COSC 425 Project 2
Ryan Pauly
University of Tennessee Knoxville

Project Description

The goal of project 2 is to apply dimensionality reduction and clustering to visualize information about a selection of universities from the provided dataset 'UTK-peers.' Part 1 of project 2 deals with data reduction and principal component analysis (PCA), a numerical analysis technique which attempts to find correlated variables into a set of variables called principal components. Part 2 of project 2 builds off part 1's data visualization by utilizing the k-means clustering method.

Pre-processing Steps

The data file 'UTK-peers.csv' is an extensive dataset with various stats for several universities. With the file open in Excel, the data file has quite a few blank cells throughout the dataset. Additionally, there were some columns which had information which was not numerically significant, such as containing only 1s or 2s like that of a true/false value, these columns were ignored. Further, some of the columns had non-numeric values, and thus were ignored as well. Upon early on file I/O, there were a plethora of "not a number" or Na values littered along the rows underneath the rows of university data, which was of course ignored as well.

With these non-numeric and inapplicable attributes ignored, the remaining data is extracted. If the attribute column did contain a missing value, then the attribute was skipped. Only columns with complete numerical values were extracted. The data for this project's analyses do not need a training and testing set, so the data that is collected, is simply the data. With all this considered, 50 of the 65 attributes (columns) of data were captured for analysis.

Description

In part 1 of this project the focus is on data reduction and the application of PCA. Singular value decomposition, a method which generalizes data for statistical analysis, is a key formula to factor the data matrix and extract singular values. The singular value decomposition (SVD) equation is represented as the following:

$$X = U\Sigma V^T \tag{1}$$

In equation 1, X represents the data matrix of $m \times n$ dimensions. U is a $m \times m$ matrix. Σ is a matrix of $m \times n$ matrix consisting of a diagonal list of values, these values are the singular values sought after in part 1. V^T, or the transpose of V, is a $m \times n$ matrix.

Part 1 requests a percentage of variance covered by the first p singular values, which is derived from the singular diagonal matrix Σ found in equation 1. To find the percentage of variance covered by the first p singular values, first find the sum of the overall variance, which is the sum of Σ^2. Then by taking the summation of each Σ_j for the *i-th* column, the variance percentages can then be found by dividing the individually summed singular value variance by the total overall variance, multiplied by 100.

Equation 1 is necessary as it produces the three factored SVD components required to reduce the data matrix. To find the reduced matrix, first take advantage of the factored V^T matrix. First, the transform of V^T needs to be undone. By simply taking the transform again, the result is the V matrix. Then by taking the dot product of the original data matrix and V where only the first *p* columns of matrix *V* are used, provides the reduced matrix. Thus, the data reduction matrix is the following in python pseudo code:

$$reduced\ matrix = X \cdot V[:, :p] \tag{2}$$

In equation 2, *X* represents the original data matrix, and *V* represents the untransformed V^T matrix consisting of the first *p* principal component columns.

In part 2 of this project, the focus shifts to cluster analysis via the k-means algorithm. The algorithm is best described in pseudo code, which in turn is a set of equations:

$$\begin{aligned}
&Initialize\ m_i, i = 1, \dots, k, for\ example, to\ k\ random\ x^t \\
&\quad Repeat \\
&\quad For\ all\ x^t \in X \\
&\quad b_i^t \leftarrow \begin{cases} 1, & if\ \| x^t - m_i \| = min_j \| x^t - m_i \| \\ 0, & otherwise \end{cases} \\
&\quad For\ all\ m_i, i = 1, \dots, k \\
&\quad m_i \leftarrow \frac{\sum_t b_i^t x^t}{\sum_t b_i^t} \\
&\quad Until\ m_i\ converge
\end{aligned} \tag{3}$$

Here, *X* is the data to be have cluster analysis performed. Then, m_i represents the mean of points in the dataset, the centroids for the clusters. Essentially, there are *k* clusters, and the goal is to cluster or 'categorize' points to a specific cluster whose mean has the nearest mean, or the closest distance to the centroids. The centroids are updated iteratively until convergence is met.

An essential component to the k-means algorithm is the Euclidean distance formula, or simply the distance formula in the k-means algorithm. This equation is also used to calculate the minimal *intercluster* distance (distance between points in different clusters), and the maximal *intracluster* distance (distance of distinct points within a cluster). The Euclidean distance formula is very straightforward, it simply calculates the distance between two 2-dimensional points. The 2-dimensional Euclidean distance formula can be described by the following equation:

$$distance(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \tag{4}$$

In equation 4, $p = (p_1, p_2)$ and $q = (q_1, q_2)$.

The Dunn Index is a formula which evaluates the performance of a clustering algorithm. The Dunn Index in part of project 2 can be represented with the following equation:

$$Dunn\ Index = \frac{minimal\ intercluster\ distance}{maximal\ intracluster\ distance} \tag{5}$$

A higher value Dunn Index represents a better clustering.

Finally, we wish to compare how standardizing the data compares to non-standardized data. To standardize the dataset, we perform a calculation which scales the variables down significantly. The goal with standardization is to scale the data down to a mean which equals 0 and a standard deviation equal to 1. To achieve this, we use the z-normalization formula:

$$z = \frac{X - \mu}{\sigma} \tag{6}$$

X represents the value we wish to standardize, and Z is the result of the standardization. We subtract from X the mean of X or μ, and lastly, we divide it by the standard deviation of X or σ.

Analysis

The first set of graphs are with no standardization taken place after the initial data is gathered. To begin, the Scree Plot is the first graph, which can be used to determine the number of attributes to keep in the PCA.

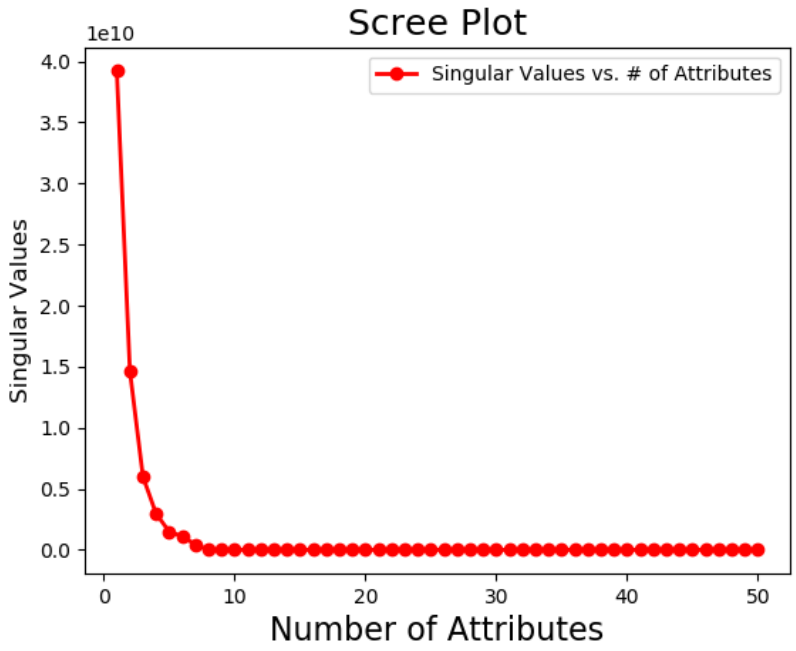

Figure 1. The Scree Plot of the singular values from Part 1 of Project 2. Non-Standardized.

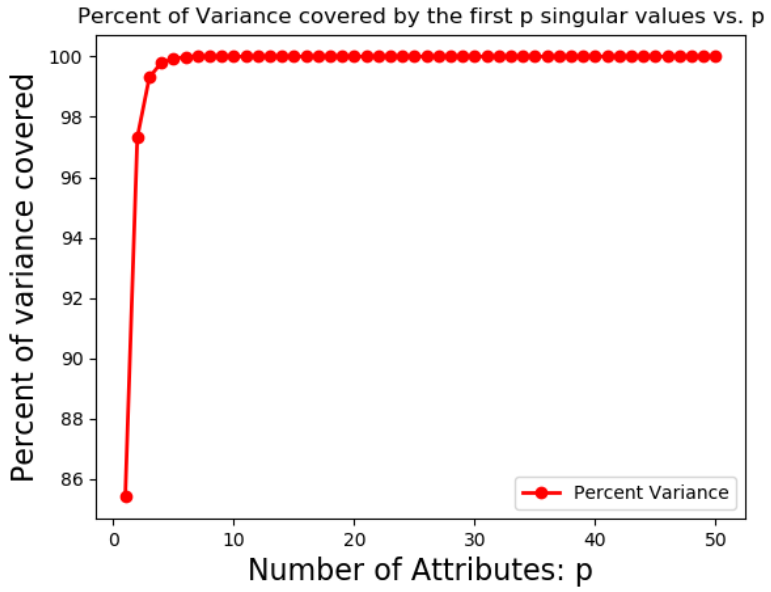

Figure 2. The percentage of variance covered by the first p singular values vs. p plot from Part 1 of Project 2. Non-Standardized.

The number of p's that should be chosen is within the "elbow" of the curve in figure 2. This shows that the p should be about 3 to 5, in this case the chosen p is 4.

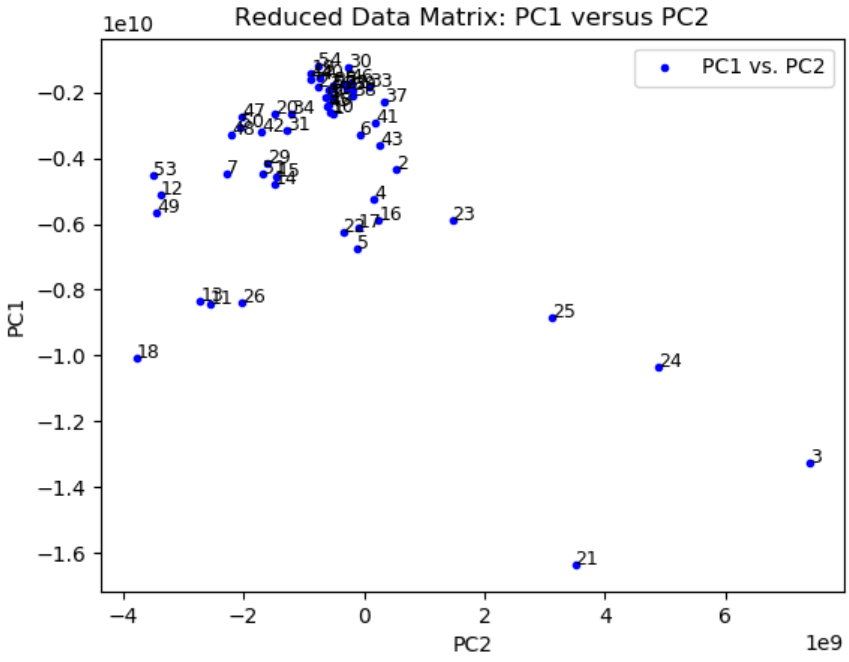

Figure 3. The reduced data matrix of the first p PCs, of the PC1 versus PC2 plot.

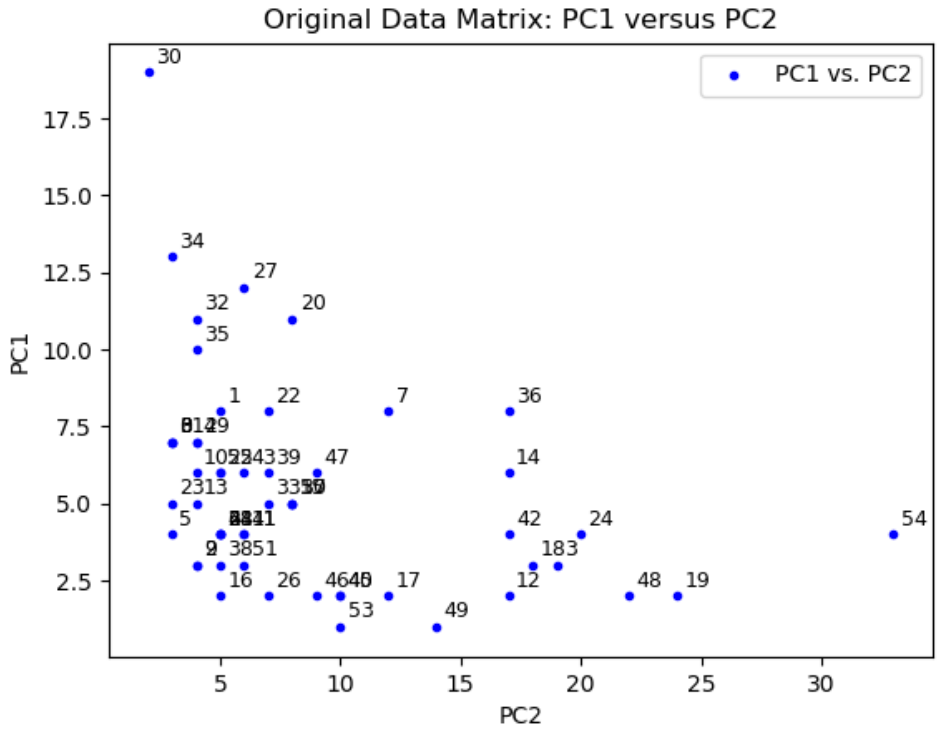

Figure 4. Original data PC1 versus PC2.

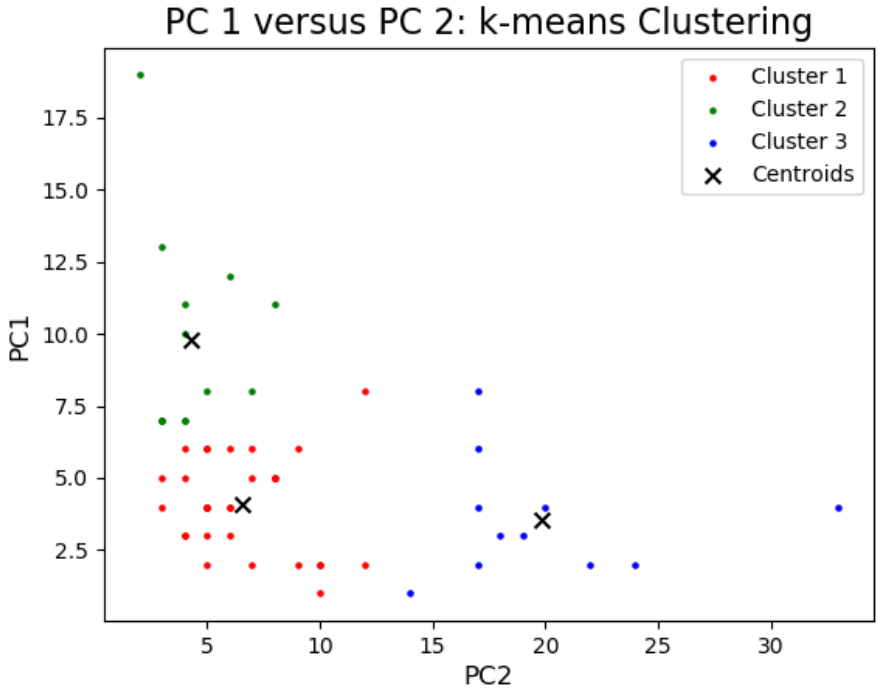

Figure 5. K-Means Clustering with three clusters, original data.

```
iterations to converge =  9

min_intercluster_distance =  1.0

max_intracluster_distance =  19.235384061671343

dunn_index =  0.05198752449100364
```

Figure 6. Calculated values with respect to the k-means clustering plot of Figure 5 above.

In figure 5, the University of Tennessee Knoxville is within cluster 1. Some of the other universities which are also in cluster 1 include: Univ. of Minnesota, Indiana Univ., Iowa, and Univ. of Virginia. Further, the Dunn Index as shown in figure 6 is very low, meaning there was poor clustering.

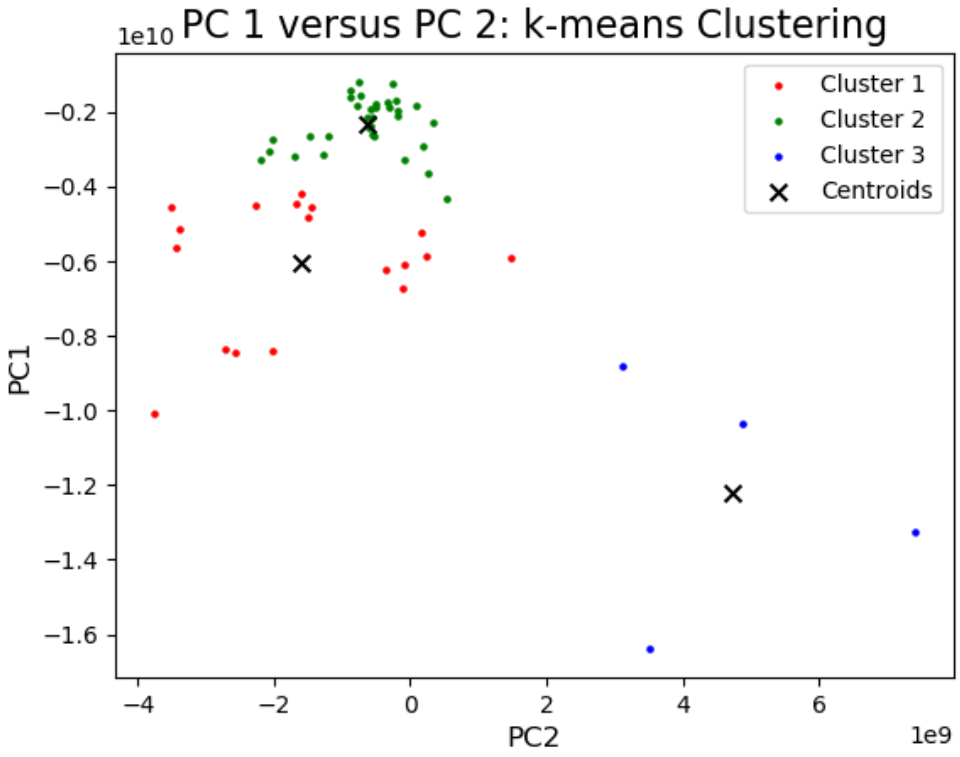

Figure 7. K-means Clustering with three clusters, 2-dimensional data.

```
iterations to converge =  9

min_intercluster_distance =  971844468.6204103

max_intracluster_distance =  7557908676.416193

dunn_index =  0.1285864265141186
```

Figure 8. Calculated values with respect to the k-means clustering plot of figure 7 above.

The clustering in figure 7 is far better than in figure 5 which used only the original data rather than the reduced data matrix in figure 7. The University of Tennessee is in cluster 2 for figure 7, and other universities within cluster 2 include: Mississippi State Univ., Louisiana State Univ., Univ. of Oklahoma, and Indiana University.

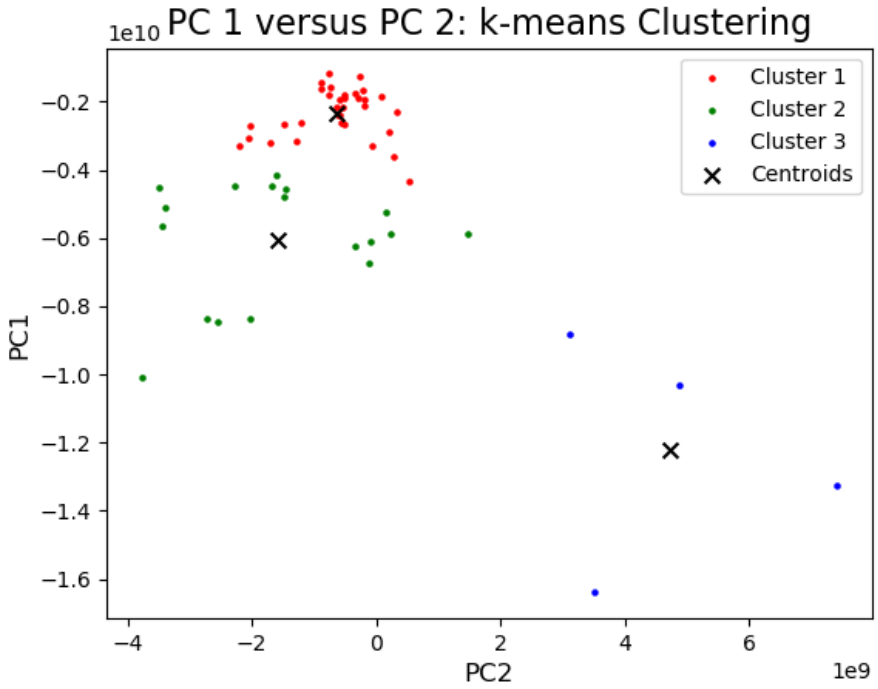

Figure 9. K-means clustering with three clusters, p-dimensional data (p=4)

```
iterations to converge =  5

min_intercluster_distance =  971844468.6204103

max_intracluster_distance =  7557908676.416193

dunn_index =  0.1285864265141186
```

Figure 10. Calculated values with respect to the k-means clustering plot of figure 9 above.

Figure 9, the p-dimensional graph changes very little compared its 2-dimensional counterpart, resulting in similar calculations as shown in figure 8 and figure 10. The University of Tennessee is in cluster 1 for figure 9, and other universities within cluster 1 include: Mississippi State Univ., Louisiana State Univ., Univ. of Oklahoma, and Indiana University.

The next set of graphs are with standardization taken place after the initial data is gathered, a z-normalization on the original data. Overall, the standardization significantly changes the Dunn Index, decreasing its value, showing mathematically that there was a worse clustering than previously with the non-standardized dataset.

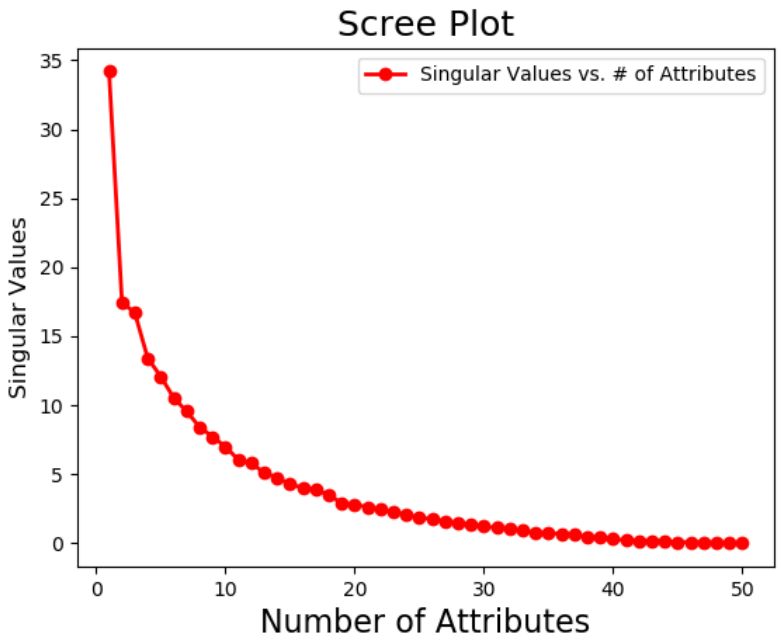

Figure 11. Scree Plot of singular values, with standardized data.

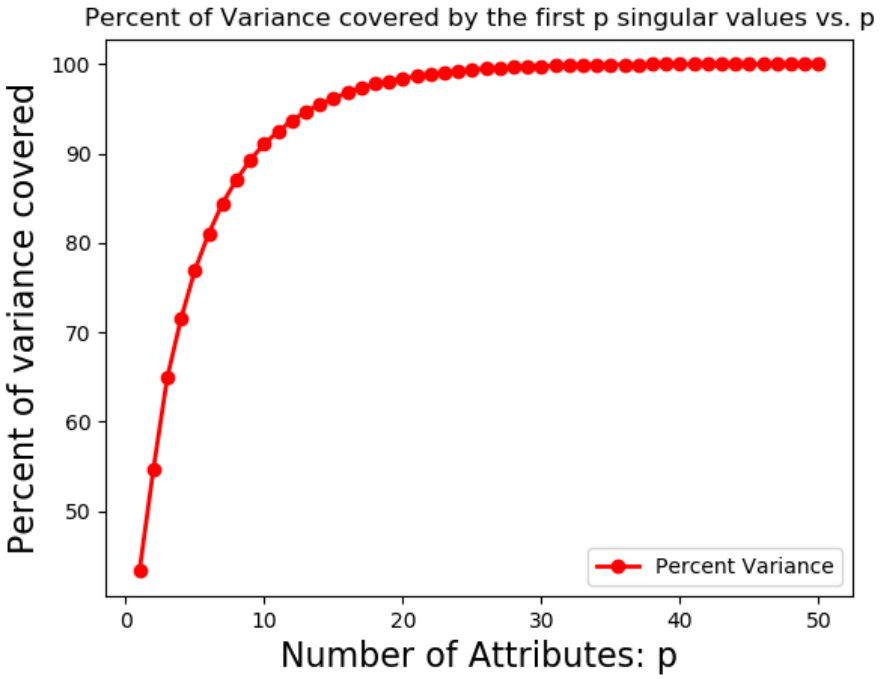

Figure 12. Percent of variance covered by the first p singular values vs. p, with standardized data.

It takes more attributes for the percentage of variance that covers the singular values to reach 100% because the data here is z-normalized, thus more attributes are required. The elbow point here is more difficult to determine as it is more curved in comparison to the non-standardized version in figure 2, but ultimately decided on a p of 30.

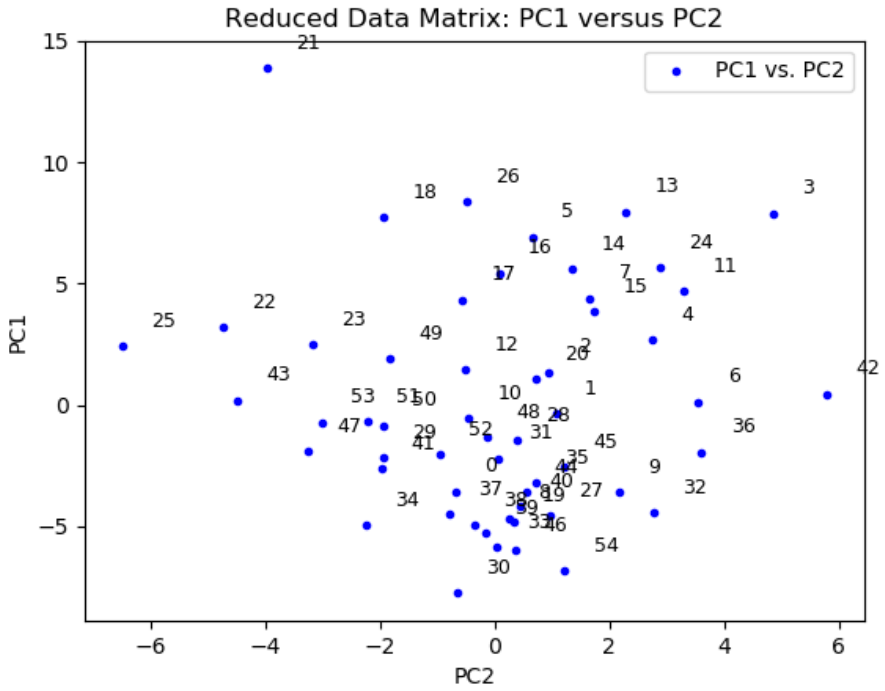

Figure 13. Reduced Data Matrix PC1 vs. PC2, with standardized original data.

There is a significant difference between the standardized reduced matrix and the non-standardized reduced matrix as seen in figure 3. The first two principal components are different in this case due to the z-normalized original data.

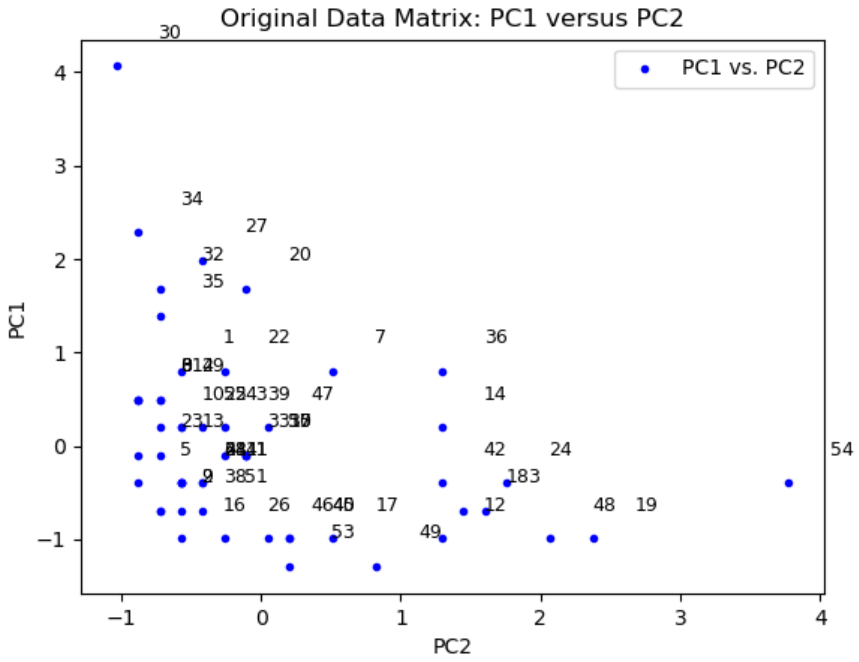

Figure 14. Original Data Matrix: PC1 versus PC2, with standardized original data.

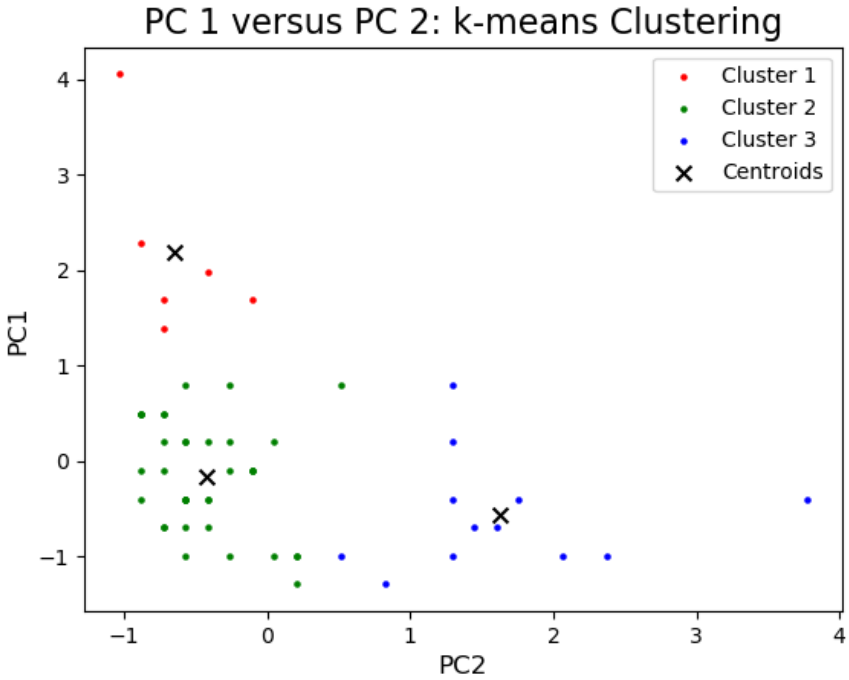

Figure 15. K-means clustering with three clusters, standardized original data.

```
iterations to converge =  5

min_intercluster_distance =  0.31038341402169545

max_intracluster_distance =  3.312896696064084

dunn_index =  0.09368943329577684
```

Figure 16. Calculations with respect to k-means clustering in figure 15 above.

Surprisingly, figure 15 and figure 5 are very similar and show nearly identical graphs. Cluster 2 or the corner cluster in figure 15, still contains the University of Tennessee as it does in figure 5 except in its "Cluster 1." The Dunn Index is also low as it is in figure 5's Dunn Index, but it is slightly larger.

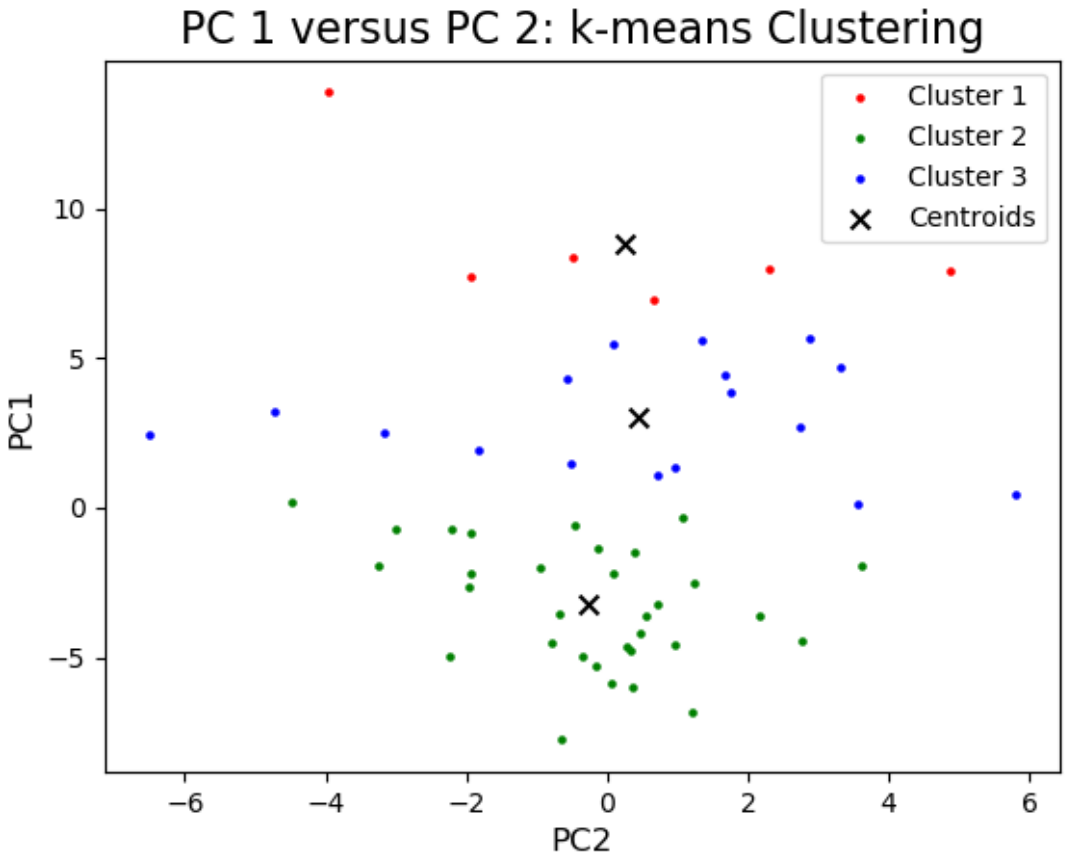

Figure 17. K-means clustering with 2-dimensional reducted matrix, with standardized original data.

```
iterations to converge =  3

min_intercluster_distance =  1.4588922788066645

max_intracluster_distance =  12.446600251779737

dunn_index =  0.11721211007785502
```

Figure 18. Calculations with respect to the k-means clustering plot in figure 17 above.

Figure 17 shows a very interesting cluster display where each of the clusters are in "layers." The University of Tennessee Knoxville is within cluster 2 in figure 17. Some of the other universities include: Univ. of Arkansas, Univ. of Kentucky, Mississippi State Univ., and Colorado State.

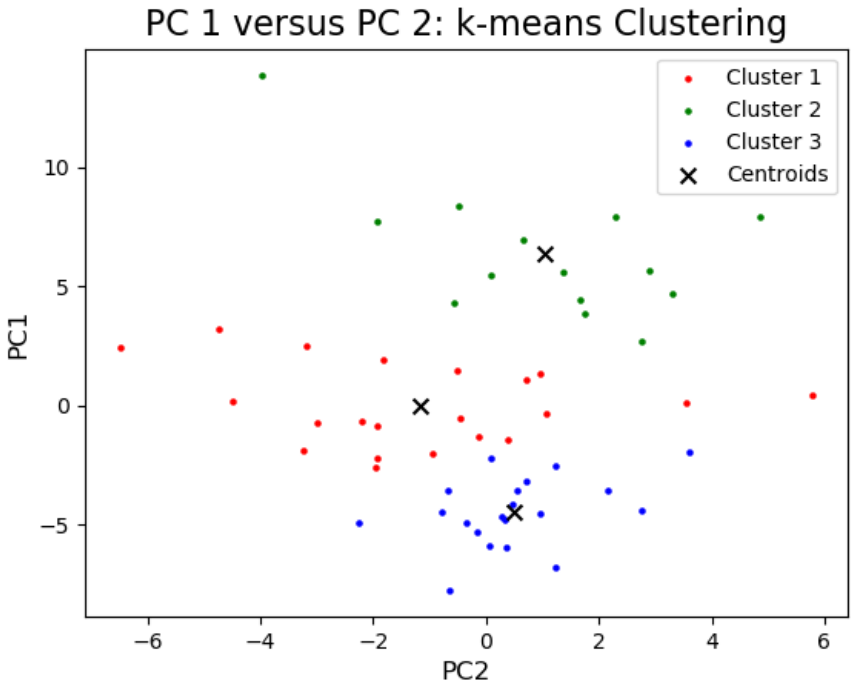

Figure 19. K-means clustering for p-dimensional reduced matrix (p=30), with standardized original data.

```
iterations to converge =  8

min_intercluster_distance =  0.8144491503719147

max_intracluster_distance =  13.050165259249024

dunn_index =  0.062409106259761066
```

Figure 20. Calculations with respect to the k-means clustering plot in figure 19 above.

With the p-dimensional reduced matrix, the Dunn Index decreased significantly compared to the 2-dimensional reduced matrix in figure 17. This is due to the minimum intercluster distance being less, leading to a much smaller Dunn Index. The University of Tennessee is surprisingly still within the lower layer cluster, or cluster 3 in figure 19. Other universities within cluster 3 include: Univ. of Kentucky, Utah, and Univ. of Missouri.

Discussion

Upon initial investigation of the data, there were several issues. High level wise, some of the data was ill-suited to this sort of analysis, such as 'x' marks or any other non-numeric data. Some of the numeric data consists of only 1s (and one 2), which is again not informative for this type of analysis. If a single row within a column was empty, the entire column was ignored in favor of using complete and hopefully correct data.

Principal component analysis is a statistical tool for 'compiling' a group of attributes that are potentially connected to each other in some form, and then determines which helps determine attributes which are uncorrelated, the principal components. Then, with SVD we were able to determine how to best reduce the original data matrix.

The k-means clustering algorithm part was the most exciting and challenging portion of this project. First, selecting k centroids to begin the k-means clustering algorithm and allowing it to converge was very interesting. The program cannot "see," and therefore it needs to take advantage of an algorithm like the Euclidean distance formula in order to sort out what each cluster consists of given k clusters. Ultimately, I believed three clusters was the best choice of k for this analysis. When testing with two clusters, it typically involved one cluster being in a small area but with a low variance such as cluster 1 in figure 9, and the other cluster taking the high variance points outside of the low variance cluster such as clusters 2 and 3 in figure 9. Though the standardized data would likely benefit more from a k value of 2 in the case of its reduced matrix plots like that of figure 17 and figure 19, where the points are dispersed more across the graph rather than congregated into certain areas like in figure 9.

Sources

Alpaydin, Ethem. *Introduction to Machine Learning*. third ed., MIT Press, 2014.

"Dunn Index." *Wikipedia*, Wikimedia Foundation, 30 July 2019, https://en.wikipedia.org/wiki/Dunn_index.

"Euclidean Distance." *Wikipedia*, Wikimedia Foundation, 5 Sept. 2019, https://en.wikipedia.org/wiki/Euclidean_distance.

"K-Means Clustering." *Wikipedia*, Wikimedia Foundation, 16 Oct. 2019, https://en.wikipedia.org/wiki/K-means_clustering.

"Principal Component Analysis." *Wikipedia*, Wikimedia Foundation, 9 Oct. 2019, https://en.wikipedia.org/wiki/Principal_component_analysis.

"Singular Value Decomposition." *Wikipedia*, Wikimedia Foundation, 9 Oct. 2019, https://en.wikipedia.org/wiki/Singular_value_decomposition.