Diamonds Exploration by Chris Saden

Univariate Plots Section

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

```
##  [1] "carat"   "cut"     "color"   "clarity" "depth"   "table"   "price"
##  [8] "x"       "y"       "z"
```

```
## '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 ...
```

```
## [1] "Fair"      "Good"      "Very Good" "Premium"   "Ideal"
```

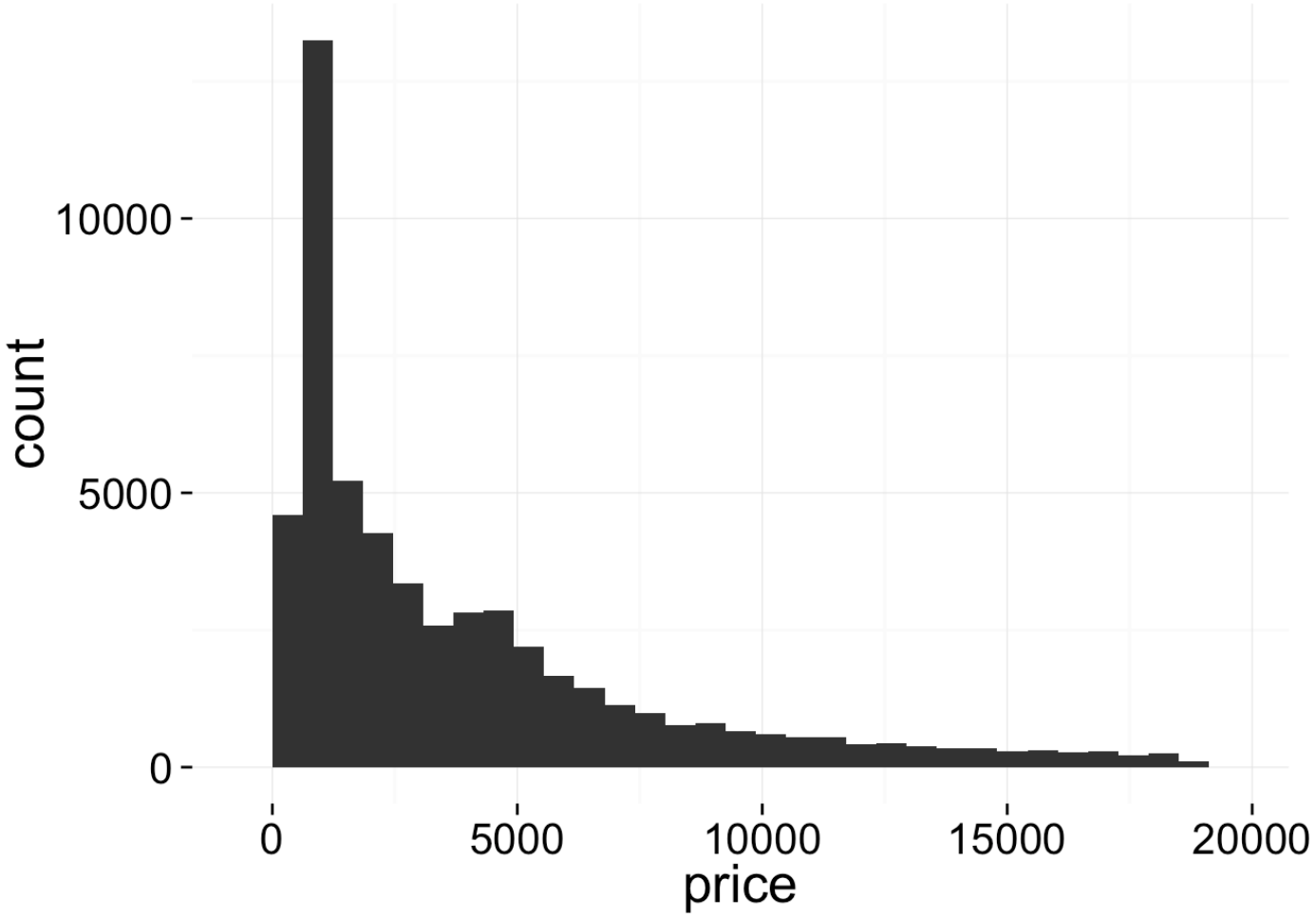
```
## [1] "D" "E" "F" "G" "H" "I" "J"
```

```
## [1] "I1"   "SI2"  "SI1"  "VS2"  "VS1"  "VVS2" "VVS1" "IF"
```

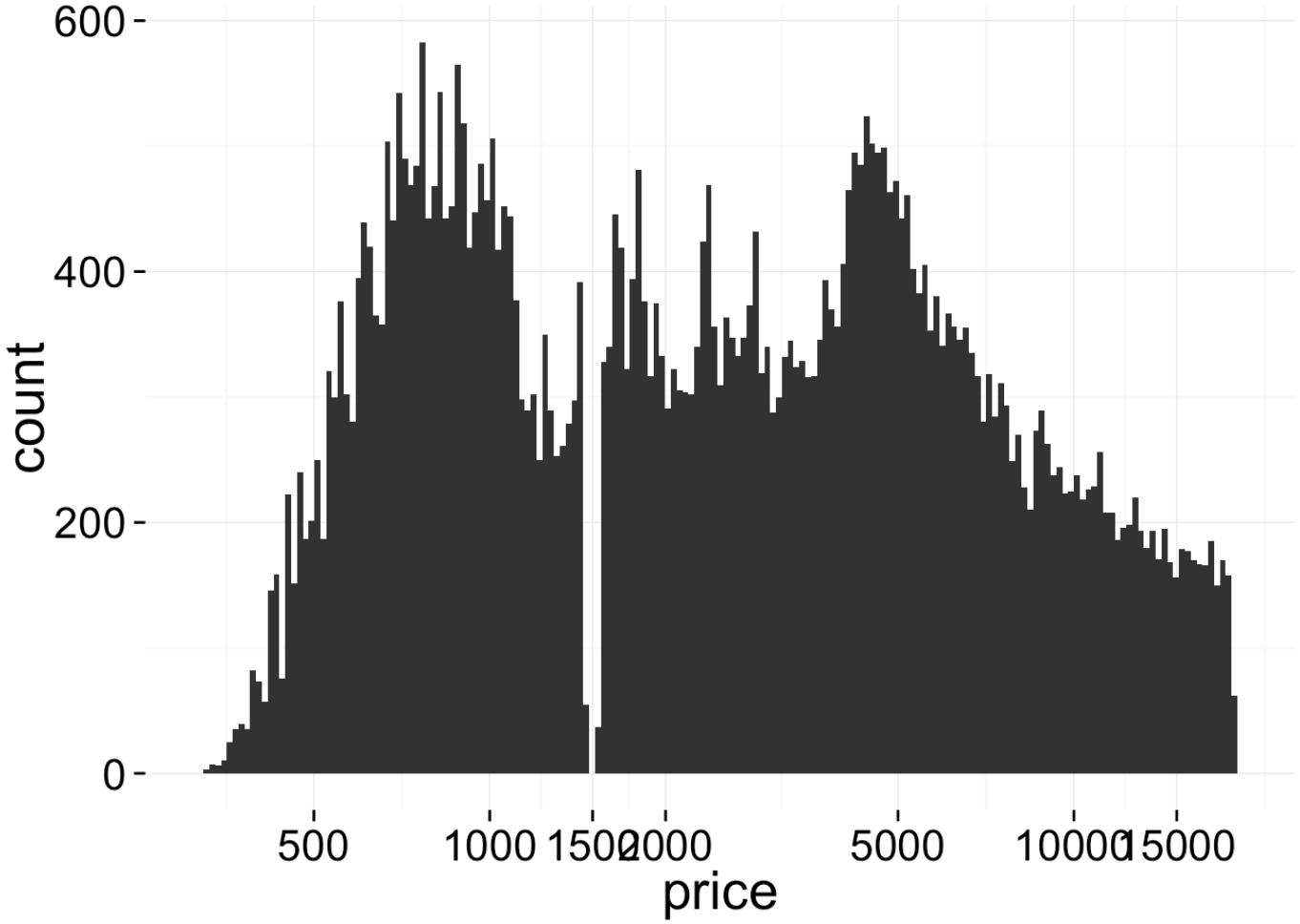
```
##      carat               cut        color        clarity     
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065  
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258  
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194  
##  Mean   :0.7979   Premium   :13791   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  
##                                   
```

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.

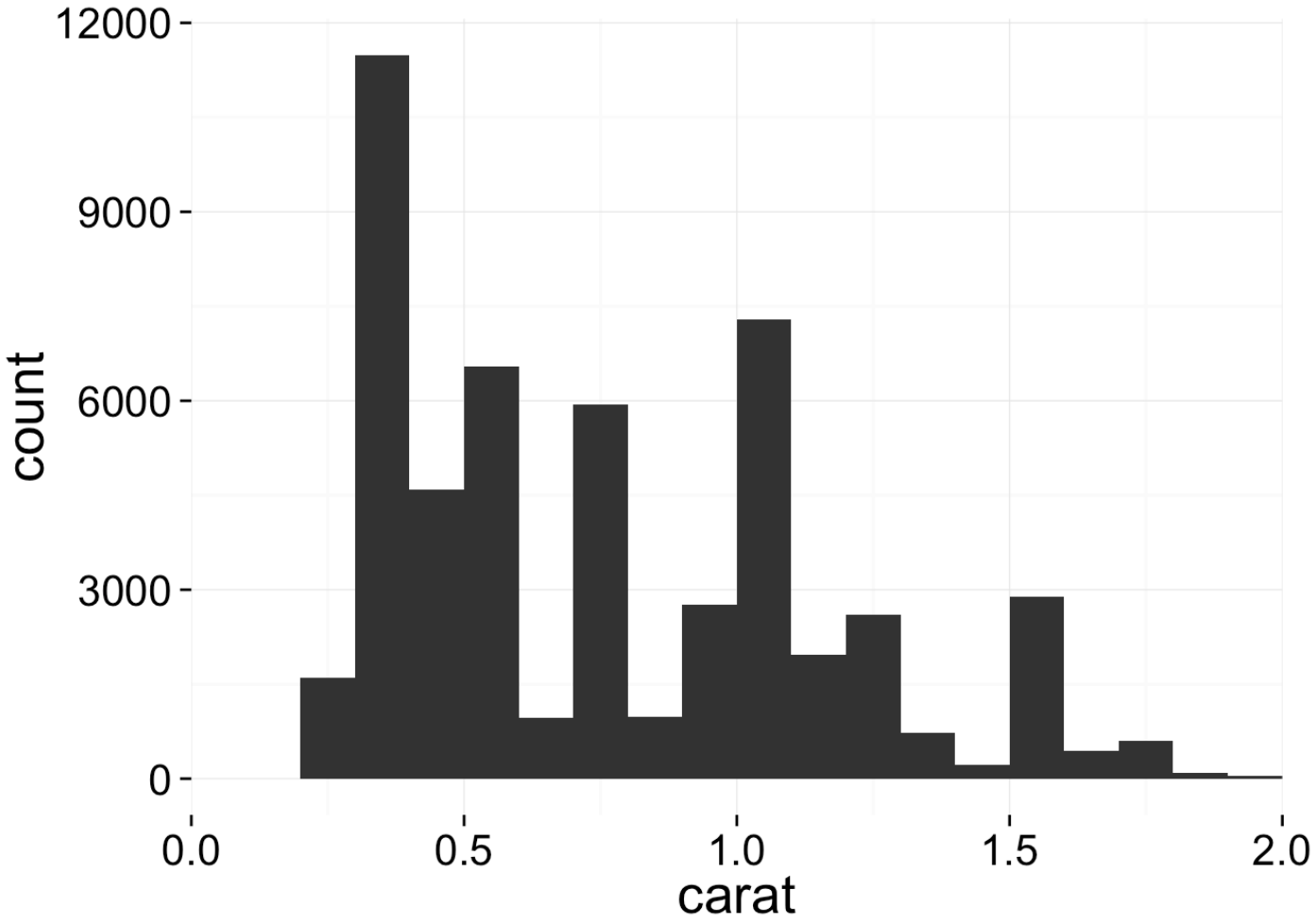
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



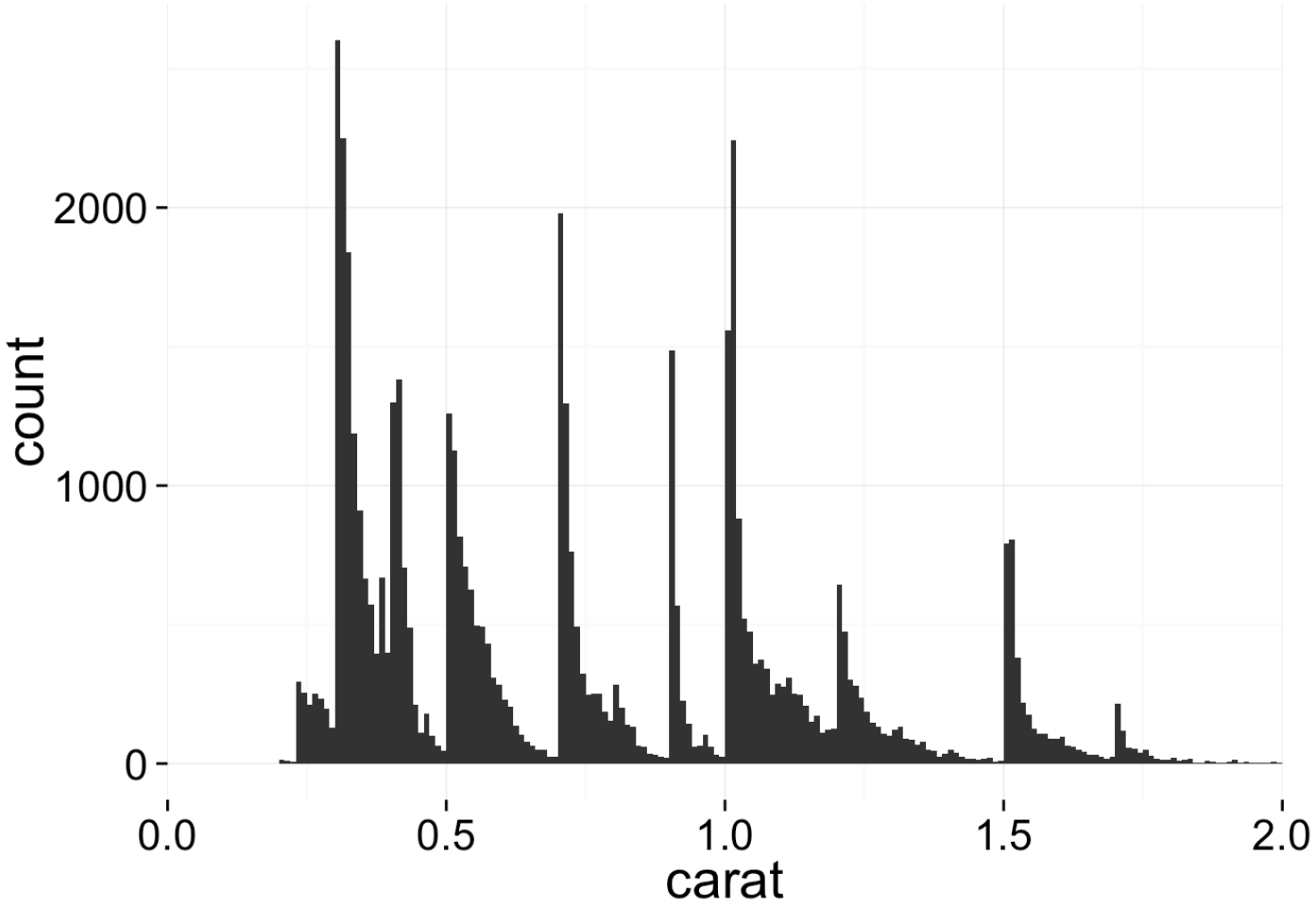

```
## Warning: position_stack requires constant width: output may be incorrect
```



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.




```
## Warning: position_stack requires constant width: output may be incorrect
```



Some carat weights occur more often than other carat weights. I wonder how carat is connected to price, and I wonder if the carat values are specific to certain cuts of diamonds. For now, I'm going to see which carat weights are most common.

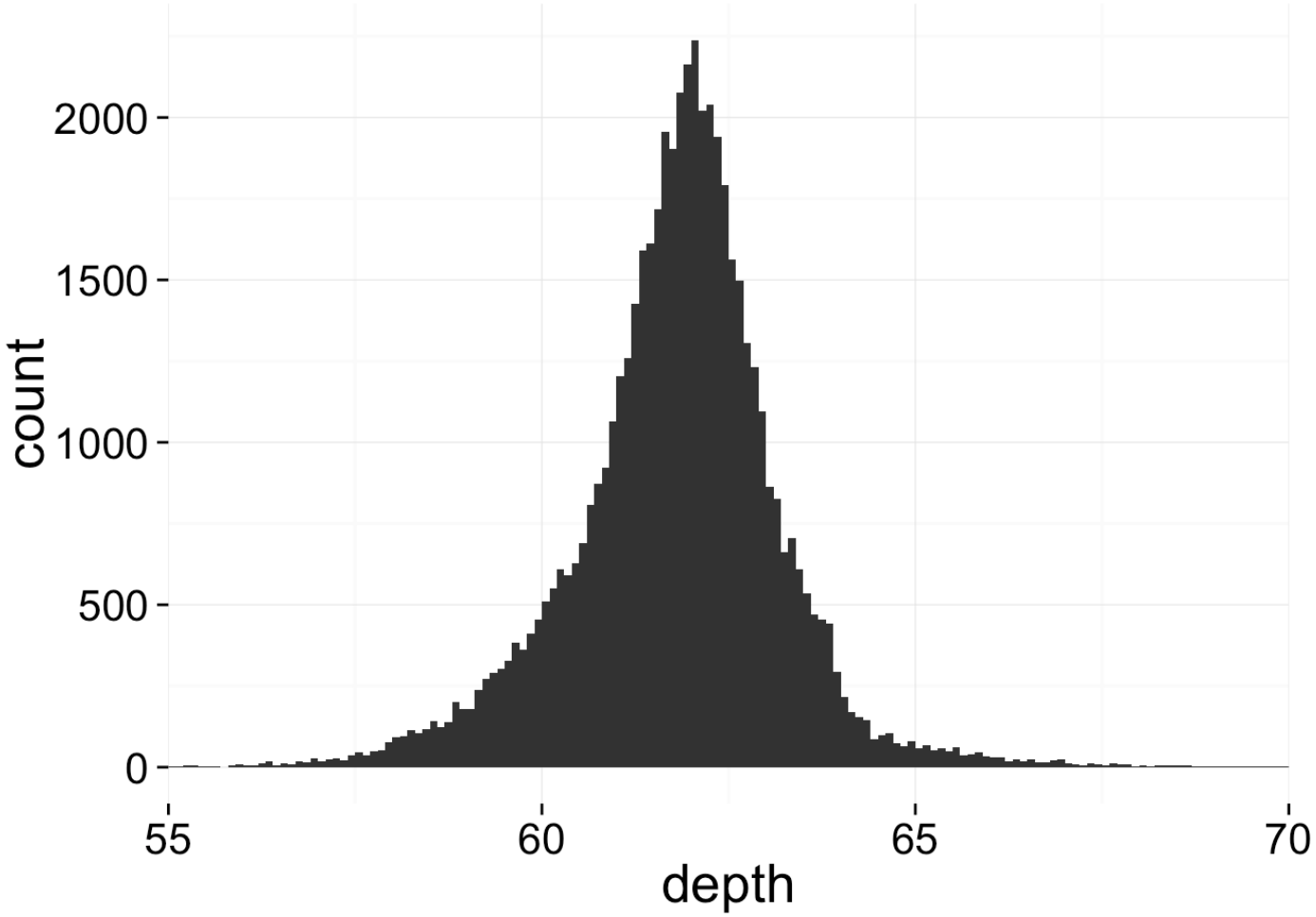
```
##
## FALSE
## 53940
```

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

No diamonds have a carat value of 0. The lightest diamond is 0.2 carat and the heaviest diamond is 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
## 0.73 0.43 1.04 1.21 2.01 0.57 0.39 0.37 1.52 1.06 1.05 1.07 0.74 0.58 1.11
##  492  488  475  473  440  430  398  394  381  373  361  342  322  310  308
## 1.22 0.23 1.09  0.8 0.59 1.23  1.1    2 0.24 0.26 0.76 0.77 1.12 0.75 1.08
##  300  293  287  284  282  279  278  265  254  253  251  251  251  249  246
## 1.13 1.24 0.27  0.6 0.92 1.53  1.7 0.25 0.44 1.14 0.61 0.81 0.28 0.78 1.25
##  246  236  233  228  226  220  215  212  212  207  204  200  198  187  187
## 0.46 2.02 1.54 1.16 0.79 1.15 1.26 0.93 0.82 0.62 1.27 1.31 0.83 0.29 1.19
##  178  177  174  172  155  149  146  142  140  135  134  133  131  130  126
## 1.55 1.18  1.3 2.03 1.71 0.45 1.17 1.56 1.28 1.57 0.96 0.63 1.29 0.47  1.6
##  124  123  122  122  119  110  110  109  106  106  103  102  101   99   95
## 1.32 1.58 1.59 1.33 2.04 0.64 1.35 1.34 2.05 0.65 0.95 0.84 1.61 0.48 0.85
##   89   89   89   87   86   80   77   68   67   65   65   64   64   63   62
## 1.62 2.06 0.94 0.97 1.72 1.73  2.1 1.36  1.4 1.63 1.75 2.07 0.66 0.67 2.14
##   61   60   59   59   57   52   52   50   50   50   50   50   48   48   48
## 1.37 0.49 2.09 1.64 2.11 2.08 1.41 1.74 1.39 0.86 1.65  2.2 0.87 0.98 2.18
##   46   45   45   43   43   41   40   40   36   34   32   32   31   31   31
## 1.66 1.76 2.22 0.69 1.38 0.68 1.42 1.67 2.12 2.16 1.69 0.88 0.99 2.21 2.15
##   30   28   27   26   26   25   25   25   25   25   24   23   23   23   22
## 2.19 0.89 1.47  1.8 2.13  2.3 2.28 1.43 1.68 1.44 1.46 1.83 2.17 2.25 1.77
##   22   21   21   21   21   21   20   19   19   18   18   18   18   18   17
## 2.29  2.5 2.51 2.24 2.32 1.45 1.79 2.26 3.01 1.82 2.23 2.31  2.4  0.2 1.78
##   17   17   17   16   16   15   15   15   14   13   13   13   13   12   12
## 1.91 2.27 1.49 0.21 1.81 1.86 2.33 2.48 2.52 2.54 2.36 2.38 2.42 2.53    3
##   12   12   11    9    9    9    9    9    9    9    8    8    8    8    8
## 1.48 1.87  1.9 2.35 2.39 1.93 2.37 2.43 0.22 1.98 2.34 2.41 1.84 1.88 1.89
##    7    7    7    7    7    6    6    6    5    5    5    5    4    4    4
## 1.96 1.97 2.44 2.45 1.85 1.94 1.95 1.99 2.46 2.47 2.49 2.55 2.56 2.57 2.58
##    4    4    4    4    3    3    3    3    3    3    3    3    3    3    3
##  2.6 2.61 2.63 2.66 2.72 2.74 1.92 2.68 2.75  2.8 3.04 4.01 2.59 2.64 2.65
##    3    3    3    3    3    3    2    2    2    2    2    2    1    1    1
## 2.67  2.7 2.71 2.77 3.02 3.05 3.11 3.22 3.24  3.4  3.5 3.51 3.65 3.67    4
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 4.13  4.5 5.01
##    1    1    1
```

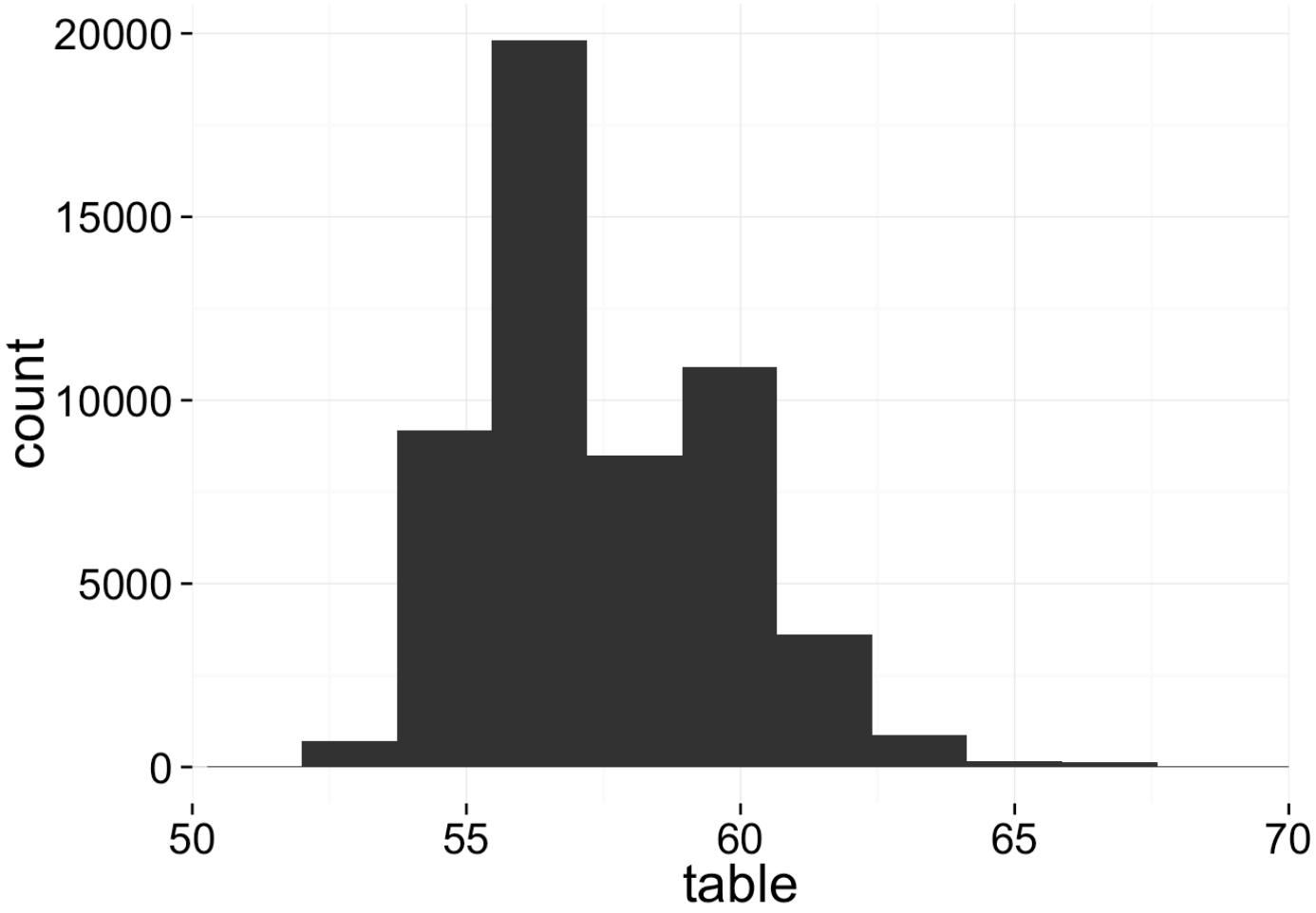
```
## Warning: position_stack requires constant width: output may be incorrect
```



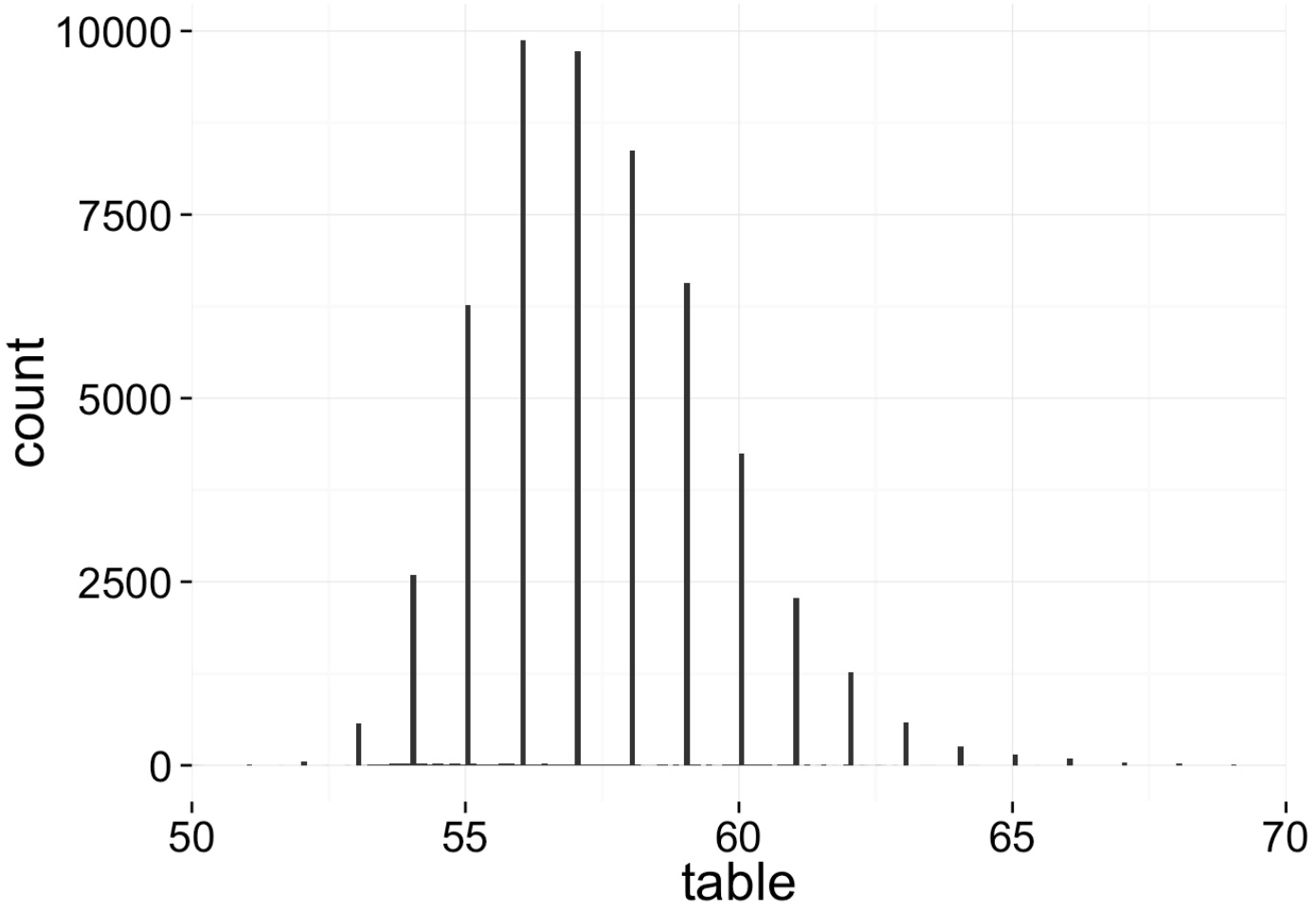

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   61.00   61.80   61.75   62.50   79.00
```

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

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
## Warning: position_stack requires constant width: output may be incorrect
```

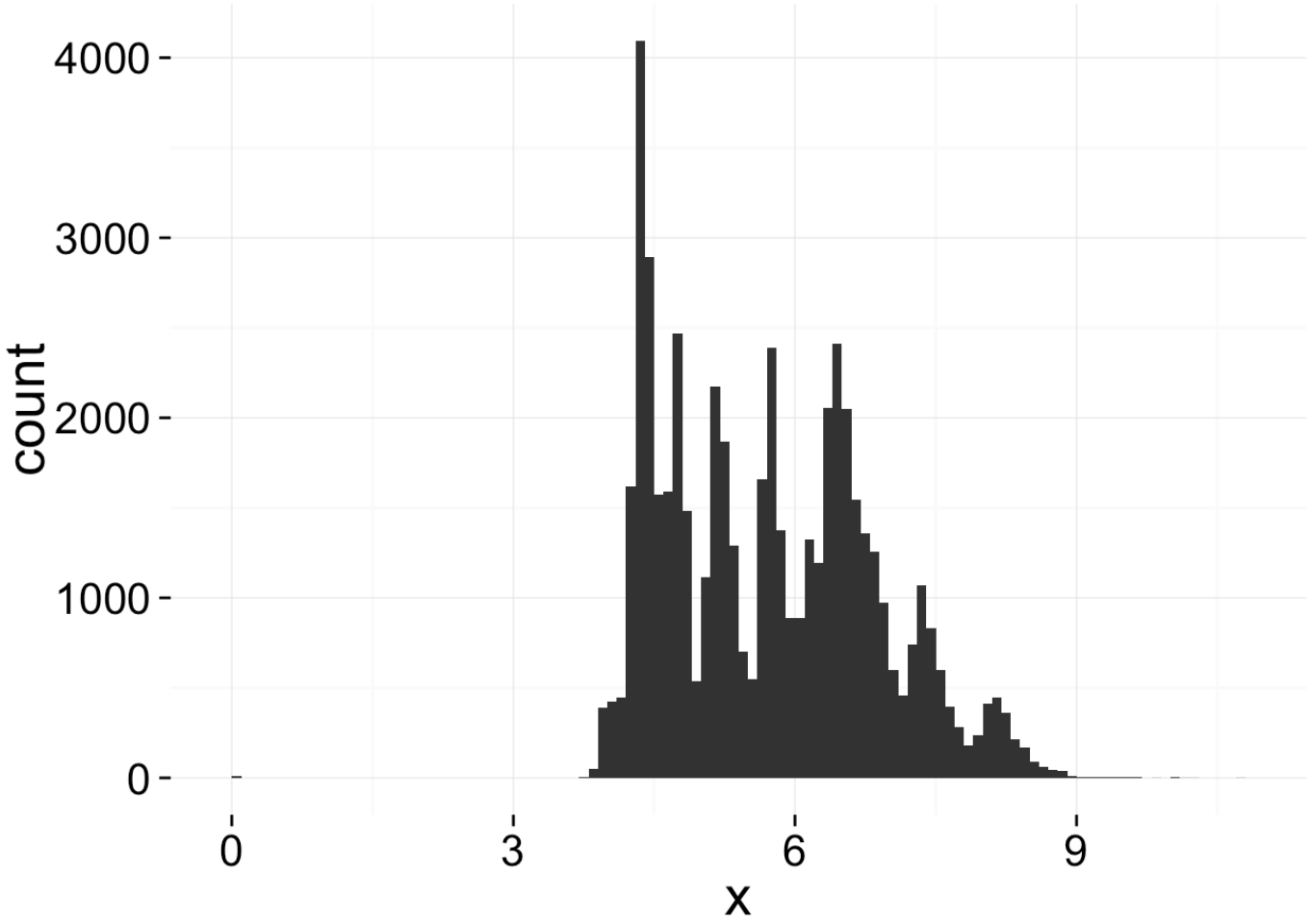


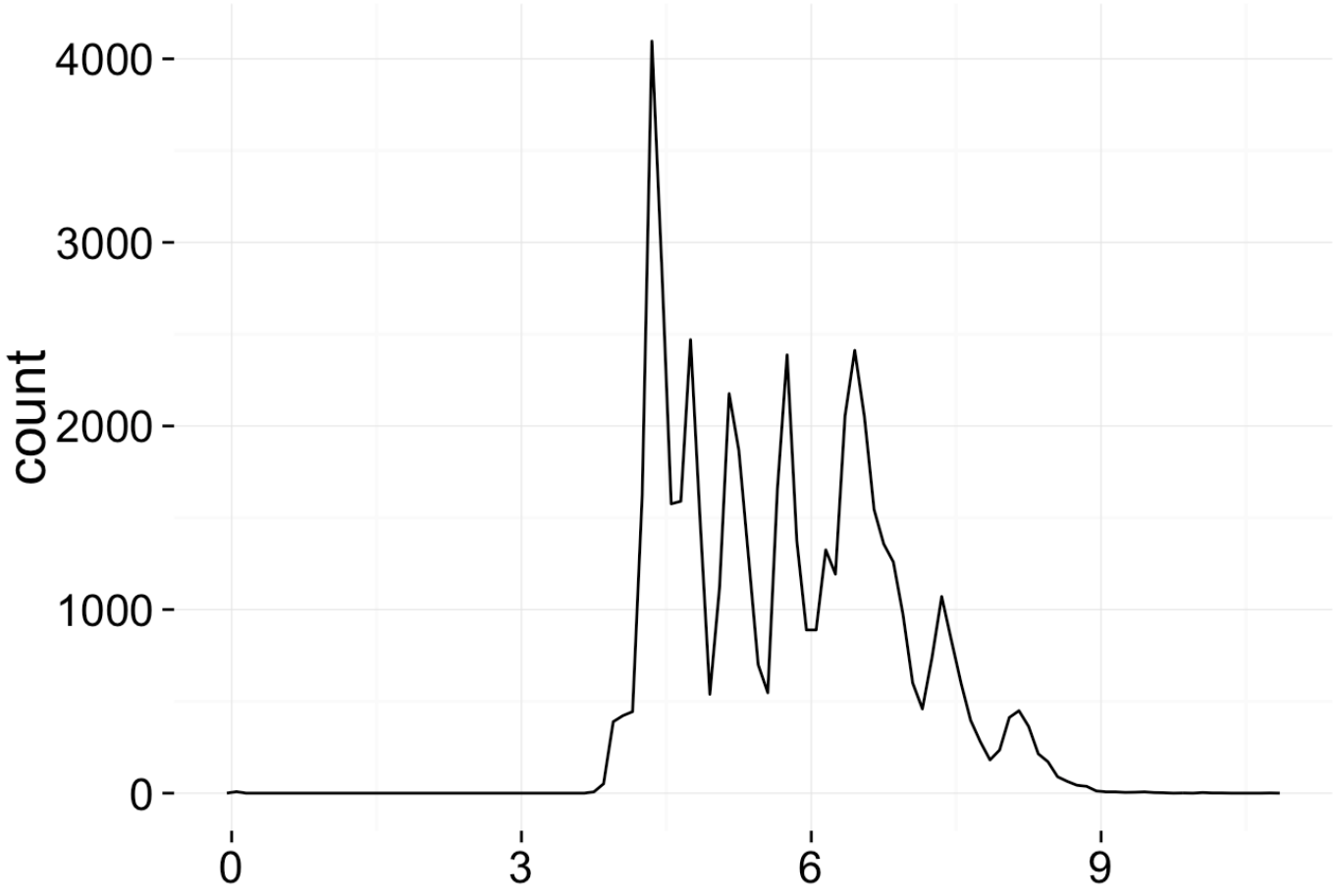

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.00   56.00   57.00   57.46   59.00   95.00
```

Setting the binwidth indicates that most table values are integers. Most diamonds have a table between 55 mm and 60 mm.

```
##
##   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
##   67 54.1 55.1 53.9 54.2 54.4 53.7 54.5 54.8 53.8 55.6 54.7   68 53.6 56.4
##   42   30   30   28   28   28   25   24   24   22   22   21   21   20   20
## 55.7 55.8 55.2 54.3 55.9 56.2 54.9 56.1 54.6 56.3 55.3 55.4 55.5 53.4 53.5
##   19   19   18   17   17   17   16   16   15   15   13   13   13   12   12
## 56.5 56.6 56.7 56.9 57.1 57.2 57.8 58.1 57.7 58.5 59.9 60.1 60.3 60.5   51
##   11   11   11   11   11   11   11   11   10   10   10   10   10   10    9
## 57.4 57.6 60.7   69   70 53.3 57.5 59.4 60.9 56.8 58.6 59.2 61.2 61.9 57.3
##    9    9    9    9    9    8    8    8    8    7    7    7    7    7    6
## 57.9 59.7 59.8 61.5 53.2 58.8 59.1 60.2 60.4 60.8 58.2 58.4 59.5 62.2 62.5
##    6    6    6    6    5    5    5    5    5    5    4    4    4    4    4
##   73 53.1 58.3 58.9 59.6 60.6 61.4   49   50 52.8 58.7 59.3 61.1 61.7 62.3
##    4    3    3    3    3    3    3    2    2    2    2    2    2    2    2
## 62.6 62.8   43   44 50.1 51.6 52.4 61.3 61.6 61.8 62.1 62.4 63.3 63.4 63.5
##    2    2    1    1    1    1    1    1    1    1    1    1    1    1    1
## 64.2 64.3 65.4   71   76   79   95
##    1    1    1    1    1    1    1
```

Again, I wonder if this has anything to do with the cut of a diamond. Cut is the quality of a diamons 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.


4000
3000
2000
1000
0
count
0
3
6
9
x

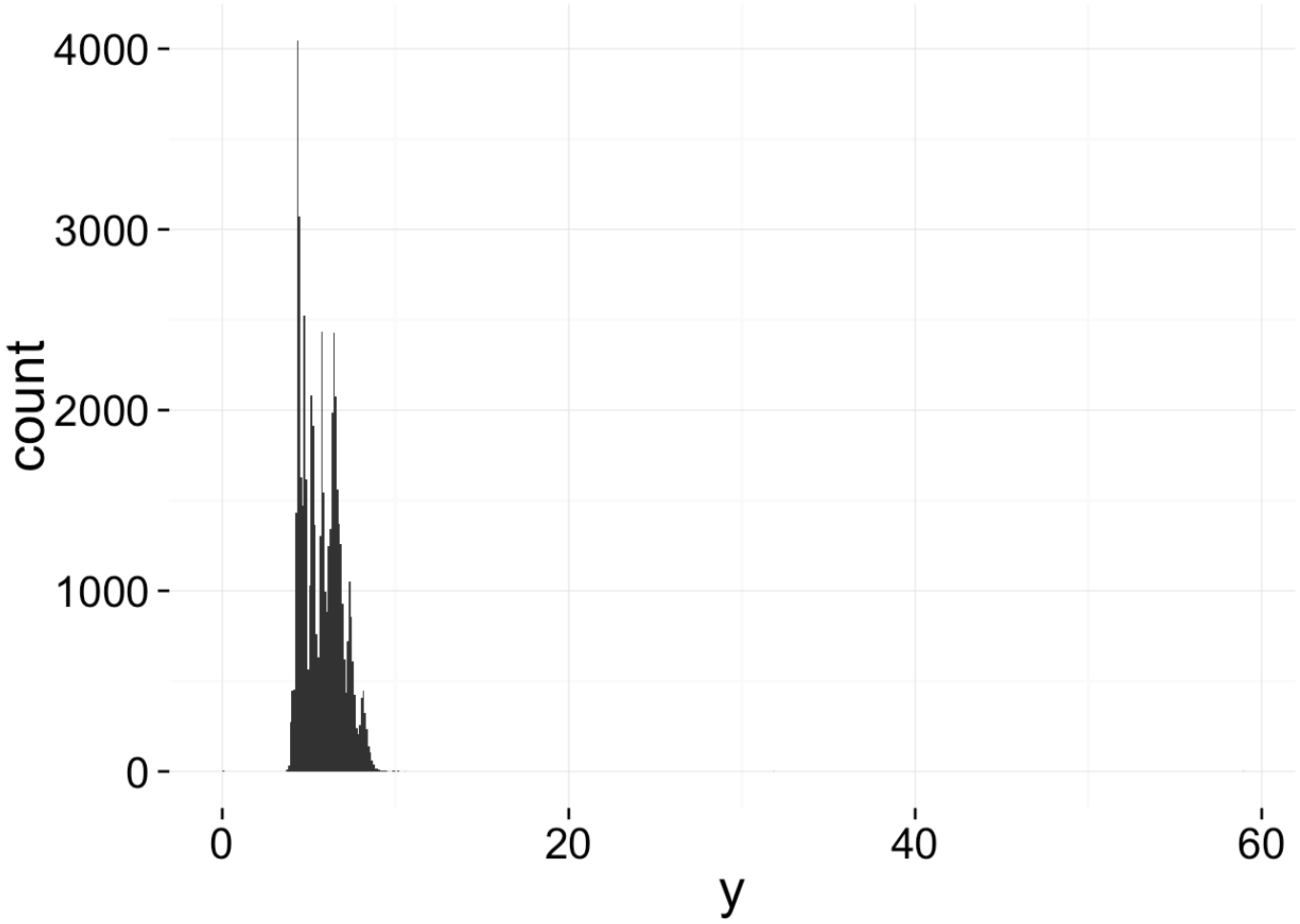

4000
3000
2000
1000
0
count
0
3
6
9

x

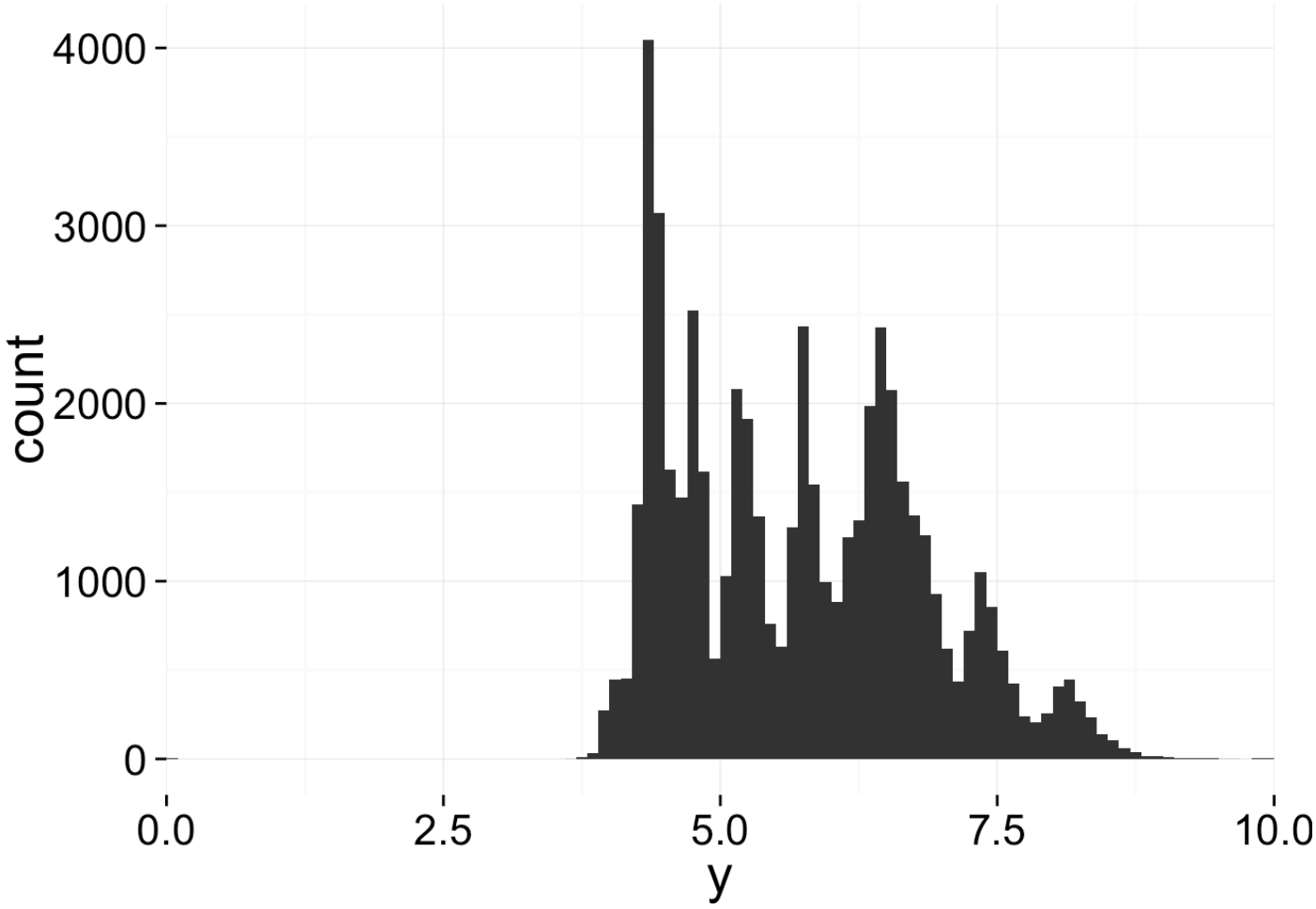
```
##
## FALSE  TRUE
## 53932     8
```

Most diamonds have an x dimension between 4 mm and 7 mm. 8 diamonds have a x dimension of 0.

```
## Warning: position_stack requires constant width: output may be incorrect
```



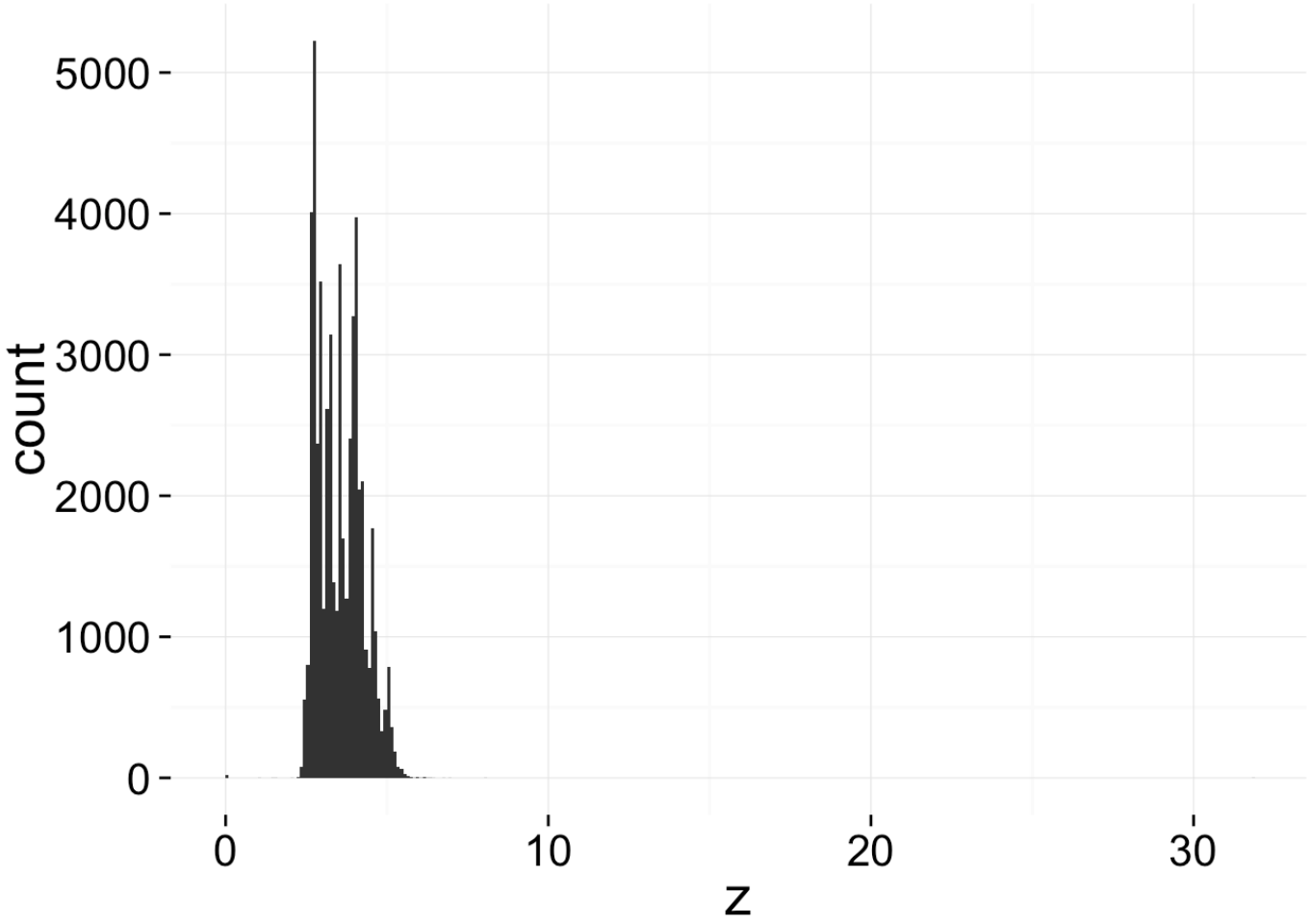

```
## Warning: position_stack requires constant width: output may be incorrect
```



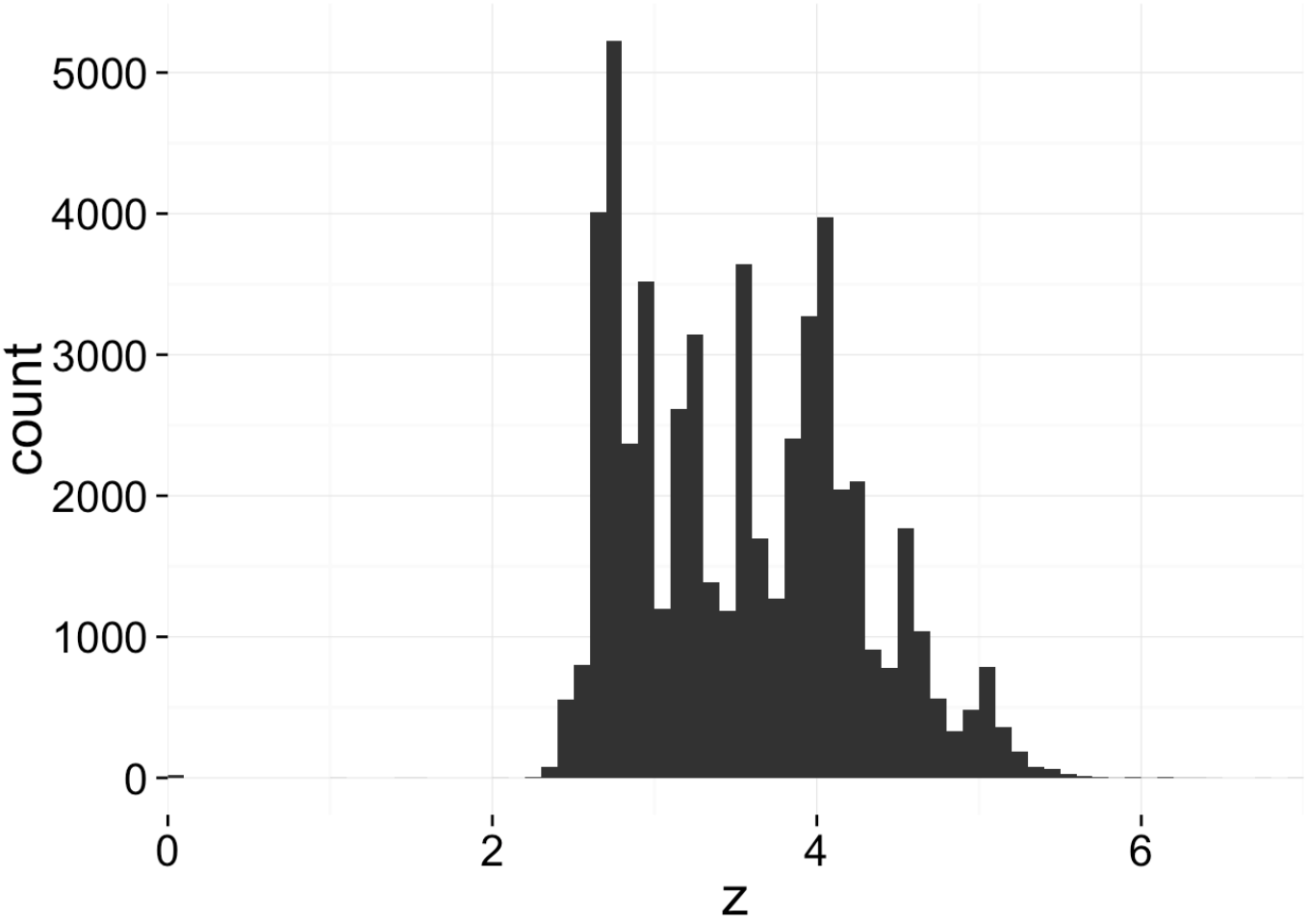

```
##
## FALSE  TRUE
## 53933     7
```

Again, most diamonds have a y dimension between 4 mm and 7 mm. There are some outliers for the y dimension. 7 diamonds have a y dimension of 0.

```
## Warning: position_stack requires constant width: output may be incorrect
```




```
## Warning: position_stack requires constant width: output may be incorrect
```




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

Most diamonds have a z dimension between 2 mm and 6 mm. There are some outliers for the z dimension too. 20 diamonds have a z dimension of 0.

```
##    carat       cut color clarity depth table price    x    y z
## 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
```

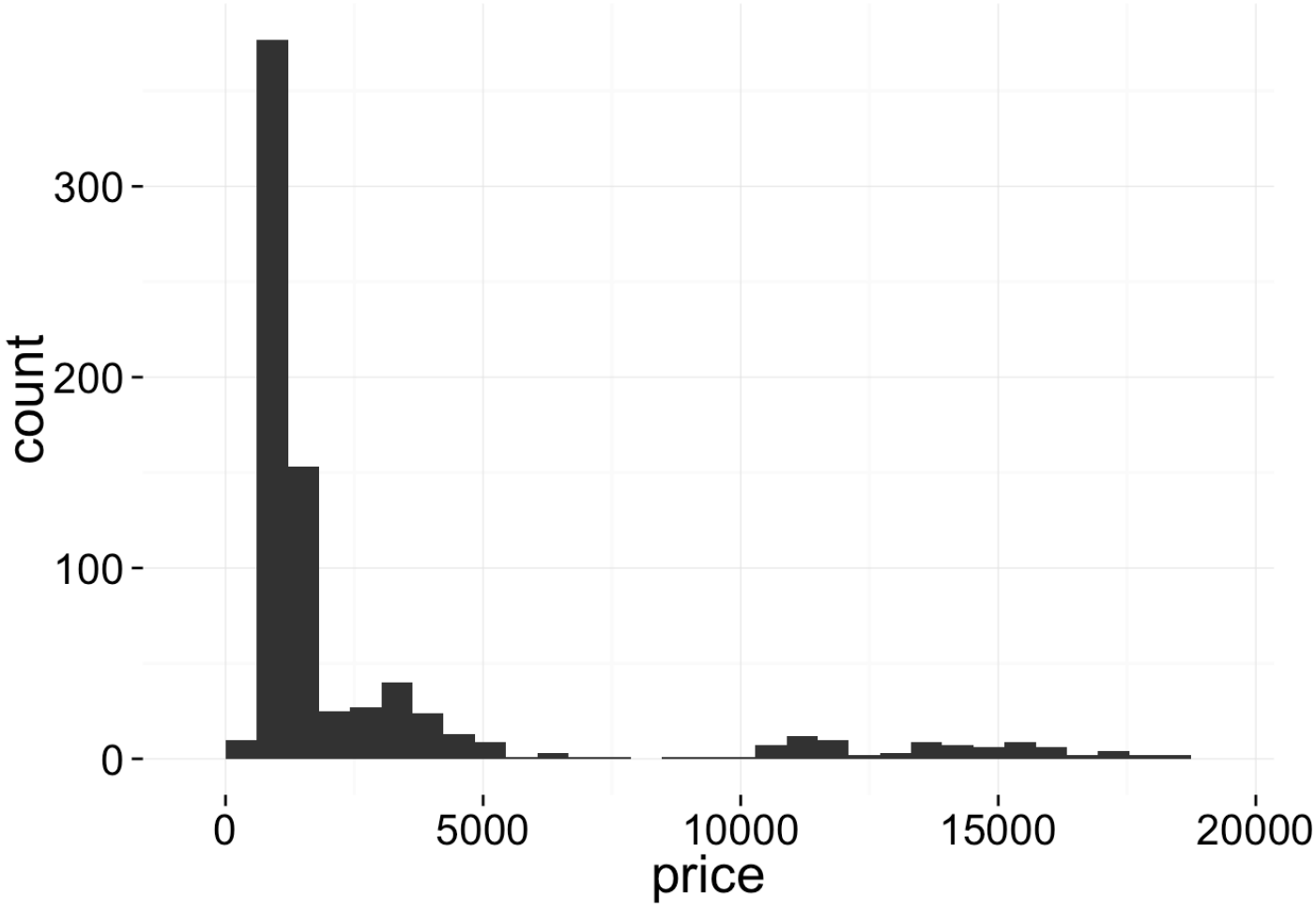
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2130    3564    5352    8803   15470   18790
```

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

The above diamonds have missing dimension values. If and only if x or y dimensions are 0, then the z dimension is 0.

The diamonds in this subset 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.

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

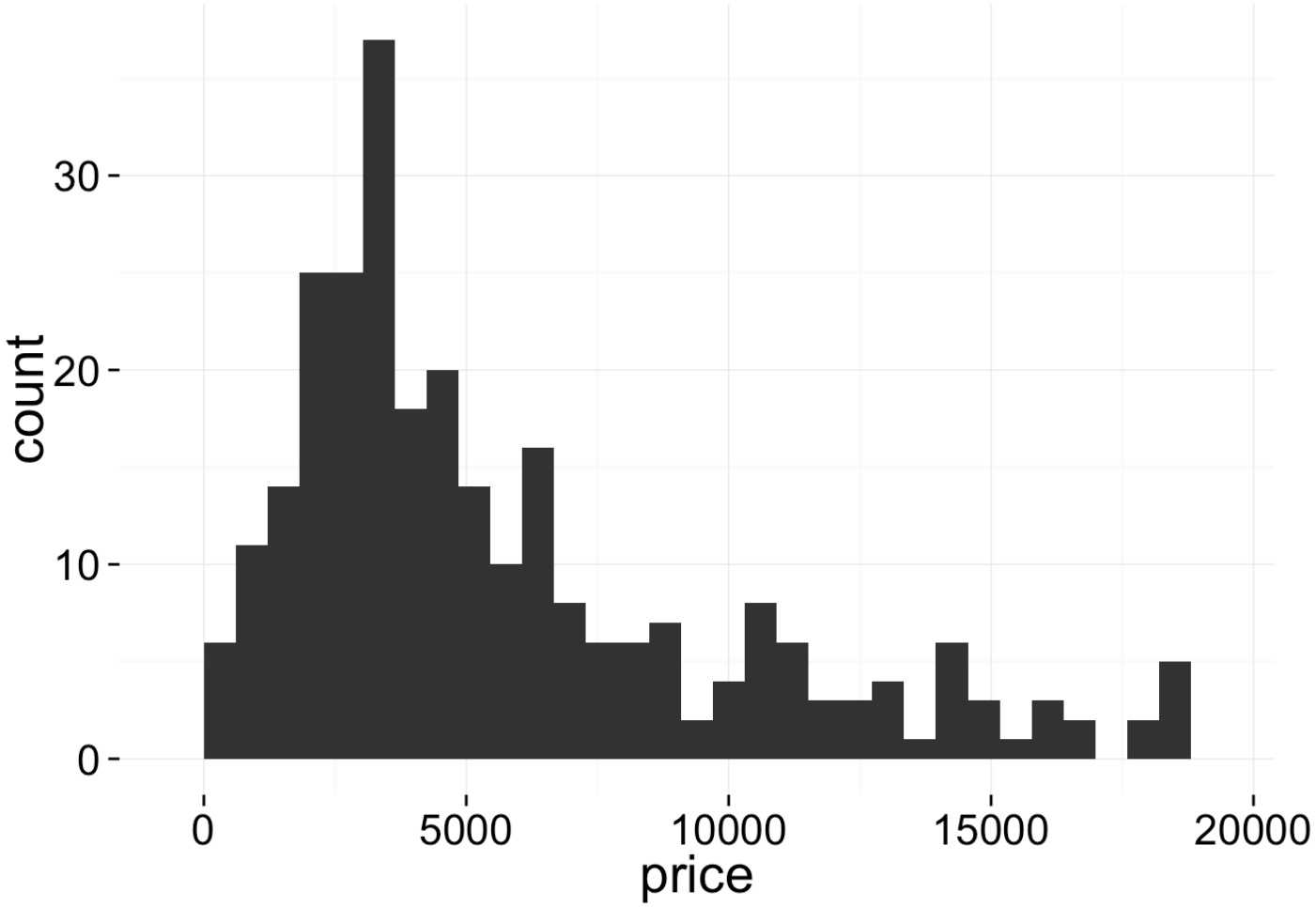



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     553     967    1207    2887    2644   18700
```

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

I'm going to compare the worst diamonds across the same variables.

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
##    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
```

This doesn't add much to my thoughts already. 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
```

```
##
## FALSE
## 53940
```

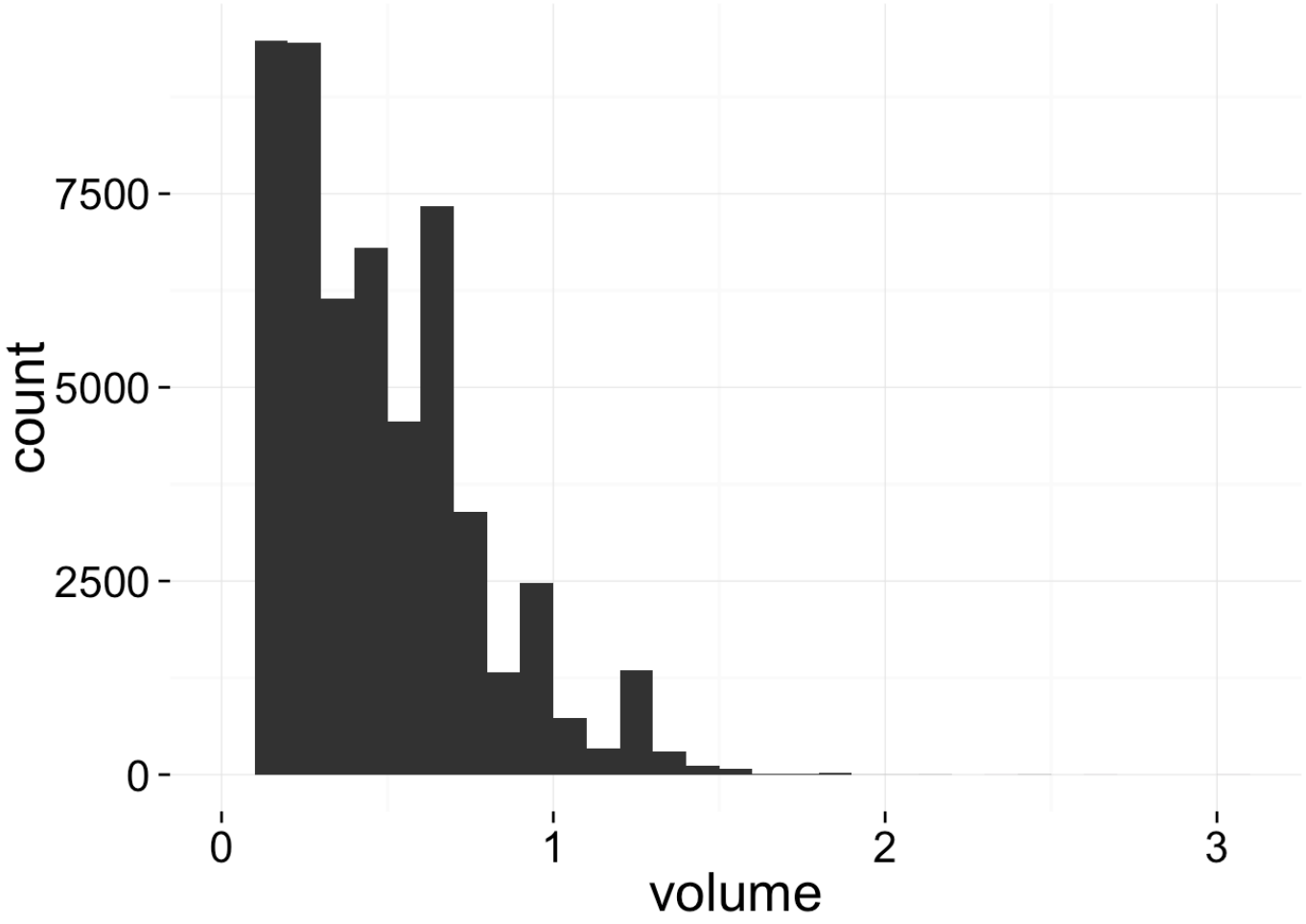
Some diamonds have a volume of 0 since they have at least one dimension with a value of 0. I'm going to use the average density of diamonds to compute the volume of a diamond instead of using the dimensions x, y, and z to compute the volume.

I can convert carat to grams and then divide by the density to get the volume of a diamond.

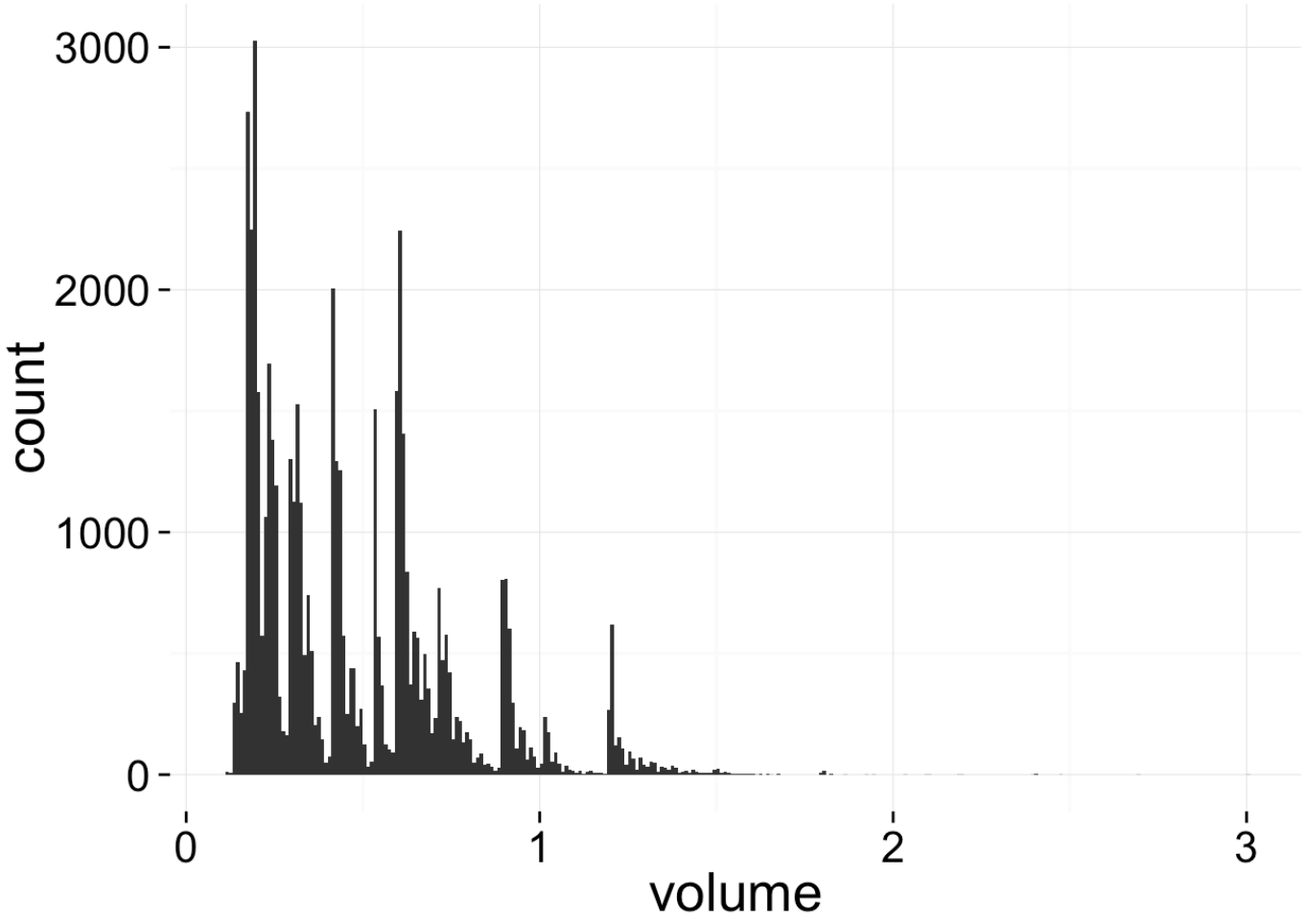
1 carat is equivalent to 2 grams

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 average density 3.34 g/cm^3 to estimate the volume of the diamonds.

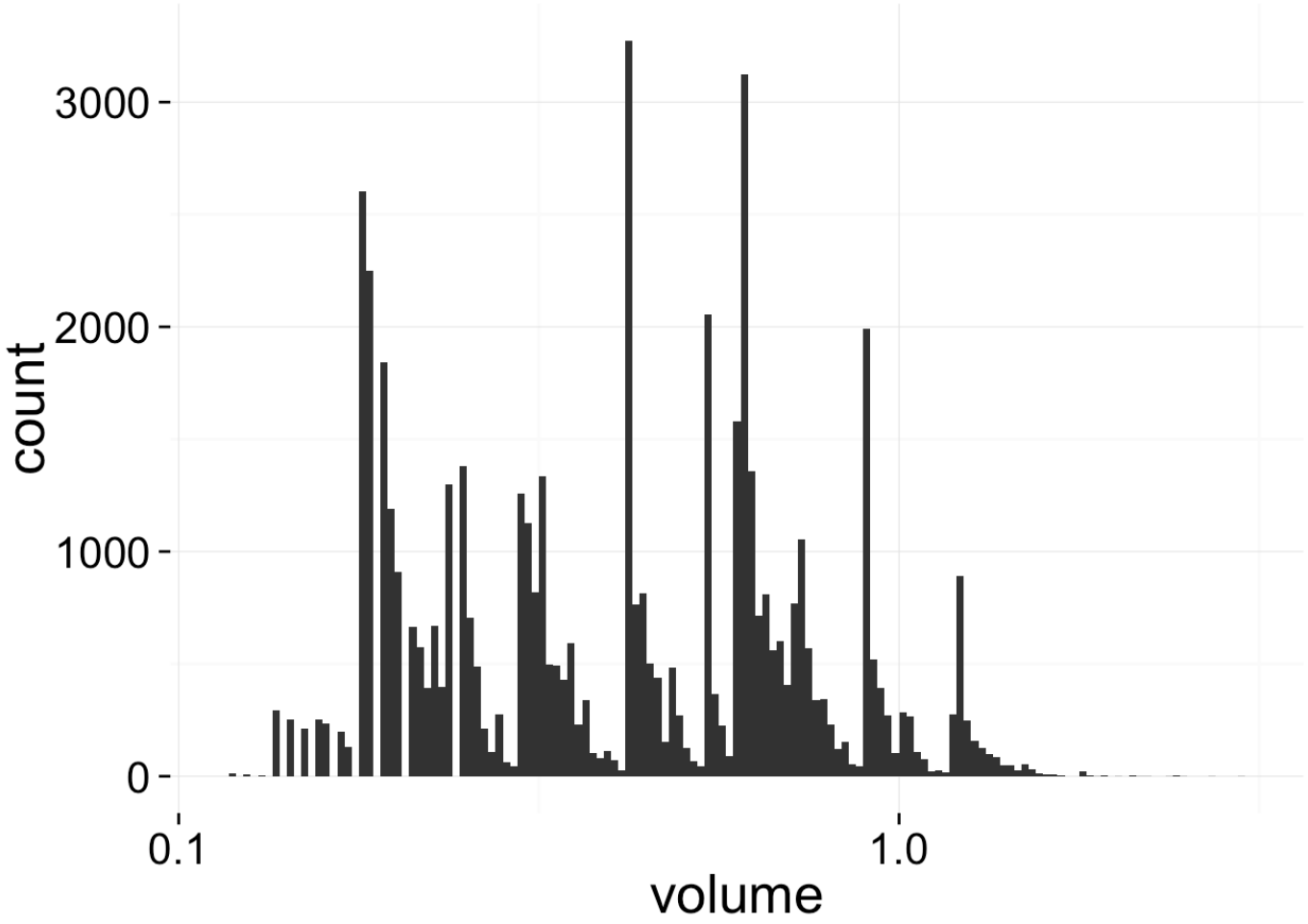
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1198  0.2395  0.4192  0.4778  0.6228  3.0000
```




```
## Warning: position_stack requires constant width: output may be incorrect
```




```
## Warning: position_stack requires constant width: output may be incorrect
```



The histogram of volume is right skewed so I'm going to transform the data using a log transform. There are some volumes that are more common than others.

```
##
##  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
##  1127   910   883   817   807   793   764   709
```

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.

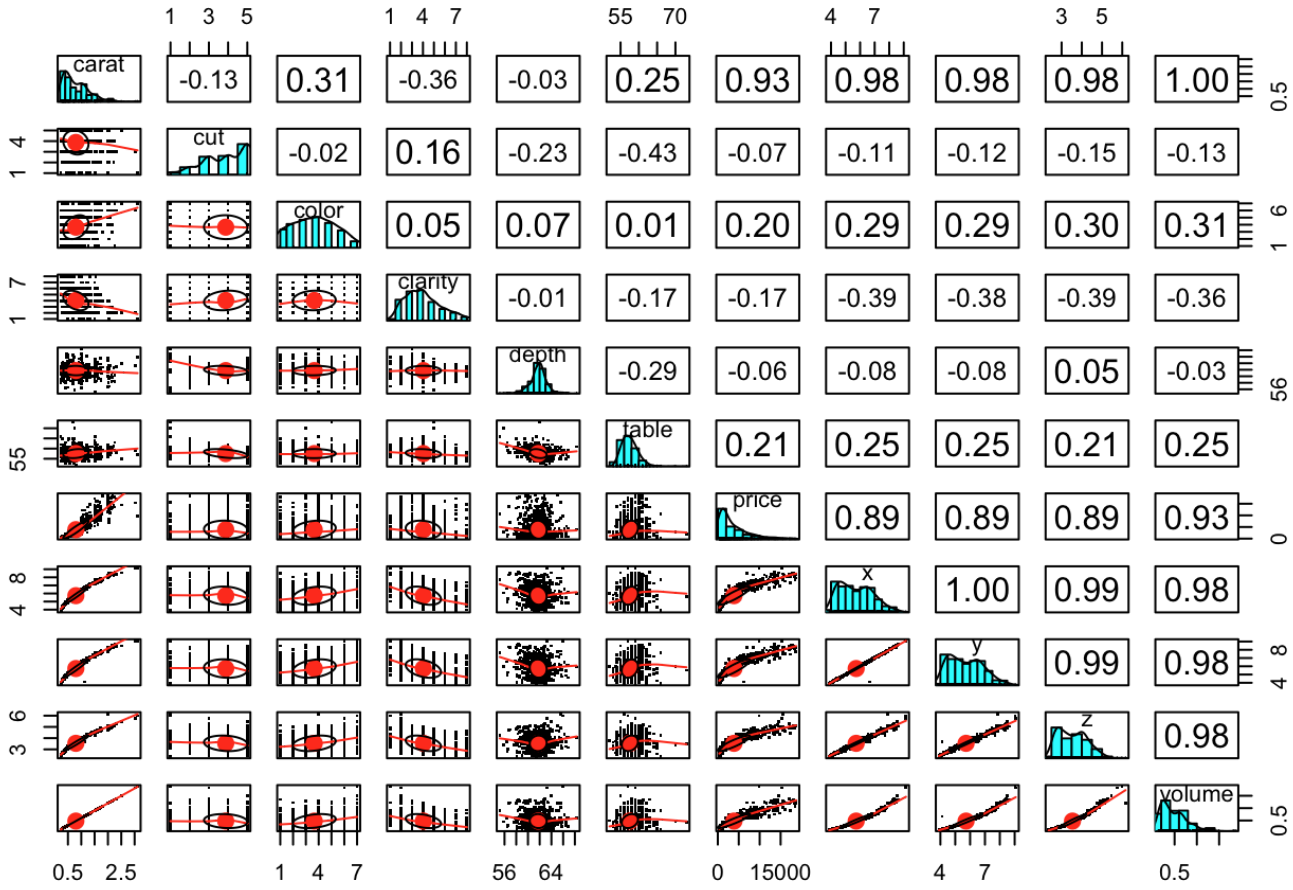
When first calculating the volume using x, y, and z, some volumes were 0 or could not be calculated because data was missing. Additionally, some values for the dimensions x, y, and z seemed too large. In the subset called noVolume, all dimensions (x, y, and z) are missing or the z value is 0. The diamonds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

this subset tend to be very expensive or fall in the third quartile of the entire diamonds data set.

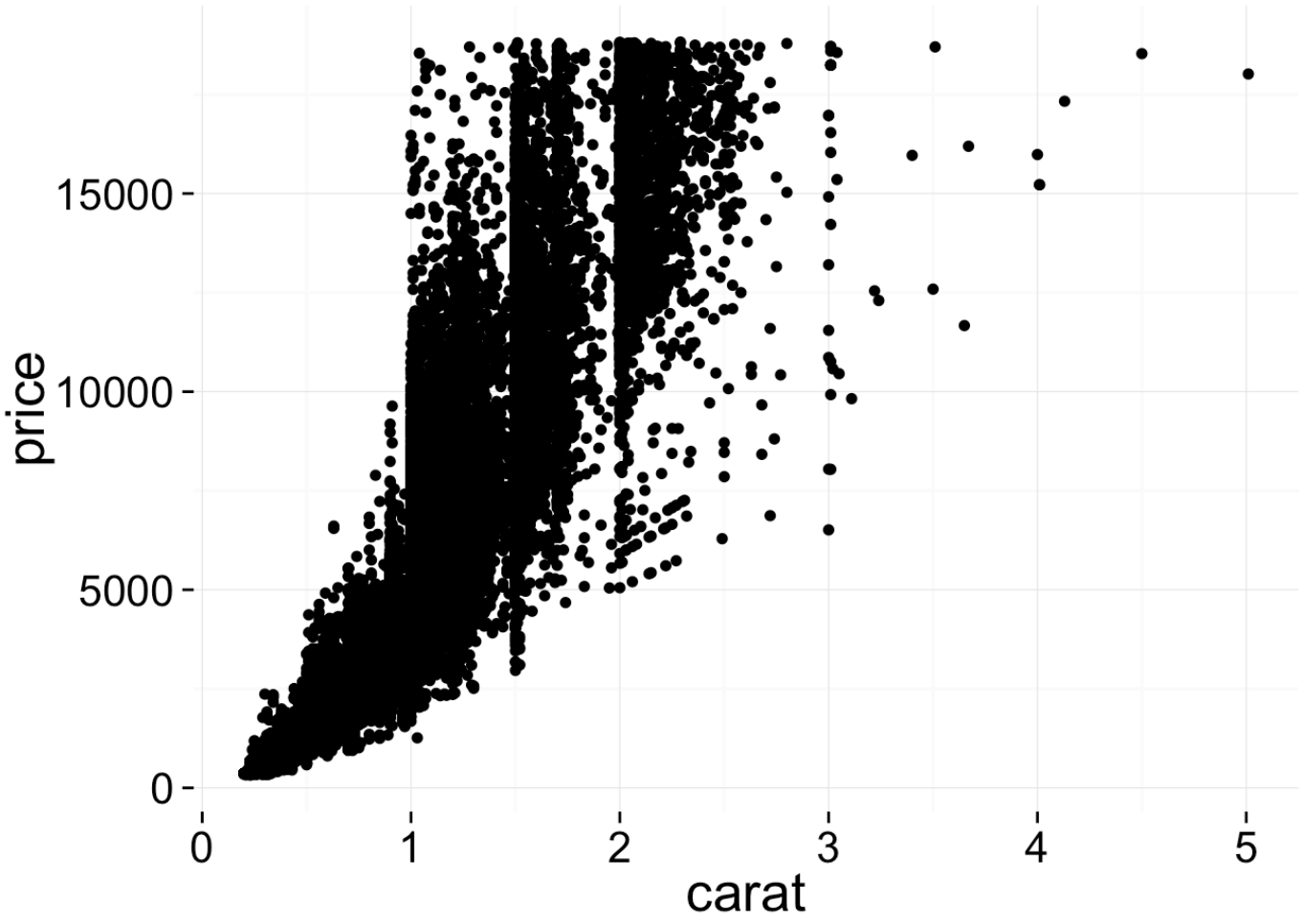
Bivariate Plots Section

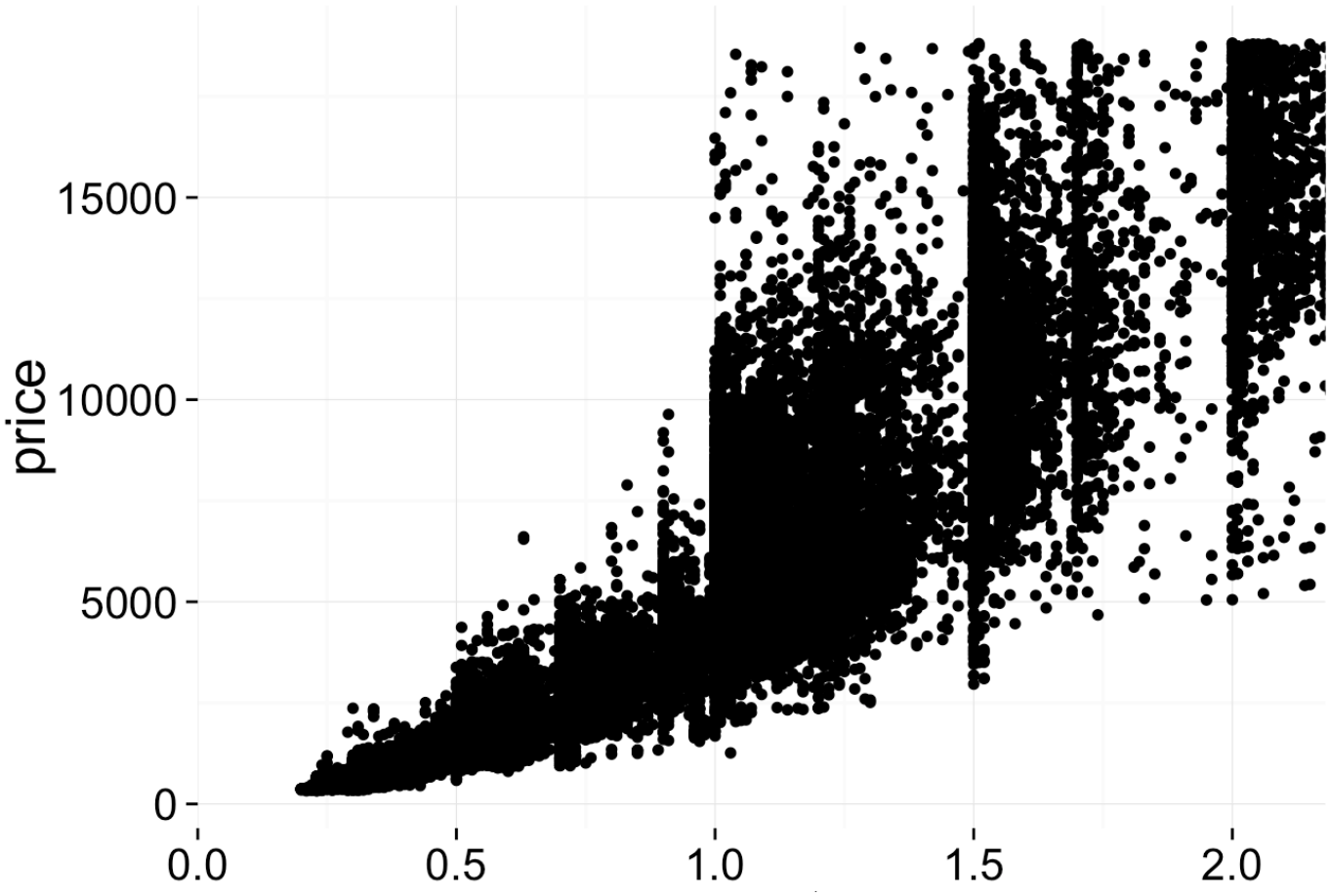
```
##              carat       depth      table      price           x
## carat   1.00000000  0.02822431  0.1816175  0.9215913  0.97509423
## depth   0.02822431  1.00000000 -0.2957785 -0.0106474 -0.02528925
## table   0.18161755 -0.29577852  1.0000000  0.1271339  0.19534428
## price   0.92159130 -0.01064740  0.1271339  1.0000000  0.88443516
## x       0.97509423 -0.02528925  0.1953443  0.8844352  1.00000000
## y       0.95172220 -0.02934067  0.1837601  0.8654209  0.97470148
## z       0.95338738  0.09492388  0.1509287  0.8612494  0.97077180
## volume  1.00000000  0.02822431  0.1816175  0.9215913  0.97509423
##                 y          z     volume
## carat   0.95172220 0.95338738 1.00000000
## depth  -0.02934067 0.09492388 0.02822431
## table   0.18376015 0.15092869 0.18161755
## price   0.86542090 0.86124944 0.92159130
## x       0.97470148 0.97077180 0.97509423
## y       1.00000000 0.95200572 0.95172220
## z       0.95200572 1.00000000 0.95338738
## volume  0.95172220 0.95338738 1.00000000
```

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



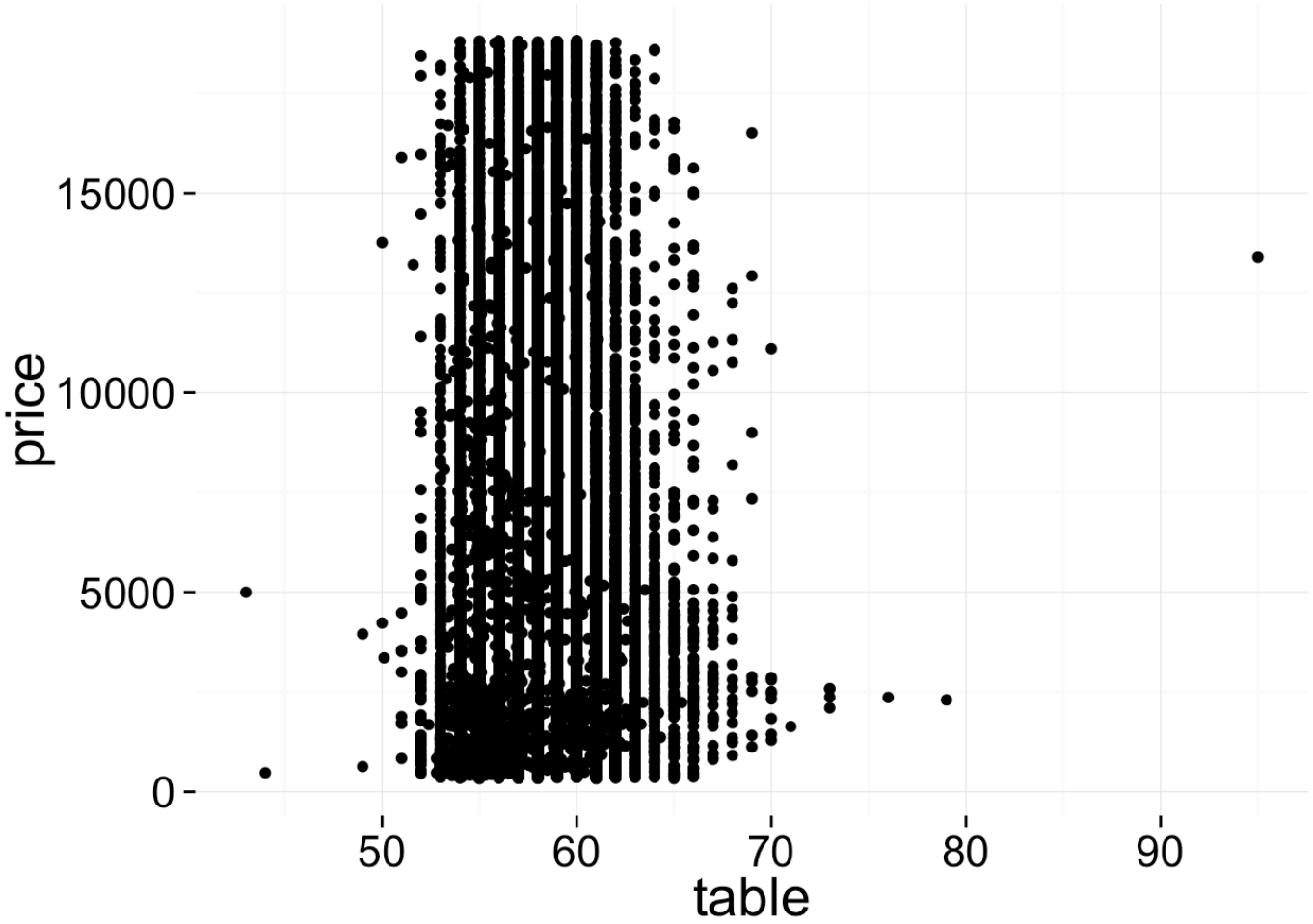
I want to look closer at scatter plots involving price and some other variables: carat, table, depth, and volume.


price
15000
10000
5000
0
0
1
2
3
4
5
carat


price
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0

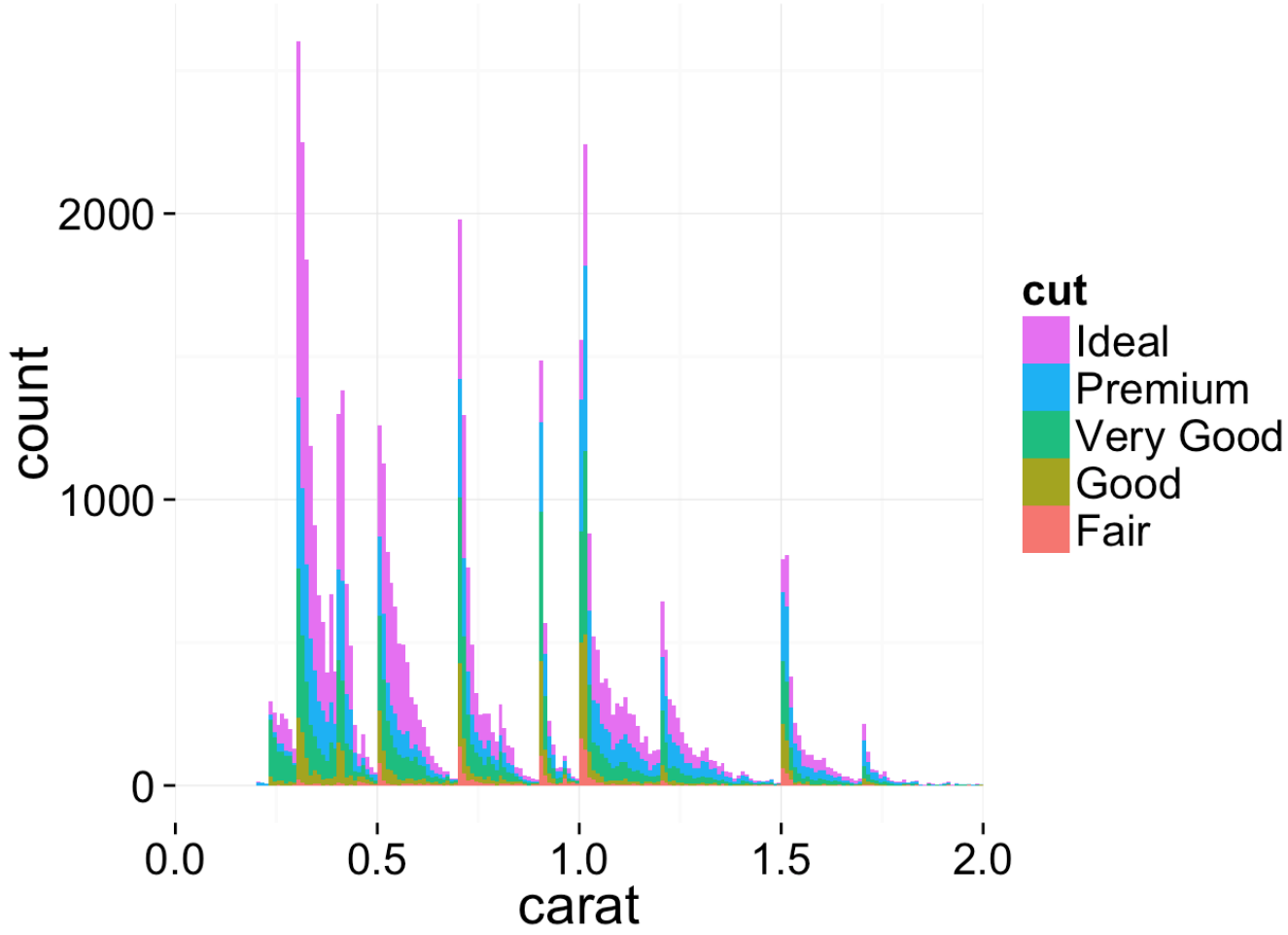
carat

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.



Again, the tall vertical strips indicate table values are mostly integers. A few outliers below 50 mm and one above 90 mm.

```
## Warning: position_stack requires constant width: output may be incorrect
```

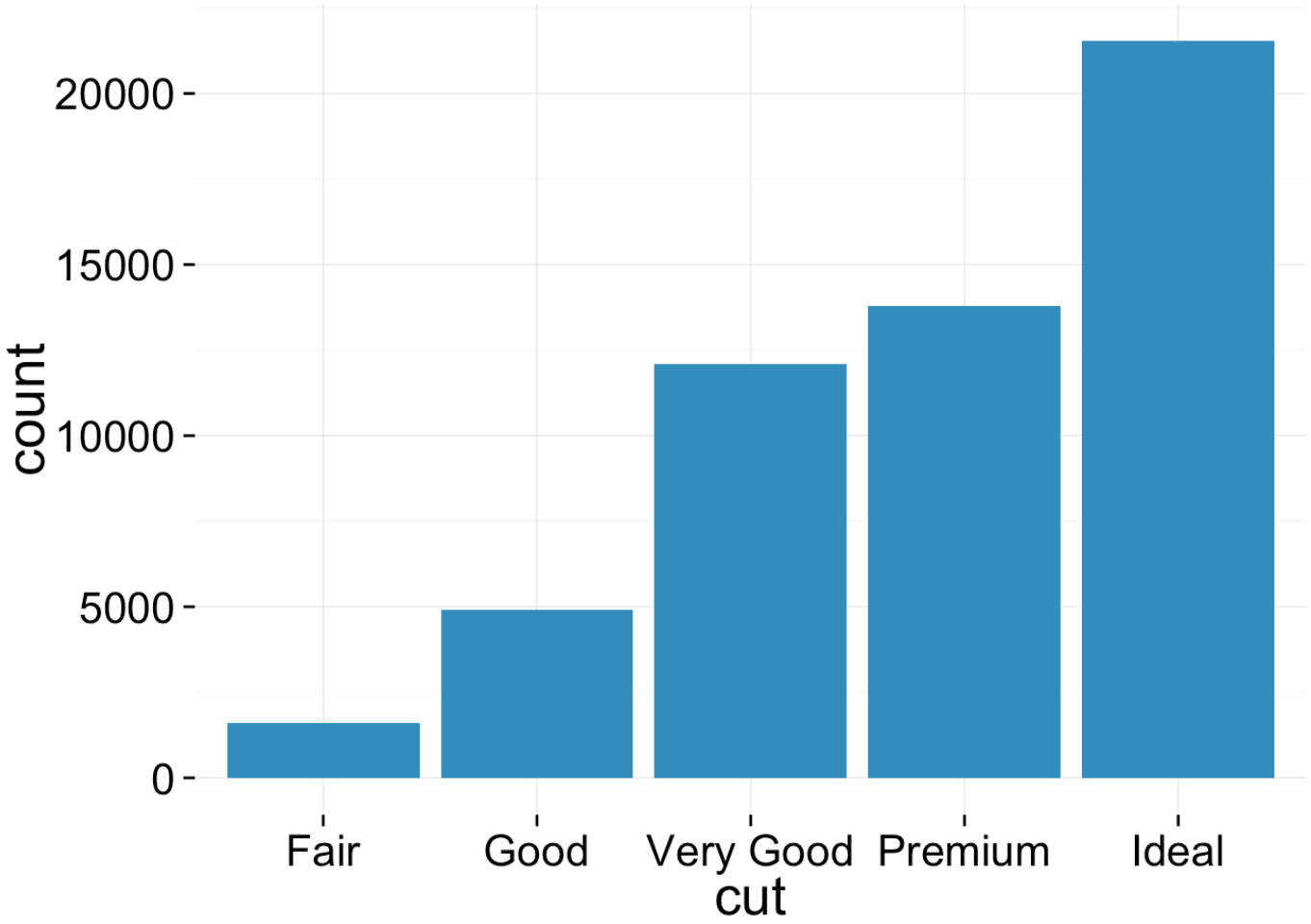



```
## cut: Fair
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.220   0.700   1.000   1.046   1.200   5.010
## ------------------------------------------------------------
## cut: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.5000  0.8200  0.8492  1.0100  3.0100
## ------------------------------------------------------------
## 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
```

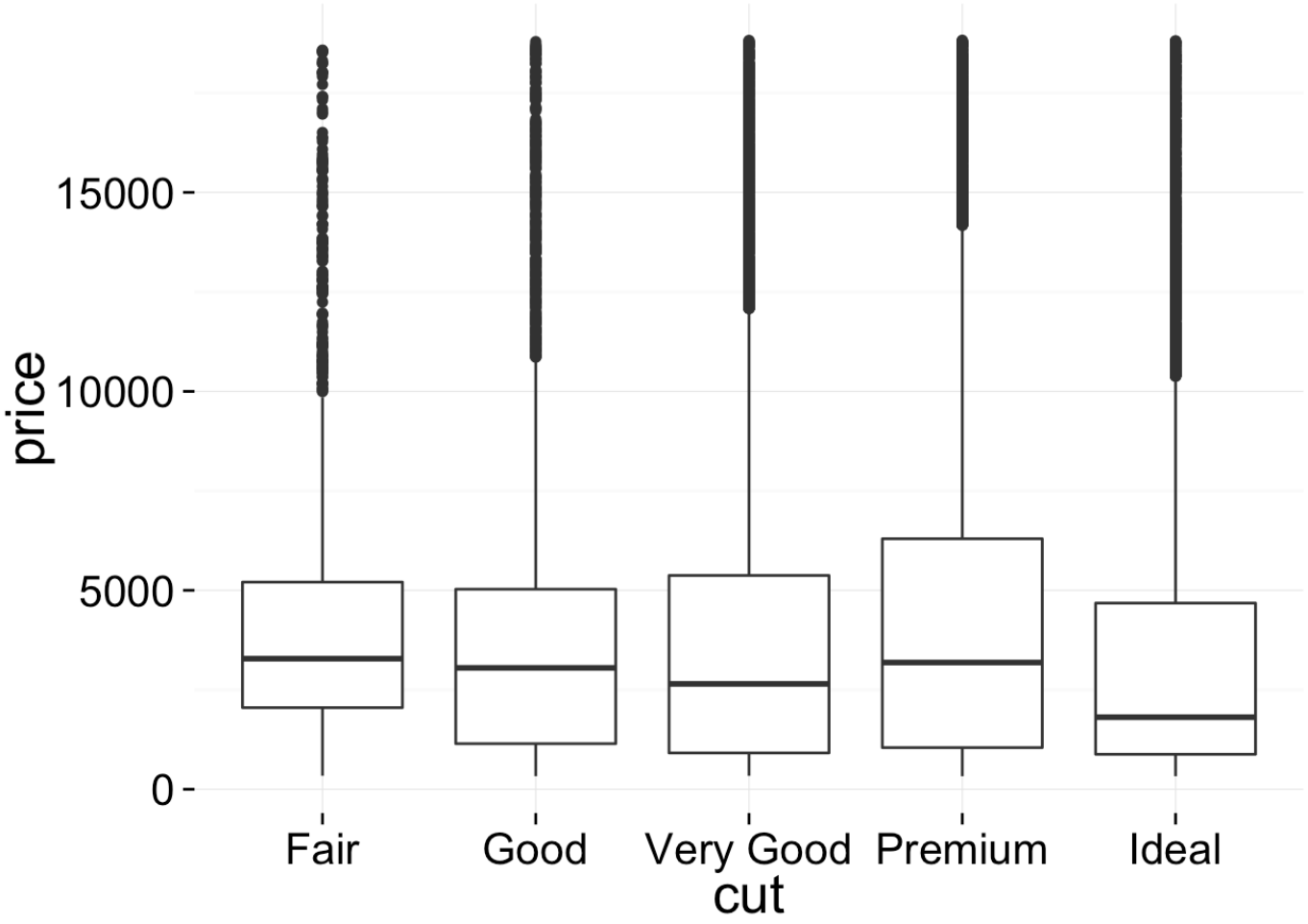
It doesn't look like particular cuts have a certain number of carats. It looks like most of the ideal cut diamonds are less than one carat. I'm going to look at those values to be sure.

```
##
##  0.3 0.31 0.32 0.33 0.41  0.7  0.4 0.51 0.34 0.71 0.52 0.53 1.01  0.5 0.42
## 1247 1209 1066  673  667  560  545  525  508  499  459  429  426  388  387
## 0.38 0.35 0.54 0.72 0.56 0.36 0.55 1.02 0.73 0.57 1.03 0.43  0.9    1  1.2
##  378  374  372  366  313  309  302  272  243  239  225  221  215  208  194
## 1.04 0.58 1.51 0.39 1.06 0.37 1.21 1.07 1.05 1.09 0.59 0.74 1.11 1.23 0.76
##  187  182  182  181  174  172  160  156  147  141  139  135  130  127  122
## 1.08 1.22  1.1  1.5 0.27 1.13 1.52  0.6  0.8 0.91 0.26 0.61 1.12 0.44 1.24
##  121  121  118  117  113  113  109  108  108  108  106  105  103   96   96
## 0.77 0.75 1.14 0.81 1.16 0.46 0.78 0.28 2.01 1.25 1.53 0.24 1.26 0.25 0.79
##   92   90   88   87   87   83   82   81   78   77   74   69   69   66   66
## 0.82 0.62 0.83  1.7 1.15 1.17 1.27 1.55 0.92 1.54 0.29 1.18 1.31 0.63 1.19
##   60   59   59   59   56   56   56   56   55   53   52   51   51   49   49
## 1.28 1.57 2.02 1.56 0.23 0.45 0.47 1.58 1.29  1.6 0.64    2  1.3 2.03 1.71
##   46   46   46   45   44   42   42   42   40   40   39   39   38   38   37
## 0.93 1.59 1.32 0.85 1.34 1.33 1.35 0.65 1.61 1.62 0.48 1.63 0.66 0.84 1.37
##   35   35   33   31   30   28   28   27   26   25   23   23   21   21   19
## 1.75 2.07 0.96 0.97 1.36 2.05 2.04 2.06 0.87 1.74 0.95 1.65 1.67  2.1 1.39
##   18   18   17   17   17   17   16   16   15   15   14   14   14   14   13
##  1.4 1.64 0.86 2.08 2.09 2.14  2.2 0.94 1.38 1.66 1.68  1.8 2.16  2.3 0.88
##   13   13   12   12   12   12   12   11   11   11   11   11   11   11   10
## 1.41 1.72 1.76 2.11 2.15 1.42 1.69 0.67 1.43 2.12 2.18 2.22 1.73 2.24  2.4
##   10   10   10   10   10    9    9    8    8    8    8    8    7    7    7
## 0.49 0.89 2.13 2.19 2.21 2.28 2.36 0.69 0.98 0.99 2.17 2.25 2.26 2.32 1.77
##    6    6    6    6    6    6    6    5    5    5    5    5    5    5    4
## 1.79 2.37  2.5  0.2 0.68 1.44 1.46 1.49 1.83 1.87 1.91 2.27 2.29 2.51 2.54
##    4    4    4    3    3    3    3    3    3    3    3    3    3    3    3
## 1.78 1.85 1.89  1.9 1.98 2.33 2.39 2.45 2.46 2.53 2.61 2.72 3.01 1.45 1.47
##    2    2    2    2    2    2    2    2    2    2    2    2    2    1    1
## 1.48 1.82 1.84 1.86 1.92 1.93 2.34 2.41 2.42 2.43 2.47 2.48 2.49 2.52 2.56
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 2.59  2.6 2.63 2.64 2.75 3.22  3.5
##    1    1    1    1    1    1    1
```

Most ideal cut diamonds are under 1.25 carats.



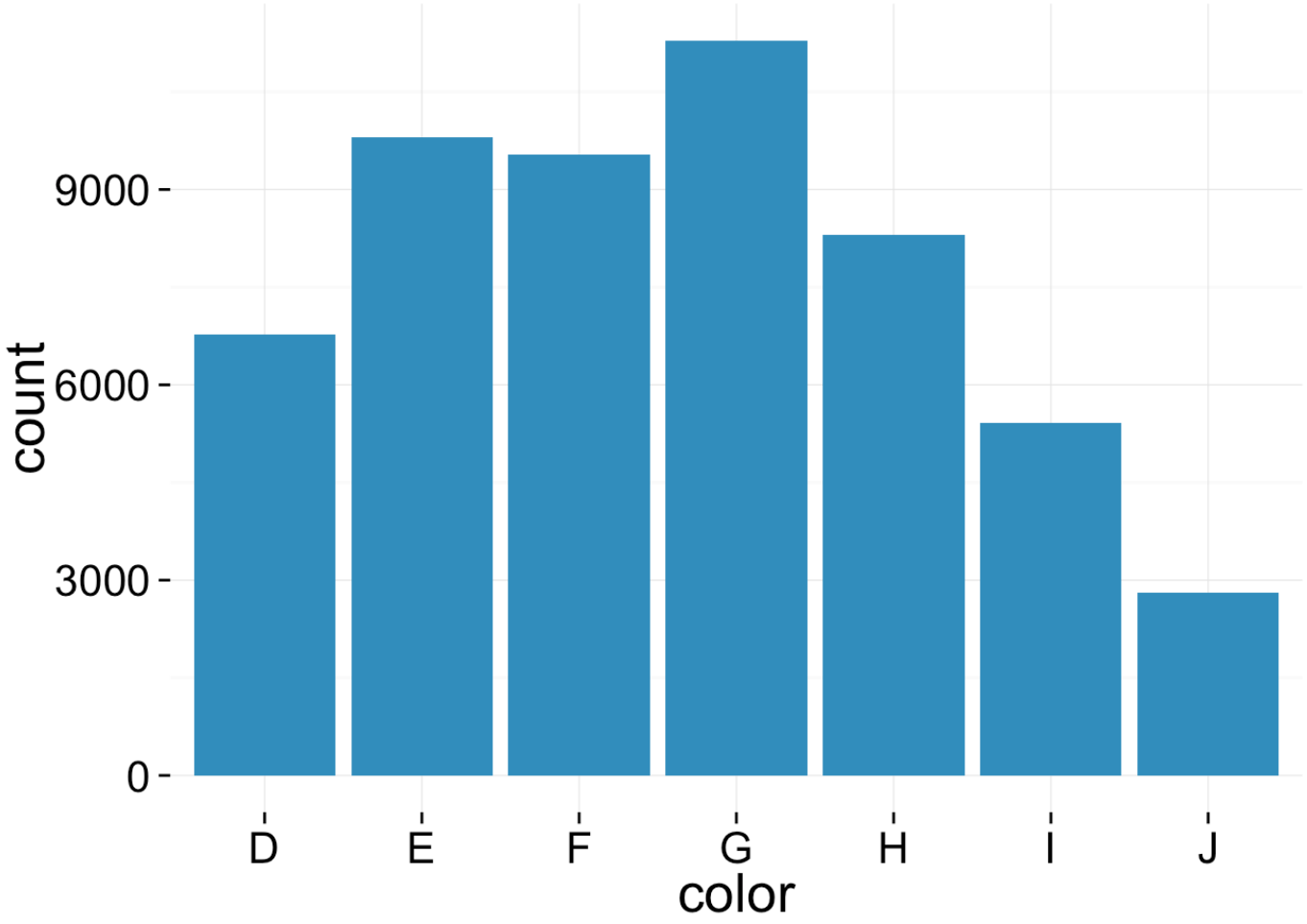
Most diamonds have ideal cut, which is almost double the amount of very good cut diamonds.



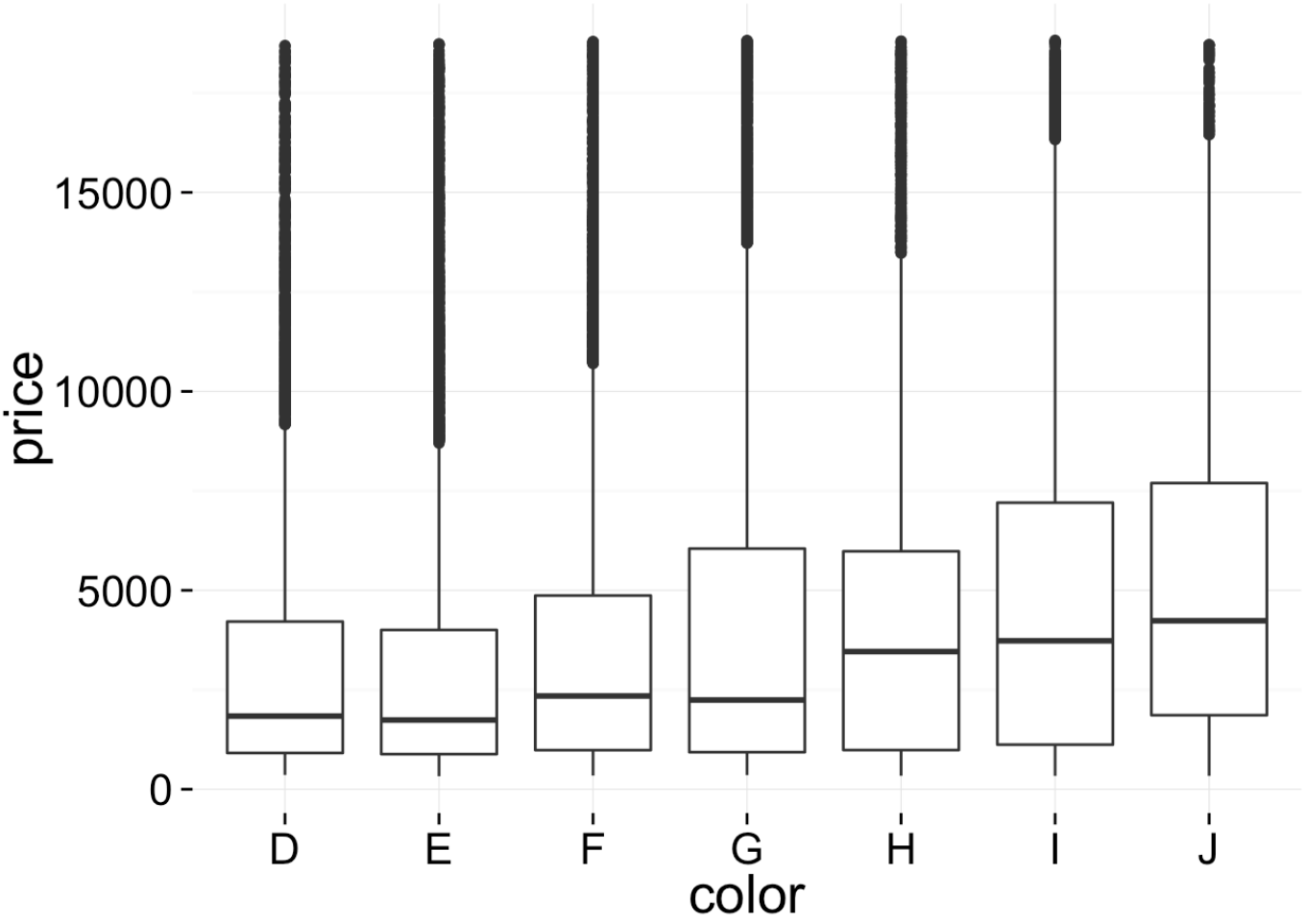

```
## diamonds$cut: Fair
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     337    2050    3282    4359    5206   18570
## ------------------------------------------------------------
## diamonds$cut: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     327    1145    3050    3929    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 about price/carat for these cuts?

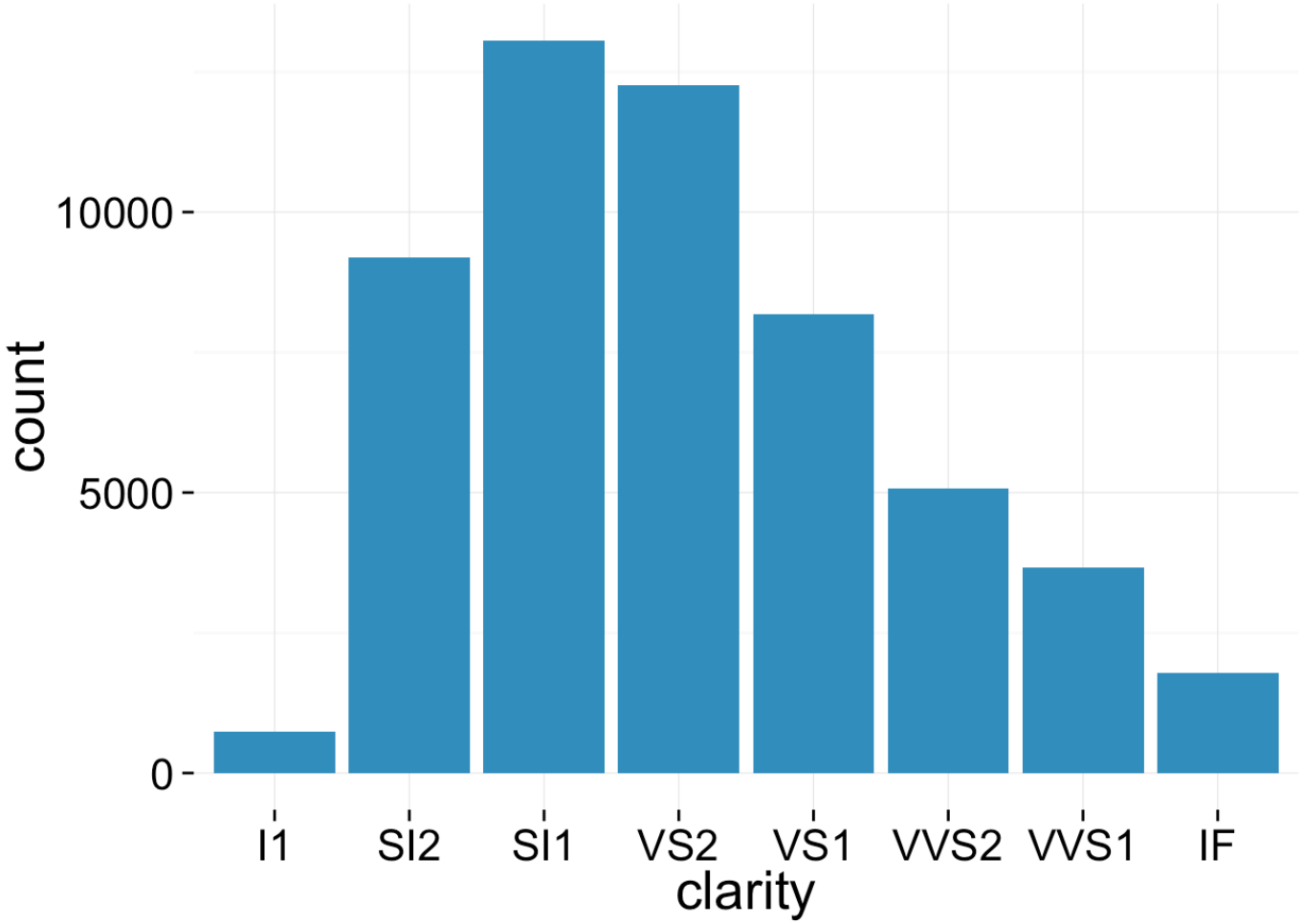


Most diamonds have have color ratings between E and H.

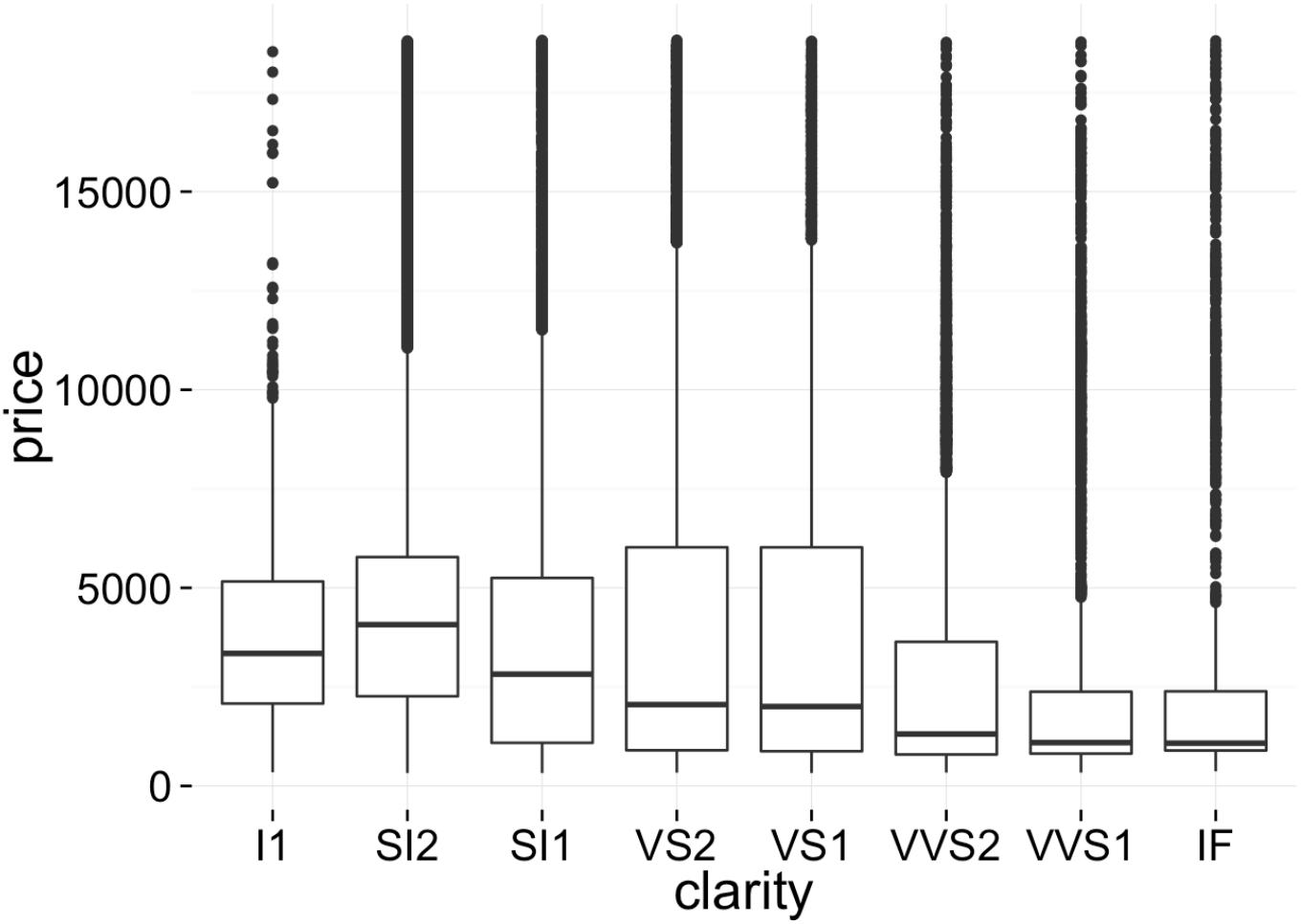

price
15000
10000
5000
0
D
E
F
G
H
I
J
color

```
## 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.



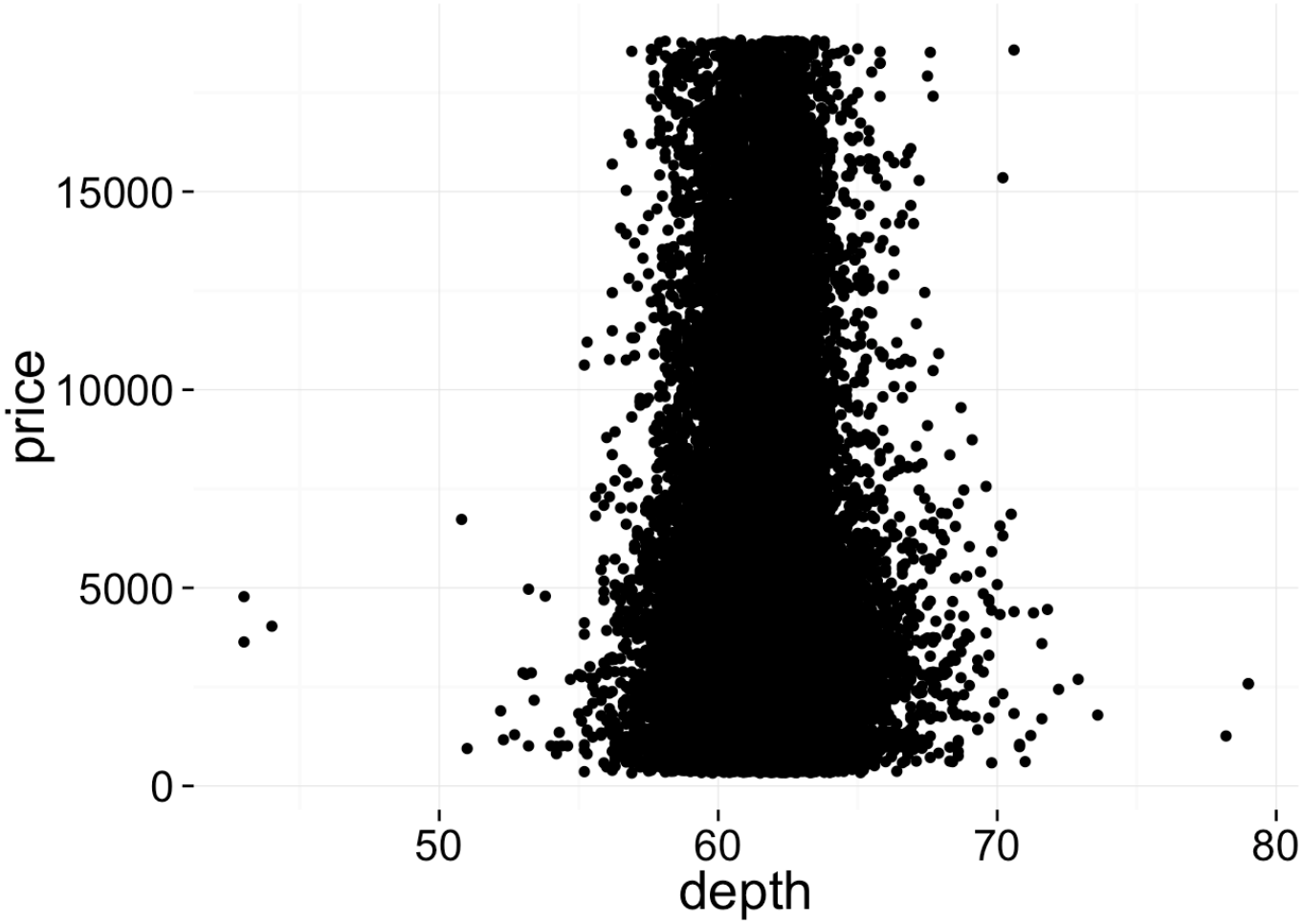
Most diamonds have average clarity ratings. Very few diamonds have the worst or best clarity rating, like the rating pattern for color.

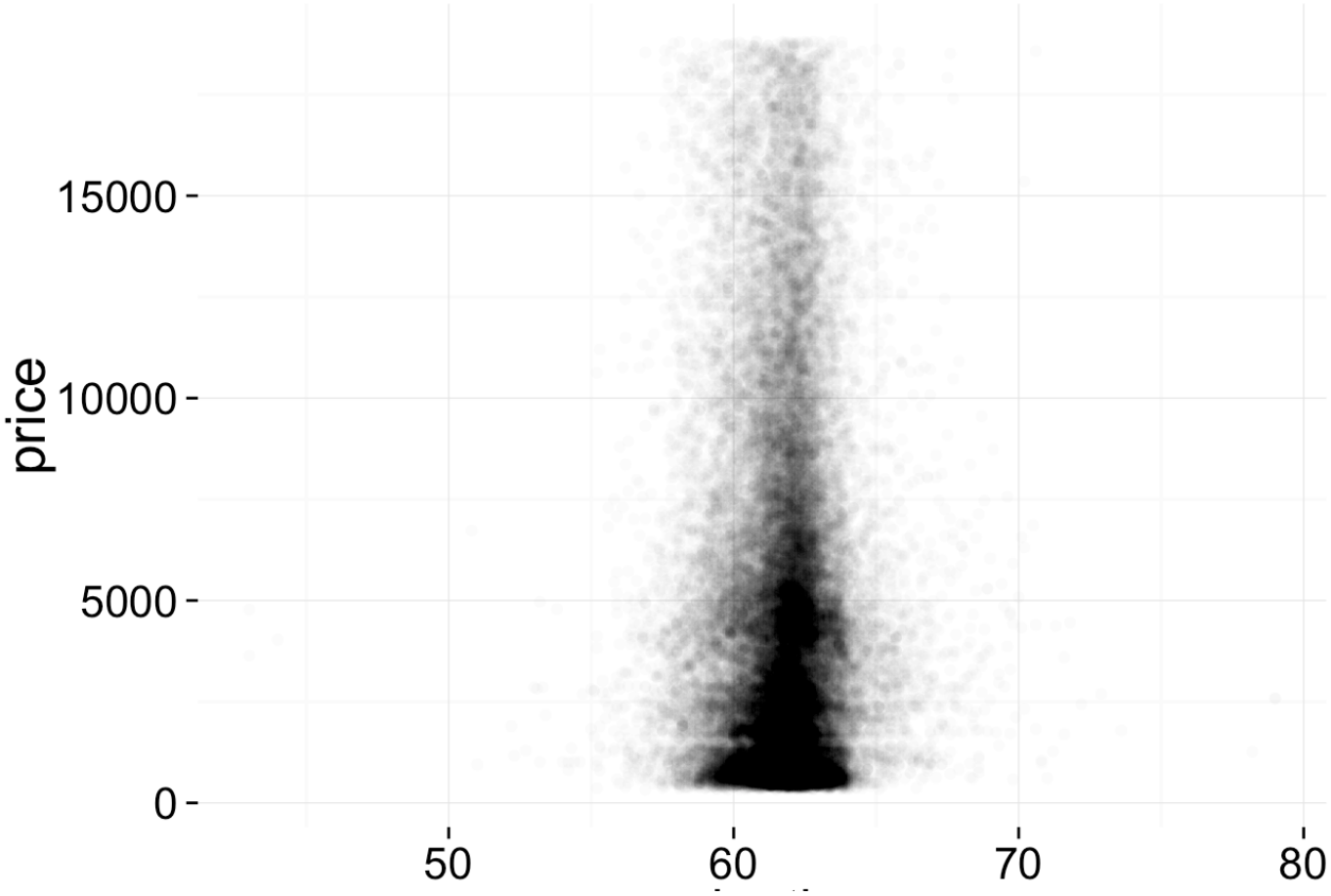

price
15000
10000
5000
0
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
clarity

```
## 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 price so low. Another trend to note here is that price variance increases then decreases significantly as the clarity improves.

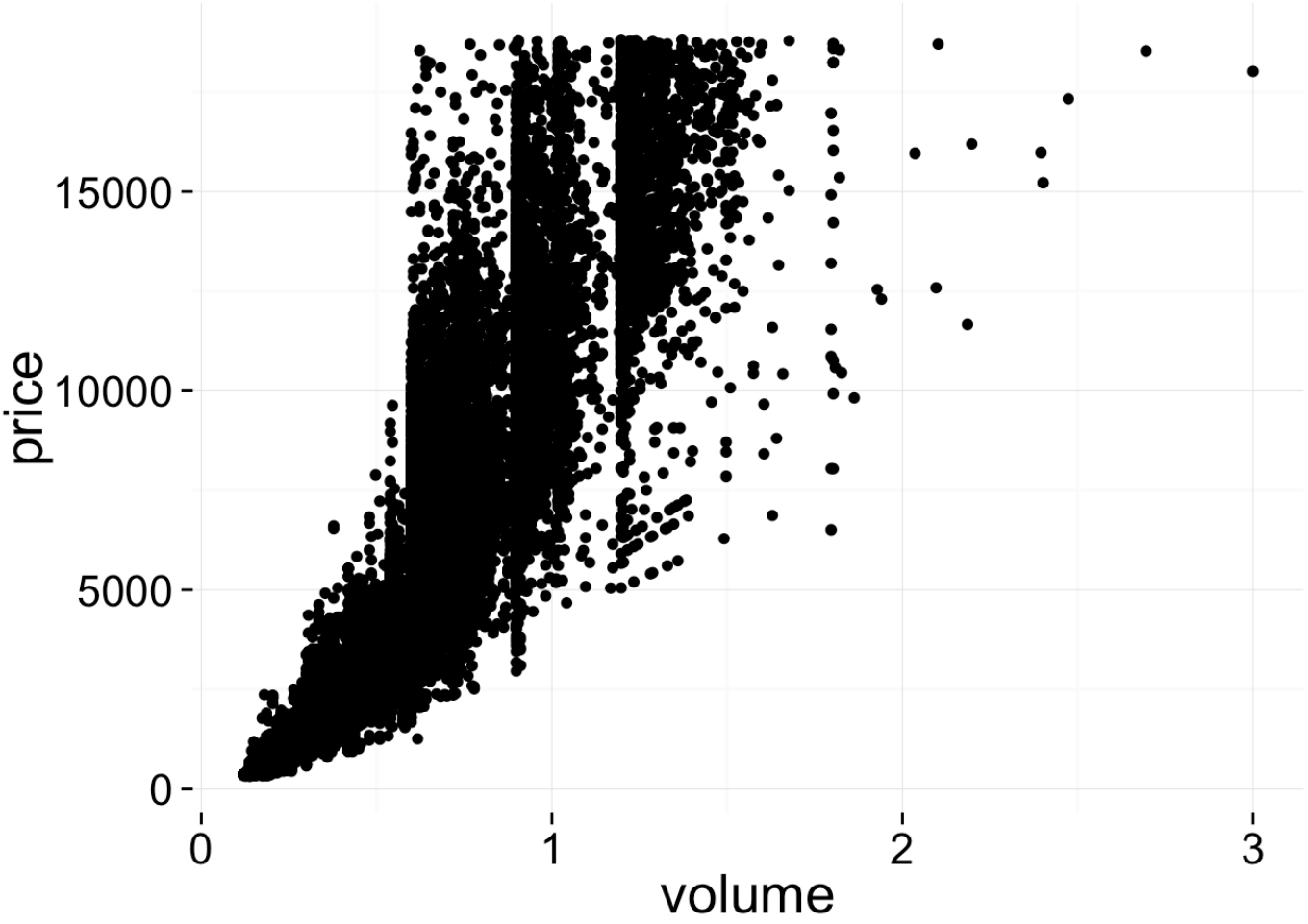
I want to look at two things: price per clarity, and the distribution of prices for diamonds with best levels of the categorical variables.


price
15000
10000
5000
0
50
60
70
80
depth


price
15000
10000
5000
0
50
60
70
80

depth

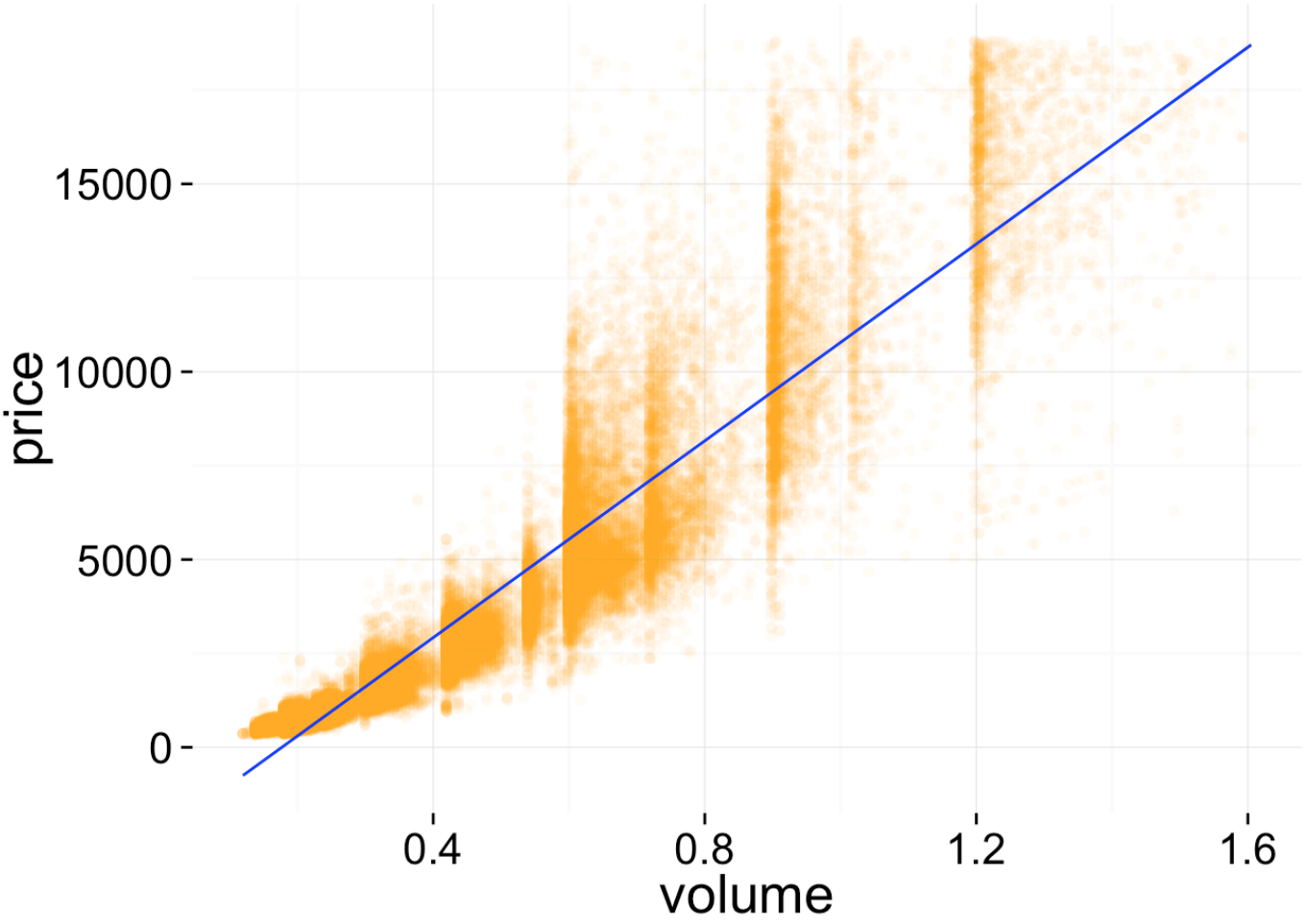
First plot suffers from overplotting. Most diamonds have a depth between 60 and 65 (no units).



No volumes that are 0. Still have some outliers, but they are less extreme.

```
##        0%        1%        2%        3%        4%        5%        6%
## 0.1197605 0.1437126 0.1616766 0.1796407 0.1796407 0.1796407 0.1796407
##        7%        8%        9%       10%       11%       12%       13%
## 0.1796407 0.1856287 0.1856287 0.1856287 0.1856287 0.1916168 0.1916168
##       14%       15%       16%       17%       18%       19%       20%
## 0.1916168 0.1916168 0.1976048 0.1976048 0.2035928 0.2035928 0.2095808
##       21%       22%       23%       24%       25%       26%       27%
## 0.2155689 0.2215569 0.2275449 0.2335329 0.2395210 0.2395210 0.2455090
##       28%       29%       30%       31%       32%       33%       34%
## 0.2455090 0.2455090 0.2514970 0.2574850 0.2694611 0.2994012 0.2994012
##       35%       36%       37%       38%       39%       40%       41%
## 0.2994012 0.3053892 0.3053892 0.3113772 0.3173653 0.3173653 0.3233533
##       42%       43%       44%       45%       46%       47%       48%
## 0.3293413 0.3353293 0.3473054 0.3592814 0.3772455 0.4191617 0.4191617
##       49%       50%       51%       52%       53%       54%       55%
## 0.4191617 0.4191617 0.4251497 0.4251497 0.4311377 0.4311377 0.4371257
##       56%       57%       58%       59%       60%       61%       62%
## 0.4491018 0.4610778 0.4790419 0.4970060 0.5389222 0.5389222 0.5389222
##       63%       64%       65%       66%       67%       68%       69%
## 0.5449102 0.5568862 0.5988024 0.5988024 0.5988024 0.6047904 0.6047904
##       70%       71%       72%       73%       74%       75%       76%
## 0.6047904 0.6047904 0.6107784 0.6107784 0.6167665 0.6227545 0.6347305
##       77%       78%       79%       80%       81%       82%       83%
## 0.6407186 0.6526946 0.6646707 0.6766467 0.6946108 0.7185629 0.7185629
##       84%       85%       86%       87%       88%       89%       90%
## 0.7305389 0.7425150 0.7544910 0.7844311 0.8323353 0.8982036 0.9041916
##       91%       92%       93%       94%       95%       96%       97%
## 0.9041916 0.9101796 0.9281437 0.9640719 1.0179641 1.1856287 1.2035928
##       98%       99%      100%
## 1.2215569 1.3053892 3.0000000
```

```
##   99.9%
## 1.60479
```



As the volume increases, the variance in price increases. That is, the data becomes more dispersed. The relationship does not look linear and appears more exponential, especially in the original plot of price vs. volume. The linear model would not be a good approximation for price since the model does not accurately predict the price at higher values diamond volumes.

```
##
## Call:
## lm(formula = price ~ volume, data = subset(diamonds, volume >
##     0 & volume <= quantile(diamonds$volume, 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 ***
## volume      13098.08      23.41   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
```

Based on the R^2 value, volume explains about 85 percent of the variance in price. Next, I'll look at other variables, including the categorical ones.

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.

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.

As the volume increases, the variance in price increases. That is, the data becomes more dispersed. The relationship does not look linear and appears exponential, especially in the plot of price vs. volume.

Based on the R^2 value, volume (the product of x, y, and z) 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.

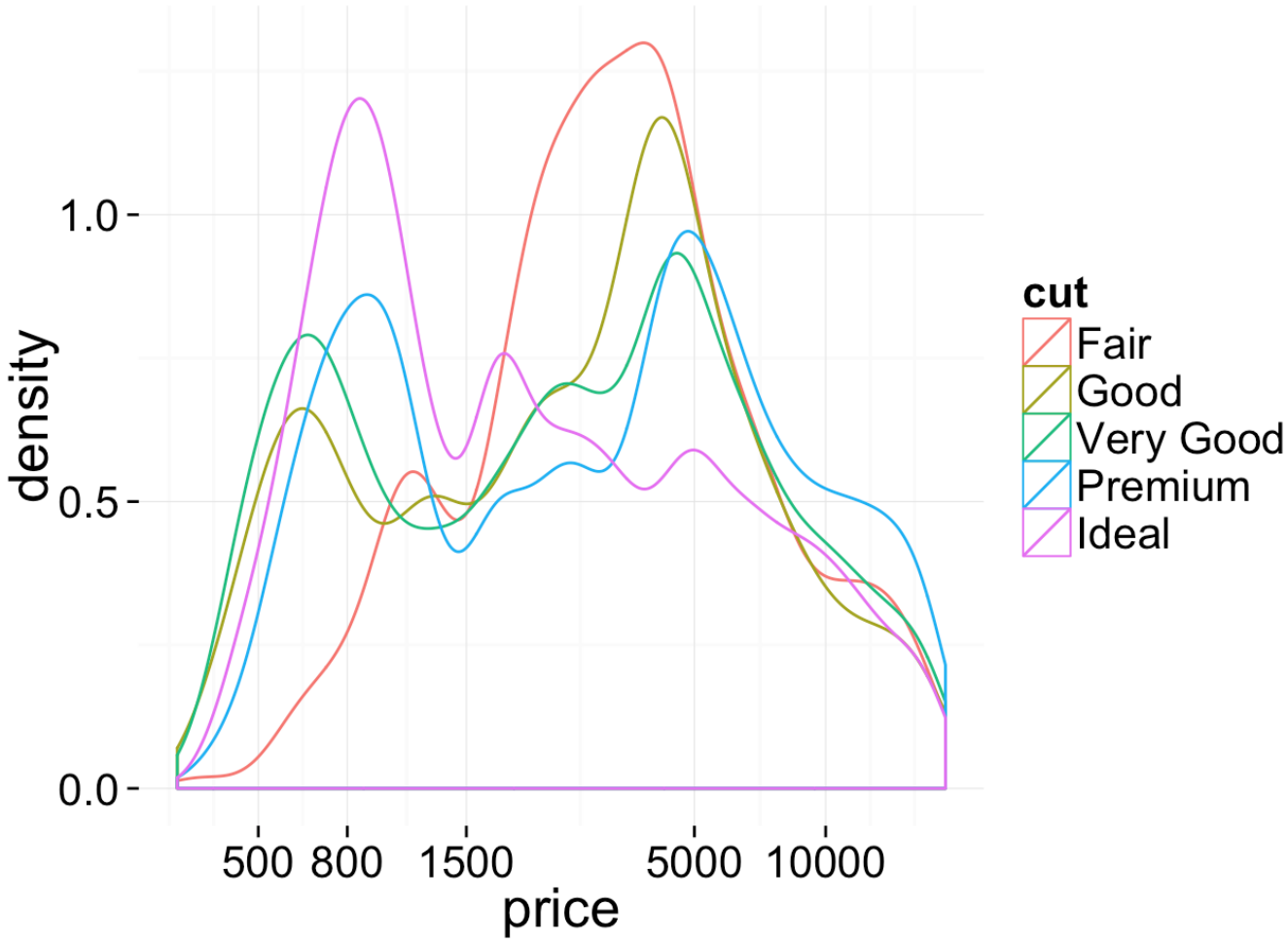
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

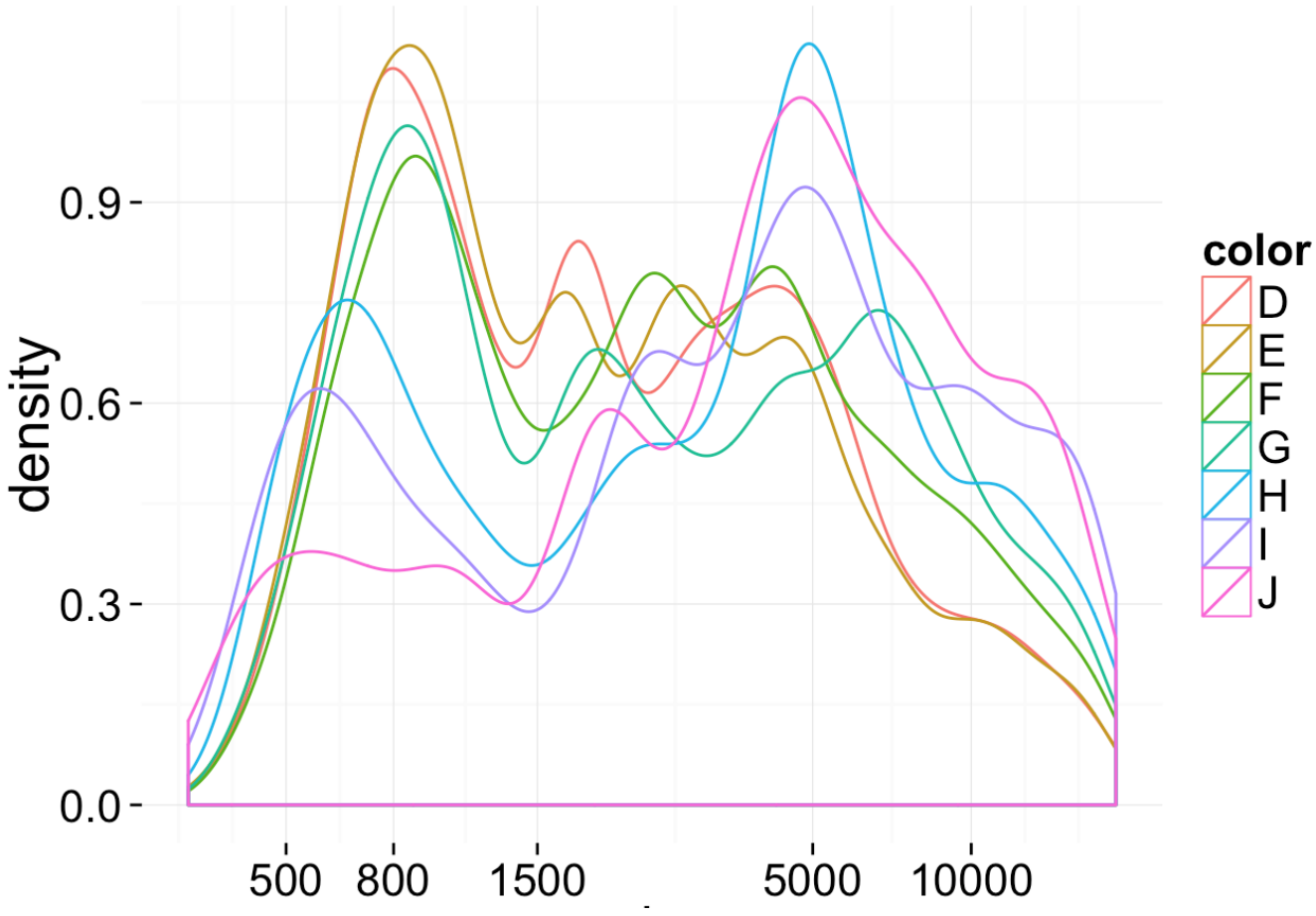
The dimensions of a diamond (x, y, and z) tend to correlate with each other. The longer one dimension, then the larger the diamond. The dimensions also correlate with carat weight which makes sense.

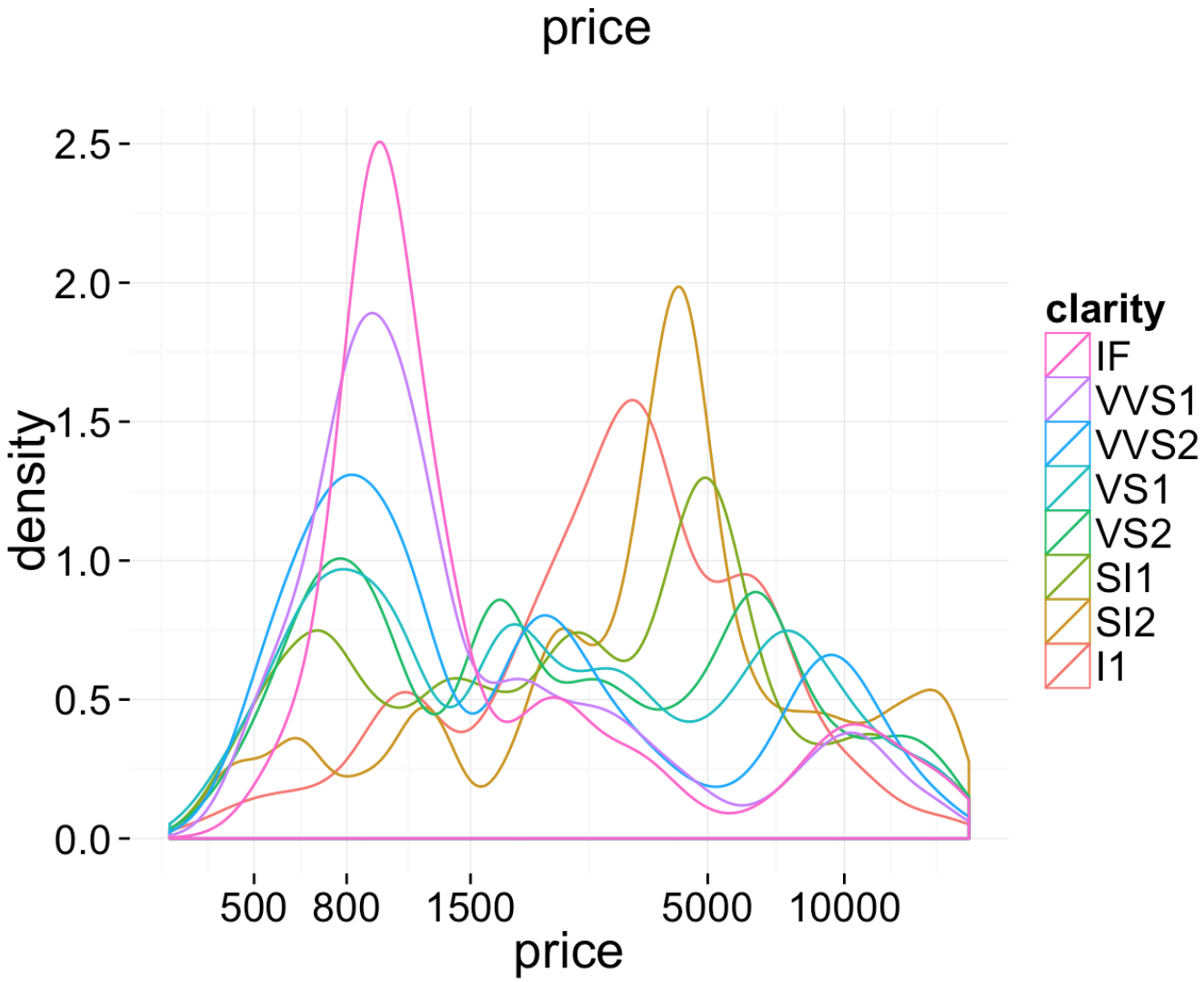
What was the strongest relationship you found?

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 essentially measuring the same quality and show perfect correlation.

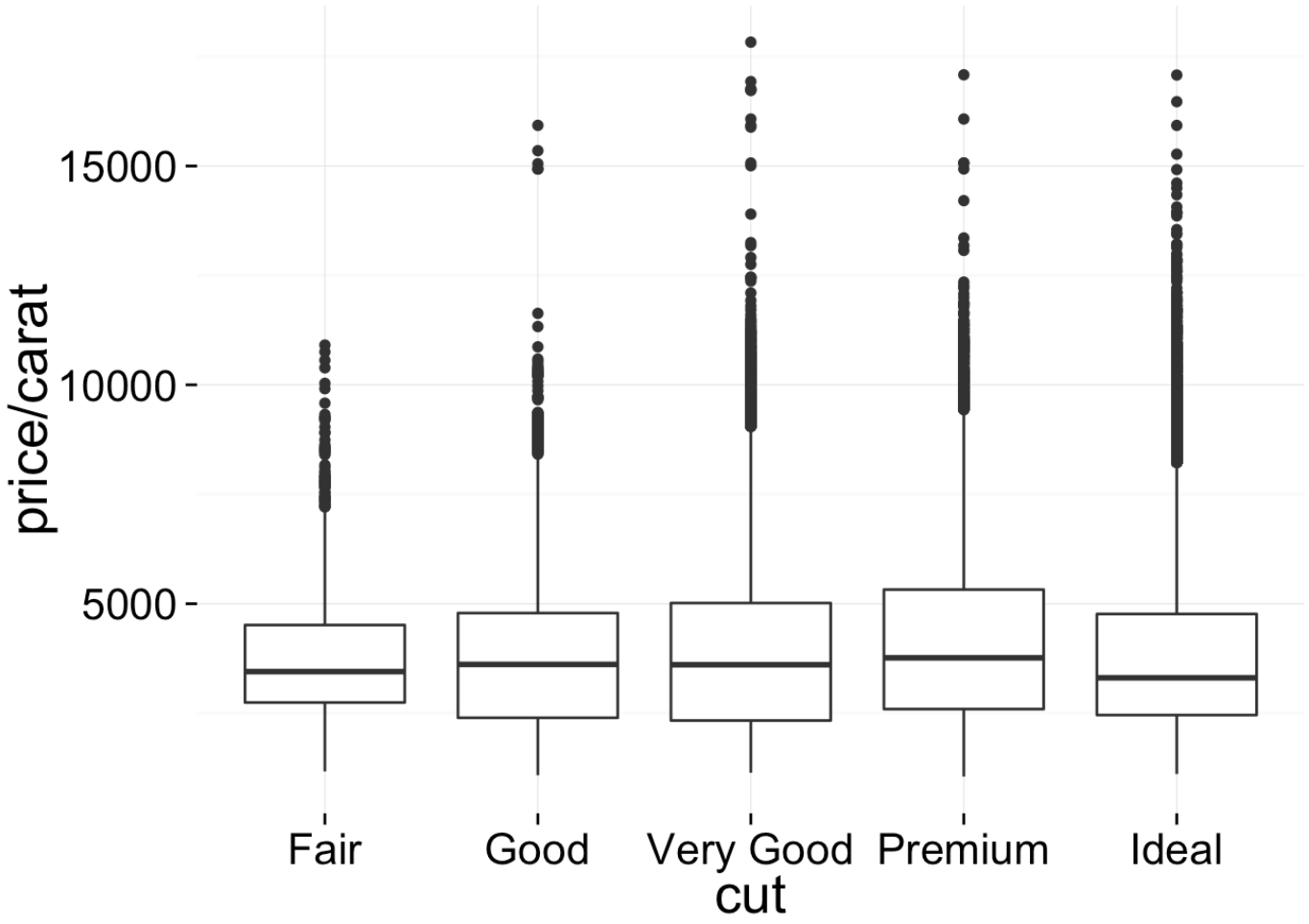
Multivariate Plots Section

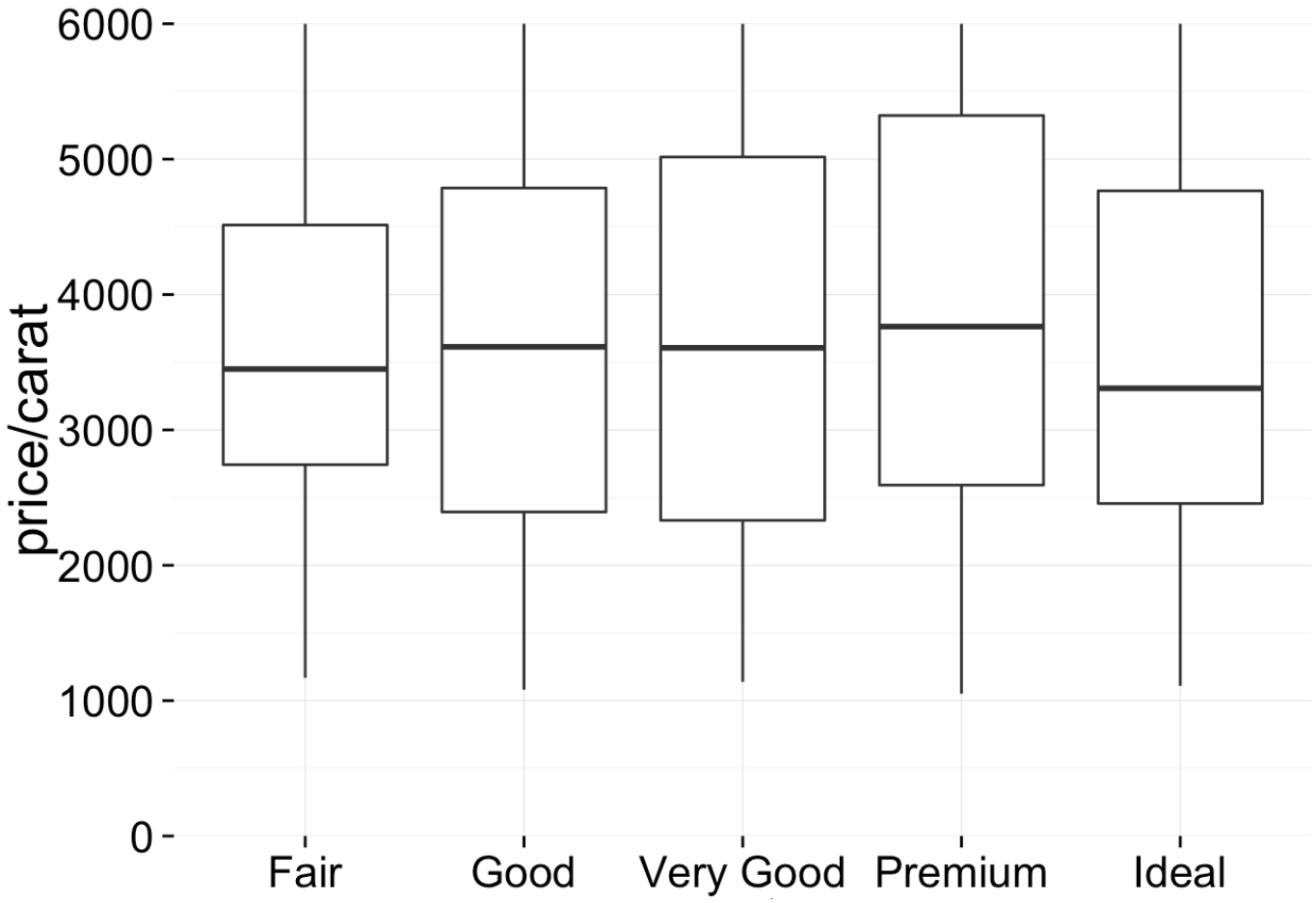

density
1.0
0.5
0.0
500
800
1500
5000
10000
price
cut
Fair
Good
Very Good
Premium
Ideal


density
0.9
0.6
0.3
0.0
500
800
1500
5000
10000
color
D
E
F
G
H
I
J



These density plots explain 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. I am wondering about price / carat too.

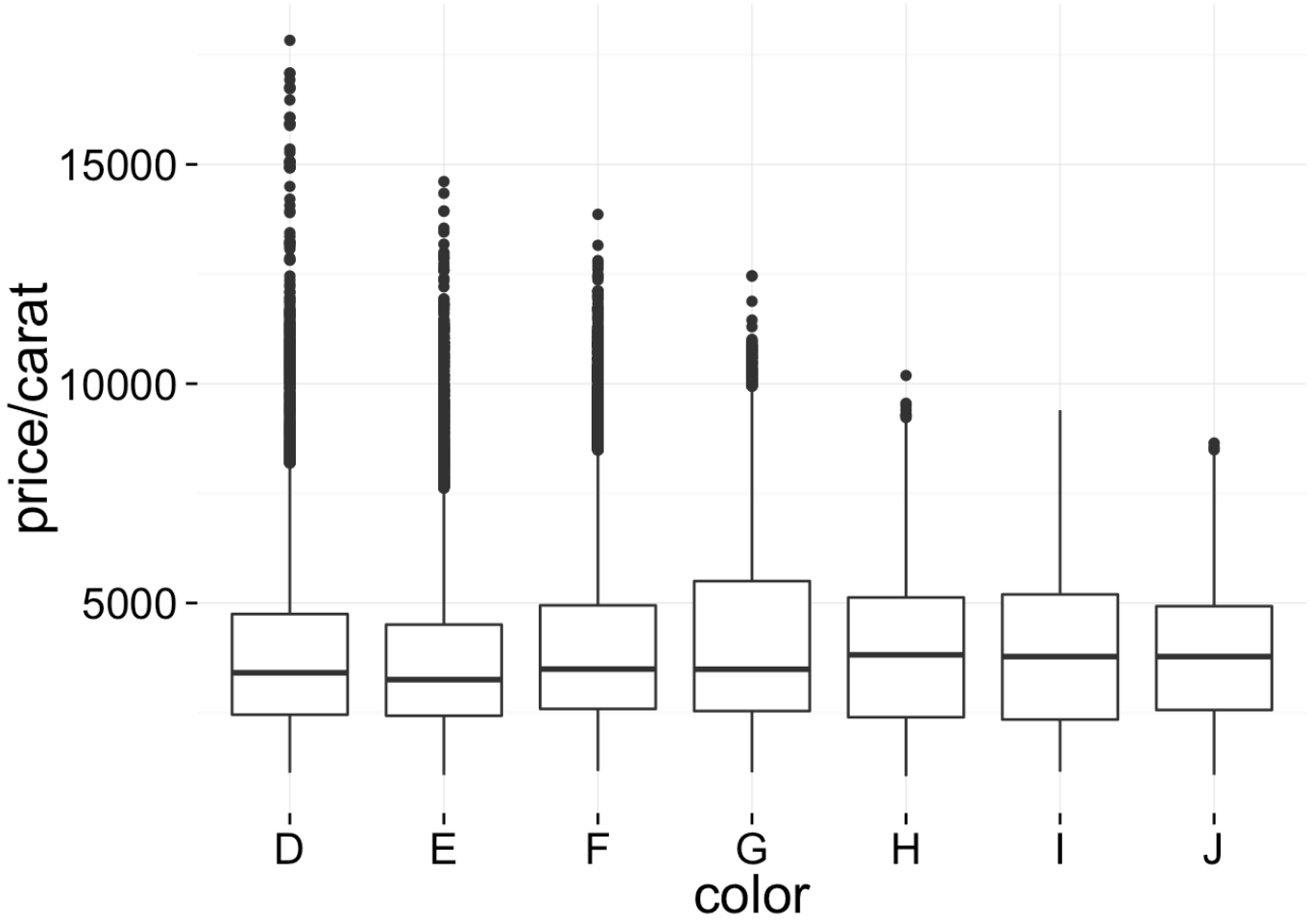

15000
10000
5000
price/carat
Fair
Good
Very Good
Premium
Ideal
cut


6000
5000
4000
3000
2000
1000
0
price/carat
Fair
Good
Very Good
Premium
Ideal

cut

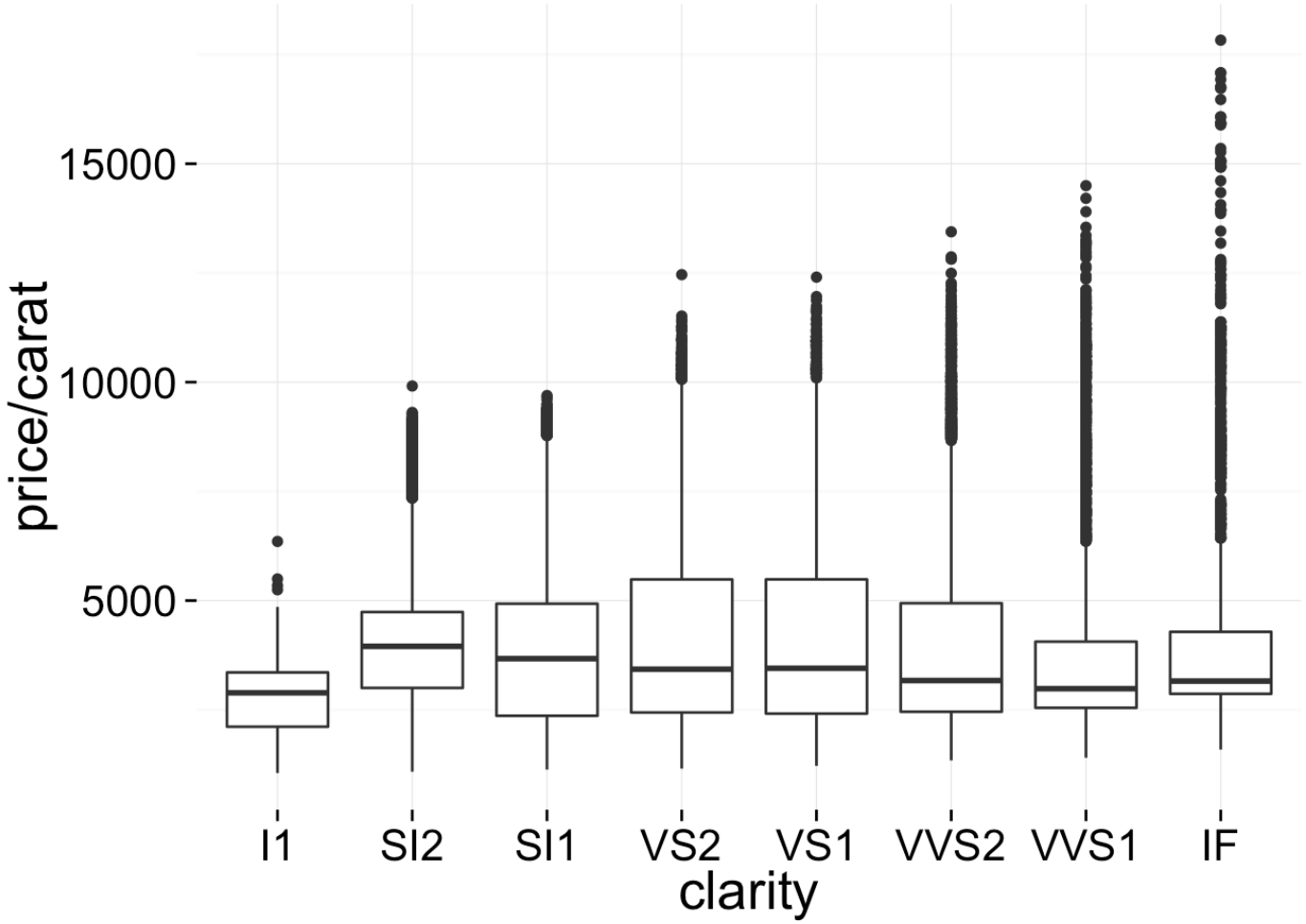
```
## cut: Fair
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1168    2743    3449    3767    4514   10910
## ------------------------------------------------------------
## cut: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1081    2394    3613    3860    4787   15930
## ------------------------------------------------------------
## 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 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.


15000
10000
5000
price/carat
D
E
F
G
H
I
J
color

```
## 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
```

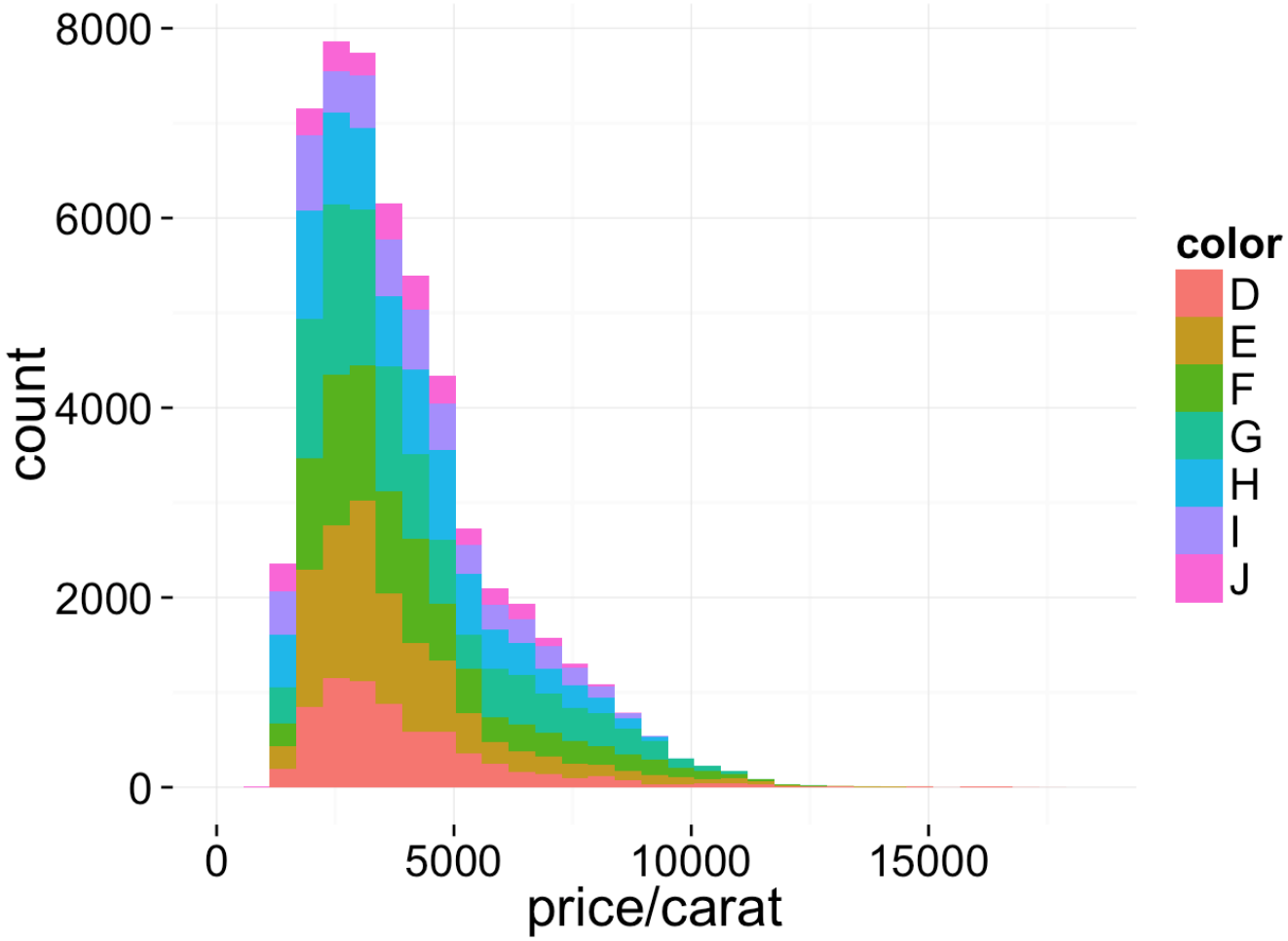
The best color diamonds (D and E) have the lowest median price. Again, this is such an unusual trend. This also seems strange since most diamonds in the data set are not of color D. I'm going to split up the price / carat distribution by color.


15000
10000
5000
price/carat
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
clarity

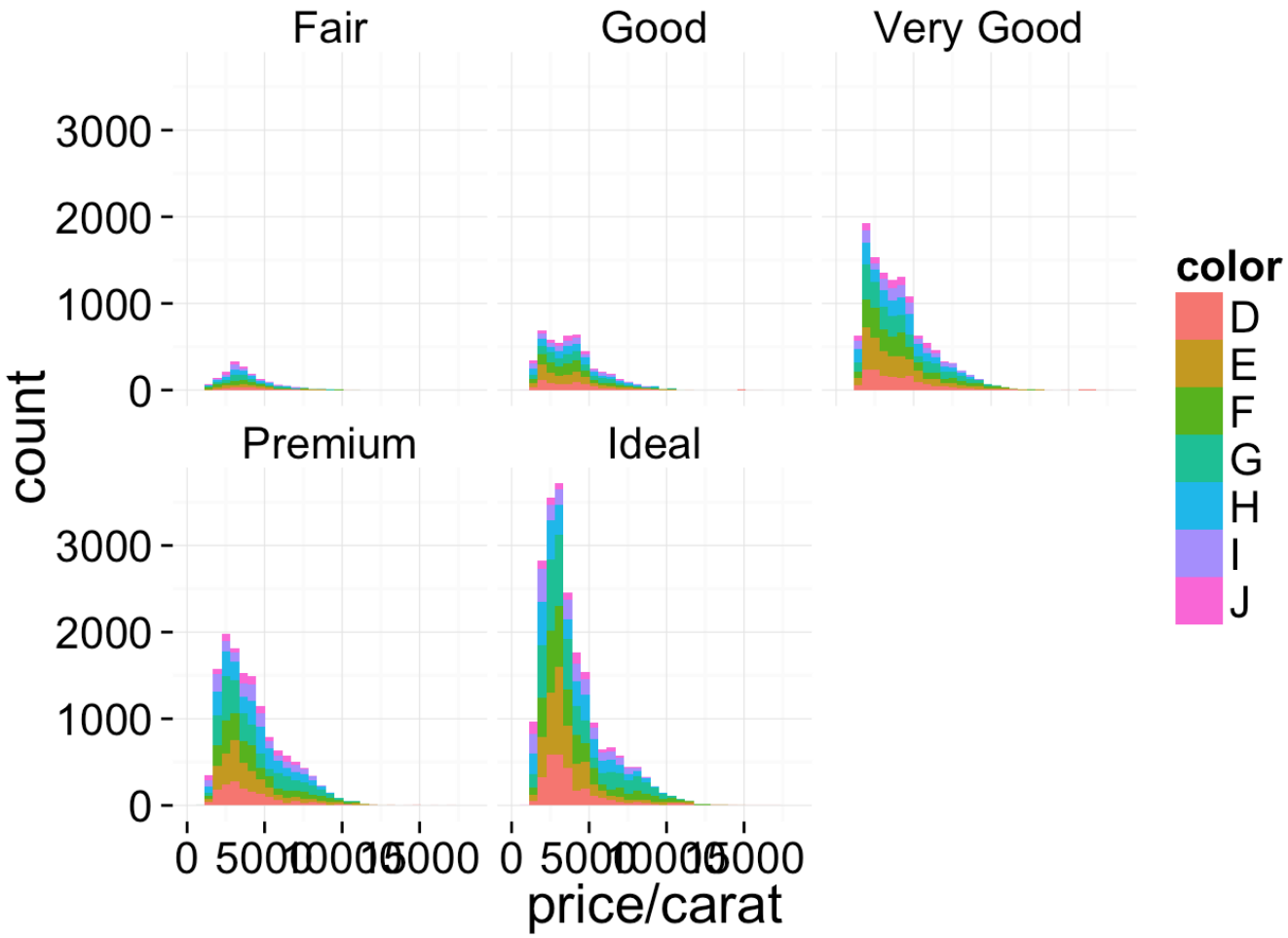
```
## clarity: I1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1051    2112    2887    2796    3354    6353
## ------------------------------------------------------------
## clarity: SI2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1081    3000    3951    4011    4738    9912
## ------------------------------------------------------------
## 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.

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

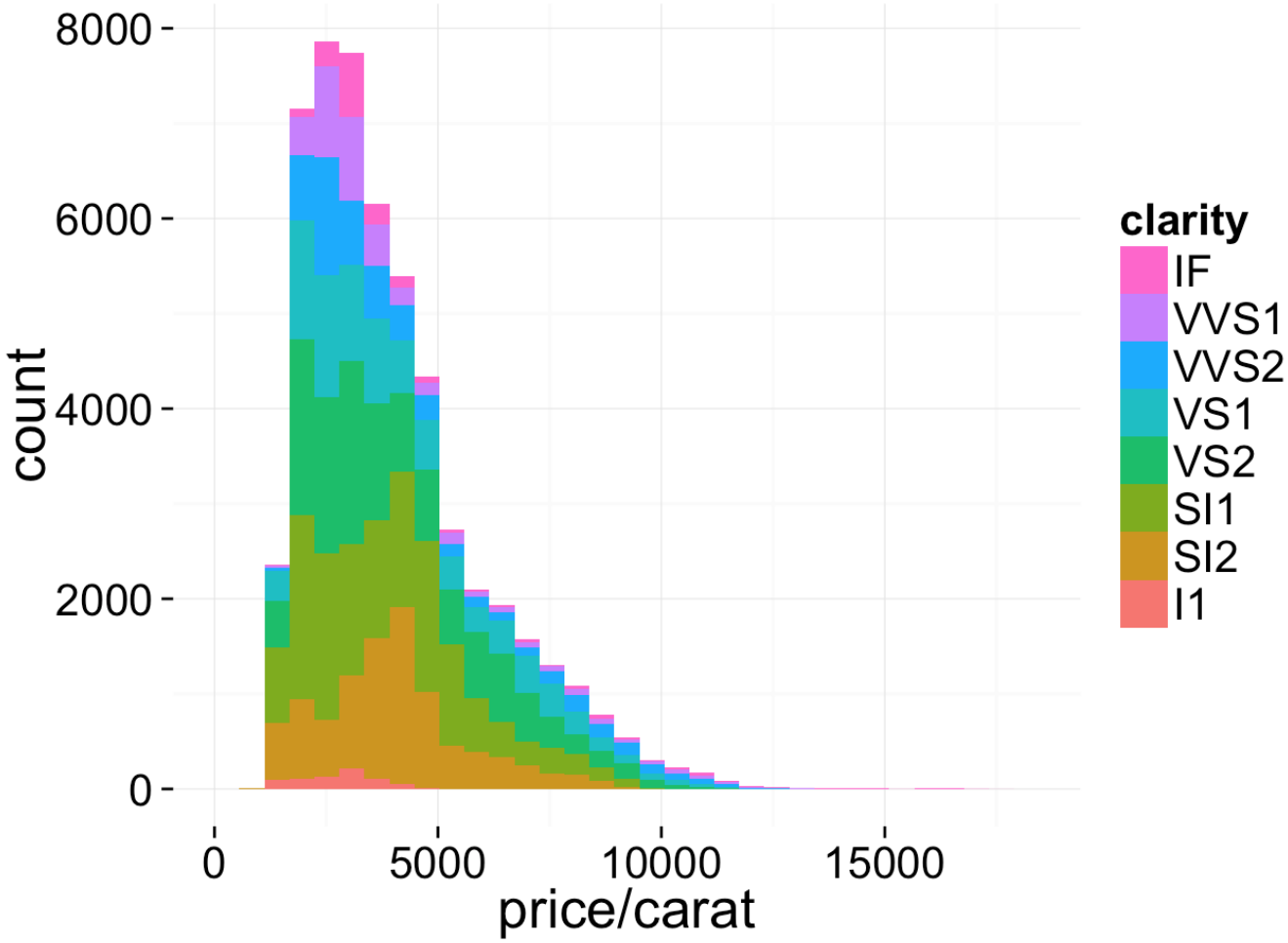


It looks like the diamonds with better cuts and color tend to have lower price / carat values. This provides some explanation for the odd low median price and price / carat for better cuts and colors, but I'm still not clear on this. I'm going to keep this in mind and try to explore the same plots for clarity.

Price per Carat Hist by Clarity

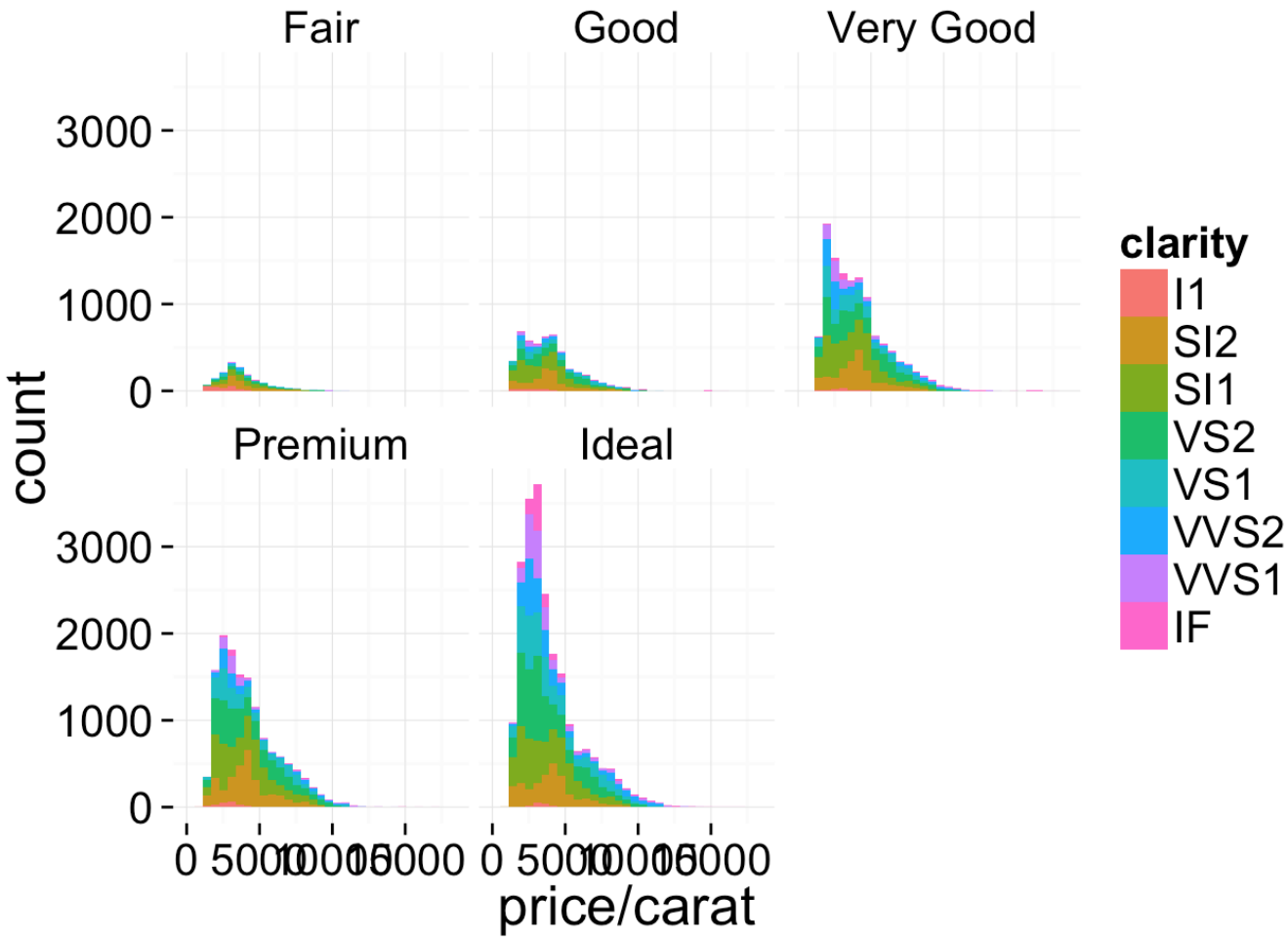
```
qplot(x = price / carat, data = diamonds, fill = clarity) +
  guides(fill = guide_legend(reverse = T))
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



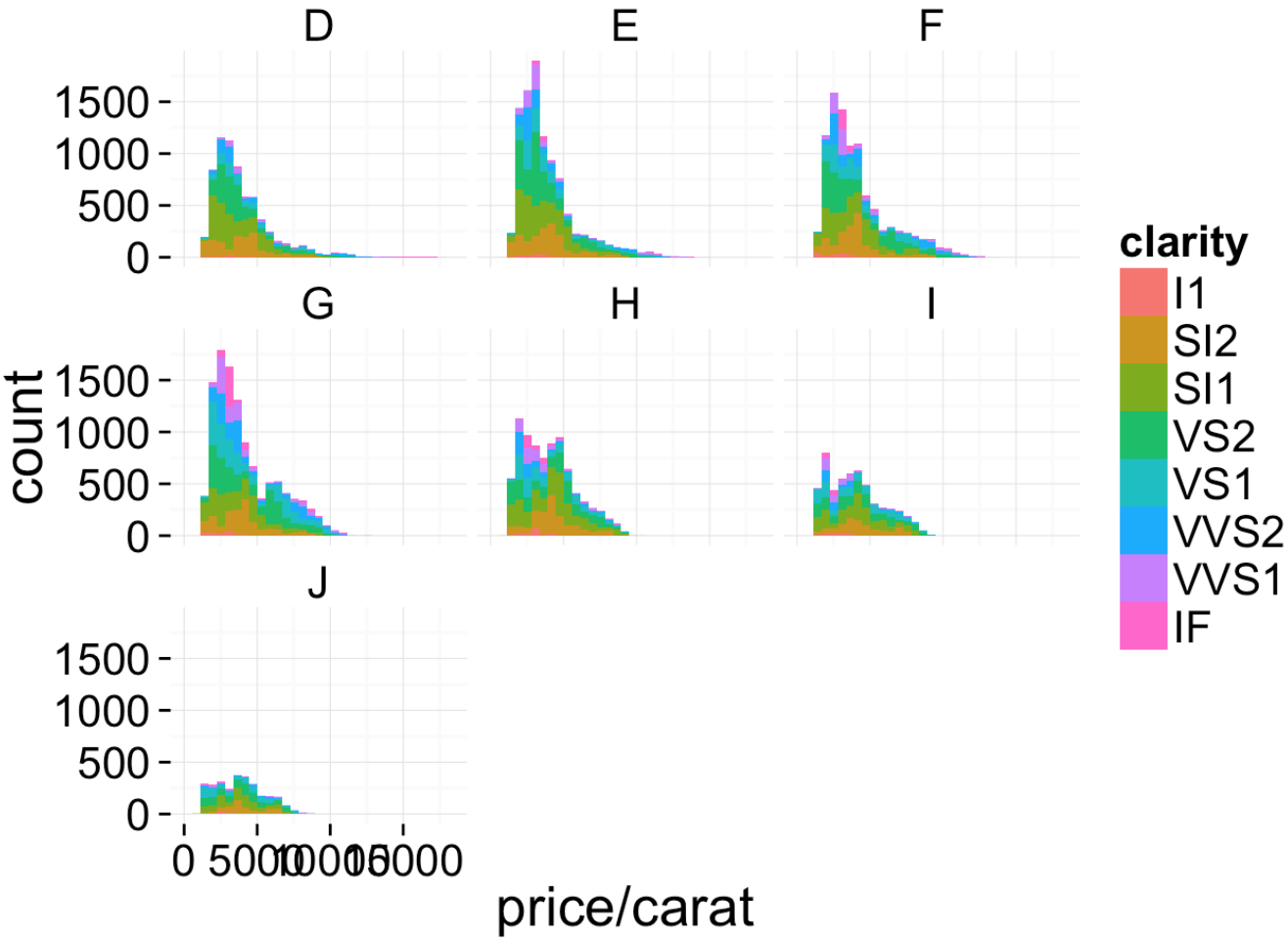

```
qplot(x = price / carat, data = diamonds, fill = clarity) +
  facet_wrap(~cut)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

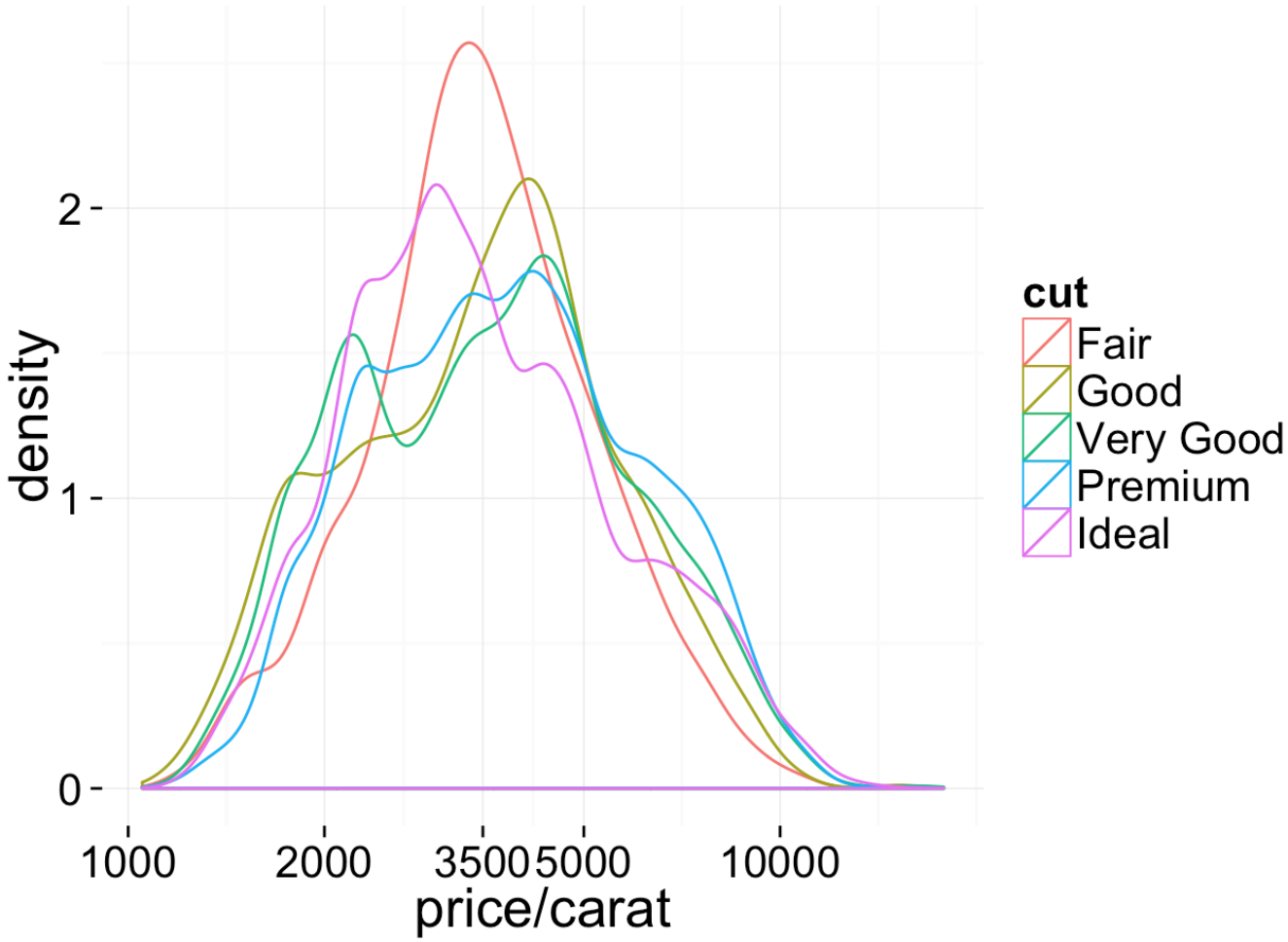


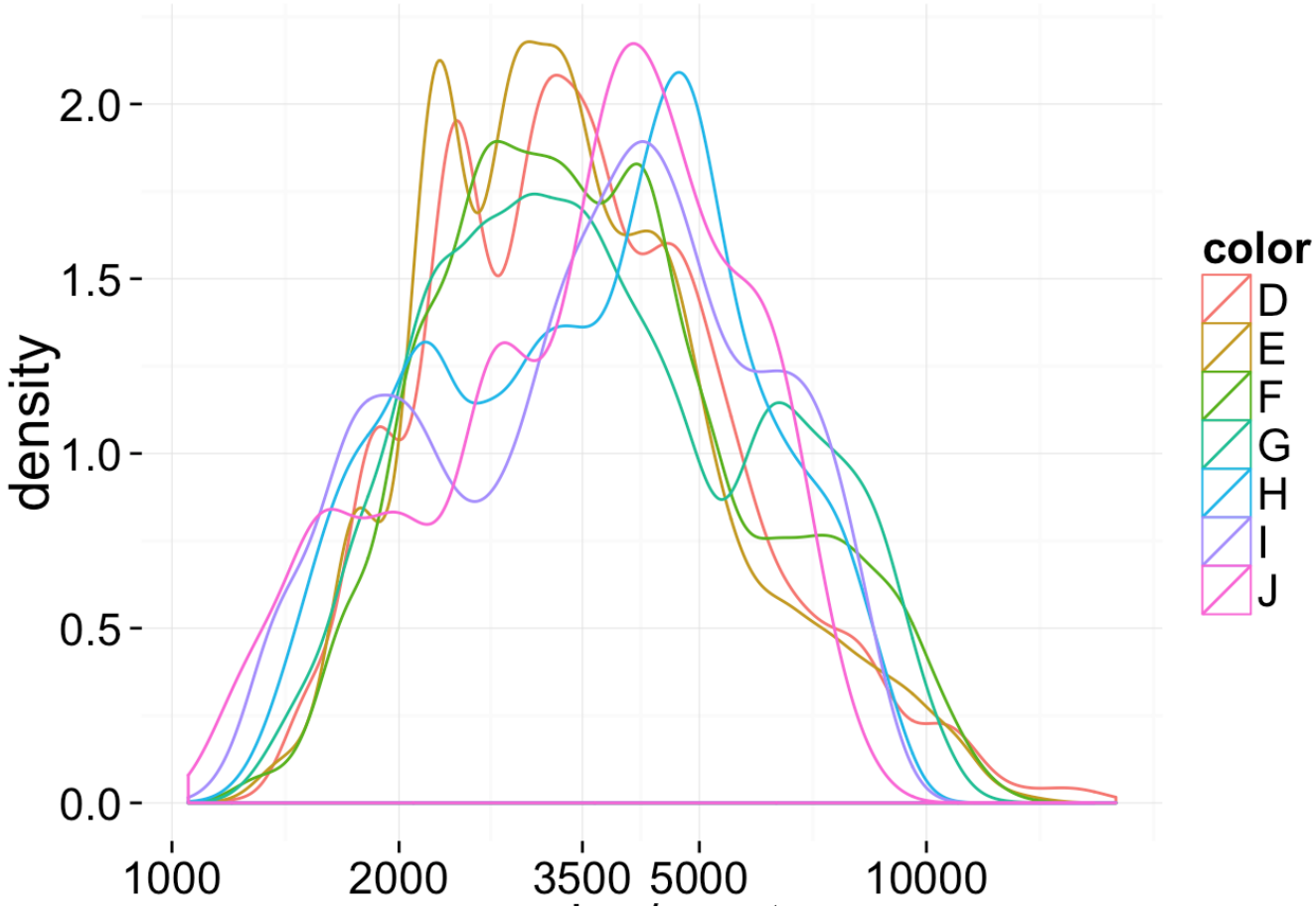

```
qplot(x = price / carat, data = diamonds, fill = clarity) +
  facet_wrap(~color)
```

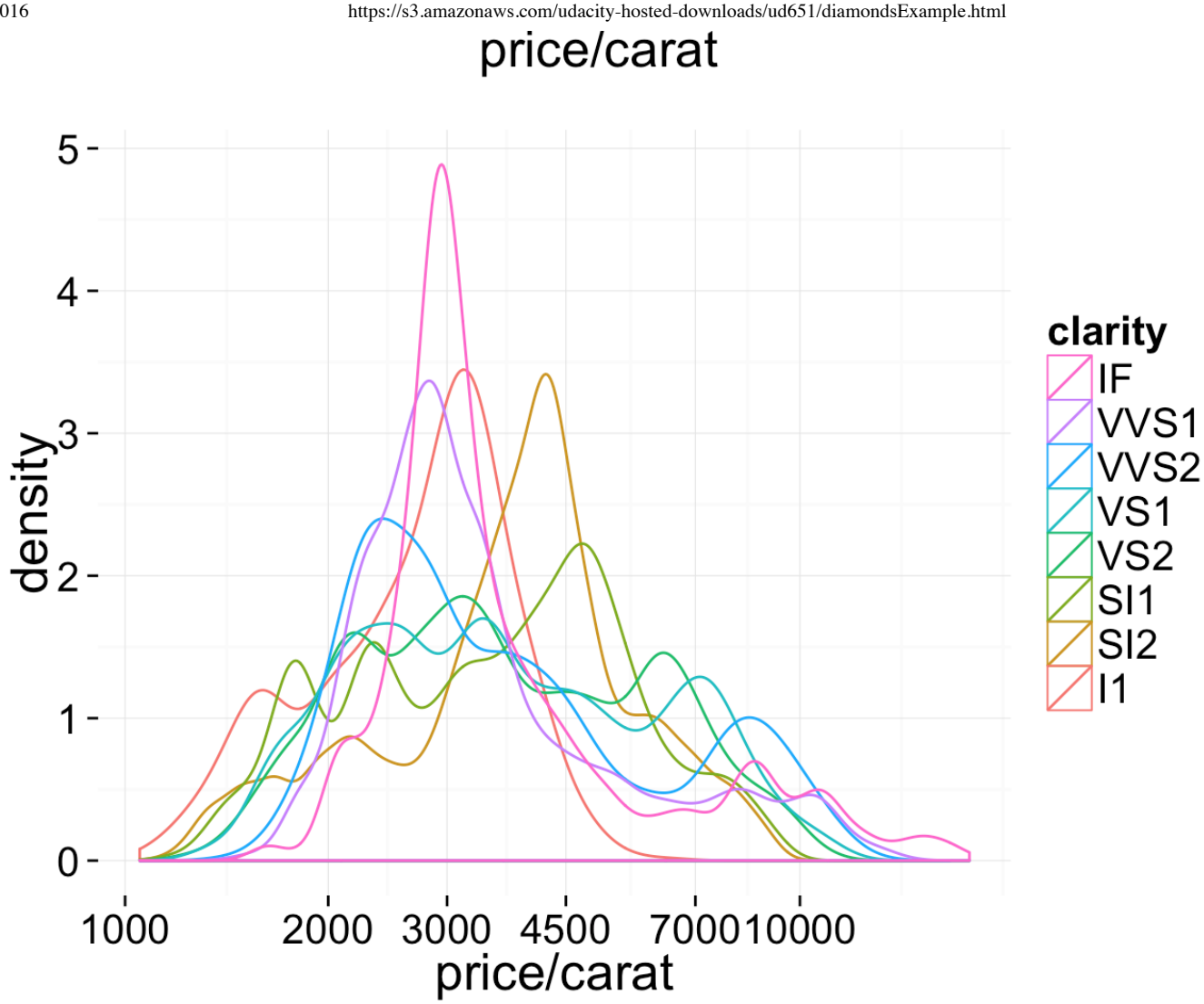
```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



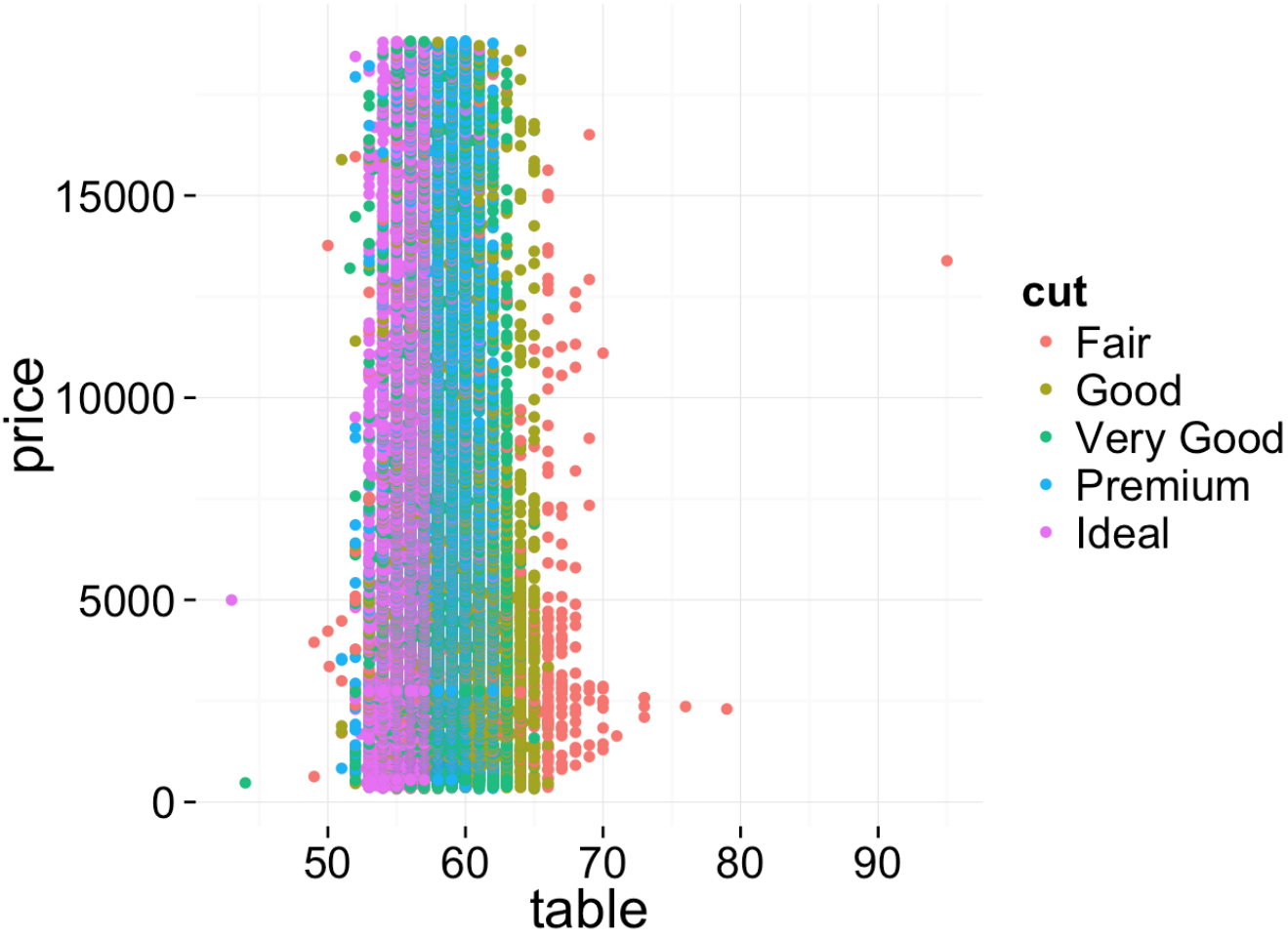
The histogram and faceted histograms somewhat explain the odd trends as again there is a greater number of ideal diamonds, color D diamonds, and clarity IF diamonds in the lower price ranges. Next, I'll look at the price distribution of the higher quality diamonds in cut, color, and clarity.

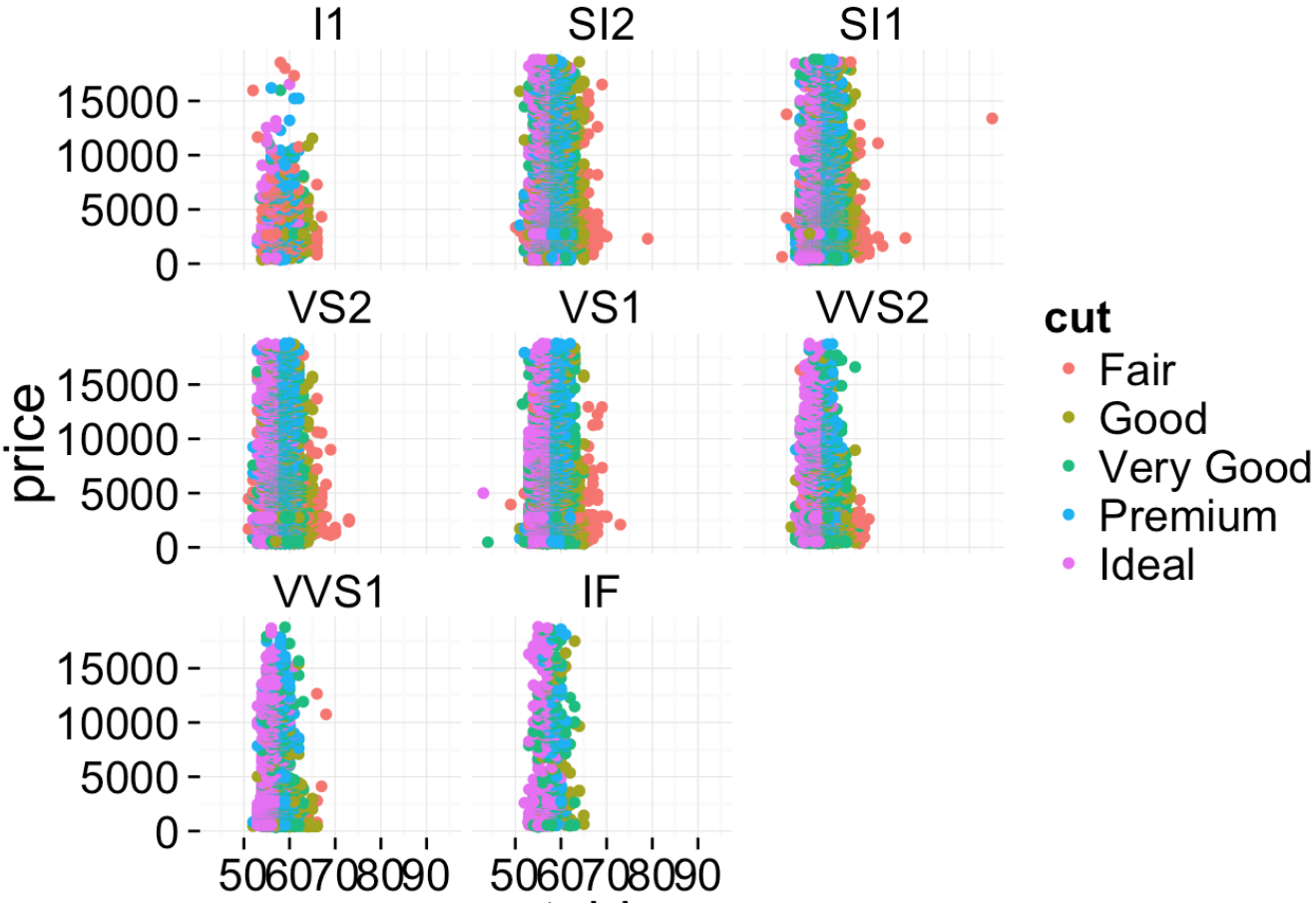

density
2
1
0
1000
2000
3500
5000
10000
price/carat
cut
Fair
Good
Very Good
Premium
Ideal

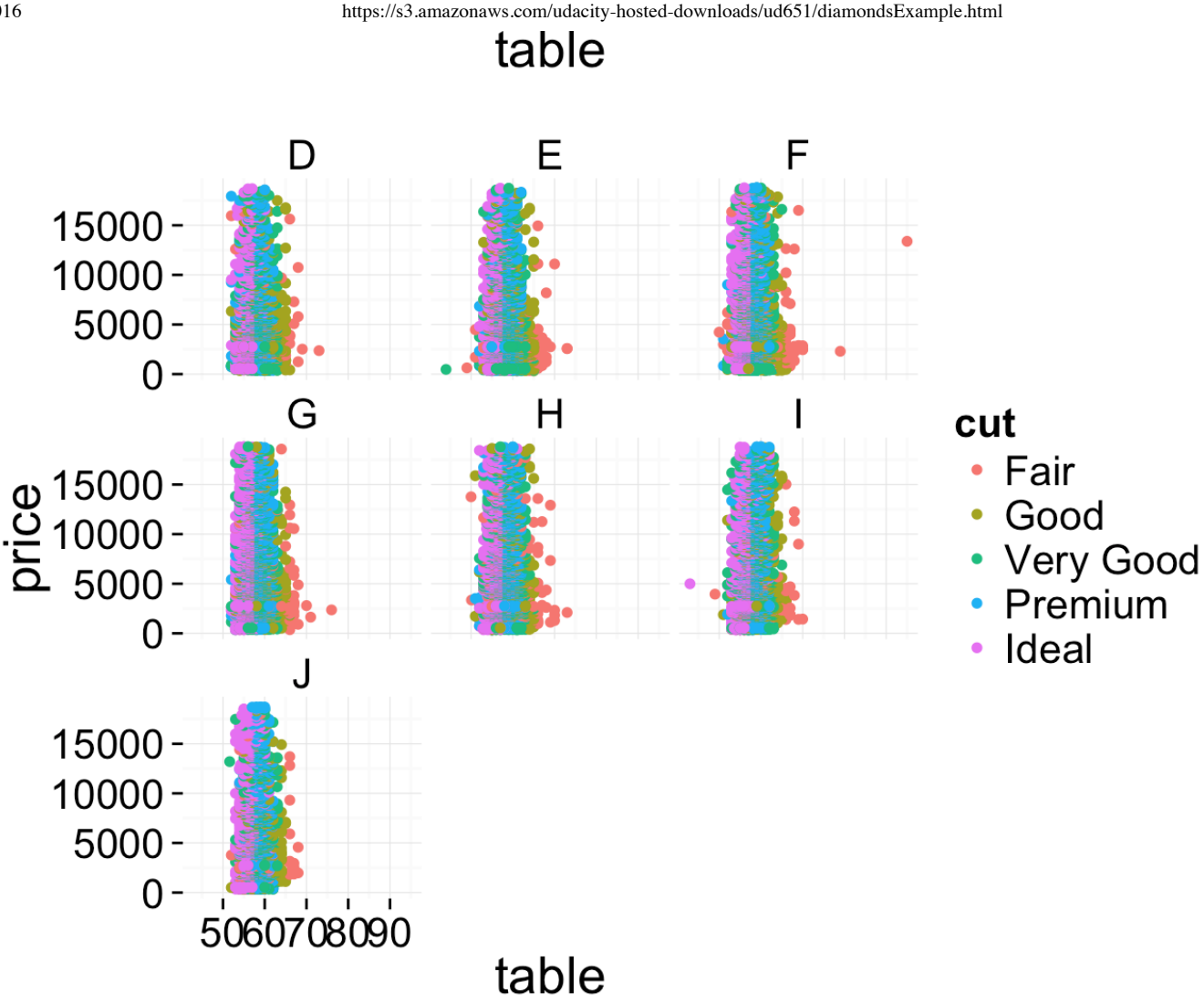

density
2.0
1.5
1.0
0.5
0.0
1000
2000
3500
5000
10000
color
D
E
F
G
H
I
J



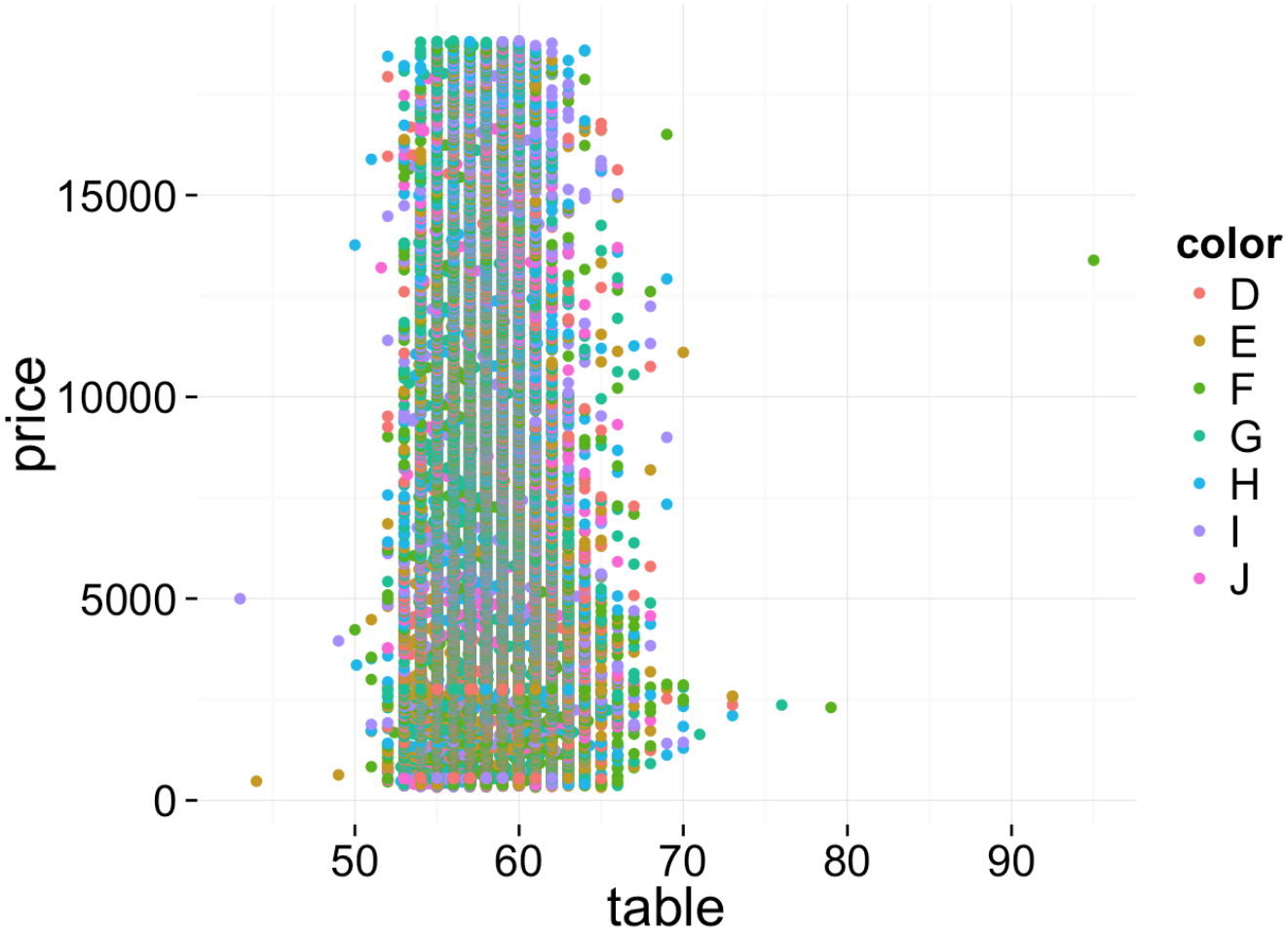
These plots support the variability and trends that the boxplots showed from before. I am going see which variables correlate with price and try to work towards building a linear model to predict price.


price
15000
10000
5000
0
50
60
70
80
90
table
cut
Fair
Good
Very Good
Premium
Ideal

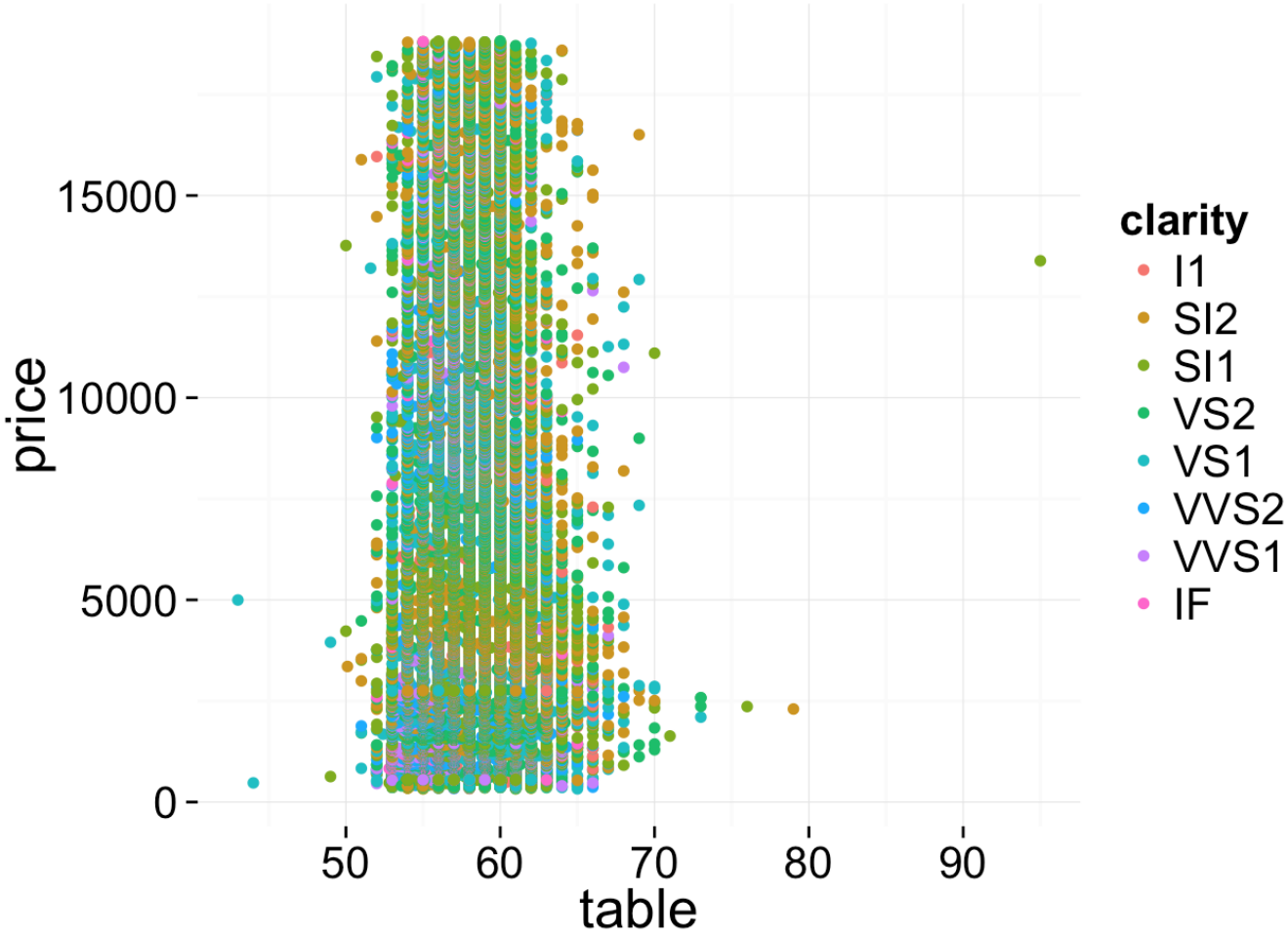

I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF
price
15000
10000
5000
0
50
60
70
80
90
cut
Fair
Good
Very Good
Premium
Ideal



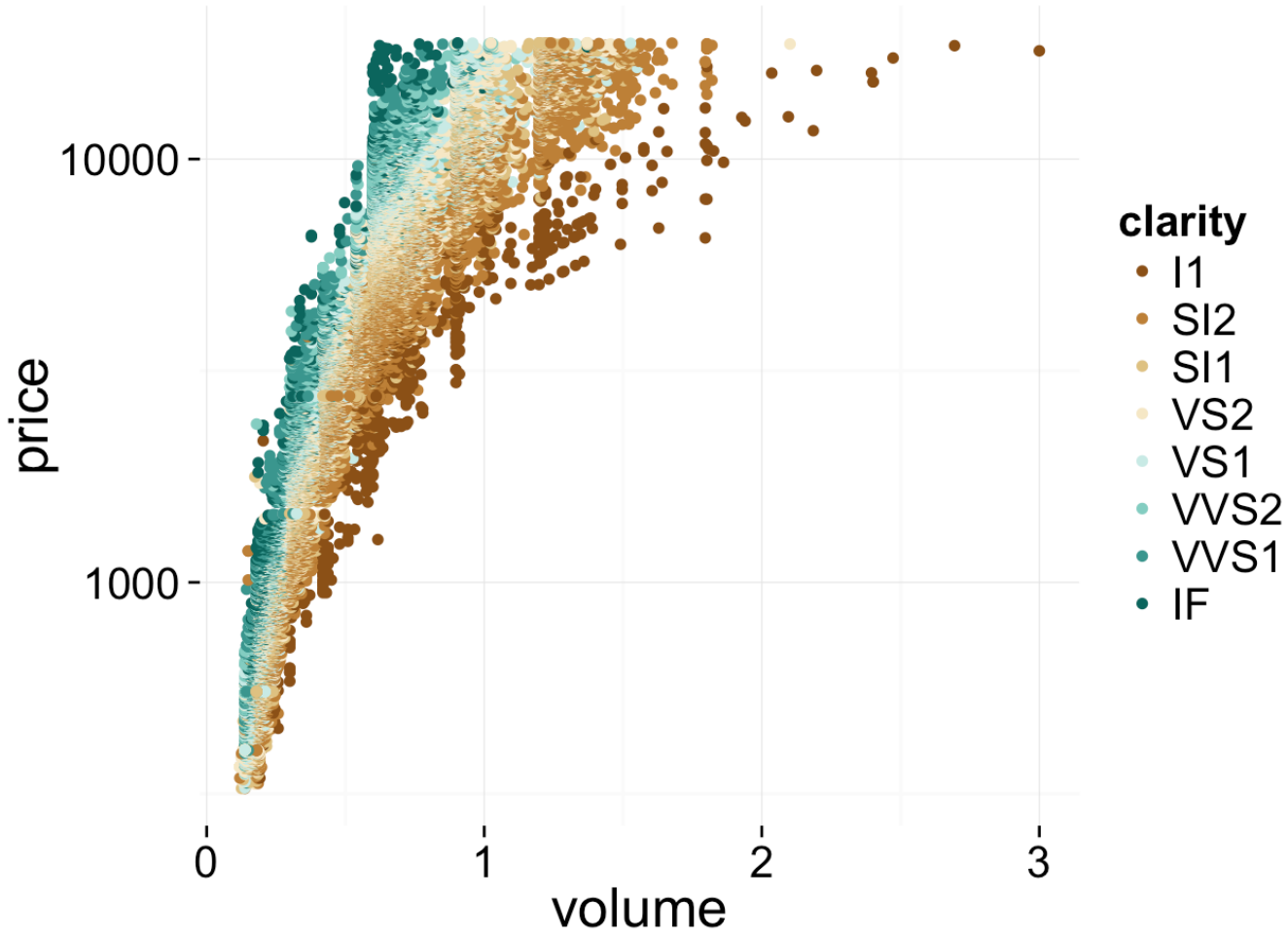
Levels of cut cluster by table value. This may make sense based on the type of cut as certain cuts produce certain dimensions. The pattern holds across each level of clarity and each level of color with the exception of the lowest clarity.

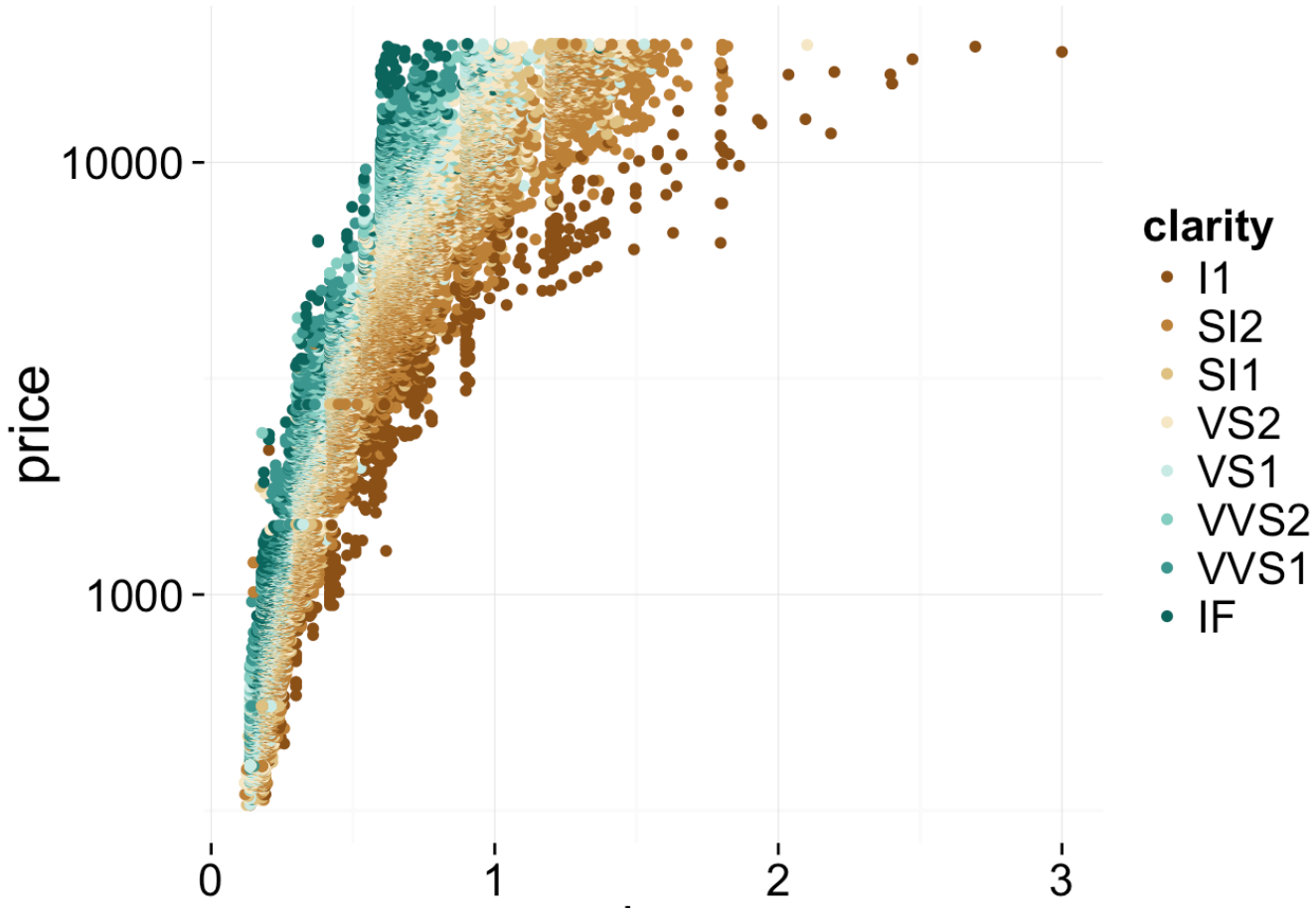


Nothing stands out in the plot above.



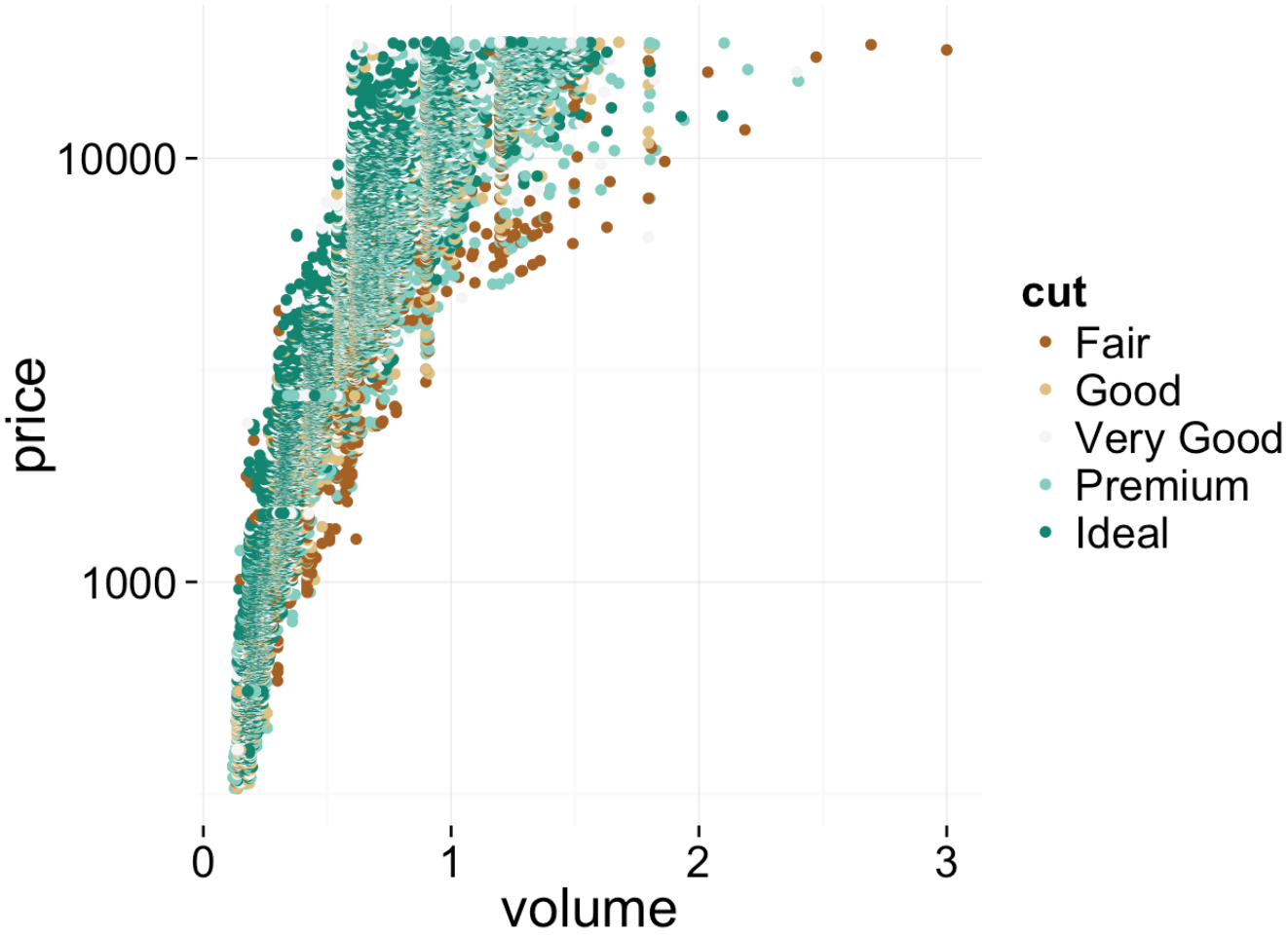
Nothing stands out in the plot above.


price
10000
1000
0
1
2
3
volume
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

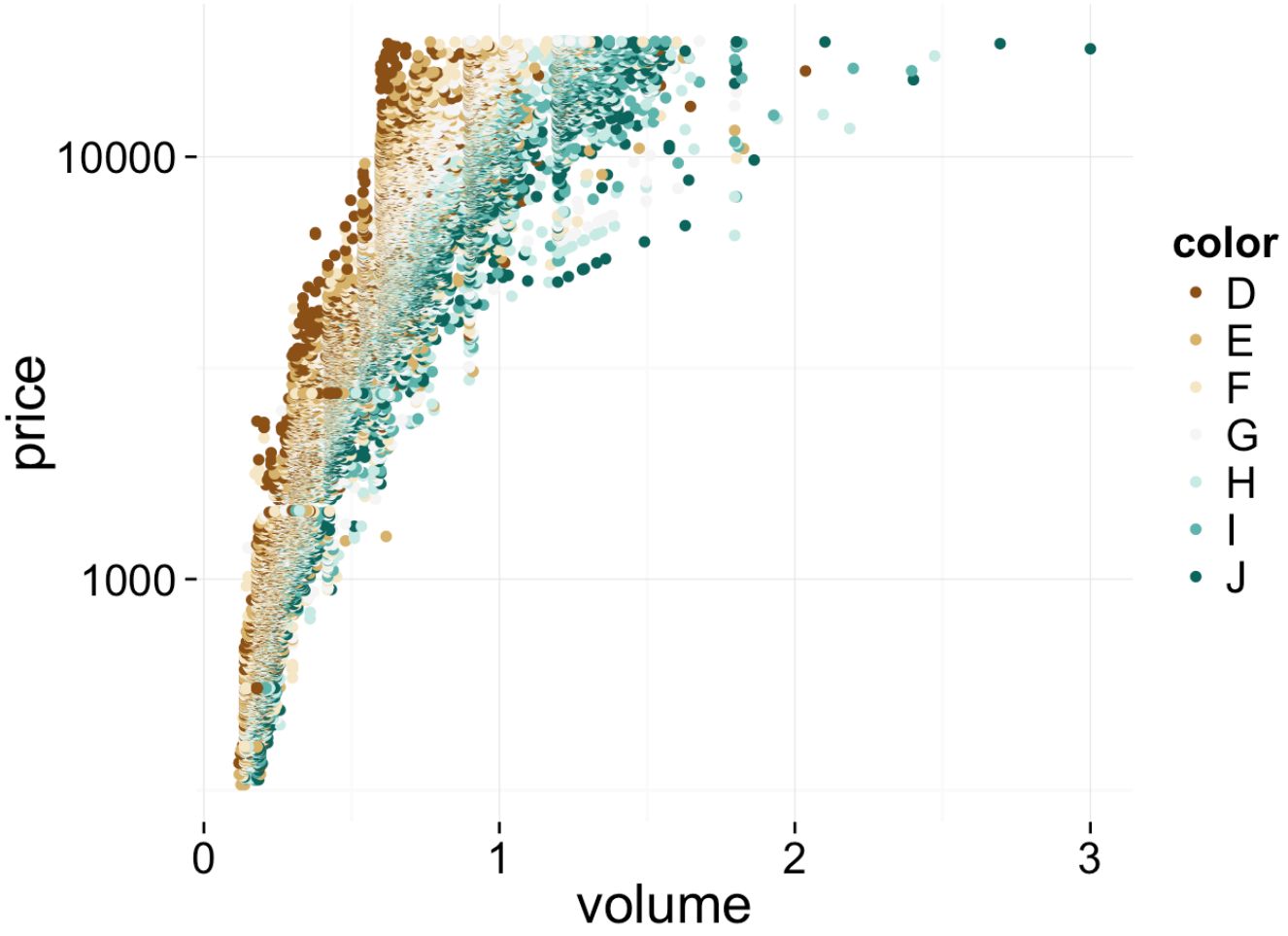

price
10000
1000
0
1
2
3
clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

volume

Diamonds are priced higher for better clarity holding volume constant.



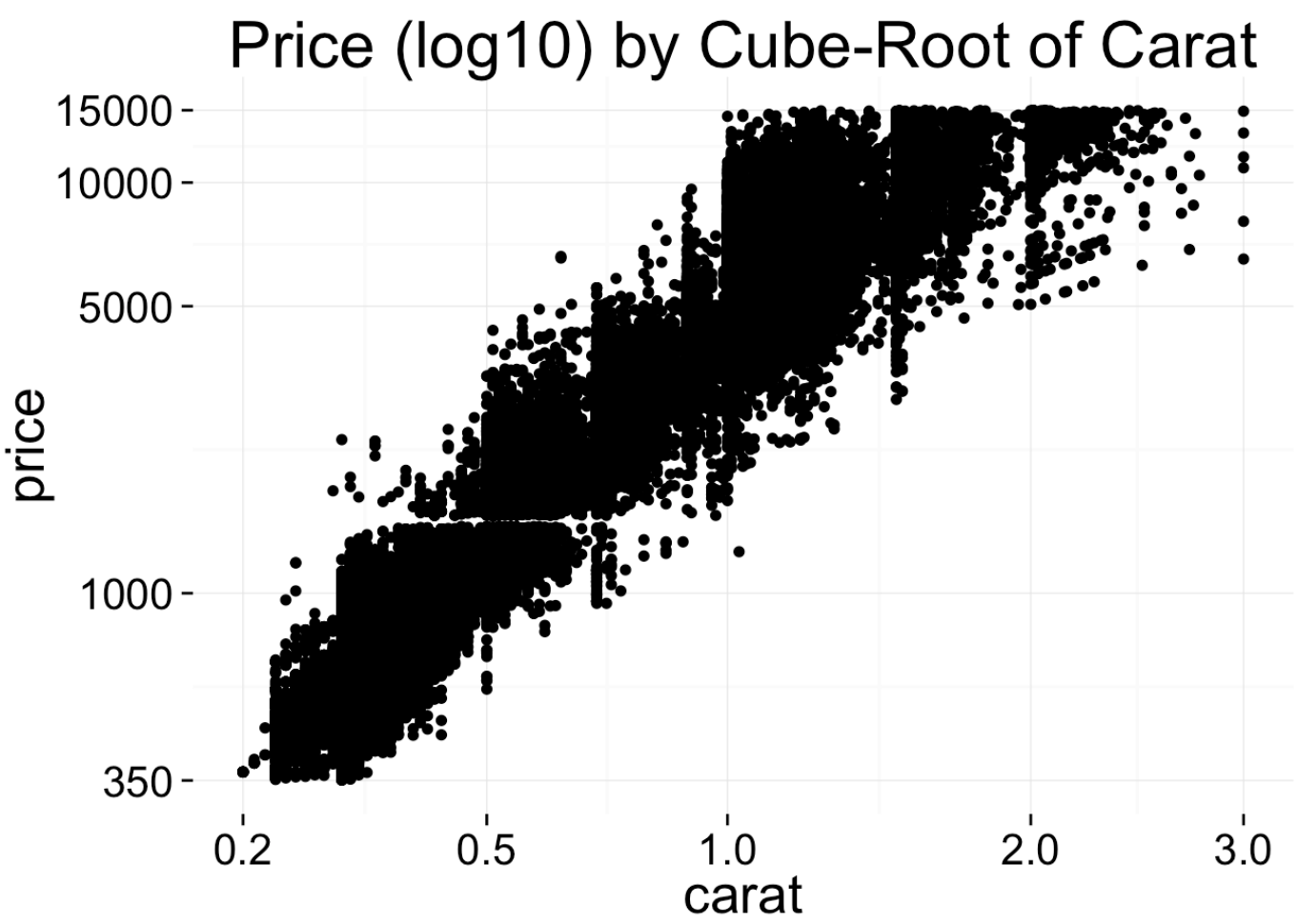
I lose the pattern when coloring by cut.



Diamonds with better color tend to be priced higher holding volume constant. This trend is not as clear when looking at prive vs volume and clarity, but the trend is still present.

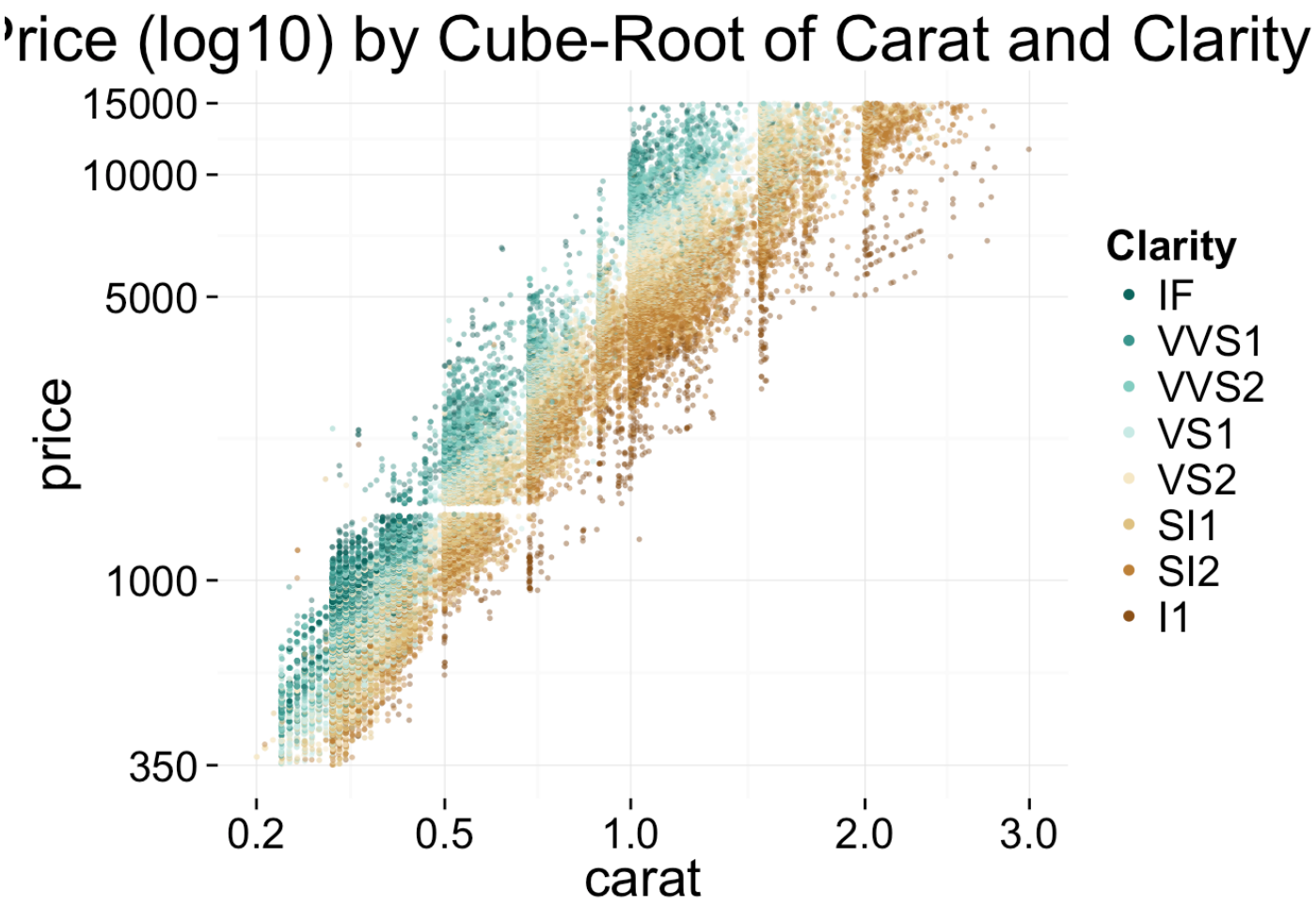
Log10 Price and Cube Root of Carat

```
## Warning: Removed 1683 rows containing missing values (geom_point).
```



Price vs Carat and Clarity

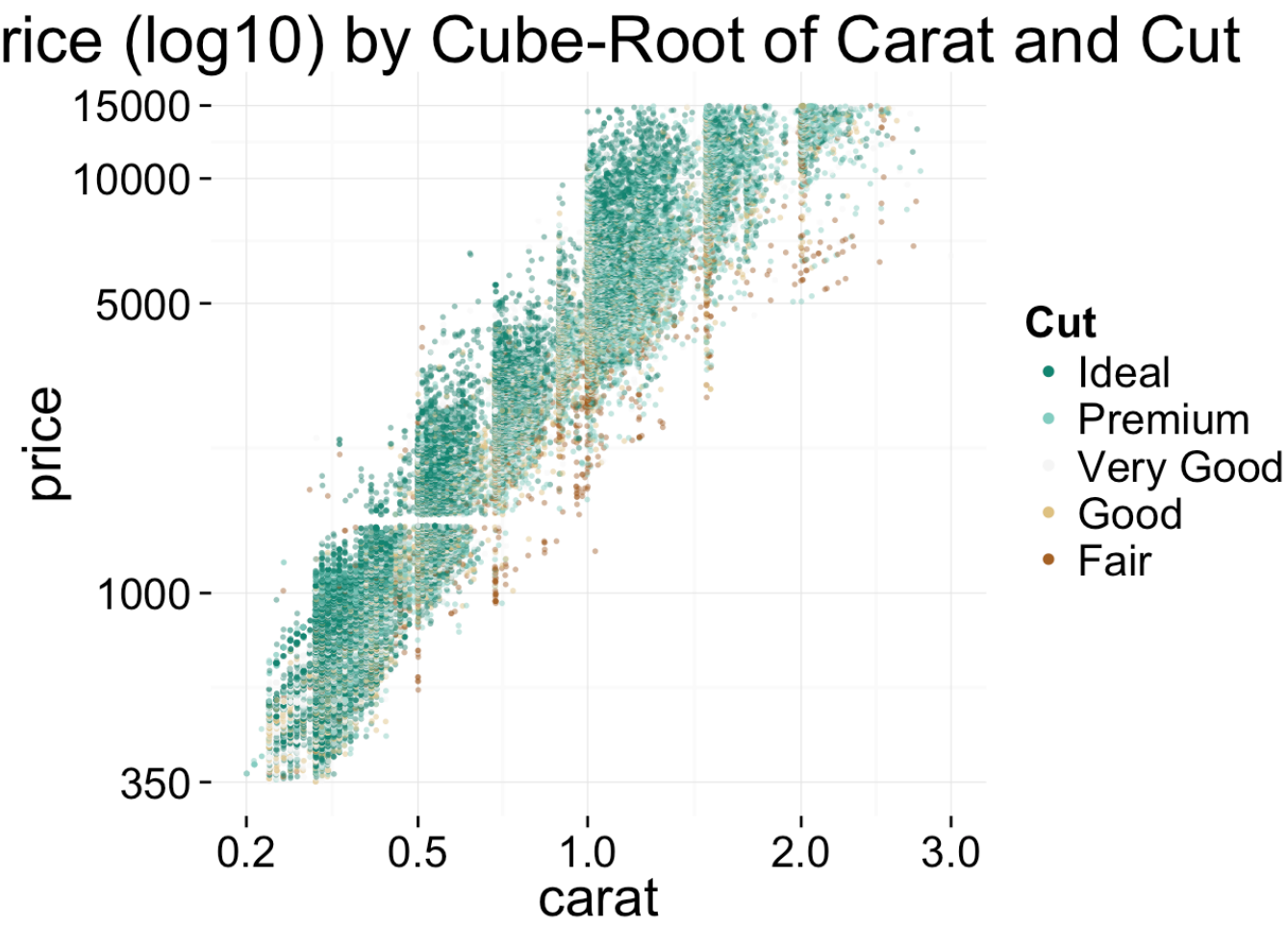
```
## Warning: Removed 1696 rows containing missing values (geom_point).
```



Holding carat weight constant, diamonds with lower clarity are almost always cheaper than diamonds with better clarity (worst clarity is I1 and best clarity is IF).

Price vs Carat and Cut

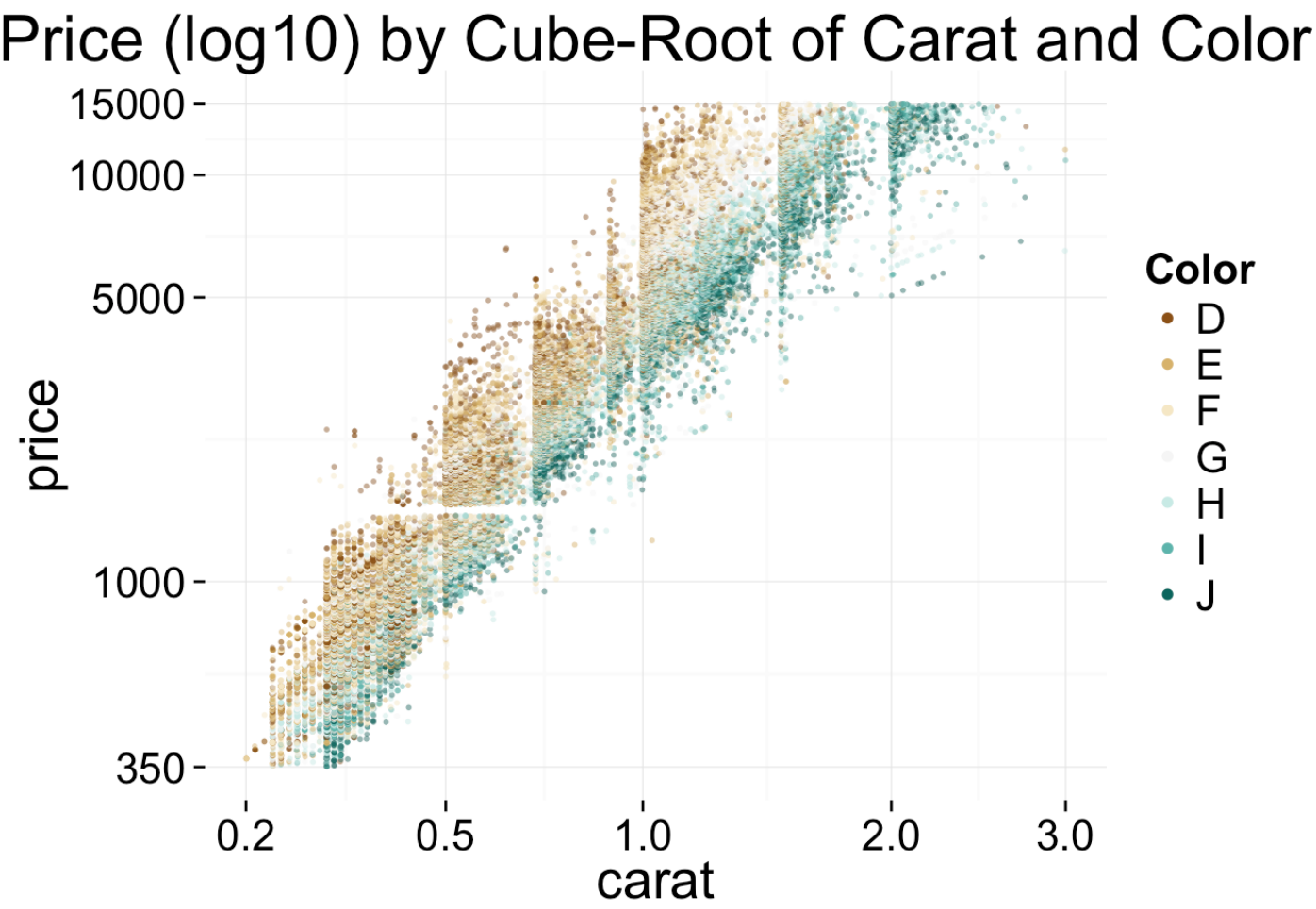
```
## Warning: Removed 1695 rows containing missing values (geom_point).
```



Price does not vary as much on cut holding carat constant; the pattern is not noticeable here.

Price vs Carat and Color

```
## Warning: Removed 1692 rows containing missing values (geom_point).
```



Color does seem to explain some of the variance in price as was the case with the clarity variable.

The last 3 plots suggest that we can build a linear model and use those variables in the linear model to predict the price of a diamond.

```
##
## Calls:
## m1: lm(formula = I(log(price)) ~ I(carat^(1/3)), data = diamonds)
## m2: lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat, data = diamonds)
## m3: lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + clarity,
##     data = diamonds)
## m4: lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + clarity +
##     cut, data = diamonds)
## m5: lm(formula = I(log(price)) ~ I(carat^(1/3)) + carat + clarity +
##     cut + color, data = diamonds)
##
## ================================================================================
##                        m1          m2          m3          m4          m5
## --------------------------------------------------------------------------------
## (Intercept)         2.821***    1.039***    0.464***    0.391***    0.415***
##                    (0.006)     (0.019)     (0.014)     (0.014)     (0.010)
## I(carat^(1/3))      5.558***    8.568***    9.319***    9.376***    9.144***
##                    (0.007)     (0.032)     (0.023)     (0.023)     (0.016)
## carat                          -1.137***   -1.260***   -1.274***   -1.093***
##                                (0.012)     (0.008)     (0.008)     (0.006)
## clarity: .L                                 0.889***    0.854***    0.907***
##                                            (0.005)     (0.005)     (0.003)
## clarity: .Q                                -0.255***   -0.239***   -0.240***
##                                            (0.005)     (0.005)     (0.003)
## clarity: .C                                 0.143***    0.129***    0.131***
##                                            (0.004)     (0.004)     (0.003)
## clarity: ^4                                -0.086***   -0.080***   -0.063***
##                                            (0.003)     (0.003)     (0.002)
## clarity: ^5                                 0.038***    0.034***    0.026***
##                                            (0.003)     (0.003)     (0.002)
## clarity: ^6                                 0.001       0.004      -0.002
##                                            (0.002)     (0.002)     (0.002)
## clarity: ^7                                 0.054***    0.051***    0.032***
##                                            (0.002)     (0.002)     (0.001)
## cut: .L                                                 0.125***    0.120***
##                                                        (0.003)     (0.002)
## cut: .Q                                                -0.034***   -0.031***
##                                                        (0.003)     (0.002)
## cut: .C                                                 0.016***    0.014***
##                                                        (0.002)     (0.002)
## cut: ^4                                                -0.001      -0.002
##                                                        (0.002)     (0.001)
## color: .L                                                          -0.441***
##                                                                    (0.002)
## color: .Q                                                          -0.093***
##                                                                    (0.002)
## color: .C                                                          -0.013***
##                                                                    (0.002)
## color: ^4                                                           0.012***
##                                                                    (0.002)
```

```
## color: ^5                                                        -0.003*
##                                                                  (0.001)
## color: ^6                                                         0.001
##                                                                  (0.001)
## ------------------------------------------------------------------------
## R-squared            0.924      0.935      0.967      0.968      0.984
## adj. R-squared       0.924      0.935      0.967      0.968      0.984
## sigma                0.280      0.259      0.185      0.181      0.129
## F               652012.063 387489.366 175093.345 125821.403 173791.084
## p                    0.000      0.000      0.000      0.000      0.000
## Log-likelihood   -7962.499  -3631.319  14605.945  15580.358  34091.272
## Deviance          4242.831   3613.360   1837.549   1772.344    892.214
## AIC              15930.999   7270.637 -29189.890 -31130.717 -68140.544
## BIC              15957.685   7306.220 -29092.038 -30997.282 -67953.736
## N                53940      53940      53940      53940      53940
## ========================================================================
```

The variables in this linear model can account for 98.4% of the variance in the price of diamonds.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

Ideal diamonds also 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.

Holding carat weight constant, diamonds with lower clarity are almost always cheaper than diamonds with better clarity (worst clarity is I1 and best clarity is IF).

The last 3 plots from the Multivariate section suggest that I can build a linear model and use those variables in the model to predict the price of a diamond.

Were there any interesting or surprising interactions between features?

Levels of cut cluster by table value. This resonates with me because I think certain diamond cuts would produce particular dimensions (x, y, and z). The pattern holds across each level of clarity and each level of color with the exception of the lowest clarity

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

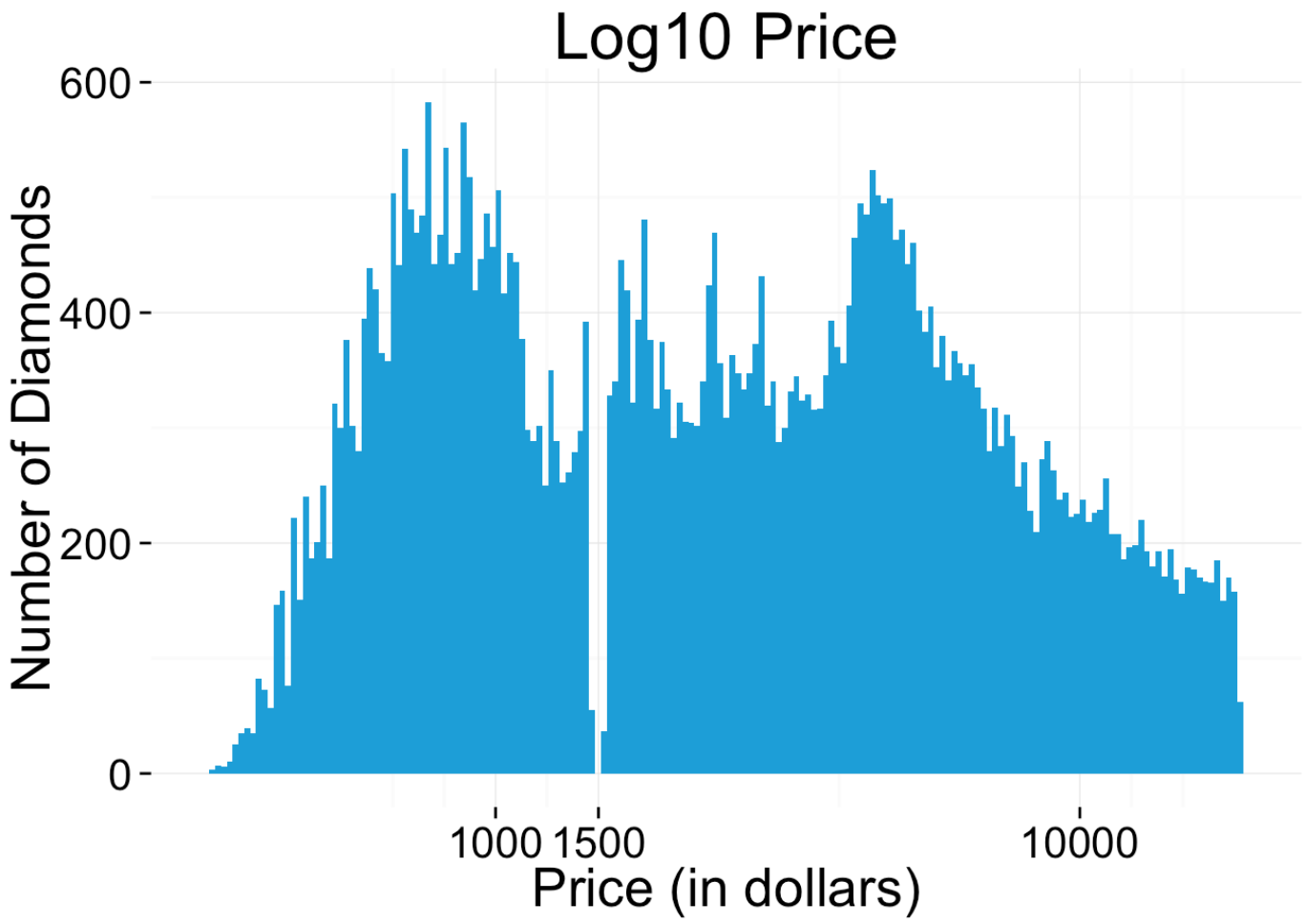
Yes, I created a linear model using the log10 of Price and the Cube-Root of Carat.

The variables in the linear model account for 98.4% of the variance in the price of diamonds. The addition of the cut variable to the model slightly improves the R^2 value by one tenth of a percent, which is expected based on the visualization above of Log10 Price vs. Cube-Root Carat and Cut.

Final Plots and Summary

Plot One

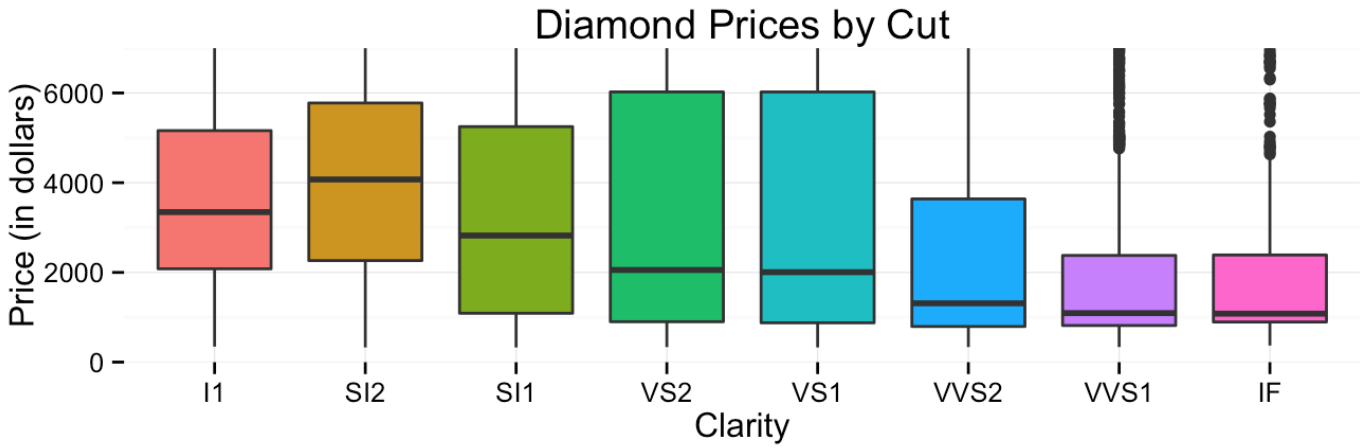
```
## Warning: position_stack requires constant width: output may be incorrect
```

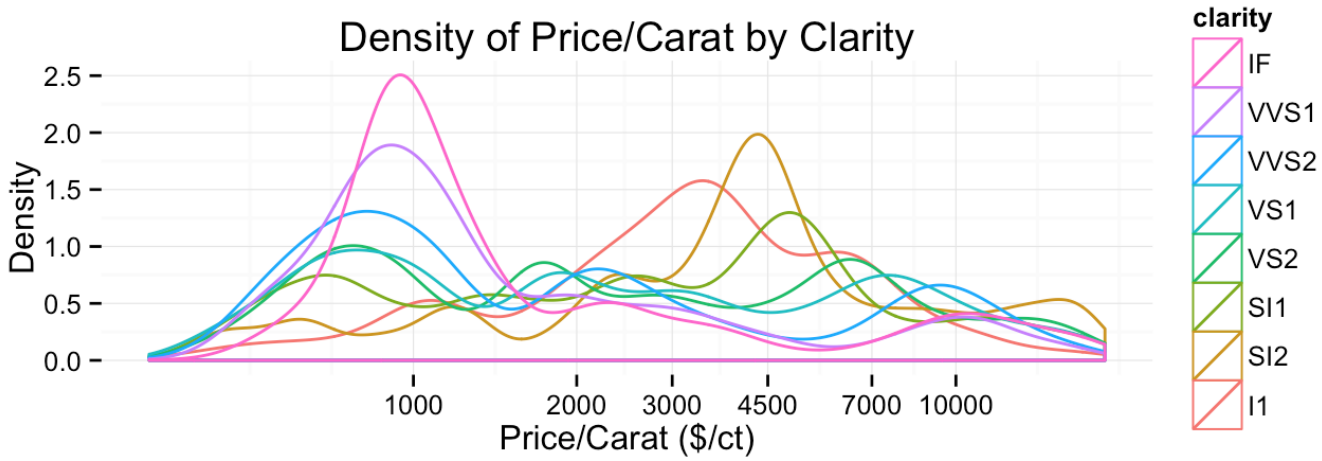


Description One

The distribution of diamond prices appears to be bimodal, perhaps due to the demand of diamonds and buyers purchasing in two different ranges of price points.

Plot Two



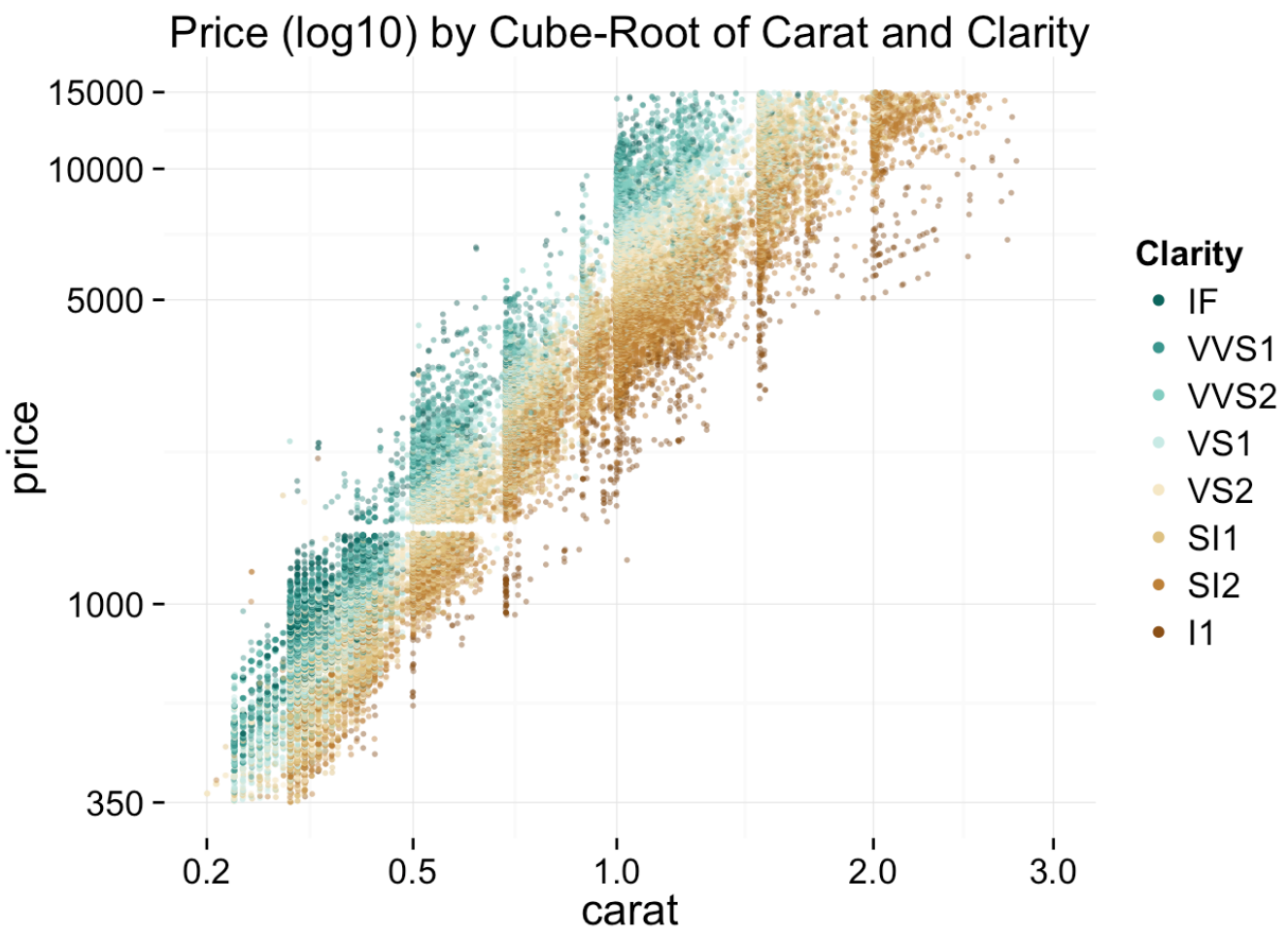


Description Two

Diamonds with the best level of clarity (IF) have the lowest median price. A greater proportion of diamonds with the best clarity are priced lowered compared to the proportion of diamonds in other price distributions for worse levels of clarity. Price variance increases as the clarity improves (worst clarity is I1).

Plot Three

```
## Warning: Removed 1695 rows containing missing values (geom_point).
```



Description Three

Holding carat weight constant, diamonds with higher clarity levels (I1 is worst and IF is best) are almost always cheaper than diamonds with better clarity. The plot indicates that a linear model could be constructed to predict the price of variables using log10(price) as the outcome variable and cube-root of carat as the predictor variable

---

Reflection

The diamonds data set contains information on almost 54,000 thousand diamonds from around 2008. I started by understanding the individual variables in the data set, and then I explored interesting questions and leads as I continued to make observations on plots. Eventually, I explored the price of diamonds across many variables and created a linear model to predict diamond prices. I was surprised that depth or table did not have a strong positive correlation with price, but these variables are likely to be represented by categorical variables: color, cut, and clarity. I struggled understanding the decrease in median price as the level of cut and clarity improved, but this became more clear when I realized that most of the data contained ideal cut diamonds. For the linear model, all diamonds were included since information on price, carat, color, clarity, and cut were available for all the diamonds. Some limitations of this model include the source of the data. Given that the diamonds date to 2008, the model would likely undervalue diamonds in

the market today, either due to changes in demand and supply or inflation rates. To investigat this data further, I would examine how values of 0 were introduced into the data set for the variables volume, x, y, and z. I would be interested in testing the linear model to predict current diamond prices and to determine to what extent the model is accurate at pricing diamonds. A more recent data would be better to make predictions of diamond prices, and comparisons might be made between the other linear models to see if other variables may account for diamond prices.