Data Analysis – Wine Quality

The dataset used in this analysis is the Wine recognition dataset, which presents a possible set of data. It provides information about the chemical composition of wines that are grown in the same region of Italy but belong to three different cultivars. The data categorizes the three cultivars into distinct classes.

There are a total of 178 datapoints in the dataset, which are distributed among the classes in the following manner:

Class 1 contains 59 datapoints, Class 2 contains 71 datapoints, and Class 3 contains 48 datapoints.

Each of these data points has 13 dimensions, besides their Class

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

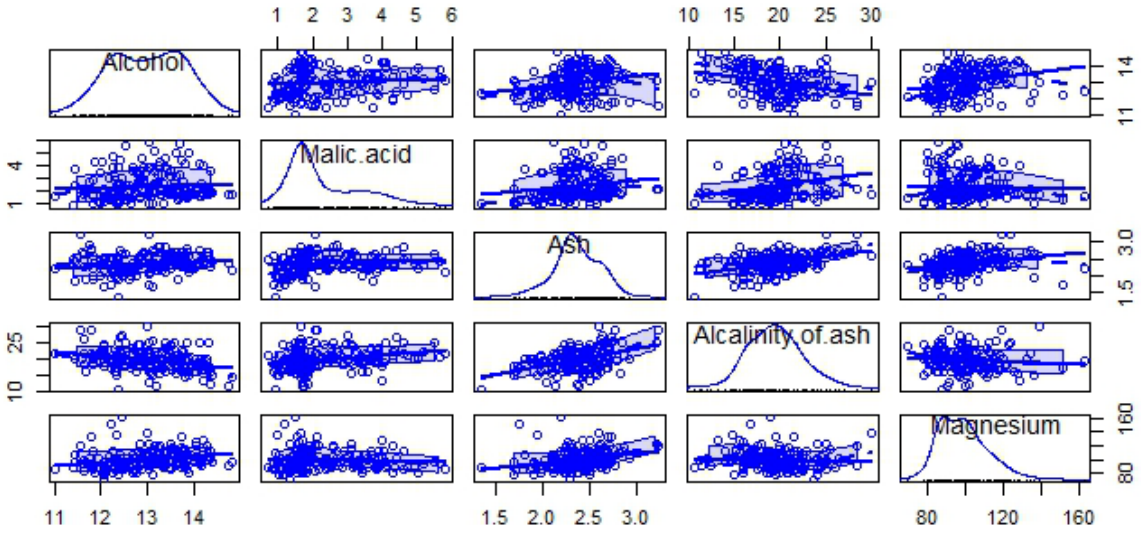
11) Hue

12) OD280/OD315 of diluted wines

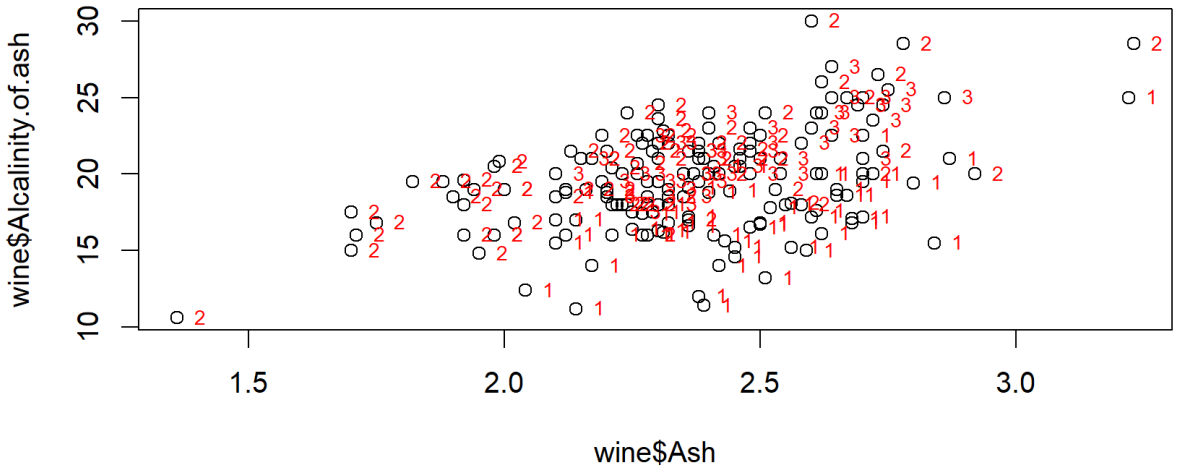
13) Proline

Goal: The objective of this task was to apply Principal Component Analysis (PCA) in a manner that reduces the problem's dimensionality while also ensuring that the new dimensions effectively differentiate among the three classes.

A preliminary look at the data



There seems to be a positive relationship between Ash and Alcalinity of ash.



Although there are discernible differences in certain dimension averages among the classes, such as Color intensity, Proline, Magnesium, and Alkalinity of Ash, there is still some level of overlap in the standard deviation. Hence, it is not possible to separate the classes based on any of these individual dimensions.

Variable	Mean Values
Alcohol	13.000618
Malic.acid	2.3363483
Ash	2.3665169
Alcalinity.of.ash	19.4949438
Magnesium	99.741573
Total.phenols	2.2951124
Flavanoids	2.0292697
Nonflavanoid.phenols	0.3618539
Proanthocyanins	1.5908989
Color.intensity	5.0580899
Hue	0.9574494
OD280.OD315.of.diluted.wines	2.6116854
Proline	746.8932584

Standard Deviation

Variable	Standard Deviation
Alcohol	0.8118265
Malic.acid	1.1171461
Ash	0.274344
Alcalinity.of.ash	3.3395638
Magnesium	14.2824835
Total.phenols	0.625851
Flavanoids	0.9988587
Nonflavanoid.phenols	0.1244533
Proanthocyanins	0.5723589
Color.intensity	2.3182859
Hue	0.2285716
OD280.OD315.of.diluted.wines	0.7099904
Proline	314.9074743

Upon further examination of the average values and standard deviation, it becomes apparent that the scales of the data points are sufficiently different, requiring the dimensions averages to be taken from them. This is a typical step in principal component

analysis. However, the varying scales of the variance also necessitate normalization of the column. The normalization is achieved by dividing each average adjusted data point by the standard deviation of the dimension's data values, which in turn sets the variance to 1. Also, there seems to be lot of correlation so PCA is a viable option for the analysis

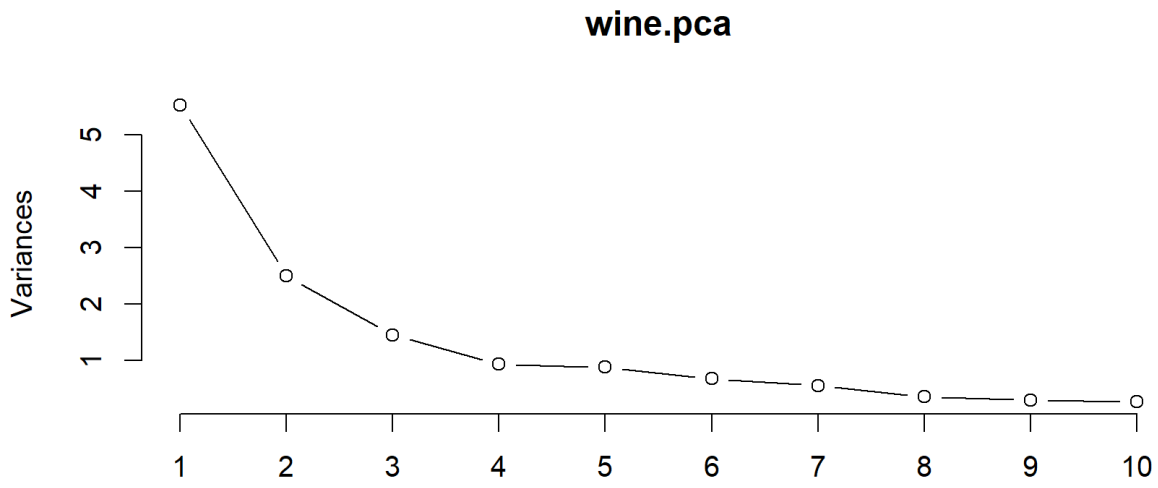
PCA output

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.3529	1.5802	1.2025	0.96328	0.93675	0.82023	0.74418	0.5916	0.54272
Proportion of Variance	0.3954	0.1784	0.1033	0.06628	0.06268	0.04806	0.03956	0.0250	0.02104
Cumulative Proportion	0.3954	0.5738	0.6771	0.74336	0.80604	0.85409	0.89365	0.9186	0.93969

	PC10	PC11	PC12	PC13	PC14
Standard deviation	0.51216	0.47524	0.41085	0.35995	0.24044
Proportion of Variance	0.01874	0.01613	0.01206	0.00925	0.00413
Cumulative Proportion	0.95843	0.97456	0.98662	0.99587	1.00000

To decide how many principal component should be retained we will use the screeplot



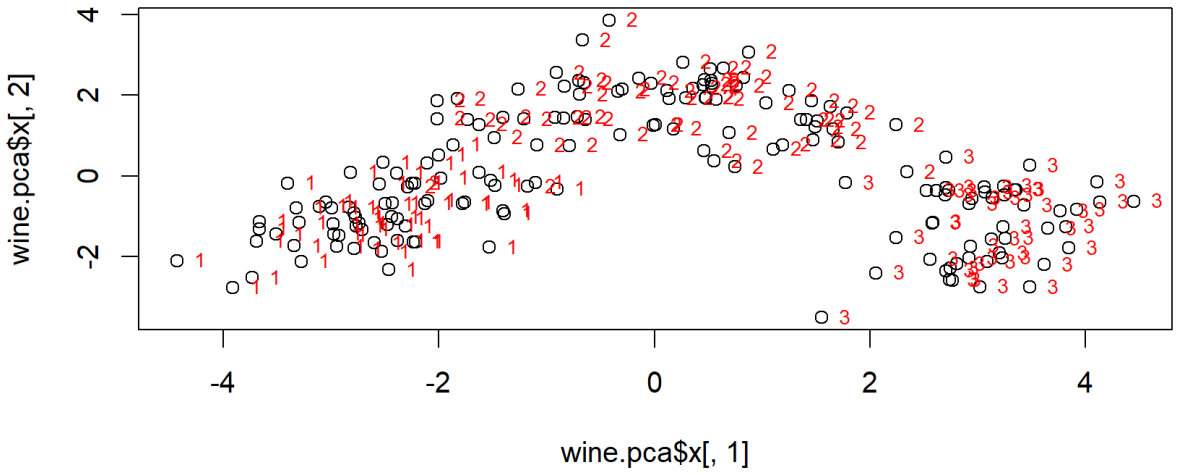
To use the scree plot to decide how many principal components to retain, we should look for the "elbow point" in the plot, which is the point where the slope of the curve changes dramatically. The idea is to retain enough principal components to explain a significant portion of the variation in the data while avoiding overfitting the model. By looking at the plot, we will choose first 3 principal components.

Now: wine.pca$rotation[,1:3] loadings of the first three principal components for each of the original variables in the wine dataset. The loadings represent the correlation coefficients between the variables and the principal components, and can be used to interpret the meaning of each principal component. The output will be a matrix with the same number of rows as the original dataset (i.e., 13 variables) and three columns, corresponding to the first three principal components.

	PC1	PC2	PC3
Class	0.393669533	-0.005690412	0.001217953
Alcohol	-0.136325011	-0.484160868	-0.207400812
Malic.acid	0.222676383	-0.223590947	0.088796064
Ash	-0.002257932	-0.315855884	0.626102363
Alcalinity.of.ash	0.224298489	0.011615737	0.611989600
Magnesium	-0.124630159	-0.300551432	0.130984580
Total.phenols	-0.359264042	-0.067119829	0.146507749
Flavanoids	-0.390711715	0.001313454	0.150962746
Nonflavanoid.phenols	0.267001203	-0.026988703	0.169975512
Proanthocyanins	-0.279062504	-0.041222563	0.149879586
Color.intensity	0.089318293	-0.529782740	-0.137266298
Hue	-0.276822650	0.277907354	0.085328539
OD280.OD315.of.diluted.wines	-0.350526181	0.162776250	0.166204360
Proline	-0.269515252	-0.366058862	-0.126686846

After analyzing the data, it was found that the first principal component has the highest absolute loadings for variables V1 through V14. Specifically, variables V1, V7, V8, V13, V10, V9, V12, V14, V3, and V5 have loadings with magnitudes ranging from 0.222 to 0.390. Among these variables, V8, V7, V13, V10, V12, and V14 have negative loadings, while V9, V3, and V5 have positive loadings. This indicates that the first principal component reflects a contrast between the concentrations of V8, V7, V13, V10, V12, and V14, and the concentrations of V9, V3, and V5..

Scatterplots of the Principal Components



The scatterplot displays the first principal component on the x-axis and the second principal component on the y-axis. It reveals that the wine samples of cultivar 1 have significantly lower values of the first principal component than the wine samples of cultivar 3, indicating that the first principal component can distinguish wine samples of cultivars 1 from those of cultivar 3. Furthermore, the scatterplot demonstrates that the wine samples of cultivar 2 have considerably higher values of the second principal component than the wine samples of cultivars 1 and 3, suggesting that the second principal component can separate the samples of cultivar 2 from those of cultivars 1 and 3. Hence, the first two principal components are reasonably useful in differentiating the wine samples of the three different cultivars.