Recipes 데이터 분석

강사 : 문성민

Text Analysis “wordcloud”

Text Analysis “wordcloud”

- 경로 지정

```
> setwd("/Users/Seongmin_M/Desktop/Class")
>
```

- 데이터 불러오기

```
> Harry_potter_1 <- file("/Users/Seongmin_M/Desktop/Class/Harry_potter_1.txt", blocking=F)
```

- 데이터 읽어들이기

```
> Harry_txtLines <- readLines(Harry_potter_1)
```

- 데이터 닫기

```
> close(Harry_potter_1)
```

- 데이터 다듬기

```
> Harry_txtLines <- gsub("()", "", Harry_txtLines)
> Harry_txtLines <- gsub("<", "", Harry_txtLines)
> Harry_txtLines <- gsub(">", "", Harry_txtLines)
> Harry_txtLines <- gsub("[ |\t]{2,}", "", Harry_txtLines)
```

Text Analysis “wordcloud”

- tm 패키지 설치 및 라이브러리 불러오기

```
> install.packages("tm")
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0100  644k  100
644k    0     0  1290k      0 --:--:-- --:--:-- --:--:-- 1292k

The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpZs3wki/downloaded_packages
> library(tm)
필요한 패키지를 로딩중입니다: NLP
```

- 텍스트 데이터를 코퍼스데이터로 변환

```
> Harry_txtLines_corpus <- Corpus(VectorSource(Harry_txtLines))
```

- 마침표 제거하기

```
> Harry_txtLines_corpus <- tm_map(Harry_txtLines_corpus, function(x)removeWords(x,stopwords()
))
```

Text Analysis “wordcloud”

- 수치형 데이터로 형 변환

```
> Harry_Tdm <- TermDocumentMatrix(Harry_txtLines_corpus, control = list(wordLengths = c(2, In
f)))
```

- wordcloud 패키지 설치 및 라이브러리 불러오기

```
> install.packages("wordcloud")
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0  1  138k    1
2168    0     0   2067      0  0:01:08  0:00:01  0:01:07  2068100  138k  100  138k    0     0
126k      0  0:00:01  0:00:01 --:--:--  126k

The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpZs3wki/downloaded_packages
>
> library(wordcloud)
필요한 패키지를 로딩중입니다: RColorBrewer
```

Text Analysis “wordcloud”

- 매트릭스 형태로 형 변환

```
> Harry_Tdm_M <- as.matrix(Harry_Tdm)
```

- 단어들의 출현빈도 카운팅

```
> Harry_wordFreq <- sort(rowSums(Harry_Tdm_M), decreasing = TRUE)
```

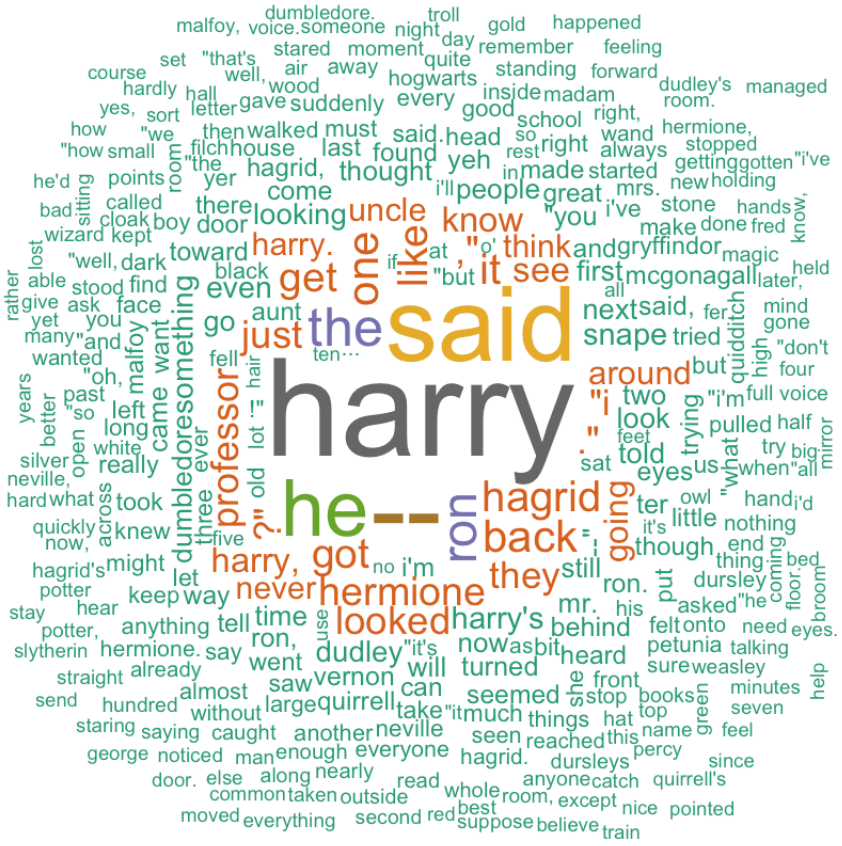
- 워드클라우드 색상 지정

```
> pal <- brewer.pal(8, "Dark2")
```

Text Analysis “wordcloud”

- 워드클라우드 생성

```
> wordcloud(words = names(Harry_wordFreq), freq = Harry_wordFreq, min.freq = 20, random.order
= F, rot.per = 0.1, colors = pal)
```



Cluster Analysis with R

What is Cluster Analysis?

- 개념(Concept)

  - N개의 관찰치를 대상으로 p개의 변수를 측정했을 때, 관측한 p개의 변수 값을 이용하여 N개의 관찰치 사이의 유사성(similarity)의 정도를 측정하여 관찰치들을 가까운 순서대로 군집화하는 통계적 분석 방법이다.
  - 두 관찰치 사이의 유사성을 측정하는 여러 방법(유클리디안, 유클리디안 제곱거리, 코사인값, 상관계수, 체비셰프, 블록, 민코브스키, 커스텀거리,,,) 중 가장 일반적으로 이용되는 방법은 유클리디안 거리 또는 상관계수 공식을 이용한 변수가 유사성 측정이다.
  - 데이터의 특징을 나타내는 변수간에 값의 차이 혹은 단위의 차이가 클 때는 데이터를 표준화시켜 사용하는 것이 일반적이다.

- 유사성 관련 수식(Formula)

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

유클리디안 거리 측정 공식

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

상관계수 거리 측정 공식

What is Cluster Analysis?

- 표준화 관련 수식(Formula)

$$Z_i = \frac{W_i - \bar{W}}{S_W} \qquad \bar{W} = \frac{\sum_{i=1}^{n} W_i}{n} \qquad S_W = \sqrt{\frac{\sum (W_i - \bar{W})^2}{n}}$$

- 군집화 방법의 개념(Concept)

  - 계층적 군집 분석 : 각 관찰치들 사이의 유사성, 거리 행렬을 구한 뒤에 관찰치들을 가까운 순서대로 연결해 가는 방법(최단연결법, 최장연결법, 중심연결법, 평균연결법, 워드의 방법 등이 있다.)

  - 비계층적 군집 분석 : 비 계층 적 군집 방법은 K-means clustering으로 불리며 관찰치들이 속할 군집의 수(K)를 미리 정한 뒤 정해진 군집으로 관찰치들을 포함시키는 방법이다.

- 예제(Example)
  - 각기 다른 브랜드 별로 생산된 시리얼의 여러 정보에 대한 데이터.

- 변수 설명
  - Manufacturer = 제조사
  - Calories = 칼로리
  - Protein = 단백질의 그램
  - Fat = 지방의 그램
  - Sodium = 나트륨의 밀리 그램
  - Fiber = 식이섬유
  - Carbohydrates = 탄수화물의 그램
  - Sugars = 설탕의 그램
  - Shelf = 판매 선반
  - Potassium = 칼륨의 밀리 그램
  - Vitamins = 비타민
  - Weight = 무게
  - Cups = 컵의 수

```
> setwd("/Users/Seongmin_M/Downloads")
>
> Cereal_total<-read.csv("Cereal_dataset.csv",head=T)
>
> head(Cereal_total)
                 Cereal Manufacturer Type Calories Protein Fat Sodium Fiber Carbohydrates
1 Apple Cinnamon Cheerios            G    C      110       2   2    180   1.5          10.5
2                Basic 4            G    C      130       3   2    210   2.0          18.0
3               Cheerios            G    C      110       6   2    290   2.0          17.0
4   Cinnamon Toast Crunch            G    C      120       1   3    210   0.0          13.0
5               Clusters            G    C      110       3   2    140   2.0          13.0
6            Cocoa Puffs            G    C      110       1   1    180   0.0          12.0
  Sugars Shelf Potassium Vitamins Weight Cups
1     10     1        70       25   1.00 0.75
2      8     3       100       25   1.33 0.75
3      1     1       105       25   1.00 1.25
4      9     2        45       25   1.00 0.75
5      7     3       105       25   1.00 0.50
6     13     2        55       25   1.00 1.00
>
```

- 데이터 확인

```
> setwd("/Users/Seongmin_M/Downloads")
>
> Cereal_total<-read.csv("Cereal_dataset.csv",head=T)
>
> head(Cereal_total)
                 Cereal Manufacturer Type Calories Protein Fat Sodium Fiber Carbohydrates
1 Apple Cinnamon Cheerios            G    C      110       2   2    180   1.5          10.5
2                Basic 4            G    C      130       3   2    210   2.0          18.0
3               Cheerios            G    C      110       6   2    290   2.0          17.0
4  Cinnamon Toast Crunch            G    C      120       1   3    210   0.0          13.0
5               Clusters            G    C      110       3   2    140   2.0          13.0
6            Cocoa Puffs            G    C      110       1   1    180   0.0          12.0
  Sugars Shelf Potassium Vitamins Weight Cups
1     10     1        70       25   1.00 0.75
2      8     3       100       25   1.33 0.75
3      1     1       105       25   1.00 1.25
4      9     2        45       25   1.00 0.75
5      7     3       105       25   1.00 0.50
6     13     2        55       25   1.00 1.00
>
> tail(Cereal_total)
   Cereal Manufacturer Type Calories Protein Fat Sodium Fiber Carbohydrates Sugars Shelf
78                             NA      NA  NA     NA    NA            NA     NA    NA
79                             NA      NA  NA     NA    NA            NA     NA    NA
80                             NA      NA  NA     NA    NA            NA     NA    NA
81                             NA      NA  NA     NA    NA            NA     NA    NA
82                             NA      NA  NA     NA    NA            NA     NA    NA
83                             NA      NA  NA     NA    NA            NA     NA    NA
   Potassium Vitamins Weight Cups
78        NA       NA     NA   NA
79        NA       NA     NA   NA
80        NA       NA     NA   NA
81        NA       NA     NA   NA
82        NA       NA     NA   NA
83        NA       NA     NA   NA
```

- 데이터 확인 결과 데이터 값 내에 NA값이 포함된 것을 확인 할 수 있다.

- 결측치(NA) 정제 과정

```
> Cereal_total_1=complete.cases(Cereal_total)
>
> head(Cereal_total_1)
[1] TRUE TRUE TRUE TRUE TRUE TRUE
>
> tail(Cereal_total_1)
[1] FALSE FALSE FALSE FALSE FALSE FALSE
>
> Cereal_total_2=Cereal_total[Cereal_total_1,]
>
> head(Cereal_total_2)
                 Cereal Manufacturer Type Calories Protein Fat Sodium Fiber Carbohydrates
1 Apple Cinnamon Cheerios            G    C      110       2   2    180   1.5          10.5
2                Basic 4            G    C      130       3   2    210   2.0          18.0
3               Cheerios            G    C      110       6   2    290   2.0          17.0
4   Cinnamon Toast Crunch            G    C      120       1   3    210   0.0          13.0
5               Clusters            G    C      110       3   2    140   2.0          13.0
6            Cocoa Puffs            G    C      110       1   1    180   0.0          12.0
  Sugars Shelf Potassium Vitamins Weight Cups
1     10     1        70       25   1.00 0.75
2      8     3       100       25   1.33 0.75
3      1     1       105       25   1.00 1.25
4      9     2        45       25   1.00 0.75
5      7     3       105       25   1.00 0.50
6     13     2        55       25   1.00 1.00
>
> tail(Cereal_total_2)
                          Cereal Manufacturer Type Calories Protein Fat Sodium Fiber
72 Muesli Raisins, Peaches, & Pecans            R    C      150       4   3    150   3.0
73                       Rice Chex              R    C      110       1   0    240   0.0
74                      Wheat Chex              R    C      100       3   1    230   3.0
75                           Maypo              A    H      100       4   1      0   0.0
76             Cream of Wheat (Quick)           N    H      100       3   0     80   1.0
77                   Quaker Oatmeal             Q    H      100       5   2      0   2.7
   Carbohydrates Sugars Shelf Potassium Vitamins Weight  Cups
72            16     11     3       170       25     -1 -1.00
73            23      2     1        30       25      1  1.13
74            17      3     1       115       25      1  0.67
75            16      3     2        95       25      1 -1.00
76            21      0     2        -1        0      1  1.00
77            -1     -1     1       110        0      1  0.67
>
```

- Complete.cases함수를 사용해서 결측치가 있는 행을 제거 해 주었다.

- 데이터 추출하기

```
> Cereal_1=Cereal_total_2[,4:15]
>
> head(Cereal_1)
  Calories Protein Fat Sodium Fiber Carbohydrates Sugars Shelf Potassium Vitamins Weight Cups
1      110       2   2    180   1.5          10.5     10     1        70       25   1.00 0.75
2      130       3   2    210   2.0          18.0      8     3       100       25   1.33 0.75
3      110       6   2    290   2.0          17.0      1     1       105       25   1.00 1.25
4      120       1   3    210   0.0          13.0      9     2        45       25   1.00 0.75
5      110       3   2    140   2.0          13.0      7     3       105       25   1.00 0.50
6      110       1   1    180   0.0          12.0     13     2        55       25   1.00 1.00
>
> row.names(Cereal_1)=Cereal_total_2[,1]
>
> head(Cereal_1)
                        Calories Protein Fat Sodium Fiber Carbohydrates Sugars Shelf Potassium
Apple Cinnamon Cheerios      110       2   2    180   1.5          10.5     10     1        70
Basic 4                      130       3   2    210   2.0          18.0      8     3       100
Cheerios                     110       6   2    290   2.0          17.0      1     1       105
Cinnamon Toast Crunch        120       1   3    210   0.0          13.0      9     2        45
Clusters                     110       3   2    140   2.0          13.0      7     3       105
Cocoa Puffs                  110       1   1    180   0.0          12.0     13     2        55
                        Vitamins Weight Cups
Apple Cinnamon Cheerios       25   1.00 0.75
Basic 4                       25   1.33 0.75
Cheerios                      25   1.00 1.25
Cinnamon Toast Crunch         25   1.00 0.75
Clusters                      25   1.00 0.50
Cocoa Puffs                   25   1.00 1.00
>
```

- 수치형 데이터와 각 시리얼의 이름만 추출하였다.

- 표준화 & 거리행렬 생성

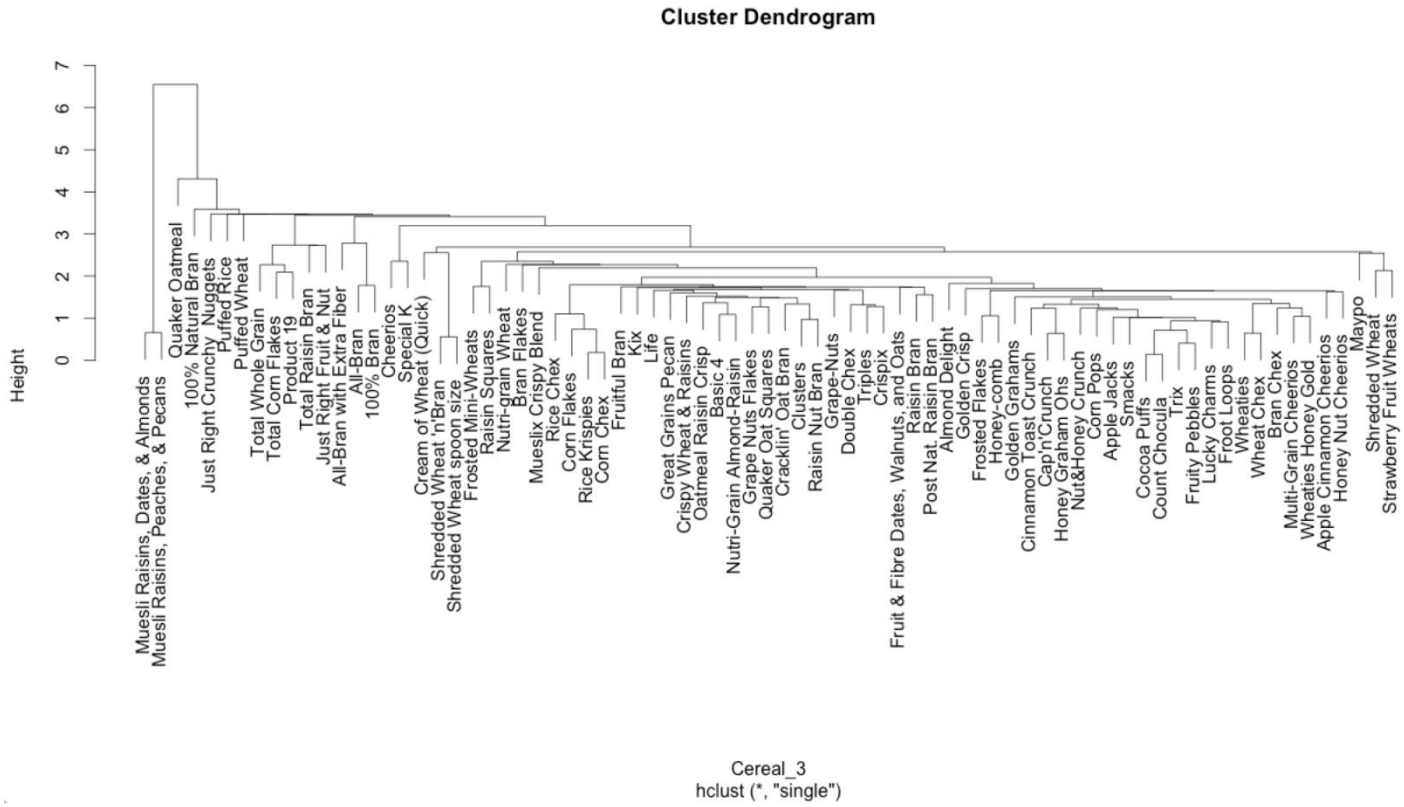
```
> Cereal_2=scale(Cereal_1)
>
> head(Cereal_2)
                          Calories    Protein         Fat     Sodium       Fiber Carbohydrates
Apple Cinnamon Cheerios 0.1599704 -0.4982277  0.98066557  0.2424445 -0.27354112    -0.9575706
Basic 4                 1.1864474  0.4151897  0.98066557  0.6003018 -0.06375361     0.7951933
Cheerios                0.1599704  3.1554419  0.98066557  1.5545879 -0.06375361     0.5614915
Cinnamon Toast Crunch   0.6732089 -1.4116451  1.97423464  0.6003018 -0.90290366    -0.3733159
Clusters                0.1599704  0.4151897  0.98066557 -0.2346986 -0.06375361    -0.3733159
Cocoa Puffs             0.1599704 -1.4116451 -0.01290349  0.2424445 -0.90290366    -0.6070178
                            Sugars      Shelf   Potassium   Vitamins     Weight       Cups
Apple Cinnamon Cheerios  0.69246377 -1.4507595 -0.36581692 -0.1453172 0.06236525  0.2613425
Basic 4                  0.24250841  0.9515734  0.05501828 -0.1453172 0.98370305  0.2613425
Cheerios                -1.33233535 -1.4507595  0.12515748 -0.1453172 0.06236525  1.0643502
Cinnamon Toast Crunch    0.46748609 -0.2495930 -0.71651292 -0.1453172 0.06236525  0.2613425
Clusters                 0.01753073  0.9515734  0.12515748 -0.1453172 0.06236525 -0.1401613
Cocoa Puffs              1.36739680 -0.2495930 -0.57623452 -0.1453172 0.06236525  0.6628463
```

```
> Cereal_3=dist(Cereal_2,method="euclidean")
```

- 변수간 측정 단위가 상이하므로 표준화를 시킨 후 거리행렬을 생성한다.

- 유클리디안 거리 , 최단 연결법

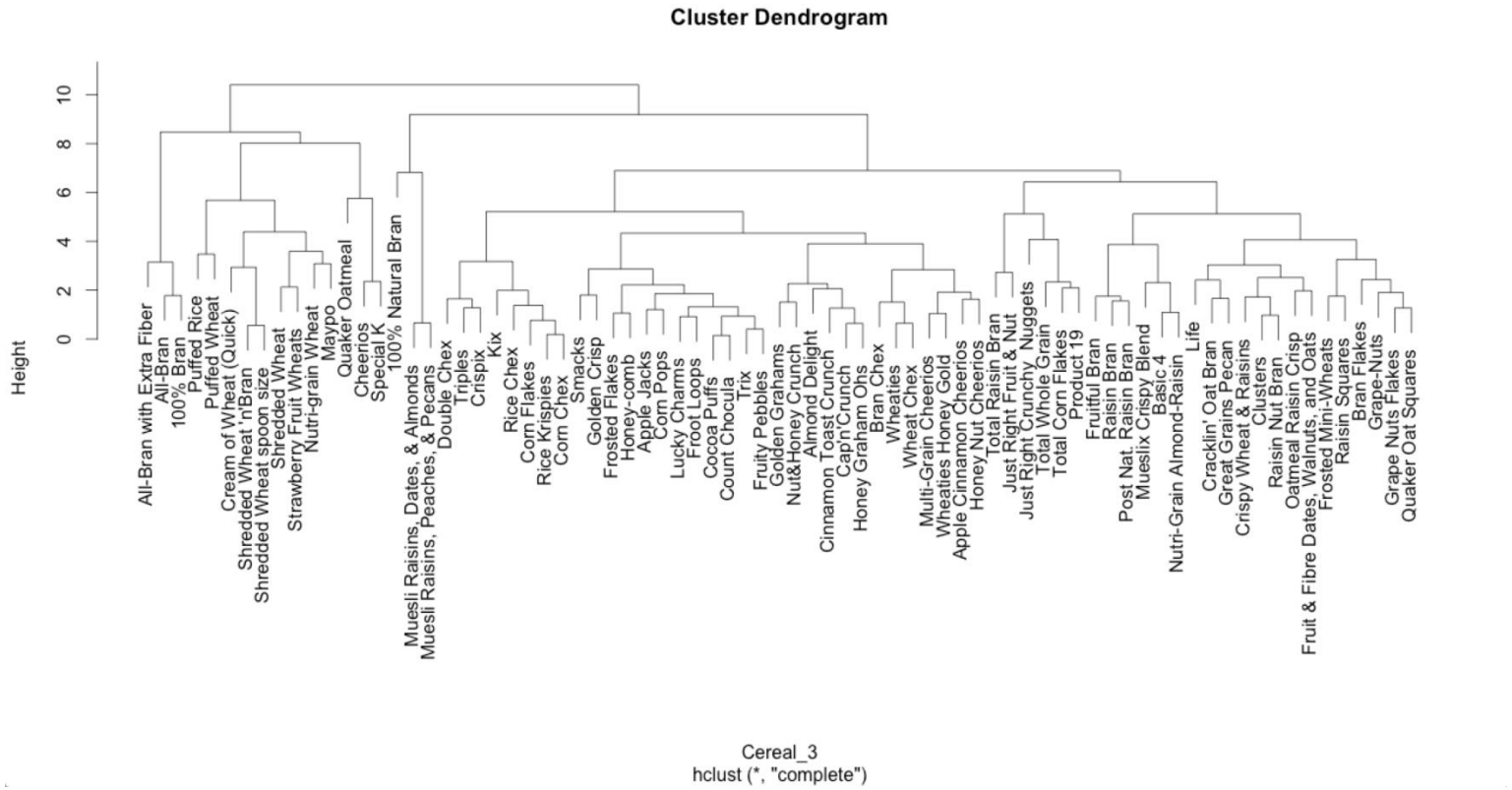
```
> Cereal_4=hclust(Cereal_3,method="single")
>
> plot(Cereal_4)
```



- 유클리디안 방법으로 거리 값을 생성 후 최단 연결법을 사용하여 데이터를 군집화하여 수형도(Dendrogram)를 생성한다.

- 유클리디안 거리 , 최장 연결법

```
> Cereal_5=hclust(Cereal_3,method="complete")
>
> plot(Cereal_5)
```



- 유클리디안 방법으로 거리 값을 생성 후 최장 연결법을 사용하여 데이터를 군집화하여 수형도(Dendrogram)를 생성한다.

- 라이브러리 설치

```
> install.packages("MASS")

  바이너리 버전을 이용할 수 있습니다 (그리고 설치되어질 것입니다) 그러나 소스 버전은
  추후에 제공될 것입니다:
     binary source
MASS 7.3-39 7.3-40

URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/MASS_7.3-39.tgz'을 시도합니다
Content type 'application/x-gzip' length 1049472 bytes (1.0 Mb)
URL을 열었습니다
==================================================
downloaded 1.0 Mb


The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmp7WjjMf/downloaded_packages
```

- 시각화 하기에 앞서 사용할 패키지를 설치하여 준다.
- Install.packages("MASS")

MDS Visualization with R

- 좌표값 생성

```
> library(MASS)
>
> Cereal_6=isoMDS(Cereal_3)
initial  value 28.785541
iter   5 value 19.376128
iter  10 value 18.753638
iter  10 value 18.742479
iter  10 value 18.738775
final  value 18.738775
converged
>
> head(Cereal_6)
$points
                                 [,1]        [,2]
Apple Cinnamon Cheerios     0.364656553  0.26230345
Basic 4                     0.004591595  1.05246966
Cheerios                   -2.403035562 -0.62517275
Cinnamon Toast Crunch       0.995680024  1.06829345
Clusters                   -0.377954097  0.31812162
Cocoa Puffs                 1.074699608  0.40892502
Count Chocula               1.022990824  0.40844767
```

- 유클리디안 거리행렬 값을 사용하여 MDS시각화에 사용될 좌표 값을 구해준다.

MDS Visualization with R

- 데이터 프레임 화 시키기

```
> Cereal_7=as.data.frame(Cereal_6)
>
> head(Cereal_7)
                          points.1   points.2   stress
Apple Cinnamon Cheerios  0.364656553  0.2623034 18.73878
Basic 4                  0.004591595  1.0524697 18.73878
Cheerios                -2.403035562 -0.6251727 18.73878
Cinnamon Toast Crunch    0.995680024  1.0682935 18.73878
Clusters                -0.377954097  0.3181216 18.73878
Cocoa Puffs              1.074699608  0.4089250 18.73878
```

- 좌표 값을 계산한 후 시각화에 사용하기 위해 데이터 프레임으로 변환시켜 준다.

MDS Visualization with R

- ggplot2를 사용한 MDS 시각화

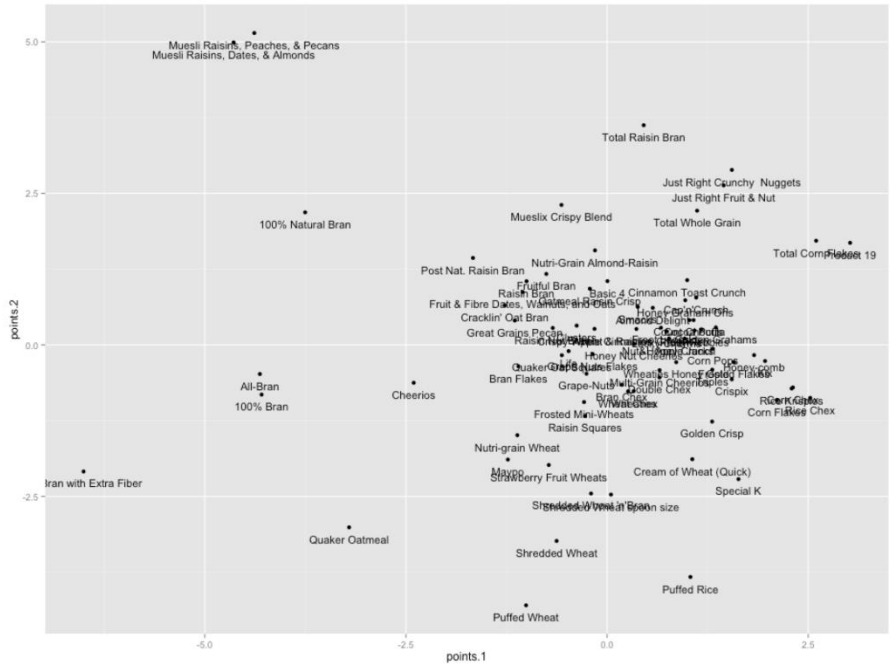
```
> library(ggplot2)

다음의 패키지를 부착합니다: 'ggplot2'

The following object is masked from 'mtcars':

    mpg

>
> ggplot(Cereal_7,aes(x=points.1,y=points.2))+geom_point()+geom_text(aes(label=row.name
s(Cereal_7)),size=4,vjust=2)
```



- ggplot2패키지를 사용하여 MDS좌표 값을 그래프 공간 위에 나타내준다.