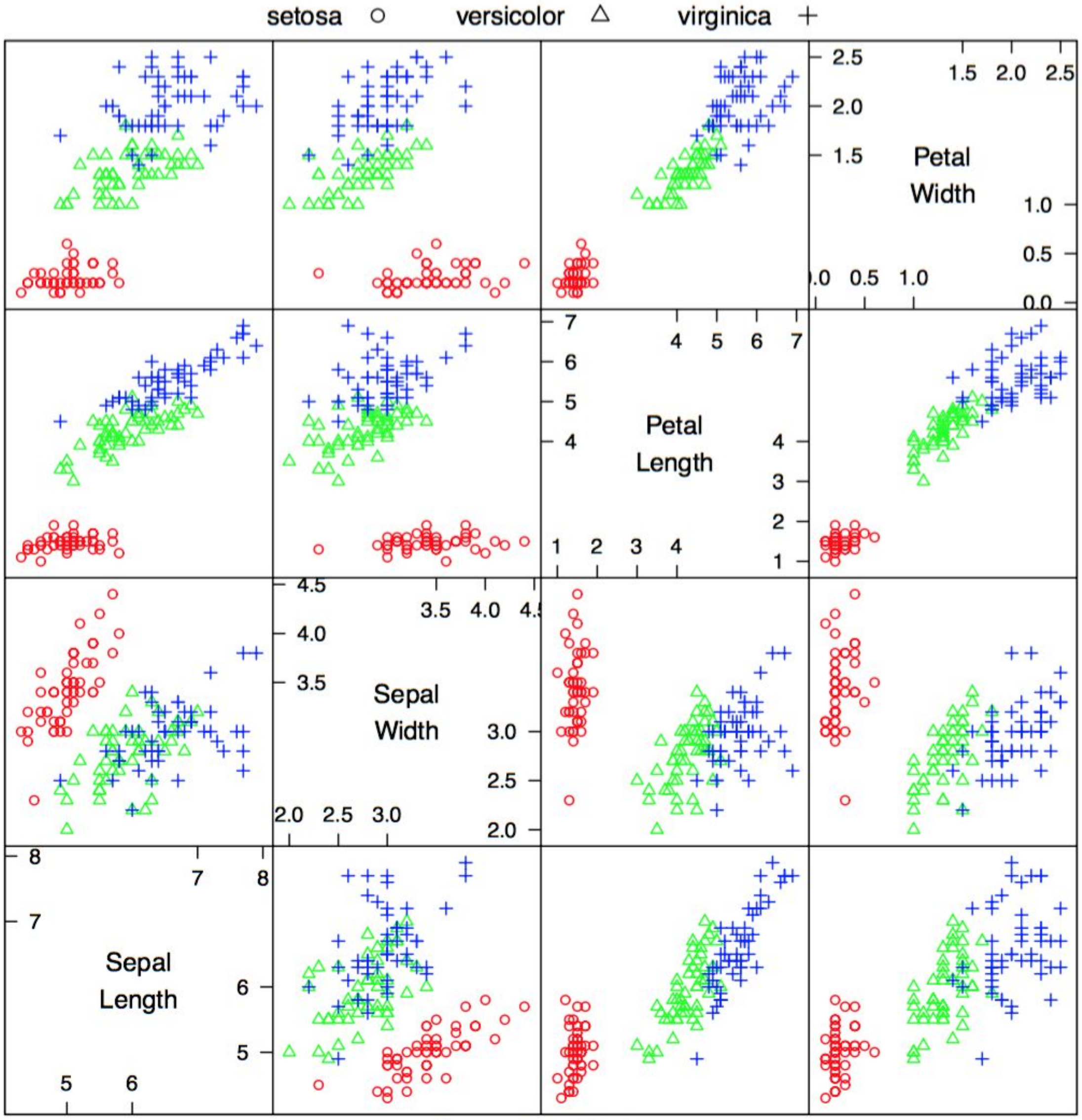
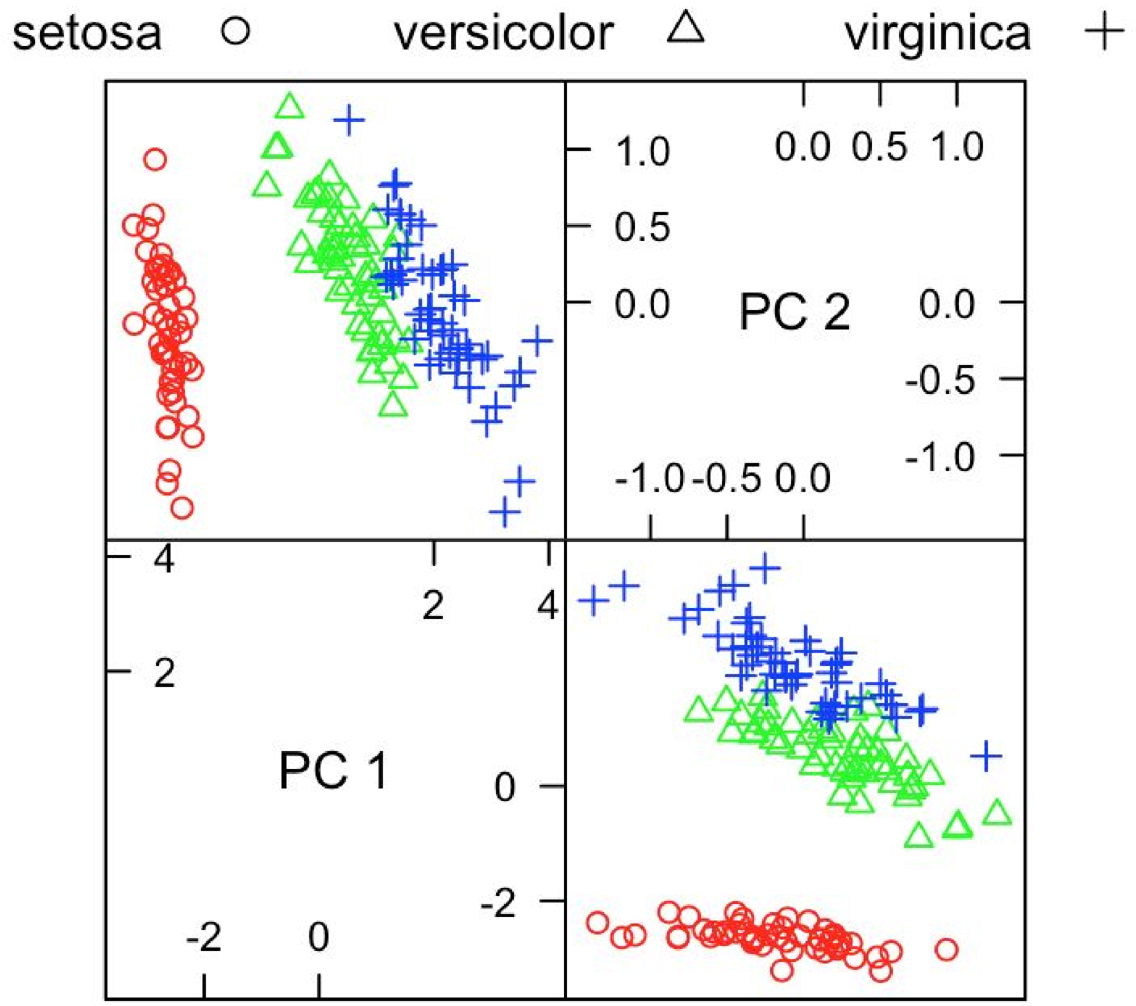
Ryley Higa
higa2

Homework 4 Report
Iris Data



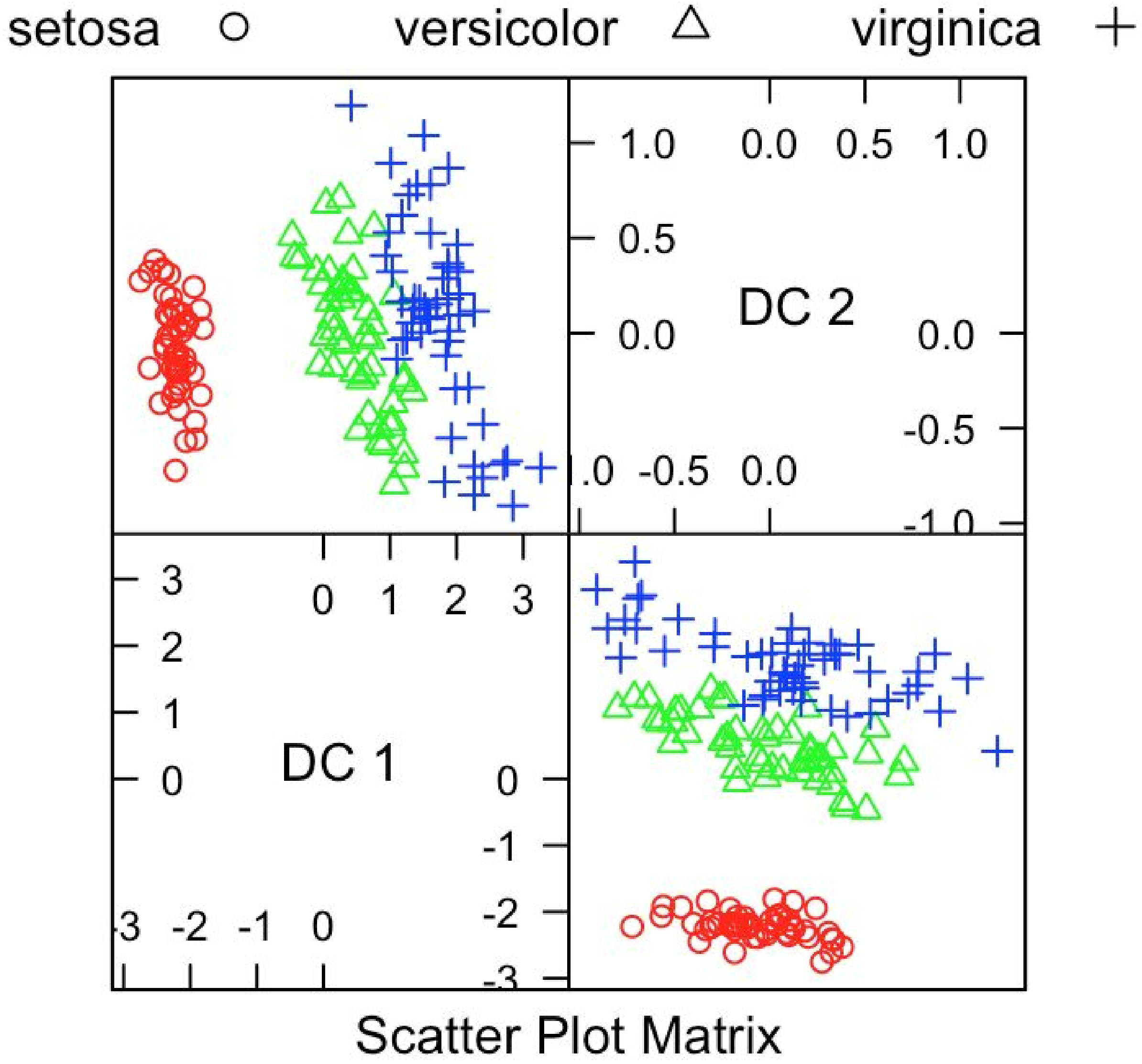
Scatter Plot Matrix

The scatterplot matrix above shows the four features plotted against one another, two features at a time. The goal is to visualize patterns and distinct separations between species of iris using features such as sepal length, sepal width, petal length, and petal width.



Scatter Plot Matrix

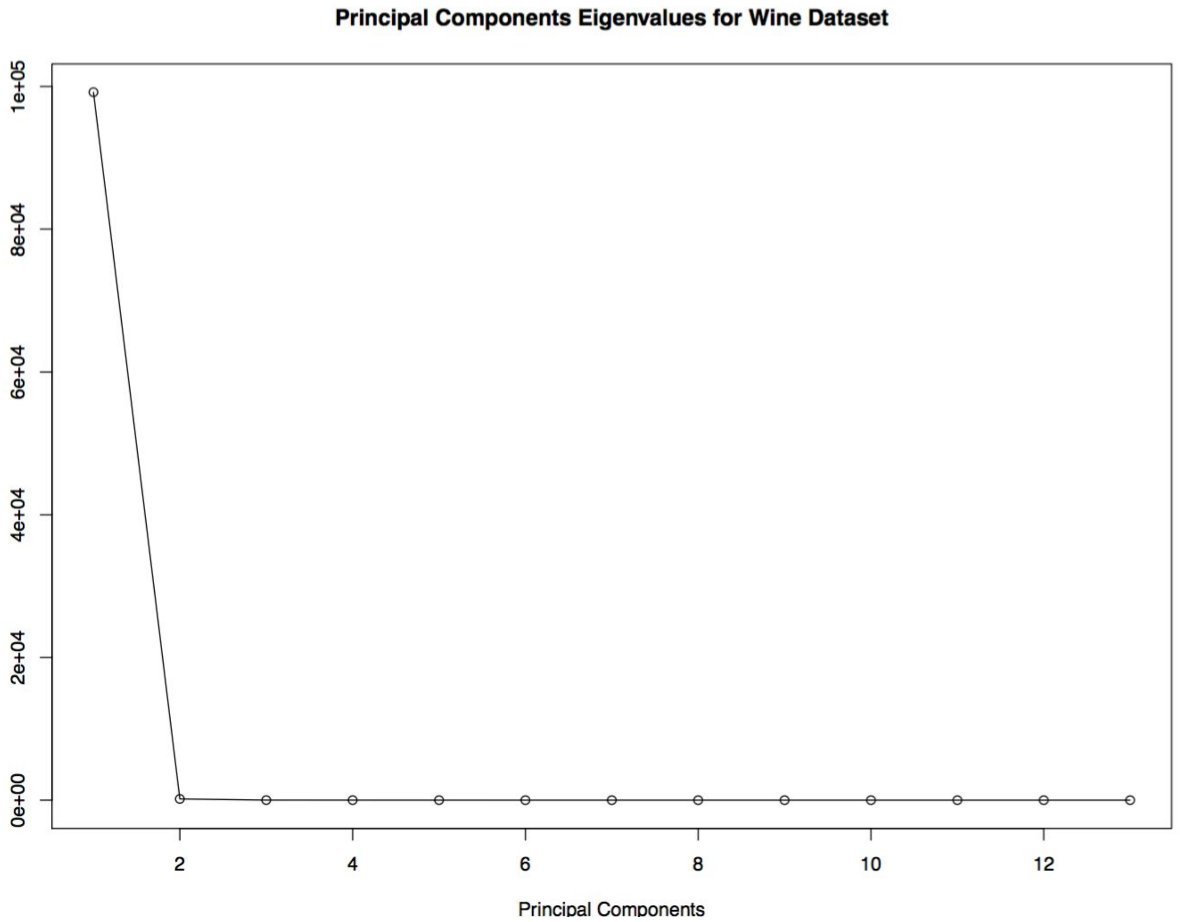
The scatter plot matrix above shows the projected data onto the first two principal components. The principal components do a decent job of discerning labels as there appears to be three clusters containing the majority of points pertaining to one label. It did not do a great job, however, since without labeled data cannot be separated cleanly by a smooth classification boundary.



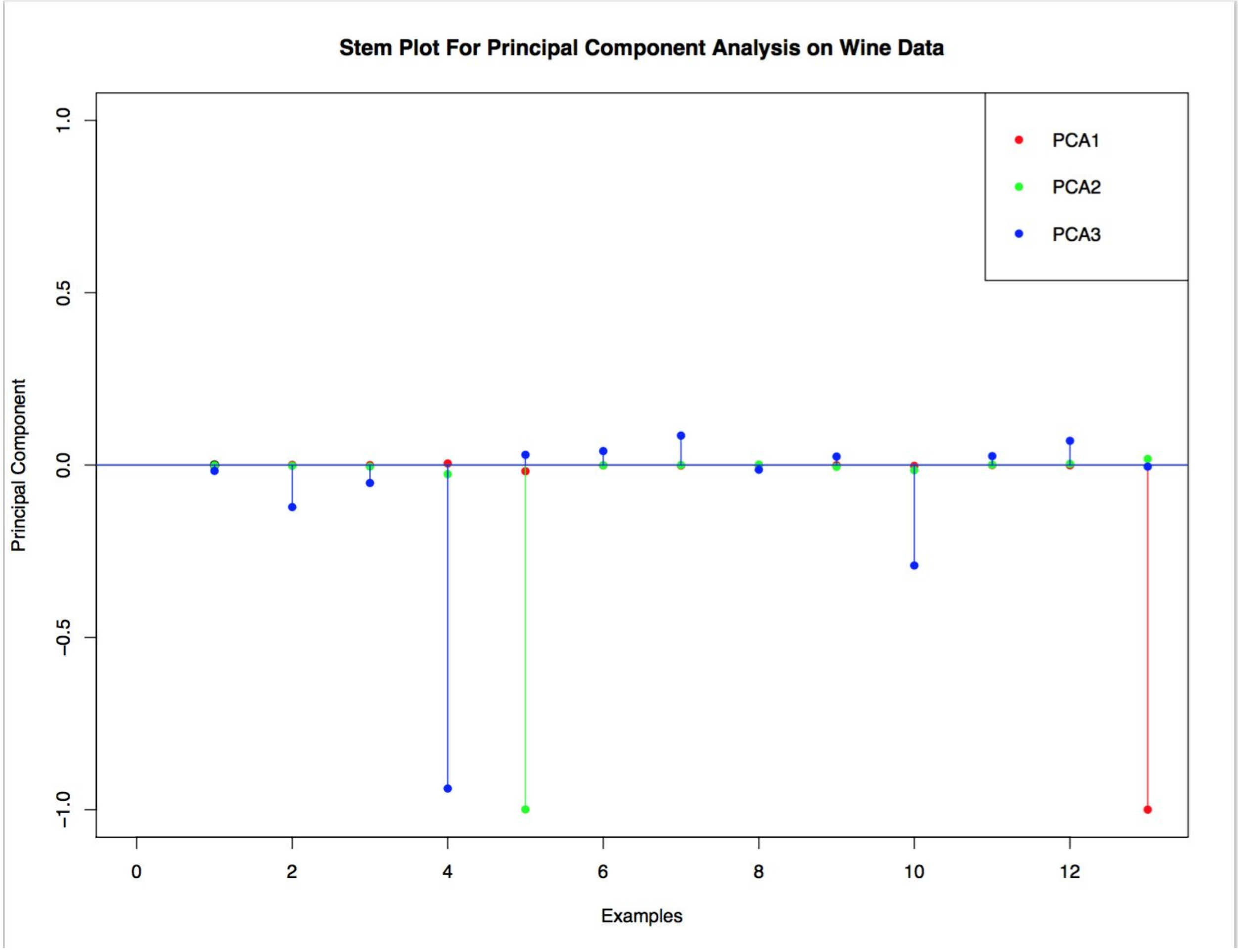
The first two discriminative directions shown in the scatterplot matrix plot show a clearer separation than the projected principal component data, but it is still quite similar to the principal component projection in terms of the position, shape, and scale of the three clusters.

Wine Data

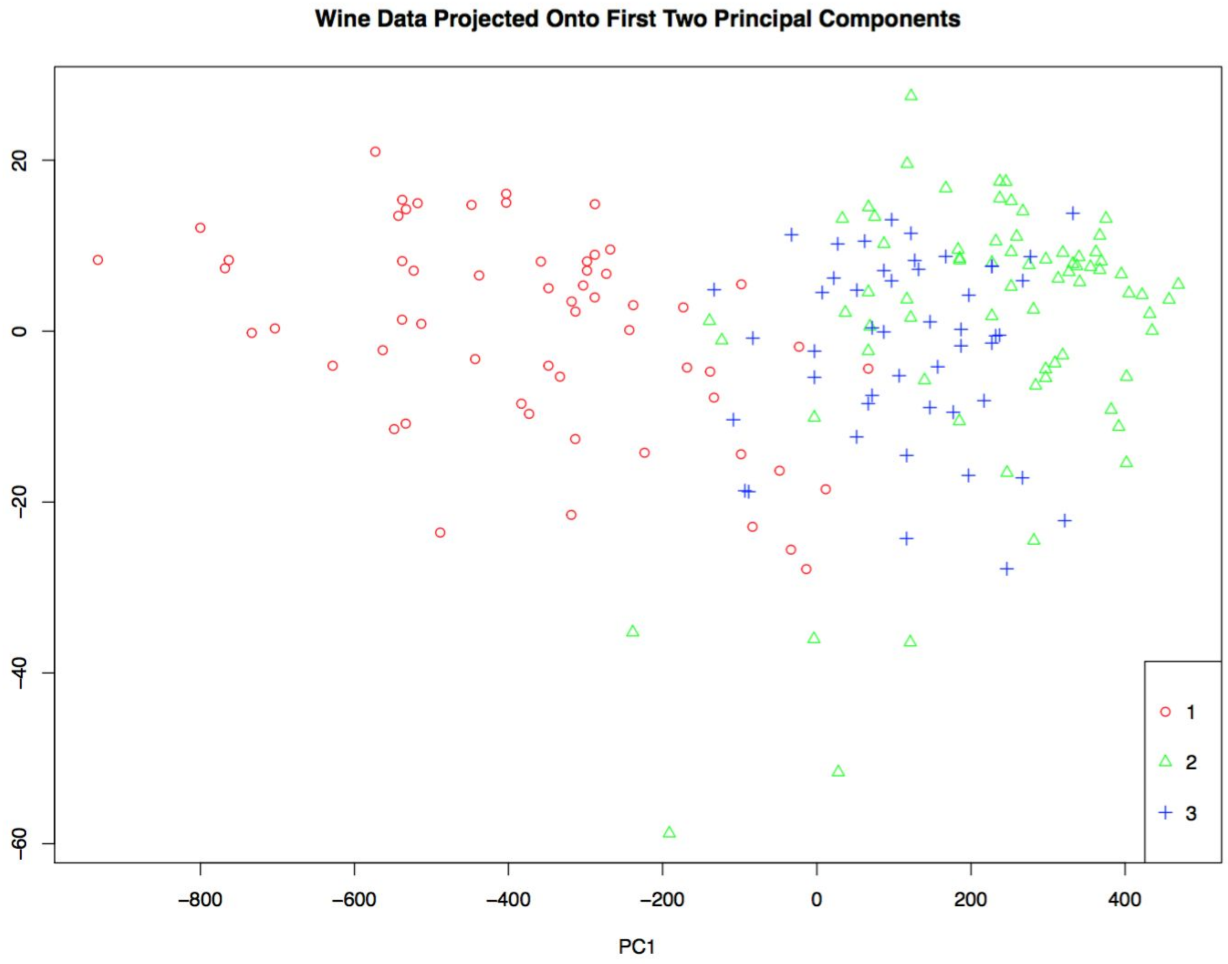
The wine data uses 13 features to classify the cultivar from which the wine was made from.



The figure above shows the amount of variance--given by the eigenvalues of the covariance matrix of the dataset--captured by each of the principal components. As expected, the explained variance seems to drop off exponentially as the principal component becomes less important. A reasonable amount of principal components that could represent the data is three since most of the variance is explained by the first three components.

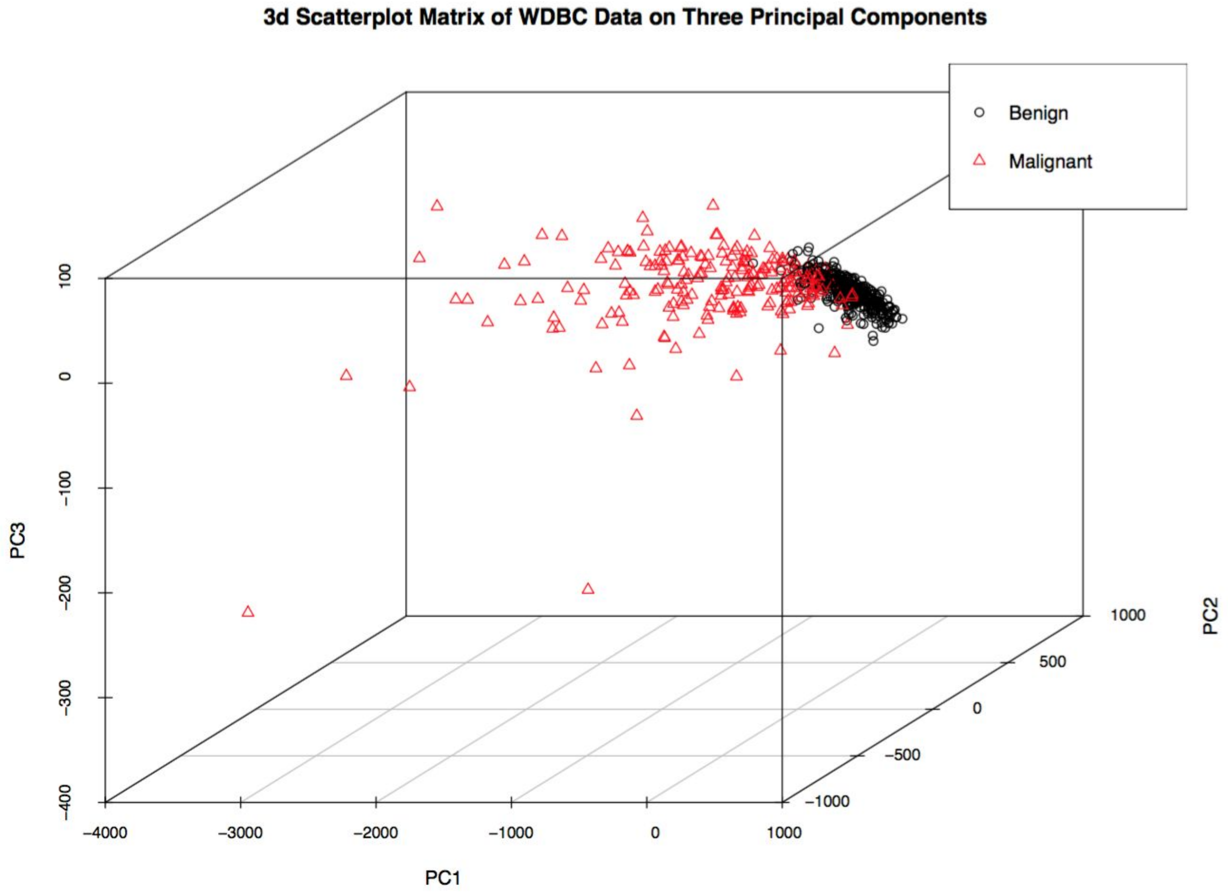


The figure above shows a stem plot of the first three principal components. The features are labeled as (1) Alcohol, (2) Malic acid, (3) Ash, (4) Alcalinity of ash , (5) Magnesium, (6) Total phenols, (7) Flavanoids, (8) Nonflavanoid phenols, (9) Proanthocyanins, (10)Color intensity, (11)Hue, (12)OD280/OD315 of diluted wines, (13)Proline. The first three principal components rely heavily on certain feature shown by the strong negative weights on certain features.

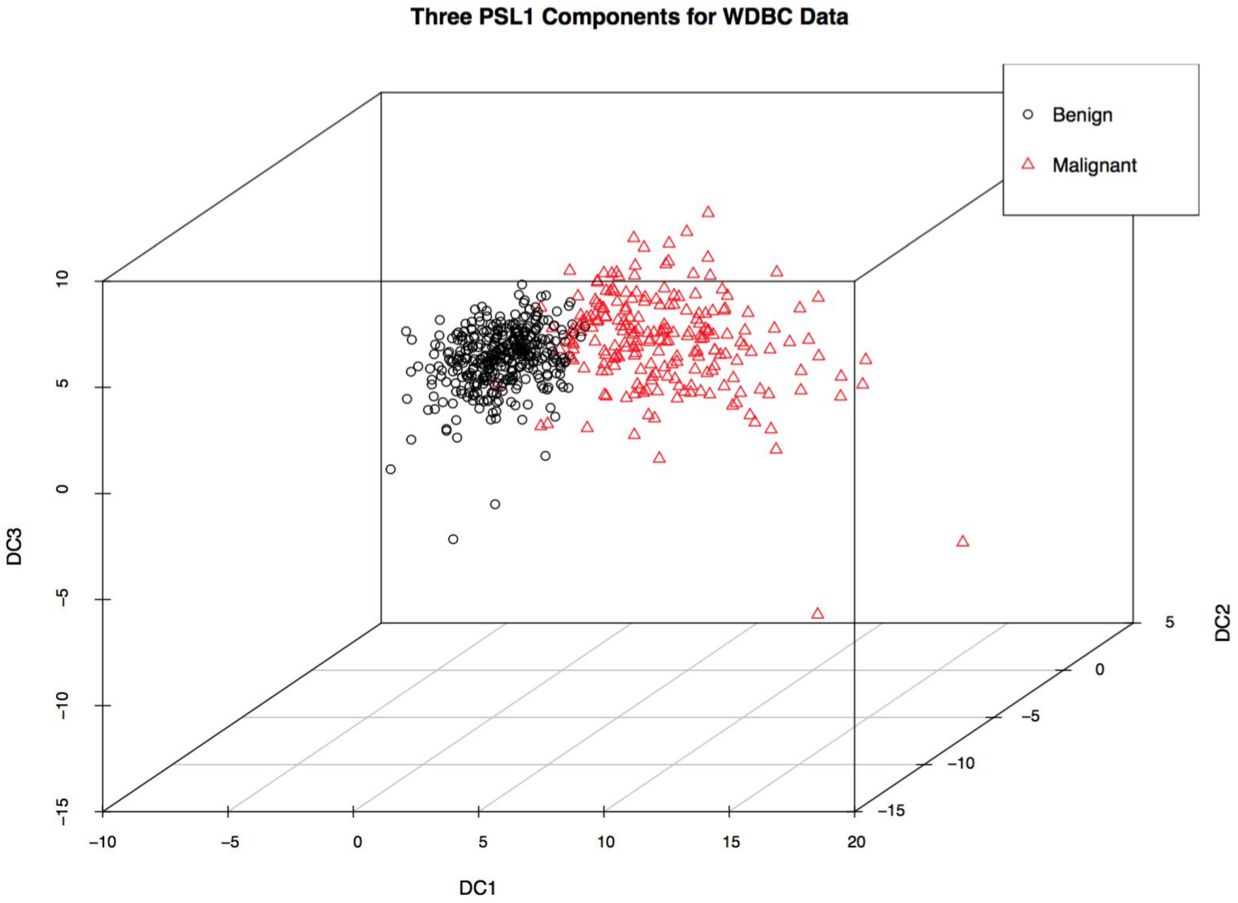


The figure above shows the projected data onto the first two principal components. Although the first two principal components captured the majority of the variance within the data, there was still a lack of discrimination between the three cultivars of wine. Without labels it appears as if the data may have two cluster centers a center that captures the majority of cultivar one and a center that captures the majority of cultivar two and three.

Wisconsin Diagnostics Breast Cancer (WDBC) Data
The WDBC data uses a list of 30 features about cancerous cells to determine whether the breast cancer is malignant or benign.



The three principal components of the WDBC data show some separation in the data amongst malignant-benign labels, but the separation is not clear without data labels. From first visual inspection, there is not a clear linear classifier that separates the data.



The discriminative directions show a clearer separation than the principal component projection of the data. About two clusters can be determined from the data without visual labels, and the malignant and benign datasets show a clearer linear separation and less intermingling.