Visualisation in R

Hadley Wickham

ssistant Professor / Dobelman Family Junior Cha

Department of Statistics / Rice University

1. Scatterplot basics
2. Adding extra variables
3. Jittering and boxplots
4. Histograms and barcharts
5. Learn more


Learning a new
language is hard!

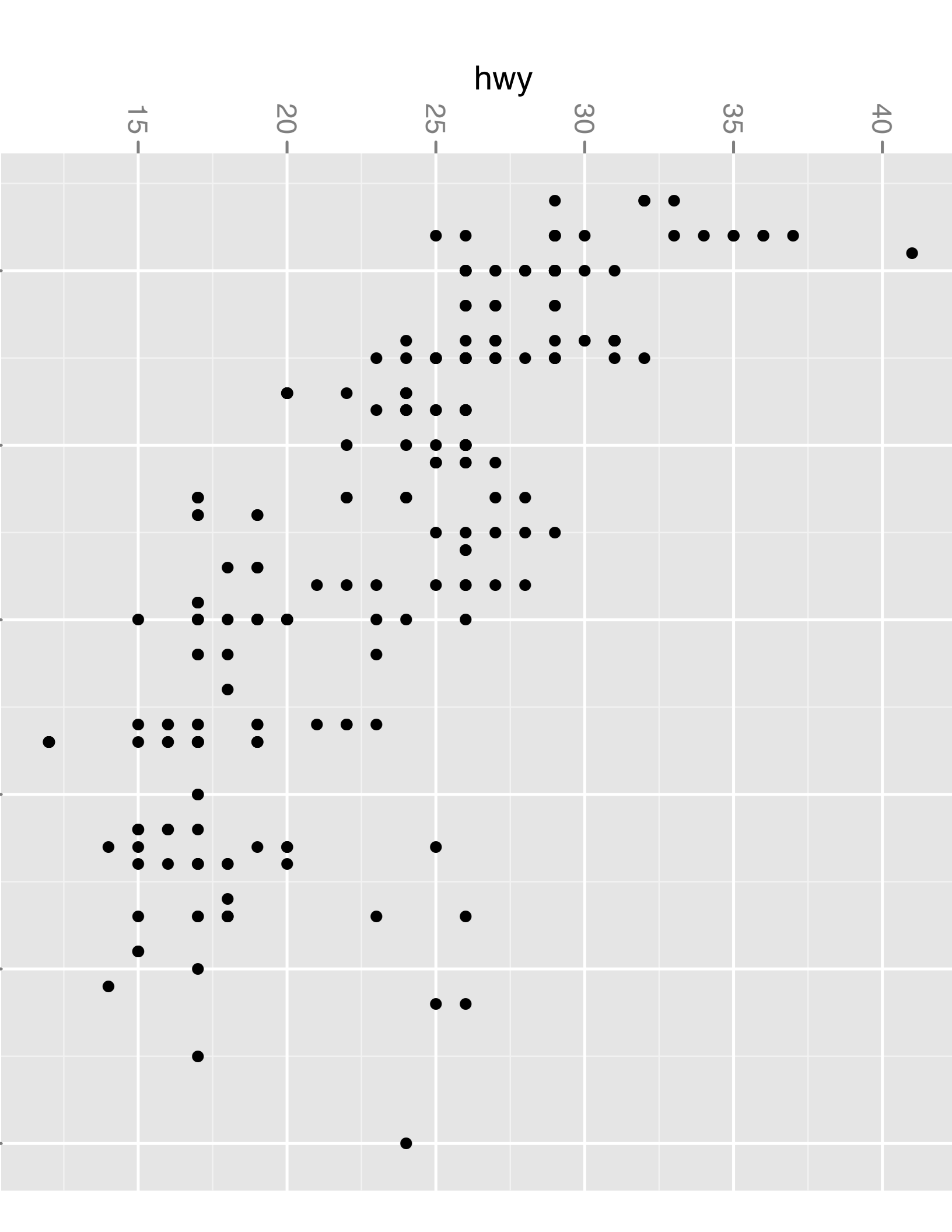
Scatterplot basics

```
install.packages("ggplot2")
library(ggplot2)

?mpg
head(mpg)
str(mpg)
summary(mpg)

qplot(displ, hwy, data = mpg)
```

Scatterplot basics

```
install.packages("ggplot2")
library(ggplot2)

?mpg
head(mpg)
str(mpg)
summary(mpg)

qplot(displ, hwy, data = mpg)
```

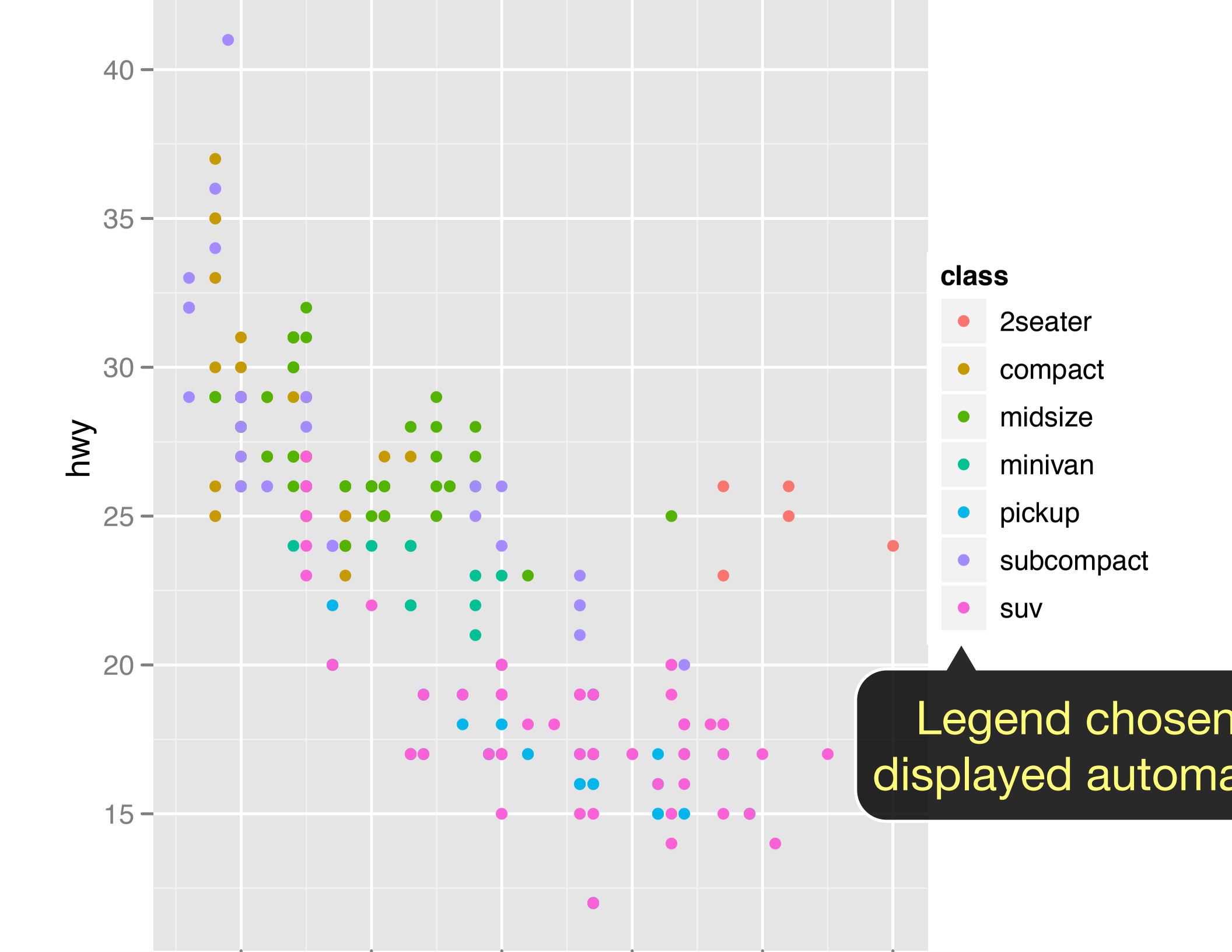
In ggplot2, we always explicitly specify the data


hwy
15
20
25
30
35
40

Additional variables

Can display additional variables with **aesthetics** (like shape, colour, size) or **facetting** (small multiples displaying different subsets)


hwy
40
35
30
25
20
15
class
2seater
compact
midsize
minivan
pickup
subcompact
suv


hwy
40
35
30
25
20
15
class
2seater
compact
midsize
minivan
pickup
subcompact
suv
Legend chosen
displayed automa

Your turn

Experiment with colour, size, and shape aesthetics.

What's the difference between discrete o continuous variables?

What happens when you combine multiple aesthetics?

	Discrete	Continuou
olour	Rainbow of colours	Gradient fro red to blue
Size	Discrete size steps	Linear mapp between rad and value
hape	Different shape for each	Doesn’t wo

Faceting

Small multiples displaying different subsets of the data.

Useful for exploring conditional relationships. Useful for large data.

Your turn

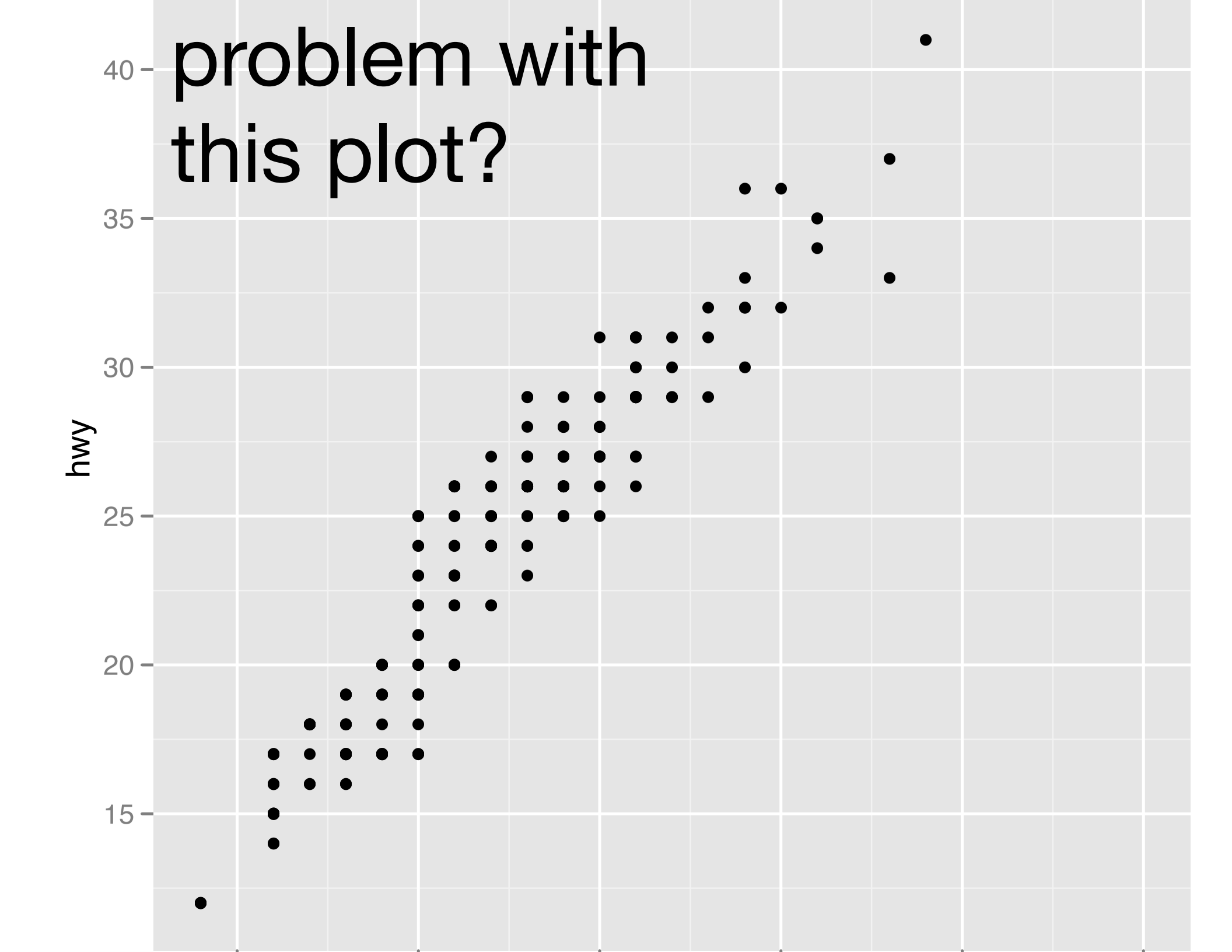
```
qplot(displ, hwy, data = mpg) +
facet_grid(. ~ cyl)

qplot(displ, hwy, data = mpg) +
facet_grid(drv ~ .)

qplot(displ, hwy, data = mpg) +
facet_grid(drv ~ cyl)

qplot(displ, hwy, data = mpg) +
facet_wrap(~ class)
```

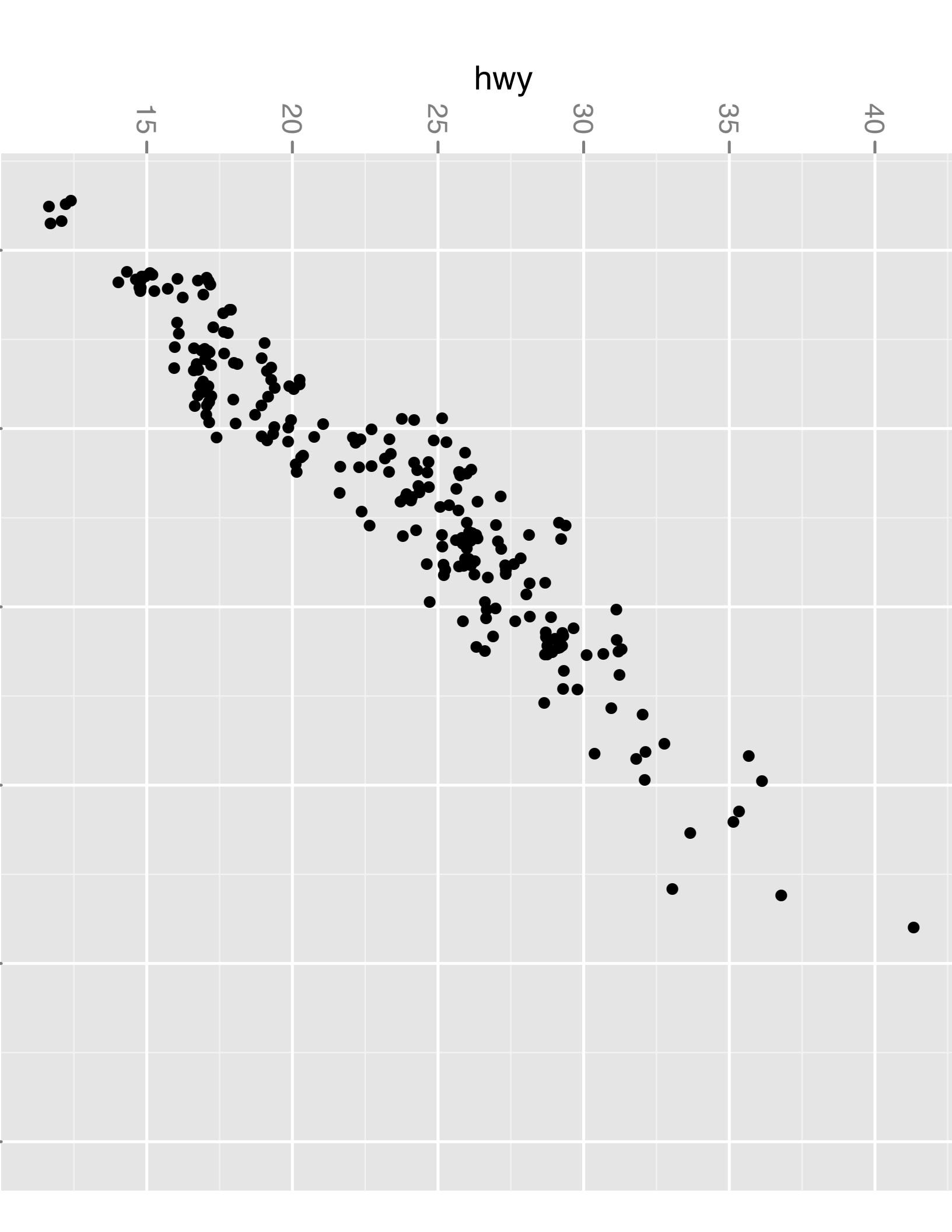
Summary

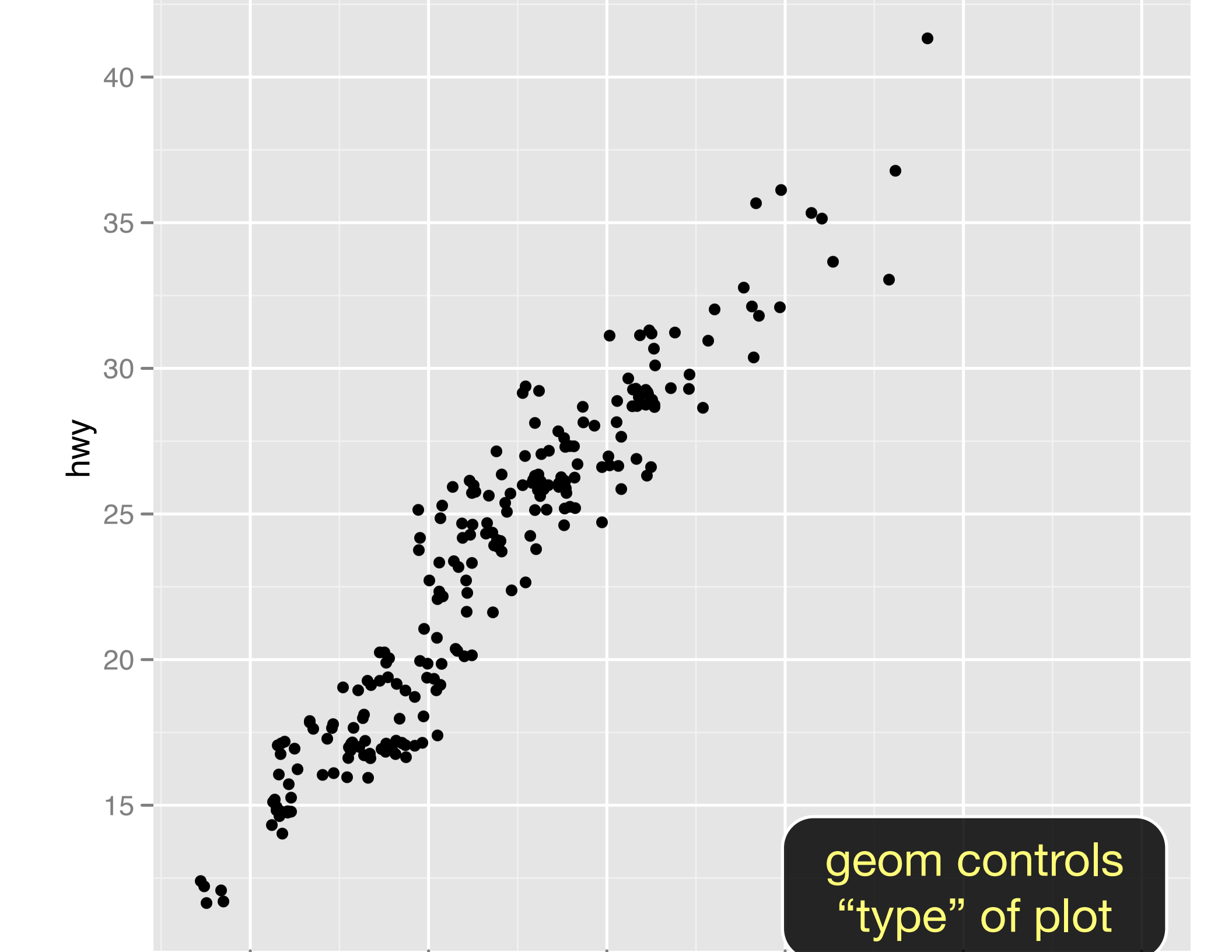
`facet_grid()`: 2d grid, rows ~ cols, . for no split

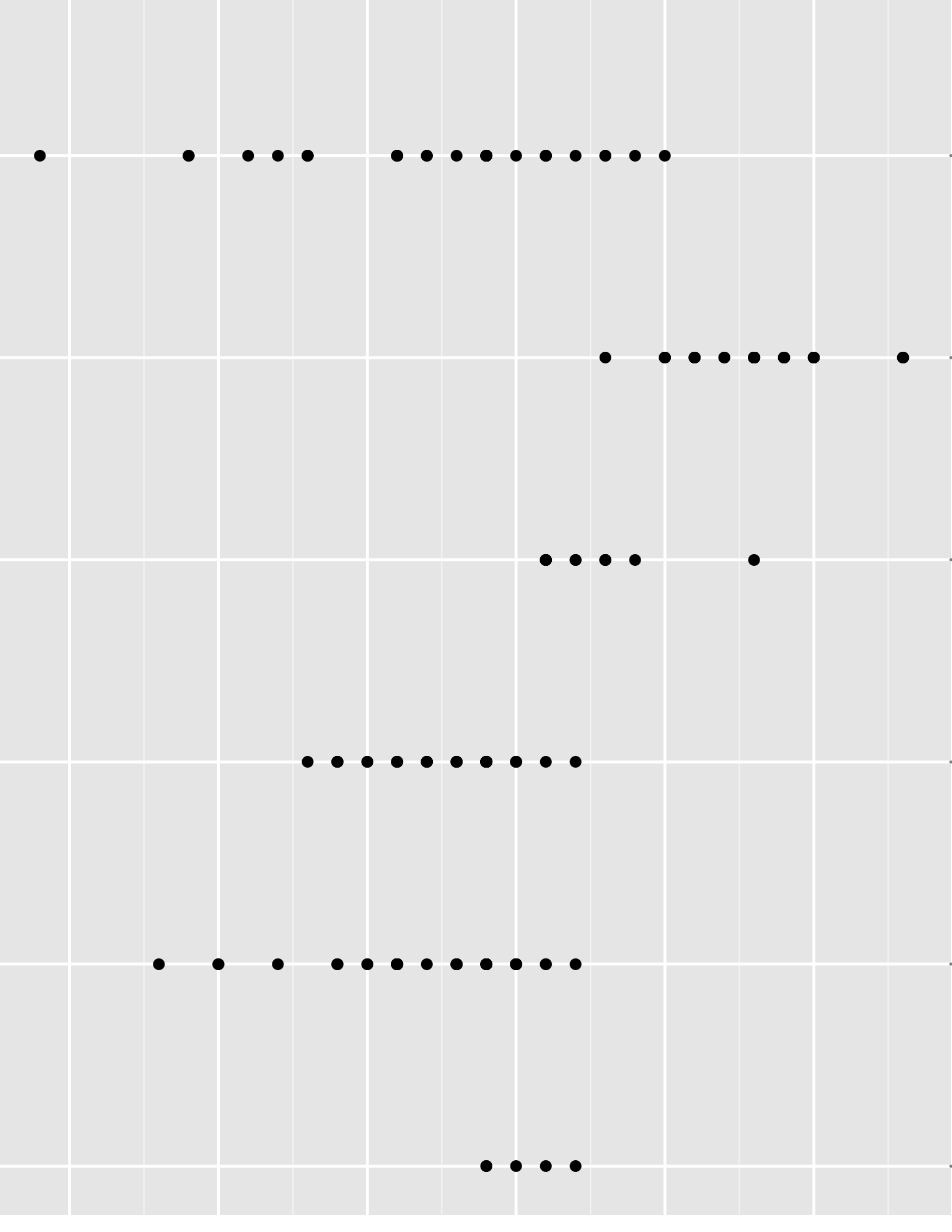
`facet_wrap()`: 1d ribbon wrapped into 2d

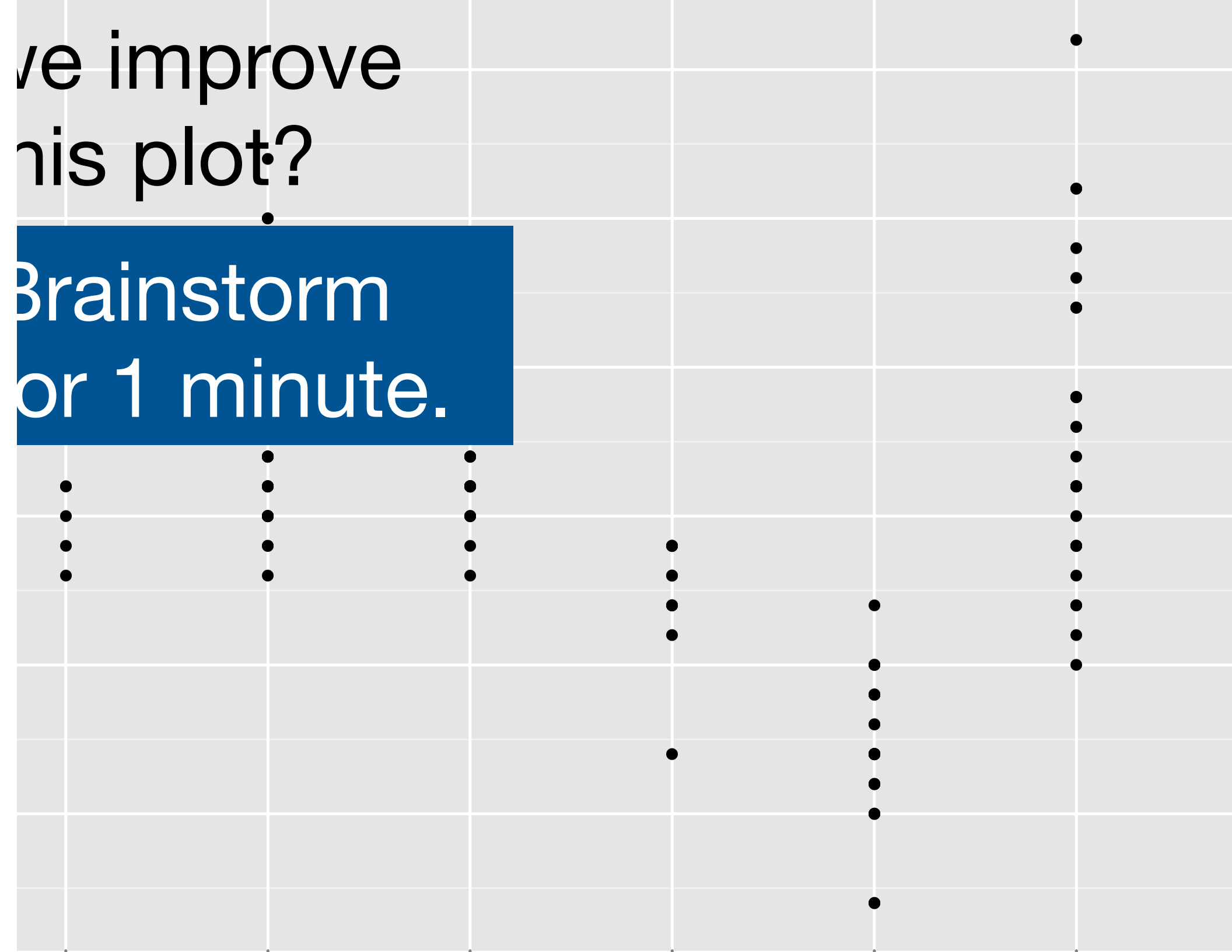
Scales argument controls whether position scales are fixed or free.

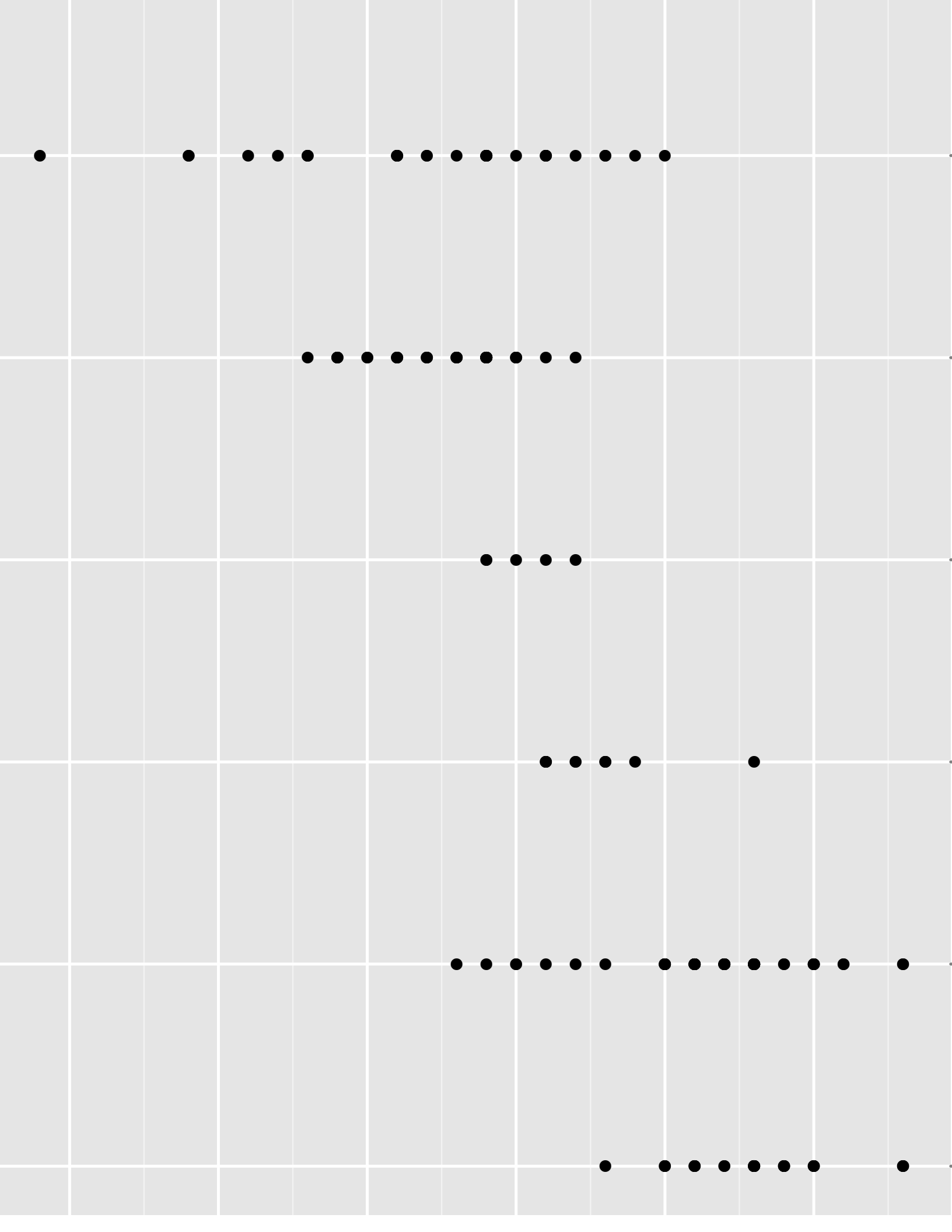

problem with
this plot?
hwy
40
35
30
25
20
15

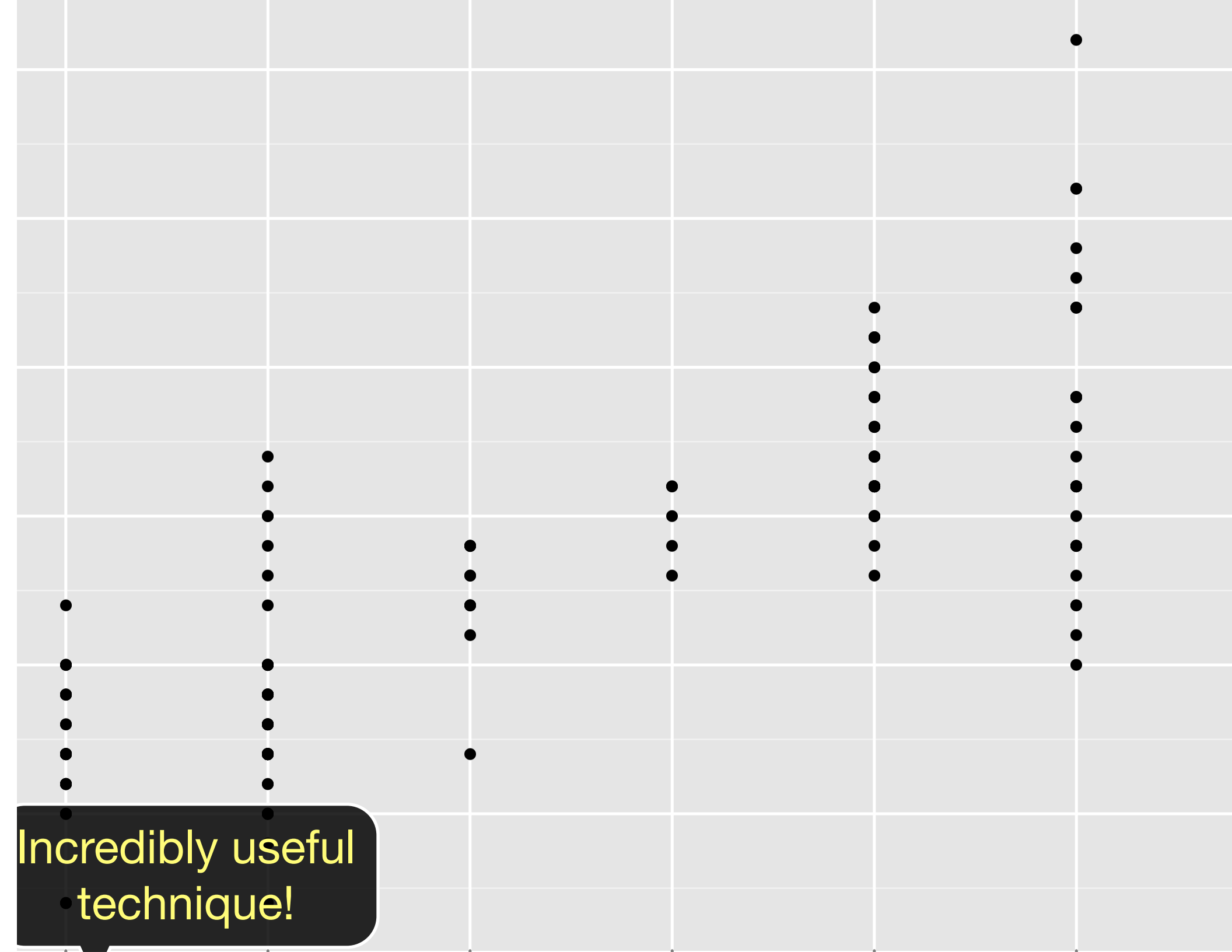

hwy
15
20
25
30
35
40

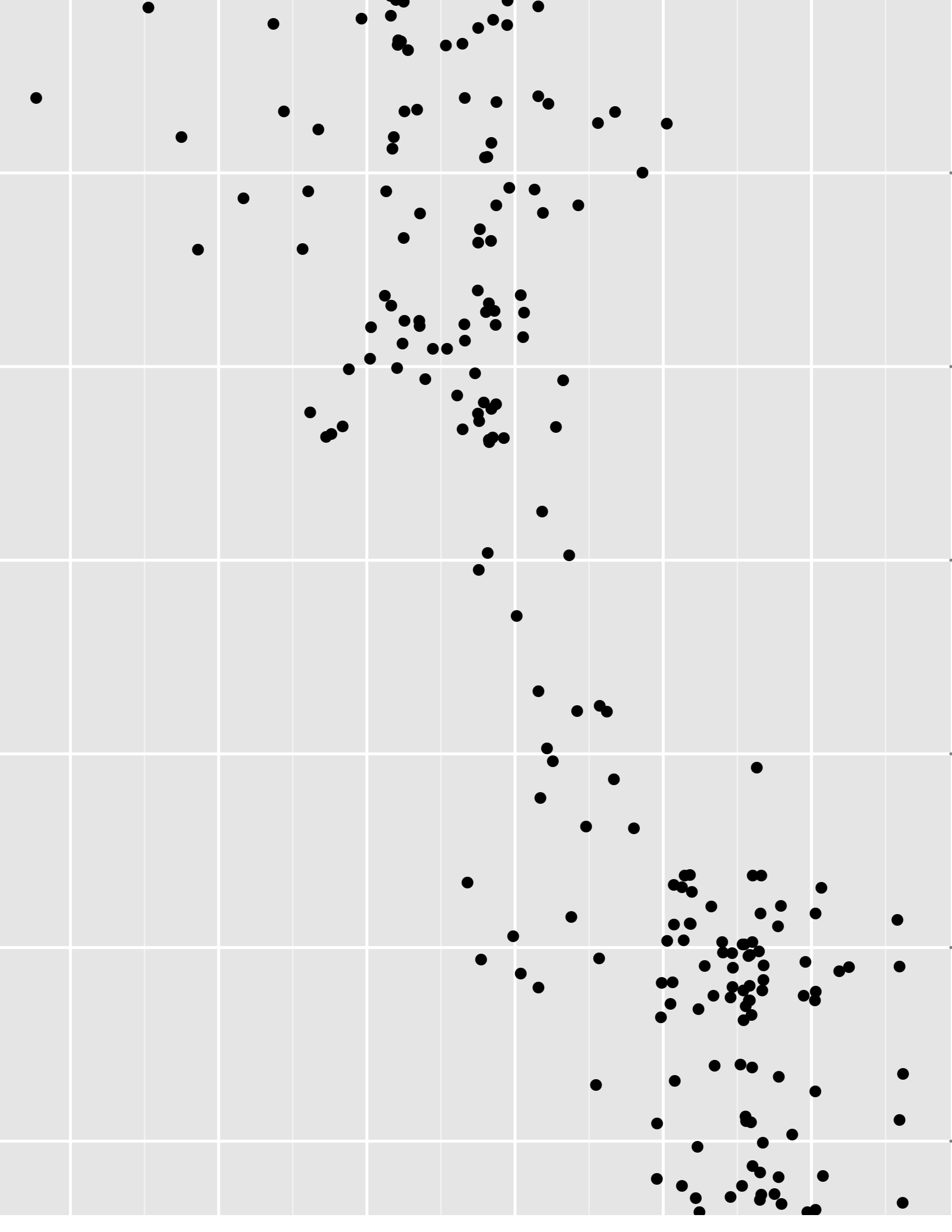

hwy
40
35
30
25
20
15
geom controls
“type” of plot

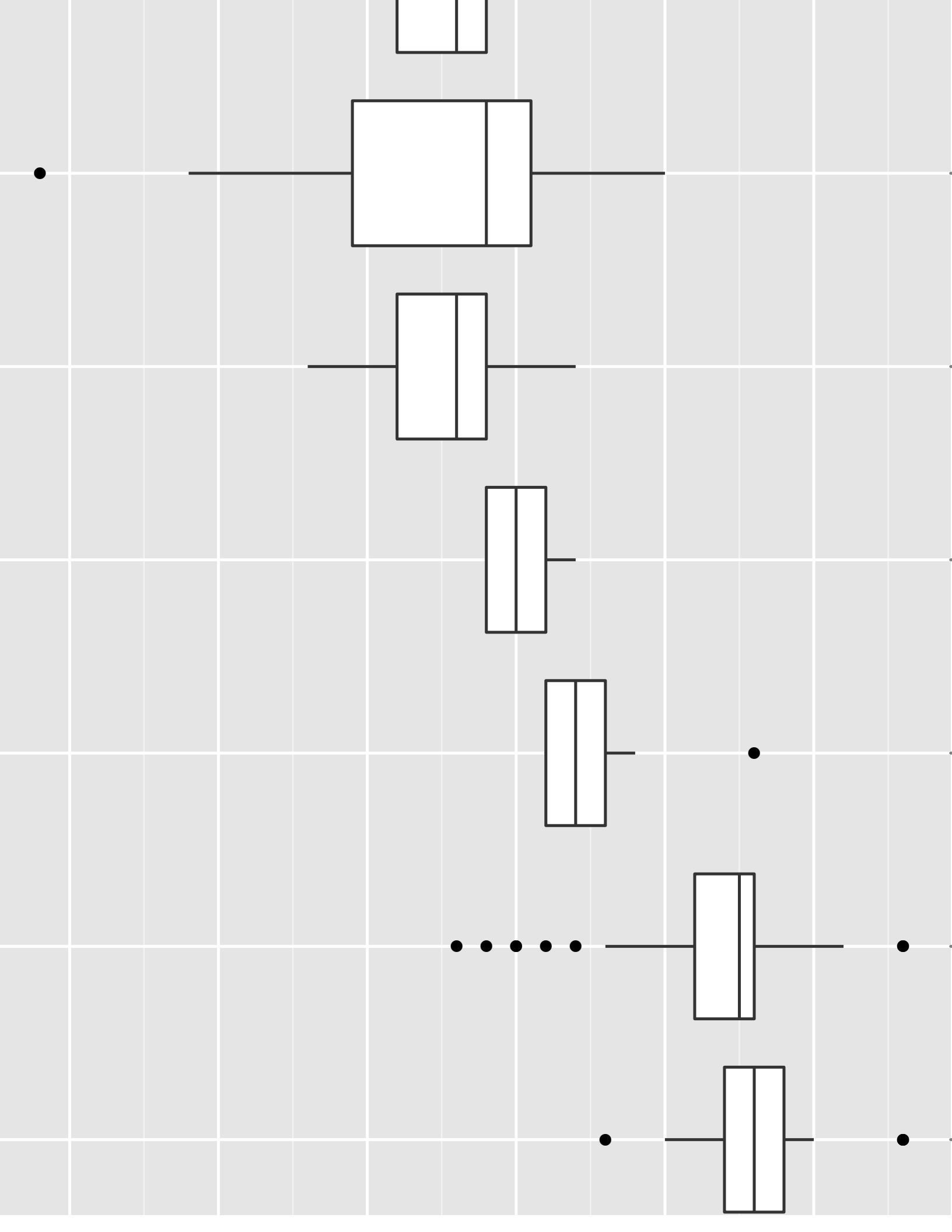


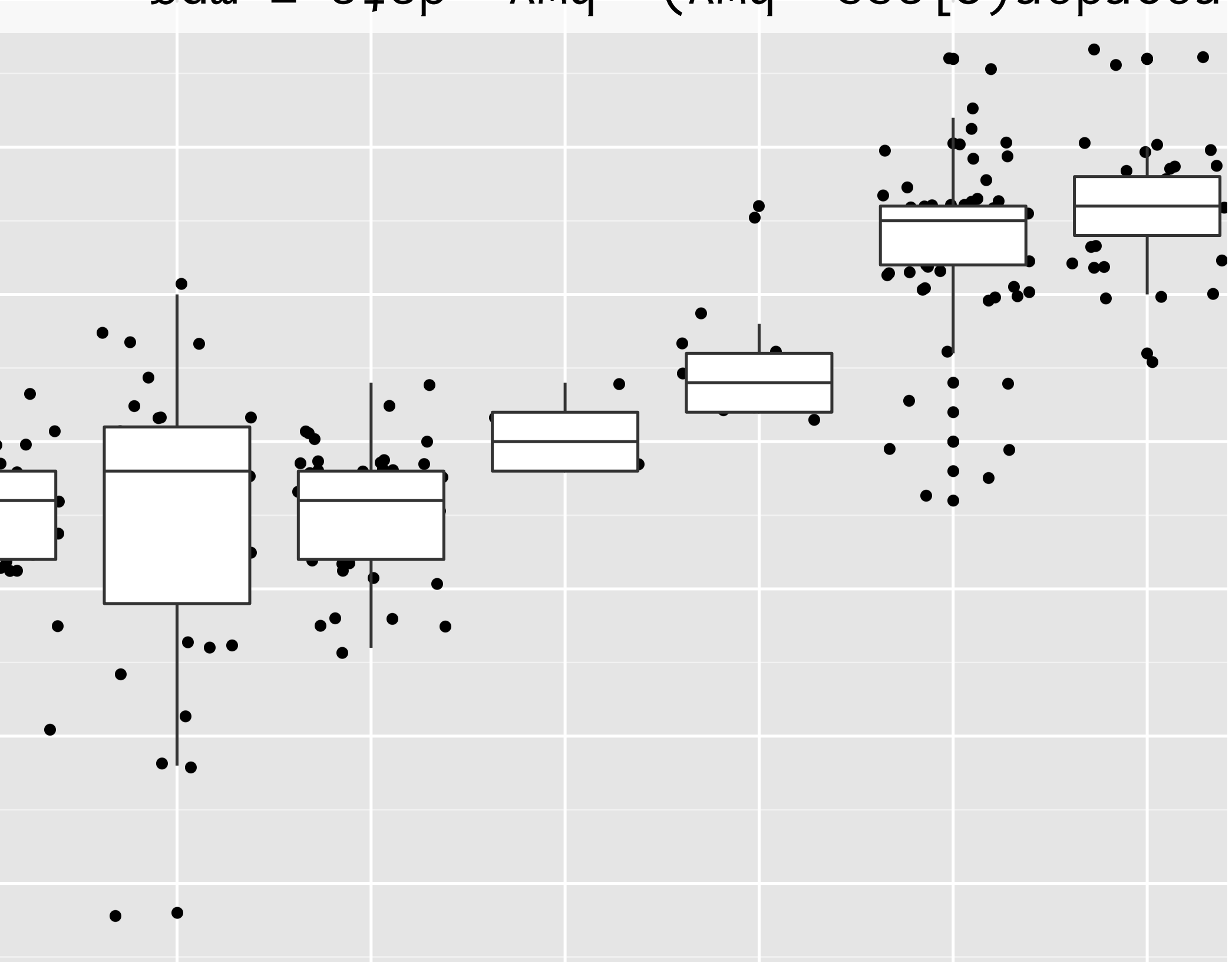

ve improve
his plot?
Brainstorm
or 1 minute.




Incredibly useful technique!







Your turn

Read the help for reorder. Redraw the previously plots with class ordered by median hwy.

How would you put the jittered points on top of the boxplots?

Aside: coding strategy

At the end of each interactive session, you want a summary of everything you did. Two options:

1. Save everything you did with `savehistory(` then remove the unimportant bits.

2. Build up the important bits as you go. (this is how I work)

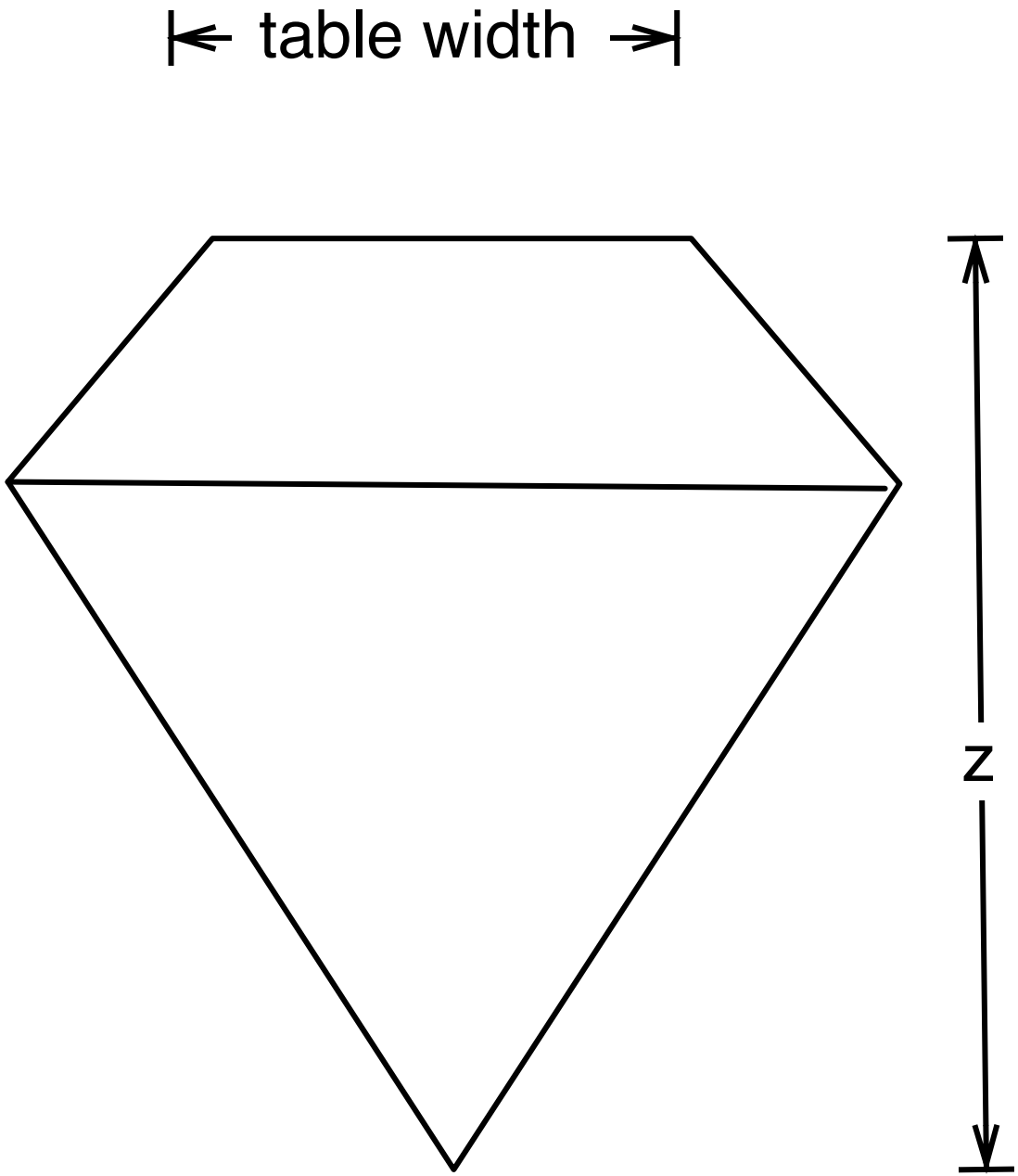
Diamonds data

~**54,000** round diamonds from http://www.diamondse.info/

Carat, colour, clarity, cut

Total depth, table, depth, width, height

Price



depth = z / diameter

table = table width / x * 100

Histogram & bar charts

Histograms and barcharts

Used to display the **distribution** of a variable

Categorical variable → bar chart

Continuous variable → histogram

Always

xperiment wit

the bin width!

Examples

```
# With only one variable, qplot guesses that
# you want a bar chart or histogram
qplot(cut, data = diamonds)

qplot(carat, data = diamonds)
qplot(carat, data = diamonds, binwidth = 1)
qplot(carat, data = diamonds, binwidth = 0.1)
qplot(carat, data = diamonds, binwidth = 0.01)
resolution(diamonds$carat)

last_plot() + xlim(0, 3)
```

Examples

```
# With only one variable, qplot guesses that
# you want a bar chart or histogram
qplot(cut, data = diamonds)

qplot(carat, data = diamonds)
qplot(carat, data = diamonds, binwidth = 1)
qplot(carat, data = diamonds, binwidth = 0.1)
qplot(carat, data = diamonds, binwidth = 0.01)
resolution(diamonds$carat)

last_plot() + xlim(0, 3)
```

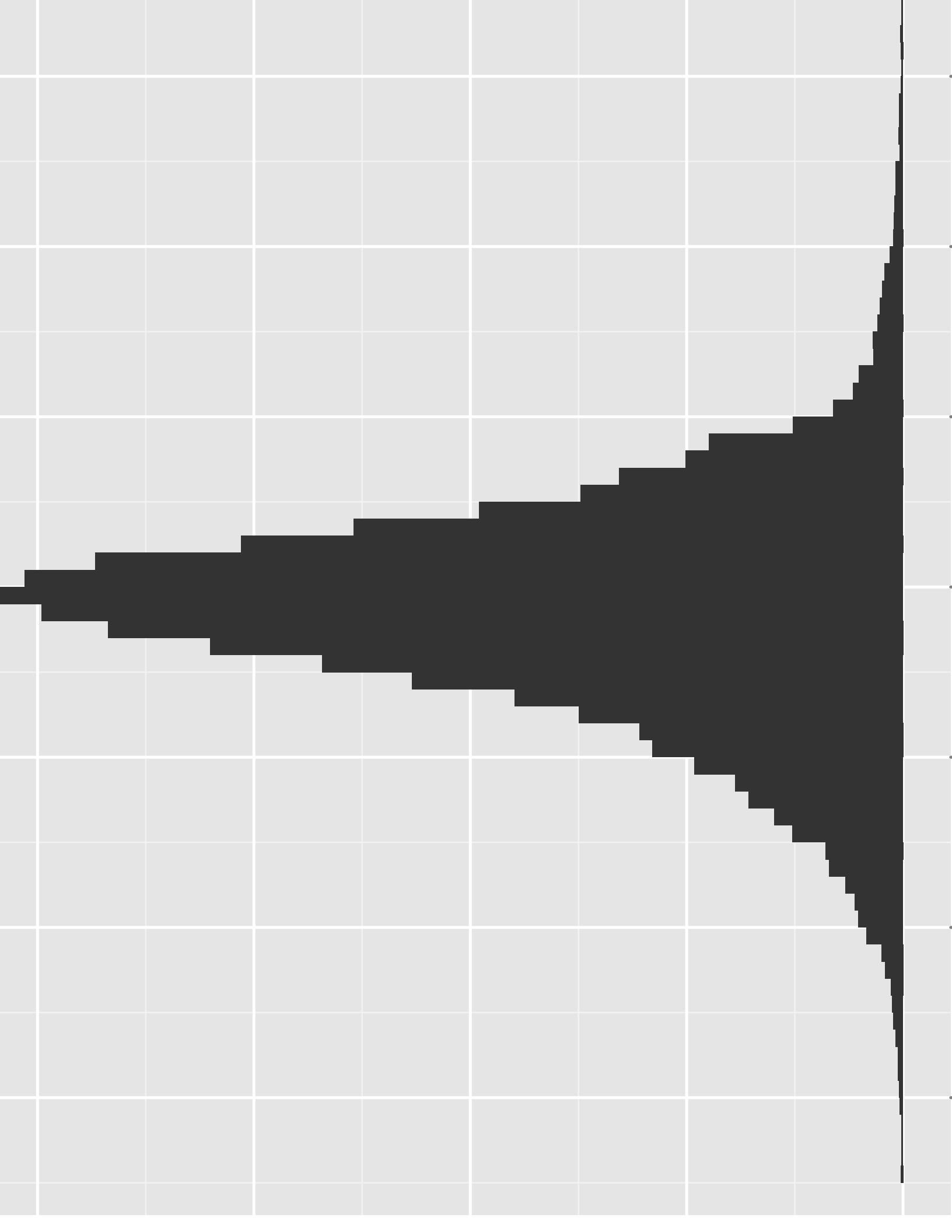
Common ggplot2 technique: adding together plot components

```
t(table, data = diamonds, binwidth = 1)


 zoom in on a plot region use xlim() and yli
t(table, data = diamonds, binwidth = 1) +
lim(50, 70)
t(table, data = diamonds, binwidth = 0.1) +
im(50, 70)
t(table, data = diamonds, binwidth = 0.1) +
im(50, 70) + ylim(0, 50)


te that this type of zooming discards data
ide of the plot regions
e coord_cartesian() for an alternative
```

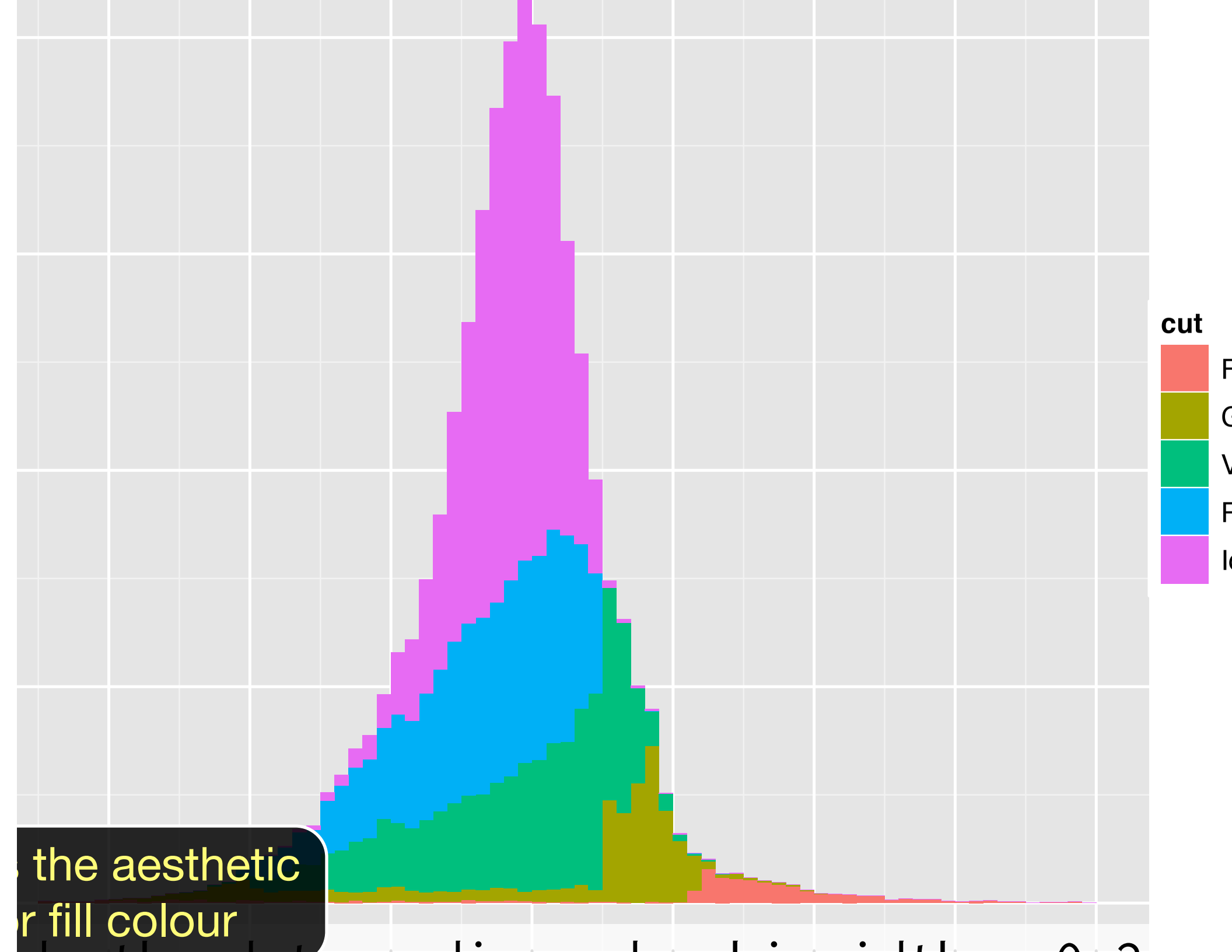
Additional variables

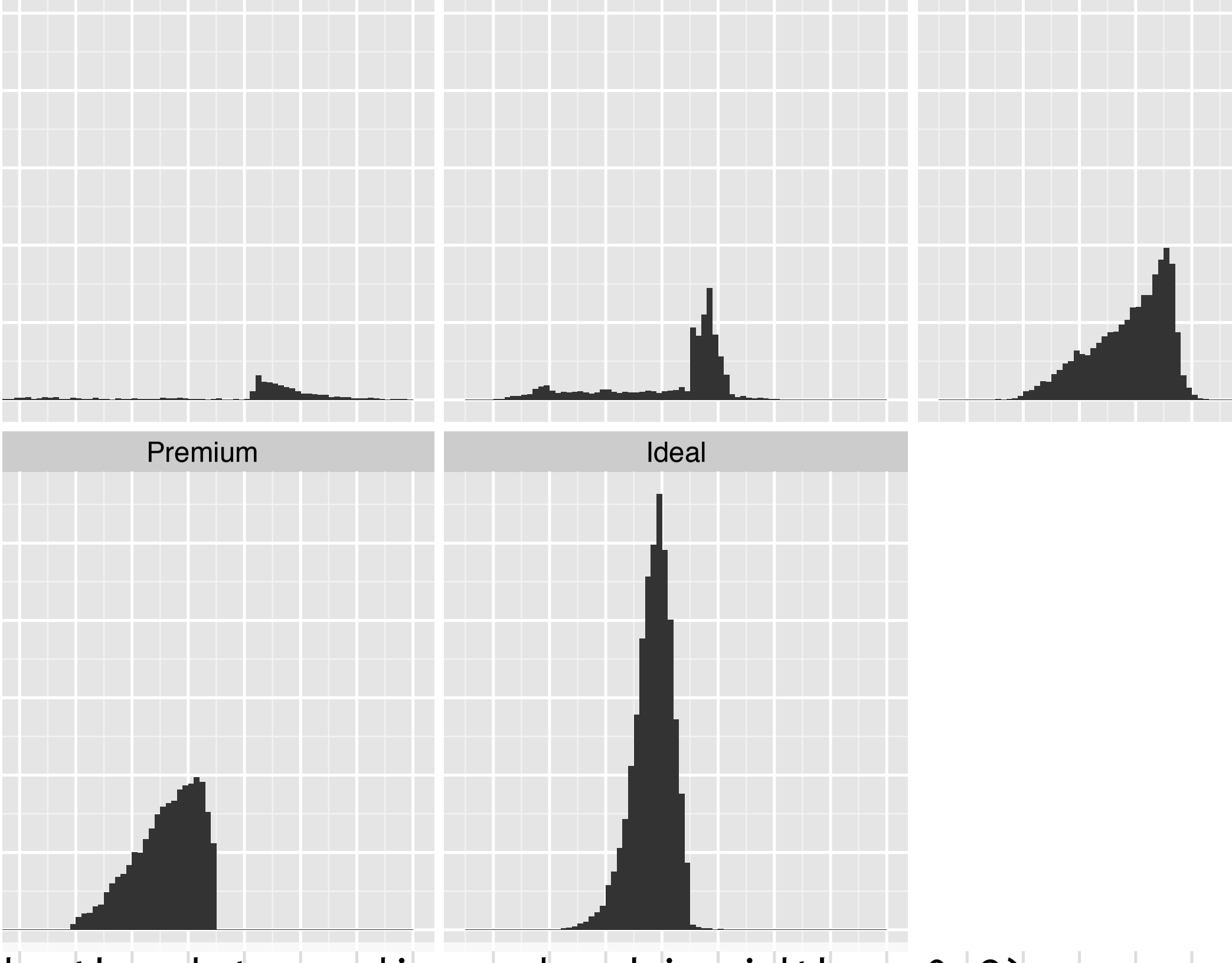
As with scatterplots can use **aesthetics** or **faceting**. Using aesthetics creates pretty, but ineffective, plots.

The following examples show the difference, when investigation the relationship between cut and depth.



cut


cut
the aesthetic
r fill colour


Premium
Ideal

Your turn

Explore the distribution of price.

How does it vary with colour, or cut, and clarity?

Learn more

About ggplot2

Graphical grammar (domain specific language), based on “The Grammar of Graphics” by Leland Wilkinson.

Specify what you want, not how to create it. Many fiddly details taken care of.

“Instead of spending time making your graph look pretty, you can focus on creating a graph that bests reveals the messages in your data.

Useful resources

http://had.co.nz/ggplot2

http://had.co.nz/ggplot2/book

http://groups.google.com/group/ggplot2

http://learnr.wordpress.com

http://ggplot2.wik.is