Problem_Set_3_Markdown

1

a.

A cubic function would overfit the traing data and lead us to reject or accept the null hypothesis incorrctly. Therefore I would say it is safe to assume that a linear model would be preferable than the cubic one as it would not overfit the training data.

b.

Since the cubic function overfits and the testing data is loaded with errors, you expect to have a lower RSS using linear as opposed to cubic.

2.

a.

There are 14 columns/variables in the data set and there are 506 observations.

```
library(MASS)
library(corrplot)
```

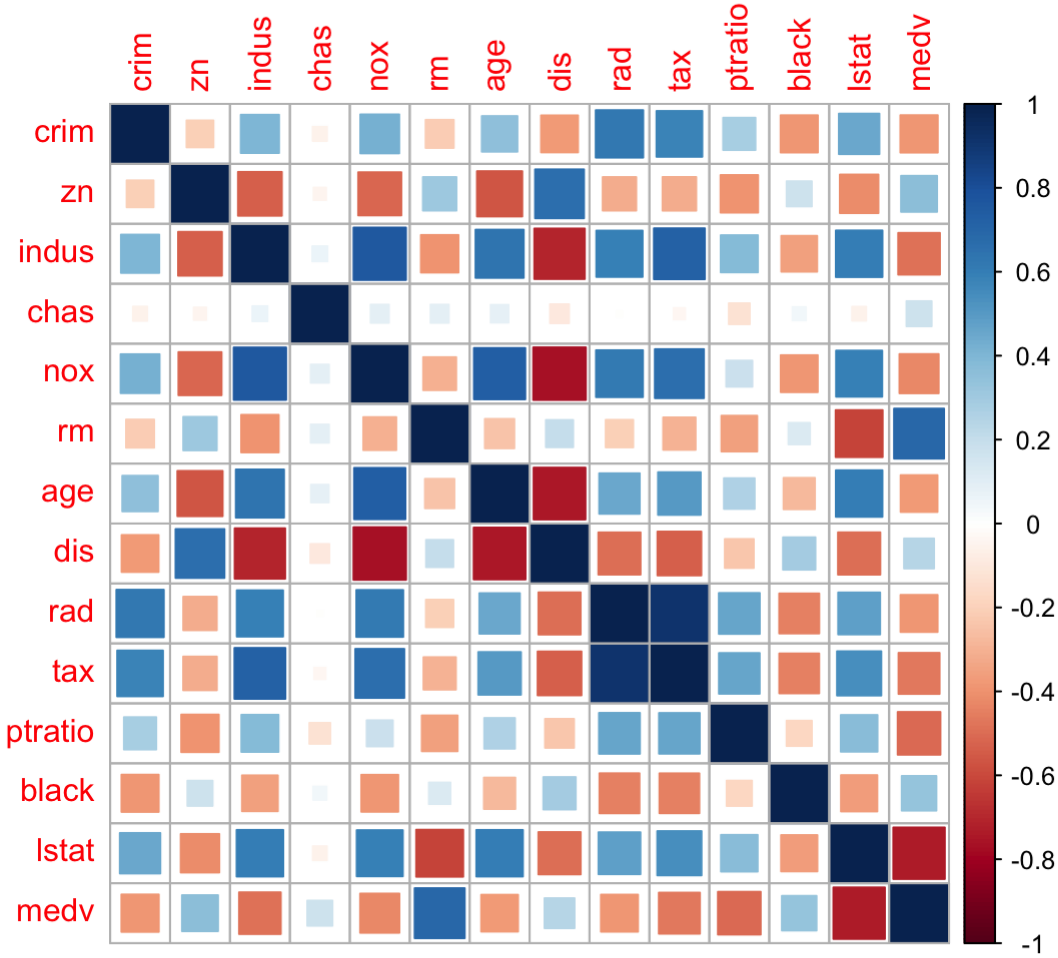
```
## corrplot 0.84 loaded
```

```
help(Boston)
```

b.

lstat ptratio rm indus

```
bostonplot <- cor(Boston)
corrplot(bostonplot, method='square')
```



c.

```
mod1 <- lm(medv ~ lstat + ptratio + rm + indus, data = Boston)
summary(mod1)
```

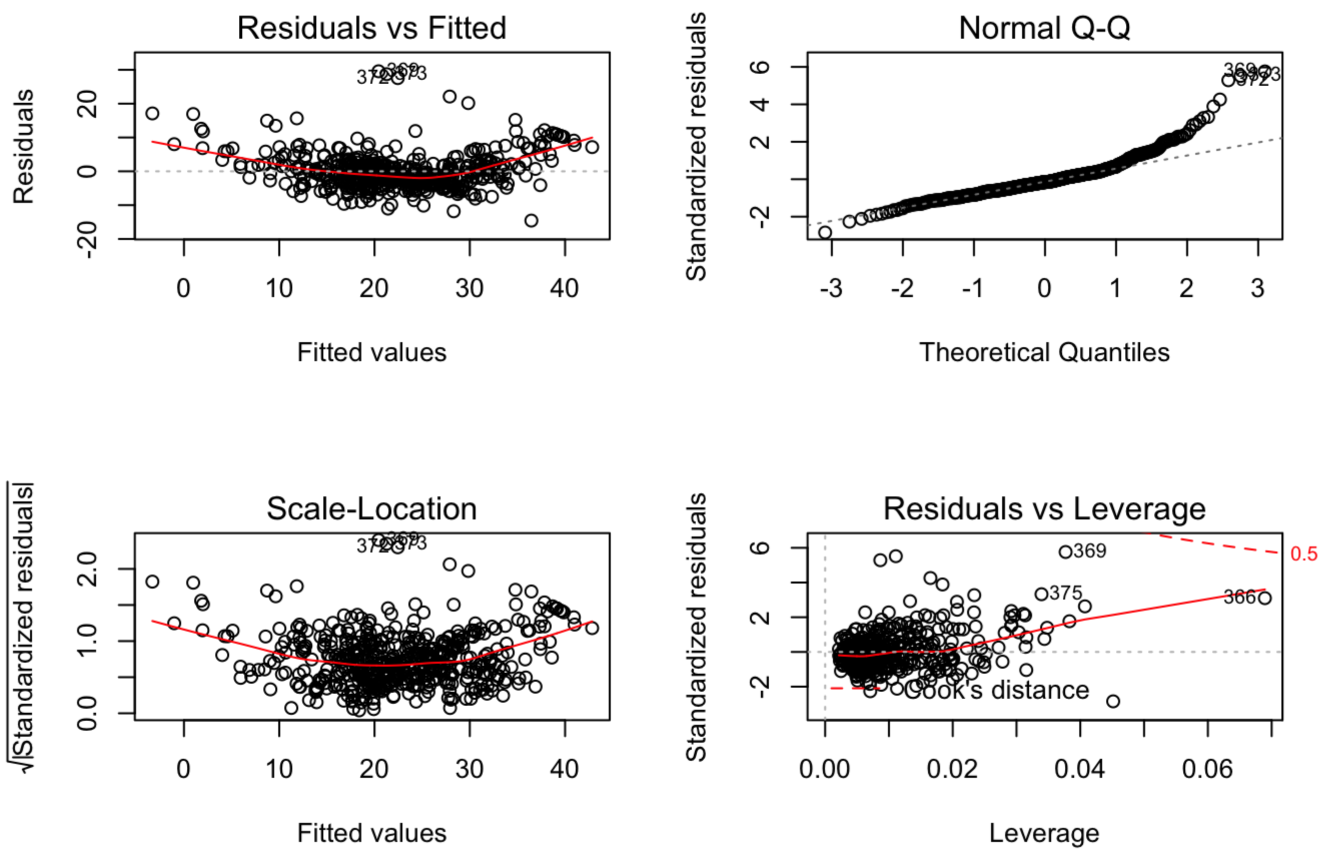
```
##
## Call:
## lm(formula = medv ~ lstat + ptratio + rm + indus, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5602  -3.1379  -0.7984   1.7783  29.5739
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.614970   3.926680   4.741 2.78e-06 ***
## lstat       -0.575711   0.047885 -12.023  < 2e-16 ***
## ptratio     -0.935122   0.120464  -7.763 4.71e-14 ***
## rm           4.515179   0.426286  10.592  < 2e-16 ***
## indus        0.007567   0.043594   0.174    0.862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.234 on 501 degrees of freedom
## Multiple R-squared:  0.6786, Adjusted R-squared:  0.6761
## F-statistic: 264.5 on 4 and 501 DF,  p-value: < 2.2e-16
```

d.

lstat, ptratio, and rm reject the null hypothesis that median value of owner-occupied homes are not affected by these variables. The only variable that is doesn't reject the null hypothesis is indus because it has a high P-Value.

e.

```
par(mfrow = c(2,2))
plot(mod1)
```



f.

Yes there is evidence of heteroscadicity on the graphs. There appears to be a U shaped curve on 2 of the Fitted Values graphs.

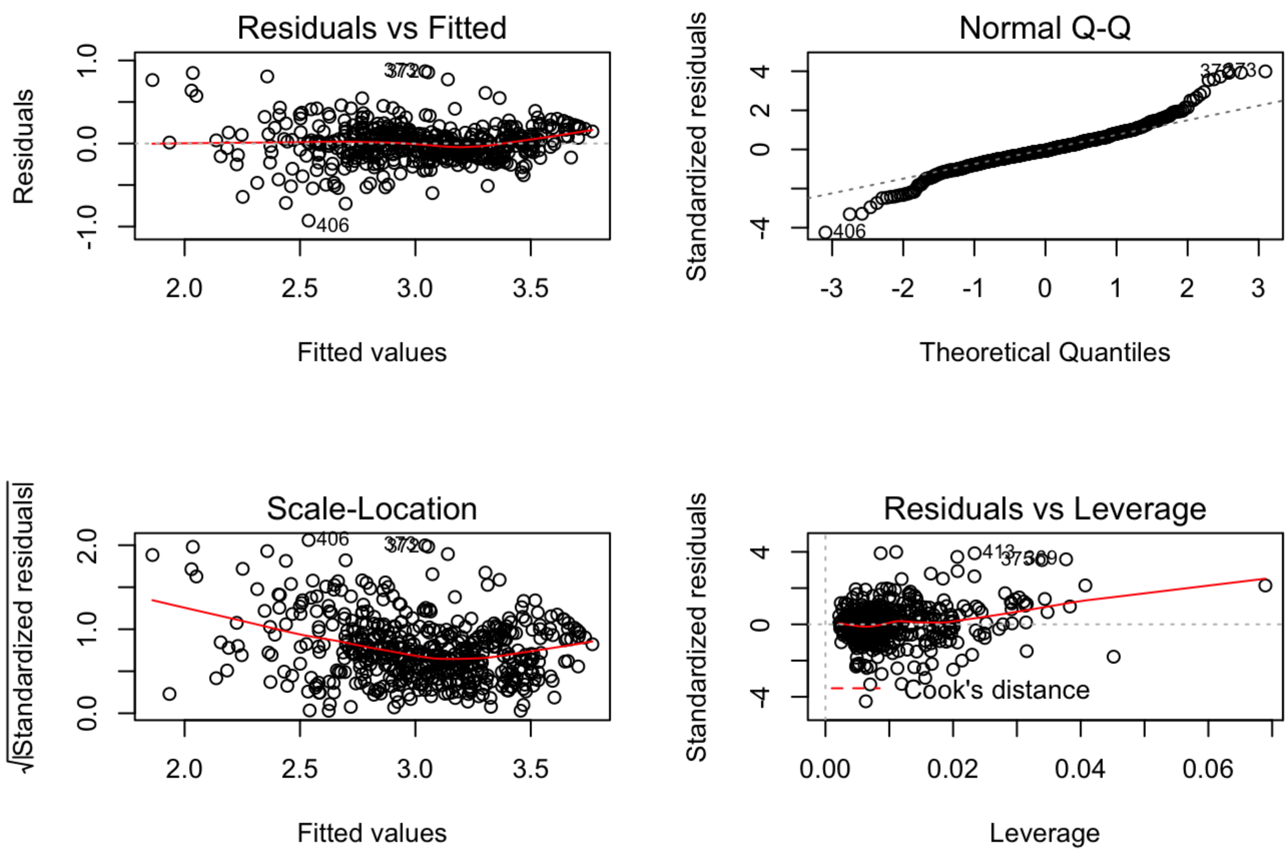
g.

```
loggeddata <- Boston$lnmedv <- log(Boston$medv)
lnmedv <- lm(loggeddata~ lstat + ptratio + rm + indus, data = Boston)
summary(lnmedv)
```

```
##
## Call:
## lm(formula = loggeddata ~ lstat + ptratio + rm + indus, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92790 -0.11001 -0.01274  0.10998  0.86993
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.535070   0.164366  21.507  < 2e-16 ***
## lstat       -0.034376   0.002004 -17.150  < 2e-16 ***
## ptratio     -0.037987   0.005042  -7.533 2.33e-13 ***
## rm           0.104442   0.017844   5.853 8.73e-09 ***
## indus       -0.001878   0.001825  -1.029    0.304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2191 on 501 degrees of freedom
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.7127
## F-statistic: 314.1 on 4 and 501 DF,  p-value: < 2.2e-16
```

h.

```
par(mfrow = c(2,2))
plot(lnmedv)
```



Yes there is still some heteroscdastity still prevalent in the data. It has been slighly reduced but there still evidence of it in the Scale-Location graph

i.

```
rmSq <-Boston$rmSq <- Boston$rm * Boston$rm
medgraph <- lm(lnmedv~ rm + rmSq + ptratio, data = Boston)
summary(medgraph)
```

```
##
## Call:
## lm(formula = lnmedv ~ rm + rmSq + ptratio, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1430 -0.1217  0.0590  0.1714  1.3125
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.855021   0.576644   6.685 6.15e-11 ***
## rm          -0.222718   0.176997  -1.258  0.20886
## rmSq         0.041036   0.013753   2.984  0.00299 **
## ptratio     -0.057533   0.006447  -8.924  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.291 on 502 degrees of freedom
## Multiple R-squared:  0.4962, Adjusted R-squared:  0.4932
## F-statistic: 164.8 on 3 and 502 DF,  p-value: < 2.2e-16
```

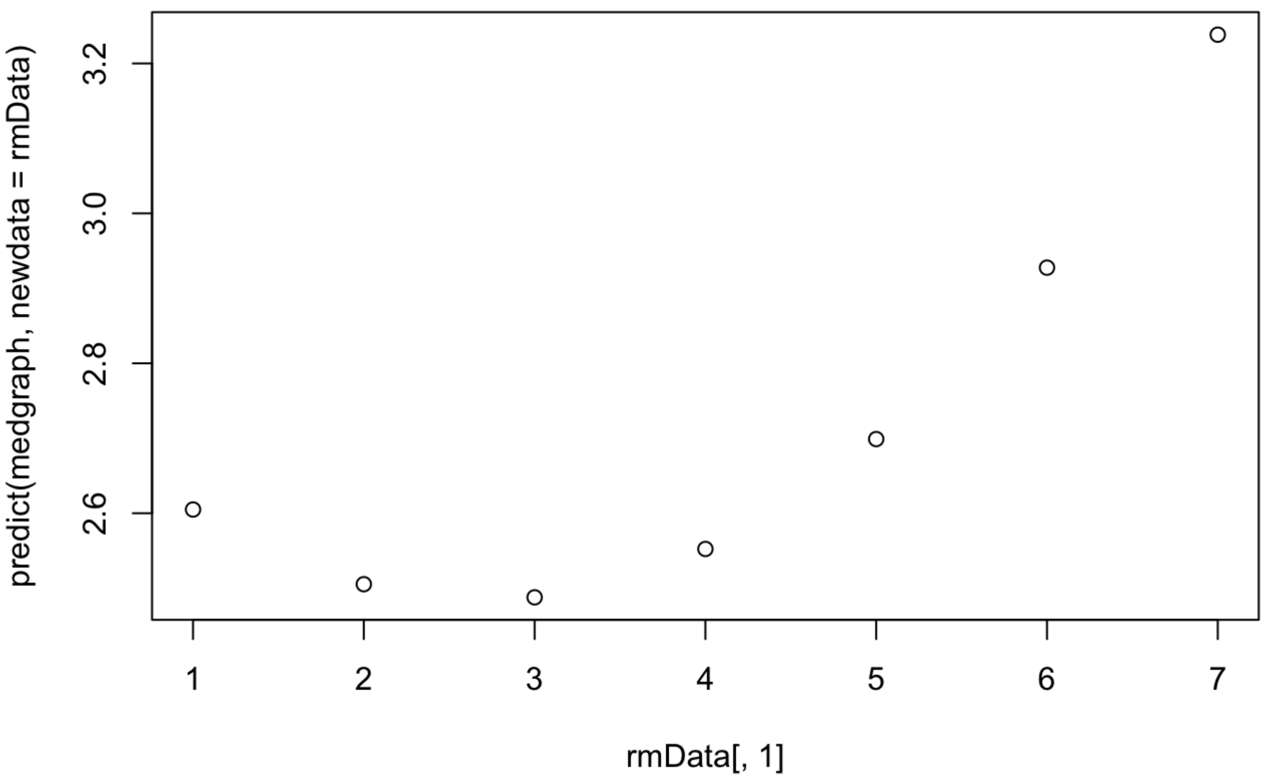
j

There is an increase in the housing values after 3 rooms. From 4 rms to 5 rms there appears to be a very drastic almost exponential increase in the predicted median value of the homes.

```
rmData <- data.frame(rm = 1:7, rmSq = 1:7 * 1:7, ptratio = rep(18.57,7))
summary(rmData)
```

```
##        rm           rmSq         ptratio
##  Min.   :1.0   Min.   : 1.0   Min.   :18.57
##  1st Qu.:2.5   1st Qu.: 6.5   1st Qu.:18.57
##  Median :4.0   Median :16.0   Median :18.57
##  Mean   :4.0   Mean   :20.0   Mean   :18.57
##  3rd Qu.:5.5   3rd Qu.:30.5   3rd Qu.:18.57
##  Max.   :7.0   Max.   :49.0   Max.   :18.57
```

```
plot(rmData[, 1], predict(medgraph, newdata = rmData))
```



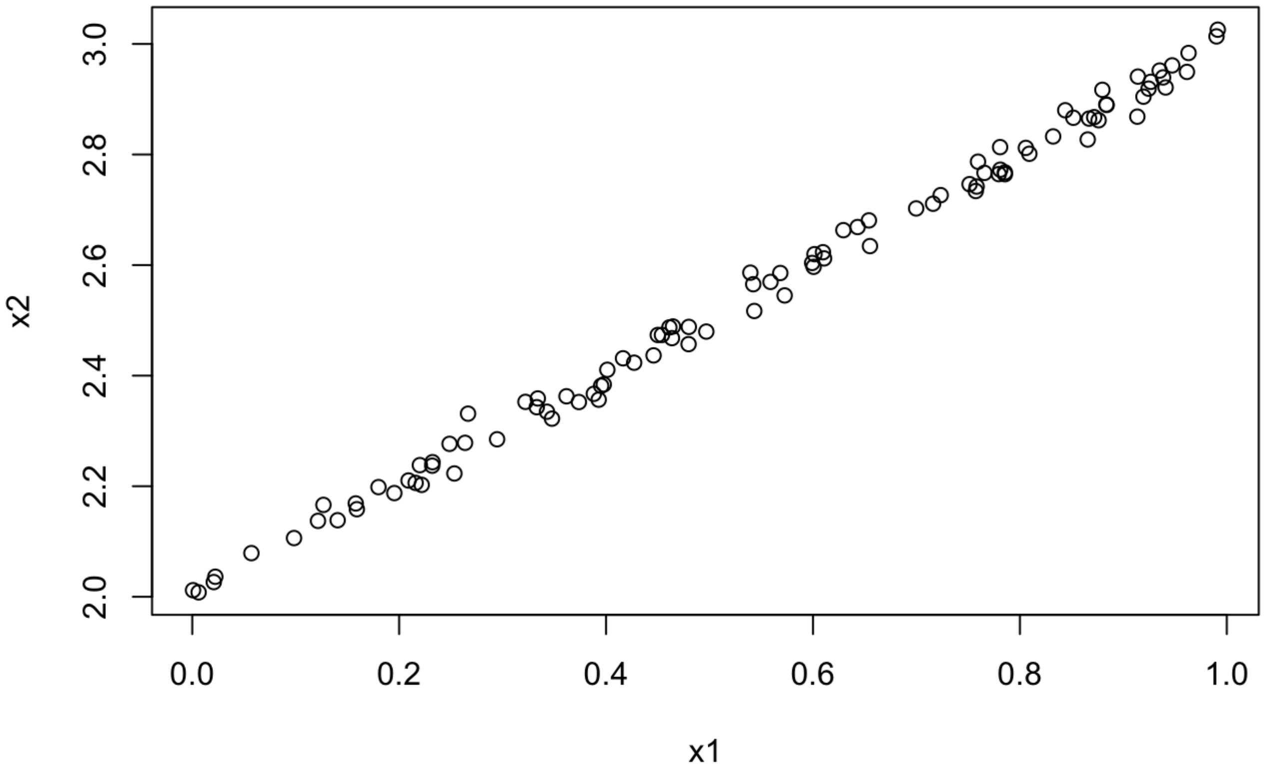
3

```
set.seed(1861)
x1 <- runif(100)
x2 <- 2 + x1 + rnorm(100, 0, 0.02)
Y <- 1 * x1 + 1 * x2 + rnorm(100)
DF <- data.frame(Y, x1, x2)
```

a.

The data, x1 and x2, appear to have a linear relationship.

```
plot(x1,x2)
```



b.

Yes X1 and X2 are very correlated. They have an r value of .99.

```
cor(x1,x2)
```

```
## [1] 0.9975058
```

c.

```
mod2 <- lm(Y~x1 + x2, data = DF)
summary(mod2)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2, data = DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58089 -0.66373 -0.02267  0.62852  2.25911
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.010      9.411   0.639    0.525
## x1             4.323      4.669   0.926    0.357
## x2            -2.114      4.687  -0.451    0.653
##
## Residual standard error: 0.9344 on 97 degrees of freedom
## Multiple R-squared:  0.3202, Adjusted R-squared:  0.3062
## F-statistic: 22.84 on 2 and 97 DF,  p-value: 7.422e-09
```

x1 = 4.323 x2 = -2.114

d.

The value of the coefficients should be 1 as indiciated on the Y equation. Y <- 1 * x1 + 1 * x2 + rnorm(100). 1 is the clear coefficient of x1and x2

e.

```
mod3 <- lm(Y~x1, data =  DF)
summary(mod3)
```

```
##
## Call:
## lm(formula = Y ~ x1, data = DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.67131 -0.62930 -0.04688  0.57545  2.33560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7672     0.2022   8.742 6.47e-14 ***
## x1            2.2224     0.3282   6.772 9.46e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9306 on 98 degrees of freedom
## Multiple R-squared:  0.3188, Adjusted R-squared:  0.3118
## F-statistic: 45.86 on 1 and 98 DF,  p-value: 9.46e-10
```

x1 = 2.22

```
mod4 <- lm(Y~x2, data = DF)
summary(mod4)
```

```
##
## Call:
## lm(formula = Y ~ x2, data = DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.76623 -0.61195 -0.04742  0.58785  2.40945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.6677     0.8484  -3.144   0.0022 **
## x2            2.2150     0.3306   6.701 1.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9338 on 98 degrees of freedom
## Multiple R-squared:  0.3142, Adjusted R-squared:  0.3072
## F-statistic:  44.9 on 1 and 98 DF,  p-value: 1.323e-09
```

x2 = 2.215

f.

X1 plotted against Y has has a higher R^2 and a lower a P value therefore X1 is my preferred data. There isn't a significant differece in the data as both have very similar R^2 I would prefer the one with a higher R value and the lowest P value.