Linear Regression: Diamond Analysis Project

Monica Klosin

11/7/2020

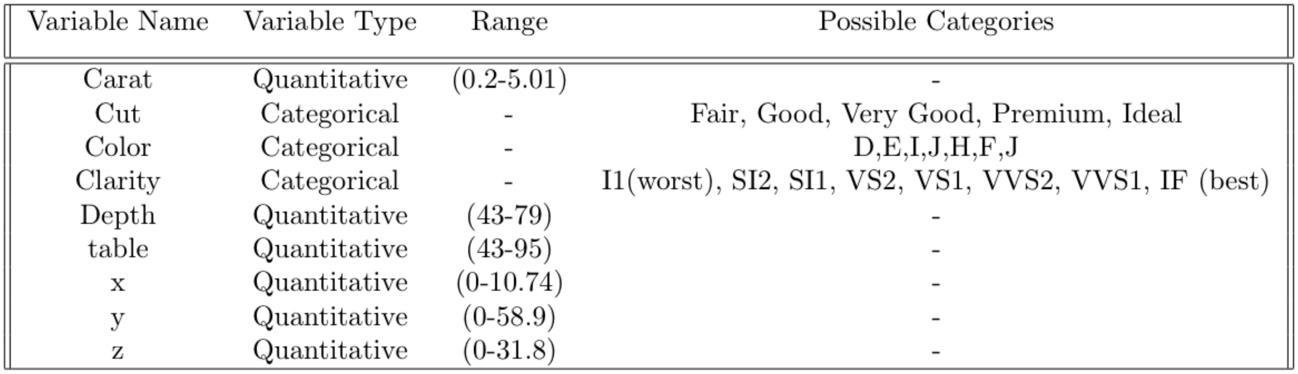
Table of Contents:

Introduction

This dataset contains the prices and other attributes of over 50,000 diamonds.
https://ggplot2.tidyverse.org/reference/diamonds.html

Research Question: what variable(s) influence the price of Diamonds?
Response Variable: Diamond price
Variables I think might be related to diamond price: Before I did any reading, I assumed the price of a diamond would mostly be related to color and carat (size). The bigger the diamond the higher the price, is what makes sense to me. As well as the color of a diamond, since a white diamond is seen as more popular in the media than a blue diamond.
After doing some reading, I found there are 4 main factors to affecting the price of diamonds: cut, clarity, color, and carat. These are called the 4Cs in the diamond community.

http://www.diamondc.com.hk/us/factors-affect-diamond-price
https://www.matthewely.com.au/journal/5-factors-affecting-the-price-of-your-diamonds/



Variable Name	Variable Type	Range	Possible Categories
Carat	Quantitative	(0.2-5.01)	-
Cut	Categorical	-	Fair, Good, Very Good, Premium, Ideal
Color	Categorical	-	D,E,I,J,H,F,J
Clarity	Categorical	-	I1(worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
Depth	Quantitative	(43-79)	-
table	Quantitative	(43-95)	-
x	Quantitative	(0-10.74)	-
y	Quantitative	(0-58.9)	-
z	Quantitative	(0-31.8)	-

Table 1: List all potential predictor variables

Variable Carat is the weight of a Diamond. Cut is a variable that indicates the quality of the cut, from a fair cut to an ideal cut (range of variable types shown in Table 1). Color is a variable indicating the color of a diamond, from D (best) to J (worst), range of variable types shown in Table 1. Variable Clarity is measures how clear the diamond is, from I1 to IF, Table 1 above shows the range. Variable table is a measurement of the width of top of a diamond relative to its widest point. Variable x,y, and z are all dimensions for the diamonds. x is the length of a diamond in mm, y is the width of a diamond in mm, and z is the depth of a diamond in mm. variable "depth" is the total depth percentage of a diamond.

This data does adhere to the rules of having more than 50 observations per predictor, the response variables is quantitative, there are 6-10 predictor variables, 3 of them are quantitative and one is categorical.

2A: Modeling – Exploratory Data Analysis

To see the distribution of our variables, we want to preform Exploratory Data Analysis on all the variables.

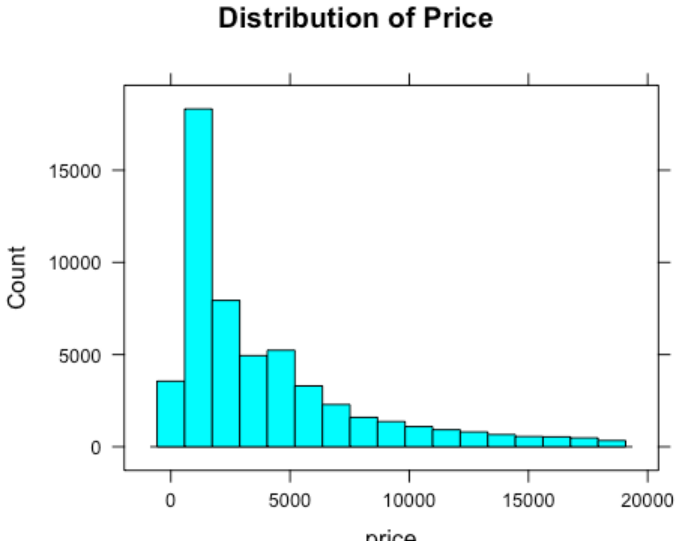
Univariate EDA:

For each Quantitative Variable var, to see the distribution in R I compiled:

```
histogram(~var, data = diamond, type="count")
```

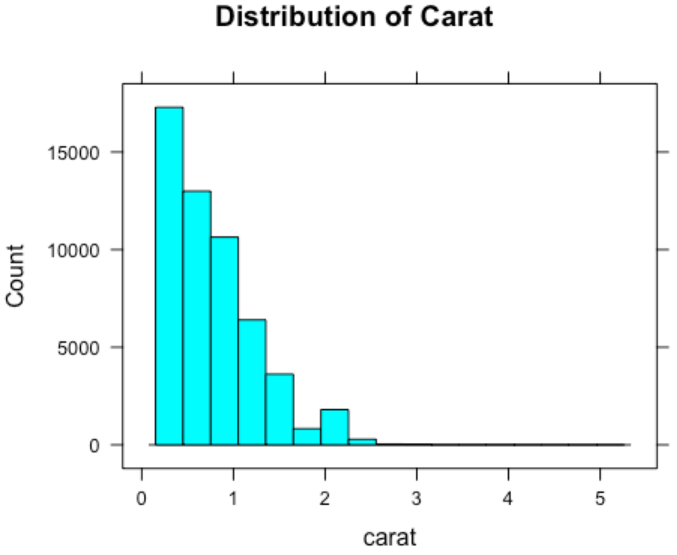
For each Categorical Variable var, to see the distribution in R I compiled:

```
tab1(diamond$var)
```



For distribution of Price:

The distribution of Price shows a uni-model and bell-shaped right skewed curve with a peak near $1000. This means that there are few outliers with a few diamonds who have a much higher price than the majority of the diamonds.

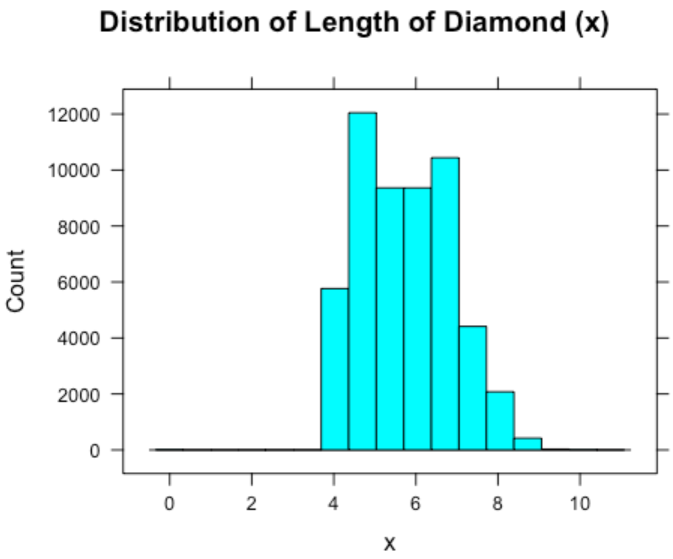


For distribution of Carat:

The distribution of Carat, or the weight of the diamonds, is uni-model and is right skewed. There are a few outliers, with some diamonds weighting more than 3 carats, causing the data to be skewed right.

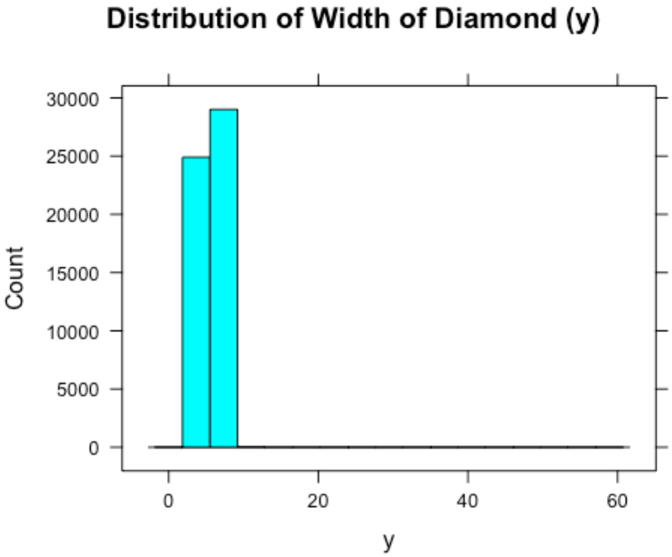
For distribution of x:

The distribution of x (length of a diamond) is uni model bell shape with no obvious skew. There are some outliers with very small lengths < 1 and some diamonds with very large lengths < 10.



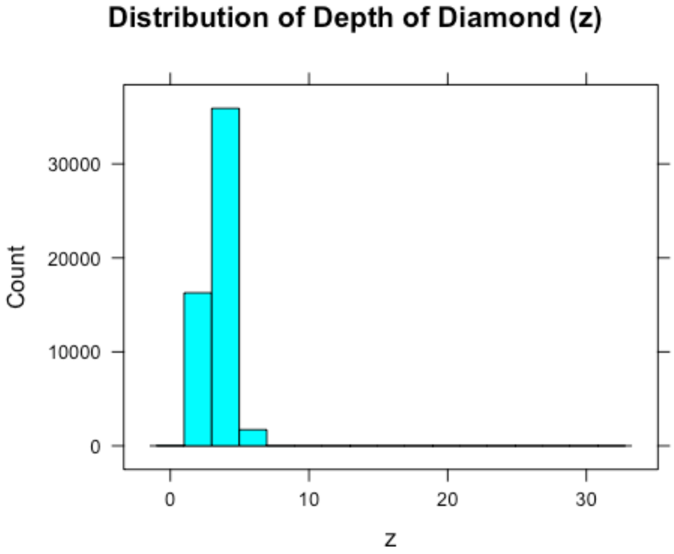
For distribution of y:

The distribution of y (width of a diamond) is unimodal with right-skew. While it is difficult to see on the graph to the right, there are a few outliers where the width of the diamond is < 40mm, while the majority widths of the diamonds are between 0-10mm.



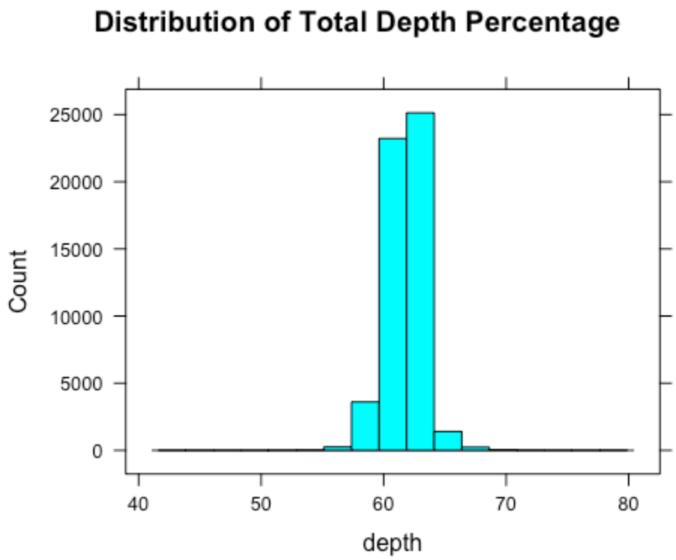
For distribution of z:

The distribution of z (depth of a diamond) is unimodal with right-skew. While it is difficult to see on the graph to the right, there are a few outliers where the width of the diamond is < 10mm, while the majority widths of the diamonds are between 0-7mm.



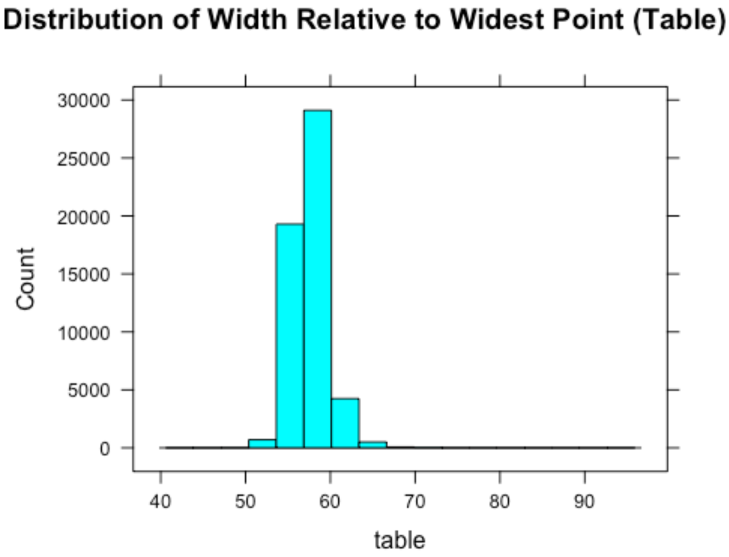
For distribution of depth:

The distribution of depth is unimodal with no obvious skew. The majority of the data is between 60-65%, and there are a few outliers where the depth % for a diamond is < 50% or > 70%.



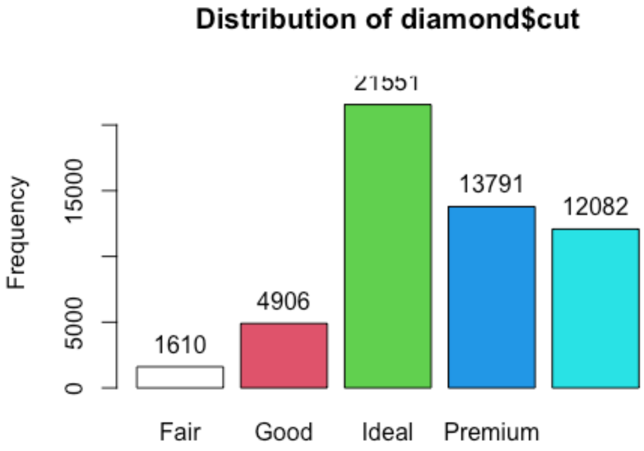
For distribution of Table:

The distribution of Table is unimodal with no obvious skew. The majority of the data is between 55-60%, and there are a few outliers where the table % for a diamond is < 50% or > 70%.



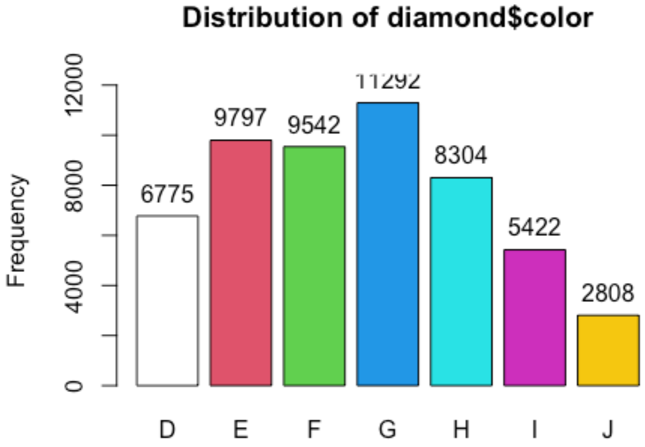
For distribution of Cut:

The majority of the diamonds have an Ideal Cut.



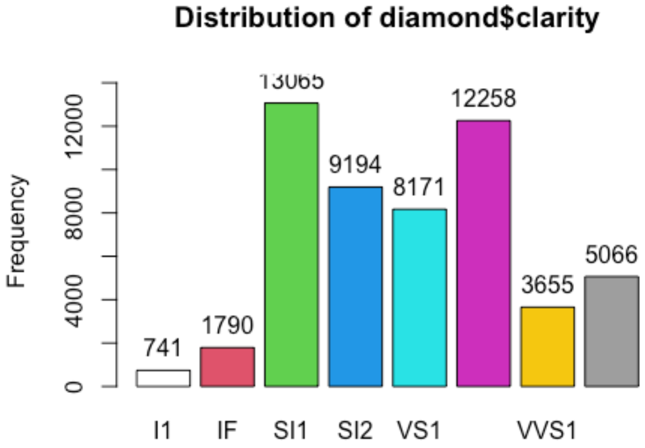
For distribution of Color:

There is a pretty even distribution of color in the dataset, with J having the fewest diamonds.



For distribution of Clarity:

There is a pretty even distribution of clarity in the dataset, with I1 having the lowest among of data points. diamonds.



Bivariate EDA:

For each Quantitative Variable var (since categorical variables would show limited information), to see the relationship between the variable and price, we perform a Bivariate EDA. To graph in R the relationship between Price and variable var:

```
diamond %>% ggplot(aes(var, LogPrice)) + geom_point(aes(color=var)) +
coord_cartesian(ylim = c(0, {limit})) + geom_smooth(method='lm', formula=
y~x, color="red", size=.5))
```

To find the correlation coefficents in R:

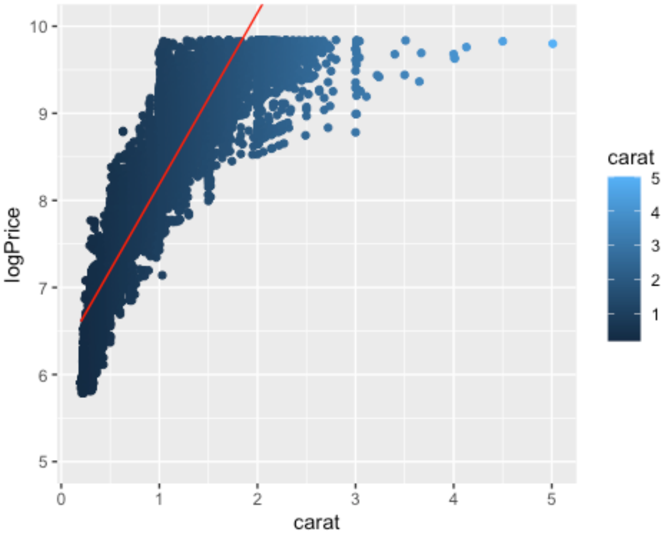
```
cor(var~logPrice, data = diamond)
```

When I plotted price against the covariants, I noticed the relationships did not seem linear, so I changed the analysis to be LogPrice plotted against the covariants.

The gradient on the right of all the graphs indicates the value of the variable. I did this to more clearly see the many dots on the graph.

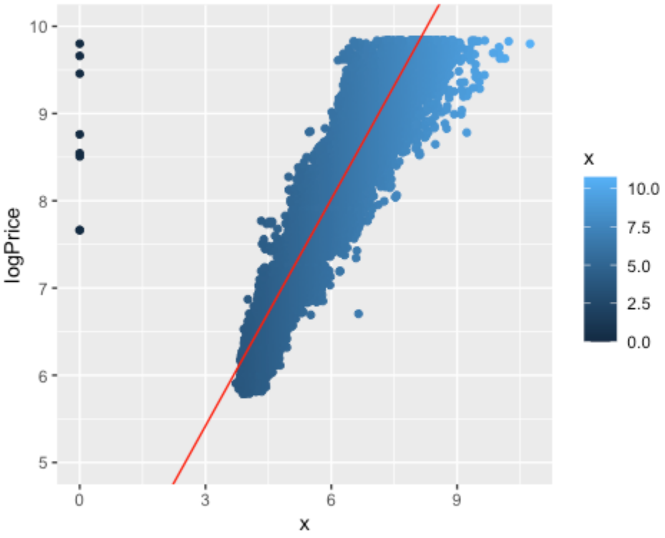
For carat and LogPrice:

The correlation coefficient for carat and LogPrice is 0.920. Based on the graph and the correlation coefficents, we see that there is a strong positive linear correlation between carat and LogPrice. There are a few outliers where carat has larger values and its LogPrice is 10, which is smaller than the model predicts it should be.



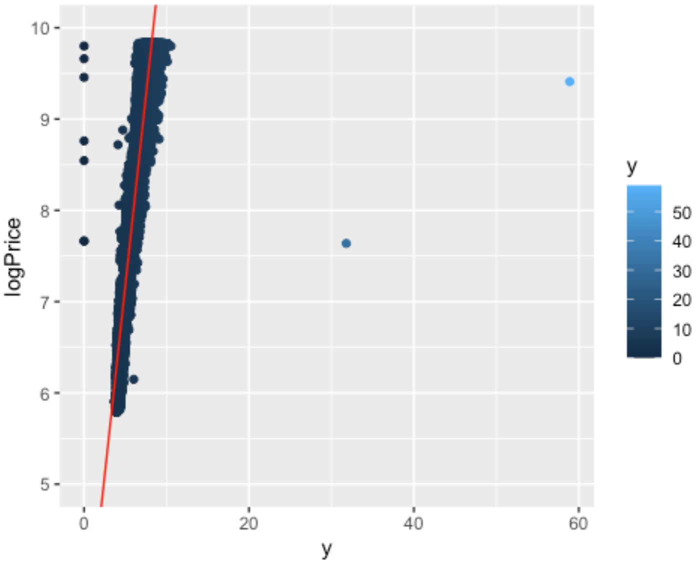
For x and LogPrice:

The correlation coefficient for x (the length) and LogPrice is 0.958. Based on the graph and the correlation coefficents , we see a strong positive linear correlation between x and LogPrice. There are a few outliers where x is much smaller than average with a higher price than the model predicts it should be.



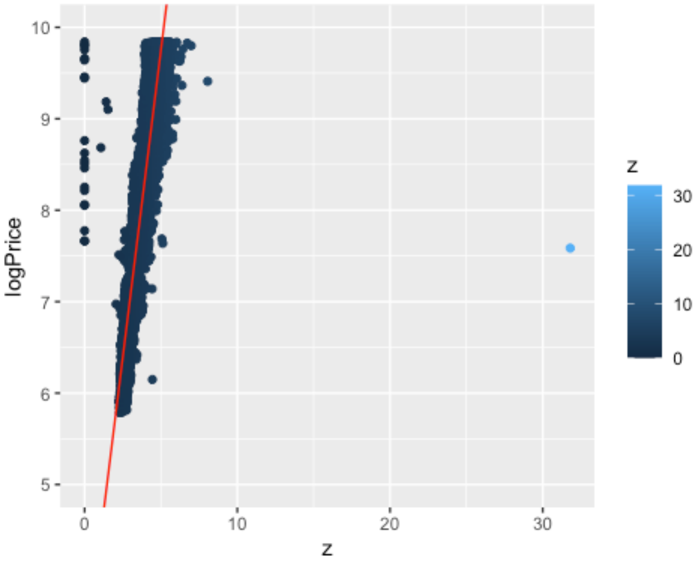
For y and LogPrice:

The correlation coefficents for y (the width) and LogPrice is 0.936. Based on the graph and the correlation coefficents , we see a strong positive linear y and LogPrice. There are a few outliers where y is larger than average with a lower price than the model predicts it should be.



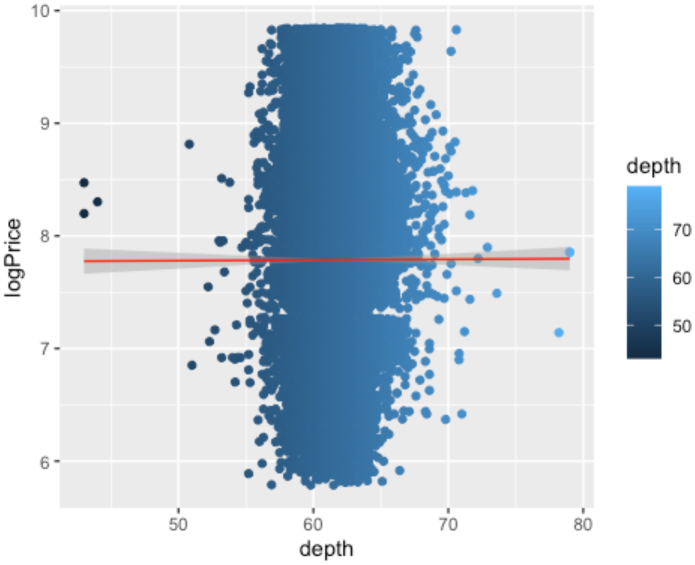
For z and LogPrice:

The correlation coefficents for z (the depth) and LogPrice is 0.935. Based on the graph and the correlation coefficient, we see a strong positive linear correlation between z and LogPrice. There are very few outliers where z is larger than average with a lower price than the model predicts it should be.



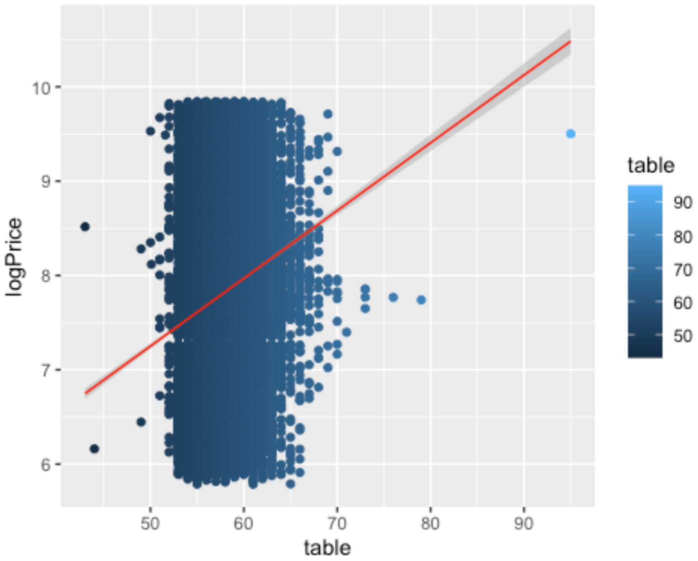
For depth and LogPrice:

The correlation coefficient for the total depth percentage and LogPrice is 0.0008. Based on the graph, we see little to no linear correlation between the depth and LogPrice.



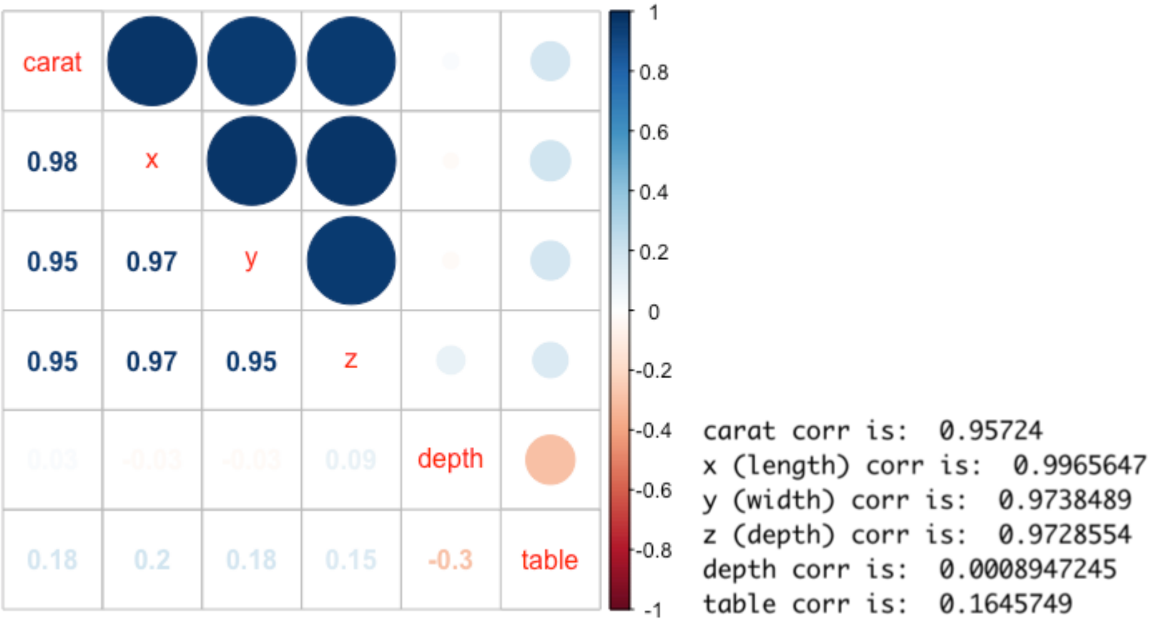
For table and LogPrice

The correlation coefficents for table and LogPrice is 0.158. Based on the graph, we see little to no linear correlation between table and LogPrice.



Collinearity Checks Among Predictors:

I used just the quantitative variables since the categorical would have limited use.



```
carat corr is:  0.95724
x (length) corr is:  0.9965647
y (width) corr is:  0.9738489
z (depth) corr is:  0.9728554
depth corr is:  0.0008947245
table corr is:  0.1645749
```

The R output shows that there is a positive strong correlation with carat and all the dimensions of a diamond (x,y, and z are the width, height, and depth of a diamond). depth and table have low correlation other variables in the model.

Determination of whether you need to transform any variables, especially the response:

Due to the Bi-variate EDA between price and the predictor variables not having a mostly linear relationship, I would transform the response, price, by logging it. I did this already above in the Bi-varatite analysis.

Since some of the variables are categorical, I will need to create dummy variables in their place. I do this in section 2B below.

2B: Modeling - Full Model Specification

Specification of Dummy Variables and the Full Model:

For this section, I have renamed the variables to have simple notation.

logY = price
X1 = carat
X2 = cut
X3 = color

X4 = clarity
X5 = depth
X6 = table
X7 = x

X8 = y
X9 = z
D_{2_i} =dummy for cut type i

D_{3_j} =dummy for color type j
D_{4_k} =dummy for clarity type k

Before we can make the full model, we must create dummy variables for the categorical variables cut, color, and clarity. The first dummy has to be removed for collinearity problems. Because of this, color_D and clarity_I1 are dropped and do not have corresponding dummy variables.
To create the dummy variables, in R:

```
dummy_cols(diamond, remove_selected_columns = TRUE, remove_first_dummy = TRUE)
```

Interactions

I think it would be really interesting to see if the best clarity and best cut together have an impact on diamond price. I created two interaction variables, one showing the interaction between the clearest diamond with the ideal cut, and one showing the 2nd worst cut and 2nd worst clarity.

```
Int.B = clarity_IF*cut_Ideal
Int.W = clarity_SI2*cut_Good
```

The full model is:

$$logY = \beta_0 + \beta_1 X_1 + \sum_i \beta_{2i} D_{2_i} + \beta_2 D_{2_5} + \sum_j \beta_{3_j} D_{3_j} + \sum_k \beta_{4_k} D_{4_k} + \beta_5 X_5 + \beta_6 X_6$$
$$+ \beta_7 X_7 + \beta_8 X_8 + \beta_9 Int.B + \beta_{10} Int.W + \epsilon$$

description of notation used above:

$\sum_i \beta_{2i} D_{2_i}$ = the sum of all the dummy variables created for cut D$_2$

Model building strategy

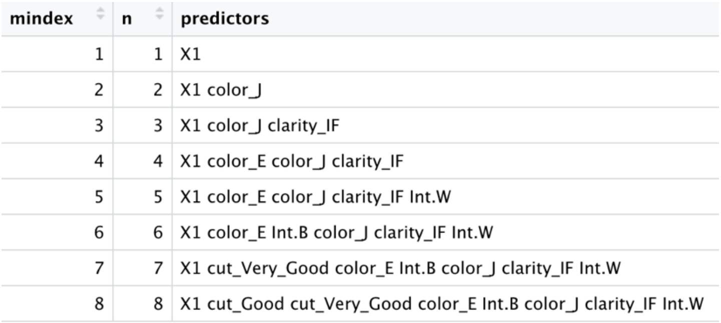
Write out how you are going to determine your candidate final model.

When I tried to use the ols_step_best_subfit to summarize the best fit for the current full model, R was running the code for hours. I instead ran a smaller version of the full model, to see of those variables, which would be the best subset. I ran this to also show that I know how to run and interpret ols_step_best_subfit.

Small Model is:

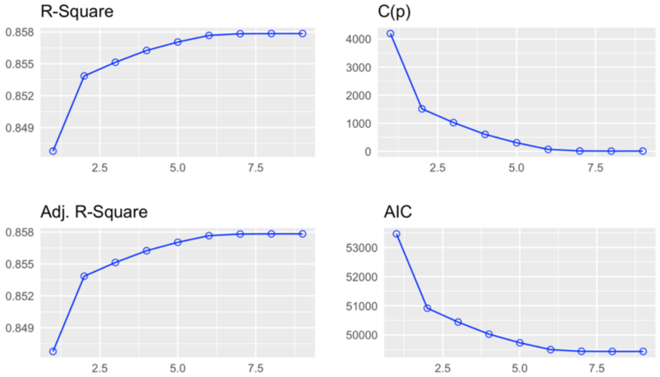
```
FMsmall =lm(logPrice~X1+cut_Good+cut_Very_Good+color_E+Int.B+color_J+
clarity_IF+ Int.W, data=dummyDiamond)
```

In R we get this table:



mindex	n	predictors
1	1	X1
2	2	X1 color_J
3	3	X1 color_J clarity_IF
4	4	X1 color_E color_J clarity_IF
5	5	X1 color_E color_J clarity_IF Int.W
6	6	X1 color_E Int.B color_J clarity_IF Int.W
7	7	X1 cut_Very_Good color_E Int.B color_J clarity_IF Int.W
8	8	X1 cut_Good cut_Very_Good color_E Int.B color_J clarity_IF Int.W

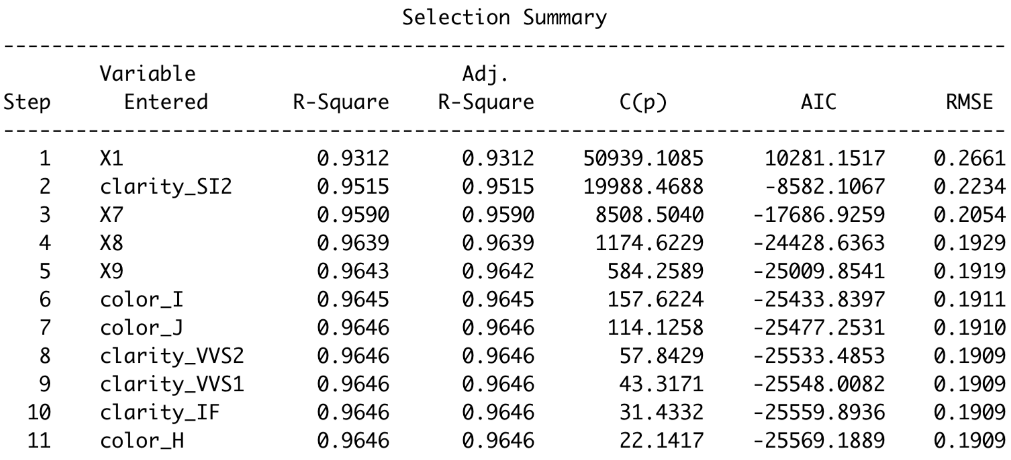
In R we get plot:



We want to add predictors up to the point where additional predictors make little difference, we can see in the graphs above that the plot starts flatlining around 6. which means we should use the variables shown in model index 6, which is X1, color_E, Int.B, color_J, clarity_IF, and Int.W.

However, since this is a smaller subset of the full model and not the full model, we do not want to use that for our analysis. By using the variable selection procedure, we can find which variables in the full model would be best to use. We use the R code below to find this:

```
ols_step_forward_p(FM, penter=0.05, details = TRUE)
```



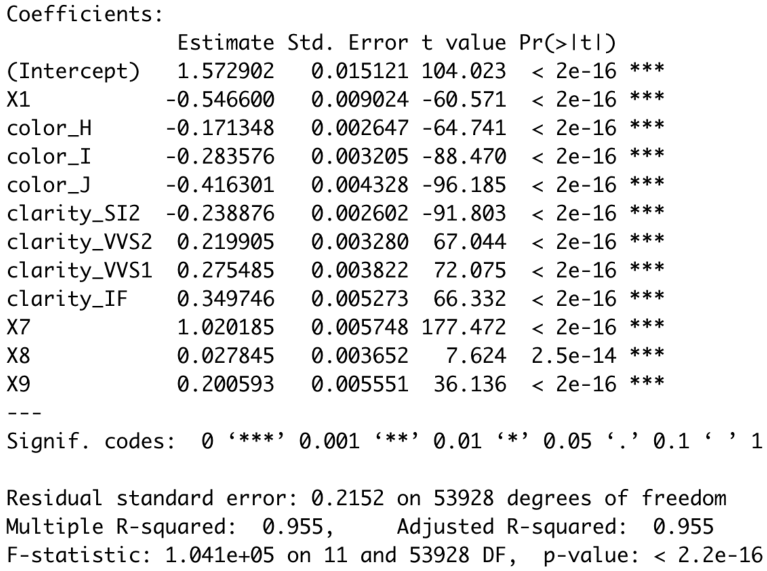
```
                              Selection Summary
-----------------------------------------------------------------------------------
       Variable                     Adj.
Step     Entered        R-Square    R-Square       C(p)            AIC         RMSE
-----------------------------------------------------------------------------------
   1    X1                0.9312      0.9312    50939.1085     10281.1517    0.2661
   2    clarity_SI2       0.9515      0.9515    19988.4688     -8582.1067    0.2234
   3    X7                0.9590      0.9590     8508.5040    -17686.9259    0.2054
   4    X8                0.9639      0.9639     1174.6229    -24428.6363    0.1929
   5    X9                0.9643      0.9642      584.2589    -25009.8541    0.1919
   6    color_I           0.9645      0.9645      157.6224    -25433.8397    0.1911
   7    color_J           0.9646      0.9646      114.1258    -25477.2531    0.1910
   8    clarity_VVS2      0.9646      0.9646       57.8429    -25533.4853    0.1909
   9    clarity_VVS1      0.9646      0.9646       43.3171    -25548.0082    0.1909
  10    clarity_IF        0.9646      0.9646       31.4332    -25559.8936    0.1909
  11    color_H           0.9646      0.9646       22.1417    -25569.1889    0.1909
```

The current full model is:

note: we restart the ordering of our β coefficients.

$$logY = \beta_0 + \beta_1 X_1 + \beta_2 D_{4_1} + \beta_3 D_{4_5} + \beta_4 D_{4_6} + \beta_5 D_{4_8} + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} D_{3_5} + \beta_{11} D_{3_6} + \beta_{12} D_{3_7} + \epsilon$$

For the overall F-Test and the individual T-tests, we will use the output from R below:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.572902   0.015121 104.023  < 2e-16 ***
X1           -0.546600   0.009024 -60.571  < 2e-16 ***
color_H      -0.171348   0.002647 -64.741  < 2e-16 ***
color_I      -0.283576   0.003205 -88.470  < 2e-16 ***
color_J      -0.416301   0.004328 -96.185  < 2e-16 ***
clarity_SI2  -0.238876   0.002602 -91.803  < 2e-16 ***
clarity_VVS2  0.219905   0.003280  67.044  < 2e-16 ***
clarity_VVS1  0.275485   0.003822  72.075  < 2e-16 ***
clarity_IF    0.349746   0.005273  66.332  < 2e-16 ***
X7            1.020185   0.005748 177.472  < 2e-16 ***
X8            0.027845   0.003652   7.624  2.5e-14 ***
X9            0.200593   0.005551  36.136  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2152 on 53928 degrees of freedom
Multiple R-squared:  0.955,	Adjusted R-squared:  0.955 
F-statistic: 1.041e+05 on 11 and 53928 DF,  p-value: < 2.2e-16
```

Overall F-Test

H$_0$: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$
H$_a$: At least one B$_x$!= 0

F-statistic = 1.041e+05 ; df = 11 ; p-value = 2.2e-16
decision: We reject H$_o$
conclusion: At least one predictor variable is useful in predicting diamond price.

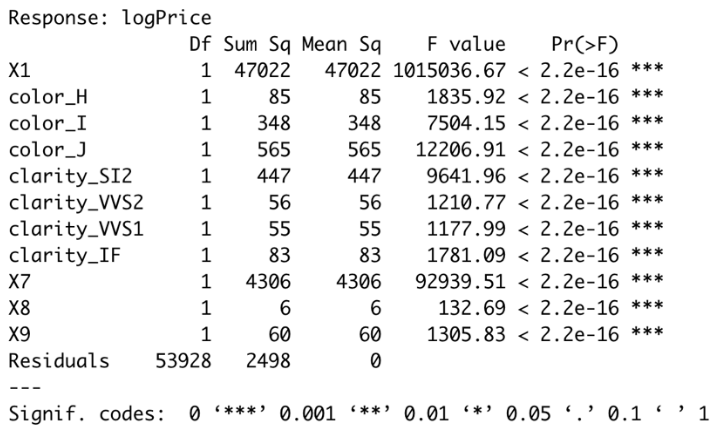
! Note: for the rest of the paper I notate the β and X with the name of the value being used, I will write X_{color_H} instead of X_{35} since it is easier to understand the analysis with names and not sub numbers.

Multiple Partial F-Test

If we don't include the color dummys, will this affect the variables/model?

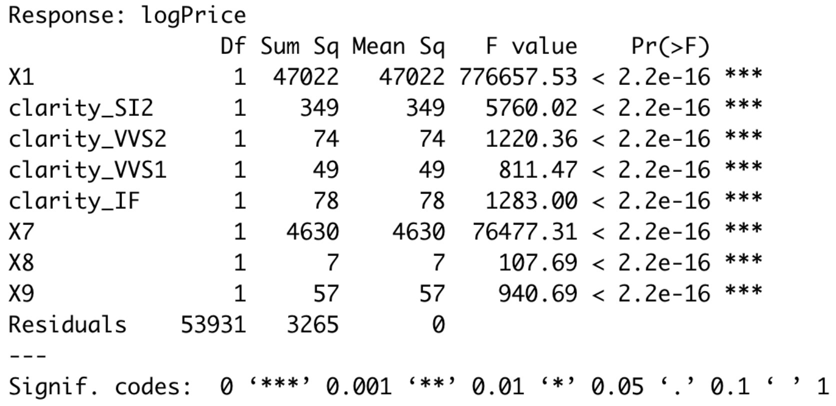
We will do a Partial F-Test for color_H, color_I, color_J in a model with all the other variables

Full Model ANOVA:



```
Response: logPrice
                 Df Sum Sq Mean Sq    F value    Pr(>F)    
X1                1  47022   47022 1015036.67 < 2.2e-16 ***
color_H           1     85      85    1835.92 < 2.2e-16 ***
color_I           1    348     348    7504.15 < 2.2e-16 ***
color_J           1    565     565   12206.91 < 2.2e-16 ***
clarity_SI2       1    447     447    9641.96 < 2.2e-16 ***
clarity_VVS2      1     56      56    1210.77 < 2.2e-16 ***
clarity_VVS1      1     55      55    1177.99 < 2.2e-16 ***
clarity_IF        1     83      83    1781.09 < 2.2e-16 ***
X7                1   4306    4306   92939.51 < 2.2e-16 ***
X8                1      6       6     132.69 < 2.2e-16 ***
X9                1     60      60    1305.83 < 2.2e-16 ***
Residuals     53928   2498       0                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Reduced Model ANOVA:



```
Response: logPrice
                Df Sum Sq Mean Sq   F value    Pr(>F)    
X1               1  47022   47022 776657.53 < 2.2e-16 ***
clarity_SI2      1    349     349   5760.02 < 2.2e-16 ***
clarity_VVS2     1     74      74   1220.36 < 2.2e-16 ***
clarity_VVS1     1     49      49    811.47 < 2.2e-16 ***
clarity_IF       1     78      78   1283.00 < 2.2e-16 ***
X7               1   4630    4630  76477.31 < 2.2e-16 ***
X8               1      7       7    107.69 < 2.2e-16 ***
X9               1     57      57    940.69 < 2.2e-16 ***
Residuals    53931   3265       0                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSE(FM) = 2498
SSE(RM) = 3265
p = 12 (p is the number of predictors in full model)
n = 53,940 (population)
k = 9 (number of predictors in reduced model + 1)

H_0: $\beta_{color\_H} = \beta_{color\_I} = \beta_{color\_J} = 0$ given X_1, $X_{clarity\_SI2}$, $X_{clarity\_VVS2}$, $X_{clarity\_VVS1}$, $X_{clarity\_IF}$, X_7, X_8, X_9 in model

H_a: β_1, $\beta_{clarity\_SI2}$, $\beta_{clarity\_VVS2}$, $\beta_{clarity\_VVS1}$, $\beta_{clarity\_IF}$, β_7, β_8, β_9 != 0 | X_{color_H},X_{color_I}, X_{color_J} in model

F-statistic = 4139.513 ; **df** = 3

$$F = \frac{\left[SSE(RM) - SSE(FM)\right]/(p+1-k)}{SSE(FM)/(n-p-1)}$$ (((3265-2498))/(12+1-9))/(2498/(53940-12-1)) = 4139.513

p-value: <2.2e-16

decision: Reject H_0

conclusion: At least one of the predictor variables, X_{color_H},$X_{color\_I,}$ X_{color_J} adds useful information to the prediction of diamond price in the model that includes the other listed predictors. The full model is better than the reduced model for predicting diamond price

Partial F-Tests

Separate individual t-tests for all variables in the final model:

For each variable var, the null and alternative hypothesis will be:

H_0: $\beta_{var} = 0$ given the X's in the model
H_a: $\beta_{var}! = 0$ given the X's in the model

Cut (X1):
t = -60.571; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, color_H is a statistically significant predictor for diamond price.

color_H:
t = -64.571; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, color_H is a statistically significant predictor for diamond price.

color_I:
t = -88.470; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, color_I is a statistically significant predictor for diamond price.

color_J:
t = -96.185; df: 53928; p = <2e-16
decision: reject H0

conclusion: after adjusting for the other predictor variables in the model, color_J is a statistically significant predictor for diamond price.

clarity_SI2:
t = -91.803; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, clarity_SI2 is a statistically significant predictor for diamond price.

clarity_VVS2:
t = 67.044; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, clarity_VVS2 is a statistically significant predictor for diamond price.

clarity_VVS1:
t = 72.075; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, clarity_VVS1 is a statistically significant predictor for diamond price.

clarity_IF:
t = 66.332; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, clarity_IF is a statistically significant predictor for diamond price.

X7:
t = 177.472; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, X7 is a statistically significant predictor for diamond price.

X8:
t = 7.624; df: 53928; p = 2.5e-14
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, X8 is a statistically significant predictor for diamond price.

X9:
t = 36.136; df: 53928; p = <2e-16
decision: reject H0
conclusion: after adjusting for the other predictor variables in the model, X9 is a statistically significant predictor for diamond price.

2C: Modeling - Model Fitting and Regression Diagnostics

Ordinary least-squares regression

Currently, the fitted model is:

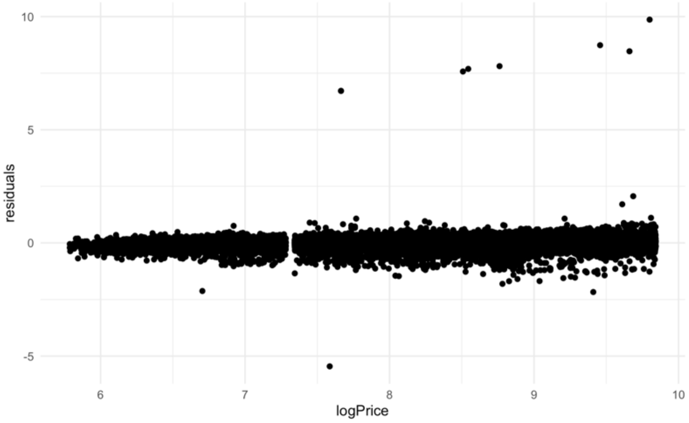
$$log\hat{Y} = 1.5729 – 0.5466X_1 – 0.1713X_{color\_H} – 0.2835X_{color\_I} – 0.4163X_{color\_J} –0.2388X_{clarity\_SI2} + 0.2199X_{clarity\_VVS2} +0.2754X_{clarity\_VVS1} + 0.3497X_{clarity\_IF} + 1.02X_7 + 0.0278X_8 + 0.2X_9$$

Hypothesis Test

Done in 2B.

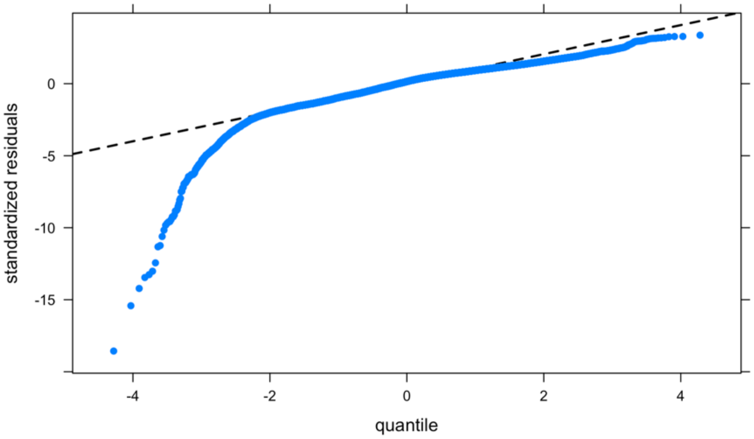
Check Conditions

residual plots



The assumption that the errors have constant variance is met because the residuals are evenly distributed across different values of LogPrice, as seen in the plot above.

Normal Probability Plot



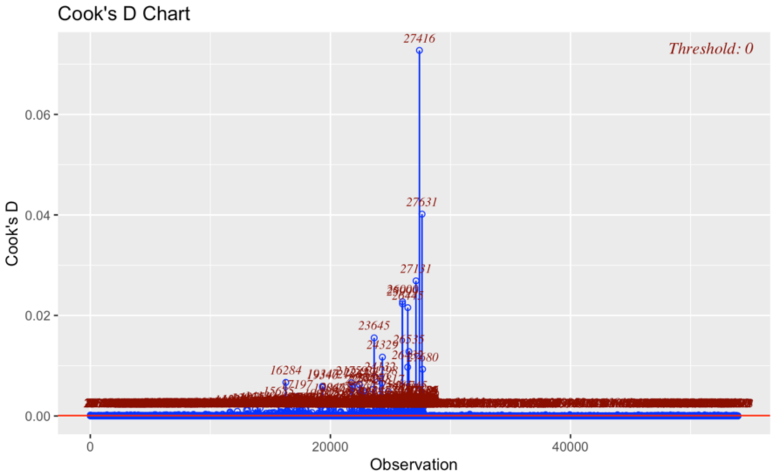
The closer the blue line is to the dotted black line the more our residuals are fitting a normal distribution. Assumption 2 checks Normality, seeing if the errors follow a normal distribution. Based on the plot, Assumption 2 Normality is not met because the residuals are heavy tailed. The blue line is pretty far from the dotted black line at very low (< -2) quantiles. This means residuals from our model at low quantiles are not fitting the normal distribution well.

Influential Observations (cooks D)

When we run Cooks D, we see one observation with a Cooks D value greater than 0.05.

When we inspect this value, we find it to be a Diamond with a price of $18018, the most expensive diamond in the data set – it is double the price of the 2nd most expensive diamond (which costs $9424).

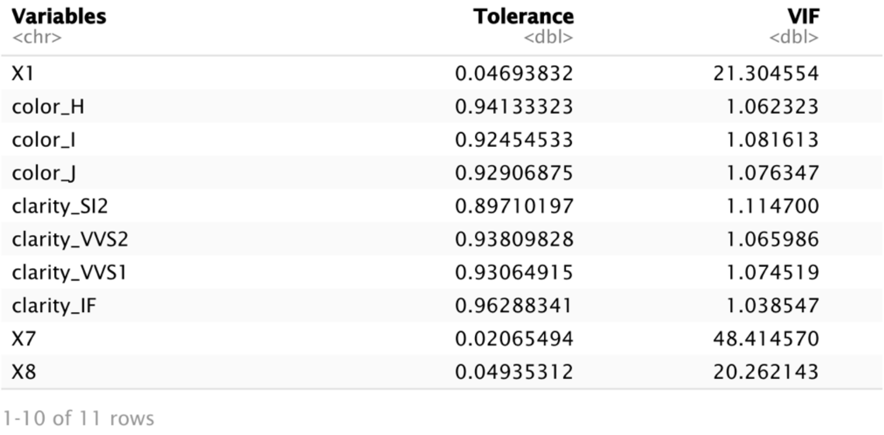
This diamond is also incredibly big, with a carat weight of 5.01, which is approx. 5 times bigger than the second most heavy diamond (whose weight is 1.20 carats).



Variance Inflation Factors

In R when we run VIF on the current model with the code below:

```
ols_vif_tol(ModelFinal2)
```



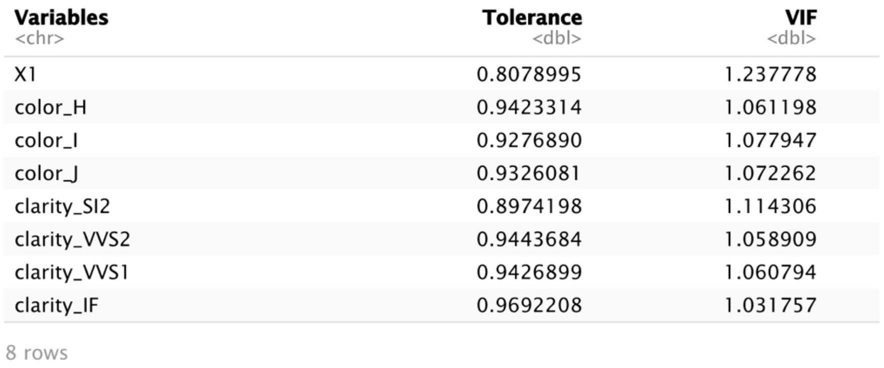
Variables <chr>	Tolerance <dbl>	VIF <dbl>
X1	0.04693832	21.304554
color_H	0.94133323	1.062323
color_I	0.92454533	1.081613
color_J	0.92906875	1.076347
clarity_SI2	0.89710197	1.114700
clarity_VVS2	0.93809828	1.065986
clarity_VVS1	0.93064915	1.074519
clarity_IF	0.96288341	1.038547
X7	0.02065494	48.414570
X8	0.04935312	20.262143

1-10 of 11 rows

We see in VIF analysis that X1, which is the size of the diamond, and the variables describing the dimensions of the diamonds (x (length) ,y (width), and z (depth)) have very high VIF values. This suggests multicollinearity.

It makes sense that there is multi-collinearity because the size of the diamond (X1) it going to be highly related to its dimensions like x, y, and z .

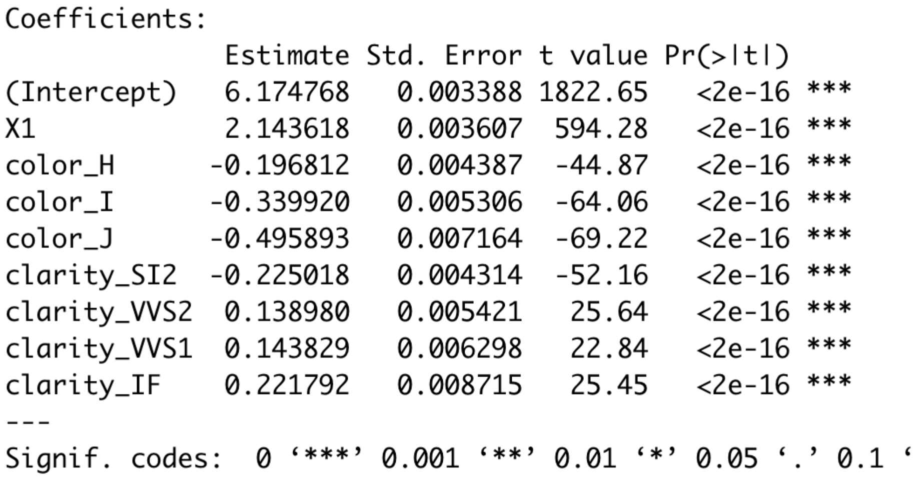
By dropping x, y, and z, the VIF analysis shows that no variables > 10, which means we no longer have any co-linearity issues.



Variables <chr>	Tolerance <dbl>	VIF <dbl>
X1	0.8078995	1.237778
color_H	0.9423314	1.061198
color_I	0.9276890	1.077947
color_J	0.9326081	1.072262
clarity_SI2	0.8974198	1.114306
clarity_VVS2	0.9443684	1.058909
clarity_VVS1	0.9426899	1.060794
clarity_IF	0.9692208	1.031757

8 rows

When we drop x,y,z and run a new regression, the coefficents of variables values stay relatively the same except X1, which from -0.5466 goes to 2.14. This large change makes sense since there is more emphasis on carat, and carat is now having a larger impact on price since there are fewer variables in the model.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.174768   0.003388 1822.65   <2e-16 ***
X1            2.143618   0.003607  594.28   <2e-16 ***
color_H      -0.196812   0.004387  -44.87   <2e-16 ***
color_I      -0.339920   0.005306  -64.06   <2e-16 ***
color_J      -0.495893   0.007164  -69.22   <2e-16 ***
clarity_SI2  -0.225018   0.004314  -52.16   <2e-16 ***
clarity_VVS2  0.138980   0.005421   25.64   <2e-16 ***
clarity_VVS1  0.143829   0.006298   22.84   <2e-16 ***
clarity_IF    0.221792   0.008715   25.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
```

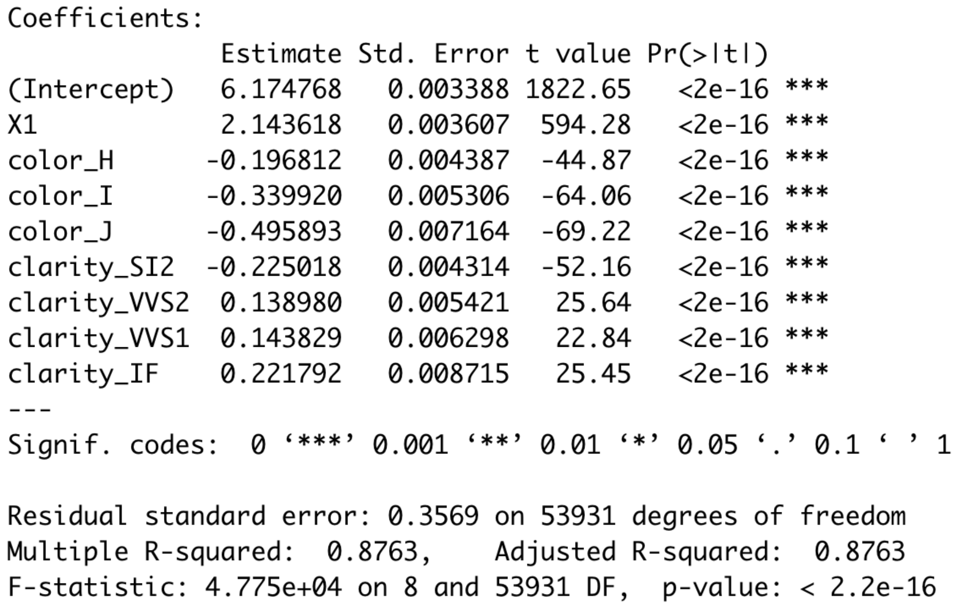
With the diagnostics done above, the final model is:

$logY = \beta_0 + \beta_1 X_1 - \beta_{color\_H} X_{color\_H} - \beta_{color\_I} X_{color\_I} - \beta_{color\_J} X_{color\_J} - \beta_{clarity\_SI2} X_{clarity\_SI2} + \beta_{clarity\_VVS2} X_{clarity\_VVS2} + \beta_{clarity\_VVS1} X_{clarity\_VVS1} + \beta_{clarity\_IF} X_{clarity\_IF}$

$log\hat{Y} = 6.174768 + 2.1436X_1 - 0.1968X_{color\_H} - 0.3399X_{color\_I} - 0.4958X_{color\_J} - 0.2250X_{clarity\_SI2} + 0.1389X_{clarity\_VVS2} + 0.1438X_{clarity\_VVS1} + 0.2217X_{clarity\_IF}$

3: Model Usage

Interpret Regression Coefficients



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.174768   0.003388 1822.65   <2e-16 ***
X1            2.143618   0.003607  594.28   <2e-16 ***
color_H      -0.196812   0.004387  -44.87   <2e-16 ***
color_I      -0.339920   0.005306  -64.06   <2e-16 ***
color_J      -0.495893   0.007164  -69.22   <2e-16 ***
clarity_SI2  -0.225018   0.004314  -52.16   <2e-16 ***
clarity_VVS2  0.138980   0.005421   25.64   <2e-16 ***
clarity_VVS1  0.143829   0.006298   22.84   <2e-16 ***
clarity_IF    0.221792   0.008715   25.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3569 on 53931 degrees of freedom
Multiple R-squared:  0.8763,	Adjusted R-squared:  0.8763 
F-statistic: 4.775e+04 on 8 and 53931 DF,  p-value: < 2.2e-16
```

X1:

If you increase the carat size by 1, the diamond LogPrice will increase by $2.14.

color_H:

In relation to the baselines color (for us D) the LogPrice of a H colored diamond is smaller by $0.19.

color_I:

In relation to the baselines color (for us D) the LogPrice of a I colored diamond is smaller by $0.33.

color_J:

In relation to the baselines color (for us D) the LogPrice of a J colored diamond is smaller by $0.49.

clarity_SI2:

In relation to the baselines color (for us I1) the LogPrice of a clarity SI2 level diamond is smaller by $0.22.

clarity_VVS2:

In relation to the baselines color (for us I1) the LogPrice of a clarity VVS2 level diamond is larger by $0.13.

clarity_VVS1:

In relation to the baselines color (for us I1) the LogPrice of a clarity VVS1 level diamond is larger by $0.14.

clarity_IF:

In relation to the baselines color (for us I1) the LogPrice of a clarity IF level diamond is larger by $0.22.

Interpret R^2:

R^2 is 0.8763, which means that 87% of the variation in the data is explained by the regression model. This is a pretty large value, our model is explaining well.

Examples of using the equation for prediction

If a diamond has a carat value of .641, is color H, and has a clarity of IF, what will its price be?

$log\hat{Y}$ $= 6.174768 + 2.1436X_1 - 0.1968X_{color\_H} - 0.3399X_{color\_I} - 0.4958X_{color\_J} -0.2250X_{clarity\_SI2} + 0.1389X_{clarity\_VVS2} +0.1438X_{clarity\_VVS1} + 0.2217X_{clarity\_IF}$

$log\hat{Y}$ $= 6.174768 + 2.1436(.641) - 0.1968(1) - 0.3399(0)- 0.4958(0) -0.2250(0) + 0.1389(0) +0.1438(0) + 0.2217(1)$

$$log\hat{Y} \; =7.573716$$

$$\exp(log\hat{Y}) \; =\exp(7.573716)$$

Using the model, a diamond with a carat value of .641, is color H, and has a clarity of IF, will cost $1946.328.

If a diamond has a carat value of 2.3, is color J, and has a clarity of SI2, what will its price be?

$log\hat{Y}$ $= 6.174768 + 2.1436X_1 - 0.1968X_{color\_H} - 0.3399X_{color\_I} - 0.4958X_{color\_J} -0.2250X_{clarity\_SI2} + 0.1389X_{clarity\_VVS2} +0.1438X_{clarity\_VVS1} + 0.2217X_{clarity\_IF}$

$log\hat{Y}$ $= 6.174768 + 2.1436(2.3) - 0.1968(0) - 0.3399(0)- 0.4958(1) -0.2250(1) + 0.1389(0) +0.1438(0) + 0.2217(0)$

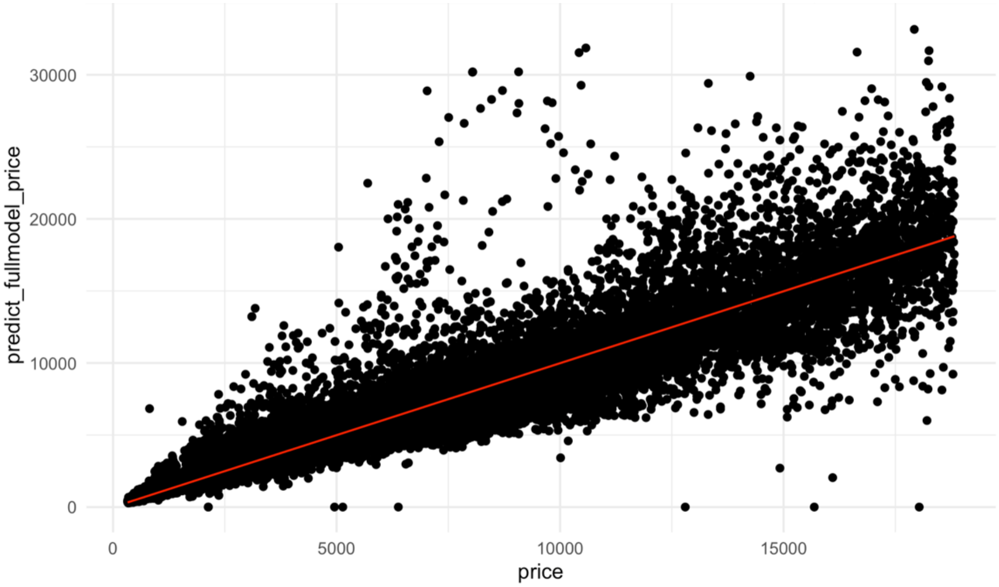
$$log\hat{Y} \; =10.38425$$

$$\exp(log\hat{Y}) \; =\exp(10.38425)$$

Using the model, a diamond with a carat value of 2.3, is color J, and has a clarity of SI2, will cost $32346.14.

Interpretation

Graph below that shows the prediction of price against actual price of diamond.



Through my analysis, I have found that the carat, color, and clarity of a diamond impacts its price. Carat impacts price the most, which makes sense since the larger the diamond, the more there is to sell. Clarity and color also impact the price of a diamond, which makes sense to me since the clearer and more colorful a diamond is, the more valuable it is to people since it is prettier to the human eye.

I am surprised that the selection algorithm determined that cut did not need to be included in the final model. I would think that cut would be highly predicted of price, since nicer cut diamonds are prettier, and I would think more valuable.