Appraising Diamonds- IDS 575 Project

I started with exploring the dataset which contains 53940 rows with 11Variables. The first variable "X" is just serial number, hence dropped. Dataset have three categorical variables as follows:

Cut: IDEAL(BEST), PREMIUM, VERY GOOD, GOOD, FAIR(WORST) but levels in R are created alphabetically.
Clarity: IF(BEST), VVS2, VVS1, VS2, VS1, SI2, SI1, I1(WORST) but levels in R are created alphabetically.
Color: D(BEST), E, F, G, H, I, J(WORST) levels works as the priority is same in alphabetical order too.

Variable Description and Statistics:

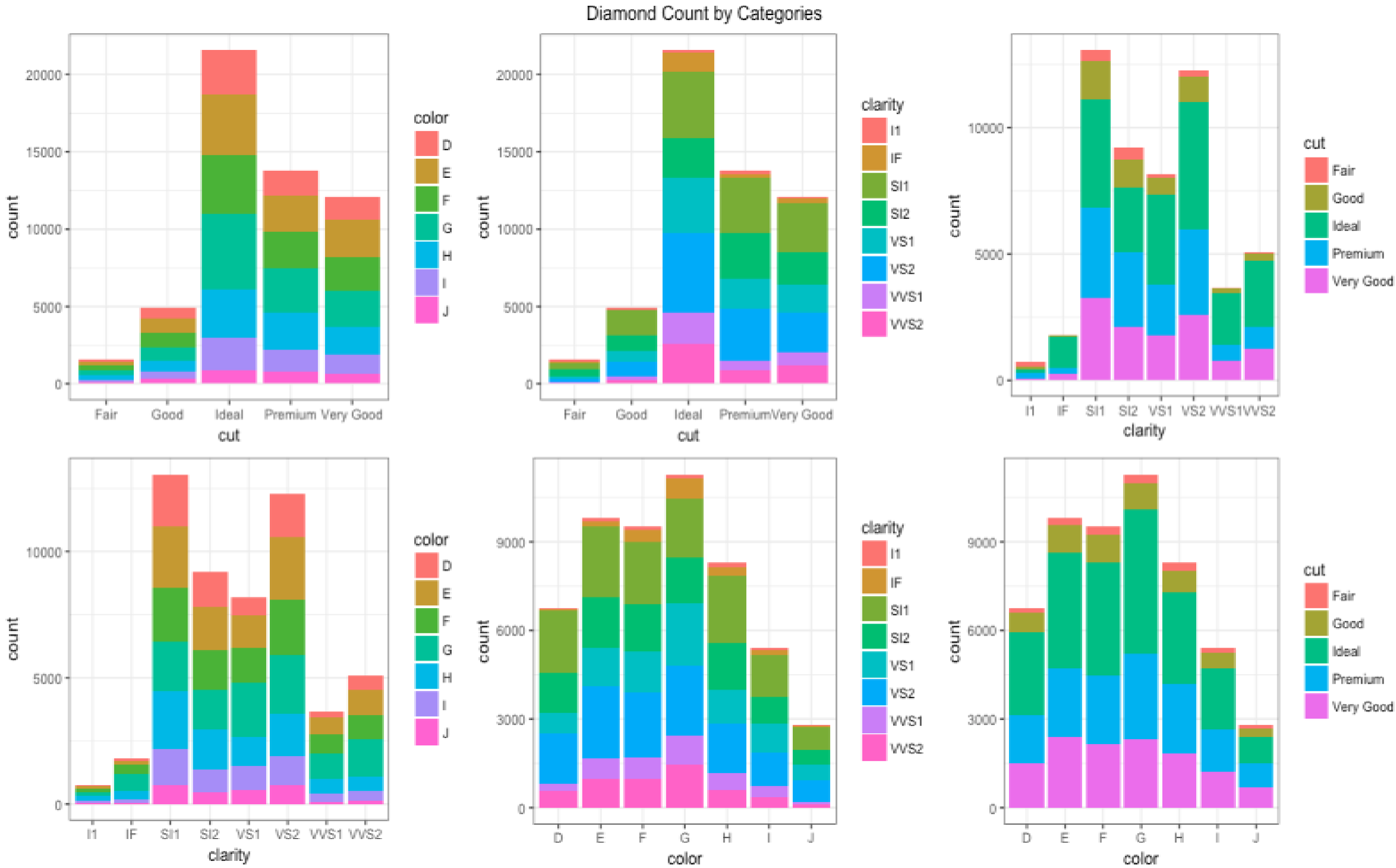
```
'data.frame':   53940 obs. of  10 variables:
 $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Factor w/ 5 levels "Fair","Good",..: 3 4 2 4 2 5 5 5 1 5 ...
 $ color  : Factor w/ 7 levels "D","E","F","G",..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Factor w/ 8 levels "I1","IF","SI1",..: 4 3 5 6 4 8 7 3 6 5 ...
 $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

```
     carat              cut        color        clarity          depth      
 Min.   :0.2000   Fair     : 1610   D: 6775   SI1     :13065   Min.   :43.00  
 1st Qu.:0.4000   Good     : 4906   E: 9797   VS2     :12258   1st Qu.:61.00  
 Median :0.7000   Ideal    :21551   F: 9542   SI2     : 9194   Median :61.80  
 Mean   :0.7979   Premium  :13791   G:11292   VS1     : 8171   Mean   :61.75  
 3rd Qu.:1.0400   Very Good:12082   H: 8304   VVS2    : 5066   3rd Qu.:62.50  
 Max.   :5.0100                     I: 5422   VVS1    : 3655   Max.   :79.00  
                                    J: 2808   (Other): 2531                   
     table           price             x                y                z         
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800  
```

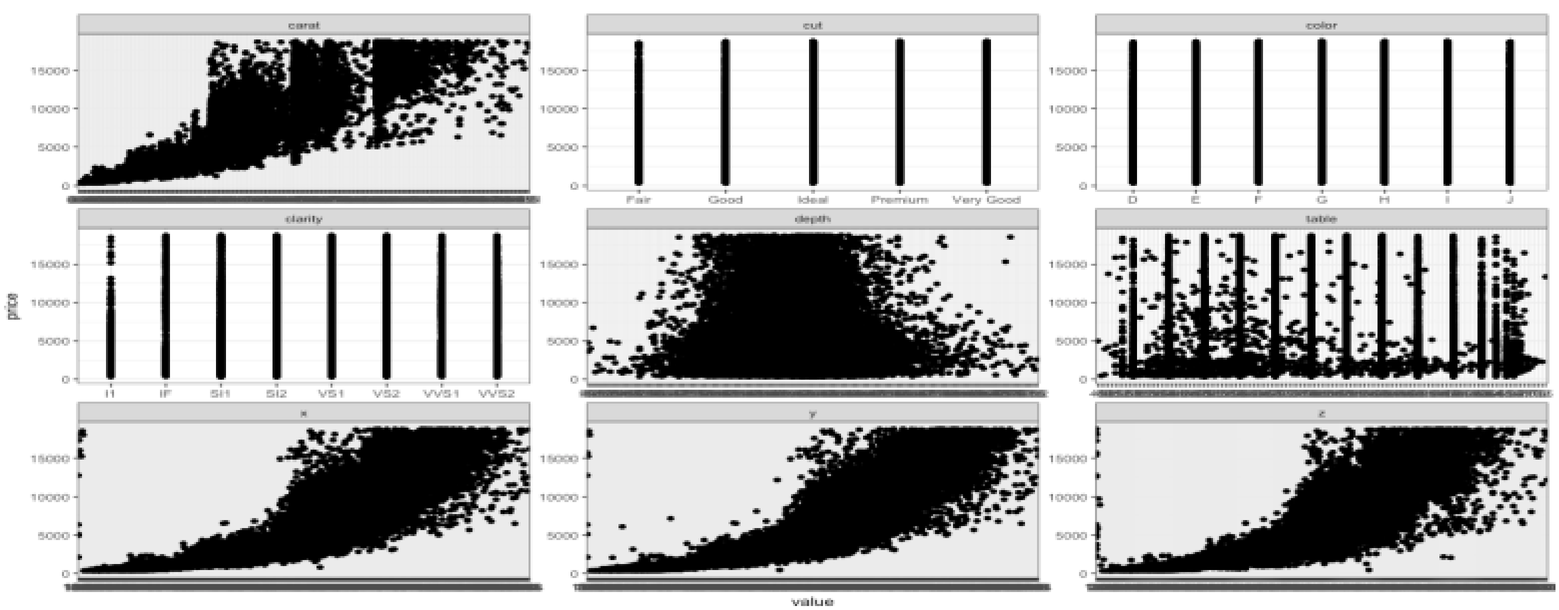
By looking at the above summary, I can comment about each category statistics. Below are few observations for some variables.
Carat: 75% diamonds are below 1.04 carat and above 0.4 carat
Cut: 40%(highest) diamonds are of Ideal cut
Color: 21%(highest) diamonds are of color G which is the average color category
Clarity: 24% diamond are of SI1 clarity which is just better than the worst
Price: Average price of diamonds is 3933 and highest price is 18823

Exploratory Analysis

Count of diamonds by plotting histogram of each category using other categories for categorizing each bin count into levels of that category. Looking at histograms, Ican see highest bar for each category to validate our summary statistics. Also, it's evident that Ideal cut of SI1 clarity and G color are the highest among all combinations followed by premium cut.



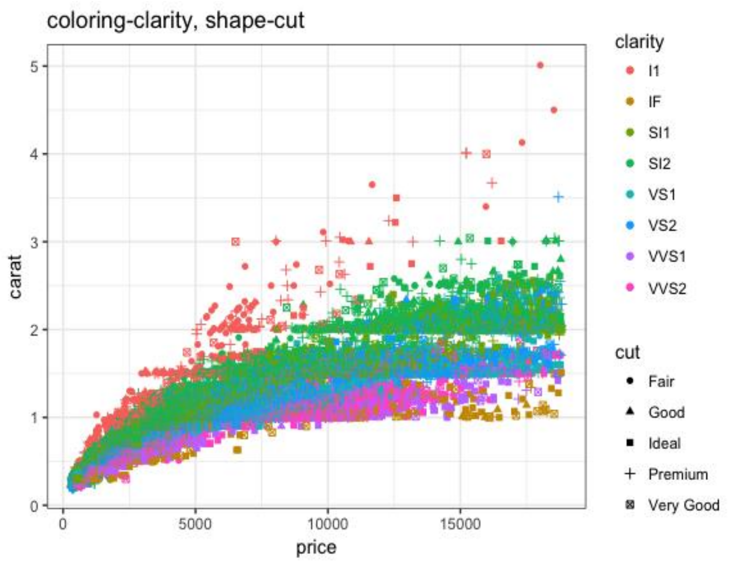
Plotting everything against price for meaning relationships

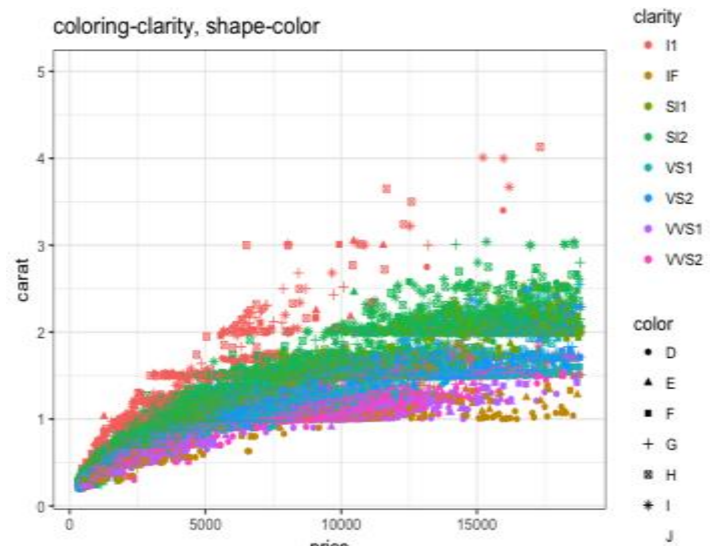


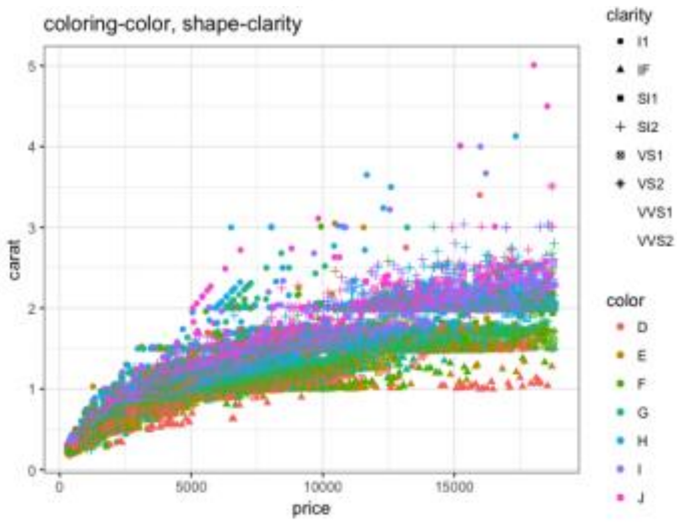
Carat, X, Y, and Z feature seems to have **decent linear relationship** with the price. I will try to explore more on the basis of this information for more meaningful relations. Correlation statistics for each variable against price confirms the above relationships in scatter plots:

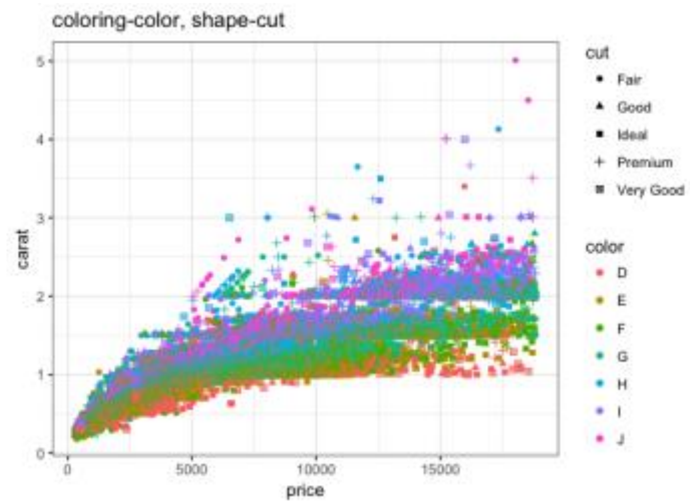
```
           carat             depth             table             price
      0.92159130       -0.01064740        0.12713390        1.00000000
               x                 y                 z   as.numeric.cut.
      0.88443516        0.86542090        0.86124944        0.03986029
 as.numeric.color. as.numeric.clarity.
      0.17251093       -0.07153497
```

Since data have three categorical variables with ordinal levels, their scatter plot with a numerical price doesn't make sense. I will explore them for these 4 highly correlated variables by assigning points shape and color by these categorical variables. Below are most important results:









Looking at plots above, I can infer that diamonds with I1 clarity and J color are the highest weight diamonds for all price ranges and cuts, indicating their inferiority from the other levels within their category. Similarly, the diamonds with IF clarity and D color are among the lowest weight diamonds for all price ranges and cuts, implying their superiority over the others. Thus, even being smaller in size and lighter in weight the IF clarity and D color diamonds are the costliest diamonds. I also looked for interactive graphs with unique information on hovering over the plot, but not functional here so not used in report.

Prices of diamond rises very quickly against carat, but it changes when carat increases
Price of diamond is linearly correlated with x,y,z, and carat

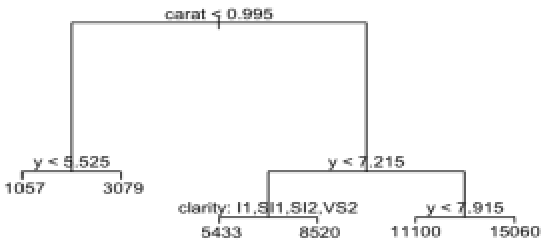
I will now try to build a model to predict the price of a diamond based on these characteristics. I will begin with Linear Model using multivariate linear regression as the observed relationship looks decently linear. Also, I will try decision trees to see which perform better. Dataset: 60% Training and 40% Testing

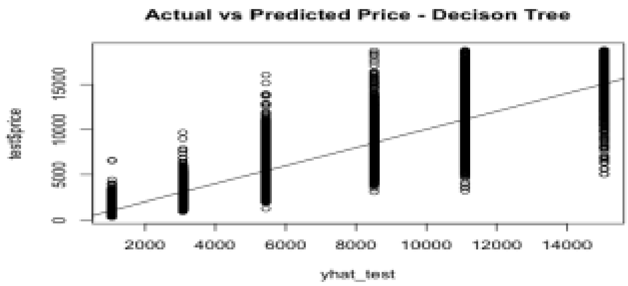
Method	**RMSE TRAIN**	**RMSE TEST**
Decision Tree	1381.633	1398.812
Linear Model (All variables)	1132.92	1126.767

Also, I explored Cross- validation with linear regression to check if RMSE changes and below are the results:

Folds	RMSE	R-Squared	MAE
K=5	1136.669	0.9188313	741.3042
K=10	1130.986	0.9197094	740.5051

The error obtained in Decision Tree is much higher than the linear model. Also, Decision Tree Model identifies carat, y, and clarity as important variables in building the tree. Whereas, according to the output of Linear Model, variable Y is insignificant(t-stats) in terms of predicting the price. And when plotting prices predicted by decision tree against actual the plots implies a linear relation, giving more reasons to explore linear models. Also running the Cross Validation model with 10 folds on entire dataset gives 1129.843 RMSE, which is almost close to the results obtained above.





This leaves me with many possibilities to explore on linear models like using log transformation, using cross-validation, removing and adding variables to decrease the test and training error. I can also find out the most important variables in predicting the price using t-static for the above model and looking at variable importance for cross-validation.

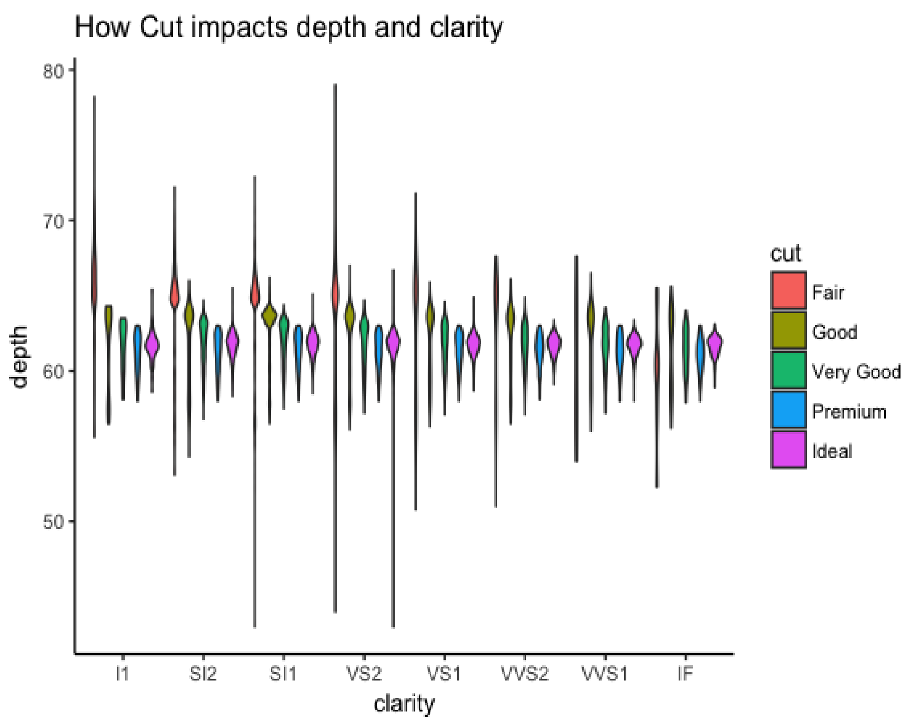
Looking T-stats, I can say that variables x, y, z, depth, table, color are very least significant in price prediction and I should try eliminating them. Building Linear models with remaining variables combinations can help me in better predictions and low test errors.

```
Call:
lm(formula = price ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-21235.1   -594.5   -182.8    379.0  10692.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2528.8645   569.4026   4.441 8.97e-06 ***
carat       11219.7128    61.8341 181.449  < 2e-16 ***
cutGood       626.5383    43.3250  14.461  < 2e-16 ***
cutIdeal      869.4421    43.0648  20.189  < 2e-16 ***
cutPremium    802.4577    41.5676  19.305  < 2e-16 ***
cutVery Good  771.1262    41.5114  18.576  < 2e-16 ***
colorE       -212.5659    23.1323  -9.189  < 2e-16 ***
colorF       -266.3933    23.3801 -11.394  < 2e-16 ***
colorG       -481.4022    22.9551 -20.971  < 2e-16 ***
colorH       -966.1312    24.3567 -39.666  < 2e-16 ***
colorI      -1461.6294    27.4563 -53.235  < 2e-16 ***
colorJ      -2347.5778    33.9903 -69.066  < 2e-16 ***
clarityIF    5347.6100    67.8698  78.792  < 2e-16 ***
claritySI1   3684.9353    58.7765  62.694  < 2e-16 ***
claritySI2   2690.2241    58.9770  45.615  < 2e-16 ***
clarityVS1   4588.3863    59.9318  76.560  < 2e-16 ***
clarityVS2   4278.8059    59.0209  72.496  < 2e-16 ***
clarityVVS1  5010.8772    63.1379  79.364  < 2e-16 ***
clarityVVS2  4991.6489    61.4935  81.174  < 2e-16 ***
depth         -69.5829     6.7974 -10.237  < 2e-16 ***
table         -28.4676     3.7776  -7.536 4.98e-14 ***
x            -955.0195    50.2249 -19.015  < 2e-16 ***
y              -0.7876    20.0409  -0.039    0.969
z             -91.1869    71.9429  -1.267    0.205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1133 on 32340 degrees of freedom
Multiple R-squared:  0.9199,	Adjusted R-squared:  0.9198
F-statistic: 1.614e+04 on 23 and 32340 DF,  p-value: < 2.2e-16
```




```
lm variable importance

  only 20 most important variables shown (out of 23)

            Overall
carat      100.0000
clarity.L   58.4063
color.L     48.5170
clarity.Q   29.3076
color.Q     18.2254
clarity.C   17.3905
```

The variable importance summary for both cross validation models gives carat as the most important variable as expected, which is predicted by decision tree and linear models also. I will explore transformation on carat and selectively chose other important varibles for my next step to build my linear regression model. I will also try taking log transformation for price variable, since the price range is quite high as compared to other variables.