Kratika Agrawal
Vandana Anand
DS502 - HW1

QUESTION 1

(a) A flexible model would work better because it would take into account all the data points and try to fit it as accurately as possible. An inflexible model would fail to fit some data points that would lead to less accuracy and subsequently introduce more bias.

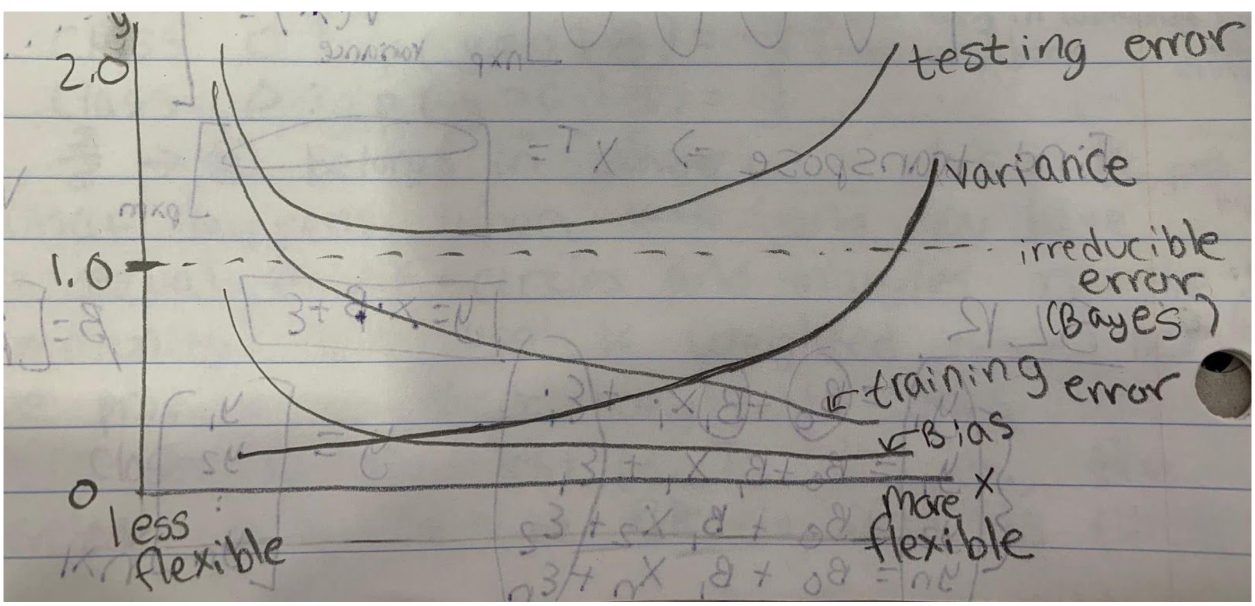
(b) A flexible model would perform worse than the inflexible model because the number of observations is small which can lead to overfitting and higher variance. The flexible model could fit the training data perfectly, but when it comes to use the model on another set of data, it would fail to achieve the same accuracy.

(c) Since nonlinear functions are more flexible and the relationship between the predictors and response if non-linear, it would be best to use a flexible model in order to be able to fit the data more accurately. The flexible model can limit the amount of bias.

(d) The flexible model would perform worse than the inflexible model because it would also take into account the variance of the error terms leading to a worse accuracy. If you already have high variance, an inflexible model would work better so that the error in estimating f could be lower if a different training dataset was given.

QUESTION 2

(a)



(b)

Bias – Due to the bias-variance trade off, as the model becomes more flexible, the bias will become lower.

Variance – Due to the bias-variance trade off, as the model becomes more flexible, the variance will become higher.

Bayes (irreducible error) – This is an independent constant line otherwise known as the irreducible error that no matter how well we estimate f, this error will still exist because no model is perfect.

Training error – The training error becomes lower as the model becomes more flexible because this is the set of data we are initially fitting our model. Since we can fit the data points, the model will be more accurate.

Testing error – The testing error becomes lower and then higher as the model becomes more flexible because of overfitting. A more flexible model has the chance that it fits the training data set accurately, but then may not fit the testing data as accurately.

QUESTION 3

A parametric approach is when we make an assumption about the functional form of f, for example, if it's linear or parabolic. After assuming the functional shape of f, we only have to estimate the coefficients, β0, β1, . . . ,βp instead of an arbitrary f(X) function. The advantage is that the problem of estimating f is brought down to estimating a set of parameters. The disadvantage is that the model that we assume will not truly match the functional form of f. This raises room for more error because if the model we assume is too far from f, the estimate will be worse. A flexible model can be used to fit many different functional forms of f, but this means we will have to estimate a greater number of parameters which is difficult for complex models. This can lead to overfitting, where the models follow the errors too much.

A non-parametric approach is the opposite of a parametric approach in that assumptions are not made about the functional shape of f. Instead, an estimate of f is taken to get as close to the data points as possible without having huge bumps. The advantage is that non-parametric techniques can accurately fit a wider range of possible shapes for f. It also avoids the problem that parametric approaches have in which the estimate of f can be very different from the true f because no assumption about f is made. However, since the non-parametric approach does not reduce the problem of estimating f to a small set of parameters, a large number of observations is needed to get an accurate estimate for f.

QUESTION 4

1. **This exercise relates to the College data set.**

(a) > college = read.csv('College.csv');
(b) > fix(college)

R Data Editor

X	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
Albertus Magnus College	Yes	353	340	103	17	45
Albion College	Yes	1899	1720	489	37	68
Albright College	Yes	1038	839	227	30	63
Alderson-Broaddus College	Yes	582	498	172	21	44
Alfred University	Yes	1732	1425	472	37	75
Allegheny College	Yes	2652	1900	484	44	77
Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64
Alma College	Yes	1267	1080	385	44	73
Alverno College	Yes	494	313	157	23	46
American International College	Yes	1420	1093	220	9	22
Amherst College	Yes	4302	992	418	83	96
Anderson University	Yes	1216	908	423	19	40
Andrews University	Yes	1130	704	322	14	23

```
>rownames(college)=college[,1];
>fix(college)
```

R Data Editor

row.names	X	Private	Apps
Abilene Christian University	Abilene Christian University	Yes	1660
Adelphi University	Adelphi University	Yes	2186
Adrian College	Adrian College	Yes	1428
Agnes Scott College	Agnes Scott College	Yes	417
Alaska Pacific University	Alaska Pacific University	Yes	193
Albertson College	Albertson College	Yes	587
Albertus Magnus College	Albertus Magnus College	Yes	353
Albion College	Albion College	Yes	1899
Albright College	Albright College	Yes	1038
Alderson-Broaddus College	Alderson-Broaddus College	Yes	582
Alfred University	Alfred University	Yes	1732
Allegheny College	Allegheny College	Yes	2652
Allentown Coll. of St. Francis de Sales	Allentown Coll. of St. Francis de Sales	Yes	1179
Alma College	Alma College	Yes	1267
Alverno College	Alverno College	Yes	494
American International College	American International College	Yes	1420
Amherst College	Amherst College	Yes	4302
Anderson University	Anderson University	Yes	1216
Andrews University	Andrews University	Yes	1130

> college=college[,-1]
> fix(college);

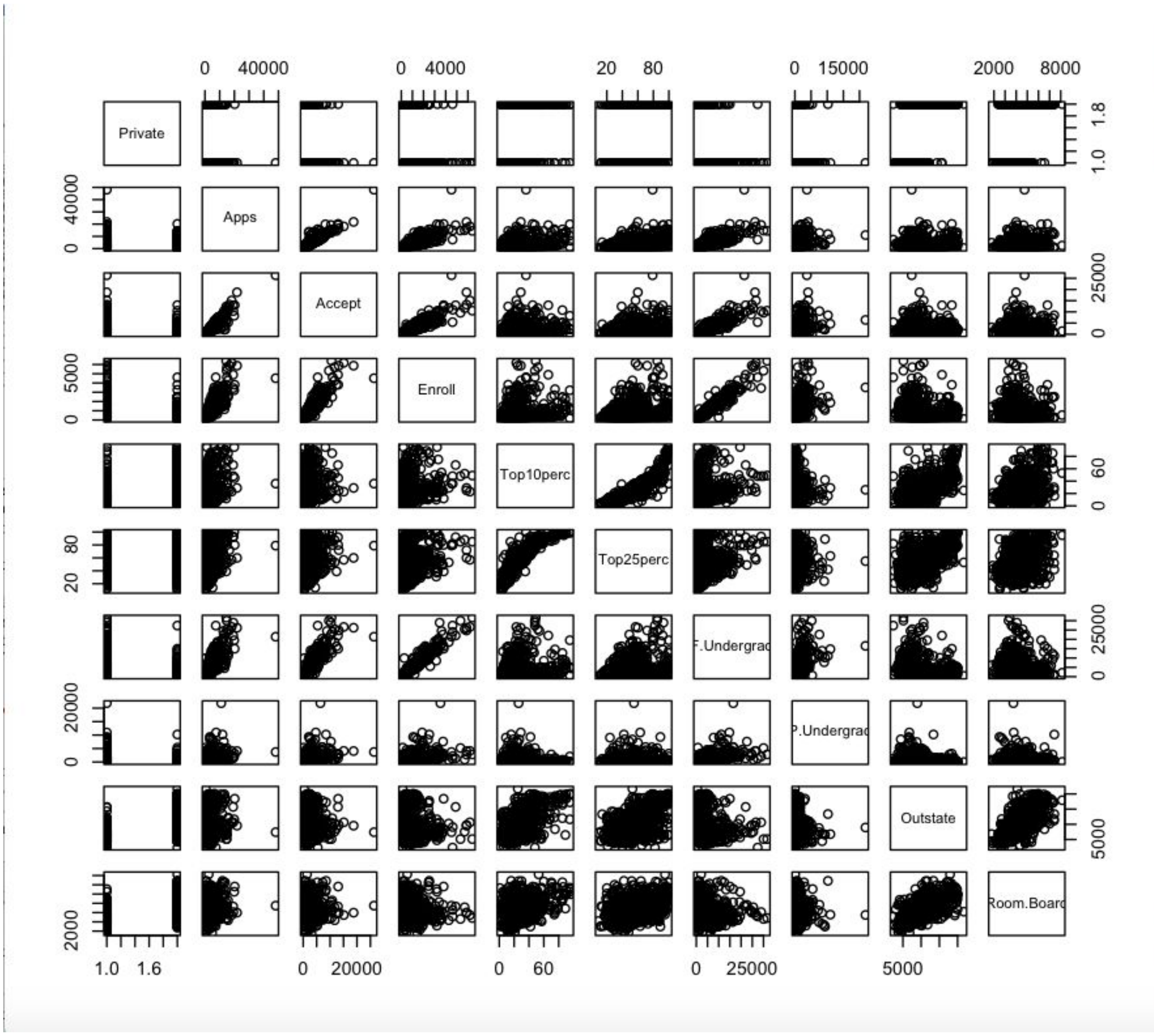
R Data Editor

row.names	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
Albertus Magnus College	Yes	353	340	103	17	45
Albion College	Yes	1899	1720	489	37	68
Albright College	Yes	1038	839	227	30	63
Alderson-Broaddus College	Yes	582	498	172	21	44
Alfred University	Yes	1732	1425	472	37	75
Allegheny College	Yes	2652	1900	484	44	77
Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64
Alma College	Yes	1267	1080	385	44	73
Alverno College	Yes	494	313	157	23	46
American International College	Yes	1420	1093	220	9	22
Amherst College	Yes	4302	992	418	83	96
Anderson University	Yes	1216	908	423	19	40
Andrews University	Yes	1130	704	322	14	23

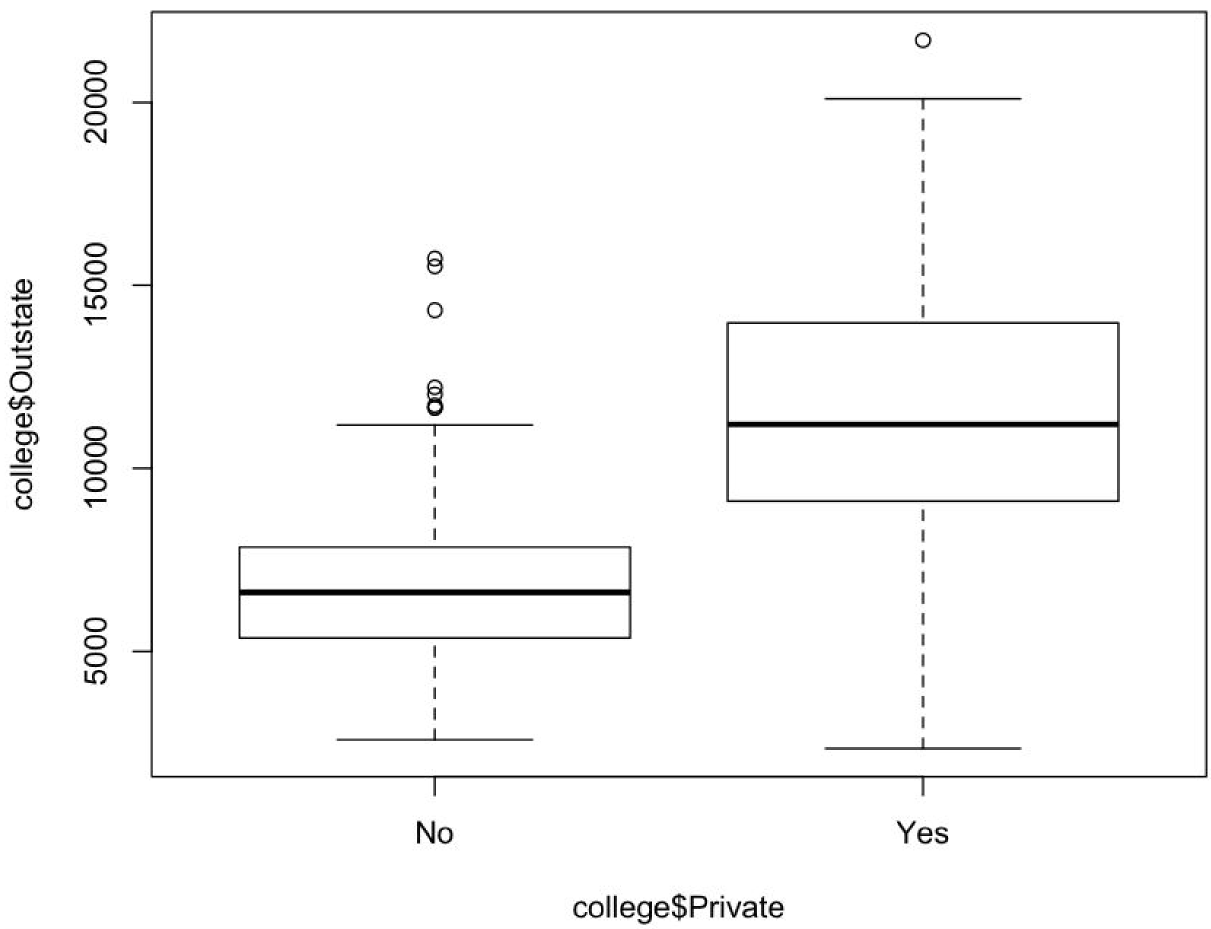
(c)

```
> 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    
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340  
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320  
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990  
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441  
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925  
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700  
   Room.Board       Books           Personal         PhD        
 Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00  
 1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00  
 Median :4200   Median : 500.0   Median :1200   Median : 75.00  
 Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66  
 3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00  
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00  
    Terminal       S.F.Ratio      perc.alumni        Expend     
 Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186  
 1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751  
 Median : 82.0   Median :13.60   Median :21.00   Median : 8377  
 Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660  
 3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830  
 Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233  
   Grad.Rate     
 Min.   : 10.00  
 1st Qu.: 53.00  
 Median : 65.00  
 Mean   : 65.46  
 3rd Qu.: 78.00  
 Max.   :118.00  
```

ii. > pairs(college[,1:10])



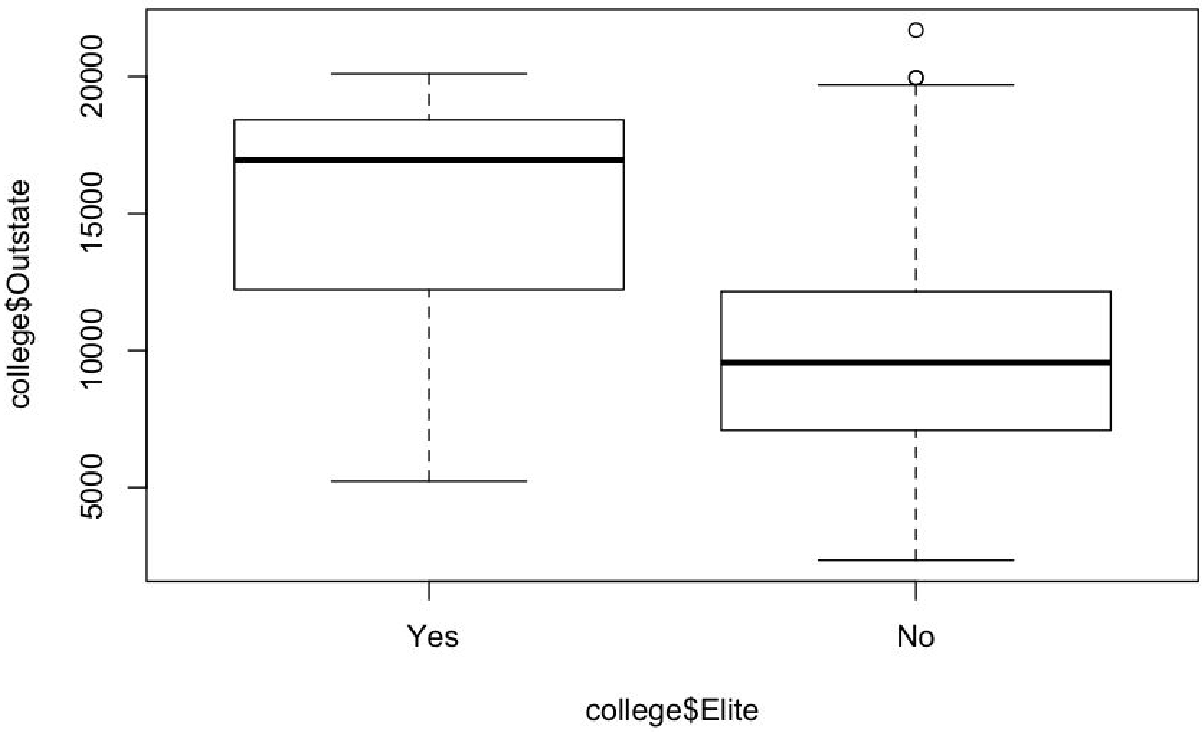
iii. > boxplot(college$Outstate ~ college$Private);



iv.

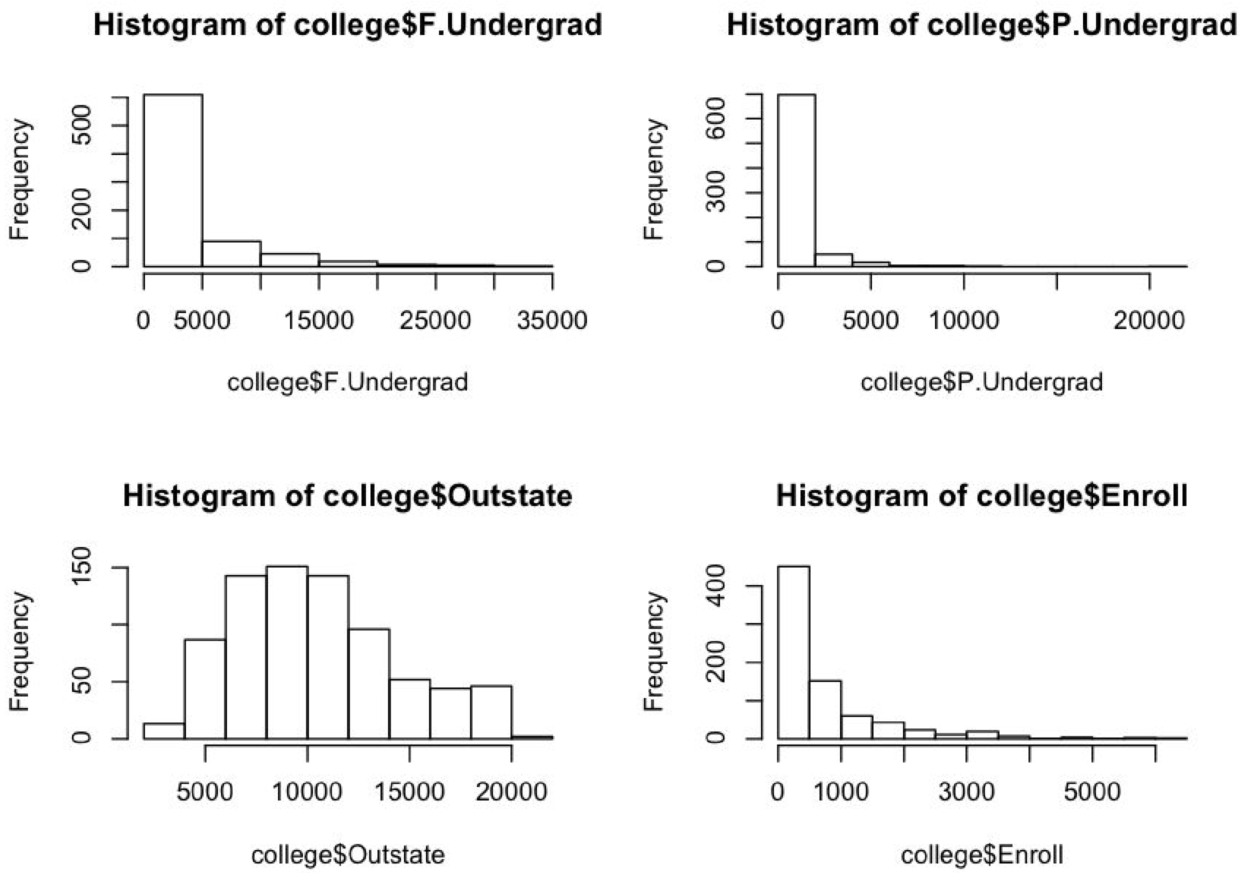
```
> summary(college$Elite)
 Yes   No
  78  699
```

```
> boxplot(college$Outstate ~ college$Elite)
```



v.

```
>  par(mfrow=c(2,2))
> hist(college$F.Undergrad);
> hist(college$P.Undergrad);
> hist(college$Outstate);
> hist(college$Enroll);
```



vi.

It is interesting how most of the people enrolled in college come from out of state. Also, most students are pursuing a full time undergraduate degree rather than part time. In addition, as shown in the picture below, the minimum amount of students accepted into college is 72 and the max was 26330, which seems like a low number considering there are many colleges listed.

```
> summary(college$Accept)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     72     604    1110    2019    2424   26330 
```

QUESTION 5

View of the Auto data:

Data Editor

File Edit Help

	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration
1	1	18	8	307	130	3504	12
2	2	15	8	350	165	3693	11.5
3	3	18	8	318	150	3436	11
4	4	16	8	304	150	3433	12
5	5	17	8	302	140	3449	10.5
6	6	15	8	429	198	4341	10
7	7	14	8	454	220	4354	9
8	8	14	8	440	215	4312	8.5
9	9	14	8	455	225	4425	10
10	10	15	8	390	190	3850	8.5
11	11	15	8	383	170	3563	10
12	12	14	8	340	160	3609	8
13	13	15	8	400	150	3761	9.5
14	14	14	8	455	225	3086	10
15	15	24	4	113	95	2372	15
16	16	22	6	198	95	2833	15.5
17	17	18	6	199	97	2774	15.5
18	18	21	6	200	85	2587	16
19	19	27	4	97	88	2130	14.5

a. quantitative variables : mpg, cylinders, displacement, horsepower, weight, acceleration
qualitative variables : year, origin, name

```
> #answer a
> #quantitative variables : mpg, cylinders, displacement, horsepower, weight, accelearation
> head(auto[,c(1:6),])
  mpg cylinders displacement horsepower weight acceleration
1  18        70          307        130   3504         12.0
2  15        70          350        165   3693         11.5
3  18        70          318        150   3436         11.0
4  16        70          304        150   3433         12.0
5  17        70          302        140   3449         10.5
6  15        70          429        198   4341         10.0
> #qualitative variables : year, origin, name
> head(auto[,c(7:9),])
  year origin                      name
1   70      1 chevrolet chevelle malibu
2   70      1         buick skylark 320
3   70      1        plymouth satellite
4   70      1             amc rebel sst
5   70      1               ford torino
6   70      1          ford galaxie 500
```

b.

```
> #answer b
> sapply(auto[,c(1:6),],range)
      mpg cylinders displacement horsepower weight acceleration
[1,]  9.0         3           68         46   1613          8.0
[2,] 46.6         8          455        230   5140         24.8
```

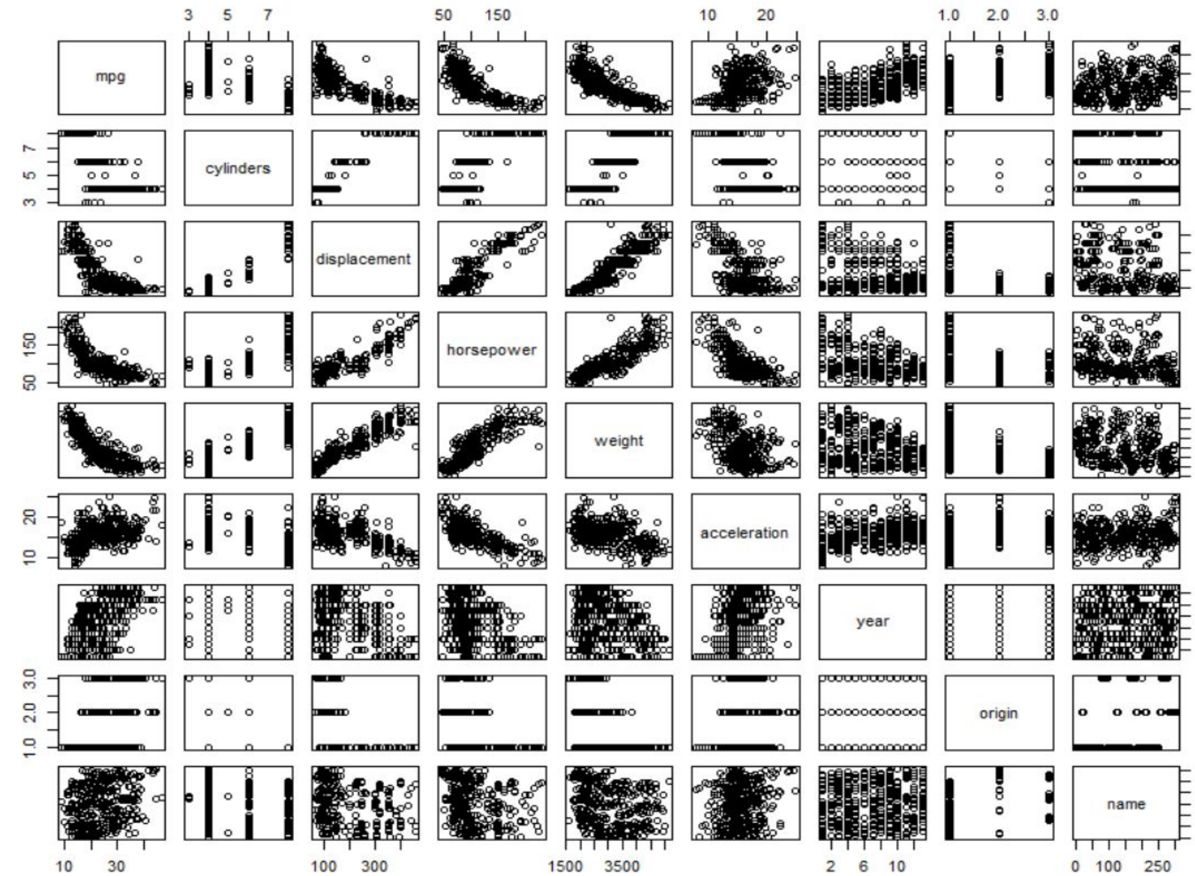
c.

```
> #answer c
> sapply(auto[,c(1:6),],mean)
         mpg    cylinders displacement   horsepower       weight acceleration 
   23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327 
> sapply(auto[,c(1:6),],sd)
         mpg    cylinders displacement   horsepower       weight acceleration 
    7.805007     1.705783   104.644004    38.491160   849.402560     2.758864 
> 
```

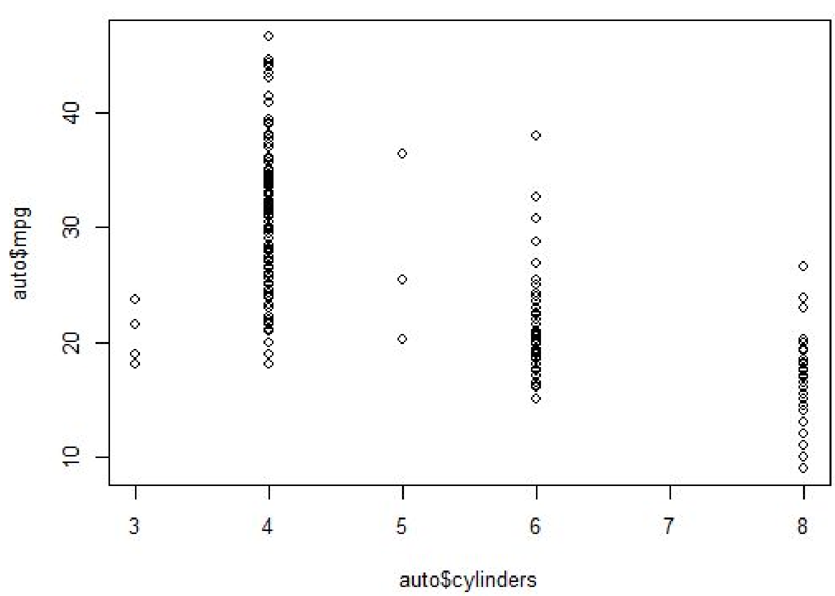
d.

```
> #answer d
> auto_X=auto[-c(10:85),]
> fix(auto_X)
> sapply(auto_X[,c(1:6),],range)
      mpg cylinders displacement horsepower weight acceleration
[1,] 11.0         3           68         46   1649          8.5
[2,] 46.6         8          455        230   4997         24.8
> sapply(auto_X[,c(1:6),],mean)
         mpg    cylinders displacement   horsepower       weight acceleration 
   24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899 
> sapply(auto_X[,c(1:6),],sd)
         mpg    cylinders displacement   horsepower       weight acceleration 
    7.867283     1.654179    99.678367    35.708853   811.300208     2.693721 
```

e. pairs(auto)

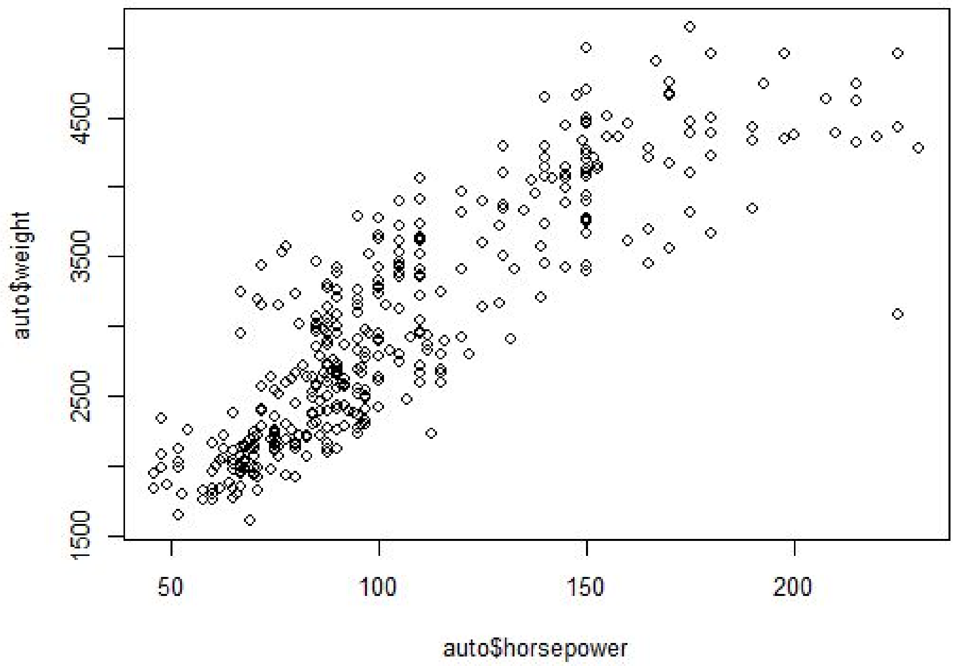


Plot of all predictors provides a clear picture of their relationship with one another.
plot(auto$cylinders,auto$mpg)

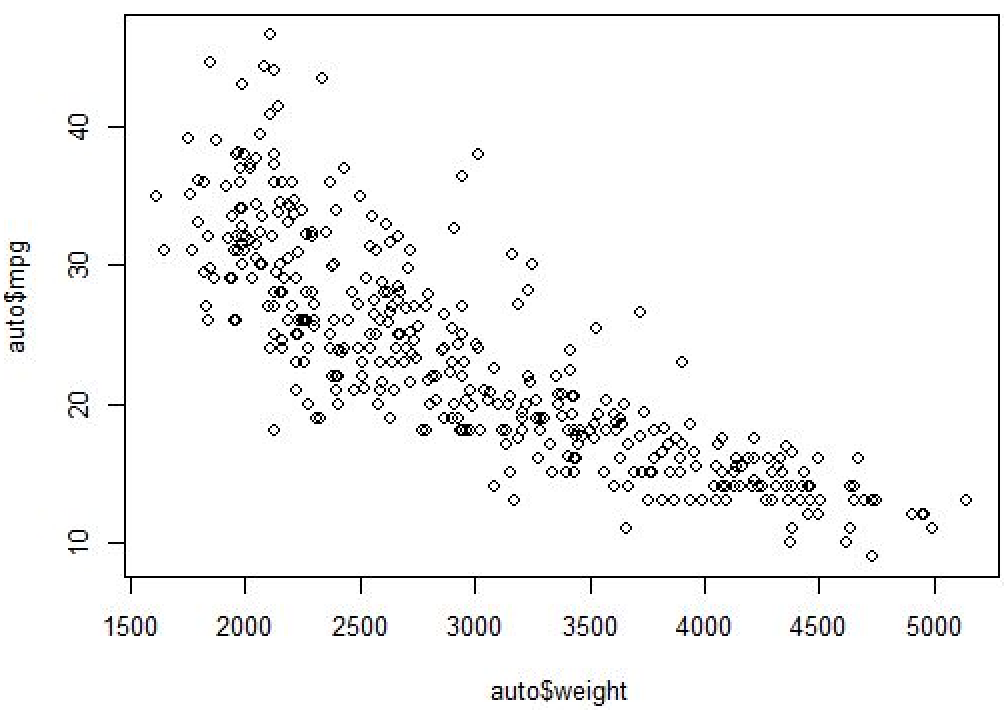


mpg is inversely related to the number of cylinders.

```
plot(auto$horsepower,auto$weight)
```




```
plot(auto$weight,auto$mpg)
```



Displacement, horsepower, weight seem to have proportional relationships with each other, however inverse relation with mpg.

f. We can see that mpg is inversely related with displacement, horsepower and weight. However with acceleration, year and origin, it increases at first and then becomes independent of these parameters, while it continues to decrease with further increase in number of cylinders.

QUESTION 6

In order to describe if there is a relationship between the response and the predictors, in this case it is sales and TV, Radio, Newspaper, respectively, we need to check if B1 = 0. In the case of multiple linear regression, the null hypothesis is that none of TV, Radio, and Newspaper are related to sales and the alternative hypothesis is that at least one of those is related to sales. The p value will tell us whether or not to reject the null hypothesis. For TV and Radio, the p value is close to zero so there is strong evidence that these two variables are related to sales. Newspaper has a very high p value so there is no evidence that it is associated with sales in the presence of the TV and Radio variables. For example, if Radio and Newspaper are held constant and TV advertising is increased, it will very likely lead to an increase in sales because the TV p value is small. However, the Newspaper will likely not have any effect on sales if TV and Radio are held constant because the p value is large.

QUESTION 7

$\hat{y}_i = x_i \hat{B}_j$ $\quad \hat{B} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)}{\left(\sum_{i=1}^{n} x_{i'}^2\right)}$ $\quad$ Show: $\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$

$$\hat{y}_i = x_i \cdot \frac{\left(\sum_{i'=1}^{n} x_{i'} y_{i'}\right)}{\sum_{i'=1}^{n} x_{i'}^2} = \sum_{i'=1}^{n} \frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2} y_{i'} = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

$$\boxed{a_{i'} = \frac{x_i x_{i'}}{\sum_{j=1}^{n} x_j^2}}$$

QUESTION 8

$y_i = b_0 + b_1 x_i \qquad (\bar{x}, \bar{y}) \qquad \bar{x} = \frac{1}{n} \sum x_i$

$y_i = b_0 + b_1 \cdot \frac{\sum x_i}{n} \qquad (x_1, y_1)$

$y_i \cdot n = n \cdot b_0 + b_1 \cdot \sum x_i \qquad (x_2, y_2)$

$y_i \cdot n = \sum_{i=1}^{n} b_0 + b_1 \sum_{i=1}^{n} x_i \qquad \vdots$

$\frac{y_i \cdot n}{n} = \frac{\sum_{i=1}^{n} \underbrace{b_0 + b_1 x_i}_{y_i}}{n} \qquad (x_n, y_n)$

$y_i = \frac{\sum y_i}{n} = \bar{y}$ ✓

y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_n	x_n

QUESTION 9

View of the Auto data:

Data Editor

File Edit Help

	row.names	mpg	cylinders	displacement	horsepower	weight	acceleration
1	1	18	8	307	130	3504	12
2	2	15	8	350	165	3693	11.5
3	3	18	8	318	150	3436	11
4	4	16	8	304	150	3433	12
5	5	17	8	302	140	3449	10.5
6	6	15	8	429	198	4341	10
7	7	14	8	454	220	4354	9
8	8	14	8	440	215	4312	8.5
9	9	14	8	455	225	4425	10
10	10	15	8	390	190	3850	8.5
11	11	15	8	383	170	3563	10
12	12	14	8	340	160	3609	8
13	13	15	8	400	150	3761	9.5
14	14	14	8	455	225	3086	10
15	15	24	4	113	95	2372	15
16	16	22	6	198	95	2833	15.5
17	17	18	6	199	97	2774	15.5
18	18	21	6	200	85	2587	16
19	19	27	4	97	88	2130	14.5

a.

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

i. Since the p-value is extremely small, i.e. <2e-16, the confidence interval is very high. Thus, we can reject the null hypothesis as β1 not equal to 0 and can say that a relationship exists between horsepower and mpg.

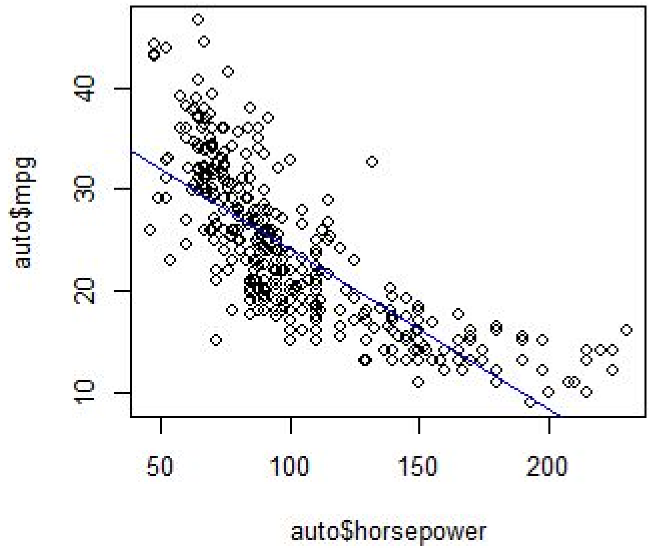
ii. The value of R-squared is 0.6059, i.e. 60.59% of variation in the model is explained by linear regression. Therefore, we can say that there is a strong relation between horsepower and mpg.

iii. As the value of horsepower coefficient is -0.157845, which is negative, the relation between horsepower and mpg is negative linear relation, i.e. with increase in horsepower value, mpg decreases.

iv.

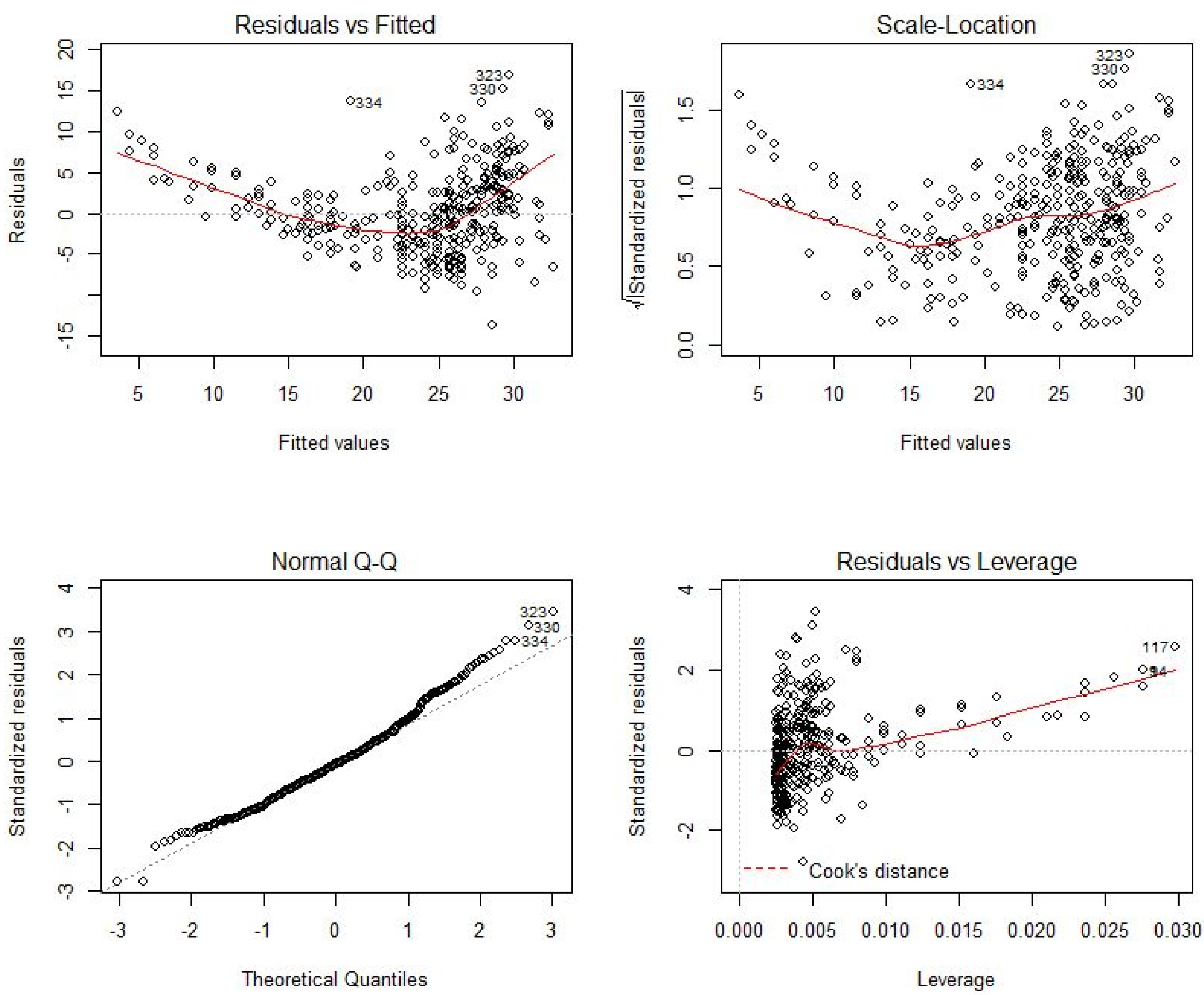
```
> #a.iv
> predict(lm.fit ,data.frame(horsepower=98),interval ="confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(lm.fit ,data.frame(horsepower=98),interval ="prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

b.



auto$mpg
auto$horsepower
10
20
30
40
50
100
150
200

c.



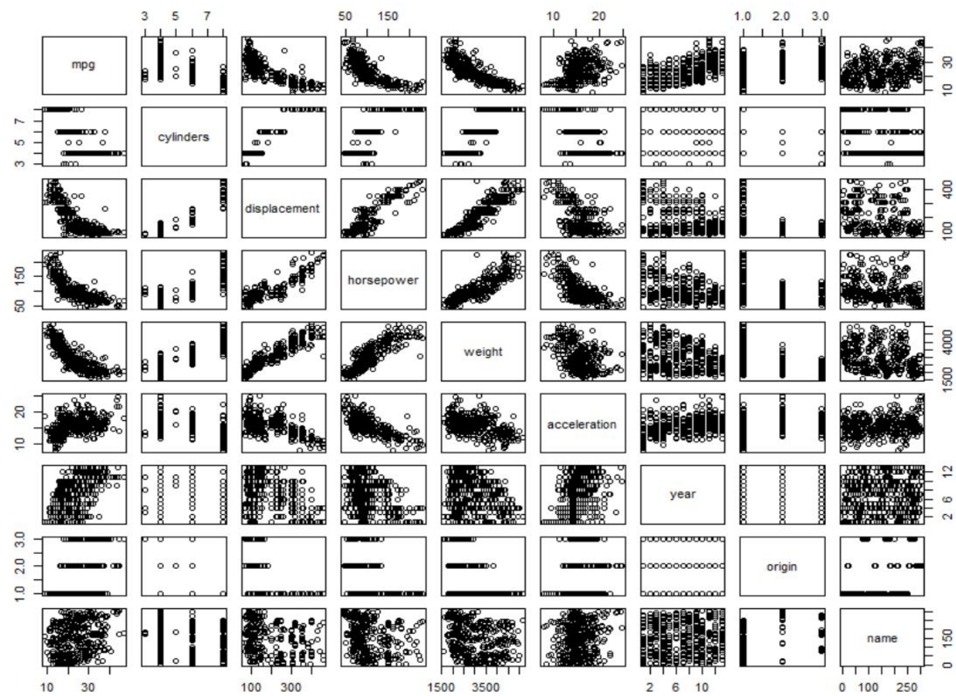
Residuals vs Fitted
Residuals
Fitted values
334
323
330
Scale-Location
√|Standardized residuals|
Fitted values
334
323
330
Normal Q-Q
Standardized residuals
Theoretical Quantiles
323
330
334
Residuals vs Leverage
Standardized residuals
Leverage
117
94
Cook's distance

Plot of least square regression fit suggest that there exists a linear relationship between horsepower and mpg, however the relation is not perfectly linear and consists of few non-linearities.

QUESTION 10

a.

```
> #answer a.
> pairs(auto)
```



b.

```
> cor(auto[,!(names(auto)=="name")])
                    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.

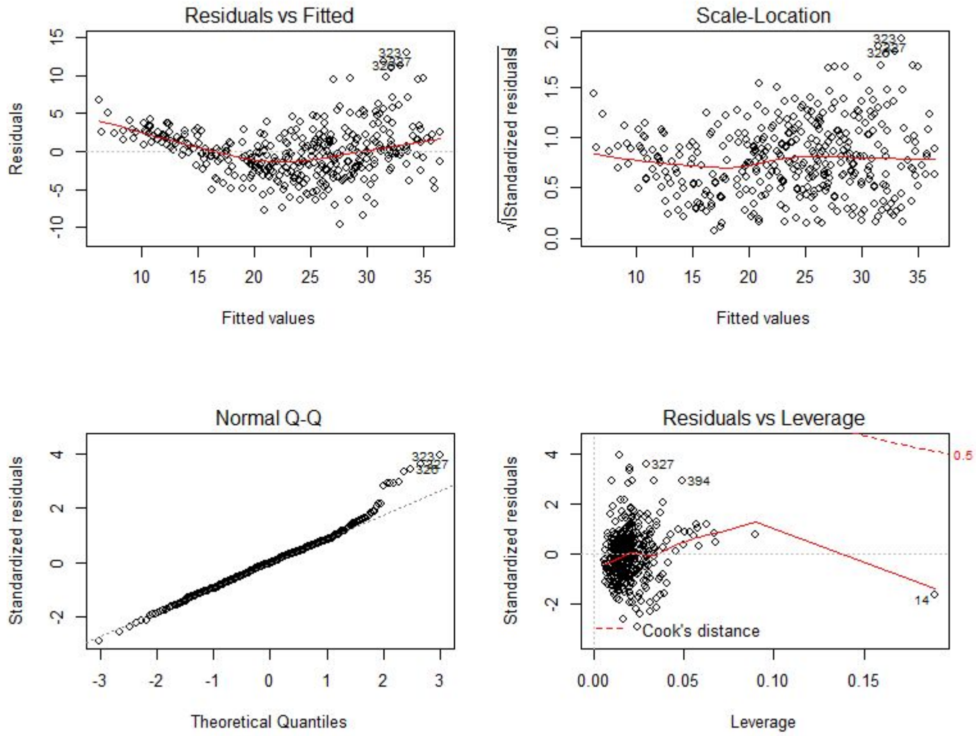
```
> #answer c.
> lm.fit = lm(formula= auto$mpg ~.,data=auto[,!(names(auto)=="name")])
> summary(lm.fit)

Call:
lm(formula = auto$mpg ~ ., data = auto[, !(names(auto) == "name")])

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)  34.584880   2.245452  15.402  < 2e-16 ***
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
```

i. Overall p-value is very small, i.e. 2.2e-16, which shows that there exists a relation between various predictors and mpg

ii. The p-values for predictors displacement, weight, year and origin are less that 0.05, thus they have statistically significant relationship to the response.

iii. Coefficient for the year is 0.750773, which shows that with an increase of each year, mpg is estimated to increase by 0.75.

d.



Residuals vs Fitted plot shows that there exists some non-linearilty in the data and there are some outliers in the plot as shown in Scale-location plot.

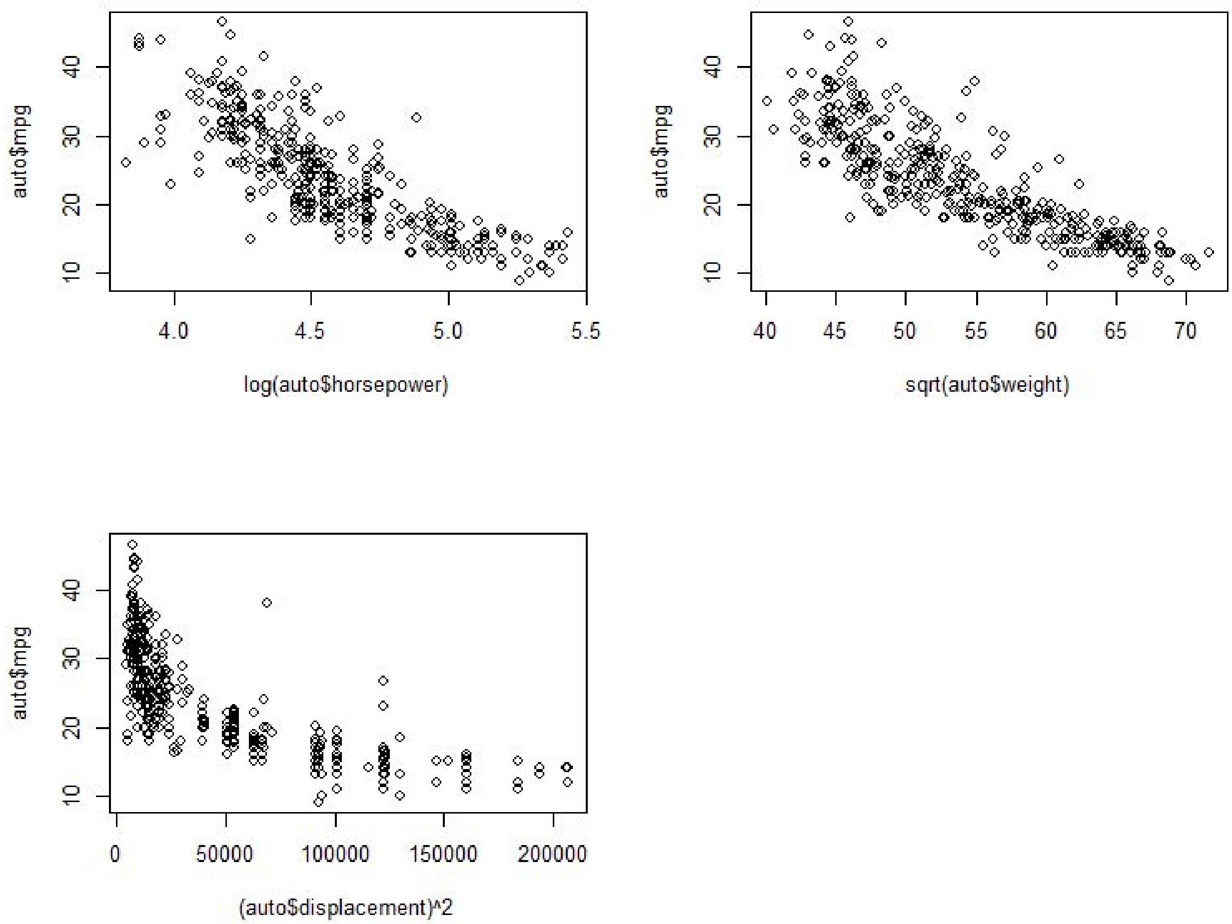
e.

```
> summary(lm.fit_interation)

Call:
lm(formula = auto$mpg ~ auto$cylinders * auto$displacement +
    auto$displacement * auto$weight, data = auto[, !(names(auto) ==
    "name")])

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 ***
auto$cylinders                    7.606e-01  7.669e-01   0.992    0.322
auto$displacement                -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
auto$weight                      -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
auto$cylinders:auto$displacement -2.986e-03  3.426e-03  -0.872    0.384
auto$displacement:auto$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
```

Interaction with weight and displacement improved the relation as the p-value is smaller with its introduction.

f.

```
> #answer f
> par(mfrow = c(2, 2))
> plot(log(auto$horsepower), auto$mpg)
> plot(sqrt(auto$weight), auto$mpg)
> plot((auto$displacement)^2, auto$mpg)
```



log and sqrt term fit the linear model well as compared to squared function