제6강
다중변수 자료의 탐색

Section 01

산점도

Section 02

상관 분석

Section 03

선그래프

Section 04

자료의 탐색 실습

4. 자료의 탐색 실습

1. Boston Housing 데이터셋 소개

- 미국 보스턴 지역의 주택 가격 정보와 주택 가격에 영향을 미치는 여러 요소들에 대한 정보를 담고 있음
- 총 14개의 변수로 구성이 되어 있는데, 여기서는 이중에 5개의 변수만 선택하여 분석
- mlbench 패키지에서 제공

변수	설명
crim	지역의 1인당 범죄율
rm	주택 1가구당 방의 개수
dis	보스턴의 5개 직업 센터까지의 거리
tax	재산세율
medv	주택 가격

표 6-1 BostonHousing 데이터셋의 변수 설명

4. 자료의 탐색 실습

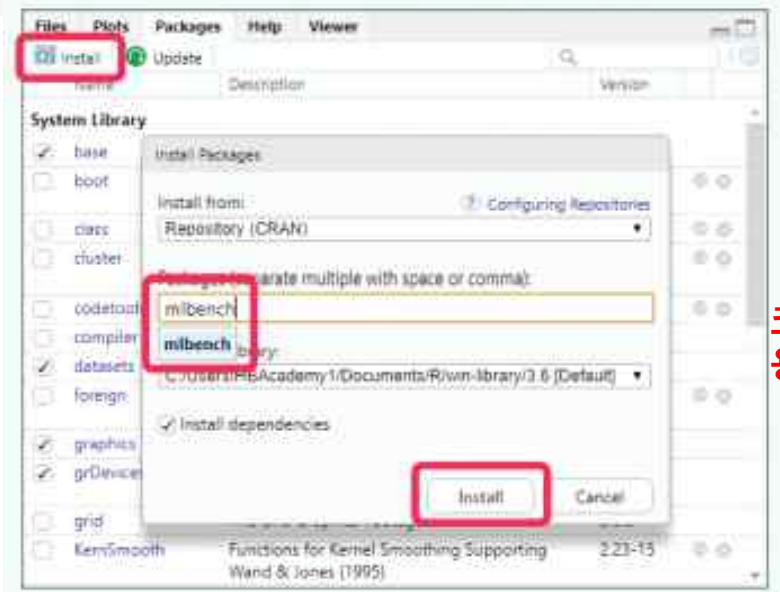
2. 탐색적 데이터 분석 과정

1.1 분석 대상 데이터셋 준비

```
> library(mlbench)
> data("BostonHousing")
> myds <- BostonHousing[,c("crim","rm","dis","tax","medv")]
```

library()는 패키지를 R프로그램에 로딩을 한다.

data()함수는 데이터를 가져오는 역할을 한다.



코드에서 install.package()함수를 이용해서 패키지를 설지해도 된다.

그림 6-8 mlbench 패키지 설치

4. 자료의 탐색 실습

1.2 grp 변수 추가

- grp는 주택 가격을 상(H), 중(M), 하(L)로 분류한 것으로 25.0 이상이면 상(H), 17.0 이하이면 하(L), 나머지를 중(M)으로 분류

```
> grp <- c()
> for (i in 1:nrow(myds)) {                    # myds$medv 값에 따라 그룹 분류
+     if (myds$medv[i] >= 25.0) {
+       grp[i] <- "H"
+     } else if (myds$medv[i] <= 17.0) {
+       grp[i] <- "L"
+     } else {
+       grp[i] <- "M"
+     }
+ }
> grp <- factor(grp)                           # 문자 벡터를 팩터 타입으로 변경
> grp <- factor(grp, levels=c("H","M","L"))    # 레벨의 순서를 H, L, M -> H, M, L

> myds <- data.frame(myds, grp)                # myds에 grp 열 추가
```

BostonHousing데이터 셋은 506행과 11개의 변수를 가지고 있다.

medv(주택가격)에 따라 grp벡터변수에 각각 H, L, M으로 분리하여 저장하고 있다.

4. 자료의 탐색 실습

1.3 데이터셋의 형태와 기본적인 내용 파악

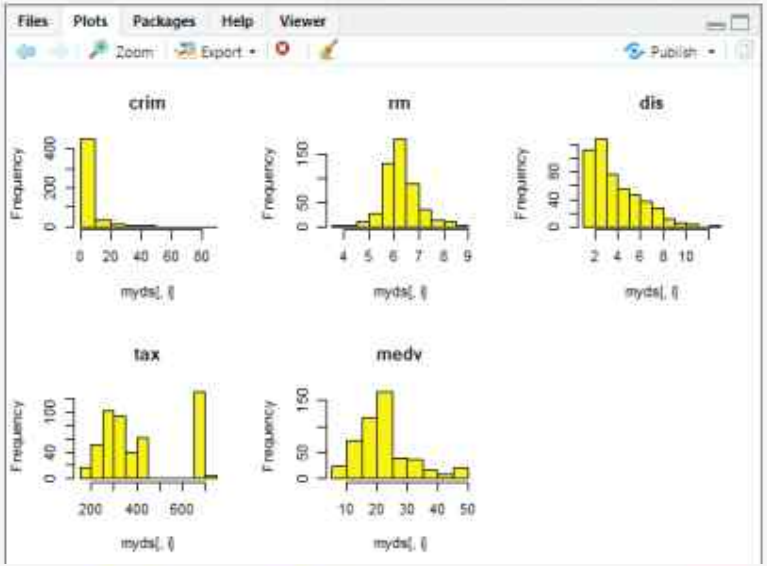
```
> str(myds)
'data.frame':506 obs. of  6 variables:
 $ crim: num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ rm  : num  6.58 6.42 7.18 7 7.15 ...
 $ dis : num  4.09 4.97 4.97 6.06 6.06 ...
 $ tax : num  296 242 242 222 222 222 311 311 311 311 ...
 $ medv: num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
 $ grp : Factor w/ 3 levels "H","L","M": 3 3 1 1 1 1 3 1 2 3 ...
> head(myds)
    crim    rm    dis tax medv grp
1 0.00632 6.575 4.0900 296 24.0   M
2 0.02731 6.421 4.9671 242 21.6   M
3 0.02729 7.185 4.9671 242 34.7   H
4 0.03237 6.998 6.0622 222 33.4   H
5 0.06905 7.147 6.0622 222 36.2   H
6 0.02985 6.430 6.0622 222 28.7   H
> table(myds$grp)                          # 주택 가격 그룹별 분포

  H   M   L
132 247 127
```

4. 자료의 탐색 실습

1.4 히스토그램에 의한 관측값의 분포 확인

```
> par(mfrow=c(2,3))                    # 2x3 가상화면 분할
> for(i in 1:5) {
+   hist(myds[,i], main=colnames(myds)[i], col="yellow")
+ }
```



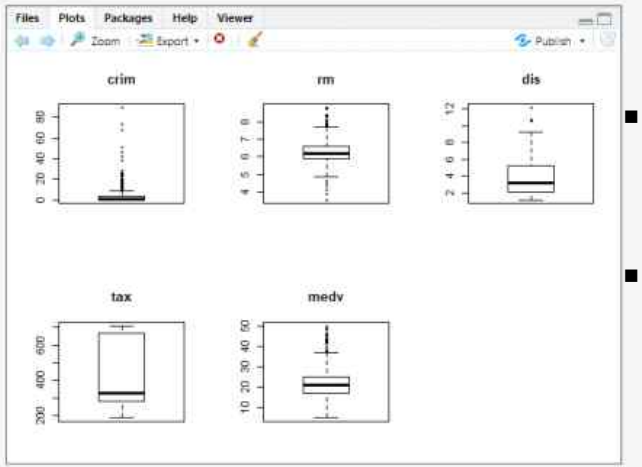

```
> par(mfrow=c(1,1))                    # 2x3 가상화면 분할 해제
```

- rm(방의 개수), medv(주택가격) 변수만 종 모양의 정규분포에 가깝고, crim(1인당 범죄율), dis(직업센터까지 거리)는 관측값들이 한쪽으로 쏠려서 분포함.
- tax는 중간에 관측값이 없는 빈 구간이 존재하는 특징임.
- 이런 경우는 데이터 분석결과를 부동산 전문가와 함께 분석을 해볼 필요성이 있는 것이다.

4. 자료의 탐색 실습

1.5 상자그림에 의한 관측값의 분포 확인

```
> par(mfrow=c(2,3))                    # 2x3 가상화면 분할
> for(i in 1:5) {
+   boxplot(myds[,i], main=colnames(myds)[i])
+ }
> par(mfrow=c(1,1))                    # 2x3 가상화면 분할 해제
```

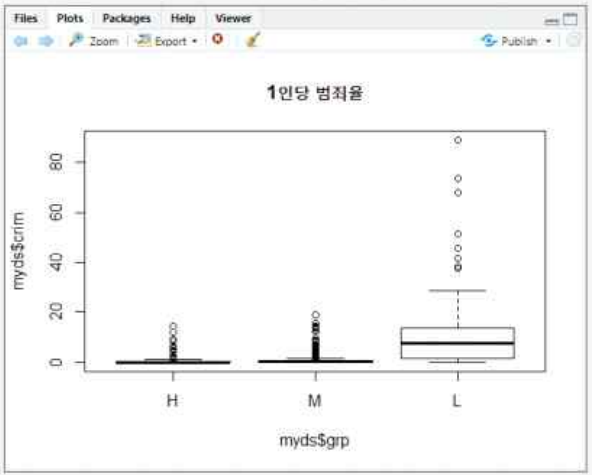


- 1인당 범죄율(crim)은 관측값들이 좁은 지역에 밀집되어 있음(관측값들의 편차가 매우 작음)
- 재산세율(tax)은 넓게 퍼져 있는 것(관측값들의 편차가 비교적 크다)을 확인

4. 자료의 탐색 실습

1.6 그룹별 관측값 분포의 확인

```
> boxplot(myds$crim~myds$grp, main="1인당 범죄율")
```

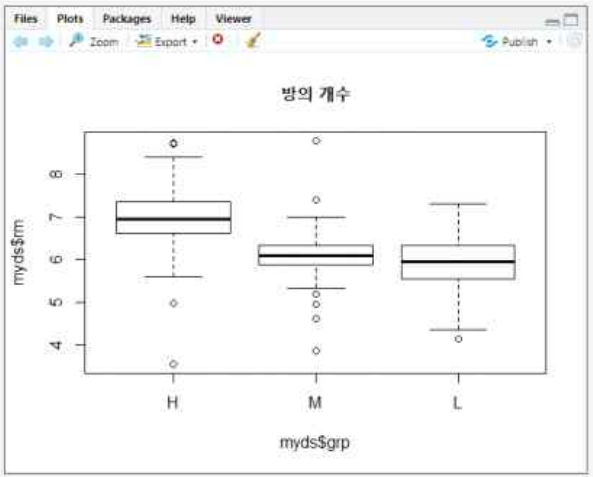


- 주택 가격이 높은 지역이나 중간 지역의 범죄율은 낮고, 주택 가격이 낮은 지역의 범죄율이 높게 나타남

4. 자료의 탐색 실습

1.6 그룹별 관측값 분포의 확인

```
> boxplot(myds$rm~myds$grp, main="방의 개수"
```

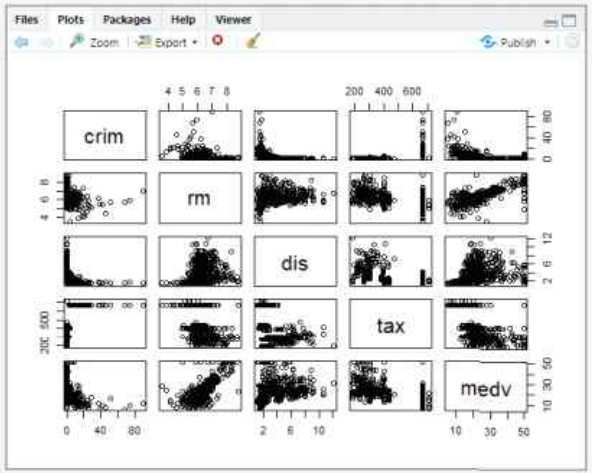


- **주택 가격이 높으면 방의 개수도 많다는 것을 알 수 있음**
- **주택 가격이 중간인 지역과 하위인 지역의 방의 개수 평균은 큰 차이가 나지 않음**
- **중간 그룹의 방의 개수가 5.2~6.8 사이로 비교적 균일한 반면 하위그룹의 방의 개수는 4.5~7.2 사이로 넓게 퍼져 있는 것을 알 수 있음**

4. 자료의 탐색 실습

1.7 다중 산점도를 통한 변수 간 상관 관계의 확인

```
> pairs(myds[,-6])
```

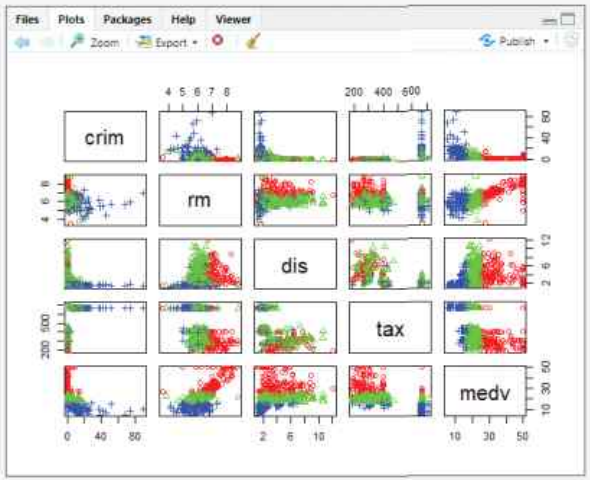


- medv(주택 가격)과 양의 상관성이 있는 변수는 rm(가구당 방의 개수) 인데, 가구당 방의 개수가 많으면 집이 넓으니 주택가격이 높을 것으로 보임
- 상대적으로 crim(1인당 범죄율)은 주택 가격과 음의 상관성이 있는 것으로 보이는데 이것은 범죄율이 높을수록 주택가격이 떨어진다는 의미로 해석이 가능함.

4. 자료의 탐색 실습

1.8 그룹 정보를 포함한 변수 간 상관 관계의 확인

```
> point <- as.integer(myds$grp)                    # 점의 모양 지정
> color <- c("red","green","blue")                 # 점의 색 지정
> pairs(myds[,-6], pch=point, col=color[point])
```



- (crim-medv), (rm-medv), (dis-medv), (tax-medv) 산점도에서 그룹별로 분포 위치가 뚜렷하게 구분
- 주택 가격 중간 그룹(녹색점들)은 상위 그룹(빨간색), 하위 그룹(파란색)에 비해 주택 가격의 변동폭이 좁음

4. 자료의 탐색 실습

1.9 변수 간 상관계수의 확인

```
> cor(myds[,-6])
           crim         rm        dis        tax       medv
crim  1.0000000 -0.2192467 -0.3796701  0.5827643 -0.3883046
rm   -0.2192467  1.0000000  0.2052462 -0.2920478  0.6953599
dis  -0.3796701  0.2052462  1.0000000 -0.5344316  0.2499287
tax   0.5827643 -0.2920478 -0.5344316  1.0000000 -0.4685359
medv -0.3883046  0.6953599  0.2499287 -0.4685359  1.0000000
```

상관계수를 살펴보면 역시 medv(주택가격)을 기준으로 보았을 때, rm(방의 개수)과 가장 높은 양의 상관계수가 높으며, crim(범죄율)은 음의 상관계수가 -0.39정도 수준으로, 상관계수로 보았을때는 상관도가 높지 않다는 것을 알 수가 있다

상관계수의 일반적 기준
1) -1.0과 -0.7 사이이면, 강한 음적 선형관계,
2) -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계,
3) -0.3과 -0.1 사이이면, 약한 음적 선형관계,
4) -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계,
5) +0.1과 +0.3 사이이면, 약한 양적 선형관계,
6) +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계,
7) +0.7과 +1.0 사이이면, 강한 양적 선형관계.

코드 6-8

```
## (1) Prepare Data ---------------------
library(mlbench)
data("BostonHousing")
myds <- BostonHousing[,c("crim","rm","dis","tax","medv")]

## (2) Add new column ---------------------
grp <- c()
for (i in 1:nrow(myds)) {                    # myds$medv 값에 따라 그룹 분류
```

4. 자료의 탐색 실습

```
if (myds$medv[i] >= 25.0) {
 grp[i] <- "H"
 } else if (myds$medv[i] <= 17.0) {
 grp[i] <- "L"
 } else {
 grp[i] <- "M"
 }
}
grp <- factor(grp)                          # 문자벡터를 팩터 타입으로 변경
grp <- factor(grp, levels=c("H","M","L")) # 레벨의 순서를 H,L,M -> H,M,L

myds <- data.frame(myds, grp)               # myds 에 grp 컬럼추가

## (3) Add new column ----------------------
str(myds)
head(myds)
table(myds$grp)                             # 주택 가격 그룹별 분포
```

4. 자료의 탐색 실습

```
## (4) histogram ----------------------
par(mfrow=c(2,3))                                  # 2x3 가상화면 분할
for(i in 1:5) {
 hist(myds[,i], main=colnames(myds)[i], col="yellow")
}
par(mfrow=c(1,1))                                  # 2x3 가상화면 분할 해제

## (5) boxplot ----------------------
par(mfrow=c(2,3))                                  # 2x3 가상화면 분할
for(i in 1:5) {
 boxplot(myds[,i], main=colnames(myds)[i])
}
par(mfrow=c(1,1))                                  # 2x3 가상화면 분할 해제

## (6) boxplot by group ------------------
boxplot(myds$crim~myds$grp, main="1인당 범죄율")
boxplot(myds$rm~myds$grp, main="방의 수")
boxplot(myds$dis~myds$grp, main="직업센터까지의 거리")
boxplot(myds$tax~myds$grp, main= " 재산세")
```

4. 자료의 탐색 실습

```
## (7) scatter plot ------------------
pairs(myds[,-6])

## (8) scatter plot with group ------------------
point <- as.integer(myds$grp)                    # 점의 모양 지정
color <- c("red","green","blue")                 # 점의 색 지정
pairs(myds[,-6], pch=point, col=color[point])

## (9) correlation coefficient ------------------
cor(myds[,-6])
```

감사합니다.