Intro to ggplot2

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

November 2010

cc BY NC

HELLO

my name is

Hadley

had.co.nz/courses/

10-tokyo

Outline

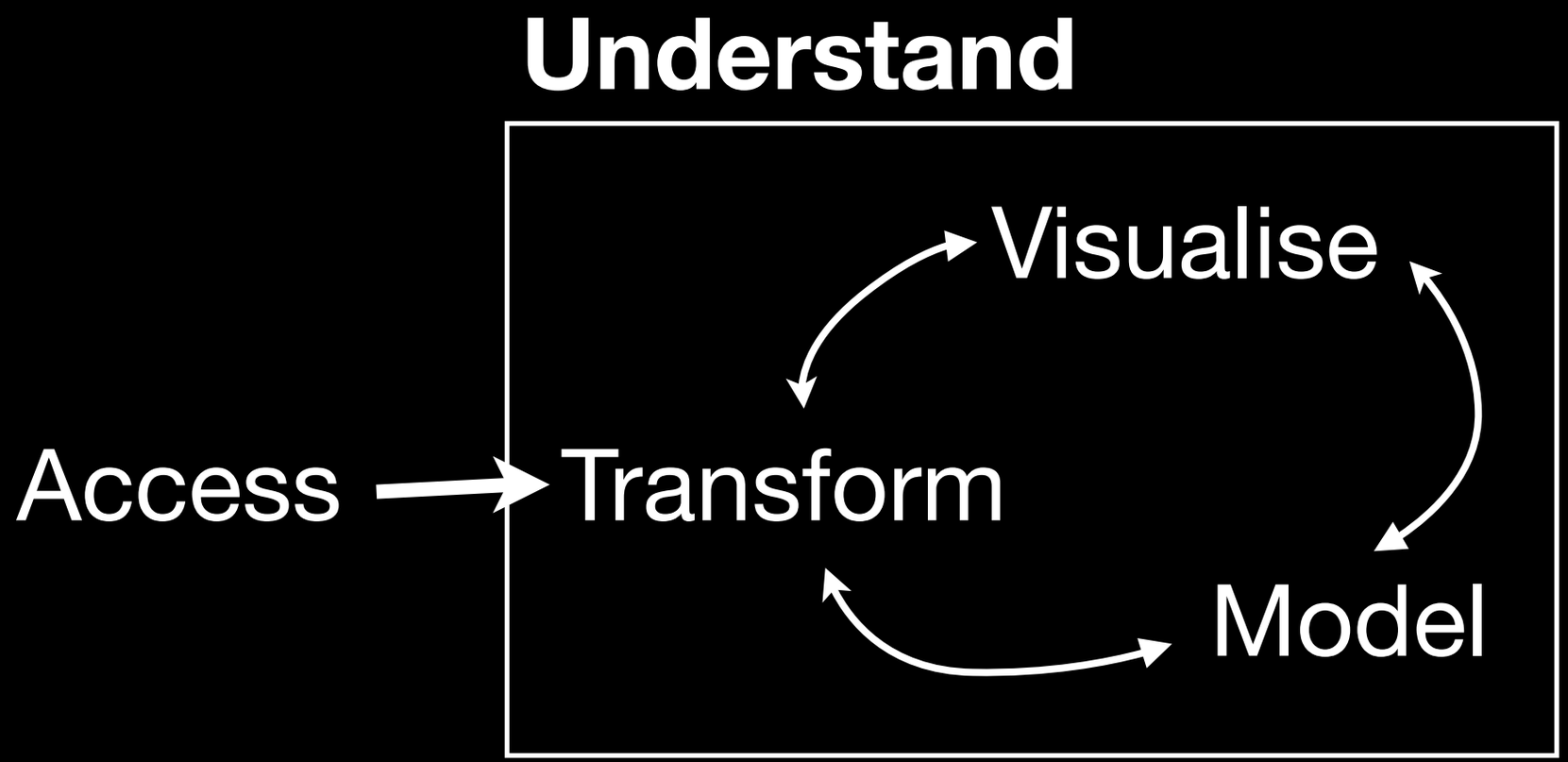
Data analysis is the process by which data becomes understanding, knowledge and insight

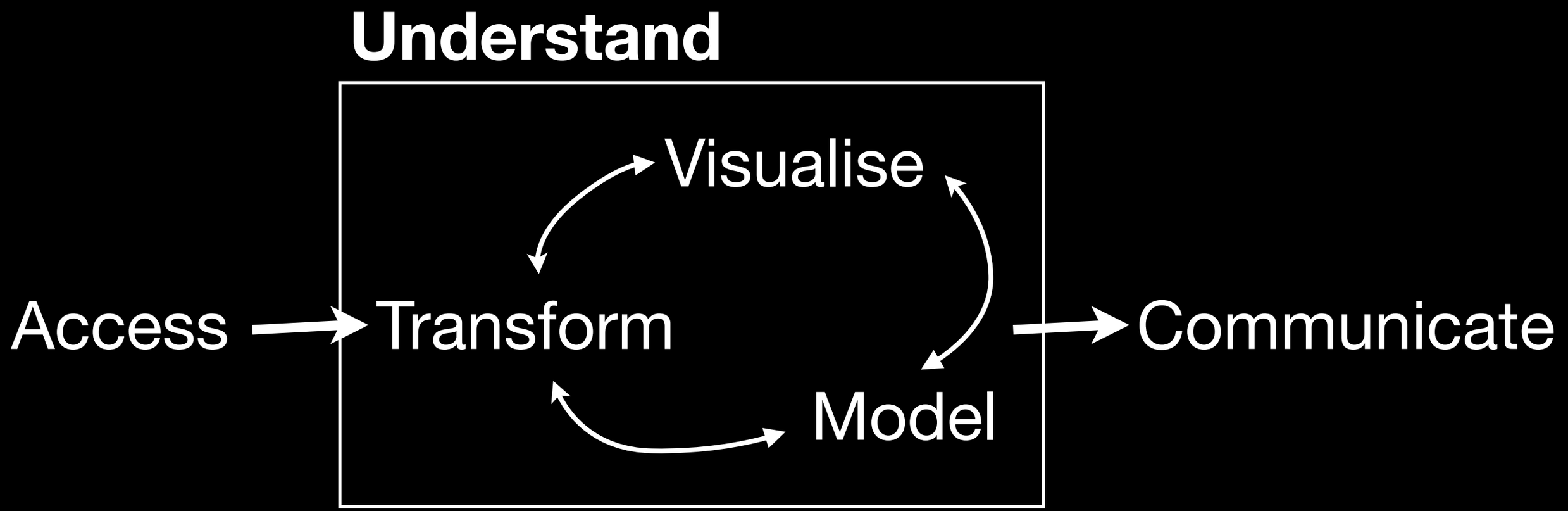
Tuesday, 23 November 2010

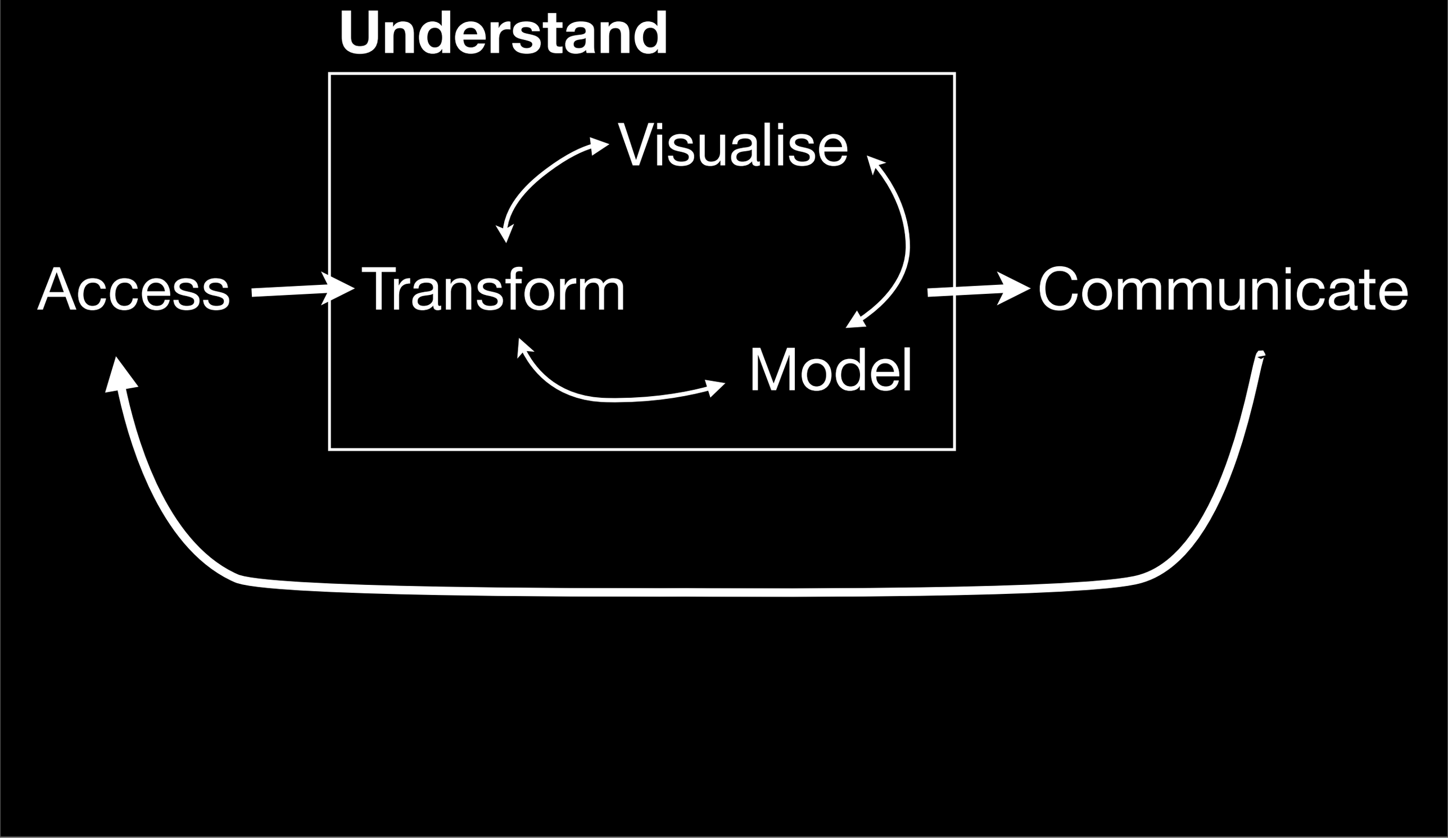
Access

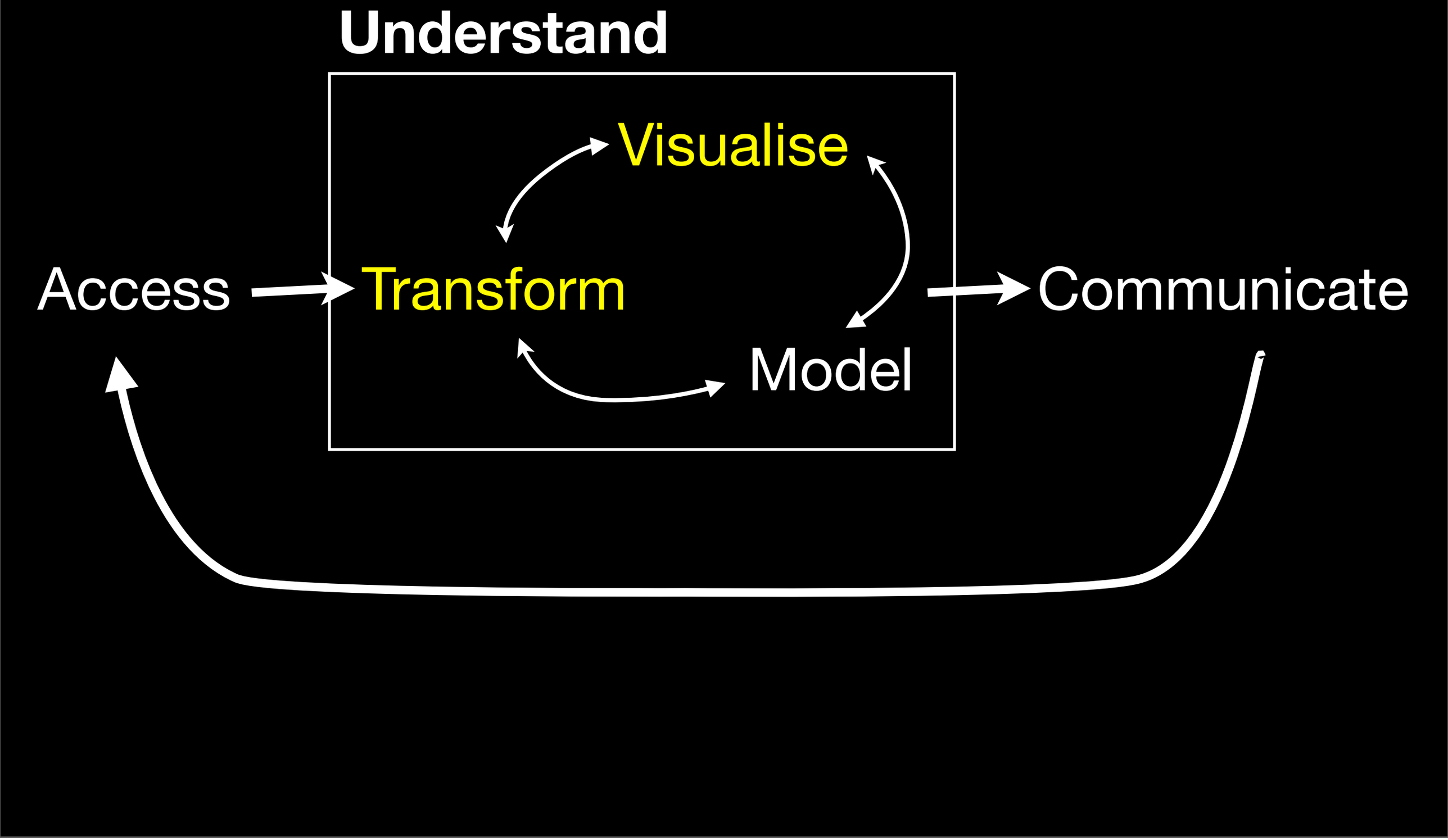
Understand

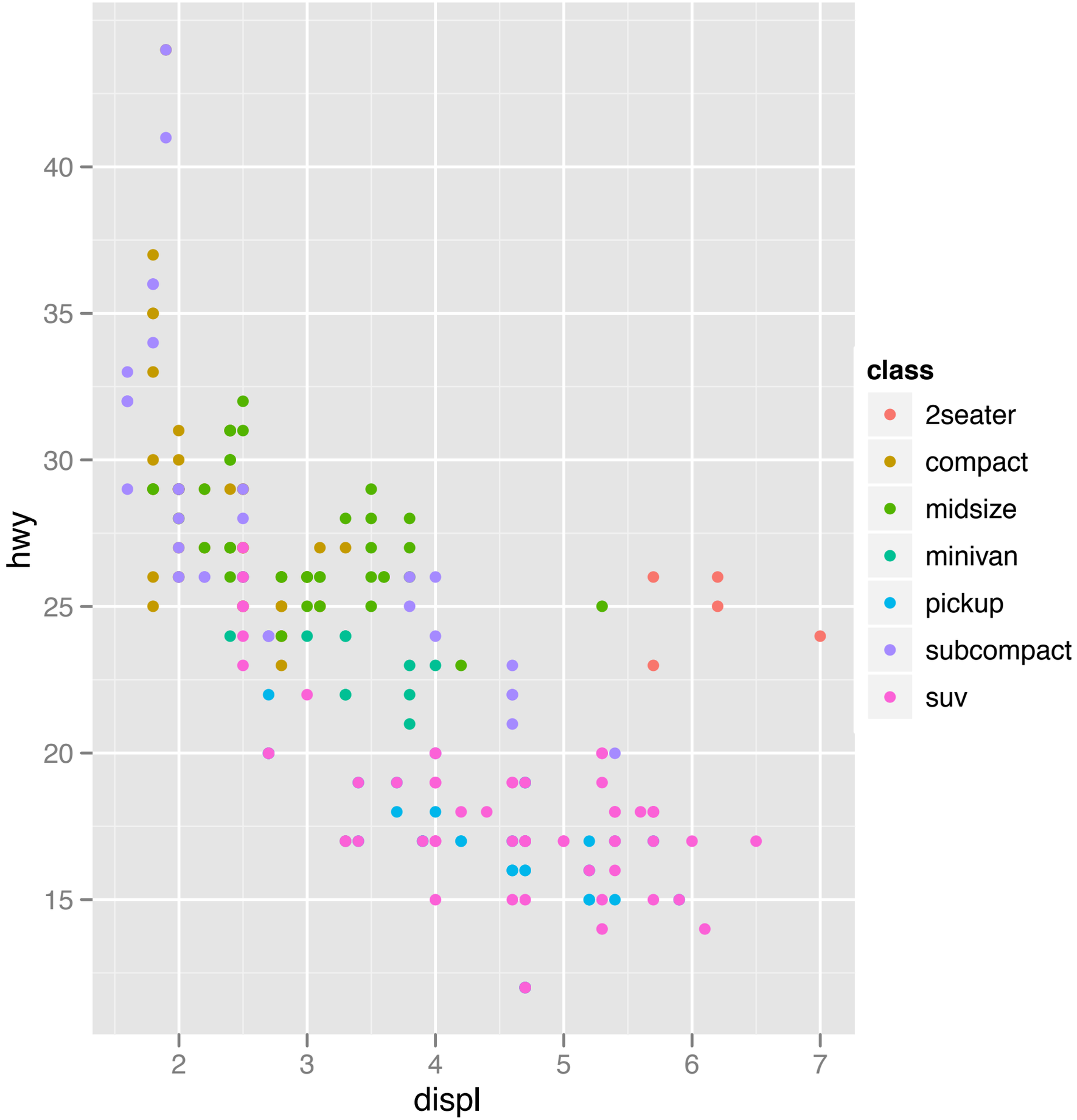
Access →

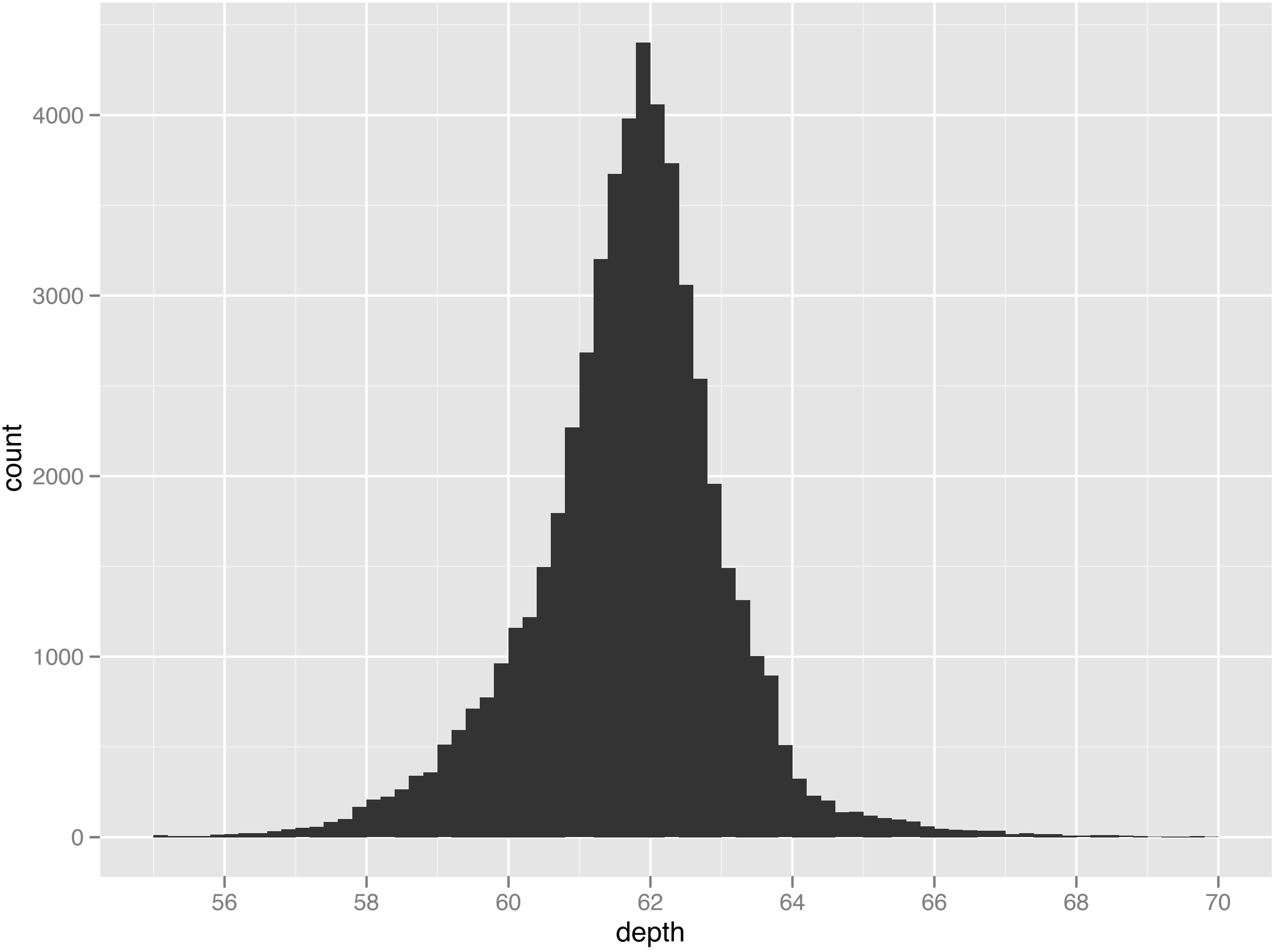

Understand
Visualise
Access
Transform
Model

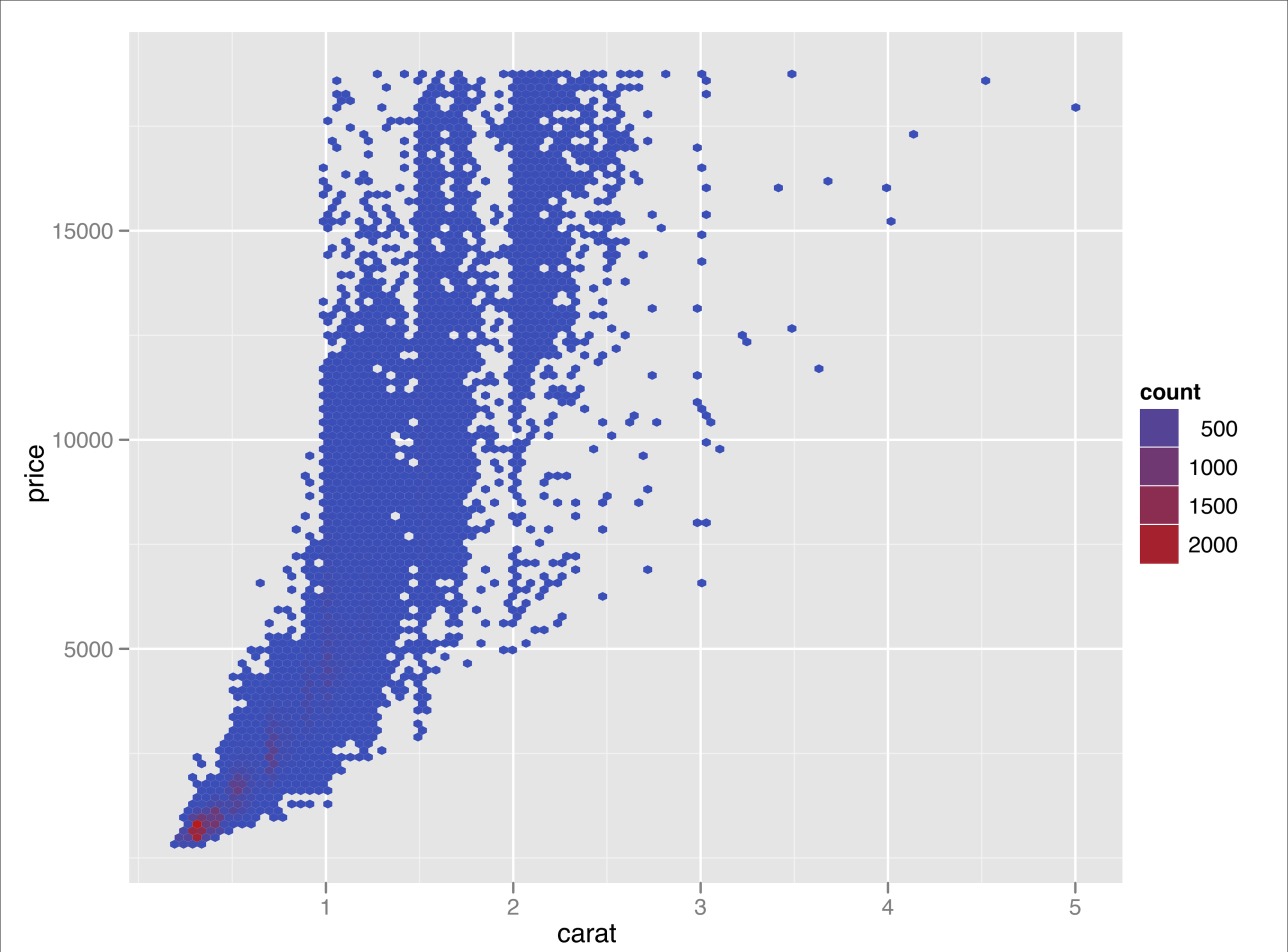

Understand
Visualise
Access
Transform
Communicate
Model


Understand
Visualise
Access
Transform
Communicate
Model

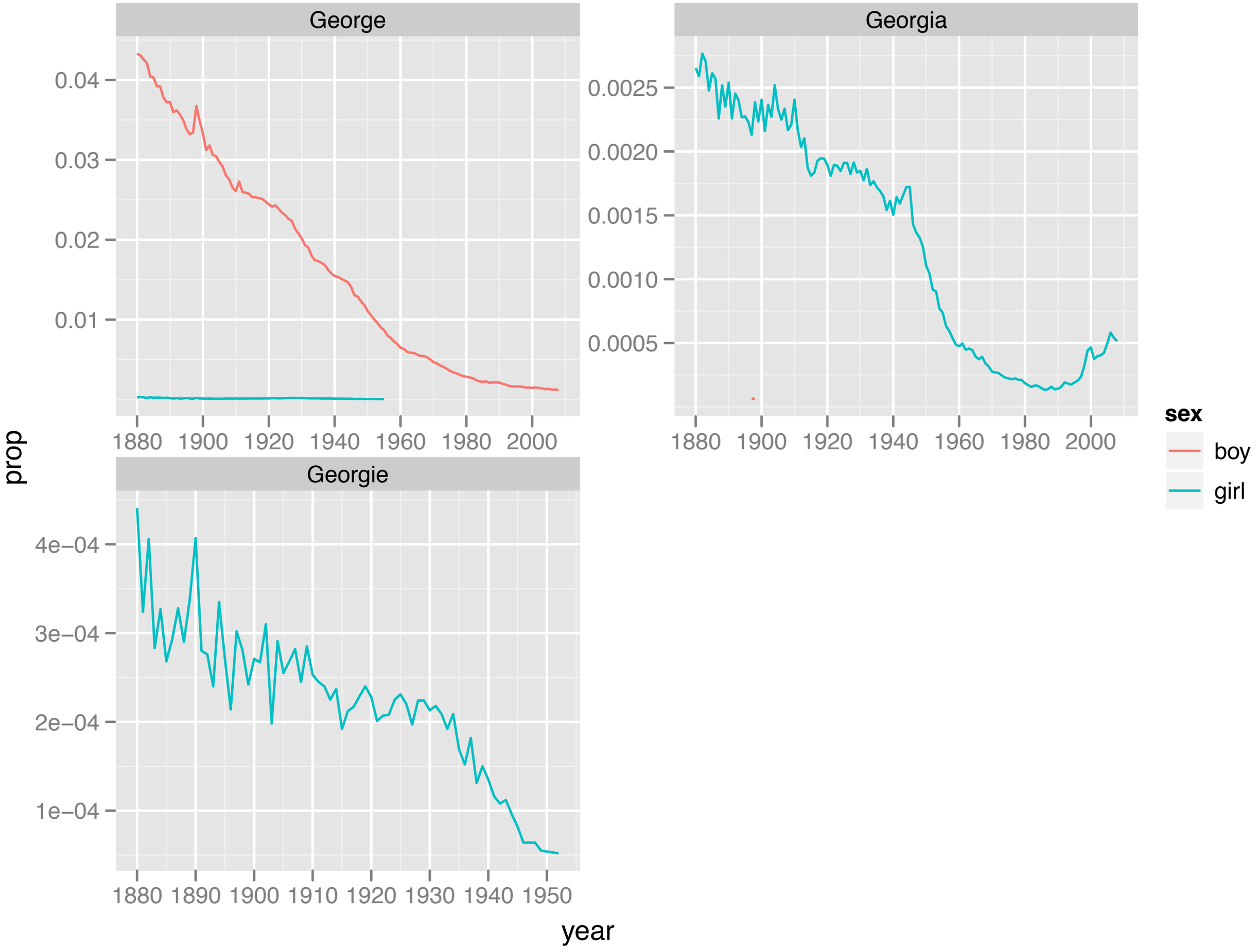

Understand
Visualise
Access
Transform
Communicate
Model

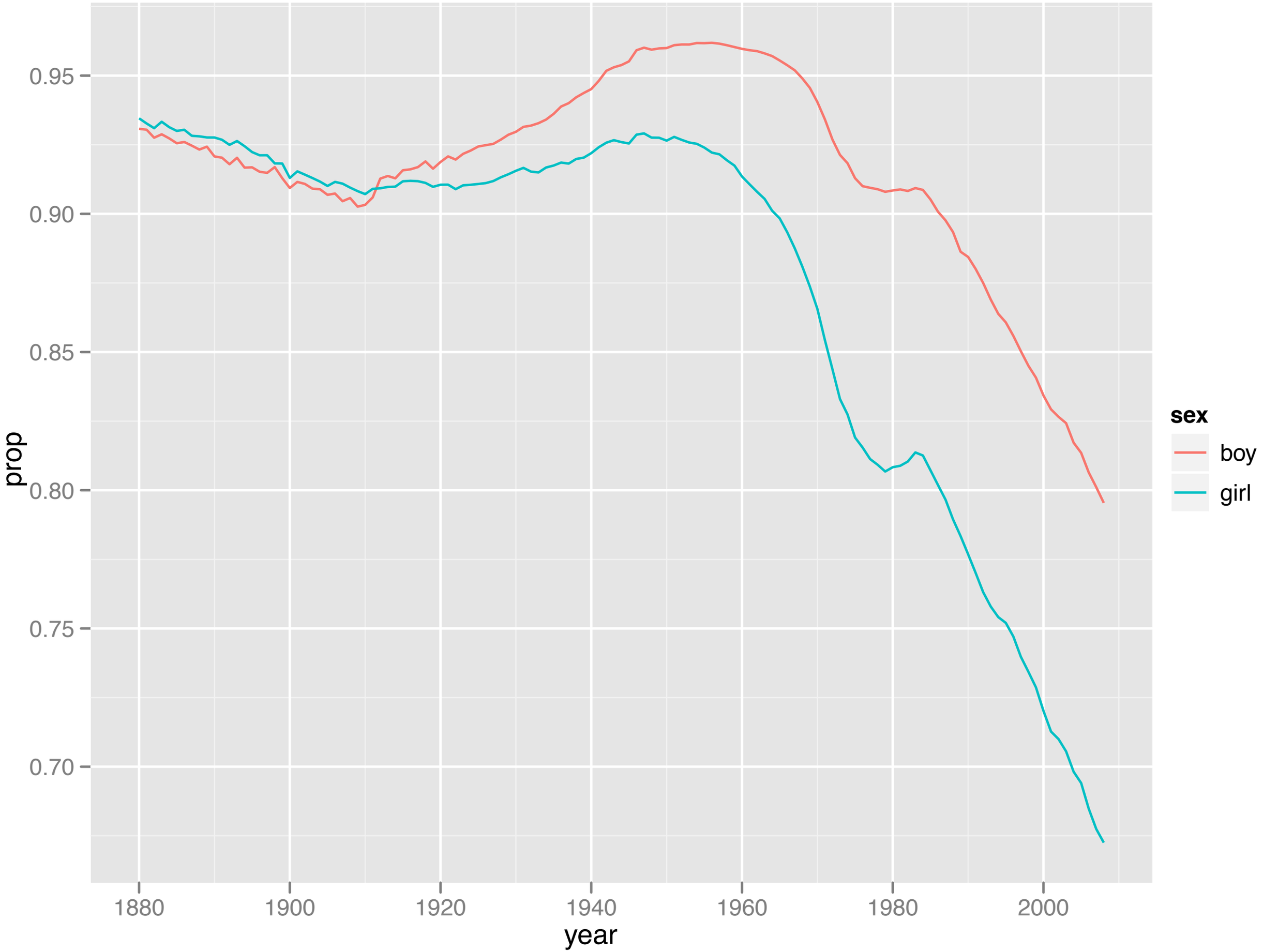

hwy
40
35
30
25
20
15
2
3
4
5
6
7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv

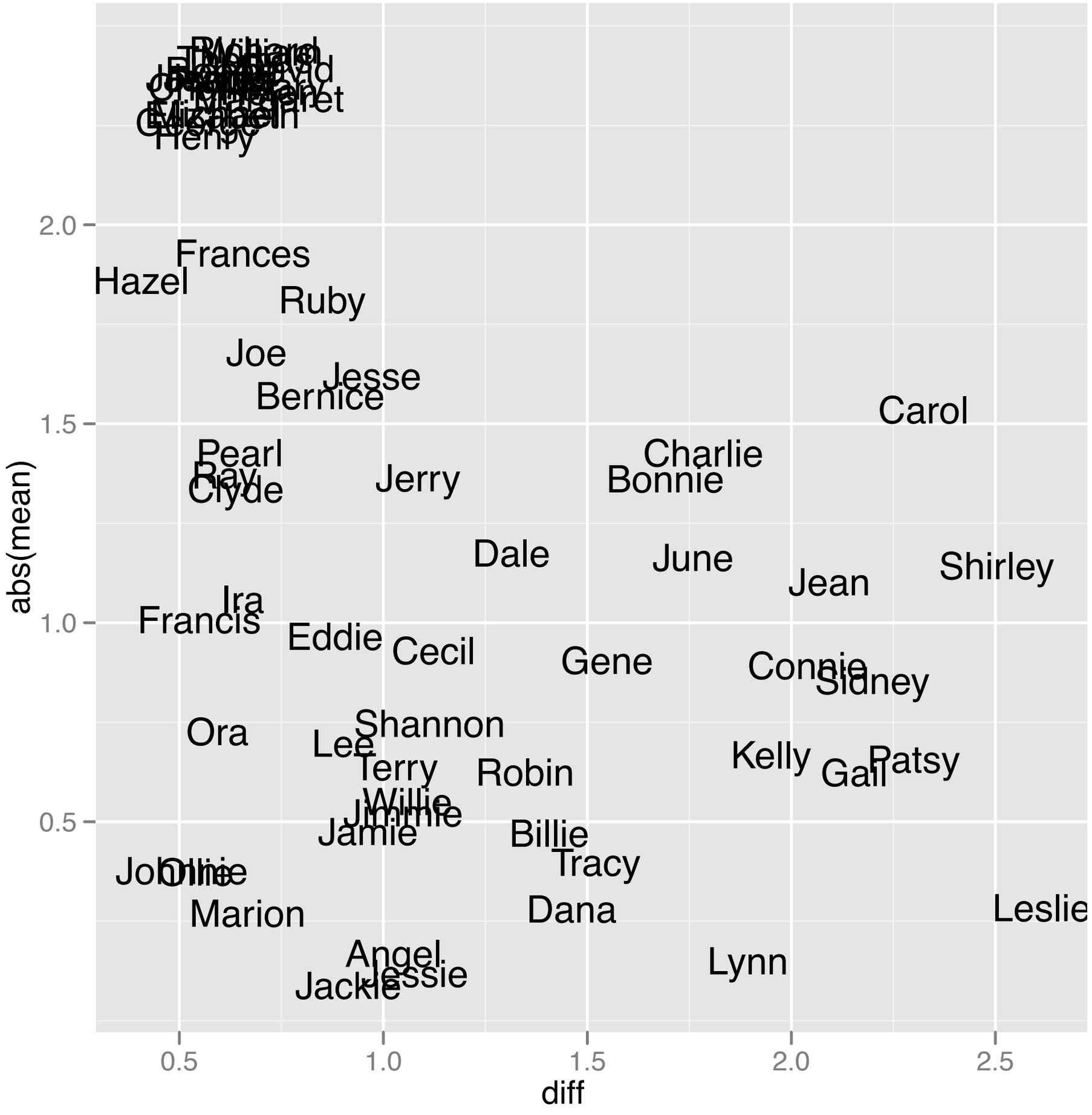

count
4000
3000
2000
1000
0
56
58
60
62
64
66
68
70
depth



Tuesday, 23 November 2010


George
Georgia
Georgie
prop
year
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
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


Frances
Hazel
Ruby
Joe
Jesse
Bernice
Carol
Pearl
Ray
Clyde
Jerry
Charlie
Bonnie
Dale
June
Jean
Shirley
Ira
Francis
Eddie
Cecil
Gene
Connie
Sidney
Ora
Shannon
Lee
Terry
Robin
Kelly
Gail
Patsy
Willie
Jimmie
Jamie
Billie
Tracy
Marion
Dana
Leslie
Angel
Jessie
Jackie
Lynn
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!

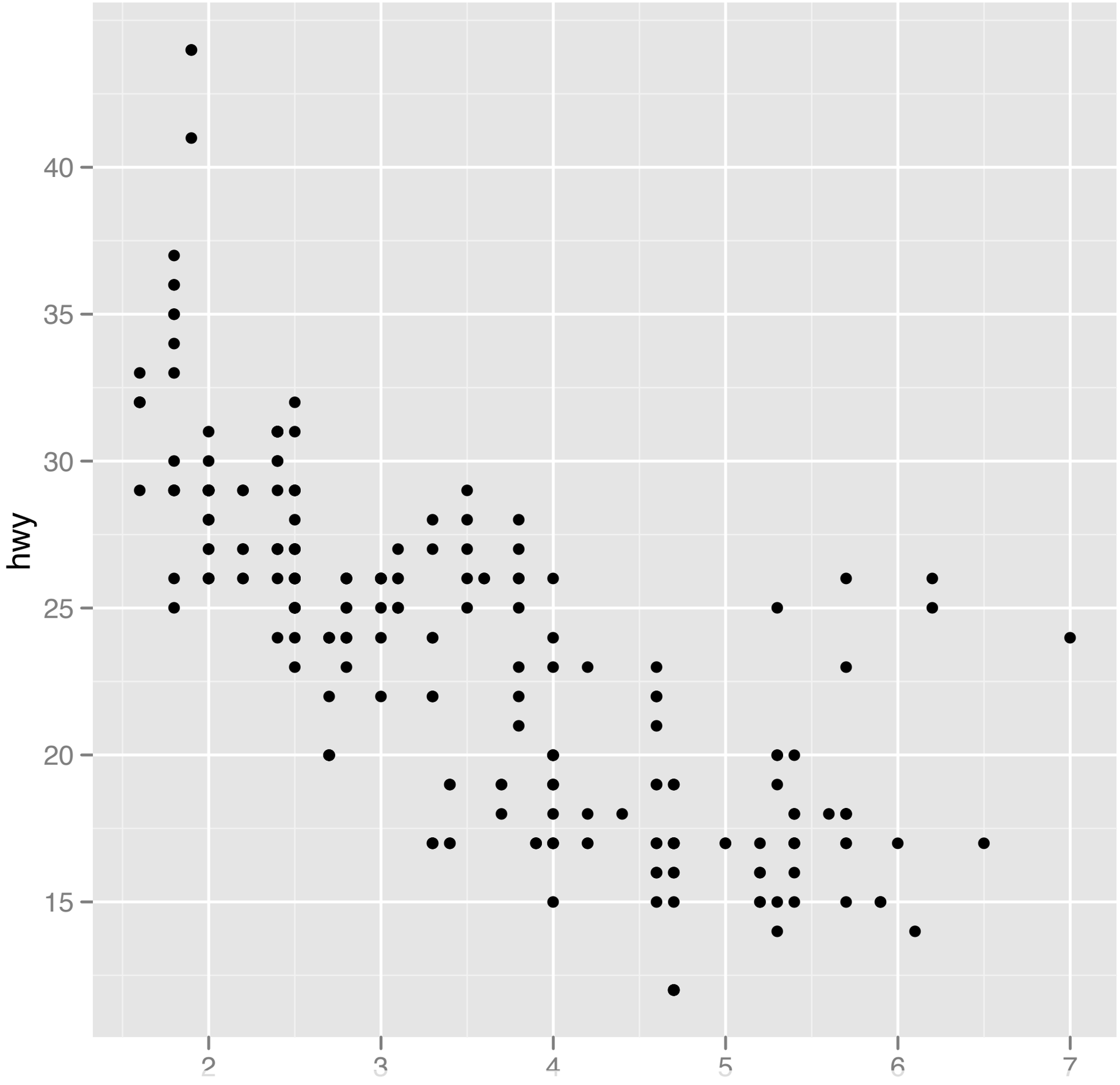
Scatterplot basics

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

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

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

Scatterplot basics

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

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

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

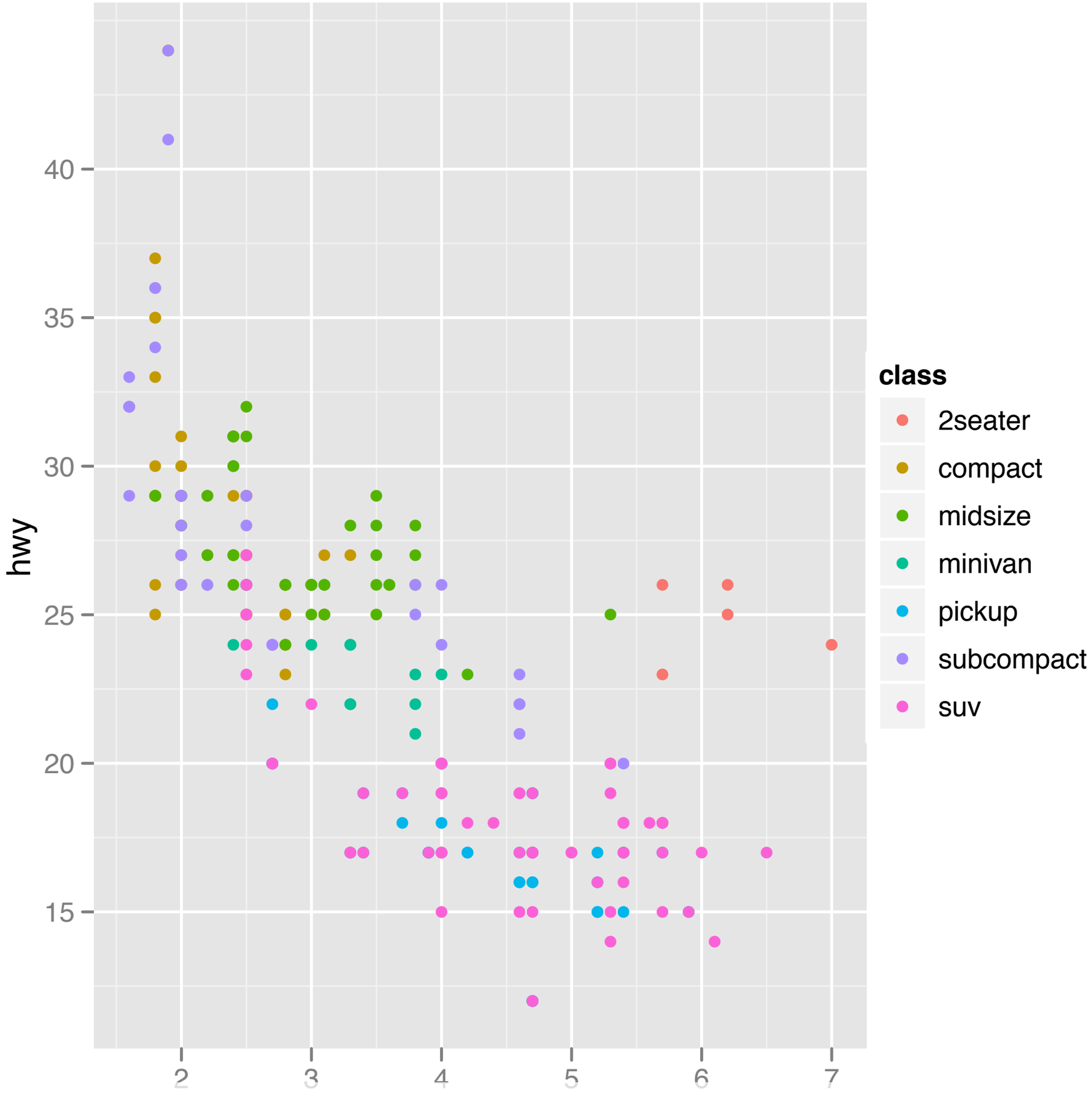
Always explicitly specify the data



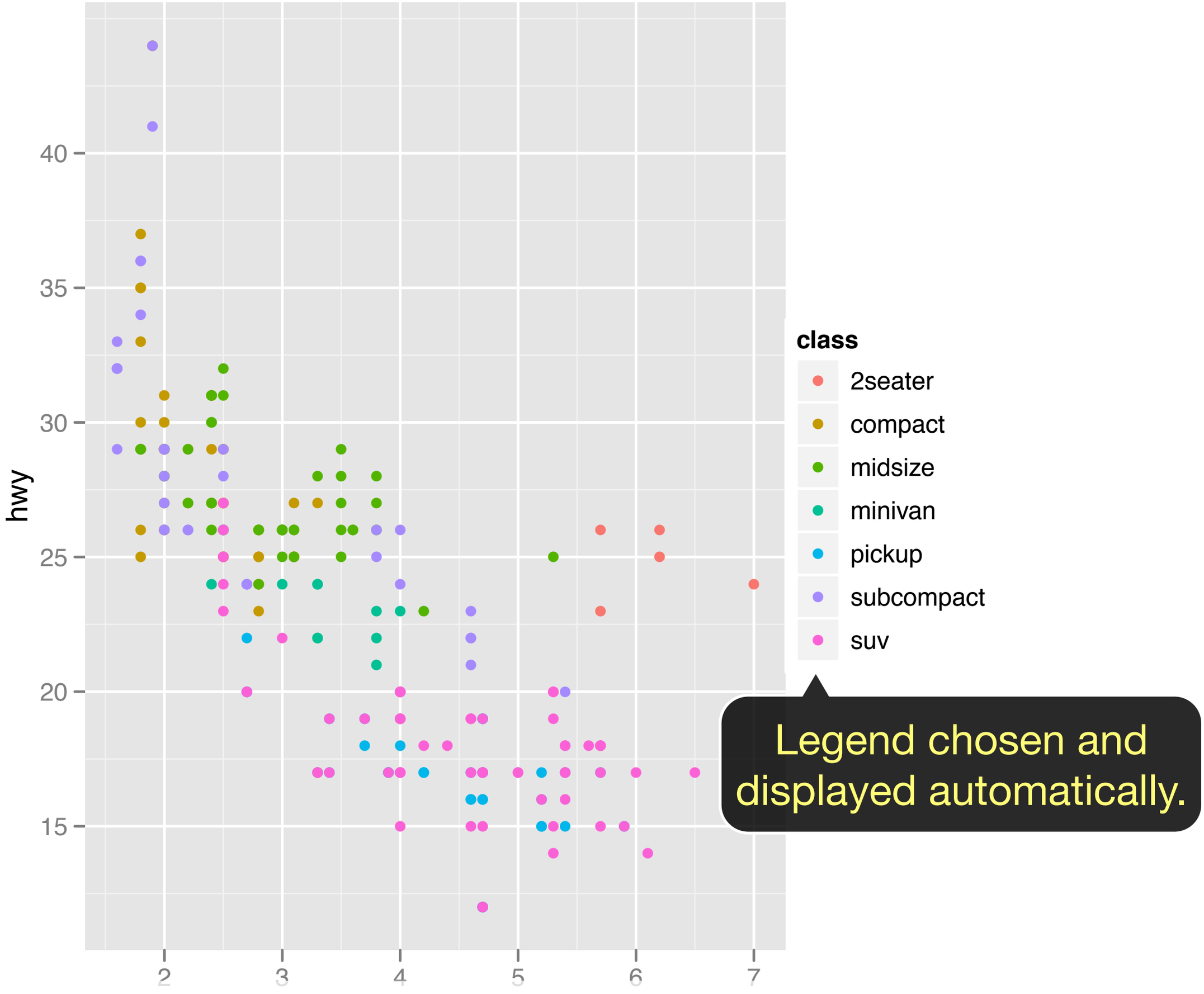

```
qplot(displ, hwy, data = mpg)
```

Additional variables

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




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




```
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

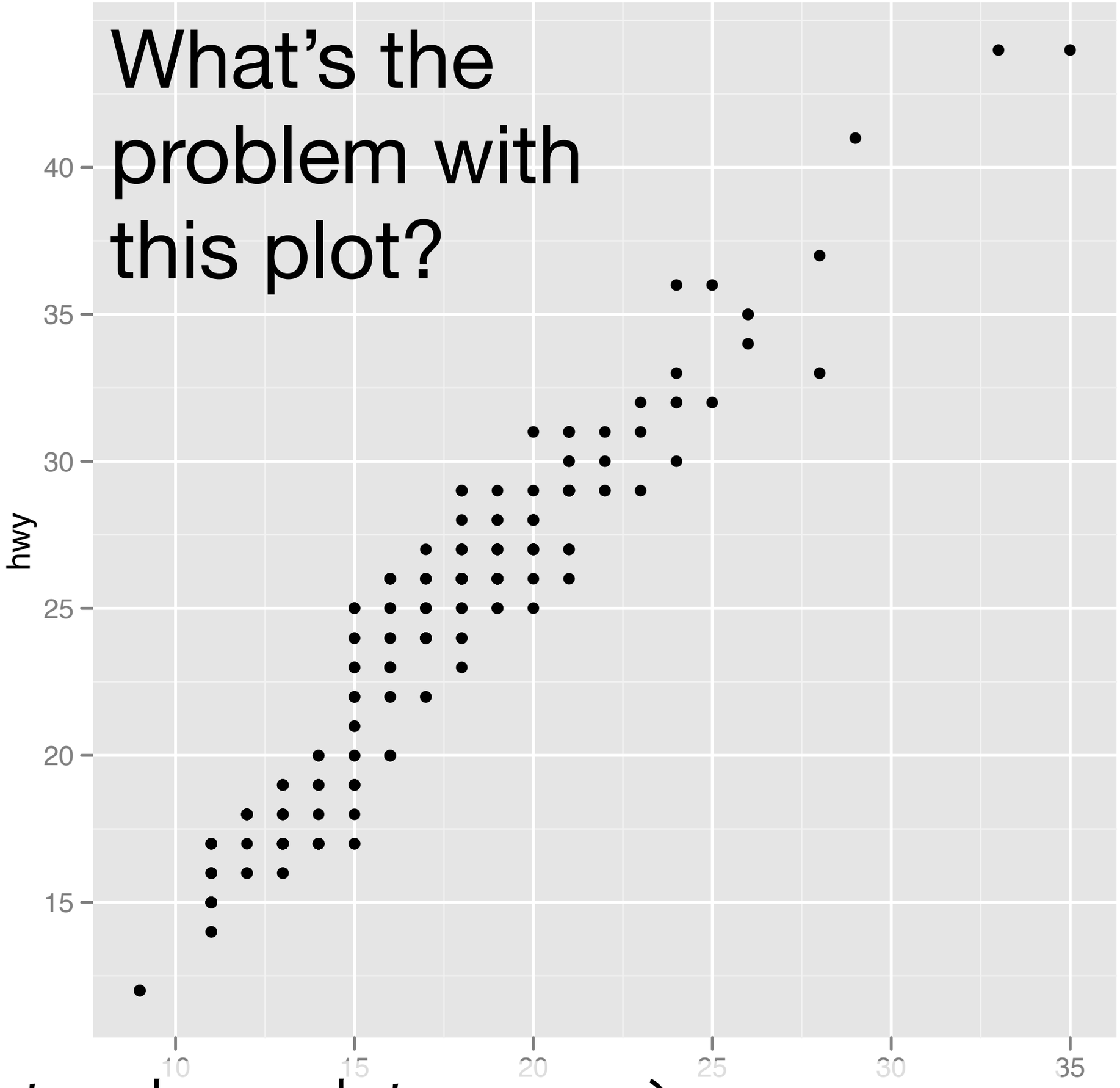
```
qplot(displ, hwy, data = mpg) +
facet_grid(. ~ cyl)

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

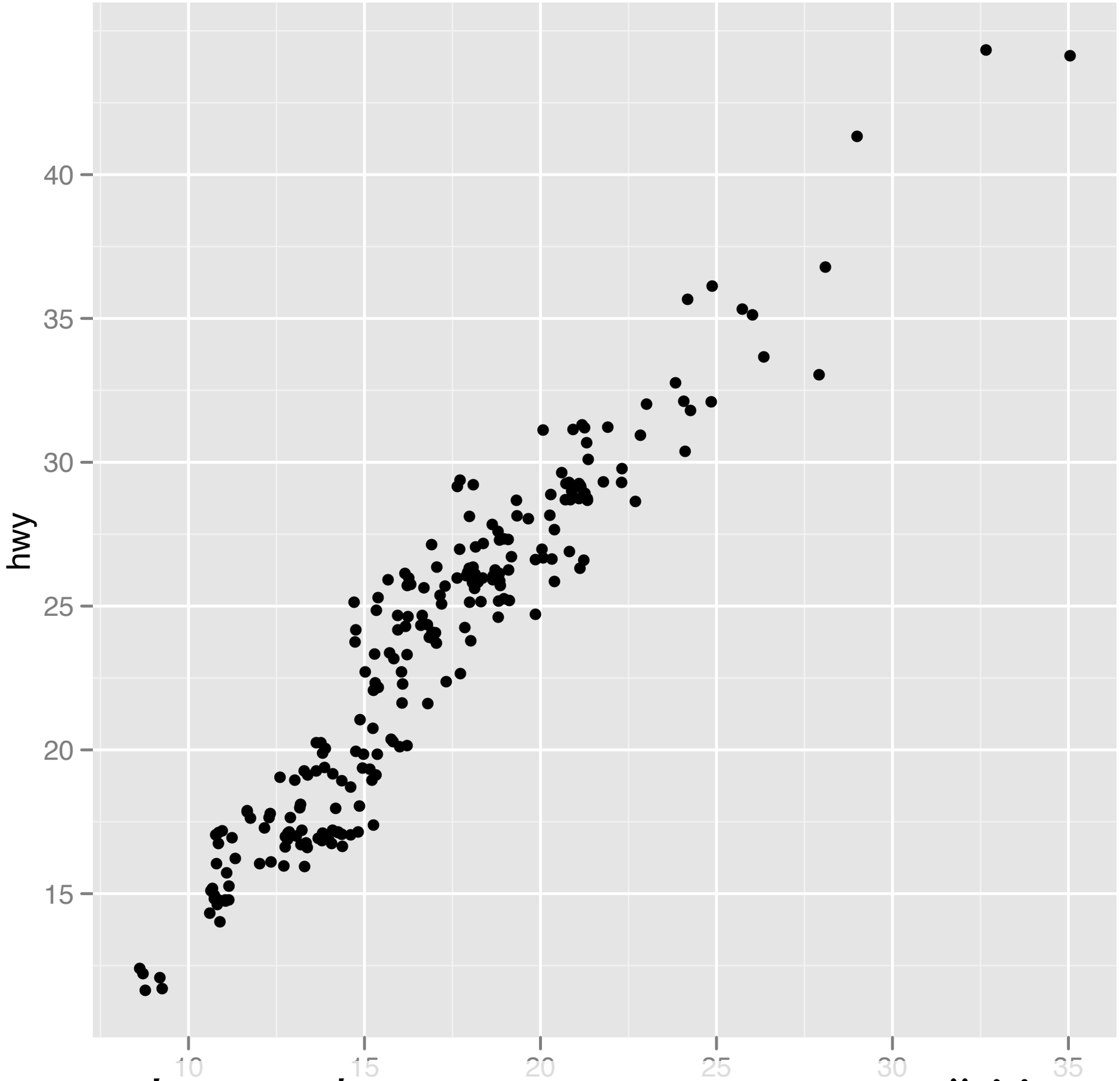
`facet_grid()`: 2d grid, rows ~ cols, . for no split

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

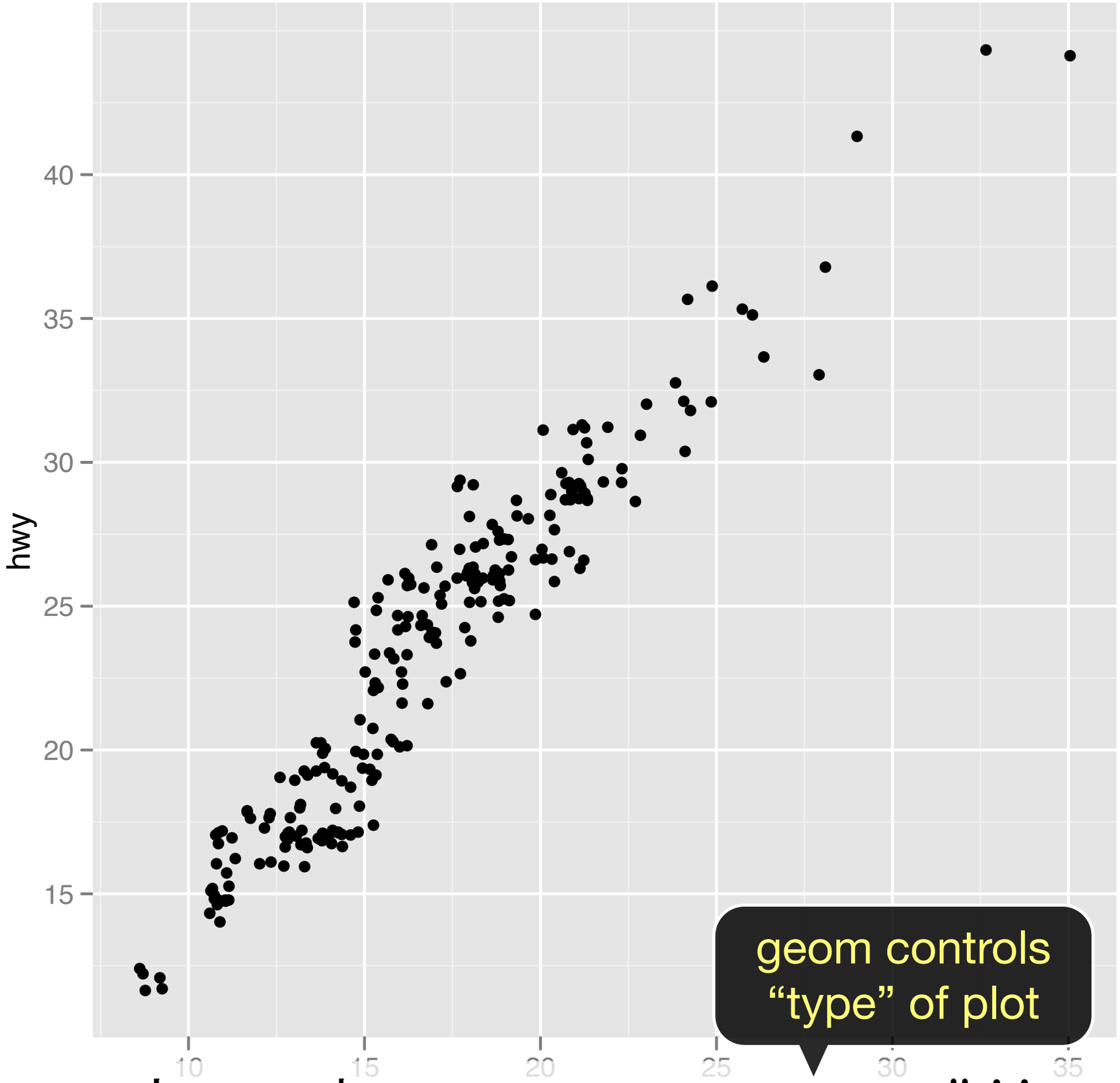
What's the problem with this plot?



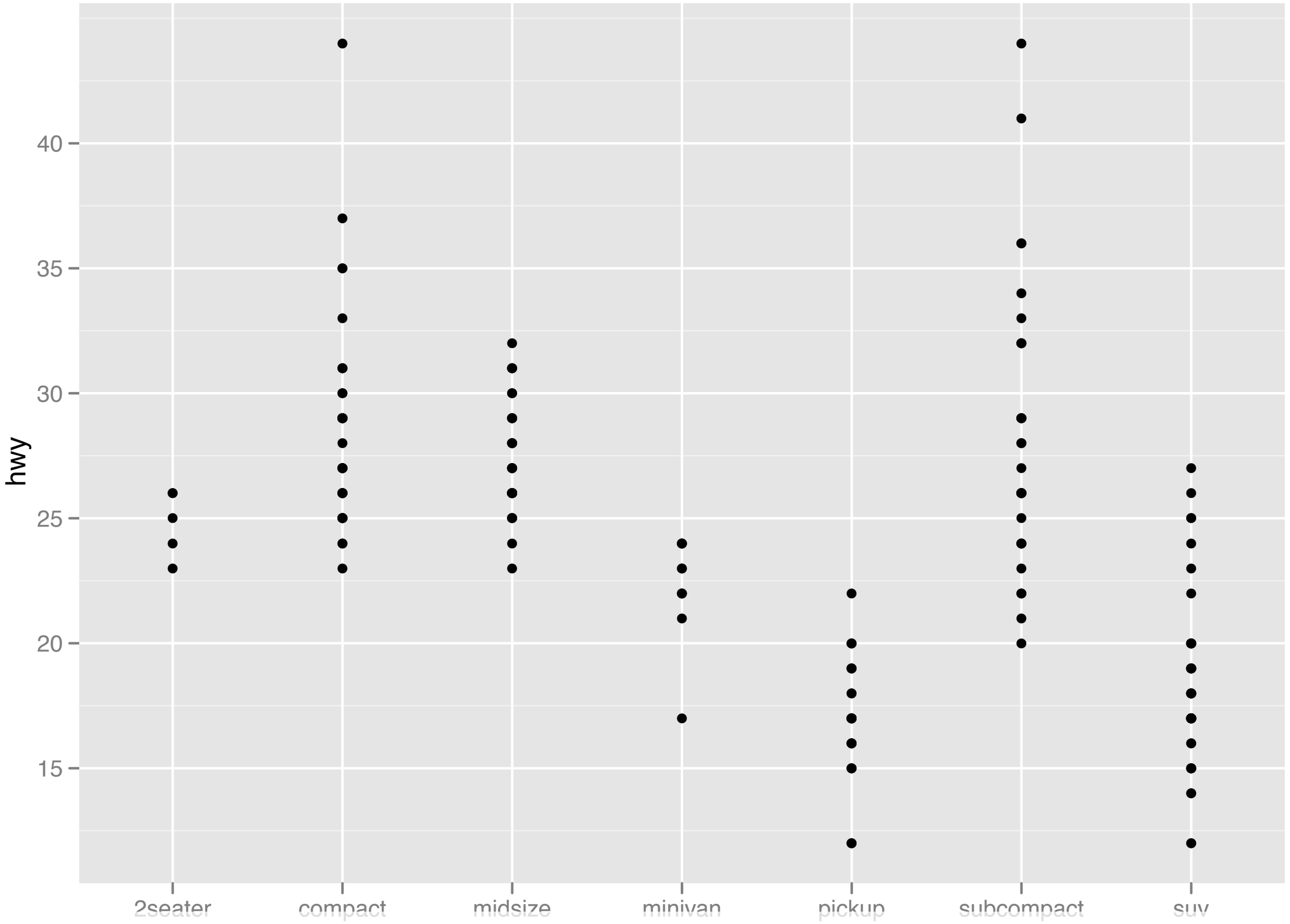

```
qplot(cty, hwy, data = mpg)
```




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



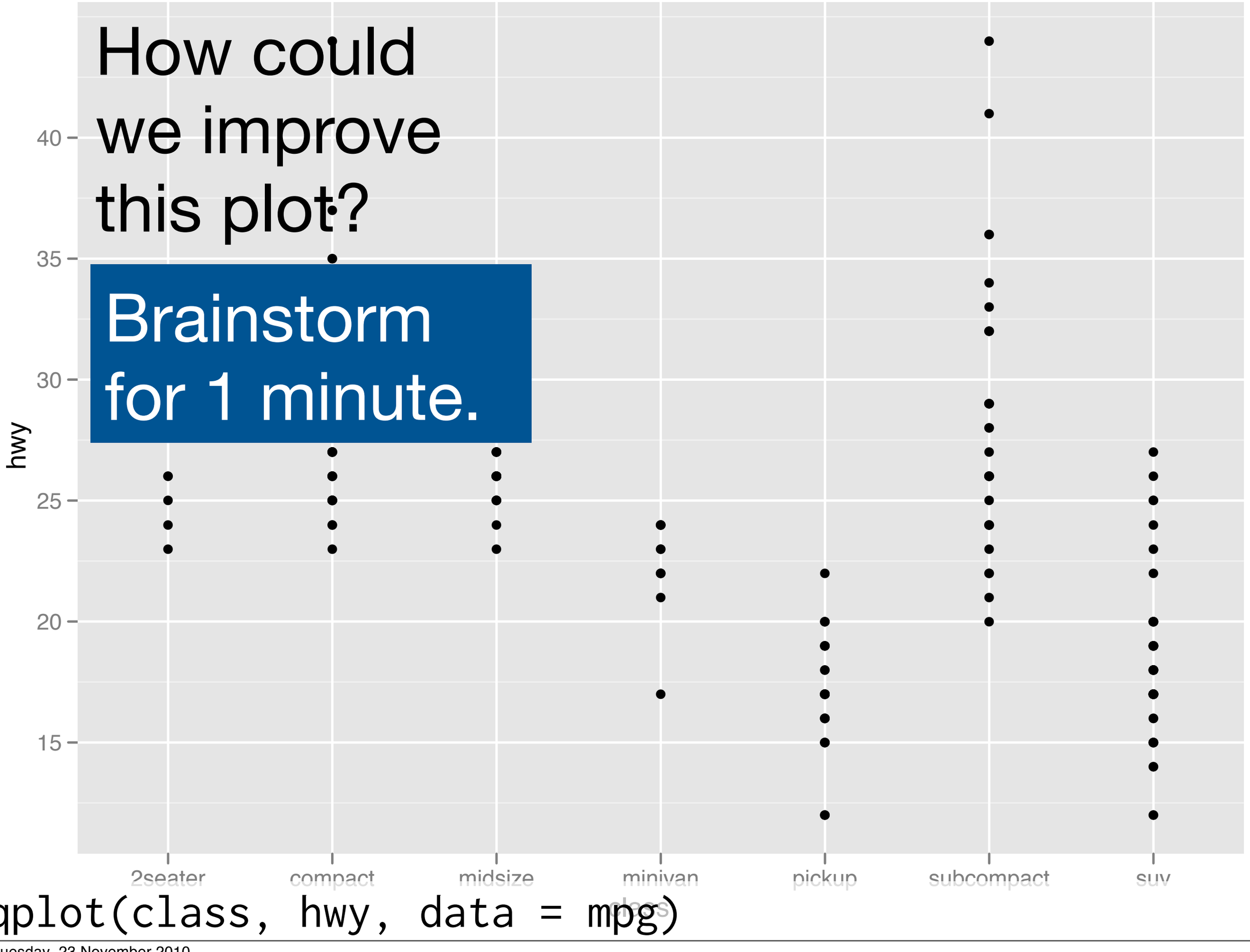

```
qplot(cty, hwy, data = mpg, geom = "jitter")
```



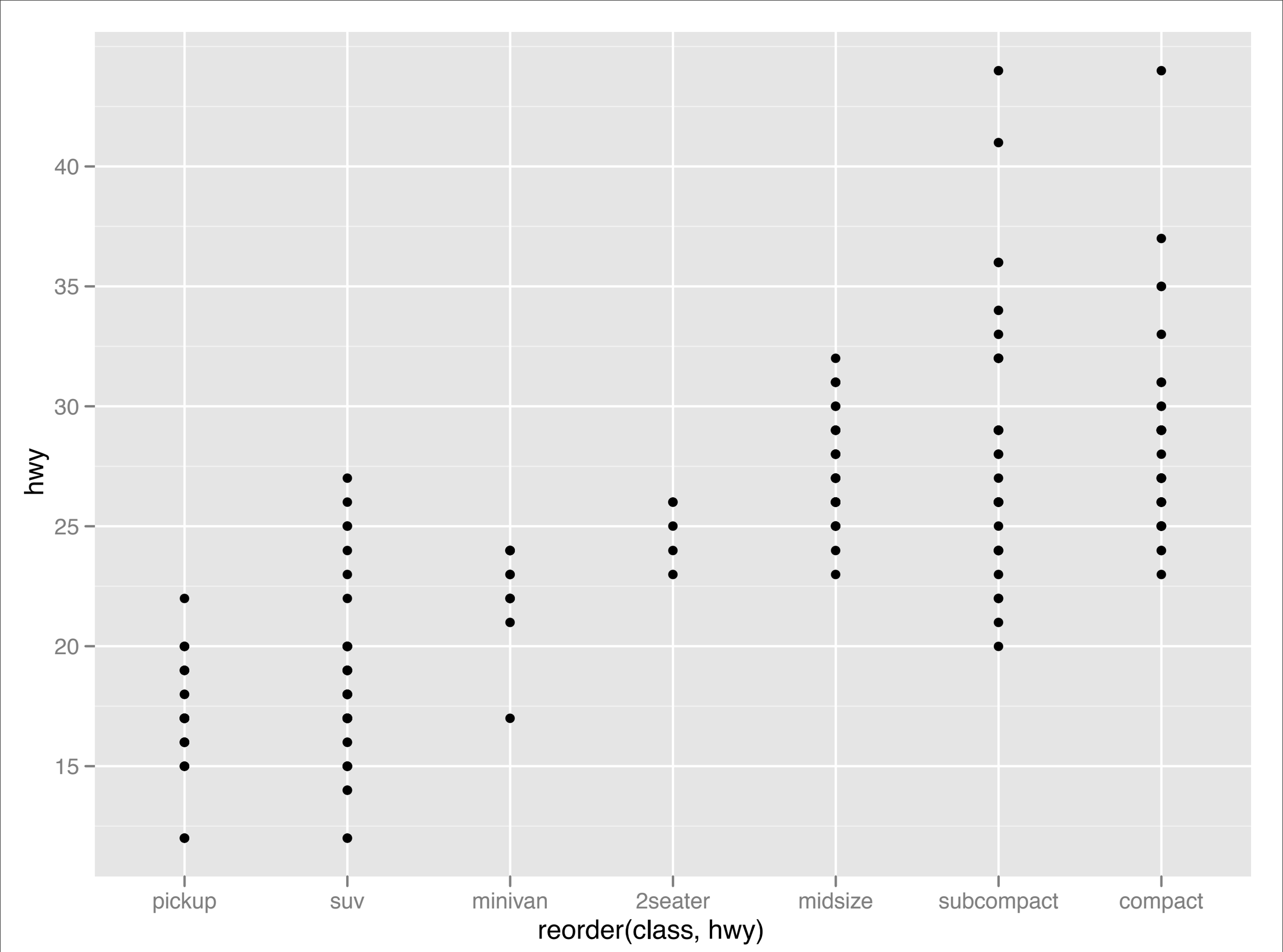

```
qplot(class, hwy, data = mpg)
```

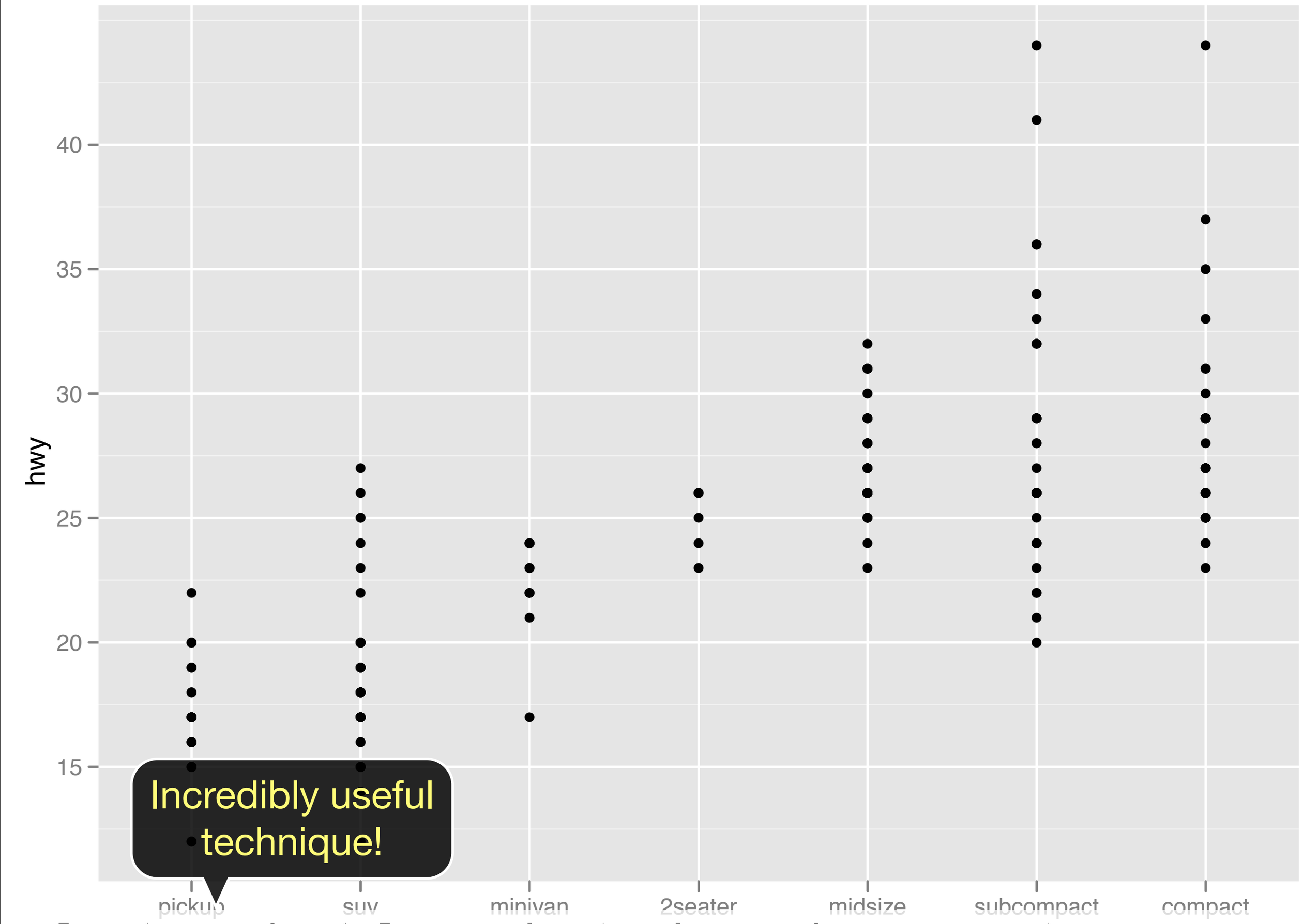
How could we improve this plot?

Brainstorm for 1 minute.

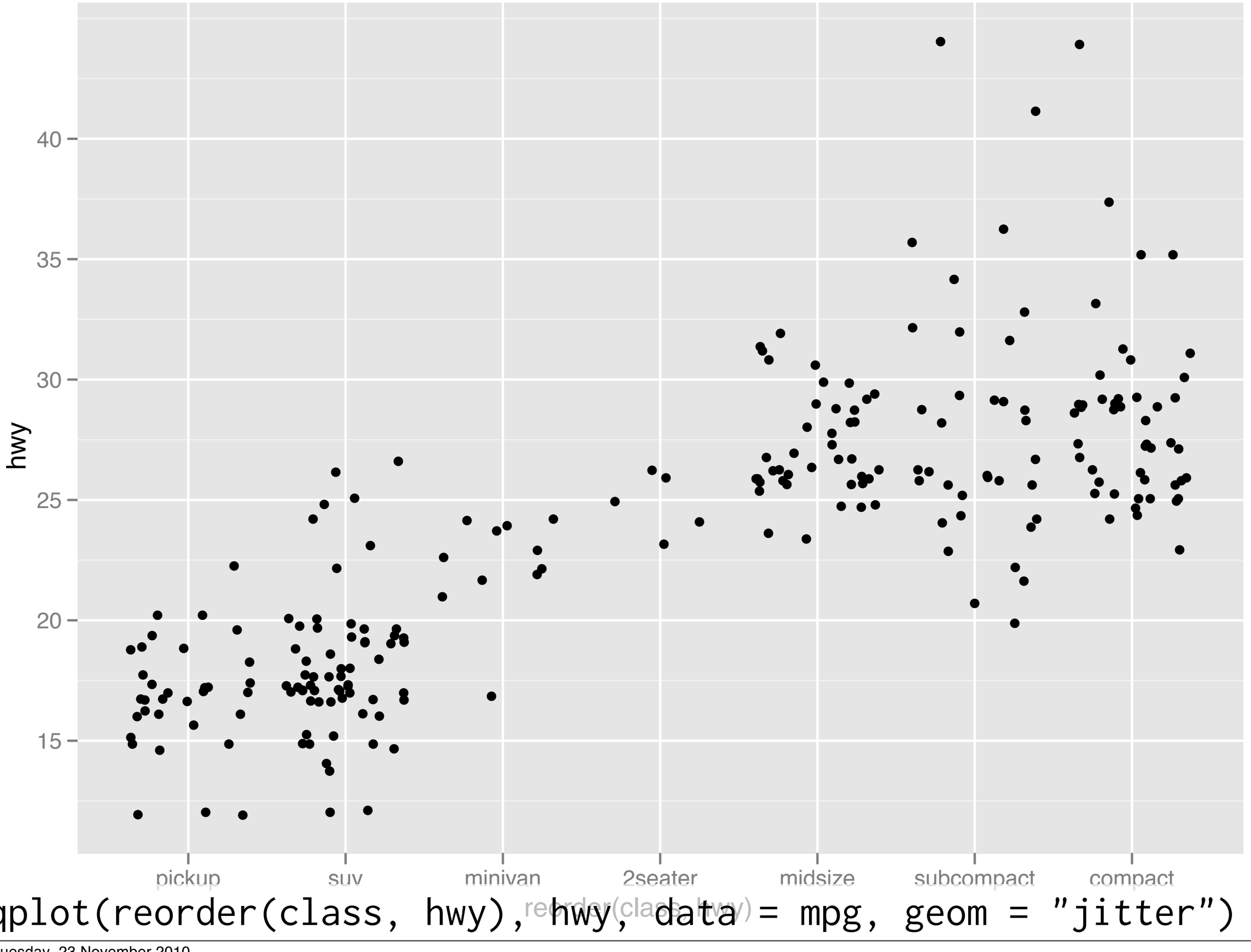



```
qplot(class, hwy, data = mpg)
```

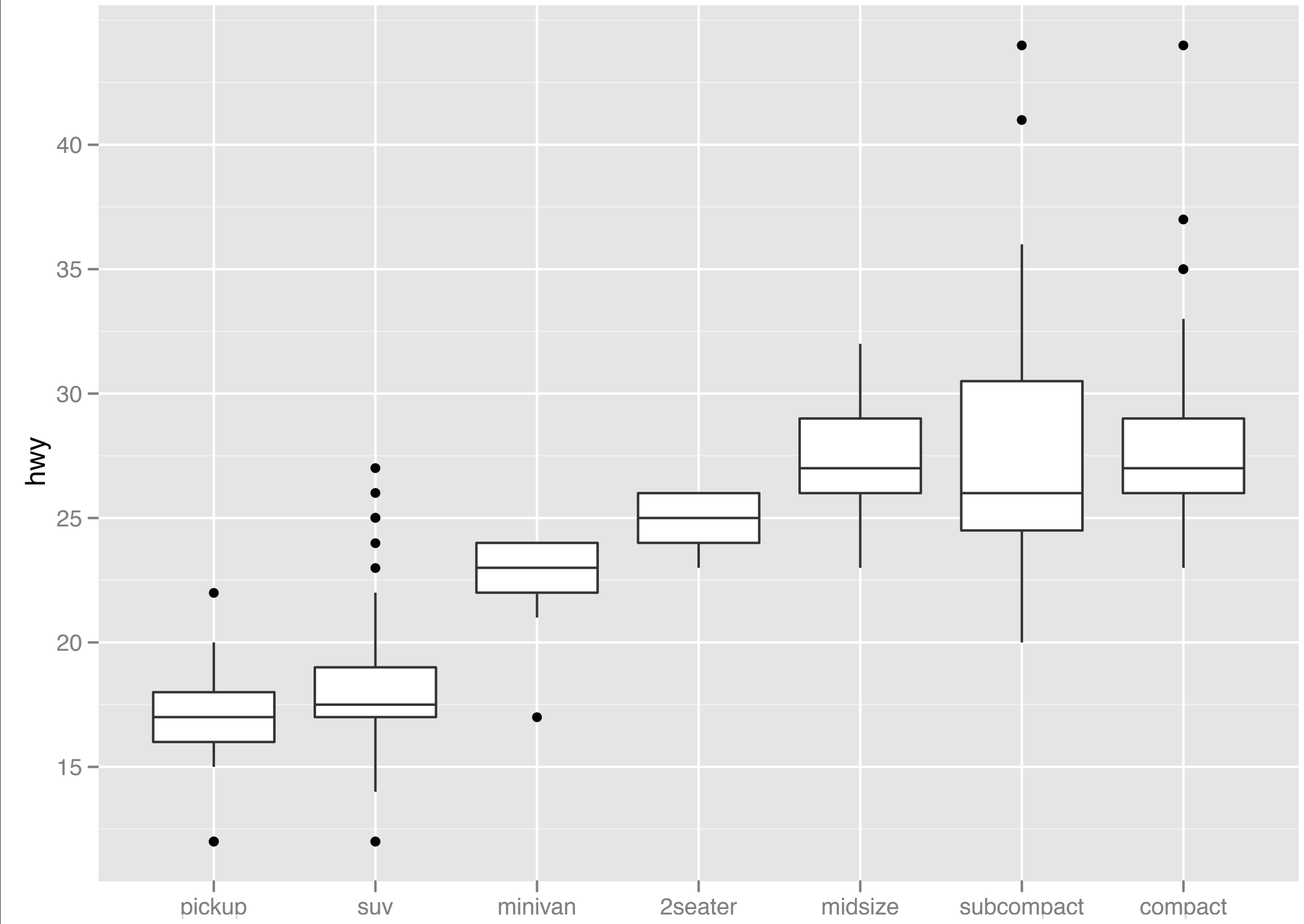

hwy
40
35
30
25
20
15
pickup
suv
minivan
2seater
midsize
subcompact
compact
reorder(class, hwy)



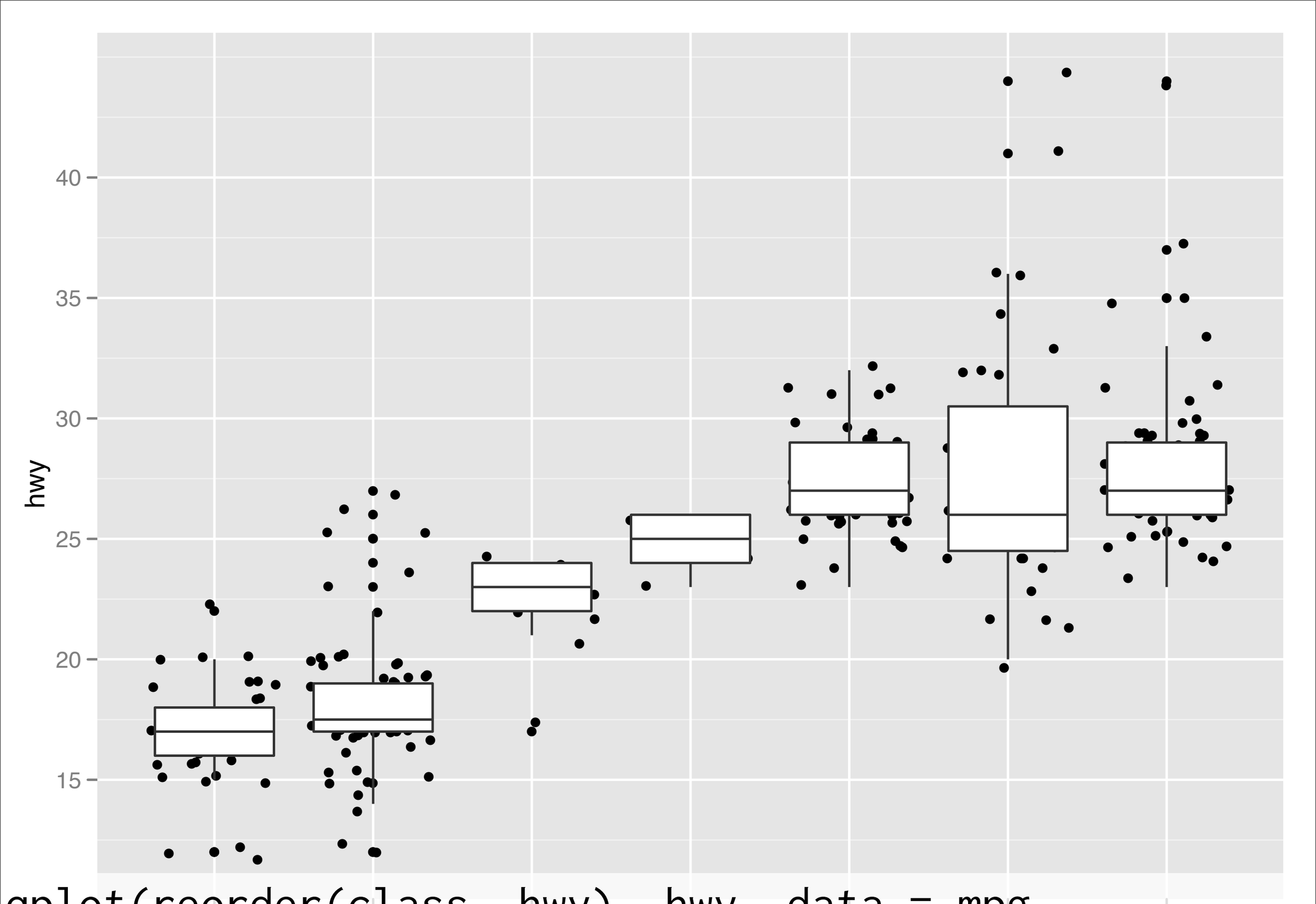

```
qplot(reorder(class, hwy), hwy, data = mpg)
```




```
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?

Tuesday, 23 November 2010

This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.