The Wine Data set

CS 7616 Home work 0

Anisha Gartia (GTID: 903136557)

The wine data set are results of a chemical analysis of wines grown in the same region in Italy but, derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes 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

The data can be classified into 3 classes (Class 1, 2 and 3). The first column of data set is the class identifier. See here (https://archive.ics.uci.edu/ml/datasets/Wine) for more information on this data set.

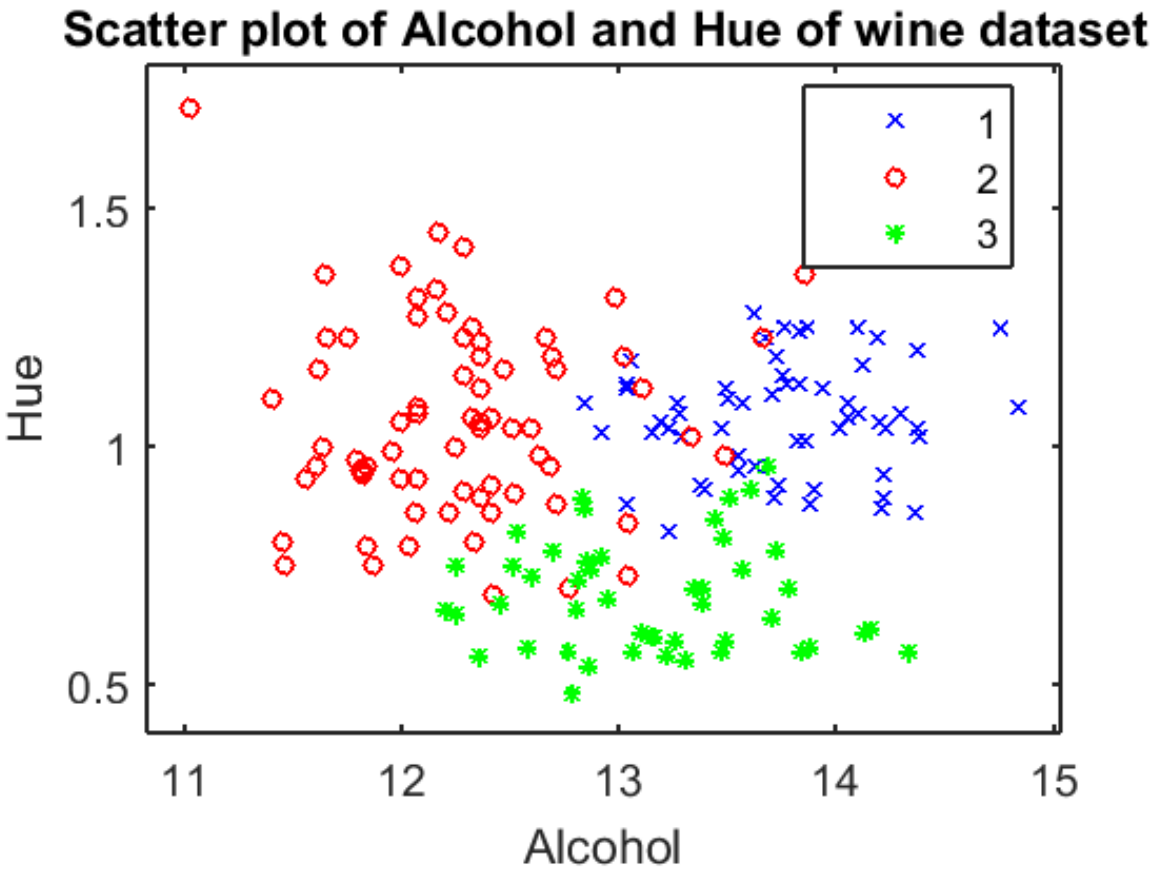
In our analysis of this data set, we have plotted how the Hue of wine varies with Alcohol content, and how the color intensity varies with dilution. We have generated scatter plots for these distributions while clearly distinguishing the classes.

Additionally, we have performed principal component analysis (PCA (https://en.wikipedia.org/wiki/Principal_component_analysis)) of the data, and plotted the first three components on a 3D scatter plot.

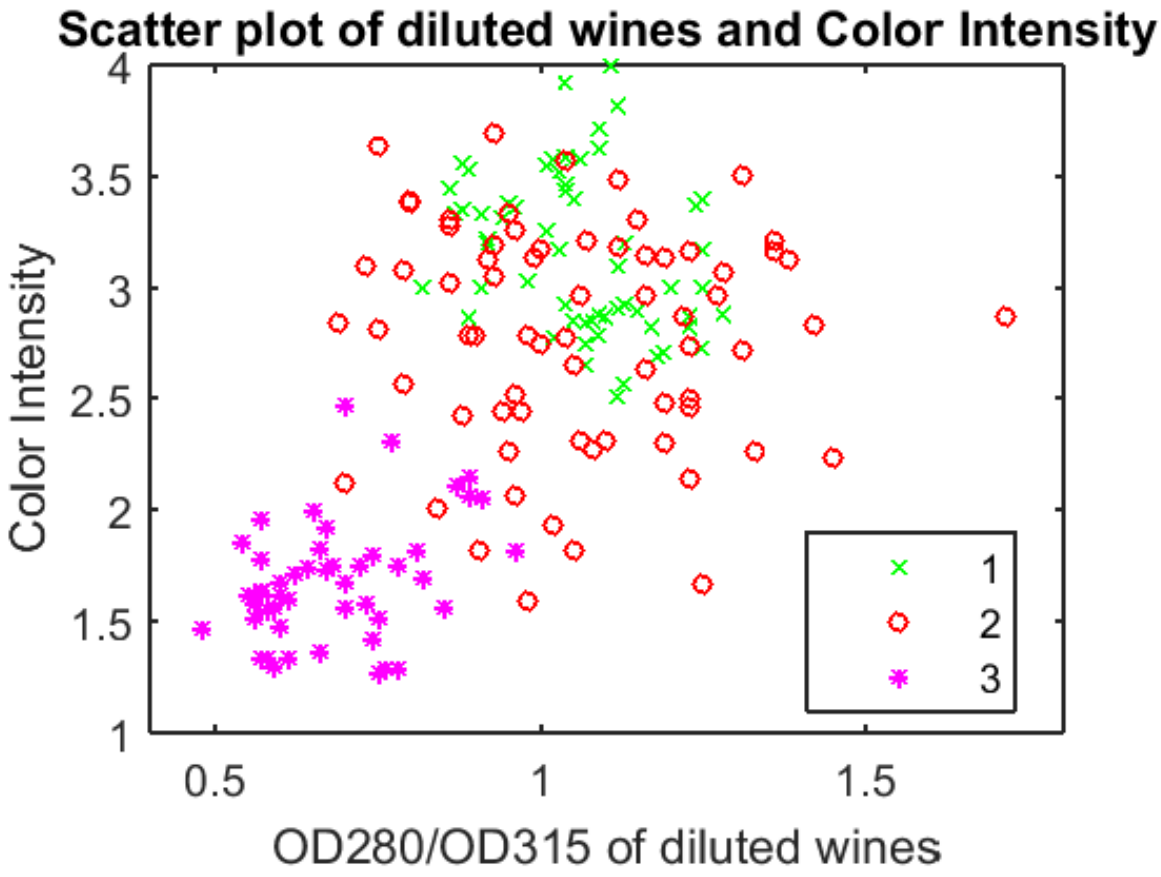
This analysis is done using Matlab.

Matlab code and implementation:

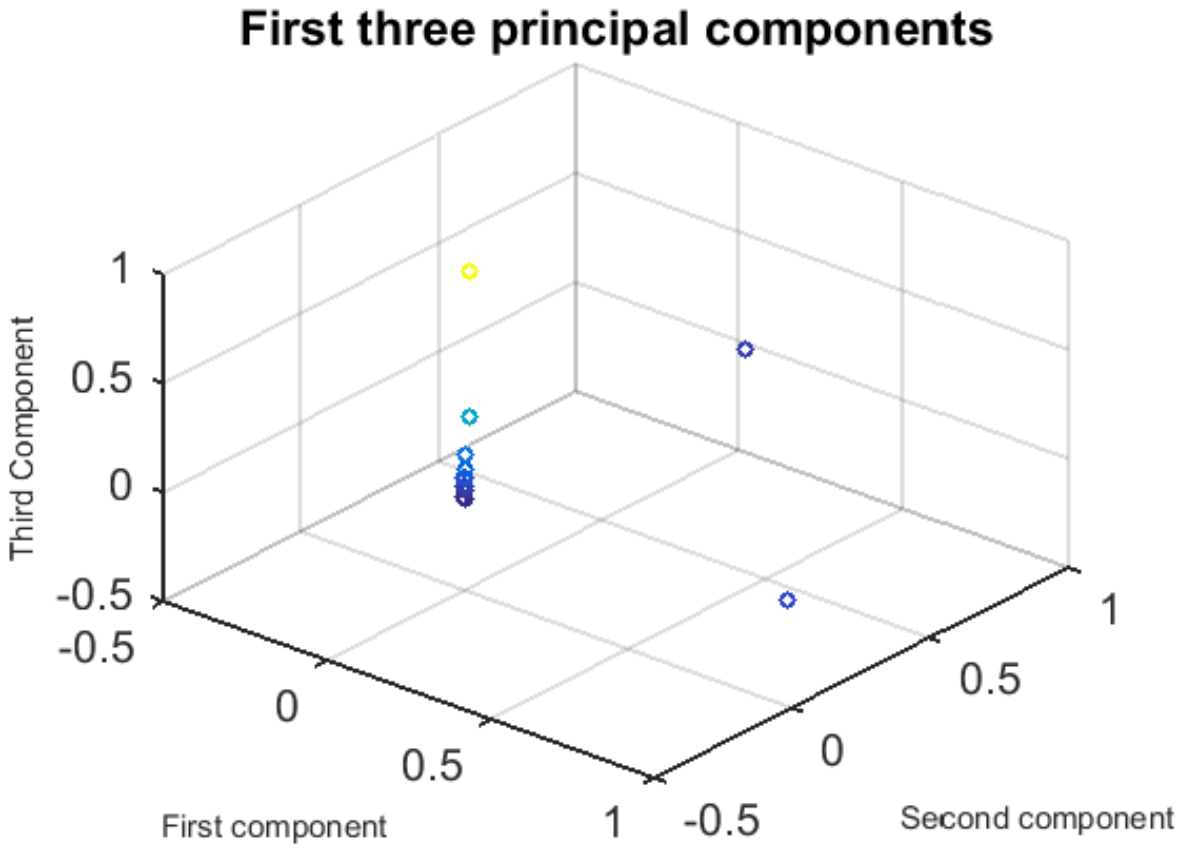
In [2]:
```
% ------------------------------- %
% Wine data set Analysis
% Author: Anisha Gartia
% Last Modified: January 23, 2016
% ------------------------------- %

%--- Import wine data set for analysis ---%
fwine = 'csv_winedata.csv';
wine = csvread(fwine);

%--- Import list of attlribute names---%
mycell = textread('attributenames.txt','%s');
attrib = char(mycell);

%--- Scatter plot for Alcohol vs Hue ---%
x_index = strmatch('Alcohol',attrib);
X = wine(:,x_index+1);
y_index = strmatch('Hue',attrib);
Y = wine(:,y_index+1);
gscatter(X,Y,wine(:,1),'brg','xo*',4);
xlabel('Alcohol');
ylabel('Hue');
title('Scatter plot of Alcohol and Hue of wine dataset');
```



In [3]:
```
%--- Scatter plot for Dilution vs Color Intensity ---%
x_index = strmatch('OD280/OD315_of_diluted_wines',attrib);
X = wine(:,y_index+1);
y_index = strmatch('Color_intensity',attrib);
Y = wine(:,x_index+1);
gscatter(X,Y,wine(:,1),'grm','xo*',4);
ylabel('Color Intensity');
xlabel('OD280/OD315 of diluted wines');
title('Scatter plot of diluted wines and Color Intensity');
```



In [36]:
```
%--- Perform Principal component analysis and plot the first three compo
nents ---%
pcacoeff = pca(wine(:,2:end));
x = pcacoeff(:,1);
y = pcacoeff(:,2);
z = pcacoeff(:,3);
scatter3(x,y,z,10,z)
xlabel('First component','FontSize',7)
ylabel('Second component','FontSize',7)
zlabel('Third Component','FontSize',7)
title('First three principal components')
view(40,40)
```



Observations:

1. We observe that from Alcohol vs Hue scatter plot, we cannot see a clear linear classification in classes.
2. In the OD280/OD315 of diluted wines vs Color Intensity scatter plot, we observe a general trend where wines of higher dilution have lower color intensity. Here although we cannot make a strict classification, a majority of samples of each class are present near each other.
3. On performing principal component analysis, we observe a much more distinct classification in scatter points in the first three principal components.