Stat405

Displaying distributions

Hadley Wickham

1. The diamonds data
2. Histograms and bar charts
3. Homework

Diamonds

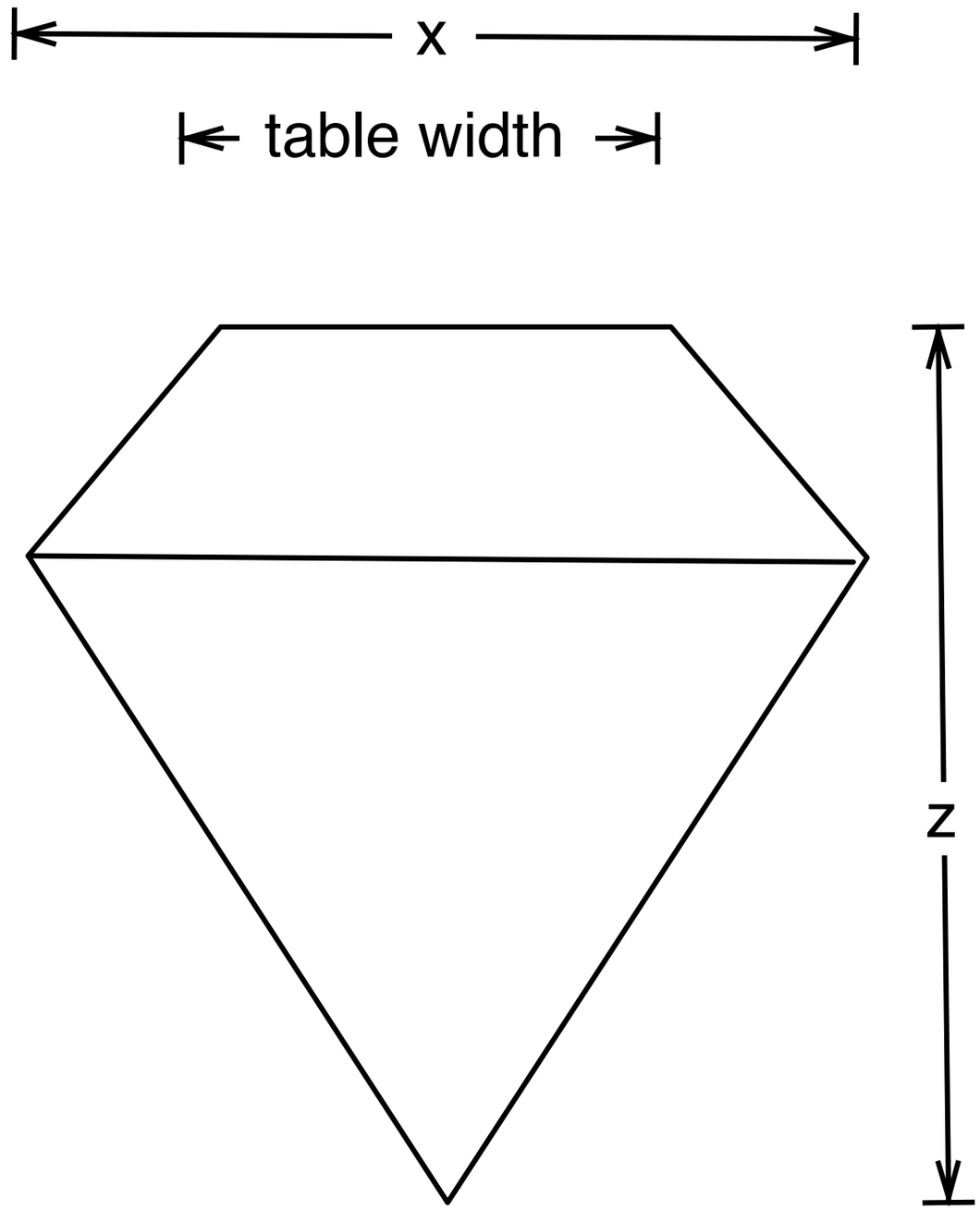
Diamonds data

~**54,000** round diamonds from http://www.diamondse.info/

Carat, colour, clarity, cut

Total depth, table, depth, width, height

Price



depth = z / diameter
table = table width / x * 100

Recall

Write down five ways to inspect the diamonds dataset.

You have one minute!

Histogram & bar charts

Histograms and barcharts

Used to display the **distribution** of a variable

Categorical variable → bar chart

Continuous variable → histogram

```
# With only one variable, qplot guesses that
# you want a bar chart or histogram
qplot(cut, data = diamonds)

qplot(carat, data = diamonds)

# Change binwidth:
qplot(carat, data = diamonds, binwidth = 1)
qplot(carat, data = diamonds, binwidth = 0.1)
qplot(carat, data = diamonds, binwidth = 0.01)
resolution(diamonds$carat)

last_plot() + xlim(0, 3)
```

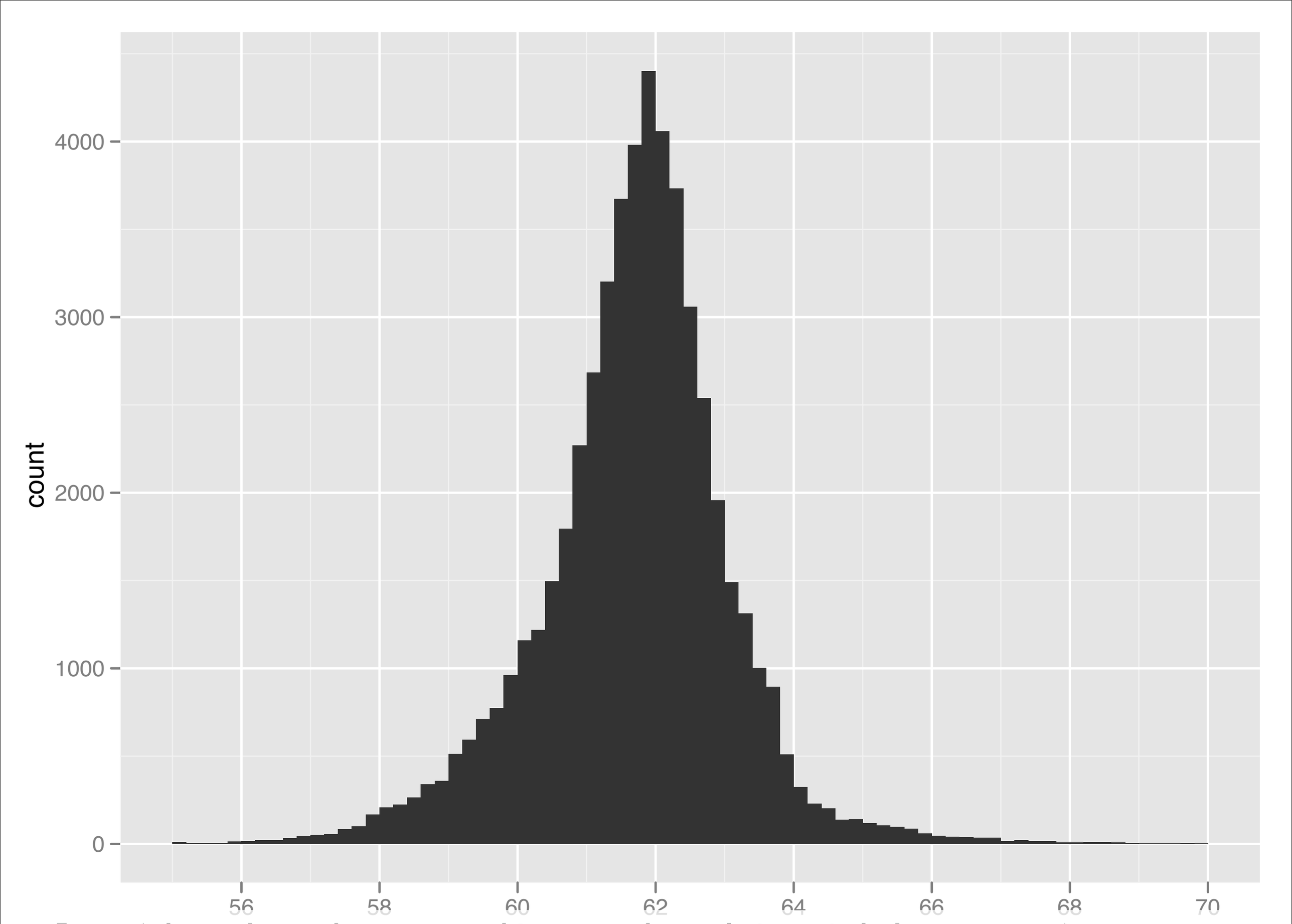
Always experiment with the bin width!

```
qplot(table, data = diamonds, binwidth = 1)

# To zoom in on a plot region use xlim() and ylim()
qplot(table, data = diamonds, binwidth = 1) +
   xlim(50, 70)
qplot(table, data = diamonds, binwidth = 0.1) +
  xlim(50, 70)
qplot(table, data = diamonds, binwidth = 0.1) +
  xlim(50, 70) + ylim(0, 50)

# Note that this type of zooming discards data
# outside of the plot regions. See
# ?coord_cartesian() for an alternative
```

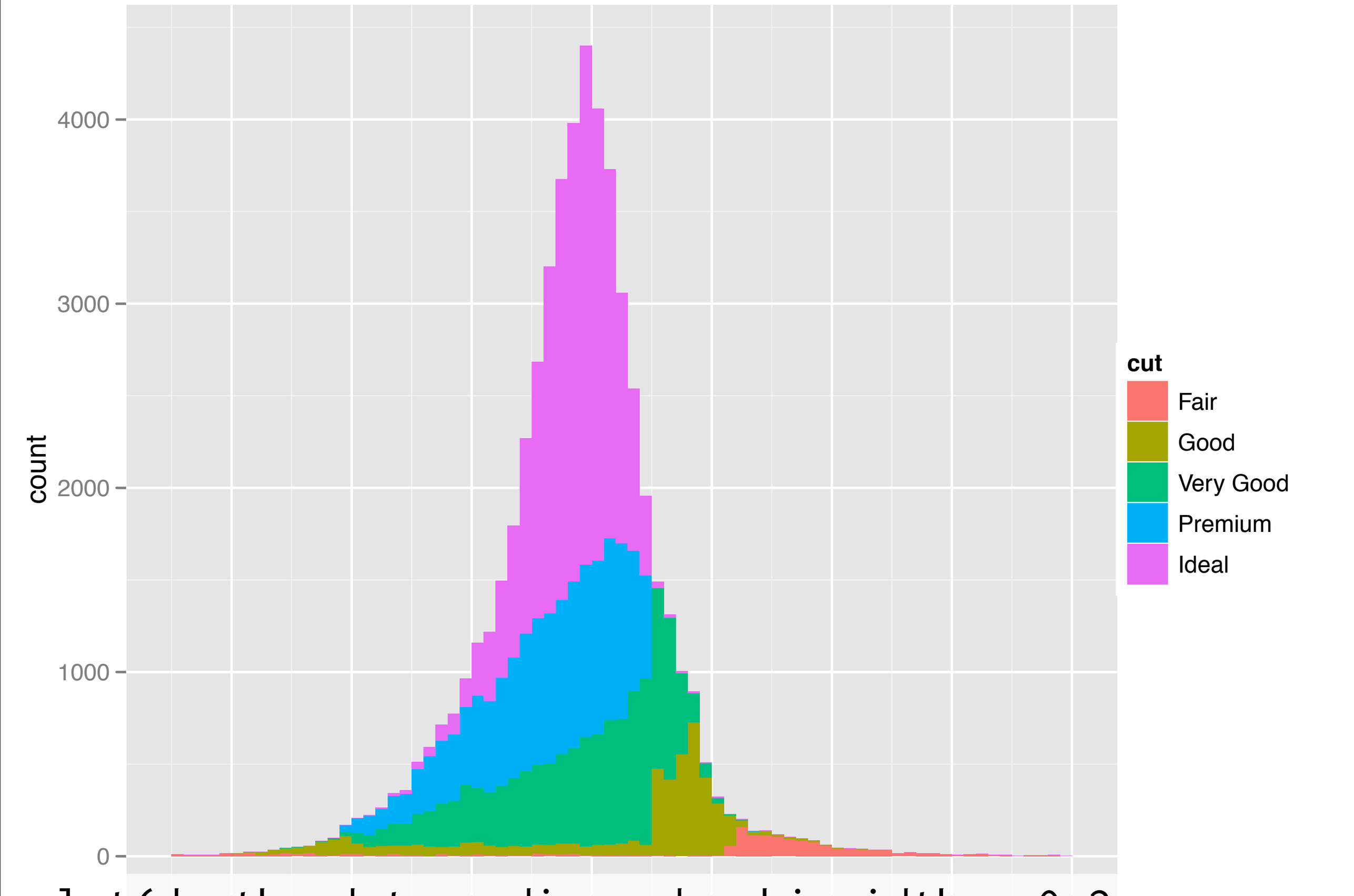
Additional variables

As with scatterplots can use **aesthetics** or **faceting**. Using aesthetics creates pretty, but ineffective, plots.

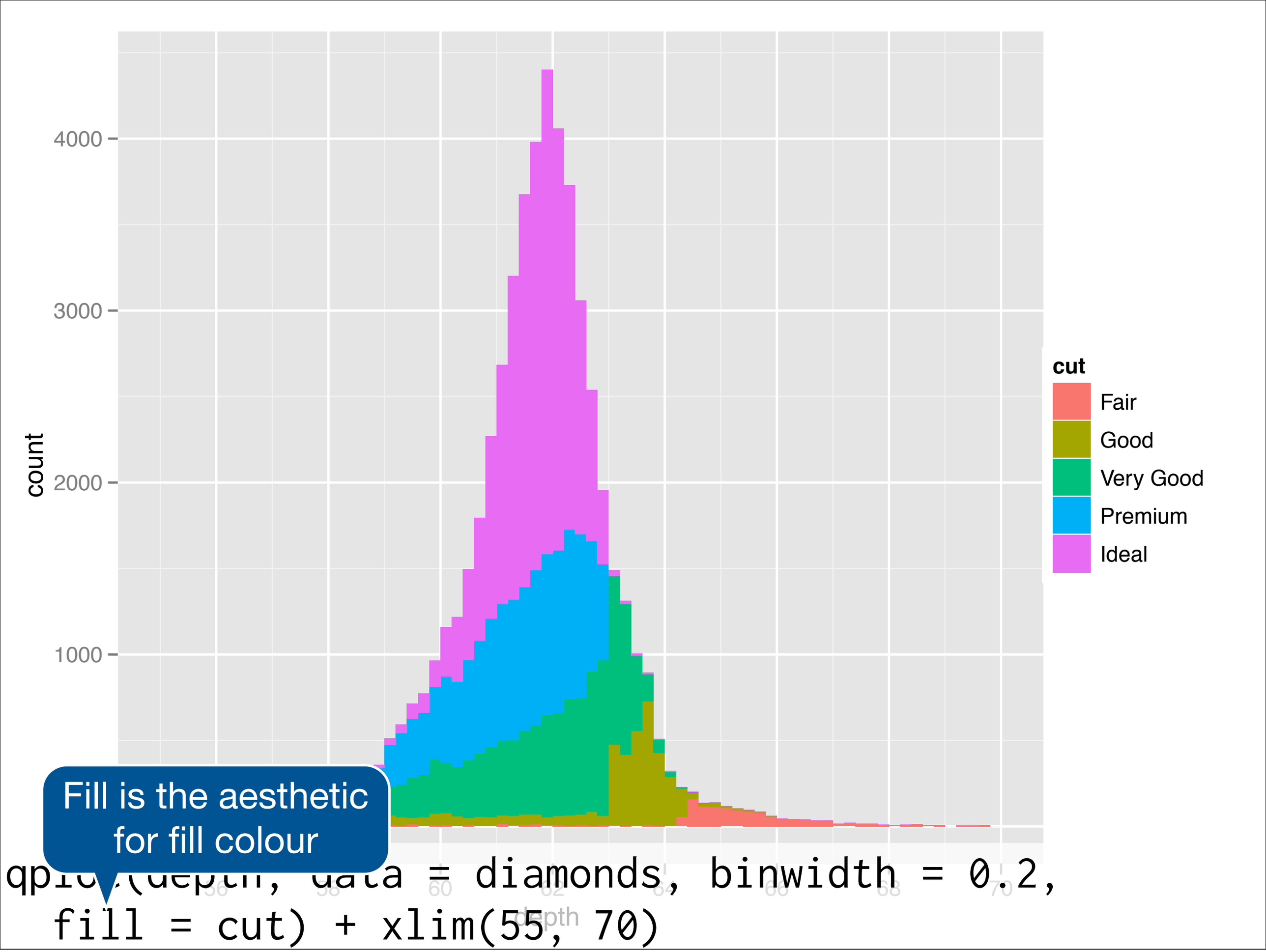
The following examples show the difference, when investigation the relationship between cut and depth.



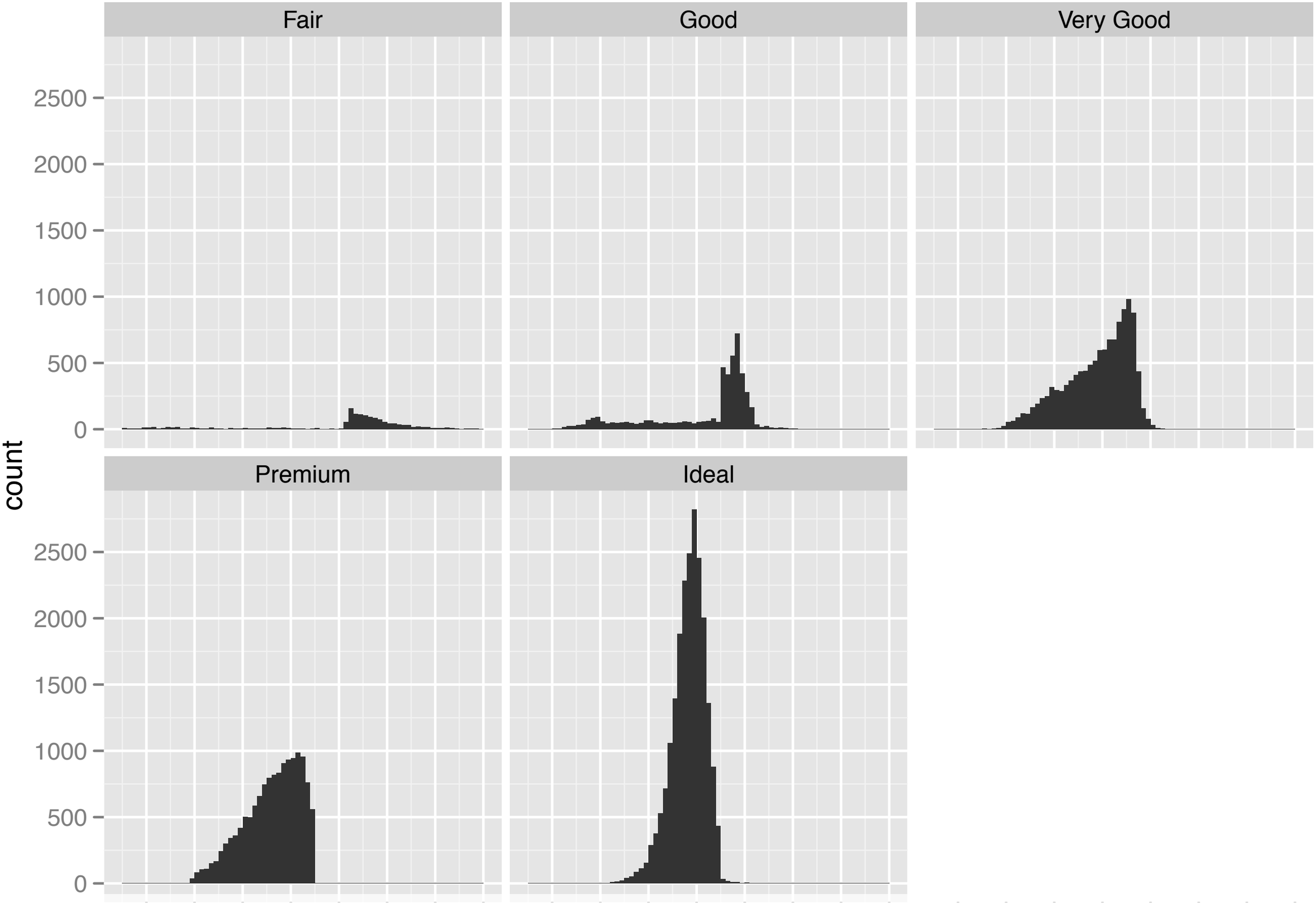

```
qplot(depth, data = diamonds, binwidth = 0.2)
```




```
qplot(depth, data = diamonds, binwidth = 0.2,
  fill = cut) + xlim(55, 70)
```




```
qplot(depth, data = diamonds, binwidth = 0.2,
  fill = cut) + xlim(55, 70)
```

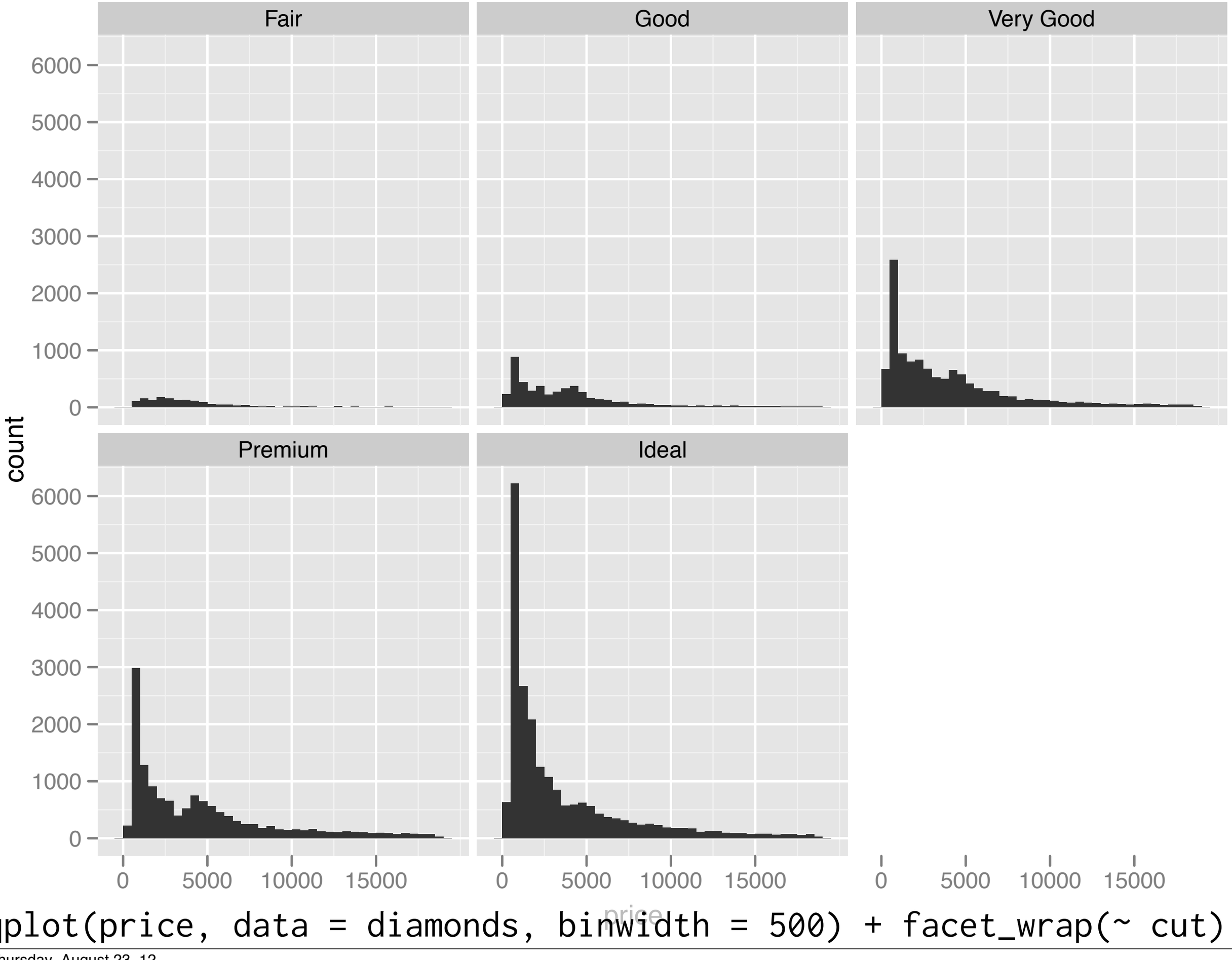



```
qplot(depth, data = diamonds, binwidth = 0.2) +
  xlim(55, 70) + facet_wrap(~cut)
```

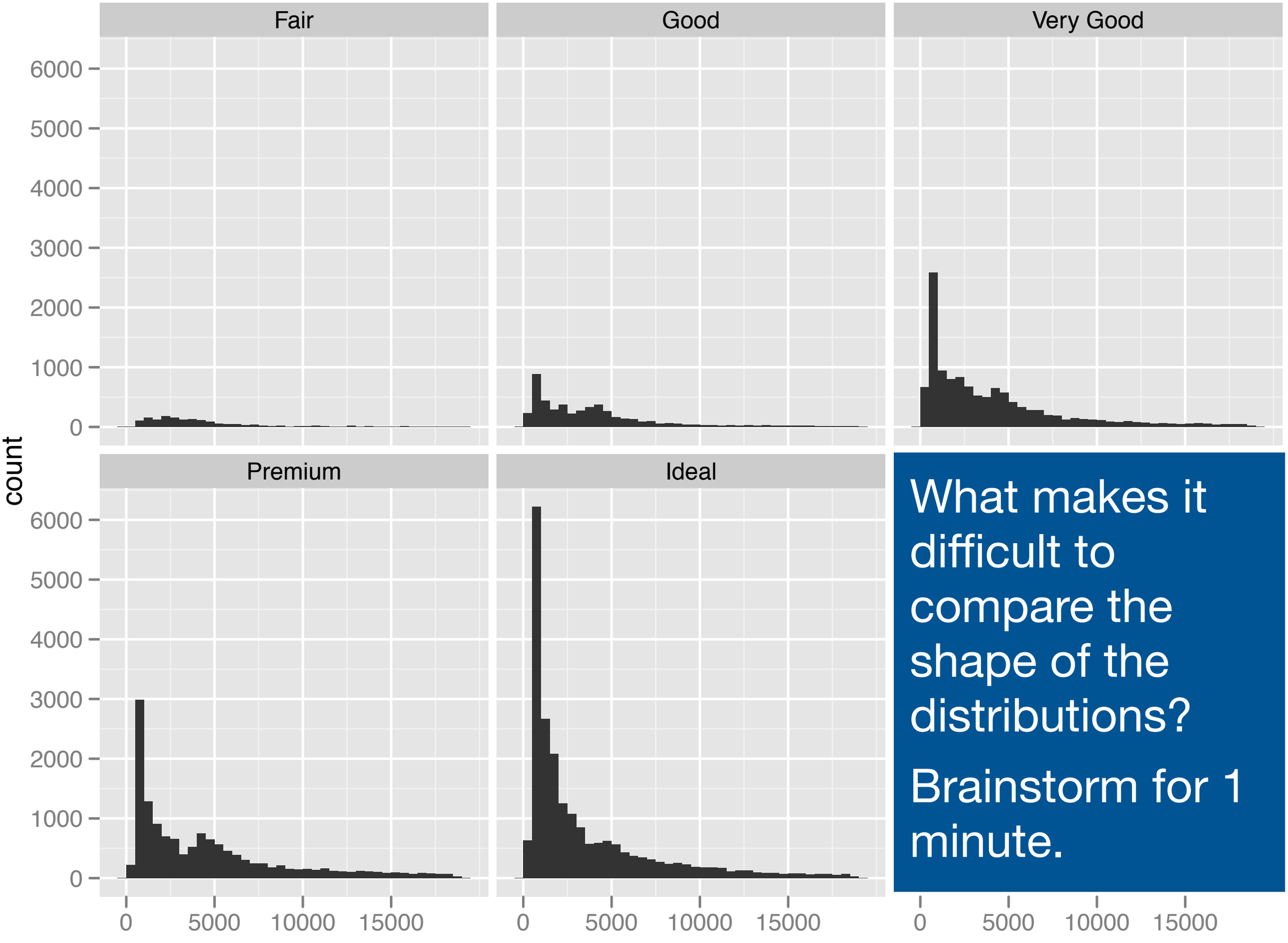
Your turn

Explore the distribution of price. What is a good binwidth to use? (**Hint**: How many bins will a binwidth of 1 give you?) Practice zooming in on regions of interest.

How does price vary with colour, cut, or clarity?




```
qplot(price, data = diamonds, binwidth = 500) + facet_wrap(~ cut)
```



What makes it difficult to compare the shape of the distributions?

Brainstorm for 1 minute.

```
qplot(price, data = diamonds, binwidth = 500) + facet_wrap(~ cut)
```

Problems

Each histogram far away from the others, but we know stacking is hard to read → *use another way of displaying densities*

Varying relative abundance makes comparisons difficult → *rescale to ensure constant area*

```
# Large distances make comparisons hard
qplot(price, data = diamonds, binwidth = 500) +
  facet_wrap(~ cut)

# Stacked heights hard to compare
qplot(price, data = diamonds, binwidth = 500, fill = cut)

# Much better - but still have differing relative abundance
qplot(price, data = diamonds, binwidth = 500,
  geom = "freqpoly", colour = cut)

# Instead of displaying count on y-axis, display density
# .. indicates that variable isn't in original data
qplot(price, ..density.., data = diamonds, binwidth = 500,
  geom = "freqpoly", colour = cut)

# To use with histogram, you need to be explicit
qplot(price, ..density.., data = diamonds, binwidth = 500,
  geom = "histogram") + facet_wrap(~ cut)
```

Aside: coding strategy

At the end of each interactive session, you want a summary of everything you did. Two options:

1. Copy from the history panel.
2. Build up the important bits as you go. (recommended)

Homework

Asking questions

You have two minutes to write down as many interesting questions as you can about the diamonds data.

Share them with your neighbour.

Try not to make them too simple (what's the biggest diamond) or too complex.

Homework

Create three plots that reveal something interesting about the diamonds or mpg data. Throwing out 90% of the data can be a valid strategy.

Include code, plot and paragraph of text.

For one plot, show the process of iteration by which you reached the final answer.

Grading

You will be graded out of five on three dispositions of a good data analyst: **curiosity**, **scepticism** and **organisation**.

The same rubric will be used throughout the semester, so it is very hard to get a 4 or 5 on your first homework.

Common mistakes

Too much in one plot.

Broad and shallow instead of deep.

Failed to cite resources used – if you get an image from the web, provide the url.

Incorrect aspect ratio.

Failed to deal with overplotting.