DSCI445 - Homework 1

Your Name

Be sure to `set.seed(400)` at the beginning of your homework.

```
#reproducibility
set.seed(400)
```

R & ggplot2

```
## load the data
library(ggplot2)

## take a look
head(diamonds)
```

```
## # A tibble: 6 x 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

```
##################################
## Continue your analysis here ##
##################################
```

Regression

```
## load the data
library(MASS)

## take a look
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
```

```
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

Start by visually inspecting the data to get an idea of relationships that might be present (**hint:** look into the `ggpairs` function in the `GGally` package.). Describe what you see.

```
## from the hint
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## make plots and describe
```

Next fit linear models using the `lm()` function:

(a) For each predictor fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response?

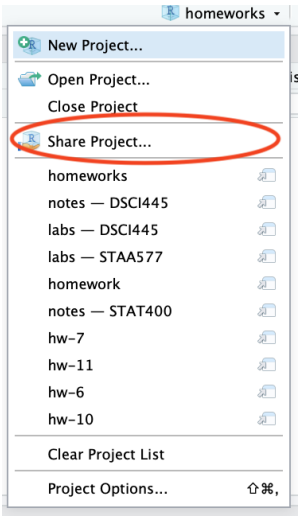
(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results (including diagnostic plots). For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point on the plot. Its coefficient in a simple linear regression model is shown as its x coordinate and its coefficient in a multiple linear regression model is shown as its y coordinate. Describe what you see.

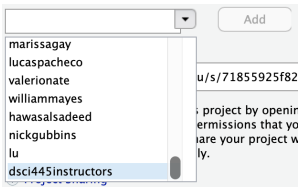
Turn in in a pdf of your analysis to canvas using the provided Rmd file as a template. Your Rmd file on the server will also be used in grading, so be sure they are identical.

Be sure to share your server project with the instructor and grader. You only need to do this once per semester.

1. Open your `homeworks` project on liberator.stat.colostate.edu
2. Click the drop down on the project (top right side) > Share Project...



3. Click the drop down and add "dsci445instructors" to your project.



This is how you **receive points** for reproducibility on your homework!