다변량통계분석 Practice 4

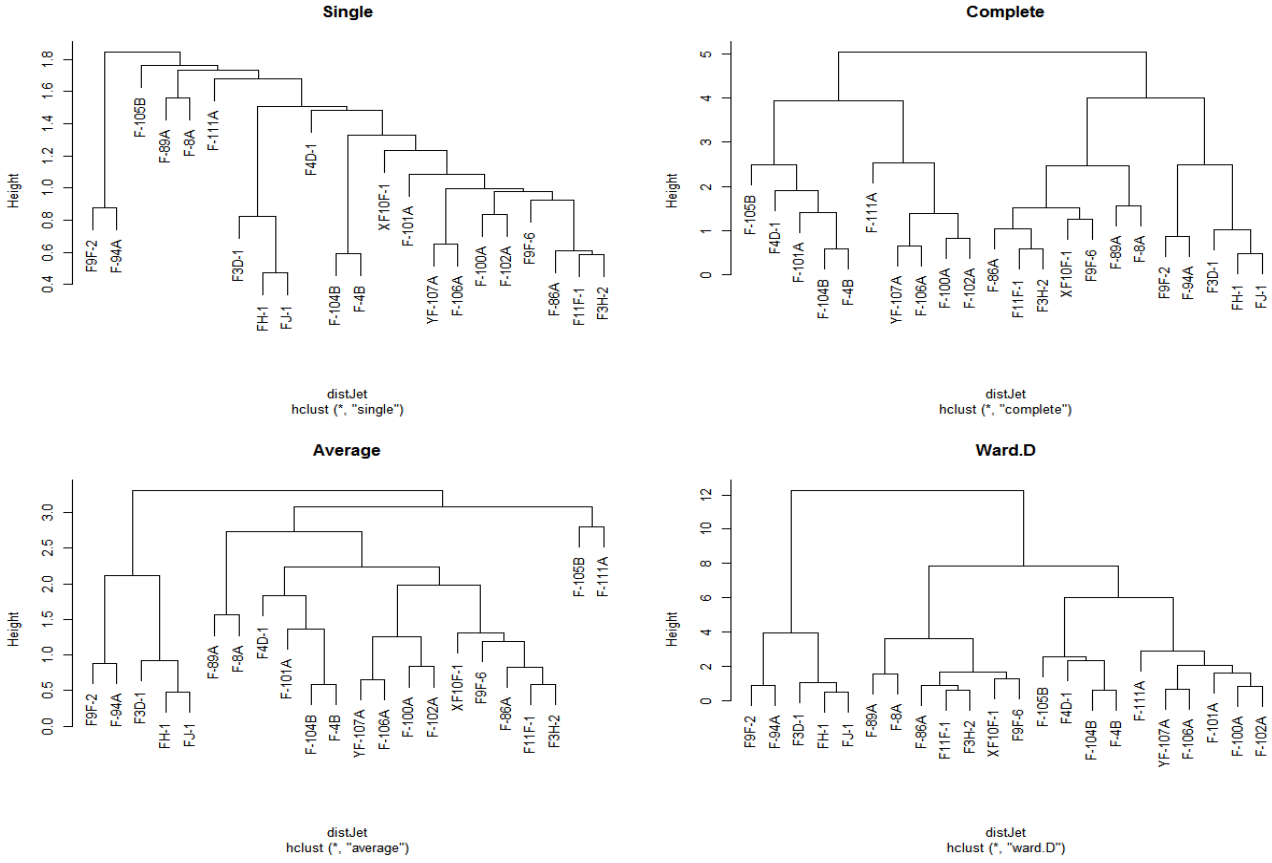
빅데이터경영MBA / U2016040 / 김우현

22개 미국 전투기에 대한 6개 변수값이 jet.csv에 저장되어 있다. 각 변수는 아래와 같다.

- FFD: 처음 비행 날짜
- SPR: 단위무게 당 출력에 비례하는 특정한 출력
- RGF: 비행범위 요인
- PLF: 비행기의 총 무게의 일부분으로서의 탑재량
- SLF: 일관된 무게 요인
- CAR: 비행기가 항공모함에 착륙 가능여부

1. 계층적군집분석
   A. FFD와 CAR를 제외한 변수를 표준화 한 후 계층적 군집화를 시행하고 덴드로그램을 그리시오.

```
jet = read.csv("jet.csv", header = T)
rownames(jet) = jet$X
jet = jet[, -c(1,2,7)]   # FFD, CAR 제외.

jet_s = scale(jet)
distJet = dist(jet_s)

hc1 = hclust(distJet, method = "single")
plot(hc1, main = "Single")
hc2 = hclust(distJet, method = "complete")
plot(hc2, main = "Complete")
hc3 = hclust(distJet, method = "average")
plot(hc3, main = "Average")
hc4 = hclust(distJet, method = "ward.D")
plot(hc4, main = "Ward.D")
```



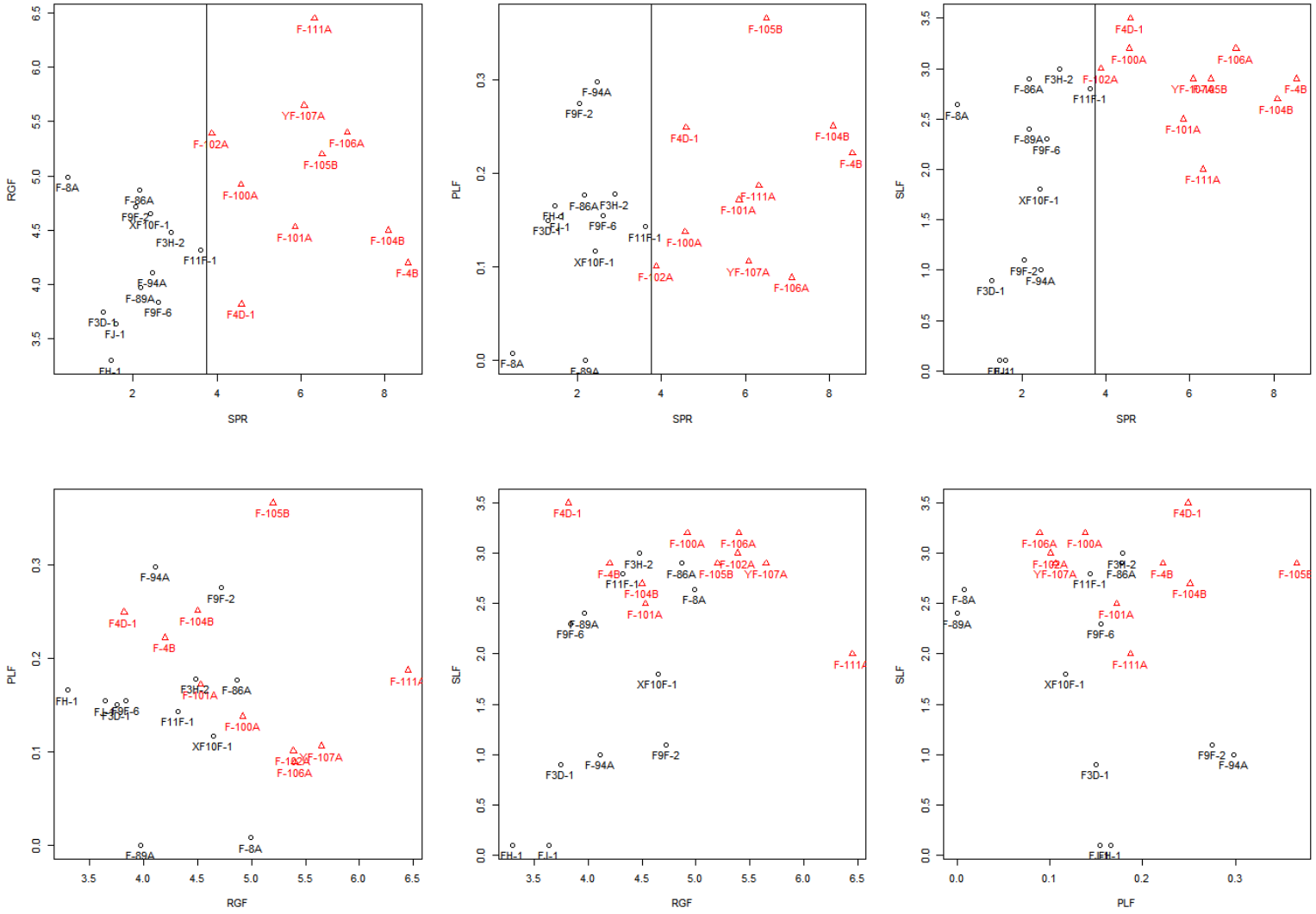
B. A의 결과를 사용해 두 개의 집단으로 관측치를 분류하고 각 집단의 특징을 원변수 관점에서 비교하시오.

2개의 군집으로 나누기 위해서는 complete linkage 또는 ward method를 선택하는 것이 좋다.

(1) complete linkage 사용하는 경우.

```
result_hc = cutree(hc2, k=2)

plot(jet$SPR, jet$RGF, col=result_hc, pch=result_hc, xlab = "SPR", ylab = "RGF")
text(jet$SPR, jet$RGF, labels = rownames(jet), col=result_hc, pos = 1)
abline(v = 3.75)
plot(jet$SPR, jet$PLF, col=result_hc, pch=result_hc, xlab = "SPR", ylab = "PLF")
text(jet$SPR, jet$PLF, labels = rownames(jet), col=result_hc, pos = 1)
abline(v = 3.75)
```

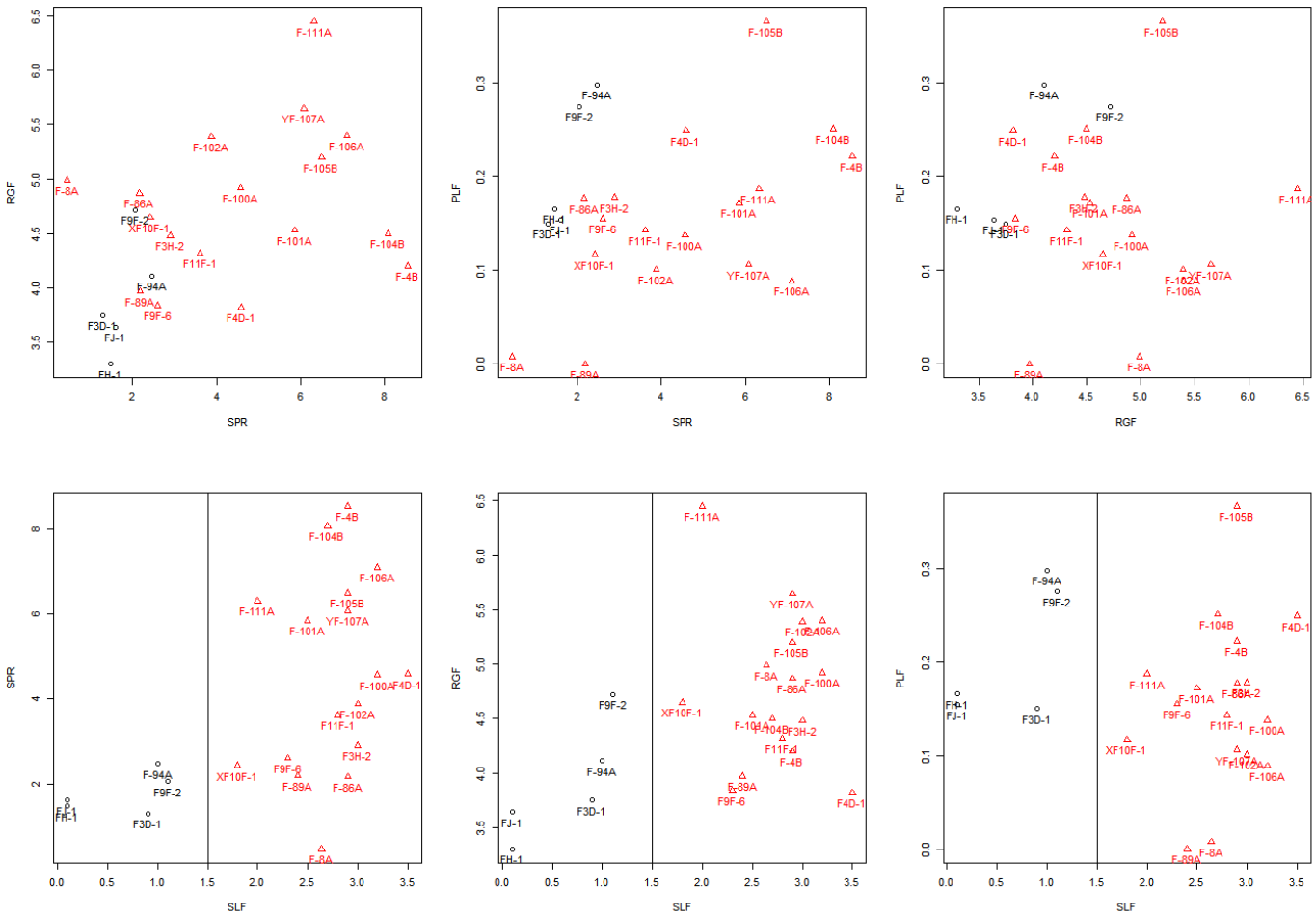
```
plot(jet$SPR, jet$SLF, col=result_hc, pch=result_hc, xlab = "SPR", ylab = "SLF")
text(jet$SPR, jet$SLF, labels = rownames(jet), col=result_hc, pos = 1)
abline(v = 3.75)
plot(jet$RGF, jet$PLF, col=result_hc, pch=result_hc, xlab = "RGF", ylab = "PLF")
text(jet$RGF, jet$PLF, labels = rownames(jet), col=result_hc, pos = 1)
plot(jet$RGF, jet$SLF, col=result_hc, pch=result_hc, xlab = "RGF", ylab = "SLF")
text(jet$RGF, jet$SLF, labels = rownames(jet), col=result_hc, pos = 1)
plot(jet$PLF, jet$SLF, col=result_hc, pch=result_hc, xlab = "PLF", ylab = "SLF")
text(jet$PLF, jet$SLF, labels = rownames(jet), col=result_hc, pos = 1)
```



single linkage 방법을 사용하여 두 개의 군집으로 나눌 경우에는 SPR 변수의 영향을 가장 크게 받았음을 알 수 있다.
SPR 값(전투기 출력)이 약 3.7 보다 작은 전투기는 cluster 1, 큰 전투기는 cluster 2로 군집이 형성되었다.

(2) ward method 사용하는 경우

```
result_hc2 = cutree(hc4, k=2)

plot(jet$SPR, jet$RGF, col=result_hc2, pch=result_hc2, xlab="SPR", ylab = "RGF")
text(jet$SPR, jet$RGF, labels = rownames(jet), col=result_hc2, pos = 1)
plot(jet$SPR, jet$PLF, col=result_hc2, pch=result_hc2, xlab="SPR", ylab = "PLF")
text(jet$SPR, jet$PLF, labels = rownames(jet), col=result_hc2, pos = 1)
plot(jet$RGF, jet$PLF, col=result_hc2, pch=result_hc2, xlab="RGF", ylab = "PLF")
text(jet$RGF, jet$PLF, labels = rownames(jet), col=result_hc2, pos = 1)
plot(jet$SLF, jet$SPR, col=result_hc2, pch=result_hc2, xlab="SLF", ylab = "SPR")
text(jet$SLF, jet$SPR, labels = rownames(jet), col=result_hc2, pos = 1)
abline(v = 1.5)
plot(jet$SLF, jet$RGF, col=result_hc2, pch=result_hc2, xlab="SLF", ylab = "RGF")
text(jet$SLF, jet$RGF, labels = rownames(jet), col=result_hc2, pos = 1)
abline(v = 1.5)
plot(jet$SLF, jet$PLF, col=result_hc2, pch=result_hc2, xlab="SLF", ylab = "PLF")
text(jet$SLF, jet$PLF, labels = rownames(jet), col=result_hc2, pos = 1)
abline(v = 1.5)
```



ward method를 사용하여 두 개의 군집으로 나눌 경우에는 SLF 변수가 가장 큰 영향을 끼쳤음을 알 수 있다.
SLF 값(전투기 무게)이 약 1.5 보다 작은 전투기는 cluster 1, 큰 전투기는 cluster 2로 군집이 형성되었다.

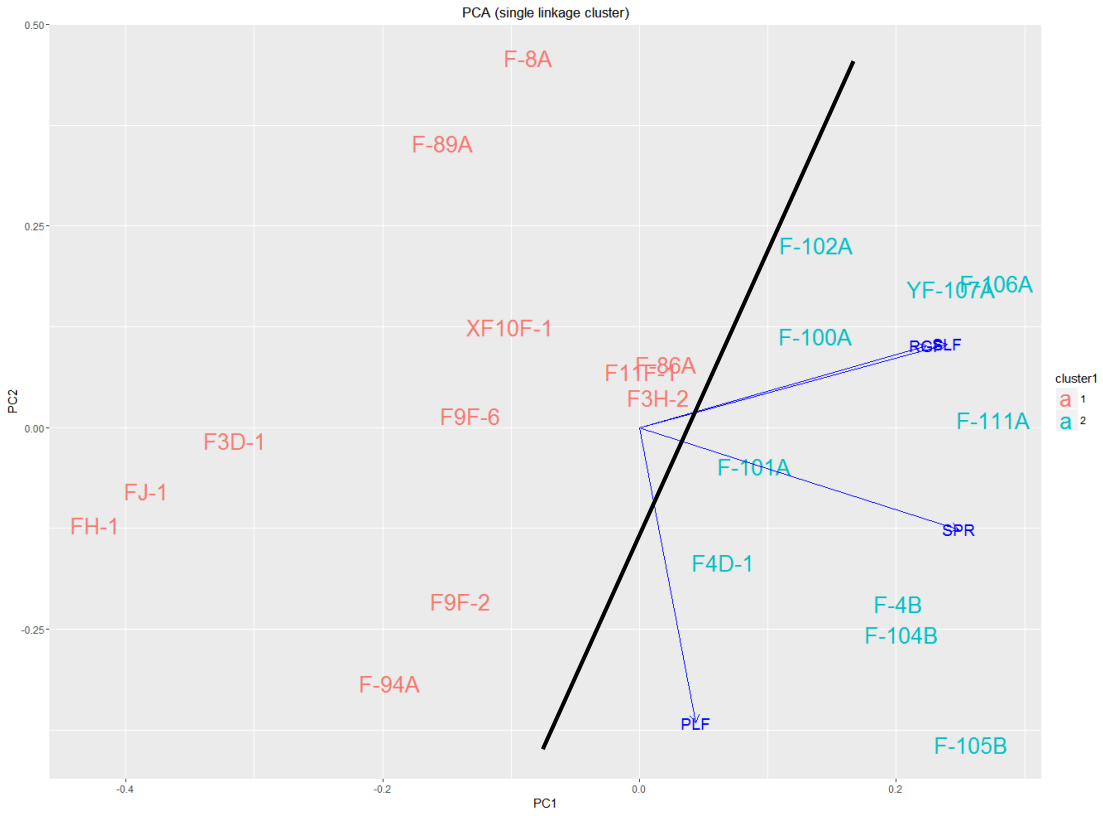
C. 두 집단을 주성분을 이용해 2차원 산점도로 표현하시오. (즉, 제1 주성분과 제2 주성분을 사용한 산점도에서 두 개의 집단을 서로 다른 마크와 색으로 표현하시오.)

```
# 주성분분석
pca <- prcomp(jet, scale = T)
summary(pca)
```

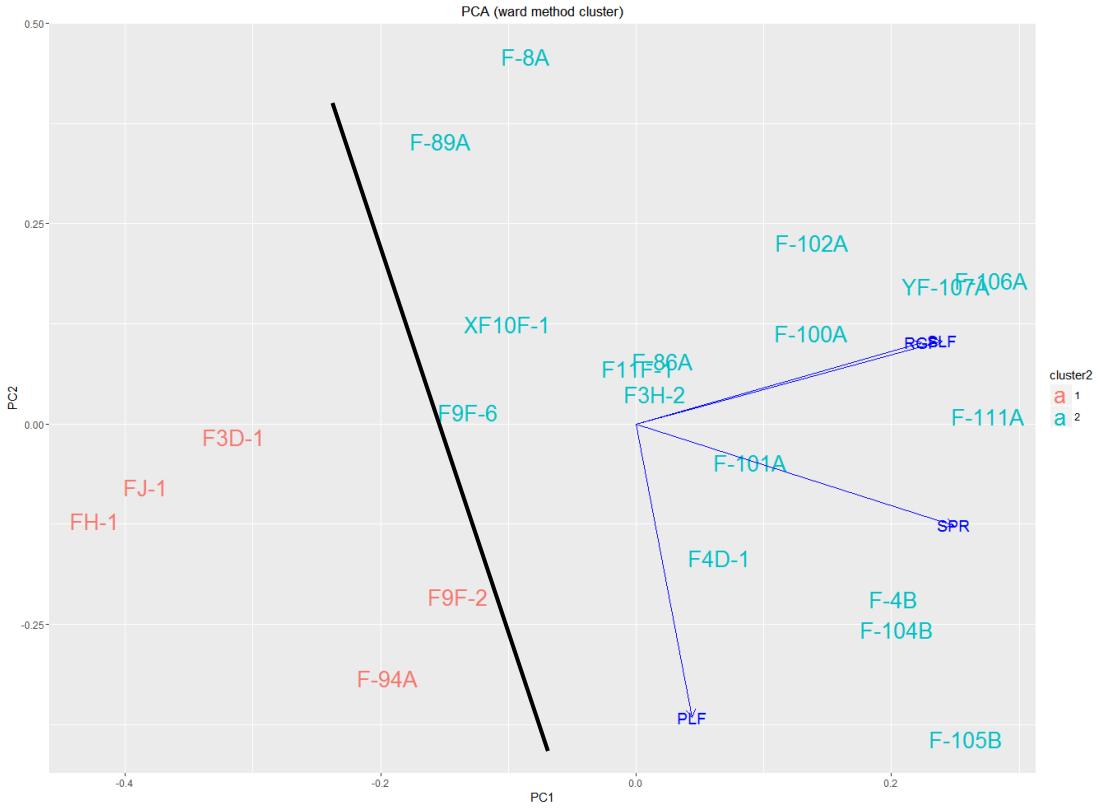
```
    mportance of components:
                            PC1    PC2    PC3     PC4
    Standard deviation     1.4064 1.0779 0.7422 0.55609
    Proportion of Variance 0.4945 0.2905 0.1377 0.07731
    Cumulative Proportion  0.4945 0.7850 0.9227 1.00000
```

처음 2개의 주성분으로 전체 변동의 78%를 설명할 수 있다.

```
library(ggfortify)

jet_c = jet
jet_c$cluster1 = factor(result_hc)   # single linkage
jet_c$cluster2 = factor(result_hc2)  # ward method
jet_c

# single linkage
autoplot(pca, data = jet_c, colour = 'cluster1', shape = F, label = T, loadings = T, label.size = 7, loadings.label.size = 5, loadings.label = T, loadings.colour = "blue", loadings.label.colour = "blue")

# ward method
autoplot(pca, data = jet_c, colour = 'cluster2', shape = F, label = T, loadings = T, label.size = 7, loadings.label.size = 5, loadings.label = T, loadings.colour = "blue", loadings.label.colour = "blue")
```



PCA : single linkage cluster - SPR 과 관련성 높음



PCA : ward method cluster - SLF 와 관련성 높음

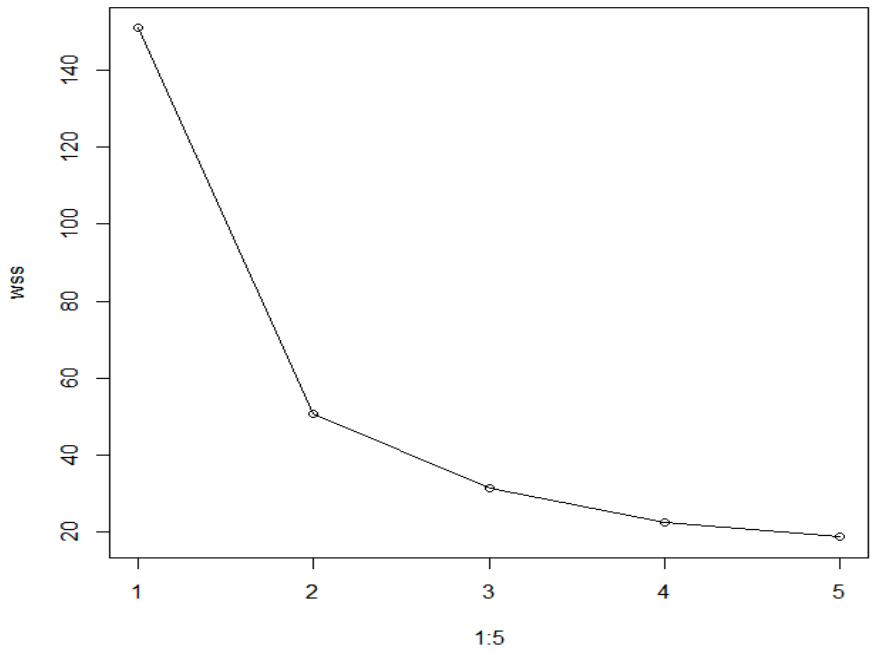
2. 비계층적군집분석

A. 군집 개수 1~5까지를 사용해 k-means clustering을 시행하고 얻은 within-group sum of squares를 저장하고 그래프로 표현하여 적절한 군집 개수를 판단하시오.

```
wss = c()
for (k in 1:5) {
    km = kmeans(jet, k)
    wss[k] = sum(km$withinss)
}
wss
```

```
[1] 150.88282  50.79957  31.32728  22.41017  18.71070
```

```
plot(1:5, wss, type = 'l')
points(wss)
```



2에서 팔꿈치 생김. 군집 3 ~ 5개 일 경우 서로 wss 차이가 많지 않음.
군집수는 2개로 결정

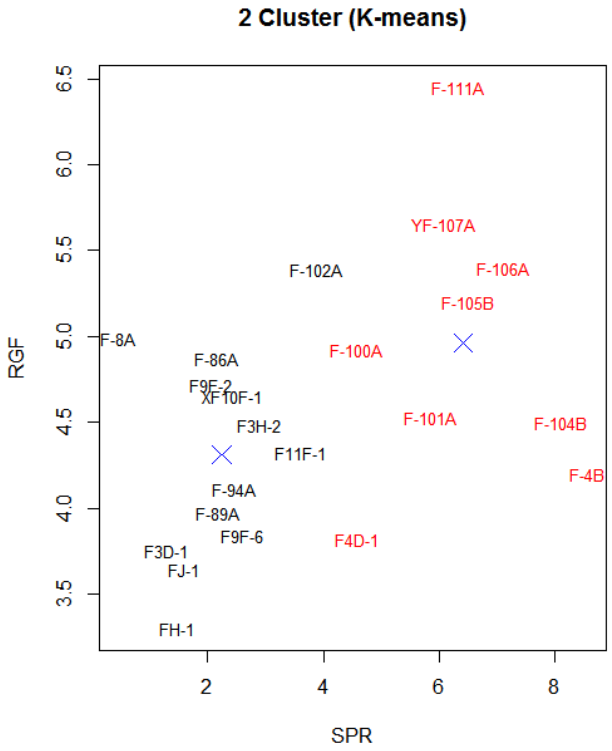
B. K-means clustering을 이용해 2개의 집단으로 군집화하고 그 결과를 1번의 B, C 와 같이 탐색하시오.

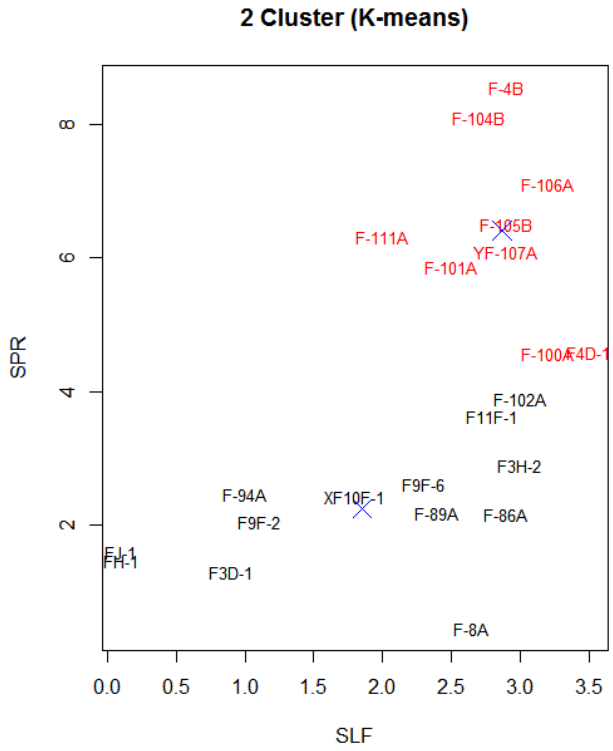
```
result_km = kmeans(jet, 2)
result_km
```

```
    K-means clustering with 2 clusters of sizes 13, 9

    Cluster means:
          SPR      RGF       PLF      SLF
    1 2.240231 4.310000 0.1478462 1.849231
    2 6.406111 4.963333 0.1977778 2.866667
```

```
plot(jet$SPR, jet$RGF, type = "n", xlab = "SPR", ylab = "RGF", main = "2 Cluster")
text(jet$SPR, jet$RGF, labels=rownames(jet), cex = 0.8, col = result_km$cluster)
points(result_km$centers[, c(1,2)], col = "blue", pch = 4, cex = 2)
```

```
plot(jet$SLF, jet$SPR, type = "n", xlab = "SLF", ylab = "SPR", main = "2 Cluster")
text(jet$SLF, jet$SPR, labels=rownames(jet), cex = 0.8, col = result_km$cluster)
points(result_km$centers[, c(4,1)], col = "blue", pch = 4, cex = 2)
```



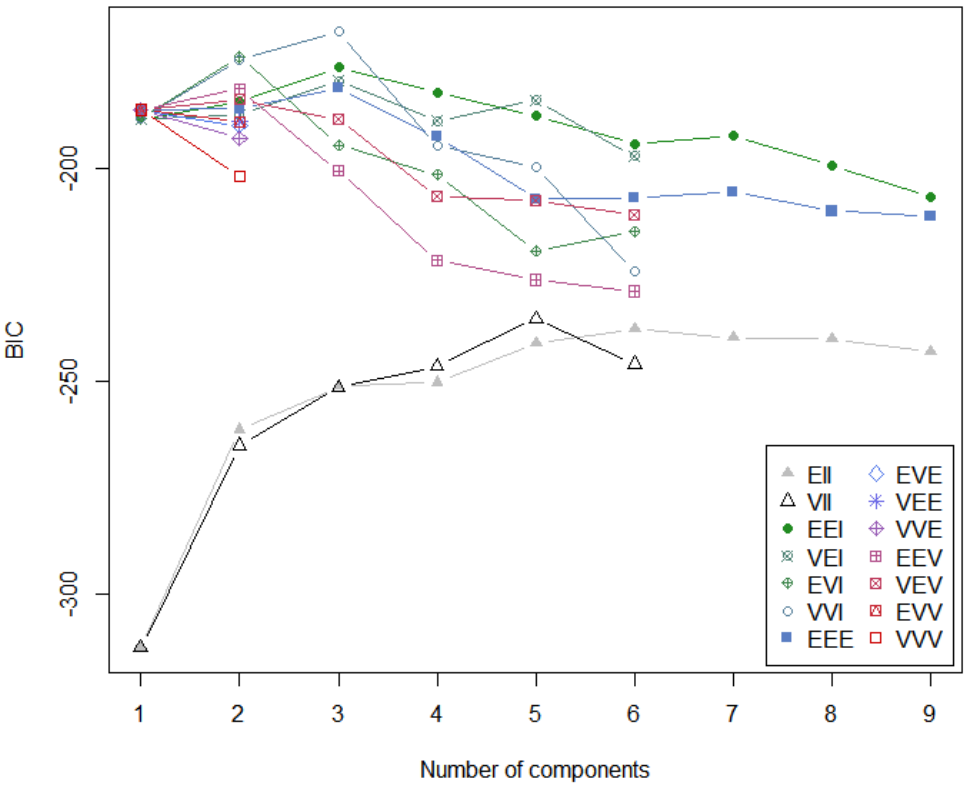


3. 모형기반 군집화를 통해 최적의 군집 개수를 찾고 그 결과를 1번의 B, C와 같이 탐색하시오.

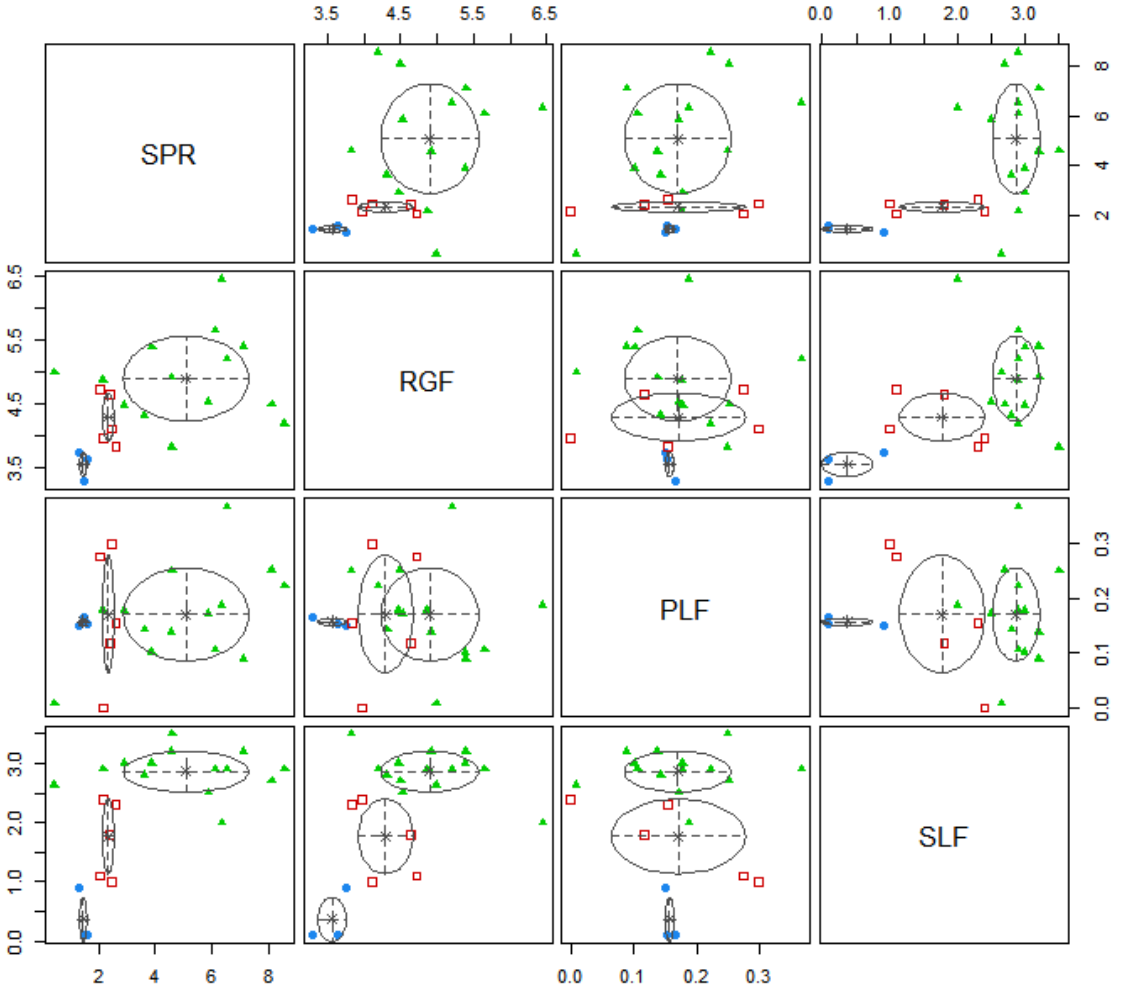
```
library(mclust)
result_mc = Mclust(jet)
summary(result_mc)
```

```
        Mclust VVI model with 3 components:
         log.likelihood  n df       BIC       ICL
              -43.76189 22 26 -167.8909 -168.7453

        Clustering table:
         1  2  3
         3  5 14
```

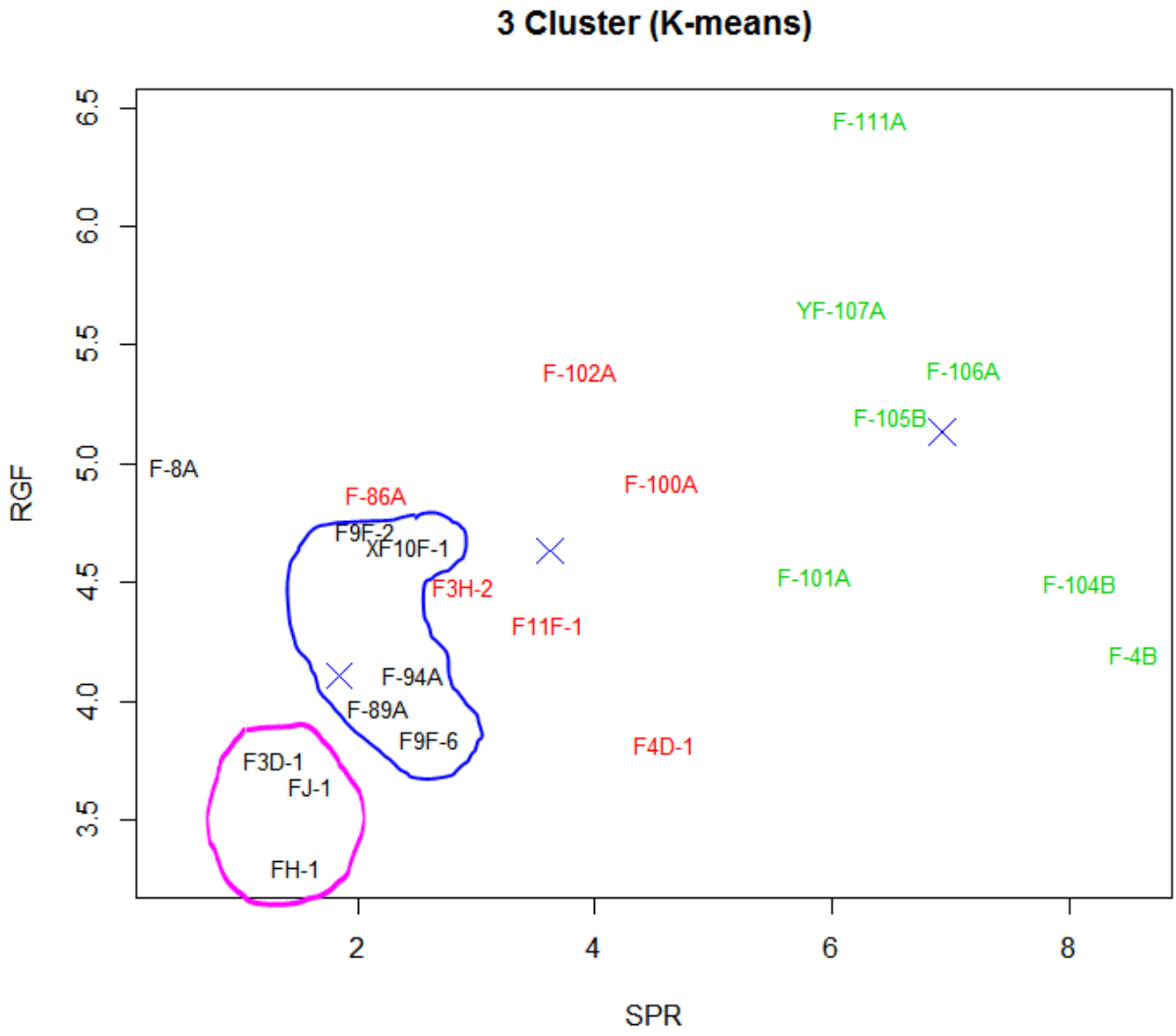
```
plot(result_mc)

# BIC plot
```



```
# Classification plot
```

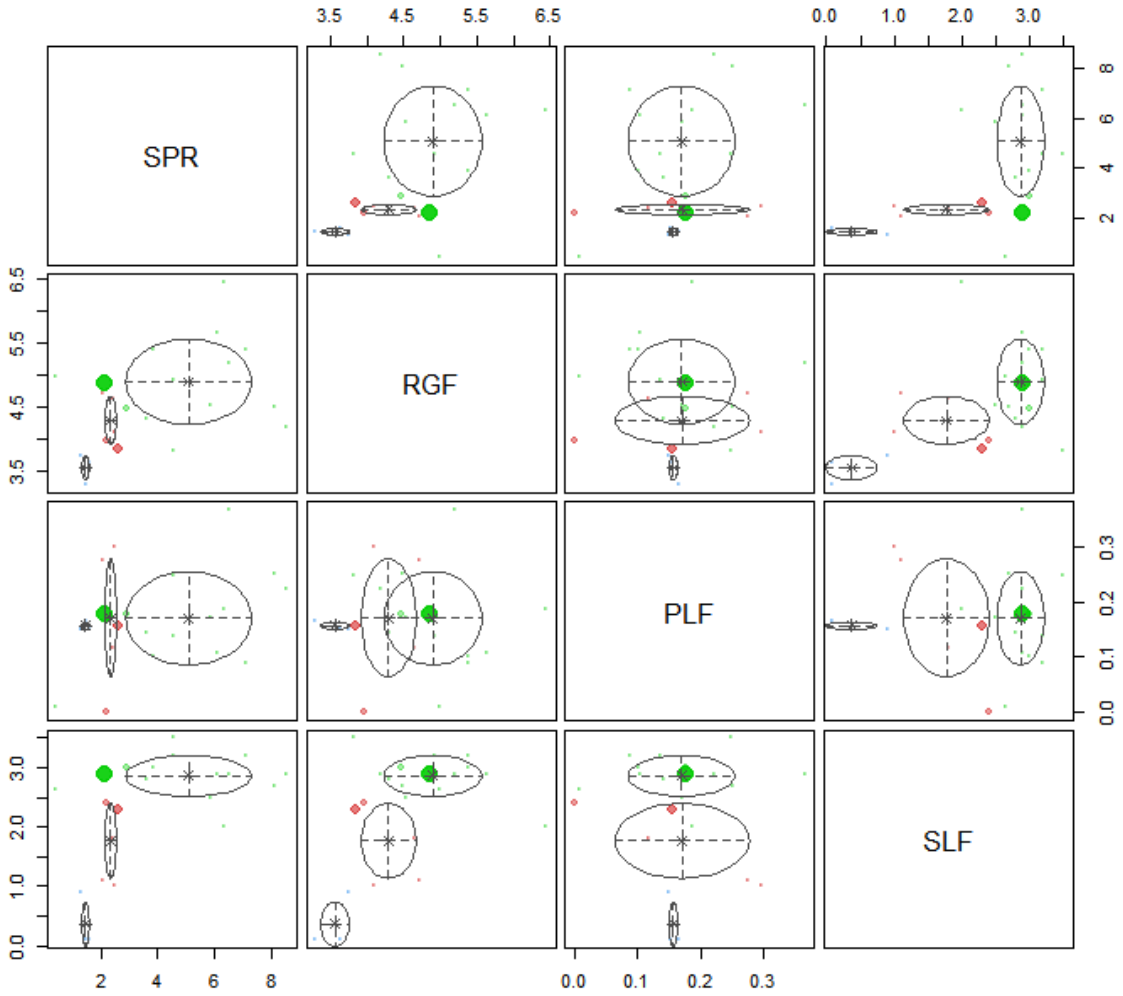



```
rownames(jet[result_mc$classification == 1, ])
       [1] "FH-1"  "FJ-1"  "F3D-1"

rownames(jet[result_mc$classification == 2, ])
       [1] "F9F-2"   "F-94A"   "F-89A"   "XF10F-1" "F9F-6"

rownames(jet[result_mc$classification == 3, ])
       [1] "F-86A"   "F-100A"  "F4D-1"   "F11F-1"  "F-101A"  "F3H-2"
           "F-102A"  "F-8A"    "F-104B"  "F-105B"  "YF-107A"
           "F-106A"  "F-4B"    "F-111A"
```



군집을 3개로 했을 경우 모형기반 군집분석과 k-means 군집분석의 차이
(펜으로 표시한 것이 모형기반 군집)

Uncertainty




```
jet[order(result_mc$uncertainty, decreasing = T), ][1, ]   # F-86A
```

F-86A 전투기의 uncertainty 가 **2.484012e-01** 으로 가장 높다.
위 그래프 상의 초록색 점.

Density