Diamonds Exploration by Chris Saden

> **Tip**: You will see quoted blocks like this throughout this example project with tips for constructing your reports. You should consider these quoted sections as outside of the example structure.

> **Tip**: Unless there is a good exception, you will want to hide code and warnings from the output of the HTML. You should try to make your visualizations and tables interpretable without needing to analyze the code. In order to format your code chunks so that they do not show up in output, you can set the following parameters as global settings for the full document or in the chunk headers, e.g.: `{r echo=FALSE, message=FALSE, warning=FALSE}`

This report explores a dataset containing prices and attributes for approximately 54,000 diamonds.

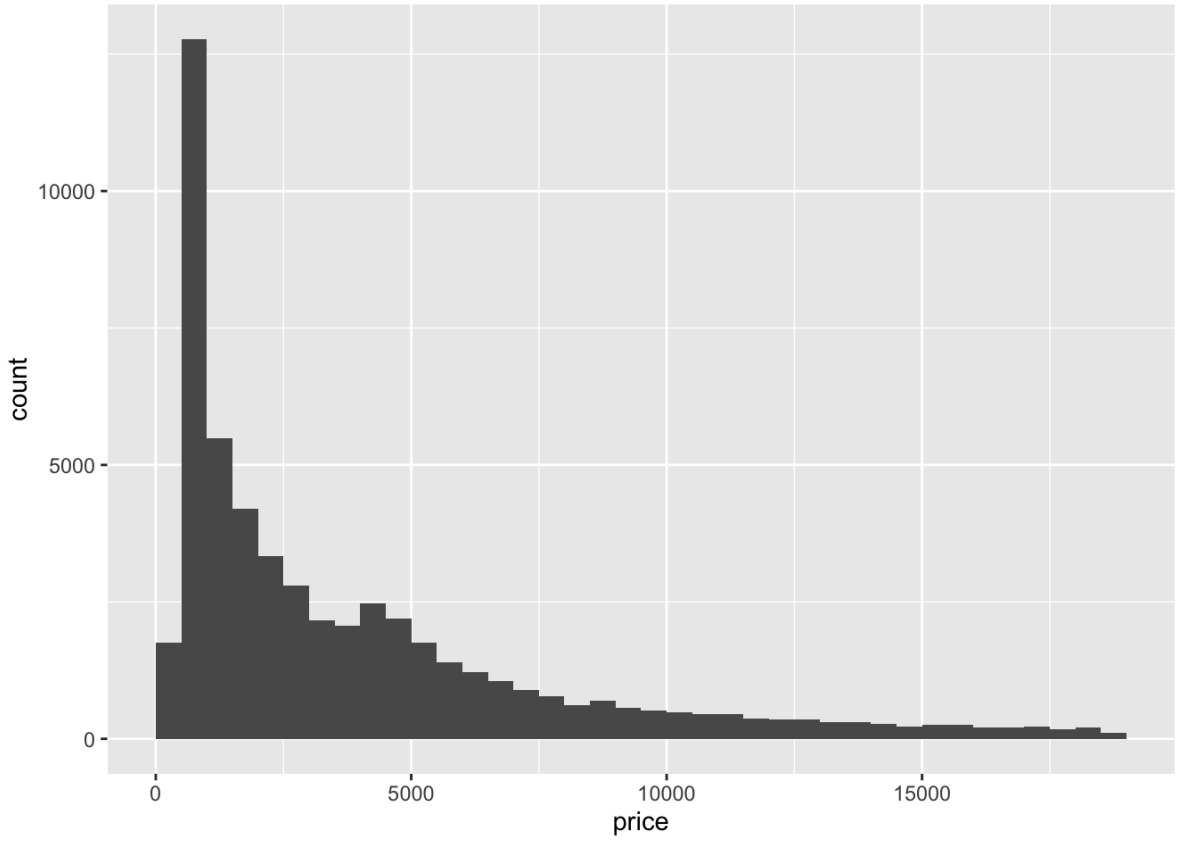
Univariate Plots Section

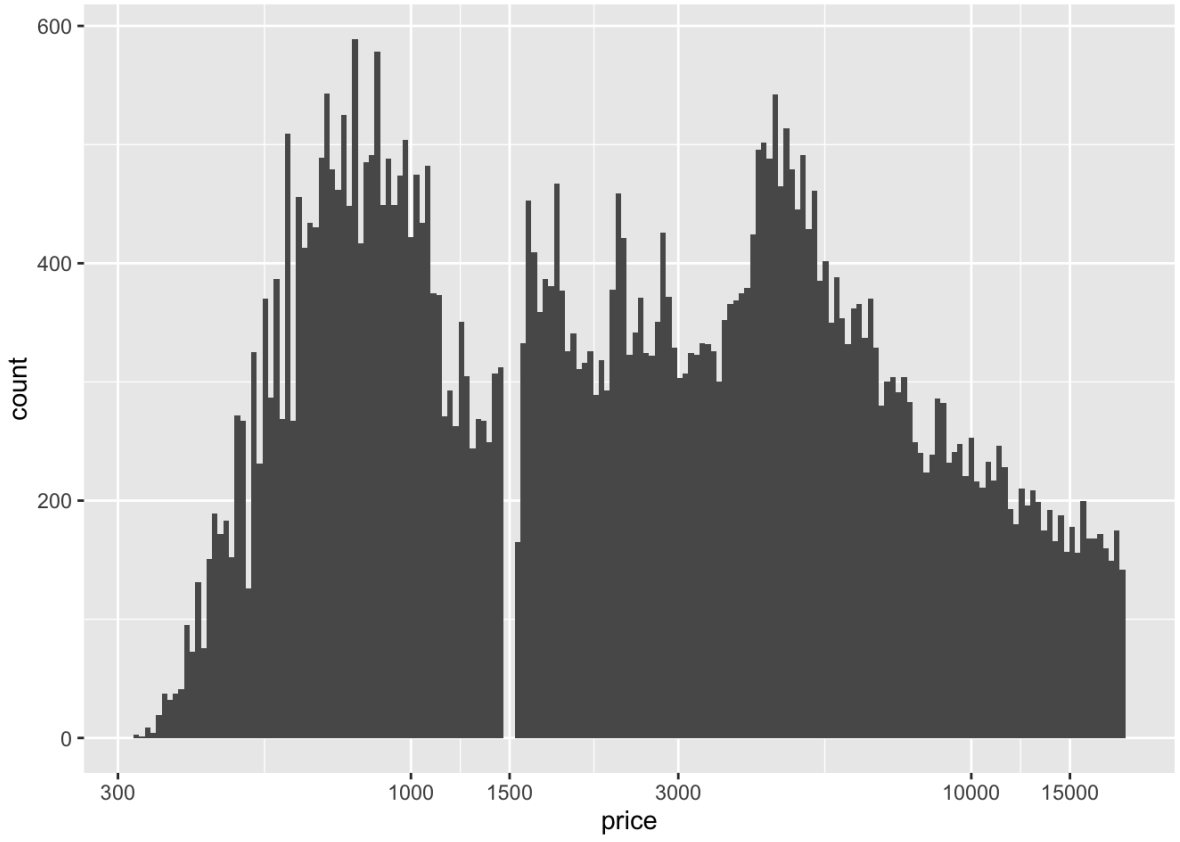
```
## [1] 53940    10
```

```
## Classes 'tbl_df', 'tbl' and '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    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 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
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066
##  Max.   :5.0100                     I: 5422   VVS1   : 3655
##                                     J: 2808   (Other): 2531
##      depth           table           price             x
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710
##  Median :61.80   Median :57.00   Median : 2401   Median : 5.700
##  Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740
##
##        y               z
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.710   Median : 3.530
##  Mean   : 5.735   Mean   : 3.539
##  3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :58.900   Max.   :31.800
##
```

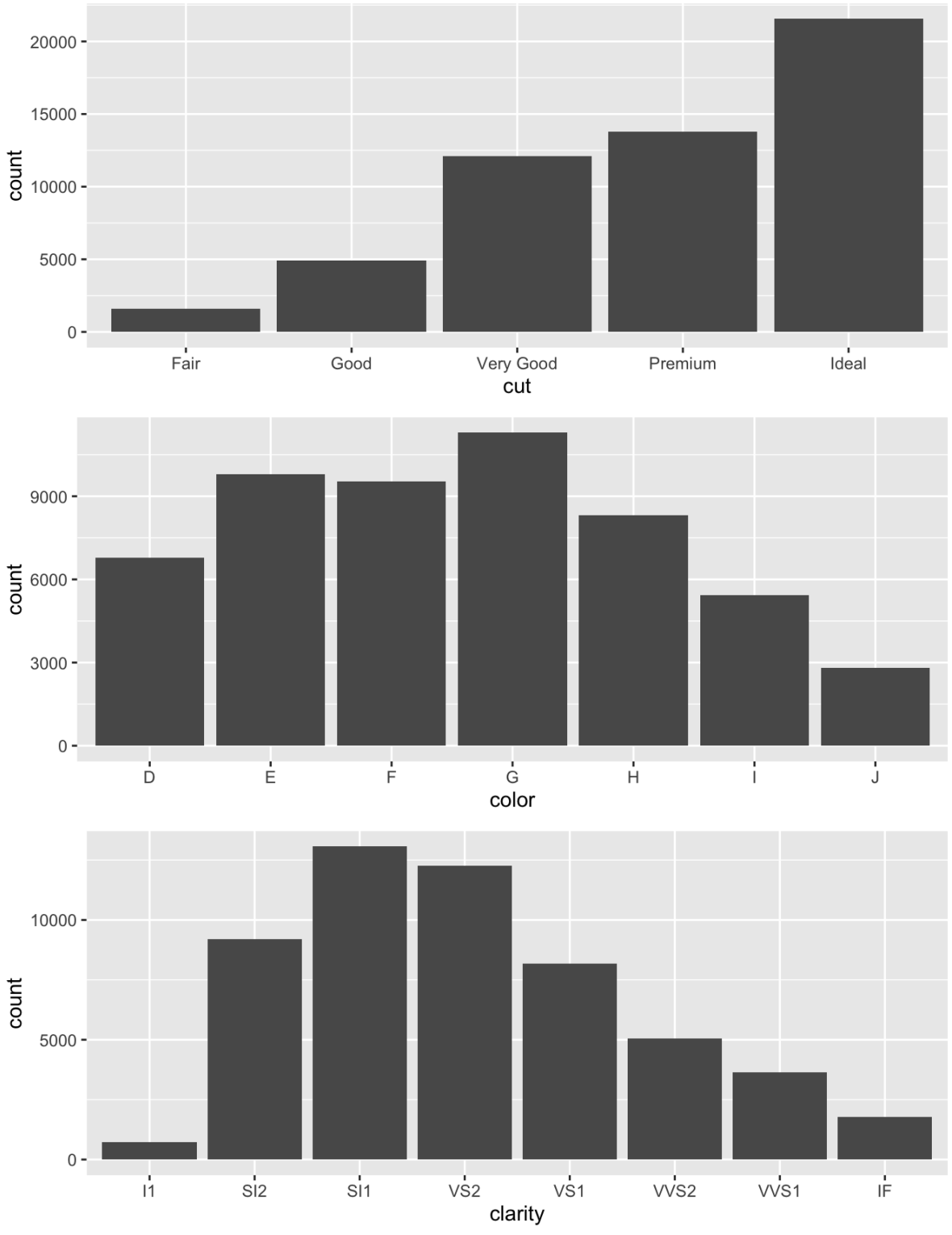
Our dataset consists of ten variables, with almost 54,000 observations.





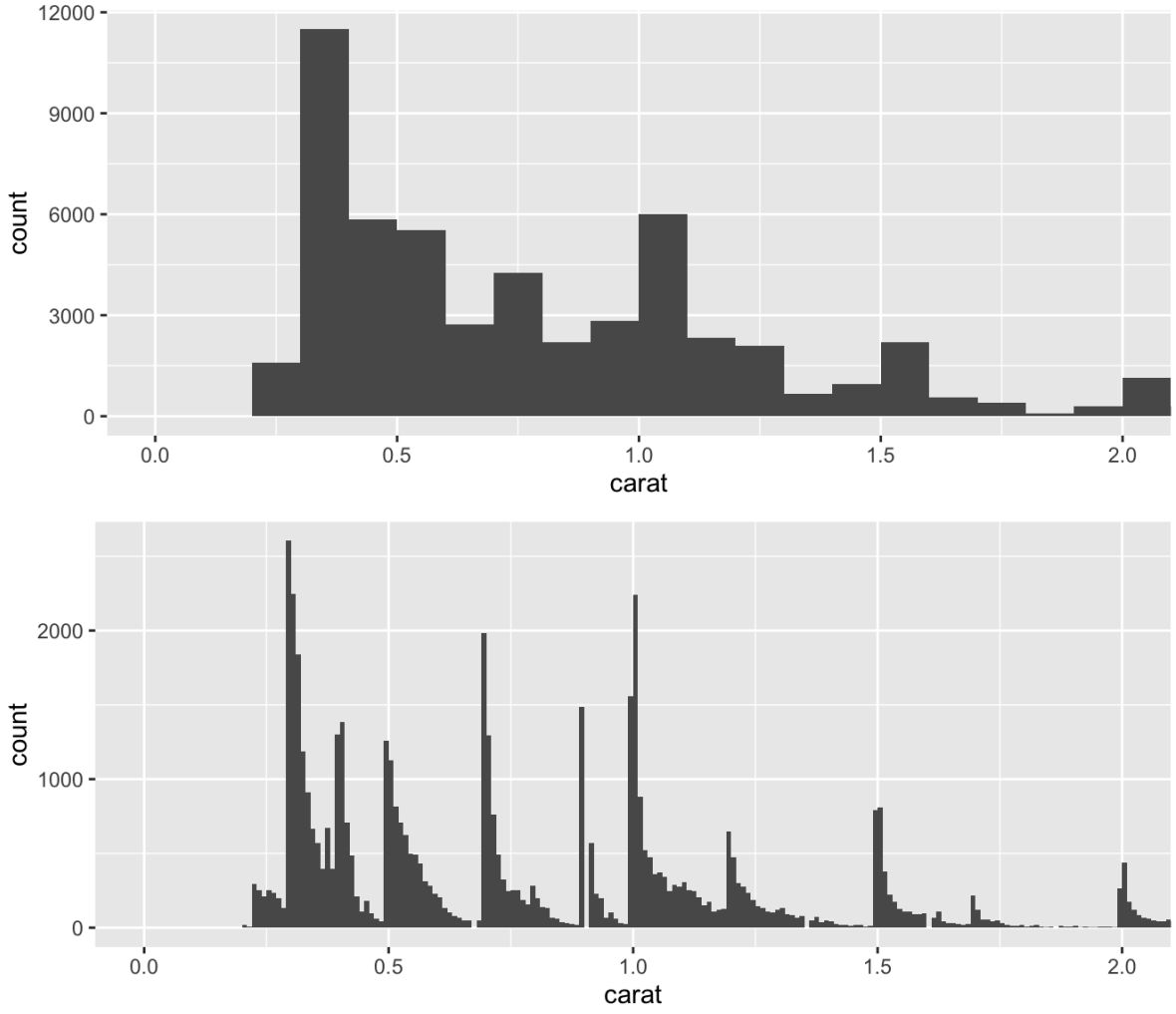
Tip: When plotting on a log scale, it is useful to note that 3 is about halfway between 1 and 10. As a side note, try not to plot counts on a log scale since counts of 0 are undefined and counts of 1 have a value of 0 (no height).

Transformed the long tail data to better understand the distribution of price. The tranformed price distribution appears bimodal with the price peaking around 800 or so and again at 5000 or so. Why is there a gap at 1500? Are there really no diamonds with that price? I wonder what this plot looks like across the categorical variables of cut, color, and clarity.



Tip: You can change the height and width of plots in code chunks with the `fig.height` and `fig.width` parameters in the chunk options.

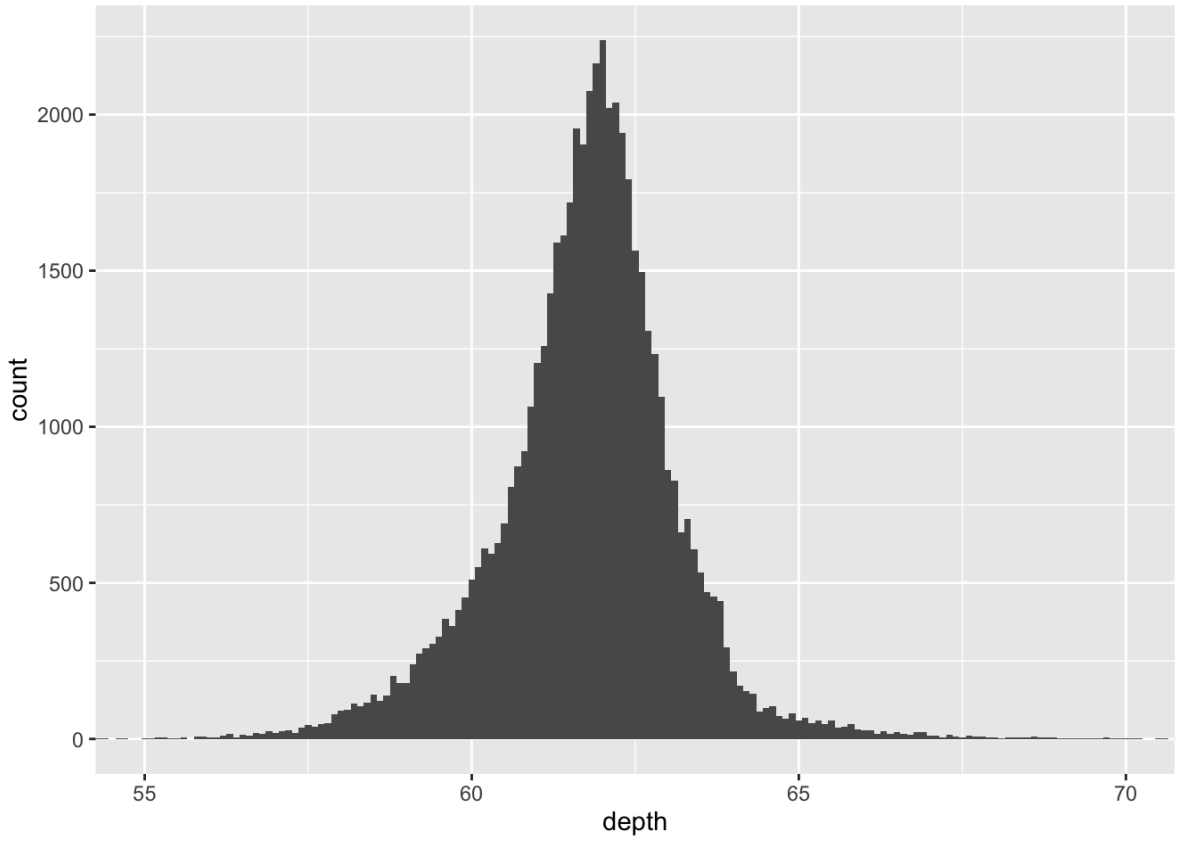
Most diamonds are of ideal cut, with gradually fewer diamonds of lesser-quality cut. A majority of diamonds are of cut G or better (lower letters are of better color). Clarity is skewed to the right, with most diamonds of lower clarity VS2 or worse.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.4000  0.7000  0.7979  1.0400  5.0100
```

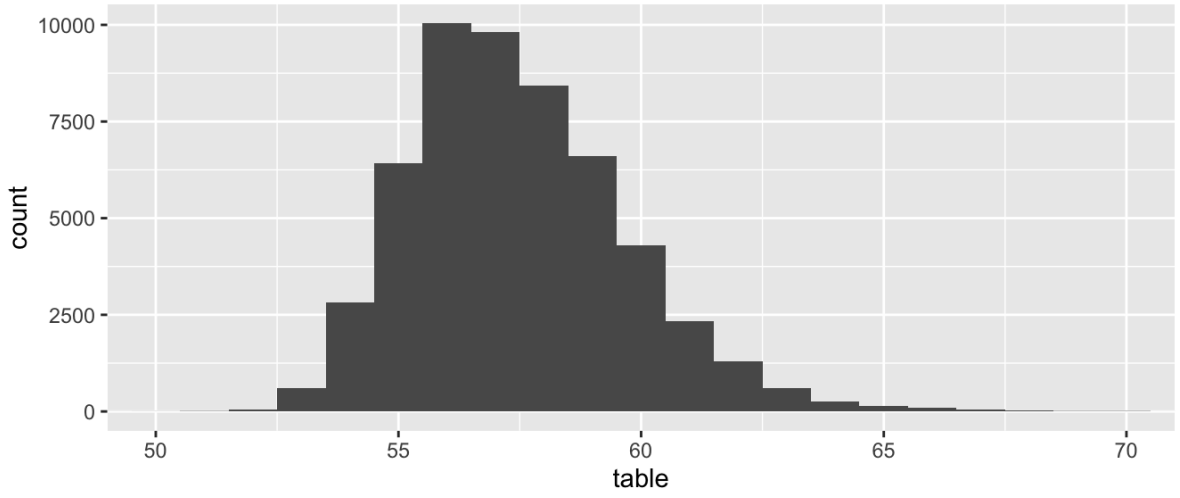
```
##
##  0.3 0.31 1.01  0.7 0.32    1  0.9 0.41  0.4 0.71  0.5 0.33 0.51 0.34 1.02
## 2604 2249 2242 1981 1840 1558 1485 1382 1299 1294 1258 1189 1127  910  883
## 0.52 1.51  1.5 0.72 0.53 0.42 0.38 0.35  1.2 0.54 0.36 0.91 1.03 0.55 0.56
##  817  807  793  764  709  706  670  667  645  625  572  570  523  496  492
```

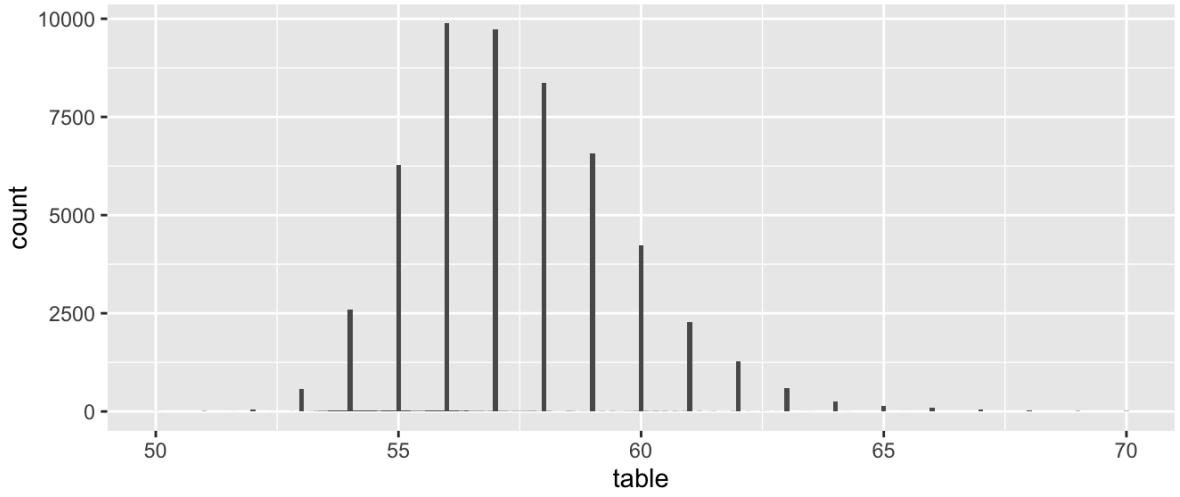
The lightest diamond is 0.2 carat and the heaviest diamond is 5.0100. Above, I plot the main body of carat weights, trimming the highest-carat diamonds. Some carat weights occur more often than other carat weights. Many of the most common carat counts end in x.x0 or x.x1. I wonder how carat is connected to price, and I wonder if the carat values are specific to certain cuts of diamonds.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   61.00   61.80   61.75   62.50   79.00
```

Most diamonds have a depth between 60 mm and 65 mm: median 61.8 mm and mean 61.75 mm.

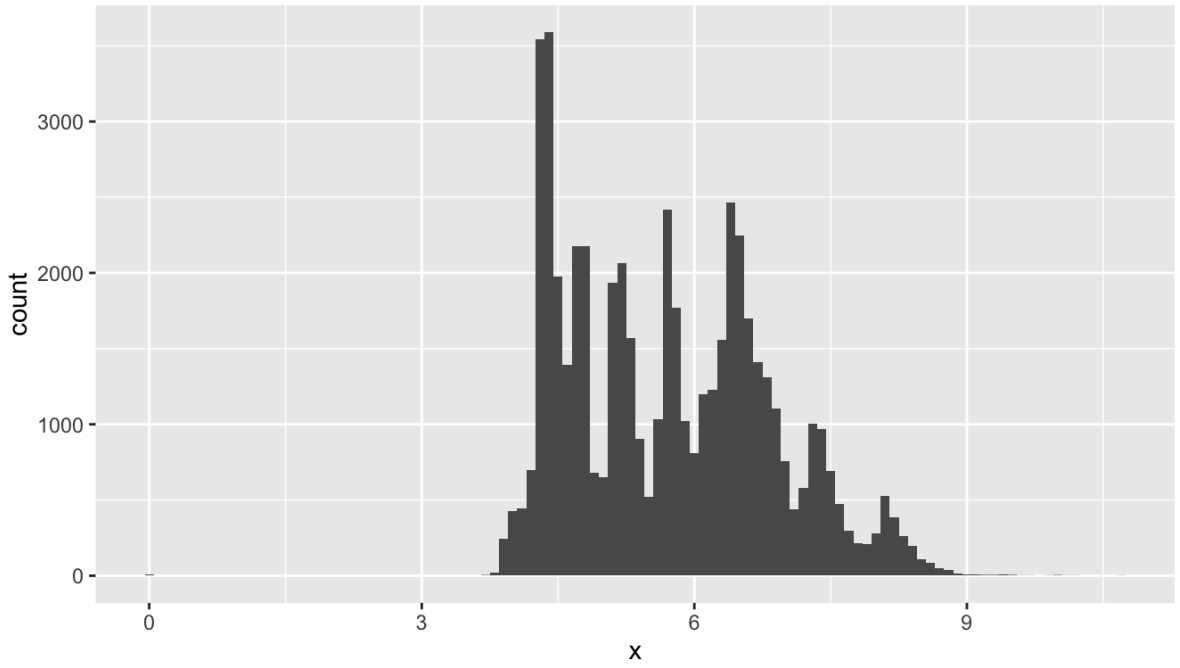


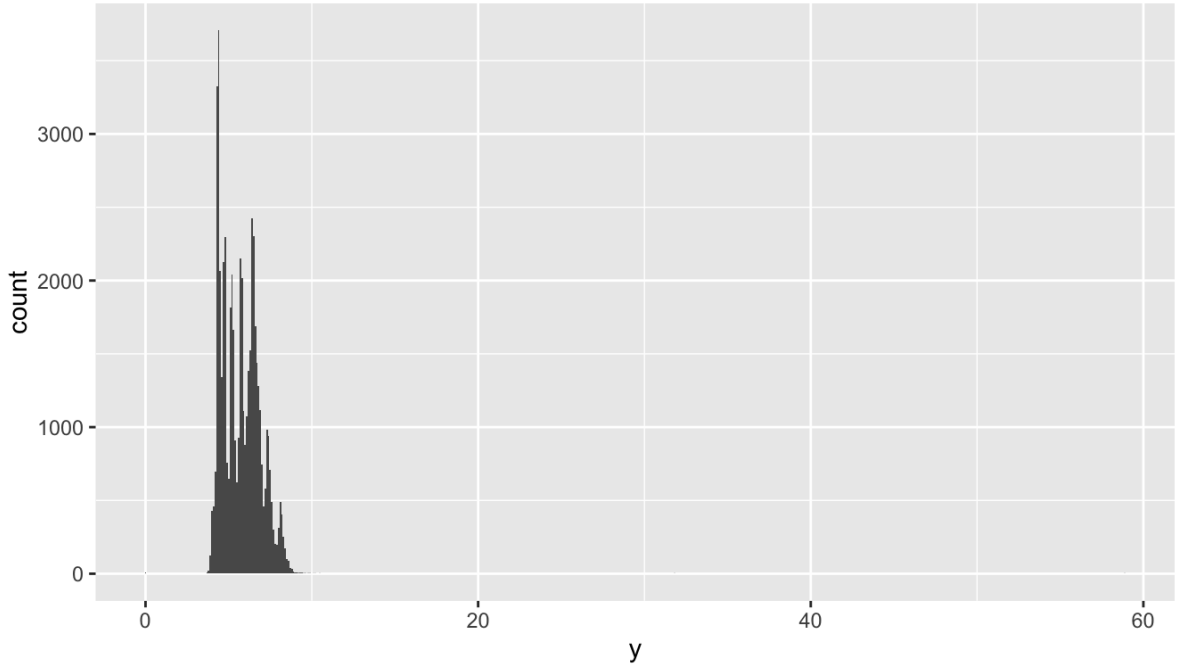


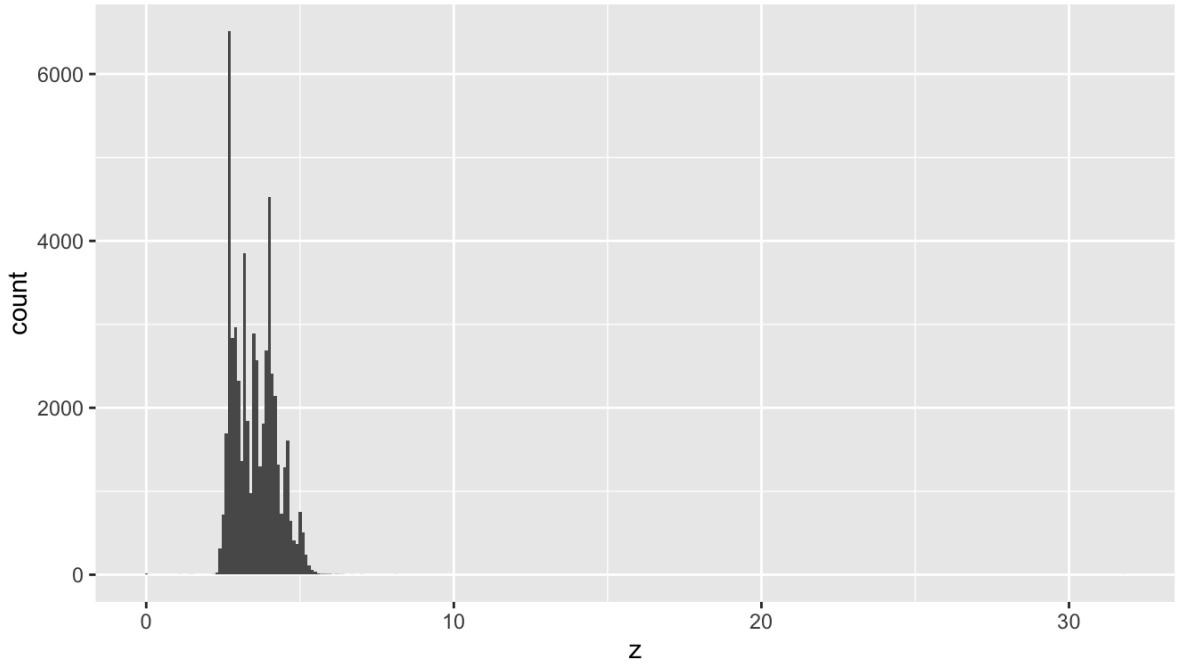
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   56.00   57.00   57.46   59.00   95.00
```

```
##
##   56   57   58   59   55   60   54   61   62   63   53   64   65   66   52
## 9881 9724 8369 6572 6268 4241 2594 2282 1273  588  567  260  146   91   56
```

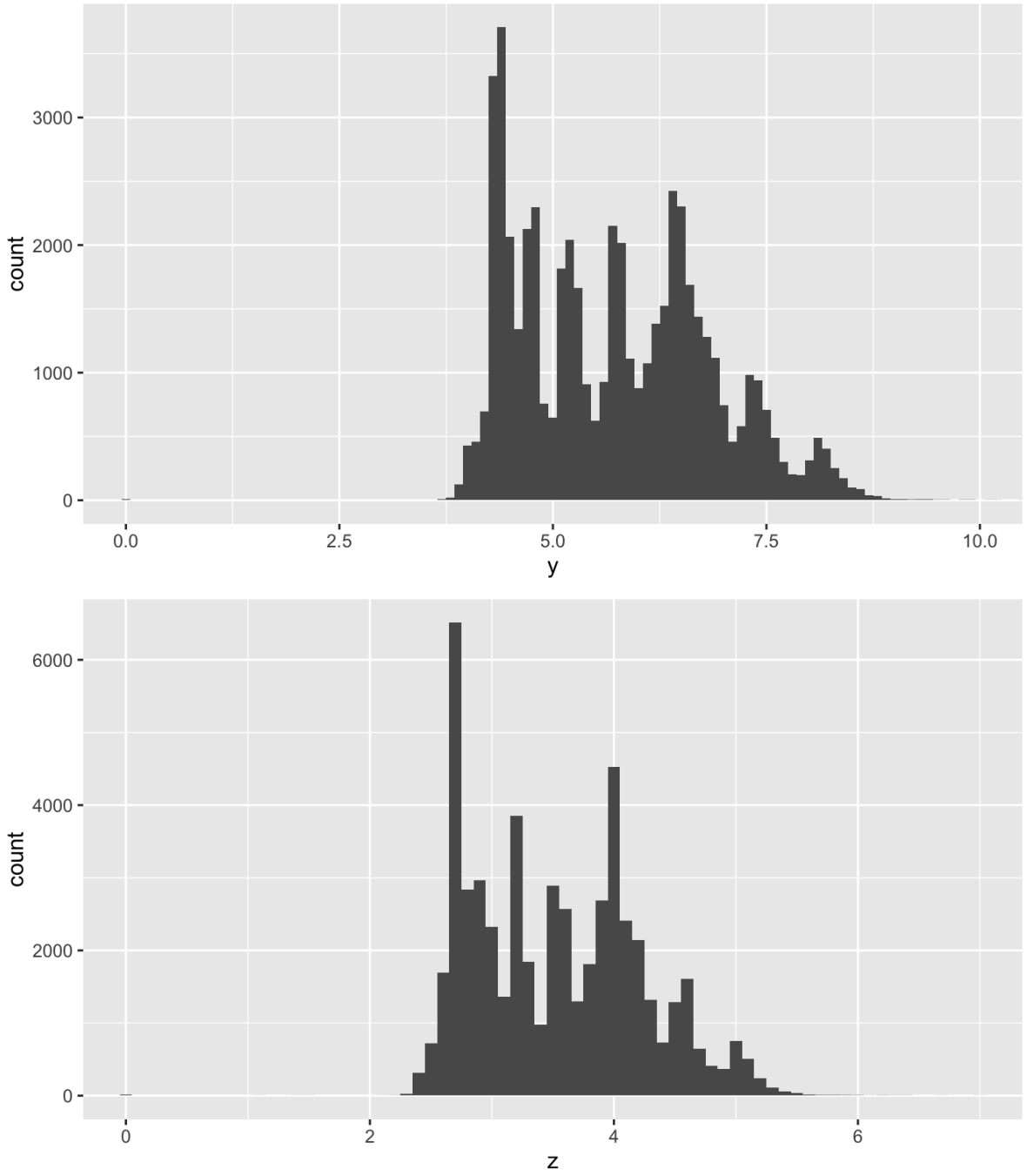
Setting the binwidth indicates that most table values are integers. Most diamonds have a table between 55 mm and 60 mm. Again, I wonder if this has anything to do with the cut of a diamond. Cut is a quality of a diamond that may influence carat weight and is responsible for making a diamond sparkle. There's likely to be strong relationships among carat, table, cut, and price.


count
3000
2000
1000
0
0
3
6
9
x


count
3000
2000
1000
0
0
20
40
60
y


count
6000
4000
2000
0
0
10
20
30
z

Most diamonds have an x dimension between 4 mm and 7 mm, a y dimension between 4 mm and 7 mm, and a z dimension between 2 mm and 6 mm. The y- and z- plots have a few high outliers so let's zoom in.



Zooming in, we see that there are a few conspicuous points at value 0 in each of the three x, y, and z plots. Let's investigate this further by finding these diamonds.

```
##
## FALSE  TRUE
## 53932     8
```

```
##
## FALSE  TRUE
## 53933     7
```

```
##
## FALSE  TRUE
## 53920    20
```

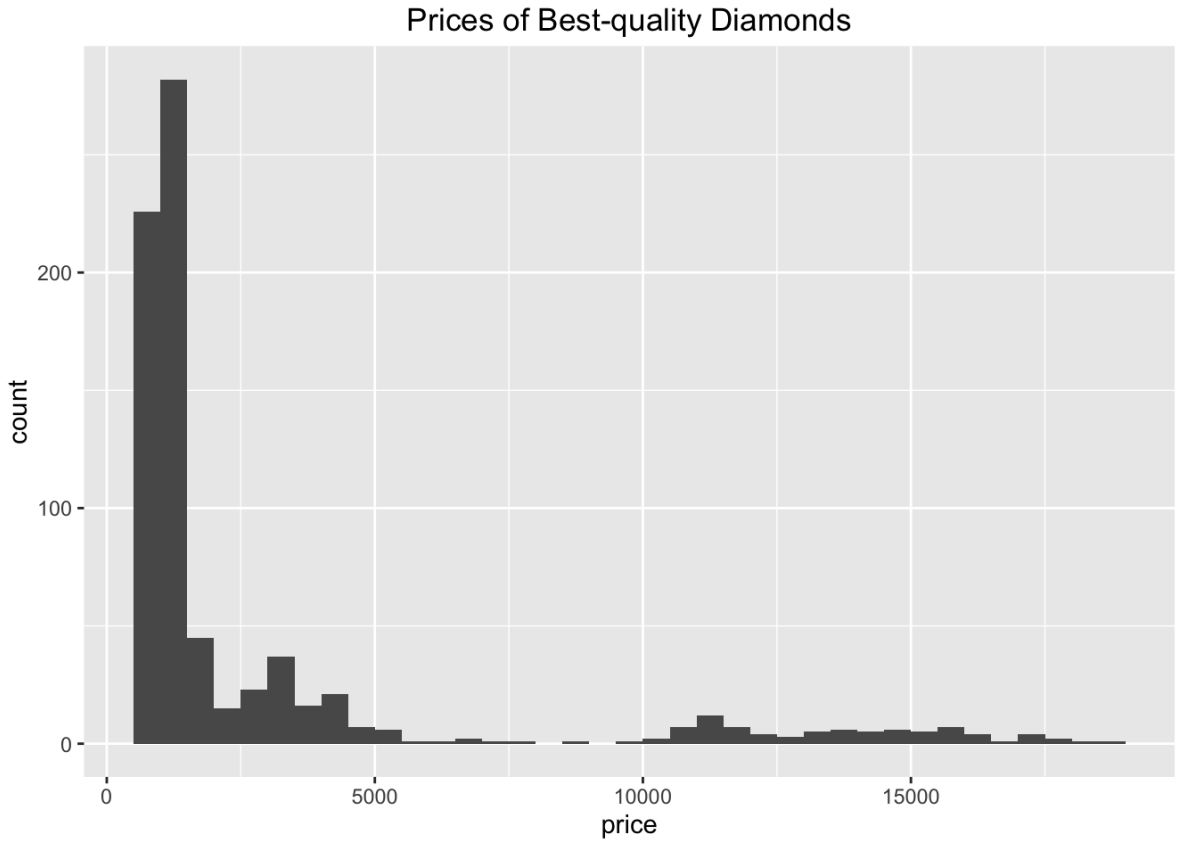
There are eight diamonds with missing x values, seven diamonds with missing y values, and twenty diamonds with missing z values.

```
## Source: local data frame [20 x 10]
##
##    carat       cut  color clarity depth table price     x     y     z
##    (dbl)    (fctr) (fctr)  (fctr) (dbl) (dbl) (int) (dbl) (dbl) (dbl)
## 1   1.00   Premium      G     SI2  59.1    59  3142  6.55  6.48     0
## 2   1.01   Premium      H      I1  58.1    59  3167  6.66  6.60     0
## 3   1.10   Premium      G     SI2  63.0    59  3696  6.50  6.47     0
## 4   1.01   Premium      F     SI2  59.2    58  3837  6.50  6.47     0
## 5   1.50      Good      G      I1  64.0    61  4731  7.15  7.04     0
## 6   1.07     Ideal      F     SI2  61.6    56  4954  0.00  6.62     0
## 7   1.00 Very Good      H     VS2  63.3    53  5139  0.00  0.00     0
## 8   1.15     Ideal      G     VS2  59.2    56  5564  6.88  6.83     0
## 9   1.14      Fair      G     VS1  57.5    67  6381  0.00  0.00     0
## 10  2.18   Premium      H     SI2  59.4    61 12631  8.49  8.45     0
## 11  1.56     Ideal      G     VS2  62.2    54 12800  0.00  0.00     0
## 12  2.25   Premium      I     SI1  61.3    58 15397  8.52  8.42     0
## 13  1.20   Premium      D    VVS1  62.1    59 15686  0.00  0.00     0
## 14  2.20   Premium      H     SI1  61.2    59 17265  8.42  8.37     0
## 15  2.25   Premium      H     SI2  62.8    59 18034  0.00  0.00     0
## 16  2.02   Premium      H     VS2  62.7    53 18207  8.02  7.95     0
## 17  2.80      Good      G     SI2  63.8    58 18788  8.90  8.85     0
## 18  0.71      Good      F     SI2  64.1    60  2130  0.00  0.00     0
## 19  0.71      Good      F     SI2  64.1    60  2130  0.00  0.00     0
## 20  1.12   Premium      G      I1  60.4    59  2383  6.71  6.67     0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2130    3564    5352    8803   15470   18790
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     949    2401    3931    5323   18820
```

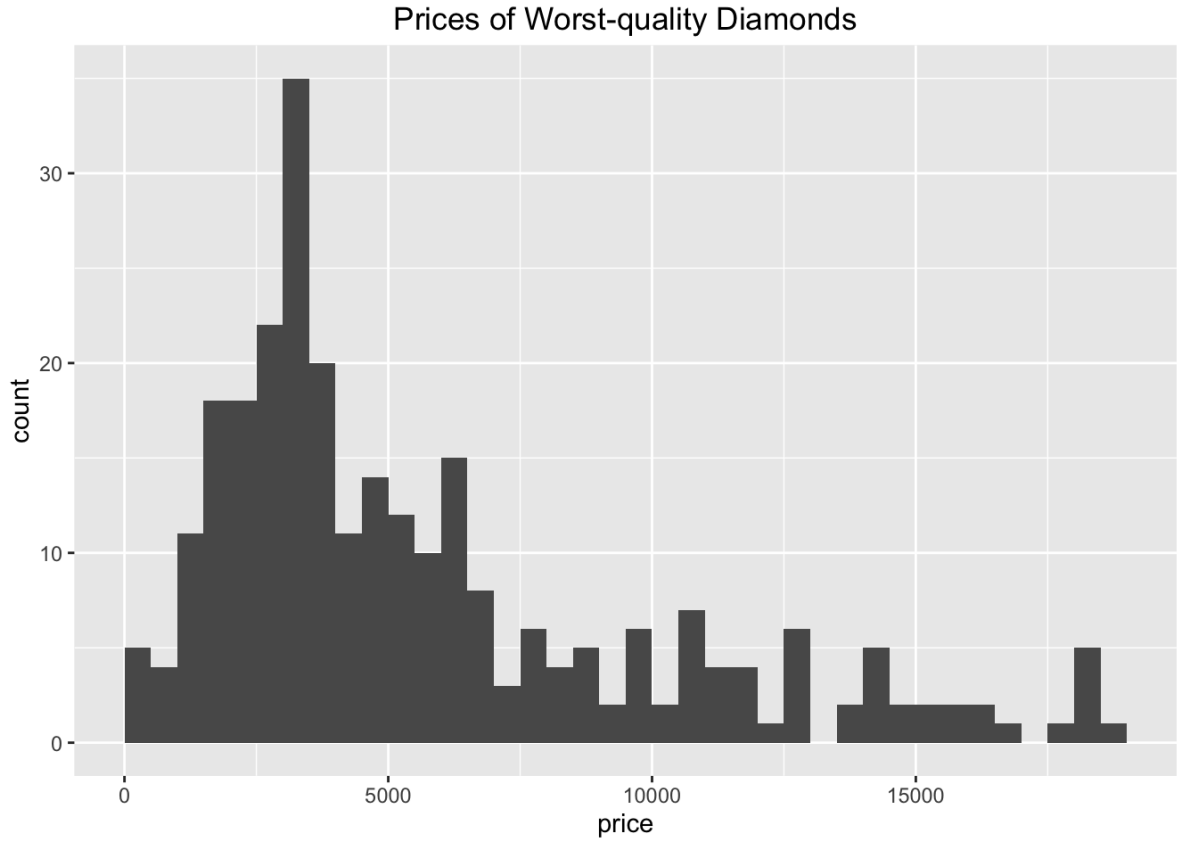
If and only if x or y dimensions are 0, then the z dimension is 0. Comparing the diamonds in this subset to all other diamonds, these diamonds tend to be very expensive or fall in the third quartile of the entire diamonds data set. Other variables such as carat, depth, table, and price are reported so I'll assume those values can be trusted.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     553     967    1207    2887    2644   18700
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2170    2983    3420    4712    5023   17080
```

Above, we subset the diamonds with high quality in color, clarity, and cut. Let's compare the prices (first summary) and prices per carat (second summary) to the diamonds with consistently low quality classes.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     335    2808    4306    5747    7563   18530
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1081    2638    3324    3579    4281    7437
```

There are a lot fewer diamonds which score low in all of color, clarity, and cut. The price per carat also seems to be significantly lower for the worst diamonds compared to the best diamonds, even if the regular price ranges are fairly similar. Later in my analysis, I'm going create density plots that are similar to the price histograms earlier to examine the price for each level of cut, color, and clarity.

What about the volume of a diamond? Does it have any relationships with price and other variables in the data set? I'm going to use a rough approximation of volume by using x * y * z to approximate a diamond as if it were a rectangular prism, basically a box.

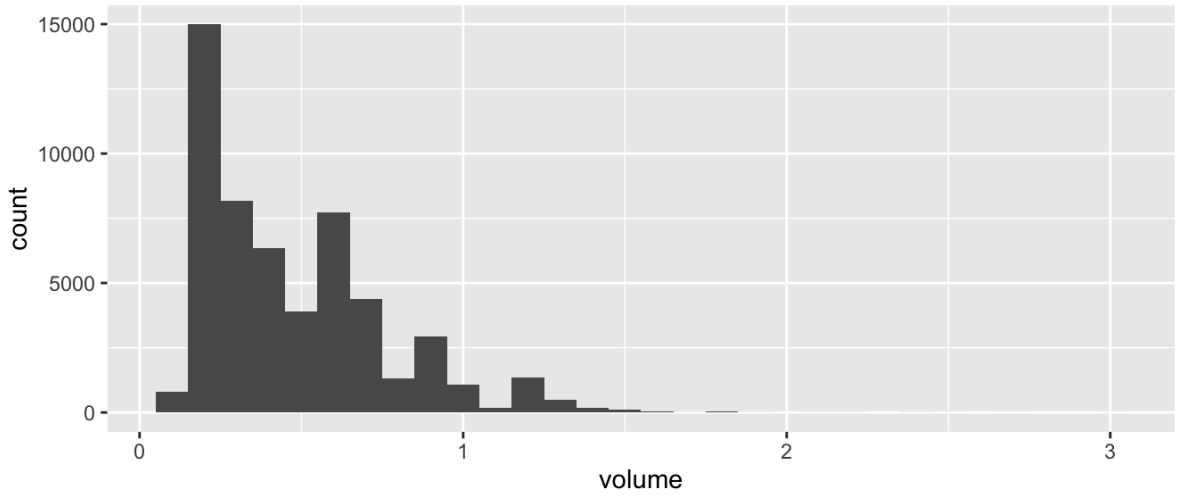
```
##
## FALSE  TRUE
## 53920    20
```

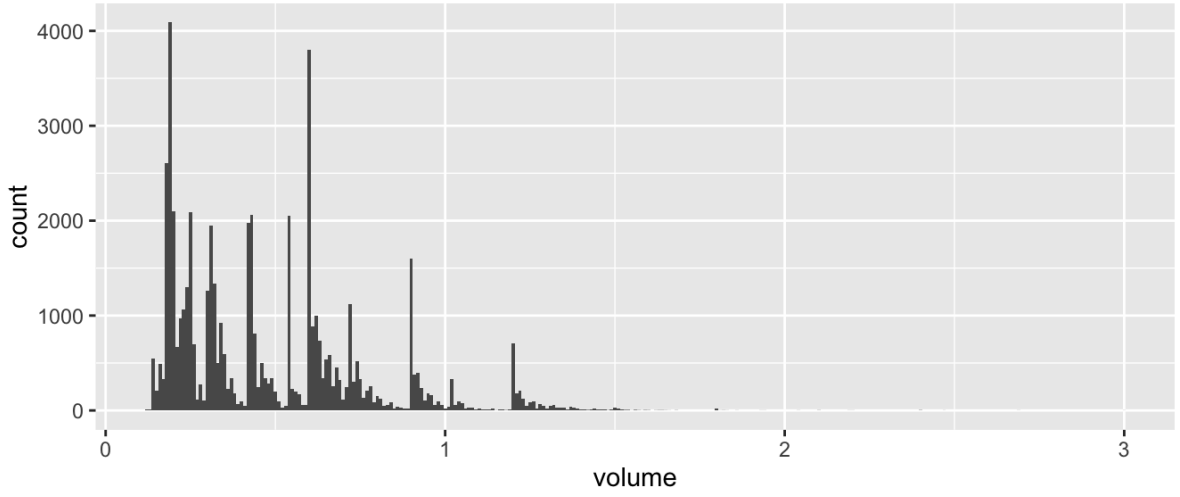
```
##       carat       cut color clarity depth table price    x    y z volume
## 2208   1.00   Premium     G     SI2  59.1    59  3142 6.55 6.48 0      0
## 2315   1.01   Premium     H      I1  58.1    59  3167 6.66 6.60 0      0
## 4792   1.10   Premium     G     SI2  63.0    59  3696 6.50 6.47 0      0
## 5472   1.01   Premium     F     SI2  59.2    58  3837 6.50 6.47 0      0
## 10168  1.50      Good     G      I1  64.0    61  4731 7.15 7.04 0      0
## 11183  1.07     Ideal     F     SI2  61.6    56  4954 0.00 6.62 0      0
## 11964  1.00 Very Good     H     VS2  63.3    53  5139 0.00 0.00 0      0
## 13602  1.15     Ideal     G     VS2  59.2    56  5564 6.88 6.83 0      0
## 15952  1.14      Fair     G     VS1  57.5    67  6381 0.00 0.00 0      0
## 24395  2.18   Premium     H     SI2  59.4    61 12631 8.49 8.45 0      0
## 24521  1.56     Ideal     G     VS2  62.2    54 12800 0.00 0.00 0      0
## 26124  2.25   Premium     I     SI1  61.3    58 15397 8.52 8.42 0      0
## 26244  1.20   Premium     D    VVS1  62.1    59 15686 0.00 0.00 0      0
## 27113  2.20   Premium     H     SI1  61.2    59 17265 8.42 8.37 0      0
## 27430  2.25   Premium     H     SI2  62.8    59 18034 0.00 0.00 0      0
## 27504  2.02   Premium     H     VS2  62.7    53 18207 8.02 7.95 0      0
## 27740  2.80      Good     G     SI2  63.8    58 18788 8.90 8.85 0      0
## 49557  0.71      Good     F     SI2  64.1    60  2130 0.00 0.00 0      0
## 49558  0.71      Good     F     SI2  64.1    60  2130 0.00 0.00 0      0
## 51507  1.12   Premium     G      I1  60.4    59  2383 6.71 6.67 0      0
```

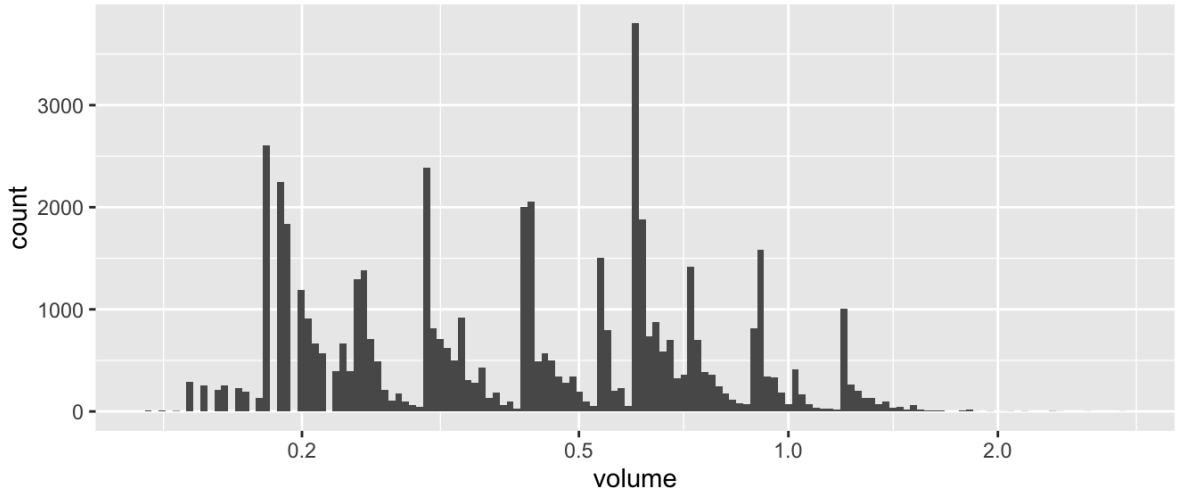
The twenty diamonds with at least one dimension with a value of 0 end up getting volumes equal to 0. Instead of using the dimensions x, y, and z to compute the volume, I now use the average density of diamonds to compute the volume instead. I can convert carat to grams and then divide by the density to get the volume of a diamond.

First, 1 carat is equivalent to 2 grams. Using Google, I found that diamond density is typically between 3.15 and 3.53 g/cm^3 with pure diamonds having a density close to 3.52 g/cm^3. I'm going to use the median density 3.34 g/cm^3 to estimate the volume of the diamonds.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1198  0.2395  0.4192  0.4778  0.6228  3.0000
```








```
##
##  0.18 0.186 0.605 0.419 0.192 0.599 0.539 0.246  0.24 0.425 0.299 0.198
##  2604  2249  2242  1981  1840  1558  1485  1382  1299  1294  1258  1189
## 0.305 0.204 0.611 0.311 0.904 0.898 0.431 0.317 0.251 0.228  0.21 0.719
##  1127   910   883   817   807   793   764   709   706   670   667   645
```

The histogram of volume is right skewed so I'm going to transform the data using a log transform. The histogram and count of most common values lines up with carat, since volume is a linear transformation of carat.

> **Tip**: Use the following section to summarize your observations during the univariate exploration of your dataset.

Univariate Analysis

What is the structure of your dataset?

There are 53,940 diamonds in the dataset with 10 features (carat, cut, color, clarity, depth, table, price, x, y, and z). The variables cut, color, and clarity, are ordered factor variables with the following levels.

(worst) ——————-> (best)
cut: Fair, Good, Very Good, Premium, Ideal
color: J, I, H, G, F, E, D
clarity: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF

Other observations:

- Most diamonds are of ideal cut.
- The median carat size is 0.7.
- Most diamonds have a color of G or better.
- About 75% of diamonds have carat weights less than 1.
- The median price for a diamonds $2401 and the max price is $18,823.

What is/are the main feature(s) of interest in your dataset?

The main features in the data set are carat and price. I'd like to determine which features are best for predicting the price of a diamond. I suspect carat and some combination of the other variables can be used to build a predictive model to price diamonds.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

Carat, color, cut, clarity, depth, and table likely contribute to the price of a diamond. I think carat (the weight of a diamond) and clarity probably contribute most to the price after researching information on diamond prices.

Did you create any new variables from existing variables in the dataset?

I created a variable for the volume of diamonds using the density of diamonds and the carat weight of diamonds. This arose in the bivariate section of my analysis when I explored how the price of a diamond varied with its volume. At first volume was calculated by multiplying the dimensions x, y, and z together. However, the volume was a crude approximation since the diamonds were assumed to be rectangular prisms in the initial calculation.

To better approximate the volume, I used the average density of diamonds. 1 carat is equivalent to 2 grams, and the average diamond density is between 3.15 and 3.53 g/cm^3 with pure diamonds having a density close to 3.52 g/cm^3. I used an average density of 3.34 g/cm^3 to estimate the volume of the diamonds.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

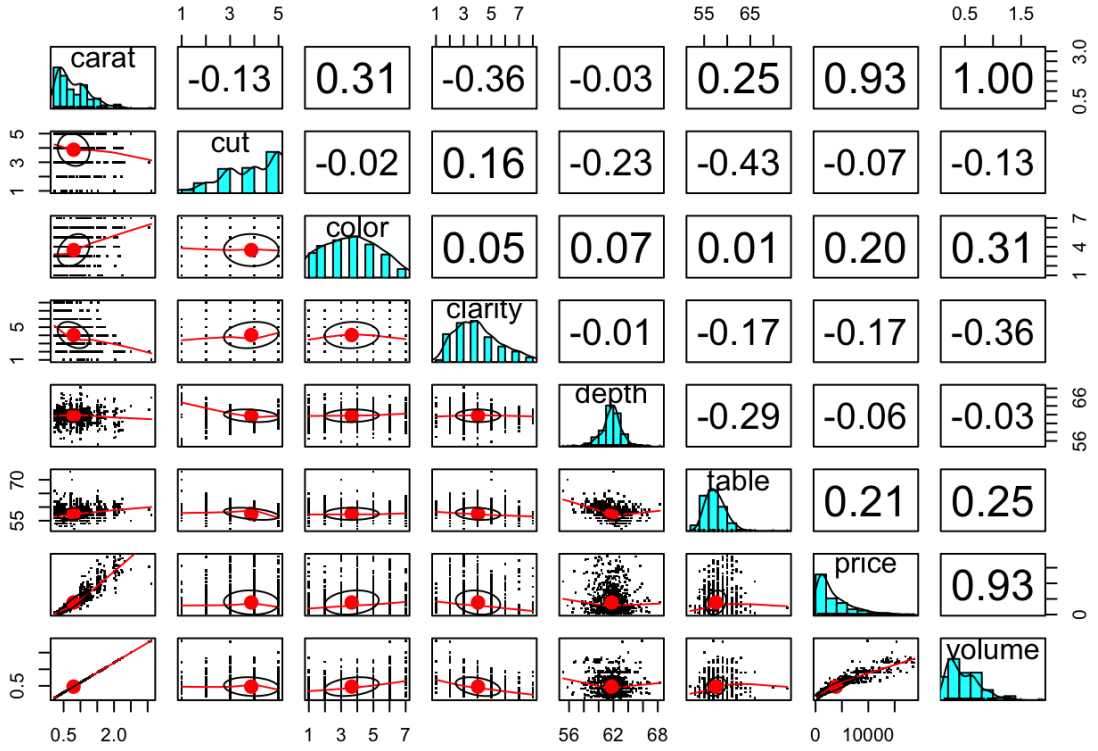
I log-transformed the right skewed price and volume distributions. The tranformed distribution for price appears bimodal with the price peaking around $800 or so and again around $5000. There's no diamonds priced at $1500.

When first calculating the volume using x, y, and z, some volumes were 0 or could not be calculated because data was missing. Additionally, some values for the dimensions x, y, and z seemed too large. In the subset called noVolume, all dimensions (x, y, and z) are missing or the z value is 0. The diamonds in this subset tend to be very expensive or fall in the third quartile of the entire diamonds data set.

Bivariate Plots Section

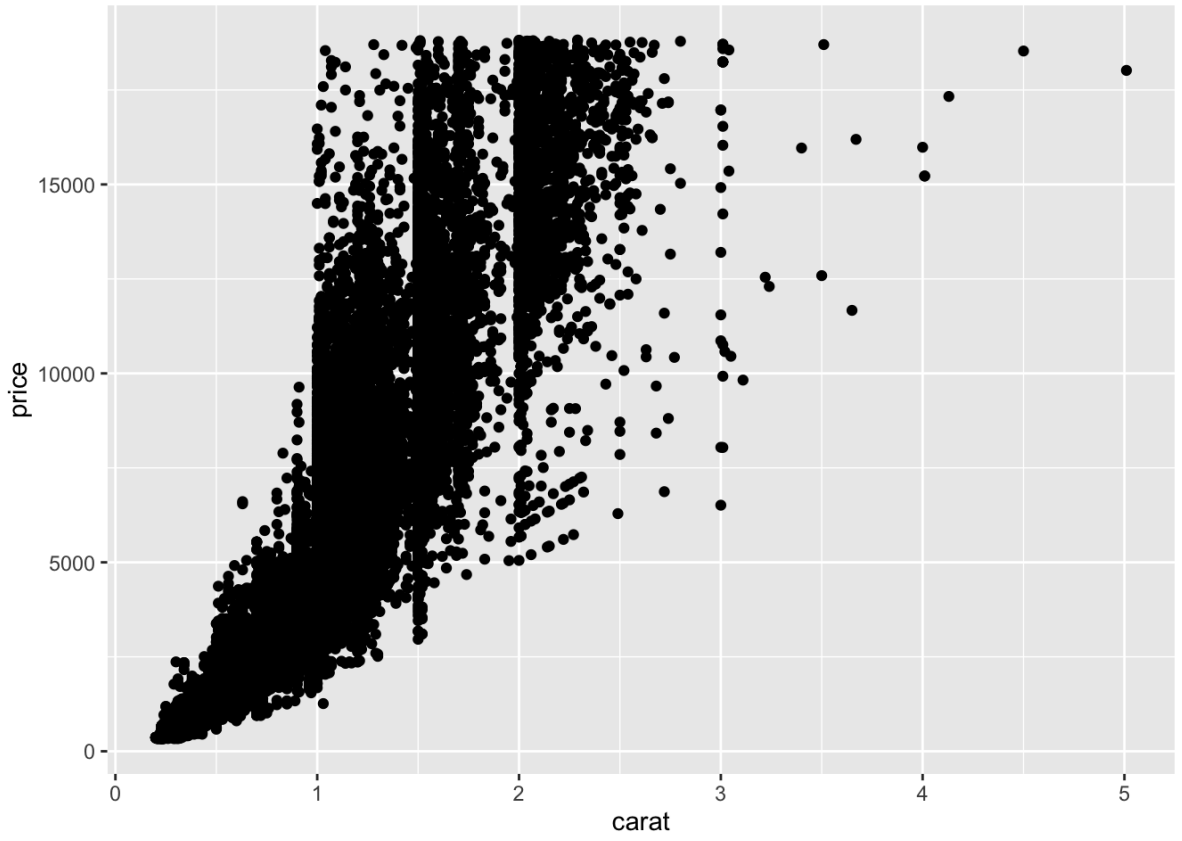
```
##         carat  depth  table  price      x      y     z volume
## carat   1.000  0.028  0.182  0.922  0.975  0.952 0.953  1.000
## depth   0.028  1.000 -0.296 -0.011 -0.025 -0.029 0.095  0.028
## table   0.182 -0.296  1.000  0.127  0.195  0.184 0.151  0.182
## price   0.922 -0.011  0.127  1.000  0.884  0.865 0.861  0.922
## x       0.975 -0.025  0.195  0.884  1.000  0.975 0.971  0.975
## y       0.952 -0.029  0.184  0.865  0.975  1.000 0.952  0.952
## z       0.953  0.095  0.151  0.861  0.971  0.952 1.000  0.953
## volume  1.000  0.028  0.182  0.922  0.975  0.952 0.953  1.000
```

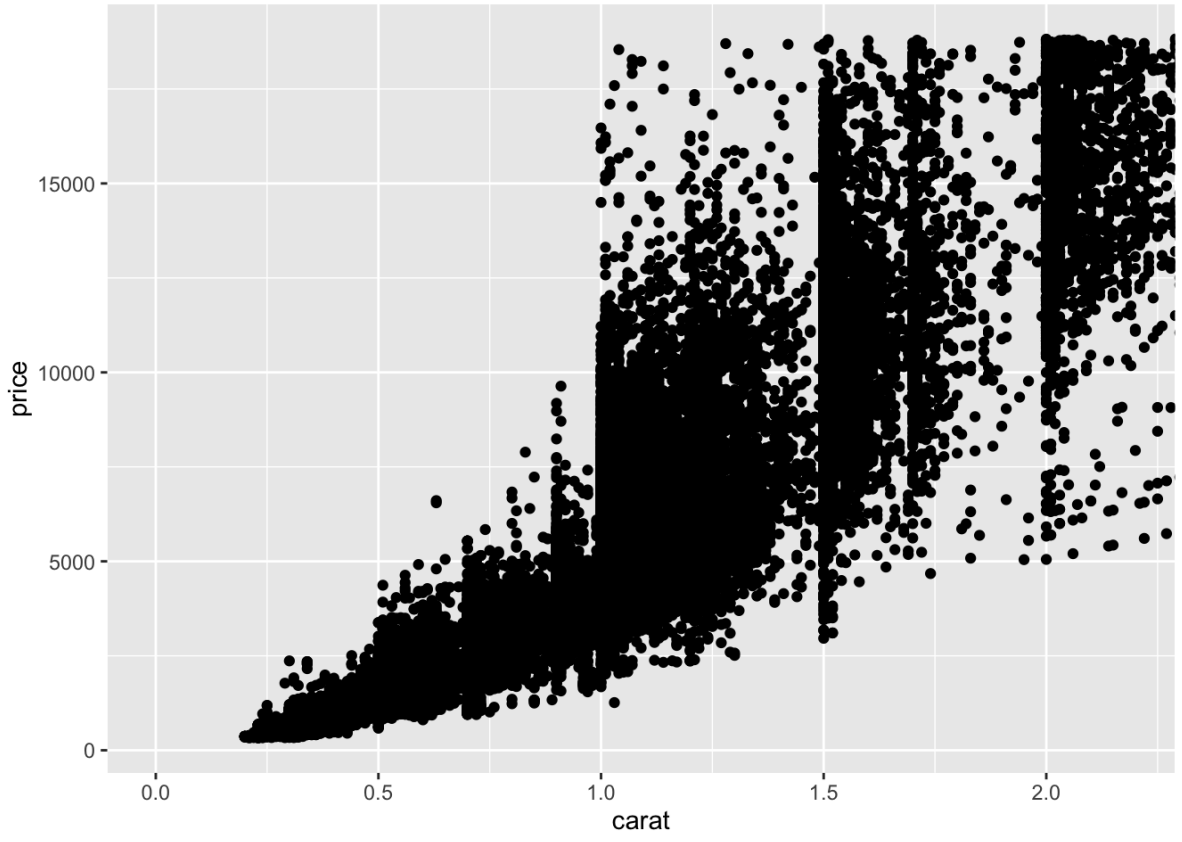
The dimensions of a diamond tend to correlate with each other. The longer one dimension, then the larger the diamond is overall. The dimensions also correlate with carat weight which makes sense. Price correlates strongly with carat weight and the three dimensions (x, y, z).



Tip: Be mindful of the number of data points and variables that you put in a correlation matrix or plot matrix: you do not need to include all variables. In addition, you can use other packages not introduced in the associated course to conduct your exploration. Make sure you load them at the beginning of your document so that it is easiest to see which packages are necessary. (The above plot matrix comes from the `psych` package.)

From a subset of the data, cut, color and clarity do not seem to have strong correlations with price, but color and clarity are moderately correlated with carat. I want to look closer at scatter plots involving price and some other variables like carat, depth, and table.

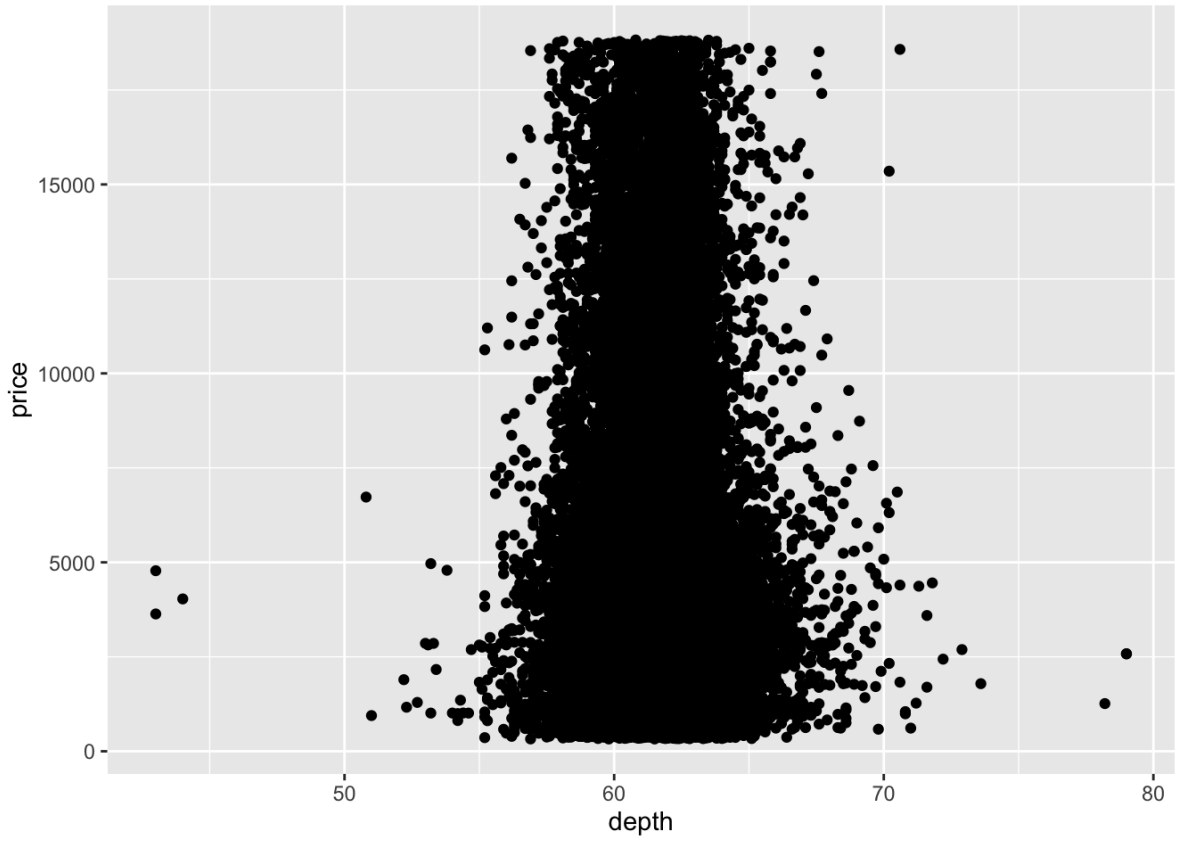


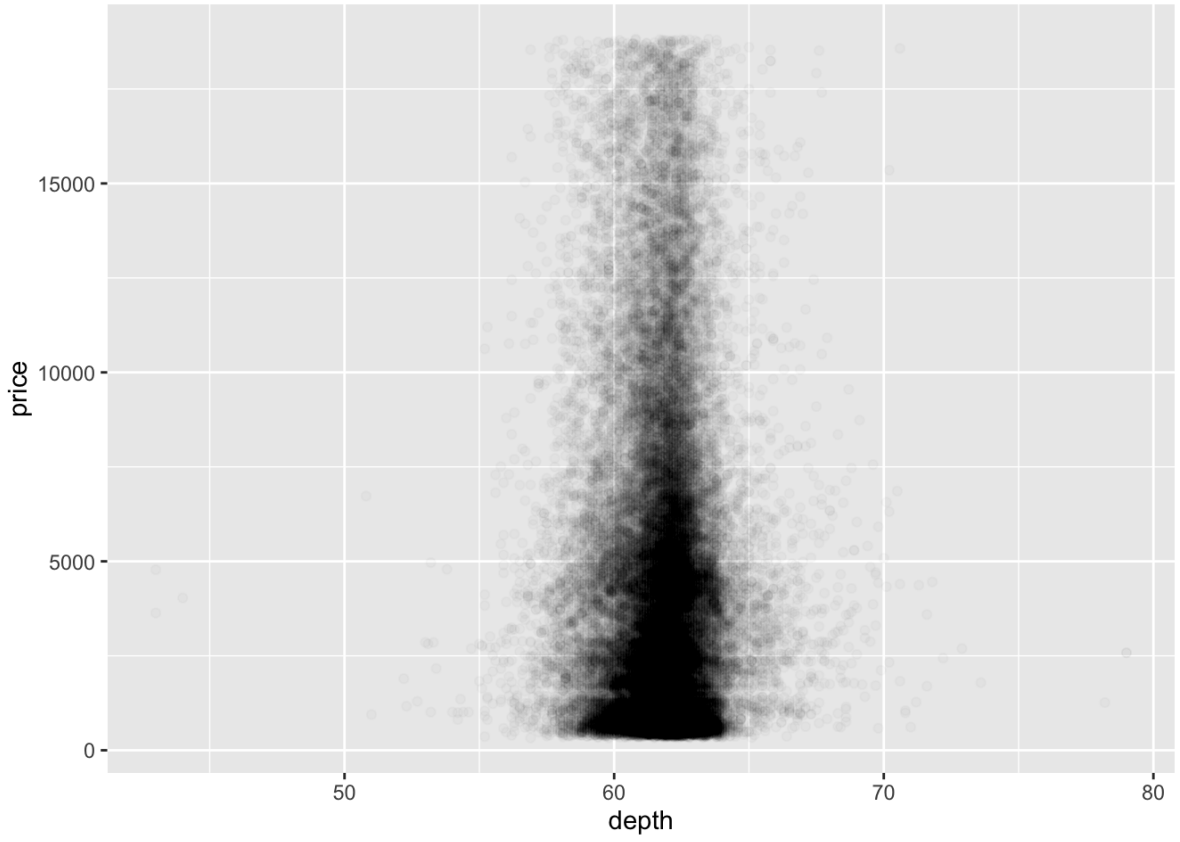


As carat size increases, the variance in price increases. We still see vertical bands where many diamonds take on the same carat value at different price points. The relationship between price and carat appears to be exponential rather than linear.

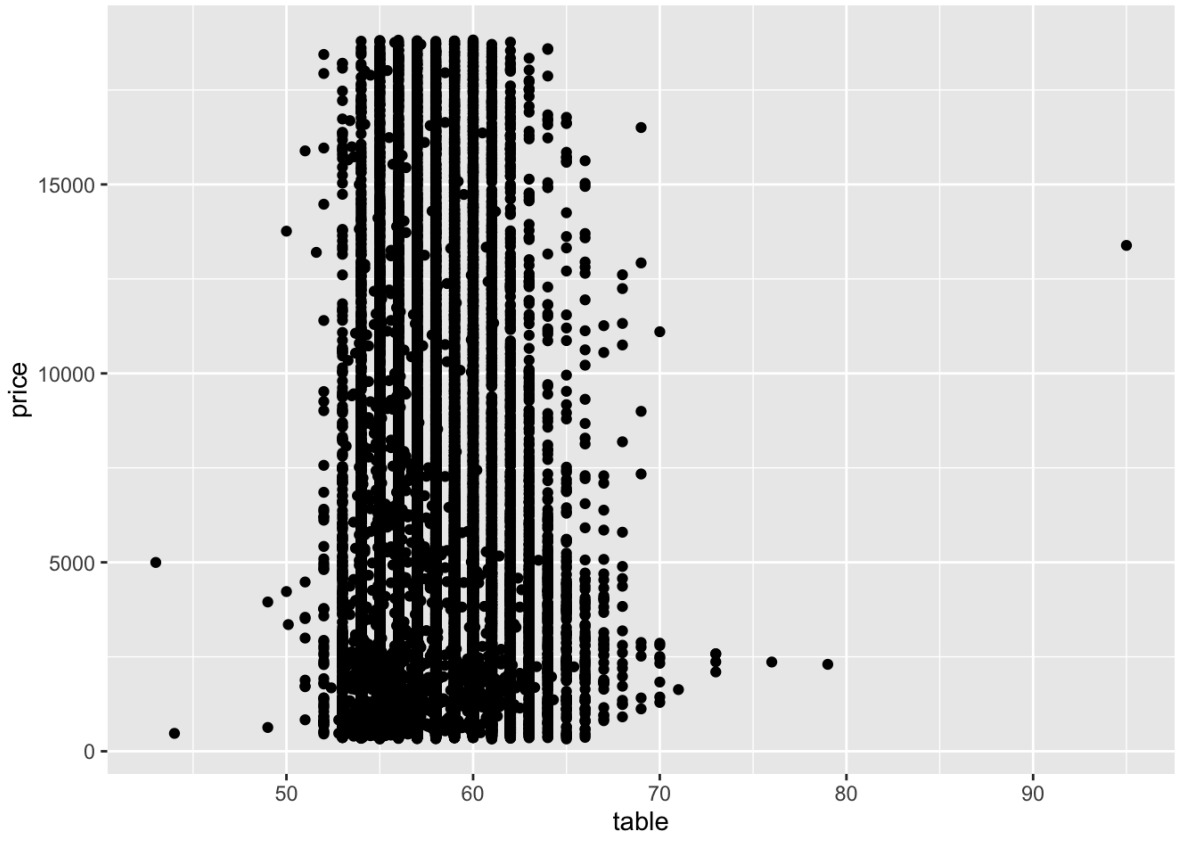
```
##
## Call:
## lm(formula = price ~ carat, data = subset(diamonds, carat <=
##     quantile(diamonds$carat, 0.999)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10922.6   -818.3     -8.3    566.5  12703.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2317.86      12.94  -179.1   <2e-16 ***
## carat        7843.16      14.02   559.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1524 on 53885 degrees of freedom
## Multiple R-squared:  0.8532, Adjusted R-squared:  0.8532
## F-statistic: 3.131e+05 on 1 and 53885 DF,  p-value: < 2.2e-16
```

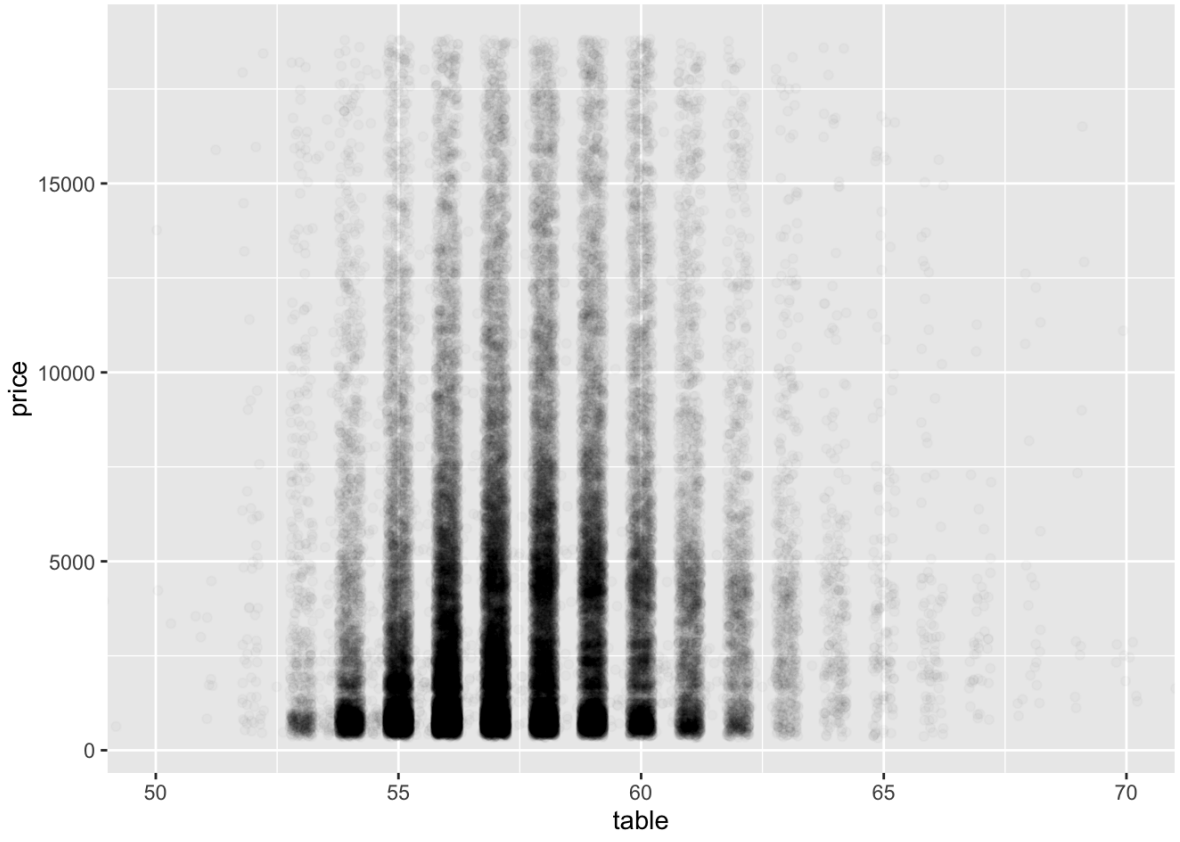
Despite the fact that the relationship looks nonlinear, based on the R^2 value, carat still explains about 85 percent of the variance in price.





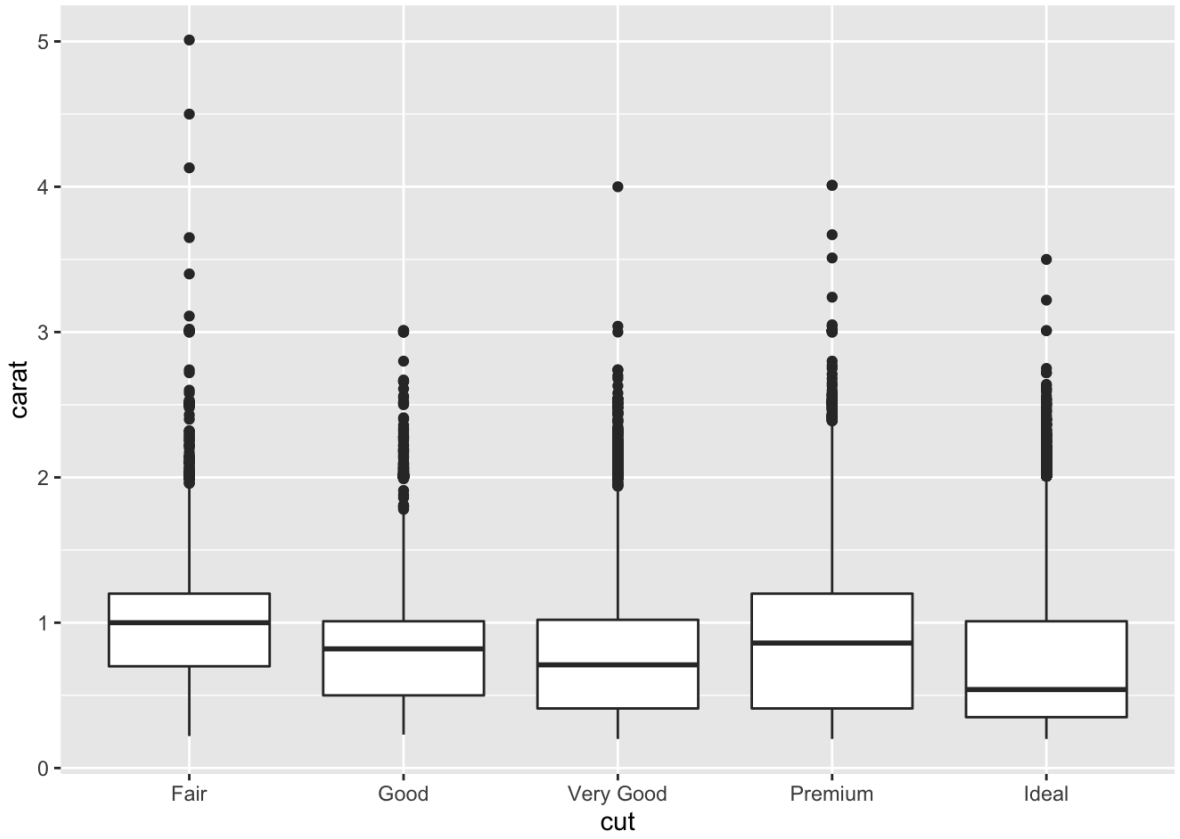
Comparing depth to price, the first plot suffers from some overplotting. Most diamonds have a depth between 60 and 65 (no units), and the lack of correlation seen in the earlier table is easy to see here.





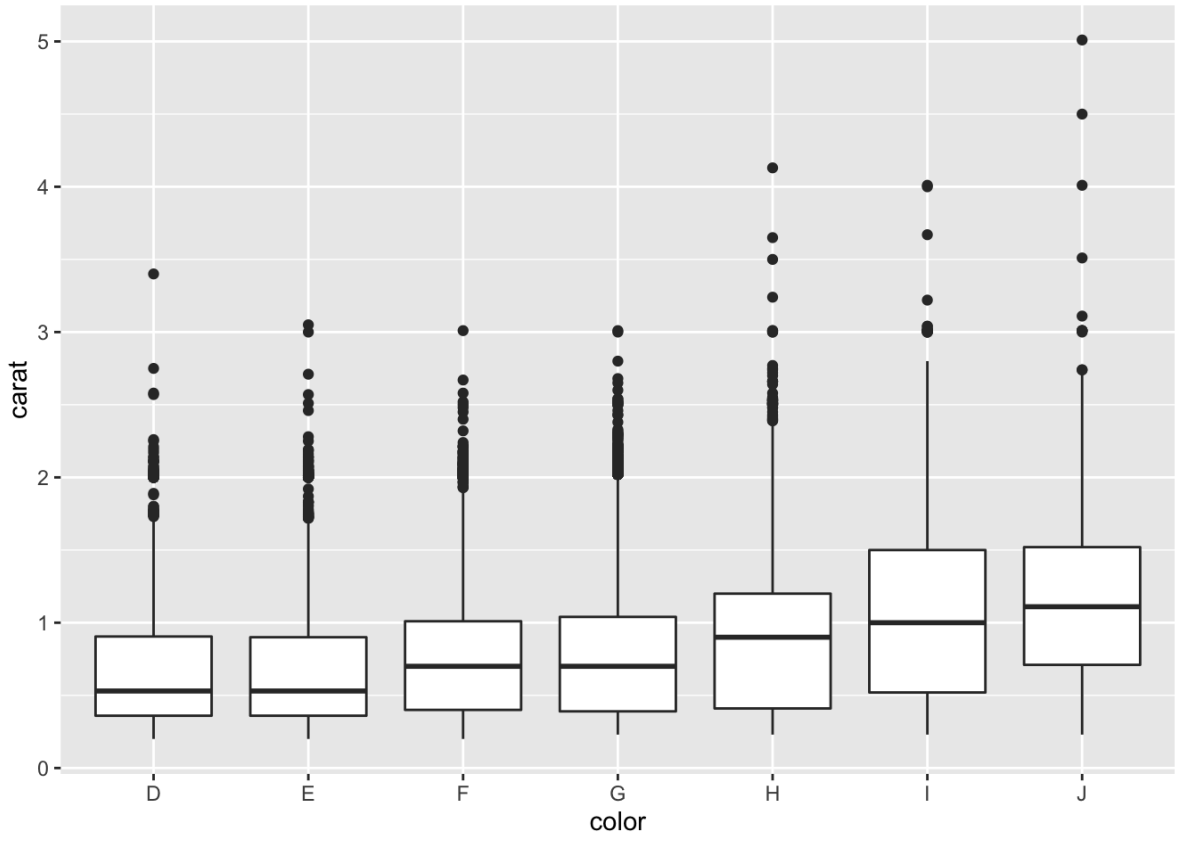
Again, the tall vertical strips indicate table values are mostly integers. Adding jitter, transparency, and changing the plot limits lets us see the slight correlation between table and price.

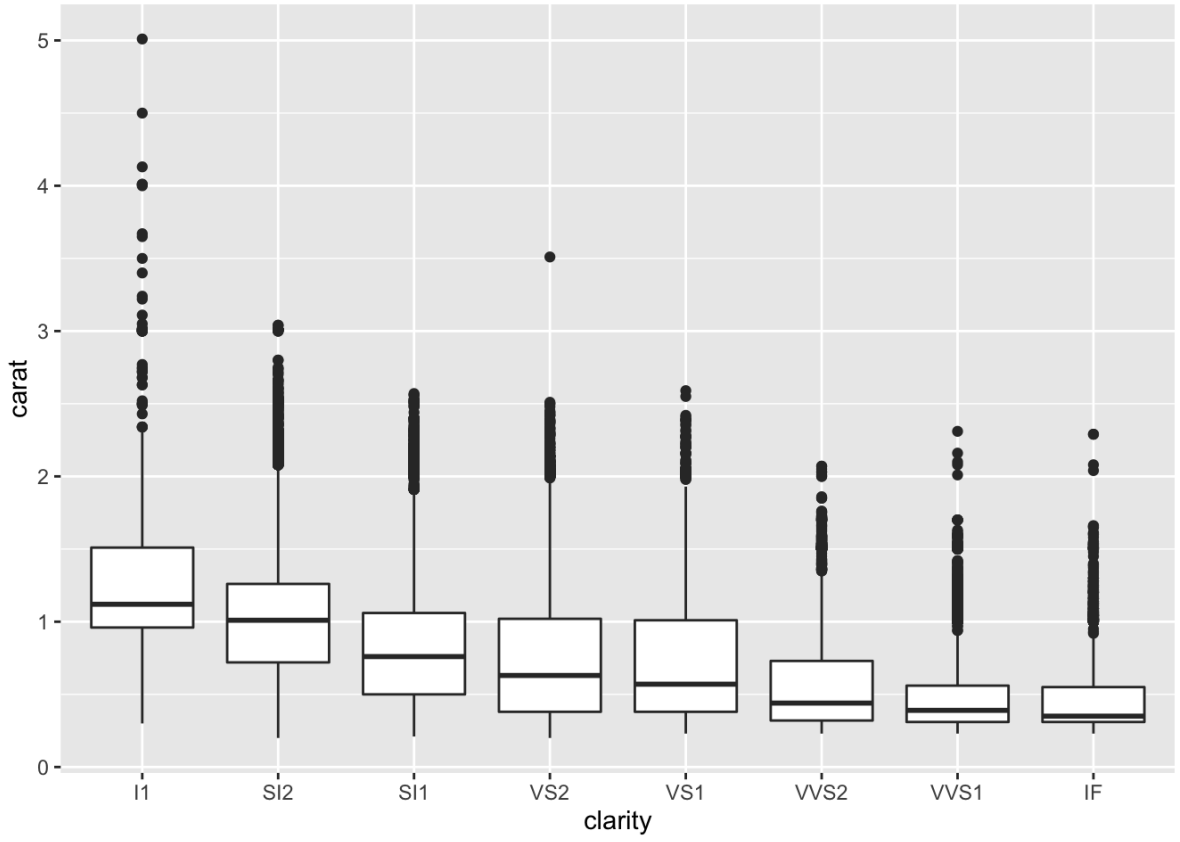
Next, I'll look at how the categorical features vary with carat and price.



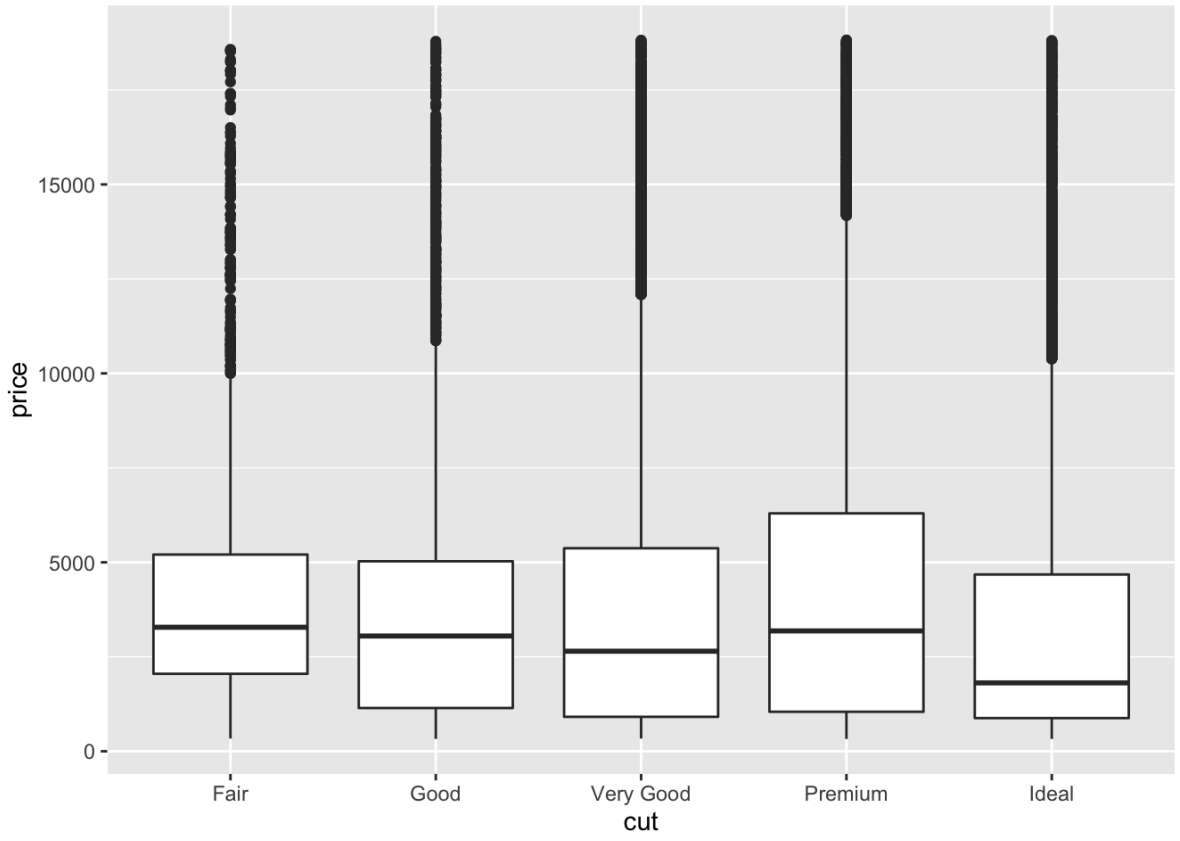

```
## cut: Fair
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.220   0.700   1.000   1.046   1.200   5.010
## ------------------------------------------------------------
## cut: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.5000  0.8200  0.8492  1.0100  3.0100
## ------------------------------------------------------------
## cut: Very Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.4100  0.7100  0.8064  1.0200  4.0000
## ------------------------------------------------------------
## cut: Premium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.200   0.410   0.860   0.892   1.200   4.010
## ------------------------------------------------------------
## cut: Ideal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.3500  0.5400  0.7028  1.0100  3.5000
```

It doesn't look like particular cuts have a certain number of carats. However, it looks like most of the ideal cut diamonds are on the smaller side, less than one carat.



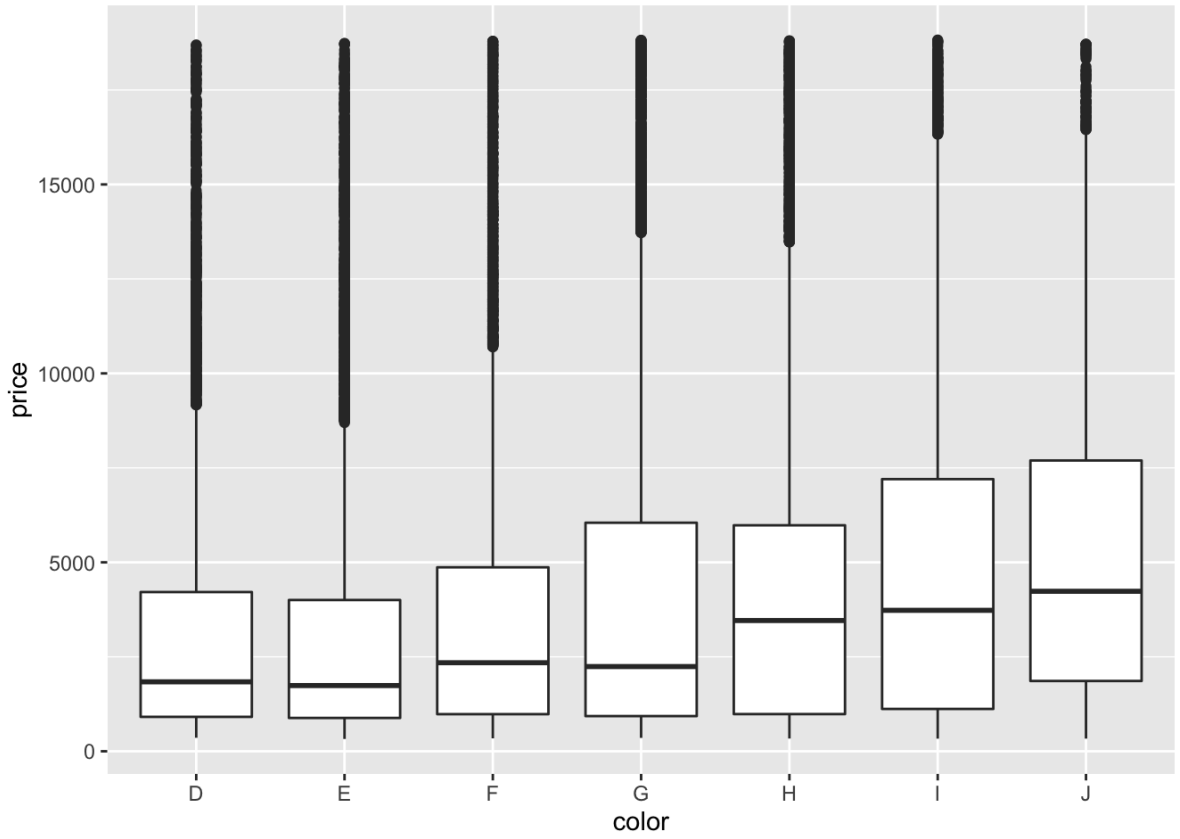


The trend between carat and color is clearer, with the worst-color diamonds (best color is D and the worst color is J) having the largest median and largest range. Clarity shows a similar trend, and most of the diamonds of 3 carats or larger fall into the worst clarity groups (I1, SI2).



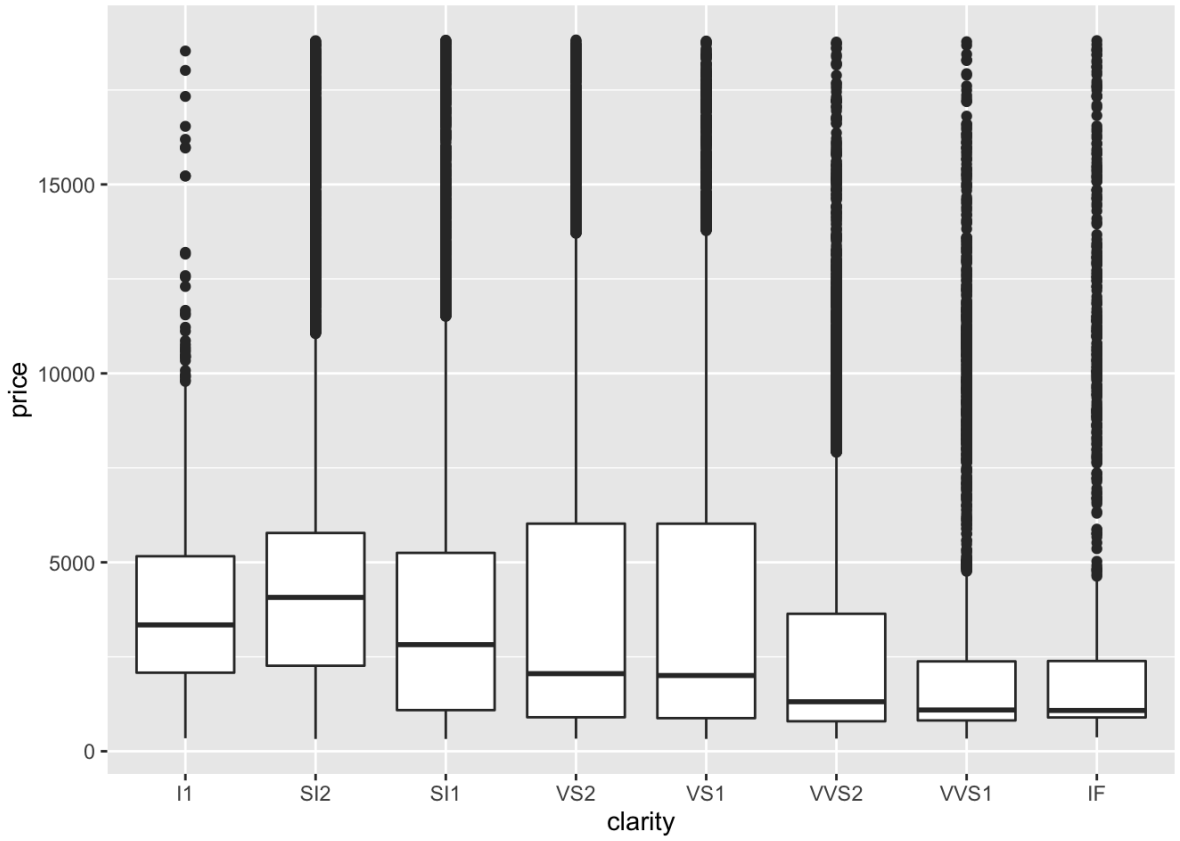

```
## diamonds$cut: Fair
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     337    2050    3282    4359    5206   18570
## ------------------------------------------------------------
## diamonds$cut: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     327    1145    3050    3929    5028   18790
## ------------------------------------------------------------
## diamonds$cut: Very Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     336     912    2648    3982    5373   18820
## ------------------------------------------------------------
## diamonds$cut: Premium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326    1046    3185    4584    6296   18820
## ------------------------------------------------------------
## diamonds$cut: Ideal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     878    1810    3458    4678   18810
```

Ideal diamonds have the lowest median price. This seems really unusual since I would expect diamonds with an ideal cut to have a higher median price compared to the other groups. There are many outliers. The variation in price tends to increase as cut improves and then decreases for diamonds with ideal cuts. What does price/carat look like for these cuts?




```
## diamonds$color: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     357     911    1838    3170    4214   18690
## ------------------------------------------------------------
## diamonds$color: E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     882    1739    3077    4003   18730
## ------------------------------------------------------------
## diamonds$color: F
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     342     982    2344    3725    4868   18790
## ------------------------------------------------------------
## diamonds$color: G
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     354     931    2242    3999    6048   18820
## ------------------------------------------------------------
## diamonds$color: H
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     337     984    3460    4487    5980   18800
## ------------------------------------------------------------
## diamonds$color: I
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334    1120    3730    5092    7202   18820
## ------------------------------------------------------------
## diamonds$color: J
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     335    1860    4234    5324    7695   18710
```

Here is another surprise. The lowest median price diamonds have a color of D, which is the best color in the data set. Price variance increases as the color decreases (best color is D and the worst color is J). The median price typically decreases as color improves. Now, I want to look at price per carat by color.




```
## diamonds$clarity: I1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     345    2080    3344    3924    5161   18530
## ------------------------------------------------------------
## diamonds$clarity: SI2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326    2264    4072    5063    5777   18800
## ------------------------------------------------------------
## diamonds$clarity: SI1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326    1089    2822    3996    5250   18820
## ------------------------------------------------------------
## diamonds$clarity: VS2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     334     900    2054    3925    6024   18820
## ------------------------------------------------------------
## diamonds$clarity: VS1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     327     876    2005    3839    6023   18800
## ------------------------------------------------------------
## diamonds$clarity: VVS2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   336.0   794.2  1311.0  3284.0  3638.0 18770.0
## ------------------------------------------------------------
## diamonds$clarity: VVS1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     336     816    1093    2523    2379   18780
## ------------------------------------------------------------
## diamonds$clarity: IF
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     369     895    1080    2865    2388   18810
```

Here again, there is a trend that goes against my intuition. The lowest median price occurs for the best clarity (IF). There also to be many more outliers for the better clarity diamonds. I'm not sure why great clarity diamonds are priced so low. Another trend to note here is that price variance increases then decreases significantly as the clarity improves.

I want to look at two things: price per carat, and the distribution of prices for diamonds with best levels of the categorical variables.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Price correlates strongly with carat weight and the three dimensions (x, y, z).

As carat size increases, the variance in price increases. In the plot of price vs carat, there are vertical bands where many diamonds take on the same carat value at different price points. The relationship between price and carat appears to be exponential rather than linear.

Based on the R^2 value, carat explains about 85 percent of the variance in price. Other features of interest can be incorporated into the model to explain the variance in the price.

Diamonds with better levels of clarity, cut, and color tend to occur more often at lower prices while diamonds with worse levels of clarity, cut, and color tend to occur more often at higher prices.

Ideal diamonds have the lowest median price. This seems really unusual since I would expect diamonds with an ideal cut to have a higher median price compared to the other groups. There are many outliers. The variation in price tends to increase as cut improves and then decreases for diamonds with ideal cuts.

The lowest median priced diamonds have a color of D, which is the best color in the data set. Price variance increases as the color decreases (best color is D and the worst color is J). The median price typically decreases as color improves.

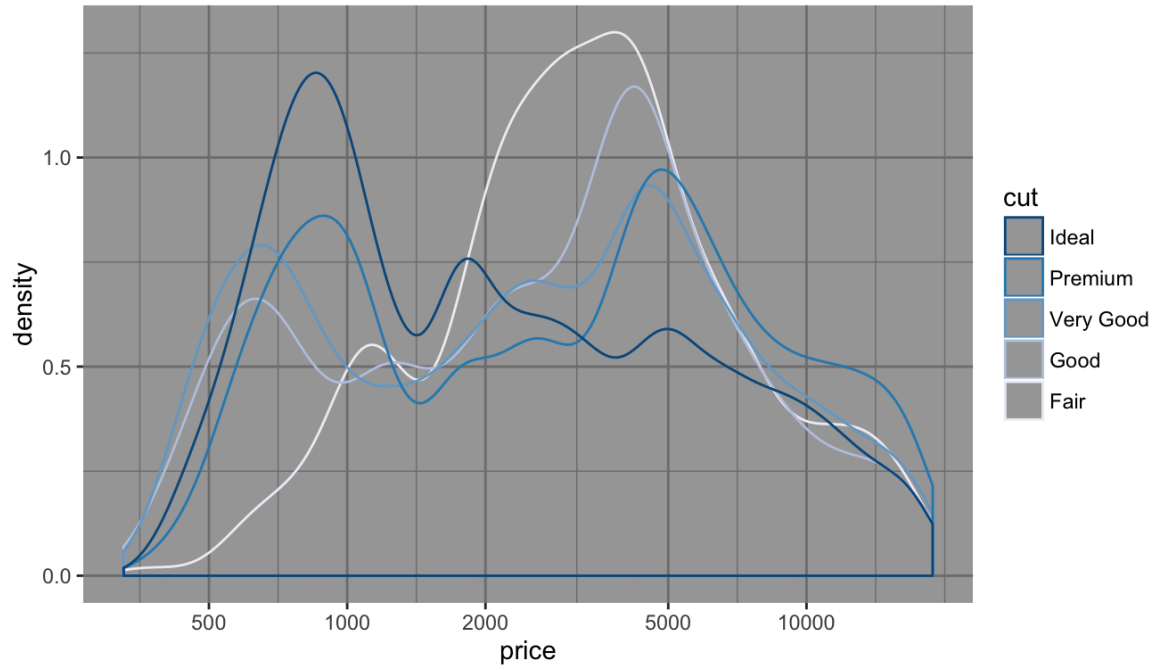
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

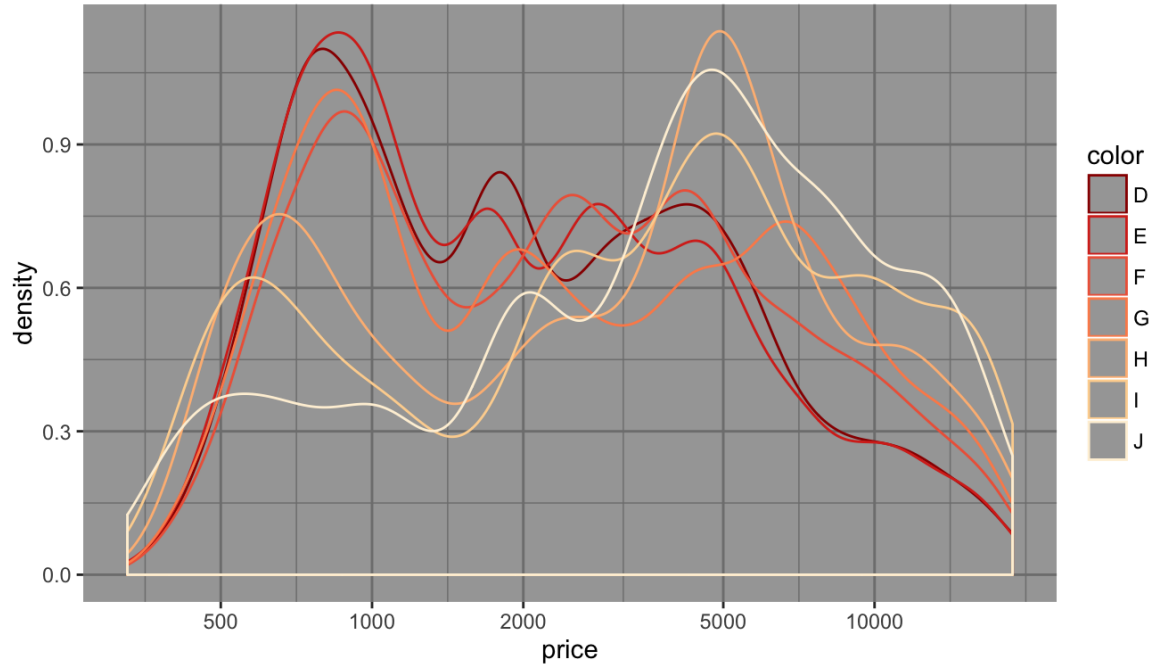
The dimensions of a diamond (x, y, and z) tend to correlate with each other. The longer one dimension, then the larger the diamond. The dimensions also correlate with carat weight which makes sense.

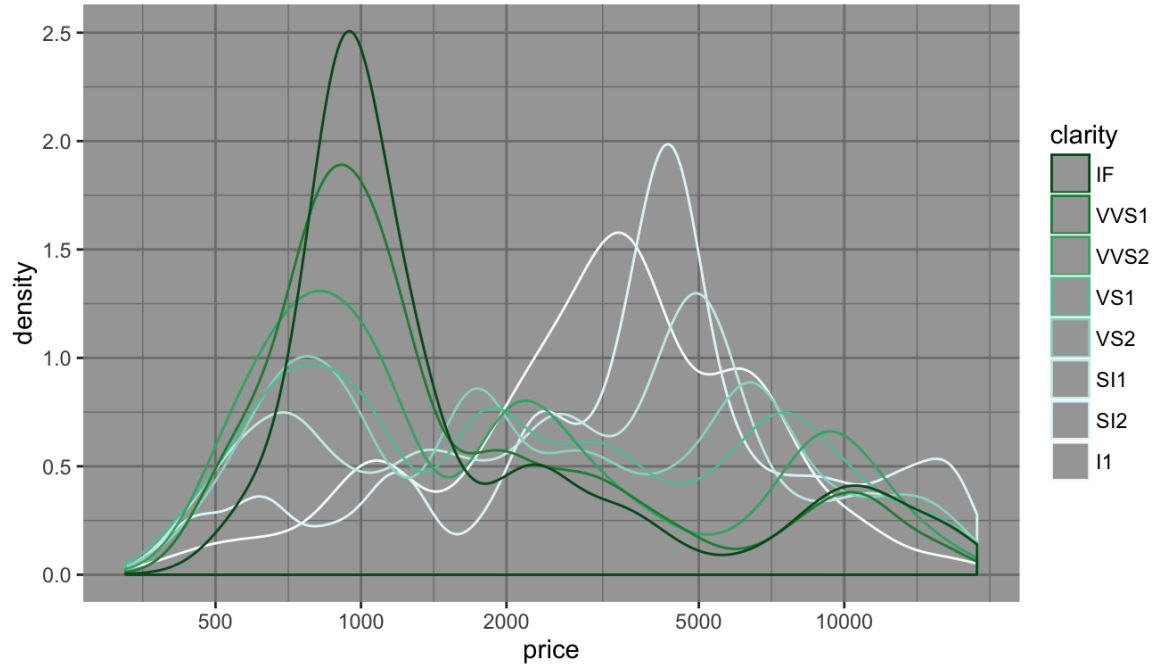
What was the strongest relationship you found?

The price of a diamond is positively and strongly correlated with carat and volume. The variables x, y, and z also correlate with the price but less strongly than carat and volume. Either carat or volume could be used in a model to predict the price of diamonds, however, both variables should not be used since they are measuring the same quality and show perfect correlation.

Multivariate Plots Section

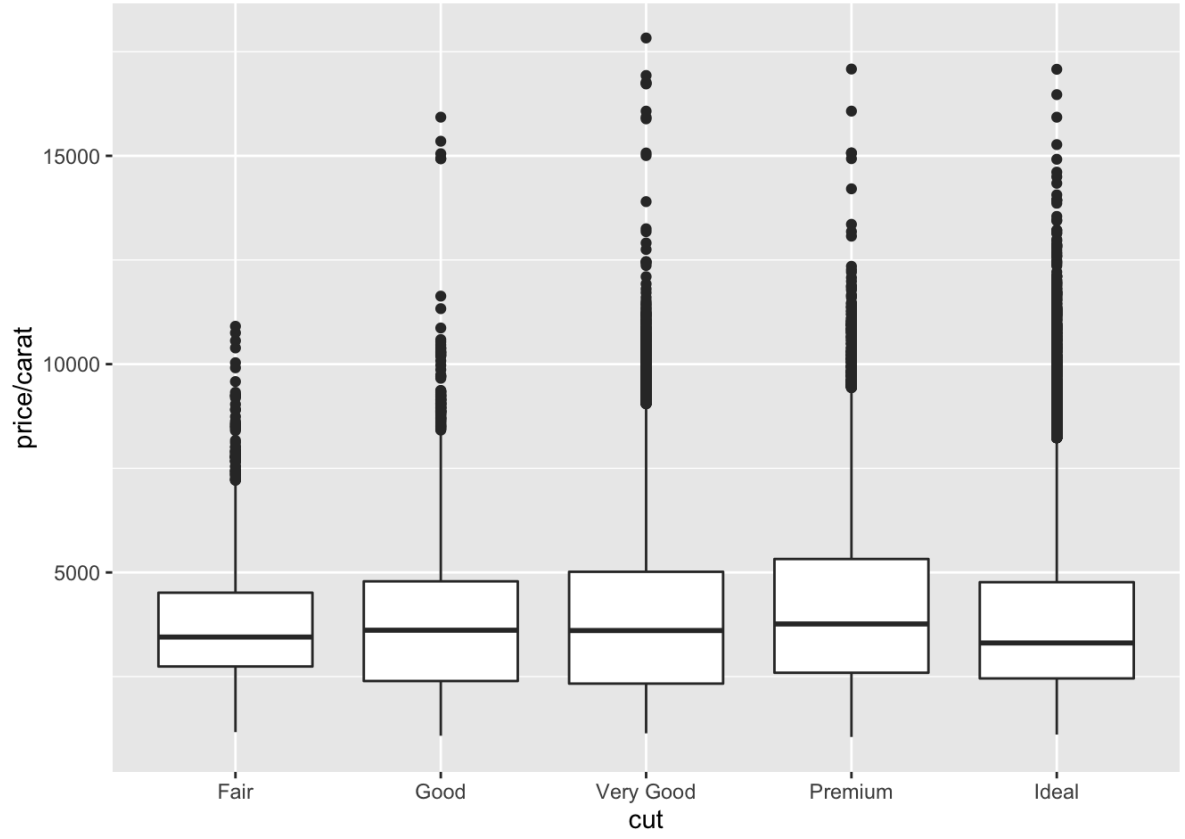

density
1.0
0.5
0.0
500
1000
2000
5000
10000
price
cut
Ideal
Premium
Very Good
Good
Fair

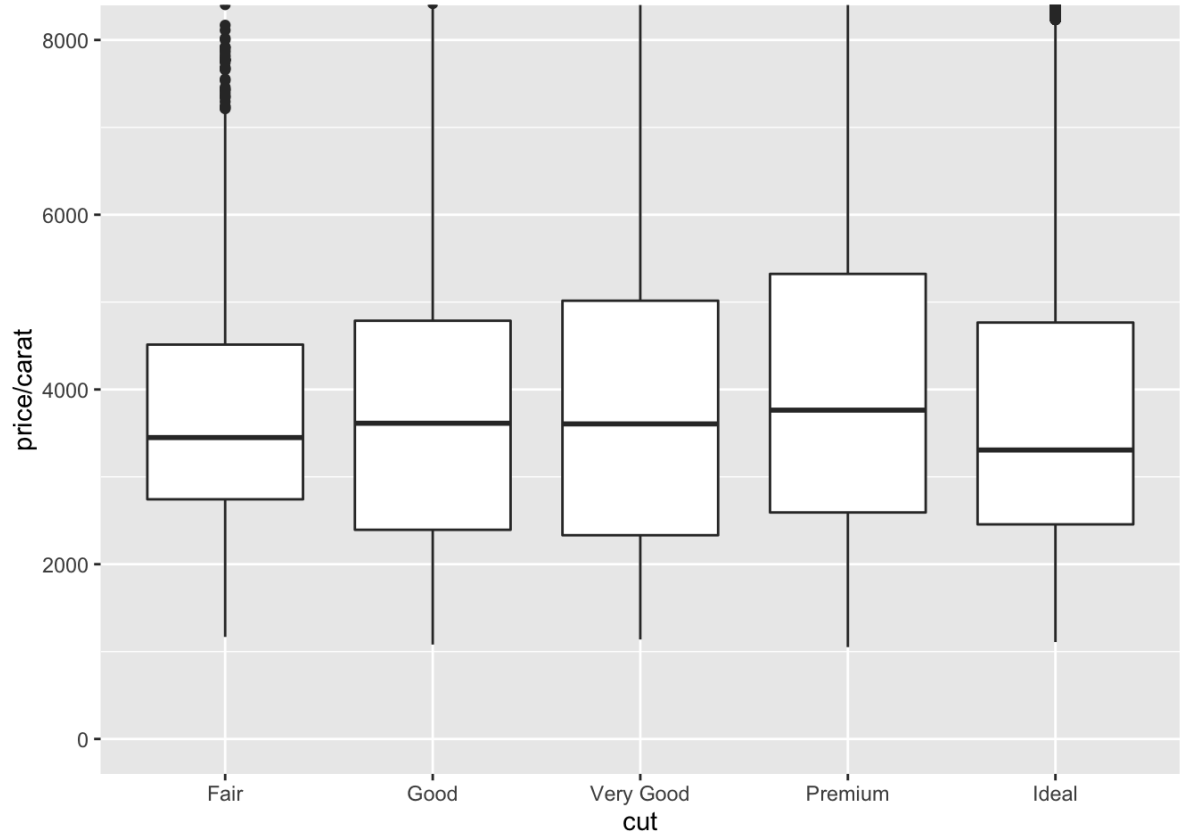

density
0.9
0.6
0.3
0.0
500
1000
2000
5000
10000
price
color
D
E
F
G
H
I
J


density
2.5
2.0
1.5
1.0
0.5
0.0
500
1000
2000
5000
10000
price
clarity
IF
VVS1
VVS2
VS1
VS2
SI1
SI2
I1

Tip: Even when doing exploration, it can be good to select appropriate color palettes and set plot themes in order to make plots more readable. (The above plots use sequential color palettes from the `RColorBrewer` package; other variables might require qualitative or diverging palettes.)

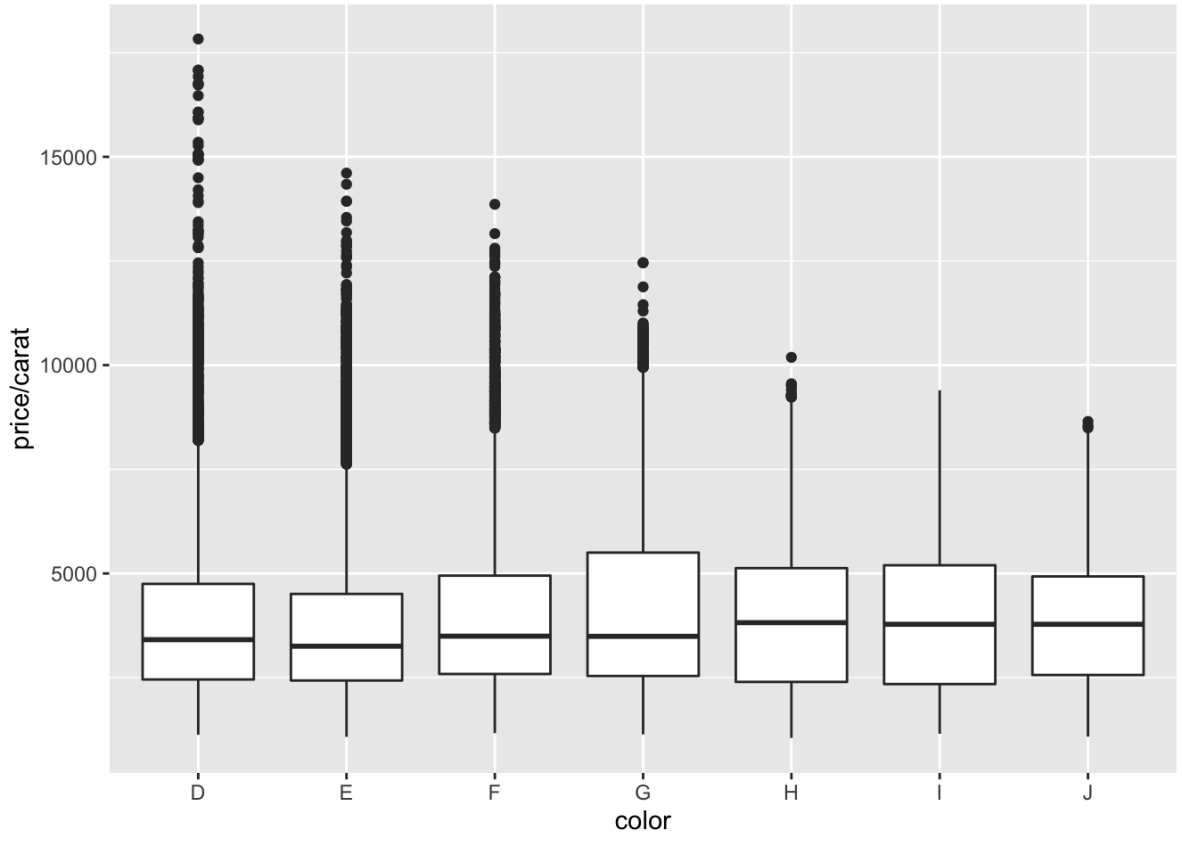
These density plots elaborate on the odd trends that were seen in the box plots earlier. Diamonds with better levels of clarity, cut, and color tend to occur more often at lower prices while diamonds with worse levels of clarity, cut, and color tend to occur more often at higher prices. Let's now take a look at price / carat.





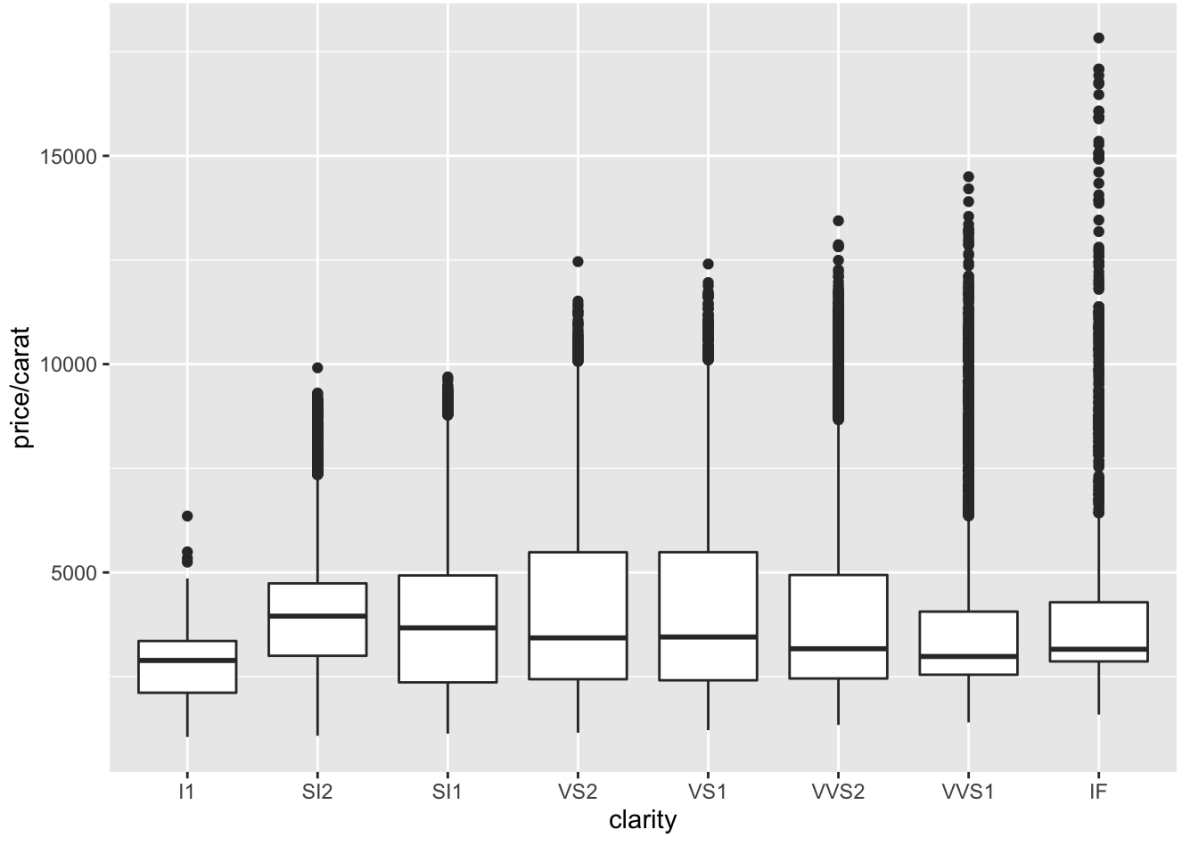
```
## cut: Fair
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1168    2743    3449    3767    4514   10910
## ------------------------------------------------------------
## cut: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1081    2394    3613    3860    4787   15930
## ------------------------------------------------------------
## cut: Very Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1139    2332    3606    4014    5016   17830
## ------------------------------------------------------------
## cut: Premium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1051    2592    3763    4223    5323   17080
## ------------------------------------------------------------
## cut: Ideal
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1109    2456    3307    3920    4766   17080
```

Wow! Ideal diamonds still have the lowest median for price per carat. The variance across the groups seems to be about the same with Fair cut diamonds having the least variation for the middle 50% of diamonds.



```
## color: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1128    2455    3411    3953    4749   17830
## ------------------------------------------------------------
## color: E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1078    2430    3254    3805    4508   14610
## ------------------------------------------------------------
## color: F
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1168    2587    3494    4135    4947   13860
## ------------------------------------------------------------
## color: G
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1139    2538    3490    4163    5500   12460
## ------------------------------------------------------------
## color: H
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1051    2397    3819    4008    5127   10190
## ------------------------------------------------------------
## color: I
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1152    2345    3780    3996    5197    9398
## ------------------------------------------------------------
## color: J
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1081    2563    3780    3826    4928    8647
```

The best color diamonds (D and E) still have the lowest medians on price per carat. Again, this is an unusual trend. This also seems strange since most diamonds in the data set are not of color D.



```
## clarity: I1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1051    2112    2887    2796    3354    6353
## ------------------------------------------------------------
## clarity: SI2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1081    3000    3951    4011    4738    9912
## ------------------------------------------------------------
## clarity: SI1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1130    2362    3669    3849    4928    9693
## ------------------------------------------------------------
## clarity: VS2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1152    2438    3429    4081    5484   12460
## ------------------------------------------------------------
## clarity: VS1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1215    2412    3450    4156    5485   12400
## ------------------------------------------------------------
## clarity: VVS2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1339    2455    3169    4204    4939   13440
## ------------------------------------------------------------
## clarity: VVS1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1400    2545    2982    3851    4060   14500
## ------------------------------------------------------------
## clarity: IF
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1588    2865    3156    4260    4284   17830
```

This plot seems more reasonable. The lowest median price per carat has clarity I1 which is the lowest clarity rating. The median increases slightly then holds relatively constant before decreasing again for the highest clarity. The variance increases then decreases across the clarity levels from worst to best.

Let’s take another look at other variables and their correlations with price and try to work towards building a linear model to predict price.