1. R 패키지 ‘arules’ 안에 있는 ‘Income’자료에 대한 연관규칙 분석을 수행하여라. (apriori 알고리즘 이용)

(a) ‘Income’ data 불러오기

-> 첨부한 R코드를 실행하면 불러오기 가능.

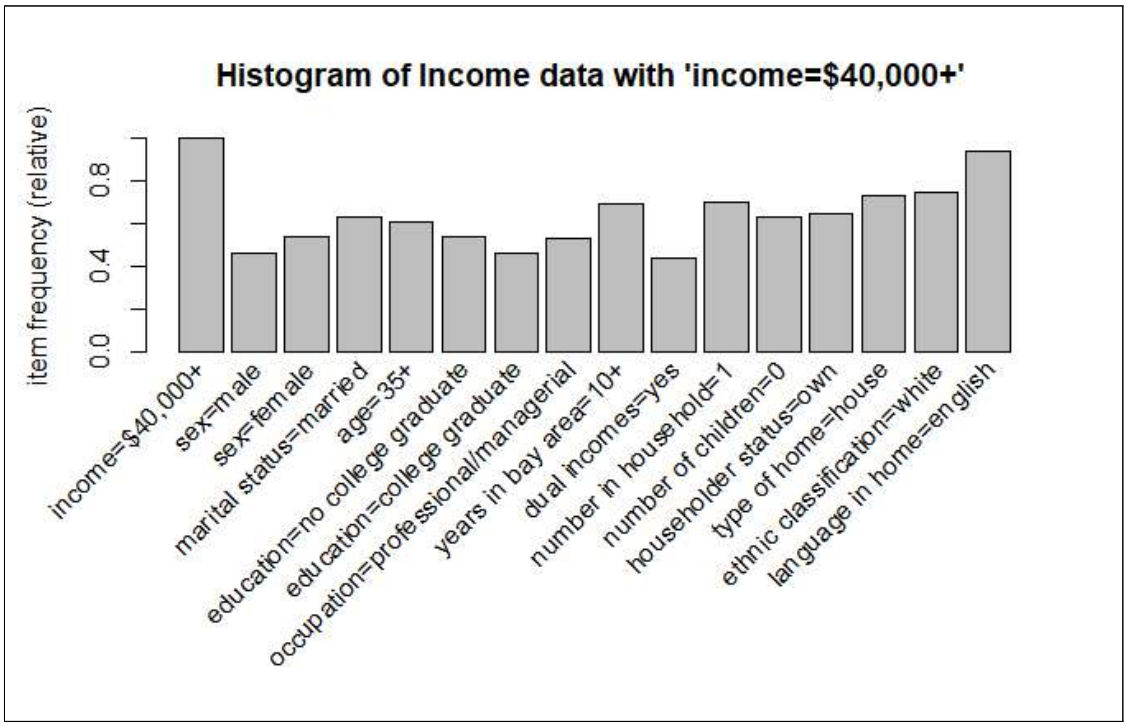
(b) ‘Income’ data 변수 확인

-> 다음과 같은 총 14개의 변수를 포함하고 있다.

income
an ordered factor with levels --> [0,10) < [10,15) < [15,20) < [20,25) < [25,30) < [30,40) < [40,50) < [50,75) < 75+
sex
a factor with levels --> male, female
marital status
a factor with levels --> married, cohabitation, divorced, widowed, single
age
an ordered factor with levels --> 14-17 < 18-24 < 25-34 < 35-44 < 45-54 < 55-64 < 65+
education
an ordered factor with levels --> grade <9 < grades 9-11 < high school graduate < college (1-3 years) < college graduate < graduate study
occupation
a factor with levels --> professional/managerial, sales, laborer, clerical/service, homemaker, student, military, retired, unemployed
years in bay area
an ordered factor with levels --> <1 < 1-3 < 4-6 < 7-10 < >10
dual incomes
a factor with levels --> not married, yes, no
number in household
an ordered factor with levels --> 1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9+
number of children
an ordered factor with levels --> 0 < 1 < 2 < 3 < 4 < 5 < 6 < 7 < 8 < 9+
householder status
a factor with levels --> own, rent, live with parents/family
type of home
a factor with levels --> house, condominium, apartment, mobile Home, other
ethnic classification
a factor with levels --> american, indian, asian, black, east indian, hispanic, pacific islander, white, other
language in home
a factor with levels --> english, spanish, other

(c) 고소득자(income="$40,000+") 그룹에 대한 itemFrequencyPlot를 그리고 설명하여라.

-> R을 이용하여 다음과 같은 plot을 그릴 수 있다. (단, 지지도가 0.4 이상인 경우에 한하여 분석을 진행하였음.)



위에 있는 plot을 통해 고소득자(income="$40,000+") 그룹은 모국어가 영어(language in home="english")이면서 자가를 소유(type of home="house")한 백인계층(ethnic classification="white")인 그룹과 가장 연관성이 높은 것을 확인할 수 있다.

(d) 연관규칙분석 : (rhs) 고소득자에 대한 연관규칙을 신뢰도 기준 상위 5개 추출하고 설명하여라. (단 최소지지도 0.1, 최소신뢰도 0.8, 향상도 1.0 적용)

-> 향상도를 기준으로 상위 5개의 연관규칙을 추출하면 다음과 같다.

lhs	rhs	support	lift
{marital status=married, occupation=professional/managerial, householder status=own}	{income=$40,000+}	0.1042757	2.318817
{marital status=married, occupation=professional/managerial, type of home=house}	{income=$40,000+}	0.1038394	2.227520
{marital status=married, dual incomes=yes, householder status=own, type of home=house, language in home=english}	{income=$40,000+}	0.1051483	2.208769
{marital status=married, dual incomes=yes, householder status=own, language in home=english}	{income=$40,000+}	0.1175102	2.199529
{dual incomes=yes, householder status=own, type of home=house, language in home=english}	{income=$40,000+}	0.1074753	2.199306

결과적으로 기혼자인 경우, 직업이 교수 또는 직급이 관리자급인 경우, 자가를 소유한 경우, 맞벌이를 하는 경우, 모국어가 영어인 경우에 고소득자일 가능성이 높다고 해석할 수 있겠다.

2. 'sample_DT.csv'는 제품의 불량 여부(DEFECT TYPE)를 분류하기 위한 데이터이다. 각 변수들은 제품을 생성하는 공정에서 관측된 값들이다. 이를 분류하기 위한 모형 적합을 하여라.

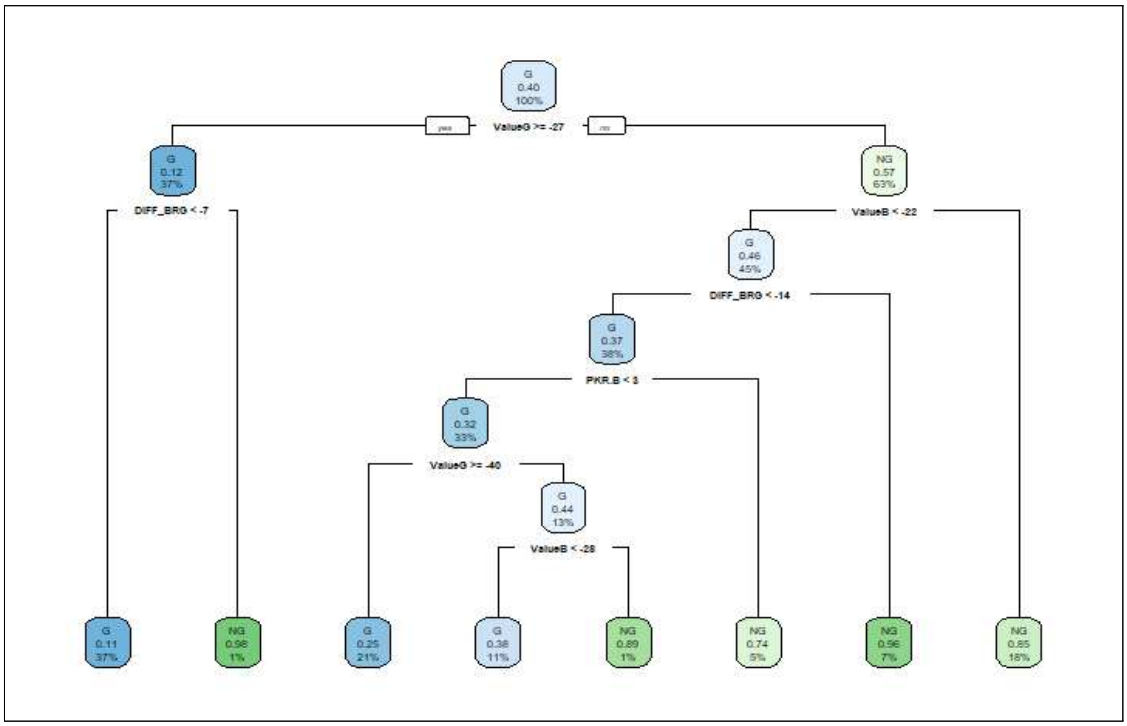
(a) 모형 검증을 위해 training data와 testing data로 나누시오.

(단, training data : testing data = 7:3, seed=1234)

-> 첨부한 R코드를 실행하여 데이터를 train과 test로 나눌 수 있음.

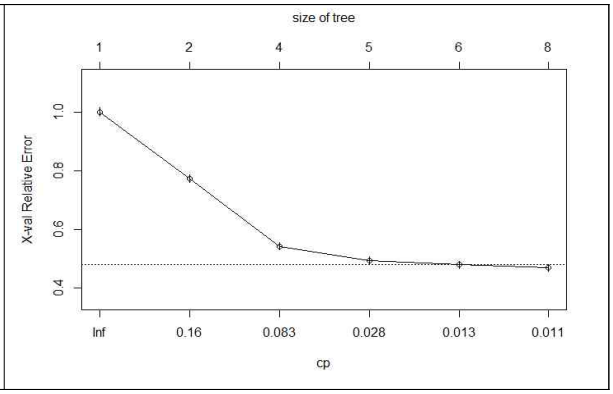
(b) 의사결정나무(가지치기 실행)

-> 제품의 불량 여부를 반응변수로 두고 의사결정나무를 적합하면 다음과 같은 형태를 보인다.

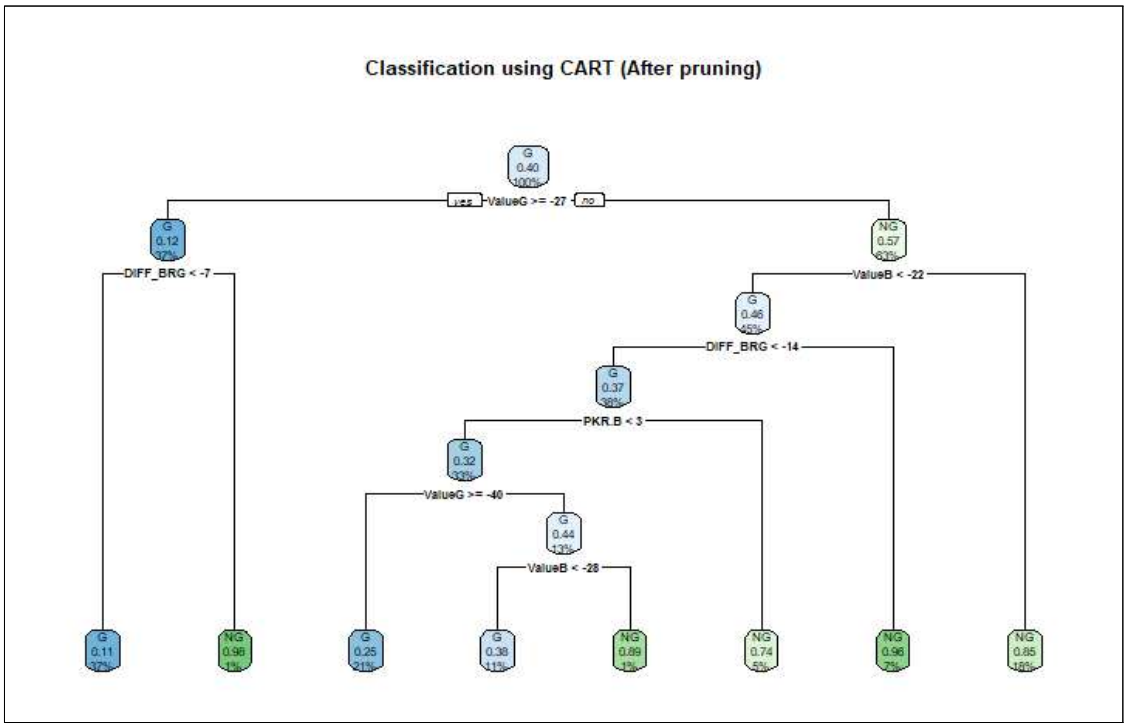


위의 의사결정나무에 대해 가지치기를 시행하기 위하여 cptable을 작성하고 이에 대한 그래프를 그리면 다음과 같다.

	CP	nsplit	rel error	xerror	xstd
1	0.22218297	0	1.0000000	1.0000000	0.01450431
2	0.11992229	1	0.7778170	0.7732250	0.01370098
3	0.05722360	3	0.5379724	0.5432709	0.01223672
4	0.01342282	4	0.4807489	0.4948781	0.01182457
5	0.01289297	5	0.4673260	0.4796892	0.01168631
6	0.01000000	7	0.4415401	0.4697987	0.01159387



위의 결과를 바탕으로 가지치기를 위한 적절한 cp는 0.012와 0.013 사이의 값인 0.0125로 둘 수 있겠다. 이를 바탕으로 가지치기를 하고 형태를 살펴보면 다음과 같다.



결과적으로 가지치기 이전의 의사결정나무와 동일한 결과임을 확인할 수 있다. 아마도 초기에 지정된 cp의 default 값이 0.01이고 이 값이 가지치기를 위해 선택한 cp의 값인 0.0125와 거의 비슷하다는 것이 주된 원인이라고 판단된다. 이 모형에 대해 test data에 대한 confusion matrix와 오분류율을 살펴보면 다음과 같다.

confusion matrix	오분류율
yhat G NG G 1653 413 NG 121 813	0.178

(c) 배깅

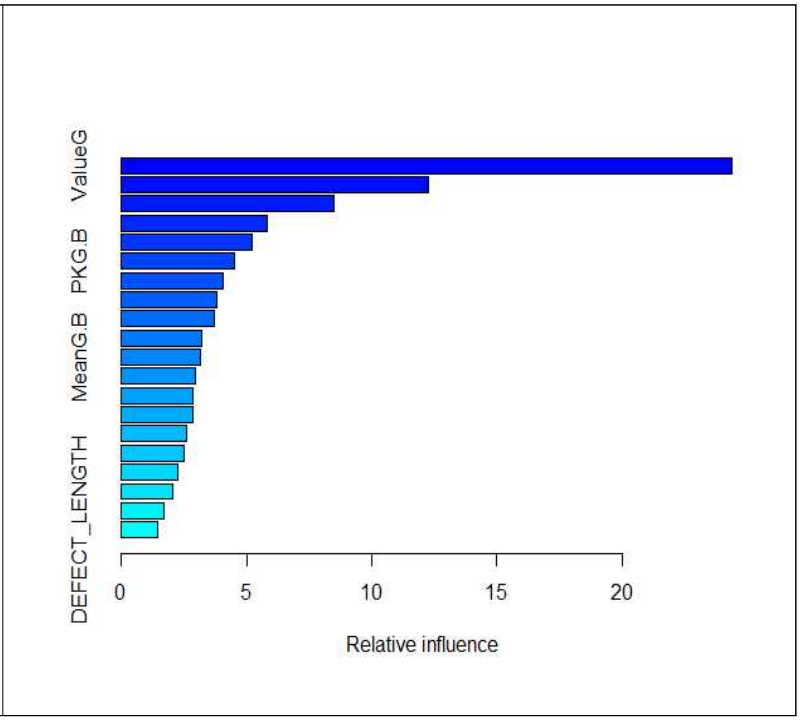
-> 분석 전 반응변수 DEFECT_TYPE을 factor로 변환하여 분석이 가능하게 하였다. 그런 다음 train data를 이용하여 총 1000개의 bootstrap sample을 추출하고 out-of-bag 방법을 통하여 Bagging을 실시한 뒤 이에 대하여 test data에 대한 confusion matrix와 오분류율을 살펴보면 다음과 같다.

confusion matrix	오분류율
yhat G NG G 1650 277 NG 124 949	0.134

(d) 부스팅

-> 분석 전 반응변수 DEFECT_TYPE을 0 또는 1의 수치형으로 변환하여 gradient boosting이 가능하게 하였다. 각 설명변수에 대한 relative influence에 대한 그래프는 다음과 같다.

var	rel.inf
ValueG	24.386961
DIFF_BRG	12.245773
ValueB	8.474496
PKR.B	5.826804
MaxB	5.244220
PKG.B	4.514696
ValueR	4.081199
DEFECT_RBG	3.819038
CTB	3.703988
MinG	3.216323
MeanG.B	3.177697
DEFECT_AREAR3	2.979433
PKB.R	2.858292
BlackRatio	2.836922
DEFECT_SIZE_Y	2.602106
CBlobCnt	2.492466
G	2.278517
DEFECT_SIZE_X	2.069606
PKR	1.720729
DEFECT_LENGTH	1.470734



위의 결과를 통해 설명변수 ValueG에 대한 relative influence가 가장 큼을 알 수 있다.
그런 다음 train data를 이용하여 총 1000개의 bootstrap sample을 추출하고 gradient boosting을 실시한 뒤 이에 대하여 test data에 대한 confusion matrix와 오분류율을 살펴보면 다음과 같다. 단, 예측은 반응변수가 "G"일 확률을 먼저 계산한 뒤 이 값이 0.5보다 크면 반응변수에 대한 최종예측을 "G", 그렇지 않으면 "NG"로 한다. 그리고 각각의 tree에 대한 maximum depth은 2로 설정하였다.

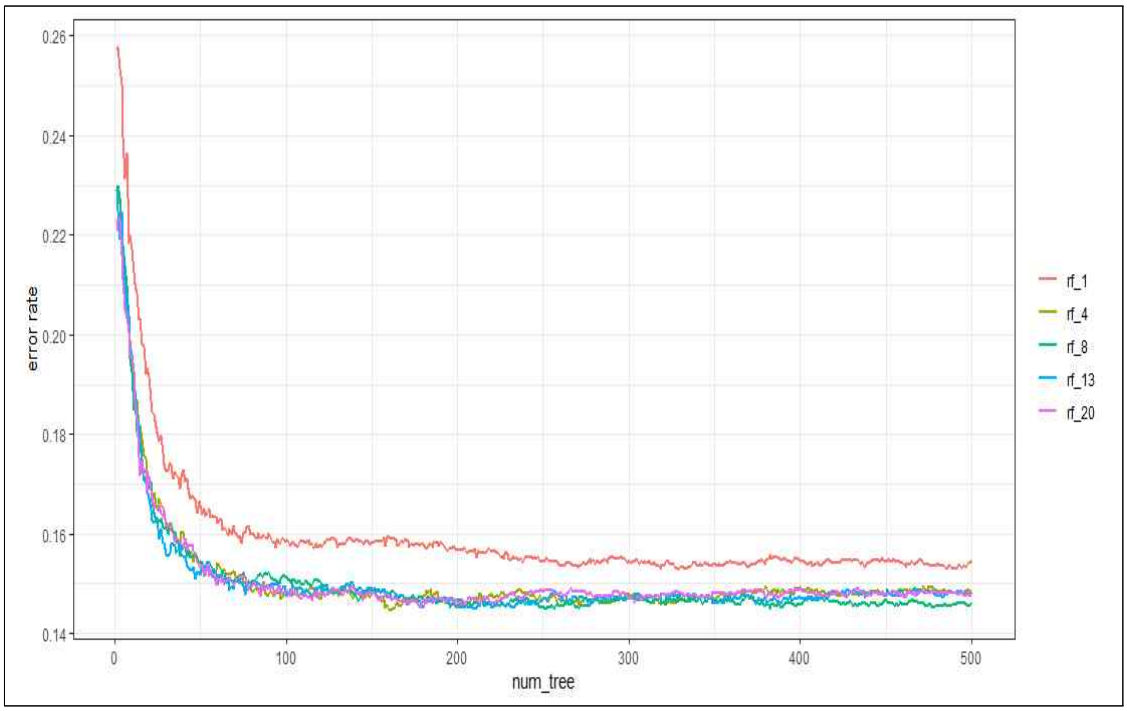
confusion matrix	오분류율
yhat G NG G 1624 277 NG 150 949	0.142

(e) 랜덤포레스트

-> 분석 전 반응변수 DEFECT_TYPE을 factor로 변환하여 분석이 가능하게 하였다. 그런 다음 train data를 이용하여 총 1000개의 bootstrap sample을 추출하고 각 sample마다 임의로 4개의 설명변수를 선택한 뒤 out-of-bag 방법을 통하여 Random Forest를 실시하고 이에 대하여 test data에 대한 confusion matrix와 오분류율을 살펴보면 다음과 같다.

confusion matrix	오분류율
yhat G NG G 1624 277 NG 150 949	0.142

각 bootstrap sample마다 임의로 선택하는 설명변수의 개수와 생성하는 tree의 개수에 따라 error rate을 계산하고 이를 그래프로 표현하면 다음과 같다.



위의 그래프를 통해 각 bootstrap sample마다 임의로 선택하는 설명변수의 개수에 따라 error rate이 미묘하게 달라짐을 확인할 수 있고, 이를 통해 적절한 설명변수의 개수를 선정할 수 있다.

(f) 위에서 적합한 모형 중 가장 좋은 모형을 선택하시오.

-> 위의 결과를 통해서 판단했을 때 test 오분류율이 가장 작은 bagging이 최적의 모형이라고 판단할 수 있다. 하지만, 다른 데이터를 분석하게 된다면 최적의 모형은 달라지게 되며, 보통 random forest의 성능이 다른 모형과 비교했을 때 좋다는 사실이 알려져 있다.

3. 'Wholesale customers data.csv'는 다양한 상품군에 대한 고객의 구매이력에 대한 자료이다. 다음에 대하여 다양한 방법으로 군집분석을 시행하여라.

- 변수설명

1) FRESH: annual spending on fresh products (Continuous)

2) MILK: annual spending on milk products (Continuous)

3) GROCERY: annual spending on grocery products (Continuous)

4) FROZEN: annual spending on frozen products (Continuous)

5) DETERGENTS PAPER: annual spending on detergents and paper products (Continuous)

6) DELICATESSEN: annual spending on and delicatessen products (Continuous)

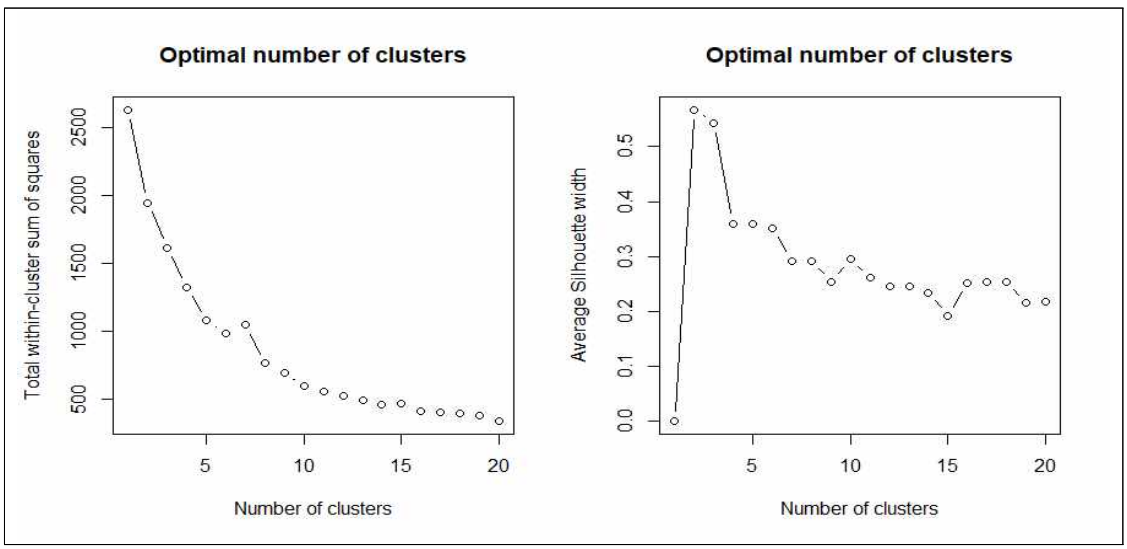
• 정답은 없습니다. 다양한 군집 분석 알고리즘을 적용해보고, 데이터 분석을 해보는 것이 목표입니다.

데이터에는 이상치가 존재합니다. 가능하면 이상치를 제거하는 단계를 포함시키는 것이 필요합니다. 이상치가 있는 경우와 없는 경우의 분석 결과 차이를 비교해 보는 것도 좋을 것 같습니다.

실제 변수는 8개입니다. 위에 제시한 6개의 변수만을 이용하여 분석합니다.

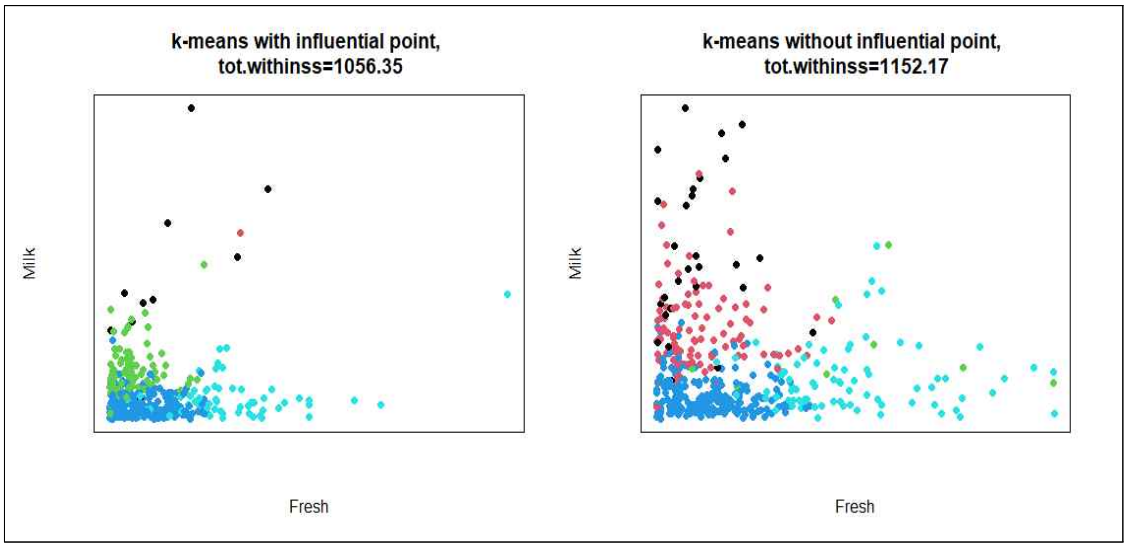
1) k-means clustering

먼저 k-means clustering 방법을 통하여 군집분석을 실시한다. 우선 변수 별 척도의 차이에 따른 분석결과의 왜곡을 보완하기 위해 데이터에 대한 표준화를 실시하고 적절한 k의 값을 선정하기 위하여 Elbow Method와 Silhouette Method를 실시하여 최적의 점을 찾아보도록 한다. Elbow Method에서는 k에 따른 total within sum of squares의 값이 안정적으로 작아지는 부분을, Silhouette Method에서는 k에 따른 average silhouette width의 평균값이 가장 큰 부분을 최적의 점으로 선택하게 된다. 두 방법에 대한 그래프는 다음과 같다.



위의 결과를 토대로 최적의 k를 5로 설정하였으며 최소 5개의 개체가 선택되었을 때 하나의 군집을 형성하도록 하였다. k-means clustering의 경우는 이상치의 영향을 많이 받기 때문에 이상치를 제거한 후 분석하게 되면 결과가 달라질 수 있다. 이에 따라 변수 Fresh의 값이 60000보다 크거나 변수 Milk의 값이 30000보다 큰 경우를 이상치로 간주하고 제거한 뒤 분

석한 결과도 추가로 비교하였다.

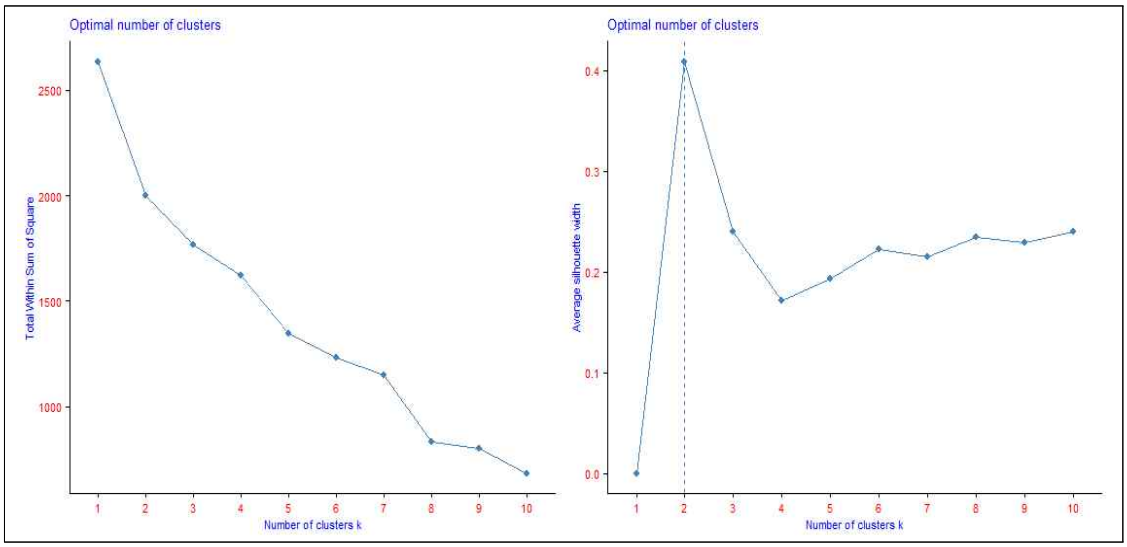


위에 있는 수치는 total within sum of squares를 나타낸 것이고 변수 Fresh와 Milk에 대해 그래프를 그려본 것이다. 그리고 색은 분류된 군집을 나타낸다. 오른쪽에 있는 그림이 이상치를 제거하고 분석한 결과인데 이상치를 포함하여 분석한 결과와 비교하면 total within sum of squares의 값이 그렇게 큰 차이가 있는 것은 아니지만 군집분석의 결과가 어느정도 안정되게 나오는 것으로 보인다. 이처럼 k-means clustering 방법은 평균을 통하여 중심점을 잡기 때문에 이상치의 영향을 많이 받게 되고 이로 인하여 왜곡된 분석결과가 나올 가능성도 있다.

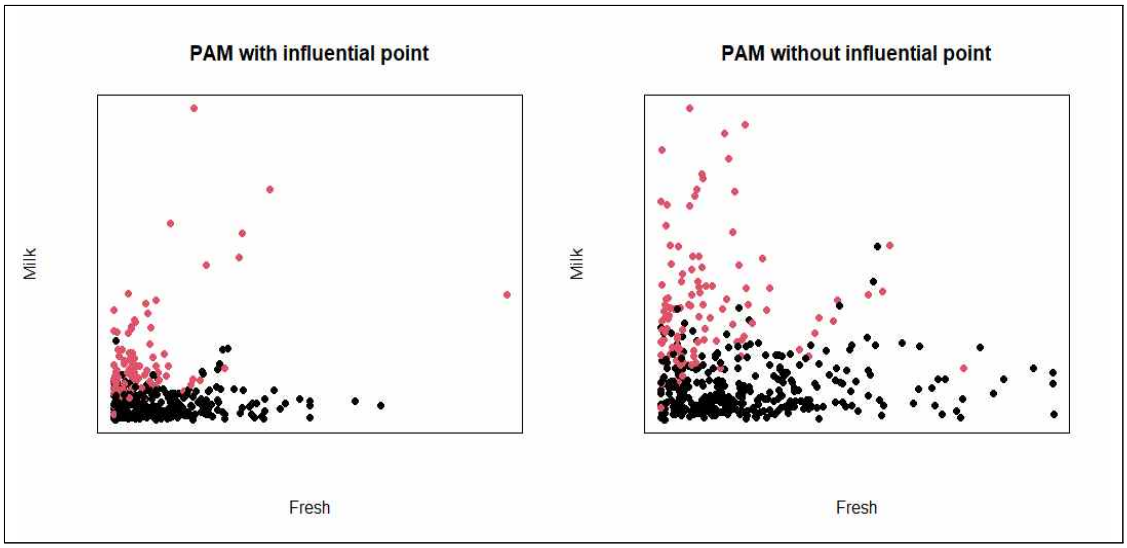
그렇다면 평균 대신 중앙값을 사용하는 k-medoids clustering의 경우는 어떠한 결과를 주게 되는지 살펴보도록 하자. 여기에 loss를 통하여 중심점의 위치를 update하는 PAM 방법까지 같이 적용해보도록 한다.

2) k-medoids clustering with PAM(partitioning around medoid)

k-means clustering과 마찬가지로 적절한 k의 값을 선정하기 위하여 Elbow Method와 Silhouette Method를 사용하여 그래프를 그리면 다음과 같다.



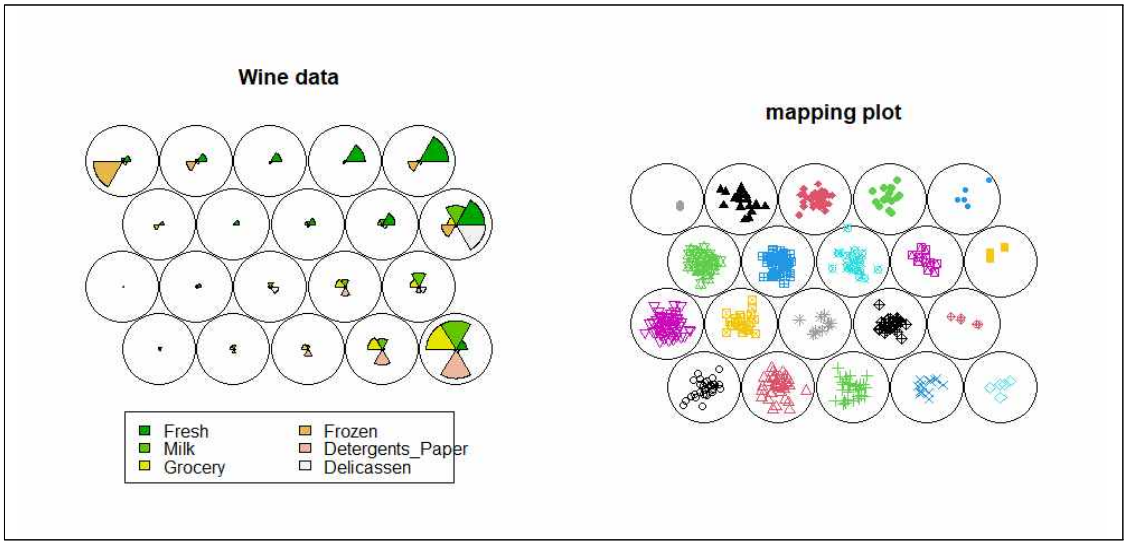
위의 결과를 바탕으로 최적의 k값을 2로 선정하도록 한다. 이를 토대로 이상치를 포함하는 경우와 그렇지 않은 경우를 나누어서 군집분석을 실시하면 다음과 같은 결과를 얻게 된다.

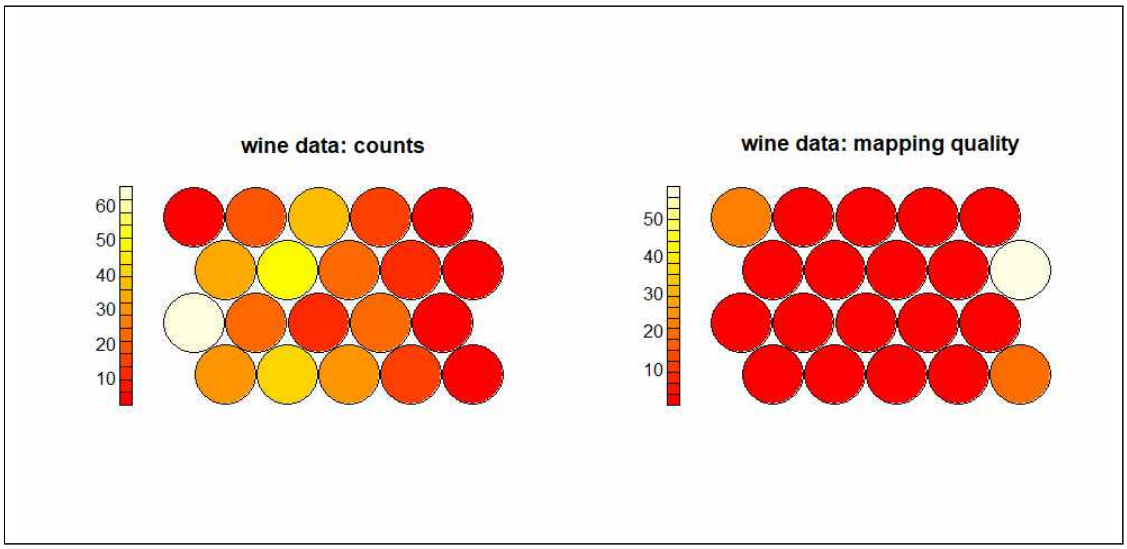


위의 결과를 보면 이상치의 유무가 군집분석의 결과에 많은 영향을 주지는 않은 것을 확인할 수 있다. 결과적으로 k-medoids의 경우는 중앙값을 바탕으로 중심점을 계산하기 때문에 분석을 하기 위한 계산량은 k-means보다 더 많지만 이상치의 영향은 덜 받게 된다. 이로 인하여 더 안정적인 군집분석 결과를 얻을 수 있다. 다음으로는 인공신경망의 방법을 이용하는 SOM(self organizing map)을 이용하여 군집분석을 실시해보도록 하겠다.

3) SOM(self organizing map)

이 방법은 비지도학습 인공신경망의 한 종류이고 경쟁 학습을 통하여 각각의 뉴런이 입력 벡터와 얼마나 가까운지 계산하여 가중 벡터를 반복적으로 재조정하여 학습하는 알고리즘을 통해 분석을 실시하게 된다. grid는 5*4 형태로 설정하여 총 20개의 군집으로 나누어지도록 하였으며 분석결과는 다음과 같다.

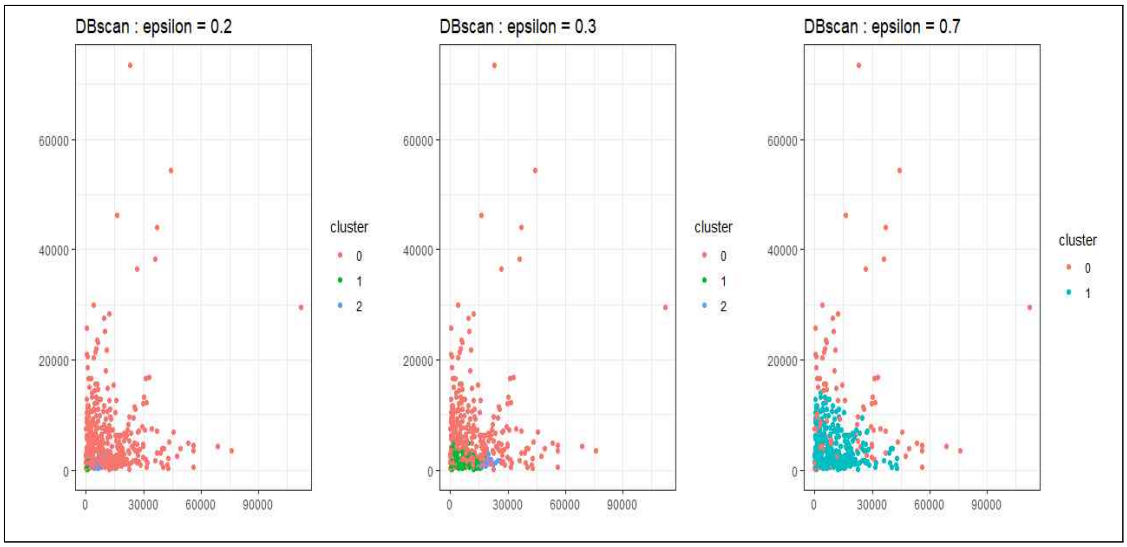




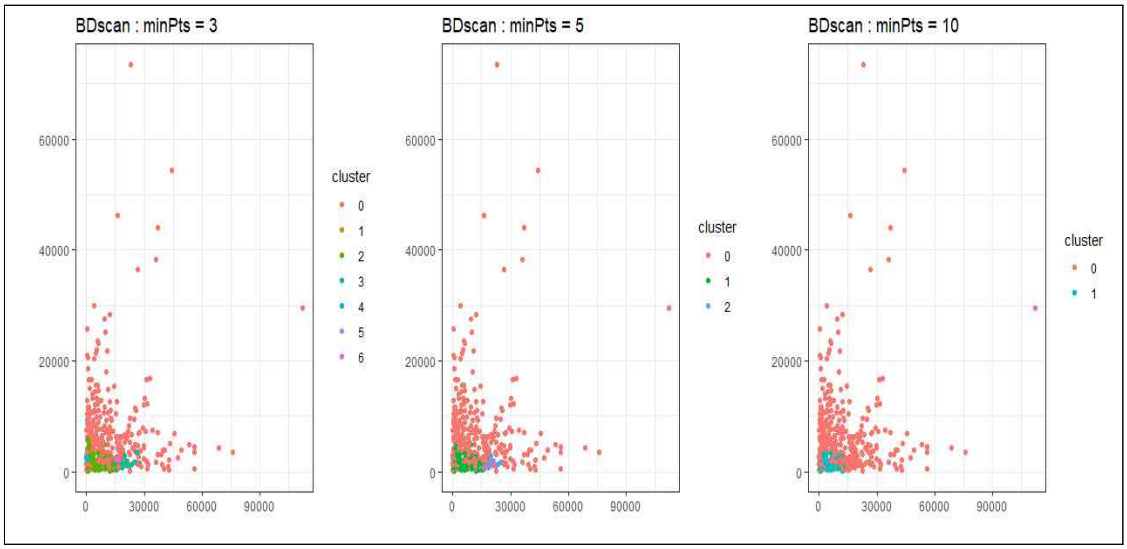
위의 결과를 통해 각 군집에 대해 변수의 분포가 어떻게 형성되어 있는지 알 수 있고 군집 당 분류된 개체의 개수도 파악할 수 있다. 마지막으로 DBSCAN의 방법을 통하여 군집분석을 실시해보도록 하겠다.

4) DBSCAN(Density-based spatial clustering of applications with noise)

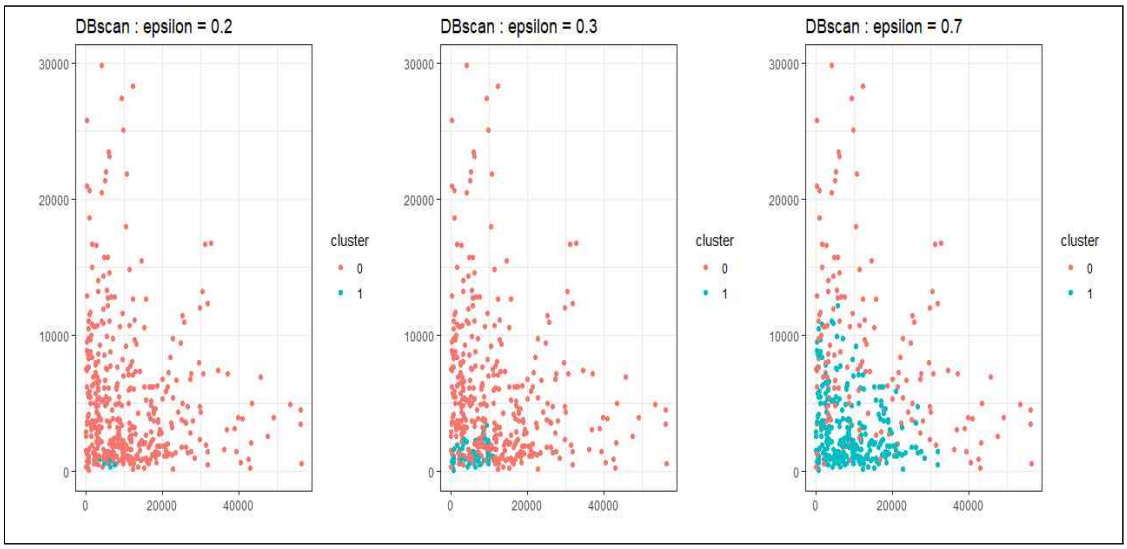
이 방법은 개체의 밀도가 높은 부분을 군집화하는 방식으로 한 점을 기준으로 ε 내에 점이 m 개 이상 있으면 하나의 군집으로 인식하게 된다. m=5로 고정하고 ε=0.2, 0.3, 0.7로 값을 바꿔가면서 분석한 결과를 살펴보면 다음과 같다.



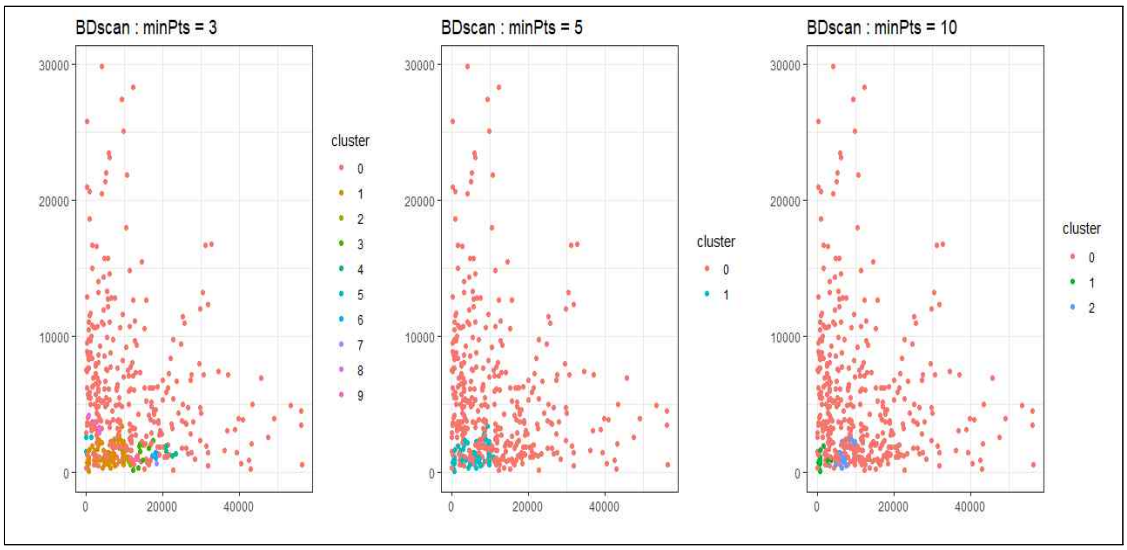
그리고 ε=0.3으로 고정하고 m=3, 5, 10으로 값을 바꿔가면서 분석한 결과를 살펴보면 다음과 같다.



즉, DBSCAN은 m과 ε의 값에 따라 군집의 개수가 달라지기 때문에 적절한 parameter를 설정하는 것이 중요하다고 할 수 있겠다. 추가적으로 이상치를 제거하고 분석을 실시하면 다음과 같은 결과를 얻게 된다. 먼저 m=5로 고정하고 ε=0.2, 0.3, 0.7로 값을 바꿔가면서 분석한 결과는 다음과 같다.



그리고 ε=0.3으로 고정하고 m=3, 5, 10으로 값을 바꿔가면서 분석한 결과는 다음과 같다.



결과적으로 이상치를 제거한 뒤 분석을 실시하면 그 결과가 다소 차이가 있음을 위의 결과를 통해서 확인할 수 있겠다.