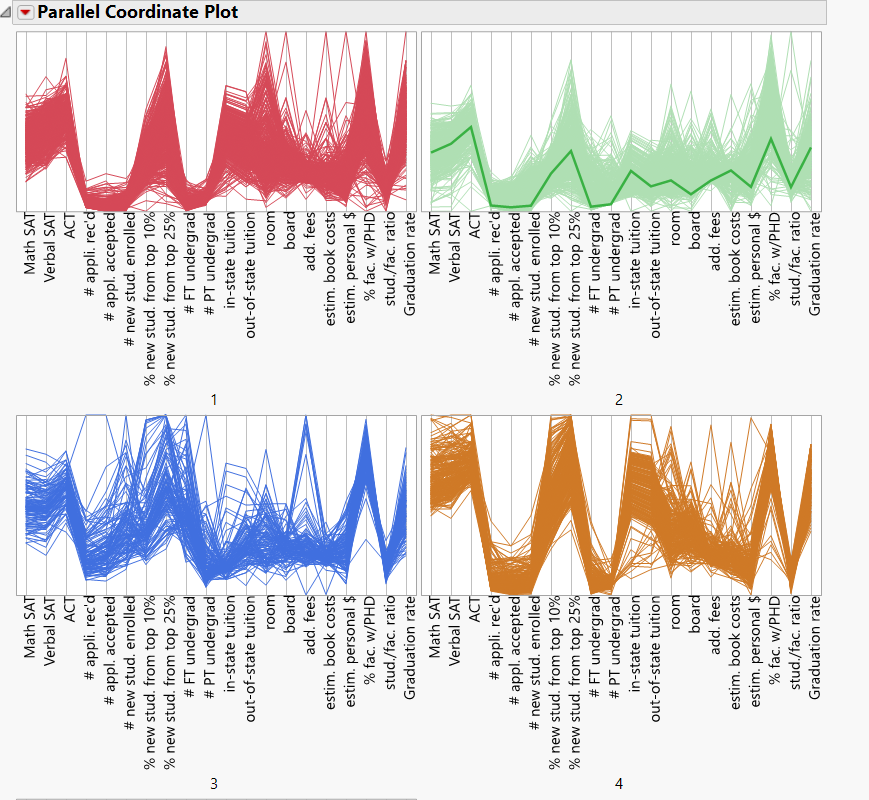
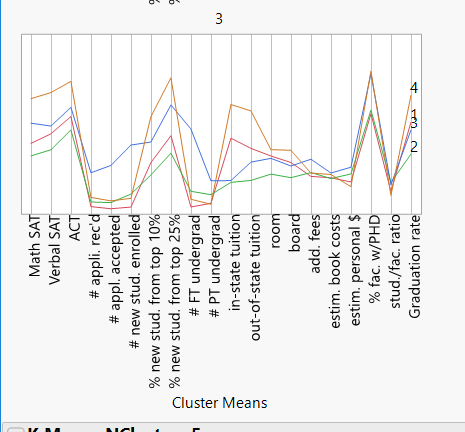


As we can see, the optimal number of clusters are 4.

**d. Use the biplot, the parallel plot, and other built-in graphical and numeric summaries to explore the clusters. Save the clusters to the data table and use other graphical tools to compare and characterize the clusters. Summarize the key characteristics of each of the four clusters, and try label or name the clusters.**

Following is the Biplot: Biplot





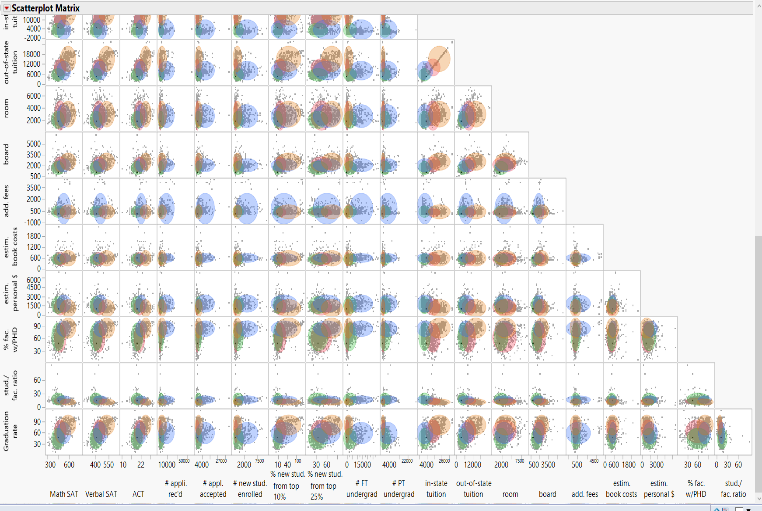
The four key characteristics of clusters are:

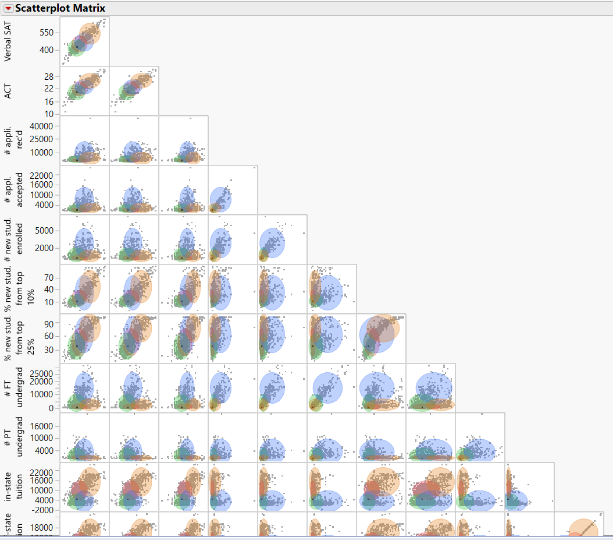
1.Cluster 4 is having highest SAT and ACT scores, also high Graduation rate.

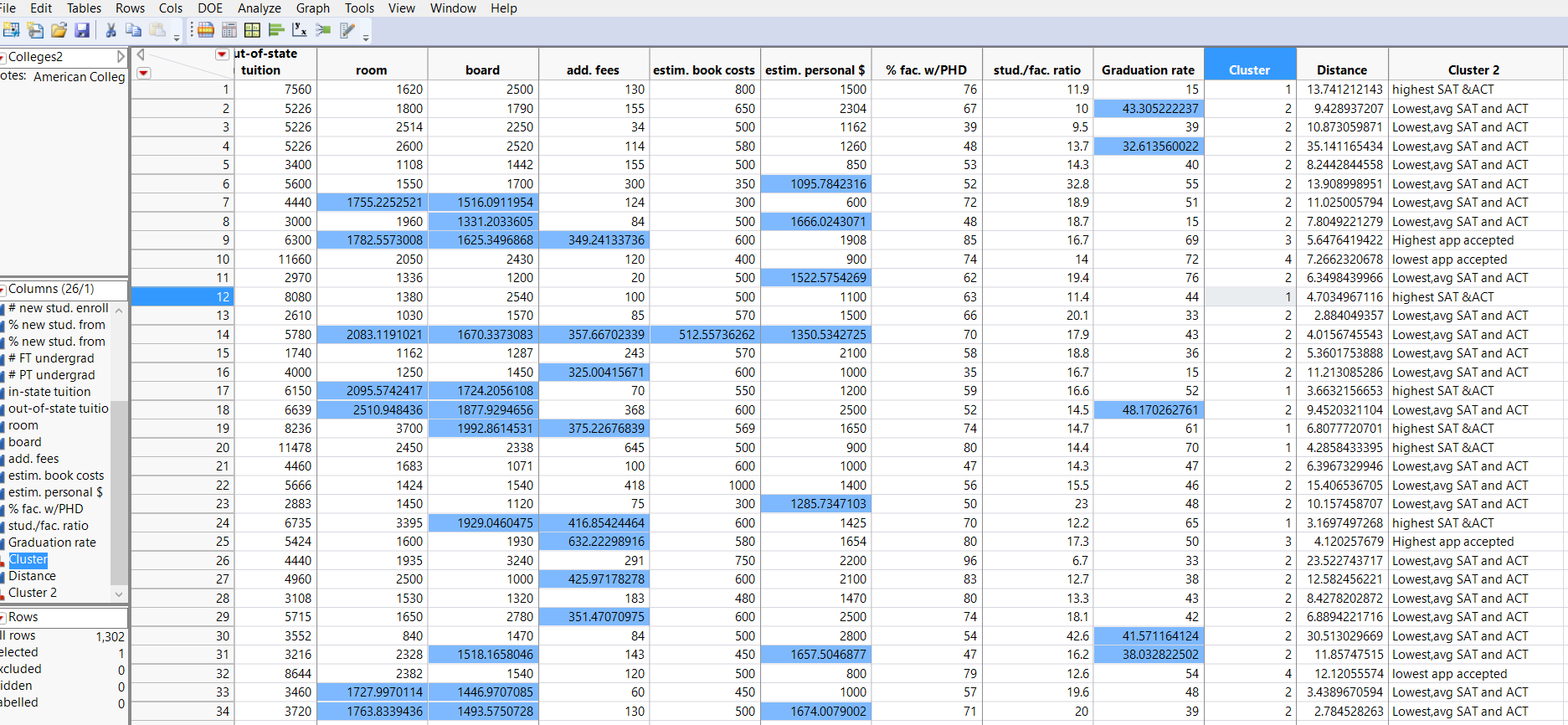
2.cluster 2 is relatively low or average in every field starting from Sat score to graduation rate.

3.Cluster 3 has high application accepted and enrolled.

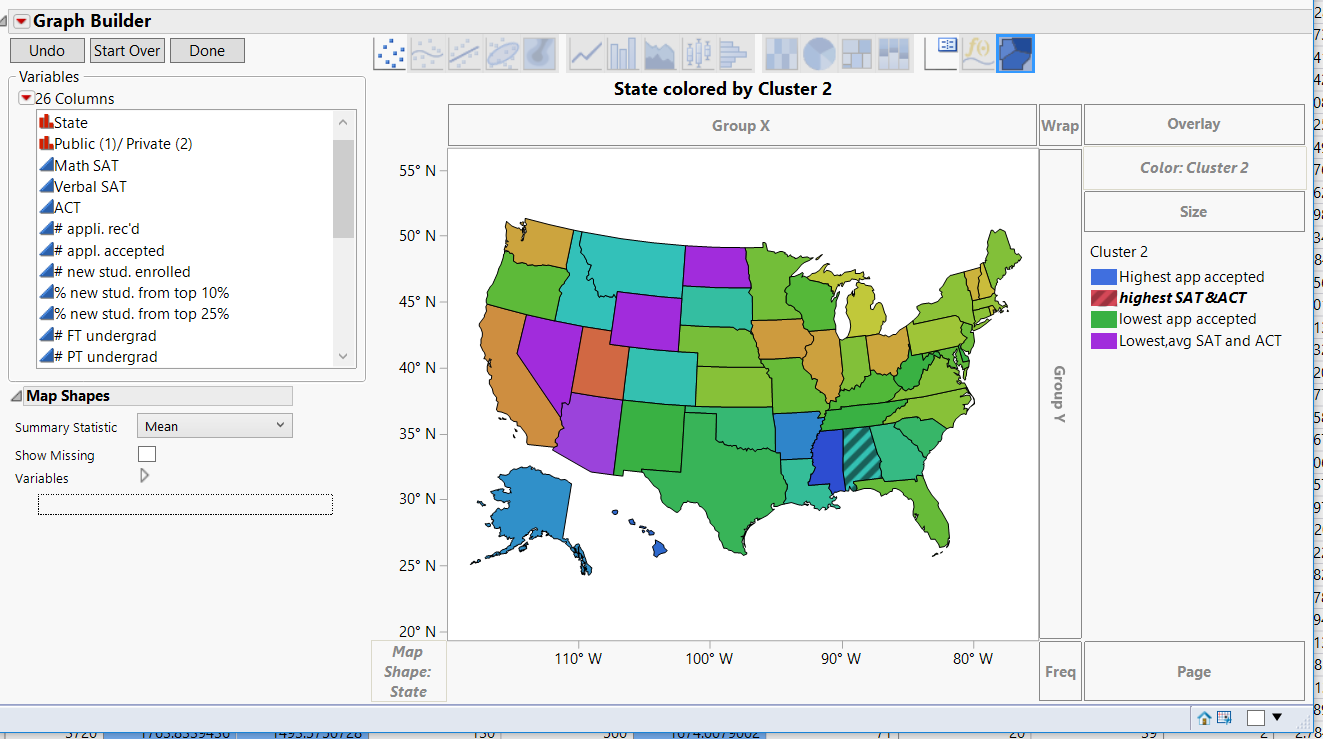
4.Cluster 1has lowest application received, accepted and enrolled.





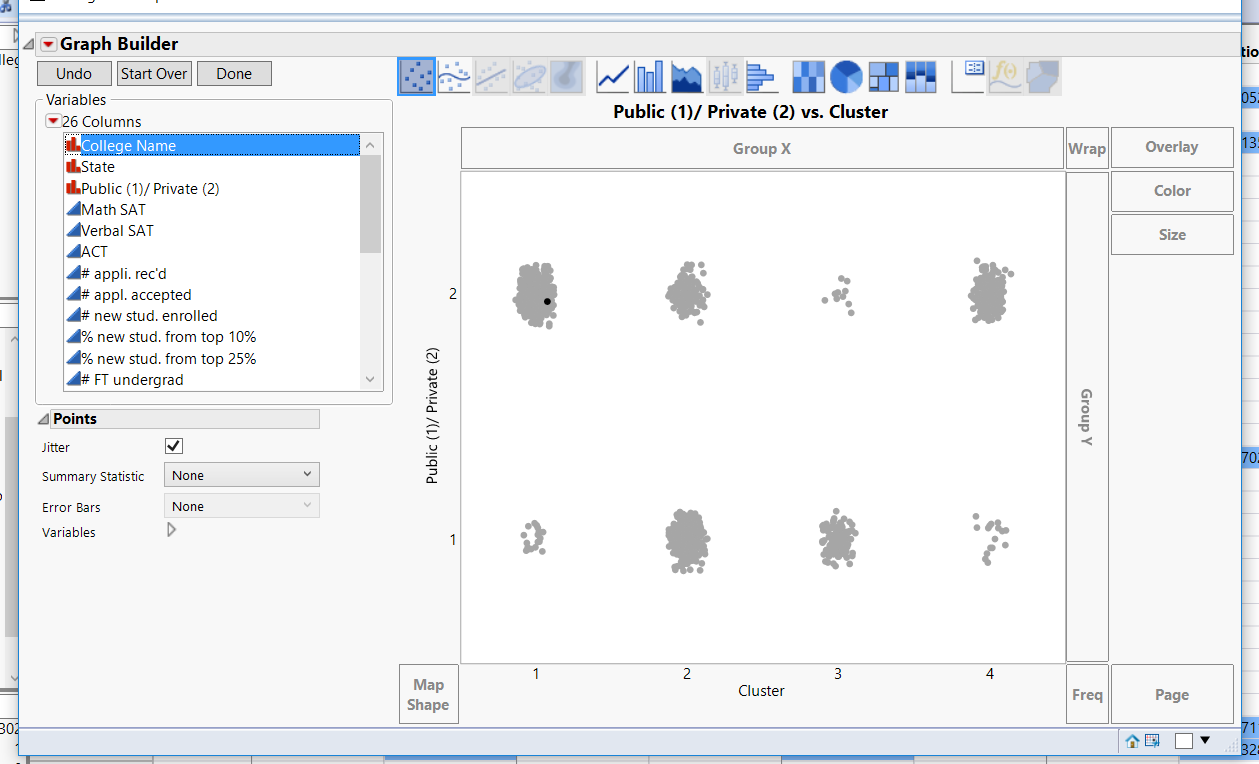


**e. Use the categorical variables that were not used in the analysis (State and Private/Public) to characterize the different clusters. (Hint: Create a geographic map to explore the clusters geographically.) Is there any relationship between the clusters and the categorical information?**



The clusters spread can be viewed in the graph beside.

The legend is rightly mentioned and denoted.



In general, the amount of observations in private colleges are high and also, clust 1 and 4 dominantly has private institutes.

**f. Can you think of other external information that might explain the contents of some or all of these clusters?**

1.Private institutes are in general have less rigorous criteria in selection than public institute. That is why we can observe more observations in private institutes.

2. In cluster 1, 2 we observed that low scores of ACT and SAT led to low acceptance rate which is a natural phenomenon.

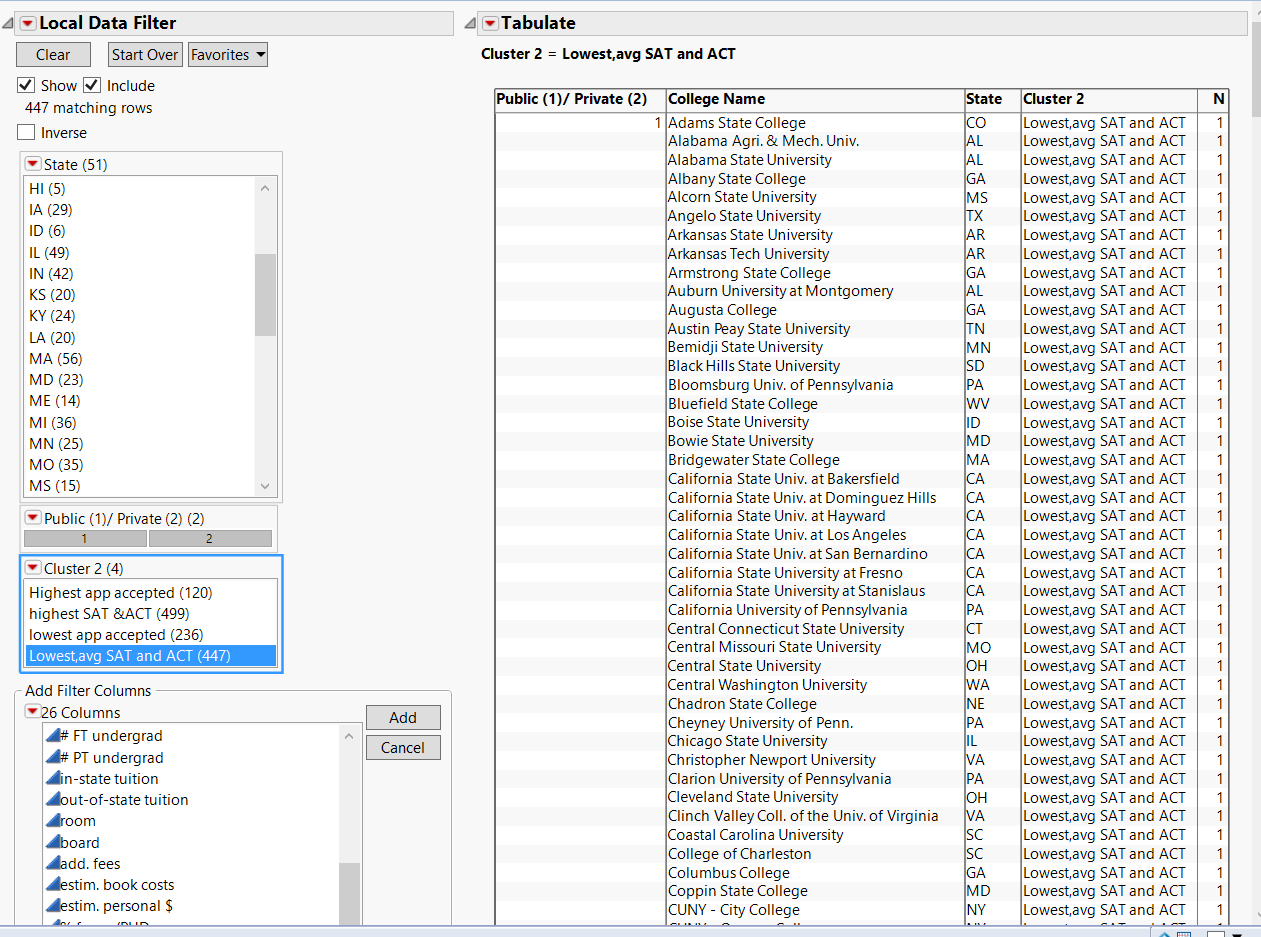
3. Similarly, high score in SAT and ACT led to more acceptance range, observed in cluster 4.

4. Cluster 4, could have the observations of good academicians.

**g. Consider Tufts University. Which cluster does Tufts belong to? Which other universities is Tufts similar to, based on the clustering and the categorical variables?**

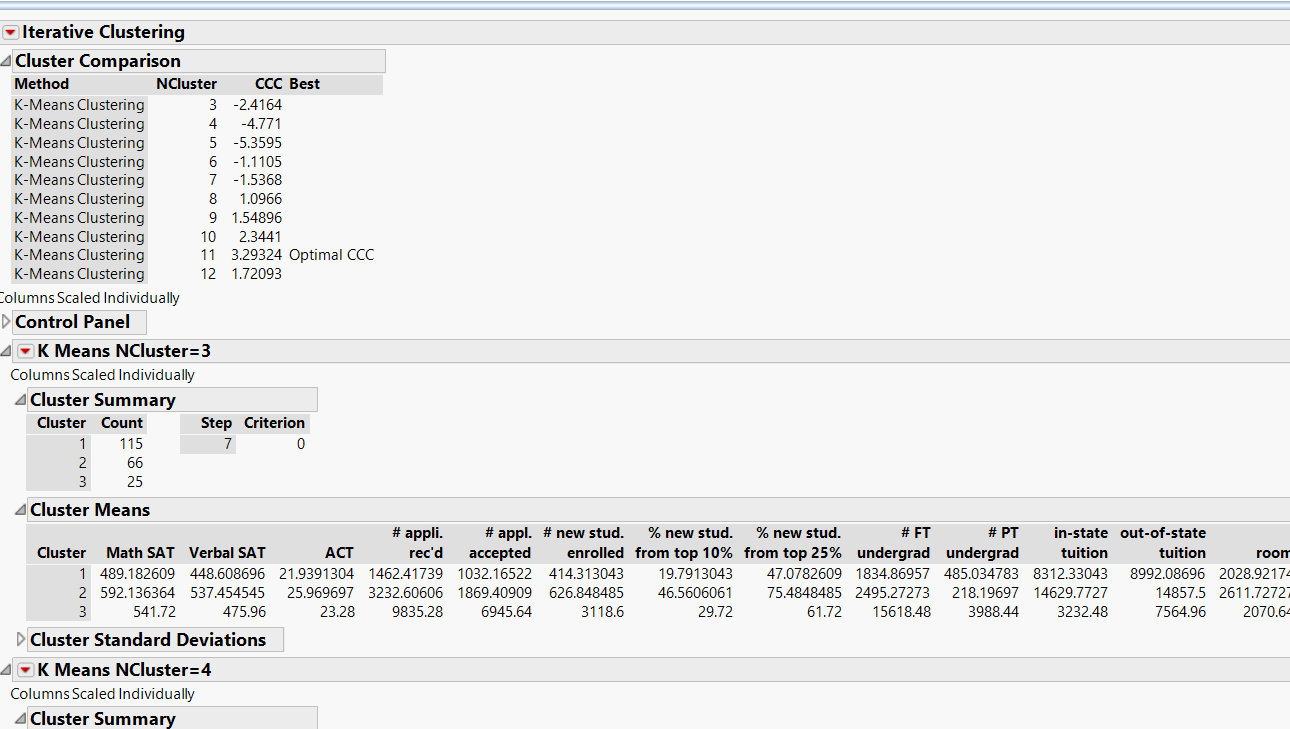
**Tufts is from cluster 4.**

**Hence, using local data filter, following are from the same cluster as Tufts.**



**h. Return to the original data table (with the missing values). Run the same *k*means cluster analysis using this data.**

**i. Compare the results to those achieved after imputing missing values in terms of the number of clusters and the characteristics of the clusters? What are the key differences?**



**The main difference is that , the number of optimum clusters has drastically increased to 11.**

**Also, due to a lot of missing values, the clusters has to be more in lesser observations due to the vagueness and obscurity.**

**ii. In the initial analysis, we assumed that the values were missing at random and imputed the missing values. Describe why this approach was, or was not, a good strategy.**

This approach is good, but the rows with missing values will be not considered for analysis. Hence it would omit some useful data.

**Compute the metrics for the following rules: (Please utilize Bank-2.jmp)**

**a. CKING 🡪 SVG**

Formula for

Support =P[A intersection B]/Total = (4329/7991) =0.543425

Confidence = P(A intersection B)/P (A) = (4329/7991)/(6855/7991)= 0.634562

Expected Confidence =P(B) = 4944/7991=0.6134

Lift =confidence/extreme confidence =0.63/0.61 =1.02001

**b. (CKING,SVG) 🡪CD**

Support =P(A intersection B)/Total = (1139/7991) =0.1467

Confidence = P(A intersection B)/P (A) = (1139/7991)/(4329/7991)= 0.2623

Extreme Confidence =P(B) = 1960/7991=0.24345

Lift =confidence/extreme confidence =0.26/0.24 =1.0700