Exploratory data analysis

Sleeping Beauty, WA

Posit Academy

posit::conf 2024

posit™

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

Diamonds dataset

```
diamonds
```

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<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,

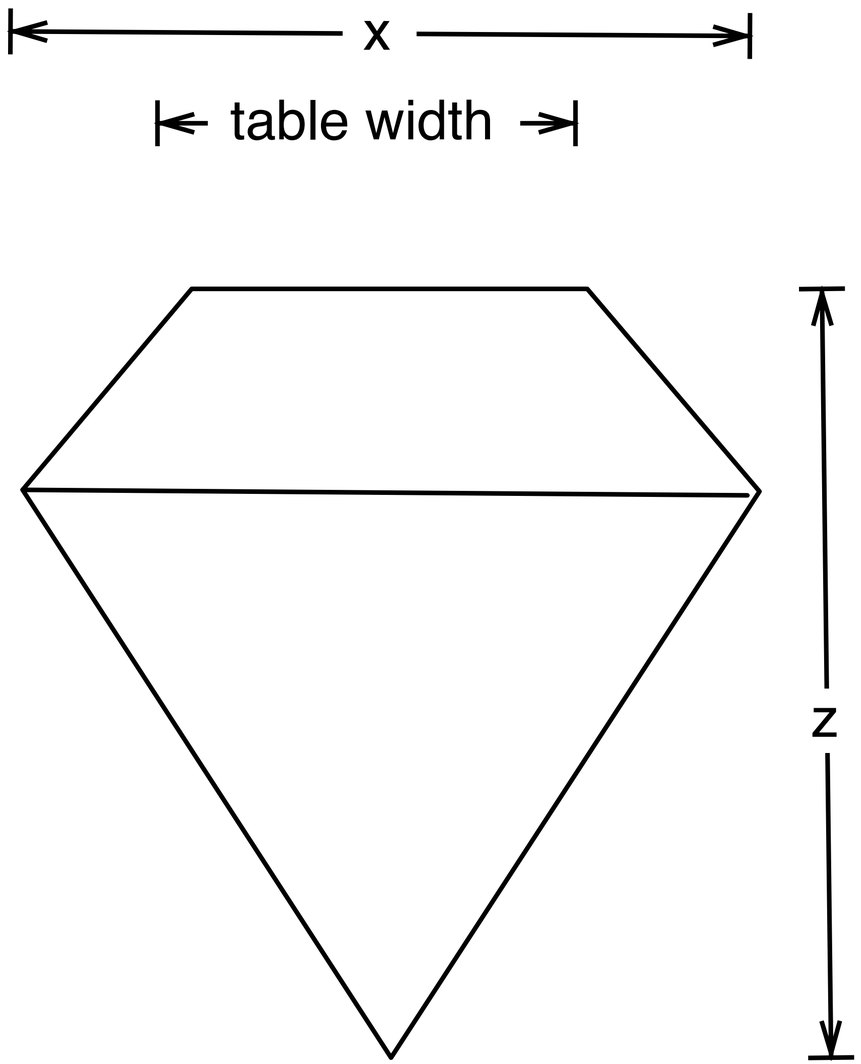
table,

length (x),

width (y),

height (z)

depth	table	price	x	y	z
<dbl>	<dbl>	<int>	<dbl>	<dbl>	<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

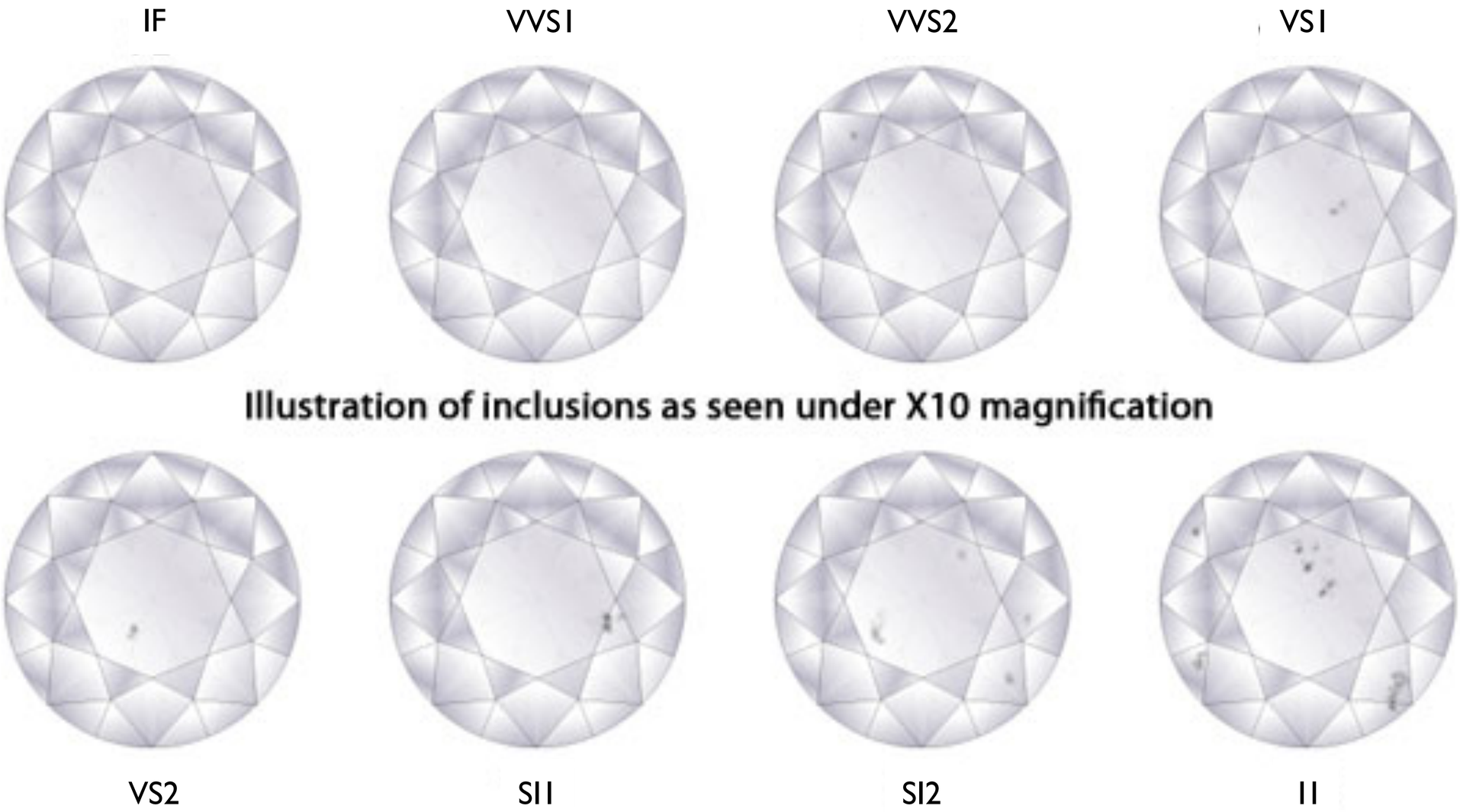
4 C's : Carat, **color**, clarity, cut

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

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

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<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

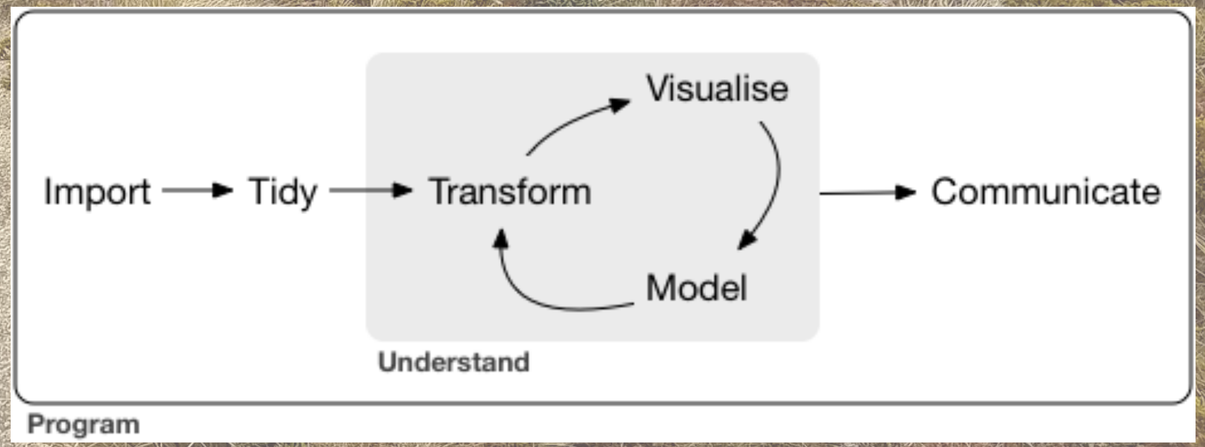
Photo by Jon Tyson on Unsplash

What is EDA?

Generate questions about data

Investigate ideas

Search for answers



Urchin Track, North Island, NZ

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

~~Your~~ Our Turn

What questions do we have about `diamonds`?

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<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

Our ~~Your~~ Turn

What questions do we have about `diamonds`?

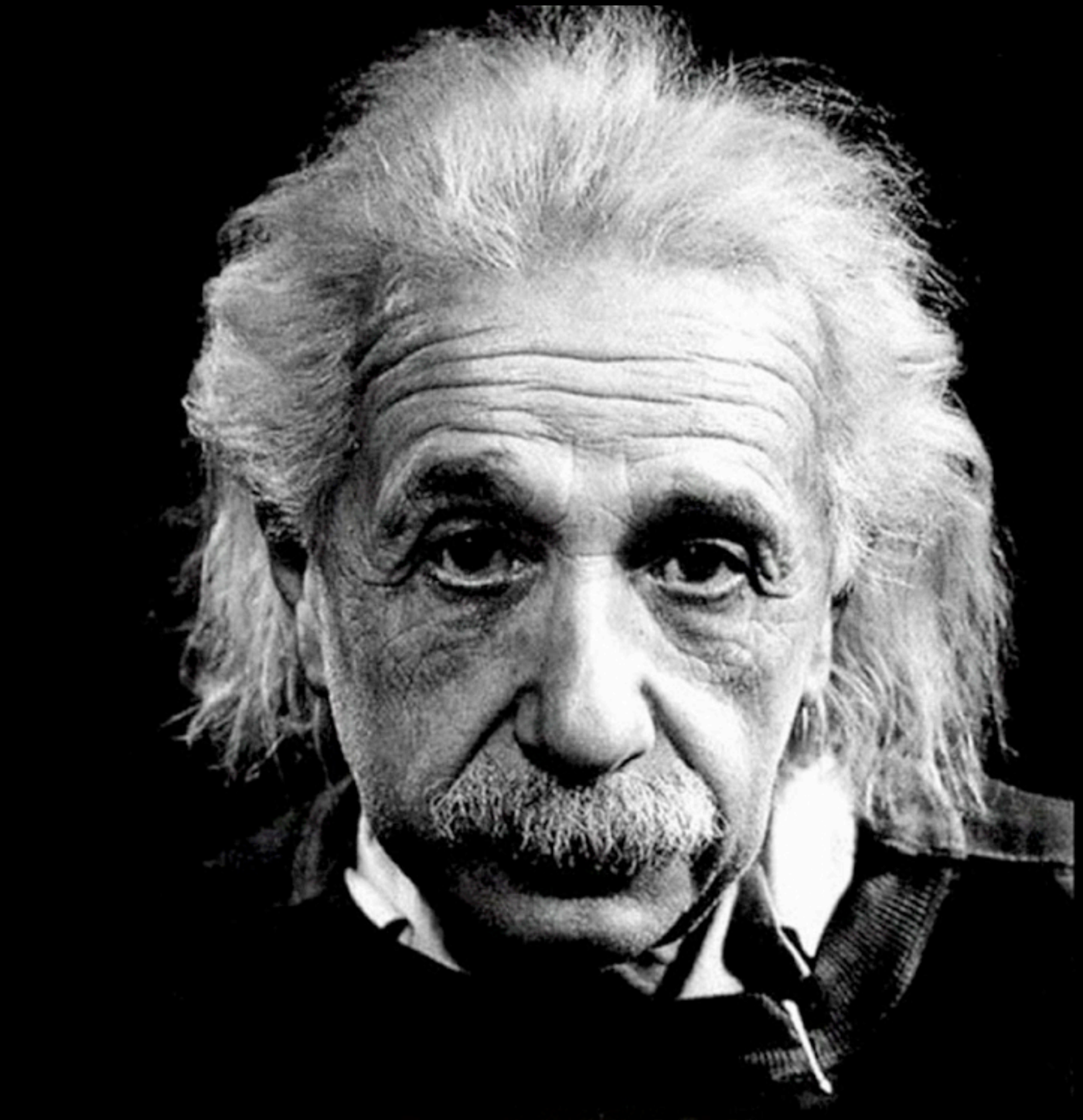
Can we *categorize* these questions?

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<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

~~Your~~ Our Turn

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

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



What did Einstein think was a universal constant?

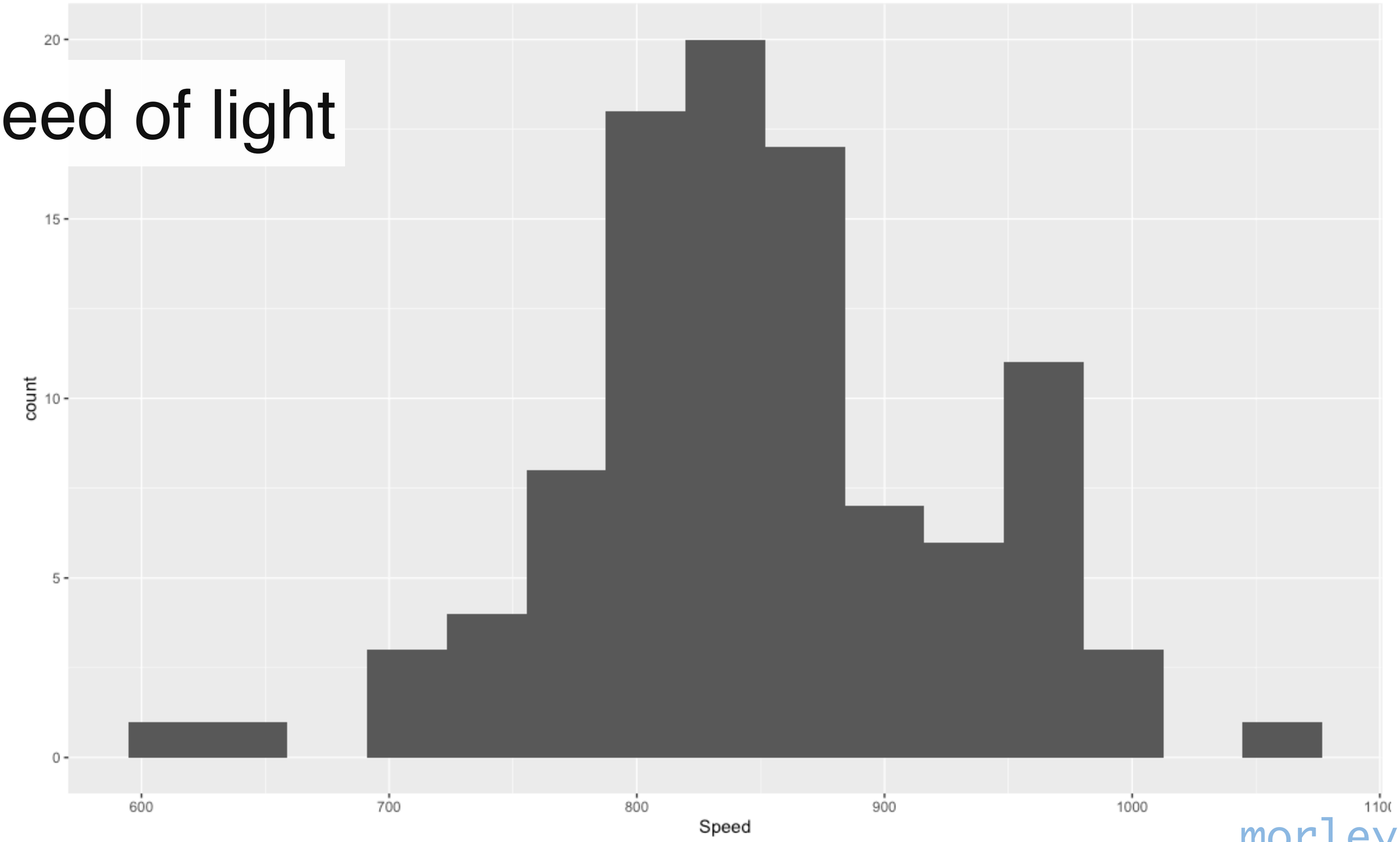
$$E = m\mathbf{c}^2$$

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
```

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

What is a **continuous** variable?

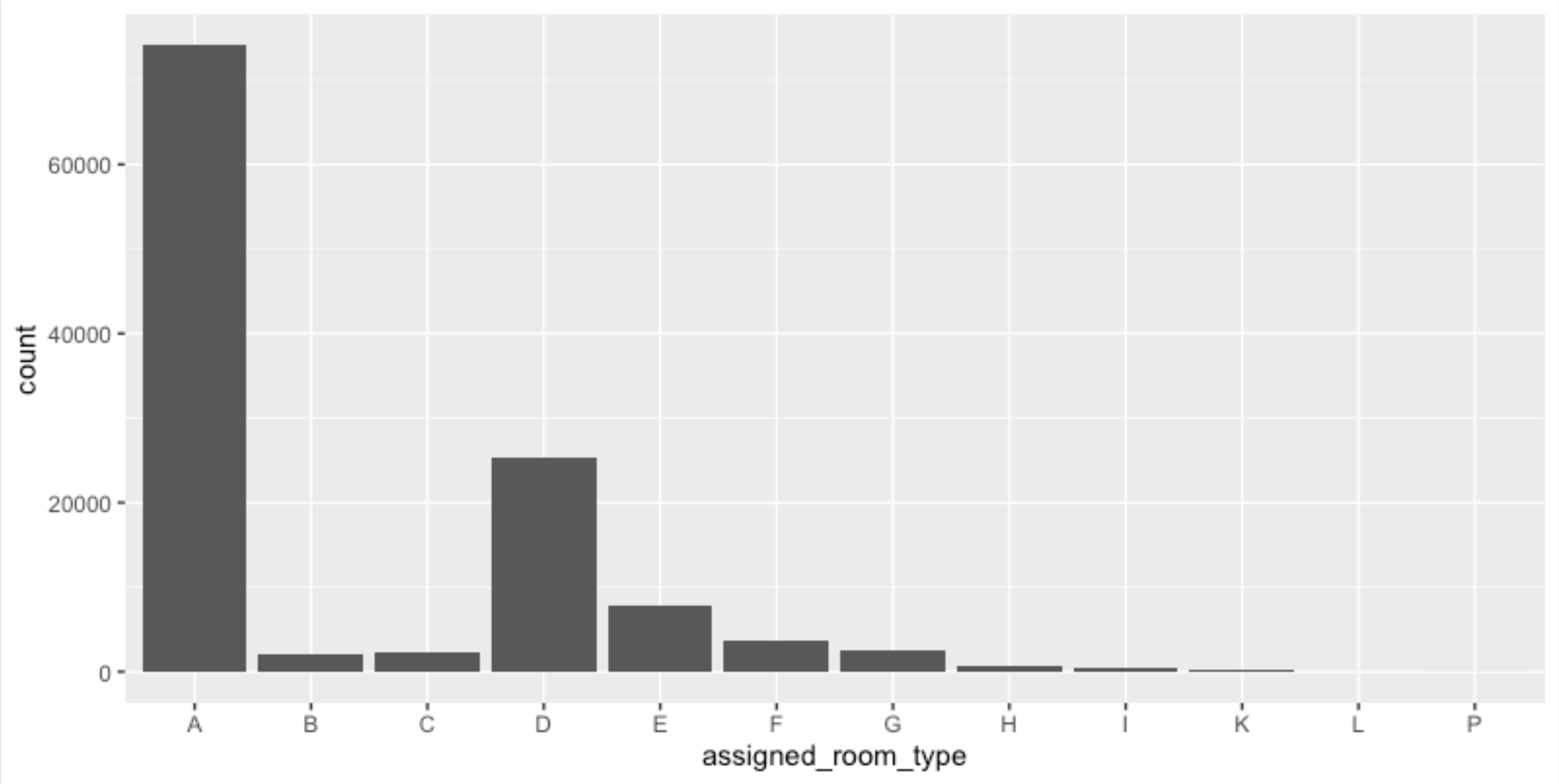
What is a **categorical** variable?

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

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

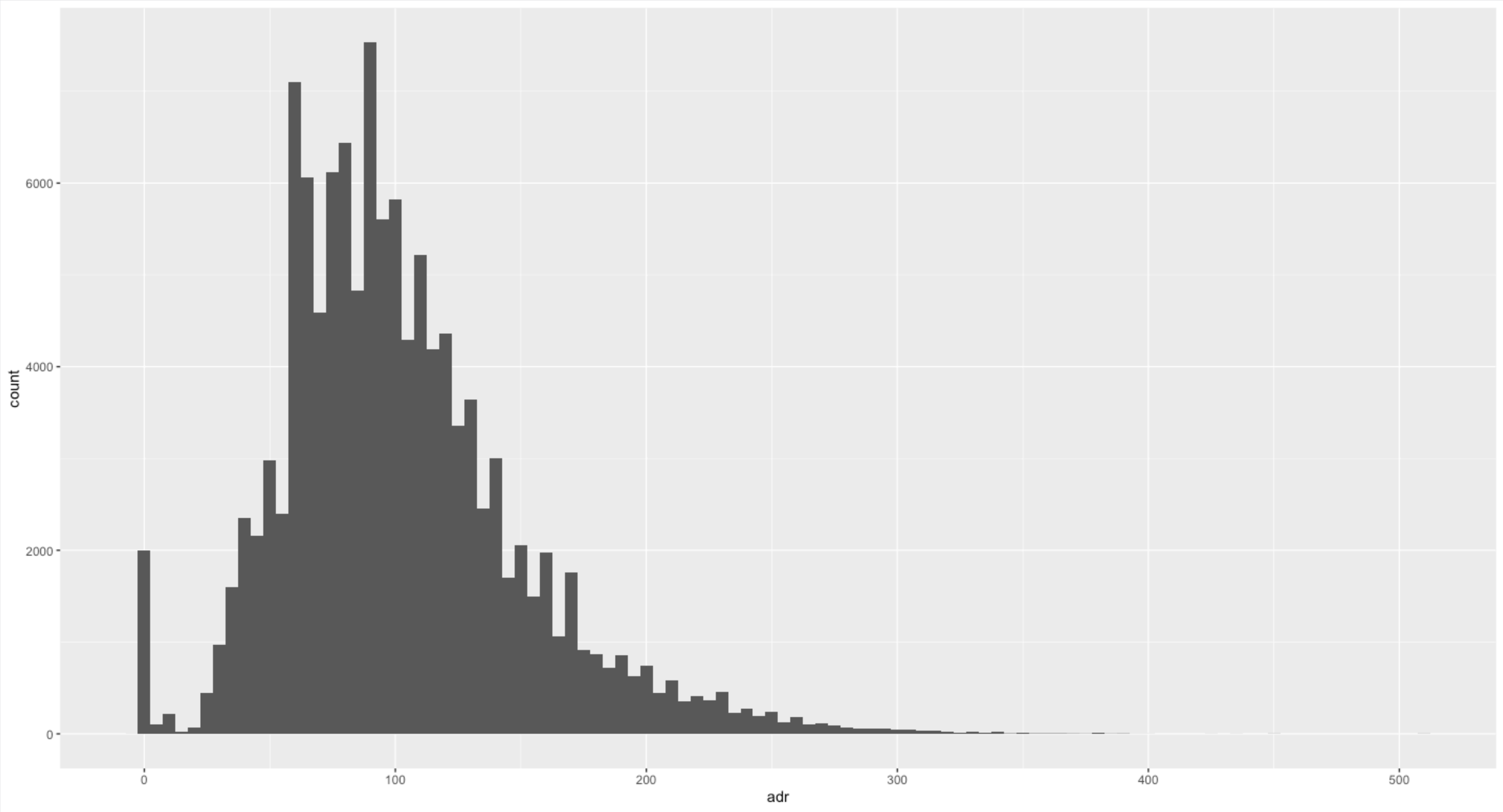
Image: Flazingo Photos | flickr.com/photos/124247024@N07/14090440354

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



`hotels$assigned_room_type`

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



hotels$adr

Your Turn 2

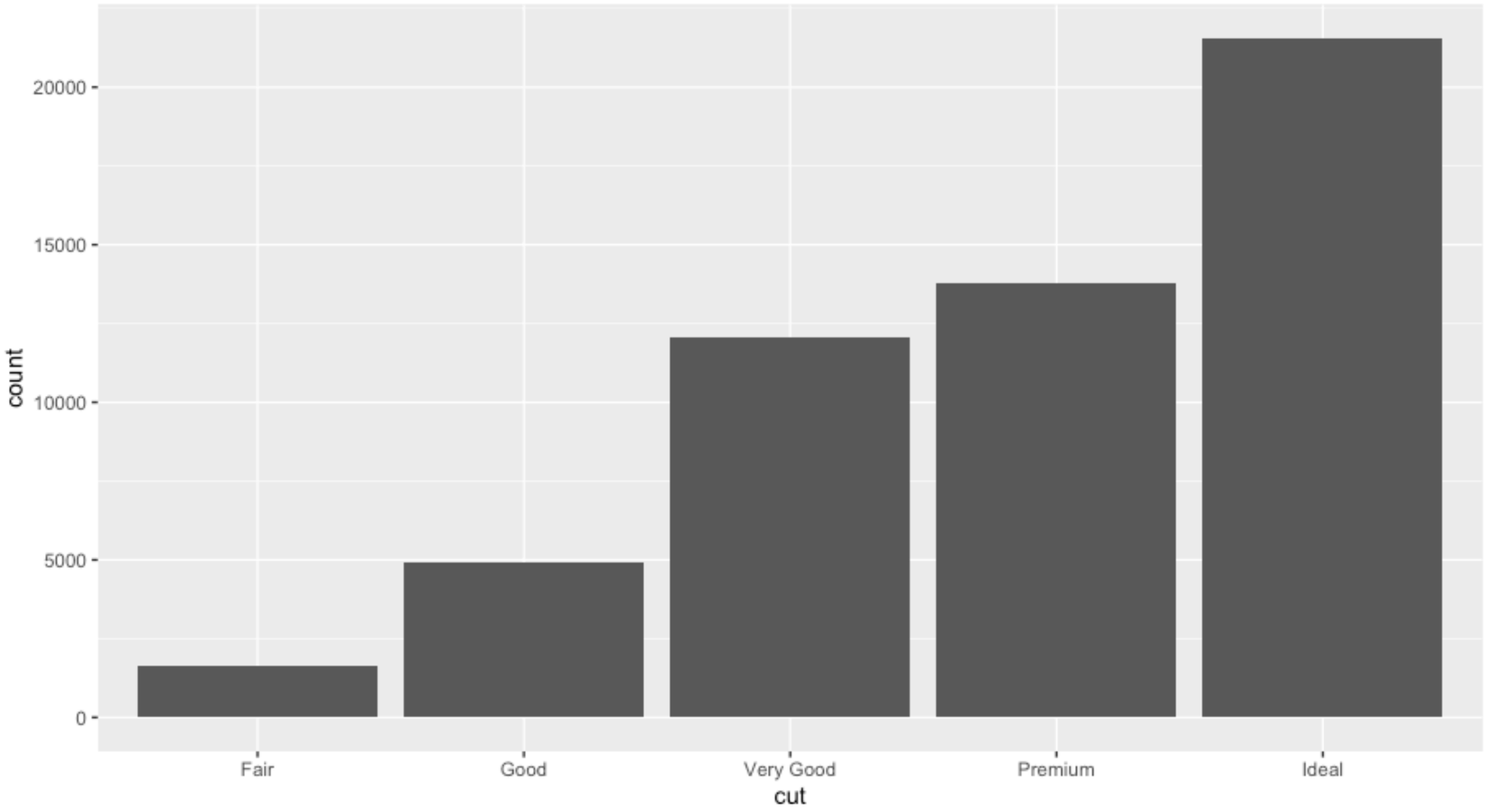
Visualize 1) the distribution for `cut`

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?*

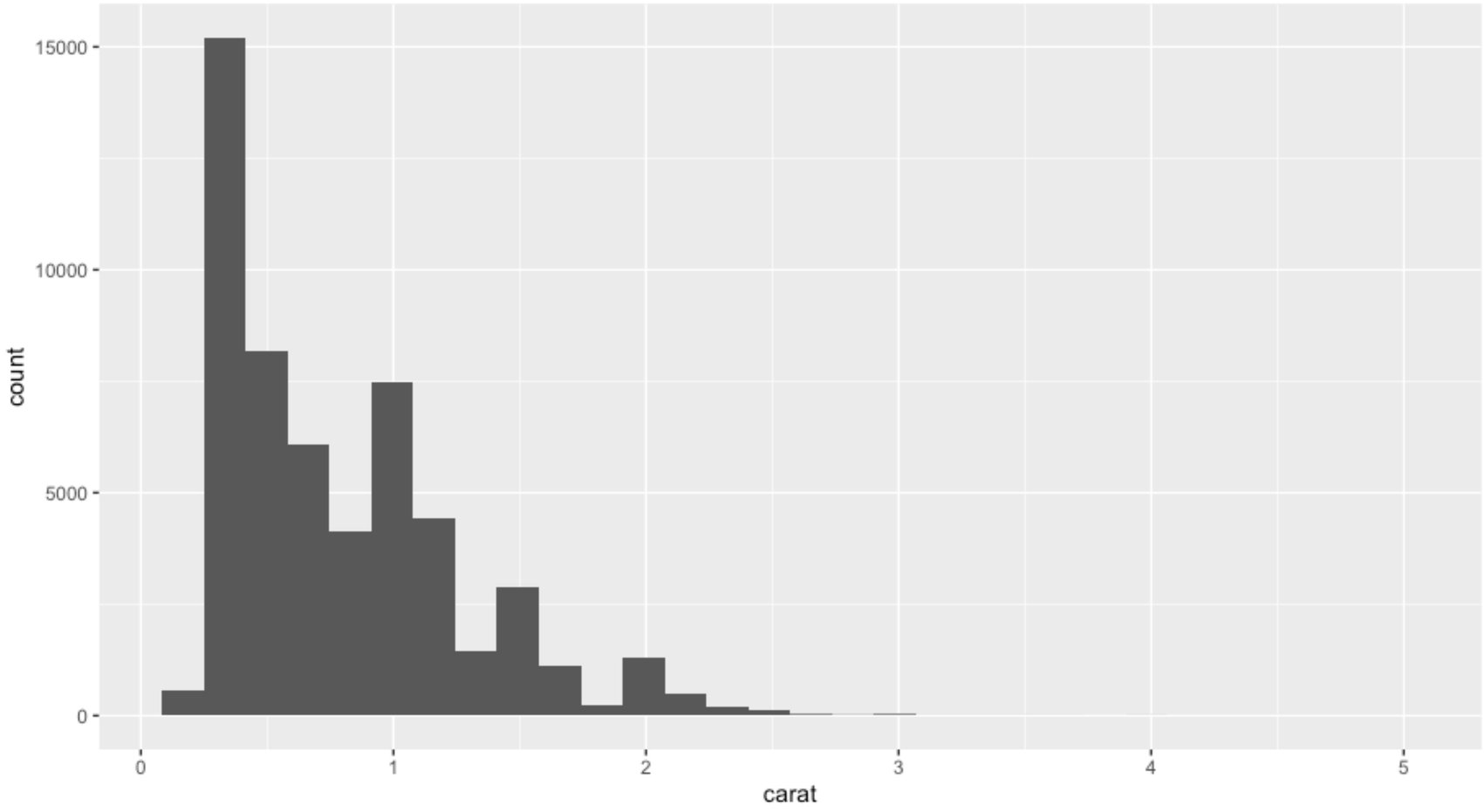
Visualize (1) the distribution for `cut`




```
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))
```

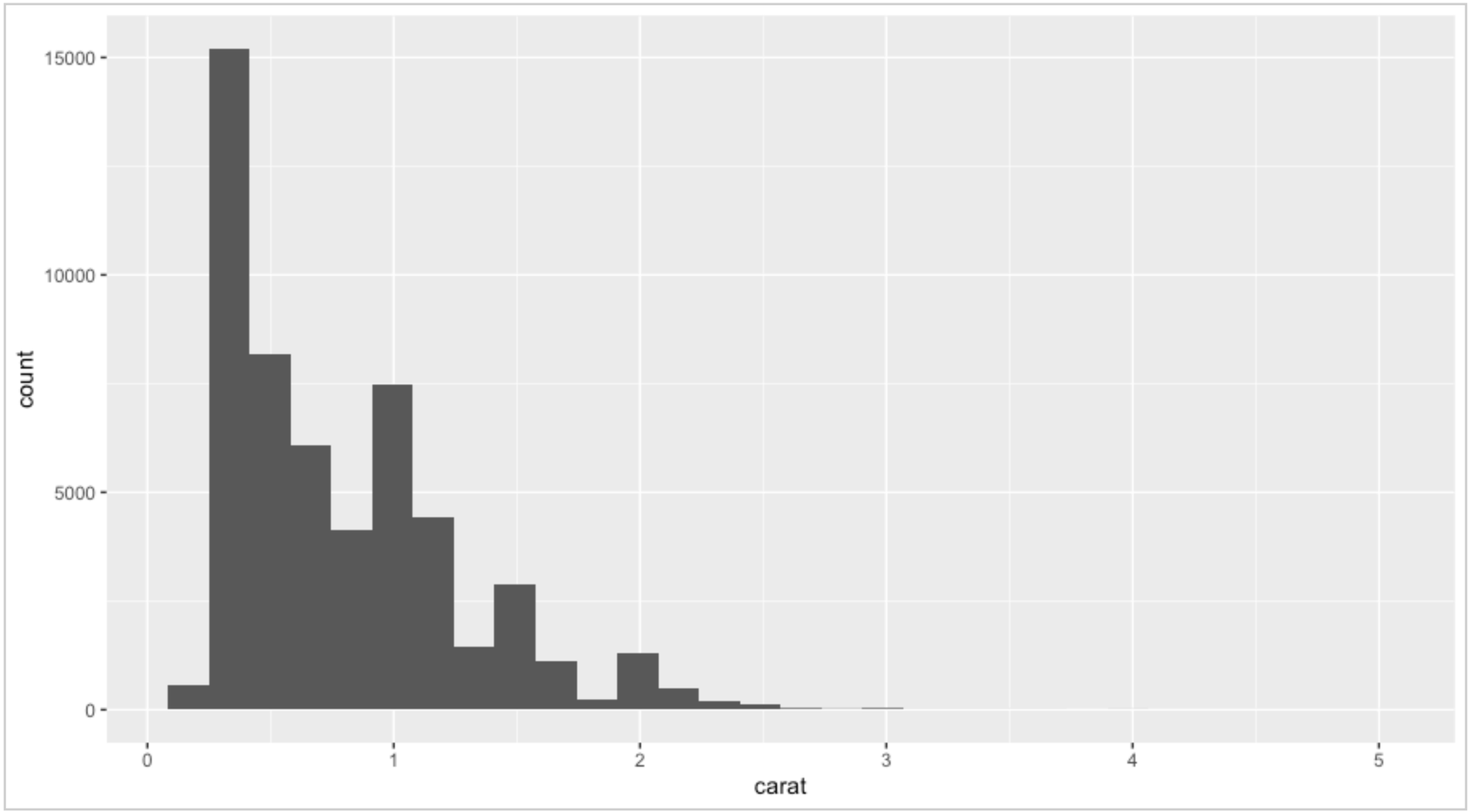
Your Turn 3

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

Share your results with your group.

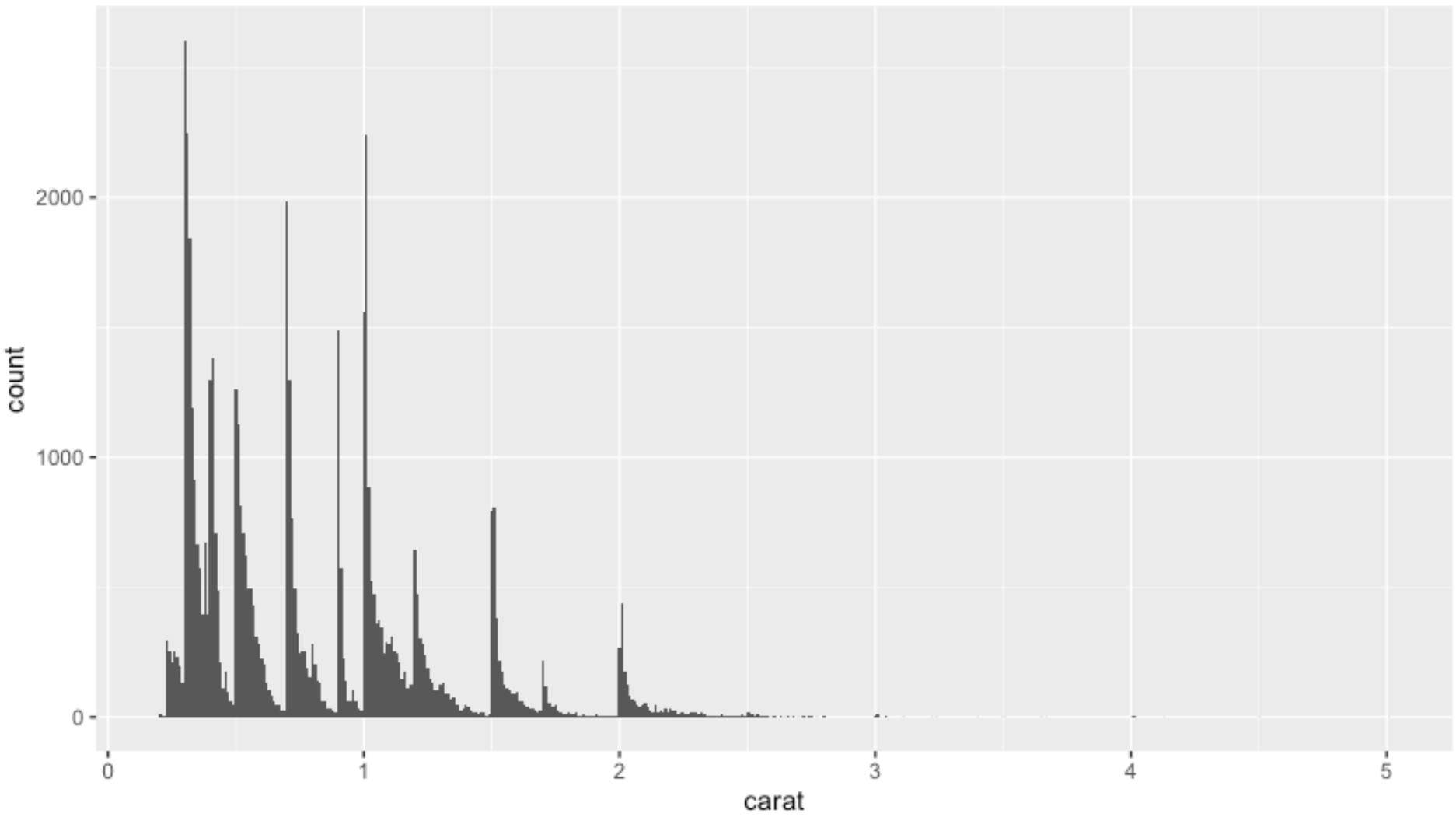
(hint: when you make a histogram, explore multiple binwidths)

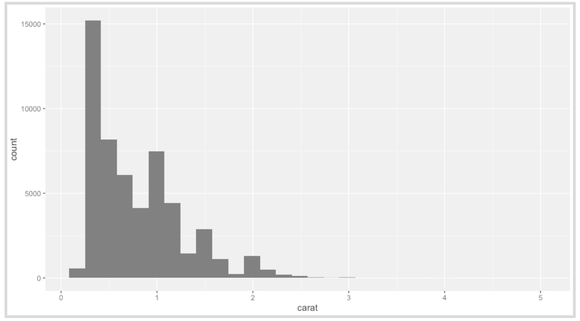
Revisit the distribution for `carat`

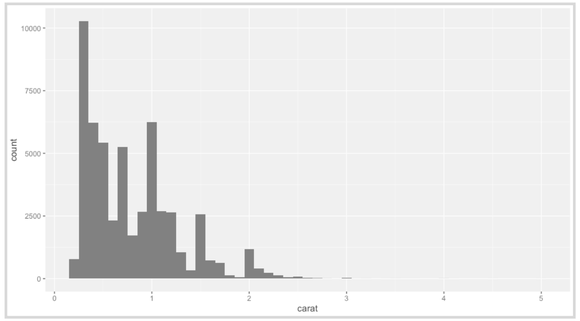


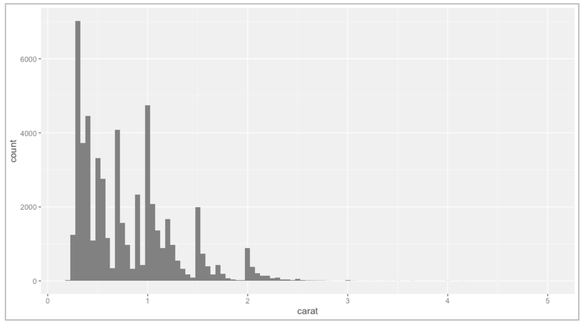

```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.1)
```

Revisit the distribution for `carat`





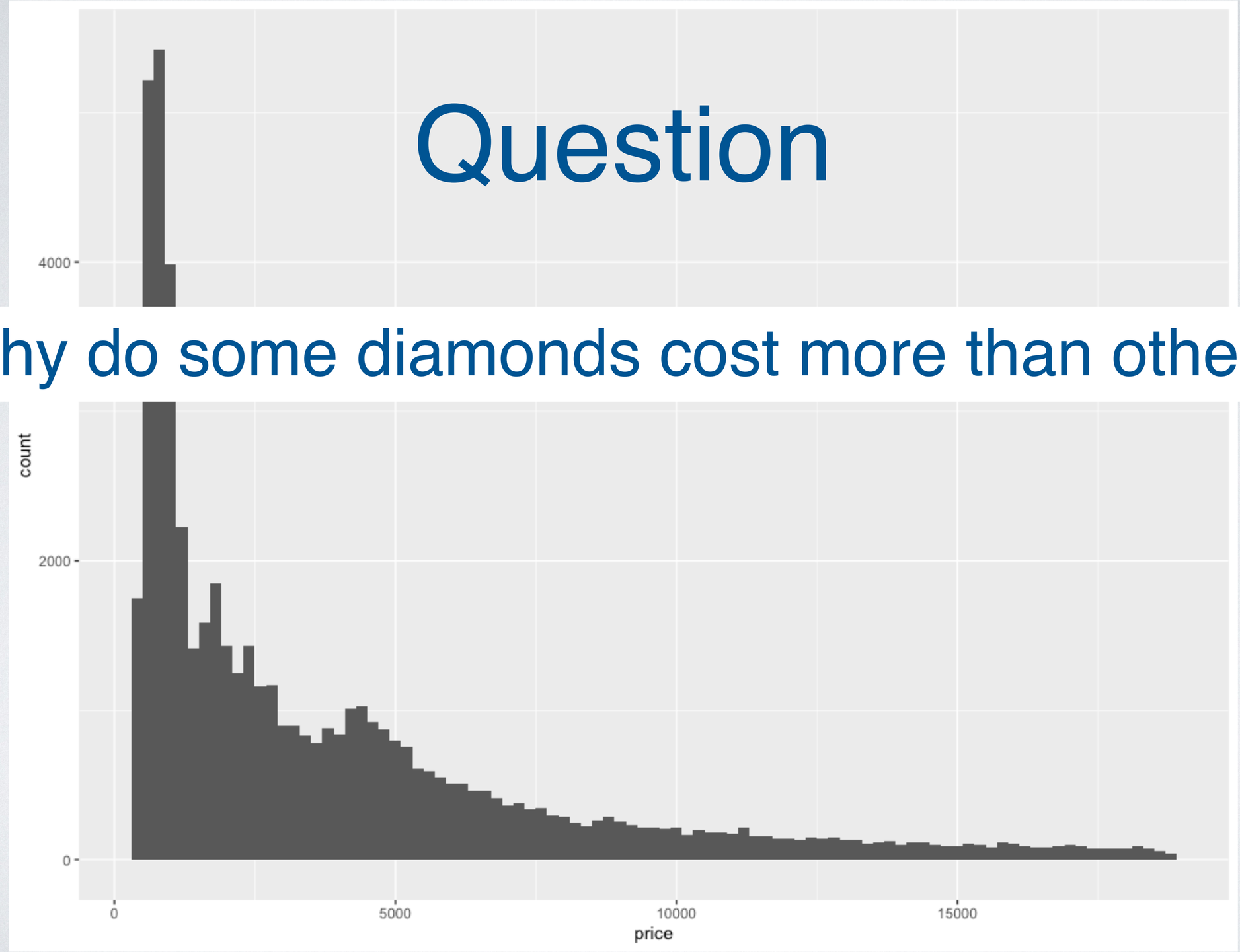




```
ggplot(data = diamonds) +
  geom_histogram(mapping = aes(x = carat), binwidth = 0.01)
```

Question

Why do some diamonds cost more than others?



Exploring Covariation

Photo by Andrew Ridley on Unsplash

$$F = MA$$

$$F = MA$$

Newton’s first law of motion : force = mass * acceleration

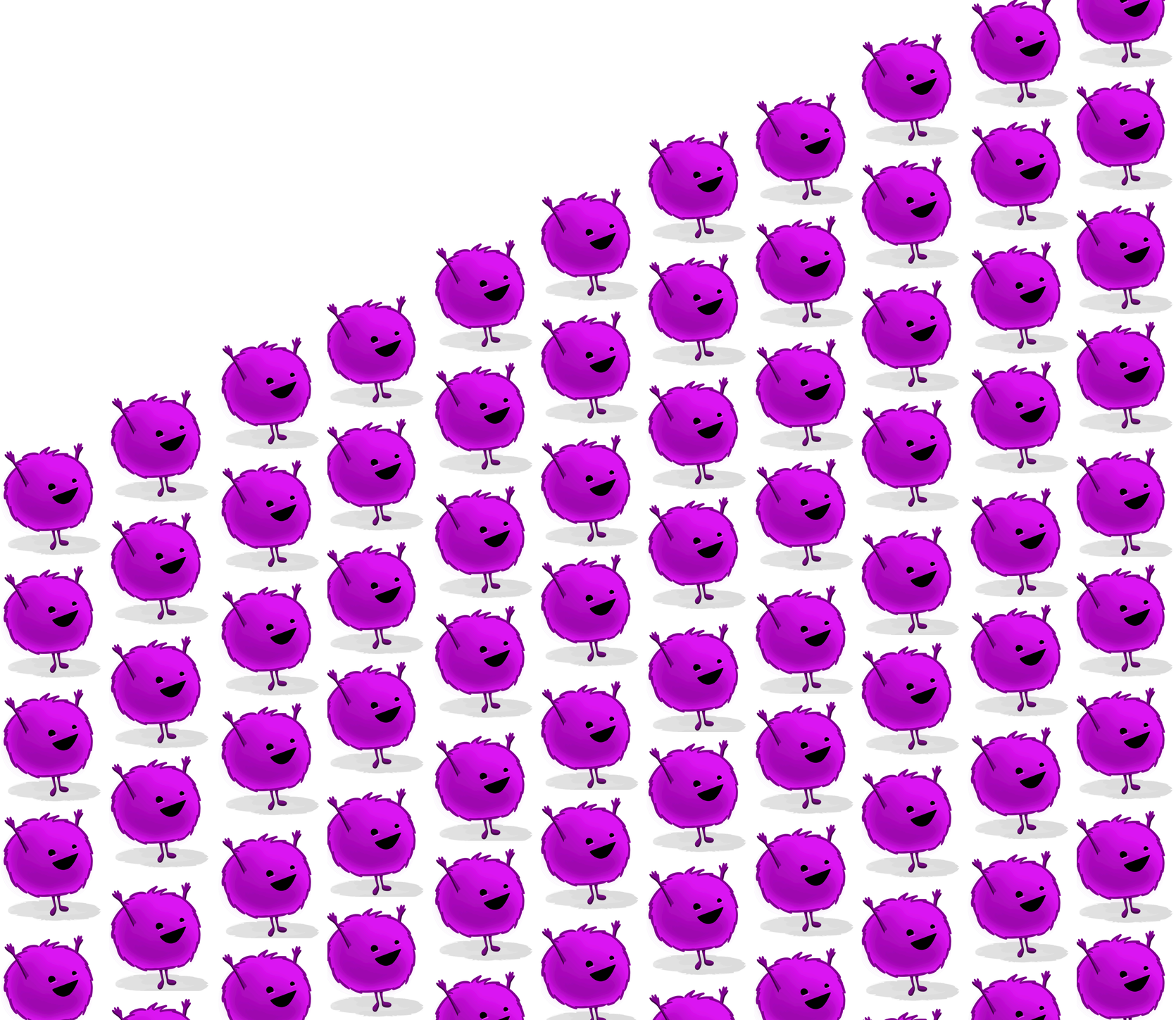
(Purple fuzzy monster)

Image credit : **Allison Horst**

https://allisonhorst.com/wickham-tidy-bakers

(Purple fuzzy monster

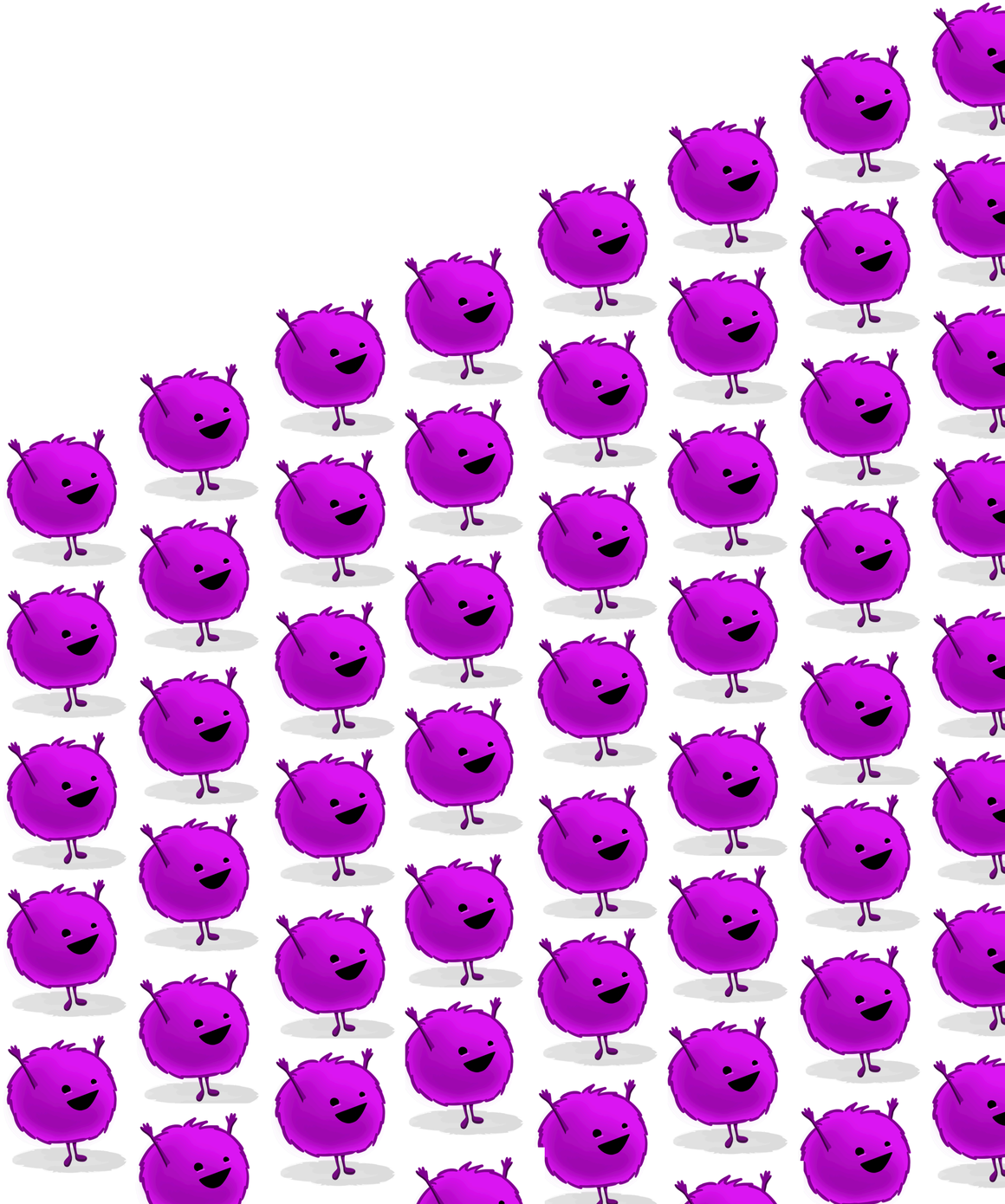
…with 99 identical twins)



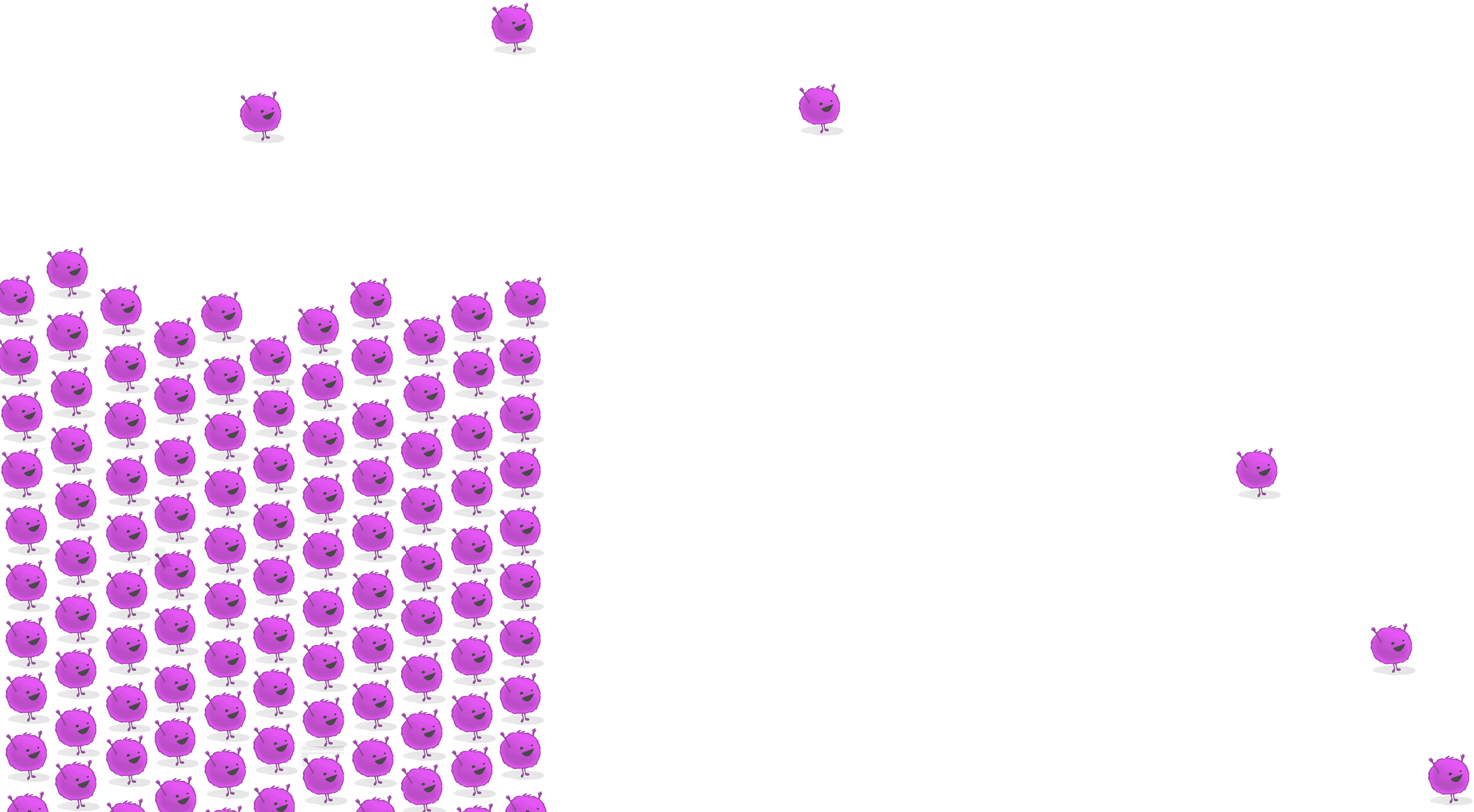
https://allisonhorst.com/wickham-tidy-bakers

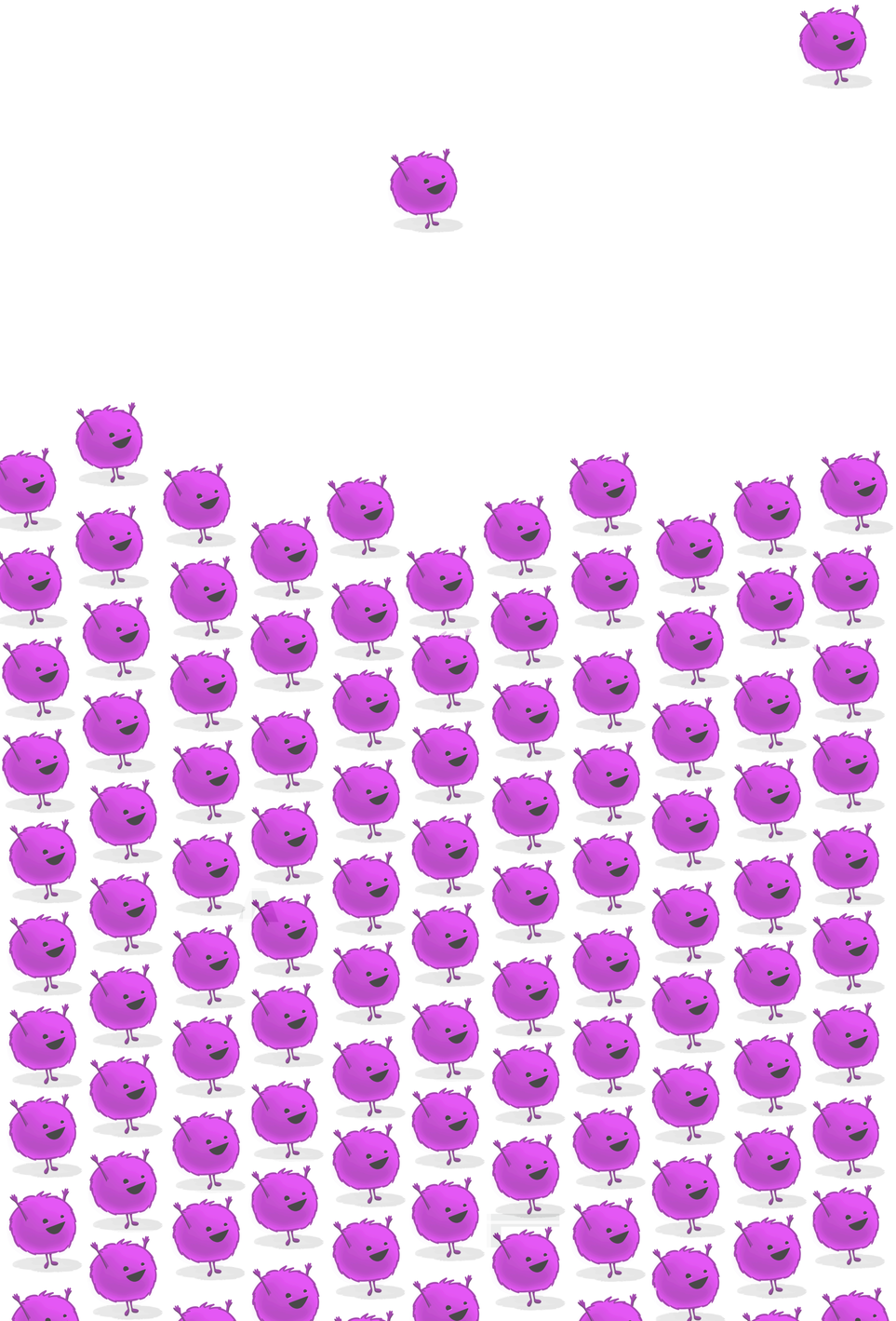
F = MA

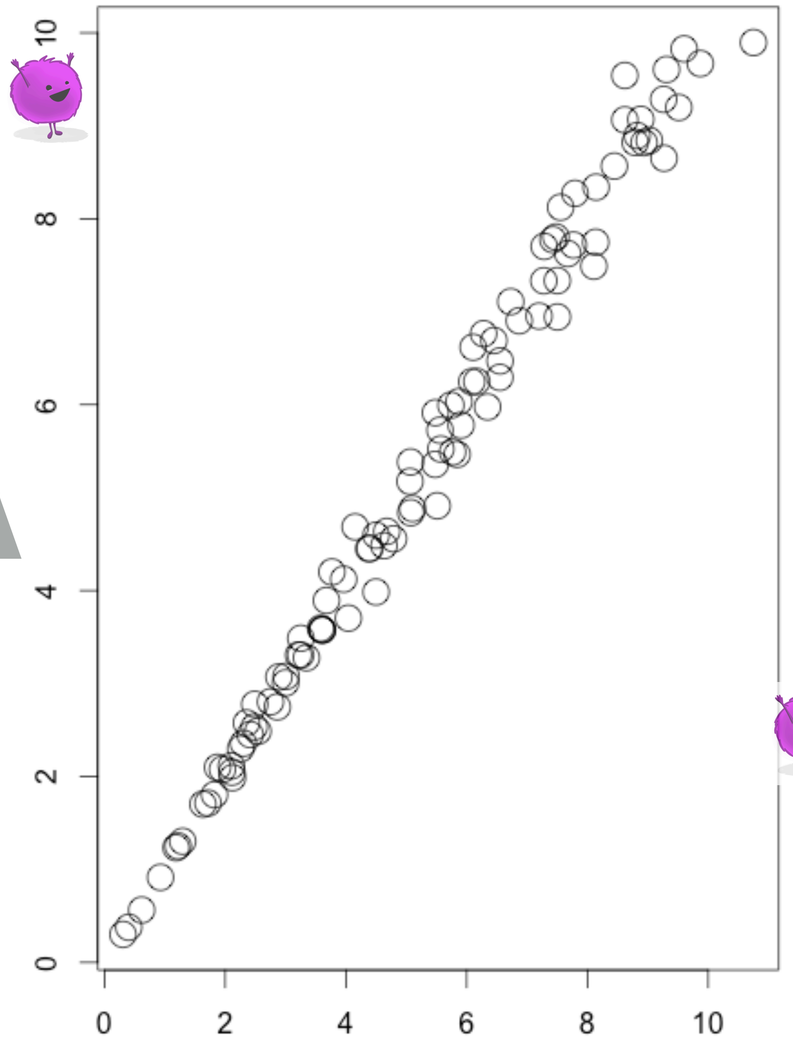
$$f_1 = m_1 * a_1$$
$$f_2 = m_2 * a_2$$
$$f_3 = m_3 * a_3$$
$$f_4 = m_4 * a_4$$
$$f_5 = m_5 * a_5$$
$$f_6 = m_6 * a_6$$
$$f_7 = m_7 * a_7$$
$$f_8 = m_8 * a_8$$
$$f_9 = m_9 * a_9$$



https://allisonhorst.com/wickham-tidy-bakers

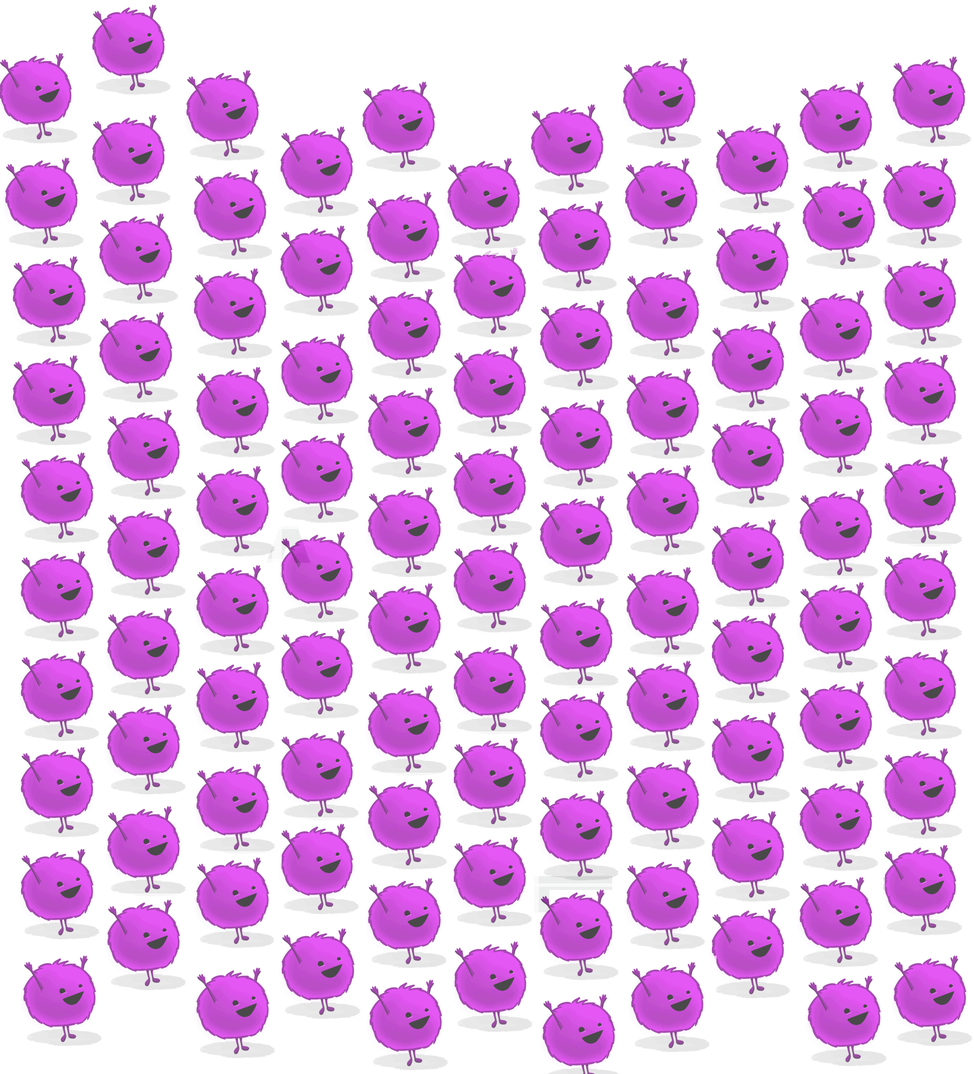




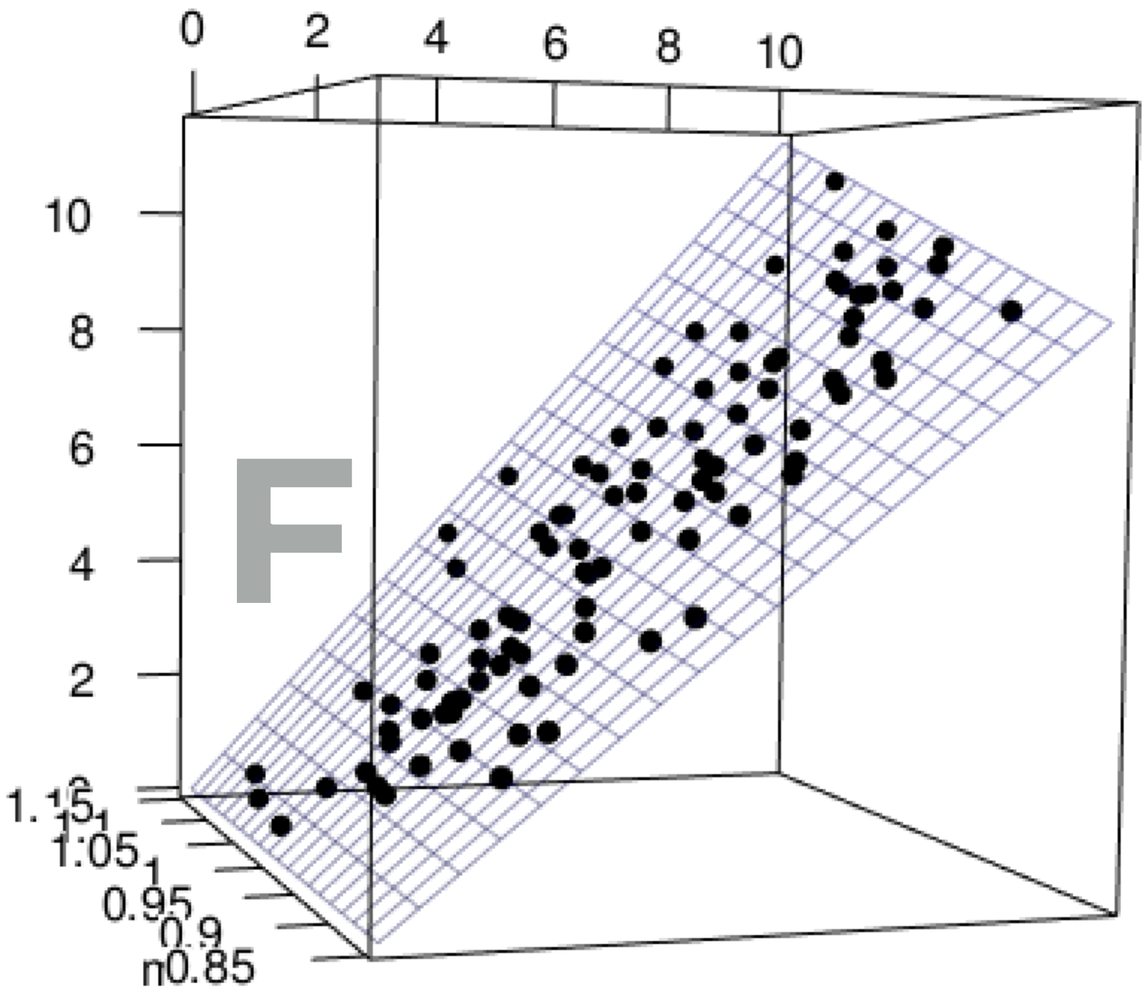


A

F



M



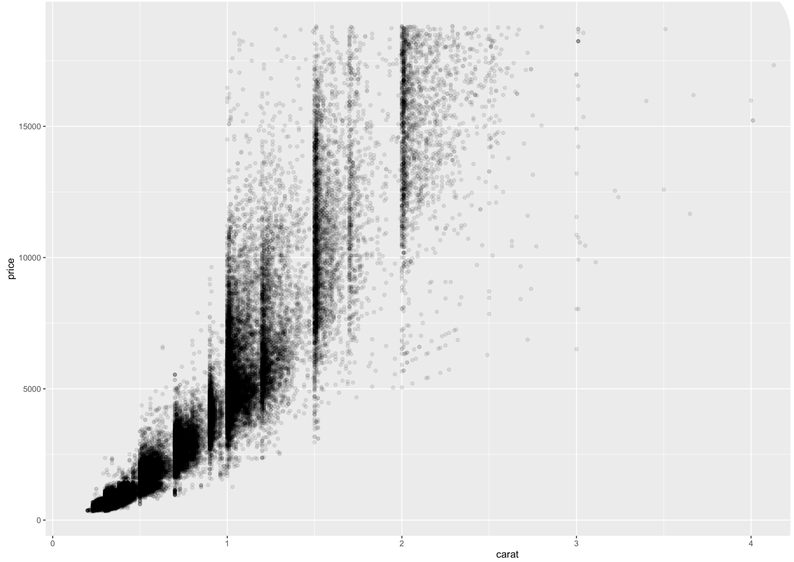
F

A

Continuous vs continuous

geom_point()

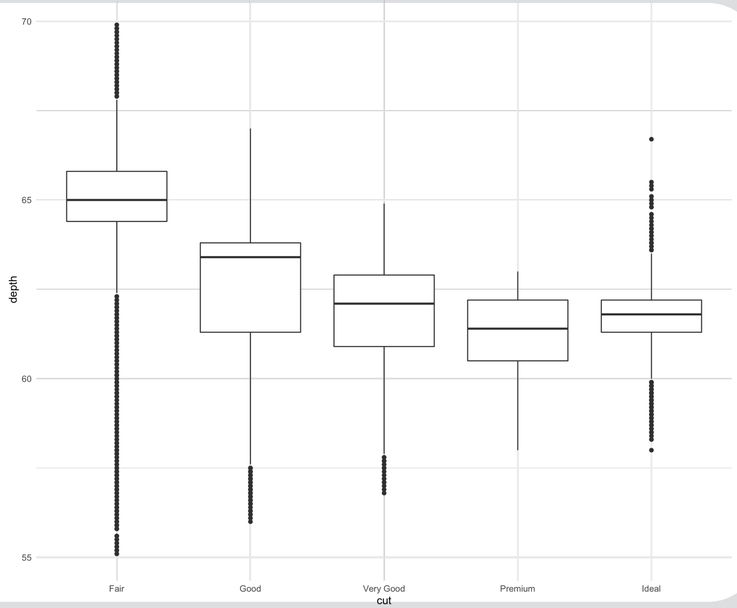
x = carat,
y = price



Continuous vs categorical

geom_boxplot()

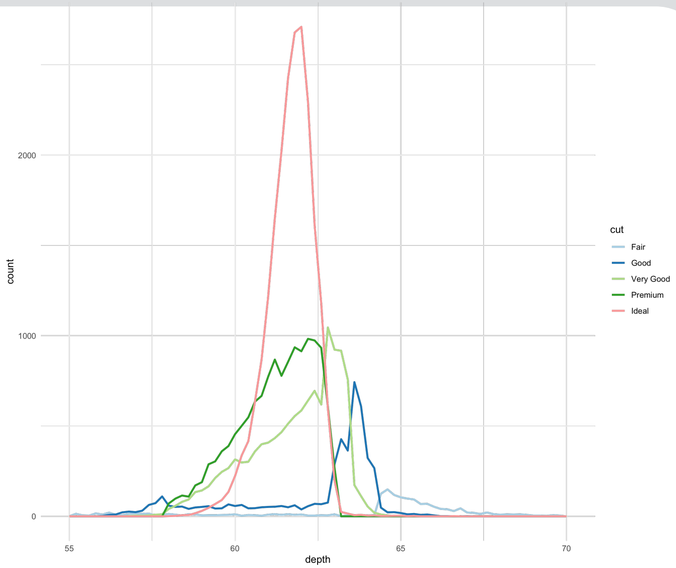
x = cut,
y = depth



Continuous vs categorical

geom_freqpoly()

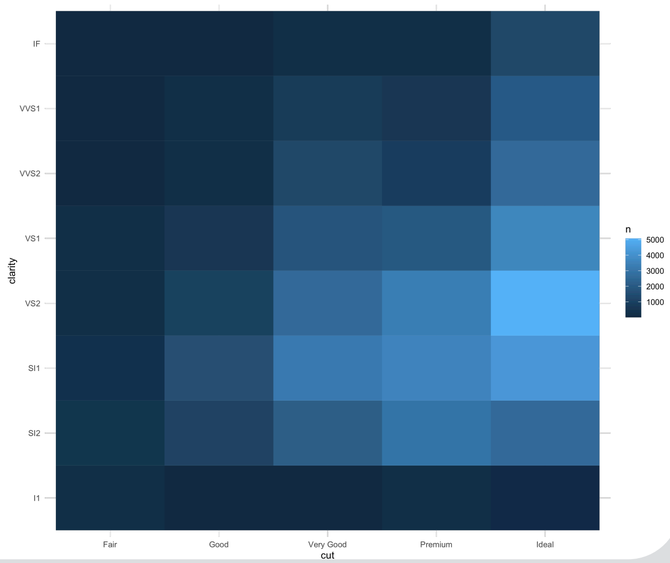
x = depth,
y = count



Categorical vs categorical

geom_tile()

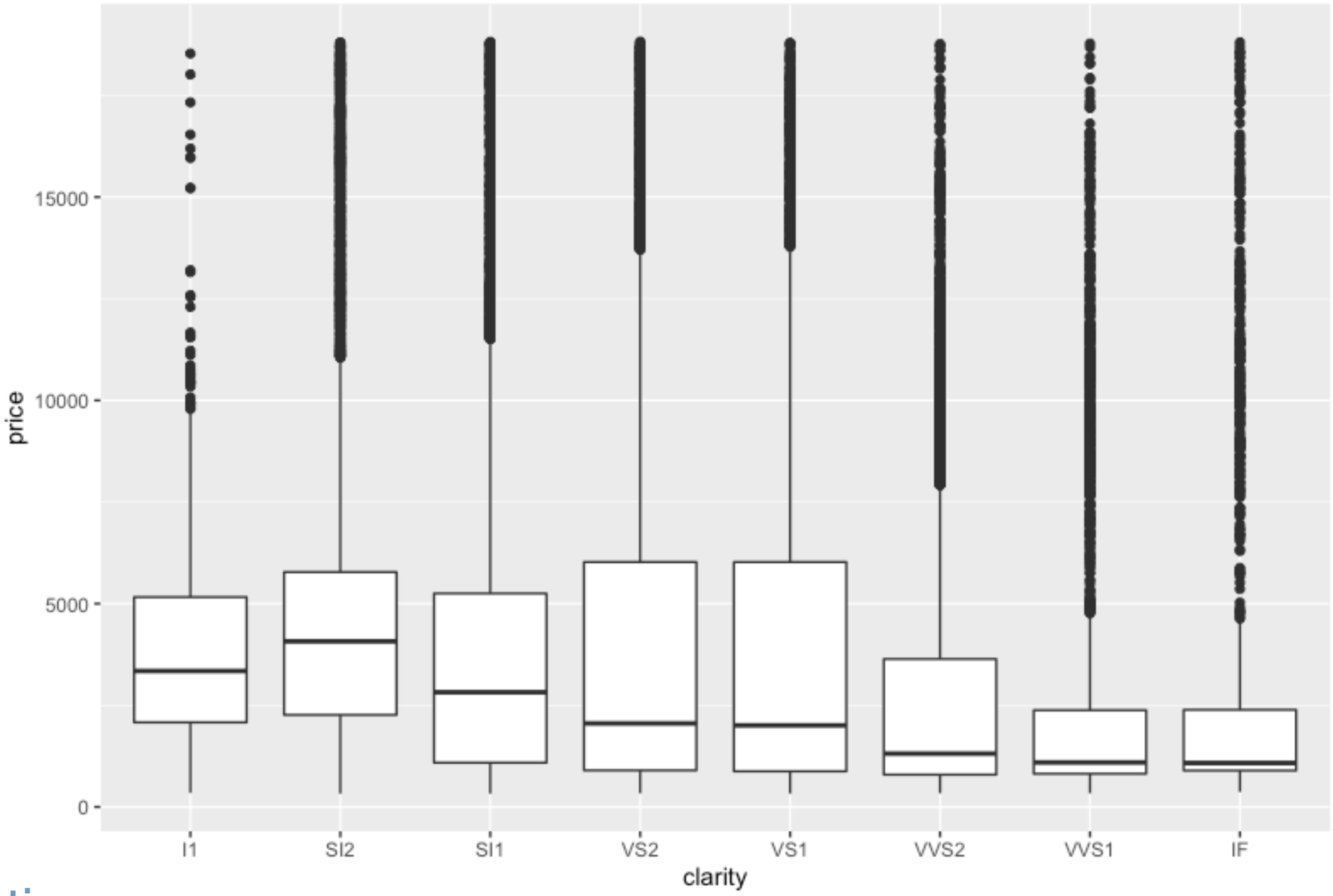
x = cut,
y = density



Your Turn 4

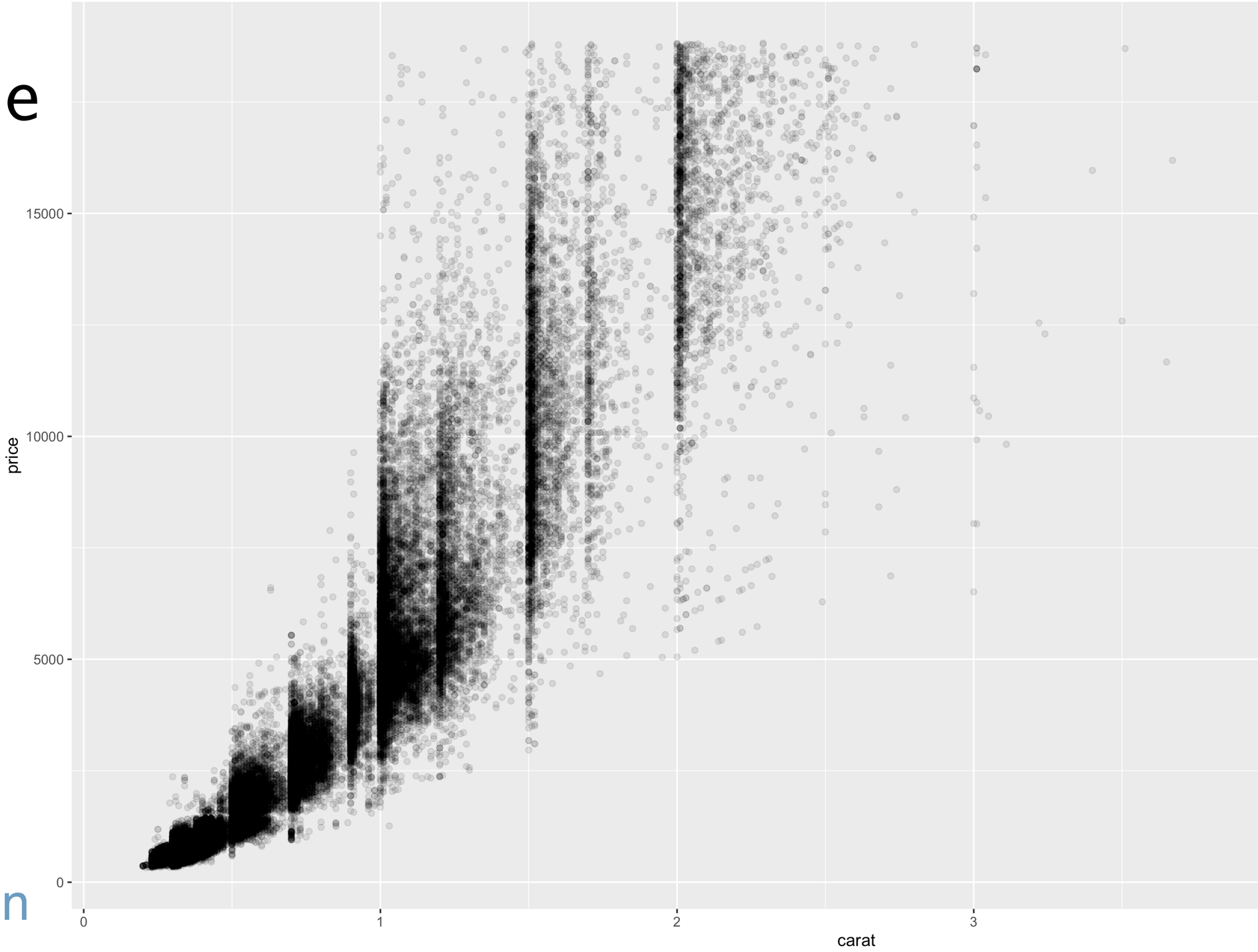
Visualize the relationships *between* different variables in `diamonds` and variable `price`. **What patterns do you notice?**

```
clarity ~
  price
```

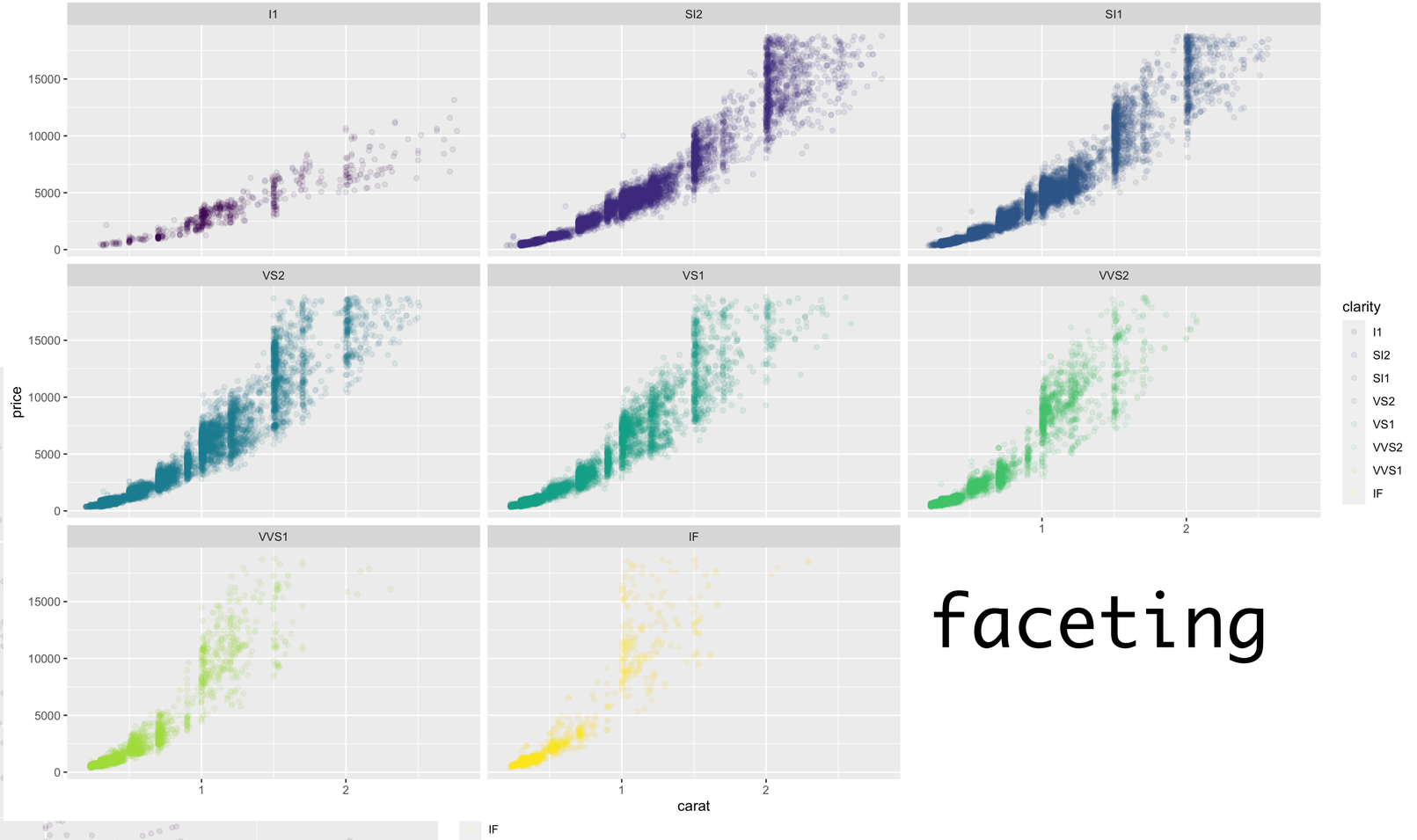


Exploring covariation

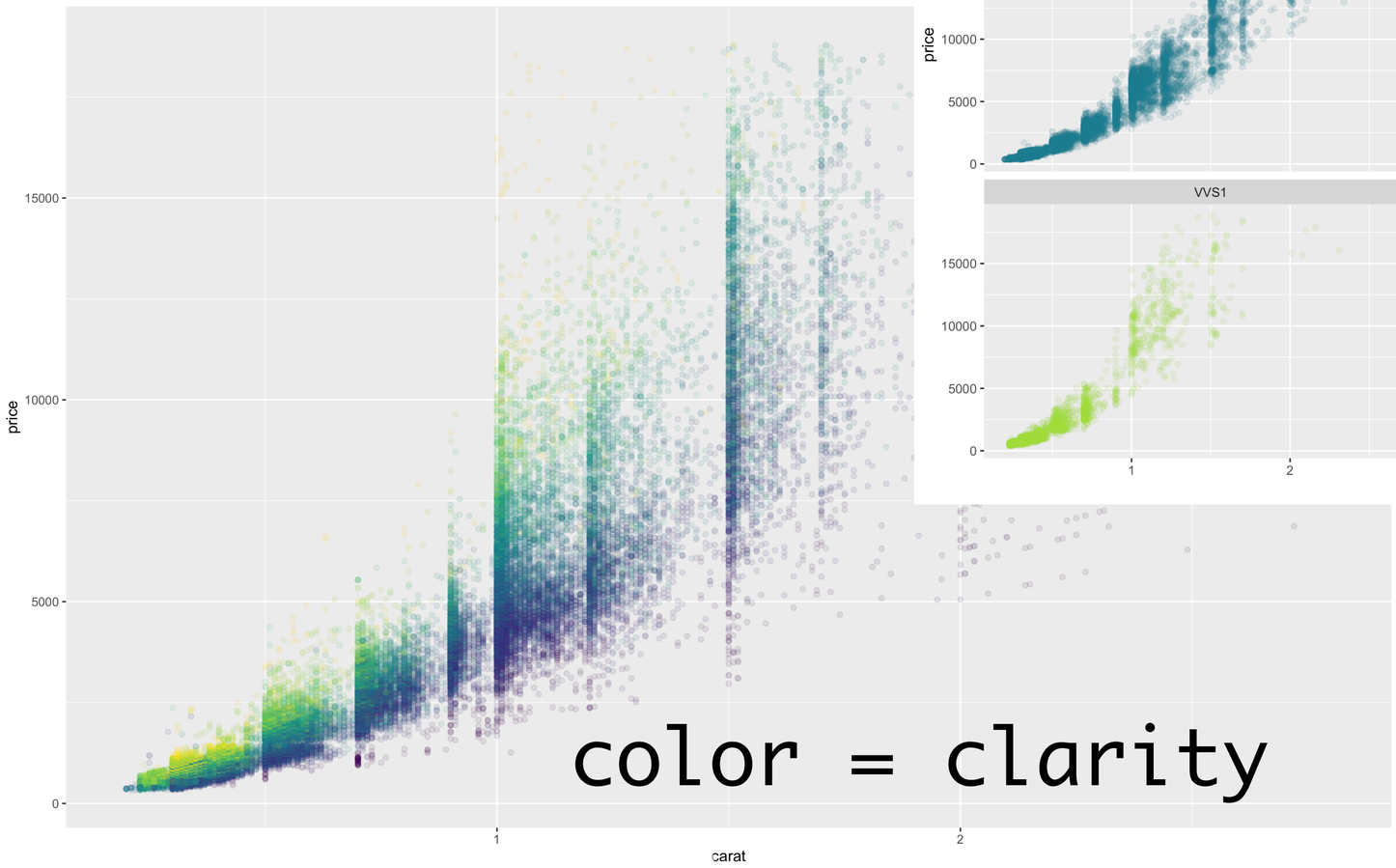
carat ~ price



Exploring covariation



faceting



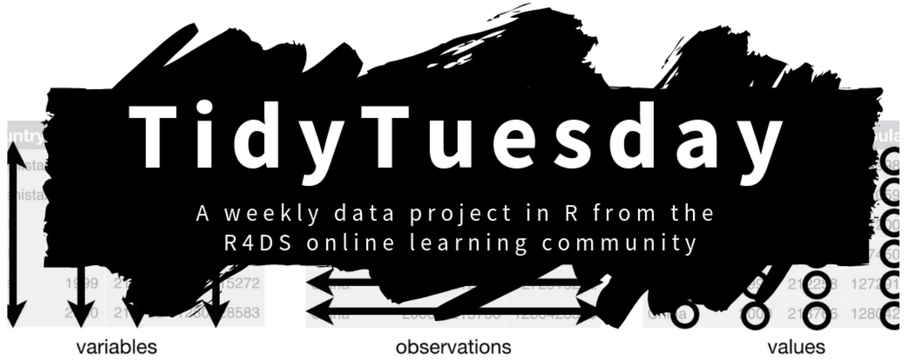
color = clarity

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

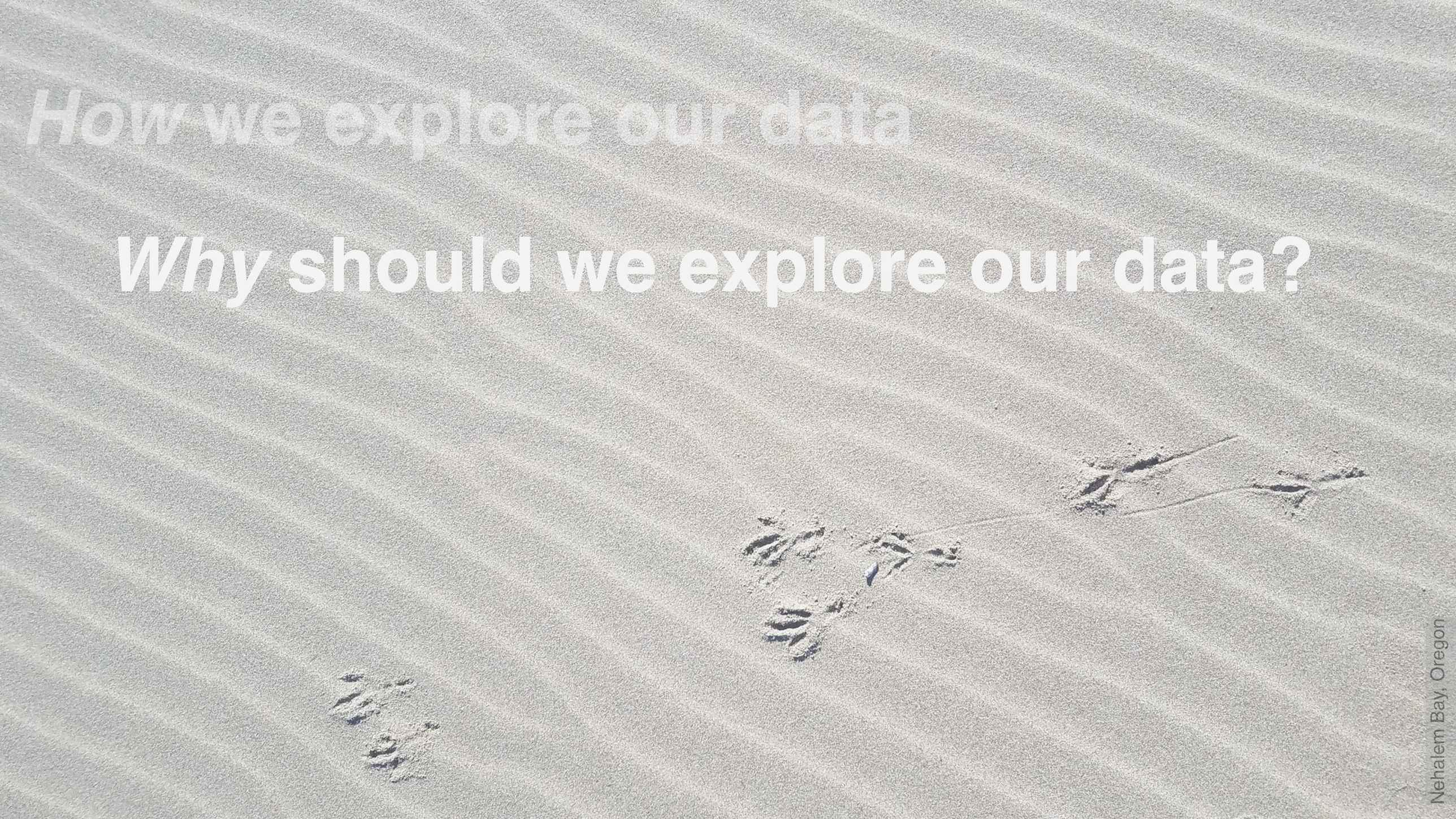
Nehalem Bay, Oregon

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 (12:30-1:30pm)

Kristin Bott

@kristin (Slack)

Nehalem Bay, Oregon

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.

Image: Flazingo Photos | flickr.com/photos/124247024@N07/14090440354

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

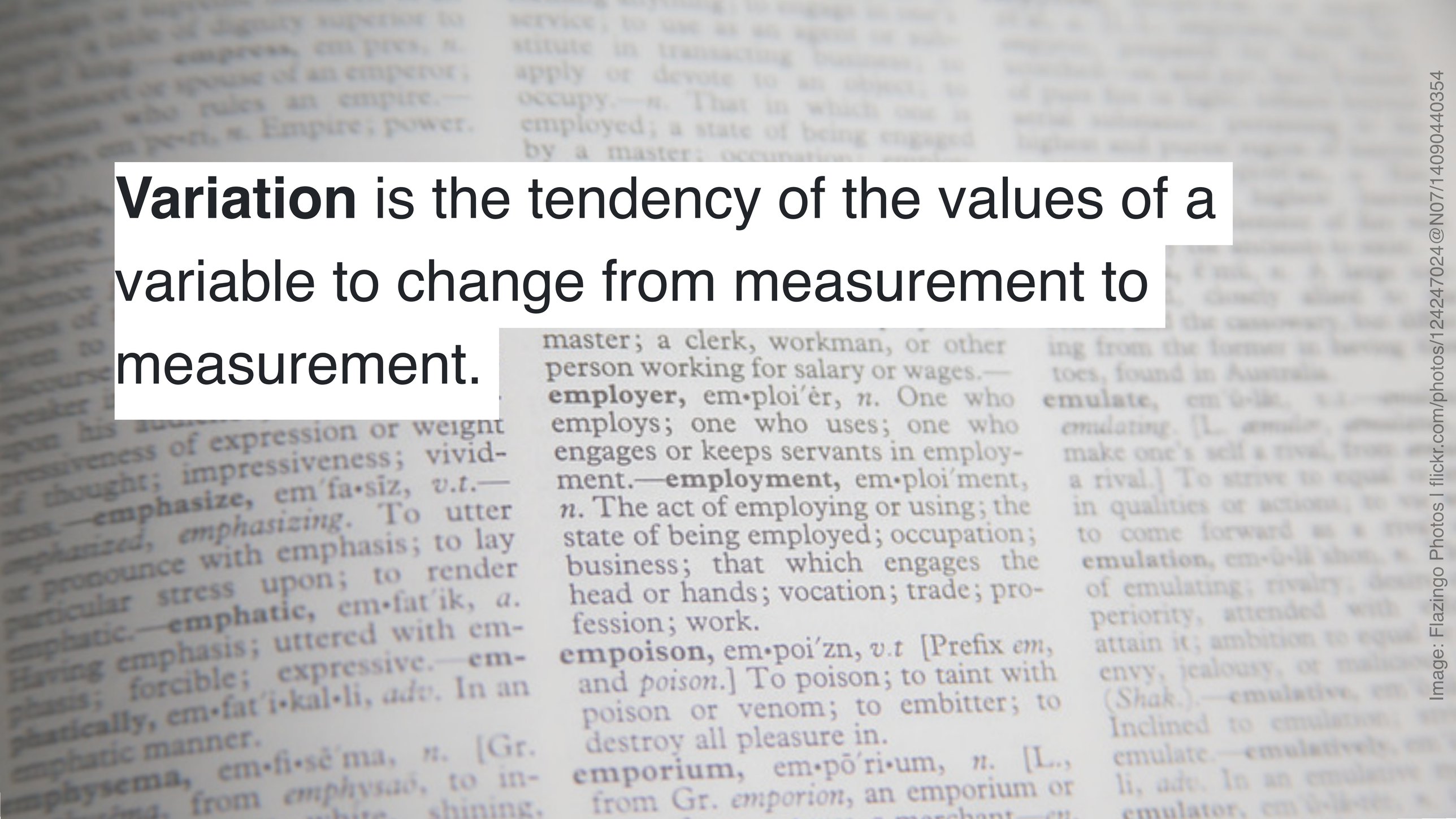


Image: Flazingo Photos | flickr.com/photos/124247024@N07/14090440354

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.

Image: Flazingo Photos | flickr.com/photos/124247024@N07/14090440354

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

Image: Flazingo Photos | flickr.com/photos/124247024@N07/14090440354