# 2025학년도 1학기 중간과제물(온라인 제출용)

- **교 과 목 명** : 다변량분석
- **학 번** : 202234-153799
- **성 명** : 한승환
- **연 락 처** : 010-2862-0200

※ A4용지 편집 사용

※ 과제물 표지등에 개인정보(주민번호,운전면허번호)가 포함될 경우 삭제처리로 과제물을 다시 제출해야 하는 경우가 발생할 수 있습니다.

---

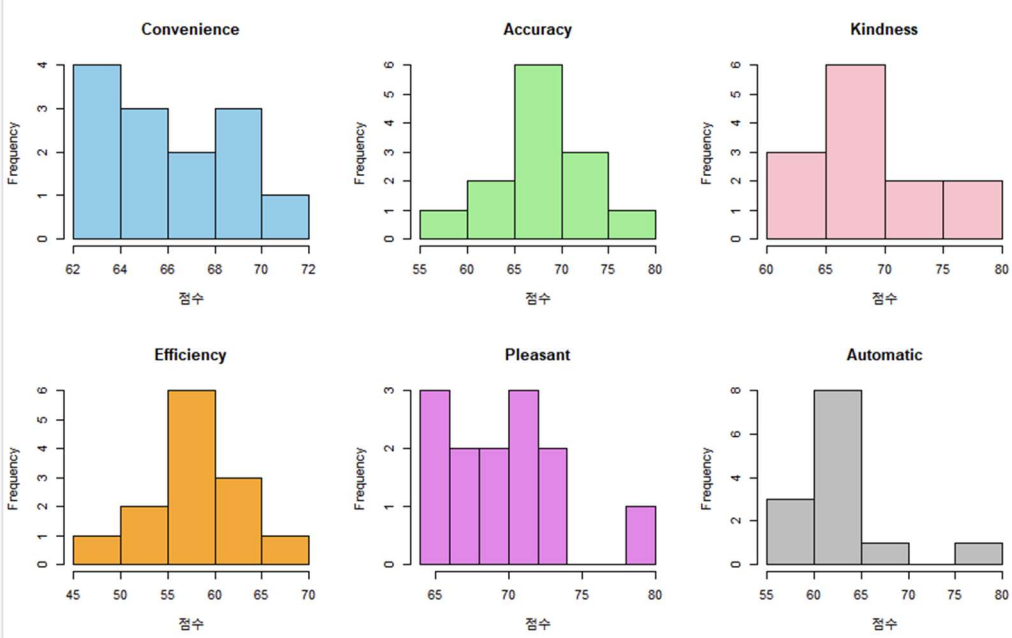
## 문제 1번

데이터 프레임 생성

```r
# 데이터프레임 생성
bank <- data.frame(
  name = c("kookmin", "enterpr", "boram", "commerce", "seoul", "shinhan", "city", "exchange",
           "first", "chohung", "hana", "hanil", "sega"),
  convenience = c(70.5, 64.8, 67.1, 62.1, 63.6, 64.2, 69.2, 68.1, 66.6, 63.5, 69.7, 63.2, 64.1),
  accuracy = c(59.4, 70.3, 79.6, 65.0, 66.5, 66.6, 72.0, 67.5, 65.4, 65.7, 74.5, 65.5, 64.8),
  kindness = c(63.7, 68.6, 78.5, 65.6, 65.7, 71.2, 71.4, 67.3, 65.2, 63.8, 75.6, 63.6, 67.8),
  efficiency = c(54.3, 55.2, 62.4, 54.4, 60.3, 59.6, 56.9, 60.5, 58.3, 56.5, 66.7, 49.6, 59.7),
  pleasant = c(66.9, 68.3, 79.4, 70.5, 67.5, 71.8, 72.8, 71.3, 68.4, 65.9, 73.6, 65.5, 65.7),
  automatic = c(62.6, 62.3, 62.4, 63.9, 63.6, 67.2, 57.8, 60.2, 61.7, 55.9, 75.4, 57.1, 61.8)
)
```

### (1) 히스토그램 그리기

```r
# 히스토그램 그리기
par(mfrow=c(2,3))  # 2행 3열로 나눔
hist(bank$convenience, main="Convenience", col="skyblue", xlab="점수")
hist(bank$accuracy, main="Accuracy", col="lightgreen", xlab="점수")
hist(bank$kindness, main="Kindness", col="pink", xlab="점수")
hist(bank$efficiency, main="Efficiency", col="orange", xlab="점수")
hist(bank$pleasant, main="Pleasant", col="violet", xlab="점수")
hist(bank$automatic, main="Automatic", col="grey", xlab="점수")
```




편리성(Convenience):

전반적으로 균형있게 분포하지만 상위점수대(70이상)는 상대적으로 적음

신속성(Accuracy):

65~70점대 사이가 가장 많은 분포를 형성하며, 일부 소수의 은행이 두드러지게 우수함

친절(Kindness):

65~70점대에 가장 많은 분포를 형성하고, 일부은행은 75점 이상의 점수로 차이가 큼

능률(Efficiency):

전반적으로 낮은 점수로 분포 있으며, 점수 차이가 가장 뚜렷함

쾌적함(Pleasant):

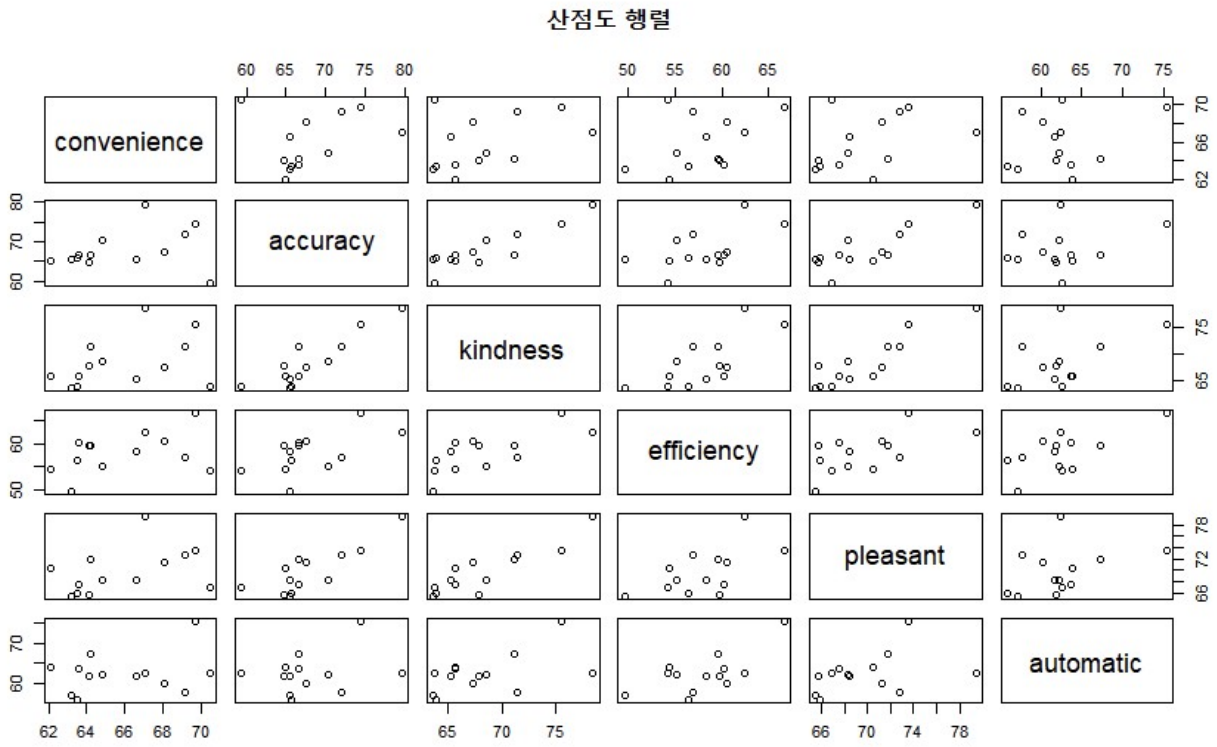
65~75점 사이에 고르게 분포해 있고, 대부분 평균이상을 유지

자동화(Automatic):

전반적으로 낮은 점수에 분포해 있고, 75점 이상의 점수를 받은 은행들은 극소수

**(2) 산점도행렬 및 상관계수행렬과 변수들의 관계 설명**

```
22  # 산점도 행렬 (pairs plot)
23  pairs(bank[,-1], main="산점도 행렬")
24
25  # 상관계수 행렬
26  cor_matrix <- cor(bank[,-1])
27  round(cor_matrix, 2)  # 소수 둘째 자리까지 보기 좋게
```



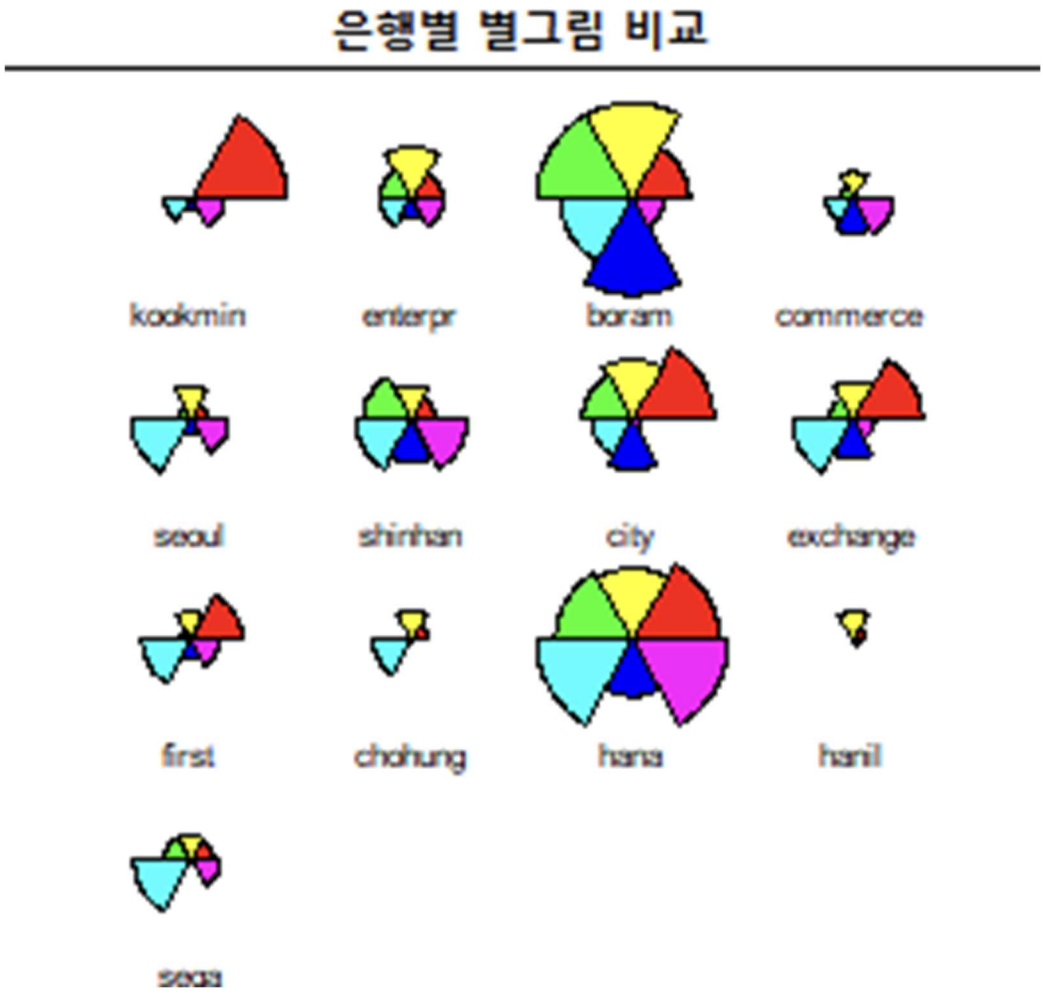

변수들의 관계를 직관적으로 관찰할 수 있으며, kindness와 pleasant, accuracy 사이에는 뚜렷한 대각선상으로 점들이 분포해 있는 것으로 보아 서로 양의 상관관계가 있는 것을 알 수 있습니다.

```
> # 상관계수 행렬
> cor_matrix <- cor(bank[,-1])
> round(cor_matrix, 2)  # 소수 둘째 자리까지 보기 좋게
            convenience accuracy kindness efficiency pleasant automatic
convenience        1.00     0.25     0.38       0.39     0.41      0.27
accuracy           0.25     1.00     0.90       0.57     0.83      0.28
kindness           0.38     0.90     1.00       0.71     0.89      0.52
efficiency         0.39     0.57     0.71       1.00     0.59      0.67
pleasant           0.41     0.83     0.89       0.59     1.00      0.39
automatic          0.27     0.28     0.52       0.67     0.39      1.00
```

| 변수 쌍 | 상관계수 | 해석 |
| --- | --- | --- |
| accuracy-kindness | 0.9 | 매우 강한 양의 상관관계<br>신속한 은행일수록 친절하다는 평가를 가짐 |
| kindness-pleasant | 0.89 | 매우 강한 양의 상관관계<br>친절한 은행일수록 쾌적한 환경을 제공 |
| accuracy-pleasant | 0.83 | 강한 양의 상관관계<br>신속한 은행일수록 쾌적한 환경을 제공 |
| kindness-efficiency | 0.71 | 강한 양의 상관관계<br>친절한 은행일수록 높은 능률을 가짐 |
| efficiency-automatic | 0.67 | 강한 양의 상관관계<br>능률이 높은 은행일수록 자동화의 수준도 높음 |

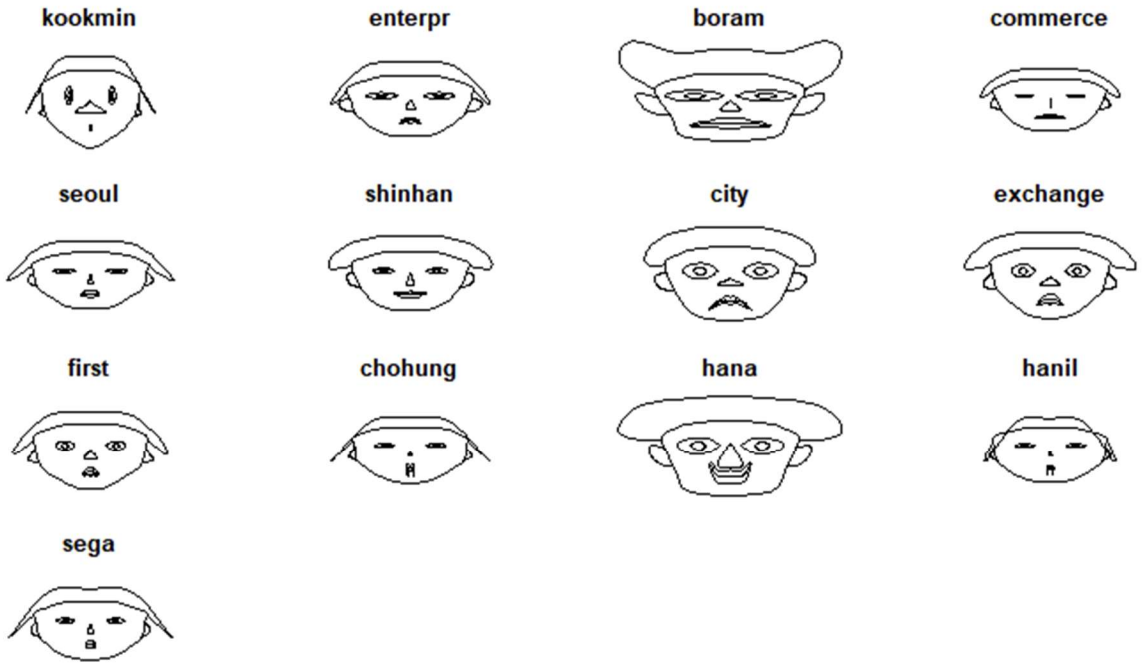
### (3) 별그림 및 얼굴그림

```
29 # 별그림 그리기기
30 stars(bank[,-1], labels=bank$name, main="은행별 별그림", draw.segments=TRUE)
31
32 # 얼굴그림 패키지 설치 및 불러오기
33 install.packages("TeachingDemos")  # 한 번만 설치
34 library(TeachingDemos)
35
36 # 얼굴그림 그리기
37 faces(bank[,-1], labels=bank$name, main="은행별 얼굴그림")
```




boram과 hana 은행이 많은 항목에서 높은 점수를 기록하였고 전반적으로 우수한 은행임을 알 수 있습니다. shinhan은 전체적으로 고른 성적이며 균형이 잘 잡힌 은행이라는 것을 볼 수 있습니다.

은행별 얼굴그림




친절함과 쾌적함이 높은 수준의 은행이 더 밝은 표정을 하는 것을 관찰가능

## 문제 2번

파이썬 코드

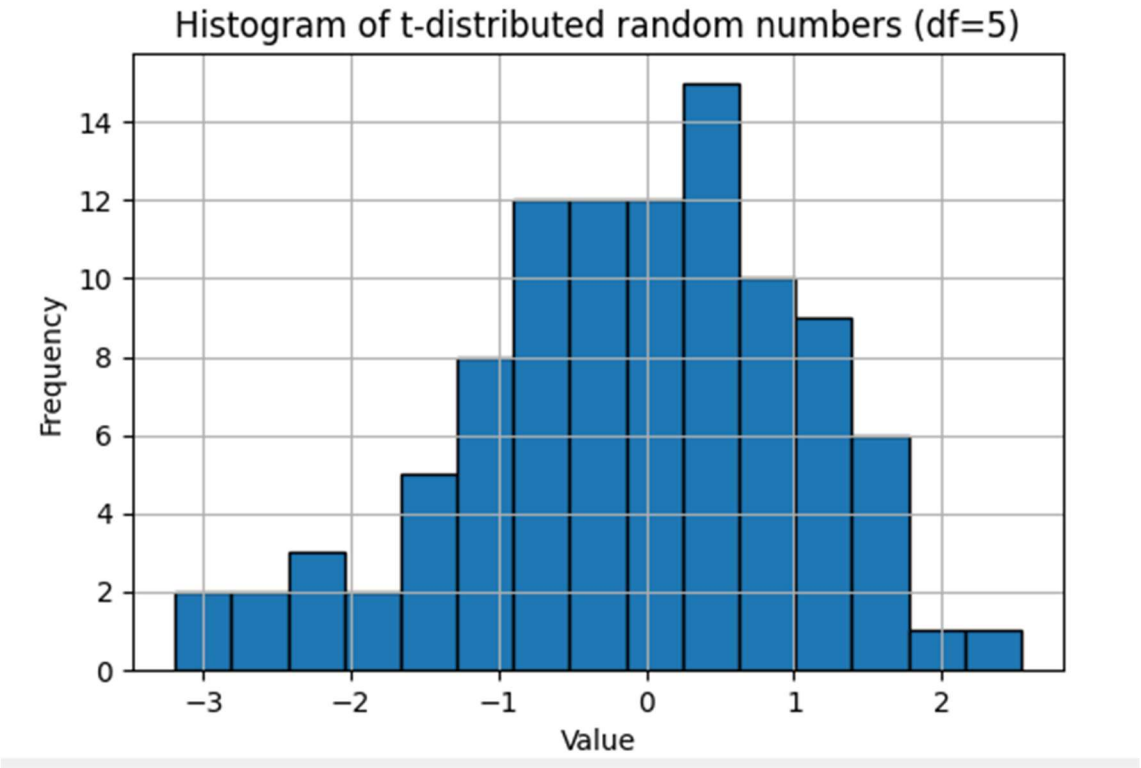
```python
import numpy as np
import matplotlib.pyplot as plt
import stemgraphic

# 난수 생성
np.random.seed(0)
data = np.random.standard_t(df=5, size=100)

# 히스토그램 그리기
plt.figure(figsize=(6, 4))
plt.hist(data, bins=15, edgecolor='black')
plt.title('Histogram of t-distributed random numbers (df=5)')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()

# 상자 그림 그리기
plt.figure(figsize=(6, 2))
plt.boxplot(data, vert=False)
plt.title('Boxplot of t-distributed random numbers (df=5)')
plt.xlabel('Value')
plt.grid(True)
plt.show()

# 줄기-잎 그림 그리기
fig, ax = stemgraphic.stem_graphic(
    data,
    scale=1,
)
fig.set_size_inches(10, 5)
plt.show()
```

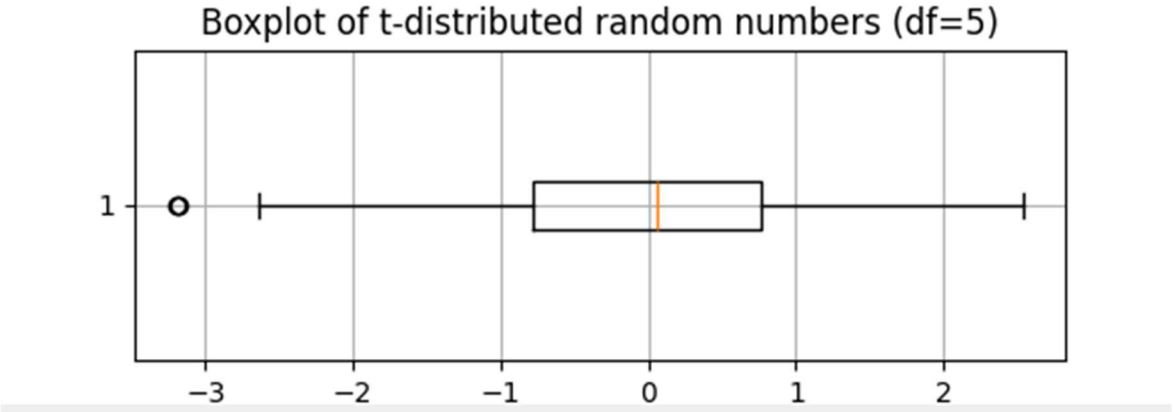



가로축: 난수 값, 자유도 5인 t-분포

세로축: 해당 구간에 속하는 데이터 개수

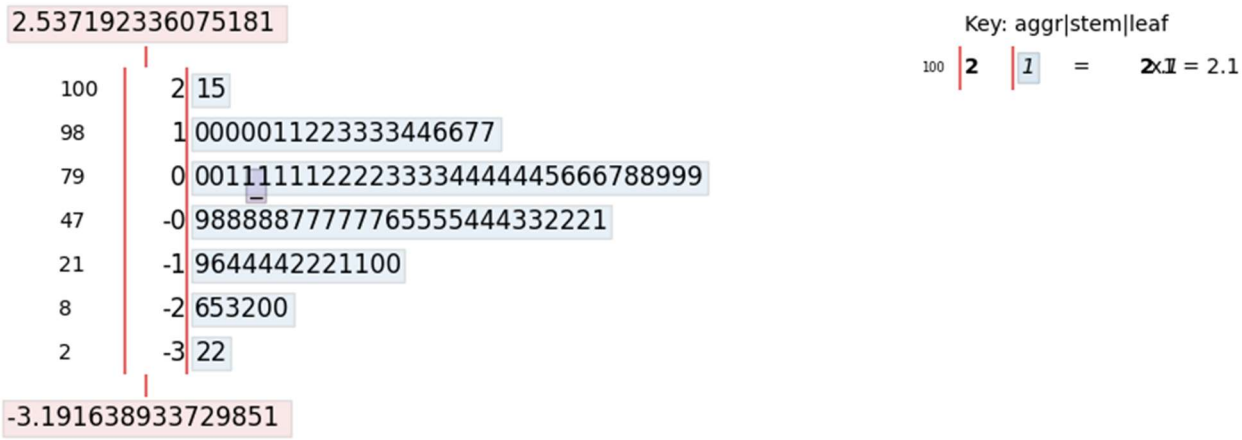
중심이 0에 가까우며 좌우 대칭에 가깝습니다.

양 끝단에 데이터가 소수 존재하며 극단 값이 나올 수 있는 분포라는 것을 시사합니다.




중앙값은 0 근처이며, 이상치가 왼쪽 -3이하에 존재합니다. t-분포의 특성으로 볼 수 있습니다.

분포가 대칭적이지만 약간의 치우침을 볼 수 있습니다.




값이 0.0~0.9 구간에 몰려 있음을 볼 수 있습니다. -1, -2에도 분포되어 있어서 좌우 대칭에 가까운 모양입니다.

## 문제 3번

### 1. R을 이용한 주성분분석

#### (1) 기술 통계량 분석

```r
library(ade4)
data(deug)
deug_tab <- deug$tab

summary(deug_tab)
```

```
> summary(deug_tab)
    Algebra         Analysis         Proba         Informatic       Economy         Option1         Option2
 Min.   : 9.00   Min.   :16.00   Min.   : 2.00   Min.   :10.50   Min.   :34.50   Min.   : 8.00   Min.   : 5.00
 1st Qu.:39.00   1st Qu.:29.00   1st Qu.:22.00   1st Qu.:21.00   1st Qu.:64.50   1st Qu.:19.75   1st Qu.:20.75
 Median :46.00   Median :33.00   Median :29.50   Median :25.75   Median :70.20   Median :23.00   Median :23.50
 Mean   :45.57   Mean   :33.99   Mean   :31.24   Mean   :26.99   Mean   :69.55   Mean   :22.75   Mean   :22.67
 3rd Qu.:52.00   3rd Qu.:40.00   3rd Qu.:41.25   3rd Qu.:30.75   3rd Qu.:76.50   3rd Qu.:26.00   3rd Qu.:26.25
 Max.   :72.00   Max.   :58.00   Max.   :65.00   Max.   :54.00   Max.   :90.90   Max.   :34.00   Max.   :30.00
    English          Sport
 Min.   : 8.50   Min.   : 0.000
 1st Qu.:18.80   1st Qu.: 8.500
 Median :21.20   Median :11.500
 Mean   :21.13   Mean   : 9.231
 3rd Qu.:23.85   3rd Qu.:12.000
 Max.   :31.00   Max.   :15.000
```

### (2) 변수 간 상관계수행렬 계산

```
cor_matrix <- cor(deug_tab)
round(cor_matrix, 2)  # 소수점 둘째 자리까지 출력
```

```
> round(cor_matrix, 2)  # 소수점 둘째 자리까지 출력
           Algebra Analysis Proba Informatic Economy Option1 Option2 English Sport
Algebra       1.00     0.44  0.50       0.39    0.37    0.54    0.20    0.11  0.23
Analysis      0.44     1.00  0.52       0.32    0.21    0.40    0.06   -0.12  0.16
Proba         0.50     0.52  1.00       0.37    0.17    0.44    0.11    0.19  0.27
Informatic    0.39     0.32  0.37       1.00    0.08    0.25    0.09    0.13  0.06
Economy       0.37     0.21  0.17       0.08    1.00    0.37    0.34    0.41  0.18
Option1       0.54     0.40  0.44       0.25    0.37    1.00    0.20    0.09  0.26
Option2       0.20     0.06  0.11       0.09    0.34    0.20    1.00    0.02  0.08
English       0.11    -0.12  0.19       0.13    0.41    0.09    0.02    1.00  0.14
Sport         0.23     0.16  0.27       0.06    0.18    0.26    0.08    0.14  1.00
```

### (3) 고유값 및 누적기여도 구하기

```
pca_result <- prcomp(deug_tab, scale. = TRUE)

# 고유값
eigenvalues <- pca_result$sdev^2
eigenvalues

# 기여율 및 누적기여율
prop_var <- eigenvalues / sum(eigenvalues)
cum_var <- cumsum(prop_var)

# 결과 정리
data.frame(PC = 1:length(eigenvalues), Eigenvalue = eigenvalues,
           Proportion = round(prop_var, 4),
           Cumulative = round(cum_var, 4))
```

```
  PC Eigenvalue Proportion Cumulative
1  1  3.1013578     0.3446     0.3446
2  2  1.3629834     0.1514     0.4960
3  3  1.0323269     0.1147     0.6107
4  4  0.9340533     0.1038     0.7145
5  5  0.7397529     0.0822     0.7967
6  6  0.5746693     0.0639     0.8606
7  7  0.5325414     0.0592     0.9197
8  8  0.4375395     0.0486     0.9684
9  9  0.2847754     0.0316     1.0000
```

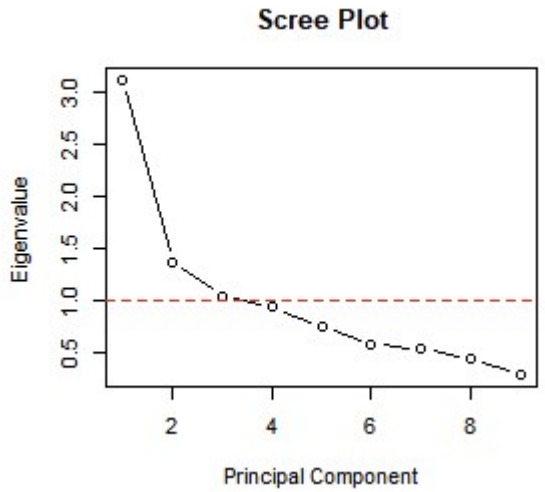
### (4) 고유값 > 1 인 주성분 개수 확인

```
which(eigenvalues > 1)
```

```
> which(eigenvalues > 1)
[1] 1 2 3
```

### (5) Scree Plot 그리기

```
27  plot(eigenvalues, type = "b", xlab = "Principal Component", ylab = "Eigenvalue",
28       main = "Scree Plot")
29  abline(h = 1, col = "red", lty = 2)
30
```



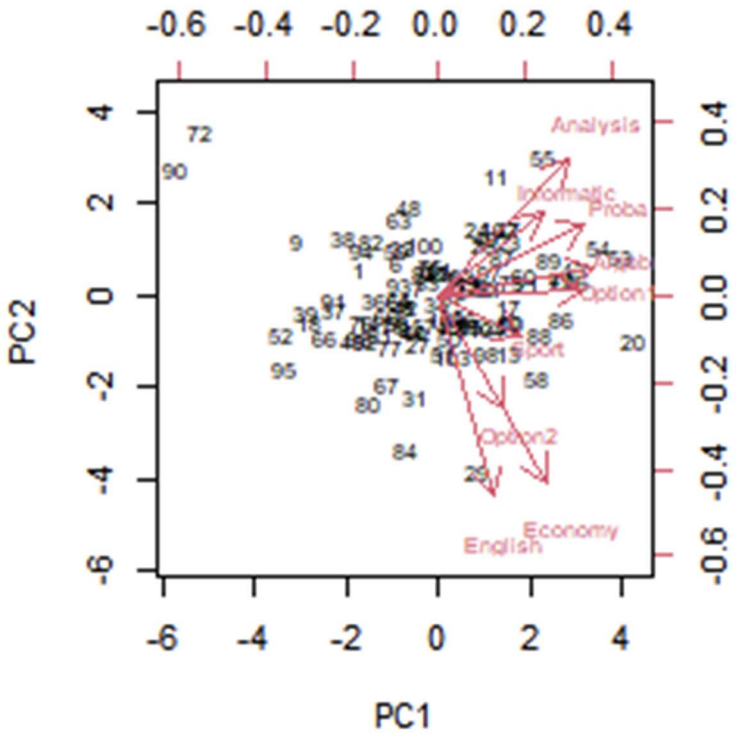

### (6) 주성분 score 구하기

```
31  pca_scores <- pca_result$x
32  head(pca_scores)
```

```
          PC1        PC2         PC3         PC4        PC5         PC6         PC7        PC8         PC9
1 -1.7350682  0.6008322  0.04581002 -0.44362273  1.5379945  0.04174513 -0.43576909  0.3705025 -0.37193406
2  0.4599338 -0.6052903  1.04002513  0.11183812  0.3544683  0.40201167  0.04969873 -0.9603250  0.03857141
3  0.5370767  0.1592945 -0.67703912  0.08174238  0.7951426  0.33559137  0.81155180 -1.1240716 -0.04119119
4  2.5297318  0.3785866  0.53126581 -0.11505442  0.5472921  0.21940170  0.03538862  1.0981909  0.95039601
5  0.2972450  0.3744154  0.85271949  0.49081432  1.5231699 -0.17113893 -0.19680178  0.8569704 -0.73343086
6 -0.8888046  0.7601720  0.73901974  0.90092739 -1.9926969  0.36019121 -0.30134908  0.7946717 -0.24344988
```

### (7) Biplot으로 시각화

```
34  biplot(pca_result, scale = 0)
```




PC1축: 언어/사회계열 중심의 성취도 축

PC2축: 수학/분석계열 중심이 성취도 축

벡터 간 각도

Economy ↔ English: 거의 같은 방향

Algebra ↔ Analysis: 유사한 방향

Analysis ↔ English: 서로 반대 방향

Sport: 비교적 중심에 위치하여 대부분 과목과 관련성 낮음

학생 분포: 서로 다른 경향으로 분포됨(수학형, 언어형 등)

**2. 파이썬을 이용한 주성분 분석 및 R의 결과와 비교**

**(1) 기술 통계량 분석**

| | Algebra | Analysis | Proba | Informatic | Economy | Option1 | Option2 | English | Sport |
|---|---|---|---|---|---|---|---|---|---|
| count | 104.000000 | 104.000000 | 104.000000 | 104.000000 | 104.000000 | 104.000000 | 104.000000 | 104.000000 | 104.000000 |
| mean | 45.567308 | 33.985577 | 31.240385 | 26.990385 | 69.550962 | 22.750000 | 22.668269 | 21.125962 | 9.230769 |
| std | 10.489722 | 8.811802 | 13.381987 | 8.247383 | 9.609744 | 5.204087 | 4.759564 | 4.261956 | 4.944313 |
| min | 9.000000 | 16.000000 | 2.000000 | 10.500000 | 34.500000 | 8.000000 | 5.000000 | 8.500000 | 0.000000 |
| 25% | 39.000000 | 29.000000 | 22.000000 | 21.000000 | 64.500000 | 19.750000 | 20.750000 | 18.800000 | 8.500000 |
| 50% | 46.000000 | 33.000000 | 29.500000 | 25.750000 | 70.200000 | 23.000000 | 23.500000 | 21.200000 | 11.500000 |
| 75% | 52.000000 | 40.000000 | 41.250000 | 30.750000 | 76.500000 | 26.000000 | 26.250000 | 23.850000 | 12.000000 |
| max | 72.000000 | 58.000000 | 65.000000 | 54.000000 | 90.900000 | 34.000000 | 30.000000 | 31.000000 | 15.000000 |

**(2) 변수 간 상관계수행렬 계산**

| | Algebra | Analysis | Proba | Informatic | Economy | Option1 | Option2 | English | Sport |
|---|---|---|---|---|---|---|---|---|---|
| Algebra | 1.000000 | 0.444965 | 0.504260 | 0.388580 | 0.365875 | 0.536707 | 0.196322 | 0.114005 | 0.234720 |
| Analysis | 0.444965 | 1.000000 | 0.516468 | 0.319952 | 0.206815 | 0.404404 | 0.061577 | -0.119735 | 0.158349 |
| Proba | 0.504260 | 0.516468 | 1.000000 | 0.372875 | 0.167062 | 0.444059 | 0.111853 | 0.186937 | 0.269368 |
| Informatic | 0.388580 | 0.319952 | 0.372875 | 1.000000 | 0.075644 | 0.249674 | 0.085804 | 0.130958 | 0.062434 |
| Economy | 0.365875 | 0.206815 | 0.167062 | 0.075644 | 1.000000 | 0.371446 | 0.339248 | 0.406349 | 0.178442 |
| Option1 | 0.536707 | 0.404404 | 0.444059 | 0.249674 | 0.371446 | 1.000000 | 0.203971 | 0.093839 | 0.259409 |
| Option2 | 0.196322 | 0.061577 | 0.111853 | 0.085804 | 0.339248 | 0.203971 | 1.000000 | 0.021272 | 0.076618 |
| English | 0.114005 | -0.119735 | 0.186937 | 0.130958 | 0.406349 | 0.093839 | 0.021272 | 1.000000 | 0.137287 |
| Sport | 0.234720 | 0.158349 | 0.269368 | 0.062434 | 0.178442 | 0.259409 | 0.076618 | 0.137287 | 1.000000 |

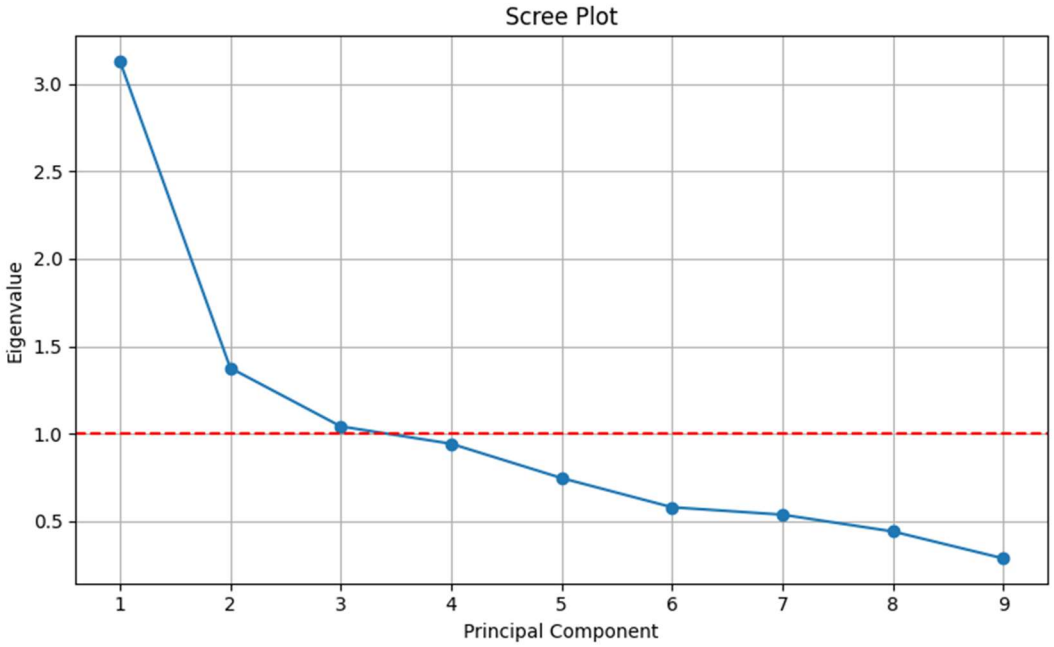
**(3) 고유값 누적 기여도 구하기**

| | PC | Eigenvalue | Proportion | Cumulative |
|---|---|---|---|---|
| 0 | 1 | 3.131468 | 0.344595 | 0.344595 |
| 1 | 2 | 1.376216 | 0.151443 | 0.496038 |
| 2 | 3 | 1.042349 | 0.114703 | 0.610741 |
| 3 | 4 | 0.943122 | 0.103784 | 0.714525 |
| 4 | 5 | 0.746935 | 0.082195 | 0.796719 |
| 5 | 6 | 0.580249 | 0.063852 | 0.860572 |
| 6 | 7 | 0.537712 | 0.059171 | 0.919743 |
| 7 | 8 | 0.441787 | 0.048616 | 0.968358 |
| 8 | 9 | 0.287540 | 0.031642 | 1.000000 |

**(4) 고유값 > 1 인 주성분 개수**

Kaiser 기준에 따라 의미 있는 주성분은 3개

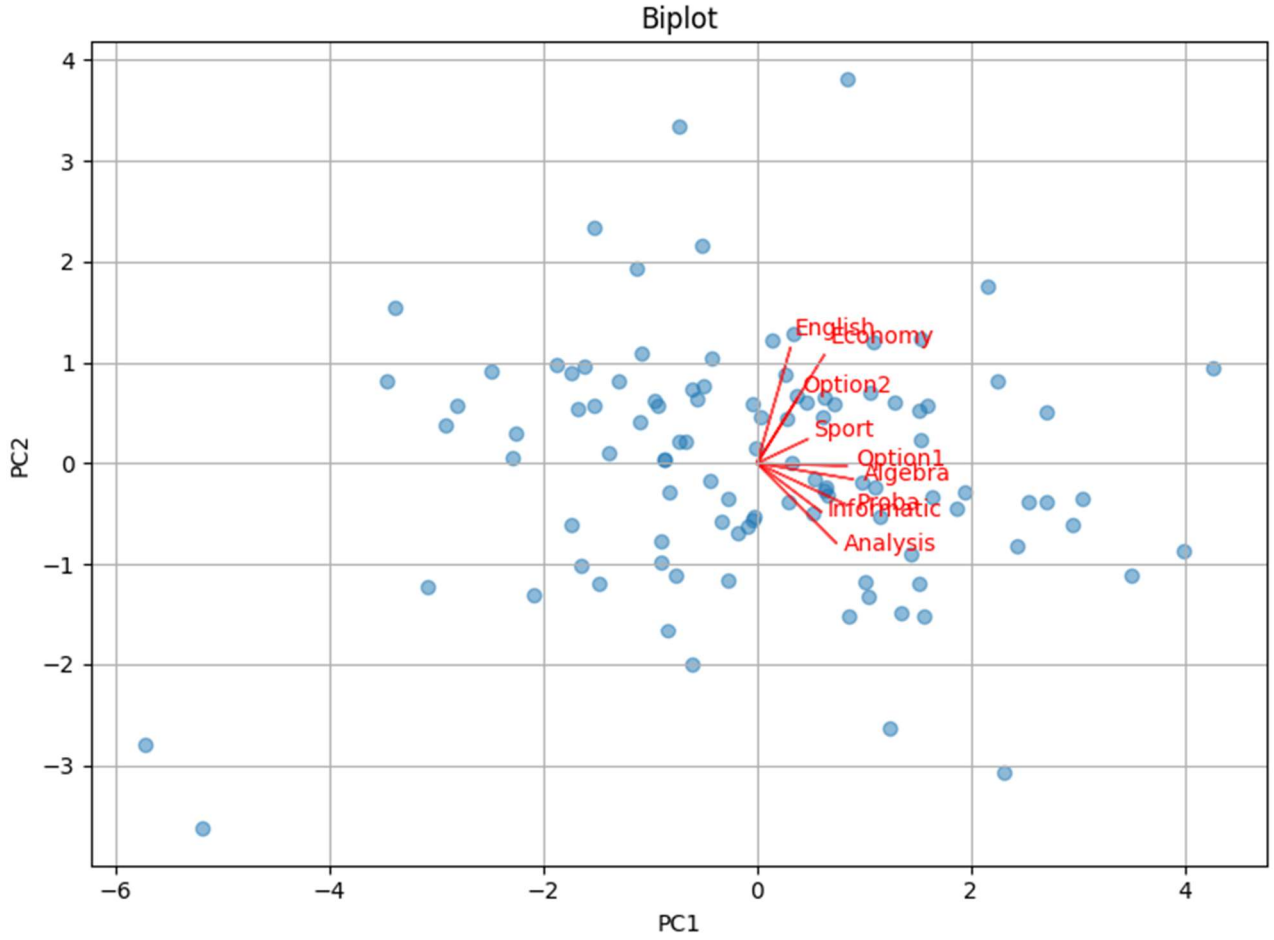
**(5) Scree plot 구하기**




**(6) 주성분 score 구하기**

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | -1.743471 | -0.603742 | -0.046032 | 0.445771 | 1.545442 | 0.041947 | 0.437879 | 0.372297 | -0.373735 |
| 1 | 0.462161 | 0.608221 | -1.045062 | -0.112380 | 0.356185 | 0.403958 | -0.049939 | -0.964975 | 0.038758 |
| 2 | 0.539678 | -0.160066 | 0.680318 | -0.082138 | 0.798993 | 0.337217 | -0.815482 | -1.129515 | -0.041391 |
| 3 | 2.541982 | -0.380420 | -0.533839 | 0.115612 | 0.549942 | 0.220464 | -0.035560 | 1.103509 | 0.954998 |
| 4 | 0.298684 | -0.376229 | -0.856849 | -0.493191 | 1.530546 | -0.171968 | 0.197755 | 0.861120 | -0.736983 |
| .. | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 99 | -0.265773 | -1.165671 | 1.243181 | -1.632266 | -0.661370 | -0.451080 | -0.540514 | 1.156825 | -0.799157 |
| 100 | -0.089057 | -0.630703 | 1.489896 | -1.740500 | -0.368198 | 0.386035 | -0.278480 | -0.043283 | -0.984427 |
| 101 | 1.341478 | -1.490834 | -0.208002 | 0.791056 | 0.655333 | 0.127995 | -0.280197 | -0.903776 | -0.055939 |
| 102 | 0.332655 | 1.276930 | 1.434766 | 0.582130 | -0.124344 | 0.009222 | -0.512870 | 0.245433 | 0.544243 |
| 103 | 1.049529 | 0.709244 | 0.353532 | 0.290474 | 0.440487 | 0.175416 | 0.904954 | -0.629660 | 0.132939 |

**(7) Biplot으로 시각화**




PC1축: 수학/분석계열 중심이 성취도 축

PC2축: 언어/사회계열 중심의 성취도 축

벡터 간 각도

Economy ↔ English: 거의 같은 방향

Algebra ↔ Analysis: 유사한 방향

Analysis ↔ English: 서로 반대 방향

Sport: 비교적 중심에 위치하여 대부분 과목과 관련성 낮음

학생 분포: 서로 다른 경향으로 분포됨(수학형, 언어형 등)

### 3. R와 Python결과 비교 분석

Biplot에서 보았을 때 벡터 방향이 반대가 되어서 해석의 기준이 바뀌었으나, 수학적으로는 동일한 결과를 보이고 있습니다.

핵심정보는 일치하는 모습을 보이며 고유값의 차이는 소수점 계산 방식에서 미세한 차이가 나타납니다.

기여율과 누적 기여율은 거의 동일한 결과를 보이고 있으며, 해석 결과 동일한 결론에 도달하였습니다.

# 문제 4번

### (1) 유의한 인자의 수와 그 인자들이 확보한 정보의 양

판단 기준: fa.parallel()

고유값 기반 병렬분석 결과: 유의한 인자 수 = 2

이 두 인자가 설명하는 전체 분산 비율:

ML2 (인자1): 35%

ML1 (인자2): 31%

누적 설명력: 66% -> 학생들의 과목 선호도 분산의 약 66%를 두 인자로 설명 가능

### (2) 인자부하행렬을 구하고 varimax와 promax 방법을 이용하여 인자회전을 실시하고 결과 비교

Varimax 회전 결과

| 과목 | ML2 | ML1 | 해석 |
|---|---|---|---|
| BIO | 0.85 | 0.13 | 인자1 |

| | | | |
|---|---|---|---|
| GEO | 0.78 | 0.13 | 인자1 |
| CHEM | 0.86 | 0.06 | 인자1 |
| ALG | 0.03 | 0.79 | 인자2 |
| CALC | 0.10 | 0.97 | 인자2 |
| STAT | 0.17 | 0.51 | 혼합 (약함) |

Promax 회전 결과

| 과목 | ML2 | ML1 | 해석 |
|---|---|---|---|
| BIO | 0.86 | 0.02 | 인자1 |
| GEO | 0.78 | 0.03 | 인자1 |
| CHEM | 0.88 | -0.06 | 인자1 |
| ALG | -0.09 | 0.81 | 인자2 |
| CALC | -0.05 | 0.99 | 인자2 |
| STAT | 0.10 | 0.50 | 약하게 인자2 |

각 과목이 두 인자 중 하나에 명확히 부하되어 복잡도가 낮고 해석이 용이합니다.

Varimax는 직교 회전이며 인자 간에 독립성을 가정하였고, Promax는 사각 회전으로써 인자 간에 상관을 허용하였습니다.

**(3) 인자들의 적절한 이름**

| 인자 | 주요 부하 과목 | 제안 이름 |
|---|---|---|
| ML2 | BIO, GEO, CHEM | 자연과학 탐구형 |
| ML1 | ALG, CALC, STAT | 수리/논리 사고형 |

**(4) 인자분석 결과 종합 정리**

| 항목 | 결과 |
|---|---|
| 유의 인자 수 | 2개 |
| 누적 설명력 | 66% |
| 회전 방식 비교 | Varimax는 명확하고 단순한 구조를 제공하며 Promax는 인자 간의 상관을 반영 |
| 적절한 인자 이름 | ML2: 자연과학 탐구형<br>ML1: 수리/논리 사고형 |
| 활용 방안 | 선호 과목 유형 분류<br>전공 적성 진단 |