Visualization (Exploring Co-variation)

Peter Ganong and Maggie Shi

January 27, 2026

Introduction

Skills hopefully acquired at the end of lecture

Take a two variables in a dataset. Visualize to learn more about how they co-vary.

Key cases of interest:

- Categorical variable and a continuous variable
- Two categorical variables
- Two continuous variables

Categorical variable and continuous variable

Categorical vs. continuous: roadmap

- `penguins` dataset
- Boxplots
- Densities
- Small multiples

penguins dataset

```python
url = ("https://raw.githubusercontent.com/mcnakhaee/palmerpenguins/master/p
penguins = pd.read_csv(url)
penguins.head()
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	boc
0	Adelie	Torgersen	39.1	18.7	181.0	375
1	Adelie	Torgersen	39.5	17.4	186.0	380
2	Adelie	Torgersen	40.3	18.0	195.0	325
3	Adelie	Torgersen	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	345

`penguins` dataset

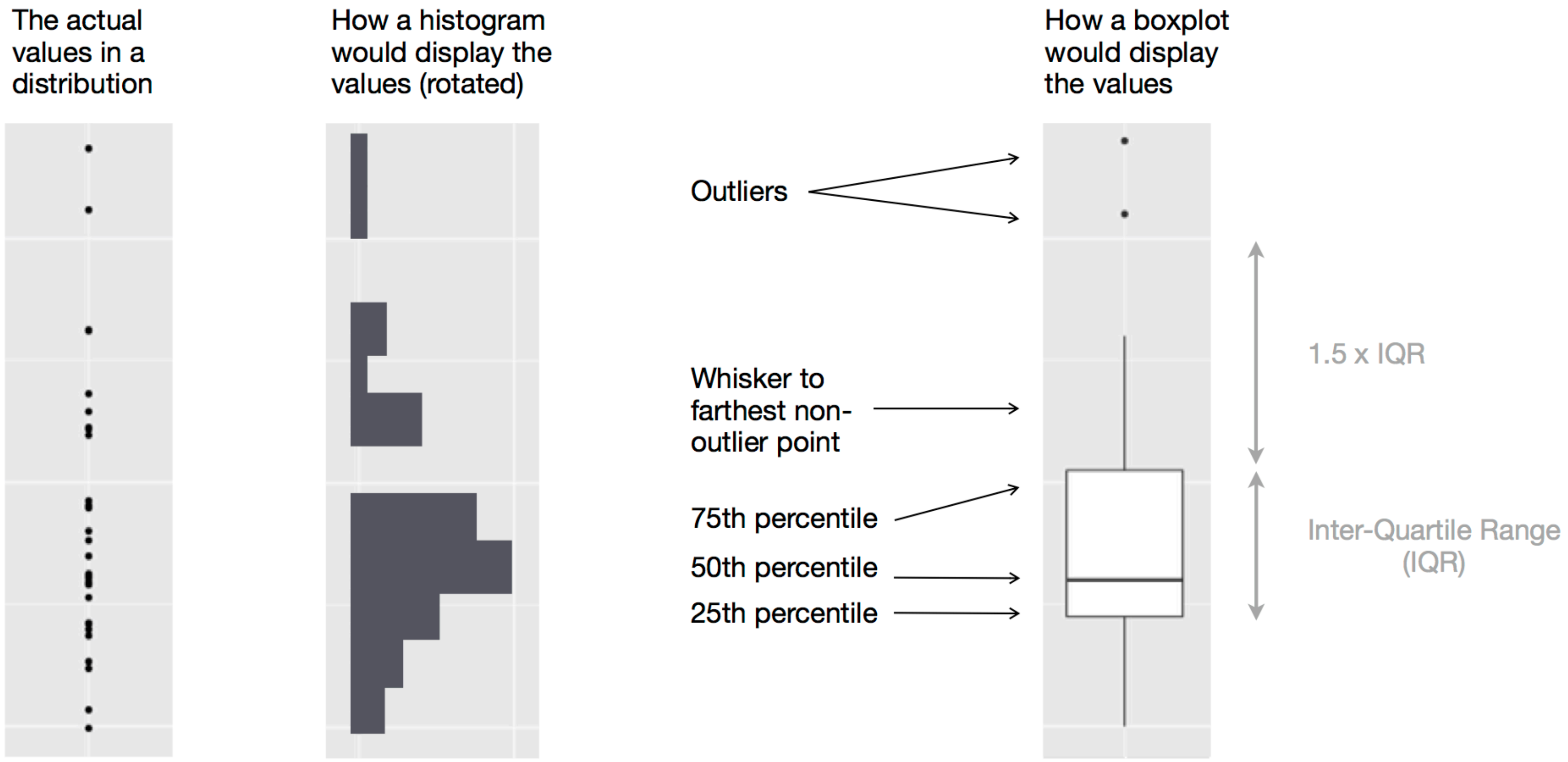
`species` appears to be a categorical variable

```python
penguins['species'].value_counts()
```

```
species
Adelie       152
Gentoo       124
Chinstrap     68
Name: count, dtype: int64
```

Discussion question: is it a Nominal or Ordinal variable?

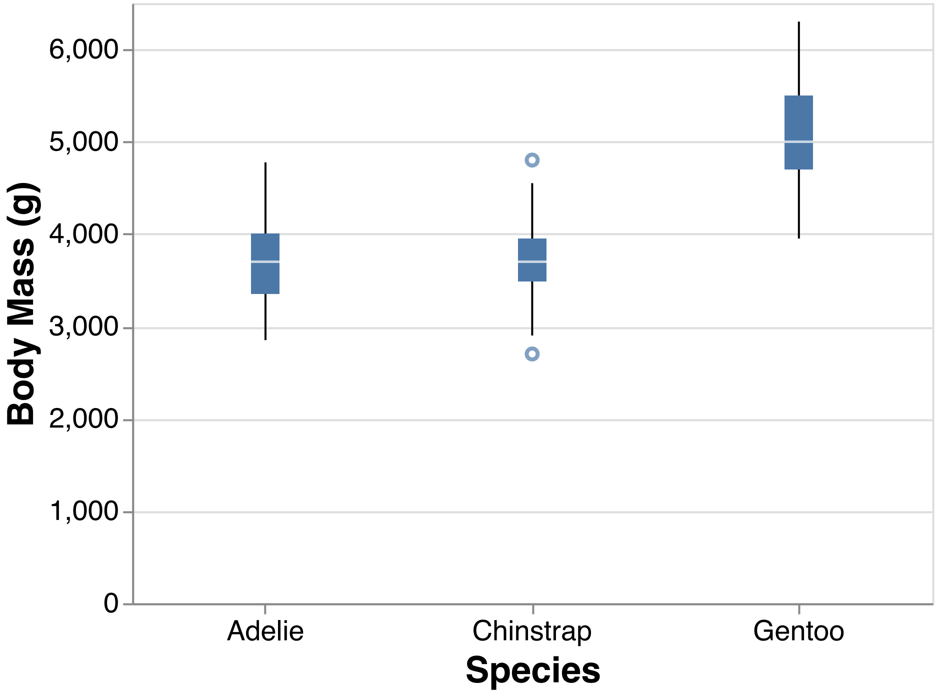
Categorical & continuous: box plot



The actual values in a distribution
How a histogram would display the values (rotated)
How a boxplot would display the values
Outliers
Whisker to farthest non-outlier point
75th percentile
50th percentile
25th percentile
1.5 x IQR
Inter-Quartile Range (IQR)

mark_boxplot()

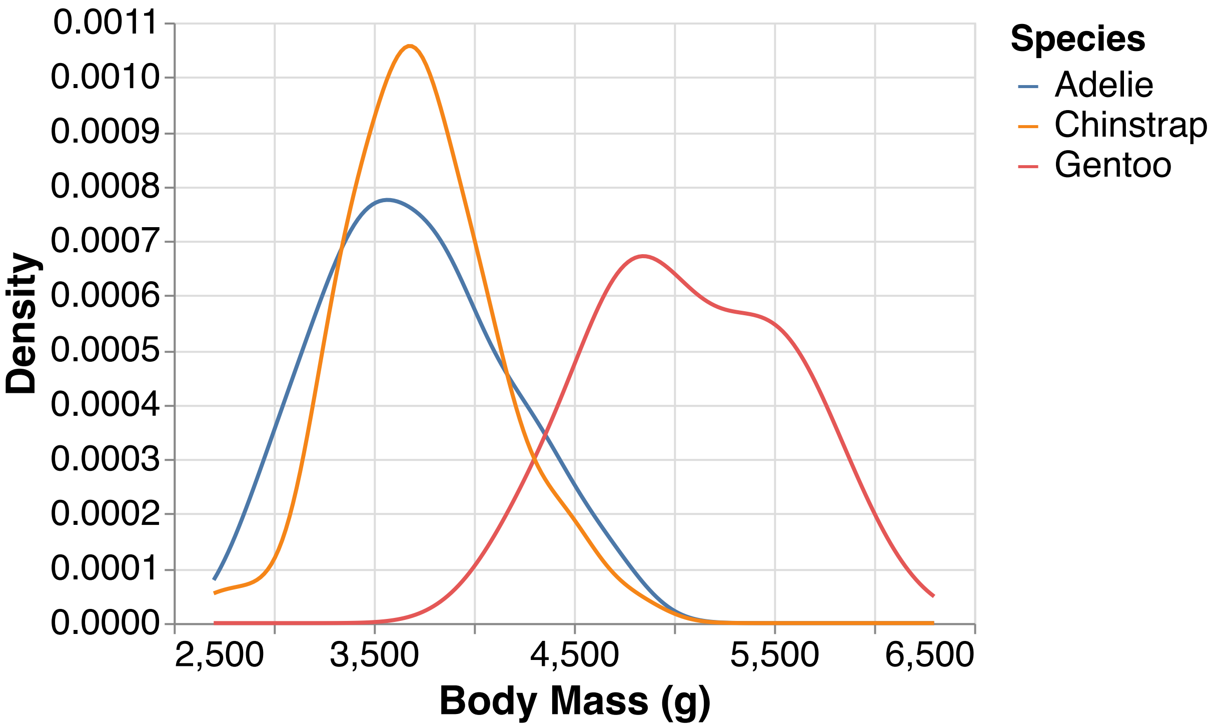
```
alt.Chart(penguins).mark_boxplot().encode(
    alt.X('species:N', title="Species"),
    alt.Y('body_mass_g:Q', title="Body Mass (g)"),
)
```



Discussion question: what is the headline message from this graph? Submessages?

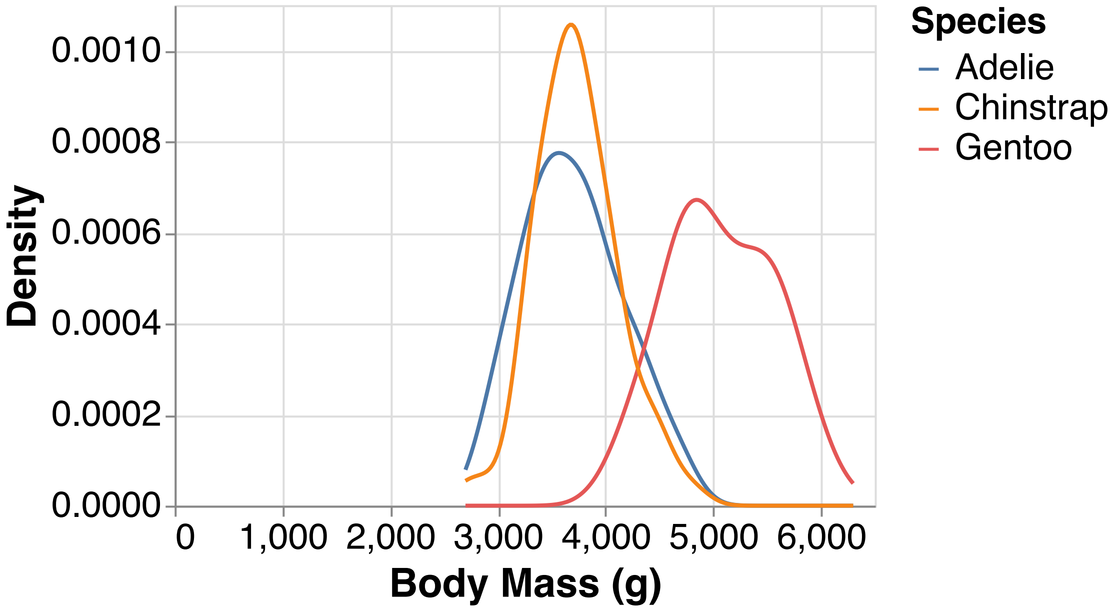
transform_density()

```python
alt.Chart(penguins).transform_density(
    'body_mass_g',
        groupby=['species'],
        as_=['body_mass_g2', 'density']
    ).mark_line().encode(
        alt.X('body_mass_g2:Q', title = "Body Mass (g)"),
        alt.Y('density:Q', title = "Density"),
        alt.Color('species:N', title = "Species")
    )
```


Species
Adelie
Chinstrap
Gentoo
Density
0.0011
0.0010
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
2,500
3,500
4,500
5,500
6,500
Body Mass (g)

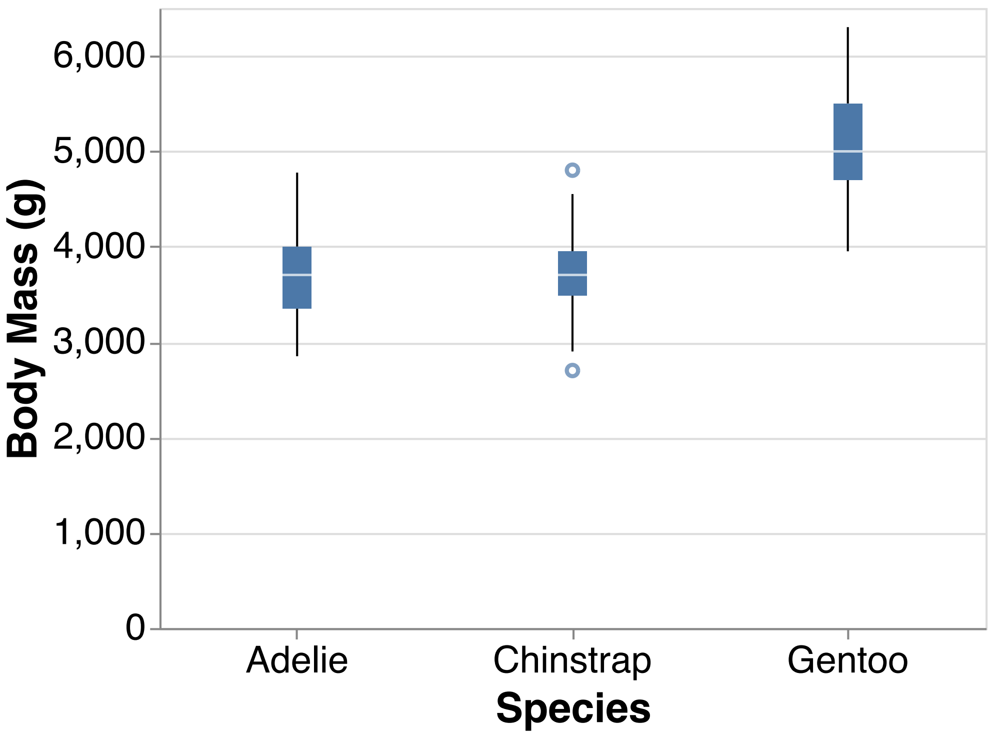
`transform_density()`, scale to 0

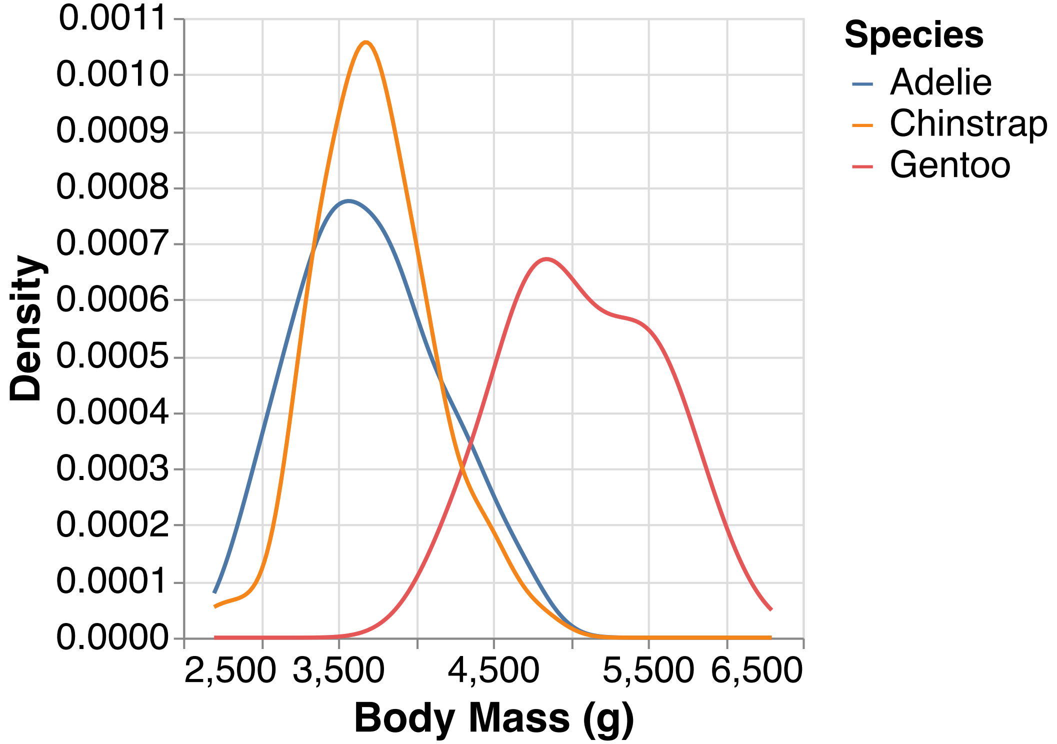
```python
alt.Chart(penguins).transform_density(
    'body_mass_g',
        groupby=['species'],
        as_=['body_mass_g', 'density']
    ).mark_line().encode(
        alt.X('body_mass_g:Q', scale=alt.Scale(zero=True), title = "Body Mass (g)"),
        alt.Y('density:Q', title = "Density"),
        alt.Color('species:N', title = "Species")
    )
```



Discussion question: what if we required the x-axis range to include zero? Would that improve or reduce clarity? Why?

Boxplot or density plots?

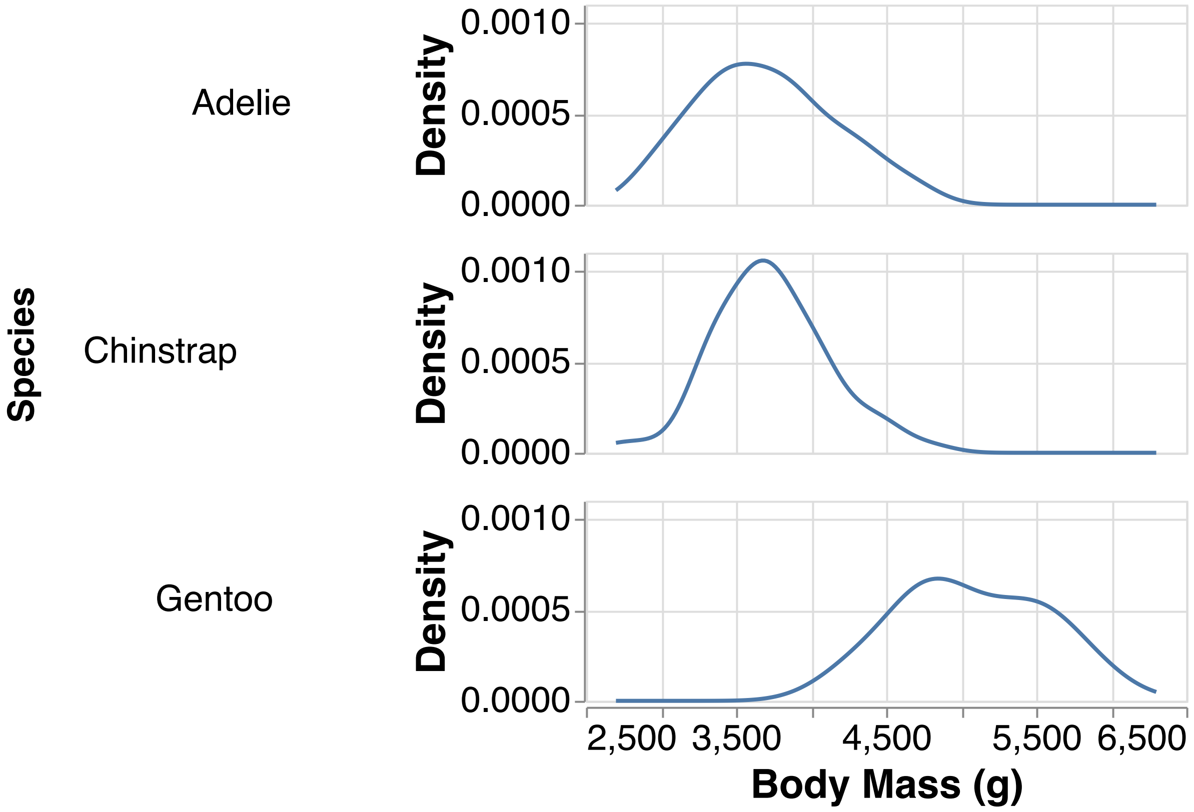




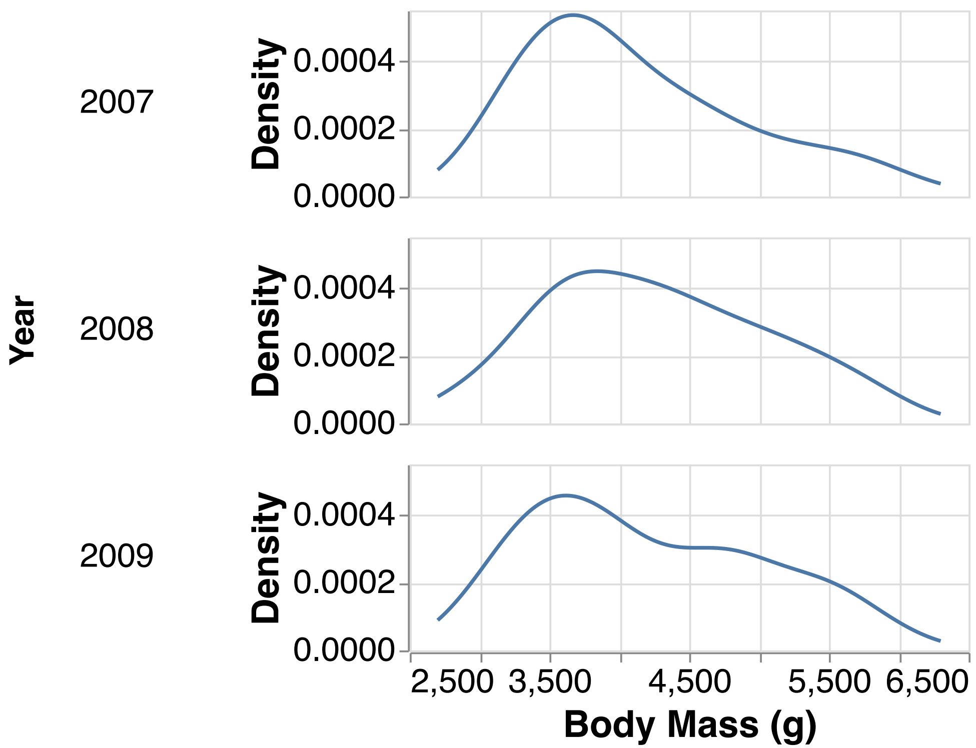
Discussion question: what messages come through more with the box plot? Through the density plot?

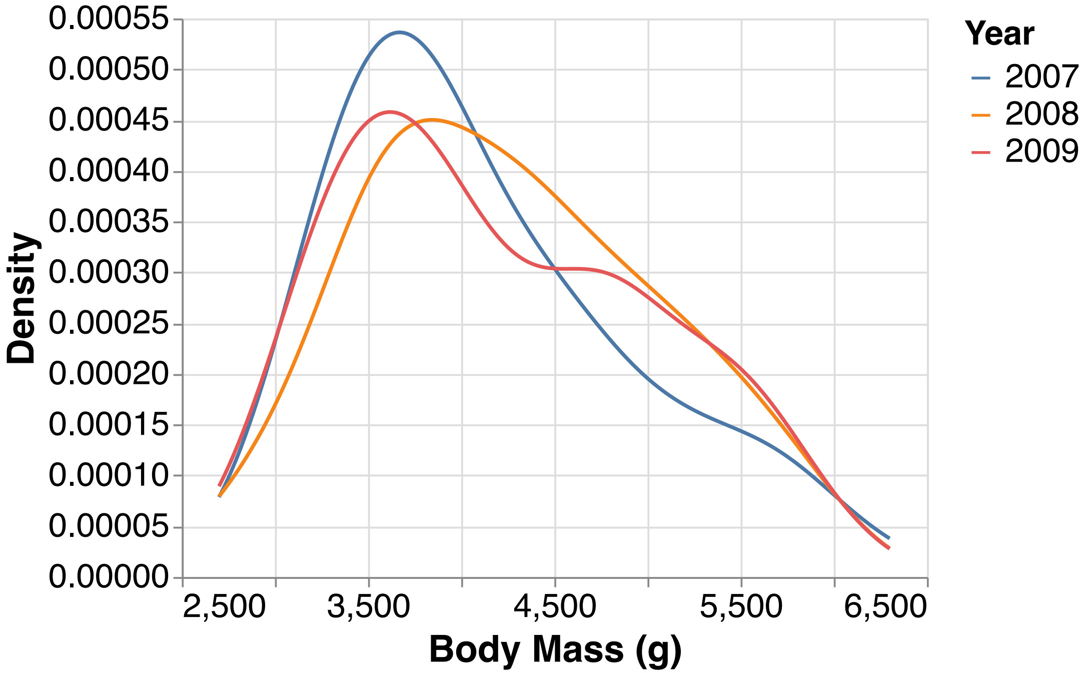
alt.Row: small multiples

```
alt.Chart(penguins).transform_density(
    'body_mass_g',
    groupby=['species'],
    as_=['body_mass_g', 'density']
).mark_line().encode(
    alt.X('body_mass_g:Q', title = "Body Mass (g)"),
    alt.Y('density:Q', title = "Density"),
    alt.Row('species:N', header=alt.Header(labelAngle=0), title = "Species")
)
```



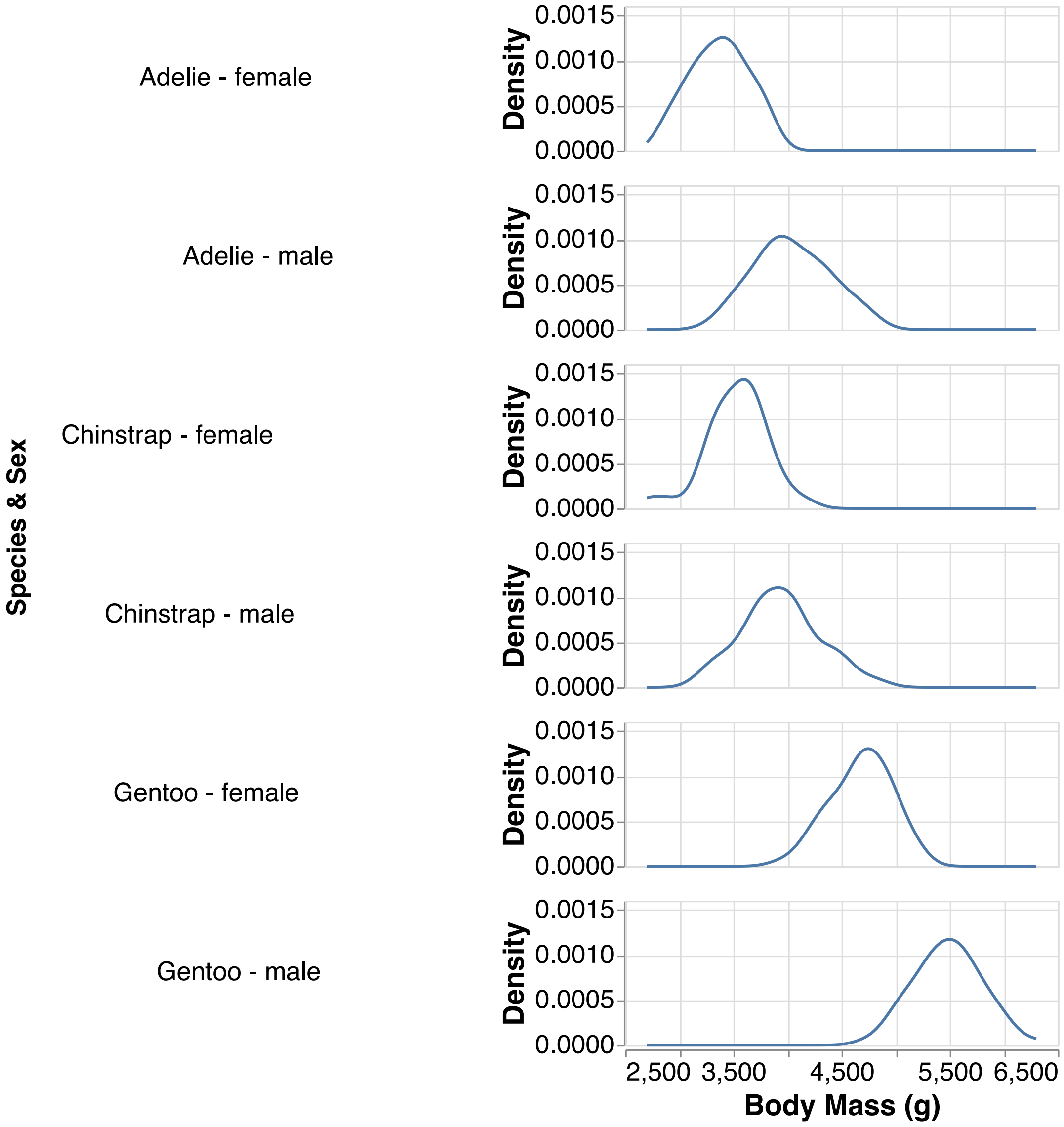
By year: colors or small multiples?

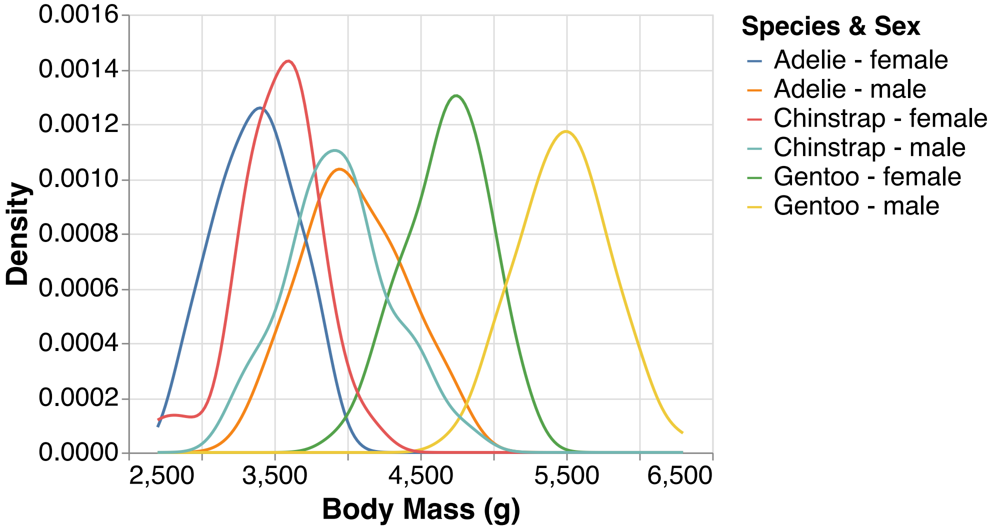




Discussion question: these two graphs show identical information. Which do you prefer, and why?

Colors or small multiples?





Two Categorical Variables

Two categorical variables: roadmap

- Two ways to encode frequency as a third dimension: `diamonds`
    - `size`
    - `color`
- A word of caution against 3D graphs

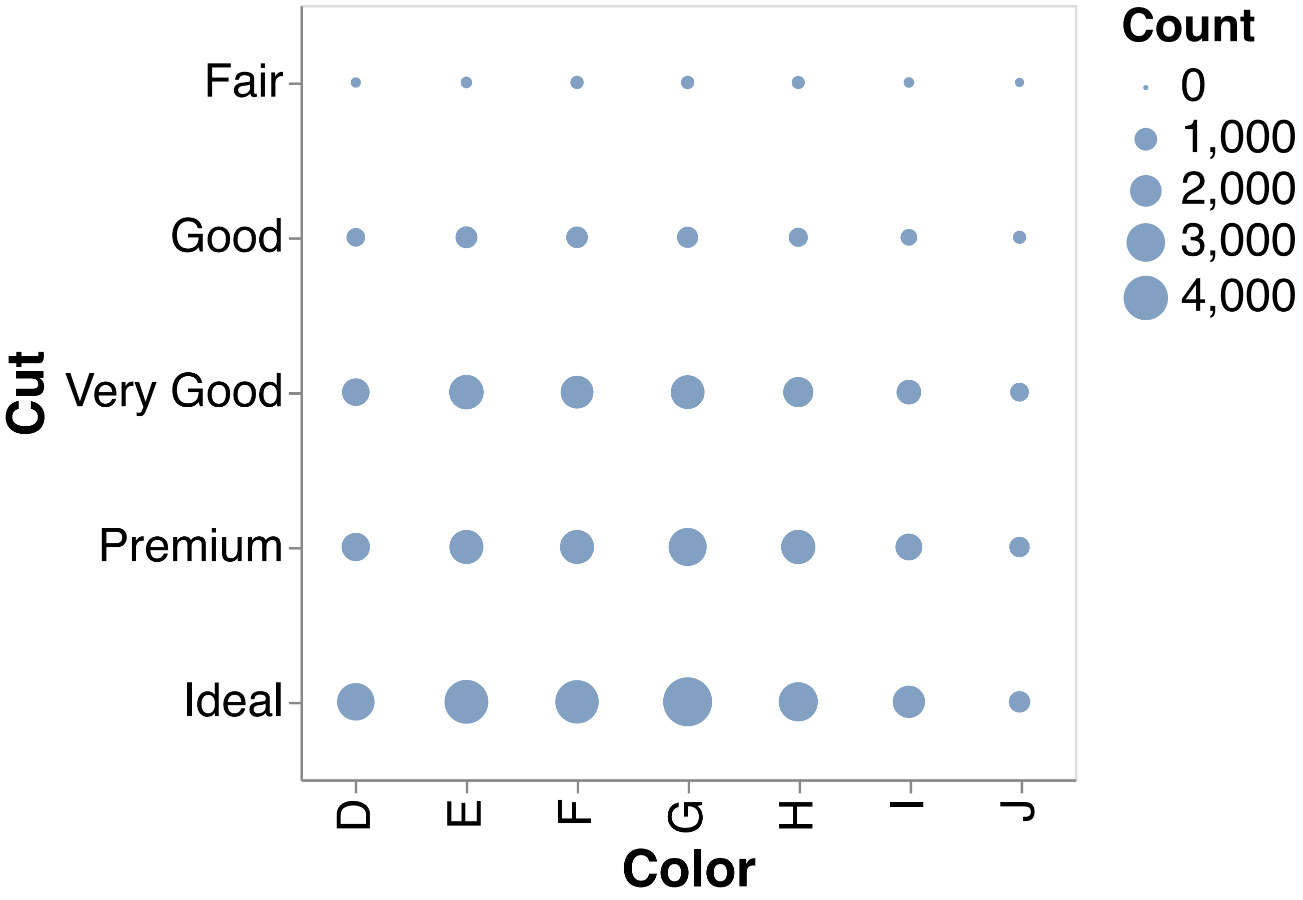
How is cut related to color? `Size`

In `diamonds` dataset, `color` and `cut` are both categorical

```python
diamonds_grouped = diamonds.groupby(['color','cut']).size().reset_index().rename(columns={0:'N'})
diamonds_grouped
```

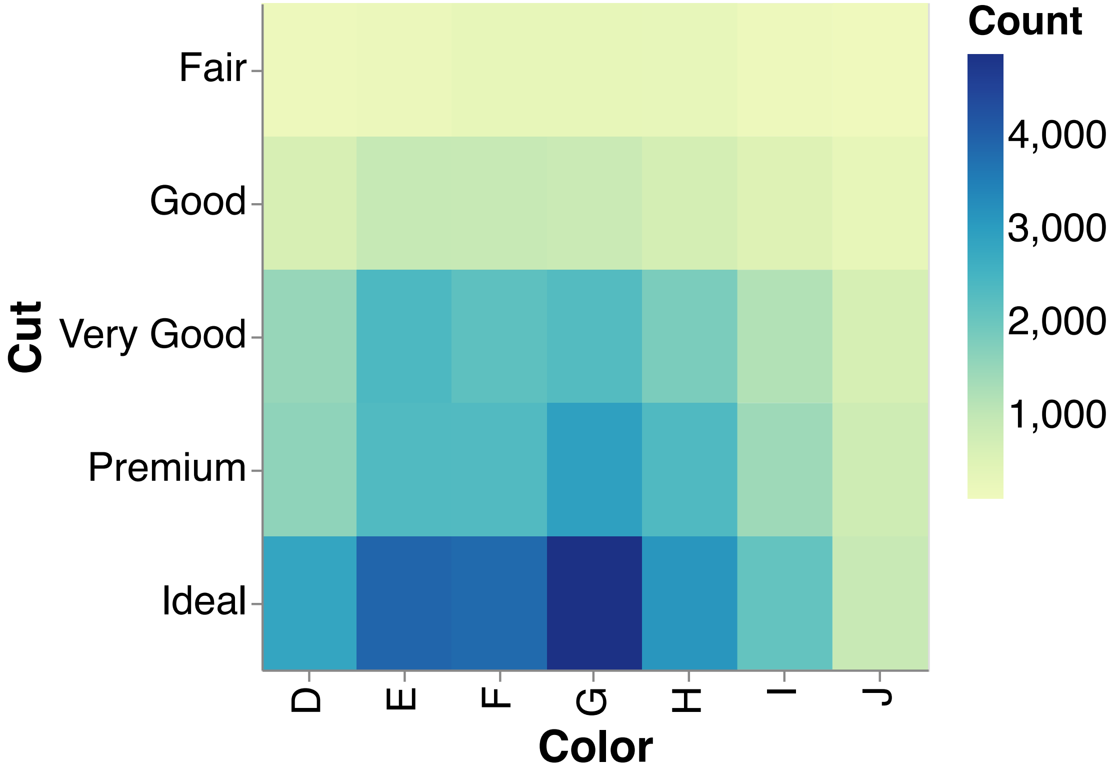
How is cut related to color? `Color`

```python
alt.Chart(diamonds_grouped).mark_circle().encode(
    alt.X('color:O', title = "Color"),
    alt.Y('cut:O', title = "Cut", sort = cut_order),
    alt.Size('N:Q', title = "Count"))
```



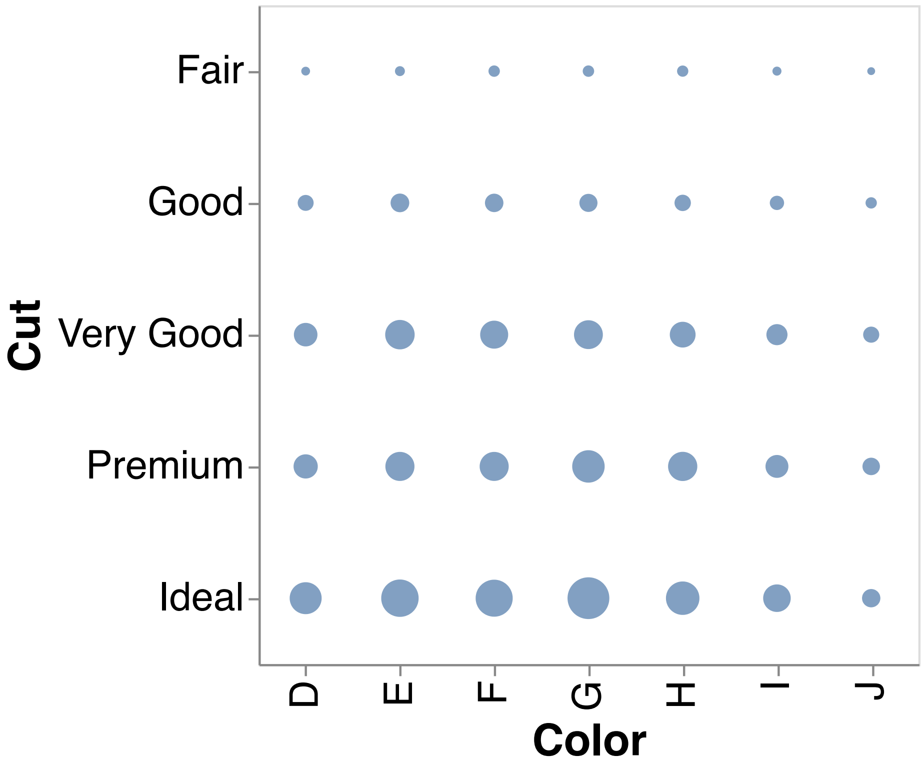
How is cut related to color?

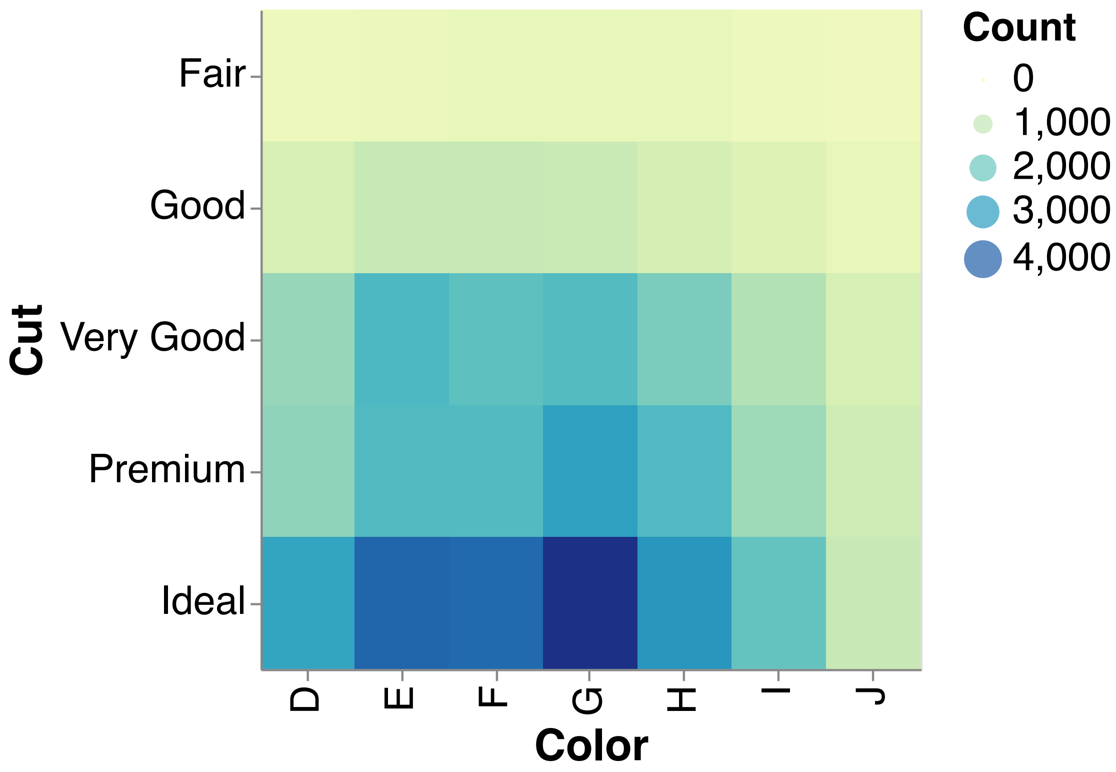
```
alt.Chart(diamonds_grouped).mark_rect().encode(
    alt.X('color:O', title = "Color"),
    alt.Y('cut:O', title = "Cut", sort = cut_order),
    alt.Color('N:Q', title = "Color"))
```



Discussion question: what diamond types are most common?

How is cut related to color?

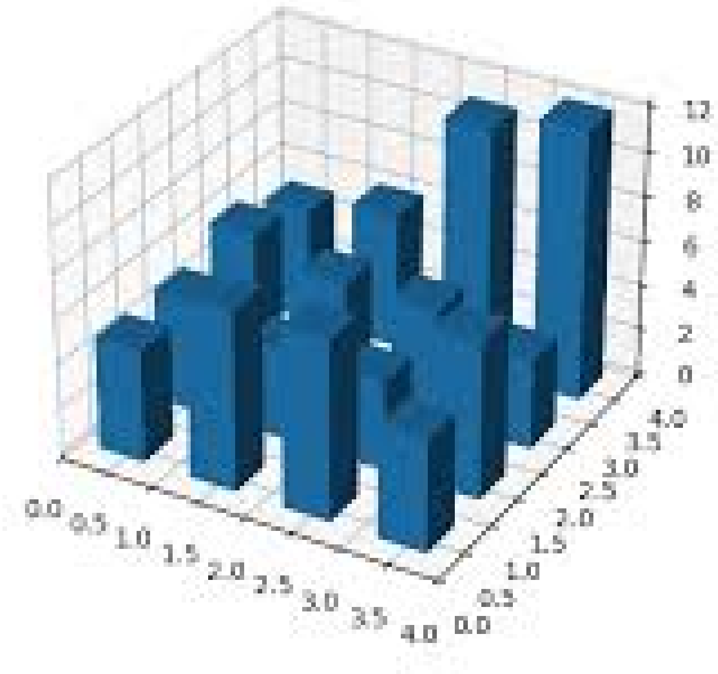




Discussion question: these two plots display the same information, but encoded differently. Which do you prefer?

A word of caution: 3D graphs

You may have seen covariation between two variables depicted as a 3D plot before



- 3D graphs are almost always *not recommended* – they distort perception and cannot accurately represent scale
- `altair` does not create 3D graphs – for good reason!

Two Categorical Variables: summary

- Encode frequency as `color` or `size`
- Avoid 3D representations!

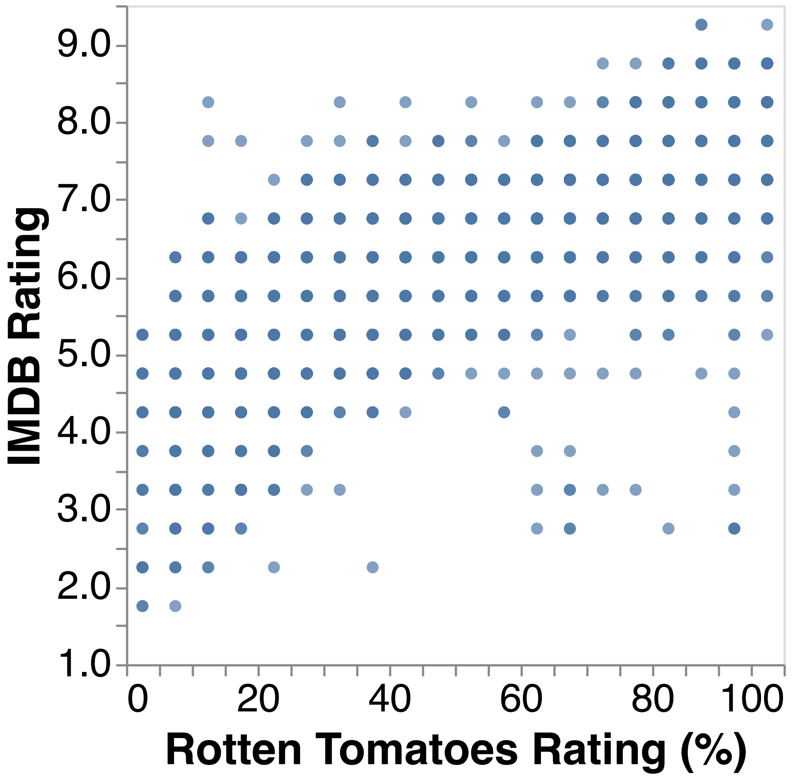
Two Continuous Variables

Two continuous variables: roadmap

- `movies` ratings from Rotten Tomatoes and IMDB
- `diamonds`: `carat` vs `price`

How are RT and IMDB ratings related?

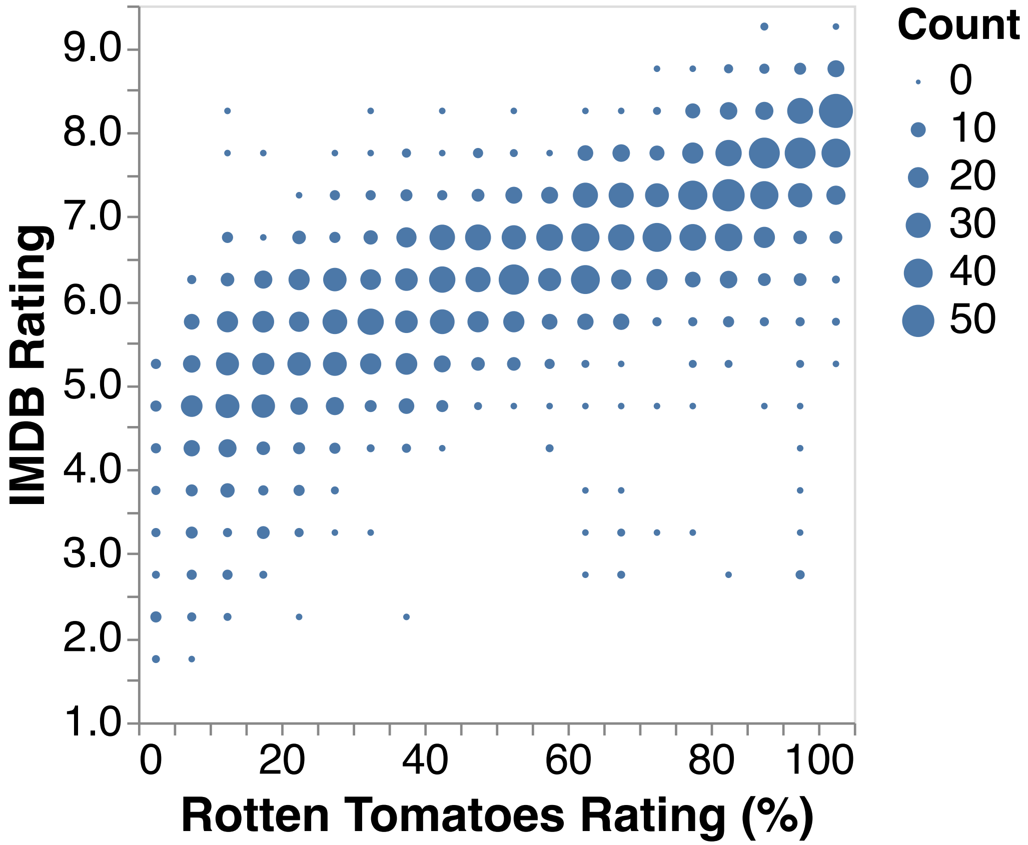
```
alt.Chart(movies).mark_circle().encode(
    alt.X('Rotten_Tomatoes_Rating:Q', bin=alt.BinParams(maxbins=20), title
    alt.Y('IMDB_Rating:Q', bin=alt.BinParams(maxbins=20), title = "IMDB Rat
)
```



Suffers from overplotting!

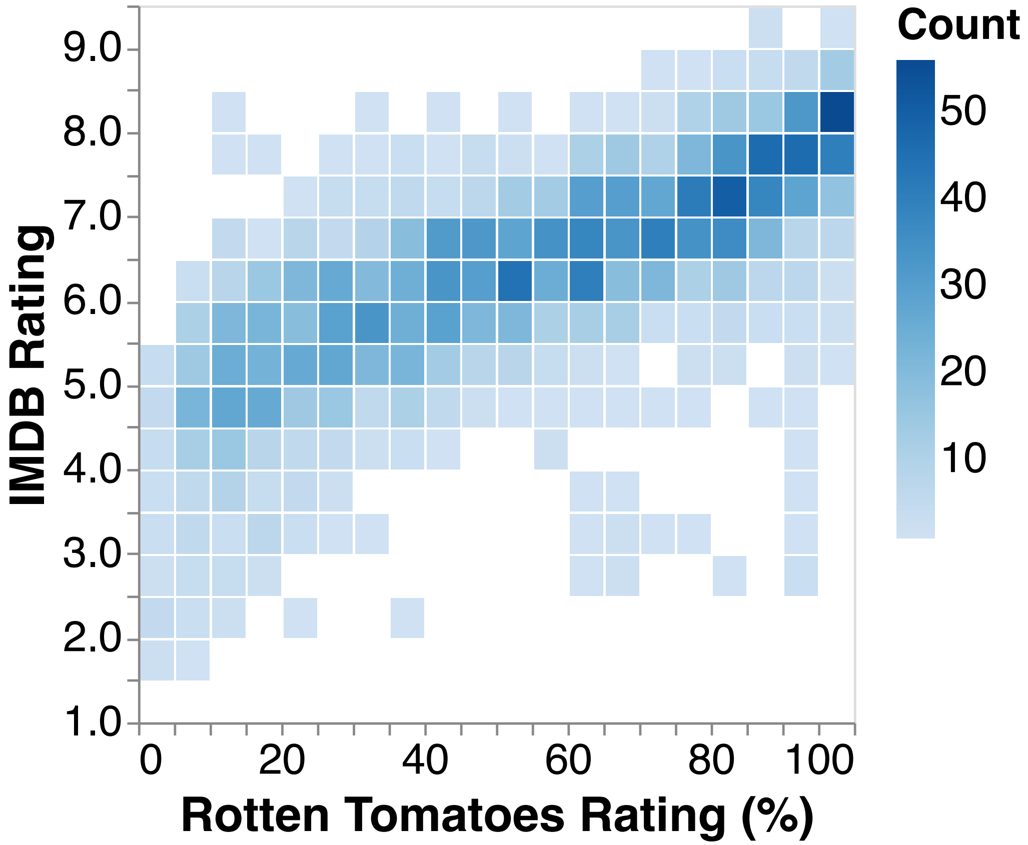
use alt.Size('count()')

```python
alt.Chart(movies_url).mark_circle().encode(
    alt.X('Rotten_Tomatoes_Rating:Q', bin=alt.BinParams(maxbins=20)),
    alt.Y('IMDB_Rating:Q', bin=alt.BinParams(maxbins=20)),
    alt.Size('count()')
)
```

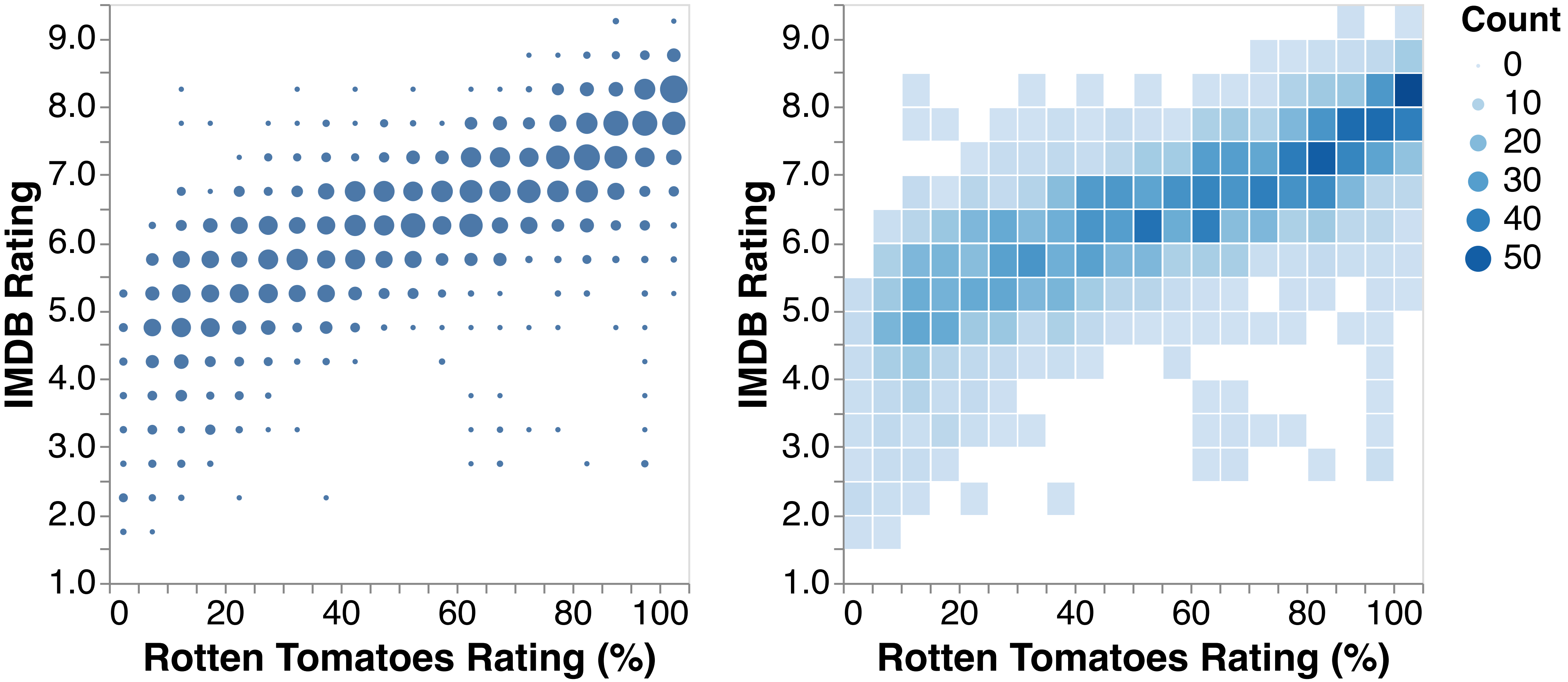


use alt.Color('count()')

```
alt.Chart(movies_url).mark_bar().encode(
    alt.X('Rotten_Tomatoes_Rating:Q', bin=alt.BinParams(maxbins=20), title
    alt.Y('IMDB_Rating:Q', bin=alt.BinParams(maxbins=20), title = "IMDB Rat
    alt.Color('count()', title = "Count")
)
```



Discussion question



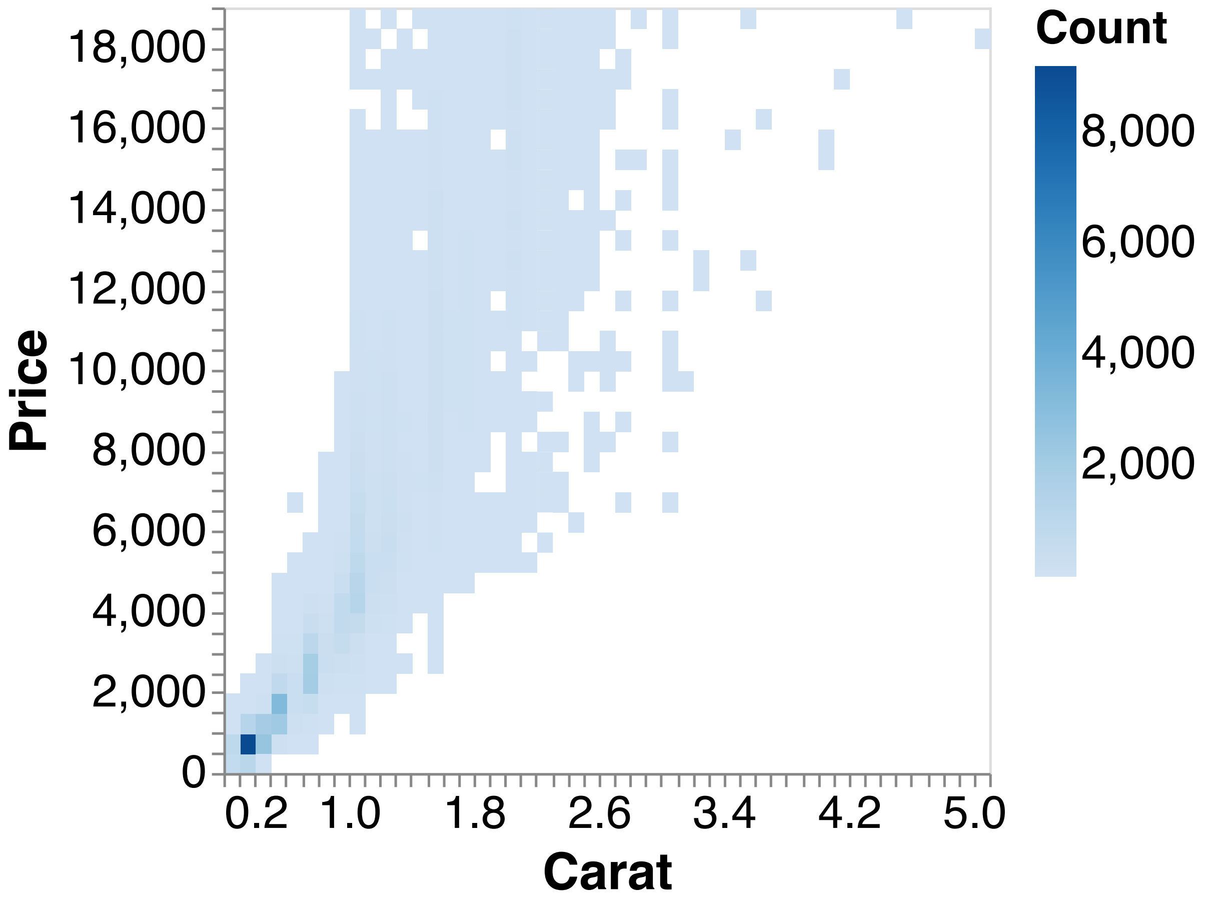
Compare the *size* and *color*-based 2D histograms above. Which encoding do you prefer? Why?

How is carat related to price? Raw data

```
alt.Chart(diamonds).mark_point().encode(
    alt.X('carat:Q', title = "Carat"),
    alt.Y('price:Q', title = "Price")
)
```

How is carat related to price? Color

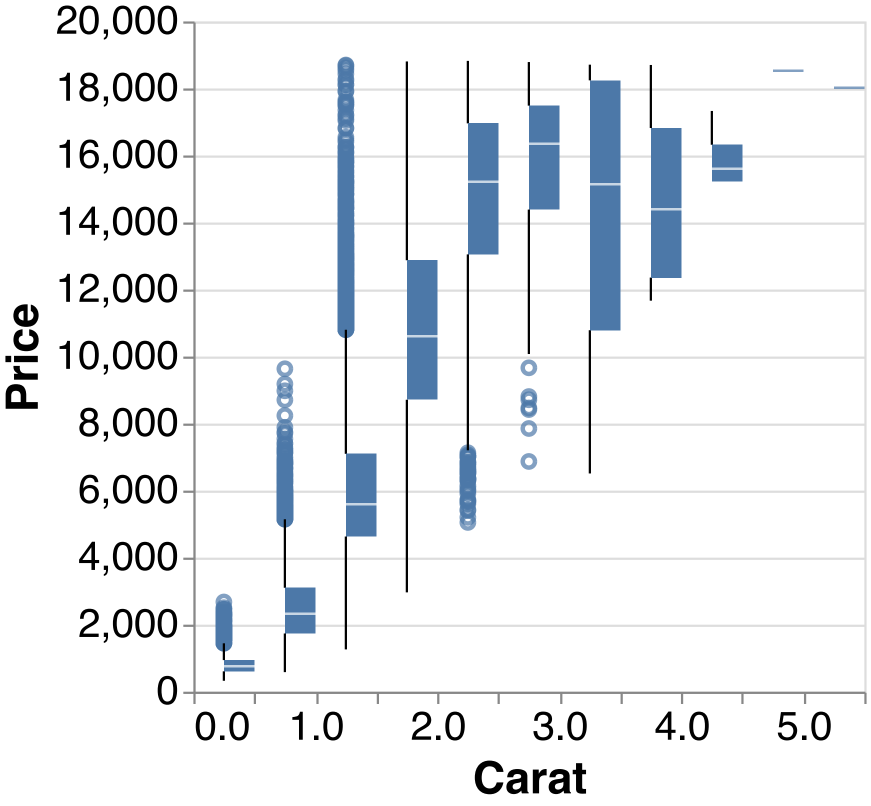
```python
alt.Chart(diamonds).mark_rect().encode(
    alt.X('carat:Q', bin=alt.Bin(maxbins=70), title = "Carat"),
    alt.Y('price:Q', bin=alt.Bin(maxbins=70), title = "Price"),
    alt.Color('count()', scale=alt.Scale(scheme='blues'), title = "Count"))
```



How is carat related to price?

`mark_boxplot()`

```python
alt.Chart(diamonds).mark_boxplot().encode(
    alt.X('carat:Q', bin=alt.Bin(maxbins=10), title = "Carat"),
    alt.Y('price:Q', title = "Price"))
```



How is carat related to price? binscatter (process)

- Can also create a binscatter: `binscatter` in stata and `binsreg` in R.
- Doesn't exist yet for `altair`, but easy to code up yourself
- What it does:

1. Computes bins using quantiles of x
2. Computes means of y within each bin

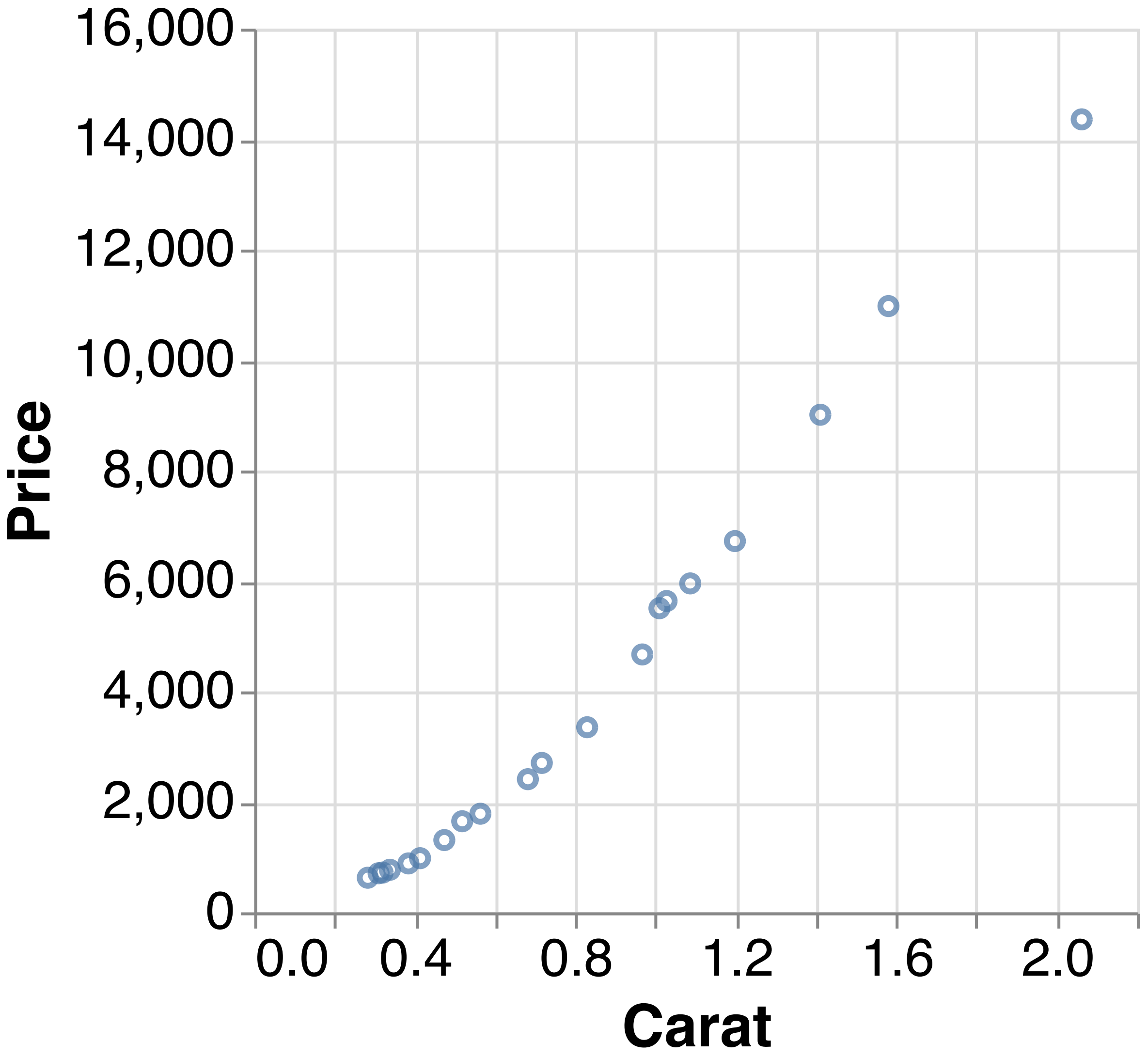
How is carat related to price? binscatter (code)

```
df = diamonds
df['carat_bin'] = pd.qcut(df['carat'], q=20, labels=(np.arange(1, 21, 1)))

df = df.groupby('carat_bin').agg(
    carat = ('carat', 'mean'),
    price = ('price', 'mean')).reset_index()

alt.Chart(df).mark_point().encode(
   alt.X('carat:Q', title = "Carat"),
   alt.Y('price:Q', title = "Price")
)
```

How is carat related to price? (plot)



Discussion question – “How is carat related to price?”

Review the `mark_rect()`, `mark_boxplot()`, and `binscatter` plots

- Headline?
- Sub-messages?

Exploring covariation: summary

Scenario	Functions
Categorical and continuous variable	`mark_boxplot()`
	`transform_density()`
	`alt.Row()`
Two categorical variables	`size`
	`color`
Two continuous variables	`alt.Size('count()')`
	`alt.Color('count()')`
	`mark_boxplot()`
	binscatter