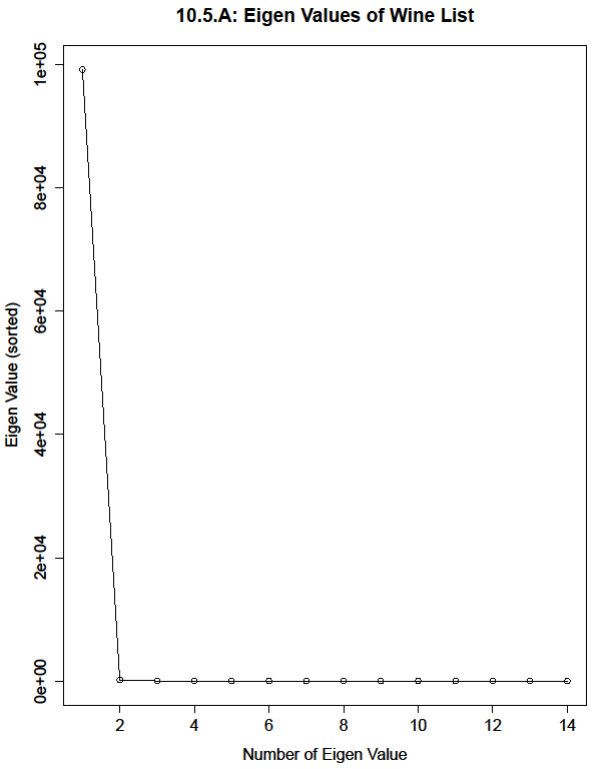
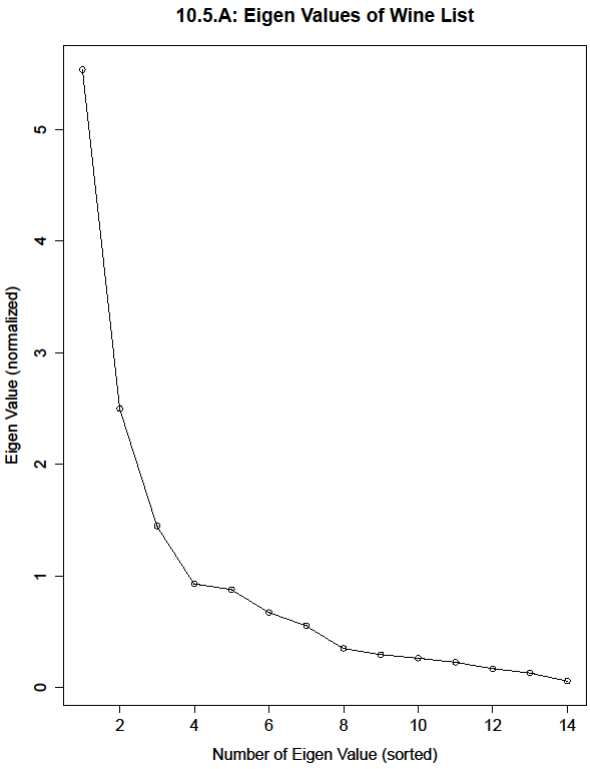
10.5) Wine Dataset

a) *Plot the eigenvalues of the covariance matrix in sorted order.*



[Not Normalized]



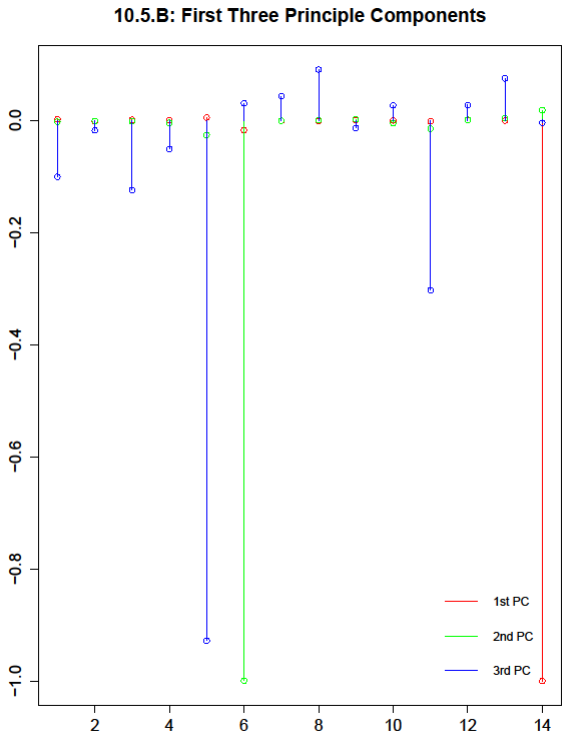
[Normalized]

a) *Continued. How many principal components should be used to represent this dataset? Why?*

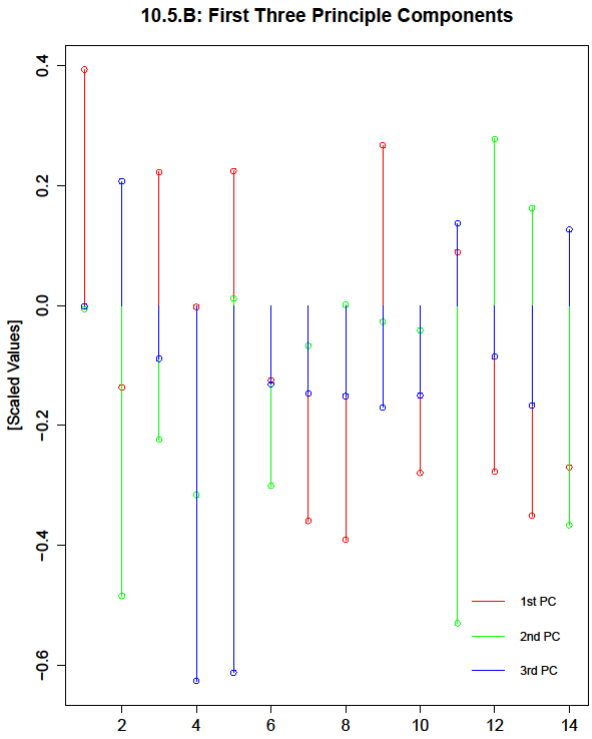
When referencing the non-normalized version of the eigenvalues, we can see that the first principal component is by far the most important it representing this dataset. But there could be other important components too, so we'll take a look at the normalized graph. From this, we see that the first 3 principal components are really very essential to representing this dataset, and show and exponential increase in eigenvalue from the other components. We can also say that components 4-7 are also important, if we wanted a more accurate representation of the data, but the dataset can rely on the first 3 fairly well.

10.5) Wine Dataset

b) *Construct a stem plot of each of the first 3 principal components.*



[Not Normalized]



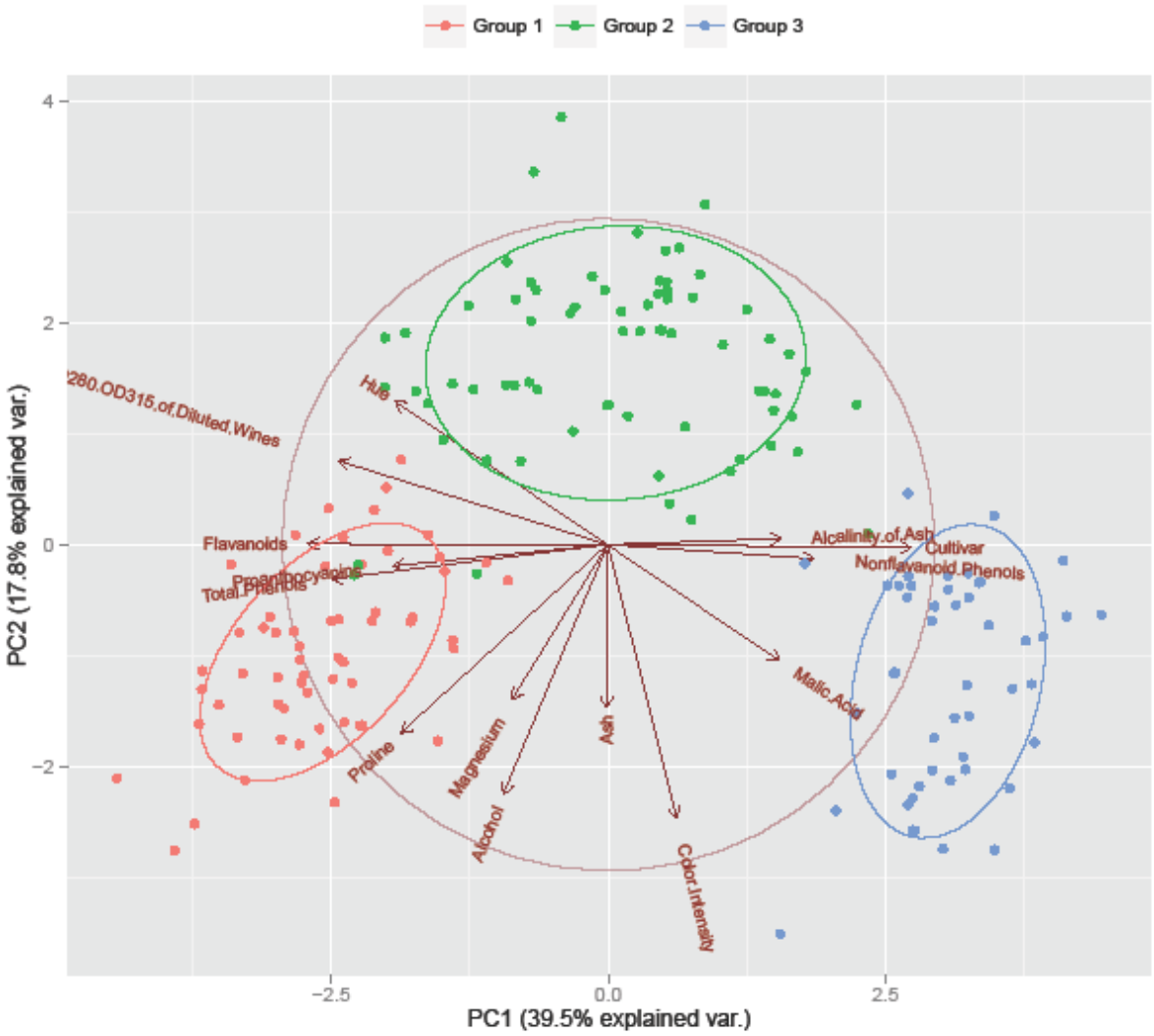
[Normalized]

b) *Continued. What do you see?*

By first looking at the non-normalized graph, we see a large difference between the values of components 5, 6, and 14 when comparing across the first 3 principal components, which means that some of these principle components may not be a good representation of some of the components alone. This same difference can be seen to a lesser extent in components 1, 3, 7, 8, and 11, with almost all of the differences coming from the third principal component, which shows that it may be important in the dataset to identifying outliers or the like. When normalizing the data, we see a largely different view of what's happening when comparing across the PCs, so it's very likely that we have a few outliers in the data according to a few principle components.

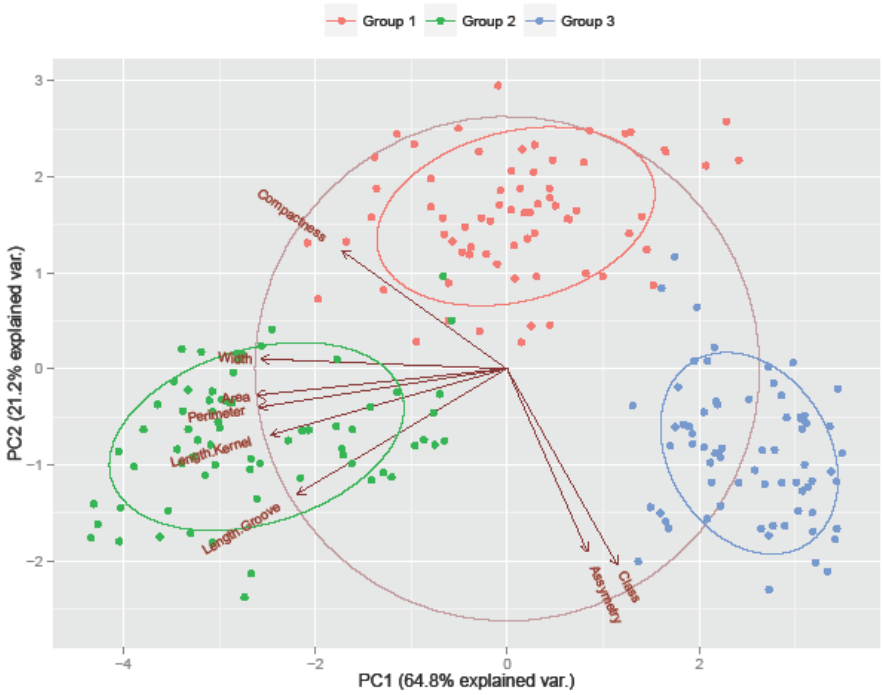
10.5) Wine Dataset

c) *Compute and project onto the first two principal components of the dataset.*



10.6) Seed Dataset

a) *Produce a scatterplot of this projection.*

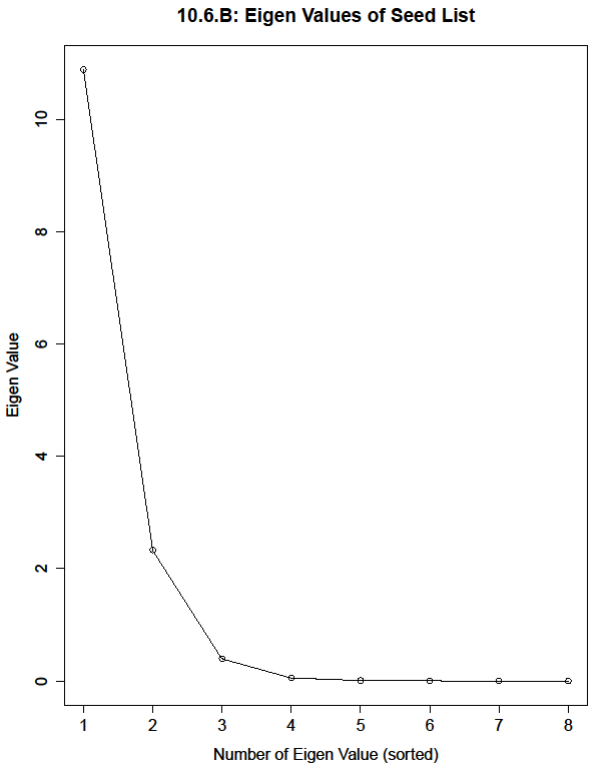


a) *Continued. Do you see any interesting phenomena?*

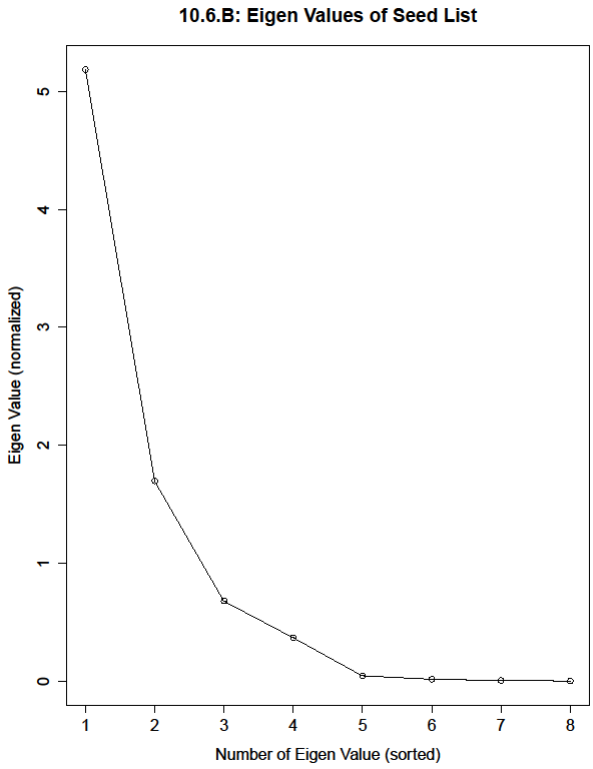
Yes, we see three distinct blobs of data separated by the seed's class when represented by the first two principle components, which means that these 3 classes may have significant differences upon relation to the other PCs. We also see a tight clustering around the second class of seeds when looking at a broad range of components, meaning that this class may be significantly different from the others.

10.6) Seed Dataset

b) *Plot the eigenvalues of the covariance matrix in sorted order.*



[Not Normalized]



[Normalized]

b) *Continued. How many principal components should be used to represent this dataset? Why?*

Since the non-normalized and normalized graphs largely tell the same story, I would say that using the first 3 principle components to represent the dataset would be a fairly good representation as the rest of the PCs don't carry too much weight in reference to the first 3. The first PC is largely the most important component, with the second also being fairly important and the third being very similar in importance to the rest of the PCs.