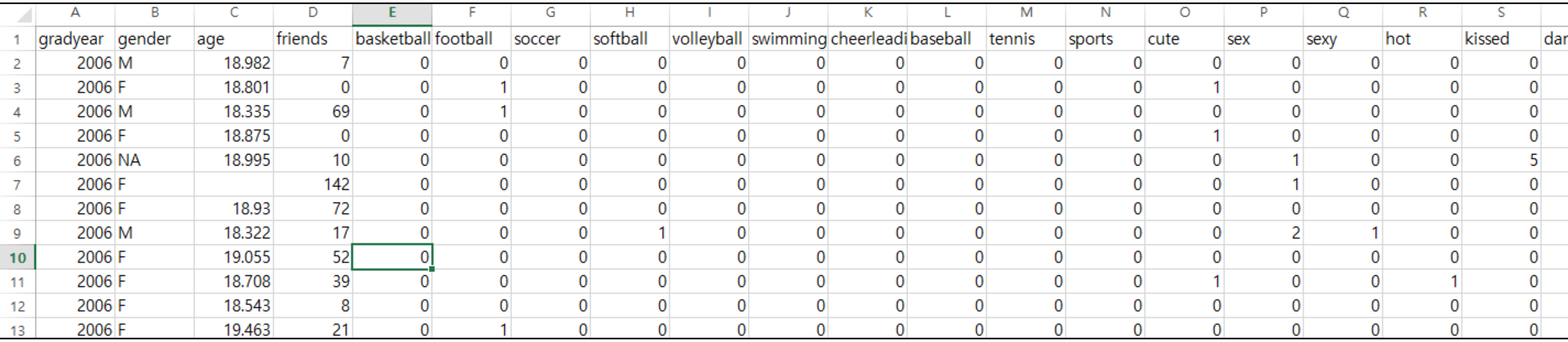
실습 3강
k평균 군집분석

실습 데이터

- 실습데이터와 실습과정은 Brett Lantz, “Machine Learning with R”에서 발췌
- 미국 10대 학생들의 소셜 네트워크 프로파일 데이터를 이용한 군집분석
- 30,000명 (여학생 비율 74%)
- 40개 변수 활용



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	gradyear	gender	age	friends	basketball	football	soccer	softball	volleyball	swimming	cheerleadi	baseball	tennis	sports	cute	sex	sexy	hot	kissed	dar
2	2006	M	18.982	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	2006	F	18.801	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	
4	2006	M	18.335	69	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	2006	F	18.875	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
6	2006	NA	18.995	10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	5	
7	2006	F		142	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
8	2006	F	18.93	72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	2006	M	18.322	17	0	0	0	1	0	0	0	0	0	0	0	2	1	0	0	
10	2006	F	19.055	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	2006	F	18.708	39	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	
12	2006	F	18.543	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	2006	F	19.463	21	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	

데이터 구조 확인

```
# 데이터 읽기
teens <- read.csv("snsdata.csv")

# 구조 확인
str(teens)
```

```
'data.frame':30000 obs. of  40 variables:
 $ gradyear    : int  2006 2006 2006 2006 2006 2006...
 $ gender      : Factor w/ 2 levels "F","M": 2 1 2 1 NA 1 ...
 $ age         : num  19 18.8 18.3 18.9 19 ...
 $ friends     : int  7 0 69 0 10 142 72 17 52 39 ...
 $ basketball  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ football    : int  0 1 1 0 0 0 0 0 0 0 ...
 $ soccer      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ softball    : int  0 0 0 0 0 0 0 1 0 0 ...
 $ volleyball  : int  0 0 0 0 0 0 0 0 0 0 ...
```

결측치

데이터 구조 확인

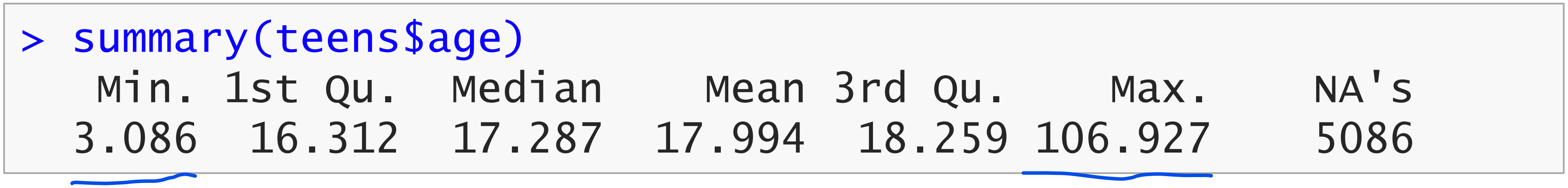
❖ 데이터 확인

```
# female 변수의 결측 데이터 확인
table(teens$gender)

# 결측값을 포함할 수 있도록 ifany 작성
table(teens$gender, useNA = "ifany")
```

```
> table(teens$gender)
    F     M
22054  5222


> table(teens$gender, useNA = "ifany")

    F     M  <NA>
22054  5222  2724
```

데이터 구조 확인

```
# age 변수의 결측 데이터 확인
summary(teens$age)
```



```
> summary(teens$age)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
  3.086   16.312   17.287   17.994   18.259  106.927     5086 
```

이상치
(Outlier)

데이터 구조 확인

이상치 (Outliers) 제거

- 연령이 13세 이상 20세 미만이면 teen$age에 값을 대입하고, 아닐 경우에는 NA로 대체

```
# age 이상치(outliers) 제거
teens$age <- ifelse(teens$age >= 13 & teens$age < 20,
                    teens$age, NA)

# age 변수의 데이터 확인
summary(teens$age)
```

```
> summary(teens$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  13.03   16.30   17.27   17.25   18.22   20.00    5523 
```

데이터 구조 확인

❖ 결측치 더미 코딩

- 남녀가 있을 때, 여자와 결측치가 있을 경우 → 최종 결측치를 남자로 추정함

```
# "unknown"인 성별값에 재부여
teens$female <- ifelse(teens$gender == "F" &
                         !is.na(teens$gender), 1, 0)

teens$no_gender <- ifelse(is.na(teens$gender), 1, 0)
```

데이터 구조 확인

❖ 데이터 확인

```
# 재지정한 작업에 대한 확인
table(teens$gender, useNA = "ifany")
table(teens$female, useNA = "ifany")
table(teens$no_gender, useNA = "ifany")
```

```
> table(teens$gender, useNA = "ifany")
    F     M  <NA> 
22054  5222  2724 

> table(teens$female, useNA = "ifany")
    0     1 
 7946 22054 

> table(teens$no_gender, useNA = "ifany")
    0     1 
27276  2724 
```

데이터 구조 확인

❖ 결측치 대체

- 졸업세대의 대표 연령을 식별할 수 있도록 나이를 추정

```
# 집단(cohort)별 나이 평균
mean(teens$age) # doesn't work
mean(teens$age, na.rm = TRUE)
```

```
> mean(teens$age) # doesn't work
[1] NA


> mean(teens$age, na.rm = TRUE)
[1] 17.25243
```

데이터 구조 확인

```
# 집단별 나이
aggregate(data = teens, age ~ gradyear, mean, na.rm = TRUE)
```

```
> aggregate(data = teens, age ~ gradyear, mean, na.rm = TRUE)

  gradyear      age
1     2006 18.65586
2     2007 17.70617
3     2008 16.76770
4     2009 15.8195
```

```
# 각 개인에 대한 예측된 나이 계산
ave_age <- ave(teens$age, teens$gradyear,
                  FUN = function(x) mean(x, na.rm = TRUE))

teens$age <- ifelse(is.na(teens$age), ave_age, teens$age)
```

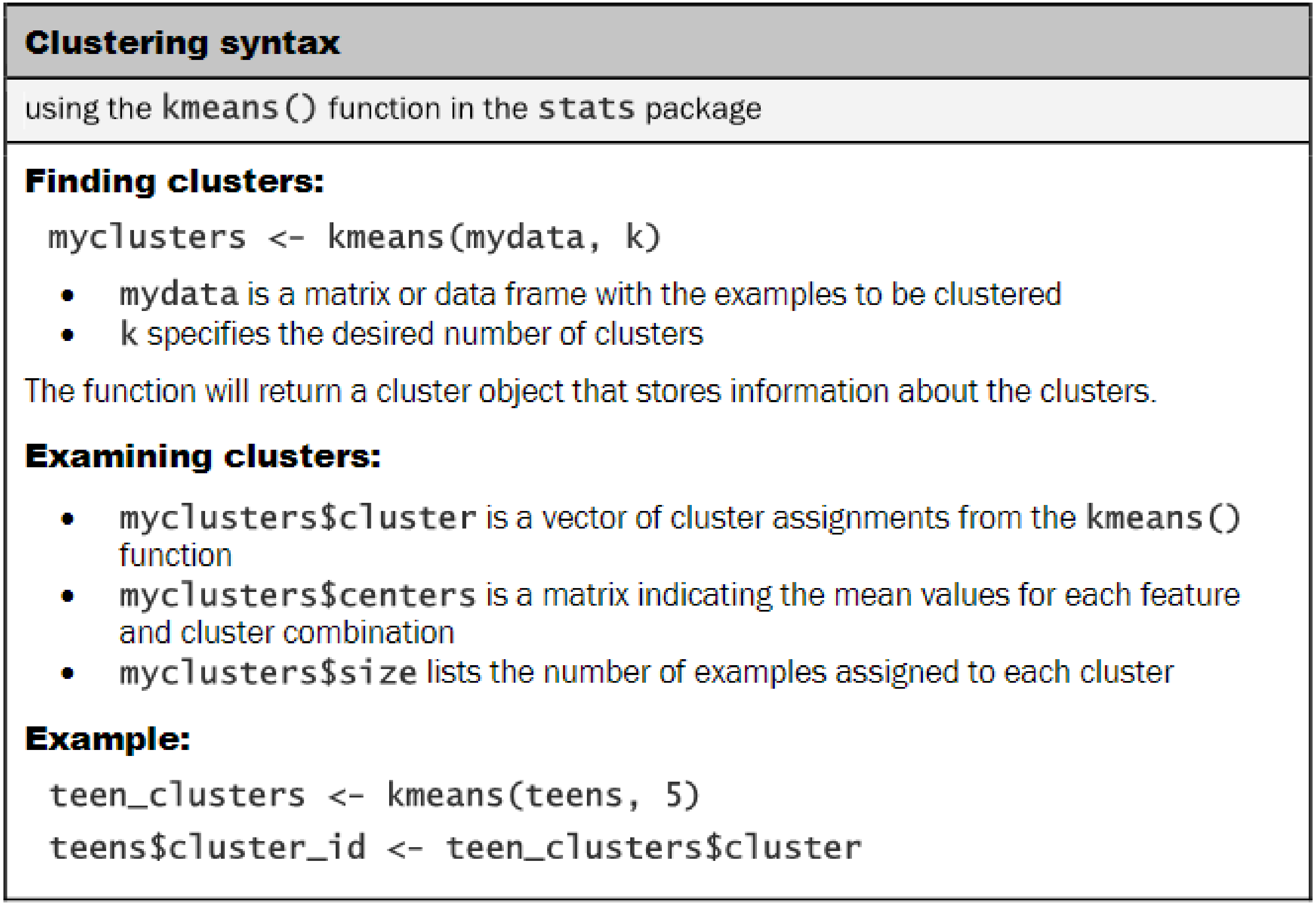
데이터 구조 확인

결측치 요약 확인

```
# 제거한 결측치에 대한 요약 결과 확인
summary(teens$age)
```

```
➢ summary(teens$age)

   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
  13.03    16.28    17.24    17.24    18.21    20.00 
```

주요 함수 문법



Clustering syntax

using the kmeans() function in the stats package

Finding clusters:

```
myclusters <- kmeans(mydata, k)
```

- mydata is a matrix or data frame with the examples to be clustered
- k specifies the desired number of clusters

The function will return a cluster object that stores information about the clusters.

Examining clusters:

- myclusters$cluster is a vector of cluster assignments from the kmeans() function
- myclusters$centers is a matrix indicating the mean values for each feature and cluster combination
- myclusters$size lists the number of examples assigned to each cluster

Example:

```
teen_clusters <- kmeans(teens, 5)
teens$cluster_id <- teen_clusters$cluster
```

kmeans 문법 구조

❖ 모델 만들기

```
myclusters <- kmeans(mydata, k)
```

↓

```
teen_clusters <- kmeans(interests_z, 5)
```

- mydata : 군집화 될 예시가 있는 행렬 또는 데이터 프레임
- k : 희망 클러스터의 개수

모형 구축

클러스터 분석

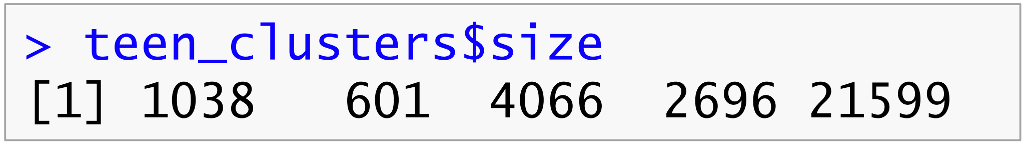
- 다양한 관심사의 횟수를 표현하는 36개의 특징만을 고려하여 클러스터 분석 시작

```
set.seed(2345)

interests <- teens[5:40]

interests_z <- as.data.frame(lapply(interests, scale))

teen_clusters <- kmeans(interests_z, 5)
```

모형 구축

❖ 모델의 성능을 평가하기 위해 kmeans 함수의 속성 이용

```
# 군집의 크기 확인
teen_clusters$size
```



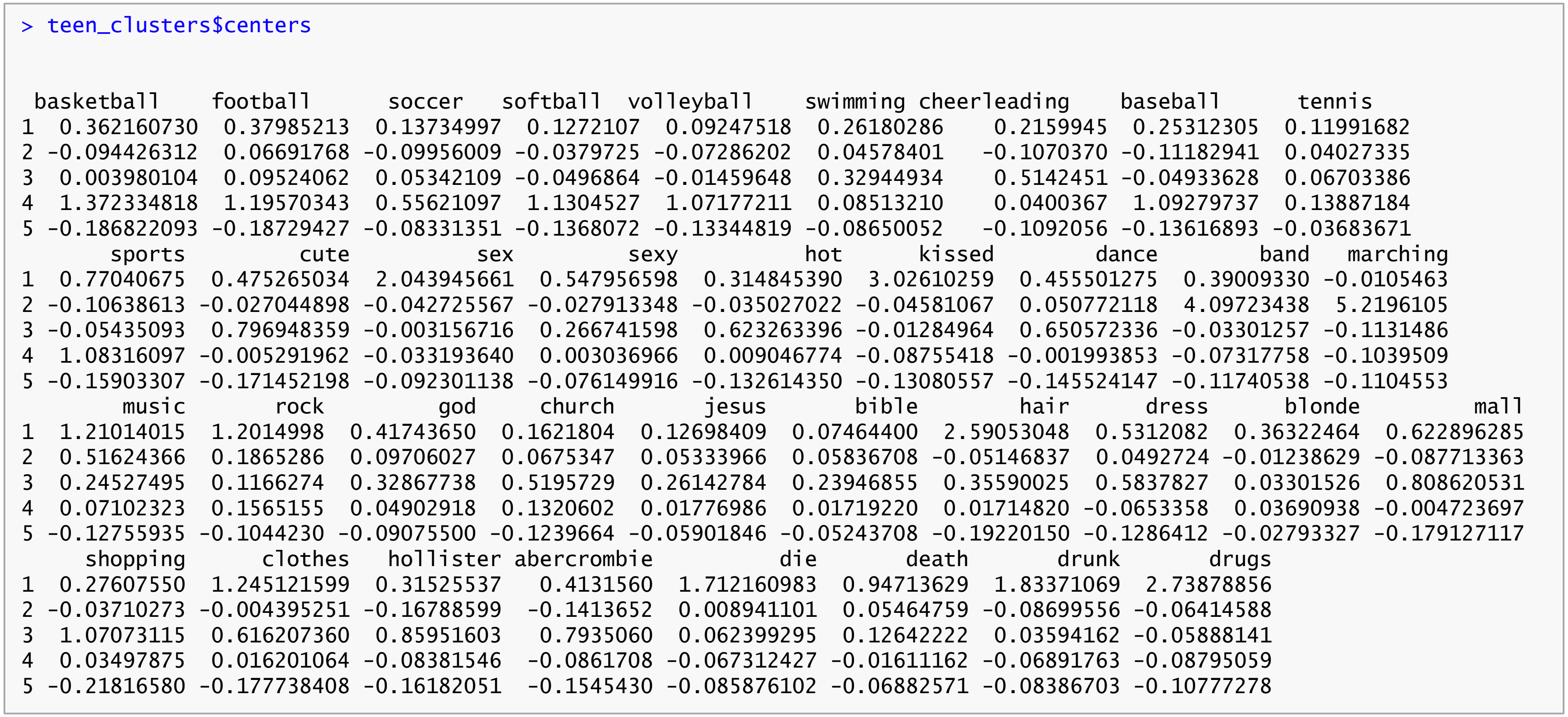
```
> teen_clusters$size
[1]  1038   601  4066  2696 21599
```

※ kmeans 의 클러스터 속성

- $cluster : kmeans() 함수에서 얻은 클러스터 할당 벡터
- $centers : 각 특징과 클러스터 조합별로 평균값을 나타내는 행렬
- $size : 각 클러스터에 할당된 예시 개수

분석 결과 확인

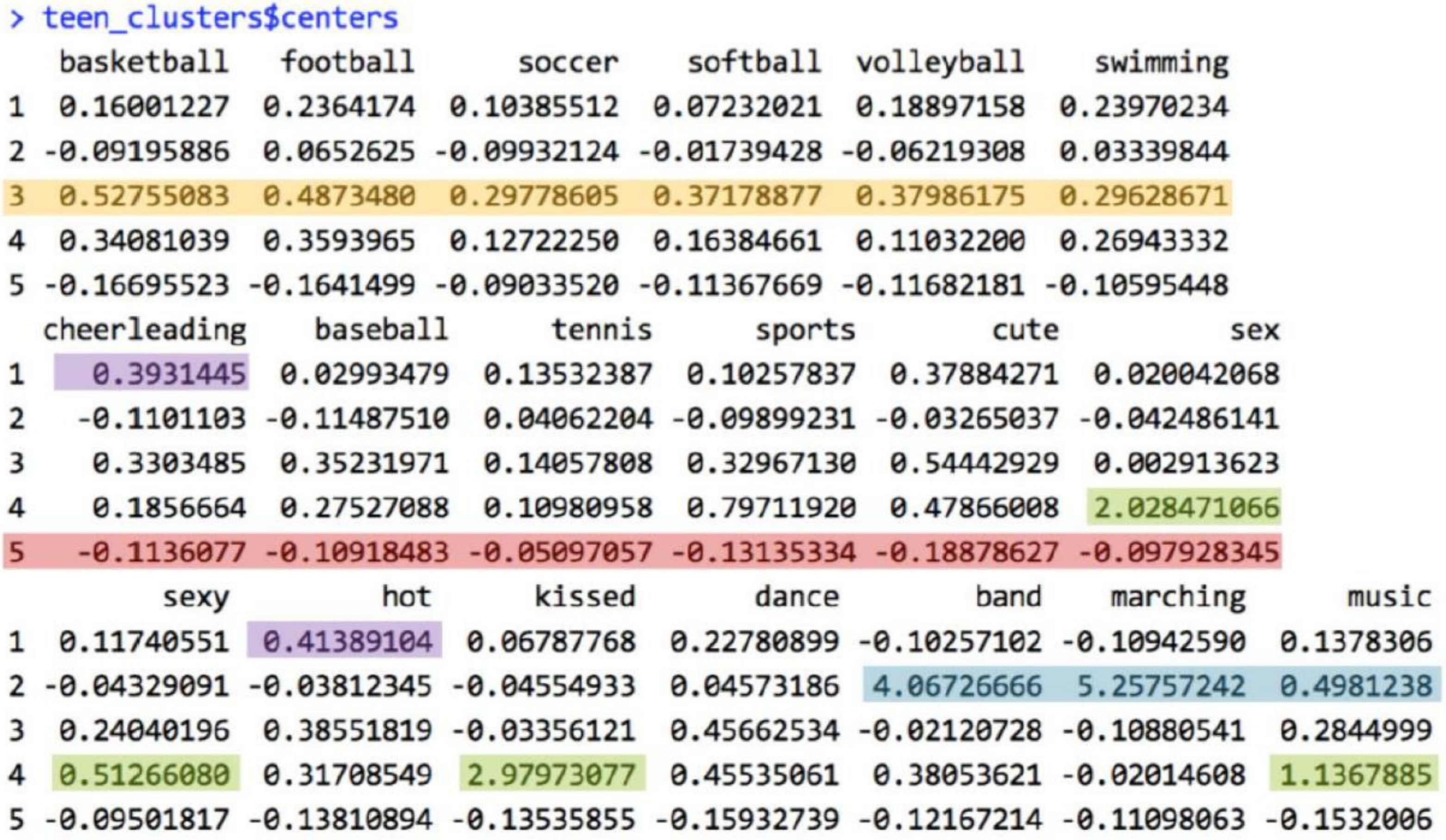
```
# 군집의 중앙점(centers) 확인
teen_clusters$centers
```



```
> teen_clusters$centers


   basketball    football      soccer    softball  volleyball    swimming cheerleading    baseball      tennis
1  0.362160730  0.37985213  0.13734997  0.1272107  0.09247518  0.26180286    0.2159945  0.25312305  0.11991682
2 -0.094426312  0.06691768 -0.09956009 -0.0379725 -0.07286202  0.04578401   -0.1070370 -0.11182941  0.04027335
3  0.003980104  0.09524062  0.05342109 -0.0496864 -0.01459648  0.32944934    0.5142451 -0.04933628  0.06703386
4  1.372334818  1.19570343  0.55621097  1.1304527  1.07177211  0.08513210    0.0400367  1.09279737  0.13887184
5 -0.186822093 -0.18729427 -0.08331351 -0.1368072 -0.13344819 -0.08650052   -0.1092056 -0.13616893 -0.03683671
        sports         cute          sex         sexy          hot      kissed        dance        band    marching
1  0.77040675  0.475265034  2.043945661  0.547956598  0.314845390  3.02610259  0.455501275  0.39009330 -0.0105463
2 -0.10638613 -0.027044898 -0.042725567 -0.027913348 -0.035027022 -0.04581067  0.050772118  4.09723438  5.2196105
3 -0.05435093  0.796948359 -0.003156716  0.266741598  0.623263396 -0.01284964  0.650572336 -0.03301257 -0.1131486
4  1.08316097 -0.005291962 -0.033193640  0.003036966  0.009046774 -0.08755418 -0.001993853 -0.07317758 -0.1039509
5 -0.15903307 -0.171452198 -0.092301138 -0.076149916 -0.132614350 -0.13080557 -0.145524147 -0.11740538 -0.1104553
        music        rock         god      church       jesus       bible        hair       dress      blonde         mall
1  1.21014015  1.2014998  0.41743650  0.1621804  0.12698409  0.07464400  2.59053048  0.5312082  0.36322464  0.622896285
2  0.51624366  0.1865286  0.09706027  0.0675347  0.05333966  0.05836708 -0.05146837  0.0492724 -0.01238629 -0.087713363
3  0.24527495  0.1166274  0.32867738  0.5195729  0.26142784  0.23946855  0.35590025  0.5837827  0.03301526  0.808620531
4  0.07102323  0.1565155  0.04902918  0.1320602  0.01776986  0.01719220  0.01714820 -0.0653358  0.03690938 -0.004723697
5 -0.12755935 -0.1044230 -0.09075500 -0.1239664 -0.05901846 -0.05243708 -0.19220150 -0.1286412 -0.02793327 -0.179127117
     shopping      clothes   hollister abercrombie          die       death       drunk       drugs
1  0.27607550  1.245121599  0.31525537   0.4131560  1.712160983  0.94713629  1.83371069  2.73878856
2 -0.03710273 -0.004395251 -0.16788599  -0.1413652  0.008941101  0.05464759 -0.08699556 -0.06414588
3  1.07073115  0.616207360  0.85951603   0.7935060  0.062399295  0.12642222  0.03594162 -0.05888141
4  0.03497875  0.016201064 -0.08381546  -0.0861708 -0.067312427 -0.01611162 -0.06891763 -0.08795059
5 -0.21816580 -0.177738408 -0.16182051  -0.1545430 -0.085876102 -0.06882571 -0.08386703 -0.10777278
```

분석 결과 확인

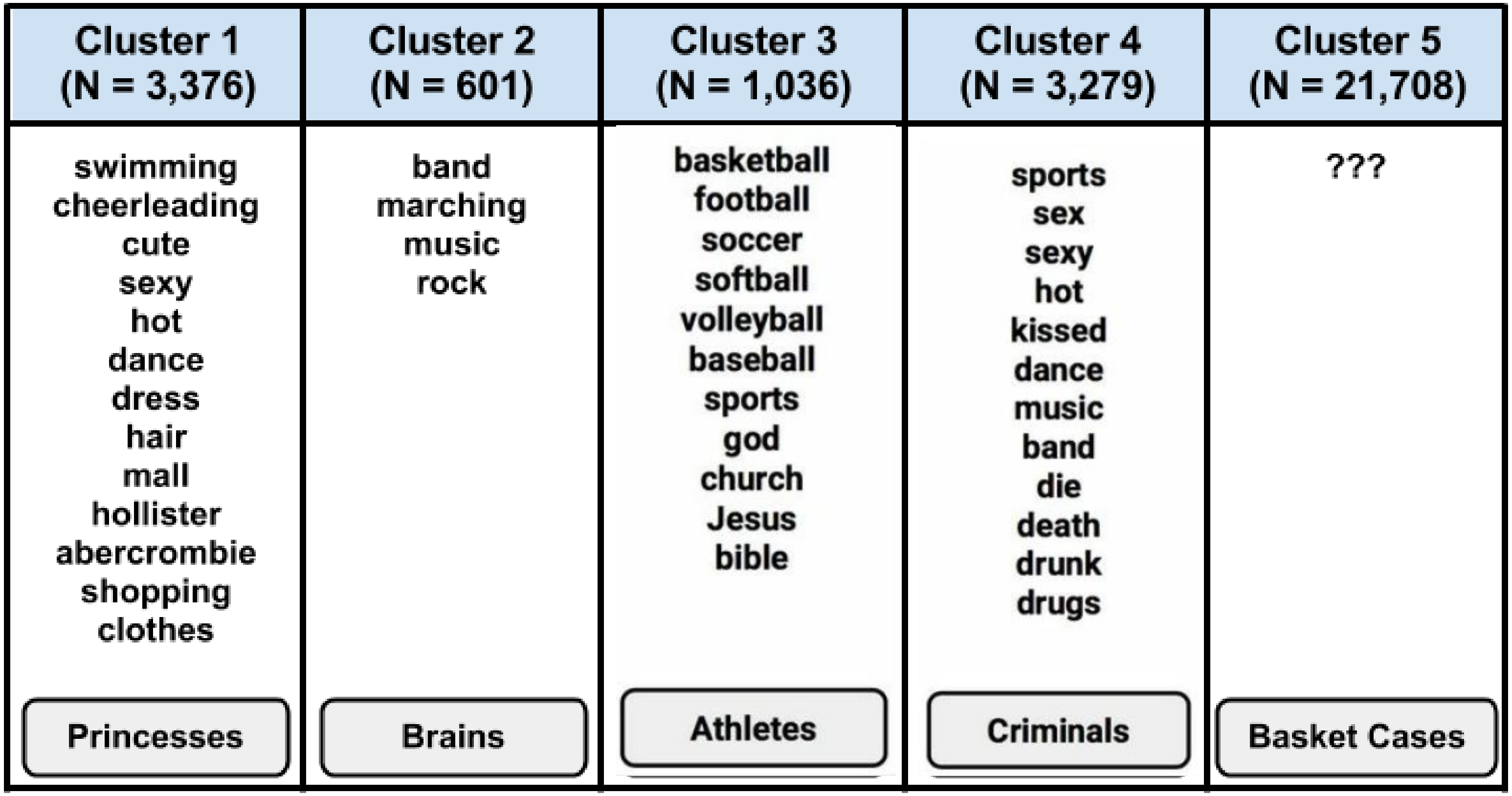
```
# 군집의 중앙점(centers) 확인
teen_clusters$centers
```



```
> teen_clusters$centers
   basketball   football      soccer    softball  volleyball    swimming
1  0.16001227  0.2364174  0.10385512  0.07232021  0.18897158  0.23970234
2 -0.09195886  0.0652625 -0.09932124 -0.01739428 -0.06219308  0.03339844
3  0.52755083  0.4873480  0.29778605  0.37178877  0.37986175  0.29628671
4  0.34081039  0.3593965  0.12722250  0.16384661  0.11032200  0.26943332
5 -0.16695523 -0.1641499 -0.09033520 -0.11367669 -0.11682181 -0.10595448
  cheerleading    baseball      tennis      sports        cute          sex
1    0.3931445  0.02993479  0.13532387  0.10257837  0.37884271  0.020042068
2   -0.1101103 -0.11487510  0.04062204 -0.09899231 -0.03265037 -0.042486141
3    0.3303485  0.35231971  0.14057808  0.32967130  0.54442929  0.002913623
4    0.1856664  0.27527088  0.10980958  0.79711920  0.47866008  2.028471066
5   -0.1136077 -0.10918483 -0.05097057 -0.13135334 -0.18878627 -0.097928345
         sexy         hot      kissed       dance        band    marching      music
1  0.11740551  0.41389104  0.06787768  0.22780899 -0.10257102 -0.10942590  0.1378306
2 -0.04329091 -0.03812345 -0.04554933  0.04573186  4.06726666  5.25757242  0.4981238
3  0.24040196  0.38551819 -0.03356121  0.45662534 -0.02120728 -0.10880541  0.2844999
4  0.51266080  0.31708549  2.97973077  0.45535061  0.38053621 -0.02014608  1.1367885
5 -0.09501817 -0.13810894 -0.13535855 -0.15932739 -0.12167214 -0.11098063 -0.1532006
```

분석 결과 확인

❖ 결과 해석



Cluster 1 (N = 3,376)	Cluster 2 (N = 601)	Cluster 3 (N = 1,036)	Cluster 4 (N = 3,279)	Cluster 5 (N = 21,708)
swimming cheerleading cute sexy hot dance dress hair mall hollister abercrombie shopping clothes	band marching music rock	basketball football soccer softball volleyball baseball sports god church Jesus bible	sports sex sexy hot kissed dance music band die death drunk drugs	???
Princesses	Brains	Athletes	Criminals	Basket Cases

분석 결과 확인

❖ 결과 해석

```
# 본래 데이터 프레임에 군집ID(cluster ID) 적용
teens$cluster <- teen_clusters$cluster

# 처음 5개 데이터 확인
teens[1:5, c("cluster", "gender", "age", "friends")]
```

```
> teens[1:5, c("cluster", "gender", "age", "friends")]

  cluster gender    age friends
1       5      M 18.982       7
2       3      F 18.801       0
3       5      M 18.335      69
4       5      F 18.875       0
5       1   <NA> 18.995      10
```

분석 결과 확인

```
# 군집 별 평균 나이
aggregate(data = teens, age ~ cluster, mean)
```

```
> aggregate(data = teens, age ~ cluster, mean)

  cluster      age
1       1 17.09319
2       2 17.38488
3       3 17.03773
4       4 17.03759
5       5 17.30265
```

```
# 군집 별 여성 비율
aggregate(data = teens, female ~ cluster, mean)
```

```
> aggregate(data = teens, female ~ cluster, mean)

   cluster    female
1       1  0.8025048
2       2  0.7237937
3       3  0.8866208
4       4  0.6984421
5       5  0.7082735
```

분석 결과 확인

```
# 군집 별 친구 수의 평균
aggregate(data = teens, friends ~ cluster, mean)
```

```
> aggregate(data = teens, friends ~ cluster, mean)

 cluster    friends
1       1  30.66570
2       2  32.79368
3       3  38.54575
4       4  35.91728
5       5  27.79221
```