2021-1 YDMS 3주차 과제

〈Subject : 차원축소〉

김하은 2019251034

1. 비즈니스 애널리틱스를 위한 데이터마이닝 (R EDITION) 차원 축소 부분 충분히 읽고 공부해보기

차원이란 공간 내의 점을 저장하는 데 필요한 축의 개수라고 생각할 수 있다. 이때 각 축은 데이터를 설명하기 위한 축이므로 변수라고 생각할 수 있다. 따라서 차원은 변수의 개수라고도 표현한다. 즉 '차원을 축소한다'는 '변수를 줄인다'와 같은 의미이다. 그렇다면 우리는 왜 변수의 개수를 줄여야하는가?

우리가 가진 변수의 개수가 너무 많은 경우, 변수의 정보를 모두 이용하는 것은 불가능하거나 불필요하다. 또한 비지도 학습 알고리즘 및 지도 학습 알고리즘에서 계산 문제가 발생하기 쉬우며, 회귀분석의 경우 변수의 개수가 관찰값의 개수보다 많으면 분석이 불가능한 경우가 발생한다. 그렇지 않다고 해도 변수가 너무 많은 경우는 비효율적이며 모델 성능에 좋은 결과를 가져오는 것도 아니다. 오히려 변수의 개수를 줄여, 모형의 복잡도를 낮춤으로써 모델 성능을 개선할 수 있다. 무엇보다 고차원에서는 2번에서 설명할 차원의 저주가 발생하기 때문에 차원을 줄여야 한다.

위에서 언급한 이유로 변수가 너무 많은 경우에는 차원축소를 진행하게 된다. 차원축소는 관측이 불가능한 대한 대상을 측정할 수 있기 때문에 심리학 같은 연구가 가능하며, 적은 변수로도 원하는 대상을 측정할 수 있게 한다는 장점이 있다. 그러나 차원축소도 항상 좋은 결과만을 보이는 것은 아니다. 차원축소 과정에서 정보의 손실이 발생할 수 있으며, 차원축소로 인해 과적합이 발생할 수 있기 때문이다. 또한 설명 불가능한 경우 연구자의 의도대로 결과를 조절할 수 있기 때문에 편협된 결과를 초래할 수 있다.

한편 차원축소의 기본적인 틀은 다음과 같다.

① Feature Selection (FS) "변수 선택"

: 말 그대로 가지고 있는 변수 중 중요한 변수만 몇 개 고르고, 나머지는 버리는 방법이다. 변수 간 중첩이 있는지, 어떤 변수가 중요한 변수인지, 타겟 변수에 주영향을 미치는 변수가 어떤 것인지를 파악하는 것이 중요한 방법이다. 따라서 상관분석을 주로 이용하게 된다.

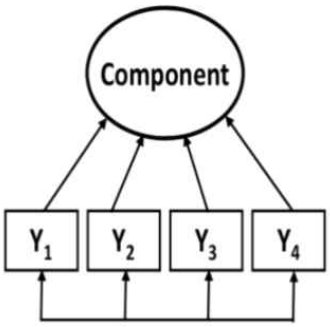
② Feature Extraction (FE) "변수 추출"

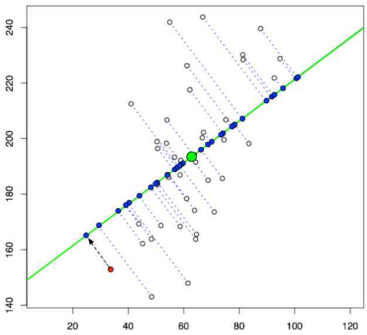
: 모든 변수를 조합하여 이 데이터를 잘 표현할 수 있는 중요 성분을 가진 새로운 변수를 추출하는 방법이다. 예를 들어, 국어 점수, 영어 점수, 독일어 점수를 묶어 언어 능력이라는 새로운 변수를 추출하는 방식이다.

③ 불필요한 변수 삭제

기본적인 틀을 바탕으로 만들어진 차원 축소 기법은 다음과 같다.

① PCA (Pricipal Component Analysis) – 주성분 분석





: 저차원의 초평면에 투영시키는 차원 축소 방법으로, 데이터를 정사영한다고 생각하면 쉽게 이해할 수 있다. 즉 데이터를 가장 잘 나타내는 PC축을 찾아 그 축 위에 데이터를 정사영한다고 생각하면 된다. 주성분 분석의 목적은 고차원의 데이터를 저차원으로 줄이는 것으로 공통된 상관관계가 높은 변수들을 줄여서 주성분을 찾는 것이다. 따라서 사용된 변수의 개수보다 주성분의 개수는 적어야 효과적이라 할 수 있다.

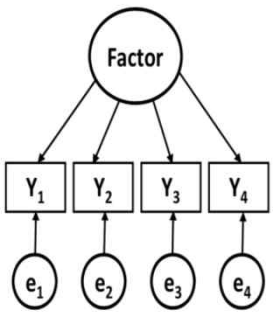
PCA에서는 분산이 가장 커지는 축을 첫 번째 주성분으로 하고, 분산이 두 번째로 커지는 축을 두 번째 주성분으로 한다. 이때 각 주성분은 공분산 행렬의 고유벡터이므로 서로 간 직교하는 특징을 가지고 있다. 주성분 분석은 가장 큰 분산을 갖는 부분 공간을 보존하는 최적의 선형변환으로 적은 성분으로 전체 분산을 설명하기에 적합한 방식이다.

주성분 분석을 진행할 때엔 가장 먼저 표준화 혹은 정규화를 통해 scaling을 해주어야 한다. 변수의 스케일 차이로 인한 왜곡을 막기 위함이다.

② LLE (Locally Linear Embedding)

: PCA와는 다르게 비선형 차원 축소 기법으로 투영(정사영)에 의존하지 않는 방법이다. 가까운 k개의 이웃들로부터 선형적으로 연관된 정도를 측정하여 관계가 가장 잘 보존되는 저차원을 찾는 방식이다.

③ FA (Factor Analysis) – 요인 분석

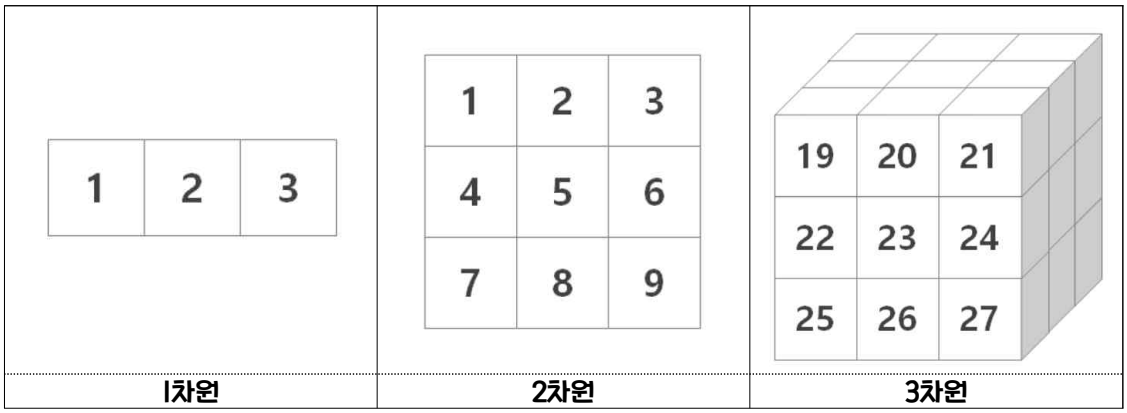


: 변수들 간의 상관관계를 고려하여 내재된 개념 요인을 추출해내는 분석방법이다. 즉 변수 간 상관관계를 고려하여 유사한 변수끼리 묶어주는 변수이다. 예를 들어 수학, 과학, 영어, 국어 점수가

있다고 하자. 수학, 과학 점수의 상관관계를 고려하여 수리 능력으로 묶고, 영어, 국어 점수의 상관관계를 고려하여 언어 능력으로 묶는 방법이다. 즉 4개의 변수에서 2개의 잠재 변수를 찾아내는 것이 요인 분석이다.

2. 차원의 저주(Curse of Dimensionality)에 대하여 조사

고차원이 된다는 것은 변수의 개수가 많아진다는 것이다. 이때 발생할 수 있는 문제가 바로 '차원의 저주'이다. 차원의 저주는 데이터보다 변수가 많아질 때 생기는 형상으로 모델의 학습 속도를 저하시키고 과적합을 발생시키기도 한다. 차원의 저주가 발생하는 이유는 아래의 그림을 보면 이해하기 쉽다.



3개의 데이터를 가진 변수가 있다고 하자. 2차원이 되어 다른 변수가 추가되면 데이터는 9개 요구된다. 3차원이 되어 또 다른 변수가 추가되면 데이터는 최소 27개 요구된다. 이처럼 차원이 올라가면 데이터 공간은 기하급수적으로 늘고, 이를 채우기 위한 데이터의 개수도 기하급수적으로 증가한다. 다시 예시로 돌아가 데이터의 수가 6개라고 가정해보자. 1차원인 경우 6/3=200%의 공간이 채워짐을 알 수 있다. 2차원인 경우 6/9=66%의 공간이 채워졌다. 3차원의 경우 6/27=22%의 공간이 채워졌다. 200%나 채우던 데이터는 고차원으로 갈수록 낮아져 3차원에는 공간의 22%만 차지하고 있음을 알 수 있다.

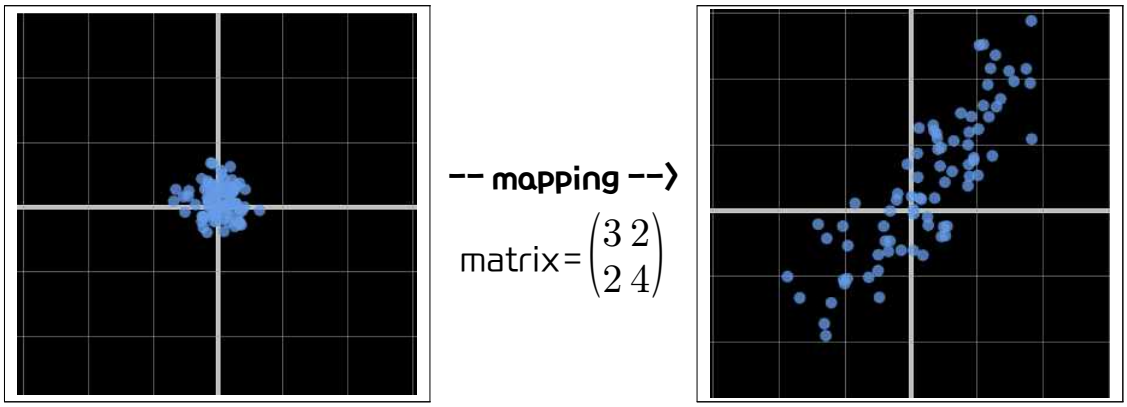
이처럼 수집된 데이터 개수가 동일한 상태에서 차원이 증가하게 되면, 차원에 비해 적은 개수의 데이터들은 특정 공간에만 채워지고 나머지 공간들은 채워지지 않은 공백의 상태가 된다. 이는 데이터 공간에서 특정 부분만 설명하는 형태임을 알 수 있다. 이로 인해 데이터가 충분히 모이지 않은 고차원 데이터셋을 모델링하게 되면 과적합이 발생하기 쉽고, 편협적인 모델이 형성될 수 있다.

정리하자면, 차원의 저주는 고차원으로 갈수록 전체 공간에서 데이터가 차지하는 영역이 매우 작아지는 현상이다. 차원의 저주는 과적합의 문제뿐만 아니라 많은 변수로 인한 잡음(noise)가 발생하여 분류모델의 정확도가 감소한다는 문제도 있으며, 변수 간 상관관계가 있는 경우 다중공산성이 발생해 모형이 불안해진다는 문제도 있다.

3. 고유벡터, 고유값에 대하여 조사

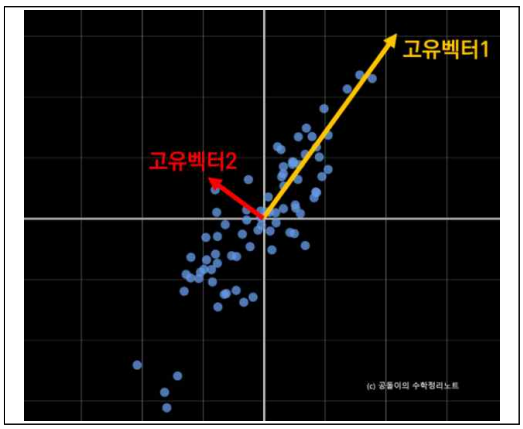
차원축소의 주요기법 중 하나인 주성분 분석은 어떤 PC축(PC벡터)에 데이터를 정사영(내적)할 것인가에 집중한다. 이때 데이터 구조를 기술하는 공분산 행렬부터 해답을 얻는데 지금부터 이에 대해 설명하고자 한다.

우선 행렬이 벡터와 결합하면 벡터를 선형변환하여 다른 벡터 공간으로 도표화(mapping)하는 특징이 있다. 즉 공분산 행렬을 통해 선형변환이 되기도 하고, 공간 속 벡터 방향이 달라지기도 한다. 이를 시각적으로 나타내면 아래의 그림과 같다.



$$\text{matrix} = \begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix}$$

활용한 행렬에서 3은 feature1의 분산으로 x축 방향으로 퍼진 정도이다. 4는 feature2의 분산으로 y축 방향으로 퍼진 정도를 나타낸다. 나머지 2는 feature1과 feature2의 공분산으로 x,y축 방향으로 함께 퍼진 정도를 나타낸다. 여기서 고유벡터의 의미를 알 수 있다.



고유벡터는 그 행렬이 벡터 변화에 작용하는 주축(principal axis)의 방향을 나타내는 벡터이다. 따라서 공분산 행렬의 고유벡터는 데이터가 x축, y축, x·y축 어느 방향으로 분산되어 있는지를 나타내준다고 할 수 있다. 실제로 고유벡터에 데이터를 정사영 해줘야 분산이 제일 커짐을 알 수 있다. 이는 데이터의 분산을 유지시켜주는 벡터이므로 데이터 정보 유지에 도움이 된다.

한편 고유값은 고유벡터 방향으로 벡터가 얼마나 늘어나있는 지를 나타내는 수치이다. 따라서 고유값이 큰 순서대로 고유벡터를 정렬하면 결과적으로 중요한 순서대로 주성분(PC)를 구성하게 된다.

4. 비즈니스 애널리틱스를 위한 데이터마이닝 (R EDITION)의 데이터셋 중 Cereals.xls 데이터를 이용하여 차원축소 진행 및 결과 해석

주어진 데이터는 77가지 아침식사용 시리얼에 관한 데이터이다. 변수는 16개이며, 각 변수의 설명을 정리하면 다음과 같다.

변수명	변수 설명
name	시리얼 이름
mfr	시리얼 제조업체
type	저온용 혹은 고온용
calories	1인분 당 칼로리
protein	단백질 (g)
fat	지방 (g)
sodium	나트륨 (mg)
fiber	식이섬유 (g)
carbo	탄수화물 (g)
sugars	설탕 (g)
potass	칼륨 (mg)
vitamins	비타민 및 미네랄 : 0, 25 또는 100으로 FDA의 일반적인 권장 비율
shelf	상품 진열 선반 (1,2 또는 3 ;바닥에서 카운트)
weight	1회분의 무게 (ounces)
cups	1회분에 제공되는 컵의 수
rating	소비자 보고서에 의해 계산된 시리얼 등급

```
> summary(cereals)
                     name    mfr    type      calories        protein           fat            sodium          fiber             carbo
 100%_Bran               : 1   A: 1   C:74   Min.   : 50.0   Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000   Min.   : 5.0
 100%_Natural_Bran       : 1   G:22   H: 3   1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0
 All-Bran                : 1   K:23          Median :110.0   Median :3.000   Median :1.000   Median :180.0   Median : 2.000   Median :14.5
 All-Bran_with_Extra_Fiber: 1   N: 6          Mean   :106.9   Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152   Mean   :14.8
 Almond_Delight          : 1   P: 9          3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0
 Apple_Cinnamon_Cheerios : 1   Q: 8          Max.   :160.0   Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000   Max.   :23.0
 (Other)                 :71   R: 8                                                                                           NA's   :1
     sugars           potass          vitamins          shelf           weight          cups            rating
 Min.   : 0.000   Min.   : 15.00   Min.   :  0.00   Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04
 1st Qu.: 3.000   1st Qu.: 42.50   1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17
 Median : 7.000   Median : 90.00   Median : 25.00   Median :2.000   Median :1.00   Median :0.750   Median :40.40
 Mean   : 7.026   Mean   : 98.67   Mean   : 28.25   Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67
 3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83
 Max.   :15.000   Max.   :330.00   Max.   :100.00   Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70
 NA's   :1        NA's   :2
```

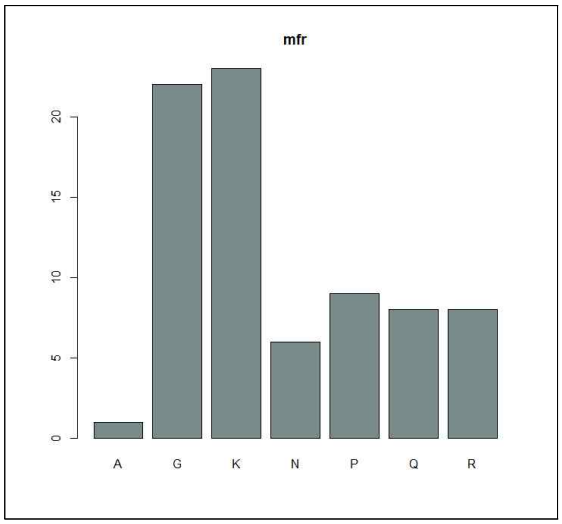
Cereals의 요약 통계량

전체적인 데이터 구성을 살피기 위해 summary 함수를 사용하였고, 3개의 문자형 변수와 13개의 수치형 변수로 구성되어 있다. name, mfr, type 의 변수는 모두 문자형 변수인데, 특히 name의 경우 전체 77개의 데이터가 모두 다른 값을 가지고 있었다. 즉 name의 경우 ID와 같은 데이

터의 고유키라고 생각하였다. 따라서 개별 분석은 하지 않았다. 나머지 변수에 대한 EDA는 다음과 같다.

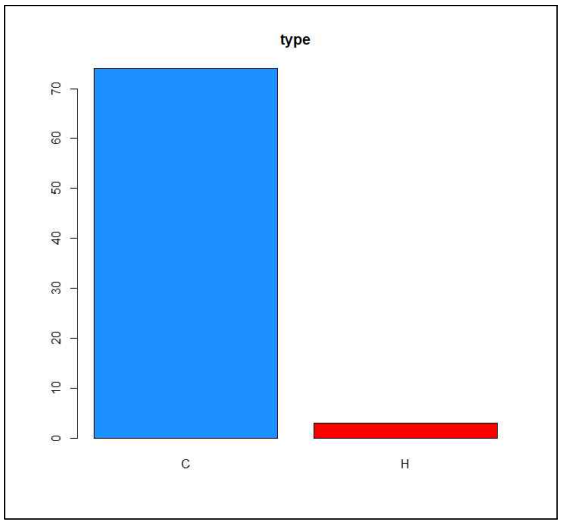
#mfr

다음은 제조업체 변수를 시각화한 그래프이다. 범주형 자료에 맞게 막대그래프를 그려 시각화하였다.



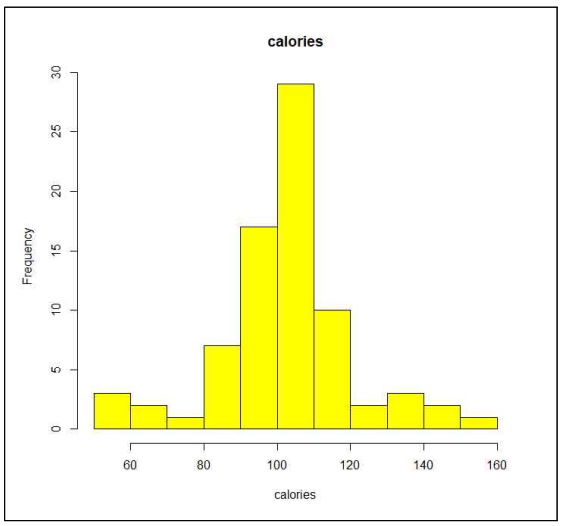
#type

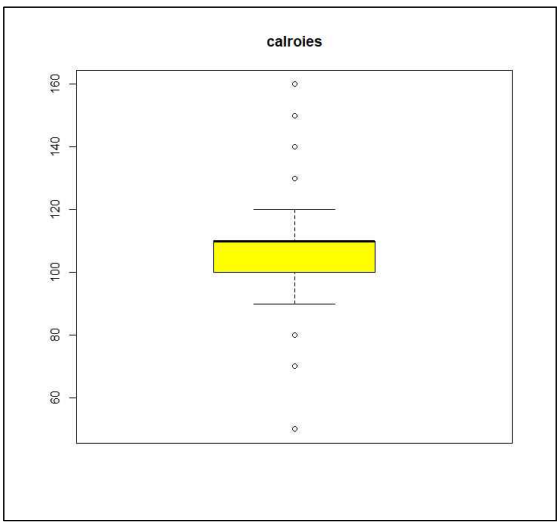
다음은 시리얼의 종류를 나타낸 변수를 시각화한 그래프이다. 저온용 시리얼의 개수가 압도적으로 많이 분포하고 있음을 알 수 있다.



#calories

다음은 시리얼 1인분 당 칼로리를 나타낸 그래프로 연속형 변수이므로 히스토그램으로 시각화하였다. Shapiro-Wilk 정규성 검정을 통해 자료를 확인하였고, 정규분포는 띠지 않음을 확인하였다.



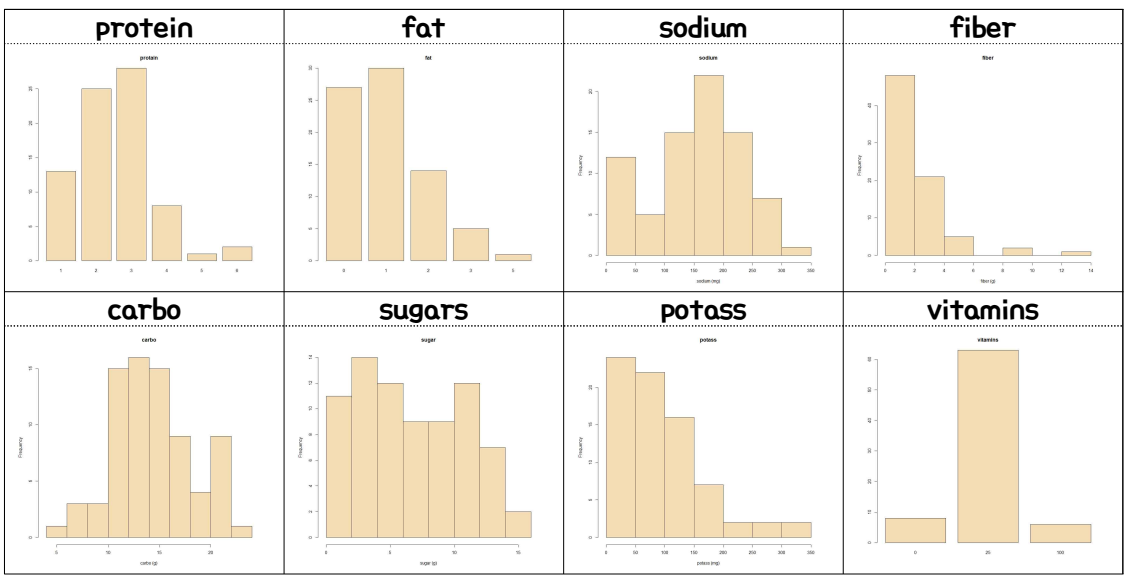



```
> shapiro.test(cereals$calories)

        Shapiro-Wilk normality test

data:  cereals$calories
W = 0.89398, p-value = 9.73e-06
```

p-value < 0.05 이므로 귀무가설(정규분포를 따른다) 기각

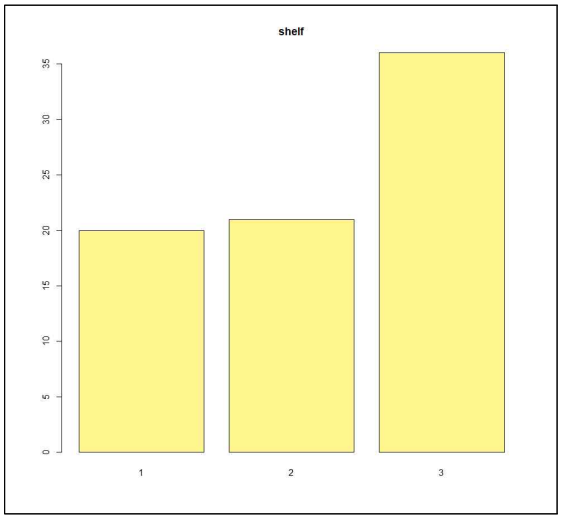
#제품 구성 성분 (protein, fat, sodium, fiber, carbo, sugars, potass, vitamins)

다음은 시리얼의 구성 성분만을 모아놓은 그래프이다. 각 변수 종류에 맞게 시각화하였다. 주목해야할 점은 각 변수 속 데이터의 분포는 고르지 않으며, 변수마다 scale이 다르게 설정되어있다.



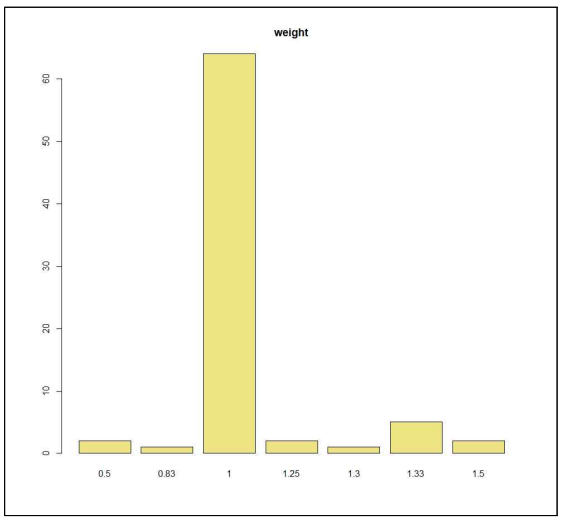
#shelf

다음은 상품 진열 선반을 바닥에서부터 카운트한 변수에 관한 자료이다. 3층 높이가 많았으며 1,2층에 분포한 데이터들의 분포는 비슷했다.



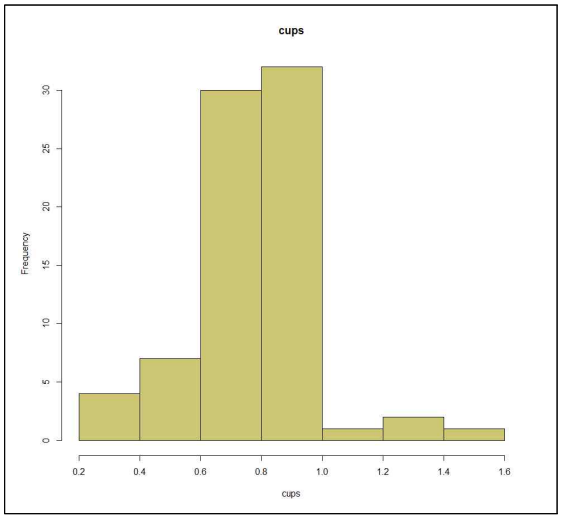
#wieght

다음은 시리얼 1회분의 무게를 나타낸 변수로, 처음에는 연속형 변수라 생각하여 히스토그램을 그려 진행했으나, 특정 무게에만 분포하고 있음을 보았다. 따라서 범주형 변수로 수정하여 막대그래프로 시각화 하였다.



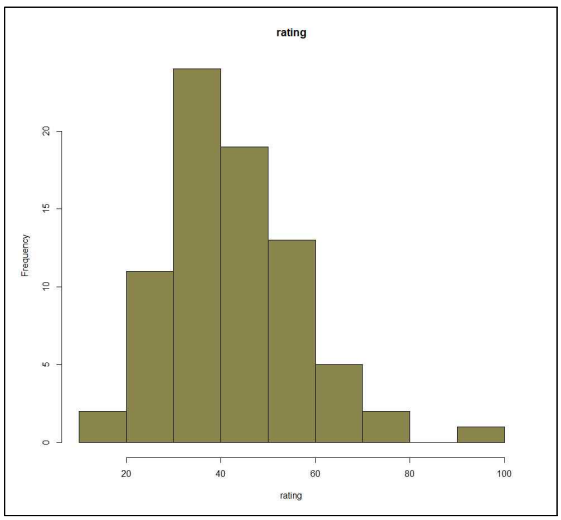
#cups

다음은 1회분에 제공되는 컵의 개수에 관한 변수를 시각화한 그래프이다. 자연수가 아닌 유리수의 형태로 분포하고 있음을 보아 연속형 변수라고 생각하였고, 이에 맞게 히스토그램을 그려 시각화하였다.



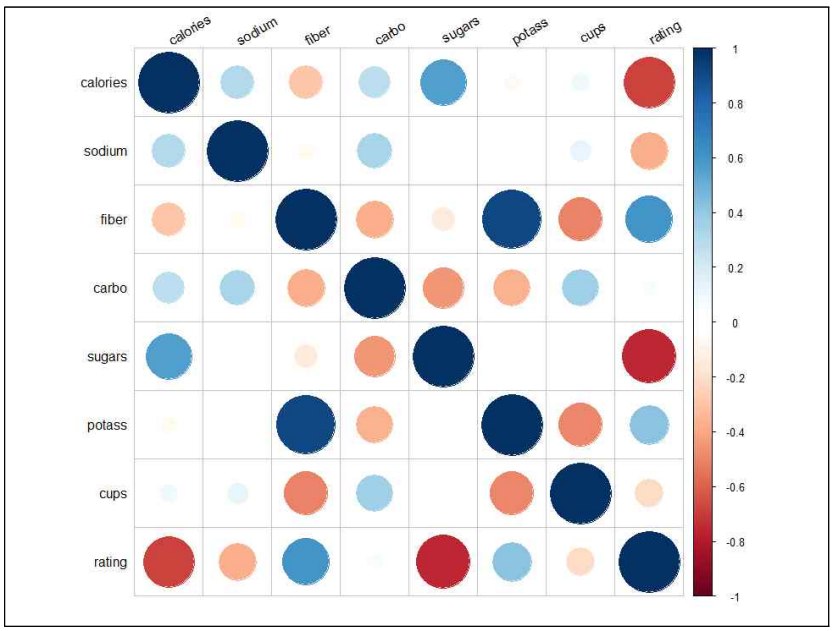
#rating

다음은 소비자 보고서에 의해 계산된 시리얼의 등급을 시각화한 그래프이다. 연속형 변수에 맞게 히스토그램으로 시각화하였다.



#연속형 변수 간 상관관계

다음은 연속형 변수간 상관관계를 나타낸 표이다. 흥미로운 점은 rating과 calories, rating과 sugers 사이의 관계이다. 칼로리가 높을수록 소비자 등급이 낮게 평가되었으며, 마찬가지로 설탕의 함유량이 높을수록 소비자 등급이 낮게 평가되었음을 알 수 있다. 건강식에 맞춰 소비자 등급이 선정되었다고 추측하였다. 또한 potass와 fiber의 관계는 강한 양의 상관관계로 단백질 함유량이 높을수록 섬유질 함유량이 높은 시리얼임을 알 수 있다.



#다중회귀분석 (PCA 안한 상태)

PCA를 진행하기 전, 성능의 비교를 위해 주성분 분석을 하지 않은 데이터셋으로 회귀분석을 진행하였다. 회귀분석 시, target 변수의 설정은 반드시 필요한 과정이므로 rating을 타겟변수로 지정하여 진행하였다. 다중회귀분석을 진행한 이유는 PCA를 했을 때, 데이터의 설명력이 어떻게 달라지나를 비교하기 위해서이다.

다음은 회귀분석을 진행한 결과를 나타낸 것으로, 데이터의 설명력은 결정계수를 통해 판단하였다. 이 경우 결정계수는 0.9543 임을 알 수 있다.

```
Call:
lm(formula = rating ~ ., data = conti2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8980 -1.5542  0.0395  0.9174 12.9703

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.409419   2.975447  19.967  < 2e-16 ***
calories    -0.215809   0.034812  -6.199 4.18e-08 ***
sodium      -0.057650   0.004687 -12.300  < 2e-16 ***
fiber        3.619270   0.434823   8.324 6.98e-12 ***
carbo        1.098653   0.178150   6.167 4.75e-08 ***
sugars      -1.085531   0.166956  -6.502 1.23e-08 ***
potass      -0.014329   0.014197  -1.009    0.317
cups         0.597028   1.803886   0.331    0.742
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3 on 66 degrees of freedom
Multiple R-squared:  0.9587,    Adjusted R-squared:  0.9543
F-statistic: 218.8 on 7 and 66 DF,  p-value: < 2.2e-16
```

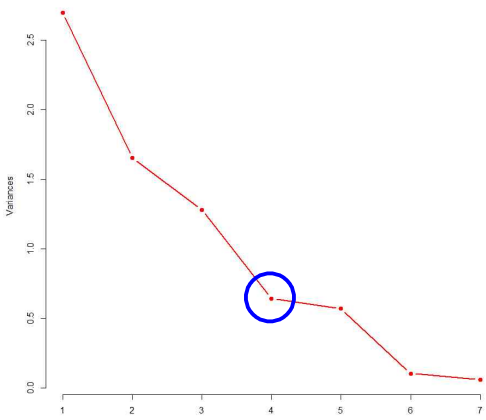
```
h(-lm(rating~.,data=conti2)
summary(h)
```

#PCA

타겟변수 rating을 제외한 나머지 연속형 변수들에 대해 PCA를 진행하였다. 이때 각 변수의 scale을 조정해주어야 하므로 표준화를 시켜주었다. 이후 각 PC축 분산을 scree plot을 통해 시각화하였다. 그래프에서는 PC4부근에서 그래프가 급격히 꺾임을 확인할 수 있다. 이와 맞게 PC4일 때 변수에 대해 약 89% 설명이 되므로 주성분 분석은 PC4에서 끊어 진행하였다.

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.6422 1.2855 1.1312 0.80126 0.75373 0.32036 0.24144
Proportion of Variance 0.3853 0.2361 0.1828 0.09172 0.08116 0.01466 0.00833
Cumulative Proportion  0.3853 0.6213 0.8041 0.89583 0.97701 0.99167 1.00000
```

```
pca<-prcomp(conti3,scale=T)
summary(pca)
```




```
plot(pca,type='l',pch=19,lwd=2,col='red',main='Scree Plot')
```

다음은 PCA를 진행한 데이터로, PC1부터 PC4까지의 데이터를 다중 회귀 분석한 결과이다. 이 결정계수는 0.9149로 PCA를 진행하기 전보다 설명력이 떨어졌다.

```
Call:
lm(formula = rating ~ ., data = cereals3)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7506 -3.1527 -0.2412  1.8882 13.0737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.84375    3.07991  22.028  < 2e-16 ***
PC1          0.02354    0.02235   1.053    0.296
PC2         -1.32349    0.06864 -19.281  < 2e-16 ***
PC3         -0.73388    0.05787 -12.682  < 2e-16 ***
PC4          0.63073    0.06279  10.046 3.86e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.095 on 69 degrees of freedom
Multiple R-squared:  0.9195,    Adjusted R-squared:  0.9149
F-statistic: 197.1 on 4 and 69 DF,  p-value: < 2.2e-16
```

```
s<-lm(rating~.,data=cereals3)
summary(s)
```

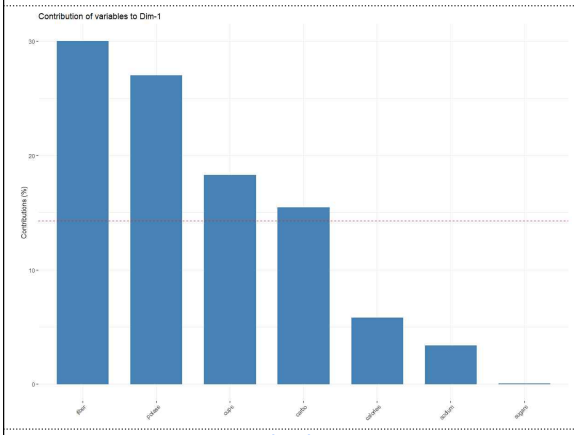
한편 PC축 전부 (PC1+PC2+PC3+PC4+PC5+PC6+PC7)를 모두 넣었을 때 결정계수가 어느 정도 나오는 지 확인하기 위해 모든 PC축에 대한 다중 회귀 분석을 진행해보았다. 흥미롭게도 PCA를 진행하지 않은 결정계수와 동일한 결정계수 값이 도출되었다. 따라서 7개의 PC축들이 각 변수들을 잘 설명하는 축이라고 생각하였다.

```
Call:
lm(formula = rating ~ ., data = cereals3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8980 -1.5542  0.0395  0.9174 12.9703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.4094     2.9754  19.967  < 2e-16 ***
PC1          -1.3729     0.7563  -1.815  0.07401 .
PC2          -1.5030     0.2292  -6.558 9.80e-09 ***
PC3          -1.6153     0.2282  -7.078 1.18e-09 ***
PC4          -0.1473     0.9974  -0.148  0.88302
PC5          -1.5287     1.2862  -1.189  0.23888
PC6          -0.8831     0.2678  -3.297  0.00157 **
PC7           2.4494     0.3268   7.496 2.11e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3 on 66 degrees of freedom
Multiple R-squared:  0.9587,    Adjusted R-squared:  0.9543
F-statistic: 218.8 on 7 and 66 DF,  p-value: < 2.2e-16
```

```
fit_pca<-lm(rating~.,data=cereals3)
summary(fit_pca)
```

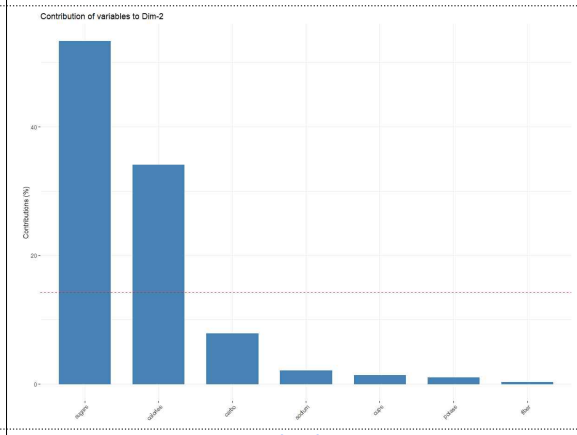
다음은 PC축의 정보를 알아보는 그래프이다. 그래프의 x축 변수의 순서가 모두 다름을 주의해야 한다. PC축 4개만 본 이유는 앞서 가장 영향력 있는 축을 4가지만 선정했기 때문이다. 우선 PC1의 경우 순서대로 fiber, potass, cups, carbo의 정보를 주로 담고 있다. PC2의 경우 sugars, calories의 정보를 주로 담고 있다. PC3의 경우 sodium, carbo에 관한 정보를 주고 담고 있다. 마지막으로 PC4는 sodium, cups, carbo, calories에 관한 정보를 담고 있다.

PC1



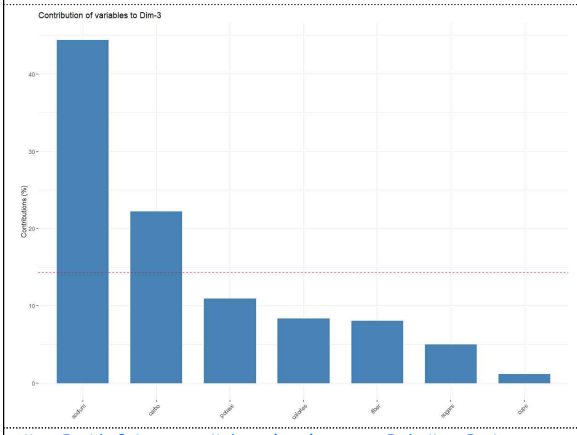

```
dim_1st<-fviz_contrib(pca,'var',axes =1 );dim_1st
```

PC2



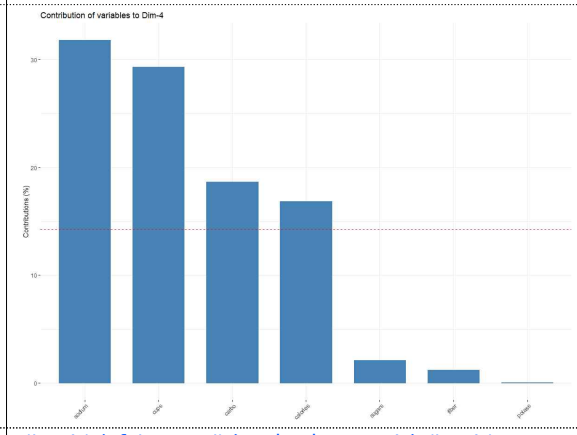

```
dim_2nd<-fviz_contrib(pca,'var',axes =2 );dim_2nd
```

PC3




```
dim_3rd<-fviz_contrib(pca,'var',axes =3 );dim_3rd
```

PC4




```
dim_4th<-fviz_contrib(pca,'var',axes =4 );dim_4th
```

5. DISCUSSION

PCA에 범주형 변수 및 순서형 변수는 포함시켜 진행해야하는가?