MULTIVARIATE ANALYSIS

PCA

주택 가격에 영향을 미치는 주요 요인을 확인

2020251005 주하연

content

데이터 소개 및 연구목적

Dataset : Bostonhousing

이 데이터셋은 보스턴 지역의 주택 가격과 관련된 속성 데이터를 포함하고 있는 데이터셋

궁금증 : 한국에서는 서울과 가까울수록 집값 UP. 그럼, 미국에서는??

-> 연구 목적 : 주택 가격에 영향을 미치는 주요 요인을 확인(PCA)

https://www.kaggle.com/datasets/huhao05133/bostonhousing?select=BostonHousing.csv

변수 설명

crim : 자치시(town)별 1인당 범죄율
zn : 25,000 평방피트를 초과하는 거주지역의 비율
indus : 비소매상업지역이 점유하고 있는 토지의 비율
chas : 찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0 (범주형)
nox : 10ppm당 농축 일산화질소
rm : 주택 1가구당 평균 방의 개수
age : 1940년 이전에 건축된 소유주택의 비율
dis : 5개의 보스턴 직업센터까지의 접근성 지수
rad : 방사형 도로까지의 접근성 지수
tax : 10,000 달러 당 재산세율
ptratio : 자치시(town)별 학생/교사 비율
black : 1000(Bk-0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함
lstat : 모집단의 하위계층 비율(%)
medv : 본인 소유의 주택가격(중앙값) (단위: $1,000)

데이터탐색

기초통계량, 결측치 확인

결측치는 존재하지 않는다.

```
> dat <- read.csv("BostonHousing.csv", header = TRUE)
> summary(dat)
      crim                zn              indus             chas        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
      nox                rm              age              dis         
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax           ptratio          black       
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
     lstat            medv      
 Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 6.95   1st Qu.:17.02  
 Median :11.36   Median :21.20  
 Mean   :12.65   Mean   :22.53  
 3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :37.97   Max.   :50.00  
```

```
> sum(is.na(dat))
[1] 0
```

데이터탐색

자료구조,관측치, 자료형 확인

*14개의 변수를 갖는 506개의 샘플로 구성

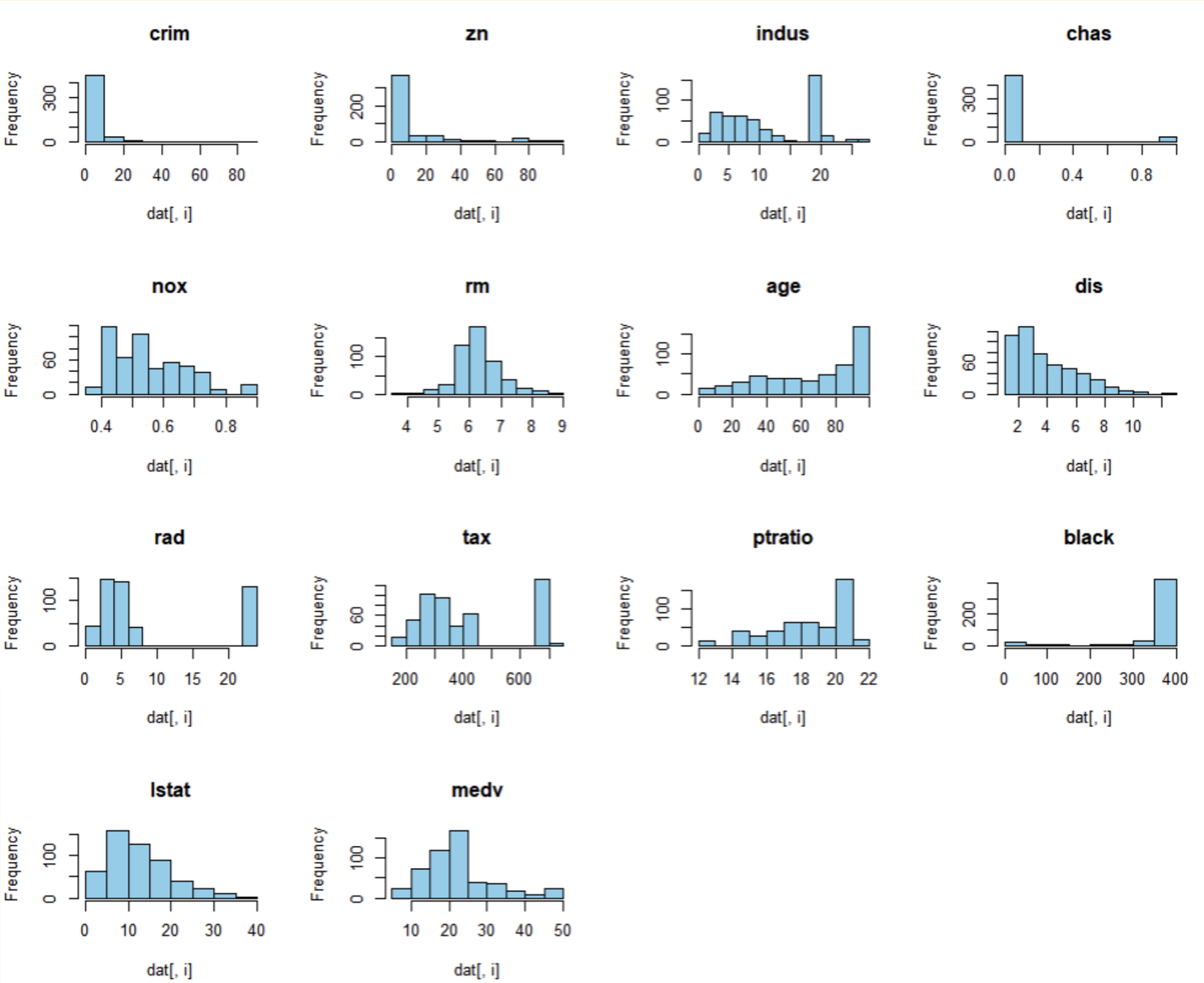
```
> str(dat)
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524
...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : int  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
> head(dat)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
  lstat medv
1  4.98 24.0
2  9.14 21.6
3  4.03 34.7
4  2.94 33.4
5  5.33 36.2
6  5.21 28.7
> tail(dat)
       crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
501 0.22438  0  9.69    0 0.585 6.027 79.7 2.4982   6 391    19.2 396.90
502 0.06263  0 11.93    0 0.573 6.593 69.1 2.4786   1 273    21.0 391.99
503 0.04527  0 11.93    0 0.573 6.120 76.7 2.2875   1 273    21.0 396.90
504 0.06076  0 11.93    0 0.573 6.976 91.0 2.1675   1 273    21.0 396.90
505 0.10959  0 11.93    0 0.573 6.794 89.3 2.3889   1 273    21.0 393.45
506 0.04741  0 11.93    0 0.573 6.030 80.8 2.5050   1 273    21.0 396.90
    lstat medv
501 14.33 16.8
502  9.67 22.4
503  9.08 20.6
504  5.64 23.9
505  6.48 22.0
506  7.88 11.9
```

데이터탐색

변수들의 분포 확인

```
par(mfrow=c(4,4))                                    #4x4 가상화면 분할
for(i in 1:14) {
  hist(dat[,i],main=colnames(dat)[i],col="skyblue")
}
```



결과를 보면 rm,mdev 변수만 종 모양의 정규분포에 가깝고, crim, zn, dis, black, lstat 은 관측값들이 한쪽으로 쏠려서 분포함을 알 수 있다.
rad, tax는 중간에 관측값이 없는 빈 구간이 존재하는 특징을 보인다.

데이터탐색

grp 변수 추가
: 주택 가격을 21.0을 기준으로 각각 H, L로 grp변수에 분류해주어서 데이터 프레임에 추가

table() 함수를 통해 도수분포표를 작성하면 가격이 높은 주택이 260채, 가격이 낮은 주택이 246채임을 알 수 있다.

```
> #주택 가격을 H,M,L로 나누기 위한 grp변수추가
> grp <- c()
> for (i in 1:nrow(dat)){  #dat$medv 값에 따라 그룹 분류
+   if (dat$medv[i] >= 21.0){
+     grp[i] <- "H"
+   } else {
+     grp[i] <- "L"
+   }
+ }
> grp <-factor(grp)  #문자벡터를 팩터 타입으로 변경
> grp <- factor(grp, levels=c("H","L"))  #레벨의 순서를 H,L -> H,L
> dat <- data.frame(dat, grp)  #myds에 grp 컬럼 추가
```

```
 $ grp    : Factor w/ 2 levels "H","L": 1 1 1 1 1 1 1 1 1 2 2 ...
> head(dat)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
  lstat medv grp
1  4.98 24.0   H
2  9.14 21.6   H
3  4.03 34.7   H
4  2.94 33.4   H
5  5.33 36.2   H
6  5.21 28.7   H
> table(dat$grp)

  H   L
260 246
```

데이터탐색

두 개의 그룹의 평균벡터가 차이가 나는지 안나는지
: Hotelling.test

```
> colMeans(groupH)
      crim         zn      indus       chas        nox         rm
 0.8729023 18.4961538  7.9236154  0.1000000  0.5025188  6.6268885
       age        dis        rad        tax    ptratio      black
55.0484615  4.4302200  6.3923077 331.9500000 17.5126923 384.8915385
     lstat       medv
 7.9924231 28.8488462
> colMeans(groupL)
      crim         zn      indus       chas        nox
 6.51011516  3.82520325 14.53280488  0.03658537  0.60984065
        rm         age        dis        rad        tax
 5.92290244 82.87113821  3.12371707 12.88617886 488.86585366
   ptratio      black      lstat       medv
19.45203252 326.85065041 17.57894309 15.85731707
```

```
> result <- hotelling.test(x = groupH, y = groupL)
> result
Test stat:  775.77
Numerator df:  14
Denominator df:  491
P-value:  0
```

귀무가설 : 두개의 mean vector가 동일하다.

p-value가 0에 수렴하기 때문에 귀무가설을 기각한다.
-> 유의미한 차이를 보인다.

데이터탐색

두 그룹의 Covariance Matrix가 동일한지 아닌지
: Box's M test

```
#두 그룹의 Covariance Matrix가 동일한지 아닌지 변수간 상관관계 분석
round(cov(groupH), 2) # High
round(cov(groupL), 2) # Low

library(heplots)

result <- boxM(cbind(crim, zn, indus, chas, nox, rm, age, dis, rad, tax,
                     ptratio, black, lstat) ~ grp, data = dat)
result
```

```
> result <- boxM(cbind(crim, zn, indus, chas, nox, rm, age, dis, rad, tax,
+                      ptratio, black, lstat) ~ grp, data = dat)
> result

        Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 1958, df = 91, p-value < 2.2e-16
```

귀무가설 : 두 그룹의 Covariance Matrix가 동일하다.

p-value가 매우 작기 때문에 귀무가설을 기각한다.
-> 유의미한 차이를 보인다.

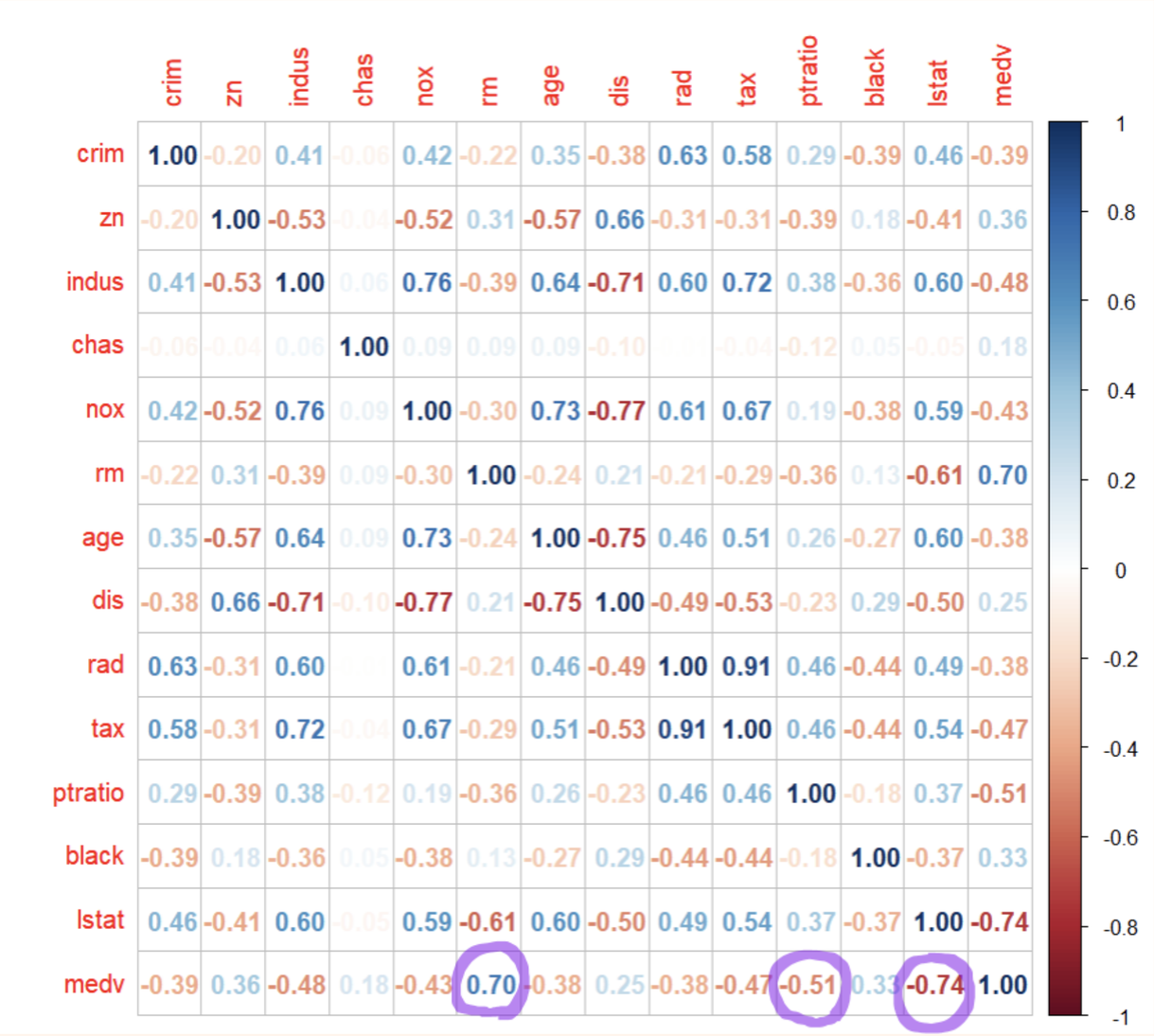
데이터탐색

변수간 상관관계 분석
: 상관행렬

빈곤층 비율(lstat) 0.74, 평균 방 갯수(rm) 0.70,
학생과교사비율(ptratio) -0.51

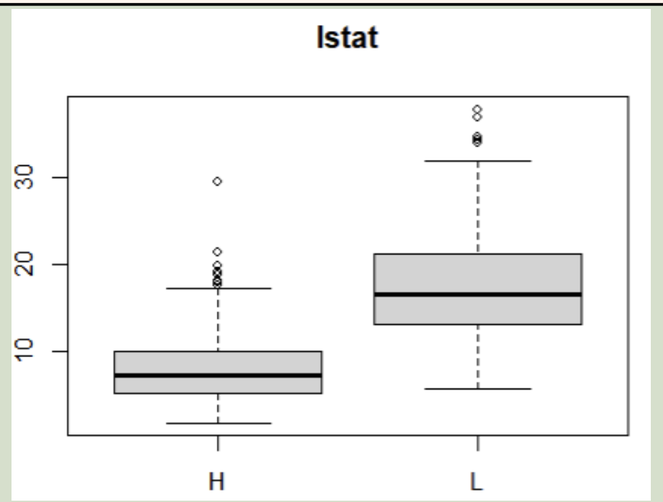
위 3개 변수가 집값에
영향을 많이 준다는 것을 확인할 수 있다.

```r
#상관행렬을 이용하여 변수간 상관관계 분석
dat_real <- read.csv("BostonHousing.csv", header = TRUE)
dat_cor <- round(cor(dat_real),2)
dat_cor
#히트맵
library(corrplot)
corrplot(dat_cor)
#숫자형태로
corrplot(dat_cor, method="number")
```



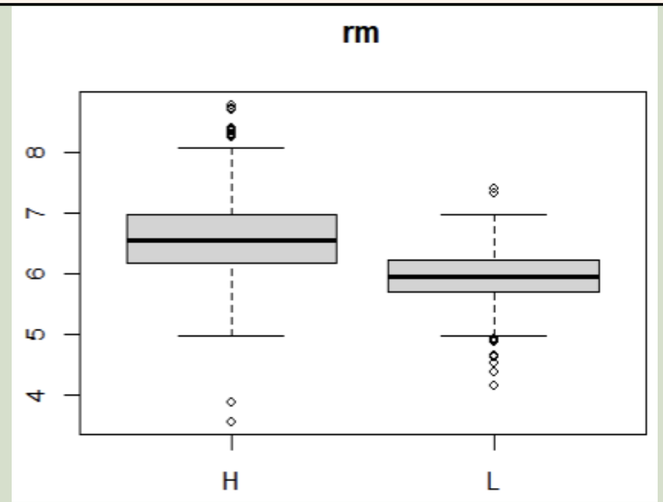
데이터 탐색

H, L 각 그룹별 관측값 분포 확인



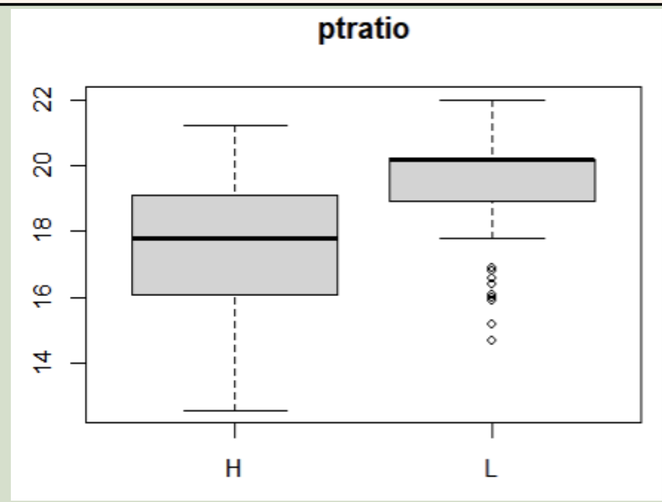
lstat

주택 가격이 낮을수록
빈곤층의 비율도 높다는 것을
알 수 있다.



rm

주택 가격이 높으면 방의 개수도
많다는 것을 알 수 있다.



ptratio

주택 가격이 높을수록
교사/학생의 비율이 낮은 것을
알 수 있다.

앞의 상관행렬에서 가장 수치가 높았던 3가지 변수들의 그룹별 분포를 확인해보았다.

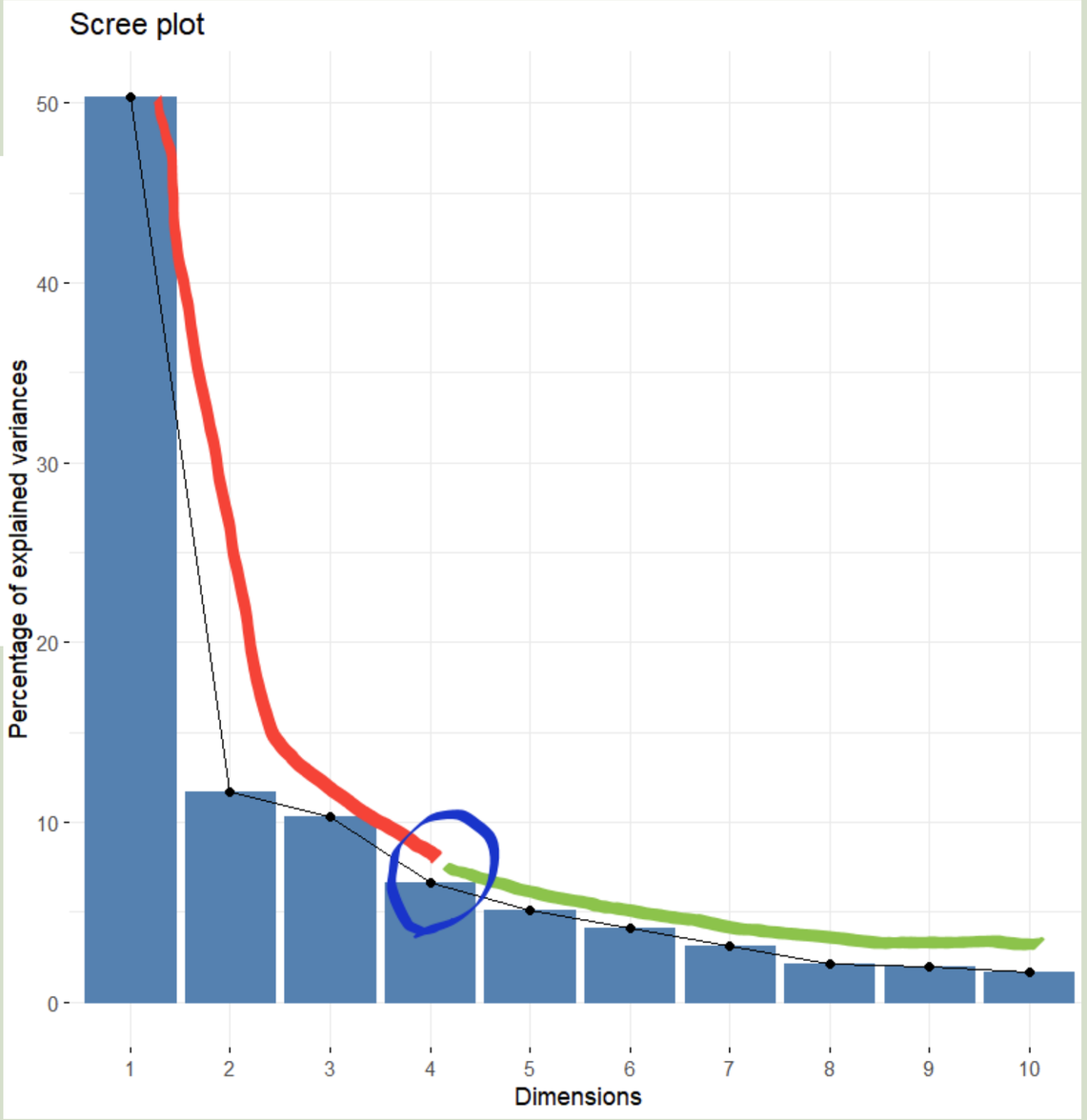
주성분분석(PCA)

```
> summary(fit_pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation     2.5584859 1.2339618 1.1557640 0.92951801 0.81654857
Proportion of Variance 0.5035269 0.1171278 0.1027531 0.06646183 0.05128858
Cumulative Proportion  0.5035269 0.6206547 0.7234078 0.78986967 0.84115825
                          Comp.6    Comp.7    Comp.8    Comp.9
Standard deviation     0.73311449 0.6353263 0.52678619 0.50343341
Proportion of Variance 0.04134284 0.0310492 0.02134644 0.01949578
Cumulative Proportion  0.88250109 0.9135503 0.93489672 0.95439251
                          Comp.10    Comp.11    Comp.12     Comp.13
Standard deviation     0.46136928 0.42809414 0.36875173 0.246563099
Proportion of Variance 0.01637397 0.01409728 0.01045983 0.004676412
Cumulative Proportion  0.97076648 0.98486375 0.99532359 1.000000000
```

- 주성분 개수를 결정하는 기준

1. 누적기여율 - PC5(Comp.5)
2. 분산 수치 - PC3(Comp.3)
3. Scree plot - PC4(Comp.4)

-> **PC4(Comp.4)**까지 활용하는 것이 적당하다고 판단



주성분분석

- 영향을 받는 변수

PC1 : indus, nox, tax, dis, rad, age, lstat

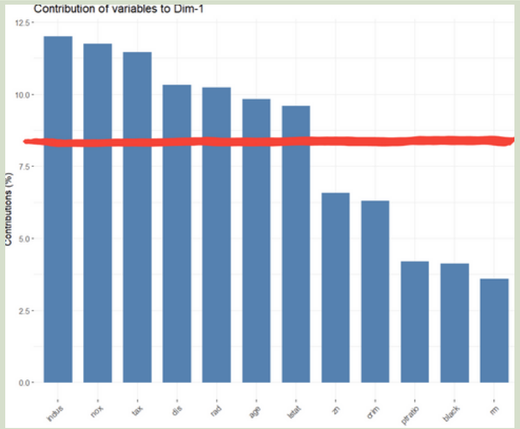
PC2 : rm, medv, dis, ptratio

PC3 : zn, rad, crim, black, tax

PC4 : ptratio, black, zn, lstat

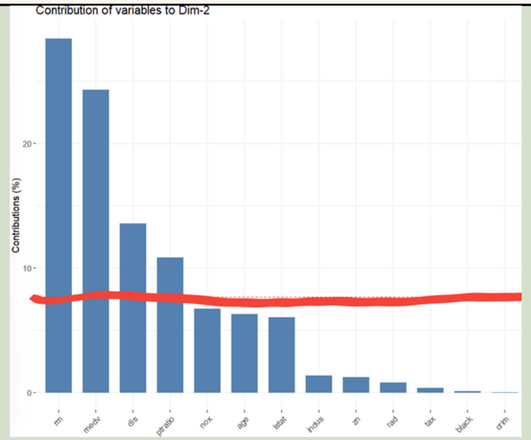
```
Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
crim     0.242          0.409          0.213  0.778  0.164  0.255
zn      -0.245  0.112  0.434  0.301  0.361 -0.270 -0.381  0.388 -0.235
indus    0.332 -0.116                        -0.341  0.170  0.622  0.265
nox      0.325 -0.259         0.193  0.140 -0.189                0.215
rm      -0.203 -0.533  0.248 -0.185 -0.168         -0.443         0.527
age      0.297 -0.250 -0.258                 0.131 -0.589        -0.248
dis     -0.298  0.368  0.240                -0.115 -0.124 -0.176  0.281
rad      0.303         0.414 -0.213  0.155 -0.139        -0.460 -0.129
tax      0.324         0.341 -0.144  0.204 -0.309        -0.182
ptratio  0.208  0.329        -0.704 -0.251         -0.276  0.281 -0.161
black   -0.197        -0.363 -0.401  0.791                       0.149
lstat    0.311  0.246 -0.113  0.288                -0.355 -0.167
medv    -0.266 -0.493        -0.143                 0.155        -0.578
        Comp.10 Comp.11 Comp.12 Comp.13
crim
zn       0.131  -0.225  -0.130
indus   -0.276   0.348          -0.232
nox      0.436  -0.439   0.531
rm      -0.227  -0.124
age      0.329   0.485
dis      0.106   0.507   0.554
rad                      -0.635
tax              0.164  -0.255   0.697
ptratio  0.100  -0.228   0.195
black
lstat   -0.684  -0.180   0.251
medv    -0.239          0.453   0.145
```

주성분분석(PCA)



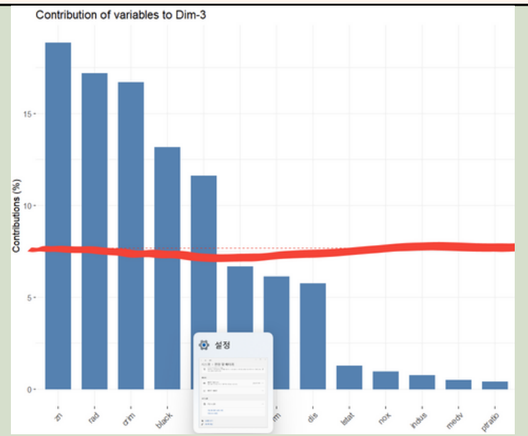
PC1

영향을 받는 변수 :
indus, nox, tax, dis, rad, age, lstat



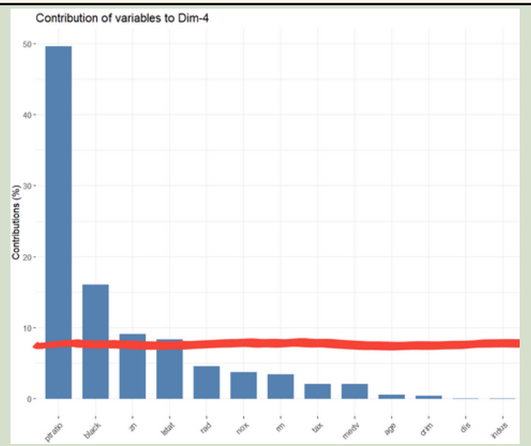
PC2

영향을 받는 변수 :
rm, medv, dis, ptratio



PC3

영향을 받는 변수 :
zn, rad, crim, black, tax



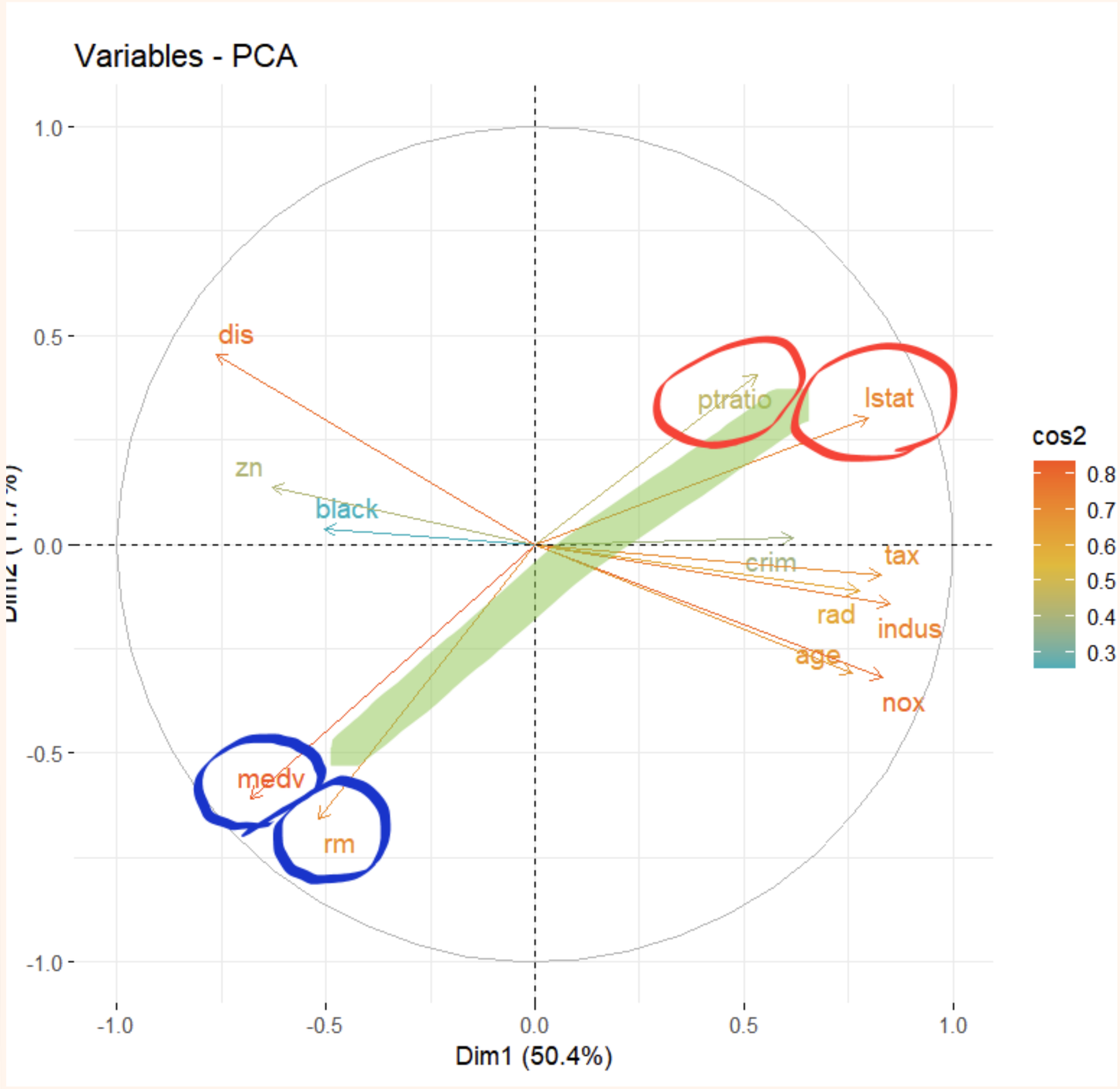
PC4

영향을 받는 변수 :
ptratio, black, zn, lstat

주성분분석

medv, rm
dis, zn, black
lstat, ptratio, crim
nox, indus, age, tax, rad

더 많은 방(rm)이 있는 더 비싼 주택(medv)은 그래프의 왼쪽 하단 모서리에 위치하는 반면, 빈곤층비율(lstat), 교사/학생비율(ptratio)은 반대쪽에 유사하게 위치한다.



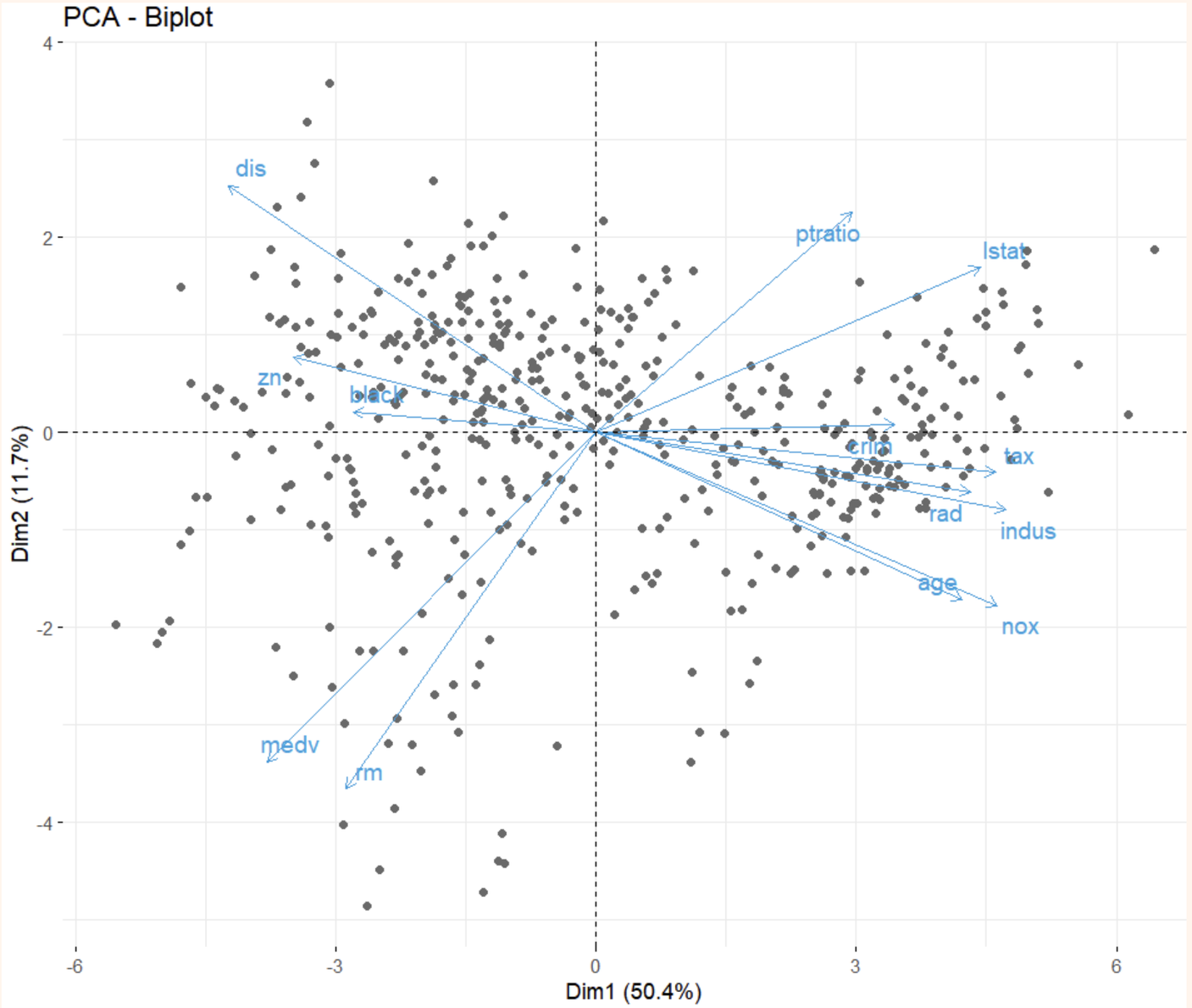
주성분분석

medv, rm

dis, zn, black

lstat, ptratio, crim

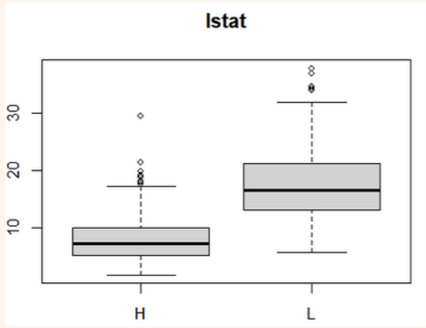
nox, indus, age, tax, rad

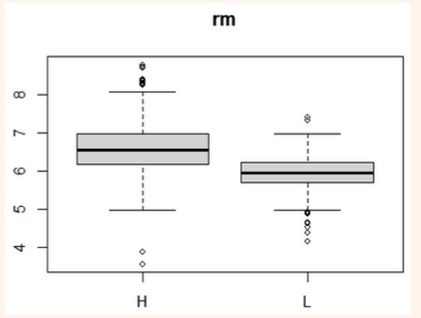


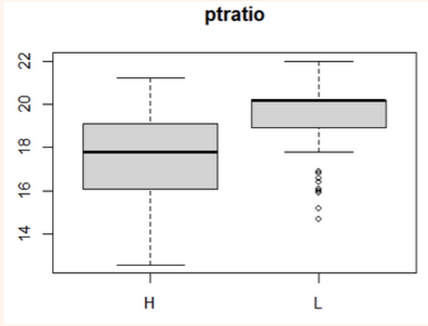
결과 해석

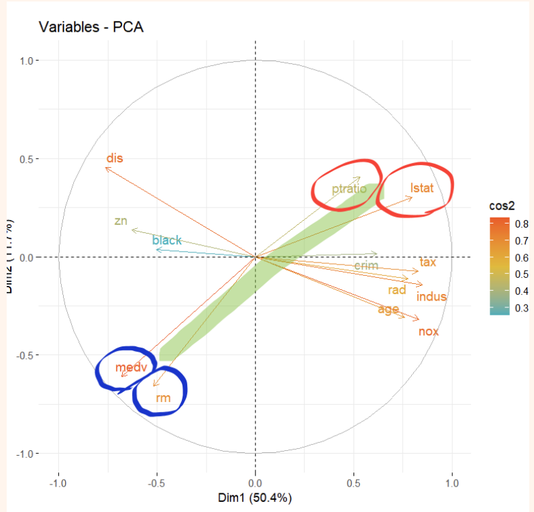
상관행렬에서도 구해보았듯이,
빈곤층 비율(lstat), 평균 방 갯수(rm), 학생과교사비율(ptratio)
이 3개 변수가 집값에 영향을 많이 준다는 것을 확인할 수 있었는데,
주성분 분석에서도
더 많은 방(rm)이 있는 더 비싼 주택(medv)은 그래프의
왼쪽 하단 모서리에 위치하는 반면,
빈곤층비율(lstat), 교사/학생비율(ptratio)은
반대쪽에 거의 대칭적으로 유사하게 위치하는 모습을 볼 수 있었다.

따라서 주택가격에는
빈곤층 비율(lstat), 평균 방 갯수(rm), 학생과교사비율(ptratio)이
집 값(medv)에 영향을 많이 준다고 결론을 내릴 수 있다.









Thank you

감 사 합 니 다