

Stat405

Statistical computing & graphics

Hadley Wickham

1. Introductions

2. Syllabus

3. Introduction to linux

4. Introduction to R

5. Basic graphics

HELLO

my name is

Hadley

had.co.nz/stat405

(if you can't remember just google stat405)

hadley@rice.edu

About me

From New Zealand

Divisional advisor for McMurtry

Major advisor for statistics

Syllabus

Introduction to linux

Essential tools

The **terminal** to run R. **gedit** to edit your R code.

To load the terminal, right-click on the desktop.

To load R, type **R** in the terminal. To load gedit, type **gedit &** in the terminal (the & tells it to run separately). To open a file in gedit, type **gedit filename &**

Setup

Work through the instructions at <http://had.co.nz/stat405/linux.html>.

I'll circulate and make sure everyone gets set up right.

Terminal essentials

Mouse select = Copy

Middle button = Paste

Ctrl + A = home

Ctrl + D = end

Alt + tab = change applications

Press tab to complete file names

Introduction to R



Learning a new language is hard!

Scatterplot basics

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

```
?mpg
```

```
head(mpg)
```

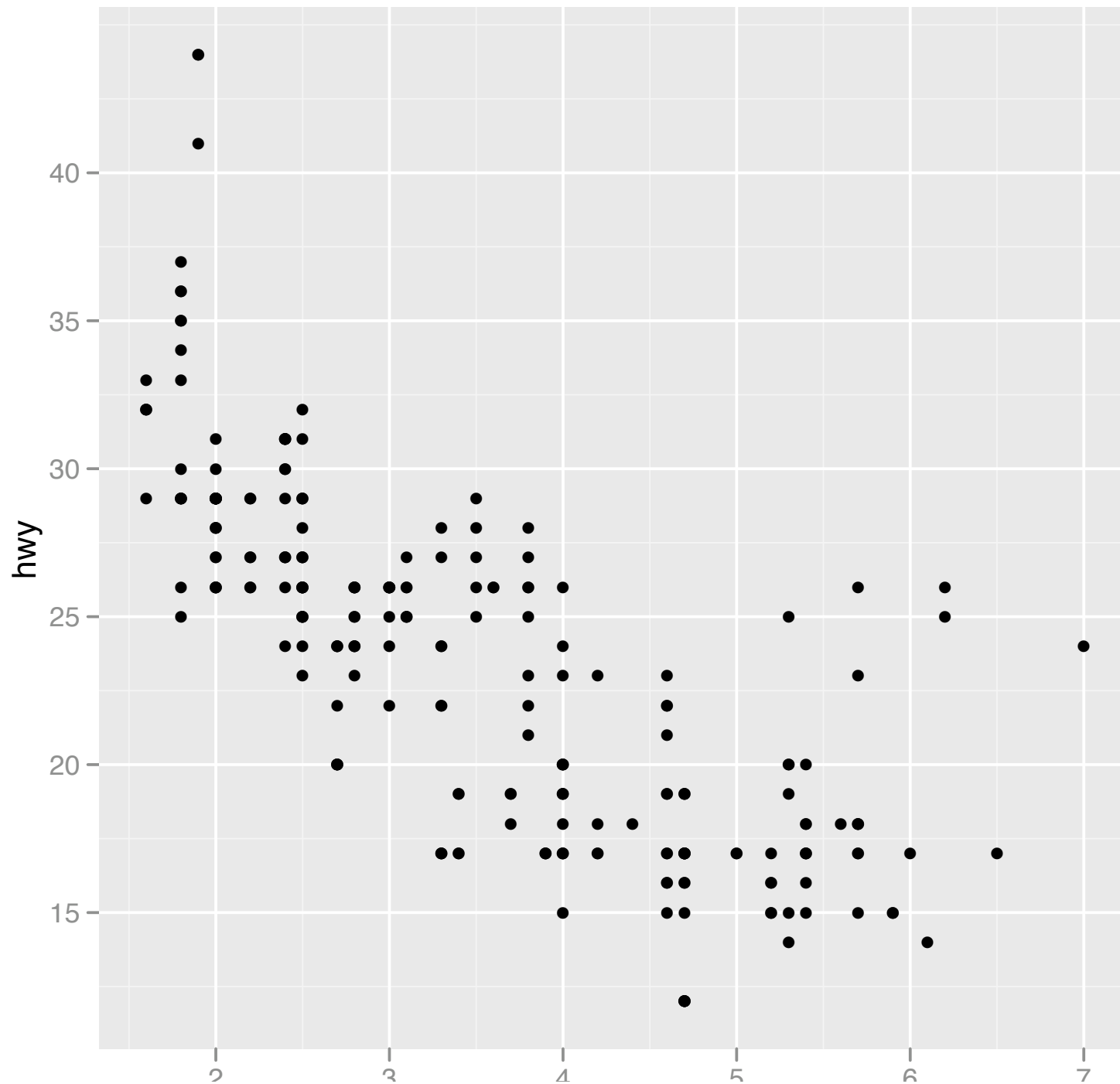
```
str(mpg)
```

```
summary(mpg)
```

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



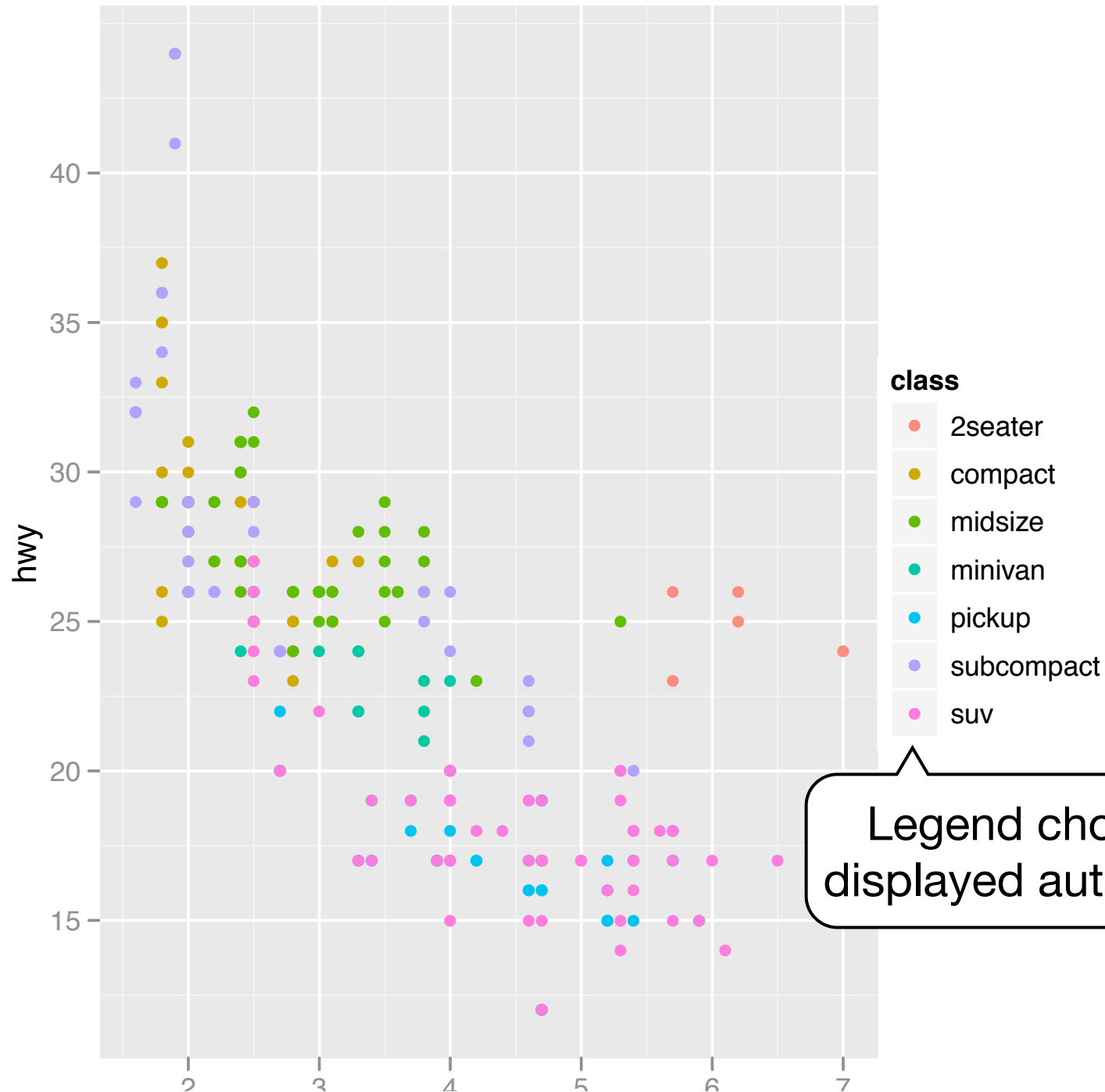
Always explicitly
specify the data



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

Additional variables

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



```
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

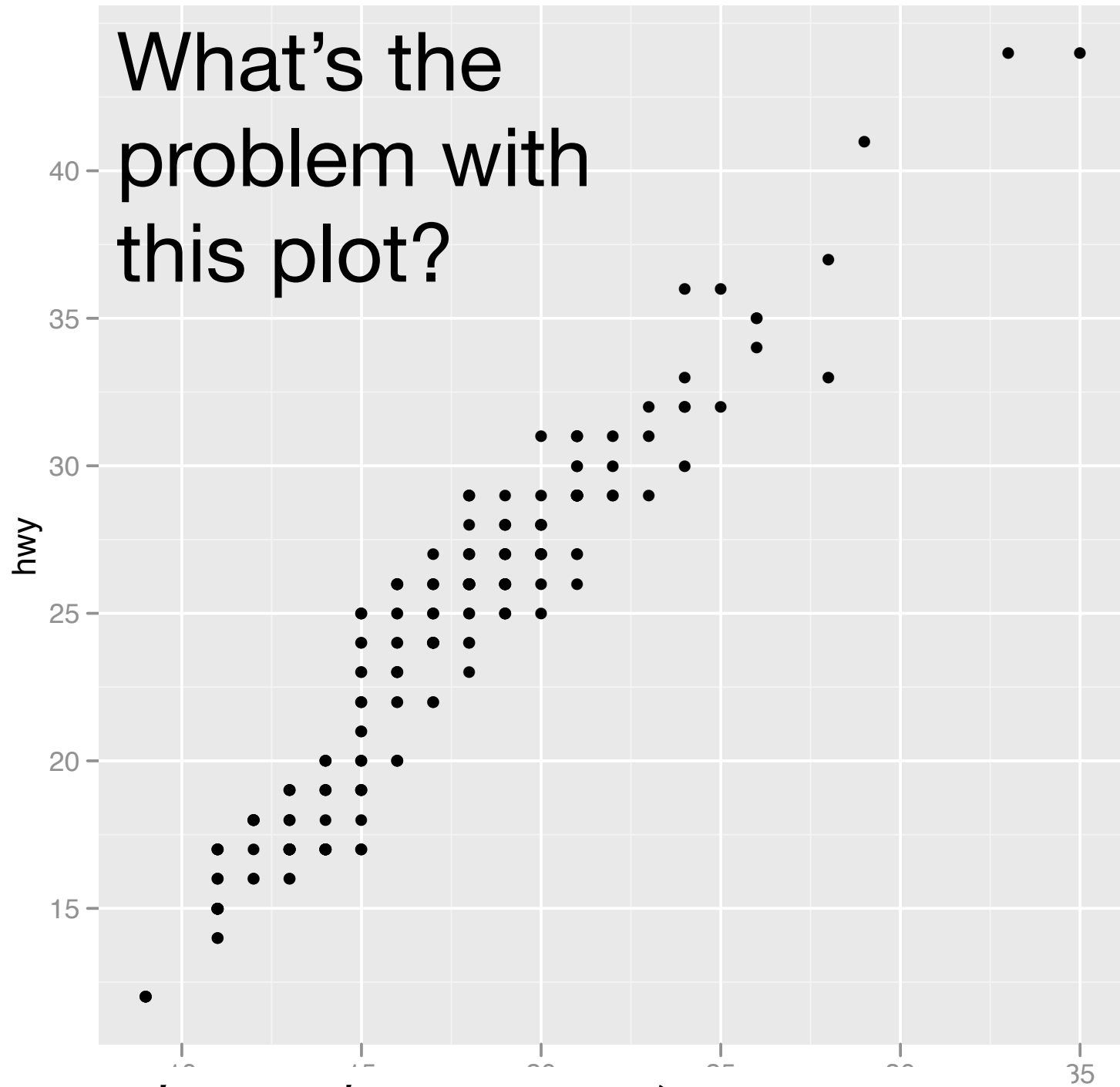
`facet_wrap()`: 1d ribbon wrapped into 2d

Aside: workflow

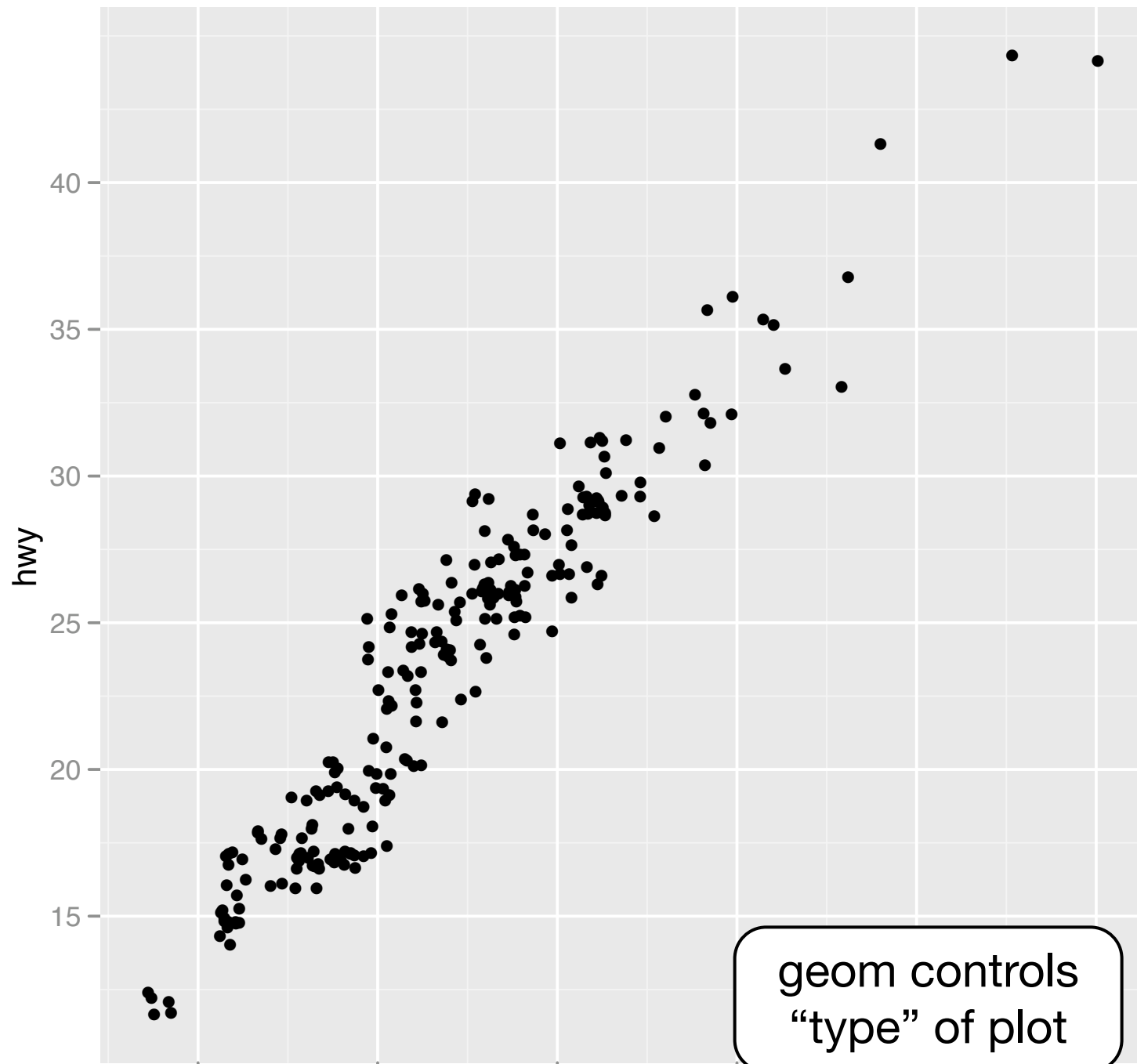
Keep a copy of the slides open so that you can copy and paste the code.

For complicated commands, write them in gedit and then copy and paste.

What's the
problem with
this plot?



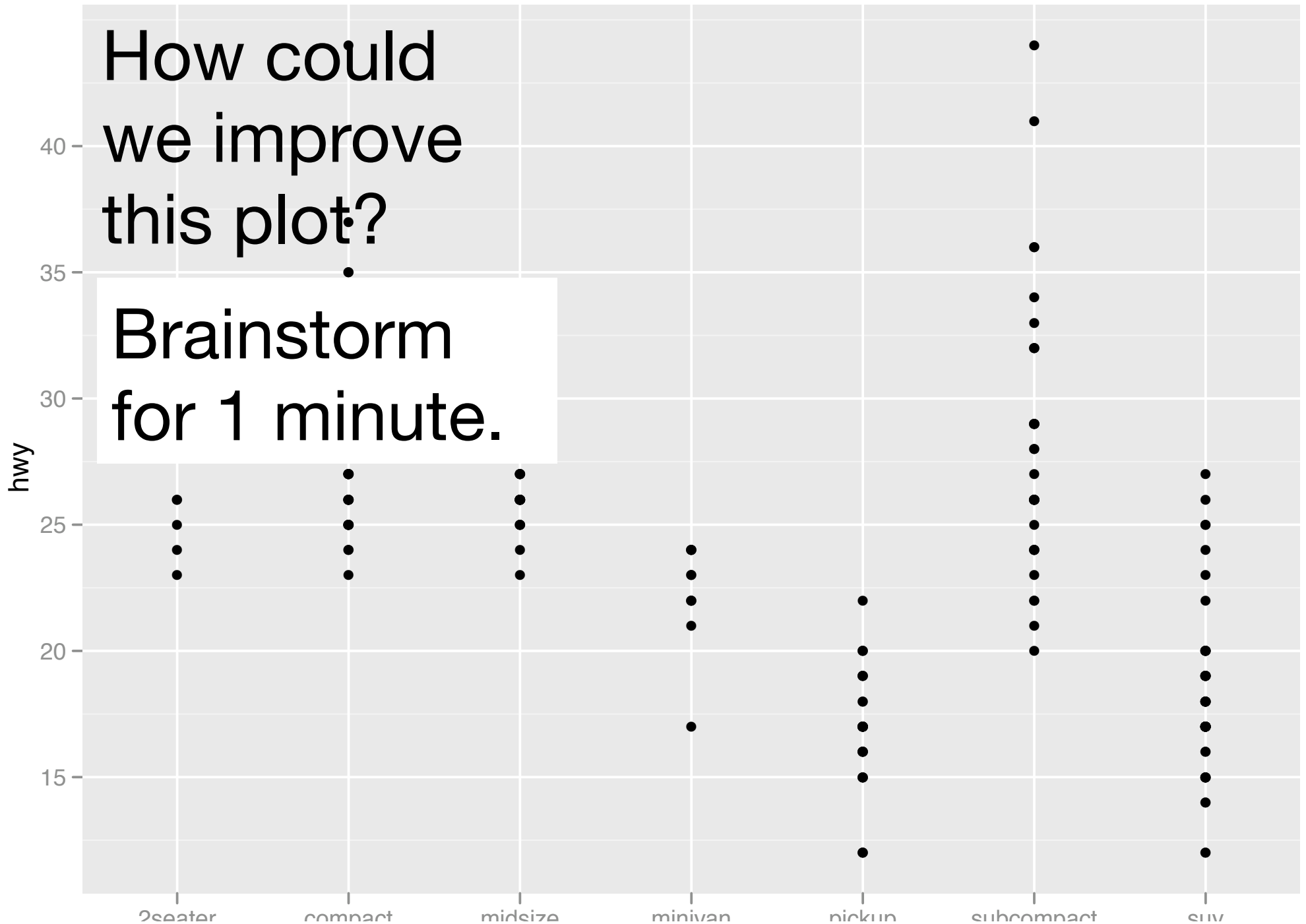
```
qplot(cty, hwy, data = mpg)
```



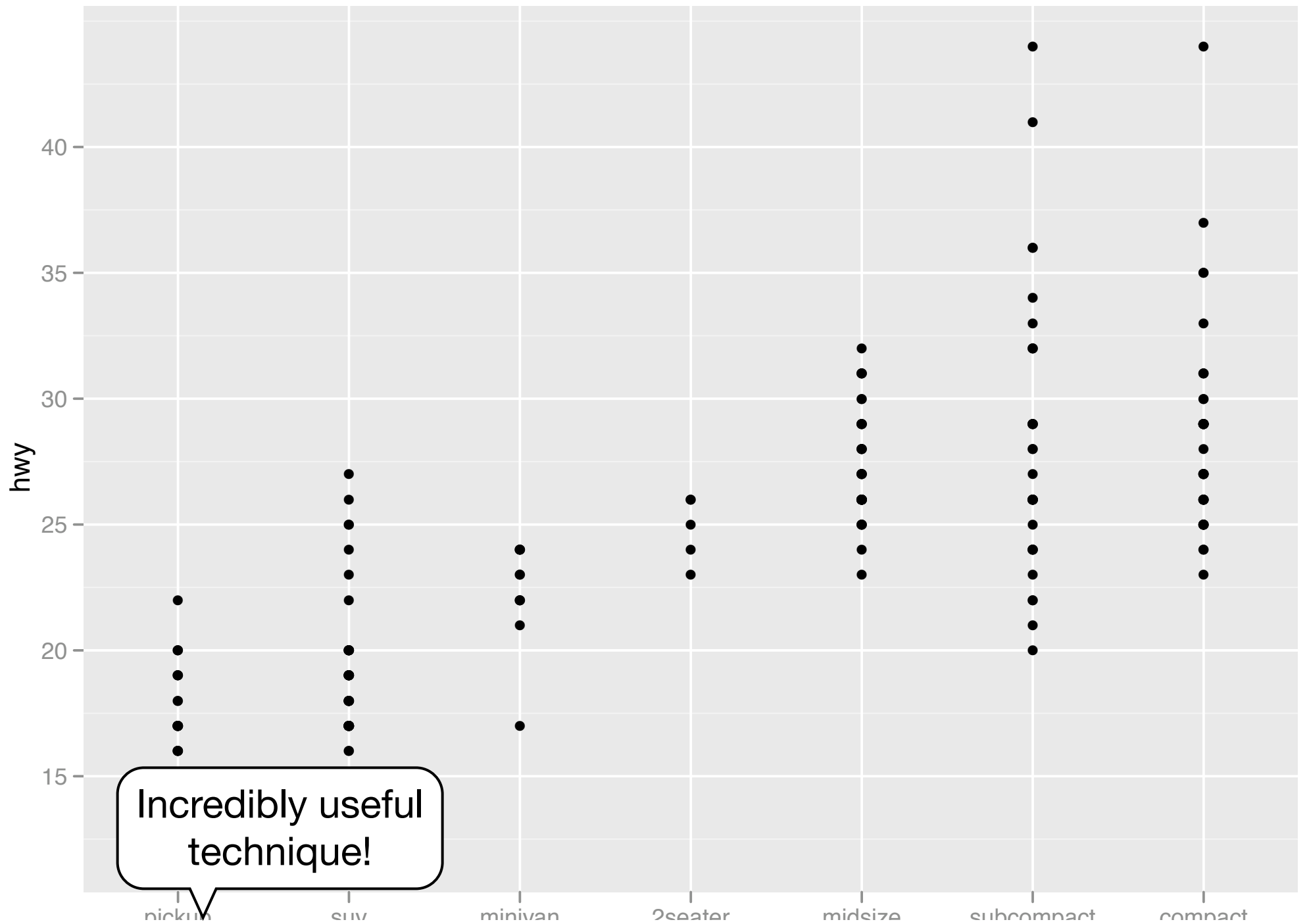
```
qplot(cty, hwy, data = mpg, geom = "jitter")
```

How could
we improve
this plot?

Brainstorm
for 1 minute.

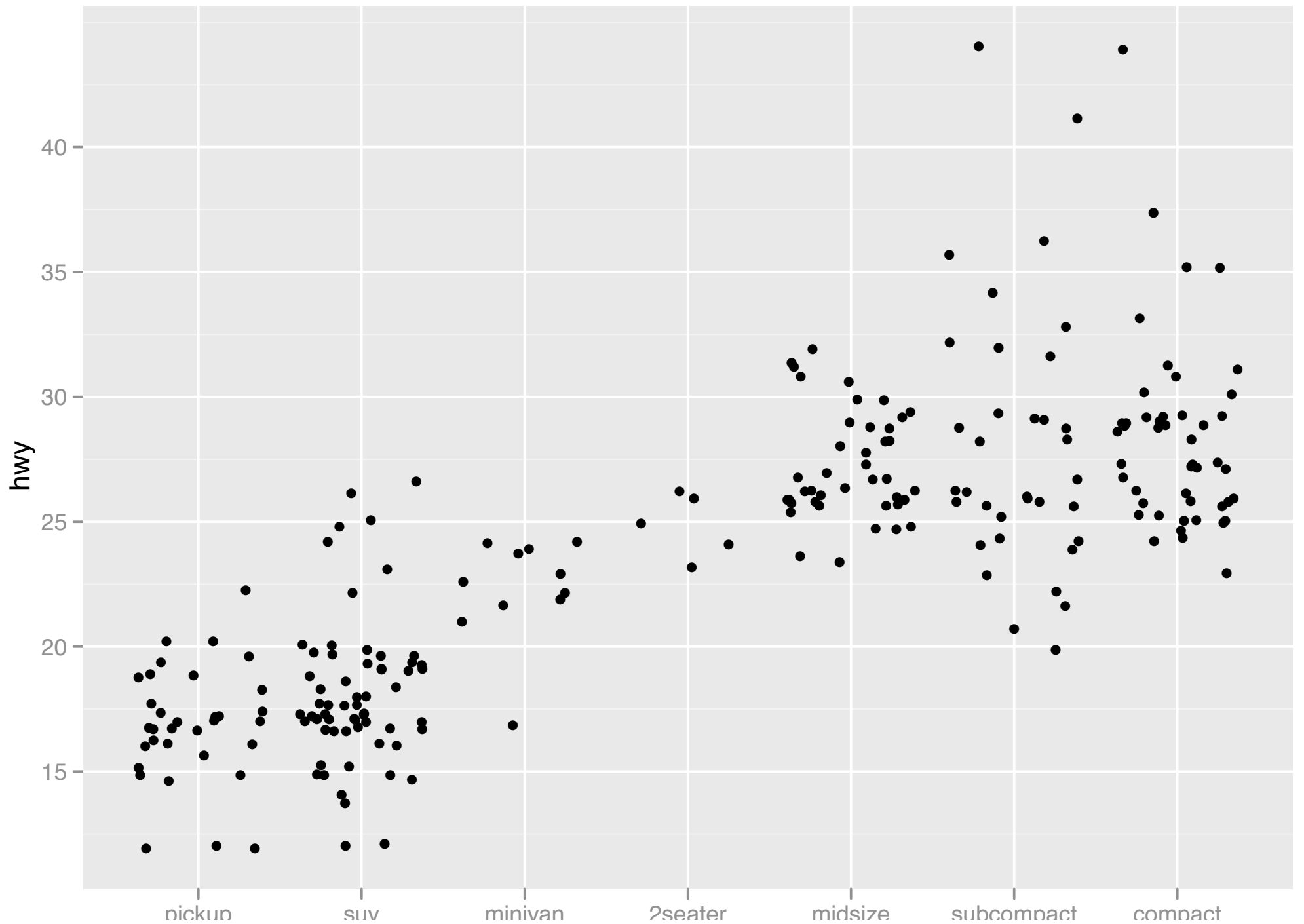


```
qplot(class, hwy, data = mpg)
```

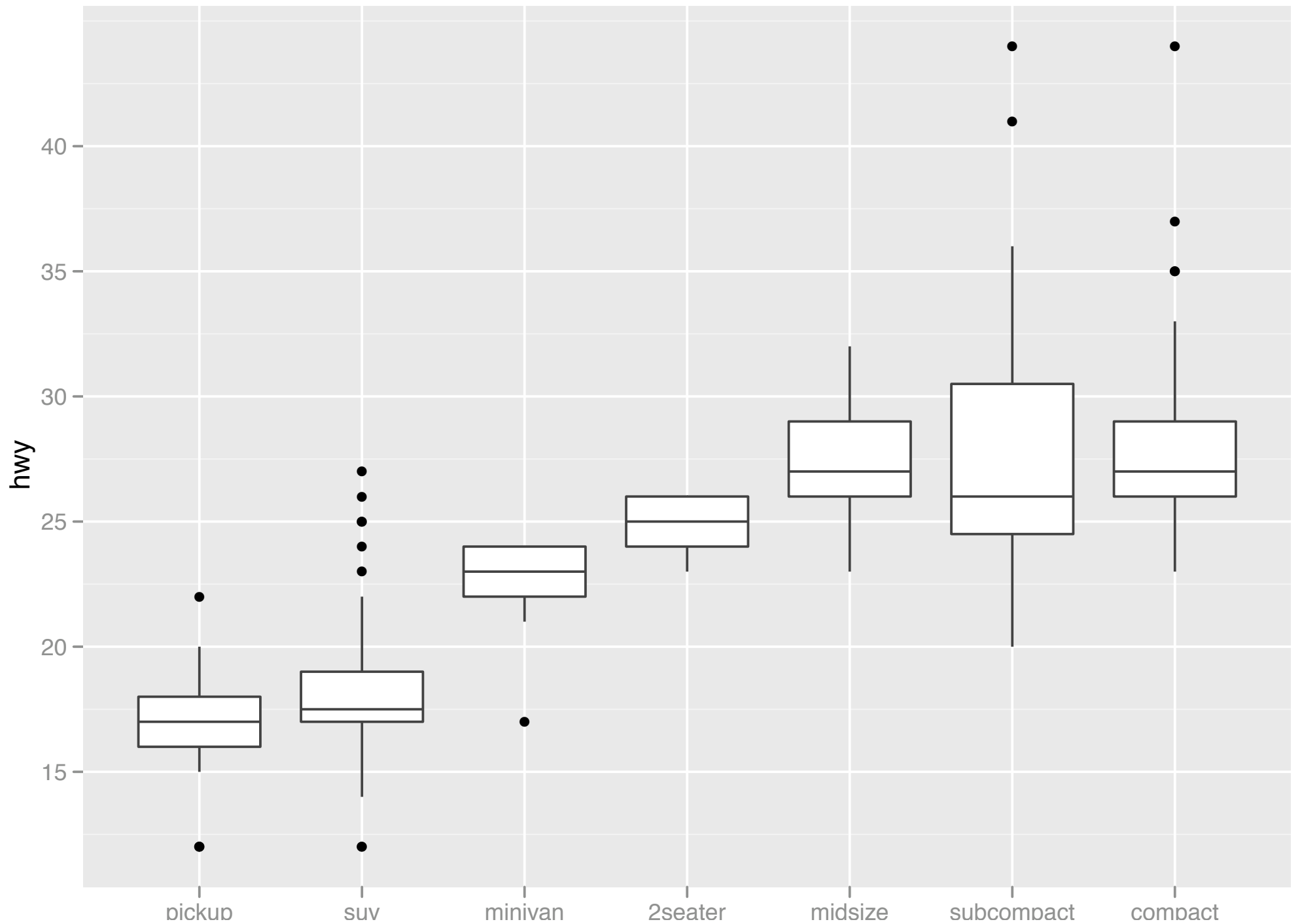


Incredibly useful
technique!

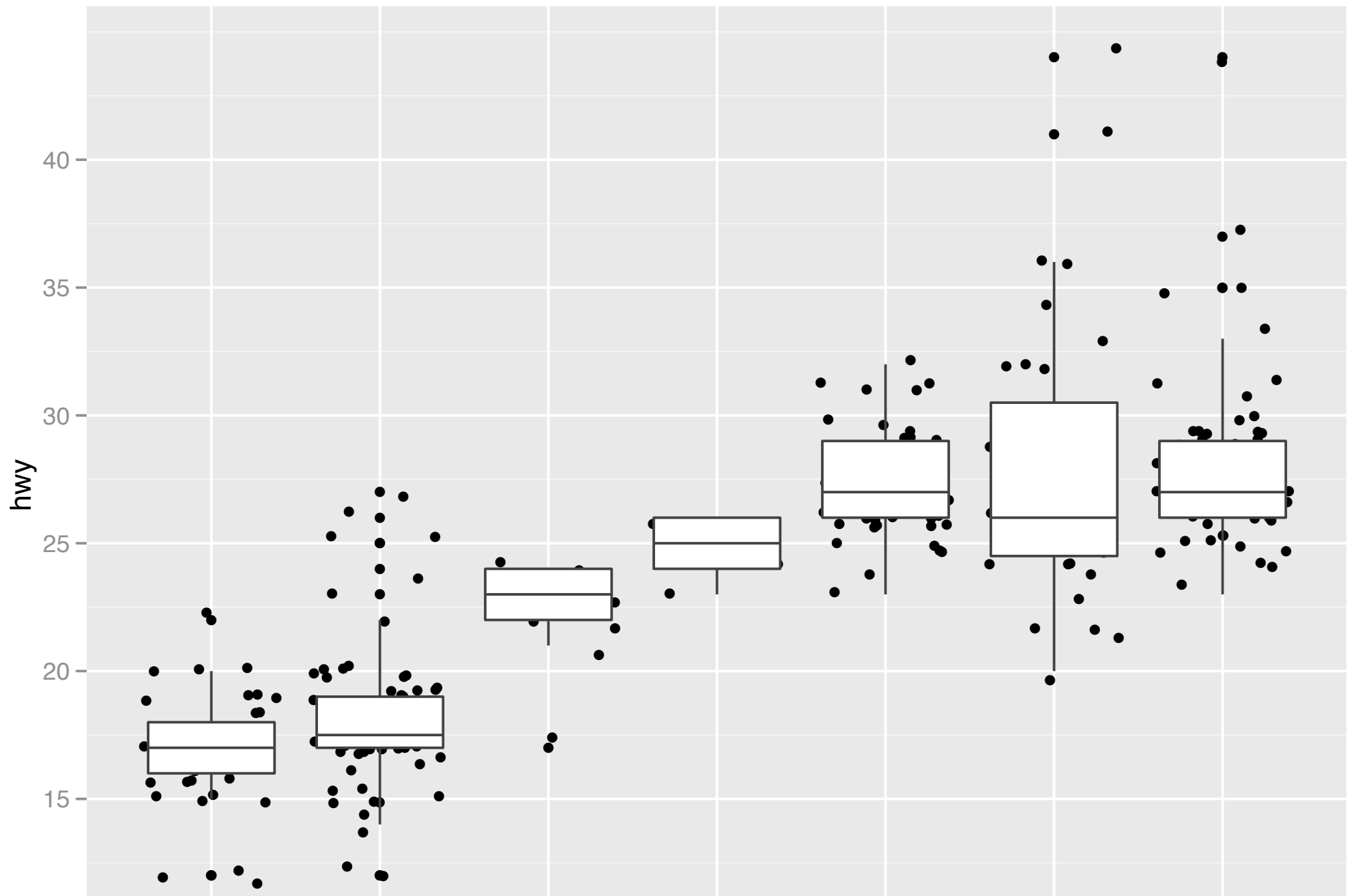
```
qplot(reorder(class, hwy), hwy, data = mpg)
```



```
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 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)