Introduction

Hadley Wickham

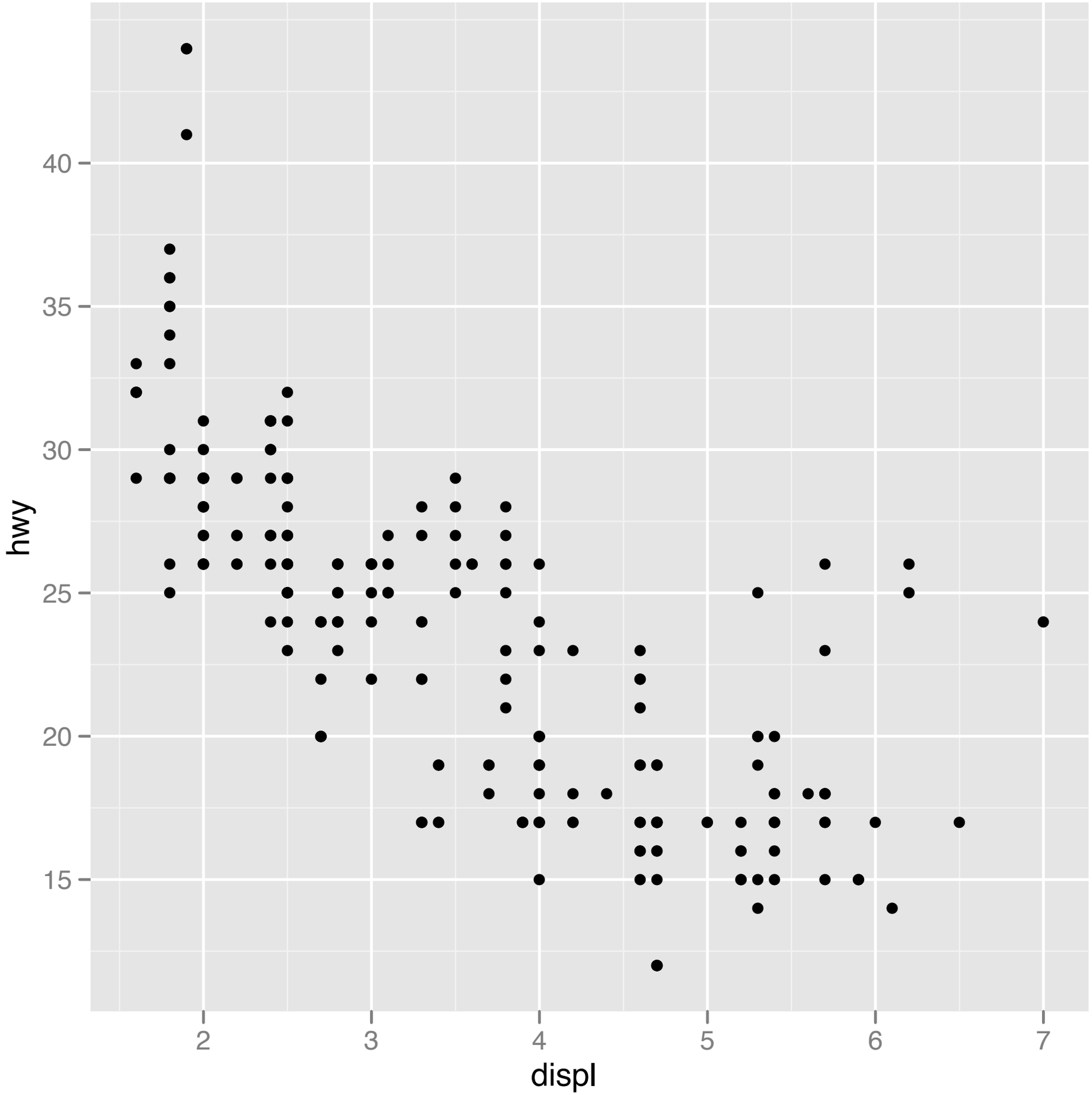
October 2009

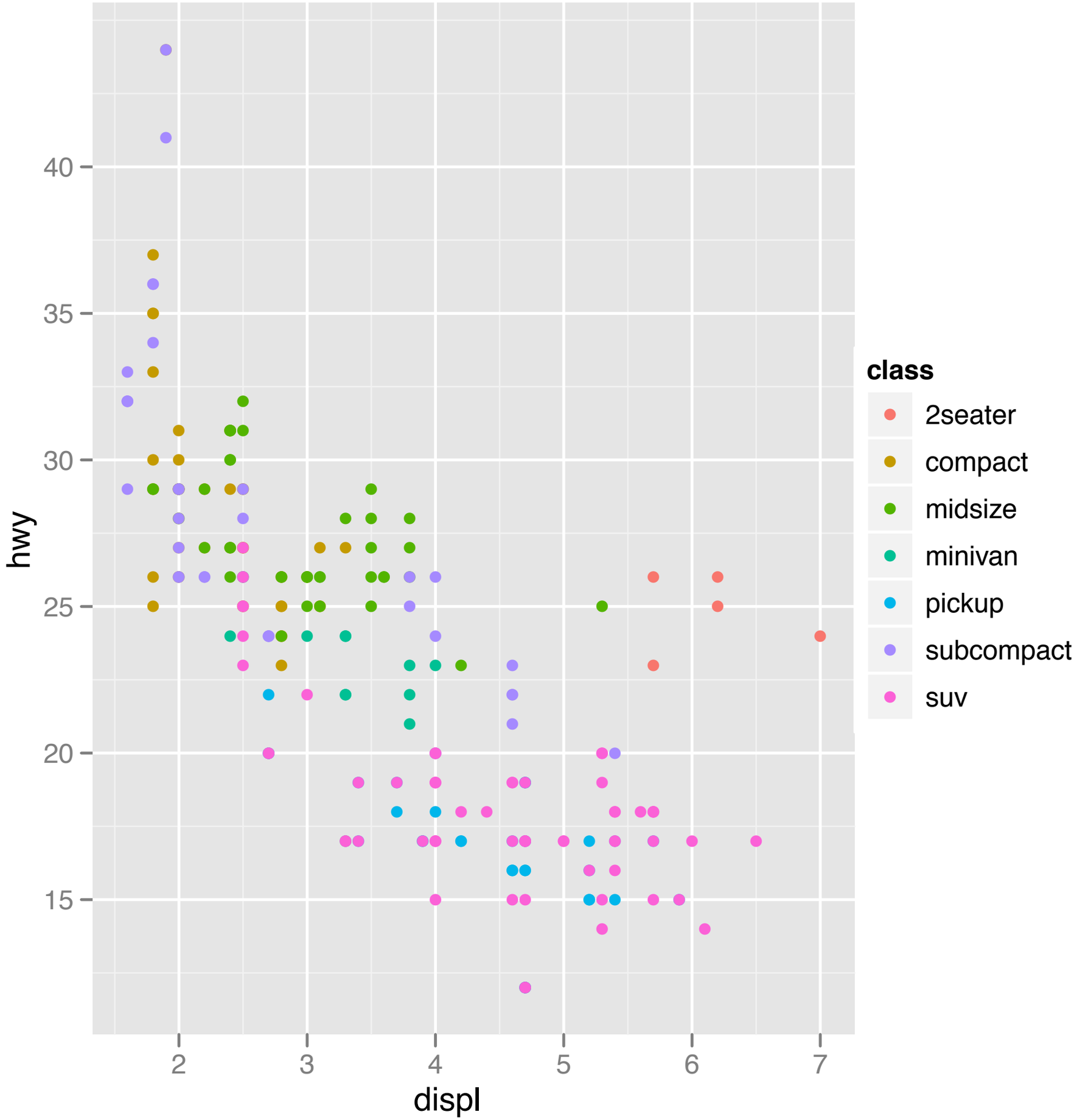
http://had.co.nz/umsl

1. Preview of today & tomorrow
2. About ggplot2
3. More resources
4. Diving in

Fuel economy

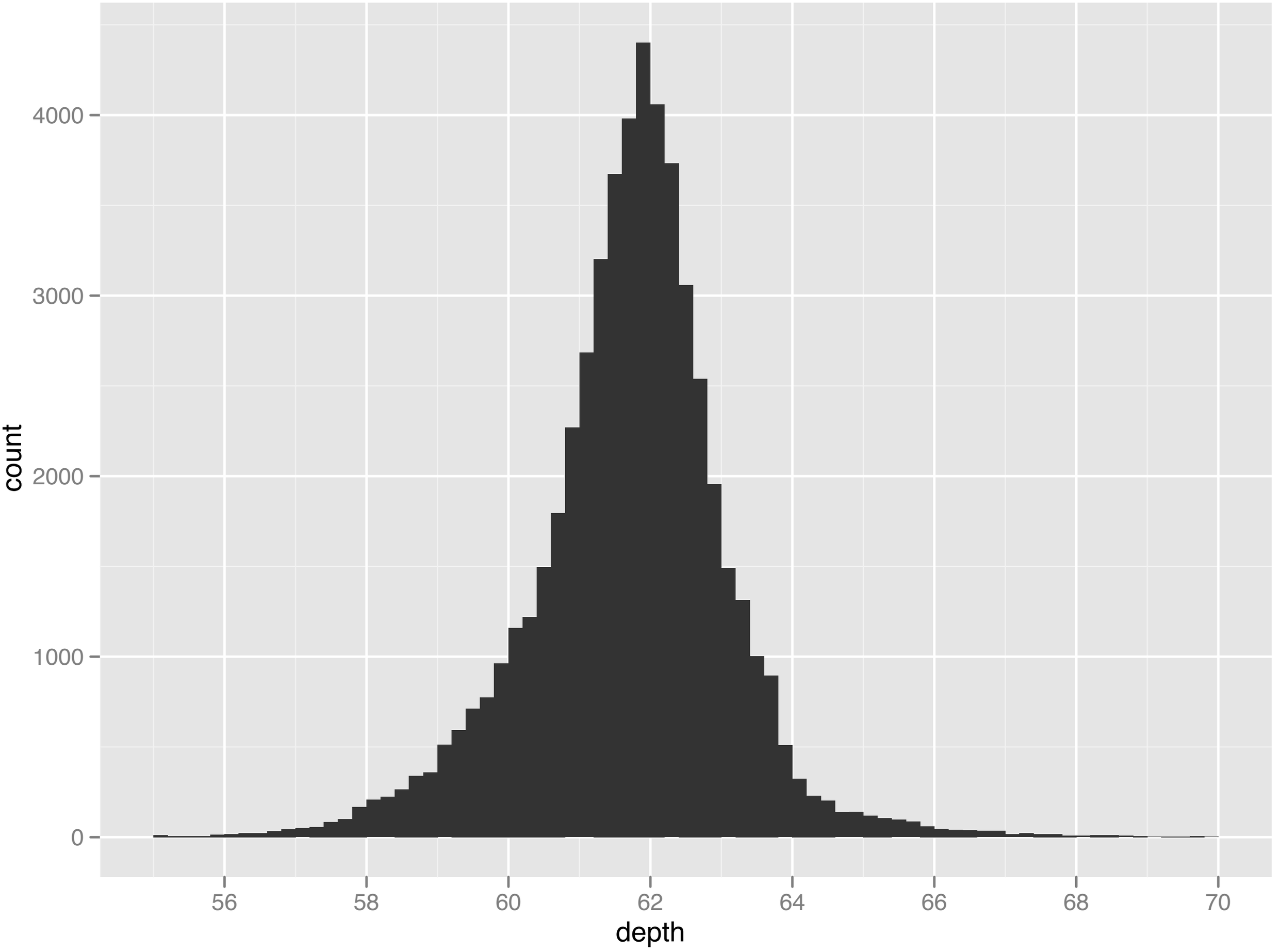
Basic graphics

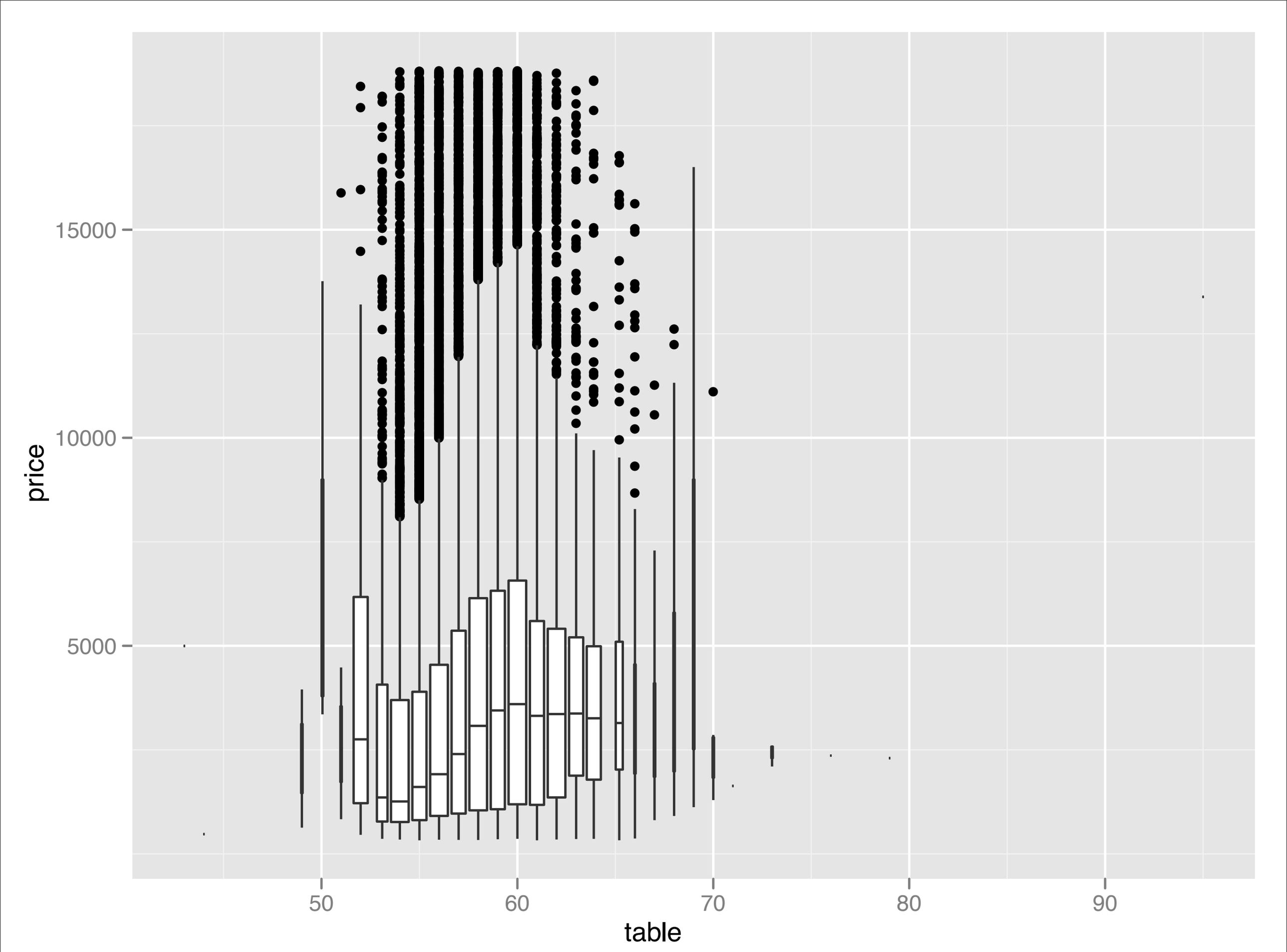

hwy
40
35
30
25
20
15
2
3
4
5
6
7
displ

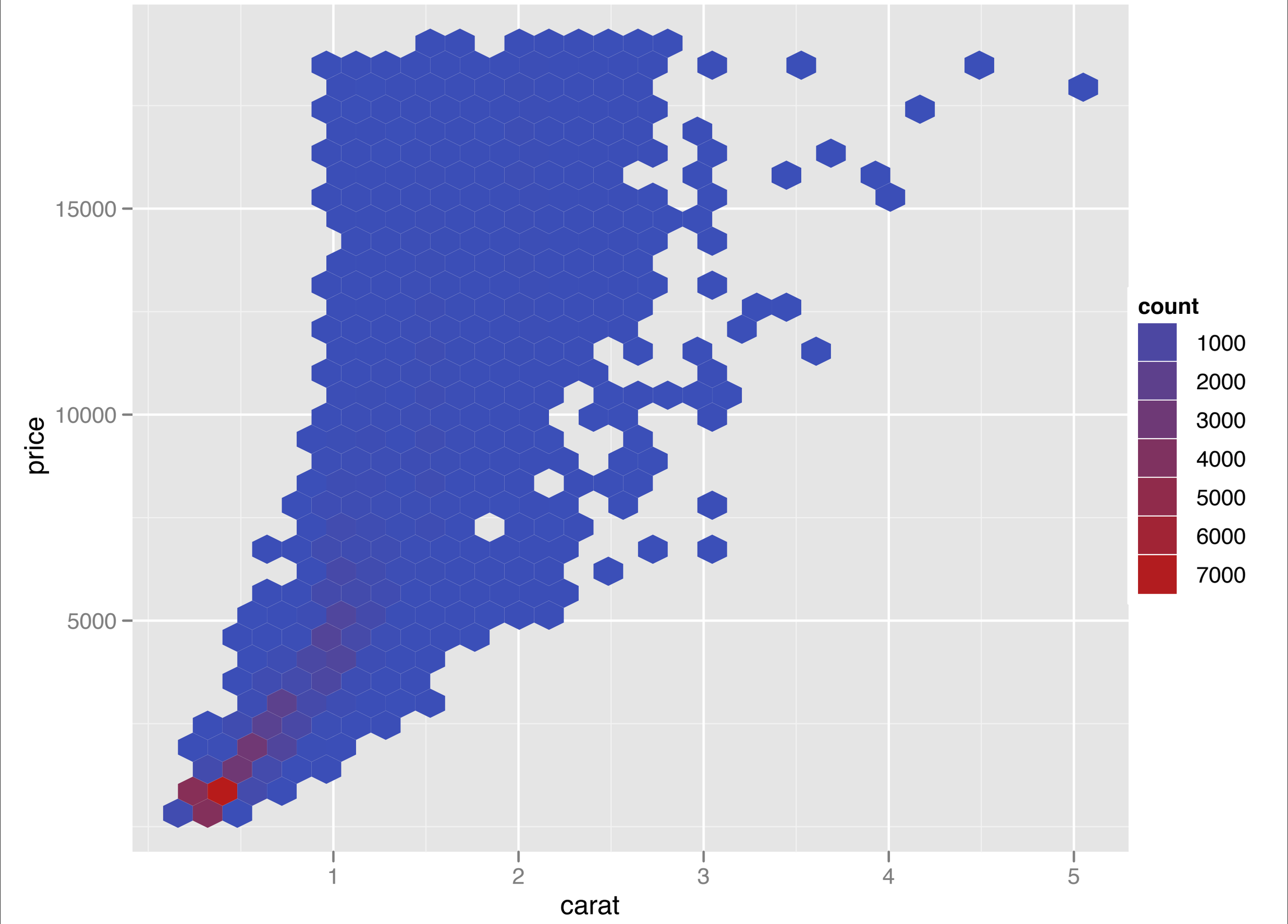

hwy
40
35
30
25
20
15
2
3
4
5
6
7
displ
class
2seater
compact
midsize
minivan
pickup
subcompact
suv

Diamond prices

Displaying large data

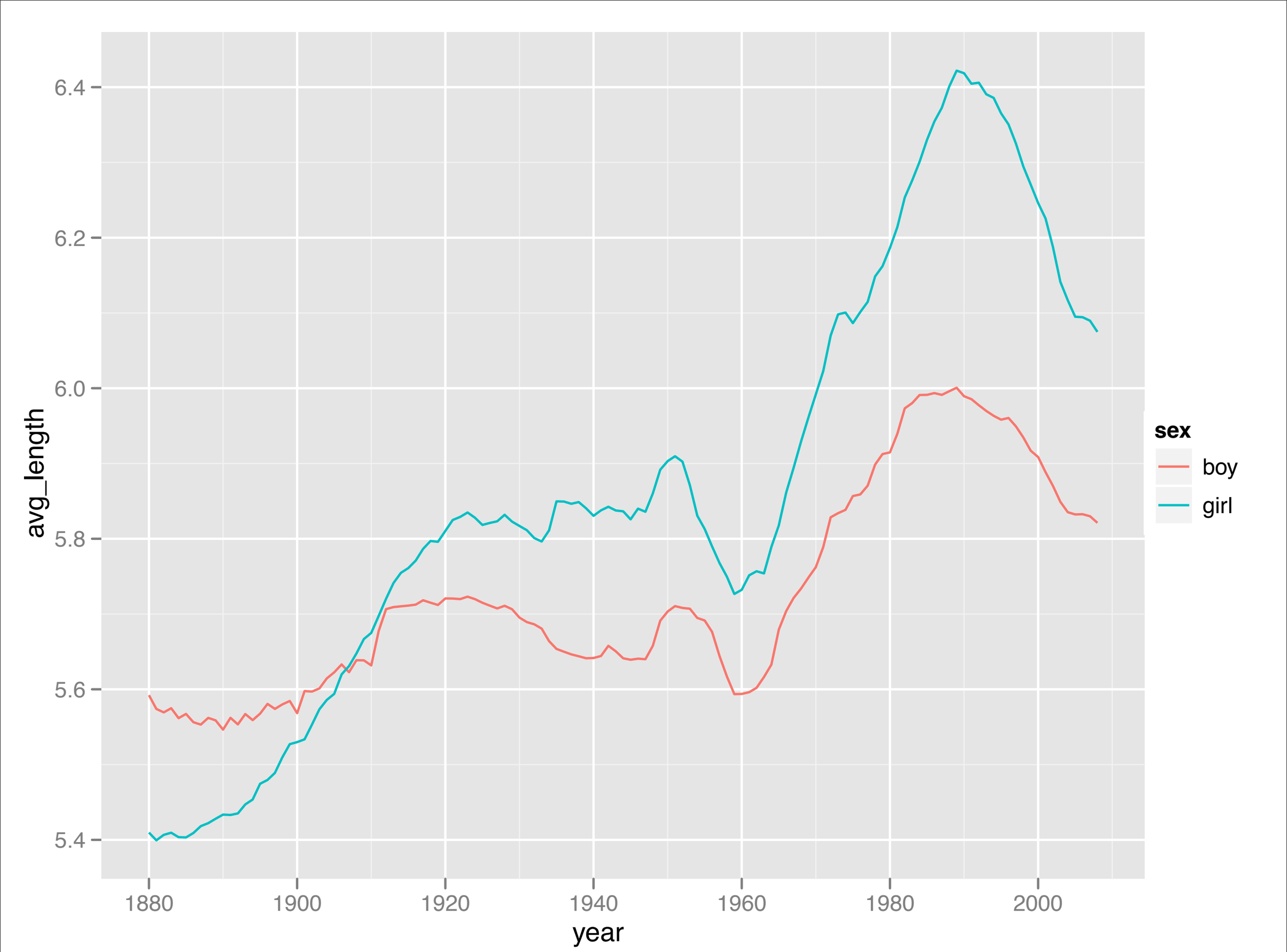

count
4000
3000
2000
1000
0
56
58
60
62
64
66
68
70
depth

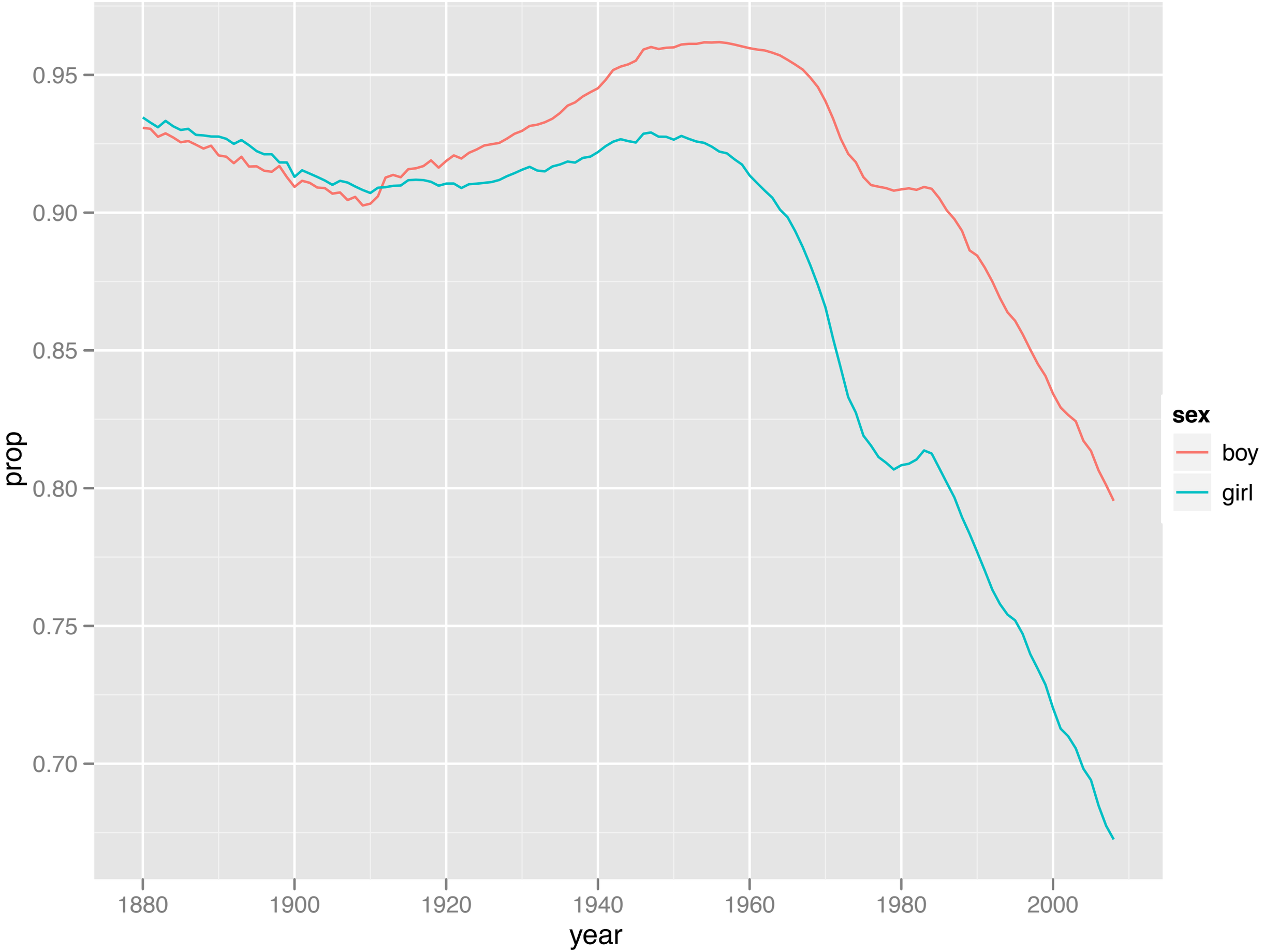

15000
10000
5000
price
50
60
70
80
90
table


15000
10000
5000
price
1
2
3
4
5
carat
count
1000
2000
3000
4000
5000
6000
7000

US baby names

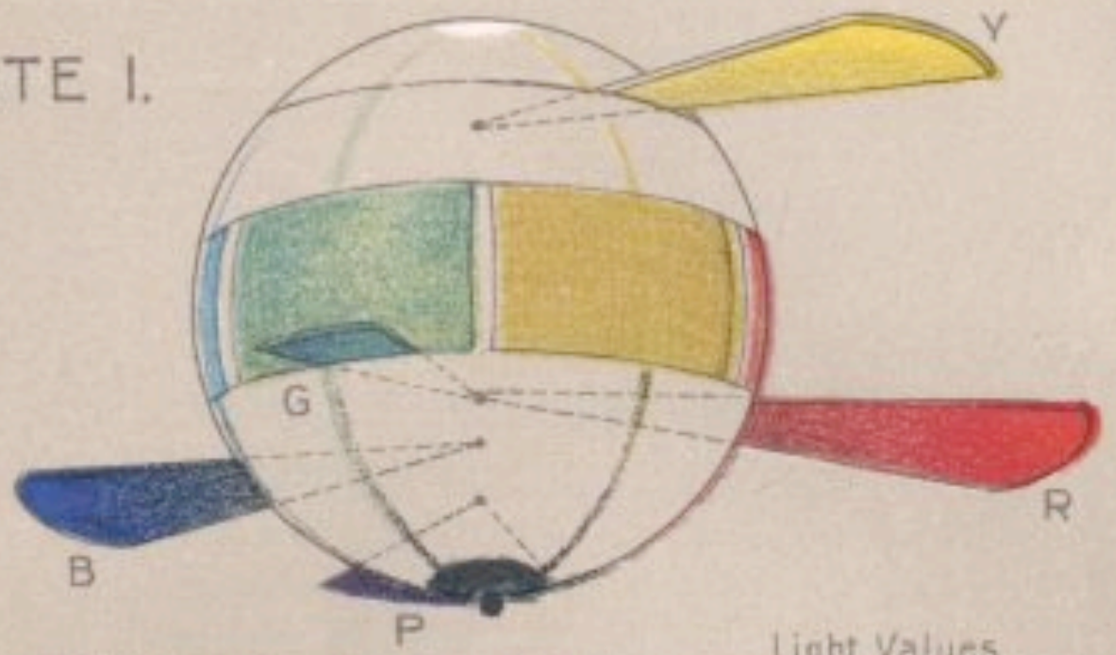
Transforming data

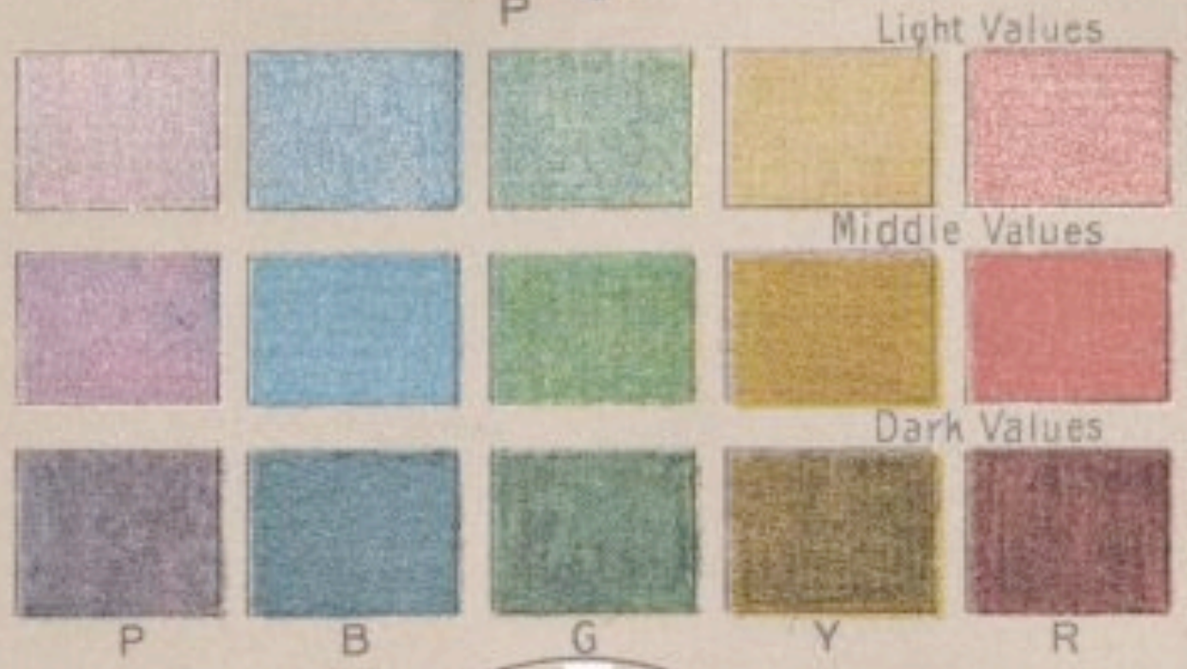

avg_length
6.4
6.2
6.0
5.8
5.6
5.4
1880
1900
1920
1940
1960
1980
2000
year
sex
boy
girl


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

Polishing your plots

PLATE I.



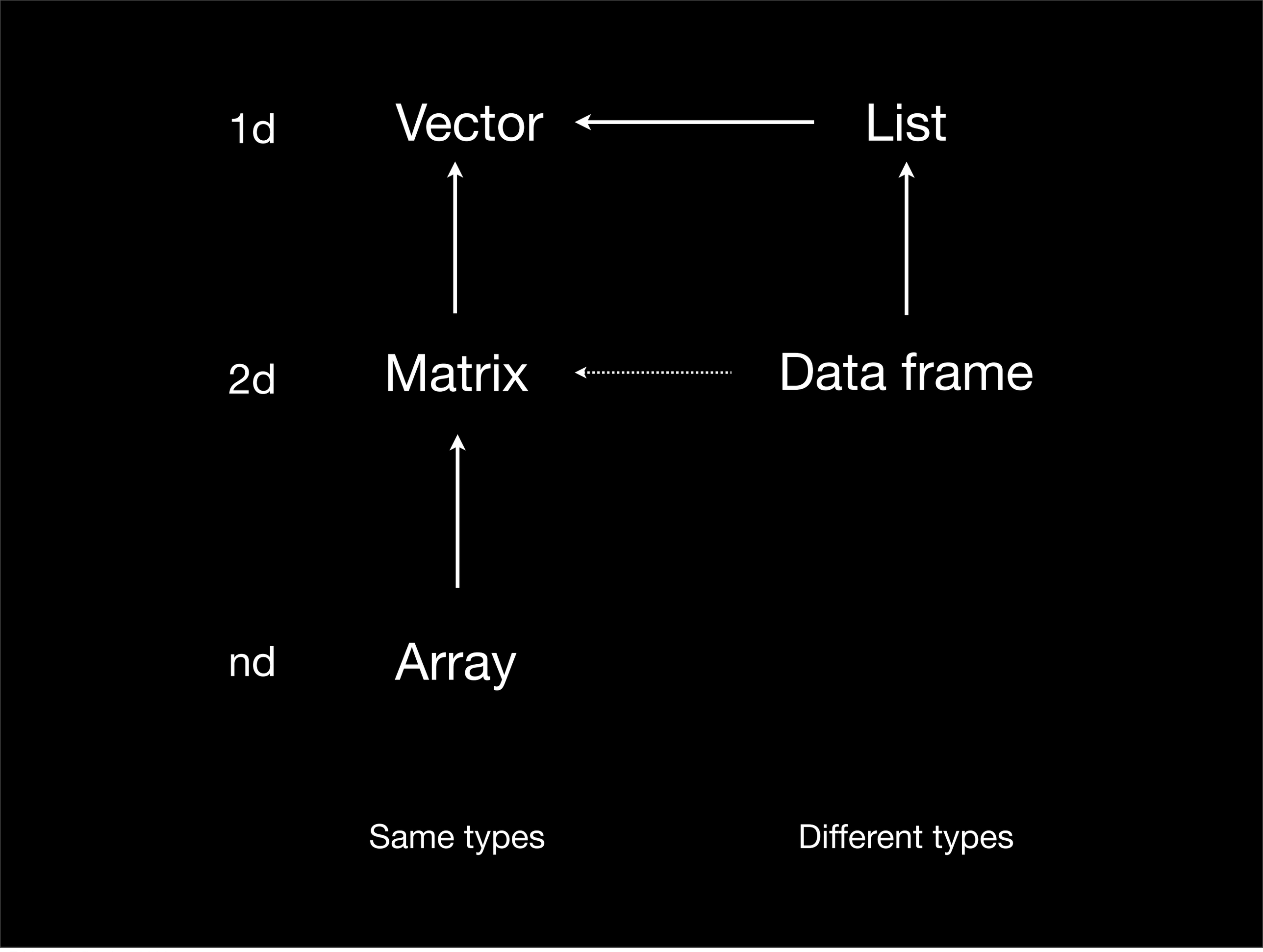


1. Saving your work
2. Labels & ticks
3. Themes

Tomorrow

Subsetting & data structures

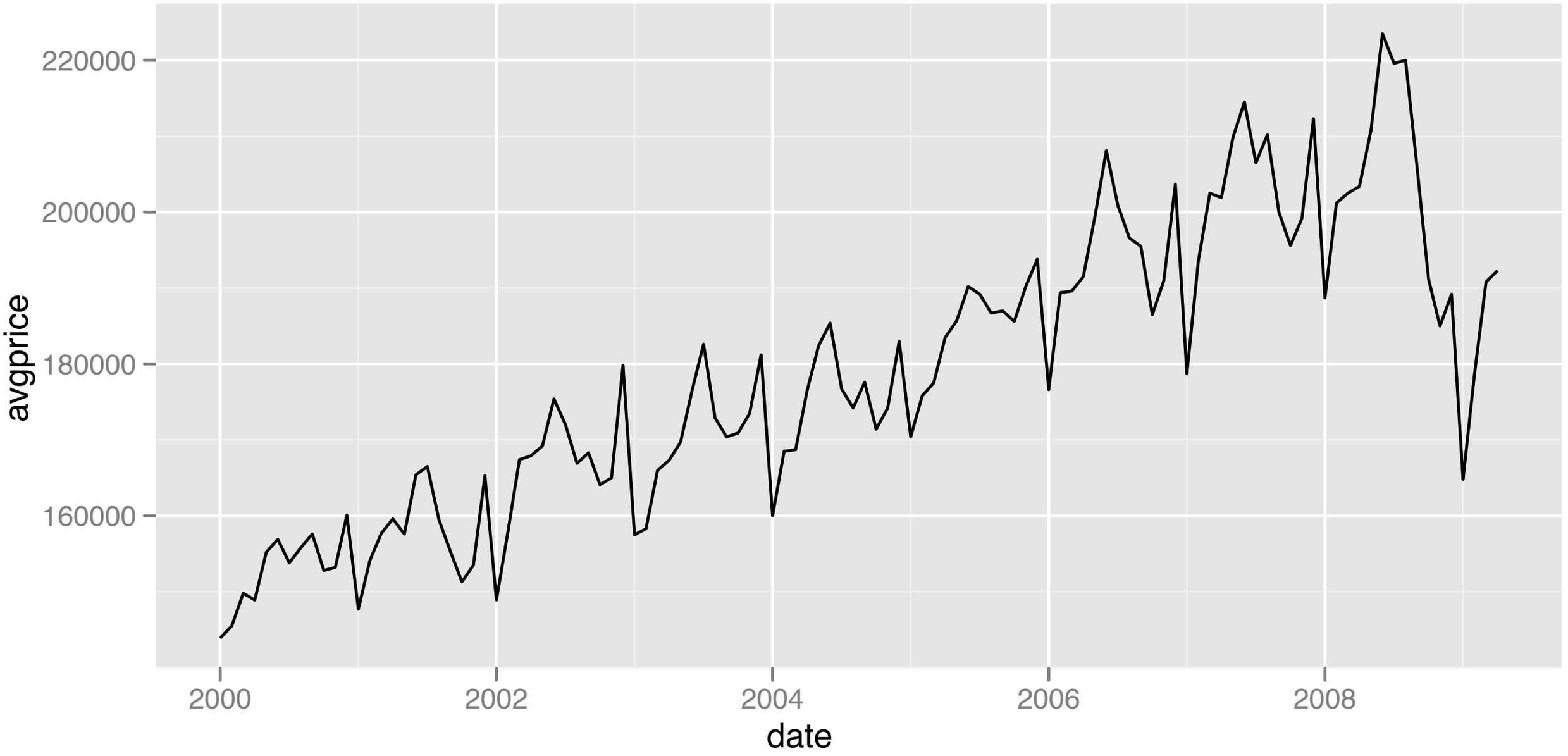
blank	include all
integer	**+ve**: include **-ve**: exclude
logical	include TRUEs
character	lookup by name


1d
Vector
List
2d
Matrix
Data frame
nd
Array
Same types
Different types

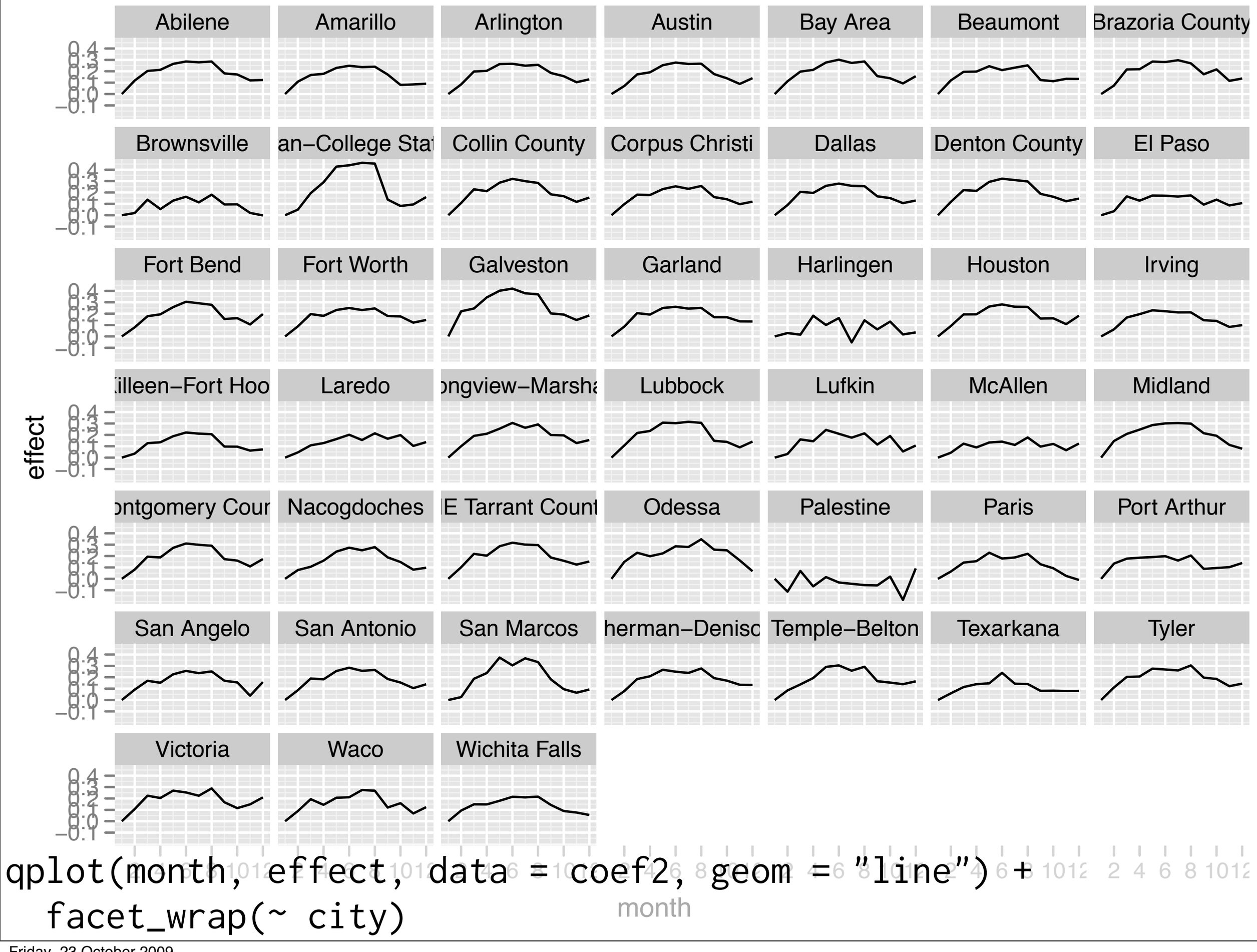
Butterfly communities & Texas house sales

Model visualisation

Friday, 23 October 2009



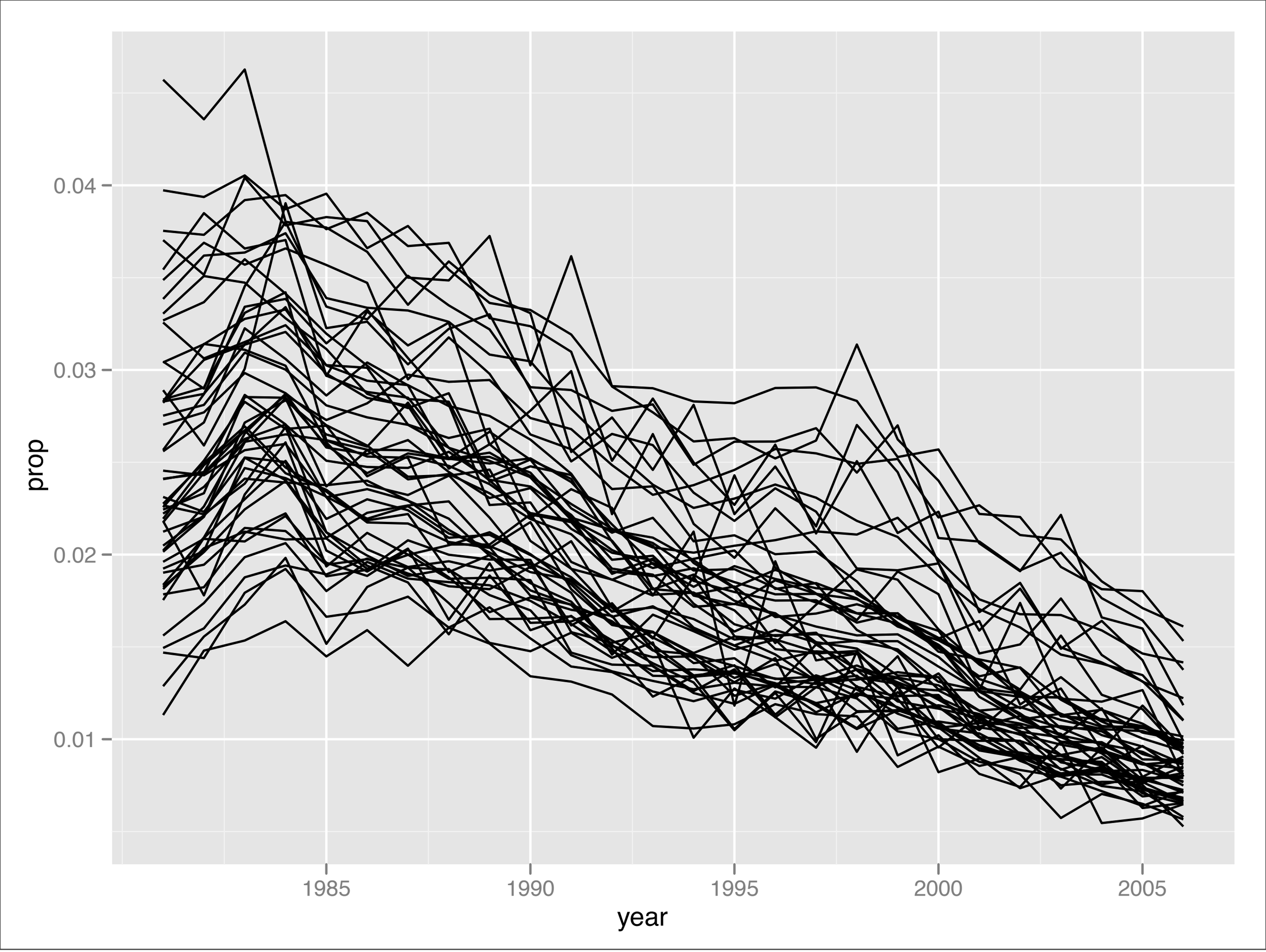

```
qplot(date, avgprice, data = houston, geom = "line")
```

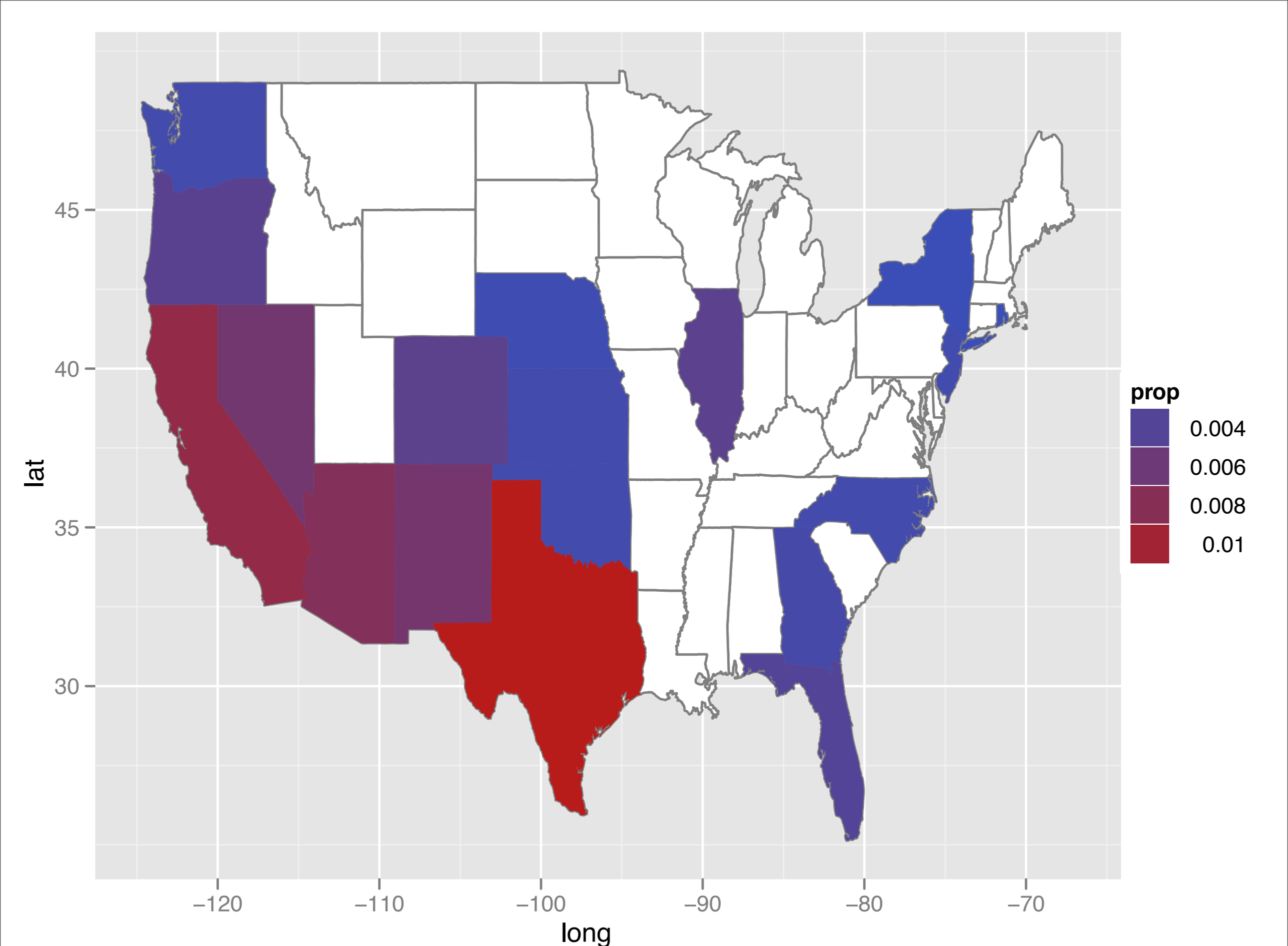



```
qplot(month, effect, data = coef2, geom = "line") +
  facet_wrap(~ city)
```

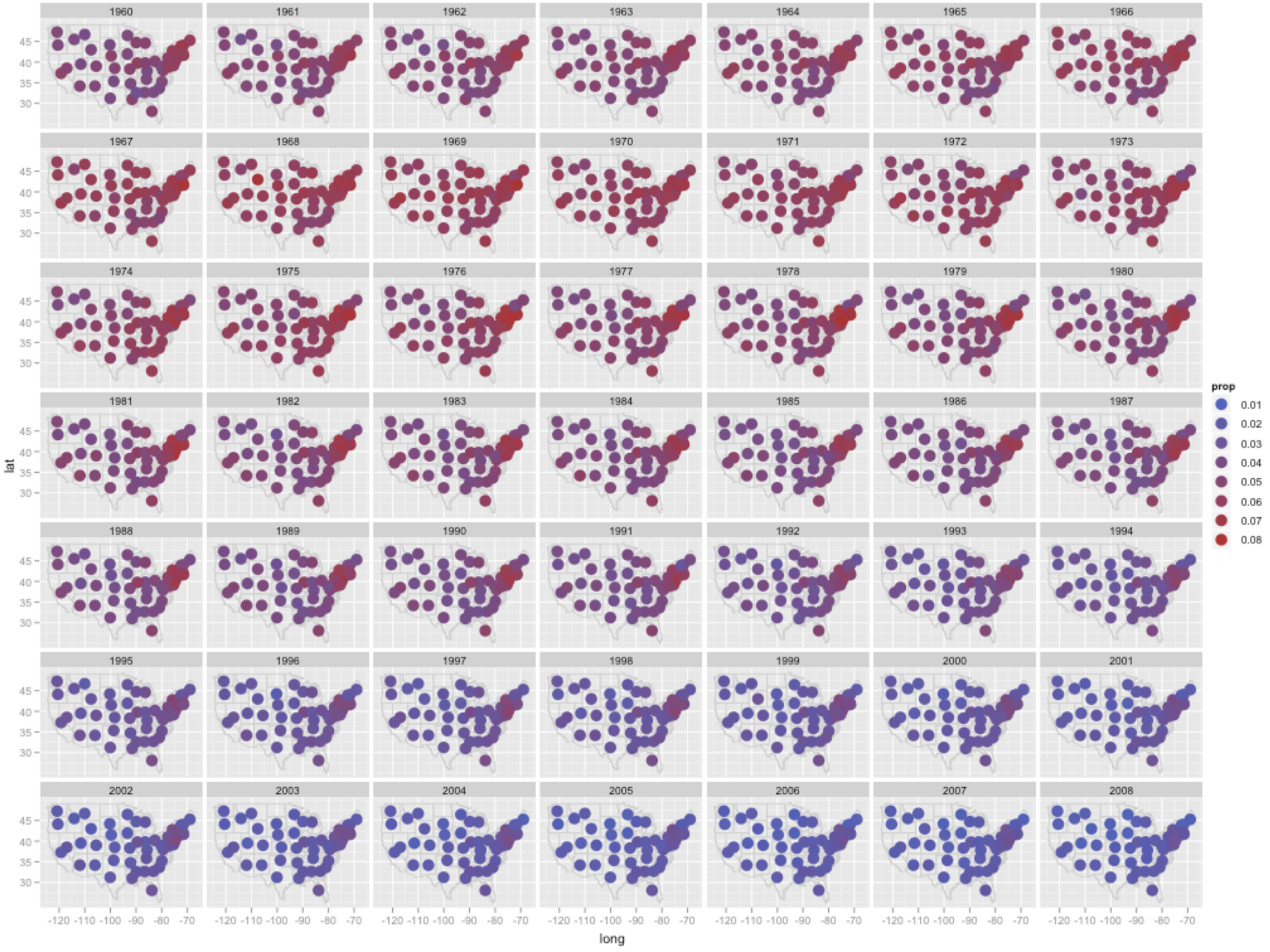
Baby names by state

Visualising time & space

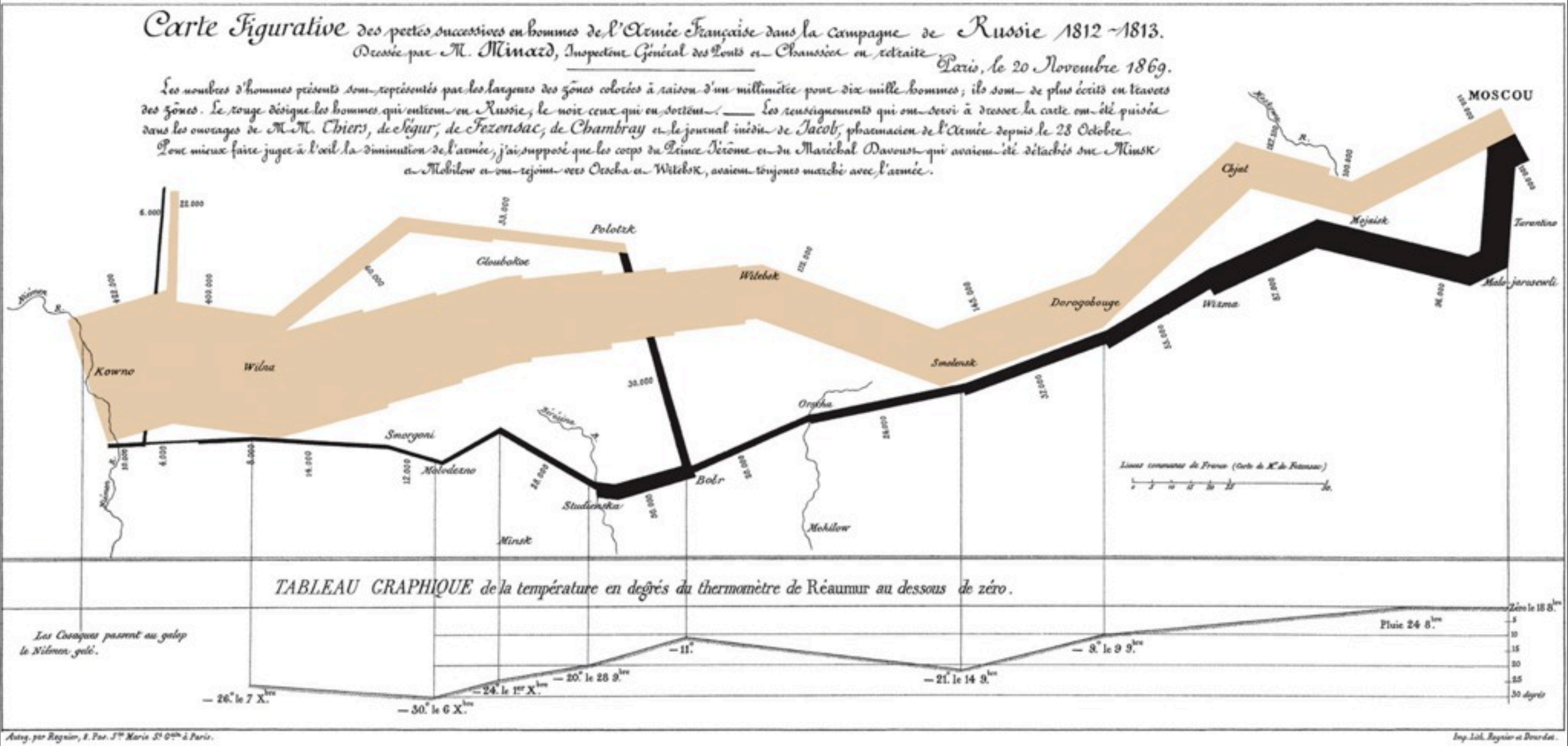

0.04
0.03
0.02
0.01
prop
1985
1990
1995
2000
2005
year



Friday, 23 October 2009


1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
lat
45
40
35
30
-120 -110 -100 -90 -80 -70
long
prop
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08

ggplot2 theory & graphical critique



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Polotzk
Gloubokoe
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Wixma
Tarentino
Malo-jarosewli
Kowno
Wilna
Smorgoni
Molodezno
Studienska
Bobr
Orscha
Minsk
Mohilow
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Zéro le 18 8.bre
Pluie 24 8.bre
– 9° le 9 9.bre
– 11°
– 21° le 14 9.bre
– 20° le 28 9.bre
– 24° le 1er X.bre
– 30° le 6 X.bre
– 26° le 7 X.bre
30 degrés

Content

Construction


Context


Consumption

ggplot2

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

Scatterplot basics

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

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

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

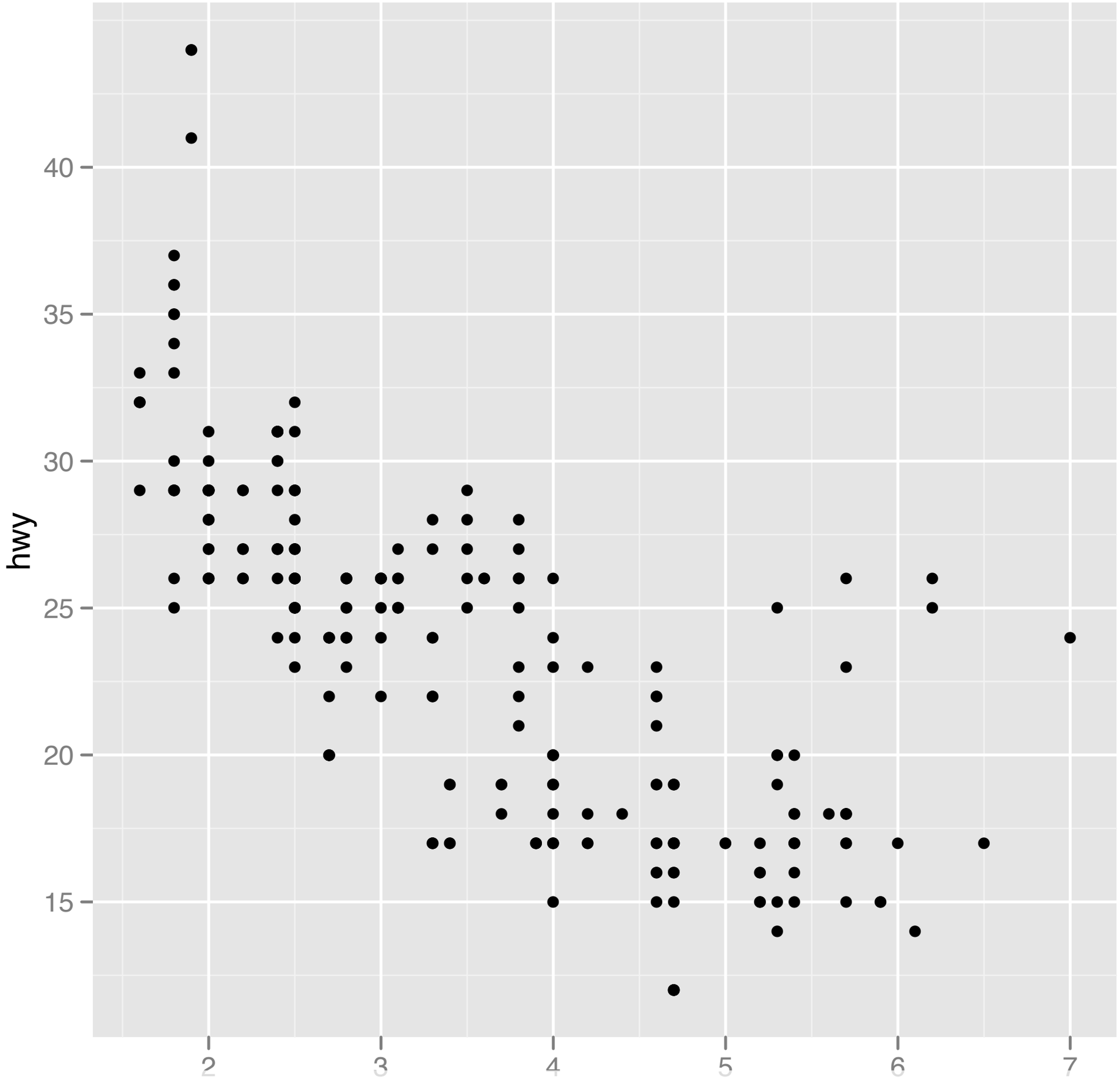
Scatterplot basics

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

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

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

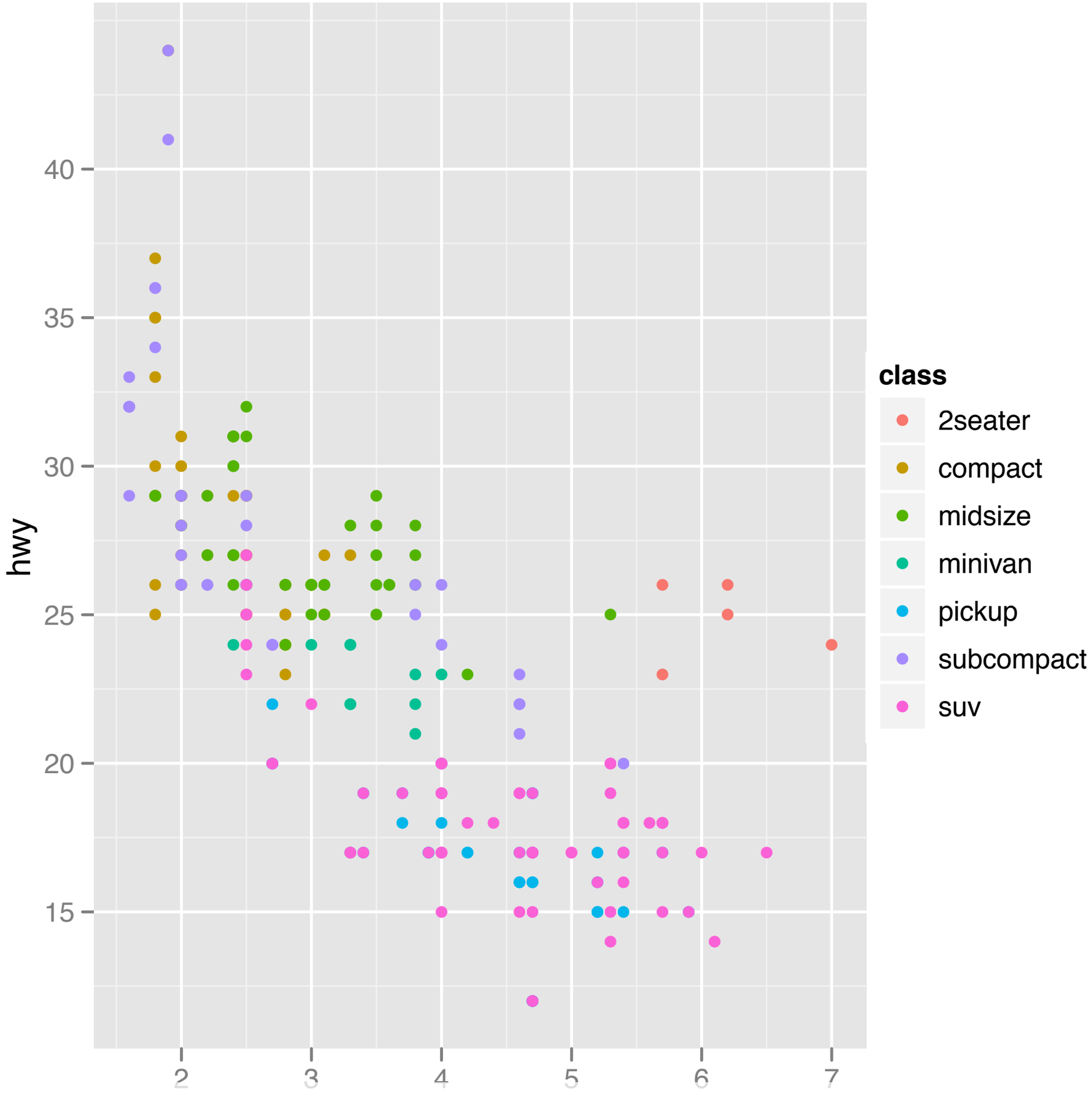
In ggplot2, we always explicitly specify the data



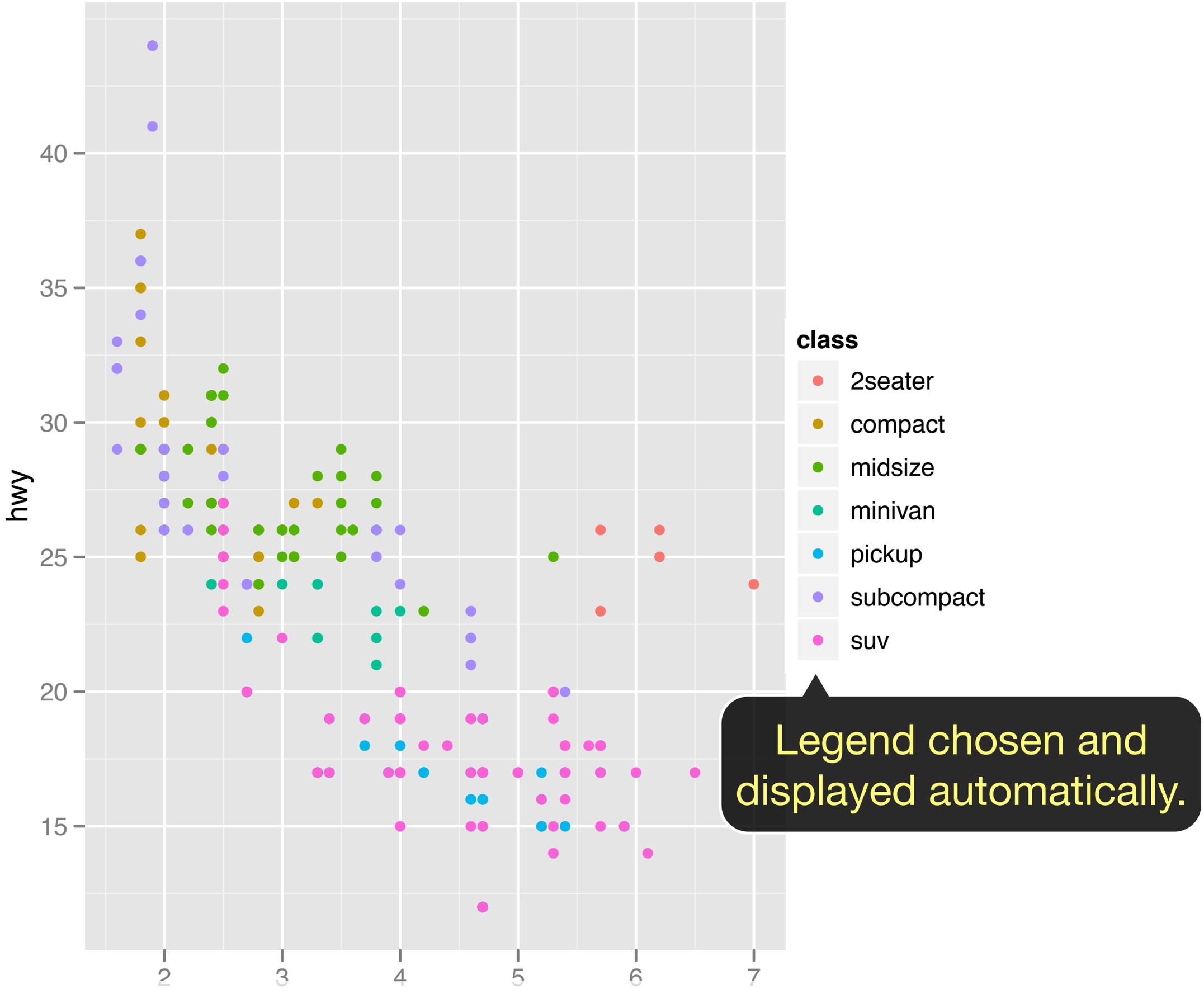

```
qplot(displ, hwy, data = mpg)
```

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

Experiment with colour, size, and shape aesthetics.

What's the difference between discrete or continuous variables?

What happens when you combine multiple aesthetics?

	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)
```

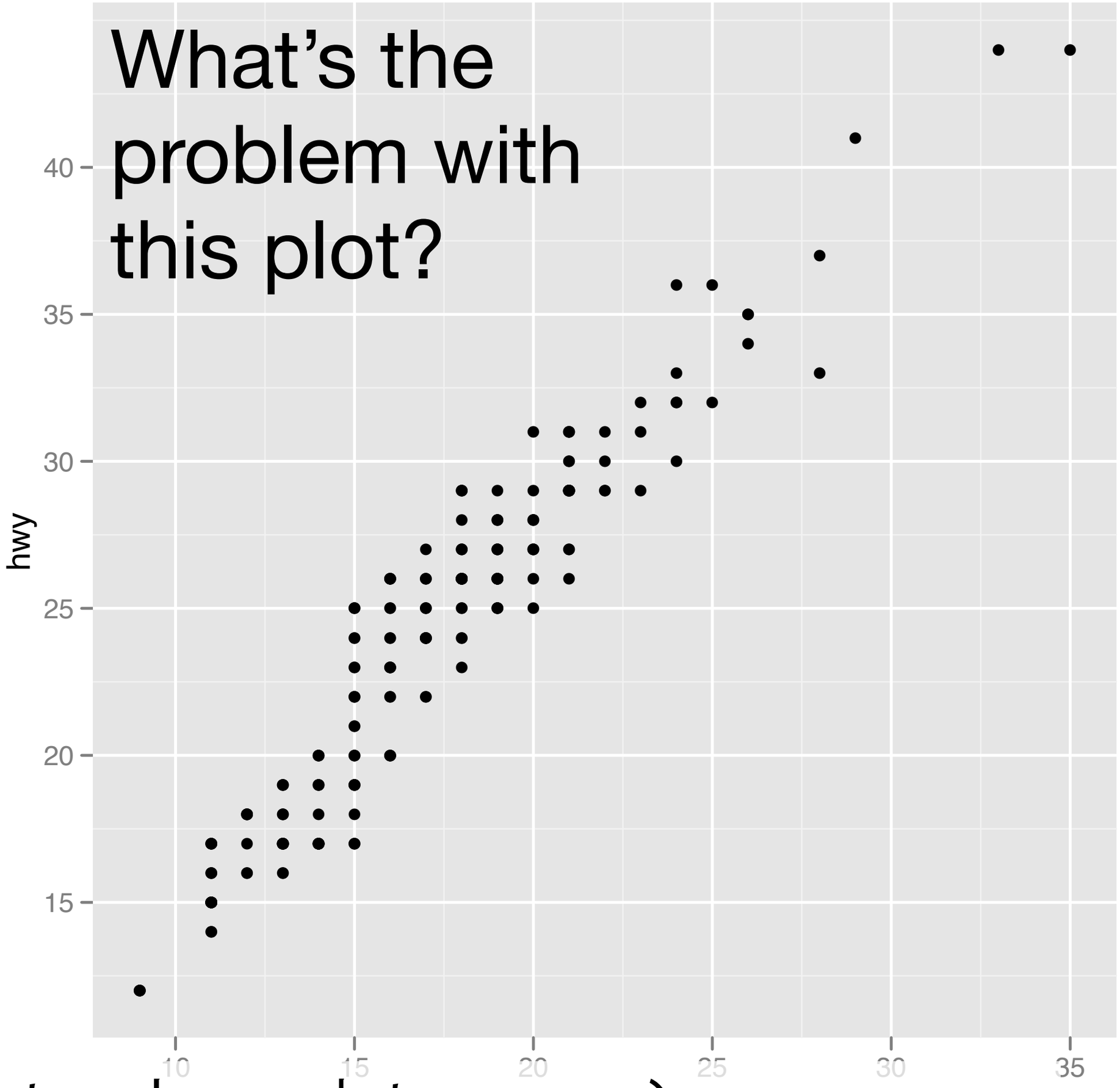
Summary

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

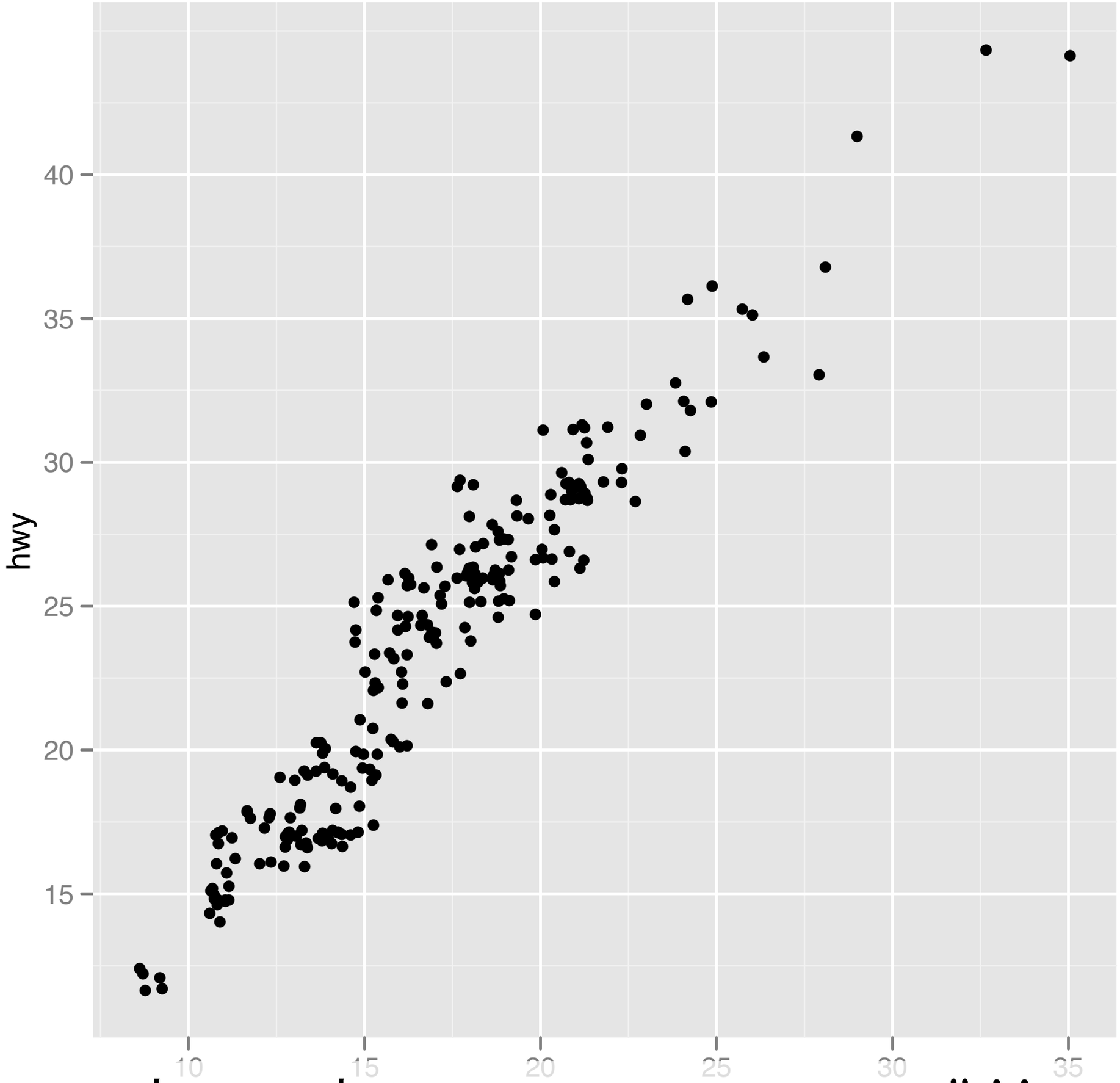
`facet_wrap()`: 1d ribbon wrapped into 2d

Scales argument controls whether position scales are fixed or free.

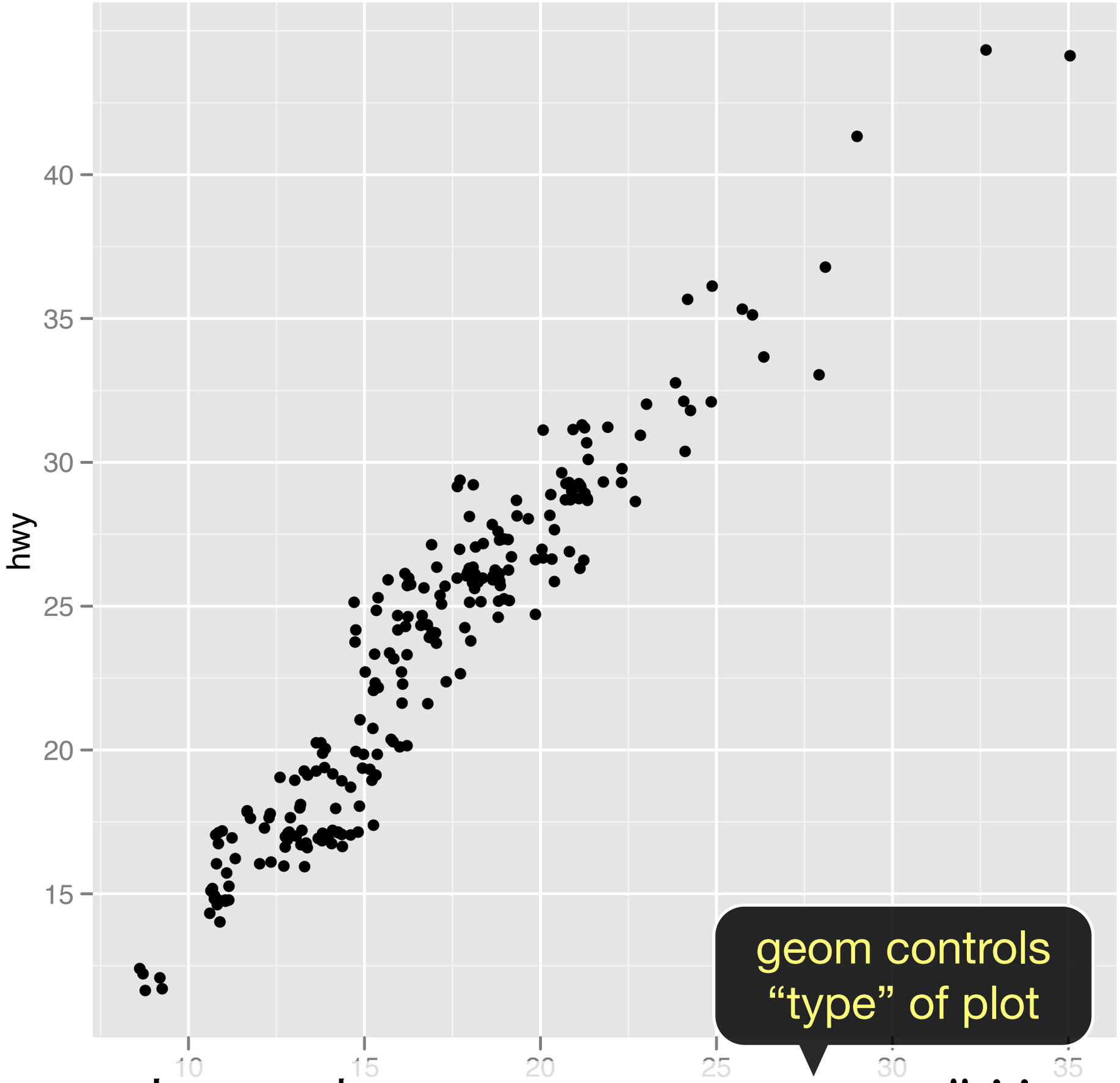
What's the problem with this plot?



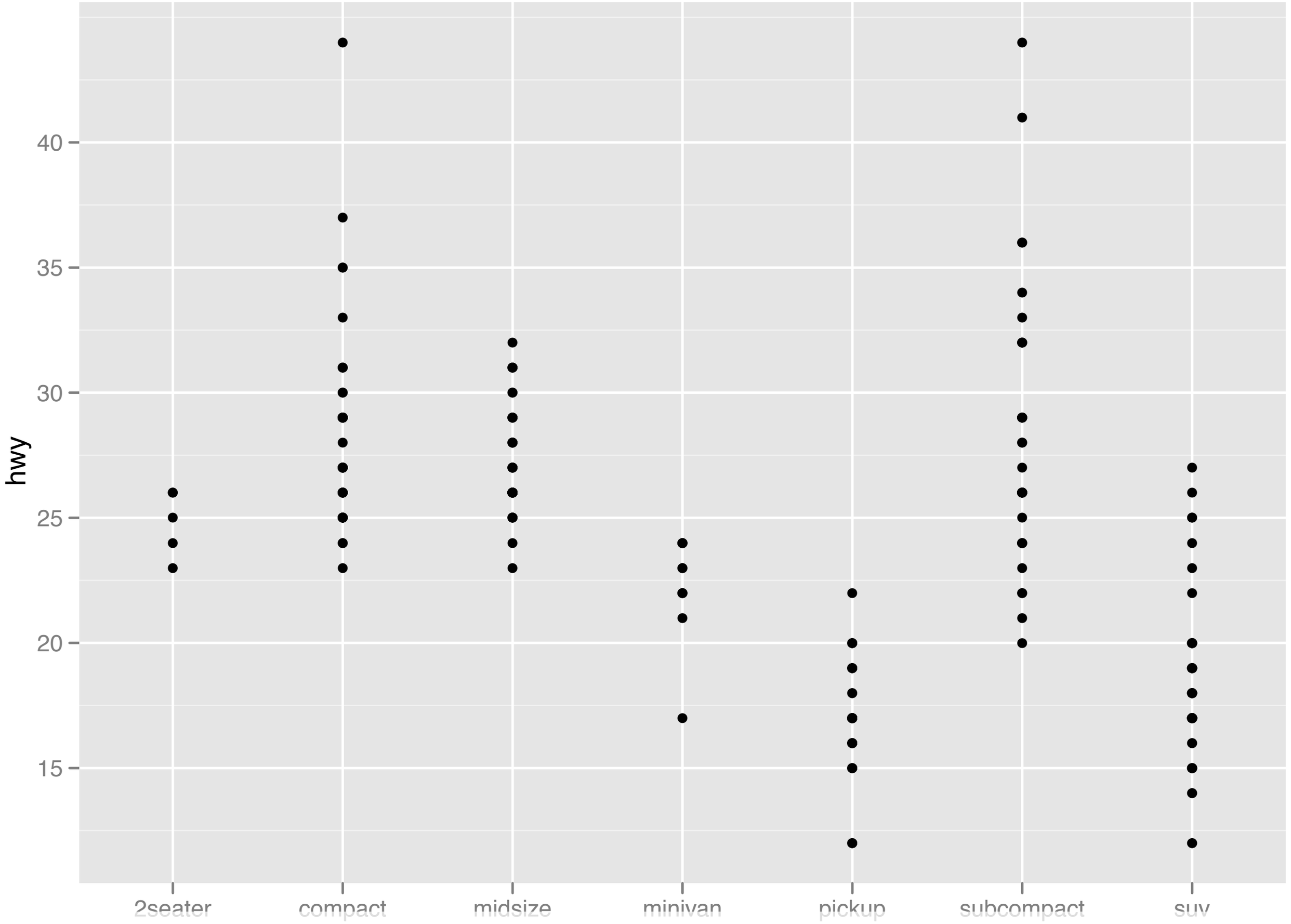

```
qplot(cty, hwy, data = mpg)
```




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



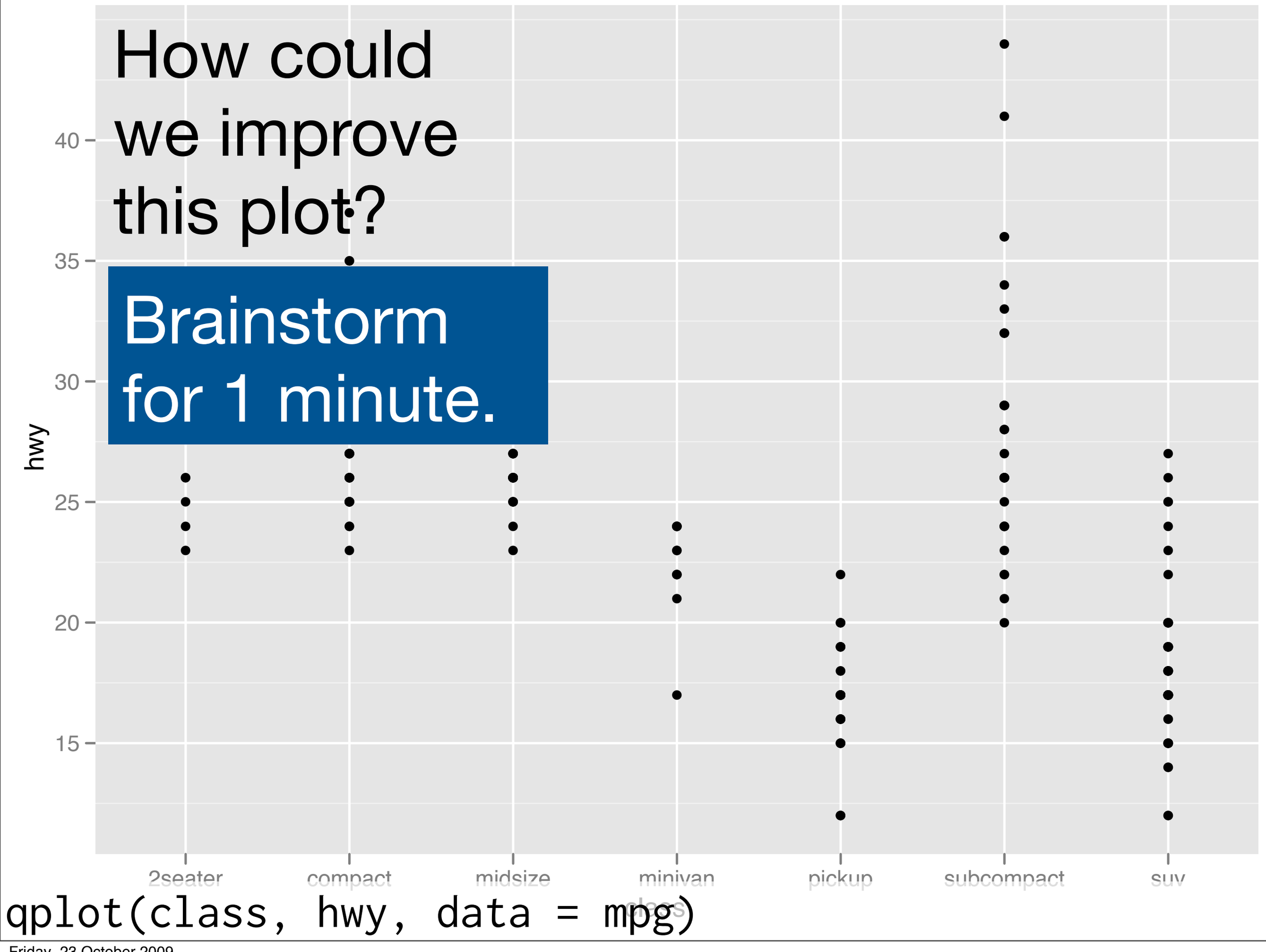

```
qplot(cty, hwy, data = mpg, geom = "jitter")
```



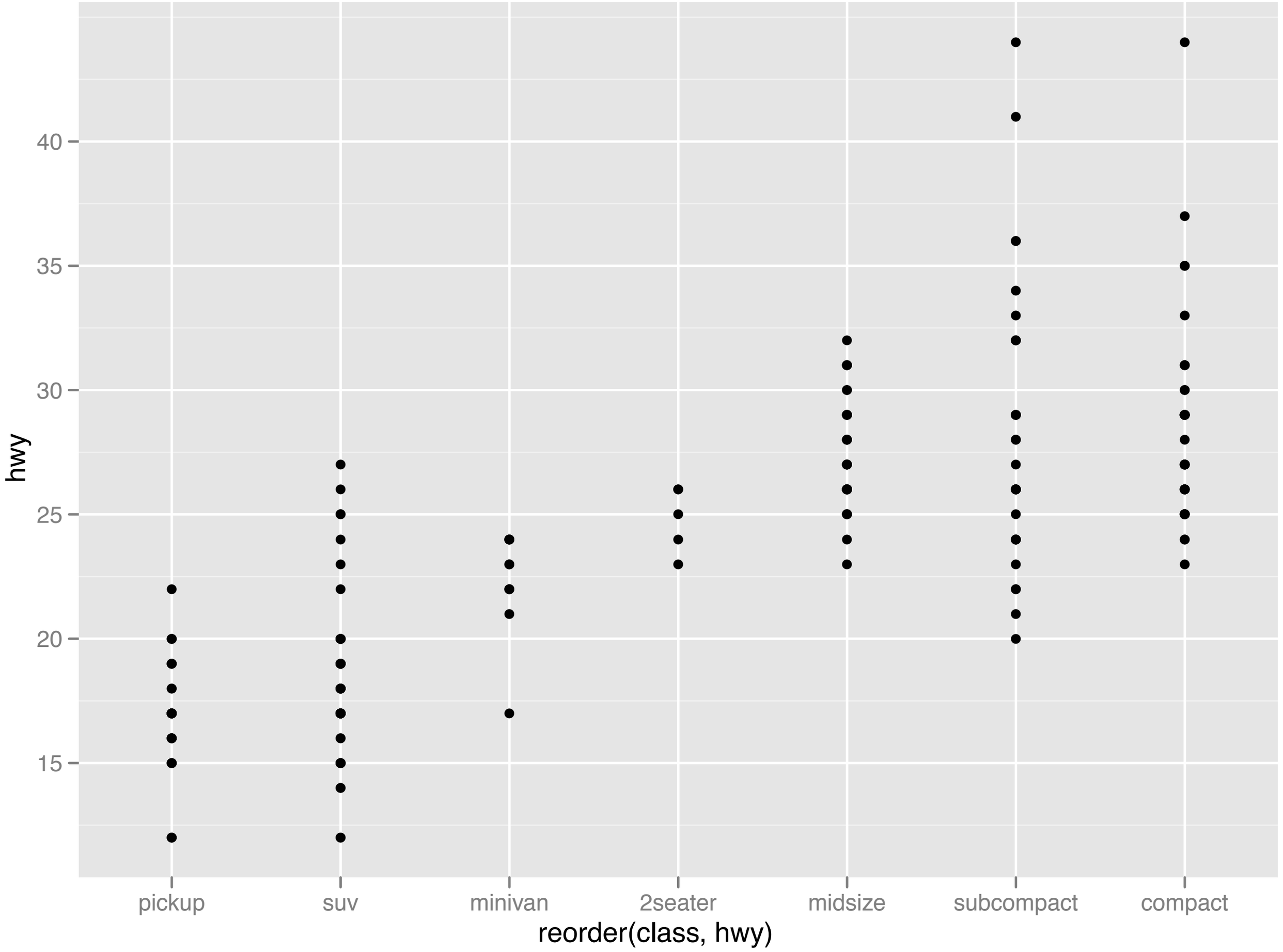

```
qplot(class, hwy, data = mpg)
```

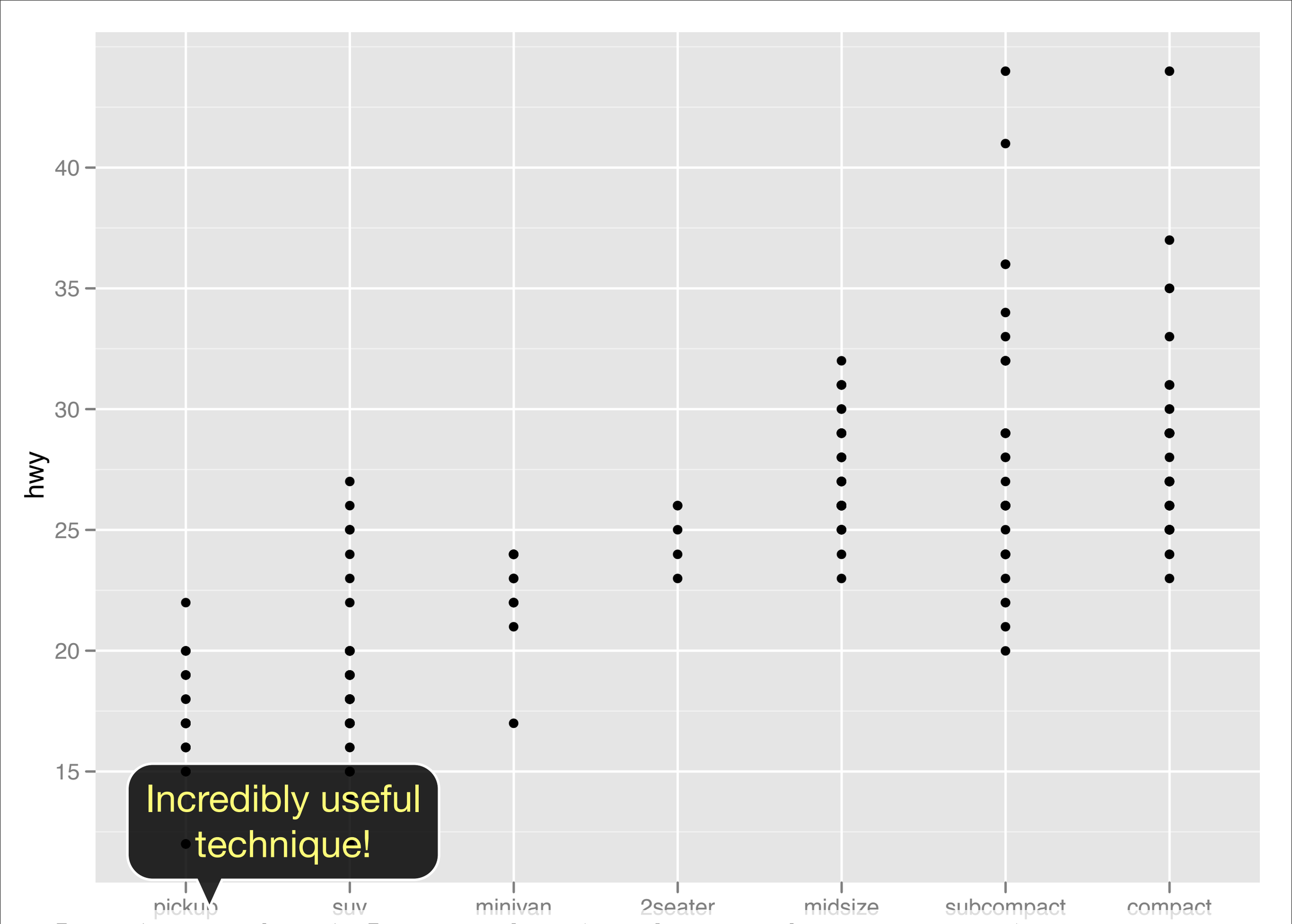
How could we improve this plot?

Brainstorm for 1 minute.

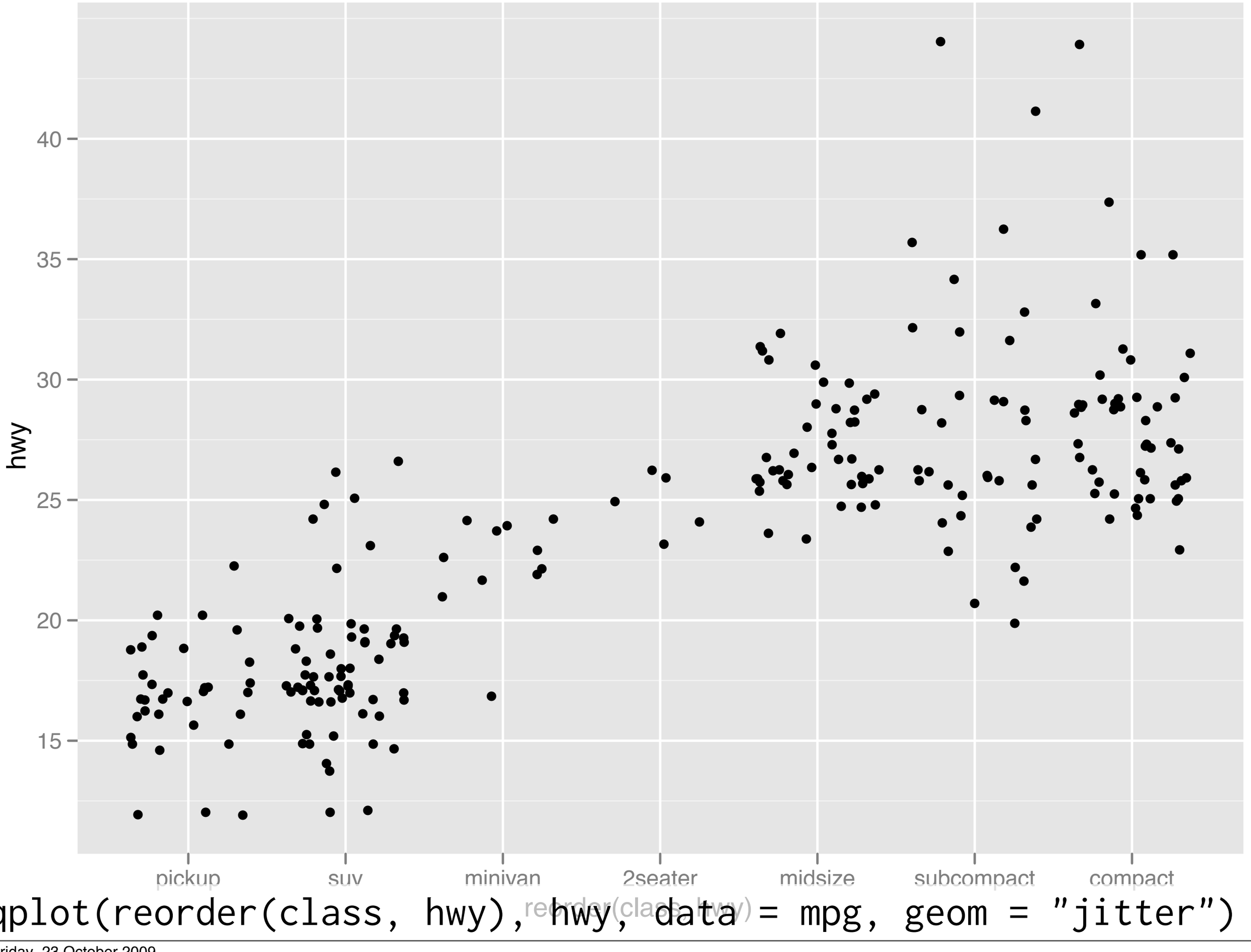



```
qplot(class, hwy, data = mpg)
```

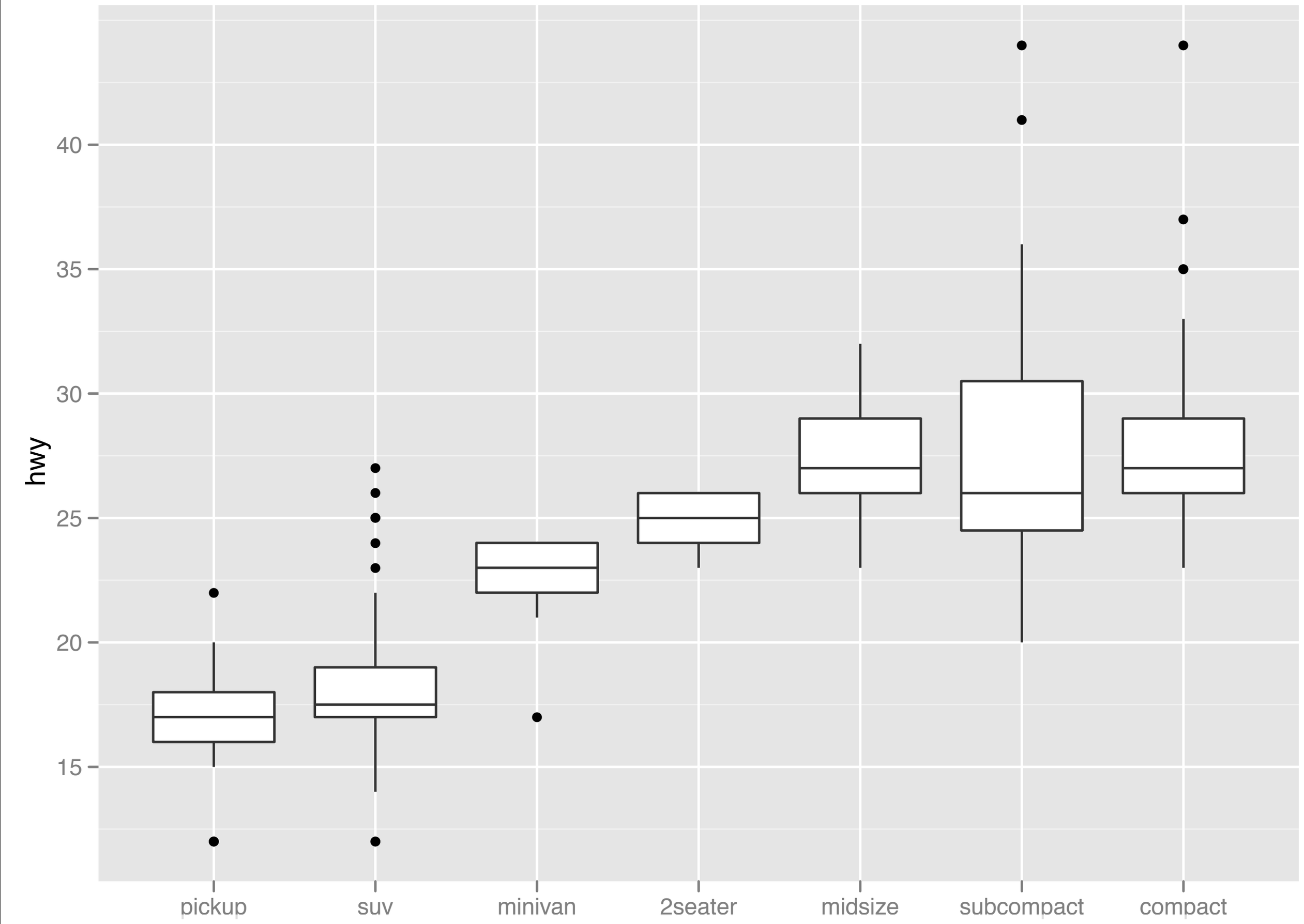

hwy
40
35
30
25
20
15
pickup
suv
minivan
2seater
midsize
subcompact
compact
reorder(class, hwy)



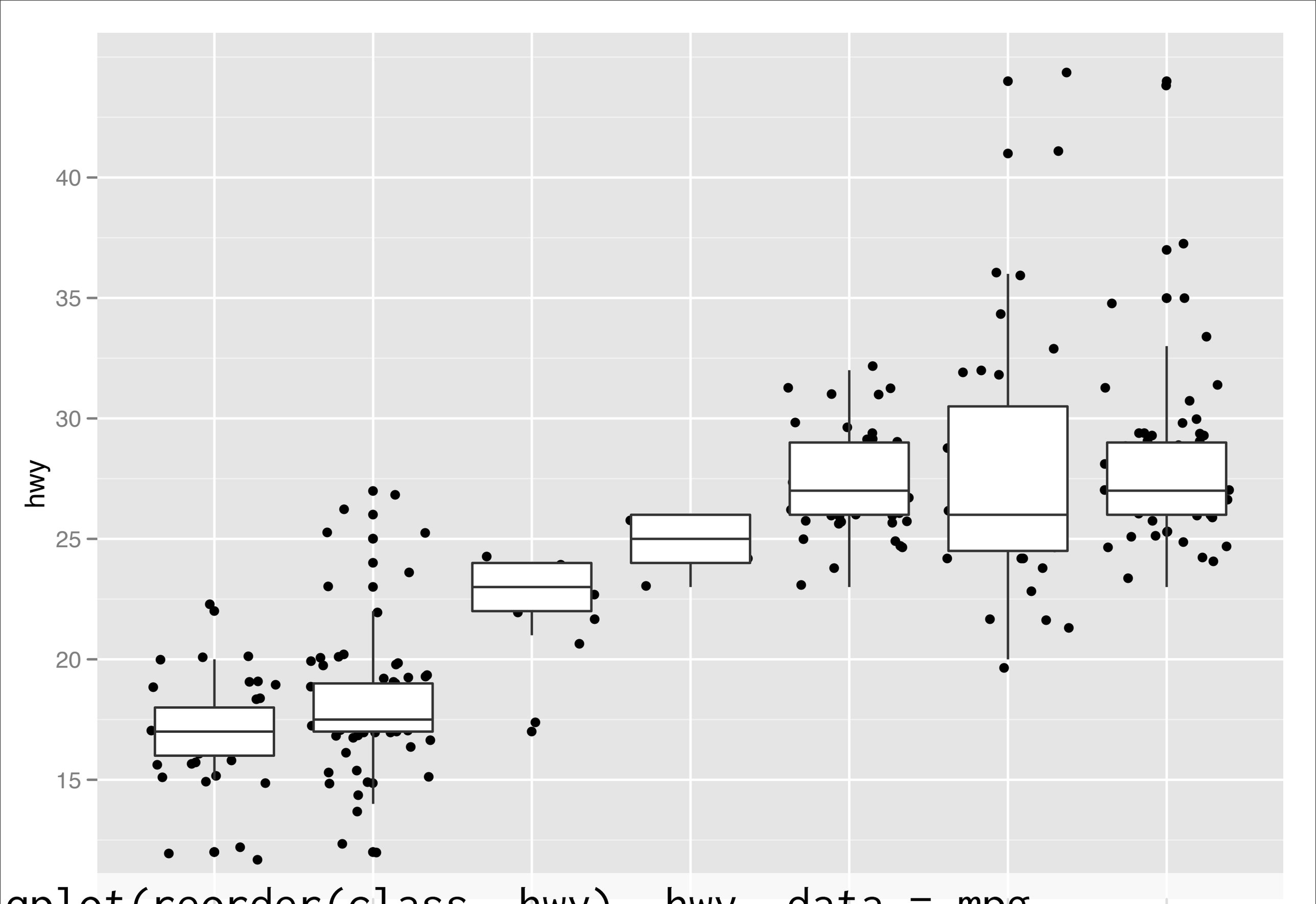

```
qplot(reorder(class, hwy), hwy, data = mpg)
```




```
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 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)