Step 1: Download the Wine data from the UCI machine learning repository (Wine dataset- UCI Repository)

There are two files that have been downloaded.
File 1: wine-data.csv
File 2: wine-names.csv

I am combining the information in both the above files into a single file to get the variable headings, which would help me with the analysis.

The first column is named Wine Class, while the columns 2 to 13 are named Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline respectively.

The columns 2-13 are the attributes of the class of wine in column 1.

The new file created is **wine-data-updated.csv**

It is in my local folder - **A:\ISB AMPBA Course\03-Term 2\Machine Learning (Unsupervised Learning 1)\Individual Assignment 2**

The new file created will be used for analysis in R.

Step 2: Do a Principal Components Analysis (PCA) on the data. Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.

The first step to perform a Principal Component Analysis is to load the data in R and perform a few basic changes. The detailed methodology followed is available in the enclosed R file with the name
Swarna Chaithanya_12210043_MLUL1_Assignment 2.

Principal Component Analysis (PCA)
The code to perform PCA is as follows:

```
## Principal Component Analysis

WD_PCA <- princomp(WD_Data_Cleaned, cor=TRUE, scores = TRUE, covmat = NULL)
summary(WD_PCA)
plot(WD_PCA)
biplot(WD_PCA)
```

The output of Summary is as follows

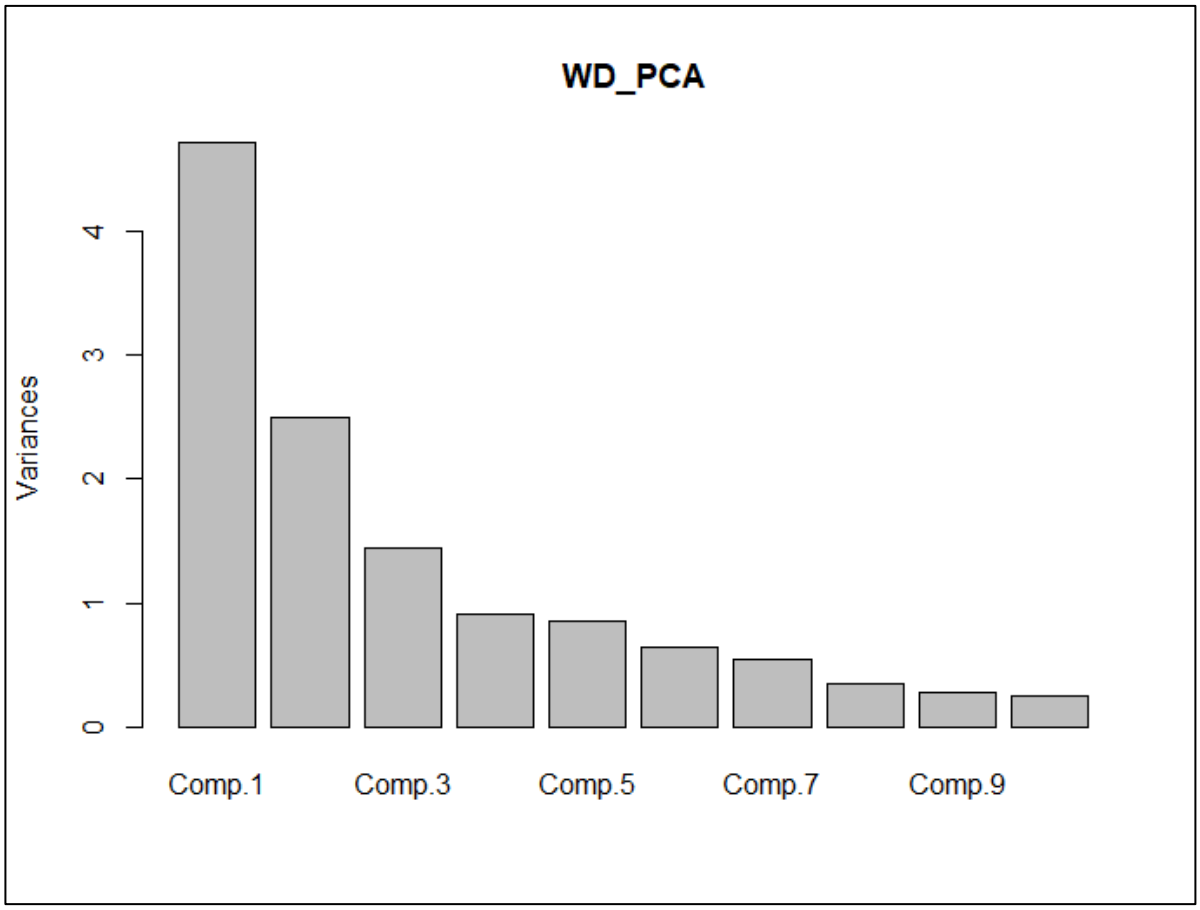
```
> summary(WD_PCA)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11
Standard deviation     2.1692972 1.5801816 1.2025273 0.9586313 0.92370351 0.80103498 0.74231281 0.59033665 0.53747553 0.50090167 0.47517222
Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294 0.04935823 0.04238679 0.02680749 0.02222153 0.01930019 0.01736836
Cumulative Proportion  0.3619885 0.5540634 0.6652997 0.7359900 0.80162293 0.85098116 0.89336795 0.92017544 0.94239698 0.96169717 0.97906553
                          Comp.12     Comp.13
Standard deviation     0.41081655 0.321524394
Proportion of Variance 0.01298233 0.007952149
Cumulative Proportion  0.99204785 1.000000000
```

The above output can be interpreted as follows:

- The first seven Principal Components have a cumulative proportion of variance of around 90%.
  - It essentially means that these 7 components are sufficient to derive all the insights that could be derived by analyzing the 13 variables in the dataset.
- When we look at the Standard Deviation values, until component 5 the values are more than 90%.
  - In component 6 the values suddenly fall to 80%

Hence, we can conclude that the analysis can be done by taking just 5 principal components.

The output of the plot is as follows:



The plot above, reiterates our findings from summary statistics. So, we go ahead with the analysis using 5 principal components.

a. Enumerate the insights you gathered during your PCA exercise. Please do not clutter your report with too many insignificant insights as it will dilute the value of your other significant findings.

The insights can be derived by looking at the loadings table.

The code and its output are as follows:

```
> ## Loading Matrix
> WD_PCA$loadings

Loadings:
                        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
Alcohol                  0.144  0.484  0.207         0.266  0.214         0.396  0.509  0.212   0.226   0.266
Malic.acid              -0.245  0.225        -0.537         0.537 -0.421                -0.309          -0.122
Ash                             0.316 -0.626  0.214  0.143  0.154  0.149 -0.170 -0.308          0.499          -0.141
Alcalinity.of.ash       -0.239        -0.612                -0.101  0.287  0.428  0.200         -0.479
Magnesium                0.142  0.300 -0.131  0.352 -0.727         -0.323 -0.156  0.271
Total.phenols            0.395        -0.146 -0.198  0.149                -0.406  0.286 -0.320  -0.304   0.304  -0.464
Flavanoids               0.423        -0.151 -0.152  0.109                -0.187        -0.163                  0.832
Nonflavanoid.phenols    -0.299        -0.170  0.203  0.501 -0.259 -0.595 -0.233  0.196  0.216  -0.117           0.114
Proanthocyanins          0.313        -0.149 -0.399 -0.137 -0.534 -0.372  0.368 -0.209  0.134   0.237          -0.117
Color.intensity                 0.530  0.137                -0.419  0.228                -0.291          -0.604
Hue                      0.297 -0.279         0.428  0.174  0.106 -0.232  0.437         -0.522          -0.259
OD280.OD315.of.diluted.wines 0.376 -0.164 -0.166 -0.184  0.101  0.266                 0.137  0.524          -0.601  -0.157
Proline                  0.287  0.365  0.127  0.232  0.158  0.120         0.120 -0.576  0.162  -0.539

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000
Proportion Var  0.077  0.077  0.077  0.077  0.077  0.077  0.077  0.077  0.077   0.077   0.077   0.077   0.077
Cumulative Var  0.077  0.154  0.231  0.308  0.385  0.462  0.538  0.615  0.692   0.769   0.846   0.923   1.000
```

Generally, the loading correlation value of 0.3 is deemed important.

Insights

- The component 1 is positively correlated to Total phenols, Flavanoids and Proanthocyanins.
- The component 2 is positively correlated to Alcohol, Ash, Magnesium, Color Intensity and Proline
- The component 3 is negatively correlated to Ash and Alkalinity of Ash
- The component 4 is negatively correlated to Malic Acid and Proanthocyanins, while it is positively correlated to Magnesium and Hue
- The component 5 is negatively correlated to Magnesium and positively correlated with Nonflavanoid Phenols.

b. What are the social and/or business values of those insights, and how the value of those insights can be harnessed—enumerate actionable recommendations for the identified stakeholder in this analysis?

For deriving the social and business insights, it is important to analyze each variable based on their contribution to health.

Good for Health	Neutral for Health	Bad for Health
Magnesium, Flavanoids, Nonflavonoid phenols, Proanthocyanins, Proline	Malic Acid, Color intensity, Hue, OD280/OD315 of diluted wines	Alcohol, Ash, Alkalinity of Ash

For a Wine Manufacturer,

- It is important to take care that the concentration of magnesium, flavonoids, Nonflavonoid phenols, Proanthocyanins and Proline are higher in quantities than alcohol and ash. Such wines could be sold with a tag of 'Better For You'. It can potentially change the traditional non-drinkers to try wine, which can be beneficial for a manufacturer.

For a Wine Consumer,

- It is important to check the composition of a wine before buying it. A few consumers with chronic diseases can even consume wine in a limited quantity to ease their life.

For Governments,

- It helps in policy advocacy and good governance to promote those kinds of wines which are Better for people, thus resulting in a good wine drinking culture as well as making higher revenue in the form of taxes.
- Alternatively, the wines with high alcohol and ash content can be heavily taxed, which, in turn can reduce the commercial viability of producing, thus, discouraging ill effects of wine consumption.

Step 3: Do a cluster analysis—you may try different algorithms or approaches and go with the one that you find most appropriate— using

(i) all chemical measurements
(ii) using two most significant PC scores.

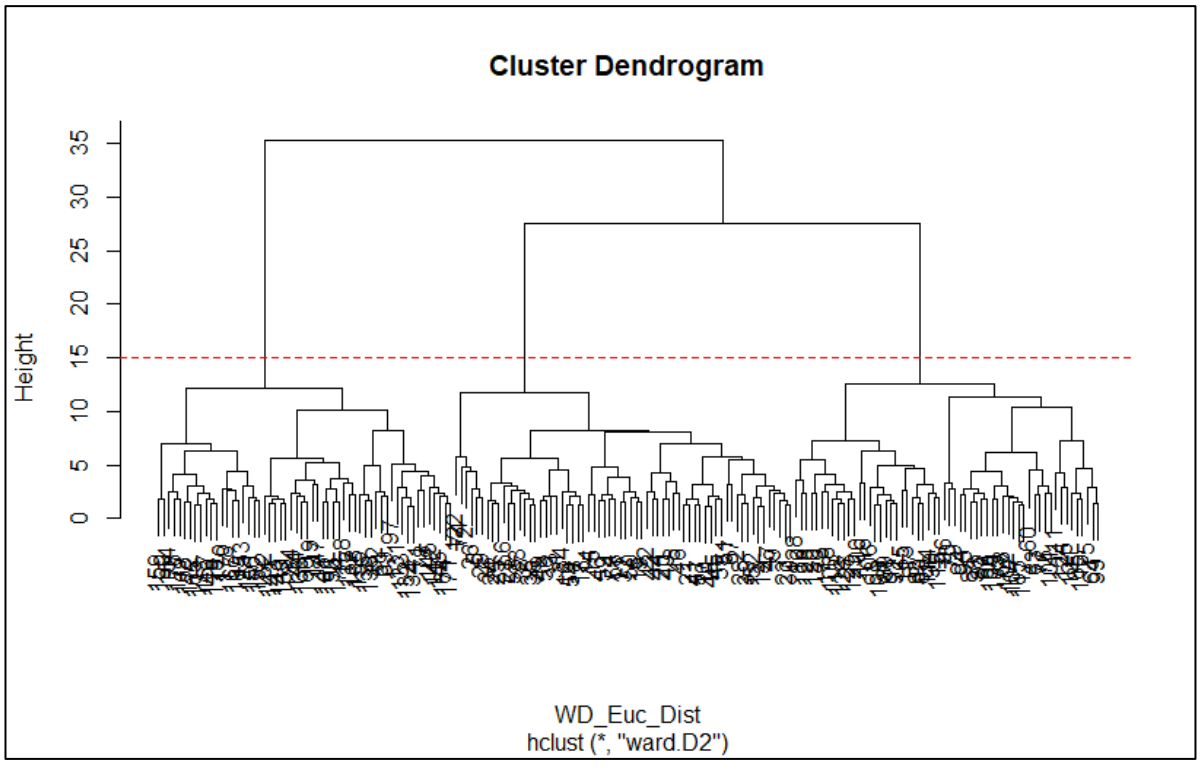
Please include (copy-paste) the relevant software outputs in your submission while answering the following questions.

Cluster Analysis by Hierarchical Clustering using all Chemical Measuements

The code is as follows:

```
## Question b> Cluster Analysis

## Calculating the Euclidean Distance using the "dist" function

## Let EWA_Euc_Dist be the variable to which Euclidean distance is assigned

WD_Euc_Dist <- dist(WD_Scaled, method = "euclidean")
str(WD_Euc_Dist)

## Applying hierarchical clustering ("hclust" function) using Ward's method and plotting the Dendrogram to see the clusters

WD_fit <- hclust(WD_Euc_Dist, method = "ward.D2")
plot(WD_fit)
abline(h=15,col="red",lty=2)
```

The output is as follows:



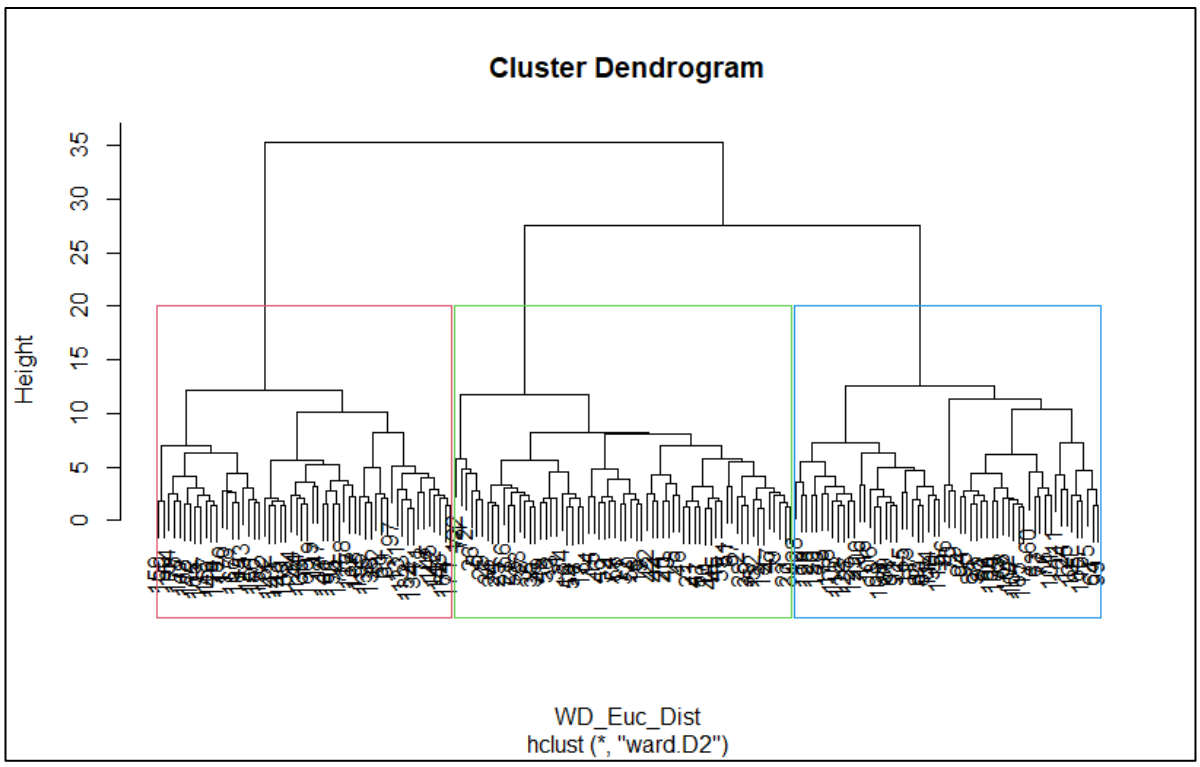
This output clearly shows that the wines can be divided into 3 clusters by using Hierarchical Clustering – Euclidean Distance and Wards method.

Visually, the three clusters are as follows:

The code is

```
## Cutting the Cluster Dendrogram into 3 different groups and finding the number of observations in each cluster

WD_Groups <- cutree(WD_fit,k=3)
plot(WD_fit)
rect.hclust(WD_fit,k=3,border = 2:5)
table(WD_Groups)
```

The output is



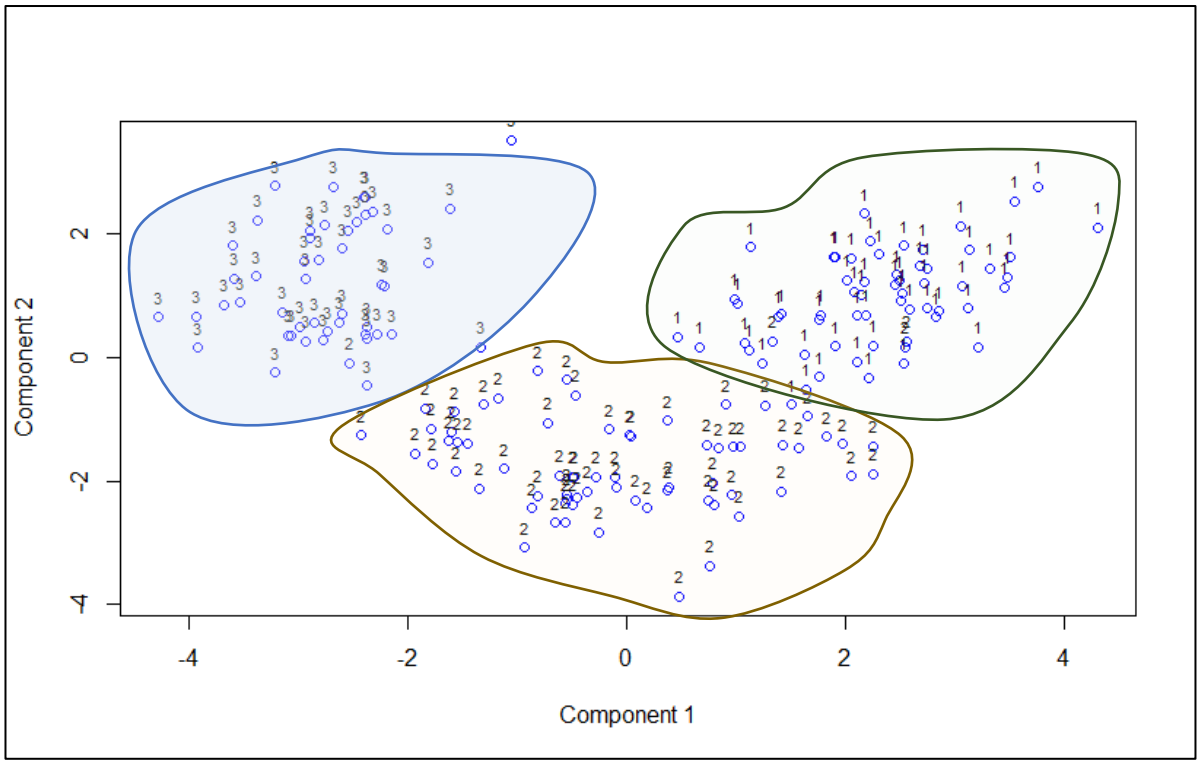
The code and output for cluster centroids is

```
> WD_Groups_Mean <- round(aggregate(WD_Scaled,list(WD_Groups),mean),2)
> print(t(WD_Groups_Mean))
                              [,1]  [,2]  [,3]
Group.1                       1.00  2.00  3.00
Alcohol                       0.82 -0.98  0.08
Malic.acid                   -0.33 -0.36  0.74
Ash                           0.35 -0.55  0.17
Alcalinity.of.ash            -0.59  0.21  0.45
Magnesium                     0.45 -0.50  0.01
Total.phenols                 0.89 -0.05 -0.96
Flavanoids                    0.98  0.06 -1.18
Nonflavanoid.phenols         -0.57 -0.05  0.70
Proanthocyanins               0.55  0.17 -0.81
Color.intensity               0.17 -0.93  0.77
Hue                           0.50  0.45 -1.03
OD280.OD315.of.diluted.wines  0.77  0.35 -1.25
Proline                       1.05 -0.78 -0.39
>
```

Cluster Analysis using two most significant PC scores

The code and the output are as follows:

```
## Cluster Analysis using two most significant PC Scores

WD_PC <- WD_PCA$scores
WD_PC
plot(WD_PC[,1], WD_PC[,2], col="blue", xlab="Component 1", ylab="Component 2")
text(WD_PC[,1:2], labels=WD_Data[,1],cex= 0.7, pos=3)
```



From the Clustering analysis using two main principal components also, there form 3 distinct clusters in the dataset.

From the above we can infer the following:

- The variance of class 1 wines ranges between 1 and 4 on principal component 1 axis and from 0 to 2.5 on principal component 2 axis.
- The variance of class 2 wines ranges between -2.2 to 2.2 on principal component 1 axis and from -4 to 0 on principal component 2 axis
- The variance of class 3 wines ranges between -4.5 to -1.5 on principal component 1 axis and from -1.7 to 2.5 on principal component 2 axis

The variance of class 2 wines is the highest on principal component 1 axis and the variance of class 1 wines is the highest on principal component 2 axis.

c. Any more insights you come across during the clustering exercise?

Looking at the cluster centroids that were arrived from hierarchical clustering

```
> WD_Groups_Mean <- round(aggregate(WD_Scaled,list(WD_Groups),mean),2)
> print(t(WD_Groups_Mean))
                              [,1]  [,2]  [,3]
Group.1                       1.00  2.00  3.00
Alcohol                       0.82 -0.98  0.08
Malic.acid                   -0.33 -0.36  0.74
Ash                           0.35 -0.55  0.17
Alcalinity.of.ash            -0.59  0.21  0.45
Magnesium                     0.45 -0.50  0.01
Total.phenols                 0.89 -0.05 -0.96
Flavanoids                    0.98  0.06 -1.18
Nonflavanoid.phenols         -0.57 -0.05  0.70
Proanthocyanins               0.55  0.17 -0.81
Color.intensity               0.17 -0.93  0.77
Hue                           0.50  0.45 -1.03
OD280.OD315.of.diluted.wines  0.77  0.35 -1.25
Proline                       1.05 -0.78 -0.39
>
```

We can infer the following:

- Wines in Cluster 1 are having the highest alcohol content, followed by wines in Cluster 3. The wines in cluster 2 have the lowest alcohol content. Manufacturers and governments can step up, reduce the production of wines in cluster 1 and increase the production of wines in cluster 2 and cluster 3.
- Ash content in wines in Cluster 1 is the highest, followed by wines in Cluster 3 and Cluster 2 respectively. Consumers should reduce the consumption of wines from Cluster 1.
- To conclude, wines in cluster 2 and cluster 3 can be promoted as Better For You wines to the consumers. Governments may increase the taxes on the wines in Cluster 1 to reduce the alcohol consumption.

d. Are there clearly separable clusters of wines? How many clusters did you go with? How the clusters obtained in part (i) are different from or similar to clusters obtained in part (ii), qualitatively?

Yes, there are three clearly separable clusters of wines. In both the type of clustering analysis (hierarchical and two main principal components), we were visually able to see the three clusters.

Initially, while I did the hierarchical clustering, intuitively I thought of going ahead with 3 clusters. At a height of 15, three distinct branches could be cut out. The same intuition was validated while using the two main principal components as well.

Qualitatively, the clusters obtained by performing clustering on all variables gives us the correlation between the variables in the dataset and the clusters. Whereas, when clustering is done on the main principal components, the data shows us the extent of variance of the different types of wines on these main principal components.

e. Could you suggest a subset of the chemical measurements that can separate wines more distinctly? How did you go about choosing that subset? How do the rest of the measurements that were not included while clustering, vary across those clusters?

The subset containing the variables Alcohol, Malic acid, Magnesium, Phenols, Flavanoids, Proanthocyanins, OD280/OD315, distinctly separates the clusters.

The subset was chosen based on the variance score (loadings) that these variables had on principal component 1 and principal component 2.

The other measurements that were not chosen for analysis carry a low score comparatively and would not have a significant impact on the analysis.