이름 (이병준)

중간고사

다변량통계분석

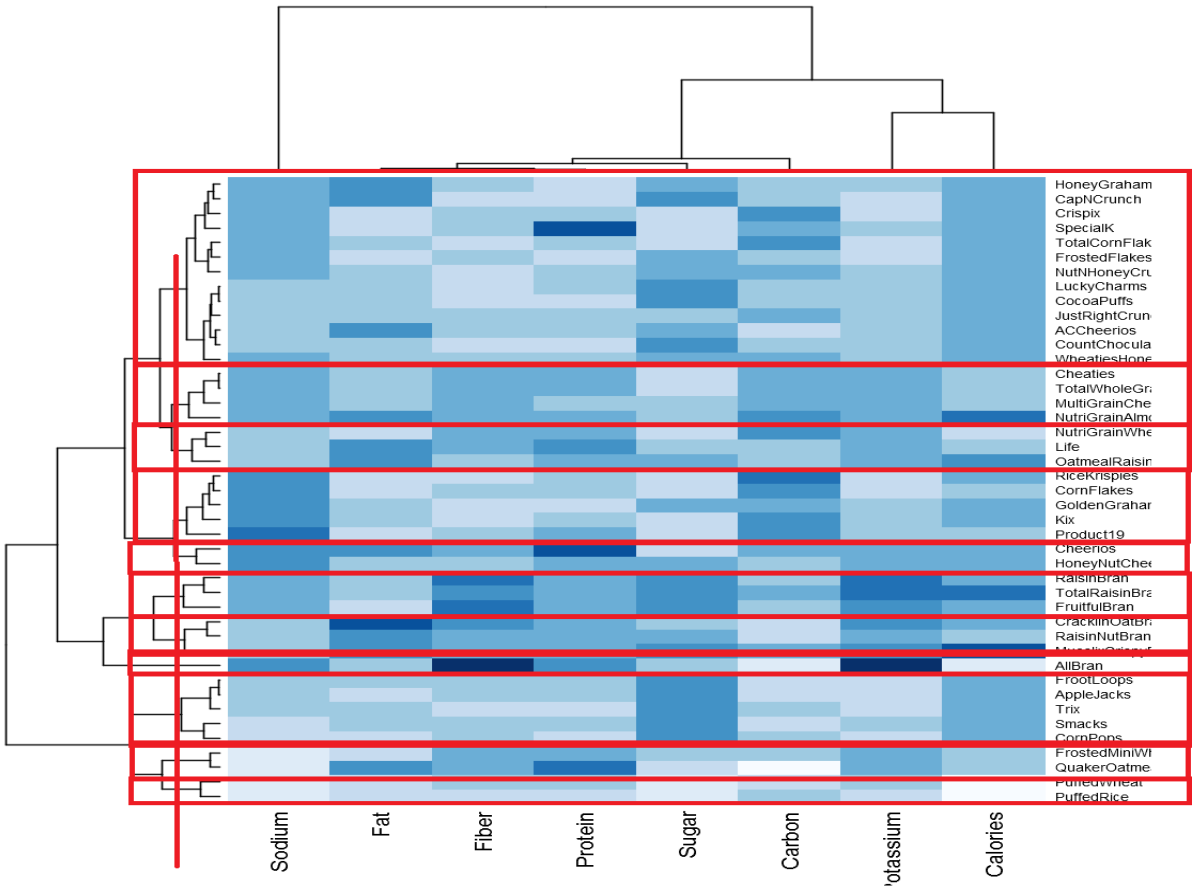
2016년 2학기

- **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.**

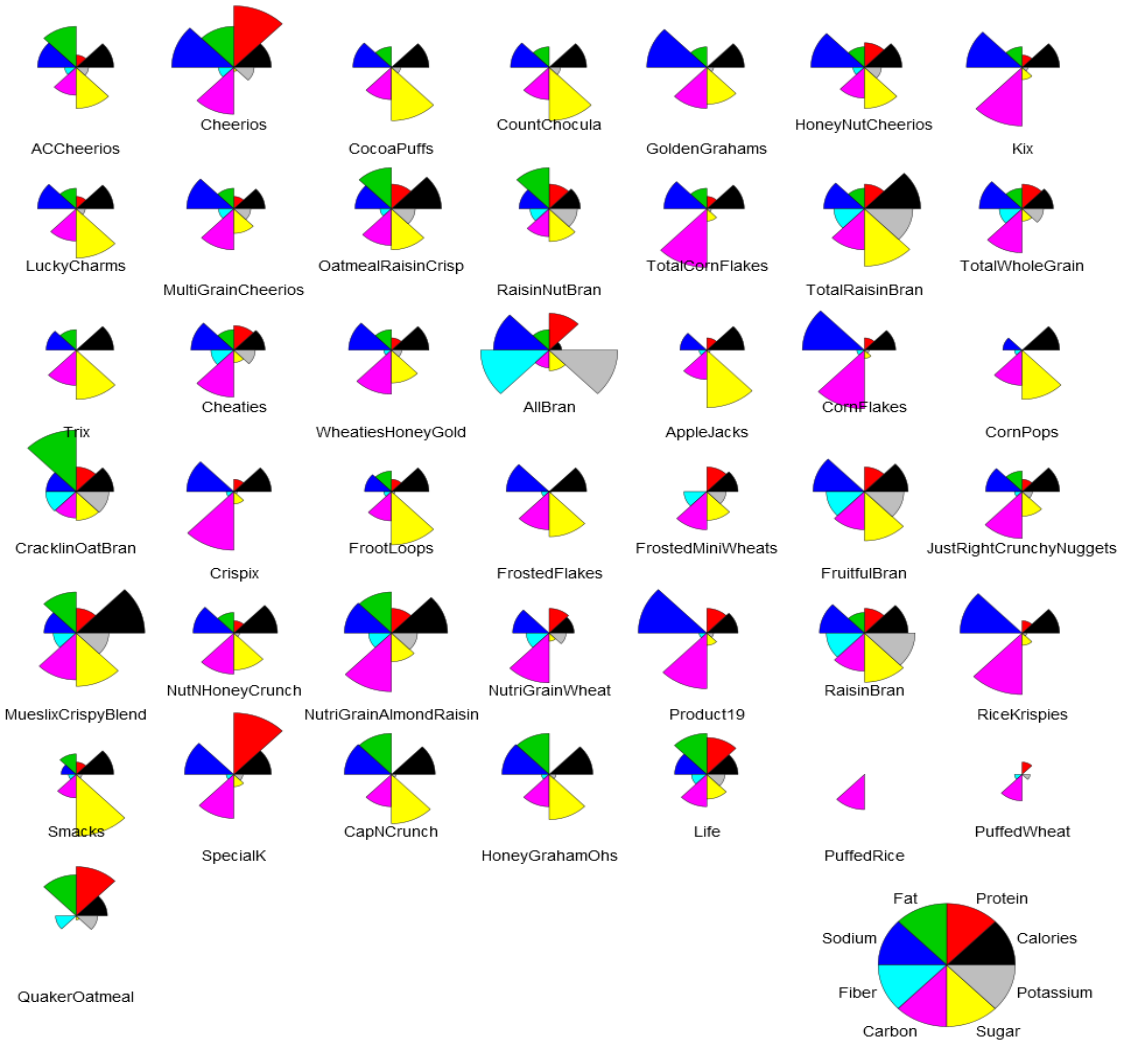
1. Cereal.csv는 3개의 미국 시리얼 제조사(General Mills: G, Kellogg: K, Quaker: Q)에 의해 생산되는 아침식사용 시리얼 각 브랜드의 영양성분 자료이다.

   A. 영양성분 상 특성을 시리얼 별로 한눈에 비교하기 위한 그래프를 그린 후 비슷한 영양성분을 가지는 시리얼들을 탐색적으로 구분하여 서술하시오.

      i. 각 성분에 대해서 전반적으로 군집화를 한번에 시각적으로 보려면 지난 수업시간에 배운 HeatMap을 이용하여 구분이 가능합니다. (나누는 기준은 분석가 기준에 따라 달라집니다.)



ii. 군집화 시킨 시리얼 끼리의 주성분의 유사도를 그래프로 보기 위해서는 Star그래프를 활용하면 해당 부분의 비율 또는 크기에 따른 비교가 가능합니다.



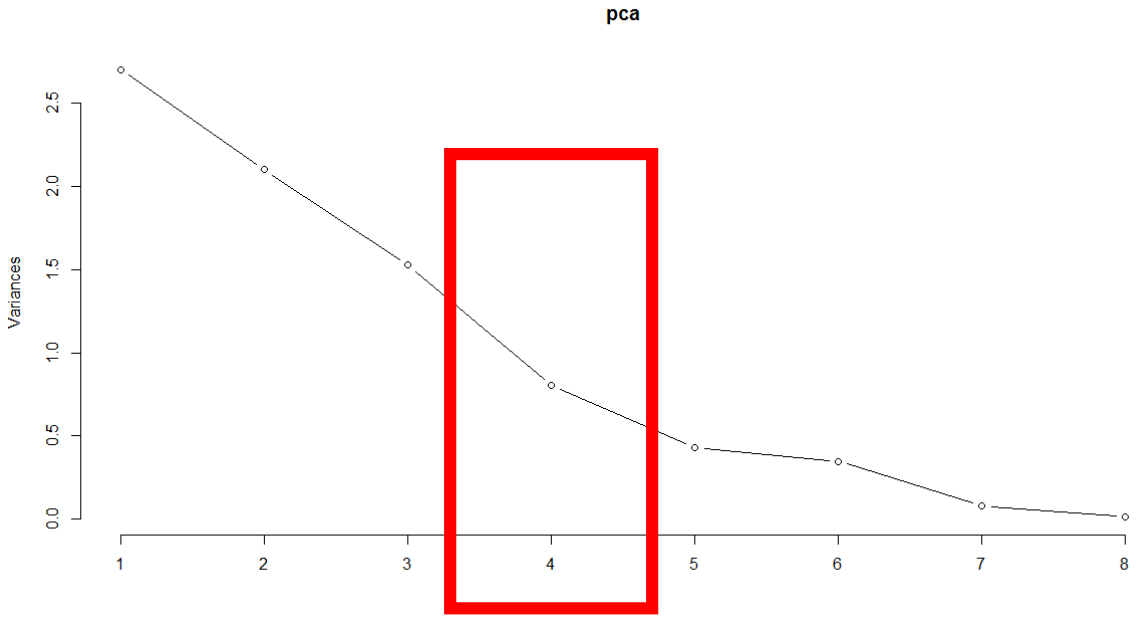
B. 8개의 영양성분 상의 특성을 보다 적은 차원에서 설명하기 위해 주성분분석을 활용하여 분석을 진행하시오. 적절한 그래프와 결과물을 사용하여 아래의 문항에 답변하시오.

i. 적절한 주성분의 개수는 무엇인가? 4개로 결정. (이상치 제거 후 돌려본 결과)

1. 1개의 이상치 제거 이후 PCA 결과 설명 비중으로 결정 : PC3~PC4에서 결정

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.6430 1.448  1.2351 0.8962 0.55722 0.58996 0.27976 0.12033
Proportion of Variance 0.3374 0.262  0.1907 0.1004 0.05399 0.04351 0.00978 0.00181
Cumulative Proportion  0.3374 0.599  0.7905 0.8909 0.94490 0.98841 0.99819 1.00000
```

2. Plot을 통한 결정 : PC1~4



3. 결론 : 4개의 주성분으로 선택.

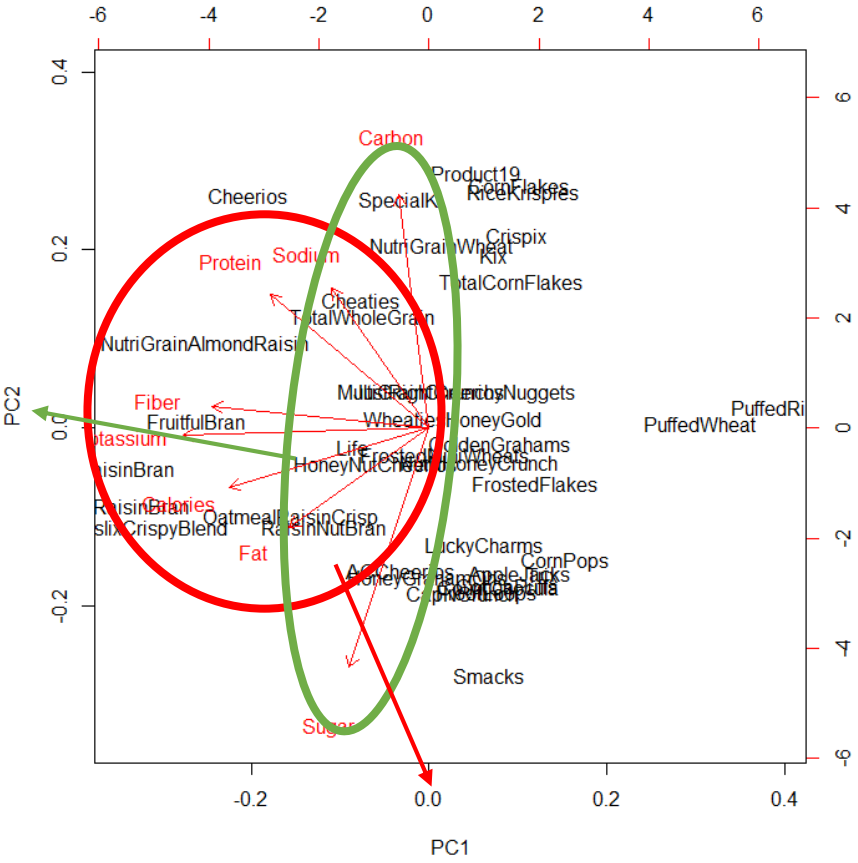
ii. 각 주성분은 어떤 의미를 가지는가?

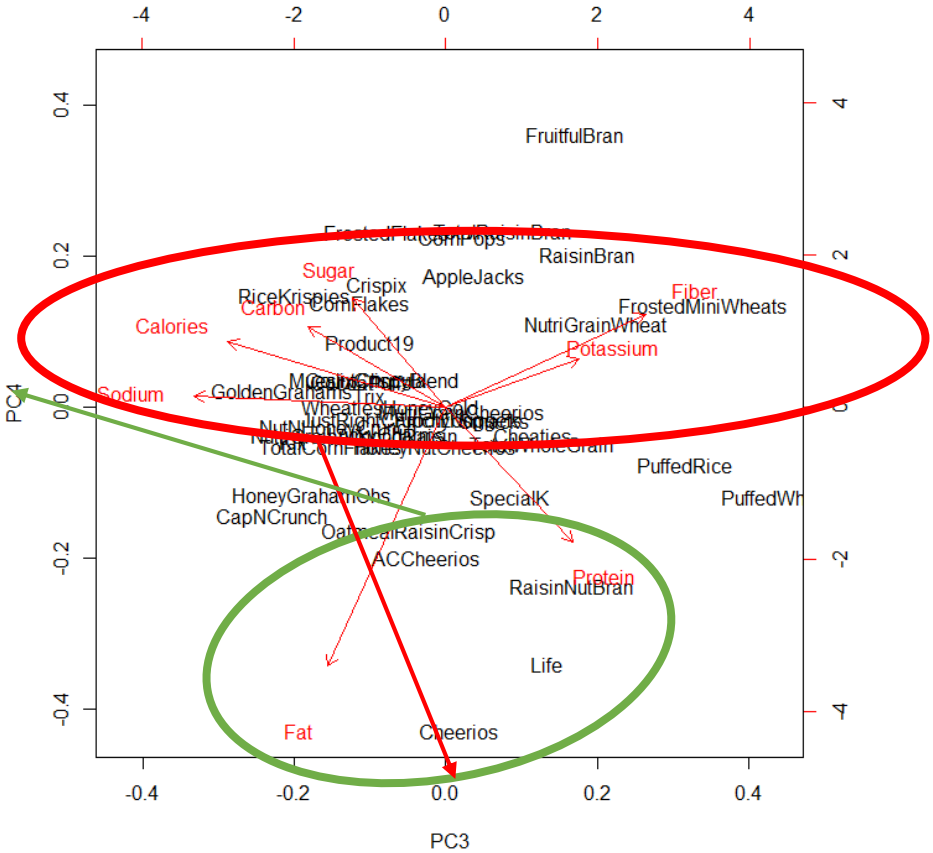
1. PC 1: PC1에서 모든 요소들이 음의 관계를 가진다. 그 중에서도 섬유소, 칼륨, 칼로리, 단백질, 지방과 강한 관계를 가지며 나머지 부분에서는 약한 관계를 가지고 있다.

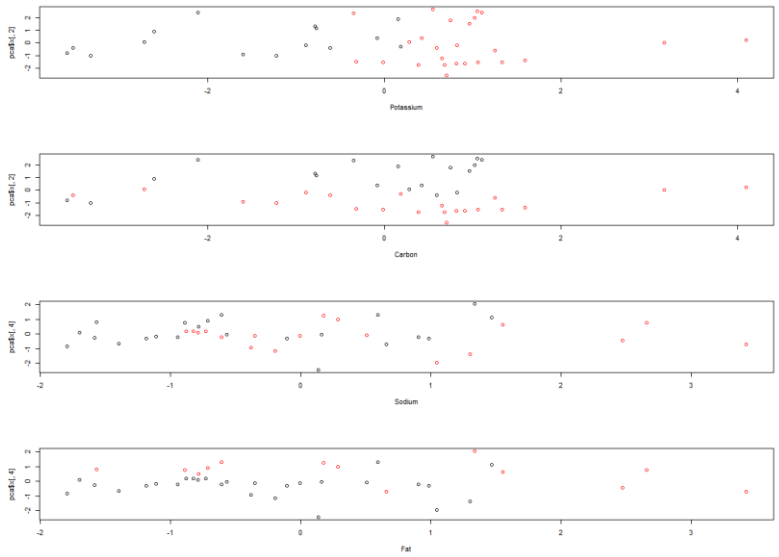
2. PC 2: 설탕과 강한 음의 관계를 가지고 있으며, 칼로리와 지방과도 약한 음의 관계를 가진다. 단백질과 나트륨, 탄수화물과는 강한 양의 관계를 가진 요소이다.

3. PC 3 : 칼로리, 나트륨, 섬유질과 강한 관계를 가진다. 칼로리와 나트륨과는 음의 관계를 가지고, 섬유질과는 양의 관계를 가진다.

4. PC 4 : 지방과 단백질에 강한 음의 관계를 가지고 있다. 섬유소, 설탕, 칼륨, 탄수화물과는 약한 양의 관계를 가진다.



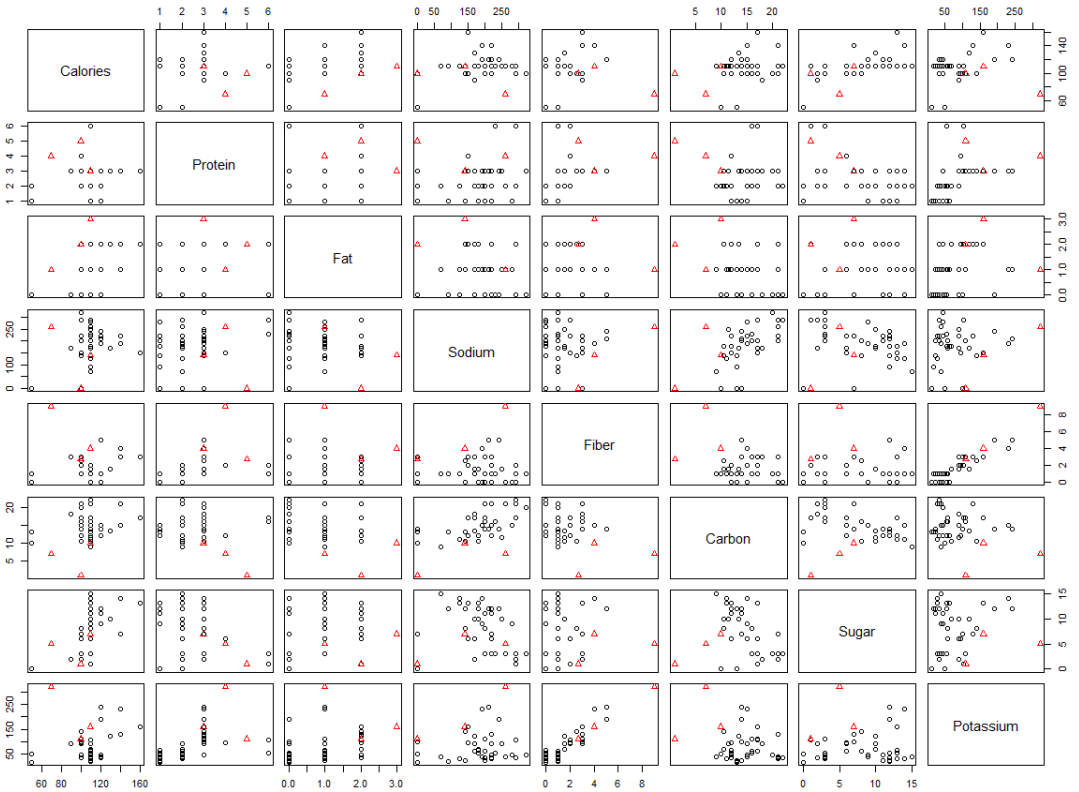




각 주성분 중 영향도 높은 성분의 평균 이상은 붉은색으로 구분 가능.

iii. 이상치가 있는가? 있다면 어떤 성질을 가지는가?

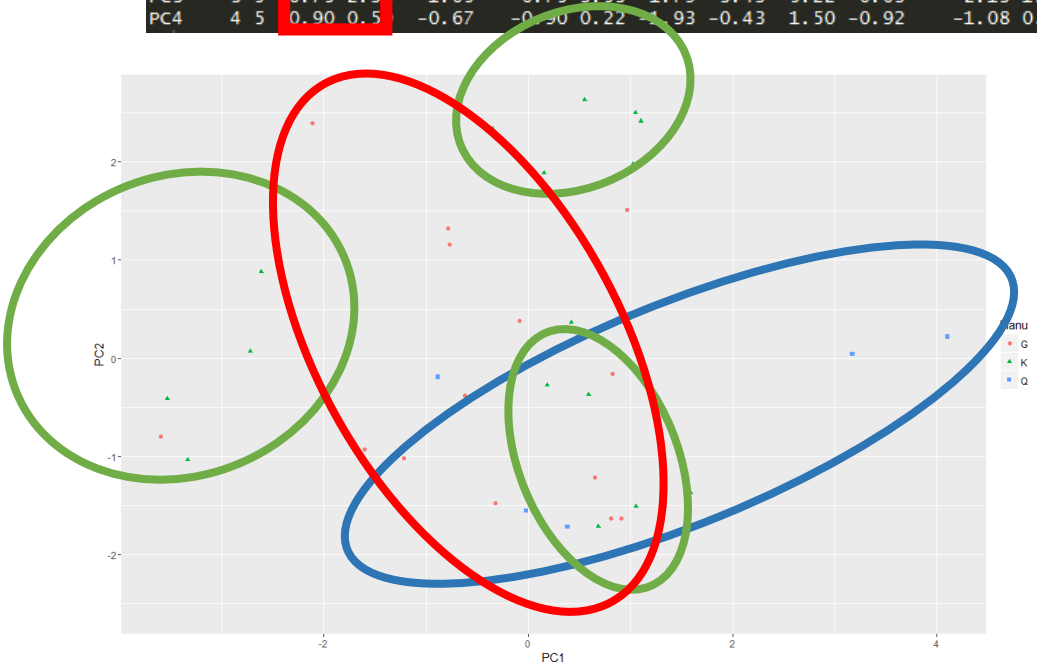
1. 존재한다. **AllBran,CracklinOatBran,QuakerOatmeal**
   A. **AllBran** : 다이어트용 시리얼 같은 구성을 가지고 있다.
   섬유소, 칼륨, 나트륨, 단백질 함량이 높고 설탕, 지방, 탄수화물, 함량이 낮아 칼로리가 낮은 모양을 하고 있다.
   B. **CracklinOatBran**
   다른 시리얼 보다 지방함량이 월등히 많다.
   C. **QuakerOatmeal**
   다른 시리얼 보다 탄수화물 함량이 월등이 낮다.



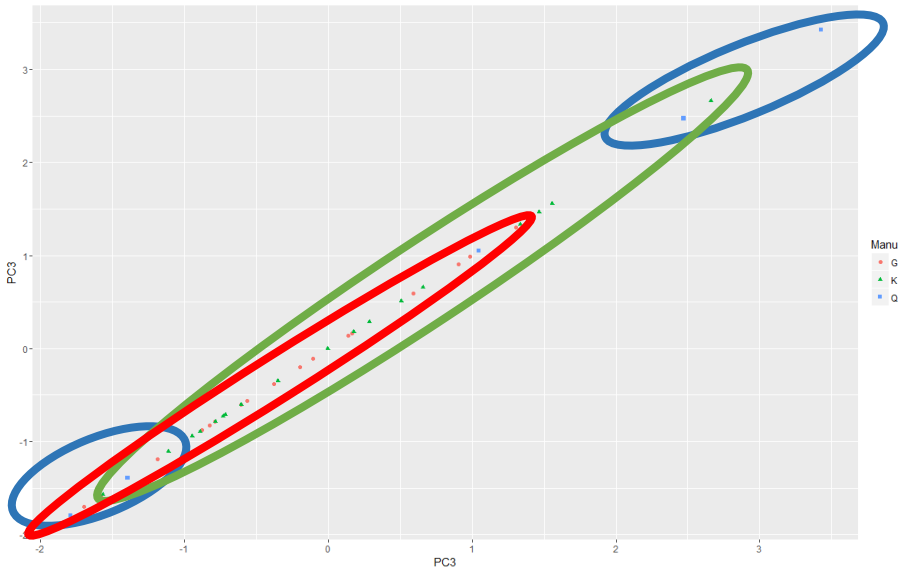
iv. 주성분 분석의 결과를 활용하여 볼 때 각 제조사가 생산하는 시리얼 별로 영양 성분 상의 특성이 다른가?

1. 아래의 결과를 확인하면 수치상으로 각 요소별로 제조하는 시리얼의 종류가 다르다고 판단됩니다.

```
INDICES: G
    vars  n  mean    d median trimmed  mad   min  max range  skew kurtosis   se
PC1    1 17 -0.27 1.2  -0.09   -0.15 1.33 -3.60 1.33  4.93 -0.88    -0.02 0.32
PC2    2 17 -0.12 1.2  -0.38   -0.19 1.62 -1.63 2.39  4.02  0.45    -1.33 0.32
PC3    3 17 -0.28 0.8  -0.38   -0.29 0.77 -1.70 1.30  3.00  0.18    -1.08 0.21
PC4    4 17 -0.32 0.0  -0.22   -0.29 0.45 -2.43 1.33  3.75 -0.69     1.06 0.20
------------------------------------------------------------------------------
INDICES: K
    vars  n  mean    d median trimmed  mad   min  max range  skew kurtosis   se
PC1    1 18 -0.12 1.8   0.57   -0.01 0.71 -3.53 1.59  5.12 -1.05    -0.59 0.40
PC2    2 18  0.29 1.6  -0.10    0.33 1.99 -2.55 2.63  5.17  0.07    -1.45 0.39
PC3    3 18  0.05 1.3  -0.18   -0.01 1.04 -1.57 2.66  4.23  0.65    -0.61 0.27
PC4    4 18  0.56 0.9   0.70    0.54 0.79 -0.69 2.04  2.73  0.14    -0.79 0.16
------------------------------------------------------------------------------
INDICES: Q
    vars n  mean   s  median trimmed  mad   min   max range  skew kurtosis   se
PC1    1 5  1.35 2.1    0.39    1.35 1.89 -0.89  4.10  4.99  0.24    -2.11 0.97
PC2    2 5  0.64 0.9   -0.19   -0.64 0.60 -1.72  0.22  1.93 -0.25    -2.21 0.41
PC3    3 5  0.75 2.3    1.05    0.75 3.53 -1.79  3.43  5.22 -0.03    -2.13 1.03
PC4    4 5  0.90 0.5   -0.67   -0.90 0.22 -1.93 -0.43  1.50 -0.92    -1.08 0.26
```



- 그래프로 다시 한 번 확인한다면 PC1 과 PC2 를 통해 주된 생성 라인이 다르게 포진되어 있음을 확인 할 수 있다.



- 각 사가 전반적으로 비슷비슷한 PC3, PC4 함유량을 가진 시리얼을 생산한다. 하지만 K,Q 사가 G 사에 비해 PC3, PC4 함류량이 높은 제품군들을 생성한다.

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

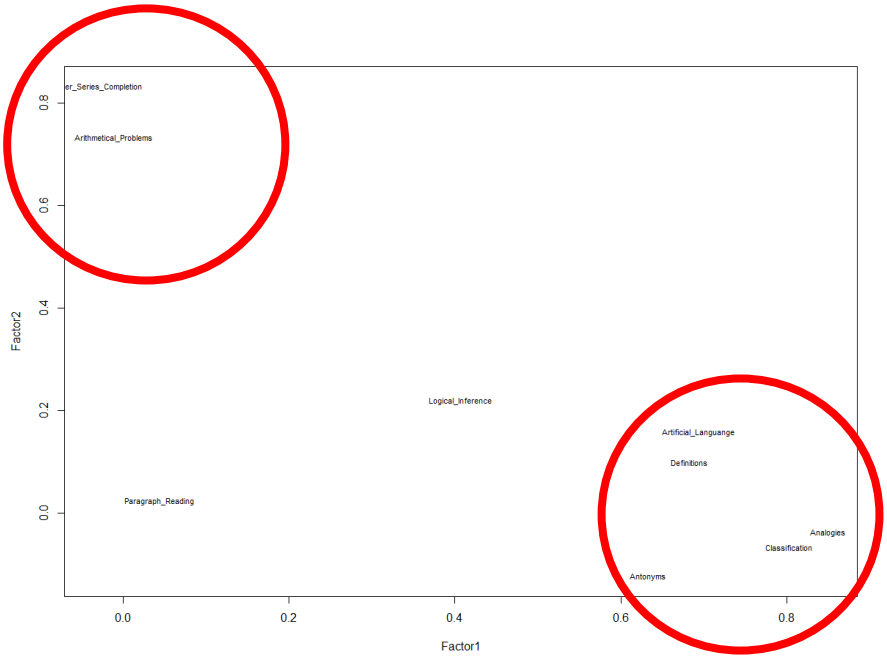
   i. 이 데이터를 사용하여 요인분석을 진행하여 9개의 테스트 결과에 영향을 주는 잠재요인을 파악하시오. (적절한 요인 개수와 요인회전 고려)

      1. 요인 개수 : 3개
      2. 회전 : 일반화에 중점을 두기 위해 직교회전 선택 (Promax)
         회전 사용 이유 : 하나의 원변수에 부하 값이 큰 인자가 2개 이상이 존재하는 값들을 발견.

```
Loadings:
                       Factor1 Factor2 Factor3
Definitions             0.68
Arithmetical_Problems           0.73
Classification          0.80
Artificial_Languange    0.69    0.16   -0.18
Antonyms                0.63   -0.12    0.36
Number_Series_Completion        0.83
Analogies               0.85
Logical_Inference       0.41    0.22    0.23
Paragraph_Reading                       0.95

                Factor1 Factor2 Factor3
SS loadings        2.88    1.33    1.15
Proportion Var     0.32    0.15    0.13
Cumulative Var     0.32    0.47    0.60
```

      3. 사회 과학문제에서는 총 분산이 60% 정도까지 허용.
      4. P-value값 : 0.0624 귀무 가설 채택. (2개인의자를 할 때도 되지만 설명도가 낮음)



비슷한 성향

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

1. 잘되는 변수 : Paragraph_Reading 개별성이 0.005로 가장 설명이 잘되는 변수이다.
2. 잘 안되는 변수 : Artificial_Language 개별성이 0.51이상으로 가장 설명이 안되는 변수이다.

```
> fa2$uniquenesses
          Definitions     Arithmetical_Problems               Classification
            0.3896150                 0.4119660                    0.4199499
Artificial_Languange                  Antonyms Number_Series_Completion
            0.5147419                 0.2781807                    0.3829876
            Analogies       Logical_Inference          Paragraph_Reading
            0.3969107                 0.4367040                    0.0050000
```

iii. 각 잠재요인이 데이터의 변동을 설명해 주는 비율을 계산하시오.

```
                Factor1 Factor2 Factor3
SS loadings        2.88    1.33    1.15
Proportion Var     0.32    0.15    0.13
Cumulative Var     0.32    0.47    0.60
```

Factor1 : 0.32

Factor2 : 0.15

Factor3 : 0.13