Data Analysis and Visualization - Assignment 5

Ma Jingchun, 2020111235

- 通过如下命令，加载数据集 wine：

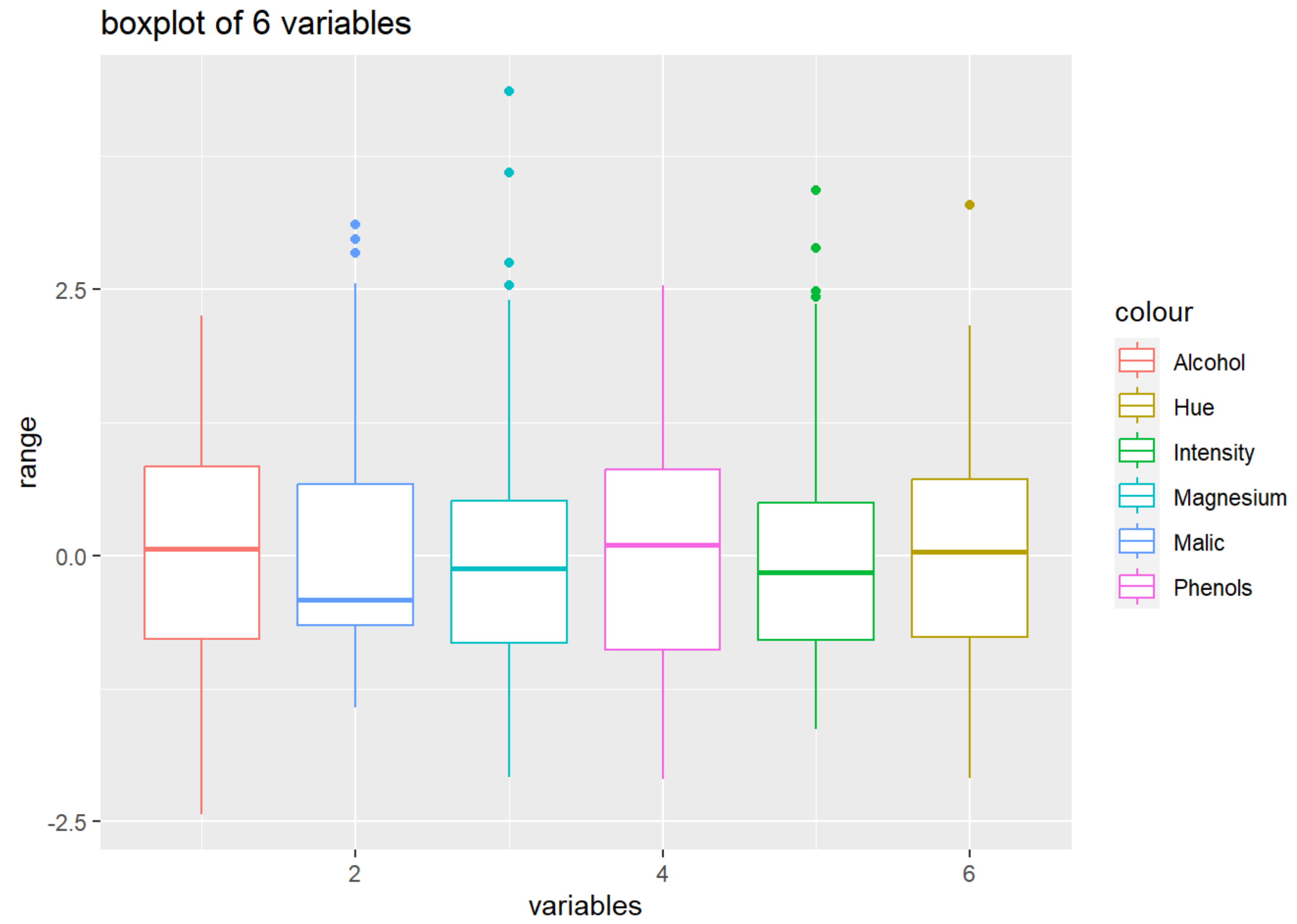
```
library(gclus)
library(tidyverse)
library(ggplot2)
library(gridExtra)
library(corrplot)
library(factoextra)
library(cluster)
library(mclust)
```

```
######### Please write your R code in this chunk #########
data(wine)
wineTrain <- wine[, which(names(wine) != "Class")]
attach(wineTrain)
```

该数据集有14个变量，178条关于酒的记录；其中，第一列 Cultivar 为一个多分类指标的标签。（该数据集是一个开源数据，有兴趣的同学可以通过数据网站查看每个变量的具体含义）。进一步，通过如下命令，生成去标签后的训练样本集 wineTrain（注意，这里我们没有选 random sample 来做后续模型估计）。

- 针对变量 Alcohol, Malic.Acid, Magnesium, Total.phenols, Color.intensity, Hue，进行描述性统计分析。请用一幅图内展示每个变量在标准化之后的箱型图，选用适当的颜色以及图片的主标题和横纵坐标的标题。从图中，有显示出可能的异常值吗？如果存在，请找出其在原始数据集中的行数。

```
######### Please write your R code in this chunk #########
### Solution to Q1
scl_wineTrain = as_tibble(scale(wineTrain))
ggplot(data = scl_wineTrain) +
  geom_boxplot(mapping = aes(x=1, y=Alcohol,color="Alcohol")) +
  geom_boxplot(mapping = aes(x=2, y=Malic, color="Malic")) +
  geom_boxplot(mapping = aes(x=3, y=Magnesium, color="Magnesium")) +
  geom_boxplot(mapping = aes(x=4, y=Phenols, color="Phenols")) +
  geom_boxplot(mapping = aes(x=5, y=Intensity, color="Intensity")) +
  geom_boxplot(mapping = aes(x=6, y=Hue, color="Hue")) +
  labs(x='variables', y='range', title='boxplot of 6 variables')
```



Malic、Magnesium、Intensity、Hue有异常值点存在。

```
# Malic
sort.list(wineTrain$Malic,decreasing=T)[1:3]
```

```
## [1] 124 174 138
```

```
# Magnesium
sort.list(wineTrain$Magnesium,decreasing=T)[1:4]
```

```
## [1] 96 70 74 79
```

```
# Intensity
sort.list(wineTrain$Intensity,decreasing=T)[1:4]
```

```
## [1] 159 160 152 167
```

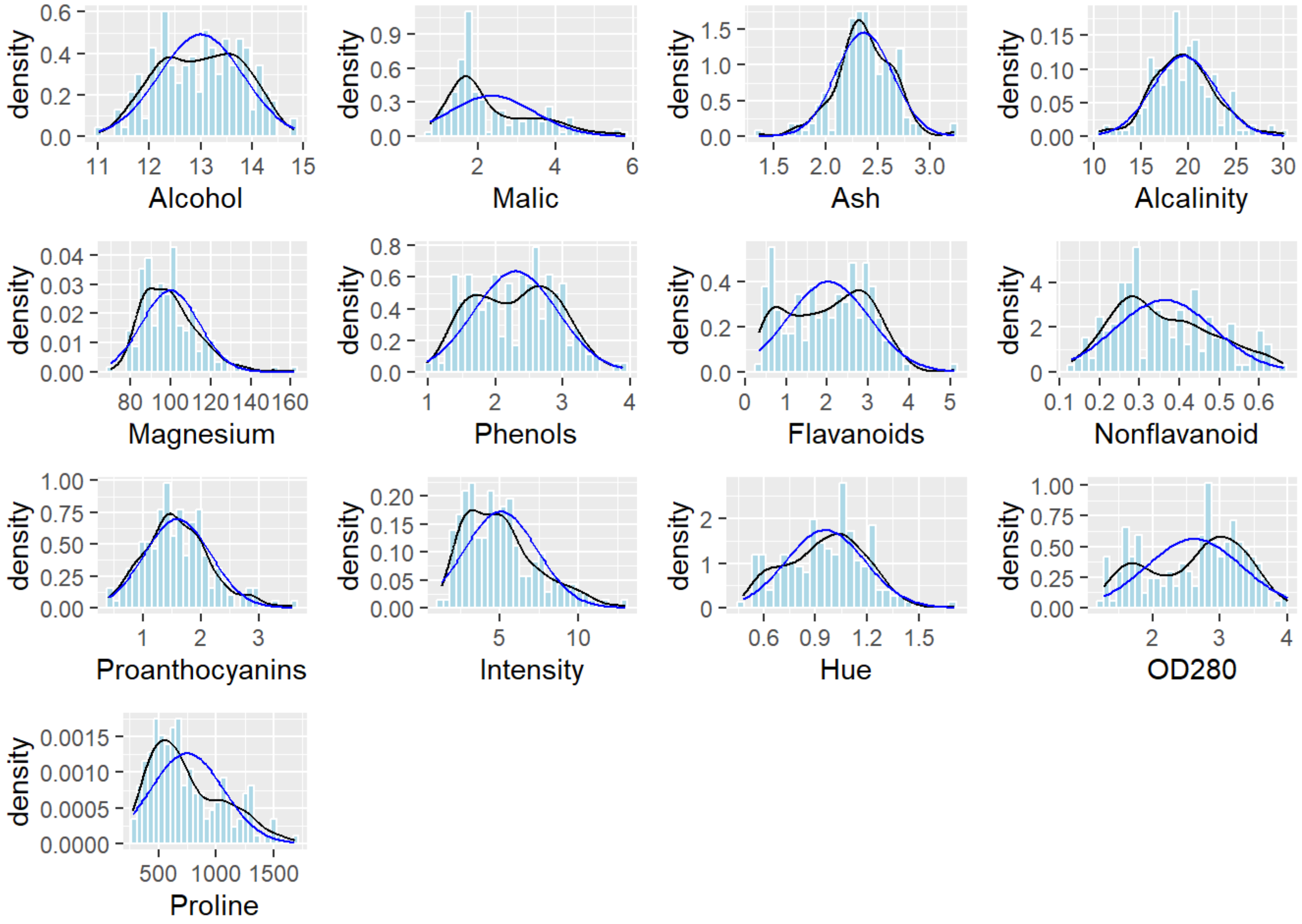
```
# Hue
sort.list(wineTrain$Hue,decreasing=T)[1]
```

```
## [1] 116
```

- 请选用ggplot2中适当的图表类型，展示每个变量的样本分布是否有偏，以及相关图标的格式，如颜色，标题，图例等等。

```
######### Please write your R code in this chunk #########
### Solution to Q2
p1 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Alcohol,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Alcohol)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Alcohol),
          sd=sd(wineTrain$Alcohol)),
    color='blue')
p2 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Malic,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Malic)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Malic),
          sd=sd(wineTrain$Malic)),
    color='blue')
p3 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Ash,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Ash)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Ash),
          sd=sd(wineTrain$Ash)),
    color='blue')
p4 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Alcalinity,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Alcalinity)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Alcalinity),
          sd=sd(wineTrain$Alcalinity)),
    color='blue')
p5 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Magnesium,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Magnesium)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Magnesium),
          sd=sd(wineTrain$Magnesium)),
    color='blue')
p6 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Phenols,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Phenols)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Phenols),
          sd=sd(wineTrain$Phenols)),
    color='blue')
p7 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Flavanoids,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Flavanoids)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Flavanoids),
          sd=sd(wineTrain$Flavanoids)),
    color='blue')
p8 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Nonflavanoid,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Nonflavanoid)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Nonflavanoid),
          sd=sd(wineTrain$Nonflavanoid)),
    color='blue')
p9 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Proanthocyanins,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Proanthocyanins)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Proanthocyanins),
          sd=sd(wineTrain$Proanthocyanins)),
    color='blue')
p10 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Intensity,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Intensity)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Intensity),
          sd=sd(wineTrain$Intensity)),
    color='blue')
p11 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Hue,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Hue)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Hue),
          sd=sd(wineTrain$Hue)),
    color='blue')
p12 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = OD280,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = OD280)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$OD280),
          sd=sd(wineTrain$OD280)),
    color='blue')
p13 <- ggplot(data = wineTrain) +
  geom_histogram(aes(x = Proline,y = ..density..), bins=30, color='white',fill='lightblue') + geom_density(mapping = aes(x = Proline)) +
  stat_function(fun=function(x)
    dnorm(x,
          mean = mean(wineTrain$Proline),
          sd=sd(wineTrain$Proline)),
    color='blue')
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13, ncol=4)
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
```



无偏：Alcalinity

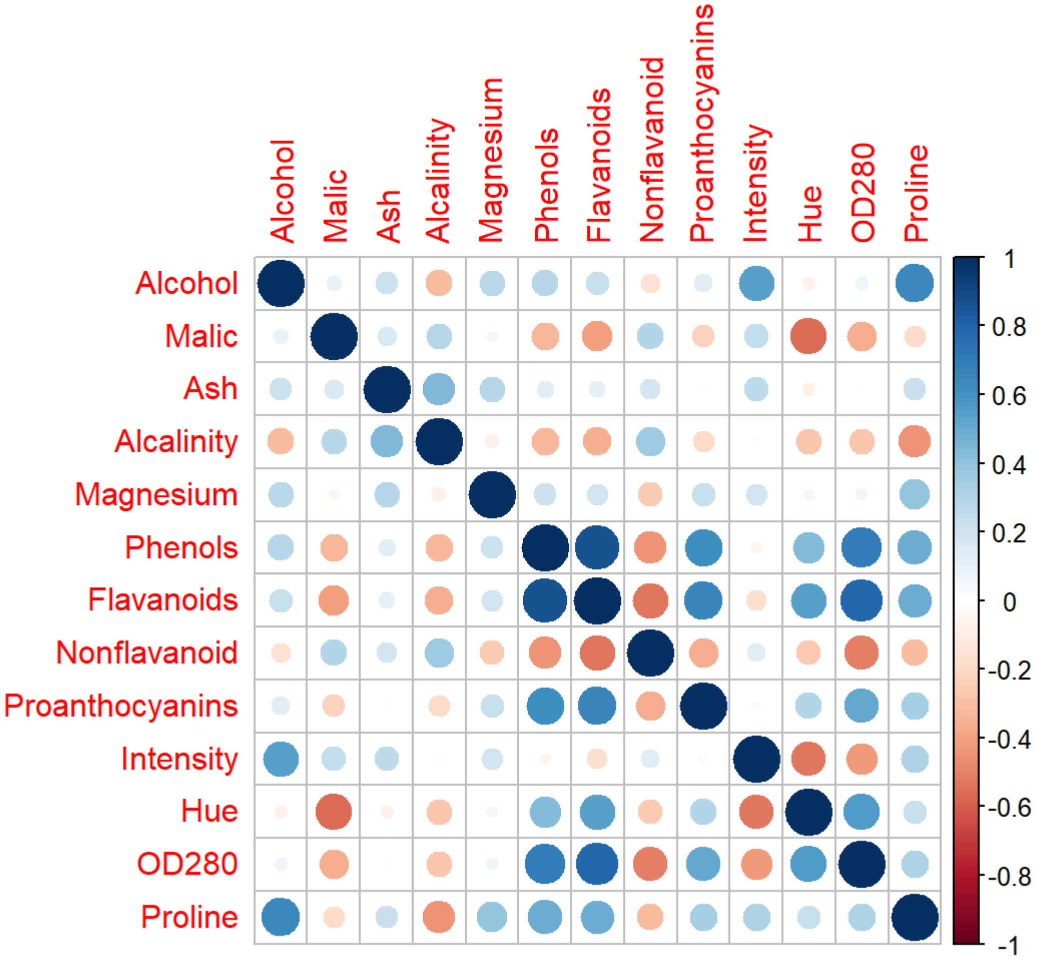
左偏：Ash Hue

右偏：Malic Magnesium Nonflavanoid Proanthocyanins Intensity Proline

双峰：Alcohol Phenols Flavanoids OD280

- 请选用合适的方式，计算并展示 wineTrain 数据集中所有变量的两两相关性。你哪些变量之间的相关性比较高？

```
########## Please write your R code in this chunk #########
### Solution to Q3
corrplot(cor(wineTrain))
```

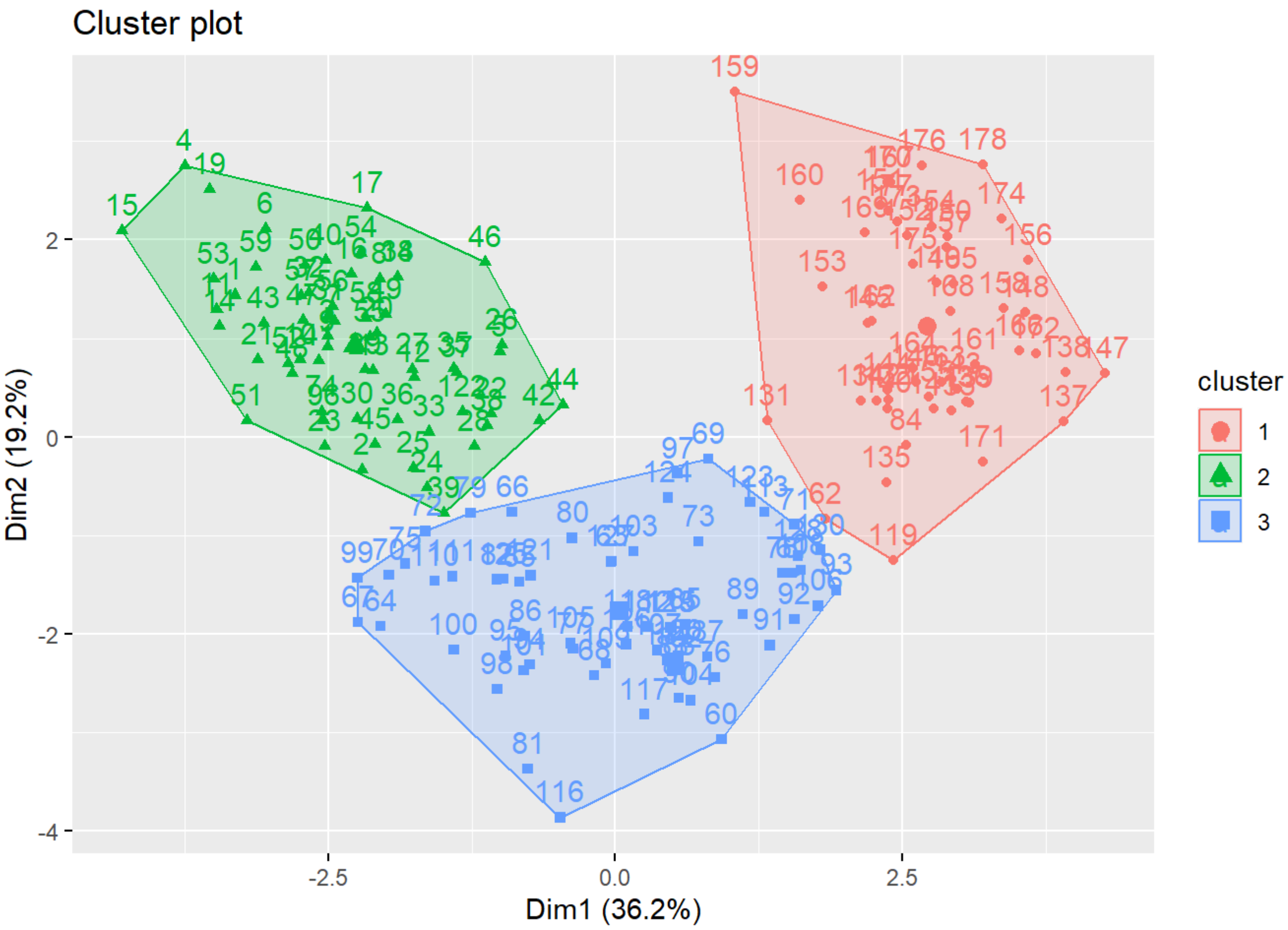


正相关性：Phenols 和 Flavanoids、Alcohol 和 Proline、OD280 和 Flavanoids

负相关性：Malic 和 Hue、Flavanoids 和 Nonflavanoid

- 设定随机数种子为你的学号，通过 k-means 方式进行聚类，其中，中心的个数定为 3 个。请通过合适的图表(建议 ggplot2 相关图表)，展示你的聚类效果。你认为 kmeans 的聚 类效果如何 ？

```
########## Please write your R code in this chunk #########
### Solution to Q4
set.seed(2020111235)
k <- kmeans(scl_wineTrain, centers = 3, nstart = 25)
scl_wineTrain$cluster = k[["cluster"]]
# fviz_cluster
fviz_cluster(k, data = scl_wineTrain[-14])
```

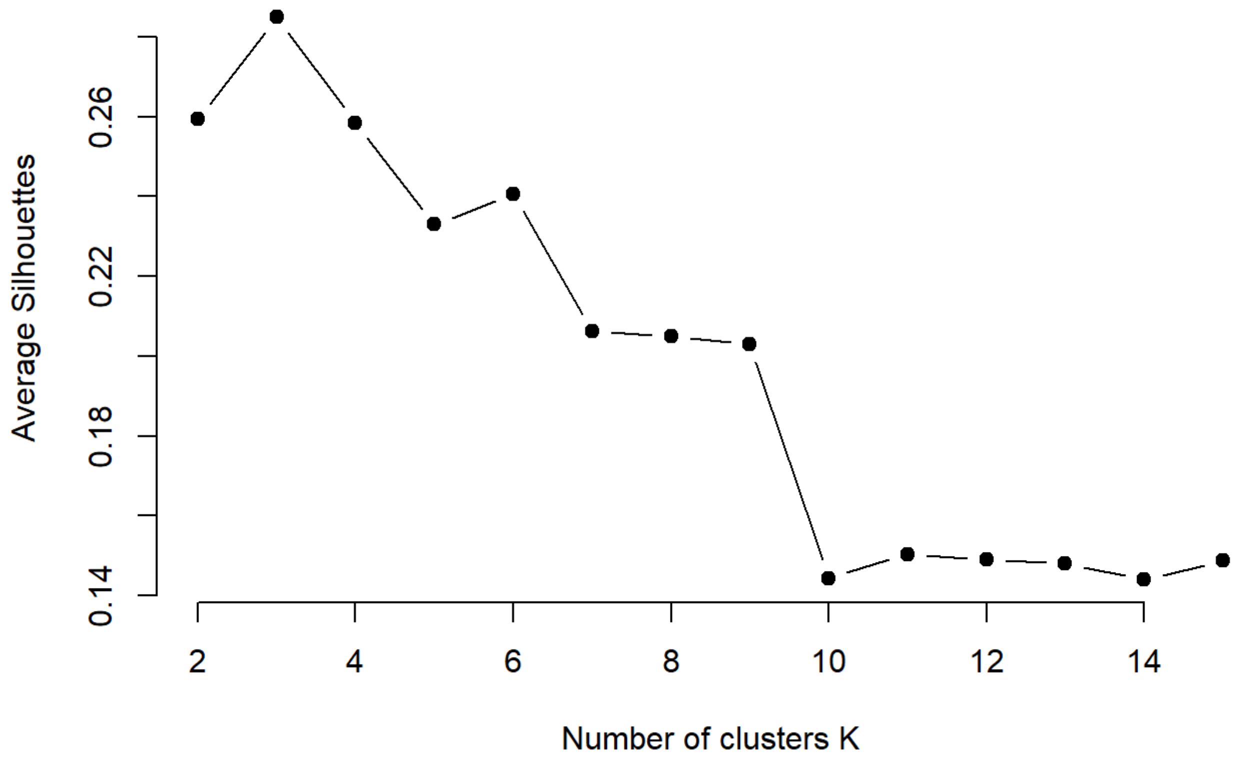


我认为聚类的效果不是很好，由上面两张图展示出聚成三类中有很多重叠的部分，且分类边界不是很明显

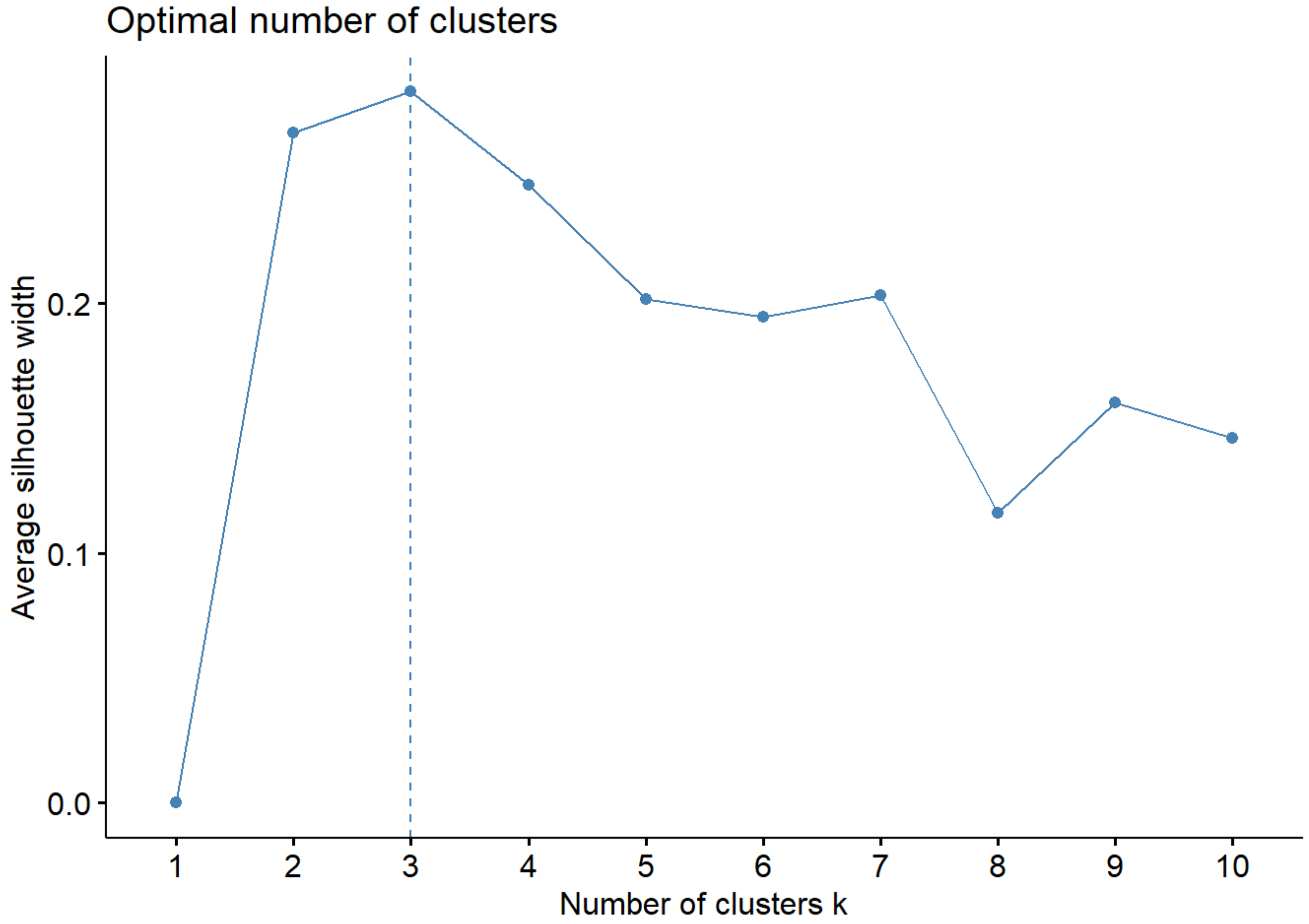
- 请通过 silhouette 统计量和 gap 统计量，分别决定 cluster 组的个数的最优值，并将你得到的结果进行展示。两种方法给出的最优组数是否相同？如果不同，你觉得哪个更合理。其中 nstart 设定为 25. 此时，组的个数与原始数据集中 wine 中的变量 Cultivar 的可能取值相比，是否相同？

```
######### Please write your R code in this chunk #########
### Solution to Q5

# silhouette method
avg_sil <- function(k) {
  km.res <- kmeans(scl_wineTrain[-14], centers = k, nstart = 25)
  ss <- silhouette(km.res$cluster, dist(scl_wineTrain[-14]))
  mean(ss[, 3])
}
k.values <- 2:15
avg_sil_values <- map_dbl(k.values, avg_sil)
plot(k.values, avg_sil_values,
     type = "b", pch = 19, frame = FALSE,
     xlab = "Number of clusters K",
     ylab = "Average Silhouettes")
```



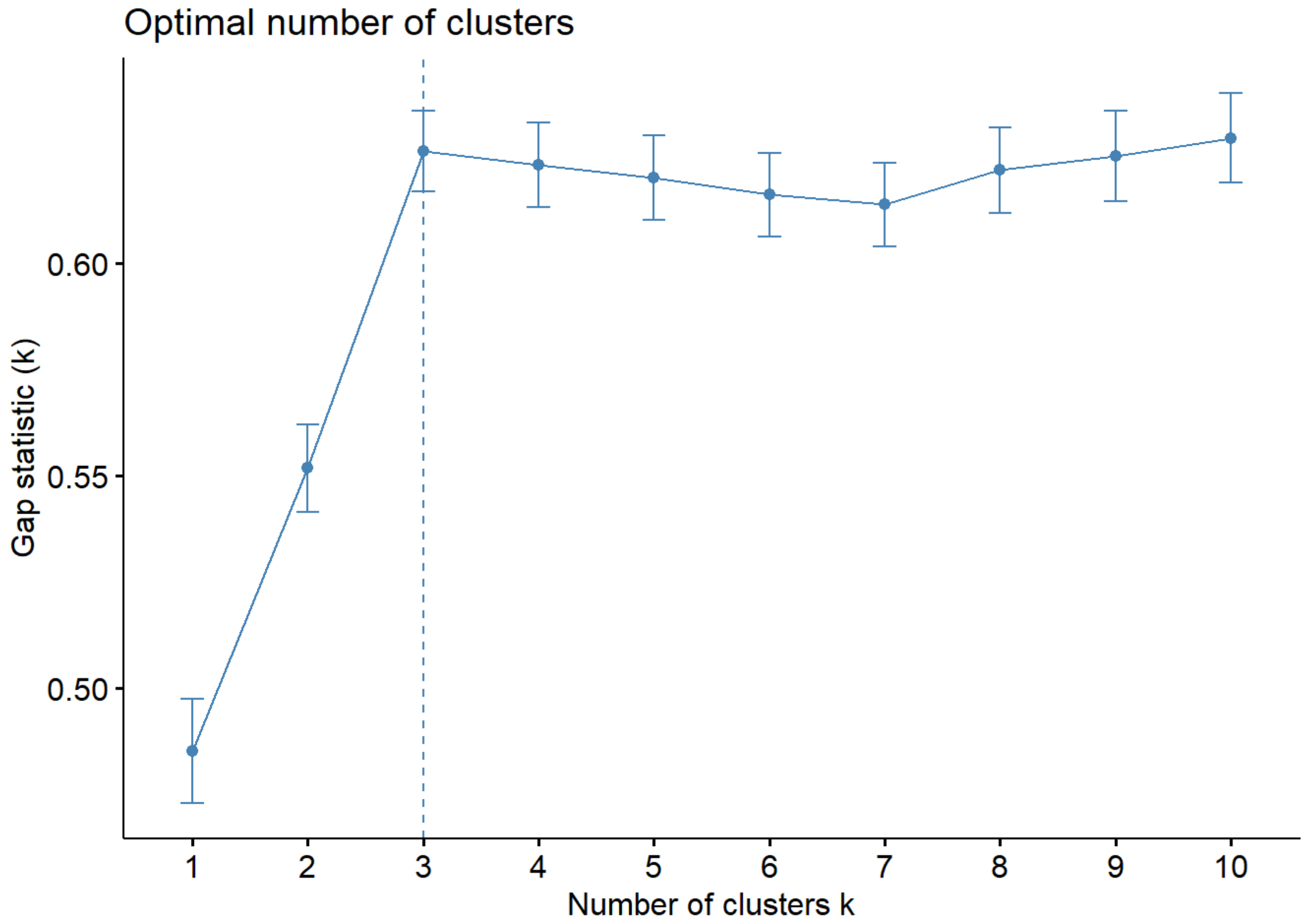
```
fviz_nbclust(scl_wineTrain[-14], kmeans, method = "silhouette")
```



```r
# Gap method
gap_stat <- clusGap(scl_wineTrain[-14], FUN = kmeans, nstart = 25,K.max = 10, B = 50)
print(gap_stat, method = "firstmax")
```

```
## Clustering Gap statistic ["clusGap"] from call:
## clusGap(x = scl_wineTrain[-14], FUNcluster = kmeans, K.max = 10, B = 50, nstart = 25)
## B=50 simulated reference sets, k = 1..10; spaceH0="scaledPCA"
##  --> Number of clusters (method 'firstmax'): 3
##           logW   E.logW       gap      SE.sim
##  [1,] 5.377557 5.862751 0.4851941 0.012320130
##  [2,] 5.203502 5.755315 0.5518138 0.010262703
##  [3,] 5.066921 5.693327 0.6264054 0.009535599
##  [4,] 5.023936 5.647073 0.6231366 0.009926301
##  [5,] 4.989510 5.609635 0.6201249 0.010005693
##  [6,] 4.961100 5.577240 0.6161406 0.009860739
##  [7,] 4.935538 5.549332 0.6137941 0.009921164
##  [8,] 4.902337 5.524279 0.6219429 0.010104241
##  [9,] 4.876049 5.501297 0.6252480 0.010541900
## [10,] 4.850382 5.479887 0.6295047 0.010445655
```

```r
fviz_gap_stat(gap_stat)
```



采用标准化后的数据进行聚类，两种方法给出的最优组数相同，获得聚类结果均为3.

- 设定随机数种子为你的学号，通过 k-means 方式进行聚类，其中，中心的个数定为 3 个。 根据每个个体的分组情况，与其对应的标签相比，吻合情况如何？你可以展示一下confusion matrix。

```
######### Please write your R code in this chunk #########
### Solution to Q6
scl_wineTrain[scl_wineTrain$cluster==1,"cluster"] <- 6
scl_wineTrain[scl_wineTrain$cluster==2,"cluster"] <- 4
scl_wineTrain[scl_wineTrain$cluster==3,"cluster"] <- 5
scl_wineTrain[scl_wineTrain$cluster==4,"cluster"] <- 1
scl_wineTrain[scl_wineTrain$cluster==5,"cluster"] <- 2
scl_wineTrain[scl_wineTrain$cluster==6,"cluster"] <- 3
table(wine$Class, scl_wineTrain$cluster)
```

```
##     
##       1  2  3
##    1 59  0  0
##    2  3 65  3
##    3  0  0 48
```

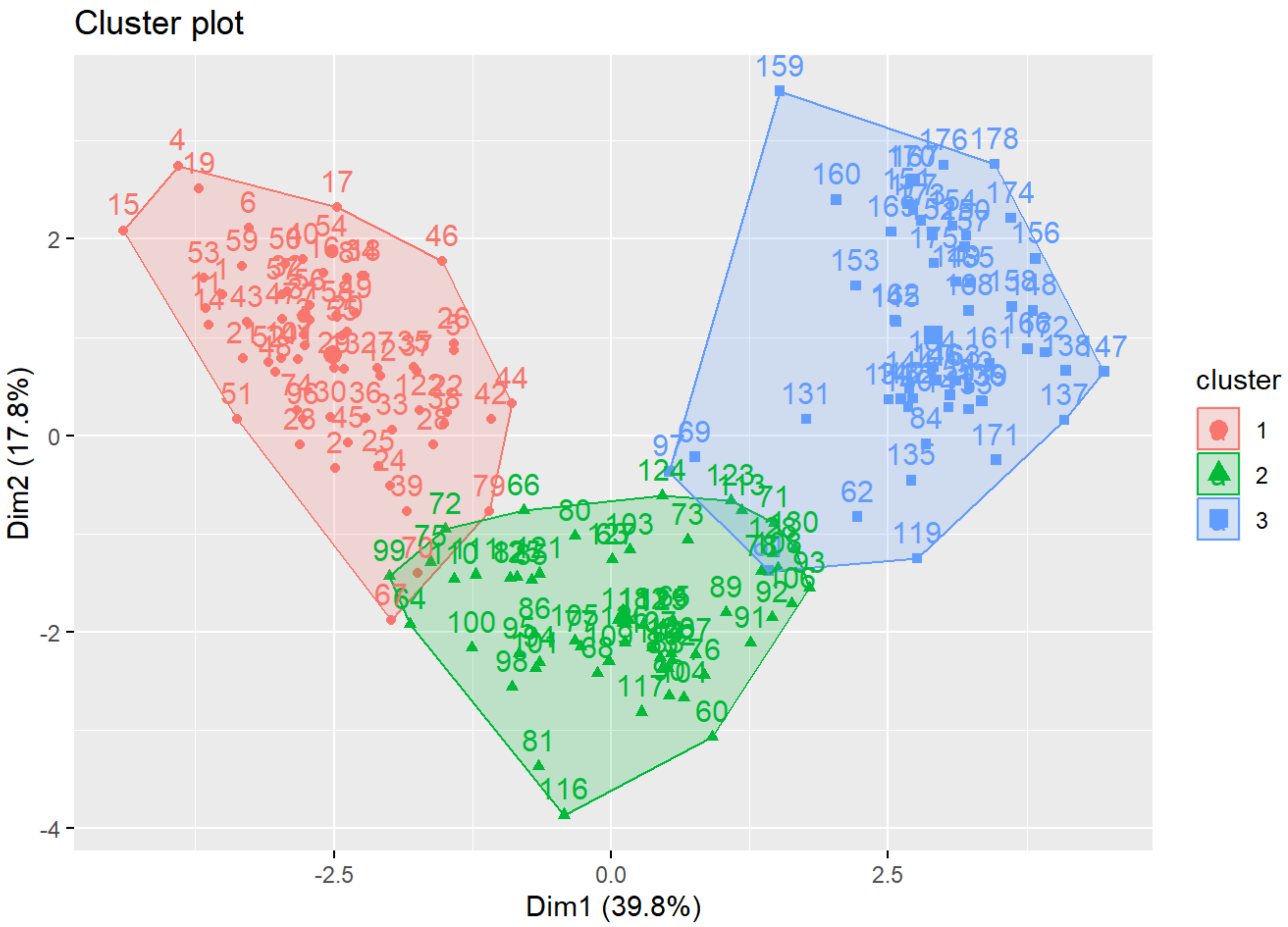
存在少量分错的情况

- 请展示通过层次聚类hclust函数进行聚类的结果，并通过合适的可视化方式进行展示。该方法与 k-means 相比，效果如何？

```
######### Please write your R code in this chunk #########
### Solution to Q7
res.hc <- eclust(scl_wineTrain, "hclust") # compute hclust
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## ℹ The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <←]8;;https://github.com/kassambara/factoextra/issues●https://g
ithub.com/kassambara/factoextra/issues←]8;;●>.
```

```
fviz_cluster(res.hc) # scatter plot
```



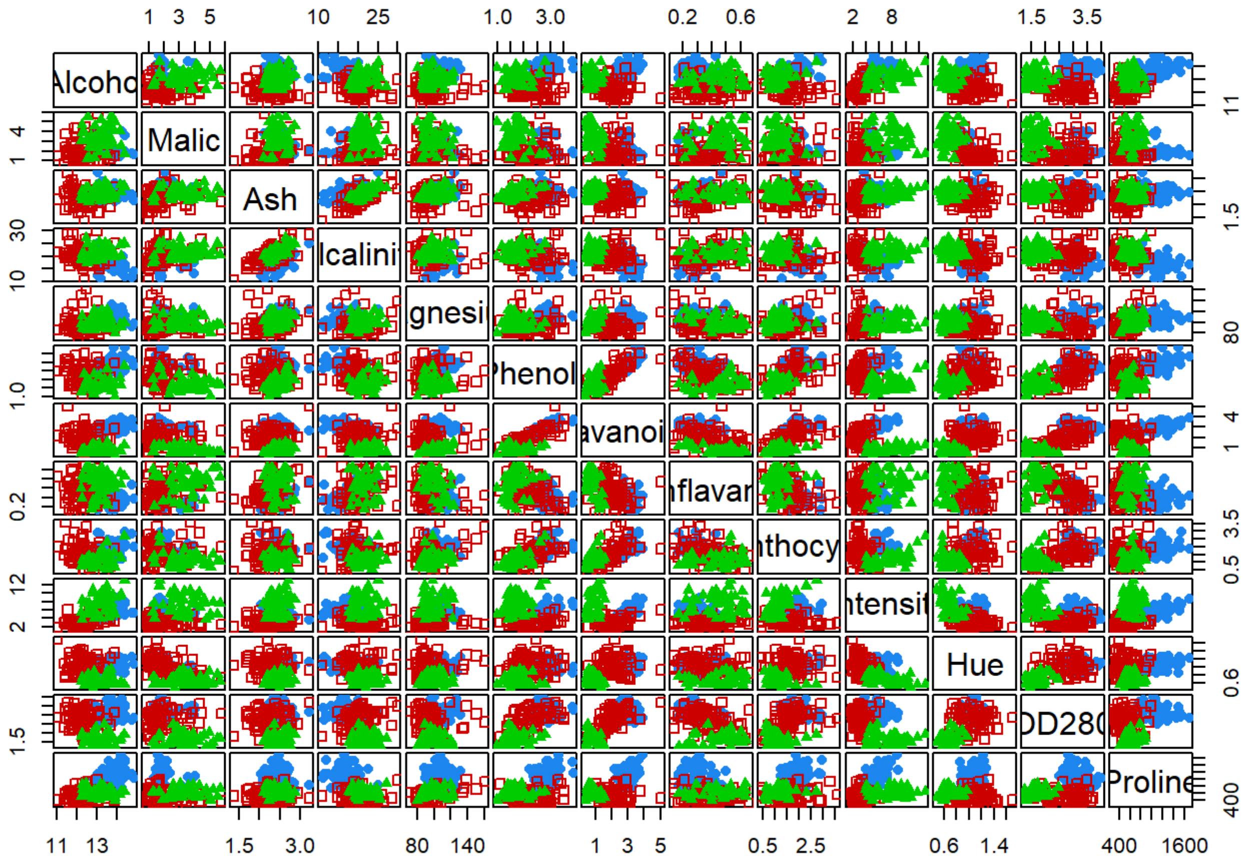
```
table(wine$Class, res.hc[["cluster"]])
```

```
##
##       1  2  3
##    1 59  0  0
##    2  6 59  6
##    3  0  0 48
```

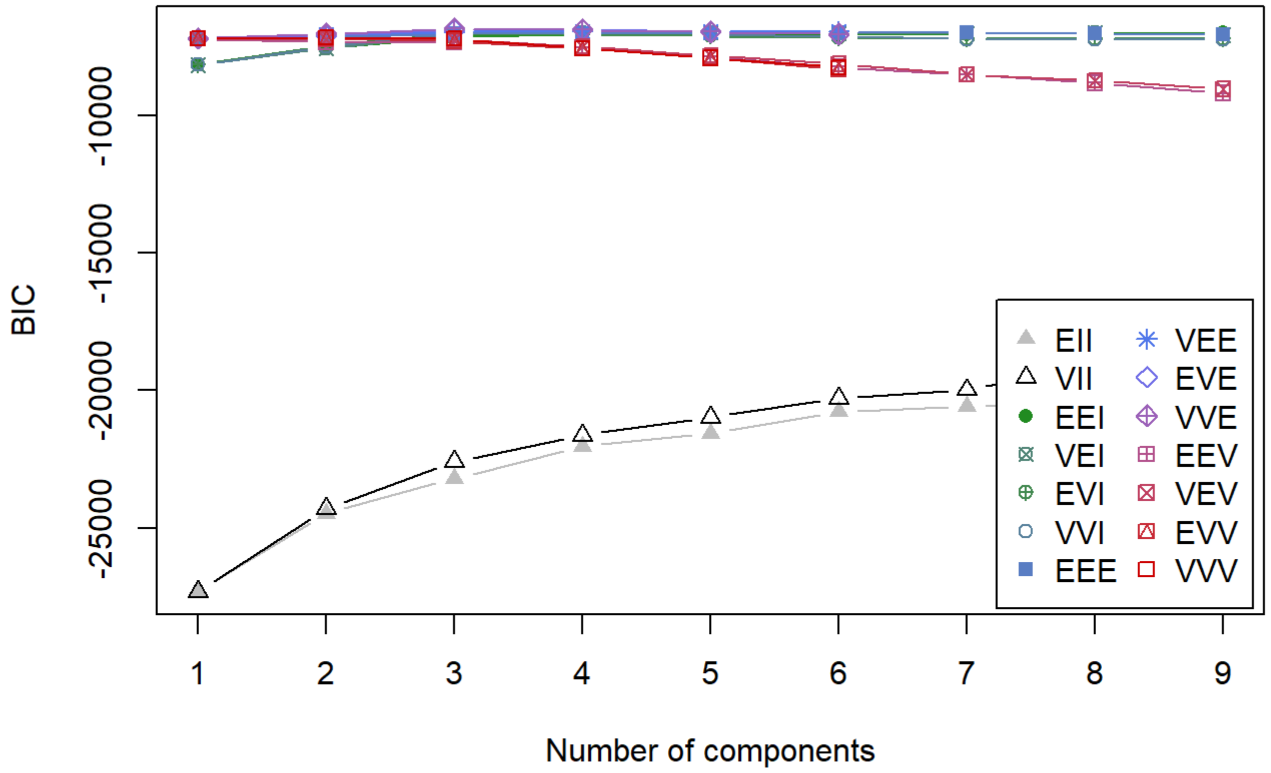
从图中看层次聚类没有kmeans方法聚类效果好，分类边界存在重叠程度更大，从Confusion matrix看，分错的总数减少，但对某两类的分类产生的错误更多

- 请通过任何一种你学过的分类方法，将 wine 进行分类，其中 Cultivar 作为响应变量，得到每个样本点的分类的预测值。对比 k-means 的 k 取 3 的时候的聚类效果，你认为通过 kmeans 方法聚类后用来做标签的预测效果怎么样？哪个更精准？你觉得可能的原因有哪些？

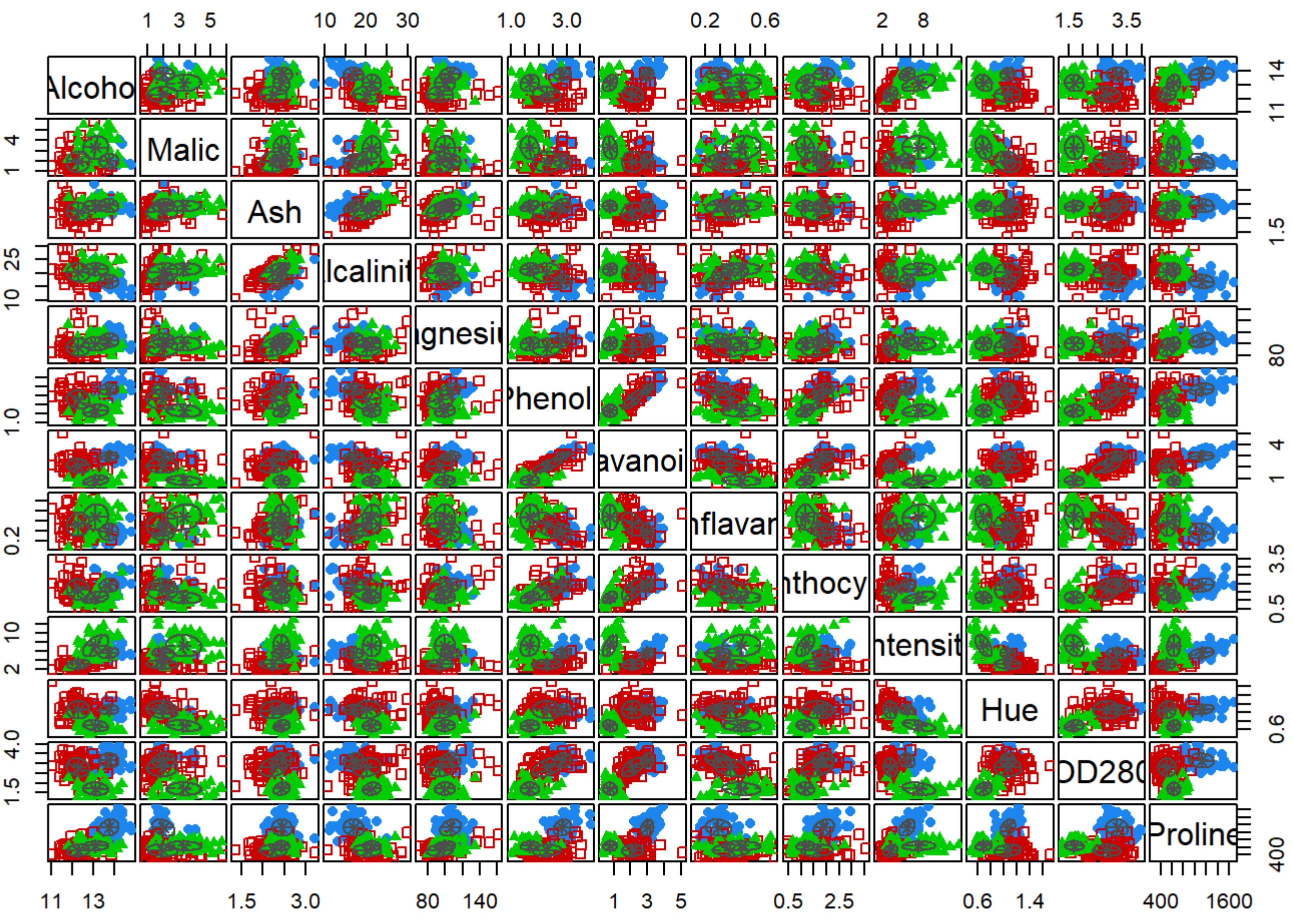
```
######### Please write your R code in this chunk #########
### Solution to Q8
clPairs(wineTrain, wine$Class)
```



```
BIC <- mclustBIC(wineTrain)
plot(BIC)
```



```
mod1 <- Mclust(wineTrain, x = BIC)
plot(mod1, what = "classification")
```



```
table(wine$Class, mod1$classification)
```

```
##
##       1  2  3
##   1 59  0  0
##   2  0 69  2
##   3  0  0 48
```

分类准确性比kmeans好，没有分错的情况