Liang, Xiyu 101086285 STAT5703 Assignment#2
Output of the code and summary of my findings

1. Data Splitting

We want to divide the dataset into 2/3 for training and 1/3 for testing. Since the wine data contains 178 samples and 14 attributes, so we can divide the original complete 178 samples into a training set which contains 119 samples and a test set which has 59 samples randomly.

When we try to split our data, it is important to maintain the distribution of every attribute the same in both training and test set to get the most realistic performance. In our case the most important attribute is the cultivar of grapes, so we just need to make sure that the distribution of the three cultivars is roughly the same in the training and test set.

We use a bar plot (Figure 1) to display the distribution of three cultivars (classes) in the full data, training set and the test set. It shows the proportion of samples of given cultivar. We run the code several times until we get a split which satisfies that in each given cultivar, the percentage of samples in the training set is roughly the same as that of in the test set.

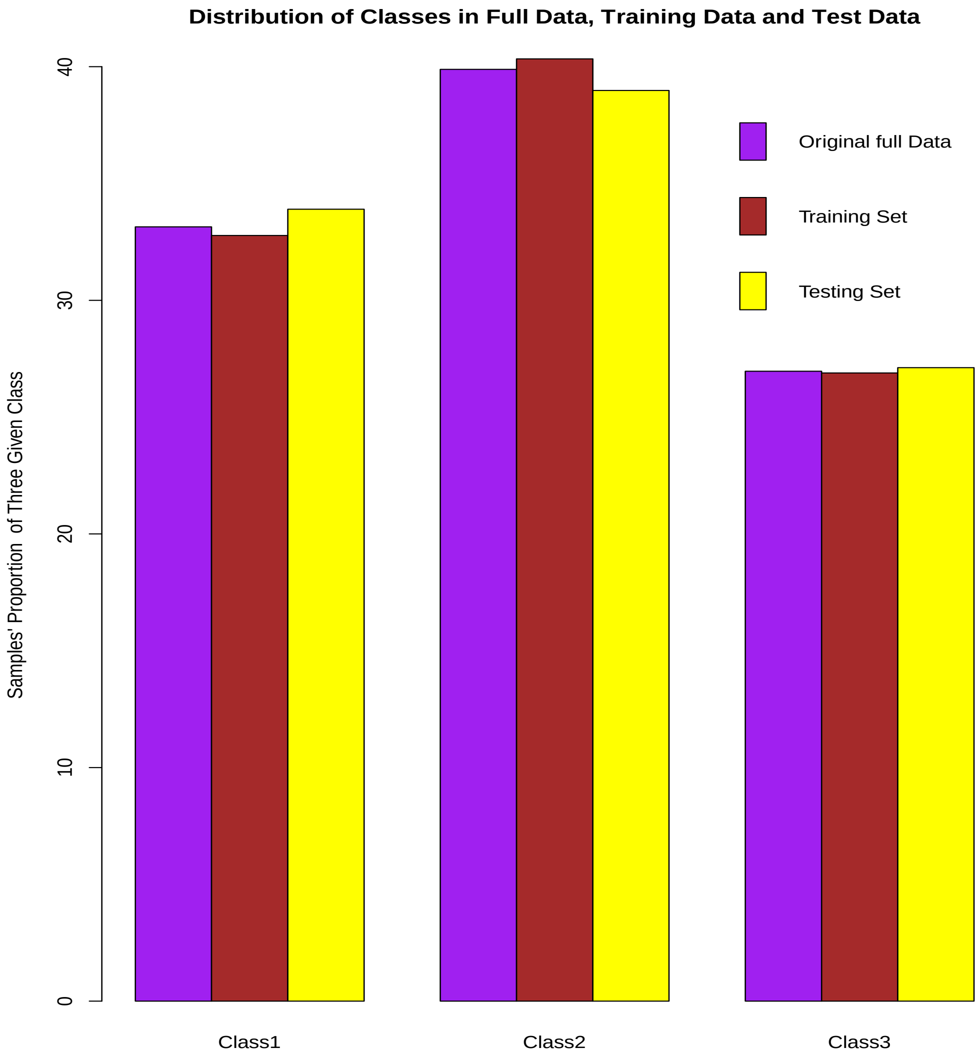


Figure 1.

We can see that the training set contains a little bit more samples which belong to the Class2 and the test set is slightly weighted towards Class1 and Class3. Since the size of our dataset is not that big and we do not need each class be splitting perfectly, therefore we think this split is acceptable and adequate.

2. Kmeans clustering

We can use R to implement a k-means clustering for our wine dataset. We know that there are three cultivars in total, so we can let k=3.

The k-means algorithm result of clustering highly depends on the selection of initial centroids. And since we want to select initial centroids randomly which means that with some inappropriate initial centroids, we may get bad clustering results. To deal with this problem, we can run a loop to try clustering with different random seeds then we select the seed that gives the most accurate result. Another possible issue is that we may get empty clusters. There are several strategies to handle this issue, we can make that centroid the point that contributes most to SSE and also, we can make that centroid the point from the cluster with the highest SSE. In our case I decide to use the second method.

Instead of using the kmeans function in R, we can use R to select three initial centroids randomly first, then use the Euclidean distance function to calculate the distance between points and each centroid. After that we can assign all samples to the closest centroid and then calculate the mean of each attributes of all samples in each cluster. In this way we can get three new centroids, then recalculate the distance between each data point and new obtained cluster centers, reassigning all samples to the closest centroid. We repeat our clustering until the centroids do not change anymore.

In order to measure the quality of our clustering we can use the SSE and the counts of cases ending up in the wrong cluster.

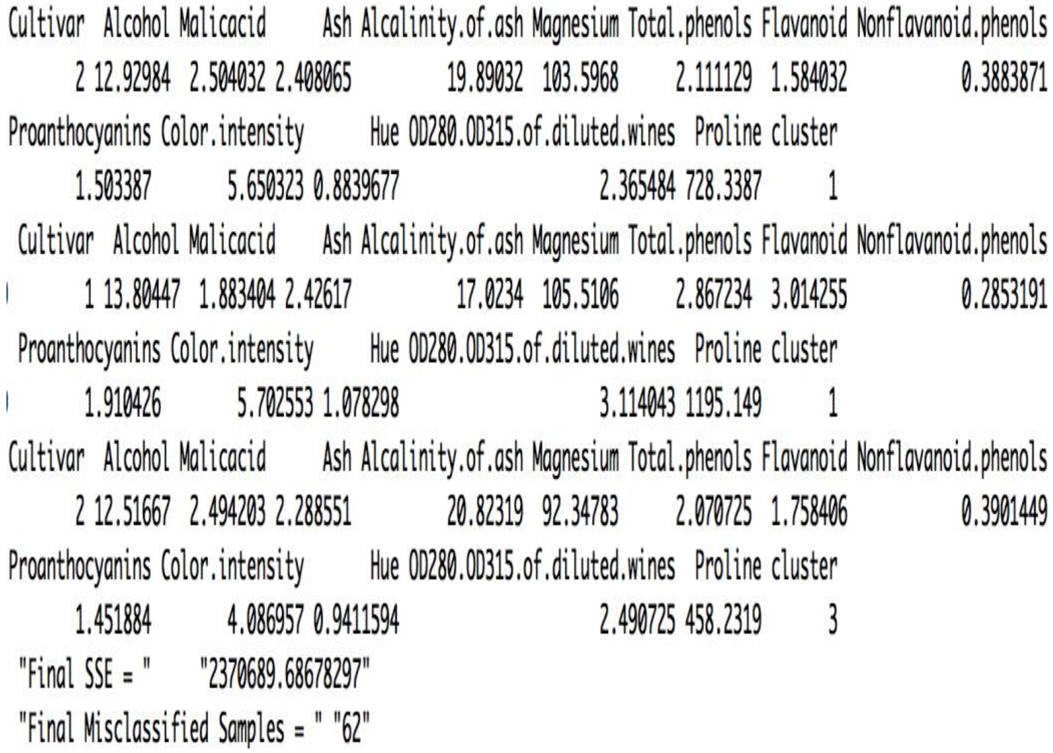
We do the analysis for the raw data, the standardized data and the whitened data respectively.

(1) Raw data

We use the raw data to do the analysis first.

We run the loop 50 times, store the final SSE and total error classes in the results matrix, then we can find the best clustering result with an SSE of 2370689.68, and 62 misclassified cases. Considering the accuracy, we note that the clustering results we got by using raw data is not very good.

We select the third seed, i.e set.seed(3). The location of centroids for the best clustering can be found in Figure 2.



```
Cultivar  Alcohol Malicacid      Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       2 12.92984  2.504032 2.408065          19.89032  103.5968      2.111129  1.584032            0.3883871
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline cluster
       1.503387        5.650323 0.8839677                     2.365484 728.3387       1
 Cultivar  Alcohol Malicacid      Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        1 13.80447  1.883404 2.42617           17.0234  105.5106      2.867234  3.014255            0.2853191
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline cluster
        1.910426        5.702553 1.078298                     3.114043 1195.149       1
Cultivar  Alcohol Malicacid      Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       2 12.51667  2.494203 2.288551          20.82319  92.34783      2.070725  1.758406            0.3901449
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline cluster
       1.451884        4.086957 0.9411594                     2.490725 458.2319       3
 "Final SSE = "     "2370689.68678297"
 "Final Misclassified Samples = " "62"
```

Figure 2.

(2) Standardized data

Clearly, it is very important that data, from different numeric variables should express in the same scale. And we know that when distances are calculated, if the units are very different, we cannot get proper results. Therefore, we need to standardize our data first.

I have read some reports and thesis talking about the data pre-processing methods before running the k-means algorithm. I realized that there are two main techniques for standardization:

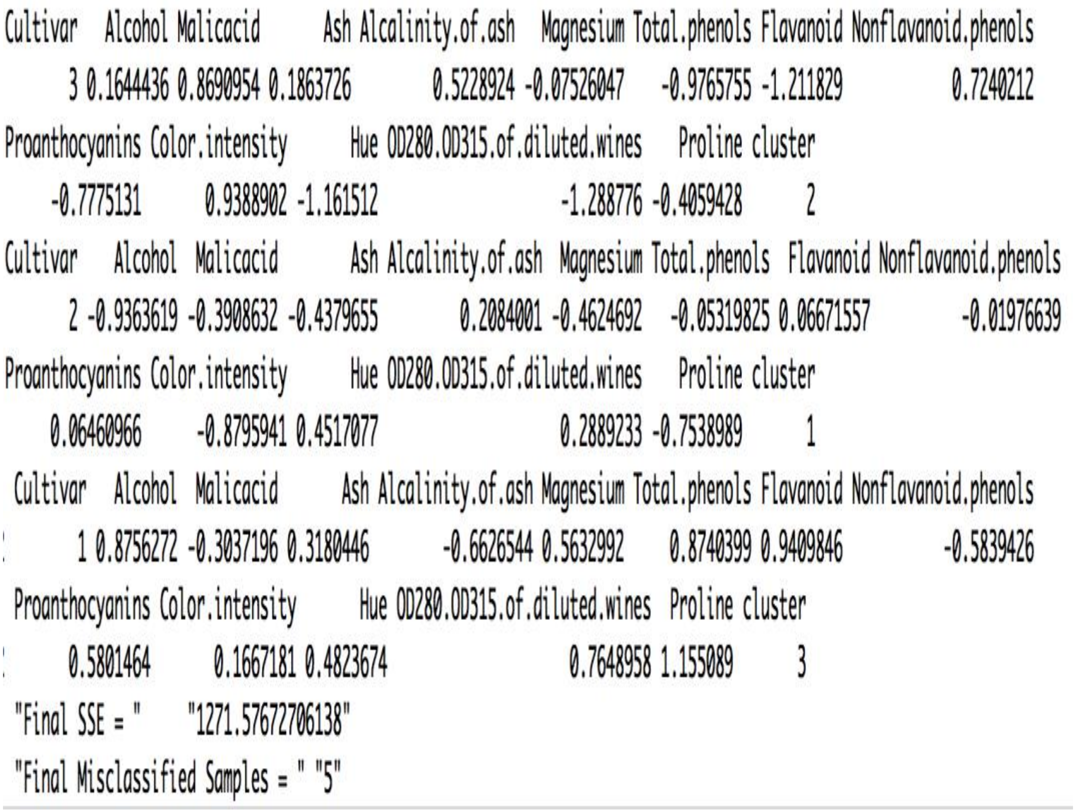
(a) Min-Max scaling: in this approach, the data is scaled to a fixed range -usually 0 and 1.

(b) Z-score standardization: the mean value of the variable is subtracted from each value and each one is then divided by the standard deviation.

It is hard to say which transformation is better since it really depends on the application. But in clustering analysis, the z-score standardization will be better because we need to compare similarities between features based on certain distance measures. Also, in PCA we usually prefer the z-score transformation. Based on the above reasons and considering that in the third part we need to use PCA to do clustering then compare the results with those found by using the actual data. So, I decided to use the z-score standardization, the resulting variable will have a mean of 0 and a standard deviation of 1.

After running the loop 50 times, we can see some repeated results. 11 seeds give us the best clustering result: the SSE is 1271.57 and only 5 misclassified samples. Therefore, we can choose any one of them as the seed that gives the most accurate result. Here we pick i=1.

The location of centroids for the best clustering can be found in Figure 3.



```
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash   Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       3 0.1644436 0.8690954 0.1863726         0.5228924 -0.07526047    -0.9765755 -1.211829            0.7240212
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline cluster
     -0.7775131       0.9388902 -1.161512                    -1.288776 -0.4059428       2
Cultivar    Alcohol  Malicacid        Ash Alcalinity.of.ash  Magnesium Total.phenols  Flavanoid Nonflavanoid.phenols
       2 -0.9363619 -0.3908632 -0.4379655         0.2084001 -0.4624692   -0.05319825 0.06671557          -0.01976639
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline cluster
     0.06460966      -0.8795941 0.4517077                    0.2889233 -0.7538989       1
 Cultivar   Alcohol  Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        1 0.8756272 -0.3037196 0.3180446        -0.6626544 0.5632992     0.8740399 0.9409846           -0.5839426
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline cluster
       0.5801464       0.1667181 0.4823674                    0.7648958 1.155089       3
 "Final SSE = "     "1271.57672706138"
 "Final Misclassified Samples = " "5"
```

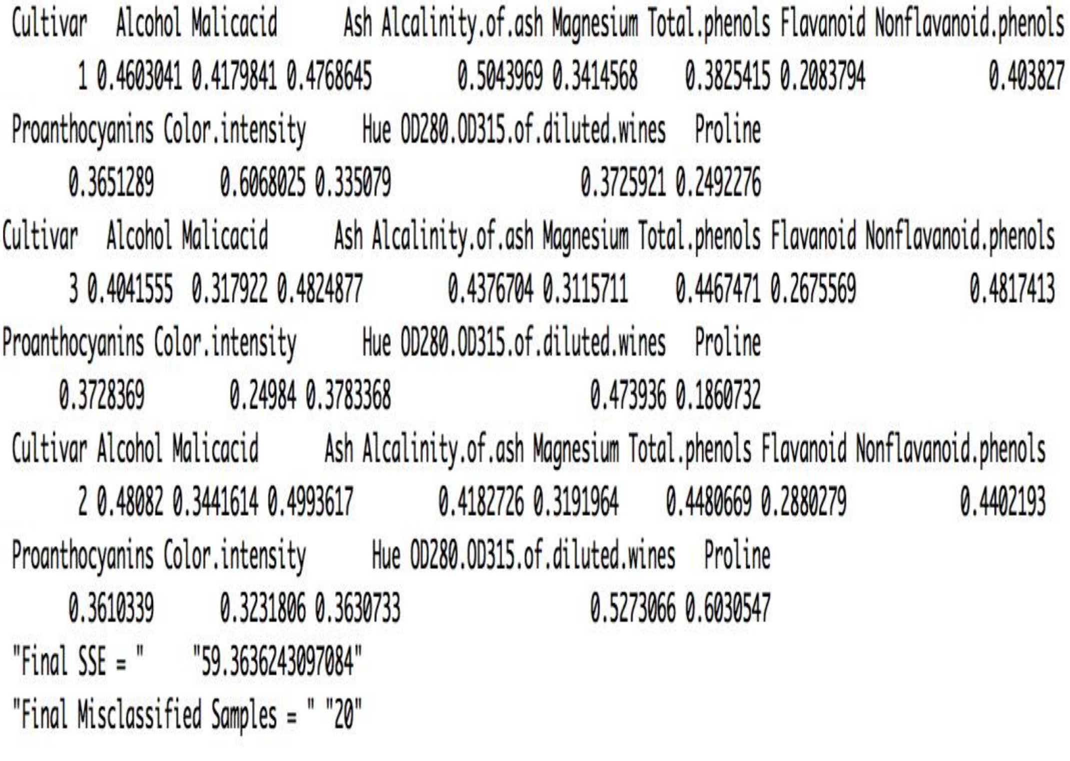
Figure 3.

(3) Whitened data

After whitening, before running the loop I also did a Min-Max transformation for the data in order to let the range of values for each variable are mapped to the range 0 to 1. By doing this, then we can use the runif function to generate initial centroids. Note that this transformation may generate smaller SSE.

We run the loop 50 times then the best clustering results (when i=15) has an SSE of 59.363, and 20 misclassified cases in total.

The location of centroids for the best clustering can be found in Figure 4.



```
 Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        1 0.4603041 0.4179841 0.4768645         0.5043969 0.3414568     0.3825415 0.2083794             0.403827
 Proanthocyanins Color.intensity      Hue OD280.OD315.of.diluted.wines   Proline
       0.3651289       0.6068025 0.335079                    0.3725921 0.2492276
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       3 0.4041555  0.317922 0.4824877         0.4376704 0.3115711     0.4467471 0.2675569            0.4817413
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
      0.3728369         0.24984 0.3783368                     0.473936 0.1860732
 Cultivar Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        2 0.48082 0.3441614 0.4993617         0.4182726 0.3191964     0.4480669 0.2880279            0.4402193
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
       0.3610339       0.3231806 0.3630733                    0.5273066 0.6030547
 "Final SSE = "     "59.3636243097084"
 "Final Misclassified Samples = " "20"
```

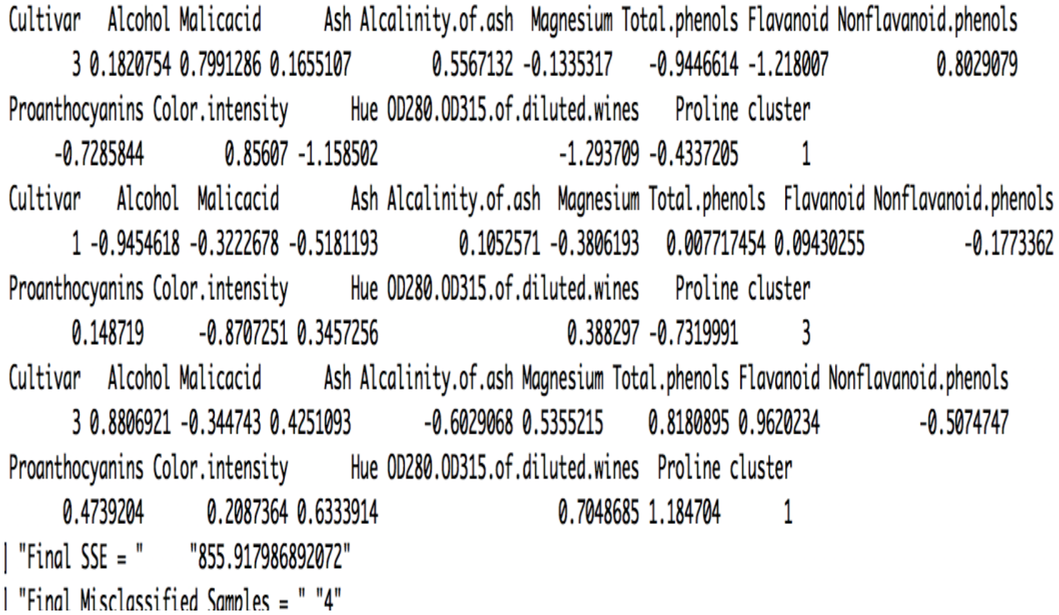
Figure 4.

In Question 1, we already did the data splitting, so now we can examine the clustering results for both the training set and the test set. We examine our training set first then we can use the centroids from our run of the clustering with the training set to do clustering for the test set. By doing this, we can evaluate our model.

Standardization is necessary for our dataset. And in Question 2 we have done analysis for raw data, standardized data and whitened data, it is clear that if we use standardized data, we can get better clustering results. Therefore, I believe we can end up sticking to standardized data most of the time, so I just used standardized data to examine and evaluate our model.

（1） Training set

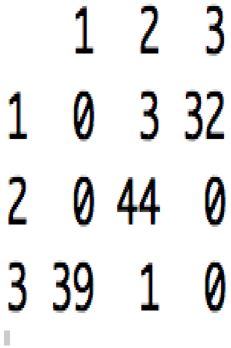
After 50 times runs, from the misclassified results (the results matrix in the R output) we know that our best clustering has an SSE of 855.9179 and 4 misclassified cases, which means the clustering accuracy is 96.6%. The location of centroids is displayed in Figure 5.



```
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash  Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       3 0.1820754 0.7991286 0.1655107         0.5567132 -0.1335317    -0.9446614 -1.218007            0.8029079
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline cluster
     -0.7285844         0.85607 -1.158502                    -1.293709 -0.4337205       1
Cultivar    Alcohol  Malicacid        Ash Alcalinity.of.ash  Magnesium Total.phenols  Flavanoid Nonflavanoid.phenols
       1 -0.9454618 -0.3222678 -0.5181193         0.1052571 -0.3806193   0.007717454 0.09430255           -0.1773362
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline cluster
       0.148719      -0.8707251 0.3457256                     0.388297 -0.7319991       3
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       3 0.8806921 -0.344743 0.4251093        -0.6029068 0.5355215     0.8180895 0.9620234           -0.5074747
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline cluster
      0.4739204       0.2087364 0.6333914                    0.7048685 1.184704       1
| "Final SSE = "     "855.917986892072"
| "Final Misclassified Samples = " "4"
```

Figure 5.

The table below displays the misclassification/classification results. Note that the cluster and cultivar numbers do not match up, but this does not matter since it is just an arbitrary “naming” of each cluster and it will not influence our clustering results.



```
   1  2  3
1  0  3 32
2  0 44  0
3 39  1  0
```

Figure 6.

As we can see, the data belonging to the Cultivar1 got grouped into cluster3, Cultivar2 into cluster2, and Cultivar3 into cluster1. The algorithm wrongly classified four data points belonging to Cultivar2.

The next plot visualizes the distribution of our clusters, colored by true class. It shows the distance of samples from their closest cluster centroid and we can see misclassified samples clearly. The horizontal axis displays which cluster the sample was assigned to and the vertical axis represents the Euclidean distance from centroid.

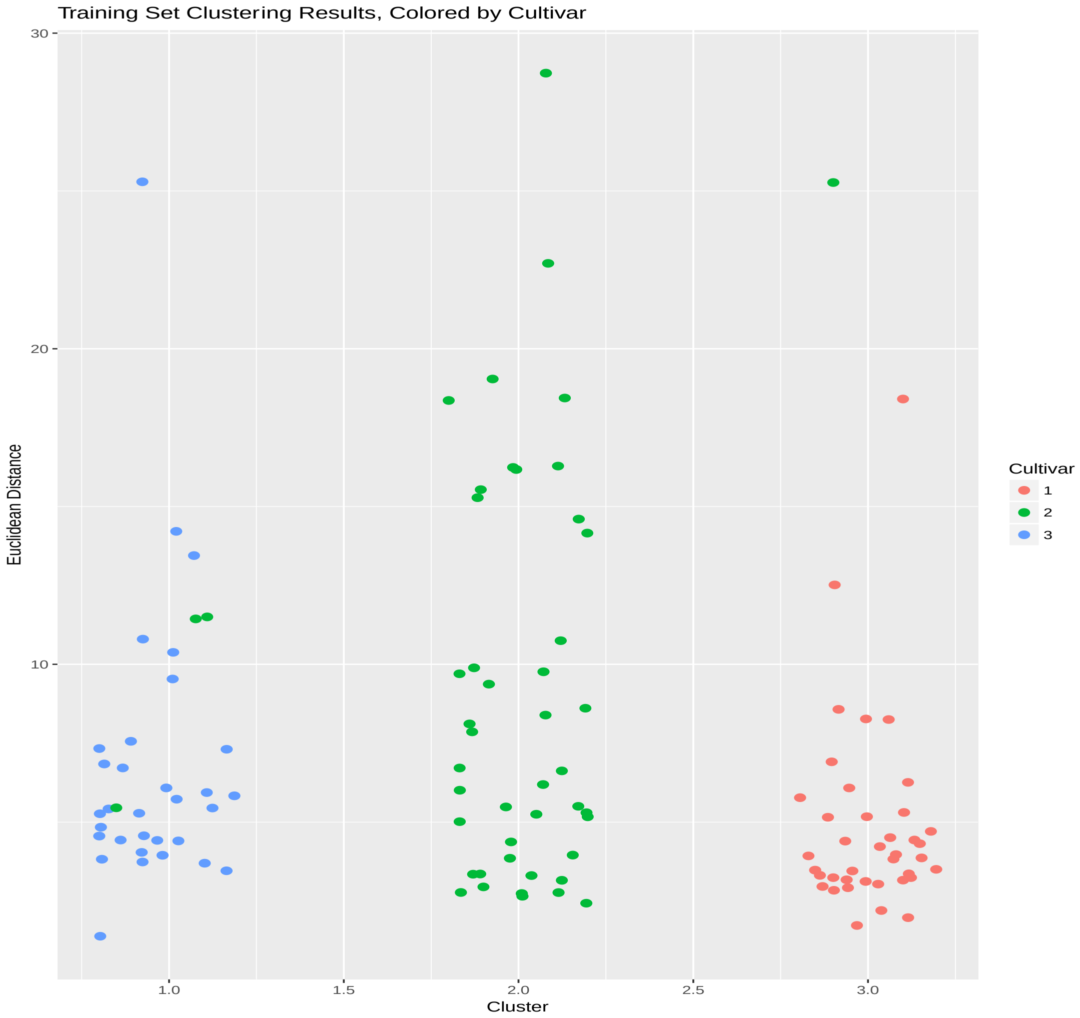


Figure 7.

We can see that there are four misclassified samples in total, they all come from Cultivar2 but ended up in the wrong cluster.

We can also see that compared with cluster2, cluster1 and 3 are more tightly.

（2） Test set

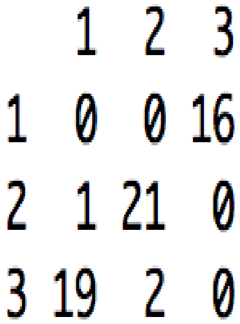
Now we can evaluate our model with the data that was not used to infer the model.

Consider the scaling process as part of the model generated by the training data, and we want to use the test data to test both the generality of the model combined with the pre-processing. Thus, we need to standardize the test data with the mean and standard deviation from each variable in the training data.

Then we can calculate the distances between each data point and three centroids we got by running of the clustering with the training set, find out the closest centroid to each of our test samples then assign the test sample to that cluster. Finally, we will get the total SSE and we will know how many samples were classified correctly.

Out of our 59 test samples, 56 of them were classified correctly.

The misclassification/classification table is shown as below:



```
   1  2  3
1  0  0 16
2  1 21  0
3 19  2  0
```

Figure 8.

It shows that two of the misclassification samples should be in cluster2, while it was assigned to cluster3, and one sample should be in cluster3 but ended up in cluster2.

Then we create a similar plot that we made with our training data, but with the test samples.

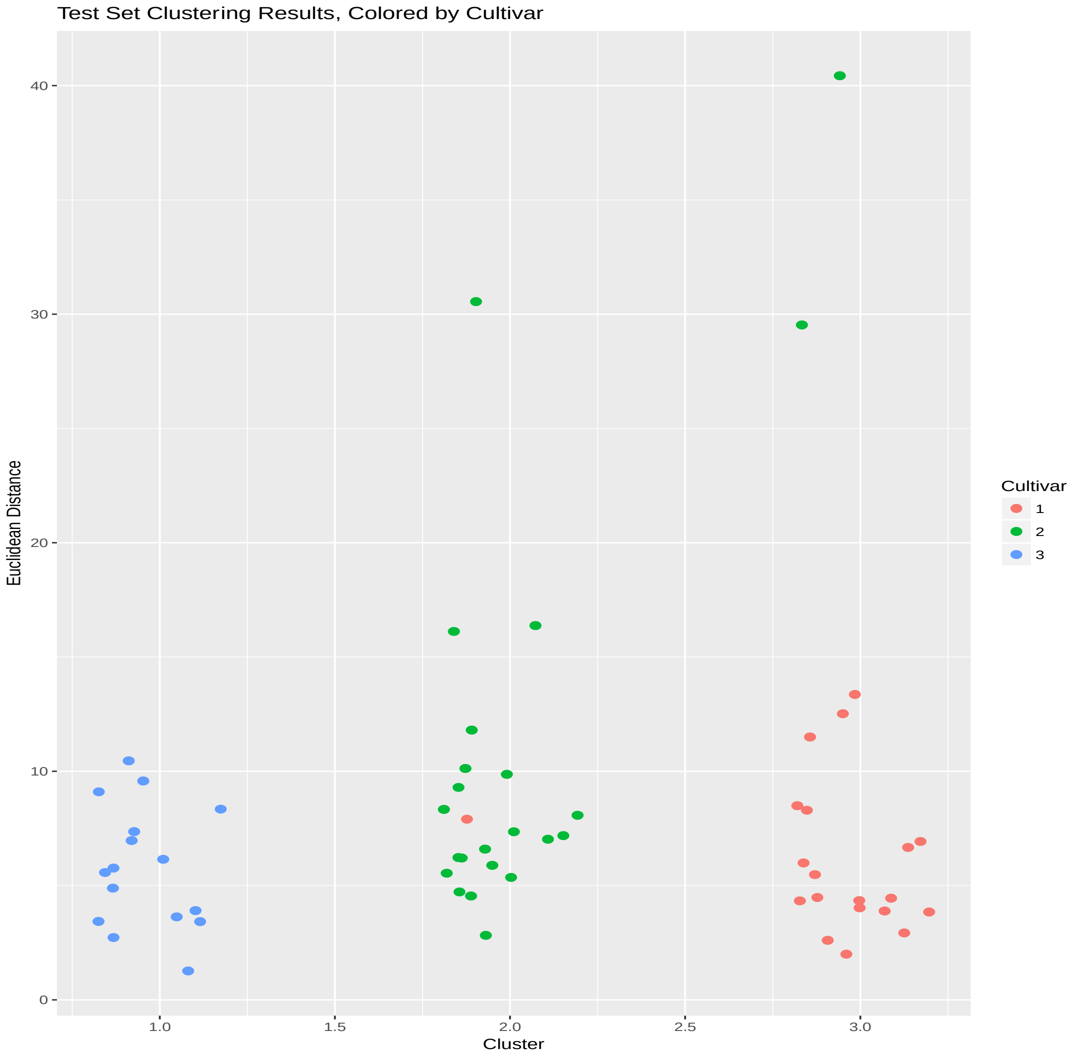


Figure 9.

We can see that if we use the 1/3 of the original data set we left out to evaluate our model, then the accuracy of clustering is almost 94.9%, the misclassified sample is not all from the Cultivar2. Compared with the clustering results we got by using the training data, the effect of the clustering of test samples is not perfect but still good.

To determine if these patterns are truly present in the data, we can run the analysis several times with different train/test splits of the data, see if we will find the same issue with Cultivar1 and Cultivar2.

By comparing this results with previous clustering results, I would like to say that this clustering produces a reasonably good model for determining the cultivar of grapes used to make a wine. And the clustering results of our model proves that wine chemistries can be used to help us to find out the grape that went into the wine.

Therefore, I believe we have enough confidence to say that our model can be used to do a chemical analysis of a wine.

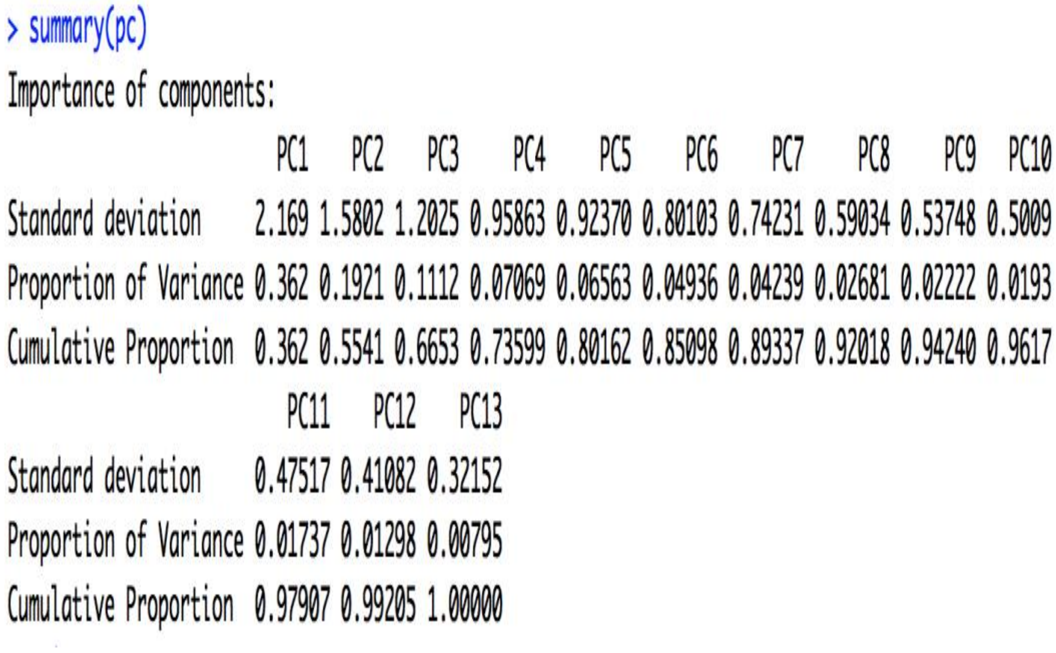
3. PCA

PCA is the most used dimensionality reduction algorithm and it reduces the p dimensions of our data set down to k principal components (PCs). PCs are ordered by the decreasing amount of variance explained. To reduce the dimension, we may drop the latter PCs which explain less of variance.

In our dataset, we have 13 attributes and we want to reduce the number of variables to two or three interpretable linear combinations of the data, each linear combination will correspond to a principal component.

PCA will perform bad if we use the original data, because all variables in our example are not measured in the same scale. So, we need to standardize our data first. Then we can use the prcomp function from the stats package to do the PCA.

If the normalization is applied on the data:

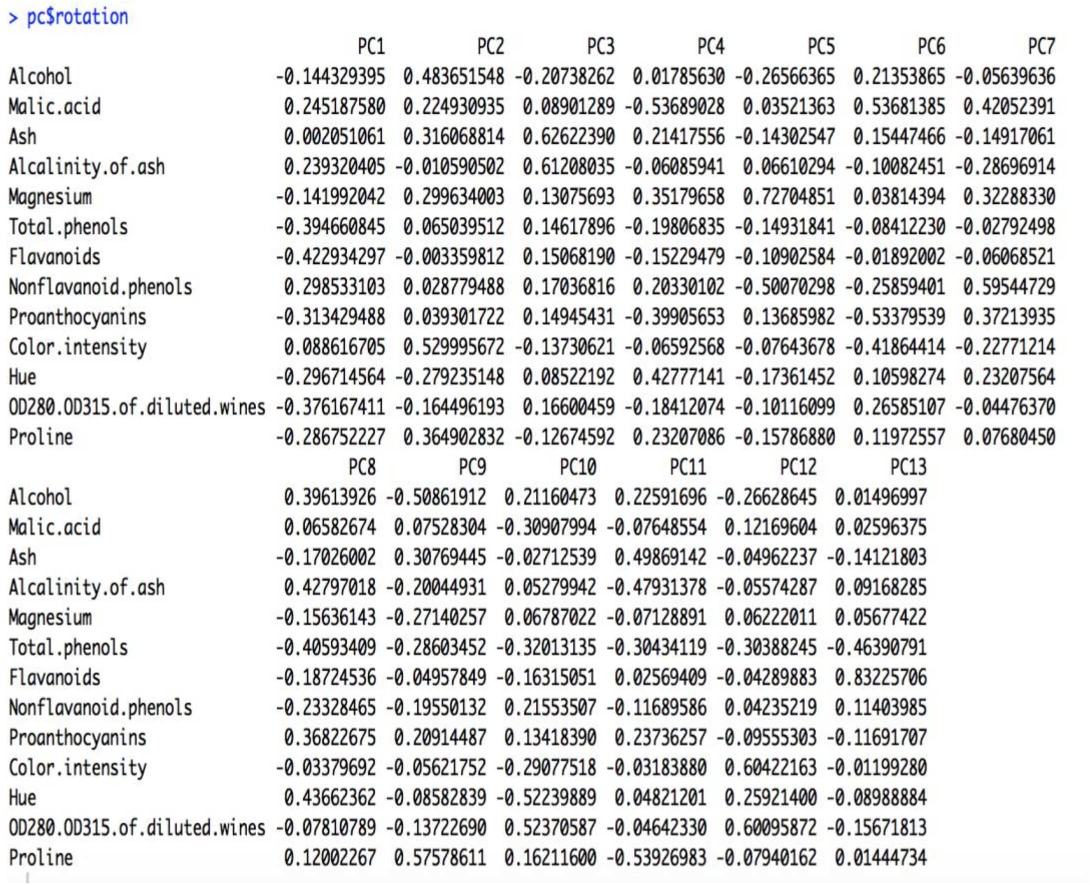


```
> summary(pc)
Importance of components:
                         PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9   PC10
Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231 0.59034 0.53748 0.5009
Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239 0.02681 0.02222 0.0193
Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337 0.92018 0.94240 0.9617
                          PC11    PC12    PC13
Standard deviation     0.47517 0.41082 0.32152
Proportion of Variance 0.01737 0.01298 0.00795
Cumulative Proportion  0.97907 0.99205 1.00000
```

Figure 10.

The Figure 10 shows the standard deviation for each PCs and how much variance they explain respectively.

We can see that the first principal component is responsible for only 36.2% of the total variance, so if we want to rise above 90% variance, we have to take 8 components.



```
> pc$rotation
                                     PC1          PC2         PC3         PC4         PC5         PC6         PC7
Alcohol                     -0.144329395  0.483651548 -0.20738262  0.01785630 -0.26566365  0.21353865 -0.05639636
Malic.acid                   0.245187580  0.224930935  0.08901289 -0.53689028  0.03521363  0.53681385  0.42052391
Ash                          0.002051061  0.316068814  0.62622390  0.21417556 -0.14302547  0.15447466 -0.14917061
Alcalinity.of.ash            0.239320405 -0.010590502  0.61208035 -0.06085941  0.06610294 -0.10082451 -0.28696914
Magnesium                   -0.141992042  0.299634003  0.13075693  0.35179658  0.72704851  0.03814394  0.32288330
Total.phenols               -0.394660845  0.065039512  0.14617896 -0.19806835 -0.14931841 -0.08412230 -0.02792498
Flavanoids                  -0.422934297 -0.003359812  0.15068190 -0.15229479 -0.10902584 -0.01892002 -0.06068521
Nonflavanoid.phenols         0.298533103  0.028779488  0.17036816  0.20330102 -0.50070298 -0.25859401  0.59544729
Proanthocyanins             -0.313429488  0.039301722  0.14945431 -0.39905653  0.13685982 -0.53379539  0.37213935
Color.intensity              0.088616705  0.529995672 -0.13730621 -0.06592568 -0.07643678 -0.41864414 -0.22771214
Hue                         -0.296714564 -0.279235148  0.08522192  0.42777141 -0.17361452  0.10598274  0.23207564
OD280.OD315.of.diluted.wines -0.376167411 -0.164496193  0.16600459 -0.18412074 -0.10116099  0.26585107 -0.04476370
Proline                     -0.286752227  0.364902832 -0.12674592  0.23207086 -0.15786880  0.11972557  0.07680450
                                    PC8         PC9        PC10        PC11        PC12        PC13
Alcohol                      0.39613926 -0.50861912  0.21160473  0.22591696 -0.26628645  0.01496997
Malic.acid                   0.06582674  0.07528304 -0.30907994 -0.07648554  0.12169604  0.02596375
Ash                         -0.17026002  0.30769445 -0.02712539  0.49869142 -0.04962237 -0.14121803
Alcalinity.of.ash            0.42797018 -0.20044931  0.05279942 -0.47931378 -0.05574287  0.09168285
Magnesium                   -0.15636143 -0.27140257  0.06787022 -0.07128891  0.06222011  0.05677422
Total.phenols               -0.40593409 -0.28603452 -0.32013135 -0.30434119 -0.30388245 -0.46390791
Flavanoids                  -0.18724536 -0.04957849 -0.16315051  0.02569409 -0.04289883  0.83225706
Nonflavanoid.phenols        -0.23328465 -0.19550132  0.21553507 -0.11689586  0.04235219  0.11403985
Proanthocyanins              0.36822675  0.20914487  0.13418390  0.23736257 -0.09555303 -0.11691707
Color.intensity             -0.03379692 -0.05621752 -0.29077518 -0.03183880  0.60422163 -0.01199280
Hue                          0.43662362 -0.08582839 -0.52239889  0.04821201  0.25921400 -0.08988884
OD280.OD315.of.diluted.wines -0.07810789 -0.13722690  0.52370587 -0.04642330  0.60095872 -0.15671813
Proline                      0.12002267  0.57578611  0.16211600 -0.53926983 -0.07940162  0.01444734
```

Figure 11.

As shown in Figure 11, we note that the weights apply to variables are different in each PC. For instance, the first principal component (PC1) is dominated by “Flavanoids” while the second (PC2) is dominated by “Color Intensity”. And weights tell us which variables are important for these PCs.

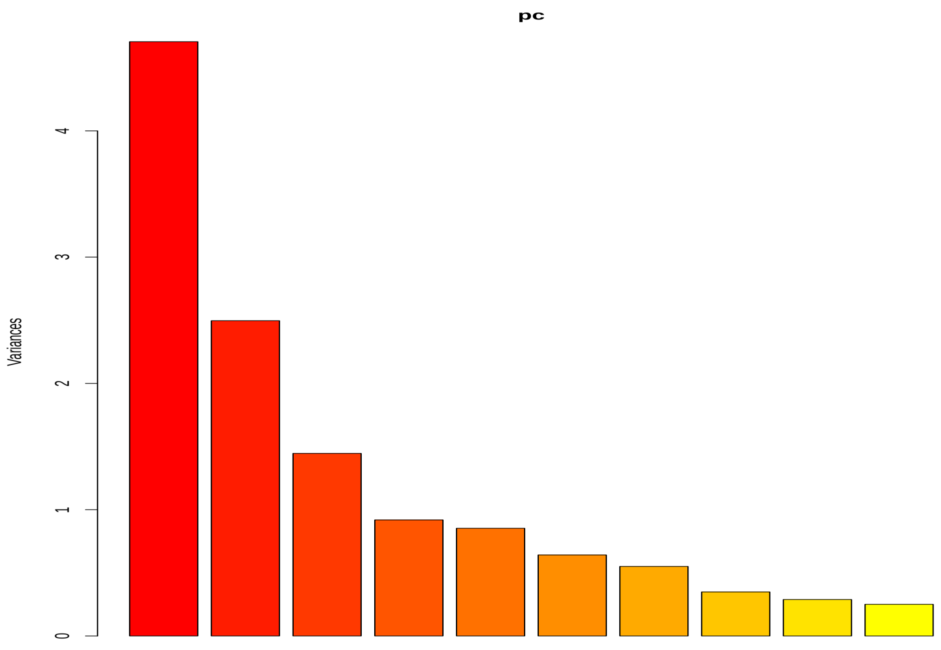


Figure 12.

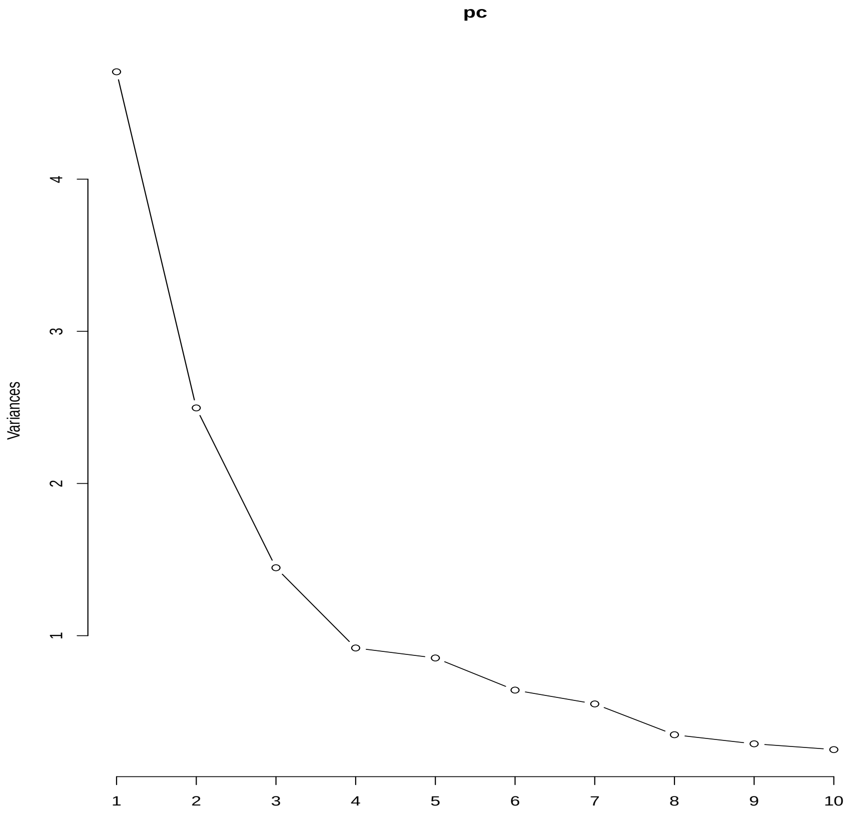
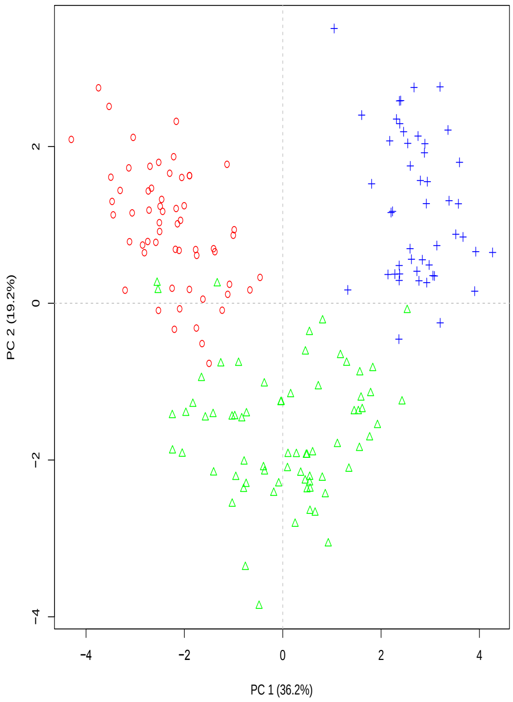


Figure 13.

The scree plots above also give an idea of the relative importance of the principal components since it plots how much variance each PCs explain.

Again, what this tells us is that the first two PCs account for over 55% of the variance in the entire dataset. While it would be difficult to justify the entire analysis based on 55% available information. So, it may be better to use the first three PCs to do clustering on our wine data. But I think it is interesting to see that we can account for 55% of the information with only 15% of the data. So, let us use the first two principal components to visualize the dataset first, then use the first three PCs to do clustering and see if we can get a better clustering results.

(1) Use the first two PCs to do clustering.



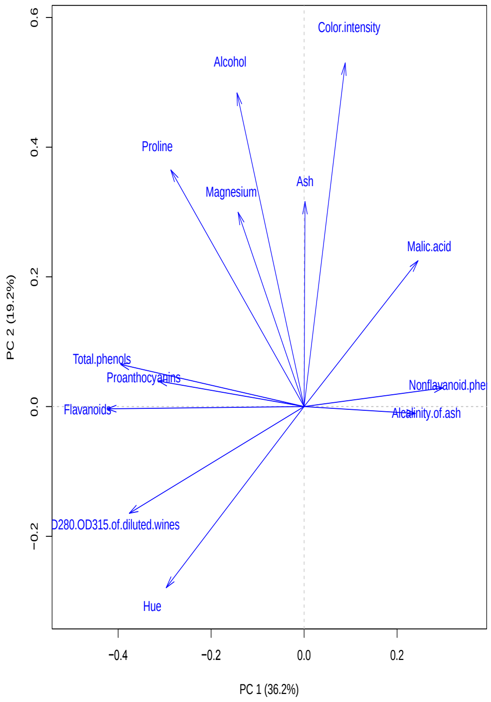


Figure 14.

Figure 15.

This is a PCA of just attributes so the pattern from the high-dimensional space of the dataset translate well to two dimensions. The color red represents Cultivar1, green for Cultivar 2 and blue for Cultivar3.

The scatterplot (Figure 14) shows a clear separation among cultivars. We can see that wine samples of Cultivar1 have much lower values of the first principal component than wine samples of Cultivar3. Therefore, the first PC separates wine samples of Cultivar1 from those of Cultivar3.

We can also see that wine samples of Cultivar2 have much lower values of the second PC than wine samples of Cultivar1 and3. So, the second principal component separates samples of Cultivar2 from samples of Cultivar1 and 3.

In addition, the data points are evenly scattered over relatively narrow ranges in both PC1 and PC2.

Therefore, the first two PCs are reasonably useful for distinguishing wine samples of the three different cultivars.

And it is not hard to see that the Cultivar2 is sparsest, at the same time, we can see that there are several attributes from Cultivar2 overlaps into the space of Cultivar1 and Cultivar3, which indicates that Cultivar1 and Cultivar3 are much better defined than Cultivar2.

In our case, the three cultivars have different characteristics. What characteristics these are can be seen in the loading plot, shown on Figure 15. The vectors show the relationship between the original variables and the principal components. The length of the vector represents the strength of the correlation between the original variable and the principal components.

By looking at these two figures together, we can claim that:

(1) Wine samples from Cultivar2 has a lighter color intensity, lower values of alcohol, compared to the wine samples from Cultivar1 and Cultivar3.

(2) Wine samples from Cultivar3 is high in non-flavanoid phenols as well as the alkalinity of ash, at the same time with a higher content of malic acid. The reverse is true for the wine samples from Cultivar1.

Now recall that when we use the actual data for clustering, the accuracy is 97%. If we use the first two principal components to do clustering, four samples of Cultivar2 overlaps to the space of Cultivar1 and Cultivar3, and one sample from Cultivar1 overlaps.

(2) Use the first three PCs to do clustering.

The first three PCs explain almost 67% variance. We can use rgl in R to generate a 3D plot, which will show the clustering results clearly.

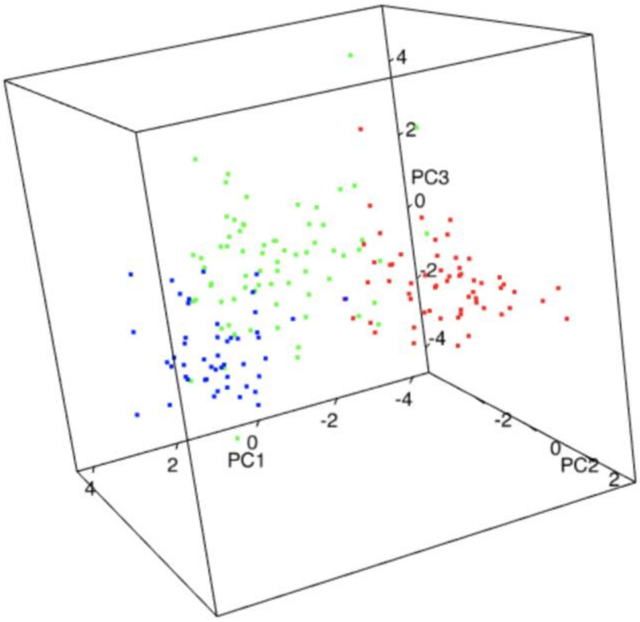


Figure 16.

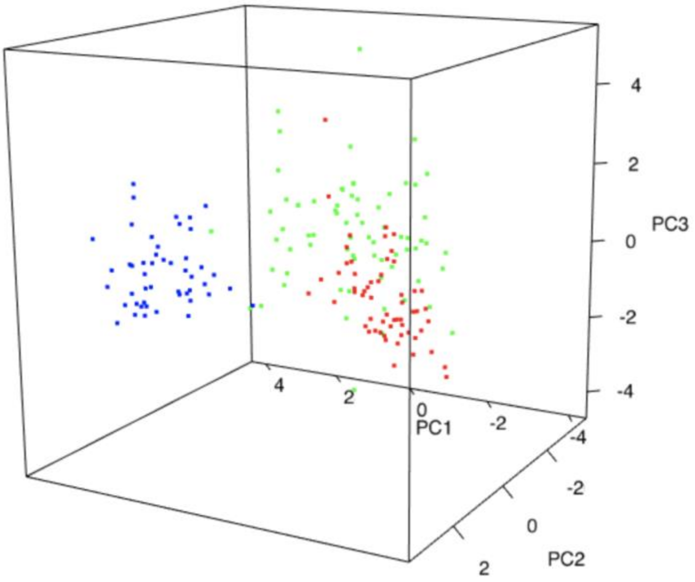


Figure 17.

With regards to separating the three classes, the 3D representation of PC1, PC2 and PC3 seems to be the most useful for us.

Clearly, PCA does not provide a higher accuracy of clustering, and by using the actual data to do clustering, we can differentiate the classes better.

However, PCA does allow us to see some general characteristics of the data. And by using the principal components to perform clustering, we can get a better 2D and 3D view of our data clustering results than we did from trying to imagine it in a 13-dimensional space. Therefore, we can treat PCA as a good tool for visualization.

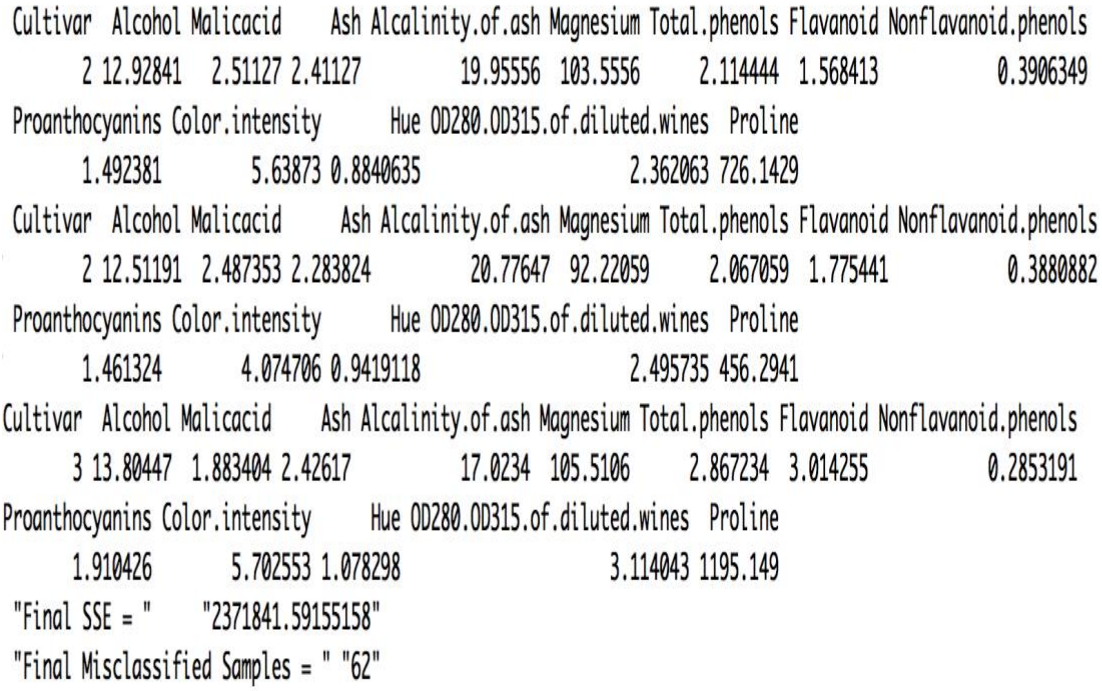
4. Kmeans clustering by using Manhattan distance

Now we use Manhattan distance to do kmeans clustering analysis for the raw data, standardized data and whitened data. Comparing with the code for Question 1, we need to change the distance function. Besides, the code used for constructing plots of the training and test set is also required to be changed.

(1) Raw data

We run the loop 50 times, the best clustering result with an SSE of 2371841.59, and 62 misclassified cases. Comparing with the result we got in Question 1, the total error cases for the best clustering results is the same (62), but the corresponding SSE is larger.

The sixth seed gives us the best results, so when we want to get the best centroids we need to pick the number of the seed first (set.seed(6)). The location of centroids for the best clustering can be found in Figure 18.



```
 Cultivar  Alcohol Malicacid      Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        2 12.92841   2.51127 2.41127          19.95556  103.5556      2.114444  1.568413            0.3906349
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline
        1.492381         5.63873 0.8840635                     2.362063 726.1429
 Cultivar  Alcohol Malicacid      Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        2 12.51191  2.487353 2.283824          20.77647  92.22059      2.067059  1.775441            0.3880882
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline
        1.461324        4.074706 0.9419118                     2.495735 456.2941
Cultivar  Alcohol Malicacid     Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       3 13.80447  1.883404 2.42617           17.0234  105.5106      2.867234  3.014255            0.2853191
Proanthocyanins Color.intensity      Hue OD280.OD315.of.diluted.wines  Proline
       1.910426        5.702553 1.078298                     3.114043 1195.149
 "Final SSE = "      "2371841.59155158"
 "Final Misclassified Samples = " "62"
```

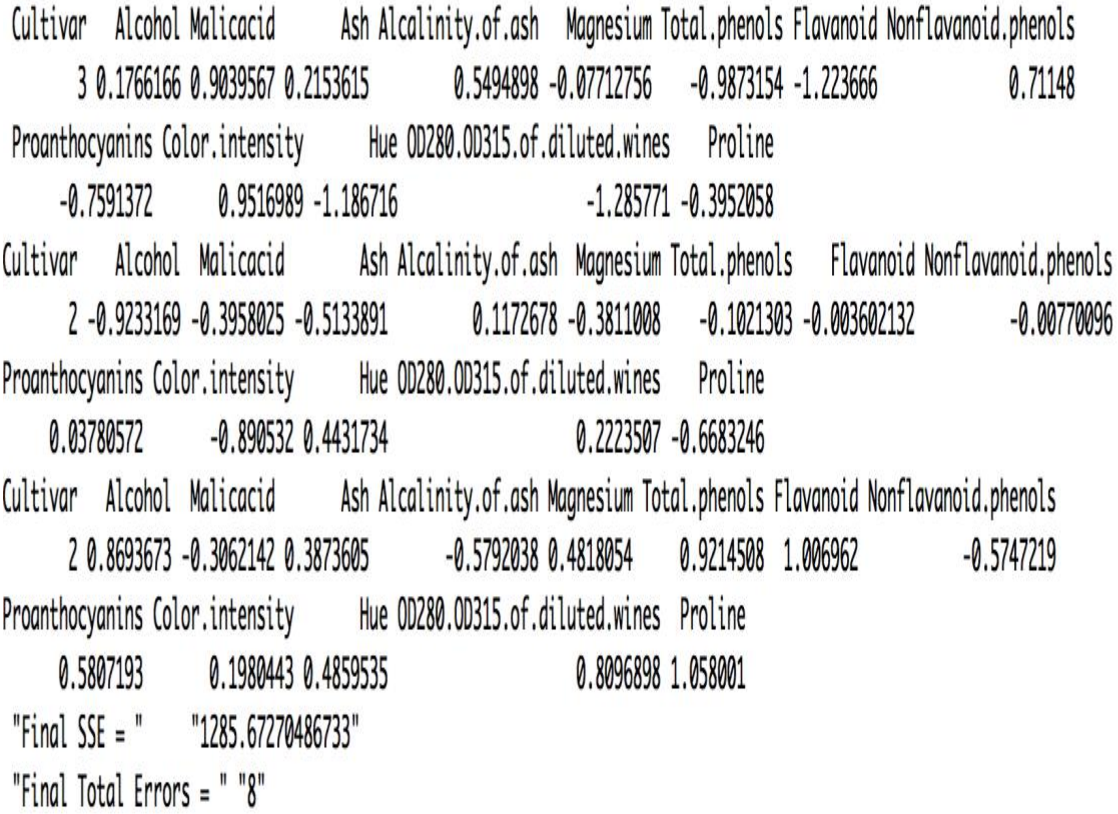
Figure 18.

(2) Standardized data

We can see some repeated results after running the loop 50 times. The best clustering result has the SSE of 1285.673, and 8 misclassified samples.

Therefore, we can see that for standardized data, the clustering results we get by using Manhattan distance is not as good as the clustering results we get by using Euclidean distance.

The location of centroids for the best clustering can be found in Figure 19.



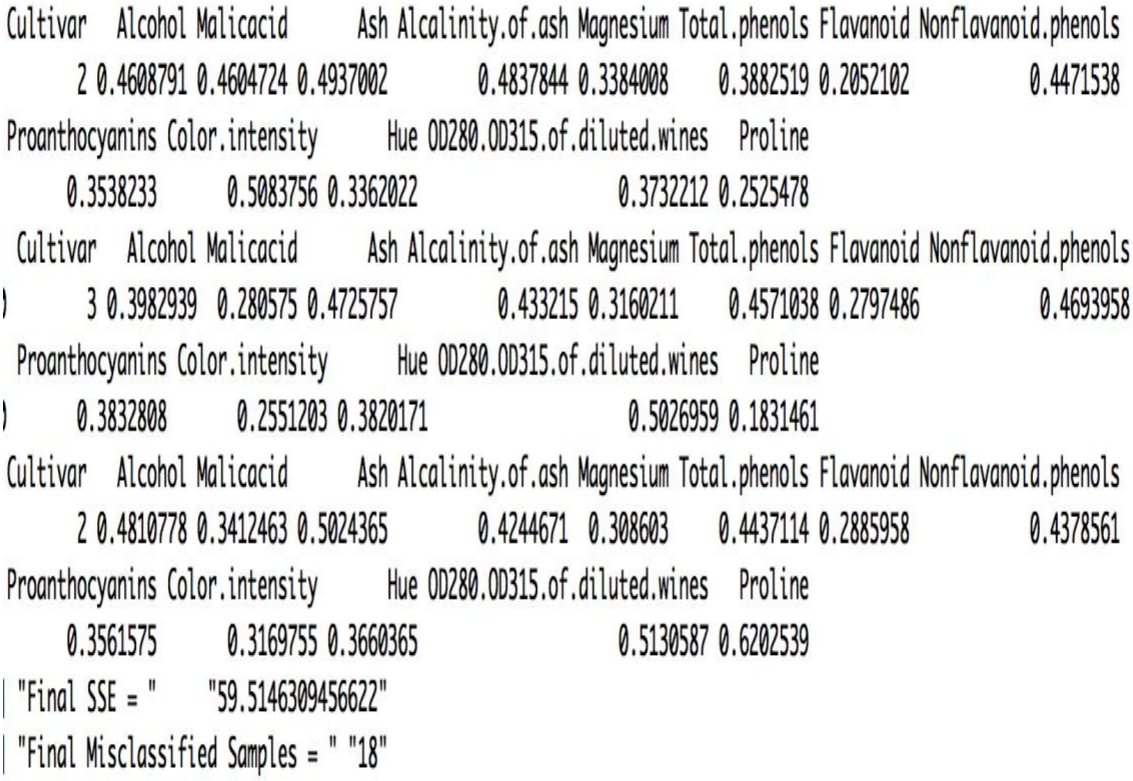
```
 Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash   Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
        3 0.1766166 0.9039567 0.2153615         0.5494898 -0.07712756    -0.9873154 -1.223666              0.71148
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline
      -0.7591372       0.9516989 -1.186716                    -1.285771 -0.3952058
Cultivar    Alcohol  Malicacid        Ash Alcalinity.of.ash  Magnesium Total.phenols    Flavanoid Nonflavanoid.phenols
       2 -0.9233169 -0.3958025 -0.5133891         0.1172678 -0.3811008    -0.1021303 -0.003602132          -0.00770096
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline
     0.03780572       -0.890532 0.4431734                    0.2223507 -0.6683246
Cultivar   Alcohol  Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       2 0.8693673 -0.3062142 0.3873605        -0.5792038 0.4818054     0.9214508  1.006962           -0.5747219
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline
      0.5807193       0.1980443 0.4859535                    0.8096898 1.058001
 "Final SSE = "     "1285.67270486733"
 "Final Total Errors = " "8"
```

Figure 19.

(3) Whitened data

We do the same thing for the whitened data, the best clustering results (when i=31) has an SSE of 59.51463, and 18 misclassified cases in total. We can see that for the whitened data, using Manhattan distance gave us a better result.

The location of centroids for the best clustering can be found in Figure 20.



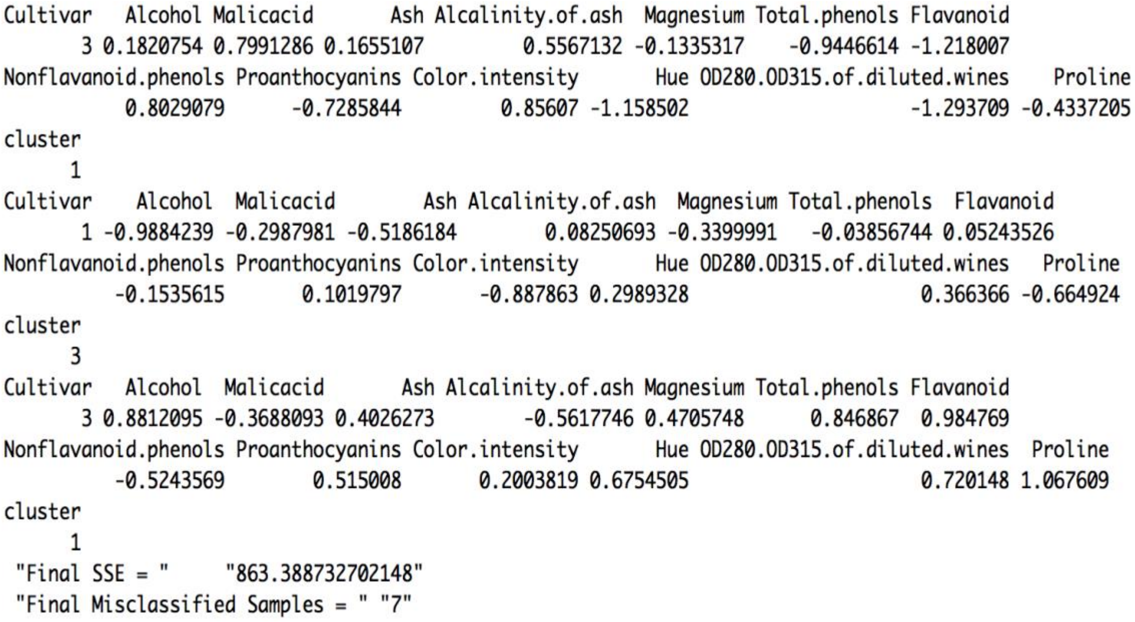
```
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       2 0.4608791 0.4604724 0.4937002         0.4837844 0.3384008     0.3882519 0.2052102            0.4471538
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
      0.3538233       0.5083756 0.3362022                    0.3732212 0.2525478
 Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
)        3 0.3982939  0.280575 0.4725757          0.433215 0.3160211     0.4571038 0.2797486            0.4693958
 Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
)       0.3832808       0.2551203 0.3820171                    0.5026959 0.1831461
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid Nonflavanoid.phenols
       2 0.4810778 0.3412463 0.5024365         0.4244671  0.308603     0.4437114 0.2885958            0.4378561
Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
      0.3561575       0.3169755 0.3660365                    0.5130587 0.6202539
 "Final SSE = "     "59.5146309456622"
 "Final Misclassified Samples = " "18"
```

Figure 20.

Again, we can see that although we changed the distance calculating method, the clustering results for the standardized data is still the best. We can end up sticking to standardized data most of the time.

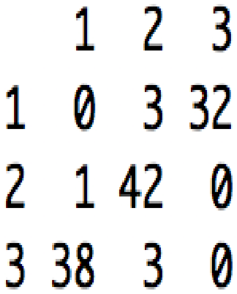
(1) Training set

After we perform our kmeans algorithm on the training dataset, the best clustering results has an SSE of 863.3887 and 7 samples were misclassified.
The location of centroids is displayed in Figure 21.



```
Cultivar   Alcohol Malicacid       Ash Alcalinity.of.ash  Magnesium Total.phenols Flavanoid
       3 0.1820754 0.7991286 0.1655107         0.5567132 -0.1335317    -0.9446614 -1.218007
Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines    Proline
           0.8029079      -0.7285844         0.85607 -1.158502                    -1.293709 -0.4337205
cluster
      1
Cultivar    Alcohol  Malicacid        Ash Alcalinity.of.ash  Magnesium Total.phenols  Flavanoid
       1 -0.9884239 -0.2987981 -0.5186184        0.08250693 -0.3399991   -0.03856744 0.05243526
Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
          -0.1535615       0.1019797       -0.887863 0.2989328                     0.366366 -0.664924
cluster
      3
Cultivar   Alcohol  Malicacid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoid
       3 0.8812095 -0.3688093 0.4026273        -0.5617746 0.4705748      0.846867  0.984769
Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines  Proline
          -0.5243569        0.515008       0.2003819 0.6754505                     0.720148 1.067609
cluster
      1
 "Final SSE = "      "863.388732702148"
 "Final Misclassified Samples = " "7"
```

Figure 21.



```
   1  2  3
1  0  3 32
2  1 42  0
3 38  3  0
```

Figure 22.

The table above shows that one sample from Cultivar1 was classified wrongly as Cultivar2, three samples from Cultivar2 were classified as Cultivar1 and another three samples from Cultivar2 were classified as Cultivar3.

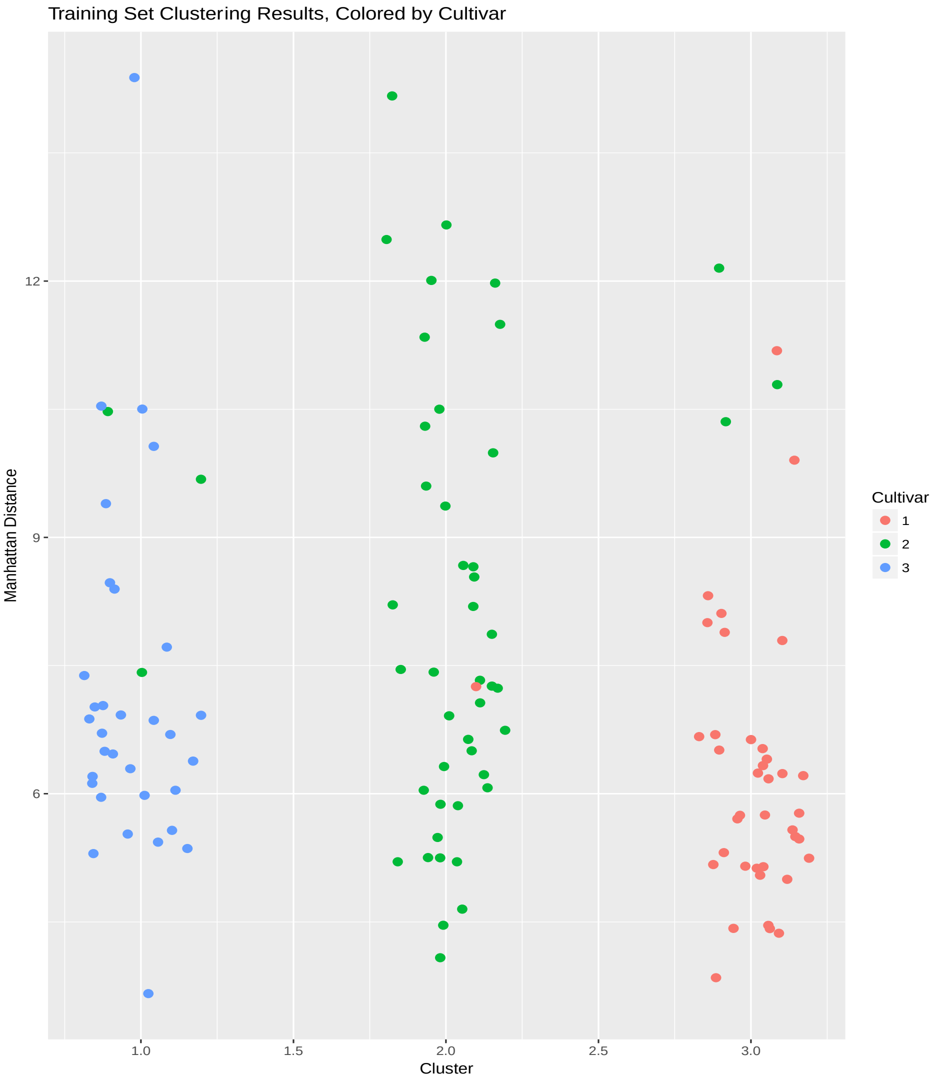


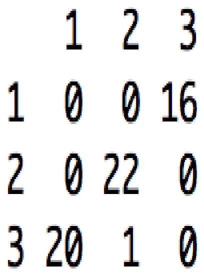
Figure 23.

The plot shows the distribution of our clusters, colored by true class, we can see misclassified samples clearly. The horizontal axis displays which cluster the sample was assigned to and the vertical axis represents the Manhattan distance from centroid.

There are 7 misclassified cases, 6 belong to Cultivar2 and 1 belongs to Cultivar1.

(2) Test set

For our test set, the best clustering results has an SSE of 481.9508, and one misclassified samples.



```
    1  2  3
1   0  0 16
2   0 22  0
3  20  1  0
```

Figure 24.

We can see from the table that the misclassified sample which comes from Cultivar2 ended up in the wrong cluster.

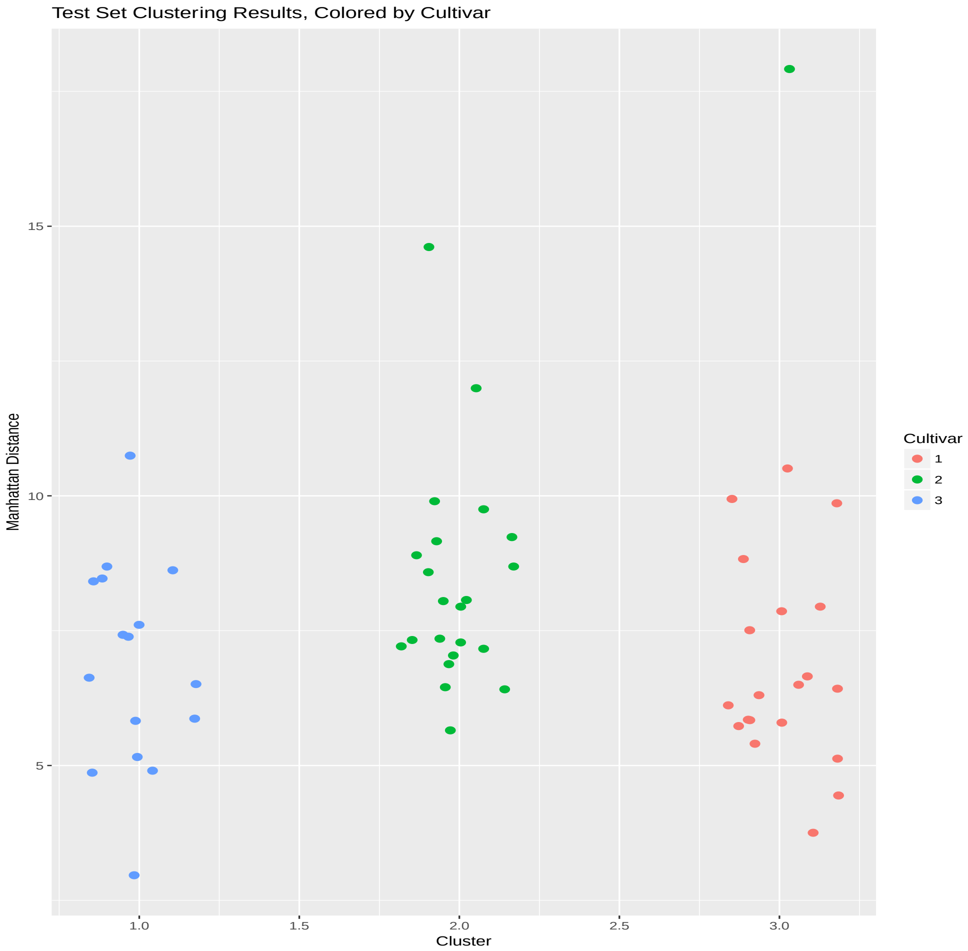


Figure 25.

If we use our test samples to evaluate the model and by using the Manhattan distance instead of the Euclidean distance, we can see that the clustering accuracy for both training set and test set is high (94.1% and 98.3%, respectively). Our model has been trained well, thus this clustering model is a reasonably good one for determining the grape cultivar used to make a wine.

5. ICA

ICA is different from PCA and independent component analysis can be seen as an extension to principal component analysis and factor analysis.

The purpose of ICA is to find independent components in the data. In PCA we will automatically get the same amounts of components as we have dimensions, while in ICA we need to specify how many independent components we want by ourselves.

And for ICA, independent components also do not have to be orthogonal and there is no "most independent component".

Since the fastICA algorithm in R will first mean-centered the data, and then performs a whitening, so we do not need to standardize our data.

Then we can use the scatterplot to display our clustering results.

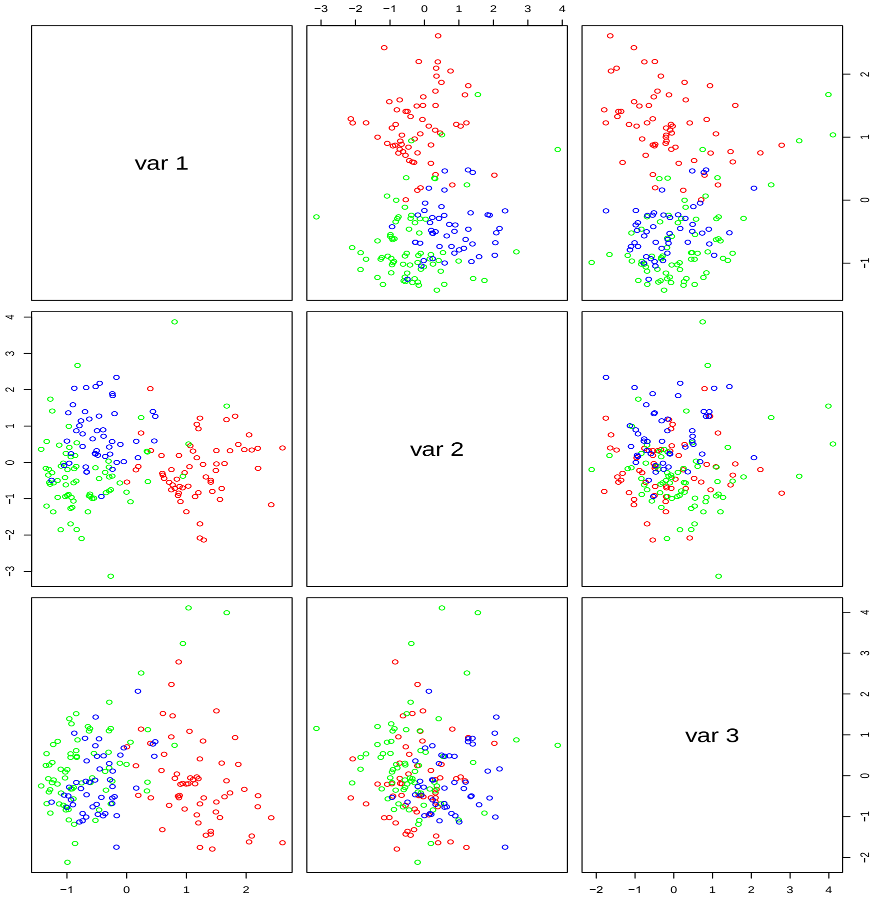


Figure 26. Results of ICA when using 3 components

We can see that when we let n.comp=3, there is no good separation. While when we use the first three PCs to do clustering, we got a high-accuracy clustering results. So maybe we need more independent components.

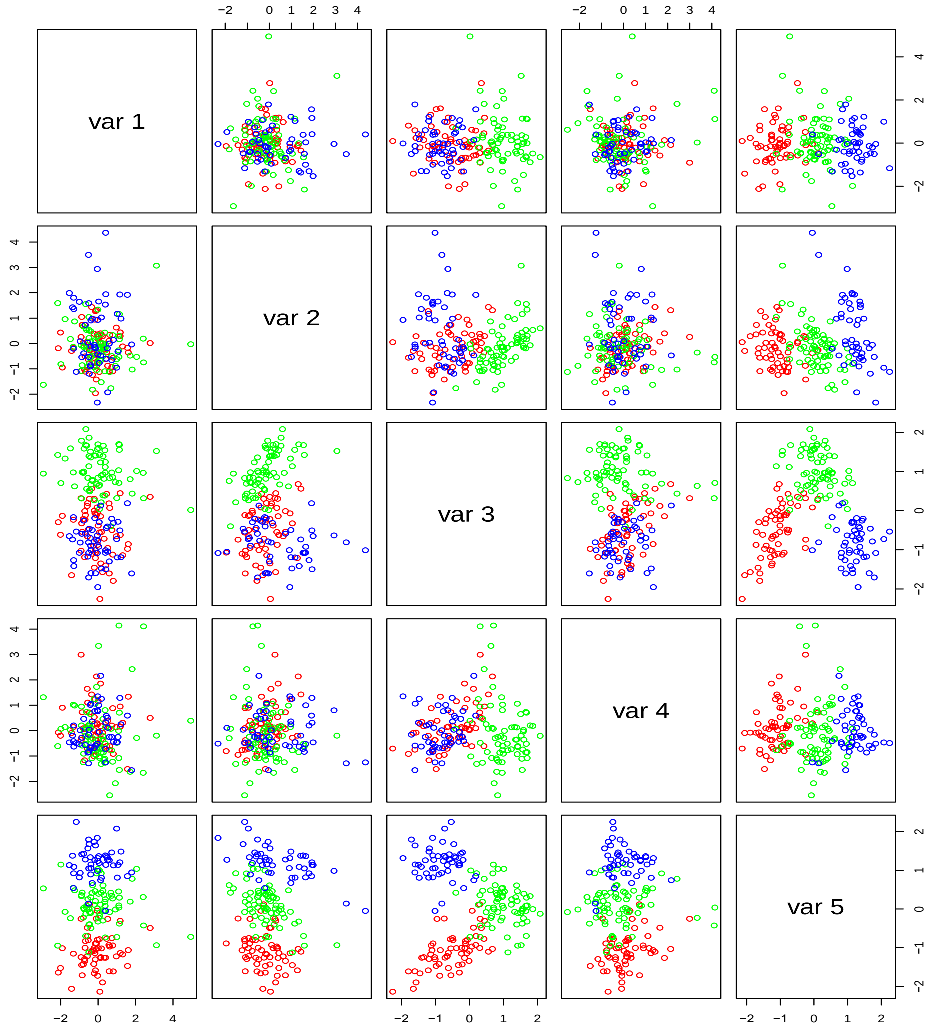


Figure 27. Results of ICA when using 5 components

If we increase the number of components to five, we can see that it is the plot of IC3 versus IC5 that shows most discrimination and is most similar to the PCA score plot shown in Figure 14. Now we try eight independent components.

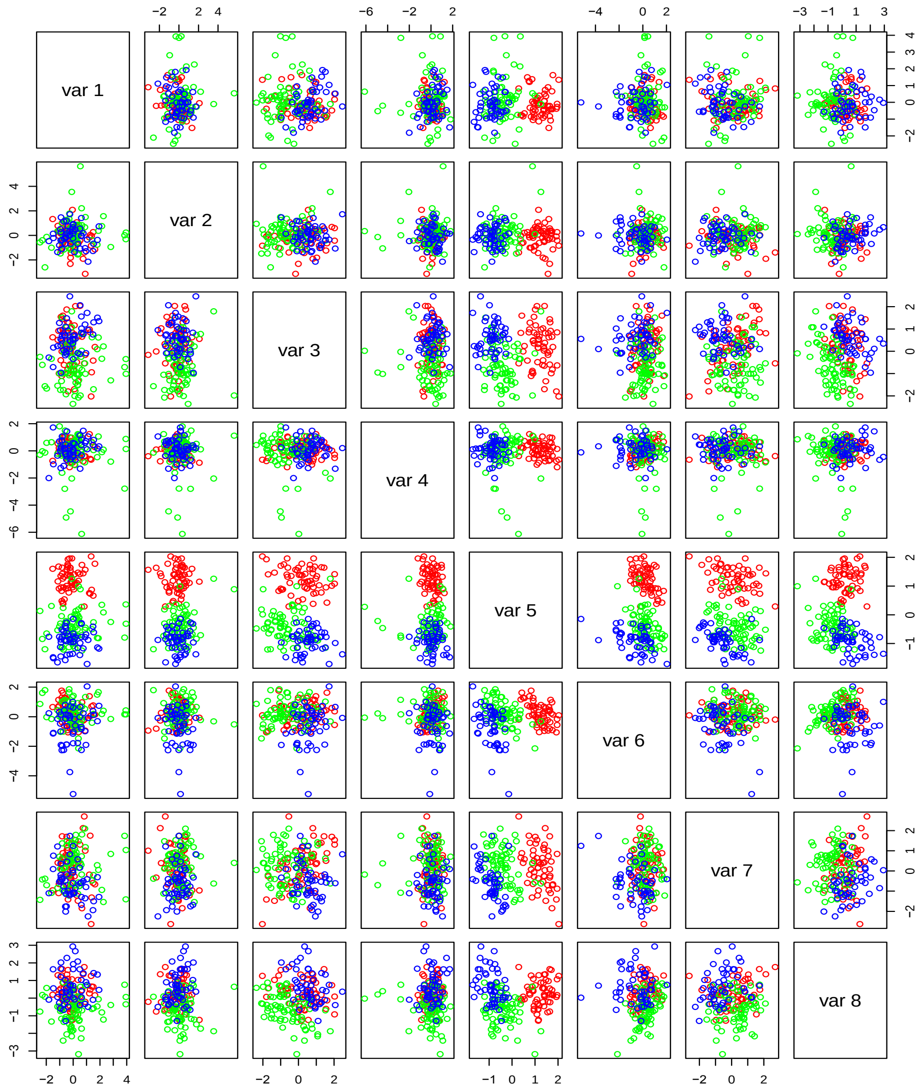


Figure 28. Results of ICA when using 8 components

When we use eight components, not many of these components seem useful and it is hard to find a good separation again.

In conclusion, one characteristic that should be noted is that in contrast to PCA, ICA components will change, depending on the number of them.

And for some datasets, ICA can do a good job on data clustering, while for some other datasets it may not perform well.

For our wine data, if we use ICA to do clustering analysis, it is better to choose five independent components.