Basic plots

Hadley Wickham

Assistant Professor / Dobelman Family Junior Chair

Department of Statistics / Rice University

July 2010

1. Scatterplots
2. Adding extra variables with facetting and aesthetics
3. Jittering and boxplots
4. Bar charts
5. Histograms

The data

Global school based healthy survey

Three countries: Uganda, The Philippines and the United Arab Emirates

Extracted variables related to diet and hand washing

Getting started

```
# If you haven't already...
install.packages("ggplot2")
# Every time you load R
library(ggplot2)

load(file.choose())

# Or if you have your working directory
# set up (very good idea!)
load("gshs.rdata")
```

Working directory

Remember to set your working directory.

From the terminal (linux or mac): the working directory is the directory you're in when you start R

On windows: `setwd(choose.dir())`

On the mac: ⌘-D

Scatterplot basics

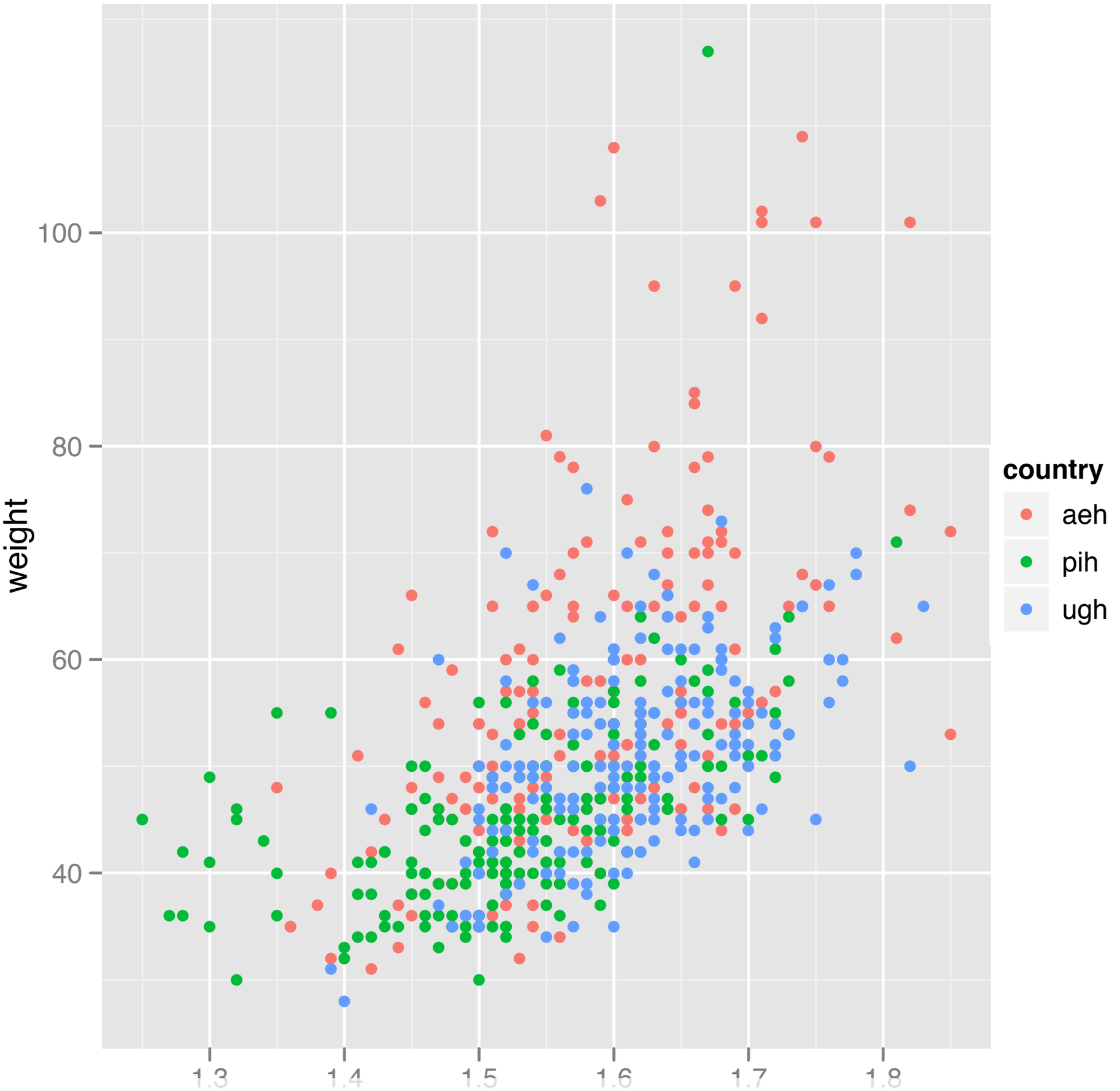
```
head(gshs)
str(gshs)
summary(gshs)

qplot(weight, height, data = gshs)
# To start with:
qplot(weight, height, data = sample)
```

Your turn

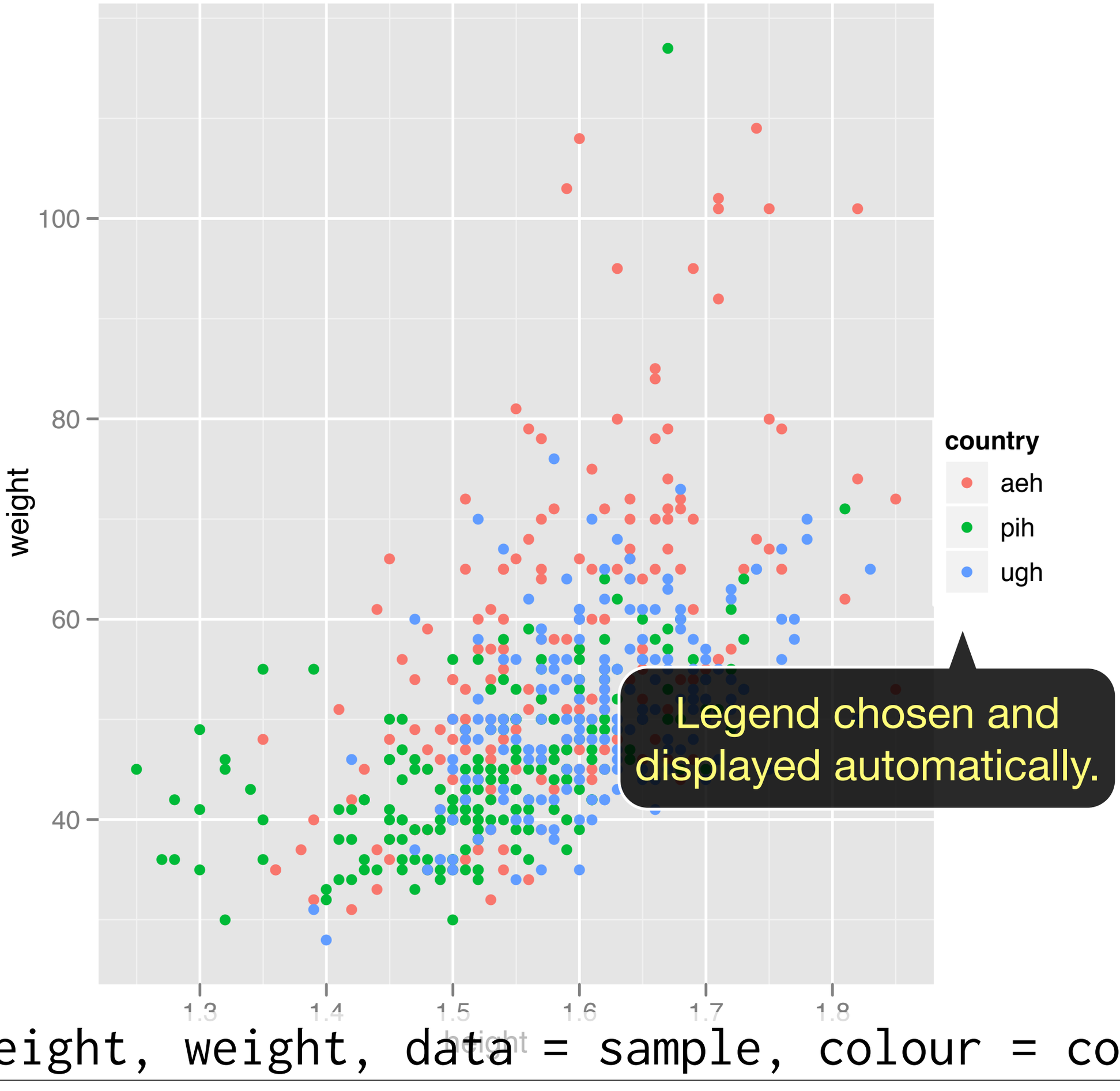
Load the data then make scatterplots of age, weight, height and bmi.

Additional variables

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




```
qplot(height, weight, data = sample, colour = country)
```




```
qplot(height, weight, data = sample, colour = country)
```

Your turn

Run the code from previous slides, then experiment with the **colour**, **size**, and **shape** aesthetics. How does the display change when you use **discrete** vs **continuous** variables? What happens when you **combine** multiple aesthetics?

	Discrete	Continuous
Colour	Evenly spaced hues	Gradient from red to blue
Size	Discrete size steps	Linear mapping between radius and value
Shape	Different shape for each	Shouldn’t work

Faceting

Small multiples display different subsets of the data.

Useful for exploring conditional relationships. Useful for large data.

Your turn

```
qplot(height, weight, data = sample) +
facet_grid(. ~ sex)

qplot(height, weight, data = sample) +
facet_grid(country ~ .)

qplot(height, weight, data = sample) +
facet_grid(country ~ sex)

qplot(height, weight, data = sample) +
facet_wrap(~ hungry)
```

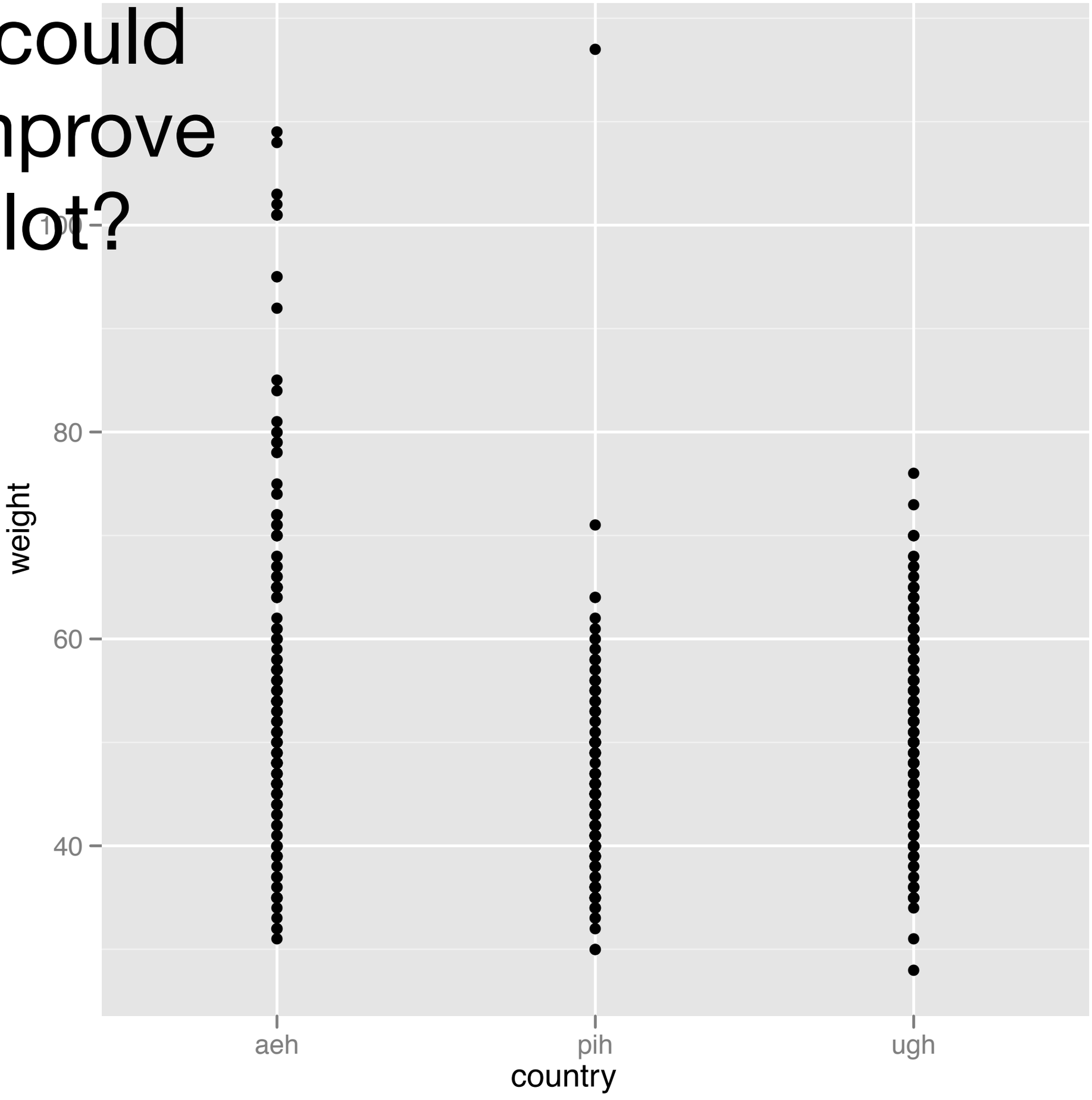
Summary

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

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

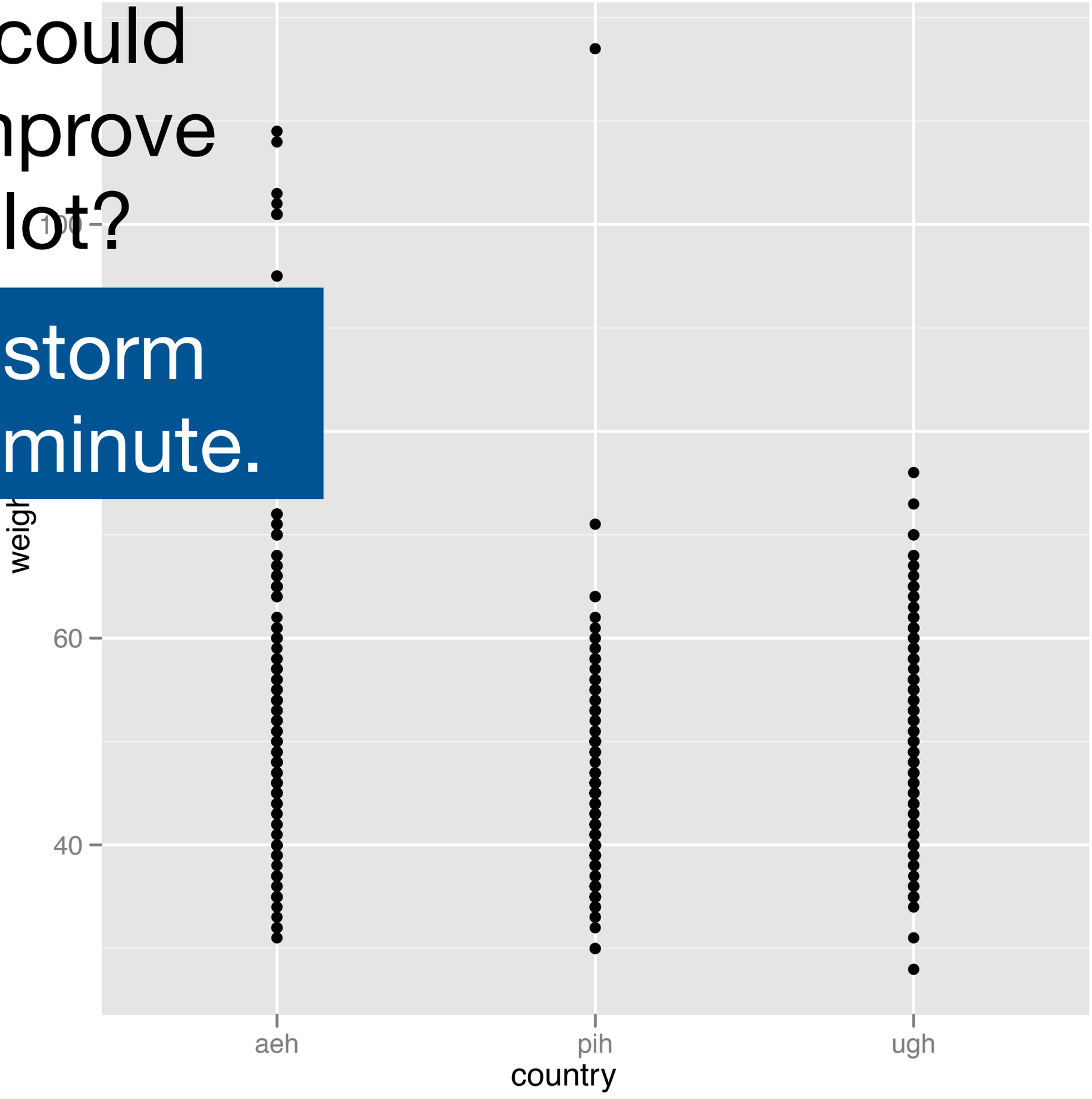
Can control whether scales are common or individual with the scales argument.

How could we improve this plot?

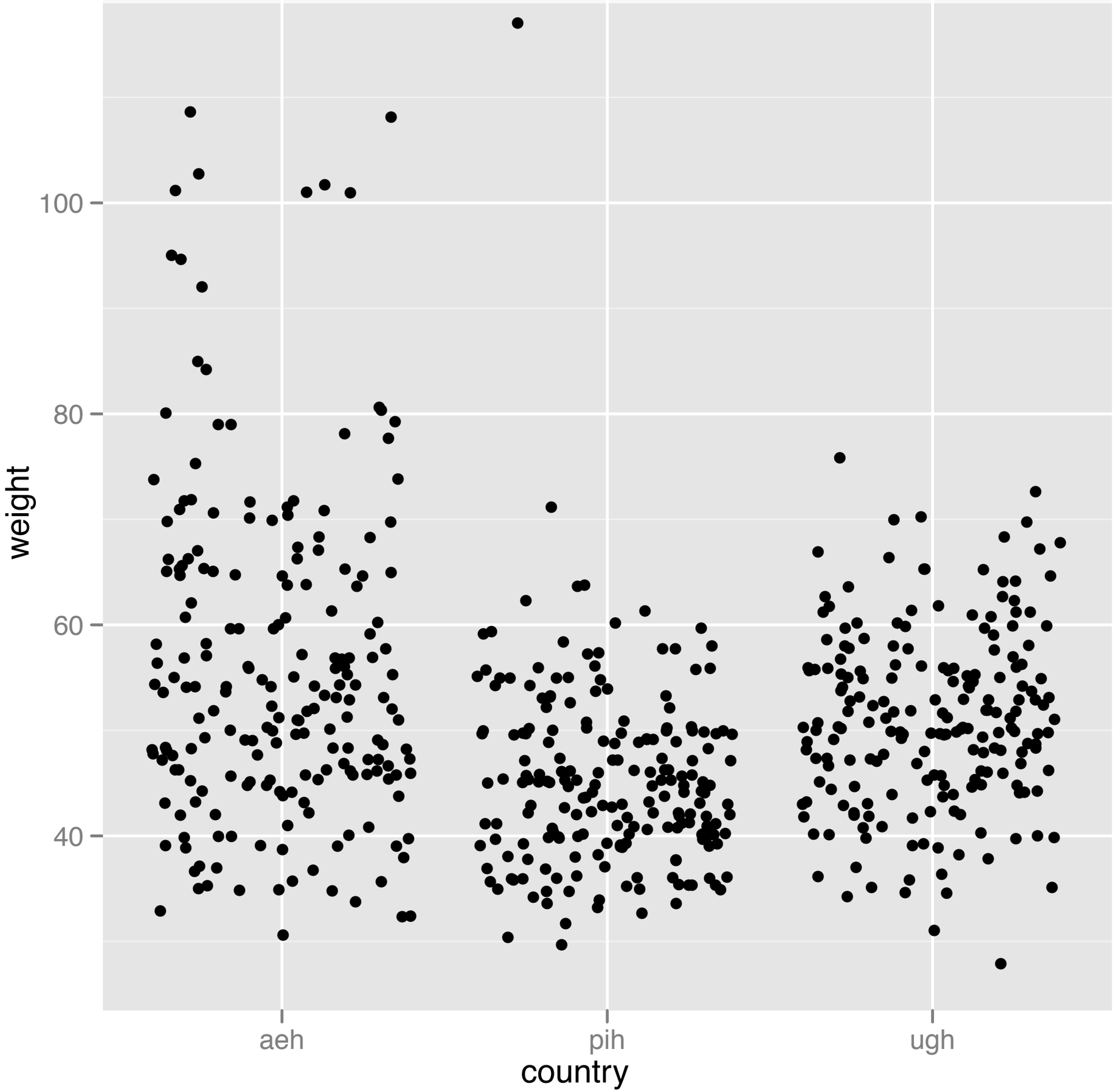


How could we improve this plot?

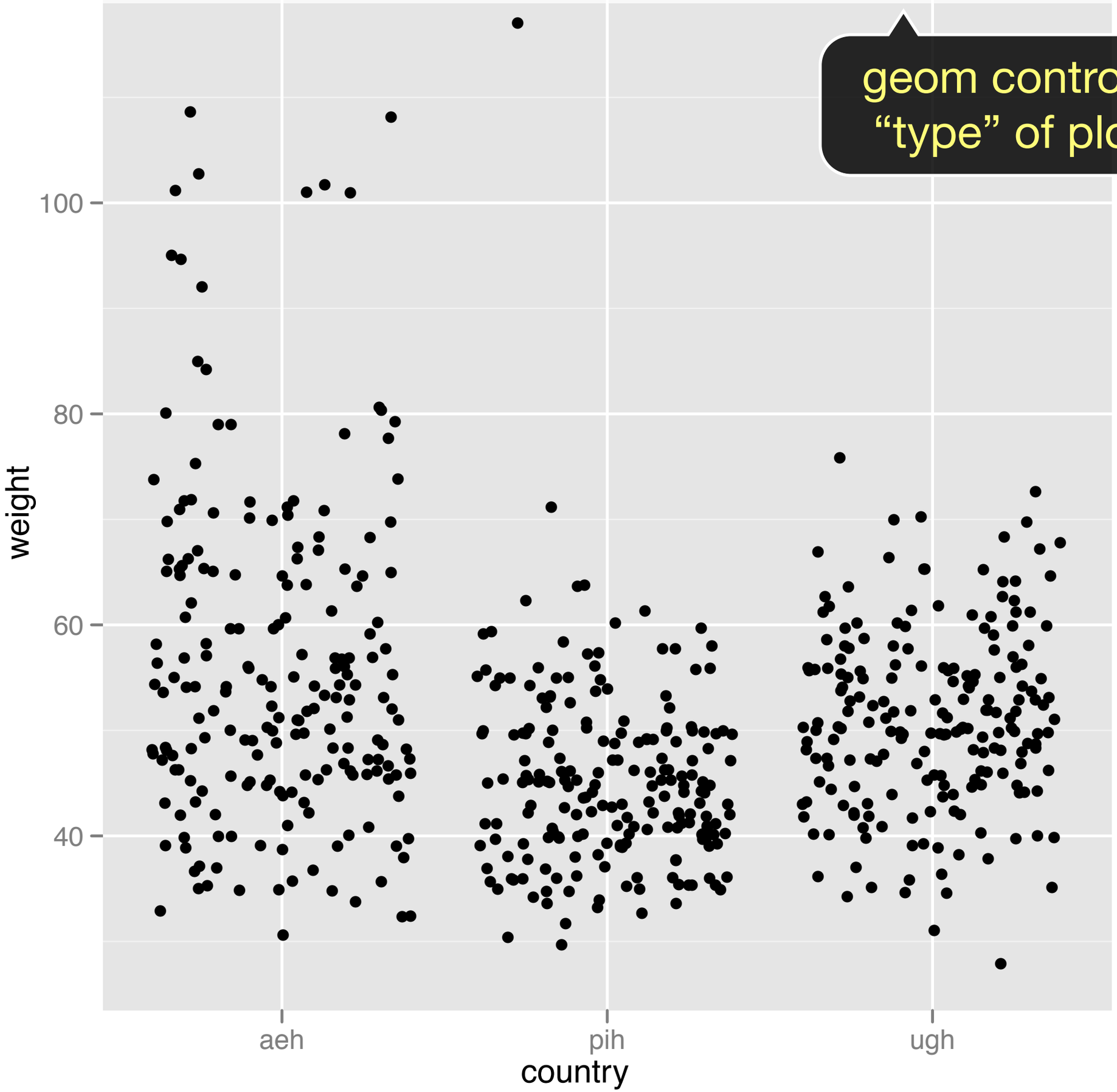
Brainstorm for 1 minute.



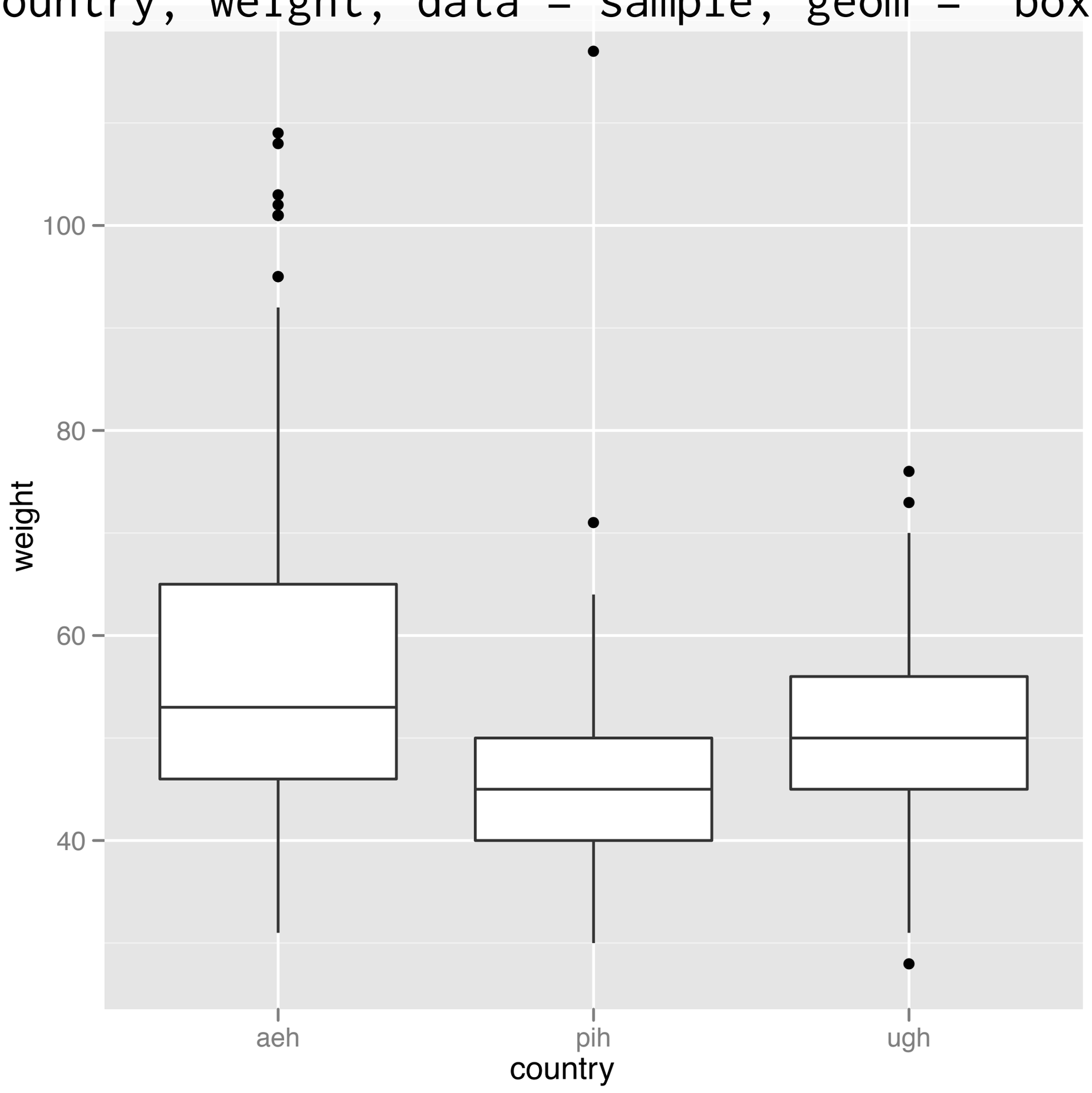
```
qplot(country, weight, data = sample, geom = "jitter")
```

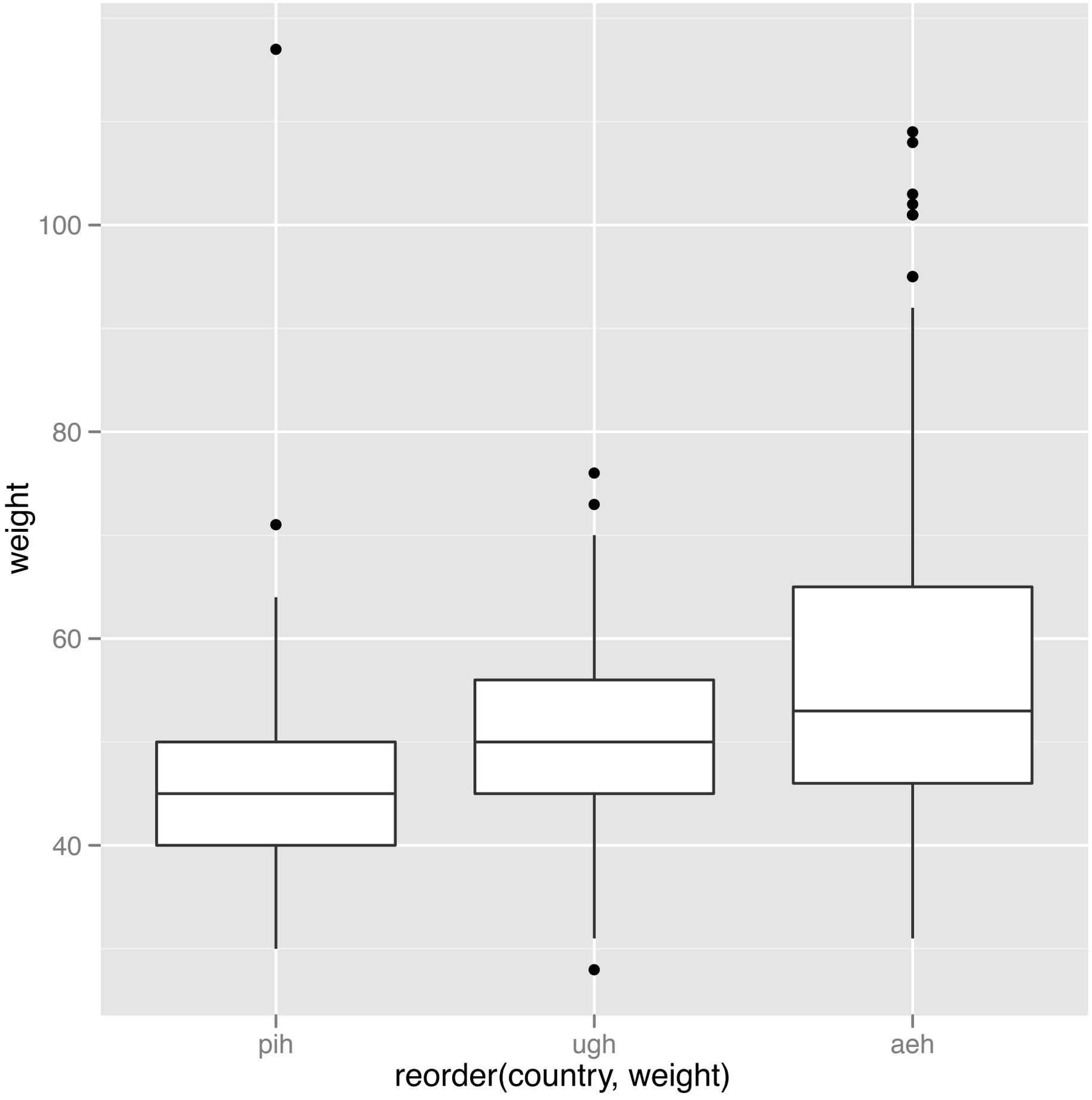


```
qplot(country, weight, data = sample, geom = "jitter")
```

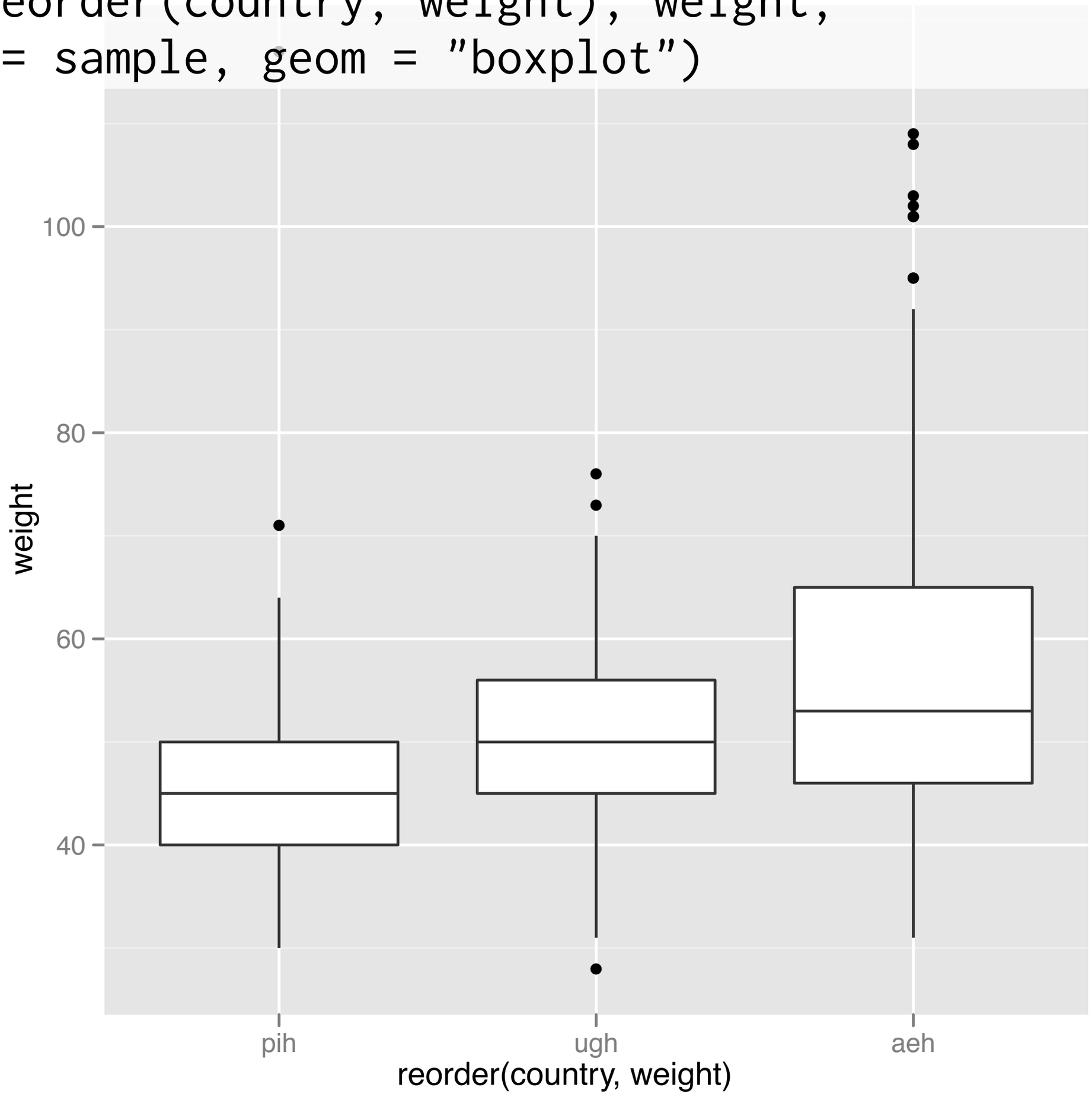


```
qplot(country, weight, data = sample, geom = "boxplot")
```



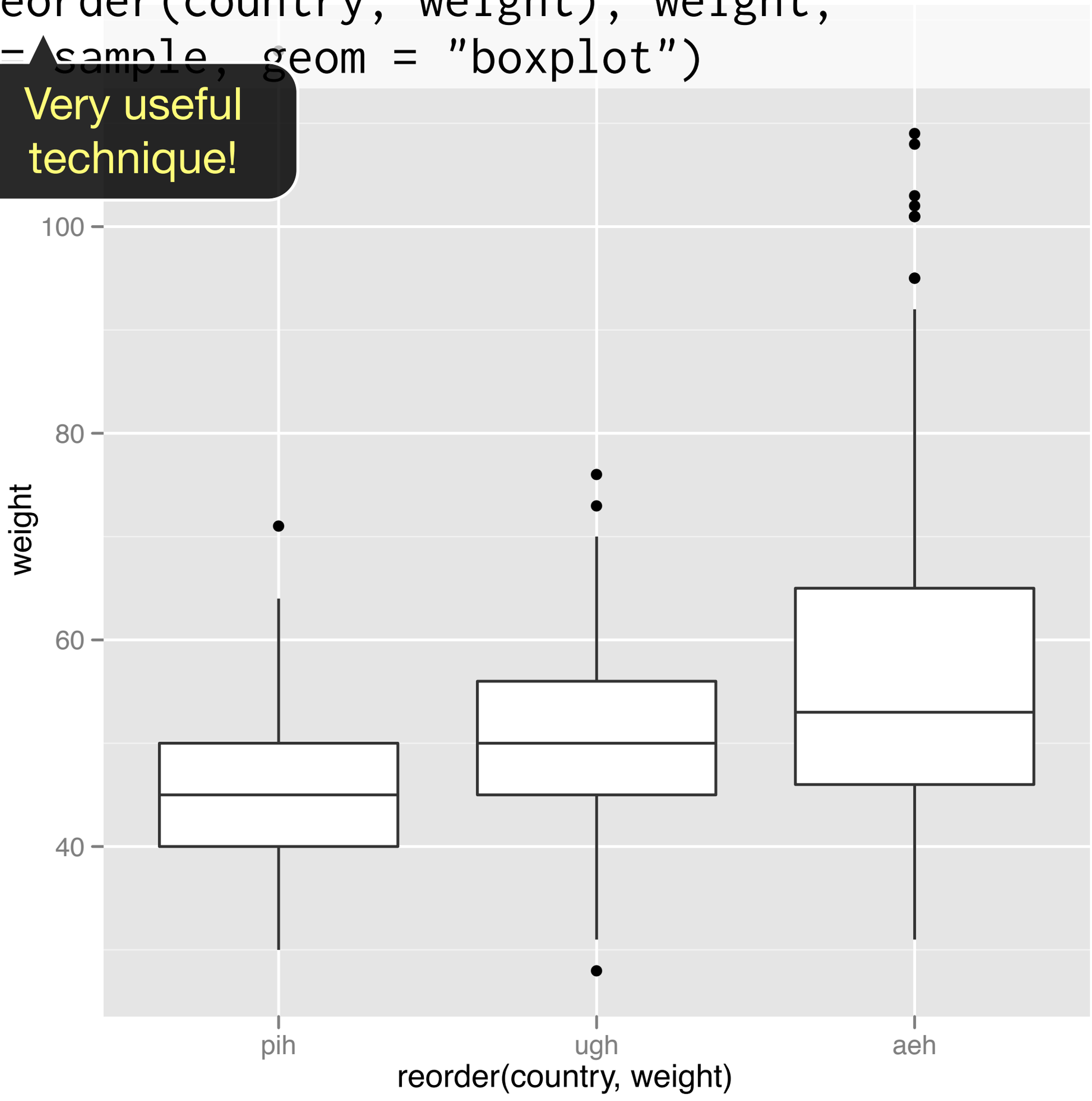

100
80
60
40
weight
pih
ugh
aeh
reorder(country, weight)

```
qplot(reorder(country, weight), weight,
  data = sample, geom = "boxplot")
```

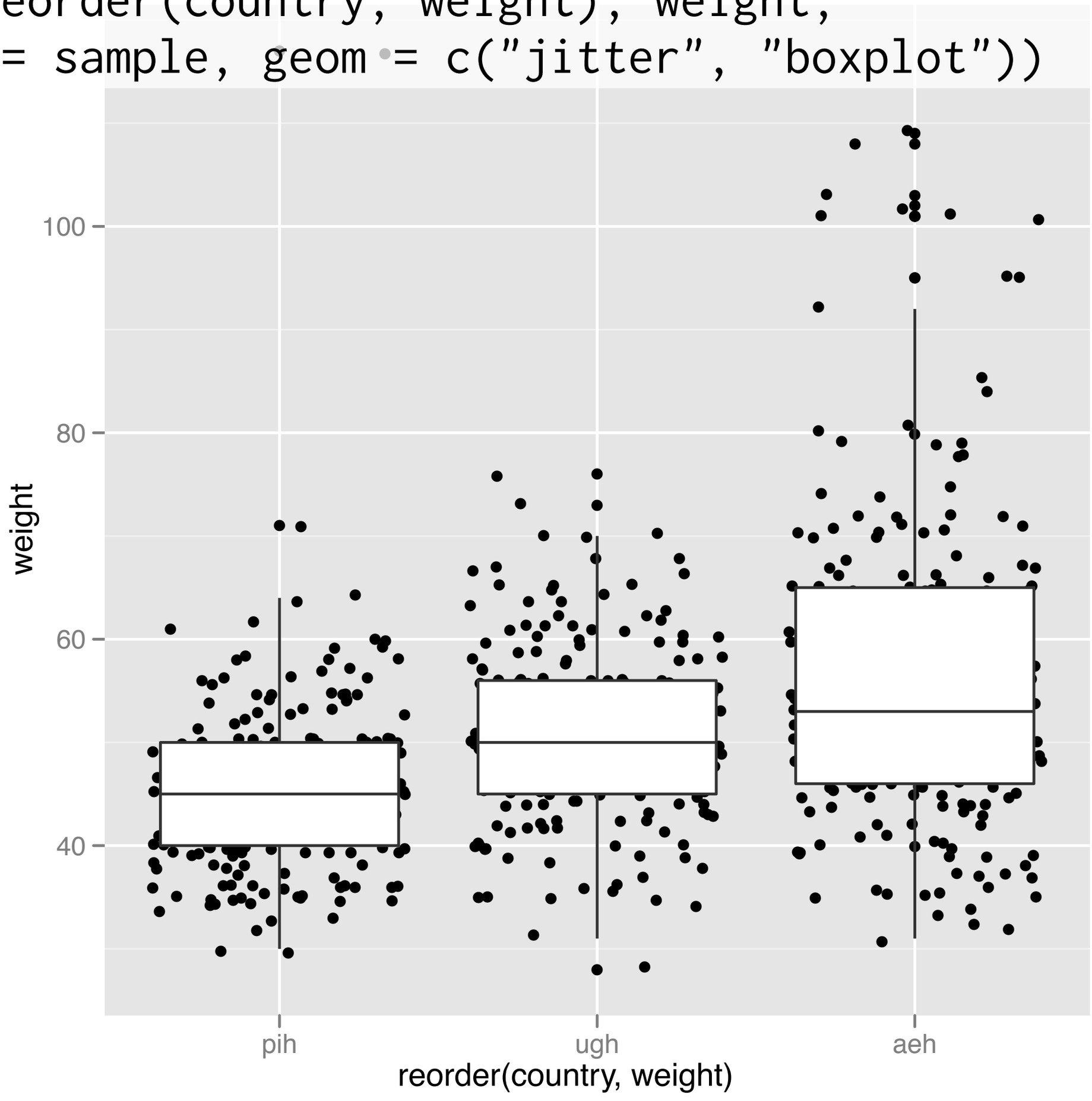


```
qplot(reorder(country, weight), weight,
  data = sample, geom = "boxplot")
```

Very useful technique!

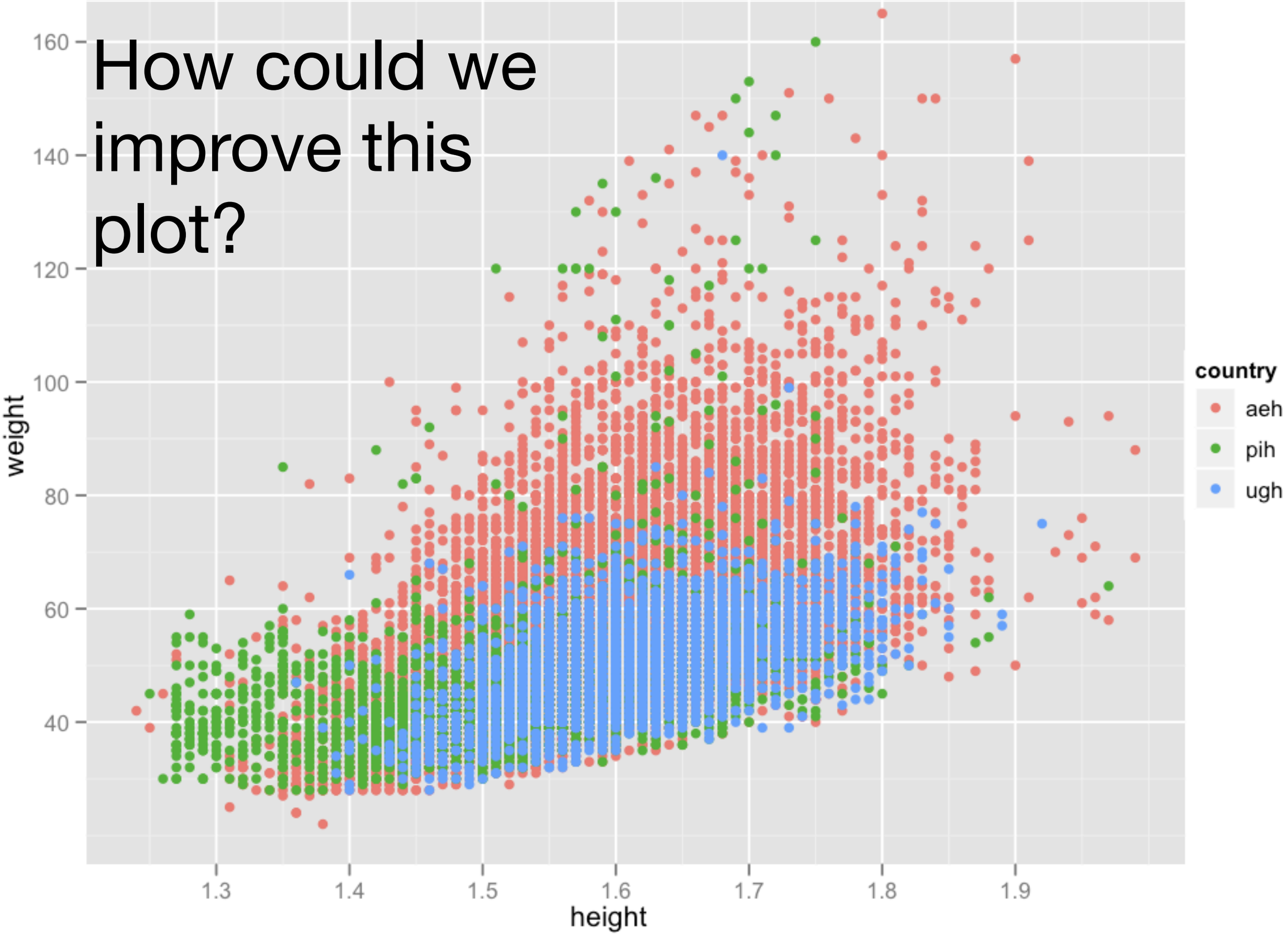


```
qplot(reorder(country, weight), weight,
  data = sample, geom = c("jitter", "boxplot"))
```



```
qplot(height, weight, data = gshs, colour = country)
```

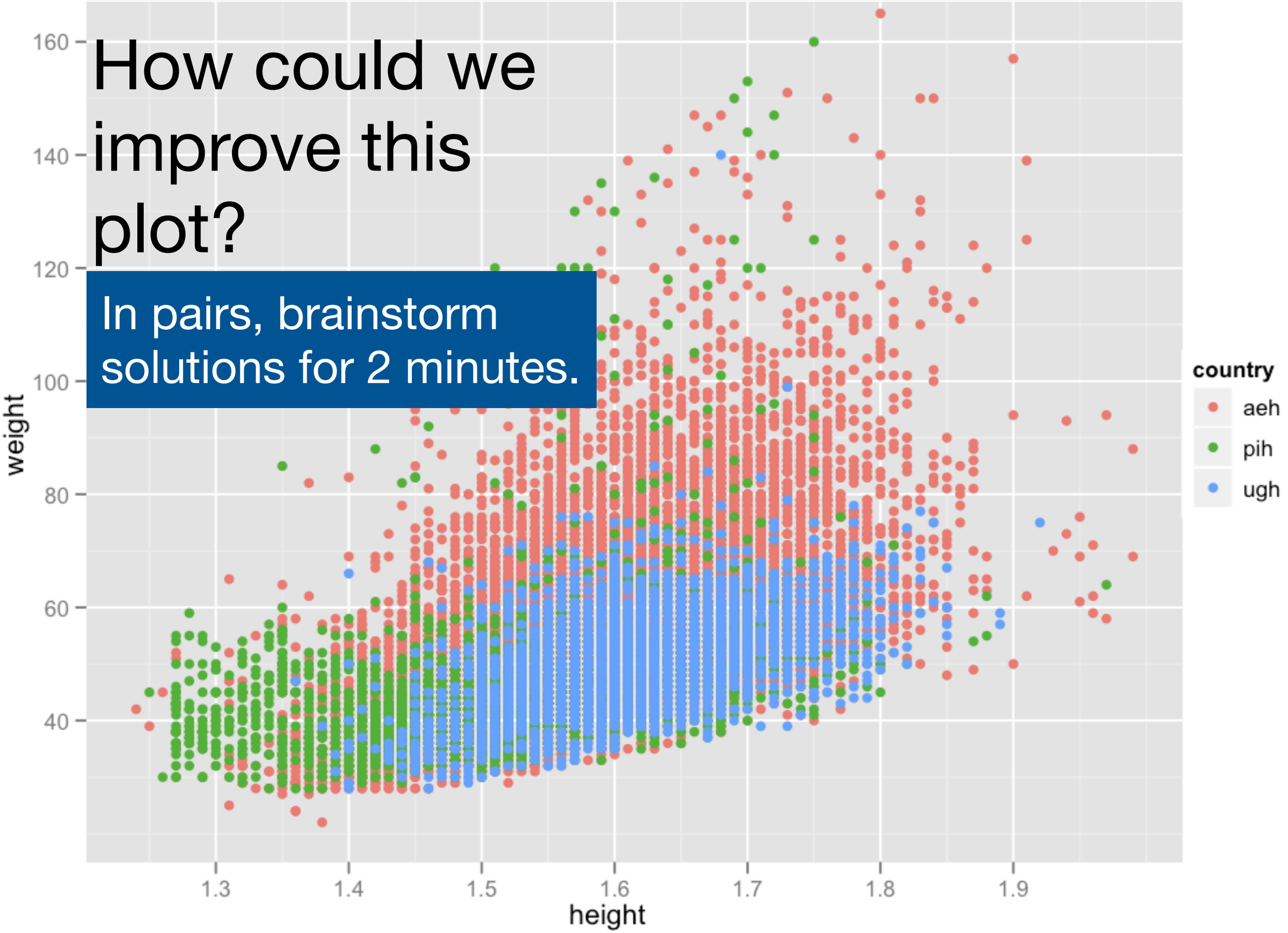
How could we improve this plot?



```
qplot(height, weight, data = gshs, colour = country)
```

How could we improve this plot?

In pairs, brainstorm solutions for 2 minutes.



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

Bar charts

```
qplot(country, data = sample)

qplot(country, data = gshs)
qplot(hungry, data = gshs)
qplot(fruit, data = gshs)
qplot(vegetables, data = gshs)

qplot(country, data = gshs, weight = sample_weight)
qplot(hungry, data = gshs, weight = sample_weight)
```

Additional variables

As with scatterplots can use **aesthetics** or **faceting**.

Using the **fill** aesthetic creates plots that are pretty, but they can be hard to read.

```
# Let's try and explore the relationship between
# country and amount of fruit eaten

qplot(country, data = gshs, fill = fruit)
qplot(fruit, data = gshs, fill = country)
```

```
# Problem: different numbers in each country
qplot(country, data = gshs, fill = fruit,
  position = "fill")
# But not easy to compare
```

```
with(gshs, table(country, fruit, exclude = NULL))
with(gshs, table(country, fruit))

table <- with(gshs, table(country, fruit))
percent <- prop.table(table, 1)
percent_df <- as.data.frame(percent)

qplot(country, data = percent_df, fill = fruit)
qplot(country, data = percent_df, weight = Freq,
  fill = fruit)
qplot(fruit, data = percent_df, weight = Freq,
  fill = country)

qplot(fruit, Freq, data = percent_df, geom = "line",
  colour = country, group = country)
```

Summary

`table`: computes counts

`prop.table`: divides out one margin

`as.data.frame`: converts to data.frame (ggplot2 only works with data frames)

Your turn

How is fruit and vegetable consumption related? Always look at marginal (1d) distributions first.

Histograms

```
qplot(weight, data = gshs)
qplot(weight, data = gshs, binwidth = 10)
qplot(weight, data = gshs, binwidth = 5)
qplot(weight, data = gshs, binwidth = 1)

# That's a bit suspicious looking.  Let's look
# at rounding more closely.
# %% is modulo operator (remainder after integer
division)
qplot(weight %% 10, data = gshs, binwidth = 1)
last_plot() + facet_wrap(~ country)
```

Always experiment with the bin width!

Your turn

Explore the distributions of height and bmi. Do you find any suspicious patterns there?

Experiment with `geom = "freqpoly"` and `geom = "density"`

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)