Homework 1 – Write Up

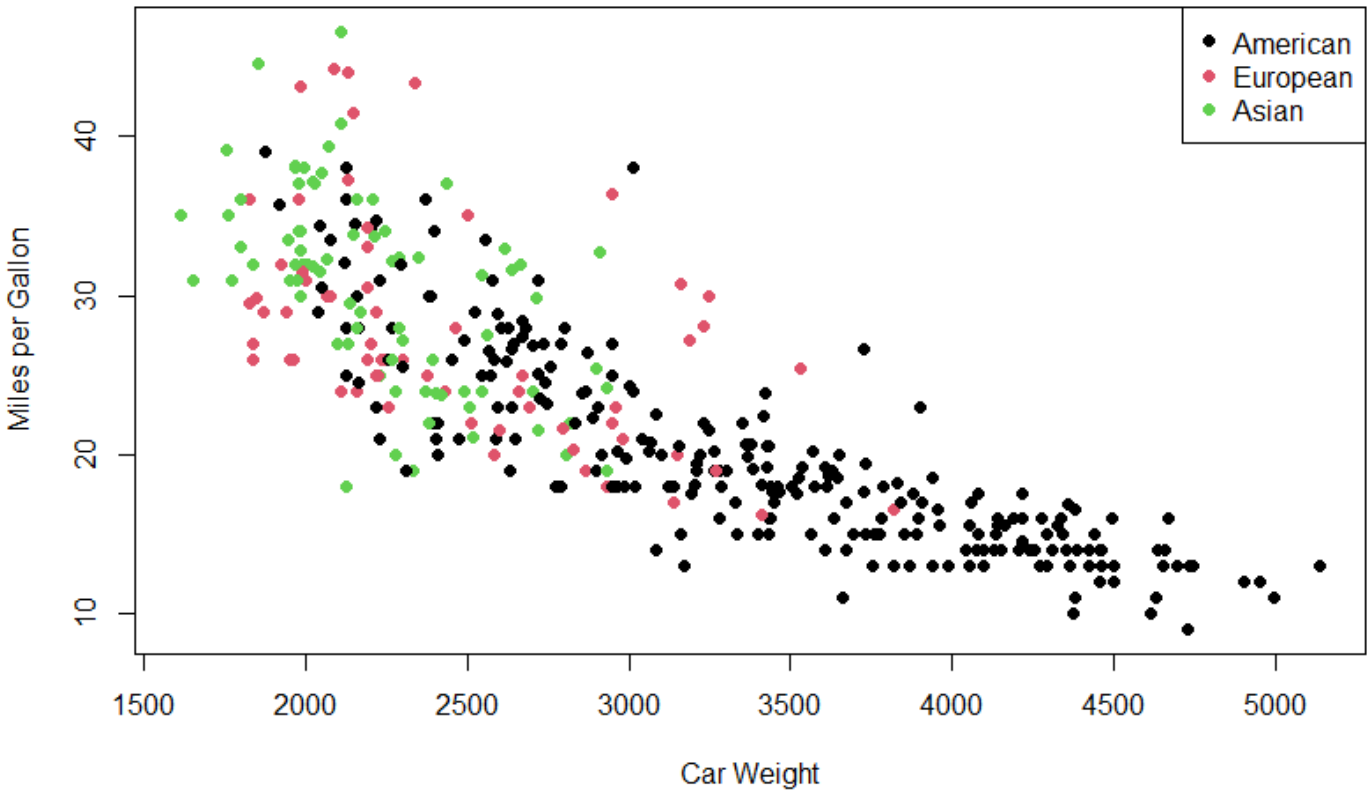
ROHITH REDDY KOLLA – rkolla@buffalo.edu

1.

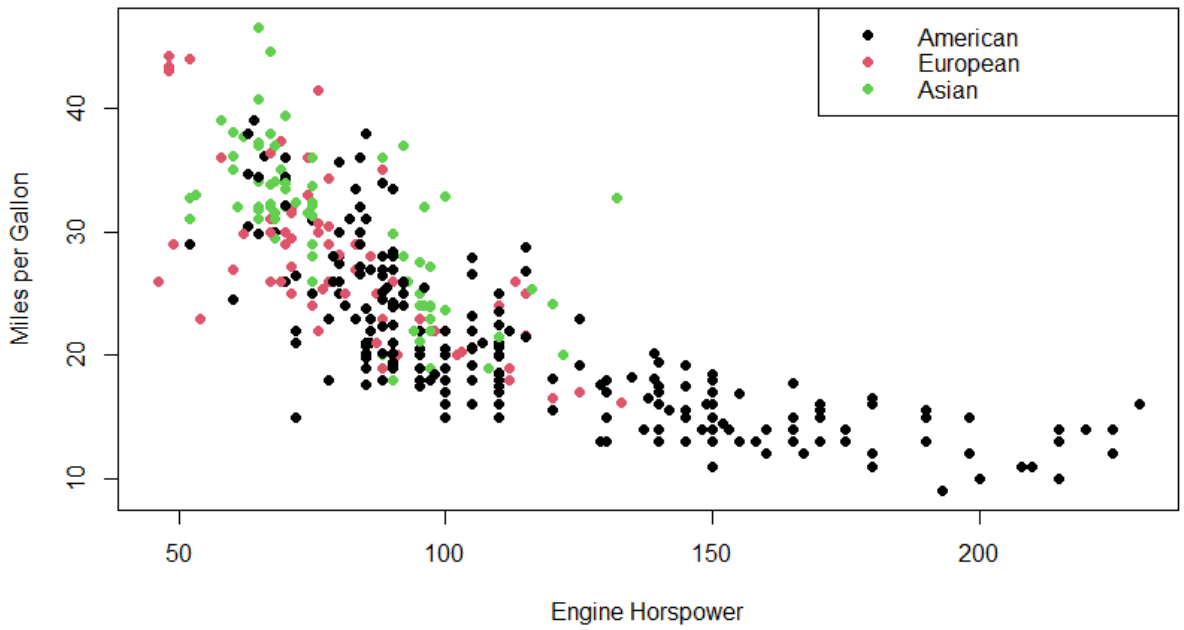
- The Auto dataset from the ISLR package contains 9 columns and 392 entries.

```
> # Finding details about the Auto dataset
> dim(Auto)
[1] 392   9
> head(Auto)
  mpg cylinders displacement horsepower weight acceleration year origin                       name
1  18         8          307        130   3504         12.0   70      1 chevrolet chevelle malibu
2  15         8          350        165   3693         11.5   70      1         buick skylark 320
3  18         8          318        150   3436         11.0   70      1        plymouth satellite
4  16         8          304        150   3433         12.0   70      1             amc rebel sst
5  17         8          302        140   3449         10.5   70      1               ford torino
6  15         8          429        198   4341         10.0   70      1          ford galaxie 500
```

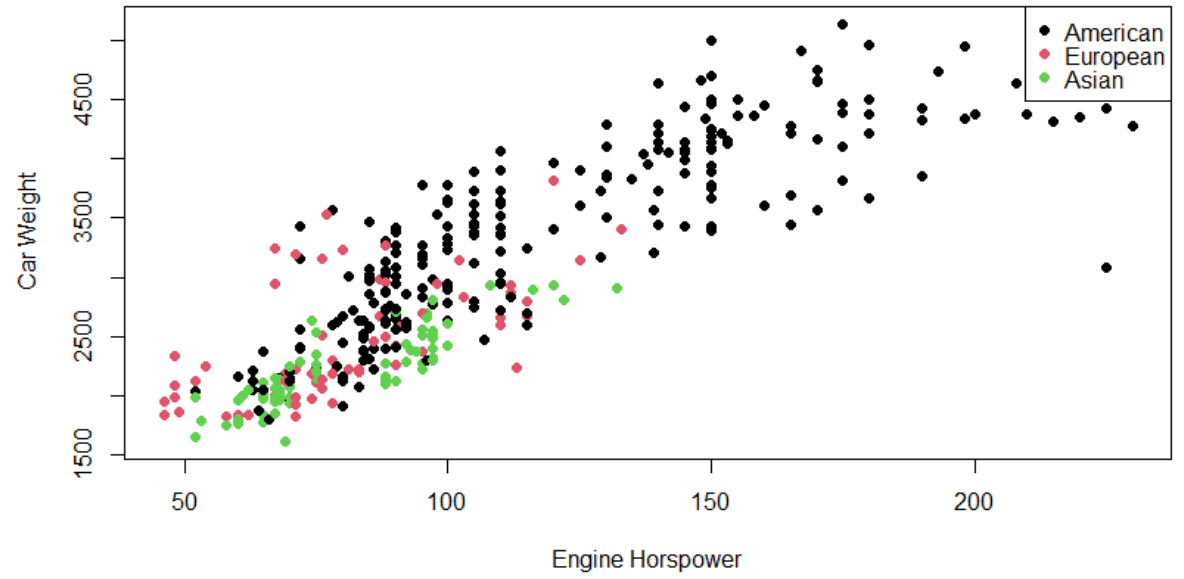
The origin column seems to indicate the continent of origin based on the names of the cars. The rest of the columns are fairly straightforward.



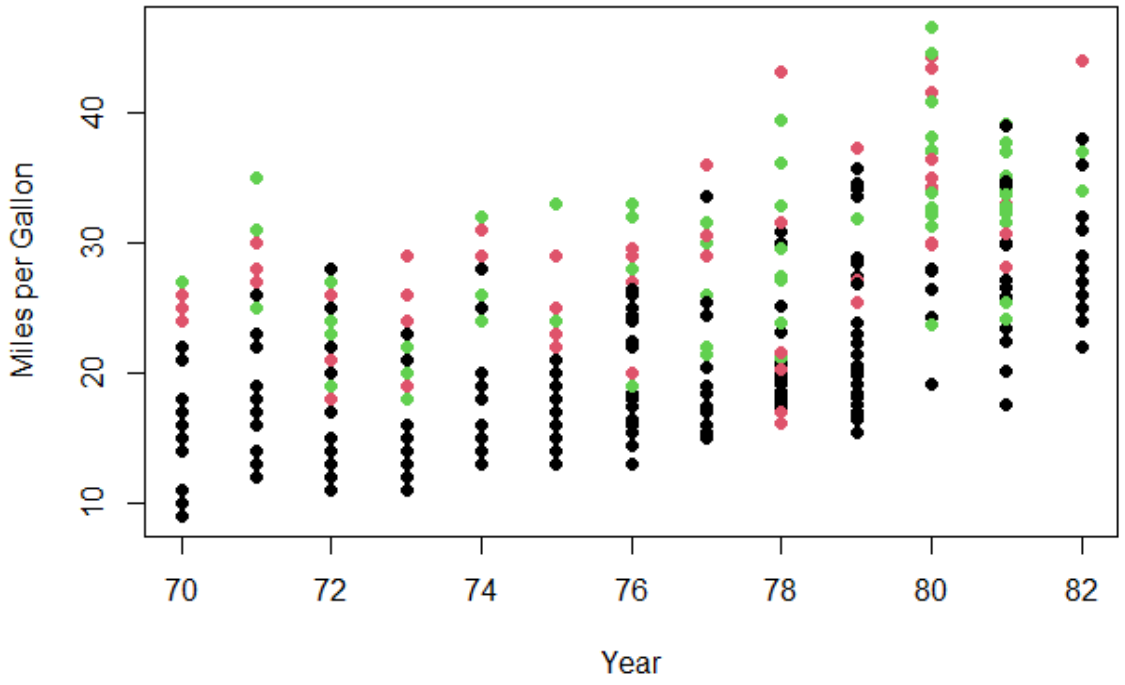
The above scatterplot is indicative of the direct impact of a vehicle's weight on its fuel efficiency. Objects with greater weight have more inertia and would require a greater force to accelerate.



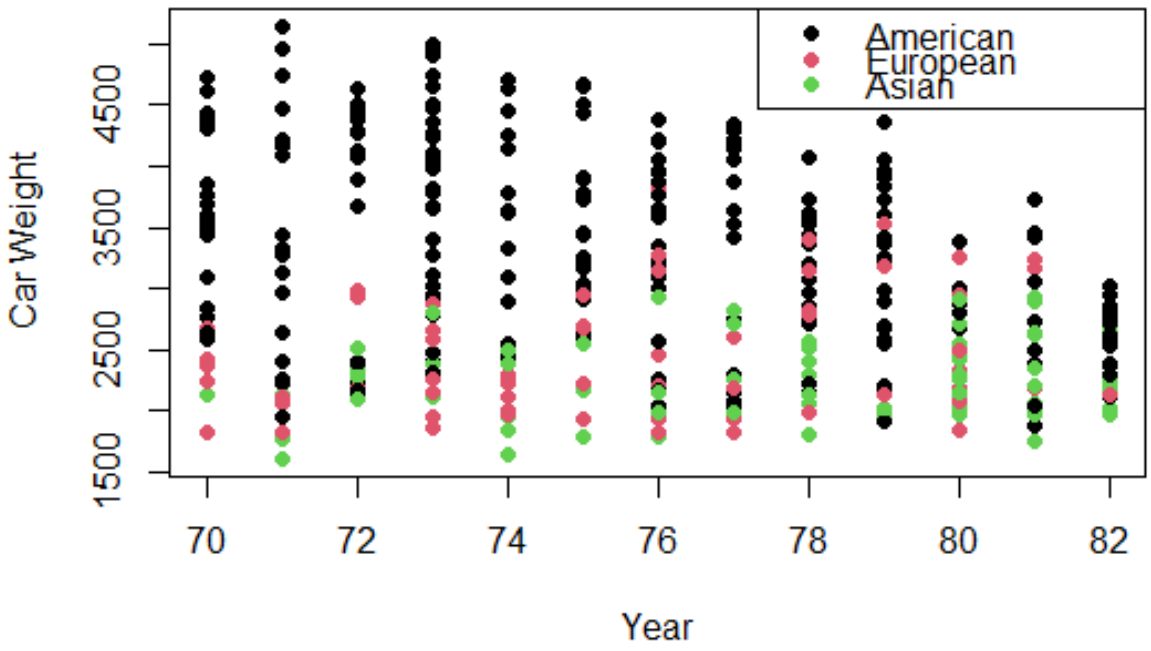
The above scatterplot paints a similar picture to the previous one. It is well understood that engine designs have to trade between power and fuel efficiency based on their purpose. Powerful engines require more fuel to produce that power, the effects of which are quickly countered by friction and air resistance.

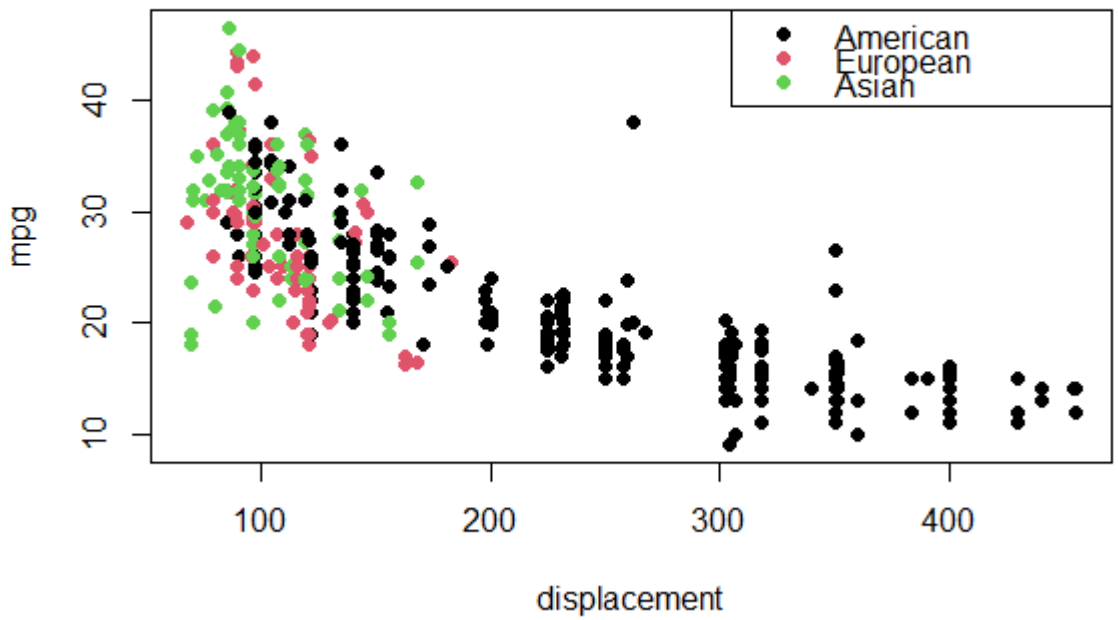


The above scatterplot confirms that Americans like their cars and engines big. More importantly, it also shows the relationship between the two. Powerful engines are often bigger and heavier than their less powerful counterparts requiring a larger vehicle to use it. On the other hand, larger vehicles need engines with greater power which add extra weight to the vehicle.

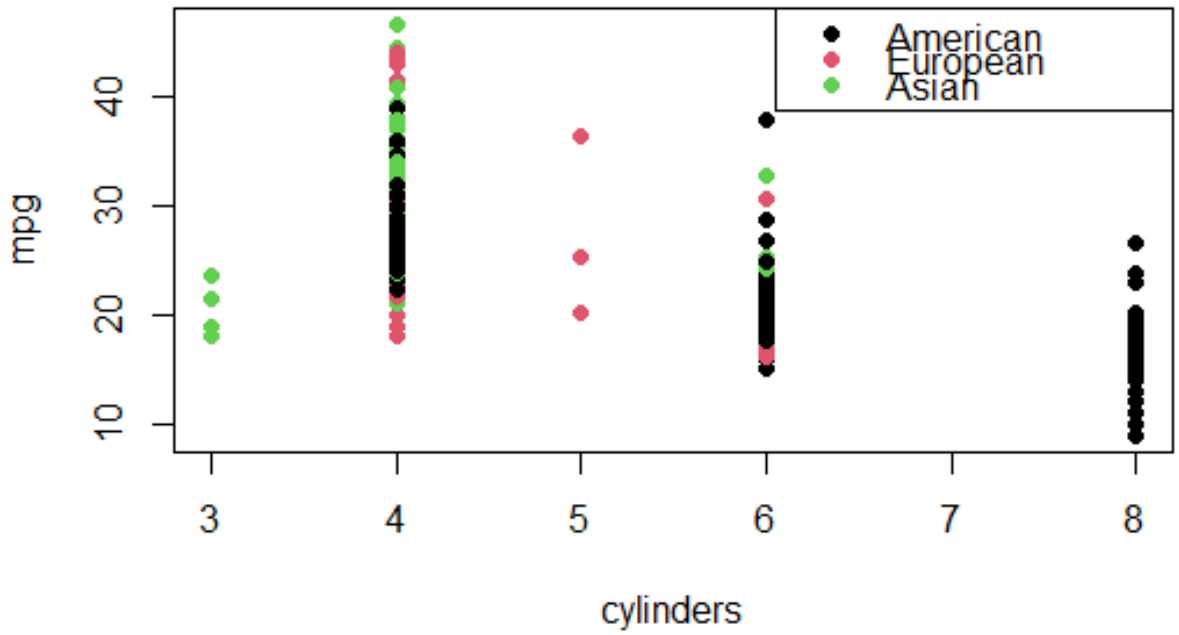


The above scatterplot shows an obvious trend of increasing fuel efficiency with time. This could be attributed to constant improvements in engine and fuel technology along with material sciences which helped develop stronger, cheaper and most importantly lighter materials for use in vehicle manufacture. The scatterplot below shows the decrease in the car weight over time.

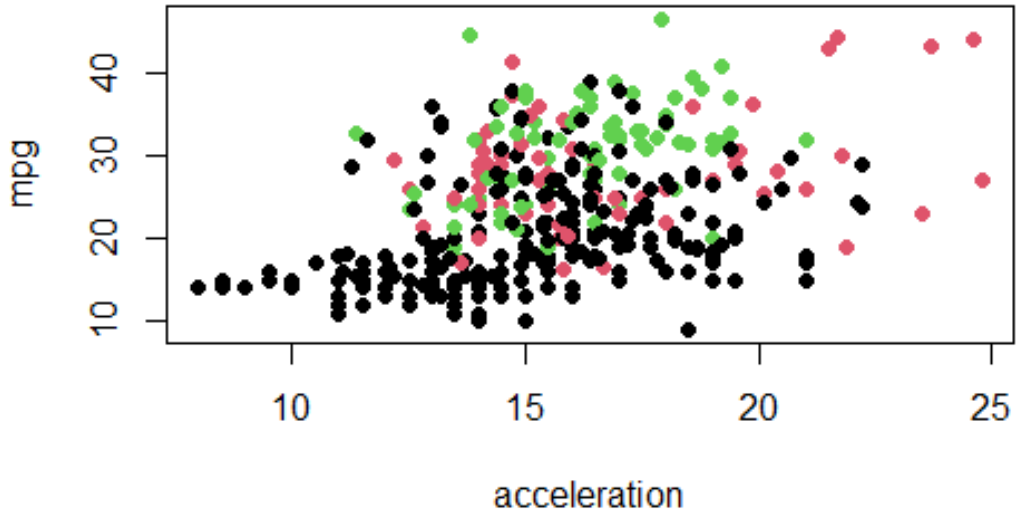




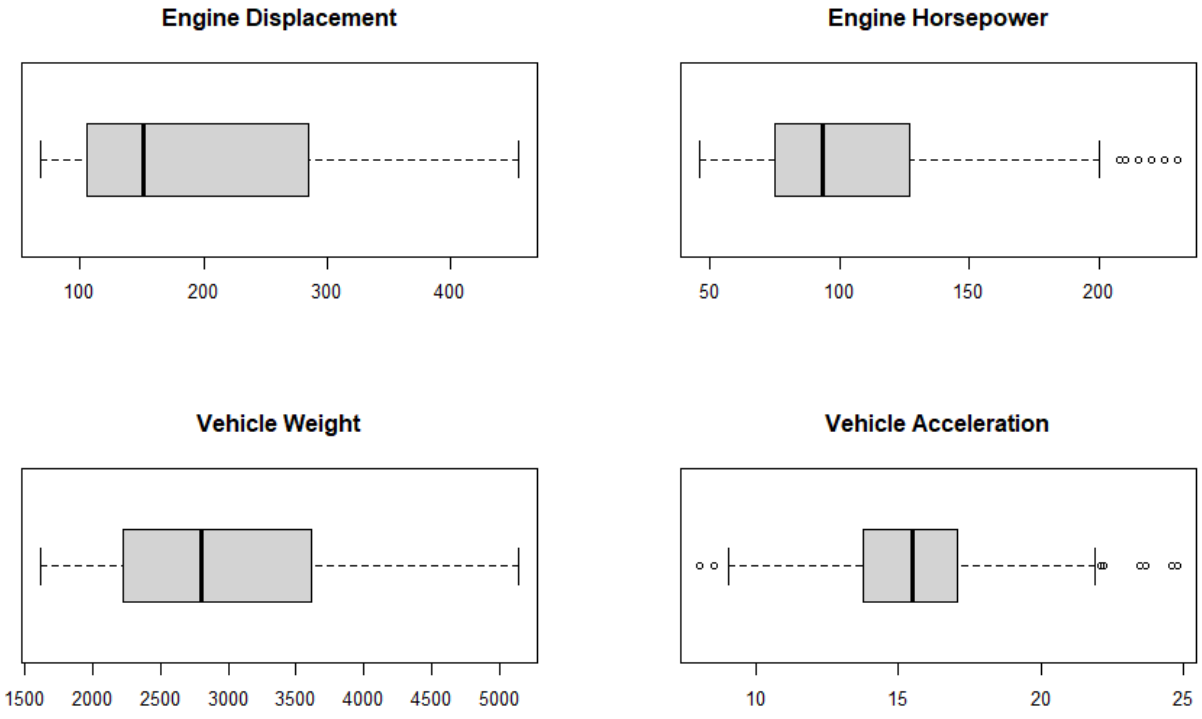
The relationship between the displacement of the engine and the mpg are clearly apparent in the above scatterplot. The exact relationship is quite complex and depends on other variables but the correlation is evident.



Fewer cylinders require lesser fuel intake resulting to better fuel efficiency but sacrifice engine power.



Finally, the acceleration of a vehicle depends majorly on engine power and weight of the vehicle. Heavier cars naturally have lower acceleration and mpg.



Outliers are present in acceleration and horsepower but removal is not necessary
Pre-Processing data for multiple regression –
Name column is removed. Furthermore, Origin column is removed because origin of manufacture is not a factor that directly affects the fuel efficiency of the vehicle.

2.

Summary of the model after loading and linear model fitting the pre-processed data –

```
Call:
lm(formula = mpg ~ ., data = pr_auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6927 -2.3864 -0.0801  2.0291 14.3607

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.454e+01  4.764e+00  -3.051  0.00244 **
cylinders    -3.299e-01  3.321e-01  -0.993  0.32122
displacement  7.678e-03  7.358e-03   1.044  0.29733
horsepower   -3.914e-04  1.384e-02  -0.028  0.97745
weight       -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
acceleration  8.527e-02  1.020e-01   0.836  0.40383
year          7.534e-01  5.262e-02  14.318  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.435 on 385 degrees of freedom
Multiple R-squared:  0.8093,    Adjusted R-squared:  0.8063
F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16
```

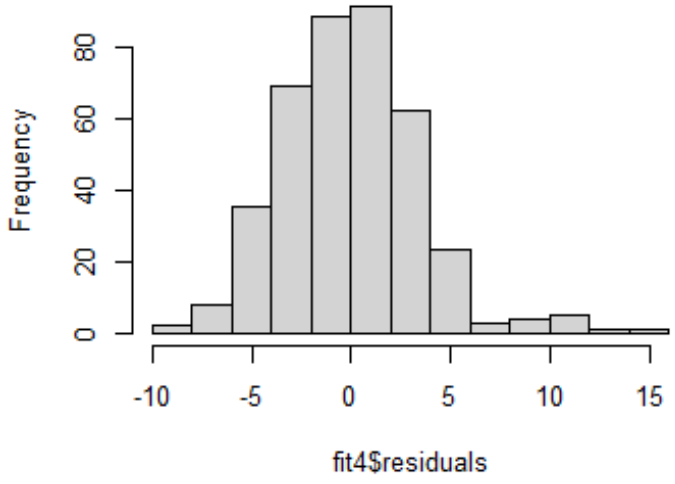
Considering 0.05 as the cutoff for significance excludes cylinders, displacement, horsepower and acceleration.
Removing acceleration based on its seemingly uncorrelated nature as seen in 1. and refitting the model.

```
> fit2 <- lm(mpg~.-acceleration, data = pr_auto)
> summary(fit2)

Call:
lm(formula = mpg ~ . - acceleration, data = pr_auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8714 -2.3852 -0.0895  2.0971 14.4267

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.278e+01  4.274e+00  -2.990  0.00297 **
cylinders    -3.437e-01  3.316e-01  -1.037  0.30058
displacement  6.996e-03  7.310e-03   0.957  0.33908
horsepower   -7.715e-03  1.070e-02  -0.721  0.47149
weight       -6.524e-03  5.866e-04 -11.122  < 2e-16 ***
year          7.499e-01  5.244e-02  14.302  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.434 on 386 degrees of freedom
Multiple R-squared:  0.8089,    Adjusted R-squared:  0.8064
F-statistic: 326.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

These changes resulted in a very slight improvement of adjusted R-squared. Further removing cylinders and displacement, acceleration and re-fitting the model.

```
> summary(fit4)

Call:
lm(formula = mpg ~ . - acceleration - displacement - cylinders -
    horsepower, data = pr_auto)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8505 -2.3014 -0.1167  2.0367 14.3555

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.435e+01  4.007e+00  -3.581 0.000386 ***
weight      -6.632e-03  2.146e-04 -30.911  < 2e-16 ***
year         7.573e-01  4.947e-02  15.308  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.427 on 389 degrees of freedom
Multiple R-squared:  0.8082,    Adjusted R-squared:  0.8072
F-statistic: 819.5 on 2 and 389 DF,  p-value: < 2.2e-16
```

Again, a slight improvement in adjusted R-squared.



The symmetry of the residuals about 0 further indicates a good fitting model.

a) Weight and Year seem to be the predictors that have a significant relationship to the response.

b) The significantly higher coefficient variable for Year as compared to Weight suggests that Weight is a better and more important predictor than Year.

Fitting model using year and horsepower*weight -

```
> fit6 <- lm(mpg~ horsepower*weight + year, data = pr_auto)
> summary(fit6)

Call:
lm(formula = mpg ~ horsepower * weight + year, data = pr_auto)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9146 -1.8987 -0.0386  1.5536 12.6333

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.577e+00  3.911e+00   0.915    0.361
horsepower        -2.236e-01  2.063e-02 -10.837   <2e-16 ***
weight            -1.185e-02  5.868e-04 -20.198   <2e-16 ***
year               7.749e-01  4.508e-02  17.190   <2e-16 ***
horsepower:weight  5.790e-05  5.020e-06  11.534   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.963 on 387 degrees of freedom
Multiple R-squared:  0.8574,    Adjusted R-squared:  0.8559
F-statistic: 581.5 on 4 and 387 DF,  p-value: < 2.2e-16
```

This model shows the best results so far with a significantly better adjusted R-squared

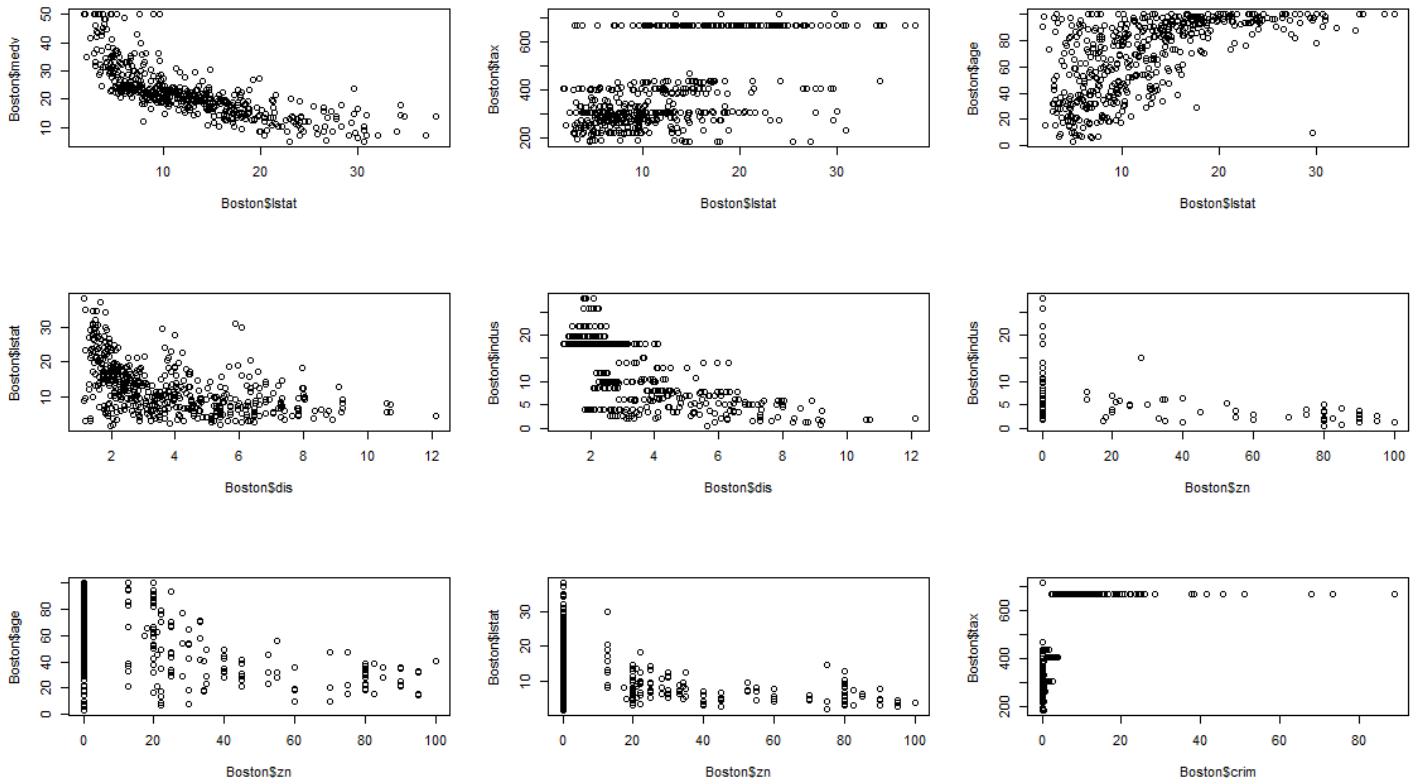
Fitting model using year and displacement*weight –

```
> fit7 <- lm(mpg~ displacement*weight + year, data = pr_auto)
> summary(fit7)

Call:
lm(formula = mpg ~ displacement * weight + year, data = pr_auto)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4068  -1.8337   0.0107   1.5543  12.8164

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -6.379e+00  3.616e+00  -1.764   0.0785 .
displacement       -7.682e-02  8.338e-03  -9.214   <2e-16 ***
weight             -1.077e-02  6.327e-04 -17.025   <2e-16 ***
year                8.185e-01  4.523e-02  18.096   <2e-16 ***
displacement:weight 2.213e-05  2.072e-06  10.683   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.019 on 387 degrees of freedom
Multiple R-squared:  0.8519,    Adjusted R-squared:  0.8503
F-statistic: 556.4 on 4 and 387 DF,  p-value: < 2.2e-16
```

This model is not as good as the previous model but still better than the rest.

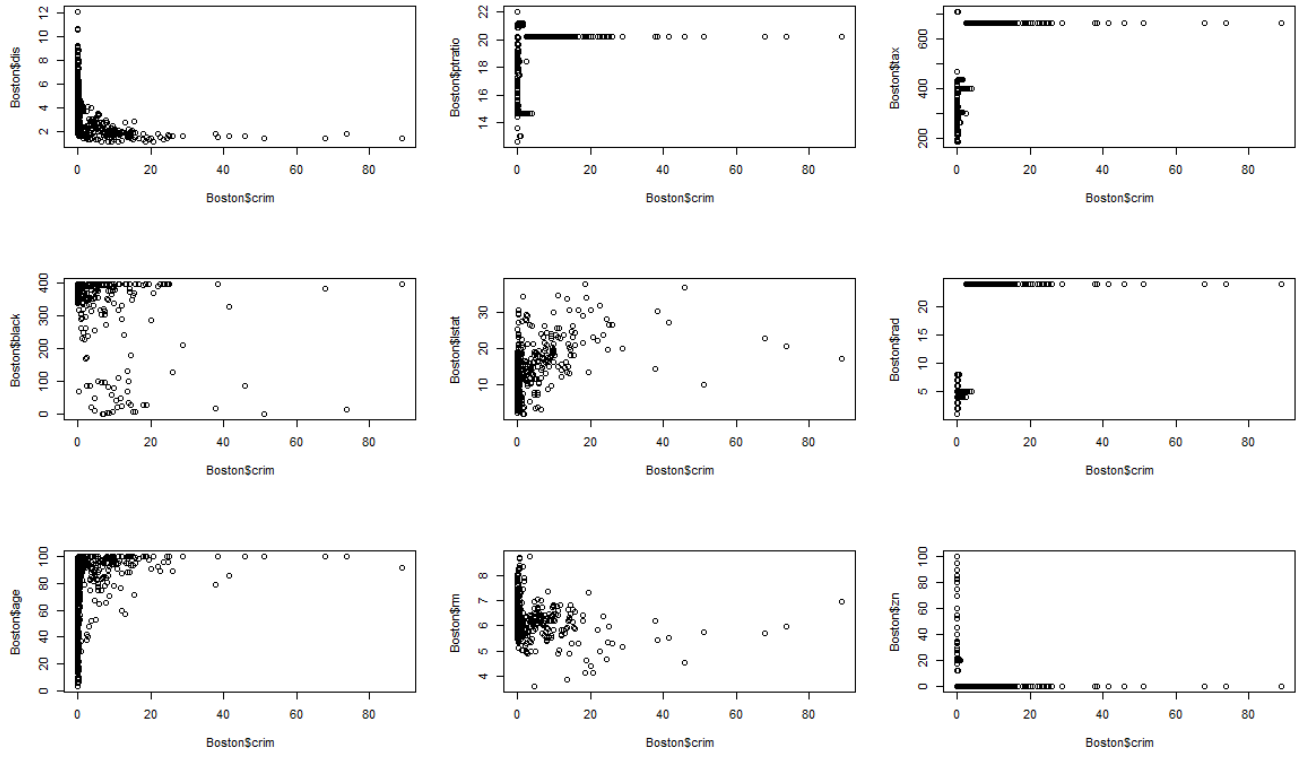
c) It appears that Weight and Horsepower interactions is quite significant along with Weight and Displacement.

3) a)



From the scatterplots above, a few trends are apparent such as the correlation between indus and dis, zn. Furthermore, zn and lstat, age are also clearly correlated. So is lstat and zn, dis.

b)



From the above scatterplots we can see that crime is associated with age, ptratio, zn, tax, dis and rad.

c)
A higher pupil teacher ratio would result in lesser quality education potentially leading to poverty. Increased taxes, distance to employment centers could result in more residents choosing the crime of life. Suburbs in Boston with these factors appear to have more crime. Furthermore, areas with older homes and easier accessibility to highways seem to be having more crime.

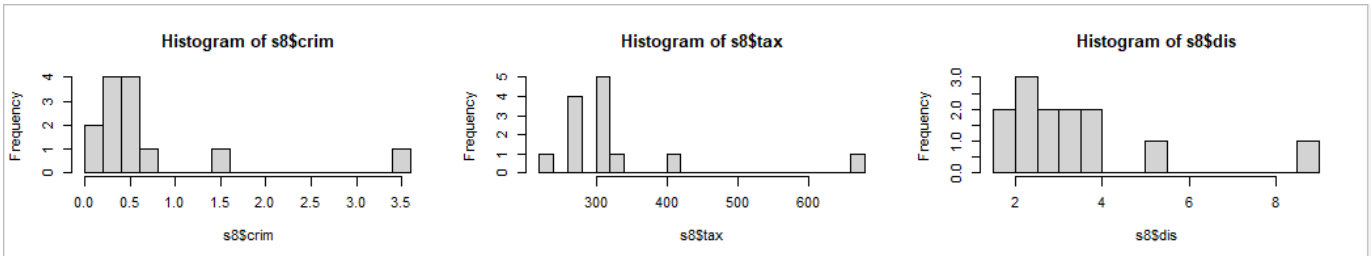
d)

```
> s7 <- Boston[which(Boston$rm > 7),]
> dim(s7)
[1] 64 14
```

64 suburbs seem to average more than 7 rooms per dwelling

```
> s8 <- Boston[which(Boston$rm > 8),]
> dim(s8)
[1] 13 14
```

13 suburbs seem to average more than 8 rooms per dwelling



The suburbs with more an average of more than 8 rooms per dwelling seem to generally have lower crime, lower taxes and easier accessibility to highways.