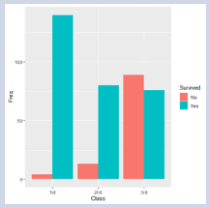
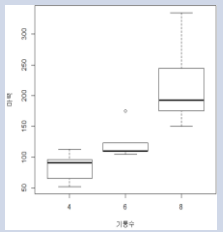
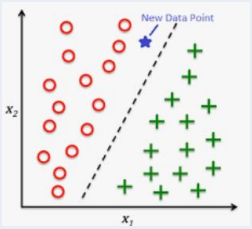
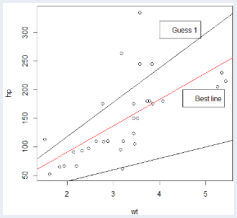
3장 이변량 기술통계
(Descriptive statistics with two variables)

최경미

자료 $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$, 표본크기 =n

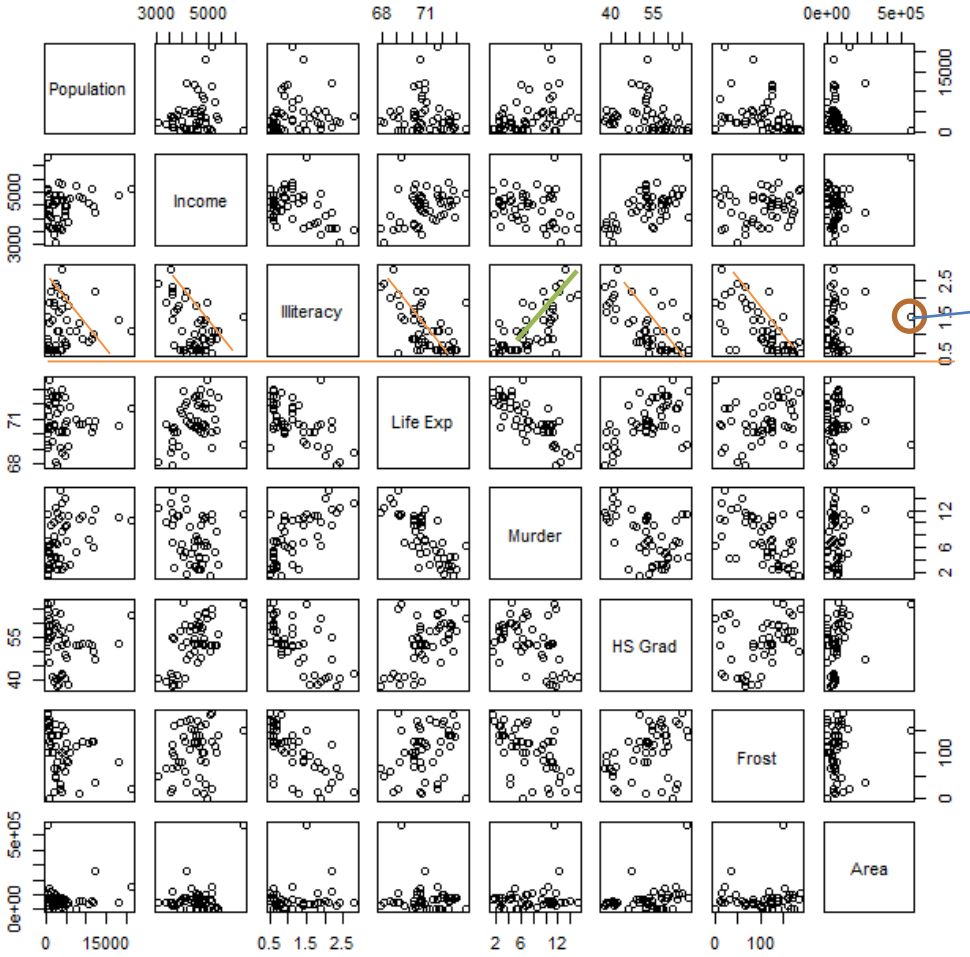
	Y 범주형	Y 연속형
X 범주형	교차표 (빈도, 상대빈도) 모자이크 그래프 집단 막대 그래프  	집단 비교 집단 평균, 표준편차 집단 상자도표 
X 연속형	AI, 머신러닝 로지스틱 회귀분석 	상관계수 산점도 직선식추정(y절편과 기울기) 

예제 (연속형, 연속형)

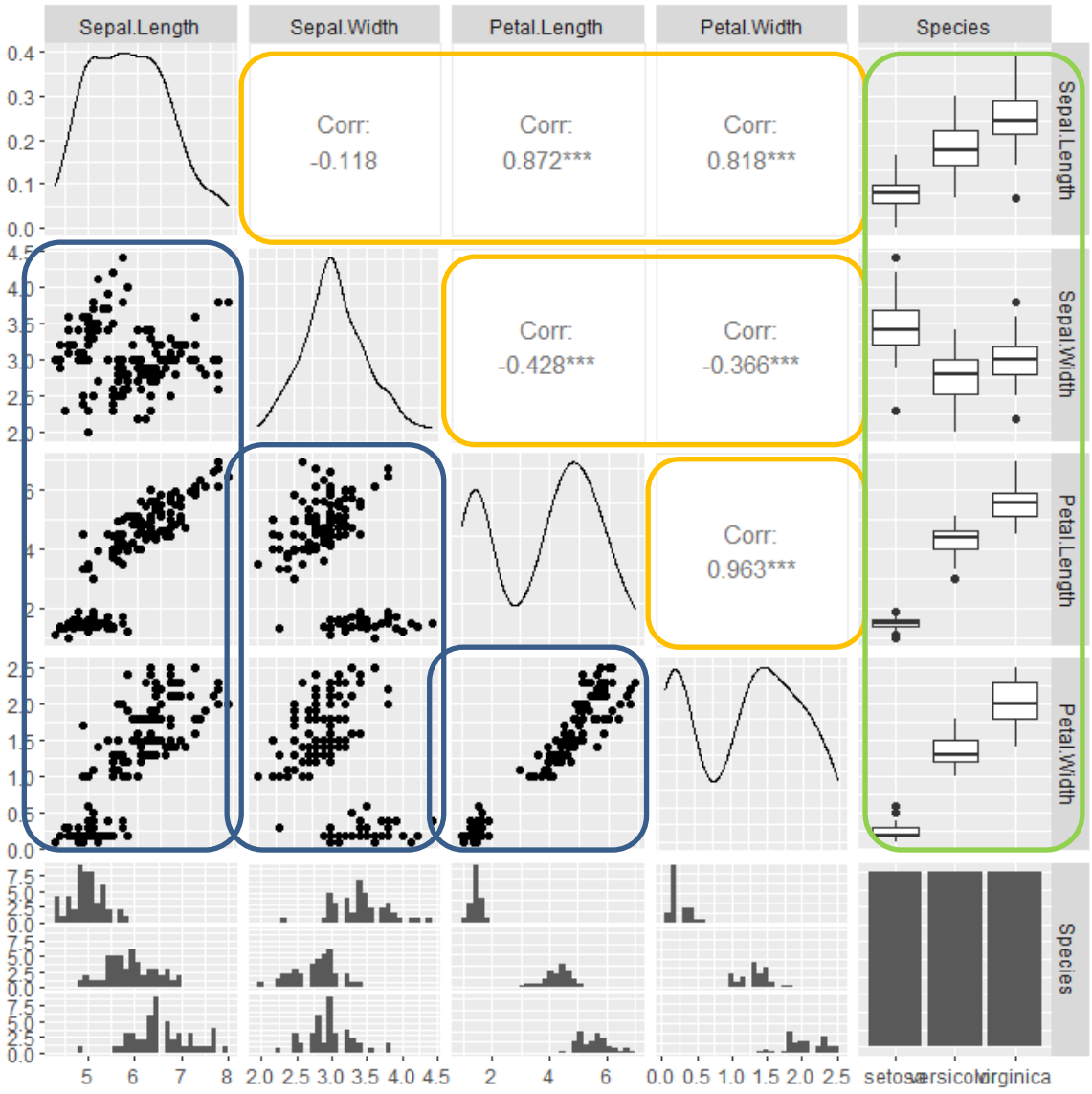
pairs(state.x77)

모든 변수들 사이의 산점도

- 1970년대
  미국 50개 주 평균
  - 인구
  - 연봉
  - 문맹률
  - 기대수명
  - 살인율
  - 고교졸업율
  - 서리일수
  - 면적



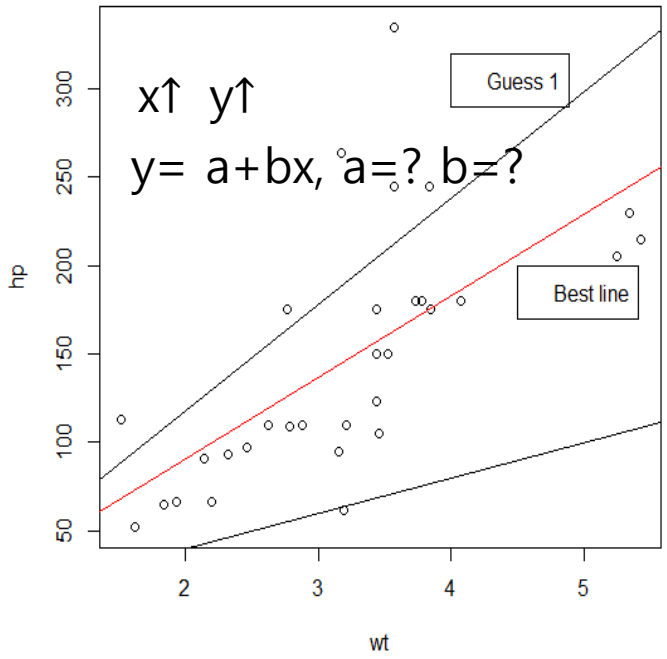
```
install.packages("GGally")
library(GGally)
ggpairs(iris)
```



예제 두 연속형 변수 사이의 산점도 (scatter plot)

- (110,2.62), (110,2.875), (293,2,32), (110,3.215), ……

Names	hp	wt
Mazda RX4	110	2.62
Mazda RX4 Wag	110	2.875
Datsun 710	93	2.32
Hornet 4 Drive	110	3.215
…	…	…



표본상관계수
(Sample correlation coefficient)

- X와 Y 둘 다 연속형.
- 표본크기 n

$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$

✓ x 편차제곱합

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 145726.9$$

✓ y 편차제곱합

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 29.67875$$

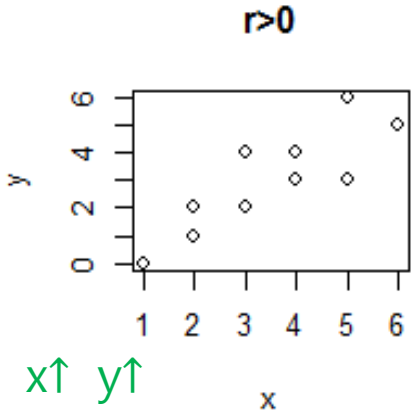
✓ x 편차, y 편차 곱의 합

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 1369.972$$

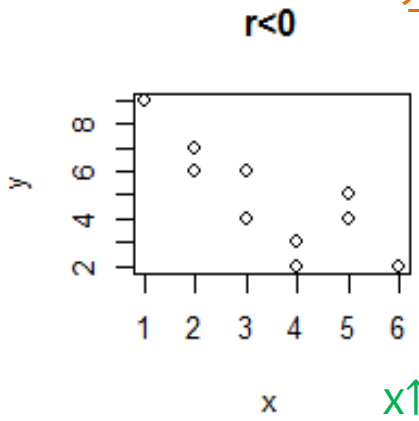
- 편차의 합 $= \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$, $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$
- $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{145726.9}{\sqrt{29.67875}\sqrt{1369.972}} = 0.6587479$

$-1 \leq r \leq 1$

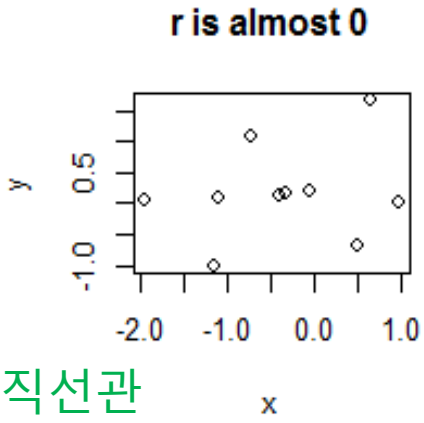
- 단위가 없음.
- 상관계수는 x와 y 사이의 직선관계 여부를 표현함.



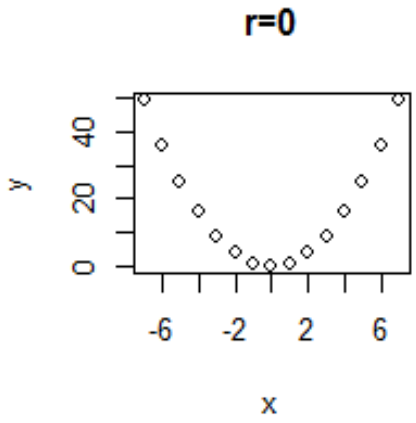
x↑ y↑



x↑ y↓



x와 y 사이에 직선관계가 보이지 않으면,
r ≈ 0



$y = x^2$의 관계가 있지만, 직선관계가 아니므로, r ≈ 0

r을 어떻게 계산할까?

id	x	x^2	y	y^2	xy
1	110	12100	2.620	6.864400	288.200
2	110	12100	2.875	8.265625	316.250
3	93	8649	2.320	5.382400	215.760
…	…	…	…	…	…
32	109	11881	2.780	7.728400	303.020
sum	$\sum_{i=1}^{10} x_i = 4694$	$\sum_{i=1}^{10} {x_i}^2 = 834278$	$\sum_{i=1}^{10} y_i = 102.952$	$\sum_{i=1}^{10} {y_i}^2 = 360.901$	$\sum_{i=1}^{10} x_i y_i = 16471.74$

$$S_{xx} = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$$
$$S_{yy} = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$
$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

```
rm(list=ls())
x <- mtcars$hp
y <- mtcars$wt


# 간단히
mean(x)
mean(y)
var(x)
sd(x)
var(y)
sd(y)
cor(x,y)

# 공식 따라서

n <- length(x)
sxx <- sum(x^2)-n*mean(x)^2

n <- length(y)
syy <- sum(y^2)-n*mean(y)^2

sxy <- sum(x*y)-n*mean(x)*mean(y)
den <- sqrt(sxx)*sqrt(syy)
r <- sxy/den
r
```

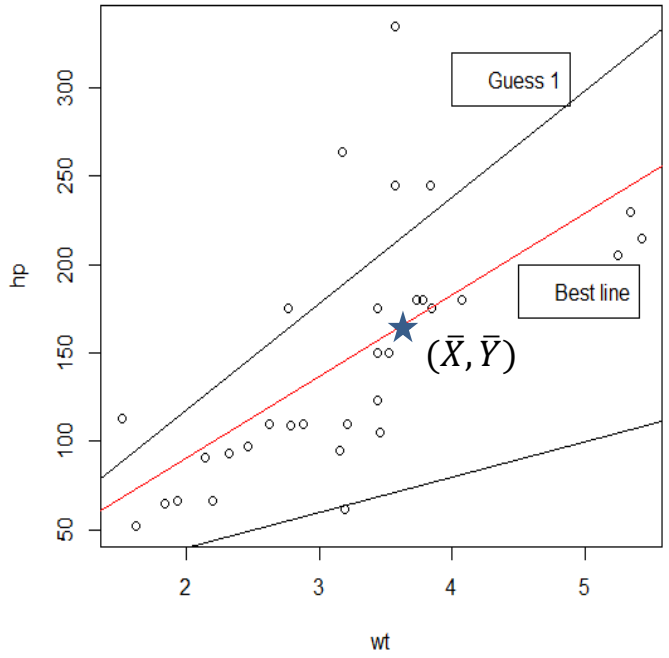
최적직선 추정
(Fit the best line)
예제 hp와 wt (둘 다 연속형)

크기 n인 표본

$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$

$$y_i = a + bx_i, \qquad i = 1,2,\ldots,n$$

Names	hp	wt
Mazda RX4	110	2.62
Mazda RX4 Wag	110	2.875
Datsun 710	93	2.32
Hornet 4 Drive	110	3.215
...	...	...



$$y = -1.821 + 46.160x$$

최적직선식 추정

```
# mtcars data
hp <- mtcars$hp
wt <- mtcars$wt
cor(hp, wt)
# 결과 r = 0.6587479

#또는
cor(mtcars$hp, mtcars$wt)
```

```
# Line equation (직선식)
plot(wt, hp)
abline(-2, 60)
abline(0, 20)

# linear model의 약자=lm
fit <- lm(hp~wt)
fit

#결과 Coefficients:
#(Intercept)          wt
#   -1.821       46.160

abline(fit)
```

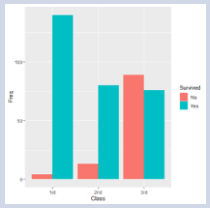
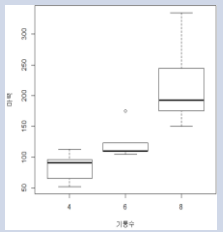
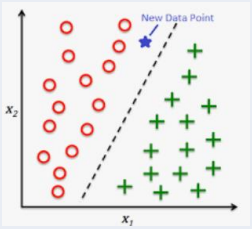
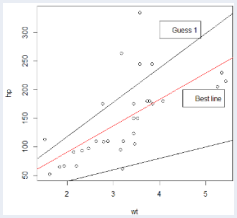
실습

1-3. 각 문항에서 주어진 R코드를 참고하여, 필요한 코드를 작성하시오.

1. R의 mtcars 중 자동 트랜스미션(am=0)이고 8기통(cyl=8) 차의 연비(mpg)와 마력(hp)를 이용하여, lm(hp~mpg)를 실행하자. 최적직선식 "마력 $= a + b$연비"에 대해 옳은 설명을 모두 고르시오.

1 $a = 185.885$ **2** $a = 285.885$ **3** $b = -6.094$ **4** $b = 6.094$ **5 위 보기 중 답 없음**

자료 $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$, 표본크기 =n

	Y 범주형	Y 연속형
X 범주형	교차표 (빈도, 상대빈도) 모자이크 그래프 집단 막대 그래프  	집단 비교 집단 평균, 표준편차 집단 상자도표 
X 연속형	AI, 머신러닝 로지스틱 회귀분석 	상관계수 산점도 직선식추정(y절편과 기울기) 

연속형 Y + 범주형 X 자료에 대한 상자도표

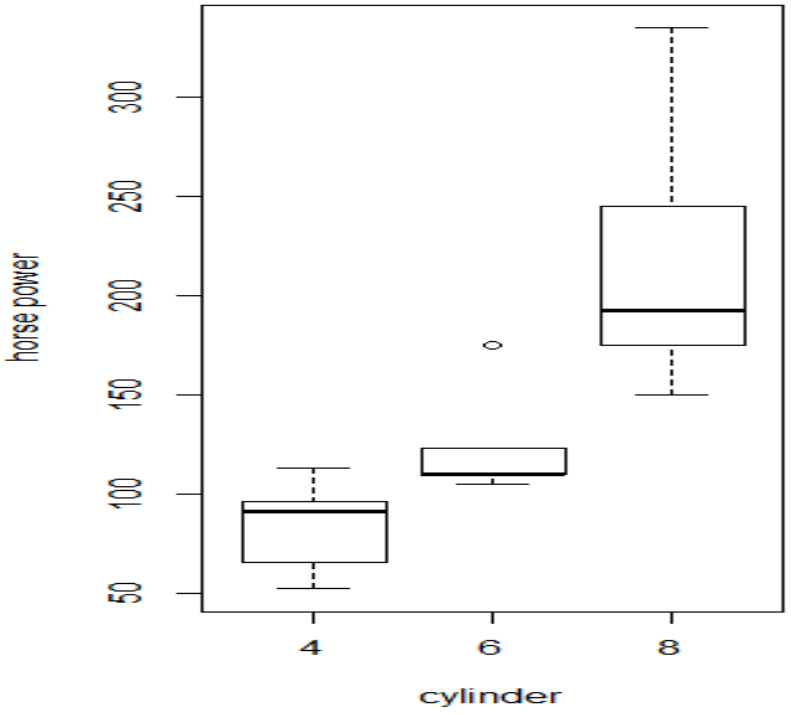
- 예제 mtcars in R:
- hp는 연속형
- cyl과 am은 범주형 (집단 표현)
- 분석목표: 집단 간 평균 비교

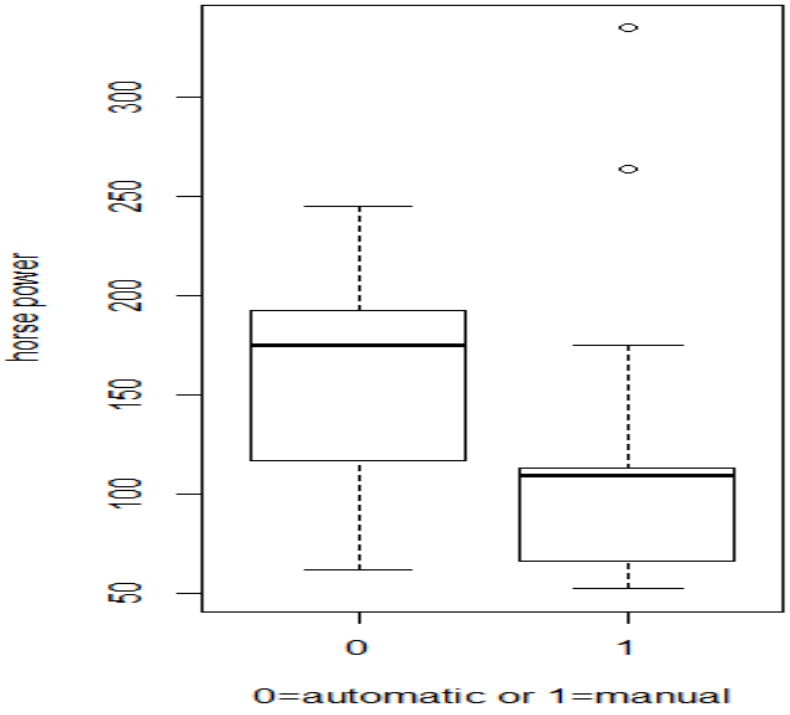
Names	cyl	hp	am
Mazda RX4	6	110	1
Mazda RX4 Wag	6	110	1
Datsun 710	4	93	1
Hornet 4 Drive	6	110	0
…	…	…	…

상자도표(Boxplots)

- 소형차 마력은 대형차 마력보다 작다. 참
- 중형차 마력은 대형차 마력보다 작다. 거짓

자동 T의 마력 중앙값은 수동 T 마력 중앙값보다 크다. 참





집단평균 (Group means)

- cyl 마다 hp의 평균계산

기통수	마력 평균 (표준편차)	표본크기(n)
4	82.6 (20.9)	11
6	122.3 (24.3)	7
8	209.2 (51.0)	14

- am (a=오토, m=매뉴얼) 마다 hp의 평균계산

트랜스미션	마력의 중앙값 (최소값, 최대값)	표본크기(n)
자동 (0)	175 (62, 245)	19
수동 (1)	109 (52, 335)	13

R 실습 cyl와 hp

- 변수 저장 후, 상자도표 그리기

```
# hp  비척도, cyl 집단 표현하는 명목척도
boxplot(hp~cyl, data=mtcars,
        xlab= " cylinder”,
        ylab= " horse power”)

# cyl=4,6,8 집단 별로 hp의 평균 표
aggregate(hp~cyl, data=mtcars, mean)
```

R 실습 am과 hp

```
# 상자도표 그리기
boxplot(hp~am, data=mtcars,
        xlab="0=automatic or 1=manual",
        ylab="horse power")

# 자동과 수동 집단 별로 hp의 평균 표
aggregate(hp~am, data=mtcars, mean)
```

R 실습 am, cyl와 hp

```
# 상자도표
 boxplot(hp~cyl+am, data=mtcars, xlab="기통수 + 트랜스미션", ylab="마력")

# 표 만들기
 aggregate(hp~cyl+am, data=mtcars, FUN=median)
 aggregate(hp~cyl+am, data=mtcars, FUN=min)
 aggregate(hp~cyl+am, data=mtcars, FUN=max)
```

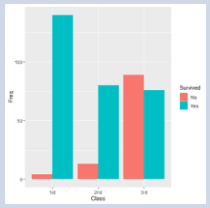
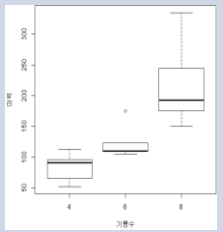
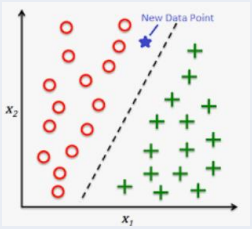
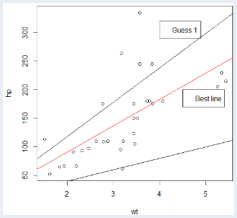
실습

2. R의 iris 자료에서 세토사(setosa), 버시컬러(versicolor), 버지니카(virginica)의 꽃받침 길이(Sepal.Length)와 넓이(Sepal.Width)를 비교한 설명의 참(T) 거짓(F)을 판별하시오. 다음의 R코드를 참고하여, 필요한 코드를 짜고 실행하자.

```
iris
names(iris)
help(iris)
boxplot(Sepal.Length~Species, data=iris)
aggregate(Sepal.Length ~Species, data=iris, mean)
boxplot(Sepal.Width~Species, data=iris)
aggregate(Sepal.Width~Species, data=iris, mean)
```

1 셋 중 꽃받침의 평균 길이가 가장 짧은 아이리스는 세토사이다.
2 버시컬러의 꽃받침 넓이 중앙값이 세 아이리스 중 제일 작다.
3 꽃받침 길이의 IQR이 가장 큰 아이리스는 세토사이다.
4 세토사의 꽃받침 넓이의 표준편차는 버지니카의 꽃받침 넓이의 표준편차보다 작다.
5 버시컬러의 평균 꽃받침 길이가 세토사의 평균 꽃받침 길이보다 크다.

자료 $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$, 표본크기 =n

	Y 범주형	Y 연속형
X 범주형	교차표 (빈도, 상대빈도) 모자이크 그래프 집단 막대 그래프  	집단 비교 집단 평균, 표준편차 집단 상자도표 
X 연속형	AI, 머신러닝 로지스틱 회귀분석 	상관계수 산점도 직선식추정(y절편과 기울기) 

두 개의 범주형 변수에 대한 a× b 교차표

- Titanic 여성어른자료

① 등급 : 1등실, 2등실, 3등실

② 생존여부 : 생존, 사망

```
# 표 만들기
Titanic
mytable <- Titanic[1:3, "Female", "Adult", ]
Install.packages("gmodels")
library(gmodels)
CrossTable(mytable)
plot(mytable)

pt <- prop.table(mytable, 1)*100
barplot(pt, beside=T,
        legend=c("1등실",    "2등실", "3등실"))
```

- mtcars

① cyl : 4기통, 6기통, 8기통

② am : am=0 오토, am=1 매뉴얼

```
# 빈도, 상대빈도 교차표
# (relative) frequency table, cross table
am<-mtcars$am
cyl<-mtcars$cyl

library(gmodels)
CrossTable(am, cyl)

# 막대그래프
barplot
mytable <- table(am,cyl)
mytable
plot(table)
pt <- prop.table(mytable,1)
barplot(pt, beside=T,
legend=c("am=0","am=1"))
```

- 다른 방법

```
CrossTable(am,cyl)$t
plot(CrossTable(am,cyl)$t)
```

예제: Titanic 성인여성 자료의 빈도표

N 전체 비율 (%) 행 비율 (%) 열 비율 (%)	사망	생존	합
1등실	4 0.995% 2.778% 3.774%	140 34.825% 97.222% 47.297%	144 35.368%
2등실	13 3.234% 13.978% 12.264%	80 19.900% 86.022% 27.027%	93 23.134%
3등실	89 22.139% 53.939% 83.962%	76 18.905% 46.061% 25.676%	165 41.045%
합	106 26.368%	296 73.632%	402 100%

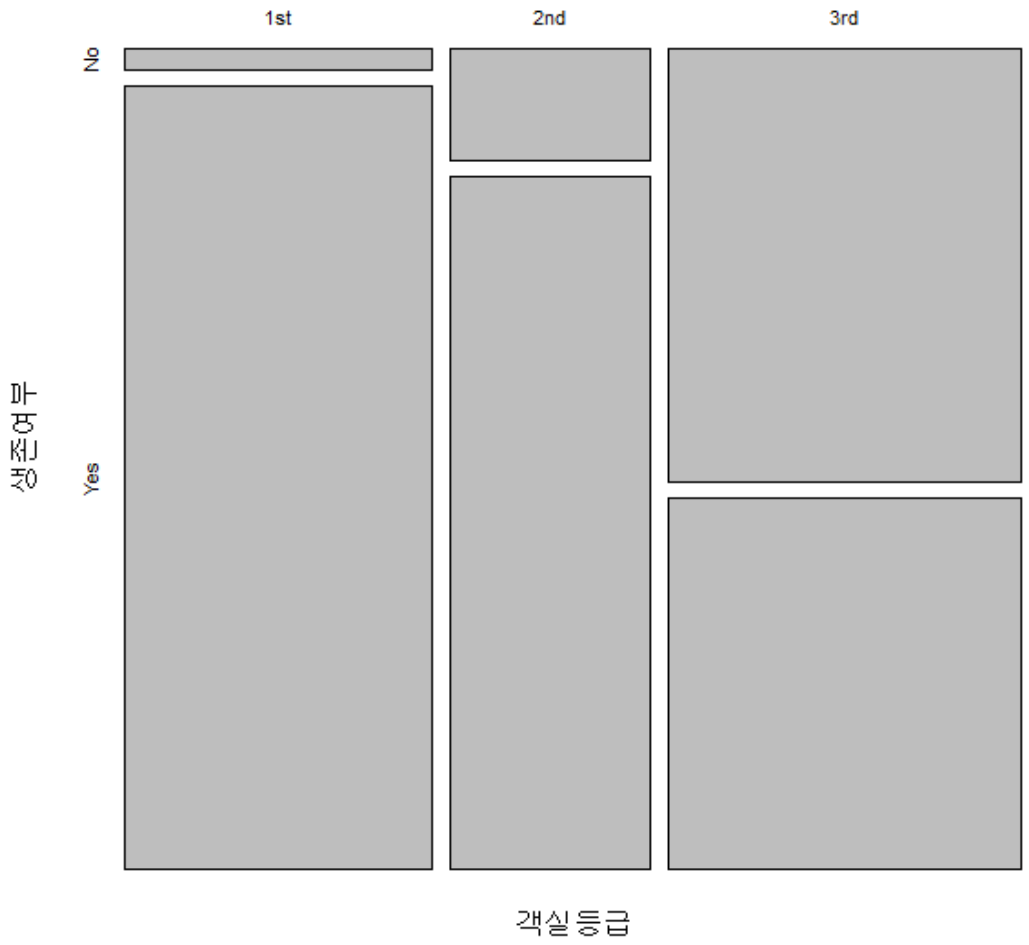
1 등실 사망자는 전체 중에서 0.995%이다.

생존자 중에서 1등실 비율은 47.29%이다.

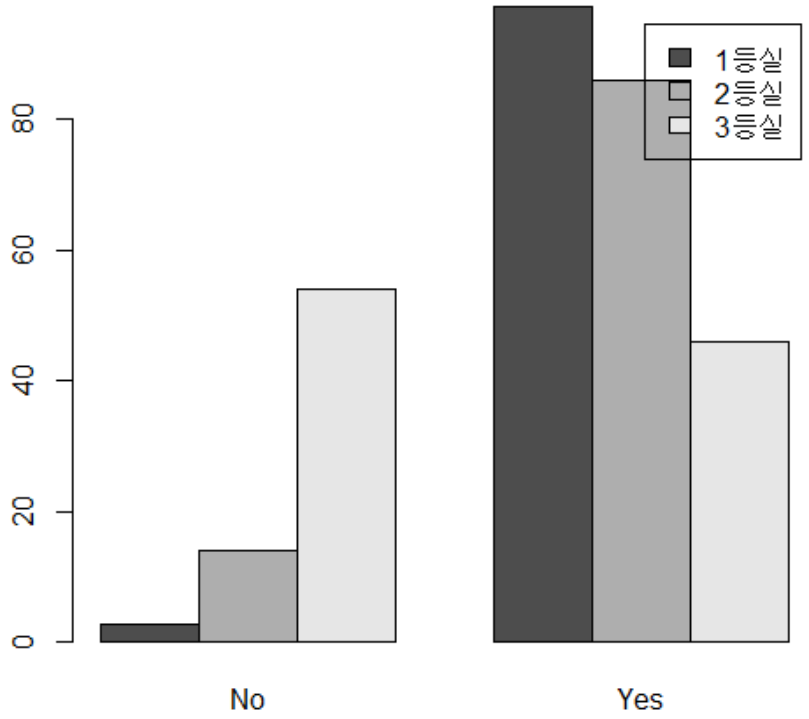
2등실 승객은 전체 중에서 23.134%이다.

3등실 승객 중에서 사망자는 53.939%이다.

모자이크 그래프



막대 그래프



실습

3. R의 mtcars에서 기어수(gear=3,4,5)와 자동/수동 트랜스미션(am=0/1)인 자동차에 대한 설명의 참(T) 거짓(F)을 판별하시오.
(힌트 gmodels 패키지의 CrossTable(am, gear)을 실행시키자.)

1 기어수가 4이고, 자동 트랜스미션인 차의 수는 4대이다. T
2 기어수가 5인 차들 중에서 자동 트랜스미션인 차의 비율은 1이다. F
3 전체 자동차의 40.6%가 수동 트랜스미션이다. T
4 수동 트랜스미션 자동차들 중에서 기어수가4개인 자동차의 비율은 0.615이다. T
5 수동 트랜스미션이고, 기어수가 5개인 자동차는 전체의 38.5%이다. F

과제

3장 연습문제 8,9,10