Part-IV. 예측 분석

13. 회귀 분석
14. 분류 분석
15. 군집 분석
16. 연관 분석

예측분석 방법 분류

1. 지도학습(Supervised Learning)

- 인간 개입에 의한 분석 방법
- 종속변수(y) 존재
- 분석방법 : 가설검정(확률/통계) → 인문.사회.심리 계열(300년)
- 분석유형 : 회귀분석, 분류분석, 시계열 분석 → 추론통계 기반

2. 비지도학습(unSupervised Learning)

- 컴퓨터 기계학습에 의한 분석 방법
- 종속변수(y) 없음
- 분석방법 : 규칙(패턴분석) → 공학.자연과학 계열(100년)
- 분석유형 : 연관분석, 군집분석 → 데이터마이닝 기반

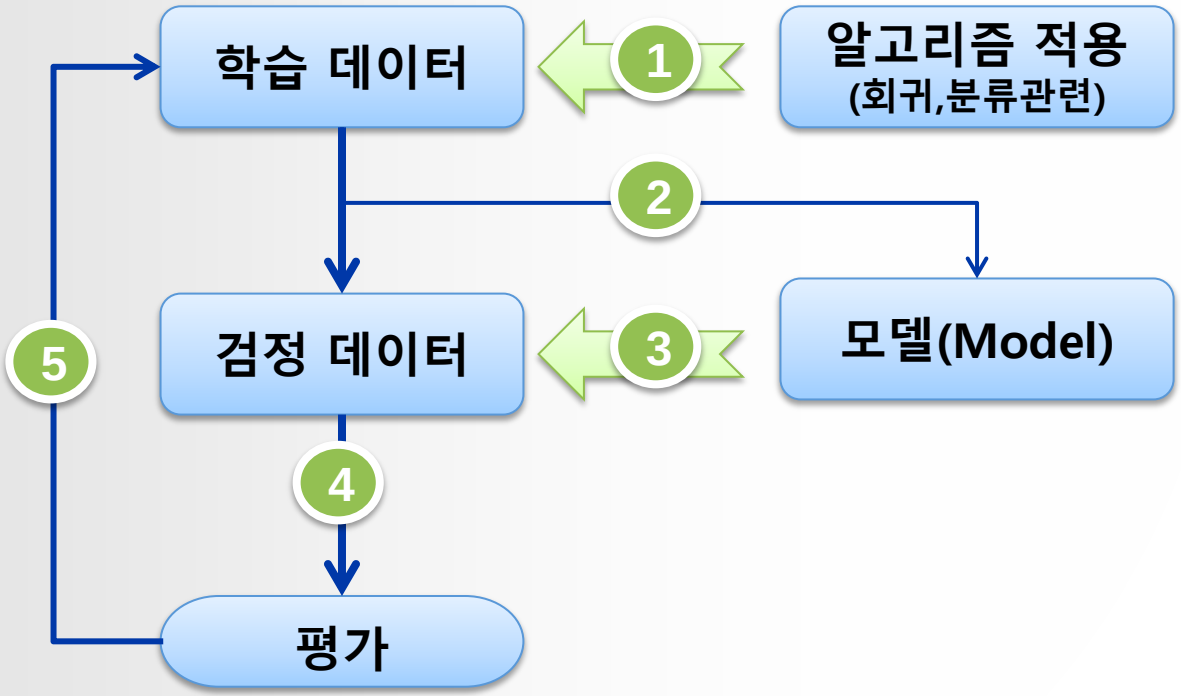
예측분석 방법 분류

지도학습과 비지도학습

분류	지도학습	비지도학습
주 관	사람의 개입에 의한 학습	컴퓨터에 의한 기계학습
기 법	확률과 통계 기반 추론통계	패턴분석 기반 데이터 마이닝
유 형	회귀분석, 분류분석(y변수 있음)	군집분석, 연관분석(y변수 없음)
분 야	인문, 사회 계열	공학, 자연 계열

예측분석 방법 분류

- 지도학습(Supervised Learning) 절차



예측분석 방법 분류

1. 예측분석 : 인과관계 예측(회귀분석 - p값 제공)
2. 분류분석 : 고객 이탈분석(번호이동, 반응고객 대상 정보 제공)
3. 군집분석 : 그룹화를 통한 예측(그룹 특성 차이 분석-고객집단 이해)
4. 연관분석 : 상품구매 규칙을 통한 구매 패턴 예측(상품 연관성)

13. 회귀 분석

Chap13_RegressionAnalysis 수업내용

1) 상관분석
2) 회귀분석
   ① 단순회귀분석
   ② 다중회귀분석

1) 상관분석

상관관계 분석(Correlation Analysis)

- 변수 간 관련성 분석 방법
- 하나의 변수가 다른 변수와 관련성 분석
- 예, 광고비와 매출액 사이의 관련성 등 분석

【상관관계분석 중요사항】

- 회귀분석 전 변수 간 관련성 분석(가설 검정 전 수행)
- 상관계수 ➔ **피어슨(Pearson) R계수** 이용 관련성 유무
  - ✓ 상관관계분석 척도:
  - ✓ 피어슨 상관계수(Pearson correlation coefficient : r)

1) 상관분석

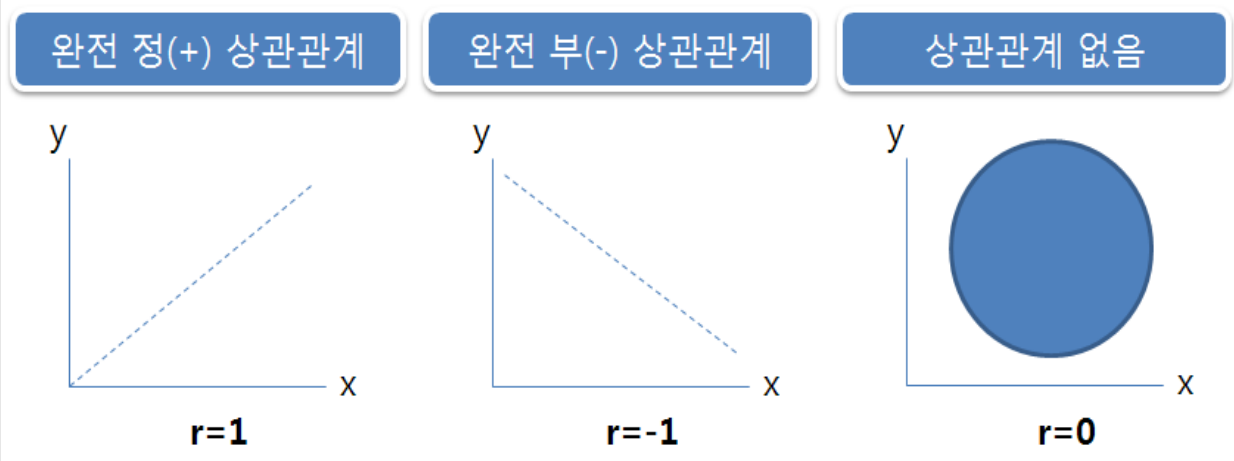
【피어슨 상관계수 R】

피어슨 상관계수 R	상관관계 정도
±0.9이상	매우 높은 상관관계
±0.9 ~ ±0.7	높은 상관관계
±0.7 ~ ±0.4	다소 높은 상관관계
±0.4 ~ ±0.2	낮은 상관관계
±0.2 미만	상관관계 없음

※ 상관계수 r은 -1에서 +1까지의 값을 가진다. 또한 가장 높은 완전 상관관계의 상관계수는 1이고, 두 변수간에 전혀 상관관계가 없으면 상관계수는 0이다.

1) 상관분석

- 상관계수 r과 상관관계 정도



1) 상관분석

```
# 데이터셋 가져오기
result <- read.csv("C:/Rwork/data/drinking_water.csv", header=T)
head(result)

# 상관계수 보기
cor(result$친밀도, result$적절성)
cor(result$친밀도, result$만족도)

# 전체 변수 간 상관계수 보기
cor(result, method="pearson") # 피어슨 상관계수 - default

cor(result, method="spearman") # spearman 상관계수(서열척도)
```

1) 상관분석

【논문에서 상관관계 분석 결과 제시 방법】

- 일반적으로 상관관계 분석 결과를 논문에서 제시할 경우 해당 기술통계량(평균과 표준편차)과 피어슨 상관계수 함께 제시

1) 상관분석

【논문에서 상관관계 분석 결과 제시 방법】

분석 단위	평균 (Mean)	표준편차 (Std. Deviation)	분석 단위 간 상관관계 (Inter-Analysis Correlations)		
			1	2	3
1. 친밀도	2.928	0.9703446	1		
2. 적절성	3.133	0.8596574	.499	1	
3. 만족도	3.095	0.8287436	.467	.767	1

2) 회귀분석

- **회귀분석(Regression Analysis)**
  - 특정 변수(독립변수)가 다른 변수(종속변수)에 어떠한 영향을 미치는가 (인과관계 분석)
  - 예) 가격은 제품 만족도에 영향을 미치는가?
  - 한 변수의 값으로 다른 변수의 값 예언

[참고] 인과관계(因果關係) : 변수A가 변수B의 값이 변하는 원인이 되는 관계(변수A : 독립변수, 변수B : 종속변수)

- 상관관계분석 : 변수 간의 관련성 분석
- 회귀분석 : 변수 간의 인과관계 분석

2) 회귀분석

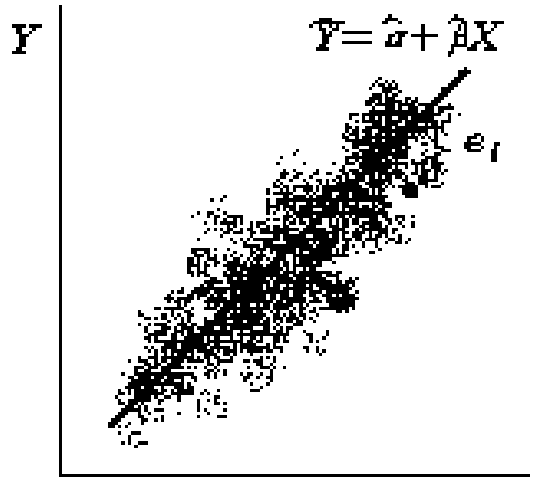
【회귀분석 중요사항】

- '통계분석의 **꽃**' ➔ 가장 강력하고, 많이 이용
- 종속변수에 영향을 미치는 변수를 규명(변수 선형 관계 분석)
- 독립변수와 종속변수의 관련성 강도
- 독립변수의 변화에 따른 종속변수 변화 예측
- **회귀 방정식**(Y=a+βX → Y:종속변수, a:상수, β:회귀계수, X:독립변수)을 도출하여 회귀선 추정
- 독립변수와 종속변수가 모두 <u>등간척도 또는 비율척도</u> 구성

2) 회귀분석

【회귀방정식】

- **회귀 방정식 -> 회귀선 추정**
  - ✓ Y=a+βX : Y:종속변수, a:상수, β:회귀계수, X:독립변수
- **회귀계수(β) : 단위시간에 따라 변하는 양(기울기)이며, 회귀선을 추정함에 있어 최소자승법 이용**
- **최소자승법 : 산포도에 위치한 각 점에서 회귀선에 수직으로 이르는 값의 제곱의 합 최소가 되는 선(정중앙을 통과하는 직선)을 최적의 회귀선으로 추정**



산포도와 회귀선

2) 회귀분석

● 단순 회귀분석

- ➢ 독립변수와 종속변수 각각 1개
- ➢ 독립변수가 종속변수에 미치는 인과관계 분석

【연구 가설】

단순 회귀분석을 수행하기 위한 연구 가설은 다음과 같다.

- 연구가설(H_1) : 음료수 제품의 당도와 가격수준을 결정하는 제품 적절성(독립변수)은 제품 만족도(종속변수)에 **정(正)**의 영향을 미친다.
- 귀무가설(H_0) : 음료수 제품의 당도와 가격수준을 결정하는 제품 적절성은 제품의 만족도에 영향을 미치지 않는다.

※ 논문에서는 **연구가설을 제시하고**, 귀무가설을 토대로 **가설 채택 또는 기각** 결정

2) 회귀분석

```
# 단순회귀분석
 형식) lm(formula= y ~ x 변수, data)
  # x:독립, y:종속, data=data.frame
  # lm() 함수 -> x변수를 대상으로 y변수 값 유추

str(result)
y = result$만족도 # 종속변수
x = result$적절성 # 독립변수
result.lm <- lm(formula=y ~ x, data=result)

# 단순선형회귀 분석 결과 보기
summary(result.lm)
```

2) 회귀분석

```
summary(result.lm)
#Coefficients: 계수
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  0.77886    0.12416   6.273 1.45e-09 ***
#x            0.73928    0.03823  19.340  < 2e-16 ***
#  ---
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Residual standard error: 0.5329 on 262 degrees of freedom
#Multiple R-squared:  0.5881,  Adjusted R-squared:  0.5865
#F-statistic:   374 on 1 and 262 DF,  p-value: < 2.2e-16
################<회귀분석 결과 해석>#########################
# 결정계수(Coefficients) : R-squared -> 0 ~ 1 사이의 값을 갖는다.
# Multiple R-squared:  0.5881:  독립변수에 의해서 종속변수가 얼마만큼 설명되었는가?
# 설명력 -> 상관(결정)계수 : 58.8% 설명력
# 1에 가까울 수록 설명변수(독립변수)가 설명을 잘한다고 판단
# 모형의 변수 선정이 우수하다는 의미.
# Adjusted R-squared:  0.5865 : 조정된 R값(오차를 감안한 값)<- 이것으로 분석
```

2) 회귀분석

################<t와 p value 추정>################

 # t value : 19.34 > ±1.96이고, p value : 2.2e-16 < 0.05(유의수준)

 # <해설> t value : 19.34는 ±1.96보다 크고, p값 2.2e-16은 유의수준(알파)

 # 보다 작다. 따라서 음료수 제품의 적절성은 제품 만족도에 정의 영향을 미친다.

 # 라는 연구가설을 채택한다.

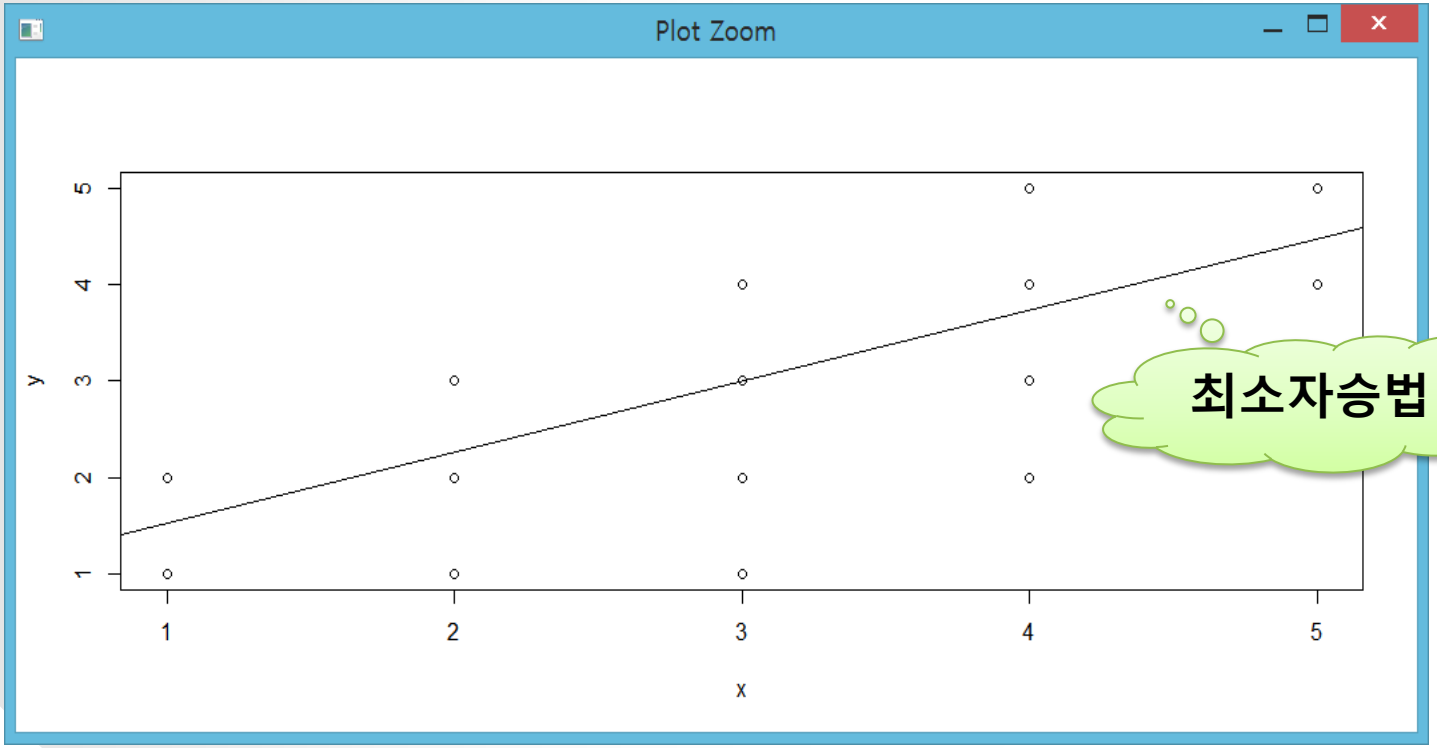
################<회귀방정식 유도>################

#회귀방정식 : Y(종속변수) = 상수 + 베타(회귀계수).X(독립변수)

#회귀방정식 : Y = 0.779 + 0.739 * 제품적절성(x1)

2) 회귀분석

- **회귀방정식에 의해서 회귀선 시각화**
  - ✓ X,Y가 선형 관계를 나타냄



2) 회귀분석

【검정통계량(t)와 유의수준(α) 관계】

더 알아보기

☞ 『t값과 유의수준α』 관계

t값(절대치)	유의수준α(양측검정 시)
t값(절대치) ≧ 2.58	α = 0.01(의생명분야)
t값(절대치) ≧ 1.96	α = 0.05(사회과학분야)
t값(절대치) ≧ 1.645	α = 0.1(기타 일반분야)

2) 회귀분석

【논문에서 단순 회귀분석 결과 제시 방법】

종속변수	독립변수	표준오차 (Std.Error)	검정통계량(t)	유의확률 (p)
제품만족도	상수	0.124	6.273	1.45e-09 ***
	제품적절성	0.038	19.340	< 2e-16 ***
분석 통계량	Multiple R-squared: 0.5881, Adjusted R-squared: 0.5865 F - statistic: 374 on 1 and 262 DF, p-value: < 2.2e-16			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

분산분석 :
회귀모델 적합성
(유의확률 0.05이상
부적합

2) 회귀분석

【논문에서 단순 회귀분석 결과 제시 방법】

- 음료수 제품의 당도와 가격수준을 결정하는 제품 적절성은 제품 만족도에 정(正)의 영향을 미칠 것이라는 연구가설을 검정한 결과 검정통계량 t=19.340, p=0.05미만으로 통계적 유의수준 하에서 영향을 미치는 것으로 나타났기 때문에 연구가설을 채택한다.
- 회귀모형은 상관계수 R=.767로 두 변수 간에 다소 높은 상관관계를 나타내며, R^2=.587로 제품 적절성 변수가 제품 만족도를 58.7% 설명하고 있다. 또한 회귀모형의 적합성은 F=374.020(p-value : < 2.2e-16)으로 회귀선이 모형에 적합하다고 볼 수 있다.

2) 회귀분석

【단순 회귀분석 결과 정리 및 기술】

▪ 가설 설정	연구가설(H_1) : 음료수 제품의 적절성은 제품 만족도에 **정(正)**의 영향을 미친다.	
	귀무가설(H_0) : 음료수 제품의 적절성은 제품 만족도에 영향을 미치지 않는다.	
1. 회귀식 모델 적합성	1) 유의순준	α = 0.05
	2) 검정통계량	*F* = 374.020
	3) 유의확률	P-value: < 2.2e-16
	4) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 회귀선이 모델에 적합하다고 볼 수 있다.
2. 가설검정	1) 유의수준	α = 0.05
	2) 검정통계량	t = 19.340
	3) 유의확률	*p* = < 2.2e-16
	4) 결과해석	유의수준 0.05에서 연구가설이 채택되었다. 따라서 제품 적절성이 높을 수록 제품 만족도가 높아지는 경향을 보이고 있다.

2) 회귀분석

● 다중 회귀분석

➢ 여러 개 독립변수가 1개의 종속변수에 미치는 영향 분석

【연구 가설】

다중 회귀분석을 수행하기 위한 연구 가설은 다음과 같다.

- 연구가설1(H_1) : 음료수 제품의 적절성(**독립변수1**)은 제품 만족도(종속변수)에 정(正)의 영향을 미친다.
- 연구가설2(H_1) : 음료수 제품의 친밀도(**독립변수2**)는 제품 만족도(종속변수)에 정(正)의 영향을 미친다.

2) 회귀분석

(1) 적절성 + 친밀도 -> 만족도

```
y <- result$만족도 # 종속변수
x1 <- result$적절성 # 독립변수
x2 <- result$친밀도 # 독립변수
result.lm <- lm(formula= y ~ x1 + x2, data=result)
summary(result.lm)
```

2) 회귀분석

(2) 학습데이터와 검증데이터 분석

```
# 단계1 : 7:3비율 데이터 샘플링
t <- sample(1:nrow(result), 0.7*nrow(result))

# 단계2 : 학습데이터와 검정데이터 생성
train <- result[t, ] # result중 70%
train # 학습데이터
test <- result[-t, ] # result중 나머지 30%
test # 검정 데이터

# 단계3 : 데이터 분석
result.lm <- lm(formula=만족도 ~ 적절성 + 친밀도, data=train)
summary(result.lm) # 학습데이터 분석
result.lm <- lm(formula=만족도 ~ 적절성 + 친밀도, data=test)
summary(result.lm) # 검정데이터 분석
```

2) 회귀분석

3) 다중공선성(Multicolinearity) 문제

- 독립변수 간의 강한 상관관계로 인해서 회귀분석의 결과를 신뢰할 수 없는 현상
- 생년월일과 나이를 독립변수로 갖는 경우
- 해결방안 : 강한 상관관계를 갖는 독립변수 제거

(1) 다중공선성 문제 확인

```
install.packages("car")
library(car)
fit <- lm(formula=Sepal.Length ~ Sepal.Width+Petal.Length+Petal.Width,
    data=train)
vif(fit)
sqrt(vif(fit))>2 # root(VIF)가 2 이상인 것은 다중공선성 문제 의심
```

2) 회귀분석

(2) iris 변수 간의 상관계수 구하기(단,Species제외)

```
cor(iris[,-5])
```

(3) 학습데이터와 검정데이터 분류

```
x <- sample(1:nrow(iris), 0.7*nrow(iris)) # 전체 70% 추출
train <- iris[x, ]
test <- iris[-x, ]
```

(4) Petal.Width 변수를 제거한 후 회귀분석

```
result.lm <- lm(formula=Sepal.Length ~ Sepal.Width+Petal.Length,
    data=train)
result.lm <- lm(formula=Sepal.Length ~ Sepal.Width+Petal.Length,
    data=test)
result.lm
summary(result.lm)
```

2) 회귀분석

【논문에서 다중 회귀분석 결과 제시 방법】

종속변수	독립변수	표준오차 (Std.Error)	검정통계량(t)	유의확률 (p)
제품만족도	상수	0.130	5.096	6.65e-07 ***
	제품적절성	0.044	15.684	< 2e-16 ***
	제품친밀성	0.039	2.478	0.0138 *
분석 통계량	Multiple R-squared: 0.5975, Adjusted R-squared: 0.5945 F-statistic: 193.8 on 2 and 261 DF, p-value: < 2.2e-16			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2) 회귀분석

【논문에서 다중 회귀분석 결과 제시 방법】

- 연구가설1(H1) : '음료수 제품의 적절성은 제품 만족도에 정(正)의 영향을 미친다.'와 연구가설2(H1) : '음료수 제품의 친밀도는 제품 만족도에 정(正)의 영향을 미친다.'를 분석을 위해서 다중 회귀분석을 실시하였다. 분석 결과를 살펴보면 제품 적절성이 제품 만족도에 미치는 영향은 t=15.684, p < 2e-16으로 유의수준 하에서 연구가설1이 채택되었으며, 제품 친밀도가 제품 만족도에 미치는 영향은 t=2.478, p=0.0138로 유의수준하에서 연구가설2가 채택되었다.
- 회귀모형은 상관계수 R=. 0.702으로 독립변수와 종속변수 간에 다소 높은 상관관계를 나타내며, R^2=.594로 독립변수가 종속변수를 59.4% 설명하고 있다. 회귀모형의 적합성은 F=374.020(p-value : < 2.2e-16)으로 나타나서 모형이 적합하다고 볼 수 있다.

2) 회귀분석

【다중 회귀분석 결과 정리 및 기술】

▪ 가설 설정	연구가설1(H_1) : 음료수 제품의 적절성은 제품 만족도에 **정(正)**의 영향을 미친다.	
	연구가설2(H_1) : 음료수 제품의 친밀도는 제품 만족도에 **정(正)**의 영향을 미친다.	
1. 회귀식 모델 적합성	1) 유의순준	α = 0.05
	2) 검정통계량	*F* = 193.8
	3) 유의확률	*P* = < 2.2e-16
	4) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 회귀선이 모델에 적합하다고 볼 수 있다.
2. 가설검정	1) 유의수준	α = 0.05
	2) 검정통계량	제품적절성(t=15.684), 제품친밀도(t=2.478)
	3) 유의확률	제품적절성(p < 2e-16), 제품친밀도(p=0.014)
	4) 결과해석	유의수준 0.05에서 연구가설이 채택되었다. 따라서 제품 적절성과 제품 친밀도가 높을 수록 제품 만족도가 높아지는 경향을 보이고 있다.

2) 회귀분석

<연습문제1> iris 데이터셋을 대상으로 다음과 같이 다중회귀분석을 수행하시오.

(결과 변수명 : result.lm)

조건1) 학습데이터(train), 검증데이터(test)를 7 : 3 비율 Sampling

조건2) y변수 : Sepal.Length, x변수 : Sepal.Width, Petal.Length, Petal.Width)

조건3) 1차분석 : train 데이터 분석, 2차 분석 : test 데이터 분석

2) 회귀분석

4) 새로운 값 예측 - predict()함수

```
# 회귀분석 결과를 대상으로 회귀방정식을 적용한 새로운 값 예측
# 형식) predict(x, data) data에 x변수(회귀분석결과) 값 존재 해야 함
result2 <- predict(result.lm, test)# x변수만 test에서 찾아서 값 예측
result2

head(test) # x,y값 확인
result2[1:5] # test 데이터에 회귀방정식이 적용된 계산결과 출력
```

2) 회귀분석

<연습문제2> 세 변수(평균구매금액,웹이용시간,구매다양성)를 대상으로
조건에 따라서 다중회귀 분석을 이용하여 총구매금액을 예측하시오.

```
# 데이터셋 가져오기
result <- read.csv("C:/Rwork/data/regression.csv", header=FALSE)

# 모양 변경 패키지 설치
install.packages("reshape")
library(reshape)
```

조건1) 변수이름 변경(reshape 패키지 이용)

```
# V1="총구매금액", V2="평균구매금액", V3="웹이용시간", V4="구매다양성“)
```

조건2) 다중회귀분석

조건3) 회귀방정식을 적용하여 total값 예측 및 출력

14. 분류 분석

Chap14_ClassificationAnalysis 수업내용

1) 데이터 셈플링
2) 분류모델 생성
3) 분류모델 예측
4) 분류모델 플로팅
5) rpart 패키지 활용

분류 분석

● 분류 분석?

분류 분석(classification analysis)은 다수의 속성(attribute) 또는 변수를 갖는 객체를 사전에 정해진 그룹 또는 범주(class, category) 중의 하나로 분류하여 분석하는 방법

● 활용분야

고객을 분류하는 변수, 규칙, 특성들을 찾아내고, 이를 토대로 미래 잠재 고객의 행동이나 반응을 예측하거나 유도하는데 활용

● 분류 분석 vs 군집 분석

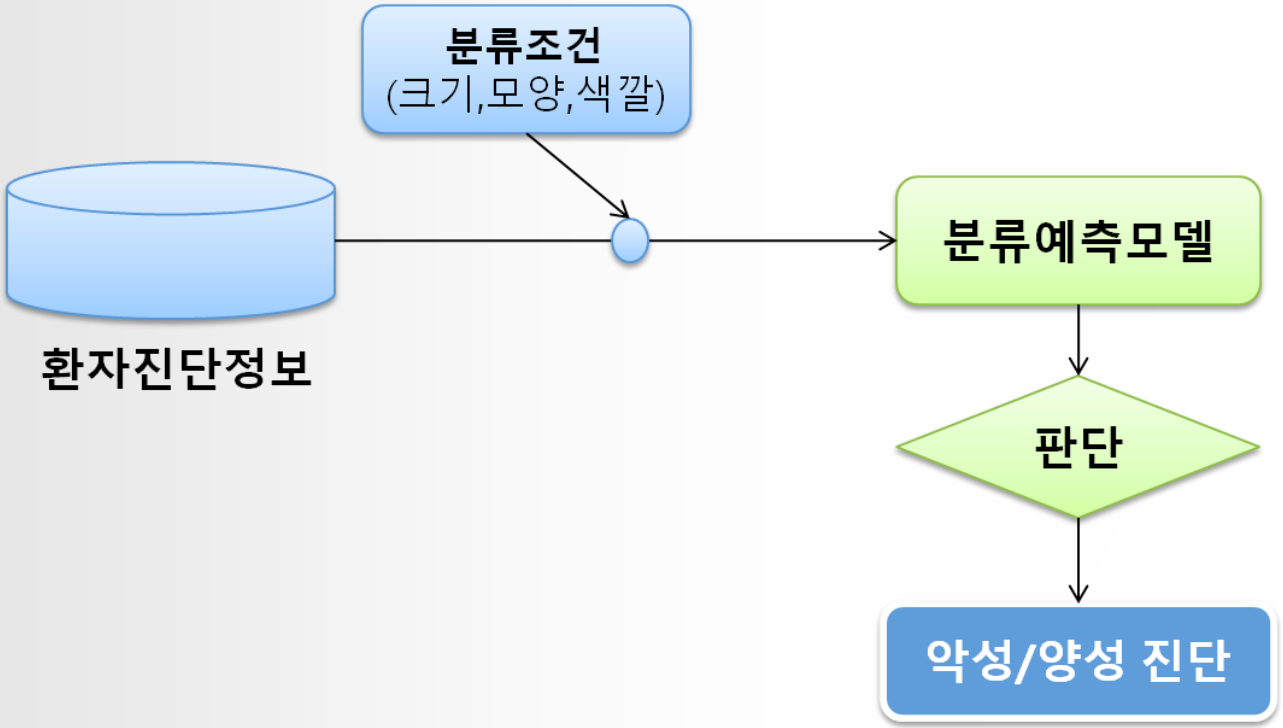
분류 분석은 이미 각 계급(클러스터)이 어떻게 정의 되는지 알고 있음

분류분석

- **분류분석 특징**
  - 종속변수(y변수) 존재
  - 종속변수 : 예측에 Focus을 두는 변수
  - 규칙(Rule)을 기반으로 의사결정트리 생성
  - 비모수 검정 : 선형성, 정규성, 등분산성 가정 필요 없음
  - 단점 : 유의수준 판단 기준 없음(추론 기능 없음)

분류분석

- 의.생명분야에서 분류분석 사례



분류 모델링

● 분류 모델링
데이터의 실체가 어떤 그룹에 속하는지 예측하는데 사용하는 데이터 마이닝 기법

● 의사결정나무
분류 모델링에 의해서 만들어진 규칙(rule)를 나무 모양으로 그리는 방법, 의사결정이 이뤄지는 시점과 성과를 한눈에 볼 수 있다.

● 의사결정나무활용
고객이 어떤 집단에 속하는지(세분화), 고객신용도에 따른 우량/불량(분류), 고객의 속성 따른 대출한도(예측), 예측변수 중 목표변수에 가장 큰 영향을 주는(변수선택)

분류분석 실습

- **실습내용**

  iris 데이터셋의 4개변수(Sepal Length,Sepal Width,Petal Length,Petal Width)값에 따라서 꽃의 종류(Species)가 분류되는 과정 알아보기

```
############### ##### 테스트 환경 ####################
# 전수 데이터를 대상으로 할 경우 error를 감안해서 학습데이터와
# 검증데이터로 분리 하여 테스트 한다.
# 학습데이터(훈련데이터) - (70%) -> 알고리즘 적용
# rule 추출 -> 검정 테스트 (30%) -> 확인(validate) 데이터
# 표본의 통계량으로 모집단의 모수 추론과 유사함
#################################################
```

1) 데이터 셈플링

- **분류분석 테스트 환경**
  - 전수 데이터를 대상으로 할 경우 error를 감안해서 학습데이터와 검정데이터로 분리 하여 테스트
  - 학습데이터(전수데이터의 70%) -> 알고리즘 적용 -> rule 발견
  - rule 적용 -> 검정데이터(전수데이터의 30%) -> 검정(validate)

  ❖ 표본의 통계량으로 모집단의 모수 추론 과정과 유사

구분	추론통계	데이터마이닝
데이터	표본	전수데이터
검정방법	통계량/추론	Rule/검증

1) 데이터 셈플링

단계1 : 학습데이터와 검증데이터 샘플링

```
result <- sample(2, nrow(iris),replace=T, prob=c(0.7,0.3))  # 7:3비율

table(result)
train <- iris[result==1,]
test <- iris[result==2,]

# formula 생성 : 형식) 변수 <- 종속변수 ~ 독립변수
formula<-Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width
```

2) 분류모델 생성

단계2 : 분류모델 생성(ctree()함수 이용)

```
# part패키지 설치
 install.packages("party") # ctree()함수 제공
 library(party)

# 학습데이터로 분류모델 생성
iris_ctree <- ctree(formula, data=train)

iris_ctree # Petal.Length,Petal.Width 중요변수
```

2) 분류모델 생성

ctree()에 의해서 생성된 분류모델 결과

Conditional inference tree with 4 terminal nodes

Response: Species
Inputs: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Number of observations: 99

1) Petal.Length <= 1.9; criterion = 1, statistic = 92.436
  2)* weights = 31
1) Petal.Length > 1.9
  3) Petal.Width <= 1.6; criterion = 1, statistic = 44.074
    4) Petal.Length <= 4.6; criterion = 0.994, statistic = 10.099
      5)* weights = 23
    4) Petal.Length > 4.6
      6)* weights = 9
  3) Petal.Width > 1.6
    7)* weights = 36

Petal.Length변수 – 중요변수
[규칙(Rule)]
1.9이상-31,
미만-32
기타-36

3) 분류모델 예측

단계3 : 분류모델 예측

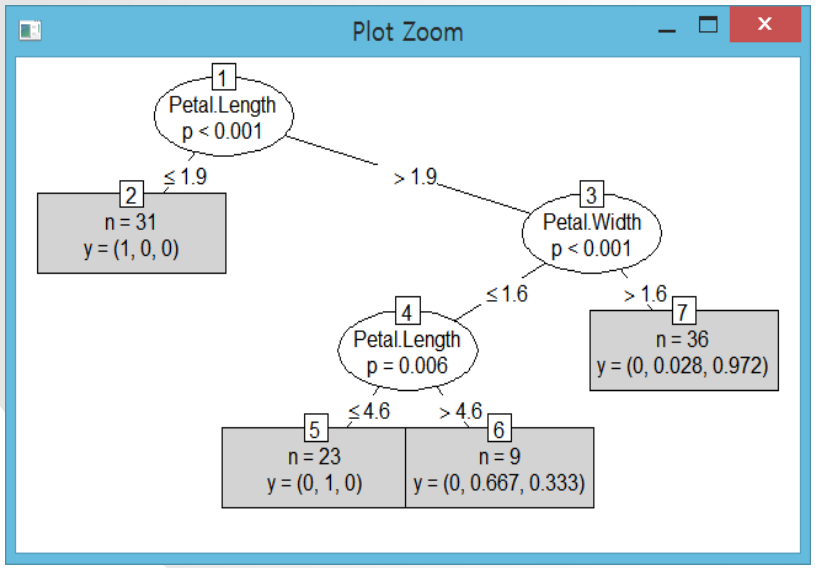
```
# predict() 이용 - 테이블 형태로 결과 제시
# 분류모델의 규칙을 적용하여 train데이터의 Species와 교차표 출력
table(predict(iris_ctree), train$Species)
```

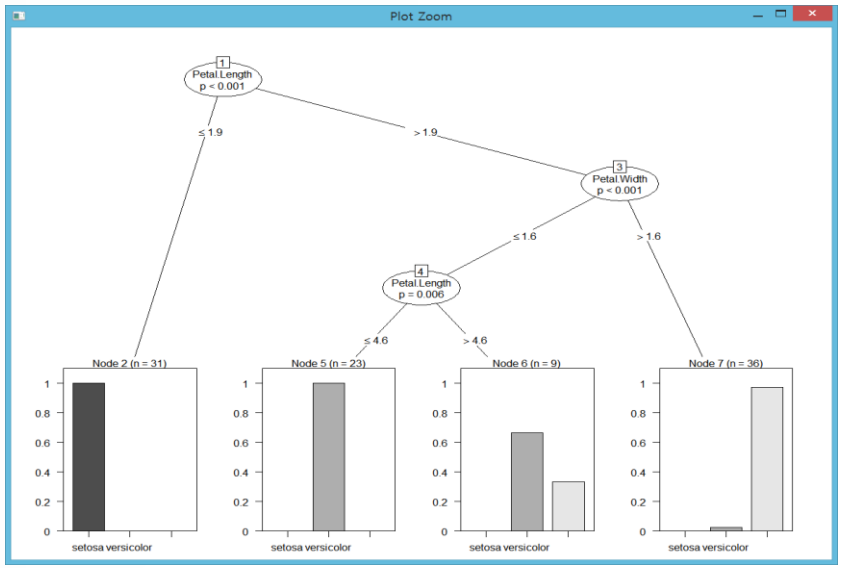
```
            setosa versicolor virginica
setosa          31          0         0
versicolor       0         29         3
virginica        0          1        35
```

4) 분류모델 플로팅

단계4 : 분류모델 플로팅

```
# plot() 이용 - 의사결정 트리 플로팅
plot(iris_ctree, type="simple")
plot(iris_ctree)
```





분류모델 실습

실습> 검증데이터를 이용하여 분류모델을 생성하고 테이블 형식과 의사결정트리를 플로팅하시오.

1) 분류모델 생성

```
iris_ctree2 <- ctree(formula, data=test)
```

2) 분류모델 예측

```
testpred <- predict(iris_ctree2, test) # 검증데이터 적용
table(testpred, test$Species)
```

3) 의사결정트리 플로팅

```
plot(iris_ctree2)
```

분류모델 연습문제

<연습문제1> classification.csv 파일을 tree로 읽어서 response가 y변수, total, price, period, variety 변수가 x변수가 되도록 하여 decision tree를 작성하시오.

1) 데이터 가져오기(header 없음)

```
result <- read.csv("C:/Rwork/data/classification.csv", header=FALSE)
```

2) 테이블 모양 변경 패키지 설치

```
install.packages("reshape")
library(reshape)
```

3) 변수명 지정

```
result <- rename(result, c(V1="total", V2="price", V3="period",
V4="variety", V5="response"))
```

분류모델 연습문제

4) sample()함수 이용 학습데이터와 검정데이터 7:3 비율로 셈플링

5) formula 생성 - 4개 변수에 대해서 NR,LOW,HIGH 구분

6) ctree() : 분류모델 생성

7) predict () : 분류모델 예측 -> 테이블 형식으로 예측결과 제시

8) 의사결정트리 플로팅

9) 의사결정트리 해석

10) 검정 데이터 적용 - 분류모델 예측(predict ()와 table()이용)

분류모델 연습문제

<연습문제2> 경력사원 채용기준에 대한 분류모델을 예측하시오.

1) 데이터 가져오기

```
result <- read.csv("C:/Rwork/data/human.csv", header=T)
head(result)
```

2) 학습데이터와 검정데이터 셈플링

```
resultsplit <- sample(2, nrow(result), replace=TRUE, prob=c(0.7, 0.3))
trainD <- result[resultsplit==1,]
testD <- result[resultsplit==2,]
```

3) formula 생성

```
formula <- Group ~ Sociability + Rating + Career + Score
```

분류모델 연습문제

4) ctree() : 학습데이터 적용 분류모델 생성 및 예측

5) 분류모델 플로팅

6) 검증데이터 적용 – 분류모델 예측

분류모델 연습문제

<연습문제3> A은행에서 실시한 개인대출 프로모션 반응에 대한 분류모델을 예측하시오.

1) 데이터 가져오기

```
result <- read.csv("C:/Rwork/data/loan.csv", header=T)
head(result)
```

2) 학습데이터와 검정데이터 셈플링

```
resultsplit <- sample(2, nrow(result), replace=TRUE, prob=c(0.7, 0.3))
trainD <- result[resultsplit==1,]
testD <- result[resultsplit==2,]
```

3) formula 생성

```
formula  <- Response ~ Age+Experience+Income+Family+CCAvg+Mortgage
  # Experience : 직장근무경력, Income: 연간급여, Family: 가족수,
  # CCAvg: 월평균 신용카드지출금액, Mortgage: 모기지론 대출금액
```

분류모델 연습문제

4) 분류모델 생성과 예측

5) 분류모델 플로팅

6) 검증데이터 적용

5) rpart 패키지 활용

- **rpart 패키지 적용 분류분석**
  - rpart() 함수
  - ctree()함수와 같은 분류모델 제공
  - 32개 트리까지 생성

<실습 내용1>

weather.csv를 weather로 읽어서 RainTomorrow가 y변수, Data, Location, RISK_MIM, RainTody를 제외한 나머지 변수가 x변수가 되도록 하여 decision tree를 작성

5) rpart 패키지 활용

1) 데이터 가져오기

```
# c:/Rwork/Part-IV/weather.csv 파일 선택
weather = read.csv(file.choose(), header=TRUE)
str(weather) # data.frame':  366 obs. of  24 variables:
names(weather) # 24개 변수명
head(weather)
```

2) 분류모델 생성 - rpart()함수 이용

```
install.packages("rpart") # 패키지 설치
library(rpart)
# 형식) rpart(y변수~., data set) # 점(.) : x변수는 data의 모든 변수
weather.df <- rpart(RainTomorrow~., data=weather[, c(-1,-2,-23,-22)])
weather.df
```

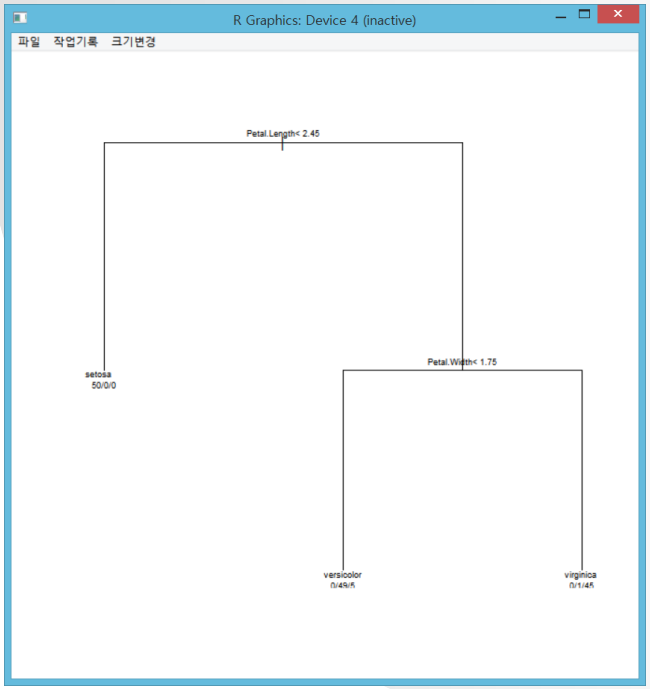
5) rpart 패키지 활용

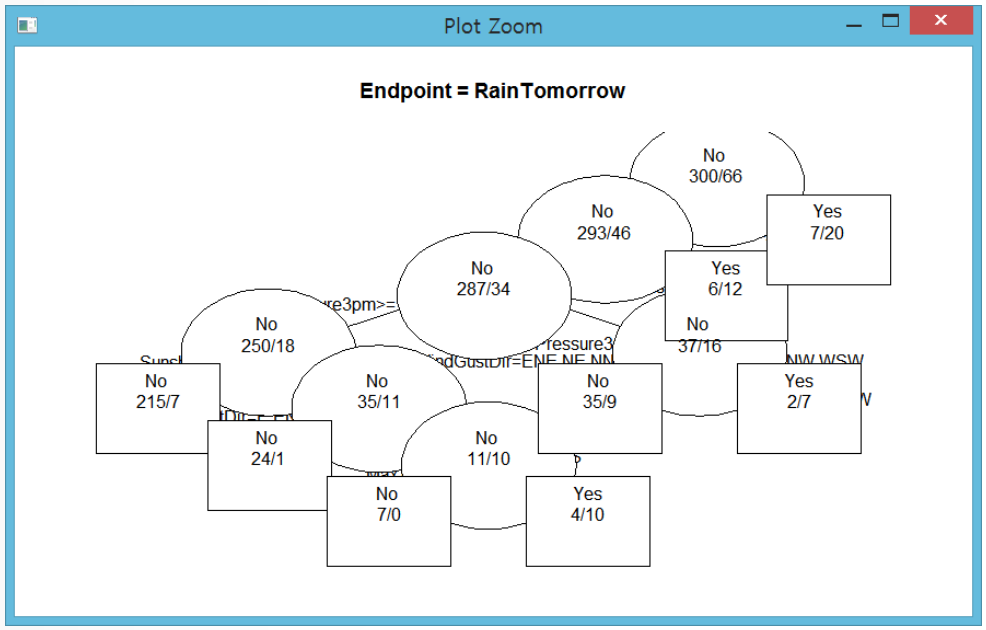
3) 분류 트리 생성

plot(weather.df) # 트리 프레임 보임

text(weather.df, use.n=T) # 텍스트 추가

post(weather.df, file="") # 타원제공 - rpart 패키지 제공





5) rpart 패키지 활용

<실습 내용2>

weather(전체 데이터)를 7:3으로 나누어 weather_train, weather_test로 저장한 후 weather_train으로 분류모델을 생성하고, weather_test로 예측하시오. 예측 결과가 50%를 기준으로 이상이면 비가 오는 것으로(Rain), 작으면 비가 안오는(No Rain)것으로 해서 테이블을 작성하시오.

5) rpart 패키지 활용

(1) 데이터 셈플링

```
index = sample(1:nrow(weather), 0.7*nrow(weather))
weather_train = weather[index,]
weather_test = weather[-index,]
```

(2) 분류모델 생성

```
weather.dt <- rpart(RainTomorrow~., data=weather_train[,c(-1,-2,-23,-22)])
weather.dt
```

5) rpart 패키지 활용

(3) 분류모델 예측 - 검정데이터로 예측

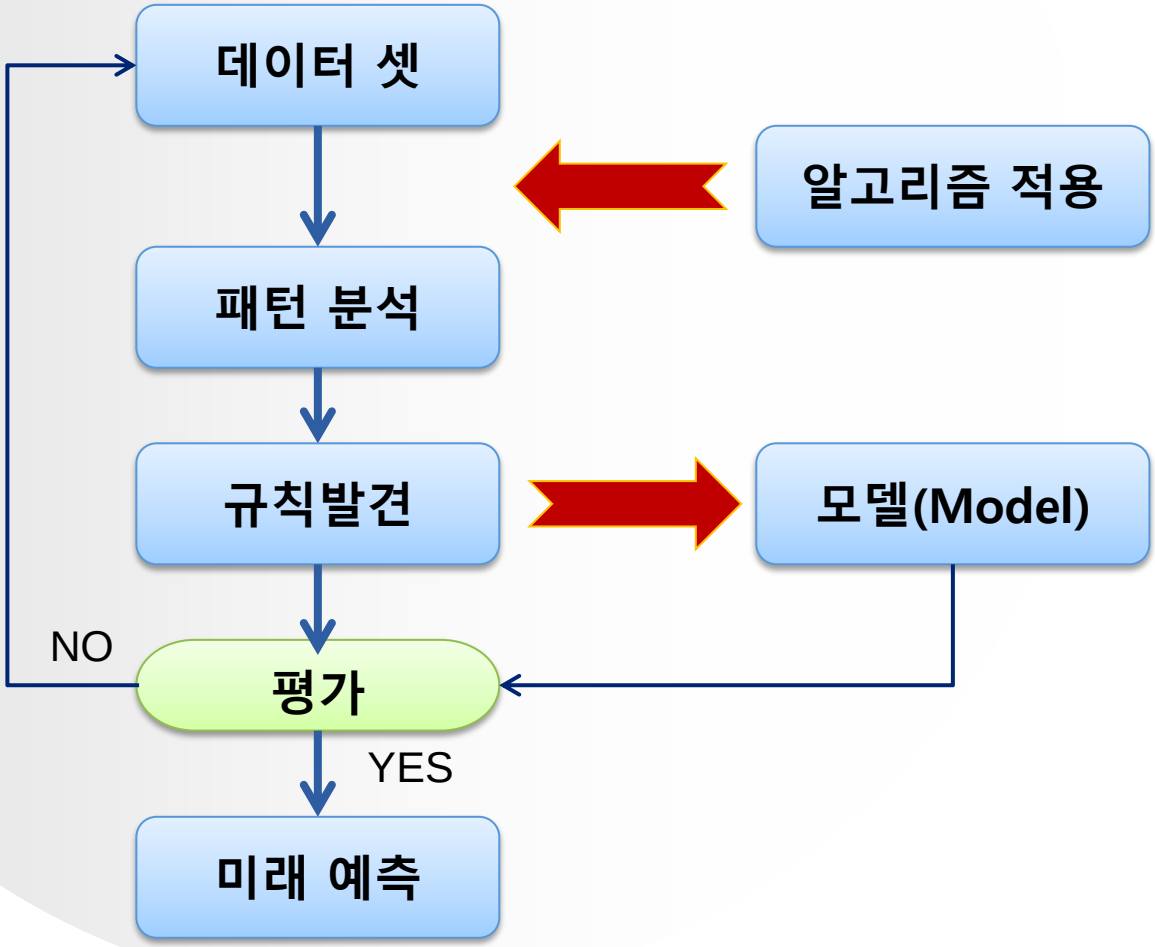
```
predict(weather.dt, weather_test[1:10,])

weather.predicted = predict(weather.dt, weather_test) # 예측
weather.predicted
result = ifelse(weather.predicted[,2]>0.5, "Rain", "No") # [,2] : Yes
#result = ifelse(as.numeric(weather.predicted[,2])>0.5, "Rain", "No")
```

(4) 분류모델 예측 결과

```
table(weather_test$RainTomorrow, result)
```

예측분석 방법 분류

- 비지도학습(unSupervised Learning) 절차



15. 군집 분석

Chap15_ClusteringAnalysis 수업내용

1) 군집분석 개요

2) 유클리드 거리

3) 계층적 군집분석

4) 비계층적 군집분석

1) 군집 분석 개요

- **군집 분석?**
  - 종속변수(y변수)가 없는 데이터 마이닝 기법
  - 유클리드 거리 기반 유사 객체 묶음
  - 고객 DB -> 알고리즘 적용 -> 패턴 추출(rule) -> 근거리 모형으로 군집형성
  - 계층적 군집분석(탐색적), 비계층적 군집분석(확인적)
  - 주요 알고리즘 : k-means, hierarchical

1) 군집 분석 개요

- **군집분석 특징**
  - 전체적인 데이터 구조를 파악하는데 이용
  - 관측대상 간 유사성을 기초로 비슷한 것 끼리 그룹화(Clustering)
  - 유사성 = 유클리드 거리
  - 분석결과에 대한 가설 검정 없음(타당성 검증 방법 없음)
  - 분야 : 사회과학, 자연과학, 공학 분야
  - 척도 : 등간, 비율척도(연속적인 양)

- **유클리드 거리 계산식**

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

관측대상 p와 q의 대응하는 변량값의 차가 작으면, 두 관측대상은 유사하다고 정의하는 식

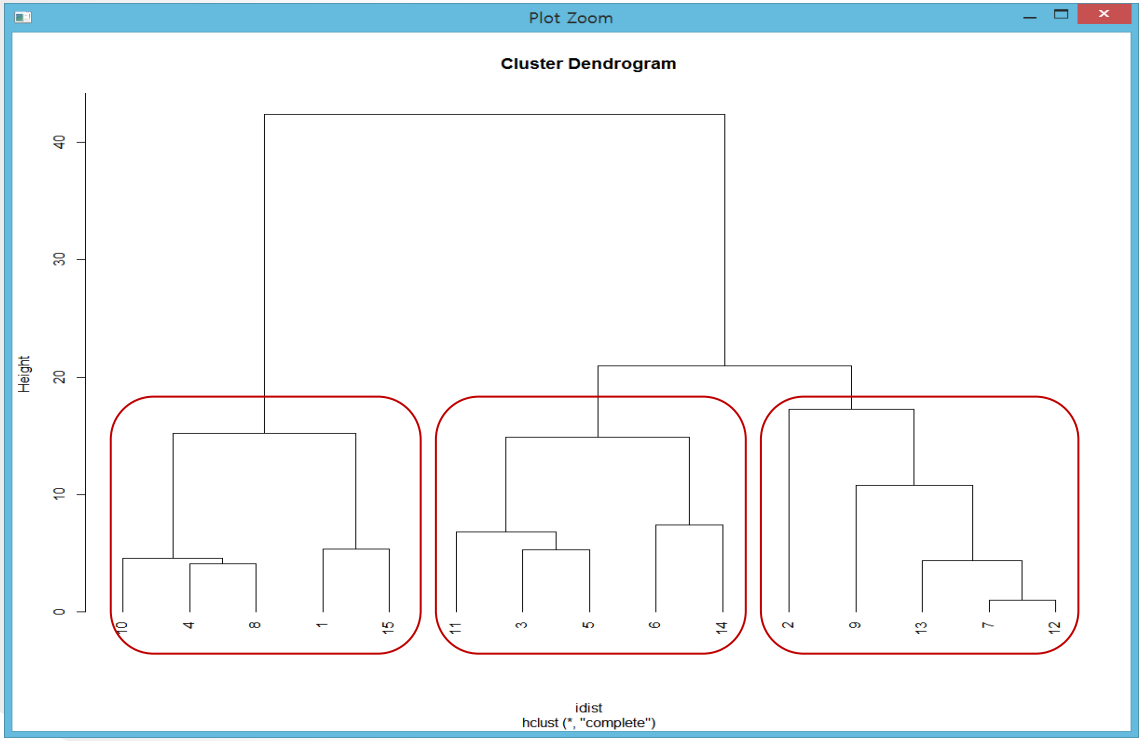
1) 군집 분석 개요

- **군집 구성법**
  - 그룹간의 유사성 계산 방법
  - 최단거리법, 최장거리법, 메디안법, 중심법, 그룹평균법

- **군집화방법**
  - 계층군집화 : 가장 가까운 대상끼리 순차적으로 묶음
  - 비계층군집화 : k-평균 군집법

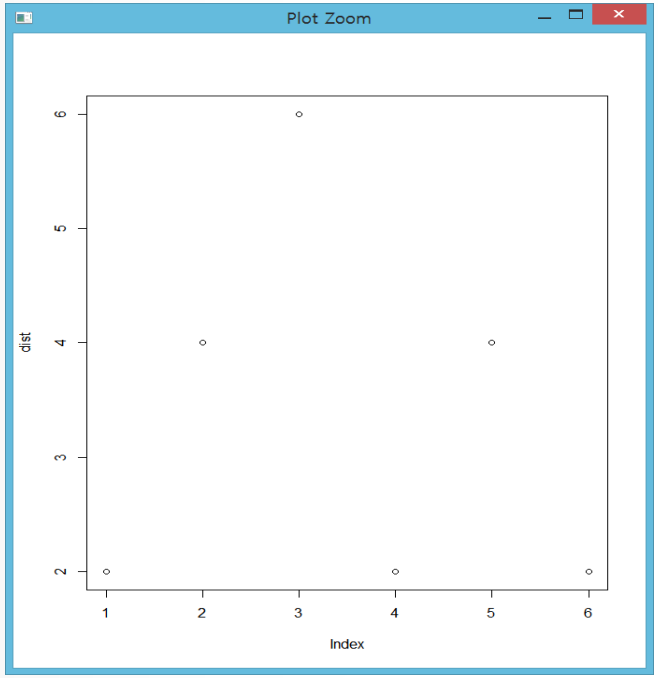
1) 군집 분석 개요

- **군집 분석 결과**
  - 평균결합방식을 적용한 덴드로그램(Dendrogram)
  - 가로축 : 학생번호, 세로축 : 상대적 거리
  - 군집수는 사용자가 정할 수 있음(2집단, 3집단 등)



2) 유클리드 거리

- **유클리드 거리(Euclidean distance)**
  - 두 점 사이의 거리를 계산하는 방법
  - 이 거리를 이용하여 유클리드 공간 정의



2) 유클리드 거리

- **유클리드 거리 실습**

(1) matrix 생성

```
x <- matrix(1:16, nrow=4)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    5    9   13
[2,]    2    6   10   14
[3,]    3    7   11   15
[4,]    4    8   12   16
```

(2) matrix 대상 유클리드 거리 생성 함수

```
# x : numeric matrix, data frame
dist <- dist(x, method="euclidean") # method 생략가능
```

```
#  1 2 3
#2 2
#3 4 2
#4 6 4 2    <- 가까운 객체 끼리 묶어줌
```

(3) 유클리드 거리 계산 식

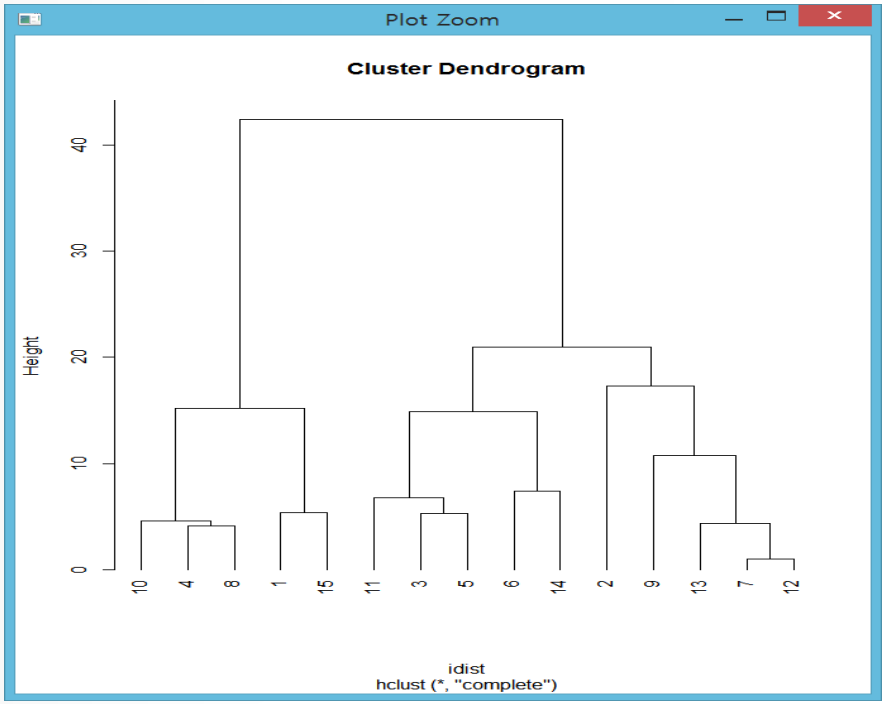
```
sqrt(sum((x[1,]-x[4,])^2)) # 6
```

<유클리드거리 계산법>

1. 두 벡터의 차이를 구한다.
2. 원소를 제곱해서 더한다.
3. 제곱근을 취한다.

3) 계층적 군집 분석

- **계층적 군집분석**
  - 유클리드 거리를 이용한 군집분석 방법
  - cluster 패키지에서 제공되는 hclust() 함수 이용
  - 계층적(hierarchical)으로 군집 결과 도출
  - 탐색적 군집분석



3) 계층적 군집 분석

- **계층적 군집분석 절차**

(1) 군집분석(Clustering)분석을 위한 패키지 설치
install.packages("cluster") # hclust() : 계층적 클러스터 함수 제공
library(cluster) # 일반적으로 3~10개 그룹핑이 적정

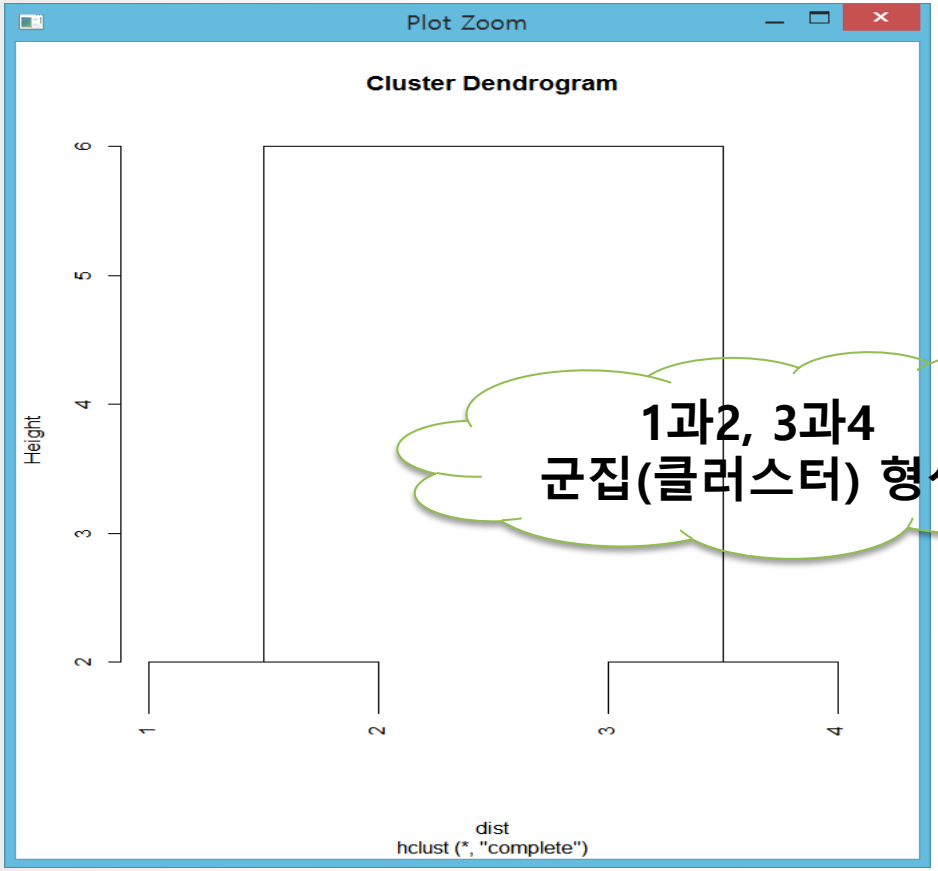
(2) 데이터 셋 생성
x <- matrix(1:16, nrow=4)

(3) matrix 대상 유클리드 거리 생성 함수
dist <- dist(x, method="euclidean") # method 생략가능

(4) 유클리드 거리 matrix를 이용한 클러스터링
hc <- hclust(dist) # 클러스터링 적용
plot(hc) # 클러스터 플로팅

3) 계층적 군집 분석

- 계층적 군집분석 결과 : 벤드로그램(dendrogram)



1과2, 3과4
군집(클러스터) 형성

3) 계층적 군집 분석

- **<실습1> 중1학년 신체검사 결과에 대한 군집분석**
  - 악력, 신장, 체중, 안경유무 칼럼 대상

(1) 데이터 셋 가져오기

```
body <- read.csv("c:/Rwork/Part-Iv/bodycheck.csv", header=TRUE)
body[,-1] # 1칼럼 제외
```

(2) 유클리드 거리

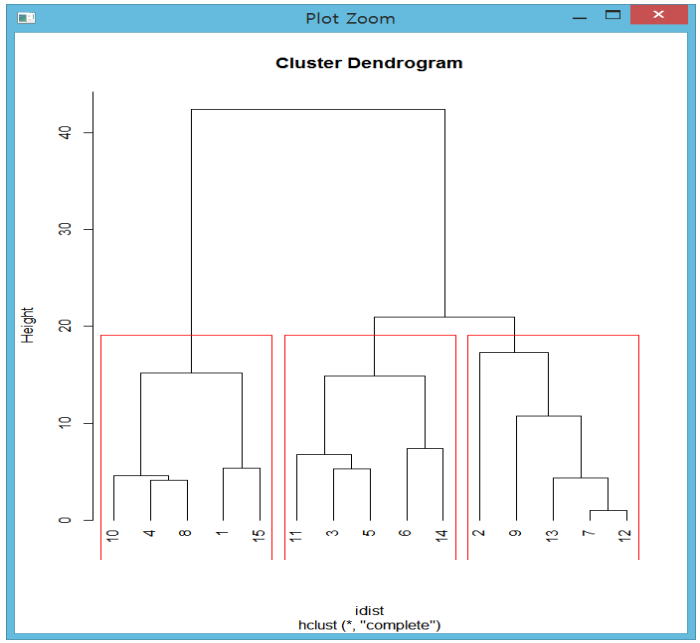
```
idist<- dist(body[, -1])
```

(3) 클러스터링

```
hc <- hclust(idist)
```

(4) 클러스터링 플로팅

```
plot(hc, hang=-1) # 음수값 제외
```



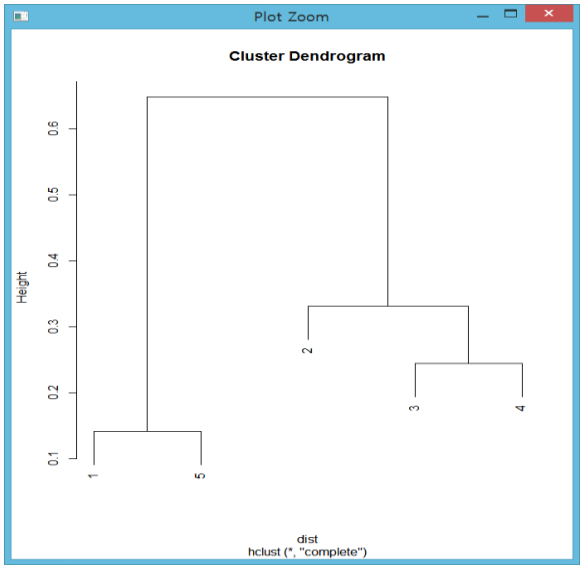
(5) 3개 그룹 선정, 선 색 지정

```
rect.hclust(hc, k=3, border="red") # 3개 그룹 선정, 선 색 지정
```

3) 계층적 군집 분석

- **<실습1> iris 데이터 셋 군집분석**
  - 대상 : 1~5행까지, 5열 제외

```
data(iris)

# 유클리드 거리구하기-iris 데이터 셋으로 유클리드 거리 계산
dist <- dist(iris[1:5, -5]) # 5컬럼 제외
#          1         2         3         4
#2 0.5385165
#3 0.5099020 0.3000000
#4 0.6480741 0.3316625 0.2449490
#5 0.1414214 0.6082763 0.5099020 0.6480741

# matrix를 이용한 클러스터링
hc <- hclust(dist)

# 계층적 clustering 그래프
plot(hc) 그래프 표현
```



해석 : 1과5, 2,3,4가 클러스터링

계층적 군집 분석 연습문제

<연습문제1> iris 1~4변수 전체를 대상으로 유클리드 거리 매트릭스를 구하여 idist에 저장하시오. 또한 저장된 idist merix로 계층적 클러스터링을 적용해서 hclust 결과를 그리시오.

조건> 4개 그룹 선정, 선 색 지정

3) 계층적 군집 분석

- **계층적 군집 결과에 그룹 수 지정**

  hclust() 함수에 의해서 군집한 결과를 지정한 그룹 수로 자르기

  <실습> iris의 계층형군집결과에 그룹수를 지정하여 그룹수 만큼 잘라서 iris의 1번째(Sepal.Length)와 3번째(Petal.Length) 변수를 대상으로 클러스터별 변수의 평균 구하기 - ddply() 이용

```
# 준비
idist<- dist(iris[1:4]) # dist(iris[, -5])
hc <- hclust(idist)
plot(hc, hang=-1)
rect.hclust(hc, k=4, border="red") # 4개 그룹수

(1) 그룹수 만들기 : cutree()함수 -> 지정된 그룹수 만큼 자르기
ghc<- cutree(hc, k=3)
#cutree(계층형군집결과, k=그룹수) -> 그룹수 만큼 자름
ghc #  150개(그룹을 의미하는 숫자(1~3) 출력)
```

3) 계층적 군집 분석

- **계층 군집 결과에 그룹 수 지정**

```
# iris에서 ghc값을 갖는 ghc라는 새로운 이름의 컬럼 추가
iris$ghc <- ghc
table(ghc) # ghc 빈도수
#ghc
#1  2  3
#50 72 28-> 150개
```

<클러스터별 평균 계산 결과>

ghc	Sepal.Length	Petal.Length
1	5.006000	1.462000
2	6.545833	5.273611
3	5.532143	3.960714

```
# (2) 패키지 설치
install.packages("plyr")
library(plyr)

# (3) ddply() 함수 이용
# 형식) ddply(dataframe, .(집단변수), 요약집계, 컬럼명=함수(변수))
ddply(iris, .(ghc), summarize, Sepal.Length=mean(Sepal.Length),
      Petal.Length=mean(Petal.Length))
```

4) 비 계층적 군집 분석

- **비계층적 군집 분석(k-means)**
  - 확인적 군집분석 방법
  - 계층적 군집분석법 보다 속도 빠름
  - 군집의 수를 알고 있는 경우 이용
  - K는 미리 정하는 군집 수
  - 계층적 군집화의 결과에 의거하여 군집 수 결정
  - 순차적 군집분석법(군집과정 반복)
  - 변수 보다 관측대상 군집화에 많이 이용
  - 군집의 중심(Cluster Center) 사용자가 정함

4) 비 계층적 군집 분석

- **계층적 vs 비 계층적 군집분석**

mydata <- read.csv("c:/Rwork/Part-IV/clustering.csv", header=TRUE)

mydata

total(총구매액), price(평균구매액)

period(웹이용시간), variety(구매다양성)

1) 계층적 군집분석(탐색적 분석)

result <- hclust(dist(mydata), method="ave")

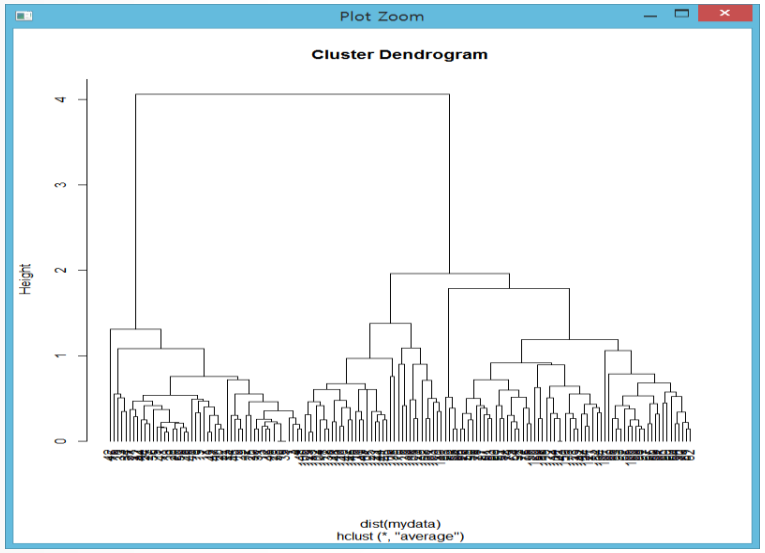
dist : 거리(4개변수 비교하여 거리구현)

method="ave" : 클러스터 방법(평균(average)거리 방식)

4) 비 계층적 군집 분석

- **계층적 vs 비 계층적 군집분석**

 # 군집분석 결과 변수
names(result) # result에서 제공하는 속성 변수 확인
result$order # 번호값
result$height # 클러스터 높이
result$method # "average"
plot(result, hang=-1) # hang : -1 이하 값 제거



탐색적 군집분석으로 군집수(3개) 확인

3) 비 계층적 군집 분석

2) 비계층적 군집분석(확인적 분석)

(1) 원형 데이터에 군집수 지정 # kmeans(data, k) : k개수: 군집수

result2 <- kmeans(mydata, 3)

result2 # 원형데이터를 대상으로 3개 군집으로 군집화

#Cluster means: 각 군집별 변수의 평균

total price period variety

#1 6.314583 4.973958 1.7031250 2.895833 <- 96

#2 4.739130 1.760870 0.3347826 2.934783 <- 23

#3 5.203226 1.477419 0.2774194 3.632258 <- 31

names(result2)

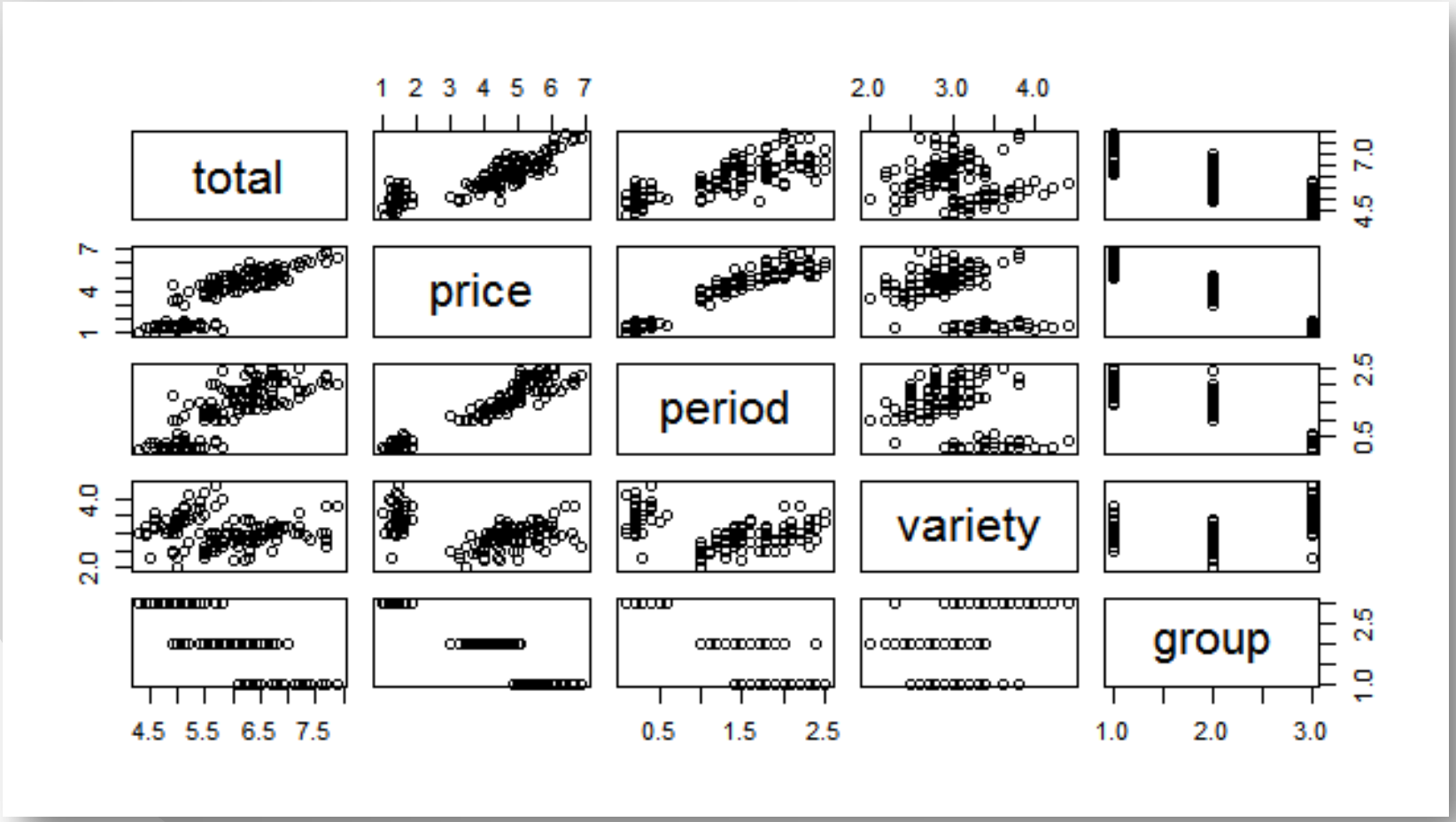
result2$cluster # 각 케이스에 대한 소속 군집수(1,2,3)

3) 비 계층적 군집 분석

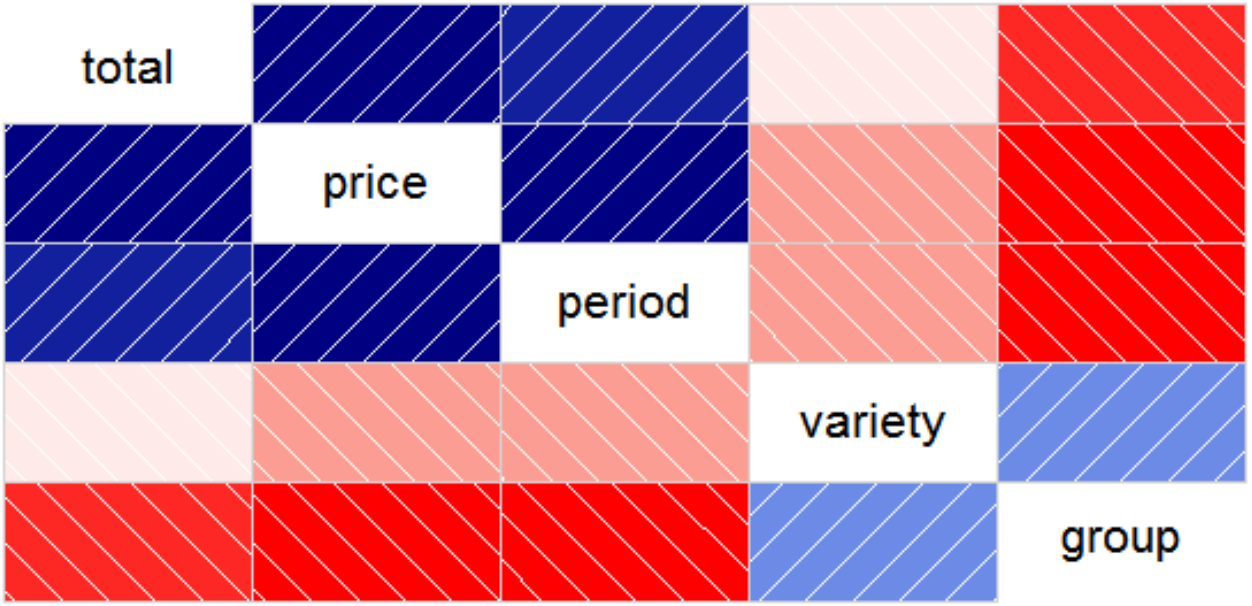
(2) 원형데이터에 군집수 추가

```
mydata$group <- result2$cluster
head(mydata)
# total price period variety group

# 변수의 관계
plot(mydata[,-5]) # 4개 변수(group 제외) 관계를 종합적으로 보여줌
```

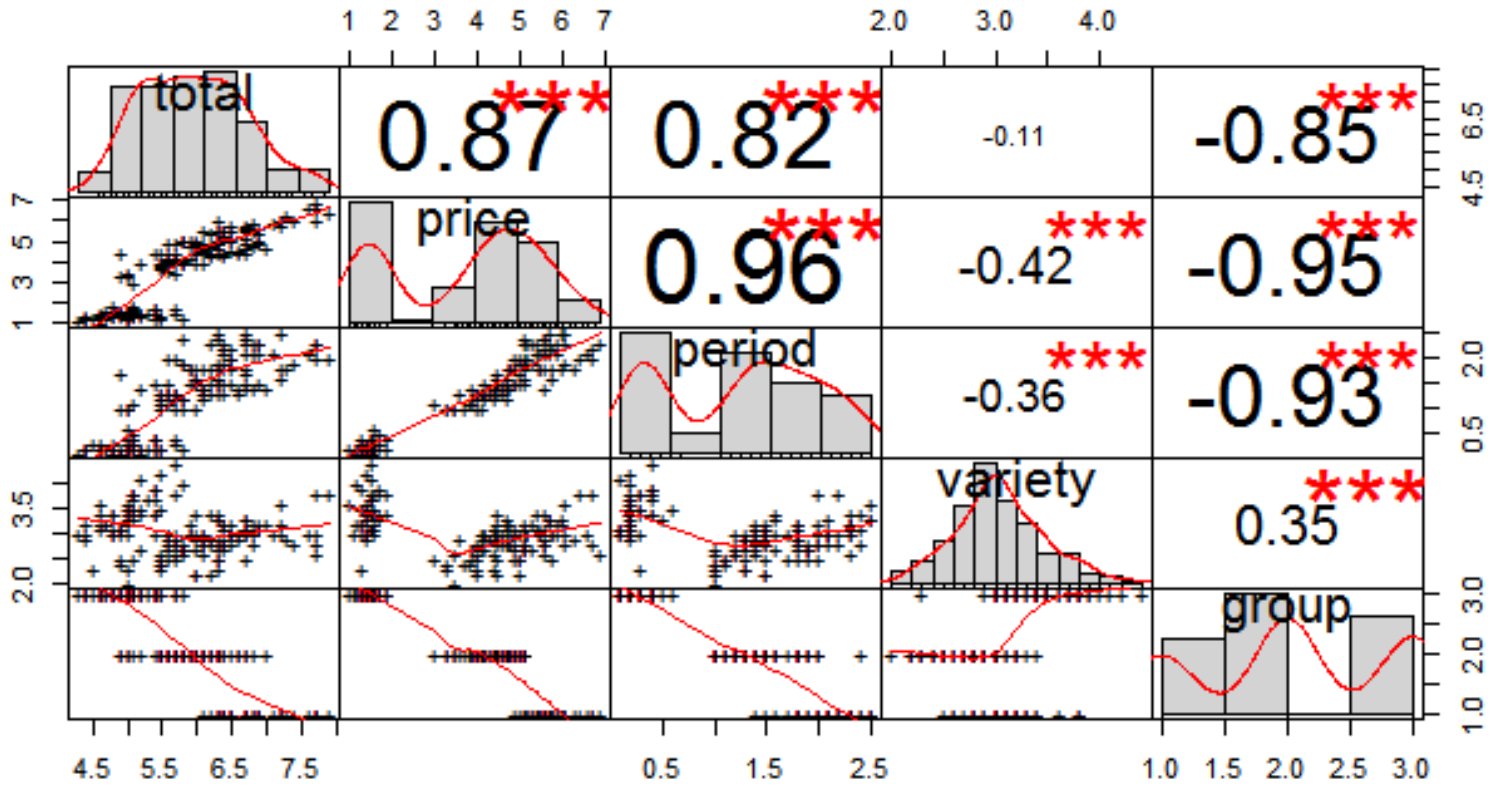
군집분석


total
price
period
variety
group
1 2 3 4 5 6 7
2.0 3.0 4.0
4.5 7.0
1 4 7
0.5 2.5
2.0 4.0
1.0 2.5
4.5 5.5 6.5 7.5
0.5 1.5 2.5
1.0 2.0 3.0

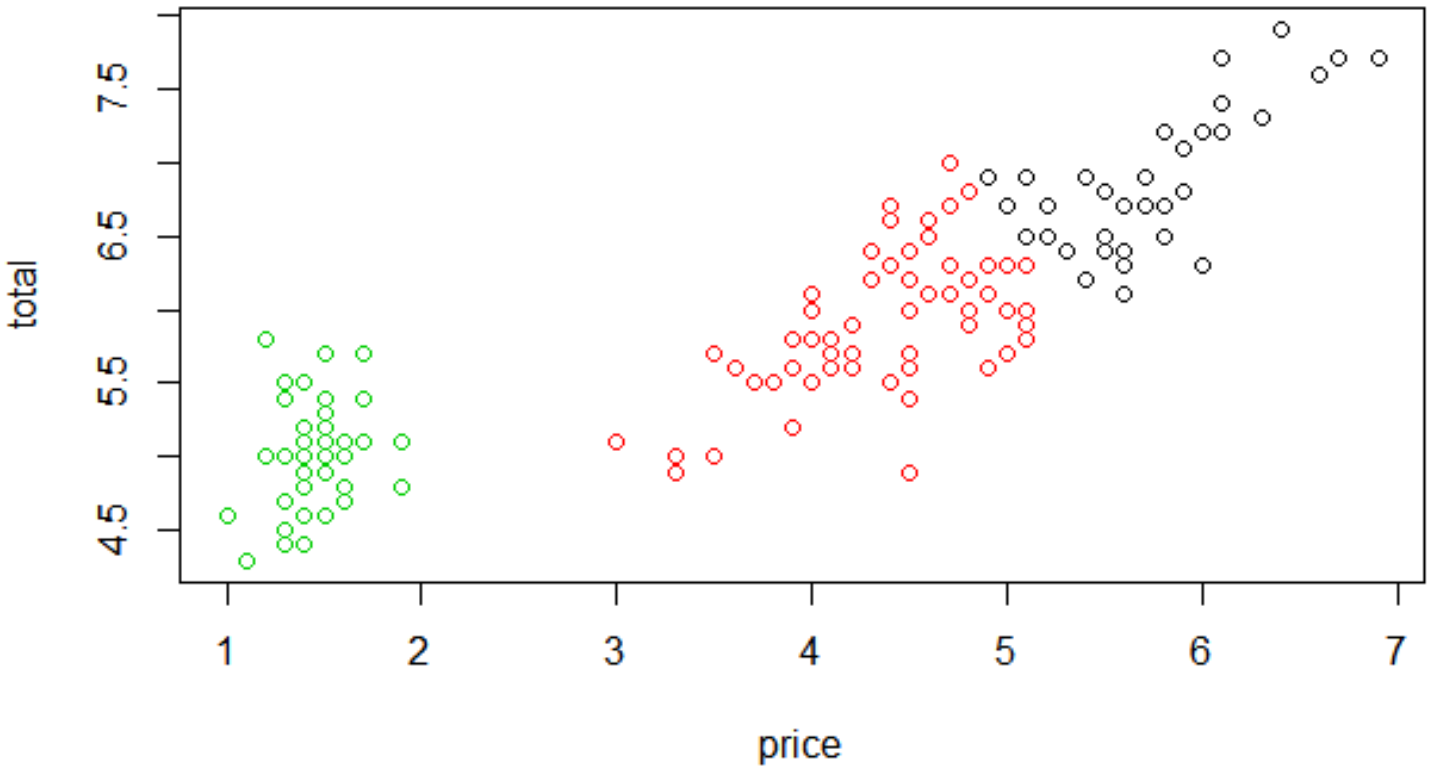
군집분석



군집분석


total
price
period
variety
group
0.87***
0.82***
-0.11
-0.85***
0.96***
-0.42***
-0.95***
-0.36***
-0.93***
0.35***
1 2 3 4 5 6 7
2.0 3.0 4.0
4.5 5.5 6.5 7.5
0.5 1.5 2.5
1.0 1.5 2.0 2.5 3.0
1 3 5 7
2.0 3.5
4.5 6.5
0.5 2.0
1.0 2.0 3.0

군집분석



16. 연관 분석

chap16_AssociationAnalysis 수업내용

1) 연관분석 개요
2) 연관규칙 생성
3) Adult 내장 데이터를 이용한 연관규칙 생성
4) single format transaction 데이터 처리
5) basket format transaction 데이터 처리

1) 연관 분석 개요

- **연관분석**
  - 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건
  - 데이터베이스에서 사건의 연관규칙을 찾는 무방향성 데이터 마이닝 기법
  - 마케팅에서 고객의 장바구니에 들어있는 품목 간의 관계 탐구
  - y변수가 없는 비지도 학습에 의한 패턴 분석
  - 사건과 사건 간 연관성(관계)를 찾는 방법(예:기저귀와 맥주)

    예) 장바구니 분석 : 장바구니 정보를 트랜잭션이라고 하며,

    트랜잭션 내의 연관성을 살펴보는 분석기법
  - 분석절차 : 거래내역 -> 품목 관찰 -> 규칙(Rule) 발견

1) 연관 분석 개요

- 관련분야 : 대형 마트, 백화점, 쇼핑몰 판매자 -> 고객 대상 상품추천

1. 고객들은 어떤 상품들을 동시에 구매하는가?
2. 라면을 구매한 고객은 주로 다른 어떤 상품을 구매하는가?

활용방안 : 위와 같은 질문에 대한 분석을 토대로 고객들에게

1) 상품정보 발송
2) 텔레마케팅를 통해서 패키지 상품 판매 기획,
3) 마트의 상품진열

1) 연관 분석 개요


What comes to your mind?
와인, 바나나, 사과 몇 알, 올리브, 오렌지주스, 귤을 구입했군!
고객의 신상 정보와 구매 내역이 서로 관계가 있을까?
일반적으로 바나나와 와인를 같이 구매하나? 그렇다면, 와인의 브랜드도 관련이 있나?
올리브와 오렌지 주스를 구입하면, 와인도 같이 사나?
보통은 같이 구매되는데 여기에는 없는 게 뭐가 있지?

1) 연관 분석 개요

- **연관규칙 평가척도**

1. 지지도(support) : 전체자료에서 관련 품목의 거래 확률
   - A->B 지지도 식 = A와 B를 포함한 거래수 / 전체 거래수
   - A를 구매한 후 B를 구매하는 거래 비율

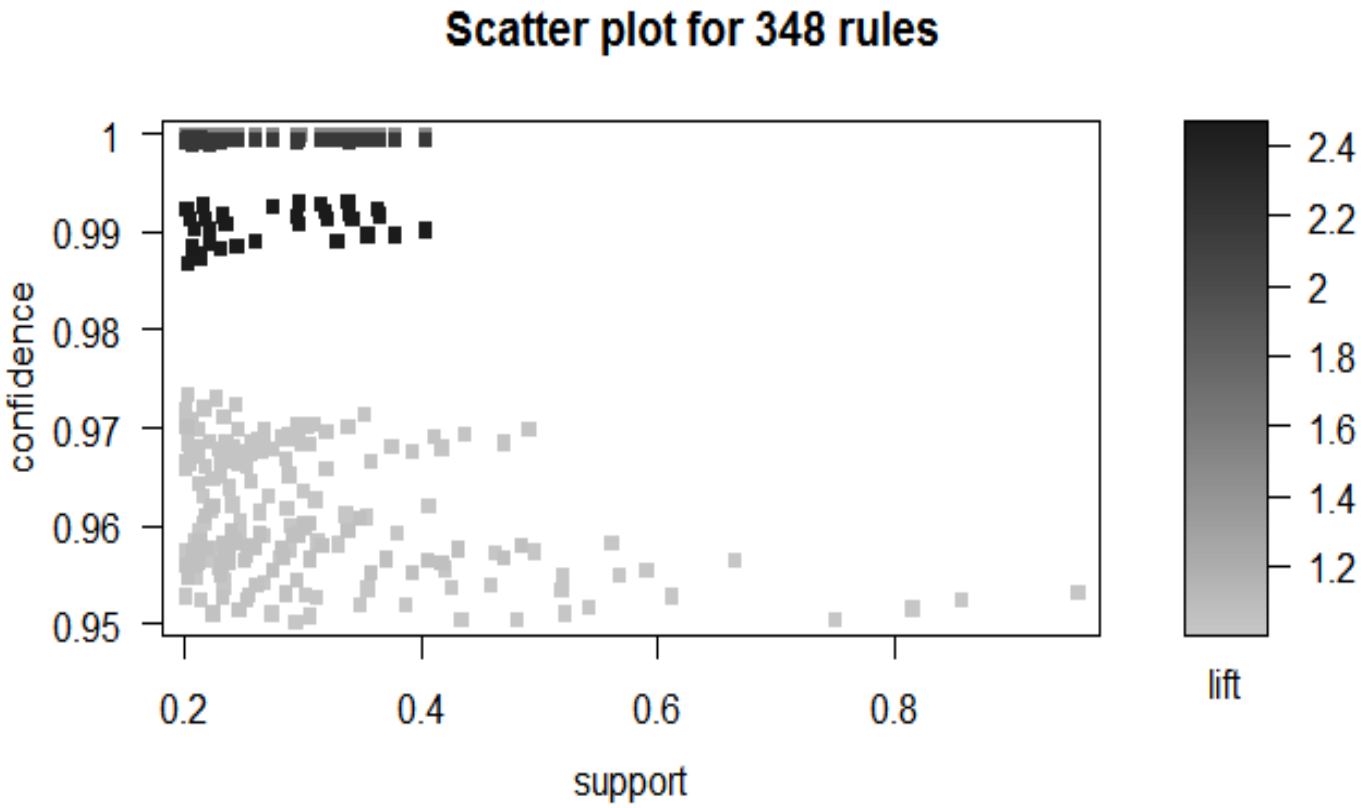
2. 신뢰도(confidence) : A가 구매될 때 B가 구매될 확률(조건부 확률)
   - A->B 신뢰도 식 = A와 B를 포함한 거래수 / A를 포함한 거래수
   - A가 포함된 거래 중에서 B를 포함한 거래의 비율

3. 향상도(Lift) : 상품 간의 독립성과 상관성을 나타내는 척도
   - 향상도 식 = 신뢰도 / B가 포함될 거래율
   - 향상도가 1에 가까우면 : 두 상품이 독립(과자와 후추)
   - 1보다 작으면 : 두 상품이 음의 상관성(설사약과 변비약)
   - 1보다 크면 : 두 상품이 양의 상관성(빵과 버터)

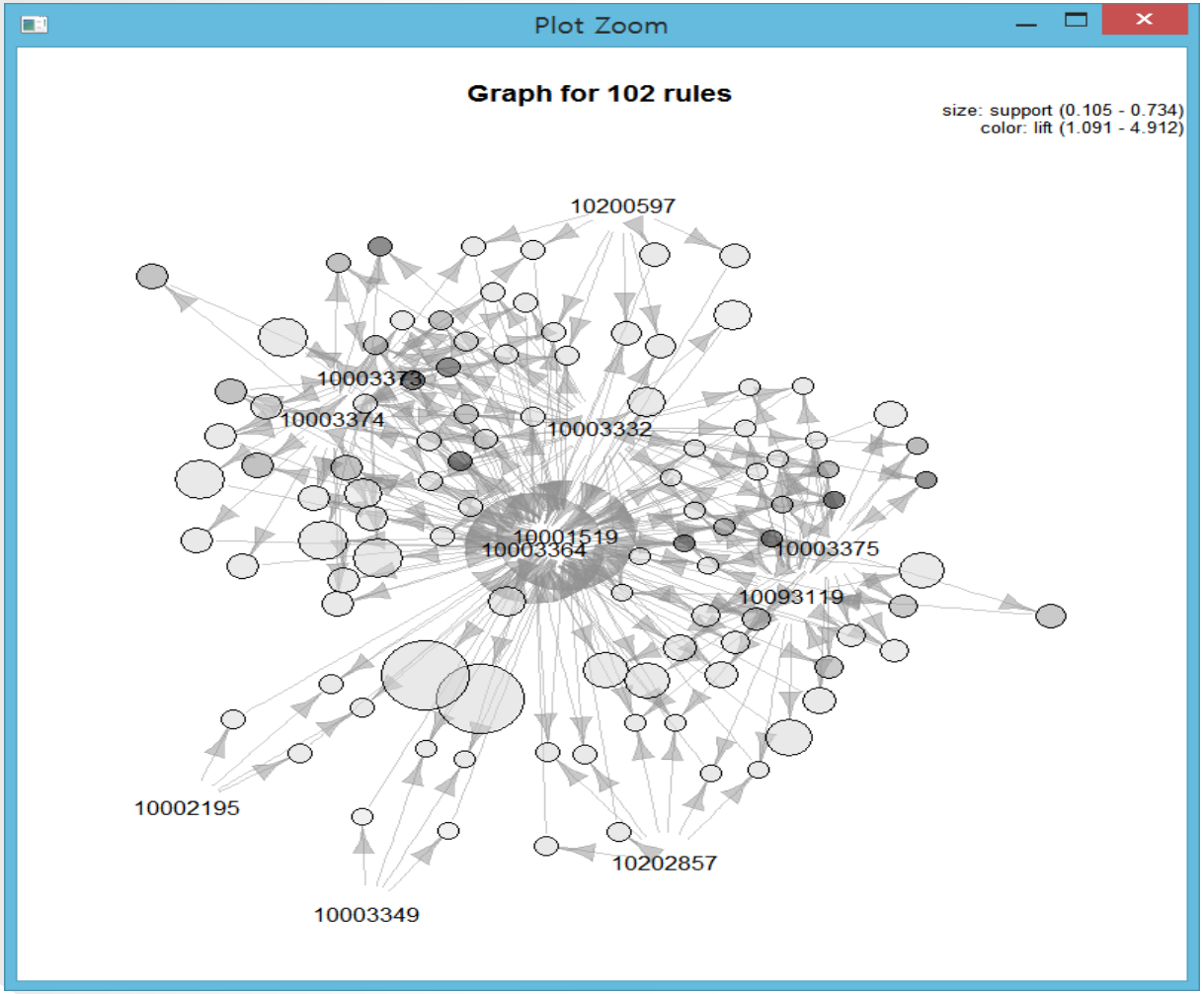
2) 연관 규칙 생성

	lhs	rhs	support	confidence	lift
1	{}	=> {1}	0.8	0.8	1.000000
2	{4}	=> {1}	0.4	1.0	1.250000
3	{2, 4}	=> {3}	0.2	1.0	1.666667
4	{3, 4}	=> {2}	0.2	1.0	1.666667
5	{2, 4}	=> {1}	0.2	1.0	1.250000
6	{3, 4}	=> {1}	0.2	1.0	1.250000
7	{2, 3, 4}	=> {1}	0.2	1.0	1.250000
8	{1, 2, 4}	=> {3}	0.2	1.0	1.666667
9	{1, 3, 4}	=> {2}	0.2	1.0	1.666667
10	{1, 2, 3}	=> {4}	0.2	1.0	2.500000

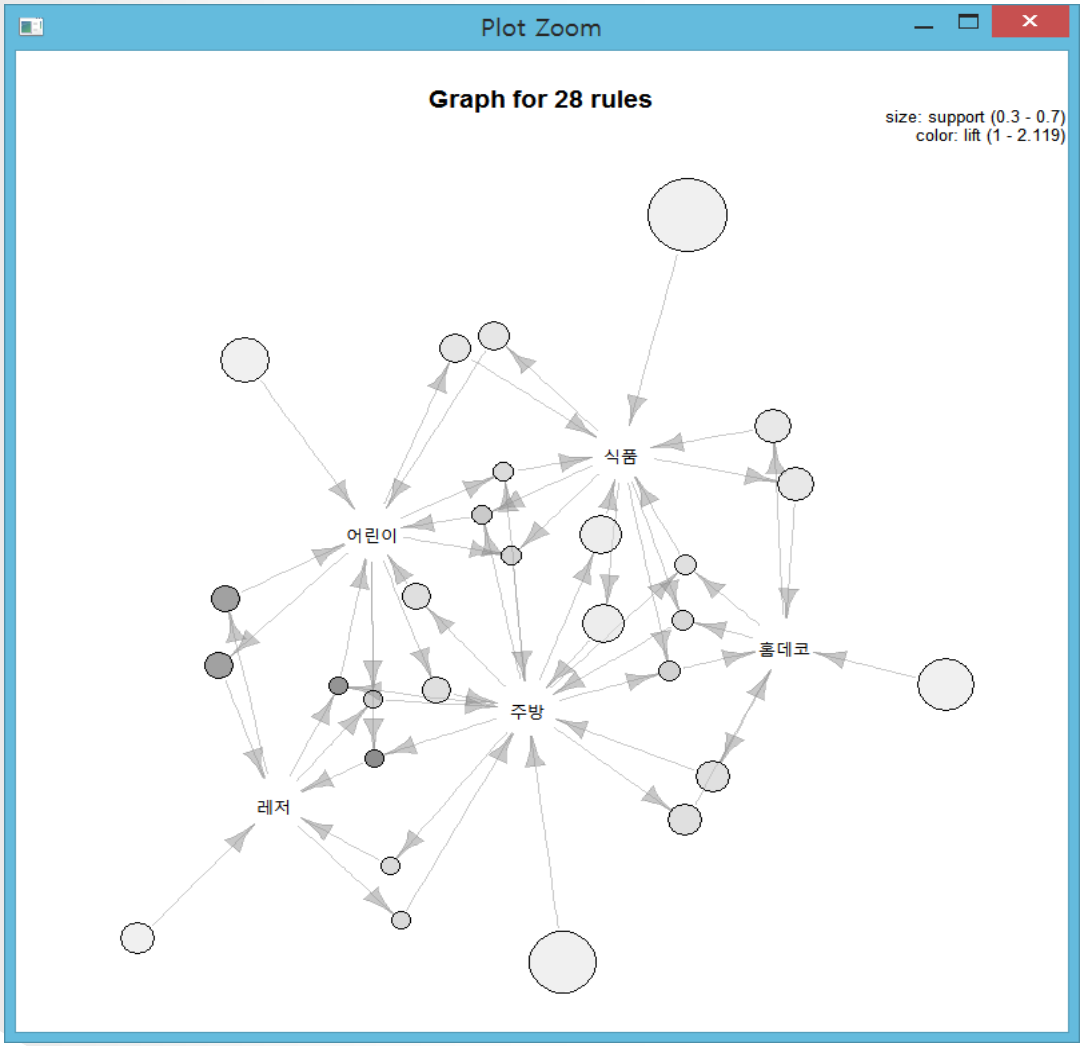
3) Adult 내장 데이터를 이용한 연관규칙 생성



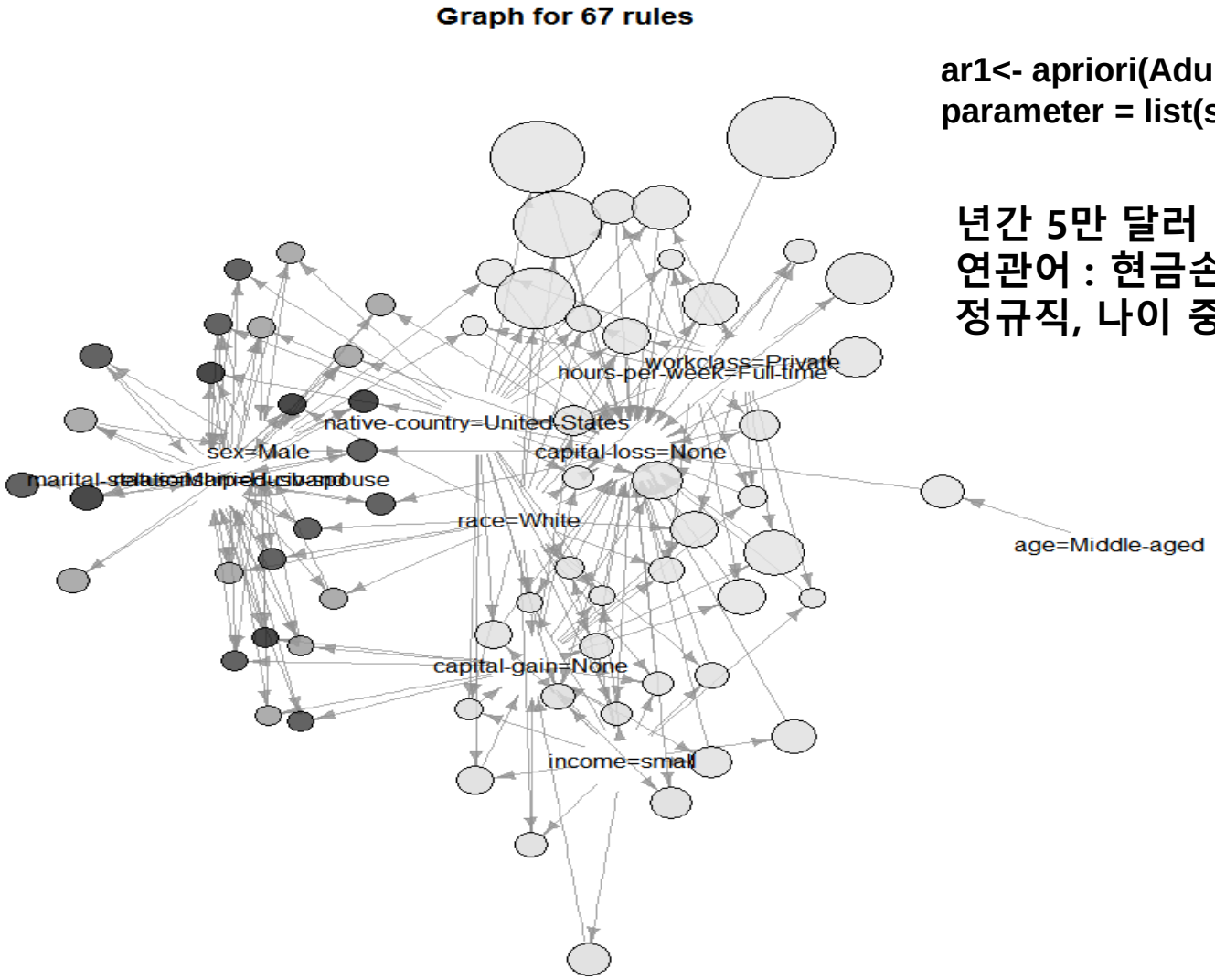
4) single format transaction 데이터 처리



5) basket format transaction 데이터 처리



연관분석



ar1<- apriori(Adult,
parameter = list(supp=0.35, conf=0.95))

년간 5만 달러 이상의 연봉 수령자 연관어 : 현금손실 무, 백인, 미국, 정규직, 나이 중년