Scatter plots

UNDERSTANDING DATA VISUALIZATION

Richie Cotton
Data Evangelist at DataCamp

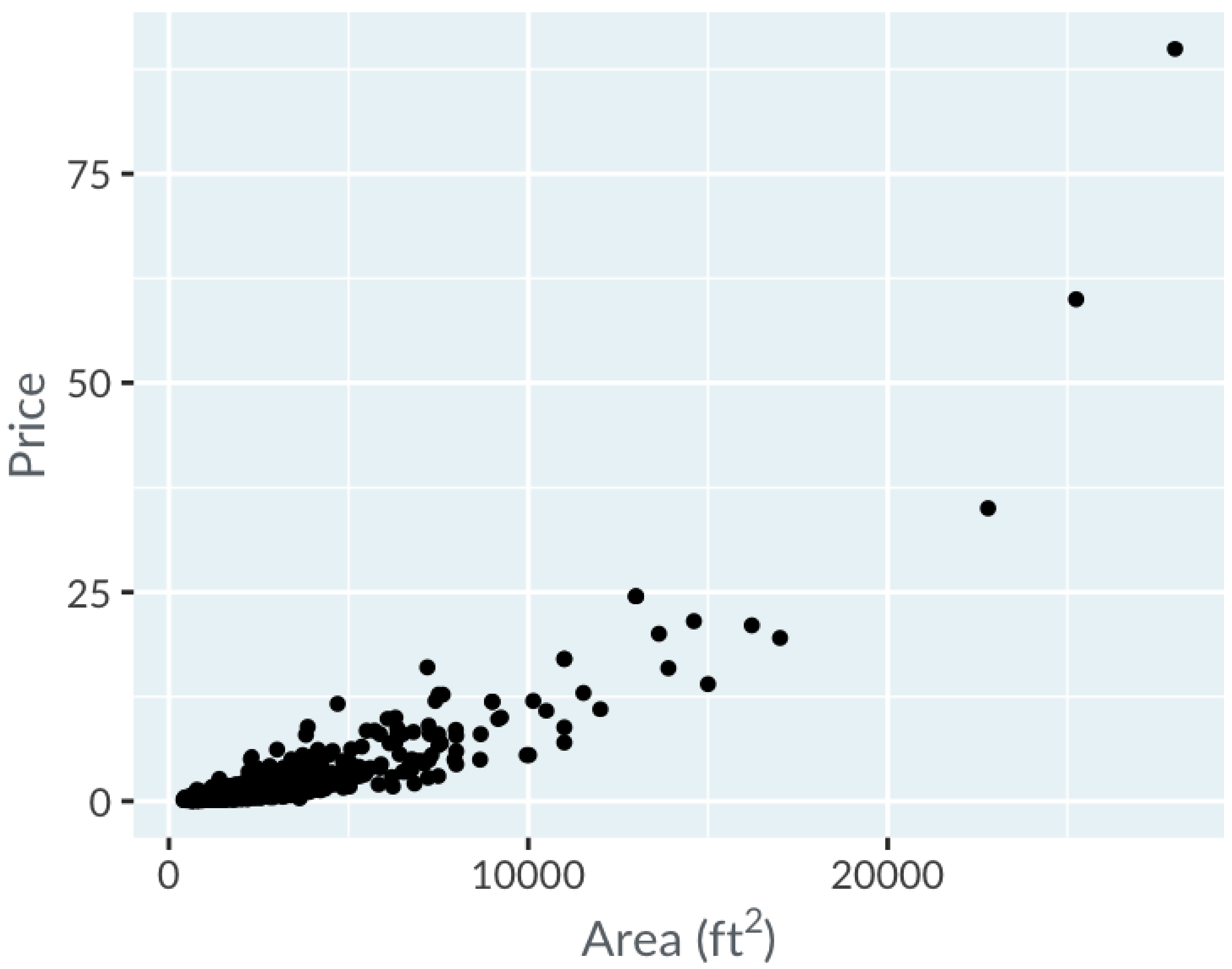
When should you use a scatter plot?

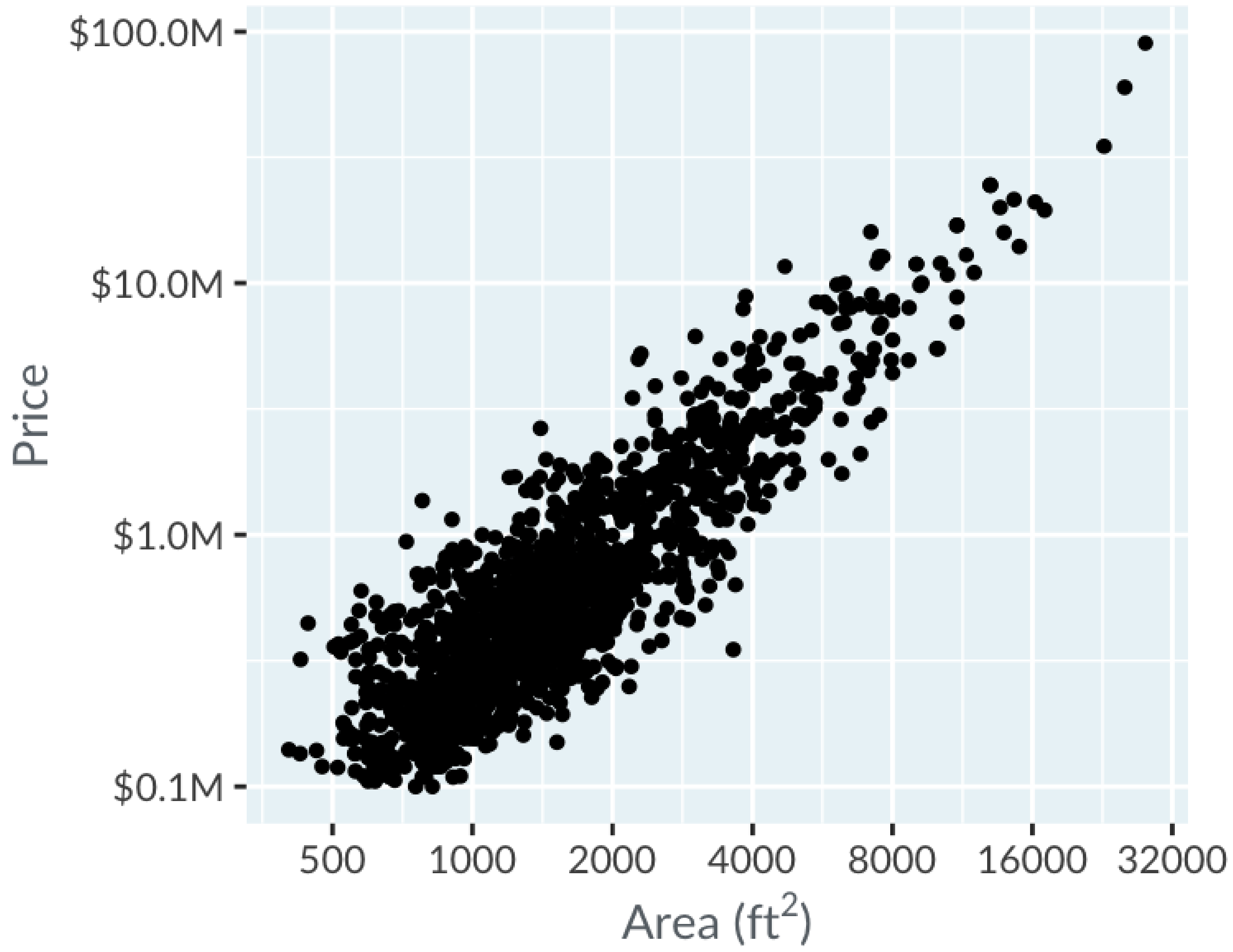
1. You have two continuous variables.
2. You want to answer questions about the relationship between the two variables.

Los Angeles County home prices

city	n_beds	price_musd	area_sqft
Long Beach	1	0.3250	846
Beverly Hills	3	2.1950	2930
Santa Monica	2	0.5740	1037
Santa Monica	1	0.5990	576
Beverly Hills	5	3.9500	5600
Long Beach	4	0.2999	1571
Westwood	3	0.6950	1913

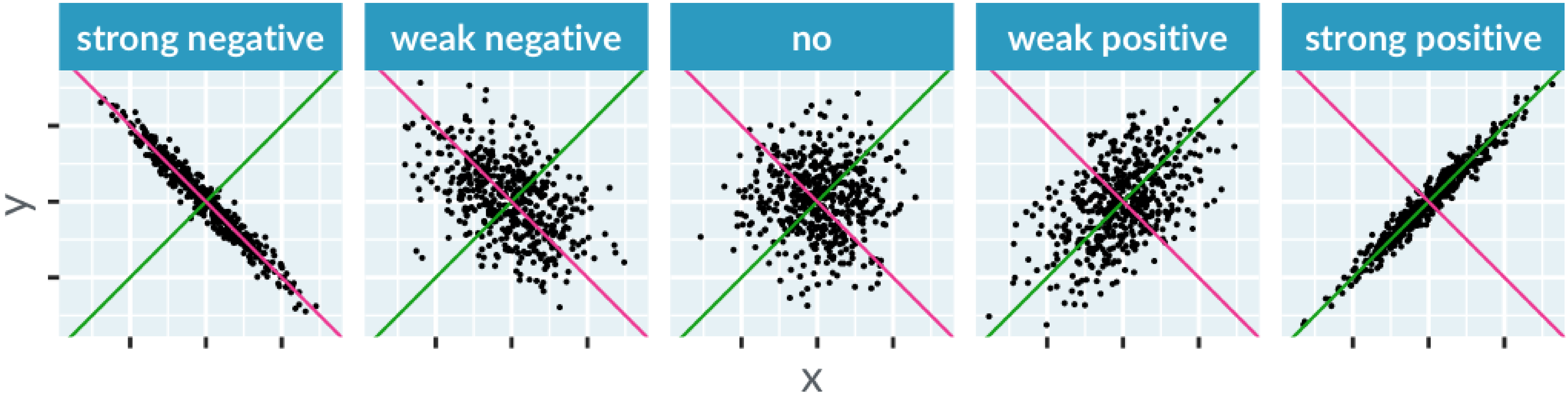
Prices vs. area



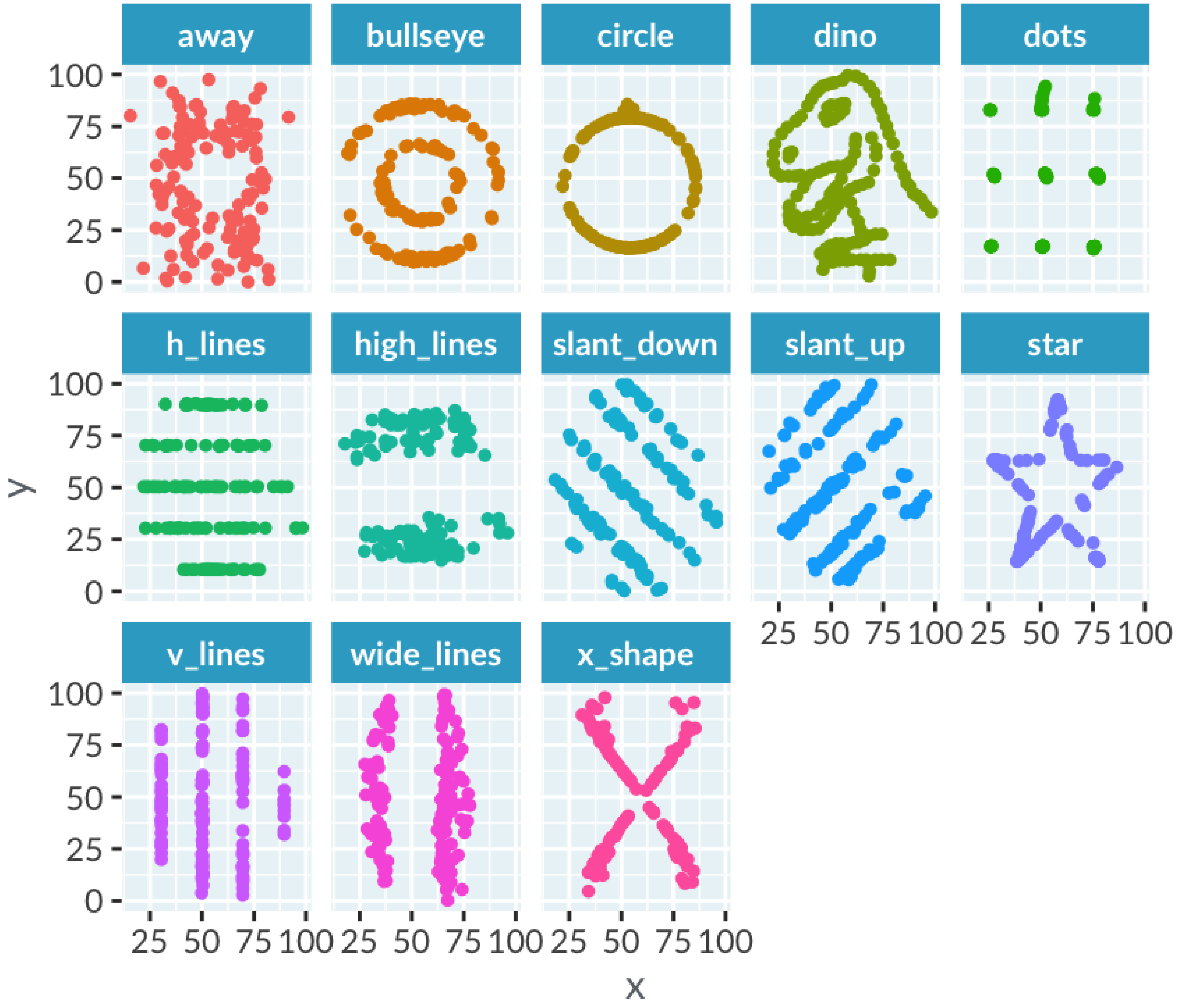


Correlation

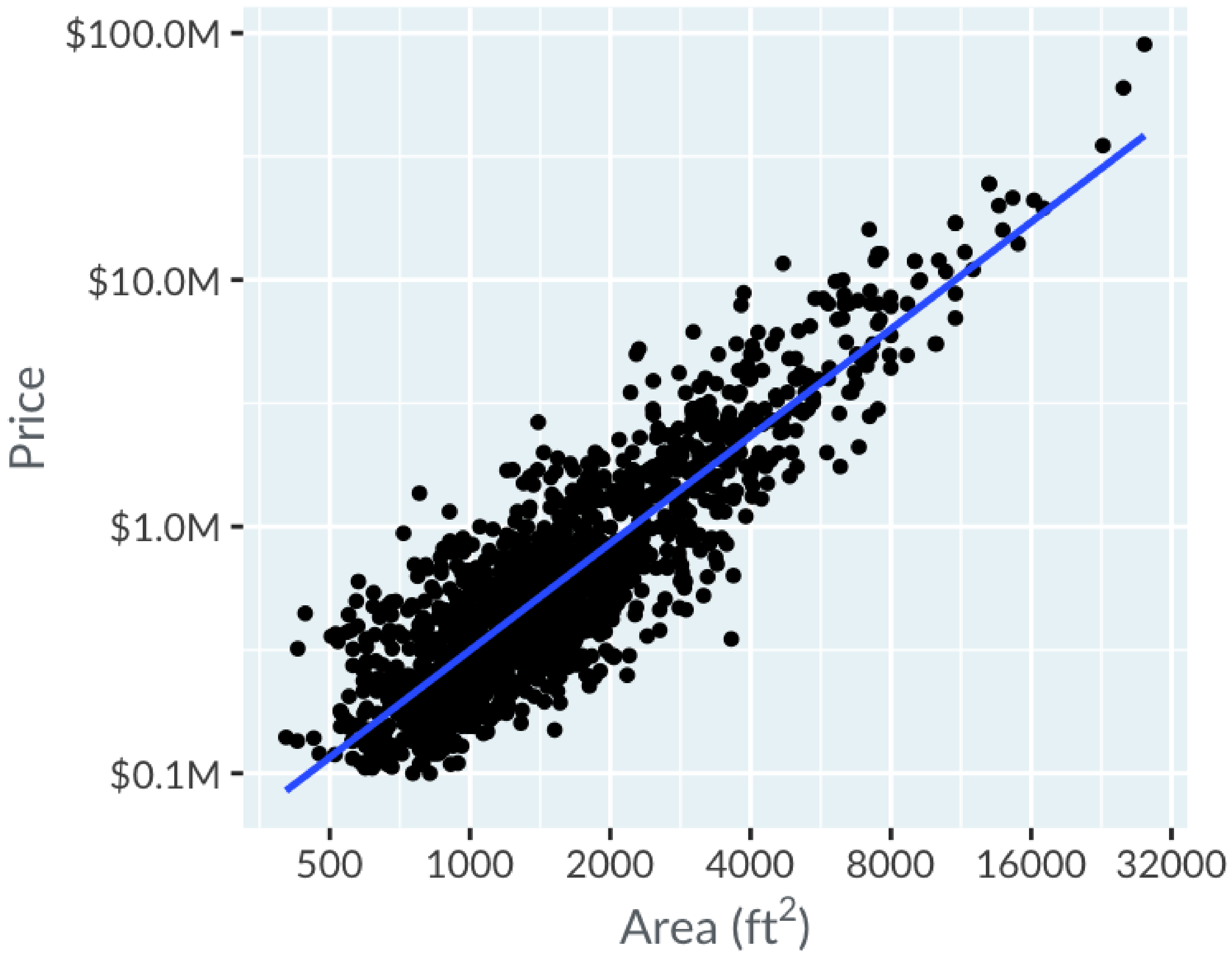
> How close are you to being able to fit a straight line through the points?



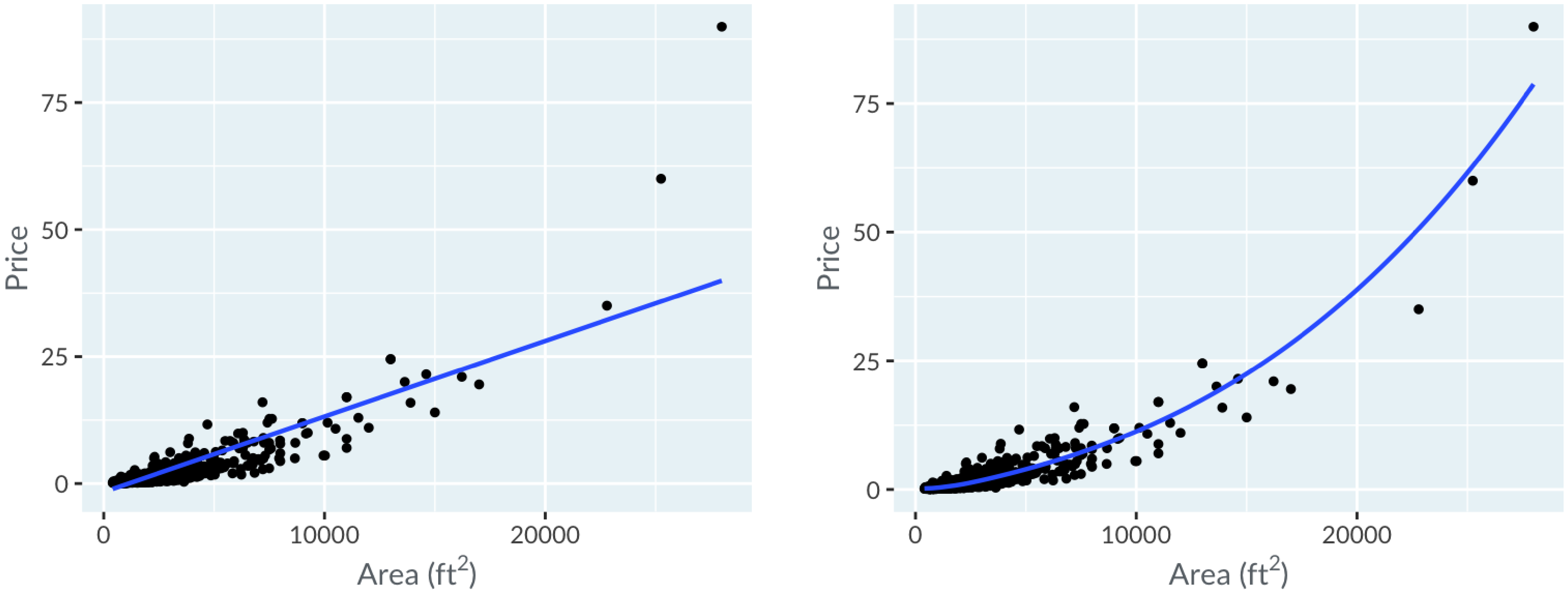
Sometimes correlation isn't helpful



Adding trend lines



Adding smooth trend lines



Let's practice!

UNDERSTANDING DATA VISUALIZATION

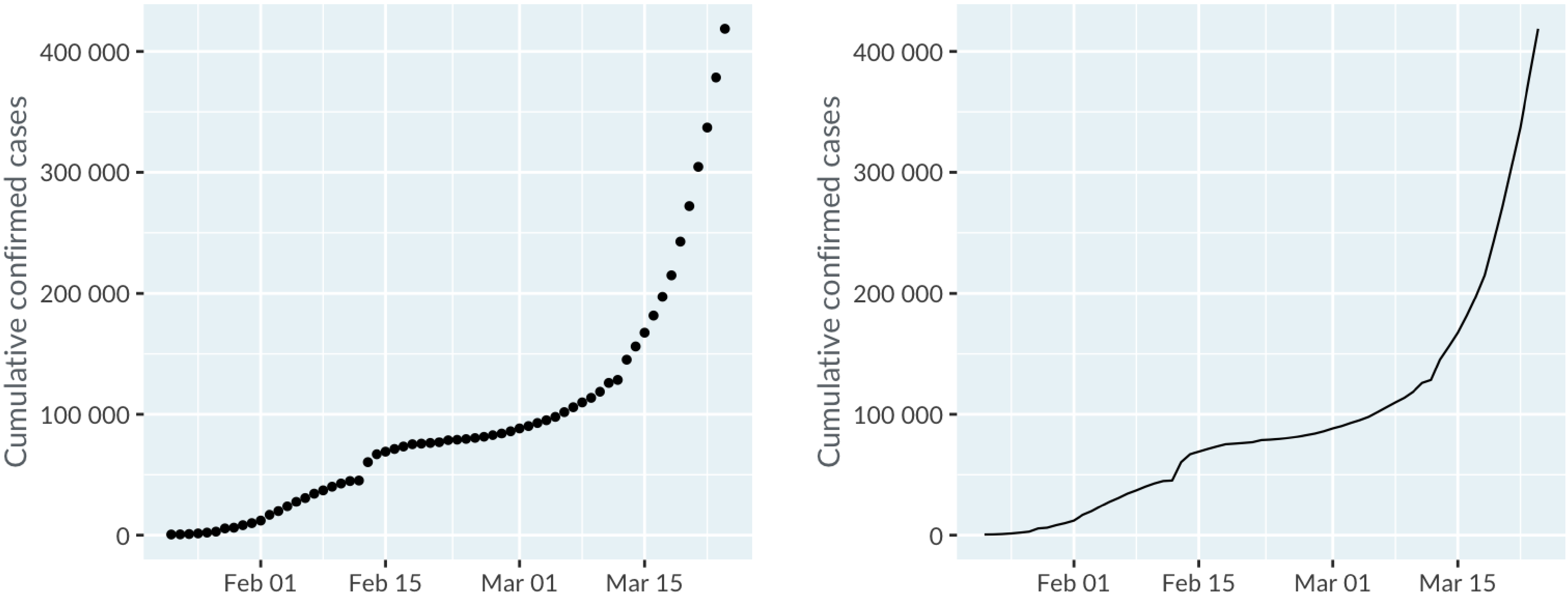
Line plots

UNDERSTANDING DATA VISUALIZATION

Richie Cotton

Data Evangelist at DataCamp

Worldwide COVID-19 coronavirus cases



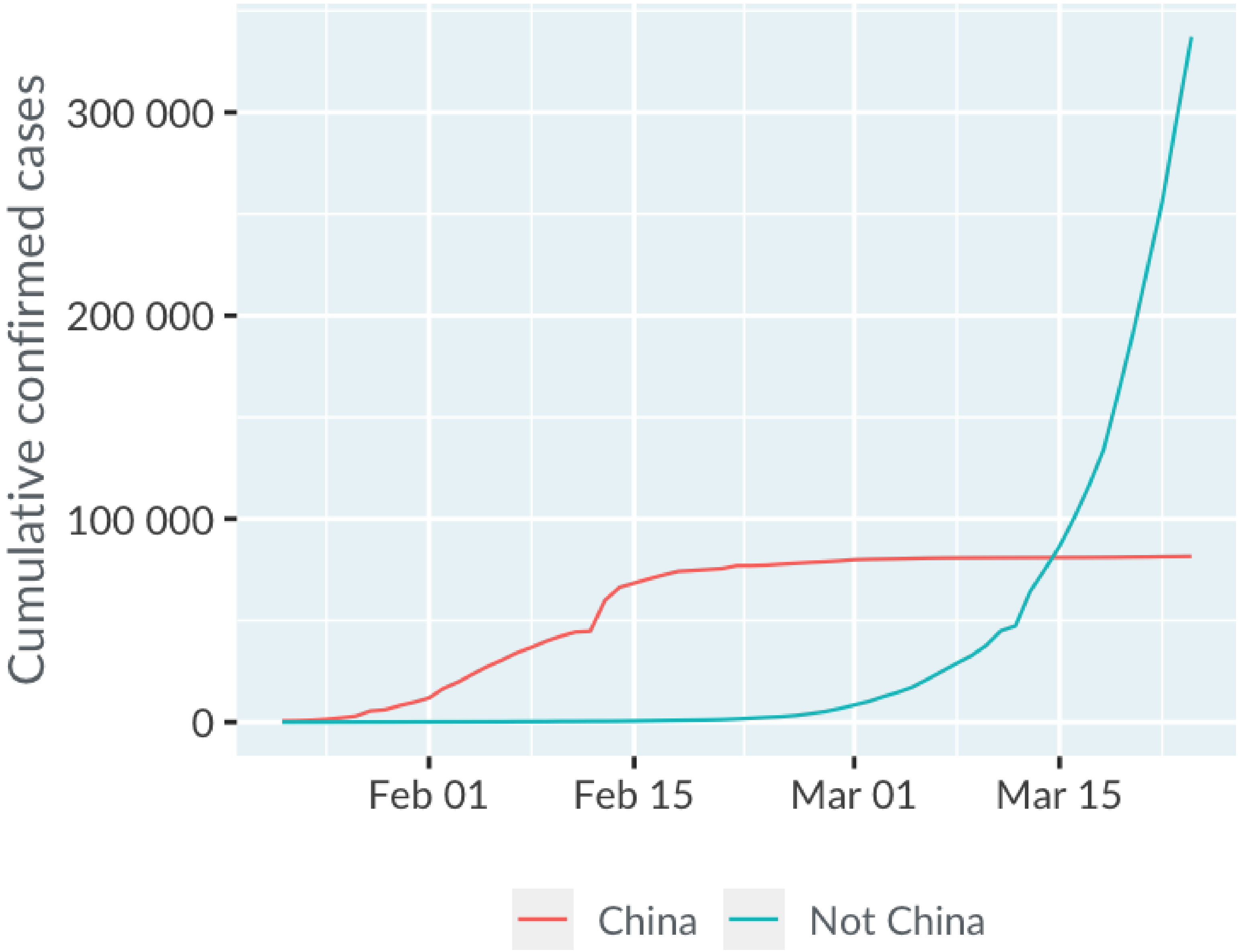
[1] https://github.com/RamiKrispin/coronavirus

When should you use a line plot?

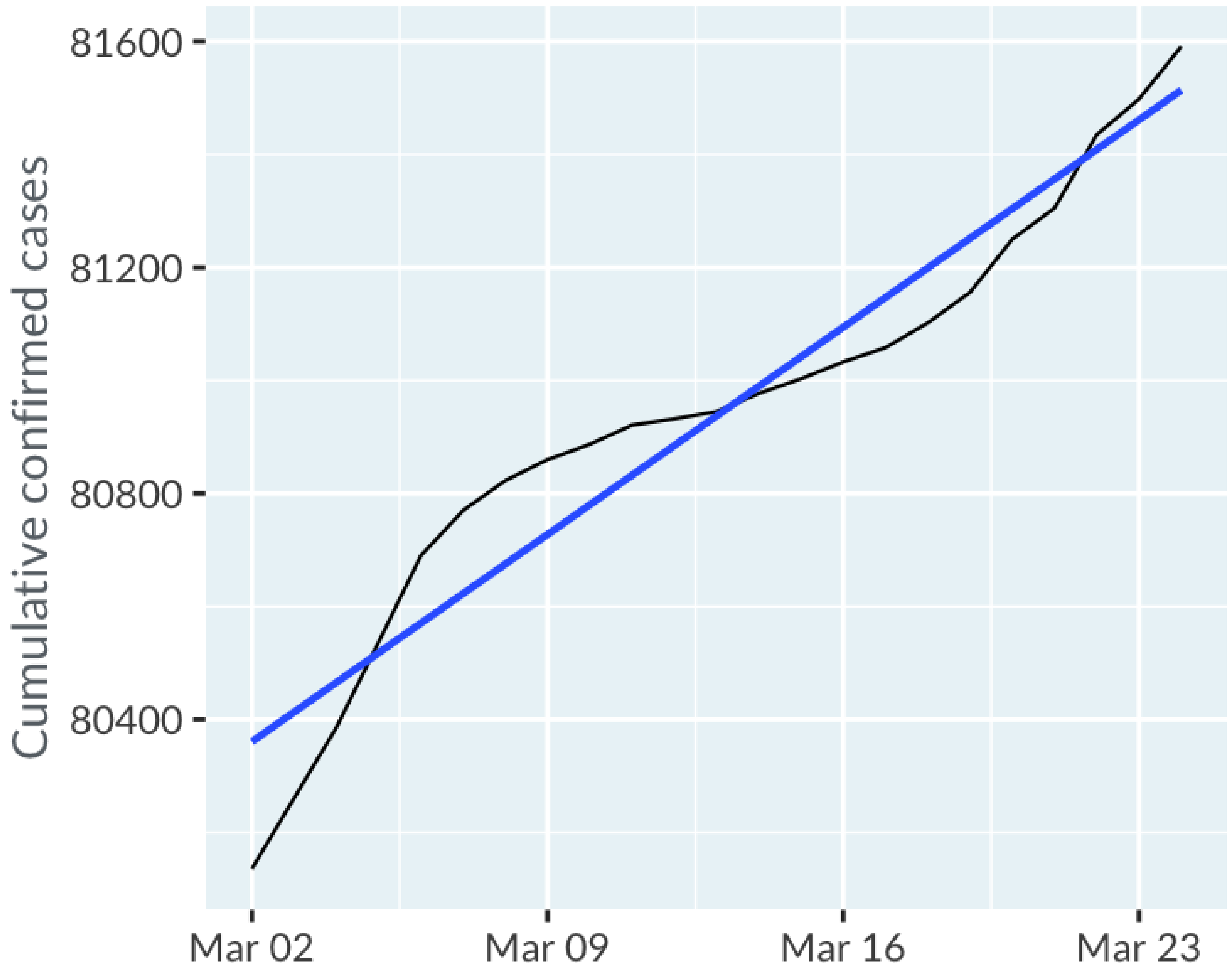
1. You have two continuous variables.
2. You want to answer questions about their relationship.
3. Consecutive observations are connected somehow.

Usually, but not always, the x-axis is dates or times.

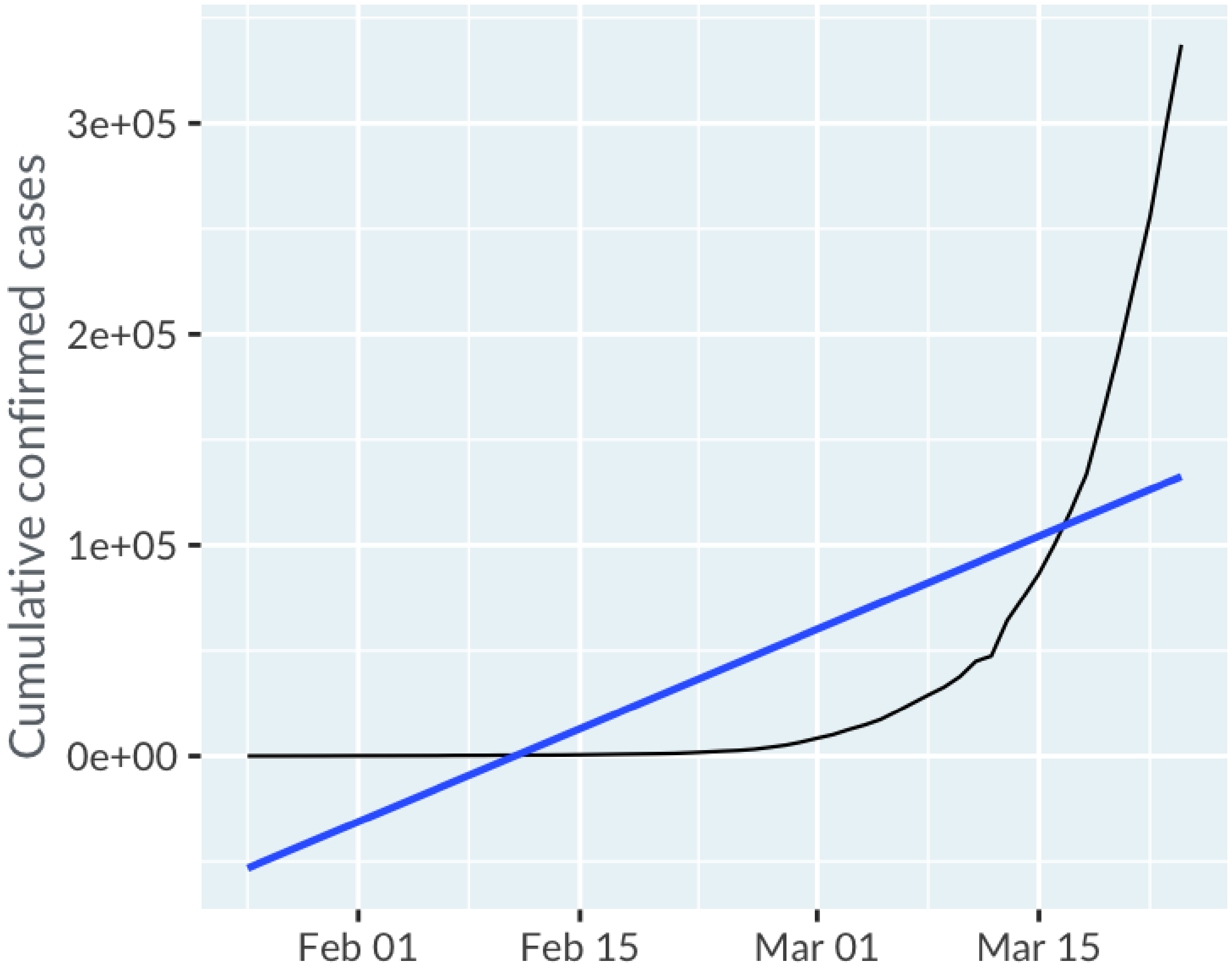
Comparing multiple lines

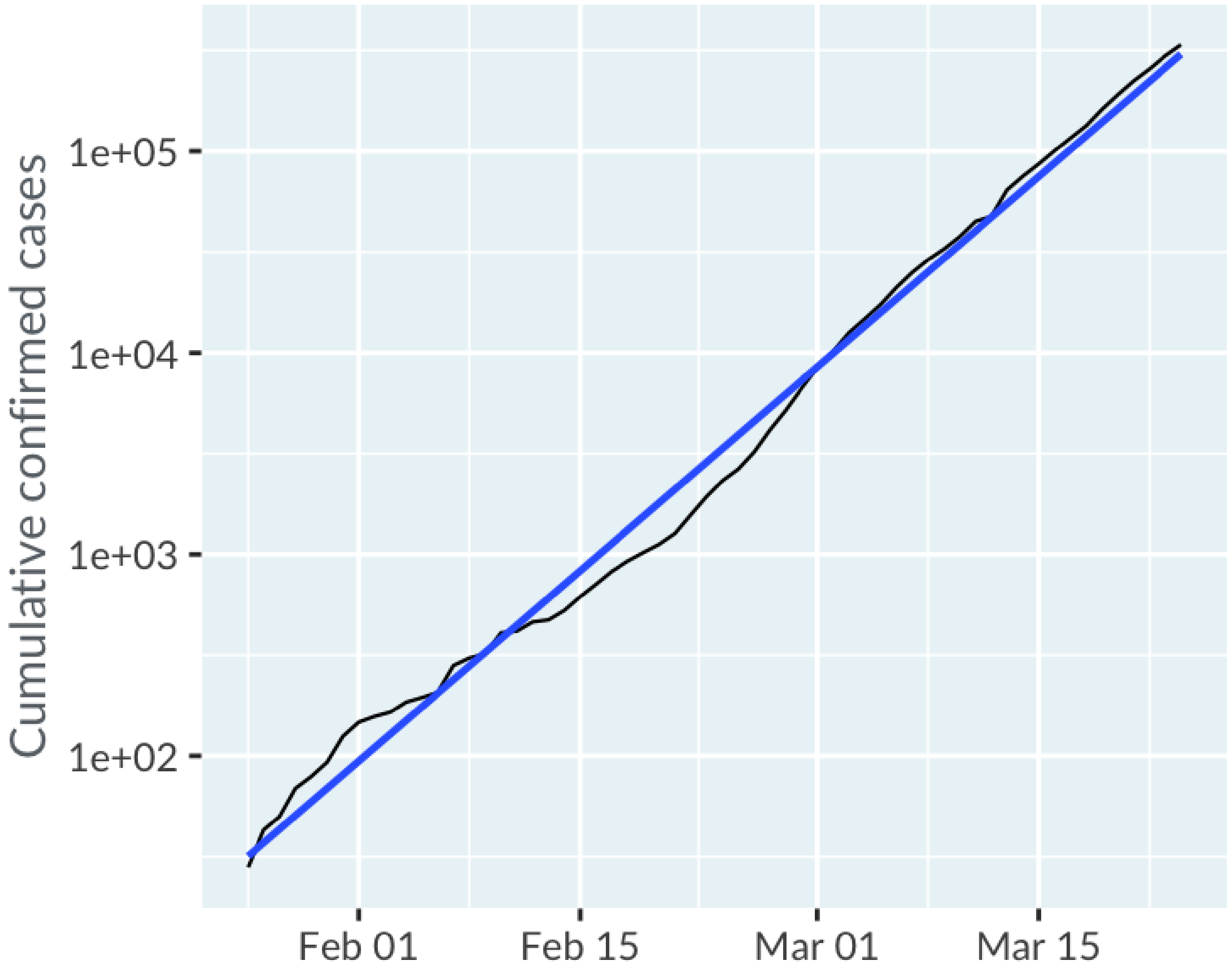


Trend lines

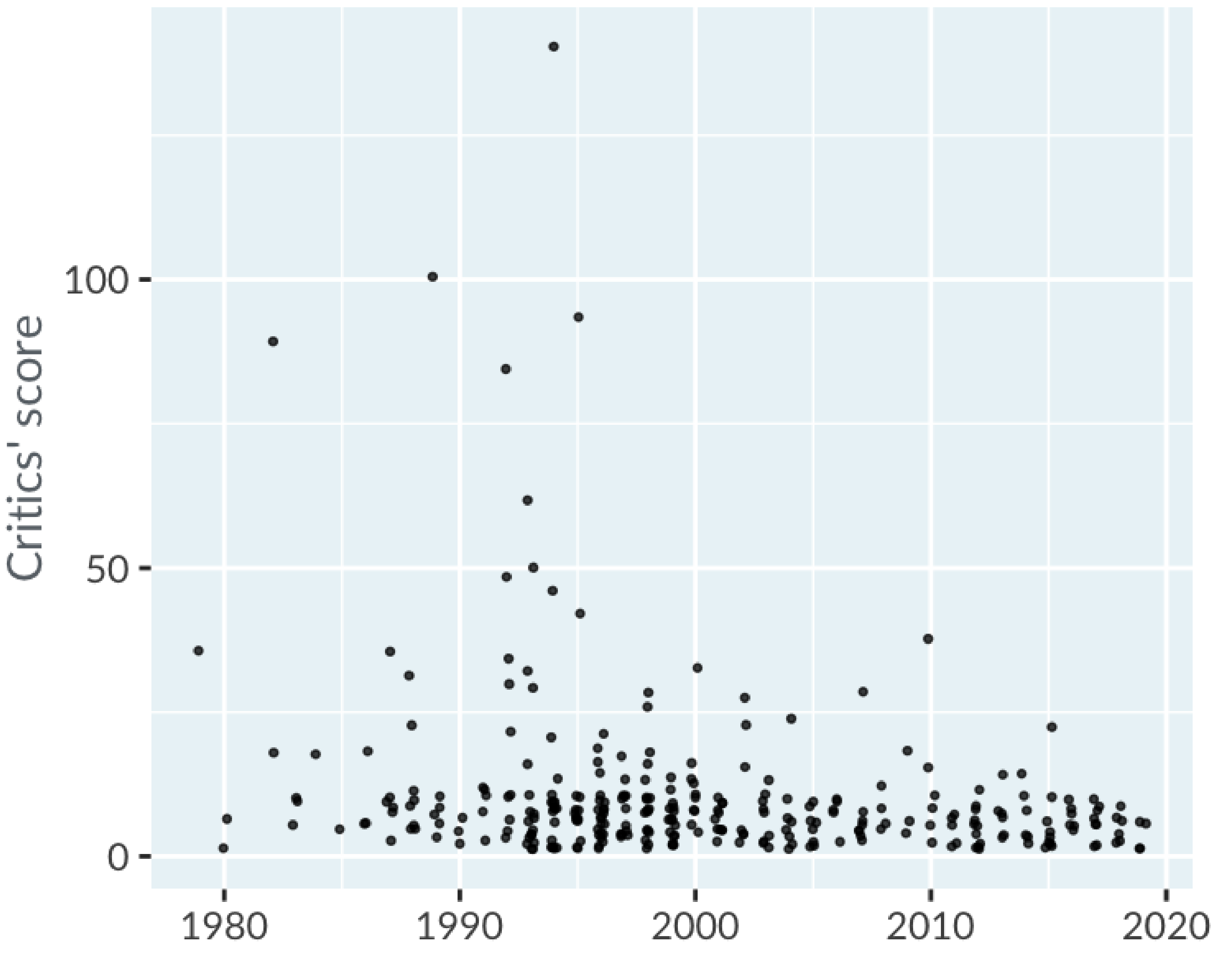

Cumulative confirmed cases
81600
81200
80800
80400
Mar 02
Mar 09
Mar 16
Mar 23

Trend lines + log scale



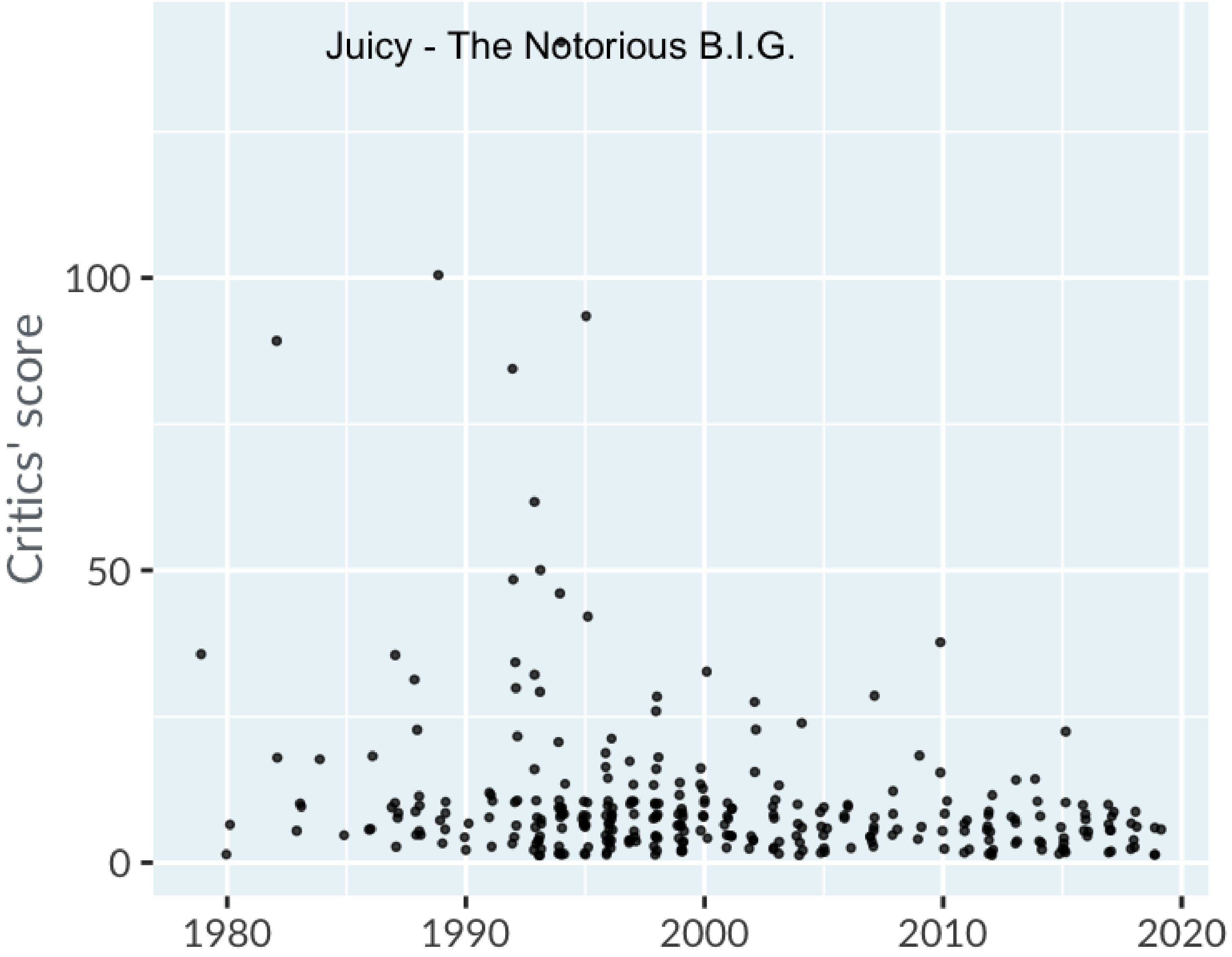


Time x-axis doesn't always imply line plot



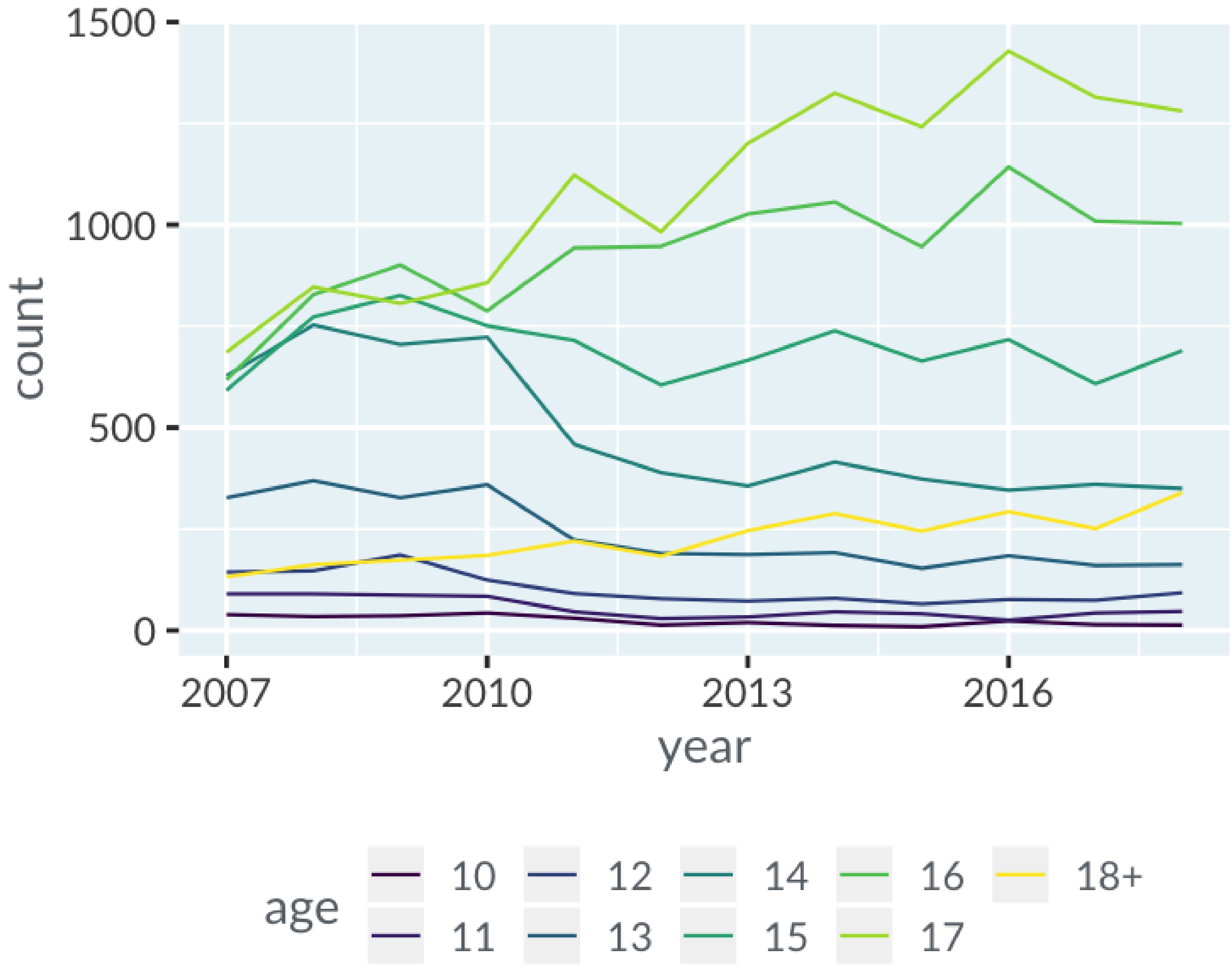
[1] https://blog.datawrapper.de/best-hip-hop-songs-of-all-time-visualized

Time x-axis doesn't always imply line plot



[1] https://blog.datawrapper.de/best-hip-hop-songs-of-all-time-visualized

Time x-axis doesn't always imply line plot



[1] https://opendata.swiss/en/dataset/statistik-der-jugendstrafrechtspflege-des-kantons-zuerich/resource/41430d20-d449-4a9b-bf91-3f9a16c9f229

Let's practice!

UNDERSTANDING DATA VISUALIZATION

Bar plots

UNDERSTANDING DATA VISUALIZATION

Richie Cotton
Data Evangelist at DataCamp

When should you use a bar plot?

Most common cases:

1. You have a categorical variable.
2. You want counts or percentages for each category.

Occasionally:

1. You want another numeric score for each category, and need to include zero in the plot.

> When will I, will I be famous?
> I can't answer, I can't answer that

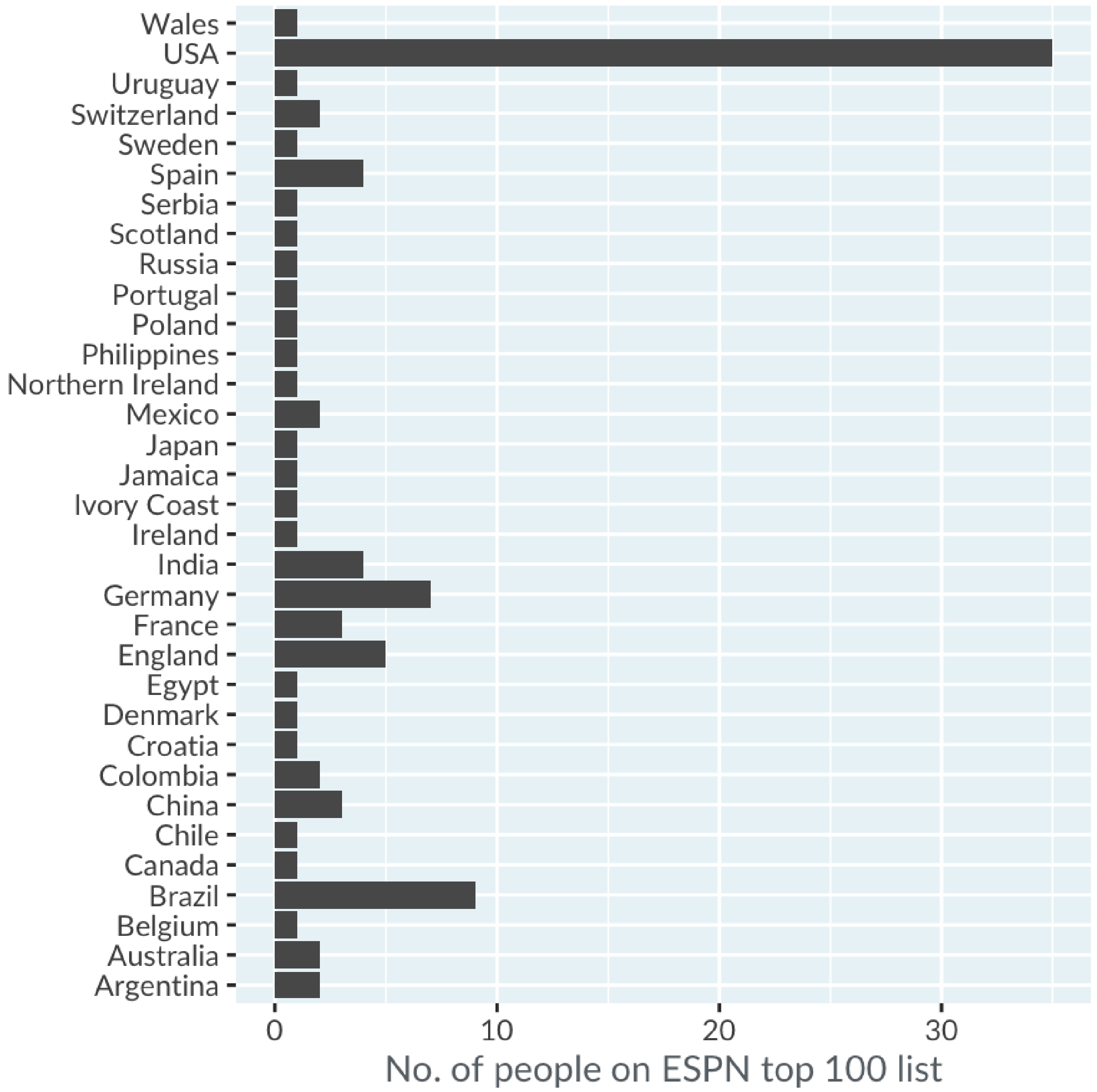
Bros – "When will I be famous?"

ESPN 100 most famous athletes from 2017

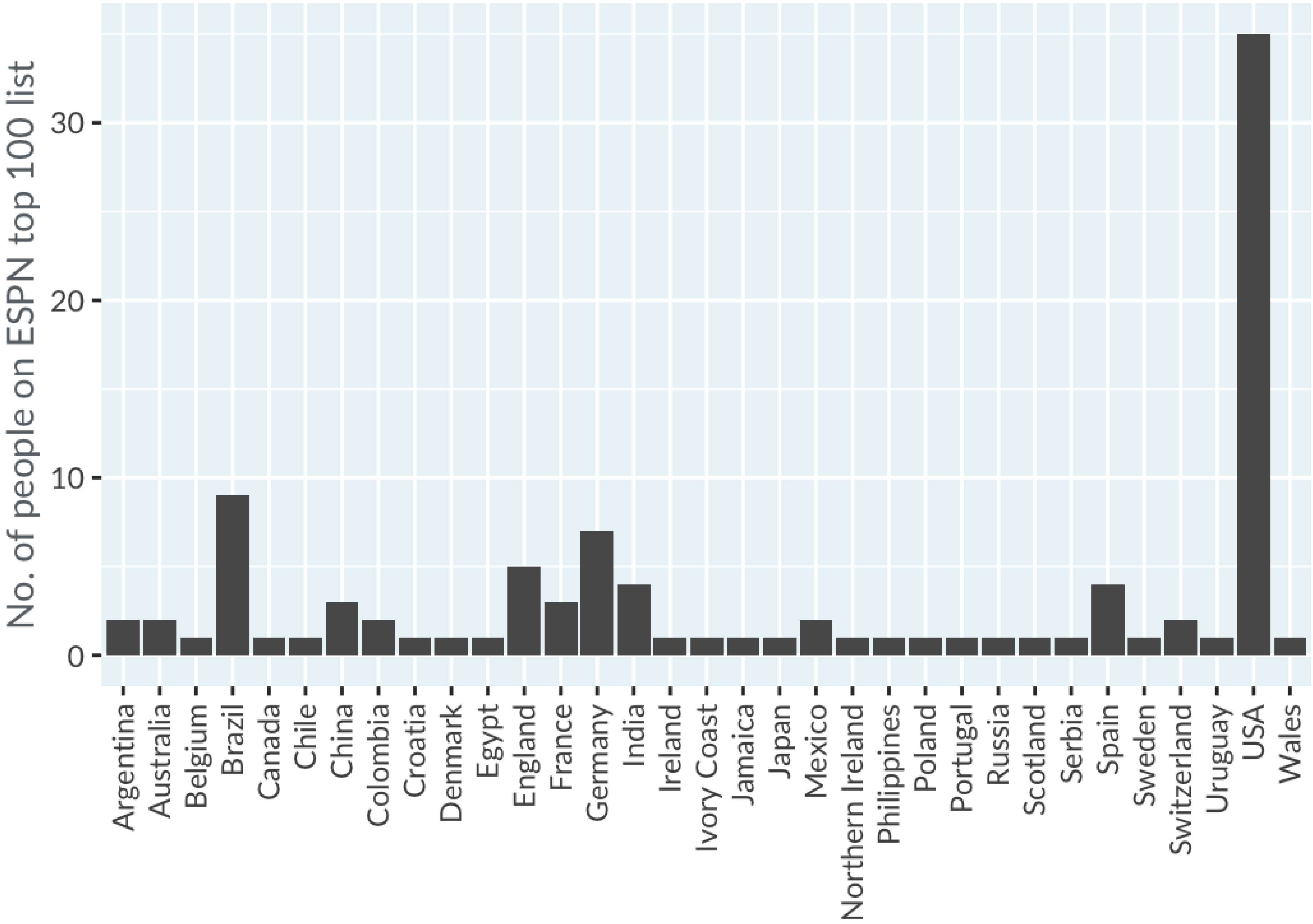
Rank	Last Name	First Name	Sport	Country
1	Ronaldo	Cristiano	Soccer	Portugal
2	James	LeBron	Basketball	USA
3	Messi	Lionel	Soccer	Argentina
4	Federer	Roger	Tennis	Switzerland
5	Mickelson	Phil	Golf	USA
...	...	...	...	...

[1] https://data.world/wesmagee/2017-espn-world-fame-100

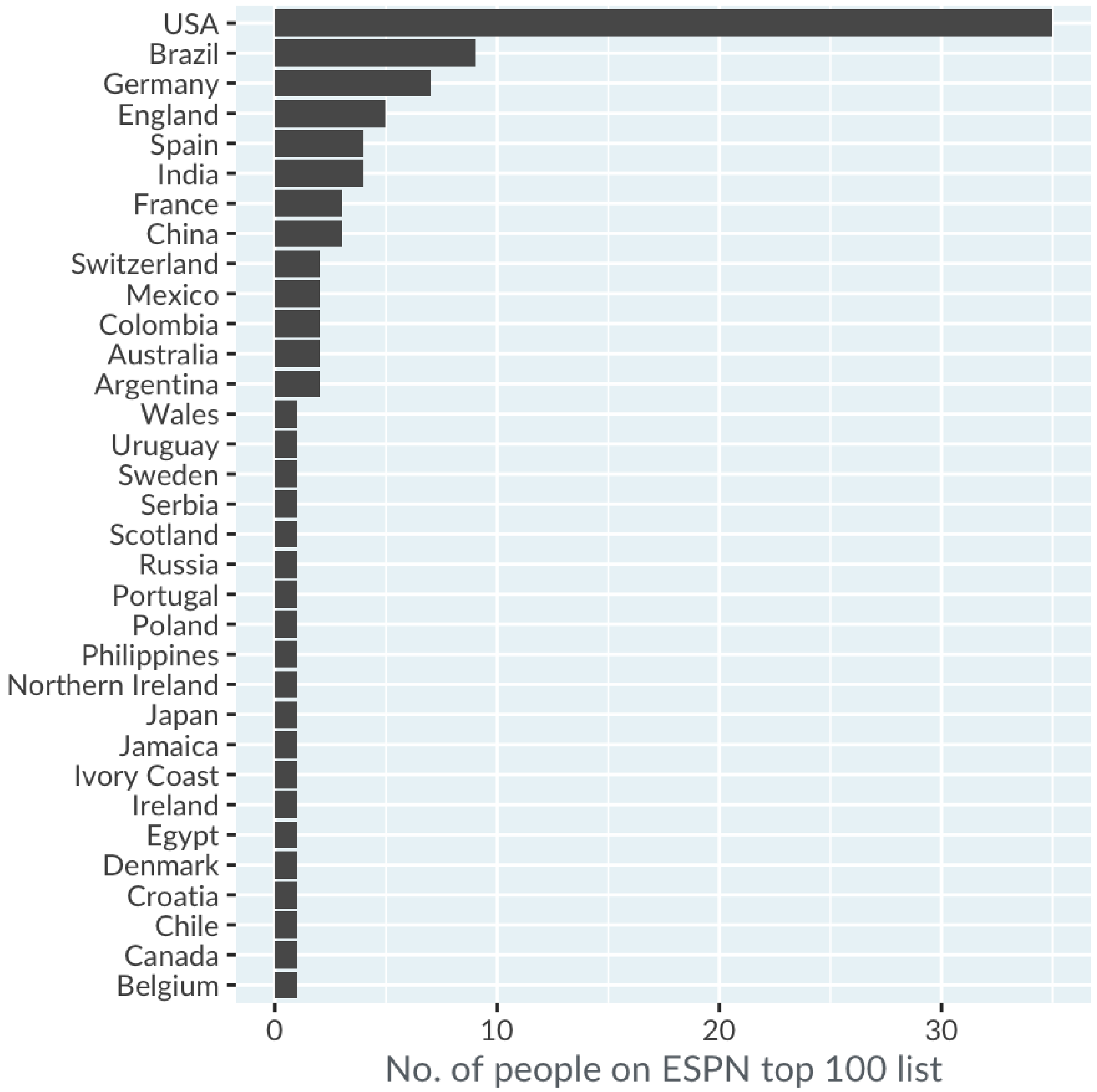
Bar plot of counts by country



Vertical bars



Sorting by count

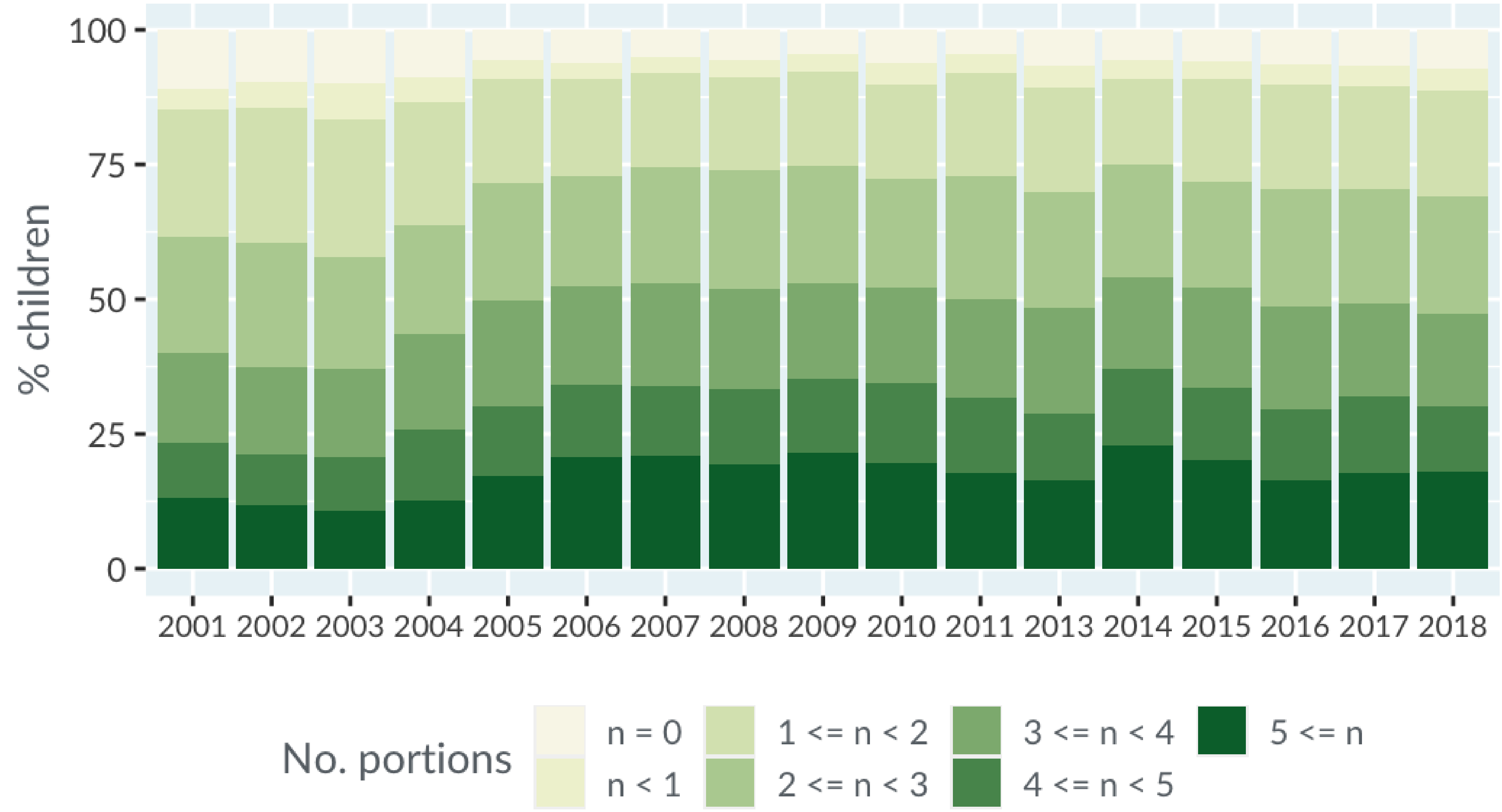


Children's fruit and veg consumption

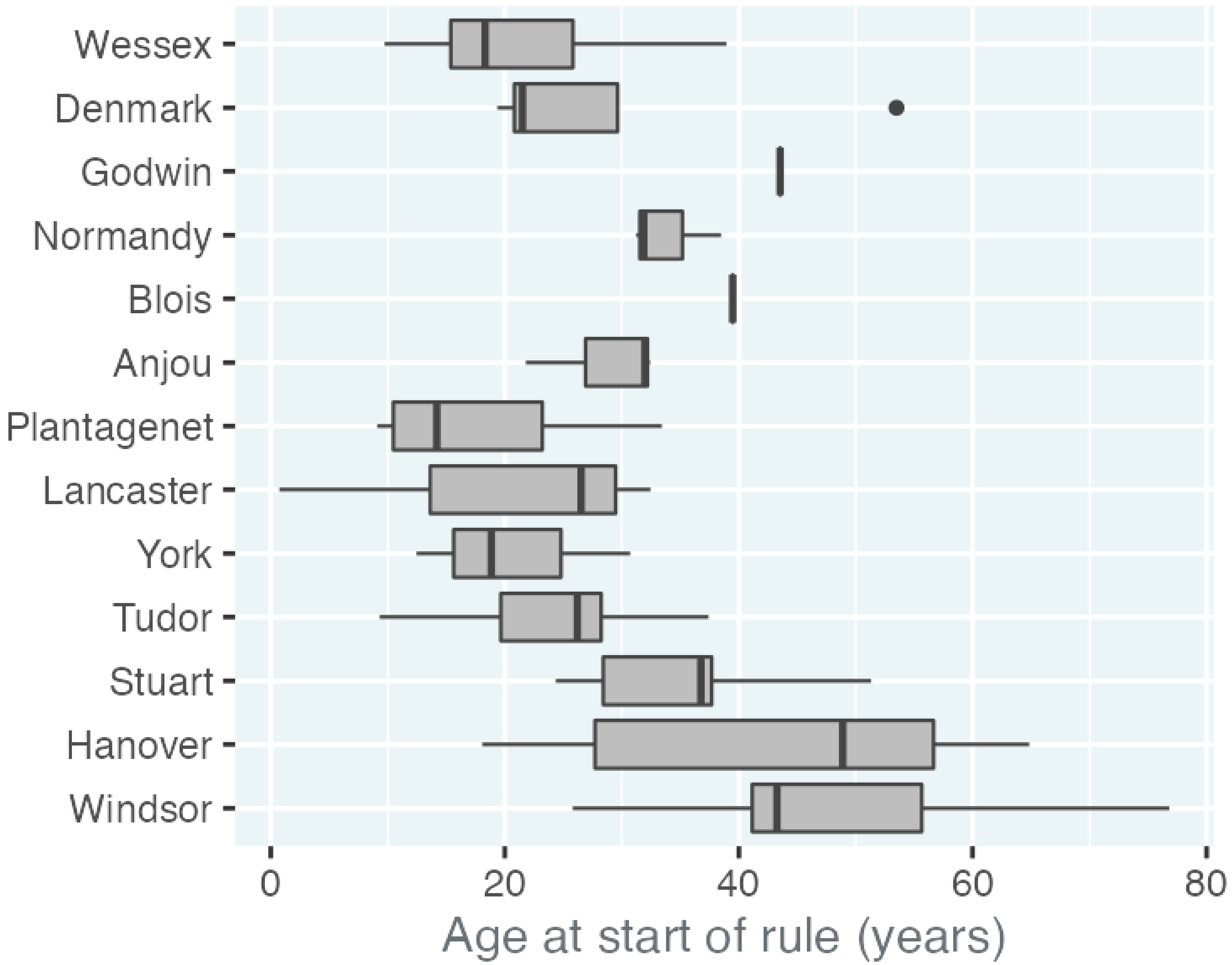
n_portions	year	pct_children
n = 0	2001	10.921779
n < 1	2001	3.843093
1 <= n < 2	2001	23.659102
...	...	...
4 <= n < 5	2018	12.28728
5 <= n	2018	17.87497

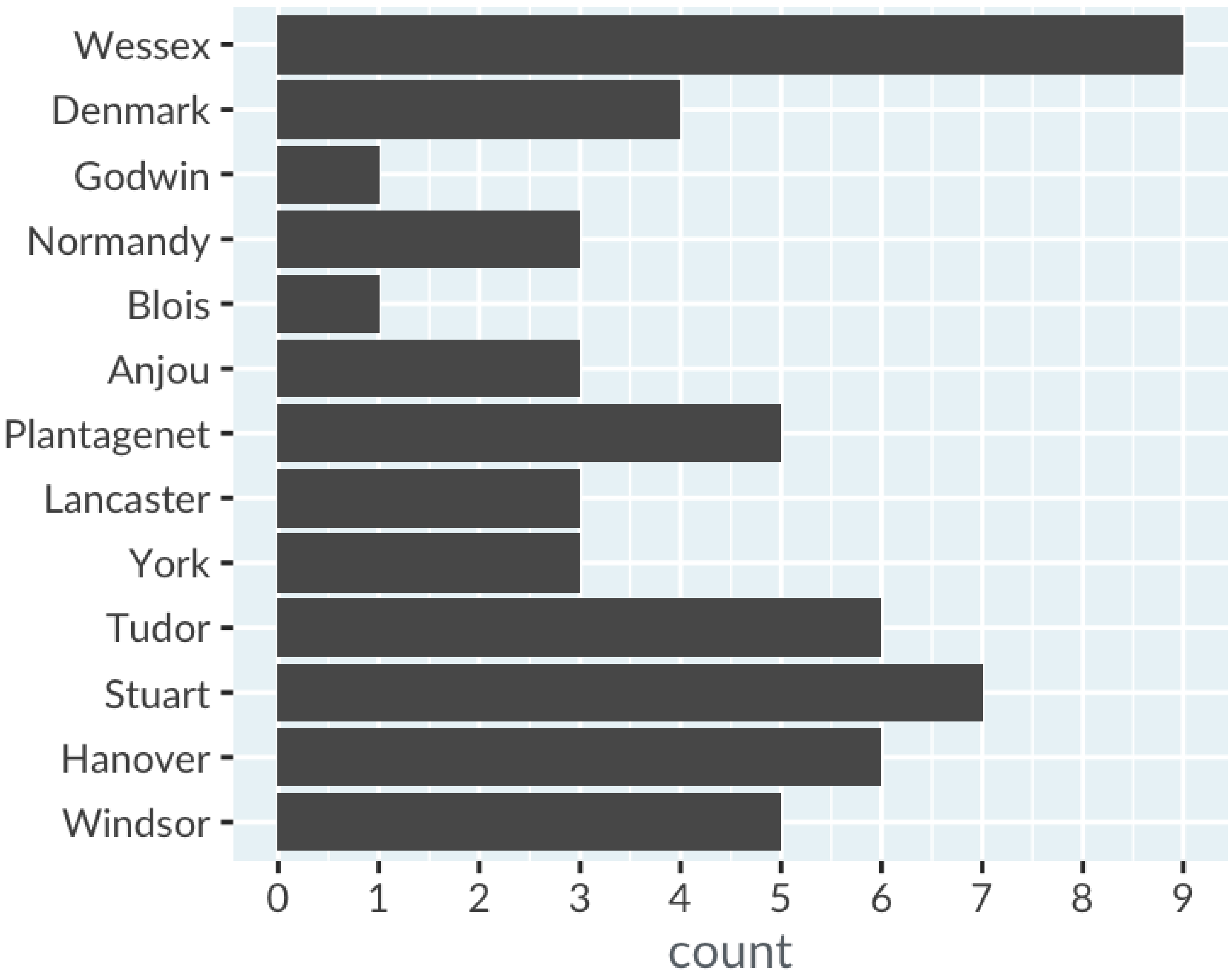
[1] https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2018/health-survey-for-england-2018-data-tables

Stacking bars

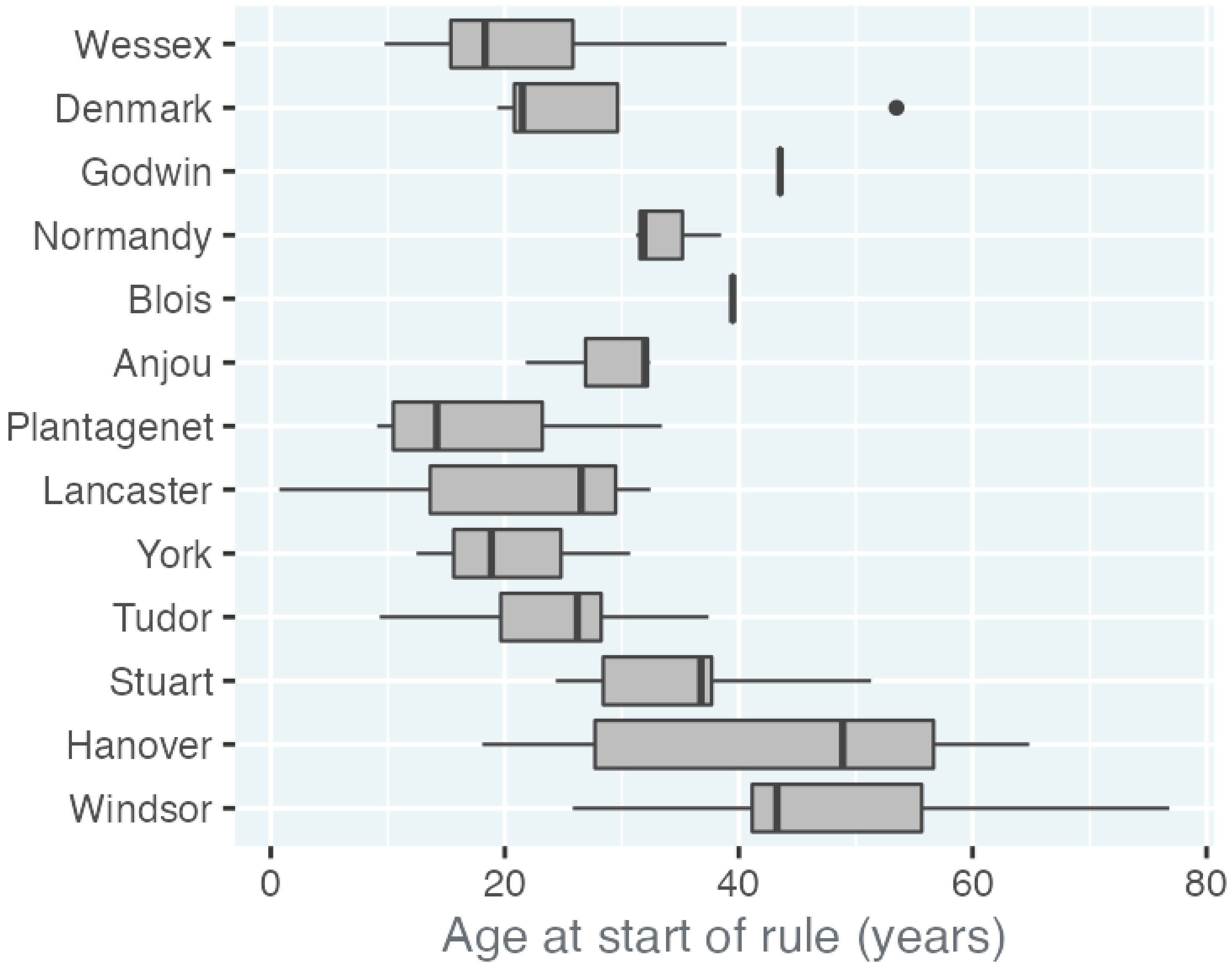


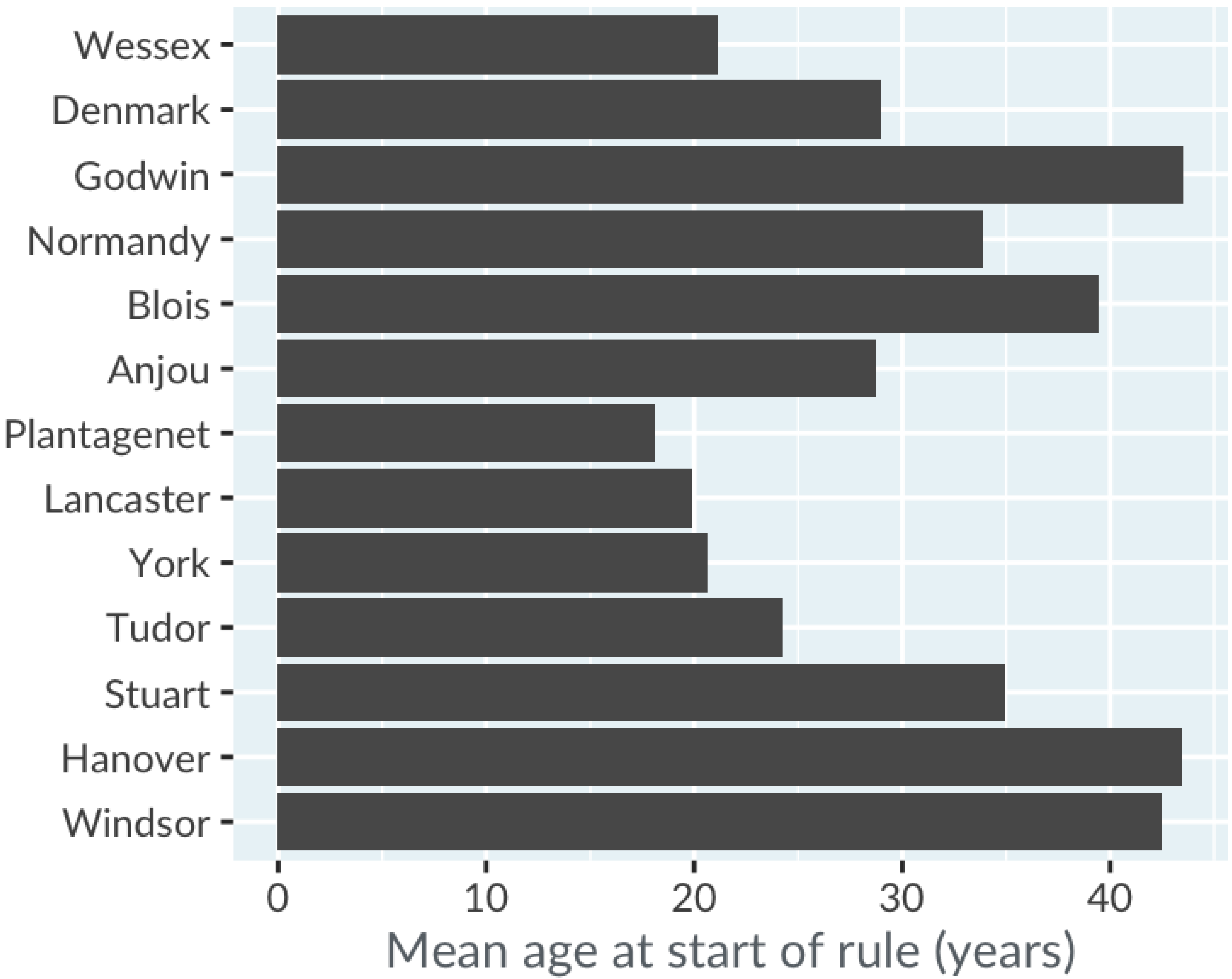
Bar plots vs. box plots





Other metrics than counts





Let's practice!

UNDERSTANDING DATA VISUALIZATION

Dot plots

UNDERSTANDING DATA VISUALIZATION

Richie Cotton

Data Evangelist at DataCamp

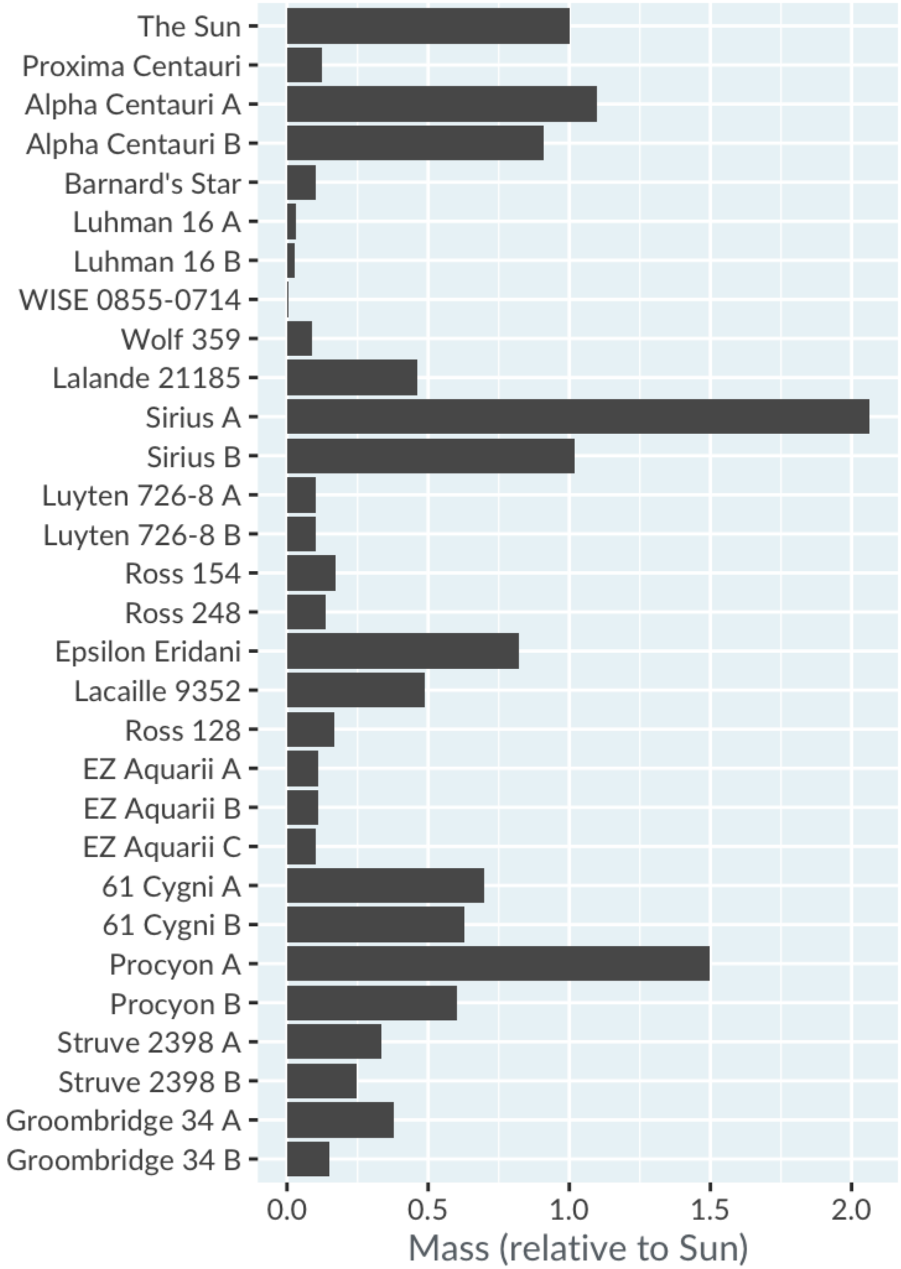
When should you use a dot plot?

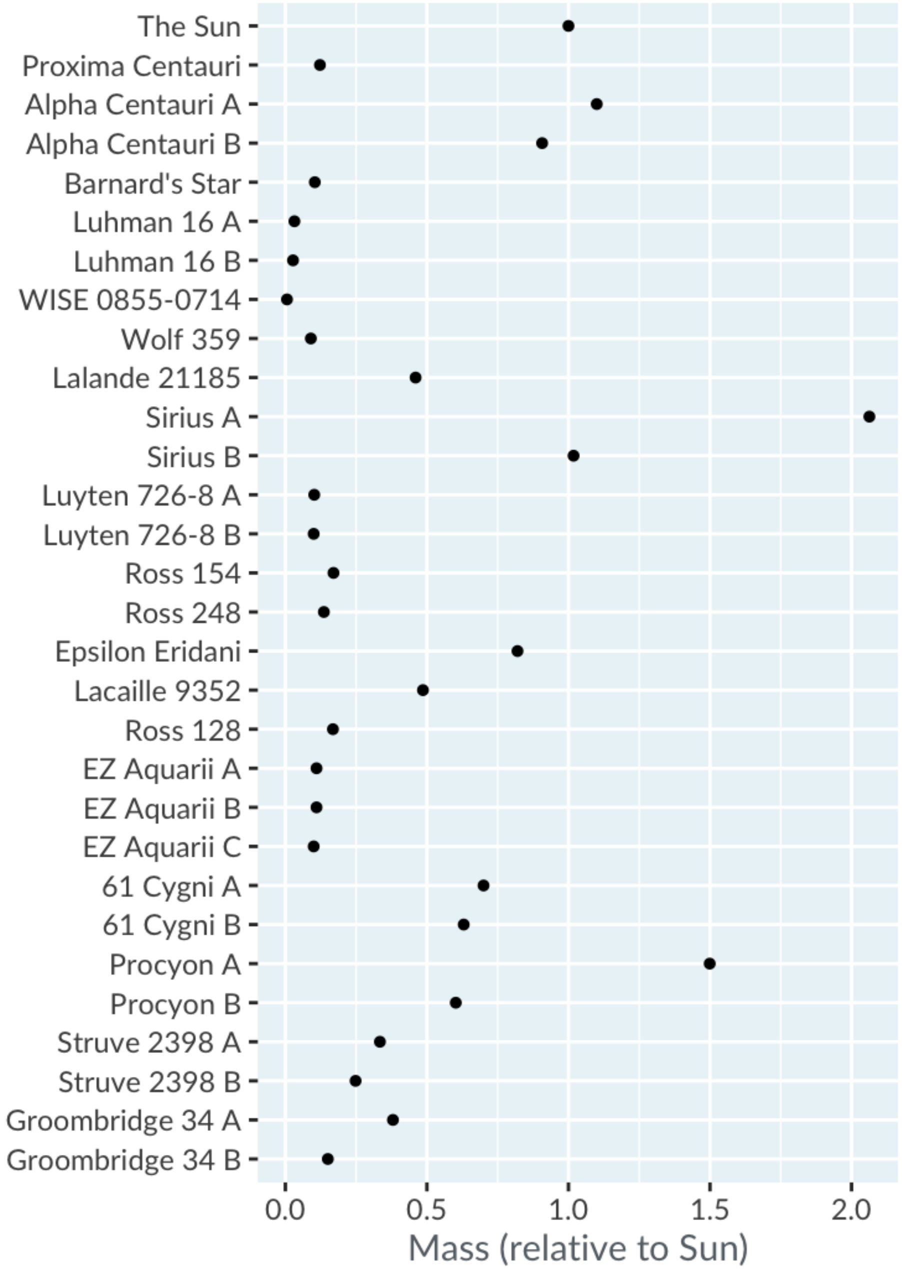
1. You have a categorical variable.
2. You want to display numeric scores for each category on a log scale, or
3. You want to display multiple numeric scores for each category.

Nearby stars and brown dwarfs

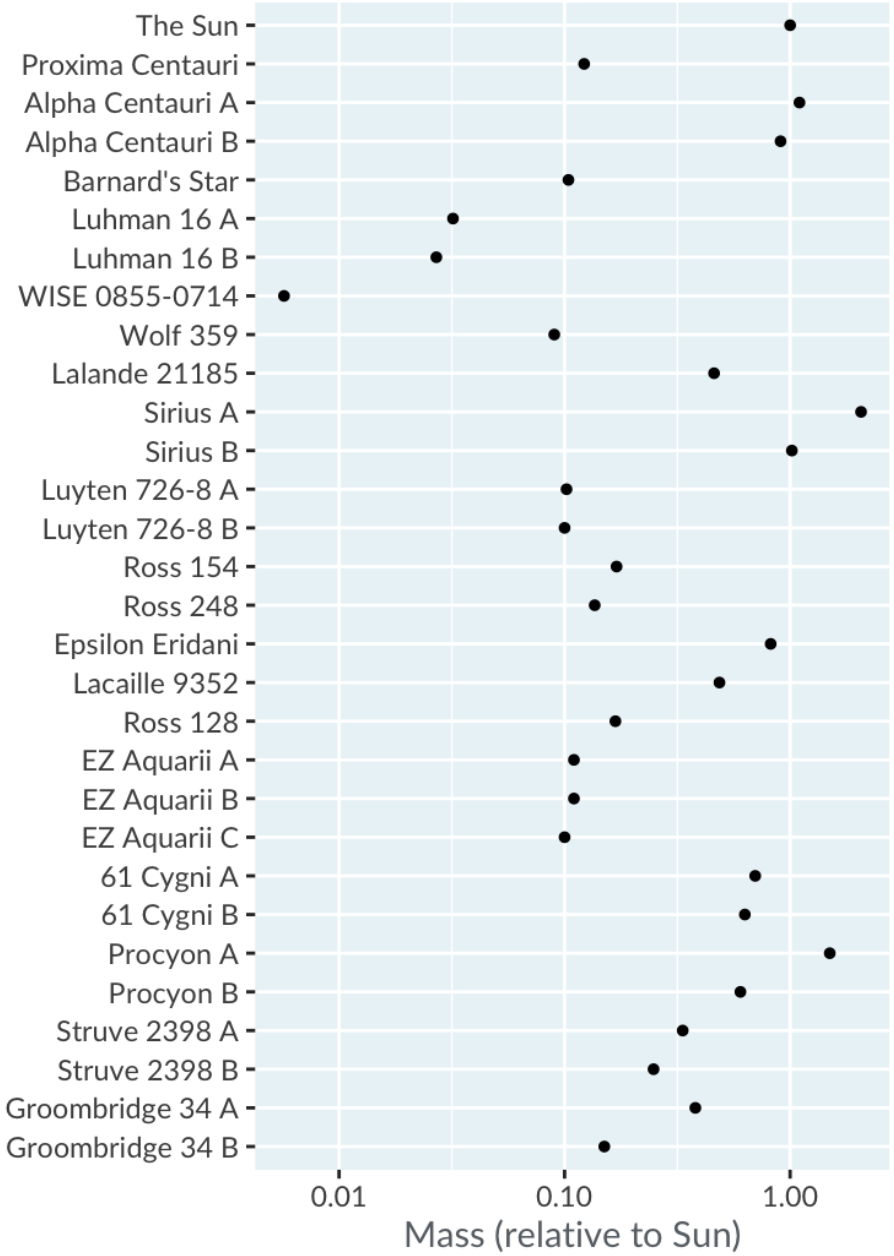
star	distance_ly	mass_sm
The Sun	0.0000158	1.0000
Proxima Centauri	4.2441000	0.1221
Alpha Centauri A	4.3650000	1.1000
Alpha Centauri B	4.3650000	0.9070
Barnard's Star	5.9577000	0.1040
Luhman 16 A	6.5029000	0.0320
...	...	...

Bar plot vs. dot plot

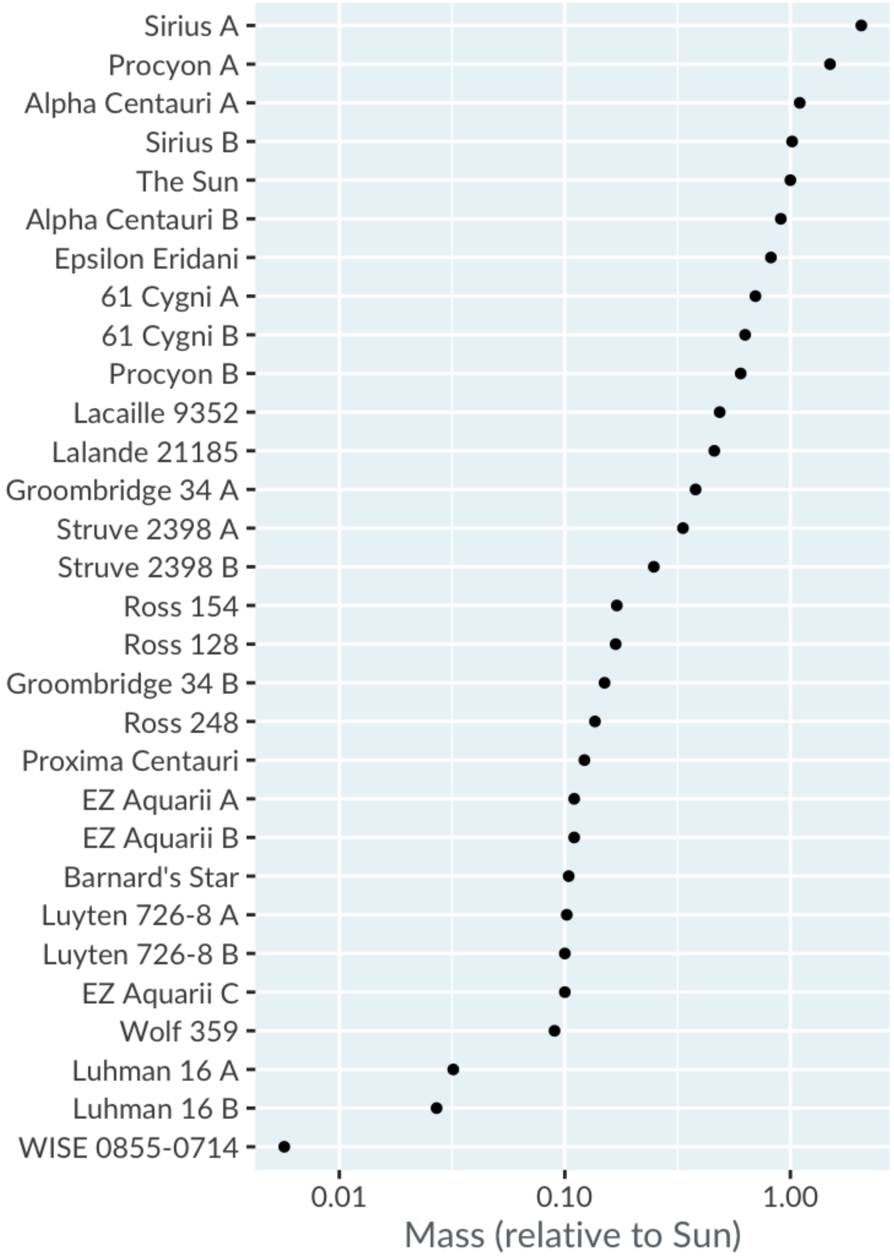




Log scales


The Sun
Proxima Centauri
Alpha Centauri A
Alpha Centauri B
Barnard's Star
Luhman 16 A
Luhman 16 B
WISE 0855-0714
Wolf 359
Lalande 21185
Sirius A
Sirius B
Luyten 726-8 A
Luyten 726-8 B
Ross 154
Ross 248
Epsilon Eridani
Lacaille 9352
Ross 128
EZ Aquarii A
EZ Aquarii B
EZ Aquarii C
61 Cygni A
61 Cygni B
Procyon A
Procyon B
Struve 2398 A
Struve 2398 B
Groombridge 34 A
Groombridge 34 B
0.01
0.10
1.00
Mass (relative to Sun)

Sorting rows



Let's practice!

UNDERSTANDING DATA VISUALIZATION