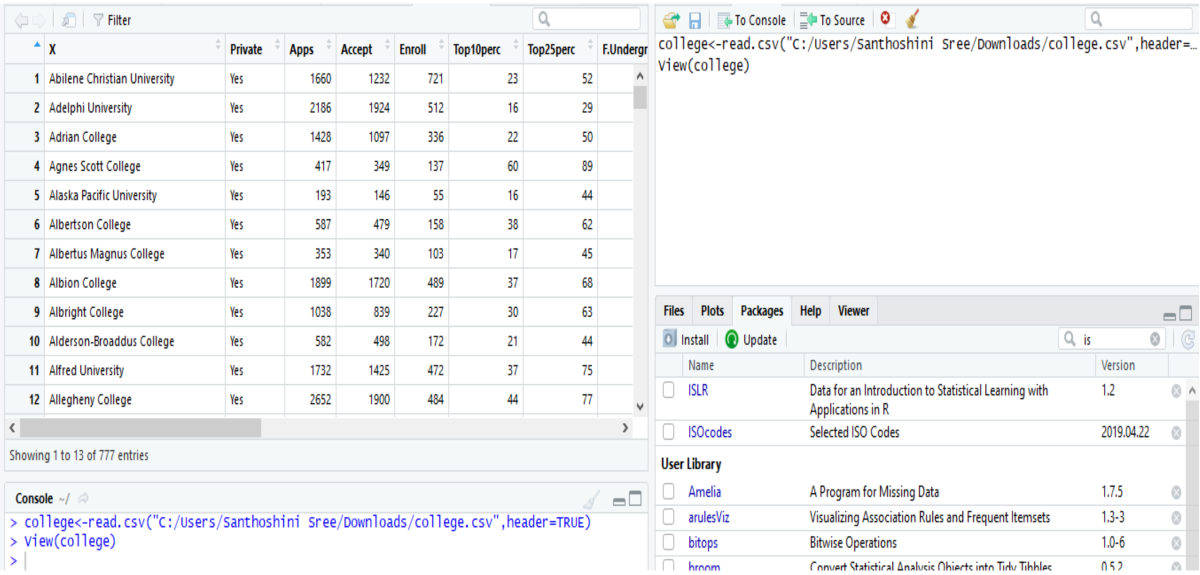
Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

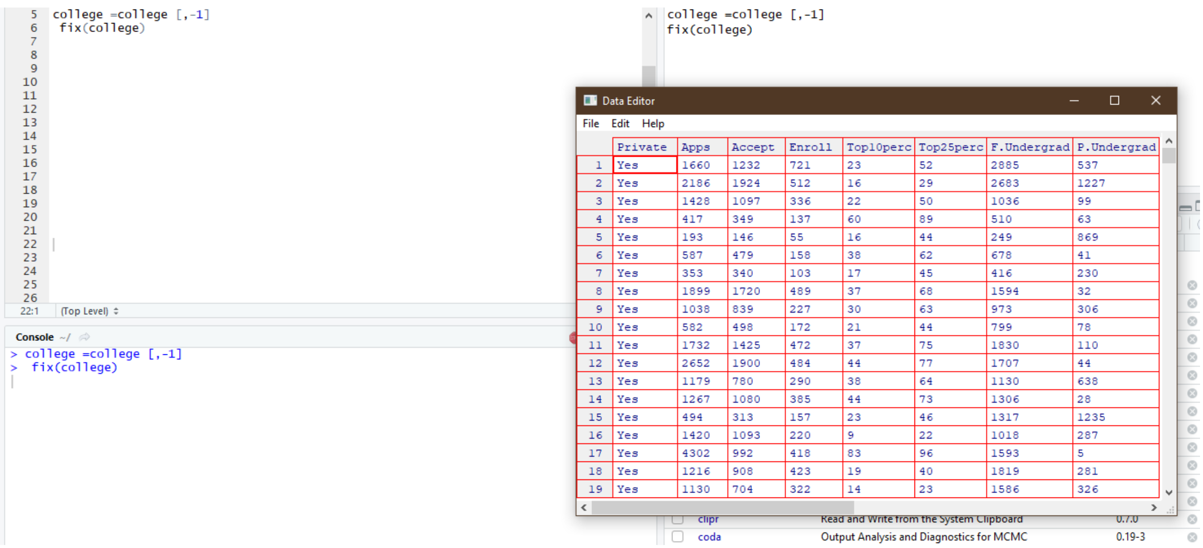
LAB#1

CHAPTER 2

8. (a)
college<-read.csv("C:/Users/SanthoshiniSree/Downloads/college.csv",header=TRUE)
View(college)



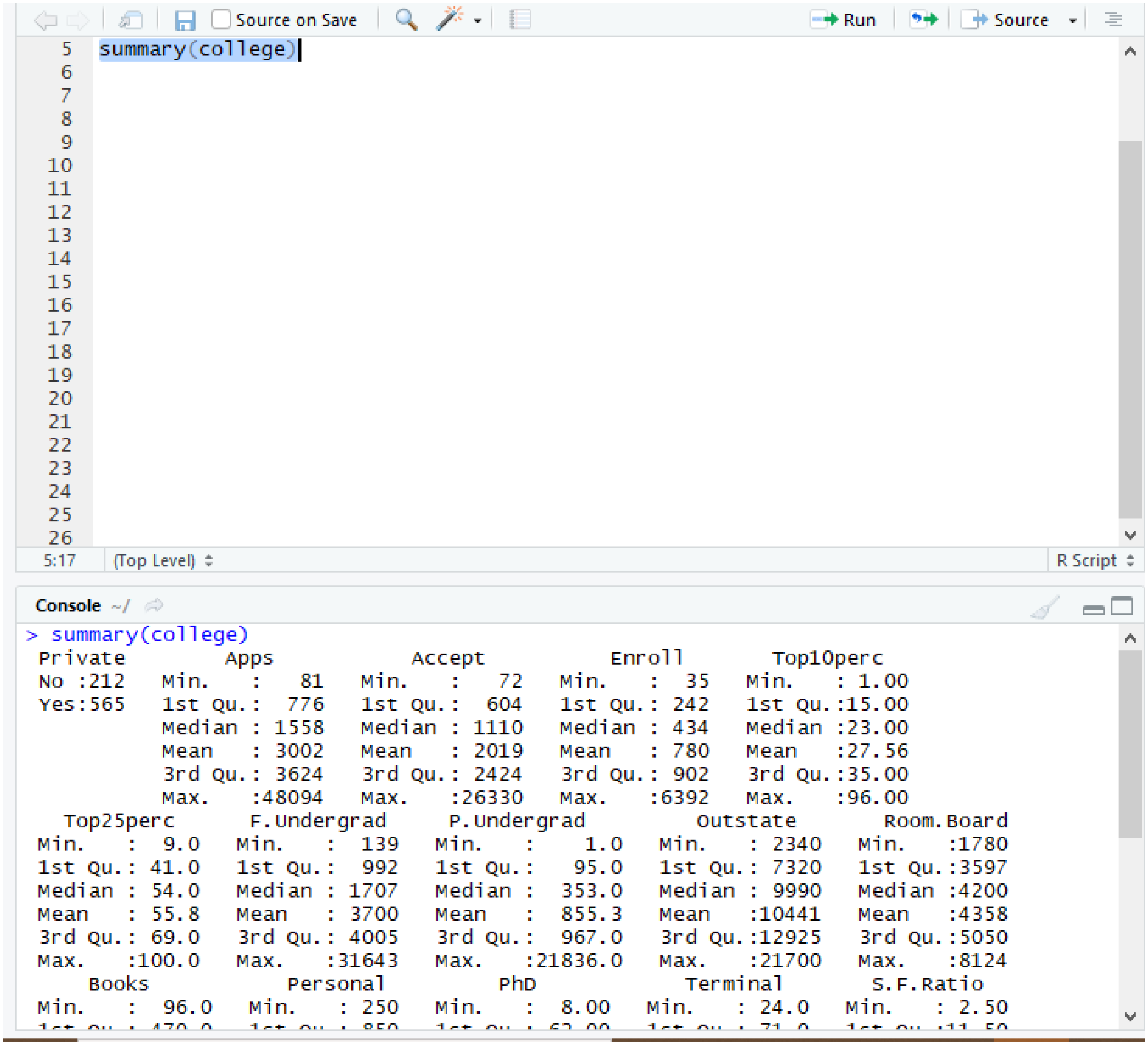
(b)
Rownames(college)=college[,1]
Fix(college)



Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(c)

(1)

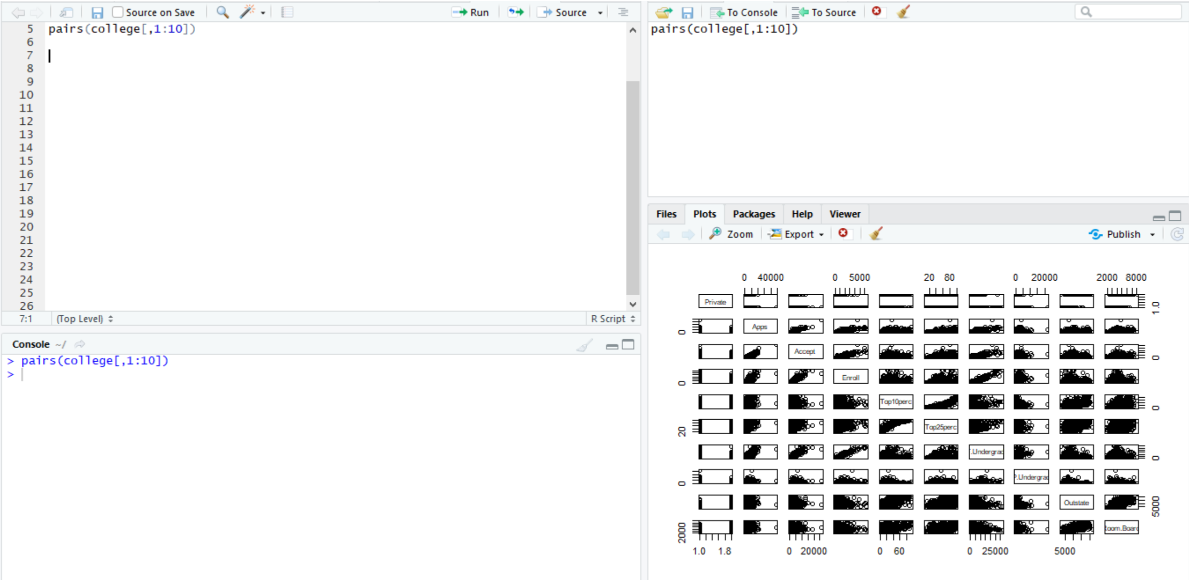


```
summary(college)
```

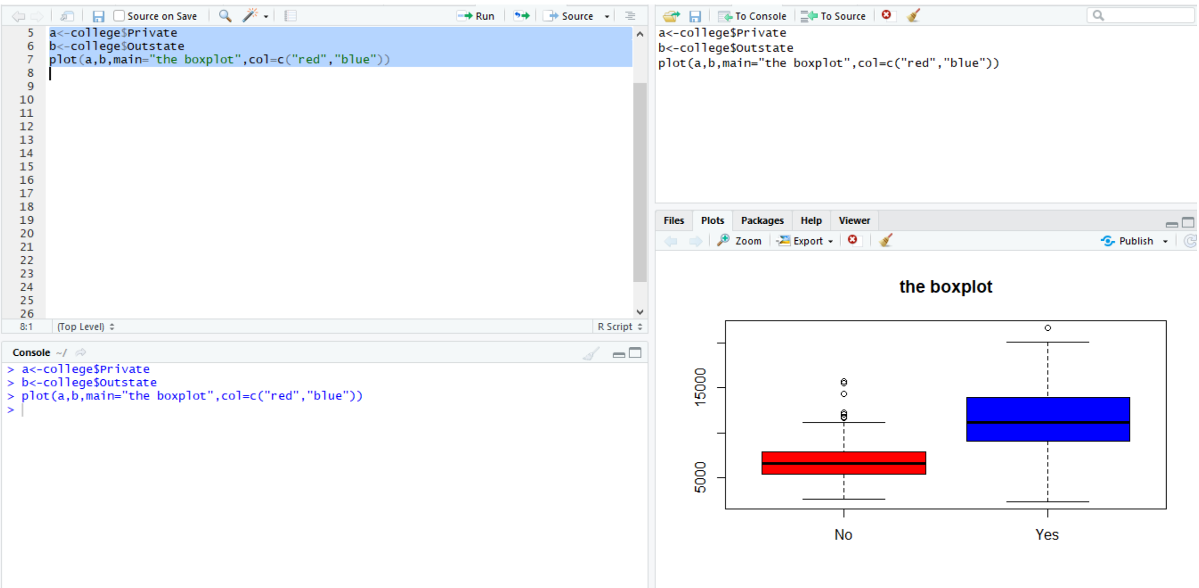
```
> summary(college)
 Private        Apps           Accept          Enroll        Top10perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
           Median : 1558   Median : 1110   Median : 434   Median :23.00
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   Top25perc      F.Undergrad     P.Undergrad         Outstate       Room.Board
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :4200
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
     Books           Personal         PhD            Terminal       S.F.Ratio
 Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(2)pairs(college[,1:10])



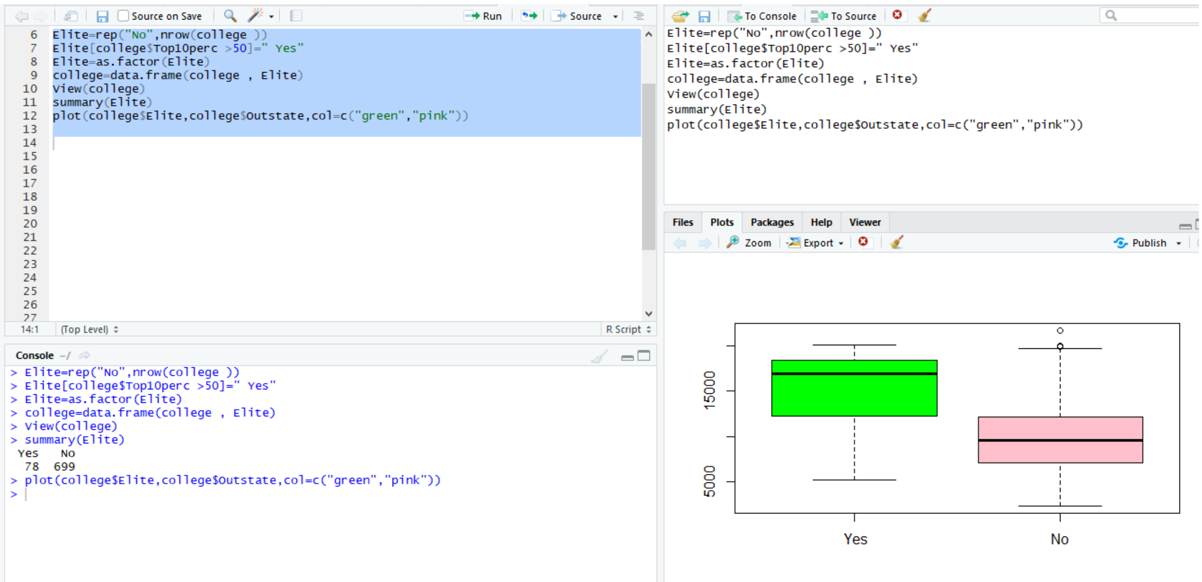
(3)
a<-college$Private
b<-college$Outstate
plot(a,b,col=c("red","yellow"))



Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

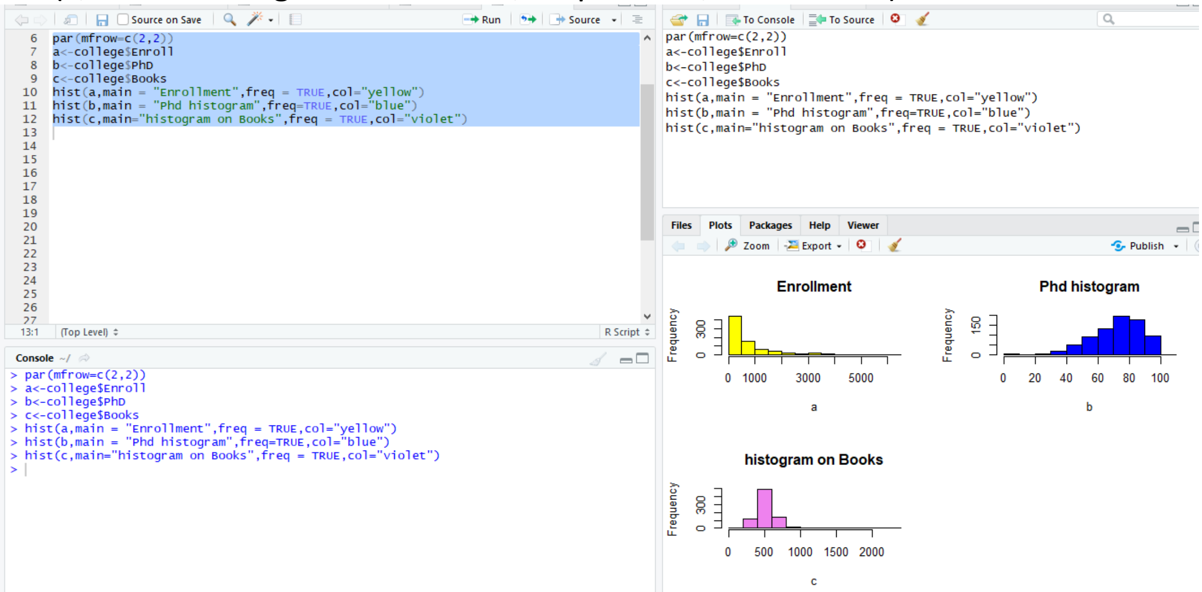
(4)

```
Elite=rep("No",nrow(college ))
Elite[college$Top10perc >50]=" Yes"
Elite=as.factor(Elite)
college=data.frame(college , Elite)
View(college)
summary(Elite)
plot(college$Elite,college$Outstate,col=c("green","pink"))
```



(5)

```
par(mfrow=c(2,2))
a<-college$Enroll
b<-college$PhD
c<-college$Books
hist(a,main = "Enrollment",freq = TRUE,col="yellow")
hist(b,main = "Phd histogram",freq=TRUE,col="blue")
hist(c,main="histogram on Books",freq = TRUE,col="violet")
```

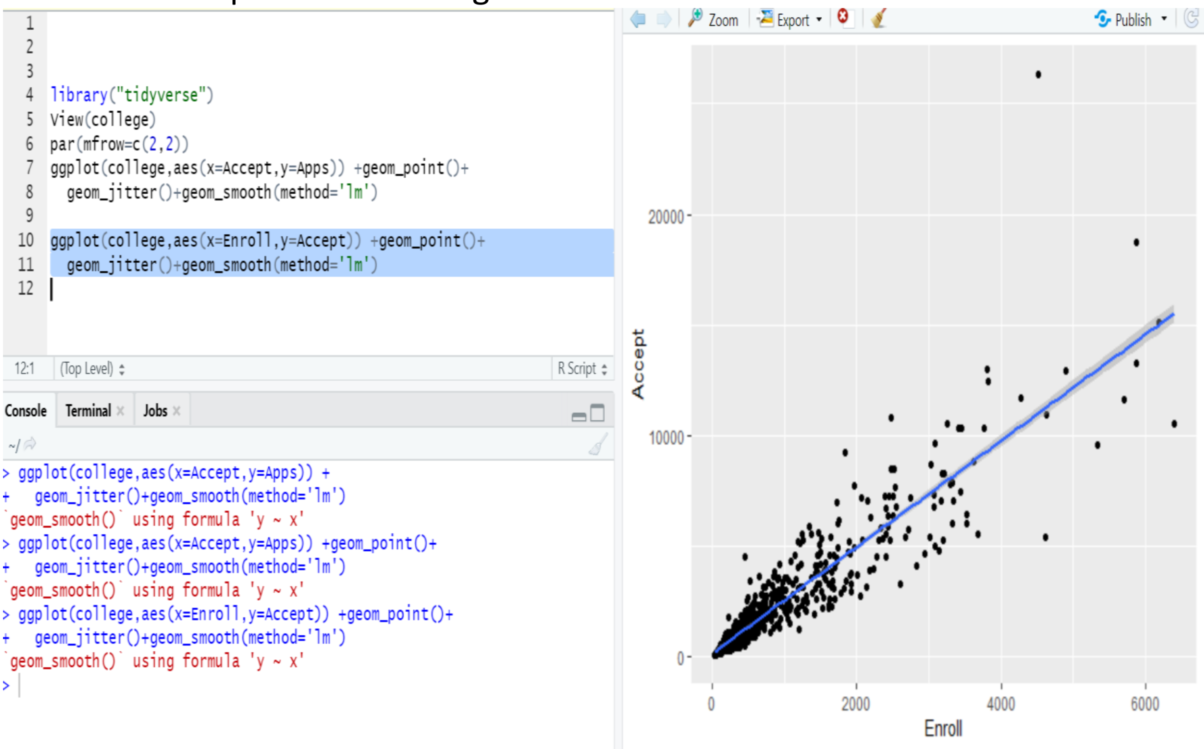


Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(6)

When the applications are more the acceptance rate is less

When the acceptance rate is high enrollment of students is less

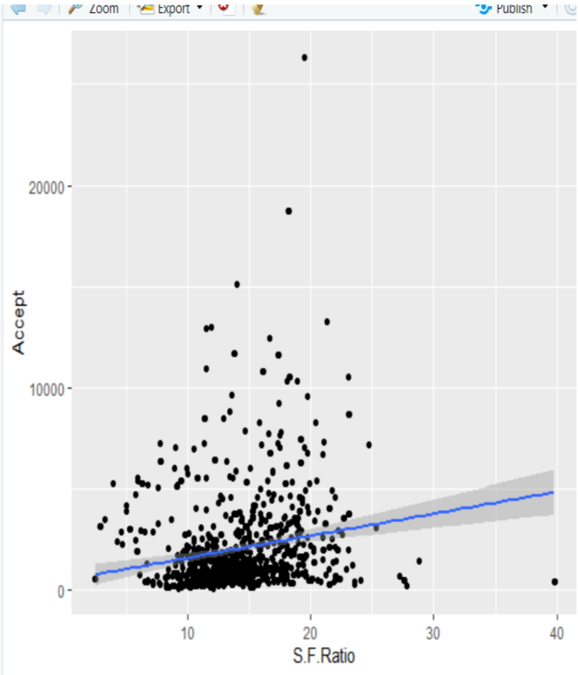


Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

Colleges which have less acceptance have less student to teacher ratio.

```
library("tidyverse")
View(college)
par(mfrow=c(2,2))
ggplot(college,aes(x=Accept,y=Apps)) +geom_point()+
  geom_jitter()+geom_smooth(method='lm')

ggplot(college,aes(x=Enroll,y=Accept)) +geom_point()+
  geom_jitter()+geom_smooth(method='lm')

ggplot(college,aes(x=S.F.Ratio,y=Accept)) +geom_point()+
  geom_jitter()+geom_smooth(method='lm')
```

```
> ggplot(college,aes(x=S.F.Ratio,y=Accept)) +geom_point()+
+   geom_jitter()+geom_smooth(method='lm')
`geom_smooth()` using formula 'y ~ x'
>
```



(9)

(a)The quantitative predictors are:

Mpg, cylinders, displacement, horsepower, weight, acceleration, year and origin

The qualitative predictor is: name

```
View(Auto)
sapply(colnames(Auto), function(x) class(Auto[[x]]))
```

```
> sapply(colnames(Auto), function(x) class(Auto[[x]]))
         mpg    cylinders displacement   horsepower       weight acceleration         year       origin 
   "numeric"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric" 
        name 
    "factor" 
>
```

(b)Range of each predictor can be shown using sapply()

```
sapply(Auto[1:8],function(x) range(x))
```

```
> sapply(Auto[1:8],function(x) range(x))
      mpg cylinders displacement horsepower weight acceleration year origin
[1,]  9.0         3           68         46   1613          8.0   70      1
[2,] 46.6         8          455        230   5140         24.8   82      3
>
```

(c)mean() and sd() function gives mean and standard deviation of the values

```
sapply(Auto[1:8],function(x) mean(x) )
sapply(Auto[1:8],function(x) sd(x) )

```

```
3:1    (Top Level) ÷

Console ~/
> sapply(Auto[1:8],function(x) mean(x) )
       mpg    cylinders displacement   horsepower       weight acceleration         year       origin
 23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327    75.979592     1.576531
> sapply(Auto[1:8],function(x) sd(x) )
       mpg    cylinders displacement   horsepower       weight acceleration         year       origin
 7.8050075    1.7057832  104.6440039   38.4911599  849.4025600    2.7588641    3.6837365    0.8055182
>
```

(d) To remove observations anti_join is used. anti_join is from the library dplyr. This function returns all rows of x where there is no matching values of y.

```
library(dplyr)
a<-anti_join(Auto,Auto[10:85,])
View(a)
sapply(a[1:8],function(x) mean(x) )
sapply(a[1:8],function(x) sd(x) )
sapply(a[1:8], function(x) max(x)-min(x))
sapply(a[1:8],function(x) range(x) )

```

```
4:36    (Top Level) ÷                                                              R Script ÷

Console ~/
> a<-anti_join(Auto,Auto[10:85,])
Joining, by = c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "year", "origin", "nam
e")
> View(a)
> sapply(a[1:8],function(x) mean(x) )
       mpg    cylinders displacement   horsepower       weight acceleration         year       origin
 24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899    77.145570     1.601266
> sapply(a[1:8],function(x) sd(x) )
       mpg    cylinders displacement   horsepower       weight acceleration         year       origin
  7.867283     1.654179    99.678367    35.708853   811.300208     2.693721     3.106217     0.819910
> sapply(a[1:8], function(x) max(x)-min(x))
       mpg    cylinders displacement   horsepower       weight acceleration         year       origin
      35.6          5.0        387.0        184.0       3348.0         16.3         12.0          2.0
> sapply(a[1:8],function(x) range(x) )
      mpg cylinders displacement horsepower weight acceleration year origin
[1,] 11.0         3           68         46   1649          8.5   70      1
[2,] 46.6         8          455        230   4997         24.8   82      3
>
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

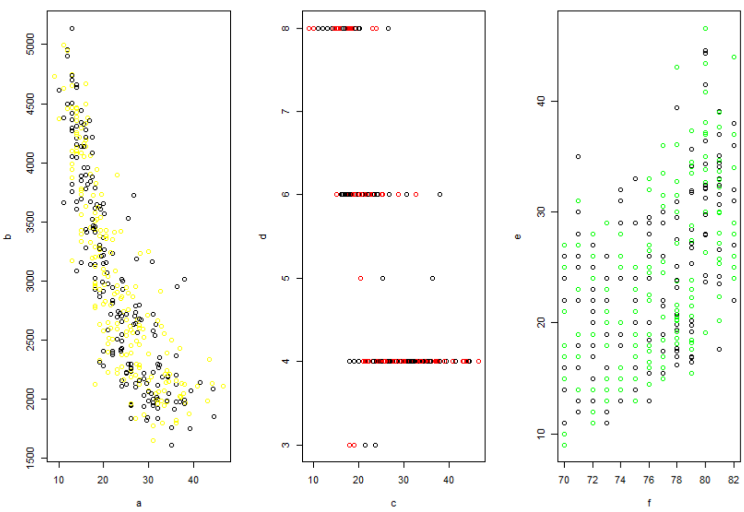
(e)

plot 1: Less mpg cylinders have high weight.

Plot2: Most of the cylinders have less mpg.

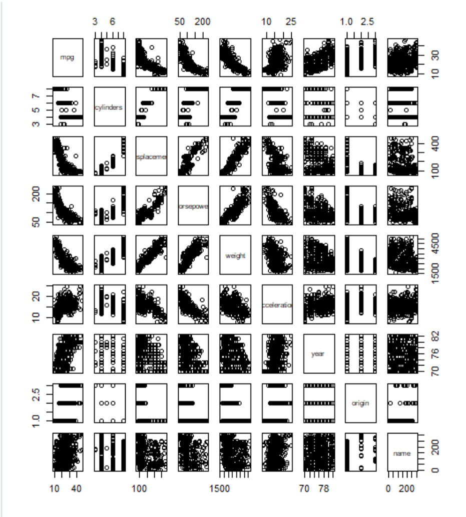
Plot 3: Over time cars are becoming systematic.

```
par(mfrow=c(1,3))

a<-Auto$mpg
b<-Auto$weight
plot(a,b,col=c("yellow","black"))

c<-Auto$mpg
d<-Auto$cylinders
plot(c,d,col=c("red","black"))


e<-Auto$mpg
f<-Auto$year
plot(f,e,col=c("green","black"))
```



(f) Every predictor correlates with mpg. This can be shown using pair() function, which return plot matrix.

```
> pairs(Auto)
> cor.test(Auto$mpg,Auto$cylinders)

        Pearson's product-moment correlation

data:  Auto$mpg and Auto$cylinders
t = -24.425, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8139723 -0.7351952
sample estimates:
       cor
-0.7776175

> cor.test(Auto$mpg,Auto$displacement)

        Pearson's product-moment correlation

data:  Auto$mpg and Auto$displacement
t = -26.808, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8373959 -0.7672653
sample estimates:
       cor
-0.8051269

> cor.test(Auto$mpg,Auto$weight)

        Pearson's product-moment correlation

data:  Auto$mpg and Auto$weight
t = -29.645, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8603702 -0.7990647
sample estimates:
       cor
-0.8322442
```

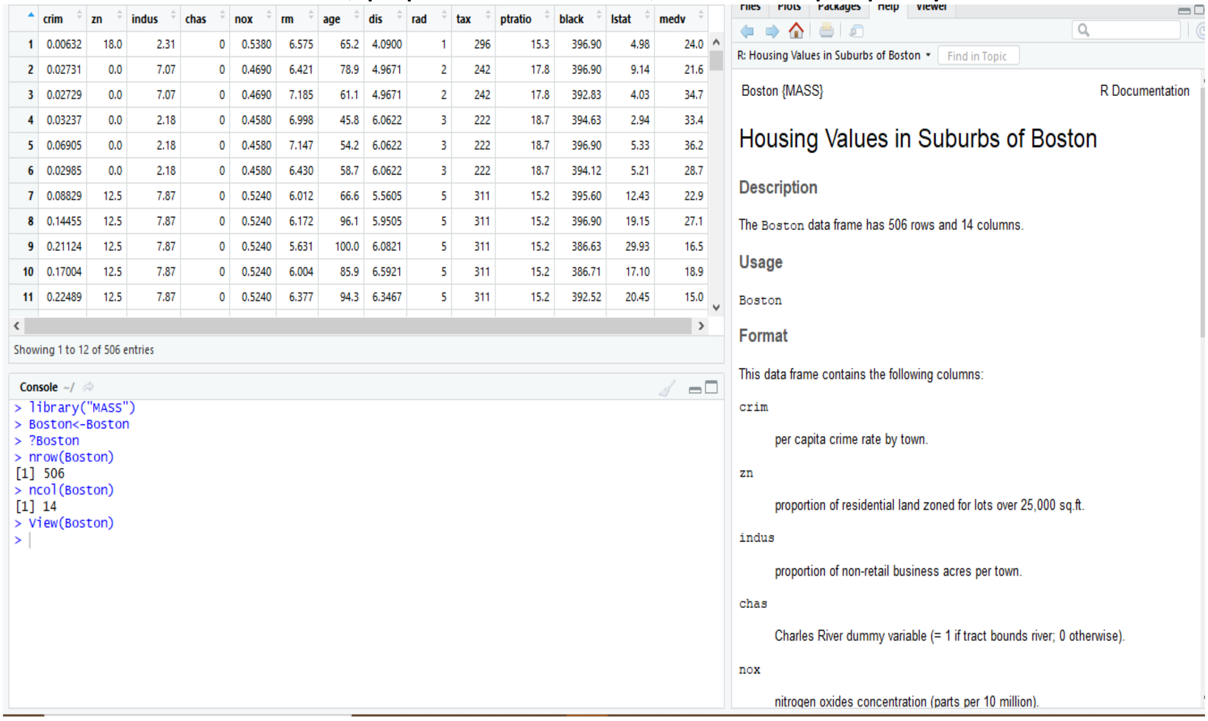


```
> cor.test(Auto$mpg,Auto$acceleration)

        Pearson's product-moment correlation

data:  Auto$mpg and Auto$acceleration
t = 9.2277, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3384724 0.5013550
sample estimates:
      cor
0.4233285

> cor.test(Auto$mpg,Auto$year)

        Pearson's product-moment correlation

data:  Auto$mpg and Auto$year
t = 14.08, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5108684 0.6426366
sample estimates:
     cor
0.580541

> cor.test(Auto$mpg,Auto$origin)

        Pearson's product-moment correlation

data:  Auto$mpg and Auto$origin
t = 13.531, df = 390, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4938051 0.6290414
sample estimates:
      cor
0.5652088
```

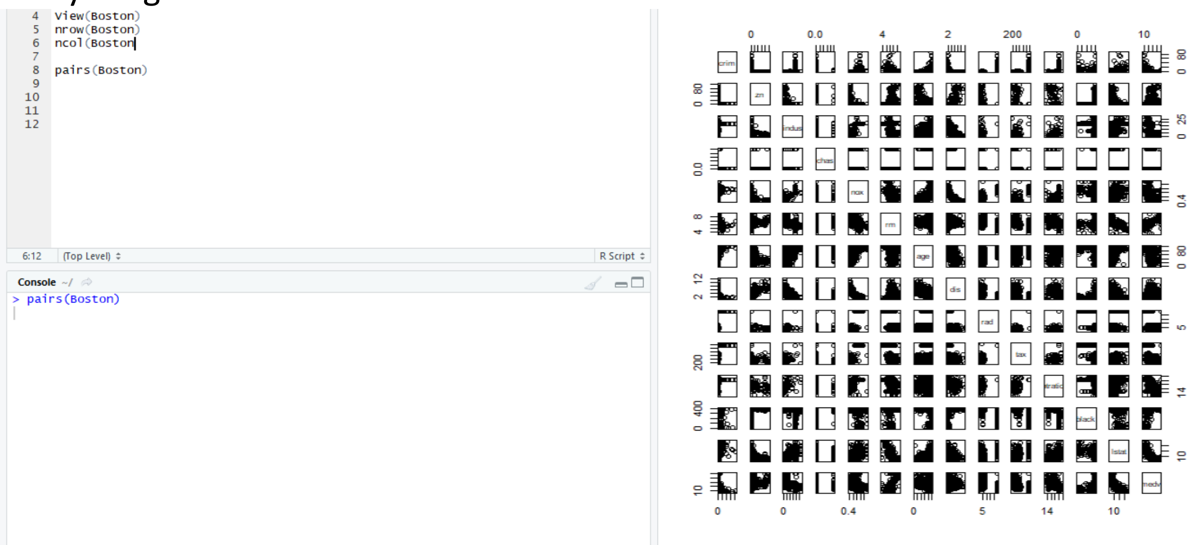
Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(10)

(a)There are 506 rows and 14 columns. The rows represent the locality of Boston. The crime rate, pupil-teacher ratio, full value property tax and more.

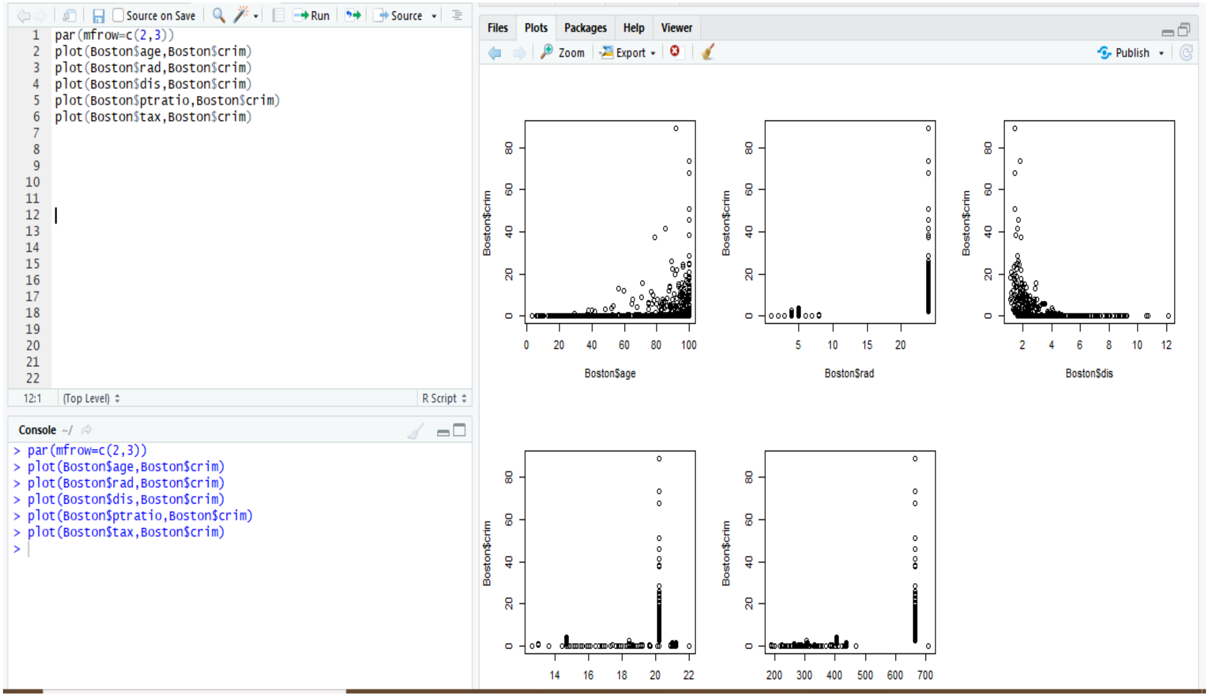


(b) As shown, there are many scatter plots and is difficult to read and clean everything.

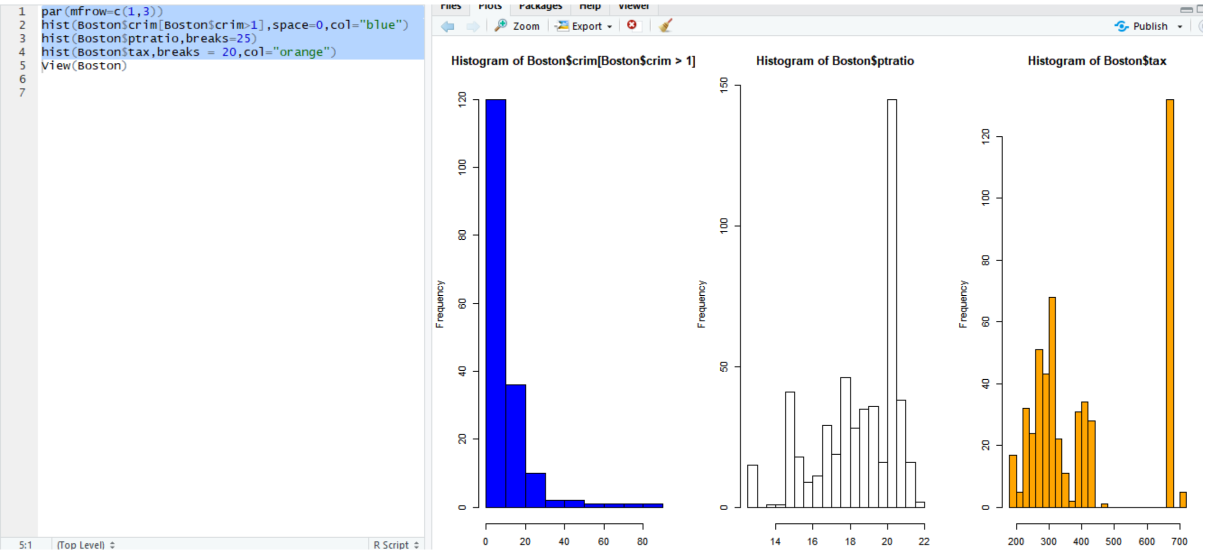


(c)

Plot(Boston$age,Boston$crim)-units build in 1940 have more crime rates
Plot(Boston$rad,Boston$crim)-radial highways have more crime rates
Plot(Boston$dis,Boston$crim)-closer to work area have more rates
Plot(Boston$ptratio,Boston$crim)-when pupil-teacher is high it has more crime rates
Plot(Boston$tax,Boston$crim)-property tax rate has more crime.



(d)
par(mfrow=c(1,3))
barplot(Boston$crim[Boston$crim>1],space=0,col="blue")- From 0-10 the the bar has significantly increased
hist(Boston$ptratio,breaks=25)- The ratio is high between 20 to 22
hist(Boston$tax,breaks = 20,col="orange")- In between 680-690 the property tax is high.



(e)
35 suburbs

```
sum(Boston$chas==1)
```

Console ~/

```
> sum(Boston$chas==1)
[1] 35
>
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(f)

```
1  median(Boston$ptratio)
2
```

1:1   (Top Level)

Console ~/

```
> median(Boston$ptratio)
[1] 19.05
>
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(g)

the crime rate for the median value of owner occupied homes is 38.3518 this means crime rate is high in that area far from highways and Charles river area. So, it is not a better place to live

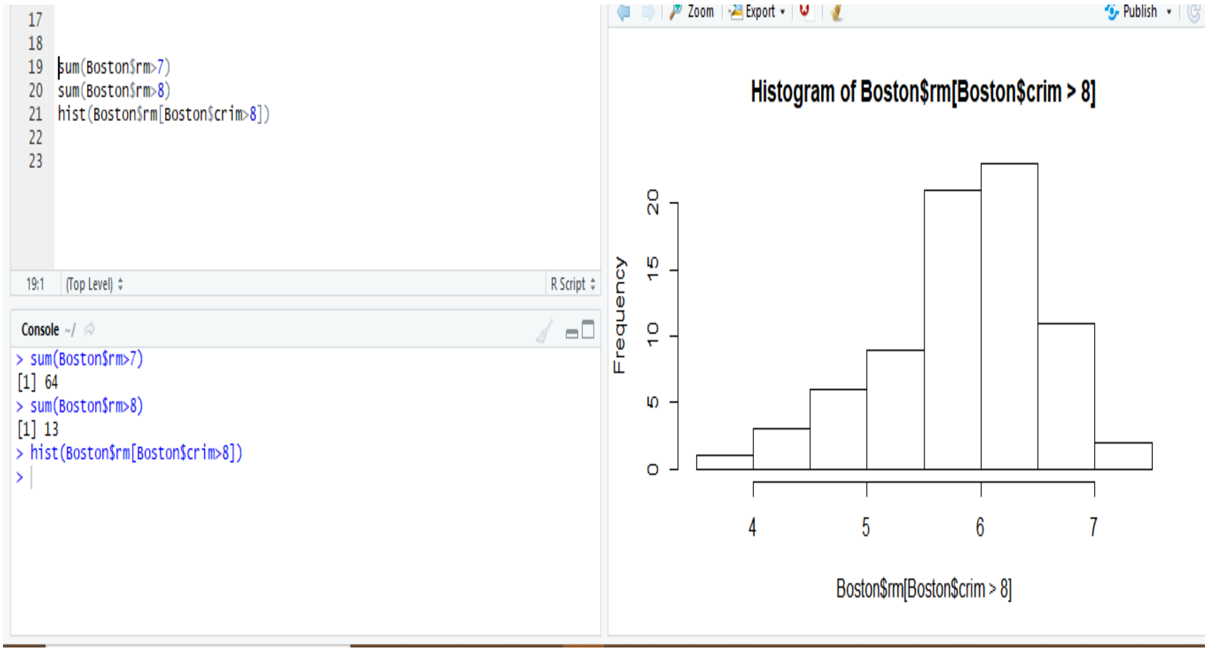
```
subset(Boston,medv==min(Boston$medv))
summary(Boston)

4:1    (Top Level) ÷
```

```
Console ~/
> subset(Boston,medv==min(Boston$medv))
       crim zn indus chas   nox    rm
399 38.3518  0  18.1    0 0.693 5.453
406 67.9208  0  18.1    0 0.693 5.683
    age    dis rad tax ptratio  black
399 100 1.4896  24 666    20.2 396.90
406 100 1.4254  24 666    20.2 384.97
    lstat medv
399 30.59    5
406 22.98    5
> summary(Boston)
      crim                zn        
 Min.   : 0.00632   Min.   :  0.00  
 1st Qu.: 0.08204   1st Qu.:  0.00  
 Median : 0.25651   Median :  0.00  
 Mean   : 3.61352   Mean   : 11.36  
 3rd Qu.: 3.67708   3rd Qu.: 12.50  
 Max.   :88.97620   Max.   :100.00  
     indus            chas        
 Min.   : 0.46   Min.   :0.00000  
 1st Qu.: 5.19   1st Qu.:0.00000  
 Median : 9.69   Median :0.00000  
 Mean   :11.14   Mean   :0.06917  
 3rd Qu.:18.10   3rd Qu.:0.00000  
 Max.   :27.74   Max.   :1.00000  
      nox               rm       
 Min.   :0.3850   Min.   :3.561  
 1st Qu.:0.4490   1st Qu.:5.886  
 Median :0.5380   Median :6.208  
 Mean   :0.5547   Mean   :6.285  
 3rd Qu.:0.6240   3rd Qu.:6.623  
 Max.   :0.8710   Max.   :8.780  
      age              dis        
 Min.   :  2.90   Min.   : 1.130  
 1st Qu.: 45.02   1st Qu.: 2.100  
 Median : 77.50   Median : 3.207  
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

(h)
sum(Boston$rm>7)
sum(Boston$rm>8)
There are 64 suburbs that average more than 7 rooms per dwelling and 13 suburbs that average more than 8 rooms per dwelling.

hist(Boston$rm[Boston$crim>8])
In the histogram, the bar simultaneously increased in between the values 5.5 to 6.5



Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

CHAPTER 3

Q8. This question involves the use of simple linear regression on the “Auto” data set.

a. Use the lm() function to perform a simple linear regression with “mpg” as the response and “horsepower” as the predictor. Use the summary() function to print the results. Comment on the output. For example :

i. Is there a relationship between the predictor and the response ?

```
#install.packages('ISLR')
library(ISLR)
data(Auto)
fit <- lm(mpg ~ horsepower, data = Auto)
summary(fit)
```

```
> data(Auto)
> fit <- lm(mpg ~ horsepower, data = Auto)
> summary(fit)

Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5710  -3.2592  -0.3435   2.7630  16.9240 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,	Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Answer: As per the hypothesis testing, The p-value from the above screenshot corresponding to the F-statistic is 7.03198910^{-81}, this indicates a clear evidence of a relationship between “mpg” and “horsepower”.

ii. How strong is the relationship between the predictor and the response?

Answer: We could see that there is a negative correlation between mpg and horsepower as the coefficient value is -0.16. The unit rise in horsepower decrease the 0.16 milage per gallon. Means there is a fair and considerable correlation between response and predictor variable. Also, we can see that RSE of the lm.fit was 4.906 which indicates a percentage error of 20.9237141%. We could see that the multiple R square is 0.6059483, which means 60.5948258% of the variability in "mpg" will be explained using "horsepower".

iii. Is the relationship between the predictor and the response positive or negative?

Answer: As the correlation coefficient is negative from the above screenshot, we can say that there is negative linear relationship between mpg and horsepower.

iv. What is the predicted mpgmpg associated with a "horsepower" of 98 ? What are the associated 95% confidence and prediction intervals ?

Answer:

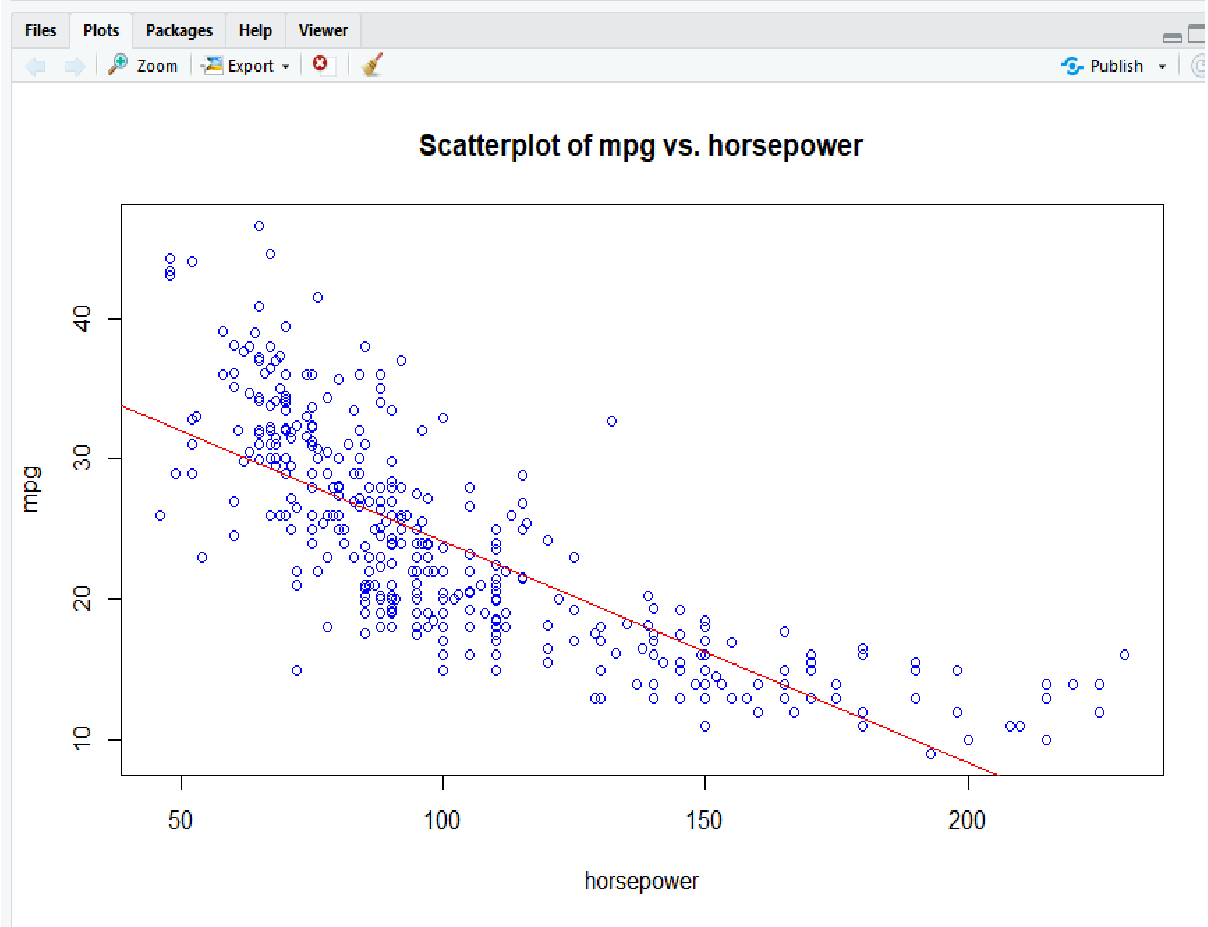
```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> 
```

```
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
> 
```

b. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

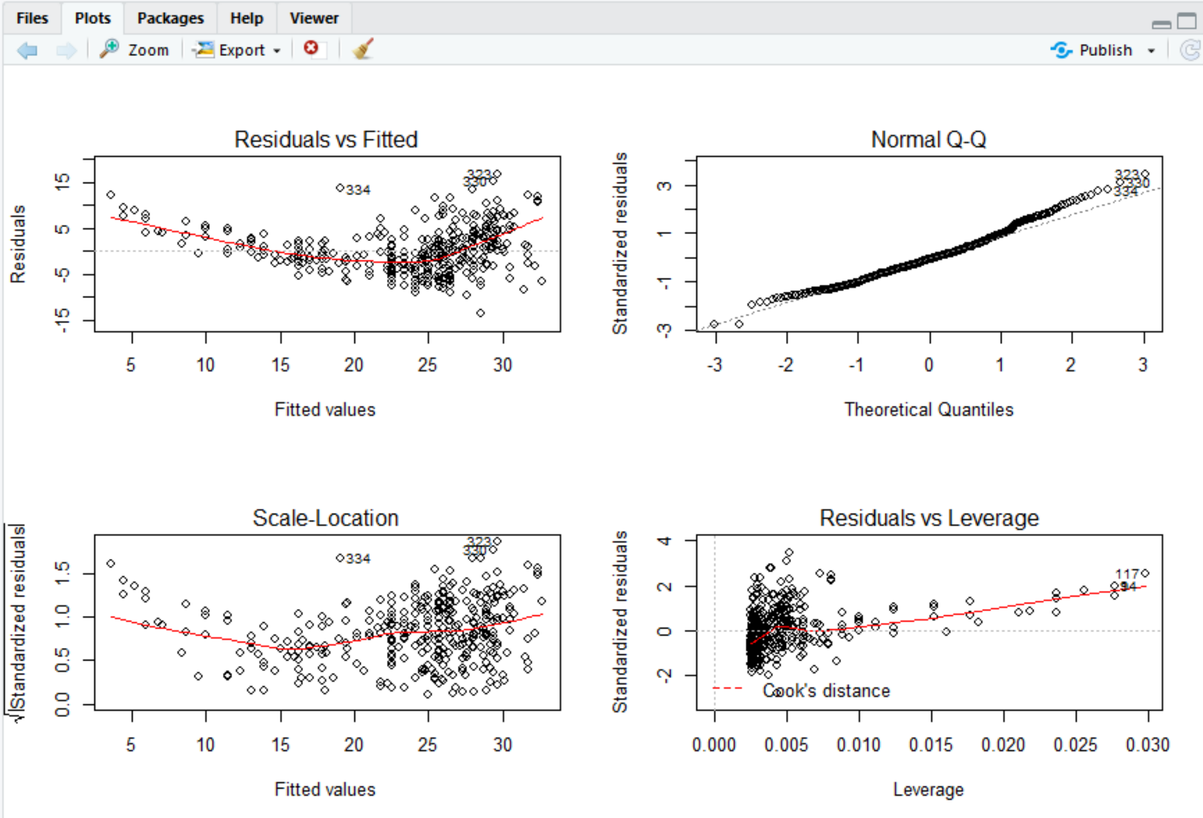
```
> plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower", xlab = "horsepower", ylab = "mpg", col = "blue
")
> abline(fit, col = "red")
> 
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565



c. Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```
> par(mfrow = c(2, 2))
> plot(fit)
>
```



The plot of residuals versus fitted values indicates the presence of non-linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and a few high leverage points.

9. This question involves the use of multiple linear regression on the "Auto" data set.

a. Produce a scatterplot matrix which include all the variables in the data set.

```
> pairs(Auto)
>
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

b. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the “name” variable, which is qualitative.

```
> names(Auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"       "acceleration" "year"
[8] "origin"       "name"
>
```

```
> cor(Auto[1:8])
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
>
```

c. Use the lm() function to perform a multiple linear regression with “mpg” as the response and all other variables except “name” as the predictors. Use the summary() function to print the results. Comment on the output. For instance :

i. Is there a relationship between the predictors and the response ?

```
> fit2 <- lm(mpg ~ . - name, data = Auto)
> summary(fit2)

Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5903 -2.1565 -0.1169  1.8690 13.0604 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16

> |
```

By considering the hypothesis testing, the p-value corresponding to the F-statistic is 2.037105910^ {-139}, this indicates a clear evidence of a relationship between “mpg” and the input predictors.

ii. Which predictors appear to have a statistically significant relationship to the response?

Answer: From the above screenshot reference, the p-value of all the predictors is below threshold that is 0.05 except cylinders, acceleration and horsepower.

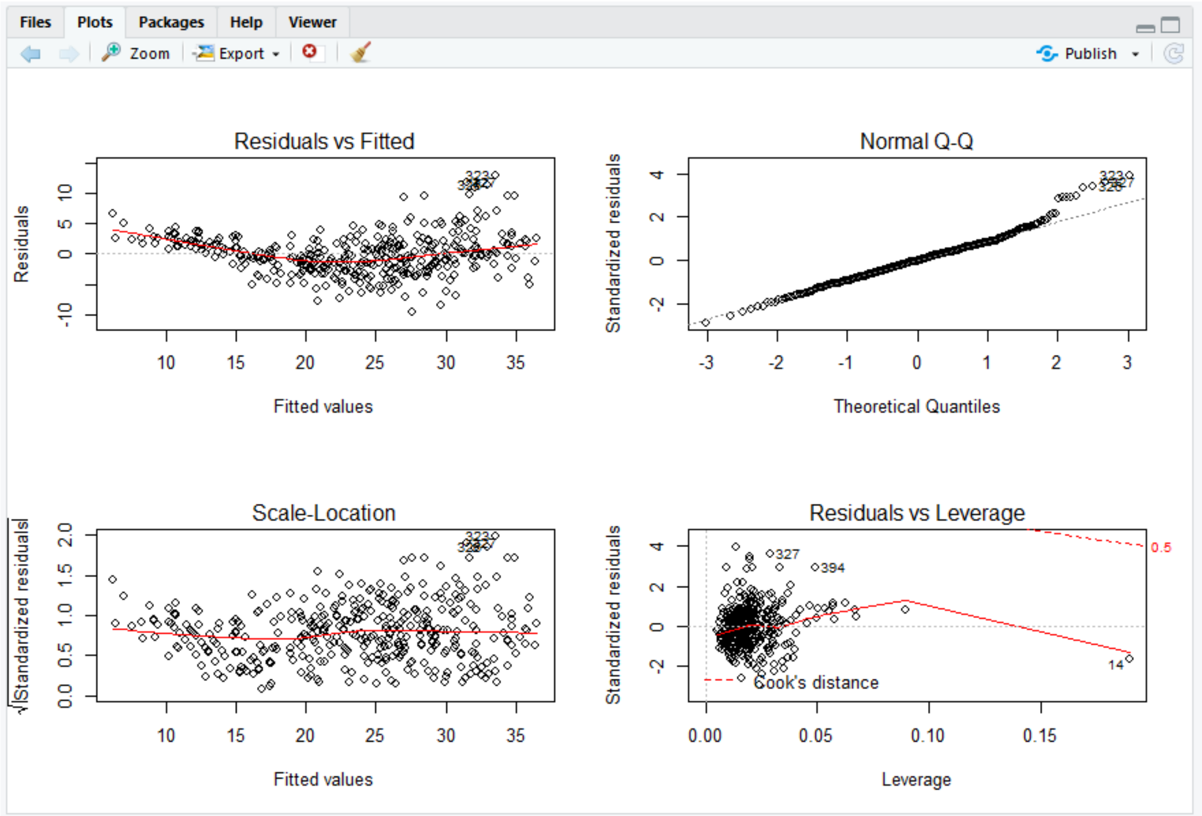
iii. What does the coefficient for the “year” variable suggest ?

Answer: The coefficient of variable “year” suggesting that a unit increase in year increases 0.75 times the miles per gallon(mpg). With we can say that, cars are becoming fuel efficient year by year.

d. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers ? Does the leverage plots identify any observations with unusually high leverages ?

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

```
> par(mfrow = c(2, 2))
> plot(fit2)
>
```



From the above screenshots, we can see that there is some trend in the distribution of residuals which disobeys the assumption of homoscedasticity. Hence, it indicates the mild non-linearity. The standardized residuals versus leverage plot indicates the presence of a few outliers.

e. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant ?

From the correlation matrix above, we have obtained the two highest correlated pairs like cylinders and displacement and weight and have used them for interaction effects.

```
> fit3 <- lm(mpg ~ cylinders * displacement+displacement * weight, data = Auto[, 1:8])
> summary(fit3)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement * 
    weight, data = Auto[, 1:8])

Residuals:
     Min       1Q   Median       3Q      Max 
-13.2934  -2.5184  -0.3476   1.8399  17.7723 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders               7.606e-01  7.669e-01   0.992    0.322    
displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384    
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237 
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16

> |
```

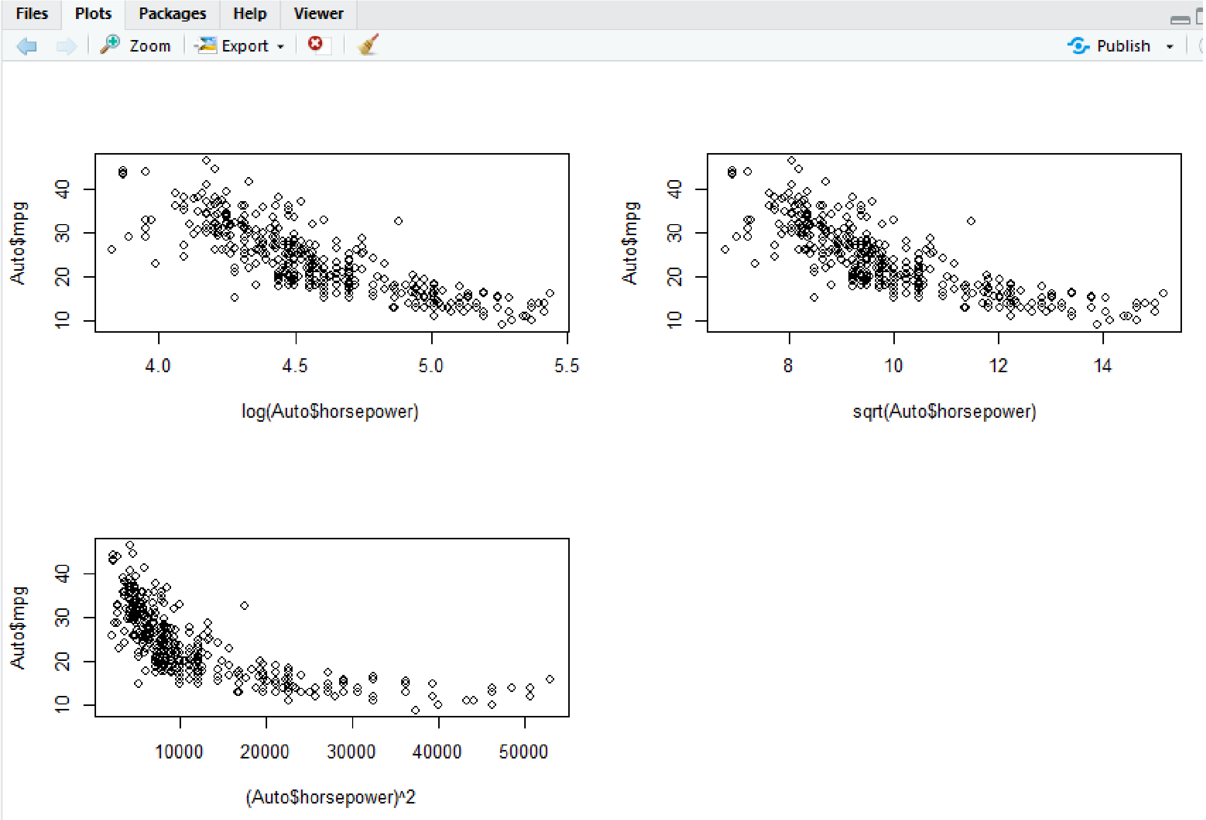
From the above p-values, we could see that the interaction between displacement and weight is statistically significant, while the interaction between cylinders and displacement is not.

f. Try a few different transformations of the variables, such as logXlog X, X−−√X, X2X2. Comment on your findings.

Answer:

```
> par(mfrow = c(2, 2))
> plot(log(Auto$horsepower), Auto$mpg)
> plot(sqrt(Auto$horsepower), Auto$mpg)
> plot((Auto$horsepower)^2, Auto$mpg)
> |
```

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565



So far, we have used horsepower as is for model fitting. But the value as is doesn't fit linear as its log transformation fits in the first plot.

Santhoshini Sree, Alexa Summers, Gireesh Kumar Muppalla
CS 5565

Q10. This question should be answered using the “Carseats” data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
1   View(Carseats)
2   ?Carseats
3   model<-lm(Sales~Price+Urban+US,data=Carseats)
4   summary(model)
5
6
6:1   (Top Level) ÷
```

```
Console ~/
> View(Carseats)
> ?Carseats
> model<-lm(Sales~Price+Urban+US,data=Carseats)
> summary(model)

Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16

> |
```

(b) Provide an interpretation of each coeffiffifficient in the model. Be careful—some of the variables in the model are qualitative!
Answer:
The coefficient of the 'price' factors might be deciphered by saying that the normal impact of a cost of 1 dollar is a reduction of 54.4588492 units in the sales any remaining indicators staying fixed. The coefficient of the 'urban' factors might be deciphered by saying that the unit deals are 21.9161 units not exactly rural area, The coefficient of the 'US' factors might be deciphered by saying that the normal deals in the US store are 1200.572 units more than in a no US store any remaining predictors.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.
Sales = 13.0434 + (-0.0544) *price + (-0.02191)*urban + (1.2005727)*US + ε

(d) For which of the predictors can you reject the null hypothesis H0: $\beta j = 0$?
Answer:
'Price' and 'US' variables can be rejected.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
View(Carseats)
?Carseats
model<-lm(Sales~Price+Urban+US,data=Carseats)
summary(model)

6:1    (Top Level) ÷
```

```
Console ~/
> View(Carseats)
> ?Carseats
> model<-lm(Sales~Price+Urban+US,data=Carseats)
> summary(model)

Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16

> 
```

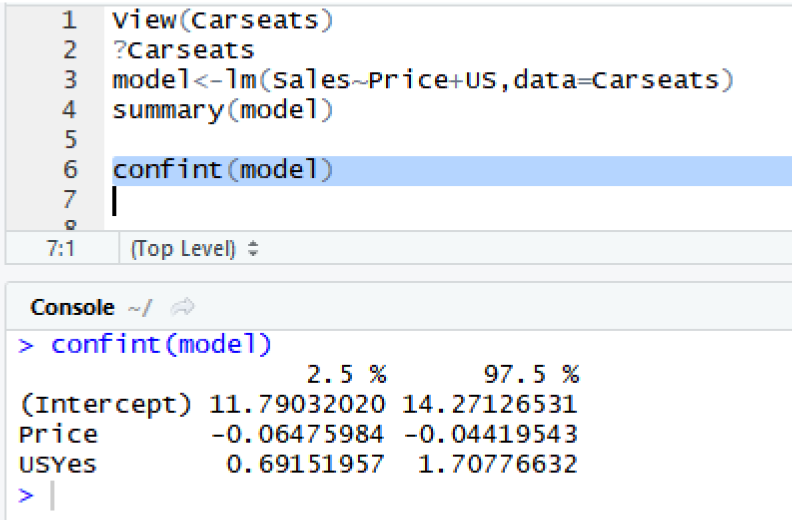
```
View(Carseats)
?Carseats
model<-lm(Sales~Price+US,data=Carseats)
summary(model)

5:1    (Top Level) ÷
```

```
Console ~/
> View(Carseats)
> ?Carseats
> model<-lm(Sales~Price+US,data=Carseats)
> summary(model)

Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9269 -1.6286 -0.0574  1.5766  7.0515 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2354 
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16

> 
```

(f) How well do the models in (a) and (e) fit the data?

The smaller model has better R square value compared to bigger model.

(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).



```
View(Carseats)
?Carseats
model<-lm(Sales~Price+US,data=Carseats)
summary(model)

confint(model)
```

```
> confint(model)
                  2.5 %      97.5 %
(Intercept) 11.79032020 14.27126531
Price       -0.06475984 -0.04419543
USYes        0.69151957  1.70776632
>
```

(h) Is there evidence of outliers or high leverage observations in the model from (e)?