2020111142_谢嘉薪_Ass5

Xie Jiaxin, 2020111142

导入包并加载数据集

```
library(gclus)
library(ggplot2)
library(gridExtra)##支持ggplot2多图并列
library(GGally)
library(factoextra)
library(mclust)
data("wine")
dim(wine)
```

```
## [1] 178  14
```

```
wineTrain <- wine[, which(names(wine) != "Class")]
dim(wineTrain)
```

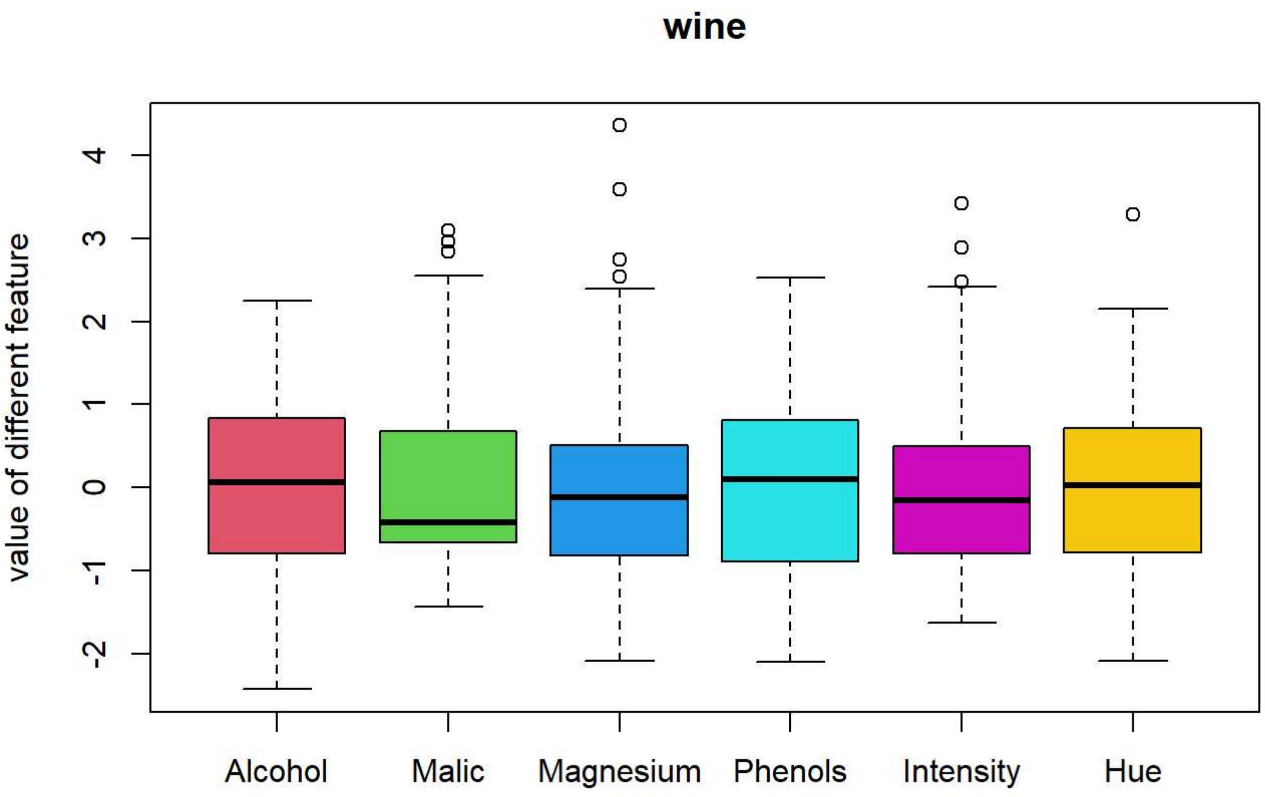
```
## [1] 178  13
```

作业五

标准化箱型图

a. 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。

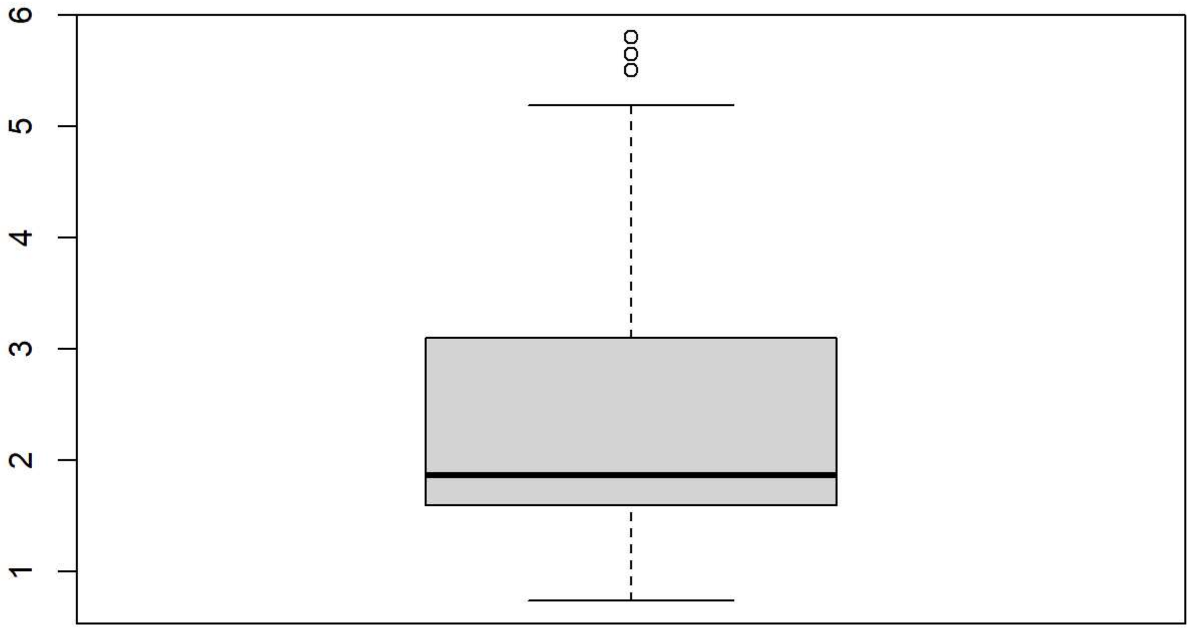
```
boxplot(apply(wine[,c(’Alcohol’, ’Malic’, ’Magnesium’, ’Phenols’, ’Intensity’, ’Hue’)],
              2,
              FUN=function(x){(x-mean(x))/sd(x)}),
        horizontal=F,
        col=c(2,3,4,5,6,7),
        main=”wine”,
        ylab=”value of different feature”)
```



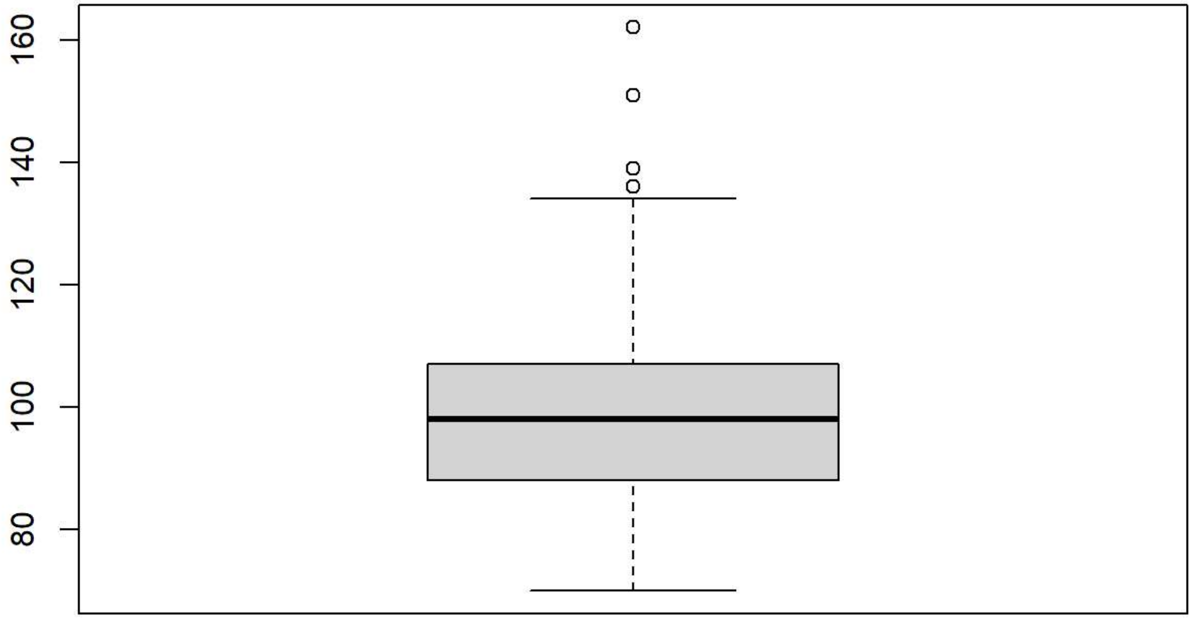
查找异常值

发现在’Malic’, ‘Magnesium’, ‘Intensity’, ’Hue’中存在离群值，认为可能是异常值，输出其行数及数值

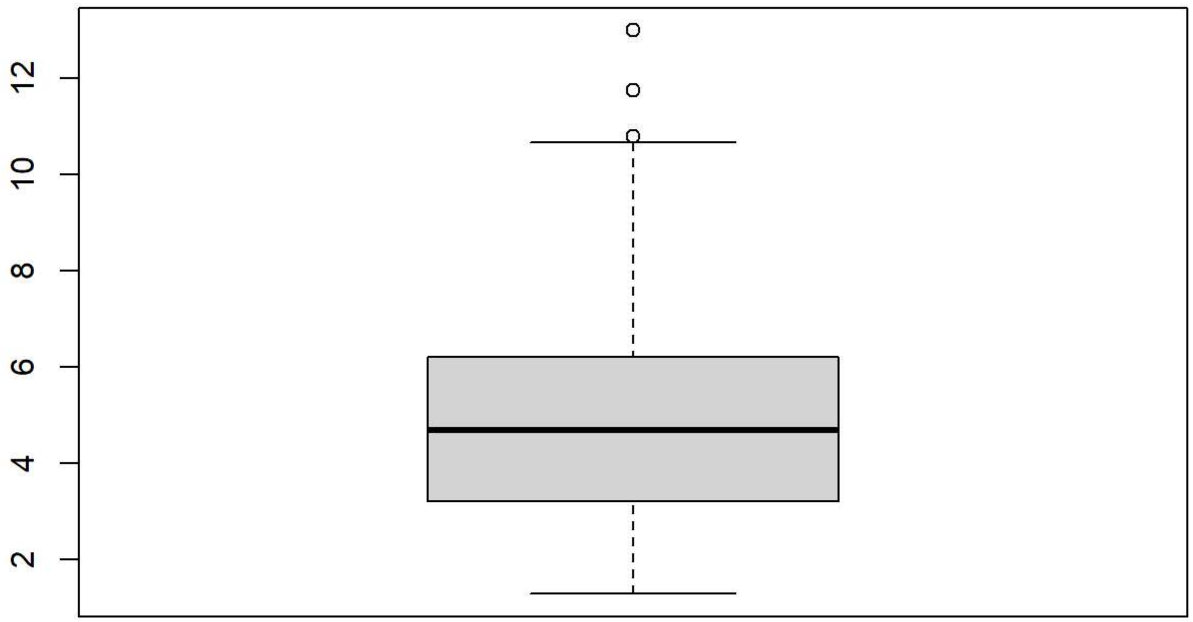
```
out_name <- c(’Malic’, ’Magnesium’, ’Intensity’, ’Hue’)
idx <- which(names(wine) %in% out_name)
name<-names(wine)
for(i in idx ){
  outVals<-boxplot(wine[,i])$out
  out<-which(wine[,i] %in% outVals)
  print("变量名:")
  print(name[i])# 输出变量名
  print("离群值:")
  print(outVals)# 输出离群值
  print("离群值行数")
  print(out)# 输出离群值行数
}
```



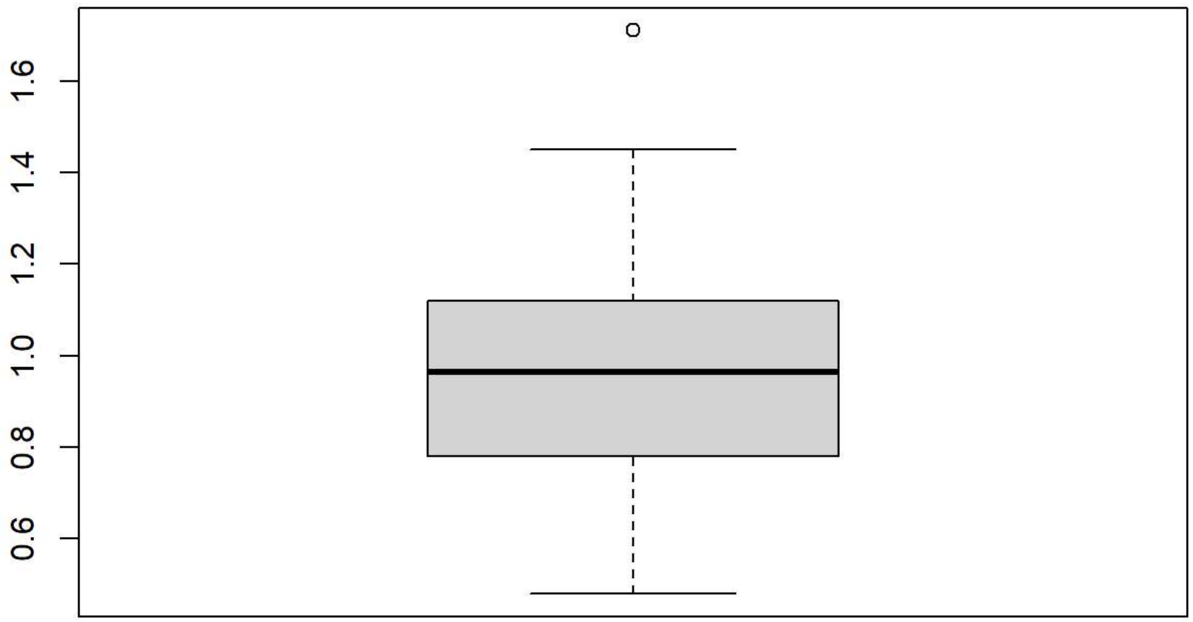
```
## [1] "变量名:"
## [1] "Malic"
## [1] "离群值:"
## [1] 5.80 5.51 5.65
## [1] "离群值行数"
## [1] 124 138 174
```



```
## [1] "变量名:"
## [1] "Magnesium"
## [1] "离群值:"
## [1] 151 139 136 162
## [1] "离群值行数"
## [1] 70 74 79 96
```



```
## [1] "变量名:"
## [1] "Intensity"
## [1] "离群值:"
## [1] 10.80 13.00 11.75
## [1] "离群值行数"
## [1] 152 159 160
```



```
## [1] "变量名:"
## [1] "Hue"
## [1] "离群值:"
## [1] 1.71
## [1] "离群值行数"
## [1] 116
```

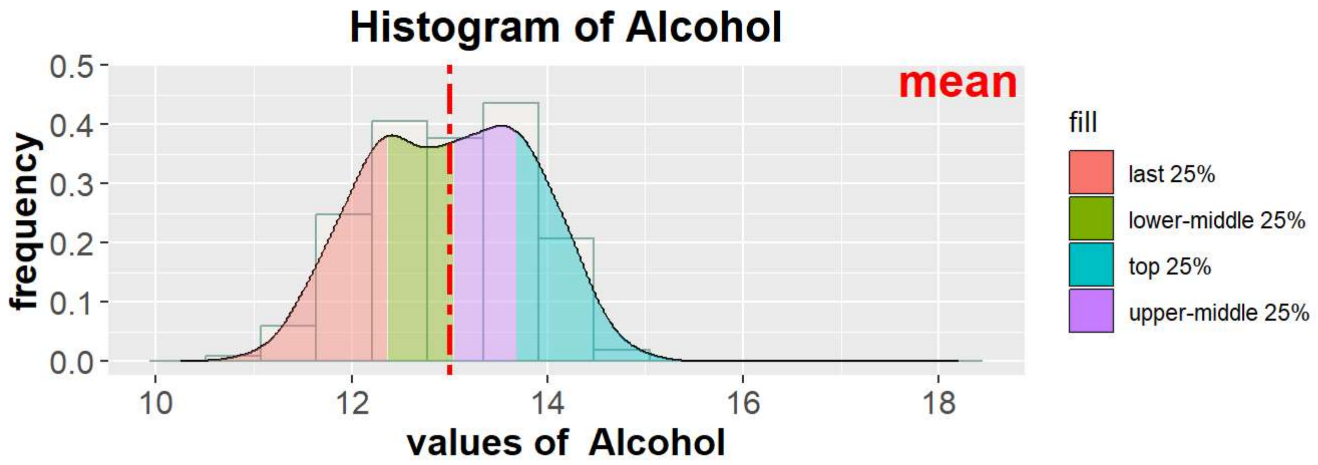
判断有偏

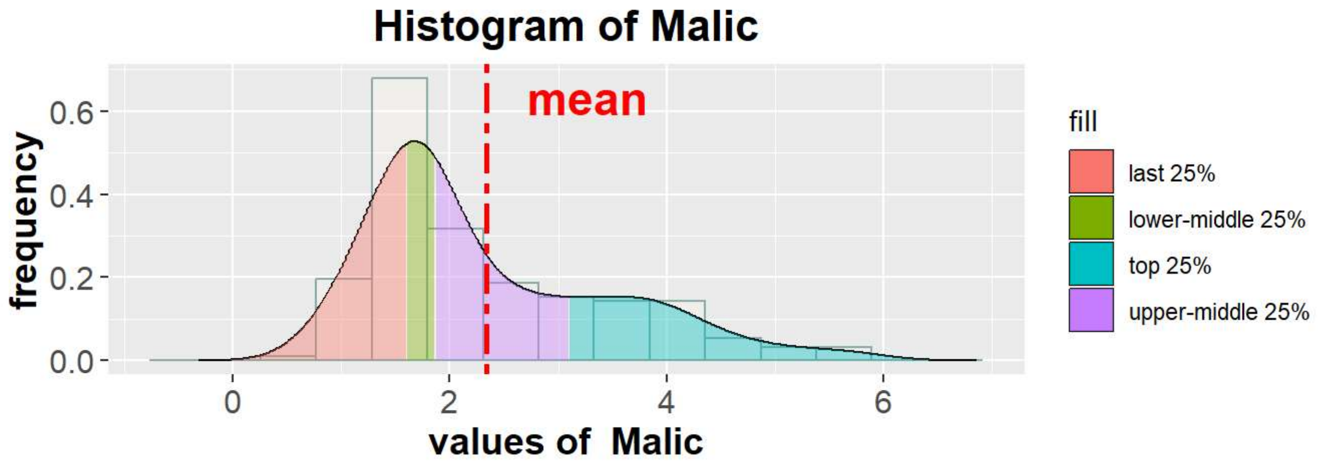
b. 请选用ggplot2中适当的图表类型，展示每个变量的样本分布是否有偏，以及相关图标的格式，如颜色，标题，图例等等。

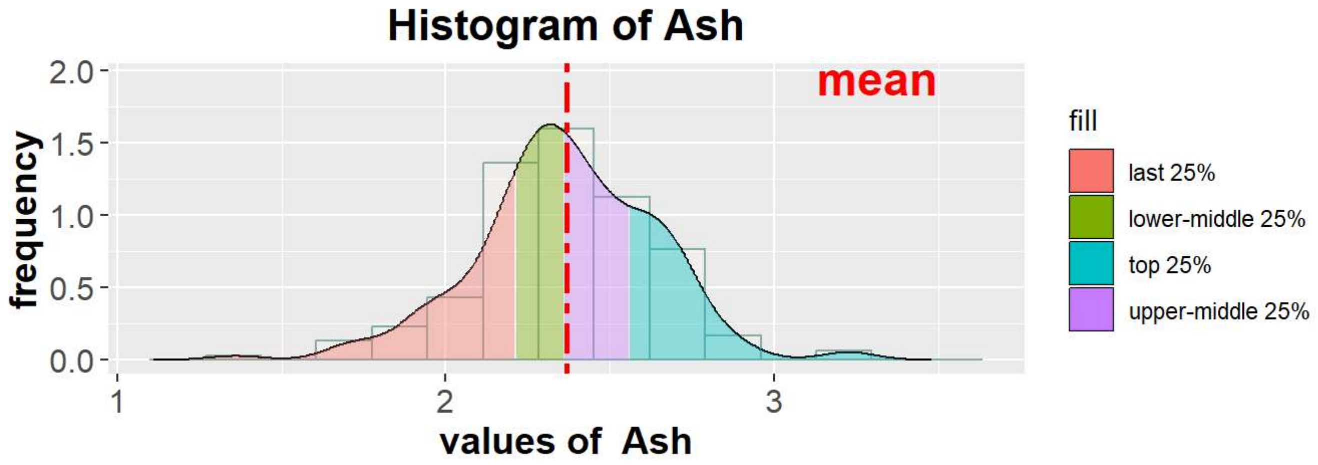
```
# 定义函数令其绘制样本分布
hist_line<-function(x,i) {
  name<- names(x);
  TM <- as.data.frame(x[,i]);
  colnames(TM) <-'tm';
  mean_vlaue<-mean(x[,i]);
  five<-fivenum(x[,i]);
  dense = data.frame(density(TM$tm)[c('x','y')]);
  pic<- ggplot(TM, aes(x =tm))+
    geom_histogram(aes(y=..density..),#纵坐标是密度。类似也可以将纵坐标设置为频数(count)
                   color="#88ada6", fill="#fffbf0",#边框与填充色，可以不设置
                   alpha=.25,# 透明度，可以不设置
                   bins =15,#柱子的宽度。类似得也可以设置柱子的个数，如bins = 30
                   center =0)+#柱子与对应横坐标的相对位置。0是指居中对齐。1是指对应数字在柱子的右侧边线。可以不设置
    geom_density() +# 密度曲线
    geom_area(data = subset(dense,x < five[2]), aes(x, y, fill ="last 25%"),alpha=.4)+
    geom_area(data = subset(dense,x >= five[2] & x < five[3]), aes(x, y, fill ="lower-middle 25%"), alpha=.4)+
    geom_area(data = subset(dense,x >= five[3] & x < five[4]), aes(x, y, fill ="upper-middle 25%"), alpha=.4)+
    geom_area(data = subset(dense,x >= five[4]), aes(x, y, fill ="top 25%"),alpha=.4)+
    labs(title=paste("Histogram of",name[i],sep =" "),
         #subtitle="with the density line",
         #caption = "caption",
         x = paste('values of ',name[i],sep =" "), y ='frequency')+
    theme(plot.title = element_text(size =16, face ="bold", hjust =0.5),
          plot.subtitle = element_text(size =12, face ="bold", hjust =0.5),
          plot.caption = element_text(size =12, face ="italic"),
          axis.text = element_text(size=12),# 坐标轴上的文字
          axis.title = element_text(size=14, face="bold"))+# 坐标轴标题
    geom_vline(xintercept = mean_vlaue,linetype ="twodash",color="red",size =1,)+
    annotate(geom ="text", fontface ="bold", color="red",
             x = mean_vlaue*1.4,y=1.2*max(dense$y),
             label ='mean', size=6)
  return(pic)
}
for(i in 1:(floor(ncol(wine)/2)-1)){
  p1<-hist_line(wine,(i*2))
  p2<-hist_line(wine,(i*2)+1)
  grid.arrange(p1, p2, nrow =2,ncol=1)
}
```

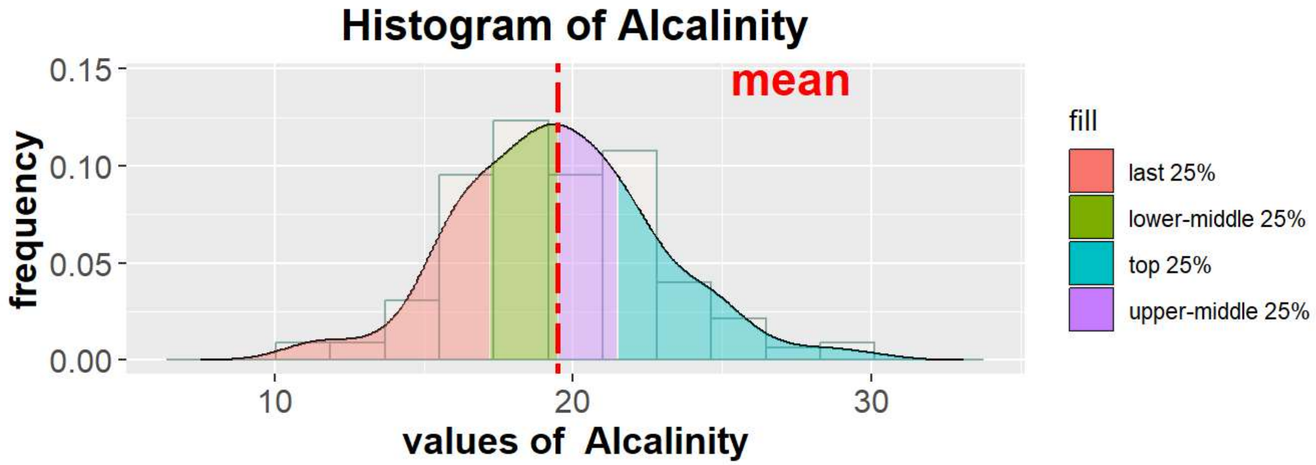
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

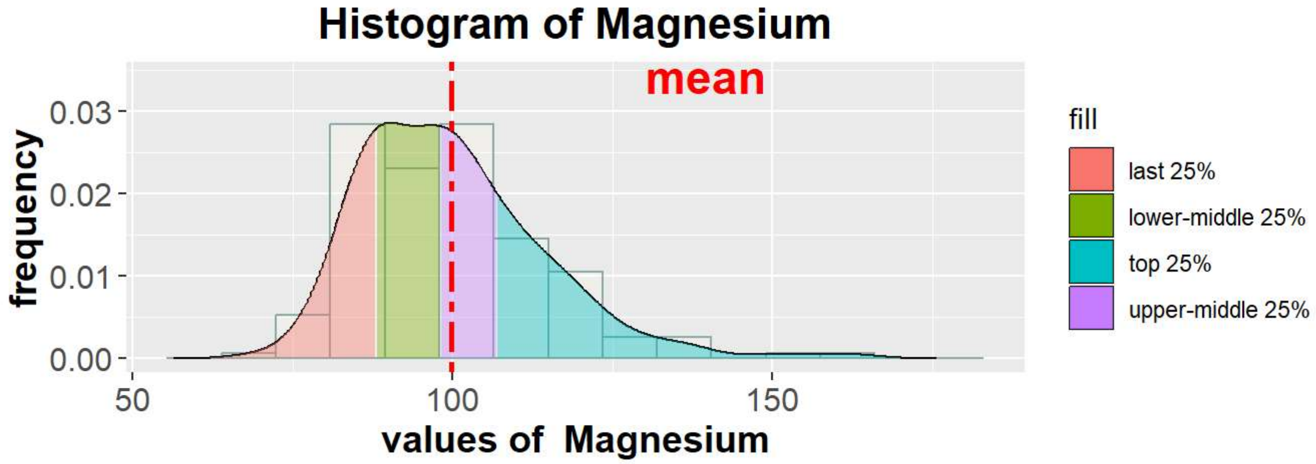
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```

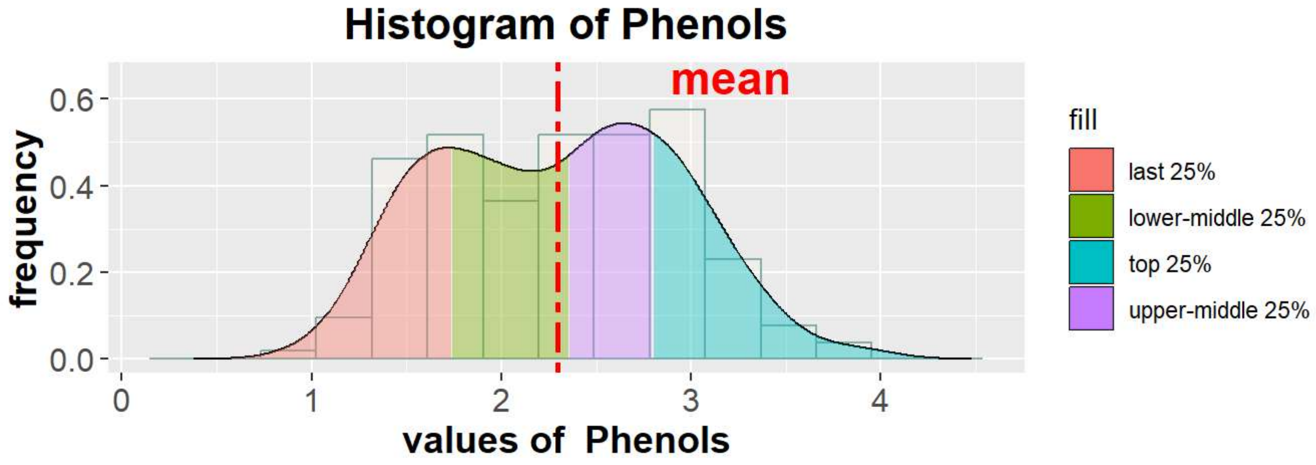

Histogram of Alcohol
mean
frequency
0.5
0.4
0.3
0.2
0.1
0.0
10
12
14
16
18
values of Alcohol
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%

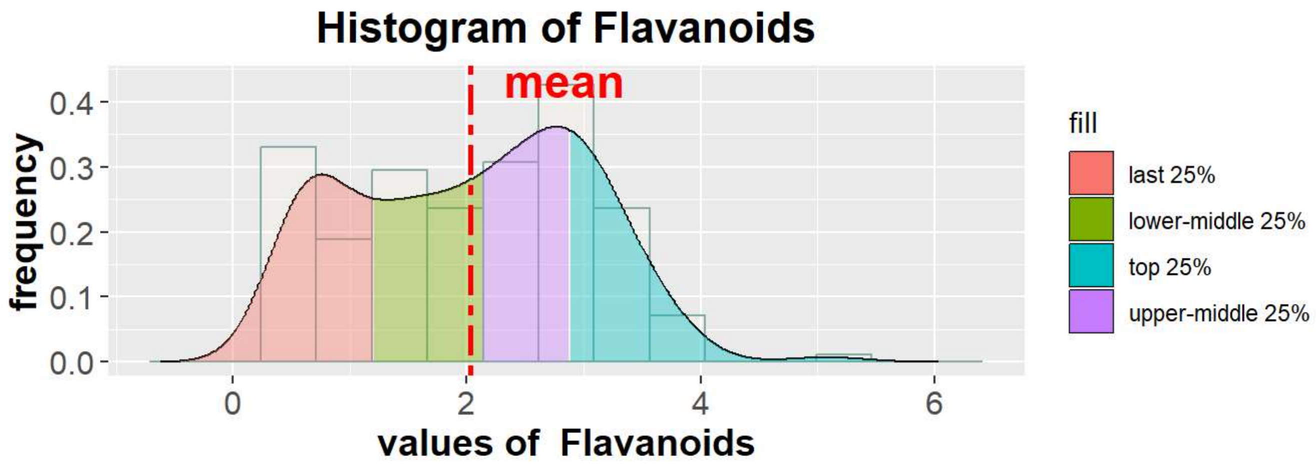

Histogram of Malic
mean
frequency
0.6
0.4
0.2
0.0
0
2
4
6
values of Malic
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%

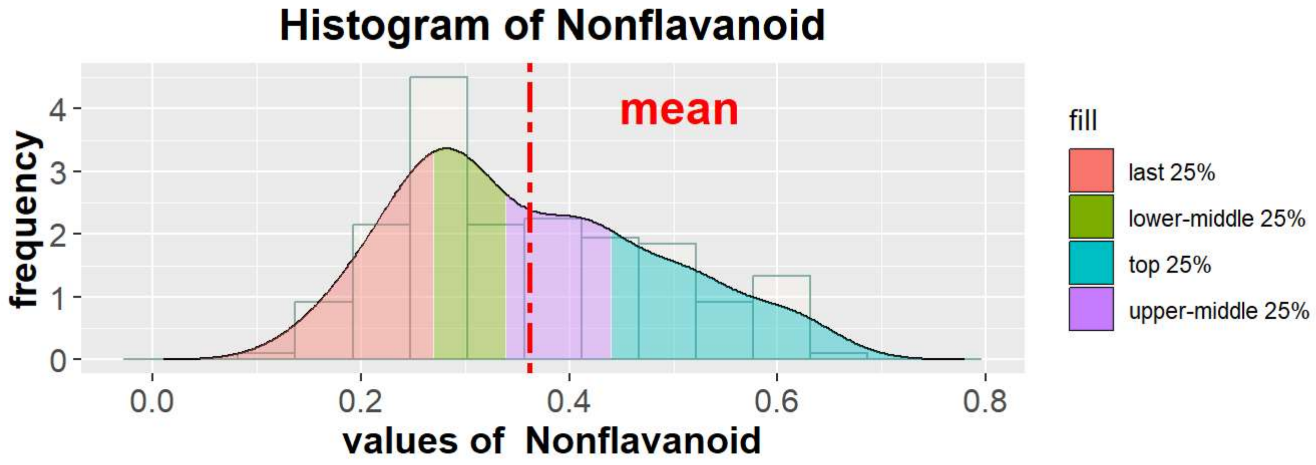

Histogram of Ash
mean
frequency
2.0
1.5
1.0
0.5
0.0
1
2
3
values of Ash
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%

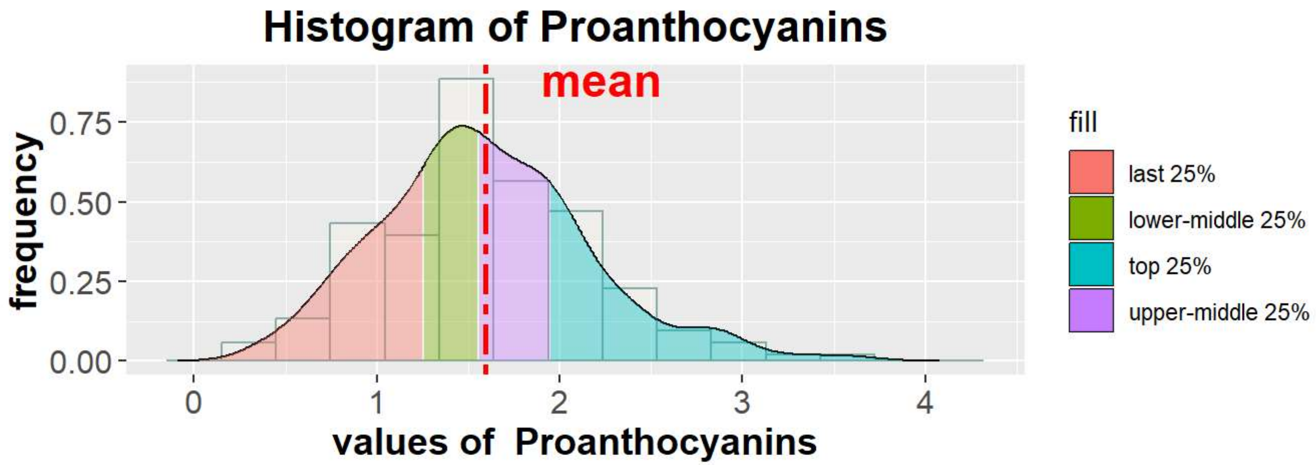

Histogram of Alcalinity
mean
frequency
0.15
0.10
0.05
0.00
10
20
30
values of Alcalinity
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%

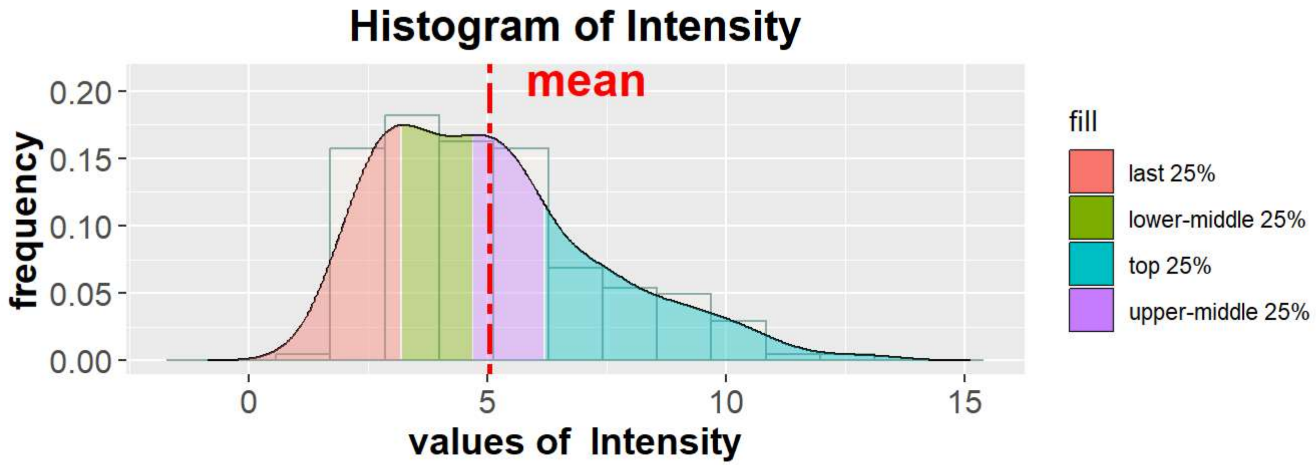

Histogram of Magnesium
mean
frequency
0.03
0.02
0.01
0.00
50
100
150
values of Magnesium
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%

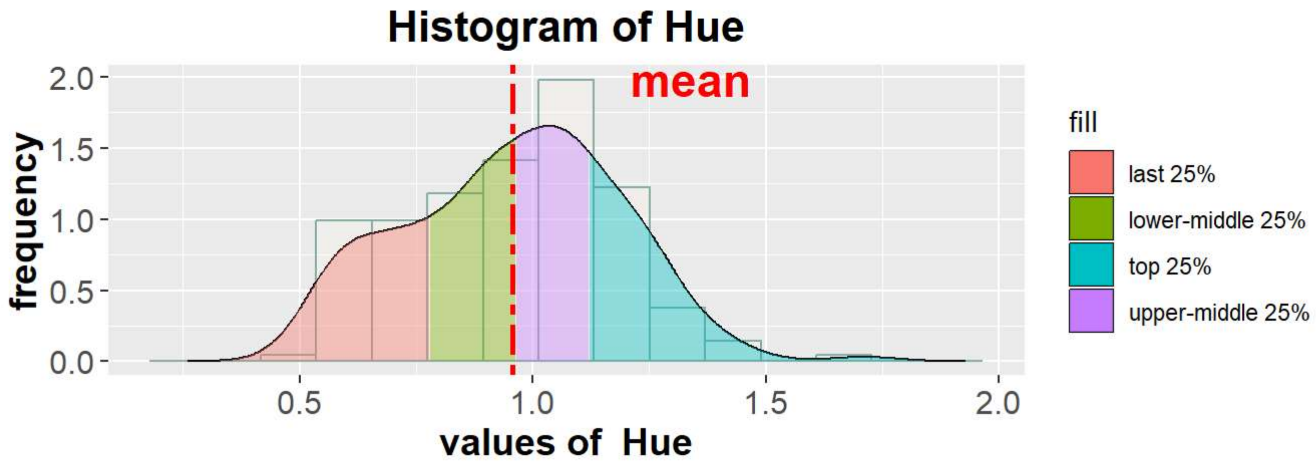

Histogram of Phenols
mean
frequency
0.6
0.4
0.2
0.0
0
1
2
3
4
values of Phenols
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%

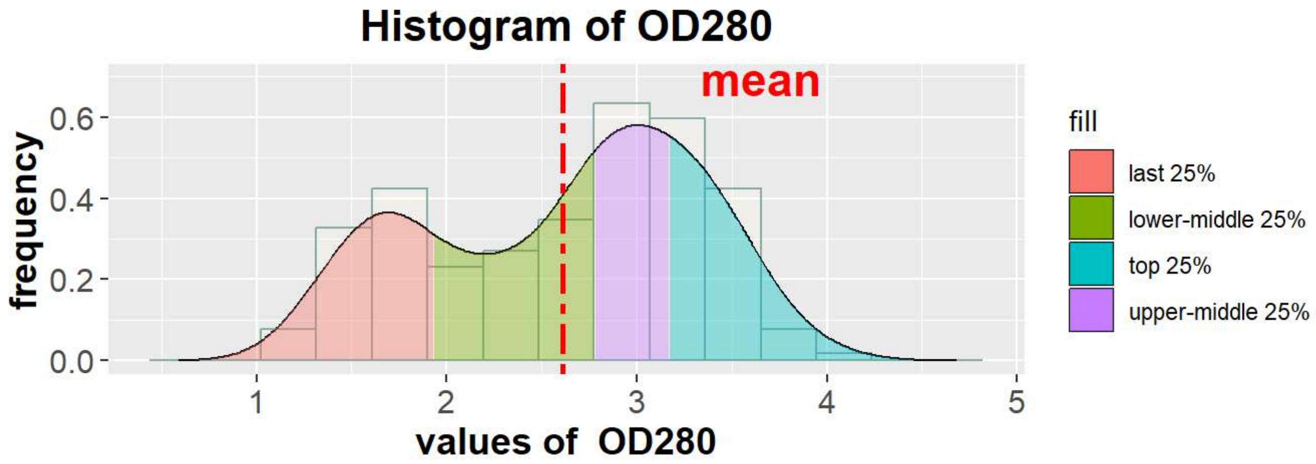

Histogram of Flavanoids
mean
frequency
0.4
0.3
0.2
0.1
0.0
0
2
4
6
values of Flavanoids
fill
🟥 last 25%
🟩 lower-middle 25%
🟦 top 25%
🟪 upper-middle 25%


Histogram of Nonflavanoid
mean
frequency
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
values of Nonflavanoid
fill
🟥 last 25%
🟩 lower-middle 25%
🟦 top 25%
🟪 upper-middle 25%


Histogram of Proanthocyanins
mean
frequency
0.75
0.50
0.25
0.00
0
1
2
3
4
values of Proanthocyanins
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%


Histogram of Intensity
mean
frequency
0.20
0.15
0.10
0.05
0.00
0
5
10
15
values of Intensity
fill
last 25%
lower-middle 25%
top 25%
upper-middle 25%



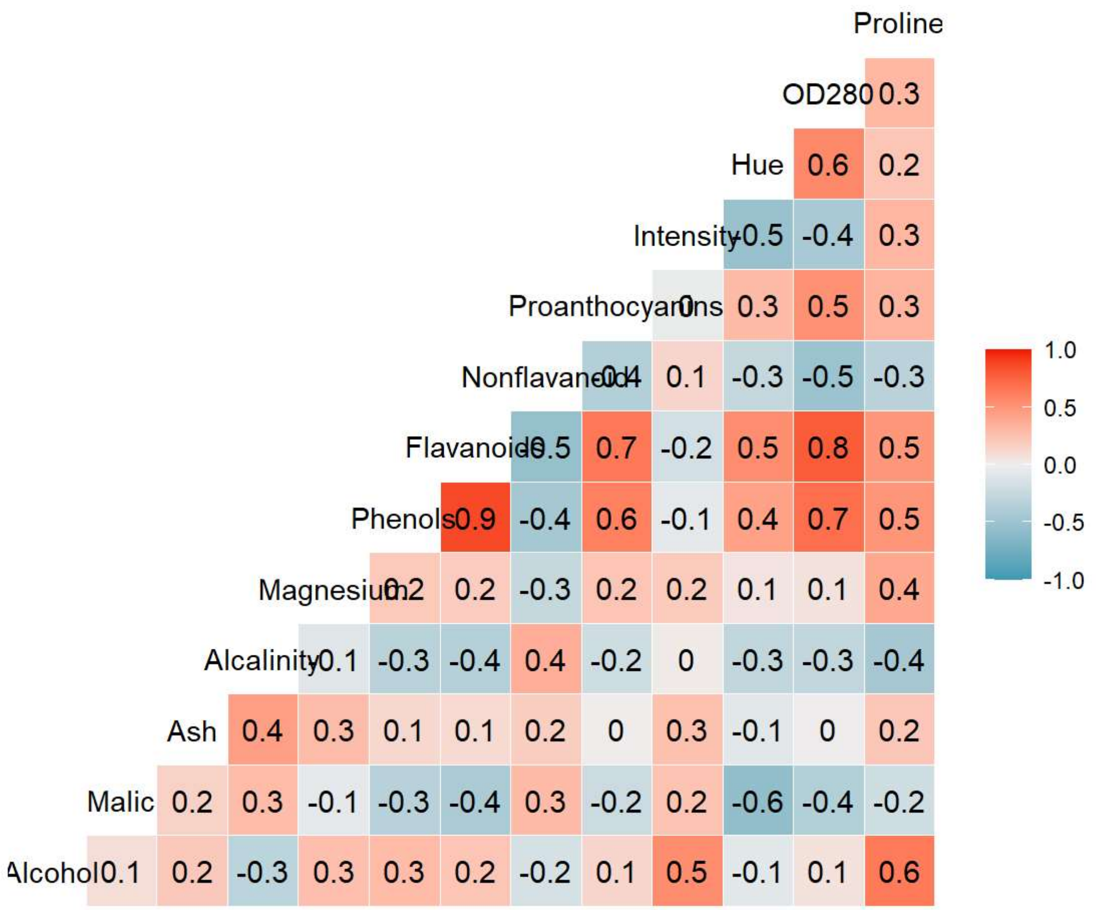


从分布图来看，中位数与均值均较为接近，认为数据不存在有偏

相关性

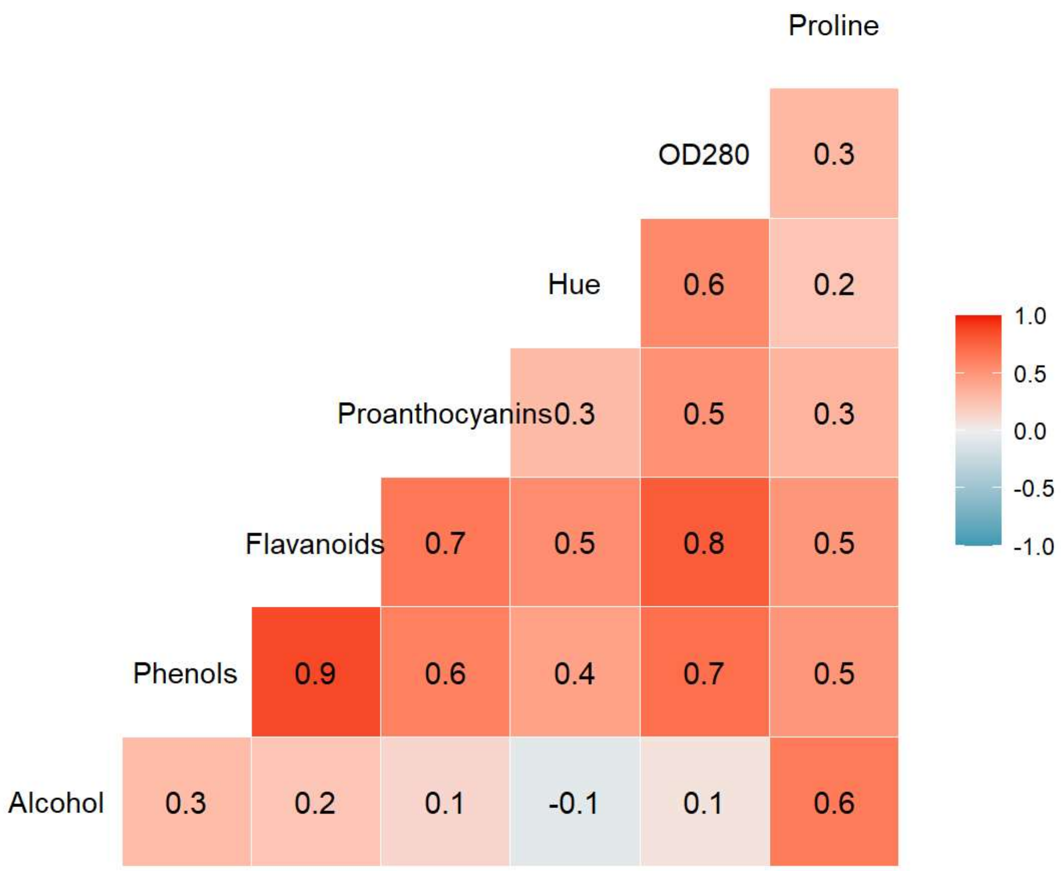
c. 请选用合适的方式，计算并展示wineTrain数据集中所有变量的两两相关性。你哪些变量之间的相关性比较高 ？

```
ggcorr(wineTrain,label = T,digits = 2,hjust=0.5)
```



> 相关性较高的变量如下图所示

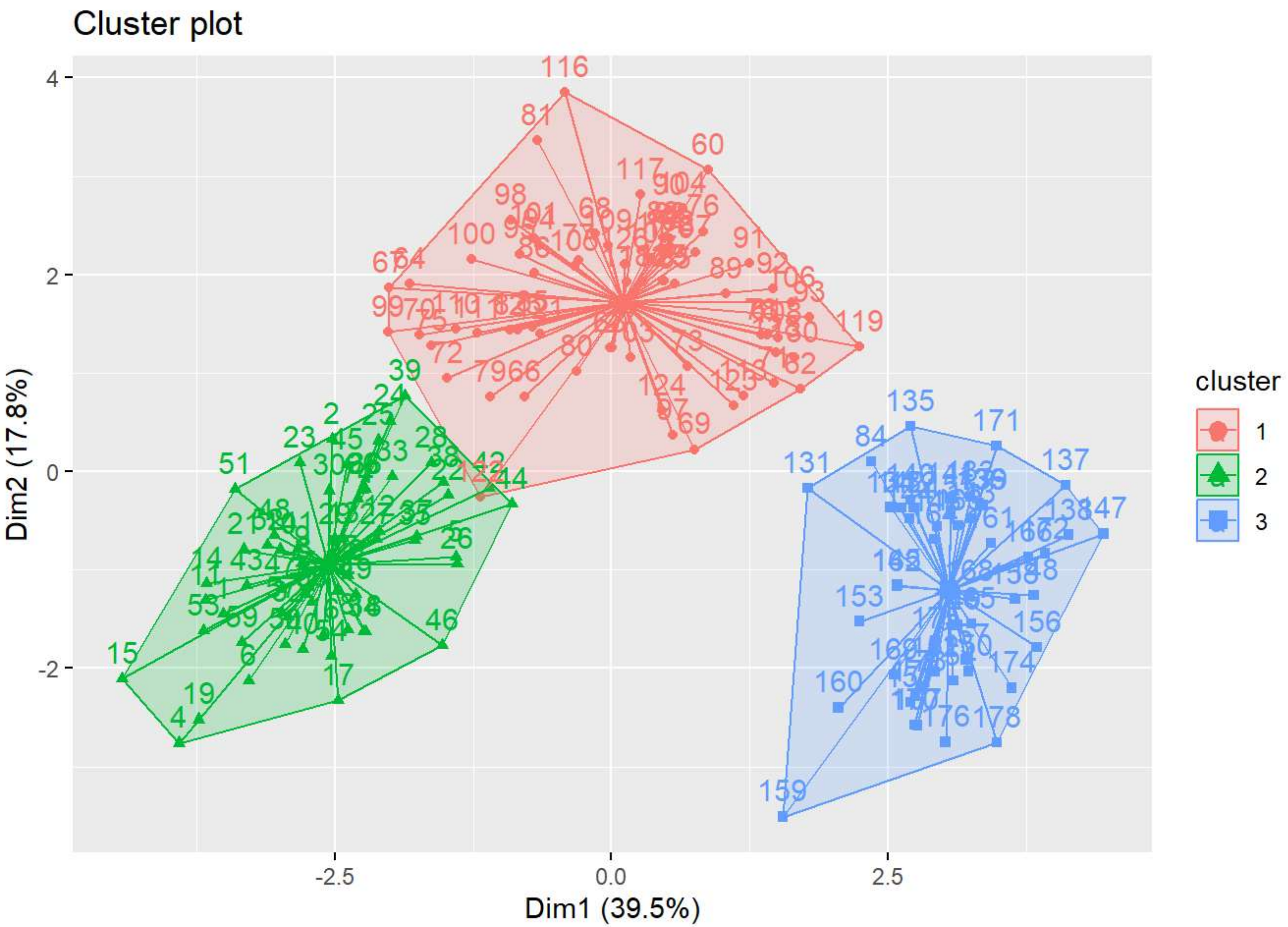
```
factor_name <- c("Alcohol","Proanthocyanins","Hue","OD280","Flavanoids","Phenols","Proline")
idx <- which(names(wineTrain) %in% factor_name)
ggcorr(wineTrain[,idx],label = T,digits = 2,hjust=0.5)
```



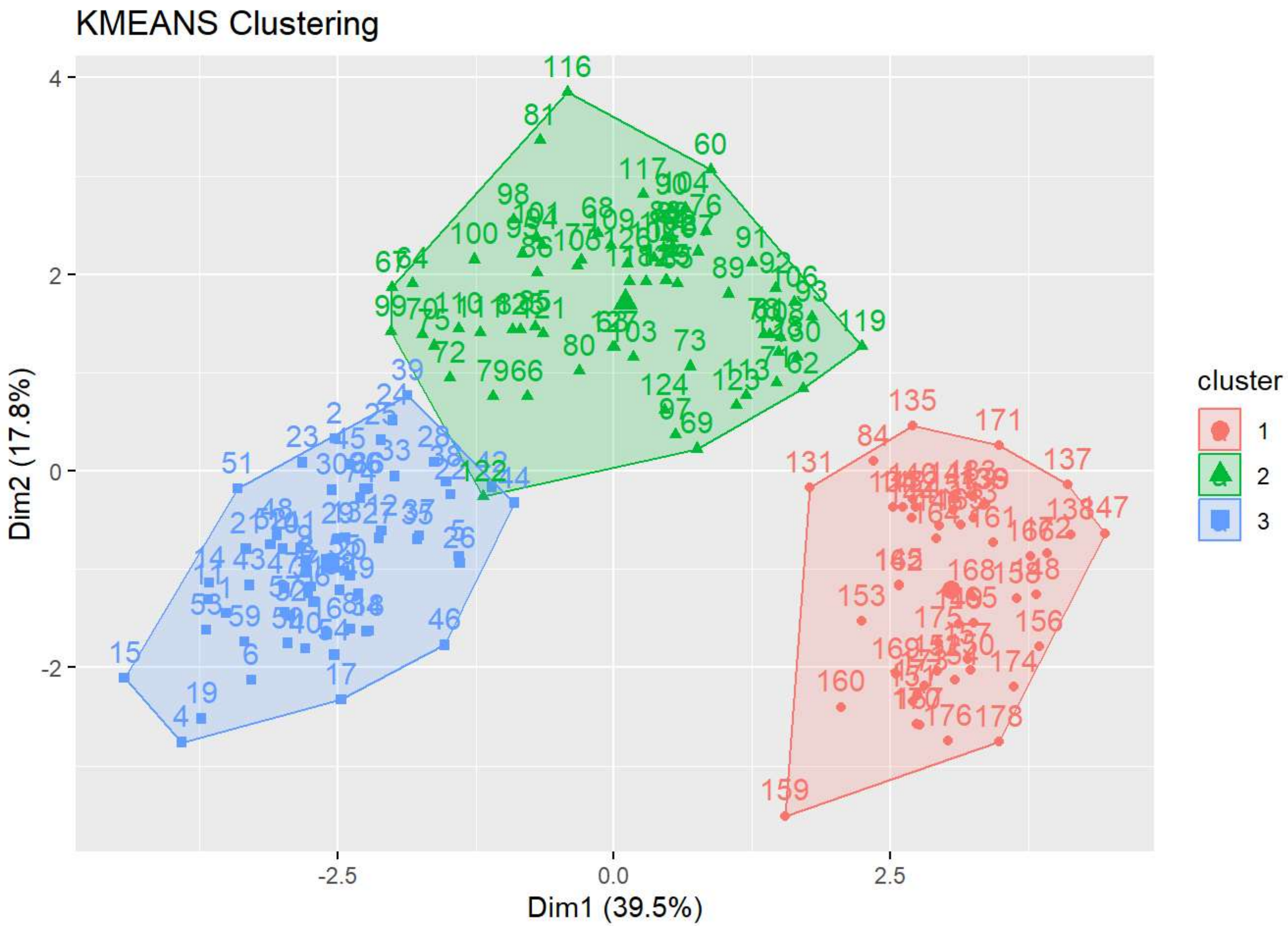
聚类

d. 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？

```
set.seed(2020111142)
df <- scale(wine)
distance <- get_dist(df)
km_result <- kmeans(df, centers =3, nstart =25)
fviz_cluster(km_result, data = df,star.plot =TRUE)
```



```
res.km <- eclust(df,″kmeans″, nstart =25)
```



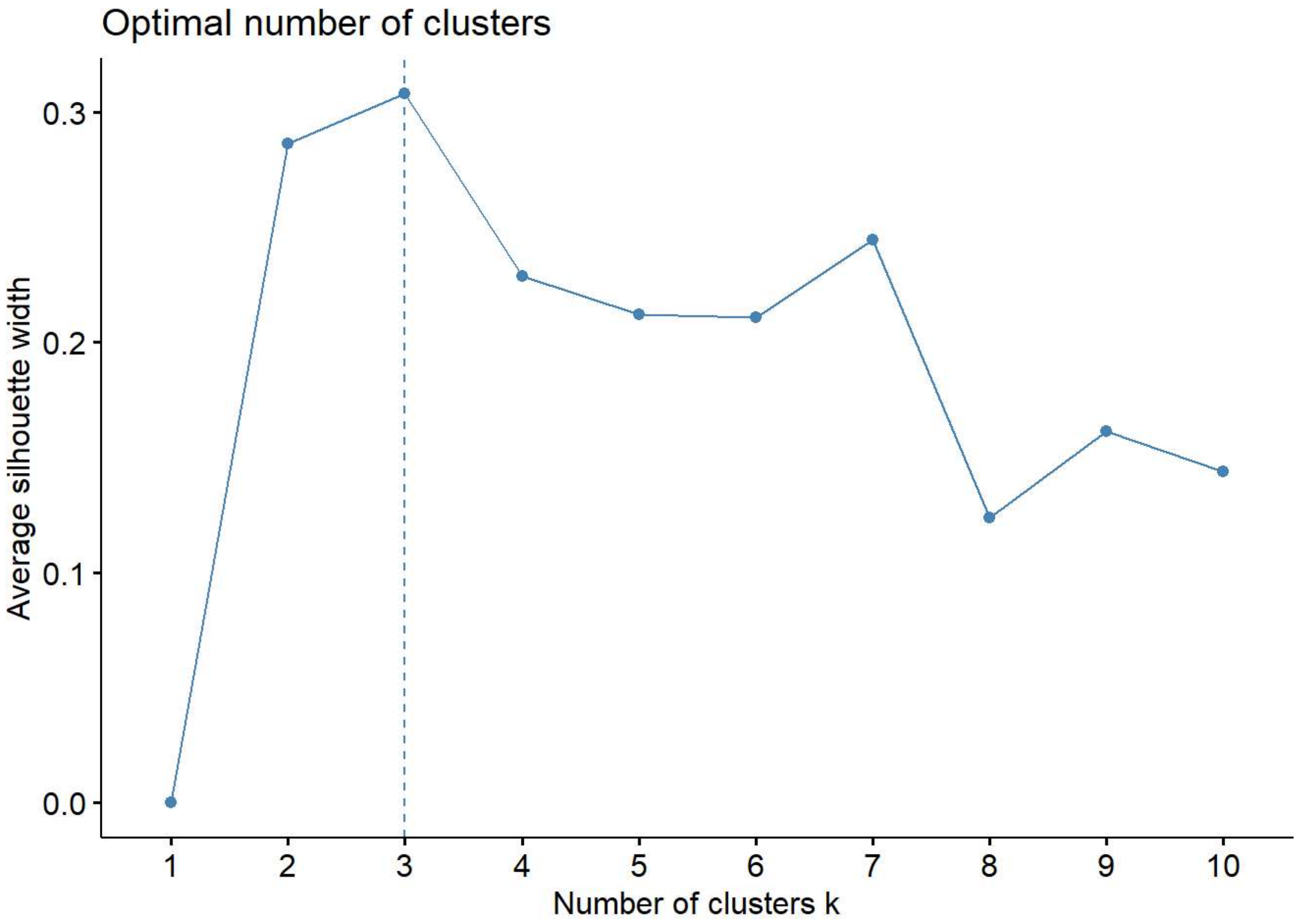
> 基本无交叉，我认为分类效果较好。

确定最优组数

e. 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 ?

silhouette 统计量

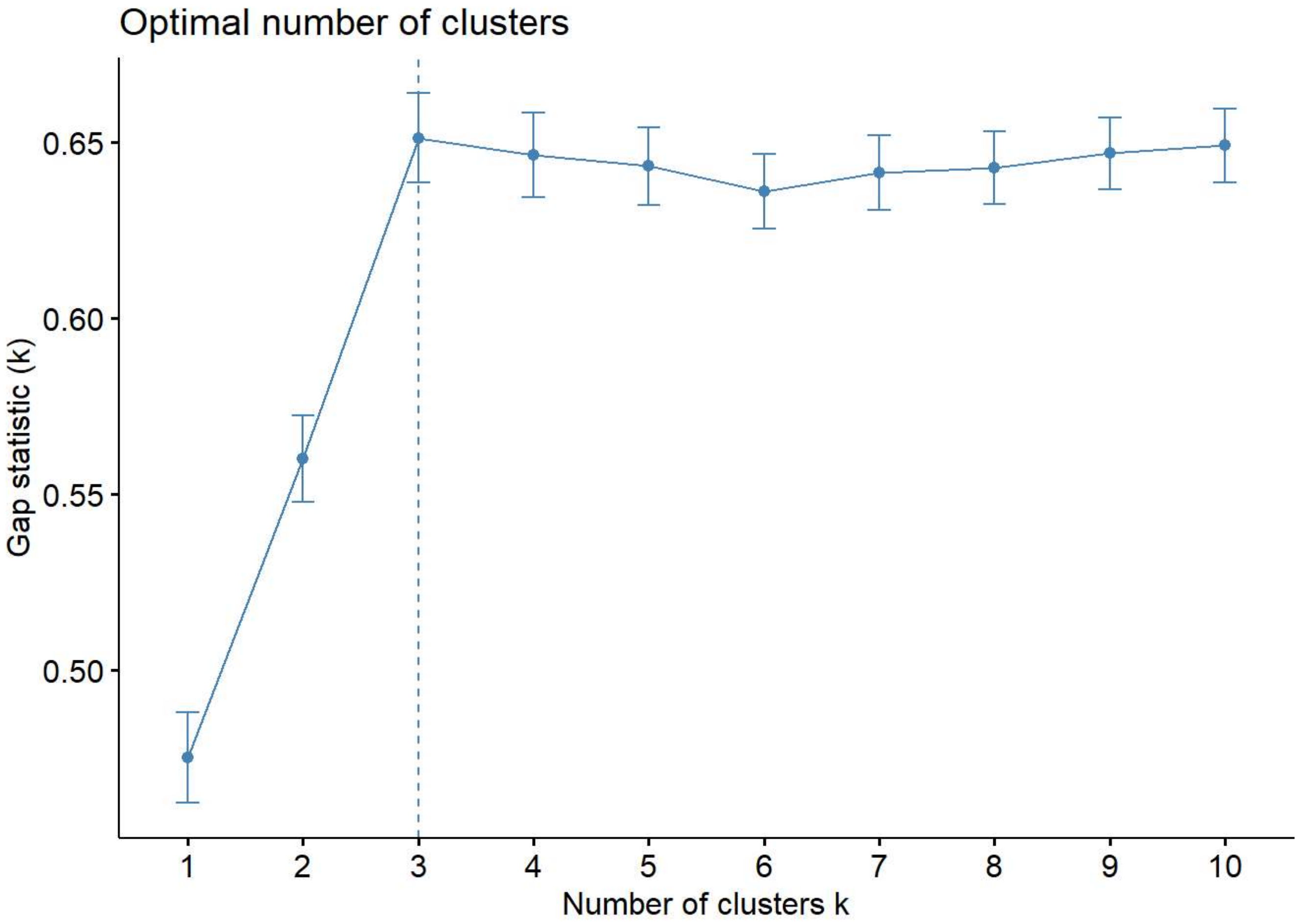
```
fviz_nbclust(df, kmeans, method ="silhouette")
```



> 最优值为 3

Gap统计量

```
gap_stat <- clusGap(df, FUN = kmeans, nstart =25,K.max =10, B =50)
fviz_gap_stat(gap_stat)
```



> 最优值为 3

> 两种方法给出的最优组数相同;组的个数与原始数据集中wine 中的变量Cultivar 的可能取值也相同

混淆矩阵

f. 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。

```
set.seed(2020111142)
table(km_result$cluster,wine$Class)
```

```
##
##      1  2  3
##   1  0 68  0
##   2 59  2  0
##   3  0  1 48
```

调整

```
cluster <- ifelse(km_result$cluster==2,1,
                  ifelse(km_result$cluster==1,2,
                         3))
table(cluster,wine$Class)
```

```
##
## cluster  1  2  3
##       1 59  2  0
##       2  0 68  0
##       3  0  1 48
```

```
sum(cluster==wine$Class)/length(cluster)
```

```
## [1] 0.9831461
```

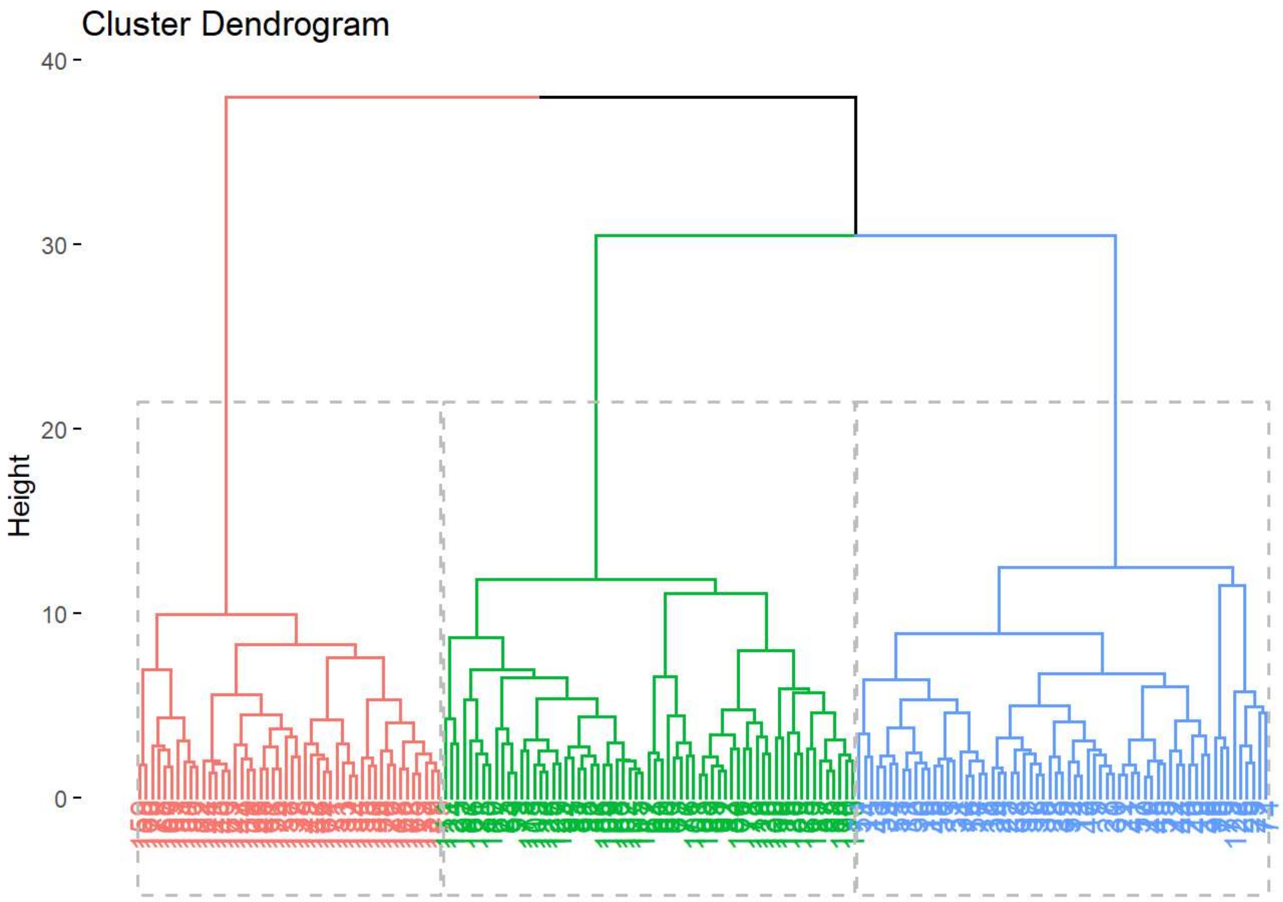
层次聚类

g. 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？

```
res.hc <- eclust(df,"hclust")# compute hclust
```

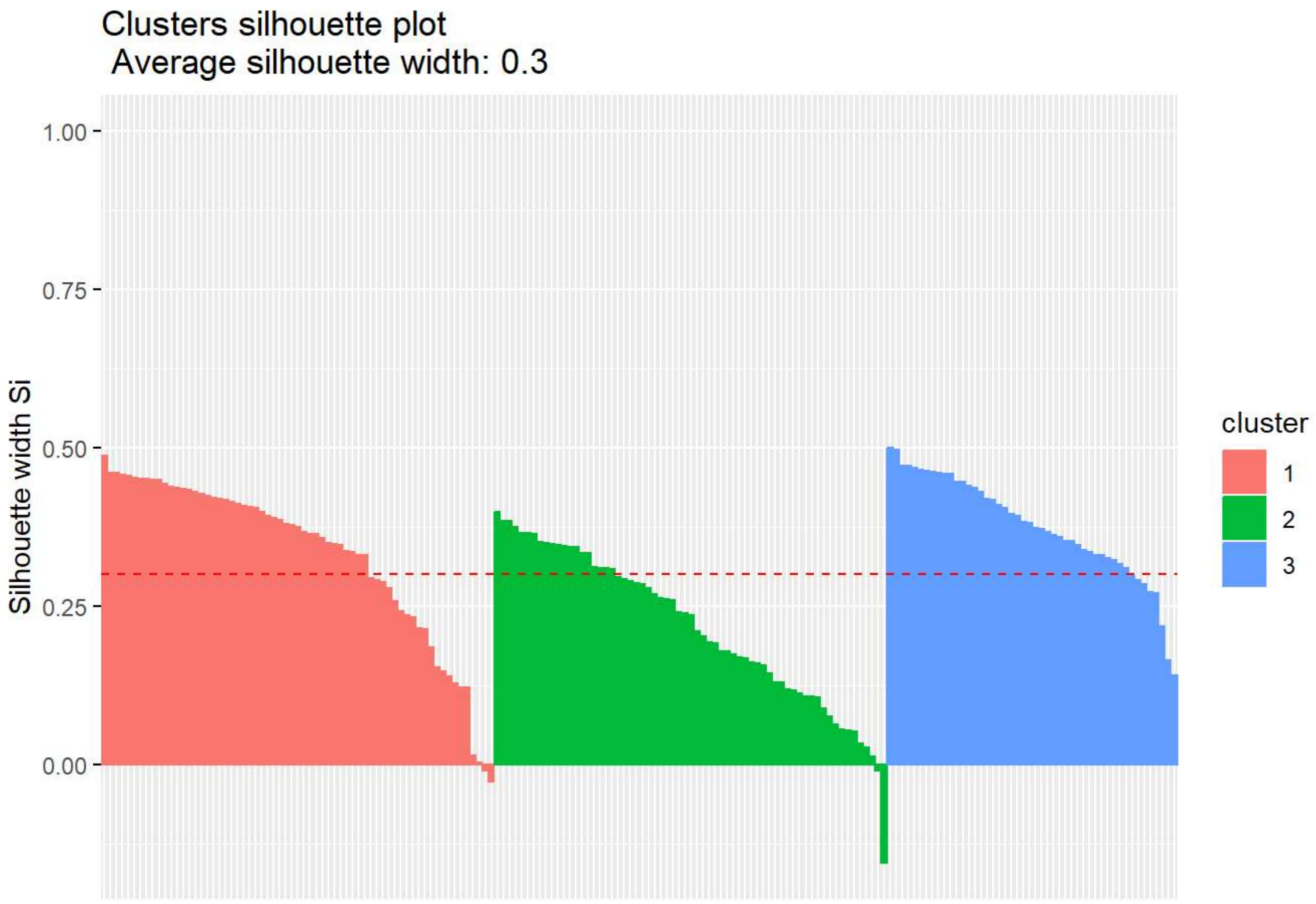
```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use ″none″ instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <←]8;;https://github.com/kassambara/factoextra/issues•https://github.com/kassambara/factoextra/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

```
fviz_dend(res.hc, rect =TRUE)# dendrogam
```

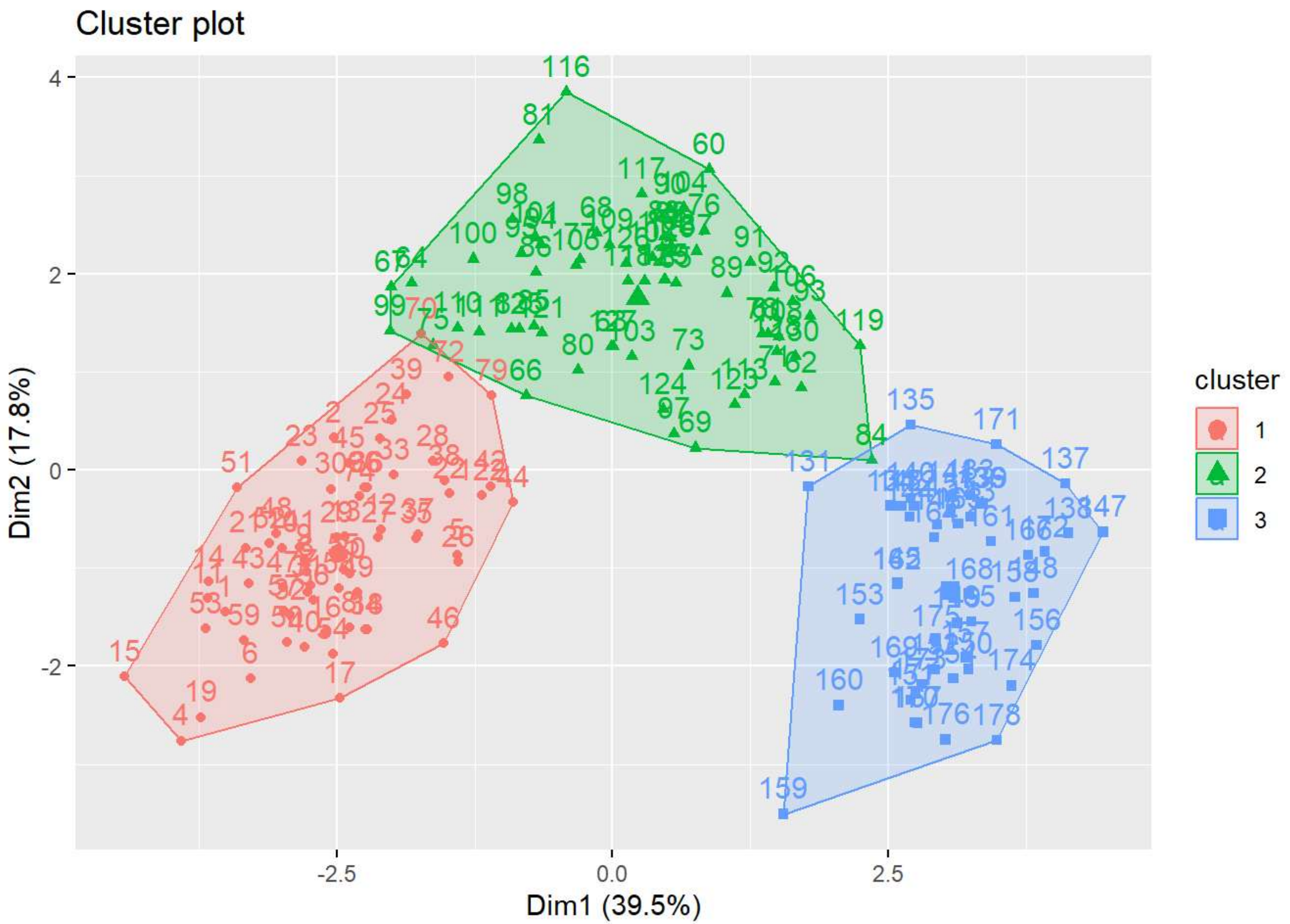


```
fviz_silhouette(res.hc)# silhouette plot
```

```
##   cluster size ave.sil.width
## 1       1   65          0.33
## 2       2   65          0.22
## 3       3   48          0.38
```



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



```
### 层次聚类法
table(res.hc$cluster,wine$Class)
```

```
##
##       1  2  3
##    1 59  6  0
##    2  0 65  0
##    3  0  0 48
```

```
sum(res.hc$cluster==wine$Class)/length(cluster)
```

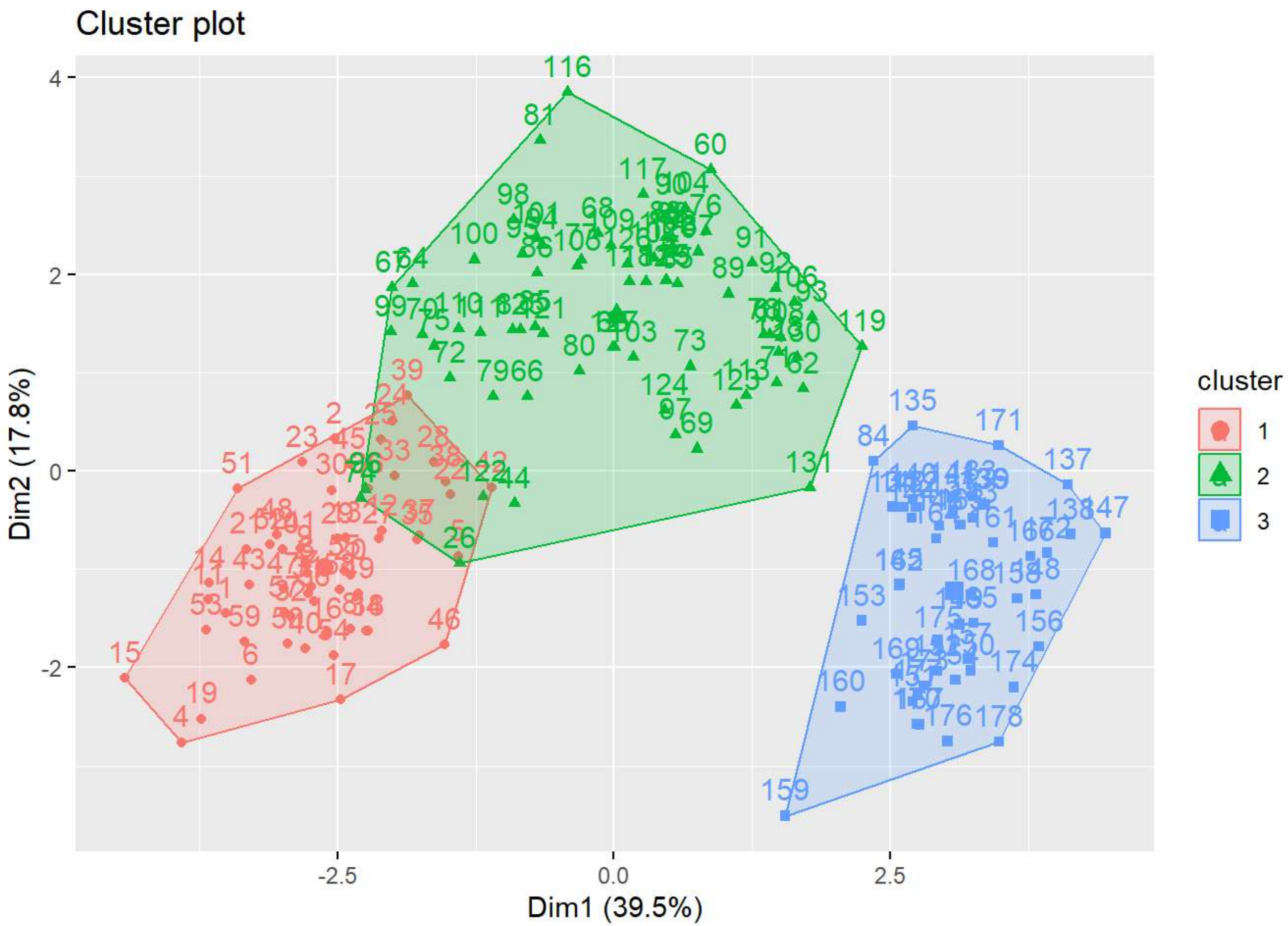
```
## [1] 0.9662921
```

> K-means 准确率较高

EM算法聚类

h. 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？

```
EM_result<-Mclust(df, G = 3)
fviz_cluster(EM_result) # scatter plot
```



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

```
##
##       1  2  3
##    1 57  0  0
##    2  2 70  1
##    3  0  1 47
```

```
sum(EM_result$classification==wine$Class)/length(cluster)
```

```
## [1] 0.9775281
```

> K-means效果更好，但两者准确率相近，都对初始值很敏感，容易陷入局部最优解。