Motor Trends

Executive Summary

In this report, we look at a data set of a collection of car, and are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). Particularly, we are interested in the following two questions:

```
"Is an automatic or manual transmission better for MPG"
"Quantify the MPG difference between automatic and manual transmissions"
```

In order to answer these two questions, we follow the steps below:

```
Load and process the data such that it makes more sense
Conduct a basic exploratory data analyses to show the relationship between mpg and am
Fit multiple models to the data and select the best model
Diagnose the model and quantify the uncertainty
Using the model we choose, draw conclusion and answer the questions
```

Libraries Required

```r
require(ggplot2)
require(dplyr)
require(reshape2)
```

Dataset

```r
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Deriving the Corelation between the variables

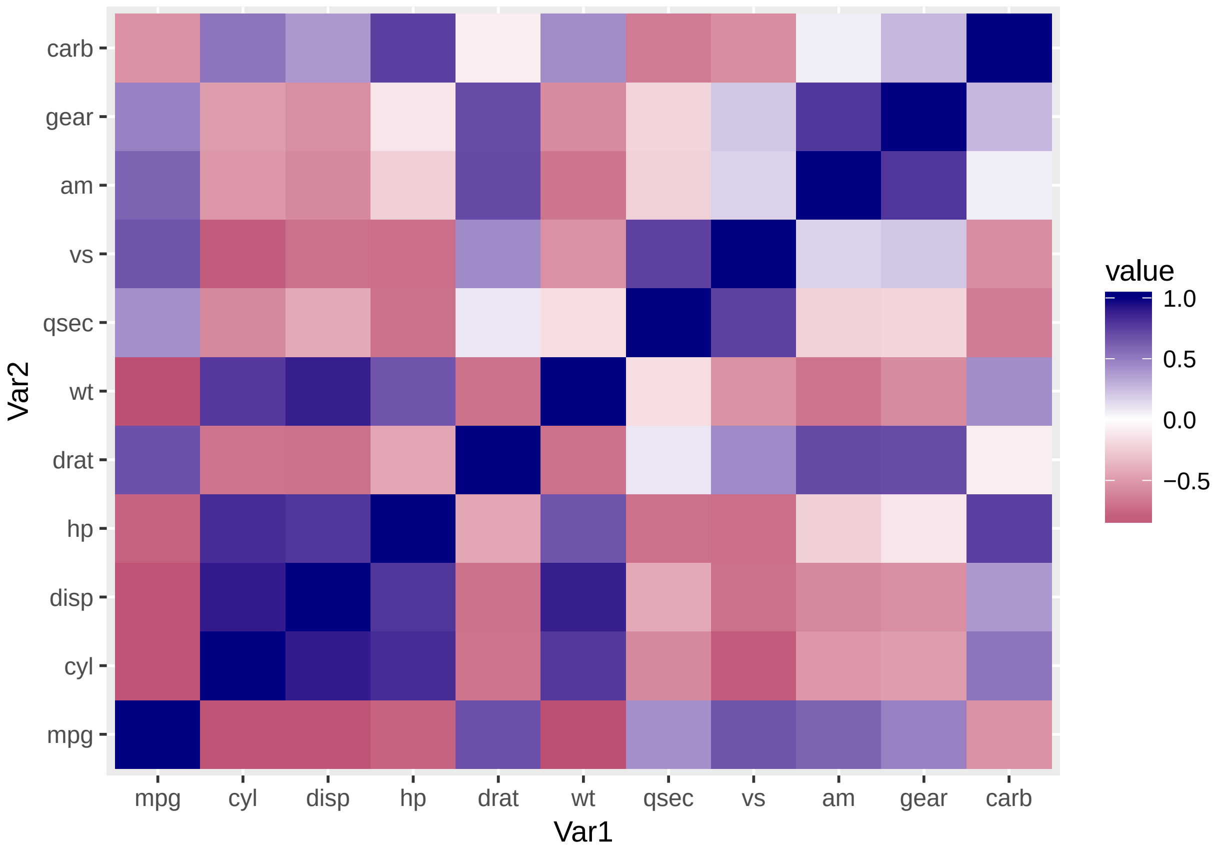
```
corMatrix <- round(cor(mtcars), 2)
corMatrix
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

```
meltedCorMatrix <- melt(corMatrix)
head(meltedCorMatrix)
```

```
##   Var1 Var2 value
## 1  mpg  mpg  1.00
## 2  cyl  mpg -0.85
## 3 disp  mpg -0.85
## 4   hp  mpg -0.78
## 5 drat  mpg  0.68
## 6   wt  mpg -0.87
```

```
G <- ggplot(data = meltedCorMatrix, aes(x = Var1, y =Var2, fill = value)) +
     geom_tile() +
     scale_fill_gradient2(low="Maroon", high="navy Blue", guide="colorbar")
G
```



Changing some variables to factor since they repersent cagtegories not continous values

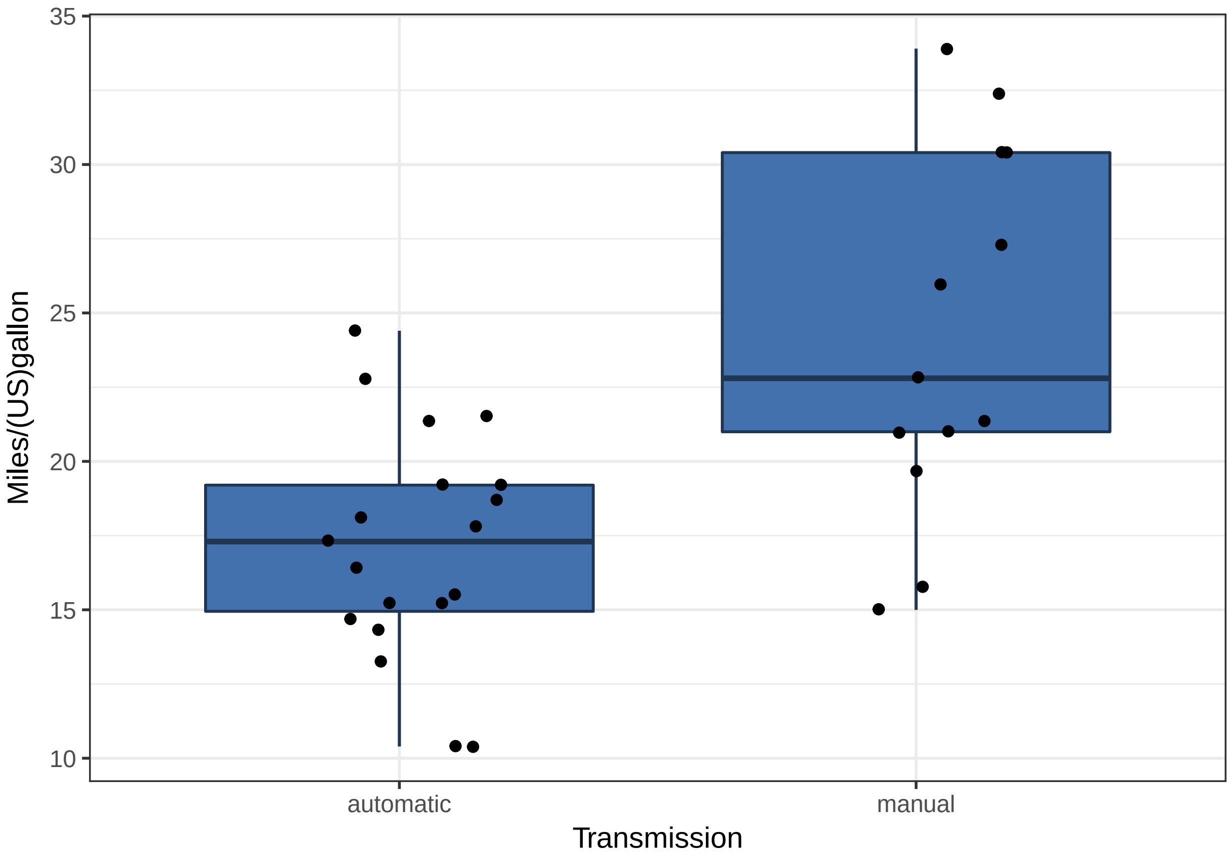
```r
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$carb <- factor(mtcars$carb)
mtcars$gear <- factor(mtcars$gear)
mtcars$cyl <- factor(mtcars$cyl)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

EDA

Relationship between mpg(miles per gallon) and am(transmission)

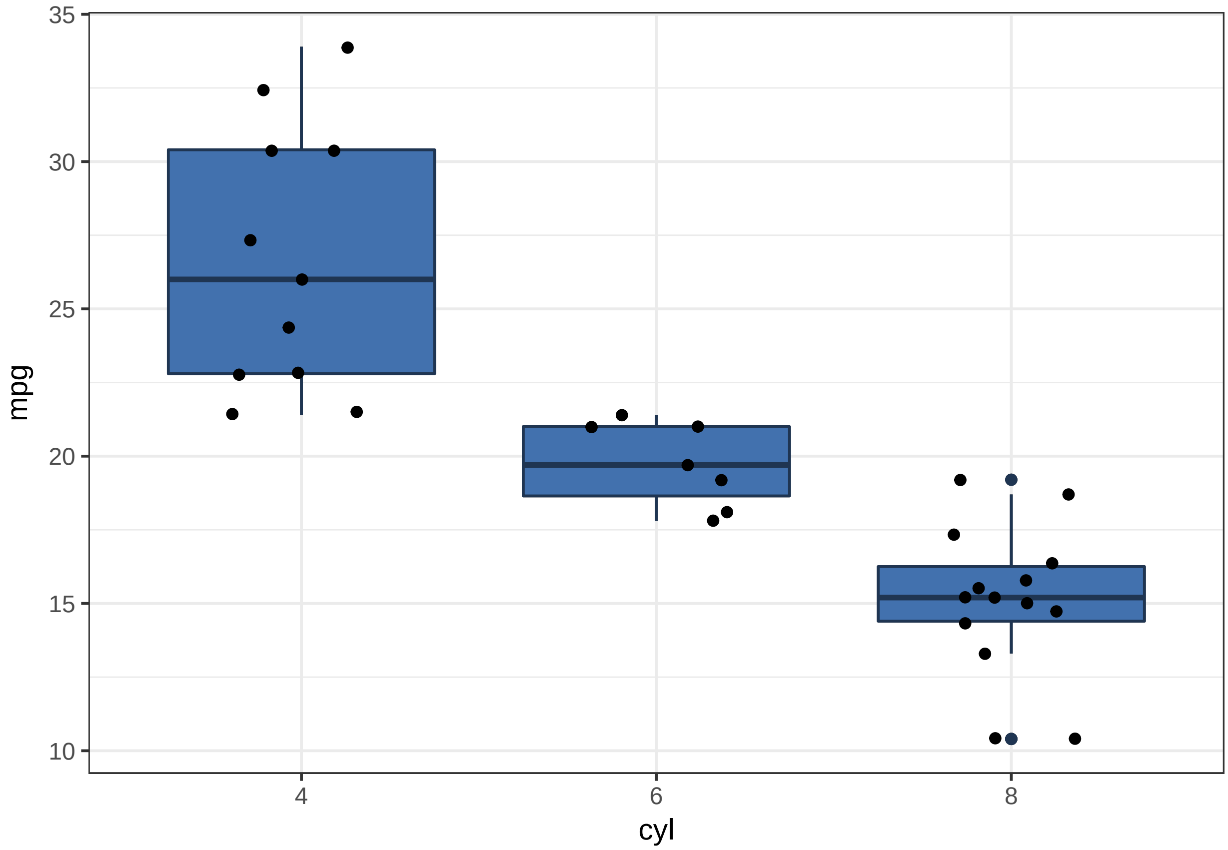
```
levels(mtcars$am) <- c("automatic", "manual")
fill <- "#4271AE"
line <- "#1F3552"
qplot(x= mtcars$am, y= mtcars$mpg, geom = "boxplot") +
    ylab("Miles/(US)gallon") +
    xlab("Transmission") +
    geom_boxplot(fill = fill, colour = line)+
    theme_bw() +

    geom_jitter(width = 0.2)
```



From the above Boxplot we can eaisly understand that, there is a diffrence between two groups, and cars with manual transmission have higher mpg so that of automatic transmission

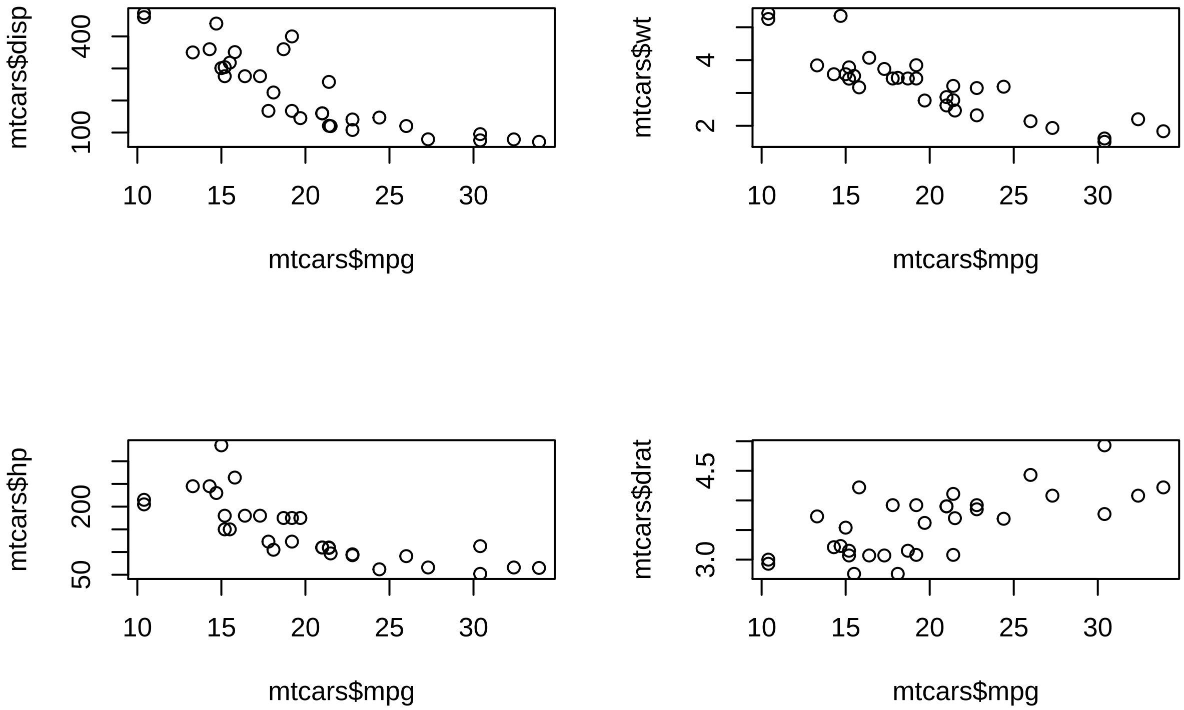
Relationship between mpg(miles per gallon) and cyl(Number of cylinders)

```
ggplot(mtcars, aes(cyl, mpg)) +
    geom_boxplot() +
    geom_boxplot(fill = fill, colour = line)+
    theme_bw() +
    geom_jitter(width = 0.2)
```

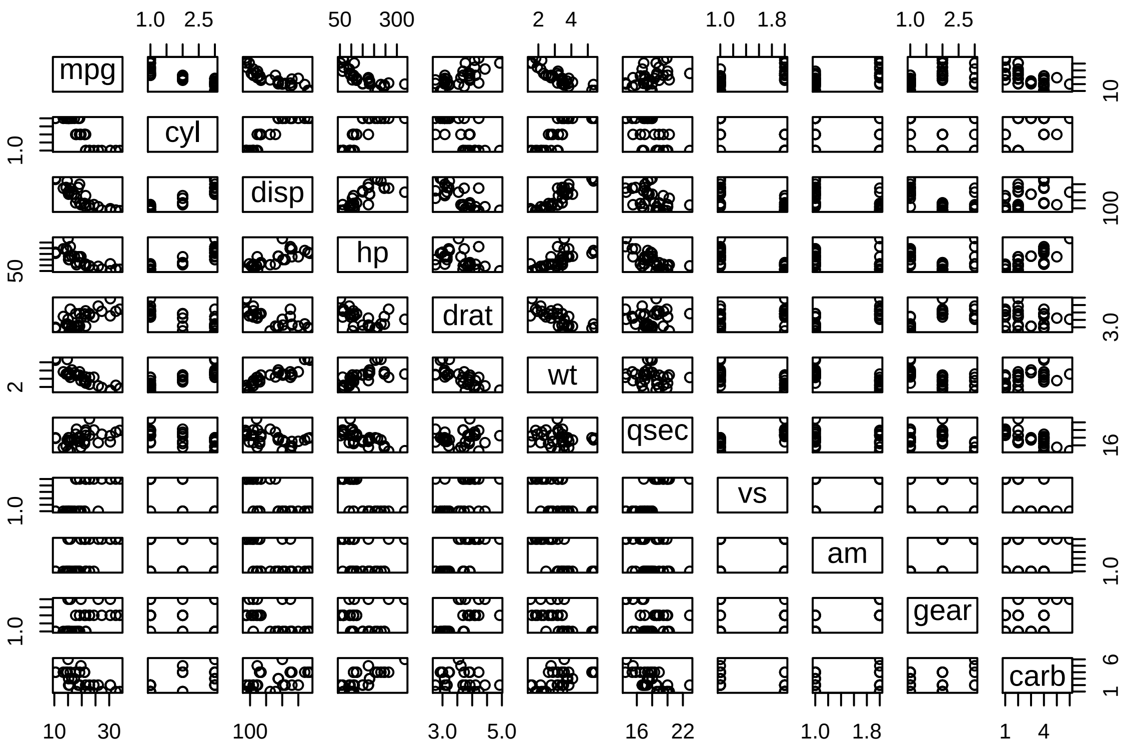


Relationship of mpg with other variable having caorelation value nearby 1 and -1

```
par(mfrow= c(2,2))
plot(mtcars$mpg, mtcars$disp)
plot(mtcars$mpg, mtcars$wt, data= mtcars)
plot(mtcars$mpg, mtcars$hp, data= mtcars)
plot(mtcars$mpg, mtcars$drat, data= mtcars)
```




```
plot(mtcars)
```



Model Buliding and Selection

Model with single variable

starting with basic model in which it depends on variable am(Transmission)

```r
basicModel <- lm(mpg~am, data = mtcars)
summary(basicModel)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## ammanual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

From the summary we can clearly see that Cars with automatic Transmission have more mileage (mpg) having a average of 17.147, wheras in cas of Manual Transmission average is 7.245. The p-value is low (~0.000285), and R-squared value is 0.3385, Which means that model can explain only 33.85% of mpg variability. Hence we need more variable take into account. ###Considring all variable for our model

```r
Full_fledgedModel <- lm(mpg~., data = mtcars)
summary(Full_fledgedModel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## ammanual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
```

```
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Here we have increase in R-squared value whichis now .779, here to improve our model efficency we will remove some insignificant model. We will use variable from our heapmap with correlation valu emore close to -1 and 1

```r
fit1 <- lm(mpg~wt+ am + cyl + disp  + hp+ drat, data =mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + cyl + disp + hp + drat, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8267 -1.4366 -0.4153  1.1649  5.0671
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.611986   6.274227   5.198 2.52e-05 ***
## wt          -2.726729   1.200207  -2.272   0.0323 *
## ammanual     1.681130   1.554386   1.082   0.2902
## cyl6        -3.026760   1.576680  -1.920   0.0669 .
## cyl8        -2.541967   3.059145  -0.831   0.4142
## disp         0.004395   0.013090   0.336   0.7400
## hp          -0.033038   0.014476  -2.282   0.0316 *
## drat         0.326616   1.471086   0.222   0.8262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.501 on 24 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8278
## F-statistic: 22.29 on 7 and 24 DF,  p-value: 4.768e-09
```

R-squared value(~0.8278) increased, means our model is now improves version of previous one Now trying to make this model more efficent by removing or adding some variable

```r
fit2 <- lm(mpg~wt+ am + cyl  + hp  , data =mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + cyl + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ammanual     1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

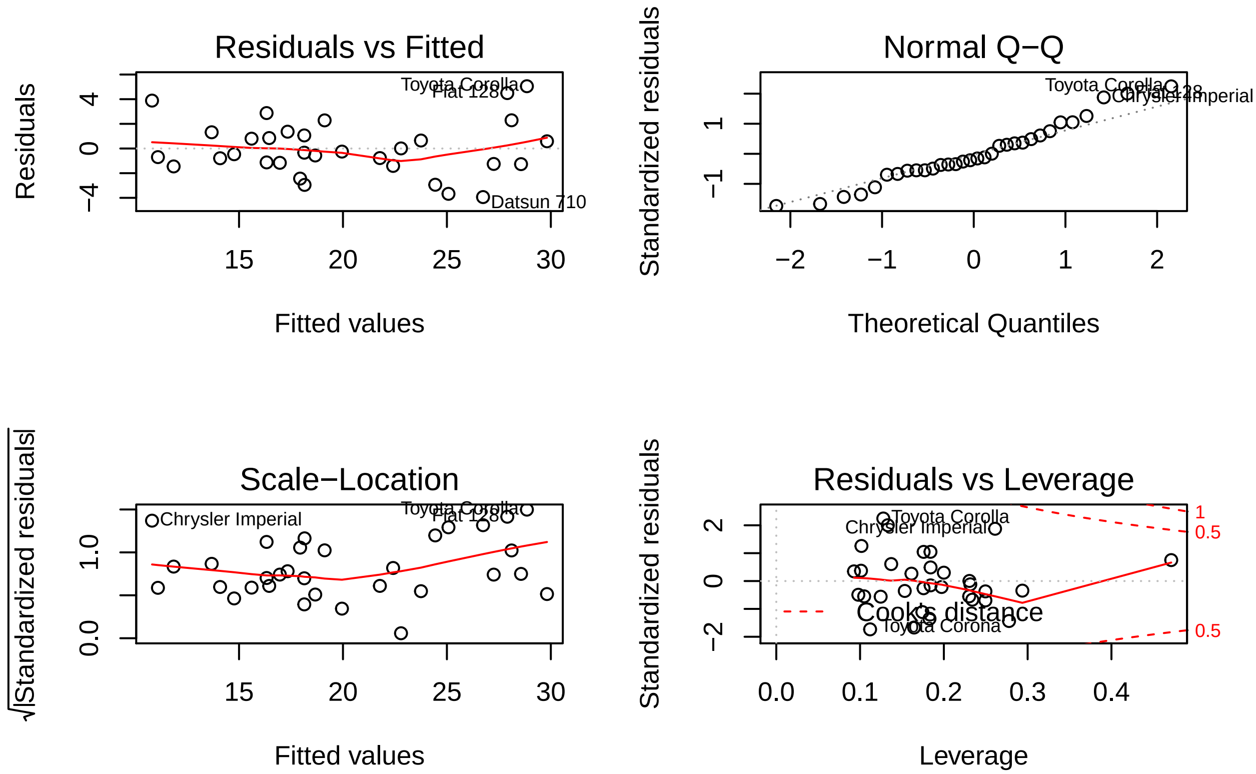
```r
fit3 <- lm(mpg~wt+ am + cyl + disp  + hp , data =mtcars)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + cyl + disp + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ammanual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

From above three model we have R-squared value as following fit1 :- 0.8278 fit2 :- 0.8401 fit3 :- 0.8344 So our best fit model is fit2 with p-value: 1.506e-10 le ss than 5% and with least Residual standard error 2.41 on 26 degrees of freedom

Let's plot the diagnosis of the model.

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



From the above plots, we can make the following observations,

The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot and verify the independence condition. The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed. The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance. There are some distinct points of interest (outliers or leverage points) in the top right of the plots. We now compute some regression diagnostics of our model to find out these interesting leverage points as shown in the following section. We compute top three points in each case of influence measures.

Infrence

We can also conduct a T-test to confirm our observation. Define the null hypothesis as manual and automatic transmissions result in the same mpg.

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group automatic    mean in group manual
##                17.14737                24.39231
```

P-value is 0.00137, and confidence interval does not include zero, so we reject the null hypothesis and accept the difference in mpg between manual and automatic transmission, which we observed earlier.

Conclusion

Based on the observations from our best fit model, we can conclude the following,

1. Cars with Manual transmission get more miles per gallon compared aganist cars with Automatic transmission. (1.8 adjusted by hp, cyl, and wt). mpg will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.
2. mpg decreases negligibly with increase of hp.
3. If number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).