2019 학년도 1학기
DATA MINING
HW2

과목명	데이터마이닝
담당교수명	송종우 교수님
제출일	2019.03.27
학번	182STG27
이름	임지연

Ⅰ. Description

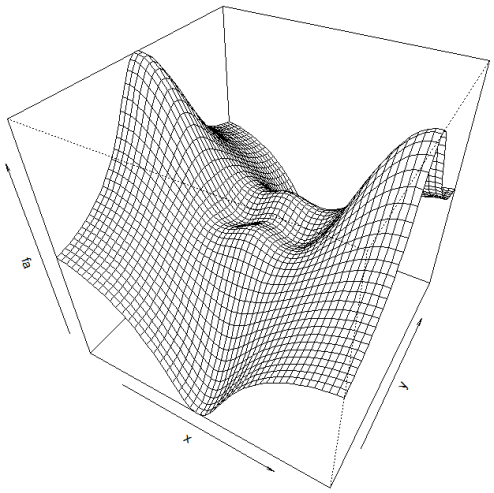
모델링을 하기 전 자료를 탐색하는 것은 매우 중요한 과정이다. 이번 과제는 총 2가지로 chapter 2.3의 R의 기본적인 vector, matrix 연산과 mtcars data set 을 이용하여 다양한 그래프를 그려보는 것과 데이터 요약하는 법에 대해 실습을 통해 공부를 한다. 예제 9, 10 번 문제를 풀며 Auto, Boston 데이터셋에 대해 EDA 및 기술통계량 산출에 대해 공부해본다.

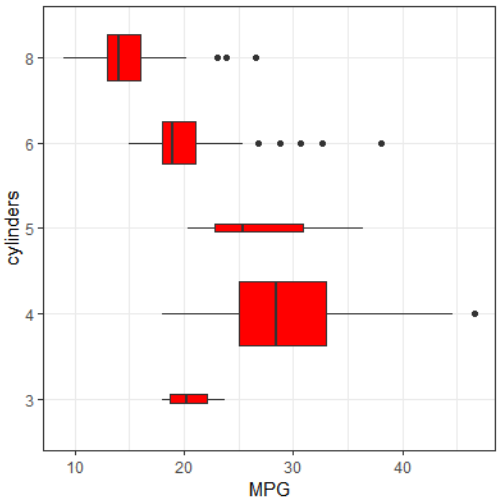
Ⅱ. Implementation

Question 1.

Lab : Introduction to R 의 코드를 실행해보고 감상문을 써라

R의 기초적인 matrix , plot 에 대해 복습할 수 있었다. Plot 함수보다는 ggplot 에 익숙하기 때문에 ggplot 을 이용하여 바꿔그려보며 공부했다. 또한 Boxplot 그래프를 그릴 때 항상 세로 방향으로만 그렸는데 이번 과제를 하며 가로 방향으로 그려보았다. 만약 x축 값의 각각의 level값의 이름이 긴 값을 가질 때 세로 방향으로 그래프를 그리면 글씨가 겹치는 경우가 생겼고 글씨를 기울여 나타냈는데 이렇게 가로방향으로 그리면 그 문제가 해결되며 더욱 효과적인 시각화가 가능할 것이라는 것을 알게 되었다. 더하여 pdf file 만들기, contour 그래프, fix 함수에 대해서 새롭게 알게 되었다. 코드를 실행해 본 그래프는 아래와 같다.





Question 2.

2.4 Excercises - Example 9,10 풀어라

[Example 9]

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

변수명	형식	구분
mpg	num	quantitative
cylinders	num	quantitative
displacement	num	quantitative
horseposer	num	quantitative
weight	num	quantitative
acceleration	num	quantitative
year	num	quantitative
origin	num	quantitative
name	Factor	qualitative

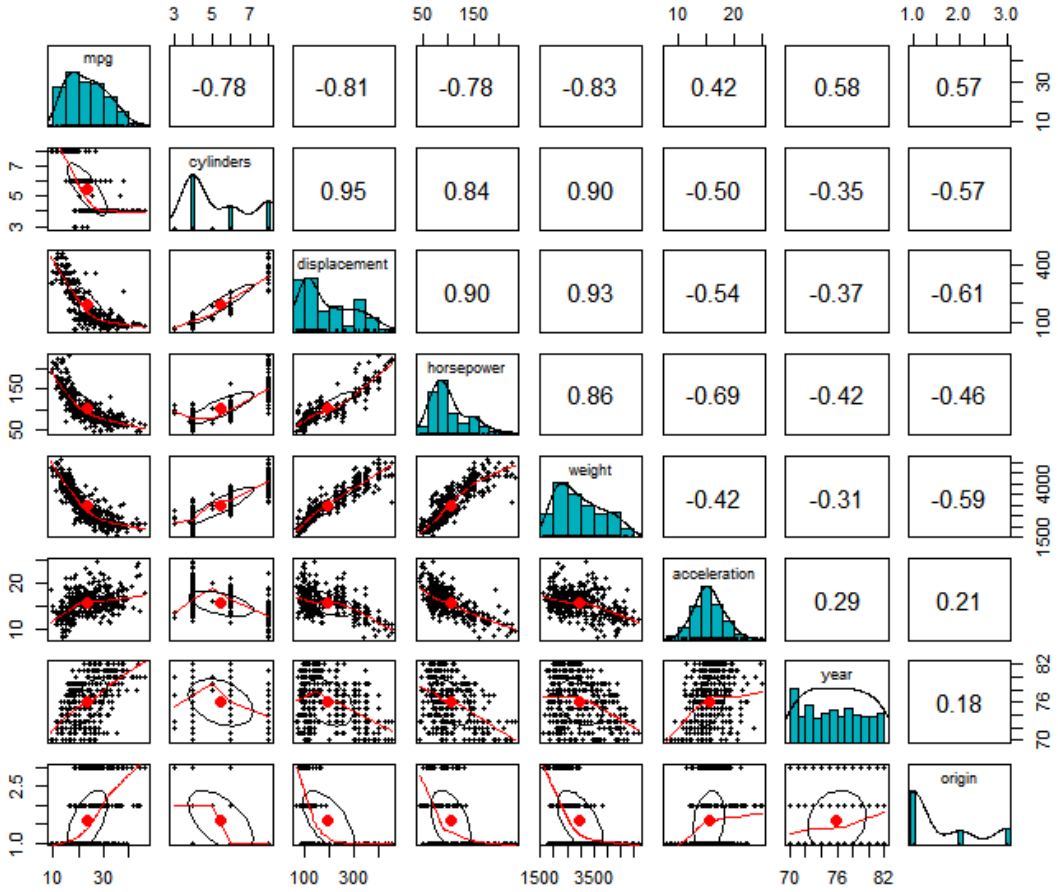
(b), (c) What is the range of each quantitative predictor? You can answer this using the range() function. What is the mean and standard deviation of each quantitative predictor?

변수명	min	max	range	mean	sd
mpg	9	46.6	37.6	23.45	7.81
cylinders	3	8	5	5.47	1.71
displacement	68	455	387	194.41	104.64
horsepower	46	230	184	104.47	38.49
weight	1613	5140	3527	2977.58	849.4
acceleration	8	24.8	16.8	15.54	2.76
year	70	82	12	75.98	3.68
origin	1	3	2	1.58	0.81

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

변수명	min	max	range	mean	sd
mpg	11	46.6	35.6	24.4	7.87
cylinders	3	8	5	5.37	1.65
displacement	68	455	387	187.24	99.68
horsepower	46	230	184	100.72	35.71
weight	1649	4997	3348	2935.97	811.3
acceleration	8.5	24.8	16.3	15.73	2.69
year	70	82	12	77.15	3.11
origin	1	3	2	1.6	0.82

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.



- Mpg : cylinder, displacement, horsepower, weight 와 높은 음의 상관관계가 나타나며, year가 증가할수록 큰 값을 갖는 경향이 나타난다.
- Horsepower, Displacement, Weight : pearson 상관계수 값이 매우 높은 것으로 보아 양의 상관관계가 있음을 알 수 있다.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

- e에서 scatter plot 을 그려본 결과 cylinder, displacement, horsepower, weight, year 변수 등 서로 관계가 있어 보인다는 것을 알 수 있었다.

[Example 10]

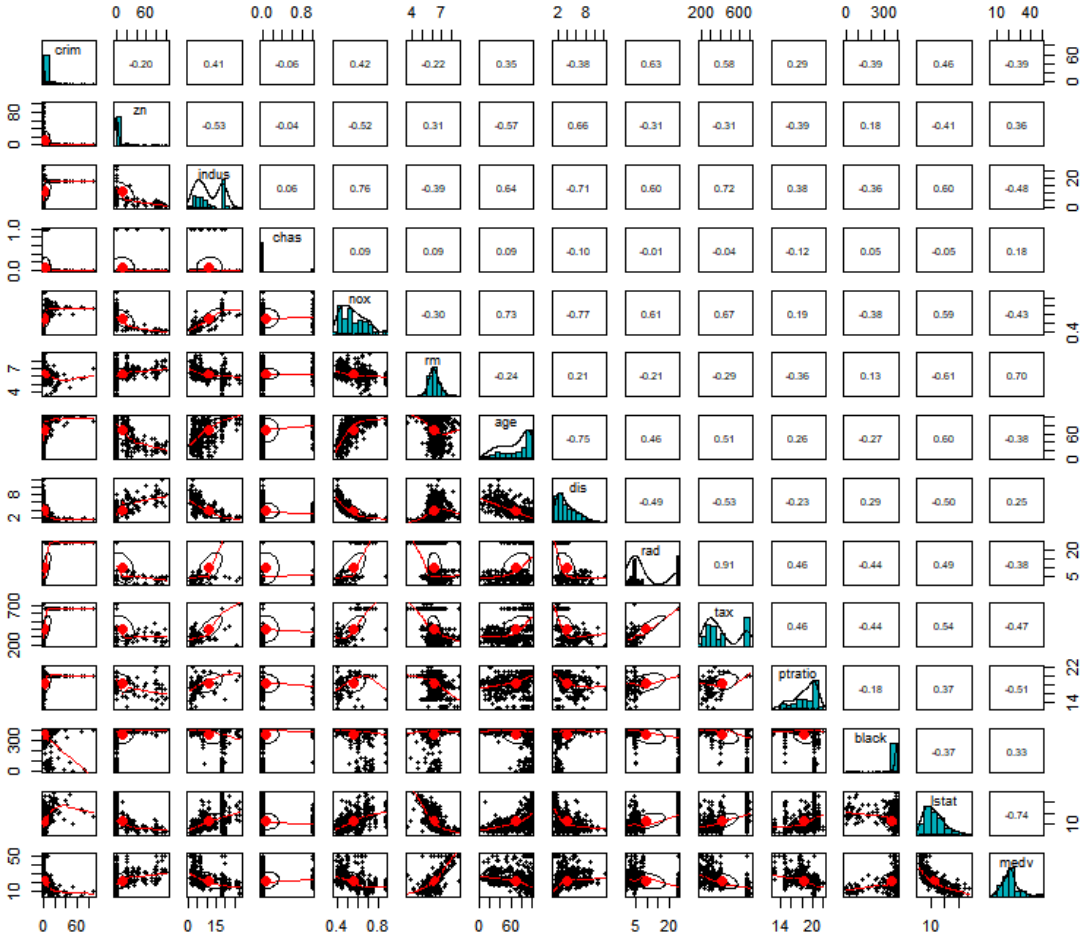
This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R. How many rows are in this data set? How many columns? What do the rows and columns represent?

- Boston 데이터셋은 506개의 행과 14개의 열로 이루어져 있다.

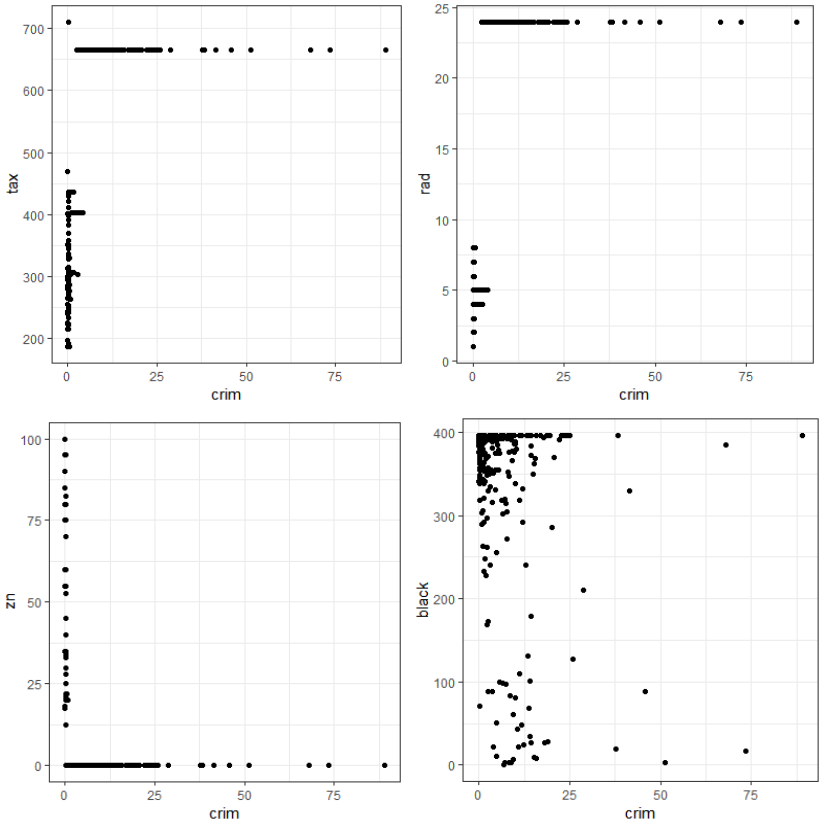
crim	per capita crime rate by town.
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town.
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox	nitrogen oxides concentration (parts per 10 million).
rm	average number of rooms per dwelling.
age	proportion of owner-occupied units built prior to 1940.
dis	weighted mean of distances to five Boston employment centres.
rad	index of accessibility to radial highways.
tax	full-value property-tax rate per ₩$10,000.
ptratio	pupil-teacher ratio by town.
black	*1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town.
lstat	lower status of the population (percent).
medv	median value of owner-occupied homes in ₩$1000s.

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.



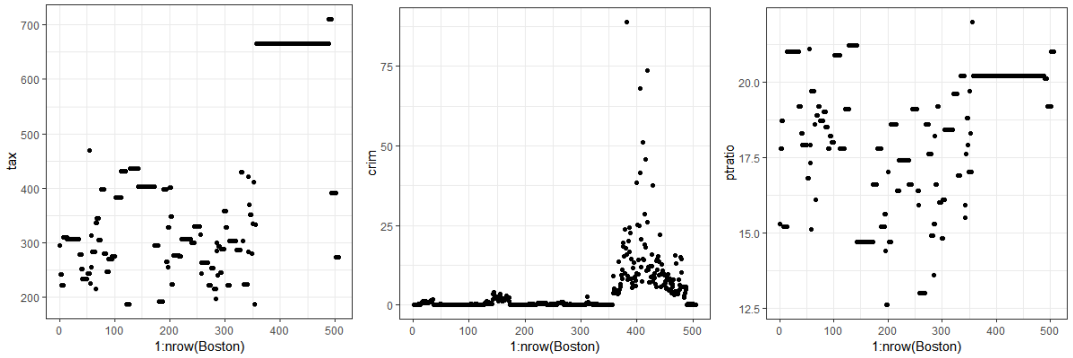
- Boston 데이터 셋에서 관심있게 살펴볼 변수인 crim 과 다른 변수와의 관계를 살펴보면 그래프의 형태나 상관계수 값에 대해 별다른 상관관계가 없어 보인다.
- 다만 rad(index of accessibility to radial highways) 변수와 tax(full-value property-tax rate per ₩$10,000) 변수의 상관계수가 0.91로 가장 높은 양의 상관관계를 가진다.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.



- 특정 지역 내에서 범죄율이 급상승하고 있는 경향을 보이고 있으며 구체적으로, Tax 변수는 600 ~ 700 사이의 값, Rad 변수는 20 ~ 25 사이의 값, Zn 변수는 0 근처에서, Black 변수는 0 값과 400 값 주변에서 범죄율이 높다는 것을 알 수 있다.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



- 관측치마다의 tax, crim, ptratio 에 대한 그래프를 그려봤을 때 특정 지역에서 tax, crim 값이 높은 것이 관측되지만, pupil- teacher ratio 에 대해서는 대체로 고르게 분포되어있다.

(e) How many of the suburbs in this data set bound the Charles river?

- Charles river 와 연결되어 있는 교외지역은 약 35 개이다.

(f) What is the median pupil-teacher ratio among the towns in this data set?

- 각 도시당 학생- 교사비율의 중앙값은 값은 약 19로 나타난다. 따라서 교사 1명 당 약 19명의 학생이 있다는 것을 알 수 있다.

(g) Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

[전체 변수 값 요약]

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
min	0	0	0	0	0	4	3	1	1	187	13	0	2	5
max	89	100	28	1	1	9	100	12	24	711	22	397	38	50
range	89	100	27	1	0	5	97	11	23	524	9	397	36	45
mean	4	11	11	0	1	6	69	4	10	408	18	357	13	23
sd	9	23	7	0	0	1	28	2	9	169	2	91	7	9

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
지역 1	38	0	18	0	1	5	100	1	24	666	20	397	31	5
지역 2	68	0	18	0	1	6	100	1	24	666	20	385	23	5

- 주거지 소유 주택의 median 값을 의미하는 medv 변수의 최소값은 5로 나타난다.
- 위 표를 살펴보면 최소값을 갖는 지역1, 지역2의 age, rad 값이 최대값을 갖는 것으로 나타나며 zn, rm. Dis 값은 대체로 작은 값을 갖는 것을 알 수 있다.

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

- 평균 7 개 이상의 방을 갖는 지역은 총 64, 평균 8 개 이상의 방을 갖는 주거지역은 총 13곳이다.
- 이 13 지역의 값을 살펴보면 lstat 변수의 값이 작은 값을 가지며, medv 값은 큰 값을 가진다.

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0	0	3	0	0	8	76	3	2	276	18	397	4	39
2	2	0	20	1	1	8	94	2	5	403	15	388	3	50
3	0	95	3	0	0	8	32	5	4	224	15	391	3	50
4	0	0	6	0	1	8	78	3	8	307	17	385	4	45
5	1	0	6	0	1	9	83	3	8	307	17	382	5	50
6	0	0	6	0	1	8	86	3	8	307	17	387	3	38
7	1	0	6	0	1	8	73	4	8	307	17	386	2	42
8	0	0	6	0	1	8	70	4	8	307	17	379	4	48
9	0	22	6	0	0	8	8	9	7	330	19	397	4	43
10	1	20	4	0	1	9	87	2	5	264	13	390	5	50
11	1	20	4	0	1	8	92	2	5	264	13	387	6	49
12	1	20	4	0	1	8	67	2	5	264	13	385	7	50
13	3	0	18	1	1	9	83	2	24	666	20	355	5	22

Ⅲ. Discussion

이번 과제를 통해 R 에서 주로 data frame, tibble 형식, ggplot 으로 데이터를 다루고 그래프를 그려는데 matrix 의 연산, plot 함수에 대해 복습할 수 있었고 새로운 함수도 새로 사용해 보았다.

예제 문제를 풀며 데이터 모델링 전 데이터 핸들링, 시각화에 대한 중요성을 알 수 있었다. 데이터 요약, 시각화 자료로부터 중요한 정보를 얻을 수 있었기 때문인데, 먼저 각각의 변수별로 전체적인 그래프를 살펴보며 어떤 특징이 발견되면 좀 더 구체적으로 살펴보면서 의미를 탐색하면 효율적인 데이터 탐색이 가능할 것이다. 새로운 모델을 만들기 전 전처리 및 시각화의 중요성을 잊지 않아야 겠다.

Ⅳ. Appendix – R code

9 번문제

```
#a
str(Auto)
write.csv(str(Auto), "C:/Users/jeeyeon/Desktop/데마/HW3/Ex9a.csv")

#b #c
library(dplyr)
mynum = select_if(Auto, is.numeric)
summary(mynum)
b = data.frame( apply(mynum, 2, range))
b[3,] = b[2,] - b[1,]
#write.csv( b, "C:/Users/jeeyeon/Desktop/데마/HW3/Ex9b.csv")
c1 = apply(mynum, 2, mean)
c2 = apply(mynum, 2, sd)
write.csv(round( rbind(b,c1,c2),2),
          "C:/Users/jeeyeon/Desktop/데마/HW3/Ex9bc.csv")

#d
Auto
fix(Auto)
quantile(Auto$mpg)
tmp = mynum[-(10:85),]   # drop rows

d = data.frame( apply(tmp, 2, range))
d[3,] = d[2,] - d[1,]
t1 = apply(tmp, 2, mean)
t2 = apply(tmp, 2, sd)
```

```
    write.csv(round( rbind(d,t1,t2),2),
              "C:/Users/jeeyeon/Desktop/데마/HW3/Ex9d.csv")
    #e
    pairs(mynum)

    library(psych)
    pairs.panels(mynum,
                 method = "pearson", # correlation method
                 hist.col = "#00AFBB",
                 density = TRUE,    # show density plots
                 ellipses = TRUE # show correlation ellipses
    )

### 10 번문제
    library(tidyverse)
    library(MASS)
    Boston
    ?Boston
    summary(Boston)

    #a
    dim(Boston)
    #b
    pairs(Boston)
    pairs.panels(Boston,
                 method = "pearson", # correlation method
                 hist.col = "#00AFBB",
                 density = TRUE,    # show density plots
                 ellipses = TRUE # show correlation ellipses
    )

    names(Boston)

    #c
    cor(Boston)[1,]
    ggplot(Boston, aes(x = crim, y = zn)) + geom_point() +theme_bw()
    ggplot(Boston, aes(x = crim, y = rad)) + geom_point() +theme_bw()
    ggplot(Boston, aes(x = crim, y = tax)) + geom_point() +theme_bw()
    ggplot(Boston, aes(x = crim, y = black)) + geom_point() +theme_bw()

    #d
    ggplot(Boston, aes(x=1:nrow(Boston), y=crim)) + geom_point()+theme_bw()
    ggplot(Boston, aes(x=1:nrow(Boston), y=tax)) + geom_point()+theme_bw()
    ggplot(Boston, aes(x=1:nrow(Boston), y=ptratio))+ geom_point()+theme_bw()
```

```
#e
sum(Boston$chas)

#f
median(Boston$ptratio)

#g
apply(Boston, 2, range)
b = data.frame( apply(Boston, 2, range))
b[3,] = b[2,] - b[1,]
c1 = apply(Boston, 2, mean)
c2 = apply(Boston, 2, sd)
write.csv(round( rbind(b,c1,c2),0),
          "C:/Users/jeeyeon/Desktop/데마/HW3/Ex10.csv")
Boston %>% filter(medv == min(medv))
write.csv(round( Boston %>% filter(medv == min(medv)),0),
          "C:/Users/jeeyeon/Desktop/데마/HW3/Ex10g.csv")
apply(Boston %>% filter(medv == min(medv)), 2, range)

#h

Boston %>% filter(rm > 7)
nrow(Boston %>% filter(rm > 7) )
Boston %>% filter(rm > 8)
nrow(Boston %>% filter(rm > 8) )
write.csv(round( Boston %>% filter(rm > 8) ,0),
          "C:/Users/jeeyeon/Desktop/데마/HW3/Ex10h.csv")
```