hw6

Getong Zhong

2023-03-31

6.1

```
crime <- read.table("crime.txt", header = TRUE)


sd_crime <- sapply(crime, sd)
crime_sd <- sweep(crime, 2, sd_crime, FUN = "/")
kmeans(crime_sd, centers = 2)$centers * sd_crime
```

```
##     Murder     Rape  Robbery    Assault  Burglary     Theft  Vehicle
## 1 10.68095 446.6573 791.2308 621.098297  55.06786  680.4368 1922.174
## 2 14.67118 255.6647 448.6922   6.276854 305.87963 1408.9155  275.400
```

From the above results we can see that, when standardized with standard deviation, we can still observe that the centroids still show significant differences in most variables, but some variables have smaller differences between the two clusters. For example, the average Murder rates are much closer between the two clusters compared to the range-standardized results. In general, the comparision between the two methods results show that the two methods have some impact on the clustering results, as range-standardized method creating more separation between clusters.

6.2

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

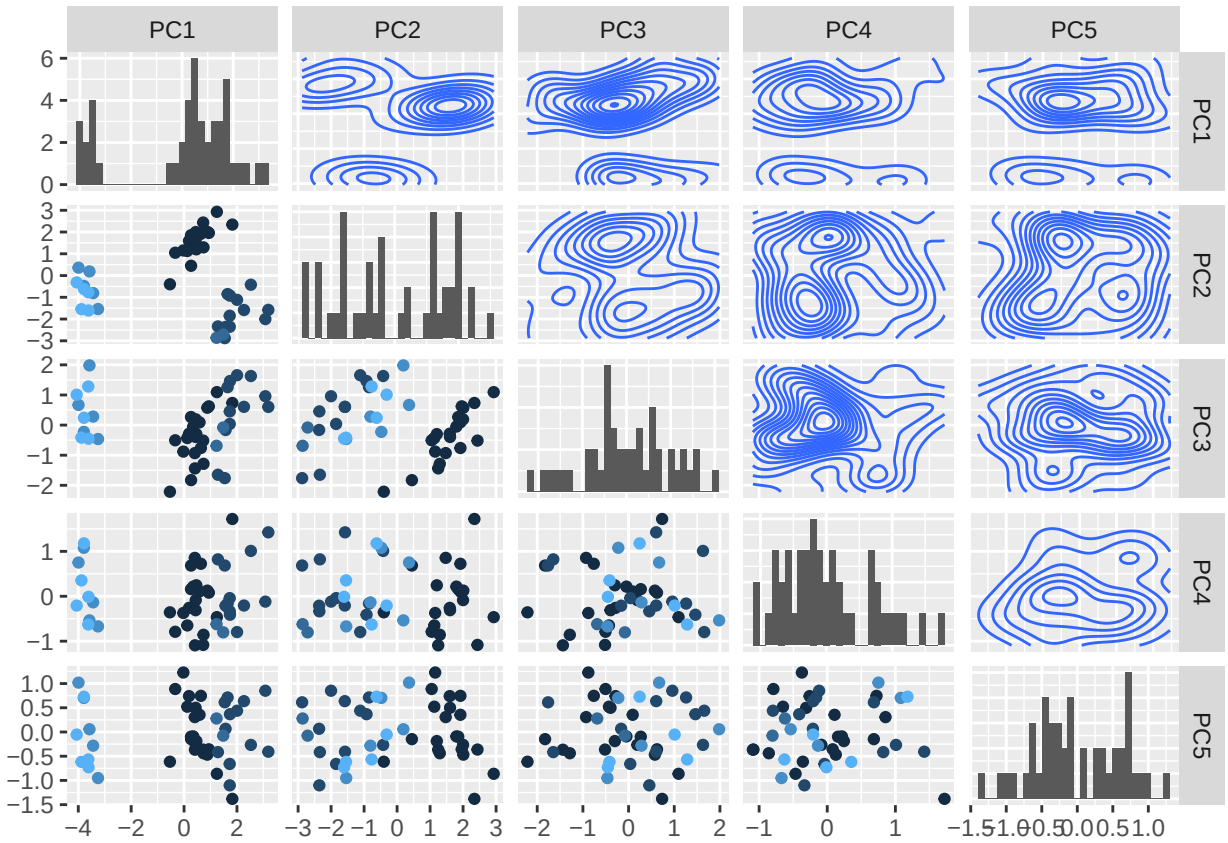
```
data("pottery", package = "HSAUR2")
pottery$kiln <- as.numeric(as.character(pottery$kiln))
attach(pottery)
pca_result <- prcomp(pottery, scale = TRUE)
pca_scores <- as.data.frame(pca_result$x[, 1:5])
pca_scores_with_kiln <- cbind(pca_scores, Kiln = kiln)
```

```
plot <- ggpairs(pca_scores_with_kiln, columns = 1:5,
                upper = list(continuous = "density"),
                diag = list(continuous = "barDiag"),
                lower = list(continuous = "points", combo = "box"),
                mapping = ggplot2::aes(color = Kiln, label = Kiln))


plot <- plot + theme(legend.position = "none")
print(plot)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
detach(pottery)
```

##6.3 From the boxplot, we can see that the level of SO2 in each cluster didn't varied much. In each of the box, the SO2 remain around the level of 20.

```r
data("USairpollution", package = "HSAUR2")
library(mclust)
```

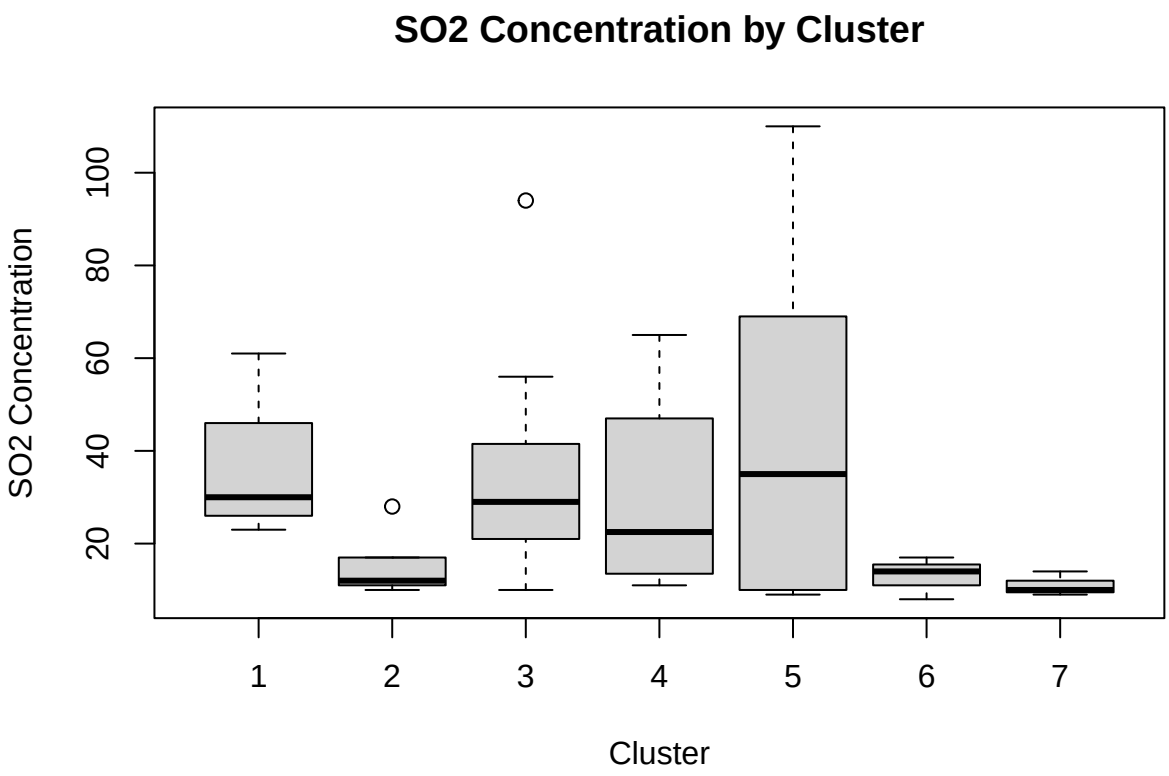
```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
```

```r
mclust_result <- Mclust(USairpollution[,-1])
summary(mclust_result)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 7 components:
##
##  log-likelihood  n  df       BIC       ICL
##       -830.9045 41 165 -2274.548 -2274.563
##
## Clustering table:
##  1  2  3  4  5  6  7
##  6  5 15  4  5  3  3
```

```r
USairpollution$Cluster <- mclust_result$classification

boxplot(SO2 ~ Cluster, data = USairpollution, main = "SO2 Concentration by Cluster"
        , xlab = "Cluster", ylab = "SO2 Concentration")
```



And then lets look at the results from a formal siginificant test, ANOVA.In the ANOVA table, the p-value of 0.41, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis and we do not have enough evidence to conclude that there is a significant difference in SO2 concentrations between the clusters based on the six climate and ecology variables.

```
anova_result <- aov(SO2 ~ Cluster, data = USairpollution)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Cluster      1    384   384.5   0.692   0.41
## Residuals   39  21653   555.2
```