Chapter 2 HW

Zahraa Alshalal

2/16/2023

## Conceptual Questions

### Exercise 2:

#### a). regression. inference. quantitative output of CEO salary based on CEO firm’s features.

#### n - 500 firms in the US

#### p - profit, number of employees, industry

#### (b) classification. prediction. predicting new product’s success or failure.

#### n - 20 similar products previously launched

#### p - price charged, marketing budget, comp. price, ten other variables

#### (c) regression. prediction. quantitative output of % change

#### n - 52 weeks of 2012 weekly data

#### p - % change in US market, % change in British market, % change in German market

### Exercise 4:

### a)

### application #1: disease X has four variances (A, B, C, D), and 10 symptoms.

#### Response: A, B, C, D: the four types of the disease.

#### predictors: The 10 symptoms.

#### The goal of each application prediction.

### application #2: The home mortgage can have two conditions: default and non-default.

#### Response: default/non-default

#### Predictors: income, education, banking balance and previous credits.

#### The goal of each application prediction.

### application #3: stock market price direction

#### Response: up, down.

#### predictors: yesterday’s price movement % change, two previous day price movement % change, etc.

#### The goal of each application prediction.

### b)

### application #1: CEO salary.

#### Response: salary.

#### predictors: age, industry experience, industry, years of education.

#### The goal of each application inference.

### application #2: The next day temperature.

#### Response: the next-day temperature.

#### Predictors: set of current daily weather factors: 5 factors.

#### The goal of each application prediction.

### application #3: The height of the children.

#### Response: the height of the children.

#### predictors: mother’s height, father’s height, daily diet and daily exercise.

#### The goal of each application inference

### b)

### application #1: social class: low-income, middle class and high-income class.

### application #2: Netflix movie recommendations. recommend movies based on users who have watched and rated similar movies.

### application #3: marketing survey. clustering of demographics for a product(s) to see which clusters of consumers buy which products.

## Applied Questions

### Exercise 8:

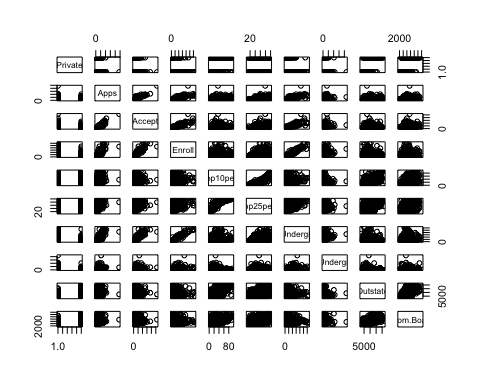
# 8. (a)  
college = read.csv("~/Desktop/spring23/math448/hw/HW#1/College.csv")  
# 8. (b)  
rownames(college) = college[, 1]  
college = college[,-1]   
head(college)

## Private Apps Accept Enroll Top10perc Top25perc  
## Abilene Christian University Yes 1660 1232 721 23 52  
## Adelphi University Yes 2186 1924 512 16 29  
## Adrian College Yes 1428 1097 336 22 50  
## Agnes Scott College Yes 417 349 137 60 89  
## Alaska Pacific University Yes 193 146 55 16 44  
## Albertson College Yes 587 479 158 38 62  
## F.Undergrad P.Undergrad Outstate Room.Board Books  
## Abilene Christian University 2885 537 7440 3300 450  
## Adelphi University 2683 1227 12280 6450 750  
## Adrian College 1036 99 11250 3750 400  
## Agnes Scott College 510 63 12960 5450 450  
## Alaska Pacific University 249 869 7560 4120 800  
## Albertson College 678 41 13500 3335 500  
## Personal PhD Terminal S.F.Ratio perc.alumni Expend  
## Abilene Christian University 2200 70 78 18.1 12 7041  
## Adelphi University 1500 29 30 12.2 16 10527  
## Adrian College 1165 53 66 12.9 30 8735  
## Agnes Scott College 875 92 97 7.7 37 19016  
## Alaska Pacific University 1500 76 72 11.9 2 10922  
## Albertson College 675 67 73 9.4 11 9727  
## Grad.Rate  
## Abilene Christian University 60  
## Adelphi University 56  
## Adrian College 54  
## Agnes Scott College 59  
## Alaska Pacific University 15  
## Albertson College 55

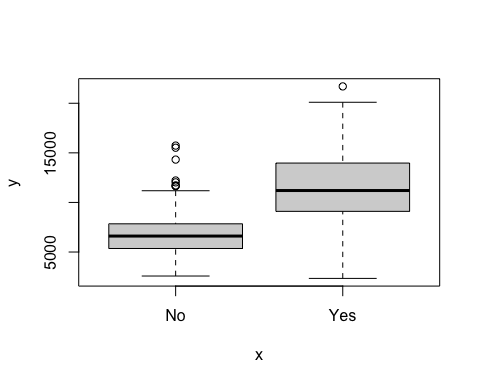
# 8. (c)  
# i.  
summary(college)

## Private Apps Accept Enroll   
## Length:777 Min. : 81 Min. : 72 Min. : 35   
## Class :character 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242   
## Mode :character Median : 1558 Median : 1110 Median : 434   
## Mean : 3002 Mean : 2019 Mean : 780   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902   
## Max. :48094 Max. :26330 Max. :6392   
## Top10perc Top25perc F.Undergrad P.Undergrad   
## Min. : 1.00 Min. : 9.0 Min. : 139 Min. : 1.0   
## 1st Qu.:15.00 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0   
## Median :23.00 Median : 54.0 Median : 1707 Median : 353.0   
## Mean :27.56 Mean : 55.8 Mean : 3700 Mean : 855.3   
## 3rd Qu.:35.00 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0   
## Max. :96.00 Max. :100.0 Max. :31643 Max. :21836.0   
## Outstate Room.Board Books Personal   
## Min. : 2340 Min. :1780 Min. : 96.0 Min. : 250   
## 1st Qu.: 7320 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850   
## Median : 9990 Median :4200 Median : 500.0 Median :1200   
## Mean :10441 Mean :4358 Mean : 549.4 Mean :1341   
## 3rd Qu.:12925 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700   
## Max. :21700 Max. :8124 Max. :2340.0 Max. :6800   
## PhD Terminal S.F.Ratio perc.alumni   
## Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00   
## 1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00   
## Median : 75.00 Median : 82.0 Median :13.60 Median :21.00   
## Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74   
## 3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00   
## Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00   
## Expend Grad.Rate   
## Min. : 3186 Min. : 10.00   
## 1st Qu.: 6751 1st Qu.: 53.00   
## Median : 8377 Median : 65.00   
## Mean : 9660 Mean : 65.46   
## 3rd Qu.:10830 3rd Qu.: 78.00   
## Max. :56233 Max. :118.00

# ii.  
college$Private = as.factor(college$Private)  
pairs(college[, 1:10])



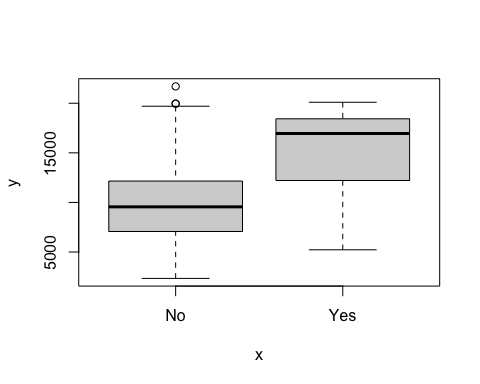
# iii.  
plot(college$Private, college$Outstate)



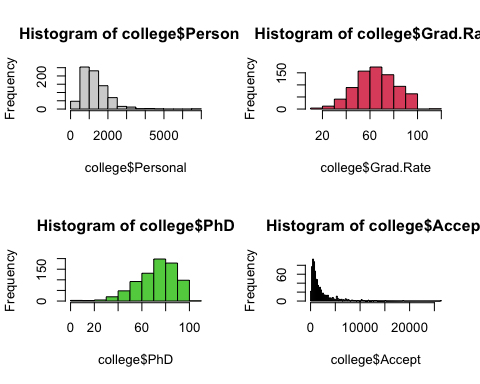
# iv.  
Elite = rep("No", nrow(college))  
Elite[college$Top10perc>50] = "Yes"  
Elite = as.factor(Elite)  
college = data.frame(college, Elite)  
summary(college$Elite)

## No Yes   
## 699 78

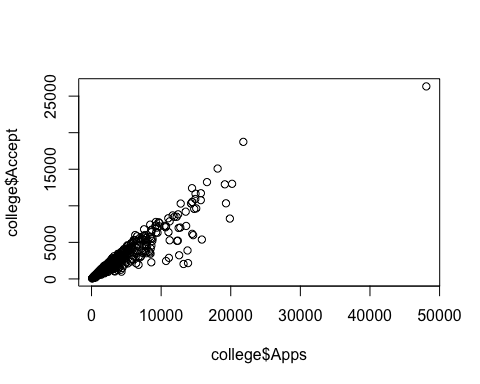
plot(college$Elite, college$Outstate)



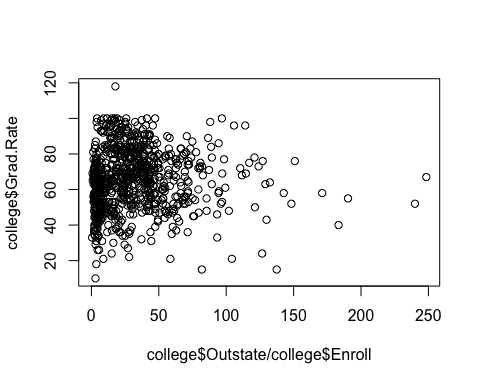
# v.  
par(mfrow=c(2,2))  
hist(college$Personal)  
hist(college$Grad.Rate, col=2)  
hist(college$PhD, col=3, breaks=10)  
hist(college$Accept, breaks=100)



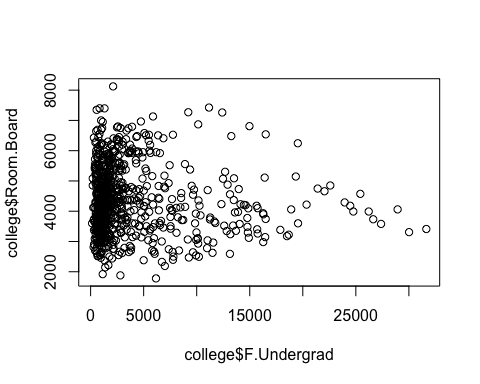
# vi.  
par(mfrow=c(1,1))  
plot(college$Apps, college$Accept)



plot(college$Outstate / college$Enroll, college$Grad.Rate)



plot(college$F.Undergrad, college$Room.Board)



### summary:

#### The relationship between the Number of applications received and the Number of applicants accepted seem to be linear.

#### High tuition correlates to high graduation rate.

### Exercise 10:

# 10.  
# (a)  
library(MASS)  
head(Boston)

## crim zn indus chas nox rm age dis rad tax ptratio black lstat  
## 1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98  
## 2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14  
## 3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03  
## 4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94  
## 5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33  
## 6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21  
## medv  
## 1 24.0  
## 2 21.6  
## 3 34.7  
## 4 33.4  
## 5 36.2  
## 6 28.7

dim(Boston)

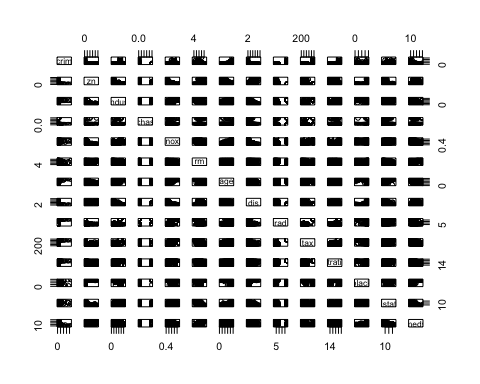
## [1] 506 14

#### 506 rows, 14 columns: 506 housing values in Boston suburbs with 14 features each.

# (b)  
str(Boston)

## '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 : num 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 ...

Boston$chas = as.numeric(Boston$chas)  
Boston$rad = as.numeric(Boston$rad)  
pairs(Boston[, 1:14])

 ##### X correlates with: a, b, c.

# (c)  
summary(Boston$crim)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00632 0.08204 0.25651 3.61352 3.67708 88.97620

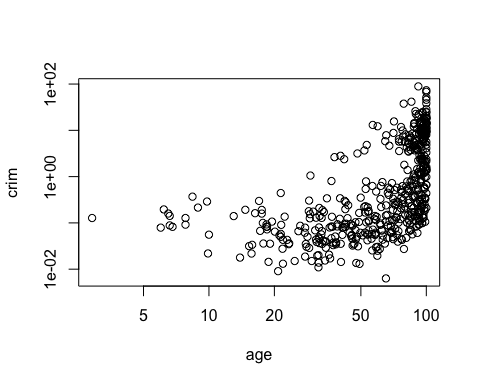
summary(Boston$tax)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 187.0 279.0 330.0 408.2 666.0 711.0

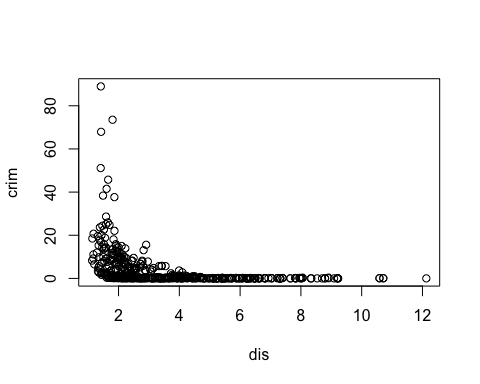
summary(Boston$ptratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.60 17.40 19.05 18.46 20.20 22.00

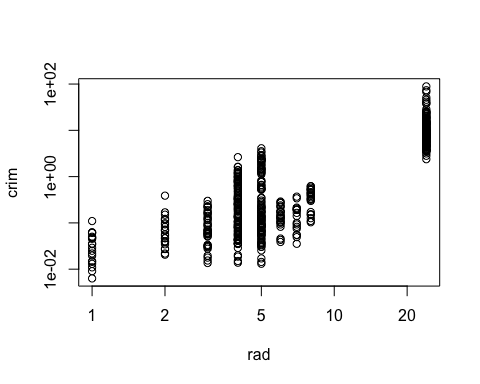
plot(crim ~ age, data = Boston, log = "xy") # Older homes, more crime



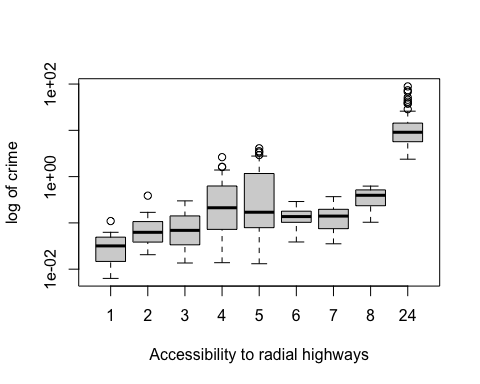
plot(crim ~ dis, data = Boston) # Closer to work-area, more crime



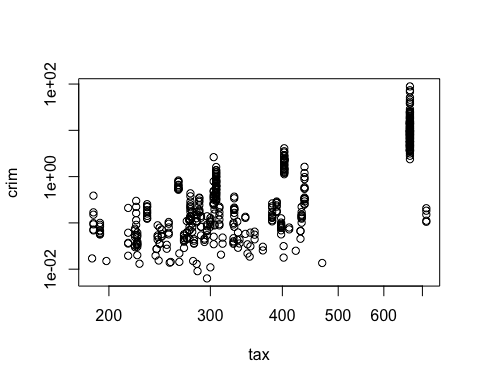
plot(crim ~ rad, data = Boston, log = "xy") # Higher index of accessibility to radial highways, more crime



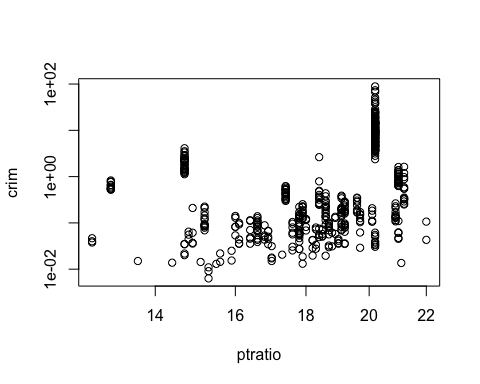
# as box plots, since rad appears to be categorical  
plot(crim ~ as.factor(rad),  
 log = "y",  
 data = Boston,  
 xlab = "Accessibility to radial highways",  
 ylab = "log of crime")



plot(crim ~ tax, log = "xy", data = Boston) # Higher tax rate, more crime



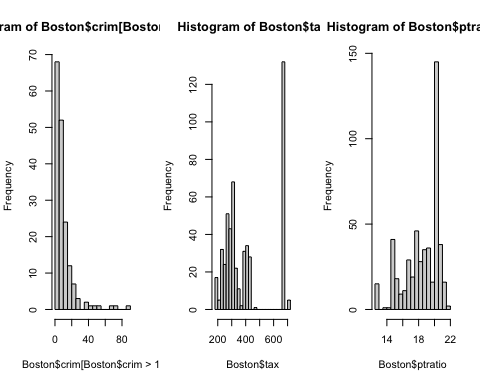
plot(crim ~ ptratio, log = "xy", data = Boston) # Higher pupil:teacher ratio, more crime



#correlations  
cor(Boston)

## crim zn indus chas nox  
## crim 1.00000000 -0.20046922 0.40658341 -0.055891582 0.42097171  
## zn -0.20046922 1.00000000 -0.53382819 -0.042696719 -0.51660371  
## indus 0.40658341 -0.53382819 1.00000000 0.062938027 0.76365145  
## chas -0.05589158 -0.04269672 0.06293803 1.000000000 0.09120281  
## nox 0.42097171 -0.51660371 0.76365145 0.091202807 1.00000000  
## rm -0.21924670 0.31199059 -0.39167585 0.091251225 -0.30218819  
## age 0.35273425 -0.56953734 0.64477851 0.086517774 0.73147010  
## dis -0.37967009 0.66440822 -0.70802699 -0.099175780 -0.76923011  
## rad 0.62550515 -0.31194783 0.59512927 -0.007368241 0.61144056  
## tax 0.58276431 -0.31456332 0.72076018 -0.035586518 0.66802320  
## ptratio 0.28994558 -0.39167855 0.38324756 -0.121515174 0.18893268  
## black -0.38506394 0.17552032 -0.35697654 0.048788485 -0.38005064  
## lstat 0.45562148 -0.41299457 0.60379972 -0.053929298 0.59087892  
## medv -0.38830461 0.36044534 -0.48372516 0.175260177 -0.42732077  
## rm age dis rad tax ptratio  
## crim -0.21924670 0.35273425 -0.37967009 0.625505145 0.58276431 0.2899456  
## zn 0.31199059 -0.56953734 0.66440822 -0.311947826 -0.31456332 -0.3916785  
## indus -0.39167585 0.64477851 -0.70802699 0.595129275 0.72076018 0.3832476  
## chas 0.09125123 0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152  
## nox -0.30218819 0.73147010 -0.76923011 0.611440563 0.66802320 0.1889327  
## rm 1.00000000 -0.24026493 0.20524621 -0.209846668 -0.29204783 -0.3555015  
## age -0.24026493 1.00000000 -0.74788054 0.456022452 0.50645559 0.2615150  
## dis 0.20524621 -0.74788054 1.00000000 -0.494587930 -0.53443158 -0.2324705  
## rad -0.20984667 0.45602245 -0.49458793 1.000000000 0.91022819 0.4647412  
## tax -0.29204783 0.50645559 -0.53443158 0.910228189 1.00000000 0.4608530  
## ptratio -0.35550149 0.26151501 -0.23247054 0.464741179 0.46085304 1.0000000  
## black 0.12806864 -0.27353398 0.29151167 -0.444412816 -0.44180801 -0.1773833  
## lstat -0.61380827 0.60233853 -0.49699583 0.488676335 0.54399341 0.3740443  
## medv 0.69535995 -0.37695457 0.24992873 -0.381626231 -0.46853593 -0.5077867  
## black lstat medv  
## crim -0.38506394 0.4556215 -0.3883046  
## zn 0.17552032 -0.4129946 0.3604453  
## indus -0.35697654 0.6037997 -0.4837252  
## chas 0.04878848 -0.0539293 0.1752602  
## nox -0.38005064 0.5908789 -0.4273208  
## rm 0.12806864 -0.6138083 0.6953599  
## age -0.27353398 0.6023385 -0.3769546  
## dis 0.29151167 -0.4969958 0.2499287  
## rad -0.44441282 0.4886763 -0.3816262  
## tax -0.44180801 0.5439934 -0.4685359  
## ptratio -0.17738330 0.3740443 -0.5077867  
## black 1.00000000 -0.3660869 0.3334608  
## lstat -0.36608690 1.0000000 -0.7376627  
## medv 0.33346082 -0.7376627 1.0000000

# (d)  
par(mfrow=c(1,3))  
hist(Boston$crim[Boston$crim>1], breaks=25)  
hist(Boston$tax, breaks=25)  
hist(Boston$ptratio, breaks=25)



#### most cities have low crime rates, but there is a long tail: 18 suburbs appear to have a crime rate > 20, reaching to above 80.

#### There is a large divide between suburbs with low tax rates and a peak at 660-680.

#### A skew towards high ratios, but no particularly high ratios.

# (e)  
dim(subset(Boston, chas == 1))

## [1] 35 14

# (f)  
median(Boston$ptratio)

## [1] 19.05

# (g)  
selection = Boston[order(Boston$medv),]  
selection[1,]

## crim zn indus chas nox rm age dis rad tax ptratio black lstat  
## 399 38.3518 0 18.1 0 0.693 5.453 100 1.4896 24 666 20.2 396.9 30.59  
## medv  
## 399 5

# (h)  
dim(subset(Boston, rm > 7))

## [1] 64 14

dim(subset(Boston, rm > 8))

## [1] 13 14

summary(subset(Boston, rm > 8))

## crim zn indus chas   
## Min. :0.02009 Min. : 0.00 Min. : 2.680 Min. :0.0000   
## 1st Qu.:0.33147 1st Qu.: 0.00 1st Qu.: 3.970 1st Qu.:0.0000   
## Median :0.52014 Median : 0.00 Median : 6.200 Median :0.0000   
## Mean :0.71879 Mean :13.62 Mean : 7.078 Mean :0.1538   
## 3rd Qu.:0.57834 3rd Qu.:20.00 3rd Qu.: 6.200 3rd Qu.:0.0000   
## Max. :3.47428 Max. :95.00 Max. :19.580 Max. :1.0000   
## nox rm age dis   
## Min. :0.4161 Min. :8.034 Min. : 8.40 Min. :1.801   
## 1st Qu.:0.5040 1st Qu.:8.247 1st Qu.:70.40 1st Qu.:2.288   
## Median :0.5070 Median :8.297 Median :78.30 Median :2.894   
## Mean :0.5392 Mean :8.349 Mean :71.54 Mean :3.430   
## 3rd Qu.:0.6050 3rd Qu.:8.398 3rd Qu.:86.50 3rd Qu.:3.652   
## Max. :0.7180 Max. :8.780 Max. :93.90 Max. :8.907   
## rad tax ptratio black   
## Min. : 2.000 Min. :224.0 Min. :13.00 Min. :354.6   
## 1st Qu.: 5.000 1st Qu.:264.0 1st Qu.:14.70 1st Qu.:384.5   
## Median : 7.000 Median :307.0 Median :17.40 Median :386.9   
## Mean : 7.462 Mean :325.1 Mean :16.36 Mean :385.2   
## 3rd Qu.: 8.000 3rd Qu.:307.0 3rd Qu.:17.40 3rd Qu.:389.7   
## Max. :24.000 Max. :666.0 Max. :20.20 Max. :396.9   
## lstat medv   
## Min. :2.47 Min. :21.9   
## 1st Qu.:3.32 1st Qu.:41.7   
## Median :4.14 Median :48.3   
## Mean :4.31 Mean :44.2   
## 3rd Qu.:5.12 3rd Qu.:50.0   
## Max. :7.44 Max. :50.0

summary(Boston)

## 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

#### There are 64 homes with 7 rooms, and there are 13 homes with 8 room. THe rates of crimes increase at neighborhoods with fewer rooms.