Exploratory data analysis

Sleeping Beauty, WA

Kristin Bott

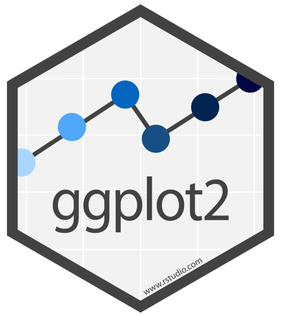
Garrett Grolemund

RStudio Academy

rstudio::conf 2022

https://commons.wikimedia.org/wiki/File:Diamonds.jpg

Diamonds dataset




```
diamonds
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

`diamonds` data

- **~54,000** round diamonds
- **4 C's :** carat, color, clarity, cut
- **Measurements:** Depth, table, length, width, height
- **Price**

https://commons.wikimedia.org/wiki/File:Diamonds.jpg

Measurements:

Depth,

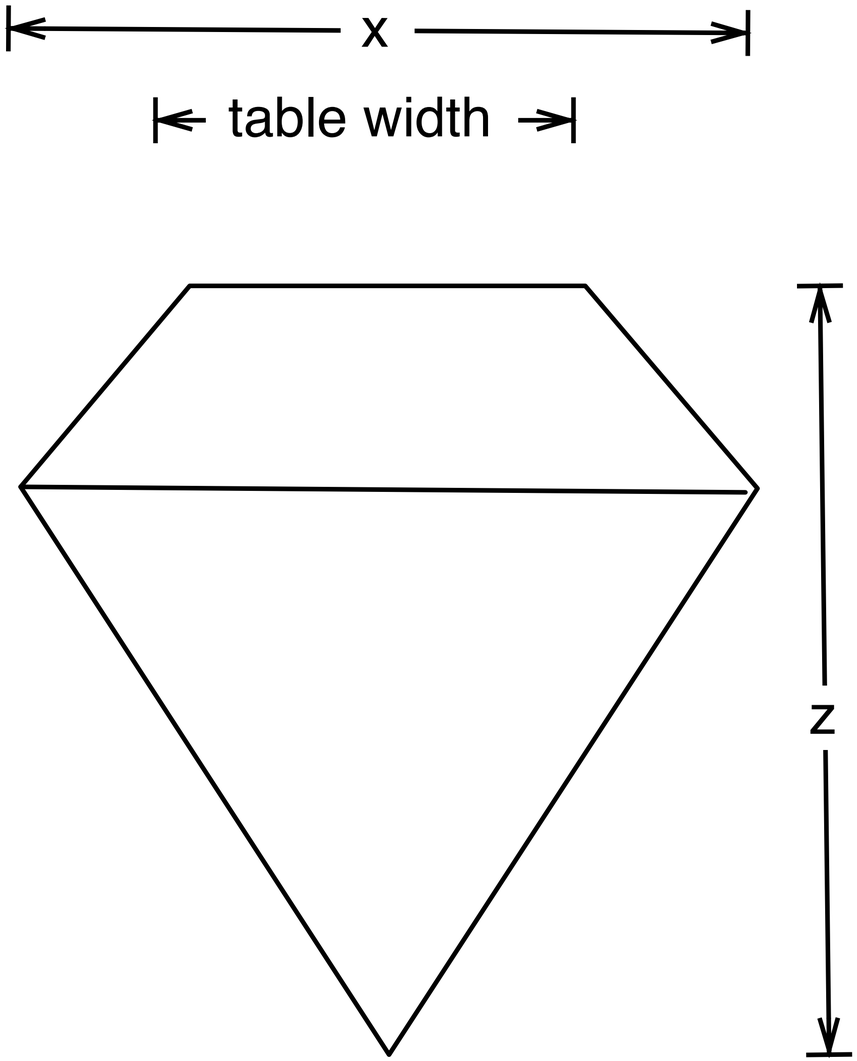
table,

length (x),

width (y),

height (z)

depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
61.5	55.0	326	3.95	3.98	2.43
59.8	61.0	326	3.89	3.84	2.31
56.9	65.0	327	4.05	4.07	2.31
62.4	58.0	334	4.20	4.23	2.63
63.3	58.0	335	4.34	4.35	2.75
62.8	57.0	336	3.94	3.96	2.48



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

4 C's : Carat, **color**, clarity, cut

COLOR GRADING SCALE

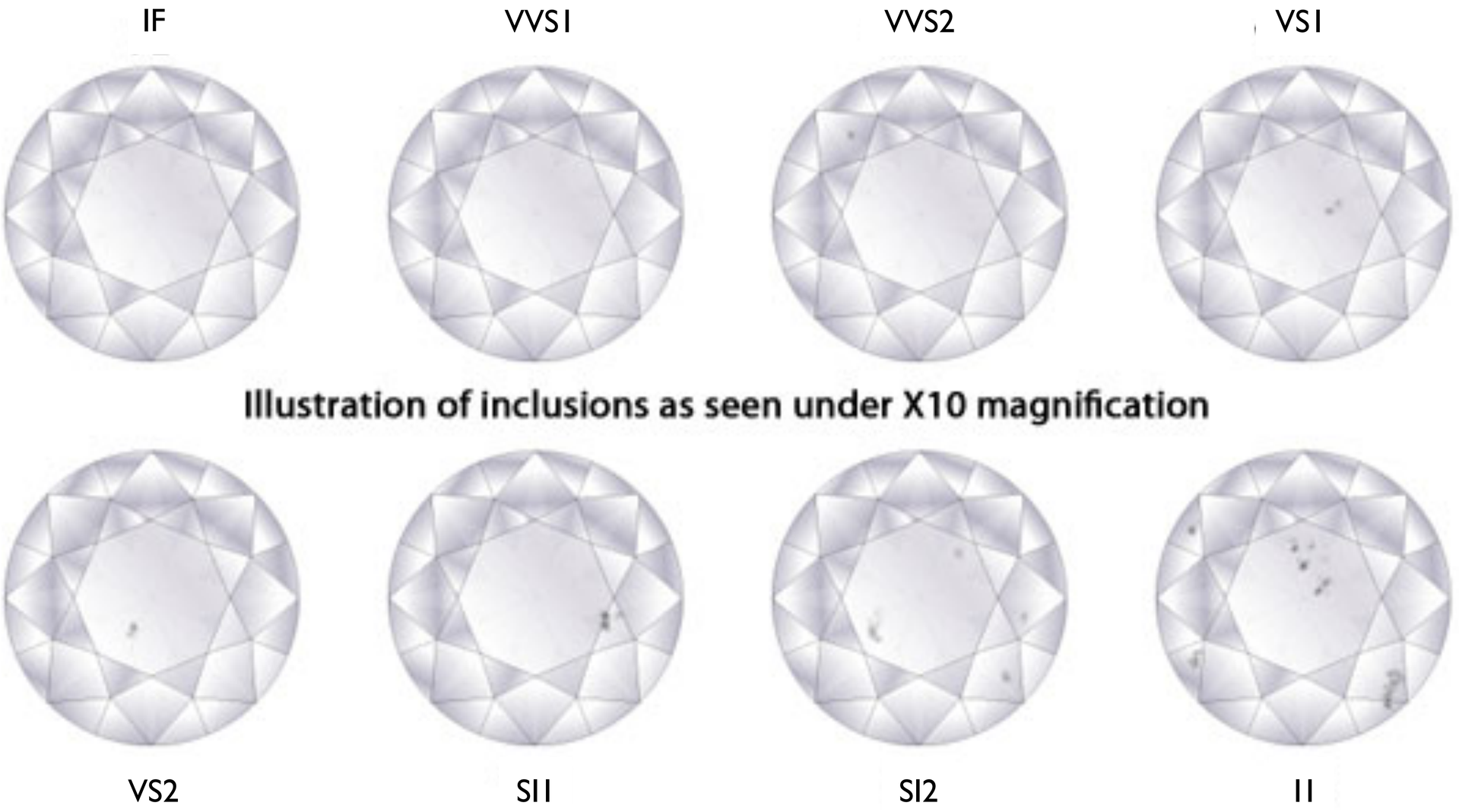




D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Colorless			Near Colorless				Faint Yellow			Very Light Yellow					Light Yellow							

http://www.mlcld.com/colour.aspx

4 C's : Carat, color, **clarity**, cut



http://www.thediamondsexperts.com/diamonds-guide

Your Turn 1

05:00

Explore the diamond `price` data. What do you learn?

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47

Exploratory data analysis

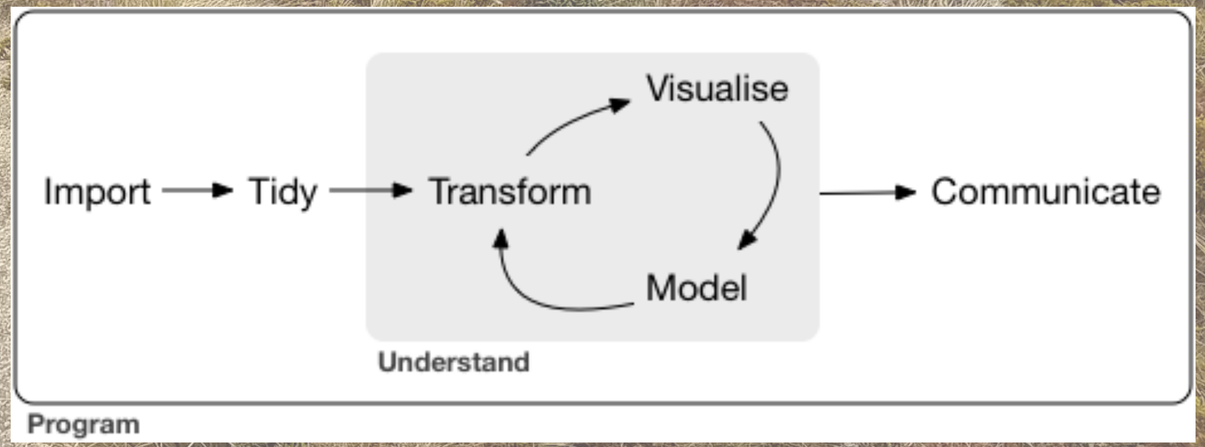
Photo by Jon Tyson on Unsplash

What is EDA?

Generate questions about data

Investigate ideas

Search for answers



Urchin Track, North Island, NZ

https://r4ds.had.co.nz/introduction.html

Our ~~Your~~ Turn

What questions do we have about `diamonds`?

Our ~~Your~~ Turn

What questions do we have about `diamonds`?

Can we *categorize* these questions?

Our ~~Your~~ Turn

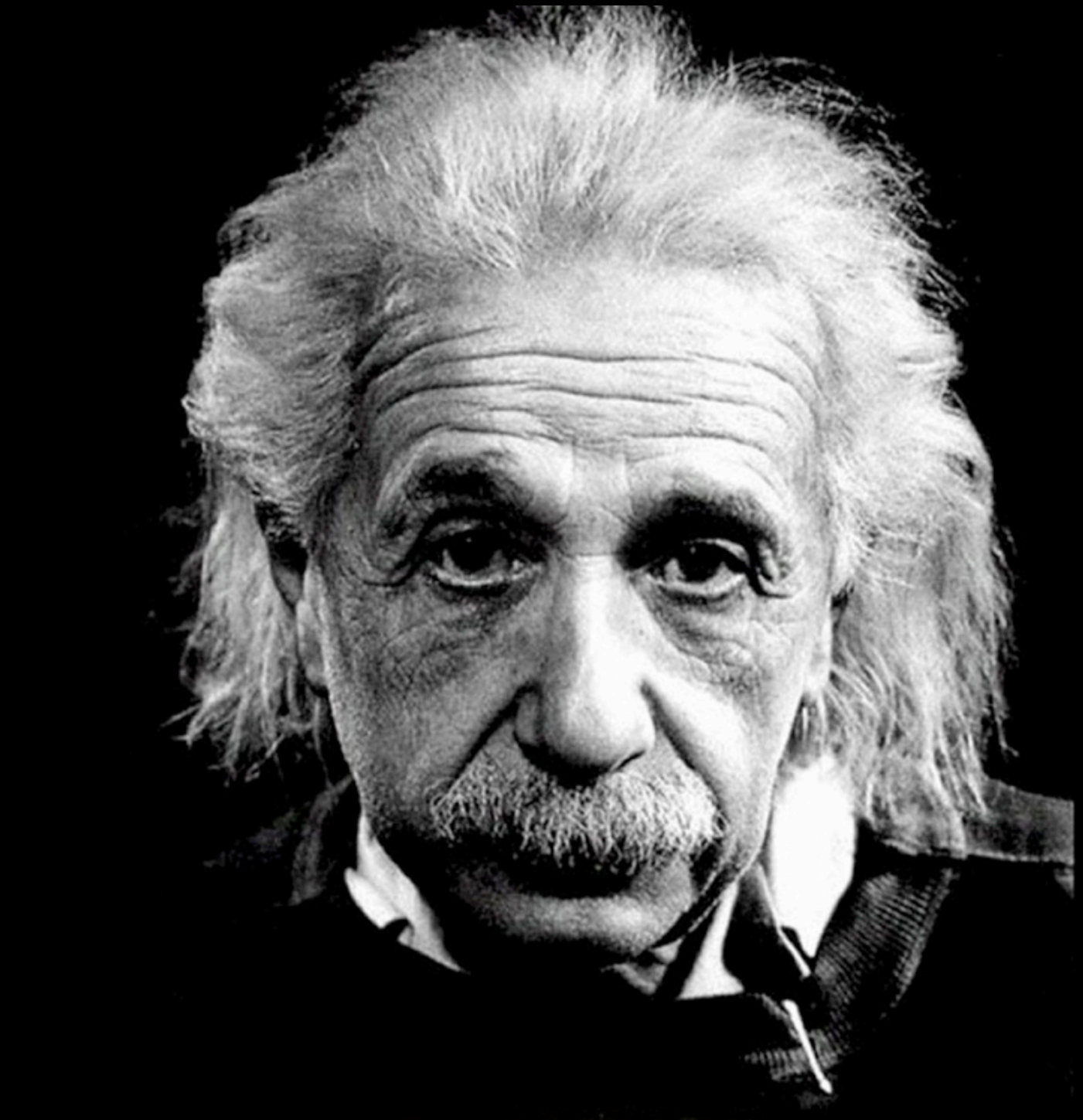
Can we *categorize* these questions about `diamonds` ?

- Questions about **distribution** [within a variable]
- Questions about **relationships** [between variables]
- Questions about **data quality**

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.

Variation is the tendency of the values of a variable to change from measurement to measurement.



What did Einstein think was a universal constant?

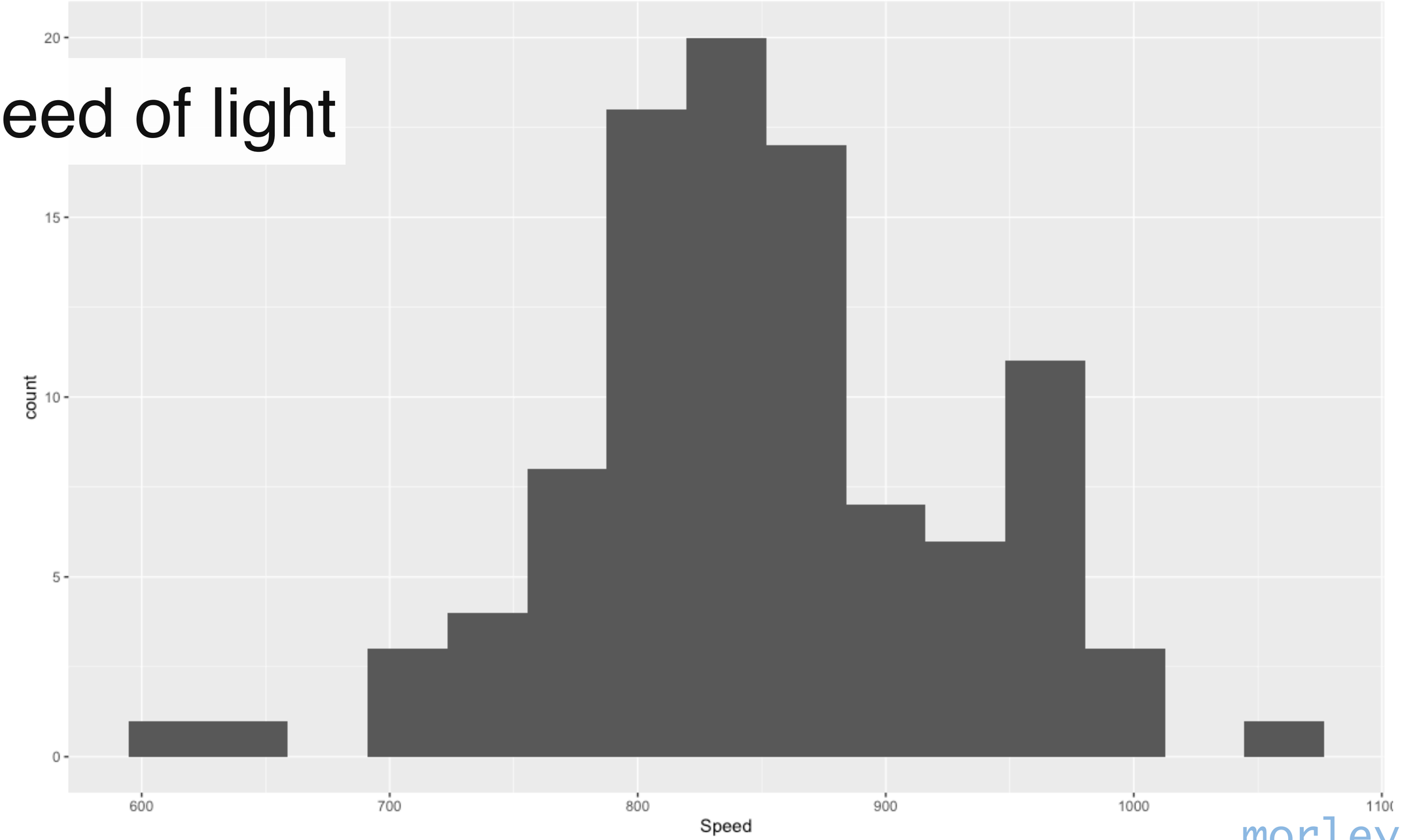
$E = mc^2$

Measurements of speed of light (1879), n = 100

```
 [1]  850  740  900 1070  930  850  950  980  980  880 1000  980
[13]  930  650  760  810 1000 1000  960  960  960  940  960  940
[25]  880  800  850  880  900  840  830  790  810  880  880  830
[37]  800  790  760  800  880  880  880  860  720  720  620  860
[49]  970  950  880  910  850  870  840  840  850  840  840  840
[61]  890  810  810  820  800  770  760  740  750  760  910  920
[73]  890  860  880  720  840  850  850  780  890  840  780  810
[85]  760  810  790  810  820  850  870  870  810  740  810  940
[97]  950  800  810  870
```

morley$Speed

speed of light



morley$Speed

Quiz

What is one of the best ways to **explore the variation** in your data?

"The simple graph has brought more information to the data analyst’s mind than any other device. "

- John Tukey

Exploring Variation

Photo by Ricardo Gomez Angel on Unsplash

Variation is the tendency of the values of a variable to change from measurement to measurement.

Distribution is the pattern of values that appear when you measure a variable many times.

Quiz

What is a **continuous** variable?

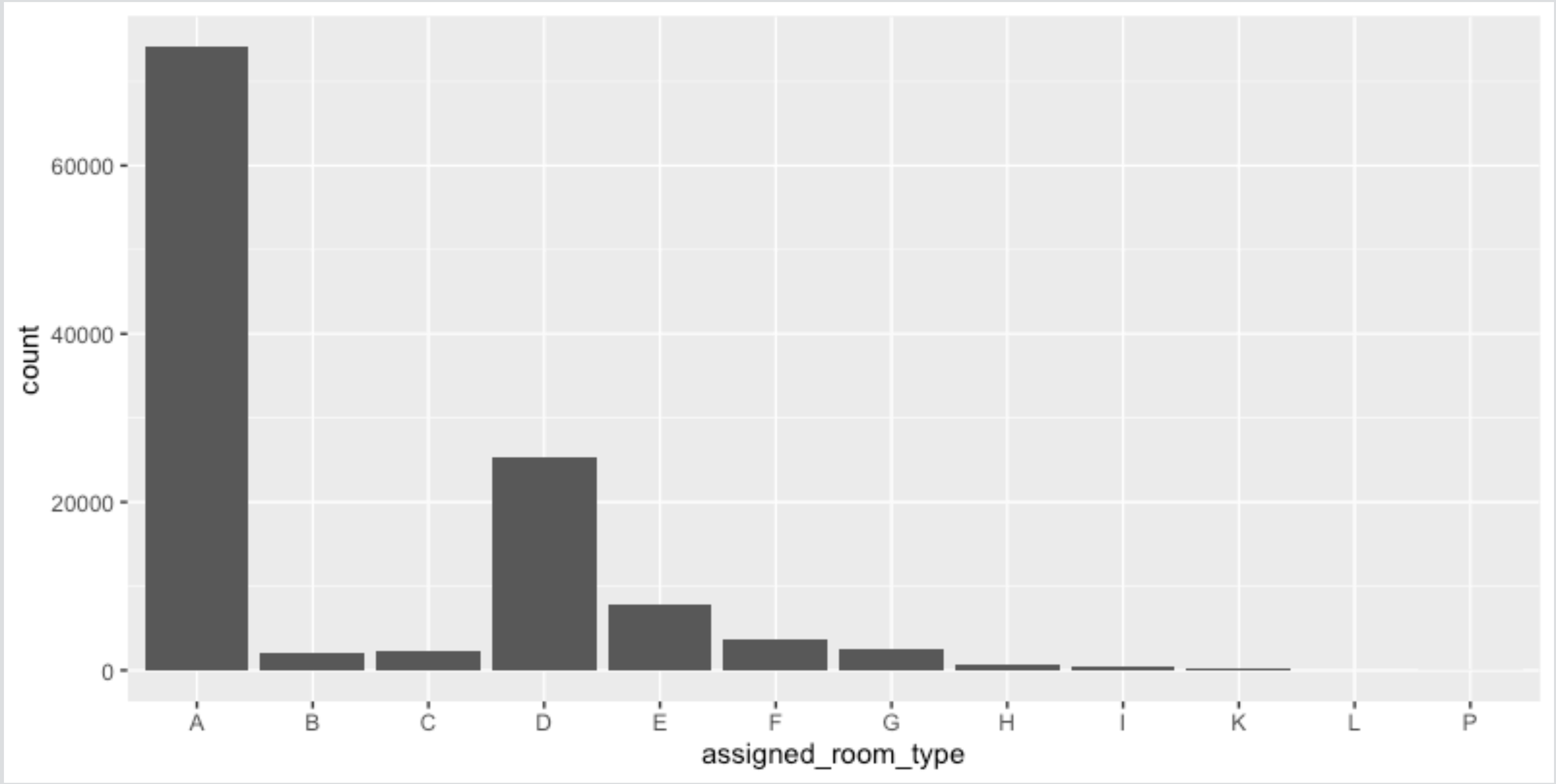
Quiz

What is a **categorical** variable?

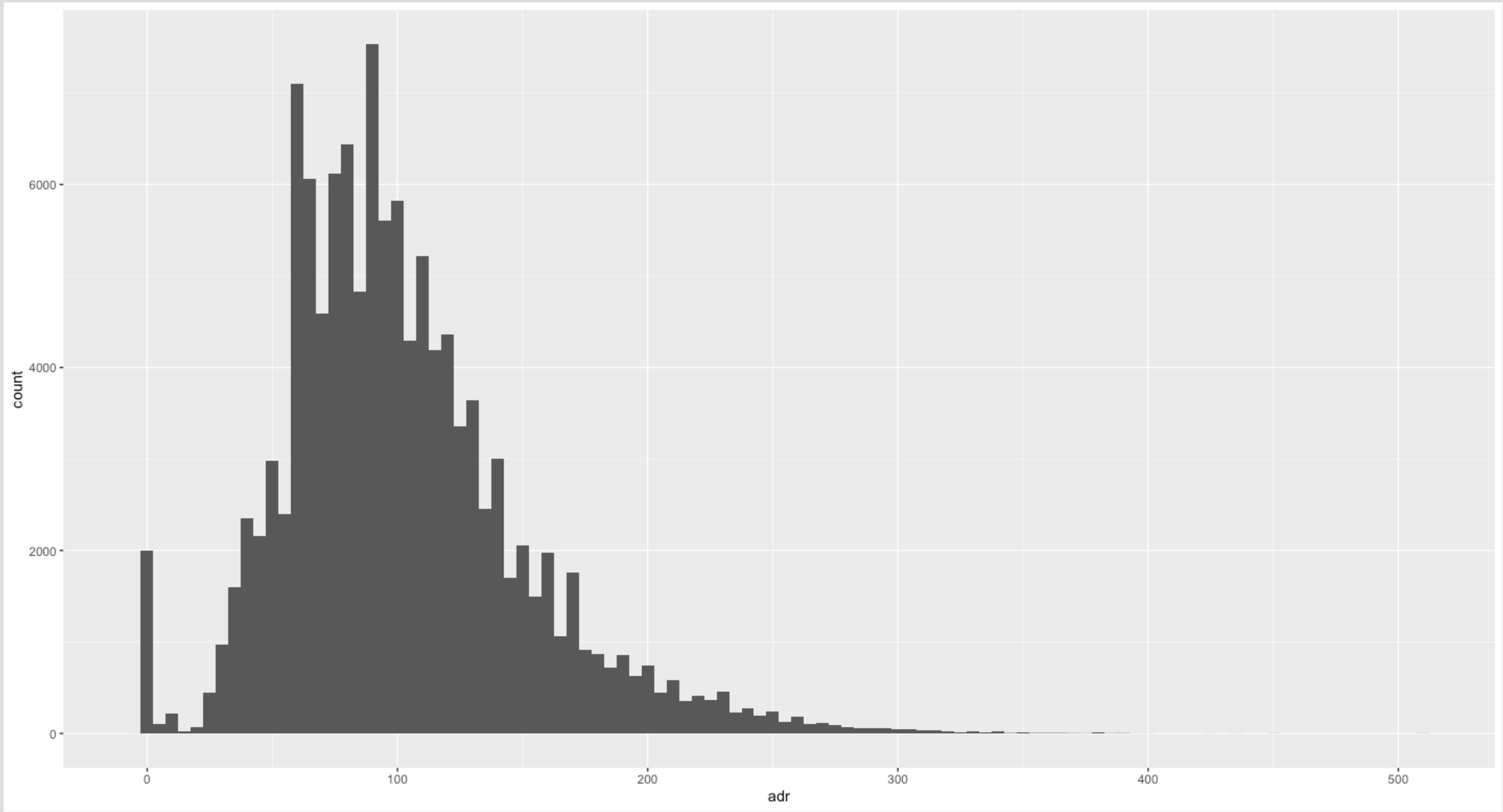
A **continuous** variable can take an infinite set of numeric values

A **categorical** variable represents groups (or categories)

Distribution of a **categorical** variable (bar plot)



Distribution of a **continuous** variable (histogram)



hotels$adr

Your Turn 2

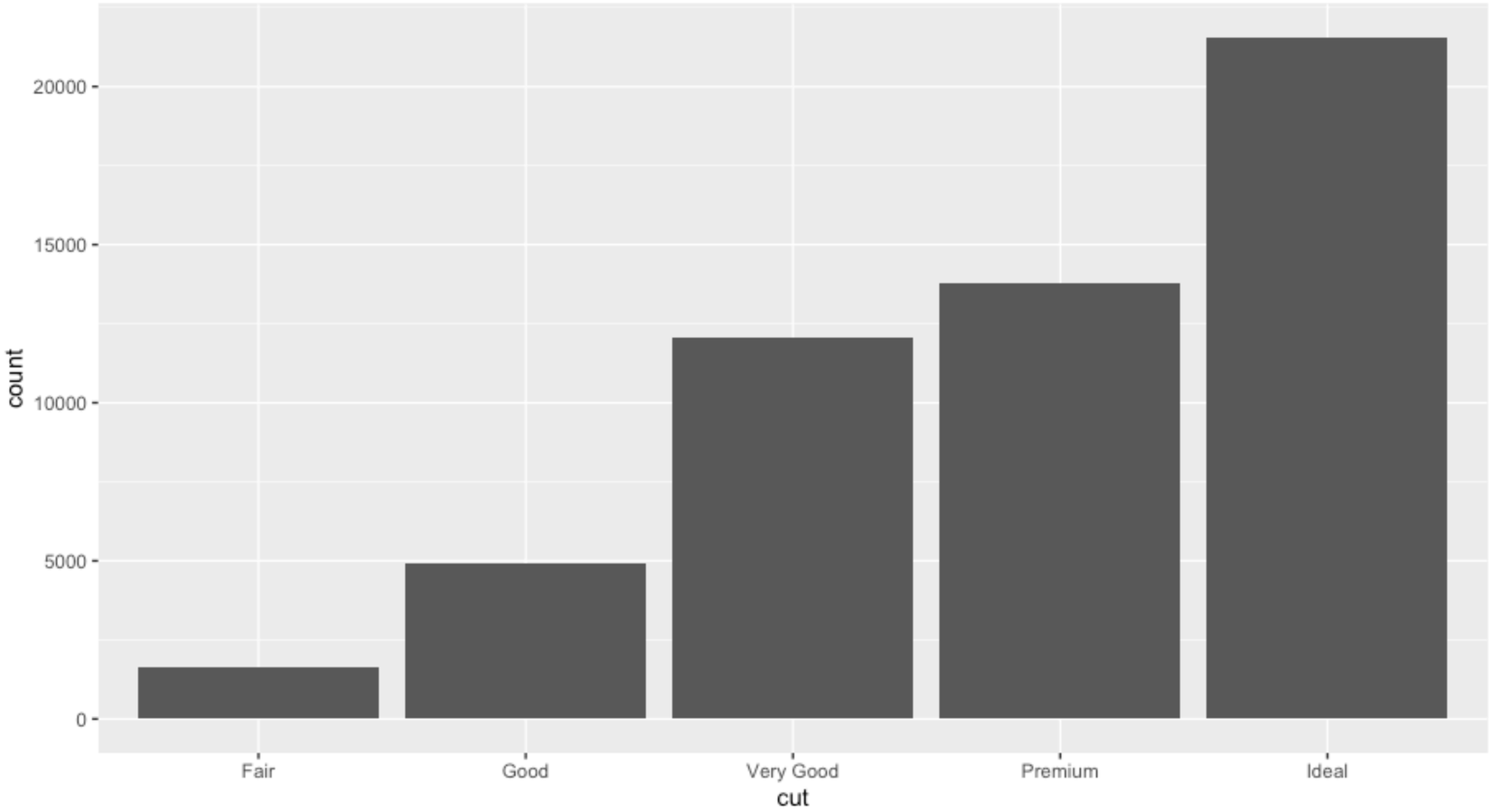
Visualize 1) the distribution for `cut`

2) the distribution for `carat`

For each, ask:

- What values are **typical**?
- What value are **rare**?
- What value are **possible**? (range)
- *What else do you notice?*

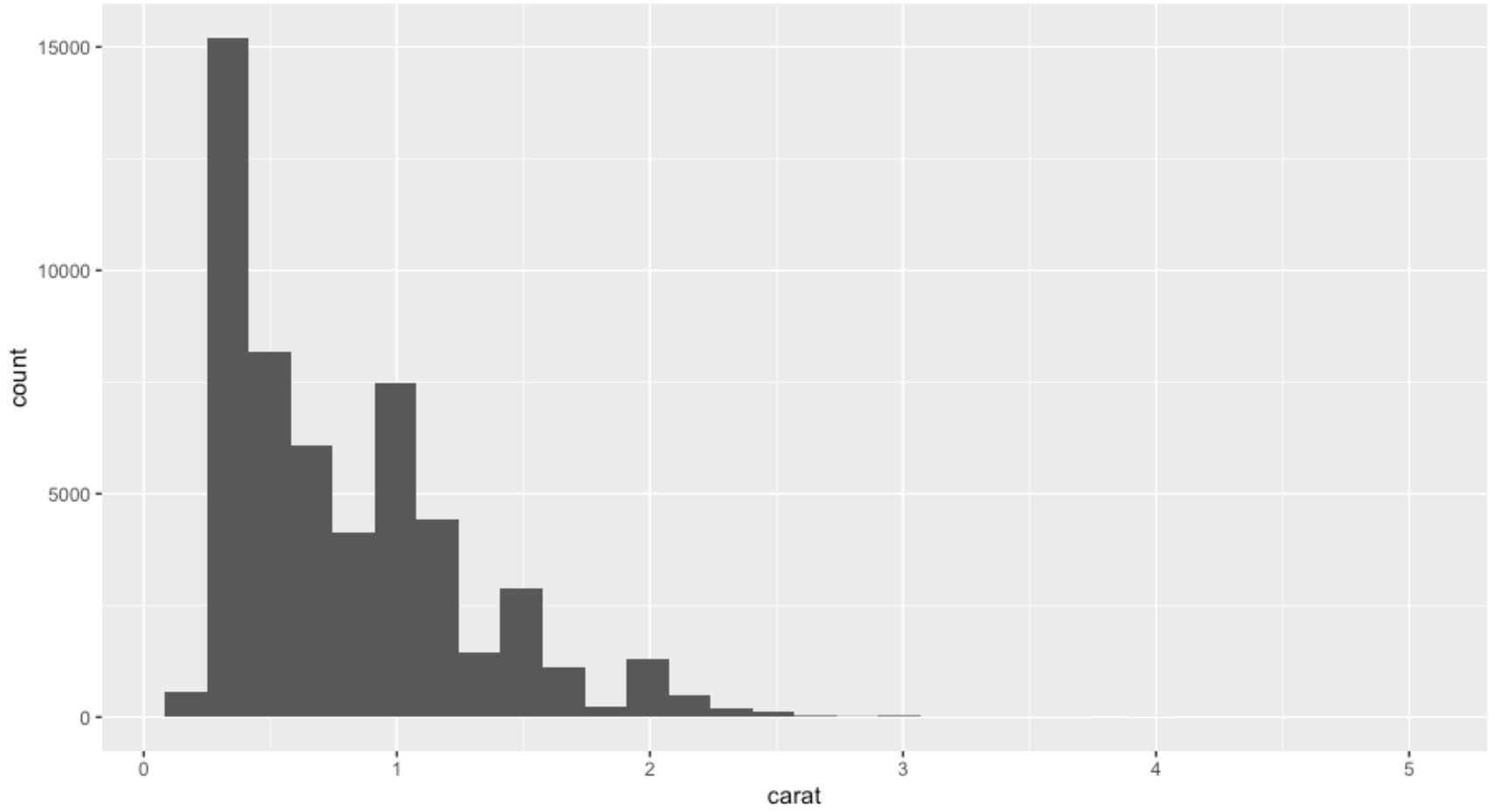
Visualize (1) the distribution for `cut`




```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
```

- **typical**?
- **rare**?
- **possible**?
- What else do you notice?

Visualize (2) the distribution for `carat`



- **typical**?
- **rare**?
- **possible**?
- What else do you notice?

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

Your Turn 3

Investigate one of your questions about `carat` by making a second plot.

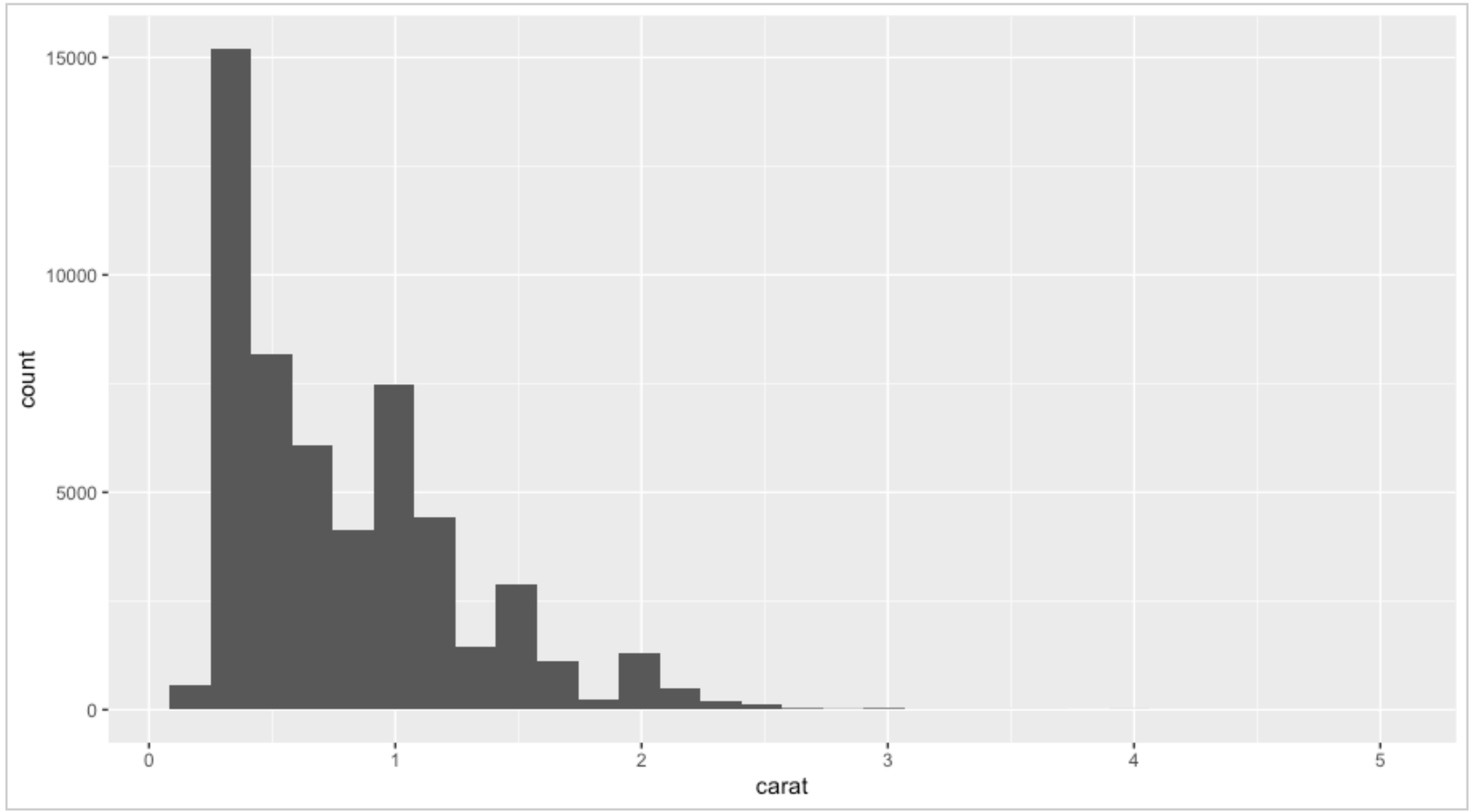
Share your results with your group.

(hint: when you make a histogram, explore multiple binwidths)

Explore 02:00

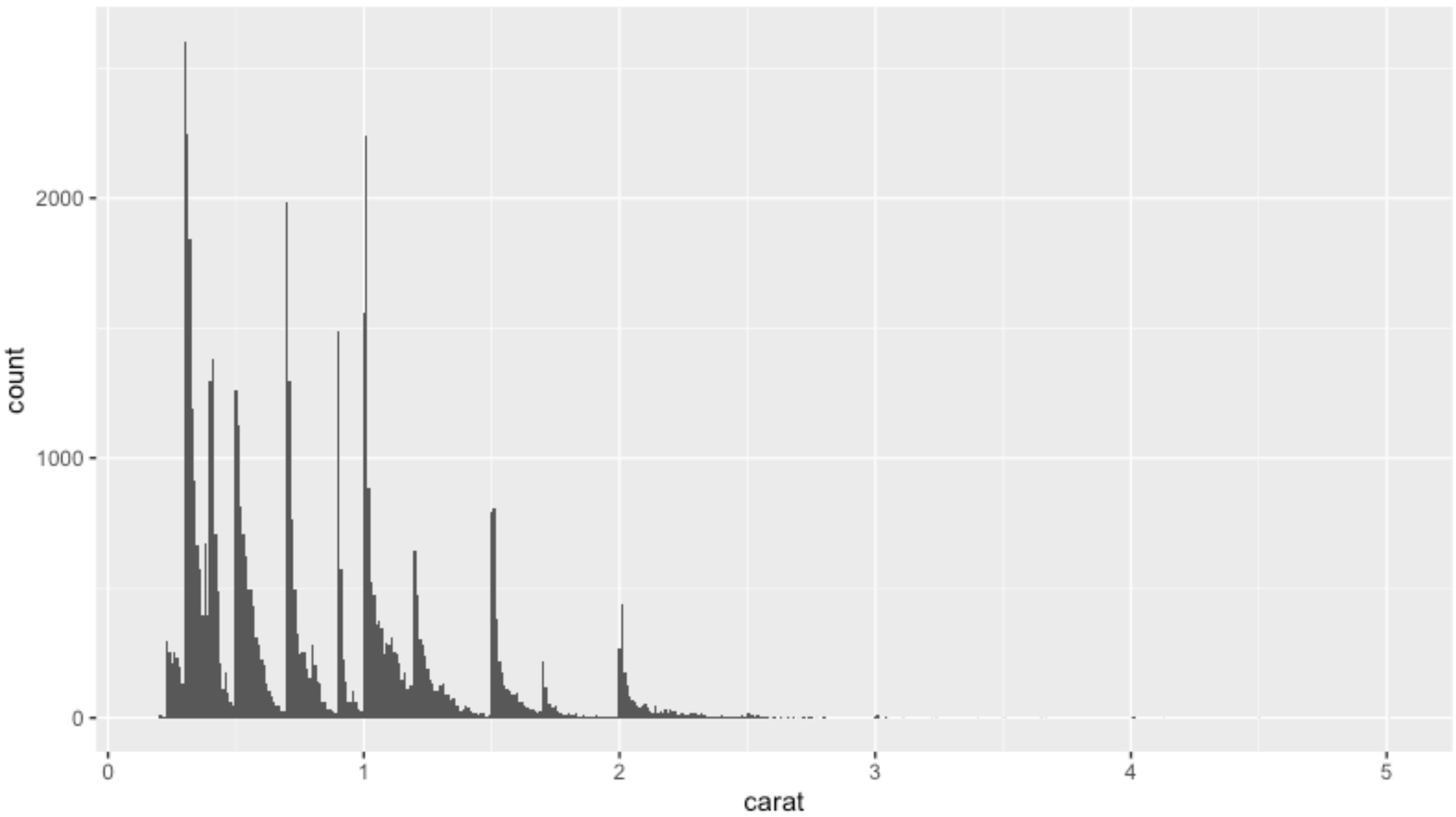
Share 02:00

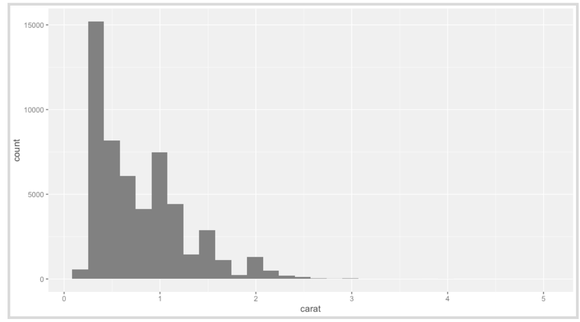
Revisit the distribution for `carat`

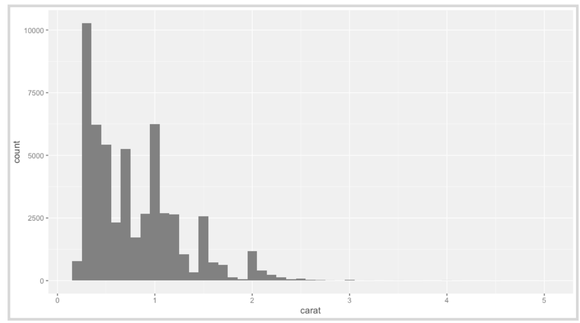


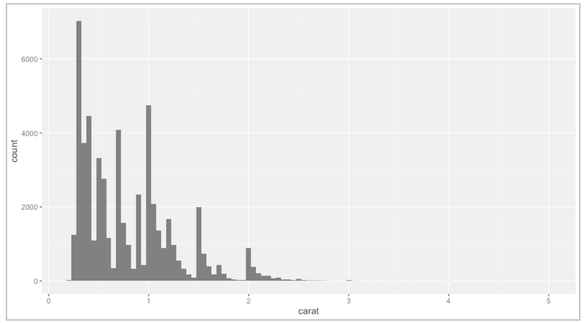

```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.1)
```

Revisit the distribution for `carat`





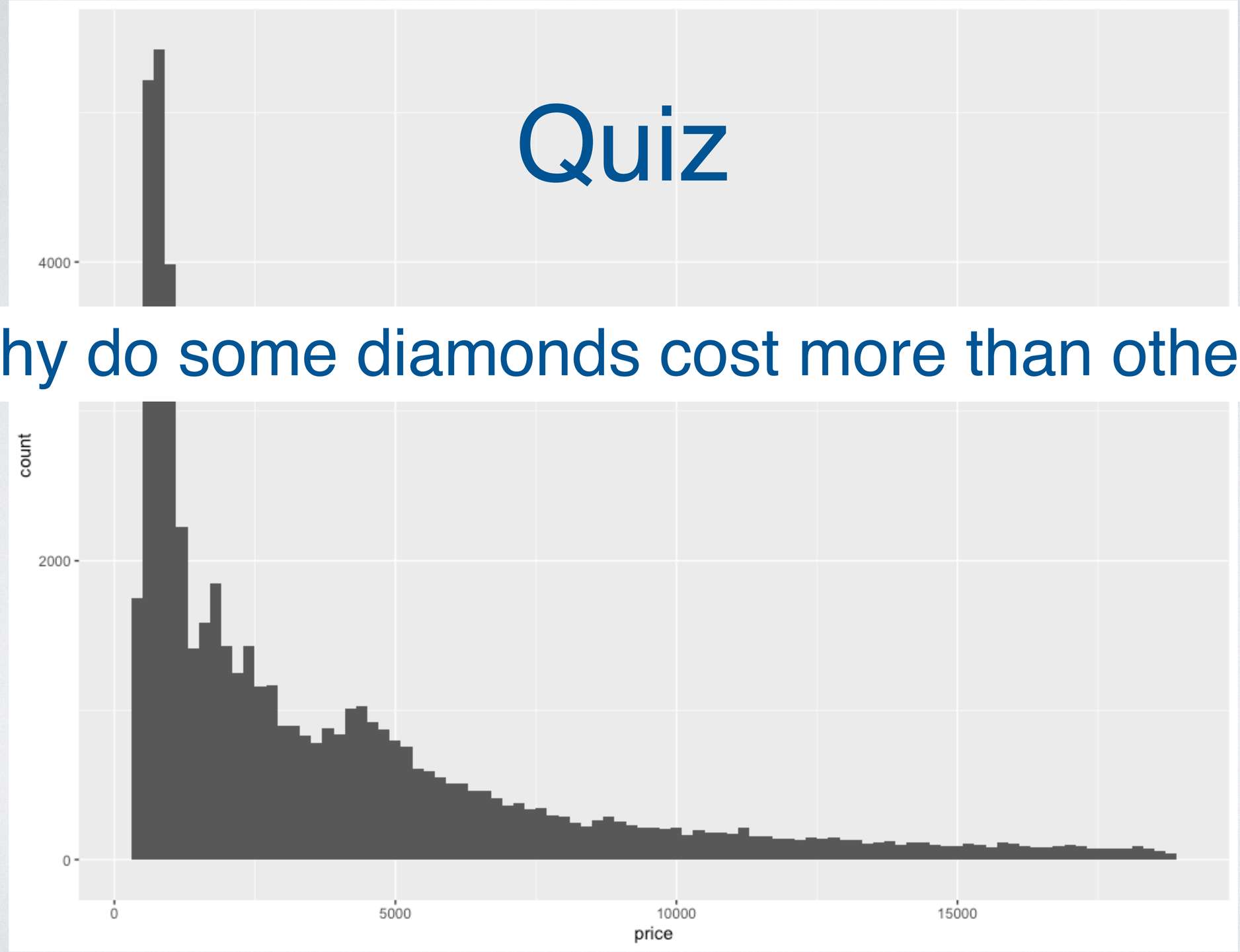





```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.01)
```

Quiz

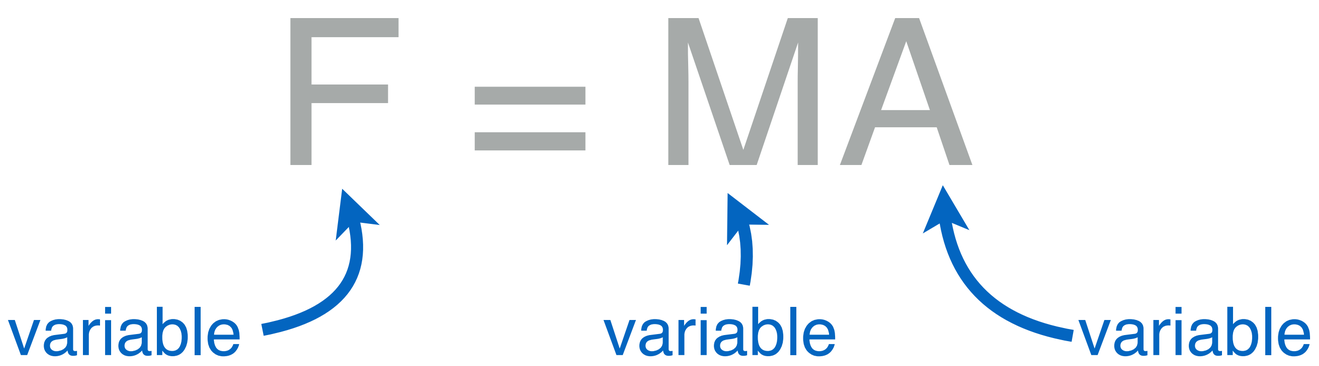
Why do some diamonds cost more than others?



Exploring Covariation

Photo by Andrew Ridley on Unsplash

An **observation** is a set of measurements made under similar conditions. An observation will contain several values, each associated with a different variable.



Variable - A quantity, quality, or property that you can measure.

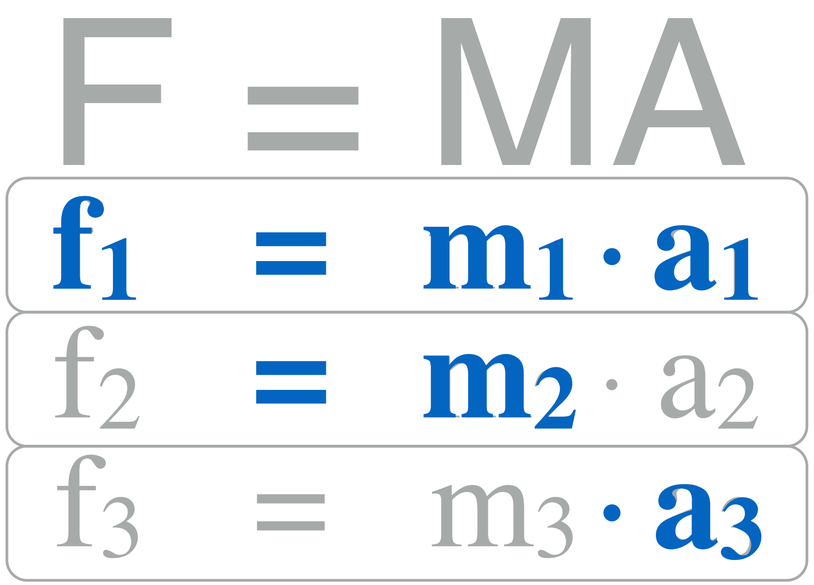
$$F = MA$$

$$f_1 \quad m_1 \quad a_1$$
$$f_2 \quad m_2 \quad a_2$$
$$f_3 \quad m_3 \quad a_3$$

Variable - A quantity, quality, or property that you can measure.

Value - The state of a variable when you measure it.

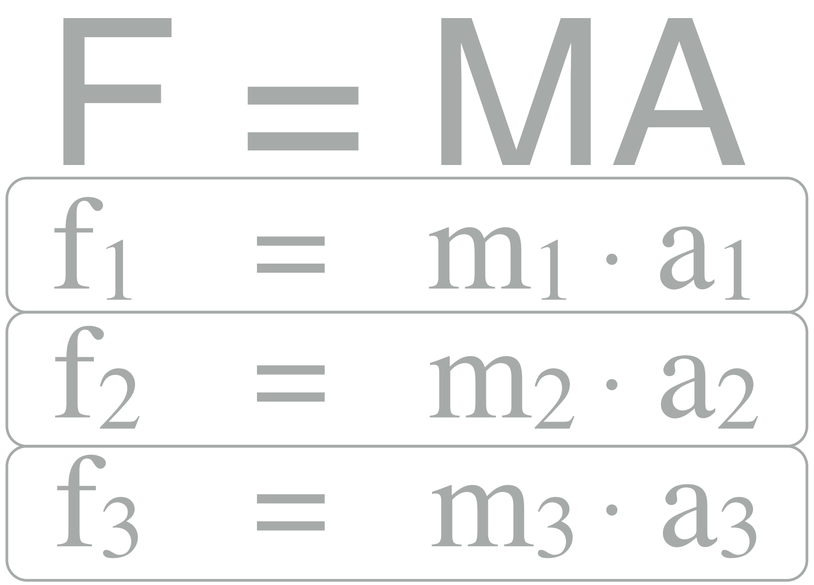
(The value can change from measurement to measurement)



Variable - A quantity, quality, or property that you can measure.

Value - The state of a variable when you measure it.

Observation - The values of several variables measured under similar conditions.



Structure of Natural Laws

Natural laws deal with **variables**,
but they operate on **values**
that appear in the same **observation**.

F	M	A	=	MA
f_1	m_1	a_1	=	$m_1 \cdot a_1$
f_2	m_2	a_2	=	$m_2 \cdot a_2$
f_3	m_3	a_3	=	$m_3 \cdot a_3$

Data = Values associated with variables and observations

Values appear as patterns

F	M	A
3.01	0.98	3.08
2.35	0.91	2.58
5.57	1.01	5.52

$$F = MA$$

$$f_1 = m_1 \cdot a_1$$

$$f_2 = m_2 \cdot a_2$$

$$f_3 = m_3 \cdot a_3$$

Laws appear as
patterns in data.

F	M	A
3.01	0.98	3.08
2.35	0.91	2.58
5.57	1.01	5.52
0.62	1.09	0.56
4.15	0.89	4.69
5.07	1.05	4.84
7.56	0.93	8.12

$$F = MA$$

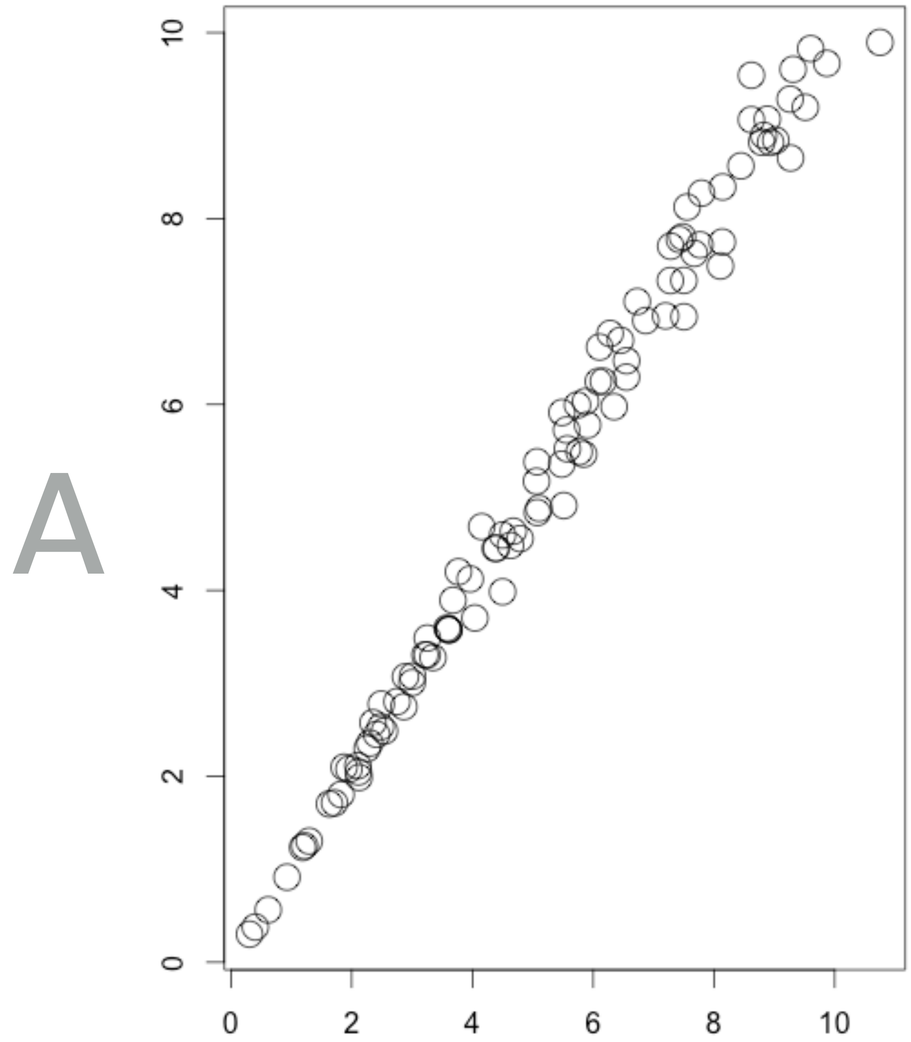
$$f_1 = m_1 \cdot a_1$$

$$f_2 = m_2 \cdot a_2$$

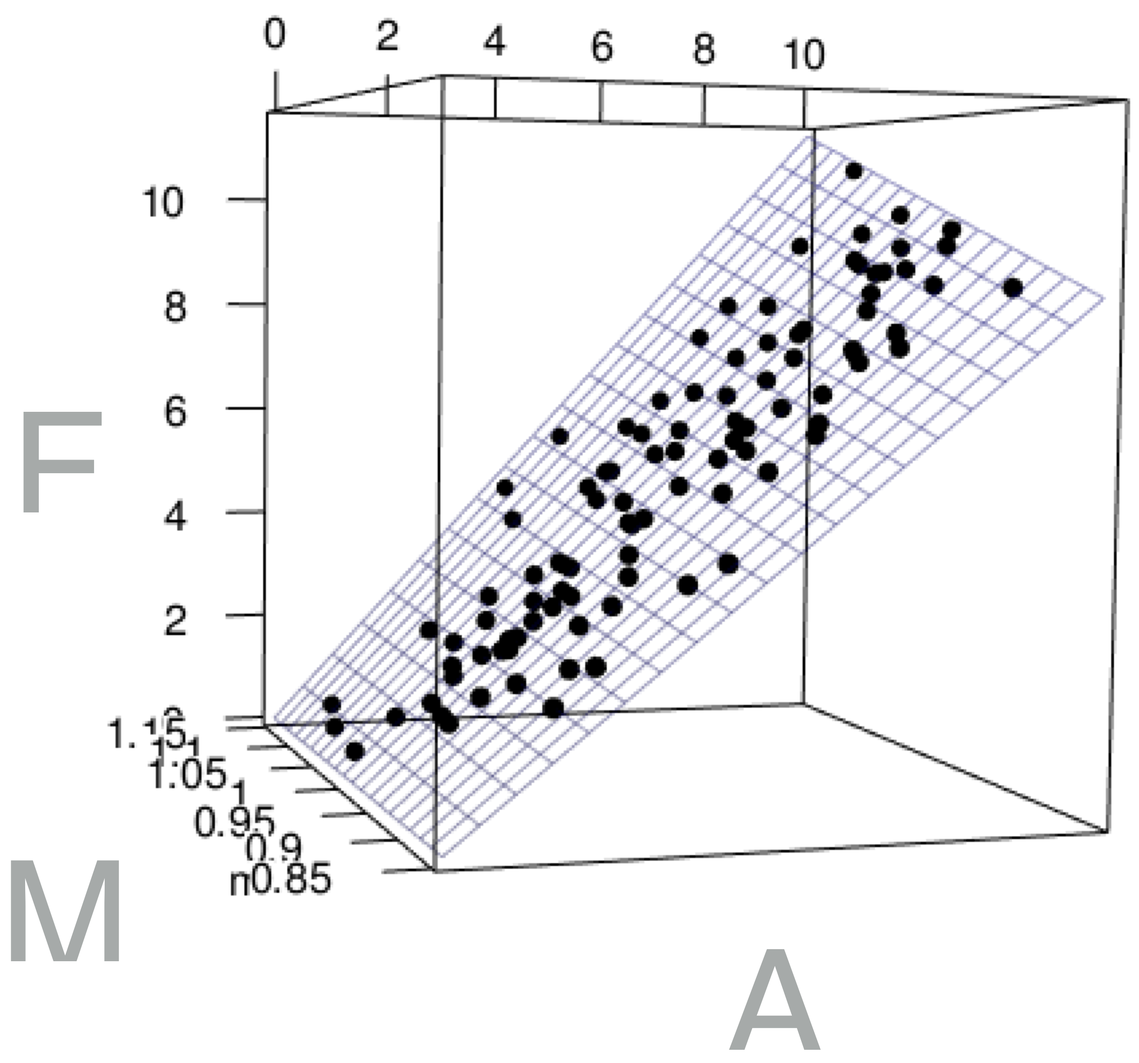
$$f_3 = m_3 \cdot a_3$$

Laws appear as **patterns** in data.

F	M	A
0.62	1.09	0.56
1.30	0.99	1.30
1.63	0.96	1.70
1.72	1.00	1.71
1.82	1.01	1.80
1.95	0.94	2.08
2.11	1.03	2.05

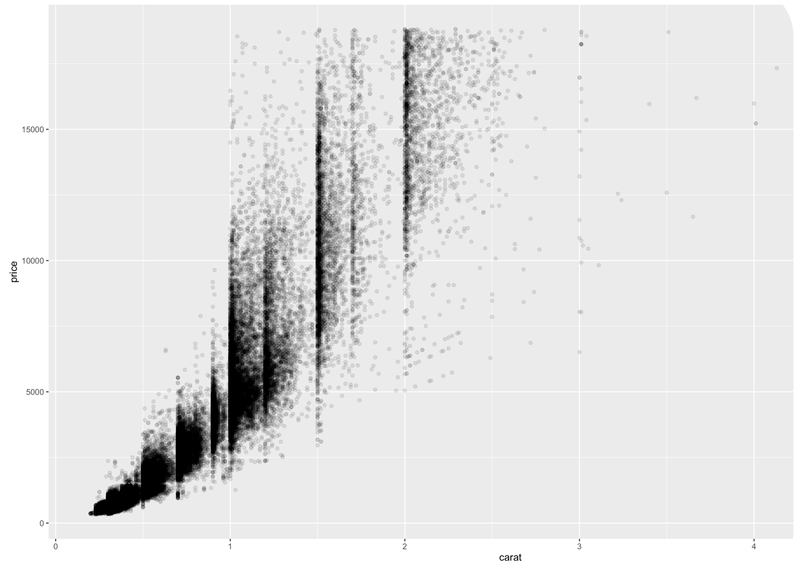


F	M	A
0.62	1.09	0.56
1.30	0.99	1.30
1.63	0.96	1.70
1.72	1.00	1.71
1.82	1.01	1.80
1.95	0.94	2.08
2.11	1.03	2.05



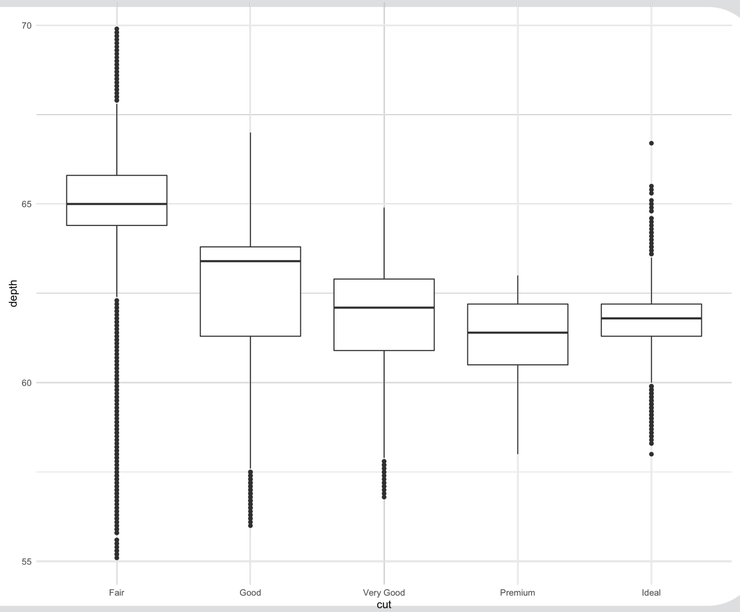
Continuous vs continuous

geom_point()



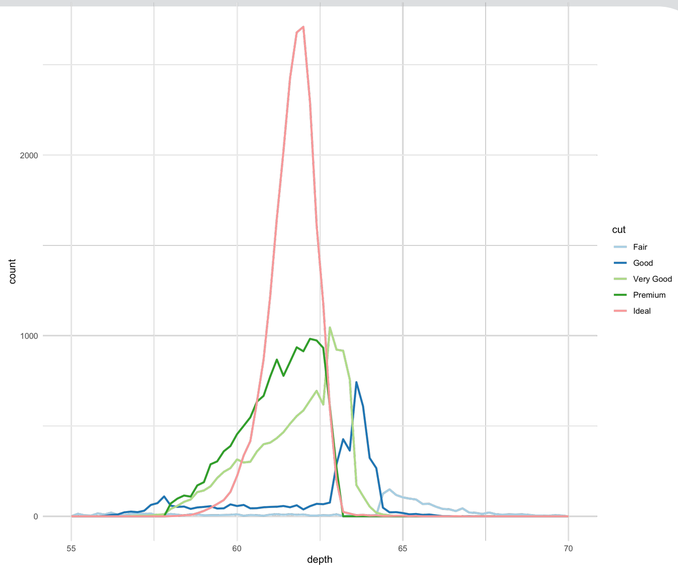
Continuous vs categorical

geom_boxplot()



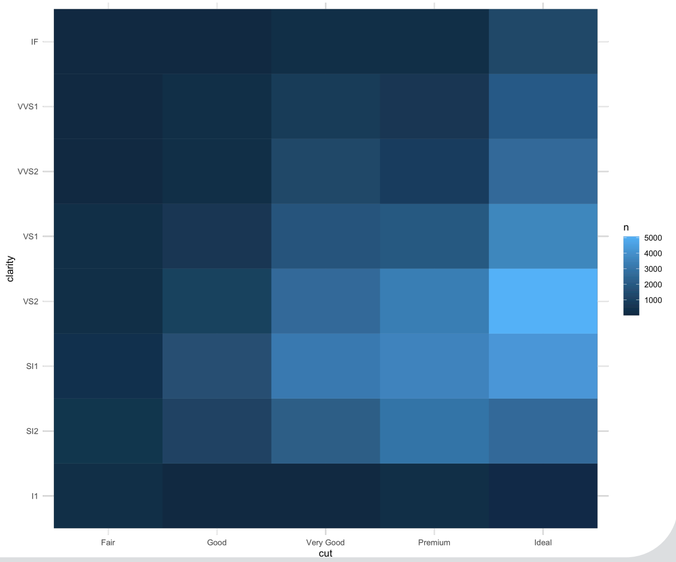
Continuous vs categorical

geom_freqpoly()



Categorical vs categorical

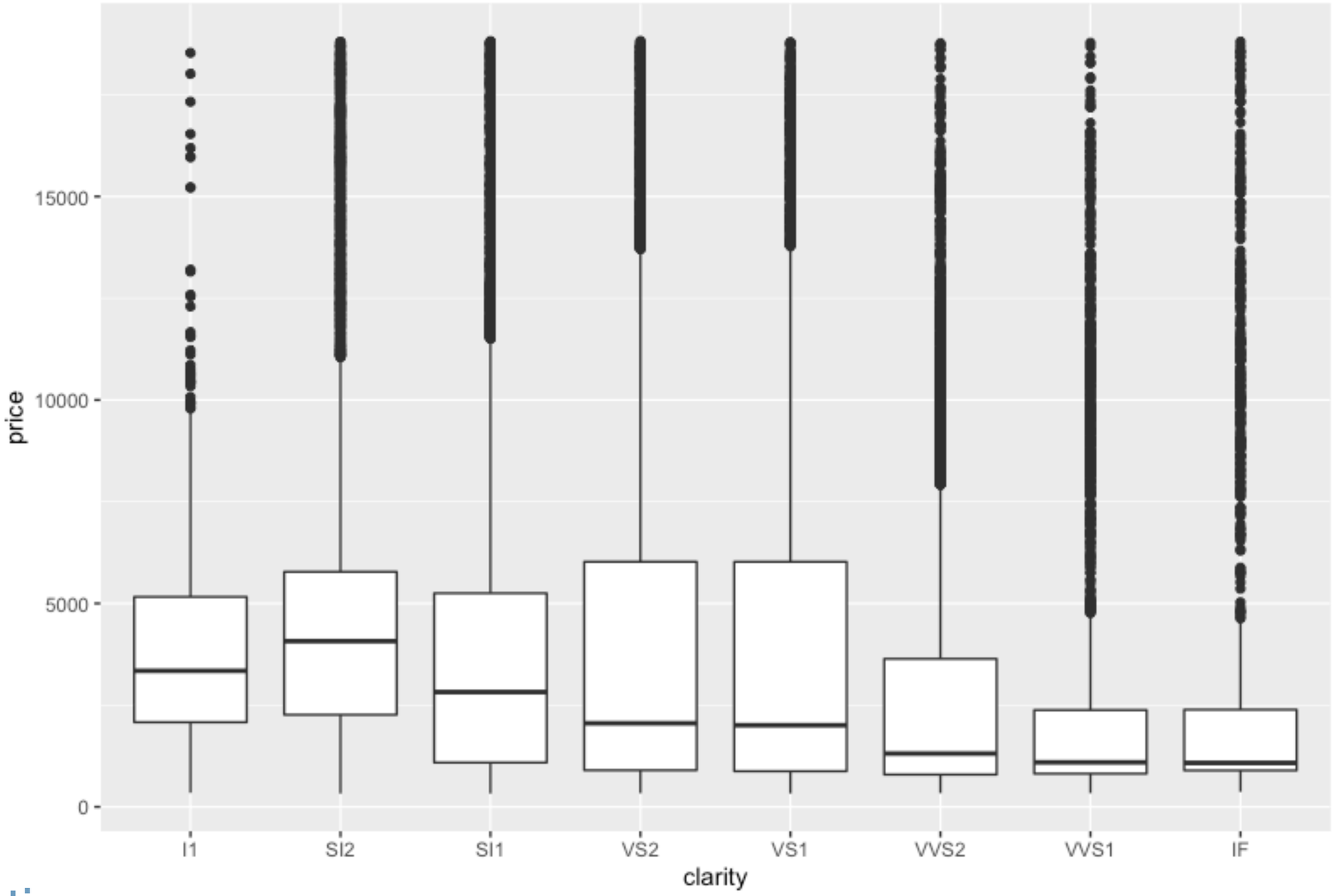
geom_tile()



Your Turn 4

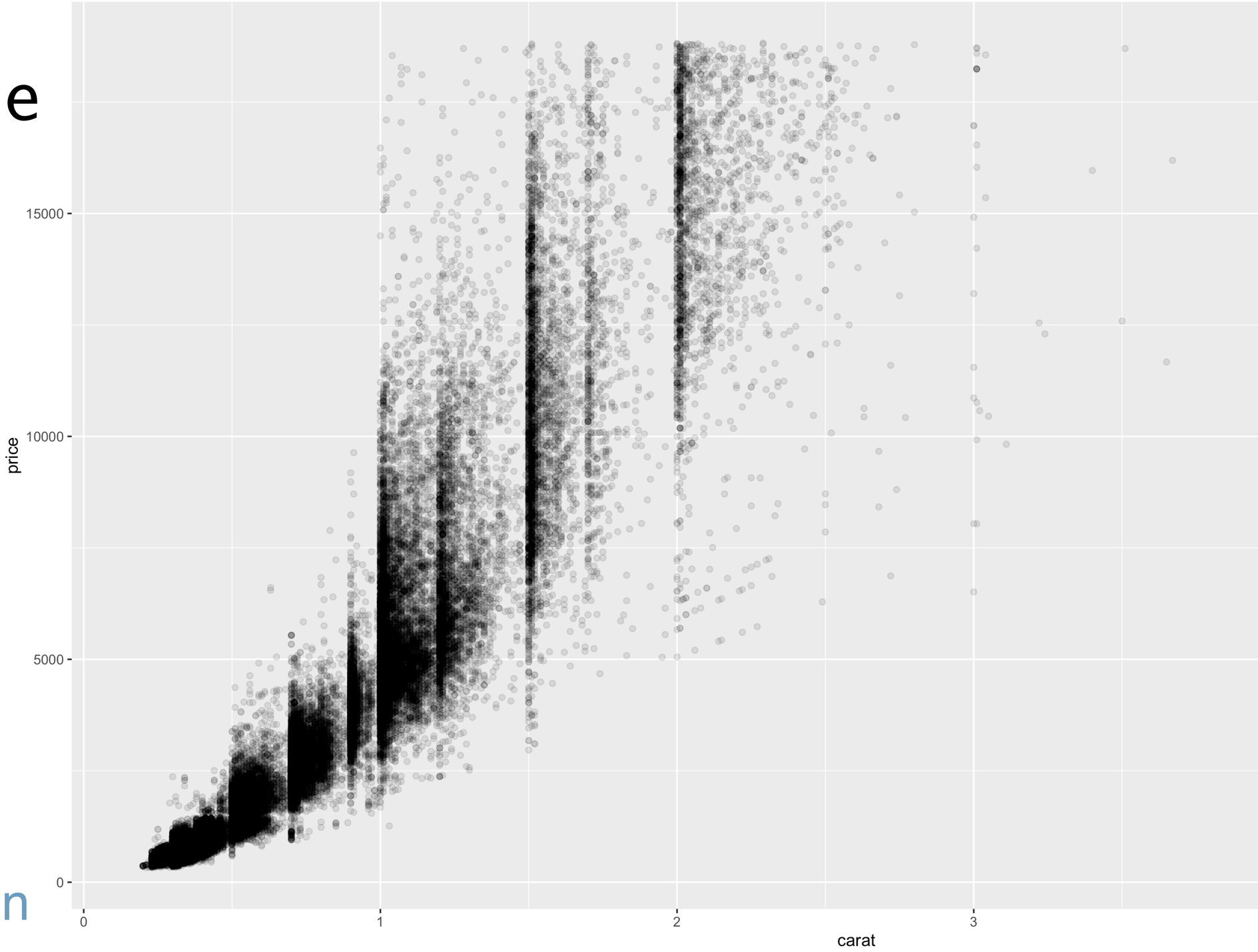
Visualize the relationships *between* different variables in `diamonds` and variable `price`. **What patterns do you notice?**

```
clarity ~
  price
```

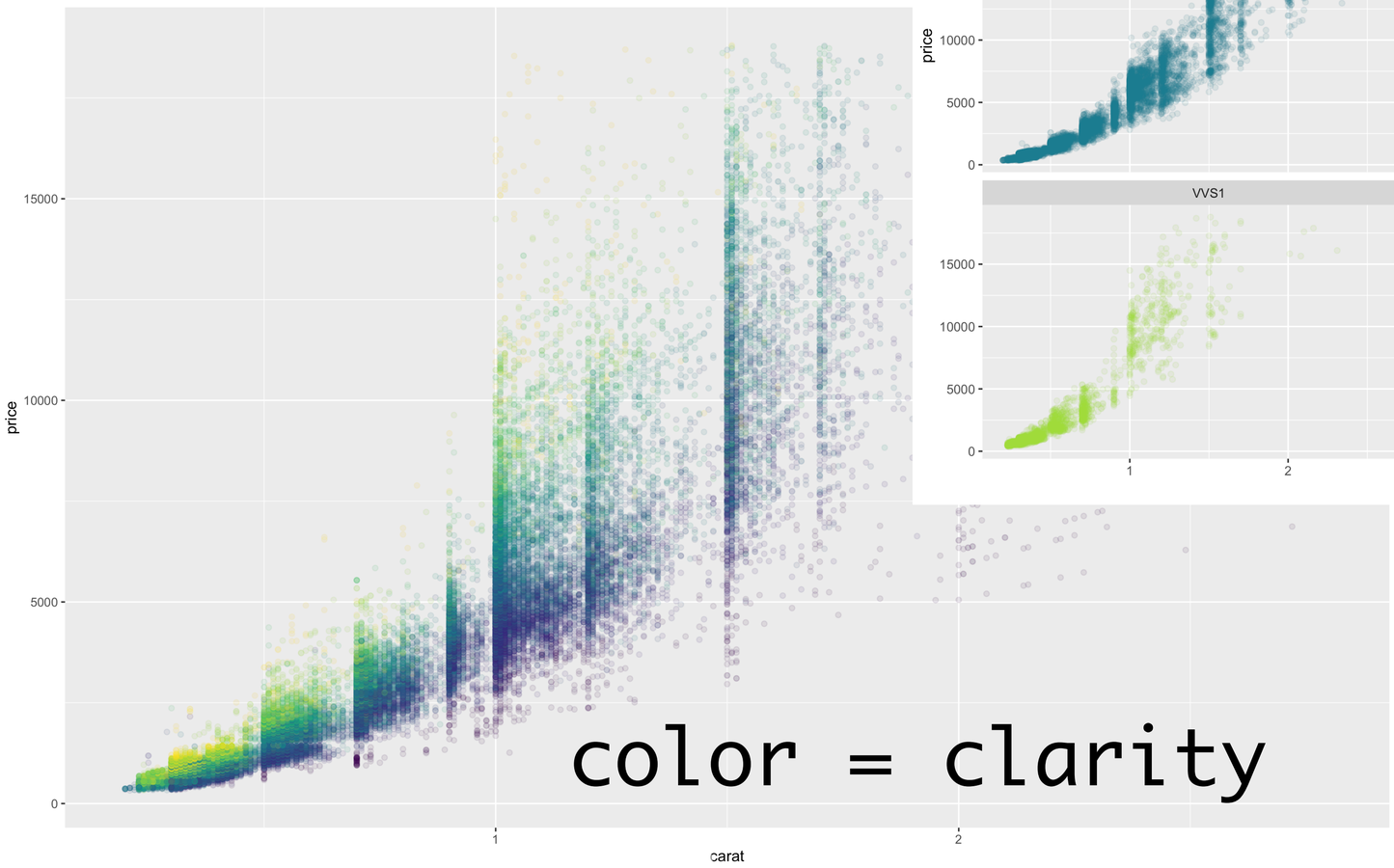


Exploring covariation

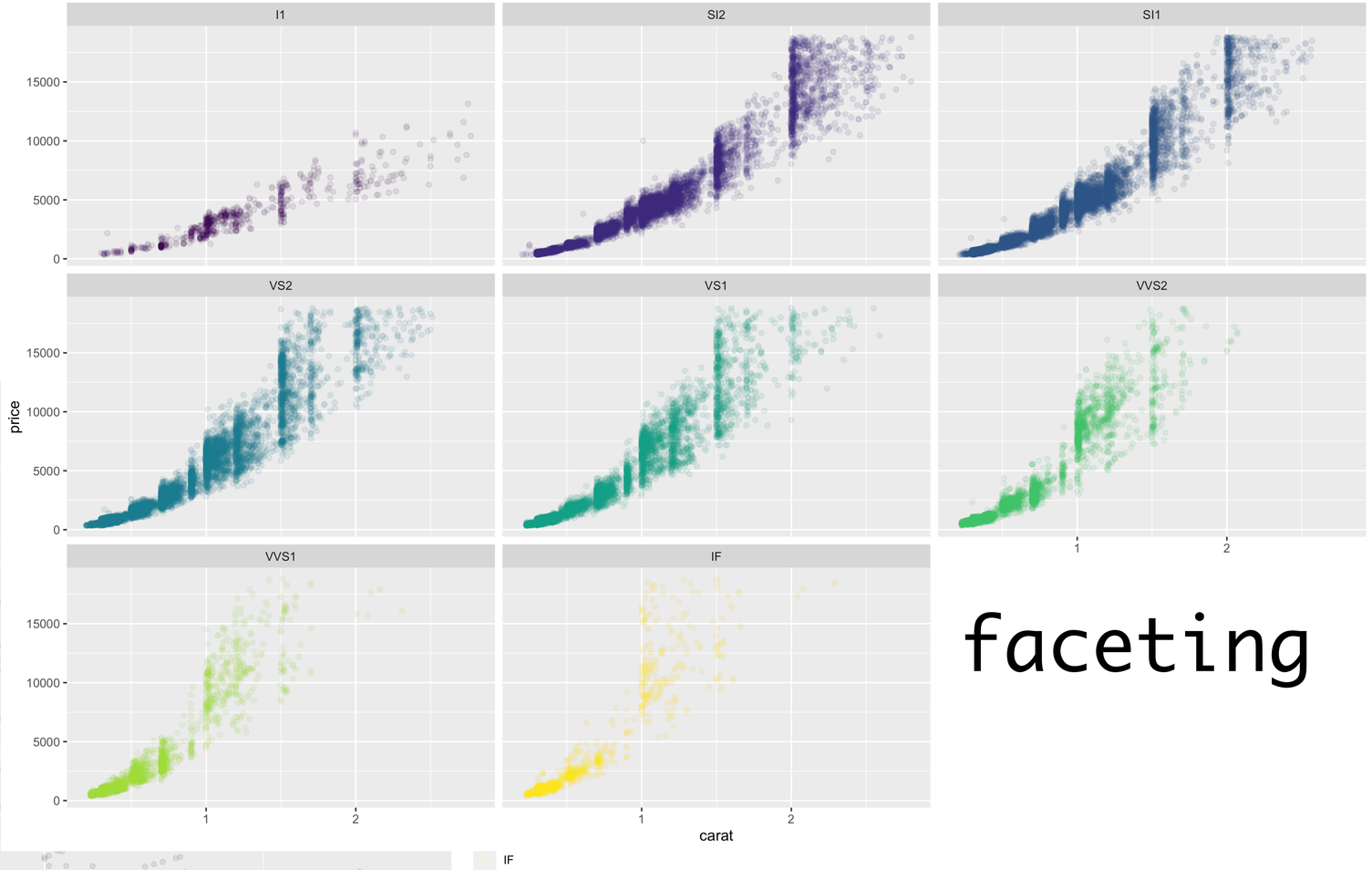
carat ~ price



Exploring covariation



color = clarity



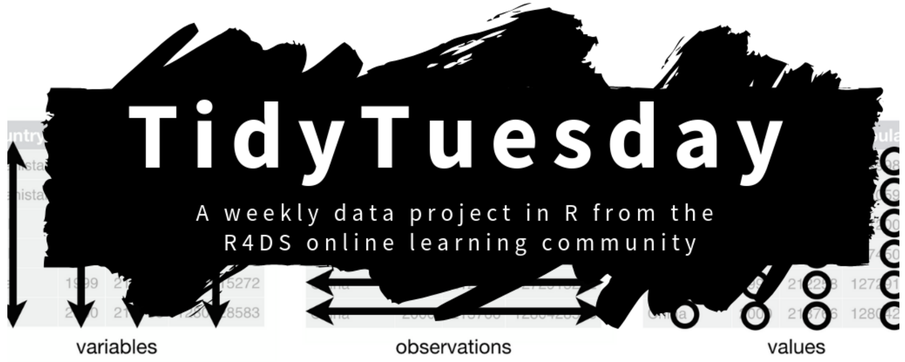
faceting

Exploring covariation

Wine ratings

Photo by Maksym Kaharlytskyi on Unsplash

wine_ratings data




```
> glimpse(wine_ratings)
Rows: 129,971
Columns: 14
$ ...1                  <dbl>
$ country               <chr>
$ description           <chr>
$ designation           <chr>
$ points                <dbl>
$ price                 <dbl>
$ province              <chr>
$ region_1              <chr>
$ region_2              <chr>
$ taster_name           <chr>
$ taster_twitter_handle <chr>
$ title                 <chr>
$ variety               <chr>
$ winery                <chr>
```

- **~130,000** wine ratings
- Wine (variety, points, price)
- Geography (wine, winery)
- Taster
- Price

Your Turn 5

Consider the `wine_ratings` data + think of a question that interests you. **Make a graph that explores this question.**

Your Turn 6

Consider the `wine_ratings` data + think of a question that interests you. **Make a graph that explores this question.**

Now think of a question your last graph raises. Iterate and **make a graph that explores further.**

Your Turn 7

Now think of a question your last graph raises. Iterate and **make a graph that explores further.**

Consider your second graph. **Ask one more question, and investigate with one more graph.**

Curiosity + skepticism

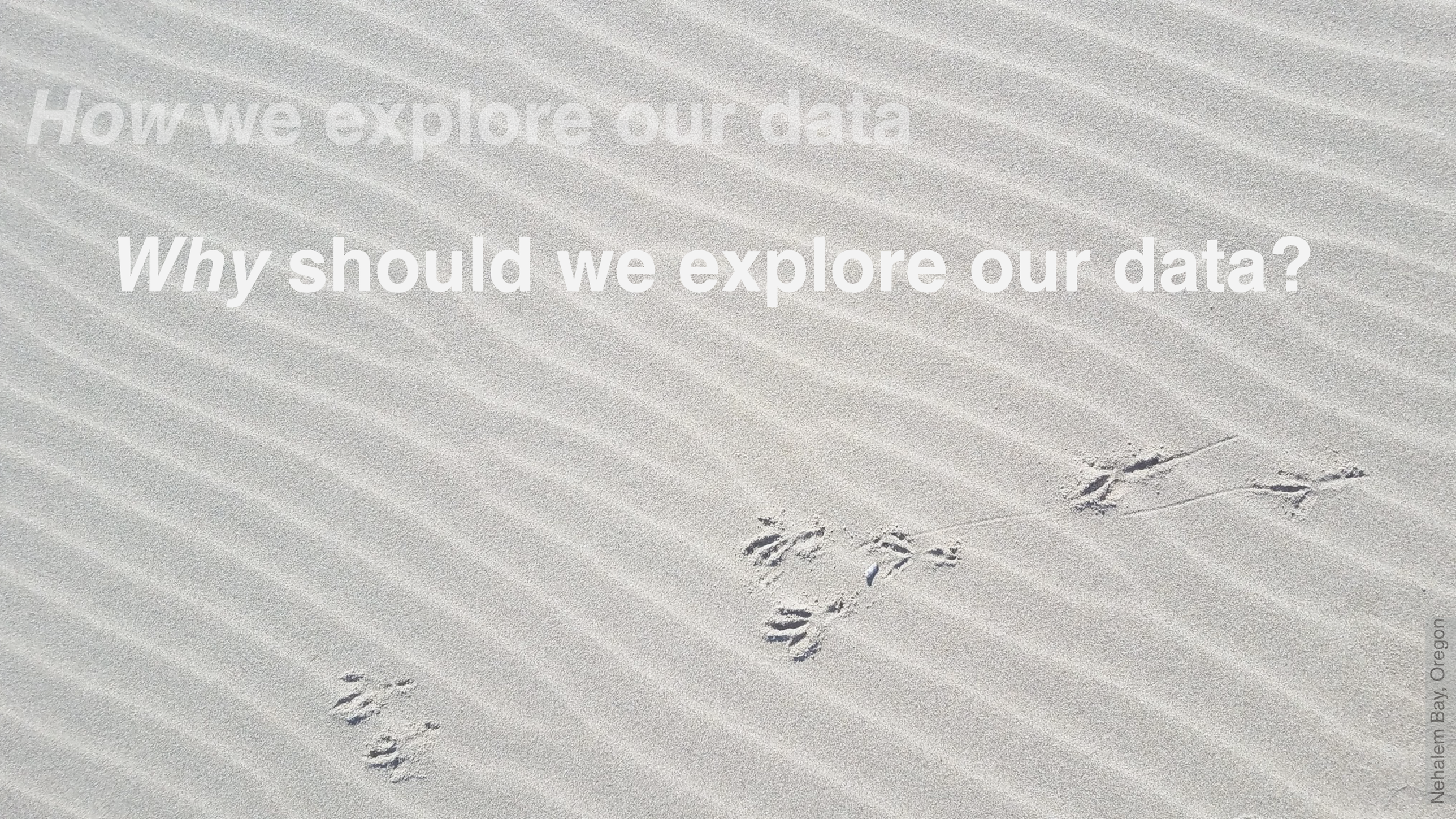
Nehalem Bay, Oregon

How we explore our data

Nehalem Bay, Oregon

How we explore our data

Why should we explore our data?



Nehalem Bay, Oregon

How we explore our data

Why should we explore our data?

Find patterns + relationships

Nehalem Bay, Oregon

Why should we explore our data?

What causes patterns in our data?

Nehalem Bay, Oregon

Why should we explore our data?

What causes patterns in our data?

covariation + coincidence

Nehalem Bay, Oregon

Curiosity + skepticism

Nehalem Bay, Oregon

Next up : lunch (12-1pm)

Kristin Bott

@kristin (Slack)

Nehalem Bay, Oregon

Next up: lunch (12-1pm)

Afternoon:

RMarkdown (Brendan)

IDE best practices (Dan)

Kristin Bott

@kristin (Slack)

Nehalem Bay, Oregon