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.

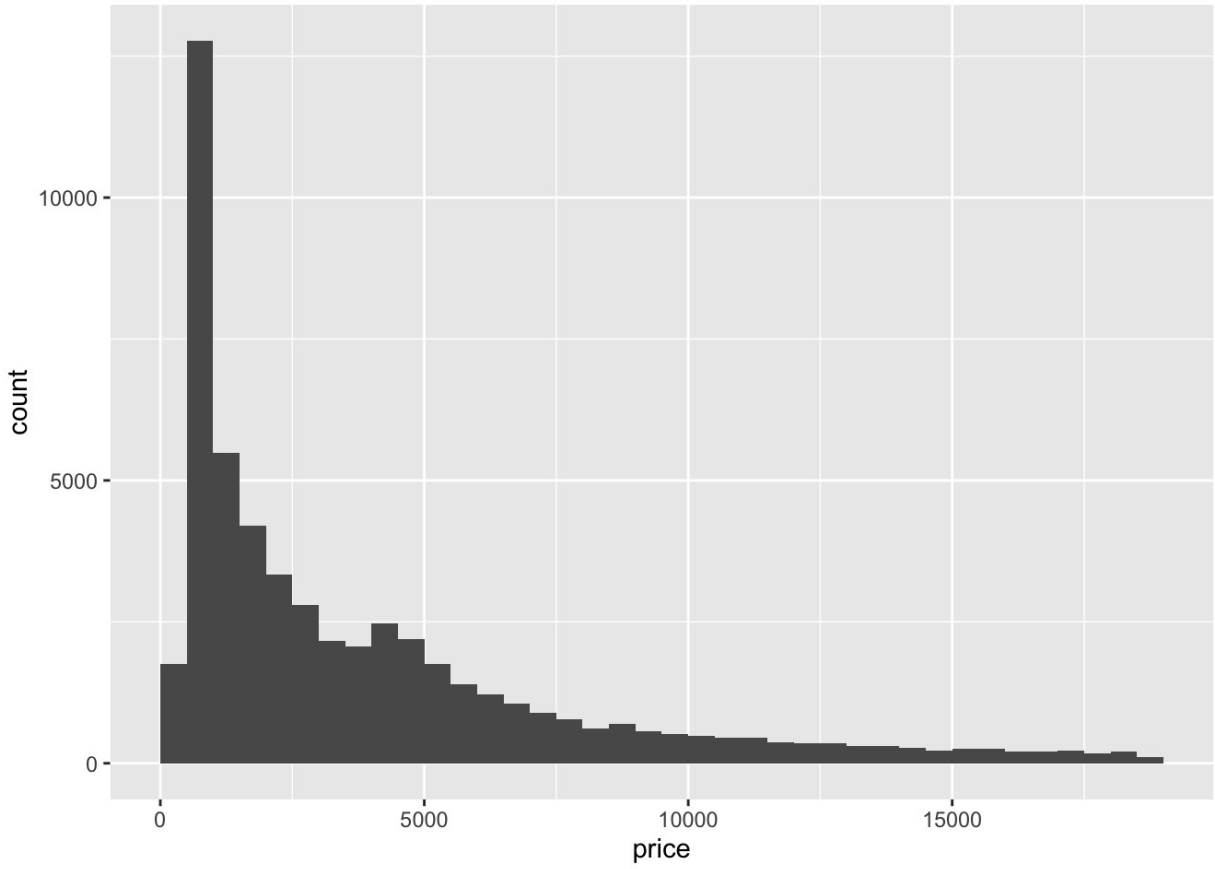
Univariate Plots Section

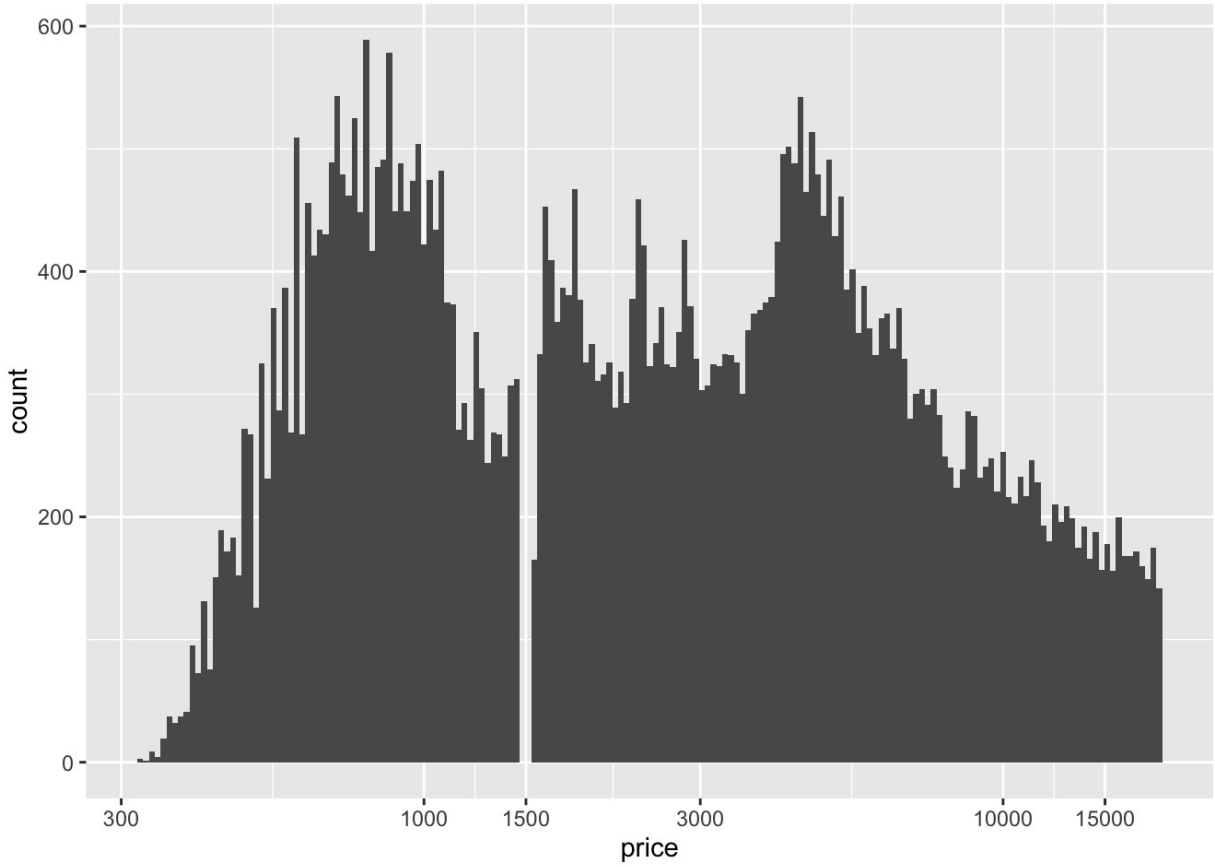
```
## [1] 53940    10
```

```
## 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
##
```

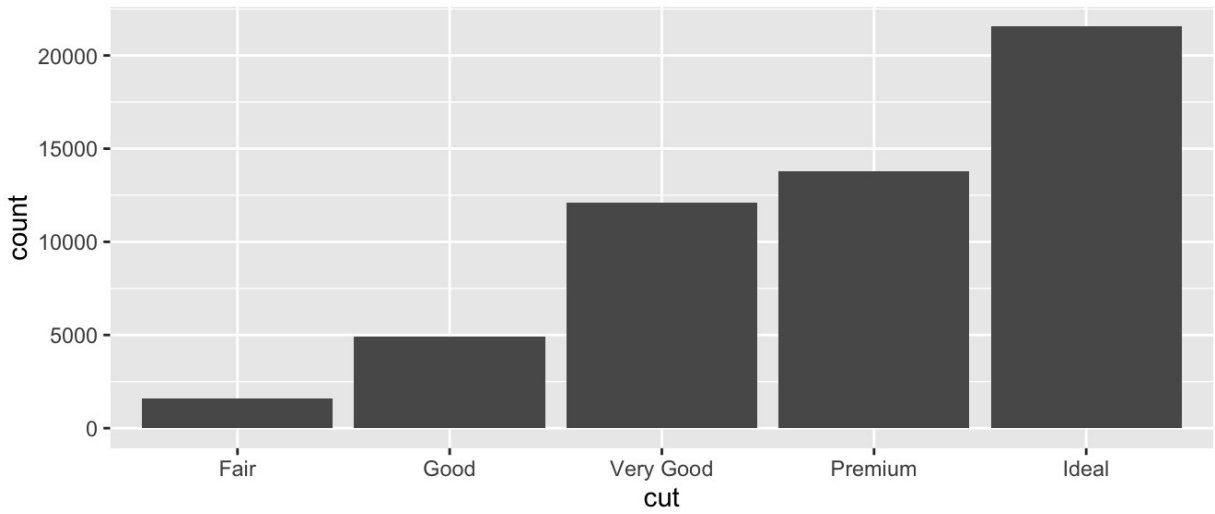
Our dataset consists of ten variables, with almost 54,000 observations.

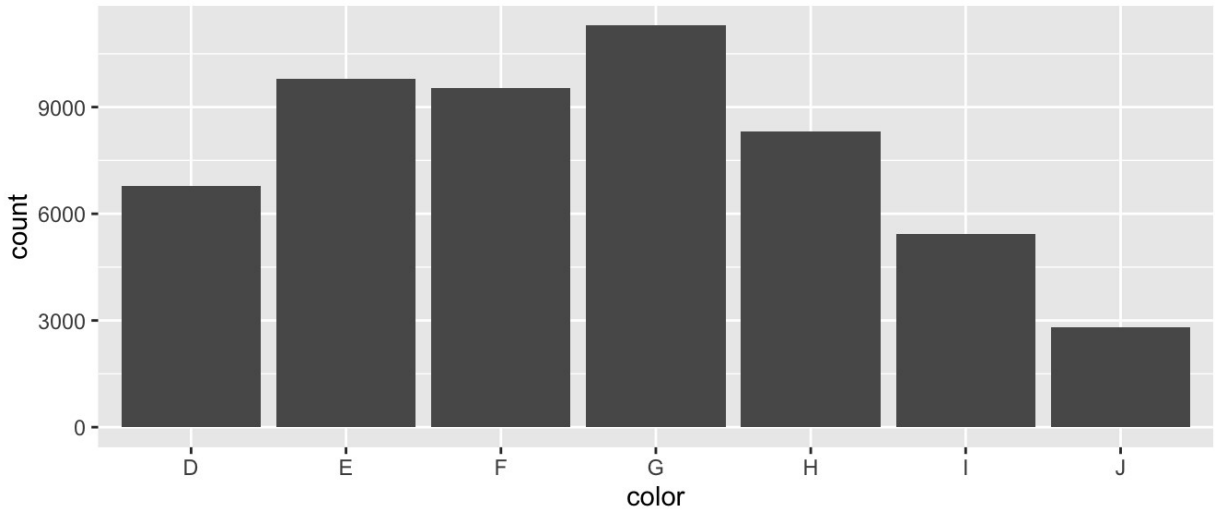


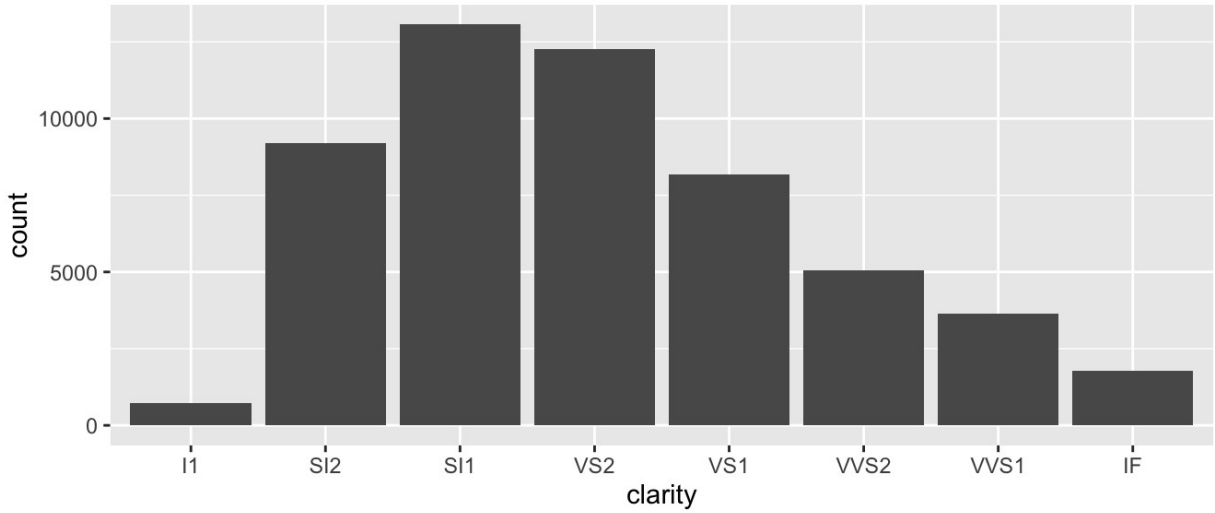


Tip: When plotting on a log scale, it is useful to note that 3 is about halfway between 1 and 10. As a side note, try not to plot counts on a log scale since counts of 0 are undefined and counts of 1 have a value of 0 (no height).

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.

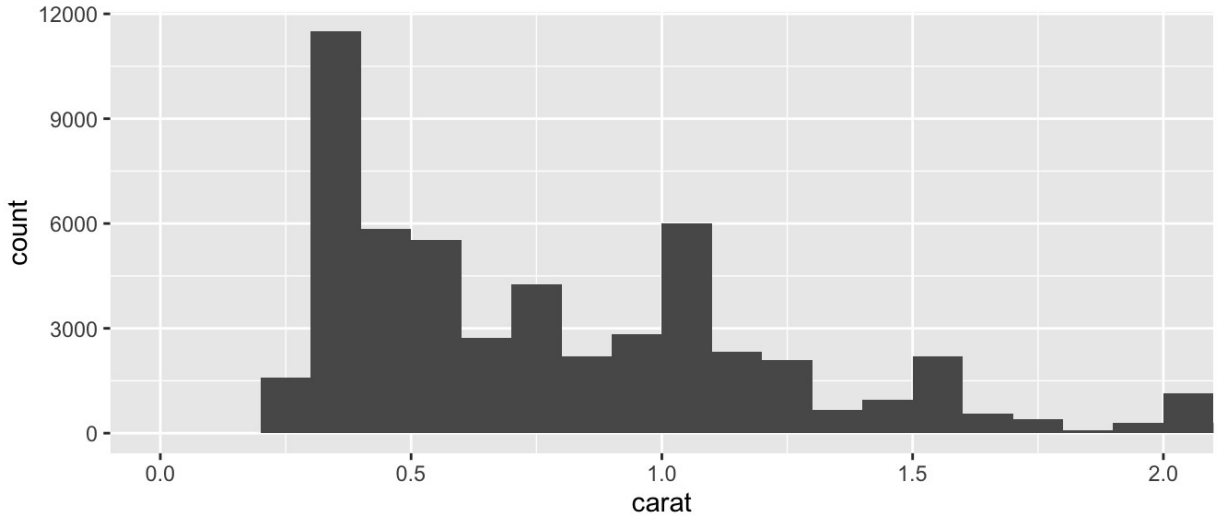


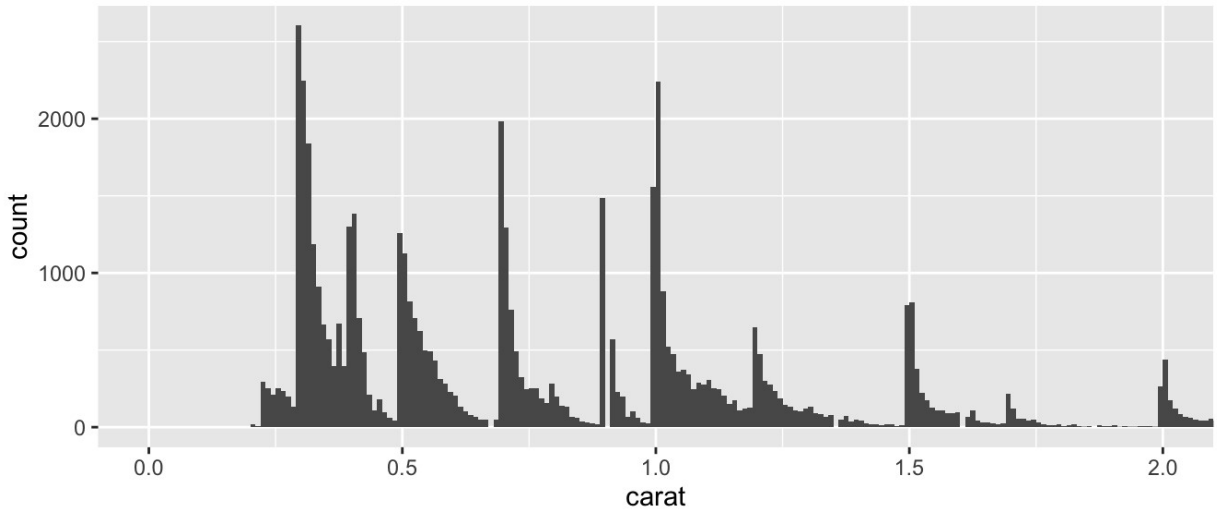




Tip: You can change the height and width of plots in code chunks with the `fig.height` and `fig.width` parameters in the chunk options.

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.

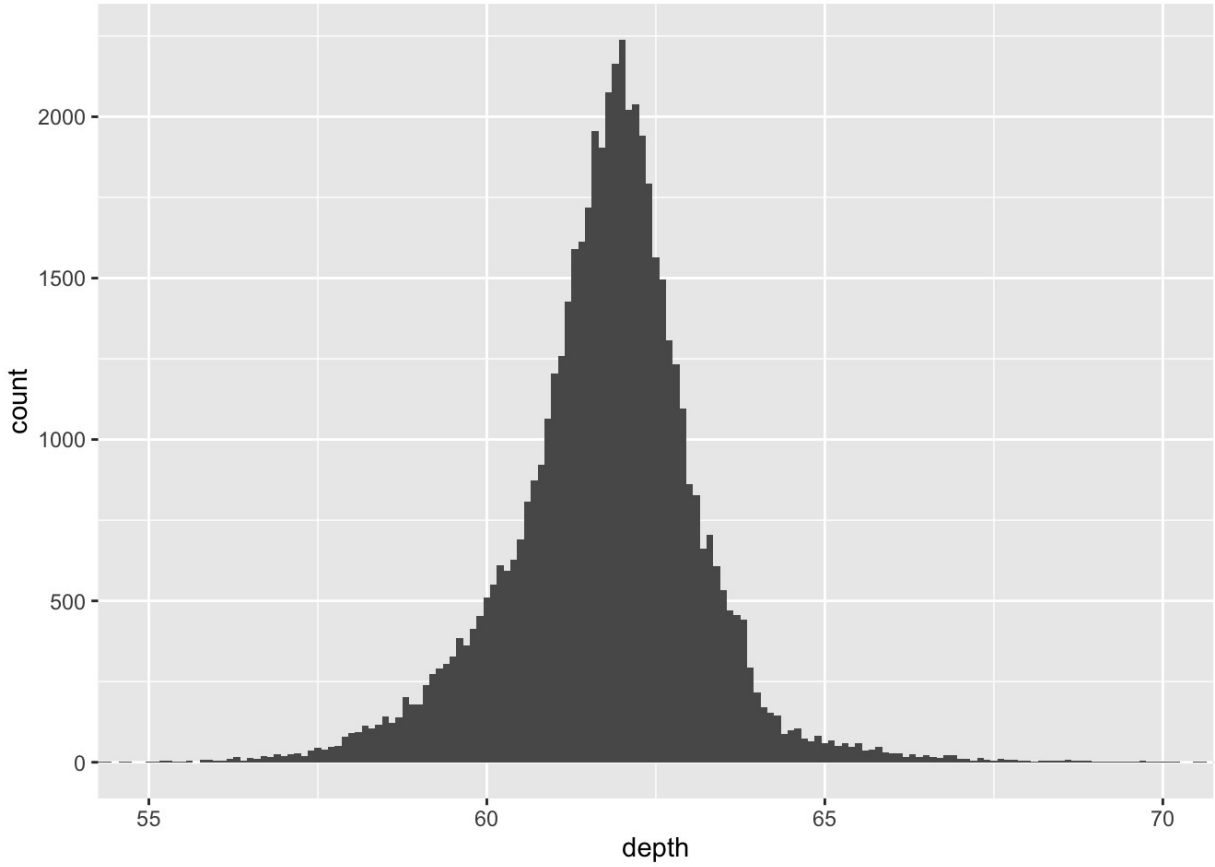





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2000  0.4000  0.7000  0.7979  1.0400  5.0100
```

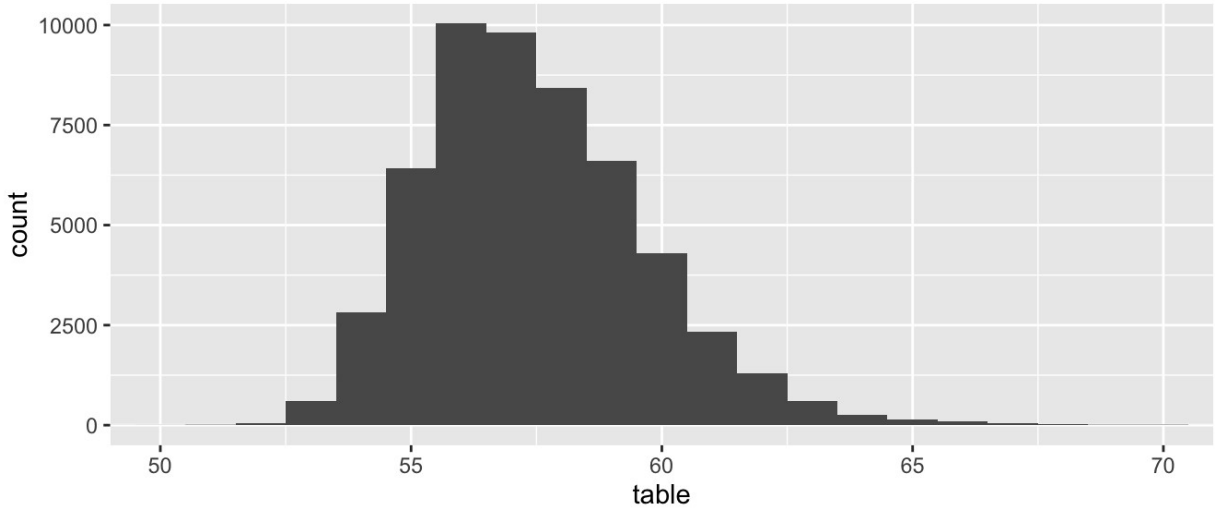
```
##
##  0.3 0.31 1.01  0.7 0.32    1  0.9 0.41  0.4 0.71  0.5 0.33 0.51 0.34 1.02
## 2604 2249 2242 1981 1840 1558 1485 1382 1299 1294 1258 1189 1127  910  883
## 0.52 1.51  1.5 0.72 0.53 0.42 0.38 0.35  1.2 0.54 0.36 0.91 1.03 0.55 0.56
##  817  807  793  764  709  706  670  667  645  625  572  570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

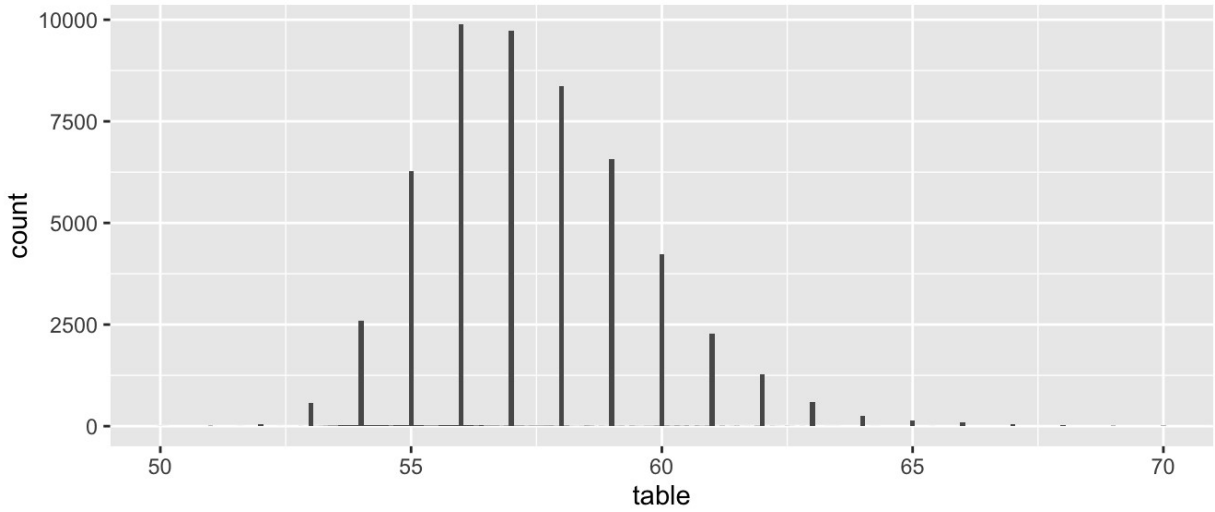
The lightest diamond is 0.2 carat and the heaviest diamond is 5.0100. Above, I plot the main body of carat weights, trimming the highest-carat diamonds. Some carat weights occur more often than other carat weights. Many of the most common carat counts end in x.x0 or x.x1. I wonder how carat is connected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

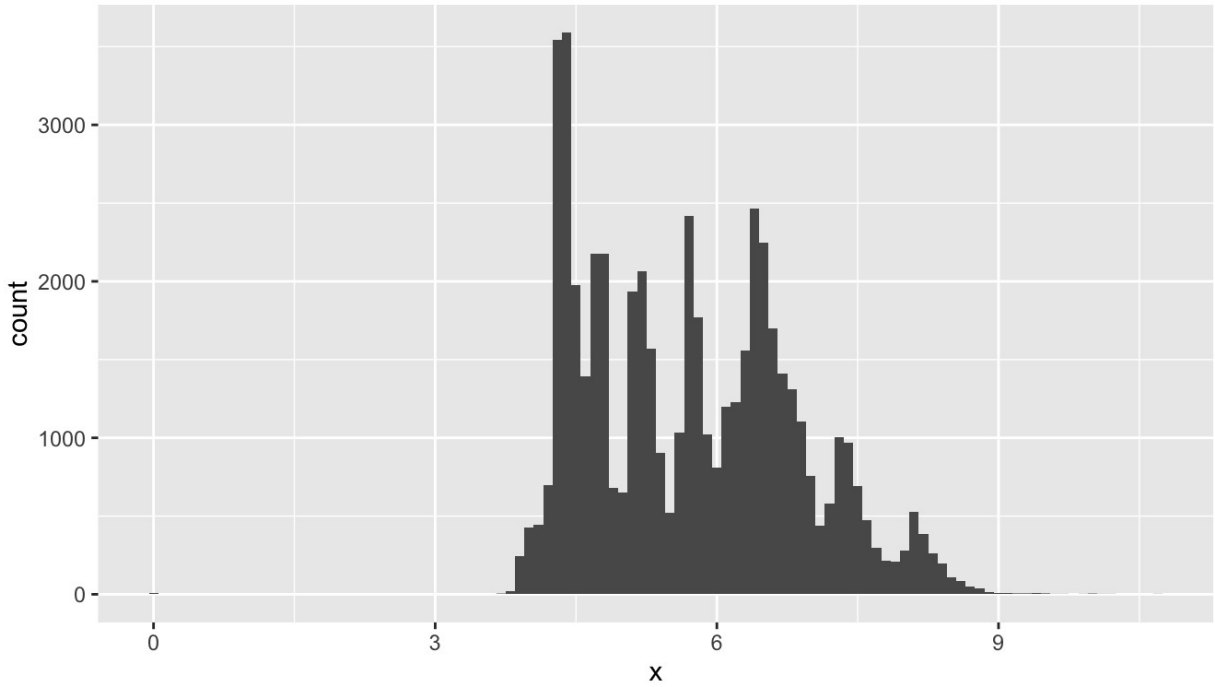


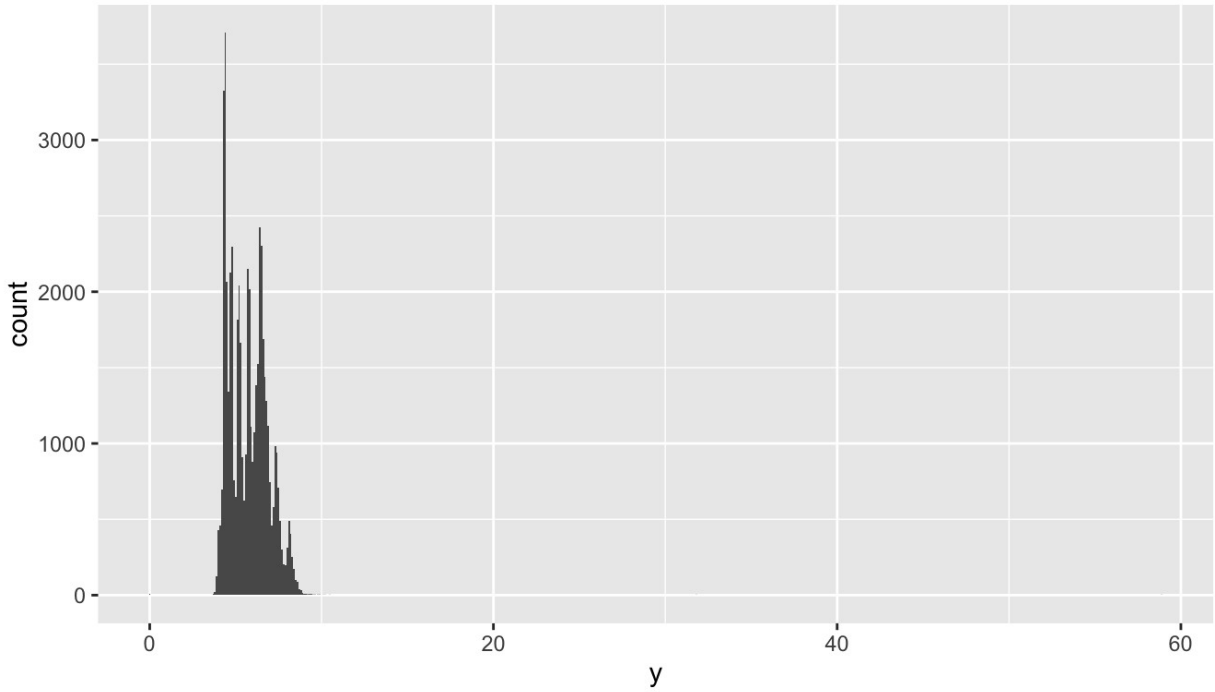


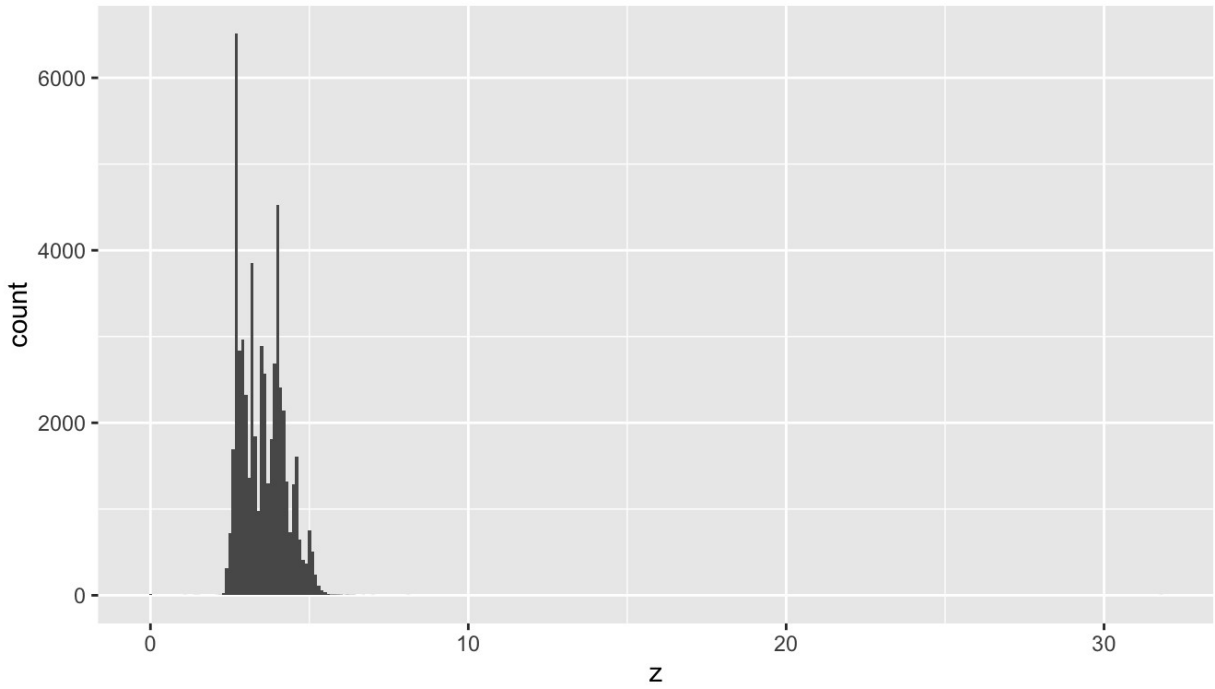

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   56.00   57.00   57.46   59.00   95.00
```

```
##
##   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
```

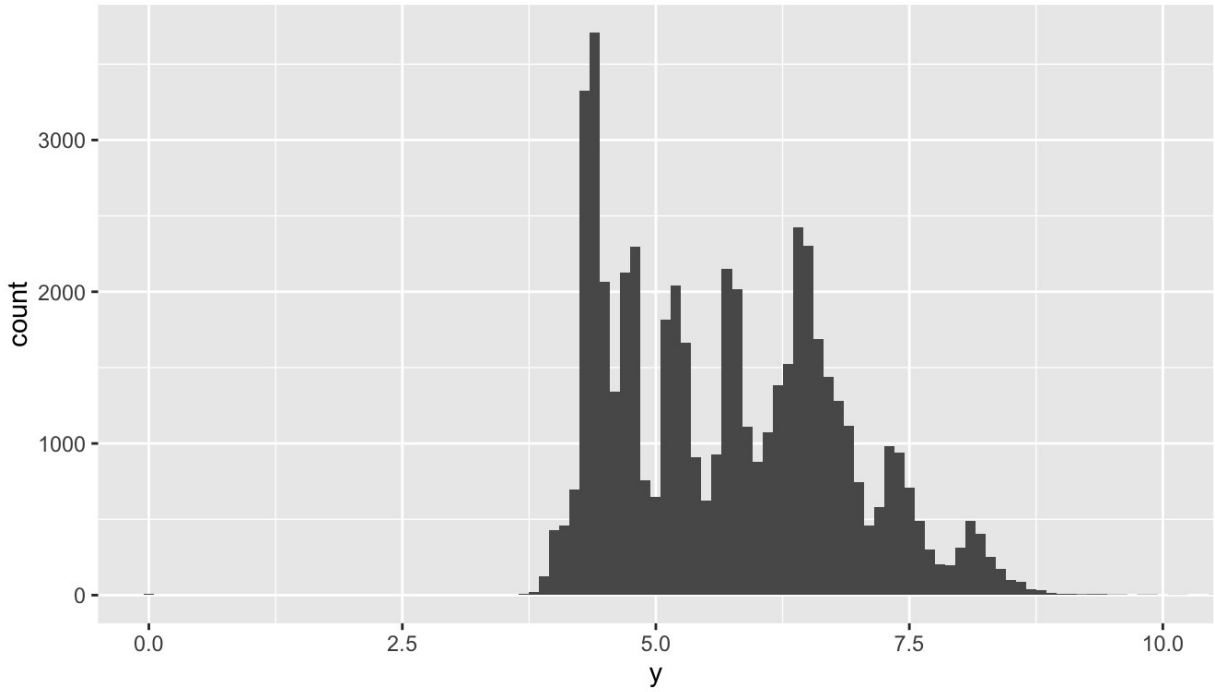
Setting the binwidth indicates that most table values are integers. Most diamonds have a table between 55 mm and 60 mm. Again, I wonder if this has anything to do with the cut of a diamond. Cut is a quality of a diamond that may influence carat weight and is responsible for making a diamond sparkle.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

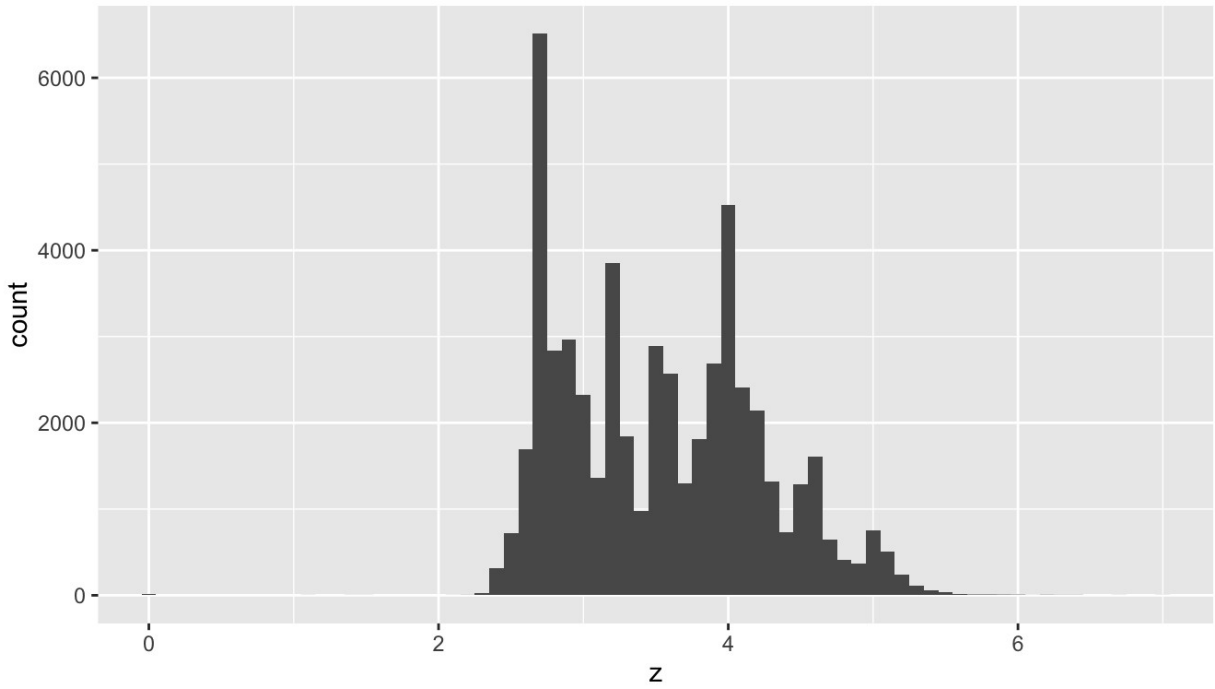






Most diamonds have an x dimension between 4 mm and 7 mm, a y dimension between 4 mm and 7 mm, and a z dimension between 2 mm and 6 mm. The y- and z- plots have a few high outliers so let’s zoom in.





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
```

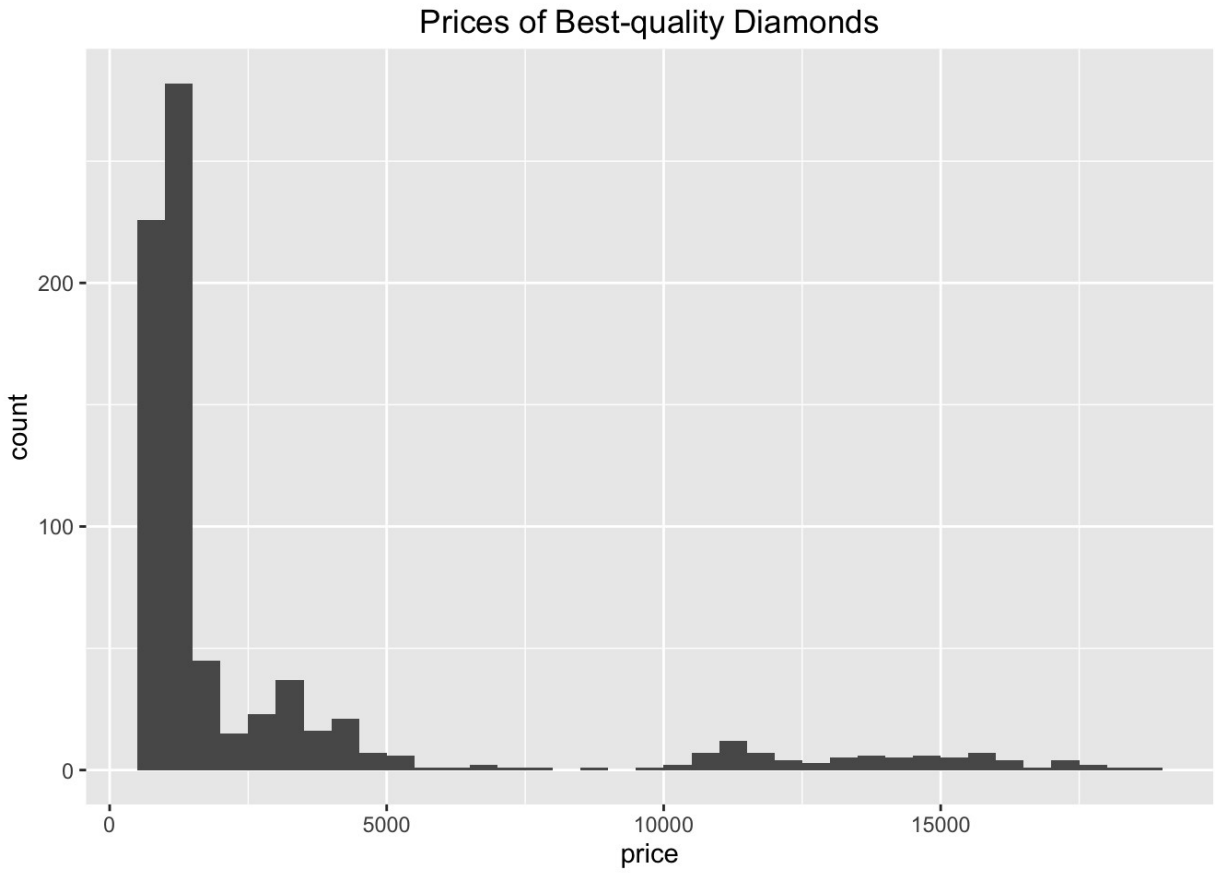
There are eight diamonds with missing x values, seven diamonds with missing y values, and twenty diamonds with missing z values.

```
## 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
```

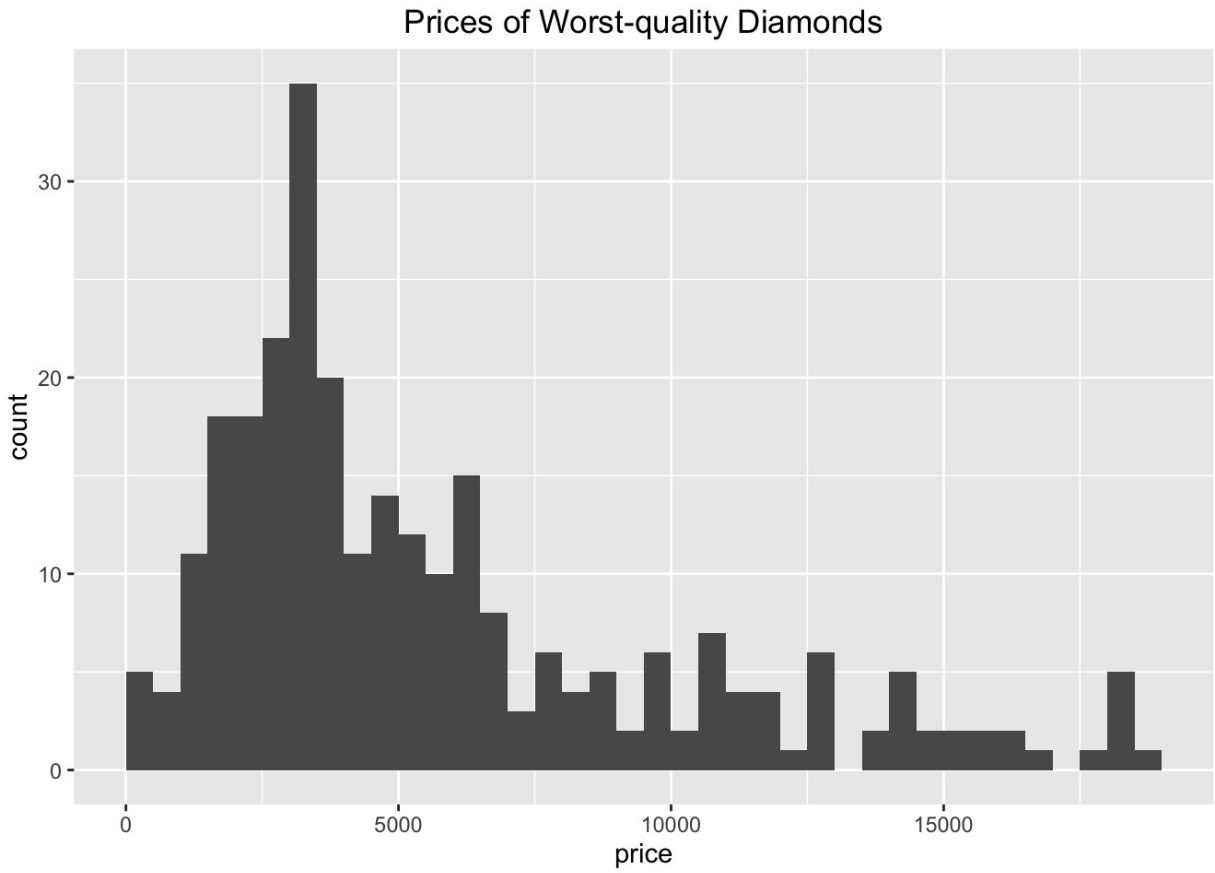
If and only if x or y dimensions are 0, then the z dimension is 0. Comparing the diamonds in this subset to all other diamonds, these diamonds tend to be very expensive or fall in the third quartile of the entire diamonds data set. Other variables such as carat, depth, table, and price are reported so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

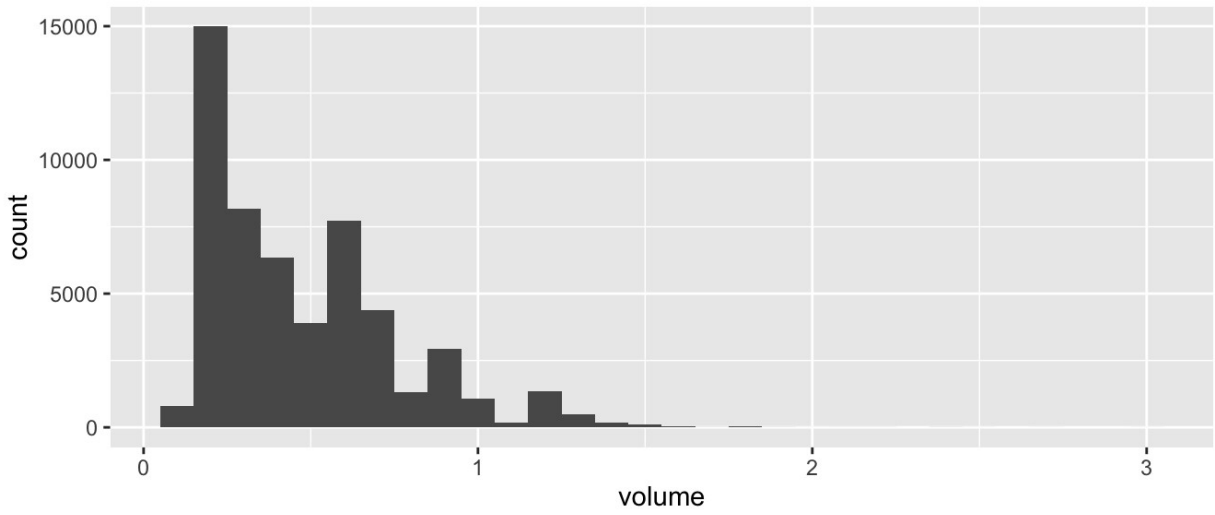
```
##
## FALSE  TRUE
## 53920    20
```

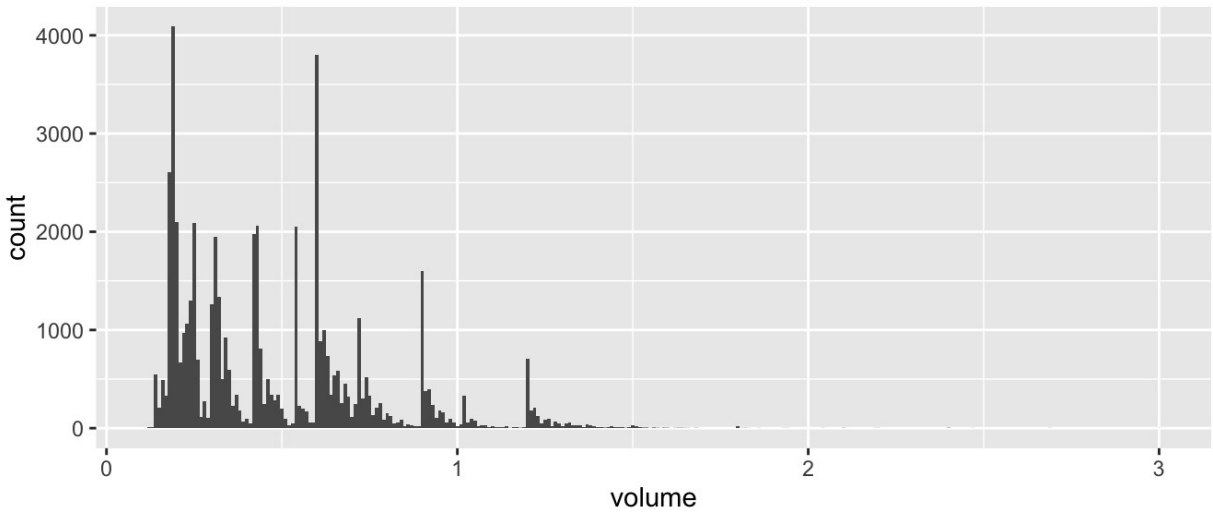
```
##       carat       cut color clarity depth table price    x    y z volume
## 2208   1.00   Premium     G     SI2  59.1    59  3142 6.55 6.48 0      0
## 2315   1.01   Premium     H      I1  58.1    59  3167 6.66 6.60 0      0
## 4792   1.10   Premium     G     SI2  63.0    59  3696 6.50 6.47 0      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
```




4000
3000
2000
1000
0
count
0
1
2
3
volume