IDA Assignment 2023

Chan Zhi Qing

2023-11-04

Data Description

The "Wine" data set obtained from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/109/wine) is a data set that contains 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 wine: Barolo, Grignolino, Barbera. The data set consists of 178 instances, each representing a sample of wine.

The first attribute of the data set is the class identifier (1-3) which represents each type of wine (Barolo, Grignolino, Barbera)

The remaining 13 attributes in the data set are:

1. Alcohol - The alcohol content of the wine.
2. Malic acid - The amount of malic acid in the wine.
3. Ash - The ash content of the wine.
4. Alcalinity of ash - The alkalinity of the ash in the wine.
5. Magnesium - The magnesium content in the wine.
6. Total phenols - The total phenolic content in the wine.
7. Flavanoids - The total flavonoid content in the wine.
8. Nonflavanoid phenols - The amount of non-flavonoid phenols in the wine.
9. Proanthocyanins - The amount of proanthocyanins in the wine.
10. Color intensity - The color intensity of the wine.
11. Hue - The hue of the wine.
12. OD280/OD315 of diluted wines - The OD280/OD315 ratio of diluted wines.
13. Proline - The proline content of the wine.

The first 6 rows of the data set looks like this:

```
##   Class Alcohol MalicAcid  Ash Alcalinity Magnesium Phenols Flavanoids
## 1     1   14.23      1.71 2.43       15.6       127    2.80       3.06
## 2     1   13.20      1.78 2.14       11.2       100    2.65       2.76
## 3     1   13.16      2.36 2.67       18.6       101    2.80       3.24
## 4     1   14.37      1.95 2.50       16.8       113    3.85       3.49
## 5     1   13.24      2.59 2.87       21.0       118    2.80       2.69
## 6     1   14.20      1.76 2.45       15.2       112    3.27       3.39
##   Nonflavanoids Proanthocyanins Color  Hue OD_Ratio Proline
## 1          0.28            2.29  5.64 1.04     3.92    1065
## 2          0.26            1.28  4.38 1.05     3.40    1050
## 3          0.30            2.81  5.68 1.03     3.17    1185
## 4          0.24            2.18  7.80 0.86     3.45    1480
## 5          0.39            1.82  4.32 1.04     2.93     735
## 6          0.34            1.97  6.75 1.05     2.85    1450
```

From the data, we may asked questions like which chemical attribute contribute the most to the principal components or in other words which chemical content is usually higher in the different type of wines?

Data Preprocessing

We will convert the categorical target variable into a factor which in this case is the first column of the data set that represents the classes of the wine.

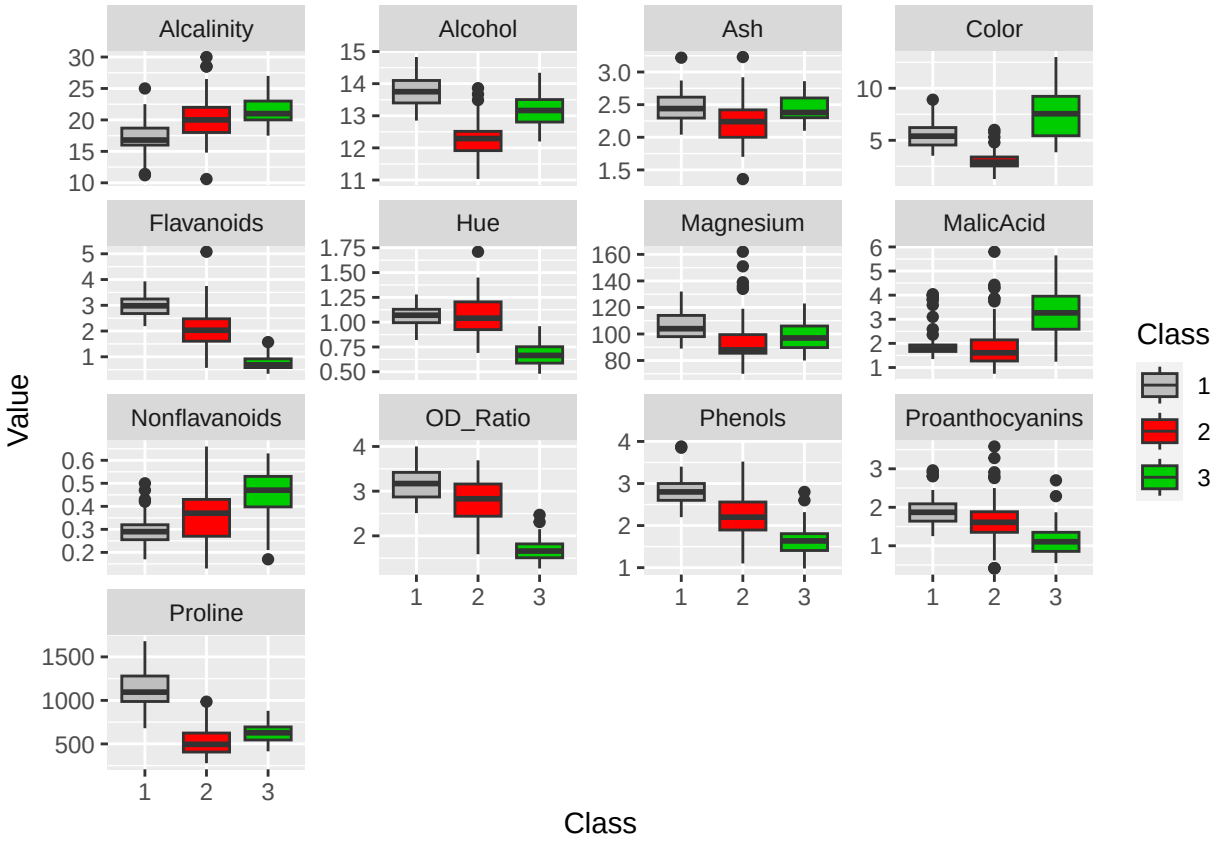
As data is classified into 3 different types of wine, we will use different markers/colours to represent the types in the plots that will be shown in the later part, where:

1 - Barolo (black)

2 - Grignolino (red)

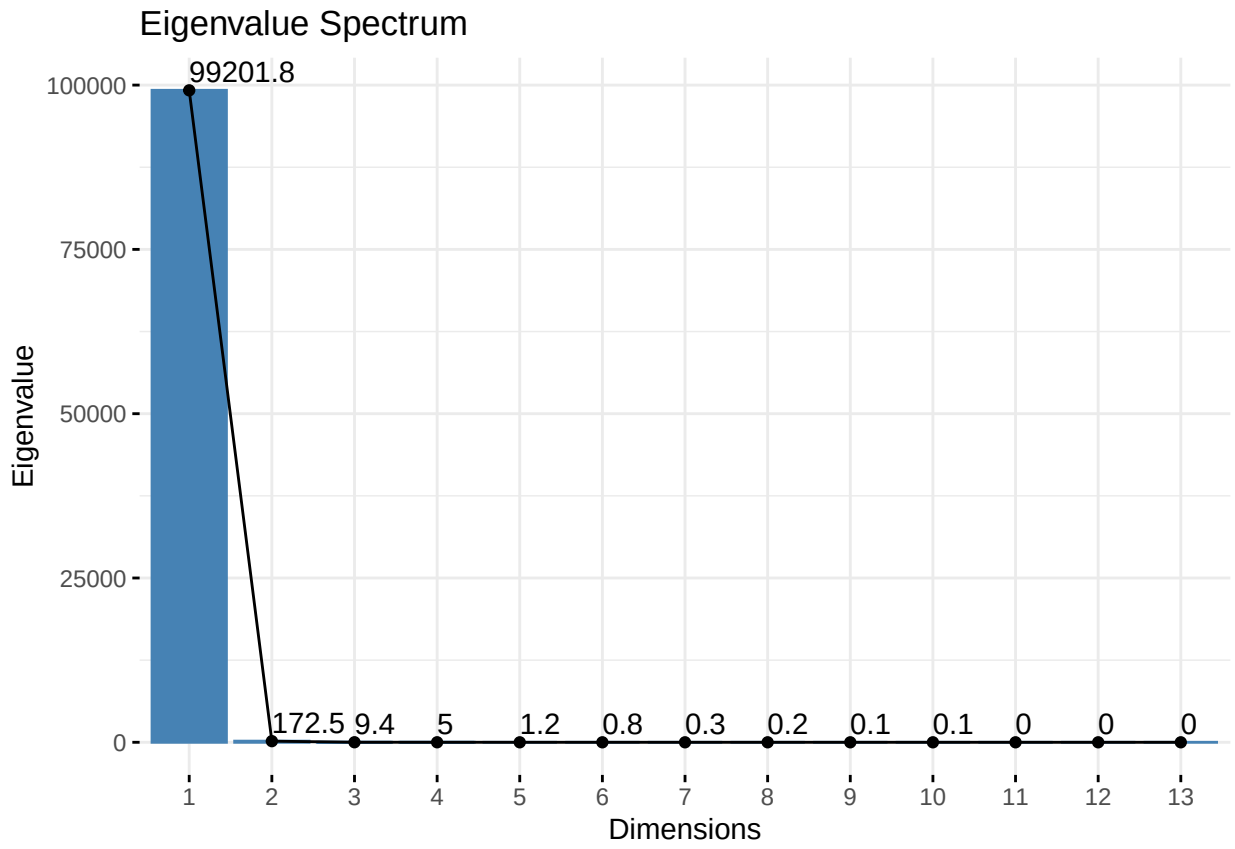
3 - Barbera (green)

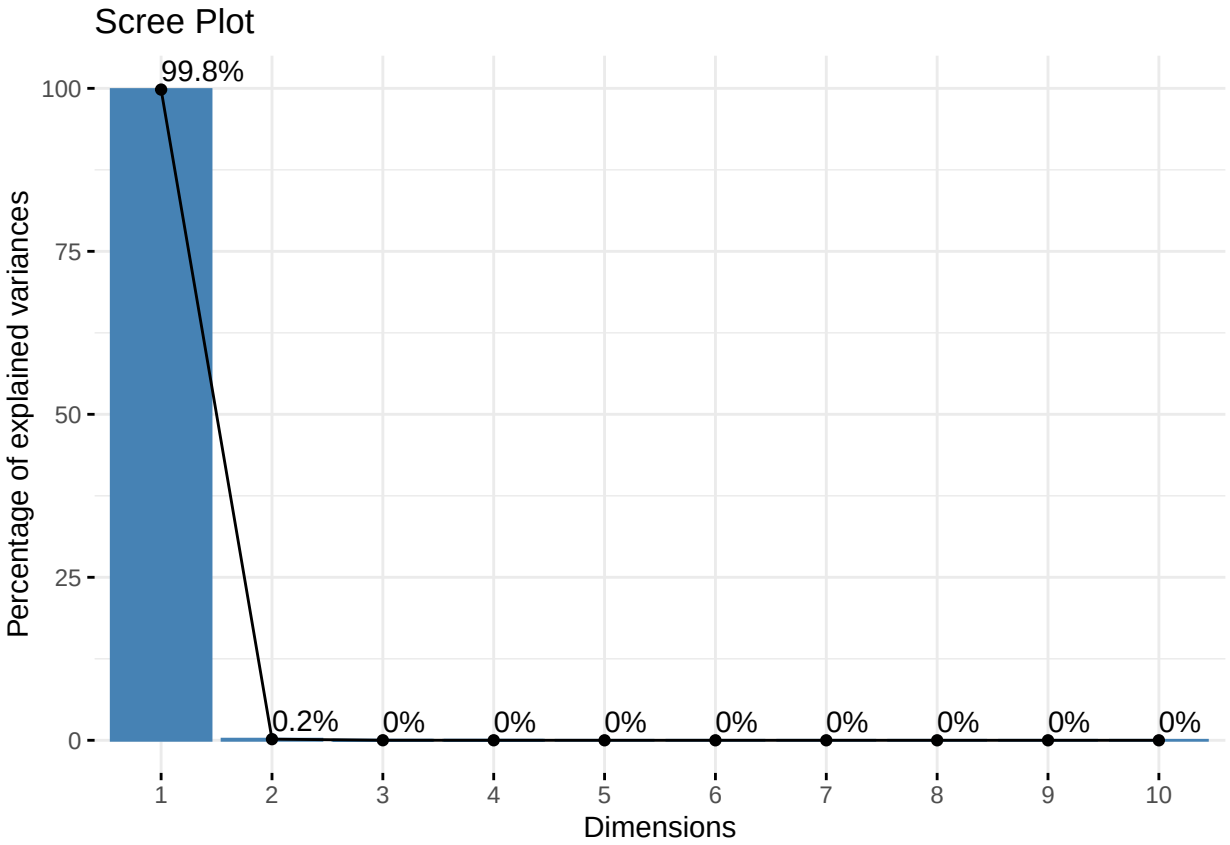
From the dataset we can plot box plots to identify the spread of each variable and potential outliers.



Component Selection & Data Visualization (PCA)

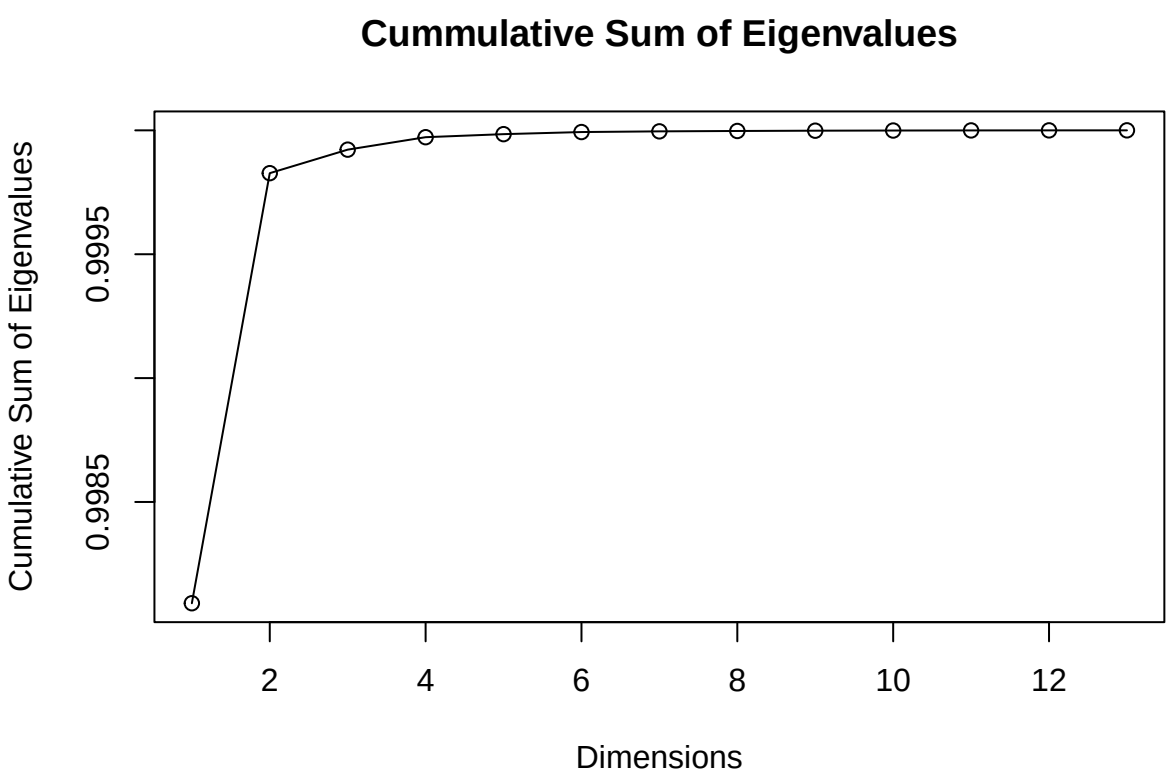
After data preprocessing, we will now attempt PCA. We will first compute the eigenvalues of the covariance matrix and plot them as below from largest to smallest.


Eigenvalue Spectrum
99201.8
172.5
9.4
5
1.2
0.8
0.3
0.2
0.1
0.1
0
0
0
100000
75000
50000
25000
0
Eigenvalue
1
2
3
4
5
6
7
8
9
10
11
12
13
Dimensions



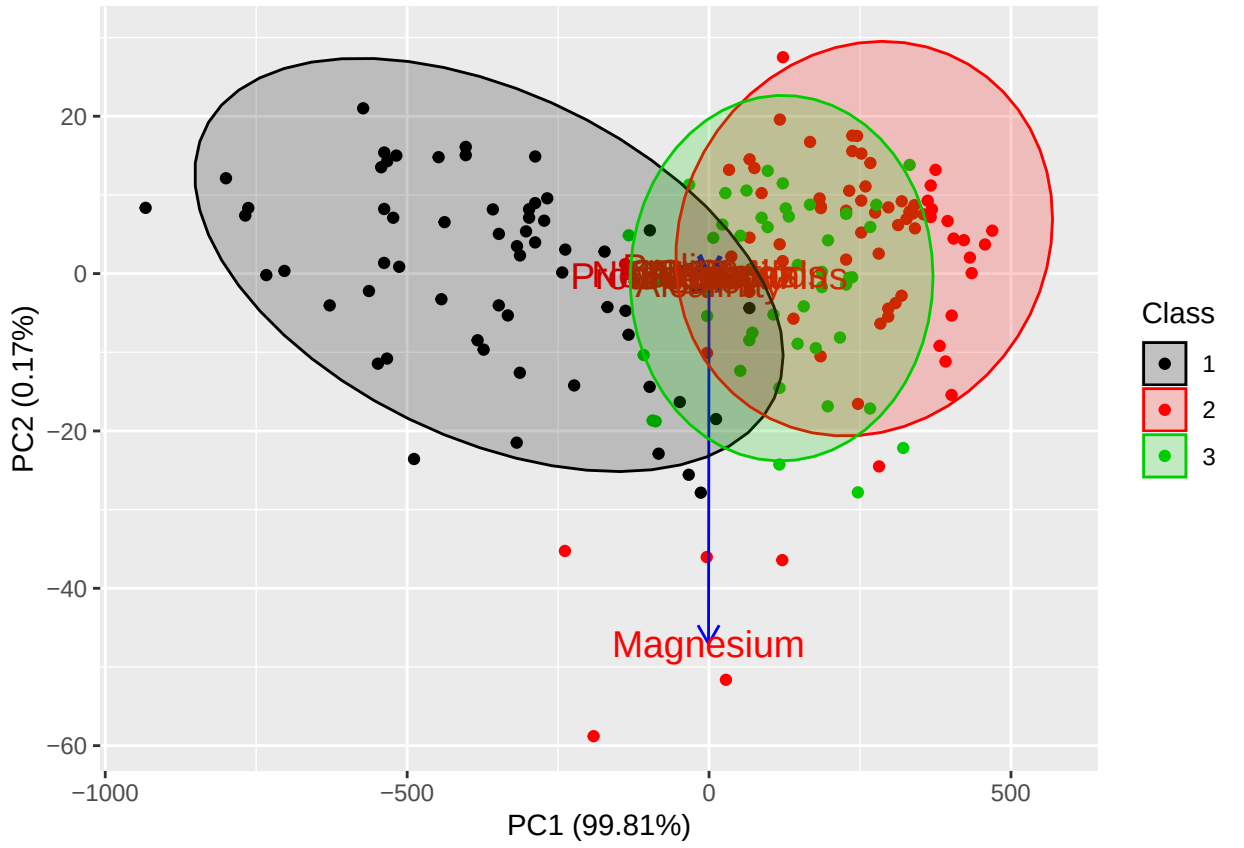
Higher eigenvalue indicates that the corresponding principal component explains more variance in the data. Therefore, based on the plot we can say that most data variance is explained by the first eigenvector of the covariance matrix where the principal component explains 99.8% of the variance.

```
##  [1] 0.9980912 0.9998271 0.9999221 0.9999723 0.9999847 0.9999931 0.9999960
##  [8] 0.9999975 0.9999986 0.9999993 0.9999997 0.9999999 1.0000000
```



The cumulative variance plot shows how much total variance is explained as we include more principal components. As we can see the curve starts to flatten out after the third component and therefore we may include just the first two component without losing too much of the variance.

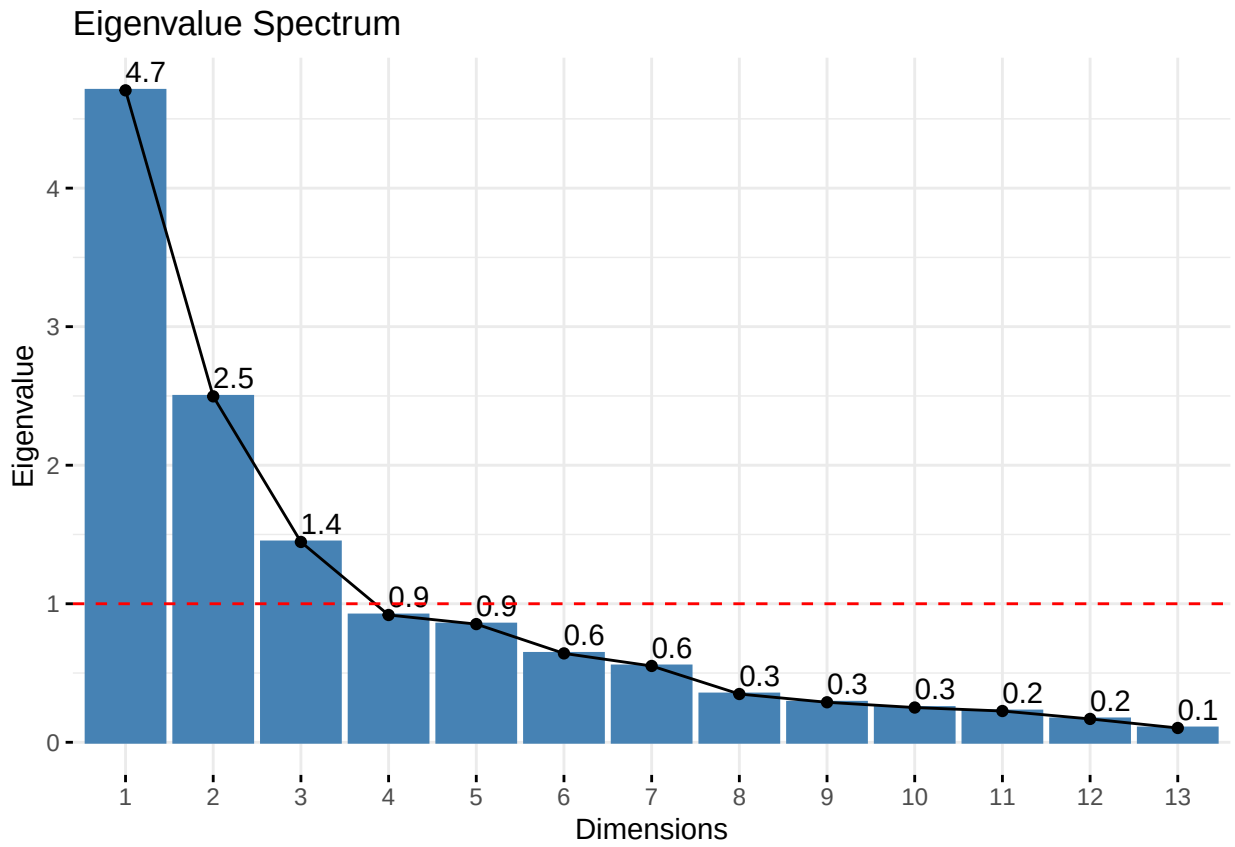
After deciding which principal component to include, we now plot the PCA as below.



However, it can be seen that the plot does not provide much information as it highly overlapped and clustered. This indicates that the data needs further processing in order to plot something more informative.

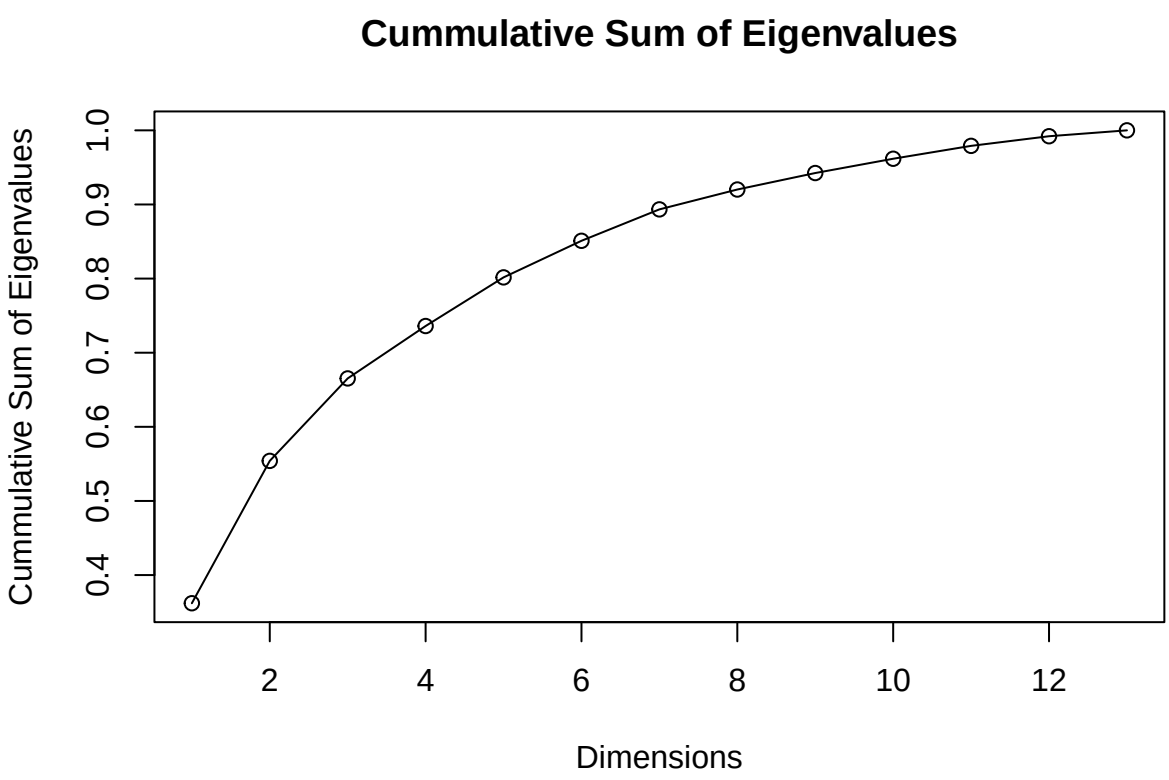
Further Data Preprocessing

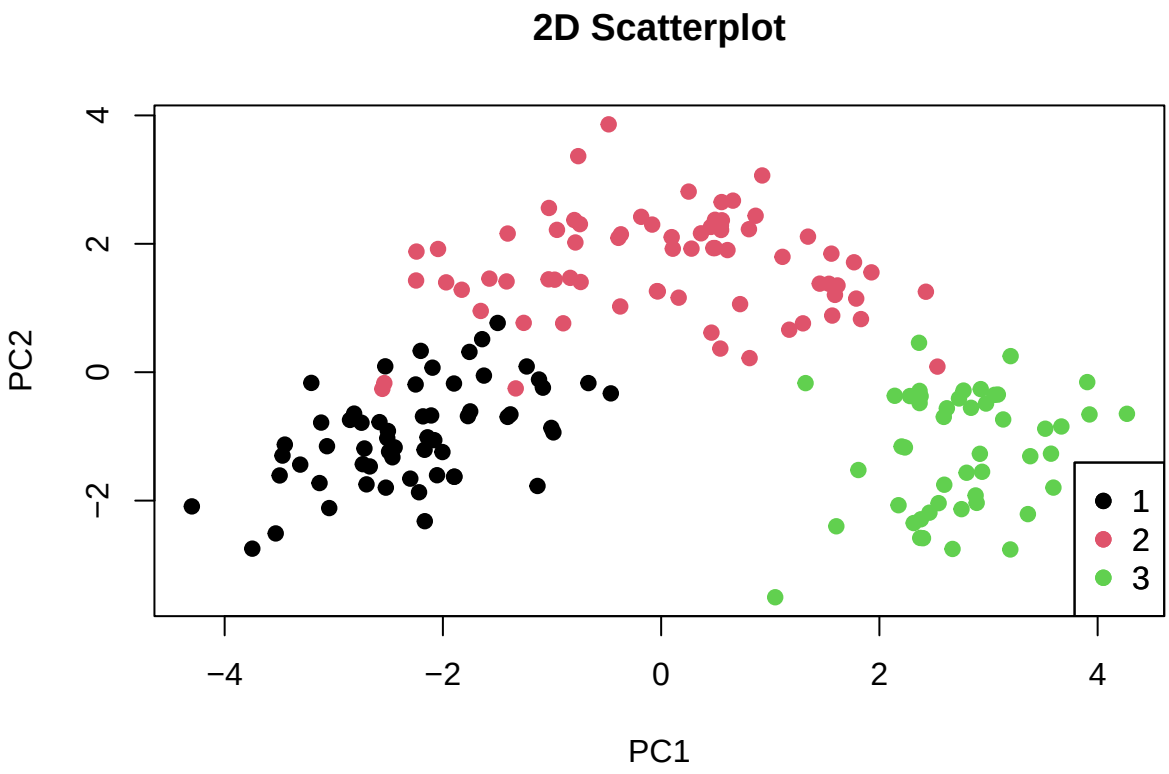
It is important that we standardize the data as each feature in the original data has different scales causing the plots to look out of portion. We can do this by centering the data around the origin and scaling the data so that all data points have a similar range. Therefore, the data now has mean of 0 and standard deviation of 1.

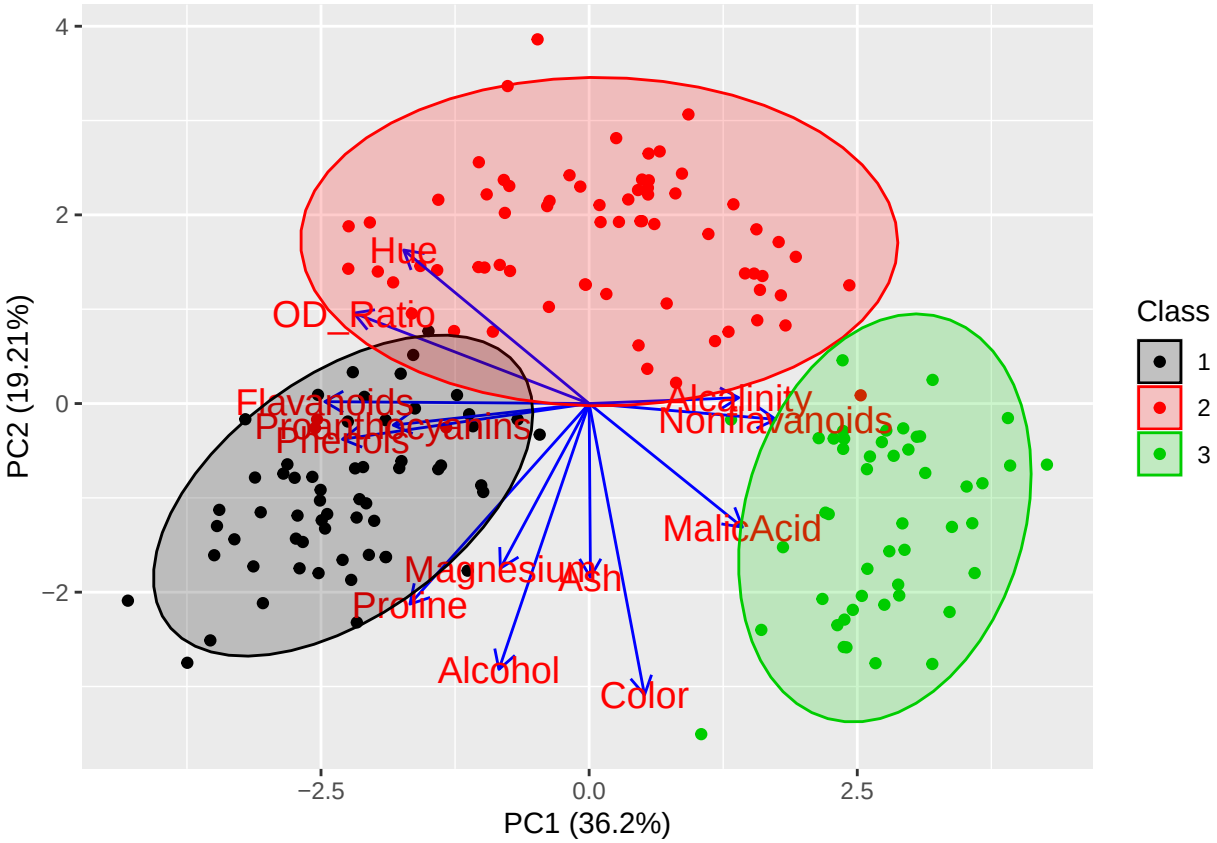


After further preprocessing of the data, we can now see that the eigenvalue decreases more smoothly than before. However, the majority of the variance is still mainly explained by the first 2 or 3 components.

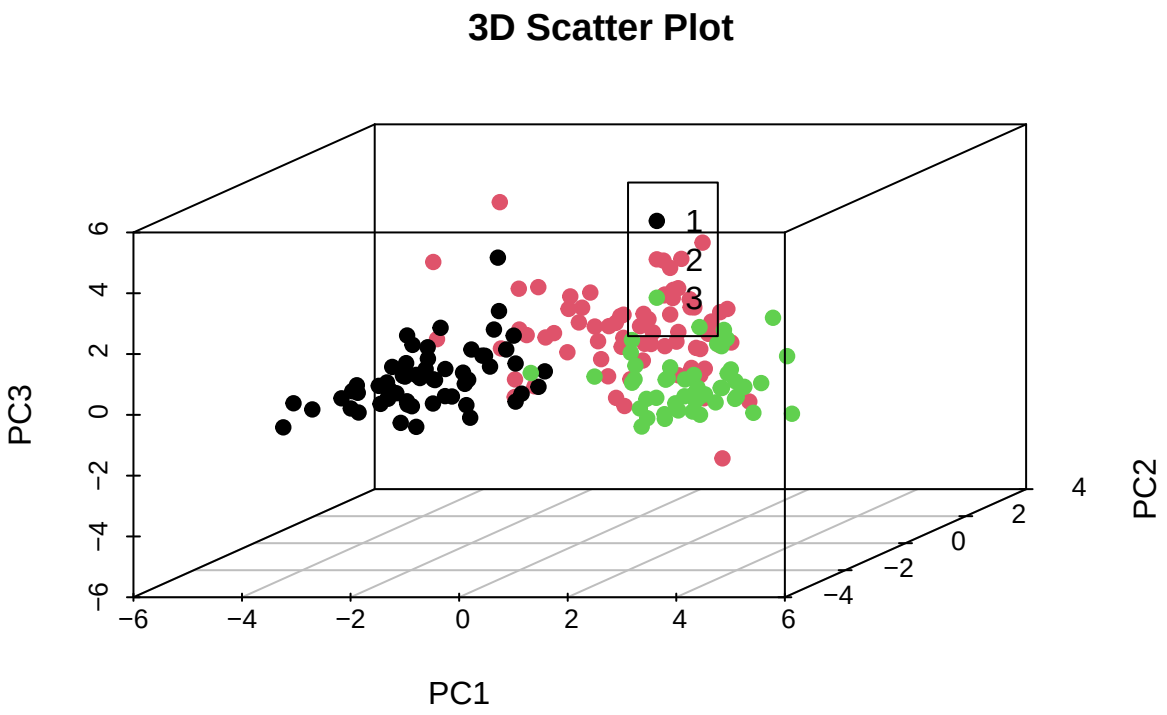
```
##  [1] 0.3619885 0.5540634 0.6652997 0.7359900 0.8016229 0.8509812 0.8933680
##  [8] 0.9201754 0.9423970 0.9616972 0.9790655 0.9920479 1.0000000
```


Cummulative Sum of Eigenvalues
Cummulative Sum of Eigenvalues
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2
4
6
8
10
12
Dimensions


2D Scatterplot
PC2
PC1
4
2
0
−2
−4
−2
0
2
4
1
2
3



From the PCA plots above, each point represents a sample of the wine and the vectors tells us about the feature relationship with the corresponding component. For instance, features such as alcalinity and content of malic acid and non-flavonoids in the wine are positively correlated with PC1. Therefore we can say that Barbera (wine class 3) tend to have higher content of those 3 chemicals.



We can also plot a 3D scatter plot to visualise the relationship between the classes and the first 3 principal components as we can see from the eigenvalues plot after data scaling, the first 3 eigenvectors have values more that 1 which explains majority of the variance in the data.

In conclusion, the analysis of eigenvalues and use of PCA allowed us to effectively reduce the dimension but still able to preserve the key chemical information in the different types of wine and identify features that play a crucial role in distinguishing between classes. It is also interesting to see different chemical composition of the wines and gain insights about the chemical constitution of the wine through analysis of the data, hence investigating which chemical attributes really differentiates the types of wine.

Full code and files can be found at https://github.com/zhiqing0923/PCA-on-wine-dataset