Task 2 - Assignment 1

Frederick Abi Chahine

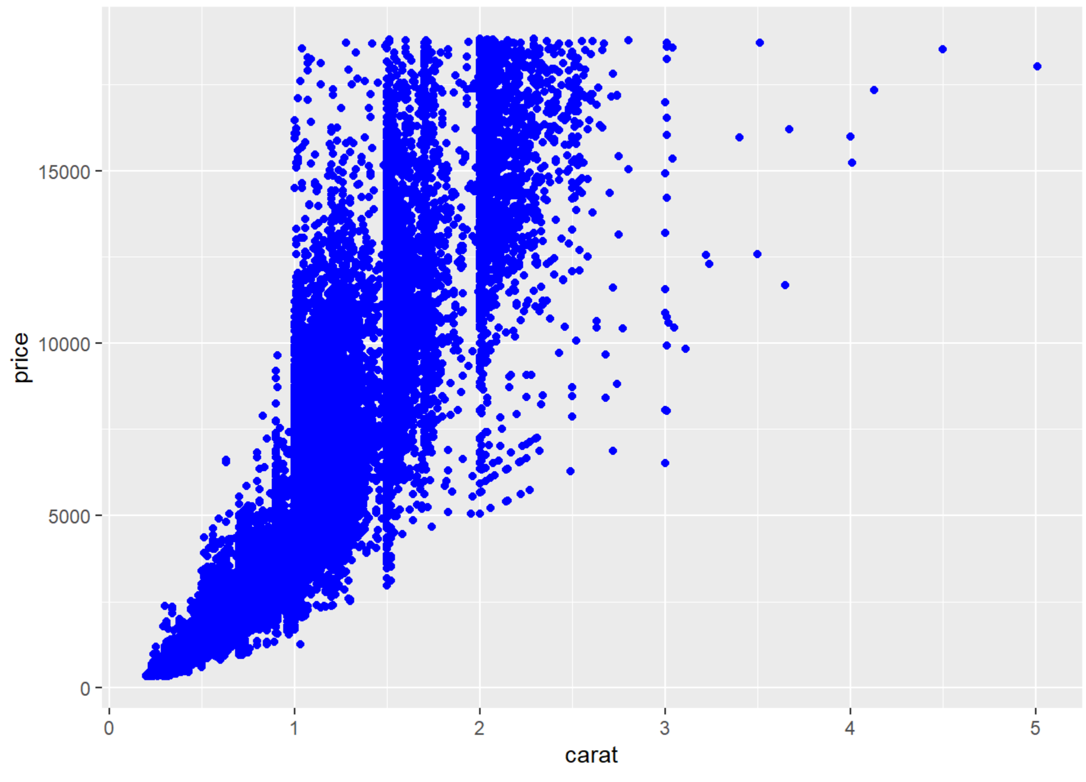
10/1/2021

```
library(ISLR2)
library(ggplot2)
```

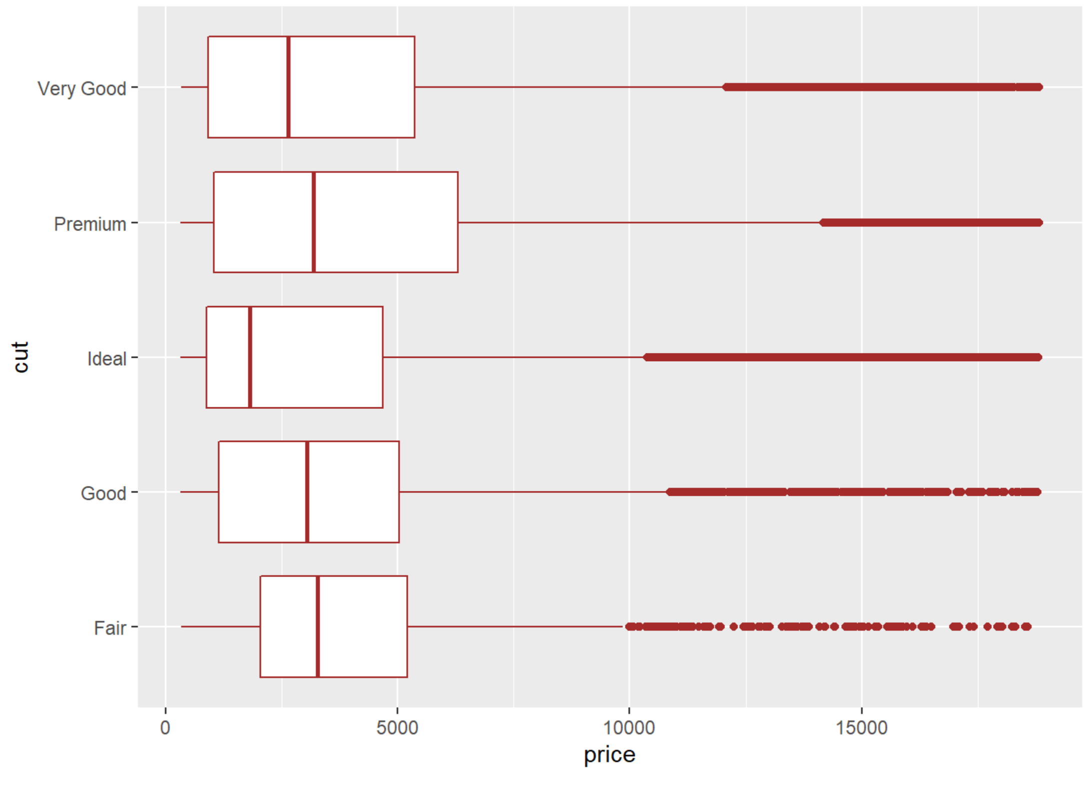
Exercise - 1

```
diamonds_data = read.csv("Diamonds.csv")

ggplot(data = diamonds_data, mapping = aes(x = carat, y = price)) + geom_point(col = "blue")
```




```
ggplot(data = diamonds_data, mapping = aes(x = price, y = cut)) + geom_boxplot(col = "brown")
```



Exercise - 2

```
lm_price_carat <- lm(price ~ carat, data = diamonds_data)
summary(lm_price_carat)
```

```
##
## Call:
## lm(formula = price ~ carat, data = diamonds_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18585.3   -804.8    -18.9    537.4  12731.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2256.36      13.06  -172.8   <2e-16 ***
## carat        7756.43      14.07   551.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1549 on 53938 degrees of freedom
## Multiple R-squared:  0.8493, Adjusted R-squared:  0.8493
## F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

```
#plot(lm_price_carat)
```

- Firstly, from the call we can see “lm” which means that we are dealing with a linear model. After that, we have the formula which shows that *price* is the Y variable and *carat* is the X variable, and the data frame that we are utilizing.
- From the residuals, we can see that the point furthest below the regression line is *-18585.3*, 25% of our residuals are less than *-804.8* (1st quartile), the median is *-18.9*, 25% of our residuals are greater than *537.4*, and that the point which is furthest above the regression line is *12731.7*.
- From the coefficients we can deduce that an increase of 1 in carat would result in a 7756.43 increase in price.
- We can see that *53938* data points went into the estimation of the parameter (DOF).
- We can also see that the standard deviation of the residuals is *1549*.
- From the Multiple R-squared, we can deduce that *84.93% of the variation in price can be explained by the carat*.
- The multiple R-squared is equal to the Adjusted R-squared in this case since we only have one predictor (simple linear regression).
- The F-statistic is *3.041e+05* which is very high and indicates that there is a relationship between the predictor (carat) variable and the response (price) variable.
- The p-value is < *2.2e-16* which is extremely low (<<0.05) and means that this model is statistically significant.

Exercise - 3

```
multiple_lm <- lm(price ~ carat + clarity + color, data = diamonds_data)
summary(multiple_lm)
```

```
##
## Call:
## lm(formula = price ~ carat + clarity + color, data = diamonds_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17310.9   -678.0   -192.2    473.0  10313.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6699.95      47.20 -141.94   <2e-16 ***
## carat        8856.23      12.10  731.86   <2e-16 ***
## clarityIF    5718.23      52.01  109.95   <2e-16 ***
## claritySI1   3795.47      44.50   85.30   <2e-16 ***
## claritySI2   2832.65      44.77   63.27   <2e-16 ***
## clarityVS1   4785.79      45.40  105.42   <2e-16 ***
## clarityVS2   4466.10      44.69   99.93   <2e-16 ***
## clarityVVS1  5351.85      48.03  111.42   <2e-16 ***
## clarityVVS2  5234.16      46.72  112.03   <2e-16 ***
## colorE       -216.45      18.53  -11.68   <2e-16 ***
## colorF       -314.92      18.72  -16.82   <2e-16 ***
## colorG       -509.09      18.33  -27.78   <2e-16 ***
## colorH       -985.01      19.49  -50.54   <2e-16 ***
## colorI      -1441.77      21.90  -65.84   <2e-16 ***
## colorJ      -2340.83      27.03  -86.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1170 on 53925 degrees of freedom
## Multiple R-squared:  0.914,  Adjusted R-squared:  0.9139
## F-statistic: 4.092e+04 on 14 and 53925 DF,  p-value: < 2.2e-16
```

- Firstly, from the call we can see “lm” which means that we are dealing with a linear model. After that, we have the formula which shows that *price* is the Y variable and that we have multiple predictors in which *carat* is X1, *clarity* is X2, and *color* is X3; and the data frame that we are utilizing.
- From the residuals, we can see that the point furthest below the regression line is *-17310.9*, 25% of our residuals are less than *-678.0* (1st quartile), the median is *-192.2*, 25% of our residuals are greater than *473.0*, and that the point which is furthest above the regression line is *10313.2*.
- From the coefficients we can deduce that an increase of 1 in carat would result in a 8856.23 increase in price. An increase of 1 in clarityIF would result in a 5718.23 increase in price etc… However, we note that all the color variables are negative since we can not truly increase color to increase price.
- We can see that *53925* data points went into the estimation of the parameter (DOF).
- We can also see that the standard deviation of the residuals is *1170*.
- From the Multiple R-squared, we can deduce that *91.4% of the variation in price can be explained by the interaction of the predictors*.
- The multiple R-squared is very close to the Adjusted R-squared here so it implies that we are NOT over fitting.
- The F-statistic is *4.092e+04* which is very high and indicates that there is a relationship between the predictor variables and the response variable.
- The p-value is < *2.2e-16* which is extremely low (<<0.05) and means that this model is statistically significant.

Exercise - 4

```
full_multiple_lm <- lm(price ~ carat + cut + clarity + color + depth + table + x + y + z, data = diamonds_data)
summary(full_multiple_lm)
```

```
##
## Call:
## lm(formula = price ~ carat + cut + clarity + color + depth +
##     table + x + y + z, data = diamonds_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21376.0   -592.4   -183.5    376.4  10694.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2184.477    408.197   5.352 8.76e-08 ***
## carat        11256.978     48.628 231.494  < 2e-16 ***
## cutGood        579.751     33.592  17.259  < 2e-16 ***
## cutIdeal       832.912     33.407  24.932  < 2e-16 ***
## cutPremium     762.144     32.228  23.649  < 2e-16 ***
## cutVery Good   726.783     32.241  22.542  < 2e-16 ***
## clarityIF     5345.102     51.024 104.757  < 2e-16 ***
## claritySI1    3665.472     43.634  84.005  < 2e-16 ***
## claritySI2    2702.586     43.818  61.677  < 2e-16 ***
## clarityVS1    4578.398     44.546 102.779  < 2e-16 ***
## clarityVS2    4267.224     43.853  97.306  < 2e-16 ***
## clarityVVS1   5007.759     47.160 106.187  < 2e-16 ***
## clarityVVS2   4950.814     45.855 107.967  < 2e-16 ***
## colorE        -209.118     17.893 -11.687  < 2e-16 ***
## colorF        -272.854     18.093 -15.081  < 2e-16 ***
## colorG        -482.039     17.716 -27.209  < 2e-16 ***
## colorH        -980.267     18.836 -52.043  < 2e-16 ***
## colorI       -1466.244     21.162 -69.286  < 2e-16 ***
## colorJ       -2369.398     26.131 -90.674  < 2e-16 ***
## depth          -63.806      4.535 -14.071  < 2e-16 ***
## table          -26.474      2.912  -9.092  < 2e-16 ***
## x            -1008.261     32.898 -30.648  < 2e-16 ***
## y                9.609     19.333   0.497    0.619
## z              -50.119     33.486  -1.497    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 53916 degrees of freedom
## Multiple R-squared:  0.9198, Adjusted R-squared:  0.9198
## F-statistic: 2.688e+04 on 23 and 53916 DF,  p-value: < 2.2e-16
```

- Firstly, from the call we can see “lm” which means that we are dealing with a linear model. After that, we have the formula which shows that *price* is the Y variable and that we have multiple predictors in which *carat* is X1, *clarity* is X2, *color* is X3, etc…; and the data frame that we are utilizing.
- From the residuals, we can see that the point furthest below the regression line is *-21376.0*, 25% of our residuals are less than *-592.4* (1st quartile), the median is *-183.5*, 25% of our residuals are greater than *376.4*, and that the point which is furthest above the regression line is *10694.2*.
- From the coefficients we can deduce that an increase of 1 in carat would result in a 11256.978 increase in price. An increase of 1 in clarityIF would result in a 5345.102 increase in price etc…
- We can see that *53916* data points went into the estimation of the parameter (DOF).
- We can also see that the standard deviation of the residuals is *1130*.
- From the Multiple R-squared, we can deduce that *91.98% of the variation in price can be explained by the interaction of the predictors*.
- The multiple R-squared is exactly equal (or extremely near) to the Adjusted R-squared here so it implies that we are NOT over fitting.
- The F-statistic is *2.688e+04* which is very high and indicates that there is a relationship between the predictor variables and the response variable.
- The *y and z predictors* show a high p-value (0.619 & 0.134 respectively) which could indicate that other predictors are masking / shadowing the significance of y and z by having correlations with them; This does NOT mean that y and z do not have a relationship with price, and it can be shown by performing 2 simple linear regressions in which both show a significant relationship with price.
- The p-value of the model is < *2.2e-16* which is extremely low (<<0.05) and means that this model is statistically significant.

```
#Proof for y and z:
lm_for_y <- lm(price ~ y, data = diamonds_data)
summary(lm_for_y)
```

```
##
## Call:
## lm(formula = price ~ y, data = diamonds_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152436   -1229    -241     838   31436
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13402.027     44.062  -304.2   <2e-16 ***
## y             3022.887      7.536   401.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1999 on 53938 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7489
## F-statistic: 1.609e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

```
lm_for_z <- lm(price ~ z, data = diamonds_data)
summary(lm_for_z)
```

```
##
## Call:
## lm(formula = price ~ z, data = diamonds_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139561   -1235    -240     825   32085
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13296.57      44.64  -297.9   <2e-16 ***
## z             4868.79      12.37   393.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2027 on 53938 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7417
## F-statistic: 1.549e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

```
#check correlations for y and each, z and each => find the predictors causing this.
```

Comparing to the previous model:
There are a few minor and subtle differences & similarities between the two, but the core comparison would be:
- They both are equally statistically significant as they have the same general p-value.
- EX4 has a lower F-stat than EX3 (although both are high) which could be due to the increase in the number of predictors.
- The R^2 of EX4 is very slightly higher (0.58 difference) than EX3 which indicates that a higher % of the variation in price comes from the interaction of predictors in EX4 than EX3.