Best practices: bar plots

INTERMEDIATE DATA VISUALIZATION WITH GGPLOT2

Rick Scavetta
Founder, Scavetta Academy

In this chapter

- Common pitfalls in Data Viz
- Best way to represent data
  - For effective explanatory (communication), and
  - For effective exploratory (investigation) plots

Bar plots

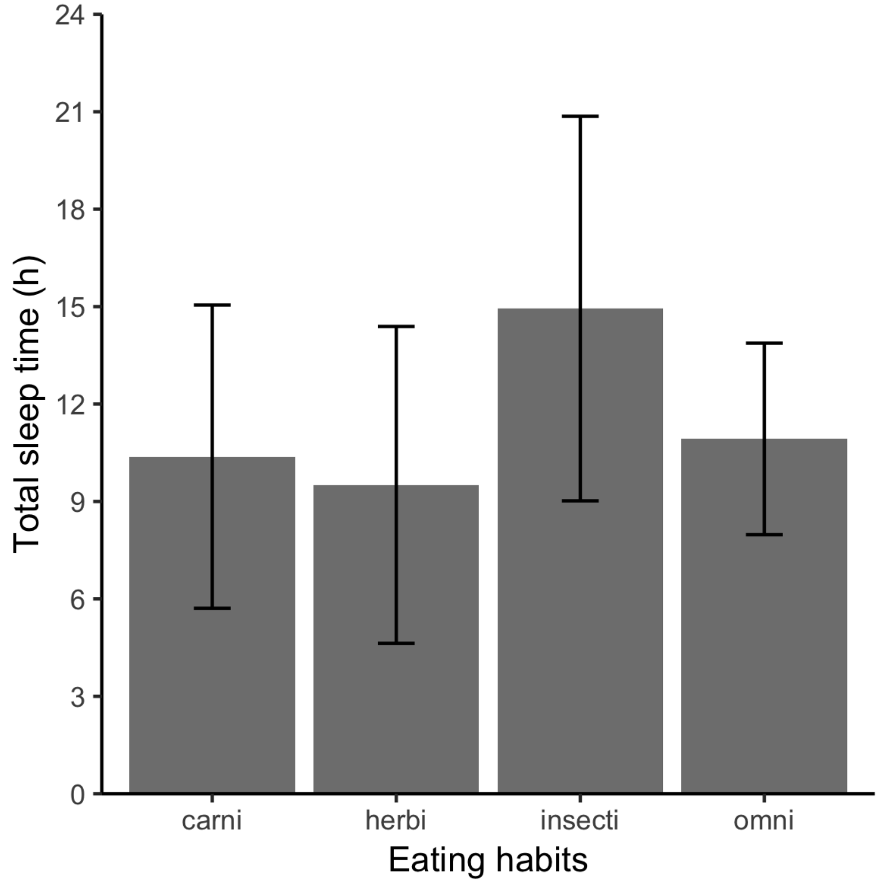
- Two types
  - Absolute values
  - Distributions

Mammalian sleep

```
Observations: 76
Variables: 3
$ vore  <chr> "carni", "omni", "herbi", "omni", "herbi", "herbi", "carni", …
$ total <dbl> 12.1, 17.0, 14.4, 14.9, 4.0, 14.4, 8.7, 10.1, 3.0, 5.3, 9.4, …
$ rem   <dbl> NA, 1.8, 2.4, 2.3, 0.7, 2.2, 1.4, 2.9, NA, 0.6, 0.8, 0.7, 1.5…
```

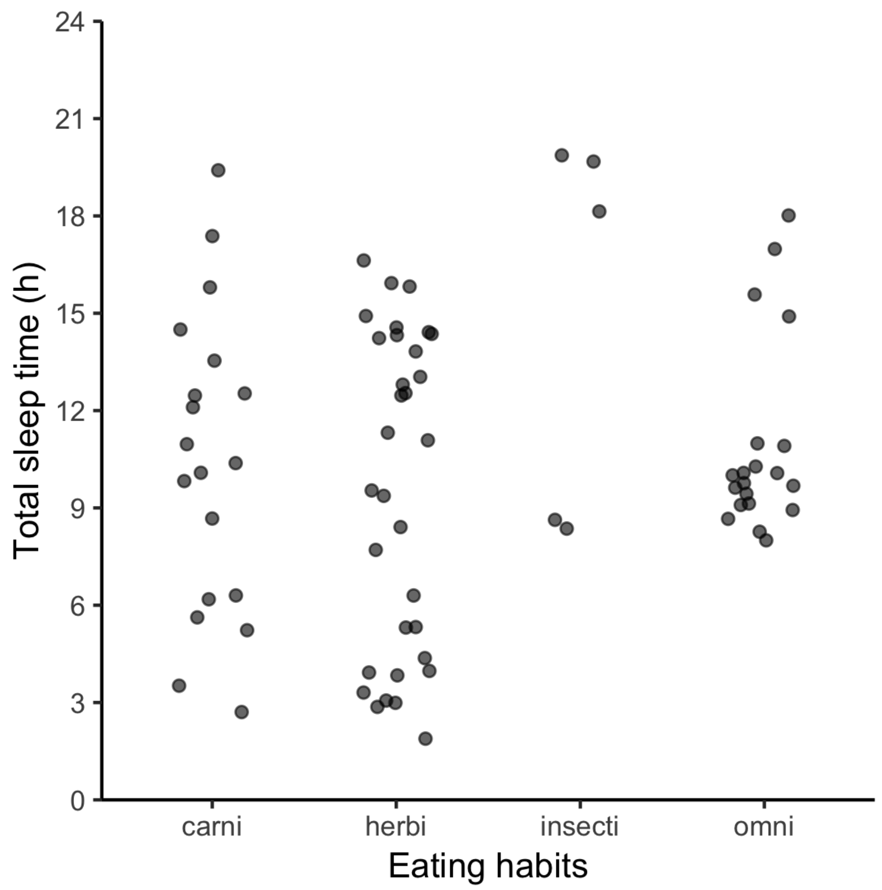
Dynamite plot

```r
d <- ggplot(sleep, aes(vore, total)) +
 # ...

d +
  stat_summary(fun.y = mean,
               geom = "bar",
               fill = "grey50") +
  stat_summary(fun.data = mean_sdl,
               fun.args = list(mult = 1),
               geom = "errorbar",
               width = 0.2)
```



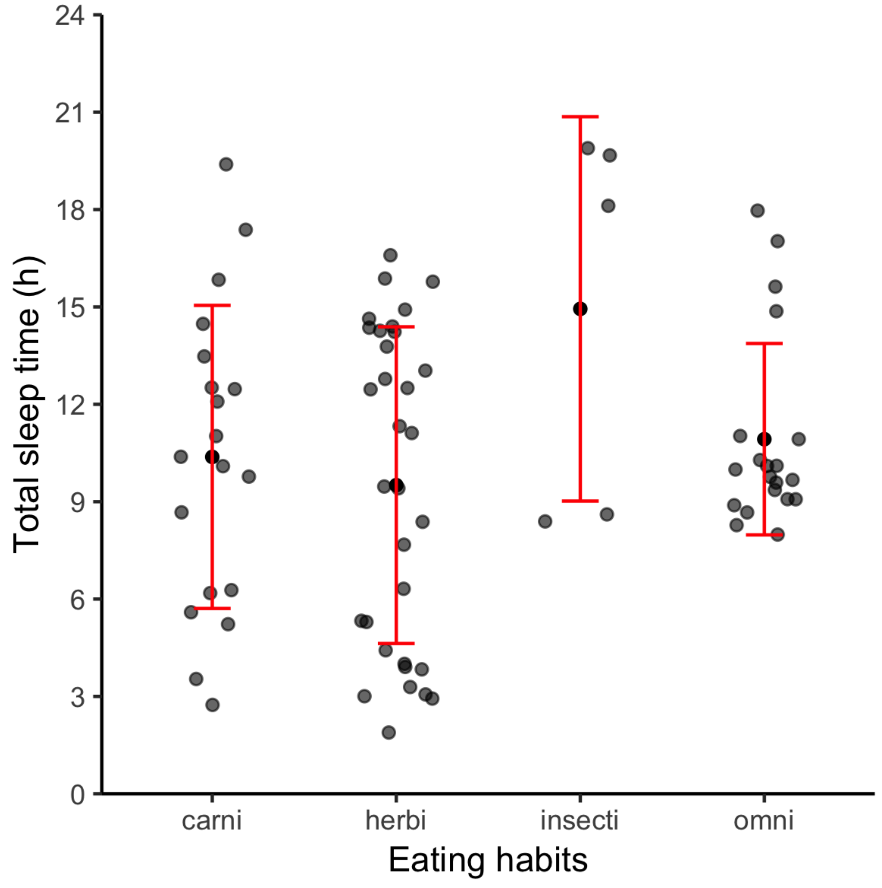
Individual data points

```
# position
posn_j <- position_jitter(width = 0.2)

# plot
d +
  geom_point(alpha = 0.6,
             position = posn_j)
```



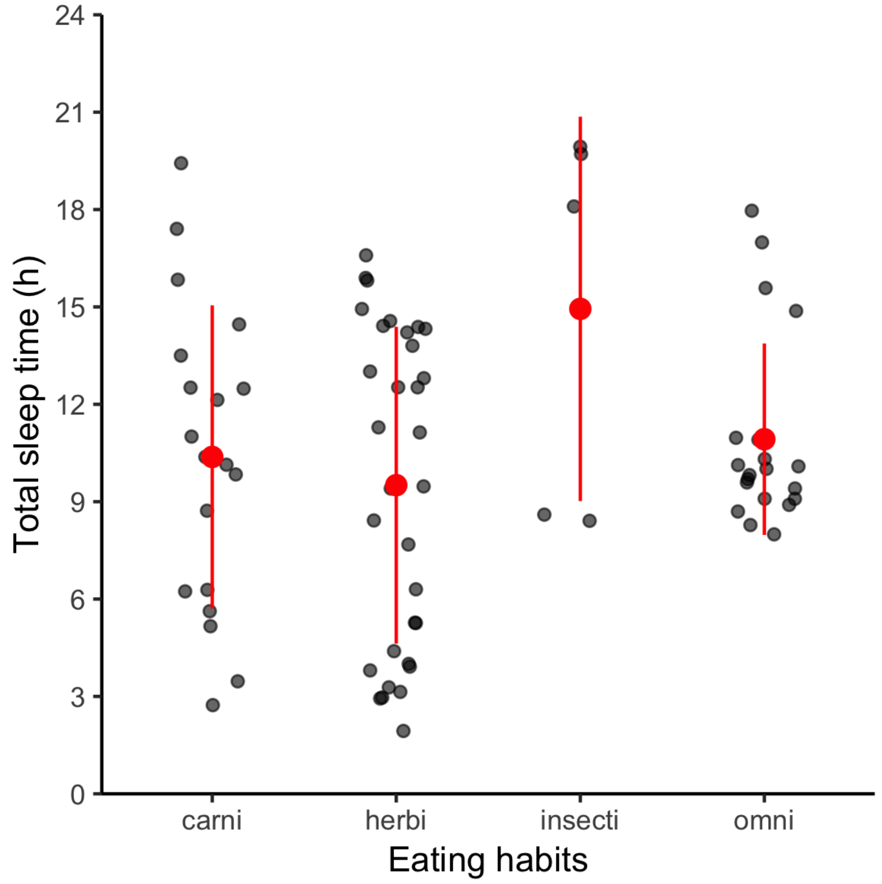
geom_errorbar()

```
d +
  geom_point(...) +
  stat_summary(fun.y = mean,
               geom = "point",
               fill = "red") +
  stat_summary(fun.data = mean_sdl,
               fun.args = list(mult = 1),
               geom = "errorbar",
               width = 0.2,
               color = "red")
```



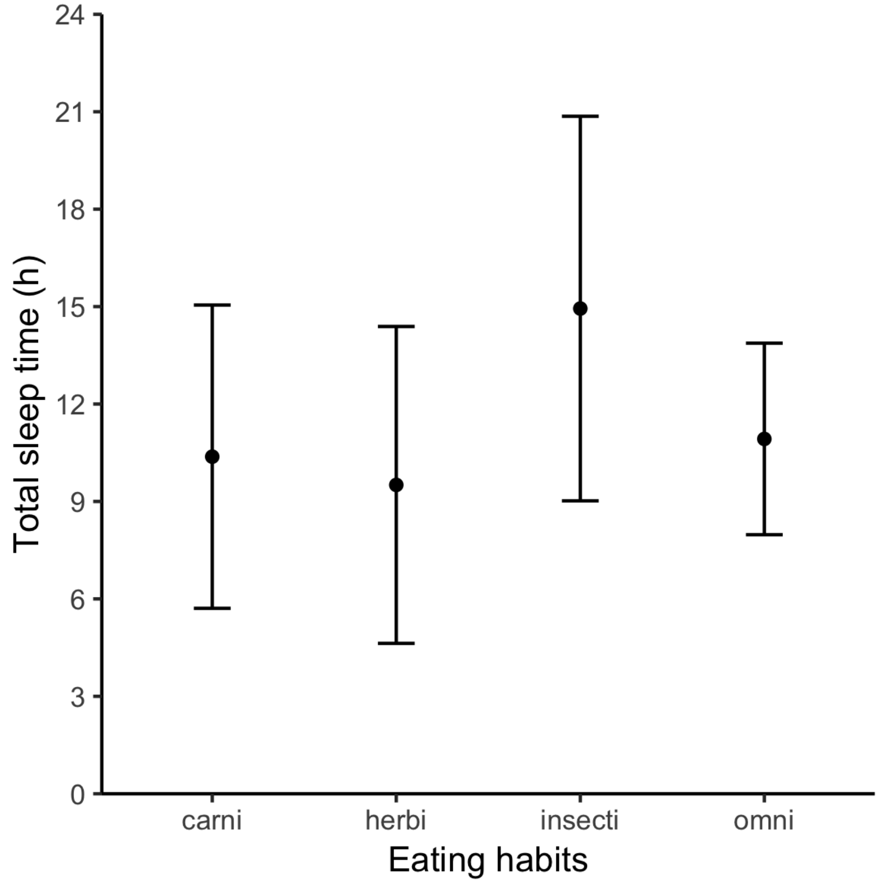
geom_pointrange()

```
d +
  geom_point(...) +
  stat_summary(fun.data = mean_sdl,
               mult = 1,
               width = 0.2,
               color = "red")
```

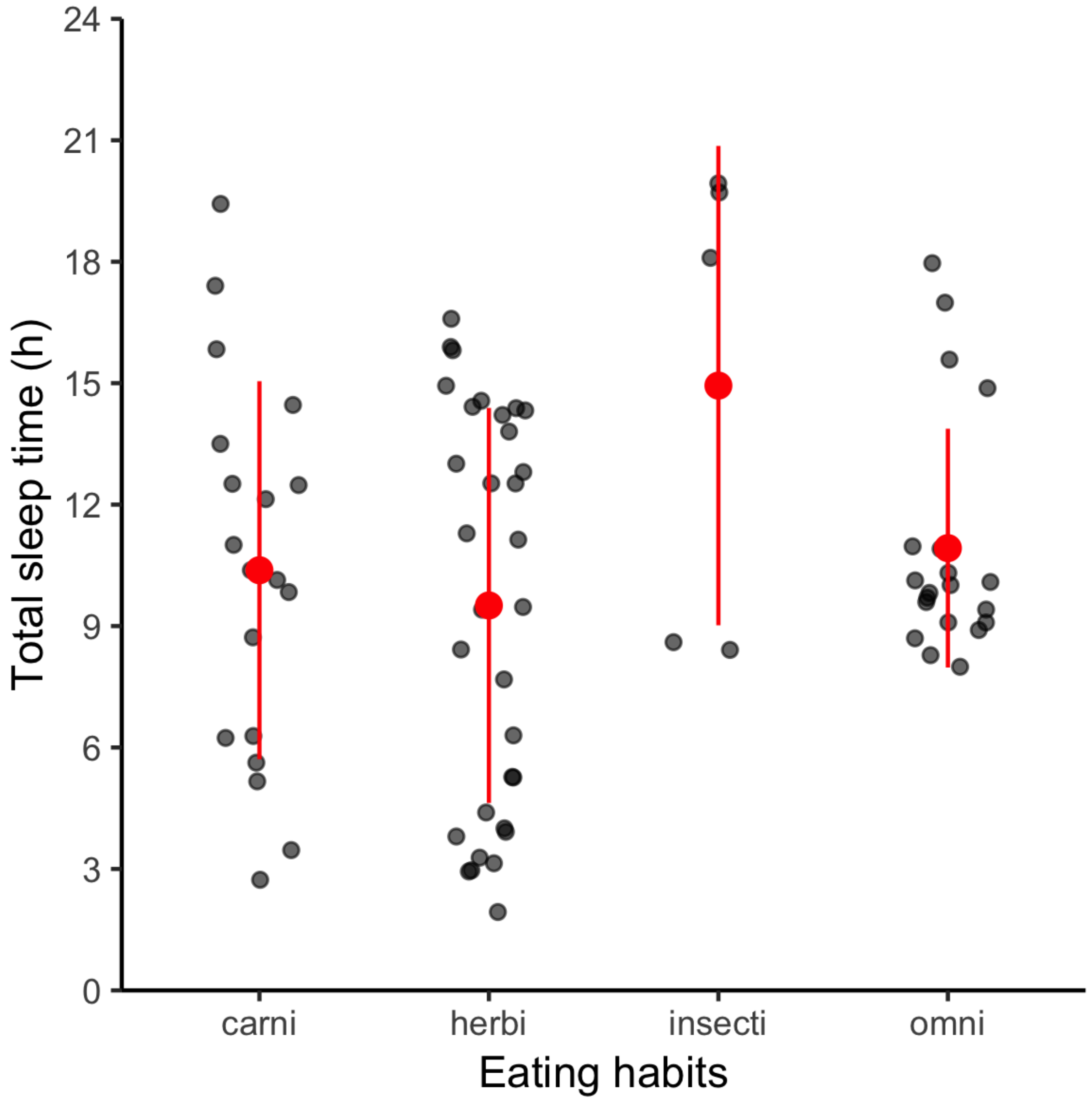


Without data points

```
d +
  stat_summary(fun.y = mean,
               geom = "point") +
  stat_summary(fun.data = mean_sdl,
               fun.args = list(mult = 1),
               geom = "errorbar",
               width = 0.2)
```



Bars are not necessary



Ready for exercises!

INTERMEDIATE DATA VISUALIZATION WITH GGPLOT2

Heatmaps use case scenario

INTERMEDIATE DATA VISUALIZATION WITH GGPLOT2

Rick Scavetta
Founder, Scavetta Academy

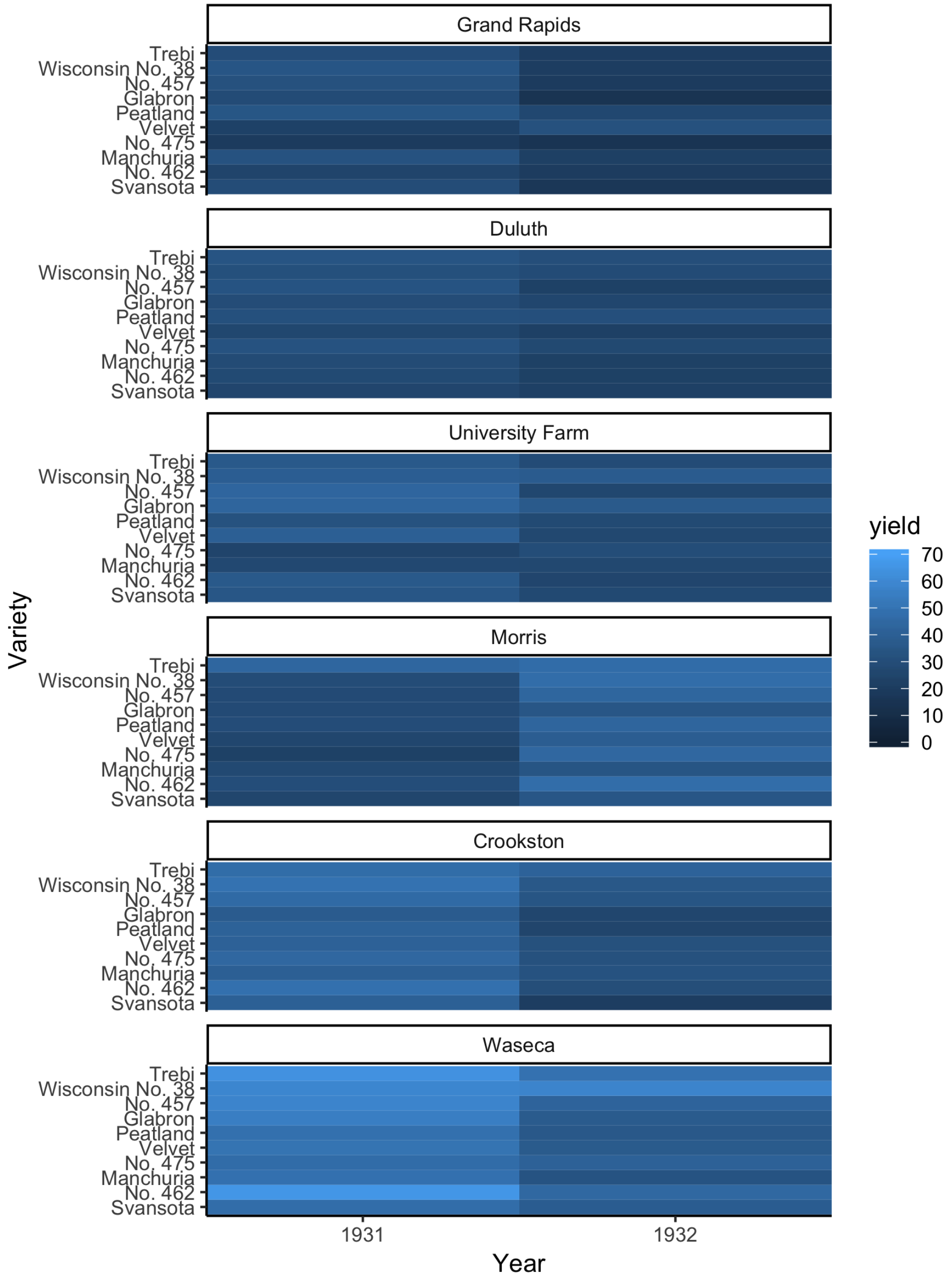
The barley dataset

```r
head(barley, 9)
```

```
     yield   variety year            site
1 27.00000 Manchuria 1931 University Farm
2 48.86667 Manchuria 1931          Waseca
3 27.43334 Manchuria 1931          Morris
4 39.93333 Manchuria 1931       Crookston
5 32.96667 Manchuria 1931    Grand Rapids
6 28.96667 Manchuria 1931          Duluth
7 43.06666   Glabron 1931 University Farm
8 55.20000   Glabron 1931          Waseca
9 28.76667   Glabron 1931          Morris
```

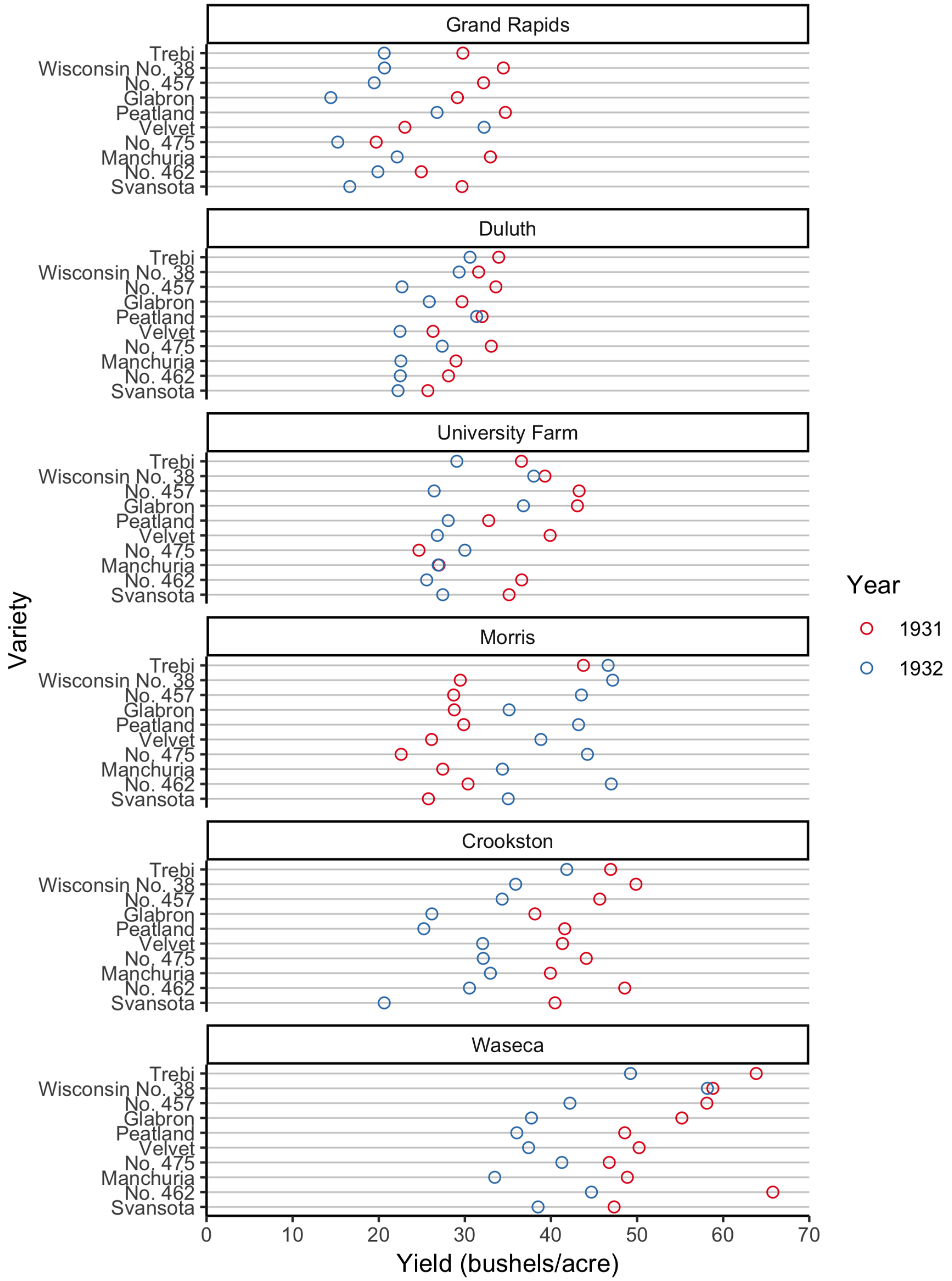
A basic heat map

```
ggplot(barley, aes(year, variety,
                   fill = yield)) +
  geom_tile() +
  facet_wrap(vars(site), ncol = 1) +
  ...
```



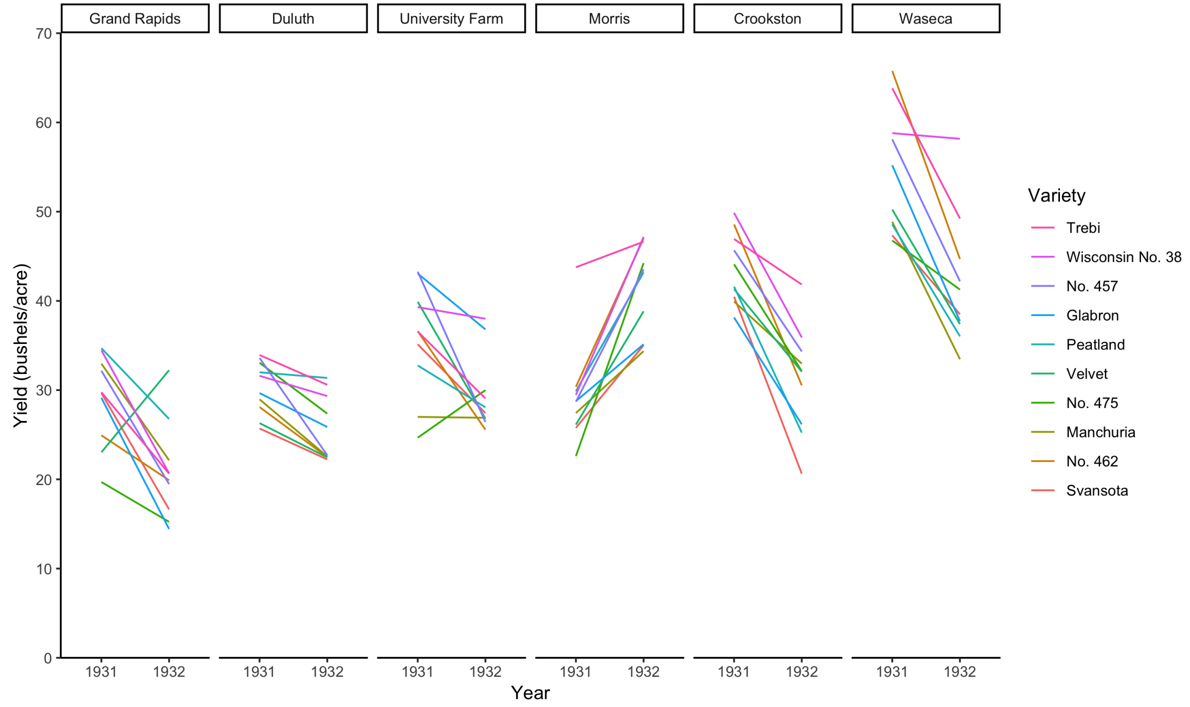
A dot plot

```
ggplot(barley, aes(yield, variety,
                   color = year)) +
  geom_point(...) +
  facet_wrap(vars(site), ncol = 1) +
  ...
```



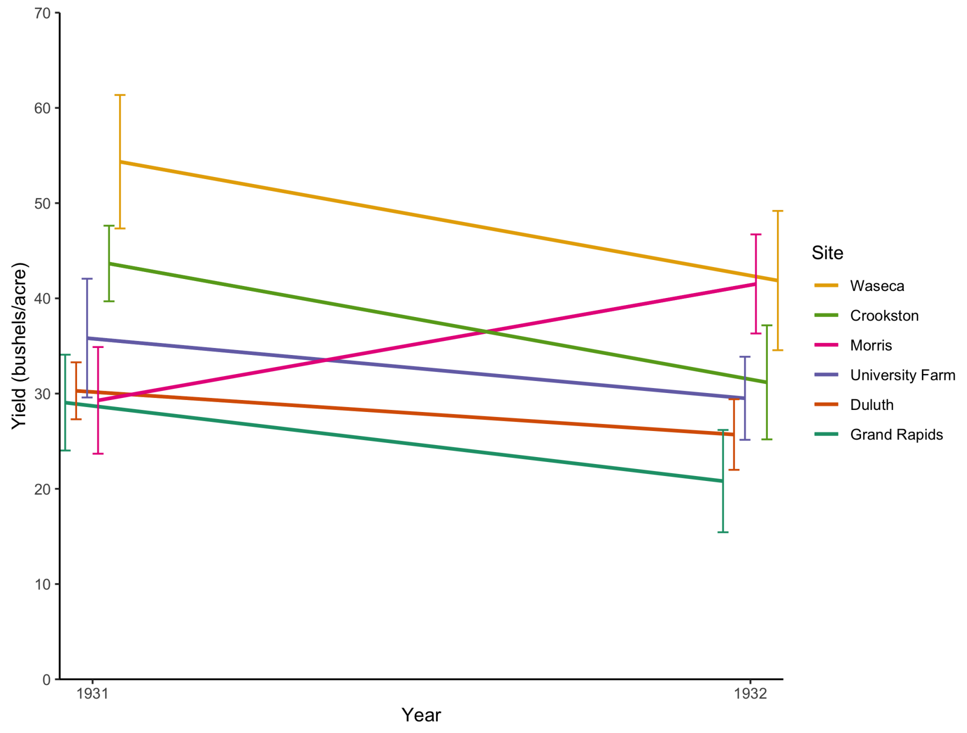
As a time series

```
ggplot(barley, aes(year, yield,
                   group = variety,
                   color = variety)) +
  geom_line() +
  facet_wrap(vars(site), nrow = 1) +
  ...
```



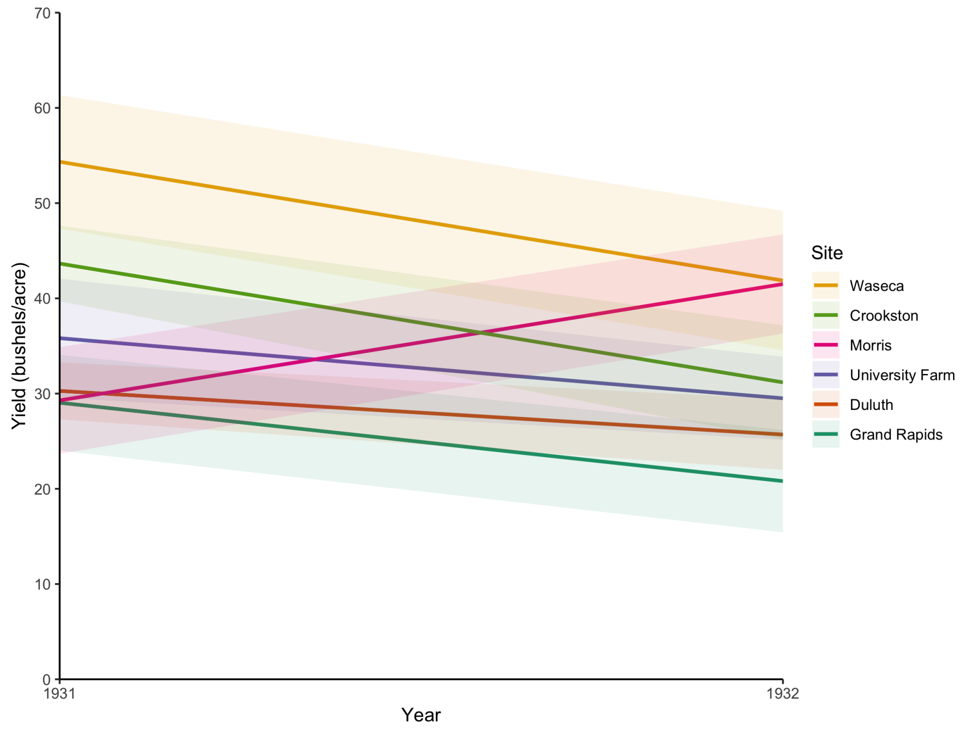
Using dodged error bars

```
ggplot(barley, aes(x = year, y = yield,
                   group = site,
                   color = site)) +
  stat_summary(fun.y = mean,
               geom = "line", ...) +
  stat_summary(fun.data = mean_sdl,
               geom = "errorbar", ...) +
  ...
```



Using ribbons for error

```
ggplot(barley, aes(x = year, y = yield,
                   group = site,
                   color = site)) +
  stat_summary(fun.y = mean,
               geom = "line", ...) +
  stat_summary(fun.data = mean_sdl,
               geom = "ribbon", ...) +
  ...
```



Coding Time!

INTERMEDIATE DATA VISUALIZATION WITH GGPLOT2

When good data makes bad plots

INTERMEDIATE DATA VISUALIZATION WITH GGPLOT2

Rick Scavetta
Founder, Scavetta Academy

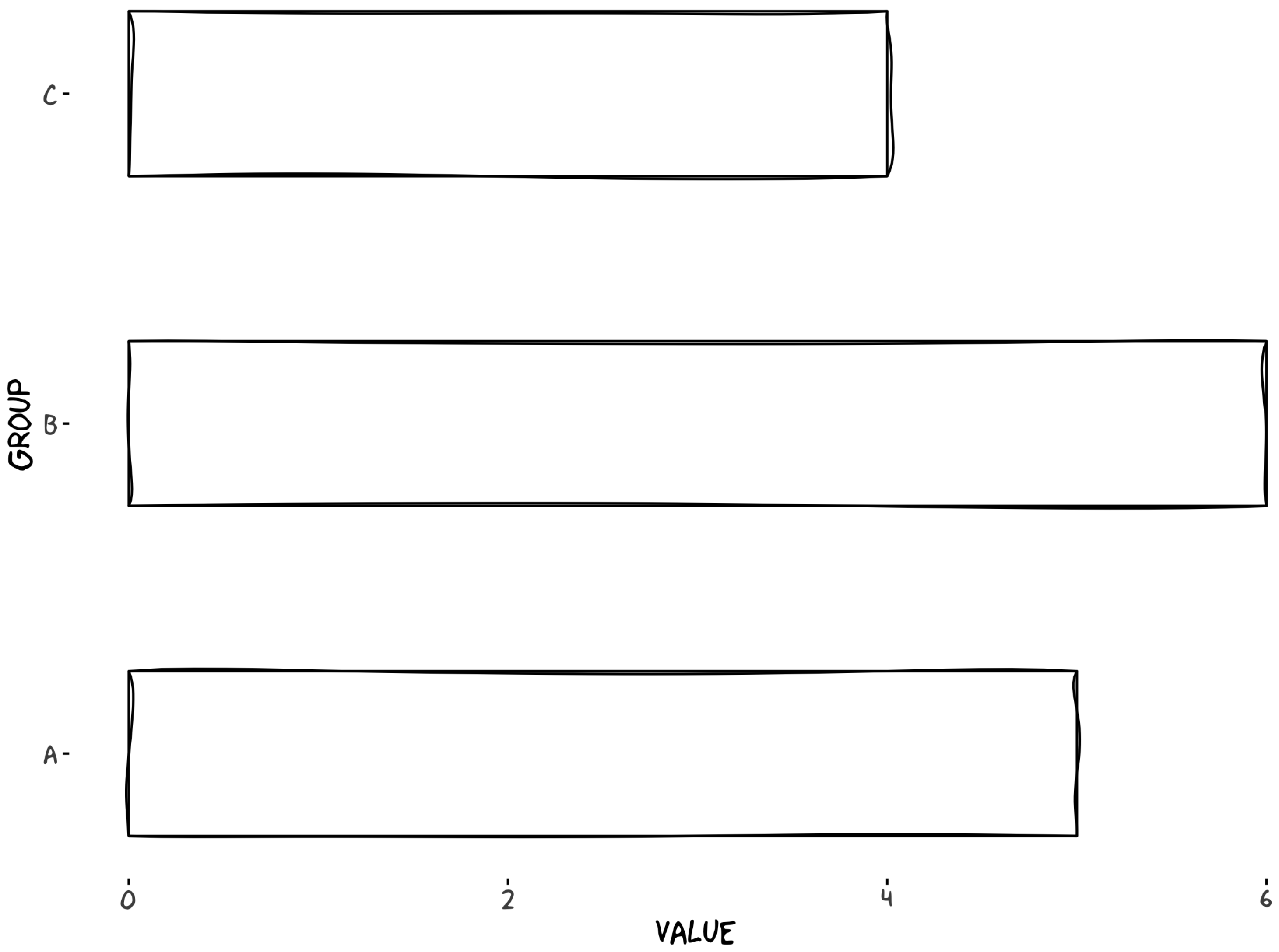
Bad plots: style

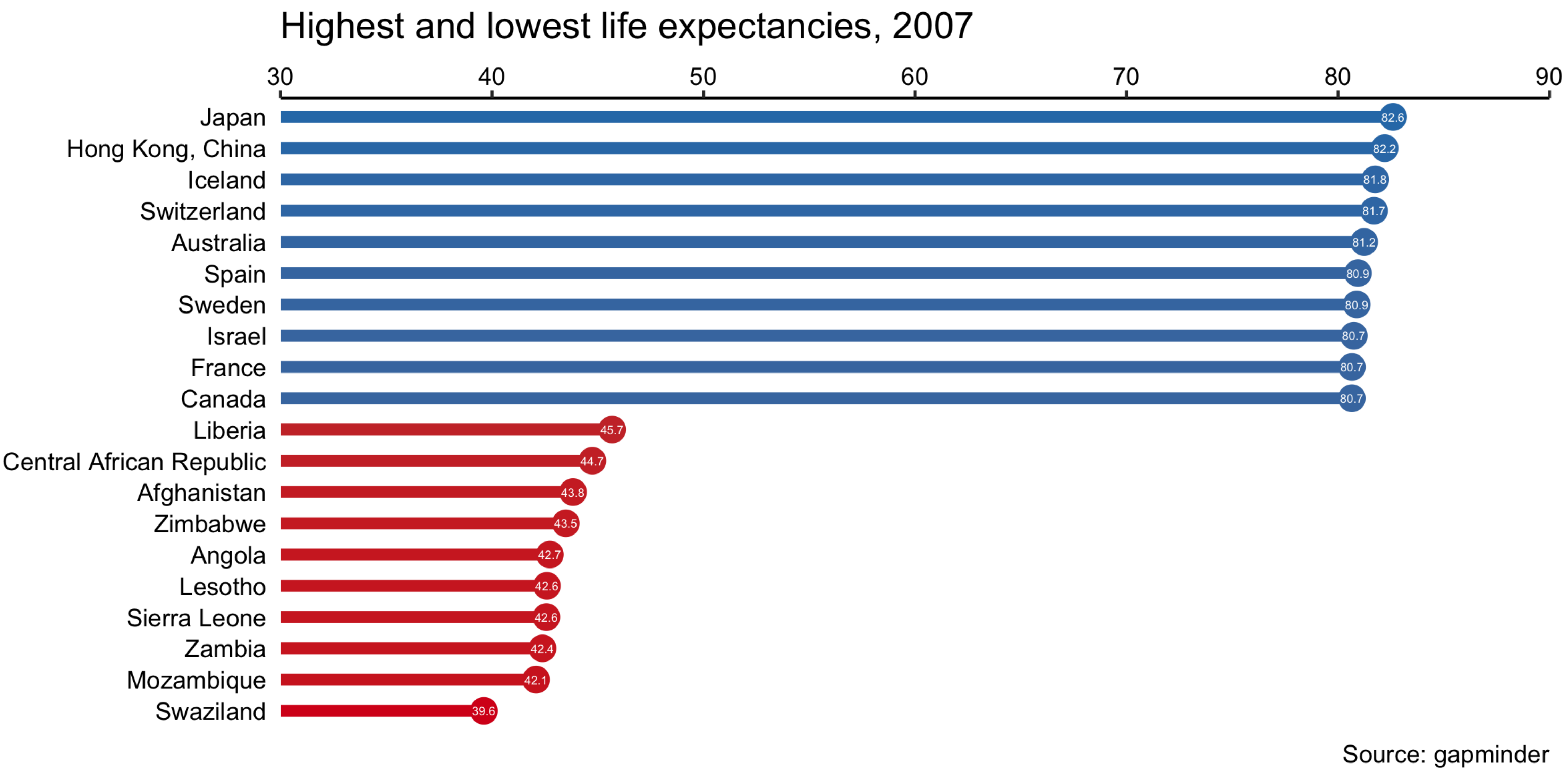
- Color
  - Not color-blind-friendly (e.g. primarily red and green)
  - Wrong palette for data type (remember sequential, qualitative and divergent)
  - Indistinguishable groups (i.e. colors are too similar)
  - Ugly (high saturation primary colors)
- Text
  - Illegible (e.g. too small, poor resolution)
  - Non-descriptive (e.g. "length" -- of what? which units?)
  - Missing
  - Inappropriate (e.g. comic sans)

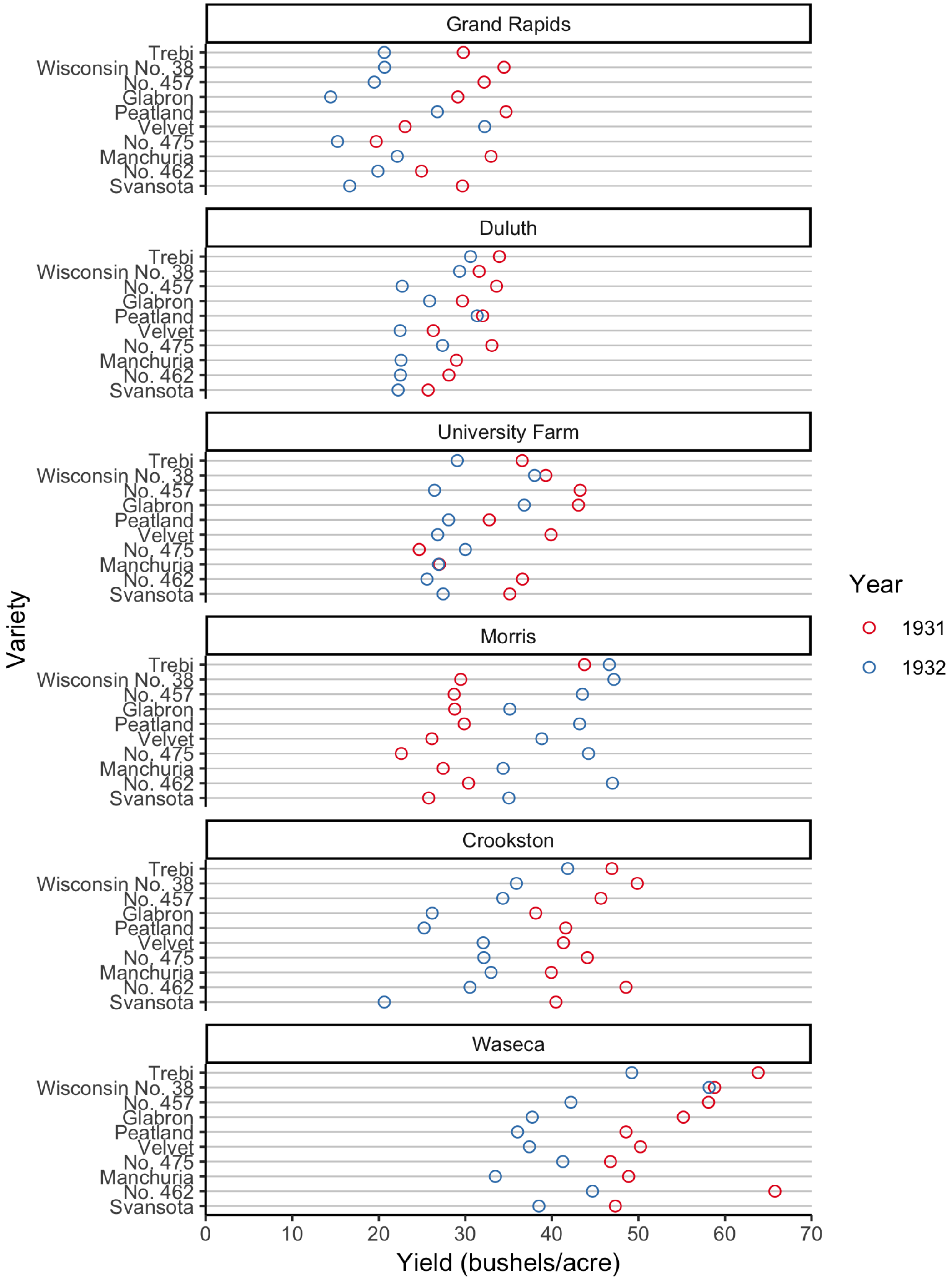
Bad plots: structure and content

- Information content
  - Too much information (TMI)
  - Too little information (TLI)
  - No clear message or purpose
- Axes
  - Poor aspect ratio
  - Suppression of the origin
  - Broken x or y axes
  - Common, but unaligned scales
  - Wrong or no transformation
- Statistics
  - Visualization doesn't match actual statistics
- Geometries
  - Wrong plot type
  - Wrong orientation
- Non-data Ink
  - Inappropriate use
- 3D plots
  - Perceptual problems
  - Useless 3rd axis

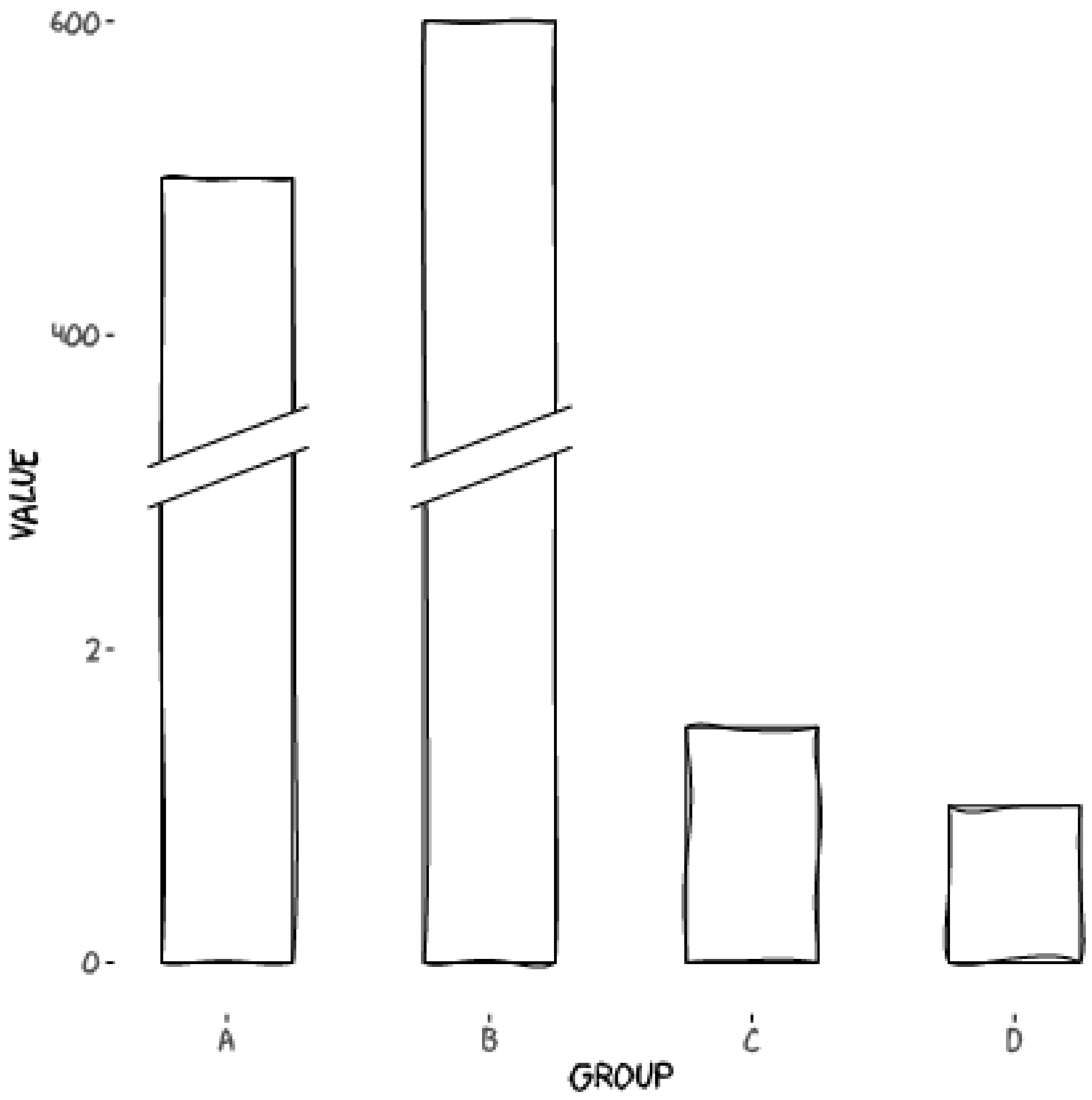
Wrong orientation



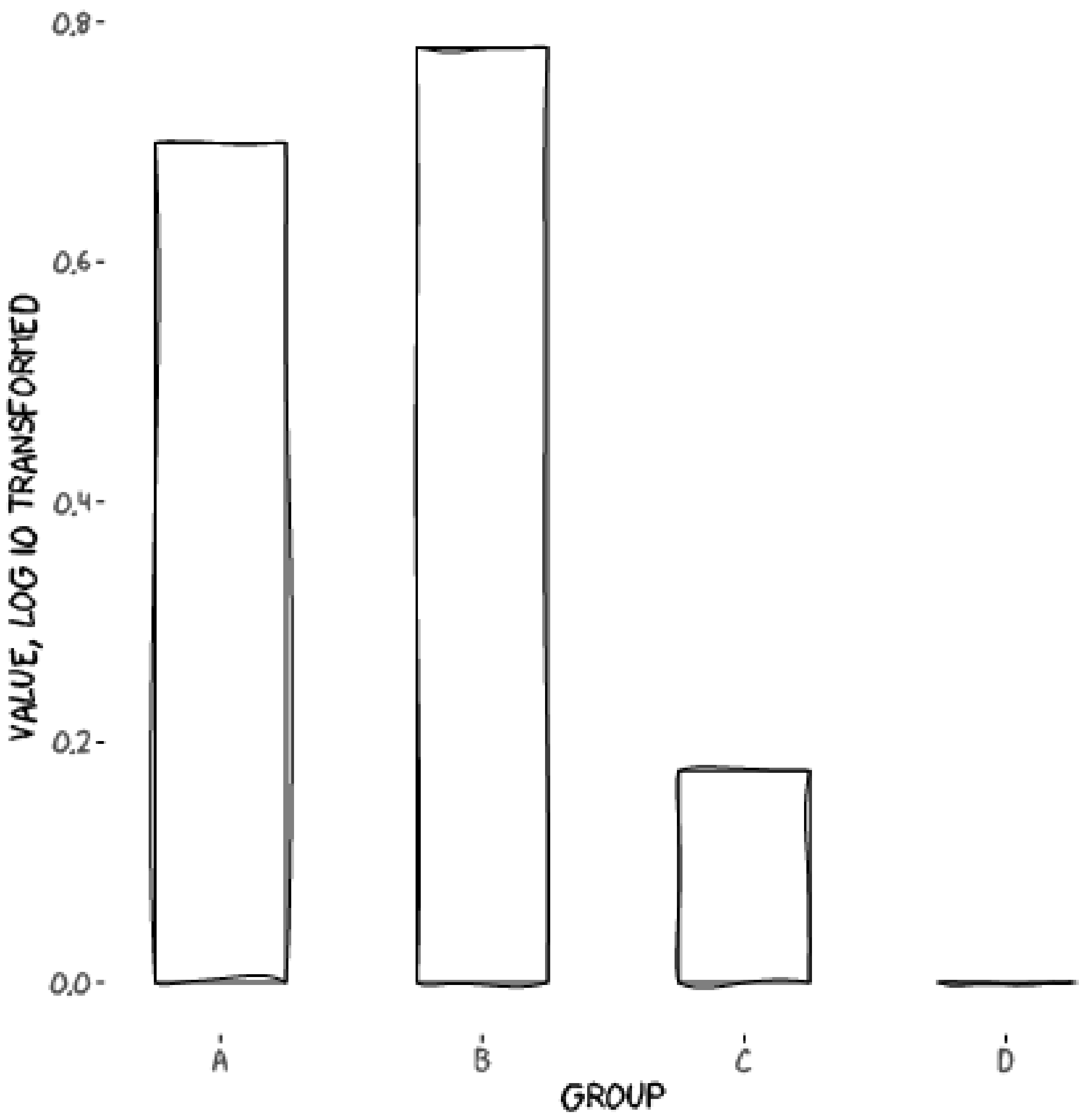

Highest and lowest life expectancies, 2007
30
40
50
60
70
80
90
Japan
82.6
Hong Kong, China
82.2
Iceland
81.8
Switzerland
81.7
Australia
81.2
Spain
80.9
Sweden
80.9
Israel
80.7
France
80.7
Canada
80.7
Liberia
45.7
Central African Republic
44.7
Afghanistan
43.8
Zimbabwe
43.5
Angola
42.7
Lesotho
42.6
Sierra Leone
42.6
Zambia
42.4
Mozambique
42.1
Swaziland
39.6
Source: gapminder


Grand Rapids
Duluth
University Farm
Morris
Crookston
Waseca
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Variety
0
10
20
30
40
50
60
70
Yield (bushels/acre)
Year
1931
1932

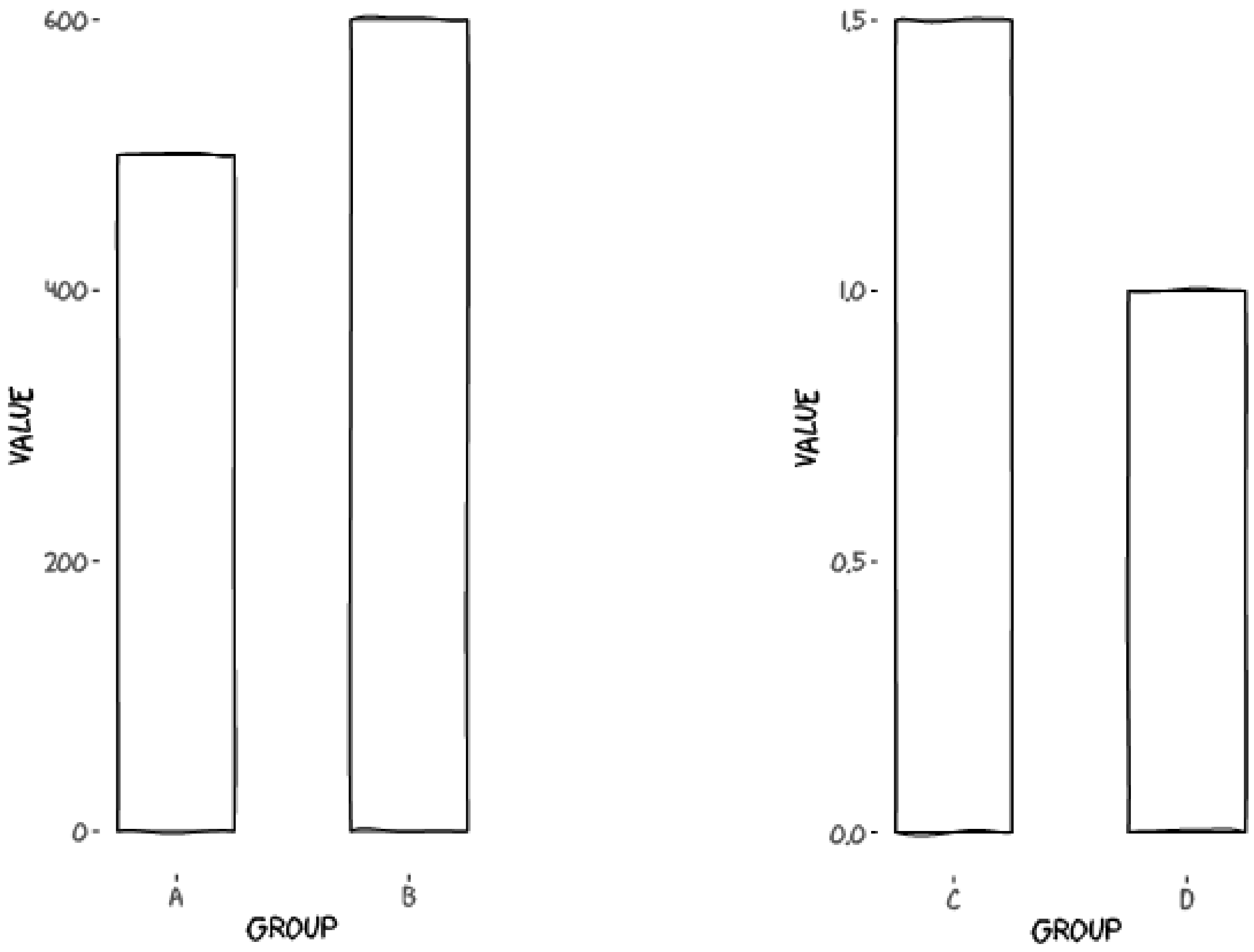
Broken y-axes



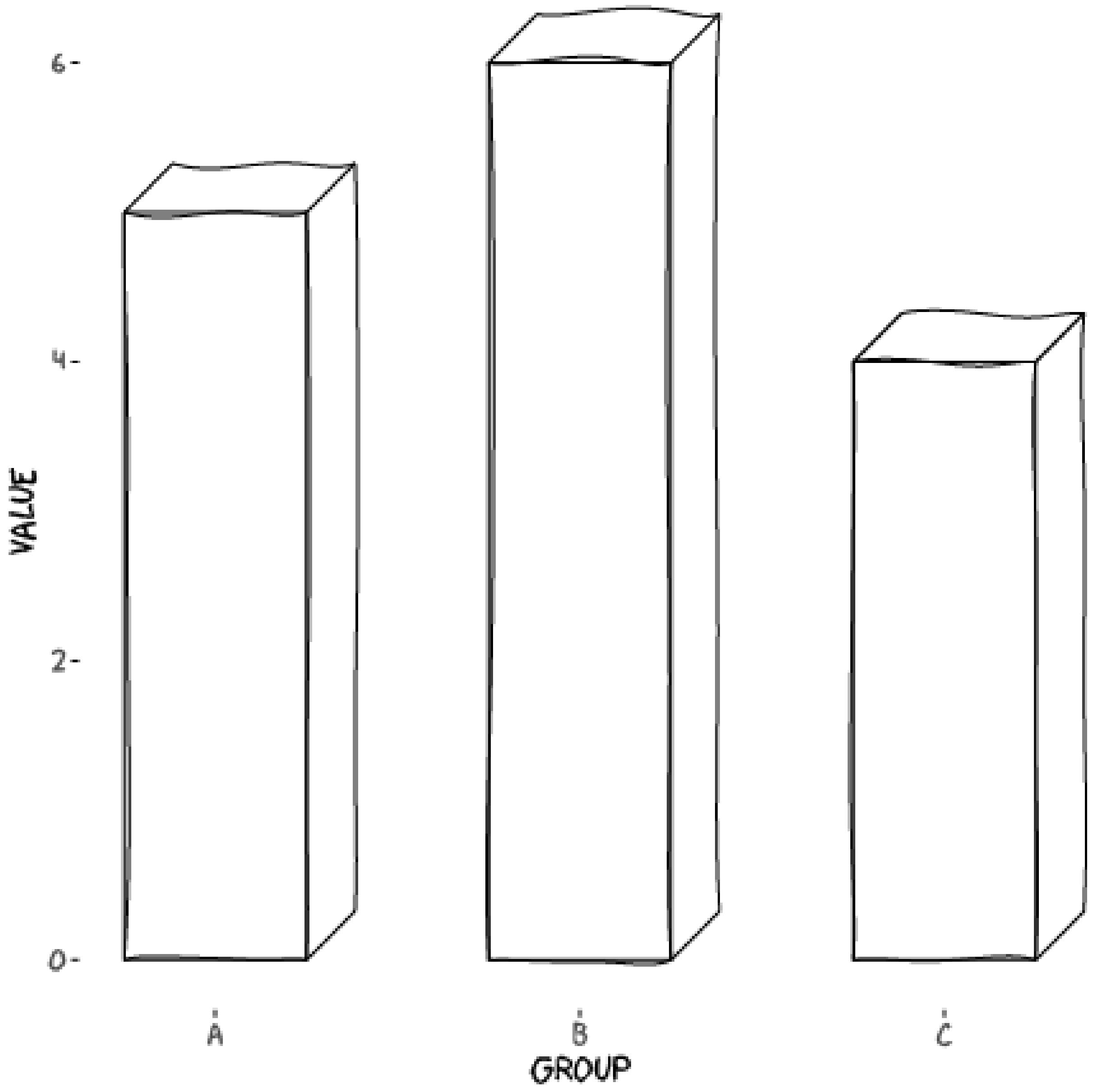
Broken y-axes, replace with transformed data



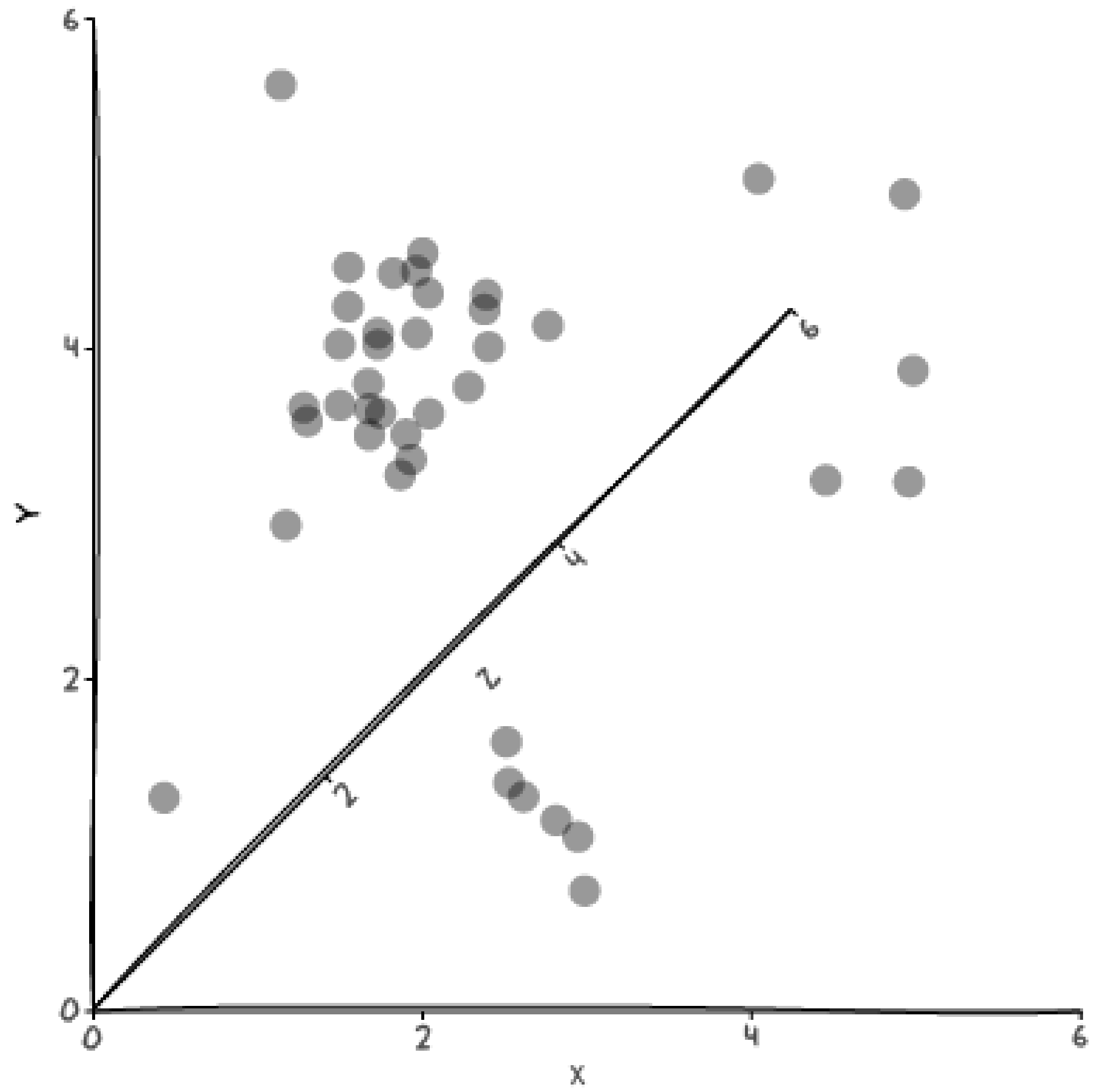
Broken y-axes, use facets



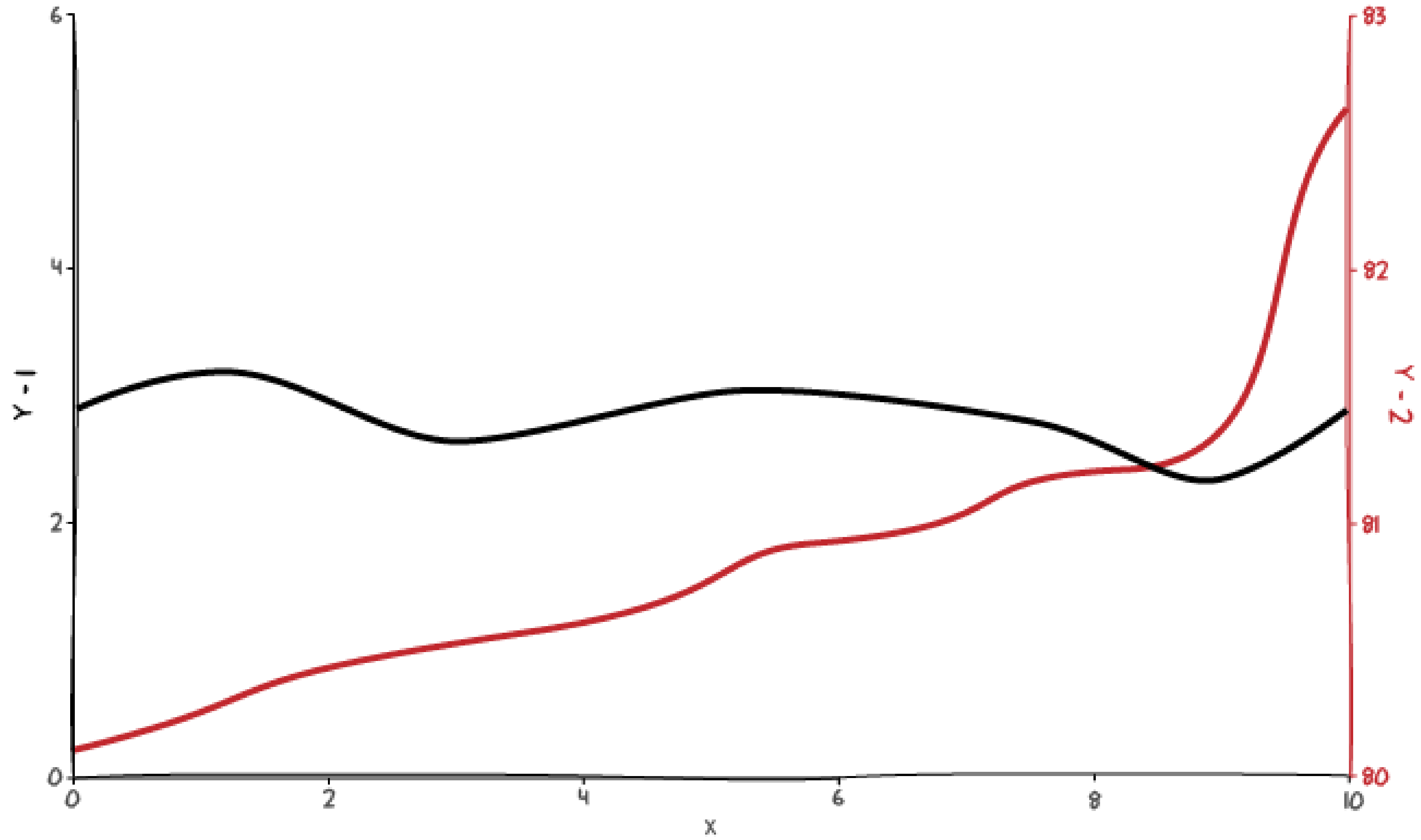
3D plots, without data on the 3rd axis



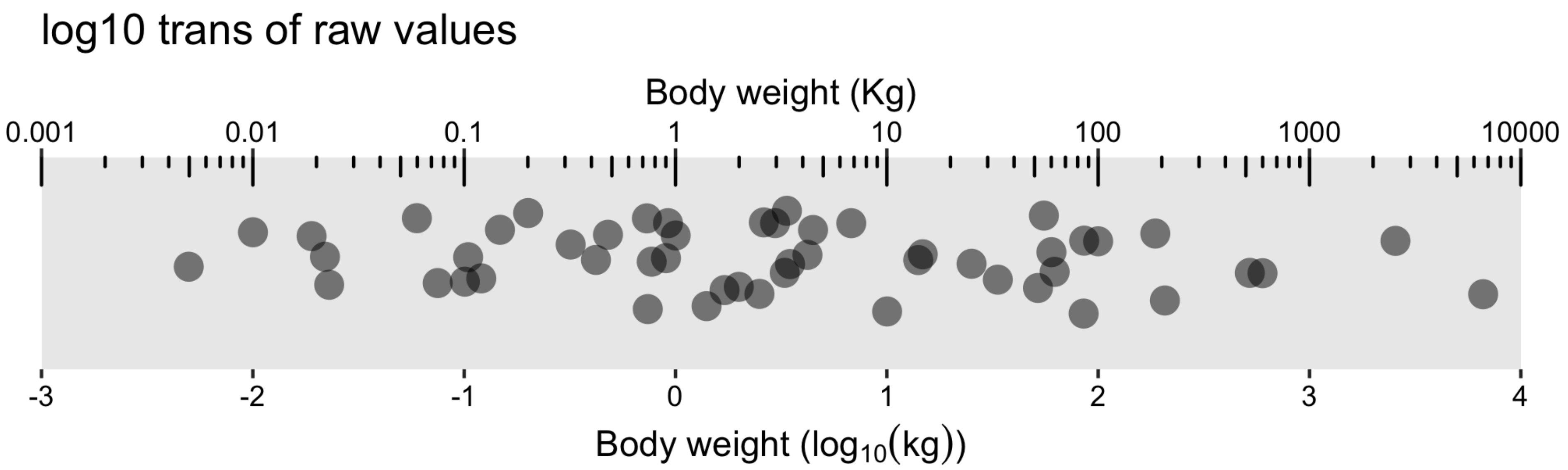
3D plots, with data on the 3rd axis



Double y-axes



Double y-axis for transformations



Guidelines not rules

- Use your common sense:
  - Is there anything on my plot that obscure a clear reading of the data or the take-home message?

Let's practice!

INTERMEDIATE DATA VISUALIZATION WITH GGPLOT2