PSTAT 126 Final Project

Nicole Xu, Srinidhi Satnish, Leah Paredes

Part 1

Data Description and Descriptive Studies The diamonds dataset is a dataset that provides information regarding various diamond properties. The dataset contains the prices and other attributes of almost 54,000 diamonds. There are three types of categorical variables and seven numeric variables. Below is a table describing the 10 variables within the dataset.

Variable	Description	Type
cut	The quality of the cuts of the diamond	Categorical
color	Diamonds are graded on a color scale from D-Z normally, and D-G in this data set with D being the best color and G being the worst	Categorical
clarity	The purity of a diamond based on internal flaws and external blemishes. Worst (I1) to Best(IF)	Categorical
carat	A unit of mass weighing 200mg for the weight of gemstones	Numeric
price	Prices of diamonds in USD	Numeric
depth	Height of the diamond in mm from the table to the tip	Numeric
table	Diameter of the largest facet of the diamond	Numeric
x	Length of the diamond in mm	Numeric
y	Width of the diamond in mm	Numeric
z	Depth of the diamond in mm	Numeric

In order to gain a better understanding of the variables and possible data within the dataset we look at the first 6 observations of diamonds within the dataset.

```
## # 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
```

Our first step in analyzing the data is to select a random sample of 1000 data entries. With this random sample we use the summary function to find the general statistics of each variable. The data provided allows us to see the mean, standard deviation, minimum, and maximum of each numeric variable.

```
## # A tibble: 7 x 5
##   Variable     Mean  Std_Dev Minimum  Maximum
##   <chr>       <dbl>    <dbl>   <dbl>    <dbl>
## 1 carat       0.795    0.482     0.2     3.01
## 2 depth      61.8      1.37     57      69.3
## 3 table      57.4      2.40     50      95
## 4 price    3852.    3956.      351   18757
## 5 x           5.71     1.15      0       9.41
## 6 y           5.71     1.14      0       9.32
## 7 z           3.52     0.716     0       5.9
```

In contrast the categorical variable presents each classification within the variable and the amount of times it is seen in the sample. The categorical variables are color, cut, and clarity.

```
##  color          cut         clarity
##  D:116   Fair     : 32   SI1    :252
##  E:203   Good     : 92   VS2    :207
##  F:165   Very Good:196   SI2    :181
##  G:193   Premium  :261   VS1    :138
##  H:161   Ideal    :419   VVS2   : 98
##  I:113                   VVS1   : 70
##  J: 49                   (Other): 54
```

Visualization

Below, we have visualized the categorical variables (color, cut, clarity) in bar charts and the quantitative variables (carat, price, clarity, table, x, y, z) through histograms, in order to understand the distribution and frequency of each variable in our random selected sample.

Categorical

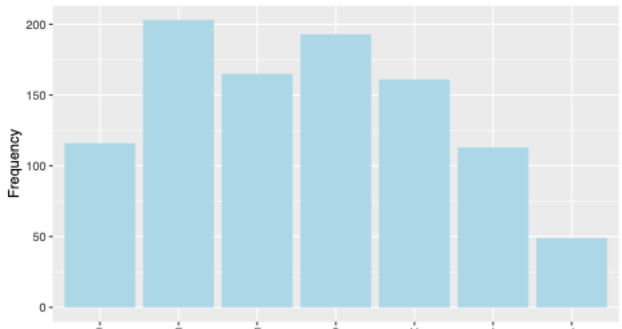


Figure 1

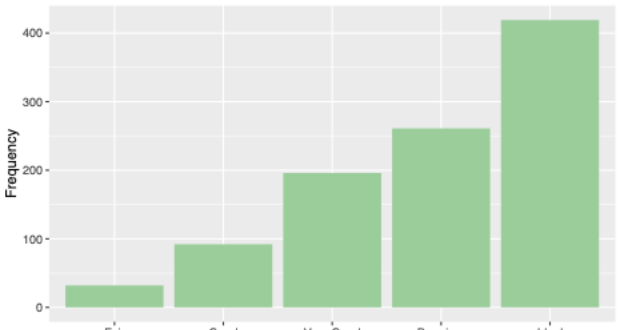


Figure 2

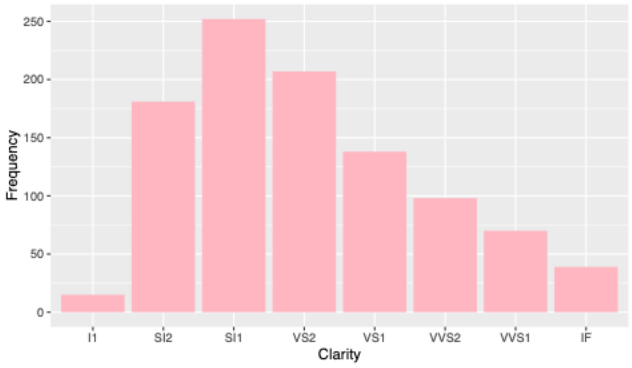


Figure 3

The graphs above show us the general distributions of the sampled categorical data. The color variable (Fig 1) had two distinct peaks with a slight right skew. This signifies the colors in the sample are generally high in grading with it gradually going down after the G color grade. The cut (Fig 2) variable has a strong and distinct left skew with most diamonds in the sample having the ideal cut and it lowering gradually as the cut worsens. The clarity (Fig 3) distribution has a distinct peak in the Sl1 grade with a strong right skew. This signifies most diamonds have a Sl1 grade and while less diamonds in the sample a better clarity grade.

Numerical

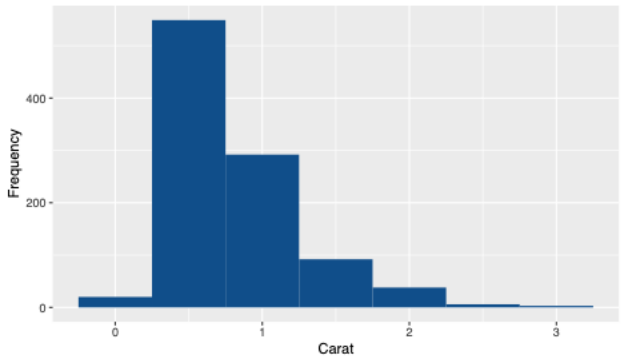


Figure 4

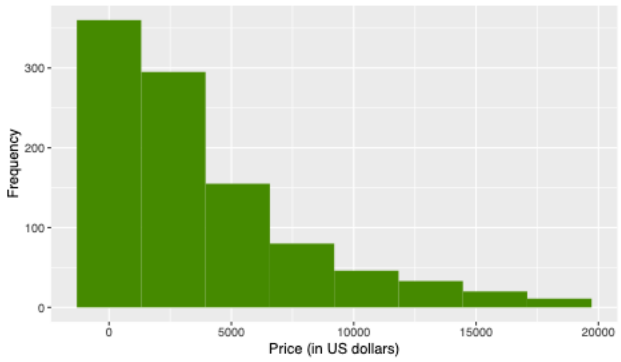


Figure 5

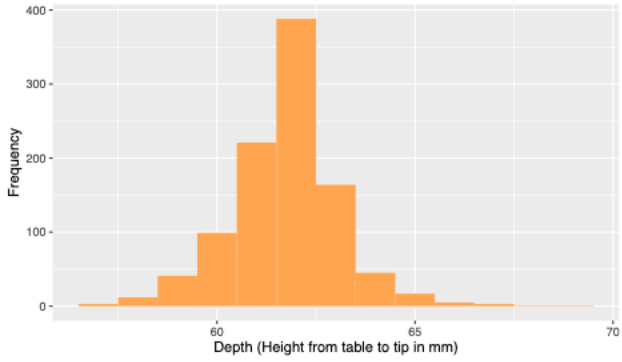


Figure 6

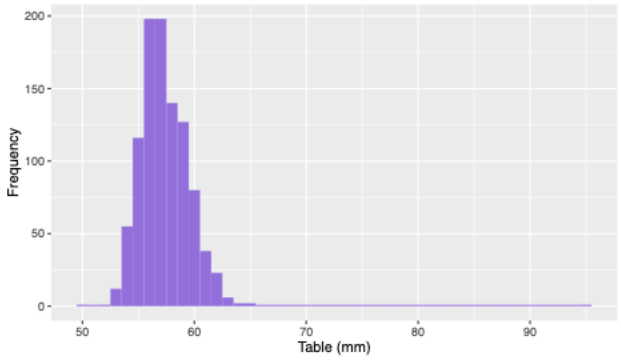


Figure 7

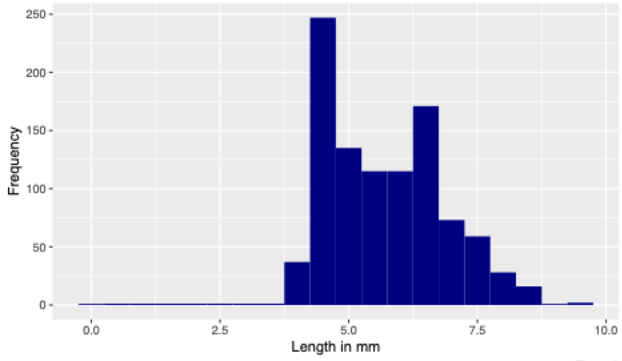


Figure 8

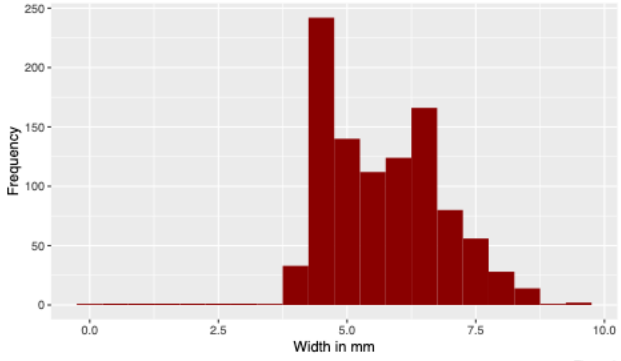


Figure 9

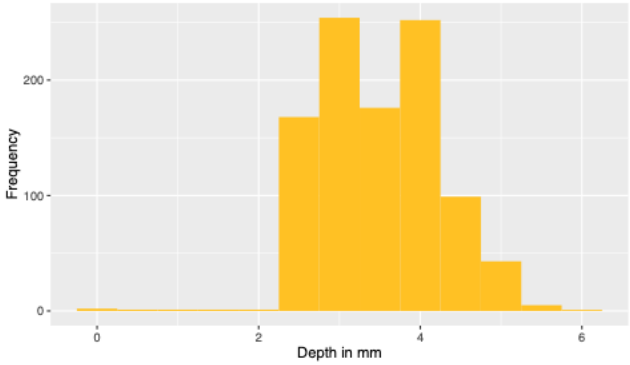


Figure 10

The numerical graphs represent many different outcomes of the sample. In the carat graph (Fig. 4) there is a strong right skew with most diamonds in the sample being between 0.5 - 1 carats. In the price graph (Fig. 5) there is also a strong right skew with the majority of the data being in between $0 - $2500. In Figure 6 the depth table resembles a normal distribution with the peak being below 62.5 mm. The table graph (Fig. 7) is also similar to a normal distribution with some data being present past the 65 mm. The length distribution (Fig. 8) had two distinct peaks in with a right skew as well. The width distribution is very similar to the length distribution (Fig. 9) with two distinct peaks and a slight right skew. The final graph (Fig. 10) focuses on the depth (mm) of the diamond. It is again similar to the other diamond measurements with two distinct peaks and a slight right skew.

Chosen Variables

After looking at the mentioned variables, we decided to further investigate **price (numerical)**, **carat(numerical)**, **depth(numerical)**, **color(categorical)**, and **cut(categorical)**. We want to know if our variables correlate so we use the cor function to see the correlation coefficients of our chosen numerical variables.

```
##            carat       price       depth
## carat 1.00000000  0.93132848  0.04534085
## price 0.93132848  1.00000000 -0.01501492
## depth 0.04534085 -0.01501492  1.00000000
```

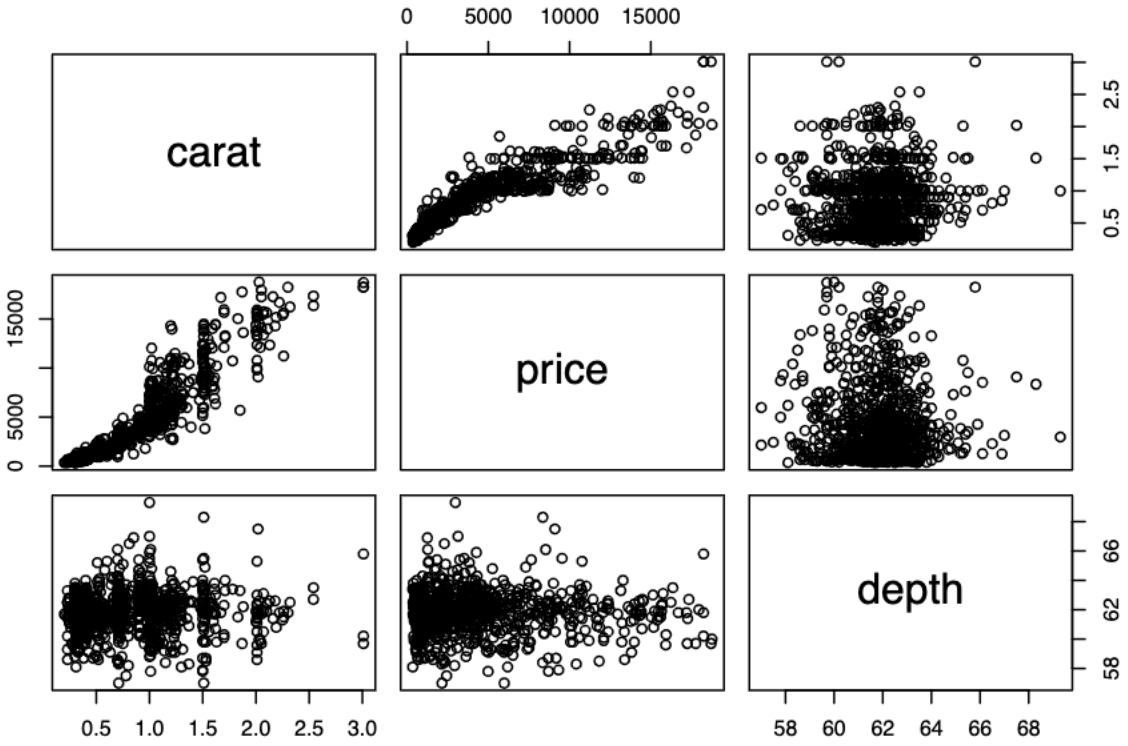
The correlation matrix helps identify linear relationships between pairs of quantitative variables. High correlations suggest that one variable can be predictive of another, while low correlations indicate little to no linear relationship. From the table, it is clear that carat is highly correlated with price as the correlation is above 0.9. In contrast, depth and price are very weakly correlated, with a correlation of - 0.0150. We can conclude studying price and depth wouldn't be too effective as they aren't closely correlated.

In order to see what variables would've worked in place of depth we decided to find the correlation between all numerical variables in the sample.

```
##            carat        depth      table       price           x           y
## carat 1.00000000  0.045340847  0.1702871  0.93132848  0.966029794  0.965047918
## depth 0.04534085  1.000000000 -0.2822655 -0.01501492 -0.006468547 -0.009074175
## table 0.17028706 -0.282265467  1.0000000  0.14448463  0.164942014  0.159615127
## price 0.93132848 -0.015014922  0.1444846  1.00000000  0.886313472  0.887271857
## x     0.96602979 -0.006468547  0.1649420  0.88631347  1.000000000  0.999003593
## y     0.96504792 -0.009074175  0.1596151  0.88727186  0.999003593  1.000000000
## z     0.93351964  0.111484527  0.1234854  0.84923683  0.965532106  0.965644969
##               z
## carat 0.9335196
## depth 0.1114845
## table 0.1234854
## price 0.8492368
## x     0.9655321
## y     0.9656450
## z     1.0000000
```

We can see that other variables with high correlations are price with x, y and z. These relationships all have correlation coefficients above a .9 for all those relationships. Additionally, x is highly correlated with y and z, and y is highly correlated with z.

Additionally, we'd like to examine relationships between categorical variables and numerical variables so we created plots to examine each variable with one another.



Multiple Linear Regression Model

Below is the summary of the multiple linear regression model for our previously chosen variables.

```
##
## Call:
## lm(formula = price ~ ., data = selected_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5998.8  -762.5     7.9   545.1  7312.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7978.69    2029.50   3.931 9.03e-05 ***
## carat        7661.40      93.53  81.911  < 2e-16 ***
## depth        -165.33      32.86  -5.031 5.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1424 on 997 degrees of freedom
## Multiple R-squared:  0.8707, Adjusted R-squared:  0.8704
## F-statistic:  3356 on 2 and 997 DF,  p-value: < 2.2e-16
```

Part 2: Simple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Under the assumptions that the error term ϵ is observed as an independent and identically distributed random variable with mean 0 and constant variance σ^2

$$\mathbb{E}[\epsilon] = 0 \ \text{Var}(\epsilon) = \sigma^2 \begin{matrix} \epsilon_i \perp \epsilon_j \text{ for } i \neq j \\ \epsilon \sim N(0, \sigma^2) \end{matrix}$$

From the correlation matrix created in part 1, we choose *carat* as our predictor because it is the most correlated with *price*. Therefore our fitted linear regression model is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

and the predicted values are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i (i = 1, 2, ..., n)$$

where y = price and x = carat.

Null hypothesis: There is no linear relationship between price and carat.

Alternate hypothesis: There is a linear relationship between price and carat.

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

In turn we run a simple liner regression model with carat as the predictor and price as the response.

```
##
## Call:
## lm(formula = price ~ carat, data = diamond_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6224.1  -838.4    24.6   552.2  7353.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2221.99      87.92  -25.27   <2e-16 ***
## carat        7640.07      94.57   80.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1441 on 998 degrees of freedom
## Multiple R-squared:  0.8674, Adjusted R-squared:  0.8672
## F-statistic:  6527 on 1 and 998 DF,  p-value: < 2.2e-16
```

Interpretation of Summary Statistics:

Since the result of the p-value is less than 2.2e-16 it also less than .05 which indicates that we reject our Null Hypothesis in favor of the Alternate Hypothesis. We can conclude there is a linear relationship between carat and price.

R^2 - The R-squared value is 0.8674, which means that approximately 86.74% of the variability in diamond price can be explained by the carat weight.

Adjusted R^2 - The adjusted R-squared value is 0.8524, which is very close to the R-squared value, indicating a good fit.

Intercept - The estimated intercept is -2221.99 This means that if the carat weight is 0, the price of the diamond would be -2221.99, which is not practically meaningful but is part of the linear equation.

Slope - The estimated slope is 7640.07. This means that for every additional carat, the price of the diamond increases by $7640.07 .

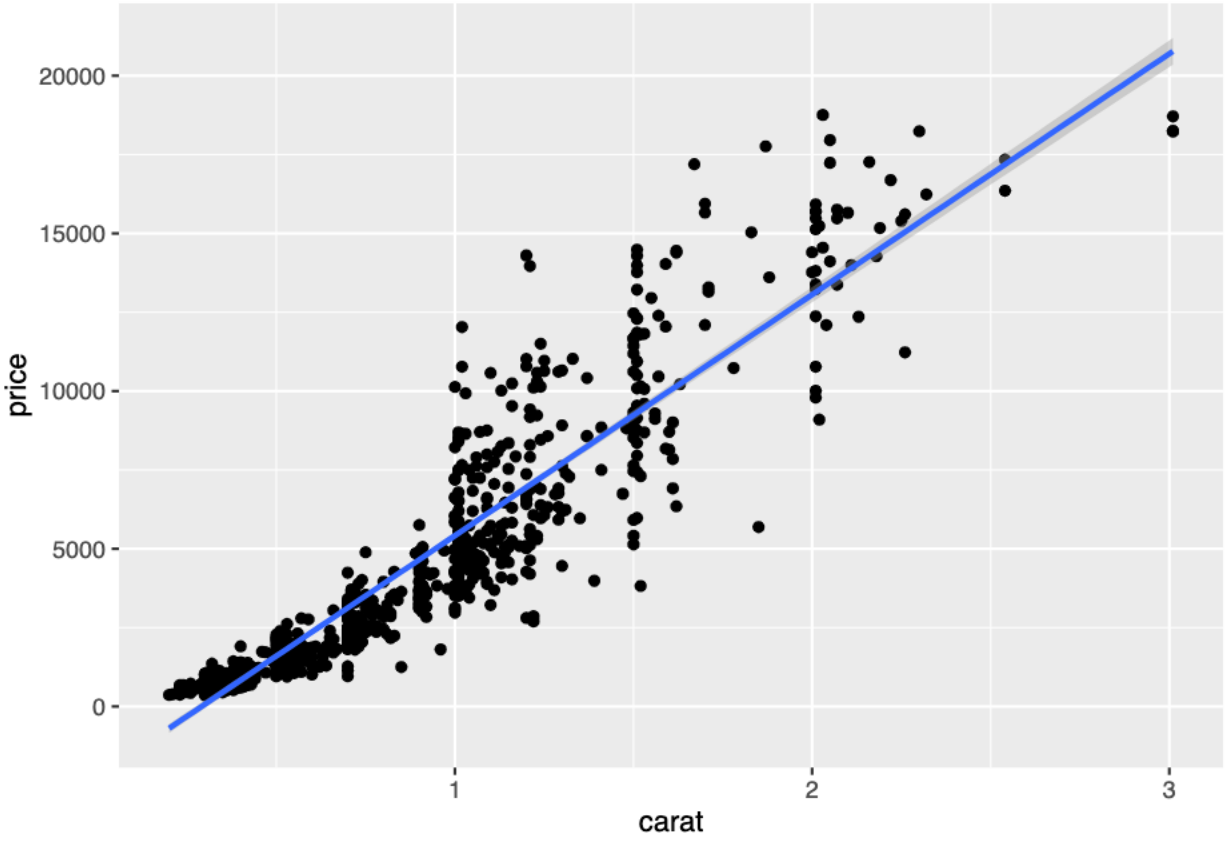
p-value - The p-value for the slope is less than 2.2e-16, which is highly significant, indicating a strong relationship between carat and price.

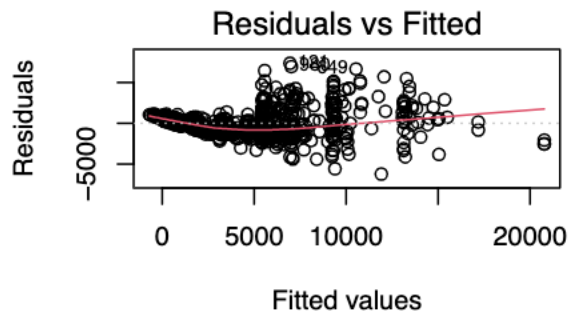
```
##                 2.5 %    97.5 %
## (Intercept) -2394.522 -2049.467
## carat        7454.490  7825.641
```

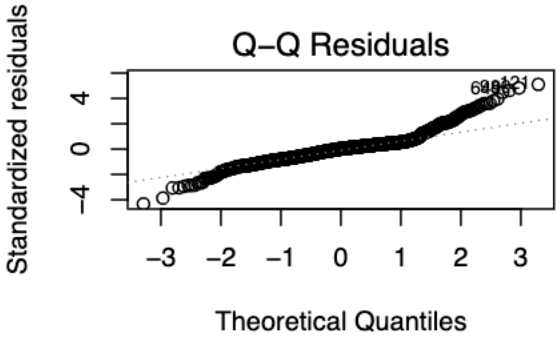
```
##        fit      lwr      upr
## 1 9238.104 6405.233 12070.98
```

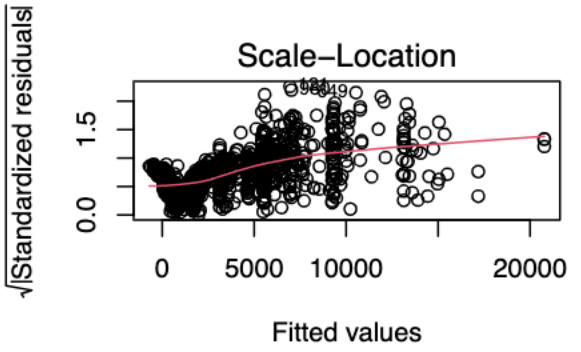
The confidence interval for β_1 is [7454.490, 7825.641]. The prediction interval for a 1.5 carat diamond is [6405.23, 12070.98]. The prediction is wider than the confidence interval because it accounts for the variability around the regression line and individual observations.

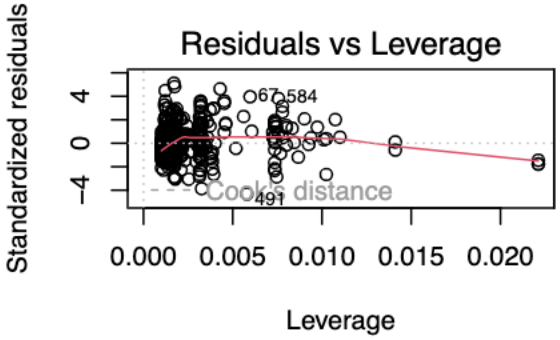
```
## `geom_smooth()` using formula = 'y ~ x'
```











The residual vs fitted graph has a slight curve with lots of data in the center of the graph which signifies it's not a linear relationship. The QQ plot has a line that is straight, but there are points towards the far

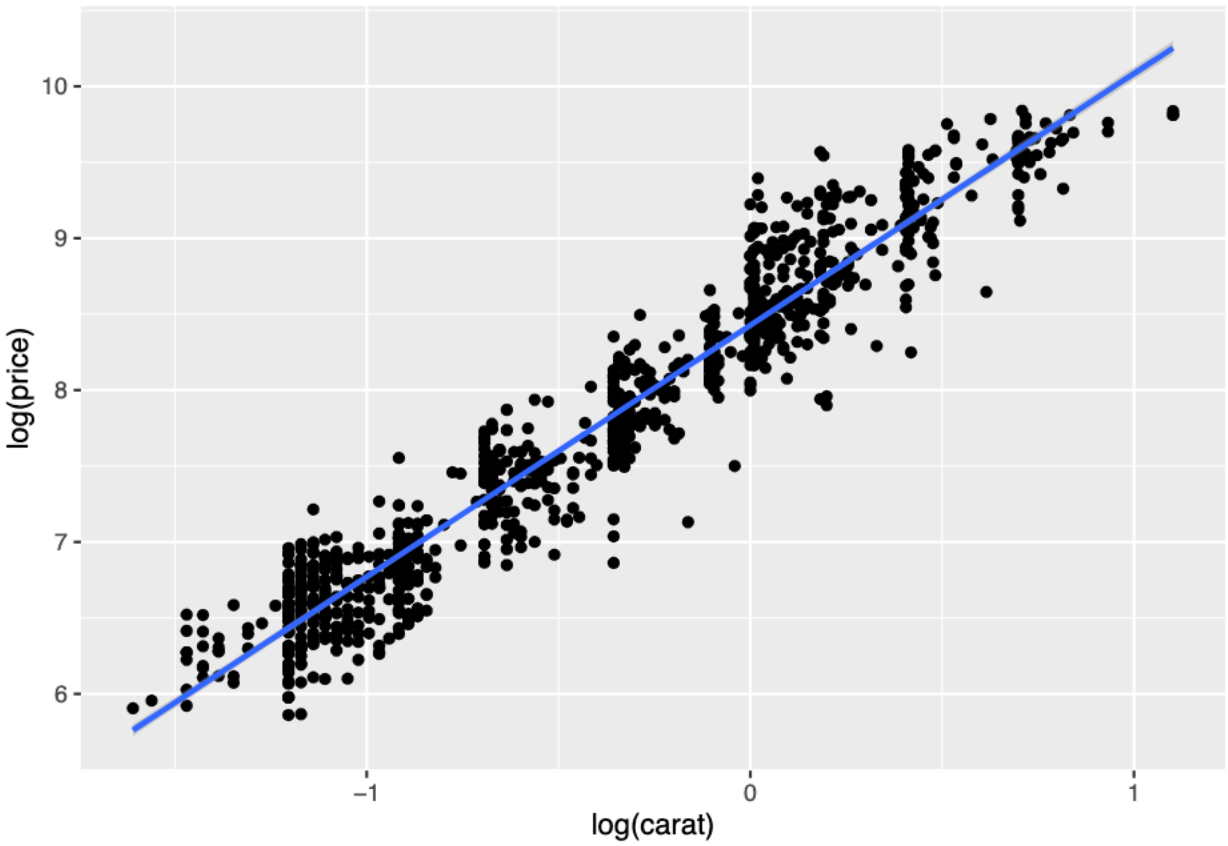
right and the far left that show it's not the most linear relationship. The scale location graph doesn't have a straight line and also has more data on the left side of the graph. In the residuals vs leverage plot there are multiple points with high leverage.

Transformations

We will transform both variables the model with the log function.

```
##
## Call:
## lm(formula = log(price) ~ log(carat), data = diamond_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0275 -0.1661 -0.0093  0.1702  0.9347
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.42760    0.01022   824.8   <2e-16 ***
## log(carat)   1.65667    0.01430   115.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2669 on 998 degrees of freedom
## Multiple R-squared:  0.9308, Adjusted R-squared:  0.9307
## F-statistic: 1.343e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



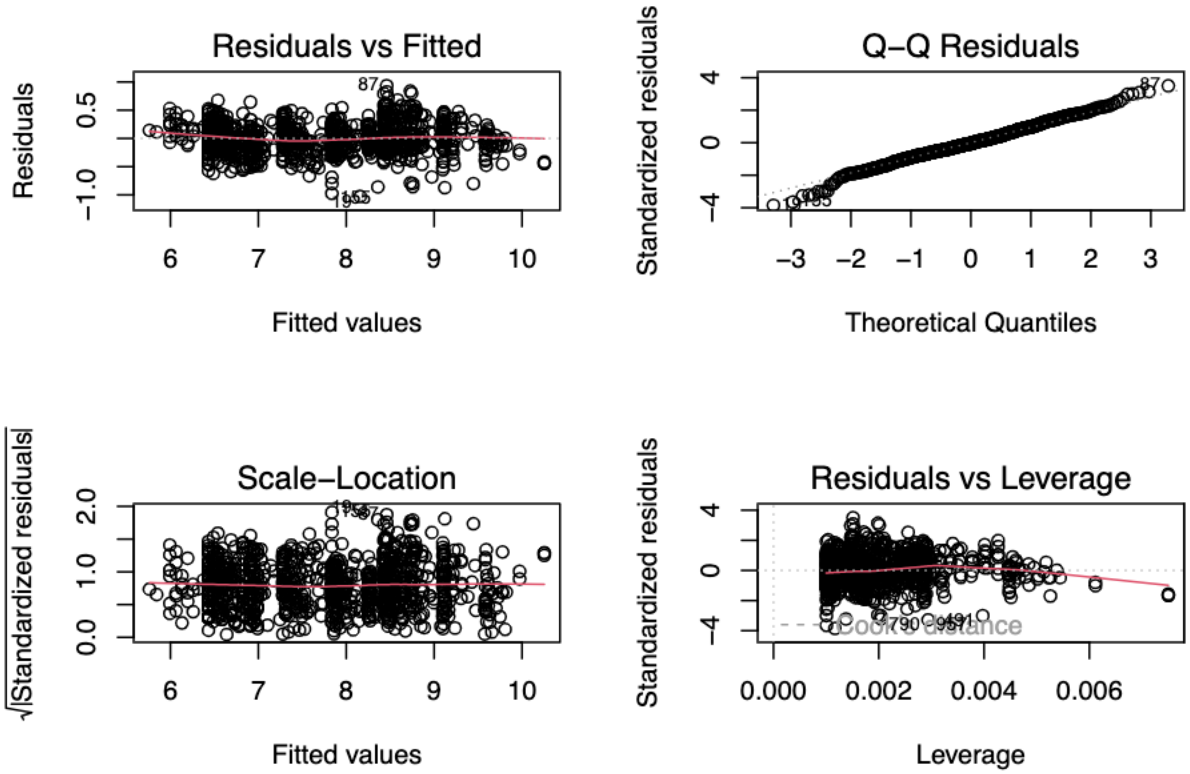
The graph shows constant variance and shows the mean response is linear in the predictors.

Slope - The estimated slope is 8.42760, meaning for every one-unit increase

in carat, the logarithm of price increases by around $8.43.

Adjusted R^2 - The adjusted R-squared value is 0.9307, indicating that about

93.07% of the variability in the logarithm of price is explained by logarithm carat. **In comparison to the previous model, it's clear the transformation produces a better model.**



The standardized residual plot shows that the residuals are evenly scattered, which satisfies the assumption that variance is constant and the expectation of error is 0. The Q-Q plot being linear satisfies the assumption that variance is normally distributed. Additionally the plot between log(carat) and log(price) shows a linear relationship, satisfying the linearity assumption.

Other Predictors After testing other predictors in the model, it was discovered that the variables carat, color, clarity and cut helped strengthen the model. From the summary provided below the adjusted R^2, .9839, has increased significantly from the previous transformed model.

```
##
## Call:
## lm(formula = log(price) ~ log(carat) + clarity + cut + color,
##     data = diamond_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38855 -0.08614  0.00160  0.08826  0.42006
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  8.4370857  0.0082450 1023.292  < 2e-16 ***
## log(carat)   1.8723209  0.0081636  229.350  < 2e-16 ***
## clarity.L    0.9575334  0.0249391   38.395  < 2e-16 ***
## clarity.Q   -0.3035950  0.0230583  -13.166  < 2e-16 ***
## clarity.C    0.1620197  0.0198216    8.174 9.16e-16 ***
## clarity^4   -0.0671327  0.0160039   -4.195 2.98e-05 ***
## clarity^5    0.0453798  0.0133140    3.408  0.00068 ***
## clarity^6    0.0065960  0.0115880    0.569  0.56934
## clarity^7    0.0213808  0.0106290    2.012  0.04454 *
## cut.L        0.1121472  0.0167068    6.713 3.23e-11 ***
```

```
## cut.Q       -0.0367886  0.0147218   -2.499  0.01262 *
## cut.C        0.0085470  0.0128427    0.666  0.50588
## cut^4       -0.0134282  0.0107125   -1.254  0.21032
## color.L     -0.4180178  0.0149025  -28.050  < 2e-16 ***
## color.Q     -0.1134473  0.0139446   -8.136 1.23e-15 ***
## color.C     -0.0184972  0.0126530   -1.462  0.14409
## color^4      0.0005899  0.0113613    0.052  0.95860
## color^5      0.0115079  0.0107818    1.067  0.28608
## color^6      0.0083661  0.0101078    0.828  0.40805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1316 on 981 degrees of freedom
## Multiple R-squared:  0.9835, Adjusted R-squared:  0.9832
## F-statistic:  3242 on 18 and 981 DF,  p-value: < 2.2e-16
```

Overfitting & Multicollinearity To test for overfitting, we generate predictions for new unseen observations not used to build the model. For this, we create a subset of "new data" from the original diamonds dataset that is not included in the current sample, to see how well our models do with data it has not been trained on

```
## Joining with `by = join_by(carat, cut, color, clarity, depth, table, price, x,
## y, z)`
```

```
## [1] 0.8276603
```

The R^2 value is about.8404 which means about84.04% of the variability in the logarithm of price can be explained by `model_simple` when using new data. The R^2 value from the original sample was .9835 and since our new R^2 value is lower, but still relatively good explanatory power we can assume there is some overfitting within the model.

```
##                GVIF Df GVIF^(1/(2*Df))
## log(carat) 1.341350  1        1.158166
## clarity    1.443853  7        1.026584
## cut        1.154042  4        1.018070
## color      1.209191  6        1.015955
```

With the vif (Variance Inflation Factor) function we are able to see if there is multicollinearity among predictors. Since no values are above or nearing ten it is safe to say there is no evidence of multicollinearity. This paired with the high adjusted R^2 value indicates there is little overfitting as well.

Part 3

From the previous step we concluded that the best model is a transformed linear model with the variables log(price), log(carat), color, cut, and clarity.

```
##
## Call:
## lm(formula = log(price) ~ log(carat) + clarity + cut + color,
##     data = diamond_sample)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38855 -0.08614  0.00160  0.08826  0.42006
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  8.4370857  0.0082450 1023.292  < 2e-16 ***
## log(carat)   1.8723209  0.0081636  229.350  < 2e-16 ***
## clarity.L    0.9575334  0.0249391   38.395  < 2e-16 ***
## clarity.Q   -0.3035950  0.0230583  -13.166  < 2e-16 ***
## clarity.C    0.1620197  0.0198216    8.174 9.16e-16 ***
## clarity^4   -0.0671327  0.0160039   -4.195 2.98e-05 ***
## clarity^5    0.0453798  0.0133140    3.408  0.00068 ***
## clarity^6    0.0065960  0.0115880    0.569  0.56934
## clarity^7    0.0213808  0.0106290    2.012  0.04454 *
## cut.L        0.1121472  0.0167068    6.713 3.23e-11 ***
## cut.Q       -0.0367886  0.0147218   -2.499  0.01262 *
## cut.C        0.0085470  0.0128427    0.666  0.50588
## cut^4       -0.0134282  0.0107125   -1.254  0.21032
## color.L     -0.4180178  0.0149025  -28.050  < 2e-16 ***
## color.Q     -0.1134473  0.0139446   -8.136 1.23e-15 ***
## color.C     -0.0184972  0.0126530   -1.462  0.14409
## color^4      0.0005899  0.0113613    0.052  0.95860
## color^5      0.0115079  0.0107818    1.067  0.28608
## color^6      0.0083661  0.0101078    0.828  0.40805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1316 on 981 degrees of freedom
## Multiple R-squared:  0.9835, Adjusted R-squared:  0.9832
## F-statistic:  3242 on 18 and 981 DF,  p-value: < 2.2e-16
```

R^2 - The R-squared value is 0.9835, which means that approximately 98.35% of the variability in diamond log price can be explained by the log carat weight.

Adjusted R^2 - The adjusted R-squared value is 0.9832, which is very close to the R-squared value, indicating a good fit.

Intercept - The estimated intercept is 8.4371 This means that if the log of carat weight is 0, the price of the diamond would be 8.4371, which is not practically meaningful but is part of the linear equation.

Coefficients - For every additional log carat the price increases by 1.8723209. Each additional predictor has it's differing weight on the price.

p-value - The p-value for the slope is less than 2.2e-16, which is highly significant, indicating a strong relationship between carat and price.

AIC

```
##
## Call:
## lm(formula = price ~ ., data = diamond_sample, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4093.3  -578.2  -178.8   376.2  5371.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3631.9434  2684.5128   1.353 0.176393
## carat       10189.7875   272.0236  37.459  < 2e-16 ***
## cut.L         517.5241   150.1800   3.446 0.000593 ***
## cut.Q        -109.6223   119.2121  -0.920 0.358032
## cut.C         159.9200   106.1877   1.506 0.132387
## cut^4        -106.7920    86.1397  -1.240 0.215364
## color.L     -1679.4432   117.1035 -14.342  < 2e-16 ***
## color.Q      -605.9038   108.2981  -5.595 2.87e-08 ***
## color.C      -111.4600    97.9659  -1.138 0.255507
## color^4       -49.1463    88.1357  -0.558 0.577232
## color^5       -57.3357    83.5935  -0.686 0.492947
## color^6       -94.1824    78.3455  -1.202 0.229600
## clarity.L    3849.5747   193.6727  19.877  < 2e-16 ***
## clarity.Q   -1881.4023   180.2987 -10.435  < 2e-16 ***
## clarity.C     860.4551   153.8455   5.593 2.90e-08 ***
## clarity^4    -186.2207   123.9233  -1.503 0.133238
## clarity^5      88.8901   103.0267   0.863 0.388467
## clarity^6      56.0287    89.8110   0.624 0.532871
## clarity^7      49.6780    82.2476   0.604 0.545980
## depth         -75.3776    32.5673  -2.315 0.020846 *
## table          -0.3694    16.9736  -0.022 0.982642
## x            -860.5327   747.1098  -1.152 0.249679
## y             179.4553   751.1162   0.239 0.811219
## z              58.2754   197.2529   0.295 0.767724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1017 on 976 degrees of freedom
## Multiple R-squared:  0.9355, Adjusted R-squared:  0.934
## F-statistic: 615.4 on 23 and 976 DF,  p-value: < 2.2e-16
```

We decided to use stepwise regression using AIC to find the best model

```
## Start:  AIC=13871.95
## price ~ carat + cut + color + clarity + depth + table + x + y +
##     z
##
##           Df  Sum of Sq        RSS   AIC
## - table    1        489 1008477022 13870
## - y        1      58981 1008535514 13870
## - z        1      90186 1008566719 13870
## - x        1    1370824 1009847357 13871
## <none>                  1008476533 13872
```

```
## - depth    1    5535254 1014011787 13875
## - cut      4   19386741 1027863273 13883
## - color    6  232463742 1240940275 14067
## - clarity  7  668614630 1677091163 14367
## - carat    1 1449883369 2458359902 14761
##
## Step:  AIC=13869.95
## price ~ carat + cut + color + clarity + depth + x + y + z
##
##           Df  Sum of Sq        RSS   AIC
## - y        1      59421 1008536443 13868
## - z        1      89902 1008566924 13868
## - x        1    1372100 1009849123 13869
## <none>                  1008477022 13870
## + table    1        489 1008476533 13872
## - depth    1    6409262 1014886285 13874
## - cut      4   24453283 1032930305 13886
## - color    6  233254184 1241731207 14066
## - clarity  7  668645844 1677122866 14365
## - carat    1 1470419834 2478896857 14767
##
## Step:  AIC=13868.01
## price ~ carat + cut + color + clarity + depth + x + z
##
##           Df  Sum of Sq        RSS   AIC
## - z        1     109945 1008646387 13866
## <none>                  1008536443 13868
## + y        1      59421 1008477022 13870
## + table    1        929 1008535514 13870
## - depth    1    6773458 1015309901 13873
## - x        1   16712735 1025249178 13882
## - cut      4   24553048 1033089491 13884
## - color    6  233334191 1241870634 14064
## - clarity  7  684872455 1693408898 14372
## - carat    1 1472127452 2480663895 14766
##
## Step:  AIC=13866.12
## price ~ carat + cut + color + clarity + depth + x
##
##           Df  Sum of Sq        RSS   AIC
## <none>                  1008646387 13866
## + z        1     109945 1008536443 13868
## + y        1      79463 1008566924 13868
## + table    1        536 1008645851 13868
## - depth    1    7288082 1015934469 13871
## - cut      4   24533351 1033179739 13882
## - x        1   33677736 1042324124 13897
## - color    6  233334624 1241981011 14062
## - clarity  7  684784104 1693430492 14370
## - carat    1 1481492524 2490138912 14768
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
```

```
##     x, data = diamond_sample, na.action = na.exclude)
##
## Coefficients:
## (Intercept)        carat        cut.L        cut.Q        cut.C        cut^4
##     3422.68     10183.12       518.96      -112.60       169.35      -100.19
##     color.L      color.Q      color.C      color^4      color^5      color^6
##    -1677.78      -605.63      -110.46       -48.29       -56.38       -94.78
##   clarity.L    clarity.Q    clarity.C    clarity^4    clarity^5    clarity^6
##     3855.42     -1887.73       863.27      -188.19        90.17        55.74
##   clarity^7        depth            x
##       49.68       -72.32      -644.66
```

The model originally started with an AIC of 13871.95, however after stepwise regression, the best model has an AIC of 13866.12. The reduction in AIC shows that the model has been improved. Thus the best predictors for price include carat, cut, color, clarity, depth, and x (length in mm).

```
##
## Call:
## lm(formula = price ~ carat + cut + color + clarity + depth +
##     x, data = diamond_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4082.7  -577.0  -177.6   387.3  5362.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3422.68    1828.15   1.872 0.061475 .
## carat       10183.12     268.54  37.920  < 2e-16 ***
## cut.L         518.96     141.93   3.657 0.000269 ***
## cut.Q        -112.60     118.11  -0.953 0.340636
## cut.C         169.35      99.29   1.706 0.088399 .
## cut^4        -100.19      82.86  -1.209 0.226888
## color.L     -1677.78     116.60 -14.389  < 2e-16 ***
## color.Q      -605.63     108.12  -5.602 2.76e-08 ***
## color.C      -110.46      97.71  -1.131 0.258540
## color^4       -48.29      87.85  -0.550 0.582654
## color^5       -56.38      83.31  -0.677 0.498682
## color^6       -94.78      78.01  -1.215 0.224666
## clarity.L    3855.42     191.95  20.085  < 2e-16 ***
## clarity.Q   -1887.73     178.02 -10.604  < 2e-16 ***
## clarity.C     863.27     153.09   5.639 2.24e-08 ***
## clarity^4    -188.19     123.54  -1.523 0.128004
## clarity^5      90.17     102.83   0.877 0.380769
## clarity^6      55.74      89.57   0.622 0.533902
## clarity^7      49.68      82.09   0.605 0.545201
## depth         -72.32      27.19  -2.660 0.007949 **
## x            -644.66     112.76  -5.717 1.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1015 on 979 degrees of freedom
## Multiple R-squared:  0.9355, Adjusted R-squared:  0.9342
## F-statistic: 709.7 on 20 and 979 DF,  p-value: < 2.2e-16
```

Looking at the best model generated from stepwise AIC regression, the adjusted

$$R^2$$

value is 0.9342 which means that 93.42% of the variability in price can be explained by the linear combination of cut, color, clarity, depth, and x. The intercept means that when all of the coefficients of the predictors are zero, the price of the diamond is 3422.68. While this doesn't make sense in real life, the intercept is important in the context of fitting a good model. The coefficient for carat is 10183.12 which means that when the unit increase by one the price increases by $10183.12.

Transformed Log Model (Part 2) Since we applied a log transformation to price and carat, we need to reverse that transformation when calculating upper and lower bounds

We're predicting the price of the following combination.

- Carat: 0.72
- Cut: Ideal
- Color: E
- Clarity: SI1
- Depth: 59.8
- Table: 57.0
- Price: 2265
- x: 5.86
- y: 5.91
- z: 3.52

```
##        fit      lwr      upr
## 1 362.0722 345.7932 379.1176
```

With 95% confidence, the mean log(price) for diamonds with these characteristics lies within [2719.704, 2862.734].

```
##        fit      lwr      upr
## 1 362.0722 278.5474 470.6427
```

With 95% confidence, a future observation of log(price) for a diamond with these characteristics lies within [2152.634, 3616.867].

Summary

The initial look at the dataset showed multiple variables relating to the diamond set. After visualizing and interpreting a sample of the available data we were able to reduce the amount of variables needed in our model and see how they affected a linear model. We were able to see the correlation between predictors and pick variables needed to make a simple linear regression model. For our simple linear regression model between carat as our predictor and price as our response, we found that our residual plot showed a pattern, so we opted to transform our model by taking the log of price and carat, which not only created a scattered residual plot, but also increased our R^2. Then, we experimented with adding different predictors, and landed on a multiple linear regression model, which increased our R^2. We then checked for multicollinearity using VIF and overfitting using AIC, concluding that our model had no signs of multicollinearity but showed signs of slight overfitting. We decided to find the best model with stepwise regression using AIC because slowly adding and removing variables would best minimize AIC. Then we constructed confidence and prediction intervals based on our final linear model from Part 2