Intro to ggplot2

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

November 2010

HELLO

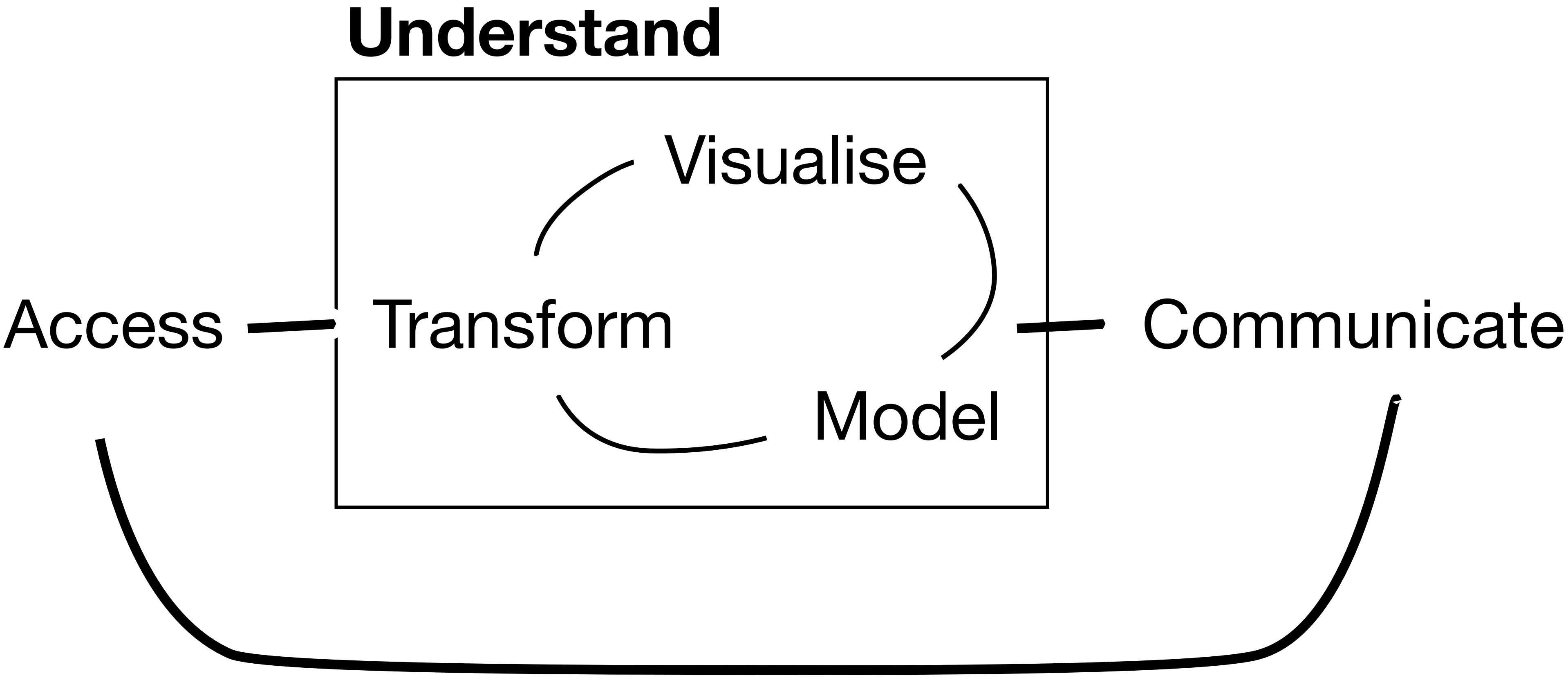
my name is

Hadley

had.co.nz/courses/
10-tokyo

Outline

Data analysis is the process by which data becomes understanding, knowledge and insight



Understand

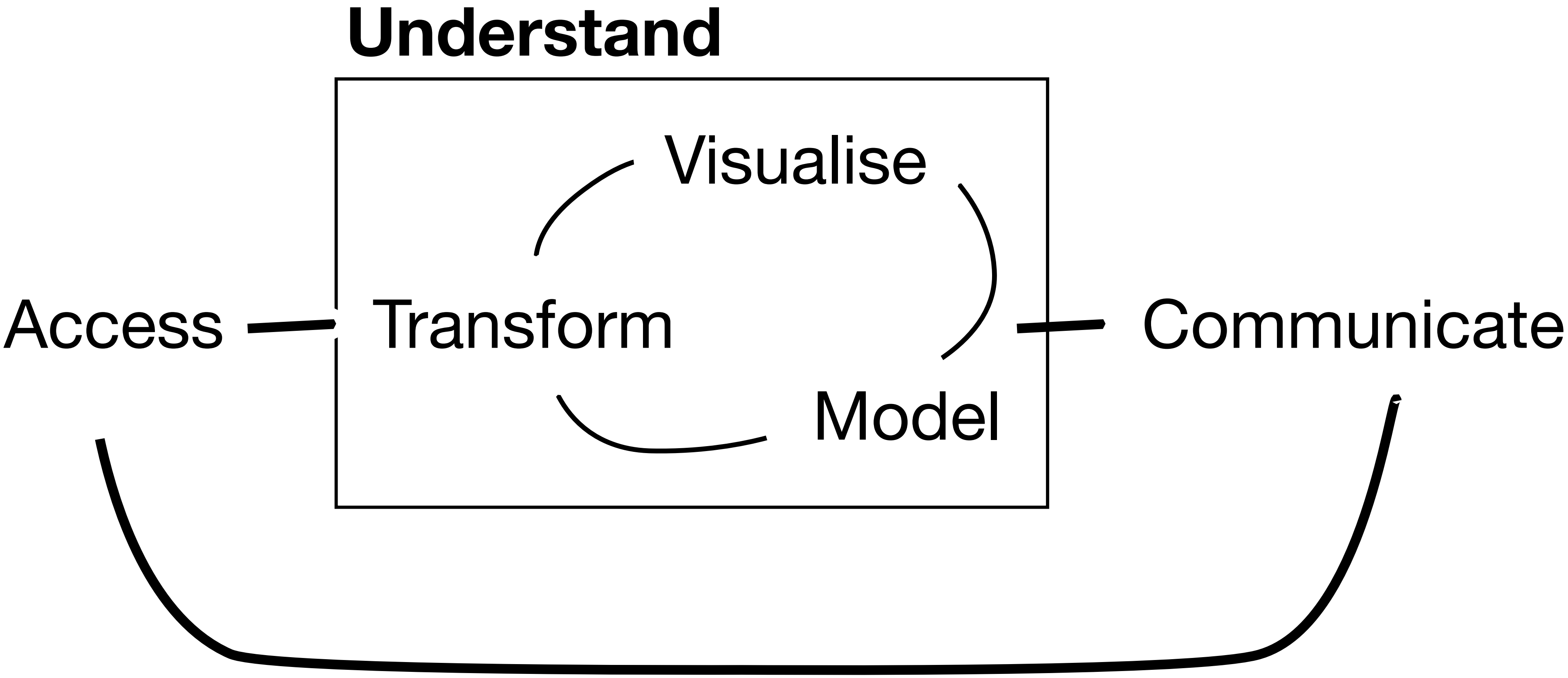
Visualise

Access

Transform

Communicate

Model



Understand

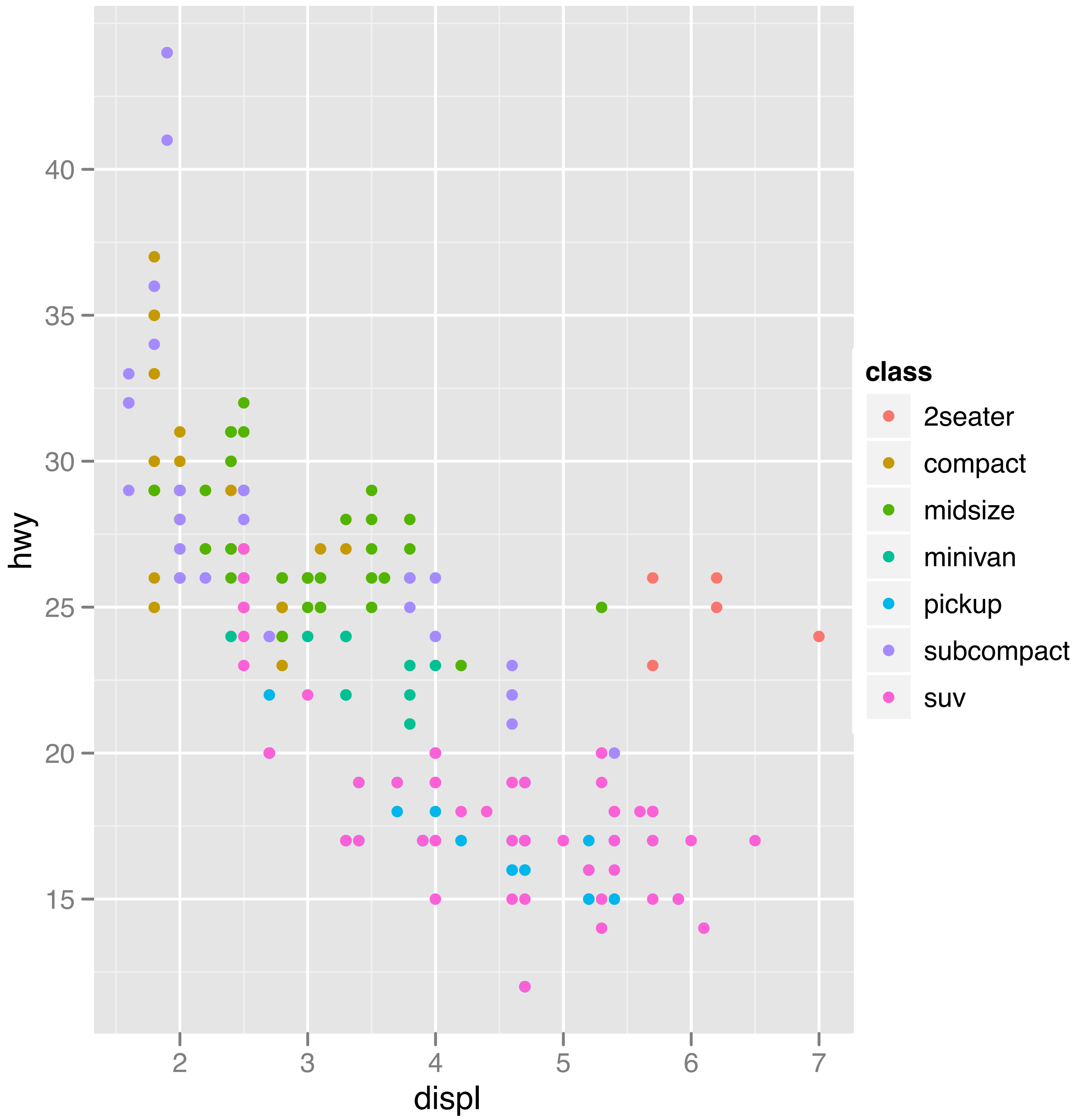
Visualise

Access

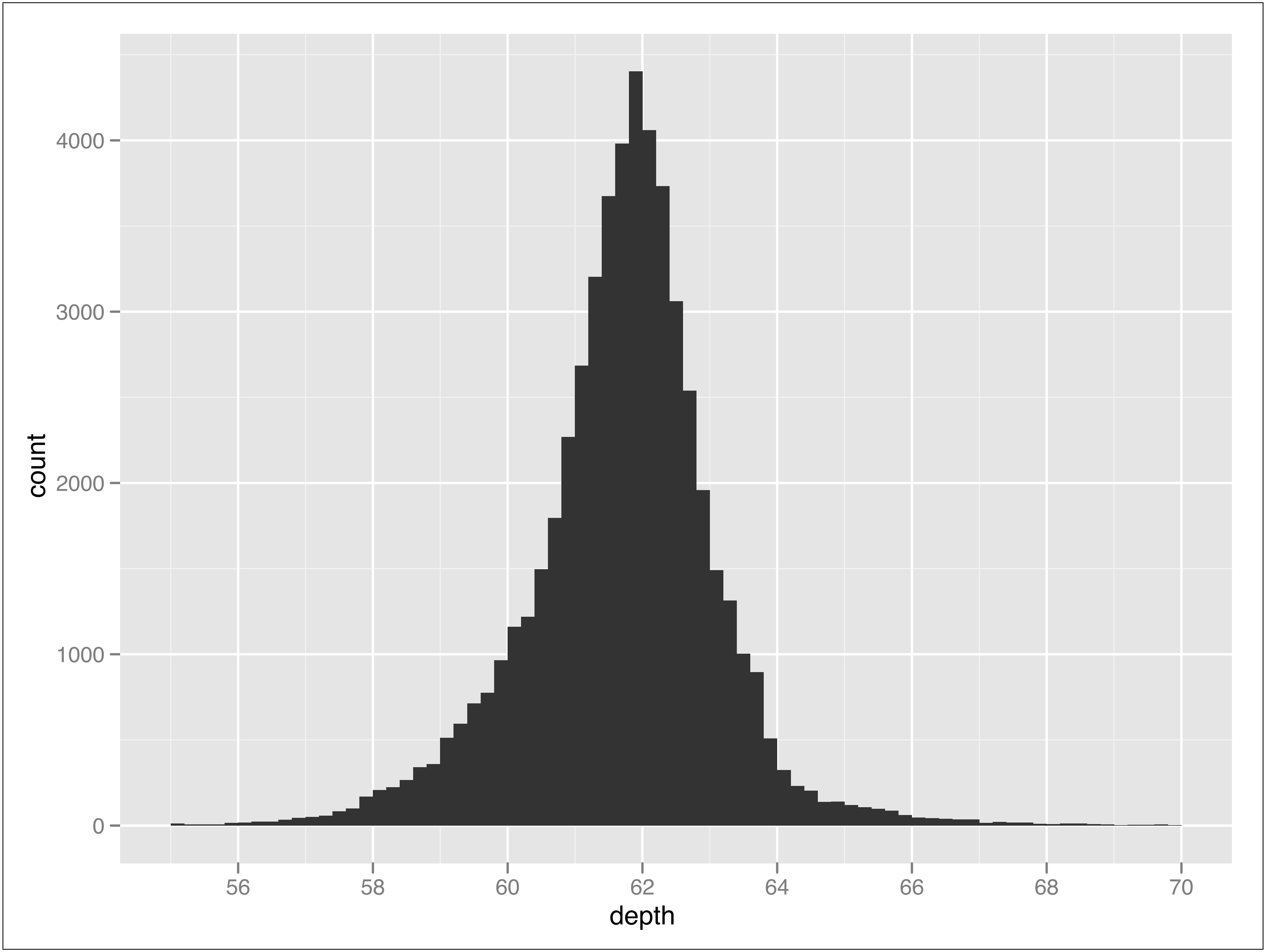
Transform

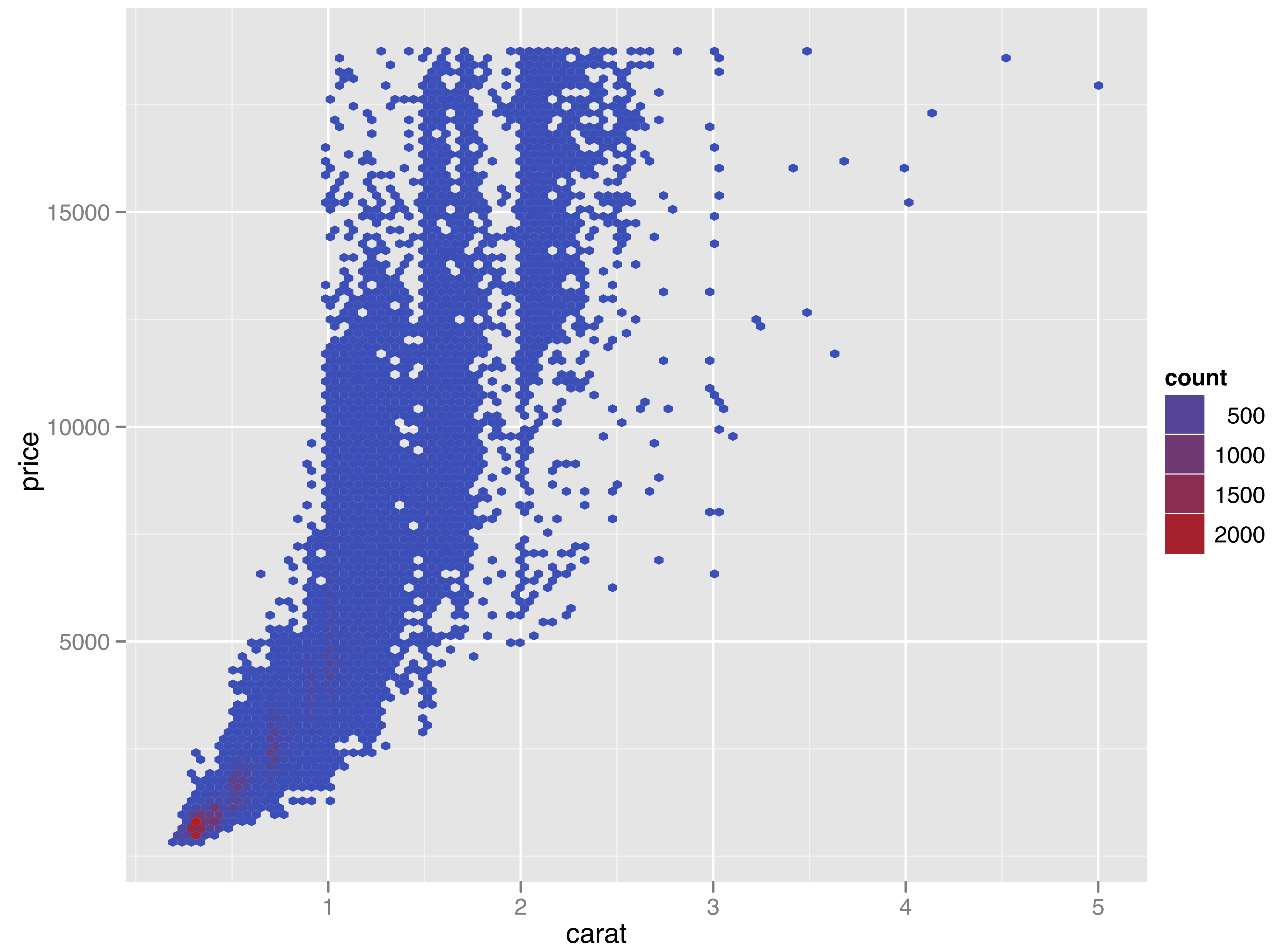
Communicate

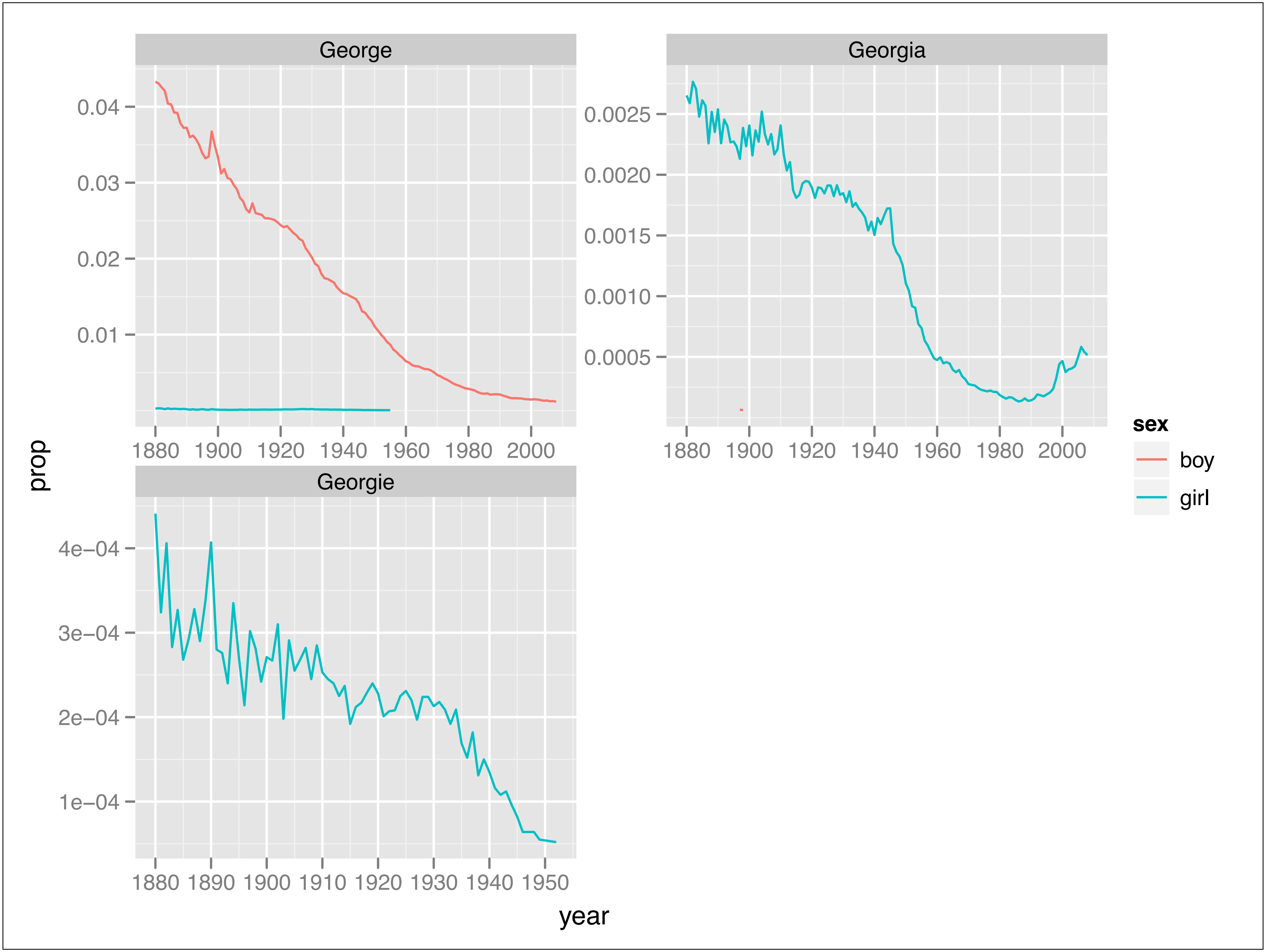
Model

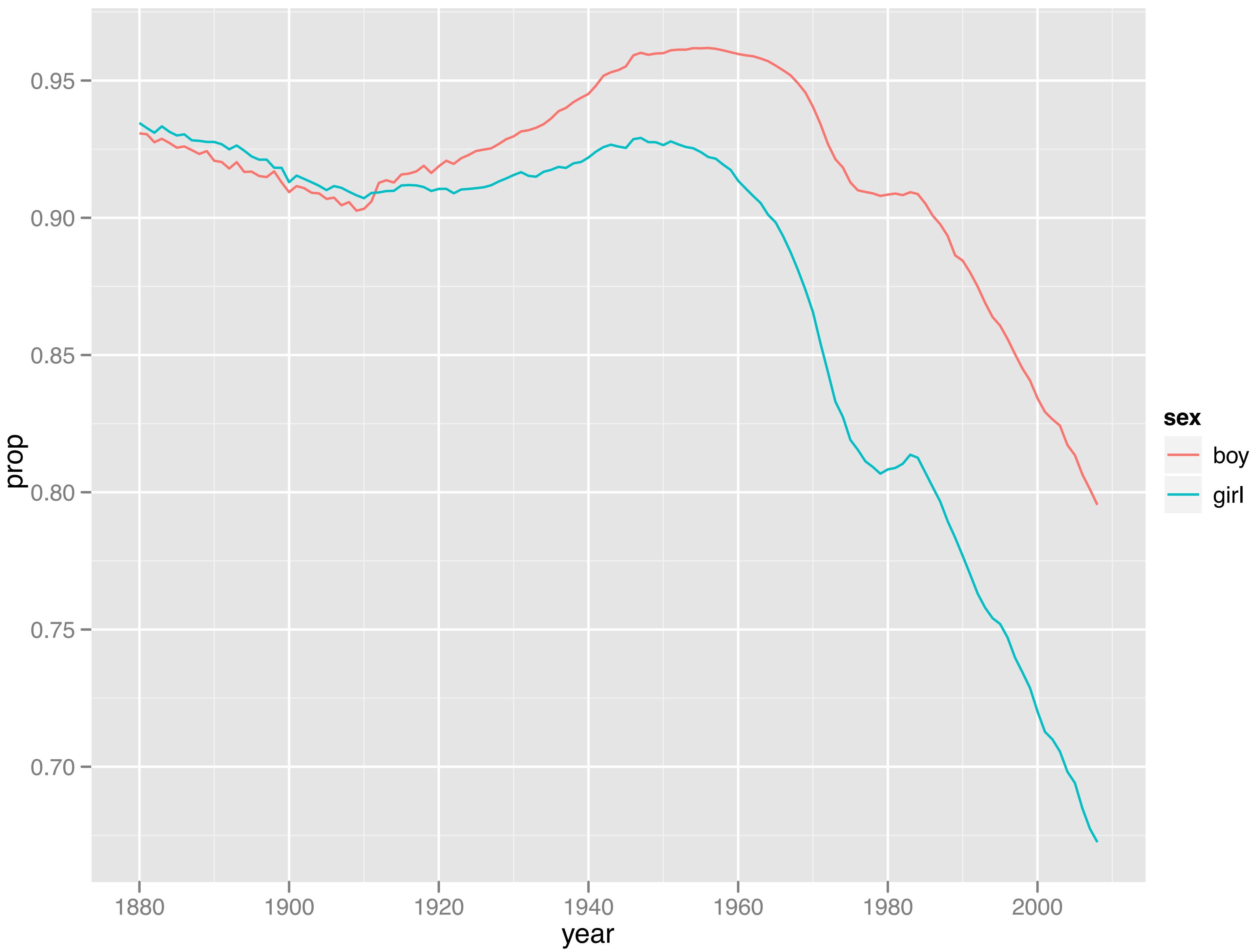


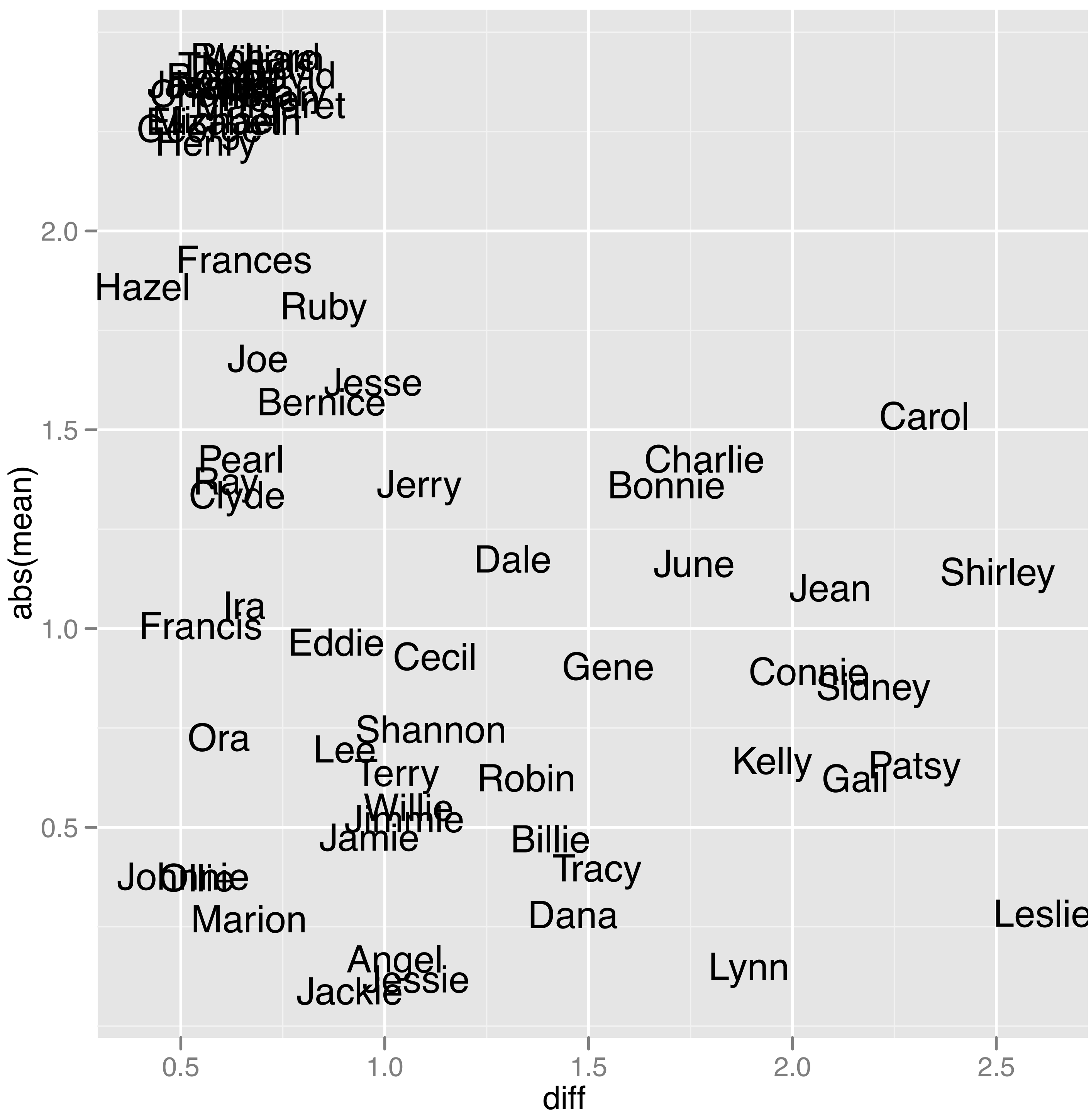
40
35
30
25
20
15
hwy
2
3
4
5
6
7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv


4000
3000
2000
1000
0
count
56
58
60
62
64
66
68
70
depth


price
15000
10000
5000
carat
1
2
3
4
5
count
500
1000
1500
2000


George
Georgia
Georgie
0.04
0.03
0.02
0.01
0.0025
0.0020
0.0015
0.0010
0.0005
4e−04
3e−04
2e−04
1e−04
1880
1900
1920
1940
1960
1980
2000
1880
1890
1900
1910
1920
1930
1940
1950
prop
year
sex
boy
girl


0.95
0.90
0.85
0.80
0.75
0.70
prop
1880
1900
1920
1940
1960
1980
2000
year
sex
boy
girl


Henry
Frances
Hazel
Ruby
Joe
Jesse
Bernice
Carol
Pearl
Charlie
Ray
Jerry
Bonnie
Clyde
Dale
June
Shirley
Jean
Ira
Francis
Eddie
Cecil
Gene
Connie
Sidney
Shannon
Ora
Lee
Kelly
Patsy
Terry
Robin
Gail
Willie
Jimmie
Jamie
Billie
Tracy
Johnnie
Ollie
Marion
Dana
Leslie
Angel
Jessie
Lynn
Jackie
2.0
1.5
1.0
0.5
abs(mean)
0.5
1.0
1.5
2.0
2.5
diff

Plotting basics

Learning a new language is hard!

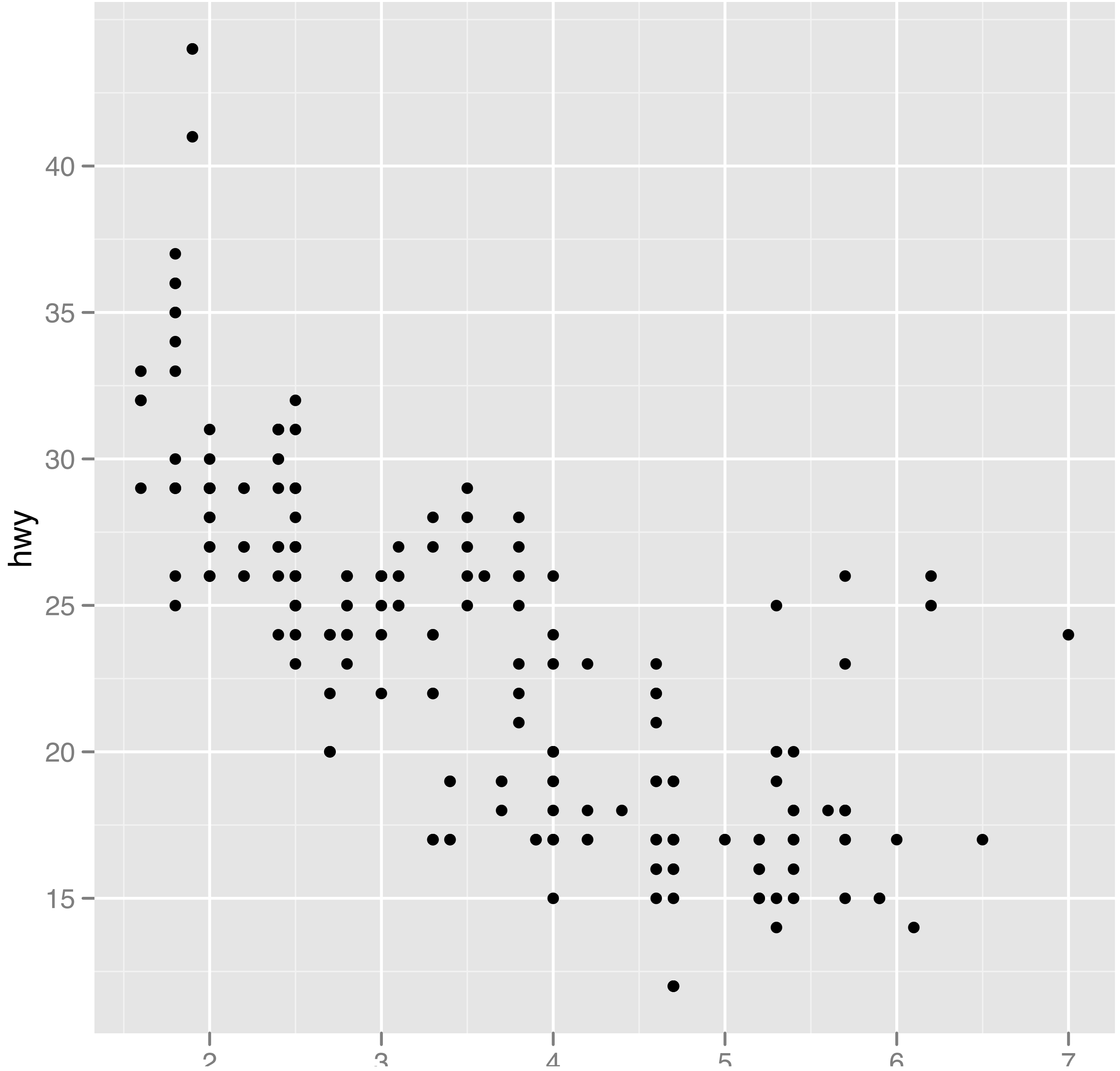
Scatterplot basics

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

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

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

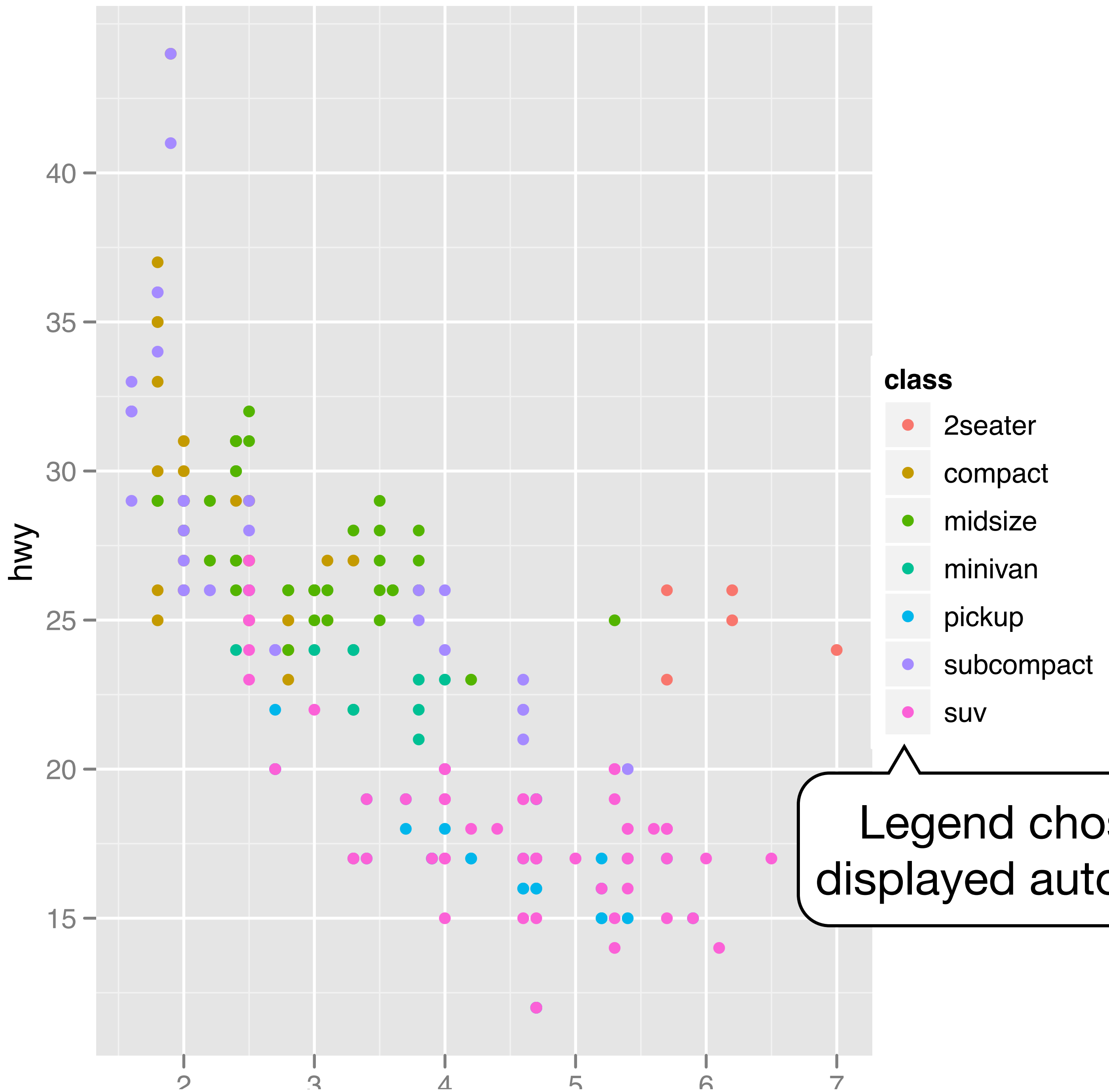
Always explicitly specify the data



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

Additional variables

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



Legend chosen and displayed automatically.

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

Your turn

Try mapping different variables to the colour, size, and shape aesthetics. Is there a difference between discrete and continuous variables? What happens when you use multiple aesthetics?

http://had.co.nz/courses/10-tokyo

Aside: workflow

Keep a copy of the slides open so that you can copy and paste the code.

For complicated commands, write them in the script editor and then copy and paste.

	Discrete	Continuous
Colour	Rainbow of colours	Gradient from red to blue
Size	Discrete size steps	Linear mapping between radius and value
Shape	Different shape for each	Doesn’t work

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 ~ .)
```

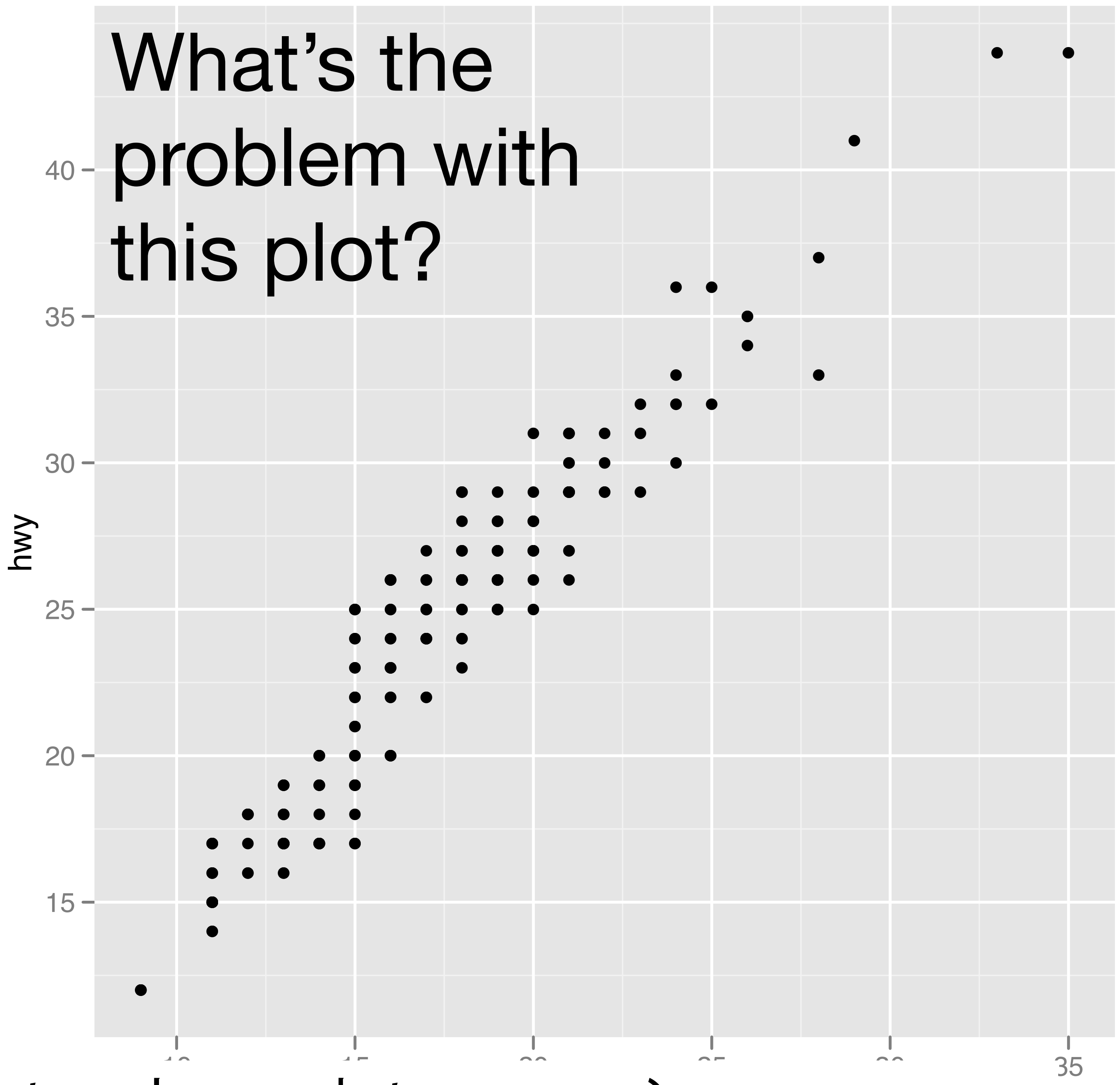
```
qplot(displ, hwy, data = mpg) +
facet_grid(drv ~ cyl)
```

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

Summary

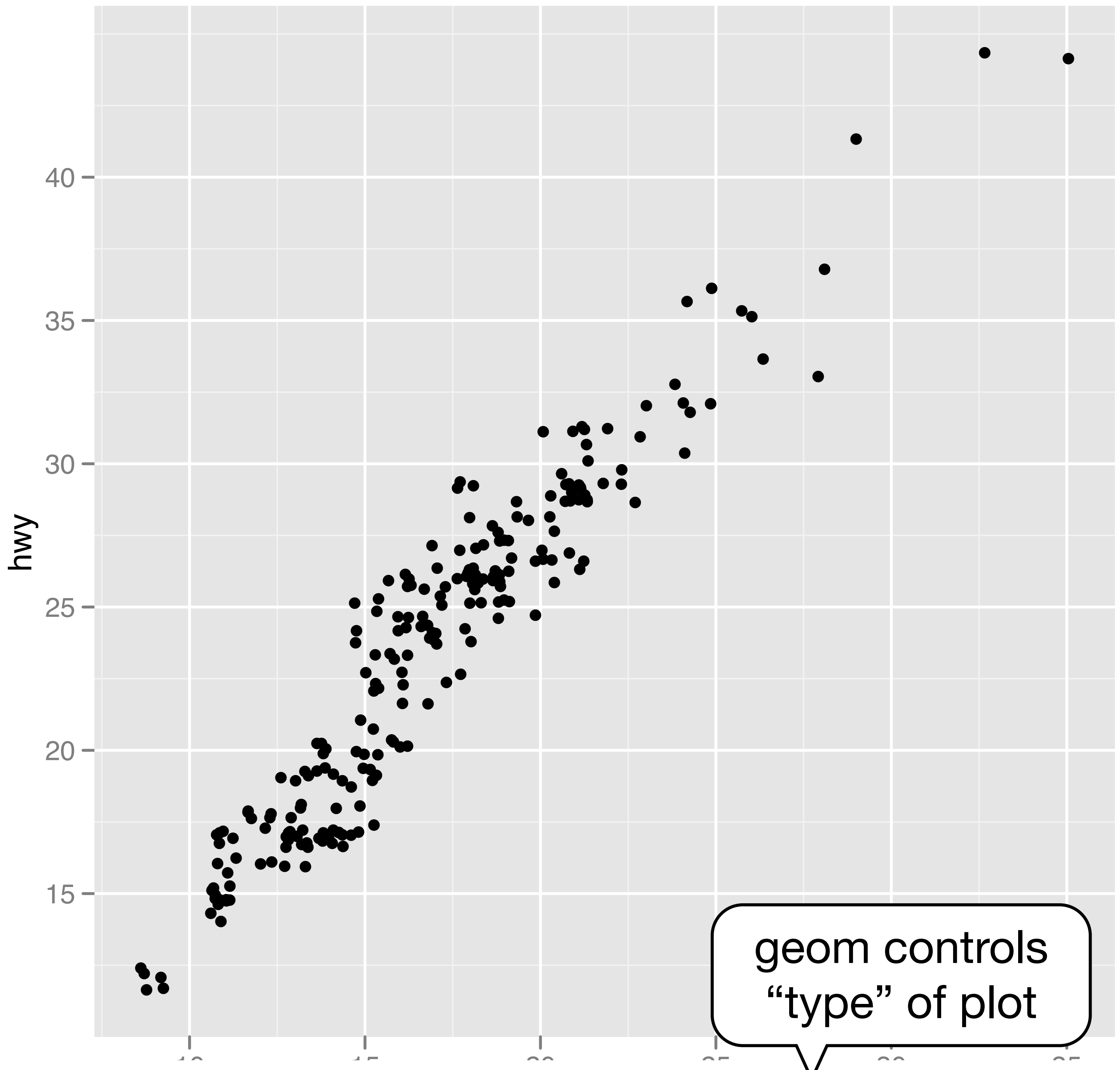
`facet_grid()`: 2d grid, rows ~ cols, . for no split

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



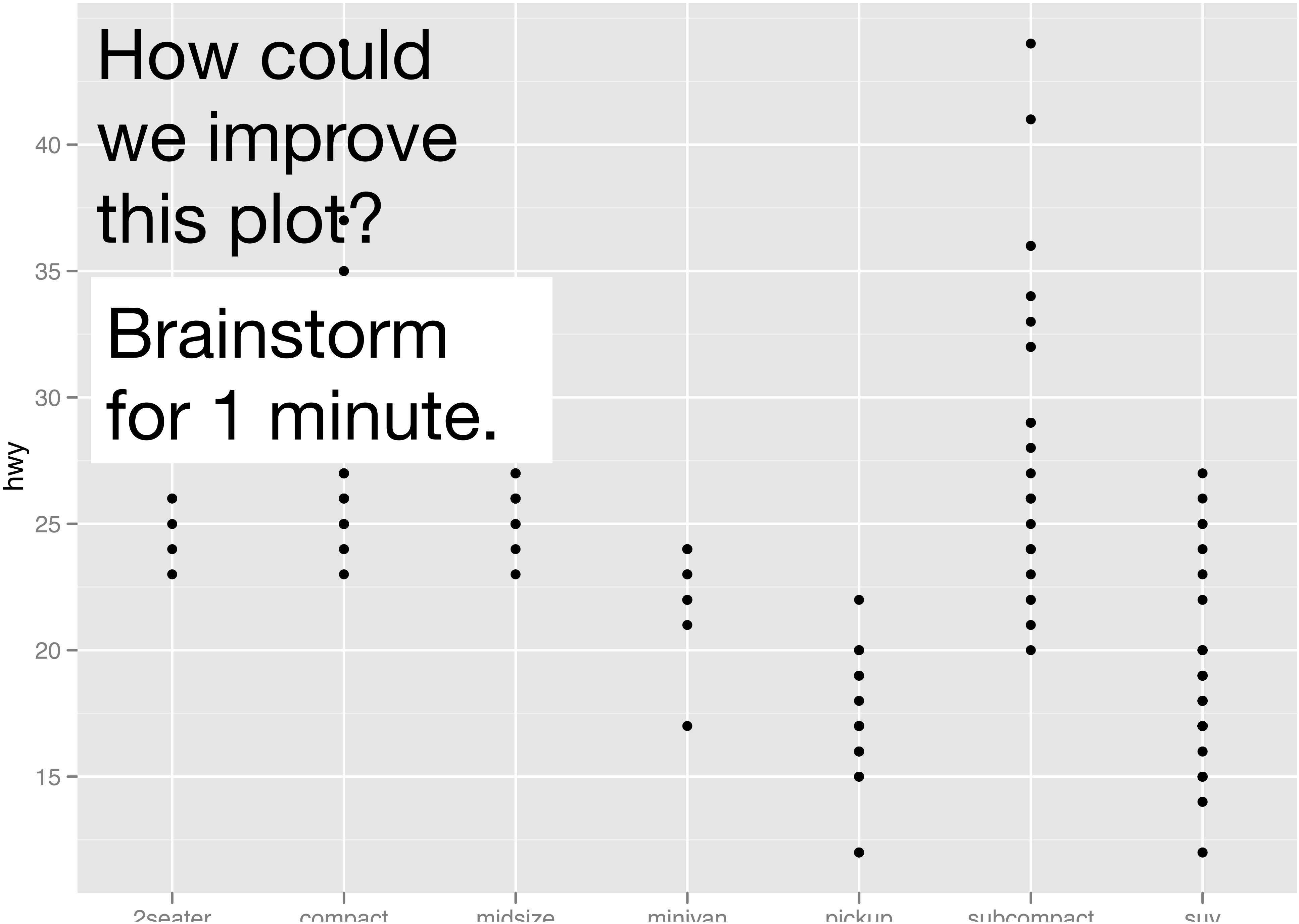
What’s the problem with this plot?

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



geom controls
“type” of plot

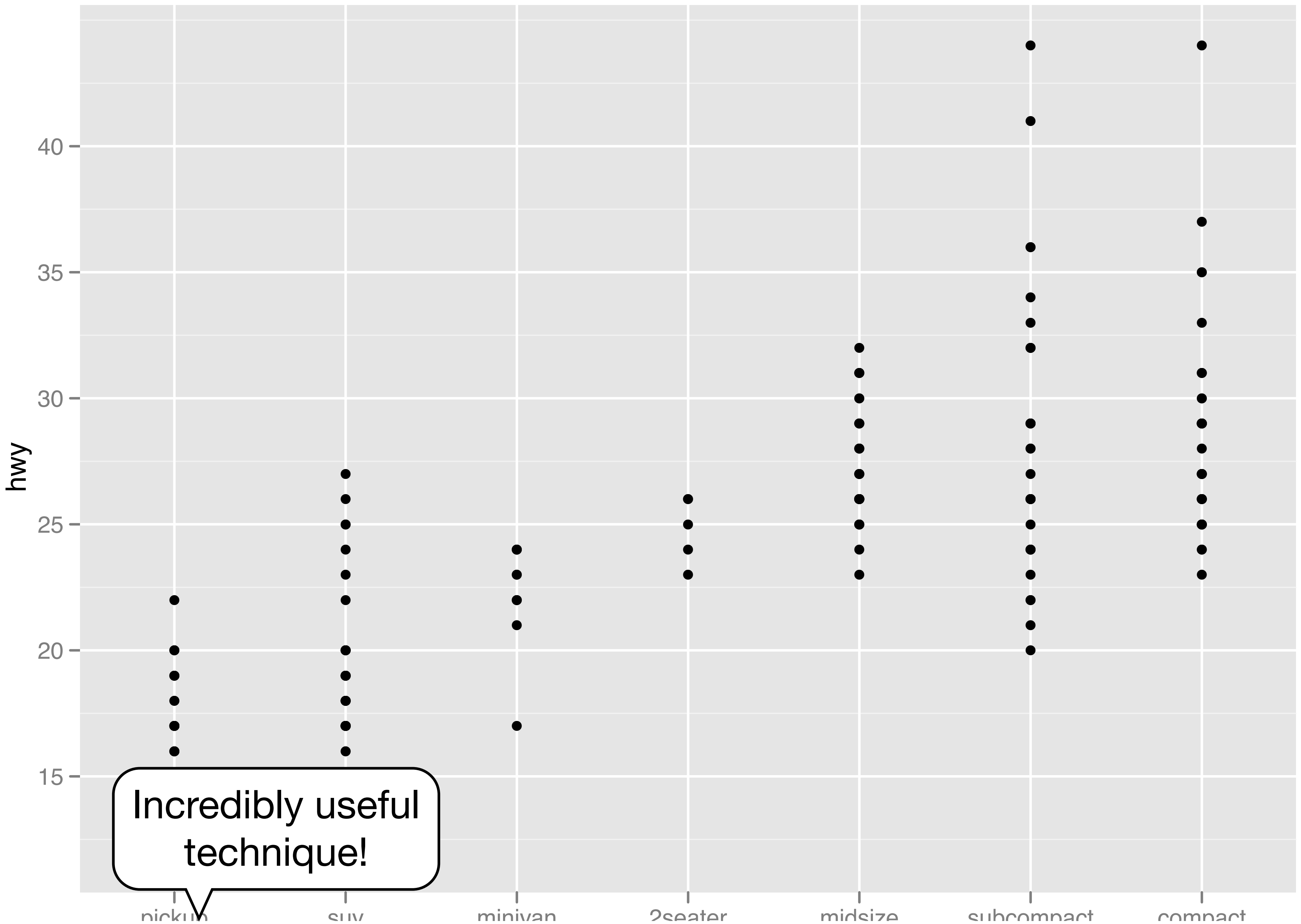
```
qplot(cty, hwy, data = mpg, geom = ”jitter”)
```



How could we improve this plot?

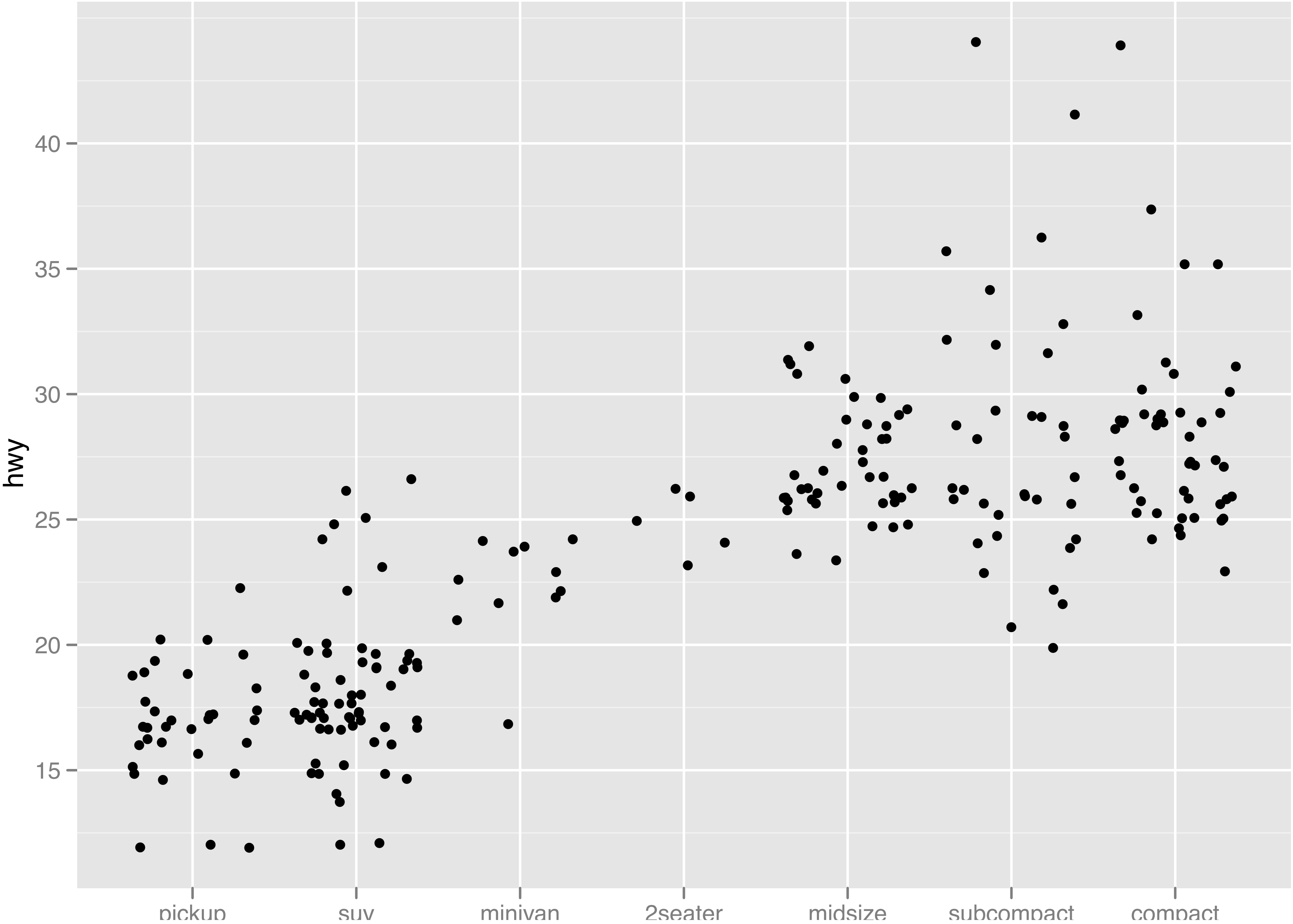
Brainstorm for 1 minute.

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

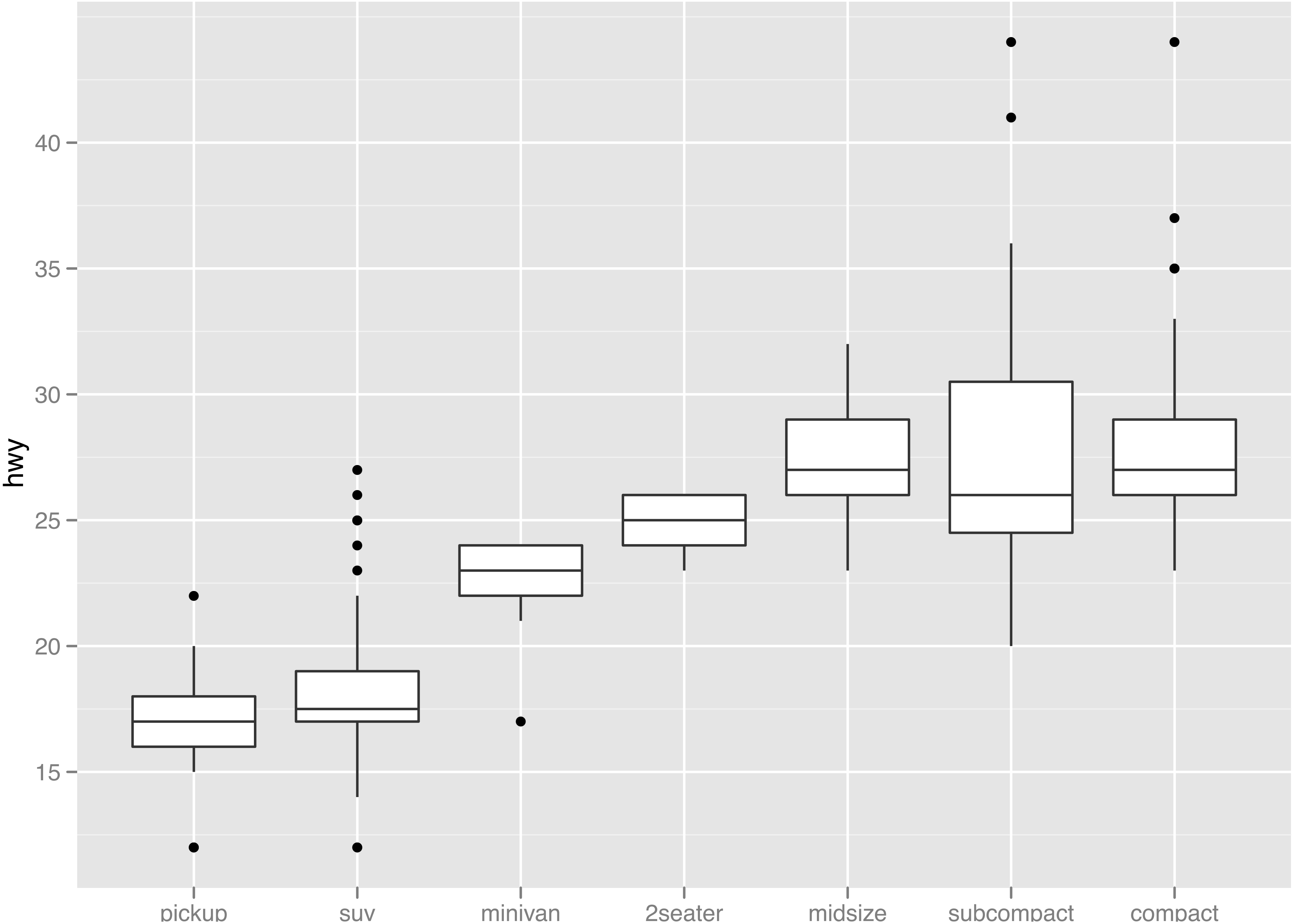


Incredibly useful technique!

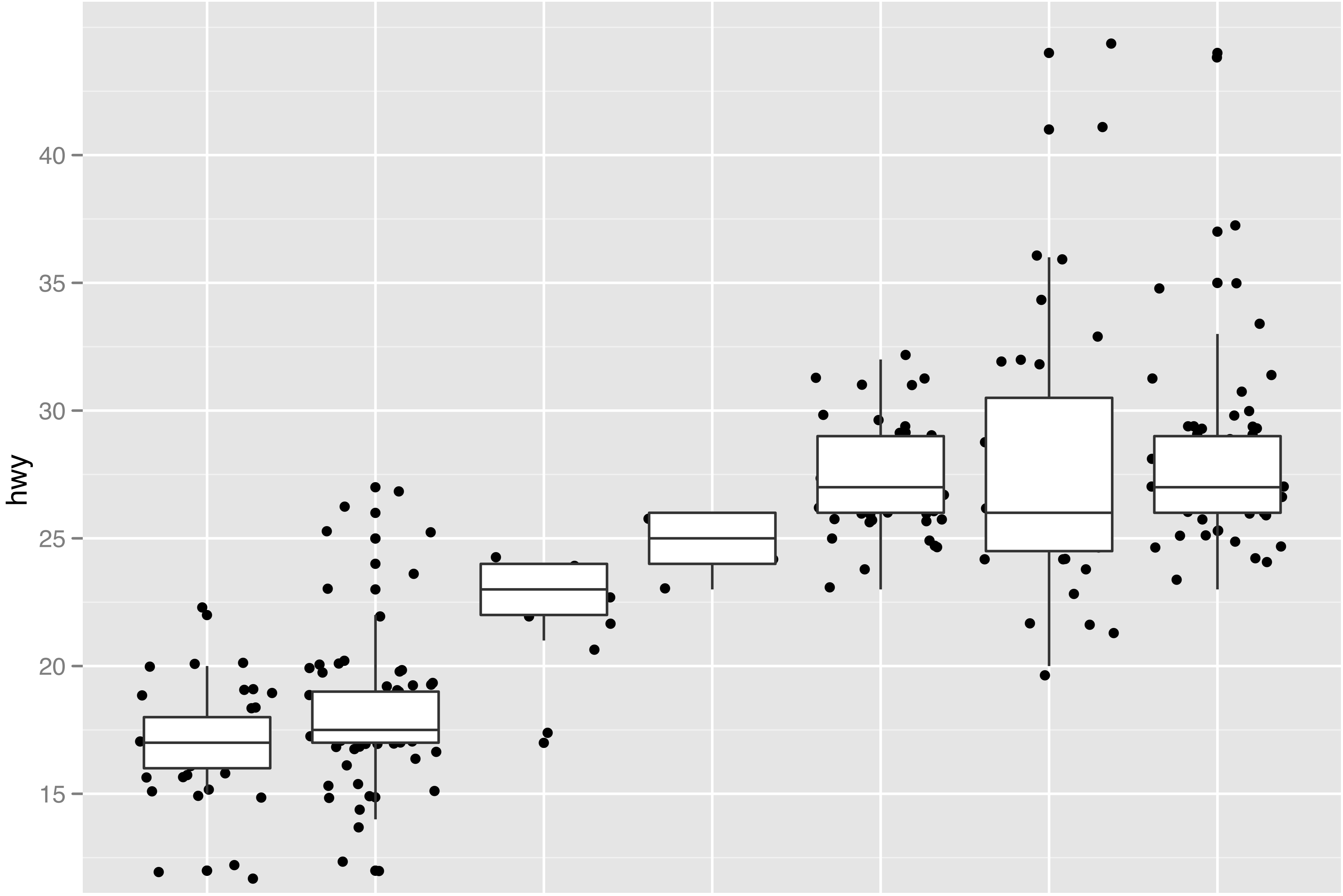
```
qplot(reorder(class, hwy), hwy, data = mpg)
```



```
qplot(reorder(class, hwy), hwy, data = mpg, geom = "jitter")
```



```
qplot(reorder(class, hwy), hwy, data = mpg, geom = ”boxplot”)
```



```
qplot(reorder(class, hwy), hwy, data = mpg,
  geom = c("jitter", "boxplot"))
```

Your turn

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

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