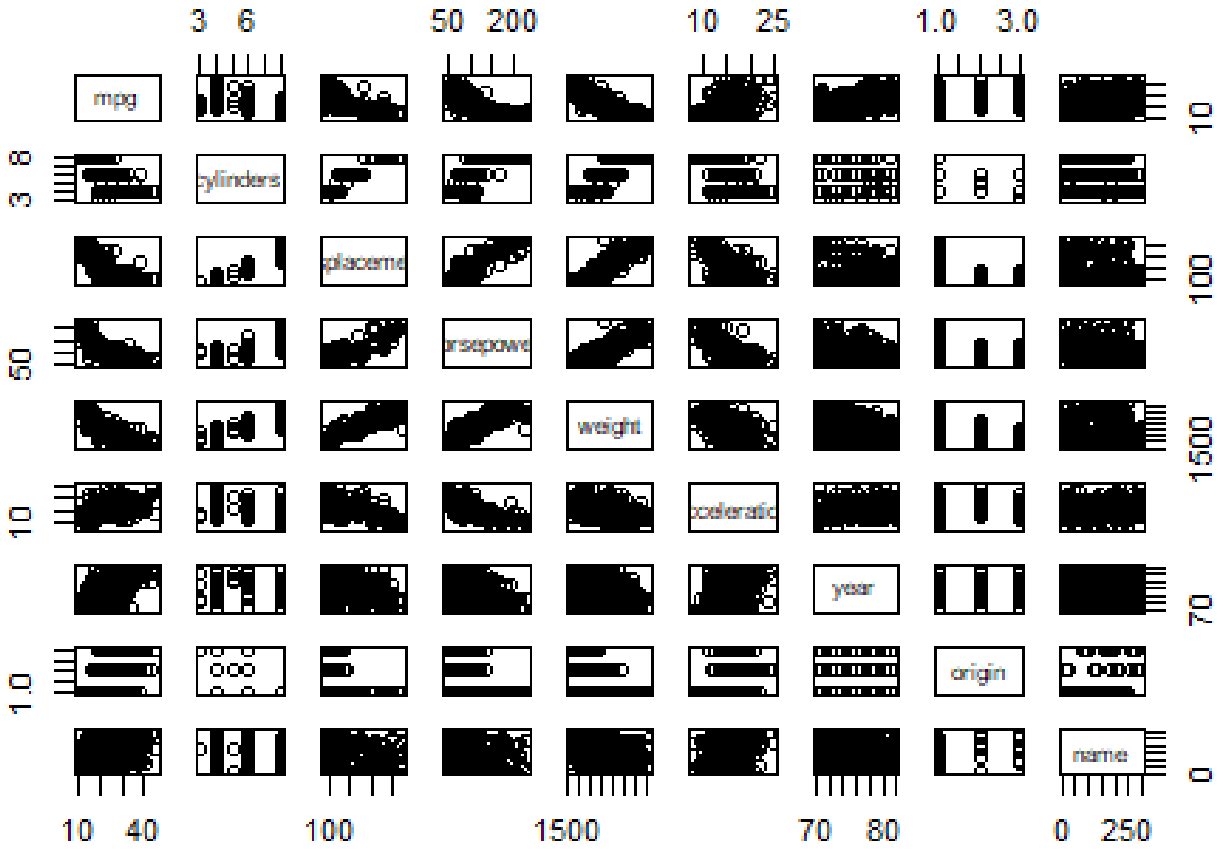
Chapter 3 Problem 9

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

```
>pairs(auto)
```



b. Compute the matrix of correlations between the variables.

```
> auto <- subset(auto, select = -c(name) ) #exclude name
> cor(auto)
                     mpg  cylinders displacement horsepower     weight acceleration       year
mpg           1.0000000 -0.7762599   -0.8044430         NA -0.8317389    0.4222974  0.5814695
cylinders    -0.7762599  1.0000000    0.9509199         NA  0.8970169   -0.5040606 -0.3467172
displacement -0.8044430  0.9509199    1.0000000         NA  0.9331044   -0.5441618 -0.3698041
horsepower           NA         NA           NA          1         NA           NA         NA
weight       -0.8317389  0.8970169    0.9331044         NA  1.0000000   -0.4195023 -0.3079004
acceleration  0.4222974 -0.5040606   -0.5441618         NA -0.4195023    1.0000000  0.2829009
year          0.5814695 -0.3467172   -0.3698041         NA -0.3079004    0.2829009  1.0000000
origin        0.5636979 -0.5649716   -0.6106643         NA -0.5812652    0.2100836  0.1843141
                 origin
mpg           0.5636979
cylinders    -0.5649716
displacement -0.6106643
horsepower           NA
weight       -0.5812652
acceleration  0.2100836
year          0.1843141
origin        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 predictors.

```
> lm.fit=lm(mpg~.,data=auto)
> summary(lm.fit)

Call:
lm(formula = mpg ~ ., 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
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8215,      Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

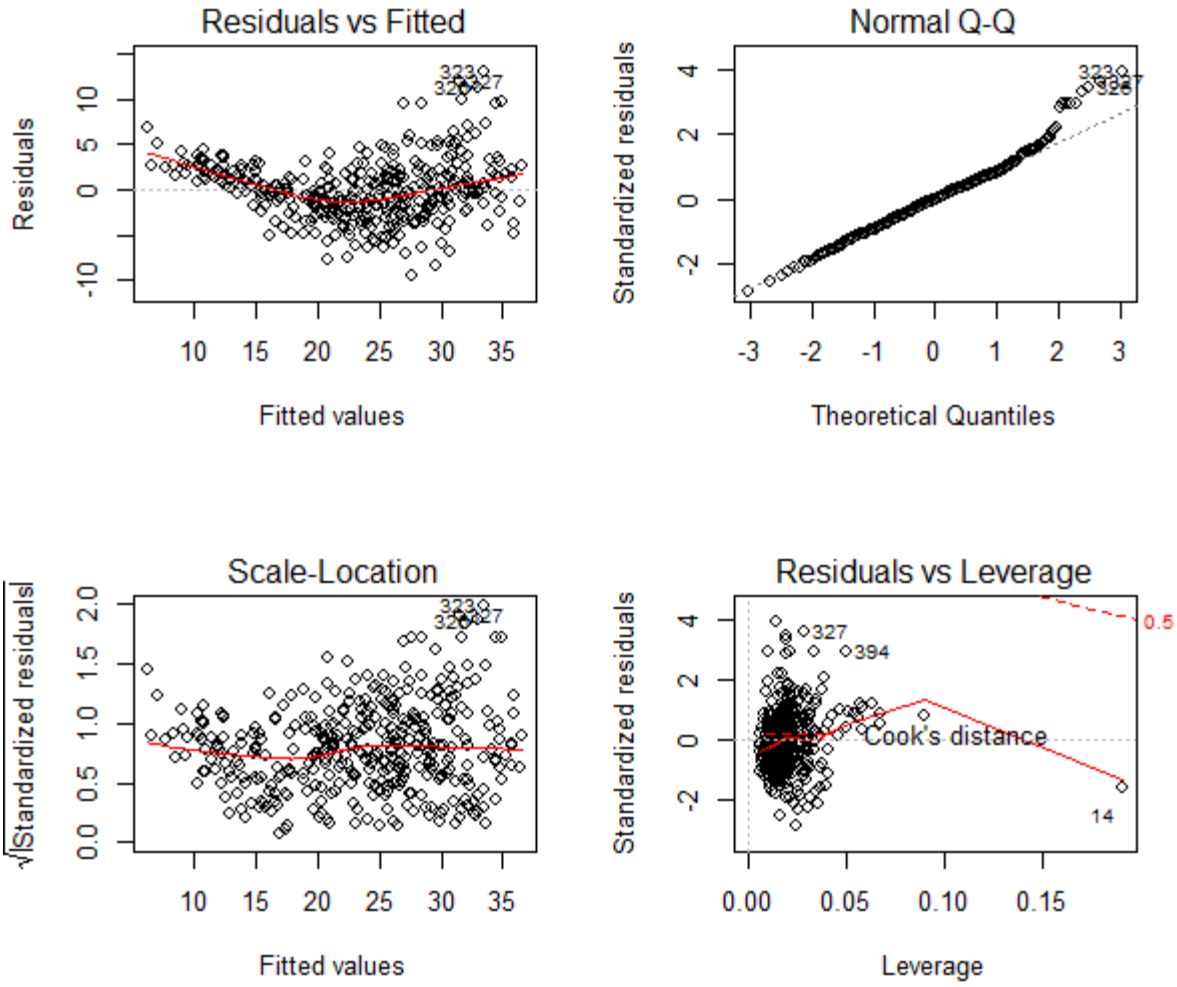
i. Is there a relationship between the predictors and the response?
Yes! Some predictors are more significant than others...

ii. Which predictors appear to have a statistically significant relationship to the response?
Displacement, weight, year, origin (low p-values)

iii. What does the coefficient for the year variable suggest?
Since the coefficient is positive, newer cars have higher mpg.

d. 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 plot identify any observations with unusually high leverage?

```
> plot(lm.fit)
```



There is a strong pattern in the Residuals vs. Fitted plot, which suggests non-linearity in the model. There are outliers, like point 323, but it's not too strong. Point 14 has very high leverage.

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

I look at the interaction between weight~horsepower, acceleration~weight, and horsepower~acceleration.

```
> lm.fit1=lm(mpg~weight*horsepower,data=auto)
> lm.fit2=lm(mpg~weight*acceleration,data=auto)
> lm.fit3=lm(mpg~horsepower*acceleration,data=auto)
```

```
> summary(lm.fit1) #Weight and horsepower

Call:
lm(formula = mpg ~ weight * horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7725  -2.2074  -0.2708   1.9973  14.7314

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.356e+01  2.343e+00  27.127  < 2e-16 ***
weight            -1.077e-02  7.738e-04 -13.921  < 2e-16 ***
horsepower        -2.508e-01  2.728e-02  -9.195  < 2e-16 ***
weight:horsepower  5.355e-05  6.649e-06   8.054 9.93e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.93 on 388 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.7484,      Adjusted R-squared:  0.7465
F-statistic: 384.8 on 3 and 388 DF,  p-value: < 2.2e-16
```

```
> summary(lm.fit2) #Weight and acceleration

Call:
lm(formula = mpg ~ weight * acceleration, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5831  -2.7125  -0.3628   2.3091  15.6577

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.855e+01  4.878e+00   5.854 1.01e-08 ***
weight              -3.254e-03  1.464e-03  -2.222 0.026844 *
acceleration         1.098e+00  3.098e-01   3.544 0.000442 ***
weight:acceleration -2.753e-04  9.704e-05  -2.837 0.004789 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.271 on 393 degrees of freedom
Multiple R-squared:  0.7044,      Adjusted R-squared:  0.7021
F-statistic: 312.1 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
> summary(lm.fit3) #Horsepower and acceleration

Call:
lm(formula = mpg ~ horsepower * acceleration, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3442  -2.7324  -0.4049   2.4210  15.8840

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             33.512440   3.420187   9.798  < 2e-16 ***
horsepower               0.017590   0.027425   0.641 0.521664
acceleration             0.800296   0.211899   3.777 0.000184 ***
horsepower:acceleration -0.015698   0.002003  -7.838 4.45e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.426 on 388 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6809,      Adjusted R-squared:  0.6784
F-statistic: 275.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

All models have statistically significant interaction terms.