이름 (김우현 U2016040)

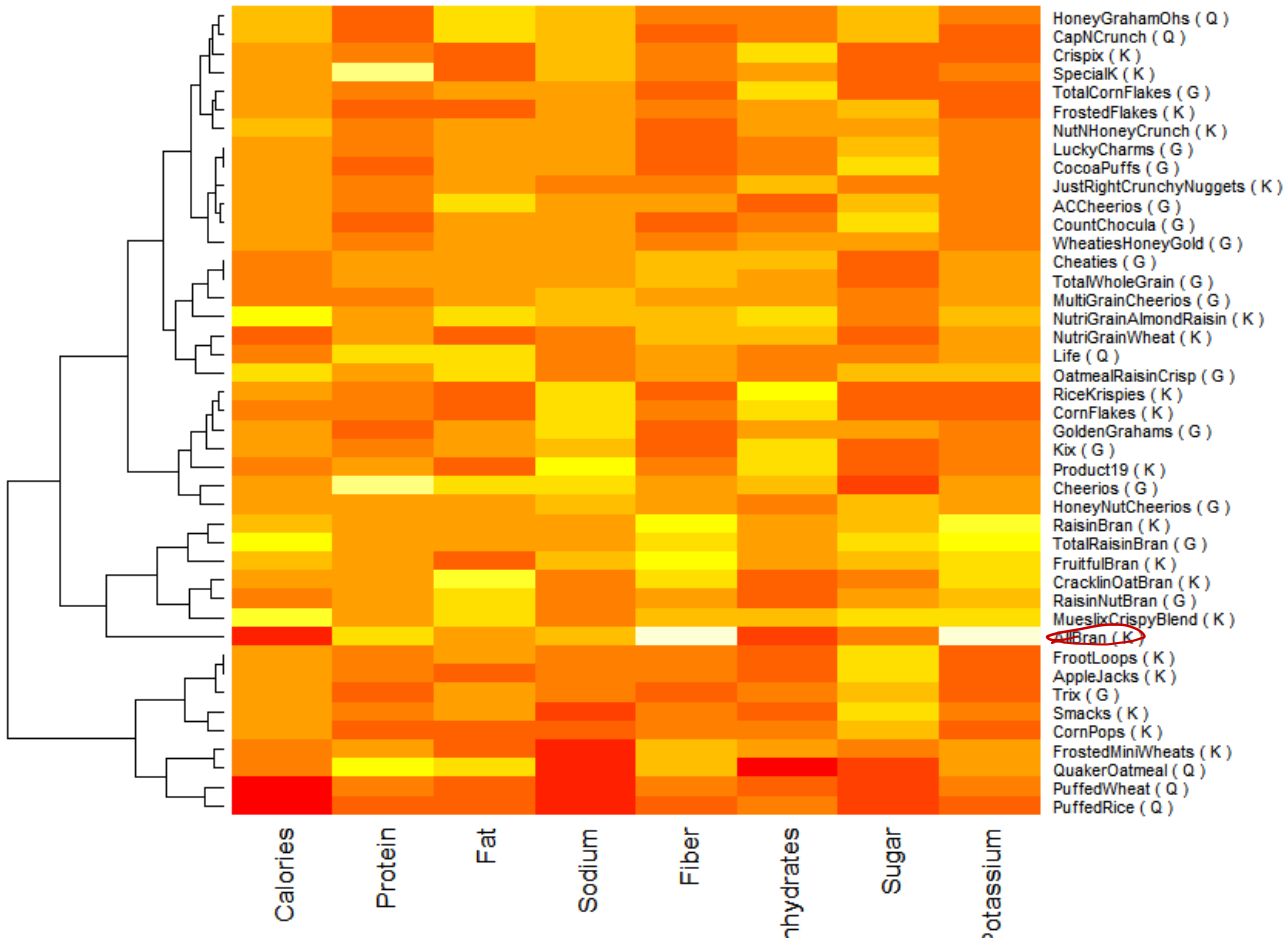
104

중간고사
다변량통계분석
2016년 2학기

- 각 문항에 답을 하기 위해 사용된 그래프, 표, 통계량 등을 반드시 모두 제시하시오.
- 각 문제에 대한 답안 파일과 문제를 해결하기 위해 사용한 R 스크립트 파일을 함께 제출하시오.

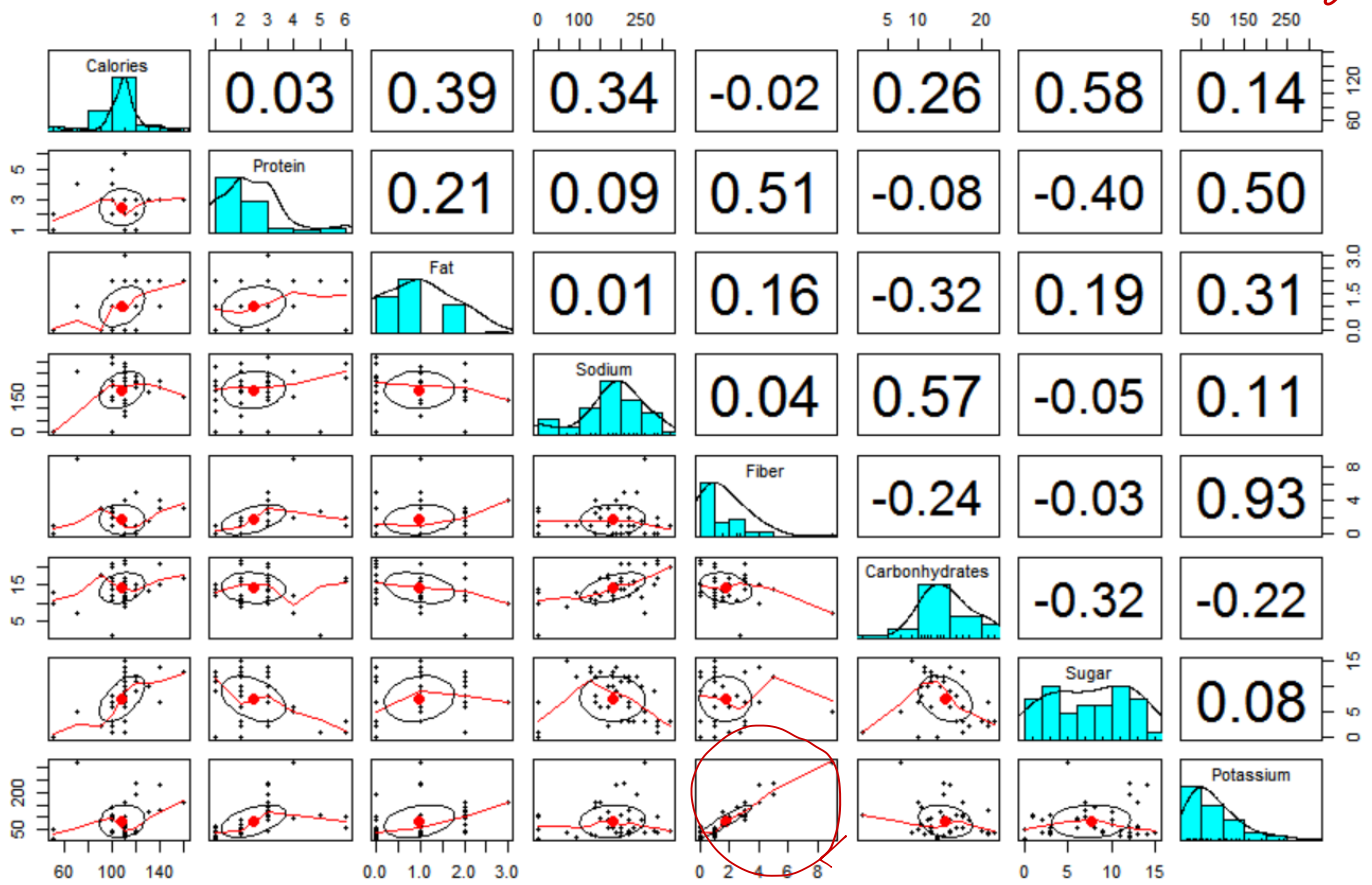
1. Cereal.csv는 3개의 미국 시리얼 제조사(General Mills: G, Kellogg: K, Quaker: Q)에 의해 생산되는 아침식사용 시리얼 각 브랜드의 영양성분 자료이다.
   A. 영양성분 상 특성을 시리얼 별로 한눈에 비교하기 위한 그래프를 그린 후 비슷한 영양성분을 가지는 시리얼들을 탐색적으로 구분하여 서술하시오.

```
cereal = read.csv("cereal.csv", header = T)
head(cereal)

# Heatmap으로 전체적인 데이터 탐색적 분석
cereal_m = cereal
rownames(cereal_m) = paste(cereal_m$Brand, '(', as.character(cereal_m$Manufacturer), ')')
heatmap(as.matrix(cereal_m[,3:10]), scale = "column", Colv = NA)
```

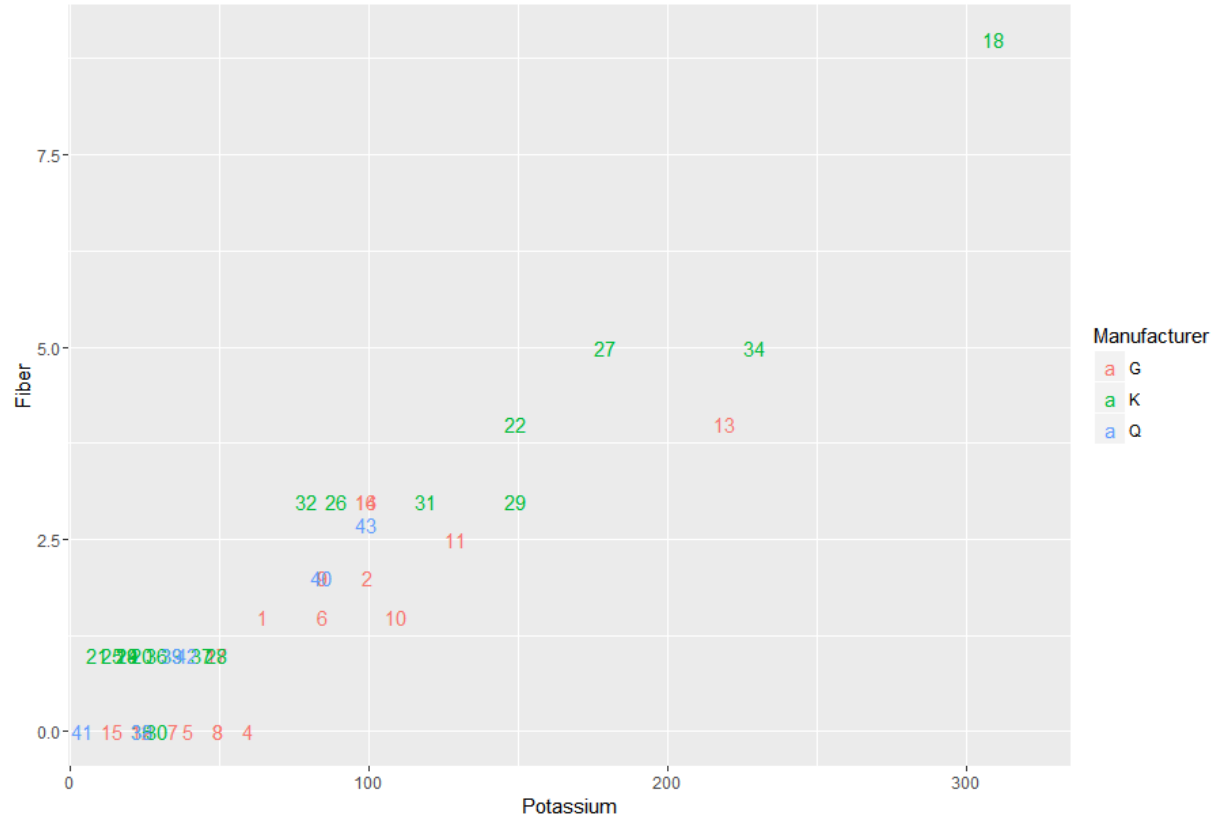


- AllBran 은 다른 제품에 비해 이례적으로 Fiber, Potassium 함량이 매우 높고, Calories 가 매우 낮은 제품임을 알 수 있다.
- HoneyGrahamOhs 에서 WheatiesHoneyGold 까지 13 개 제품은 Calories 와 Sodium 이 중간치 정도이면서 Fiber, Potassium 함량이 낮은 제품군으로 보인다.
- RiceKrispies 에서 Product19 까지의 5 개 제품은 Sodium 과 Carbonhydreates 의 함량이 매우 높고, Fat, Fiber, Sugar 의 함량이 상대적으로 낮은 제품군이다.
- RaisinBran 에서 MueslixCrispyBlend 까지 6 개 제품은 Fiber 와 Potassium 의 함량이 매우 높으면서 Protein 이 중간치 정도를 포함한 제품군이다.ㅇㅇ
- 가장 아래쪽에 있는 PuffedRice, PuffedWheat 두 가지 제품은 Calories와 Sodium 함량이 매우 낮고, 다른 영양 성분도 전체적으로 낮은 제품이다.

```
# Find outlier
library(psych)
library(ggplot2)

pairs.panels(cereal[, 3:10])
```

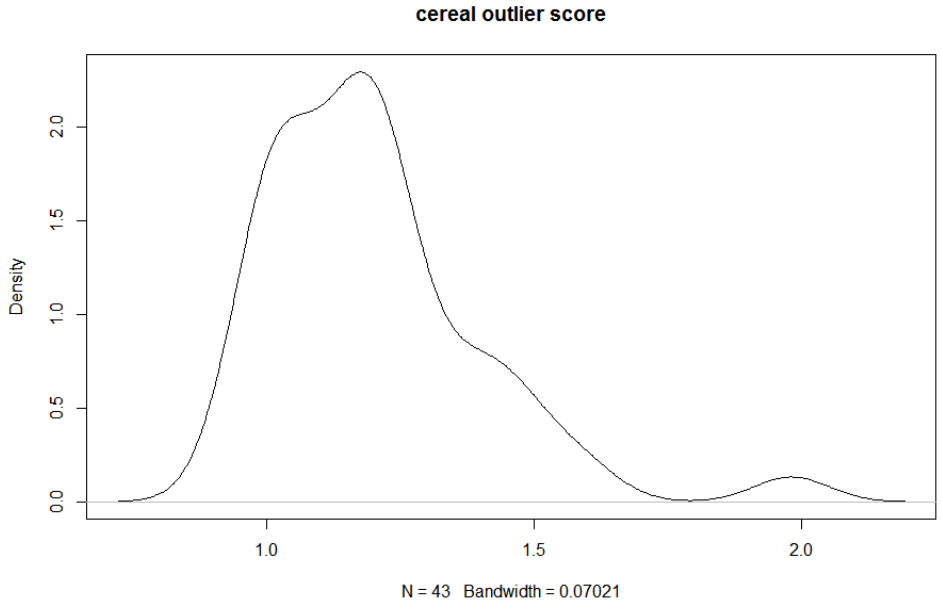



```
# Fiber, Potassium 산점도에서 이상치 데이터 보임
# 산점도로 이상치 확인

ggplot(cereal, aes(Potassium, Fiber)) +
geom_text(aes(Potassium, Fiber, label = rownames(cereal), colour = Manufacturer), hjust = 2)
```



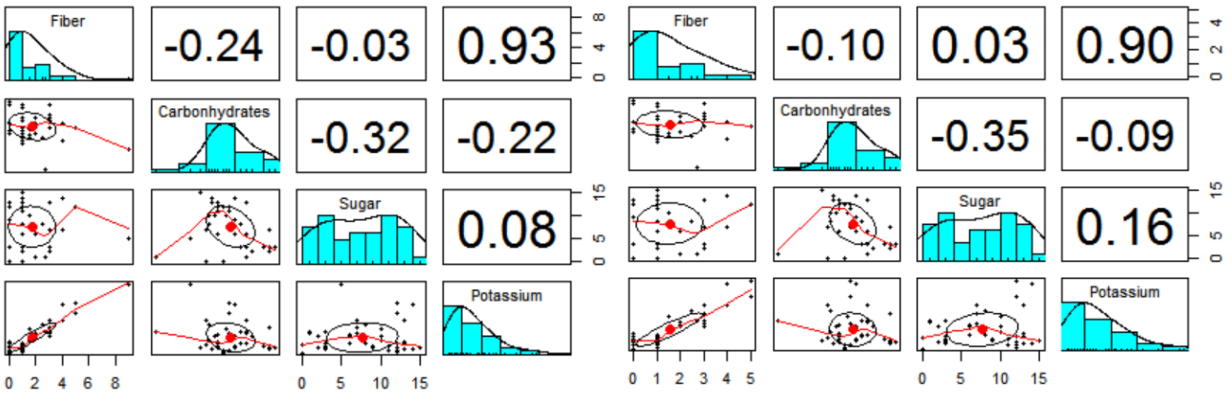

```
outlier = 18
```

```
# 전체 변수를 기준으로 이상치 확인

library(DMwR)
outlier.score <- lofactor(cereal[,3:10], k = 5)  # 이웃 5개 데이터 기준
plot(density(outlier.score), main = "cereal outlier score")
```




```
sort(outlier.score, decreasing = T)[1:10]
outliers <- order(outlier.score, decreasing = T)[1]
# score > 1.9 인 데이터를 outlier로 결정
```

```
outliers              # 18
cereal[outliers, ]    # 역시 outlier 는 AllBran 제품
pairs.panels(cereal[, 3:10])
pairs.panels(cereal[-outlier, 3:10])
```



원본 데이터 이상치 제거 데이터

```
# outlier 제거하기 전 Fiber, Potassium의 correlation = 0.93
# outlier 제거 후 correlation = 0.90
# 두 변수가 correlation의 변동량이 크지 않고, 다른 변수에서 큰 영향이 없기 때문에 outlier 포함하고 주성분 분석 진행.
```

B. 8개의 영양성분 상의 특성을 보다 적은 차원에서 설명하기 위해 주성분분석을 활용하여 분석을 진행하시오. 적절한 그래프와 결과물을 사용하여 아래의 문항에 답변하시오.

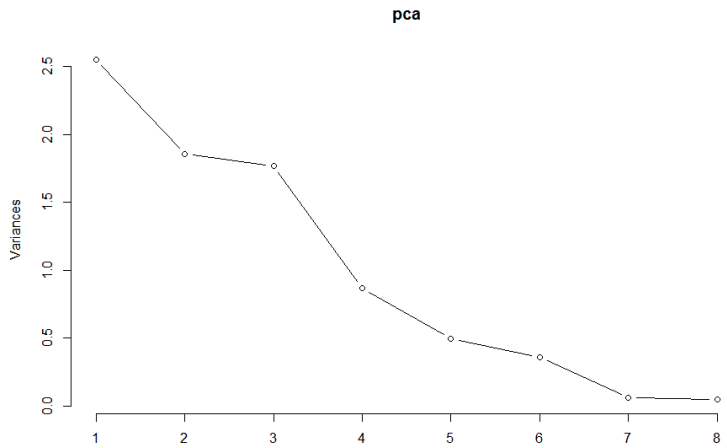
i. 적절한 주성분의 개수는 무엇인가?

ii. 각 주성분은 어떤 의미를 가지는가?

iii. 이상치가 있는가? 있다면 어떤 성질을 가지는가?

iv. 주성분 분석의 결과를 활용하여 볼 때 각 제조사가 생산하는 시리얼 별로 영양성분 상의 특성이 다른가?

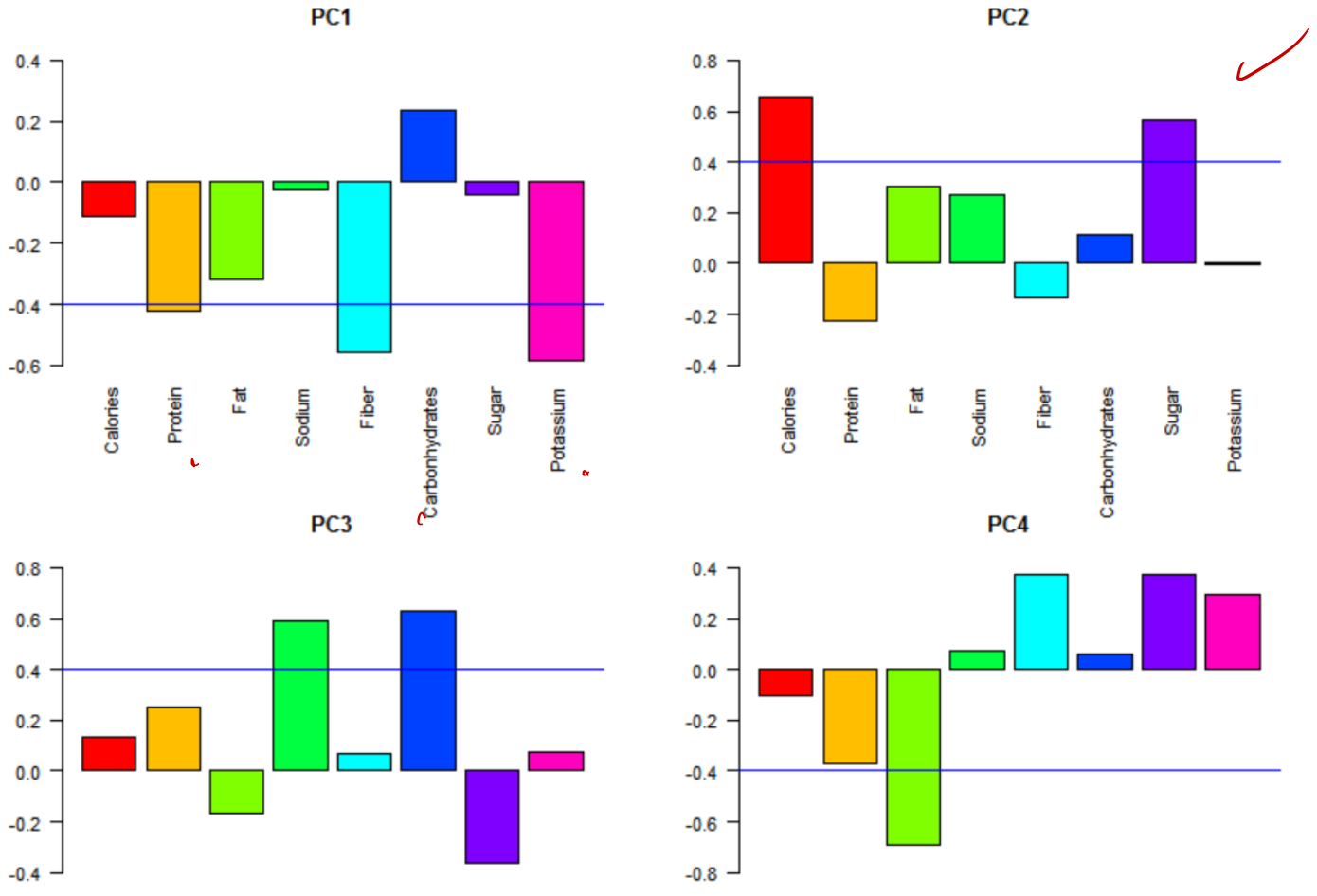
```
pca <- prcomp(cereal[, 3:10], scale = T)   # 상관계수 행렬 이용
plot(pca, type = "l")
```



```
summary(pca)
```

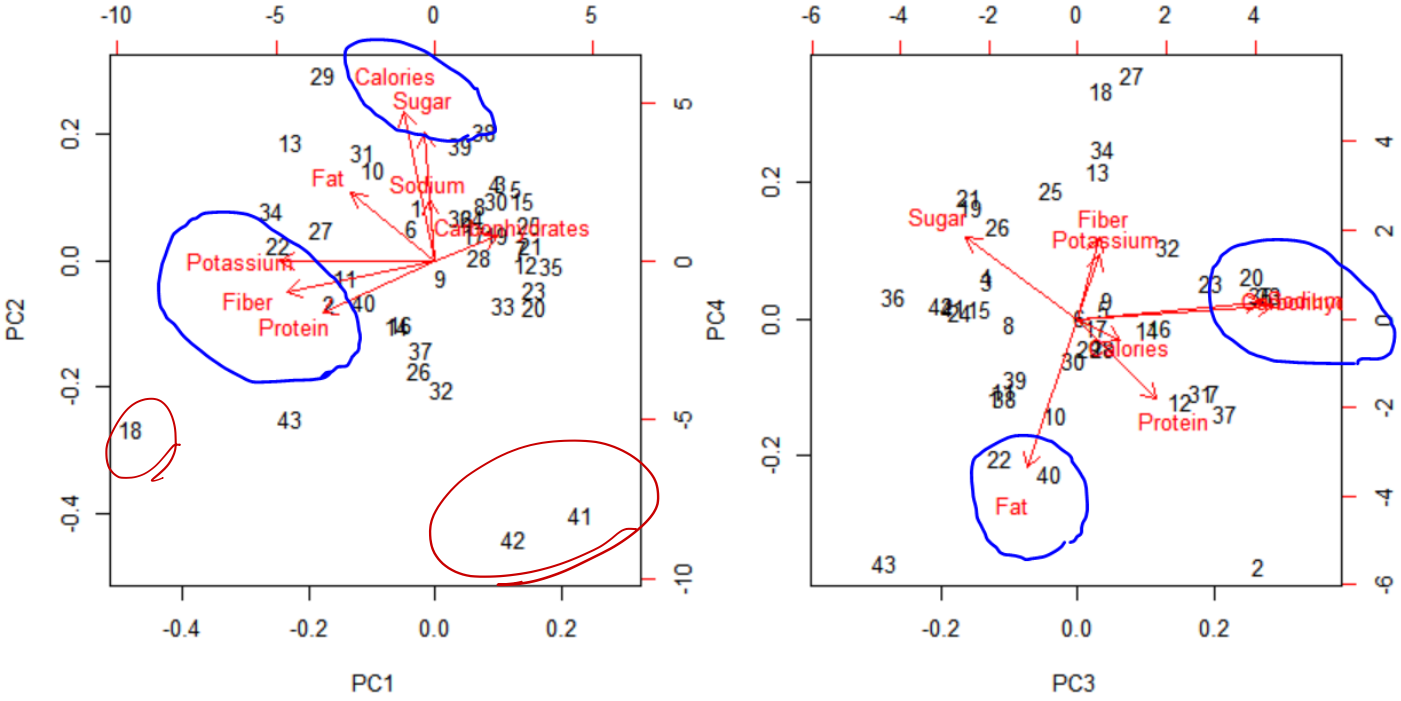
```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.5961 1.3619 1.3300 0.9318 0.70520 0.59785 0.24597 0.21289
Proportion of Variance 0.3185 0.2318 0.2211 0.1085 0.06216 0.04468 0.00756 0.00567
Cumulative Proportion  0.3185 0.5503 0.7714 0.8799 0.94209 0.98677 0.99433 1.00000
```

```
# 주성분의 개수
# Cumulative Proportion을 기준으로 87.99 %까지 설명이 가능한 PC4까지 4개의 주성분 선택


# 각 주성분은 어떤 의미를 가지는가?

par(mfrow=c(2,2))
barplot(pca$rotation[,1], col = rainbow(8), ylim = c(-0.6,0.4), las = 2, main = "PC1")
abline(h = -0.4, col="blue")
barplot(pca$rotation[,2], col = rainbow(8), ylim = c(-0.4,0.8), las = 2, main = "PC2")
abline(h = 0.4, col="blue")
barplot(pca$rotation[,3], col = rainbow(8), ylim = c(-0.4,0.8), las = 2, main = "PC3")
abline(h = 0.4, col="blue")
barplot(pca$rotation[,4], col = rainbow(8), ylim = c(-0.8,0.4), las = 2, main = "PC4")
abline(h = -0.4, col="blue")
par(mfrow=c(1,1))
```



- 제1 주성분은 시리얼의 영양성분 중 단백질(protein), 섬유소(fiber), 칼륨(potassium)의 함량이 높은 제품들의 특성을 나타낸다.
- 제2 주성분은 설탕이 많이 들어있고 칼로리가 높은 제품의 특성을 설명하고 있다.
- 제3 주성분은 나트륨(sodium)과 탄수화물(carbonhydrate) 함량이 높은 제품의 특성을 보여주고 있다.
- 마지막으로 제4 주성분은 지방의 함량이 높은 제품의 특성을 나타낸다.

```
par(mfcol=c(1,2))
biplot(pca)
biplot(pca, choices = c(3,4))
par(mfcol=c(1,1))
```




```
# 이상치가 있는가? 있다면 어떤 성질을 가지는가?
```

PC1과 PC2의 행렬도에서 18, 41, 42번 제품이 이상치라고 할 수 있다.

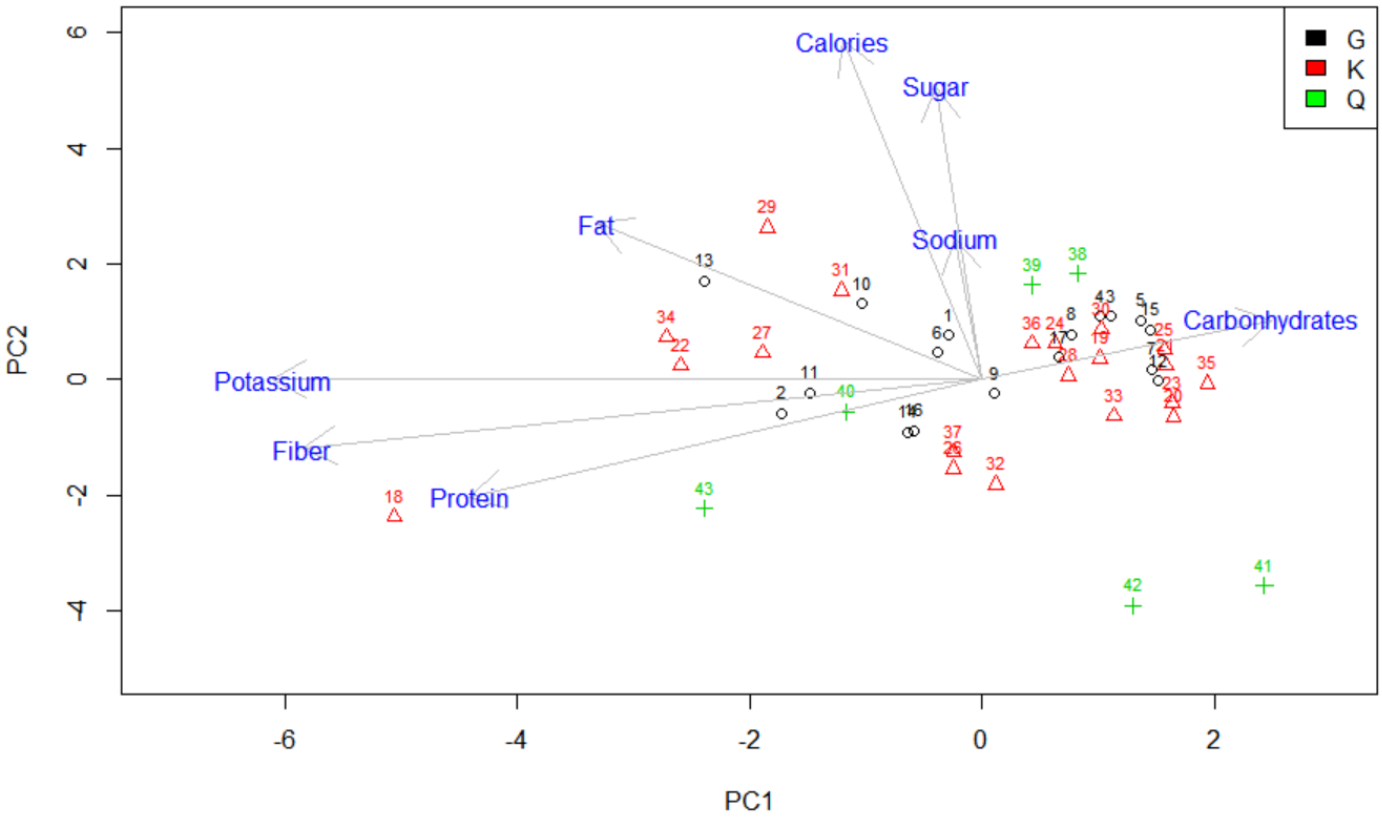
```
ord = order(cereal$Potassium, decreasing = T)[1:10]
cereal[ord, ]   # 18: potassium, fiber 함량이 다른 제품에 비해 상당히 높다.
```

	Brand	Manufacturer	Calories	Protein	Fat	Sodium	Fiber	Carbonhydrates	Sugar	Potassium
18	AllBran	K	70	4	1	260	9.0	7.0	5	320
34	RaisinBran	K	120	3	1	210	5.0	14.0	12	240
13	TotalRaisinBran	G	140	3	1	190	4.0	15.0	14	230
27	FruitfulBran	K	120	3	0	240	5.0	14.0	12	190
22	CracklinOatBran	K	110	3	3	140	4.0	10.0	7	160
29	MueslixCrispyBlend	K	160	3	2	150	3.0	17.0	13	160
11	RaisinNutBran	G	100	3	2	140	2.5	10.5	8	140
31	NutriGrainAlmondRaisin	K	140	3	2	220	3.0	21.0	7	130
10	OatmealRaisinCrisp	G	130	3	2	170	1.5	13.5	10	120
14	TotalWholeGrain	G	100	3	1	200	3.0	16.0	3	110

```
ord = order(cereal$Calories)[1:6]
cereal[ord, ]   # 41, 42 : Calories, Sugar 함량이 다른 제품에 비해 상당히 낮다.
                           또한, Sodium, Fat의 함량이 0이다.
```

	Brand	Manufacturer	Calories	Protein	Fat	Sodium	Fiber	Carbonhydrates	Sugar	Potassium
41	PuffedRice	Q	50	1	0	0	0.0	13.0	0	15
42	PuffedWheat	Q	50	2	0	0	1.0	10.0	0	50
18	AllBran	K	70	4	1	260	9.0	7.0	5	320
32	NutriGrainWheat	K	90	3	0	170	3.0	18.0	2	90
9	MultiGrainCheerios	G	100	2	1	220	2.0	15.0	6	90
11	RaisinNutBran	G	100	3	2	140	2.5	10.5	8	140

```
# 주성분 분석 결과를 볼 때 각 제조사가 생산하는 시리얼 별로 영양성분 상의 특성이 다른가

# 색깔을 구분하기 위해 제조사별로 번호 지정
pcolors = ifelse(cereal$Manufacturer == 'G', 1, ifelse(cereal$Manufacturer == 'K', 2, 3))
plot(pca$x[,1], pca$x[,2], xlab = "PC1", ylab = "PC2",
     xlim = c(-7,3), ylim = c(-5,6), pch = pcolors, col = pcolors)
text(pca$x[,1], pca$x[,2], labels = rownames(cereal), cex = 0.7, pos = 3, col = pcolors)
legend("topright", c("G","K","Q"), col = c("black", "red", "green"),
                                   fill = c("black", "red", "green"))
# 주성분 화살표의 방향과 길이
lambda <- pca$sdev * sqrt(nrow(pca$x))
rot <- t(t(pca$rotation)*lambda)
arrows(rep(0,nrow(pca$rotation)), rep(0,nrow(pca$rotation)), rot[,1], rot[,2], col = "grey")
text(rot[,1:2], rownames(rot), col = "blue")   # 화살표 제목
```



- 영양성분 만으로는 제조사를 특징지을 수 없는 분포를 가지고 있다.
- General Mills (G) 제품의 경우 중심점과 가까운 거리에 분포하고 있어 제품들의 영양성분 함량이 별로 좋지 않다고 판단할 수 있다.
- Kellogg (K) 제품들의 경우 고급 제품인 18번을 제외하면, 지방이 많은 제품군과 탄수화물이 많은 제품군으로 나뉘는 것을 알 수 있다.
- Quaker (Q) 제품들의 경우에는 저지방, 저칼로리에 특화된 제품군을 가지고 있으며, 나머지는 다른 회사 제품들의 틈새에 잘 포지셔닝 되어 있다.

---

2. Psych package 안에 포함되어 있는 Thurstone.33 데이터셋은 4175명의 학생의 인지능력 테스트로부터 계산된 상관계수 행렬이다.

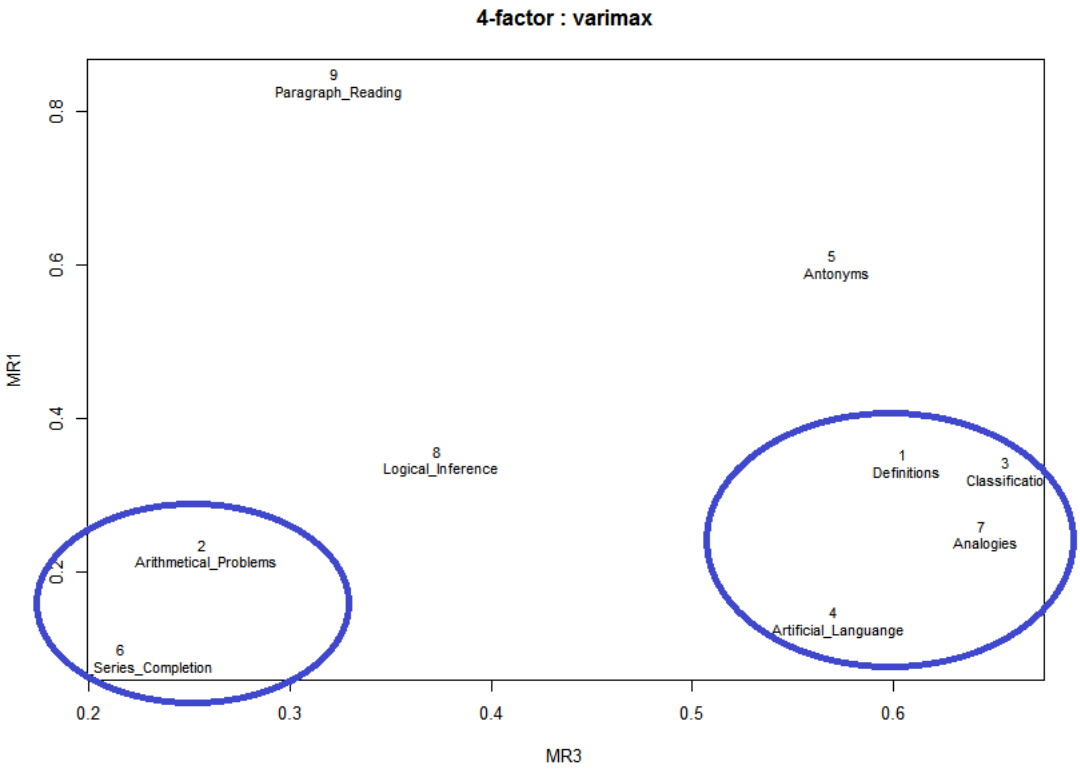
   A. 이 데이터를 사용하여 요인분석을 진행하여 9개의 테스트 결과에 영향을 주는 잠재요인을 파악하시오. (적절한 요인 개수와 요인회전 고려)

```
library(psych)
data("Thurstone.33")
df = Thurstone.33   # 상관계수 행렬. 이미 표준화되어 있기 때문에 scaling 할 필요 없음.

library(GPArotation)

fa1 = fa(df, 4, rotate = "varimax")
fa2 = fa(df, 5, rotate = "varimax")
fa3 = fa(df, 4, rotate = "quartimax")
fa4 = fa(df, 5, rotate = "quartimax")

# 4-factor varimax
print(fa1, digits = 2, sort = T)
```

```
Factor Analysis using method =  minres
Call: fa(r = df, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                        item  MR3  MR1  MR2  MR4   h2    u2 com
Classification             3 0.66 0.33 0.21 0.11 0.60 0.402 1.8
Analogies                  7 0.64 0.25 0.21 0.29 0.61 0.392 2.0
Definitions                1 0.61 0.34 0.30 0.19 0.61 0.391 2.4
Artificial_Languange       4 0.57 0.14 0.33 0.17 0.49 0.513 2.0
Paragraph_Reading          9 0.32 0.84 0.20 0.24 0.90 0.099 1.6
Antonyms                   5 0.57 0.60 0.15 0.18 0.74 0.255 2.3
Number_Series_Completion   6 0.22 0.09 0.83 0.09 0.76 0.242 1.2
Arithmetical_Problems      2 0.26 0.23 0.60 0.22 0.52 0.479 2.0
Logical_Inference          8 0.37 0.35 0.30 0.68 0.82 0.183 2.6
```

   # 변수들이 아이템번호를 기준으로 (3,7,1,4) (9,5) (6,2) (8)로 나뉘어 각 요인에 대한 설명력을 갖는다.

```
plabel = paste(c(1:9), "\n", colnames(df))
plot(fa1$loadings, type = "n", main = "4-factor : varimax")
text(fa1$loadings, labels = plabel, cex = 0.8)
```



+2.

```
# 5-factor varimax
print(fa2, digits = 2, sort = T)
```

```
Factor Analysis using method =  minres
Call: fa(r = df, nfactors = 5, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                       item  MR3  MR4  MR5  MR1  MR2   h2    u2 com
Paragraph_Reading         9 0.75 0.23 0.29 0.10 0.26 0.77 0.228 1.8
Antonyms                  5 0.72 0.17 0.41 0.23 0.18 0.80 0.204 2.1
Arithmetical_Problems     2 0.22 0.75 0.17 0.10 0.17 0.68 0.324 1.4
Number_Series_Completion  6 0.10 0.67 0.18 0.20 0.13 0.54 0.456 1.5
Classification            3 0.36 0.22 0.66 0.17 0.13 0.66 0.343 2.1
Analogies                 7 0.30 0.22 0.53 0.27 0.29 0.58 0.416 3.2
Definitions               1 0.38 0.34 0.52 0.22 0.19 0.61 0.386 3.4
Artificial_Languange      4 0.19 0.26 0.26 0.90 0.14 1.00 0.005 1.5
Logical_Inference         8 0.34 0.31 0.26 0.17 0.81 0.97 0.033 2.0
```

변수들이 아이템번호를 기준으로 (9,5) (6,2) (3,7,1) (4) (8)로 나뉘어 각 요인에 대한 설명력을 갖는다.

```
# 4-factor / 5-factor quartimax
print(fa3, digits = 2, sort = T)
print(fa4, digits = 2, sort = T)
```

quartimax 회전의 경우 하나의 요인에 너무 많은 변수들이 영향을 미치는 것으로 나타나기 때문에 이 데이터에 대한 요인 회전으로는 적당하지 않다.

```
Factor Analysis using method =  minres
Call: fa(r = df, nfactors = 4, rotate = "quartimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                        item  MR1   MR2   MR3   MR4   h2    u2 com
Antonyms                   5 0.85 -0.07 -0.09 -0.10 0.74 0.255 1.1
Paragraph_Reading          9 0.85 -0.04 -0.43 -0.01 0.90 0.099 1.5
Logical_Inference          8 0.76  0.12  0.01  0.47 0.82 0.183 1.7
Definitions                1 0.76  0.12  0.12 -0.06 0.61 0.391 1.1
Analogies                  7 0.74  0.04  0.23  0.04 0.61 0.392 1.2
Classification             3 0.74  0.03  0.16 -0.15 0.60 0.402 1.2
Artificial_Languange       4 0.62  0.20  0.25 -0.03 0.49 0.513 1.6
Arithmetical_Problems      2 0.53  0.48 -0.03  0.09 0.52 0.479 2.0
Number_Series_Completion   6 0.44  0.75  0.02  0.01 0.76 0.242 1.6
```

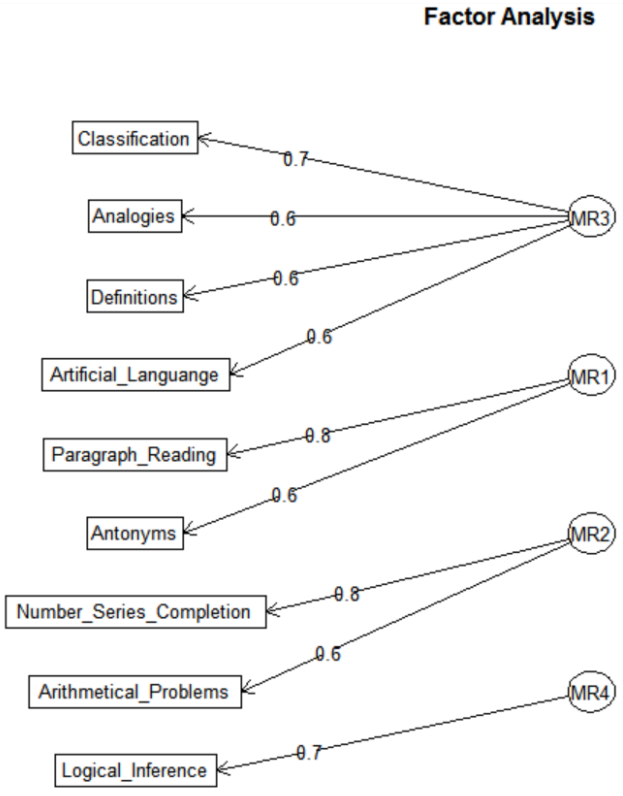
이 분석에서는 4-factor varimax 요인분석 결과가 가장 좋은 설명력을 갖는 것으로 생각되어, 해당 분석의 결과를 채택한다.

```
print(fa1, digits = 2, sort = T)
```

```
                       MR3  MR1  MR2  MR4
SS loadings           2.22 1.55 1.49 0.77
Proportion Var        0.25 0.17 0.17 0.09
Cumulative Var        0.25 0.42 0.59 0.67
Proportion Explained  0.37 0.26 0.25 0.13
Cumulative Proportion 0.37 0.62 0.87 1.00
```

처음 3개의 공통요인으로 전체 데이터의 87%를 설명할 수 있다.

```
fa.diagram(fa1)
```



공통요인 MR3는 "분석능력", MR1는 "독해능력", MR2는 "수리능력", MR4는 "논리력" 으로 정의할 수 있다.

B. 잠재요인에 의해 가장 설명이 잘되는 원변수와 가장 설명이 안되는 원변수를 찾으시오.

```
Factor Analysis using method =  minres
Call: fa(r = df, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                        item  MR3  MR1  MR2  MR4   h2    u2 com
Classification             3 0.66 0.33 0.21 0.11 0.60 0.402 1.8
Analogies                  7 0.64 0.25 0.21 0.29 0.61 0.392 2.0
Definitions                1 0.61 0.34 0.30 0.19 0.61 0.391 2.4
Artificial_Languange       4 0.57 0.14 0.33 0.17 0.49 0.513 2.0
Paragraph_Reading          9 0.32 0.84 0.20 0.24 0.90 0.099 1.6
Antonyms                   5 0.57 0.60 0.15 0.18 0.74 0.255 2.3
Number_Series_Completion   6 0.22 0.09 0.83 0.09 0.76 0.242 1.2
Arithmetical_Problems      2 0.26 0.23 0.60 0.22 0.52 0.479 2.0
Logical_Inference          8 0.37 0.35 0.30 0.68 0.82 0.183 2.6
```

위 결과에서 h2는 각 변수가 전체 공통요인에 대해 차지하는 공통성(communality)을 나타내며, u2는 특정요인(specific factor)의 분산 즉, uniquiness를 나타낸다.
그러므로 잠재요인에 의해 가장 설명이 잘되는 원변수는 communality 값이 가장 큰 Paragraph Reading 이며, 가장 설명이 안되는 원변수는 uniquiness 값이 가장 큰 Artificial Language 이다.

C. 각 잠재요인이 데이터의 변동을 설명해 주는 비율을 계산하시오.

```
print(fa1, digits = 2, sort = T)
```

```
                       MR3  MR1  MR2  MR4
SS loadings           2.22 1.55 1.49 0.77
Proportion Var        0.25 0.17 0.17 0.09
Cumulative Var        0.25 0.42 0.59 0.67
Proportion Explained  0.37 0.26 0.25 0.13
Cumulative Proportion 0.37 0.62 0.87 1.00
```

각 잠재요인이 데이터의 변동을 설명해 주는 비율은 위 도표에서 Proportion Explained 항목이 나타내고 있다.
첫번째 잠재요인 MR3는 전체의 37%, MR1은 26%, MR2는 25%, 그리고 MR4는 13% 만큼 전체 데이터의 변동을 설명하고 있다.

감사합니다.