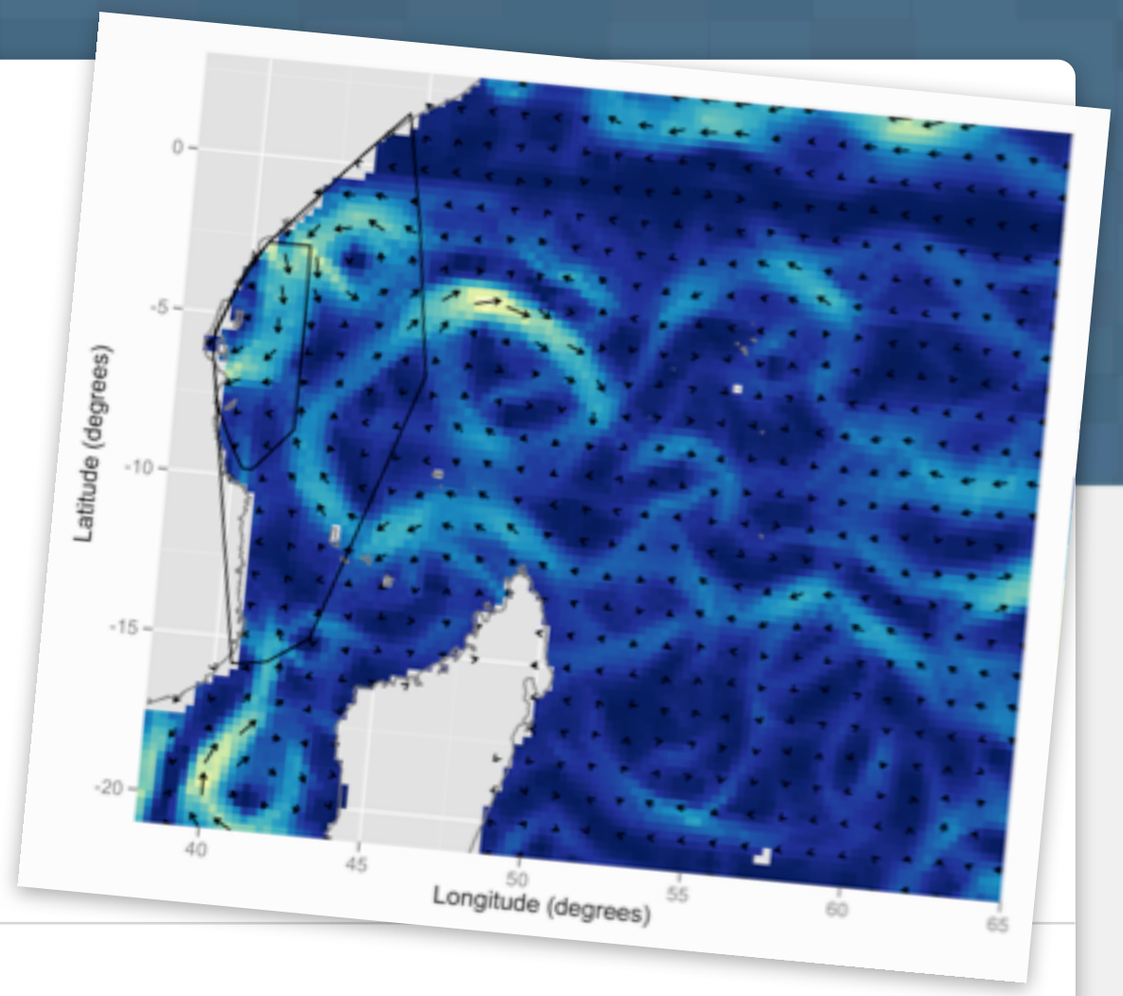
RStudio

Intro to ggplot2

http://bit.ly/ggplot2ubc



Hadley Wickham

Chief Scientist, RStudio

May 2013

1. Rstudio
2. Diving in: scatterplots & aesthetics
3. Facetting and geoms
4. Diamonds data
5. Bar charts and histograms
6. Big(ger) data
7. Where next

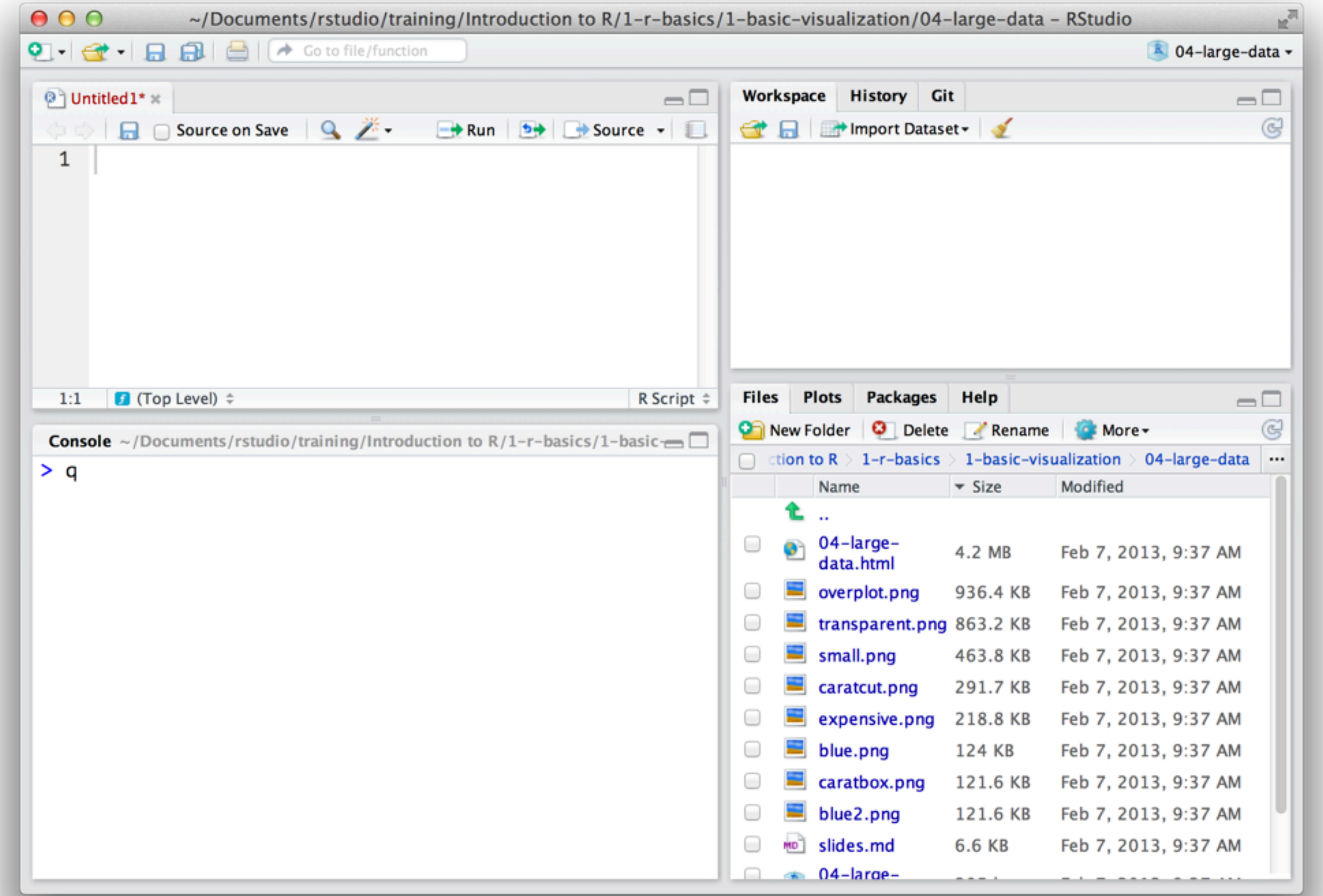
http://bit.ly/ggplot2ubc

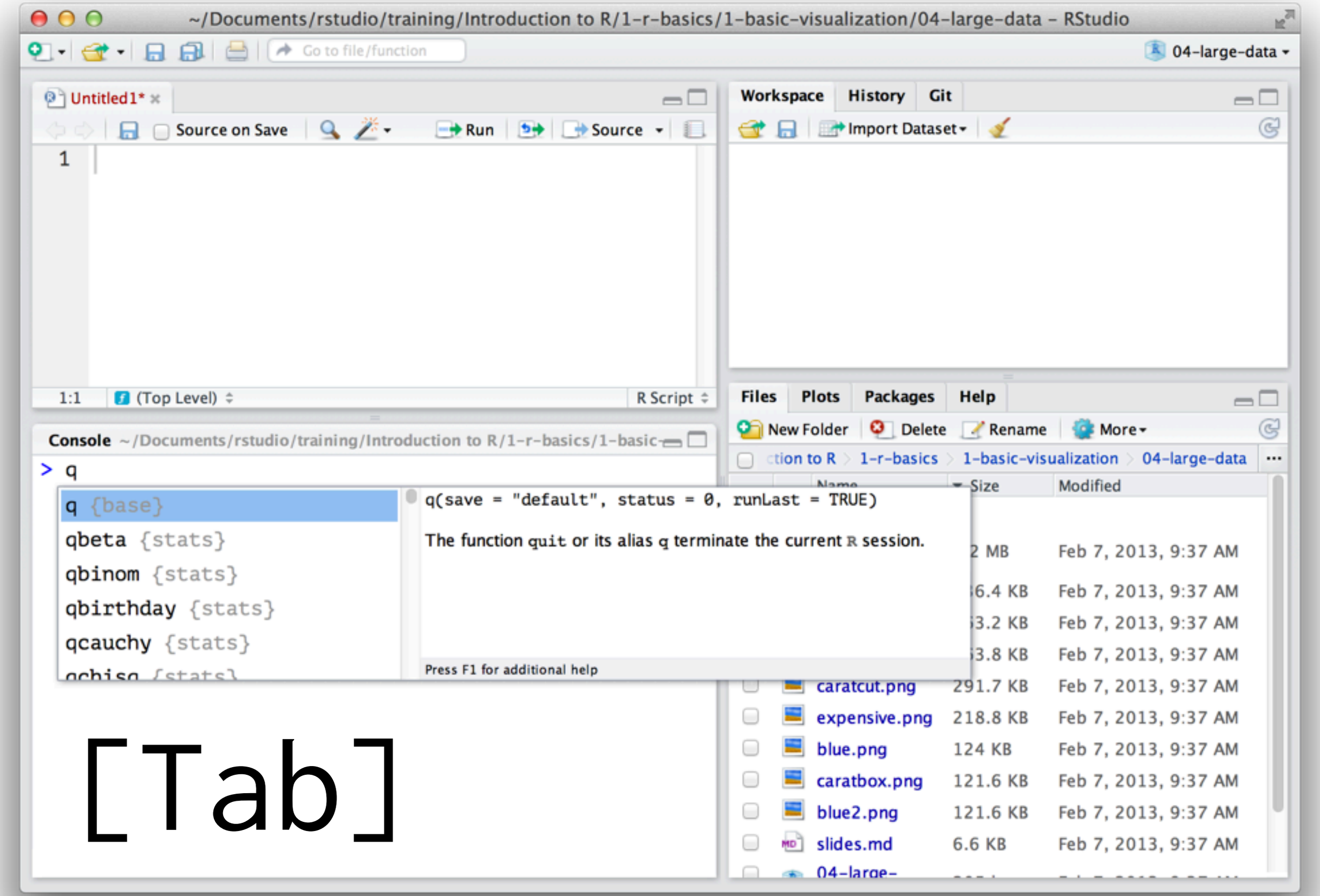
Your turn

http://bit.ly/ggplot2ubc

Please introduce yourself to your neighbours. You'll be working with them! Why are you interested in ggplot2?

Rstudio


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-
> q
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
tion to R
1-r-basics
1-basic-visualization
04-large-data
Name
Size
Modified
..
04-large-data.html 4.2 MB Feb 7, 2013, 9:37 AM
overplot.png 936.4 KB Feb 7, 2013, 9:37 AM
transparent.png 863.2 KB Feb 7, 2013, 9:37 AM
small.png 463.8 KB Feb 7, 2013, 9:37 AM
caratcut.png 291.7 KB Feb 7, 2013, 9:37 AM
expensive.png 218.8 KB Feb 7, 2013, 9:37 AM
blue.png 124 KB Feb 7, 2013, 9:37 AM
caratbox.png 121.6 KB Feb 7, 2013, 9:37 AM
blue2.png 121.6 KB Feb 7, 2013, 9:37 AM
slides.md 6.6 KB Feb 7, 2013, 9:37 AM
04-large-


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
1-r-basics
1-basic-visualization
04-large-data
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-
> q
q {base}
qbeta {stats}
qbinom {stats}
qbirthday {stats}
qcauchy {stats}
q(save = "default", status = 0, runLast = TRUE)
The function quit or its alias q terminate the current R session.
Press F1 for additional help
Size
Modified
2 MB
Feb 7, 2013, 9:37 AM
6.4 KB
Feb 7, 2013, 9:37 AM
3.2 KB
Feb 7, 2013, 9:37 AM
3.8 KB
Feb 7, 2013, 9:37 AM
caratcut.png
291.7 KB
Feb 7, 2013, 9:37 AM
expensive.png
218.8 KB
Feb 7, 2013, 9:37 AM
blue.png
124 KB
Feb 7, 2013, 9:37 AM
caratbox.png
121.6 KB
Feb 7, 2013, 9:37 AM
blue2.png
121.6 KB
Feb 7, 2013, 9:37 AM
slides.md
6.6 KB
Feb 7, 2013, 9:37 AM
[Tab]

~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio

Go to file/function

04-large-data

Untitled1*

Source on Save | Run | Source

```
qplot(table - depth, data = diamonds,
qplot(day, data = email)
qplot(day, mails, data = daily, geom = "line", colo
qplot(day, mails, data = daily, geom = "smooth", co
qplot(day, variants, data = daily, geom = "line", c
qplot(wday, hour, data = wh, size = freq)
qplot(mpg, wt, data = mtcars)
qplot(mpg, wt, data = mtcars, colour = cyl)
```

```
> q
```

[Cmd + ↑]

Workspace | History | Git

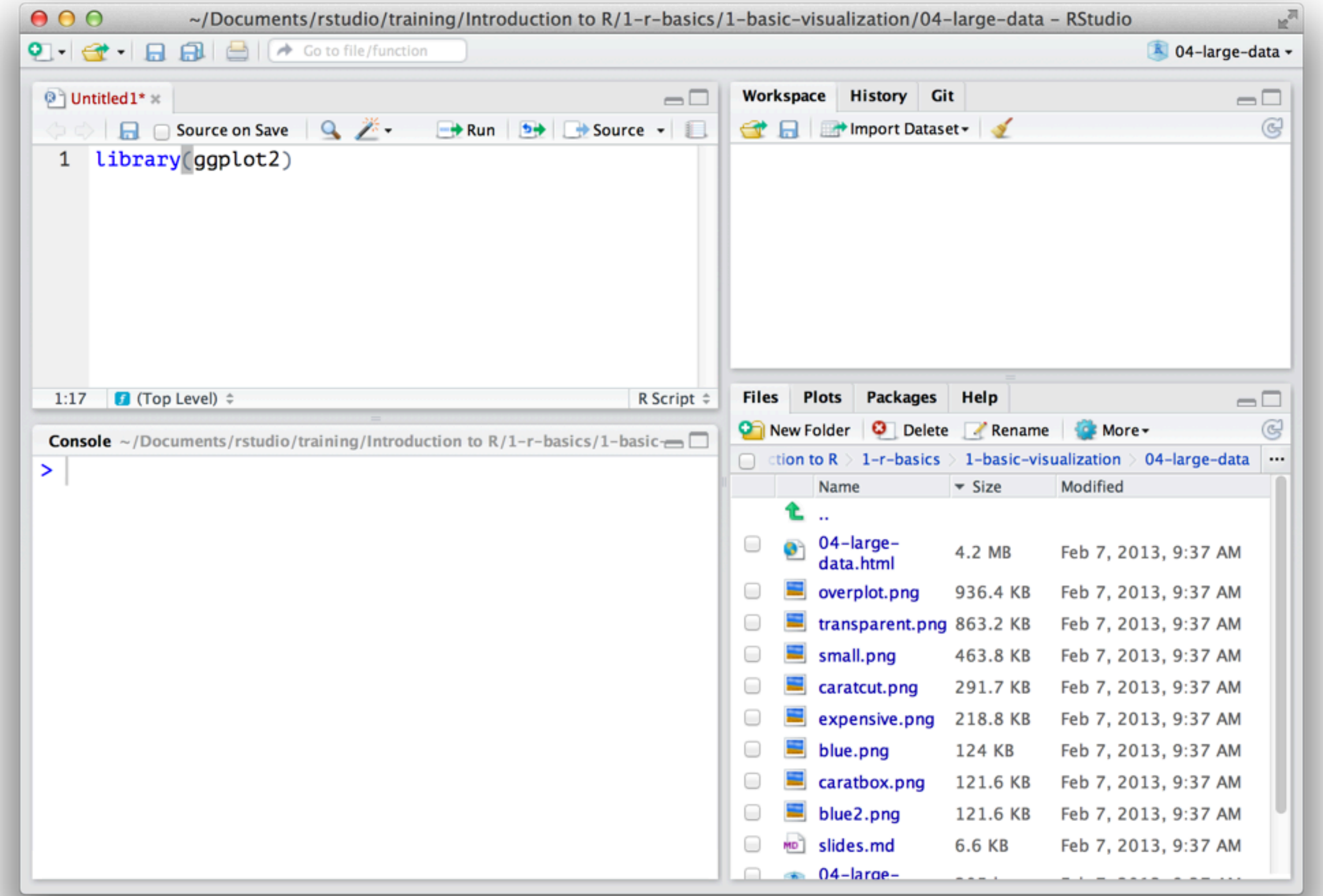
Import Dataset

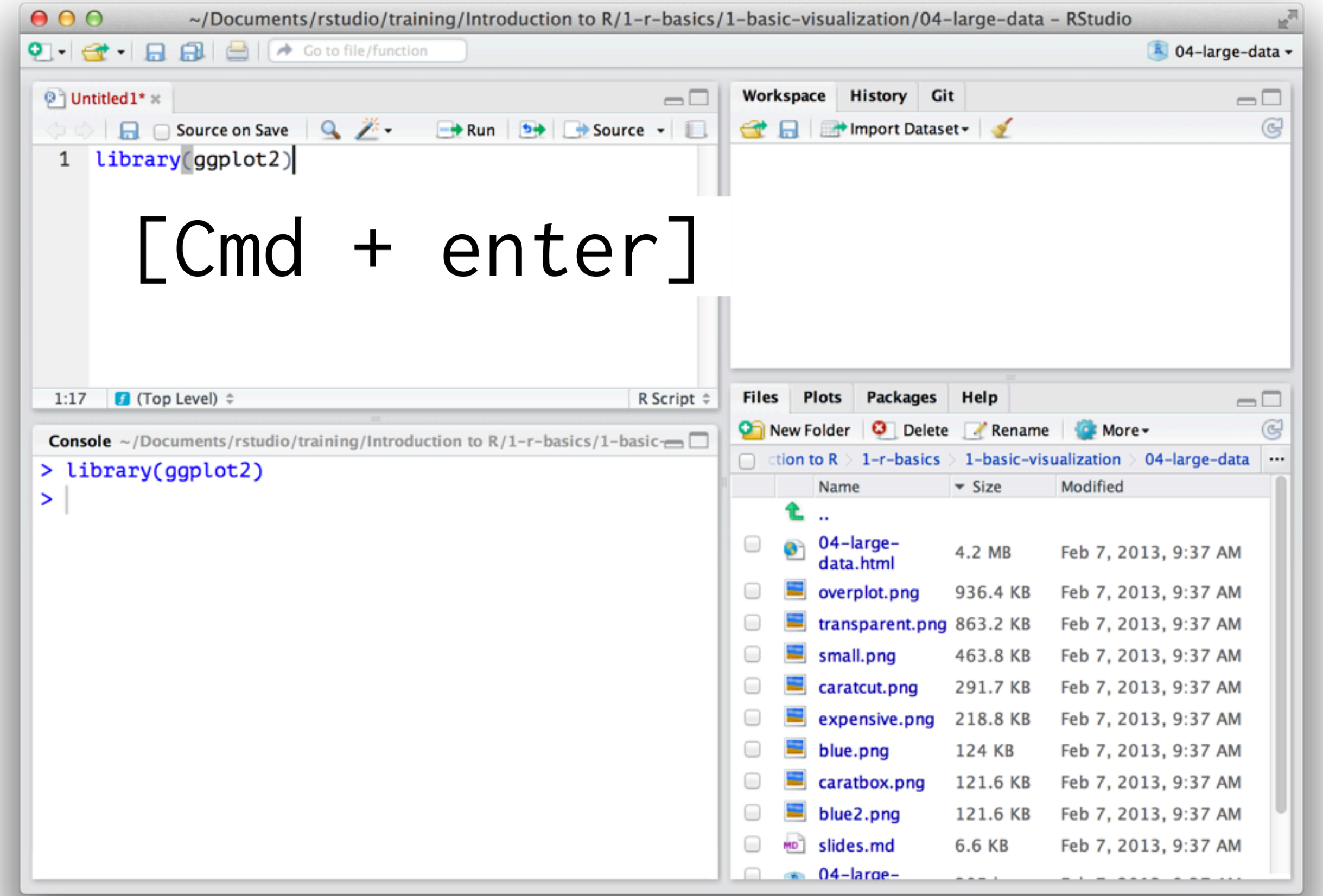
Files | Plots | Packages | Help

New Folder | Delete | Rename | More

tion to R › 1-r-basics › 1-basic-visualization › 04-large-data

Name	Size	Modified
..		
04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
small.png	463.8 KB	Feb 7, 2013, 9:37 AM
caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
blue.png	124 KB	Feb 7, 2013, 9:37 AM
caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
04-large-		


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1 library(ggplot2)
1:17
(Top Level)
R Script
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-
>
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
1-r-basics
1-basic-visualization
04-large-data
Name
Size
Modified
..
04-large-data.html 4.2 MB Feb 7, 2013, 9:37 AM
overplot.png 936.4 KB Feb 7, 2013, 9:37 AM
transparent.png 863.2 KB Feb 7, 2013, 9:37 AM
small.png 463.8 KB Feb 7, 2013, 9:37 AM
caratcut.png 291.7 KB Feb 7, 2013, 9:37 AM
expensive.png 218.8 KB Feb 7, 2013, 9:37 AM
blue.png 124 KB Feb 7, 2013, 9:37 AM
caratbox.png 121.6 KB Feb 7, 2013, 9:37 AM
blue2.png 121.6 KB Feb 7, 2013, 9:37 AM
slides.md 6.6 KB Feb 7, 2013, 9:37 AM


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1 library(ggplot2)
[Cmd + enter]
1:17
(Top Level)
R Script
Workspace
History
Git
Import Dataset
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
1-r-basics
1-basic-visualization
04-large-data
Name
Size
Modified
04-large-data.html 4.2 MB Feb 7, 2013, 9:37 AM
overplot.png 936.4 KB Feb 7, 2013, 9:37 AM
transparent.png 863.2 KB Feb 7, 2013, 9:37 AM
small.png 463.8 KB Feb 7, 2013, 9:37 AM
caratcut.png 291.7 KB Feb 7, 2013, 9:37 AM
expensive.png 218.8 KB Feb 7, 2013, 9:37 AM
blue.png 124 KB Feb 7, 2013, 9:37 AM
caratbox.png 121.6 KB Feb 7, 2013, 9:37 AM
blue2.png 121.6 KB Feb 7, 2013, 9:37 AM
slides.md 6.6 KB Feb 7, 2013, 9:37 AM
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic
> library(ggplot2)
>

Diving in


Learning a new
language is hard!

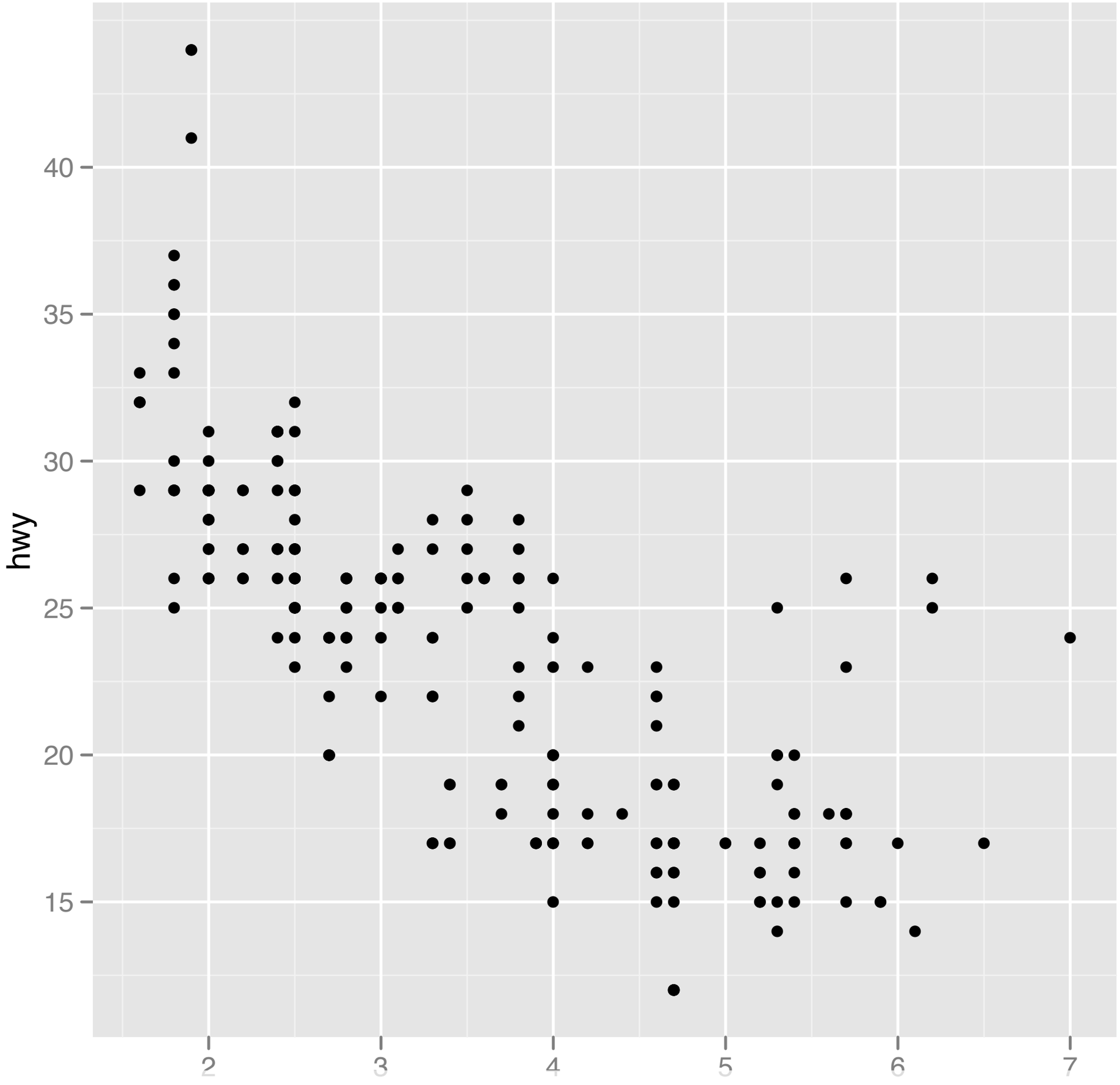
Scatterplot basics

```
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)
```

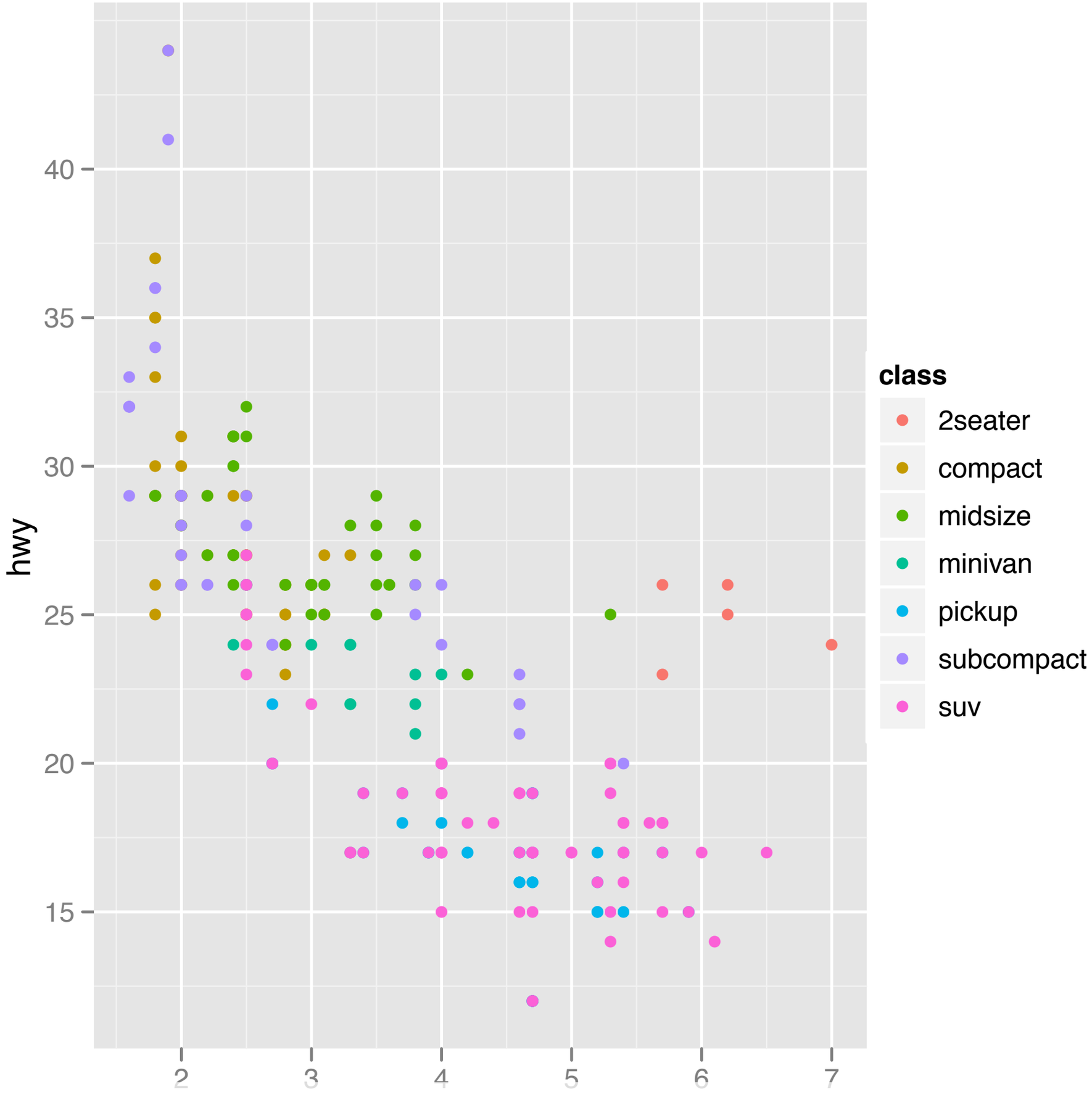
Always explicitly specify the data



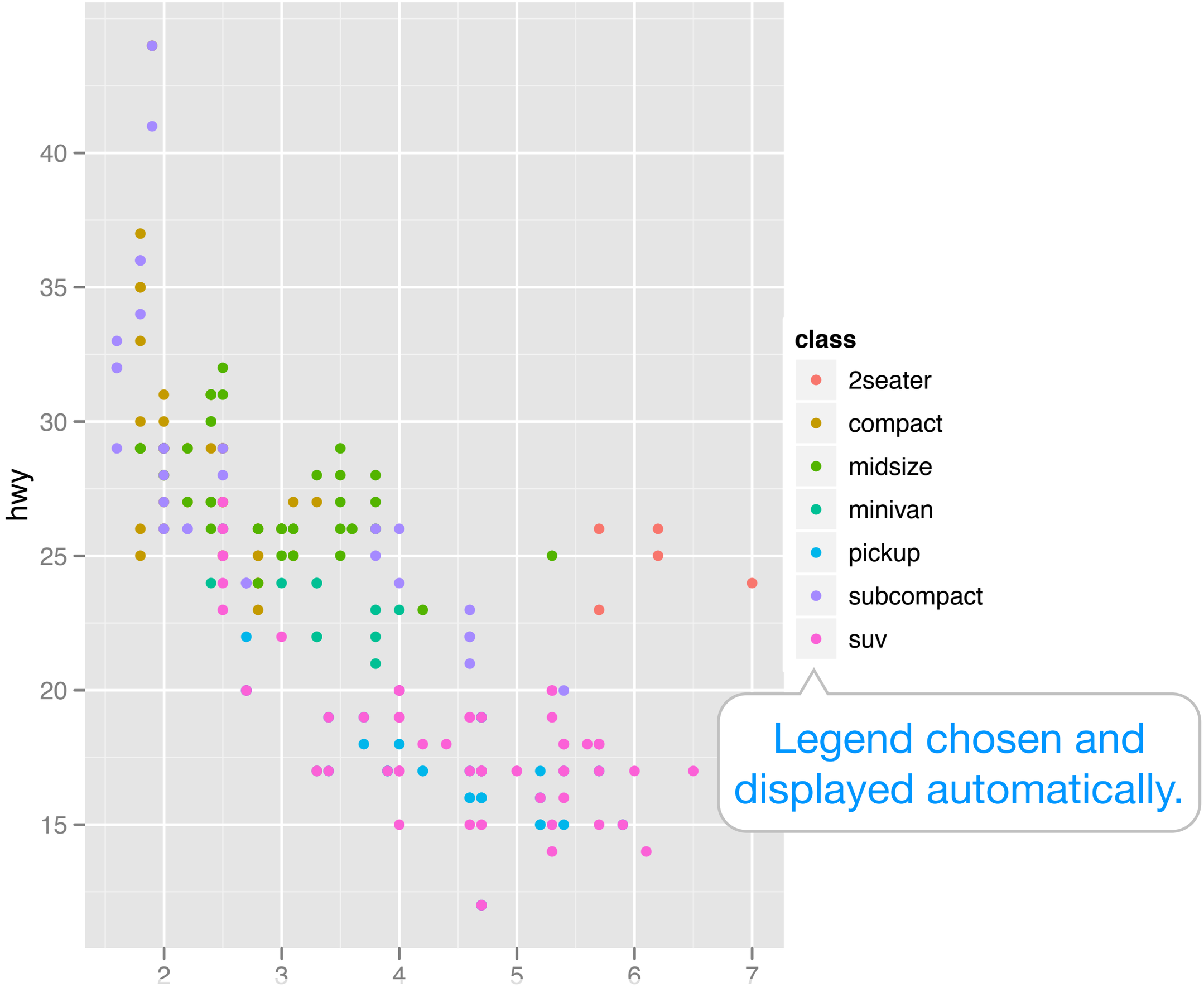

```
qplot(displ, hwy, data = mpg)
```

Additional variables

Can display additional variables with **aesthetics** (like shape, colour, size) or **faceting** (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	Colour gradient
Size	Discrete size steps	Linear mapping between radius and value
Shape	Different shape for each	Doesn’t work

Facetting & Geoms

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

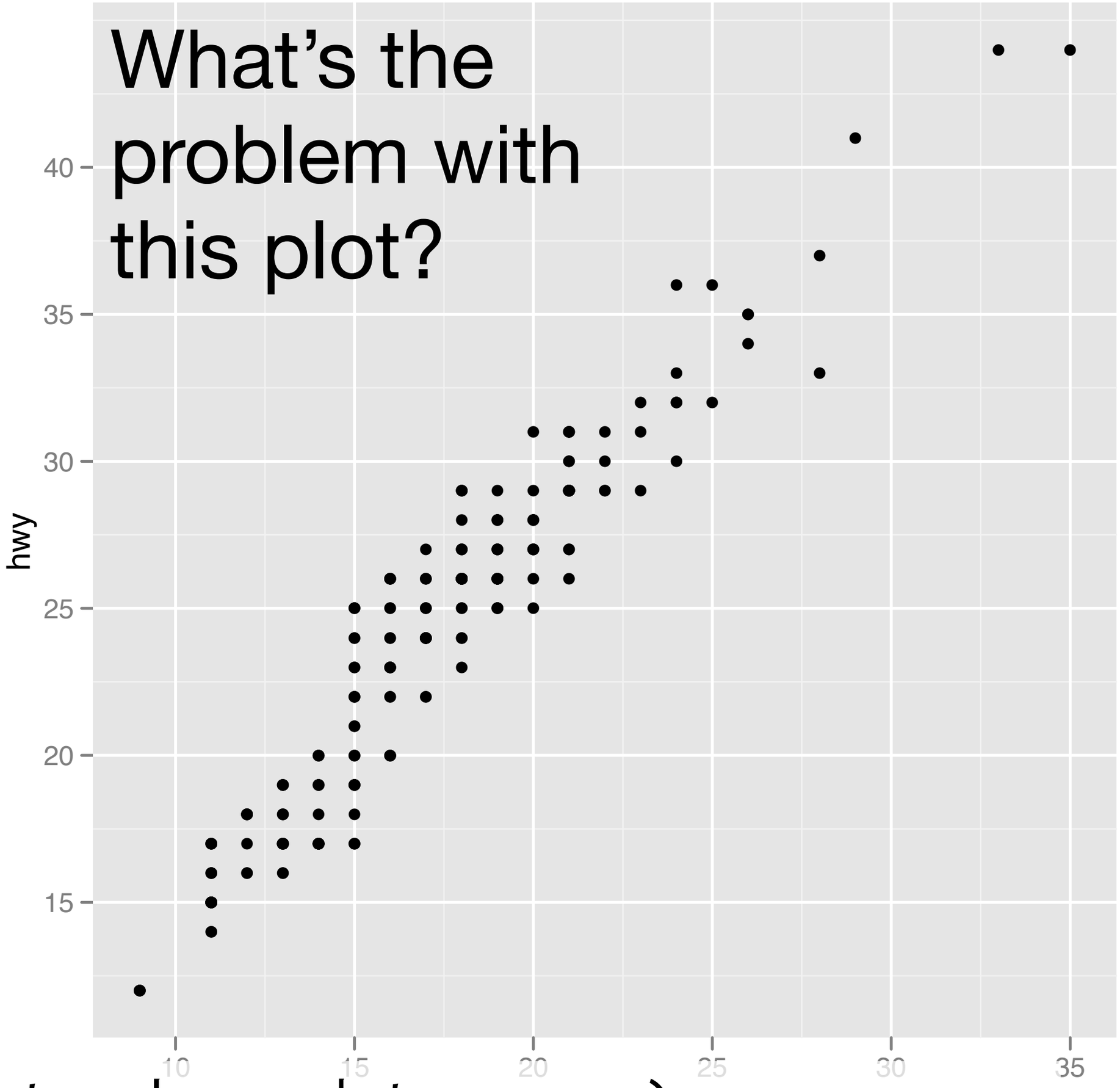
`facet_wrap()`: 1d ribbon wrapped into 2d

Aside: workflow

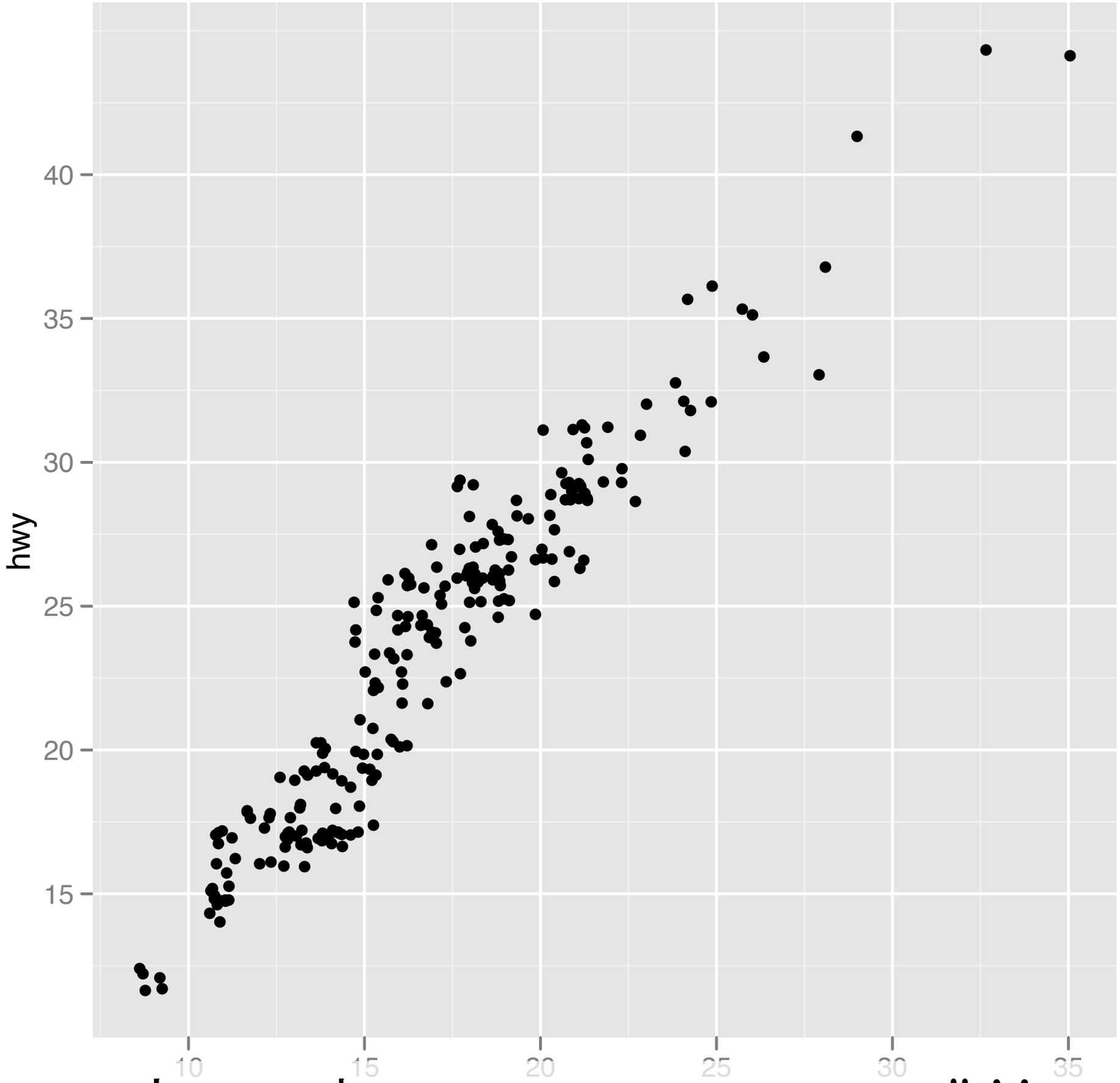
Keep a copy of the slides open so that you can copy and paste the code.

For complicated commands, write them in the editing area and then run.

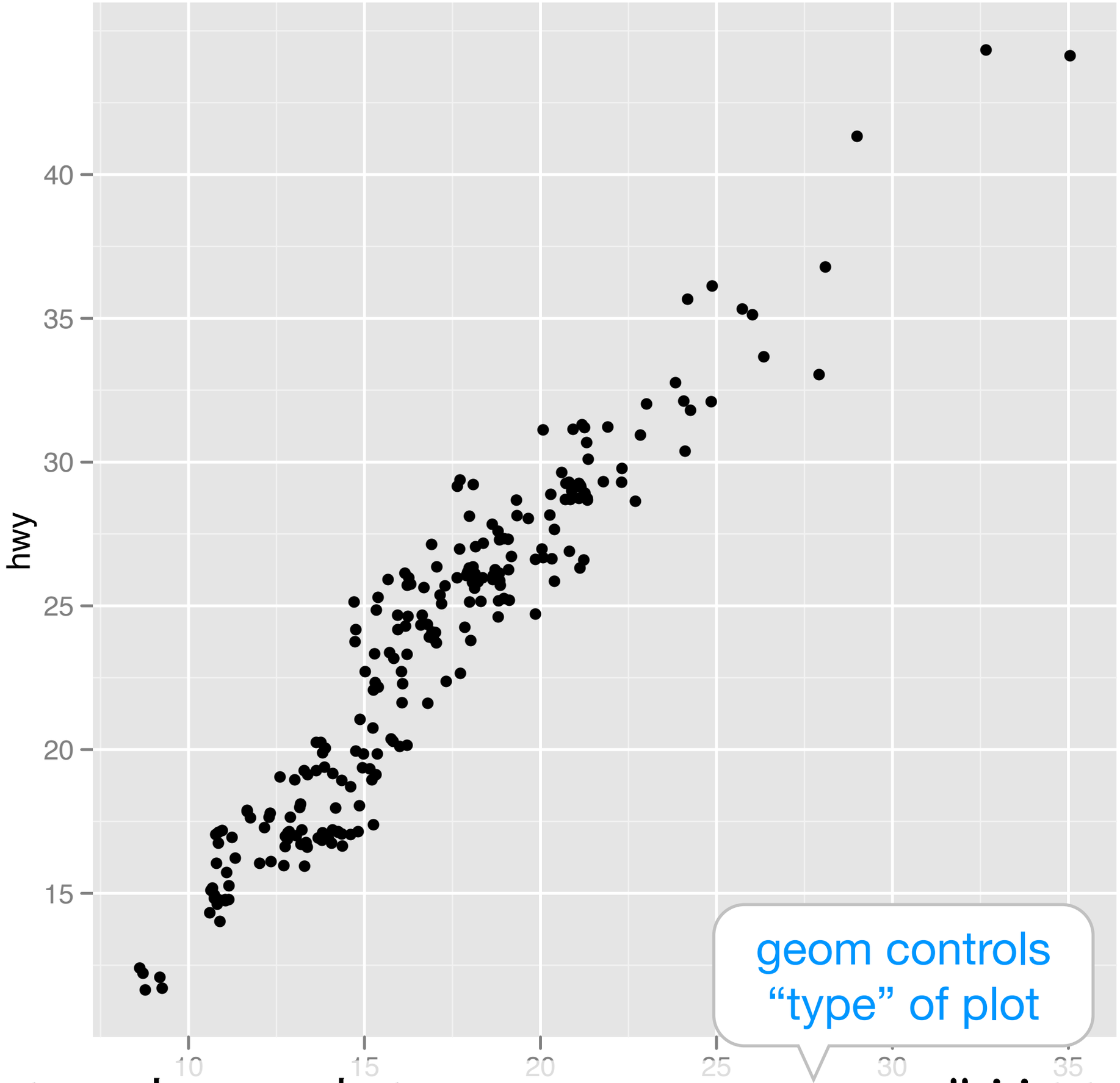
What's the problem with this plot?



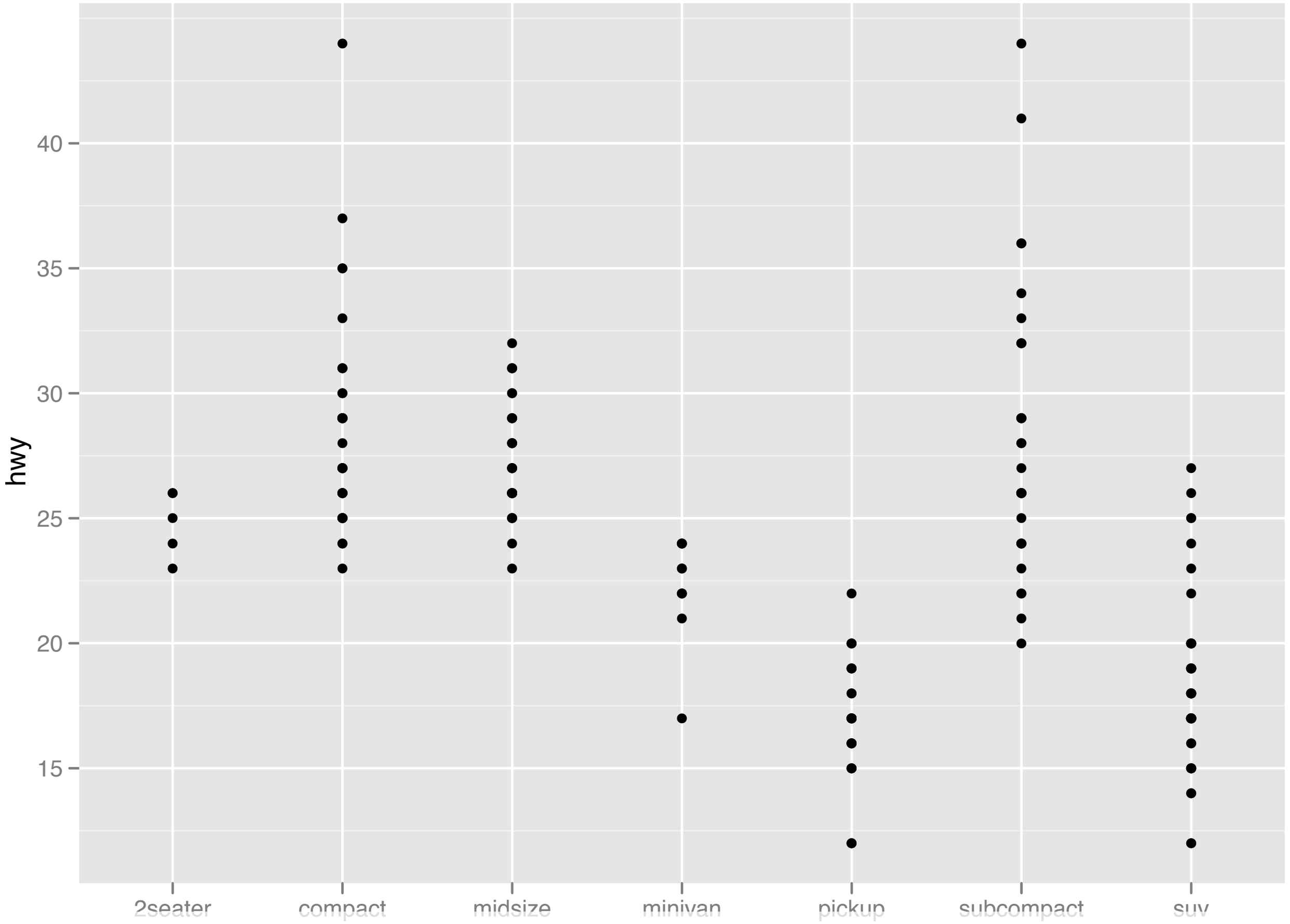

```
qplot(cty, hwy, data = mpg)
```




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



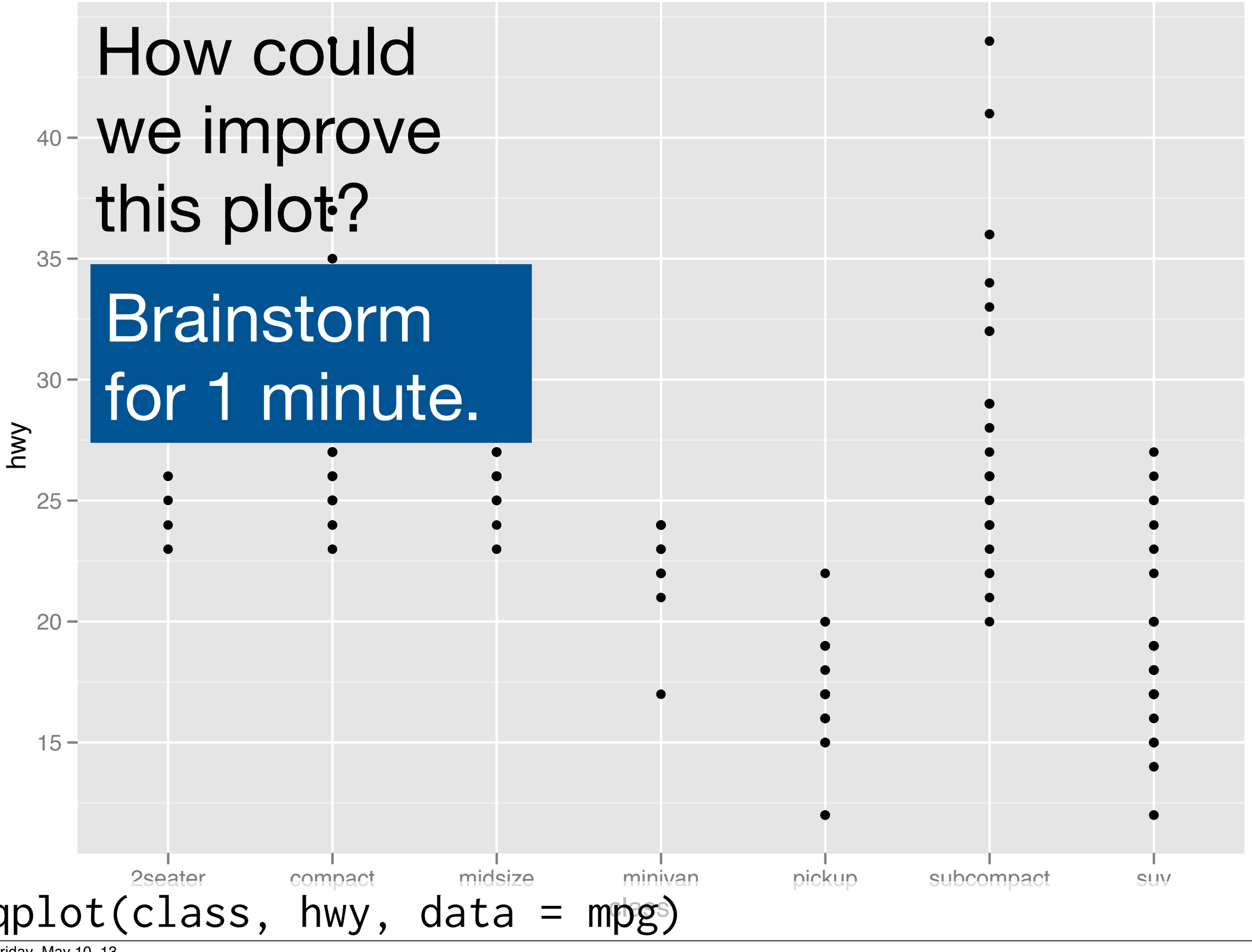

```
qplot(cty, hwy, data = mpg, geom = "jitter")
```



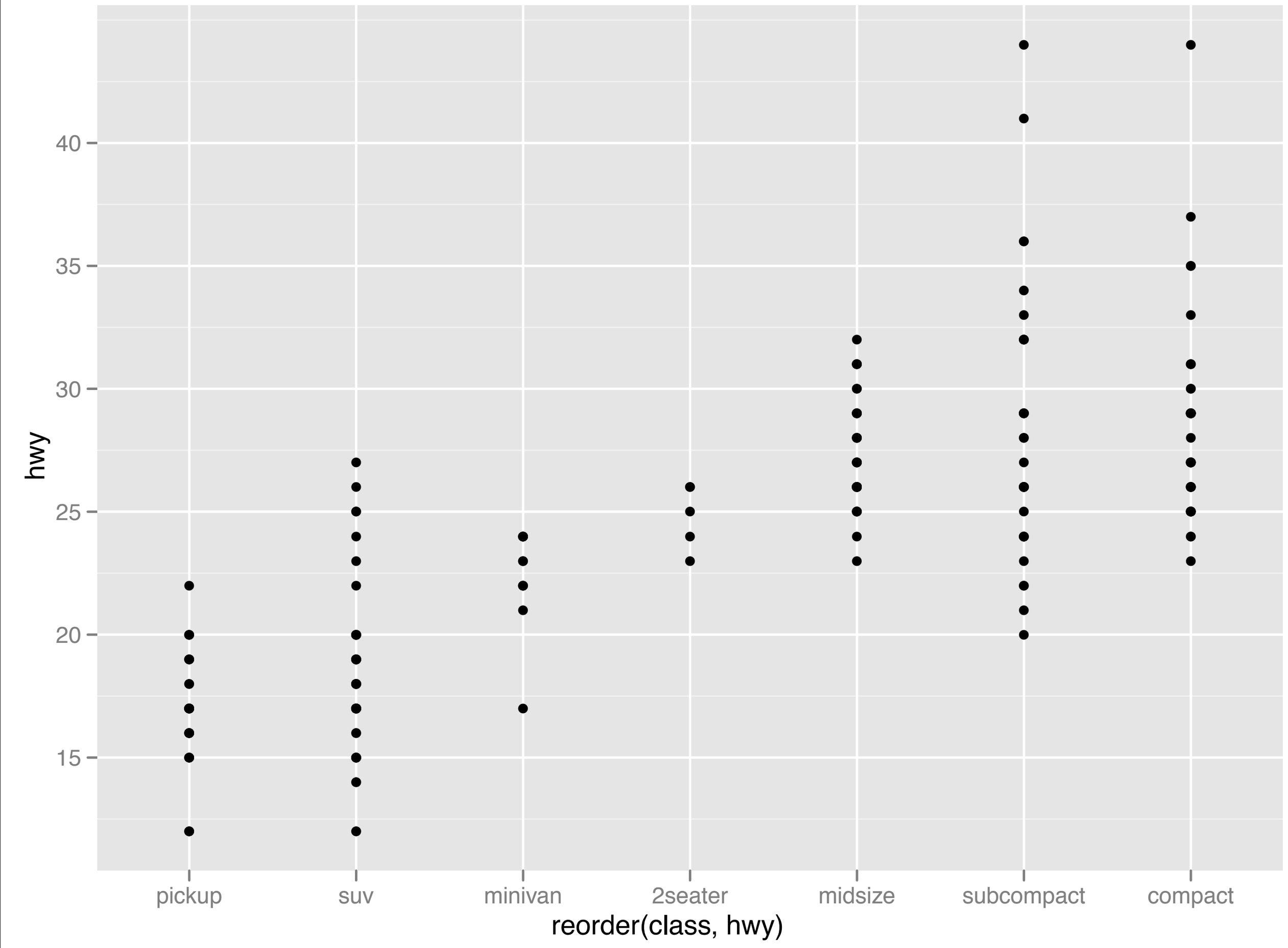

```
qplot(class, hwy, data = mpg)
```

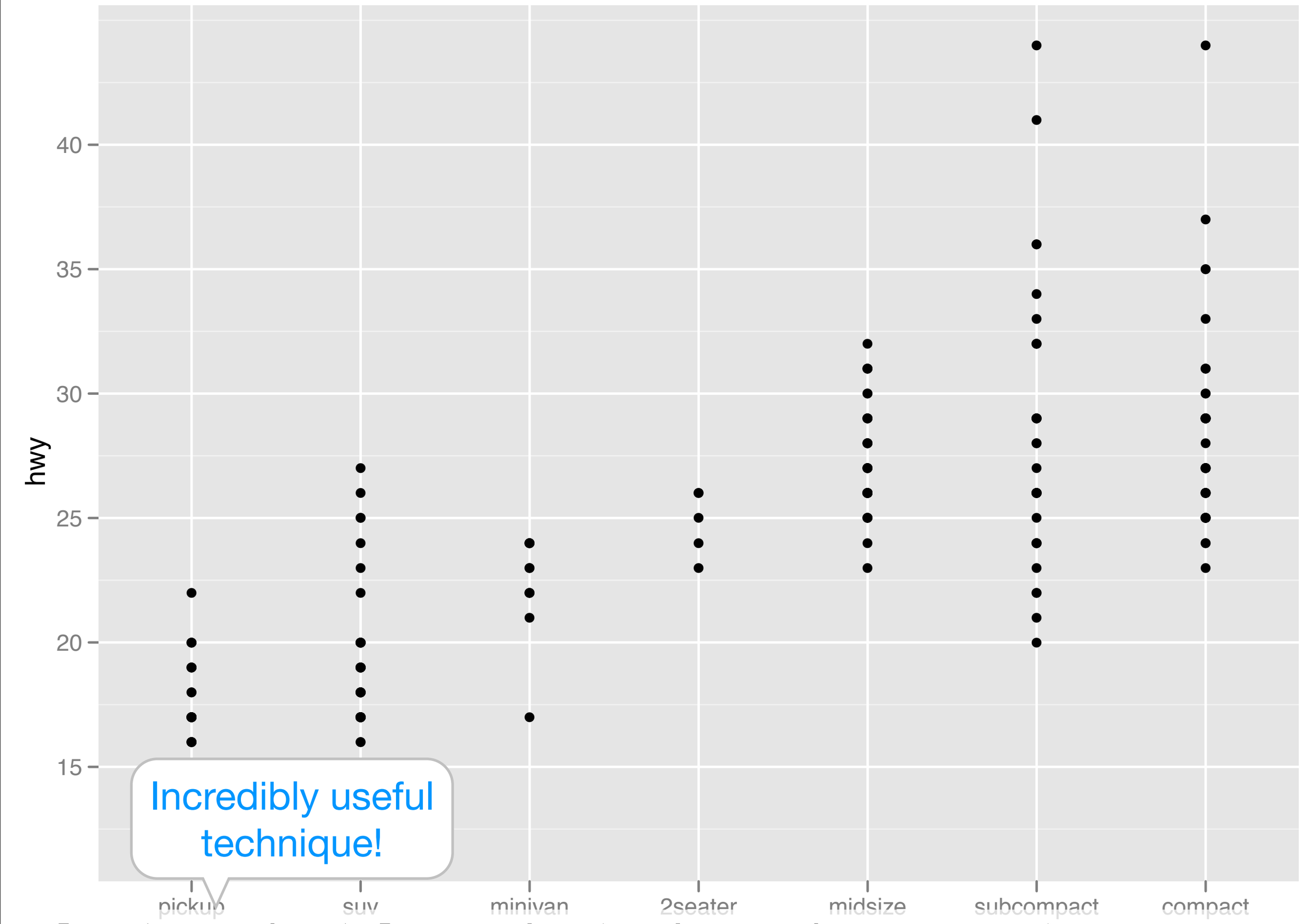
How could we improve this plot?

Brainstorm for 1 minute.

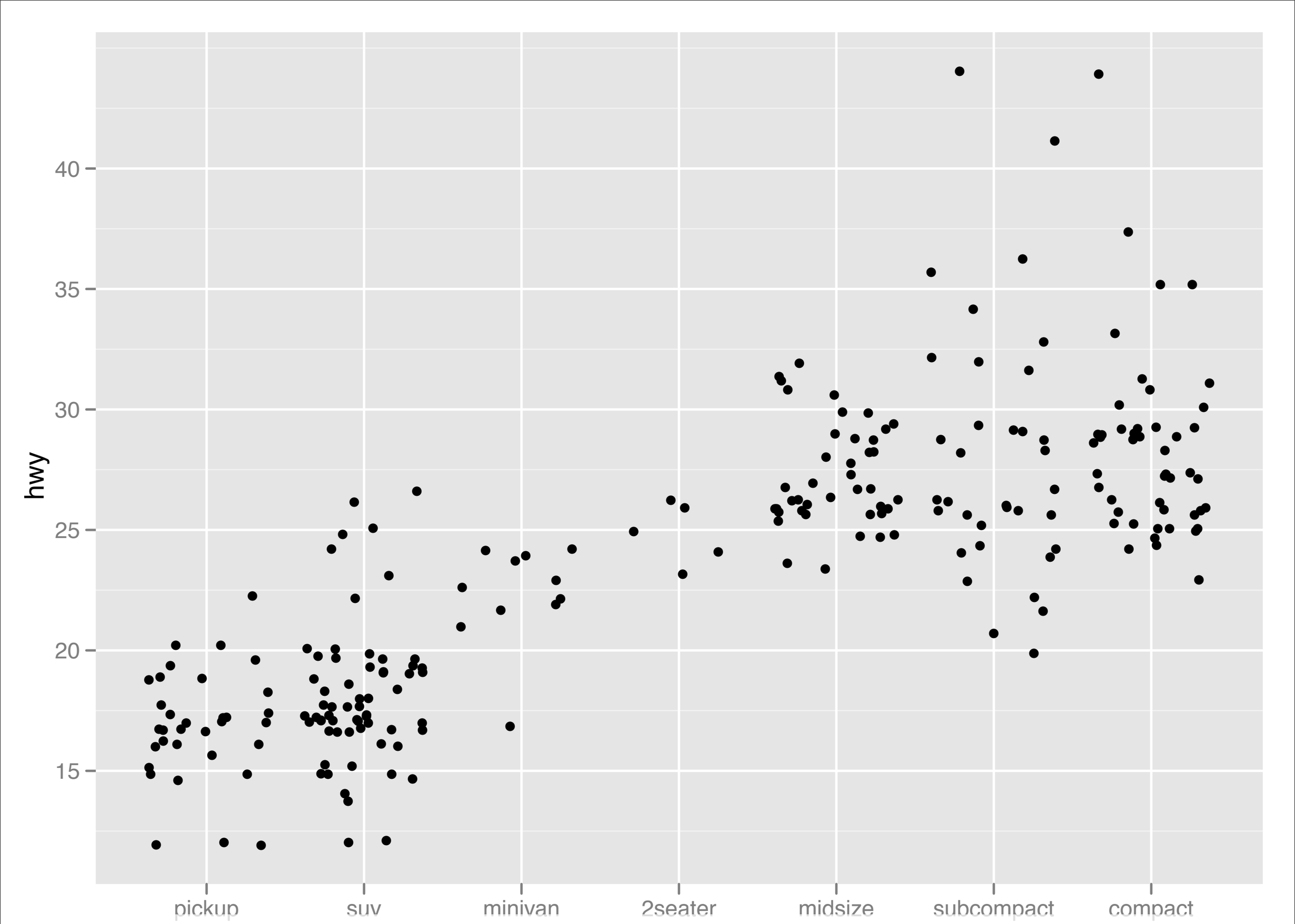



```
qplot(class, hwy, data = mpg)
```

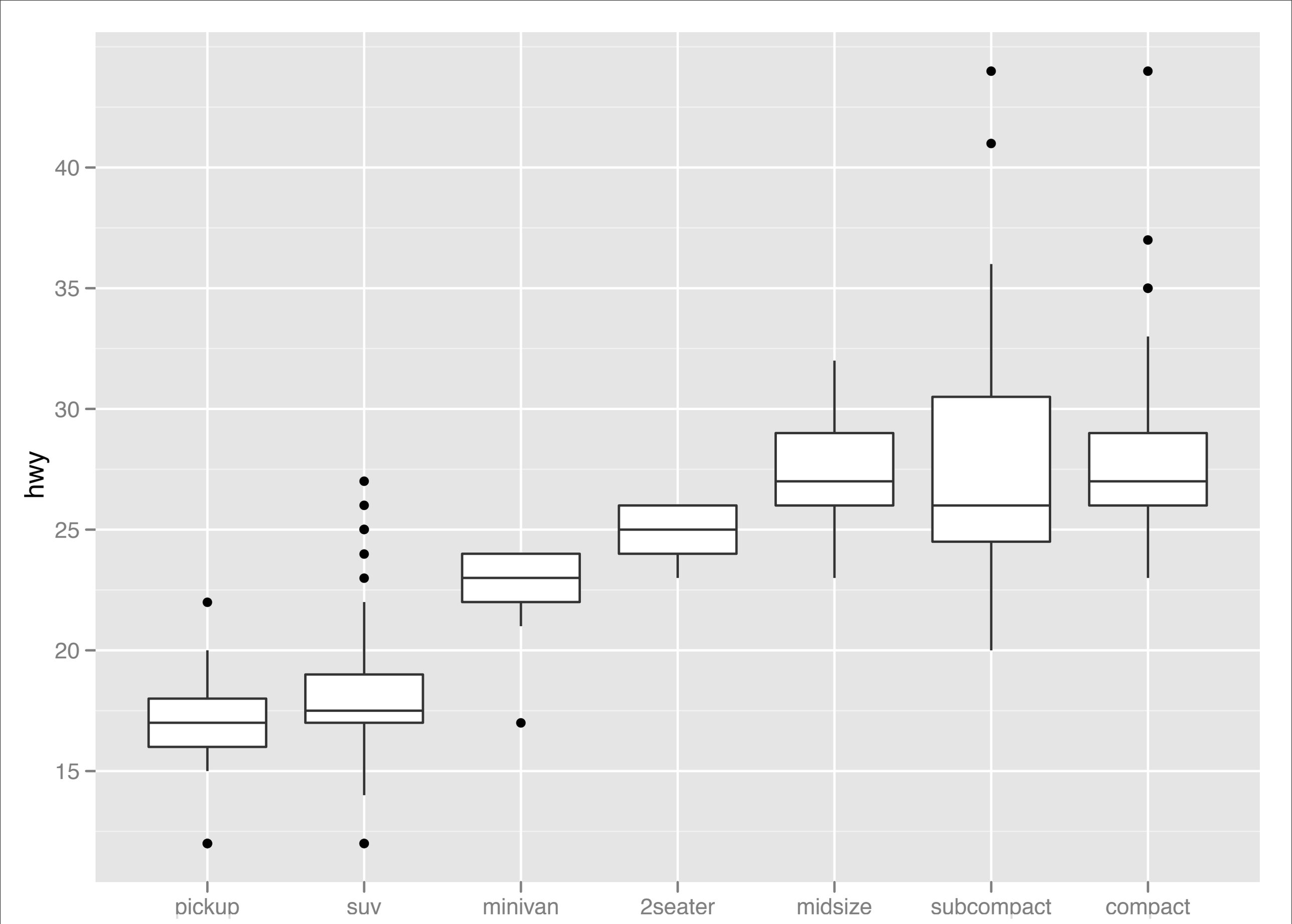

hwy
40
35
30
25
20
15
pickup
suv
minivan
2seater
midsize
subcompact
compact
reorder(class, hwy)



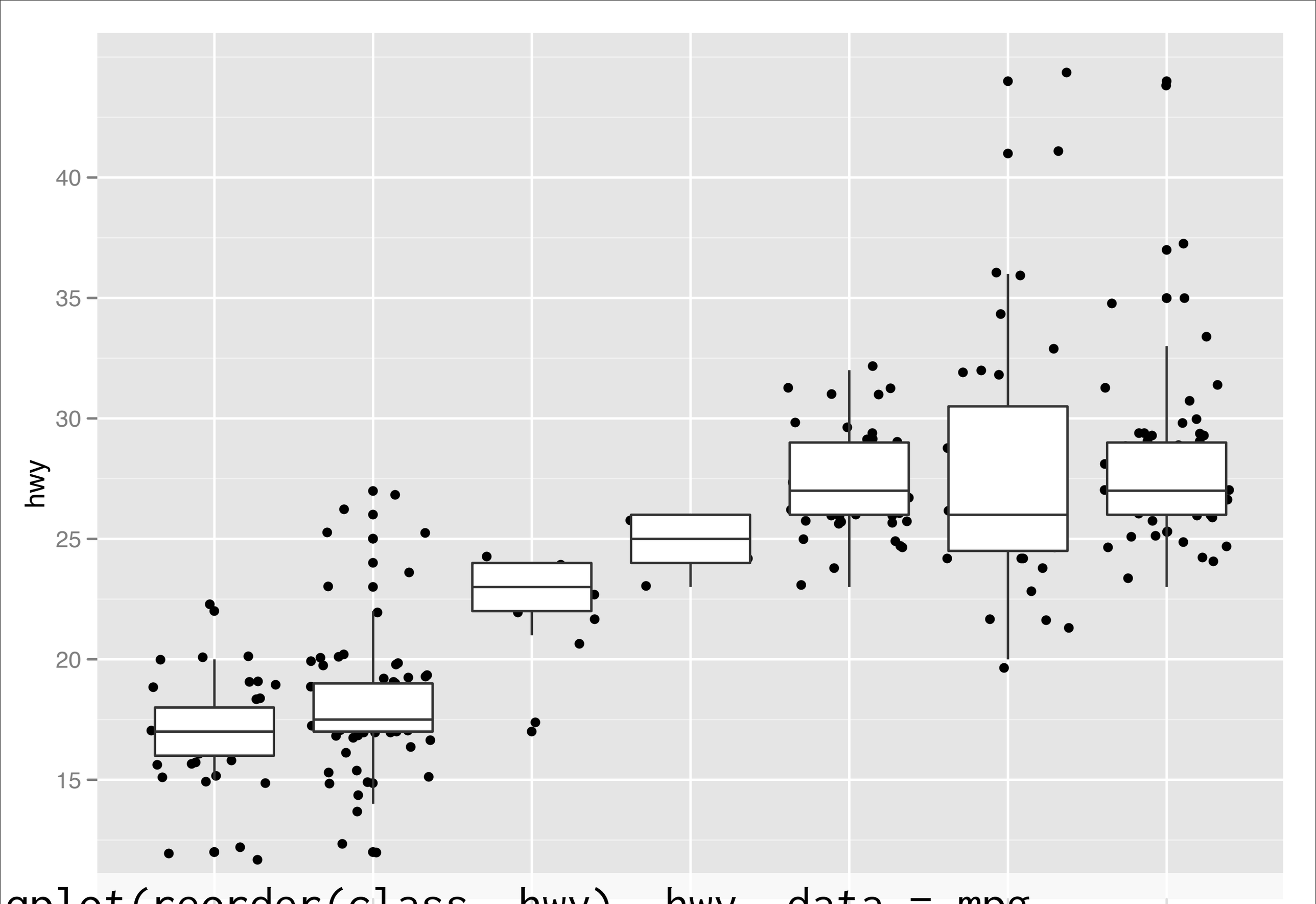

```
qplot(reorder(class, hwy), hwy, data = mpg)
```




```
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?

Aside: coding strategy

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

1. Copy from the history panel.
2. Build up the important bits as you go. (recommended)

Diamonds

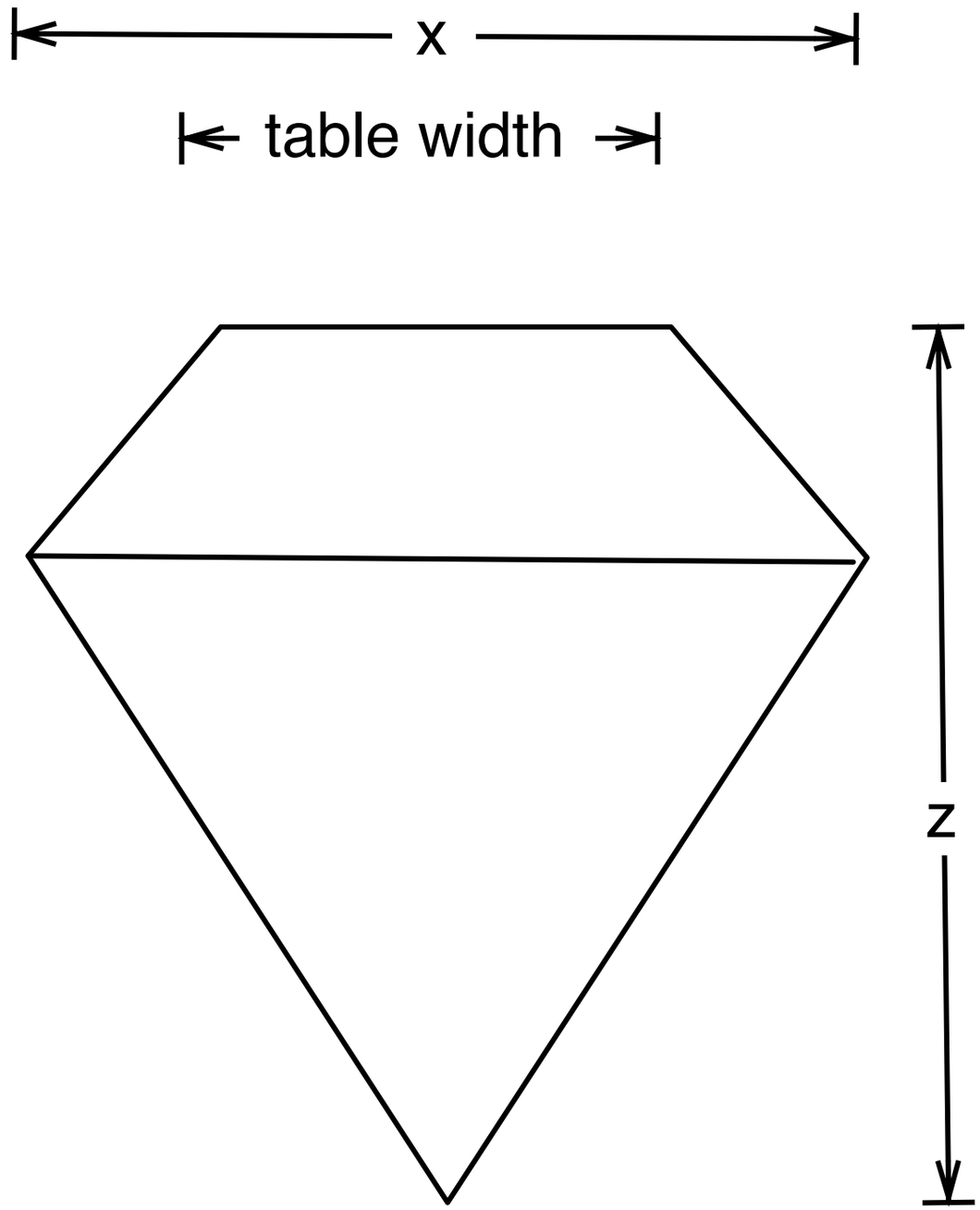
Diamonds data

~**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

Recall

Write down five ways to inspect the diamonds dataset.

You have one minute!

Histogram & bar charts

Histograms and barcharts

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

Categorical variable → bar chart

Continuous variable → histogram

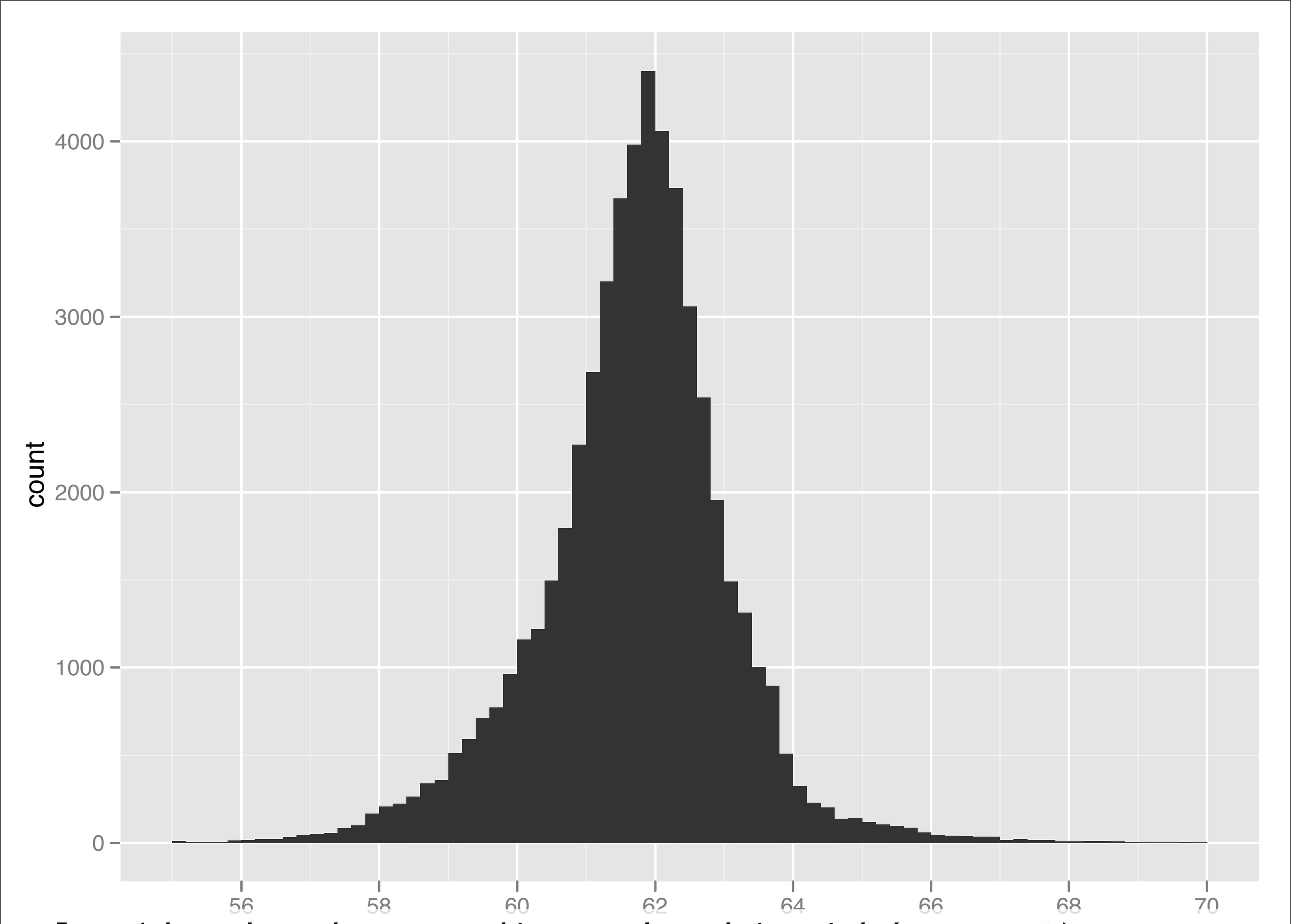
```
# With only one variable, qplot guesses that
# you want a bar chart or histogram
qplot(cut, data = diamonds)

qplot(carat, data = diamonds)

# Change binwidth:
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)
```

Always experiment with the bin width!

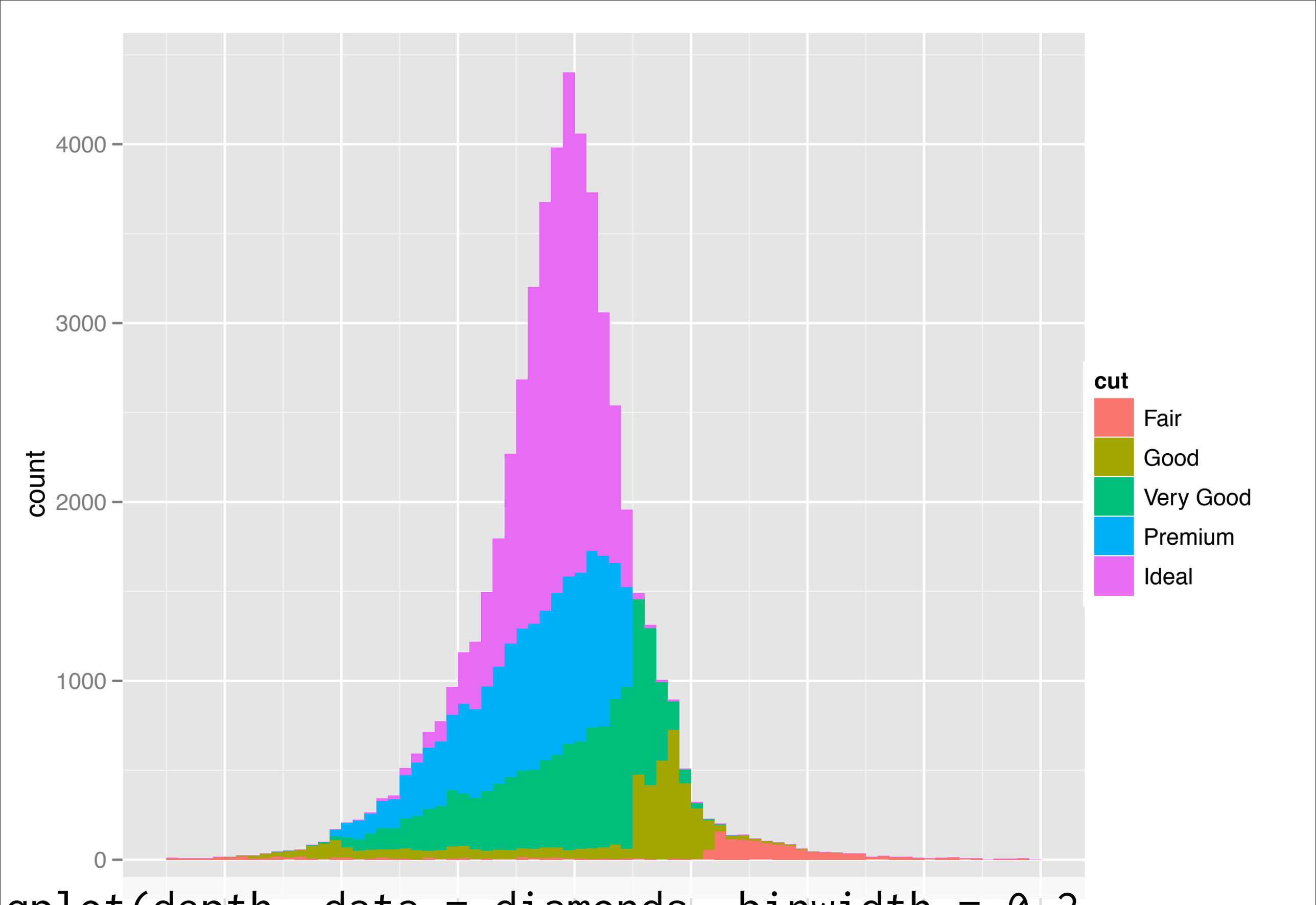
Additional variables

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

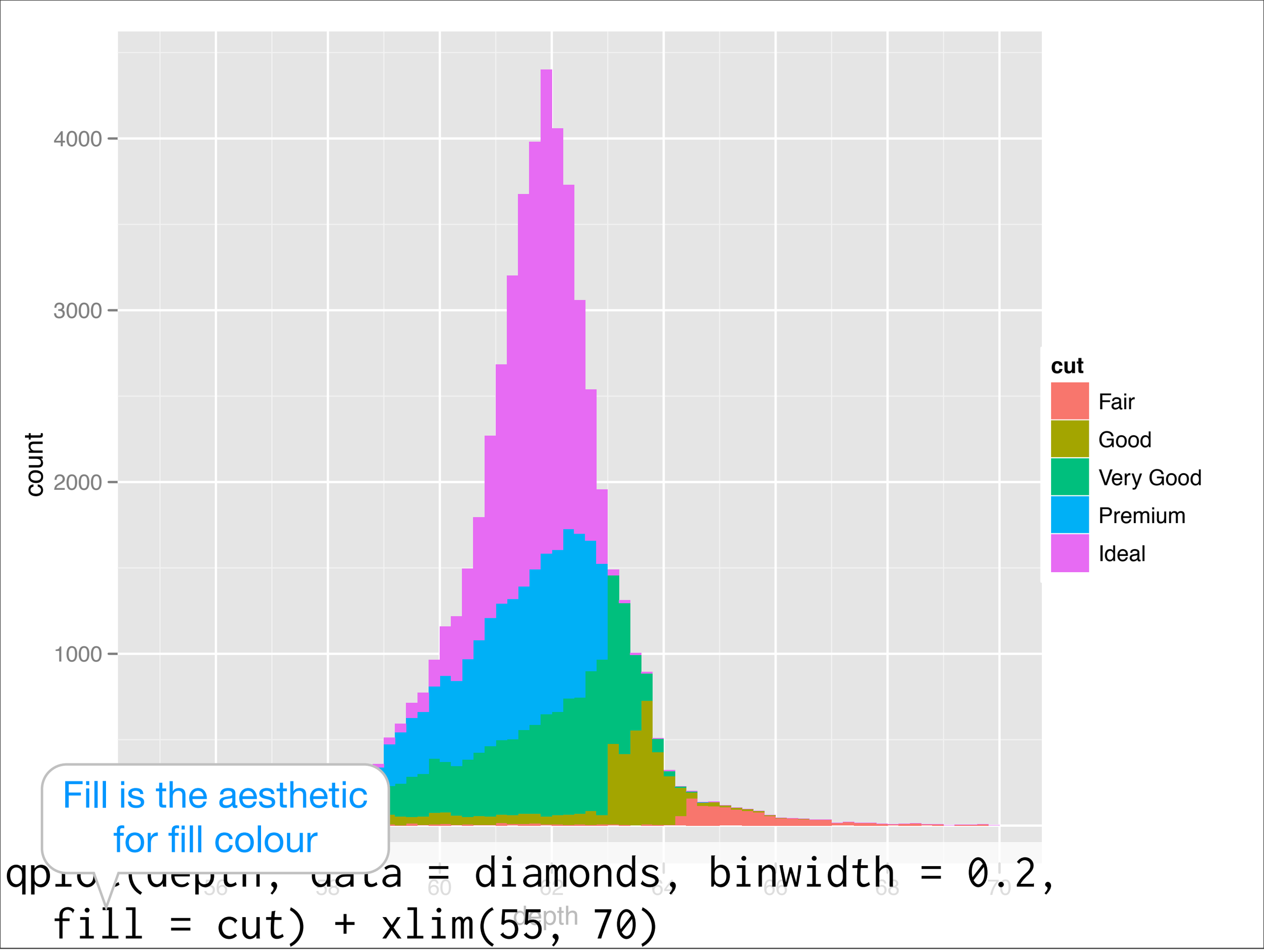
The following examples show the difference, when investigation the relationship between cut and depth.



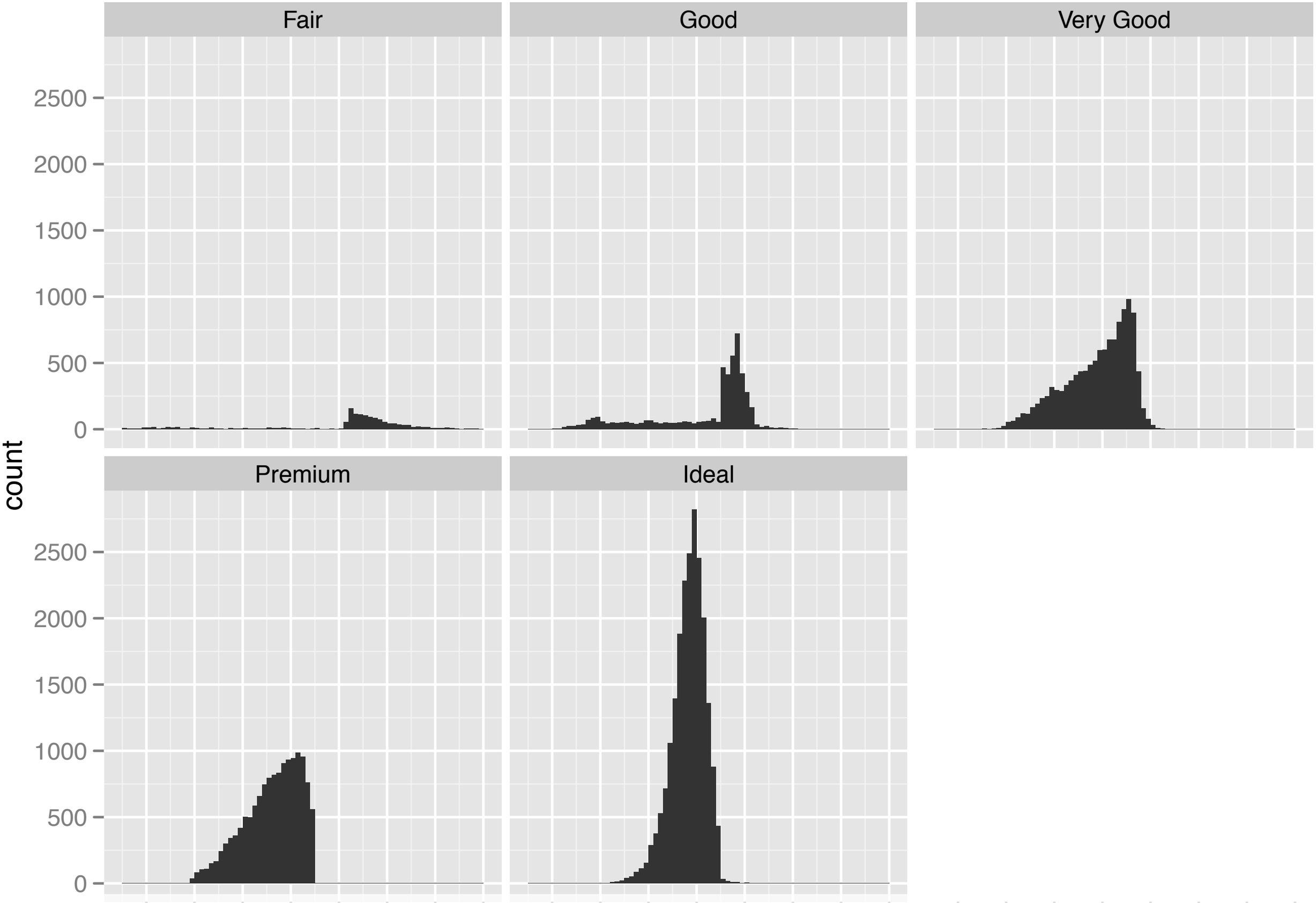

```
qplot(depth, data = diamonds, binwidth = 0.2)
```




```
qplot(depth, data = diamonds, binwidth = 0.2,
  fill = cut) + xlim(55, 70)
```




```
qplot(depth, data = diamonds, binwidth = 0.2,
  fill = cut) + xlim(55, 70)
```

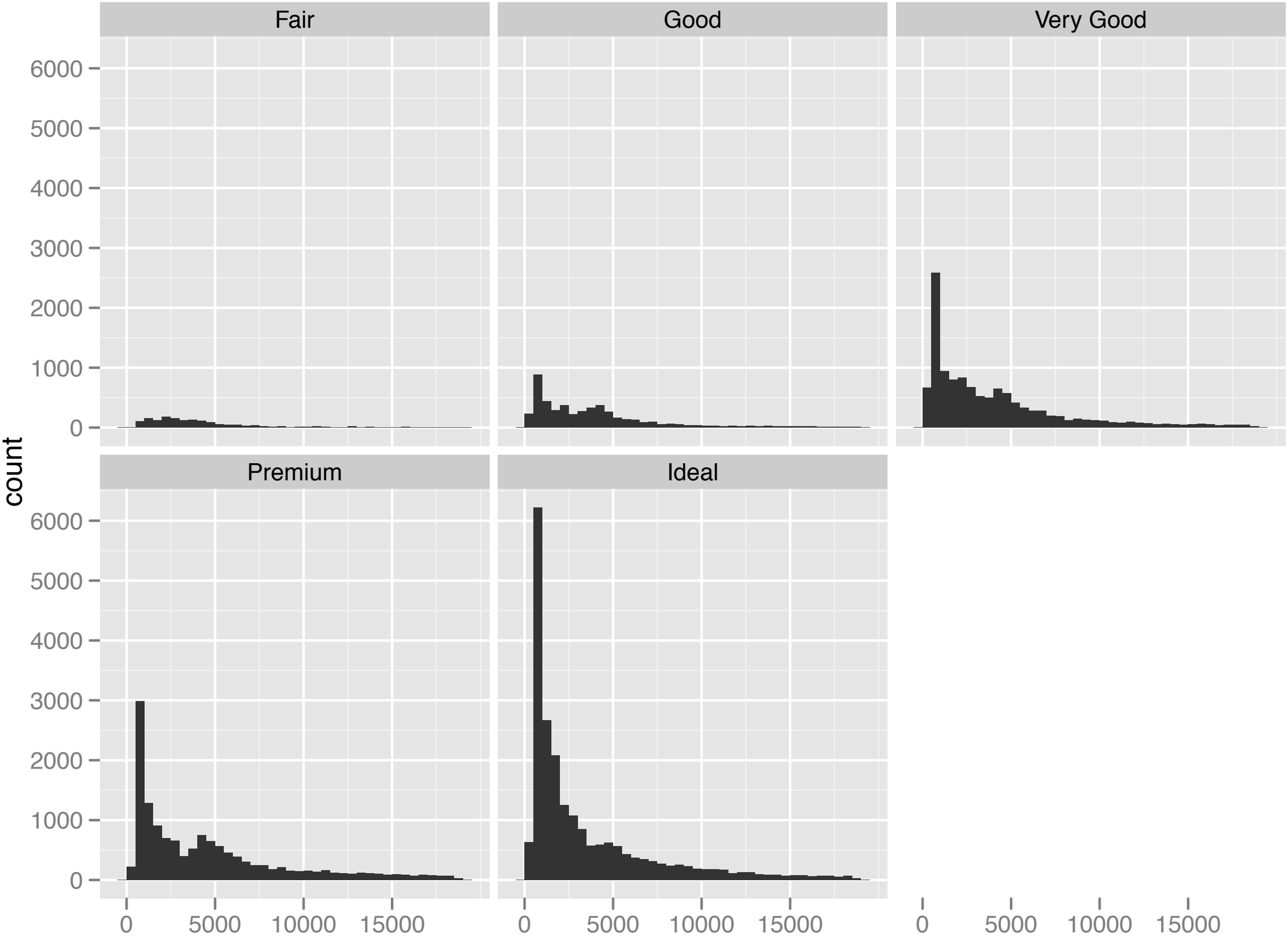



```
qplot(depth, data = diamonds, binwidth = 0.2) +
  xlim(55, 70) + facet_wrap(~cut)
```

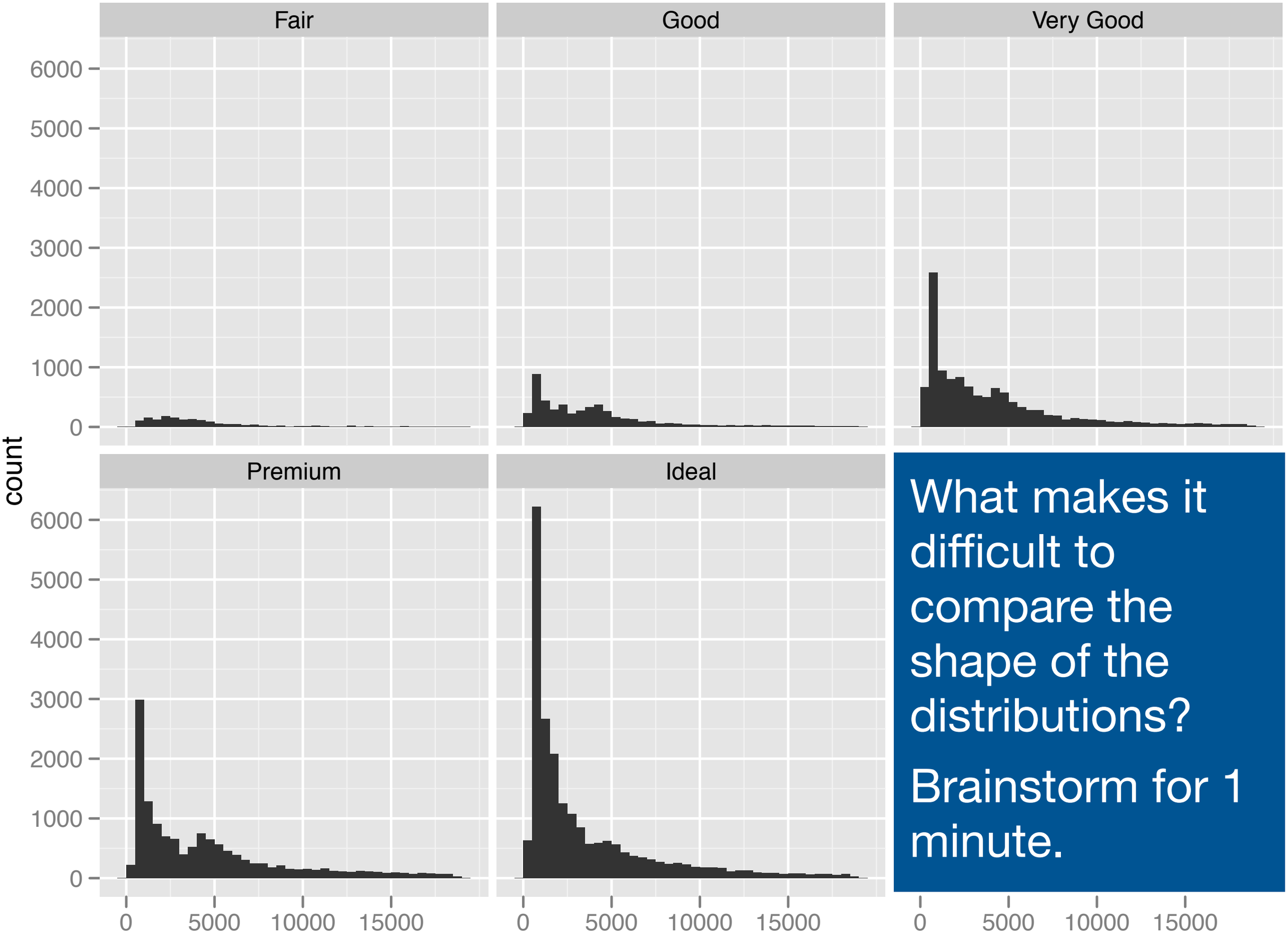
Your turn

Explore the distribution of price. What is a good binwidth to use? (**Hint**: How many bins will a binwidth of 1 give you?) Practice zooming in on regions of interest.

How does price vary with colour, cut, or clarity?




```
qplot(price, data = diamonds, binwidth = 500) + facet_wrap(~ cut)
```



What makes it difficult to compare the shape of the distributions?

Brainstorm for 1 minute.

```
qplot(price, data = diamonds, binwidth = 500) + facet_wrap(~ cut)
```

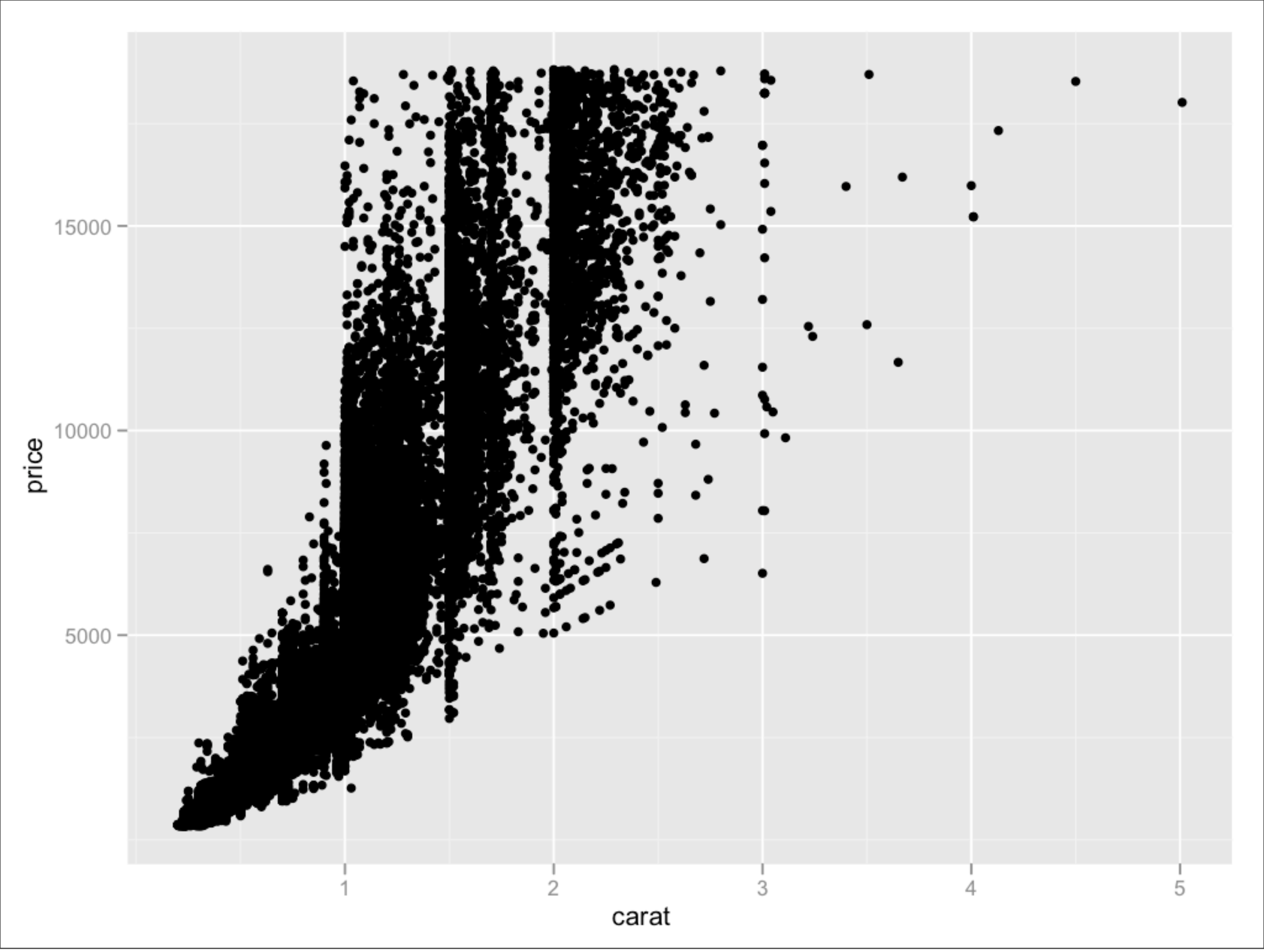
Problems

Each histogram far away from the others, but we know stacking is hard to read → *use another way of displaying densities*

Varying relative abundance makes comparisons difficult → *rescale to ensure constant area*

```
# Large distances make comparisons hard
qplot(price, data = diamonds, binwidth = 500) +
  facet_wrap(~ cut)

# Stacked heights hard to compare
qplot(price, data = diamonds, binwidth = 500, fill = cut)

# Much better - but still have differing relative
abundance
qplot(price, data = diamonds, binwidth = 500,
  geom = "freqpoly", colour = cut)

# Instead of displaying count on y-axis, display density
# .. indicates that variable isn't in original data
qplot(price, ..density.., data = diamonds, binwidth = 500,
  geom = "freqpoly", colour = cut)
```

Visualizing Big(ger) Data


15000
10000
5000
price
1
2
3
4
5
carat

Your turn

Take two minutes to brainstorm possible solutions to the overplotting problem.

Idea	ggplot
Small points	`size = I(0.25)`
Transparency	`alpha = I(1/50)`
Jittering	`geom = "jitter"`
Smooth curve	`geom = "smooth"`
2d bins	`geom = "bin2d"` or `geom = "hex"`
Density contours	`geom = "density2d"`

```
# To set aesthetics to a particular value, you
# need to wrap that value in I() (otherwise qplot will
# try to map your input to the data set)

qplot(carat, price, data = diamonds, colour = "blue")
qplot(carat, price, data = diamonds, colour = I("blue"))
```

```
# Practical application: varying alpha
qplot(carat, price, data = diamonds, alpha = I(1/10))
qplot(carat, price, data = diamonds, alpha = I(1/50))
qplot(carat, price, data = diamonds, alpha = I(1/100))
qplot(carat, price, data = diamonds, alpha = I(1/250))
```

Your turn

Create a plot that reveals the relationship between carat and price without overplotting.

Which method do you prefer and why?

```
# There are two ways to add additional geoms
# 1) A vector of geom names:
qplot(price, carat, data = diamonds,
   geom = c("point", "smooth"))

# 2) Add on extra geoms
qplot(price, carat, data = diamonds) +
   geom_smooth()

# This is how you get help about a specific geom:
?geom_smooth
# or go to http://docs.ggplot2.org/current/
```

Your turn

Explore the relationship between carat, price and clarity, using these techniques.

(i.e. make this plot more informative:

```
qplot(carat, price, data = diamonds, colour = clarity)
```

Which did you find most useful?

```
qplot(carat, price, data = diamonds,
  colour = cut)
qplot(log10(carat), log10(price),
  data = diamonds, colour = cut)

# install.packages("hexbin")
qplot(log10(carat), log10(price), data = diamonds,
  geom = "bin2d", bins = 50) + facet_wrap(~ cut)
qplot(log10(carat), log10(price), data = diamonds,
  colour = cut, geom = "smooth")

lm(log10(price) ~ log10(carat), data = diamonds)
qplot(log10(carat), log10(price),
  data = diamonds) + facet_wrap(~ cut) +
  geom_abline(colour = "red", intercept = 3.6, slope = 1.7)
```

Where to go from here

Index. ggplot2 0.9.2.1

docs.ggplot2.org/current/

ggplot2 0.9.2.1 Index

Help topics

Geoms

Geoms, short for geometric objects, describe the type of plot you will produce.

- `geom_abline`
  Line specified by slope and intercept.
- `geom_area`
  Area plot.
- `geom_bar`
  Bars, rectangles with bases on x-axis
- `geom_bin2d`
  Add heatmap of 2d bin counts.
- `geom_blank`
  Blank, draws nothing.
- `geom_boxplot`
  Box and whiskers plot.
- `geom_contour`
  Display contours of a 3d surface in 2d.
- `geom_crossbar`
  Hollow bar with middle indicated by horizontal line.
- `geom_density`
  Display a smooth density estimate.
- `geom_density2d`
  Contours from a 2d density estimate.
- `geom_dotplot`
  Dot plot


Dependencies

- **Depends**: stats, methods
- **Imports**: plyr, digest, grid, gtable, reshape2, scales, memoise, proto, MASS
- **Suggests**: quantreg, Hmisc, mapproj, maps, hexbin, maptools, multcomp, nlme, testthat
- **Extends**: sp

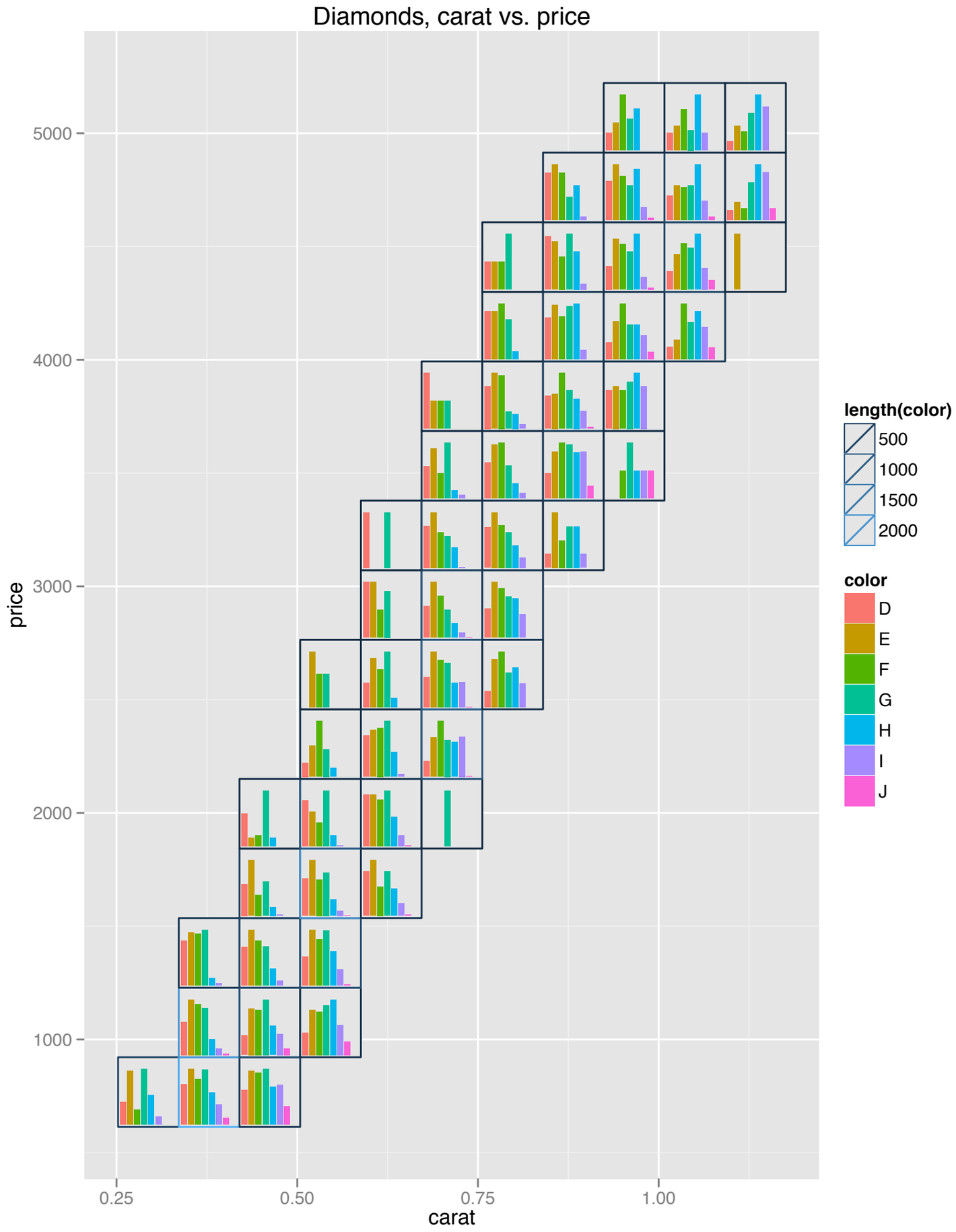
Learning ggplot2

R graphics cookbook
http://amzn.com/1449316956

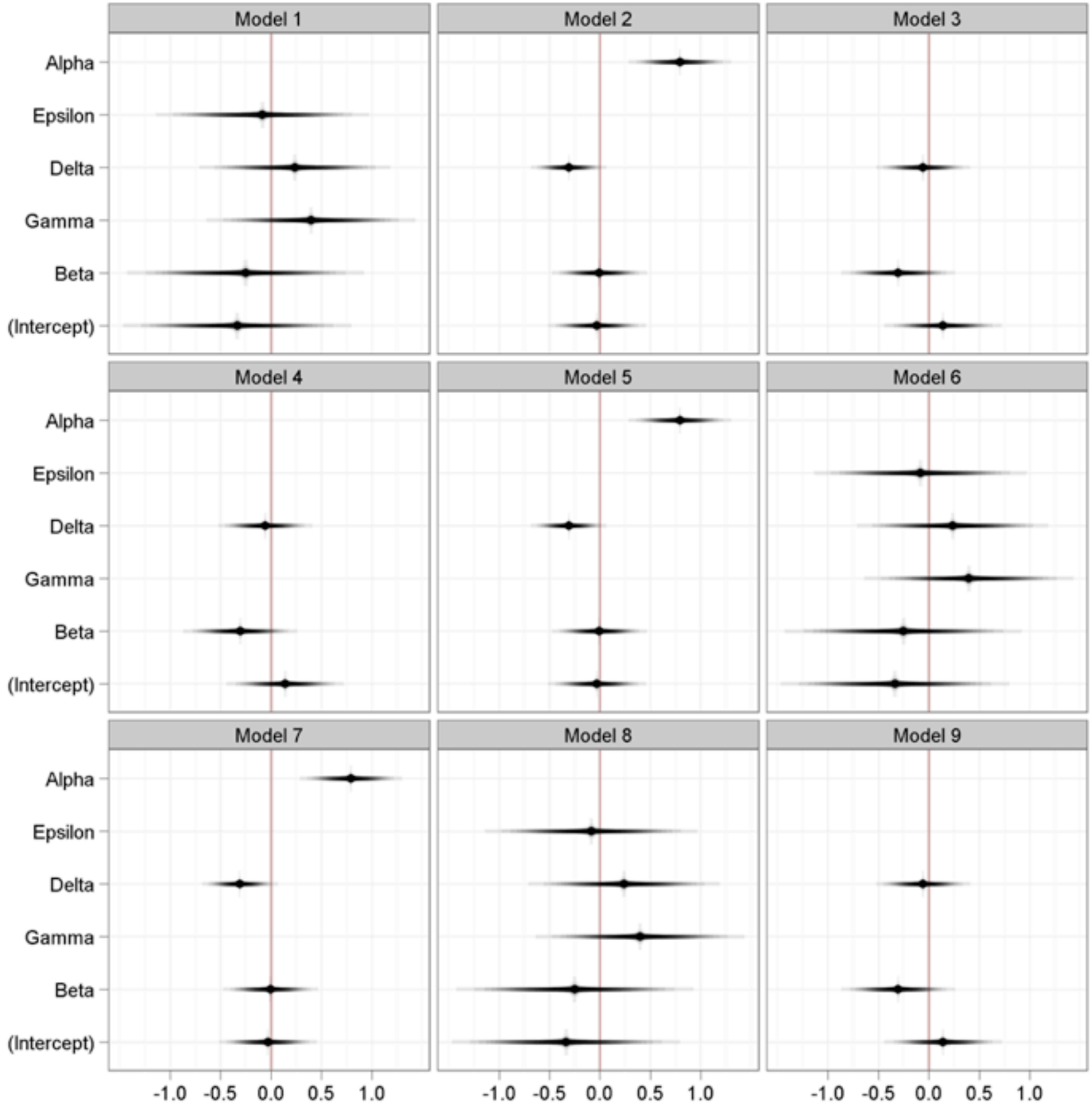
ggplot2 book
http://amzn.com/0387981403

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

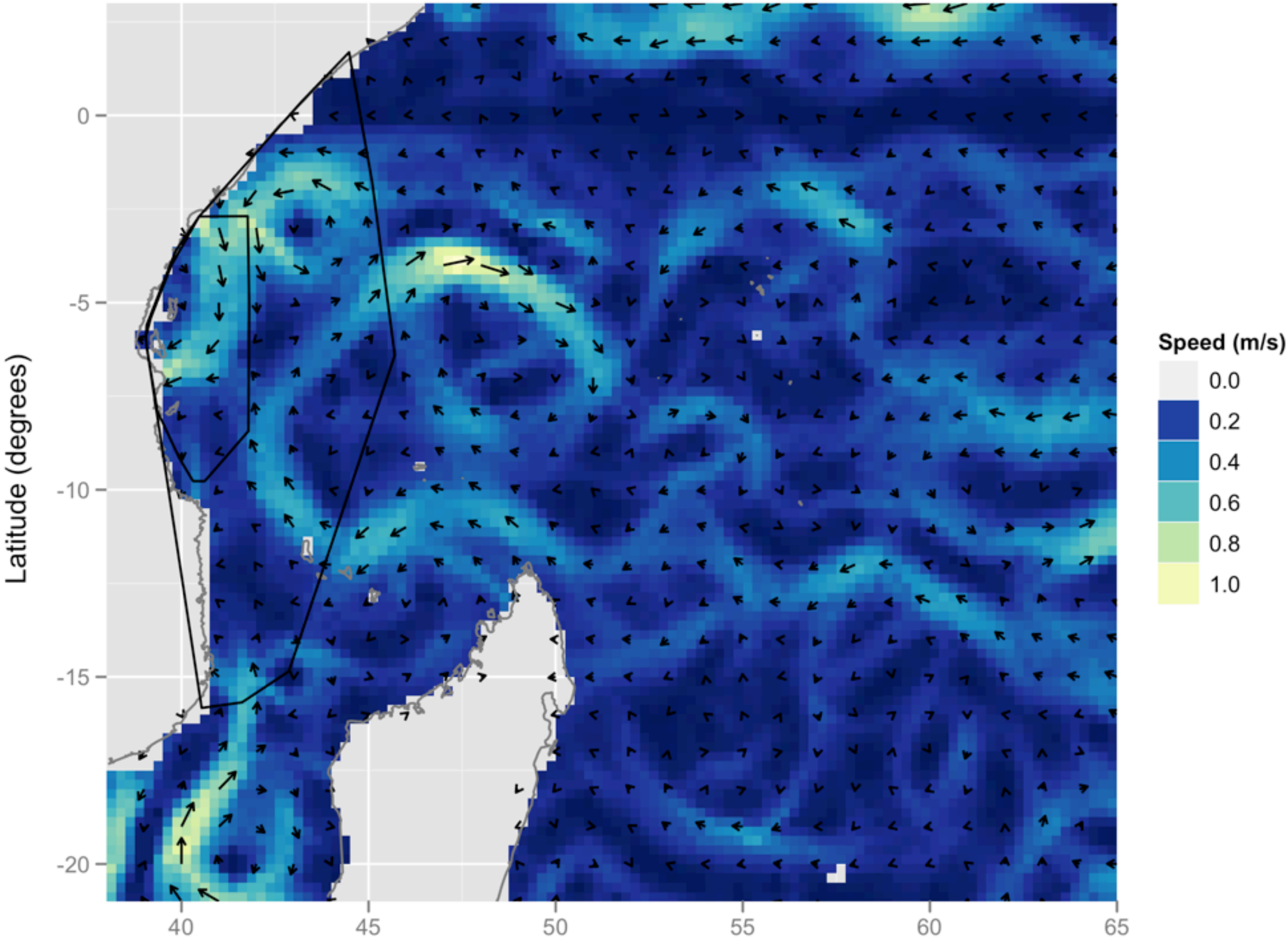
stackoverflow
http://stackoverflow.com/tags/ggplot2



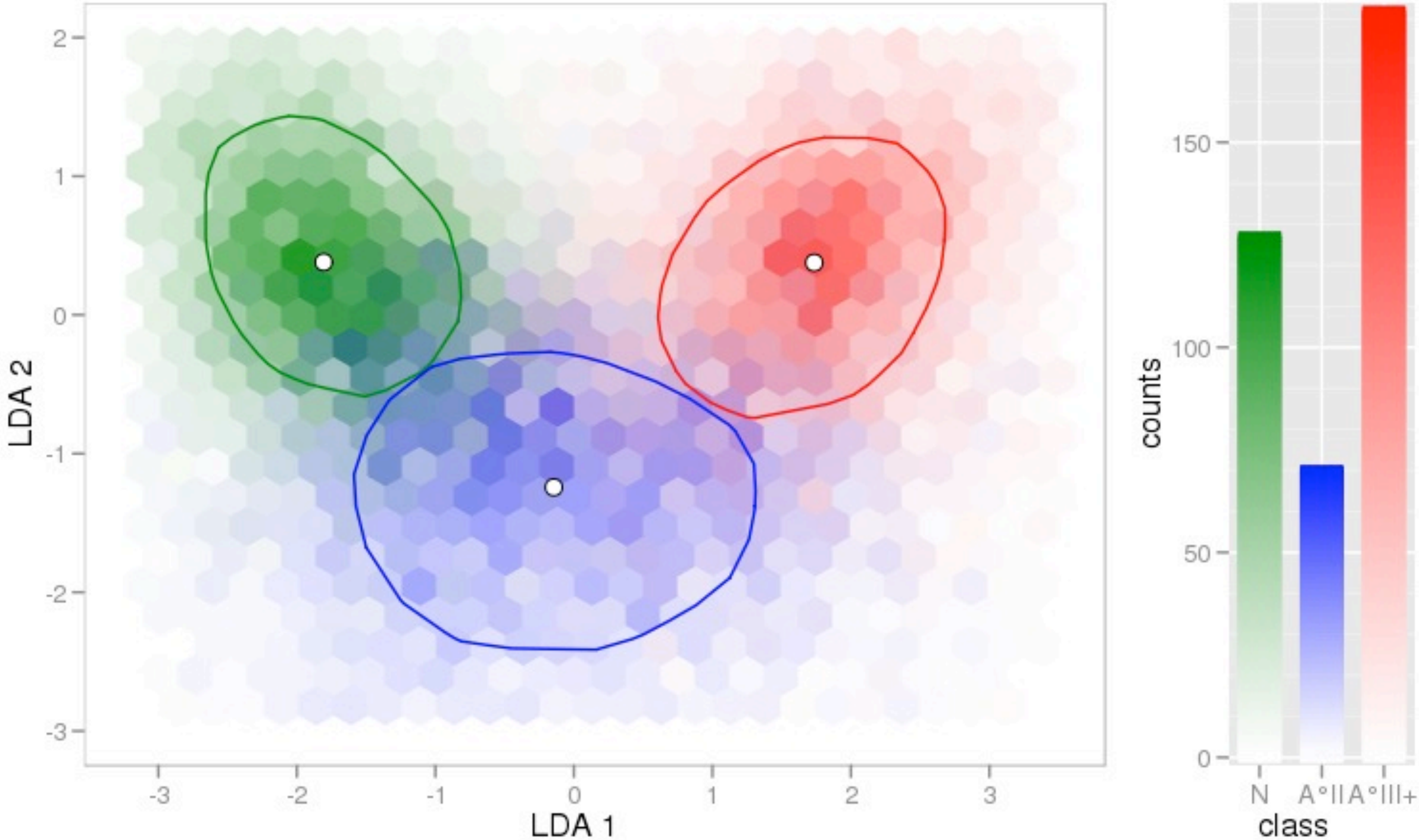
Garrett Grolemund, http://blog.revolutionanalytics.com/2012/09/visualize-complex-data-with-subplots.html



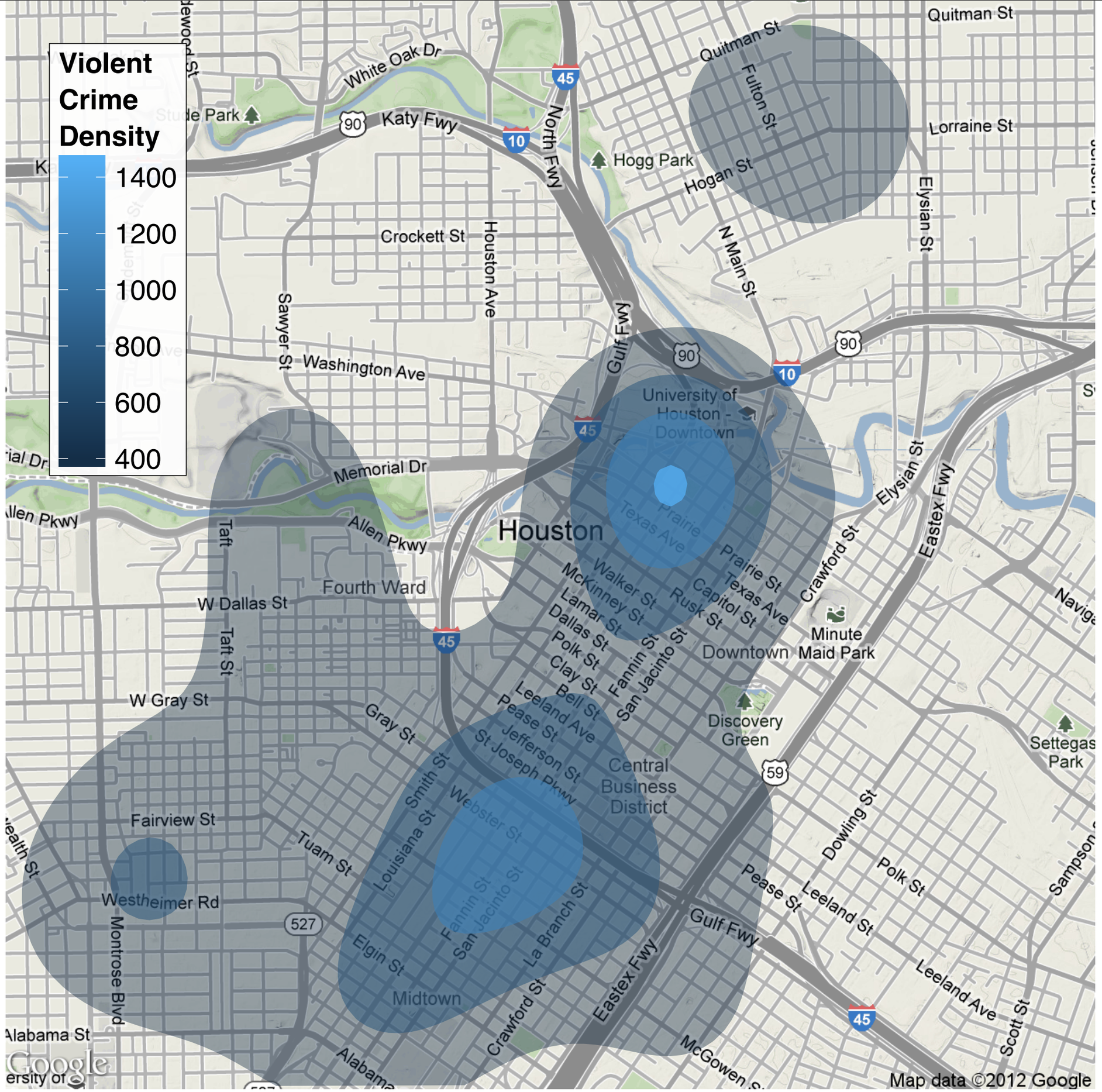
David B Sparks, http://bit.ly/hn54NW



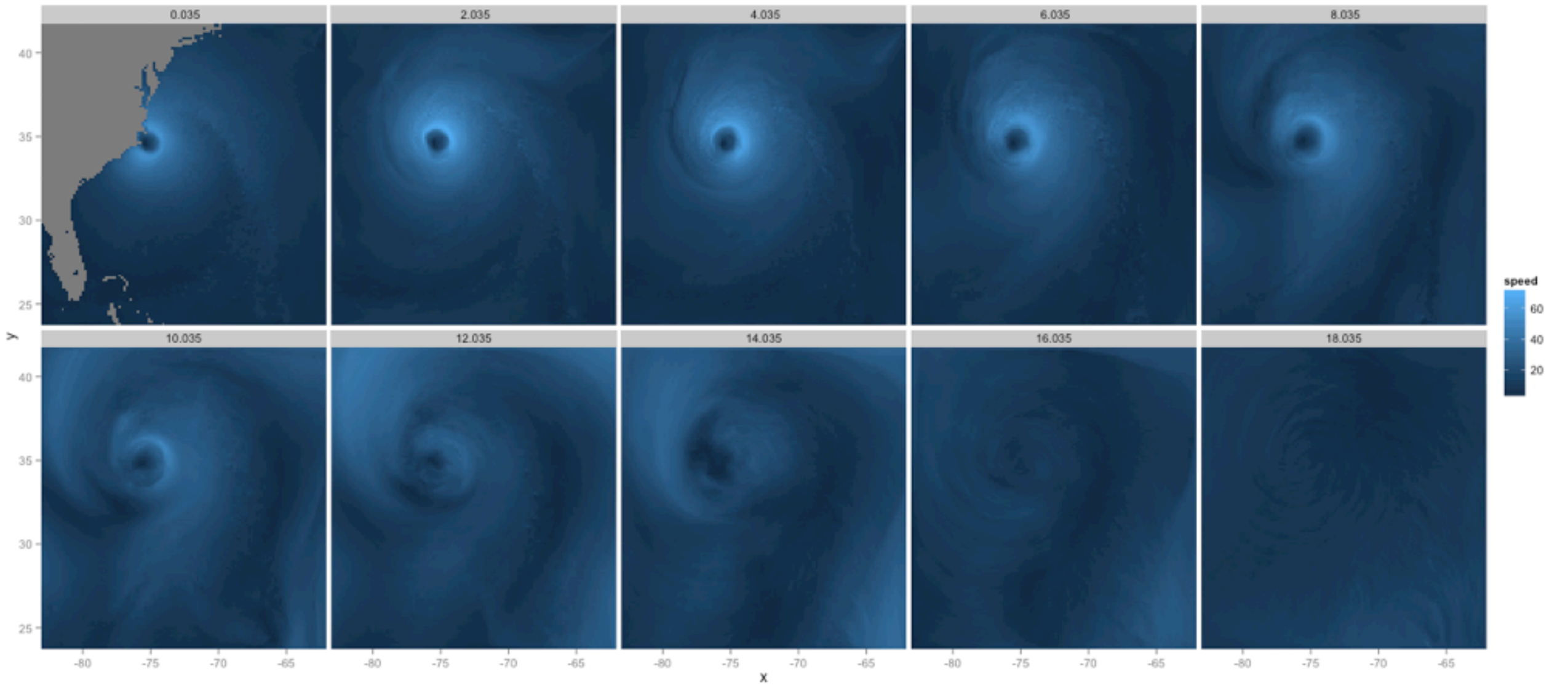
Charlotte Wickham, http://cwick.co.nz/



Claudia Beleites, http://bit.ly/yNqlpz



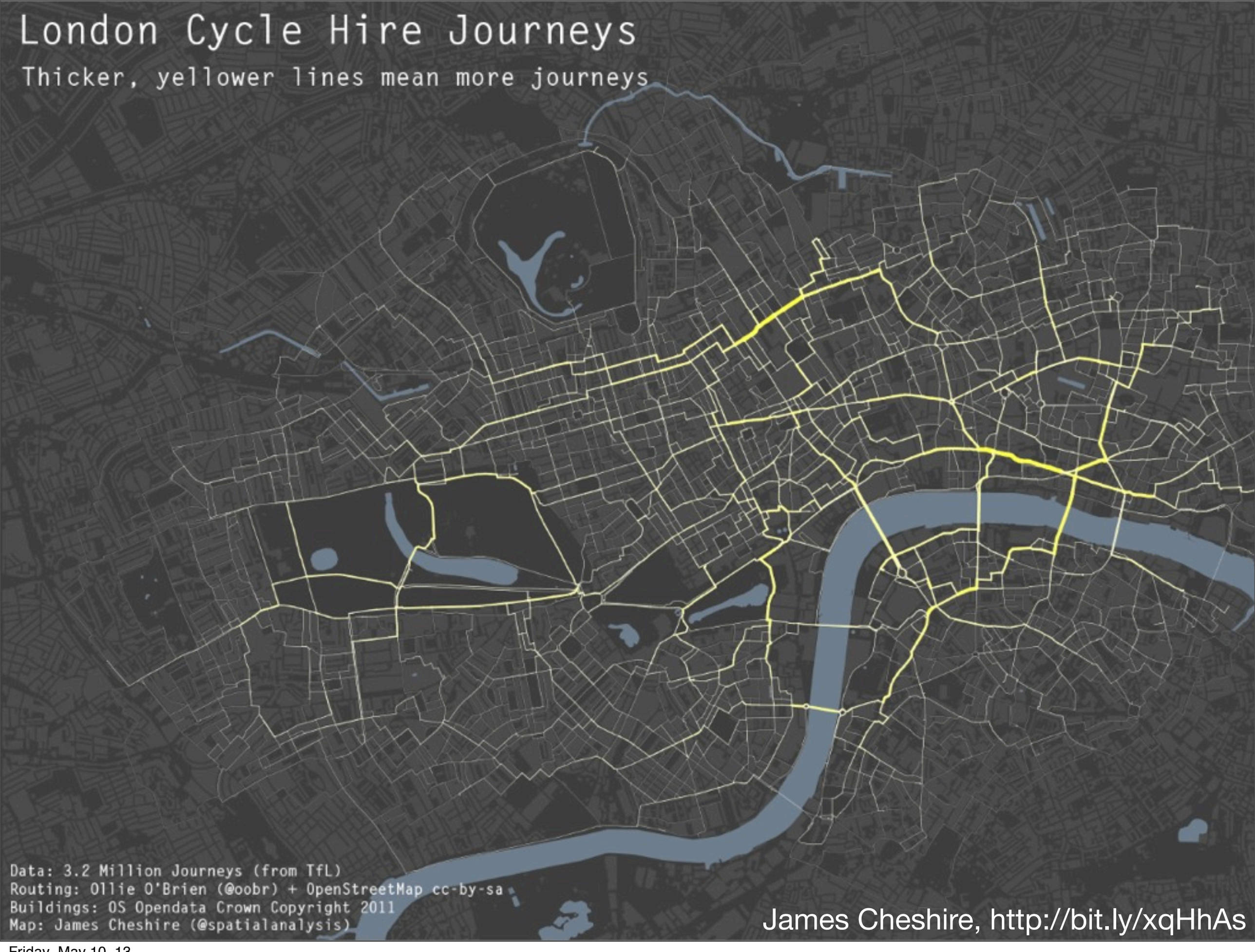
David Kahle, https://dl.dropbox.com/u/24648660/ggmap%20useR%202012.pdf



Winston Chang, http://shop.oreilly.com/product/0636920023135.do

London Cycle Hire Journeys

Thicker, yellower lines mean more journeys



James Cheshire, http://bit.ly/xqHhAs