<연습문제 해답>

[제1장 연습문제]

01. 현재 작업공간을 확인하고 "c:/Rwork/Part-I"으로 변경하시오.

```
getwd()
setwd("c:/Rwork/Part-I")
```

02. 다음 조건에 맞게 name, age, address 변수를 생성하고 처리하시오.

조건1) 각 변수의 특성에 맞게 값을 초기화하고 결과를 확인한다.

```
name <-"홍길동"
age <- 35
address <-"서울시 용산구"
name; age; address
```

조건2) 다음 함수를 이용하여 각 변수의 자료형(data type)을 확인한다.

```
mode(name); mode(age); mode(address)
is.character(name); is.numeric(age); is.character(address)
```

03. R에서 제공하는 women 데이터 셋을 다음과 같이 처리하시오.

조건1) women 데이터 셋은 어떤 데이터의 모음인가?

```
data() # Average Heights and Weights for American Women
```

조건2) women 데이터 셋의 자료 유형과 자료구조는?

```
mode(women); class(women)
```

조건3) plot() 함수를 이용하여 기본 차트 그리기

```
plot(women)
```

04. R에서 제공하는 c() 함수를 이용하여 벡터를 생성하고, 데이터를 처리하시오.

조건1) 1~100까지 벡터를 생성한다.

```
c(1:100)
```

조건2) 생성된 벡터를 대상으로 평균을 구한다.

```
num <- c(1:100)
mean(num) # 평균
```

05. R 프로그래밍 언어의 특징을 2가지만 기술하시오.

① 일반 데이터, 함수, 차트 등 모든 데이터가 객체 형태로 관리된다.
② 모든 객체는 메모리로 로딩 되어 고속으로 처리되고 재사용이 가능하다.
③ 데이터 분석에 필요한 최신의 알고리즘 및 방법론을 제공한다.
④ 데이터 분석과 표현을 위한 다양한 그래픽 도구를 제공한다.

[제2장 연습문제]

01. 다음과 같은 벡터 객체를 생성하시오.

조건1) 벡터 변수 Vec1을 만들고, "R" 문자가 5회 반복되도록 하시오.

```
Vec1 <- rep("R",5)
```

조건2) 벡터 변수 Vec2에 1~10까지 3간격으로 연속된 정수를 만드시오.

```
Vec2 <- seq(1,10, by=3)
```

조건3) 벡터 변수 Vec3에 1~10까지 3간격으로 연속된 정수가 3회 반복되도록 만드시오.

```
Vec3 <- rep(Vec2, 3)
```

조건4) 벡터 변수 Vec4에는 Vec2~Vec3가 모두 포함되는 벡터를 만드시오.

```
Vec4 <- c(Vec2, Vec3)
```

조건5) 25 ~ -15까지 5를 간격으로 seq() 함수를 이용하여 벡터를 생성하시오.

```
seq(25, -15, -5)
```

조건6) 벡터 변수 Vec4에서 홀수 번째 값들만 선택하여 Vec5에 할당하시오.(첨자 이용)

```
Vec5 <- Vec4[seq(1, 16, by=2)] # 홀수 번째만 저장
```

02. 다음과 같은 벡터를 칼럼으로 갖는 데이터프레임을 생성하시오.

조건1) 위 7개의 벡터를 칼럼으로 갖는 user 데이터프레임을 생성하시오.

```
user <- data.frame(name,age,gender,job,sat,grade,total)
```

조건2) gender 변수를 이용하여 히스토그램 그리기시오.

```
hist(user$gender)
```

조건3) user에서 짝수행만 선택해서 user2에 넣으시오.

```
user2 <- user[seq(2, 4, 2), ]
nrow(user) # 6 -> 전체 행
user2 <- user[seq(2, nrow(user), 2),] # 2~끝행 까지 짝수 행 추출
```

03. Data를 대상으로 apply()를 적용하여 행/열 방향으로 조건에 맞게 통계량을 구하시오.

```
kor <- c(90,85,90)
eng <- c(70,85,75)
mat <- c(86,92,88)
```

조건1) 3개의 과목점수를 이용하여 데이터프레임(Data)을 생성한다.

```
Data <- data.frame(kor=kor, eng=eng, mat=mat)
Data
```

조건2) 행/열 방향으로 max() 함수를 적용하여 최대값을 구하시오.

```
apply(Data, 1, max)
apply(Data, 2, max)
```

조건3) 행/열 방향으로 mean() 함수를 적용하여 평균 구하기 소숫점 2자리까지 표현하시오. 힌트 : round(data, 자릿수)

```
round(apply(Data, 1, mean), 2)
round(apply(Data, 2, mean), 2)
```

조건4) 행 단위 분산과 표준편차를 구하시오.

```
apply(Data, 1, var) # 분산
apply(Data, 1, sd) # 표준편차
```

04. 다음의 Data2 객체를 대상으로 조건에 맞게 정규표현식을 적용하여 문자열을 처리하시오.

```
Data2 <- c("2017-02-05 수입3000원", "2017-02-06 수입4500원", "2017-02-07 수입2500원")
library(stringr)
```

조건1) 날짜별 수입을 다음과 같이 출력하시오.
출력 결과) "3000원" "4500원" "2500원"

```
income <- str_extract_all(Data2, '[0-9]{4}[가-힣]')
income
unlist(income) # "3000원" "4500원" "2500원"
```

조건2) 위 벡터에서 연속하여 2개 이상 나오는 모든 숫자를 제거하시오.
출력 결과) "-- 수입원" "-- 수입원" "-- 수입원"

```
str_replace_all(Data2, '[0-9]{2}', '')
```

조건3) 위 벡터에서 -를 /로 치환하시오.
출력 결과) "2017/02/05 수입3000원" "2017/02/06 수입4500원" "2017/02/07 수입2500원"

```
str_replace_all(Data2, '-', '/')
```

조건4) 모든 원소를 쉼표(,)에 의해서 하나의 문자열로 합치시오.
출력 결과) "2017-02-05 수입3000원,2017-02-06 수입4500원,2017-02-07 수입2500원"

```
paste(Data2, collapse = ',')
```

[제3장 연습문제]

01. 본문에서 작성한 titanic 변수를 다음과 같은 단계를 통해서 "titanic.csv" 파일로 저장한 후 파일을 불러오시오.

[단계 1] "C:/Rwork/output" 폴더에 "titanic.csv"로 저장한다.

```
#힌트: write.csv() 함수 사용
write.csv(titanic, "C:/Rwork/output/titanic.csv")
```

[단계 2] "tanic.csv" 파일을 titanicData 변수로 가져와서 결과를 확인하고, titanicData의 관측치와 칼럼수를 확인한다.

```
#힌트: str() 함수 사용
titanicData = read.csv("C:/Rwork/output/titanic.csv")
str(titanicData)
```

[단계 3] 1번, 3번 칼럼을 제외한 나머지 칼럼을 대상으로 상위 6개의 관측치를 확인한다.

```
head(titanicData[, -c(1,3)])
```

02. R에서 제공하는 CO2 데이터셋을 대상으로 다음과 같이 파일로 저장하시오.

```
data(CO2)
```

[단계1] Treatment 칼럼 값이 'nonchilled'인 경우 'CO2_df1.csv' 파일로 저장

```
df1 <- subset(CO2, Treatment=='nonchilled')
write.csv(df1, "CO2_df1.csv", row.names = F)
```

[단계2] Treatment 칼럼 값이 'chilled'인 경우 'CO2_df2.csv' 파일로 저장

```
df2 <- subset(CO2, Treatment=='chilled')
write.csv(df2, "CO2_df2.csv", row.names = F)
```

[제4장 연습문제]

01. 다음 조건에 맞게 client 데이터프레임을 생성하고, 조건에 맞게 처리하시오.

<vector 준비>

```
name <-c("유관순","홍길동","이순신","신사임당")
gender <- c("F","M","M","F")
price <-c(50,65,45,75)
```

조건1) 다음 3개 벡터 객체를 이용하여 client 데이터프레임을 생성하시오.

```
client <- data.frame(name, gender, price)
```

조건2) price 변수의 값이 65만원 이상이면 문자열 "Best", 65만원 미만이면 문자열 "Normal"을 변수 result에 추가하시오. 힌트) ifelse() 함수 이용

```
client$result <-  ifelse(client$price >= 65, "Best", "Normal")
client
```

조건3) result 변수를 대상으로 빈도수를 구하시오.

```
table(client$result)
```

02. 다음 벡터 EMP는 '입사년도이름급여'순으로 사원의 정보가 기록된 데이터 있다. 이 벡터 데이터를 이용하여 다음과 같은 출력결과가 나타나도록 함수를 정의하시오.

<vector 준비>

```
EMP <- c("2014홍길동220", "2002이순신300", "2010유관순260")
```

<출력 결과>

전체 급여 평균 : 260

평균 이상 급여 수령자

이순신 => 300

유관순 => 260

힌트) 사용 함수
stringr 패키지 : str_extract(), str_replace() 함수
숫자변환 함수 : as.numeric() 함수
한글 문자 인식 정규표현식 : [가-힣]

```
# 함수 정의
emp_pay <- function(x) {
  library(stringr) # stringr 패키지 메모리 로딩
  epay <- numeric() # 급여 저장 vector
  ename <- character() # 이름 저장 vector
  idx <- 1 # index 변수

  for(n in x){ # EMP vector 원소 처리
    ename[idx] <- str_extract(n, '[가-힣]{3}') # 이름 추출/ 저장
    spay <- str_extract(n, '[가-힣]{3}[0-9]{3}') # 이름+숫자 추출
    spay <- str_replace(spay, '[가-힣]{3}', '') # 이름 제거
    npay <- as.numeric(spay) # 숫자형 변환
    epay[idx] <- npay # 급여 저장
    idx <- idx + 1 # index 카운터
  }
  avg <- mean(epay)
  cat('전체 급여 평균 :', avg, '\n')
  cat('평균 이상 급여 수령자\n')

  # 평균 급여 이상자 출력
  n <- 1: length(x)
  for(idx in n){
    if(epay[idx] >= avg){  # 평균 비교
      cat(ename[idx], '=>', epay[idx],'\n')
    }
  }
}

# 함수 호출
emp_pay(EMP)
```

03. 함수 y = f(x)에서 x의 값이 a에서 b까지 변할 때 △x = b − a를 x의 증분이라고 하며, △y = f(b) − f(a)를 y의 증분이라고 한다. 여기서 평균변화율 = △y/△x = f(b) − f(a)/ b−a 이다.

조건) 함수 f(x) = x^3 + 4에서 x의 값이 1에서 3까지 변할 때 평균변화율(mean ratio of change)을 구하는 함수를 작성하시오. ∴ 평균변화율 = 31-5/2 = 13

```
mrc <- function(x) {
    x <- x^3 + 4
}
a <- 1
b <- 3
mrc_result <- { mrc(b) - mrc(a) } / {b-a}
mrc_result
```

04. RSADBE 패키지에서 제공되는 Bug_Metrics_Software 데이터 셋을 대상으로 소프트웨어 발표 후 행 단위 합계와 열 단위 평균을 구하고, 칼럼 단위로 요약통계량을 구하시오.

```
library('RSADBE')
data('Bug_Metrics_Software')
rowSums(Bug_Metrics_Software[,,2])
colMeans(Bug_Metrics_Software[,,2])
summary(Bug_Metrics_Software[,,2])
```

[제5장 연습문제]

01. iris3 데이터 셋을 대상으로 다음 조건에 맞게 산점도를 그리시오.

조건1) iris3 데이터 셋의 칼럼명을 확인합니다.

```
attributes(iris3)
```

조건2) iris3 데이터 셋의 구조를 확인합니다.

```
str(iris3)
```

조건3) 꽃의 종별로 다음과 같이 4개의 그래프 이미지를 참고하여 산점도 그래프를 그립니다.

```
plot(iris3, main="iris3 전체 데이터 셋의 분포현황")
plot(iris3[,c(3,1),1], main="iris3 데이터 셋의 Setosa 분포현황")
plot(iris3[,c(4,2),2], main="iris3 데이터 셋의 Versicolor 분포현황")
plot(iris3[,c(2,3),3], main="iris3 데이터 셋의 Virginica 분포현황")
```

02. iris 데이터 테이블을 대상으로 다음 조건에 맞게 시각화하시오.

조건1) 1번 칼럼(x축), 3번 칼럼(y축)

```
plot(iris[,1], iris[,3])
plot(iris$Sepal.Length, iris$Petal.Length, col="red") # 모두 빨강색
```

조건2) 5번 칼럼으로 색상 지정

```
plot(iris[,1], iris[,3], col=iris[,5])  # 5칼럼으로 색상 구분
names(iris)
plot(iris$Sepal.Length, iris$Petal.Length, col=iris$Species)
```

조건3) "iris 데이터 테이블 산포도 차트" 제목 추가

```
plot(iris$Sepal.Length, iris$Petal.Length, col=iris$Species)
title(main="iris 데이터 테이블 산포도 차트")
```

조건4) 다음 조건에 맞게 작성한 차트를 파일에 저장하기

```
setwd("C:/Rwork/output") # 작업 디렉터리 지정
jpeg("iris.jpg", width=720, height=480) # 픽셀 지정 가능
plot(iris$Sepal.Length, iris$Petal.Length, col=iris$Species)
title(main="iris 데이터 테이블 산포도 차트")
dev.off() # 장치 종료
```

[제6장 연습문제]

```
# dplyr 패키지 관련 문제
library(dplyr)
data(iris)
```

1. iris의 꽃잎의 길이(Petal.Length) 칼럼을 대상으로 1.5 이상의 값만 필터링하시오.

```
#힌트) %>% 기호와 filter() 함수 이용
q1 <- iris %>% filter(Petal.Length >= 5.0)
```

2. 문제1번 결과에서 1,3,5번 칼럼을 선택하시오.

```
#힌트) %>% 기호와 select() 함수 이용
q2 <- q1 %>% filter(Petal.Length >= 5.0) %>% select(Sepal.Length, Petal.Length, Species)
```

3. 문제2번 결과에서 1번 칼럼에서 3번 칼럼의 값을 뺀 diff 파생변수를 만들고, 앞부분 6개만 출력하시오.(diff = 1번 칼럼 - 3번 칼럼)

```
#힌트) %>% 기호와 mutate() 함수 이용
q3 <- q2 %>% mutate(diff = Sepal.Length - Petal.Length) %>% head()
```

4. 문제3번 결과에서 꽃의 종(Species)별로 그룹화하여 Sepal.Length와 Petal.Length 변수의
평균을 계산하시오.

```
#힌트) %>% 기호와 group_by() 함수, summarise() 함수 이용
q4 <- q3 %>% group_by(Species) %>% summarise(sepal_mean=mean(Sepal.Length),
                                             petal_mean=mean(Petal.Length))
```

```
# reshape2 패키지 관련 문제
library('reshape2')
```

5. reshape2 패키지를 이용하여 단계별로 iris 데이터 셋을 처리하시오.

[단계1] 꽃의 종류(Species)를 기준으로 '넓은 형식'을 '긴 형식'으로 변경하기
#힌트) melt() 함수 이용
```
melt <- melt(iris, id=c("Species"), na.rm=TRUE)
head(melt)
```

[단계2] 꽃의 종별로 나머지 4가지 변수의 합계 구하기
#힌트) dcast() 함수 이용
```
dcast <- dcast(melt, Species ~ variable, sum)
```

[제7장 연습문제]

01. 본문에서 생성된 dataset2의 직급(position) 칼럼을 대상으로 1급 -> 5급, 5급 -> 1급 형식으로 역코딩하여 position2 칼럼에 추가하시오.
```
pos <- dataset2$position
cpos <- 6 - pos
dataset2$position <- cpos
dataset2$position2[dataset2$position==1] <-'1급'
dataset2$position2[dataset2$position==2] <-'2급'
dataset2$position2[dataset2$position==3] <-'3급'
dataset2$position2[dataset2$position==4] <-'4급'
dataset2$position2[dataset2$position==5] <-'5급'
```

02. dataset2의 resident 칼럼을 대상으로 NA 값을 제거한 후 dataset2 변수에 저장하시오.
```
range(dataset2$resident, na.rm = T) # 1 5
dataset2 <- subset(dataset2, !is.na(dataset2$resident))
head(dataset2)
dim(dataset2)
```

03. dataset2의 gender 칼럼을 대상으로 1->"남자", 2->"여자" 형태로 코딩 변경하여 gender2 칼럼에 추가하고, 파이 차트로 결과를 확인하시오.
```
dataset2$gender2[dataset2$gender == 1] <-'남자'
dataset2$gender2[dataset2$gender == 2] <-'여자'
pie(table(dataset2$gender2))
```

04. 나이를 30세 이하 -> 1, 30~55 -> 2, 56이상 -> 3 으로 리코딩하여 age3 칼럼에 추가한 후 age, age2, age3 칼럼만 확인하시오.
```
dataset2$age3[dataset2$age <= 30] <-1
dataset2$age3[dataset2$age > 30 & dataset2$age <= 55] <-2
dataset2$age3[dataset2$age > 55] <-3
head(dataset2[c('age', 'age2', 'age3')])
```

05. 정제된 data를 대상으로 작업 디렉터리(c:/Rwork/Part-II)에 cleanData.csv 파일명으로

따옴표와 행 이름을 제거하여 저장하고, new_data변수로 읽어오시오.

```
setwd("C:/Rwork/Part-II")
# (1) 정제된 데이터 저장
write.csv(dataset2,"cleanData.csv ", quote=F, row.names=F)
# (2) 저장된 파일 불러오기/확인
new_data <- read.csv("cleanData.csv", header=TRUE)
new_data
dim(new_data)
str(new_data)
```

06. user_data.csv와 return_data.csv 파일을 이용하여 각 고객별 반품사유코드(return_code)를 대상으로 다음과 같이 파생변수를 추가하시오.

```
단계1: 고객 정보 파일 가져오기
u_data <- read.csv('user_data.csv', header = T)
head(u_data) # user_id age house_type resident job child

단계2: 반품 정보 파일 가져오기
r_data <- read.csv('return_data.csv', header = T)
head(r_data) # user_id return_code

단계3: 고객별 반품사유코드에 따른 파생변수 생성
library(reshape2)
u_return <- dcast(r_data, user_id ~ return_code, length) # 행 ~ 열
head(u_return, 3) # 행(고객 id) 열(반품사유코드)
names(u_return)<-c('user_id','return_code1','return_code2','return_code3',
  'return_code4')
head(u_return, 3) # 칼럼명 확인

단계4: 파생변수 추가 : 고객정보에 반품사유 칼럼 추가
library(plyr) # 패키지 로딩
user_return_data <- join(u_data, u_return, by='user_id')
head(user_return_data,10) # NA : 해당 고객 반품이력이 없는 경우
```

07. iris 데이터를 이용하여 5겹 2회 반복하는 교차검정 데이터를 샘플링하시오.

```
단계1: K겹 교차검정 데이터 생성
cross <- cvFolds(nrow(iris), K=5, R=2)

단계2: K겹 교차검정 데이터 보기
str(cross) # 구조 보기
table(cross$which) # 5겹 빈도수
1  2  3  4  5
```

```
30 30 30 30 30
cross # 5겹 교차검정 데이터 보기

단계3: 샘플링 관측치 행번호 추출
# K=1, R=1인 경우 샘플링 관측치 행번호 추출
datas_idx <- cross$subsets[cross$which==1, 1]
datas_idx # test set
length(datas_idx) # 30
train <- iris[-datas_idx, ]
test <- iris[datas_idx, ]

# K=2, R=1인 경우 샘플링 관측치 행번호 추출
datas_idx <- cross$subsets[cross$which==2, 1]
datas_idx # test set
length(datas_idx) # 30
train <- iris[-datas_idx, ]
test <- iris[datas_idx, ]

# K=5, R=2인 경우 샘플링 관측치 행번호 추출
cross$subsets[cross$which==5, 2]
length(cross$subsets[cross$which==5, 2]) # 30

단계4: iris 데이터프레임 적용 데이터 셋 생성
R=1:2
K=1:5

for(r in R){ # 2회
  cat('R=',r, '\n')
  for(k in K){ # 5회
    datas_idx <- cross$subsets[cross$which==k, r]
    cat('K=',k,'검정데이터 \n')
    print(iris[datas_idx, ])  # 검정데이터 생성

    cat('K=',k,'훈련데이터 \n') # 학습데이터 생성
    print(iris[-datas_idx, ])
  } # outer for K
} # outer for R
```

[제8장 연습문제]

01. 다음 조건에 맞게 quakes 데이터 셋의 수심(depth)과 리히터규모(mag)가 동일한 패널에 지진의 발생지를 산점도로 시각화하시오.

조건1) 수심 3개 영역으로 범주화

```
depthgroup<-equal.count(quakes$depth, number=3, overlap=0)
```

조건2) 리히터규모 2개 영역으로 범주화

```
magnitudegroup<-equal.count(quakes$mag, number=2, overlap=0)
```

조건3) 수심과 리히터규모가 3행 2열 구조의 패널로 산점도 그래프 그리기

```
xyplot(lat ~ long | magnitudegroup*depthgroup, data=quakes,
main="Fiji Earthquakes", ylab="latitude", xlab="longitude",
pch="@",col=c("red","blue"))
```

02. latticeExtra패키지에서 제공되는 SeatacWeather 데이터 셋에서 월 별로 최저기온과 최고기온을 선 그래프로 플로팅 하시오.

힌트) lattice 패키지의 xyplot() 함수 이용

힌트) 선 그래프 : type="l"

```
xyplot(min.temp + max.temp ~ day | month,
data=SeatacWeather, type="l", layout=c(3,1)) # type=line
```

03. diamonds 데이터 셋을 대상으로 x축에 carat변수, y축에 price변수를 지정하고, clarity 변수를 선 색으로 지정하여 미적 요소 맵핑 객체를 생성한 후 산점도 그래프 주변에 부드러운 곡선이 추가되도록 레이아웃을 추가하시오.

```
p<- ggplot(data=diamonds, aes(x=carat, y=price, colour=clarity))
p + geom_point() + geom_smooth()
```

04. 서울지역에 있는 주요 대학교의 위치 정보를 이용하여 레이어 기법으로 다음과 같이 시각화 하시오.

조건1) 지도 중심 지역 SEOUL, zoom=11, maptype="watercolor"

조건2) 데이터 셋("C:/Rwork/Part-II/university.csv")

조건3) 지도 좌표 : 위도(LAT), 경도(LON)

조건3) 각 학교명으로 포인터의 크기와 텍스트 표시

조건5) 파일명을 "university.png"로하여 이미지 파일로 결과 저장
이미지의 가로/세로 픽셀 크기(width=10.24,height=7.68)

```
library(ggmap)
setwd("d:/Rwork/Part-II") # 파일 경로 지정

# 서울지역 4년제 대학교 위치 정보 자료 가져오기
university <- read.csv("university.csv")
university # # 학교명","LAT","LON"

# 지도정보 생성
seoul <- c(left = 126.77, bottom = 37.40,
```

```
            right = 127.17, top = 37.70)

map <- get_stamenmap(seoul, zoom=11,  maptype='watercolor')

# (1)레이어1 : 정적 지도 생성
layer1 <- ggmap(map)
layer1

# (2)레이어2 : 지도위에 포인트
layer2 <- layer1 + geom_point(data=university,
                              aes(x=LON,y=LAT, color=학교명), size=3)
layer2

# (3)레이어3 : 지도위에 텍스트 추가
layer3 <- layer2 + geom_text(data=university,
                             aes(x=LON+0.01, y=LAT+0.01,label=학교명), size=5)
layer3

# (4)지도 저장 : 넓이, 폭 적용 파일 저장
ggsave("university.png",width=10.24,height=7.68)
```

[제9장 연습문제]

01. 다음과 같은 단계를 통해서 테이블을 생성하고, SQL문을 이용하여 레코드를 조회하시오.

단계1: 상품정보(GoodsInfo)테이블 생성(SQLPLUS 이용)

```
create table GoodsInfo(
  proCode char(5) primary key,
  proName varchar2(30) not null,
  price number(6) not null,
  maker varchar2(25) not null,
);
```

단계2: 레코드 추가

```
insert into GoodsInfo values(1001, '냉장고', 1800000, 'SM');
insert into GoodsInfo values(1002, '세탁기', 500000, 'LN');
insert into GoodsInfo values(1003, 'HDTV', 2500000, 'HP');
```

단계3: 전체 레코드 검색(R 코드 이용)

```
query = "SELECT * FROM GoodsInfo"
dbGetQuery(conn, query)
```

[제10장 연습문제]

01	④	02	①	03	①
04	②	05	④	06	④
07	②	08	②	09	①
10	▪ 예비조사와 사전조사의 목적 ○ 예비조사 : 조사설계를 확정하기 전에 연구문제에 대한 사전지식이 부족할 경우 질문지 및 면접조사를 통해서 연구문제 확인, 변수 및 변수 간 상관관계 등을 파악하여 가설을 정립하기 위해서 실시하는 조사방법이다.(예: 문헌조사, 경험자조사(파일럿조사),사례조사 등) ○ 사전조사 : 조사설계를 마친 후 조사계획의 적절성을 알아보기 위해서 본 조사에 들어가기 전에 실시하는 소규모 조사방법이다.(표본추출방법이나 표본의 대표성은 크게 고려하지 않음)				
11	①	12	④	13	②
14	④	15	①	16	중심극한정리
17	①	18	④ 신뢰수준이 높이면 신뢰구간은 좁아진다.	19	①
20	③	21	②	22	[2948.4, 3051.6]

[제11장 연습문제]

01. MASS 패키지에 있는 Animals 데이터 셋을 이용하여 각 단계에 맞게 기술 통계량을 구하시오.

단계1: MASS 패키지 설치 및 메모리 로딩

```
library(MASS) # MASS 패키지 불러오기
data(Animals) # Animals 데이터 셋 로딩
head(Animals) # Animals 데이터 셋 보기
```

단계2: R의 기본 함수를 이용하여 brain 칼럼을 대상으로 다음 기술 통계량 구하기

```
summary(Animals$brain) # 요약통계량
mean(Animals$brain) # 평균
median(Animals$brain) # 중위수
sd(Animals$brain) # 표준편차
var(Animals$brain) # 분산
max(Animals$brain) # 최댓값
min(Animals$brain) # 최솟값
```

단계3: 패키지에서 제공되는 describe()과 freq() 함수를 이용하여 Animals 데이터 셋 전체를 대상으로 기술 통계량 구하기

```
describe(Animals)
freq(Animals)
```

02. descriptive.csv 데이터 셋을 대상으로 다음 조건에 맞게 빈도분석 및 기술 통계량 분석을 수행하시오

조건1) 명목척도 변수인 학교유형(type), 합격여부(pass) 변수에 대해 빈도분석을 수행하고 결과를 막대그래프와 파이차트로 시각화

```
type <- data$type
table(type)
barplot(table(type))
pie(table(type))
```

조건2) 비율척도 변수인 나이 변수에 대해 요약치(평균, 표준편차)와 비대칭도(왜도와 첨도)통계량을 구하고, 히스토그램 작성하여 비대칭도 통계량 설명

```
age <- data$age
range(age) #  40 69
mean(age) # 53.88
sd(age) # 6.813247
skewness(age) # 0.3804892(왜도)
kurtosis(age) # 1.866623(첨도)
hist(age)
```

<설명> 왜도가 0보다 조금 크기 때문에 오른쪽 방향으로 비대칭 꼬리가 치우치고, 첨도는 3보다 작기 때문에 표준정규분포보다 완만한 형태를 갖는다.

조건3) 나이 변수에 대한 밀도분포곡선과 정규분포 곡선으로 정규분포 검정

```
hist(age, freq = F)
lines(density(age), col='blue')
x <- seq(35, 80, 0.1)
curve(dnorm(x, mean(age), sd(age)), col='red', add = T)
```

[제12장 연습문제]

01. 교육수준(education)과 흡연율(smoking) 간의 관련성을 분석하기 위한 연구가설을 수립하고, 각 단계별로 가설을 검정하시오. [독립성 검정]

귀무가설(H_0) : **교육수준과 흡연율 간의 관련성은 없다.**

연구가설(H_1) : **교육수준과 흡연율 간의 관련성은 있다.**

단계1: 파일 가져오기

```
setwd("c:/Rwork/Part-III")
smoke <- read.csv("smoke.csv", header=TRUE)
head(smoke)
```

단계2: 코딩 변경(변수 리코딩)

education(독립변수) : 1:대졸, 2:고졸, 3:중졸

smoke(종속변수): 1:과다흡연, 2:보통흡연, 3:비흡연

```
smoke$education2[smoke$education==1] <- "1.대졸"
smoke$education2[smoke$education==2] <- "2.고졸"
smoke$education2[smoke$education==3] <- "3.중졸"
smoke$smoking2[smoke$smoking==1] <- "1.과대흡연"
smoke$smoking2[smoke$smoking==2] <- "2.보통흡연"
smoke$smoking2[smoke$smoking==3] <- "3.비흡연"
```

단계3: 교차 분할표 작성

```
table(smoke$education2, smoke$smoking2)
```

단계4: 독립성 검정

```
library(gmodels) # CrossTable() 함수 사용
CrossTable(smoke$education2, smoke$smoking2, chisq = TRUE)
```

단계5: 검정결과 해석

<해설> 검정 결과 p-value가 0.003110976 이므로 유의미한 수준에서 '**교육수준과 흡연율 간의 관련성은 있다.**' 라고 볼 수 있다.

02. 나이(age3)와 직위(position) 간의 관련성을 단계별로 분석하시오. [독립성 검정]

단계1: 파일 가져오기

```
setwd("c:/Rwork/Part-III")
data <- read.csv("cleanData.csv", header=TRUE)
head(data)
```

단계2: 코딩 변경(변수 리코딩)

```
x <- data$position # 행 - 직위 변수 이용
y <- data$age3 # 열 - 나이 리코딩 변수 이용
```

단계3: 산점도를 이용한 변수간의 관련성 보기 - plot(x,y) 함수 이용

```
plot(x, y) #나이와 직위에 대한 산점도
 # 경향선과 제목 추가 -> 직위가 높을 수록 나이가 많음(1급-장년층)
plot(x,y,abline(lm(y~x)),main="나이와 직위에 대한 산점도")
```

단계4: 독립성 검정

```
CrossTable(x,y, chisq = TRUE)
```

단계5: 검정결과 해석

<해설> '나이와 직위는 관련성이 있다.'를 분석하기 위해서 A회사 223명을 표본으로 추출한 후 설문조사하여 교차분석과 카이제곱검정을 실시하였다. 분석결과를 살펴보면 나이와 직위의 관련성은 유의미한 수준에서 차이가 있는 것으로 나타났다.(X^2=309.369, p<0.05) 따라서 연구가설을 채택한다.

03. 직업유형에 따른 응답정도에 차이가 있는가를 단계별로 검정하시오.[동질성 검정]

단계1: 파일 가져오기

```
setwd("c:/Rwork/Part-III")
response <- read.csv("response.csv", header=TRUE)
```

단계2: 코딩 변경 - 리코딩

job : 1:학생, 2:직장인, 3:주부

response : 1:무응답, 2:낮음, 3:높음

```
# job2 칼럼 추가
result$job2[result$job==1] <- "1.학생"
result$job2[result$job==2] <- "2.직장인"
result$job2[result$job==3] <- "3.주부"
# response2 칼럼 추가
result$response2[result$response==1] <- "1.무응답"
result$response2[result$response==2] <- "2.낮음"
result$response2[result$response==3] <- "3.높음"
```

단계3: 교차 분할표 작성

```
table(result$job2, result$response2)
```

단계4: 동일성 검정

```
chisq.test(result$job3, result$response3)  #p-value = 6.901e-12
# 귀무가설 : 세 집단의 비율은 동일하다.
```

단계5: 검정결과 해석

<해설> 귀무가설 기각 : 세 집단 간의 응답율이 서로 다르다고 할 수 있다.

[제13장 연습문제]

01. 중소기업에서 생산한 HDTV 판매율을 높이기 위해서 프로모션을 진행한 결과 기존 구매비율 보다 15% 향상되었는지를 각 단계별로 분석을 수행하여 검정하시오.

연구가설(H_1) : 기존 구매비율과 차이가 있다.

귀무가설(H_0) : 기존 구매비율과 차이가 없다.

조건) 구매여부 변수 : buy (1: 구매하지 않음, 2: 구매)

단계1: 데이터셋 가져오기

```
setwd("c:/Rwork/Part-III")
hdtv <- read.csv("hdtv.csv", header=TRUE)
```

단계2: 빈도수와 비율 계산

```
summary(hdtv)
length(hdtv$buy) # 50개
install.packages('prettyR')
library(prettyR) # freq() 함수 사용
freq(hdtv$buy) # 1:40, 2:10
table(hdtv$buy)
table(hdtv$buy, useNA="ifany") # NA 빈도수 표시

단계3: 가설검정
binom.test(c(10,40), p=0.15) #15% 비교 ->p-value = 0.321
binom.test(c(10,40), p=0.15, alternative="two.sided", conf.level=0.95)
<해설> 귀무가설 채택 : 기존 구매비율(15%)과 차이가 없다.

[실습] 방향성이 있는 단측가설 검정
binom.test(c(10,40), p=0.15, alternative="greater", conf.level=0.95)
#p-value=0.2089
binom.test(c(10,40), p=0.15, alternative="less", conf.level=0.95) #p-value =
0.8801
<해설> 방향성이 잇는 단측가설은 모두 기각된다.

[실습] 11% 기준 : 방향성이 있는 연구가설 검정
binom.test(c(10,40), p=0.11, alternative="greater", conf.level=0.95)
#p-value=0.04345
<해설> 구매비율은 11%을 넘지 못한다.
```

02. 우리나라 전체 중학교 2학년 여학생 평균 키가 148.5cm로 알려져 있는 상태에서 A중학교 2학년 전체 500명을 대상으로 10%인 50명을 표본으로 선정하여 표본평균신장을 계산하고 모집단의 평균과 차이가 있는지를 각 단계별로 분석을 수행하여 검정하시오.

```
단계1: 데이터셋 가져오기
setwd("c:/Rwork/Part-III")
stheight<- read.csv("student_height.csv", header=TRUE)
stheight
height <- stheight$height
head(height)

단계2: 기술 통계량/결측치 확인
length(height) #50
summary(height) # 149.4
x1 <- na.omit(height)
x1 # 정제 데이터
mean(x1) # 149.4 : 평균신장
```

단계3: 정규성 검정
shapiro.test(x1) # p-value = 0.0001853 -> 정규분포 아님
정규분포(모수검정) - t.test()
비정규분포(비모수검정) - wilcox.test()

단계4: 가설검정 - 양측검정
wilcox.test(x1, mu=148.5) # p-value = 0.067
wilcox.test(x1, mu=148.5, alter="two.side", conf.level=0.95) # p-value = 0.067
<해설> 귀무가설을 기각할 수 없다.

03. 대학에 진학한 남학생과 여학생을 대상으로 진학한 대학에 대해서 만족도에 차이가 있는가를 검정하시오.
힌트) 두 집단 비율 차이 검정
조건1) 파일명 : two_sample.csv
조건2, 변수명 : gender(1,2), survey(0,1)

단계1: 실습데이터 가져오기
getwd()
setwd("c:/Rwork/Part-III")
data <- read.csv("two_sample.csv", header=TRUE)
data
head(data) # 변수명 확인

단계2: 두 집단 subset 작성
data$gender
data$survey # 1(만족), 0(불만족)
데이터 정체/전처리
x<- data$gender # 성별 추출
y<- data$survey # 만족도 추출

교차테이블 확인
table(x) # 성별 구분 (1 : 174, 2 : 126)
table(y) # # 대학진학 만족도(0 : 55, 1 : 245)
table(x, y, useNA="ifany") # 결측치 까지 출력

단계3: 두 집단 비율차이검증 : prop.test()
help(prop.test) # prop.test(x,n,p, alternative, conf.level, correct)
prop.test(c(138,107),c(174,126)) # 남학생과 여학생의 만족도 차이 검정
prop.test(c(138,107),c(174,126), alternative="two.sided", conf.level=0.95)
<해설> p-value = 0.2765 : 남학생과 여학생의 만족도에 차이가 없다.

04. 교육방법에 따라 시험성적에 차이가 있는지 검정하시오.
힌트) 두 집단 평균 차이 검정
조건1) 파일 : twomethod.csv
조건2) 변수 : method(교육방법), score(시험성적)
조건3) 모델 : 교육방법(명목) -> 시험성적(비율)
조건4) 전처리 : 결측치 제거

단계1: 실습파일 가져오기
```
setwd("C:/Rwork/Part-III")
edumethod <- read.csv("twomethod.csv", header=TRUE)
head(edumethod) #3개 변수 확인 -> id method score
```

단계2: 두 집단 subset 작성(데이터 정제, 전처리)
```
# 데이터 전처리(score의 NA 처리)
result <- subset(edumethod, !is.na(score), c(method, score))
result
```

단계3: 데이터 분리
```
# 1) 교육방법별로 분리
a <- subset(result,method==1)
b <- subset(result,method==2)

# 2) 교육방법에서 영업실적 추출
a1 <- a$score
b1 <- b$score

# 3) 기술 통계량
length(a1); # 22
length(b1); # 35
```

단계4: 분포모양 검정
```
var.test(a1, b1) # p-value = 0.8494 : 차이가 없다.
```
<해설> 동질성 분포와 차이가 없다. 모수검정 방법 수행

단계5: 가설검정
```
t.test(a1, b1) # p-value = 1.303e-06
t.test(a1, b1, alter="greater", conf.int=TRUE, conf.level=0.95) # p-value = 1
```
<해설> a1 교육방법과 b1 교육방법 간의 시험성적에 차이가 있다.
```
t.test(b1, a1, alter="greater", conf.int=TRUE, conf.level=0.95) #
p-value=6.513e-07
```

<해설> b1 교육 방법이 a1 교육방법 보다 시험성적이 더 좋다.

[제14장 연습문제]

01. 다음은 drinkig_water_example.sav 파일의 데이터 셋이 구성된 테이블이다. 전체 2개의 요인에 의해서 7개의 변수로 구성되어 있다. 아래에서 제시된 각 단계에 맞게 요인분석을 수행하시오.

단계 1 : 데이터 파일 가져오기

```
library(memisc)
setwd("C:\\Rwork\\Part-III")
data.spss <- as.data.set(spss.system.file('drinking_water_example.sav'))
data.spss
drinkig_water_exam <- data.spss[1:7]
drinkig_water_exam_df <- as.data.frame(drinkig_water_exam)
str(drinkig_water_exam_df)
```

단계 2 : 베리멕스 회전법, 요인수 2, 요인점수 회귀분석 방법을 적용하여 요인분석

```
result <- factanal(drinkig_water_exam_df, factors = 2,
                   rotation = "varimax",
                   scores = "regression")
```

단계 3 : 요인적재량 행렬의 칼럼명 변경

```
loadings <- result$loadings
colnames(loadings) <- c("제품친밀도","제품만족도")
loadings
```

단계 4 : 요인점수를 이용한 요인적재량 시각화

```
plot(result$scores[,c(1,2)], main="Factor1과 Factor2 요인점수 행렬")

# 관측치별 이름 매핑(rownames mapping)
text(result$scores[,1], result$scores[,2],
     labels = rownames(result$scores),
     cex = 0.7, pos = 3, col = "blue")

# 요인적재량 plotting
points(result$loadings[,c(1,2)], pch=19, col = "red")
# Factor1, Factor3 요인적재량 표시

text(result$loadings[,1], result$loadings[,2],
     labels = rownames(result$loadings),
     cex = 0.8, pos = 3, col = "red")
```

단계 5 : 요인별 변수 묶기
# 제품만족도 데이터프레임 - q1,q2,q3
s <- data.frame(drinkig_water_exam_df$q1, drinkig_water_exam_df$q2, drinkig_water_exam_df$q3)
# 제품친밀도 데이터프레임 - q4,q5,q6,q7
c <- data.frame(drinkig_water_exam_df$q4, drinkig_water_exam_df$q5, drinkig_water_exam_df$q6, drinkig_water_exam_df$q7)

# 요인별 산술평균 계산
satisfaction <- round((s$drinkig_water_exam_df.q1+s$drinkig_water_exam_df.q2 + s$drinkig_water_exam_df.q3)/ncol(s),2)
closeness <- round((c$drinkig_water_exam_df.q4 + c$drinkig_water_exam_df.q5 + c$drinkig_water_exam_df.q6 + c$drinkig_water_exam_df.q7) / ncol(c), 2)

02. 문제 01에서 생성된 두 개의 요인을 데이터프레임으로 생성한 후 이를 이용하여 두 요인 간의 상관관계 계수를 제시하시오.
new_drinking_water <- data.frame(satisfaction, closeness)
cor(new_drinking_water)

[제15장 연습문제]

01. product.csv 파일의 데이터를 이용하여 다음과 같은 단계로 다중회귀분석을 수행하시오.
product <- read.csv("C:/Rwork/Part-IV/product.csv", header=TRUE)

단계1 : 학습데이터(train), 검정데이터(test)를 7 : 3 비율로 샘플링
idx <- sample(1:nrow(product), 0.7*nrow(product))
train <- product[idx,] # result중 70%
dim(train) # [1] 184 3
train # 학습데이터
test <- product[-idx,] # result중 나머지 30%
dim(test) # [1] 80 3
test # 검정 데이터

단계2 : 학습데이터 이용 회귀모델 생성
변수 모델링) y변수 : 제품_만족도, x변수 : 제품_적절성, 제품_친밀도
model <- lm(formula=제품_만족도 ~ 제품_적절성 + 제품_친밀도, data=train)
summary(model) # 학습데이터 분석 -> p-value: < 2.2e-16

단계3 : 검정데이터 이용 모델 예측치 생성
pred <- predict(model, test) # 예측치 생성

단계4 : 모델 평가 : cor() 함수 이용

```
cor(pred, test$제품_만족도) # 모델 평가
```

02. ggplot2패키지에서 제공하는 diamonds 데이터 셋을 대상으로 carat, table, depth 변수 중 다이아몬드의 가격(price)에 영향을 미치는 관계를 다중회귀 분석을 이용하여 예측하시오.

조건1) 다이아몬드 가격 결정에 가장 큰 영향을 미치는 변수는?

조건2) 다중회귀 분석 결과를 정(+)과 부(-) 관계로 해설

```
library(ggplot2)
data(diamonds)
# diamonds에서 비율척도 대상으로 식 작성
formula <- price ~ carat +  table + depth
head(diamonds)
result <- lm(formula, data=diamonds)
summary(result) # 회귀분석 결과
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)   13003.441    390.918   33.26   <2e-16 ***
  carat        7858.771     14.151  555.36   <2e-16 ***
  table        -104.473      3.141  -33.26   <2e-16 ***
  depth         -151.236      4.820  -31.38   <2e-16 ***
```

<해설>carat은 price에 정(+)의 영향을 미치지만, table과 depth는 부(-)의 영향을 미친다.

03. mpg 데이터 셋을 대상으로 7:3 비율로 학습데이터와 검정데이터로 각각 샘플링한 후 각 단계별로 분류분석을 수행하시오.

조건) 변수모델링 : x변수(displ + cyl + year), y변수(cty)

```
library(ggplot2)
data(mpg)
str(mpg)
```

단계1 : 학습데이터와 검증데이터 샘플링

```
idx <- sample(1: nrow(mpg), nrow(mpg) * 0.7)
train <- mpg[idx, ] # 학습데이터
dim(train)
test <- mpg[-idx, ]  # 검정데이터
dim(test)
```

단계2 : formula 생성

```
# 도시 주행마일수 <- 실린더, 엔진크기, 제조년도
formula <- cty ~ displ + cyl + year
```

단계3 : 학습데이터에 분류모델 적용

```
mpg_train <- ctree(formula, data=train)
```

단계4 : 검정데이터에 분류모델 적용
```
mpg_test <- ctree(formula, data=test)
```

단계5 : 분류분석 결과 시각화
```
plot(mpg_test)
```

단계6 : 분류분석 결과 해설
<해설> 실린더가 5이하이면 엔진크기에 의해서 23개가 분류되고, 실린더가 5이상이고, 6이하이면 27개가 분류되고, 6을 초과한 경우 21개가 분류된다.

04. weather 데이터를 이용하여 다음과 같은 단계별로 분류분석을 수행하시오.
조건1) rpart() 함수 이용 분류모델 생성
조건2) 변수 모델링 :
y변수(RainTomorrow), x변수(Date와 RainToday 변수 제외한 나머지 변수)
조건3) 비가 올 확률이 50% 이상이면 'Yes Rain', 50% 미만이면 'No Rain'으로 범주화

단계1 : 데이터 가져오기
```
library(rpart)
weather = read.csv("c:/Rwork/Part-IV/weather.csv", header=TRUE)
```

단계2 : 데이터 샘플링
```
weather.df <- weather[, c(-1,-14)]
nrow(weather.df)
idx <- sample(1:nrow(weather.df), nrow(weather.df)*0.7)
weather_train <- weather.df[idx, ]
weather_test <- weather.df[-idx, ]
```

단계3 : 분류모델 생성
```
weather_model <- rpart(RainTomorrow ~ ., data = weather.df)
weather_model # Humidity 중요변수
```

단계4 : 예측치 생성 : 검정데이터 이용
```
weater_pred <- predict(weather_model, weather_test)
weater_pred
```

단계5 : 예측 확률 범주화('Yes Rain', 'No Rain')
```
weater_class <- ifelse(weater_pred[,1] >=0.5, 'No Rain', 'Rain')
```

단계6 : 혼돈 행렬(confusion matrix) 생성 및 분류정확도 구하기
```
table(weater_class, weather_test$RainTomorrow)
```

```
> table(weater_class, weather_test$RainTomorrow)
weater_class No Yes
    No Rain 83   6
    Rain     2  19
(83 + 19) / nrow(weather_test)
[1] 0.9272727
```

[제16장 연습문제]

01. iris 데이터 셋의 1~4번째 변수를 대상으로 유클리드 거리 매트릭스를 구하여 idist에 저장한 후 계층적 클러스터링을 적용하여 결과를 시각화 하시오.

단계1. 유클리드 거리 계산

```
idist<- dist(iris[1:4]) # or dist(iris[, -5])
head(idist)
```

단계2. 계층형 군집분석(클러스터링)

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

단계3. 분류결과를 대상으로 음수값을 제거하여 덴드로그램 시각화

```
plot(hc, hang=-1) # 계층적 clustering 그래프(덴드로그램)
```

단계4. 그룹수를 4개로 지정하고 그룹별로 테두리 표시

```
rect.hclust(hc, k=4, border="red") # 4개 그룹 선정, 선 색 지정
```

02. 다음과 같은 조건을 이용하여 각 단계별로 비계층적 군집분석을 수행하시오.

작업 파일경로 : c:/Rwork/Part-IV/product_sales.csv

```
sales <- read.csv("c:/Rwork/Part-IV/product_sales.csv", header=TRUE)
```

단계1: 비계층적 군집분석 : 3개 군집으로 군집화

```
model <- kmeans(sales, 3) # 형식) kmeans(data, k) : k(군집수)
model # 원형데이터를 대상으로 3개 군집으로 군집화
# 각 케이스에 대한 소속 군집수(1,2,3) 확인
model$cluster # 각 케이스에 대한 소속 군집수(1,2,3)
```

단계2: 원형데이터에 군집수 추가

```
sales$group <- model$cluster
head(sales)# group 추가
```

단계3 : tot_price 변수와 가장 상관계수가 높은 변수와 군집분석 시각화

```
# (1) 상관관계 분석
cor(sales[,-5], method="pearson")
```

<해설> tot_price에 가장 큰 영향을 미치는 변수는 avg_price
(2) 비계층적 군집분석 시각화 : 그룹으로 색상 표시
```
plot(sales[c("tot_price", "avg_price")], col=sales$group)
```

단계4. 군집의 중심점 표시
```
points(result2$centers[,c("tot_price", "avg_price")], col=1:3, pch=8, cex=2)
```

03. tranExam.csv 파일을 대상으로 중복된 트랜잭션 없이 1~2칼럼만 single 형식으로 트랜잭션 객체를 생성하시오. (작업 파일경로 : C:/Rwork/Part-IV/tranExam.csv)

단계1 : 트랜잭션 객체 생성 및 확인
```
library(arules)
tranExam <- read.transactions("C:/Rwork/Part-IV/tranExam.csv", format="single",
                              sep=",", cols=c(1,2), rm.duplicates=T)
# 트랜잭션 데이터 보기
inspect(tranExam)
```

단계2 : 각 item별로 빈도수 확인
```
summary(tranExam)
```

단계3 : 파라미터(supp=0.3, conf=0.1)를 이용하여 규칙(rule) 생성
```
ruleExam<- apriori(tranExam) #set of 10 rules
ruleExam <- apriori(tranExam, parameter = list(supp=0.3, conf=0.1)) # 12 rule
```

단계4 : 연관규칙 결과 보기
```
inspect(ruleExam) # 12개 연관규칙
```

04. Adult 데이터셋을 대상으로 다음 단계별로 연관분석을 수행하시오.

조건 1) 최소 support=0.5, 최소 confidence=0.9를 지정하여 연관규칙을 생성한다.
```
data(Adult)
library(arules)
adult <- apriori(Adult, parameter = list(supp=0.5, conf=0.9))
adult # set of 52 rules
```

조건 2) 수행한 결과를 lift 기준으로 정렬하여 상위 10개 규칙을 기록한다.
```
rules<- inspect(head (sort(adult, by="lift"), 10))
```

조건 3) 연관분석 결과를 LHS와 RHS의 빈도수로 시각화한다.
```
library(arulesViz)
plot(adult) # 지지도, 신뢰도 향상도 산점도
```

```
plot(adult, method="grouped") # LHS와 RHS 간의 빈도수 시각화
```

조건 4) 연관분석 결과를 연관어 네트워크 형태로 시각화한다.

```
plot(adult, method="graph", control=list(type="items"))
```

조건 5) 연관어 중심 단어를 해설한다.

<해설> 자본이익과 자본손실의 단어를 중심으로 연관어가 형성되어 있다.

[제17장 연습문제]

01. 시계열 자료를 대상으로 다음 단계별로 시계열 모형을 생성하고, 미래를 예측하시오.

<데이터 셋 준비>

```
data(EuStockMarkets)
head(EuStockMarkets)
EuStock<- data.frame(EuStockMarkets)
head(EuStock)
Second <- c(1:500) # 초단 단위로 벡터 생성
DAX <- EuStock$DAX[1001:1500] # DAX 칼럼으로 백터 생성
EuStock.df <- data.frame(Second, DAX) # 데이터프레임 생성
```

단계1 : 시계열자료 생성 : EuStock.df$DAX 칼럼을 대상으로 2001년1월 단위

```
tsdata <- ts(EuStock.df$DAX, start=c(2001, 1), frequency=12)
```

단계2 : 시계열 자료 분해

(1) stl() 함수 이용 분해 시각화

(2) decompose() 함수 이용 분해 시각화, 불규칙요인만 시각화

```
plot(stl(tsdata, "periodic")) # 주기적인
m <- decompose(tsdata)
# 추세요인, 계절요인, 불규칙요인이 포함된 그래프
plot(m)
```

(3) 계절요인추세요인 제거 그래프-불규칙요인만 출력

```
plot(tsdata - m$seasonal - m$trend)
```

단계3 : ARIMA 시계열 모형 생성

```
library(forecast)
auto.arima(tsdata)  # 자동으로 최적의 ARIMA모형 제공
model<- auto.arima(tsdata)
```

단계4 : 향후 3년의 미래를 90%와 95% 신뢰수준으로 각각 예측 및 시각화

```
fore <- forecast(model, level=c(90, 95), h=36) # 향후 36개월(3년) 예측
plot(fore) # 시각화
```

02. Sales.csv 파일을 대상으로 시계열 자료를 생성하고, 각 단계별로 시계열 모형을 생성하여 예측하시오.

```
setwd("C:/Rwork/Part-IV")
goods <- read.csv("Sales.csv", header = TRUE)

단계1 : 시계열 자료 생성 : goods$Goods 칼럼으로 2015년 1월 기준 12개월 단위
tsGoods <- ts(goods$Goods, start=c(2015, 1), frequency=12)
tsGoods # 2015.01 ~ 2018.05

단계2 : 시계열모형 추정과 모형 생성
model <- auto.arima(tsGoods) # 시계열 예측모형 추정
model

단계3 : 시계열모형 진단 : Box-Ljung 검정 이용
Box.test(model$residuals, lag=1, type = "Ljung")

단계4 : 향후 7개월 예측
fore <- forecast(model, level=c(80, 95), h=7)
# 80% 신뢰구간(Lo 80~Hi80), 95% 신뢰구간(Lo 95 ~  Hi 95)

단계5 : 향후 7개월 예측결과 시각화
plot(fore)
```