Data Mining

HW#2

학번	182STG18
이름	이하경
제출일	2019.03.27

Description

Regression & Classification

일반적으로 모형에 포함된 변수의 형태는 정량변수와 정성변수 중 하나로 나눌 수 있다. 특히 반응변수 Y 가 수치형으로 크기의 비교가 가능하며 연속적인 값 또는 수량 등의 형태를 가지는 정량변수(=양적변수, Quantitative)일 때 단순 또는 다중 선형 회귀모형을 적합하여 해당 숫자를 예측하는 모형을 만들 수 있다. 반면 반응변수가 양적으로 비교할 수 없는 질적 정보를 담고 있는 정성변수(=질적변수, Qualitative)일 때는 관측치별 반응변수의 Class(Category)를 예측하고자 분류 모형을 이용한다. 회귀분석에서 실제 값과 예측 값의 차이를 이용하여 MSE 를 계산하는 것처럼 분류 모형에서는 예측된 Class Label 과 실제 Label 을 비교하여 정확도 또는 오분류율*을 계산하여 예측력을 평가할 수 있다.

$$* \ \mathrm{Error}_{\mathrm{Test}} = \mathrm{Ave}_{i \in \mathrm{Test}} I[y_i \neq \hat{C}(x_i)]$$

만약 반응변수에 K 개의 Class 가 존재할 경우, 설명변수 X 가 x 로 주어졌을 때 Class k 에 속할 확률은 다음과 같이 조건부 확률로 표현할 수 있다.

$$\mathrm{p_k}(x) = \Pr(Y = k | X = x), \qquad k = 1, 2, \ldots, K$$

설명변수에 기반한 Bayes Optimal Classifier 는 $\mathrm{p_j}(x) = \max\{\mathrm{p}_1(x), \ldots, p_K(x)\}$, 즉 K 개의 Class 로 분류될 확률을 각각 계산하고 가장 큰 값을 찾고 해당 Class j 로 Label 을 분류한다.

HW1 에서 중점적으로 다루었던 K-Nearest Neighbor 추정 방법은 반응변수가 수치형일 때 설명변수 값 X=x 주변의 k 개의 관측치들을 이용해 반응변수의 기댓값을 추정하였다. 만약 K-NN 을 분류모형에서 이용할 경우 주어진 X=x 주변의 k 개의 가까운 점들에서 Label 의 Frequency 를 계산하여 가장 빈도가 높은 Label 로 분류하는 'Voting' 형식이 된다. 정량변수 예측모형과 마찬가지로 설명변수의 차원이 커질 경우 K-NN 을 이용한 모형은 불안정해지며 예측된 Class 의 정확도가 떨어진다.

Results

2.3 Lab: Introduction to R (Comment)

ISLR 의 2.3 에는 R 에서 기본적으로 사용하는 몇 가지 함수 및 명령어가 소개되어 있다. 가장 기초적인 Vector, Matrix 등의 형태를 다루는 명령어부터 실제 데이터를 테이블로 불러와 그래프 및 요약 통계량을 탐색하는 명령어까지 R 에서 직접 실습해보았다. 실습을 통해 평소 R 을 사용하며 필수적으로 사용하는 함수 및 명령어들을 복습할 수 있었고, 새롭게 알게 되거나 잘 사용하지 않아 생소했던 몇 가지 함수와 기능들을 정리하게 되는 계기가 되었다.

- **pdf(), jpeg() 또는 png()**
  R 에서 그리는 plot 을 pdf, jpeg, png 의 파일 형태로 저장하는 함수로 **dev.off()**을 선언하기 이전까지의 plot 들을 파일로 저장할 수 있다.
- **contour(), persp()**
  3 차원(x, y 와 각각의 (x, y)에 대한 z 좌표)로 이루어진 데이터에 대해 평면 또는 입체 형태의 등고선을 그리는 함수이다.
- **fix()**
  데이터 프레임을 별도의 창에서 스프레드시트의 형태로 확인하고 직접 수정할 수 있다.
- **identify()**
  출력한 plot 에서 선언 후 클릭한 point 에 대해 변수의 값을 plot 위에 나타낼 수 있다.

Exercise 2.9 Auto Data

(a) Quantitative & Qualitative Predictors

변수	의미	타입
mpg (response)	miles per gallon	Quantitative
cylinders	Number of cylinders	Quantitative (or Qualitative)
displacement	Engine displacement	Quantitative
horsepower	Engine horsepower	Quantitative
weight	Vehicle weight	Quantitative
acceleration	Time to accelerate from 0 to 60 mph	Quantitative
year	Model year	Quantitative
origin	Origin of car (1=American, 2=European, 3=Japanese)	Qualitative
name	Vehicle name	Qualitative

(b) Range & (c) Mean, SD of each Quantitative predictor

Variable	Lower Range	Upper Range	Mean	SD
cylinders	3.0	8.0	5.47	1.71
displacement	68.0	455.0	194.41	104.64
horsepower	46.0	230.0	104.47	38.49
weight	1613.0	5140.0	2977.58	849.40
acceleration	8.0	24.8	15.54	2.76
year	70.0	82.0	75.98	3.68

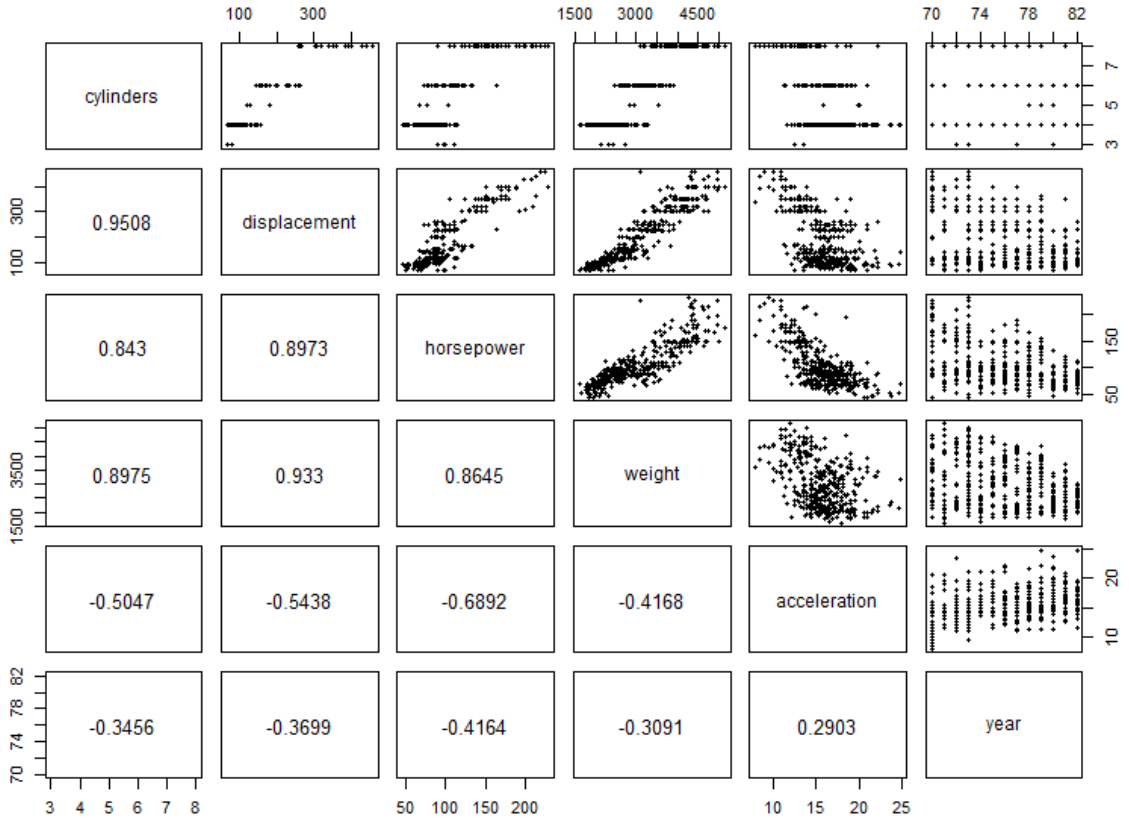
(d) Range & Mean, SD in the Subset (Removed 10th-85th Observations)

Variable	Lower Range	Upper Range	Mean	SD
cylinders	3.0	8.0	5.37	1.65
displacement	68.0	455.0	187.24	99.68
horsepower	46.0	230.0	100.72	35.71
weight	**1649.0**	**4997.0**	2935.97	811.30
acceleration	**8.5**	24.8	15.73	2.69
year	70.0	82.0	77.15	3.11

- weight의 범위가 기존에 비해 좁아지고 편차가 줄어들었다. 제외된 관측치들의 weight가 평균에서 멀리 떨어진 곳에 위치하였음을 알 수 있다. 또한 평균은 감소하였다.
- acceleration의 최소값과 평균이 증가하였으므로 제외된 관측치들의 acceleration이 작은 구간에 위치하였음을 알 수 있다.

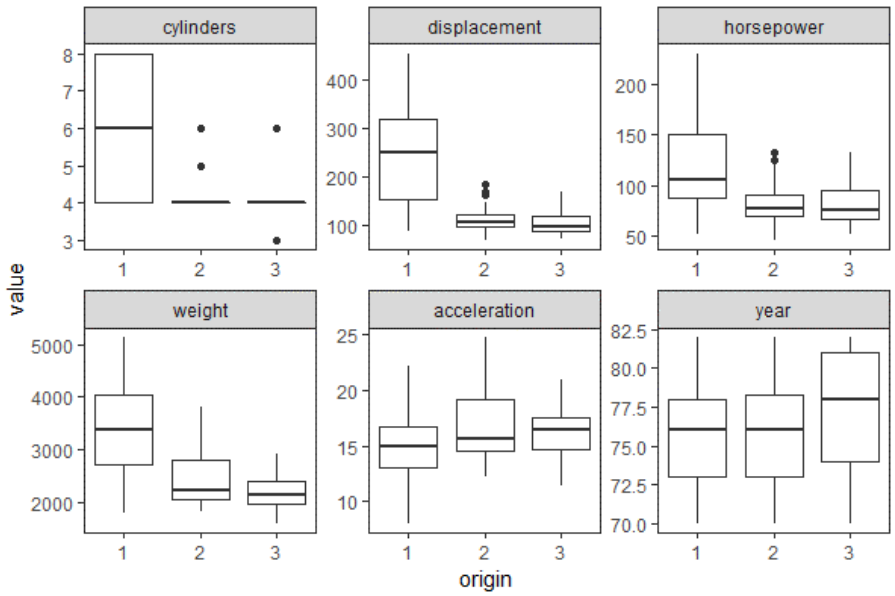
(e) Graphical Investigation of the predictors

▷ **Pairwise Scatterplots of Quantitative Predictors**



대부분의 연속형 설명변수들 사이에 선형관계가 있음을 확인하였다. 특히 cylinders, displacement, horsepower, weight 사이의 상관계수는 모두 0.8 이상으로 매우 높으며 산점도에서도 눈에 띄게 직선의 형태를 보인다. acceleration, year는 이 네 가지 변수와 음의 상관관계가 있으나 year의 경우 다른 변수들 사이의 관계보다 정도가 약하다.

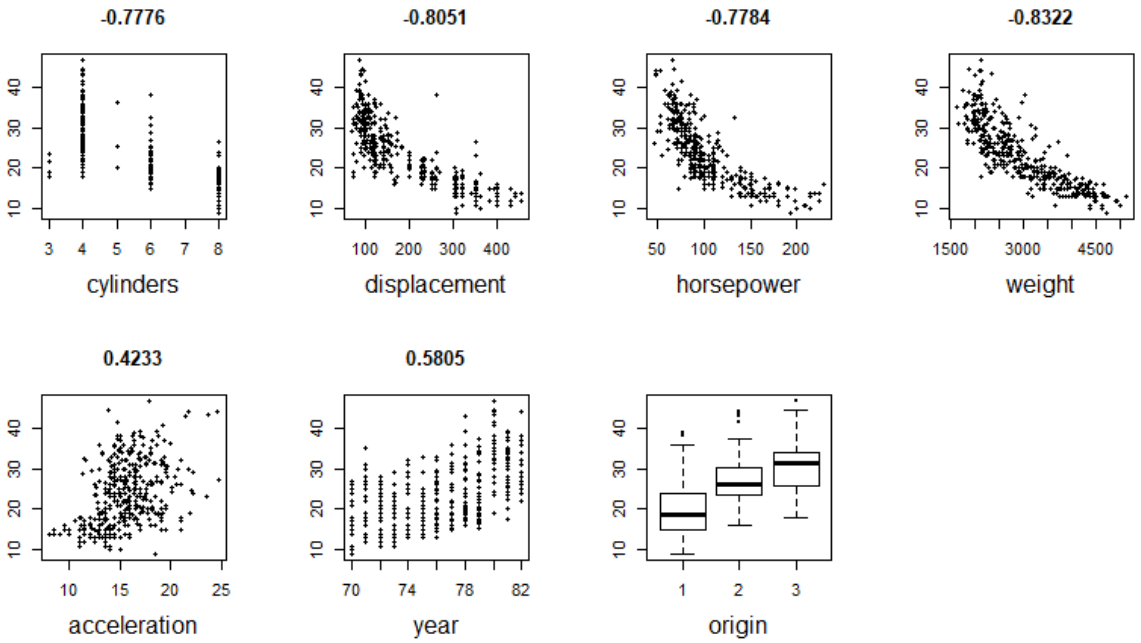
▷ Boxplots of 'origin' vs Quantitative Predictors



범주형 변수인 origin (출시국가)에 따른 다른 양적 설명변수들의 값의 차이를 확인하기 위한 Boxplot이다. acceleration을 제외하고는 origin의 그룹간 각 설명변수들의 차이가 확연히 나타났다. cylinders부터 weight까지 4개의 설명변수에 대해서는 origin이 1(American) 인 자동차들이 2(European)과 3(Japanese)의 자동차들에 비해 높은 값을 가진다. Japanese 자동차들의 경우 출시년도(year)가 다른 두 국가에 비해 늦은 편이다.

(f) Plots & Interpretation to predict 'mpg' based on other Predictors

▷ Scatterplot (or Boxplot) of mpg vs Predictors



반응변수 mpg와 다른 설명변수들의 산점도 또는 Boxplot을 그리고 양적 설명변수들에 대해서는 상관계수를 구해보았다. cylinders, displacement, horsepower, weight는 모두 값이 커질수록 mpg가 작아지며, acceleration의 경우 상대적으로 미미하지만 year와 함께 mpg와 양의 상관관계가 있음을 확인하였다. orgin이 1인 그룹의 mpg 분포에 비해 2와 3그룹으로 갈수록 분포가 높은 곳에서 형성되어 있다. 따라서 모든 그래프를 통해 각 설명변수가 mpg를 예측하는 데 유용할 것이라고 예상하였다.

Exercise 2.10 Boston Housing Data

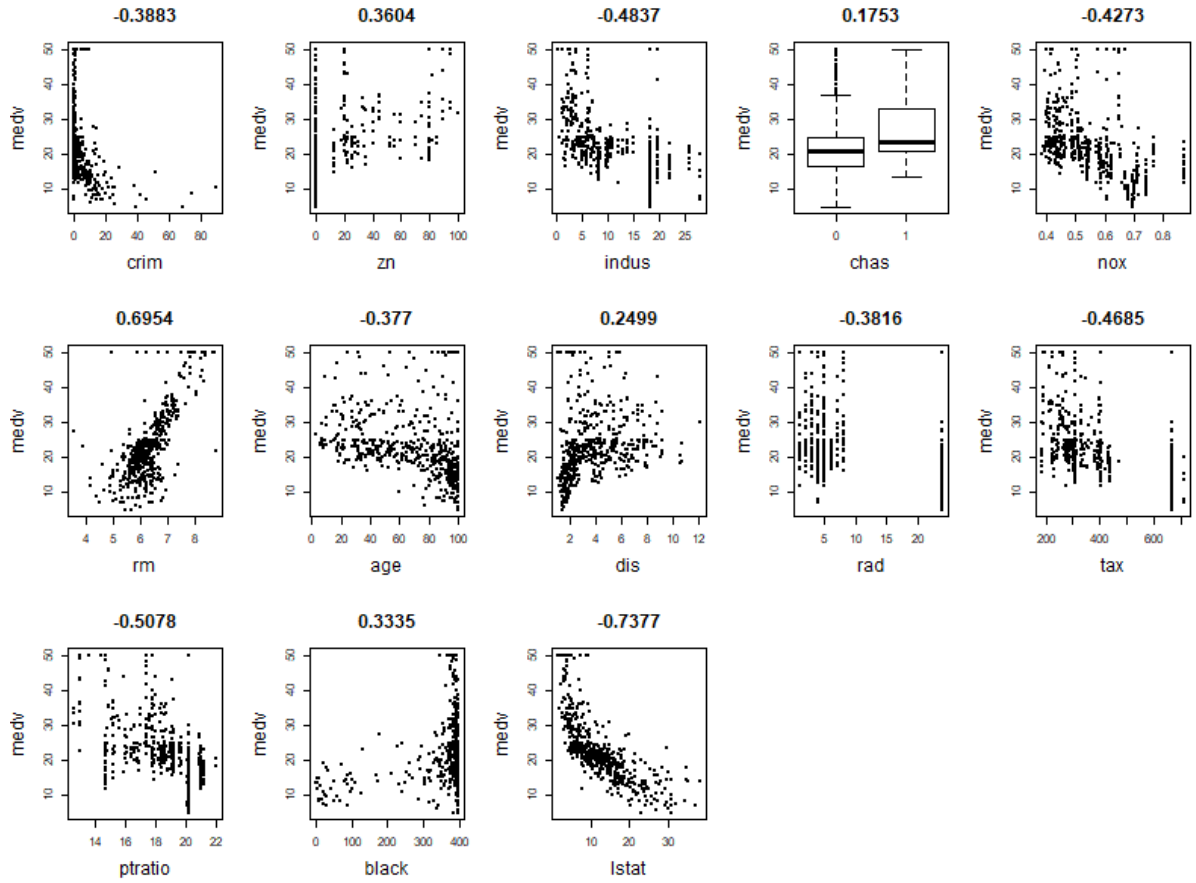
(a) Dimension of Boston Data Set & Meaning

- 데이터는 총 506개의 행과 14개의 열을 포함하고 있다. 각각의 행은 14개 변수에 대한 개별 도시의 관측치를 나타내며 각각의 열은 14개의 변수를 각각 나타낸다. 다음은 14개 변수의 의미와 변수의 타입을 나타낸 것이다. 찰스 강 경계 위치 여부를 나타내는 chas 변수를 제외하고는 반응변수와 모든 설명변수가 양적 정보를 담고 있는 정량변수이다.

변수	의미	타입
crim	범죄율	Quantitative
zn	25,000 평방피트를 초과하는 거주지역의 비율	Quantitative
indus	비소매상업지역이 점유하는 토지의 비율	Quantitative
chas	찰스 강의 경계 위치 여부 (1=True, 0=False)	Qualitative (Categorical)
nox	10ppm당 일산화질소 농도	Quantitative
rm	주택 1가구당 평균 방 수	Quantitative
age	1940년 이전에 건축된 소유주택의 비율	Quantitative
dis	보스턴 5대 직업센터까지 거리들의 가중평균	Quantitative
rad	방사형 고속도로까지의 접근성 지수	Quantitative
tax	$10,000 당 재산세율	Quantitative
ptratio	학생-교사의 비율	Quantitative
black	$1000(Bk - 0.63)^2$ (BK=인구 중 흑인의 비율)	Quantitative
lstat	인구 중 하위계층의 비율	Quantitative
***medv**	소유주택 가격의 중앙값	Quantitative

(b) Pairwise Scatterplots of the Predictors

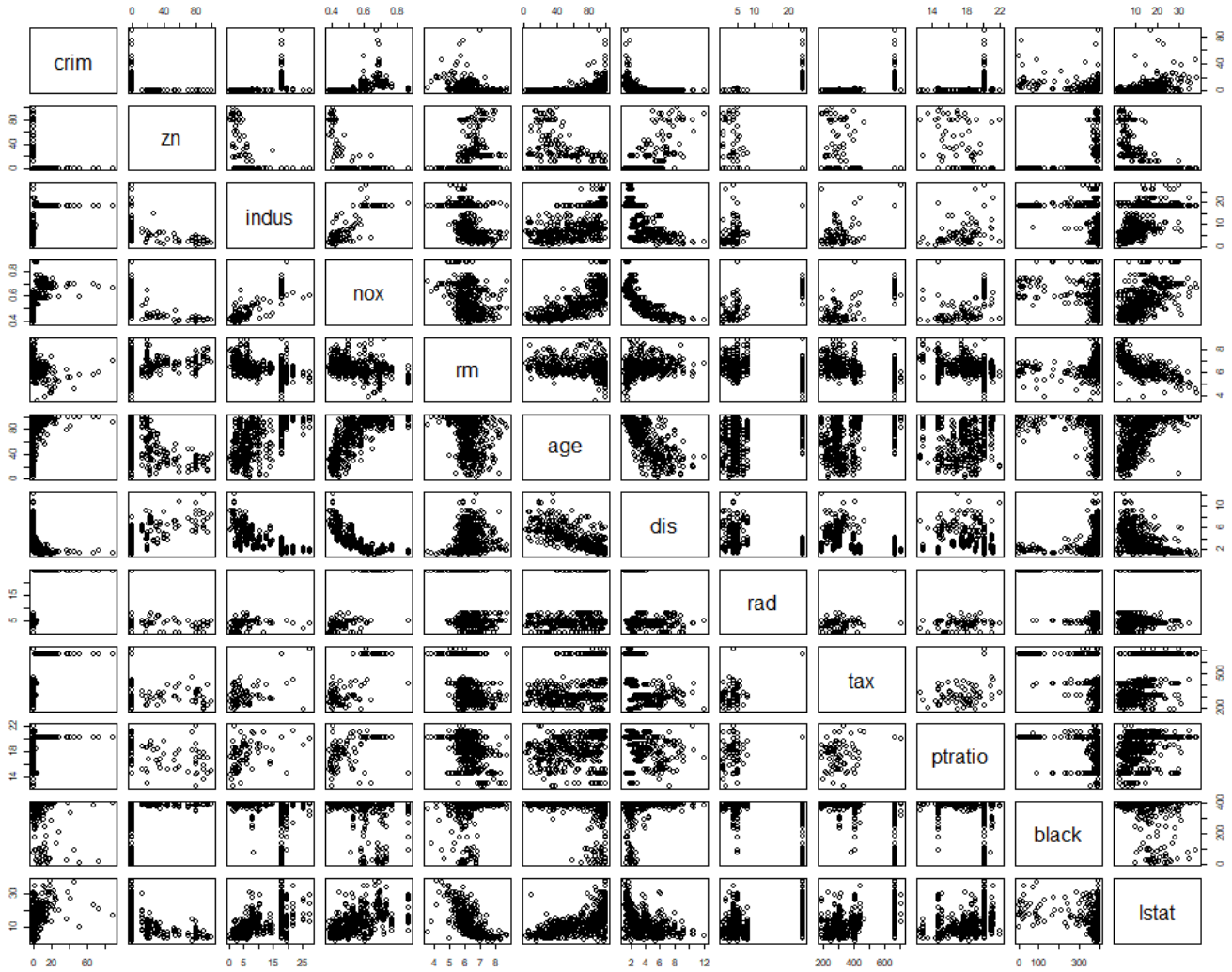
▷ **Response 'medv' vs other Predictors**



일부 변수들에서 반응변수 medv와 비선형 또는 선형 상관관계가 있음을 산점도를 통해 확인하였다. zn, rm, dis, black, lstat은 medv와 양의 상관관계가 있으며 crim, indus, nox, age, rad, tax, ptratio, lstat은 음의 상관관계가 있다. 범주형 변수인 chas의 경우 찰스 강의 경계에 위치한 도시(1)들이 그렇지 않은 도시(0)들에 비해 주택가격이 대체로 더 높다. rm의 경우 상관계수가 0.6954로 설명변수 중 가장 높아 주택 1가구 당 평균 방 수가 많을수록 주택가격이 증가한다는 것을 알 수 있다.

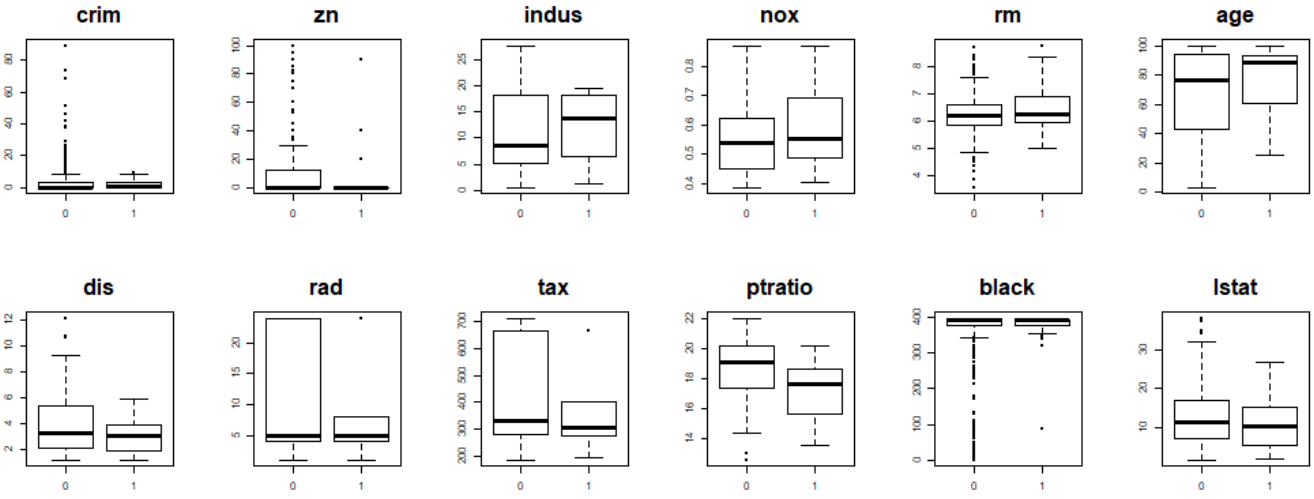
▷ **Pairwise Scatterplots of Quantitative Predictors**

반응변수인 medv 와 범주형 변수 chas 을 제외한 나머지 13 개 설명변수들에 대해 pairwise scatterplot 을 그려보았다.



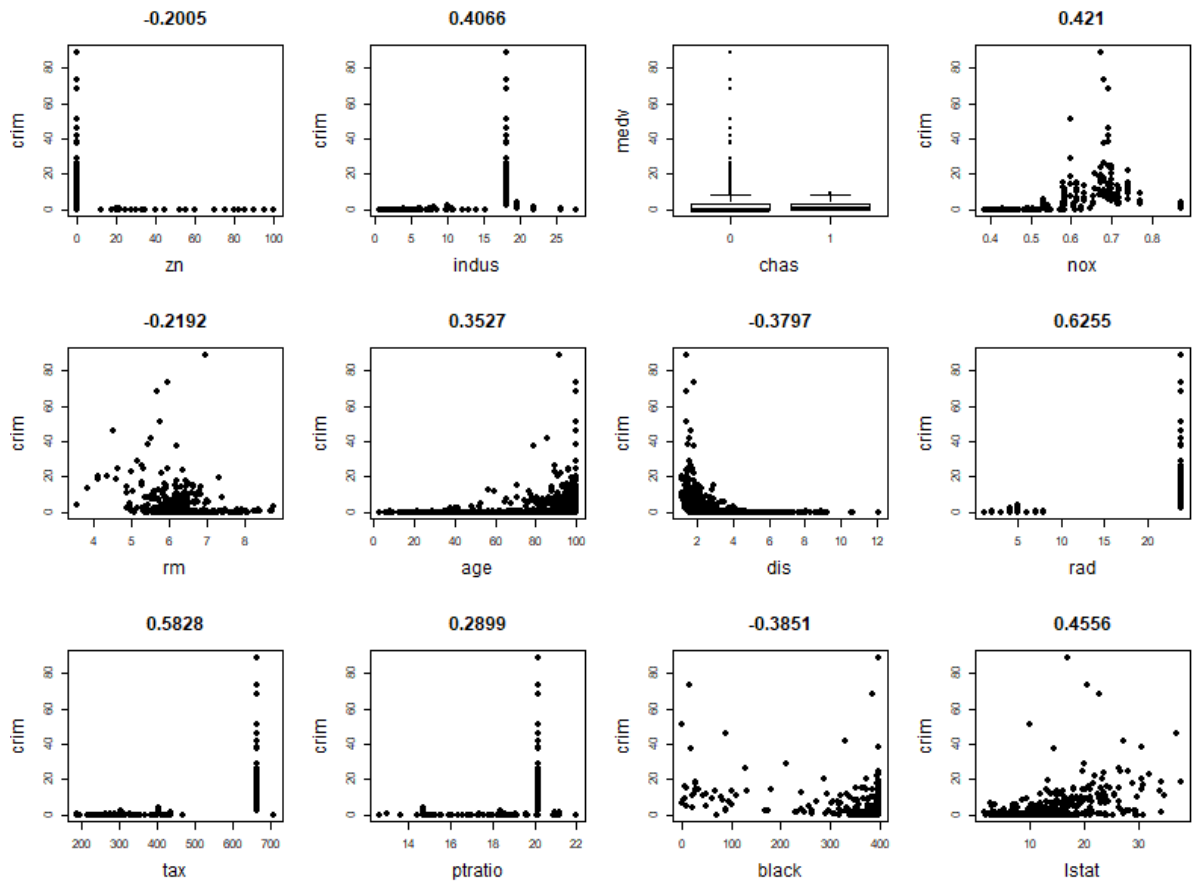
- 다른 변수의 범위 중 범죄율(crim)이 눈에 띄게 높은 특정 구간이 존재한다.
- indus 와 dis, nox 와 dis 사이의 비선형 관계가 존재하는 것으로 보인다.

▷ **Boxplot of 'chas' vs Other Quantitative Predictors**



- indus, nox, age, dis, rad, tax, ptratio 에서 찰스 강 경계 위치 여부에 따른 값의 분포의 차이가 존재하는 것으로 보인다.

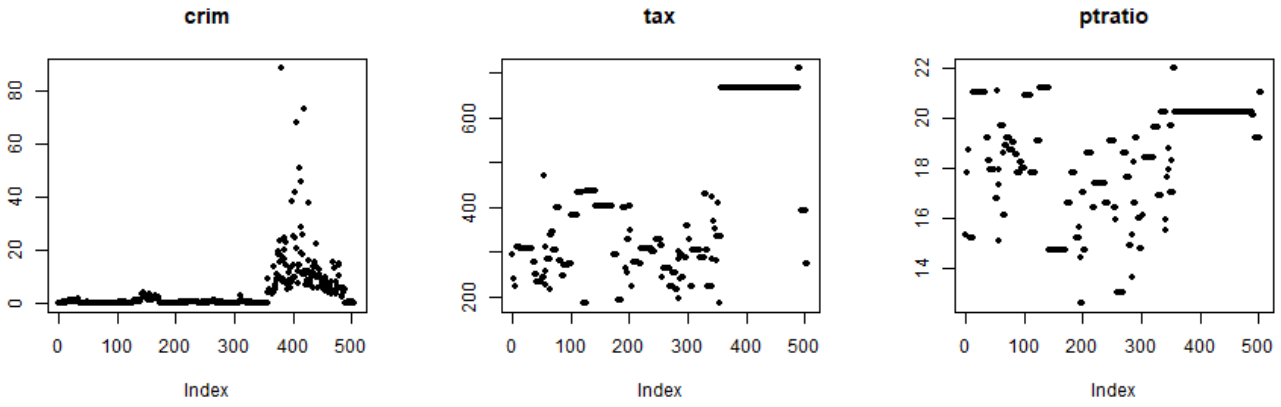
(c) Correlations of Capita Crime Rate and Other Predictors



범죄율과 다른 설명변수와의 관계를 산점도와 상관계수를 통해 확인하였을 때 zn, indus, nox, dis, rad, tax, ptratio에서 값이 극단적으로 높은 일부 관측치들이 존재하는 특정 수준이 있는 것으로 나타났다.

(d) Outliers of Crime Rates, Tax Rates, Pupil-Teacher Ratio

범죄율, 재산세율, 학생-교사비율의 관측치 별 값을 산점도에 나타내었다.



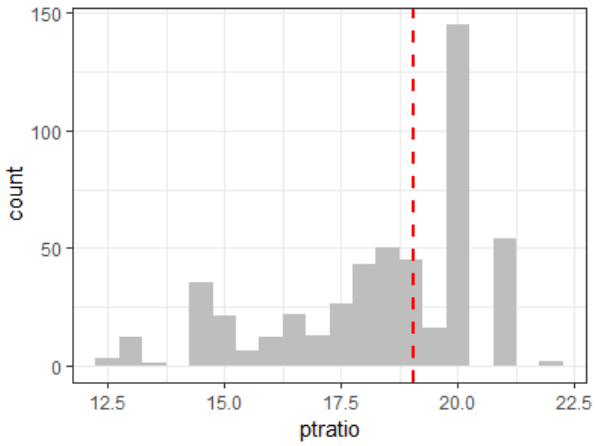
- 약 357~488 번째 행(도시)의 범죄율, 재산세율이 다른 관측치에 비해 특히 높음을 확인하였고, 해당 관측치들의 학생/교사 비율이 모두 666 으로 동일하며 높은 편에 속한다.

(e) Number of suburbs bound the Charles River

0 (False)	1 (Bounds)
471	35

- 35 개의 도시가 찰스 강의 경계에 위치해 있다.

(f) Median Pupil-Teacher Ratio =19.05



- 학생/교사 비율의 중앙값은 19.05 로 히스토그램을 보았을 때 약 20 에서 빈도가 특히 높은 구간이 존재하여 왼쪽으로 꼬리가 긴 분포를 보이고 있다.

Min	Q1	Median	Mean	Q3	Max
12.60	17.40	19.05	18.46	20.20	22.00

(g) Suburb which has lowest median value of owner-occupied homes

id	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.90	30.59	5
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24	666	20.2	384.97	22.98	5

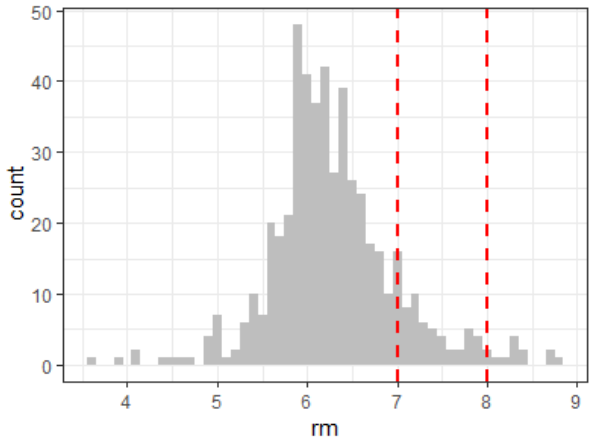
주택 가격의 중앙값이 $5,000 로 관측치들 중 가장 작은 도시는 2 개로 다른 설명변수들의 값은 위와 같다. 두 도시 모두 찰스 강의 경계에 위치하지 않으며, crime, lstat 을 제외한 대부분의 변수 값이 비슷한 경향이 있다.

▷ Range of Predictors

Summary	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
Min	0.0063	0	0.46	0 (471)	0.385	3.561	2.9	1.1296	1	187	12.6	0.32	1.73
Max	88.9762	100	27.74	1 (35)	0.871	8.780	100.0	12.1265	24	711	22.0	396.90	37.97

설명변수의 범위를 확인하여 두 관측치의 변수 값과 비교하였을 때 zn 의 경우 두 관측치 모두 최소값인 0 으로 25,000 피트를 초과하는 거주지역이 존재하지 않으며, age 는 최대값으로 도시 내 주택이 모두 1940 년 이전에 건축되었다. 보스턴 직업센터들까지의 평균 거리는 다른 도시들에 비해 가까운 편이며, 고속도로 접근성 지수는 24 로 최대이다. 재산세율과 학생/교사 비율이 높은 편이며 인구 중 흑인의 비율이 압도적으로 높다.

(h) Number of the Room per Dwelling



	Total	more than 7	more than 8
Nobs	506	64	13

그래프에서도 확인할 수 있듯이 방의 개수가 7 개보다 많은 도시는 64 개, 8 개보다 많은 도시는 13 개로 전체 관측치 506 개에 비하여 매우 적다.

Summary	crim	zn	indus	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
Min	0.01	0.00	0.46	0.38	3.56	2.90	1.13	1.00	187.00	12.60	0.32	1.73	5.00
Q2	0.08	0.00	5.19	0.45	5.89	45.02	2.10	4.00	279.00	17.40	375.38	6.95	17.02
Median	0.26	0.00	9.69	0.54	6.21	77.50	3.21	5.00	330.00	19.05	391.48	11.36	21.20
Mean	3.61	11.36	11.14	0.55	6.28	68.57	3.80	9.55	408.24	18.46	356.67	12.65	22.53
Q3	3.68	12.50	18.10	0.62	6.62	98.07	5.19	24.00	666.00	20.20	396.23	16.96	25.00
Max	88.98	100.00	27.74	0.87	8.78	100.00	12.13	24.00	711.00	22.00	396.90	37.97	50.00
Mean (rm > 8)	0.72	13.62	7.08	0.54	8.35	71.54	3.43	7.46	325.08	16.36	385.21	4.31	44.20

- crim 과 zn 변수는 일부 도시들에서 극단적으로 높은 값이 포함되어 있다. 범죄율의 경우 평균 방 수가 8 개보다 많은 도시들의 평균이 전체 중앙값의 약 3 배에 가깝다. zn 의 경우 상위 50% 중앙값이 0 인 반면 방 수가 8 개보다 많은 도시들의 평균은 월등히 높다. 직업센터와 고속도로까지의 거리가 비교적 가까운 편이며 재산세율과 학생/교사 비율, 흑인의 비율은 중앙값보다 작았다. 인구 중 하위계층의 비율은 전체 중앙값에 비해 상당히 낮다. 이 도시들의 주택 가격(중앙값)은 전체 상위 50%에 비해 대략 2 배 이상 높다.

Discussion

Example 2.9와 2.10에서 실제 데이터에 포함된 변수에 대해 정성변수, 정량변수 별로 반응변수와의 관계 또는 설명변수 사이의 관계 등을 확인하고, 데이터 내 이상치가 있는지 등을 다양한 그래프와 요약 통계량을 통해 탐색하였다. 구체적으로, 정량변수들 사이의 관계는 산점도(scatterplot)과 상관계수를 계산하여 확인하였고 정성변수와 정량변수 사이의 관계는 정성변수의 Class에 따른 정량변수 분포에 차이가 있는지 Boxplot을 통해 확인하였다. 본 과제를 통하여 변수의 유형에 따른 적절한 그래프를 그리는 방법에 대해 더욱 익숙해질 수 있었다.

변수 간 상관관계를 탐색하는 과정은 직접적인 모델링 이전에 반응변수를 설명하는 데에 영향력이 있을 것으로 기대되는 변수들을 선별하고, 데이터의 분포와 동떨어진 극단적인 이상치에 대해서는 원인을 탐색하여 제거 여부를 결정함으로써 예측력 향상을 기대할 수 있다.

[Appendix] R code

2.3 Lab: Introduction to R (생략)

Exercise 2.9

```
# (a)
  str(Auto)

# (b)
  quan <- select(Auto, -mpg, -origin, -name)
  t(apply(quan, 2, range))

# (c)
  cbind(mean = apply(quan, 2, mean), sd = apply(quan, 2, sd)) %>% round(2)

# (d)
  quan.s <- quan[-c(10:85),]
  t(apply(quan.s, 2, range))
  cbind(mean = apply(quan.s, 2, mean), sd = apply(quan.s, 2, sd)) %>% round(2)

# (e)-(f)
  panel.cor <- function(x, y){
    usr <- par('usr')
    on.exit(par(usr))
    par(usr=c(0,1,0,1))
    r <- round(cor(x,y), digits=4)
    text(0.5, 0.5, r, cex=1.2)
  }
  pairs(quan, upper.panel=function(x, y) points(x, y, pch=20), lower.panel=panel.cor)

  Auto %>%
    select(-mpg,-name) %>%
    gather(var, value, -origin) %>%
    mutate(var=factor(var, levels=c('cylinders','displacement','horsepower','weight','acceleration','year'))) %>%
    ggplot() + theme_test() +
    geom_boxplot(aes(x=origin, y=value, group=origin)) +
    facet_wrap(~var, scales='free', ncol=3)
  Auto %>%
    select(-name) %>%
    gather(var, value, -mpg) %>%
    mutate(var=factor(var, levels=colnames(Auto)[2:8])) %>%
    ggplot() + theme_test() +
    geom_point(aes(x=value, y=mpg)) +
    facet_wrap(~var, scales='free', ncol=4)

  cormat1 <- cor(Auto[,1:7])
  ctmp <- cormat1[1,] %>% round(4)
  par(mfrow=c(2,4))
  for (i in 2:8) {
    if (i==8) boxplot(mpg~origin, Auto, xlab='origin', pch=20, cex.lab=1.5)
    else plot(Auto[,i], Auto$mpg, xlab=colnames(Auto)[i], ylab='', main=ctmp[i], pch=20, cex.lab=1.5)
  }
```

Exercise 2.10

```
# (a)
  str(Boston)
  dim(Boston)

# (b)
  png('Pairwise Scatterplot of Predictors.png', width=1000, height=800)
  pairs(select(Boston, -medv, -chas))
  dev.off()

  tmp <- select(Boston, -chas, -medv)
  colnames(tmp)

  png('predictors-chas.png', width=1000, height=400)
  par(mfrow=c(2,6))
  for(i in 1:12) {
  boxplot(tmp[,i]~chas, Boston,
          main=colnames(tmp)[i], pch=20, cex.main=2)
  }
  dev.off()

  cormat <- cor(Boston)
  ctmp <- cormat[14,-14] %>% round(4)

  png('predictors-medv.png', width=800, height=600)
  par(mfrow=c(3,5))
  for (i in 1:13) {
    if(i==4) {
      boxplot(medv~chas, Boston, main=ctmp[i], cex.main=1.5,
              xlab='chas', ylab='medv', pch=20, cex.lab=1.5)
    } else {
      plot(Boston[,i], Boston$medv, pch=20, main=ctmp[i], cex.main=1.5,
           xlab=colnames(Boston)[i], ylab='medv', cex.lab=1.5)
    }
  }
  dev.off()

# (c)
  ctmp <- cormat[1,-1] %>% round(4)
  png('predictors-crim.png', width=800, height=600)
  par(mfrow = c(3, 4))
  for (i in 2:13) {
    if(i==4) {
      boxplot(crim~chas, Boston, main=ctmp[i-1], cex.main=1.5,
              xlab='chas', ylab='medv', pch=19, cex.lab=1.5)
    } else {
      plot(Boston[,i], Boston$crim, pch=19, main=ctmp[i-1], cex.main=1.5,
           xlab=colnames(Boston)[i], ylab='crim', cex.lab=1.5)
    }
  }
  dev.off()
```

```
  tmp <- gather(select(Boston, -medv, -chas), var, value, -crim)
  tmp$var <- factor(tmp$var, levels=colnames(Boston)[-1])
  ggplot(tmp) + theme_test() +
    geom_point(aes(value, crim)) +
    facet_wrap(~var, scales='free', ncol=4)

  ggplot(Boston) + theme_test() +
    geom_boxplot(aes(x = factor(chas), y = medv, group = chas)) + labs(x = 'chas')

  cor(select(Boston, -medv))[1,-1] %>% round(4)

# (d)
  png('index-crim.png', 300, 300) ; plot(Boston$crim, pch=20, main='crim', ylab='') ; dev.off()
  png('index-tax.png', 300, 300) ; plot(Boston$tax, pch=20, main='tax', ylab='') ; dev.off()
  png('index-ptratio.png', 300, 300) ; plot(Boston$ptratio, pch=20, main='ptratio', ylab='') ; dev.off()

  t(apply(select(Boston, crim, tax, ptratio), 2, range))

# (e)
  table(Boston$chas)

# (f)
  median(Boston$ptratio)
  g + geom_histogram(aes(ptratio), fill = 'gray', binwidth = 0.5) +
    geom_vline(aes(xintercept = median(ptratio)), color = 'red', linetype = 2, size = 1)

  hist(Boston$ptratio, xlab='ptratio', breaks=50, col='gray', main='')
  abline(v=median(Boston$ptratio), lty=2)

# (g)
  filter(Boston, medv==min(medv))

# (h)
  nrow(filter(Boston, rm > 7))
  nrow(filter(Boston, rm > 8))
  nrow(filter(Boston, rm <= 8))

  rbind(apply(Boston,2, summary),
        apply(filter(Boston, rm > 8), 2, mean)) %>% round(2)

  g <- ggplot(Boston) + theme_bw()
  g + geom_histogram(aes(rm), binwidth = 0.1, fill = 'gray') +
    geom_vline(aes(xintercept = 7), linetype = 2, color = 'red', size = 1) +
    geom_vline(aes(xintercept = 8), linetype = 2, color = 'red', size = 1)
```