Exploratory data analysis

- Generate questions about your data.
- Search for answers by visualizing, transforming, and modelling your data.
- Use what you learn to refine your questions and/or generate new questions.

Exploratory data analysis

"There are no routine statistical questions, only questionable statistical routines."

— Sir David Cox

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

— John Tukey

Exploratory data analysis

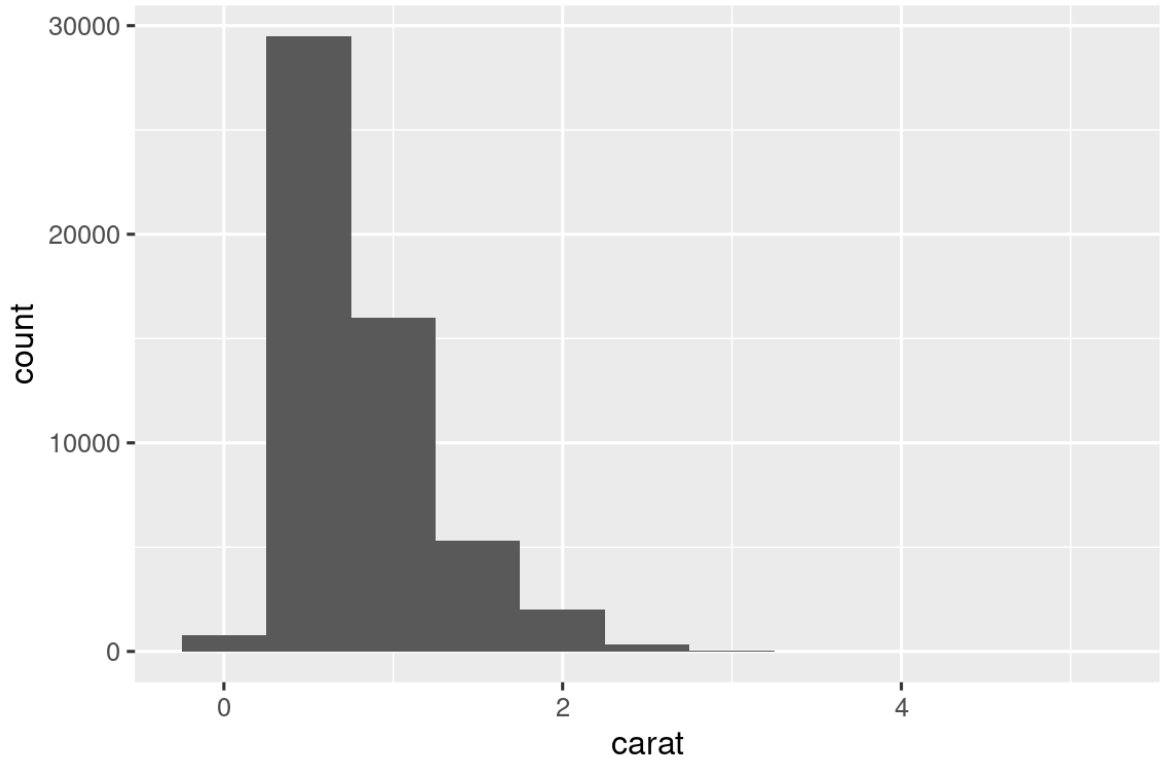
The key to asking *quality* questions is to generate a large *quantity* of questions.

The basic questions

- What type of variation occurs within my variables?
- What type of covariation occurs between my variables?

diamonds

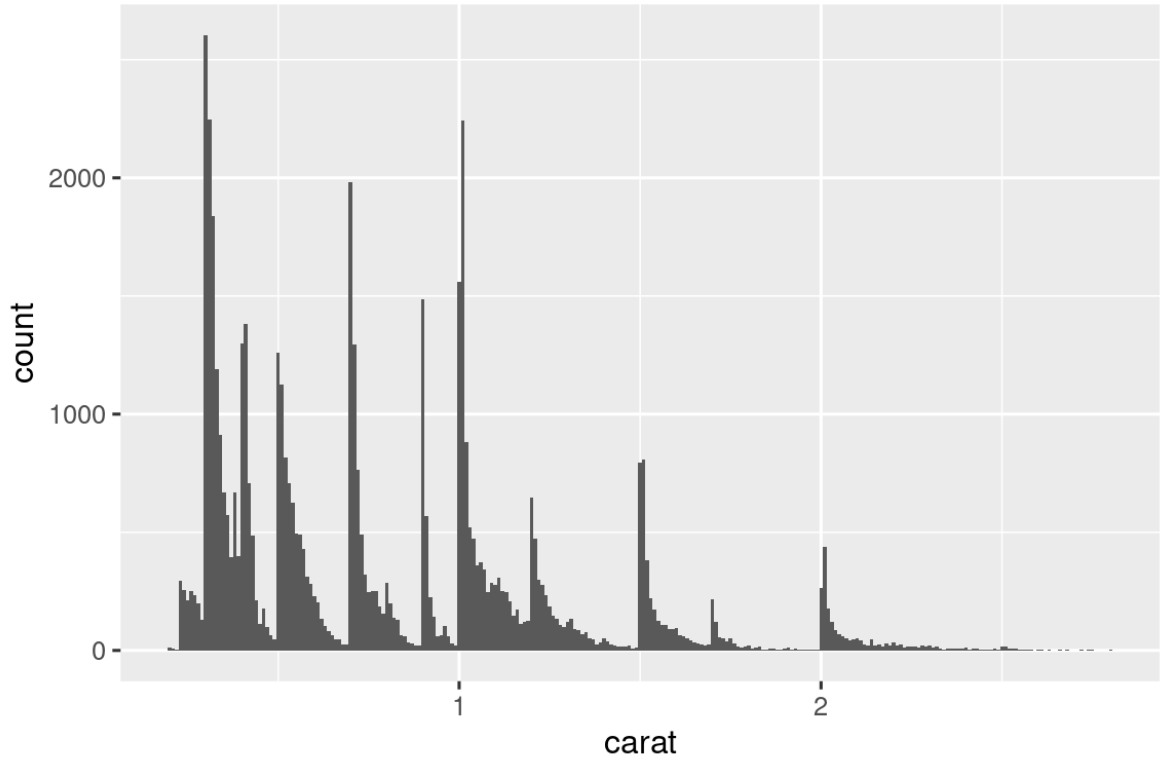
```
ggplot(diamonds, aes(x = carat)) +
  geom_histogram(binwidth = 0.5)
```



- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

diamonds

```
smaller <- diamonds |>
  filter(carat < 3)

ggplot(smaller, aes(x = carat)) +
  geom_histogram(binwidth = 0.01)
```



- Why are there more diamonds at whole carats and common fractions of carats?
- Why are there more diamonds slightly to the right of each peak than there are slightly to the left of each peak?

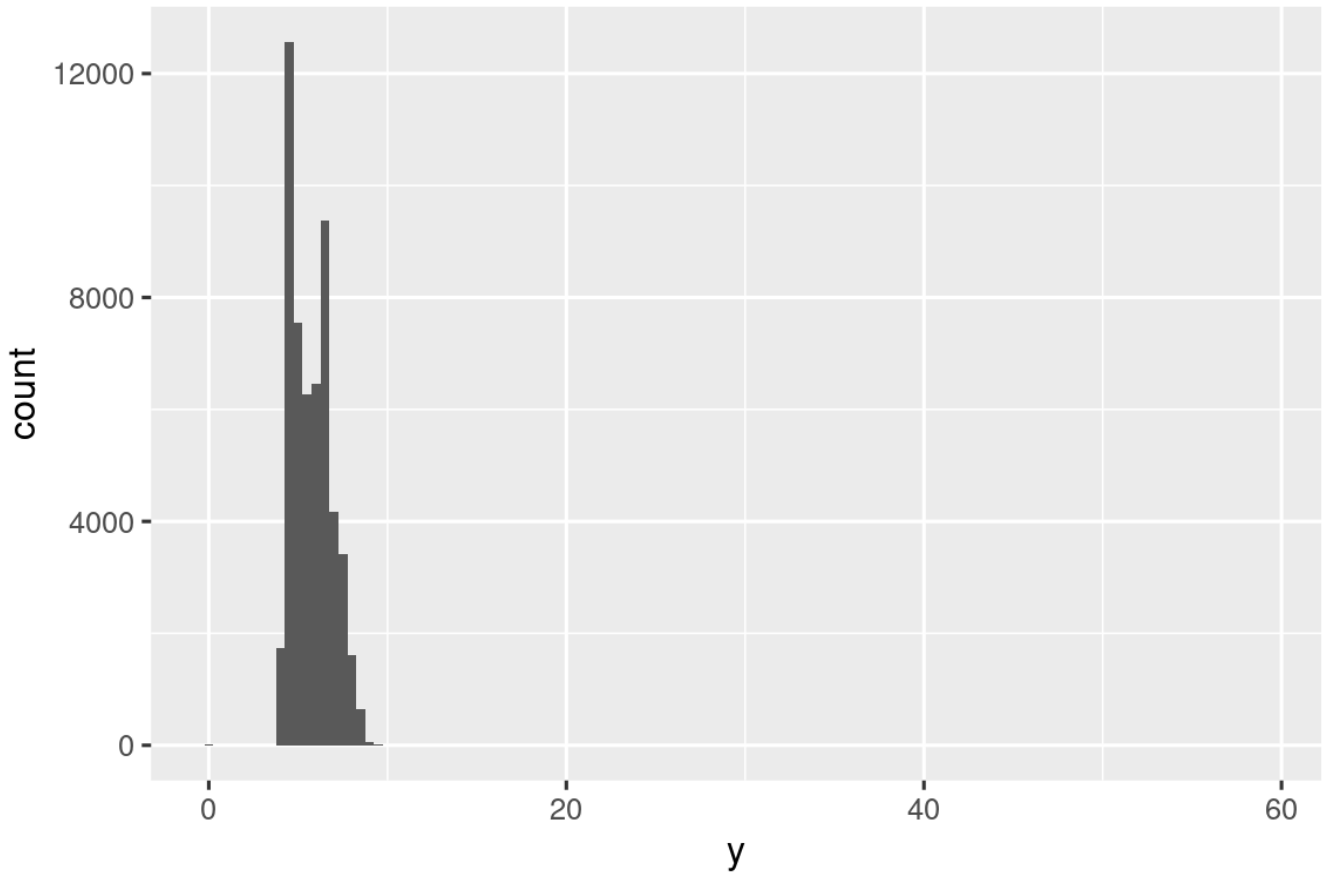
Understand subgroups

- How are the observations within each subgroup similar to each other?
- How are the observations in separate clusters different from each other?
- How can you explain or describe the clusters?
- Why might the appearance of clusters be misleading?

Some of these questions can be answered with the data while some will require domain expertise about the data.

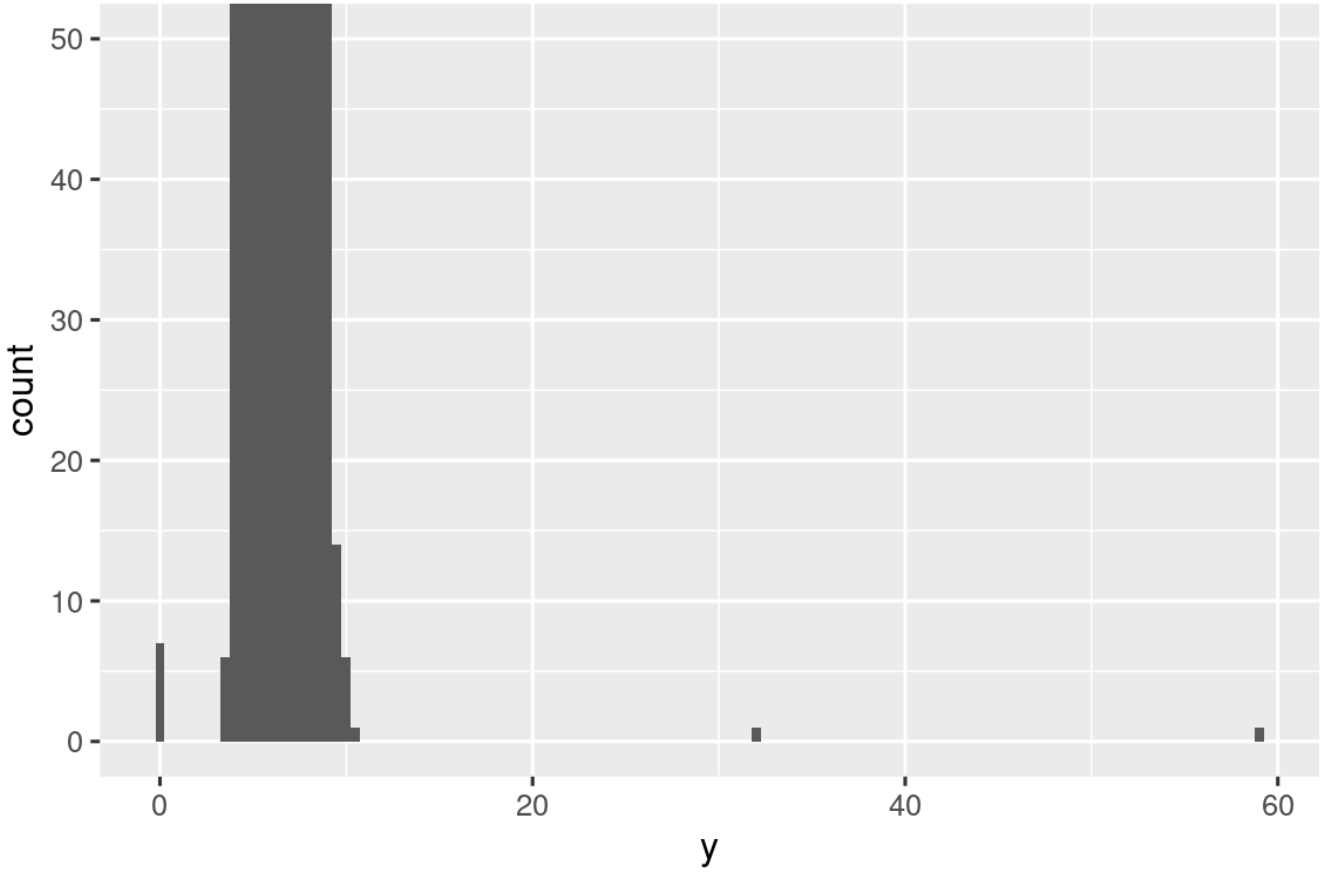
Outliers

```
ggplot(diamonds, aes(x = y)) +
  geom_histogram(binwidth = 0.5)
```



Zoom in to see small boxes

```
ggplot(diamonds, aes(x = y)) +
  geom_histogram(binwidth = 0.5) +
  coord_cartesian(ylim = c(0, 50))
```



Examine the outlier data

```
unusual <- diamonds |>
  filter(y < 3 | y > 20) |>
  select(price, x, y, z) |>
  arrange(y)
unusual
#> # A tibble: 9 × 4
#>   price     x     y     z
#>   <int> <dbl> <dbl> <dbl>
#> 1  5139  0      0     0
#> 2  6381  0      0     0
#> 3 12800  0      0     0
#> 4 15686  0      0     0
#> 5 18034  0      0     0
#> 6  2130  0      0     0
#> 7  2130  0      0     0
#> 8  2075  5.15  31.8   5.12
#> 9 12210  8.09  58.9   8.06
```

- We know dimensions can't be zero.
- 31.8 and 58.9 are highly suspicious.
- We might want to replace all these with NAs.

Options for suspicious data

- Drop the entire row

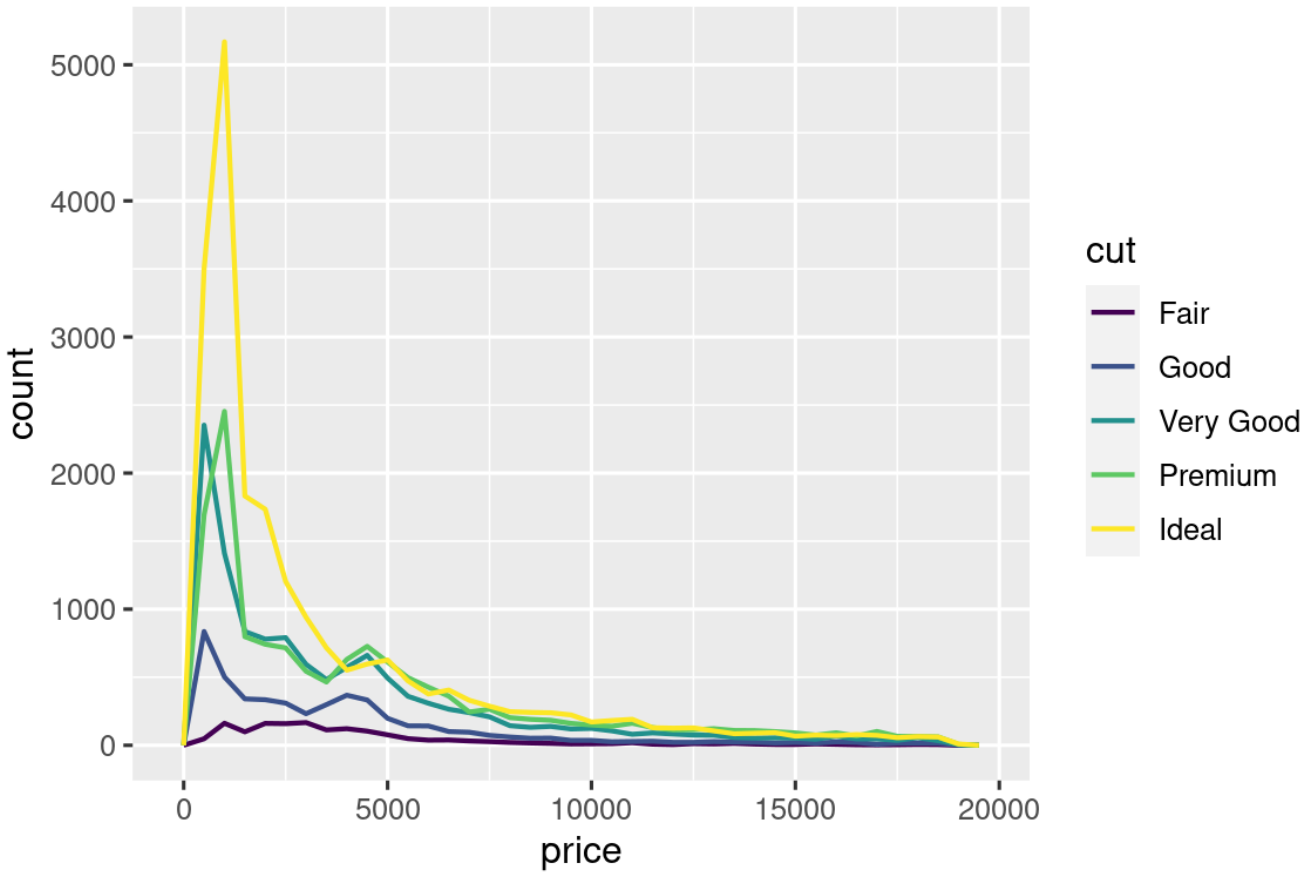
```
diamonds2 <- diamonds |>
  filter(between(y, 3, 20))
```

- Replace with NAs

```
diamonds2 <- diamonds |>
  mutate(y = if_else(y < 3 | y > 20, NA, y))
```

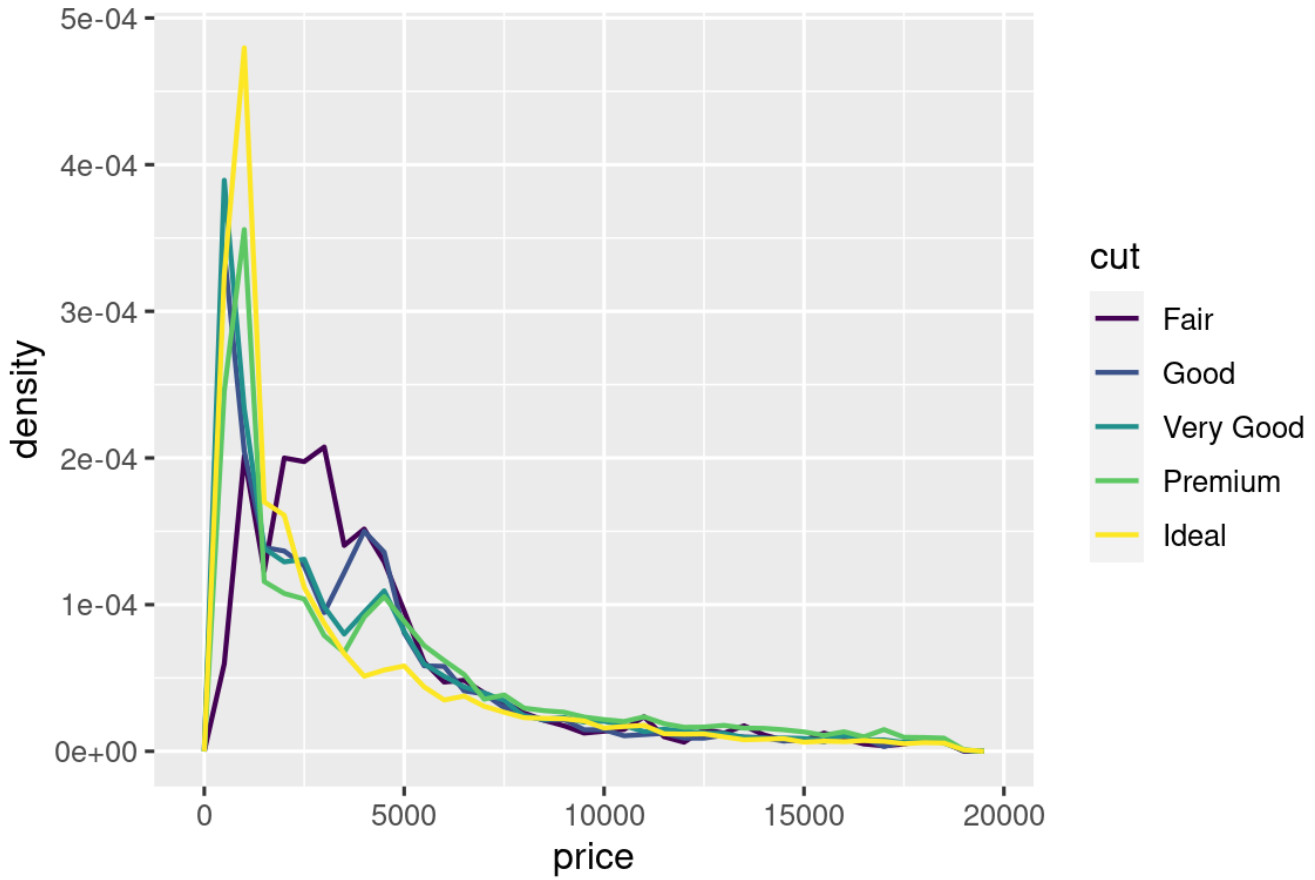
Covariation between categorical and a numerical variable

```
ggplot(diamonds, aes(x = price)) +
  geom_freqpoly(aes(color = cut), binwidth = 500, linewidth = 0.75)
```



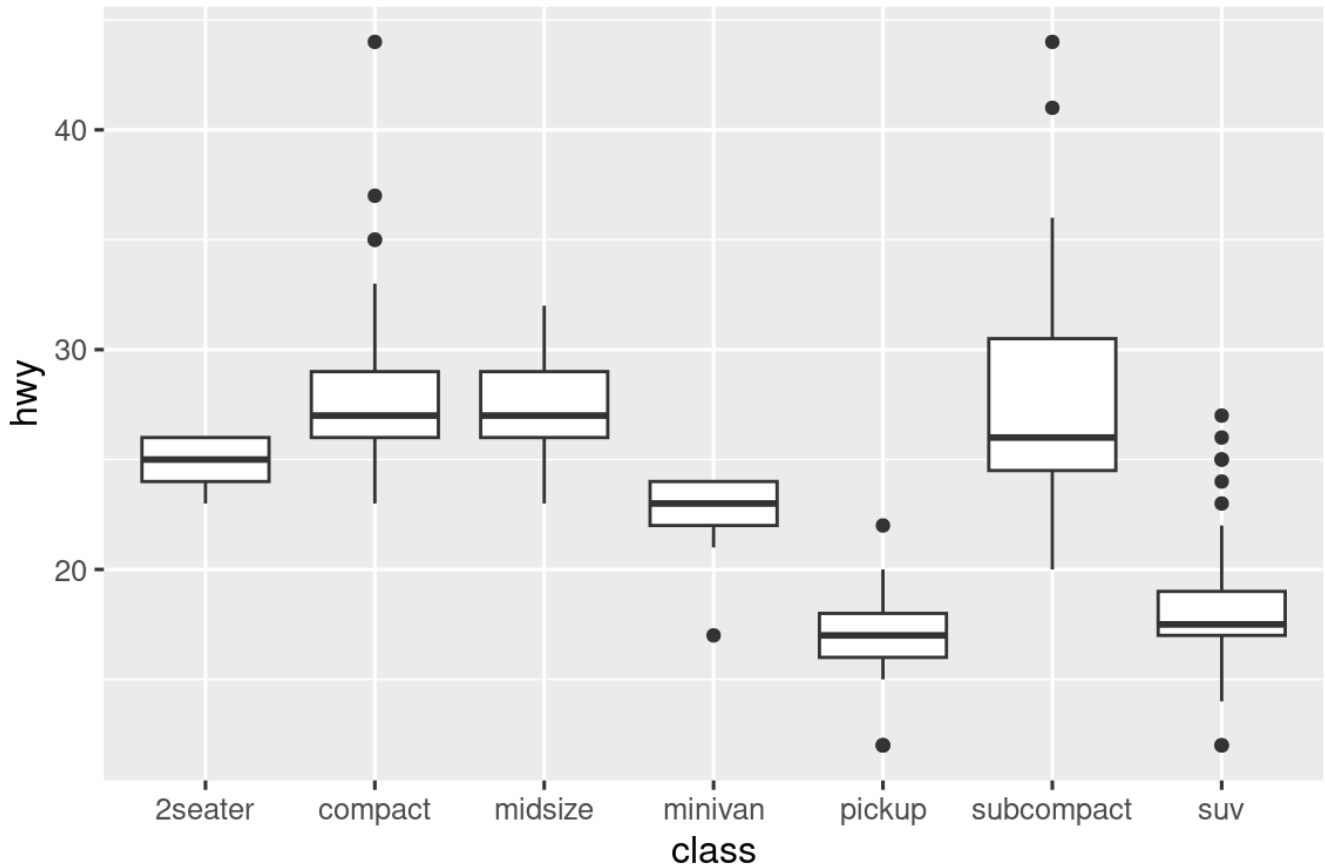
Covariation between categorical and a numerical variable

```
ggplot(diamonds, aes(x = price, y = after_stat(density))) +
  geom_freqpoly(aes(color = cut), binwidth = 500, linewidth = 0.75)
```



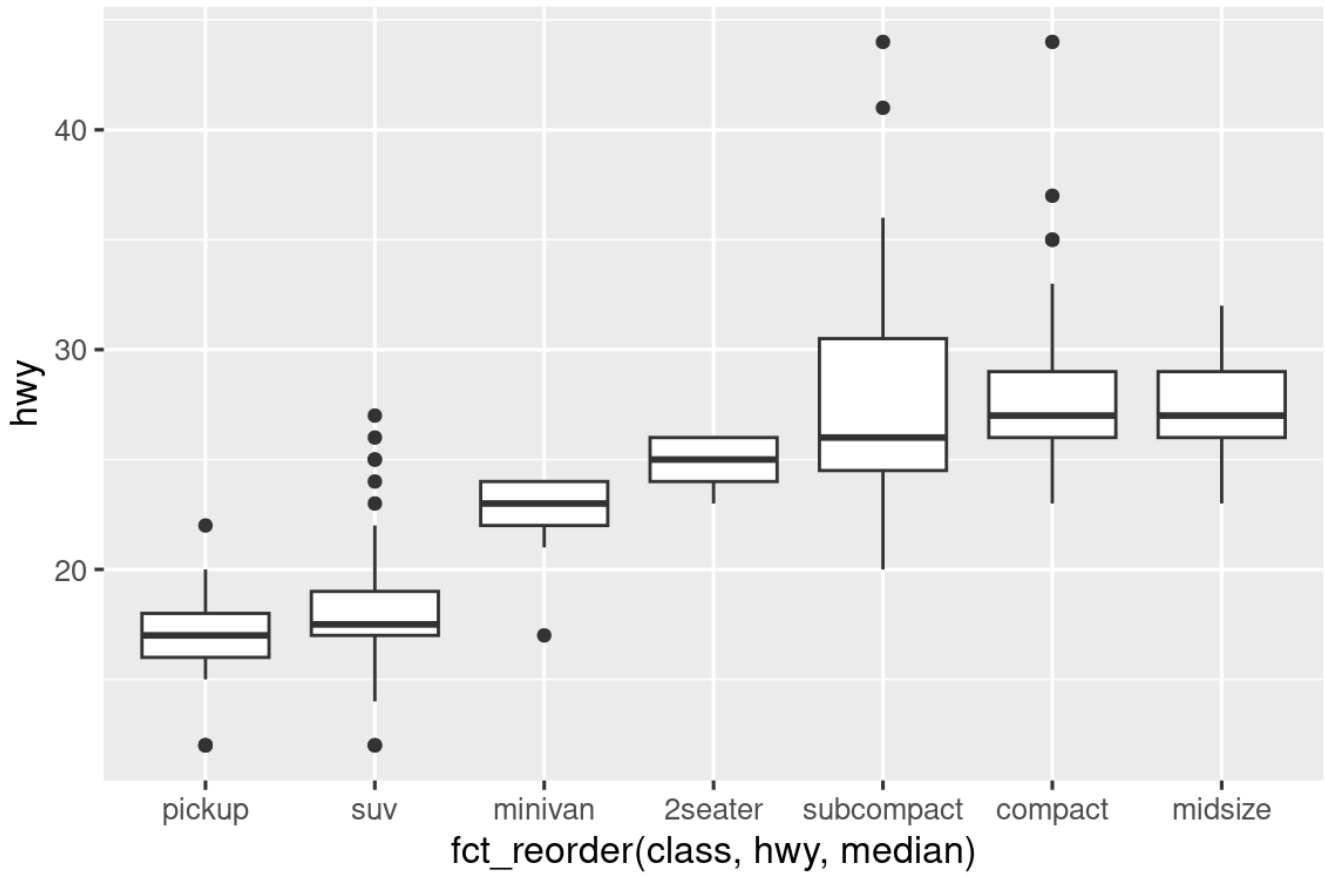
Reorder factors

```
ggplot(mpg, aes(x = class, y = hwy)) +
  geom_boxplot()
```



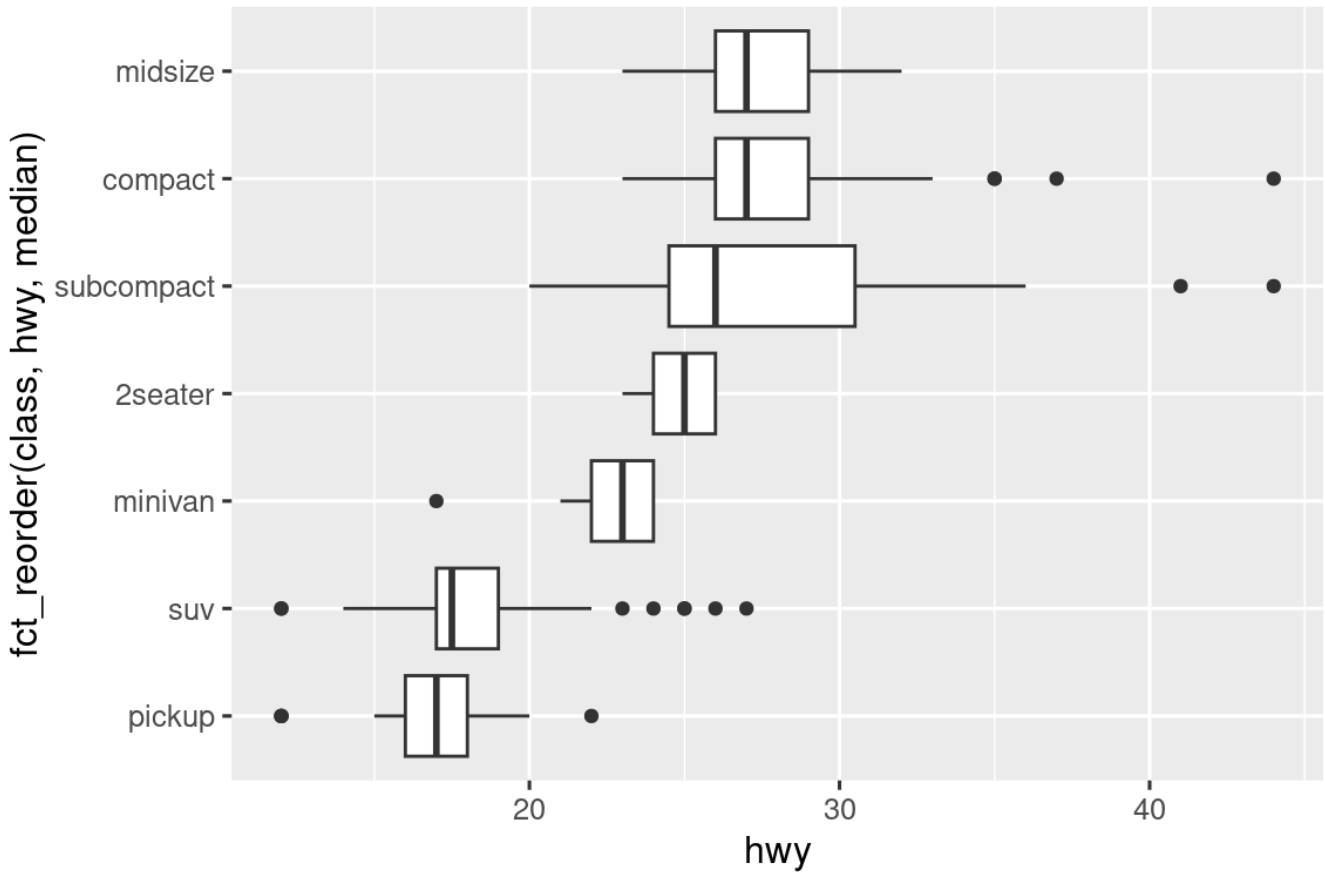
Reorder factors

```
ggplot(mpg, aes(x = fct_reorder(class, hwy, median), y = hwy)) +
  geom_boxplot()
```



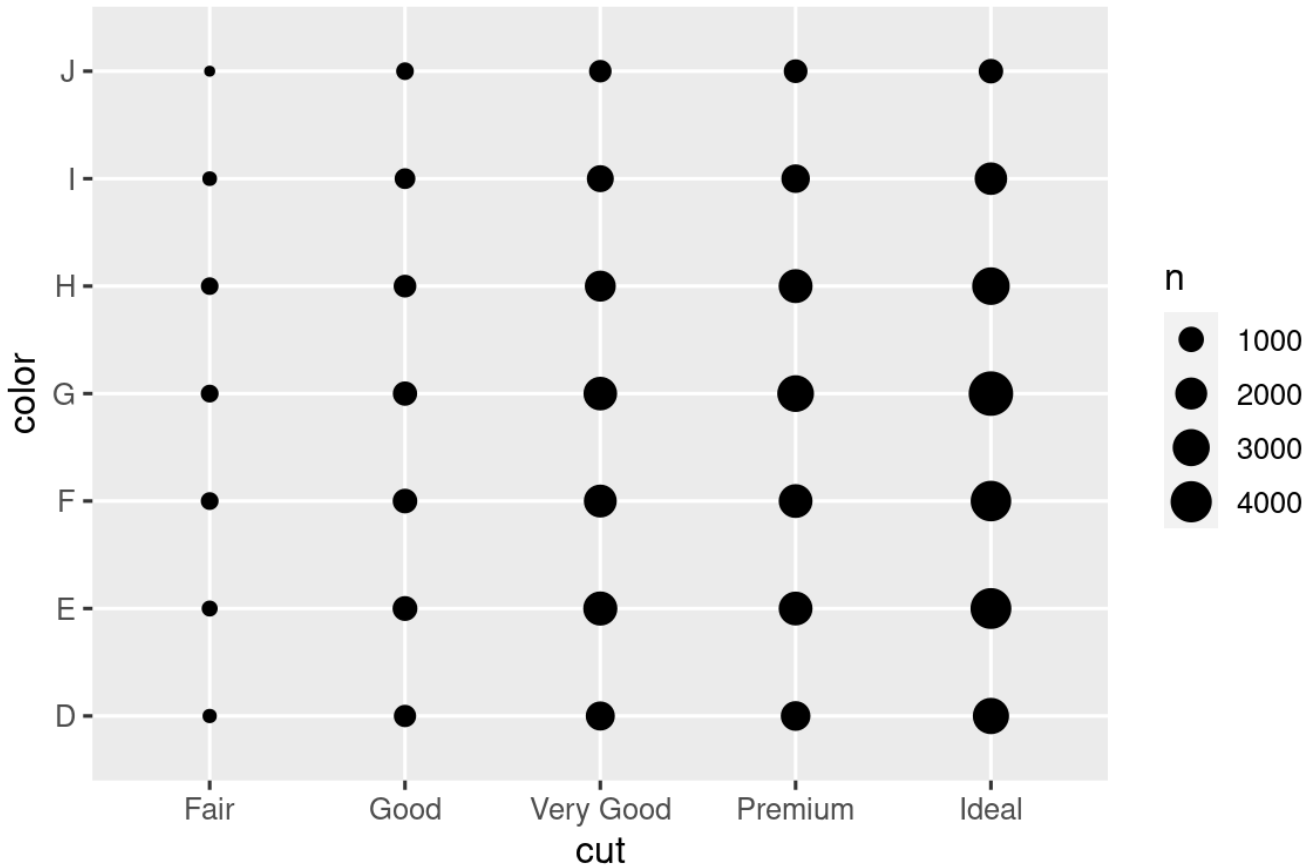
Long names

```
ggplot(mpg, aes(x = hwy, y = fct_reorder(class, hwy, median))) +
  geom_boxplot()
```



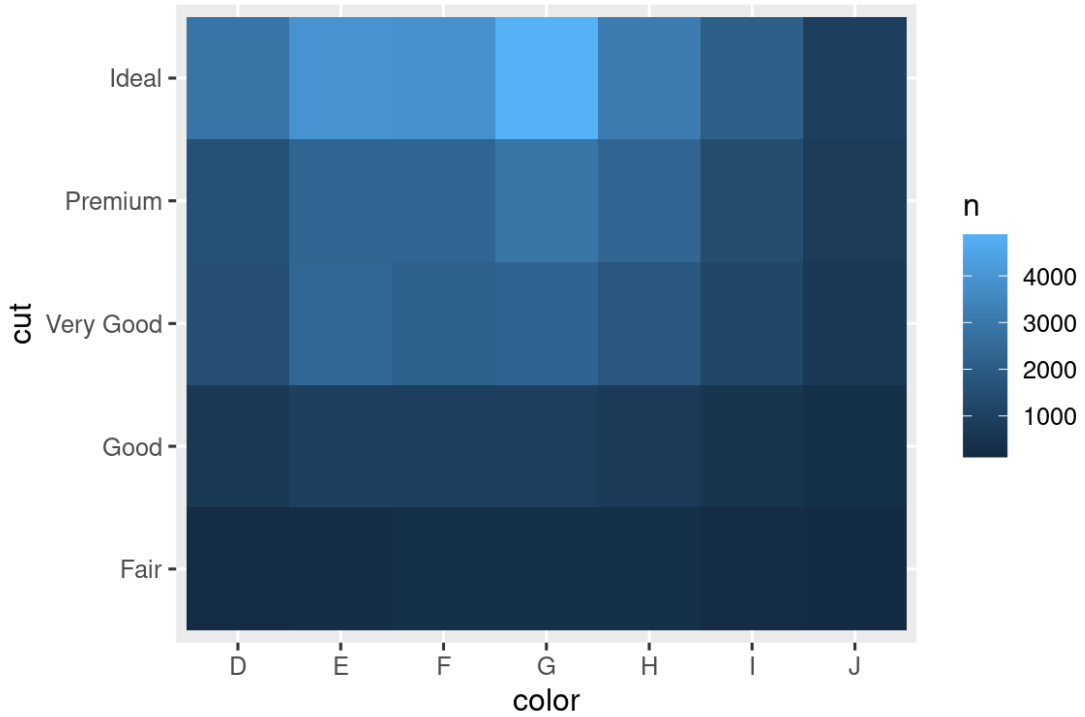
Two categorical variables

```
ggplot(diamonds, aes(x = cut, y = color)) +
  geom_count()
```



count

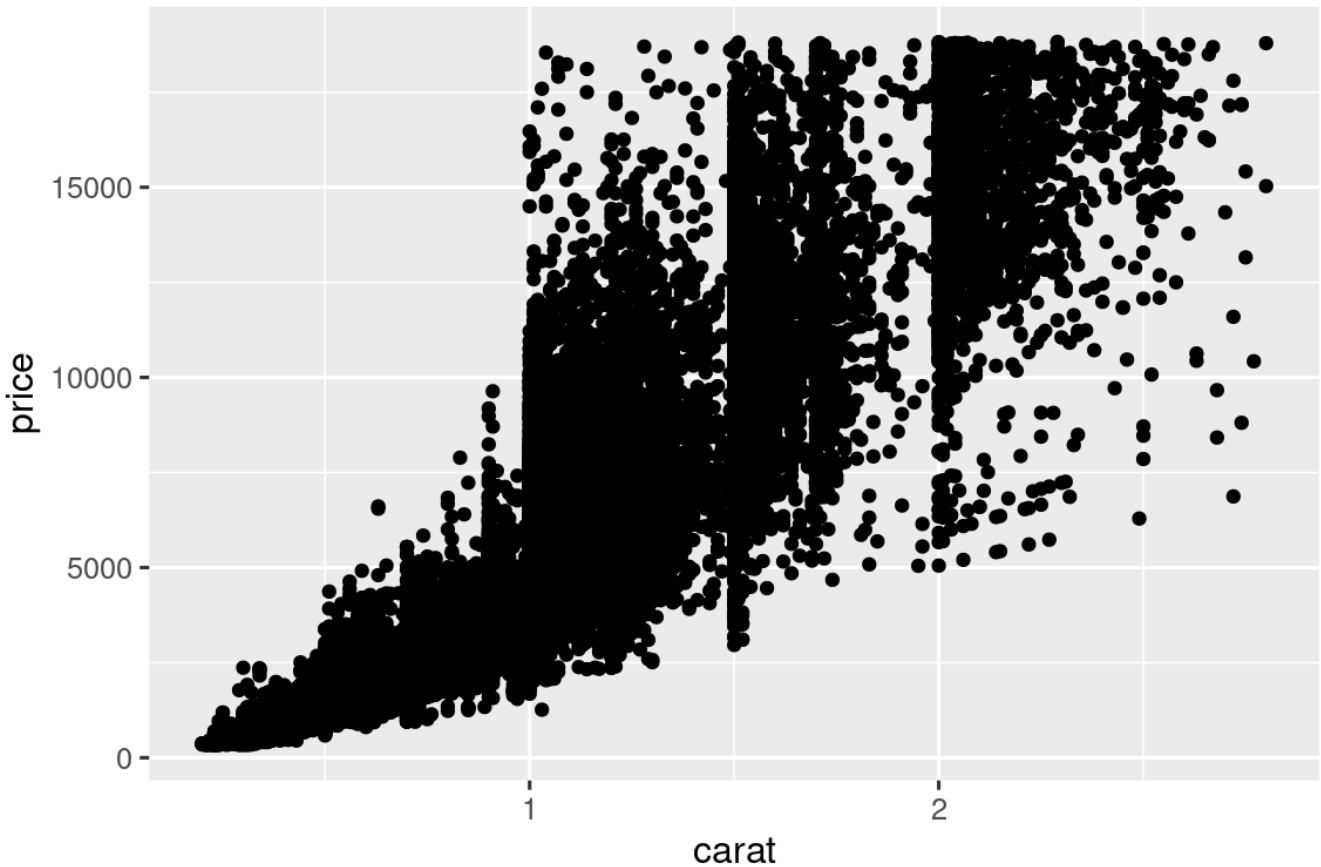
```
diamonds |>
  count(color, cut)
#> # A tibble: 35 × 3
#>   color cut            n
#>   <ord> <ord>      <int>
#> 1 D     Fair         163
#> 2 D     Good         662
#> 3 D     Very Good   1513
#> 4 D     Premium     1603
#> 5 D     Ideal       2834
#> 6 E     Fair         224
#> # i 29 more rows
```




```
diamonds |>
  count(color, cut) |>
  ggplot(aes(x = color, y = cut)) +
  geom_tile(aes(fill = n))
```

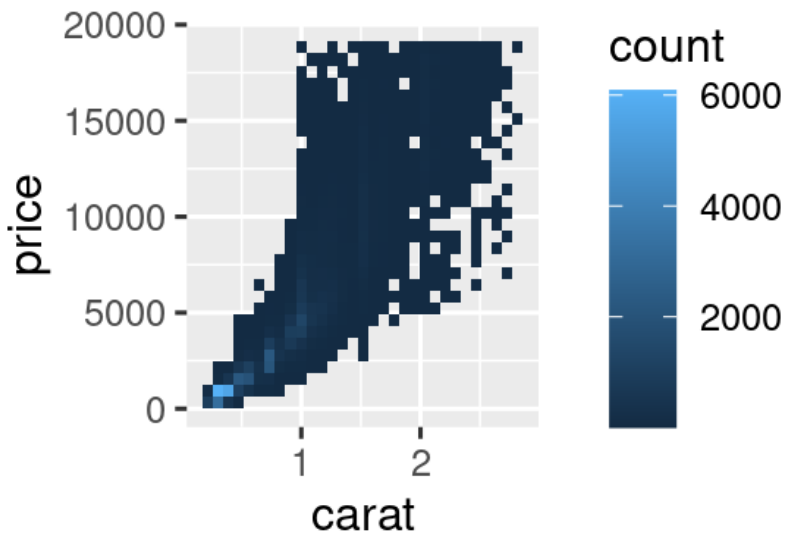
Two numerical variables

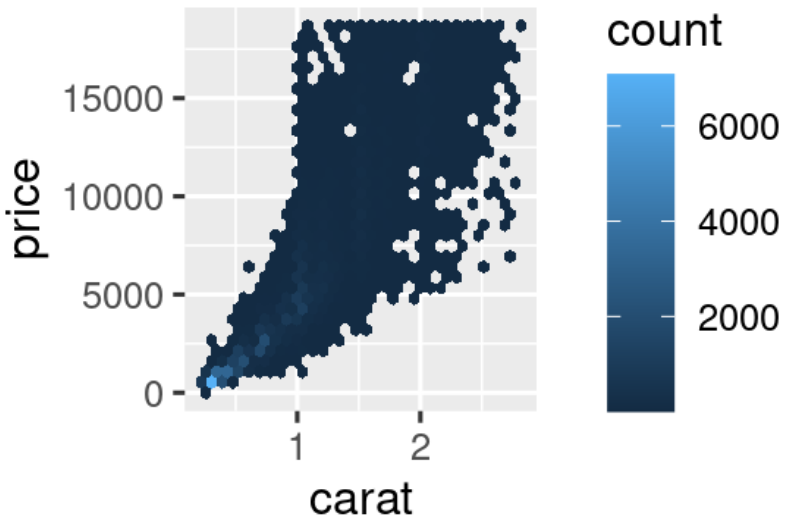
```
ggplot(smaller, aes(x = carat, y = price)) +
  geom_point()
```



geom_bin2d and geom_hex

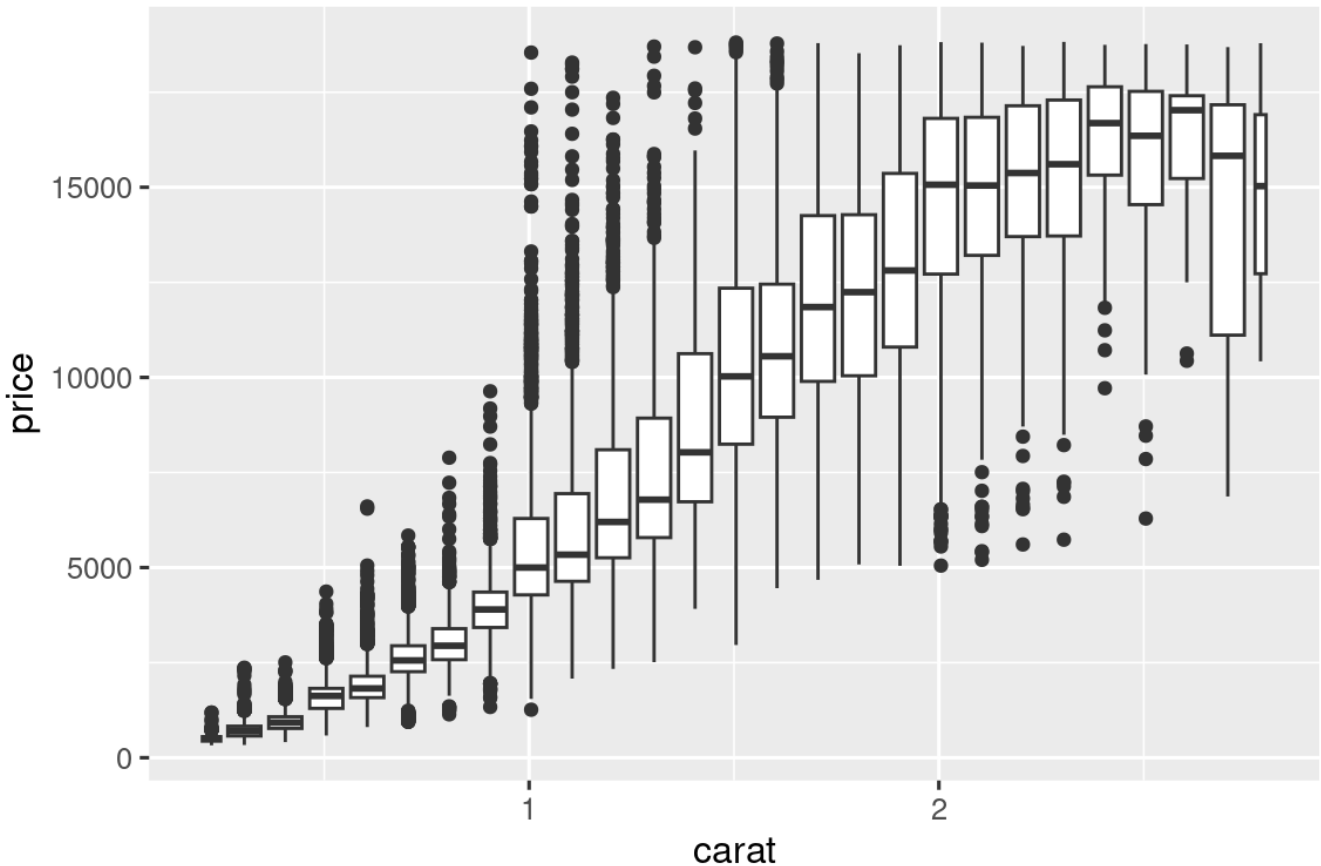
```
ggplot(smaller, aes(x = carat, y = price)) +
  geom_bin2d()

# install.packages("hexbin")
ggplot(smaller, aes(x = carat, y = price)) +
  geom_hex()
```





Bin one continuous variable

```
ggplot(smaller, aes(x = carat, y = price)) +
  geom_boxplot(aes(group = cut_width(carat, 0.1)))
```



Patterns in data

- Could this pattern be due to coincidence (i.e. random chance)?
- How can you describe the relationship implied by the pattern?
- How strong is the relationship implied by the pattern?
- What other variables might affect the relationship?
- Does the relationship change if you look at individual subgroups of the data?