Exploratory Data Analysis

Matthew McDonald

Exporatory Data Analysis (EDA)

EDA is an iterative cycle.

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

Questions

Your goal during EDA is to develop an understanding of your data.

The easiest way to do this is to use questions as tools to guide your investigation.

EDA is fundamentally a creative process.

And like most creative processes, the key to asking *quality* questions is to generate a large *quantity* of questions.

Useful Questions

1. What type of variation occurs within my variables?
2. What type of covariation occurs between my variables?

Tidy Data and Definitions

- A **variable** is a quantity, quality, or property that you can measure.
- A **value** is the state of a variable when you measure it. The value of a variable may change from measurement to measurement.
- An **observation** (data point) is a set of measurements made under similar conditions. An observation will contain several values, each associated with a different variable.
- **Tabular data** is a set of values, each associated with a variable and an observation.

  Tabular data is *tidy* if each value is placed in its own "cell", each variable in its own column, and each observation in its own row.

Variation

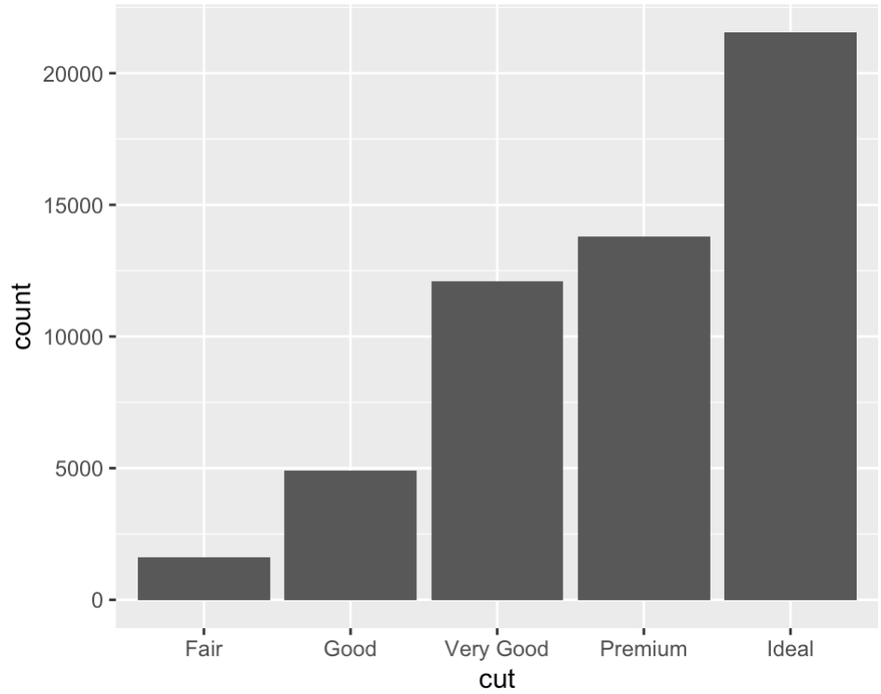
Variation is the tendency of the values of a variable to change from measurement to measurement. You can see variation easily in real life; if you measure any continuous variable twice, you will get two different results.

Every variable has its own pattern of variation, which can reveal interesting information. The best way to understand that pattern is to visualise the distribution of the variable’s values.

Categorical Variables

A variable is **categorical** if it can only take one of a small set of values. In R, categorical variables are usually saved as factors or character vectors. To examine the distribution of a categorical variable, use a bar chart.

```
1 library(tidyverse)
2
3 ggplot(data = diamonds) +
4   geom_bar(mapping = aes(x = cut))
```



Calculating Counts

```
1 diamonds %>%
2   count(cut)
```

```
# A tibble: 5 × 2
  cut           n
  <ord>     <int>
1 Fair       1610
2 Good       4906
3 Very Good 12082
4 Premium   13791
5 Ideal     21551
```

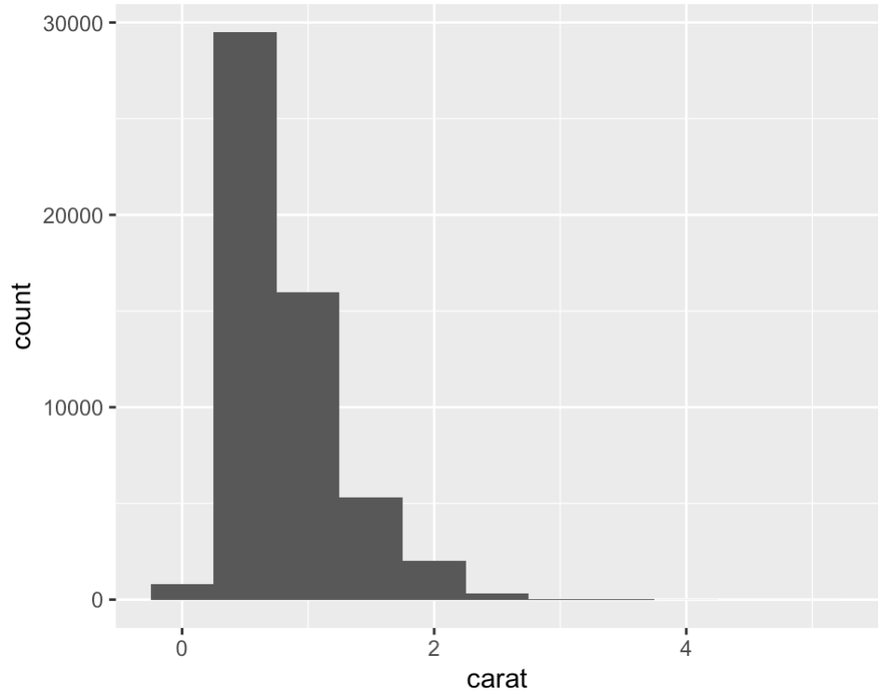
```
1 diamonds %>%
2   group_by(cut) %>%
3   summarise(n=n())
```

```
# A tibble: 5 × 2
  cut           n
  <ord>     <int>
1 Fair       1610
2 Good       4906
3 Very Good 12082
4 Premium   13791
5 Ideal     21551
```

Continuous Variables

A variable is **continuous** if it can take any of an infinite set of ordered values. Numbers and date-times are two examples of continuous variables. To examine the distribution of a continuous variable, use a histogram

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



Using Count with Continuous Variables

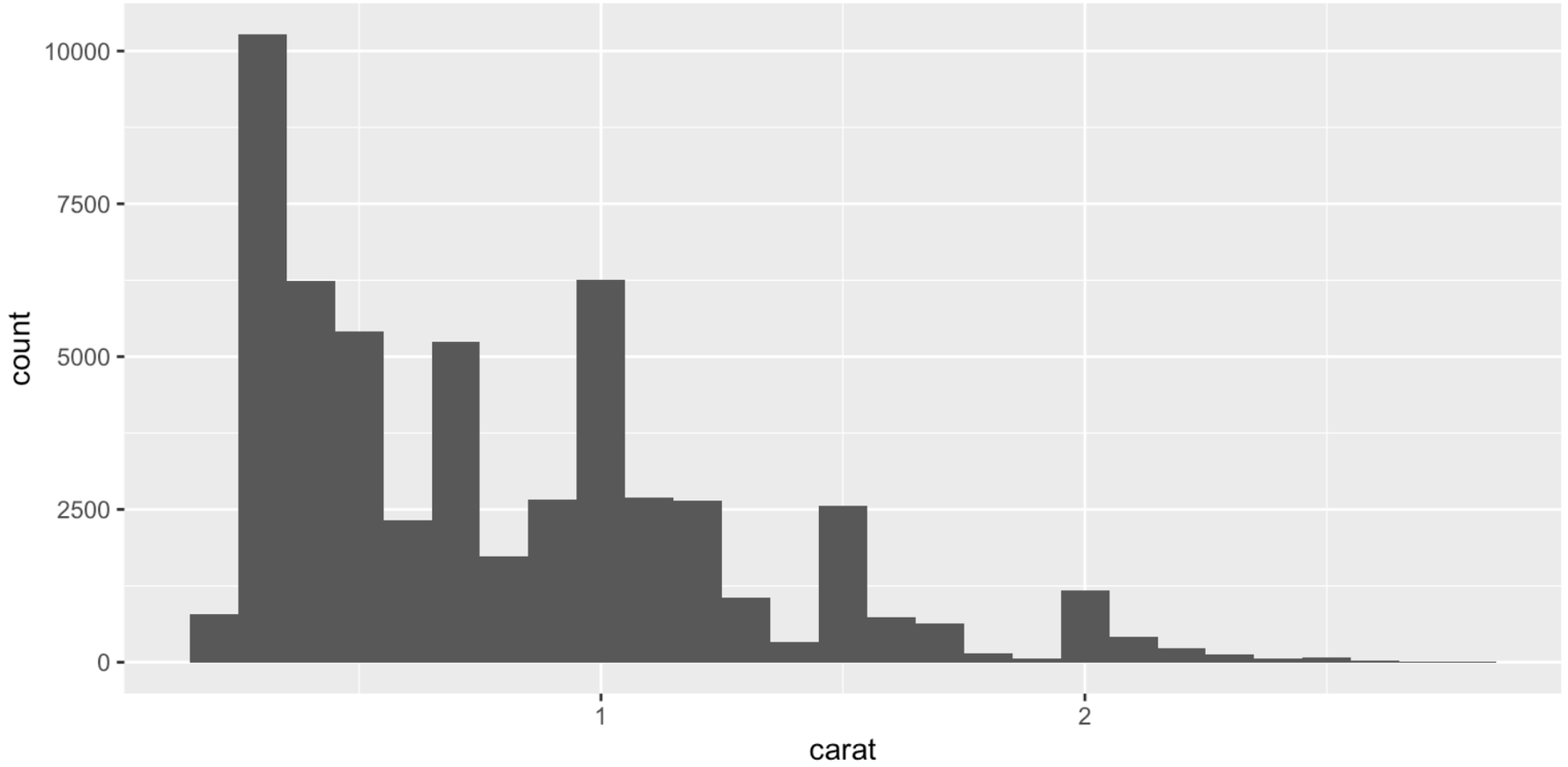
```
1 diamonds %>%
2   count(cut_width(carat, 0.5))
```

```
# A tibble: 11 × 2
   `cut_width(carat, 0.5)`     n
   <fct>                   <int>
 1 [-0.25,0.25]              785
 2 (0.25,0.75]             29498
 3 (0.75,1.25]             15977
 4 (1.25,1.75]              5313
 5 (1.75,2.25]              2002
 6 (2.25,2.75]               322
 7 (2.75,3.25]                32
 8 (3.25,3.75]                 5
 9 (3.75,4.25]                 4
10 (4.25,4.75]                 1
11 (4.75,5.25]                 1
```

Bin Width

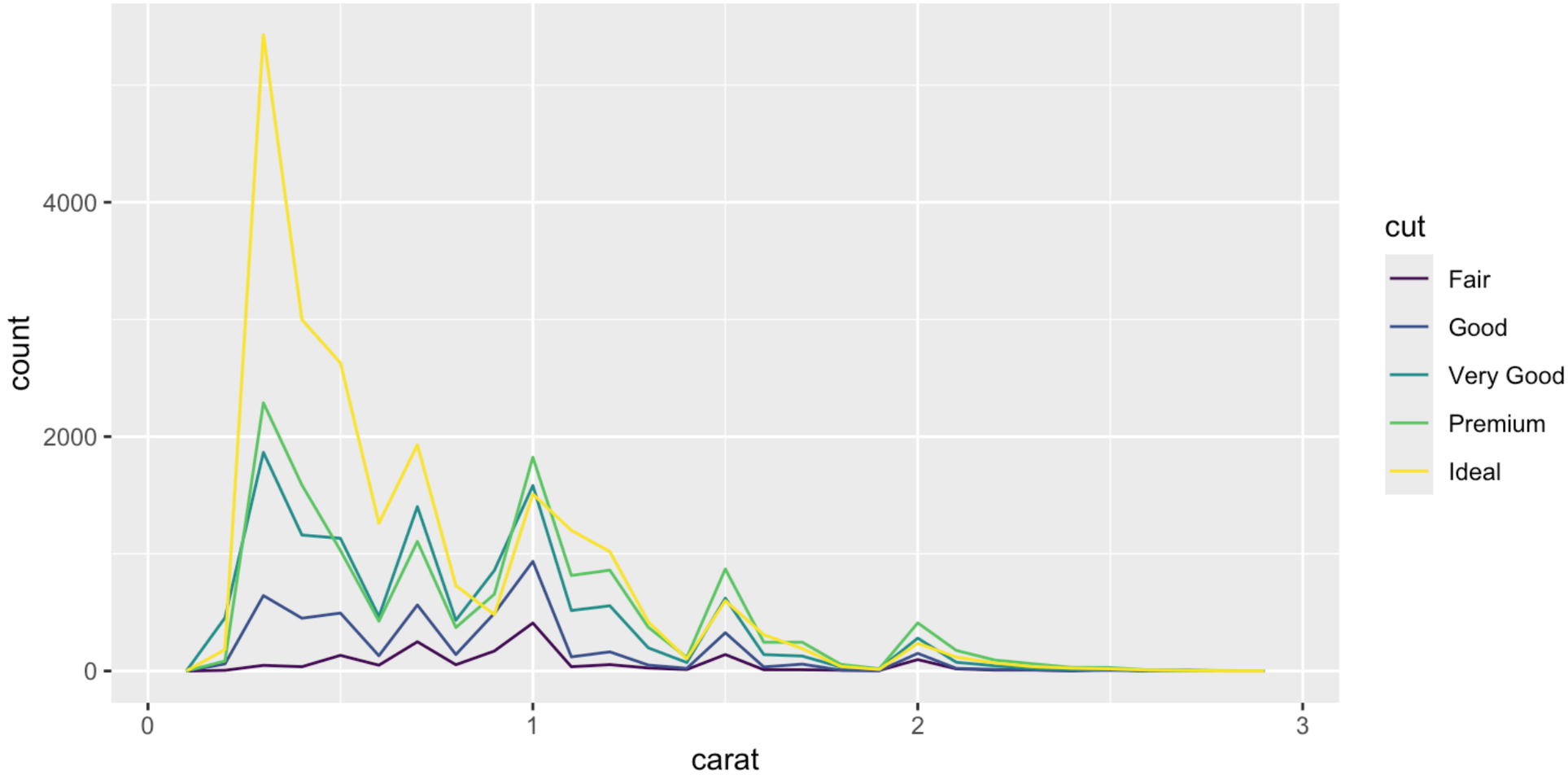
```r
smaller <- diamonds %>%
  filter(carat < 3)

ggplot(data = filter(diamonds, carat < 3), mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.1)
```



Plotting Multiple Histograms

```
1 ggplot(data = smaller, mapping = aes(x = carat, colour = cut)) +
2   geom_freqpoly(binwidth = 0.1)
```



Questions

Now that you can visualise variation, what should you look for in your plots?

And what type of follow-up questions should you ask?

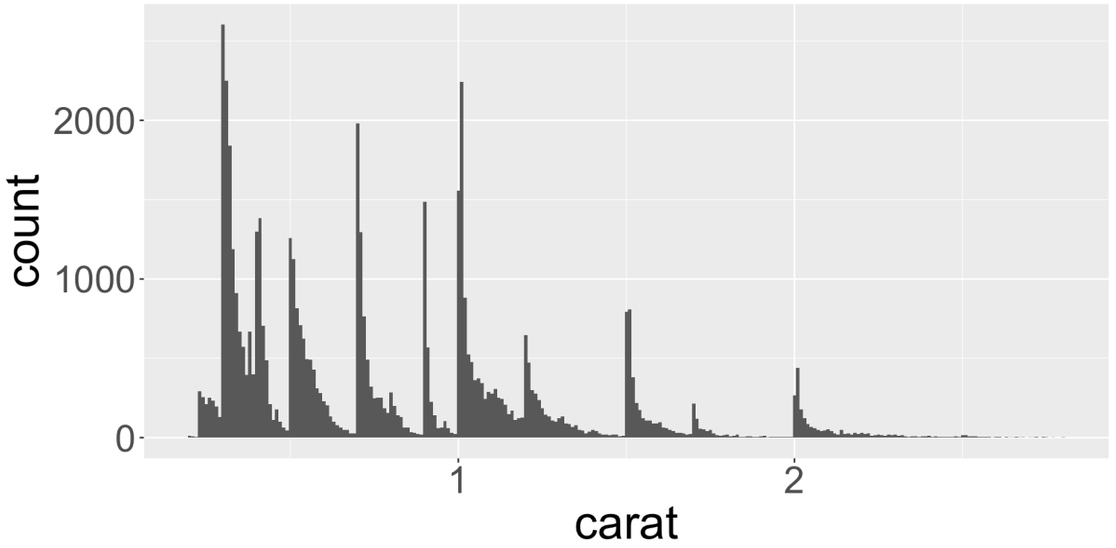
The key to asking good follow-up questions will be to rely on your curiosity (What do you want to learn more about?) as well as your skepticism (How could this be misleading?).

Typical Values

- 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?

Example

- 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?



Clusters

Clusters of similar values suggest that subgroups exist in your data. To understand the subgroups, ask:

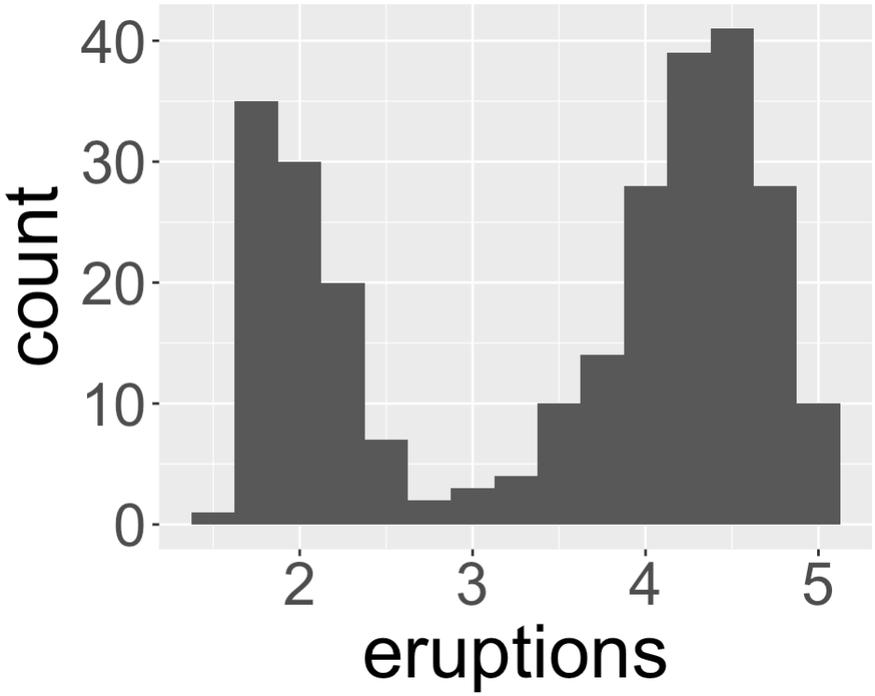
- How are the observations within each cluster similar to each other?
- How are the observations in separate clusters different from each other?

Example

The histogram shows the length (in minutes) of 272 eruptions of the Old Faithful Geyser in Yellowstone National Park.

Eruption times appear to be clustered into two groups: there are short eruptions (of around 2 minutes) and long eruptions (4-5 minutes), but little in between.

- How can you explain or describe the clusters?
- Why might the appearance of clusters be misleading?



Unusual Values

Outliers are observations that are unusual; data points that don’t seem to fit the pattern.

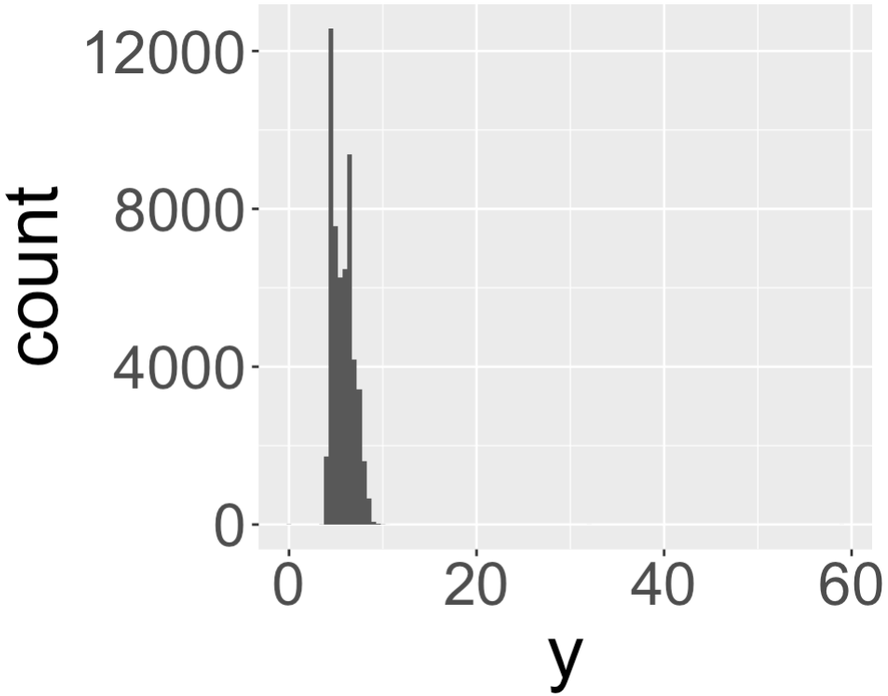
Sometimes outliers are data entry errors; other times outliers suggest important new science.

When you have a lot of data, outliers are sometimes difficult to see in a histogram.

Example

For example, take the distribution of the y variable from the diamonds dataset.

The only evidence of outliers is the unusually wide limits on the x-axis.



Unusual Observations

The y variable measures one of the three dimensions of these diamonds, in mm.

We know that diamonds can't have a width of 0mm, so these values must be incorrect.

We might also suspect that measurements of 32mm and 59mm are implausible: those diamonds are over an inch long, but don't cost hundreds of thousands of dollars!

```r
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
```

Covariation

If variation describes the behavior *within* a variable, covariation describes the behavior *between* variables.

Covariation is the tendency for the values of two or more variables to vary together in a related way.

The best way to spot covariation is to visualise the relationship between two or more variables.

How you do that should again depend on the type of variables involved.

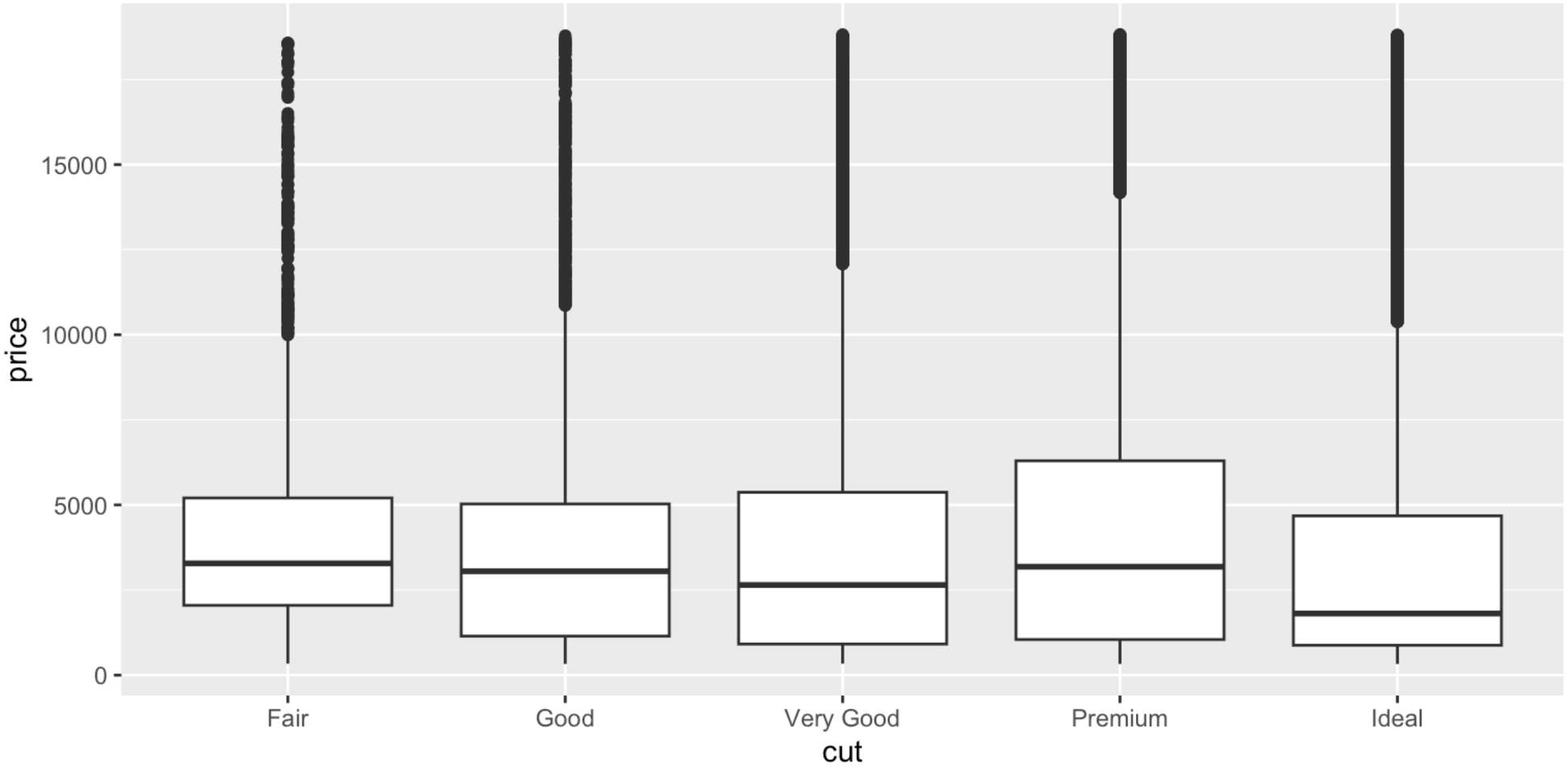
Boxplots

A **boxplot** is a type of visual shorthand for a distribution of values that is popular among statisticians. Each boxplot consists of:

- A box that stretches from the 25th percentile of the distribution to the 75th percentile, a distance known as the interquartile range (IQR). In the middle of the box is a line that displays the median, i.e. 50th percentile, of the distribution. These three lines give you a sense of the spread of the distribution and whether or not the distribution is symmetric about the median or skewed to one side.
- Visual points that display observations that fall more than 1.5 times the IQR from either edge of the box.
- A line (or whisker) that extends from each end of the box and goes to the farthest non-outlier point in the distribution.

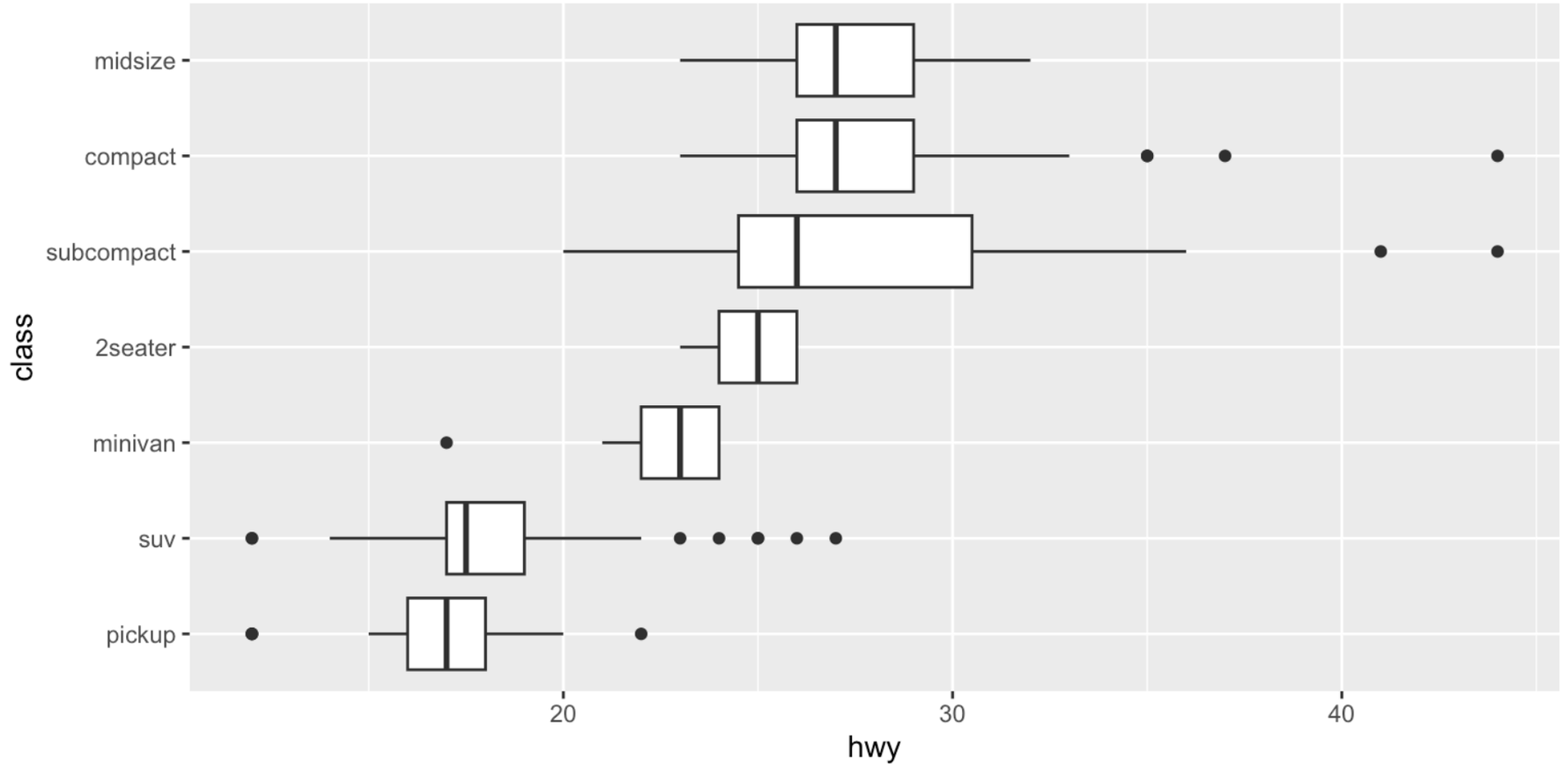
Example Boxplot

```
1 ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
2   geom_boxplot()
```



Cleaning Up A Boxplot

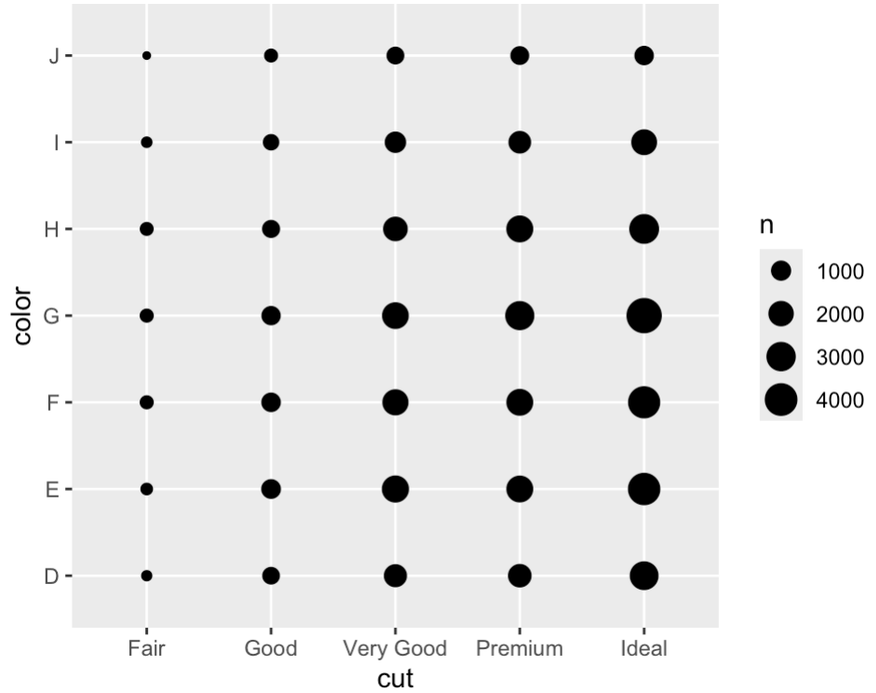
```
mpg %>%
  mutate(class = reorder(class, hwy, FUN = median)) %>%
  ggplot() +
    geom_boxplot(mapping = aes(x = class, y = hwy)) +
    coord_flip()
```



Two Categorical Variables

To visualise the covariation between categorical variables, you'll need to count the number of observations for each combination. One way to do that is to rely on the built-in `geom_count()`:

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



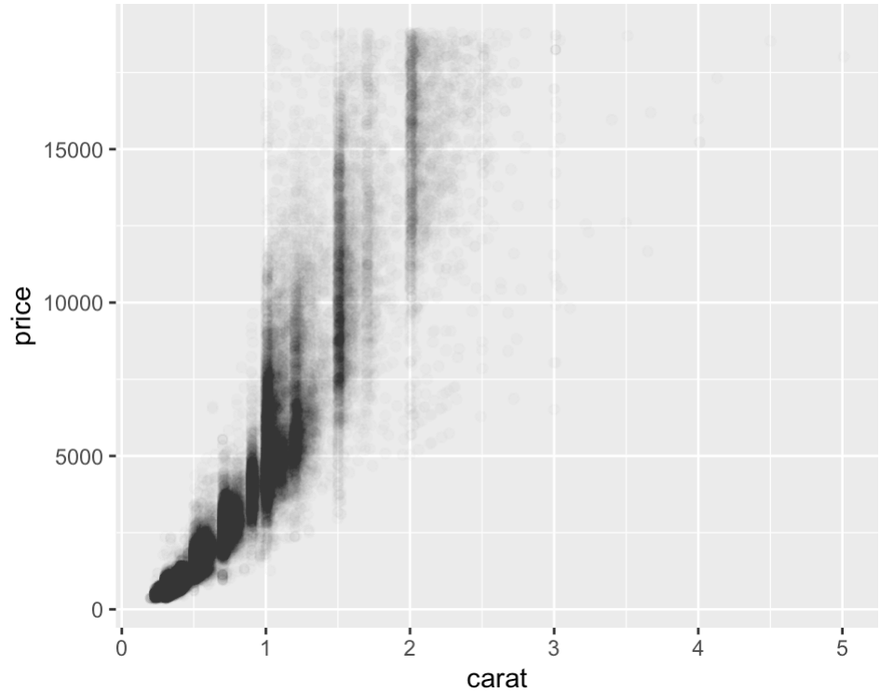
Two Continuous Variables

Overplotting

Scatterplots become less useful as the size of your dataset grows, because points begin to overplot, and pile up into areas of uniform black (as above).

Using the `alpha` aesthetic to add transparency can address the problem

```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price),
             alpha = 1 / 100)
```



Patterns and models

Patterns in your data provide clues about relationships.

If a systematic relationship exists between two variables it will appear as a pattern in the data.

If you spot a pattern, ask yourself:

- 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?

Patterns and Covariation

Patterns provide one of the most useful tools for data scientists because they reveal covariation.

If you think of variation as a phenomenon that creates uncertainty, covariation is a phenomenon that reduces it.

If two variables covary, you can use the values of one variable to make better predictions about the values of the second.

If the covariation is due to a causal relationship (a special case), then you can use the value of one variable to control the value of the second.

Models

Models are a tool for extracting patterns out of data.

For example, consider the diamonds data.

It's hard to understand the relationship between cut and price, because cut and carat, and carat and price are tightly related.

It's possible to use a model to remove the very strong relationship between price and carat so we can explore the subtleties that remain.

Fitting a Model

This code fits a model that predicts `price` from `carat` and then computes the residuals.

The residuals give us a view of the price of the diamond, once the effect of carat has been removed.

Once you've removed the strong relationship between carat and price, you can see what you expect in the relationship between cut and price: relative to their size, better quality diamonds are more expensive.

```
1 library(modelr)
2
3 mod <- lm(log(price) ~ log(carat), data = diamonds)
4
5 diamonds2 <- diamonds %>%
6   add_residuals(mod) %>%
7   mutate(resid = exp(resid))
```

The Updated Diamonds Plot

```
1 ggplot(data = diamonds2) +
2   geom_boxplot(mapping = aes(x = cut, y = resid))
```