R Exercise

Seminar 9

Instructor: Prof. Lee, Gun-woong
Nanyang Business School

Cluster Analysis

Procedures in a Cluster Analysis

1. **Identify Business Problem(s)**
2. **Understand Data**
3. **Prepare Data**
4. **Build a Clustering model**
5. **Train a Model on the Data**
6. **Validate Clusters**
7. **Evaluate the Business Problem(s)**

Step1: Identify Business Problem(s)

- **Background**
  - Interacting with friends on a social networking service, such as Facebook and Instagram has become a rite of passage for teenagers around the world.
  - The many millions of teenage consumers using such sites have attracted the attention of marketers struggling to find an edge in an increasingly competitive market.
- **Main Problem:** Find teen market segments
  - Identify segments of teenagers who share similar tastes, so that companies can avoid targeting advertisements to teens with no interest in the product being sold.
  - Discover the natural segments in this population

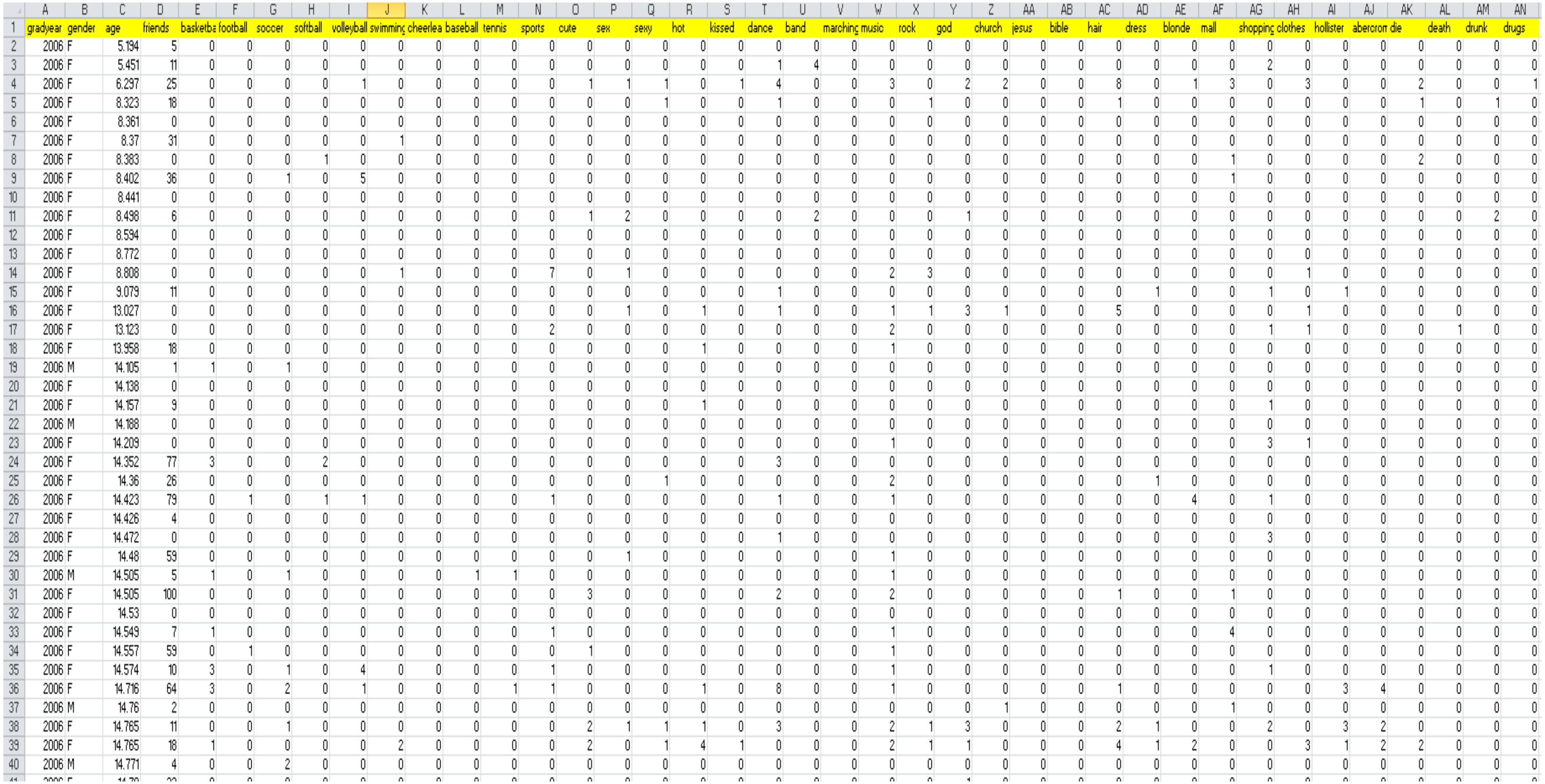
Step2: Understand Data

- **Data Description**
  - Describe the Characteristics of Data

“We used a dataset representing a random sample of 27,276 U.S. high school students who had profiles on a well-known SNS in 2006. The data was sampled across four high-school graduation years (2006 – 2009) representing the senior, junior, sophomore, and freshman classes at the time of data collection. The data includes the full text of the SNS profiles, and each teen’s gender, age, and number of SNS friends…”

“A text-mining tool was used to divide the SNS page content into words. From the top 500 words appearing across all the pages, 36 words were chosen to represent five categories of interests: extracurricular activities, fashion, religion, romance, and antisocial behavior. The 36 words include terms such as *football, sexy, kissed, bible, shopping, death*, and *drugs*. The final dataset indicates, for each person, how many time each word appeared in the person’s SNS profile…”

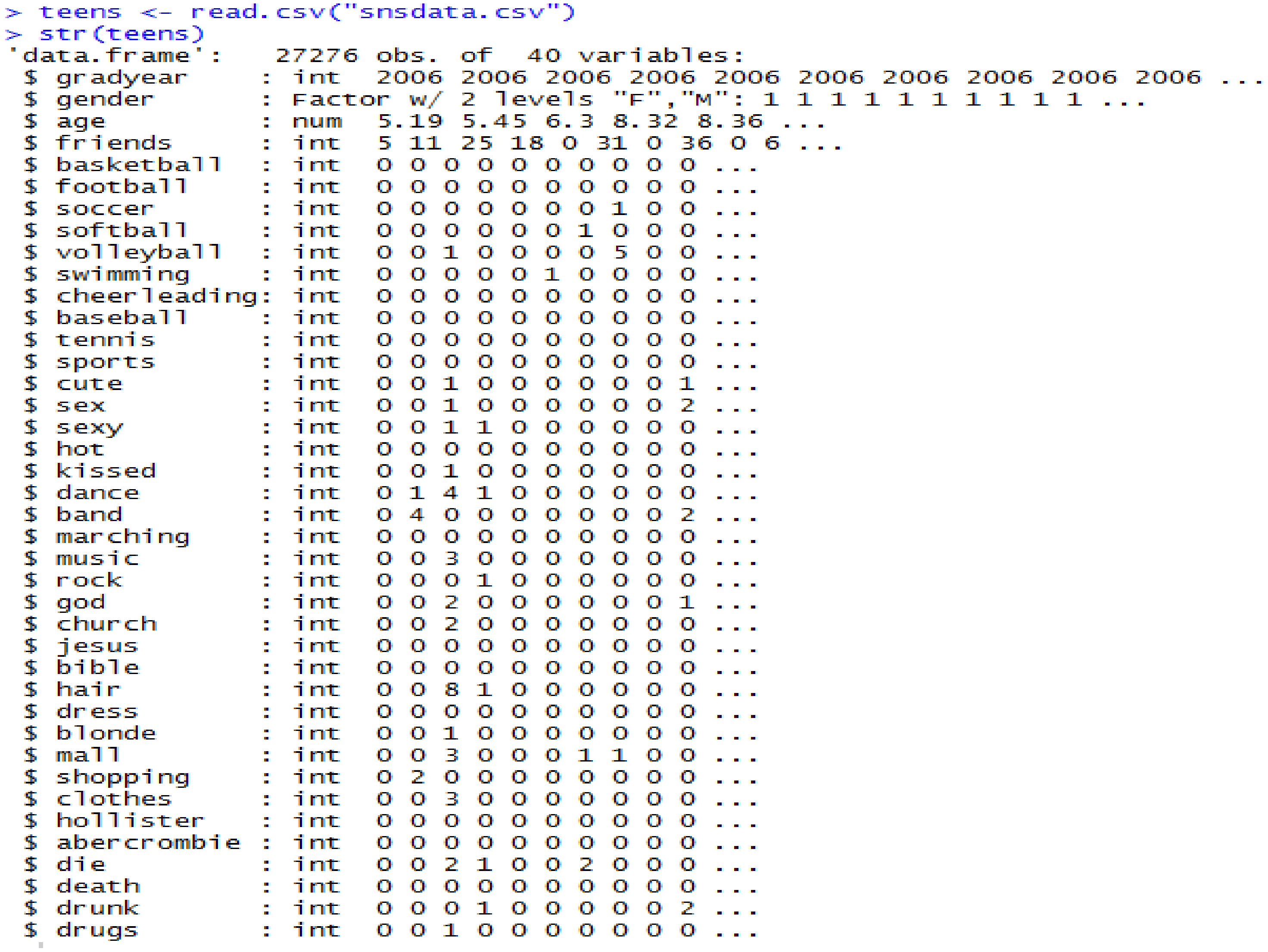
Step2: Understand Data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN
1	gradyear	gender	age	friends	basketba	football	soccer	softball	volleyball	swimming	cheerlea	baseball	tennis	sports	cute	sex	sexy	hot	kissed	dance	band	marching	music	rock	god	church	jesus	bible	hair	dress	blonde	mall	shopping	clothes	hollister	abercrom	die	death	drunk	drugs
2	2006	F	5.194	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2006	F	5.451	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
4	2006	F	6.297	25	0	0	0	0	1	0	0	0	0	0	1	1	1	0	1	4	0	0	3	0	2	2	0	0	8	0	1	3	0	3	0	0	2	0	0	1
5	2006	F	8.323	18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0
6	2006	F	8.361	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	2006	F	8.37	31	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	2006	F	8.383	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0
9	2006	F	8.402	36	0	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
10	2006	F	8.441	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	2006	F	8.498	6	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
12	2006	F	8.594	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	2006	F	8.772	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	2006	F	8.808	0	0	0	0	0	0	1	0	0	0	7	0	1	0	0	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
15	2006	F	9.079	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
16	2006	F	13.027	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1	3	1	0	0	5	0	0	0	0	1	0	0	0	0	0	0
17	2006	F	13.123	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
18	2006	F	13.958	18	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	2006	M	14.105	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	2006	F	14.138	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	2006	F	14.157	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
22	2006	M	14.188	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	2006	F	14.209	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0
24	2006	F	14.352	77	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	2006	F	14.36	26	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
26	2006	F	14.423	79	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	4	0	1	0	0	0	0	0	0	0
27	2006	F	14.426	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	2006	F	14.472	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
29	2006	F	14.48	59	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	2006	M	14.505	5	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	2006	F	14.505	100	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	2	0	0	2	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
32	2006	F	14.53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	2006	F	14.549	7	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
34	2006	F	14.557	59	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	2006	F	14.574	10	3	0	1	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
36	2006	F	14.716	64	3	0	2	0	1	0	0	0	1	1	0	0	0	1	0	8	0	0	1	0	0	0	0	0	1	0	0	0	0	0	3	4	0	0	0	0
37	2006	M	14.76	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
38	2006	F	14.765	11	0	0	1	0	0	0	0	0	0	0	2	1	1	1	0	3	0	0	2	1	3	0	0	0	2	1	0	0	2	0	3	2	0	0	0	0
39	2006	F	14.765	18	1	0	0	0	0	2	0	0	0	0	2	0	1	4	1	0	0	0	2	1	1	0	0	0	4	1	2	0	0	3	1	2	2	0	0	0
40	2006	M	14.771	4	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- 27,276 teenagers with four variables indicating personal characteristics and 36 words indicating interests

Step2: Understand Data

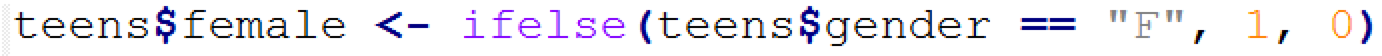


```
> teens <- read.csv("snsdata.csv")
> str(teens)
'data.frame':   27276 obs. of  40 variables:
 $ gradyear    : int  2006 2006 2006 2006 2006 2006 2006 2006 2006 2006 ...
 $ gender      : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ age         : num  5.19 5.45 6.3 8.32 8.36 ...
 $ friends     : int  5 11 25 18 0 31 0 36 0 6 ...
 $ basketball  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ football    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ soccer      : int  0 0 0 0 0 0 0 1 0 0 ...
 $ softball    : int  0 0 0 0 0 0 1 0 0 0 ...
 $ volleyball  : int  0 0 1 0 0 0 0 5 0 0 ...
 $ swimming    : int  0 0 0 0 0 1 0 0 0 0 ...
 $ cheerleading: int  0 0 0 0 0 0 0 0 0 0 ...
 $ baseball    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ tennis      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ sports      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ cute        : int  0 0 1 0 0 0 0 0 0 1 ...
 $ sex         : int  0 0 1 0 0 0 0 0 0 2 ...
 $ sexy        : int  0 0 1 1 0 0 0 0 0 0 ...
 $ hot         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ kissed      : int  0 0 1 0 0 0 0 0 0 0 ...
 $ dance       : int  0 1 4 1 0 0 0 0 0 0 ...
 $ band        : int  0 4 0 0 0 0 0 0 0 2 ...
 $ marching    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ music       : int  0 0 3 0 0 0 0 0 0 0 ...
 $ rock        : int  0 0 0 1 0 0 0 0 0 0 ...
 $ god         : int  0 0 2 0 0 0 0 0 0 1 ...
 $ church      : int  0 0 2 0 0 0 0 0 0 0 ...
 $ jesus       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ bible       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hair        : int  0 0 8 1 0 0 0 0 0 0 ...
 $ dress       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ blonde      : int  0 0 1 0 0 0 0 0 0 0 ...
 $ mall        : int  0 0 3 0 0 0 1 1 0 0 ...
 $ shopping    : int  0 2 0 0 0 0 0 0 0 0 ...
 $ clothes     : int  0 0 3 0 0 0 0 0 0 0 ...
 $ hollister   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ abercrombie : int  0 0 0 0 0 0 0 0 0 0 ...
 $ die         : int  0 0 2 1 0 0 2 0 0 0 ...
 $ death       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ drunk       : int  0 0 0 1 0 0 0 0 0 2 ...
 $ drugs       : int  0 0 1 0 0 0 0 0 0 0 ...
```

Step3: Prepare Data

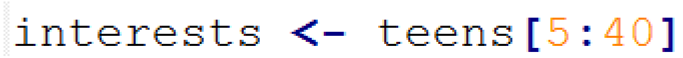
- **Data Cleaning and Pre-Processing**
  - Cleaning: Missing Values, Duplicates, and Outliers
  - Pre-processing: Variable Standardisation
  - DO NOT impute any missing values for this exercise

- **Convert Characters to Numbers**
  - Clustering only takes numerical variables



```
teens$female <- ifelse(teens$gender == "F", 1, 0)
```

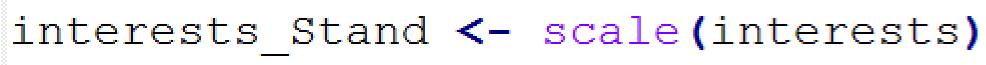
- **Select the variables that will be used for clustering**
  - 36 words indicating interests will be used



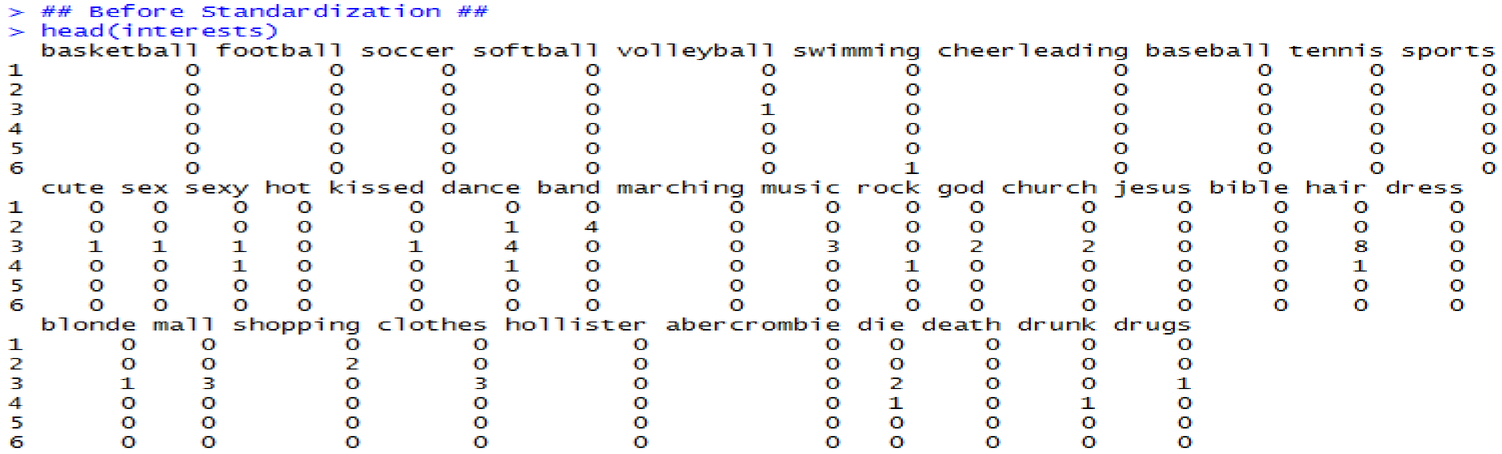
```
interests <- teens[5:40]
```

Step3: Prepare Data

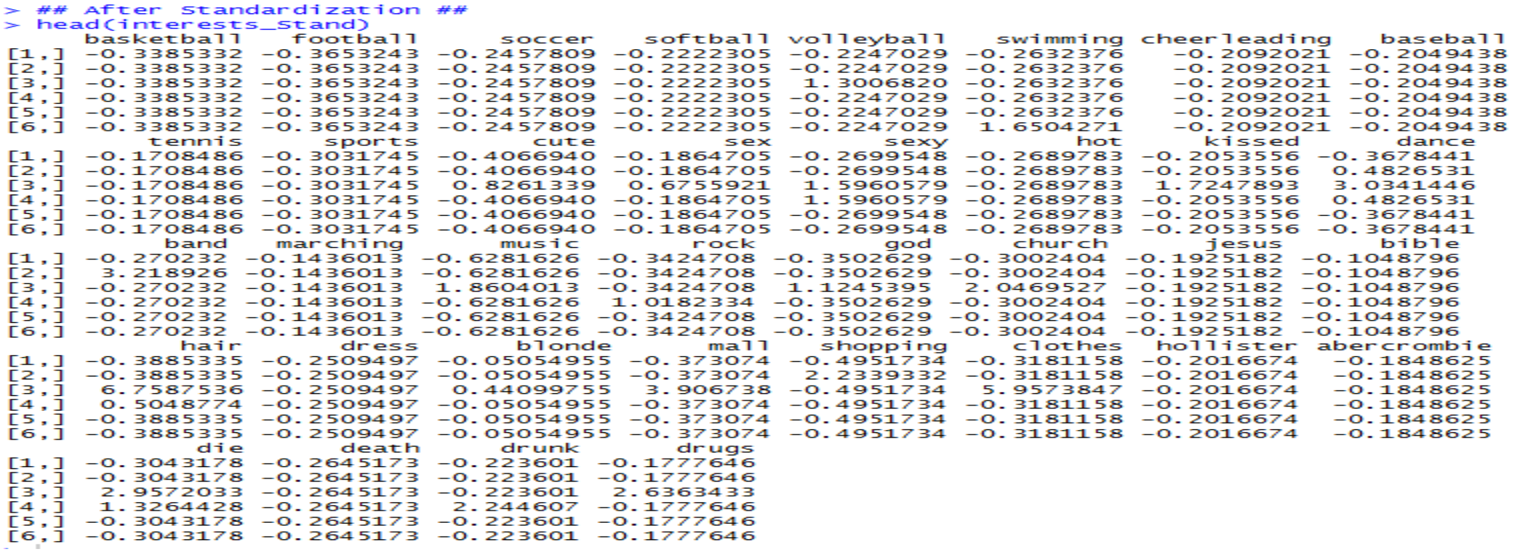
- **Standardize the variables**



```
interests_Stand <- scale(interests)
```



```
> ## Before Standardization ##
> head(interests)
  basketball football soccer softball volleyball swimming cheerleading baseball tennis sports
1          0        0      0        0          0        0            0        0      0      0
2          0        0      0        0          0        0            0        0      0      0
3          0        0      0        0          1        0            0        0      0      0
4          0        0      0        0          0        0            0        0      0      0
5          0        0      0        0          0        0            0        0      0      0
6          0        0      0        0          0        1            0        0      0      0
  cute sex sexy hot kissed dance band marching music rock god church jesus bible hair dress
1    0   0    0   0      0     0    0        0     0    0   0      0     0     0    0     0
2    0   0    0   0      0     1    4        0     0    0   0      0     0     0    0     0
3    1   1    1   0      1     4    0        0     3    0   2      2     0     0    8     0
4    0   0    1   0      0     1    0        0     0    1   0      0     0     0    1     0
5    0   0    0   0      0     0    0        0     0    0   0      0     0     0    0     0
6    0   0    0   0      0     0    0        0     0    0   0      0     0     0    0     0
  blonde mall shopping clothes hollister abercrombie die death drunk drugs
1      0    0        0       0         0           0   0     0     0     0
2      0    0        2       0         0           0   0     0     0     0
3      1    3        0       3         0           0   2     0     0     1
4      0    0        0       0         0           0   1     0     1     0
5      0    0        0       0         0           0   0     0     0     0
6      0    0        0       0         0           0   0     0     0     0
```



```
> ## After Standardization ##
> head(interests_Stand)
     basketball   football     soccer   softball volleyball   swimming cheerleading   baseball
[1,] -0.3385332 -0.3653243 -0.2457809 -0.2222305 -0.2247029 -0.2632376   -0.2092021 -0.2049438
[2,] -0.3385332 -0.3653243 -0.2457809 -0.2222305 -0.2247029 -0.2632376   -0.2092021 -0.2049438
[3,] -0.3385332 -0.3653243 -0.2457809 -0.2222305  1.3006820 -0.2632376   -0.2092021 -0.2049438
[4,] -0.3385332 -0.3653243 -0.2457809 -0.2222305 -0.2247029 -0.2632376   -0.2092021 -0.2049438
[5,] -0.3385332 -0.3653243 -0.2457809 -0.2222305 -0.2247029 -0.2632376   -0.2092021 -0.2049438
[6,] -0.3385332 -0.3653243 -0.2457809 -0.2222305 -0.2247029  1.6504271   -0.2092021 -0.2049438
         tennis     sports       cute        sex       sexy        hot     kissed      dance
[1,] -0.1708486 -0.3031745 -0.4066940 -0.1864705 -0.2699548 -0.2689783 -0.2053556 -0.3678441
[2,] -0.1708486 -0.3031745 -0.4066940 -0.1864705 -0.2699548 -0.2689783 -0.2053556  0.4826531
[3,] -0.1708486 -0.3031745  0.8261339  0.6755921  1.5960579 -0.2689783  1.7247893  3.0341446
[4,] -0.1708486 -0.3031745 -0.4066940 -0.1864705  1.5960579 -0.2689783 -0.2053556  0.4826531
[5,] -0.1708486 -0.3031745 -0.4066940 -0.1864705 -0.2699548 -0.2689783 -0.2053556 -0.3678441
[6,] -0.1708486 -0.3031745 -0.4066940 -0.1864705 -0.2699548 -0.2689783 -0.2053556 -0.3678441
          band   marching      music       rock        god     church      jesus      bible
[1,] -0.270232 -0.1436013 -0.6281626 -0.3424708 -0.3502629 -0.3002404 -0.1925182 -0.1048796
[2,]  3.218926 -0.1436013 -0.6281626 -0.3424708 -0.3502629 -0.3002404 -0.1925182 -0.1048796
[3,] -0.270232 -0.1436013  1.8604013 -0.3424708  1.1245395  2.0469527 -0.1925182 -0.1048796
[4,] -0.270232 -0.1436013 -0.6281626  1.0182334 -0.3502629 -0.3002404 -0.1925182 -0.1048796
[5,] -0.270232 -0.1436013 -0.6281626 -0.3424708 -0.3502629 -0.3002404 -0.1925182 -0.1048796
[6,] -0.270232 -0.1436013 -0.6281626 -0.3424708 -0.3502629 -0.3002404 -0.1925182 -0.1048796
           hair      dress      blonde      mall   shopping    clothes  hollister abercrombie
[1,] -0.3885335 -0.2509497 -0.05054955 -0.373074 -0.4951734 -0.3181158 -0.2016674  -0.1848625
[2,] -0.3885335 -0.2509497 -0.05054955 -0.373074  2.2339332 -0.3181158 -0.2016674  -0.1848625
[3,]  6.7587536 -0.2509497  0.44099755  3.906738 -0.4951734  5.9573847 -0.2016674  -0.1848625
[4,]  0.5048774 -0.2509497 -0.05054955 -0.373074 -0.4951734 -0.3181158 -0.2016674  -0.1848625
[5,] -0.3885335 -0.2509497 -0.05054955 -0.373074 -0.4951734 -0.3181158 -0.2016674  -0.1848625
[6,] -0.3885335 -0.2509497 -0.05054955 -0.373074 -0.4951734 -0.3181158 -0.2016674  -0.1848625
            die      death     drunk      drugs
[1,] -0.3043178 -0.2645173 -0.223601 -0.1777646
[2,] -0.3043178 -0.2645173 -0.223601 -0.1777646
[3,]  2.9572033 -0.2645173 -0.223601  2.6363433
[4,]  1.3264428 -0.2645173  2.244607 -0.1777646
[5,] -0.3043178 -0.2645173 -0.223601 -0.1777646
[6,] -0.3043178 -0.2645173 -0.223601 -0.1777646
```

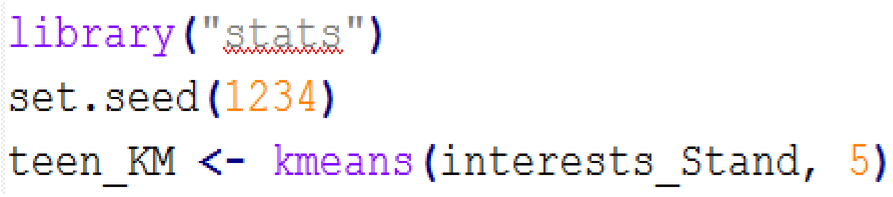
Step4: Build a Model

- **Problem:** Identify segments of teenagers who share similar tastes
- **Variables**
  - 36 words indicating interests/tastes
    - Basketball
    - Shopping
    - Abercrombie
    - Drunk
    - Drugs
    - …
- **Expected Clusters**
  - Five categories of interests
    - Extracurricular activities
    - Fashion
    - Religion
    - Romance
    - Antisocial behavior

K-means Clustering

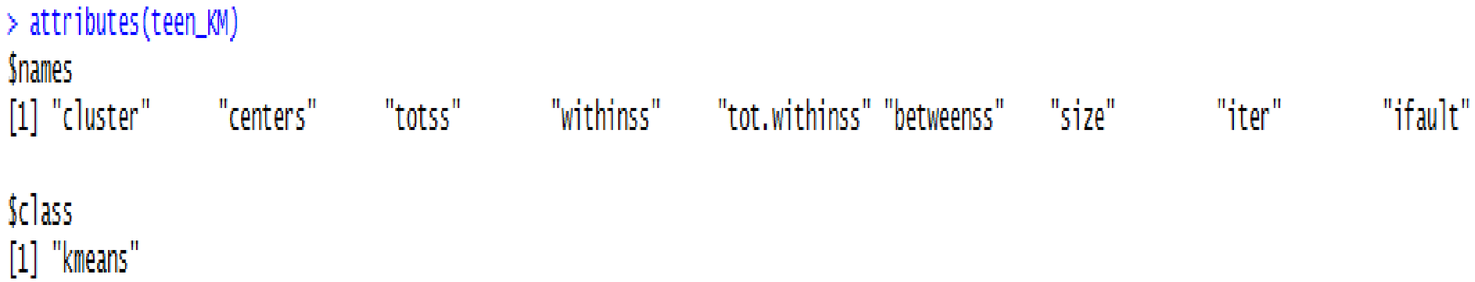
Step5: Train a Model on the Data

- K-means Clustering



```
library("stats")
set.seed(1234)
teen_KM <- kmeans(interests_Stand, 5)
```

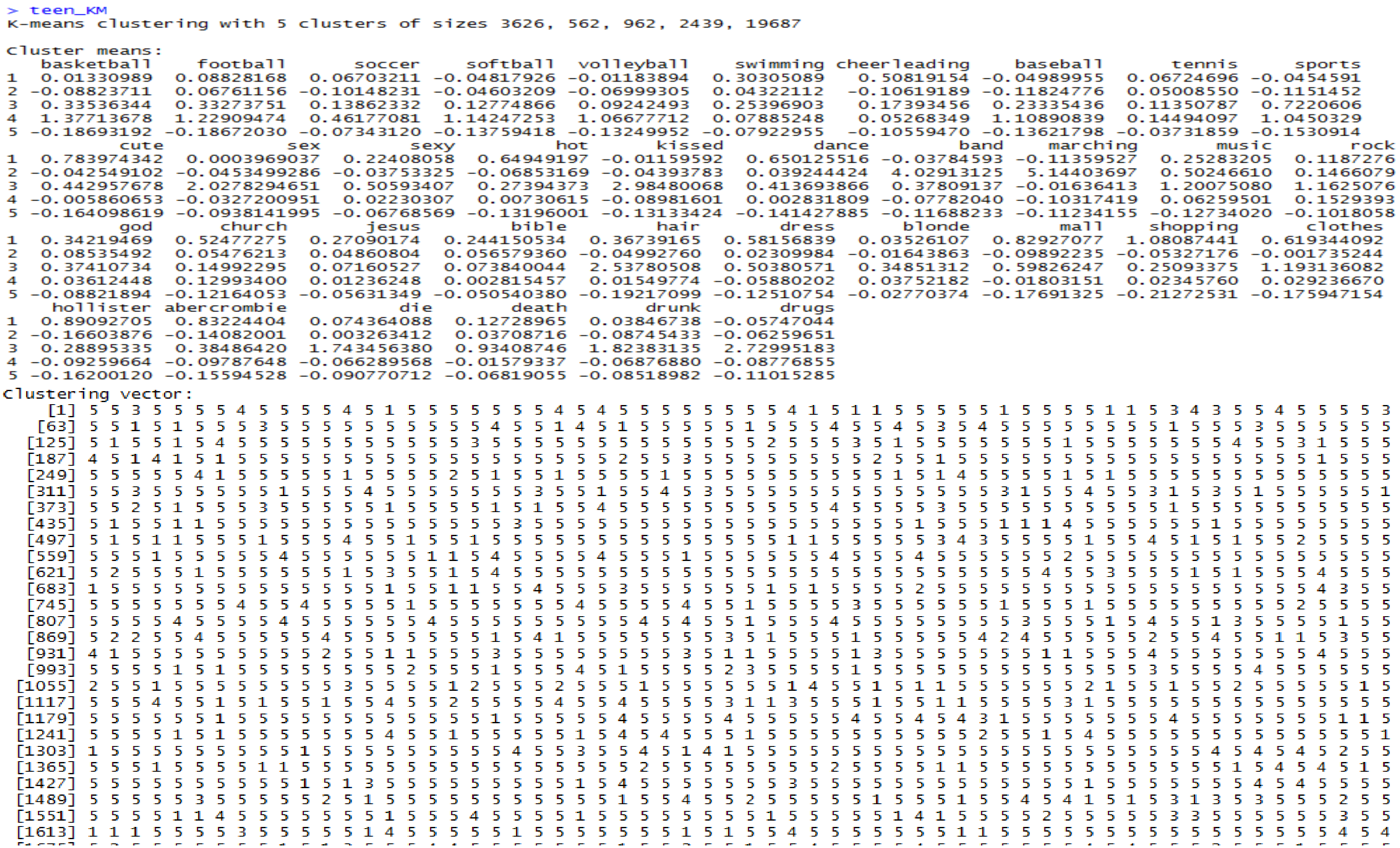
  - The results of k-means clustering process is a list named teen-KM that stores the properties of each of the five clusters.
    - cluster memberships, centroids, sum of squares (within, between, total), cluster sizes …



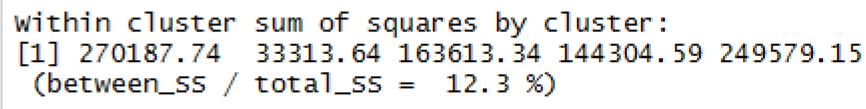
```
> attributes(teen_KM)
$names
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

$class
[1] "kmeans"
```

Step5: Train a Model on the Data

- K-means Clustering Output



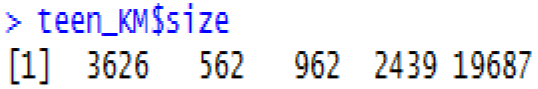
```
> teen_KM
K-means clustering with 5 clusters of sizes 3626, 562, 962, 2439, 19687

Cluster means:
   basketball    football      soccer    softball  volleyball    swimming cheerleading    baseball      tennis     sports
1  0.01330989  0.08828168  0.06703211 -0.04817926 -0.01183894  0.30305089   0.50819154 -0.04989955  0.06724696 -0.0454591
2 -0.08823711  0.06761156 -0.10148231 -0.04603209 -0.06999305  0.04322112  -0.10619189 -0.11824776  0.05008550 -0.1151452
3  0.33536344  0.33273751  0.13862332  0.12774866  0.09242493  0.25396903   0.17393456  0.23335436  0.11350787  0.7220606
4  1.37713678  1.22909474  0.46177081  1.14247253  1.06677712  0.07885248   0.05268349  1.10890839  0.14494097  1.0450329
5 -0.18693192 -0.18672030 -0.07343120 -0.13759418 -0.13249952 -0.07922955  -0.10559470 -0.13621798 -0.03731859 -0.1530914
          cute           sex       sexy         hot      kissed        dance        band    marching       music       rock
1  0.783974342  0.0003969037  0.22408058  0.64949197 -0.01159592  0.650125516 -0.03784593 -0.11359527  0.25283205  0.1187276
2 -0.042549102 -0.0453499286 -0.03753325 -0.06853169 -0.04393783  0.039244424  4.02913125  5.14403697  0.50246610  0.1466079
3  0.442957678  2.0278294651  0.50593407  0.27394373  2.98480068  0.413693866  0.37809137 -0.01636413  1.20075080  1.1625076
4 -0.005860653 -0.0327200951  0.02230307  0.00730615 -0.08981601  0.002831809 -0.07782040 -0.10317419  0.06259501  0.1529393
5 -0.164098619 -0.0938141995 -0.06768569 -0.13196001 -0.13133424 -0.141427885 -0.11688233 -0.11234155 -0.12734020 -0.1018058
          god      church       jesus        bible        hair       dress      blonde        mall    shopping      clothes
1  0.34219469  0.52477275  0.27090174  0.244150534  0.36739165  0.58156839  0.03526107  0.82927077  1.08087441  0.619344092
2  0.08535492  0.05476213  0.04860804  0.056579360 -0.04992760  0.02309984 -0.01643863 -0.09892235 -0.05327176 -0.001735244
3  0.37410734  0.14992295  0.07160527  0.073840044  2.53780508  0.50380571  0.34851312  0.59826247  0.25093375  1.193136082
4  0.03612448  0.12993400  0.01236248  0.002815457  0.01549774 -0.05880202  0.03752182 -0.01803151  0.02345760  0.029236670
5 -0.08821894 -0.12164053 -0.05631349 -0.050540380 -0.19217099 -0.12510754 -0.02770374 -0.17691325 -0.21272531 -0.175947154
    hollister abercrombie          die        death       drunk       drugs
1  0.89092705  0.83224404  0.074364088  0.12728965  0.03846738 -0.05747044
2 -0.16603876 -0.14082001  0.003263412  0.03708716 -0.08745433 -0.06259651
3  0.28895335  0.38486420  1.743456380  0.93408746  1.82383135  2.72995183
4 -0.09259664 -0.09787648 -0.066289568 -0.01579337 -0.06876880 -0.08776855
5 -0.16200120 -0.15594528 -0.090770712 -0.06819055 -0.08518982 -0.11015285
Clustering vector:
   [1] 5 5 3 5 5 5 5 4 5 5 5 5 4 5 1 5 5 5 5 5 5 5 4 5 4 5 5 5 5 5 5 5 5 4 1 5 1 1 5 5 5 5 5 1 5 5 5 5 1 1 5 3 4 3 5 5 4 5 5 5 5 3
  [63] 5 5 1 5 1 5 5 5 3 5 5 5 5 5 5 5 5 5 5 4 5 5 1 4 5 1 5 5 5 5 5 1 5 5 5 4 5 5 4 5 3 5 4 5 5 5 5 5 5 5 5 1 5 5 5 3 5 5 5 5 5 5
 [125] 5 1 5 5 1 5 4 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 2 5 5 5 3 5 1 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 4 5 5 3 1 5 5 5
 [187] 4 5 1 4 1 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 2 5 5 3 5 5 5 5 5 5 5 5 2 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5
 [249] 5 5 5 5 5 4 1 5 5 5 5 5 1 5 5 5 5 2 5 1 5 5 1 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5 1 5 1 4 5 5 5 5 1 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5
 [311] 5 5 3 5 5 5 5 5 5 1 5 5 5 4 5 5 5 5 5 5 5 3 5 5 1 5 5 4 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 3 1 5 5 4 5 5 3 1 5 3 5 1 5 5 5 5 5 1
 [373] 5 5 2 5 1 5 5 5 3 5 5 5 5 5 1 5 5 5 5 1 5 1 5 5 4 5 5 5 5 5 5 5 5 5 5 4 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5
 [435] 5 1 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 1 1 1 4 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5
 [497] 5 1 5 1 1 5 5 5 1 5 5 5 4 5 5 1 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 1 5 5 5 5 5 3 4 3 5 5 5 5 1 5 5 4 5 1 5 1 5 5 2 5 5 5 5
 [559] 5 5 5 1 5 5 5 5 5 4 5 5 5 5 5 5 1 1 5 4 5 5 5 5 4 5 5 5 1 5 5 5 5 5 5 4 5 5 5 4 5 5 5 5 5 5 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [621] 5 2 5 5 5 1 5 5 5 5 5 5 1 5 3 5 5 1 5 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 5 3 5 5 5 1 5 1 5 5 5 4 5 5 5
 [683] 1 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 1 1 5 5 4 5 5 5 3 5 5 5 5 5 5 1 5 1 5 5 5 5 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 3 5 5
 [745] 5 5 5 5 5 5 5 4 5 5 4 5 5 5 5 1 5 5 5 5 5 5 5 4 5 5 5 5 4 5 5 1 5 5 5 5 3 5 5 5 5 5 5 1 5 5 5 1 5 5 5 5 5 5 5 5 5 2 5 5 5 5
 [807] 5 5 5 5 4 5 5 5 5 4 5 5 5 5 5 5 4 5 5 5 5 5 5 5 5 5 4 5 4 5 5 1 5 5 5 4 5 5 5 5 5 5 5 5 3 5 5 5 1 5 4 5 5 1 3 5 5 5 5 1 5 5
 [869] 5 2 2 5 5 4 5 5 5 5 5 4 5 5 5 5 5 5 5 1 5 4 1 5 5 5 5 5 5 5 3 5 1 5 5 5 1 5 5 5 5 5 4 2 4 5 5 5 5 5 2 5 5 4 5 5 1 1 5 3 5 5
 [931] 4 1 5 5 5 5 5 5 5 5 5 2 5 5 1 1 5 5 5 3 5 5 5 5 5 5 5 5 3 5 1 1 5 5 5 5 1 3 5 5 5 5 5 5 5 1 1 5 5 5 4 5 5 5 5 5 5 5 4 5 5 5
 [993] 5 5 5 5 1 5 1 5 5 5 5 5 5 5 5 2 5 5 5 1 5 5 5 4 5 1 5 5 5 5 2 3 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 4 5 5 5 5 5 5 5
[1055] 2 5 5 1 5 5 5 5 5 5 5 5 3 5 5 5 5 1 2 5 5 5 2 5 5 5 1 5 5 5 5 5 5 1 4 5 5 1 5 1 1 5 5 5 5 5 5 2 1 5 5 1 5 5 2 5 5 5 5 5 1 5
[1117] 5 5 5 4 5 5 1 5 1 5 5 1 5 5 4 5 5 2 5 5 5 5 4 5 5 4 5 5 5 5 3 1 1 3 5 5 5 1 5 5 1 1 5 5 5 5 3 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[1179] 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 4 5 5 5 5 4 5 5 5 5 5 4 5 5 4 5 4 3 1 5 5 5 5 5 5 5 4 5 5 5 5 5 5 5 1 1 5
[1241] 5 5 5 5 1 5 1 5 5 5 5 5 5 5 4 5 5 1 5 5 5 5 5 1 5 4 5 4 5 5 5 1 5 5 5 5 5 5 5 5 5 5 2 5 5 1 5 4 5 5 5 5 5 5 5 5 5 5 5 5 5 1
[1303] 1 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 4 5 5 3 5 5 4 5 1 4 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 4 5 4 5 2 5 5
[1365] 5 5 5 1 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 2 5 5 5 5 5 5 5 5 2 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 1 5 4 5 4 5 1 5
[1427] 5 5 5 5 5 5 5 5 5 5 1 5 1 3 5 5 5 5 5 5 5 5 5 1 5 4 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 4 5 4 5 5 5 5
[1489] 5 5 5 5 5 3 5 5 5 5 5 2 5 1 5 5 5 5 5 5 5 5 5 5 5 1 5 5 4 5 5 2 5 5 5 5 5 1 5 5 5 1 5 5 4 5 4 1 5 1 5 3 1 3 5 3 5 5 5 2 5 5
[1551] 5 5 5 5 1 1 4 5 5 5 5 5 5 5 1 5 5 5 4 5 5 5 5 1 5 5 5 5 5 5 5 5 1 5 5 5 5 5 1 4 1 5 5 5 5 2 5 5 5 5 5 3 3 5 5 5 5 5 5 3 5 5
[1613] 1 1 1 5 5 5 5 3 5 5 5 5 5 1 4 5 5 5 5 5 1 5 5 5 5 5 5 5 1 5 1 5 5 4 5 5 5 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 4
```



```
Within cluster sum of squares by cluster:
[1] 270187.74  33313.64 163613.34 144304.59 249579.15
 (between_SS / total_SS =  12.3 %)
```

Step6: Validate Clusters

- Evaluating clustering results can be somewhat subjective. Ultimately, the success of failure of the model hinges on whether the clusters are useful for the intended purpose.
- As the goal of this analysis was to identify clusters of teenagers with similar interests for marketing purposes, we will largely measure our success in qualitative terms.
  - Look at the size of the clusters (i.e., #teens for each cluster)

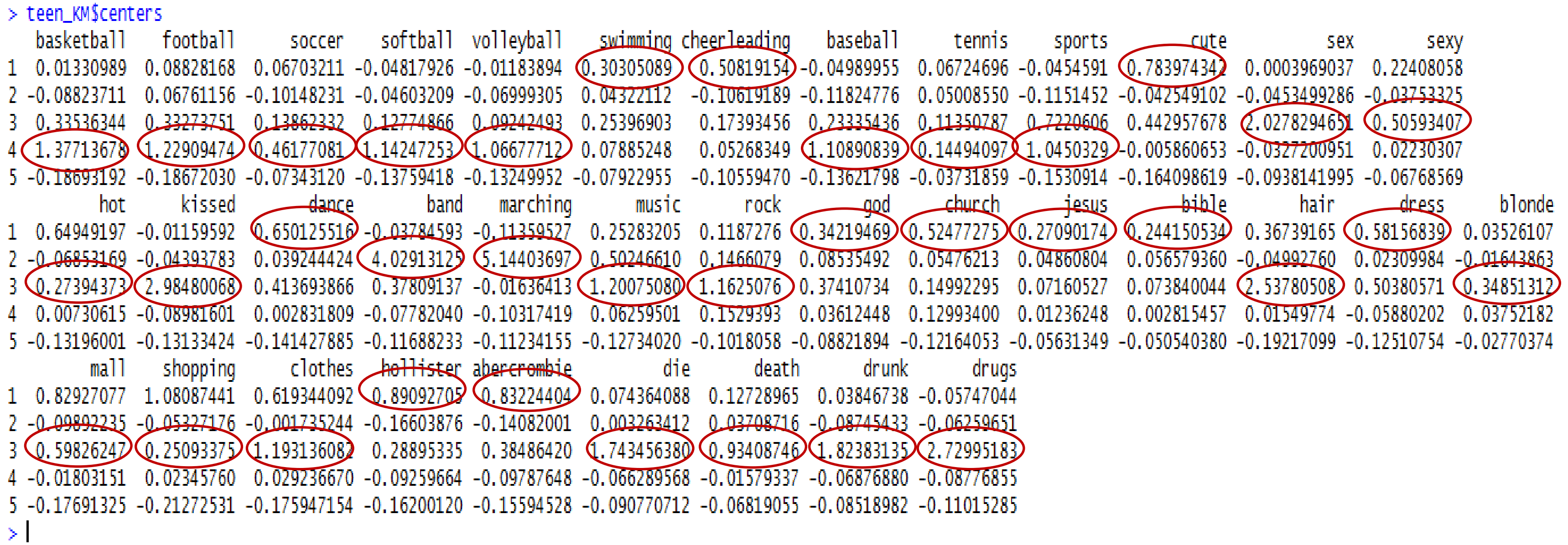


```
> teen_KM$size
[1]  3626   562   962  2439 19687
```

  - The smallest cluster has 562 teenagers while the largest cluster has 19,687.
  - The clusters' size disparity indicates something real.
    - A big group of teens that share similar interests
    - Or, a random fluke caused by the initial k-means cluster centers (centroids)

Step6: Validate Clusters

- Return the centers for SNS Keywords

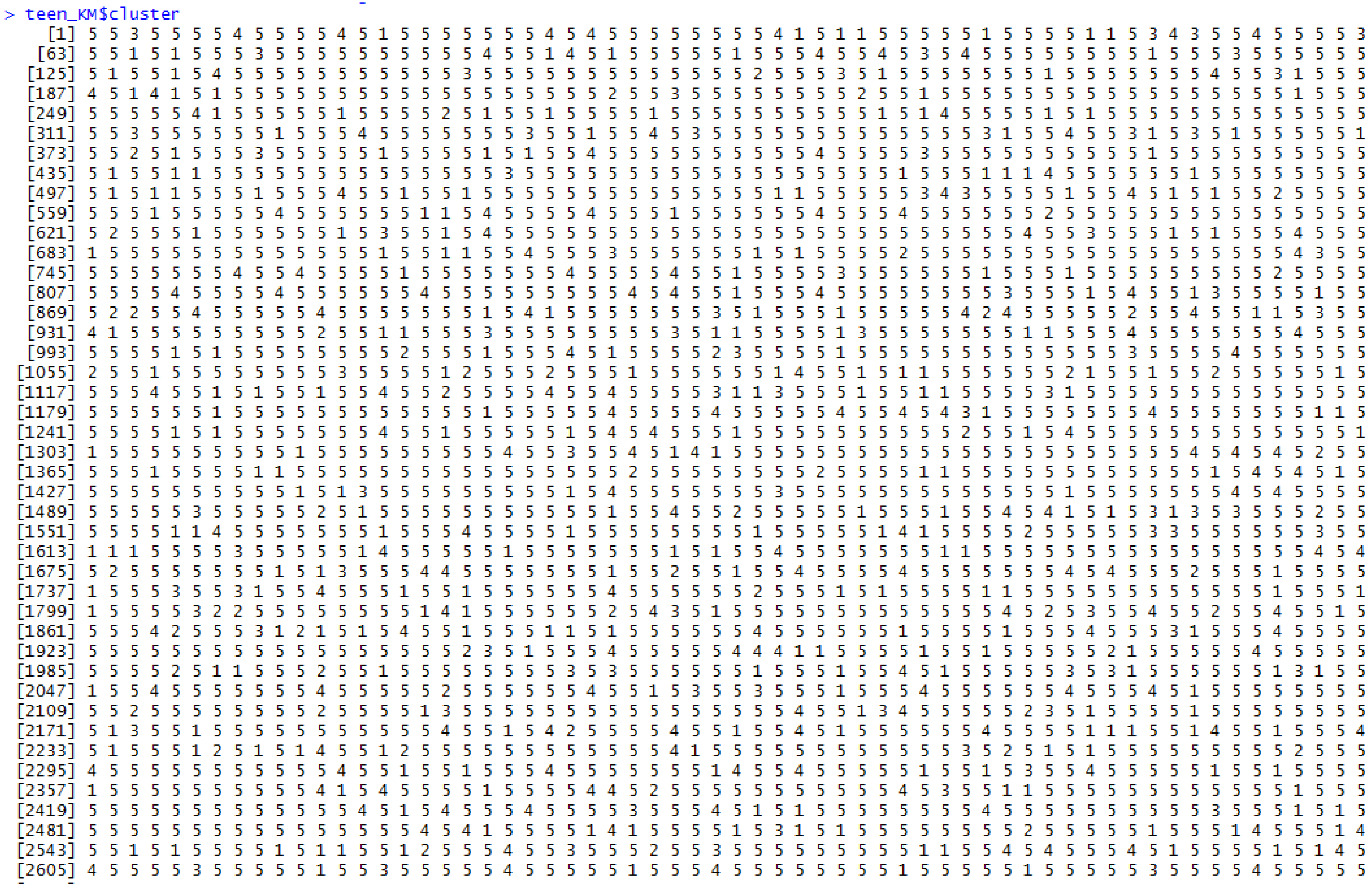


```
> teen_KM$centers
   basketball    football      soccer    softball  volleyball    swimming cheerleading    baseball     tennis     sports         cute           sex        sexy
1  0.01330989  0.08828168  0.06703211 -0.04817926 -0.01183894  0.30305089  0.50819154 -0.04989955 0.06724696 -0.0454591  0.783974342  0.0003969037  0.22408058
2 -0.08823711  0.06761156 -0.10148231 -0.04603209 -0.06999305  0.04322112 -0.10619189 -0.11824776 0.05008550 -0.1151452 -0.042549102 -0.0453499286 -0.03753325
3  0.33536344  0.33273751  0.13862332  0.12774866  0.09242493  0.25396903  0.17393456  0.23335436 0.11350787  0.7220606  0.442957678  2.0278294651  0.50593407
4  1.37713678  1.22909474  0.46177081  1.14247253  1.06677712  0.07885248  0.05268349  1.10890839 0.14494097  1.0450329 -0.005860653 -0.0327200951  0.02230307
5 -0.18693192 -0.18672030 -0.07343120 -0.13759418 -0.13249952 -0.07922955 -0.10559470 -0.13621798 -0.03731859 -0.1530914 -0.164098619 -0.0938141995 -0.06768569
          hot      kissed       dance        band    marching       music       rock         god     church      jesus        bible        hair       dress      blonde
1  0.64949197 -0.01159592 0.650125516 -0.03784593 -0.11359527  0.25283205  0.1187276  0.34219469 0.52477275 0.27090174  0.244150534  0.36739165  0.58156839  0.03526107
2 -0.06853169 -0.04393783 0.039244424  4.02913125  5.14403697  0.50246610  0.1466079  0.08535492 0.05476213 0.04860804  0.056579360 -0.04992760  0.02309984 -0.01643863
3  0.27394373  2.98480068 0.413693866  0.37809137 -0.01636413  1.20075080  1.1625076  0.37410734 0.14992295 0.07160527  0.073840044  2.53780508  0.50380571  0.34851312
4  0.00730615 -0.08981601 0.002831809 -0.07782040 -0.10317419  0.06259501  0.1529393  0.03612448 0.12993400 0.01236248  0.002815457  0.01549774 -0.05880202  0.03752182
5 -0.13196001 -0.13133424 -0.141427885 -0.11688233 -0.11234155 -0.12734020 -0.1018058 -0.08821894 -0.12164053 -0.05631349 -0.050540380 -0.19217099 -0.12510754 -0.02770374
         mall    shopping      clothes   hollister abercrombie          die       death       drunk       drugs
1  0.82927077  1.08087441  0.619344092  0.89092705  0.83224404  0.074364088  0.12728965  0.03846738 -0.05747044
2 -0.09892235 -0.05327176 -0.001735244 -0.16603876 -0.14082001  0.003263412  0.03708716 -0.08745433 -0.06259651
3  0.59826247  0.25093375  1.193136082  0.28895335  0.38486420  1.743456380  0.93408746  1.82383135  2.72995183
4 -0.01803151  0.02345760  0.029236670 -0.09259664 -0.09787648 -0.066289568 -0.01579337 -0.06876880 -0.08776855
5 -0.17691325 -0.21272531 -0.175947154 -0.16200120 -0.15594528 -0.090770712 -0.06819055 -0.08518982 -0.11015285
> |
```

- Rows: 5 clusters
- Columns: the clusters' average values
  - The fourth row has the highest value in the basketball column, which means that cluster 4 has the highest average interest in basketball among all the clusters.
- Cluster 1: God, Church, Jesus, Bible, Swimming, Cheerleading, Cute, Dance, Dress, Hollister, Abercrombie
- Cluster 2: Band, Marching
- Cluster 3: Die, Death, Drunk, Drugs, Sex, Sexy, Hot, Kissed, Hair, Blonde, Rock, Music, Mall, Shopping, Clothes
- Cluster 4: Basketball, Football, Soccer, Softball, Volleyball, Baseball, Tennis, Sports
- Cluster 5: ??

Step6: Validate Clusters

- Return the clusters for teenagers



```
> teen_KM$cluster
   [1] 5 5 3 5 5 5 5 4 5 5 5 5 4 5 1 5 5 5 5 5 5 5 4 5 4 5 5 5 5 5 5 5 5 4 1 5 1 1 5 5 5 5 5 1 5 5 5 5 1 1 5 3 4 3 5 5 4 5 5 5 5 3
  [63] 5 5 1 5 1 5 5 5 3 5 5 5 5 5 5 5 5 5 5 4 5 5 1 4 5 1 5 5 5 5 5 1 5 5 5 4 5 5 4 5 3 5 4 5 5 5 5 5 5 5 5 1 5 5 5 3 5 5 5 5 5 5
 [125] 5 1 5 5 1 5 4 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 2 5 5 5 3 5 1 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 4 5 5 3 1 5 5 5
 [187] 4 5 1 4 1 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 2 5 5 3 5 5 5 5 5 5 5 5 2 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5
 [249] 5 5 5 5 5 4 1 5 5 5 5 5 1 5 5 5 5 2 5 1 5 5 1 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5 1 5 1 4 5 5 5 5 1 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5
 [311] 5 5 3 5 5 5 5 5 5 1 5 5 5 4 5 5 5 5 5 5 5 3 5 5 1 5 5 4 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 3 1 5 5 4 5 5 3 1 5 3 5 1 5 5 5 5 5 1
 [373] 5 5 2 5 1 5 5 5 3 5 5 5 5 5 1 5 5 5 5 1 5 1 5 5 4 5 5 5 5 5 5 5 5 5 5 4 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5
 [435] 5 1 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 1 1 1 4 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5
 [497] 5 1 5 1 1 5 5 5 1 5 5 5 4 5 5 1 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 1 1 5 5 5 5 5 3 4 3 5 5 5 5 1 5 5 4 5 1 5 1 5 5 2 5 5 5 5
 [559] 5 5 5 1 5 5 5 5 5 4 5 5 5 5 5 5 1 1 5 4 5 5 5 5 4 5 5 5 1 5 5 5 5 5 5 4 5 5 5 4 5 5 5 5 5 5 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [621] 5 2 5 5 5 1 5 5 5 5 5 5 1 5 3 5 5 1 5 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 5 3 5 5 5 1 5 1 5 5 5 4 5 5 5
 [683] 1 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 1 1 5 5 4 5 5 5 3 5 5 5 5 5 5 1 5 1 5 5 5 5 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 3 5 5
 [745] 5 5 5 5 5 5 5 4 5 5 4 5 5 5 5 1 5 5 5 5 5 5 5 4 5 5 5 5 4 5 5 1 5 5 5 5 3 5 5 5 5 5 5 1 5 5 5 1 5 5 5 5 5 5 5 5 5 2 5 5 5 5
 [807] 5 5 5 5 4 5 5 5 5 4 5 5 5 5 5 5 4 5 5 5 5 5 5 5 5 5 4 5 4 5 5 1 5 5 5 4 5 5 5 5 5 5 5 5 3 5 5 5 1 5 4 5 5 1 3 5 5 5 5 1 5 5
 [869] 5 2 2 5 5 4 5 5 5 5 5 4 5 5 5 5 5 5 5 1 5 4 1 5 5 5 5 5 5 5 3 5 1 5 5 5 1 5 5 5 5 5 4 2 4 5 5 5 5 5 2 5 5 4 5 5 1 1 5 3 5 5
 [931] 4 1 5 5 5 5 5 5 5 5 5 2 5 5 1 1 5 5 5 3 5 5 5 5 5 5 5 5 3 5 1 1 5 5 5 5 1 3 5 5 5 5 5 5 5 1 1 5 5 5 4 5 5 5 5 5 5 5 4 5 5 5
 [993] 5 5 5 5 1 5 1 5 5 5 5 5 5 5 5 2 5 5 5 1 5 5 5 4 5 1 5 5 5 5 2 3 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 4 5 5 5 5 5 5
[1055] 2 5 5 1 5 5 5 5 5 5 5 5 3 5 5 5 5 1 2 5 5 5 2 5 5 5 1 5 5 5 5 5 5 1 4 5 5 1 5 1 1 5 5 5 5 5 5 2 1 5 5 1 5 5 2 5 5 5 5 5 1 5
[1117] 5 5 5 4 5 5 1 5 1 5 5 1 5 5 4 5 5 2 5 5 5 5 4 5 5 4 5 5 5 5 3 1 1 3 5 5 5 1 5 5 1 1 5 5 5 5 3 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[1179] 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 4 5 5 5 5 4 5 5 5 5 5 4 5 5 4 5 4 3 1 5 5 5 5 5 5 5 4 5 5 5 5 5 5 5 1 1 5
[1241] 5 5 5 5 1 5 1 5 5 5 5 5 5 5 4 5 5 1 5 5 5 5 5 1 5 4 5 4 5 5 5 1 5 5 5 5 5 5 5 5 5 5 2 5 5 1 5 4 5 5 5 5 5 5 5 5 5 5 5 5 5 1
[1303] 1 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 5 5 4 5 5 3 5 5 4 5 1 4 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 4 5 4 5 2 5 5
[1365] 5 5 5 1 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 2 5 5 5 5 5 5 5 5 2 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 1 5 4 5 4 5 1 5
[1427] 5 5 5 5 5 5 5 5 5 5 1 5 1 3 5 5 5 5 5 5 5 5 5 1 5 4 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 5 5 5 5 4 5 4 5 5 5 5
[1489] 5 5 5 5 5 3 5 5 5 5 5 2 5 1 5 5 5 5 5 5 5 5 5 5 5 1 5 5 4 5 5 2 5 5 5 5 5 1 5 5 5 1 5 5 4 5 4 1 5 1 5 3 1 3 5 3 5 5 5 2 5 5
[1551] 5 5 5 5 1 1 4 5 5 5 5 5 5 5 1 5 5 5 4 5 5 5 5 1 5 5 5 5 5 5 5 5 1 5 5 5 5 5 1 4 1 5 5 5 5 2 5 5 5 5 5 3 3 5 5 5 5 5 5 3 5 5
[1613] 1 1 1 5 5 5 5 3 5 5 5 5 5 1 4 5 5 5 5 5 1 5 5 5 5 5 5 5 1 5 1 5 5 4 5 5 5 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 4
[1675] 5 2 5 5 5 5 5 5 5 1 5 1 3 5 5 5 4 4 5 5 5 5 5 5 5 1 5 5 2 5 5 1 5 5 4 5 5 5 5 4 5 5 5 5 5 5 5 4 5 4 5 5 5 2 5 5 5 1 5 5 5 5
[1737] 1 5 5 5 3 5 5 3 1 5 5 4 5 5 5 1 5 5 1 5 5 5 5 5 5 4 5 5 5 5 5 5 5 2 5 5 5 1 5 1 5 5 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5 1
[1799] 1 5 5 5 5 3 2 2 5 5 5 5 5 5 5 5 1 4 1 5 5 5 5 5 5 2 5 4 3 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 2 5 3 5 5 4 5 5 2 5 5 4 5 5 1 5
[1861] 5 5 5 4 2 5 5 5 3 1 2 1 5 1 5 4 5 5 1 5 5 5 1 1 5 1 5 5 5 5 5 5 4 5 5 5 5 5 5 1 5 5 5 5 1 5 5 5 4 5 5 5 3 1 5 5 5 4 5 5 5 5
[1923] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 2 3 5 1 5 5 5 4 5 5 5 5 5 4 4 4 1 1 5 5 5 5 1 5 5 1 5 5 5 5 5 2 1 5 5 5 5 5 4 5 5 5 5 5
[1985] 5 5 5 5 2 5 1 1 5 5 5 2 5 5 1 5 5 5 5 5 5 5 5 3 5 3 5 5 5 5 5 5 1 5 5 5 1 5 5 4 5 1 5 5 5 5 5 3 5 3 1 5 5 5 5 5 5 1 3 1 5 5
[2047] 1 5 5 4 5 5 5 5 5 5 5 4 5 5 5 5 5 2 5 5 5 5 5 5 4 5 5 1 5 3 5 5 3 5 5 5 1 5 5 5 4 5 5 5 5 5 5 4 5 5 5 4 5 1 5 5 5 5 5 5 5 5
[2109] 5 5 2 5 5 5 5 5 5 5 5 2 5 5 5 5 1 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 5 1 3 4 5 5 5 5 5 2 3 5 1 5 5 5 5 1 5 5 5 5 5 5 5 5
[2171] 5 1 3 5 5 1 5 5 5 5 5 5 5 5 5 5 5 4 5 5 1 5 4 2 5 5 5 5 4 5 5 1 5 5 4 5 1 5 5 5 5 5 5 4 5 5 5 5 1 1 1 5 5 1 4 5 5 1 5 5 5 4
[2233] 5 1 5 5 5 1 2 5 1 5 1 4 5 5 1 2 5 5 5 5 5 5 5 5 5 5 5 5 4 1 5 5 5 5 5 5 5 5 5 5 5 5 3 5 2 5 1 5 1 5 5 5 5 5 5 5 5 5 2 5 5 5
[2295] 4 5 5 5 5 5 5 5 5 5 5 5 4 5 5 1 5 5 1 5 5 5 4 5 5 5 5 5 5 5 1 4 5 5 4 5 5 5 5 5 1 5 5 1 5 3 5 5 4 5 5 5 5 5 1 5 5 1 5 5 5 5
[2357] 1 5 5 5 5 5 5 5 5 5 5 4 1 5 4 5 5 5 5 1 5 5 5 5 4 4 5 2 5 5 5 5 5 5 5 5 5 5 5 4 5 3 5 5 1 1 5 5 5 5 5 5 5 5 5 5 5 5 1 5 5 5
[2419] 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 1 5 4 5 5 5 4 5 5 5 5 3 5 5 5 4 5 1 5 1 5 5 5 5 5 5 5 5 4 5 5 5 5 5 5 5 5 5 5 3 5 5 5 1 5 1 5
[2481] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 4 1 5 5 5 5 1 4 1 5 5 5 5 1 5 3 1 5 1 5 5 5 5 5 5 5 5 2 5 5 5 5 5 1 5 5 5 1 4 5 5 5 1 4
[2543] 5 5 1 5 1 5 5 5 5 1 5 1 1 5 5 1 2 5 5 5 4 5 5 3 5 5 5 2 5 5 3 5 5 5 5 5 5 5 5 5 1 1 5 5 4 5 4 5 5 5 4 5 1 5 5 5 5 1 5 1 4 5
[2605] 4 5 5 5 5 3 5 5 5 5 5 1 5 5 3 5 5 5 5 5 4 5 5 5 5 5 1 5 5 5 4 5 5 5 5 5 5 5 5 1 5 5 5 5 5 1 5 5 5 5 5 3 5 5 5 5 4 5 5 5 5 5
```

Step6: Validate Clusters

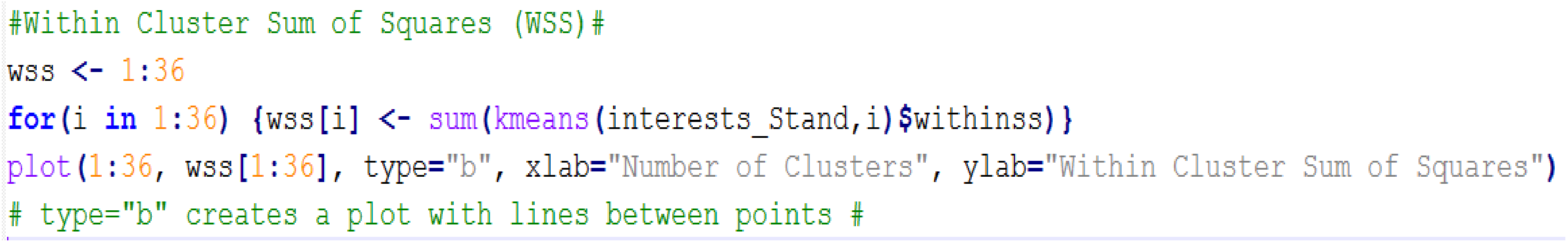
- Characterize the clusters
  - Expected Clusters: Extracurricular activities, Fashion, Religion, Romance, and Antisocial behavior
  - Output Clusters:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
N = 3,626	N = 562	N = 962	N = 2,439	N = 19,687
God Church Jesus Bible Swimming Cheerleading Cute Dance Dress Hollister Abercrombie	Band Marching	Die Death Drunk Drugs Sex Sexy Hot Kissed Hair Blonde Rock Music Mall Shopping Clothes	Basketball Football Soccer Softball Volleyball Baseball Tennis Sports	??
Religion + Fashion	E.A.1	Anti-social	E.A.2 (Sports)	Nothing

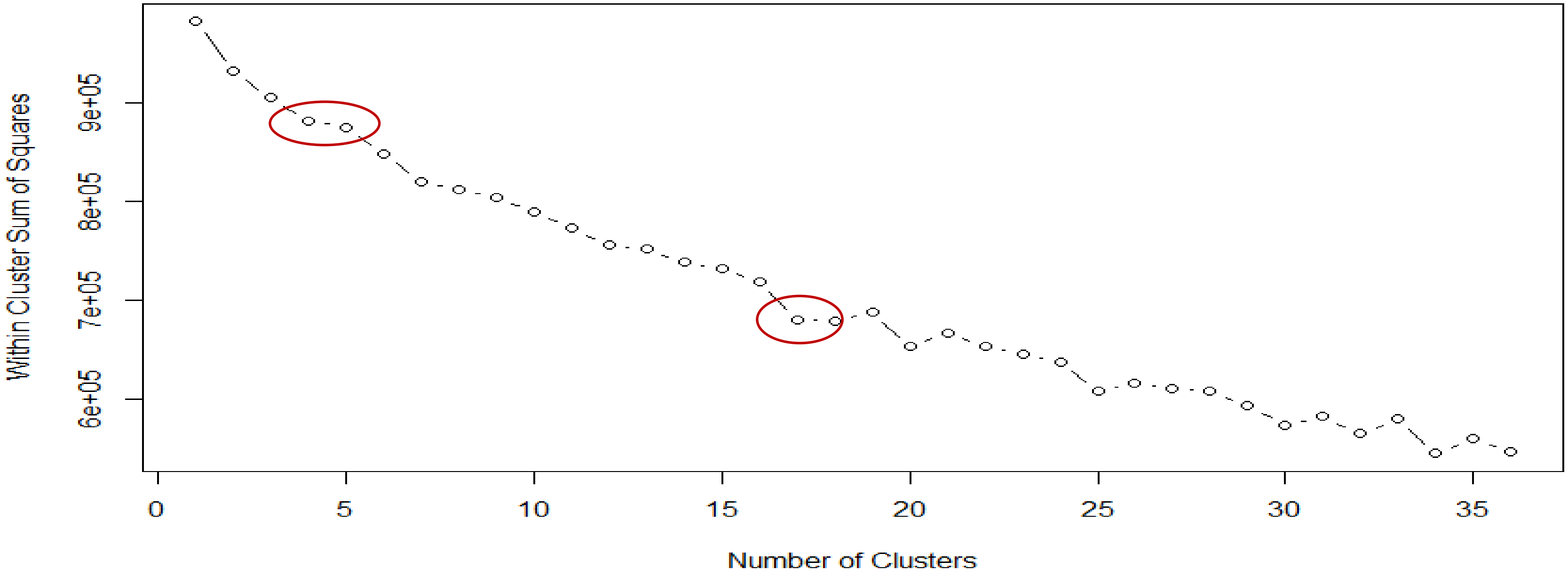
  - Cluster 5:
    - It's members had lower-than-average levels of interest in every measured activity.
    - It is also the single largest group in terms of the number of members.
    - Possibly, these teens created a profile on a SNS but never posted any interests

Step6: Validate Clusters

- Determine the appropriate value of k (Elbow Method)



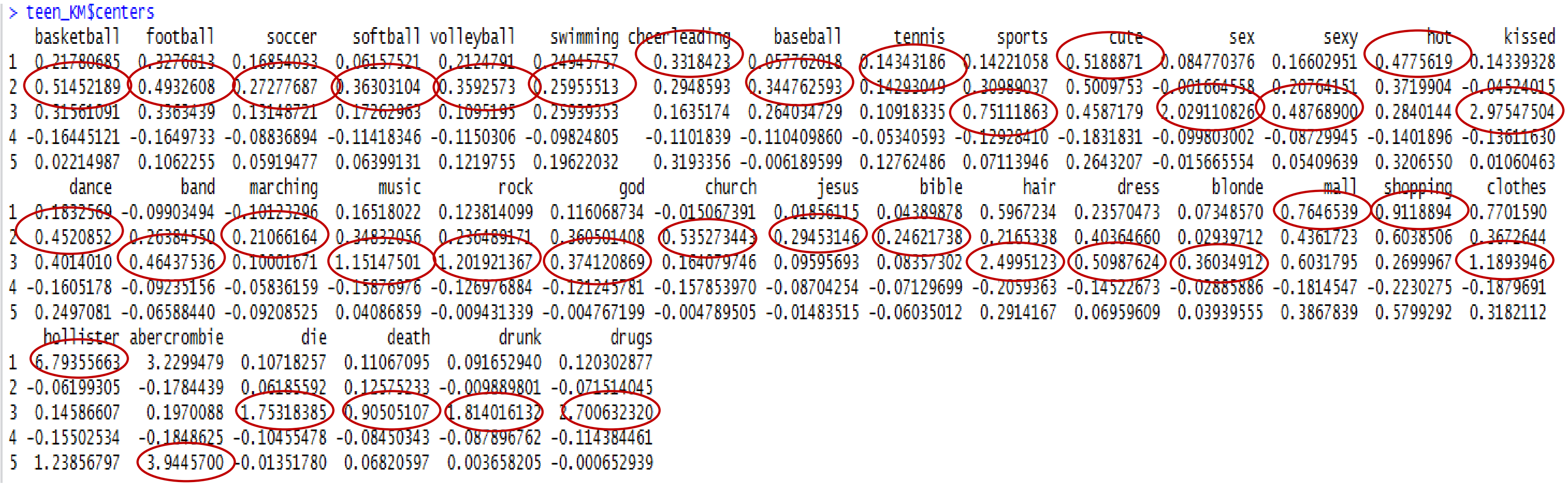
```
#Within Cluster Sum of Squares (WSS)#
wss <- 1:36
for(i in 1:36) {wss[i] <- sum(kmeans(interests_Stand,i)$withinss)}
plot(1:36, wss[1:36], type="b", xlab="Number of Clusters", ylab="Within Cluster Sum of Squares")
# type="b" creates a plot with lines between points #
```



- Which k is better?

Step6: Validate Clusters

- With set.seed(2406)



```
> teen_KM$centers
   basketball   football      soccer    softball volleyball    swimming cheerleading     baseball      tennis      sports      cute          sex        sexy       hot      kissed
1  0.21780685  0.3276813  0.16854033  0.06157521  0.2124791  0.24945757    0.3318423  0.057762018  0.14343186  0.14221058 0.5188871  0.084770376  0.16602951 0.4775619  0.14339328
2  0.51452189  0.4932608  0.27277687  0.36303104  0.3592573  0.25955513    0.2948593  0.344762593  0.14293049  0.30989037 0.5009753 -0.001664558  0.20764151 0.3719904 -0.04524015
3  0.31561091  0.3363439  0.13148721  0.17262963  0.1095195  0.25939353    0.1635174  0.264034729  0.10918335  0.75111863 0.4587179  2.029110826  0.48768900 0.2840144  2.97547504
4 -0.16445121 -0.1649733 -0.08836894 -0.11418346 -0.1150306 -0.09824805   -0.1101839 -0.110409860 -0.05340593 -0.12928410 -0.1831831 -0.099803002 -0.08729945 -0.1401896 -0.13611630
5  0.02214987  0.1062255  0.05919477  0.06399131  0.1219755  0.19622032    0.3193356 -0.006189599  0.12762486  0.07113946 0.2643207 -0.015665554  0.05409639 0.3206550  0.01060463
        dance        band    marching       music         rock          god       church       jesus       bible       hair       dress      blonde       mall  shopping   clothes
1   0.1832569 -0.09903494 -0.10123296  0.16518022  0.123814099  0.116068734 -0.015067391  0.01856115  0.04389878  0.5967234  0.23570473  0.07348570  0.7646539 0.9118894 0.7701590
2   0.4520852  0.26384550  0.21066164  0.34832056  0.236489171  0.360501408  0.535273443  0.29453146  0.24621738  0.2165338  0.40364660  0.02939712  0.4361723 0.6038506 0.3672644
3   0.4014010  0.46437536  0.10001671  1.15147501  1.201921367  0.374120869  0.164079746  0.09595693  0.08357302  2.4995123  0.50987624  0.36034912  0.6031795 0.2699967 1.1893946
4  -0.1605178 -0.09235156 -0.05836159 -0.15876976 -0.126976884 -0.121245781 -0.157853970 -0.08704254 -0.07129699 -0.2059363 -0.14522673 -0.02885886 -0.1814547 -0.2230275 -0.1879691
5   0.2497081 -0.06588440 -0.09208525  0.04086859 -0.009431339 -0.004767199 -0.004789505 -0.01483515 -0.06035012  0.2914167  0.06959609  0.03939555  0.3867839 0.5799292 0.3182112
    hollister abercrombie         die       death        drunk        drugs
1  6.79355663   3.2299479  0.10718257  0.11067095  0.091652940  0.120302877
2 -0.06199305  -0.1784439  0.06185592  0.12575233 -0.009889801 -0.071514045
3  0.14586607   0.1970088  1.75318385  0.90505107  1.814016132  2.700632320
4 -0.15502534  -0.1848625 -0.10455478 -0.08450343 -0.087896762 -0.114384461
5  1.23856797   3.9445700 -0.01351780  0.06820597  0.003658205 -0.000652939
```

- Cluster 1: Cheerleading, Tennis, Cute, Hot, Mall, Shopping, Hollister
- Cluster 2: Basketball, Football, Soccer, Softball, Baseball, Dance, Marching, Church, Jesus, Bible
- Cluster 3: Sports, Sex, Sexy, Kissed, Band, Music, Rock, God, Hair, Dress, Blonde, Clothes, Die, Drunk, Drugs
- Cluster 4: ??
- Cluster 5: Abercrombie

Step6: Validate Clusters

- Characterize the clusters (set.seed(2406))
  - Expected Clusters: Extracurricular activities, Fashion, Religion, Romance, and Antisocial behavior
  - Output Clusters:

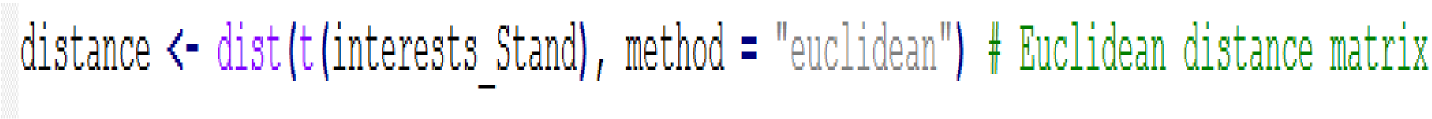
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
N = 321	N = 5,509	N = 963	N =19,625	N = 858
Cheer leading Tennis Cute Hot Mall Shopping Hollister	Basketball Football Soccer Softball Baseball Dance Marching Church Jesus Bible	Sports, Sex, Sexy, Kissed, Band, Music, Rock, God, Hair, Dress, Blonde, Clothes, Die, Drunk, Drugs	??	Abercrombie
Fashion or Girlish	E.A. or Boyish	Anti-Social Wannabes	Nothing	Fashion or Luxurious

Hierarchical Clustering

Case 1: Identify the Clusters of SNS Keywords

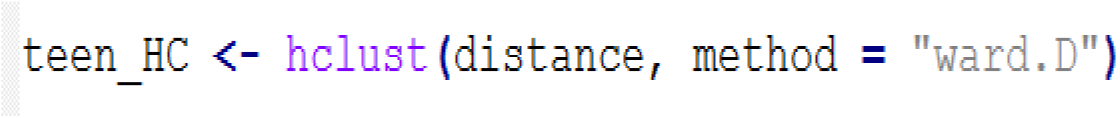
Step5: Train a Model on the Data

- Hierarchical Clustering
  - Measure distance between records
    - Use the Euclidean distance as an input for clustering
    - As we want to cluster the SNS keywords, the interests_Stand matrix should be transposed.
      - Identify the clusters of SNS keywords



```
distance <- dist(t(interests_Stand), method = "euclidean") # Euclidean distance matrix
```

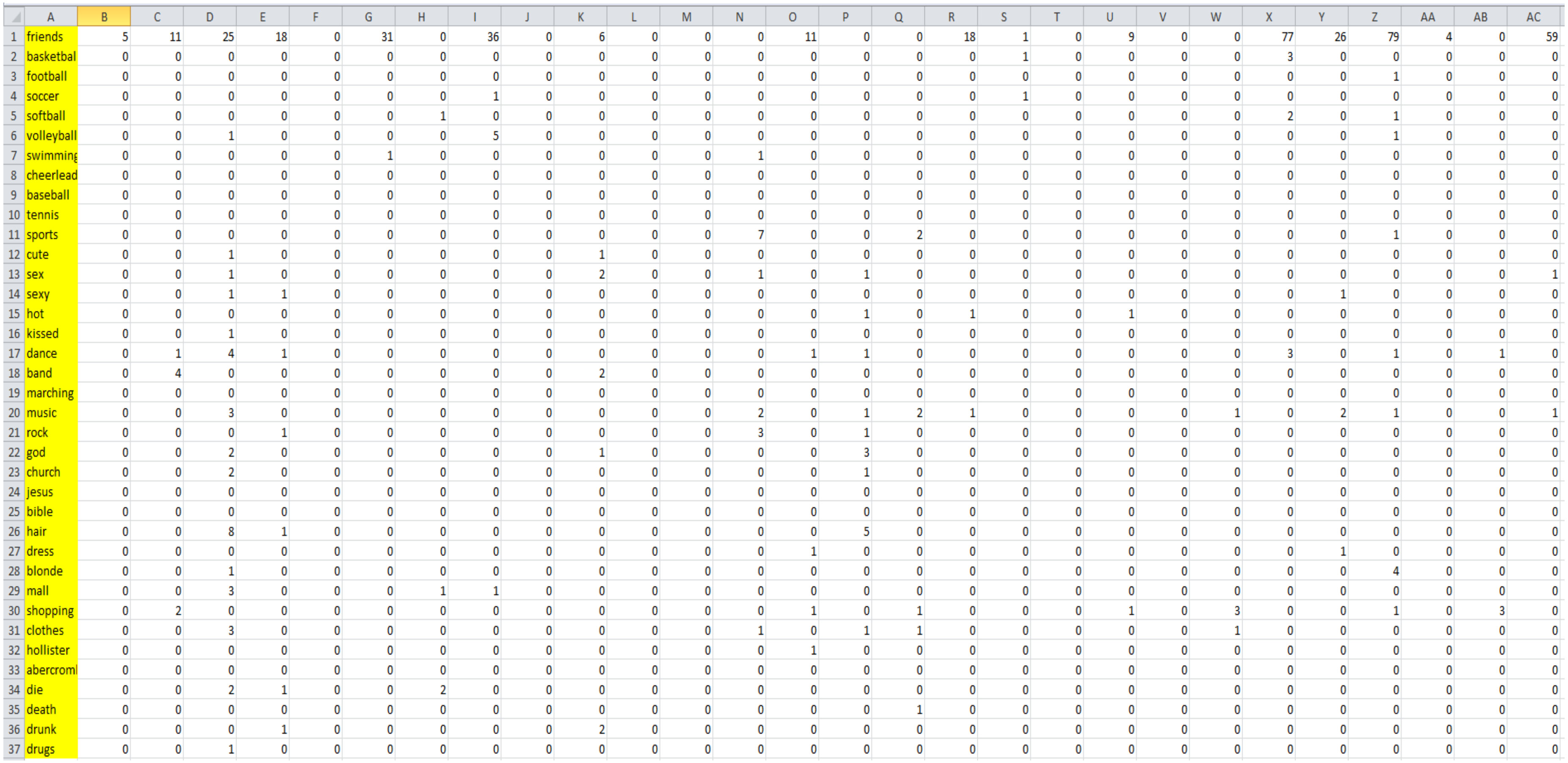
  - Run a hierarchical clustering model
    - With Ward’s distance



```
teen_HC <- hclust(distance, method = "ward.D")
```

Step5: Train a Model on the Data

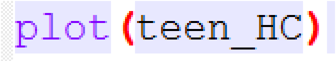
- Transposed Matrix
  - Rows: Instances (SNS Keywords)
  - Columns: Variables (Teens)



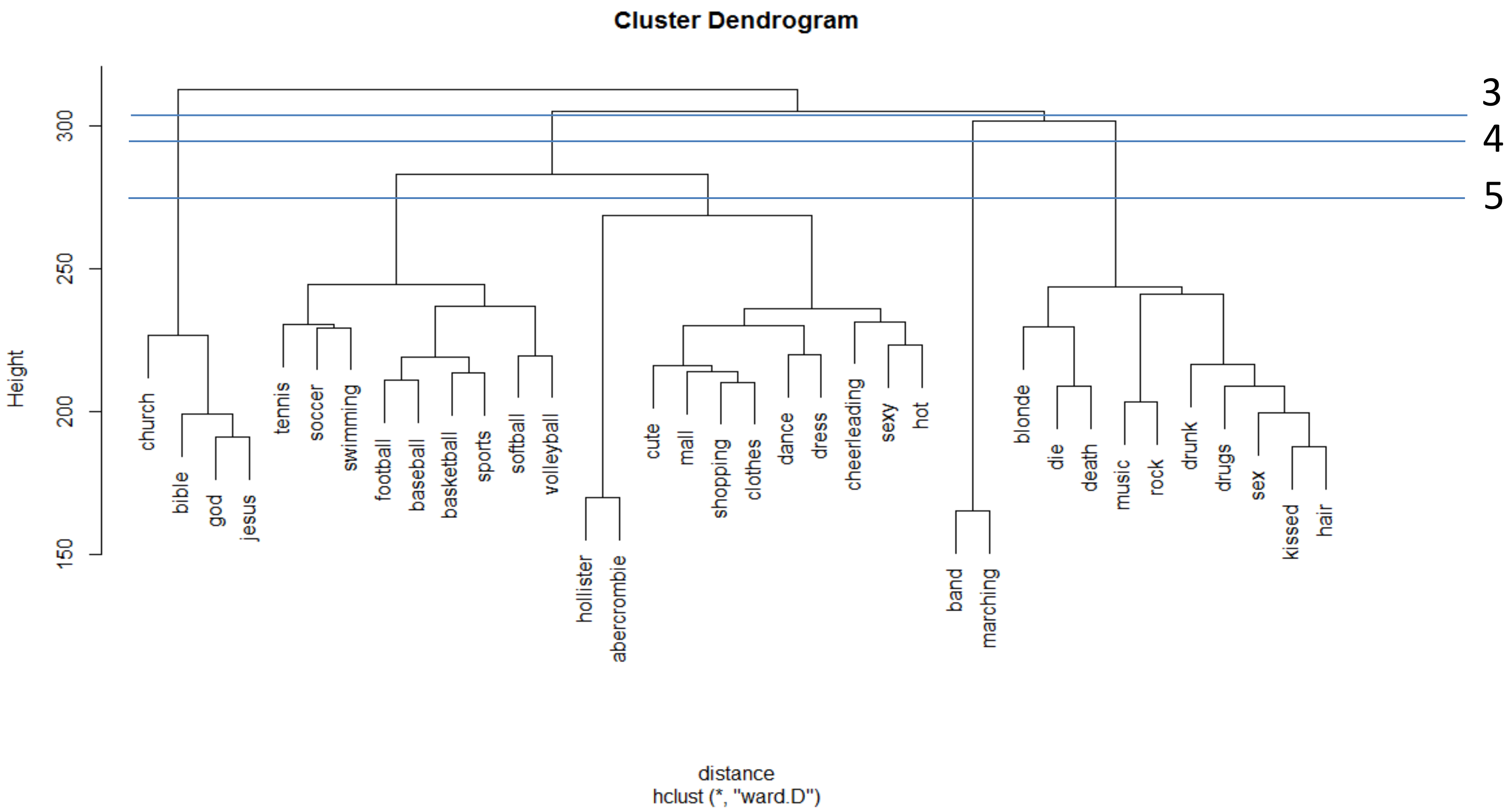
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	friends	5	11	25	18	0	31	0	36	0	6	0	0	0	11	0	0	18	1	0	9	0	0	77	26	79	4	0	59
2	basketbal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0
3	football	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4	soccer	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	softball	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0
6	volleyball	0	0	1	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7	swimming	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	cheerlead	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	baseball	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	tennis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	sports	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0
12	cute	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	sex	0	0	1	0	0	0	0	0	0	2	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
14	sexy	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
15	hot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0
16	kissed	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	dance	0	1	4	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	3	0	1	0	1	0
18	band	0	4	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	marching	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	music	0	0	3	0	0	0	0	0	0	0	0	0	2	0	1	2	1	0	0	0	0	1	0	2	1	0	0	1
21	rock	0	0	0	1	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
22	god	0	0	2	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
23	church	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
24	jesus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	bible	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	hair	0	0	8	1	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
27	dress	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
28	blonde	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
29	mall	0	0	3	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	shopping	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	3	0	0	1	0	3	0
31	clothes	0	0	3	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
32	hollister	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	abercroml	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	die	0	0	2	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	death	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
36	drunk	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	drugs	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Step6: Validate Clusters

- Display the clustering output in a dendrogram

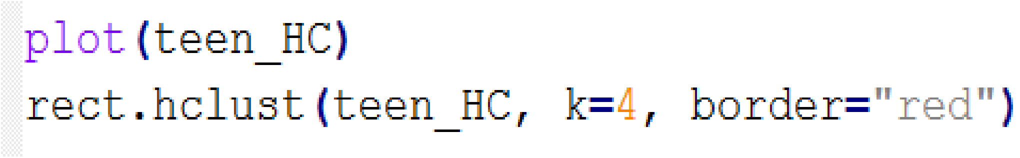


```
plot(teen_HC)
```

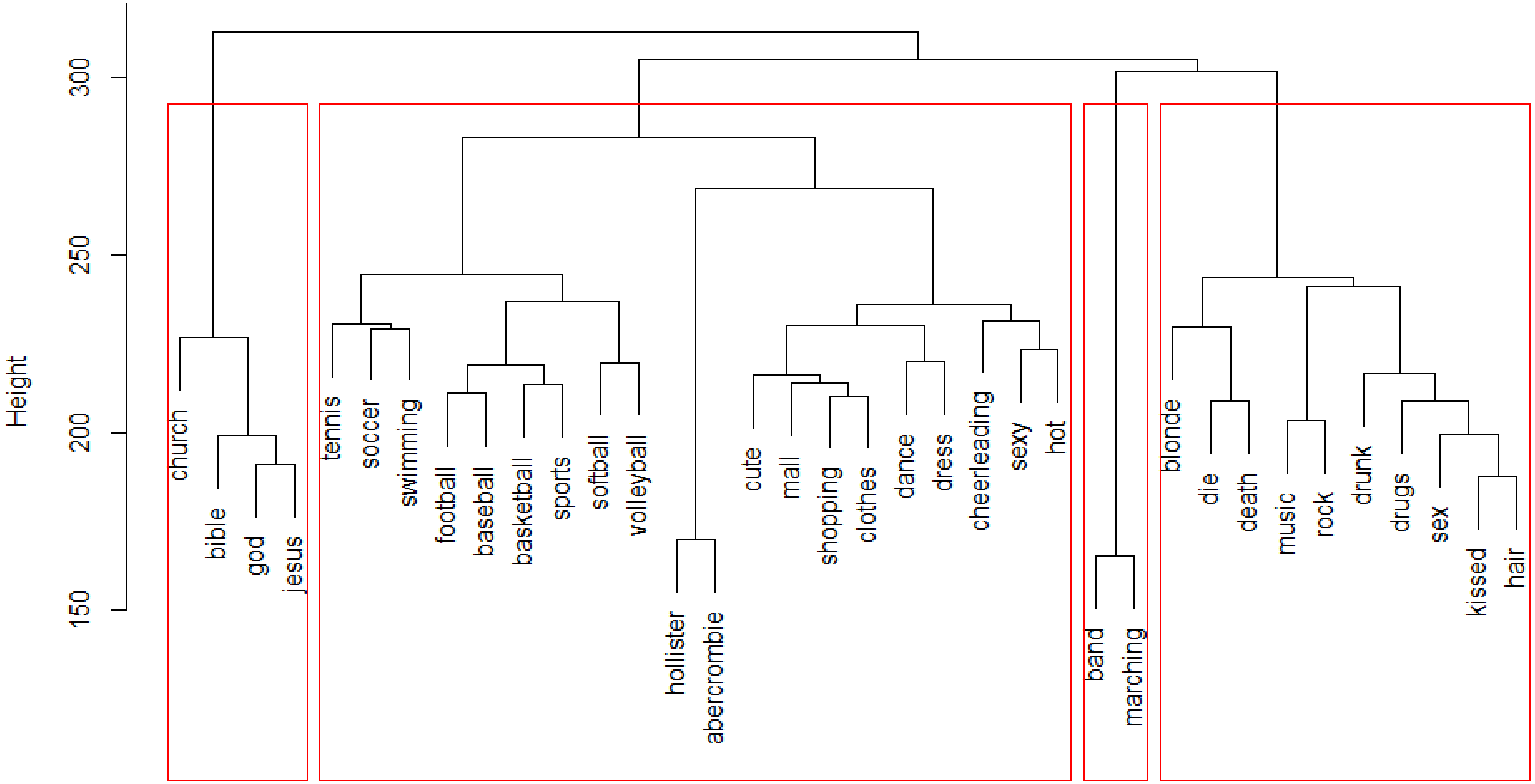


Step6: Validate Clusters

- With 4 Clusters

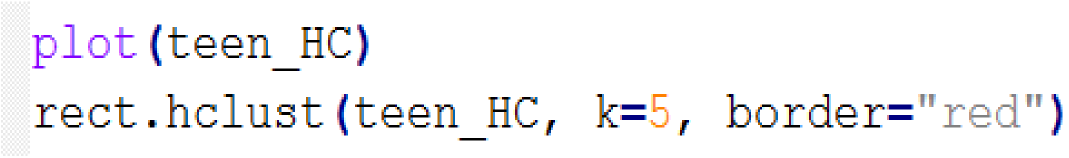


```
plot(teen_HC)
rect.hclust(teen_HC, k=4, border="red")
```

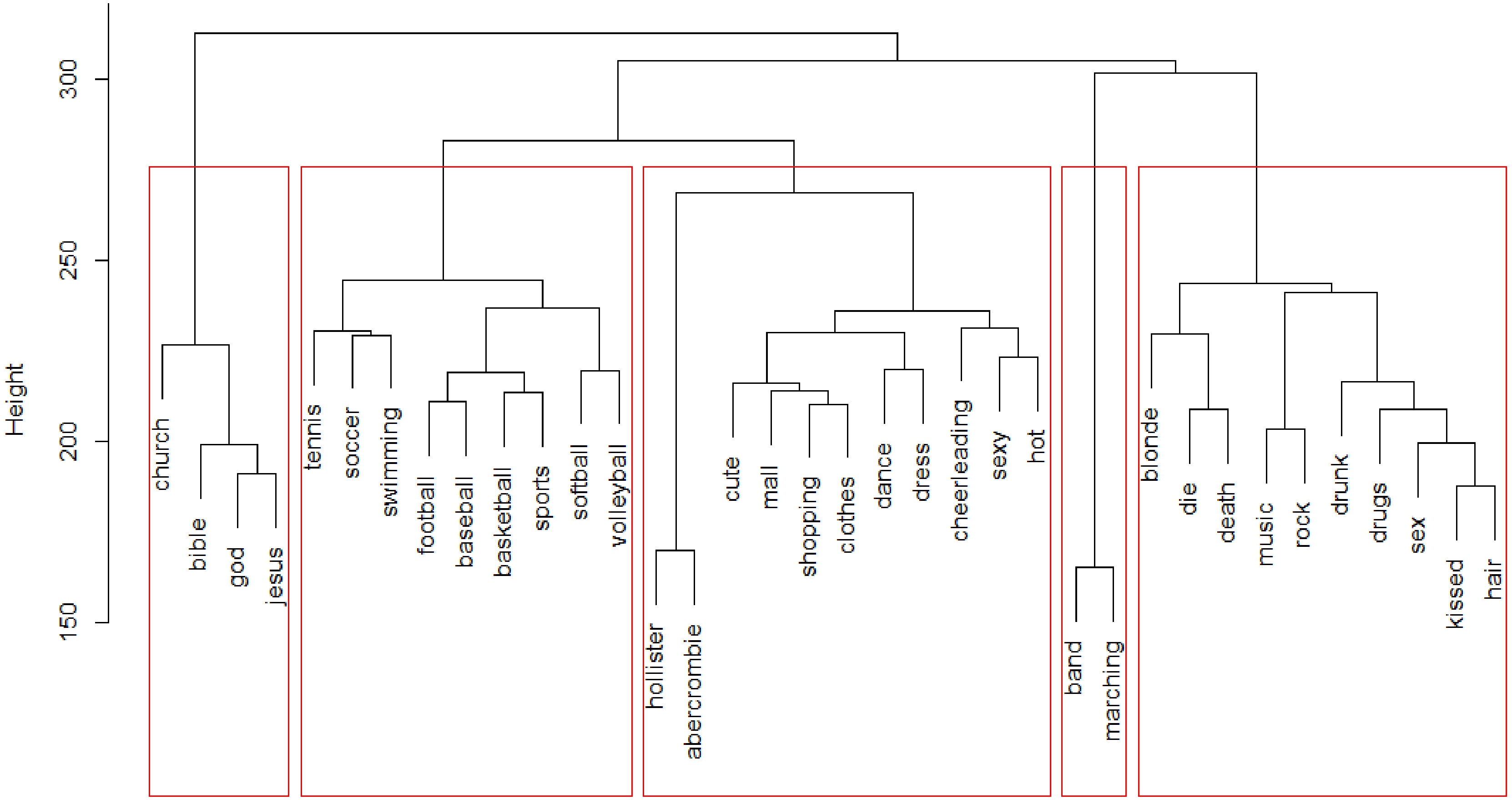


Step6: Validate Clusters

- With 5 Clusters

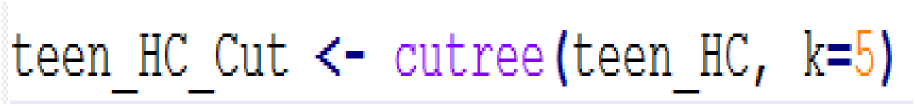


```
plot(teen_HC)
rect.hclust(teen_HC, k=5, border="red")
```



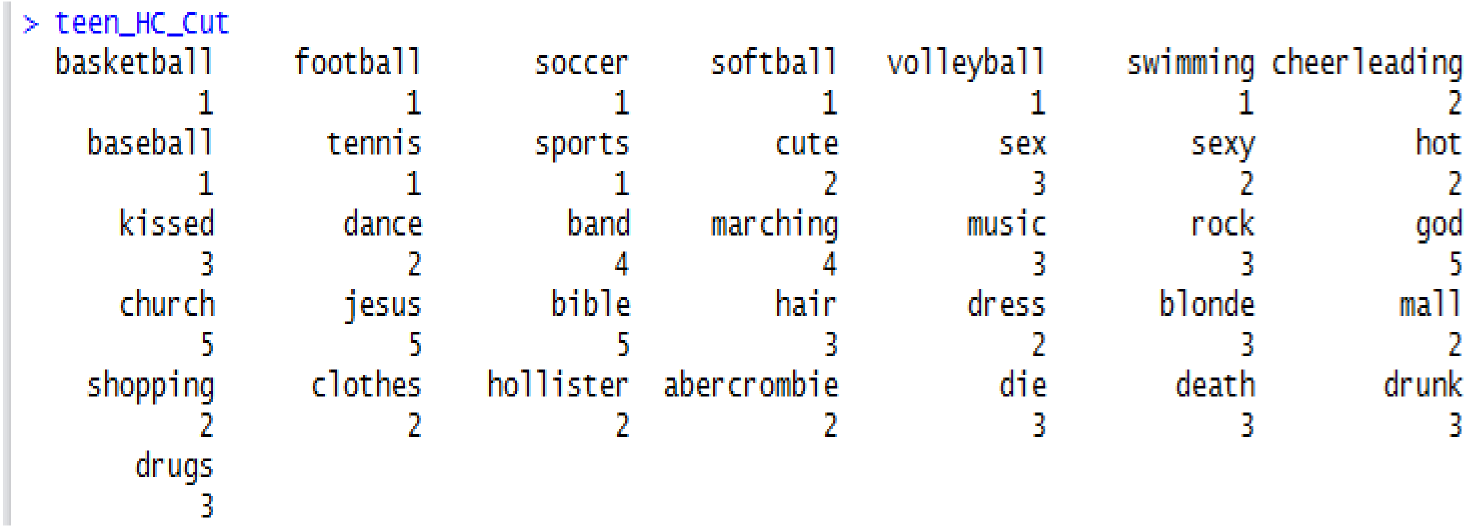
Step6: Validate Clusters

- Cut off the tree at the desired number of clusters (k=5)



```r
teen_HC_Cut <- cutree(teen_HC, k=5)
```

- Examine the clusters for SNS keywords



```
> teen_HC_Cut
  basketball     football       soccer     softball   volleyball     swimming cheerleading 
           1            1            1            1            1            1            2 
    baseball       tennis       sports         cute          sex         sexy          hot 
           1            1            1            2            3            2            2 
      kissed        dance         band     marching        music         rock          god 
           3            2            4            4            3            3            5 
      church        jesus        bible         hair        dress       blonde         mall 
           5            5            5            3            2            3            2 
    shopping      clothes    hollister  abercrombie          die        death        drunk 
           2            2            2            2            3            3            3 
       drugs 
           3 
```

Step6: Validate Clusters

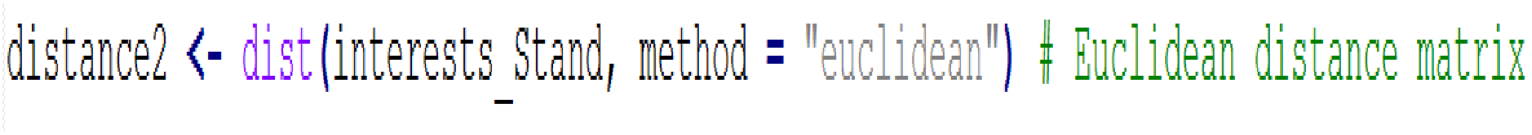
- Characterize the Clusters
  - Expected Clusters: Extracurricular activities, Fashion, Religion, Romance, and Antisocial behavior
  - Output Clusters:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Basketball Football Soccer Softball Volleyball Swimming Baseball Tennis Sports	Cheer leading Cute Sexy Hot Dance Shopping Clothes Hollister Abercrombie	Sex Kissed Music Rock Hair Blonde Die Death Drunk Drugs	Band Marching	God Church Jesus Bible
E.A. 1	Fashion	Anti-Social	E.A. 2	Religion

Case 2: Identify the Clusters of Teens

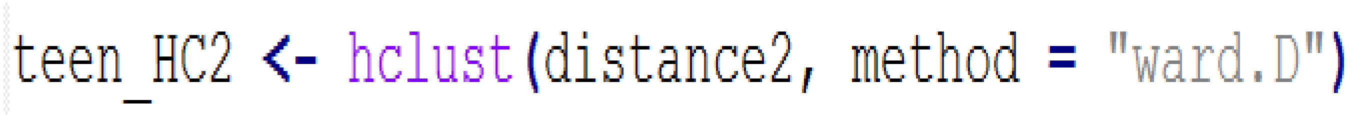
Step5: Train a Model on the Data

- Hierarchical Clustering
  - Measure distance between records
    - Use the Euclidean distance as an input for clustering
    - As we want to cluster the teenagers, the interests_Stand matrix doesn't need to be transposed.



```
distance2 <- dist(interests_Stand, method = "euclidean") # Euclidean distance matrix
```

  - Run a hierarchical clustering model
    - With Ward’s distance

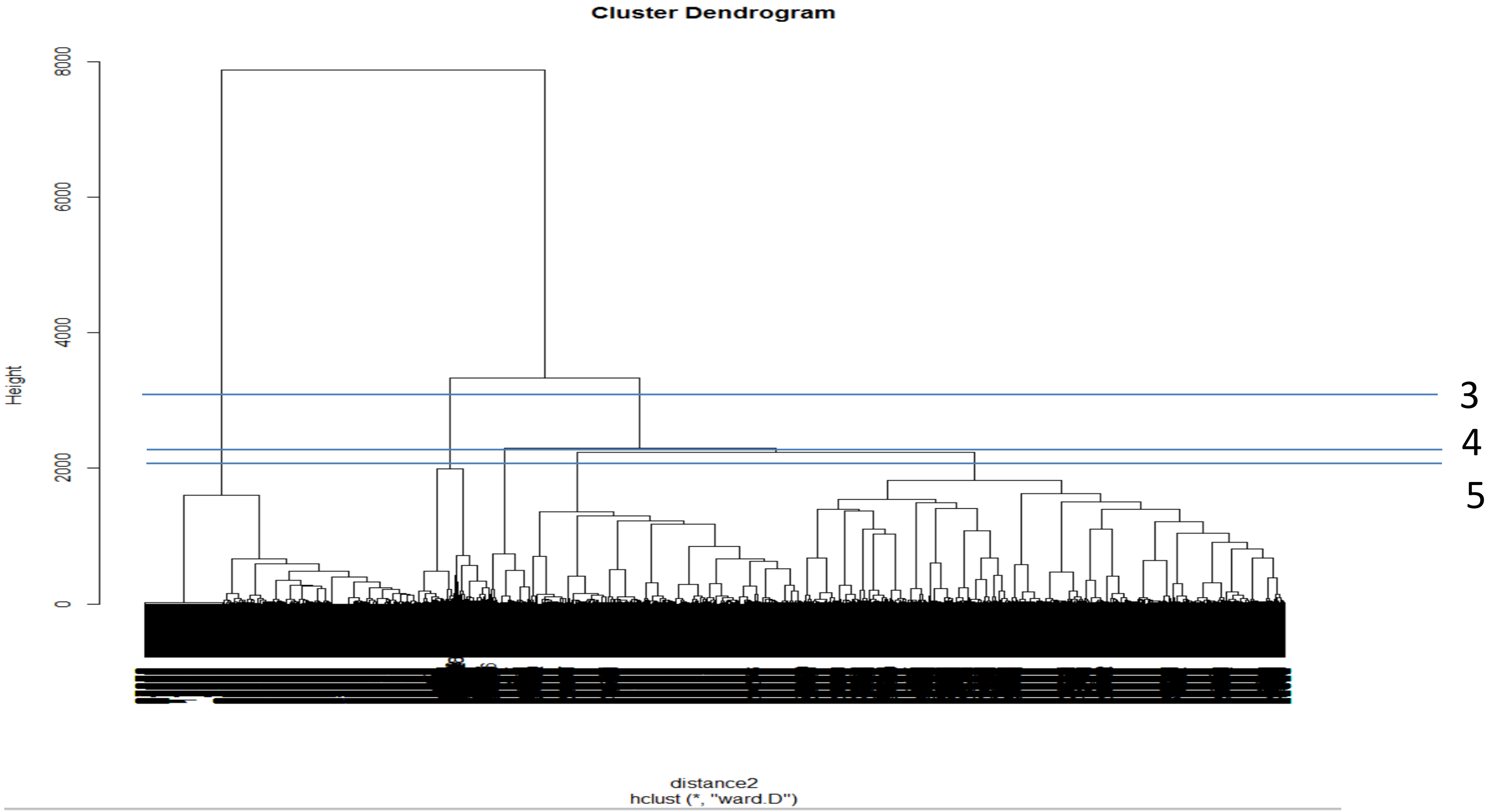


```
teen_HC2 <- hclust(distance2, method = "ward.D")
```

Step6: Validate Clusters

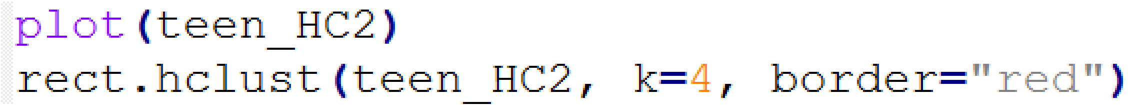
- Display the clustering output in a dendrogram

```
plot(teen_HC2)
```

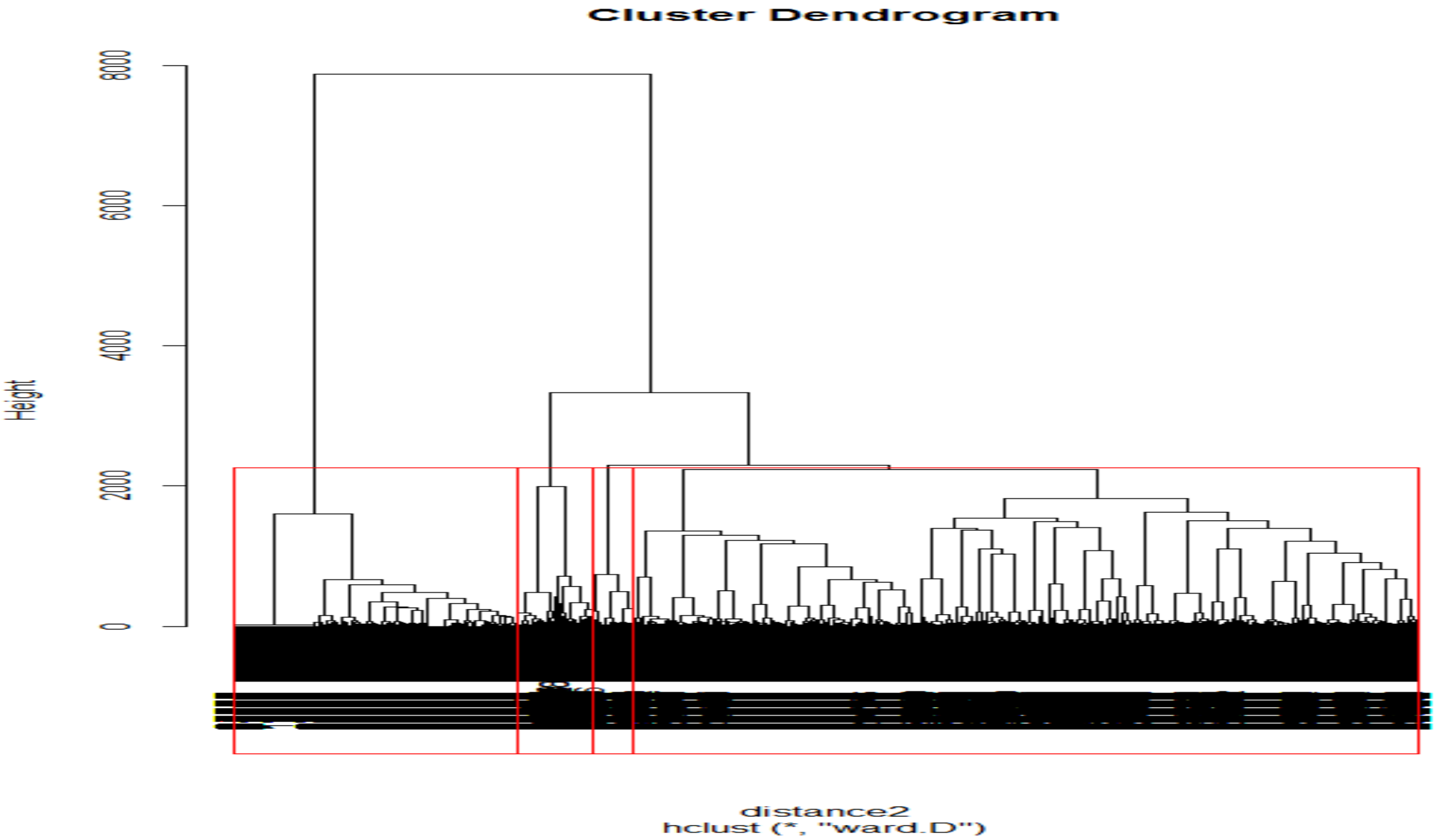


Step6: Validate Clusters

- With 4 Clusters

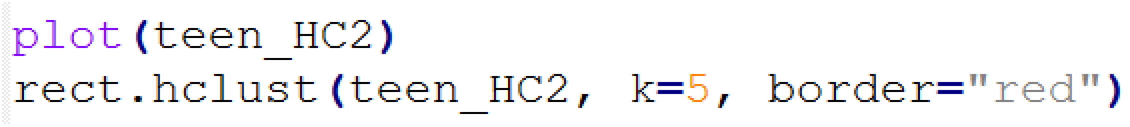


```
plot(teen_HC2)
rect.hclust(teen_HC2, k=4, border="red")
```

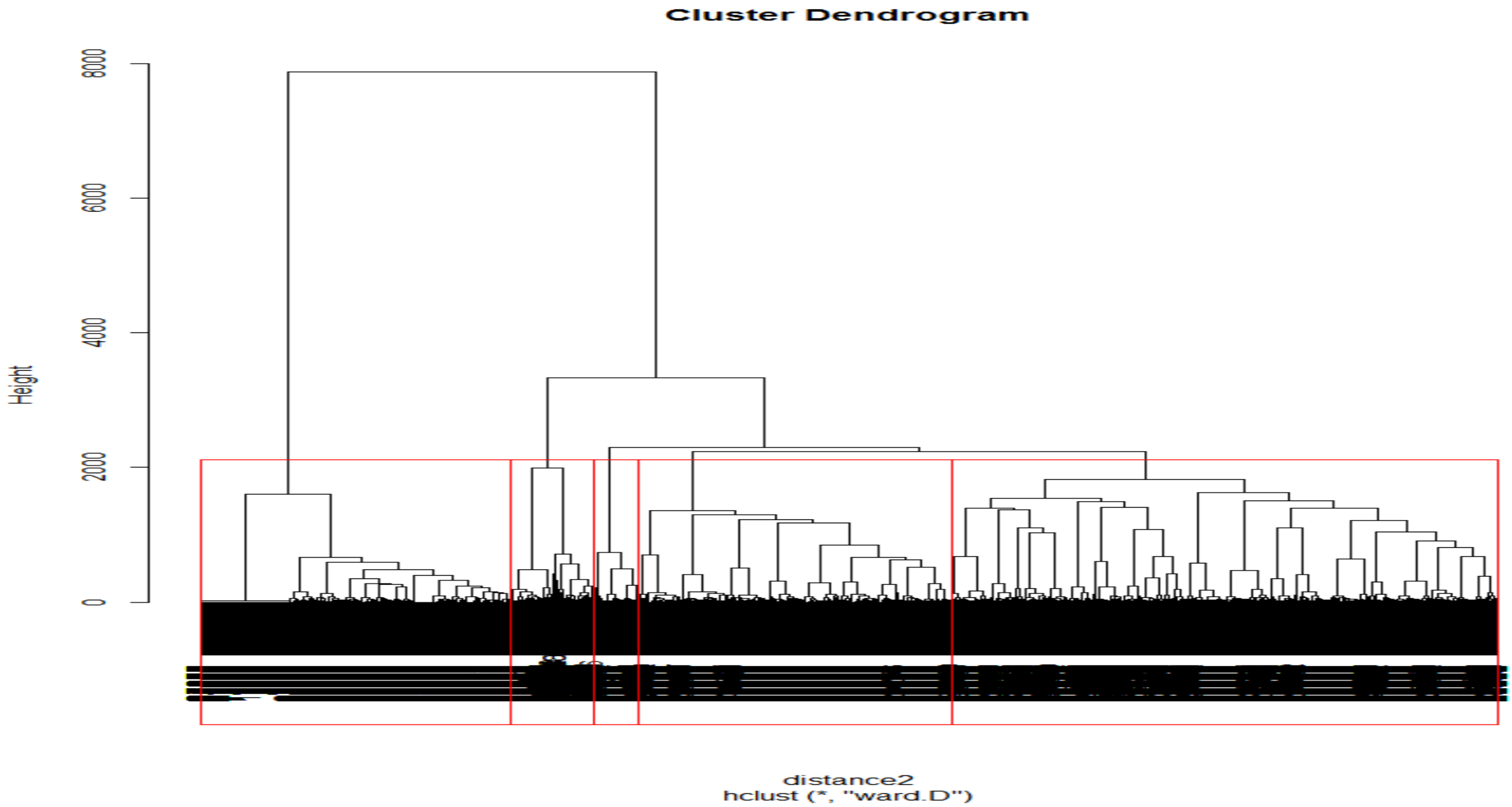


Step6: Validate Clusters

- With 5 Clusters

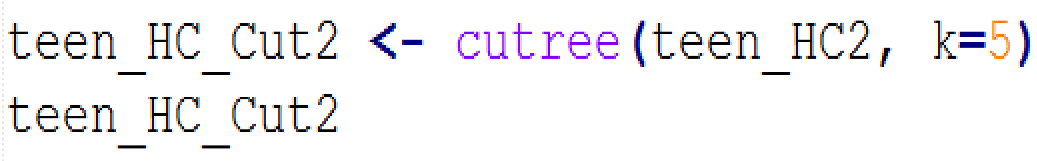


```
plot(teen_HC2)
rect.hclust(teen_HC2, k=5, border="red")
```



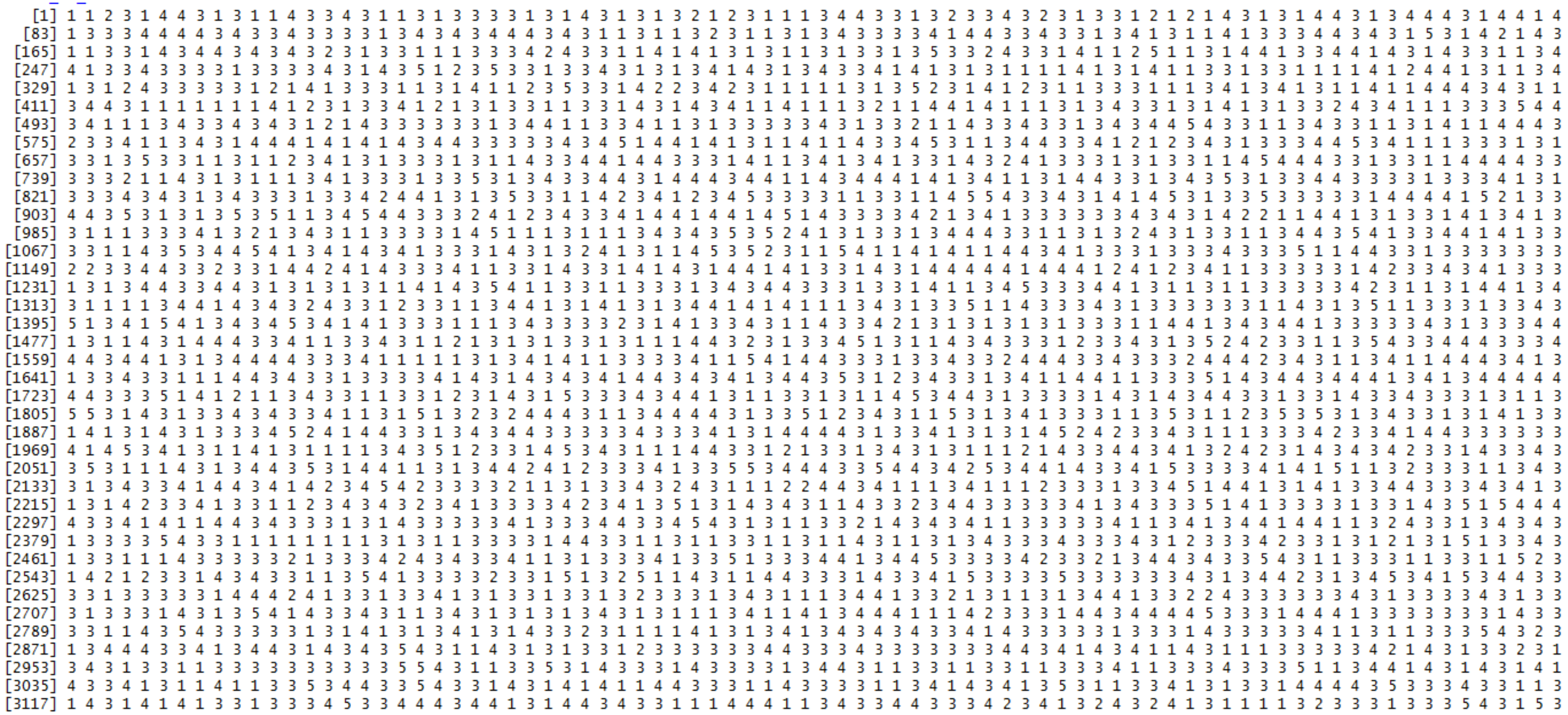
Step6: Validate Clusters

- Cut off the tree at the desired number of clusters (k=5)



```
teen_HC_Cut2 <- cutree(teen_HC2, k=5)
teen_HC_Cut2
```

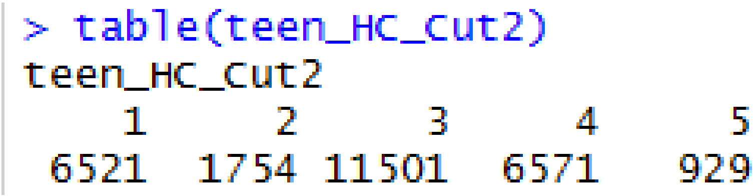
- Examine the clusters for teens



```
   [1] 1 1 2 3 1 4 4 3 1 3 1 1 4 3 3 4 3 1 1 3 1 3 3 3 3 1 3 1 4 3 1 3 1 3 2 1 2 3 1 1 1 3 4 4 3 3 1 3 2 3 3 4 3 2 3 1 3 3 1 2 1 2 1 4 3 1 3 1 4 4 3 1 3 4 4 4 3 1 4 4 1 4
  [83] 1 3 3 3 4 4 4 4 3 4 3 3 4 3 3 3 3 1 3 4 3 4 3 4 4 4 3 4 3 1 1 3 1 1 3 2 3 1 1 3 1 3 4 3 3 3 3 4 1 4 4 3 3 4 3 3 1 3 4 1 3 1 1 4 1 3 3 3 4 4 3 4 3 1 5 3 1 4 2 1 4 3
 [165] 1 1 3 3 1 4 3 4 4 3 4 3 4 3 2 3 1 3 3 1 1 1 3 3 3 4 2 4 3 3 1 1 4 1 4 1 3 1 3 1 1 3 1 3 3 1 3 5 3 3 2 4 3 3 1 4 1 1 2 5 1 1 3 1 4 4 1 3 3 4 4 1 4 3 1 4 3 3 1 1 3 4
 [247] 4 1 3 3 4 3 3 3 3 1 3 3 3 3 4 3 1 4 3 5 1 2 3 5 3 3 1 3 3 4 3 1 3 1 3 4 1 4 3 1 3 4 3 3 4 1 4 1 3 1 3 1 1 1 1 4 1 3 1 4 1 1 3 3 1 3 3 1 1 1 1 4 1 2 4 4 1 3 1 1 3 4
 [329] 1 3 1 2 4 3 3 3 3 3 1 2 1 4 1 3 3 3 1 1 3 1 4 1 1 2 3 5 3 3 1 4 2 2 3 4 2 3 1 1 1 1 1 3 1 3 5 2 3 1 4 1 2 3 1 1 3 3 3 1 1 1 3 4 1 3 4 1 3 1 1 4 1 1 4 4 4 3 4 3 1 1
 [411] 3 4 4 3 1 1 1 1 1 1 1 4 1 2 3 1 3 3 4 1 2 1 3 1 3 3 1 1 3 3 1 4 3 1 4 3 4 1 1 4 1 1 1 3 2 1 1 4 4 1 4 1 1 1 3 1 3 4 3 3 1 3 1 4 1 3 1 3 3 2 4 3 4 1 1 1 3 3 3 5 4 4
 [493] 3 4 1 1 1 3 4 3 3 4 3 4 3 1 2 1 4 3 3 3 3 3 3 1 3 4 4 1 1 3 3 4 1 1 3 1 3 3 3 3 3 4 3 1 3 3 2 1 1 4 3 3 4 3 3 1 3 4 3 4 4 5 4 3 3 1 1 3 4 3 3 1 1 3 1 4 1 1 4 4 4 3
 [575] 2 3 3 4 1 1 3 4 3 1 4 4 4 1 4 1 4 1 4 3 4 4 3 3 3 3 3 4 3 4 5 1 4 4 1 4 1 3 1 1 4 1 1 4 3 3 4 5 3 1 1 3 4 4 3 3 4 1 2 1 2 3 4 3 1 3 3 3 4 4 5 3 4 1 1 1 3 3 3 1 3 1
 [657] 3 3 1 3 5 3 3 1 1 3 1 1 2 3 4 1 3 1 3 3 3 1 3 1 1 4 3 3 4 4 1 4 4 3 3 3 1 4 1 1 3 4 1 3 4 1 3 3 1 4 3 2 4 1 3 3 3 1 3 1 3 3 1 1 4 5 4 4 4 3 3 1 3 3 1 1 4 4 4 4 3 3
 [739] 3 3 3 2 1 1 4 3 1 3 1 1 1 3 4 1 3 3 3 1 3 3 5 3 1 3 4 3 3 4 4 3 1 4 4 4 3 4 4 1 1 4 3 4 4 4 1 4 1 3 4 1 1 3 1 4 4 3 3 1 3 4 3 5 3 1 3 3 4 4 3 3 3 3 1 3 3 3 4 1 3 1
 [821] 3 3 3 4 3 4 3 1 3 4 3 3 3 1 3 3 4 2 4 4 1 3 1 3 5 3 3 1 1 4 2 3 4 1 2 3 4 5 3 3 3 3 1 1 3 3 1 1 4 5 5 4 3 3 4 3 1 4 1 4 5 3 1 3 3 5 3 3 3 3 3 1 4 4 4 4 1 5 2 1 3 3
 [903] 4 4 3 5 3 1 3 1 3 5 3 5 1 1 3 4 5 4 4 3 3 3 2 4 1 2 3 4 3 3 4 1 4 4 1 4 4 1 4 5 1 4 3 3 3 3 4 2 1 3 4 1 3 3 3 3 3 3 4 3 4 3 1 4 2 2 1 1 4 4 1 3 1 3 3 1 4 1 3 4 1 3
 [985] 3 1 1 1 3 3 3 4 1 3 2 1 3 4 3 1 1 3 3 3 3 1 4 5 1 1 1 3 1 1 1 3 4 3 4 3 5 3 5 2 4 1 3 1 3 3 1 3 4 4 4 3 3 1 1 3 1 3 2 4 3 1 3 3 1 1 3 4 4 3 5 4 1 3 3 4 4 1 4 1 3 3
[1067] 3 3 1 1 4 3 5 3 4 4 5 4 1 3 4 1 4 3 4 1 3 3 3 1 4 3 1 3 2 4 1 3 1 1 4 5 3 5 2 3 1 1 5 4 1 1 4 1 4 1 1 4 4 3 4 1 3 3 3 1 3 3 3 4 3 3 3 5 1 1 4 4 3 3 1 3 3 3 3 3 3 3
[1149] 2 2 3 3 4 4 3 3 2 3 3 1 4 4 2 4 1 4 3 3 3 4 1 1 3 3 1 4 3 3 1 4 1 4 3 1 4 4 1 4 1 3 3 1 4 3 1 4 4 4 4 4 1 4 4 4 1 2 4 1 2 3 4 1 1 3 3 3 3 3 1 4 2 3 3 4 3 4 1 3 3 3
[1231] 1 3 1 3 4 4 3 3 4 4 3 1 3 1 3 1 3 1 1 4 1 4 3 5 4 1 1 3 3 1 1 3 3 3 1 3 4 3 4 4 3 3 3 1 3 3 1 4 1 1 3 4 5 3 3 3 4 4 1 3 1 1 3 1 1 3 3 3 3 3 4 2 3 1 1 3 1 4 4 1 3 4
[1313] 3 1 1 1 1 3 4 4 1 4 3 4 3 2 4 3 3 1 2 3 3 1 1 3 4 4 1 3 1 4 1 3 1 3 4 4 1 4 1 4 1 1 1 3 4 3 1 3 3 5 1 1 4 3 3 3 4 3 1 3 3 3 3 3 3 1 1 4 3 1 3 5 1 1 3 3 3 1 3 3 4 3
[1395] 5 1 3 4 1 5 4 1 3 4 3 4 5 3 4 1 4 1 3 3 3 1 1 1 3 4 3 3 3 3 2 3 1 4 1 3 3 4 3 1 1 4 3 3 4 2 1 3 1 3 1 3 1 3 1 3 3 3 1 1 4 4 1 3 4 3 4 4 1 3 3 3 3 3 4 3 1 3 3 3 4 4
[1477] 1 3 1 1 4 3 1 4 4 4 3 3 4 1 1 3 3 4 3 1 1 2 1 3 1 3 1 3 3 1 3 1 1 1 4 4 3 2 3 1 3 3 4 5 1 3 1 1 4 3 4 3 3 3 1 2 3 3 4 3 1 3 5 2 4 2 3 3 1 1 3 5 4 3 3 4 4 4 3 3 3 4
[1559] 4 4 3 4 4 1 3 1 3 4 4 4 4 3 3 3 4 1 1 1 1 1 3 1 3 4 1 4 1 1 3 3 3 3 4 1 1 5 4 1 4 4 3 3 3 1 3 3 4 3 3 2 4 4 4 3 3 4 3 3 3 2 4 4 4 2 3 4 3 1 1 3 4 1 1 4 4 4 3 4 1 3
[1641] 1 3 3 4 3 3 1 1 1 4 4 3 4 3 3 1 3 3 3 3 4 1 4 3 1 4 3 4 3 4 1 4 4 3 4 3 4 1 3 4 4 3 5 3 1 2 3 4 3 3 1 3 4 1 1 4 4 1 1 3 3 3 5 1 4 3 4 4 3 4 4 4 1 3 4 1 3 4 4 4 4 4
[1723] 4 4 3 3 3 5 1 4 1 2 1 1 3 4 3 3 1 1 3 3 1 2 3 1 4 3 1 5 3 3 3 4 3 4 4 1 3 1 1 3 3 1 3 1 1 4 5 3 4 4 3 1 3 3 3 3 1 4 3 1 4 3 4 4 3 3 1 3 3 1 4 3 3 4 3 3 3 1 3 1 1 3
[1805] 5 5 3 1 4 3 1 3 3 4 3 4 3 3 4 1 1 3 1 5 1 3 2 3 2 4 4 4 3 1 1 3 4 4 4 4 3 1 3 3 5 1 2 3 4 3 1 1 5 3 1 3 4 1 3 3 3 1 1 3 5 3 1 1 2 3 5 3 5 3 1 3 4 3 3 1 3 1 4 1 3 3
[1887] 1 4 1 3 1 4 3 1 3 3 3 4 5 2 4 1 4 4 3 3 1 3 4 3 4 4 3 3 3 3 3 4 3 3 3 4 1 3 1 4 4 4 4 3 1 3 3 4 1 3 1 3 1 4 5 2 4 2 3 3 4 3 1 1 1 3 3 3 4 2 3 3 4 1 4 4 3 3 3 3 3 3
[1969] 4 1 4 5 3 4 1 3 1 1 4 1 3 1 1 1 1 3 4 3 5 1 2 3 3 1 4 5 3 4 3 1 1 1 4 4 3 3 1 2 1 3 3 1 3 4 3 1 3 1 1 1 2 1 4 3 3 4 4 3 4 1 3 2 4 2 3 1 4 3 4 3 4 2 3 3 1 4 3 3 4 3
[2051] 3 5 3 1 1 1 4 3 1 3 4 4 3 5 3 1 4 4 1 1 3 1 3 4 4 2 4 1 2 3 3 3 4 1 3 3 5 5 3 4 4 4 3 3 5 4 4 3 4 2 5 3 4 4 1 4 3 3 4 1 5 3 3 3 3 4 1 4 1 5 1 1 3 2 3 3 3 1 1 3 4 3
[2133] 3 1 3 4 3 3 4 1 4 4 3 4 1 4 2 3 4 5 4 2 3 3 3 3 2 1 1 3 1 3 3 4 3 2 4 3 1 1 1 2 2 4 4 3 4 1 1 1 3 4 1 1 1 2 3 3 3 1 3 3 4 5 1 4 4 1 3 1 4 1 3 3 4 4 3 3 3 4 3 4 1 3
[2215] 1 3 1 4 2 3 3 4 1 3 3 1 1 2 3 4 3 4 3 2 3 4 1 3 3 3 3 4 2 3 4 1 3 5 1 3 1 4 3 4 3 1 1 4 3 3 2 3 4 4 3 3 3 3 3 4 1 3 4 3 3 3 5 1 4 1 3 3 3 3 1 3 3 1 4 3 5 1 5 4 4 4
[2297] 4 3 3 4 1 4 1 1 4 4 3 4 3 3 3 1 3 1 4 3 3 3 3 3 4 1 3 3 3 4 4 3 3 4 5 4 3 1 3 1 1 3 3 2 1 4 3 4 3 4 1 1 3 3 3 3 3 4 1 1 3 4 1 3 4 4 1 4 4 1 1 3 2 4 3 3 1 3 4 3 4 3
[2379] 1 3 3 3 3 5 4 3 3 1 1 1 1 1 1 1 1 3 1 3 1 1 3 3 3 3 1 4 4 3 3 1 1 3 1 1 3 3 1 1 3 1 1 4 3 1 1 3 1 3 4 3 3 3 4 3 3 3 4 3 1 2 3 3 3 4 2 3 3 1 3 1 2 1 3 1 5 1 3 3 4 3
[2461] 1 3 3 1 1 1 4 3 3 3 3 3 2 1 3 3 3 4 2 4 3 4 3 3 4 1 1 3 1 3 3 3 4 1 3 3 5 1 3 3 3 4 4 1 3 4 4 5 3 3 3 3 4 2 3 3 2 1 3 4 4 3 4 3 3 5 4 3 1 1 3 3 3 1 1 3 3 1 1 5 2 3
[2543] 1 4 2 1 2 3 3 1 4 3 4 3 3 1 1 3 5 4 1 3 3 3 3 2 3 3 1 5 1 3 2 5 1 1 4 3 1 1 4 4 3 3 3 1 4 3 3 4 1 5 3 3 3 3 5 3 3 3 3 3 3 4 3 1 3 4 4 2 3 1 3 4 5 3 4 1 5 3 4 4 3 3
[2625] 3 3 1 3 3 3 3 3 1 4 4 4 2 4 1 3 3 1 3 3 4 1 3 1 3 3 1 3 3 1 3 2 3 3 3 1 3 4 3 1 1 1 3 4 4 1 3 3 2 1 3 1 1 3 1 3 4 4 1 3 3 2 2 4 3 3 3 3 3 3 4 3 1 3 3 3 3 4 3 1 3 3
[2707] 3 1 3 3 3 1 4 3 1 3 5 4 1 4 3 3 4 3 1 1 3 4 3 1 3 1 3 1 3 4 3 1 3 1 1 1 3 4 1 1 4 1 3 4 4 4 1 1 1 4 2 3 3 3 1 4 4 3 4 4 4 4 5 3 3 3 1 4 4 4 1 3 3 3 3 3 3 3 1 4 3 3
[2789] 3 3 1 1 4 3 5 4 3 3 3 3 3 1 3 1 4 1 3 1 3 4 1 3 1 4 3 3 2 3 1 1 1 1 4 1 3 1 3 4 1 3 4 3 4 3 4 3 3 4 1 4 3 3 3 3 3 1 3 3 3 1 4 3 3 3 3 3 4 1 1 3 1 1 3 3 3 5 4 3 2 3
[2871] 1 3 4 4 4 3 3 4 1 3 4 4 3 1 4 3 4 3 5 4 3 1 1 4 3 1 3 1 3 3 1 2 3 3 3 3 3 3 4 4 3 3 3 4 3 3 3 3 3 3 3 4 4 3 4 1 4 3 4 1 1 4 3 1 1 1 3 3 3 3 3 4 2 1 4 3 1 3 3 2 3 1
[2953] 3 4 3 1 3 3 1 1 3 3 3 3 3 3 3 3 3 3 5 5 4 3 1 1 3 3 5 3 1 4 3 3 3 1 4 3 3 3 3 1 3 4 4 3 1 1 3 3 1 1 3 3 1 1 3 3 3 4 1 1 3 3 3 4 3 3 3 5 1 1 3 4 4 1 4 3 1 4 3 1 4 1
[3035] 4 3 3 4 1 3 1 1 4 1 1 3 3 5 3 4 4 3 3 5 4 3 3 1 4 3 1 4 1 4 1 1 4 4 3 3 3 1 1 4 3 3 3 3 1 1 3 4 1 4 3 4 1 3 5 3 1 1 3 3 4 1 3 1 3 3 1 4 4 4 4 3 5 3 3 3 4 3 3 1 1 3
[3117] 1 4 3 1 4 1 4 1 3 3 1 3 3 3 4 5 3 3 4 4 4 3 4 4 1 3 1 4 4 3 4 3 3 1 1 1 4 4 4 1 1 3 4 3 3 4 4 3 3 3 4 2 3 4 1 3 2 4 3 2 4 1 3 1 1 1 1 3 2 3 3 3 1 3 3 3 5 4 3 1 5 3
```

Step6: Validate Clusters

- Characterize the Clusters
  - Expected Clusters: Extracurricular activities, Fashion, Religion, Romance, and Antisocial behavior
  - Output Clusters:



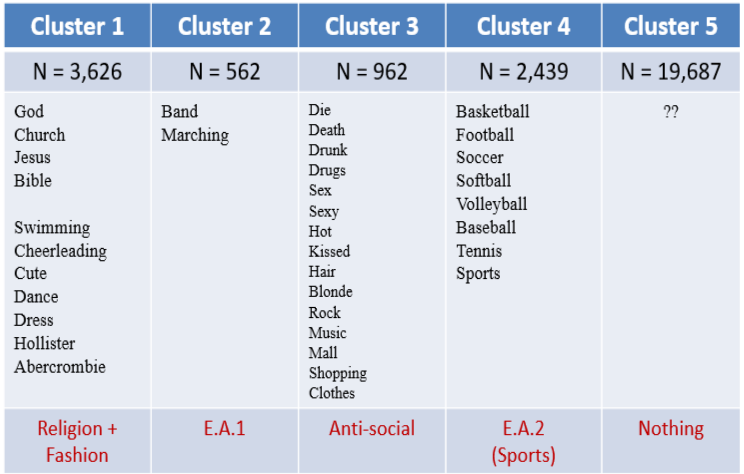
```
> table(teen_HC_Cut2)
teen_HC_Cut2
    1     2     3     4     5 
 6521  1754 11501  6571   929 
```

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
6,521	1,754	11,501	6,571	929
?	?	?	?	?

Step7: Evaluate the Business Problem(s)

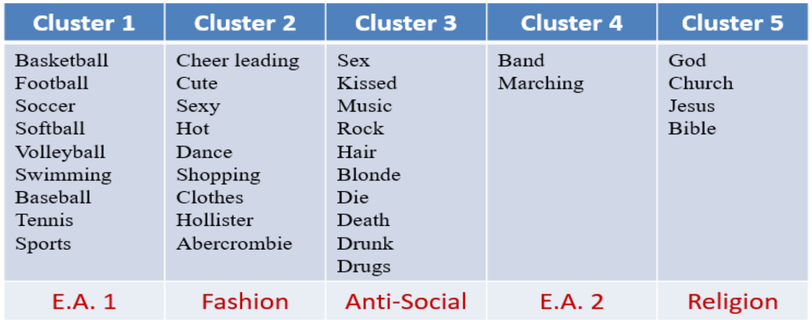
- Problem: Identify segments of teenagers who share similar tastes
  - Expected Clusters: Extracurricular activities, Fashion, Religion, Romance, and Antisocial behavior

K-means



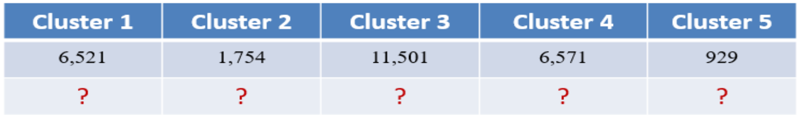
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
N = 3,626	N = 562	N = 962	N = 2,439	N = 19,687
God Church Jesus Bible Swimming Cheerleading Cute Dance Dress Hollister Abercrombie	Band Marching	Die Death Drunk Drugs Sex Sexy Hot Kissed Hair Blonde Rock Music Mall Shopping Clothes	Basketball Football Soccer Softball Volleyball Baseball Tennis Sports	??
Religion + Fashion	E.A.1	Anti-social	E.A.2 (Sports)	Nothing

HC 1



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Basketball Football Soccer Softball Volleyball Swimming Baseball Tennis Sports	Cheer leading Cute Sexy Hot Dance Shopping Clothes Hollister Abercrombie	Sex Kissed Music Rock Hair Blonde Die Death Drunk Drugs	Band Marching	God Church Jesus Bible
E.A. 1	Fashion	Anti-Social	E.A. 2	Religion

HC 2



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
6,521	1,754	11,501	6,571	929
?	?	?	?	?

  - Which one is better?
- What can we do with the clusters?
  - Given the clusters, a marketing manager would have a clear depiction of different types of teenage visitors to the SNS.
  - The manager could sell targeted advertising impressions to businesses with products relevant to one or more of the clusters.

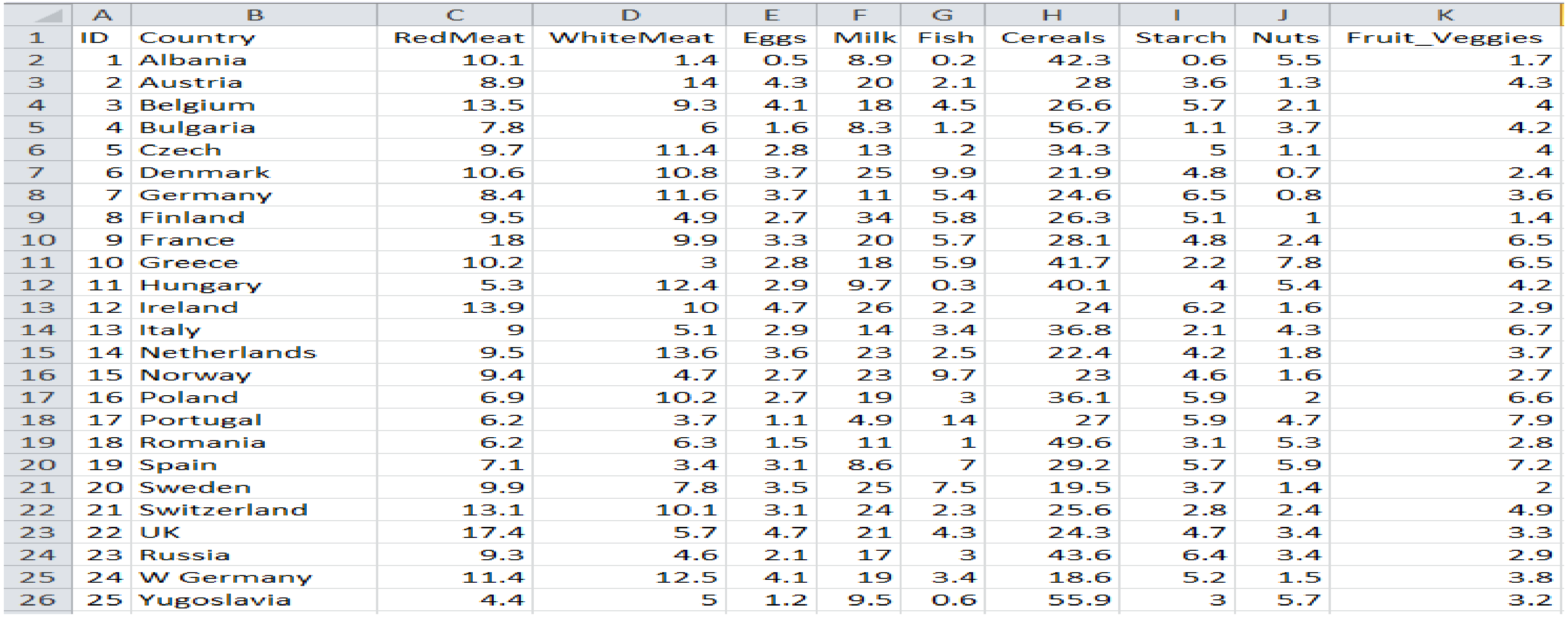
Individual Exercise Task (Class Participation)

Seminar 9

Instructor: Prof. Lee, Gun-woong
Nanyang Business School

Problem and Data

- Problem: Identify European Protein Consumption
  - Cluster Counties based their Protein Consumption Sources
- Your Task
  - Assume you are a sales manager at Walmart. What will you do with the clusters?
- Data
  - 25 European countries and their protein intakes (in %) from nine major food sources. The data is listed below
    - Download protein.csv from the Course Site (Seminar 9 – R Exercise)



	A	B	C	D	E	F	G	H	I	J	K
1	ID	Country	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fruit_Veggies
2	1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
3	2	Austria	8.9	14	4.3	20	2.1	28	3.6	1.3	4.3
4	3	Belgium	13.5	9.3	4.1	18	4.5	26.6	5.7	2.1	4
5	4	Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
6	5	Czech	9.7	11.4	2.8	13	2	34.3	5	1.1	4
7	6	Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
8	7	Germany	8.4	11.6	3.7	11	5.4	24.6	6.5	0.8	3.6
9	8	Finland	9.5	4.9	2.7	34	5.8	26.3	5.1	1	1.4
10	9	France	18	9.9	3.3	20	5.7	28.1	4.8	2.4	6.5
11	10	Greece	10.2	3	2.8	18	5.9	41.7	2.2	7.8	6.5
12	11	Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
13	12	Ireland	13.9	10	4.7	26	2.2	24	6.2	1.6	2.9
14	13	Italy	9	5.1	2.9	14	3.4	36.8	2.1	4.3	6.7
15	14	Netherlands	9.5	13.6	3.6	23	2.5	22.4	4.2	1.8	3.7
16	15	Norway	9.4	4.7	2.7	23	9.7	23	4.6	1.6	2.7
17	16	Poland	6.9	10.2	2.7	19	3	36.1	5.9	2	6.6
18	17	Portugal	6.2	3.7	1.1	4.9	14	27	5.9	4.7	7.9
19	18	Romania	6.2	6.3	1.5	11	1	49.6	3.1	5.3	2.8
20	19	Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
21	20	Sweden	9.9	7.8	3.5	25	7.5	19.5	3.7	1.4	2
22	21	Switzerland	13.1	10.1	3.1	24	2.3	25.6	2.8	2.4	4.9
23	22	UK	17.4	5.7	4.7	21	4.3	24.3	4.7	3.4	3.3
24	23	Russia	9.3	4.6	2.1	17	3	43.6	6.4	3.4	2.9
25	24	W Germany	11.4	12.5	4.1	19	3.4	18.6	5.2	1.5	3.8
26	25	Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Clustering

- Data Preparation
  - Use the nine variables of major food sources for clustering
  - Standardize the selected variables
- K-means Clustering
  - Run a K-means clustering
    - Cluster Countries based on their protein consumption sources
  - Summarize the clustering outcomes using the table below:

	Cluster 1	Cluster 2	…
#Instances	N=	N=	N =
Countries			
Centers (Foods)			

  - Conduct "Elbow Method". What is the appropriate value of k?

Clustering

- Hierarchical Clustering
  - Cluster Countries based on their protein consumption sources
  - Cluster Food Sources based on Countries
    - Measure the distance between Countries with Euclidean Distance
    - Measure the distance between Clusters with any distance measures
  - Cut tree into the desired number of clusters
  - Present the Dendrograms

Summarize the clustering outcomes using the tables below:

	Cluster 1	Cluster 2	…
#Instances	N=	N=	N =
Countries			

	Cluster 1	Cluster 2	…
#Instances	N=	N=	N =
Foods			

Clustering

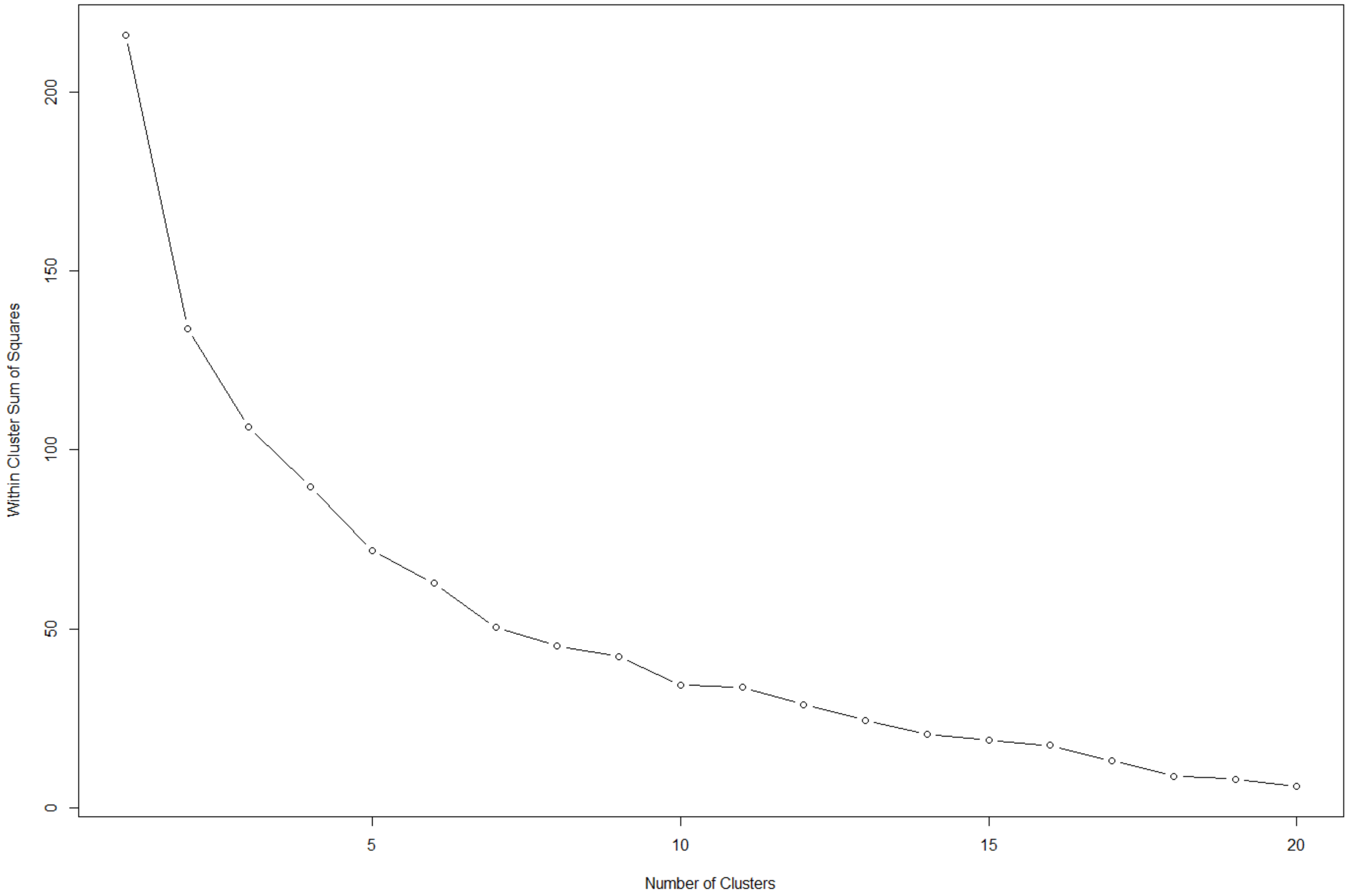
- Instructions (Due by Next Wednesday midnight)
  - Summarize your answers including the Three Cluster tables in two pages document.
    - The document should not exceed 3 pages (A4 size, single-spaced, Times New Roman 12-point font)
  - Describe your clustering approach
    - How do you determine the number of clusters (K)?
    - Which distance measure(s) is used for clustering (for hierarchical clustering)?
    - Interpret the clustering outcomes (i.e., Clusters)
  - Briefly explain what you can do with the clusters as a sales manager at Walmart.
  - Deliverables
    - R Code file: BC2406_Sxx_Gyy_NAME_Clustering.r
    - MS-Word document: BC2406_ Sxx-Gyy_NAME_Clustering.docx
  - Email the files to the TA ONLY

Sample

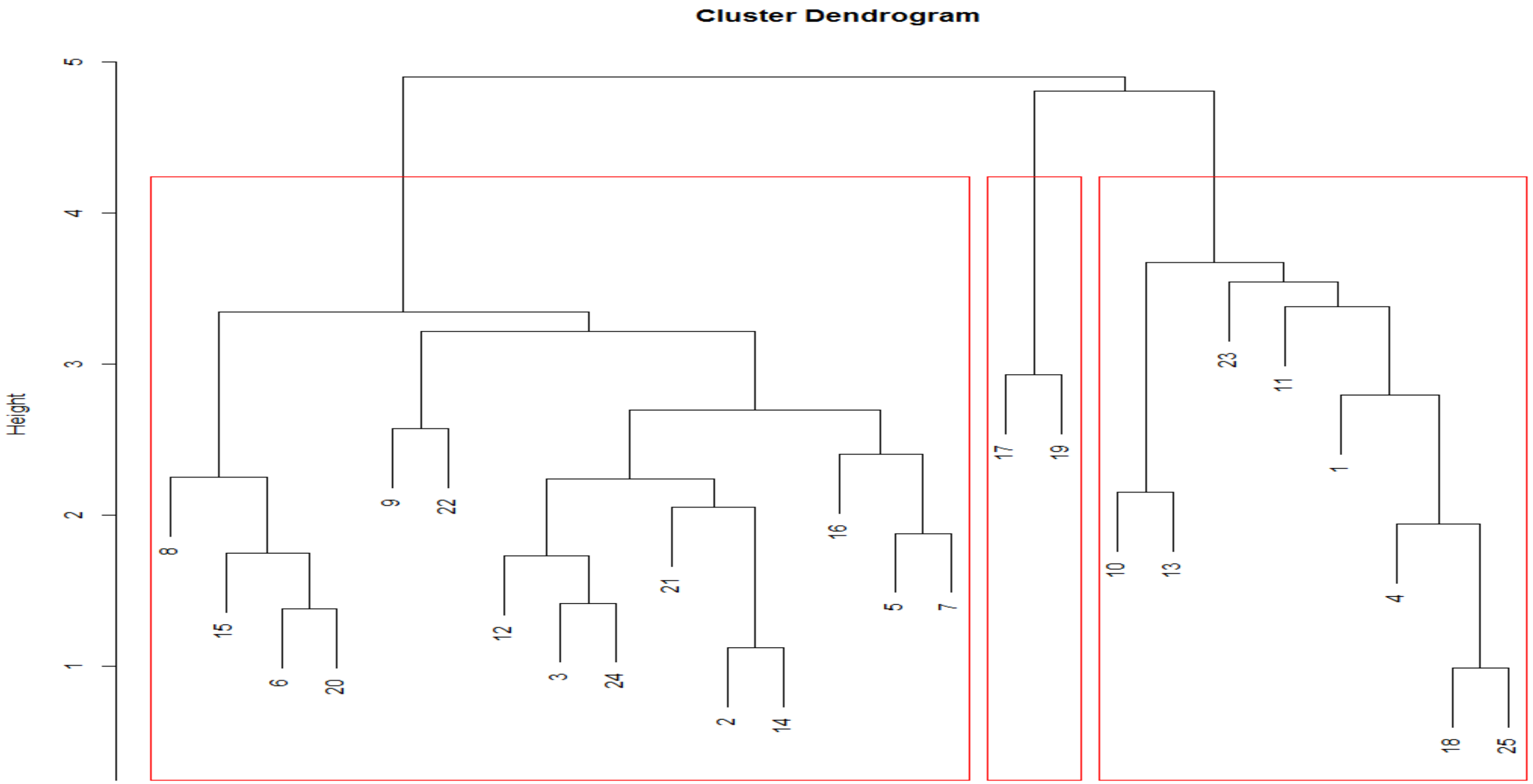
- K-means Clustering: K=3, set.seed(2406)

	Cluster 1	Cluster 2	Cluster 3
#Instances	N= 6	N= 9	N = 10
Countries	Austria Czech Germany Netherlands Poland W Germany	Belgium Denmark Finland France Ireland Norway Sweden Switzerland UK	Albania Bulgaria Greece Hungary Italy Portugal Romania Spain Russia Yugoslavia
Foods	White Meat Starch	Red Meat Milk Fish Fruit and Veggies	Eggs Cereals Nuts

- <u>Elbow Method: </u>2 or 3 clusters look relevant



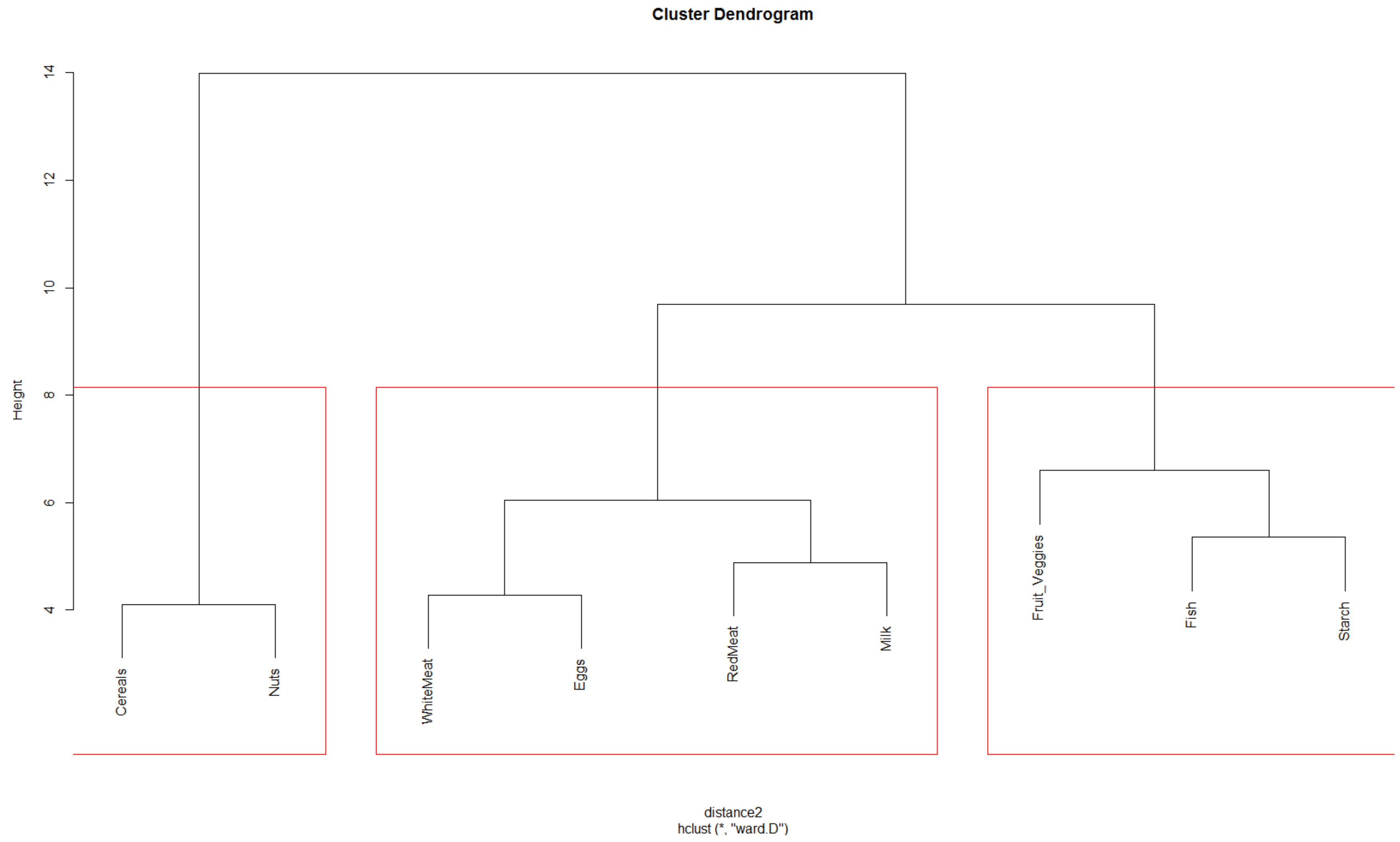
- <u>Hierarchical Clustering by Countries (3 Clusters, Average Distance)</u>



- Hierarchical Clustering by Countries (3 Clusters, Average Distance)

	Cluster 1	Cluster 2	Cluster 3
#Instances	N= 8	N= 15	N = 2
Countries	Albania Bulgaria Greece Hungary Italy Romania Russia Yugoslavia	Austria Belgium Czech Denmark Germany Finland France Ireland Netherlands Norway Poland Sweden Switzerland UK W Germany	Portugal Spain

- <u>Hierarchical Clustering by Food Sources (3 Clusters, Ward’s Method)</u>



- Hierarchical Clustering by Food Sources (3 Clusters, Ward’s Method)

	Cluster 1	Cluster 2	Cluster 3
#Instances	N= 4	N= 3	N = 2
Countries	White Meat Eggs Red Meat Milk	Fruits & Veggies Fish Starch	Cereals Nuts

- Evaluate the Business Problem
  - What will you do with the clusters??