Classification of Wine

Fangfang Lee

**Introduction**

In this project, I will perform classification algorithm support vector machine (SVM), k-nearest neighbors (KNN), and naïve Bayes on the famous wine dataset. I will briefly describe the dataset, show data visualization, describe my models, and evaluate their performances. Of the three models, SVM performs the best with a 98.8% accuracy rate. Naïve Bayes has an accuracy rate of 94% and KNN also has an accuracy rate of 94%. The models are evaluated based on their accuracy rate which is examined by viewing confusion matrix.

**Data**

The wine dataset is acquired through UCI Machine Learning Repository. The dataset comprises of 13 features and 1 outcome variable with three classes. The data are result of the chemical analysis found in wine grown in the same region in Italy but from three cultivars. The outcome variable is *wine,* which has three levels: type 1, 2 or 3. The goal of the classifiers is to correctly predict the type of wine based on their chemical composition. The data are measured on different scales depending on the type of chemical composition. Therefore, before performing the analysis, we will need to standardize the data.

The features of the wine dataset are:

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

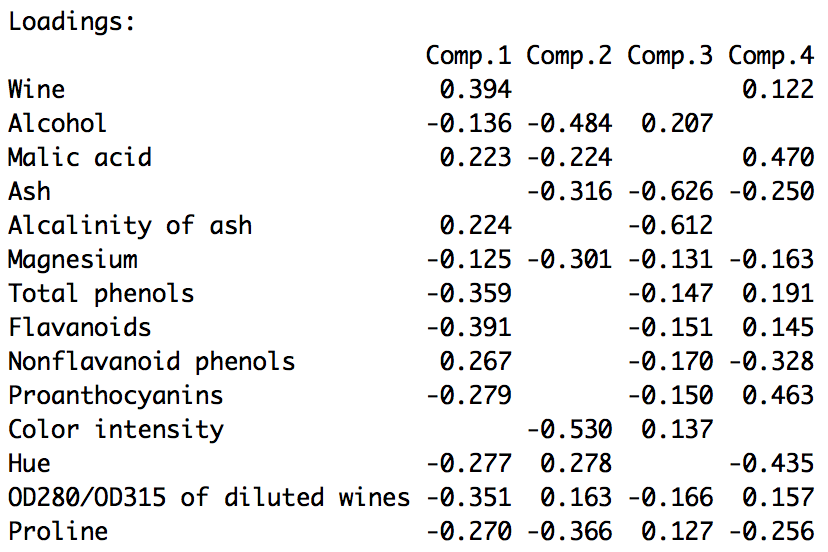
**Analysis**

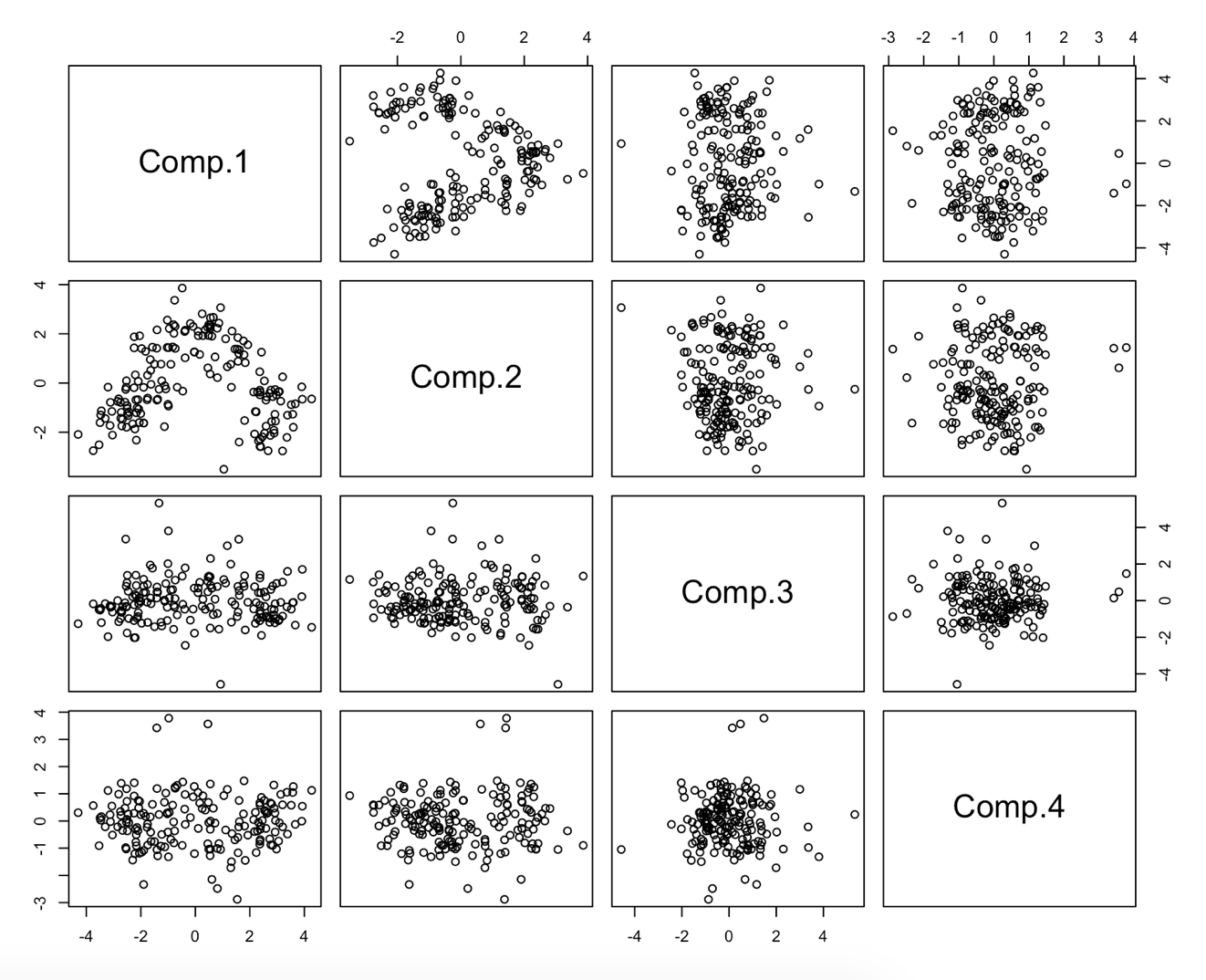
*Pre-processing detail*

The data will need to be standardized since some of the measurement are in different unit. However, the outcome should be treated as factor throughout the analysis. Let’s look at some visualization of the data.

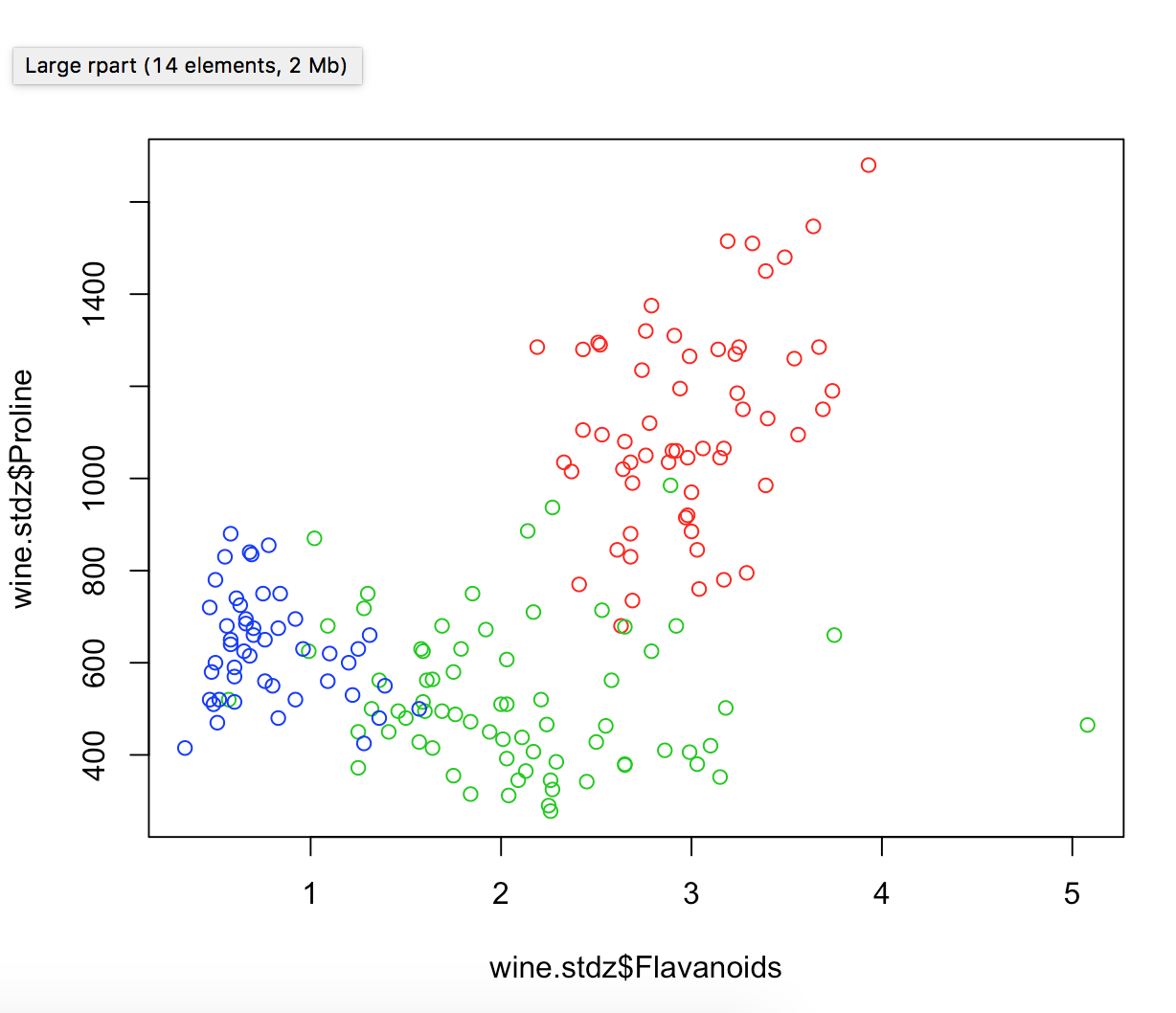
*Principal Component Analysis*

After conducting a Principal Component Analysis, we can see that in the first component, *Proline*, *Flavanoids, Total phenols, and OD280/OB315 of diluted wines* account for a huge proportion of the variance in the outcome. In other words, these variables account for the largest amount of variance (i.e. heaviest loadings) in the outcome variable *wine.*

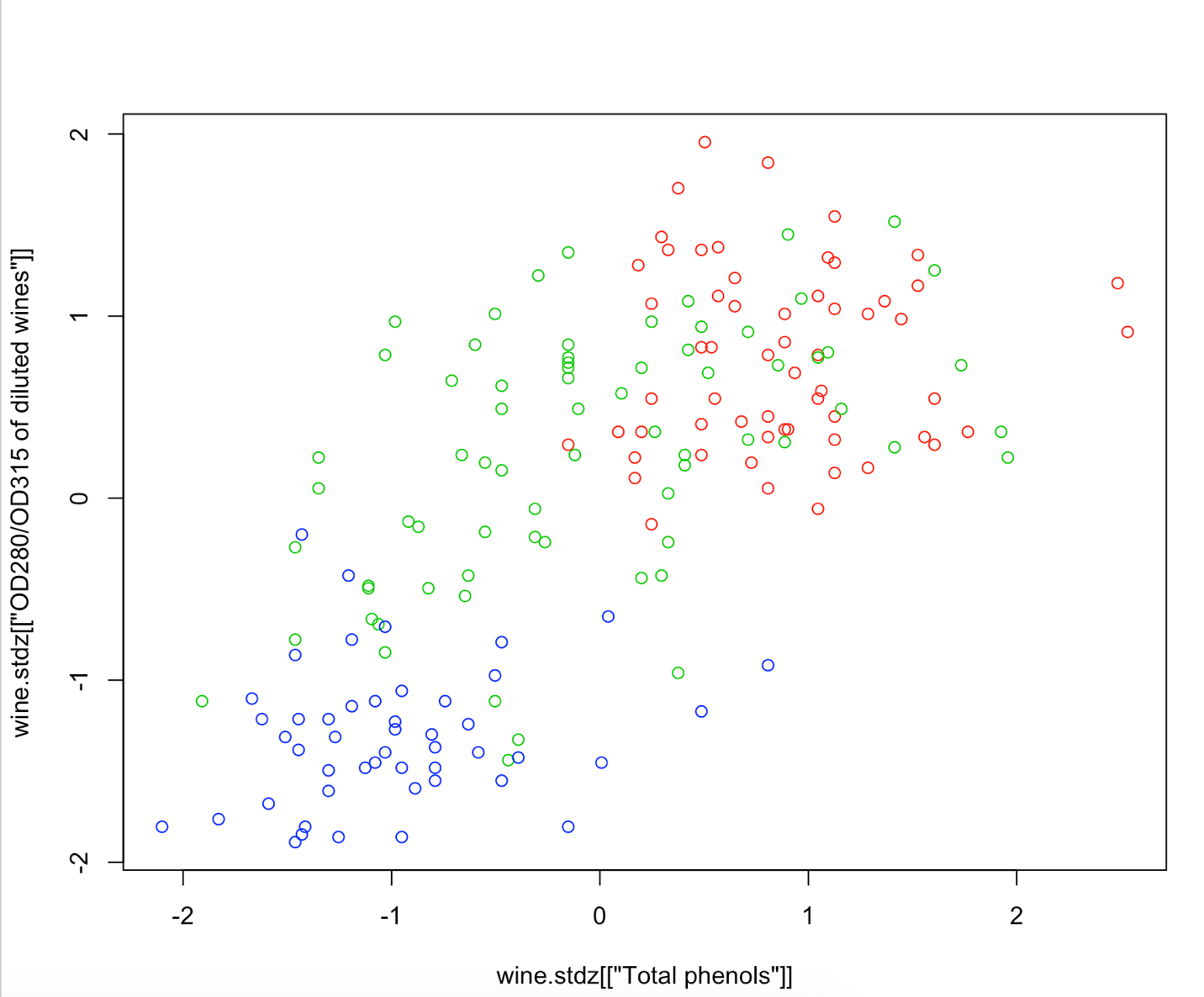




It will be interesting to plot the two variables, *proline* and *flavonoid,* colored by three different types of wine.



To examine the visualization further, we will also plot the other two variables, *Total phenols,* and *OD280/OB315 of diluted wine,* colored by groups of three types of wine.



Based on the scatterplot, we should feel hopeful because we see very well separated clusters. We have good reason to believe that our classifiers will perform well on the data.

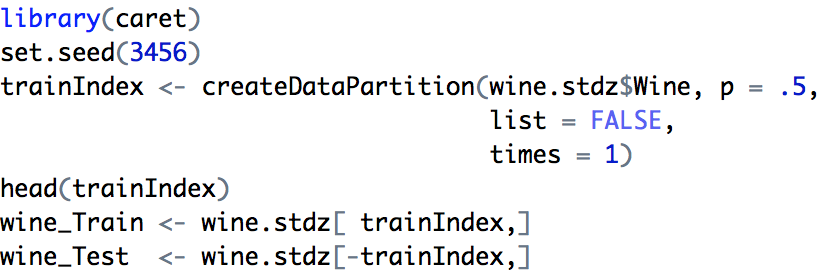
*Support Vector Machine(SVM)*

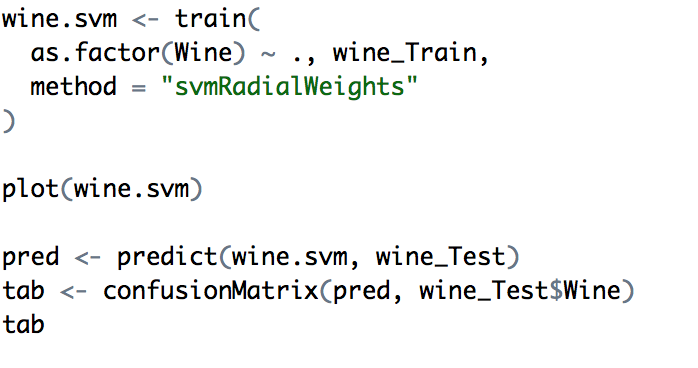
SVM is a creative and useful classification algorithm that often work very well with classes that can be linearly separated. SVM creates a hyperplane (p-1 dimension) that separates the different classes. If the data are not in fact separable, the algorithm will either inflates the feature space by creating a higher dimension, or softens the definition of separation.

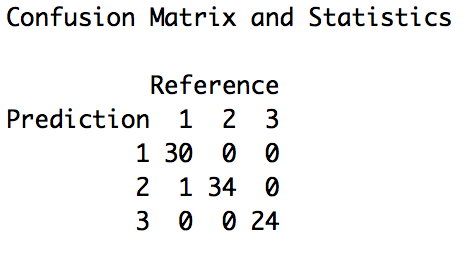
For the models in this project, I use the caret package, which trains and tunes the model with great level of flexibility. The model was implemented twice. The first time it was implemented with 10-fold cross validation and the second time the model was implemented with 50% manually split training and testing dataset. With 10-fold cross-validation, the model yields a 98.08% accuracy rate, and for manually split, the model yields a 98.8% accuracy rate. Each yields exceedingly well results. The output below shows the cross tabulation of the model with manual split. As we can see, there is only one prediction error where one observation in type 2 is misclassified in type 1.

I chose the support vector machine algorithm for this case because of the well separated classes we can witness on the colored scatterplot. The fact that SVM performed so well leads me to form the hypothesis that the data are linearly separable based on all the features, which I will examine

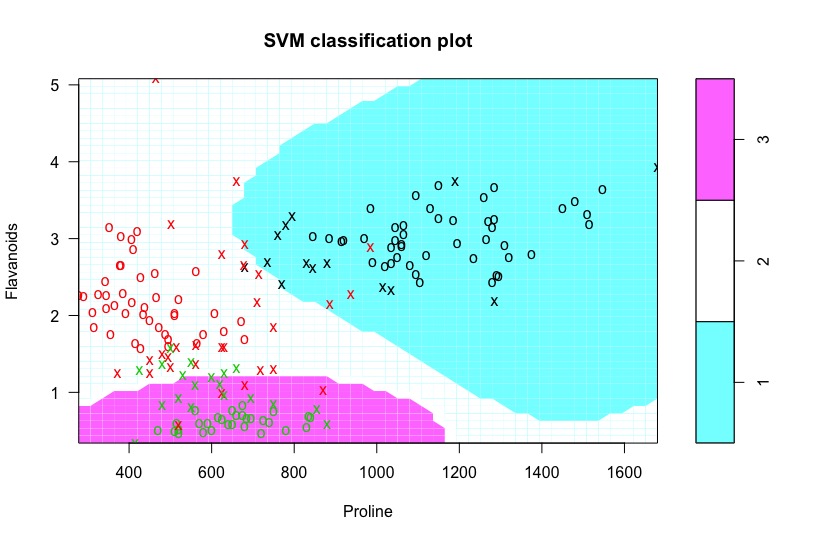
Later.







It will be helpful to view a visualization of the svm model with the two most important variables, *flavonoids* and *proline*.

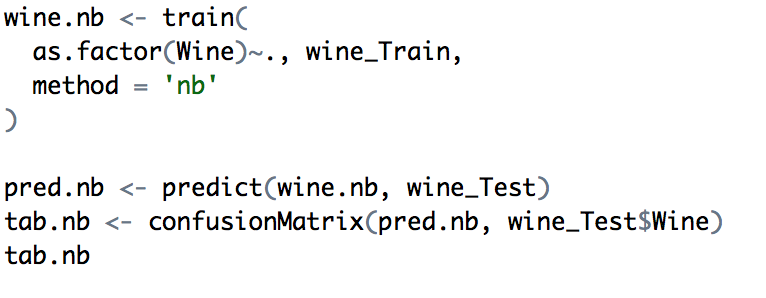


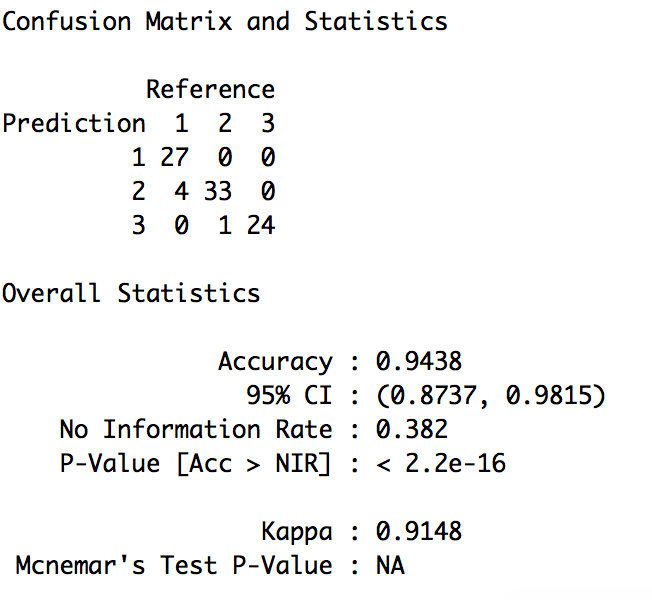
Based on the plot, we can see that even with only two predictors, the svm model performs fairly well with classifying the three classes of wine.

*Naïve Bayes*

Traditionally, Naïve Bayes classifier performs very well on most classification problems, especially with text data. It will be interesting to perform such model on our dataset, which seems to have fairly well separated groups.

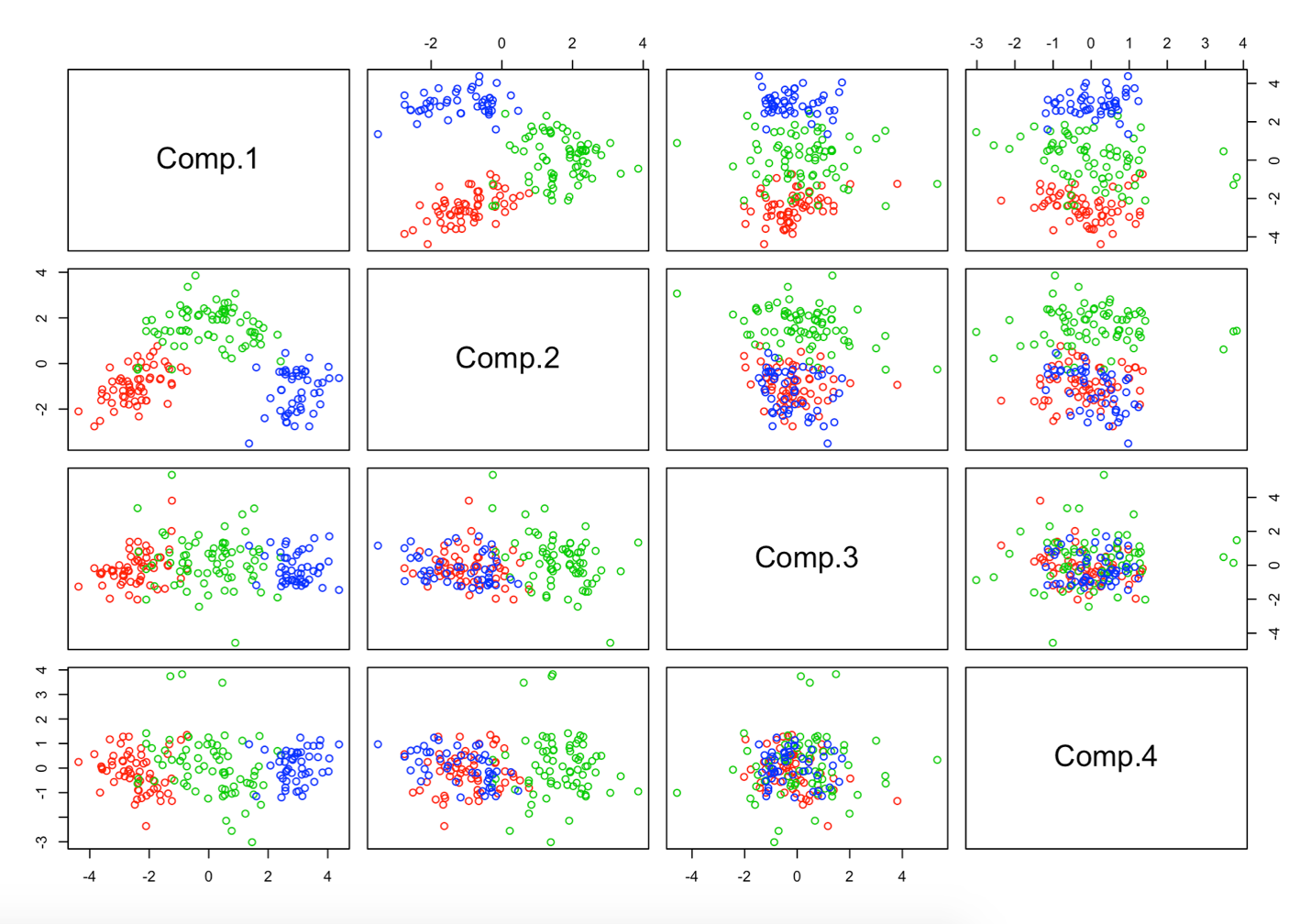
Unfortunately, even though Naïve Bayes classifier performs reasonably well, with an accuracy rate over 94%, it did not perform as well as SVM.



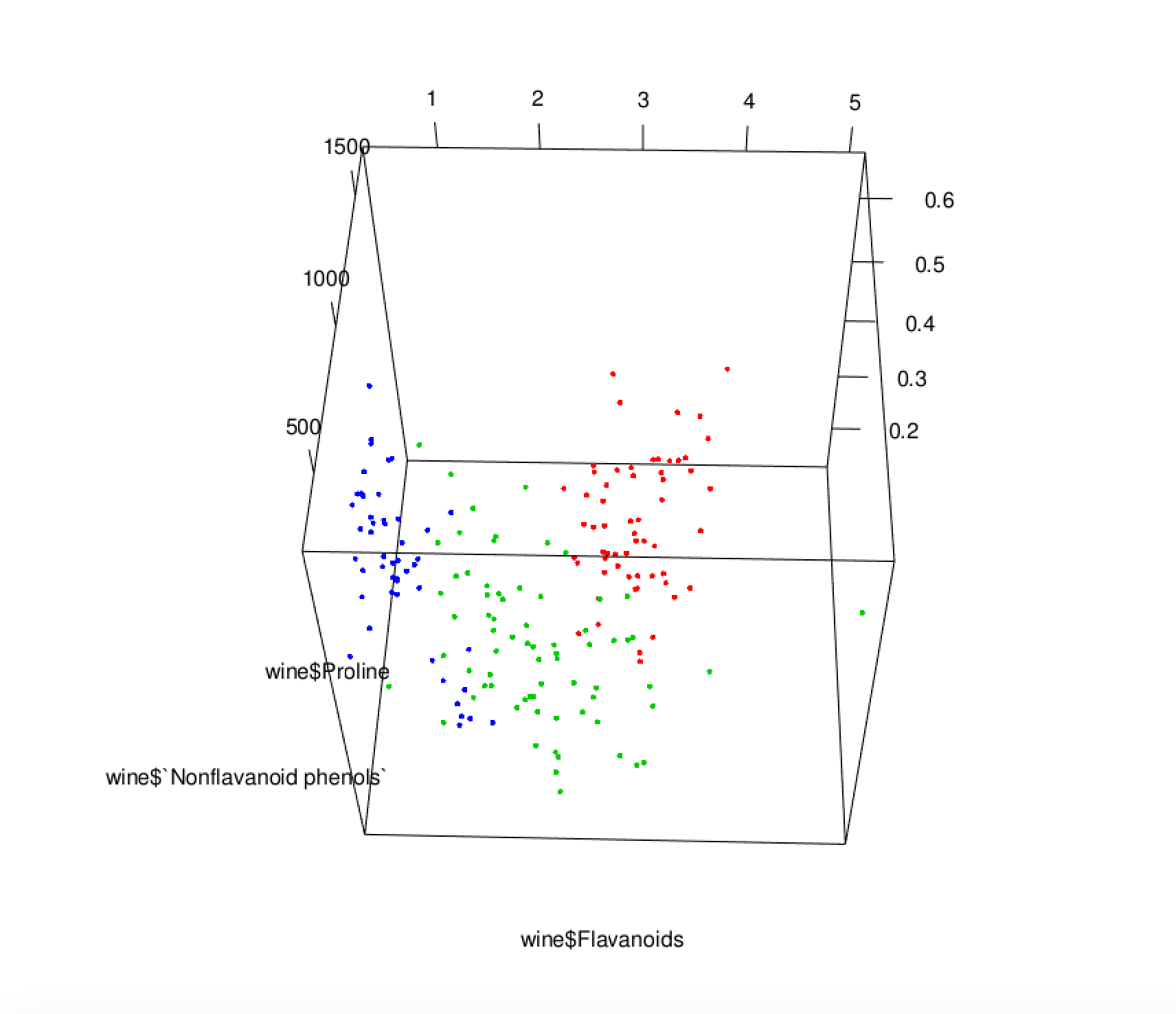


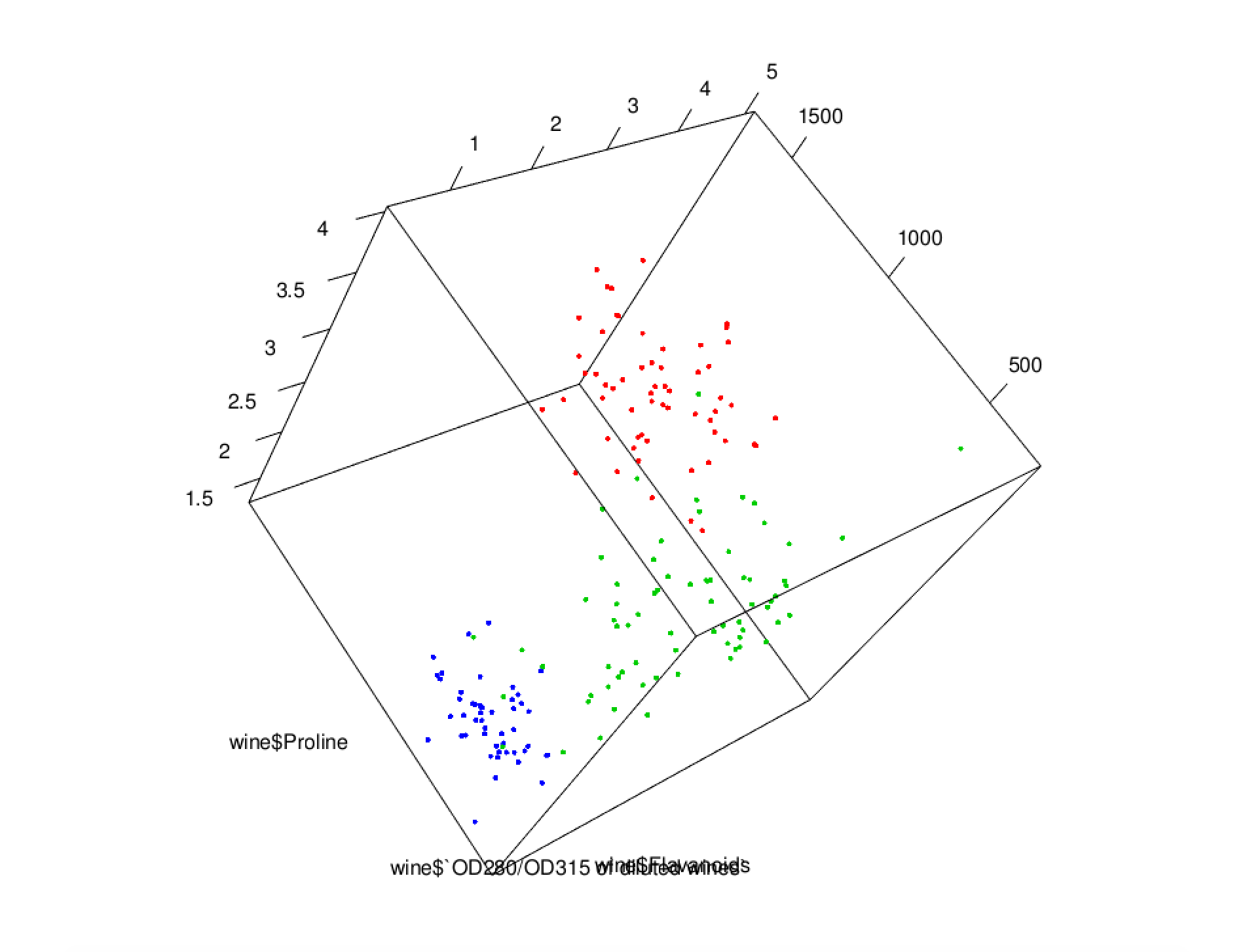
**Discussion**

Since SVM performed so well, I have reason to believe that the classes are very well separated. Taken all the features into consideration, we might find linear separation between the classes. We shall examine the principal components colored by groups of wine first.



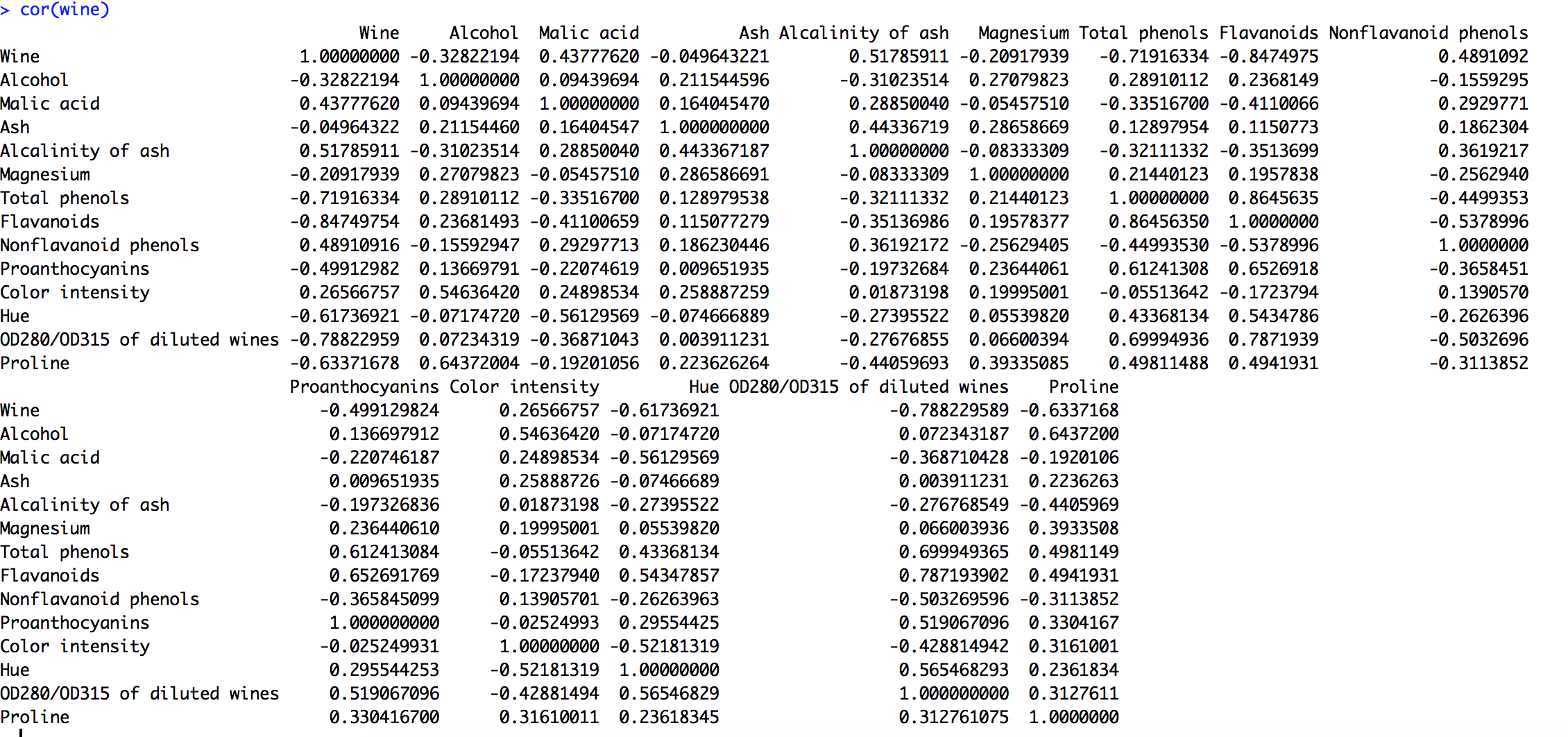
From the colored principal components, we can observe that the three classes are very well separated. Now, we will examine the variables *proline*, flavonoids and diluted wines in 3-D plots to see if these three features can be separated by hyperplane.





Based on the above 3-D plots, we can see that even with only three features (*Proline, Flavanoids*, and *diluted wines*), there seem to exist a hyperplane that can perfectly separate the three classes. The three dimensional images further corroborated my hypothesis earlier regarding linear separation between the classes.

Next, it is important to examine whether there is a feature disguised as the outcome variable. Based on examination of the correlation matrix, there is not a feature that correlates perfectly with the outcome variable. All the variables except for *Ash* has a medium to high correlation with *Wine*, the response.



In conclusion, the wine dataset has very well defined groups and very informative features. In the future, more works can be done in order to examine whether the data are linearly separable throughout the feature space. Algorithms such as multilayer perceptron can be employed for such purpose.