성분을 통한 과자 특징 분류

INDEX

1. 개요

1) 동기

저는 평소 집에서 쉴 때도, 공부를 할 때도 항상 과자를 먹습니다. 어릴 때부터 밥은 걸러도 과자는 꾸준히 먹었던 것 같습니다. 그만큼 과자를 좋아하고 다양한 먹어보았습니다.
요즘도 신제품이 출시되면 설레는 마음으로 마트로 달려가곤 합니다.
그런데 요즘은 기술력의 발달과 삶의 질이 상승함에 따라 수많은 종류가 과자가 쏟아져 나옵니다. 수많은 과자들 속에 제가 먹고 싶은 과자를 선택하기가 쉽지 않았습니다.
그래서 저는

"그 선택을 간단한 알고리즘을 통해서 어느저도의 카테고리를 잡을 수는 없을까?"

라는 호기심이 생겼습니다. 그래서 저는 이번 프로젝트를 통해 그동안 제가 품어왔던 궁금증을 해소하고자 과자내의 포함된 영양성분과 열량을 가지고 과자의 종류(ex. 크래커, 샌드, 감자칩)를 분류하는 분석을 진행했습니다.

2) 목적

시중에서 쉽게 구할 수 있고 실제 제가 먹어본 경험이 있는 과자를 바탕으로 각 과자마다 표기되어 있는 영양 성분표를 통해 과자의 맛과 영양성분 사이에서 유의미한 공통점과 차이점을 찾고 이를 통해 영양성분을 기준으로 과자의 종류를 나누어보려 합니다. **원하는 과자를 고르는 데 있어서 선택의 폭을 줄이고 신속하며 합리적인 선택을 유도하려 합니다.**

3) 개요

먼저 R – Tech., 즉 PCA와 FA(PCFA,MLFA)를 통해 차원축소 시켜 새롭게 생성된 변수들과 관측치들 간의 관계 속에서 과자의 종류를 분류해보았습니다.
이후 Q – Tech.의 한 방법인 CA를 통해 이전에 R –Tech.를 통해 군집을 나눠보았던 것을 비교해보는 과정을 걸쳐서 과자를 종류별로 분류해보았습니다.

2. 데이터 소개

1) 데이터 설명

본 데이터는 실제 제가 먹어본 경험이 있는 과자들을 대상으로 식품안전정보포털(http://www.foodsafetykorea.go.kr)를 통해 직접 검색하여 얻은 자료입니다. 과자류로 분류된 가공제품군을 대상으로 총 25개의 과자에 대해 열량과 총 8개의 영양성분을 조사한 데이터입니다. 실제 제가 경험해 본 과자를 통해 비교해야 보다 정확한 분석이 가능하다고 생각하기 때문에 다소 관측치의 개수가 적더라도 먹어본 것을 위주로 데이터를 생성했습니다.

변수의 개수 : 10개

관측치 개수 : 25개

	제품명	1회제공량	열량	탄수화물	단백질	지방	당류	나트륨	콜레스테롤	포화지방산	트랜스지방산
2	치토스매콤한맛	100	570	56	5	35	5	400	0	23	0
3	치토스바베큐맛	100	575	59	5	36	5	445	0	18	0
4	도리토스나쵸치즈맛	30	154.5	18.9	2.1	8.1	2.1	159	0	3.6	0
5	도도한나쵸 멕시칸타코맛	30	144.76	19.26	2.49	6.8	0.68	172.31	0.03	1.89	0.06
6	썬칩 오리지날	30	144.6	17.7	1.5	7.5	3.9	201.6	0.03	2.1	0.1
7	스윙칩	30	168	16.8	1.5	10.2	0	102	0	0	0
8	칩포테이토	30	170.44	16.53	1.36	10.98	0.44	77.81	0.94	3.4	0
9	포카칩오리지날	30	174	16.5	1.8	10.5	0	102	0	0	0
10	포카칩 양파맛	30	168	16.8	1.8	10.2	0	102	0	0	0
11	듀팝스팝콘	20	101.43	11.86	1.8	5.62	0.76	123.48	0.82	2.23	0.03
12	조청유과	30	155	18	0.8	9	8	85	0	2.7	0
13	꿀꽈배기	30	144.9	19.8	1.86	6.12	0	70.8	0	0	0
14	고구마깡	30	147.9	18.96	1.77	6.75	0	165.3	0	0	0
15	새우깡	30	151.2	18.12	1.8	7.35	0	210	0	0	0
16	오잉	30	145.5	21.3	1.5	6	1.5	180	0	2.7	0
17	오징어집	30	155.45	18.55	2.18	8.18	1.64	147.27	1.5	2.56	0
18	꽃게랑	30	156	20.1	2.4	7.2	1.2	250.2	0.9	2.52	0
19	초코파이	70	208.25	31.36	1.96	8.33	16.8	78.4	N/A	N/A	N/A
20	몽쉘	70	244.51	23.67	2.21	15.58	15.31	60.27	N/A	N/A	N/A
21	마가렛트	30	145.97	17.63	1.71	7.63	8.71	48.36	9.96	3.62	0.05
22	후레쉬베리	30	156.13	20.25	1.54	8.05	2.89	61.92	20.51	4.04	0.05
23	에이스	30	158.1	17.67	2.1	8.7	0	95.1	0	0	0
24	제크	30	147	19.8	1.8	6.6	1.8	193.27	0.32	3.6	0.29
25	콘칲	30	157.9	19.23	1.49	8.33	0.83	122.63	0	2.53	0.07
26	꼬깔콘	30	162.3	17.97	1.83	9	0	175.2	0	0	0

2) 변수 설명

① 제품명 : 실제 출시되어 있는 과자의 제품명입니다.
(하나의 제품명에 여러 맛이 존재하는 경우 맛 또한 표기했습니다.)

② 1회제공량 : 식품의약품안전처에서 고시하도록 한 "식품등의 표시기준"에 의거하여 각 사측에서 표시한 각 제품의 1회 제공량입니다. (30, 70, 100)
기준단위 : (g)

③ 열량 : 각 제품의 1회 제공량에 대한 열량입니다.
기준단위 : (kcal)

④ 탄수화물 : 각 제품의 1회 제공량에 대한 탄수화물 함량입니다.
기준단위 : (g)

⑤ 단백질 : 각 제품의 1회 제공량에 대한 단백질 함량입니다.
기준단위 : (g)
⑥ 지방 : 각 제품의 1회 제공량에 대한 지방 함량입니다.
기준단위 : (g)
⑦ 당류 : 각 제품의 1회 제공량에 대한 당류 함량입니다.
기준단위 : (g)
⑧ 나트륨 : 각 제품의 1회 제공량에 대한 나트륨 함량입니다.
기준단위 : (mg)
⑨ 콜레스테롤 : 각 제품의 1회 제공량에 대한 콜레스테롤 함량입니다.
기준단위 : (mg)
⑩ 포화지방산 : 각 제품의 1회 제공량에 대한 포화지방산 함량입니다.
기준단위 : (g)
⑪ 트랜스지방산 : 각 제품의 1회 제공량에 대한 트랜스지방산 함량입니다.
기준단위 : (g)

3) 데이터 수정 및 기초통계량 확인

먼저 본 데이터의 경우 N/A가 존재함을 알 수 있습니다. 이것은 그 성분에 대하여 제품에 포함되어 양이 극소량인 경우 절삭하여 표기하지 않은 것이므로, 이 경우 모두 0으로 수정하였습니다.

또한 factor형으로 되어있는 변수의 경우에는 모두 numeric형으로 바꿔주었습니다.

본 데이터의 경우 모두 대부분의 과자가 1회제공량 기준이 30g이고 몇몇의 관측치의 경우는 70g과 100g이 존재함을 알 수 있습니다. 이것은 분석을 시행할 때 동일한 기준에서 시행하기 위하여 모두 30g으로 통일 시킨 후 그 변수를 삭제했습니다.

그리고 제품명의 경우는 rowname으로 지정해주고 삭제해주었습니다.

```
> data
                      열량 탄수화물    단백질     지방     당류 나트륨 콜레스테롤 포화지방산 트랜스지방산
치토스매콤한맛       171.000 16.80000 1.5000000 10.500000 1.500000 120.00     0.00     6.900      0.000
치토스바베큐맛       172.500 17.70000 1.5000000 10.800000 1.500000 133.50     0.00     5.400      0.000
도리토스나쵸치즈맛   154.500 18.90000 2.1000000  8.100000 2.100000 159.00     0.00     3.600      0.000
도도한나쵸 멕시칸타코맛 144.760 19.26000 2.4900000  6.800000 0.680000 172.31     0.03     1.890      0.060
썬칩 오리지날        144.600 17.70000 1.5000000  7.500000 3.900000 201.60     0.03     2.100      0.100
스윙칩               168.000 16.80000 1.5000000 10.200000 0.000000 102.00     0.00     0.000      0.000
칩포테이토           170.440 16.53000 1.3600000 10.980000 0.440000  77.81     0.94     3.400      0.000
포카칩오리지날       174.000 16.50000 1.8000000 10.500000 0.000000 102.00     0.00     0.000      0.000
포카칩 양파맛        168.000 16.80000 1.8000000 10.200000 0.000000 102.00     0.00     0.000      0.000
듀팝스팝콘           152.145 17.79000 2.7000000  8.430000 1.140000 185.22     1.23     3.345      0.045
조청유과             155.000 18.00000 0.8000000  9.000000 8.000000  85.00     0.00     2.700      0.000
꿀꽈배기             144.900 19.80000 1.8600000  6.120000 0.000000  70.80     0.00     0.000      0.000
고구마깡             147.900 18.96000 1.7700000  6.750000 0.000000 165.30     0.00     0.000      0.000
새우깡               151.200 18.12000 1.8000000  7.350000 0.000000 210.00     0.00     0.000      0.000
오잉                 145.500 21.30000 1.5000000  6.000000 1.500000 180.00     0.00     2.700      0.000
오징어집             155.450 18.55000 2.1800000  8.180000 1.640000 147.27     1.50     2.560      0.000
꿀게랑               156.000 20.10000 2.4000000  7.200000 1.200000 250.20     0.90     2.520      0.000
초코파이              89.250 13.44000 0.8400000  3.570000 7.200000  33.60     0.00     0.000      0.000
몽쉘                 104.790 10.14429 0.9471429  6.677143 6.561429  25.83     0.00     0.000      0.000
마가렛트             145.970 17.63000 1.7100000  7.630000 8.710000  48.36     9.96     3.620      0.050
후레쉬베리           156.130 20.25000 1.5400000  8.050000 2.890000  61.92    20.51     4.040      0.050
에이스               158.100 17.67000 2.1000000  8.700000 0.000000  95.10     0.00     0.000      0.000
제크                 147.000 19.80000 1.8000000  6.600000 1.800000 193.27     0.32     3.600      0.290
콘칲                 157.900 19.23000 1.4900000  8.330000 0.830000 122.63     0.00     2.530      0.070
꼬깔콘               162.300 17.97000 1.8300000  9.000000 0.000000 175.20     0.00     0.000      0.000
```

```
> summary(data)
      열량          탄수화물         단백질           지방            당류           나트륨          콜레스테롤        포화지방산
 Min.   : 89.25   Min.   :10.14   Min.   :0.800   Min.   : 3.570   Min.   :0.000   Min.   : 25.83   Min.   : 0.000   Min.   :0.000
 1st Qu.:145.97   1st Qu.:16.80   1st Qu.:1.500   1st Qu.: 6.800   1st Qu.:0.000   1st Qu.: 85.00   1st Qu.: 0.000   1st Qu.:0.000
 Median :155.00   Median :17.97   Median :1.770   Median : 8.100   Median :1.200   Median :122.63   Median : 0.000   Median :2.520
 Mean   :151.89   Mean   :17.83   Mean   :1.713   Mean   : 8.127   Mean   :2.064   Mean   :128.80   Mean   : 1.417   Mean   :2.086
 3rd Qu.:162.30   3rd Qu.:19.23   3rd Qu.:1.860   3rd Qu.: 9.000   3rd Qu.:2.100   3rd Qu.:175.20   3rd Qu.: 0.320   3rd Qu.:3.400
 Max.   :174.00   Max.   :21.30   Max.   :2.700   Max.   :10.980   Max.   :8.710   Max.   :250.20   Max.   :20.510   Max.   :6.900
  트랜스지방산
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.0266
 3rd Qu.:0.0450
 Max.   :0.2900
```

->콜레스테롤의 max., 나트륨의 min., max값이 평균보다 차이가 많은 것을 확인할 수 있다. 이를 통해 이후 분석 과정에서 문제 발생시 조치해야 할 수 있기 때문에 미리 각 관측치를 파악해놓을 필요성이 있습니다.

-> 트랜스지방산의 경우는 거의 0에 밀집되어 있기 때문에 이후 분석 과정에 큰 영향을 미치지 않을 가능성이 있음을 예측해볼 수 있습니다.

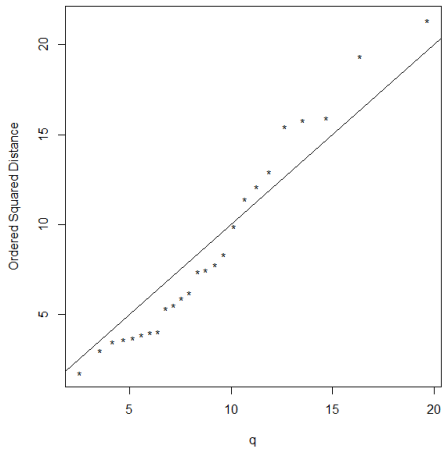
```
data=read.csv("과자DATA.csv",header = T)

#관측치에 N/A로 표시된 곳은 0 이하라는 의미이므로 모두 0으로 통일
for (i in c(1:25)){
  if (data[i,9]=="N/A"){
    data[i,9]=0
  }
}

for (i in c(1:25)){
  if (data[i,10]=="N/A"){
    data[i,10]=0
  }
}

for (i in c(1:25)){
  if (data[i,11]=="N/A"){
    data[i,11]=0
  }
}

# 세 변수에 대해서 factor형임을 확인한 후 numeric형으로 전환,
str(data)
data$콜레스테롤=as.character(data$콜레스테롤)
data$콜레스테롤=as.numeric(data$콜레스테롤)
data$포화지방산=as.character(data$포화지방산)
data$포화지방산=as.numeric(data$포화지방산)
data$트랜스지방산=as.character(data$트랜스지방)
data$트랜스지방산=as.numeric(data$트랜스지방)
```

```
#1열의 상품명을 행이름으로 지정
rownames(data)=data[,1]
data=data[-1]

#1회제공량이 다른 관측치의 경우 해석의 편의성을 위해 30g으로 통일
summary(data$X1회제공량)
which(data[,1]>30 | data[,1]<30)

for (i in c(1:10)){
  data[1,i]=data[1,i]*0.3
}
data[1,]

for (i in c(1:10)){
  data[2,i]=data[2,i]*0.3
}
data[2,]

for (i in c(1:10)){
  data[10,i]=data[10,i]*1.5
}
data[10,]

for (i in c(1:10)){
  data[18,i]=(data[18,i]/7)*3
}
data[18,]

for (i in c(1:10)){
  data[19,i]=(data[19,i]/7)*3
}
data[19,]


#1회제공량을 모두 30으로 통일했으므로 제거.
data=data[,-1]
```

3. 분석 및 결과 해석

1) 정규성 분석 (Normality Test)

MLFA의 기본가정이 정규성 만족이므로 이를 위해 본 자료에 대해 다변량 정규성을 검정을 실시했습니다. 또한 위의 기초통계량을 통해 대략적으로 파악했던 outlier로 의심되는 관측치 파악을 위해 Q-Q plot을 그렸습니다.



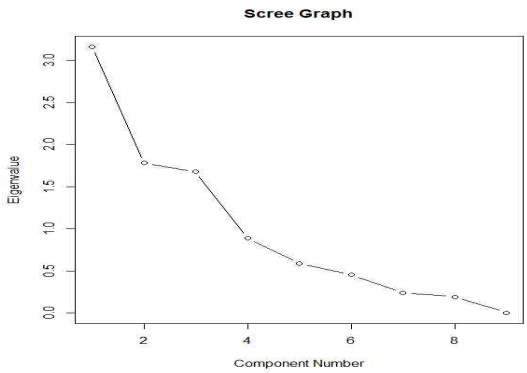

```
> X=data
> n=dim(X)[1]
> p=dim(X)[2]
> S=cov(X)
> Xbar=colMeans(X)
> m=mahalanobis(X, Xbar, S)
> m=sort(m)
> id=seq(1, n)
> pt=(id-0.5)/n
> q=qchisq(pt, p)
> plot(q, m, pch="*", ylab="Ordered Squared Distance")
> abline(0, 1)
```

```
> # Correlation Coefficient Test for Normailty
> rq=cor(cbind(q, m))[1,2]
> rq
[1] 0.9833517
```

그래프 상으로는 관측치 개수가 적어 다소 정규성을 만족하지 않는 것으로 보일 수 있으나 전체적으로 형태가 정규성을 만족하는 형태이고 더군다나 분위수와 마할라노비스 거리의 상관계수 r_Q=0.9833517 의 값을 보면 거의 1이 되어 카이제곱그림의 직선성이 매우 인정되어 다변량 정규성을 만족한다고 볼 수 있습니다.

2) 주성분 분석 (PCA)

상관계수행렬 R에 대한 Spectral decomposition을 통해 PCA를 실시했습니다.
먼저 주성분 개수를 정하기 위해 Eigenvalues를 이용하여 Scree Graph를 그리고 각각의 설명비율을 구했습니다.



```
> #[Step 4] Choice of Eigenvalues and Eigenvectors
> gof=eigen.R$values/sum(eigen.R$values)*100 # Goodness-of fit
> round(gof, 2)
[1] 35.11 19.82 18.63  9.89  6.59  5.05  2.70  2.16  0.07
> plot(eigen.R$values, type="b", main="Scree Graph", xlab="Component Number", ylab="Eigenvalue")
> round(sum(gof[1:3]),2)
[1] 73.55
```

Scree Graph를 통해 팔꿈치가 2, 4, 7 정도에서 이루어짐을 확인할 수 있었고 총 설명비율 또한 그 부근인 3개의 고유값의 설명비율 합이 약 73.55%이므로 주성분 개수를 3개로 하는 것이 적합하다고 판단했습니다.

```
> V3=V[,1:3]
> V3
      [,1]  [,2]  [,3]
 [1,] -0.46  0.40  0.08
 [2,] -0.41 -0.25  0.23
 [3,] -0.41 -0.24 -0.10
 [4,] -0.26  0.62  0.04
 [5,]  0.44 -0.05  0.34
 [6,] -0.40 -0.37 -0.12
 [7,]  0.05  0.05  0.58
 [8,] -0.16  0.11  0.59
 [9,] -0.07 -0.42  0.35
```

3개의 eigenvalue에 대응하는 eigenvector를 활용하여 표준화 변수(Z)의 선형결합인 주성분(P)을 구합니다.

$$p_1 = -0.46z_1 - 0.41z_2 - 0.41z_3 - 0.26z_4 + 0.44z_5 - 0.40z_6 + 0.05z_7 - 0.16z_8 - 0.07z_9$$
$$p2 = 0.40z_1 - 0.25z_2 - 0.24z_3 + 0.62z_4 - 0.05z_5 - 0.37z_6 + 0.05z_7 + 0.11z_8 - 0.42z_9$$
$$p3 = 0.08z_1 + 0.23z_2 - 0.10z_3 + 0.04z_4 + 0.34z_5 - 0.12z_6 + 0.58z_7 + 0.59z_8 + 0.35z_9$$

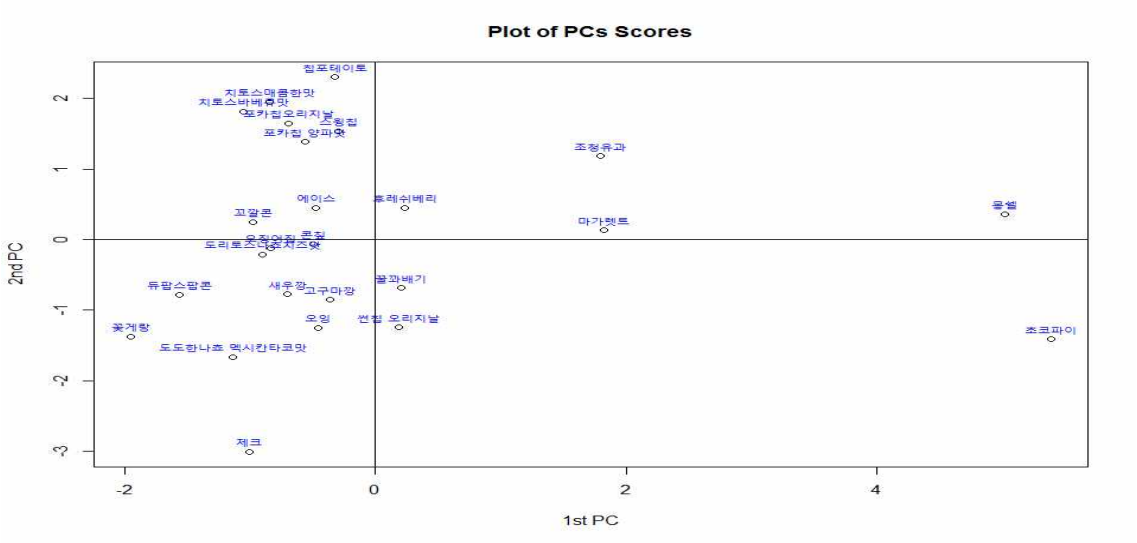
위의 식에서의 주성분계수를 통해서 각 주성분이 무엇을 의미하는지 해석해보았습니다.

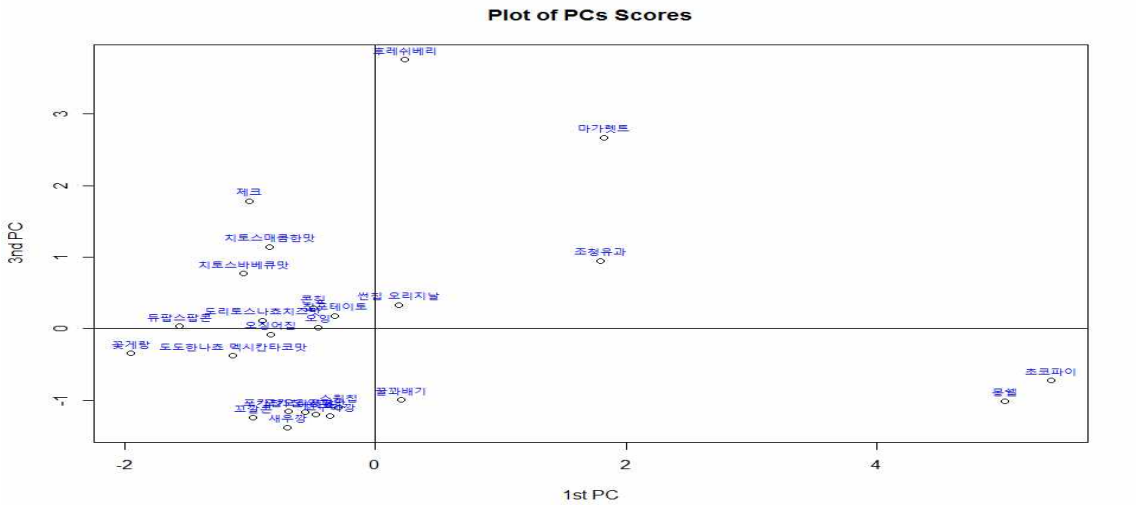
제 1 주성분 : (-)부호의 계수 중에서도 z1, z2, z3, z6에 대한 계수 값들이 비슷하면서도 큰 값을 가지고 있고 (+)부호의 계수로는 z5에 대한 계수의 값이 컸기 때문에 제 1 주성분은 "열량", "탄수화물", "단백질", "나트륨"의 함량과 "당류"의 함량에 대한 대비를 나타냅니다. 미각으로써 "열량", "탄수화물", "단백질"은 쉽게 판단할 수 없기에 **크게는 "나트륨"과 "당류"의 대비, 즉 짠맛과 단맛의 대비로 해석했습니다. (설명비율 : 35.11%)**

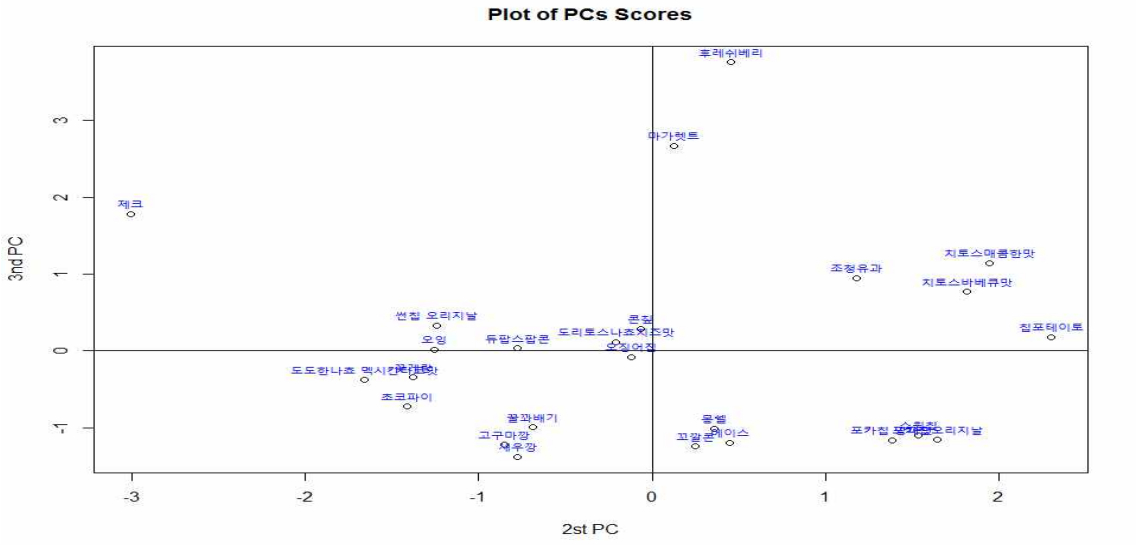
제 2 주성분 : (+)부호의 계수 중에서도 z1, z4에 대한 계수 값이 다른 계수들에 비해 큰 값이기 때문에 **제 2 주성분은 "열량", "지방"의 함량을 나타내는 값으로 해석했습니다. (설명비율 : 19.82%)**

제 3 주성분 : (+)부호의 계수 중에서도 z7, z8에 대한 계수 값이 다른 계수들에 비해 큰 값이기 때문에 **제 3 주성분은 "콜레스테롤", "포화지방산"의 함량을 나타내는 값으로 해석했습니다. (설명비율 : 18.63%)**

다음으로는 각 주성분을 통해 구한 PC scores를 주성분을 축으로 하는 plot 상에 찍어 확인해보았습니다.


Plot of PCs Scores
칩포테이토
치토스맵콤한맛
치토스바베큐맛
포카칩오리지널
스윙칩
포카칩 양파맛
조청유과
에이스
후레쉬베리
몽쉘
꼬깔콘
마가렛트
콘칲
도리토스나쵸치즈맛
새우깡
꿀꽈배기
고구마깡
듀팜스팜콘
오잉
썬칩 오리지널
꽃게랑
초코파이
도도한나쵸 멕시칸타코맛
제크
2nd PC
1st PC


Plot of PCs Scores
후레쉬베리
마가렛트
제크
조청유과
치토스맵콤한맛
치토스바베큐맛
콘칲
썬칩 오리지널
듀팜스팜콘
칩포테이토
오잉
꽃게랑
도도한나쵸 멕시칸타코맛
초코파이
몽쉘
꿀꽈배기
새우깡
3nd PC
1st PC

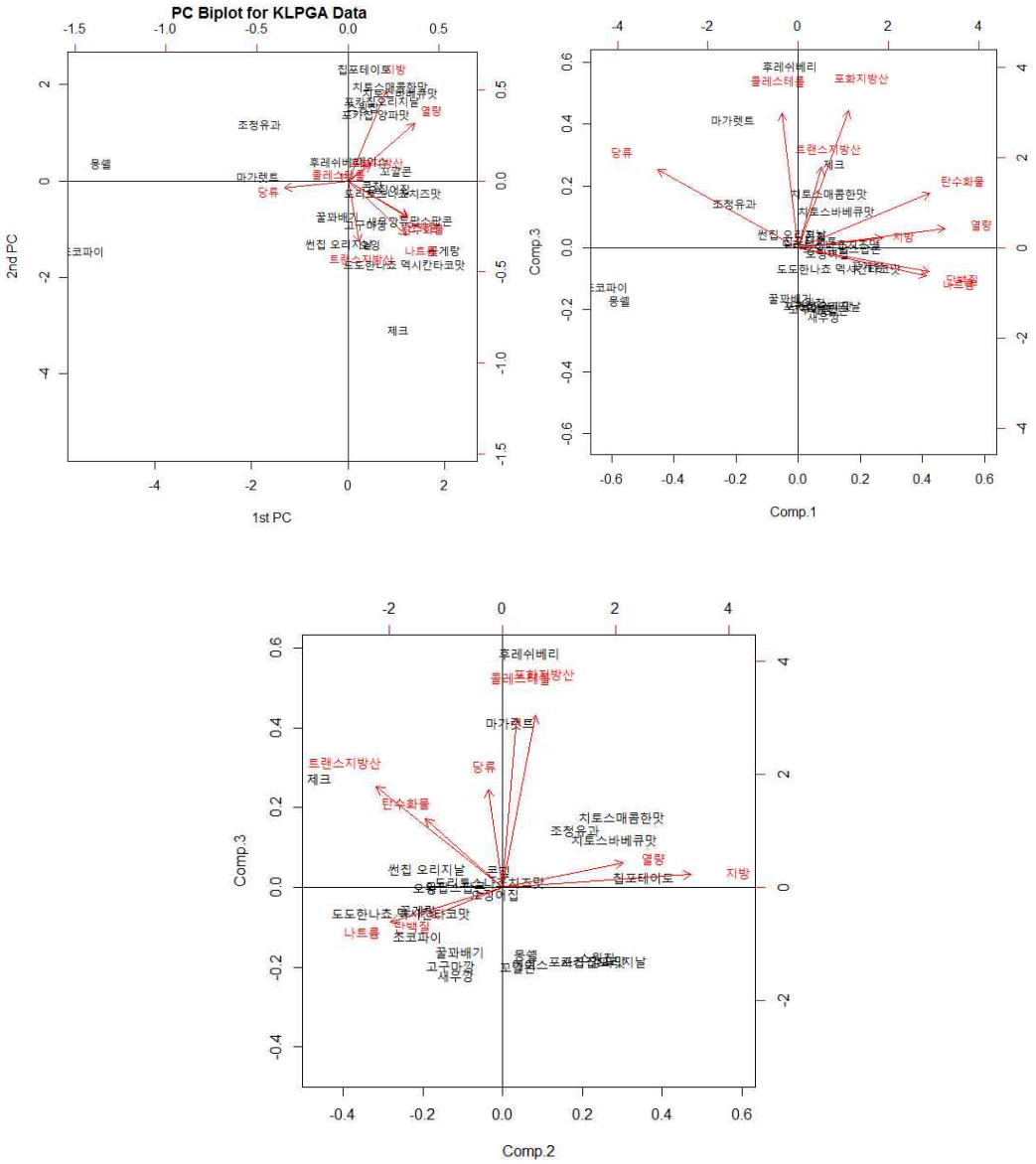

Plot of PCs Scores
후레쉬베리
마가렛트
제크
치토스맵콤한맛
조청유과
치토스바베큐맛
칩포테이토
썬칩 오리지널
콘칲
도리토스나쵸치즈맛
오잉
듀팜스팜콘
도도한나쵸 멕시칸타코맛
초코파이
꿀꽈배기
고구마깡
새우깡
몽쉘
에이스
꼬깔콘
포카칩 오리지널
3nd PC
2st PC

```
> round(P, 3)
                          [,1]   [,2]   [,3]
치토스매콤한맛          -0.846  1.947  1.142
치토스바베큐맛          -1.057  1.815  0.770
도리토스나쵸치즈맛      -0.902 -0.208  0.113
도도한나쵸 멕시칸타코맛 -1.137 -1.661 -0.379
썬칩 오리지날            0.185 -1.243  0.327
스윙칩                  -0.294  1.534 -1.098
칩포테이토              -0.326  2.300  0.178
포카칩오리지날          -0.691  1.643 -1.161
포카칩 양파맛           -0.557  1.380 -1.162
듀팝스팝콘             -1.568 -0.779  0.034
조청유과                1.795  1.181  0.949
꿀꽈배기                0.205 -0.686 -0.995
고구마깡                -0.362 -0.853 -1.225
새우깡                  -0.703 -0.776 -1.379
오잉                    -0.455 -1.251  0.013
오징어집                -0.838 -0.123 -0.084
꿀게랑                  -1.949 -1.375 -0.347
초코파이                 5.389 -1.410 -0.726
몽쉘                     5.014  0.362 -1.015
마가렛트                 1.823  0.129  2.668
후레쉬베리               0.232  0.451  3.759
에이스                  -0.475  0.446 -1.199
제크                    -1.007 -3.002  1.779
콘칲                    -0.498 -0.066  0.284
꼬깔콘                  -0.975  0.246 -1.248
```

제1주성분 & 제2주성분에 대한 그림을 통해서,
대부분의 과자들이 제 1주성분에 대해서 (-)값을 가지고 있으므로 짠맛을 가진다는 것을 해석할 수 있었다. 썬칩과 낱개형 과자(몽쉘,초코파이,마가렛트, 후레쉬베리)가 단맛을 주로 가지고 있음을 알 수 있습니다. 뿐만 아니라 옛날 유과(꿀꽈배기, 조청유과)들도 상품명에서 알 수 있듯이 주로 단맛을 가지고 있음을 알 수 있습니다.
감자칩류와 치토스가 주로 열량과 지방이 많이 포함되어 있고 나트륨이 많이 포함되어 있음을 확인할 수 있습니다. 파이류(초코파이,몽쉘)가 달고 열량이 낮음을 알 수 있었습니다. 그 외의 대부분의 기름에 튀긴 과자들은 탄수화물이 다소 많이 포함되어 있으며 짜고 열량이 높음을 알 수 있습니다.

제1주성분 & 제3주성분에 대한 그림을 통해서,
후레쉬베리, 마가렛트가 짠맛보다는 단맛을 많이 가지고 있고 콜레스테롤과 불포화지방산이 다른 과자류들에 대해 많이 포함되어 있음을 알 수 있습니다.
또한 대부분의 과자류가 x축(횡축) 근처와 원점에 몰려있는 것으로 보아 콜레스테롤과 포화지방산을 많지 않은 양을 포함하고 있으며 짠맛을 가짐을 확인할 수 있습니다.
짠맛과 단맛의 차이로는 콜레스테롤, 포화지방산의 함량과는 큰 연관성이 없음을 확인할 수 있습니다.

제2주성분 & 제3주성분에 대한 그림을 통해서,
감자칩류와 치토스가 주로 지방이 많이 들어있음을 확인할 수 있습니다.
제크가 지방이 적게 들어가 있음을 확인할 수 있습니다.
그리고 지방과 맛의 차이에는 큰 상관관계가 없음을 유추할 수 있습니다.
각 주성분의 설명비율이 크게 낮아진 만큼 과자의 종류를 잘 분류해내기는 힘들었습니다.
위의 해석한 것들을 biplot 통해 한 번 더 확인해보았습니다.



biplot을 통해 확인해본 결과,
관측치에 대한 해석은 동일하게 이루어짐을 알 수 있습니다.
각 변수가 주성분에 미치는 영향력과 상관관계에 대한 측면에서 살펴보면,
1st & 2nd PC를 통해서 ,
1st PC에 대해 "열량", "탄수화물", "단백질", "나트륨"과 "당류"가 상관관계와 영향력이 높음을 확인할 수 있습니다. 그리고 2nd PC에 대해서는 "열량", "지방"이 상관관계와 영향력이 높음을 확인할 수 있습니다.
1st & 3rd PC를 통해 ,
3rd PC에 대해 "콜레스테롤", "포화지방산"이 상관관계와 영향력이 높음을 확인할 수 있습니다.
2nd & 3rd에서 각 주성분과 변수간의 관계는 위의 2개로써 설명이 됨을 확인할 수 있으므로 일단 생략하고 변수들 간의 관계를 간략하게 살펴보았습니다.

또한 열량과 지방의 상관관계가 높음을 알 수 있습니다.
그리고 (나트륨과 탄수화물, 단백질), 세 변수 간의 상관관계가 큽니다.
지방과 열량간의 상관관계가 높고, 나트륨과 당류 간의 각이 둔각(>90°)임을 통해 어느정도 대비가 되면서도 과자들이 짠맛과 단맛을 같이 가지고 있음을 파악할 수 있습니다.
콜레스테롤과 포화지방산 간에도 매우 큰 상관관계가 있음을 파악할 수 있습니다.

이러한 결과를 통해 관측치들의 군집을 나눠보면,

파이류	감자칩과 치토스
낱개형 과자류(마가렛트, 후레쉬베리)	그 외의 짠 맛을 가진 유과

이와 같이 되겠습니다.

```
R=round(cor(X),3)
R


#[Step 3] Spectral Decomposition
eigen.R=eigen(R)
round(eigen.R$values, 2) # Eigenvalues
V=round(eigen.R$vectors, 2) # Eigenvectors

#[Step 4] Choice of Eigenvalues and Eigenvectors
gof=eigen.R$values/sum(eigen.R$values)*100 # Goodness-of fit
gof
round(gof, 2)
plot(eigen.R$values, type="b", main="Scree Graph", xlab="Component Number", ylab="Eigenvalue")
sum(gof[1:3])

#[Step 5] PCs : liner combination of original variables
V3=V[,1:3]
V3
colnames(Z)
#[Step 6] PCS, PCs Scores and New Data Matrix P
Z=scale(X, scale=T) # Standardized Data Matrix
Z
P=Z%*%V3              # PCs Scores
round(P, 3)

#[Step 7] Plot of PCs Scores
par(mfrow=c(2,2))
plot(P[,1], P[, 2], main="Plot of PCs Scores", xlab="1st PC", ylab="2nd PC")
text(P[,1], P[, 2], labels=rownames(P), cex=0.8, col="blue", pos=3)
abline(v=0, h=0)

plot(P[,1], P[, 3], main="Plot of PCs Scores", xlab="1st PC", ylab="3nd PC")
text(P[,1], P[, 3], labels=rownames(P), cex=0.8, col="blue", pos=3)
abline(v=0, h=0)

plot(P[,2], P[, 3], main="Plot of PCs Scores", xlab="2st PC", ylab="3nd PC")
text(P[,2], P[, 3], labels=rownames(P), cex=0.8, col="blue", pos=3)
abline(v=0, h=0)

# PCA based on the SD using princomp( )
par(mfrow=c(1,1))
pca.R<-princomp(X, cor=T)
summary(pca.R, loadings=T) # explanation, coefficient
round(pca.R$scores, 3)  # PC score
screeplot(pca.R, type="lines")

# Principle component biplot (SD)
biplot(pca.R,choices = c(1,2))
abline(v=0, h=0)

biplot(pca.R,choices = c(1,3))
abline(v=0, h=0)

biplot(pca.R,choices = c(2,3))
abline(v=0, h=0)
```

3) 인자 분석 (FA)

이번에는 FA 중에서 PCFA와 MLFA를 실시하고 두 분석방법 간의 결과 차이를 살펴보았습니다.

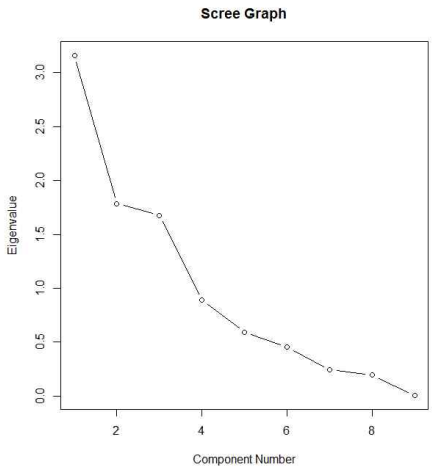
3-1) PCFA

먼저 PCFA를 실시했습니다.

```
> pcfa<-principal(Z, nfactors=3, rotate="varimax")
> pcfa
Principal Components Analysis
Call: principal(r = Z, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
               RC1   RC2   RC3   h2    u2 com
열량          0.40  0.87  0.17 0.96 0.044 1.5
탄수화물      0.77  0.08  0.37 0.74 0.259 1.5
단백질        0.80  0.11 -0.06 0.65 0.351 1.1
지방         -0.05  0.95  0.09 0.92 0.082 1.0
당류         -0.66 -0.48  0.36 0.80 0.203 2.4
나트륨        0.88 -0.04 -0.08 0.78 0.223 1.0
콜레스테롤   -0.18 -0.02  0.74 0.57 0.426 1.1
포화지방산    0.10  0.25  0.78 0.68 0.320 1.2
트랜스지방산  0.37 -0.42  0.47 0.53 0.471 2.9

                       RC1  RC2  RC3
SS loadings           2.77 2.16 1.69
Proportion Var        0.31 0.24 0.19
Cumulative Var        0.31 0.55 0.74
Proportion Explained  0.42 0.33 0.26
Cumulative Proportion 0.42 0.74 1.00

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.09
 with the empirical chi square  16.24  with prob <  0.18

Fit based upon off diagonal values = 0.92> round(pcfa$values, 3)
[1] 3.161 1.783 1.677 0.890 0.593 0.454 0.242 0.195 0.006
```

scree graph와 설명비율의 합 모두 위의 PCA와 유사하게 나왔으므로 같은 이유로 주성분 개수는 3개를 선택했습니다.

```
> gof=pcfa$values/p*100 # Goodness-of fit
> round(gof, 3)
[1] 35.117 19.811 18.634  9.886  6.585  5.048  2.693  2.162  0.064
> sum(gof[1:3])
[1] 73.56152
```



아래의 두 행렬은 인자적재행렬과 잔차행렬 입니다.

```
> L=pcfa$loading[, 1:3]
> round(L, 3)
               RC1    RC2    RC3
열량         0.402  0.874  0.174
탄수화물     0.771  0.081  0.375
단백질       0.795  0.112 -0.059
지방        -0.045  0.953  0.085
당류        -0.663 -0.476  0.362
나트륨       0.877 -0.045 -0.079
콜레스테롤  -0.175 -0.020  0.737
포화지방산   0.096  0.246  0.781
트랜스지방산 0.366 -0.421  0.467
> Psi=pcfa$uniquenesses
> Rm = R-(L%*%t(L) + diag(Psi))
> round(Rm, 3)
               열량 탄수화물 단백질   지방   당류 나트륨 콜레스테롤 포화지방산 트랜스지방산
열량          0.000    0.038 -0.055  0.019  0.023 -0.006     -0.014     -0.068        0.062
탄수화물      0.038    0.000 -0.099 -0.100 -0.006 -0.070      0.069     -0.097       -0.169
단백질       -0.055   -0.099  0.000 -0.011  0.037 -0.098      0.159     -0.020       -0.114
지방          0.019   -0.100 -0.011  0.000  0.038  0.041     -0.076     -0.002        0.157
당류          0.023   -0.006  0.037  0.038  0.000  0.112     -0.117      0.060       -0.059
나트륨       -0.006   -0.070 -0.098  0.041  0.112  0.000     -0.102      0.113       -0.038
콜레스테롤   -0.014    0.069  0.159 -0.076 -0.117 -0.102      0.000     -0.262       -0.177
포화지방산   -0.068   -0.097 -0.020 -0.002  0.060  0.113     -0.262      0.000       -0.050
트랜스지방산  0.062   -0.169 -0.114  0.157 -0.059 -0.038     -0.177     -0.050        0.000
```

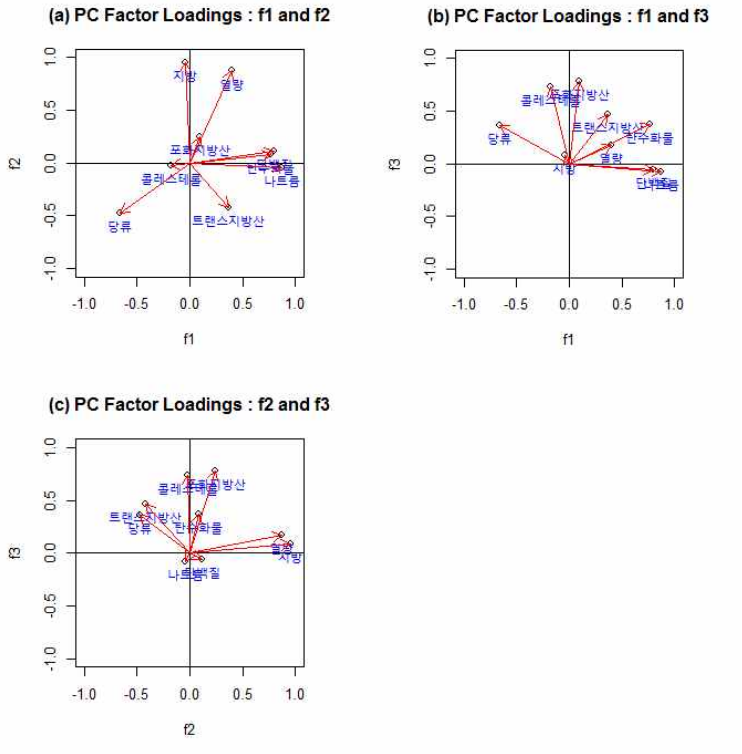
PCFA에서 인자적재를 살펴보면,

공통인자 f_1 : (+)부호의 인자적재값은 탄수화물(0.771), 단백질(0.795), 나트륨(0.877)로 가장 다른 인자들에 비해 크고 (-)부호의 인자적재값은 당류(-0.663)으로 다른 인자들에 비해 크기 때문에, **공통인자 f_1은 (탄수화물, 단백질, 나트륨)과 당류와의 대비를 나타내는 공통인자라고 해석할 수 있다. 특히 이것 또한 해석의 편의성을 위해 짠맛과 단맛의 대비로 해석할 수도 있습니다.**

공통인자 f_2 : (+),(-) 부호는 고루 있으나 (+)부호에서 열량(0.874)과 지방(0.953)의 인저적재값이 월등히 크기 때문에 **공통인자 f_2는 열량과 지방에 대한 공통인자라고 해석할 수 있습니다.**

공통인자 f_3 : (+)부호가 대부분이고 인자적재값 또한 콜레스테롤(0.737)과 포화지방산(0.781), 트랜스지방산(0.467)이 크기 때문에 **공통인자 f_3는 콜레스테롤, 포화지방산, 그리고 트랜스지방산에 대한 공통인자라고 해석할 수 있습니다.**

다음은 해석을 위해 인자적재그림을 그려보았습니다. 해석의 용의성을 위해 varimax 인자회전을 적용했습니다.



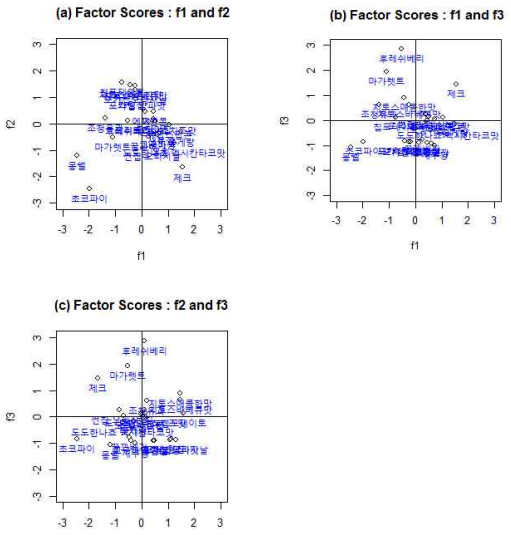
(a) PC Factor Loadings : f1 and f2를 통해,
(단백질, 탄수화물, 나트륨)이 f1에 대하여 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다. (당류) 또한 위의 세 변수 만큼은 아니지만 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다. (지방, 열량)이 f2에 대하여 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다.

(b) PC Factor Loadings : f1 and f3를 통해,
(포화지방산, 콜레스테롤)이 f3에 대하여 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다.

(c) PC Factor Loadings : f2 and f3를 통해,
위에서와 동일하게 (지방,열량)이 f2에 대하 높은 상관,영향력을 가지고 (포화지방산, 콜레스테롤)이 f3에 대하여 높은 상관, 영향력을 가짐을 확인할 수 있습니다.

다음으로는 인자점수그림을 통해 각 관측치들의 군집 형성과 특성을 살펴보았습니다.

```
> round(fpc, 3)
                      RC1    RC2    RC3
치토스매콤한맛     -0.441  1.463  0.901
치토스바베큐맛     -0.262  1.450  0.631
도리토스나쵸치즈맛  0.504  0.126  0.138
도도한나쵸 멕시칸타코맛  1.221 -0.719 -0.213
썬칩 오리지날       0.381 -0.851  0.251
스윙칩             -0.392  1.097 -0.837
칩포테이토         -0.763  1.562  0.133
포카칩오리지날     -0.242  1.285 -0.866
포카칩 양파맛      -0.202  1.078 -0.872
뉴팝스팝콘          1.050 -0.087  0.115
조청유과           -1.381  0.206  0.618
꿀꽈배기            0.238 -0.479 -0.764
고구마깡            0.590 -0.411 -0.906
새우깡              0.732 -0.356 -1.007
오잉                0.710 -0.675  0.053
오징어집            0.453  0.167 -0.017
꿀게랑              1.494 -0.304 -0.143
초코파이           -1.966 -2.457 -0.847
몽쉘               -2.468 -1.200 -1.068
마가렛트           -1.095 -0.523  1.938
후레쉬베리         -0.535  0.108  2.867
에이스              0.129  0.456 -0.896
제크                1.548 -1.659  1.443
콘칲                0.246  0.097  0.248
꼬깔콘              0.451  0.475 -0.901
```



(a) Factor Scores : f1 and f2

: 주로 제 2,4 사분면에 밀집되어 있음을 확인할 수 있습니다.

감자칩류와 치토스는 f1 값이 (-)이지만 0 근처이고, f2값이 높음을 알 수 있습니다.

즉 짠맛과 단맛이 공존하면서, 지방과 열량이 높은 종류의 군집으로 해석할 수 있습니다.

파이류는 f1값이 매우 낮고, f2값도 낮음을 확인할 수 있습니다.

즉 단맛을 주로 가지고, 지방과 열량은 낮은 종류의 군집으로 해석할 수 있습니다.

그 외의 과자들의 군집은 열량이 크게 높지 않음을 알 수 있습니다.

(b) Factor Scores : f1 and f3

: 주로 제 2,4 사분면에 밀집되어 있음을 확인할 수 있습니다.

낱개형과자류(후레쉬베리,마가렛트)가 (콜레스테롤, 포화지방산)이 높게 함유되어 있음을 확인할 수 있고, 파이류는 (콜레스테롤, 포화지방산)이 적게 함유되어 있음을 확인할 수 있습니다. "제크"의 경우에는 짠맛이 강한 과자 중 독특하게 (콜레스테롤, 포화지방산)이 높음을 알 수 있습니다.

그 외의 과자들은 큰 하나의 군집을 이루어 원점 근처에서 타원형을 이루고 있습니다.

(c) Factor Scores : f2 and f3

: **주로 제 1,3 사분면에 밀집되어 있음을 확인할 수 있습니다.**

낱개형과자류와 제크는 역시 (콜레스테롤, 포화지방산)이 많이 함유된 군집으로 나뉘고, 파이류는 (지방,열량)이 낮게 함유되면서 (콜레스테롤, 포화지방산)이 낮게 함유된 군집을 이루고 있습니다.

간자칩류는 대부분 (지방, 열량)이 높고, (콜레스테롤, 포화지방산)은 크게 높지 않음을 알 수 있습니다. 치토스를 비롯한 그 외의 군집은 타원형으로 하나의 큰 군집을 이룸을 확인할 수 있습니다.

3-2) MLFA

이번에는 MLFA를 시행해보았습니다. 이후 시각화를 할 경우에 해석의 용의성을 위해 varimax를 적용했습니다.

```
> mlfa

Call:
factanal(x = Z, factors = 3, scores = "regression", rotation = "varimax")

Uniquenesses:
    열량   탄수화물     단백질       지방       당류     나트륨 콜레스테롤 포화지방산 트랜스지방산
   0.005    0.005    0.558    0.005    0.233    0.522    0.688    0.522    0.868

Loadings:
           Factor1 Factor2 Factor3
열량        0.478   0.853   0.197
탄수화물    0.850           0.521
단백질      0.654
지방                0.997
당류       -0.713  -0.380   0.337
나트륨      0.690
콜레스테롤                  0.553
포화지방산          0.264   0.637
트랜스지방산 0.142  -0.228   0.244

               Factor1 Factor2 Factor3
SS loadings      2.395   1.996   1.208
Proportion Var   0.266   0.222   0.134
Cumulative Var   0.266   0.488   0.622

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 12.08 on 12 degrees of freedom.
The p-value is 0.443
```

Factor1 : (탄수화물, 단백질, 나트륨)와 (당류)간의 대비를 나타내는 인자임을 알 수 있습니다. 이후 해석의 용의성을 위해 단맛과 짠맛에 대한 대비로도 유추해낼 수 있습니다.

Factor2 : (열량, 지방)에 대한 인자임을 알 수 있습니다.

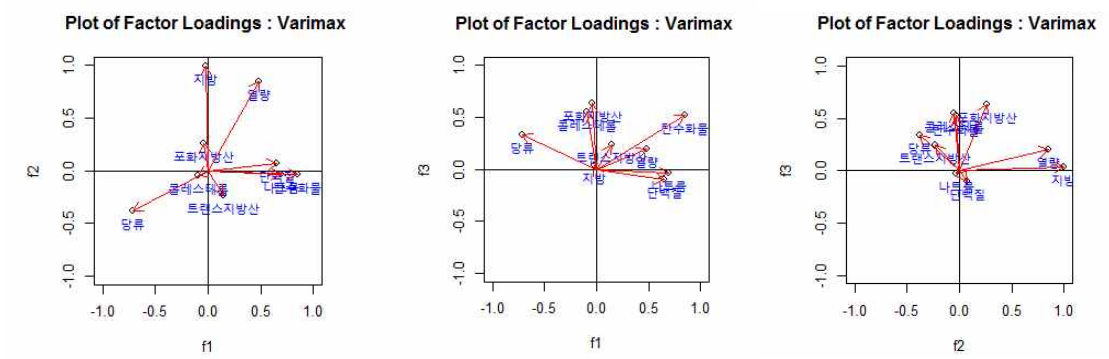
Factor3 : (탄수화물, 콜레스테롤, 포화지방산)에 대한 인자임을 알 수 있습니다.

=> 각 Factor에 대한 해석은 위의 PCFA와 거의 유사함을 알 수 있습니다.

=> **factor1에서는 당류에 대한 값이 PCFA보다 높게 나왔으므로 나트륨과 대비를 나타내기에는 PCFA보다 적합하다고 말할 수 있습니다.**

=> **그러나 직관적으로 factor3의 경우는 MLFA에서는 PCFA보다 더 많은 변수에 대한 설명력이 사용되었음을 알 수 있습니다. 이를 통해 이러한 측면에서 PCFA가 더 이 데이터를 잘 설명한다고 말할 수 있습니다.**

다음은 해석을 위해 인자적재그림을 그려보았습니다. 해석의 용의성을 위해 varimax 인자회전을 적용했습니다.

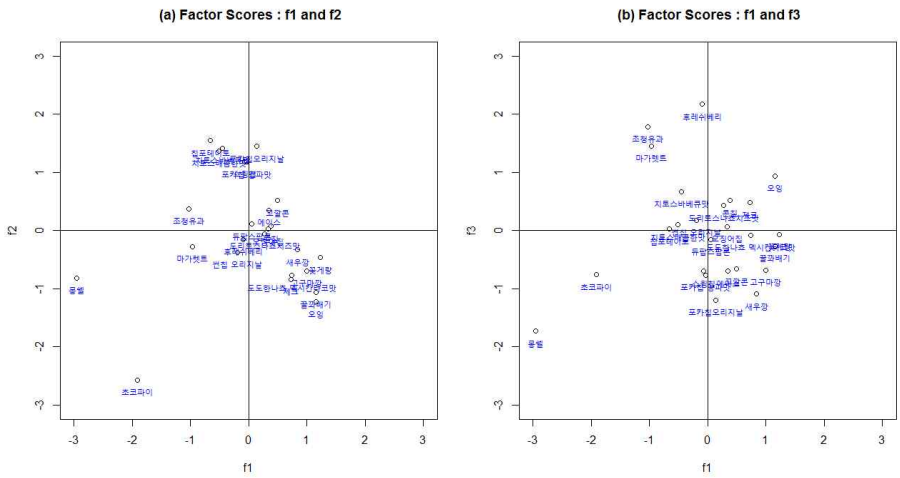


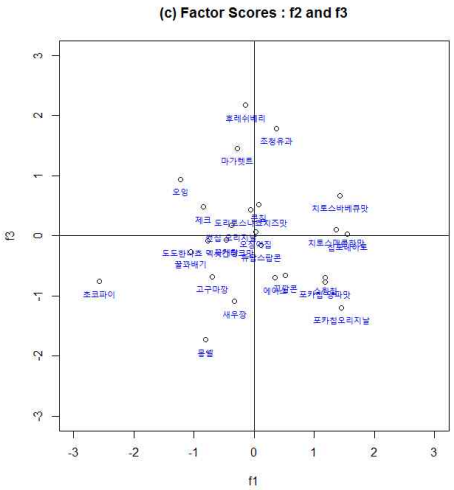
(a) PC Factor Loadings : f1 and f2를 통해,
(단백질, 탄수화물, 나트륨)이 f1에 대하여 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다. (당류) 또한 위의 세 변수 만큼 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다. (지방, 열량)이 f2에 대하여 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다.

(b) PC Factor Loadings : f1 and f3를 통해,
(포화지방산, 콜레스테롤)이 f3에 대하여 높은 상관관계와 높은 영향력을 가짐을 확인할 수 있습니다. (탄수화물)은 f1, f3 둘 다에 대하여 높은 상관관계와 영향력을 가짐을 확인할 수 있습니다.

(c) PC Factor Loadings : f2 and f3를 통해,
위에서와 동일하게 (지방,열량)이 f2에 대하 높은 상관,영향력을 가지고 (탄수화물, 포화지방산, 콜레스테롤)이 f3에 대하여 높은 상관, 영향력을 가짐을 확인할 수 있습니다.

다음은 factor score로 plot을 그려보았습니다.





(a) Factor Scores : f1 and f2
: 감자칩류와 치토스는 f1 값이 (-)이지만 0 근처이고, f2값이 높음을 알 수 있습니다.
즉 짠맛과 단맛이 공존하면서, 지방과 열량이 높은 종류의 군집으로 해석할 수 있습니다.
파이류는 f1값이 매우 낮고, f2값도 낮음을 확인할 수 있습니다.
즉 단맛을 주로 가지고, 지방과 열량은 낮은 종류의 군집으로 해석할 수 있습니다.
그 외의 과자들의 군집은 열량이 크게 높지 않음을 알 수 있습니다.
=> PCFA와 해석은 동일하나 그림상에서 점들이 더욱 밀집되어 있음을 확인할 수 있습니다.

(b) Factor Scores : f1 and f3
: 낱개형과자류(후레쉬베리,마가렛트)와 조청유과가 (탄수화물, 콜레스테롤, 포화지방산)이 높게 함유되어 있음을 확인할 수 있고, 파이류는 (탄수화물, 콜레스테롤, 포화지방산)이 적게 함유되어 있음을 확인할 수 있습니다. "제크"의 경우에는 짠맛이 강한 과자 중 독특하게 (탄수화물, 콜레스테롤, 포화지방산)이 높음을 알 수 있습니다.
그 외의 과자들은 큰 하나의 군집을 이루어 원점 근처에서 타원형을 이루고 있습니다.
=> PCFA와 해석은 유사하나 그림상에서 점들이 다소 퍼져 있음을 확인할 수 있습니다.

(c) Factor Scores : f2 and f3
: **(**낱개형과자류, 조청유과)와 제크는 역시 (탄수화물, 콜레스테롤, 포화지방산)이 많이 함유된 군집으로 나뉘고, 파이류는 (지방,열량)이 낮게 함유되면서 (탄수화물, 콜레스테롤, 포화지방산)이 낮게 함유된 군집을 이루고 있습니다.
감자칩류는 대부분 (지방, 열량)이 높고, (탄수화물, 콜레스테롤, 포화지방산)은 크게 높지 않음을 알 수 있습니다.
=> PCFA와 해석은 유사하나 그림상에서 원점 주변을 제외하고는 직관적으로 군집을 파악하기에는 다소 어려울 정도로 퍼져있음을 확인할 수 있습니다.

3-3) PCFA, MLFA 비교

PCFA

```
> pcfa
Principal Components Analysis
Call: principal(r = Z, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                 RC1   RC2   RC3   h2    u2 com
열량            0.40  0.87  0.17 0.96 0.044 1.5
탄수화물        0.77  0.08  0.37 0.74 0.259 1.5
단백질          0.80  0.11 -0.06 0.65 0.351 1.1
지방           -0.05  0.95  0.09 0.92 0.082 1.0
당류           -0.66 -0.48  0.36 0.80 0.203 2.4
나트륨          0.88 -0.04 -0.08 0.78 0.223 1.0
콜레스테롤     -0.18 -0.02  0.74 0.57 0.426 1.1
포화지방산      0.10  0.25  0.78 0.68 0.320 1.2
트랜스지방산    0.37 -0.42  0.47 0.53 0.471 2.9

                       RC1  RC2  RC3
SS loadings           2.77 2.16 1.69
Proportion Var        0.31 0.24 0.19
Cumulative Var        0.31 0.55 0.74
Proportion Explained  0.42 0.33 0.26
Cumulative Proportion 0.42 0.74 1.00

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.09
 with the empirical chi square  16.24  with prob <  0.18

Fit based upon off diagonal values = 0.92
```

```
> Psi
      열량     탄수화물      단백질        지방        당류      나트륨  콜레스테롤  포화지방산 트랜스지방산
 0.04360679 0.25921596 0.35138525 0.08195493 0.20284990 0.22337347 0.42596150 0.32005937 0.47115570
> Rm
                  열량       탄수화물        단백질          지방          당류          나트륨  콜레스테롤  포화지방산 트랜스지방산
열량       -6.661338e-16  3.785018e-02 -5.531331e-02  1.874169e-02  2.281133e-02 -6.338809e-03 -0.01421780 -0.068219142  6.151349e-02
탄수화물    3.785018e-02 -4.440892e-16 -9.905054e-02 -9.964190e-02 -5.901202e-03 -7.028208e-02  0.06886923 -0.096994675 -1.694861e-01
단백질     -5.531331e-02 -9.905054e-02 -4.440892e-16 -1.083841e-02  3.682984e-02 -9.783203e-02  0.15906054 -0.020499368 -1.142912e-01
지방        1.874169e-02 -9.964190e-02 -1.083841e-02 -4.440892e-16  3.835309e-02  4.099074e-02 -0.07626137 -0.001741386  1.568539e-01
당류        2.281133e-02 -5.901202e-03  3.682984e-02  3.835309e-02 -6.661338e-16  1.123124e-01 -0.11748504  0.059714207 -5.895701e-02
나트륨     -6.338809e-03 -7.028208e-02 -9.783203e-02  4.099074e-02  1.123124e-01 -4.440892e-16 -0.10139649  0.113067909 -3.847825e-02
콜레스테롤 -1.421780e-02  6.886923e-02  1.590805e-01 -7.626137e-02 -1.174850e-01 -1.013965e-01  0.00000000 -0.251524453 -1.774974e-01
포화지방산 -6.821914e-02 -9.699468e-02 -2.043997e-02 -1.741366e-03  5.971421e-02  1.130679e-01 -0.25152445  0.000000000 -4.952501e-02
트랜스지방산 6.151349e-02 -1.694861e-01 -1.142912e-01  1.568539e-01 -5.895701e-02 -3.847825e-02 -0.17749741 -0.049525007 -4.440892e-16
```

MLFA

```
> mlfa

Call:
factanal(x = Z, factors = 3, scores = "regression", rotation = "varimax")

Uniquenesses:
   열량 탄수화물   단백질    지방    당류  나트륨 콜레스테롤 포화지방산 트랜스지방산
  0.005   0.005   0.558   0.005   0.233   0.522   0.683   0.522   0.868

Loadings:
             Factor1 Factor2 Factor3
열량          0.478   0.853   0.197
탄수화물      0.850           0.521
단백질        0.654
지방                  0.997
당류         -0.713  -0.380   0.337
나트륨        0.690
콜레스테롤                    0.553
포화지방산            0.264   0.637
트랜스지방산  0.142  -0.228   0.244

                Factor1 Factor2 Factor3
SS loadings       2.395   1.995   1.208
Proportion Var    0.266   0.222   0.134
Cumulative Var    0.266   0.488   0.622

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 12.03 on 12 degrees of freedom.
The p-value is 0.443
```

```
> Psi
    열량   탄수화물    단백질      지방      당류    나트륨 콜레스테롤 포화지방산 트랜스지방산
0.005000 0.005000 0.5577262 0.005000 0.2330296 0.5223685 0.6826999 0.5218398  0.8684313
> Rm
                  열량       탄수화물        단백질          지방          당류          나트륨  콜레스테롤  포화지방산 트랜스지방산
열량       -1.449635e-05 -3.332593e-04 -1.402863e-03  2.838898e-04  1.978963e-03 -1.260314e-03  2.204634e-03 -8.089972e-03  1.230917e-04
탄수화물   -3.332593e-04 -3.499115e-05 -1.156244e-03  1.682898e-04 -9.499722e-04 -4.587675e-04  5.164773e-04  2.758708e-03 -1.649397e-03
단백질     -1.402863e-03 -1.156244e-03 -3.805370e-06  1.921907e-04 -3.858802e-02  1.477678e-01  9.509675e-02  1.110927e-01  4.923112e-02
지방        2.838898e-04  1.682898e-04  1.921907e-04 -7.493017e-05 -1.268790e-03  5.903906e-04 -1.716231e-03  7.084830e-03 -5.464393e-04
당류        1.978963e-03 -9.499722e-04 -3.858802e-02 -1.268790e-03 -4.082772e-03  1.188197e-02  2.097536e-03  1.516372e-02  4.345510e-04
나트륨     -1.260314e-03 -4.587675e-04  1.477678e-01  5.903906e-04  1.188197e-02  3.055952e-06 -2.323241e-01  1.826049e-01  1.648603e-01
콜레스테롤  2.204634e-03  5.164773e-04  9.509675e-02 -1.716231e-03  2.097536e-03 -2.323241e-01  5.503121e-06 -5.270976e-02 -2.058417e-02
포화지방산 -8.089972e-03  2.758708e-03  1.110927e-01  7.084830e-03  1.516372e-02  1.826049e-01 -5.270976e-02 -1.330318e-06  1.579235e-01
트랜스지방산 1.230917e-04 -1.649397e-03  4.923112e-02 -5.464393e-04  4.345510e-04  1.648603e-01 -2.058417e-02  1.579235e-01  6.561372e-06
```

위의 표로 비교했을 때,
잔차행렬의 경우는 PCFA, MLFA 둘 다 별 차이가 없는 것으로 보이며, 총기여율의 경우는
PCFA 총기여율 : 약 0.74%
MLFA 총기여율 : 약 0.62%
인자 적재값은 각 기법의 f1에 대해서 MLFA는 나트륨과 당류에 대한 대비가 더욱 잘 나타나는 장점이 있으나 f3에 대해서 영향력 높은 변수가 "탄수화물"이 추가되어 해석에 있어 더 모호해집니다.
=> 결론 : PCFA가 더 적합한 기법으로 판단됩니다.

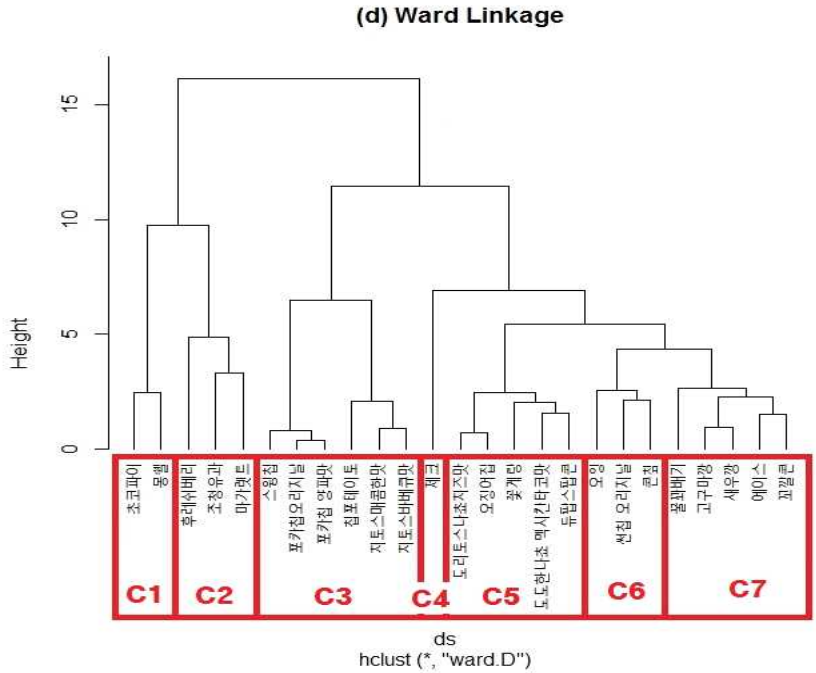
```
##PCFA##

library(psych)
pcfa<-principal(Z, nfactors=3, rotate="varimax")
pcfa
round(pcfa$values, 3)
gof=pcfa$values/p*100 # Goodness-of fit
round(gof, 3)
sum(gof[1:3])

# Residual Matrix
L=pcfa$loading[, 1:3]
round(L, 3)
Psi=pcfa$uniquenesses
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm, 3)

# Plot of PC Factor Loadings (171p)
par(mfrow=c(2,2))
lim<-range(pretty(L))
plot(L[,1], L[,2],main="(a) PC Factor Loadings : f1 and f2", xlab="f1", ylab="f2", xlim=lim, ylim=lim)
text(L[,1], L[, 2], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[,2], col=2, code=2, length=0.1)

plot(L[,1], L[,3],main="(b) PC Factor Loadings : f1 and f3", xlab="f1", ylab="f3", xlim=lim, ylim=lim)
text(L[,1],L[,3],labels=rownames(L),cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[,3], col=2, code=2, length=0.1)
plot(L[,2], L[,3],main="(c) PC Factor Loadings : f2 and f3", xlab="f2", ylab="f3", xlim=lim, ylim=lim)
text(L[,2], L[,3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,2], L[,3], col=2, code=2, length=0.1)

# Factor Scores : Regression Method
fpc=pcfa$scores
round(fpc, 3)

# Plot of Factor Scores : PFA (173p)
par(mfrow=c(1,1))
par(pty="s")
lim<-range(pretty(fpc))
plot(fpc[,1], fpc[,2],main=" (a) Factor Scores : f1 and f2", xlab="f1", ylab="f2",
     xlim=lim, ylim=lim)
text(fpc[,1], fpc[,2], labels=rownames(fpc), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fpc[,1], fpc[,3],main=" (b) Factor Scores : f1 and f3", xlab="f1", ylab="f3", xlim=lim, ylim=lim)
text(fpc[,1], fpc[,3], labels=rownames(fpc), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fpc[,2], fpc[,3],main="(c) Factor Scores : f2 and f3", xlab="f2", ylab="f3", xlim=lim, ylim=lim)
text(fpc[,2], fpc[,3], labels=rownames(fpc), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
```

```
##MLFA##

# ML Estimation using the factanal( )
mlfa<-factanal(Z, factors = 3, rotation="varimax", score="regression")
mlfa

# Residual Matrix
L=mlfa$loading[, 1:3]
L
Psi=mlfa$uniquenesses
Rm = R-(L%*%t(L) + diag(Psi))
round(Rm, 3)

# Factor Loadings Plot(after rotation)
par(mfrow=c(2,2))
# par(mfrow=c(1,1))

lim<-range(pretty(L))
plot(L[,1], L[,2],main="Plot of Factor Loadings : Varimax ", xlab="f1", ylab="f2",xlim=lim, ylim=lim)
text(L[,1],L[,2],labels=rownames(L),cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[,2], col=2, code=2, length=0.1)

plot(L[,1], L[,3],main="Plot of Factor Loadings : Varimax ", xlab="f1", ylab="f3",xlim=lim, ylim=lim)
text(L[,1], L[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
arrows(0,0, L[,1], L[,3], col=2, code=2, length=0.1)

plot(L[,2], L[,3],main="Plot of Factor Loadings : Varimax ", xlab="f2", ylab="f3",xlim=lim, ylim=lim)
text(L[,2], L[, 3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

arrows(0,0, L[,2], L[,3], col=2, code=2, length=0.1)
# Factor Scores : Regression Method
fml=mlfa$scores
round(fml, 3)

# Plot of Factor Scores : MLFA
par(mfrow=c(1,1))
par(pty="s")
lim<-range(pretty(fml))
plot(fml[,1], fml[,2],main=" (a) Factor Scores : f1 and f2", xlab="f1", ylab="f2", xlim=lim, ylim=lim)
text(fml[,1], fml[,2], labels=rownames(fml),cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fml[,1], fml[,3],main=" (b) Factor Scores : f1 and f3", xlab="f1", ylab="f3", xlim=lim, ylim=lim)
text(fml[,1], fml[,3],labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)

plot(fml[,2], fml[,3],main="(c) Factor Scores : f2 and f3", xlab="f1", ylab="f3", xlim=lim, ylim=lim)
text(fml[,2],fml[,3], labels=rownames(fml), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
```

4) 군집 분석 (CA)

4-1) Hierarchical CA

먼저 Hierarchical CA를 실시했습니다. 그 중에서도 군집평균과 개체 간의 ESS에 의해 정보손실을 측정하고 이를 최소하하는 군집을 병합하는 방법인 "와드연결법"을 사용했습니다.



와드연결법의 경우에는 위의 PCA, FA를 통해 대략적으로 나눠보았던 군집과 유사하게 나왔으며 더욱 디테일하게 군집화가 되어있었다.

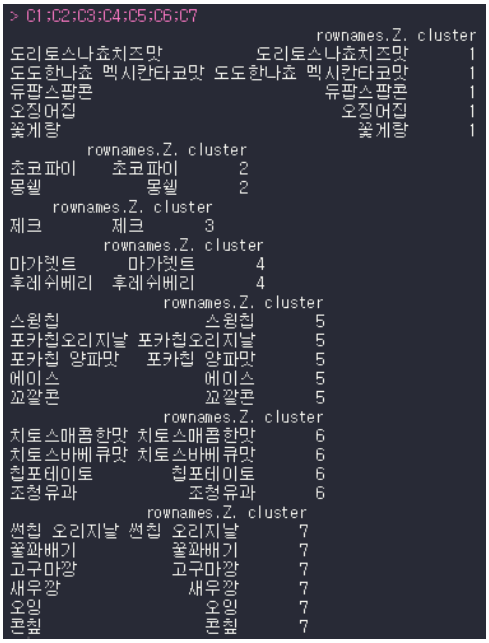
군집	개체	군집특성
C1	**초코파이 몽쉘**	파이류, 당류가 높게 함유되어 단맛이 주로 나는 과자. 콜레스테롤, 지방, 열량, 탄수화물이 적게 함유되어 있음.
C2	**후레쉬베리 조청유과 마가렛트**	낱개형 과자와 조청유과, 단맛이 나는 과자. (탄수화물, 콜레스테롤, 포화지방산)이 높게 함유되어 있음.
C3	**스윙칩 포카칩오리지날 포카칩 양파맛 칩 포테이토 치토스 매콤한맛 치토스 바베큐맛**	감자칩과 치토스, 짠맛과 단맛이 공존하면서, 지방과 열량이 높음.
C4	**제크**	짠맛이 강한 과자 중 독특하게 (탄수화물, 콜레스테롤, 포화지방산)이 높음.
C5	**도리토스 나쵸치즈맛 오징어집 꽃게랑 도도한나쵸 멕시칸타코맛 듀팝스팝콘**	다른 과자들에 비해 짠 과자에 속함. 열량,지방은 다소 적은 편이고 탄수화물, 콜레스테롤, 포화지방산은 보통 수준임.
C6	**오잉 썬칩 오리지날 콘칲**	다른 과자들에 비해 덜 짠 과자에 속함. 열량,지방은 다소 적은 편이고 탄수화물, 콜레스테롤, 포화지방산은 보통 수준임.
C7	**꿀꽈배기 고구마깡 새우깡 에이스 꼬깔콘**	열량, 지방은 보통 수준이며, 탄수화물, 콜레스테롤, 포화지방산은 적게 포함됨.

결론 : 위의 R –Tech.들을 통해 차원축소 시켜 분류했던 관측치들의 군집들이 CA를 통해 만들어진 군집과 거의 유사하며 C5 ~ C7의 경우는 더욱 더 디테일하게 군집화 됐습니다.

4-2) Non-Hierarchical CA

Non-Hierarchical CA에서는 자료의 측정변수가 모두 연속형이고 outlier로 판단되는 관측치가 없으므로 K-평균법을 사용했습니다. 이 때, 와드연결법에서 7개의 군집으로 판단함으로써 이전에 R-Tech.를 통해 제가 판단했던 군집보다 더욱 정교하게 군집을 나눌 수 있었기에 7개의 군집으로 동일한 조건하에 시행했습니다.

```
> kmeans <- kmeans(Z, 7) # 7 cluster solution
> cluster=data.frame(rownames(Z),cluster=kmeans$cluster)
> C1=cluster[(cluster[,2]==1),]
> C2=cluster[(cluster[,2]==2),]
> C3=cluster[(cluster[,2]==3),]
> C4=cluster[(cluster[,2]==4),]
> C5=cluster[(cluster[,2]==5),]
> C6=cluster[(cluster[,2]==6),]
> C7=cluster[(cluster[,2]==7),]
```



결론 : 그 결과, cluster2(파이류), cluster4(낱개포장류)를 제외하면 기존의 분류된 군집들과는 매우 다르게 군집화가 되어 분석 목적에 맞지 않는 군집이 형성되었습니다.

4-3) 와드연결법, K-평균법 비교

같은 군집의 수로 비교해보았을 때, Hierarchical CA의 '와드연결법'이 더욱 분석 목적에 맞는 군집화가 진행되었다고 판단했습니다. 왜냐하면 R-Tech.를 통해 제가 판단했던 군집과 유사하며 오히려 더욱 정교하게 군집화가 되었기 때문입니다.

결론 : 와드연결법을 통해 얻은 C1 ~ C7로 군집이 적합한 군집.

```
## Hierarchical CA
# 표준화 유클리드거리
ds <- dist(Z, method="euclidean")
round(ds, 3)

#와드연결법
ward=hclust(ds, method="ward.D")
plot(ward, hang=-1, main="(d) Ward Linkage")


## Non-Hierarchical CA
#All Indices
library(NbClust)
allindex<-NbClust(Z,distance="euclidean", min.nc = 2, max.nc = 8,method = "kmeans", index = "all", )
rownames(Z)
```

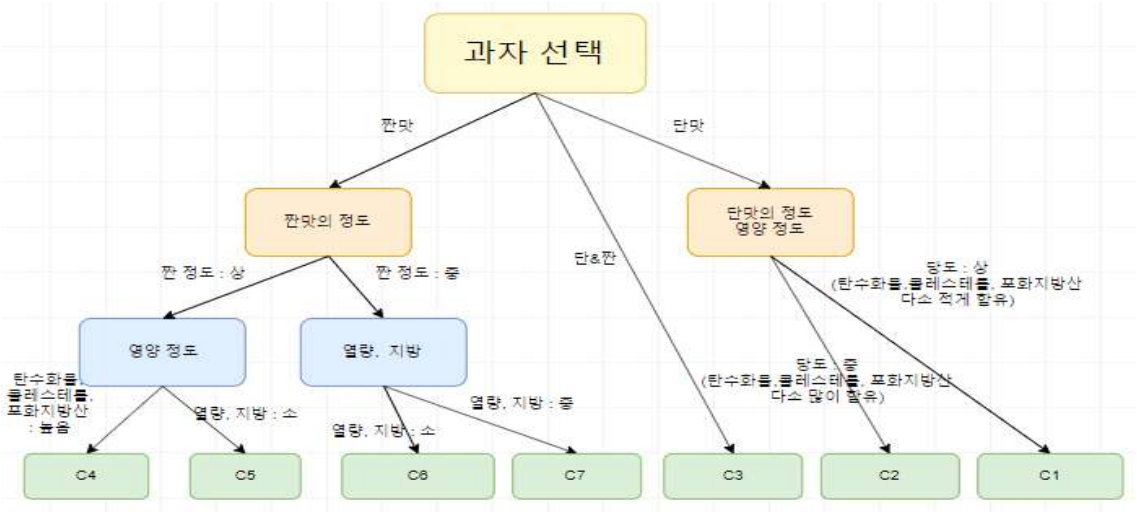
```
# K-means Method
kmeans <- kmeans(Z, 7) # 7 cluster solution
cluster=data.frame(rownames(Z),cluster=kmeans$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C5=cluster[(cluster[,2]==5),]
C6=cluster[(cluster[,2]==6),]
C7=cluster[(cluster[,2]==7),]

C1;C2;C3;C4;C5;C6;C7

# Get cluster means
aggregate(X, by=list(kmeans$cluster),FUN=mean)
```

4. 결론

시중에 판매 중인 과자들 중 제가 먹어본 과자를 바탕으로 그들의 영양 성분표를 이용하여 과자를 군집화 해보았습니다. 이를 통해 원하는 과자를 선택함에 있어서 군집화 된 과자 데이터를 통해 선택의 폭을 줄일 수 있도록 시도해보았습니다.

결과적으로는 와드연결법을 통해 얻은 군집화 자료를 가지고 가장 효율적인 과자 선택을 할 수 있는 알고리즘을 만들었습니다.

과자를 선택할 때는 먼저 자신이 원하는 맛의 구분(짠맛/단맛)을 정하고, 이후 그 맛의 정도와 영양정도를 통해 자신의 원하는 과자를 최소한의 카테고리 내에 결정합니다.



C1	**파이류**	**C4**	**제크**
C2	**낱개포장류** **조청유과**	**C5**	**도리토스 나쵸치즈맛** **오징어집** **꽃게랑** **도도한나쵸 멕시칸타코맛** **듀팝스팝콘**
C3	**감자칩류, 치토스**	**C6**	**오잉** **썬칩 오리지날** **콘칲**

5. 참고자료

1) 분석 방법 참고 : R과 함께하는 다변량 자료분석, 최용석 지음, KM 경문사

2) 코드 참고 : https://stat.pusan.ac.kr/stat/49709/subview.do

3) 데이터 출처 : http://www.foodsafetykorea.go.kr

4) 상품 출처 : 오리온 (http://www.orionworld.com/Snak/company/ceo01.asp)
Kraft (http://www.kraftheinzcompany.com/)
농심 (www.nongshim.com)
해태제과 (http://www.ht.co.kr/)
롯데제과 (http://www.lotteconf.co.kr/)
크라운제과 (http://www.crown.co.kr/)
빙그레 (http://www.bing.co.kr/)