# Math 189: Midterm Project 1

## Motor Trend Car Road Tests

Working alone, prepare a R Markdown Notebook report based on examining the Motor Trend Car Road Tests dataset (available from GitHub). The file contains data extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). The dataset contains 32 observations on 11 variables.

### Introduction

In this project, we were asked the same question as in Homework 2: to find the relationship between the weight of a car and miles per gallon, and whether or not the amount of cylinders in the car affected this correlation. However, in this project, I used more concrete methods to fully determine the effect such as regression lines, sample mean, sample variance and co-variance, and sample correlation while also maintaining the more basic methods from the previous homework.

### Data

This data was pulled from a 1974 Motor Trend US magazine. (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html) "Motor Trend Car Road Tests". It contains data about the model and different aspects such as weight, miles per gallon, number of cylinders, and 8 other variables. The three listed were the ones mainly used in this task.

### Methods

Tables were used to show the relevant data, and the models used were 2D and 3D scatterplots containing regression lines to show the information. The rest of the data was outputted in matrix form.

Analysis and Results will be discussed above each block of R code to more accurately demonstrate what was being done.

### Task

What is the relationship between weight (wt) and miles per gallon (mpg)? Does this relationship depend on the number of cylinders (cyl)?

Read the full data from the source and give the head of that data.

```
"C:\\Users\\Mark's PC\\Desktop\\Math 189\\ma189-main\\Data\\mtcars.csv" = paste0(getwd(), "/mtcars.csv"
carsfull = read.csv("C:\\Users\\Mark's PC\\Desktop\\Math 189\\ma189-main\\Data\\mtcars.csv", header = T
head(carsfull)
```

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

Narrow down the data to what will be used (ie. mpg, cyl, wt) for this project.

```
cars = carsfull[,c(1,2,3,7)]
cars
```

```
##                 model  mpg cyl    wt
## 1           Mazda RX4 21.0   6 2.620
## 2       Mazda RX4 Wag 21.0   6 2.875
## 3          Datsun 710 22.8   4 2.320
## 4      Hornet 4 Drive 21.4   6 3.215
## 5   Hornet Sportabout 18.7   8 3.440
## 6             Valiant 18.1   6 3.460
## 7          Duster 360 14.3   8 3.570
## 8           Merc 240D 24.4   4 3.190
## 9            Merc 230 22.8   4 3.150
## 10           Merc 280 19.2   6 3.440
## 11           Merc 280C 17.8   6 3.440
## 12          Merc 450SE 16.4   8 4.070
## 13          Merc 450SL 17.3   8 3.730
## 14         Merc 450SLC 15.2   8 3.780
## 15  Cadillac Fleetwood 10.4   8 5.250
## 16 Lincoln Continental 10.4   8 5.424
## 17   Chrysler Imperial 14.7   8 5.345
## 18            Fiat 128 32.4   4 2.200
## 19         Honda Civic 30.4   4 1.615
## 20      Toyota Corolla 33.9   4 1.835
## 21       Toyota Corona 21.5   4 2.465
## 22    Dodge Challenger 15.5   8 3.520
## 23         AMC Javelin 15.2   8 3.435
## 24          Camaro Z28 13.3   8 3.840
## 25    Pontiac Firebird 19.2   8 3.845
## 26           Fiat X1-9 27.3   4 1.935
## 27       Porsche 914-2 26.0   4 2.140
## 28        Lotus Europa 30.4   4 1.513
## 29      Ford Pantera L 15.8   8 3.170
## 30        Ferrari Dino 19.7   6 2.770
## 31       Maserati Bora 15.0   8 3.570
## 32          Volvo 142E 21.4   4 2.780
```

Pull subsets from the narrowed down list to create data sets for each of (4 cylinder, 6 cylinder, and 8 cylinder) cars to be used later on.

```
cars4 = subset(cars,cyl == 4)
cars4
```

```
##             model  mpg cyl    wt
## 3      Datsun 710 22.8   4 2.320
## 8       Merc 240D 24.4   4 3.190
## 9        Merc 230 22.8   4 3.150
## 18       Fiat 128 32.4   4 2.200
## 19    Honda Civic 30.4   4 1.615
## 20 Toyota Corolla 33.9   4 1.835
## 21  Toyota Corona 21.5   4 2.465
## 26      Fiat X1-9 27.3   4 1.935
## 27  Porsche 914-2 26.0   4 2.140
## 28   Lotus Europa 30.4   4 1.513
## 32     Volvo 142E 21.4   4 2.780
```

```
cars6 = subset(cars,cyl == 6)
cars6
```

```
##             model  mpg cyl    wt
## 1       Mazda RX4 21.0   6 2.620
## 2   Mazda RX4 Wag 21.0   6 2.875
## 4  Hornet 4 Drive 21.4   6 3.215
## 6         Valiant 18.1   6 3.460
## 10       Merc 280 19.2   6 3.440
## 11      Merc 280C 17.8   6 3.440
## 30   Ferrari Dino 19.7   6 2.770
```

```
cars8 = subset(cars,cyl == 8)
cars8
```

```
##                  model  mpg cyl    wt
## 5    Hornet Sportabout 18.7   8 3.440
## 7            Duster 360 14.3   8 3.570
## 12           Merc 450SE 16.4   8 4.070
## 13           Merc 450SL 17.3   8 3.730
## 14          Merc 450SLC 15.2   8 3.780
## 15  Cadillac Fleetwood 10.4   8 5.250
## 16 Lincoln Continental 10.4   8 5.424
## 17   Chrysler Imperial 14.7   8 5.345
## 22    Dodge Challenger 15.5   8 3.520
## 23         AMC Javelin 15.2   8 3.435
## 24          Camaro Z28 13.3   8 3.840
## 25    Pontiac Firebird 19.2   8 3.845
## 29      Ford Pantera L 15.8   8 3.170
## 31       Maserati Bora 15.0   8 3.570
```

Plot a single scatter-plot with points and regression lines for each subgroup of cars (red = 4cylinders, blue = 6cylinders, green = 8cylinders). Already one notices that the regression lines are not equal.

```
plot(x = cars$wt, y = cars$mpg, xlab = "Weight", ylab = "Miles per gallon",
     main = "Weight vs Miles per Gallon")
points(cars4$wt, cars4$mpg, col = "red", pch = 19)
points(cars6$wt, cars6$mpg, col = "blue", pch = 19)
points(cars8$wt, cars8$mpg, col = "green", pch = 19)

lm(formula =cars4$mpg ~ cars4$wt )
```

```
##
## Call:
## lm(formula = cars4$mpg ~ cars4$wt)
##
## Coefficients:
## (Intercept)     cars4$wt
##      39.571       -5.647
```

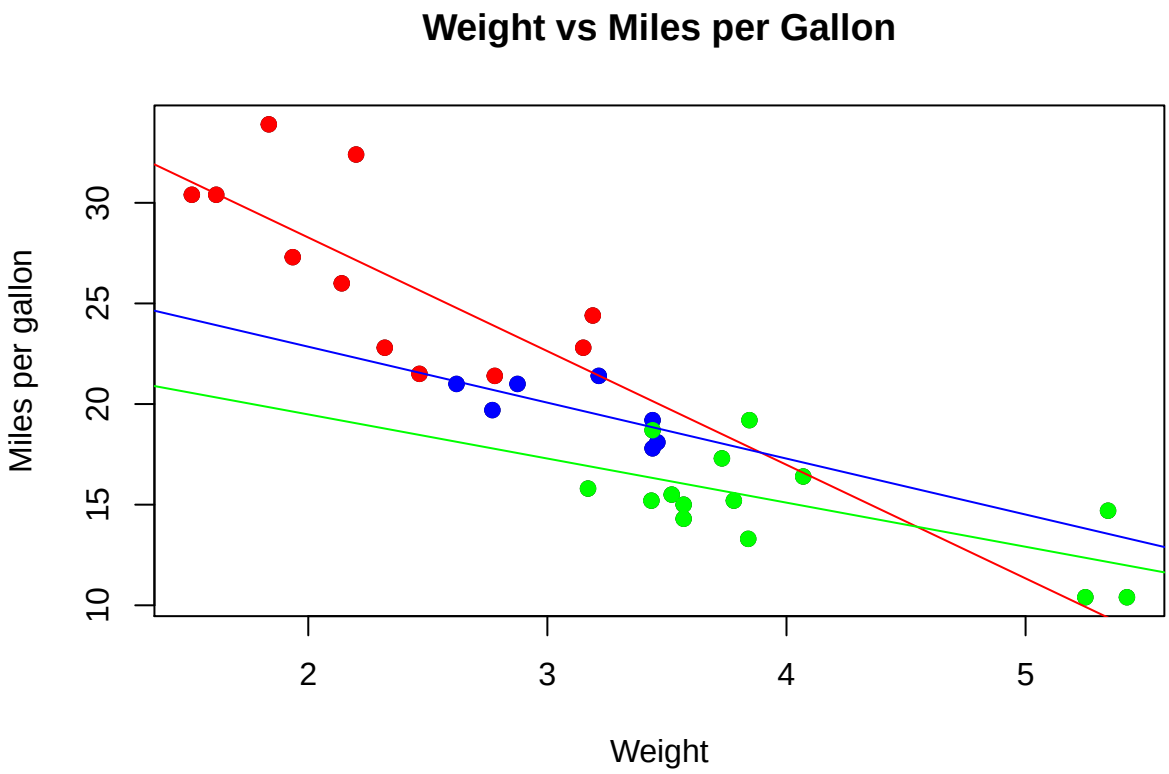
```
abline(39.571,-5.647,col = "red")
lm(formula =cars6$mpg ~ cars6$wt )
```

```
##
## Call:
## lm(formula = cars6$mpg ~ cars6$wt)
##
## Coefficients:
## (Intercept)     cars6$wt
##       28.41        -2.78
```

```
abline(28.41,-2.78,col = "blue")
lm(formula =cars8$mpg ~ cars8$wt )
```

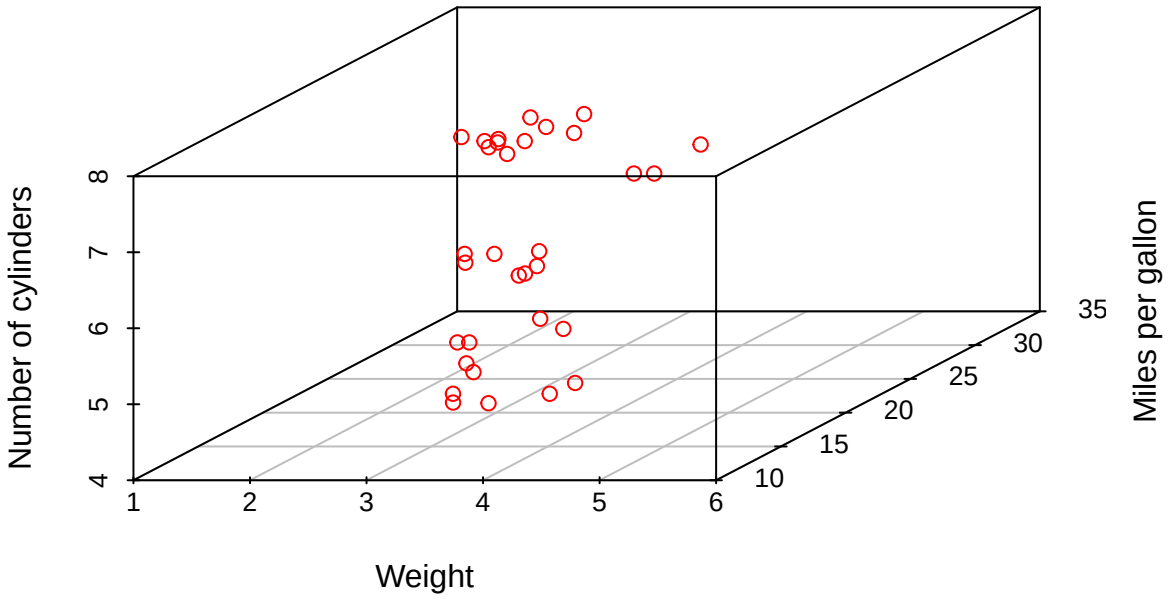
```
##
## Call:
## lm(formula = cars8$mpg ~ cars8$wt)
##
## Coefficients:
## (Intercept)     cars8$wt
##      23.868       -2.192
```

```
abline(23.868,-2.192,col = "green")
```

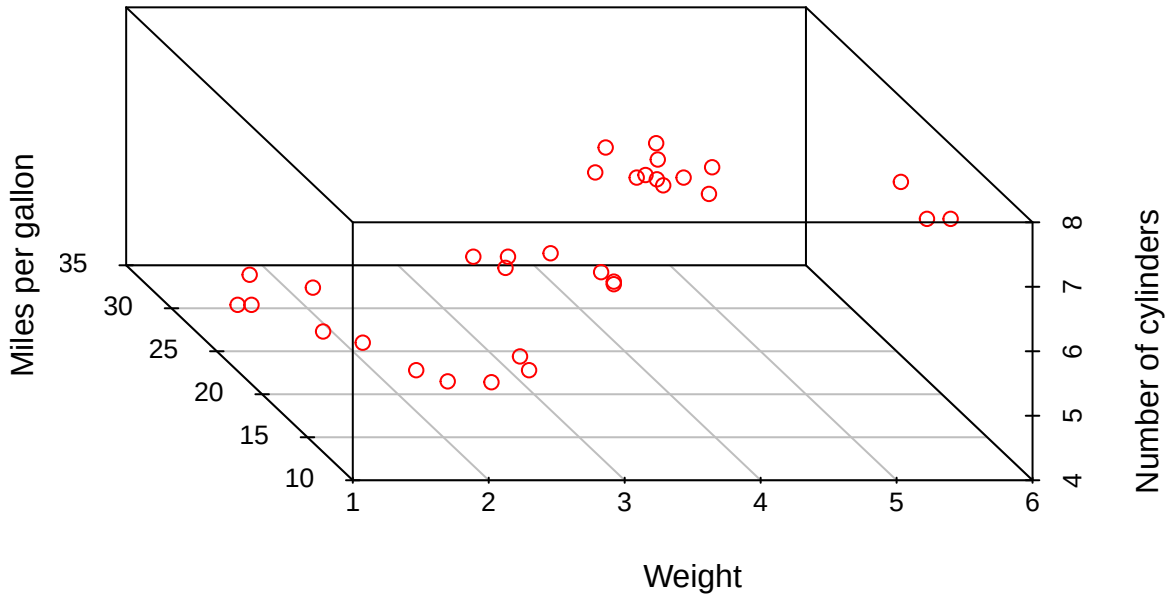



Plot 3D scatterplots from different angles including all of the variables (weight, miles per gallon, and number of cylinders).

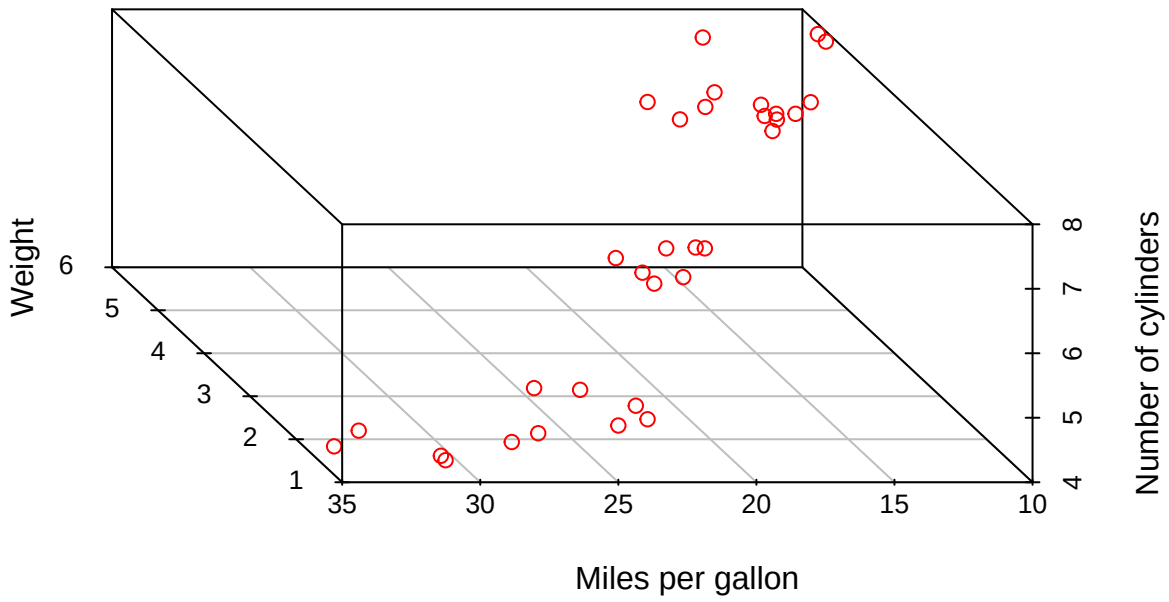
```
library("scatterplot3d")
scatterplot3d(x = cars$wt, y = cars$mpg, z = cars$cyl, xlab = "Weight",
              ylab = "Miles per gallon", zlab = "Number of cylinders",
              color = "red")
```




```
scatterplot3d(x = cars$wt, y = cars$mpg, z = cars$cyl, xlab = "Weight",
              ylab = "Miles per gallon", zlab = "Number of cylinders",
              color = "red", angle = 120)
```

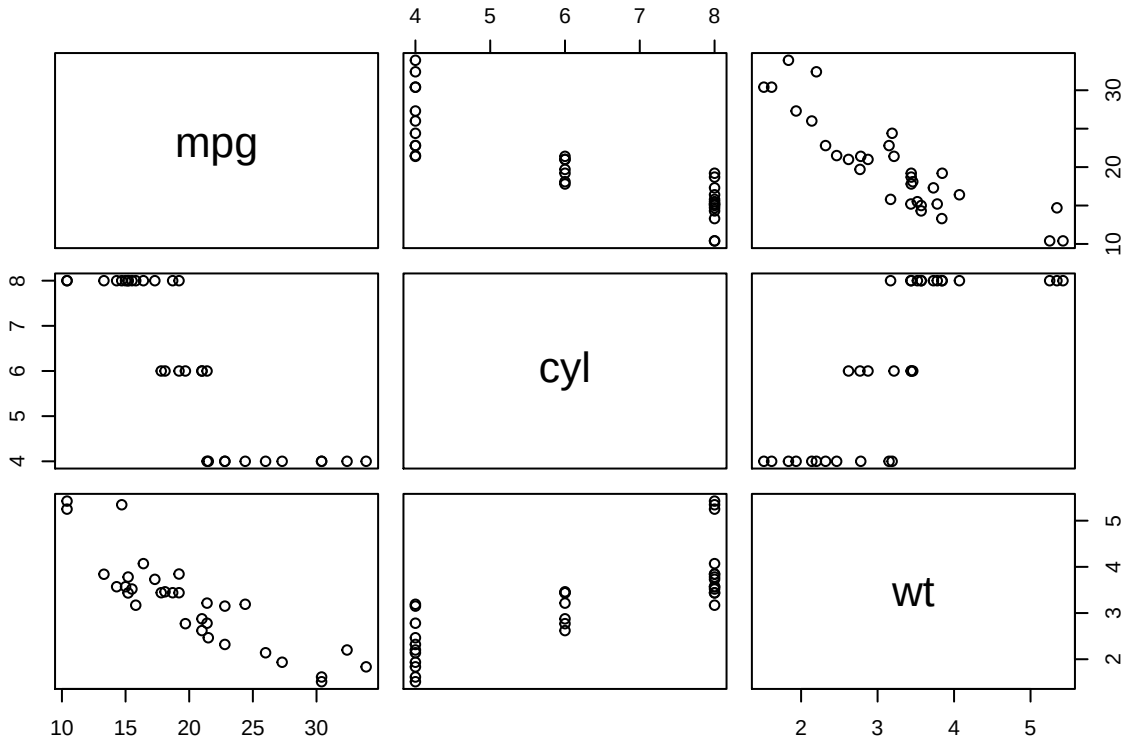



```
scatterplot3d(x = cars$wt, y = cars$mpg, z = cars$cyl, xlab = "Weight",
              ylab = "Miles per gallon", zlab = "Number of cylinders",
              color = "red", angle = 300)
```




Plot a group of scatterplot graphs based on comparing each of the variables with one another.

```
pairs(cars[,c(-1)])
```




Find the sample means for each column. Check that this is equivalent to the population mean by dividing the sum of the columns by the number of variables (n). Notice they are equivalent, and thus the sample mean is unbiased.

```
colMeans(cars[,-1])
```

```
##      mpg      cyl       wt
## 20.09062  6.18750  3.21725
```

```
n = dim(cars)[1]
sum(cars$mpg)/n
```

```
## [1] 20.09062
```

```
sum(cars$cyl)/n
```

```
## [1] 6.1875
```

```
sum(cars$wt)/n
```

```
## [1] 3.21725
```

This is simply to put the mean and standard deviation next to each other, and make sure the data isn't too outrageous or contains many outliers. For each variable the standard deviation is around 1/3 the mean, which is an appropriate amount. We put these results in a matrix for later use.

```r
muSD = as.matrix(cbind(apply(cars[,-1],2,mean),apply(cars[,-1],2,sd)))
muSD
```

```
##         [,1]      [,2]
## mpg 20.09062 6.0269481
## cyl  6.18750 1.7859216
## wt   3.21725 0.9784574
```

Here, we calculate the co-variance between each variable and put the results in a matrix for later use. The sample variances are on the diagonal of the matrix. Notice, for more legitimacy, that the diagonal variances here correspond to the square of the standard deviations from the previous section. Lastly, we must take the expectation of the sample variances to estimate the population variances. Notice they are equivalent and thus these sample variances are unbiased. Negative, non-diagonal entries are for negative correlation, as positive non-diagonal entries are for positive correlation.

```r
CovMat = as.matrix(cov(cars[,-1]))
CovMat
```

```
##           mpg       cyl        wt
## mpg 36.324103 -9.172379 -5.116685
## cyl -9.172379  3.189516  1.367371
## wt  -5.116685  1.367371  0.957379
```

```r
mean(var(cars[,2]))
```

```
## [1] 36.3241
```

```r
mean(var(cars[,3]))
```

```
## [1] 3.189516
```

```r
mean(var(cars[,4]))
```

```
## [1] 0.957379
```

Here we take the sample correlation between each of the variables. Negative values equate to a negative correlation while positive values equate to a positive correlation. The closer the absolute value of the entry is to 1 the stronger the correlation. A value of 0 would indicate no correlation at all, but we see none here. We must also estimate the population correlation by plugging in the sample co-variance and sample variances into the correlation equation. Notice that the numbers are equivalent, and thus the sample correlation is unbiased. (Use the coordiantes next to CovMat to determine which correlation we're calculating for).

```r
cor(cars[,-1])
```

```
##            mpg        cyl         wt
## mpg  1.0000000 -0.8521620 -0.8676594
## cyl -0.8521620  1.0000000  0.7824958
## wt  -0.8676594  0.7824958  1.0000000
```

```
CovMat[1,2]/(muSD[1,2]*muSD[2,2])
```

```
##       mpg
## -0.852162
```

```
CovMat[1,3]/(muSD[1,2]*muSD[3,2])
```

```
##        mpg
## -0.8676594
```

```
CovMat[2,3]/(muSD[2,2]*muSD[3,2])
```

```
##       cyl
## 0.7824958
```

Here, we check if the number of cylinders has an effect on the relationship between the weight and miles per gallon. We notice that there is a different sample correlation for each of the 3 different data sets (4 cylinders, 6 cylinders, and 8 cylinders).

```
cor(cars4[,c(2,4)])
```

```
##            mpg         wt
## mpg  1.0000000 -0.7131848
## wt  -0.7131848  1.0000000
```

```
cor(cars6[,c(2,4)])
```

```
##            mpg         wt
## mpg  1.0000000 -0.6815498
## wt  -0.6815498  1.0000000
```

```
cor(cars8[,c(2,4)])
```

```
##           mpg        wt
## mpg  1.000000 -0.650358
## wt  -0.650358  1.000000
```

## Conclusion

We were trying to determine the relationship between the weight of a car and the miles per gallon it gets. We also wanted to know if cylinders had an effect on that relationship. While it was fairly simple to see from a scatterplot that weight and miles per gallon had an inverse relationship, we couldn't quantify it and the effect of cylinders was even harder to determine. Through the use of sample mean, variance, and correlation, we got a better understanding that weight and miles per gallon had a very strong negative correlation. By separating the data into subsets based on cylinders, we noticed that although not extraordinarily, the number of cylinders did affect that weight:miles per gallon correlation. Namely, the more cylinders there were, the weaker the negative correlation. One issue is that the correlations of weight:miles per gallon based on cylinders were relatively close, so it wasn't fully conclusive that they caused the change. Another setback is that there was quite a bit of data left out of this project, and thus, there were almost definitely other factors that influenced the correlation. However, one thing we know, is that based on the equations from class, we were able to determine that these sample means, variances, and correlations were unbiased compared to the population and were therefore accurate measurements. All in all, we can nearly guarantee that more weight correlates to less miles per gallon, and it seems cylinders do affect this ratio even if by very little.