Introduction to R: mtcars

Sean Cho

Load dataset

We will be working with the `mtcars` dataset. We will first load the dataset and then find out more about the dataset.

We can use `?` or `help` to get information about `functions`. We can also use them to find out more about datasets, including `mtcars`.

```
data('mtcars')
?mtcars

## A data frame with 32 observations on 11 (numeric) variables.
##
## [, 1]    mpg Miles/(US) gallon
## [, 2]    cyl Number of cylinders
## [, 3]    disp    Displacement (cu.in.)
## [, 4]    hp  Gross horsepower
## [, 5]    drat    Rear axle ratio
## [, 6]    wt  Weight (1000 lbs)
## [, 7]    qsec    1/4 mile time
## [, 8]    vs  Engine (0 = V-shaped, 1 = straight)
## [, 9]    am  Transmission (0 = automatic, 1 = manual)
## [,10]    gear    Number of forward gears
## [,11]    carb    Number of carburetors
```

Examine dataset

We can look at the structure of `mtcars` using the `str()`.

```
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 ...
```

We can see that the `data.frame` has 11 variables, or columns, and 32 observations, or rows. Although all of these are numeric, we know that `vs` and `am` are binary columns of whether the car has a V-shaped engine and whether the car has an automatic or manual transmission.

Next, we will use `summary` to summarise the `mtcars` data frame.

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

We see that `mtcars$mpg` ranges from 10.40 to 33.90 with a mean of 20.09 and a median of 19.20.

Now, we'll take a look at the first few rows of mtcars.

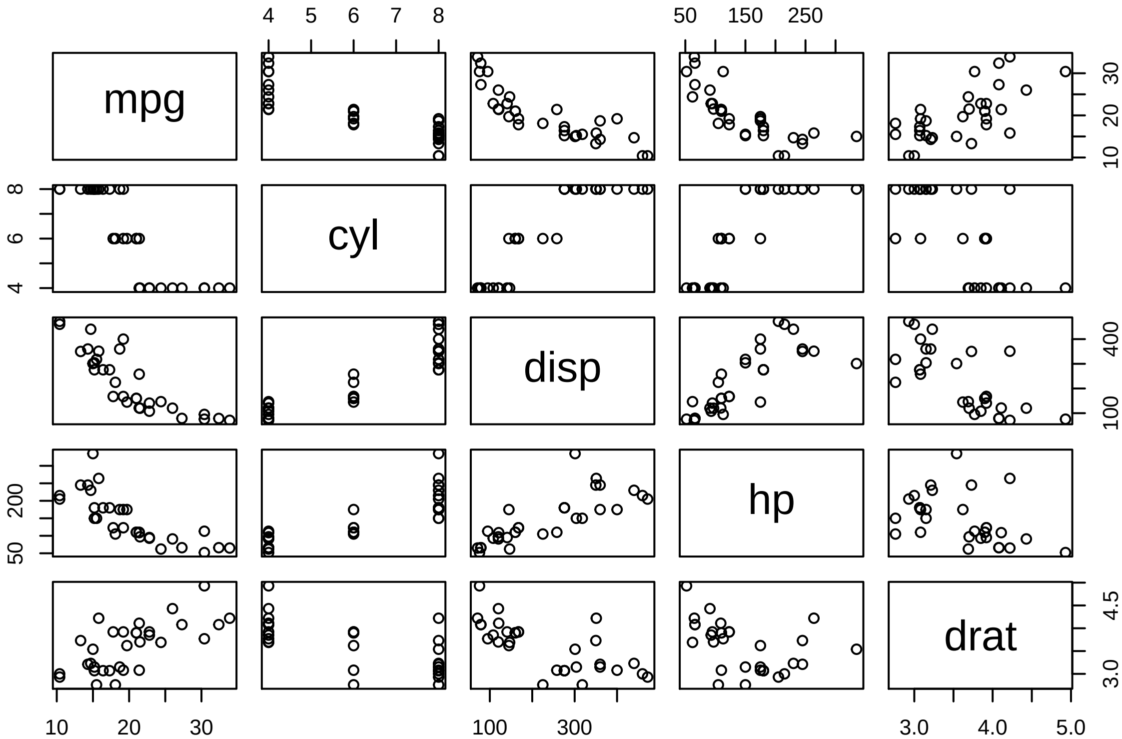
```r
head(mtcars)
```

```
##                    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
```

Exploratory analysis

By using the `plot` function on a `data.frame`, we can make pair-wise scatterplots for the columns in the `data.frame`. Here, we will `plot` the first five columns of `mtcars`.

```r
plot(mtcars[,1:5])
```



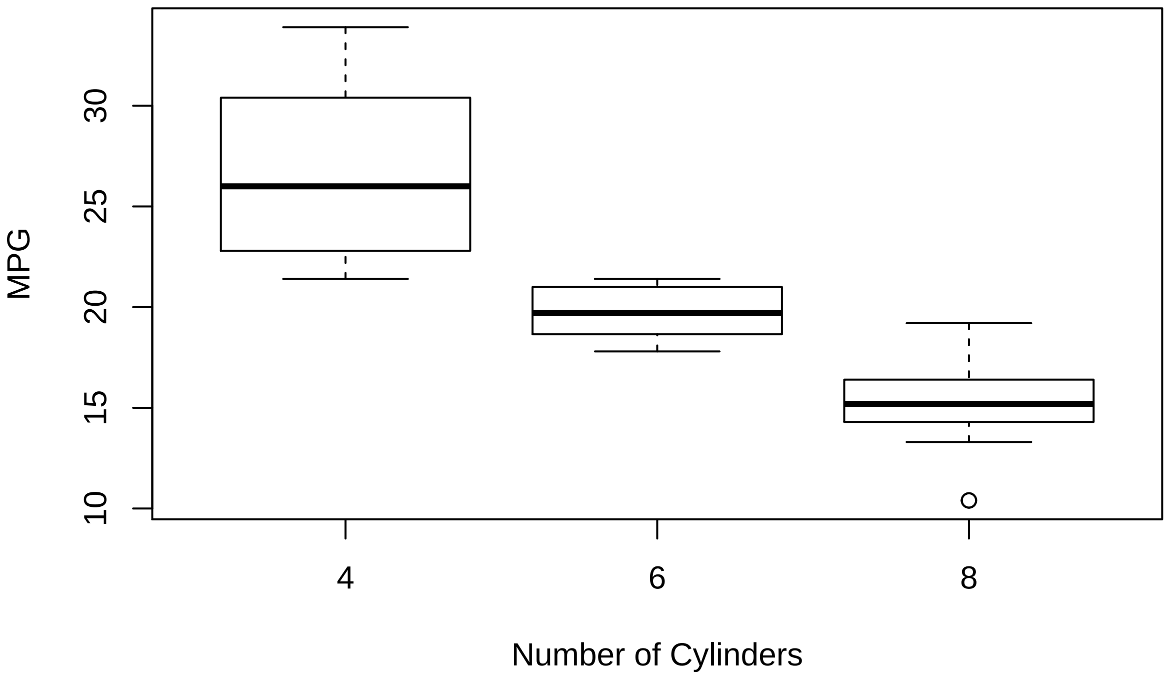
We can see a few relationships between `mpg` and the other variables. One of them is `hp`, or horsepower.

Does this make sense? Do we expect that miles/gallon is inversely related to horsepower? Yes. We would expect that a car with greater horsepower would tend to be less efficient.

Evaluating categorical variables

There are several categorical variables in the `mtcars` dataset, including `cyl`, `gear`, and `carb`. We can examine the relationship between `mpg` and `cyl` and visualise that using a boxplot.

```r
boxplot(mtcars$mpg ~ mtcars$cyl, xlab = 'Number of Cylinders', ylab = 'MPG')
```



From the boxplot, we can observe that there is a stepwise decrease in mpg with increasing cylinders. There is very little overlap between the boxplots across cylinders and we can test that if there are statistically significant differences.

We will run two tests. (1) an ANOVA to identify if any of the `cyl` groups have different mpg values and (2) a pairwise t-test to identify differences across groups.

```
## ANOVA analysis
anova(aov(mpg ~ cyl, data = mtcars))
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        1 817.71  817.71  79.561 6.113e-10 ***
## Residuals 30 308.33   10.28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## pairwise t-test
pairwise.t.test(mtcars$mpg, g = mtcars$cyl)
```

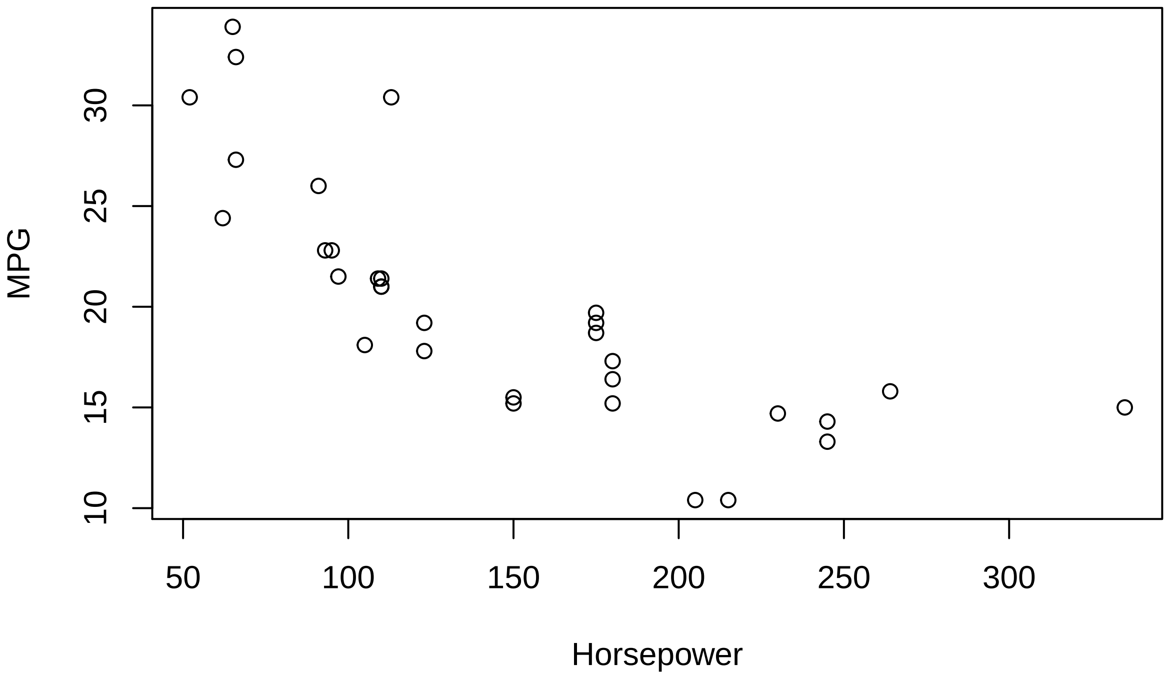
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  mtcars$mpg and mtcars$cyl
##
##   4       6
## 6 0.00024 -
## 8 2.6e-09 0.00415
##
## P value adjustment method: holm
```

In this sample, we can conclude that there are pairwise differences of mpg across number of cylinders of the car's engine.

Comparing numerical variables

We will continue to explore the relationship between mpg and horsepower that we observed earlier by making a scatterplot of the two variables.

```
## plot( y ~ x , data = dataset)
plot(mpg ~ hp, data = mtcars, ylab = 'MPG', xlab = 'Horsepower')
```




```
## plot(x = mtcars$mpg, y = mtcars$mpg) will work as well
```

We can observe what appears to be an inverse relationship between mpg and horsepower. We can fit a linear model that

```
## lm( y ~ x1 + x2 + ... + xn , data = dataset)
summary(lm(mpg ~ hp, data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

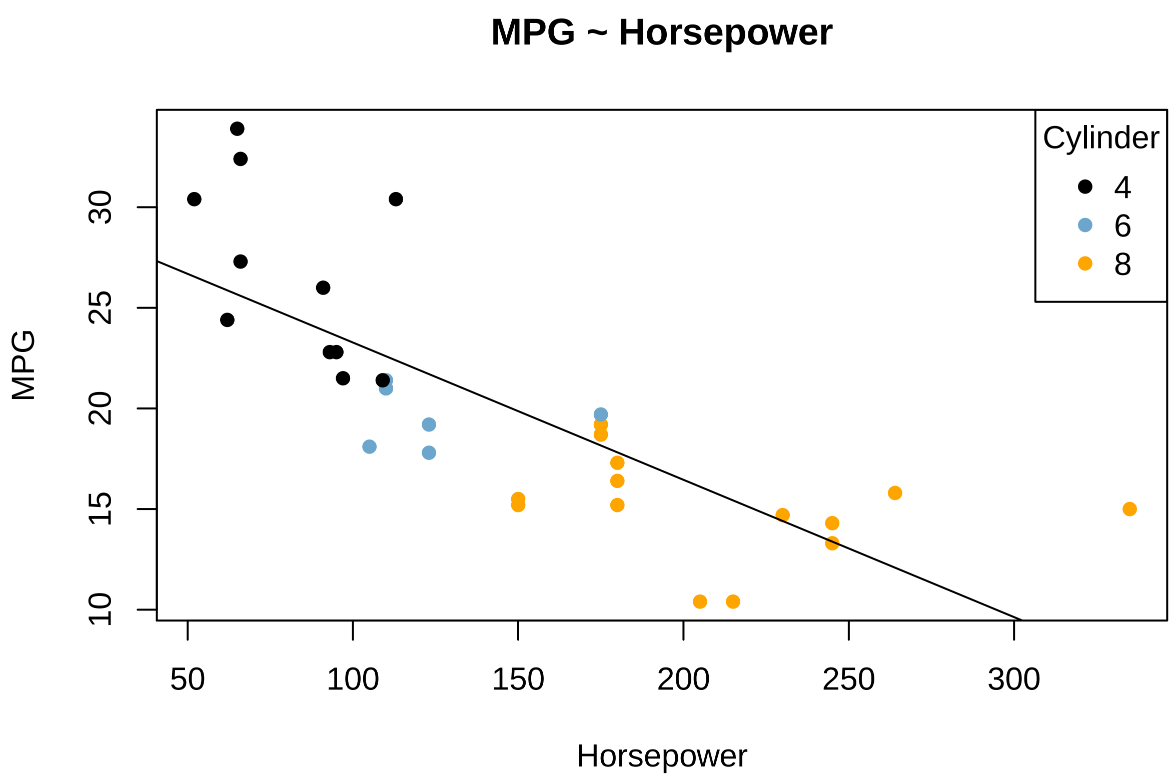
From this model, we observe that the estimate is that for every 1 point increase of horsepower, there is a decrease of -0.06823 mpg.

So now we know that there is an inverse relationship between mpg and horsepower on top of the relationship between mpg and the number of cylinders. Let's include all that information in our plot.

```
## create colors
mapcolors <- c('4'='black','6'='skyblue3','8'='orange')
mtcolors <- mapcolors[as.character(mtcars$cyl)]

## model
mtmodel <- lm(mpg ~ hp, data = mtcars)

## make basic plot
plot(mpg ~ hp, data=mtcars, col = mtcolors, pch = 16, xlab = 'Horsepower', ylab = 'MPG',
     main = 'MPG ~ Horsepower')
## add fitted line
abline(mtmodel)
## add legend
legend('topright', legend = c(4,6,8), col = c('black','skyblue3','orange'), pch = 16, title = 'Cylinder
```



Multiple regression

We will explore the relationship between `vs`, V-shaped engine, and mpg. Let's first fit a model to assess that.

```
unique(mtcars$vs)
```

```
## [1] 0 1
```

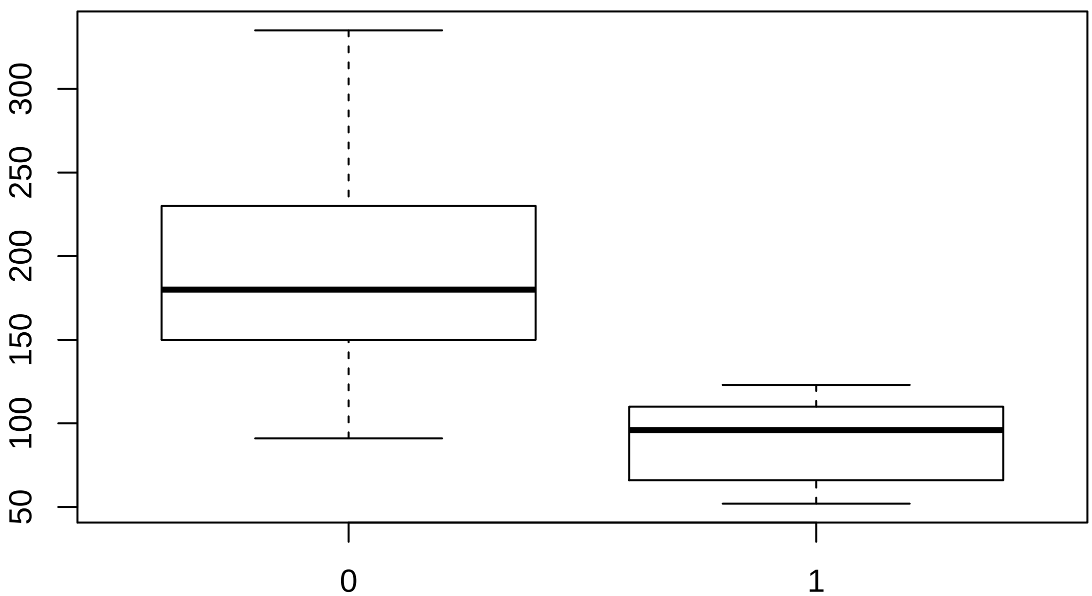
```
## V-shaped engine
summary(lm(mpg ~ vs ,data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ vs, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.757 -3.082 -1.267  2.828  9.383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.617      1.080  15.390 8.85e-16 ***
## vs             7.940      1.632   4.864 3.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.581 on 30 degrees of freedom
## Multiple R-squared:  0.4409, Adjusted R-squared:  0.4223
## F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05
```

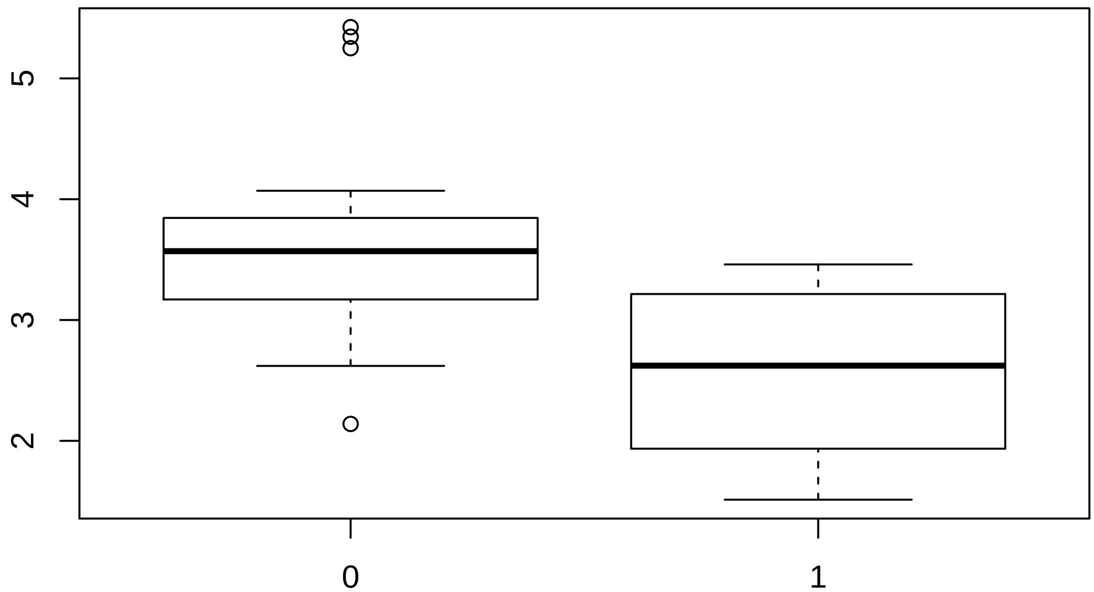
From this analysis, we can see that there is a significant difference between the mpg of cars with a V-shaped versus an inline engine. Do we believe that this relationship is real? Maybe. Or maybe there are other covariates that we have not considered.

In this dataset, we will show here that cars with V-shaped engines differ from those with inline engines in horsepower and weight, and that difference captures more mpg variability than engine shape.

```
boxplot(mtcars$hp ~ mtcars$vs)
```




```
boxplot(mtcars$wt ~ mtcars$vs)
```




```r
## V-shaped engine and hp and weight
summary(lm(mpg ~ vs + hp + wt ,data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ vs + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4667 -1.4857 -0.4296  1.0341  5.7384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.38267    2.42564  14.587 1.31e-14 ***
## vs           1.36771    1.35296   1.011   0.3207
## hp          -0.02542    0.01100  -2.312   0.0284 *
## wt          -3.78003    0.63985  -5.908 2.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.592 on 28 degrees of freedom
## Multiple R-squared:  0.8329, Adjusted R-squared:  0.815
## F-statistic: 46.52 on 3 and 28 DF,  p-value: 5.276e-11
```