1.
   a. Max = 100
   b. Min = 37
   c. Q1 = 68
   d. Q3 = 87
   e. Median = 77
   f. Sample mean = 76.681
   g. Sample mode = 77
   h. Sample variance = 173.531

   These were all calculated using my program [**1.py**] (included with submission). Comments in the file provide implementation details.

2.
   a. Jacquard coefficient = 0.339
      i. I calculated this by taking (q / q + r + s) where q is the cell that is 1 for both supermarkets and r and s are the cells where one supermarket has the item, and one does not.

   b.
      i. Manhattan distance = 5700
      ii. Euclidian distance = 695.950
      iii. Supremum distance = 166

   These were calculated using my program [**1.py**] (included with submission). Comments in the file provide implementation details.

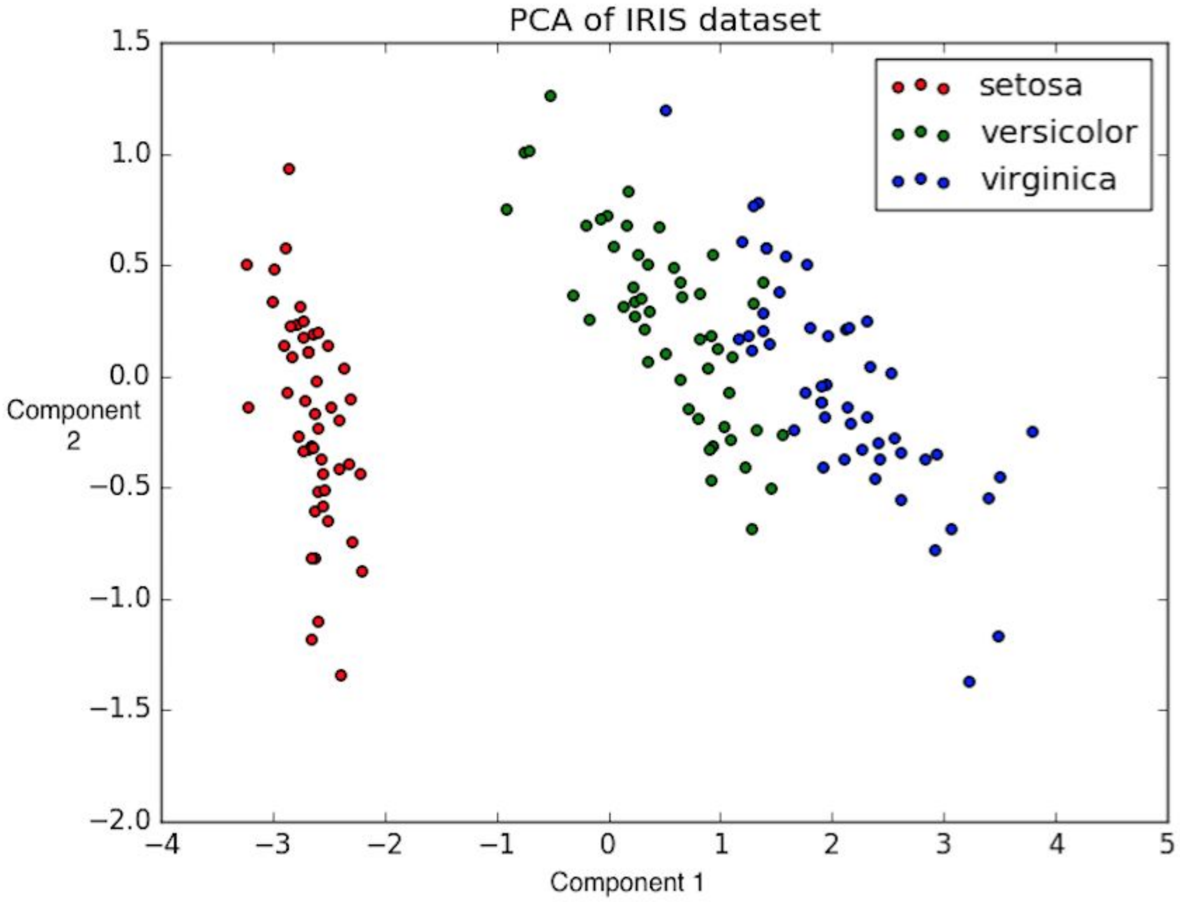
   c. Cosine similarity = .845

3.
   a.
      i. Sample mean (pre-normalization) = 76.681
      ii. Sample variance (pre-normalization) = 173.531
      iii. Sample mean (normalized) = 0
      iv. Sample variance (normalized) = 1

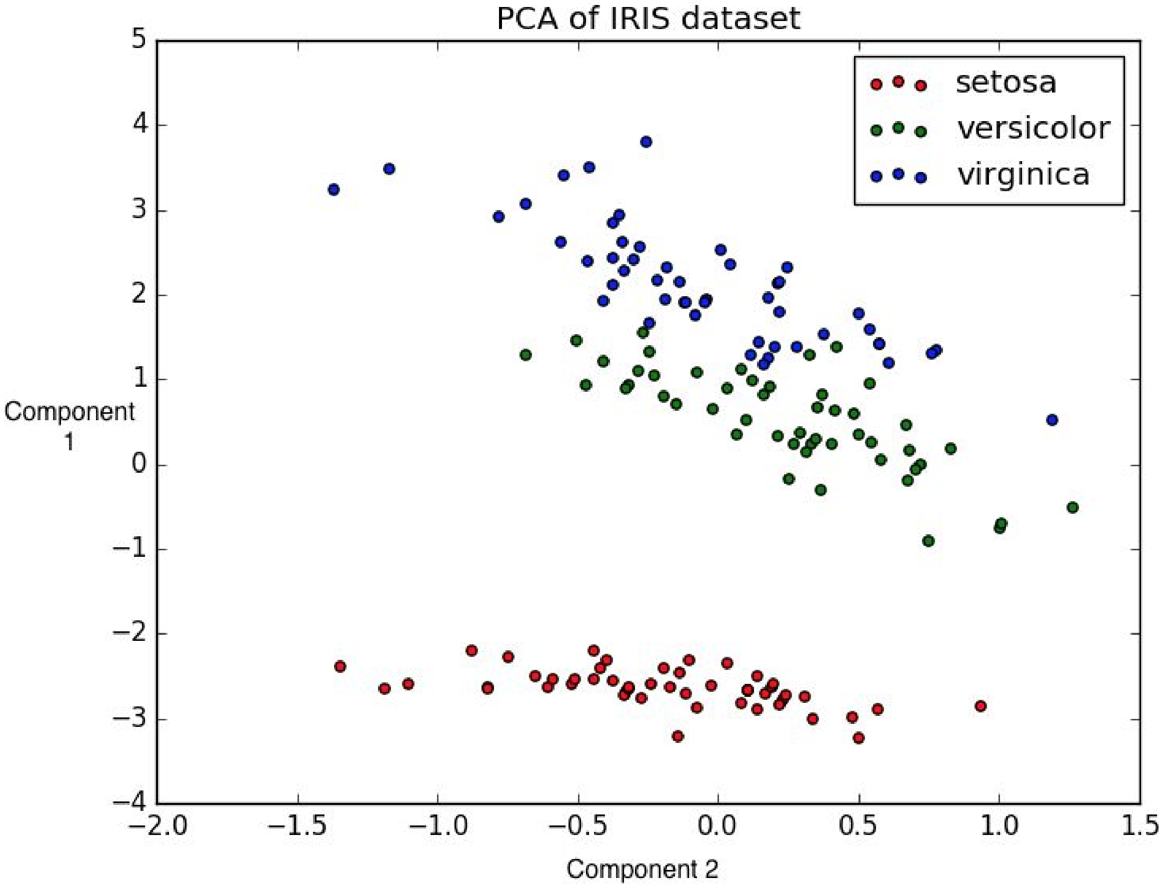
   b. Score of 90 after normalization = 1.011

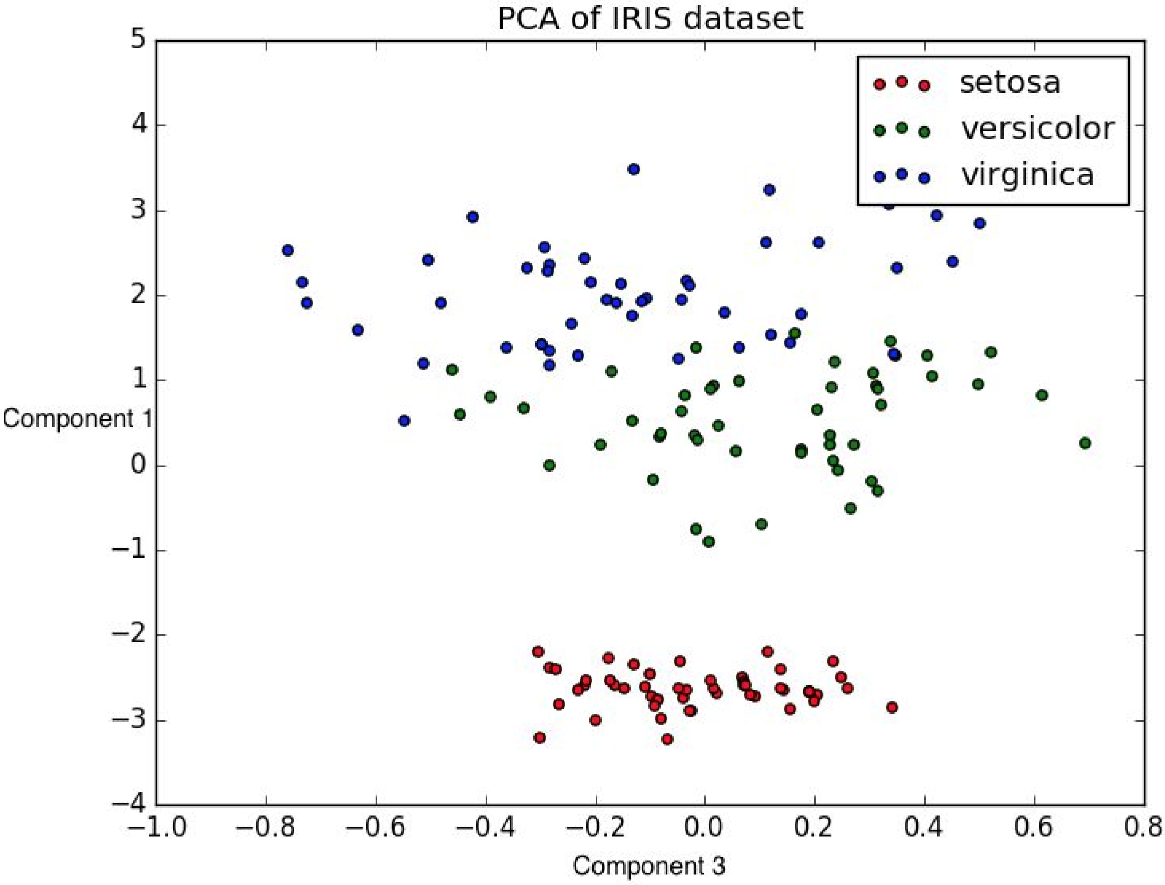
   These were all calculated using my program [**1.py**] (included with submission). Comments in the file provide implementation details.

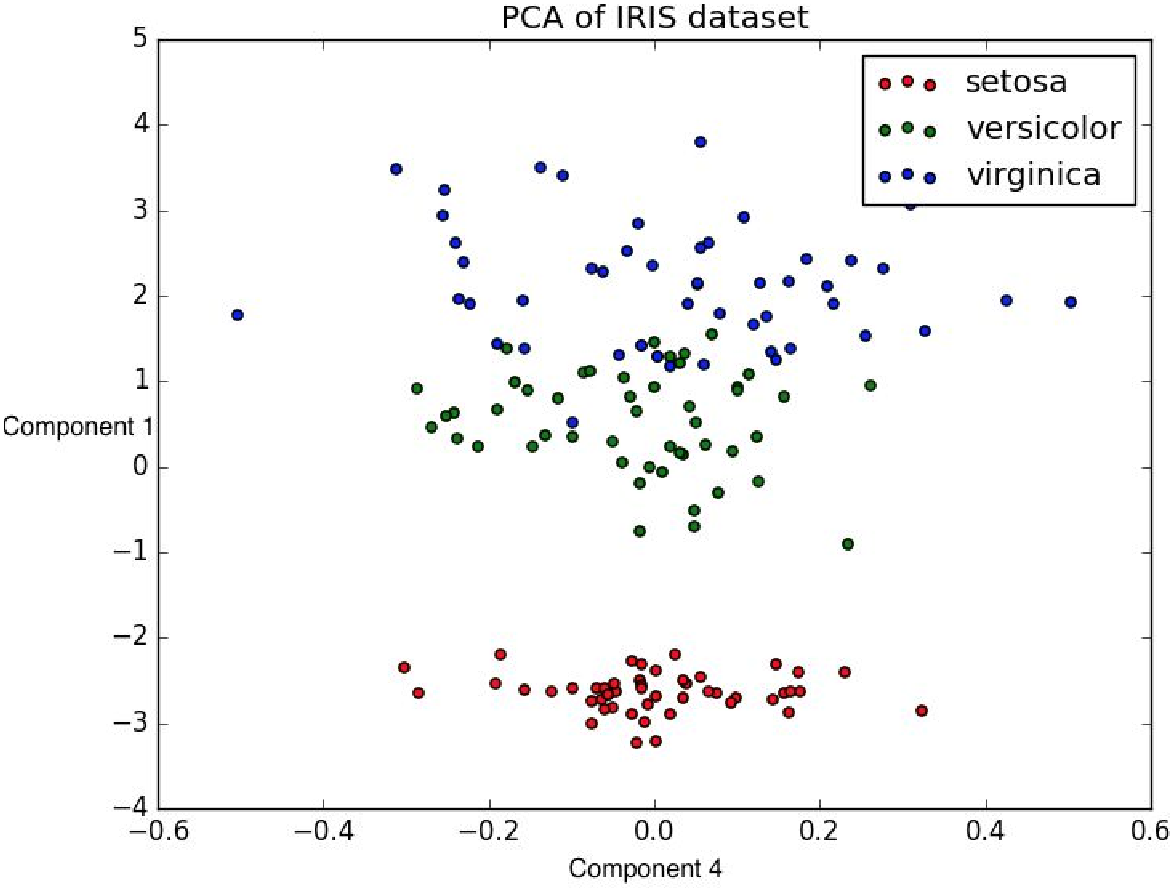
4.
   a. I used the sklearn PCA package in order to do the principal component analysis for problem 4. I broke the provided dataset up into data and labels and then performed principal component analysis on 4 features. This was the resulting scatter plot on the projections of the first two principal components:



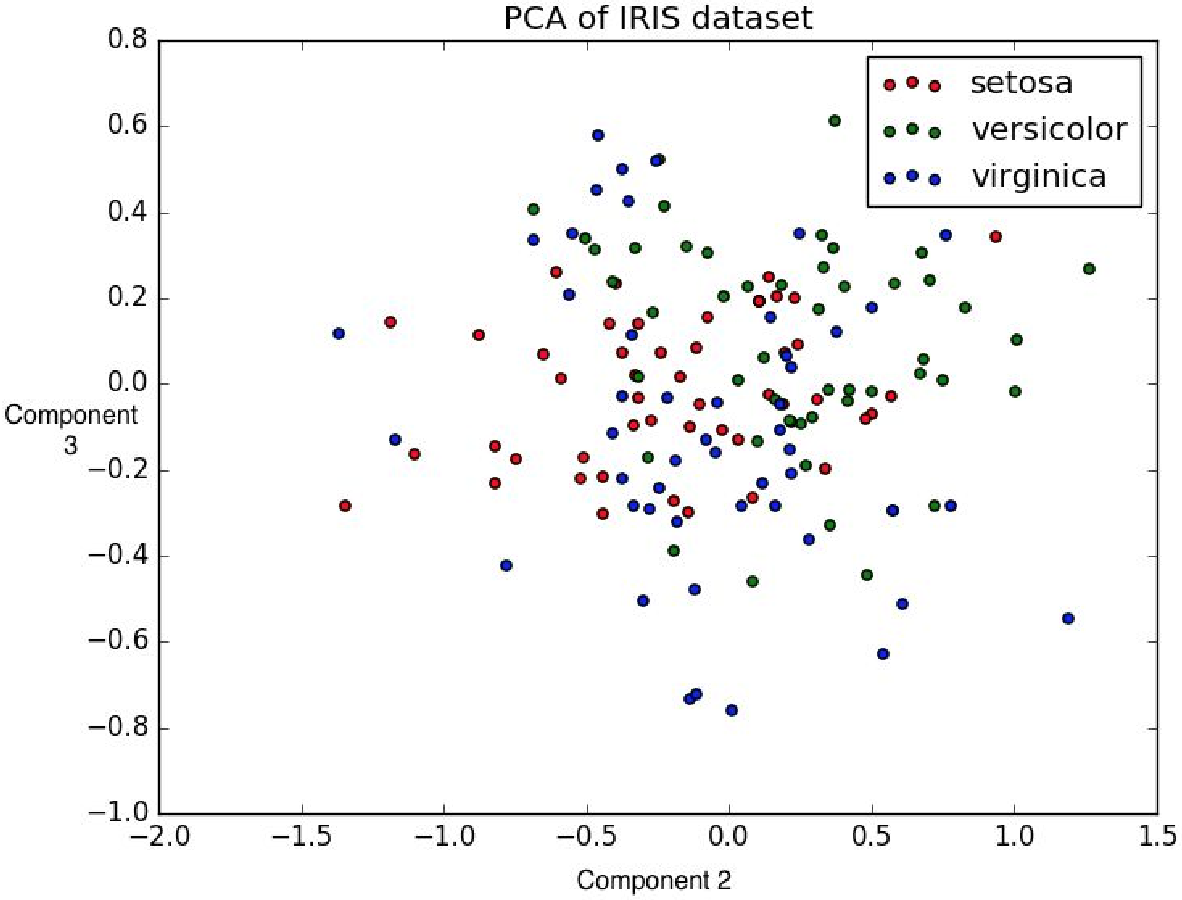
b. Looking at the projections of the dataset onto different pairs of components provides some interesting information. Looking at the following three scatter plots:

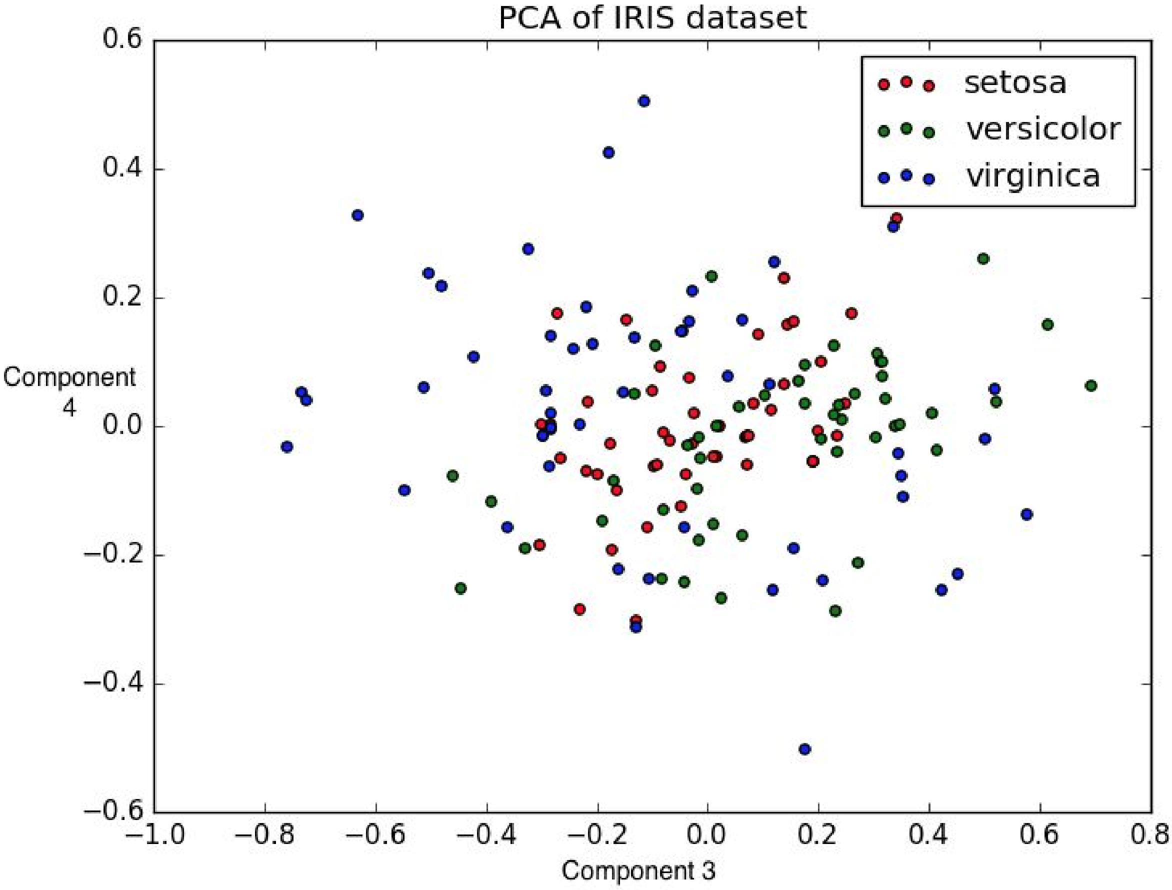

PCA of IRIS dataset
setosa
versicolor
virginica
Component 1
Component 2
5
4
3
2
1
0
−1
−2
−3
−4
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


PCA of IRIS dataset
setosa
versicolor
virginica
Component 1
Component 3
5
4
3
2
1
0
−1
−2
−3
−4
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

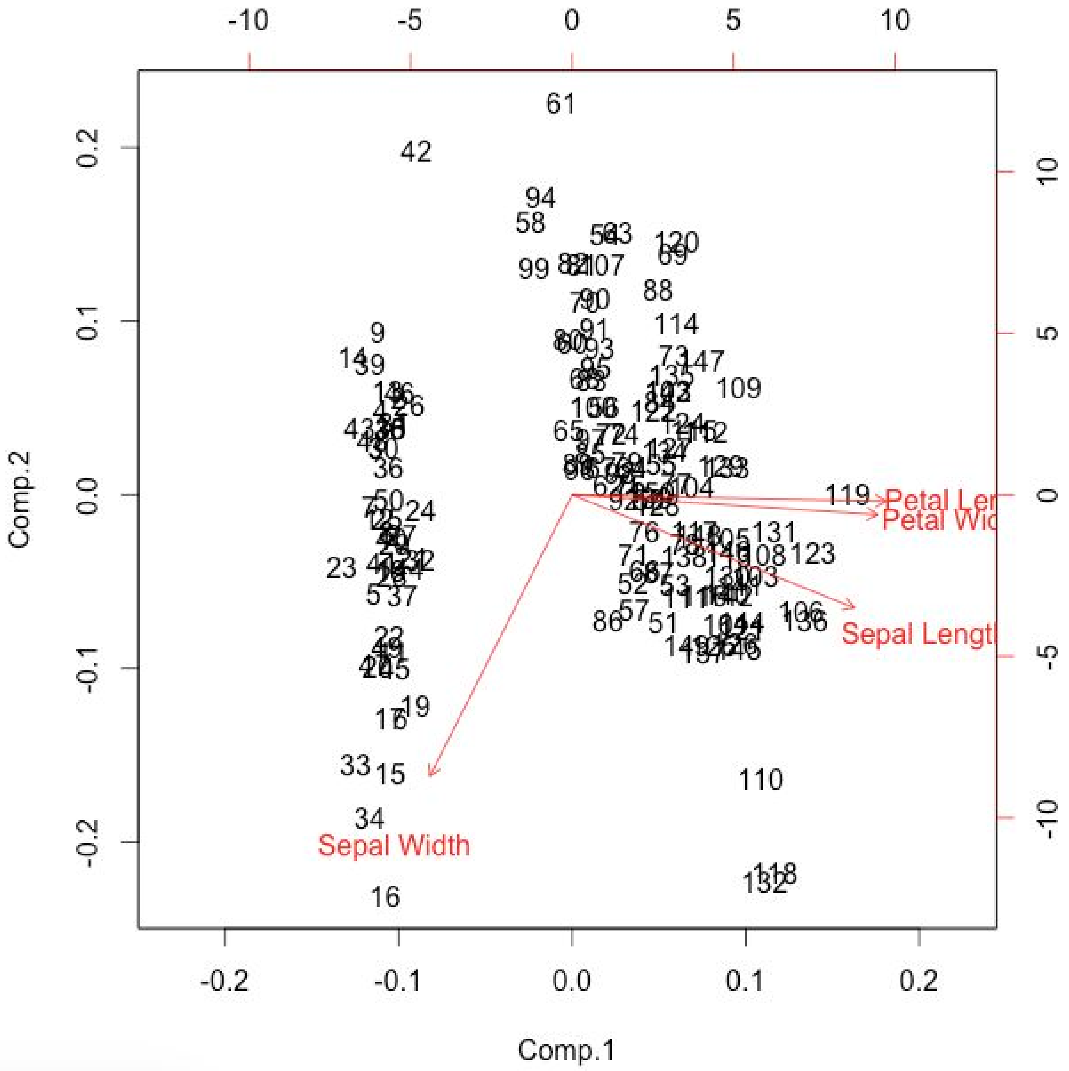


We see that Component 1 is somehow correlated to all of the other components and thus was a good choice to use for reducing the dimensionality of the dataset. Looking at the following scatterplots:


PCA of IRIS dataset
setosa
versicolor
virginica
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
Component 3
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Component 2



We see very little correlation between the other components. This is what we want from dimensionality reduction because it means that redundant data can be summarized by fewer attributes than in the original dataset. Looking back at the original scatter plots we see that the different categories of Irises can be identified mostly through component 1 (with help from 2). The biplot of this dataset confirms this:



One attribute encodes much of the pertinant information in the dataset.