mbruner3_assign2

Mark Bruner

10/27/2020

```
rm(list=ls())
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(colorspace)
```

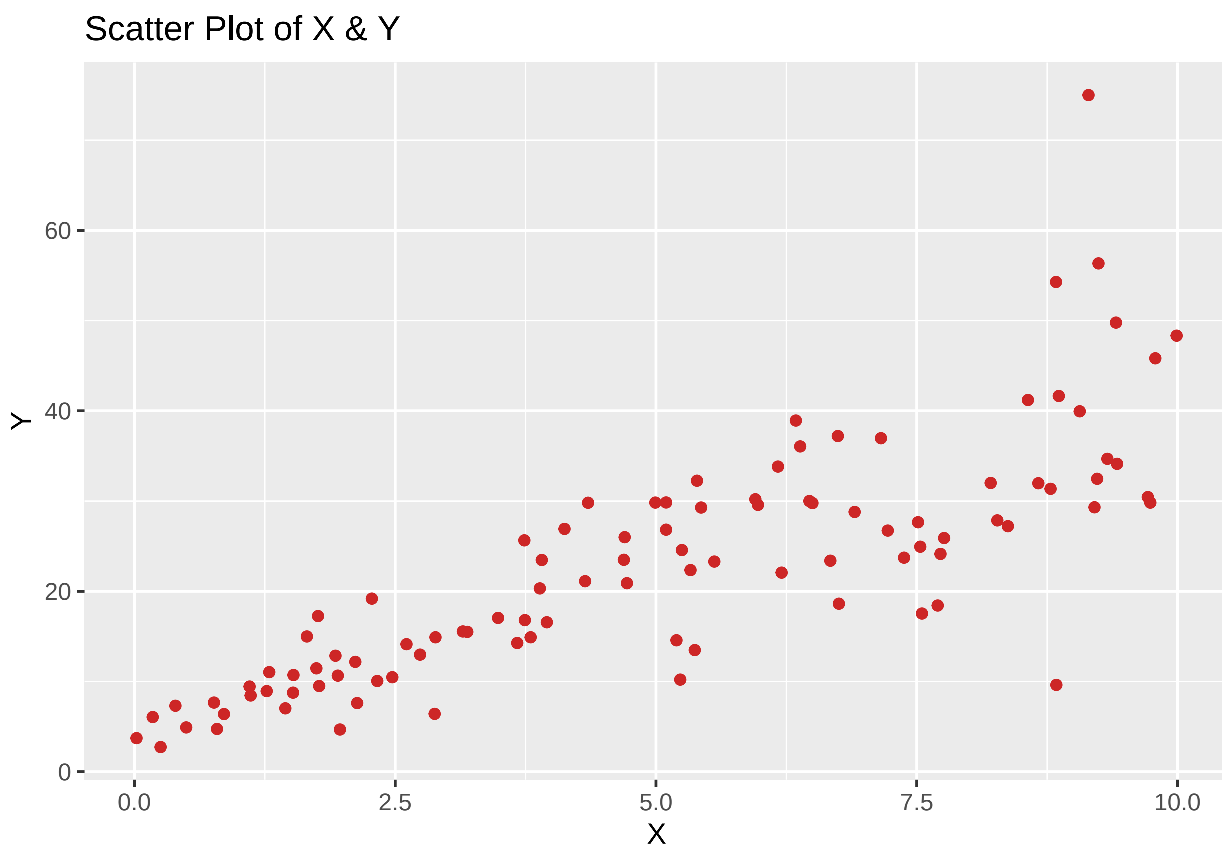
QUESTION 1

part a

```
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y

X <- as.data.frame(X)
Y <- as.data.frame(Y)
table <- cbind(X, Y)

table %>%
ggplot(mapping = aes(x = X, y = Y)) +
  geom_point(colour = "firebrick3") +
  labs(title = "Scatter Plot of X & Y")
```



Yes we will be able to fit a linear model to this data. The reason is, in general, as x increases so does y. Therefore, that implies that there is a relationship between x and y making it possible to create a linear mapping function that fits the data.

part b

```
lin_reg <- lm(Y~ X, table)
lin_reg
```

```
##
## Call:
## lm(formula = Y ~ X, data = table)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611
```

The model equation that explains y to x: $\mathbf{y = 3.611x + 4.465}$. **For accuracy of the model see part c.**

part c: note: includes accuracy from part b

```r
summary(lin_reg)
```

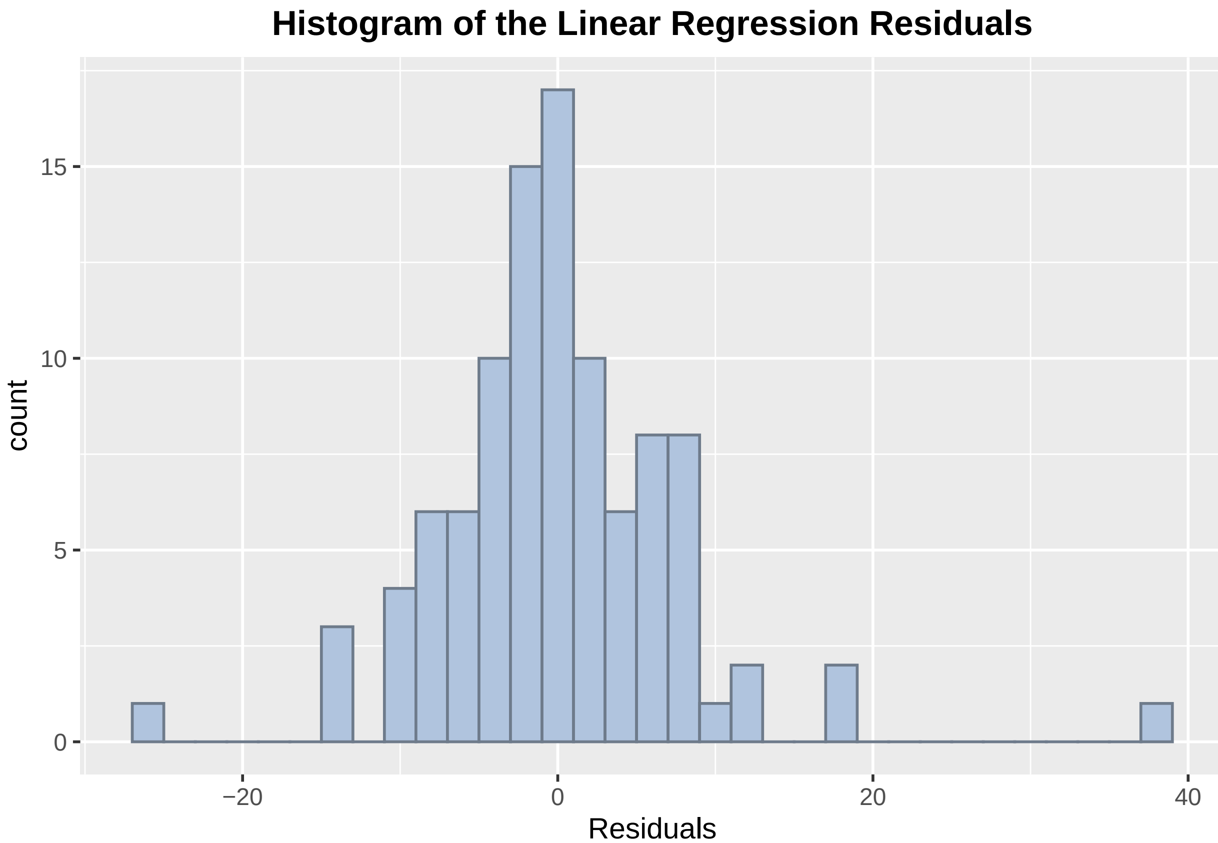
```
##
## Call:
## lm(formula = Y ~ X, data = table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

The r^2 is 65%, meaning that 65% of the variability of Y is captured by it captured by X.

EXTRA INVESTIGATIONS/EXPLORATION

I decided to use some of the concepts in class to further explore and practice. You can skip the next couple of graphs as they do not pertain to this assignment.

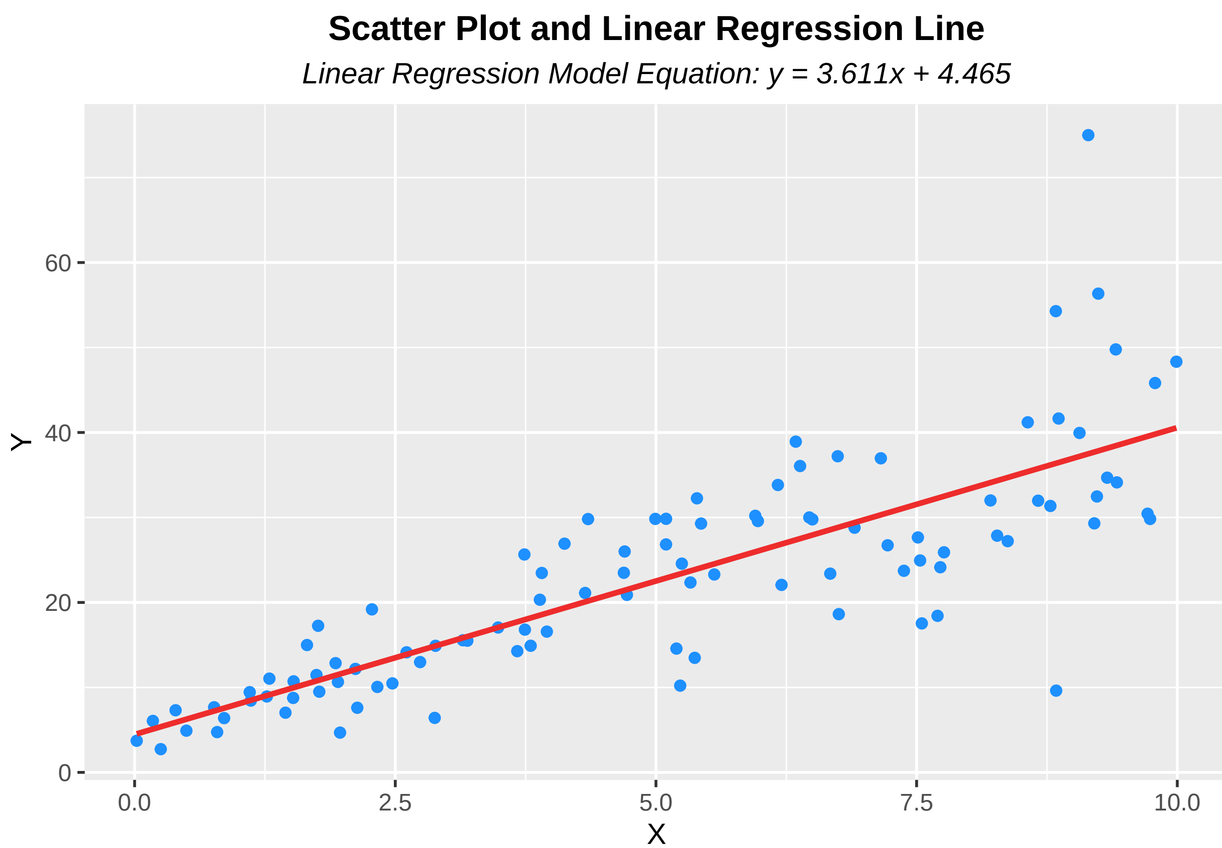
```r
lin_reg %>%
ggplot(mapping = aes(x = lin_reg$residuals)) +
  geom_histogram(colour = "lightsteelblue4", fill = "lightsteelblue", binwidth = 2) +
  labs(title = "Histogram of the Linear Regression Residuals") +
  xlab("Residuals") +
  theme(plot.title = element_text(face = "bold", hjust = .5))
```



The above graph shows a fairly normal residual distribution with maybe a couple of outliers.

```
table %>%
ggplot(mapping = aes(x = X, y = Y), ) +
  geom_point(colour = "dodgerblue") +
   stat_smooth(method = "lm", colour = "firebrick2", se = FALSE) +
  labs(title = "Scatter Plot and Linear Regression Line", subtitle = "Linear Regression Model Equation: 
  theme(plot.title = element_text(face = "bold", hjust = .5), plot.subtitle = element_text(face = "itali
```

```
## `geom_smooth()` using formula 'y ~ x'
```



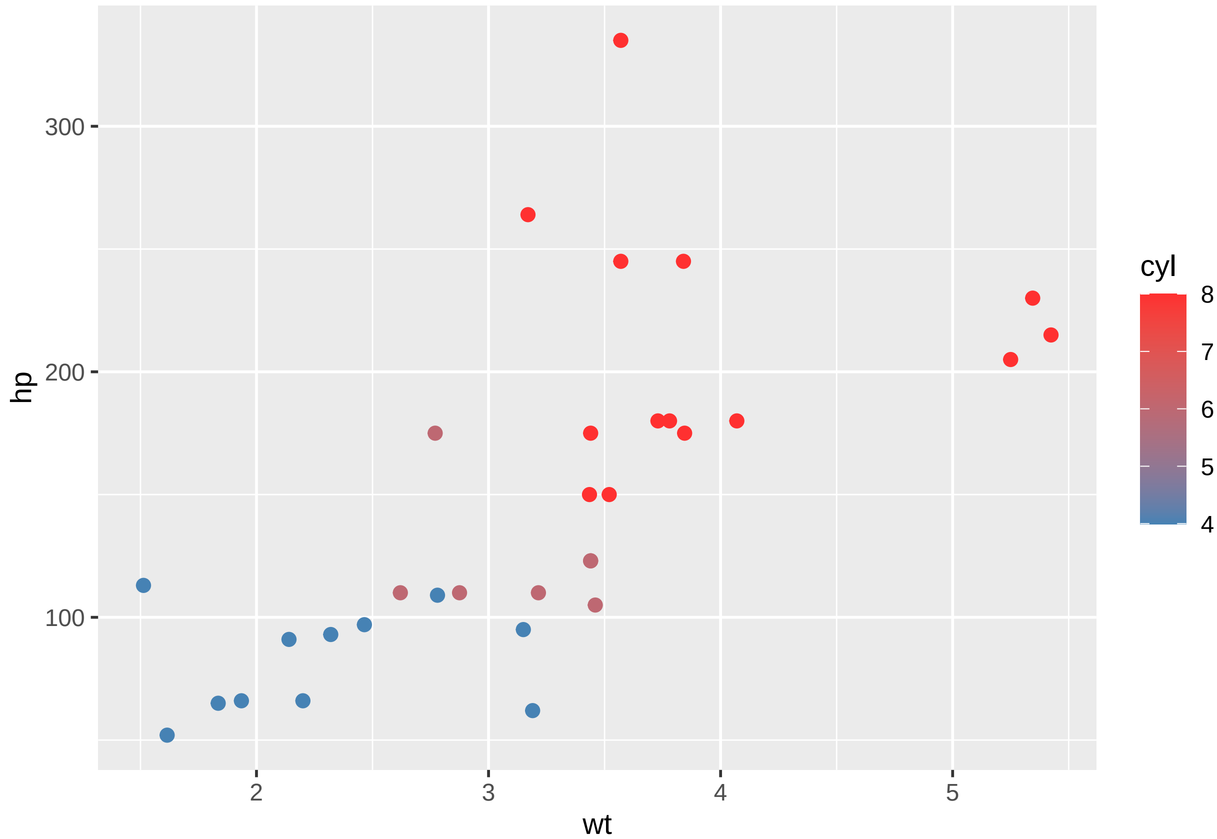
QUESTION 2

part a

HP as a function of Weight

```
cars <- mtcars

cars %>%
ggplot(mapping = aes(x = wt, y = hp, colour = cyl)) +
  geom_point(size = 2) +
    scale_color_gradient(low = "steelblue", high = "firebrick1")
```



My initial observation on the above graph is that the two are not strongly related. As x increases y increase to about x = 3 there seems to be a relationship but after 3 the points become more scattered and more spread out.

Linear regression formula for hp ~ wt

```
lin_reg <- lm(hp ~ wt, cars)
lin_reg
```

```
##
## Call:
## lm(formula = hp ~ wt, data = cars)
##
## Coefficients:
## (Intercept)           wt
##      -1.821       46.160
```

R^2

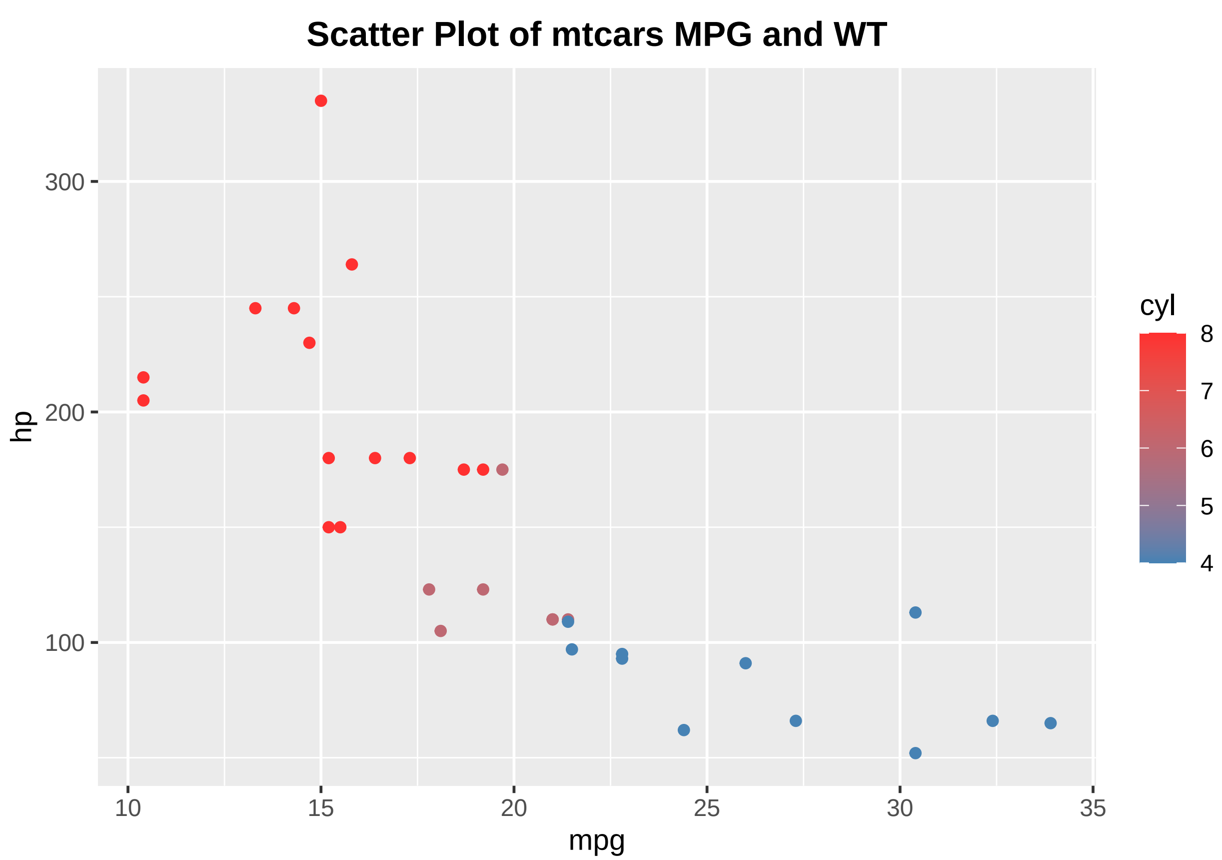
```
summary(lin_reg)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = cars)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## wt            46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

HP as a function of MPG

```r
cars %>%
ggplot(mapping = aes(x = mpg, y = hp, colour = cyl)) +
  geom_point() +
  labs(title = "Scatter Plot of mtcars MPG and WT") +
  theme(plot.title = element_text(face = "bold", hjust = .5)) +
    scale_color_gradient(low = "steelblue", high = "firebrick1")
```



There seems to be a stronger correlation between hp ~ mpg due to as x increases y decreases, generally.

```r
lin_reg <- lm(hp ~ mpg, cars)
lin_reg
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = cars)
##
## Coefficients:
## (Intercept)          mpg
##      324.08        -8.83
```

```r
summary(lin_reg)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mpg            -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 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
```

The answer is that MPG is a better predictor for HP than weight. 60% of the variance in the HP can be explained by the MPG of a car. Comparatively, only 43% of the variance in HP can be explained by the weight of a car.

part b

```r
lin_reg <- lm(hp ~ cyl + mpg, cars)
lin_reg
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = cars)
##
## Coefficients:
## (Intercept)          cyl          mpg
##      54.067       23.979       -2.775
```

y = 23.979x1 - 2.775x2 + 54.067

```
summary(lin_reg)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## cyl           23.979      7.346   3.264  0.00281 **
## mpg           -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

71% of the variance in HP can be explained by the number of cylinders and mpg of a car. Adding cylinders as a variable increased the predictive power of this model by ~10%. I would say that is an improvement!

```
23.979*4 - 2.775*22 + 54.067
```

```
## [1] 88.933
```

A car with 4 cylinders and 22 MPG will have about 89 HP.

QUESTION 3

```
library(mlbench)
data(BostonHousing)
```

```
BostonHousing %>%
  select(medv, crim, zn, ptratio, chas) -> bos_median

lm(medv ~., data = bos_median) -> bos_reg
bos_reg
```

```
##
## Call:
## lm(formula = medv ~ ., data = bos_median)
##
## Coefficients:
## (Intercept)         crim           zn      ptratio        chas1
##    49.91868     -0.26018      0.07073     -1.49367      4.58393
```

```r
summary(bos_reg)
```

```
##
## Call:
## lm(formula = medv ~ ., data = bos_median)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

I would say probably not based on the r-squared for the model, which only 36% of the variance in the median house price is accounted for by the crime, zoning, teacher-student ratio, and the Chas River. All of the variables are statistically significant at significant levels at 0!

** part b.I**

The house that bounds the Chas River would be $4,580 more expensive than the house that does not bound the Chas River.

** part b.II**

```r
-1.4937*15
```

```
## [1] -22.4055
```

```r
-1.4937*18
```

```
## [1] -26.8866
```

```r
-1.4937*15 - -1.4937*18
```

```
## [1] 4.4811
```

The house that resides in the neighborhood where the stud/teacher ratio is lower (15:1) would be 4.48 (thousand dollars) more expensive than the one that has 18:1 student/teacher.

part c

```r
summary(lm(medv ~., data = bos_median))
```

```
##
## Call:
## lm(formula = medv ~ ., data = bos_median)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

Crime, zone, teacher/student rati, and chas river are statiscally significant at a significance level at 0 (***).

```r
anova(lm(medv ~., data = bos_median))
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The order of importance of these variables is as follow (most importance to least): 1) Crime rate 2) Student:Teacher 3) Zone 4) Chas River