Project 2: Dimensionality Reduction and Clustering

CS425

Samuel Steinberg

October 9th, 2019

Description:

This paper analyzes the application of dimensionality reduction and *k*-means clustering to visualize information about a set of universities. Dimensionality reduction is the process of reducing the number of random variables under consideration, which in this project was done through feature selection and feature extraction. Feature extraction was performed with Principal Component Analysis (PCA) to map the data to a lower dimensional space, while feature selection was used to lower the dimensionality and reduce overfitting (reduction of variance). In the PCA, Singular Value Decomposition is used to factor the data matrix and extract singular values. The *k*-means clustering algorithm was performed to cluster the original data, which is where vectorized data is partitioned into clusters depending on its mean. The dataset used for the purposes of this project is called *UTK-peers.csv*, which consisted of 65 attributes for 57 schools and universities across the nation. The dataset contained both numeric and non-numeric attributes along with numerous illegal or non-existent values.

Data Pre-Processing:

As mentioned in the Description, data was given in the form of both numeric and non-numeric attributes. Non-numeric string attributes were dropped. Numeric values read in as strings that contained illegal values were stripped of any prepending dollar-signs, commas, dashed, spaces, etc. They were then converted to their proper numerical values. To keep the type of the data consistent, all data was converted to its float value. Additionally, there were numerical columns with missing data. To best preserve the bulk of the data, mean imputation was applied to still be able to use the other data that the school or university had, while providing a realistic number as a placeholder.

Solution Description:

After data pre-processing was completed, it was possible to begin computation. First, the data was placed in a matrix and SVD was performed:

$$X = U\Sigma V^T$$

Figure 1: SVD Formula

Here, X represents the matrix derived from the dataset. Performing SVD produces three matrices, U, Σ, and V^T. U represents the left singular vectors, Σ represents the singular values,

and V^T represents the right singular vectors. The singular value matrix is returned as a diagonal matrix, and to extract the actual single values the following formula is used:

$$Singular\ Value = \frac{Cumulative\ Sum\ Singular\ Matrix\ Values\ Squared}{Sum\ of\ Singular\ Matrix\ Values\ Squared}$$

Figure 2: Extraction of singular values from S matrix

Here, the cumulative sum at the position in the matrix over the sum of the entire matrix squared will yield the singular values necessary to plot Scree graphs of the singular values and provides the data necessary to plot the percentage variance. This information is necessary to find a good *p* to limit the number of PC's.

After the matrix calculations and dimensionality reduction came *k*-means clustering. The general outline of the algorithm for performing this algorithm is provided in Figure 3 below:

Initialize $\boldsymbol{m}_i, i = 1, \ldots, k$, for example, to k random $\boldsymbol{x}^t$
Repeat
  For all $\boldsymbol{x}^t \in \mathcal{X}$
    $b_i^t \leftarrow \begin{cases} 1 & \text{if } \|\boldsymbol{x}^t - \boldsymbol{m}_i\| = \min_j \|\boldsymbol{x}^t - \boldsymbol{m}_j\| \\ 0 & \text{otherwise} \end{cases}$
  For all $\boldsymbol{m}_i, i = 1, \ldots, k$
    $\boldsymbol{m}_i \leftarrow \sum_t b_i^t \boldsymbol{x}^t / \sum_t b_i^t$
Until $\boldsymbol{m}_i$ converge

Figure 3: k-means algorithm

Here, m_i represents a set of means, initialized randomly. Then, until convergence (a limit of 1000 iterations until convergence was placed) for each item x^t in set X. The value at b_i^t depends heavily on m_i, and the Euclidean distance formula is used to find the mean closest to the item. See Figure 4 below:

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Figure 4: Euclidean Distance Formula for n-dimensional space

Here, q_i represents singular values with p_i represents the mean value at the index. The Euclidean is necessary here because we are in *n*-dimensional space. After this items are then assigned to the mean (see Figure 5) and the mean is updated.

$$\boldsymbol{m}_i = \frac{\sum_t b_i^t \boldsymbol{x}^t}{\sum_t b_i^t}$$

Figure 5: Mean is assigned to this division of summations for each element

After the mean is updated, a check is performed to determine convergence between the "old" means and the "new" means. If the mean arrays are equivalent, convergence has been reached and the loop will break. If not, further iterations will be performed until the algorithm either convergences or is halted due to too many iterations.

After the algorithm is completed, it is possible to perform calculations for minimal *intercluster* distances, maximal *intracluster* distances, and the Dunn Index. The minimal *intercluster* distance is the minimal distance between two points in different clusters, while the maximal *intracluster* distance is the maximal distance distinct points within a cluster. See Figures 6 and 7 below:

$$SSW(C,k) = \sum_{i=1}^{N} \| x_i - c_{p(i)} \|^2$$

Figure 6: Intercluster calculation

$$SSB(C,k) = \sum_{j=1}^{k} n_j \| c_j - \overline{x} \|^2$$

Figure 7: Intracluster calculation

Here, *c* and *x* represent sub-cluster points for distance computation. *C and* k represent the main clusters and number of clusters respectively. After calculating these there is a check to see if the calculated distance is either a maximum or minimum. If so, this distance is set as the new metric for all further calculations. This occurs until iteration is complete, and the metric most qualified is passed as either the minimal *intercluster* or maximal *intracluster*.

The Dunn Index is a metric for evaluating clustering (such as compact clusters or more heavily separated clusters). This is an internal evaluation, and is found through the ratio of the minimal *intercluster* data to the maximal *intracluster* data:

$$Dunn\ Index = \frac{Minimum\ Intercluster}{Maximum\ Intracluster}$$

Figure 8: Dunn Index Calculation

Analysis and Discussion:

After performing SVD on the data matrix, it was possible to use the singular value matrix to plot singular values along with the percentage variance (square of the singular value). Plotting these to visualize the data was an important decision-maker in determining the appropriate *p* best value. See Figures 9 and 10 below:

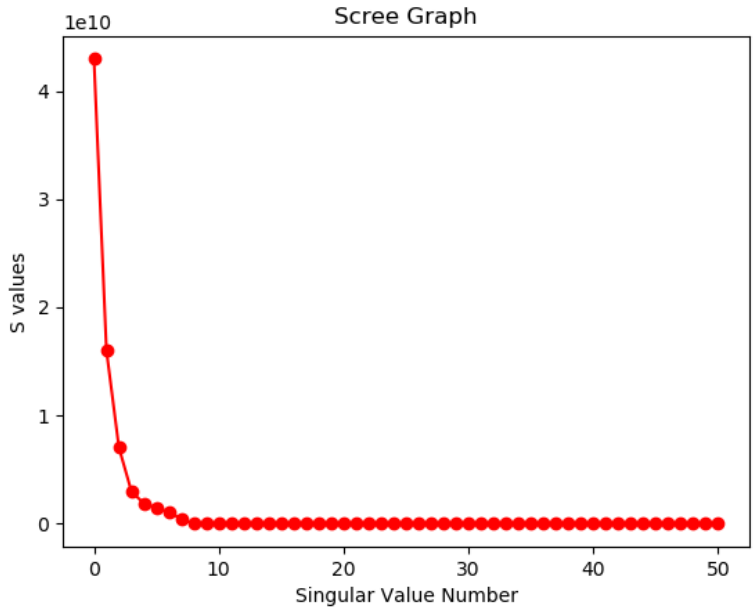

Figure 9: Scree Graph for Singular Values

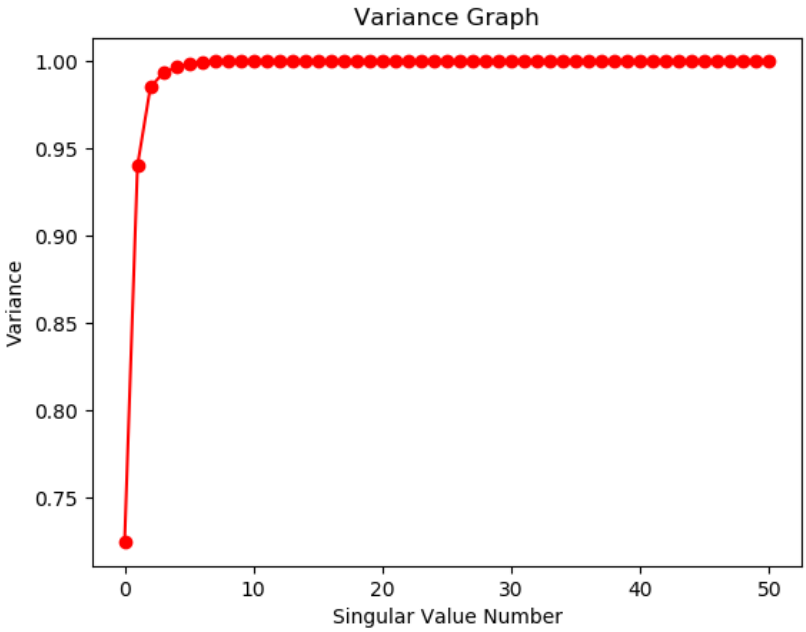

Figure 10: Scree Graph for Percentage Variance

After considering the Figures 9 and 10, it became clear that the optimal *p* value would be four. This is due to a higher percentage variance at around 98%, though not quite at the constant 100% mark that would signify overfitting the data. After finding this value, the data matrix was then reduced to this number of *p* PC's. The first two PC's are graphed below in Figures 10 and 11:

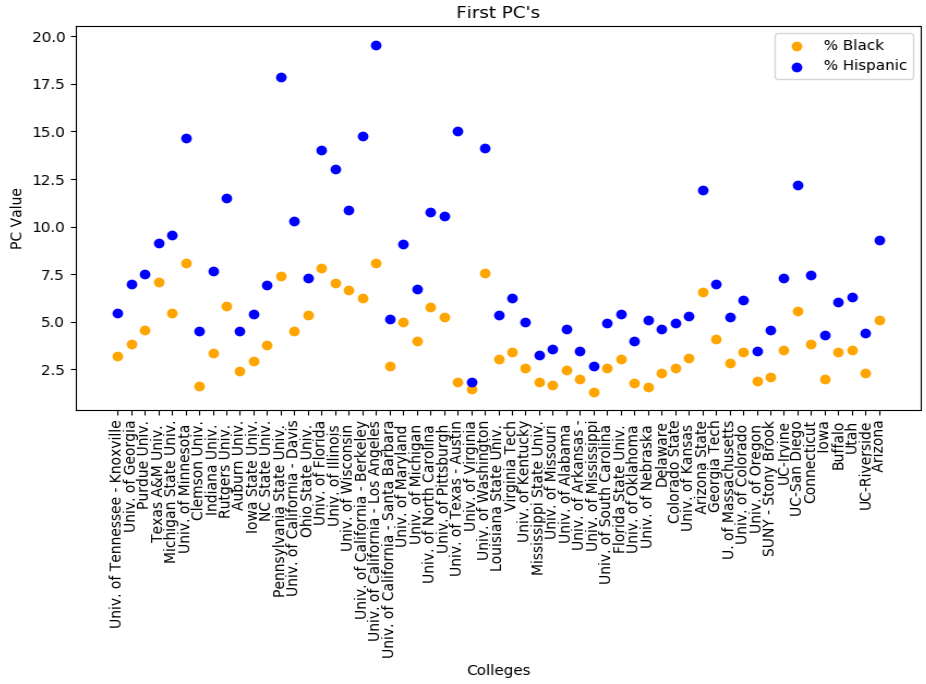

Figure 11: First Two PC's (Black and Hispanic Enrollment) plotted on top of each other

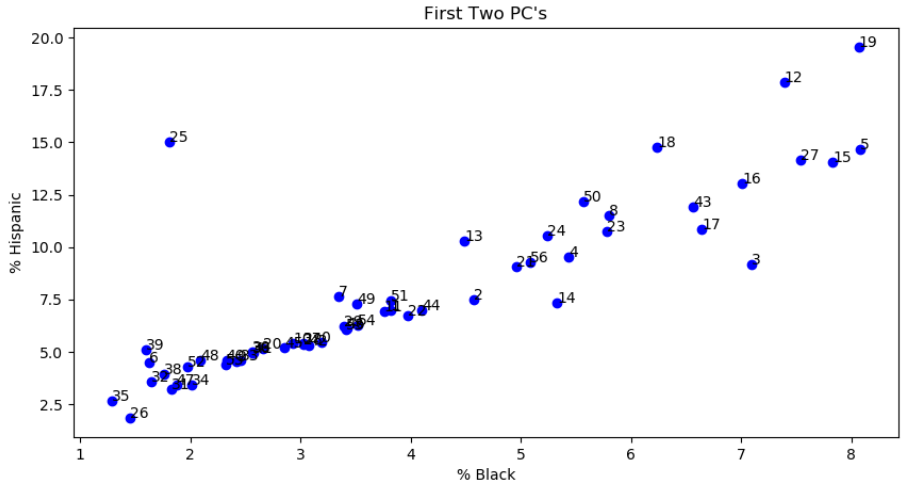

Figure 12: First Two PC's plotted against each other. Note that the numbered annotations represent the school in the order of their appearance in the dataset.

When implementing the *k*-means aspect of the project, there were different combinations of *k* number of clusters that were tested. Ideally, the best data was produced with *k*'s of 2, 3, and 4. Statistics were kept for the number of iterations to converge in the *k*-means algorithm, minimal *intercluster* distance, maximal *intracluster* distance, and the Dunn Index for the original data, *p*-dimensional data, and the first two PC's. Their results are displayed in Tables 1, 2, and 3 below:

Table 1: Data Clustering Statistics for Original Data

k	Iterations to Converge	Minimal Intercluster	Maximal Intracluster	Dunn Index
2	2	5044455563.477482	10524479017.398981	0.47930691439814077
3	2	5044455563.477482	9669286117.139622	0.5216988619806958
4	5	955542216.0995947	8460508782.803454	0.11294146021593851

Table 2: Data Clustering Statistics for p-Dimensional Data

k	Iterations to Converge	Minimal Intercluster	Maximal Intracluster	Dunn Index
2	3	5619.903762394874	25476.4527834719	0.22059208203588068
3	5	3739.091263034262	20148.81336938451	0.18557377025069346
4	5	2542.3791340421535	17009.06313194602	0.14947202643202126

Table 3: Data Clustering Statistics for First Two PC's

k	Iterations to Converge	Minimal Intercluster	Maximal Intracluster	Dunn Index
2	5	1.627915410003039	10.937187988885237	0.1488422263252113
3	9	0.5763435073272482	9.113302776077784	0.06324200144432125
4	7	0.19397845864051086	7.7361828955512655	0.025074182094642468

The minimal *intercluster* distance is the minimal distance between two objects belonging to two different clusters, while the maximal *intracluster* distance is the maximal distance between two objects in the same cluster. These values are expected to be fairly large as clusters are quite far separated from each other. The correct values for these measurables were confirmed by examining the Dunn Index (their ratio between one another, see Figure 8).

After the gathering of statistics, the clusters were plotted for the original data, *p*-dimensional data, and the first two PC's. See the Figures 13-18 on the following pages:

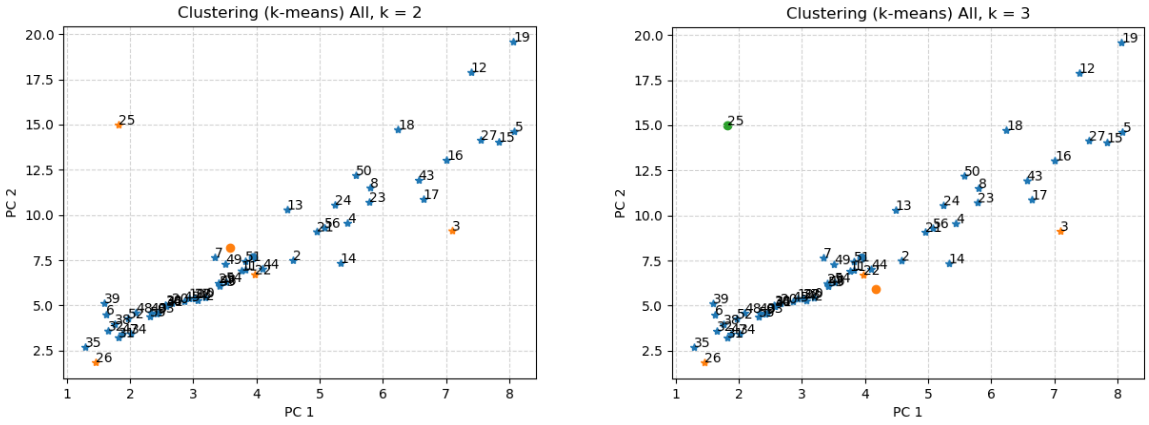

Figure 13: Original data, the left graph represents the results for two clusters and the right graph represents three clusters

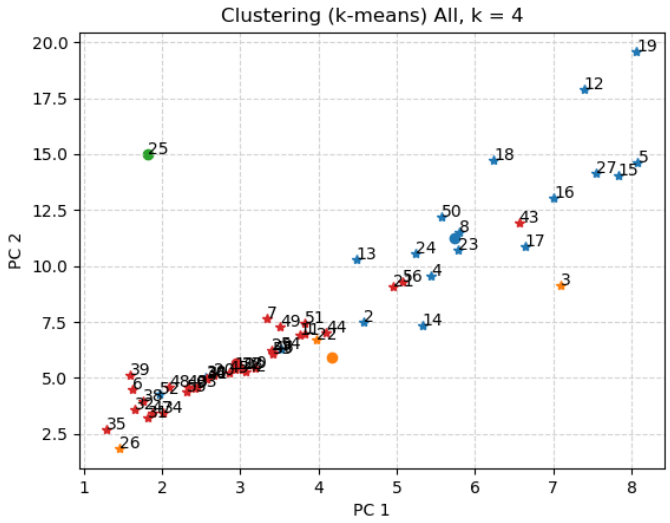

Figure 14: Original data, four clusters. Note that k-means is prone to outliers (see point 25).

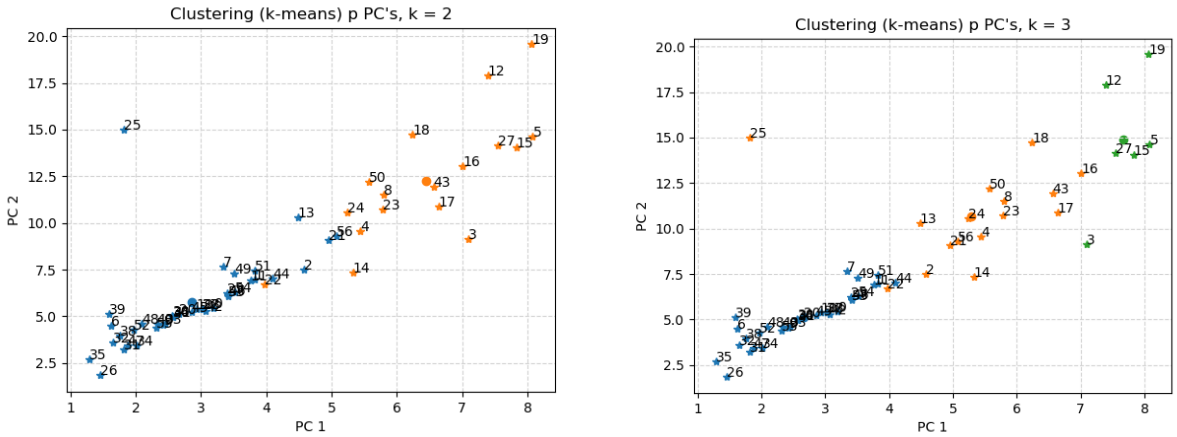

Figure 15: p-dimensional data, the left graph represents the results for two clusters and the right graph represents three clusters

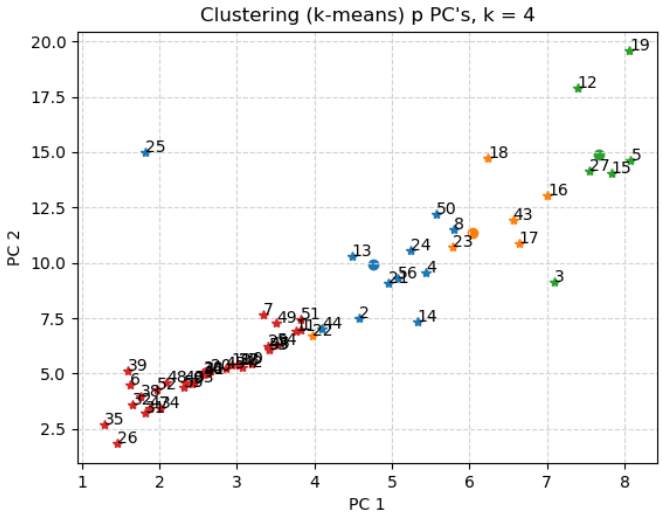

Figure 16: p-dimensional data, four clusters

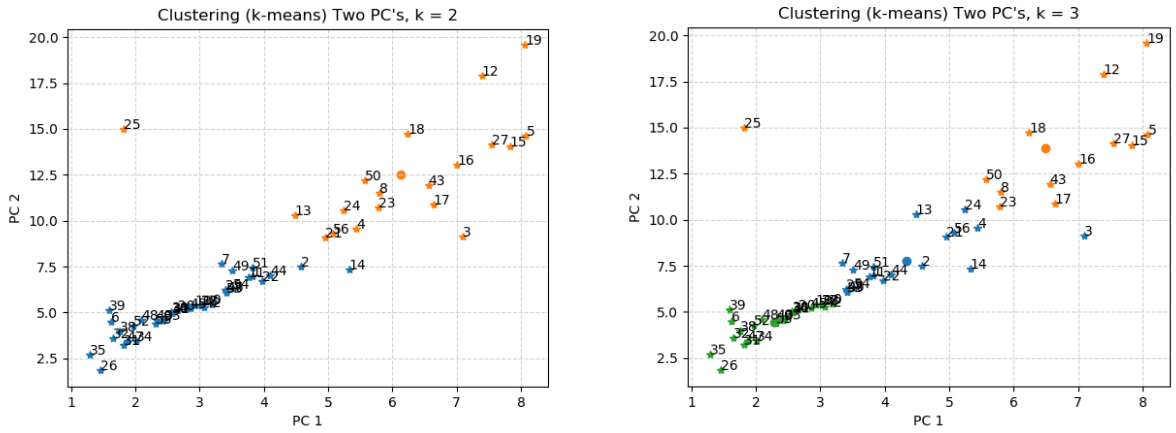

Figure 17: First Two PC data, the left graph represents the results for two clusters and the right graph represents three clusters

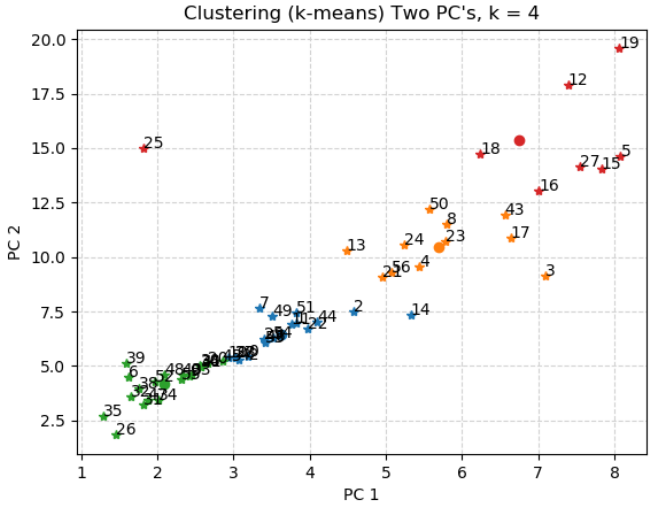

Figure 18: First Two PC data, four clusters

Overall, the best clustering was performed with *p*-dimensional data. The original data proved to be prone to outliers when the number of clusters was greater than two, as seen in Figure 14 where a cluster was formed around a single school (University of Texas – Austin has a high Hispanic population coupled with a low Black population). Additionally, even with two clusters this school pulled the centroid towards it and slightly skewed the data. The first two PC's, like the *p*-dimensional data, performed exceptionally well. The difference, however, was that the cluster centroid points were stronger with *p*-dimensional data and resulted in better performance and results.

For the most part, UTK was clustered with schools such as Rutgers, University of Oregon, University of Georgia, University of Minnesota, University of Colorado, Purdue University, University of Missouri, and University of Kansas. This changed according to the different number of clusters, though these schools largely stayed consistent.

In summation, dimensionality reduction using SVD revealed that the ideal *p*-dimensions for this project would be four. The visualization with singular values and percentage variance confirmed this value would not be as prone to overfitting. The implementation of the *k*-means algorithm with cluster sizes (*k*) of 2, 3, and 4 proved to best capture the structure of the data. Through this algorithm it became possible to collect iteration data, minimal *intercluster* distance, maximal *intracluster* distance, and the Dunn Index for the original data, *p*-dimensional data, and first two PC data. These measurables are of great importance in analyzing special data and distances. Lastly, clustering was performed with *k* clusters on the original data, *p*-dimensional data, and first two PC's. The most optimal data for the clustering proved to be the *p*-dimensional data, with accurate clustering and centrally placed centroids. The schools most associated with UTK after analyzing the clustering data were Rutgers, University of Oregon, University of Georgia, University of Minnesota, University of Colorado, Purdue University, University of Missouri, and University of Kansas.