**Week 8 DB1**

What are the similarities and differences between PCA and multidimensional scaling (MDS)? When is it appropriate to use each method?

Similarities:

PCA and MDS transform data into a new form that allows the analyst to gather new observations and insights. Both methods look at similarities and differences in the data. Matrixes are used in both setting to create new insights. In my opinion, both methods are used to work with difficult data to draw out meaningful insights that are not available at first glance.

Differences:

PCA reduces the number of uncorrelated variables into principal components. The principal components represent the variability in the data set. Each component in PCA is a: new summary variable, linearly related to the original variables, does not suffer from collinearity, maximizes the variation from the prior data set.

MDS analyzes dissimilarity. The dissimilarity is constructed in a matrix that shows the differences between points. I relate this method to a compressed file in that it reduced the space between data points and it is usually shown on a two dimensional plot. Another difference between the two methods is PCA uses eigenvalues and eigenvectors as a metric while MDS utilizes Euclidean distance between data points.

I would use PCA in the presence of collinear data, and use MDS to analyze data sets with large disparities or outliers.

**DB2**

What are the basic ideas behind the principal component analysis (PCA)? How would the concepts of Eigen value and Eigen vector apply to PCA? What are the Eigen vectors and values given a matrix M when implementing a linear transformation?

PCA is a great technique for transforming a set of variables into a new set of variables that can be analyzed from a different perspective. There are useful metrics for identifying the presence of collinearity, but few techniques for working with data that suffers from collinearity. Principal component analysis (PCA) is a method for transforming collinear variables into uncorrelated variables. While PCA focuses on the components with the largest variance, the minor variances are of interest as well. The minor variances can be helpful in finding outliers that are obscuring the overall data. According to RABE “The raw data [is] processed through a principal components subroutine that operates on the correlation matrix of the predictor variables in order to compute the eigenvalues, the eigenvectors, and the principal components.” Based on the resulting eigenvalues we can determine how many components to keep (RABE, 263, Drake). I understand the eigenvalues as well as the eigenvectors to be the metrics utilized to select the principal components in the overall analysis.