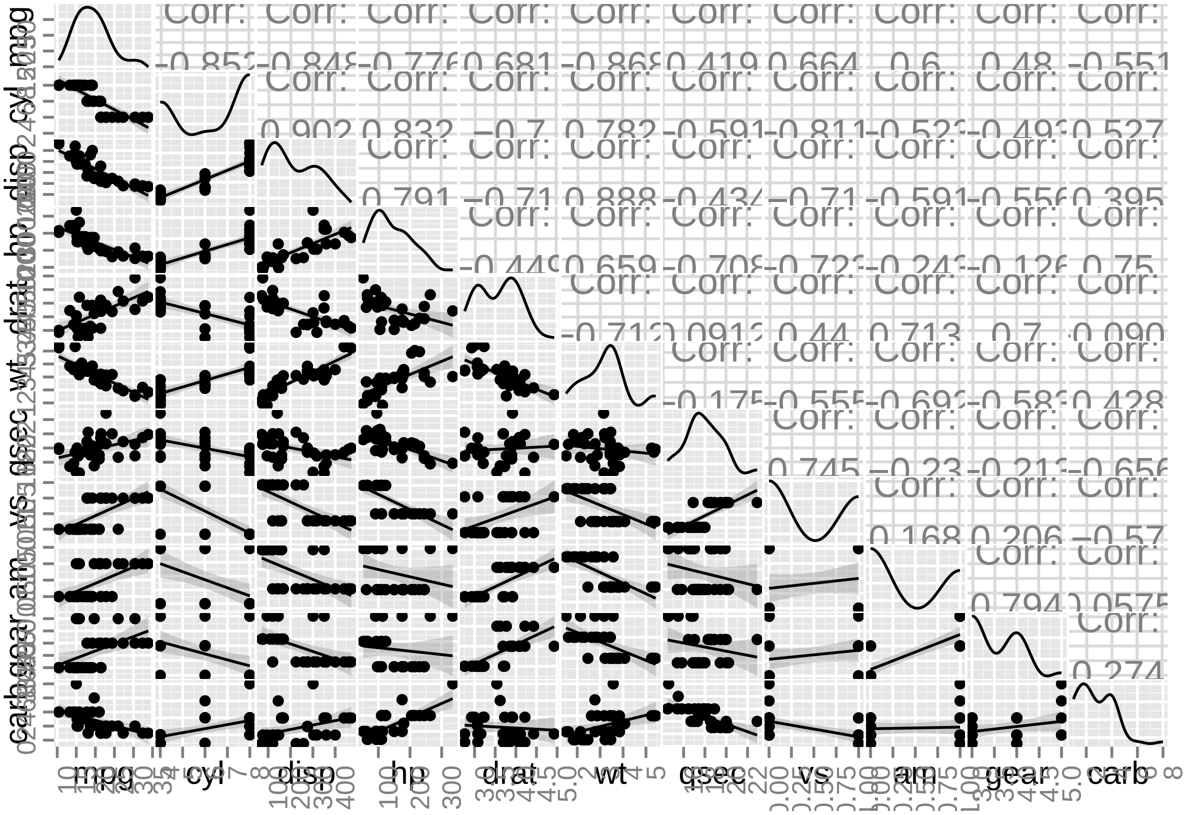
Best Fit Model Selection

Initial Preprocessing and Investigation

Loading the data and seeing how the data is related to each other by using GGally package in R. This plot shows the correlation between various variables.

```
library(datasets)
data("mtcars")

library(GGally)
library(ggplot2)
g <- ggpairs(mtcars, lower = list(continuous = "smooth"), params = c(binwidth = 1))
g <- g + theme(axis.text.x = element_text(angle = 90, hjust = 1), axis.text.y = element_text(angle = 90
g
```

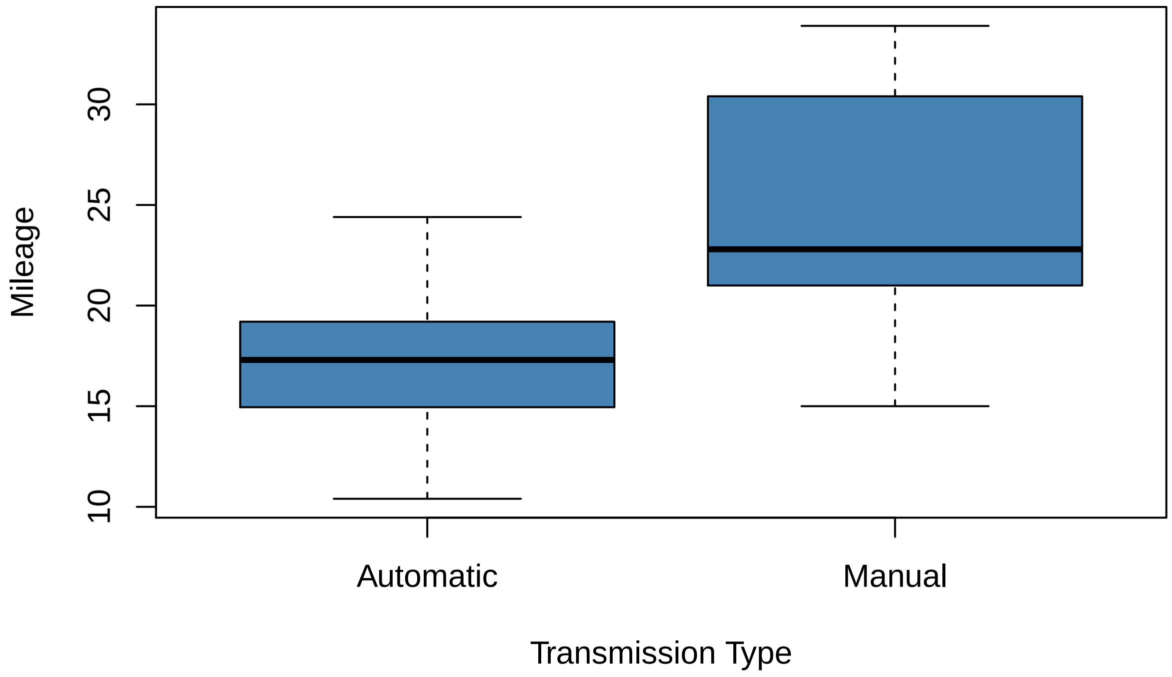


```
#Changing data to include "Automatic" and "Manual" instead of 0 and 1
mtcars$am <- ifelse(mtcars$am == 0, "Automatic", "Manual")
```

Modeling

Now ploting the data to see which one has better effect on Mileage, automatic or manual transmission

```
boxplot(mtcars$mpg ~ mtcars$am, col = "steelblue", ylab = "Mileage", xlab = "Transmission Type")
```



As it can be clearly seen from the plot that Manual transmission cars have better performance than Automatic transmission cars. But this is only when taking Transmission as the criteria and keeping other factors constant. We need to include other factors also to see wehther transmission is only the major factor or not.

First we will consider how MPG varies when all the variables are included into the model.

```
allDataFit <- lm(data = mtcars, mpg ~ .)
summary(allDataFit)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## amManual     2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

From the above summary of coefficients, it can be easily said that mileage depends on wt and transmission, majorly, but this cant be said with 100% surety as this relation is computed with taking all the coefficients into consideration which can lead to overfitting of the data.

Further if we try to fit linear model between mileage and transmission and see what it shows:-

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

```
##
## 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 Manual transmission cars give more mileage than Automatic tranmission cars as the μ value for Manual cars is 7.245 more than Automatic. But as we can see from summary that R^2 value is approximately 35%, which implies that this model only captures 35% of the total variance, so we need to further investigate with other variables as well.

So now we fit various models with the step function to find out the best modeling variables. As this function tries to find out the best variables that define this model using steps in decreasing order of the number of variables present in the data.

```
bestFit <- step(lm(data = mtcars, mpg ~.), trace = 0)
summary(bestFit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

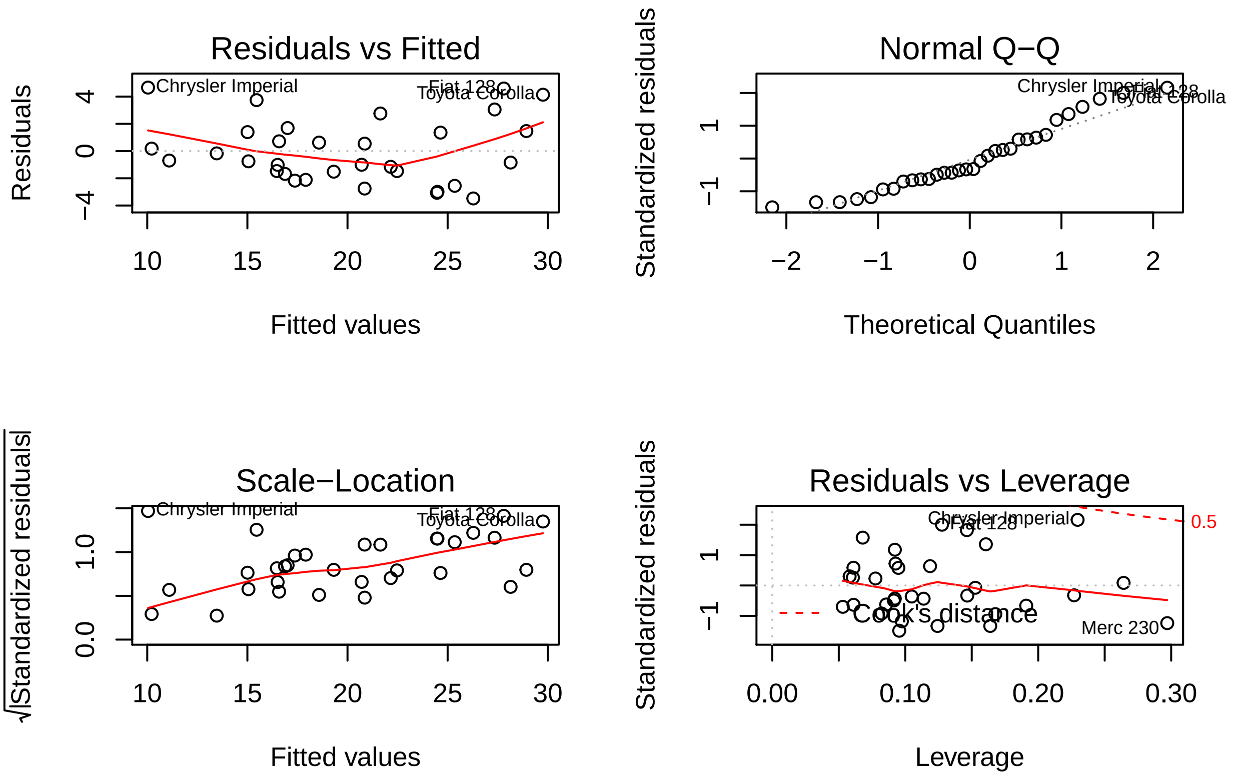
As we can see from the summary that the variables which can be used to fit the best model are wt, qsec and am. So not only weight and transmission are important factors which was seen from the above mulitvariable model fit but qsec also plays a major role. R^2 value is also approximately 85% percent which means that these three variables define the maximum variance.

Appendix

Conclusion

As it can be seen that mileage not only depends on the type of transmission of the car but on other factors as well. Further we plot the residual graphs below:-

```
#Plotting residual graphs
par(mfrow = c(2,2))
plot(bestFit)
```



From the first plot between Residuals and Fitted values, we can see that there isn't any abnormal variance between the fitted values and residuals to show that the model doesn't fit. So this plot shows that this model is a good fit. Further more, from the last plot we can find out whether there are any residuals which has high levarage on the regression line to impact the analysis and we can see that there is no such point in the data which suggests this kind of behaviour.